STATISTICA E ANALISI DEI DATI

Capitolo 12 – Intervalli di confidenza: grandi campioni

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026

Intervalli di confidenza: grandi campioni

- I metodi per la ricerca degli intervalli di confidenza per una popolazione normale non dipendono dalla dimensione del campione osservato
- Se invece la dimensione del campione è elevata $(n \geq 30)$ è possibile utilizzare il **teorema centrale di convergenza** per determinare un intervallo di confidenza di grado $1 - \alpha$ per il parametro non noto ϑ di una popolazione
- Teorema centrale di Convergenza:
  - Considerata $X = N(\mu, \sigma^2)$ con $E(X) = \mu$ e $Var(X) = \sigma^2$
  - Sia $X_1, X_2, \dots, X_n$ il campione casuale analizzato di lunghezza n
  - Il teorema centrale di convergenza afferma che la variabile aleatoria

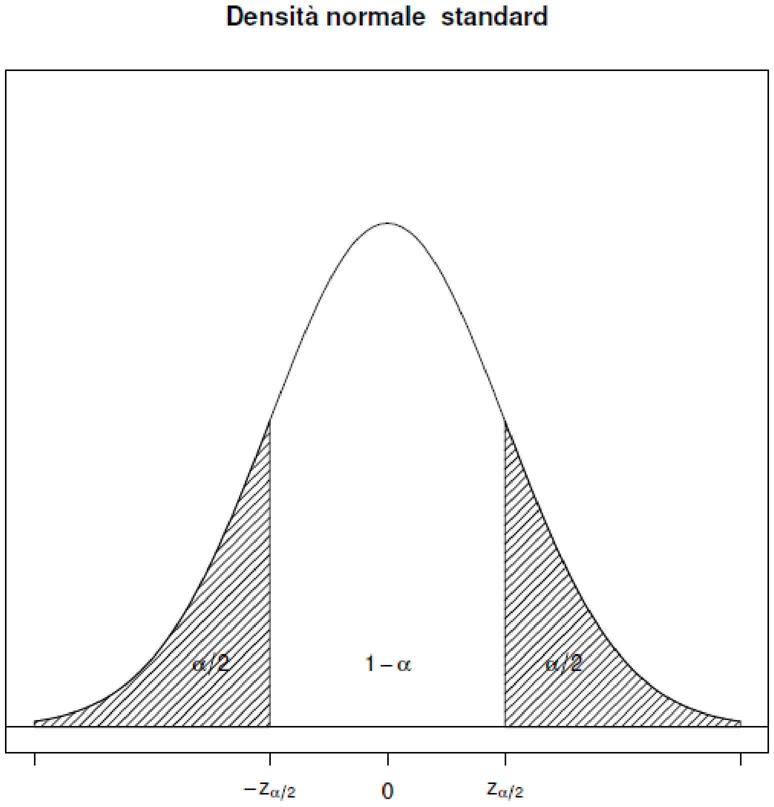
$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard

Intervalli di confidenza: grandi campioni

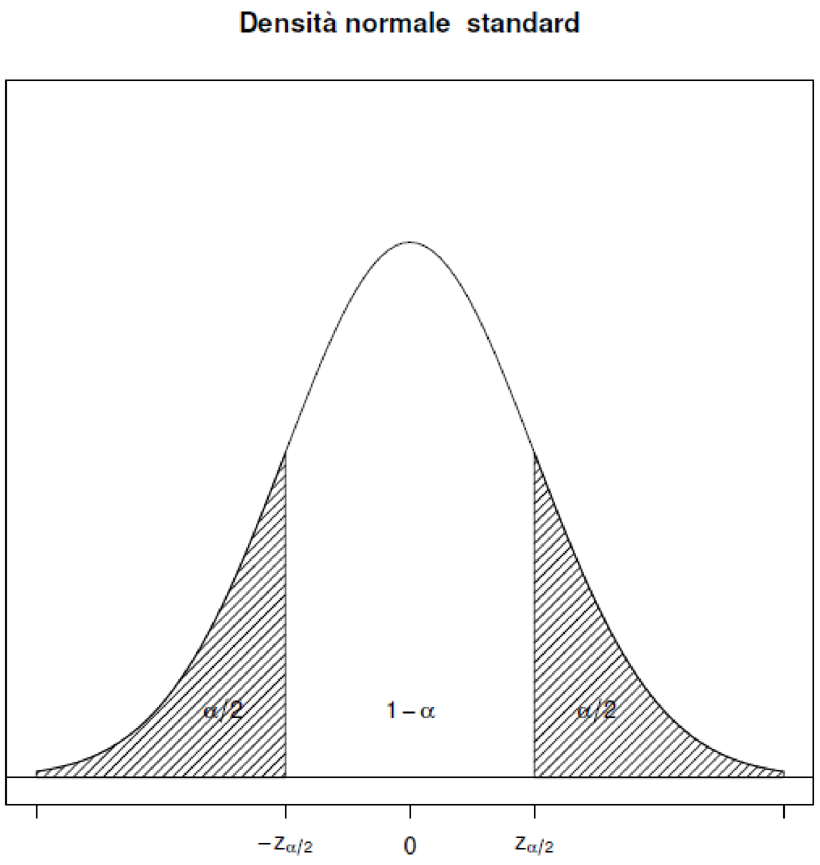
- Se $E(X) = \mu$ e $Var(X) = \sigma^2$ dipendono da un parametro non noto ϑ, Z_n è una variabile pivot, poiché:
  - Dipende dal campione $X_1, X_2, \dots, X_n$
  - Dipende dal ϑ attraverso $E(X) = \mu$ e $Var(X) = \sigma^2$
    - per grandi campioni la sua funzione di distribuzione è approssimativamente normale standard e quindi non contiene il parametro ϑ da stimare
- Per campioni di grandi dimensioni, si può applicare il metodo pivotale in forma approssimata:

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$



Intervalli di confidenza: grandi campioni

- Quando la dimensione del campione è elevata, utilizzeremo il metodo pivotale in forma approssimata nei seguenti casi:
  - intervallo di confidenza per:
    - il parametro p di una popolazione di Bernoulli;
    - il parametro p di una popolazione binomiale;
    - il parametro p di una popolazione geometrica modificata;
    - il parametro λ di una popolazione di Poisson;
    - il parametro ϑ di una popolazione uniforme;
    - il parametro λ di una popolazione esponenziale



STATISTICA E ANALISI DEI DATI

Parametro p di una popolazione di Bernoulli

I.C. per il parametro p di una popolazione di Bernoulli

- Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria:

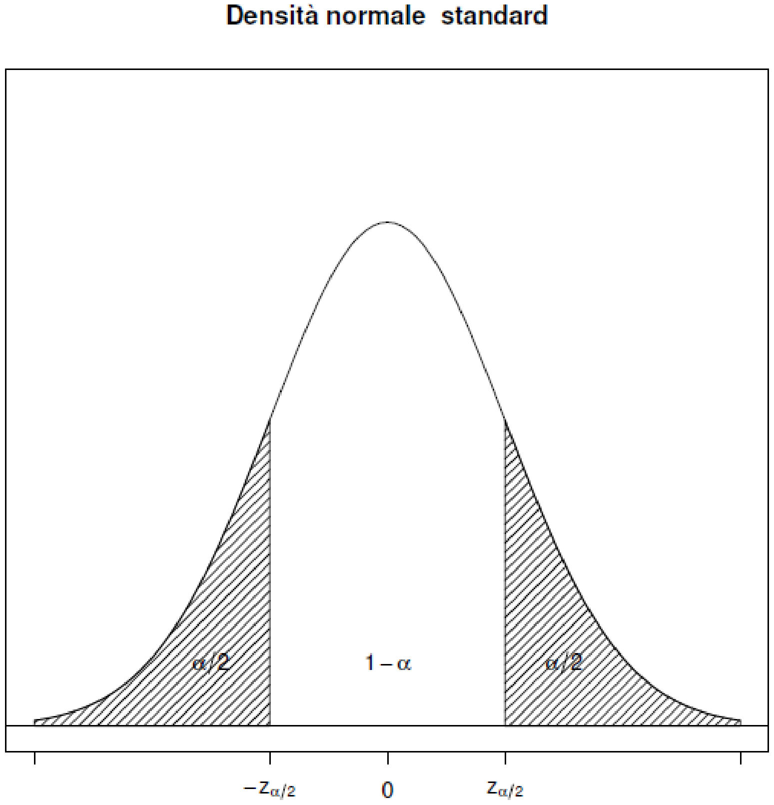
$$p^x(1-p)^{1-x} \quad x = 0,1 \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = p \qquad\qquad \sigma^2 = Var(X) = p(1-p)$$

- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



I.C. per il parametro p di una popolazione di Bernoulli

- Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria:

