Statistica inferenziale – Esercizi #2

Dario Comanducci, 21 febbraio 2024

Esercizio 1

Data un campione $\mathbf{X} = [X_1 \ \ldots \ X_n]^\top$ di $n = 100$ osservazioni i.i.d $X_i \sim N(\mu, \sigma^2)$ (con $i = 1 \ldots n$, $\mu = 170$, $\sigma = 20$), possiamo analizzare il comportamento delle statistiche

$$\overline{X} = \frac{\sum_i X_i}{n}$$
$$S^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1}$$

attraverso le simulazioni riportate nel listato 1 in Appendice.

Analisi di $\overline{X}$

Sapendo che $\overline{X} \sim N(\mu, \sigma^2/n)$, si verifica che su $N = 10^4$ istanze del campione aleatorio $\mathbf{X}$ otteniamo a partire dai valori $\overline{X}_k$ ($k = 1 \ldots N$) che:

$$\hat{\mu}_{\overline{X}} = \frac{\sum_{k=1}^N \overline{X}_k}{N} \approx 169.986 \approx \mu = 170$$
$$\hat{\sigma}^2_{\overline{X}} = \frac{\sum_{k=1}^N (\overline{X}_k - \bar{\mu})^2}{N-1} \approx 3.982 \approx \frac{\sigma^2}{n} = 4$$

Fig. 1a illustra la distribuzione campionaria osservata sugli N campioni $\mathbf{X}$, mostrando come l'istogramma sia sovrapponibile alla curva teorica attesa.

Analisi di S^2

Per analizzare la distribuzione campionaria di S^2, partiamo dalle seguenti considerazioni:

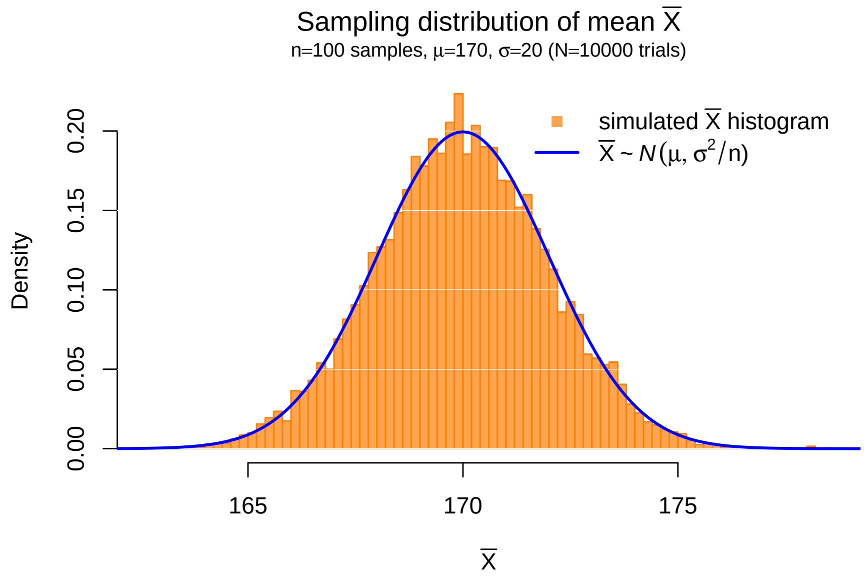
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \overline{X})^2}{n} \sim \frac{\sigma^2}{n} \chi^2_{n-1}$$
$$S^2 = \frac{n}{n} \frac{\sum_{=1}^n (x_i - \overline{X})^2}{n-1} = \frac{n}{n-1} \hat{\sigma}^2$$

da cui ricaviamo che

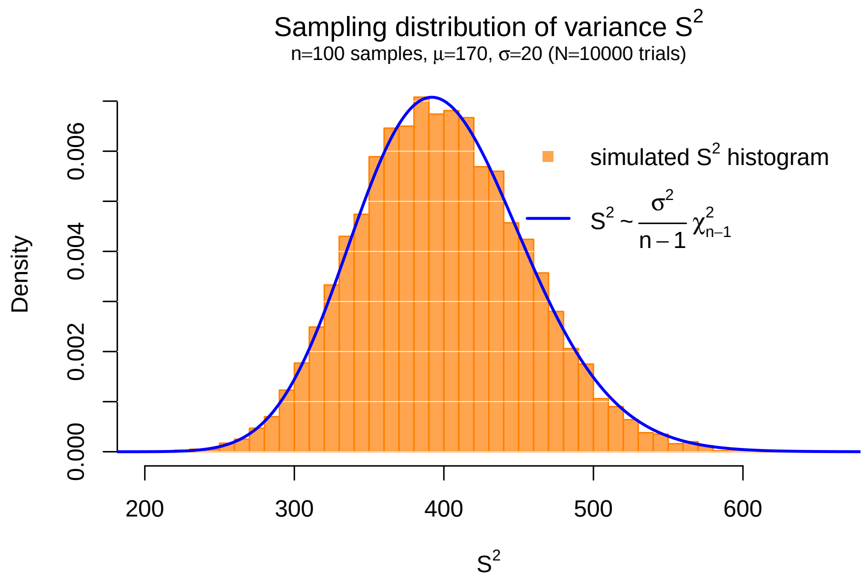
$$\hat{\sigma}^2 = \frac{n-1}{n} S^2 \sim \frac{\sigma^2}{n} \chi^2_{n-1} \quad \Rightarrow \quad Y = \frac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$$

