Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 10 11

Piano della presentazione

Curve di densità

Funzione di densità

- Questa curva continua si chiama funzione di densità e ha le seguenti proprietà:
  - l'area totale sotto la curva di densità è uguale a 1.0;
  - la proporzione di dati della distribuzione con un valore compreso in un certo intervallo è uguale all'area sottesa alla curva di densità in quell'intervallo;
  - la curva di densità non assume mai valori negativi (...sull'ordinata, naturalmente).

Funzione di densità

- Per trovare l'area sottesa alla funzione di densità sono necessari calcoli matematici complessi (integrali).
- Per i nostri scopi, tuttavia, dobbiamo soltanto capire l'idea di base:

$$area \quad \rightarrow \quad proporzione\ di\ casi$$

Funzione di densità

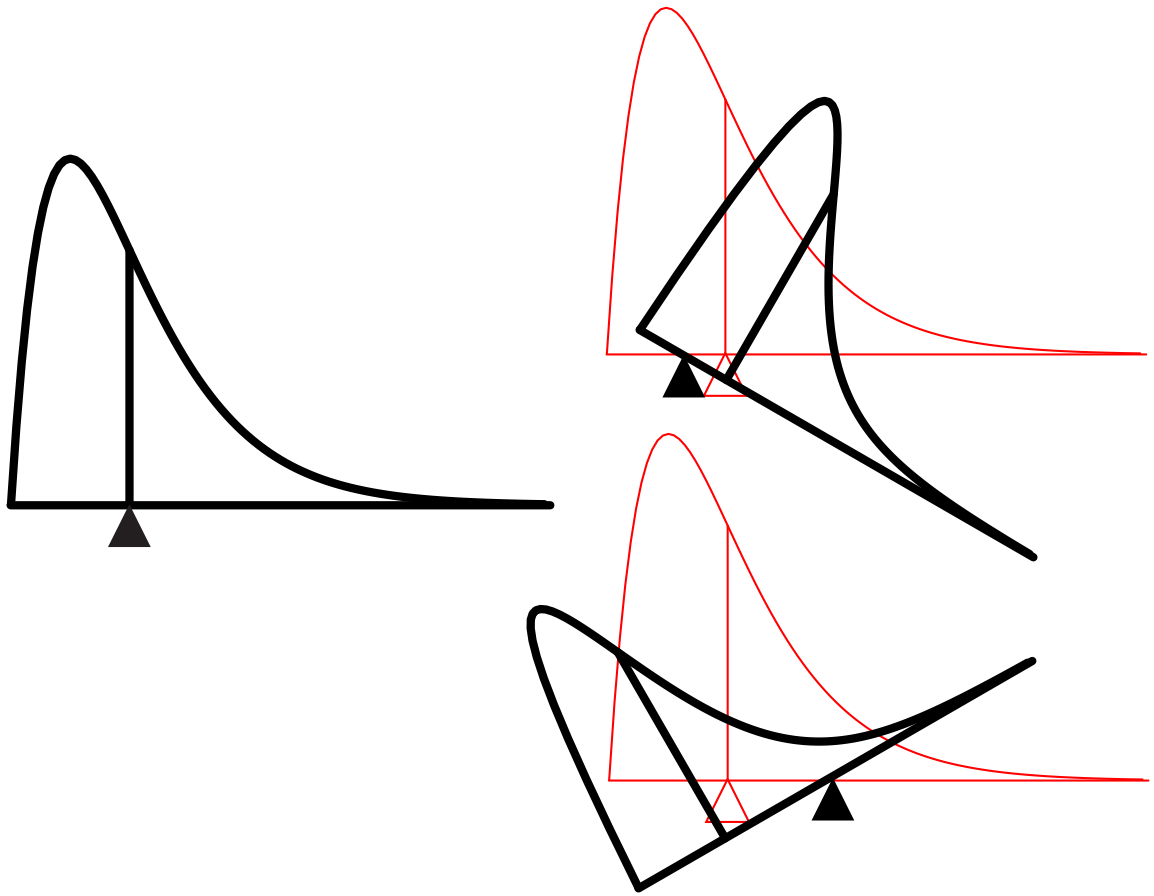
- Abbiamo imparato a calcolare la media, mediana e deviazione standard di un insieme di dati. Queste concetti hanno lo stesso significato nel caso di una funzione di densità.
  - La **media** di una curva di densità è il valore della variabile tale per cui metà dell'area sottesa alla curva si trova al di sotto di quel valore e metà si trova al di sopra di esso.
  - Se la curva di densità venisse ritagliata in un materiale solido (per esempio, un foglio di cartone), allora la media sarebbe il punto d'equilibrio.

Funzione di densità

notazione La media di una curva di densità è denotata da μ.

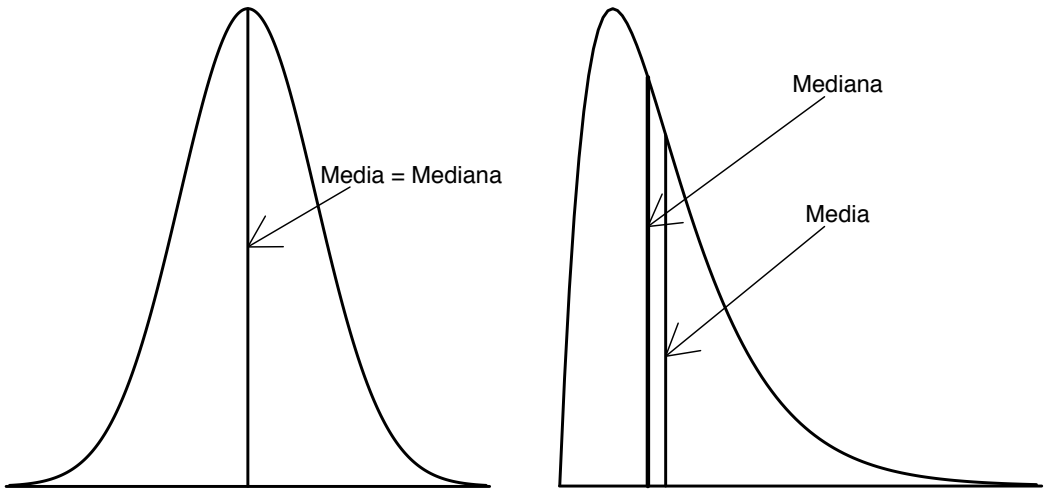
Le lettere greche sono tradizionalmente usate per indicare le caratteristiche delle funzioni di densità, per aiutarci a distinguere la media $\bar{x}$ di un campione di osservazioni dalla media μ di una funzione di densità.
Lo stesso dicasi per la varianza campionaria s^2 e la varianza della popolazione σ^2.

μ = punto di equilibrio



Media e Mediana

- In una distribuzione simmetrica la media e la mediana coincidono.
- In una distribuzione asimmetrica sia la media che la mediana si spostano nella direzione della coda più lunga della distribuzione –la media più della mediana.



Deviazione standard

- È possibile definire la deviazione standard, denotata dalla lettera σ anche per una curva di densità.
- La deviazione standard di una funzione di densità è più difficile da visualizzare della media ma, anche in questo caso, possiamo pensare alla deviazione standard di una serie di osservazioni come ad una sorta di scarto medio delle osservazioni dalla media della distribuzione.

