Statistica inferenziale – Esercizi #3

Dario Comanducci, 26 aprile 2024

Esercizio 1

Avendo ottenuto $s = 2$ successi con distribuzione di Bernoulli su un campione casuale composto da $n = 100$ tentativi abbiamo che la stima per la probabilità p dei successi vale

$$\hat{p} = \frac{s}{n} = \frac{2}{100} = 0.02$$

1.1

Per n grande ($n \geqslant 100$) possiamo approssimare la distribuzione di $\hat{p}$ secondo

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1)$$

da cui ricaviamo che l'intervallo di confidenza I_{N} al 95% per p vale

$$\left|\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}}\right| \leqslant 1.96 \iff |p - \hat{p}| \leqslant 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \approx 0.027$$

ossia $I_{\mathrm{N}} \approx [-0.007,\, 0.047]$. Notiamo che in questo caso l'estremo inferiore di I_{N} è negativo, cioè si estende al di fuori dell'intervallo teorico per $p \in (0,1)$.

1.2

Per determinare gli intervalli di confidenza basati sulla verosimiglianza, conviene valutare la log-verosimiglianza relativa $r(p)$ rispetto a $\hat{p} = 0.2$ (curva blu in Fig. 1):

$$r(p) = s \ln \frac{p}{\hat{p}} + (n - s) \ln \frac{1-p}{1-\hat{p}}$$

In particolare, occorre determinare

$$I_{\mathrm{L}} = \{\, p : r(p) \geqslant \underbrace{-(1/2)\,\chi^2_{1,0.95}}_{\approx -1.92} \,\}$$

Nel caso in esame un possibile modo per determinare gli estremi di I_{L} è il metodo della bisezione[1], ottenendo

$$I_{\mathrm{L}} \approx [0.003, 0.060]$$

Notiamo che in questo caso l'intervallo è "ben definito", in quanto contiene solo valori ammissibili per p; inoltre non è simmetrico rispetto a $\hat{p}$.

[1] A esempio, tamite la lunzione `bisect` del package `pracma` in R.

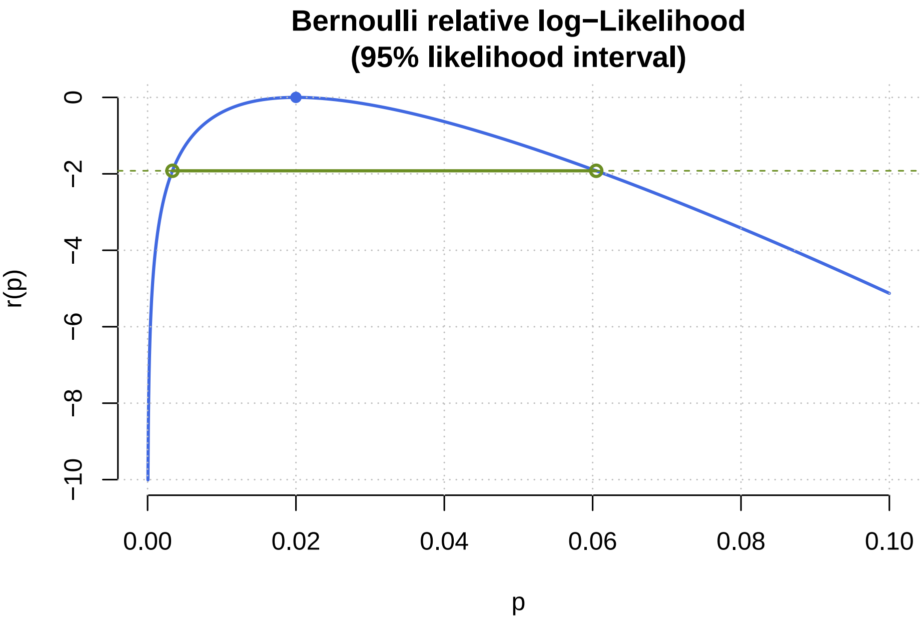


Fig. 1: Log-verosimiglianza relativa $r(p)$ (in blu) per l'esercizio 1, limitando il grafico in un intorno di $\hat{p} = 0.02$. L'intersezione tra $r(p)$ e la linea corrispondente alla soglia $-\chi^2_1/2$ fornisce gli estremi per l'intervallo di confidenza al 95%.