Il cambio di variabile in Y è necessario per poter tracciare in Fig. 1b, attraverso la funzione `dchisq(...)` di R, l'andamento della distribuzione teorica prevista tenendo anche conto del fatto che [1, p. 34]

$$X \sim f_X(x) \Rightarrow Y = aX \sim f_Y(y) = 1/|a| \, f_X(y/a)$$



(a)



(b)

Fig. 1: Istogrammi della simulazione per l'esercizio 1 a confronto con le distribuzioni teoriche previste.

Notiamo inoltre che la media per le N istanze di S^2 vale

$$\overline{S^2} = \frac{\sum_{k=1}^{N} S_k^2}{N} \approx 399.865 \approx \sigma^2 = 400$$

Il risultato è in accordo alla distribuzione teorica in quanto $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ e $\mathbb{E}[\chi^2_{n-1}] = n - 1$.

Esercizio 2

Sia $\mathbf{X} = [X_1 \dots X_n]^\top$ un campione di $n = 5$ elementi tratti da una distribuzione esponenziale $f(x; \lambda) = \lambda e^{-\lambda x}$ $(x > 0)$ con $\lambda = 3$. Analizziamo

1. la distribuzione campionaria della media
2. la distribuzione campionaria della stima MLE per λ
3. il bias^2 e la varianza dello stimatore MLE

Lo script R è allegato come listato 2 in appendice.

2.1

La media per una distribuzione esponenziale vale $\mu = \lambda^{-1}$, e la media campionaria $\hat{\mu}$ si comporta come

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \sim \Gamma\left(n, \frac{1}{n\lambda}\right)$$

come mostrato in Fig. 2a, dove si osserva la concordanza tra i risultati della simulazione e la curva teorica prevista.

2.2

La stima MLE per il tasso di fallimento vale $\hat{\lambda} = n / \sum_{i=1}^{n} x_i$. Fig. 2b riporta l'istogramma dei suoi valori assunti su $N = 10^4$ esperimenti indipendenti. Per valutare se la simulazione rispetta la distribuzione teorica, è necessario determinarne l'espressione per la v.a. $\hat{\lambda} = \hat{\mu}^{-1}$ considerando che:

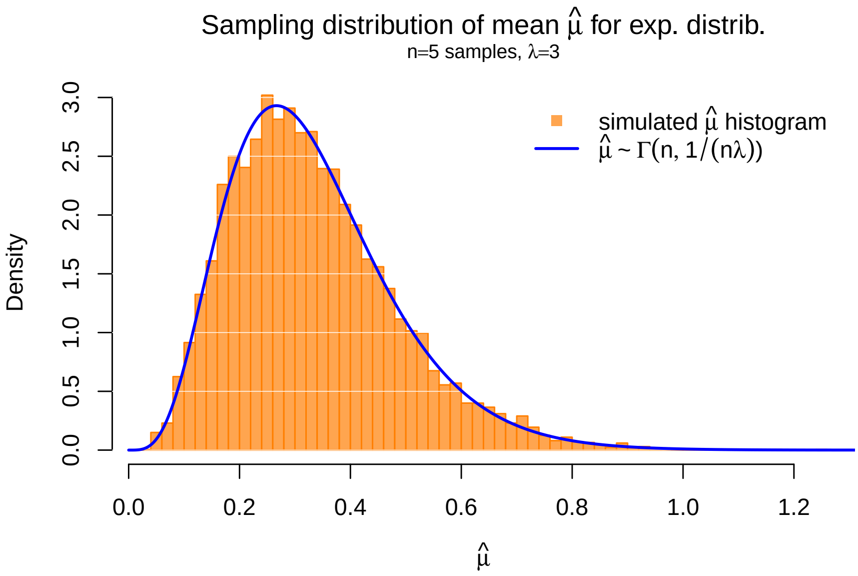
- $\hat{\mu} \sim f_{\hat{\mu}} = \Gamma(x, \alpha, \beta)$ con $\alpha = n$ (parametro di forma) e $\beta = 1/(n\lambda) = \mu/n$ (parametro di scala)
- se una v.a. $X \sim f_X(x)$, allora $Y = 1/X \sim f_Y(y) = 1/y^2\, f_X(1/y)$ [1, p. 36].