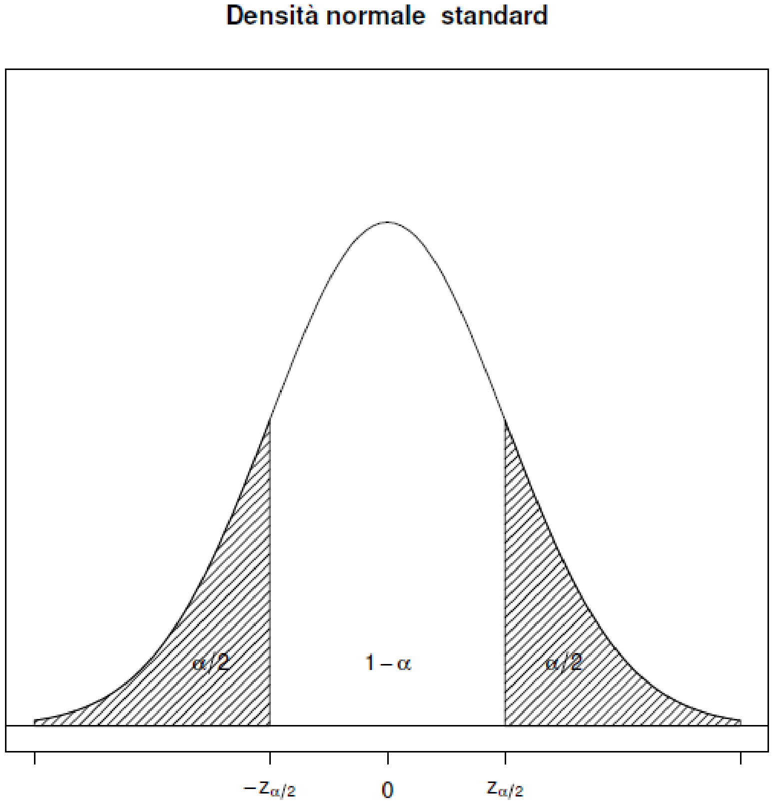
$$p^x(1-p)^{1-x} \quad x = 0{,}1 \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = p \qquad\qquad \sigma^2 = Var(X) = p(1-p)$$

- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}$$



I.C. per il parametro p di una popolazione di Bernoulli

- Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria:

$$p^x(1-p)^{1-x} \quad x = 0,1 \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = p \qquad\qquad \sigma^2 = Var(X) = p(1-p)$$

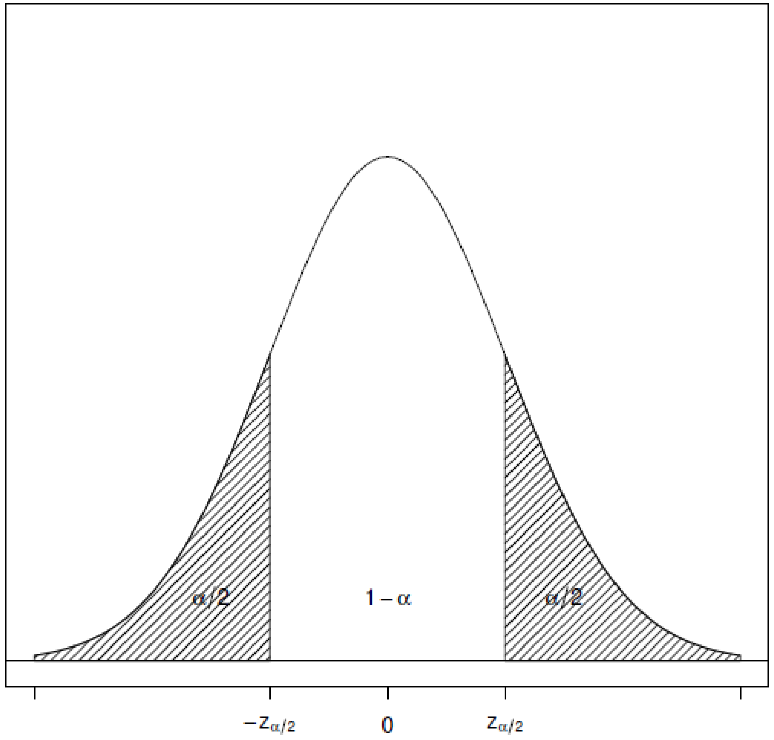
- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_n - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$$

- Quindi per campioni sufficientemente numerosi si ha che:

$$P\left(-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} < z_{\frac{\alpha}{2}}\right) \cong 1 - \alpha$$

Densità normale standard



I.C. per il parametro p di una popolazione di Bernoulli

- Da cui si può risolvere la disuguaglianza:

$$-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} < z_{\frac{\alpha}{2}}$$

- Che rappresenta il **valore assoluto** ed è equivalente a:

$$\left[\sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}}\right]^2 < z^2_{\frac{\alpha}{2}}$$

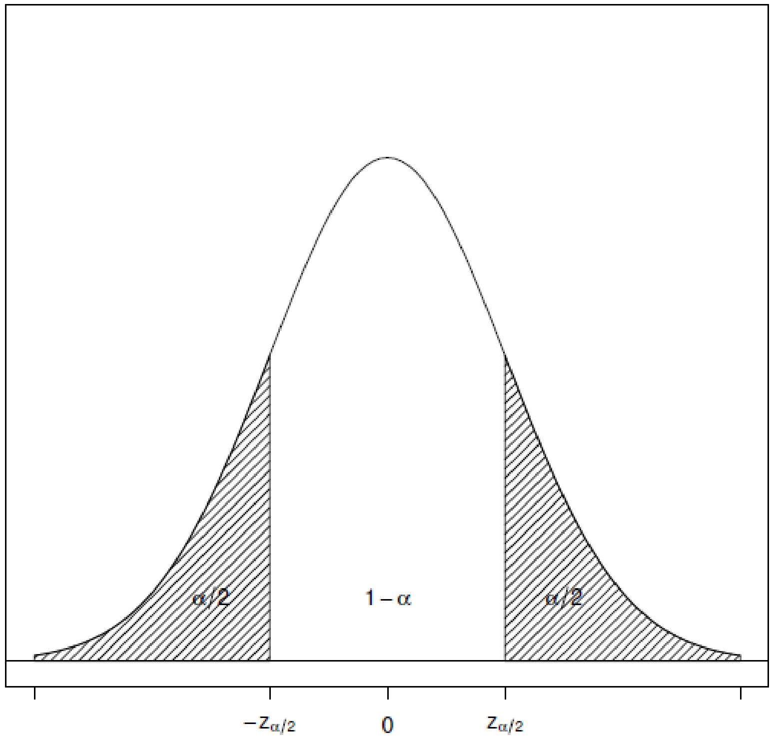
che conduce alla disuguaglianza di secondo grado in p:

$$p^2\left(n + z^2_{\frac{\alpha}{2}}\right) - p\left(2n\bar{X}_n + z^2_{\frac{\alpha}{2}}\right) + n\bar{X}_n{}^2 < 0$$

- Dato che p^2 è positivo, si ha che le soluzioni sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia:

$$C_n < p < \overline{C_n}$$

Densità normale standard



I.C. per il parametro p di una popolazione di Bernoulli

- Il sistema R mette a disposizione la funzione:

$$\textbf{polyroot}(\mathrm{c}(a_1, a_2, \dots, a_{n-1}, a_n))$$

per calcolare le radici reali e complesse di un'equazione:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x^1 + a_0 = 0$$

- Denotando con:

$$a_2 = n + z_{\frac{\alpha}{2}}^2 \qquad a_1 = -\left(2n\bar{x}_n + z_{\frac{\alpha}{2}}^2\right) \qquad a_0 = n\bar{x}_n^2$$

Le radici dell'equazione:

$$a_2 p^2 + a_1 p^1 + a_0 = 0$$

possono essere calcolate con:

$$\textbf{polyroot}(\mathrm{c}(a_0, a_1, a_2))$$

Esempio

- Consideriamo un campione campbern di ampiezza 30 contenente i risultati di lanci indipendenti di una moneta

```
> campbern<-c(0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0,
+ 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1)
```

- Il metodo dei momenti e della massima verosimiglianza hanno fornito come stima del parametro p la media campionaria $\bar{X}_n$

```
> stimap<-mean(campbern)
> stimap
[1] 0.5666667
```

- la stima del parametro p con il metodo dei momenti e con il metodo della massima verosimiglianza è $\hat{p} = 0.5667$

Esempio

- Determiniamo un intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro p

```
campbern <-c(0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0,
             1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1)
[1]  1.959964
alpha <- 1 - 0.95
qnorm(1 - alpha/2, mean = 0, sd = 1)
zalpha <- qnorm(1 - alpha/2, mean = 0, sd = 1)

n <- length(campbern)

a2 <- n + zalpha^2
a1 <- -(2 * n * mean(campbern) + zalpha^2)
a0 <- n * (mean(campbern))^2

polyroot(c(a0, a1, a2))
[1]  0.3919731+0i  0.7262251-0i
```

- Una stima dell'intervallo di confidenza per p è $(0.3919731, 0.7262251)$
- Si nota che la stima puntuale di p, $\hat{p} = 0.5667$ è **contenuta** nell'intervallo

Esempio (ii)

- Una ditta farmaceutica è interessata a stabilire l'efficacia di un nuovo farmaco per curare una data malattia
- Da un'indagine condotta su 900 pazienti affetti da questa malattia trova che il farmaco è efficace in 740 casi
- Determiniamo una stima dell'intervallo di confidenza di grado $1-\alpha=0.95$ per la probabilità p
- Supponiamo che la popolazione sia distribuita secondo Bernoulli, con p che denota la probabilità che il farmaco sia efficace
  - Il campione è di ampiezza $n=900$, dove 900 rappresenta il numero di pazienti esaminati
  - La media campionaria è $\overline{X}_{900}=\frac{740}{900}=0.822$
  - Poiché $\alpha=1-0.95=0.05$ si ha che $\frac{\alpha}{2}=0.025$

