Distribuzione campionaria

Tecniche di Ricerca Psicologica e di Analisi dei Dati

Corrado Caudek

A.A. 2007-2008

Indice

1 Inferenza statistica

- L'inferenza statistica è il processo che consente di formulare delle conclusioni relative ad una popolazione sulla base di un campione di osservazioni estratte a caso dalla popolazione.
- Centrale all'inferenza statistica classica è la nozione di distribuzione campionaria, ovvero la descrizione di come variano le statistiche dei campioni, se campioni casuali aventi la stessa grandezza n vengono ripetutamente estratti dalla popolazione.

- Anche se, in ciascuna applicazione pratica dell'inferenza statistica, il ricercatore dispone solamente di un unico campione casuale di grandezza n, la possibilità che il campionamento venga ripetuto fornisce, in principio, la fondazione concettuale per decidere quanto il campione osservato sia informativo della popolazione nel suo complesso.

1.1 Parametri e statistiche

Si ricordi che

- Un parametro è un numero che descrive un qualche aspetto della popolazione.
  - Per esempio, il reddito italiano medio μ è un parametro. Supponiamo che $\mu = €43,236$.
  - In qualsiasi situazione concreta, i parametri sono sconosciuti.

- Una statistica è un numero che può essere calcolato utilizzando i dati forniti da un campione, senza alcuna conoscenza dei parametri della popolazione.
  - Supponiamo che, per un campione casuale di $n = 1000$ famiglie italiane, il reddito medio sia uguale a €$42,586$. La media del campione $\bar{x}$=€$42,586$ è una statistica.

1.2 Stima e test d'ipotesi

- Solitamente, non siamo interessati alle statistiche in sè, ma a quello che le statistiche ci dicono della popolazione.
  - Potremmo usare la media di un campione di famiglie italiane, per esempio, per stimare il reddito medio (sconosciuto) della popolazione.
  - Oppure, potremmo usare la media del campione per stabilire se il reddito medio italiano sia mutato dall'ultimo censimento.
- Questi due tipi di domande sono propri dei due pricipali approcci all'inferenza statistica classica:
  1. la stima di parametri;
  2. il test d'ipotesi statistiche.

1.3 Variabilità campionaria

Un aspetto fondamentale delle statistiche campionarie riguarda il fatto che variano da campione a campione.

- Nel caso dell'esempio precedente, sarebbe molto improbabile trovare, per un secondo campione casuale di 1000 famiglie italiane, un reddito medio esattamente uguale a €$42,586$.

La variazione di una statistica campionaria da campione a campione viene detta variabilità campionaria.

- Quando la variabilità campionaria è molto grande, il campione è poco informativo a proposito del parametro della popolazione.
- Quando la variabilità campionaria è piccola, invece, la statistica del campione è informativa del parametro della popolazione, anche se è impossibile che la statistica di un qualsiasi campione sia esattamente uguale al parametro della popolazione.

2 Simulazione 1

La variabilità campionaria verrà illustrata nel modo seguente:

1. verrà considerata una variabile discreta che può assumere soltanto un piccolo numero di valori possibili ($N = 4$);
2. verrà fornito l'elenco di tutti i possibili campioni di grandezza $n = 2$;
3. verrà calcolata la media di ciascuno dei possibili campioni di grandezza $n = 2$;
4. verrà esaminata la distribuzione delle medie di tutti i possibili campioni di grandezza $n = 2$.

La media μ e la varianza σ della popolazione verranno calcolate.

- μ e σ sono dei parametri, mentre la media $\bar{x}_i$ e la varianza s_i^2 di ciascun campione sono delle statistiche.

- L'esperimento di questo esempio consiste in $n = 2$ estrazioni con rimessa di una pallina x_i da un'urna che contiene $N = 4$ palline.
- Le palline sono numerate nel modo seguente:

$$\{2, 3, 5, 9\}$$

- L'estrazione con rimessa corrisponde ad una popolazione di grandezza infinita (è sempre possibile infatti estrarre una nuova pallina dall'urna).

Per ciascun campione di grandezza $n = 2$ viene calcolata la media dei valori delle palline estratte $\bar{x} = \sum_{i=1}^{2} x_i/2$.

