Statistica

Capitolo 7

Campionamento e
Distribuzioni Campionarie

Obiettivi del Capitolo

Dopo aver completato il capitolo, sarete in grado di:

- Descrivere un campione casuale semplice e spiegare perché il campionamento è importante
- Spiegare la differenza tra statistica descrittiva e statistica inferenziale
- Definire il concetto di distribuzione campionaria
- Determinare la media e lo scarto quadratico medio per la distribuzione della media campionaria $\overline{X}$
- Enunciare il Teorema del limite centrale e capire la sua importanza
- Determinare la media e lo scarto quadratico medio della distribuzione della proporzione campionaria $\hat{P}$
- Descrivere la distribuzione della varianza campionaria

Strumenti della Statistica Economico-Aziendale

- **Statistica Descrittiva**
  - Raccogliere, presentare, e descrivere i dati
- **Statistica Inferenziale**
  - Trarre conclusioni e/o prendere decisioni riguardanti una popolazione sulla base dei soli dati campionari

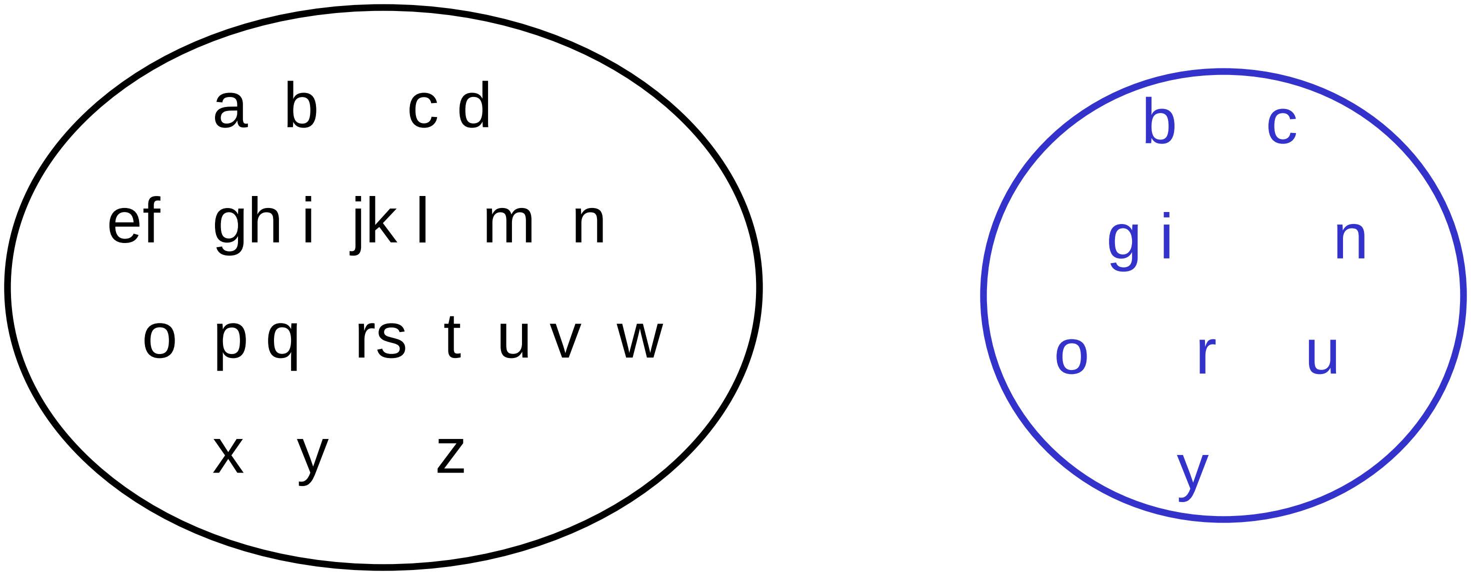
Popolazione e Campione

- 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 distruttivo
    Scontrini selezionati a caso per una verifica

Popolazione vs. Campione

Popolazione **Campione**

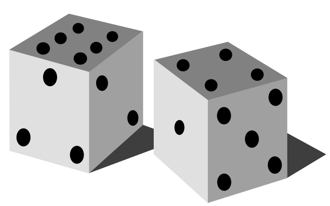


Perché Usare un Campione?

- 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.

Campione Casuale Semplice

- Ciascuna unità della popolazione ha la stessa opportunità di essere scelta
- Le unità sono scelte in modo indipendente
- I campioni possono essere ottenuti usando le tavole dei numeri casuali o i generatori di numeri casuali forniti dai computer



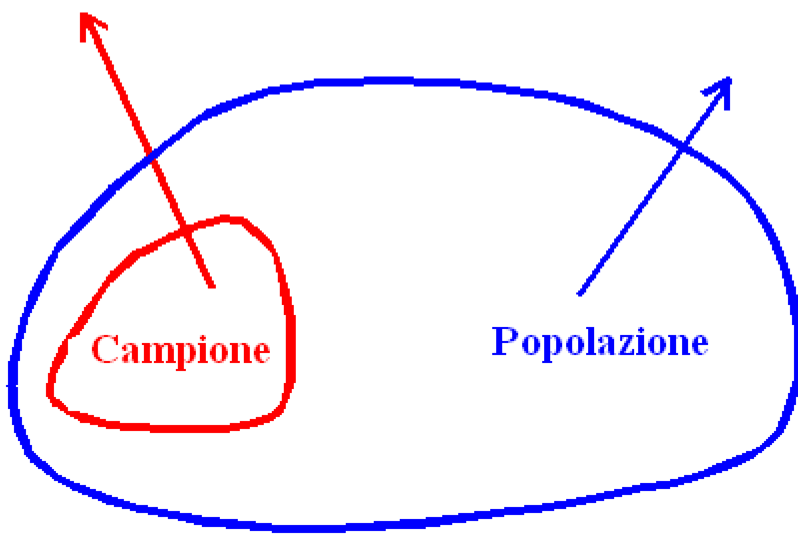
- Un campione casuale semplice è il metodo ideale e costituisce il termine di paragone per gli altri metodi di campionamento

Statistica Inferenziale

- Facciamo inferenza sulla popolazione esaminando i risultati campionari

Statistiche campionarie ➡ Parametri della popolazione

(note) **Inferenza** (non noti, ma possono essere stimati usando un campione)



Statistica Inferenziale

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 di ipotesi**
  - Esempio: usare l'evidenza campionaria per verificare l'affermazione che il peso medio della popolazione sia di 60 Kg