Esempio (ii)

- Usando R:

```
# Calcolo alpha per intervallo di confidenza 95%
alpha <- 1 - 0.95

# Quantile della normale standard per 97.5%
qnorm(1 - alpha/2, mean = 0, sd = 1)
# [1] 1.959964

zalpha <- qnorm(1 - alpha/2, mean = 0, sd = 1)

# Dimensione del campione
n <- 900

# Proporzione campionaria (740 successi su 900)
medcamp <- 740/900
medcamp
# [1] 0.8222222

# Coefficienti dell'equazione quadratica per l'intervallo di confidenza
a2 <- n + zalpha^2
a1 <- -(2 * n * medcamp + zalpha^2)
a0 <- n * medcamp^2

# Risoluzione equazione quadratica: a2*p^2 + a1*p + a0 = 0
polyroot(c(a0, a1, a2))
# [1] 0.7958901+0i 0.8458153-0i
```

Esempio (ii)

- Usando R:

```
# Calcolo alpha per intervallo di confidenza 95%
alpha <- 1 - 0.95

# Quantile della normale standard per 97.5%
qnorm(1 - alpha/2, mean = 0, sd = 1)
# [1] 1.959964
```

$z_{\frac{\alpha}{2}}$

```
zalpha <- qnorm(1 - alpha/2, mean = 0, sd = 1)

# Dimensione del campione
n <- 900

# Proporzione campionaria (740 successi su 900)
medcamp <- 740/900
medcamp
# [1] 0.8222222
```

Media campionaria e stima puntuale di p $(\hat{p})$

```
# Coefficienti dell'equazione quadratica per l'intervallo di confidenza
a2 <- n + zalpha^2
a1 <- -(2 * n * medcamp + zalpha^2)
a0 <- n * medcamp^2

# Risoluzione equazione quadratica: a2*p^2 + a1*p + a0 = 0
polyroot(c(a0, a1, a2))
# [1] 0.7958901+0i 0.8458153-0i
```

Intervallo di confidenza

STATISTICA E ANALISI DEI DATI

Parametro p di una popolazione Binomiale

I.C. per il parametro p di una popolazione Binomiale

- Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria:

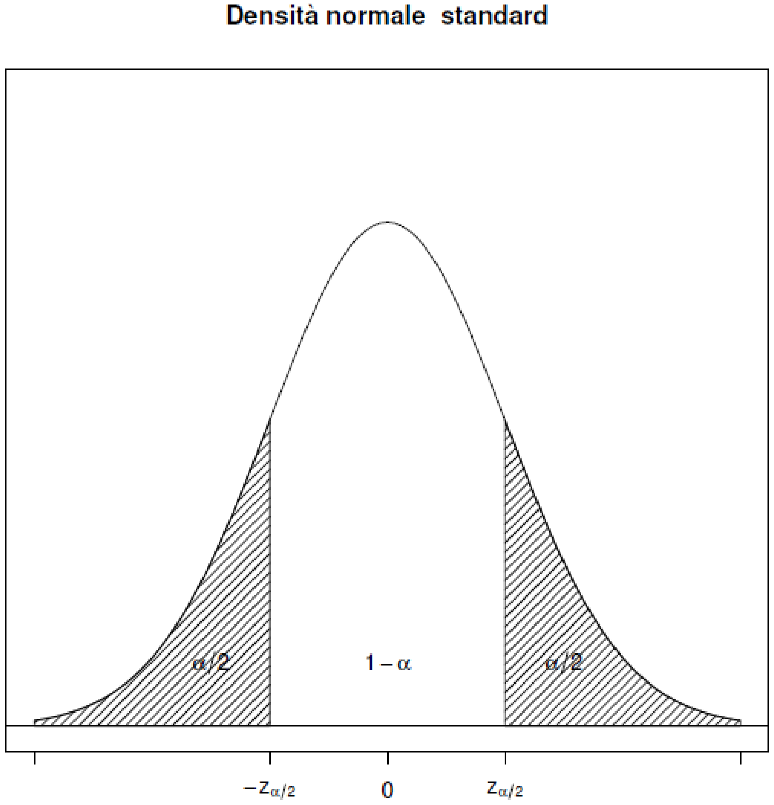
$$p^x(1-p)^{k-x} \quad x = 0,1,\dots,k \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = kp \qquad \sigma^2 = Var(X) = kp(1-p)$$

- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$



I.C. per il parametro p di una popolazione Binomiale

- Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria:

$$p^x(1-p)^{k-x} \quad x = 0,1,\ldots,k \quad (0 < p < 1)$$

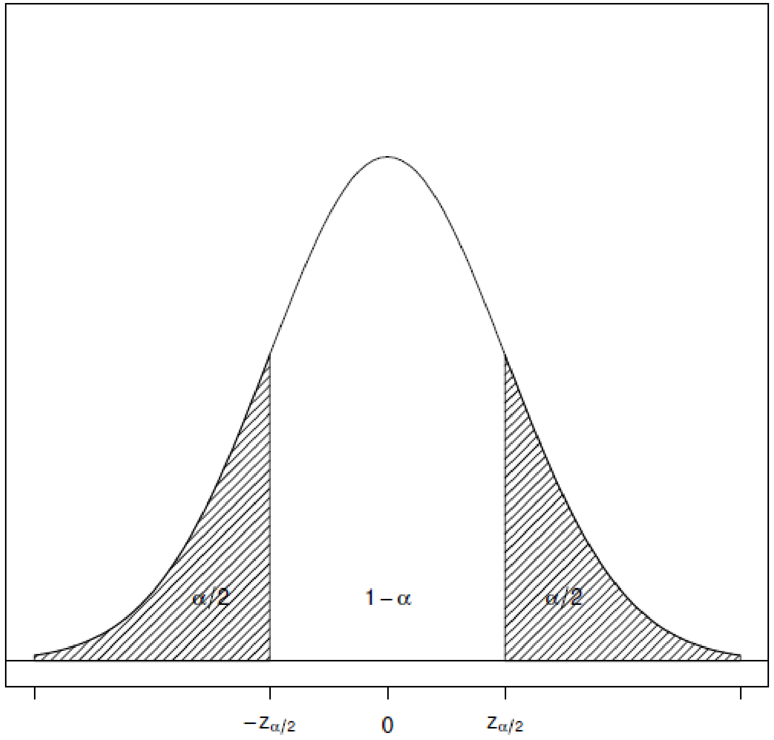
- Ricordando che:

$$\mu = E(X) = kp \qquad\qquad \sigma^2 = Var(X) = kp(1-p)$$

- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_n - kp}{\frac{\sqrt{kp(1-p)}}{\sqrt{n}}} =$$

Densità normale standard



I.C. per il parametro p di una popolazione Binomiale

- Consideriamo una popolazione di Bernoulli descritta da una variabile aleatoria:

$$p^x(1-p)^{k-x} \quad x = 0,1,\dots,k \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = kp \qquad\qquad \sigma^2 = Var(X) = kp(1-p)$$

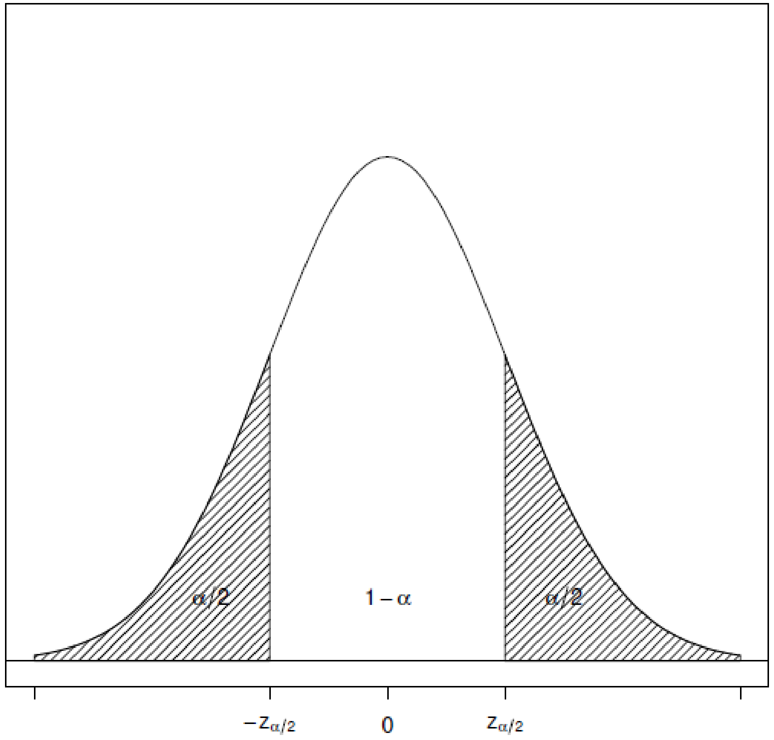
- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n}-kp}{\frac{\sqrt{kp(1-p)}}{\sqrt{n}}} = \sqrt{n}\frac{\overline{X_n}-kp}{\sqrt{kp(1-p)}}$$