- Per esempio, se le palline estratte sono $x_1 = 2$ e $x_2 = 3$, allora

$$\bar{x} = (2 + 3)/2 = 5/2 = 2.5$$

2.1 Tre distribuzioni

Dobbiamo distinguere tre distribuzioni:

1. la distribuzione della popolazione,
2. la distribuzione di un particolare campione,
3. la distribuzione campionaria delle medie di tutti i possibili campioni.

2.1.1 Distribuzione della popolazione

Distribuzione della popolazione: la distribuzione di X (il valore della pallina estratta) nella popolazione. In questo caso la popolazione è infinita e ha la seguente distribuzione di probabilità:

x_i	p_i
2	$\frac{1}{4}$
3	$\frac{1}{4}$
5	$\frac{1}{4}$
9	$\frac{1}{4}$
somma	1.0

- La media della popolazione è

$$\mu = \sum x_i p_i = 4.75$$

- La varianza della popolazione è

$$\sigma^2 = \sum (x_i - \mu)^2 p_i = 7.1875$$

2.1.2 Distribuzione di un campione

Distribuzione di un campione: la distribuzione di X in un particolare campione.

- Per esempio, se $x_1 = 2$ e $x_2 = 3$, allora la media di questo campione sarà $\bar{x} = 2.5$ e la varianza sarà $s^2 = 0.5$.

2.1.3 Distribuzione campionaria della media

Distribuzione campionaria della media: la distribuzione delle medie di tutti i possibili campioni.

- Se $n = 2$, ci sono $4 \times 4 = 16$ possibili campioni. Possiamo dunque elencarli, insieme alle loro medie.

campione	media $\bar{x}_i$	campione	media $\bar{x}_i$
$\{2, 3\}$	2.5	$\{3, 2\}$	2.5
$\{5, 2\}$	3.5	$\{2, 5\}$	3.5
$\{9, 2\}$	5.5	$\{2, 9\}$	5.5
$\{5, 3\}$	4.0	$\{3, 5\}$	4.0
$\{9, 3\}$	6.0	$\{3, 9\}$	6.0
$\{9, 5\}$	7.0	$\{5, 9\}$	7.0
$\{2, 2\}$	2	$\{3, 3\}$	3
$\{5, 5\}$	5	$\{9, 9\}$	9

La distribuzione campionaria della media ha la seguente distribuzione di probabilità:

$\bar{x}_i$	p_i
2.0	$1/16$
2.5	$2/16$
3.0	$1/16$
3.5	$2/16$
4.0	$2/16$
5.0	$1/16$
5.5	$2/16$
6.0	$2/16$
7.0	$2/16$
9.0	$1/16$
somma	1.0

- La media della distribuzione campionaria della media è

$$\mu_{\bar{x}} = \sum \bar{x}_i p_i = 4.75$$

- La varianza della distribuzione campionaria della media è

$$\sigma^2_{\bar{x}} = \sum (\bar{x}_i - \mu_{\bar{x}})^2 p_i = 3.59375$$

- L'esercizio presente ha a che fare con una situazione particolare, quella in cui la distribuzione della popolazione è conosciuta.
- In pratica, la distribuzione della popolazione non è mai conosciuta.

Con questo esercizio possiamo però di notare come la distribuzione campionaria della media possieda due importanti proprietà.

- La media $\mu_{\bar{x}}$ della distribuzione campionaria della media è uguale alla media della popolazione μ.
- La varianza $\sigma^2_{\bar{x}}$ della distribuzione campionaria della media è uguale al rapporto tra la varianza della popolazione σ^2 e la numerosità n del campione:

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{7.1875}{2} = 3.59375$$

Si noti che:

1. la media e la varianza della distribuzione campionaria sono determinate dalla media e varianza della popolazione:

$$\mu_{\bar{x}} = \mu \qquad \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

2. la varianza della distribuzione campionaria della media è più piccola della varianza della popolazione.

In seguito utilizzeremo le proprietà della distribuzione campionaria per fare delle inferenze a proposito dei parametri della popolazione anche quando la distribuzione della popolazione non è conosciuta.

Tre distribuzioni

Si noti inoltre che abbiamo distinto tra tre diverse distribuzioni.