Pertanto

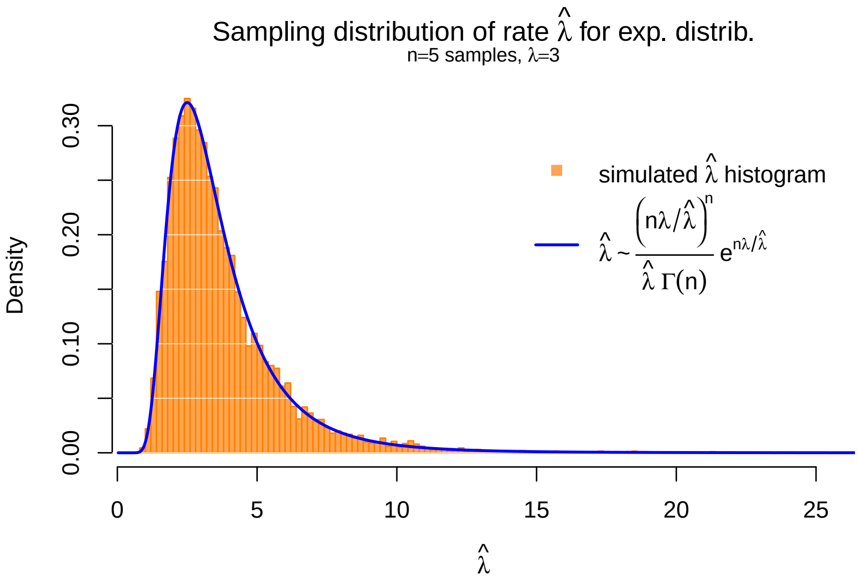
$$\hat{\lambda} \sim f_{\hat{\lambda}}(y) = 1/y^2\, f_{\hat{\mu}}(1/y) = 1/y^2\, \Gamma(1/y, \alpha, \beta)\big|_{\alpha=n, \beta=\mu/n}$$

Esplicitando l'equazione per $f_{\hat{\mu}}$:

$$f_{\hat{\mu}}(x) = \Gamma(\alpha, \beta)\big|_{\alpha=n,\, \beta=\mu/n} = \frac{1}{\beta\Gamma(\alpha)} \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} = \frac{n}{\mu\Gamma(n)} \left(\frac{xn}{\mu}\right)^{n-1} e^{-nx/\mu}$$



(a)



(b)

Fig. 2: Istogrammi della simulazione per l'esercizio 2.

Ne segue che

$$\hat{\lambda} = 1/\hat{\mu} \sim f_{\hat{\lambda}}(y) = 1/y^2\, f_{\hat{\mu}}(1/y) = \frac{n\lambda}{y^2\Gamma(n)} \left(\frac{n\lambda}{y}\right)^{n-1} e^{-n\lambda/y}$$
$$= \frac{1}{y\Gamma(n)} \left(\frac{n\lambda}{y}\right)^{n} e^{-n\lambda/y}$$

L'andamento di $f_{\hat{\lambda}}(\lambda)$ è riportato in Fig. 2b, sovrapposto all'istogramma ottenuto dalla simulazione, mostrando un buon accordo con i risultati ottenuti.[1]

2.3

La stima MLE per il tasso λ presenta un bias:

$$\mathbb{E}_\lambda[\text{MLE}] = \frac{n\lambda}{n-1} \Rightarrow \text{bias} = (\mathbb{E}_\lambda[\text{MLE}] - \lambda) = \frac{\lambda}{n-1} = \frac{3}{4} \quad (\text{per } \lambda = 3,\, n = 5)$$