- Quindi per campioni sufficientemente numerosi si ha che:

$$P\left(-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{X_n}-kp}{\sqrt{kp(1-p)}} < z_{\frac{\alpha}{2}}\right) \cong 1-\alpha$$

Densità normale standard



I.C. per il parametro p di una popolazione Binomiale

- Da cui si può risolvere la disuguaglianza:

$$-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} < z_{\frac{\alpha}{2}}$$

- Che è equivalente a:

$$\sqrt{n}\frac{\left(\overline{X_n} - kp\right)^2}{\sqrt{kp(1-p)}} < z_{\frac{\alpha}{2}}^2$$

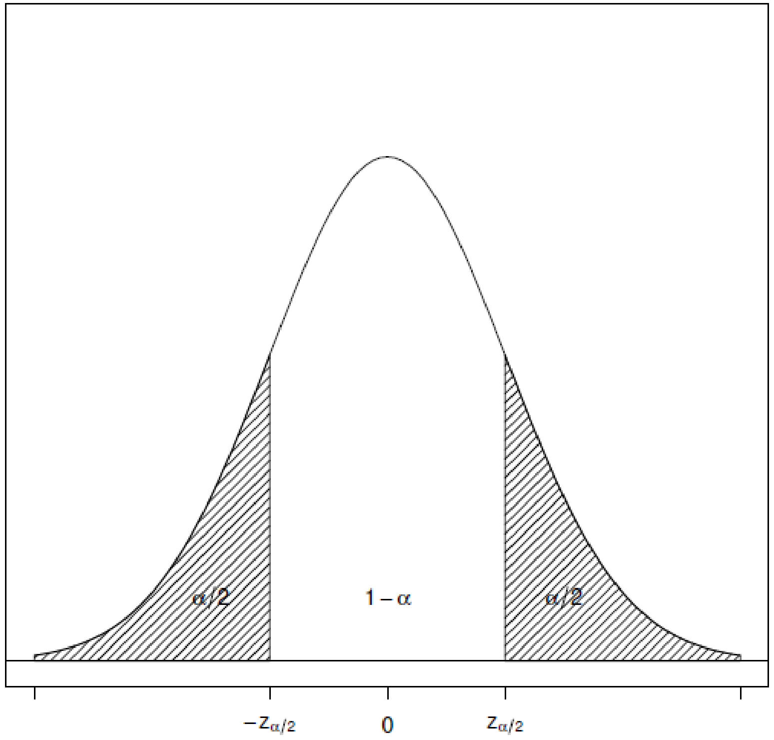
- che conduce alla disuguaglianza di secondo grado in p:

$$k\left(nk + z_{\frac{\alpha}{2}}^2\right)p^2 - k\left(2n\overline{X_n} + z_{\frac{\alpha}{2}}^2\right)p + n\overline{X_n}^2 < 0$$

- Dato che p^2 è positivo, si ha che le **soluzioni** sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia:

$$C_n < p < \overline{C_n}$$

Densità normale standard



I.C. per il parametro p di una popolazione Binomiale

- Il sistema R mette a disposizione la funzione:

$$\textbf{polyroot}(\mathrm{c}(a_1, a_2, \dots, a_{n-1}, a_n))$$

per calcolare le radici reali e complesse di un'equazione:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x^1 + a_0 = 0$$

- Denotando con:

$$a_2 = k\left(nk + z_{\frac{\alpha}{2}}^2\right) \qquad a_1 = -k\left(2n\overline{X_n} + z_{\frac{\alpha}{2}}^2\right) \qquad a_0 = n\overline{X_n}^2$$

- Le radici dell'equazione:

$$a_2 p^2 + a_1 p^1 + a_0 = 0$$

possono essere calcolate con:

$$\textbf{polyroot}(\mathrm{c}(a_0, a_1, a_2))$$

Esempio

- Consideriamo un campione campbinom di ampiezza 30 contenente come risultati il numero di successi ottenuti in k = 10 lanci indipendenti di una moneta

```
> campbinom<-c(3, 2, 6, 2, 4, 4, 7, 4, 6, 6, 5, 4, 5, 4, 8,
+ 1, 3, 7, 4, 0, 3, 7, 4, 4, 3, 2, 5, 5, 3, 2)
```

- Il metodo dei momenti e della massima verosimiglianza hanno fornito come stima del parametro p la media campionaria $\frac{\bar{x}_n}{k}$

```
> lanci<-10
> stimap<-mean(campbinom)/lanci
> stimap
[1] 0.41
```

- la stima del parametro p con il metodo dei momenti e con il metodo della massima verosimiglianza è $\hat{p} = 0.41$

Esempio

- Determiniamo un intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro p

```
alpha <- 1 - 0.95
qnorm(1 - alpha/2, mean = 0, sd = 1)
zalpha <- qnorm(1 - alpha/2, mean = 0, sd = 1)
 [1] 1.959964
n <- length(camphinom)

a2 <- lanci * (n * lanci + zalpha^2)
a1 <- lanci * (2 * n * mean(camphinom) + zalpha^2)
a0 <- n * (mean(camphinom))^2

polyroot(c(a0, a1, a2))
 [1] 0.3558239-0i 0.4664518+0i
```

- Una stima dell'intervallo di confidenza per p è $(0.3558239, 0.4664518)$
- Si nota che la stima puntuale di p, $\hat{p} = 0.41$ è contenuta nell'intervallo

STATISTICA E ANALISI DEI DATI

Parametro p di una popolazione geometrica modificata

I.C. per il parametro p di una Pop. Geometrica Mod.

- Consideriamo una popolazione di geometrica modificata descritta da una variabile aleatoria:

$$p(1-p)^{x-1} \quad x = 1,2,\dots \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = \frac{1}{p} \qquad\qquad \sigma^2 = Var(X) = \frac{(1-p)^2}{p^2}$$

- Possiamo applicare il teorema centrale di convergenza e si ha che:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

I.C. per il parametro p di una Pop. Geometrica Mod.

- Consideriamo una popolazione di geometrica modificata descritta da una variabile aleatoria:

$$p(1-p)^{x-1} \quad x = 1,2,\dots \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = \frac{1}{p} \qquad\qquad \sigma^2 = Var(X) = \frac{(1-p)^2}{p^2}$$

- Possiamo applicare il teorema centrale di convergenza e si ha che:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n} - \frac{1}{p}}{\sqrt{\frac{(1-p)}{np^2}}}$$

I.C. per il parametro p di una Pop. Geometrica Mod.

- Consideriamo una popolazione di geometrica modificata descritta da una variabile aleatoria:

$$p(1-p)^{x-1} \quad x = 1,2,\ldots \quad (0 < p < 1)$$

- Ricordando che:

$$\mu = E(X) = \frac{1}{p} \qquad \sigma^2 = Var(X) = \frac{(1-p)^2}{p^2}$$

- Possiamo applicare il teorema centrale di convergenza e si ha che:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n} - \frac{1}{p}}{\sqrt{\frac{(1-p)}{np^2}}} = \sqrt{n}\,\frac{p\overline{X_n} - 1}{\sqrt{1-p}}$$

- converge in distribuzione ad una variabile aleatoria normale standard

$$P\left(-z_{\frac{\alpha}{2}} < \sqrt{n}\,\frac{p\overline{X_n} - 1}{\sqrt{1-p}} < z_{\frac{\alpha}{2}}\right) \cong 1 - \alpha$$

I.C. per il parametro p di una Pop. Geometrica Mod.

- Da cui si può risolvere la disuguaglianza:

$$-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{p\overline{X_n} - 1}{\sqrt{1-p}} < z_{\frac{\alpha}{2}}$$

- Che è equivalente a:

$$\left[\sqrt{n}\frac{p\overline{X_n} - 1}{\sqrt{1-p}}\right]^2 < z_{\frac{\alpha}{2}}^2$$

- che conduce alla disuguaglianza di secondo grado in p:

$$n\bar{X}_n{}^2 p^2 - p\left(2n\bar{X}_n - z_{\frac{\alpha}{2}}^2\right) + n - z_{\frac{\alpha}{2}}^2 < 0$$

- Dato che p^2 è positivo, si ha che le soluzioni sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia:

$$\underline{C_n} < p < \overline{C_n}$$

I.C. per il parametro p di una Pop. Geometrica Mod.