Le distribuzioni normali

Le distribuzioni normali

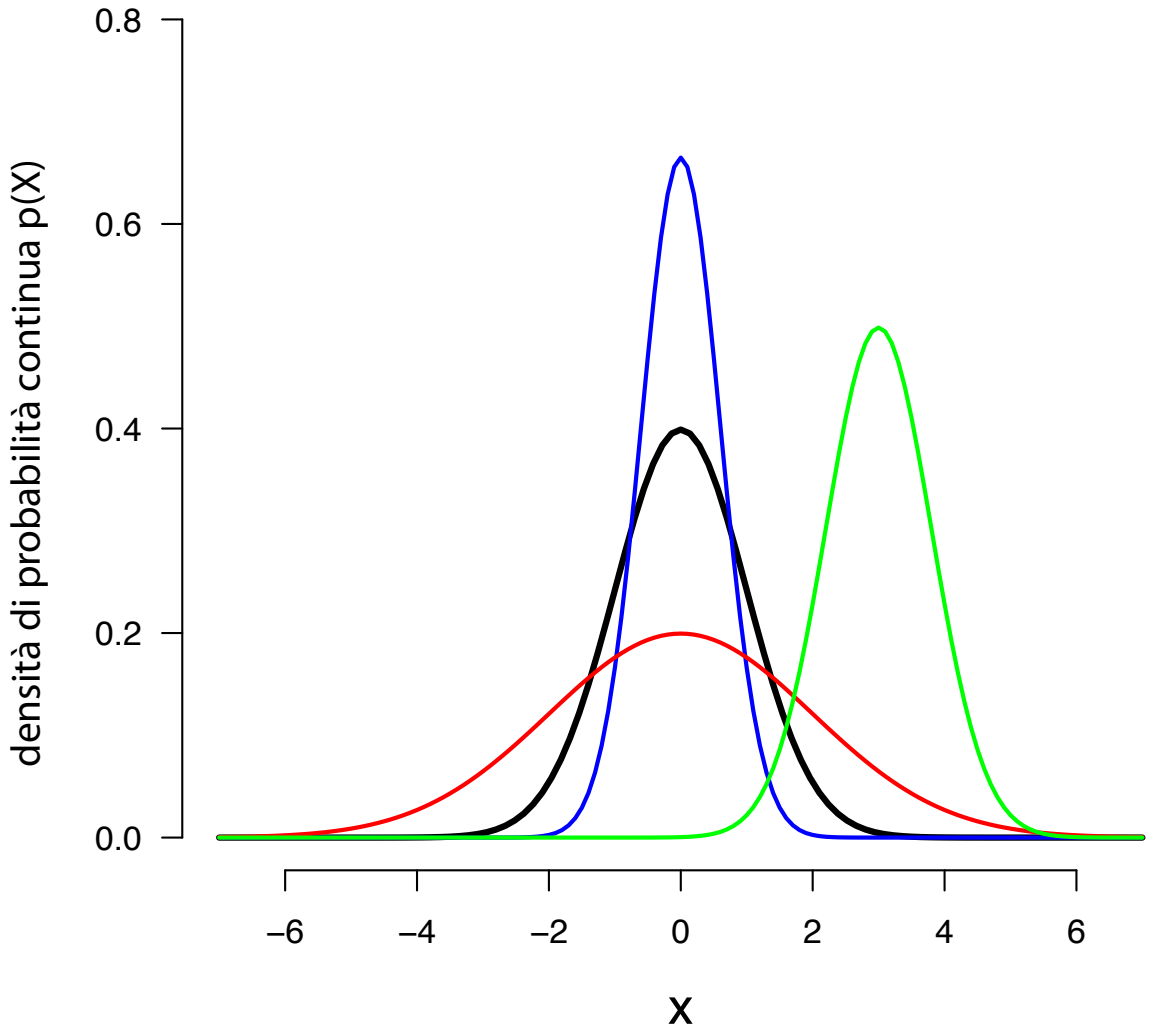
- Le cosiddette distribuzioni normali sono le distribuzioni più importanti in statistica. Le distribuzioni normali sono anche dette gaussiane in onore di Carl Friedrich Gauss, uno dei matematici che le ha scoperte.
- Questo è un nome migliore perché, come le persone "normali", anche le variabili distribuite normalmente sono difficili da trovare.
- Perché allora le distribuzioni normali sono importanti per la statistica? Sono importanti per il ruolo centrale che hanno nell'inferenza statistica.

Micceri, T. (1989) *The unicorn, the normal curve, and other improbable creatures*. Psychological Bulletin, 105(1), 156-166.

Le distribuzioni normali

- Tutte le distribuzioni gaussiane hanno la stessa forma: sono simmetriche, unimodali e a forma di campana.
- Le distribuzioni normali hanno medie e deviazioni standard diverse.
- C'è una diversa distribuzione normale per ciascuna coppia μ e σ.
- Una distribuzione normale con media μ e deviazione standard σ viene denotata da $N(\mu, \sigma)$

Le distribuzioni normali

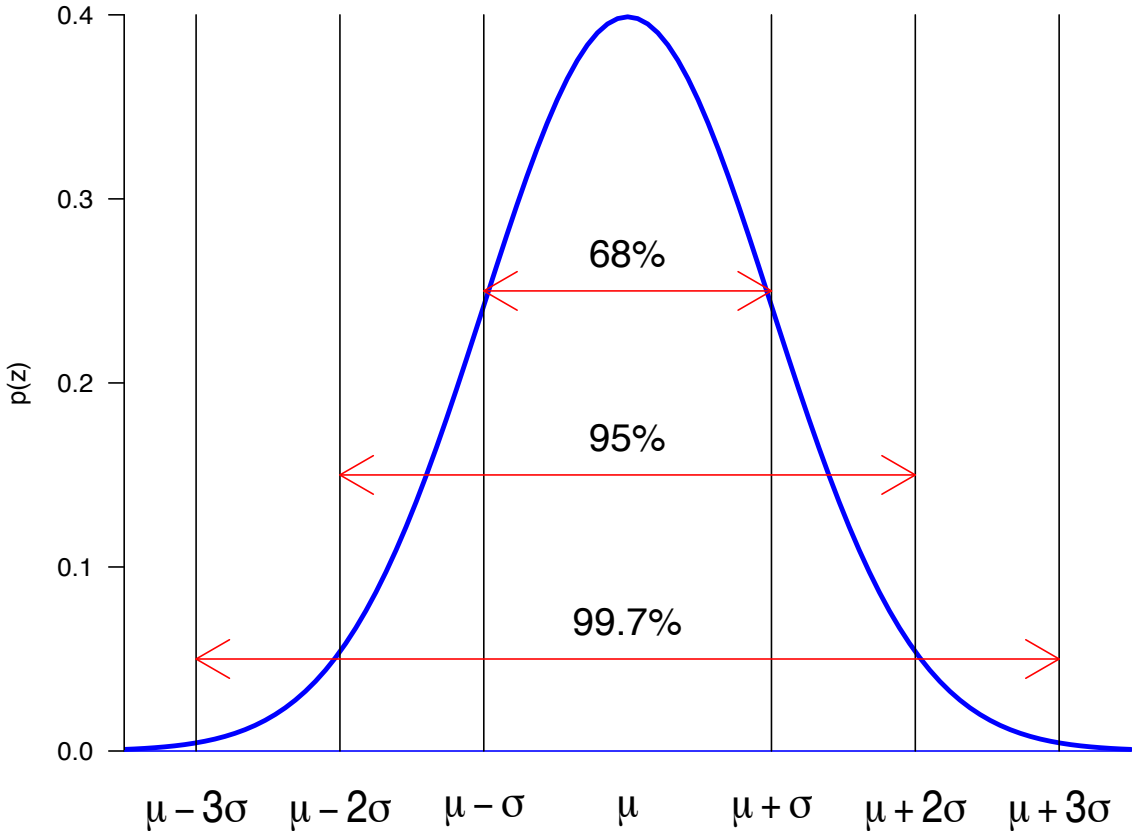


Proprietà delle distribuzioni normali

Tutte le distribuzioni normali hanno le seguenti proprietà:

- Il 68% (circa due terzi) delle osservazioni è compreso nell'intervallo tra ± 1 deviazione standard σ dalla media μ.
- Il 95% delle osservazioni è compreso nell'intervallo tra ± 2 deviazioni standard σ dalla media μ.
- Il 99.7% delle osservazioni è compreso nell'intervallo tra ± 3 deviazioni standard σ dalla media μ

Proprietà delle distribuzioni normali



Proprietà delle distribuzioni normali

Illustrazione

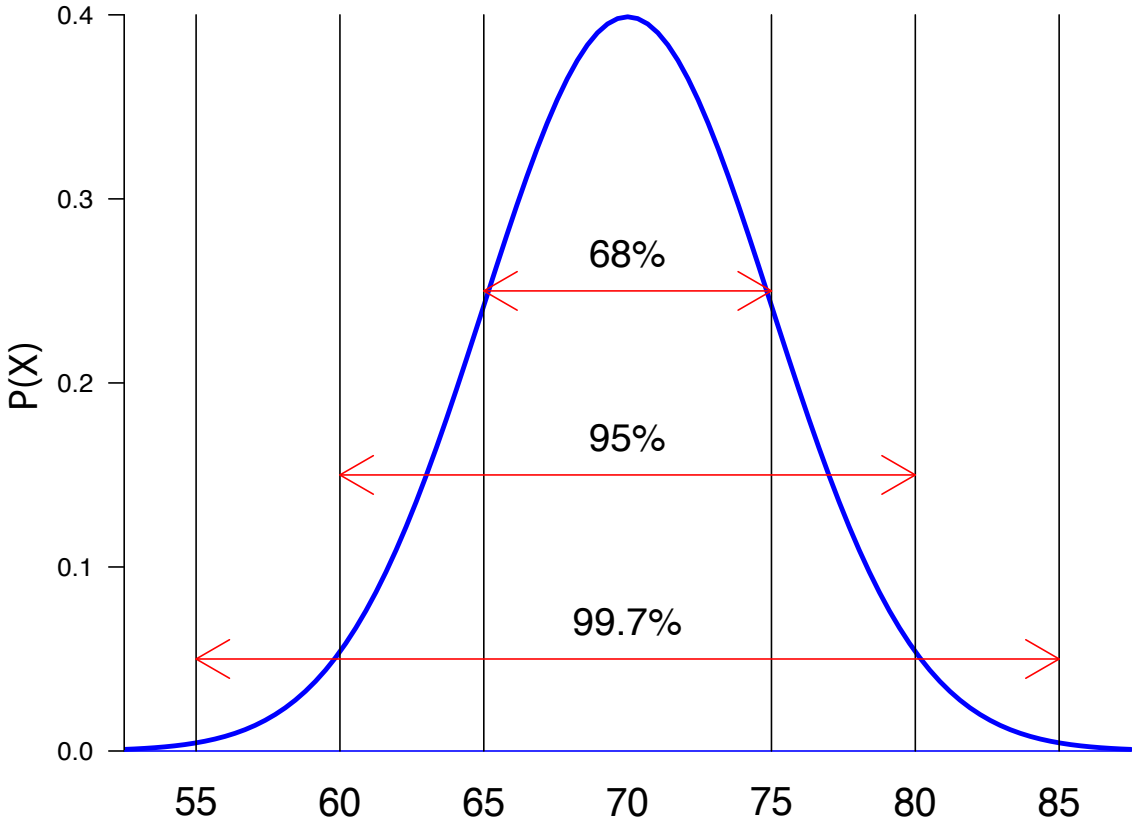
Si considerino i punteggi d'esame di un grande numero di studenti. Si supponga che il punteggio medio sia 70 (in centesimi), con una deviazione standard di 5. Si supponga, inoltre, che i punteggi siano distribuiti in maniera approssimativamente normale.

Che percentuale di studenti riceve un punteggio compreso tra 60 e 80?

Illustrazione

- Dato che 60 è 2 deviazioni standard sotto la media e 80 è 2 deviazioni standard sopra la media, circa il 95% degli studenti ha un punteggio compreso in questo intervallo.
- Circa il 2.5% degli studenti ha punteggi inferiori a 60 e circa il 2.5% ha punteggi superiori a 80.

$N(70, 5)$



Definizione

- In termini formali, diciamo che una variabile aleatoria continua ha una distribuzione normale con parametri μ e σ, dove $-\infty < \mu < \infty$ e $0 < \sigma$, se la funzione di densità di X è

$$f(x; \mu; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right\}}; \quad (\vee x, -\infty < x < \infty)$$

- Può essere dimostrato che $E(X) = \mu$ e $V(X) = \sigma^2$.

La distribuzione normale standardizzata

La distribuzione normale standardizzata

- Tutte le distribuzioni normali si riducono alla stessa distribuzione se vengono misurate in unità σ attorno alla media μ. Tale mutamento dell'unità di misura della distribuzione viene detto standardizzazione.
- Anche se la nozione di standardizzazione è stata introdotta nel contesto delle distribuzioni normali, una variabile può essere standardizzata che se non è distribuita normalmente.

La distribuzione normale standardizzata

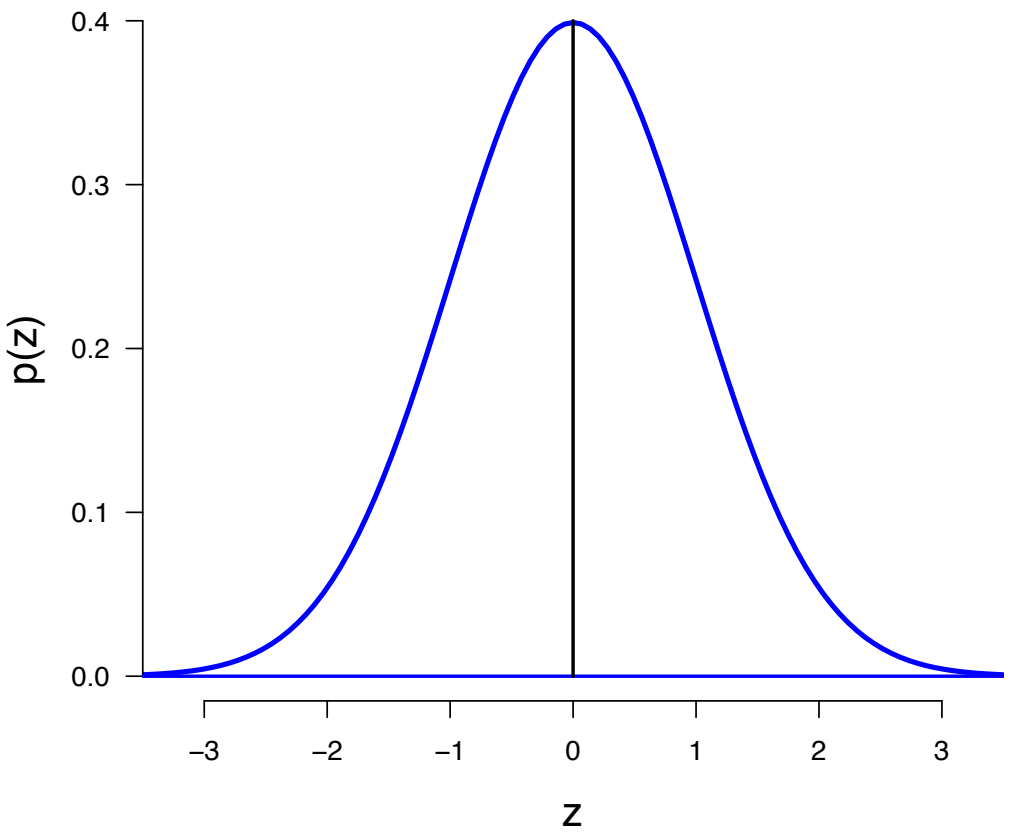
- Se x è un'osservazione facente parte di una distribuzione avente media μ e deviazione standard σ, allora il valore standardizzato di x si calcola come:

$$\frac{x - \mu}{\sigma}$$

- Le variabili standardizzate, dette punti z, hanno media 0 e varianza 1.

$N(0, 1)$

La distribuzione normale standardizzata si indica con $N(0, 1)$. Una variabile aleatoria normale standardizzata viene denotata da Z.



Motivazione

- Perchè è utile standardizzare una variabile distribuita normalmente?
  - La standardizzazione consente di trovare le aree sottese alla distribuzione normale.
  - Queste aree non possono essere calcolate con una semplice formula.
  - Invece, dobbiamo utilizzare delle tabelle o un software statistico.

Motivazione

- La standardizzazione consente di risolvere questo problema utilizzando un'unica tabella – quella della distribuzione normale standardizzata.
- Sarebbe impossibile, infatti, avere tabelle per tutte le infinite possibili combinazioni di μ e σ.

Proprietà di $N(0, 1)$

- Nel contesto della distribuzione normale, la standardizzazione è utile in quanto trasforma qualsiasi distribuzione normale – con media μ e deviazione standard σ – in una distribuzione normale con media 0 e deviazione standard 1.
- Se la distribuzione di x è $N(\mu, \sigma)$, allora la distribuzione di $z = (x - \mu)/\sigma$ è $N(0, 1)$.

Distribuzione cumulativa

- La distribuzione cumulativa (cdf) di Z è denotata da

$$\Phi(z) = P(Z < z)$$

- La tavola A del manuale di Agresti e Finlay può essere usata per trovare $\Phi(z)$ e i percentili di $N(0, 1)$.

Distribuzione cumulativa

Notazione z_α denota il valore z che ha alla sua destra una proporzione α dell'area sottesa alla curva di densità.

- In altri termini, z_α è il $[100(1-\alpha]$esimo percentile di $N(0,1)$.
- Per esempio $z_{0.05} = 1.645$, rappresenta il $[100(1-0.05] = 95^o$percentile della normale standard

Tipi di problemi

- Ci sono due tipi di problemi relativi alle distribuzioni normali.

1. Trovare l'area sottesa alla funzione di densità normale tra due valori x, o per valori maggiori o minori di x.
2. Trovare il valore x che corrisponde ad una data area.

Trovare l'area sottesa alla funzione di densità normale

Come trovare l'area sottesa alla normale?

1. Si disegni una curva gaussiana con la media ? al centro e i valori x che definiscono l'area da trovare.
2. Si standardizzino i valori x usando la formula

$$z_i = \frac{x_i - \mu}{\sigma}$$

   Si noti che la media μ corrisponde ad un valore $z = 0$
3. Si trovino i valori z_i nella tabella della distribuzione normale. La tabella A del libro fornisce l'area sottesa alla curva alla destra di Z.
4. Sapendo che l'area totale sottesa alla curva è 1, si usi l'informazione fornita dalla tabella per trovare l'area richiesta.

Trovare l'area sottesa alla funzione di densità normale

Illustrazione

si considerino nuovamente i punteggi d'esame di un gruppo di studenti e supponiamo che il punteggio medio sia $\mu = 70$ con una deviazione standard $\sigma = 5$.

Qual è il punteggio standardizzato corrispondente ai punteggi "grezzi" $x_1 = 65$ e $x_2 = 80$?

Illustrazione

- Si standardizzino i valori x

$$z_1 = \frac{65 - 70}{5} = -1$$

$$z_i = \frac{80 - 70}{5} = 2$$

- $x_1 = 65$ è 1 deviazione standard sotto la media
- $x_2 = 80$ è 2 deviazioni standard sopra la media

Trovare l'area sottesa alla funzione di densità normale

Illustrazione

si considerino ancora i punteggi d'esame di un grande numero di studenti e supponiamo che i punteggi siano distribuiti normalmente. La media della distribuzione è $\mu = 70$ e la deviazione standard è $\sigma = 5$.

Che percentuale di studenti riceve un punteggio minore di 60?