Confronto tra i due intervalli di confidenza

Gli intervalli ottenuti con i due metodi sono piuttosto diversi:

$$I_\mathrm{N} \approx [-0.007,\, 0.047]$$
$$I_\mathrm{L} \approx [\ \ 0.003,\, 0.060]$$

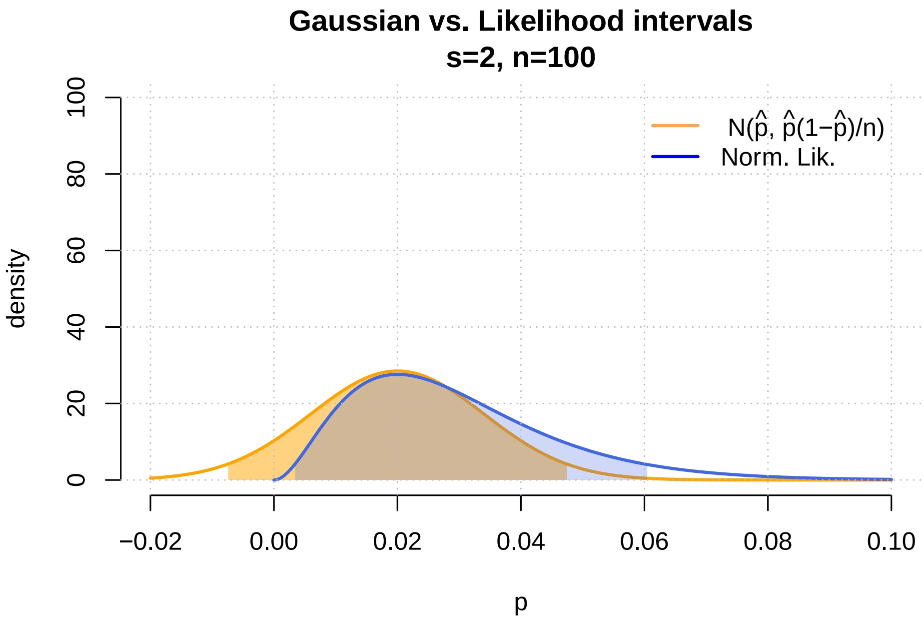
Tuttavia da Fig. 2 osserviamo che, aumentando n da 100 a 1000 (e proporzionalmente anche s), i due intervalli tendono a sovrapporsi.

L'impiego della densità Gaussiana è un'approssimazione della reale distribuzione che assume $\hat{p}$ ($\hat{p} \sim 1/n\,\mathrm{Bin}(n, p)$). Assumendo che il vero valore di p coincida con $\hat{p}$, possiamo vedere da Fig. 3a che l'approssimazione Gaussiana è in questo caso ancora piuttosto grossolana rispetto alla distribuzione teorica; se però passiamo a $n = 1000$ e $s = 20$, osserviamo che invece le due distribuzioni sono molto più in accordo (Fig. 3b).

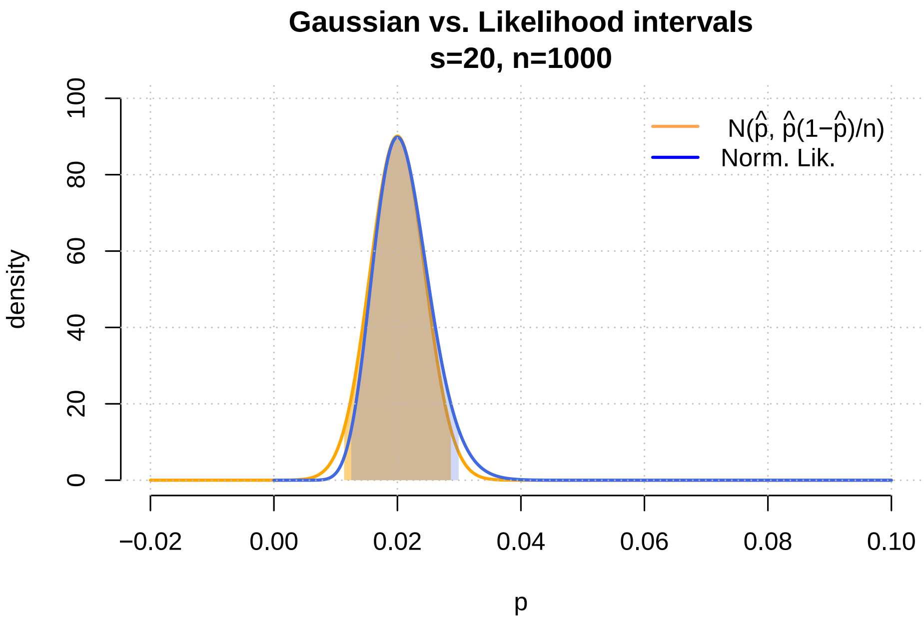
Un altro modo di analizzare il fenomeno è chiedersi cosa sarebbe accaduto per un maggior numero di successi, ad esempio con $s = 20$, lasciando invece $n = 100$; in tal caso i due intervalli sono molto più sovrapponibili:

$$I_\mathrm{N} \approx [0.122,\, 0.278]$$
$$I_\mathrm{L} \approx [0.130,\, 0.285]$$

In tale situazione la curva della verosimiglianza assume una forma più simile all'approssimazione Gaussiana impiegata per stimare I_N, meno asimmetrica a ttorno a $\hat{p} = 20/100 = 0.2$ (Fig. 4).



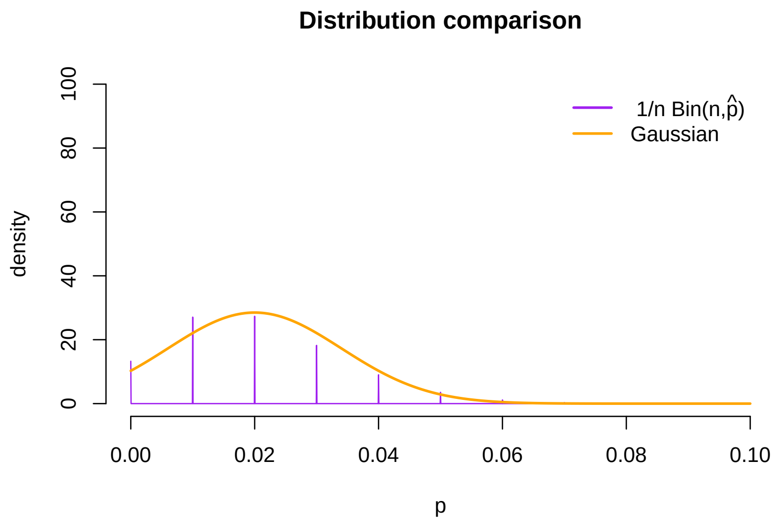
(a) $s = 2$, $n = 100$



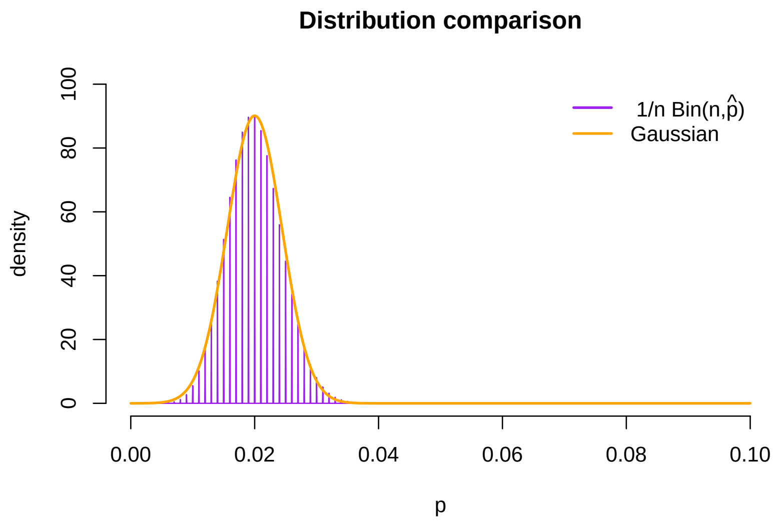
(b) $s = 20$, $n = 1000$

Fig. 2: Confronto, al crescere di n, tra gli intervalli di confidenza al 95% forniti dall'approssimazione Gaussiana e dalla verosimiglianza. La curva della verosimiglianza nei due grafici è stata normalizzata con area unitaria per poterla confrontare meglio con la Gaussiana.