1. **Distribuzione della popolazione**:
   $\Omega = \{2, 3, 5, 9\}$, $\mu = 4.75, \sigma^2 = 7.1875$

2. **Distribuzione di un particolare campione**:
   $\Omega_i = \{2, 3\}$, $\bar{x} = 2.5$, $s^2 = 0.5$

3. **Distribuzione campionaria della media**:
   $\Omega_{\bar{x}} = \{2.5, 3.5, 5.5, 4, 6, 7, 2.5, 3.5, 4, 6, 7, 2, 5, 3, 9\}$,
   $\mu_{\bar{x}} = 4.75, \sigma^2_{\bar{x}} = 3.59375$

Distribuzione della popolazione La distribuzione che contiene tutte le osservazioni. Media e varianza di questa distribuzione si indicano con μ e σ^2.

Distribuzione del campione La distribuzione dei valori della popolazione che fanno parte di un particolare campione casuale di grandezza n. Le singole osservazioni si indicano con $x_1, \ldots, x_n$, e hanno media $\bar{x}$ e varianza s^2.

Distribuzione campionaria delle medie dei campioni La distribuzione di $\bar{x_i}$ per tutti i possibili campioni di grandezza n che si possono estrarre dalla popolazione considerata. Media e varianza della distribuzione campionaria della media si indicano con $\mu_{\bar{x}}$ e $\sigma^2_{\bar{x}}$.

La distribuzione che sta alla base dell'inferenza statistica è la distribuzione campionaria.

Definizione: la distribuzione campionaria di una statistica è la distribuzione dei valori che quella statistica assume in tutti i campioni di numerosità n che possono essere estratti dalla popolazione.

- Si noti che, se in una simulazione consideriamo un numero di campioni minore di quello che teoricamente è possibile, la distribuzione risultante ci fornirà soltanto un'approssimazione alla vera distribuzione campionaria.

3 Simulazione 2

Consideriamo ora un'altro esempio in cui la variabilità campionaria verrà illustrata nel modo seguente:

1. la stessa popolazione dell'esempio precedente verrà usata;
2. utilizzando **R**, verranno estratti con rimessa da questa popolazione 50000 campioni causali di grandezza $n = 2$;
3. verrà calcolata la media di ciascuno di questi campioni di grandezza $n = 2$;
4. verranno calcolate la media e la varianza della distribuzione delle medie dei 50000 campioni di grandezza $n = 2$.

```
N <- 4
n <- 2
nSamples <- 50000
X <- c(2, 3, 5, 9)

Mean <- mean(X)
Var <- var(X)*(N-1)/N

SampDistr <- rep(0, nSamples)

for (i in 1:nSamples){
  samp <- sample(X, n, replace=T)
  SampDistr[i] <- mean(samp)
}

MeanSampDistr <- mean(SampDistr)
VarSampDistr <- var(SampDistr)*(nSamples-1)/nSamples
```

Risultati della simulazione

```
> Mean
[1] 4.75
> Var
[1] 7.1875
> MeanSampDistr
[1] 4.73943
> VarSampDistr
[1] 3.578548
> Var/n
[1] 3.59375
```

- Popolazione: $\mu = 4.75, \sigma^2 = 7.1875$.
- Distribuzione campionaria della media: $\mu_{\bar{x}} = 4.75, \sigma^2_{\bar{x}} = 3.59375$.
- Risultati della simulazione: $\hat{\mu}_{\bar{x}} = 4.73943, \hat{\sigma}^2_{\bar{x}} = 3.578548$.

4 Simulazione 3

In un terzo esempio, considereremo la distribuzione campionaria della media nel caso di una variabile continua.

1. Verrà utilizzata una popolazione teorica distribuita normalmente con media e varianza conosciute: $\mathcal{N}(125, 33)$.
2. Usando **R**, verranno estratti da questa popolazione 50000 campioni causali di grandezza $n = 10$.
3. Verrà calcolata la media di ciascuno di questi campioni di grandezza $n = 10$;
4. Verranno calcolate la media e la varianza della distribuzione delle medie dei 50000 campioni di grandezza $n = 10$.

```
n <- 10
nSamples<- 50000
Mean <- 125
SD <-  sqrt(33)

SampDistr <- rep(0,nSamples)

for (i in 1:nSamples){
  samp <- rnorm(n, Mean, SD)
  SampDistr[i] <- mean(samp)
}

MeanSampDistr <- mean(SampDistr)
VarSampDistr <- var(SampDistr)*(nSamples-1)/nSamples
```