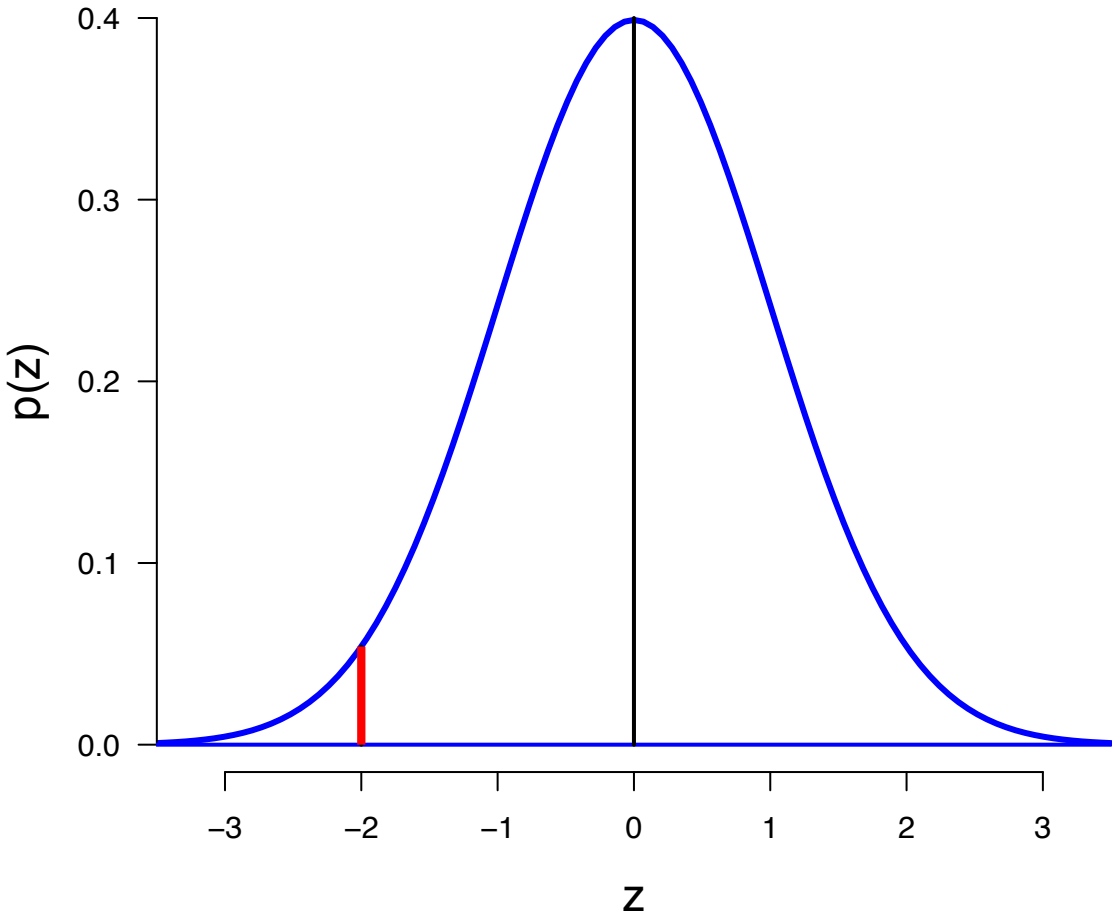
Illustrazione

- Si standardizzi il valore di $x = 60$

$$z_x = \frac{60 - 70}{5} = -2$$

- Dato che la distribuzione è simmetrica, l'area nell'intervallo $[2; \infty]$ è uguale all'area nell'intervallo $[-\infty; -2]$.
- Dalla tabella troviamo $\Phi(z = 2) = 0.0228$.
- Dunque, circa il 2.3% degli studenti riceve un punteggio minore di 60.

Illustrazione



Trovare l'area sottesa alla funzione di densità normale

Illustrazione

un milionario eccentrico offre un premio agli studenti che ottengono un punteggio maggiore di 83.75 nel test precedente.

Che percentuale di studenti otterrà il premio?

Illustrazione

- Si standardizzi il valore $x = 83.75$

$$z_1 = \frac{83.75 - 70}{5} = 2.75$$

- Dalla tabella troviamo $\Phi(z = 2.75) = 0.0030$.
- Dunque, soltanto 3 studenti su 1000 vinceranno il premio.

Trovare l'area sottesa alla funzione di densità normale

Illustrazione

che percentuale di studenti ottiene punteggi compresi tra 70 e 80?

Illustrazione

- Si standardizzino i valori x

$$z_\mu = \frac{70 - 70}{5} = 0$$

$$z_2 = \frac{80 - 70}{5} = 2$$

- Per la simmetria della distribuzione normale, l'area compresa nell'intervallo $[-\infty; 0]$ è uguale a 0.5.
- Dalla tabella troviamo l'area nell'intervallo $[2; \infty]$, $\Phi(z = 2) = 0.0228$.
- Quindi, l'area compresa tra 0 e 2 è $0.5 - 0.0228 = 0.4772$. In altri termini, circa il 48% degli studenti ottiene punteggi compresi tra 70 e 80.

Dalla probabilità al punteggio grezzo

Dalla probabilità al punteggio grezzo

Trovare il punteggio x conoscendo l'area sottesa alla curva sostanzialmente inverte il processo descritto sopra.

1. Si disegni una curva gaussiana con la media μ al centro e i valori che definiscono l'area considerata.
2. Si usi la tabella della distribuzione normale standardizzata per trovare i valori z_i corrispondenti all'area in questione.
3. Si trasformino i punteggi standardizzati nei punteggi grezzi usando la formula

$$x_i = \mu + z_i\sigma$$

Dalla probabilità al punteggio grezzo

Illustrazione

essendosi convinto che la sua offerta non è particolarmente generosa, il milionario eccentrico dell'esempio precendente decide di offrire un premio al 5% degli studenti coni voti più alti.

Che voto deve ottenere uno studente per ricevere il premio?

Illustrazione

- Il problema richiede il valore z_i tale per cui l'intervallo $[z; \infty]$ sottenda il 5% dell'area. Dalla tabella troviamo che $z_{0.05}$, ovvero il $95esimo$ percentile di $N(0, 1)$, è $z_{0.05} = 1.645$.
- Quindi, circa il 95% dell'area è sottesa alla curva di densità normale standardizzata nell'intervallo $[-\infty; 1.645]$.
- Infine, per trasformare il punteggio z_i nel punteggio grezzo, avremo

$$x_{0.05} = \mu + z_{0.05}\sigma = 70 + 1.645 \times 5 = 78.23$$

R e la distribuzione normale

R e la distribuzione normale

R offre le seguenti funzioni per le distribuzioni normali:

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

dnorm()

La funzione `dnorm(x, mean, sd)` calcola il valore della densità in x.

```
 dnorm(0,0,1)
# [1] 0.3989423
 dnorm(x=70,mean=70,sd=5)
# [1] 0.07978846
```

pnorm()

La funzione `pnorm(x, mean, sd)` calcola il valore della ripartizione in x – ovvero, calcola l'area sottesa alla normale nell'intervallo $[-\infty, x]$.

```
 pnorm(0, 0, 1)
## [1] 0.5
 pnorm(-1, 0, 1)
## [1] 0.1586553
 pnorm(1, 0, 1)
## [1] 0.8413447
 pnorm(1, 0, 1)-pnorm(-1, 0, 1)
## [1] 0.6826895
 pnorm(-1.65, 0, 1)
## [1] 0.04947147
 pnorm(-1.96, 0, 1)
## [1] 0.0249979
 pnorm(1.96, 0, 1)
## [1] 0.9750021
```

Illustrazione

. . . Torniamo ai nostri punteggi al test:

si trovi l'area compresa tra 70 e 80 per una funzione normale con media 70 e deviazione standard 5.

```
pnorm(80, 70, 5) - pnorm(70, 70, 5)
## [1] 0.4772499
```

```
## ovvero, esprimendo i due valori in punti Z
pnorm(2,0,1)-pnorm(0,0,1)
## [1] 0.4772499
```

rnorm()

La funzione `rnorm(n, mean, sd)` genera un campione (casuale) da una normale standard, di dimensione n.