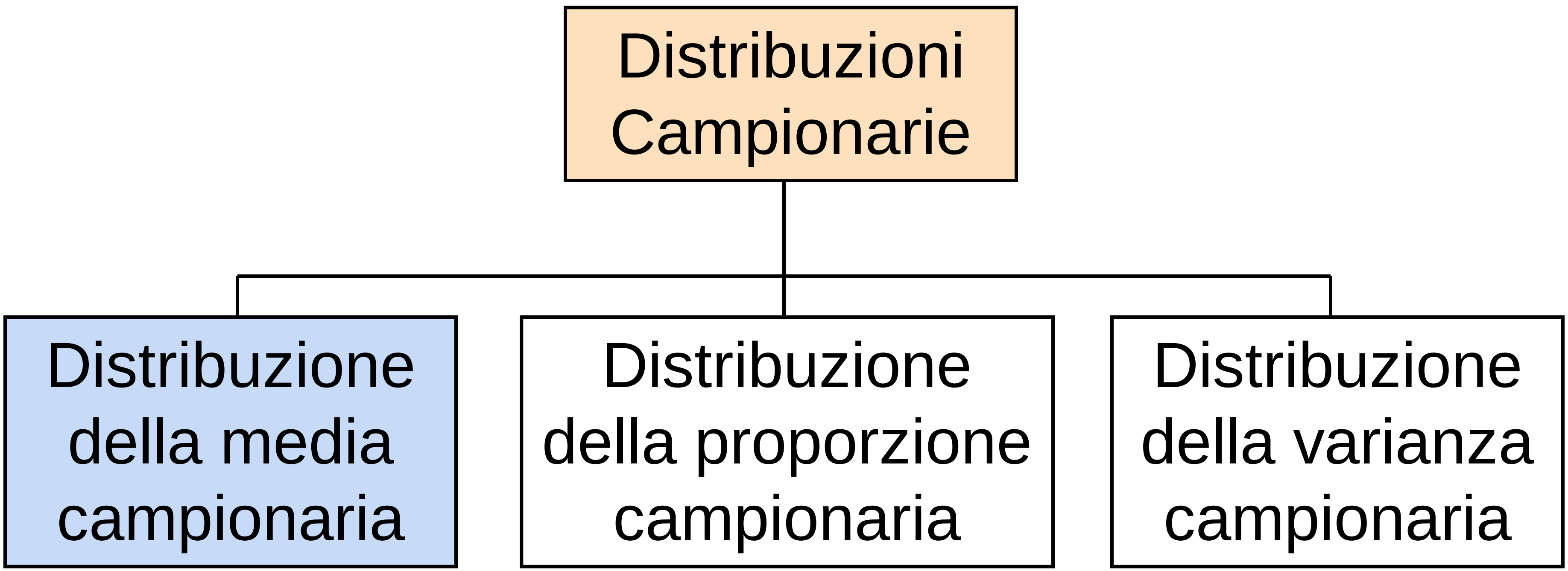
Distribuzione Campionaria

- Una distribuzione campionaria è una distribuzione di tutti i possibili valori di una statistica ottenuti da campioni della stessa ampiezza estratti dalla popolazione

Sommario del Capitolo

- Distribuzioni Campionarie
  - Distribuzione della media campionaria
  - Distribuzione della proporzione campionaria
  - Distribuzione della varianza campionaria

Distribuzione della Media Campionaria



Sviluppo della Distribuzione della Media Campionaria

- **Assumiamo che ci sia una popolazione …**
- Dimensione della popolazione N = 4
- Variabile *X*:

  età degli individui
- Valori di *X*:

  18, 20, 22, 24 (in anni)

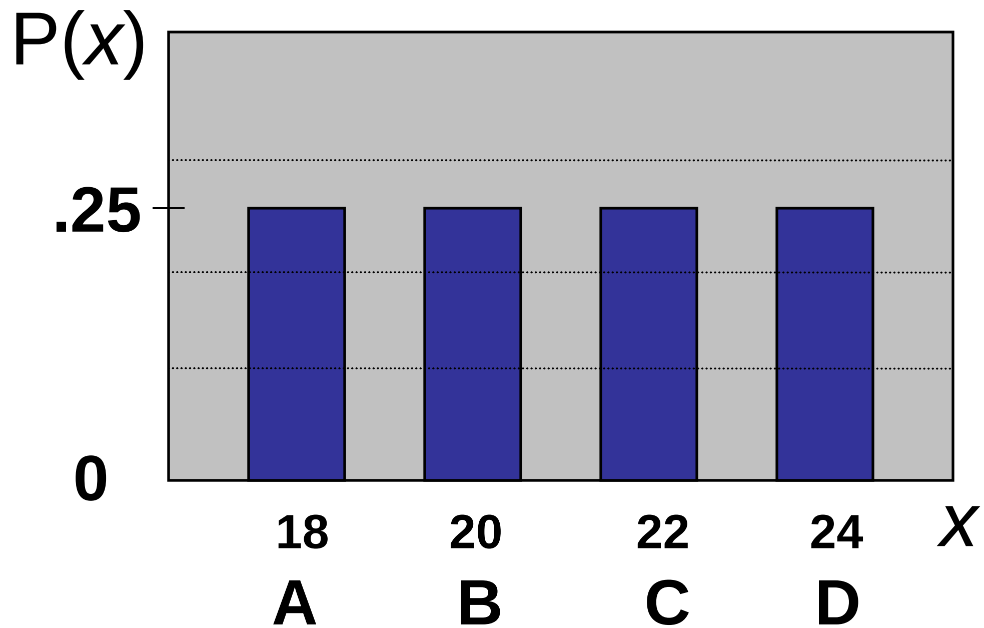
Sviluppo della Distribuzione della Media Campionaria

(continuazione)

Misure di sintesi della distribuzione della Popolazione:

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

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



Distribuzione Uniforme

Sviluppo della Distribuzione della Media Campionaria

(continuazione)

Adesso consideriamo tutti i possibili campioni di dimensione n = 2

1ª Oss	2ª 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

1ª Oss	2ª 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 della Media Campionaria