Conclusione In termini generali l'intervallo di confidenza fornito dall'approssimazione Gaussiana può non comportarsi bene quando p si trova agli estremi del proprio insieme ammissibile (nel nostro caso $p \approx 0$): in tal caso occorrono davvero molte osservazioni affinché l'intervallo di confidenza Gaussiano sia affidabile; viceversa quando p è sufficientemente lontano da 0 (e da 1), allora anche con $n = 100$ osservazioni tale approssimazione fornisce risultati attendibili. Viceversa, poiché la verosimiglianza è analiticamente legata alla distribuzione vera, e non ad una sua approssimazione, gli intervalli di confidenza che essa fornisce non hanno il rischio di oltrepassare l'insieme di valori ammissibili per p.



(a) $s = 2$, $n = 100$



(b) $s = 20$, $n = 1000$

Fig. 3: Confronto al variare di n tra la distribuzione teorica per $\hat{p}$, assumendo che il vero valore di p corrisponda a $\hat{p} = 0.02$, e la sua approssimazione Gaussiana

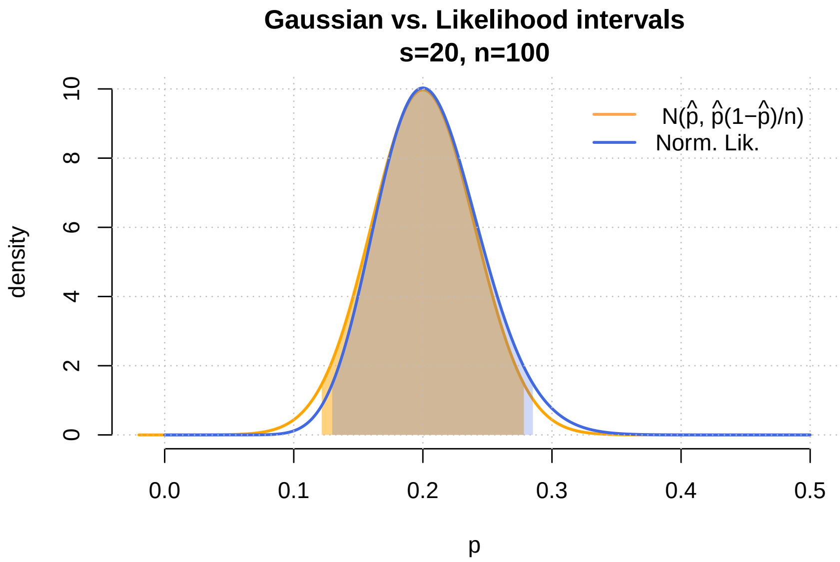


Fig. 4: Confronto tra gli intervalli di confidenza tramite approssimazione Gaussiana e verosimiglianza per $s = 20$, $n = 100$.

Tab. 1: Tempi di rottura per l'esercizio 2

y_i	41.53	18.73	2.99	30.34	12.33	117.52	73.02	223.63	4.00	26.78

Esercizio 2

Dati gli $n = 10$ valori y_i $(i = 1 \dots n)$ in Tab. 1 di un tempo di rottura, si vuole stimarne il tempo medio μ nell'ipotesi che esso soddisfi una distribuzione esponenziale:

$$\hat{\mu} = \bar{Y} = \frac{\sum_i y_i}{n} \approx 55.087$$

2.1

Sapendo che

$$X = \frac{n\bar{Y}}{\mu} \sim \text{Gamma}(n, 1)$$

si vogliono determinare gli estremi esatti L e U t.c. $P(L < \mu < U) = 0.95$,

$$X < \tau_a = \texttt{qgamma(0.975,n,1)} \iff \frac{n\bar{Y}}{\mu} < \tau_a \underset{\mu>0}{\iff} \frac{n\bar{Y}}{\tau_a} < \mu$$

$$X > \tau_b = \texttt{qgamma(0.025,n,1)} \iff \frac{n\bar{Y}}{\mu} > \tau_b \underset{\mu>0}{\iff} \frac{n\bar{Y}}{\tau_b} > \mu$$