- Il sistema R mette a disposizione la funzione:

$$\textbf{polyroot}(\mathrm{c}(a_1, a_2, \dots, a_{n-1}, a_n))$$

per calcolare le radici reali e complesse di un'equazione:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x^1 + a_0 = 0$$

- Denotando con:

$$a_2 = n\bar{X}_n^{\ 2} \qquad a_1 = -\left(2n\bar{X}_n - z_{\frac{\alpha}{2}}^2\right) \qquad a_0 = n - z_{\frac{\alpha}{2}}^2$$

- Le radici dell'equazione:

$$a_2 p^2 + a_1 p^1 + a_0 = 0$$

possono essere calcolate con:

$$\textbf{polyroot}(\mathrm{c}(a_0, a_1, a_2))$$

Esempio

- In una produzione di aghi con una macchina automatica, vengono scartati quelli la cui lunghezza è inferiore a 2 cm
- Numerando gli aghi prodotti, denotiamo con X la variabile aleatoria che descrive il numero associato al primo ago imperfetto prodotto
  - la distribuzione di X è geometrica modificata di parametro p (probabilità che l'ago sia imperfetto in una singola produzione)
- Se si effettuano 100 osservazioni di X, si ha $\bar{X}_{100} = 10.5$
- Determinare un intervallo di confidenza per il parametro p con un grado di confidenza $1 - \alpha = 0.96$

Esempio

- Determinare un intervallo di confidenza per il parametro p con un grado di confidenza $1-\alpha = 0.96$
- Si ha quindi:
  - $n = 100$
  - $\bar{x}_{100} = 10.5$ (Stima puntuale di $\frac{1}{p}$)
  - $\alpha = 0.04 \Rightarrow \frac{\alpha}{2} = 0.02$
- Una stima dell'intervallo di confidenza per p è $(0.0764, 0.1137)$
- Si nota che la stima puntuale di p, $\hat{p} = \frac{1}{\bar{x}_{100}} = 0.095$ è contenuta nell'intervallo

```r
# Calcolo alpha per intervallo di confidenza 96%
alpha <- 1 - 0.96
# alpha = 0.04

# Quantile della normale standard per 1 - alpha/2 = 0.98
qnorm(1 - alpha/2, mean = 0, sd = 1)
# [1] 2.053749

zalpha <- qnorm(1 - alpha/2, mean = 0, sd = 1)

# Dimensione del campione
n <- 100

# Media campionaria
medcamp <- 10.5

# Coefficienti dell'equazione quadratica
a2 <- n * medcamp^2
a1 <- -(2 * n * medcamp - zalpha^2)
a0 <- n - zalpha^2

# Risoluzione equazione quadratica: a2*p^2 + a1*p + a0 = 0
polyroot(c(a0, a1, a2))
# [1] 0.07644102+0i 0.11365260-0i
```

STATISTICA E ANALISI DEI DATI

Intervallo di confidenza per il parametro λ di una popolazione di Poisson

I.C. per il parametro λ di una popolazione di Poisson

- Consideriamo una popolazione di geometrica modificata descritta da una variabile aleatoria:

$$\frac{\lambda^x}{x!}e^{-\lambda} \quad x = 0,1,\ldots \quad (\lambda > 0)$$

- Ricordando che:

$$E(X) = \lambda \qquad\qquad Var(X) = \lambda$$

- Possiamo applicare il teorema centrale di convergenza e si ha che:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n} - \lambda}{\sqrt{\frac{\lambda}{n}}} = \sqrt{n}\frac{\overline{X_n} - \lambda}{\sqrt{\lambda}}$$

- converge in distribuzione ad una variabile aleatoria normale standard

$$P\left(-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{X_n} - \lambda}{\sqrt{\lambda}} < z_{\frac{\alpha}{2}}\right) \cong 1 - \alpha$$

I.C. per il parametro λ di una popolazione di Poisson

- Da cui si può risolvere la disuguaglianza:

$$-z_{\frac{\alpha}{2}} < \sqrt{n}\frac{\overline{x_n} - \lambda}{\sqrt{\lambda}} < z_{\frac{\alpha}{2}}$$

- Che è equivalente a:

$$\left[\sqrt{\frac{n}{\lambda}}(\overline{x_n} - \lambda)\right]^2 < z_{\frac{\alpha}{2}}^2$$

- che conduce alla disuguaglianza di secondo grado in p:

$$n\lambda^2 - \lambda\left(2n\bar{X}_n + z_{\frac{\alpha}{2}}^2\right) + n\bar{X}_n{}^2 < 0$$

- Dato che λ^2 è positivo, si ha che le soluzioni sono interne all'intervallo delle radici della corrispondente equazione di secondo grado, ossia:

$$C_n < \lambda < \overline{C_n}$$

I.C. per il parametro λ di una popolazione di Poisson

- Il sistema R mette a disposizione la funzione:

$$\textbf{polyroot}(\mathrm{c}(a_1, a_2, \dots, a_{n-1}, a_n))$$

per calcolare le radici reali e complesse di un’equazione:

$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x^1 + a_0 = 0$$

- Denotando con:

$$a_2 = n \qquad a_1 = -\left(2n\bar{X}_n + z_{\frac{\alpha}{2}}^2\right) \qquad a_0 = n\bar{X}_n^{\ 2}$$

- Le radici dell’equazione:

$$a_2 \lambda^2 + a_1 \lambda^1 + a_0 = 0$$

possono essere calcolate con:

$$\textbf{polyroot}(\mathrm{c}(a_0, a_1, a_2))$$

Esempio

- Si supponga che il numero $N(t)$ di chiamate che arrivano ad un centralino telefonico nell'intervallo $(0, t)$ sia distribuito secondo Poisson

$$P(N(t) = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad x = 0,1, \dots \quad (\lambda > 0)$$

con valore medio $E[N(t)] = \lambda t$ e $Var[N(t)] = \lambda t$

- Se in 100 osservazioni effettuate in intervalli di tempo di $t = 10$ minuti si riscontra che <u>in media sono state effettuate 4</u> chiamate, si determini una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ per il parametro λ

- Si ha quindi:
  - $n = 100$
  - $t = 10$
  - $\bar{X}_{100} = 4$ (Stima puntuale di 10λ)
  - $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$
  - $1 - \frac{\alpha}{2} = 0.975$

Esempio

- Si ha quindi:
  - $n = 100$
  - $t = 10$
  - $\bar{X}_{100} = 4$ (Stima puntuale di 10λ)
  - $\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$
  - $1 - \frac{\alpha}{2} = 0.975$

```r
# Calcolo alpha per intervallo di confidenza 95%
alpha <- 1 - 0.95
# alpha = 0.05

# Quantile della normale standard per 97.5%
qnorm(1 - alpha/2, mean = 0, sd = 1)
# [1] 1.959964

zalpha <- qnorm(1 - alpha/2, mean = 0, sd = 1)

# Dimensione del campione
n <- 100

# Media campionaria
medcamp <- 4

# Parametro tempo
tempo <- 10

# Coefficienti dell'equazione quadratica
a2 <- n
a1 <- -(2 * n * medcamp + zalpha^2)
a0 <- n * medcamp^2

# Risoluzione equazione quadratica e divisione per tempo
polyroot(c(a0, a1, a2)) / tempo
# [1] 0.3626744+0i 0.4411670-0i
```

$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

- Una stima dell'intervallo di confidenza per λ è $(0.3627, 0.4412)$
- Si nota che la stima puntuale di λ, $\hat{\lambda} = \frac{4}{10} = 0.04$ è contenuta nell'intervallo

STATISTICA E ANALISI DEI DATI

Intervallo di confidenza per il parametro ϑ di una popolazione uniforme

I.C. per il parametro ϑ di una popolazione uniforme

- Consideriamo una popolazione uniforme descritta da una variabile aleatoria:

$$F(x) = \frac{1}{\vartheta} \qquad (0 < x < \vartheta)$$

- Ricordando che:

$$E(X) = \frac{\vartheta}{2} \qquad Var(X) = \frac{\vartheta^2}{12n}$$

- Possiamo applicare il teorema centrale di convergenza e si ha che:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n} - \frac{\vartheta}{2}}{\frac{\vartheta}{\sqrt{12n}}} = \sqrt{3n}\left(\frac{2\overline{X_n}}{\vartheta} - 1\right)$$

- converge in distribuzione ad una variabile aleatoria normale standard

$$P\left(2\overline{X_n}\left(1 + \frac{z_{\frac{\alpha}{2}}}{\sqrt{3n}}\right)^{-1} < \vartheta < 2\overline{X_n}\left(1 - \frac{z_{\frac{\alpha}{2}}}{\sqrt{3n}}\right)^{-1}\right) \cong 1 - \alpha$$

Esempio

- Supponiamo di considerare i tempi misurati in ore, e supposti uniformi in un intervallo $(0, \vartheta)$, necessari per soddisfare le richieste di 100 utenti che accedono ad un centro di calcolo
- Se si riscontra che il tempo medio per soddisfare le richieste è di 1.5 ore
  - Determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.98$ per ϑ
- Si ha quindi:
  - $n = 100$
  - $\bar{X}_{100} = 1.5$ (Stima puntuale di $\frac{\vartheta}{2}$)
  - $\alpha = 0.02 \Rightarrow \frac{\alpha}{2} = 0.01$
  - $1 - \frac{\alpha}{2} = 0.99$