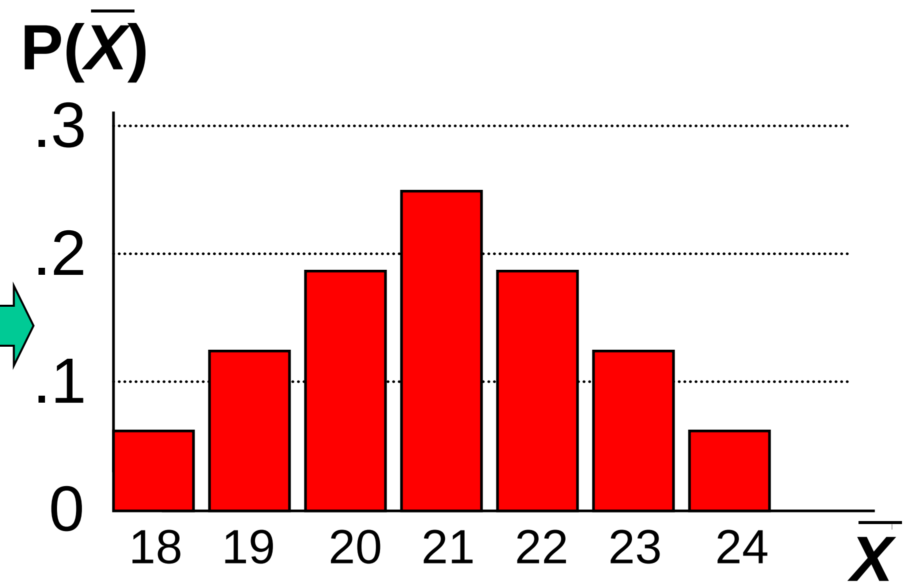
(continuazione)

Distribuzione di tutte le medie campionarie

16 Medie Campionarie

1ª Oss	2ª 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

Distribuzione Medie Campionarie



(non è più uniforme)

Sviluppo della Distribuzione della Media Campionaria

(continuazione)

Misure di sintesi della distribuzione campionaria:

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

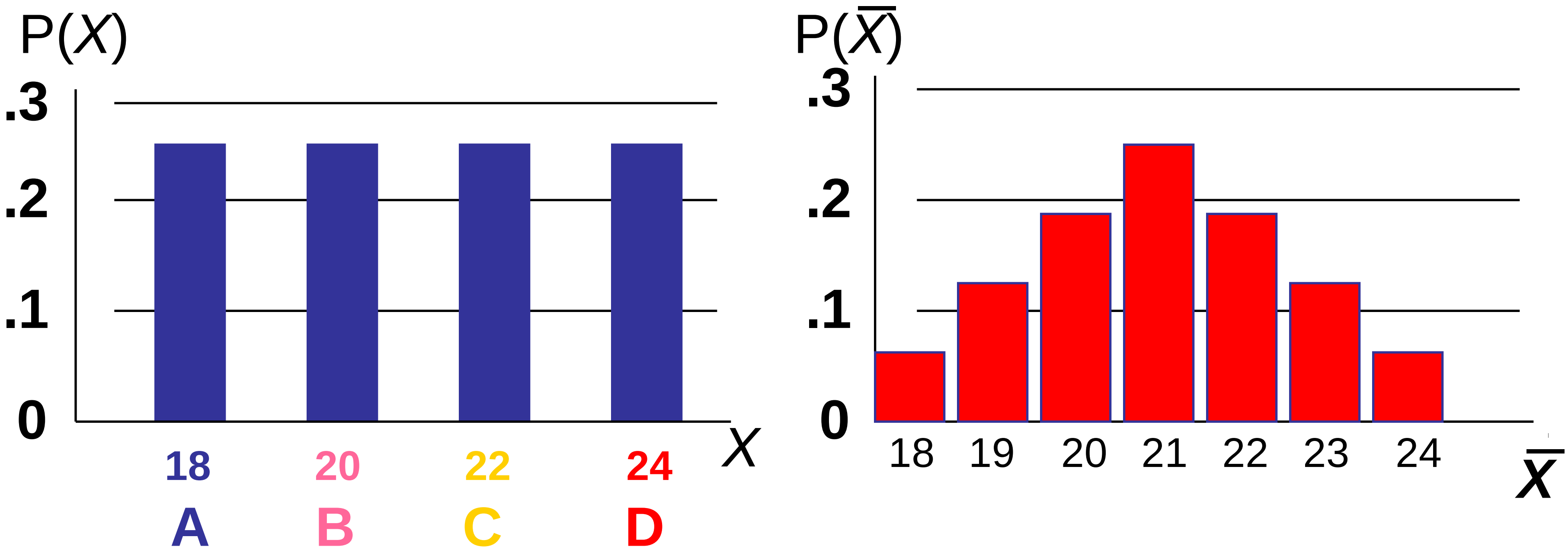
$$\sigma_{\bar{X}} = \sqrt{\frac{\sum(\bar{x}_i - \mu)^2}{N}}$$

$$= \sqrt{\frac{(18-21)^2 + (19-21)^2 + \cdots + (24-21)^2}{16}} = 1.58$$

Confronto tra Distribuzioni: Popolazione e Media Campionaria

Popolazione
$N = 4$
$\mu = 21 \quad \sigma = 2.236$

Distrib. Media Campionaria
$n = 2$
$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$



Media Campionaria

- Sia $(X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione

- La media campionaria è definita come:

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

Scarto Quadratico Medio della Media Campionaria

- Campioni diversi della stessa dimensione estratti dalla stessa popolazione produrranno in genere medie campionarie diverse
- Una misura della variabilità nel valore della media da campione a campione è data dall' Errore Standard della Media:

$$\boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$$

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

Se la Popolazione è 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

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

e

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

Z per la Distribuzione della Media Campionaria

- Z per la distribuzione di $\bar{X}$:

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

dove: $\bar{X}$ = media campionaria

μ = media della popolazione

σ = scarto quadratico medio della popolazione

n = dimensione del campione

Fattore di Correzione per Popolazioni Finite

- Applicare il **Fattore di Correzione per Popolazioni Finite se**:
  - un elemento della popolazione non può essere incluso nel campione più di una volta (il campione è senza reintroduzione), e
  - il campione è ampio rispetto alla popolazione

    (n è superiore al 5% di N)
- Allora

$$\boxed{\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}}$$

oppure

$$\boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

Fattore di Correzione per Popolazioni Finite

(continuazione)