La media $\overline{\lambda}$ su $N = 10^4$ istanze di $\mathbf{X}$ per la stima MLE $\hat{\lambda}$ è in accordo a $\mathbb{E}_\lambda[\text{MLE}]$ in quanto vale

$$\overline{\lambda} = \frac{\sum_{k=1}^N \hat{\lambda}_k}{N} \approx 3.766 \approx \mathbb{E}_\lambda[\text{MLE}] = \frac{15}{4} = 3.75$$

essendo $\hat{\lambda}_k$ la stima MLE di λ sul k-esimo trial, di $n = 5$ campioni, della simulazione ($k = 1 \dots N$); ne segue che

$$\widehat{\text{bias}} = \overline{\lambda} - \lambda \approx 0.766 \approx \frac{3}{4} \qquad (\text{inoltre } \widehat{\text{bias}}^2 \approx 0.586)$$

La stima della varianza dello stimatore MLE per λ può essere ottenuta calcolando il valore di $S^2_{\hat{\lambda}}$ sugli N esperimenti:

$$S^2_{\hat{\lambda}} = \frac{\sum_{k=1}^N \left(\hat{\lambda}_k - \overline{\lambda}\right)^2}{N-1} \approx 4.682$$

Combinando $\widehat{\text{bias}}$ e $S^2_{\hat{\lambda}}$ possiamo ottenere un'idea dell'errore quadratico medio MSE su λ commesso con lo stimatore MLE:

$$\text{MSE}_\lambda[\text{MLE}] \approx \widehat{\text{bias}}^2 + S^2_{\hat{\lambda}} \approx 5.2683$$

e tale valore è in linea con la media

$$\overline{\text{MSE}}_\lambda = \frac{\sum_{k=1}^N \left(\hat{\lambda}_k - \lambda\right)^2}{N} \approx 5.2678$$

Esercizio 3

Data X con funzione di distribuzione $F_X(x)$, il quantile di ordine $p \in [0,1]$ è quella determinazione x_p t.c. $F_X(x_p) = \mathbb{P}[X \leqslant x_p] = p$. Tab. 1 riporta i quantili richiesti e le istruzioni `R` impiegate per calcolarli.

[1] Nella creazione del grafico di Fig. 2b per semplicità è stato comunque usata la forma "implicita" $f_{\hat{\lambda}}(y) = 1/y^2\ \Gamma(1/y, \alpha, \beta)|_{\alpha=n, \beta=\mu/n}$.

Tab. 1: Calcolo dei quantili per l'esercizio 3.

quantile	istruzione R	valore
$z_{0.9}$	`qnorm(p=0.9)`	1.2816
$t_{0.9}(3)$	`qt(p=0.9, df=3)`	1.6377
$t_{0.9}(10)$	`qt(p=0.9, df=10)`	1.3721
$-t_{0.1}(10)$	`−qt(p=0.1, df=10)`	1.3721
$-t_{0.1}(3)$	`−qt(p=0.1, df=3)`	1.6377

Riferimenti bibliografici

[1] G. Iuculano. *Introduzione alla probabilità e statistica nell'ingegneria e nelle scienze fisiche*. Pitagora Editrice Bologna, 1991.

A Listati R

Listato 1: Codice R per l'esercizio 1

```
rm(list=ls())

library(latex2exp)

n = 100
mu = 170
sigma = 20
N = 10000 # number of samples

# Check that the sampling distribution of the mean is normal
# Check using simulation that the mean and the standard deviation of
# the sampling distribution of the mean correspond to the theoretical
    values
barX = rep (0, length = N)
S2 = rep (0, length = N)
for(k in 1:N)
{
  x = rnorm(n, mean = mu, sd = sigma)
  barX[k] = mean(x)
  S2[k] = var(x)
}


# \hat{\mu}_{\bar{X}}: estimated mean of the sampling distrib. for \mu
hatmu_barX = mean(barX)
# \hat{\sigma}^2_{\bar{X}}: estim. mean of the sampling distrib. for \mu
hatsigma2_barX = var(barX)
# \bar{S^2}: average value for S^2
barS2 = mean(S2)


hist(barX, breaks="Freedman-Diaconis", probability = TRUE,
```

```
    border = "darkorange1", col="tan1",
    main=TeX(r"(Sampling distribution of mean $\bar{X}$)"),
    xlab = TeX(r"($\bar{X}$)"))
title(TeX(r"(\small{$n=100$ samples, $\mu=170$, $\sigma=20$ ($N=10000$
    trials)})"),
     line=1)
grid(lty="solid", lwd=0.5, col="white", nx=NA, ny=NULL)
curve(dnorm(x,mean=mu,sd=sigma/sqrt(n)),
  from=100,to=240, n = 2001,
  ylab="density", lwd=2,col="blue", add=T)
legend("topright", bty = "n",
  legend = c(TeX(r"(simulated $\bar{X}$ histogram)"),
            TeX(r"($\bar{X} \sim \textit{N}(\mu,\sigma^2/n$))")),
  col = c("tan1", "blue"), pch = c(15, NA), lty = c(NA, 1), lwd=c(NA,2),
  inset = 0.03
)

# Study using simulation the sampling distribution of the variance S^2
# solution: Y=(n-1)*S^2/\sigma^2 ~ \chi^2_{n-1}
hist(S2, breaks="Freedman-Diaconis", probability = TRUE,
    border = "darkorange1", col="tan1",
    main=TeX(r"(Sampling distribution of variance $S^2$)"),
    xlab = TeX(r"($S^2$)"))
grid(lty="solid", lwd=0.5, col="white", nx=NA, ny=NULL)
title(TeX(r"(\small{$n=100$ samples, $\mu=170$, $\sigma=20$ ($N=10000$
    trials)})"),
     line=1)
curve((n-1)/sigma^2*dchisq(x*(n-1)/sigma^2, df=n-1),
  from=100,to=700, n = 2001,
  ylab="density", lwd=2,col="blue", add=T)
legend("topright", bty = "n",
      legend = c(TeX(r"(simulated $S^2$ histogram)"),
                TeX(r"($S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$)")),
      col = c("tan1", "blue"), pch = c(15, NA), lty = c(NA, 1),
          lwd=c(NA,2),
      inset = 0.03,
)
```