Esempio

- Si ha quindi:
  - $n = 100$
  - $\bar{X}_{100} = 1.5$ (Stima puntuale di $\frac{\vartheta}{2}$)
  - $\alpha = 0.02 \Rightarrow \frac{\alpha}{2} = 0.01$
  - $1 - \frac{\alpha}{2} = 0.99$
- Una stima dell'intervallo di confidenza per $\frac{\vartheta}{2}$ è

$$\left(\frac{2.644}{2}, \frac{3.465}{2}\right) = (1.322, 1.733)$$

- Si nota che la stima puntuale di ϑ, $\hat{\vartheta} = \frac{\vartheta}{2} = 1.5$ è contenuta nell'intervallo

```r
# Calcolo alpha per intervallo di confidenza 98%
alpha <- 1 - 0.98
# alpha = 0.02

# Quantile della normale standard per 1 - alpha/2 = 0.99
qnorm(1 - alpha/2, mean = 0, sd = 1)
# Risultato: [1] 2.326348

# Dimensione del campione
n <- 100

# Media campionaria
m <- 1.5

# Calcolo limite inferiore dell'intervallo di confidenza
2 * m / (1 + qnorm(1 - alpha/2, mean = 0, sd = 1) / sqrt(3 * n))
# [1] 2.644776

# Calcolo limite superiore dell'intervallo di confidenza
2 * m / (1 - qnorm(1 - alpha/2, mean = 0, sd = 1) / sqrt(3 * n))
# [1] 3.465451
```

$z_{\frac{\alpha}{2}} = z_{0.01} = 2.32$

STATISTICA E ANALISI DEI DATI

Intervallo di confidenza per il parametro λ di una popolazione esponenziale

I.C. per il parametro λ di una popolazione esponenziale

- Consideriamo una popolazione esponenziale descritta da una variabile aleatoria:

$$f_X(x) = \lambda e^{-\lambda x} \qquad x > 0 \; (\lambda > 0)$$

- Ricordando che:

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

- Possiamo applicare il teorema centrale di convergenza e si ha che:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X_n} - \frac{1}{\lambda}}{\frac{1}{\lambda\sqrt{n}}} = \sqrt{n}\,\frac{\overline{X_n} - \frac{1}{\lambda}}{\frac{1}{\lambda}} = \sqrt{n}(\lambda\overline{X_n} - 1)$$

- converge in distribuzione ad una variabile aleatoria normale standard

- Per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1 - \alpha$ per il parametro $\frac{1}{\lambda}$ può essere determinato richiedendo che

$$P\left(-z_{\frac{\alpha}{2}} < \sqrt{n}(\lambda\overline{X_n} - 1) < z_{\frac{\alpha}{2}}\right) \cong 1 - \alpha \text{ ossia } P\left(\overline{X_n}\left(1 + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\right)^{-1} < \lambda < \overline{X_n}\left(1 - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\right)^{-1}\right) \cong 1 - \alpha$$

Esempio

- Si supponga che la durata delle conversazioni effettuate ad un telefono pubblico sia distribuita esponenzialmente con valore medio non noto $\frac{1}{\lambda}$

$$P(N(t) = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \quad x = 0,1, \dots \quad (\lambda > 0)$$

- Se in 100 osservazioni si riscontra che in media la durata delle conversazioni degli utenti è di 3 minuti, determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.94$ per la durata media delle conversazioni
- Si ha quindi:
  - $n = 100$
  - $\bar{X}_{100} = 3$ (Stima puntuale di $\frac{1}{\lambda}$)
  - $\alpha = 0.06 \Rightarrow \frac{\alpha}{2} = 0.03$
  - $1 - \frac{\alpha}{2} = 0.97$

Esempio

- Si ha quindi:
  - $n = 100$
  - $\bar{X}_{100} = 3$ (Stima puntuale di $\frac{1}{\lambda}$)
  - $\alpha = 0.06 \Rightarrow \frac{\alpha}{2} = 0.03$
  - $1 - \frac{\alpha}{2} = 0.97$
- Una stima dell'intervallo di confidenza per $\frac{1}{\lambda}$ è

$$(2.525, 3.694)$$

- Si nota che la stima puntuale di λ, $\hat{\lambda} = 3$ è contenuta nell'intervallo

```r
# Calcolo alpha per intervallo di confidenza 94%
alpha <- 1 - 0.94
# alpha = 0.06

# Quantile della normale standard per 1 - alpha/2 = 0.97
qnorm(1 - alpha/2, mean = 0, sd = 1)
# [1] 1.880794
```

$\boxed{1.880794} \rightarrow z_{\frac{\alpha}{2}} = z_{0.03} = 1.88$

```r
# Dimensione del campione
n <- 100

# Media campionaria
m <- 3

# Calcolo limite inferiore dell'intervallo di confidenza
m / (1 + qnorm(1 - alpha/2, mean = 0, sd = 1) / sqrt(n))
# [1] 2.525084

# Calcolo limite superiore dell'intervallo di confidenza
m / (1 - qnorm(1 - alpha/2, mean = 0, sd = 1) / sqrt(n))
# [1] 3.694942
```

STATISTICA E ANALISI DEI DATI

Confronto tra due popolazioni

Confronto tra due popolazioni

- Spesso i ricercatori sono interessati a stimare la differenza tra le medie di due distinte popolazioni
- In questo caso occorre considerare due campioni casuali indipendenti:

$$X_1, X_2, \dots, X_{n_1} \quad \text{e} \quad Y_1, Y_2, \dots, Y_{n_2}$$

Di ampiezza n_1 e n_2 e calcolare la media campionaria per ciascun campione $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$

- Come posso confrontare le due popolazioni?

Confronto tra due popolazioni

- Spesso i ricercatori sono interessati a stimare la differenza tra le medie di due distinte popolazioni
- In questo caso occorre considerare due campioni casuali indipendenti:

$$X_1, X_2, \dots, X_{n_1} \quad \text{e} \quad Y_1, Y_2, \dots, Y_{n_2}$$

Di ampiezza n_1 e n_2 e calcolare la media campionaria per ciascun campione $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$

- Come posso confrontare le due popolazioni?
  - Soluzione 1:
    - Si può calcolare la differenza tra le due medie campionarie $\overline{X_{n_1}} - \overline{Y_{n_2}}$
  - Problema:
    - Non si può essere certi che la differenza tra le medie campionarie corrisponda alla differenza effettiva tra le medie delle due popolazioni

Confronto tra due popolazioni

- Spesso i ricercatori sono interessati a stimare la differenza tra le medie di due distinte popolazioni
- In questo caso occorre considerare due campioni casuali indipendenti:

$$X_1, X_2, \dots, X_{n_1} \quad \text{e} \quad Y_1, Y_2, \dots, Y_{n_2}$$

Di ampiezza n_1 e n_2 e calcolare la media campionaria per ciascun campione $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$

- Come posso confrontare le due popolazioni?
  - <u>Soluzione 1</u>:
    - Si può calcolare la differenza tra le due medie campionarie $\overline{X_{n_1}} - \overline{Y_{n_2}}$
  - <u>Problema</u>:
    - Non si può essere certi che la differenza tra le medie campionarie corrisponda alla differenza effettiva tra le medie delle due popolazioni
  - <u>Soluzione 2</u>:
    - Costruire un **intervallo di confidenza** per la differenza tra le due medie con un certo grado di fiducia $1 - \alpha$, scelto dal decisore

Confronto tra due popolazioni

- Consideriamo due popolazioni descritte dalle variabili aleatorie X e Y indipendenti aventi:

$$E(X) = \mu_1 \qquad E(Y) = \mu_2$$
$$Var(X) = {\sigma^2}_1 \qquad Var(Y) = {\sigma^2}_2$$

la distribuzione della differenza $X - Y$ avrà valore medio e varianza:

$$E(X - \mathrm{Y}) = E(X) - E(Y) = \mu_1 - \mu_2$$
$$Var(X - \mathrm{Y}) = Var(X) - Var(Y) = {\sigma^2}_1 - {\sigma^2}_2$$

Un intervallo di confidenza $(C_n; \overline{C_n})$ per $\mu_1 - \mu_2$:

$$P\left(C_n < \mu_1 - \mu_2 < \overline{C_n}\right) = 1 - \alpha$$

Dove C_n e $\overline{C_n}$ sono due statistiche dipendenti dai campioni estratti dalle due popolazioni

Interpretazione intervallo di confidenza

- Un intervallo di confidenza $(C_n; \overline{C_n})$ per $\mu_1 - \mu_2$:

$$P\left(C_n < \mu_1 - \mu_2 < \overline{C_n}\right) = 1 - \alpha$$

Dove C_n e $\overline{C_n}$ sono due statistiche dipendenti dai campioni estratti dalle due popolazioni

- Se il limite inferiore e il limite superiore sono entrambi negativi allora $\mu_1 - \mu_2 < 0$
  - ciò implica che la media della prima popolazione è **inferiore** alla media della seconda popolazione con un grado di confidenza $1 - \alpha$

Interpretazione intervallo di confidenza

- Un intervallo di confidenza $(C_n; \overline{C_n})$ per $\mu_1 - \mu_2$:

$$P\left(C_n < \mu_1 - \mu_2 < \overline{C_n}\right) = 1 - \alpha$$

Dove C_n e $\overline{C_n}$ sono due statistiche dipendenti dai campioni estratti dalle due popolazioni

- Se il limite inferiore e il limite superiore sono entrambi negativi allora $\mu_1 - \mu_2 < 0$
  - ciò implica che la media della prima popolazione è **inferiore** alla media della seconda popolazione con un grado di confidenza $1 - \alpha$
- Se il limite inferiore e il limite superiore sono entrambi positivi allora $\mu_1 - \mu_2 > 0$
  - ciò implica che la media della prima popolazione è **superiore** alla media della seconda popolazione con un grado di confidenza $1 - \alpha$