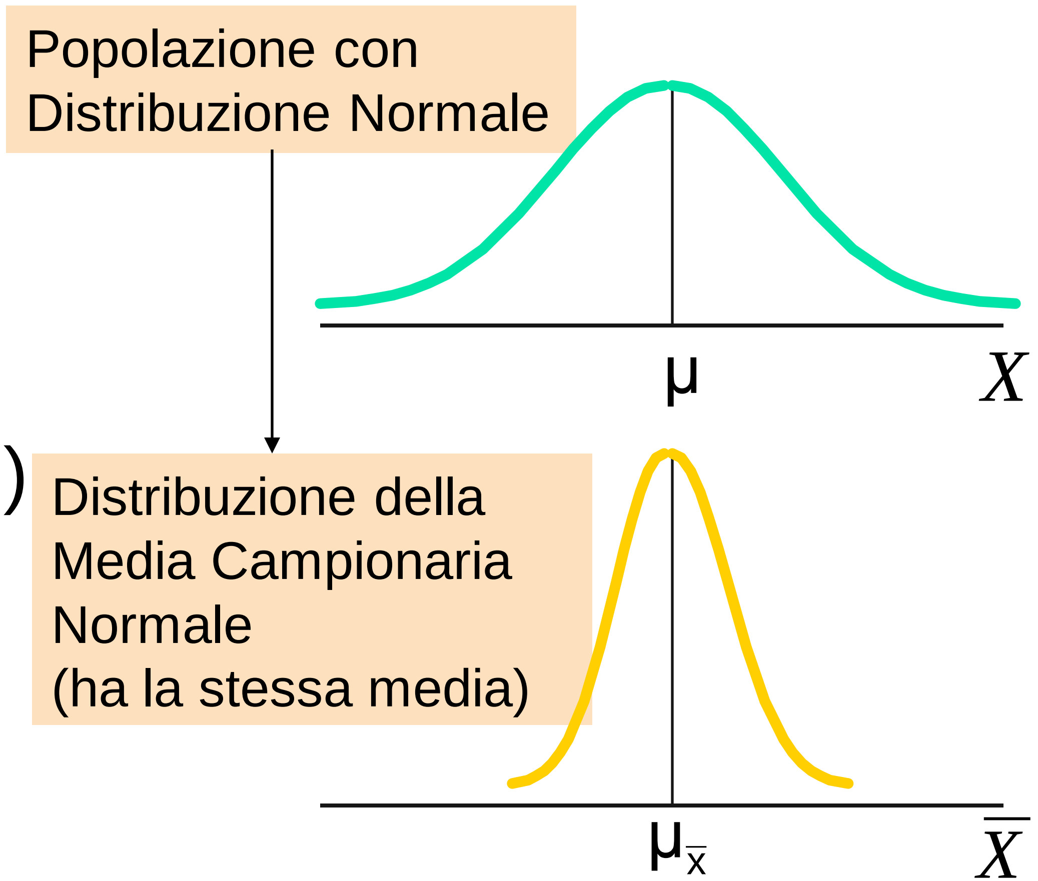
- Se la dimensione del campione n non è abbastanza piccola rispetto alla dimensione della popolazione N, allora vale

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

Proprietà della Media Campionaria

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

(i.e. $\bar{X}$ è non distorto)



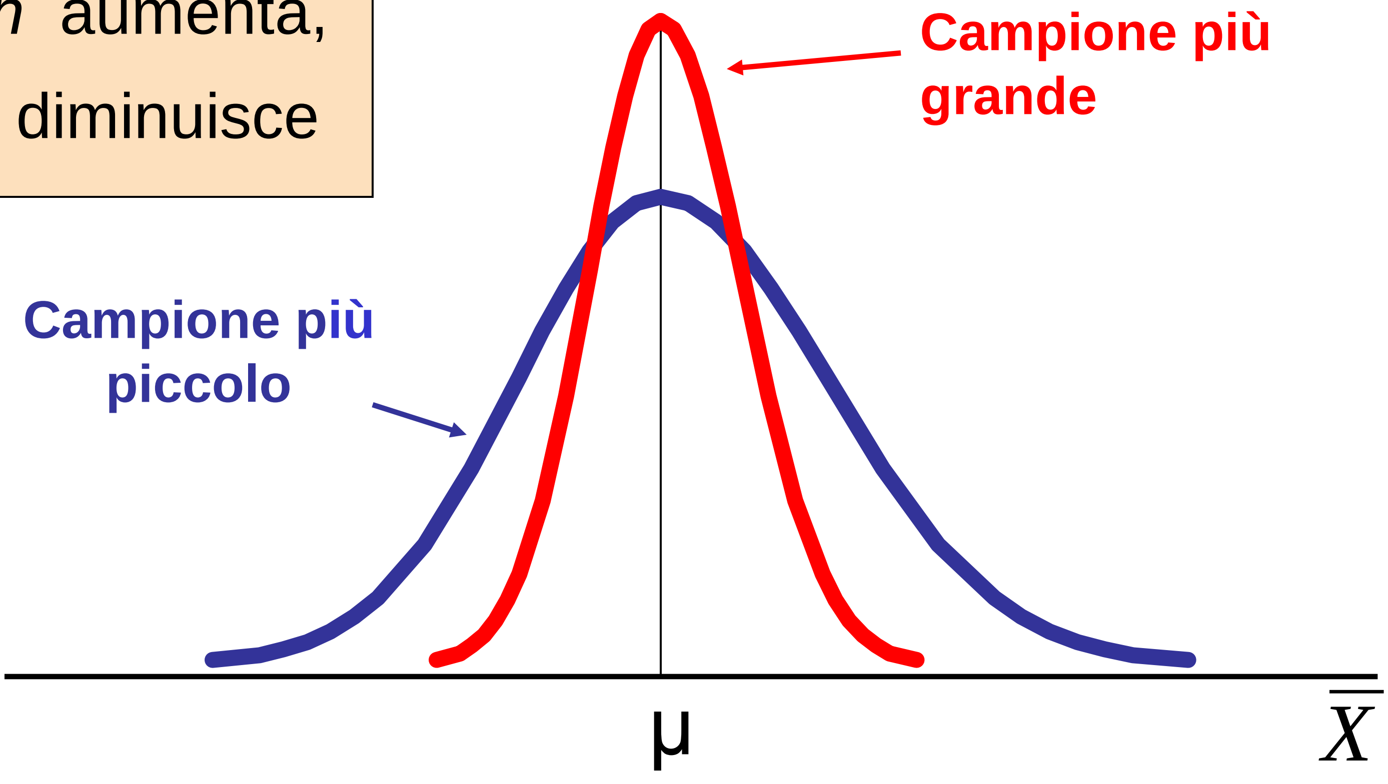
Proprietà della Media Campionaria

(continuazione)

- Per campionamenti con reintroduzione:

Se n aumenta,

$\sigma_{\bar{x}}$ diminuisce



Se la Popolazione **non** è Normale

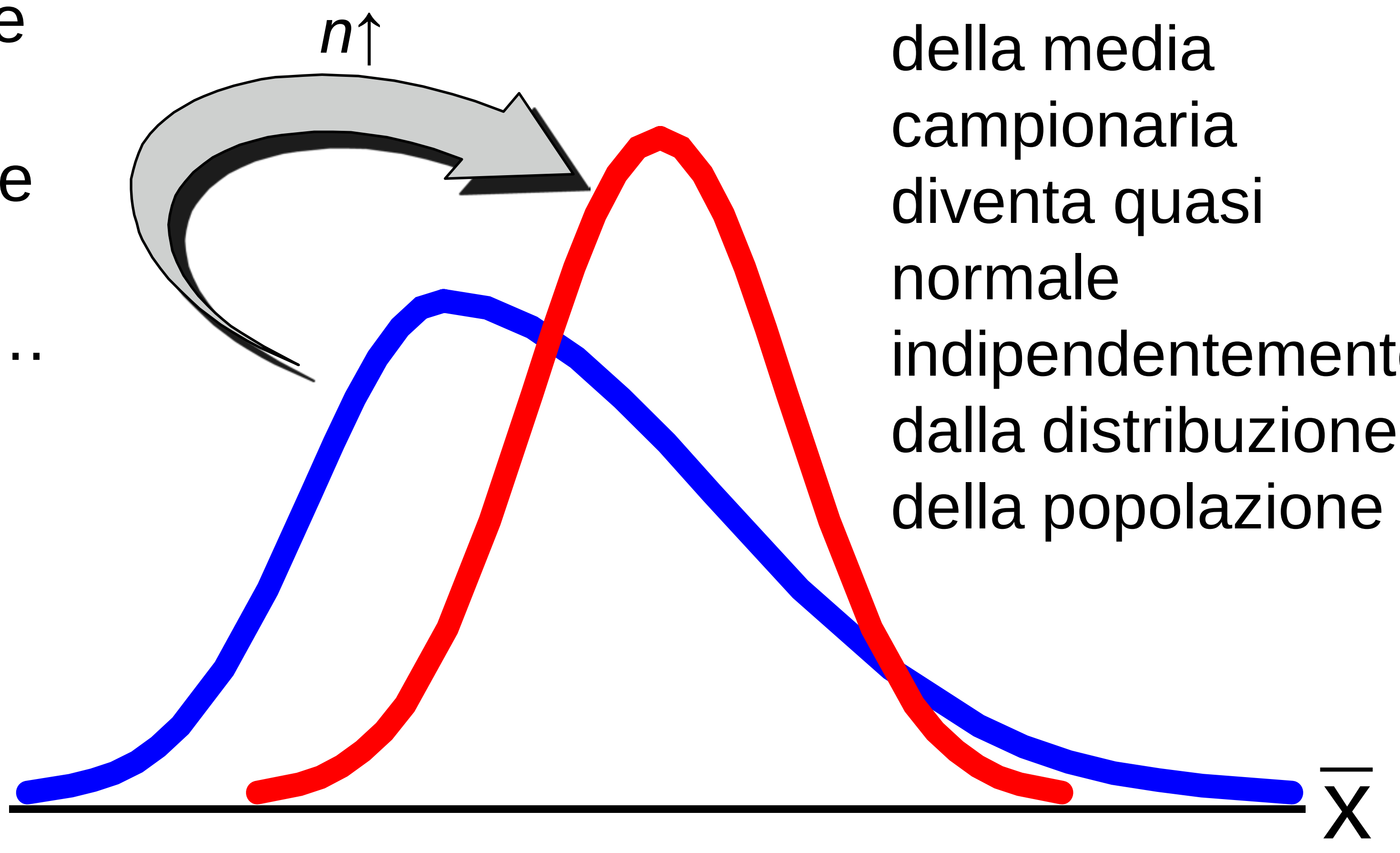
- Possiamo applicare il Teorema del limite centrale:
  - Anche se la popolazione non è normale,
  - …la media campionaria della popolazione sarà approssimativamente normale, purché l'ampiezza del campione sia abbastanza grande.

Proprietà della distribuzione della Media Campionaria:

$$\boxed{\mu_{\bar{x}} = \mu} \quad \text{e} \quad \boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

Teorema del Limite Centrale

Al crescere della dimensione del campione…



la distribuzione della media campionaria diventa quasi normale indipendentemente dalla distribuzione della popolazione

Se la Popolazione **non** è Normale

(continuazione)

Proprietà della distribuzione della Media Campionaria:

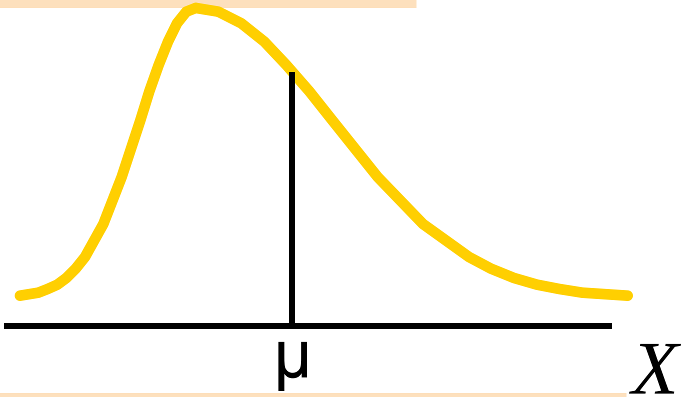
Tendenza Centrale

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

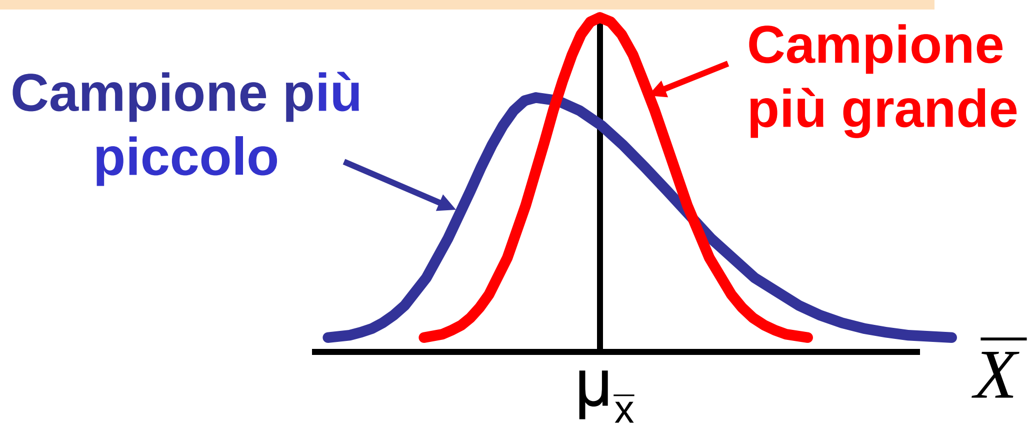
Variabilità

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

Distribuzione Popolazione



Distribuzione della Media 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 ampiezza $n = 36$.