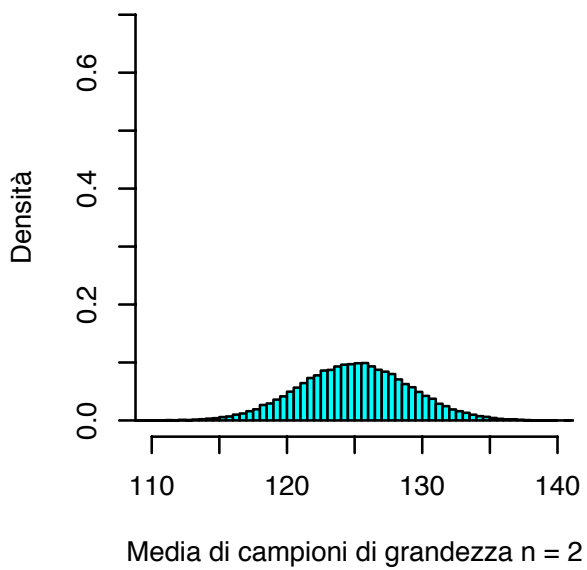
Risultati della simulazione

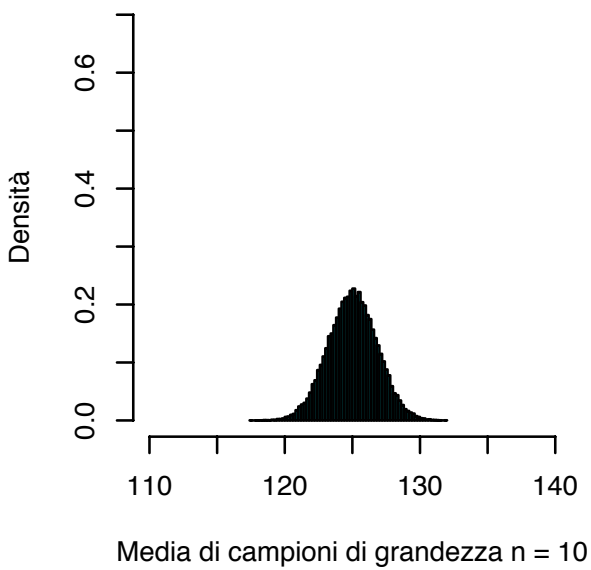
```
> Mean
[1] 125
> Var
[1] 33
> MeanSampDistr
[1] 125.0029
> VarSampDistr
[1] 3.277463
> Var/n
[1] 3.300000
```

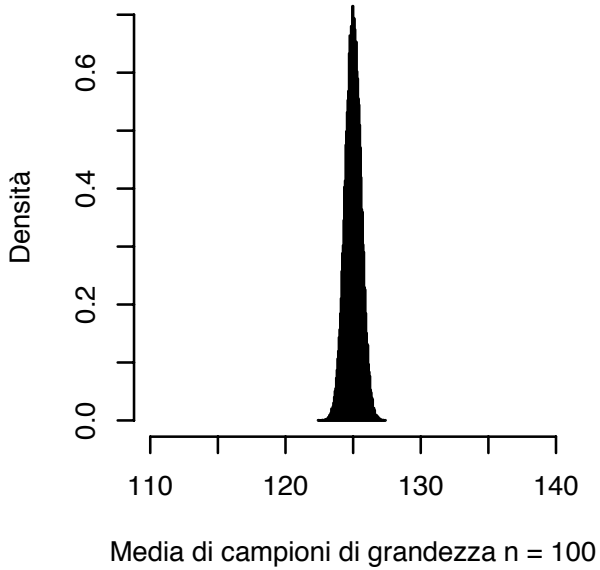
- Popolazione: $\mu = 125, \sigma^2 = 33$.

- Distribuzione campionaria della media: $\mu_{\bar{x}} = 125, \sigma^2_{\bar{x}} = 3.3$.

- Risultati della simulazione: $\hat{\mu}_{\bar{x}} = 125.0029, \hat{\sigma}^2_{\bar{x}} = 3.277463$.


