6. Intervalli di confidenza

La media campionaria e la varianza campionaria ci offrono una stima dei parametri media e varianza del campione statistico in esame. Abbiamo però bisogno di sapere *quanto ci si possa fidare di questa stima* ovvero quale sia la probabilità che il *vero* valore del parametro incognito non sia *troppo distante* dalla stima trovata.

Diamo perció la seguente definizione:

Definizione 6.0.1 (Intervallo di confidenza)**.** Sia $X_1, X_2, \ldots, X_n$ un campione statistico e sia θ un parametro (ignoto) che caratterizza la distribuzione del campione.
Siano $L_i = l_i(X_1, X_2, \ldots, X_n)$ e $L_s = l_s(X_1, X_2, \ldots, X_n)$ due statistiche del campione e sia $\alpha \in (0,1)$. Dico che l'intervallo (L_i, L_s) è un *intervallo di confidenza* (o di fiducia) di livello $1-\alpha$ se $\mathbb{P}(\theta \in (L_i, L_s)) \geq 1-\alpha$, ovvero che (L_i, L_s) è un intervallo di confidenza (o di fiducia) di errore α se $\mathbb{P}(\theta \notin (L_i, L_s)) \leq \alpha$.

Dico che la semiretta $(L_i, +\infty)$ è un *intervallo di confidenza unilaterale superiore* di livello $1-\alpha$ se $\mathbb{P}(\theta > L_i) \geq 1-\alpha$

Dico che la semiretta $(-\infty, L_s)$ è un *intervallo di confidenza unilaterale inferiore* di livello $1-\alpha$ se $\mathbb{P}(\theta < L_s) \geq 1-\alpha$

Osservazione 6.0.3.

1. La scelta dei nomi delle due statistiche non è casuale: L_i sta per limitazione inferiore mentre L_s sta per limitazione superiore.
2. Di solito si è interessati a *piccoli* valori di α, più precisamente a $\alpha \in (10^{-2}, 10^{-1})$.
3. La disuguaglianza di Chebyshev ci ha fornito un intervallo di confidenza per la media μ del campione nel caso in cui la varianza σ^2 sia nota

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| \geq t\right) \leq \frac{\sigma^2}{t^2} \qquad \forall t > 0$$

ovvero

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| < t\right) \geq 1 - \frac{\sigma^2}{t^2} \qquad \forall t > 0$$

cioè

$$\mathbb{P}\left(\overline{X} - t < \mu < \overline{X} + t\right) \geq 1 - \frac{\sigma^2}{t^2} \qquad \forall t > 0.$$

Fissato $\alpha \in (0,1)$ scelgo $t = \dfrac{\sigma}{\sqrt{\alpha}}$. La disuguaglianza di Chebyshev si legge allora

$$\mathbb{P}\left(\overline{X} - \frac{\sigma}{\sqrt{\alpha}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{\alpha}}\right) \geq 1 - \alpha \qquad \forall \alpha \in (0,1).$$

Dunque l'intervallo $\left(\overline{X} - \frac{\sigma}{\sqrt{\alpha}}, \overline{X} + \frac{\sigma}{\sqrt{\alpha}}\right)$ è un intervallo di confidenza di livello $1-\alpha$ per la media μ del campione.

6.1. Stima per intervalli della media di campioni gaussiani

6.1.1. Campione gaussiano di cui è nota la varianza

Intervallo bilaterale

Sia $X_1, X_2, \ldots, X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota.

Sia Z una v.a. gaussiana standard e sia $\alpha \in (0,1)$. Calcolo $\mathbb{P}\left(|Z| \leq z_{1-\frac{\alpha}{2}}\right)$:

$$\begin{aligned}
&\mathbb{P}\left(|Z| \leq z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(Z \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \leq -z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(Z \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \leq z_{\frac{\alpha}{2}}\right) \qquad (6.1)\\
&= \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.
\end{aligned}$$

Sappiamo che $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ e che dunque $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. Applichiamo quindi la disuguaglianza (6.1) a $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$. Si ha:

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(\frac{\left|\overline{X}-\mu\right|}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\mu-\overline{X}}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(\frac{-\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu - \overline{X} \leq \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) \qquad (6.2)\\
&= \mathbb{P}\left(\overline{X} - \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)
\end{aligned}$$

L'intervallo

$$\left(\overline{X} - \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{X} + \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per la media μ del campione.