- Qual è la probabilità che la media campionaria sia compresa fra 7.8 e 8.2?

Esempio

(continuazione)

Soluzione:

- Anche se la popolazione non ha distribuzione normale, può essere usato il teorema del limite centrale (n > 25)
- … quindi la distribuzione campionaria di $\bar{X}$ è approssimativamente normale
- … con media $\mu_{\bar{x}} = 8$
- …e scarto quadratico medio $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

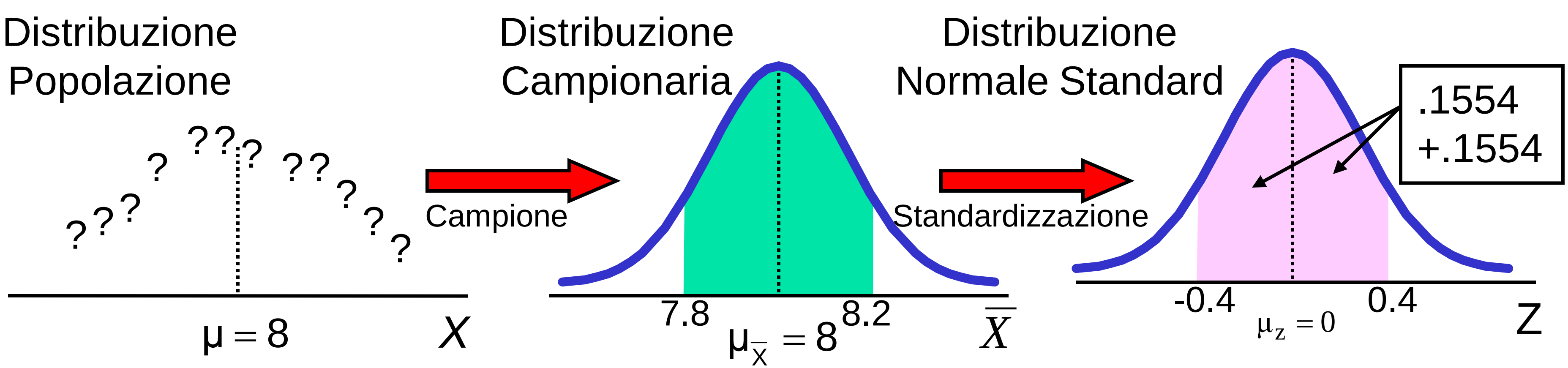
Esempio

(continuazione)

Soluzione (continuazione):

$$P(7.8 < \mu_{\bar{X}} < 8.2) = P\left(\frac{7.8 - 8}{3/\sqrt{36}} < \frac{\mu_{\bar{X}} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}} \right)$$

$$= P(-0.4 < Z < 0.4) = \boxed{0.3108}$$



Intervalli di Accettazione

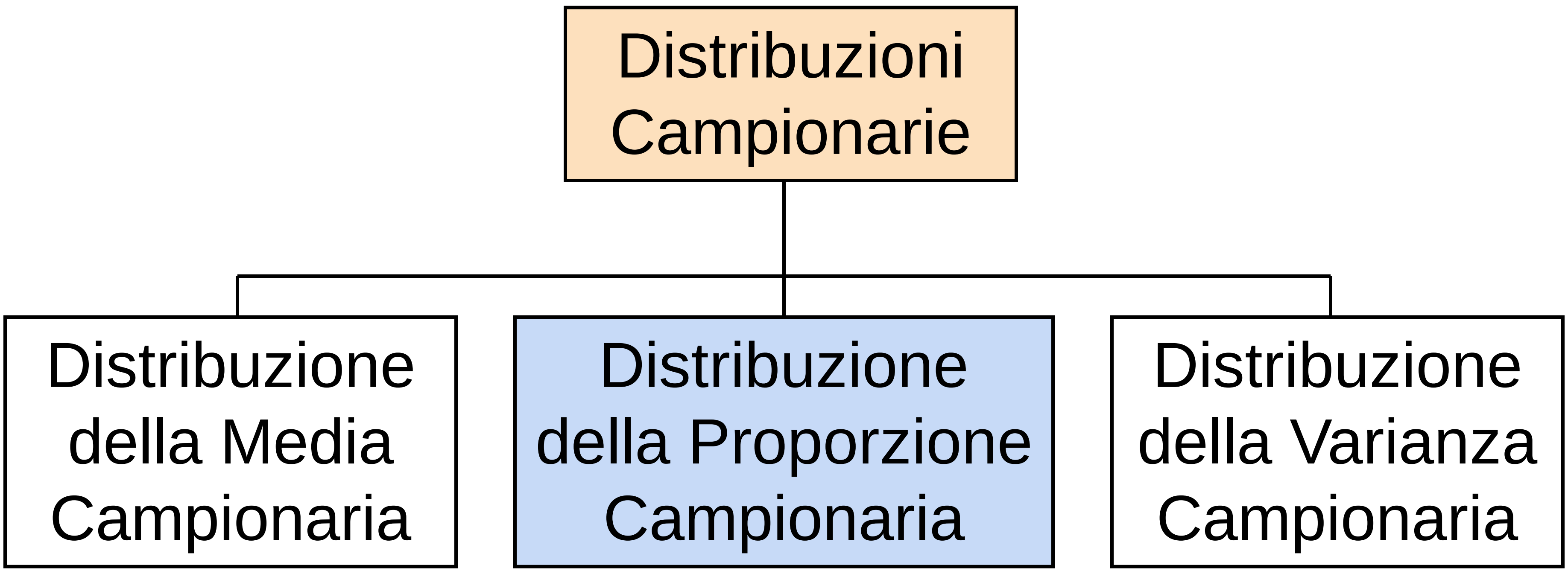
- Obiettivo: determinare un intervallo entro il quale verosimilmente cadono i valori delle medie campionarie, per una data media e varianza della popolazione

  - 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 α/2 (i.e., l'intervallo da - $z_{\alpha/2}$ a $z_{\alpha/2}$ racchiude una probabilità 1 – α)
  - Allora

$$\boxed{\mu \pm z_{\alpha/2}\sigma_{\bar{X}}}$$

è l'intervallo che include $\bar{X}$ con probabilità 1 – α

Distribuzione della Proporzione Campionaria



Proporzione della Popolazione, *p*

p = proporzione della popolazione che possiede le caratteristiche oggetto di studio