Interpretazione intervallo di confidenza

- Un intervallo di confidenza $(C_n; \overline{C_n})$ per $\mu_1 - \mu_2$:

$$P\left(C_n < \mu_1 - \mu_2 < \overline{C_n}\right) = 1 - \alpha$$

Dove C_n e $\overline{C_n}$ sono due statistiche dipendenti dai campioni estratti dalle due popolazioni

- Se il limite inferiore e il limite superiore sono entrambi negativi allora $\mu_1 - \mu_2 < 0$
  - ciò implica che la media della prima popolazione è **inferiore** alla media della seconda popolazione con un grado di confidenza $1 - \alpha$
- Se il limite inferiore e il limite superiore sono entrambi positivi allora $\mu_1 - \mu_2 > 0$
  - ciò implica che la media della prima popolazione è **superiore** alla media della seconda popolazione con un grado di confidenza $1 - \alpha$
- se l'intervallo contiene lo zero, ossia il limite inferiore risulta negativo e il limite superiore positivo
  - allora con un grado di confidenza $1 - \alpha$ non si può affermare che la media di una popolazione sia superiore alla media dell'altra popolazione

STATISTICA E ANALISI DEI DATI

Confronto tra due popolazioni Normali

Confronto tra due popolazioni

- Siano:

$$X_1, X_2, \dots, X_{n_1} \quad \text{e} \quad Y_1, Y_2, \dots, Y_{n_2}$$

due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad e \quad Y \sim \mathrm{N}(\mu, \sigma^2)$$

Vogliamo analizzare i seguenti problemi:

- determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 sono note
- determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 sono NON note

I.C. di grado $1-\alpha$ per $\mu_1 - \mu_2$ con varianze σ_1^2 e σ_2^2 note

- Consideriamo le medie campionarie di X e Y:

$$\overline{X}_{n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \overline{Y}_{n_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

- Poiché $X_1, X_2, \dots, X_{n_1}$ e $Y_1, Y_2, \dots, Y_{n_2}$ sono indipendenti, la statistica $\overline{X_{n_1}} - \overline{Y_{n_2}}$ è distribuita normalmente con valore medio e varianza

$$E(\overline{X}_{n_1} - \overline{Y}_{n_2}) = \mu_1 - \mu_2, \qquad \mathrm{Var}(\overline{X}_{n_1} - \overline{Y}_{n_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

ottenute considerando la proprietà di linearità del valore medio e le proprietà della varianza per combinazioni lineari di variabili aleatorie indipendenti

- **Proprietà**:
  - La proprietà di linearità del valore medio afferma che il valore medio di una combinazione lineare di variabili casuali è uguale alla combinazione lineare dei loro valori medi

I.C. di grado 1−α per $\mu_1 - \mu_2$ con varianze σ_1^2 e σ_2^2 note

- Consideriamo la variabile **pivot** che dipende da $\mu_1 - \mu_2$ non noti:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

I.C. di grado $1-\alpha$ per $\mu_1 - \mu_2$ con varianze σ_1^2 e σ_2^2 note

- Consideriamo la variabile **pivot** che dipende da $\mu_1 - \mu_2$ non noti:

$$\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow Z_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Z_n è caratterizzata da una densità normale standard

I.C. di grado $1-\alpha$ per $\mu_1 - \mu_2$ con varianze σ_1^2 e σ_2^2 note

- Consideriamo la variabile **pivot** che dipende da $\mu_1 - \mu_2$ non noti:

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \longrightarrow Z_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Z_n è caratterizzata da una densità normale standard
- Pertanto, utilizzando il metodo pivotale si ha

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) = 1 - \alpha$$

- Isolando $\mu_1 - \mu_2$ si ha che la stima dell'intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ è:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Esempio

- Osservando un campione di 150 lampadine prodotte dall'**industria A** si riscontra che la durata media di una lampadina è 1400 ore
- Osservando un campione di 100 lampadine prodotte dall'**industria B** si riscontra che la durata media di una lampadina è 1200 ore
- Supponendo che i campioni casuali siano stati estratti indipendentemente da due popolazioni normali $X \sim \mathrm{N}(\mu, \sigma^2)$ e $Y \sim \mathrm{N}(\mu, \sigma^2)$ con rispettive deviazioni standard $\sigma_1 = 120$ e $\sigma_2 = 80$
  - determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per $\mu_1 - \mu_2$ delle lampadine prodotte dalle due industrie
- Si ha quindi:
  - $\overline{x_{150}} = 1400$
  - $\overline{y_{100}} = 1200$
  - $\sigma_1^2 = 120 * 120 = 14400$ e $\sigma_2^2 = 80 * 80 = 6400$
- Dobbiamo determinare $z_{\frac{\alpha}{2}} = z_{0.005}$

Esempio

- Dobbiamo determinare $z_{\frac{\alpha}{2}} = z_{0.005}$

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
> n1<-150
> n2<-100
> m1<-1400
> m2<-1200
> sigma1<-120
> sigma2<-80
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] 167.4181
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] 232.5819
```

Esempio

- Dobbiamo determinare $z_{\frac{\alpha}{2}} = z_{0.005}$

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
> n1<-150
> n2<-100
> m1<-1400
> m2<-1200
> sigma1<-120
> sigma2<-80
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] 167.4181
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
[1] 232.5819
```

$$z_{\frac{\alpha}{2}} = z_{0.005} = 2.575$$

$$\overline{x}_{n_1} - \overline{y}_{n_2} - \boxed{z_{\alpha/2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + \boxed{z_{\alpha/2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Si ha che la stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per $\mu_1 - \mu_2$ delle lampadine prodotte dalle due industrie è $(167.42, 232.582)$
- Poiché il limite inferiore ed il limite superiore sono positivi, si deduce che le lampadine prodotte dall'industria A hanno una durata media superiore a quella delle lampadine prodotte dall'industria B con un grado di fiducia del 99%

I.C. di grado $1-\alpha$ per $\mu_1 - \mu_2$ con varianze σ_1^2 e σ_2^2 NON note

- Siano:

$$X_1, X_2, \dots, X_{n_1} \quad \text{e} \quad Y_1, Y_2, \dots, Y_{n_2}$$

due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni normali

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad e \quad Y \sim \mathrm{N}(\mu, \sigma^2)$$

- Determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ con grandi valori di n_1 e n_2
- Denotiamo con $S_{n_1}^2$ e $S_{n_2}^2$ le varianze campionarie delle due popolazioni
  - Essendo

$$E(S_{n_1}^2) = \sigma_1^2,$$

$$E(\tilde{S}_{n_2}^2) = \sigma_2^2,$$

  - le varianze campionarie delle due popolazioni normali sono stimatori corretti e consistenti delle varianze delle due popolazioni

I.C. di grado $1-\alpha$ per $\mu_1 - \mu_2$ con varianze σ_1^2 e σ_2^2 NON note

- Applicando il metodo pivotale in forma approssimata si ha:

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{S^2_{n_1}/n_1 + \tilde{S}^2_{n_2}/n_2}} < z_{\alpha/2}\right) \simeq 1 - \alpha$$

- Da cui si ricava:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{\tilde{s}^2_{n_2}}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{\tilde{s}^2_{n_2}}{n_2}}$$

dove $\bar{X}_{n_1}$ e $\bar{Y}_{n_2}$ sono le medie campionarie dei due campioni e $S^2_{n_1}$ e $S^2_{n_2}$ sono le varianze campionarie dei due campioni

Esempio

- Una ditta farmaceutica è interessata a stabilire l'efficacia di un nuovo tipo di sonnifero
  - Un'indagine condotta su 50 pazienti mostra che il **nuovo** sonnifero conduce ad un numero medio di ore di sonno per individuo di 7.82 ore con una deviazione standard campionaria di 0.24 ore
  - Un'indagine condotta su altri 100 pazienti mostra che il **vecchio** tipo di sonnifero conduce ad un numero medio di ore di sonno per individuo di 6.75 ore con una deviazione standard campionaria di 0.30 ore
- Supponendo che i campioni casuali siano stati estratti indipendentemente da due popolazioni normali $X \sim \mathrm{N}(\mu, \sigma_1^2)$ e $Y \sim \mathrm{N}(\mu, \sigma_2^2)$:
  - determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per $\mu_1 - \mu_2$ tra i numeri medi di ore di sonno degli individui delle due popolazioni
- Si ha quindi:
  - $\overline{X_{50}} = 7.82$
  - $\overline{Y_{100}} = 6.75$
  - $s_{50}^2 = 0.24 * 0.24 = 0.0576$ e $s_{100}^2 = 0.3 * 0.3 = 0.09$
- Dobbiamo determinare $z_{\frac{\alpha}{2}} = z_{0.005}$