Osservazione 6.1.1 (Dimensionamento del campione)**.** Fissato il livello di confidenza $1-\alpha$, supponiamo di voler controllare l'ampiezza dell'intervallo di confidenza $L_s - L_i$. Nel caso in esame l'ampiezza dell'intervallo di confidenza è $\frac{2\,\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. Se fissiamo una limitazione superiore $2\,\delta$ per l'ampiezza di tale intervallo, deve dunque essere

$$\frac{2\,\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq 2\,\delta$$

ovvero

$$n \geq \left(\frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\delta}\right)^2.$$

Intervallo unilaterale superiore

Sia $Z \sim \mathcal{N}(0,1)$. Sappiamo che

$$\mathbb{P}(Z \leq t) = 1 - \alpha \quad \text{se e solo se} \quad t = z_{1-\alpha}.$$

Abbiamo dunque

$$1 - \alpha = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha}\right) = \mathbb{P}\left(\overline{X} - \mu \leq \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \geq \overline{X} - \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(\overline{X} - \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}, +\infty\right)$$

è un intervallo di confidenza unilaterale superiore di livello $1 - \alpha$.

Intervallo unilaterale inferiore

Sia $Z \sim \mathcal{N}(0,1)$. Sappiamo che

$$\mathbb{P}(Z \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(Z \leq t) = \alpha \quad \text{se e solo se} \quad t = z_{\alpha}.$$

Abbiamo dunque

$$1 - \alpha = \mathbb{P}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq z_{\alpha}\right) = \mathbb{P}\left(\overline{X} - \mu \geq \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \leq \overline{X} - \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(-\infty, \overline{X} - \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right) = \left(-\infty, \overline{X} + \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right)$$

è un intervallo di confidenza unilaterale inferiore di livello $1 - \alpha$.

6.1.2. Campione gaussiano di cui non è nota la varianza

Intervallo bilaterale

Sia $X_1, X_2, \ldots, X_n$ un campione gaussiano di media μ varianza σ^2, entrambe incognite.

Sappiamo che la v.a. $T := \dfrac{(\overline{X} - \mu)\sqrt{n}}{S}$ segue la distribuzione t di Student con $n - 1$ gradi di libertà:

$$T \sim t(n-1).$$

Sia $t_{n-1,1-\frac{\alpha}{2}}$ il relativo quantile di livello $1 - \frac{\alpha}{2}$:

$$\mathbb{P}\left(T \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Calcolo $\mathbb{P}\left(|T| \leq t_{n-1,1-\frac{\alpha}{2}}\right)$:

$$\begin{aligned}\mathbb{P}\left(|T| \leq t_{n-1,1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} \leq T \leq t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(T \leq t_{n-1,1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(T \leq -t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(T \leq t_{n-1,1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(T \leq t_{n-1,\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

Abbiamo dunque

$$\begin{aligned}1-\alpha = \mathbb{P}\left(|T| \leq t_{n-1,1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(\frac{\left|\overline{X}-\mu\right|\sqrt{n}}{S} \leq t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\left|\overline{X}-\mu\right| \leq \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\frac{-S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \leq \mu - \overline{X} \leq \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\end{aligned}$$

L'intervallo

$$\left(\overline{X} - \frac{S\, t_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{X} + \frac{S\, t_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per la media μ del campione.

Intervallo unilaterale superiore

Sappiamo che

$$\mathbb{P}\left(T \leq t\right) = 1-\alpha \quad \text{ se e solo se } \quad t = t_{n-1,1-\alpha}.$$

Abbiamo dunque

$$\begin{aligned}1-\alpha = \mathbb{P}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S} \leq t_{n-1,1-\alpha}\right) &= \mathbb{P}\left(\overline{X} - \mu \leq \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(\mu \geq \overline{X} - \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right).\end{aligned}$$

Quindi la semiretta

$$\left(\overline{X} - \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}, +\infty\right)$$

è un intervallo di confidenza unilaterale superiore di livello $1-\alpha$.