ossia

$$L = \frac{n\bar{Y}}{\tau_a} < \mu < \frac{n\bar{Y}}{\tau_b} = U$$

Attraverso la funzione `qgamma(...)` di R, si ottiene

$$\tau_a = \texttt{qgamma(0.975, shape=n, scale=1)} \approx 17.0848$$
$$\tau_b = \texttt{qgamma(0.025, shape=n, scale=1)} \approx 4.795389$$

da cui,

$$\underbrace{32.24327}_{\approx L} < \mu < \underbrace{114.8749}_{\approx U}$$

Un procedimento alternativo, basato sul calcolo esplicito della densità di distribuzione di μ è riportato in appendice B.1

2.2

Impiegando l'approssimazione Gaussiana

$$Z = \frac{\bar{Y} - \mu}{\bar{Y}/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

abbiamo che l'intervallo di confidenza al 95% è dato da $q_L < Z < q_U$, con $q_L = \texttt{qnorm(p=0.025)} \approx -1.96$ e $q_U = \texttt{qnorm(p=0.975)} \approx 1.96$, da cui

$$\begin{aligned}
q_L < \frac{\bar{Y} - \mu}{\bar{Y}/\sqrt{n}} < q_U \underset{\bar{Y}>0}{\iff} & \; q_L \frac{\bar{Y}}{\sqrt{n}} < \bar{Y} - \mu < q_U \frac{\bar{Y}}{\sqrt{n}} \\
\iff & \; q_L \frac{\bar{Y}}{\sqrt{n}} - \bar{Y} < -\mu < q_U \frac{\bar{Y}}{\sqrt{n}} - \bar{Y} \\
\iff & \; -q_L \frac{\bar{Y}}{\sqrt{n}} + \bar{Y} > +\mu > -q_U \frac{\bar{Y}}{\sqrt{n}} + \bar{Y} \\
\iff & \; \underbrace{\bar{Y}\left(1 - \frac{q_U}{\sqrt{n}}\right)}_{\approx 20.94372} < \mu < \underbrace{\bar{Y}\left(1 - \frac{q_L}{\sqrt{n}}\right)}_{\approx 89.23028}
\end{aligned}$$

Esercizio 3

3.1

Assumendo un modello di regressione lineare

$$y_i = \alpha + \beta z_i + \epsilon_i \qquad \epsilon_i \text{ iid } \sim N(0, \sigma^2)$$

per gli $n = 12$ dati in Fig. 5 ($i = 1 \dots n$), i valori MLE per β e α sono:

$$\hat{\beta} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})^2} \approx 0.112$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{z} \approx -1.269$$

Il valore $\hat{\beta} > 0$ concorda col trend visivamente crescente di y_i all'aumentare di z_i.

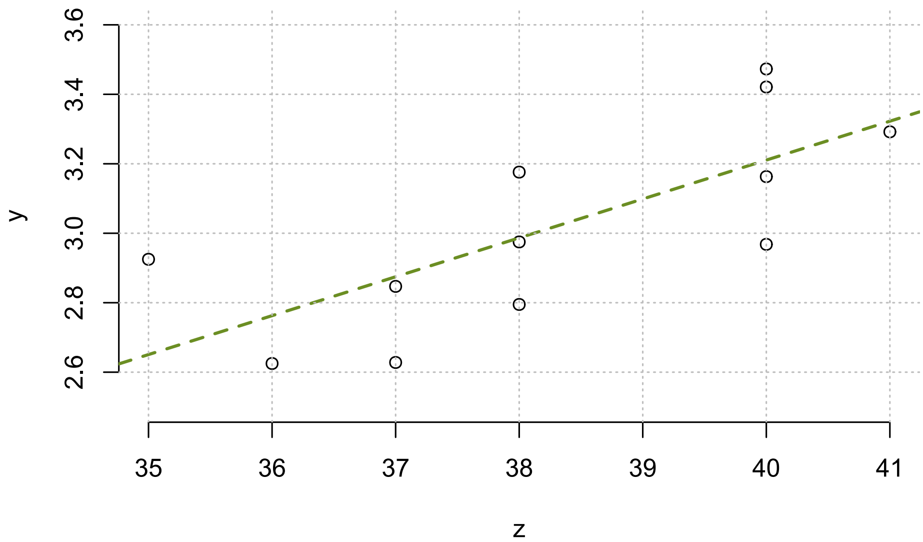


Fig. 5: Retta di regressione MLE per l'esercizio 3.