Esempio

- Dobbiamo determinare $z_{\frac{\alpha}{2}} = z_{0.005}$

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
>
> n1<-50
> n2<-100
> m1<-7.82
> m2<-6.75
> s1<-0.24
> s2<-0.30
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 0.9533175
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 1.186683
```

$z_{\frac{\alpha}{2}} = z_{0.005} = 2.575$

- Si ha che la stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per $\mu_1 - \mu_2$ tra i numeri medi di ore di sonno dalle due popolazioni è $(0.9533, 1.1866)$

- Poiché il limite inferiore ed il limite superiore sono positivi, può dedurre che il nuovo sonnifero è più efficace rispetto al precedente sonnifero con un grado di fiducia del 99%

STATISTICA E ANALISI DEI DATI

Confronto tra due popolazioni di Bernoulli

Confronto tra due popolazioni di Bernoulli

- Consideriamo due popolazioni di Bernoulli descritte da una variabile aleatoria:

$$p_1{}^x(1-p_1)^{1-x} \quad x = 0{,}1 \quad (0 < p_1 < 1)$$
$$p_2{}^y(1-p_2)^{1-y} \quad y = 0{,}1 \quad (0 < p_2 < 1)$$

- E siano $X_1, X_2, \dots, X_{n_1}$ e $Y_1, Y_2, \dots, Y_{n_2}$ due campioni casuali indipendenti
- Ricordando che:

$$E(X) = p \qquad\qquad Var(X) = p(1-p)$$

- Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $p_1 - p_2$ tra i parametri delle due popolazioni per grandi valori di n_1 e n_2
- Denotando con $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$ le medie campionarie delle due popolazioni
- Possiamo applicare il teorema centrale di convergenza:

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (p_1 - p_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \xrightarrow{d} Z;$$

Confronto tra due popolazioni di Bernoulli

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (p_1 - p_2)}{\sqrt{p_1\,(1-p_1)/n_1 + p_2\,(1-p_2)/n_2}} \xrightarrow{d} Z;$$

- Poiché $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$ le medie campionarie sono stimatori corretti e consistenti di p_1 e p_2 per campioni sufficientemente numerosi, l'intervallo di confidenza di grado $1-\alpha$ per la differenza $p_1 - p_2$ può essere determinato supponendo che

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (p_1 - p_2)}{\sqrt{\overline{X}_{n_1}\,(1-\overline{X}_{n_1})/n_1 + \overline{Y}_{n_2}\,(1-\overline{Y}_{n_2})/n_2}} < z_{\alpha/2}\right) \simeq 1-\alpha$$

- Quindi una stima approssimata è data da:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}\,(1-\overline{x}_{n_1})}{n_1} + \frac{\overline{y}_{n_2}\,(1-\overline{y}_{n_2})}{n_2}} < p_1 - p_2$$
$$< \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}\,(1-\overline{x}_{n_1})}{n_1} + \frac{\overline{y}_{n_2}\,(1-\overline{y}_{n_2})}{n_2}}$$

Esempio

- In un sondaggio su una certa trasmissione televisiva sono stati intervistati due campioni:
  - uno di adulti (400 individui)
  - uno di giovani (600 individui)
- I giovani che hanno espresso gradimento per la trasmissione televisiva sono stati 300, gli adulti invece sono stati 100
- Si desidera determinare l'intervallo di confidenza di grado $1-\alpha=0.95$ e di grado $1-\alpha=0.99$ per la differenza tra le frequenze relative degli adulti e dei giovani favorevoli alla trasmissione televisiva
- Stimiamo che le popolazioni siano distribuite in modo Bernoulliano, occorre stimare l'intervallo di confidenza per p_1-p_2
- Si ha quindi:
  - $\overline{x_{400}}=\frac{100}{400}=0.25$
  - $\overline{y_{600}}=\frac{300}{600}=0.50$

Esempio

- Determiniamo l'intervallo di confidenza di grado $1 - \alpha = 0.95$:

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
>
> n1<-400
> n2<-600
> m1<-100/400
> m2<-300/600
> rad<-sqrt(m1*(1-m1)/n1+m2*(1-m2)/n2)
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.3083206
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.1916794
```

Esempio

- Determiniamo l'intervallo di confidenza di grado $1 - \alpha = 0.95$:

```
> alpha<-1-0.95
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964
>
> n1<-400
> n2<-600
> m1<-100/400
> m2<-300/600
> rad<-sqrt(m1*(1-m1)/n1+m2*(1-m2)/n2)
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.3083206
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.1916794
```

- Determiniamo l'intervallo di confidenza di grado $1 - \alpha = 0.99$:

```
> alpha<-1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
>
> n1<-400
> n2<-600
> m1<-100/400
> m2<-300/600
> rad<-sqrt(m1*(1-m1)/n1+m2*(1-m2)/n2)
>
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.3266463
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*rad
[1] -0.1733537
```

- Si nota che aumentando il grado di confidenza da 0.95 a 0.99 aumenta l'ampiezza dell'intervallo di confidenza stimato
- Inoltre, essendo $p_1 - p_2 < 0$, è possibile concludere che, relativamente alla trasmissione televisiva oggetto dell'indagine, il gradimento degli adulti è inferiore al gradimento dei giovani con un grado di fiducia del 99%

STATISTICA E ANALISI DEI DATI

Confronto tra due popolazioni di Poisson

Confronto tra due popolazioni di Poisson

- Consideriamo due popolazioni di Poisson descritte dalle variabili:

$$p_X(x) = \frac{(\lambda_1)^x}{x!} e^{-\lambda_1}, \qquad x = 0, 1, \ldots \qquad (\lambda_1 > 0)$$

$$p_Y(x) = \frac{(\lambda_2)^x}{x!} e^{-\lambda_2}, \qquad x = 0, 1, \ldots \qquad (\lambda_2 > 0)$$

- Siano $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti rispettivamente da due popolazioni di Poisson
- Vogliamo determinare un intervallo di confidenza di grado $1 - \alpha$ per $\lambda_1 - \lambda_2$
- Denotiamo con $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$ le medie campionarie delle popolazioni
  - Dal teorema centrale di convergenza si ha che:

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1/n_1 + \lambda_2/n_2}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard

Confronto tra due popolazioni di Bernoulli

- Inoltre, poiché:

$$E(\overline{X}_{n_1}) = \lambda_1 \quad E(\overline{Y}_{n_2}) = \lambda_2$$

Cioè le medie campionarie $\overline{X_{n_1}}$ e $\overline{Y_{n_2}}$ sono stimatori corretti e consistenti di λ_1 e λ_2, per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1 - \alpha$ possiamo determinare $\lambda_1 - \lambda_2$ con:

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\overline{X}_{n_1}/n_1 + \overline{Y}_{n_2}/n_2}} < z_{\alpha/2}\right) \simeq 1 - \alpha$$

Da cui:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}}{n_1} + \frac{\overline{y}_{n_2}}{n_2}} < \lambda_1 - \lambda_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}}{n_1} + \frac{\overline{y}_{n_2}}{n_2}}$$

Esempio

- Due incroci stradali A e B sono analizzati in base al numero di incidenti per un fissato numero di giorni
- Si registrano il numero di incidenti:
  - nell'incrocio A per 50 giorni distinti
  - nell'incrocio B per 40 giorni distinti
- Supponendo che i numeri di incidenti all'incrocio A e B siano descritti da variabili aleatorie di Poisson con parametri λ_1 e λ_2 rispettivamente
- Si vuole determinare una stima dell'intervallo di confidenza di grado $1-\alpha=0.99$ per $\lambda_1-\lambda_2$

```
> camppoisA<-c(4, 5, 8, 0, 3, 1, 8, 2, 3, 0, 1, 2, 0, 1, 3, 1, 3,
   4, 2, 1,
+  5, 2, 0, 0, 1, 1, 3, 3, 1, 4, 5, 1, 3, 5, 0, 1, 1, 1, 4, 2,
+  6, 3, 1, 0, 2, 5, 1, 5, 1, 4)
>  length(camppoisA)
[1]  50
>
>  camppoisB<-c(1, 5, 2, 3, 2, 0, 3, 3, 0, 2, 5, 8, 1, 3, 1, 4, 2,
   6, 4, 2, 3, 2,
+  7, 5, 1, 3, 3, 4, 1, 4, 3, 3, 3, 2, 0, 2, 3, 7, 2, 1)
>  length(camppoisB)
[1]  40
```

Esempio

- Le frequenze assolute degli incidenti nei due incroci sono:

```
>  table(camppoisA) # frequenze assolute  incidenti in incrocio A
camppoisA
 0  1  2  3  4  5  6  8
 7 15  6  8  5  6  1  2
>  mean(camppoisA)
[1]  2.46
>
 >  table(camppoisB) # frequenze assolute  incidenti incrocio B
camppoisB
 0  1  2  3  4  5  6  7  8
 3  6  9 11  4  3  1  2  1
>  mean(camppoisB)
[1]  2.9
```

- Determiniamo una stima dell'intervallo di confidenza di grado $1-\alpha=0.99$ per $\lambda_1-\lambda_2$

```
>  alpha<-1-0.99
>  qnorm(1-alpha/2,mean=0,sd=1)
[1]  2.575829
>  n1<- length(camppoisA)
>  n2<- length(camppoisB)
>  m1<-mean(camppoisA)
>  m2<-mean(camppoisB)
>  rad<-sqrt(m1/n1+m2/n2)
>  m1-m2- qnorm(1-alpha/2,mean=0,sd=1)*rad
[1]  -1.338592
>  m1-m2+ qnorm(1-alpha/2,mean=0,sd=1)*rad
[1]  0.4585916
```

Una stima dell'intervallo di confidenza di grado per $\lambda_1-\lambda_2$ è (−1.3386, 0.4586)