Listato 2: Codice R per l'esercizio 2

```
rm(list=ls())

library(latex2exp)

n = 5
lambda = 3
N = 10000 # number of samples

barX = rep (0, length = N)
lambdaMLE = rep (0, length = N)
for(k in 1:N)
{
  x = rexp(n, lambda)
```

```
  barX[k] = mean(x)
  lambdaMLE[k] = n/sum(x)
}

barLambdaMLE = mean(lambdaMLE)
bias = barLambdaMLE - lambda
S2MLE = var(lambdaMLE)
estMSE = bias^2 + S2MLE
barMSE = mean((lambdaMLE-lambda)^2)

# pdf for \hat{\lambda}
lambdaPdf <- function(x,lambda,n)
{
  v = n*lambda/x
  pdf = 1/(x*gamma(n)) *v^n * exp(-v)
  return (pdf)
}

hist(barX, breaks="Freedman-Diaconis", probability = TRUE,
     border = "darkorange1", col="tan1",
     main=TeX(r"(Sampling distribution of mean $\hat{\mu}$ for exp.
         distrib.)"),
     xlab = TeX(r"($\hat{\mu}$)"))
title(TeX(r"(\small{$n=5$ samples, $\lambda=3$})"), line=1)
grid(lty="solid", lwd=0.5, col="white", nx=NA, ny=NULL)
curve(dgamma(x, shape=n, scale=1/(n*lambda)),
      from=0,to=2, n = 2001,
      ylab="density", lwd=2,col="blue", add=T)
legend("topright", bty = "n",
       legend = c(TeX(r"(simulated $\hat{\mu}$ histogram)"),
                  TeX(r"($\hat{\mu} \sim \Gamma(n,1/(n\lambda)$))")),
       col = c("tan1", "blue"), pch = c(15, NA),
      lty = c(NA, 1), lwd=c(NA,2), inset = 0.03)

hist(lambdaMLE, breaks="Freedman-Diaconis", probability = TRUE,
     border = "darkorange1", col="tan1",
     main=TeX(r"(Sampling distribution of rate $\hat{\lambda}$ for exp.
         distrib.)"),
     xlab = TeX(r"($\hat{\lambda}$)"))
title(TeX(r"(\small{$n=5$ samples, $\lambda=3$})"), line=1)
grid(lty="solid", lwd=0.5, col="white", nx=NA, ny=NULL)
curve(1/x^2*dgamma(1/x, shape=n, scale=1/(n*lambda)),
# curve(lambdaPdf(x, lambda=3,n=n),
      from=0,to=60, n = 2001,
      ylab="density", lwd=2,col="blue", add=T)
legend("topright", bty = "n",
       legend = c(TeX(r"(simulated $\hat{\lambda}$ histogram)"),
                  TeX(r"($\hat{\lambda} \sim
                      \frac{\left(n\lambda/\hat{\lambda}\right)^n}
                      {\hat{\lambda}\, \Gamma(n)}
                      e^{n\lambda/\hat{\lambda}} $)")),
       col = c("tan1", "blue"), pch = c(15, NA),
      lty = c(NA, 1), lwd=c(NA,2), inset = 0.03)
```