3.2

L'intervallo di confidenza al 95% per β è dato da (0.040, 0.184) ottenuto come

$$\begin{aligned}
&\hat{\beta} \pm t_{n-2,0.975}\,\hat{se} = \hat{\beta} \pm t_{10,0.975}\,\hat{se} \approx \hat{\beta} \pm 0.0720 \quad \text{dove} \\
&\hat{se} = \frac{S}{\sum_i (z_i - \bar{z})} \approx 0.032 \\
&S = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}} \approx 0.201 \\
&\hat{y}_i = \hat{\alpha} + \hat{\beta} z_i \\
&t_{10,0.975} \approx 2.228
\end{aligned}$$

3.3

Come statistica per validare l'ipotesi $H_0 : \beta = 0$ ($H_a : \beta \neq 0$) impieghiamo

$$T = \frac{\hat{\beta} - 0}{\hat{se}} \sim t_{10} \quad \Rightarrow \quad t_{\text{obs}} \approx 3.4657$$

Per testare l'ipotesi H_0 occorre valutare $P(T_{10} \geqslant t_{\text{obs}}) = 1 - P(T_{10} < t_{\text{obs}}) =$ `1-pt(q=3.4657,df=10)` ≈ 0.003, a cui corrisponde il p-value $= 2P(T_{10} \geqslant t_{\text{obs}}) \approx 0.006$: il valore ottenuto è molto inferiore alla soglia 0.01, al di sotto della quale c'è una forte evidenza contro l'ipotesi $\beta = 0$.

A Listati R

Listato 1: Codice R per l'esercizio 1

```
rm(list=ls())
library(pracma) #per bisect
n = 100
s = 2

hatp = s/n
delta95 = 1.96 * sqrt(hatp*(1-hatp)/n)
(rangeInfN95 = hatp-delta95)
(rangeSupN95 = hatp+delta95)

r <- function(p,hatp,s,n)
{
  return (s*log(p/hatp) + (n-s)*log((1-p)/(1-hatp)))
}

rint <- function(p,hatp,s,n,thr)
{
  return( r(p,hatp,s,n)-thr)
}

thr = -qchisq(p=0.95, df=1)/2
rangeInfLR95 = bisect(rint, a=0.0, b=hatp, hatp=hatp, s=s, n=n ,thr=thr)
rangeSupLR95 = bisect(rint, a=hatp, b=0.5, hatp=hatp, s=s, n=n, thr=thr)
(pinf = rangeInfLR95[[1]])
(psup = rangeSupLR95[[1]])
```

Listato 2: Codice R per l'esercizio 2

```
rm(list=ls())
y = c(41.53, 18.73, 2.99, 30.34, 12.33, 117.52, 73.02, 223.63, 4.00,
    26.78)

n = length(y)
hatmu = mean(y)

(tauA = qgamma(0.975, shape=n, scale=1))
(tauB = qgamma(0.025, shape=n, scale=1))
(L= n*hatmu/tauA)
(U= n*hatmu/tauB)

(gaussL = hatmu*(1-1.96/sqrt(n)))
(gaussU = hatmu*(1+1.96/sqrt(n)))
```

Listato 3: Codice R per l'esercizio 3

```
rm(list=ls())
z = c(40 , 38, 40, 35, 36, 37, 41, 40, 37, 38, 40, 38)
```

```
y = c(2.968, 2.795, 3.163, 2.925, 2.625, 2.847, 3.292, 3.473, 2.628,
    3.176, 3.421, 2.975)

n = numel(z)
barz = mean(z)
bary = mean(y)
numBeta = sum((z-barz)*(y-bary))
denBeta = sum((z-barz)^2)
hatbeta = numBeta/denBeta
hatalpha = bary - hatbeta*barz

t_n = qt(0.975, n-2)
fity = hatalpha + hatbeta*z;
S2 = sum((y-fity)^2)/(n-2)
se = sqrt(S2)/sqrt(sum((z-barz)^2))
(betainf = hatbeta -t_n*se)
(betasup = hatbeta +t_n*se)

tobs = hatbeta/(se)
(pH0 = 1 - pt(q=tobs, df=10))
(pval = 2*pH0)
```