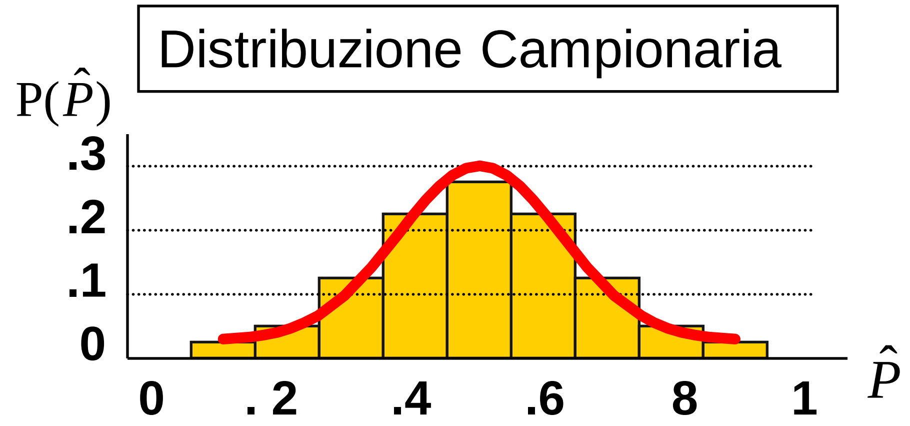
- Proporzione campionaria ($\hat{p}$) fornisce una stima di p:

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

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

Distribuzione della Proporzione Campionaria, $\hat{P}$

- Approssimazione Normale:



Proprietà:

$$E(\hat{P}) = p$$

e

$$\sigma^2_{\hat{P}} = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{p(1-p)}{n}$$

(dove p = proporzione della popolazione)

Z per la distribuzione della Proporzione Campionaria

Passiamo da $\hat{P}$ a Z usando la formula:

$$Z = \frac{\hat{P} - p}{\sigma_{\hat{P}}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Esempio

- Se la vera proporzione di votanti che sostiene la Proposta A è p = .4, qual è la probabilità che un campione di dimensione 200 produca una proporzione campionaria compresa tra .40 e .45?

  - i.e.: **se p = .4 e n = 200, quanto vale**

    $P(.40 \leq \hat{P} \leq .45)$?

Esempio

(continuazione)

- **se p = .4 e *n* = 200, quanto vale $P(.40 \leq \hat{P} \leq .45)$?**

Si trova:

$$\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.4(1-.4)}{200}} = .03464$$

In termini della distribuzione normale stardard:

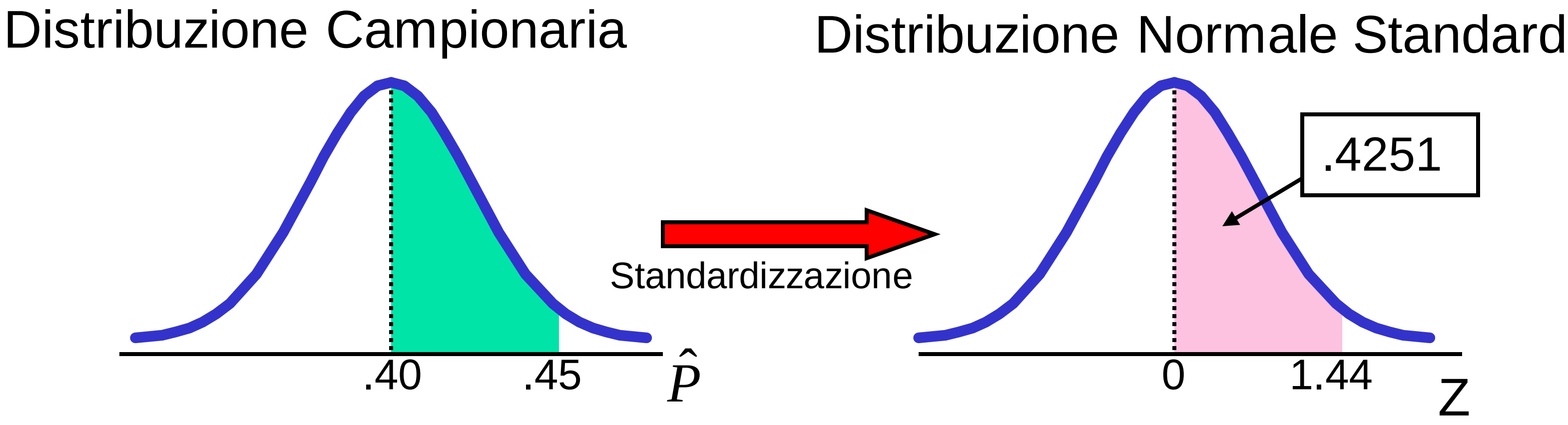
$$P(.40 \leq \hat{P} \leq .45) = P\left(\frac{.40-.40}{.03464} \leq Z \leq \frac{.45-.40}{.03464}\right)$$
$$= P(0 \leq Z \leq 1.44)$$

Esempio

(continuazione)

- **se p = .4 e n = 200, quanto vale $P(.40 \le \hat{P} \le .45)$?**

Dalla tavola della normale standard: $P(0 \le Z \le 1.44) = \boxed{.4251}$



Distribuzione della Varianza Campionaria

- Distribuzioni Campionarie
  - Distribuzione della Media Campionaria
  - Distribuzione della Proporzione Campionaria
  - Distribuzione della Varianza Campionaria

Varianza Campionaria

- Sia (X_1, X_2, . . . , X_n) un campione casuale estratto da una popolazione. La varianza campionaria è

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

- La radice quadrata della varianza campionaria è detta deviazione standard campionaria

- La varianza campionaria è diversa per i vari campioni casuali estratti dalla stessa popolazione