Densità
0.0
0.2
0.4
0.6
110
120
130
140
Media di campioni di grandezza n = 2


Densità
0.0
0.2
0.4
0.6
110
120
130
140
Media di campioni di grandezza n = 10

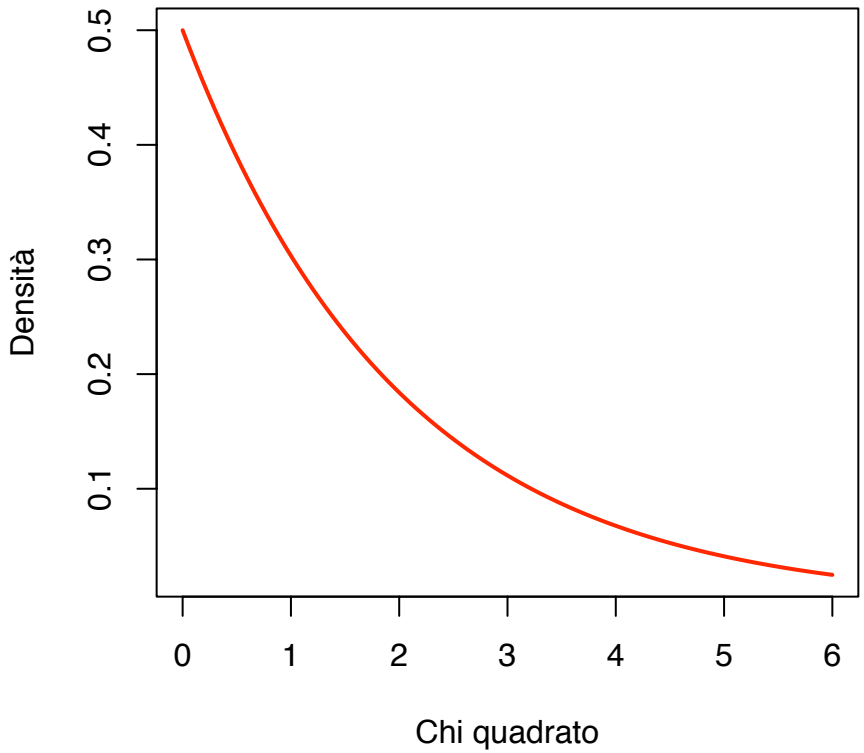

Densità
0.0
0.2
0.4
0.6
110
120
130
140
Media di campioni di grandezza n = 100

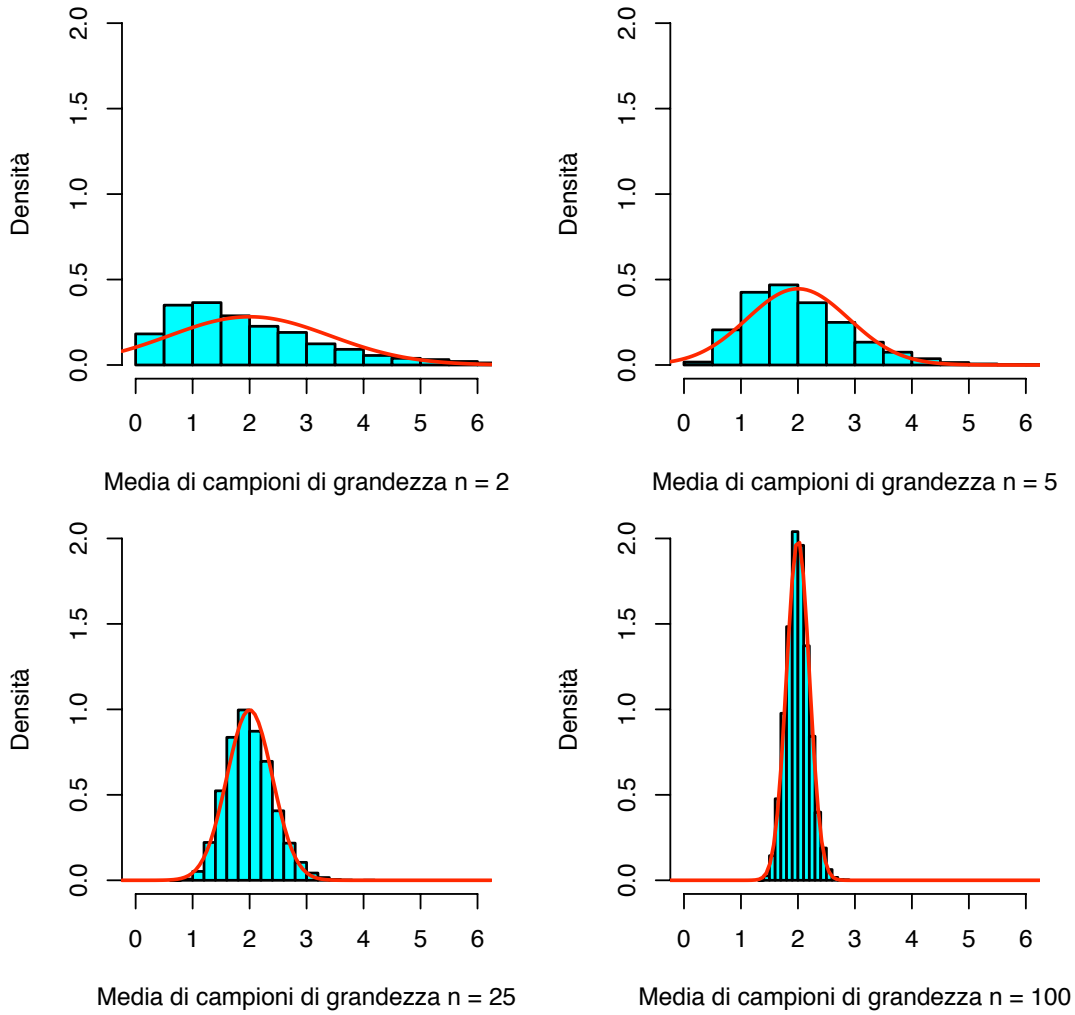
5 Simulazione 4

- Consideriamo ora una popolazione asimmetrica, $\chi^2_{\nu=2}$.
- La distribuzione χ^2 con parametro $\nu = 2$ ha una media $\mu = \nu$ e una varianza uguale a $\sigma^2 = 2\nu$.
- A differenza della distribuzione normale, la distribuzione $\chi^2_{\nu=2}$ è dotata di un'asimmetria positiva.

- Usando **R**, verranno estratti da questa popolazione 10000 campioni causali di grandezza $n = 2, 5, 25, 100$ e verrà calcolata la media di ciascuno di questi campioni di grandezza n.

- All'istogramma che rappresenta la distribuzione delle medie dei campioni di grandezza n verrà sovrapposta la distribuzione normale con parametri $\mu = \nu$ e $\sigma^2 = (2\nu)/n$.


Densità
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
Chi quadrato


Densità
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
6
Media di campioni di grandezza n = 2
Media di campioni di grandezza n = 5
Media di campioni di grandezza n = 25
Media di campioni di grandezza n = 100

6 Conclusioni

- Da questi esempi possiamo concludere le seguenti regole generali. Supponiamo che $\bar{x}$ sia la media di un campione casuale estratto da una popolazione avente media μ e varianza σ^2.

  - La media della distribuzione campionaria di $\bar{x}$ è uguale alla media della popolazione: $\mu_{\bar{x}} = \mu$.
  - La varianza della distribuzione campionaria di $\bar{x}$ è uguale a $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$.

Legge dei grandi numeri

- Di conseguenza, al crescere della numerosità del campione, la media del campione $\bar{x}$ diventa via via più simile alla media della popolazione μ.
  - In un campione molto grande, $\bar{x}$ sarà quasi certamente molto simile a μ. Tale fatto è chiamato legge dei grandi numeri.

Teorema del limite centrale

- Indipendentemente dalla forma della distribuzione della popolazione, la distribuzione campionaria di $\bar{x}$ è approssimativamente normale e quest'approssimazione è tanto migliore quanto maggiori sono le dimensioni del campione: $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$. Tale fatto è chiamato teorema del limite centrale.

  - Quanto debba essere grande n affinché questa approssimazione sia accettabile dipende dalla forma della distribuzione della popolazione – in generale, comunque, $n = 30$ è sufficiente.

Distribuzione campionaria nel caso di una popolazione gaussiana

- Se la distribuzione della popolazione è gaussiana allora la distribuzione campionaria di $\bar{x}$ sarà normale, indipendentemente dalla numerosità n del campione.