B Extra

B.1 2.1

Poiché per $Y = c/X$ la sua densità di probabilità vale $f_Y(y) = |c|/y^2 f_X(c/y)$, se $X = (n\bar{Y})/\mu \sim \text{Gamma}(n, 1)$ vogliamo ricavare la pdf $f_\mu(y)$ per $\mu = (n\bar{Y})/X$:

$$
\begin{aligned}
f_\mu(y) &= \frac{n\bar{Y}}{\mu^2}\frac{1}{\Gamma(n)}\left(\frac{n\bar{Y}}{y}\right)^{n-1} e^{-(n\bar{Y})/y} \\
&= \frac{1}{y\Gamma(n)}\left(\frac{n\bar{Y}}{y}\right)^{n} e^{-(n\bar{Y})/y} \\
&= \frac{1}{y\Gamma(n)}\left(\frac{s}{y}\right)^{n} e^{-s/y} \quad \text{avendo posto } s = n\bar{Y} = \sum_{i=1}^{n} y_i \\
&= \frac{1}{s\Gamma(n)}\left(\frac{s}{y}\right)^{n+1} e^{-s/y}
\end{aligned}
$$

Una volta ottenuta la formula per la densità di probabilità abbiamo come conseguenza anche la funzione cumulativa di probabilità $F_\mu(y) = P(\mu \leqslant y)$:

$$
F_\mu(y) = \int_0^y f_\mu(t)\, dt = \frac{1}{s\Gamma(n)}\int_0^y \left(\frac{s}{t}\right)^{n+1} e^{-s/t}\, dt
$$

Pertanto, per determinare gli estremi dell'intervallo di confidenza al 95% per μ dobbiamo calcolare i quantili L e U tali che:

$$
\begin{aligned}
P(\mu \leqslant U) &= F_\mu(U) = 0.975 \\
P(\mu \leqslant L) &= F_\mu(L) = 0.025
\end{aligned}
$$

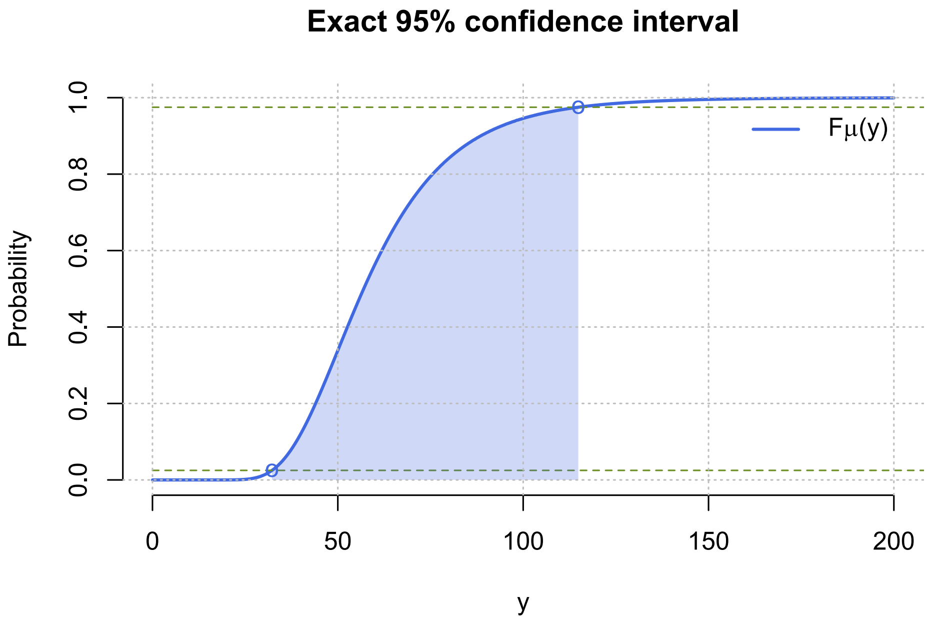


Fig. 6: Intervallo di confidenza al 95% "esatto" $[L, U]$ per l'esercizio 2.