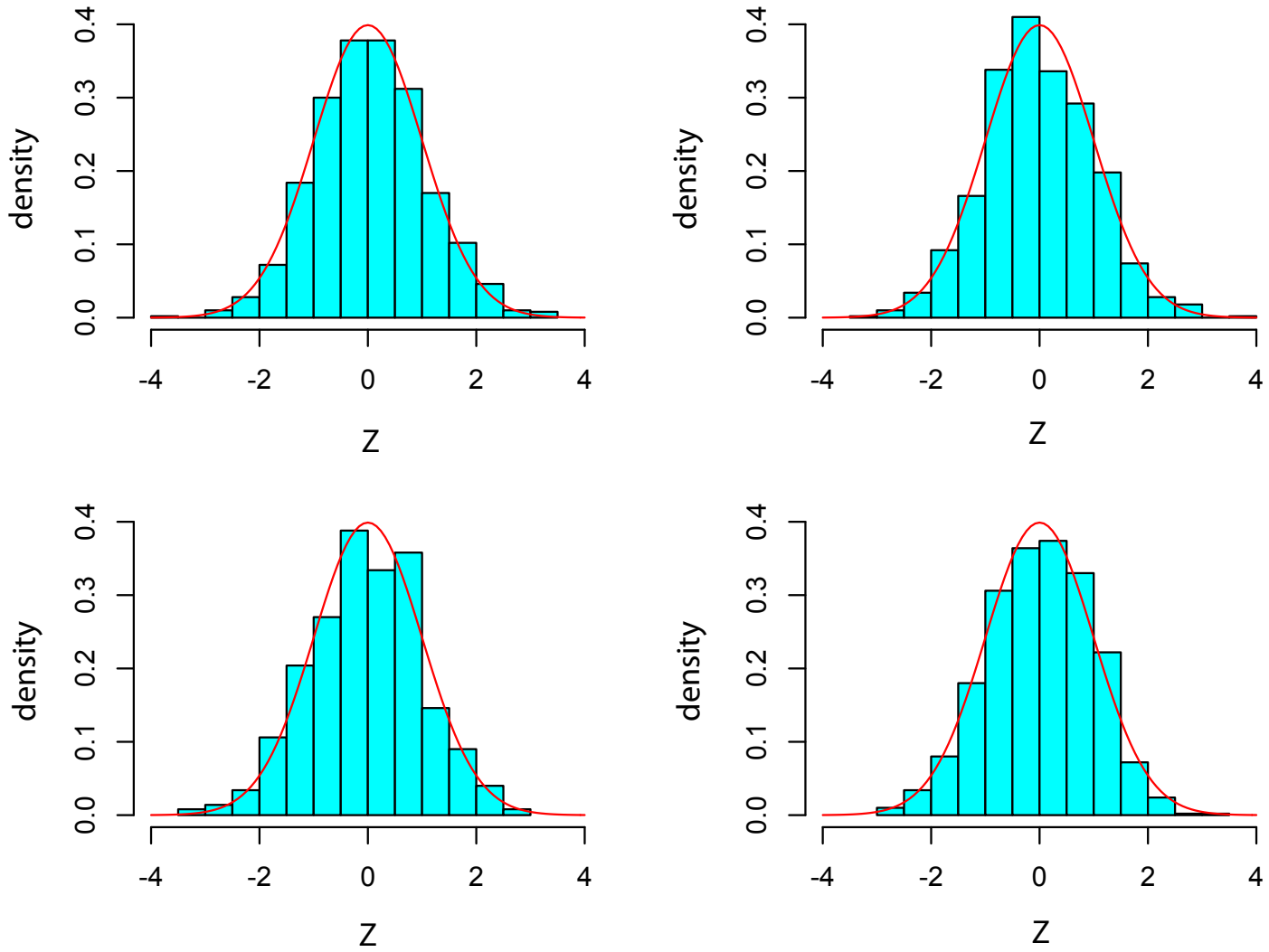
Le figure seguenti, per esempio, sono state create con il seguente comando:

```
x<-rnorm(1000,0,1)
 truehist(x, ylim=c(0, .41),xlim=c(-4,4)
                      xlab=c("z"), ylab=c("density"))

curve(dnorm(x, 0, 1), add=TRUE, col="red")
```

Si noti l'uso di dnorm(x,0,1) per disegnare la curva (`curve()`) continua di densità di probabilità.

4 campioni casuali da $N(0, 1)$



Stima e test di ipotesi

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.

Legame concettuale tra statistiche campionarie e caratteristiche della popolazione

In ciascuna applicazione pratica dell'inferenza statistica, un solo campione casuale di grandezza n viene estratto, ma la possibilità che il campionamento venga ripetuto ci fornisce, in principio, la fondazione concettuale per decidere quanto un particolare campione sia informativo a proposito della popolazione.

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.

Parametri e statistiche

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.

Stima e test d'ipotesi

- Solitamente, non siamo interessati alle statistiche in sé, 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 principali approcci all'inferenza statistica classica
  1. la stima di parametri (*puntuale* e *intervallare*)
  2. il test d'ipotesi statistiche.

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 $€42,586$.

Variabilità campionaria

- 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 molto raro che la statistica di un qualsiasi campione sia esattamente uguale al parametro della popolazione.

Simulazione 1

Simulazione 1

- La variabilità campionaria verrà illustrata nel modo seguente:

1. verrà considerata una variabile discreta (generica) che può assumere soltanto un piccolo numero di valori possibili (N = 4). Il modello probabilistico associato ai 4 valori sarà uniforme $(p = 1/N)$.
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 $n = 2$.

Simulazione 1

- La media μ e la varianza σ^2 della popolazione verranno anch'esse calcolate.
- μ e σ^2 sono dei parametri, mentre la media aritmetica $\bar{x}$ e la varianza s^2 di ciascun campione sono delle statistiche.

Simulazione 1

- 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: $\Omega = \{2, 3, 5, 9\}$
- L'estrazione con rimessa corrisponde ad una popolazione di grandezza infinita (è sempre possibile infatti estrarre una nuova pallina dall'urna).

Simulazione 1

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

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.

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

Distribuzione della popolazione

- Il valore atteso della popolazione è

$$\mu = \sum_{i=1}^{4} x_i p_i = 4.75$$

- La varianza della popolazione è

$$\sigma^2 = \sum_{i=1}^{4} (x_i - 4.75)^2 p_i = 7.1875$$

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 $s^2 = 0.25$.

Distribuzione campionaria della media

Distribuzione campionaria della media: la distribuzione delle medie $\bar{x}$ di tutti i possibili campioni.

- Se $n = 2$, ci sono $4 \times 4 = 16$[1] possibili campioni. Possiamo dunque elencarli, insieme alle loro medie.

[1] Con il calcolo combinatorio abbiamo $\frac{4!}{2!} = 12$ estrazioni possibili senza rimessa più le 4 coppie dovute alla rimessa $\{22, 33, 55, 99\}$

Distribuzione campionaria della media

campione	$\bar{x}_i$	campione	$\bar{x}_i$
$\{2; 3\}$	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

Ciascuna coppia di osservazioni $\{x_i; x_j\}$ per $i : 1, \ldots, 4$, e $j : 1, \ldots, 4$, ha la stessa probabilità $p = 1/4 \times 1/4 = 1/16$. Per costruire la distribuzione di probabilità della media campionaria sarà sufficiente contare le frequenze (relative) di ciascun valore $\bar{x}_i$.

Distribuzione campionaria della media

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

$\bar{x}$	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	1.00

Distribuzione campionaria della media

- 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

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

Distribuzione campionaria della media

- 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.

Distribuzione campionaria della media

- Questo esercizio ci permette 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 alla varianza della popolazione σ^2 divisa per la grandezza del campione n:

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

Distribuzione campionaria della media

Si noti che:

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

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

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

Distribuzione campionaria della media

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:**
   $\Omega_i = \{2, 3\},\ \bar{x} = 2.5,\ s^2 = 0.25$

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

Tre distribuzioni

Distribuzione della popolazione La distribuzione che contiene tutte le possibili modalità della variabile aleatoria. 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 $x_1, x_2, \ldots, x_n$, 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 $\sigma^2_{\bar{x}}$.

Definizione

• 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 grandezza 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.

Simulazione 2

Simulazione 2

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

1. La stessa variabile aleatoria della simulazione precedente verrà utilizzata come modello di popolazione
2. utilizzando R, verranno estratti con rimessa da questa popolazione 50000 campioni casuali 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 $n = 2$.