Intervallo unilaterale inferiore

Sappiamo che

$$\mathbb{P}(T \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(T \leq t) = \alpha \quad \text{se e solo se} \quad t = t_{n-1,\alpha}.$$

Abbiamo dunque

$$1 - \alpha = \mathbb{P}\left(\frac{(\overline{X} - \mu)\sqrt{n}}{S} \geq t_{n-1,\alpha}\right) = \mathbb{P}\left(\overline{X} - \mu \geq \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \leq \overline{X} - \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(-\infty, \overline{X} - \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right) = \left(-\infty, \overline{X} + \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right)$$

è un intervallo di confidenza unilaterale inferiore di livello $1 - \alpha$.

6.2. Stima per intervalli della varianza di campioni gaussiani

Intervallo bilaterale

Sia $X_1, X_2, \ldots, X_n$ un campione gaussiano di media μ (incognita o nota)e varianza σ^2 incognita.

Sappiamo che la v.a. $V := (n-1)\dfrac{S^2}{\sigma^2}$ segue la distribuzione χ^2 a $n-1$ gradi di libertà. Per ogni $\alpha \in (0,1)$ indico con $\chi^2_{n-1,\alpha}$ il quantile di livello α della v.a. V:

$$F_V\left(\chi^2_{n-1,\alpha}\right) = \alpha \qquad \forall \alpha \in (0,1).$$

Osservazione 6.2.1. $\chi^2_{n-1,\alpha} > 0$ per ogni $\alpha \in (0,1)$.

Calcolo $\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$:

$$\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) - \\ - \mathbb{P}\left(V < \chi^2_{n-1,\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.$$

Dunque

$$1 - \alpha = \mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$$
$$= \mathbb{P}\left(\frac{1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = \mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

Quindi l'intervallo

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

è un intervallo di confidenza di livello $1 - \alpha$ per la varianza σ^2 del campione.

Intervallo unilaterale superiore

Sappiamo che

$$\mathbb{P}(V \leq t) = 1 - \alpha \quad \text{se e solo se} \quad t = \chi^2_{n-1,1-\alpha}.$$

Dunque

$$1 - \alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha}\right) = \mathbb{P}\left(\sigma^2 > (n-1)\frac{S^2}{\chi^2_{n-1,1-\alpha}}\right).$$

Quindi la semiretta

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}, +\infty\right)$$

è un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 del campione.

Intervallo unilaterale inferiore

Sappiamo che

$$\mathbb{P}(V \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(V \leq t) = \alpha \quad \text{se e solo se} \quad t = \chi^2_{n-1,\alpha}.$$

Dunque

$$1 - \alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} > \chi^2_{n-1,\alpha}\right) = \mathbb{P}\left(\sigma^2 \leq (n-1)\frac{S^2}{\chi^2_{n-1,\alpha}}\right).$$

Quindi l'intervallo

$$\left(0, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}\right)$$

è un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 del campione.