Distribuzione della Varianza Campionaria

- La distribuzione campionaria di S^2 ha media σ^2

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

- Se la popolazione ha distribuzione normale, allora

$$\boxed{\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}}$$

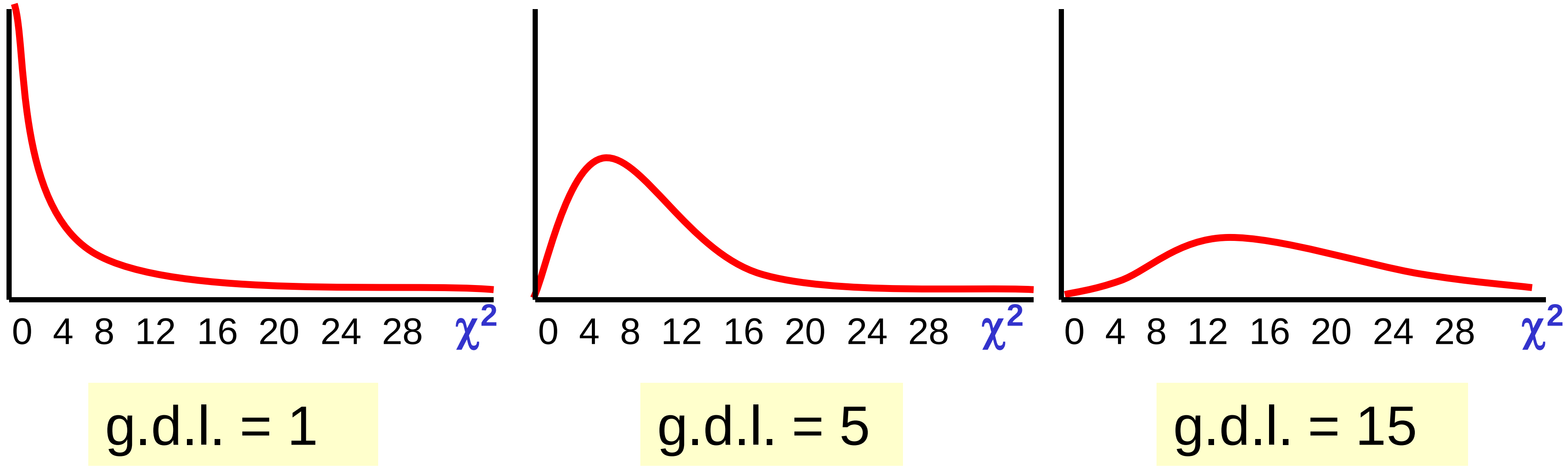
- Se la popolazione ha distribuzione normale allora

$$\boxed{\frac{(n-1)\,S^2}{\sigma^2}}$$

ha una distribuzione χ^2 con $n - 1$ gradi di libertà

Distribuzione Chi-Quadrato

- La distribuzione chi-quadrato è una famiglia di distribuzioni, che dipende dai gradi di libertà:
- g.d.l. = $n - 1$



- La Tavola 3 dell'Appendice contiene i quantili della distribuzione chi-quadrato

Gradi di Libertà (g.d.l.)

Idea: il numero di osservazioni che sono libere di variare dopo che è stata calcolata la media campionaria

Esempio: Supponiamo la media di 3 numeri sia 8

Sia $x_1 = 7$
Sia $x_2 = 8$
Quanto vale X_3?

Se la media di questi tre valori è 8, allora x_3 deve essere 9 (i.e., x_3 non è libero di variare)

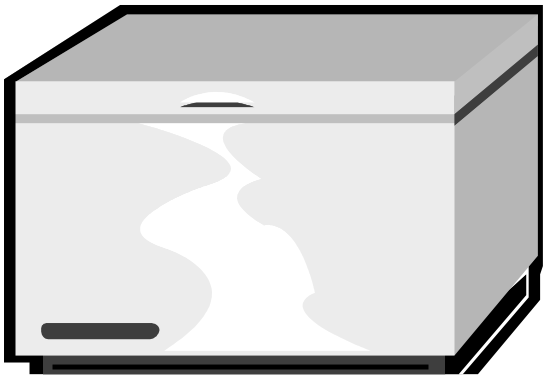
Qui $n = 3$, allora gradi di libertà $= n – 1 = 3 – 1 = 2$

(2 osservazioni possono assumere qualsiasi valore, ma dato il valore della media campionaria, la terza non è libera di variare)

Esempio: Chi-Quadrato

- Un congelatore industriale deve mantenere la temperatura selezionata con bassa variabilità. Si richiede una deviazione standard non superiore a 4 gradi (una varianza di 16 gradi^2).

- Viene controllato un campione di 14 congelatori
- Data una deviazione standard della popolazione di 4, qual è il limite superiore (*K*) per la varianza campionaria che ha probabilità pari a 0.05 di essere superato?



Esempio: Chi-Quadrato

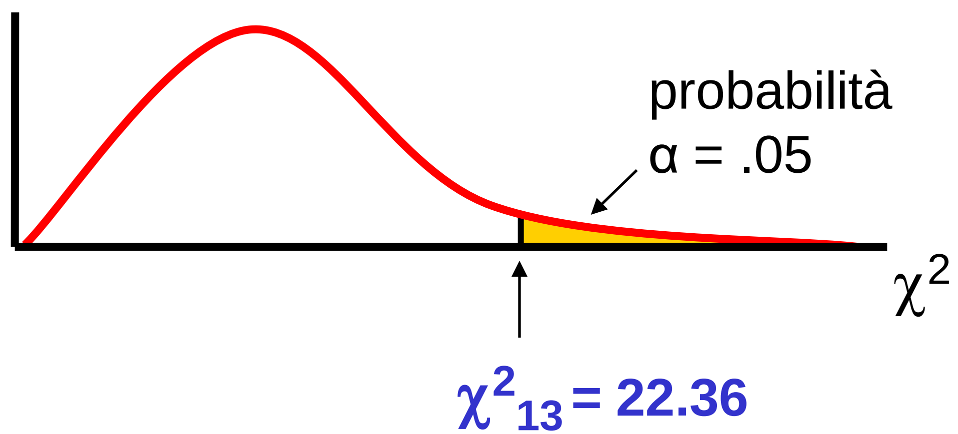
$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

Distribuzione chi-quadrato con

(n – 1) = 13 gradi di libertà

- Usiamo la distribuzione chi-quadrato con area 0.05 nella coda destra:

χ^2_{13} = 22.36 (α = .05 e 14 – 1 = 13 g.d.l.)



Esempio: Chi-Quadrato

(continuazione)

χ^2_{13} = 22.36 (α = .05 e 14 – 1 = 13 g.d.l.)

Allora:

$$P(S^2 > K) = P\left(\frac{(n-1)S^2}{16} > \chi^2_{13}\right) = 0.05$$

oppure $$\frac{(n-1)K}{16} = 22.36$$ (dove *n* = 14)

Allora $$K = \frac{(22.36)(16)}{(14-1)} = 27.52$$

Se, sulla base di un campione casuale di dimensione *n* = 14, si osservasse s^2 maggiore di 27.52, ci sarebbero buoni motivi per ipotizzare una varianza della popolazione superiore a 16.

Riepilogo del Capitolo

- Introdotte le distribuzioni campionarie
- Descritta la distribuzione della media campionaria
  - Per popolazioni con distribuzione normale
  - Usando il Teorema del limite centrale
- Descritta la distribuzione della proporzione campionaria
- Introdotta la distribuzione chi-quadrato
- Esaminata la distribuzione della varianza campionaria
- Calcolate probabilità usando le distribuzioni campionarie