Simulazione 2

```
N <- 4
n <- 2
nCampioni <- 50000
X <- c(2, 3, 5, 9)
Media <- mean(X)
Var <- var(X)*(N-1)/N
DistrCamp <- rep(0, nCampioni)

for (i in 1:nCampioni){

samp <- sample(X, n, replace=TRUE)
DistrCamp[i] <- mean(samp)

}

MediaDistrCamp <- mean(DistrCamp)
VarDistrCamp <- var(DistrCamp)*(nCampioni-1)/nCampioni
```

Simulazione 2

- Risultati della simulazione

```
 Media
## [1] 4.75
 Var
## [1] 7.1875
 MediaDistrCamp
## [1] 4.75844
 VarDistrCamp
## [1] 3.622919
 Var/n
## [1] 3.59375
```

Risultati della simulazione

- Popolazione:
  $\mu = 4.75$, $\sigma^2 = 7.1875$
- Distribuzione campionaria della media;
  $\mu_{\bar{x}} = 4.75$, $\sigma^2_{\bar{x}} = 7.1875/2 = 3.59375$
- Risultati della simulazione:
  $\hat{\mu}_{\bar{x}} = 4.75844$, $\hat{\sigma^2}_{\bar{x}} = 3.622919$

Simulazione 3

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: $N(\mu = 125; \sigma = \sqrt{33})$.
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 $n = 10$.

Simulazione 3

```
n <- 10
nCampioni<- 50000
Media <- 125; Var<-33
DS <- sqrt(Var)
DistrCamp <- rep(0,nCampioni)

for (i in 1:nCampioni){
samp <- rnorm(n, Media, DS)
DistrCamp[i] <- mean(samp)
}

MediaDistrCamp <- mean(DistrCamp)
VarDistrCamp<- var(DistrCamp)*(nCampioni-1)/nCampioni
```

Simulazione 3

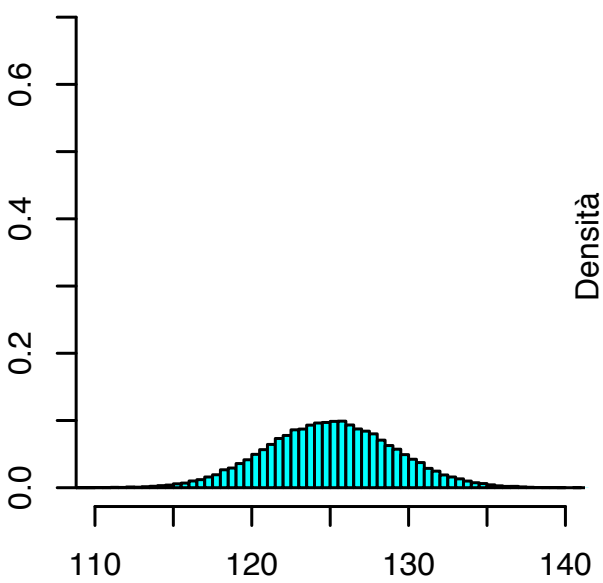
- Risultati della simulazione

```
 Media
## [1] 125
 Var
## [1] 33
 MediaDistrCamp
## [1] 125.001
 VarDistrCamp
## [1] 3.326924
 Var/n
## [1] 3.3
```

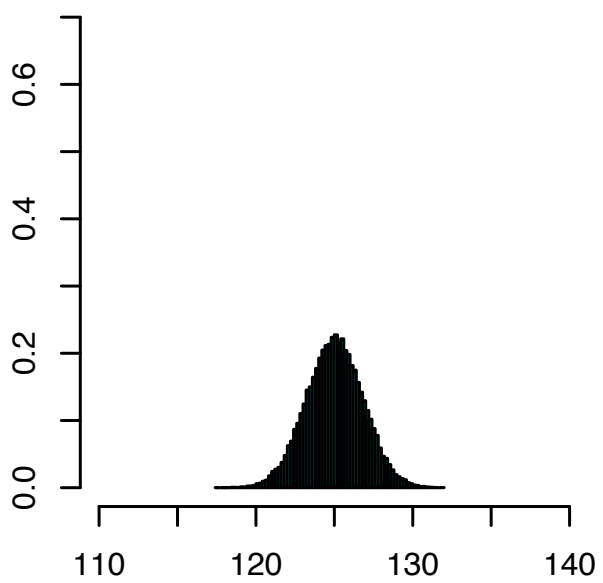
Simulazione 3

- Popolazione:
  $\mu = 125$, $\sigma^2 = 33$
- Distribuzione campionaria della media;
  $\mu_{\bar{x}} = 125$, $\sigma^2_{\bar{x}} = 33/10 = 3.3$
- Risultati della simulazione:
  $\hat{\mu}_{\bar{x}} = 125.001$, $\hat{\sigma^2}_{\bar{x}} = 3.326924$

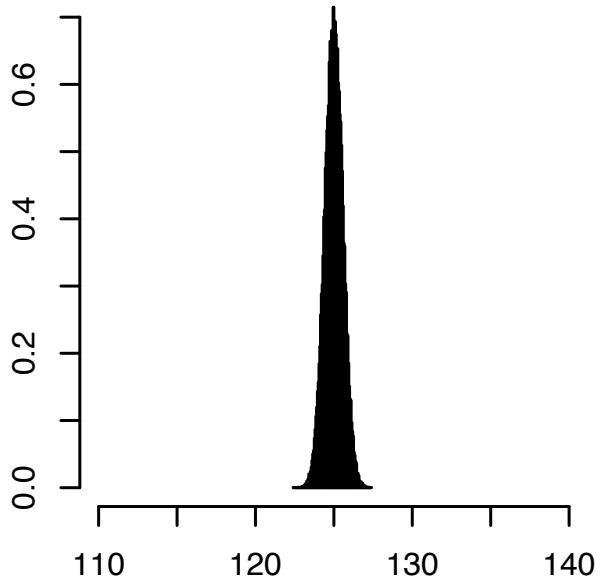
Distribuzione Campionaria al variare di n