Esempio 6.2.1. Calcoliamo gli intervalli di confidenza per il carattere CO2SBW dei dati tratti da [3], nell'ipotesi che si tratti della realizzazione di v.a. normali.

```
> setwd("/home/laura/Documents/didattica/2012-13_elaborazioni_B194")

> X <-
+   read.table("/home/laura/Documents/didattica/2012-13_elaborazioni_B194/
table2_noR5.csv",
+    header=TRUE, sep="\t", na.strings="NA", dec=".", strip.white=TRUE)

> X
   Totpor   PRA   PV Densi TenStr CO2SBW FirTemp
1   41.46 0.528 80.0  1.55  0.403   0.38     740
2   47.21 0.467 81.2  1.65  0.645   0.70     740
3   43.67 0.697 78.5  1.71  0.527   0.46     740
4   52.39 0.422 77.3  1.52  0.143   0.48     740
5   44.70 0.411 87.4  1.50  0.593   0.29     740
```

```
6   51.33 0.422 88.6  1.48  0.463   0.33     740
7   31.46 0.718 80.6  1.90  0.955   0.23     740
8   40.90 0.458 80.4  1.68  0.195   0.41     740
9   45.54 0.492 80.8  1.62  1.328   0.50     750
10  45.62 0.734 86.2  1.62  1.405   0.34     750
11  44.14 0.730 85.7  1.59  0.256   0.42     750
12  40.71 0.543 87.8  1.75  0.309   0.20     750
13  35.70 0.686 84.3  1.52  0.472   0.05     740
14  40.29 0.306 43.5  1.76  0.520   0.43     740
15  36.57 0.625 42.3  1.75  0.738   0.36     740
16  42.13 0.249 63.2  1.63  0.410   0.25     740
17  37.83 0.731 47.9  2.02  0.601   0.28     740
18  42.18 0.407 59.4  1.58  0.376   0.34     740
19  41.60 0.446 42.8  1.85  0.473   0.26     740
20  32.66 0.664 64.3  1.85  0.695   0.25     740
21  36.07 0.673 58.2  1.78  0.624   0.29     740
22  36.04 1.397 55.6  1.73  0.582   0.38     740
23  36.64 0.861 45.2  1.75  0.650   0.47     740
24  42.89 0.785 10.2  1.54  0.453   1.04     850
25  26.85 0.315 14.7  2.01  1.124   1.86     960
26  28.55 0.158 18.6  1.92  0.937   1.96     850
27  29.86 0.158 15.3  1.89  1.020   1.48     850
28  54.64 1.525 12.5  1.34  0.267   0.67     750
29  27.55 2.657 14.6  1.92  0.892   0.40     730
30  40.82 0.622 15.3  1.57  0.502   1.94     860

> numSummary(X[,c("CO2SBW", "Densi", "FirTemp", "PRA", "PV", "TenStr", "Totpor")],
statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
               mean         sd      0%       25%      50%       75%    100%  n
CO2SBW    0.5816667  0.5259152   0.050   0.29000   0.3900   0.49500   1.960 30
Densi     1.6993333  0.1691548   1.340   1.57250   1.6950   1.83250   2.020 30
FirTemp 763.6666667 53.4649955 730.000 740.00000 740.0000 750.00000 960.000 30
PRA       0.6629000  0.4806106   0.158   0.42200   0.5825   0.72700   2.657 30
PV       56.7466667 27.9061201  10.200  42.42500  61.3000  80.75000  88.600 30
TenStr    0.6186000  0.3153048   0.143   0.42075   0.5545   0.72725   1.405 30
Totpor   39.9333333  7.0795326  26.850  36.04750  40.8600  44.02250  54.640 30

> ## definisco la funzione che calcola l'intervallo bilaterale con varianza nota

> bilat.norm = function(x,sigma,conf) { n = length(x); xbar=mean(x);
+ alpha = 1 - conf;
+ zstar = qnorm(1-alpha/2);
+ SE = sigma/sqrt(n);
+ xbar + c(-zstar*SE,zstar*SE)
+ }

> bilat.norm(X[,c("CO2SBW")],1,.9) ## supponiamo deviazione standard = 1
[1] 0.2813589 0.8819745
```

```
> bilat.norm(X[,c("CO2SBW")],2,.9) ## supponiamo deviazione standard = 2
[1] -0.01894896  1.18228229

> bilat.norm(X[,c("CO2SBW")],1,.95) ## supponiamo deviazione standard = 1
[1] 0.2238278 0.9395055

> bilat.norm(X[,c("CO2SBW")],2,.95) ## supponiamo deviazione standard = 2
[1] -0.134011  1.297344

> ## definisco la funzione che calcola l'intervallo bilaterale con varianza ignota

> bilat.stud = function(x,conf) { n = length(x); m = n-1; xbar=mean(x);
+ alpha = 1 - conf;
+ zstar = qt(1-alpha/2, m, lower.tail=TRUE);
+ SE = sd(x)/sqrt(n);
+ xbar + c(-zstar*SE,zstar*SE)
+ }

> bilat.stud(X[,c("CO2SBW")],.9)
[1] 0.4185190 0.7448144

> bilat.stud(X[,c("CO2SBW")],.95)
[1] 0.3852867 0.7780466

> ## definisco la funzione che calcola l'intervallo bilaterale per la varianza

> bilat.chi = function(x,conf) { n = length(x); m = n-1;
+ alpha = 1 - conf;
+ zsup = qchisq(alpha/2, m, lower.tail=TRUE);
+ zinf = qchisq(1 - alpha/2, m, lower.tail=TRUE);
+ SE = sd(x)*sd(x)*m;
+ c(SE/zinf,SE/zsup)
+ }

> bilat.chi(X[,c("CO2SBW")],.9)
[1] 0.1884772 0.4529507

> bilat.chi(X[,c("CO2SBW")],.95)
[1] 0.175429 0.499843
```