L'andamento di $F_\mu(y)$, assieme all'individuazione di L e U è mostrata in Fig. 6, da cui ricaviamo che[2]

$$\underbrace{32.24327}_{L} < \mu < \underbrace{114.87494}_{U}$$

Impiegando gli intervalli di verosimiglianza abbiamo che per $\lambda = 1/\mu$, vale

$$l(\lambda) = \ln L(\lambda) = n \ln \lambda - \lambda \overbrace{\sum_i y_i}^{s} = \ln \lambda^n - s\lambda \quad \Rightarrow \quad L(\lambda) = \lambda^n e^{-s\lambda}$$

Applicando la proprietà di invarianza per $L(\lambda)$ abbiamo che, per $\mu = g(\lambda) = 1/\lambda$ (e posto $\hat{\mu} = s/n$) vale

$$L^*(\mu) = L(g^{-1}(\lambda)) = \left(\frac{1}{\mu}\right)^n e^{-s/\mu} = \mu^{-n} e^{-s/\mu}$$

$$R(\mu) = \frac{L^*(\mu)}{L^*(\hat{\mu})} = \frac{\mu^{-n} e^{-s/\mu}}{(s/n)^{-n} e^{-s/(s/n)}} = \frac{\mu^{-n} e^{-s/\mu}}{(s/n)^{-n} e^{-n}} = \left(\frac{n\mu}{s}\right)^{-n} e^{n-s/\mu} = \left(\frac{s}{n\mu}\right)^n e^{n-s/\mu}$$

Fig. 7 riporta il grafico per $R(\mu)$, assieme alla soglia $\tau_R = 0.15$ per individuare l'intervallo di confidenza al 95% basato sulla verosimiglianza: i due punti evidenziati su $R(\mu)$ corrispondono rispettivamente a $R(L)$ e $R(U)$; come si vede dal grafico, i valori trovati per L e U corrispondono indicativamente a quelli forniti dal criterio $R(\mu) > \tau_R$.

[2] Per il calcolo numerico di $F_\mu(y)$ è stata impiegata la funzione `integrate(...)` applicata su $f_\mu(y)$; per il calcolo dei quantili L e U, è stata impiegata la funzione `bisect(...)` per risolvere $F_\mu(y) - \tau_p = 0$ (con $\tau_p \in \{0.025, 0.975\}$).

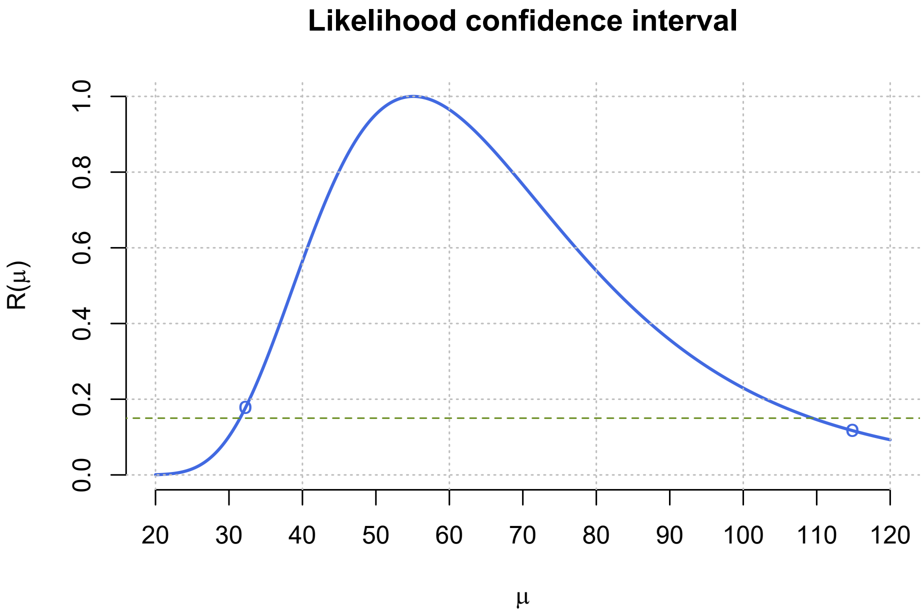


Fig. 7: Intervallo di confidenza al 95% per l'esercizio 2 basato sulla verosimiglianza a confronto con quello $[L, U]$ trovato tramite $F_\mu(y)$ (punti $\circ$).