Media di campioni di grandezza n = 2



Media di campioni di grandezza n = 10



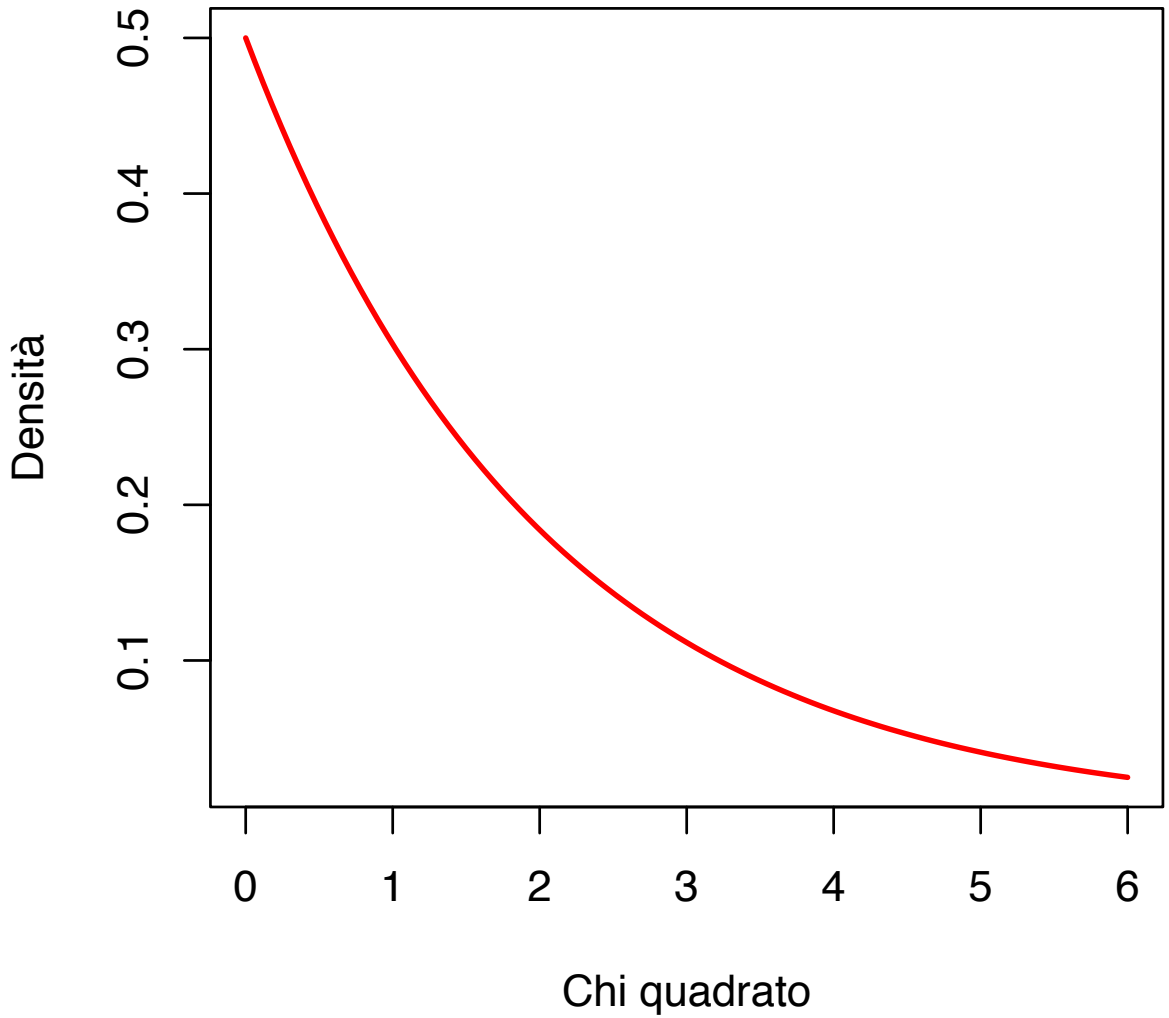
Media di campioni di grandezza n = 100

Simulazione 4

Simulazione 4

- Consideriamo ora una popolazione asimmetrica, $\chi^2_{v=2}$.
- 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 $\chi^2_{v=2}$ è dotata di un'asimmetria positiva.

Distribuzione χ^2 con parametro $\nu = 2$

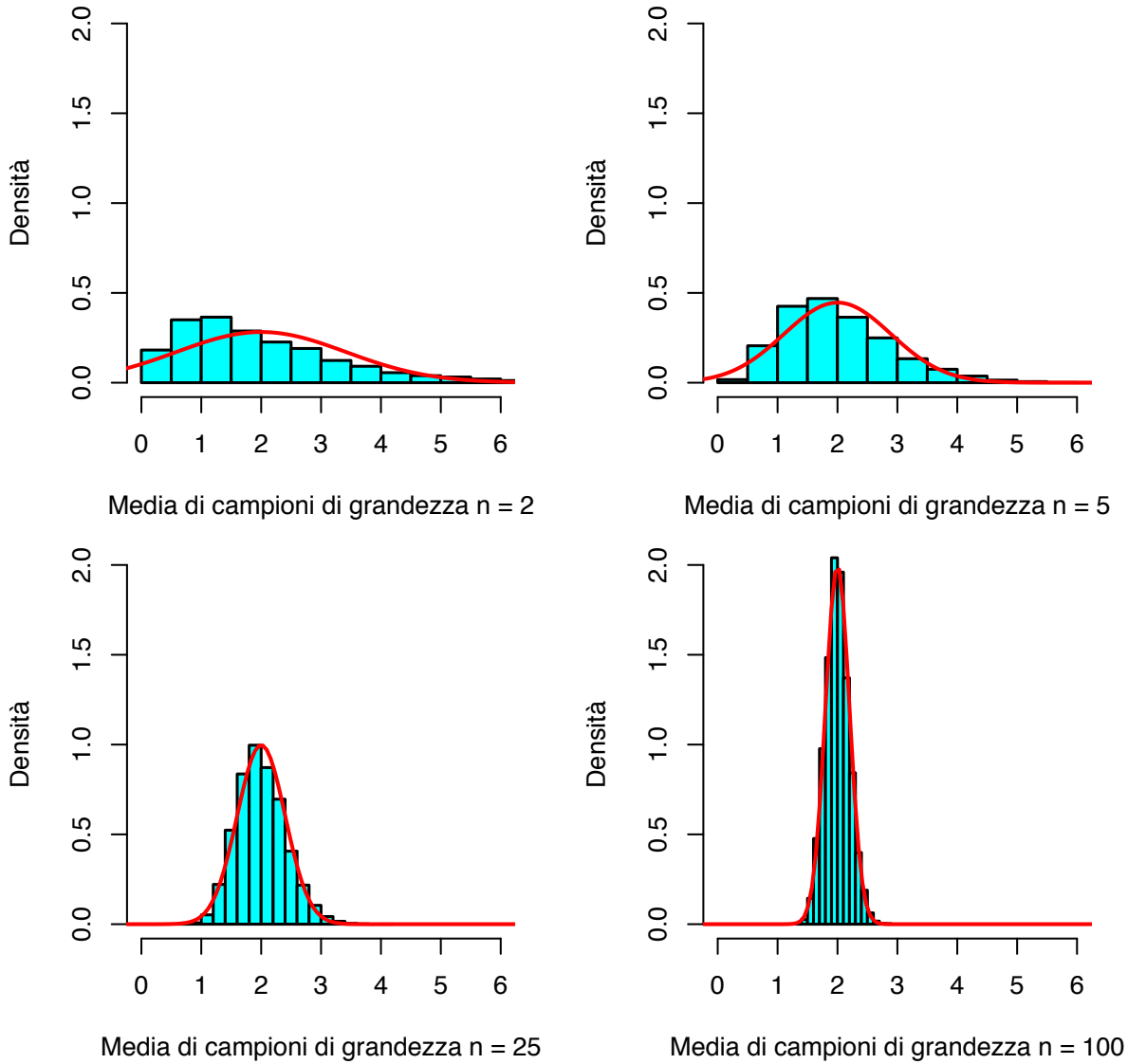


Simulazione 4

- 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

$$\mu = \nu \quad \sigma = \sqrt{2\nu/n}$$

Distribuzione χ^2 con parametro $\nu = 2$



Conclusioni

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 $\sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}$.

Conclusioni

• Di conseguenza, al crescere della grandezza del campione, la media del campione $\bar{x}$ diventa sempre 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.

Conclusioni

• 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 (n) del campione: $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. 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, $n = 100$ è sufficiente.

Conclusioni

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