Inferenza Statistica

Note in R

V. Gioia (e R. Pappadà, N. Torelli,)

12/12/2024

Contents

Esercizio 3 - Esercitazione 9

Il peso (in grammi) delle mozzarelle prodotte da un'azienda segue una distribuzione normale $\mathcal{N}(\mu, \sigma^2)$. Un campione di $n = 12$ mozzarelle ha fornito i seguenti valori

$$\sum_{i=1}^{n} y_i^2 = 60794 \quad \sum_{i=1}^{n} y_i = 852$$

dove le $y_i (i = 1, \dots, 12)$ denotano le osservazioni nel campione. Si verifichi l'ipotesi che la varianza sia $\sigma^2 = 36$ contro l'alternativa che sia inferiore a 36, ponendo il livello del test pari a 0.01

Sia Y la v.a. che descrive il peso (in grammi) delle mozzarelle prodotte da un'azienda ed è noto ch si distribuisca secondo una $\mathcal{N}(\mu, \sigma^2)$. Si vuole verificare a livello $\alpha = 0.01$ il seguente sistema di ipotesi

$$\begin{cases} H_0 : \sigma^2 = 36 \\ H_1 : \sigma^2 < 36 \end{cases}$$

Essendo μ ignota, uno stimatore non distorto per σ^2 è la varianza campionaria

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

e data la normalità della popolazione si ha che

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Pertanto sotto H_0 possiamo dire che

$$\frac{1}{\sigma_0^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$$

Quindi, essendo l'ipotesi alternativa unilaterale la regione di rifiuto risulterà del tipo $\mathcal{R} = (0, k)$

$$\alpha = P\left(\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 < k | H_0 \right) = P\left(\frac{1}{\sigma_0^2} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 < \chi^2_{n-1;\alpha} \right) = P\left(\sum_{i=1}^{n} (Y_i - \overline{Y})^2 < \sigma_0^2 \chi^2_{n-1;\alpha} \right)$$

avendo dentotato con $k = \chi^2_{n-1;\alpha}$. Pertanto la regione di rifiuto del test può essere equivalentemente espressa come

$$\mathcal{R} = \left\{ (y_1, \ldots, y_{12}) : \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (y_i - \overline{y})^2 < \chi^2_{n-1;\alpha} \right\} = \left\{ (y_1, \ldots, y_{12}) : \sum_{i=1}^{n} (y_i - \overline{y})^2 < \sigma_0^2 \chi^2_{n-1;\alpha} \right\}$$

che con $\alpha = 0.01$ risulta

$$\mathcal{R} = \left\{ (y_1, \ldots, y_{12}) : \frac{1}{\sigma_0^2} \sum_{i=1}^{n} (y_i - \overline{y})^2 < 3.05 \right\} = \left\{ (y_1, \ldots, y_{12}) : \sum_{i=1}^{n} (y_i - \overline{y})^2 < 109.8 \right\}$$

Quindi, poichè (equivalentemente)

- $$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 60794 - 12 \times (852/12)^2 = 302 \notin \mathcal{R}$$
- $$\frac{1}{\sigma_0^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 = 8.38 \notin \mathcal{R}$$

non vi è evidenza empirica per rifiutare H_0.

Si consideri il seguente esempio, supponiamo di generare 12 valori da una Normale con media 70 (grammi) e varianza 36 (grammi al quadrato). Intuitivamente, ci aspettiamo di avere evidenza in favore di $H_0 : \sigma^2 = 36$. Esploriamo il risultato della procedura in R:

- La funzione usata è *varTest* del pacchetto *EnvStats*
- Qui viene fornito soltanto il valore corrispondente a $\frac{1}{\sigma_0^2} \sum_{i=1}^{n} (y_i - \overline{y})^2$, il quale è di immediato utilizzo per il calcolo del p-value.
- Nota: I valori numerici sono leggermente diversi da quelli riportati nell'esercizio svolto in classe (considerate i dati usati nell'esercitazione come se fossero ottenuti da un idverse seme di generazione)

```r
set.seed(1)
n <- 12

mean <- 70
sigma <- 6

y <- rnorm(n, mean, sigma)
# Le seguenti sono simili a quelle riportate ell'esercizio
sum(y)
```

```
## [1] 859.3419
```

```r
sum(y^2)
```

```
## [1] 61799.33
```

```r
library(EnvStats)
varTest(y, sigma.squared = 36, alternative = "less")
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 36
##
## Alternative Hypothesis:          True variance is less than 36
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 23.66205
##
## Data:                            y
##
## Test Statistic:                  Chi-Squared = 7.230071
##
## Test Statistic Parameter:        df = 11
##
## P-value:                         0.2198425
##
## 95% Confidence Interval:         LCL =  0.00000
##                                  UCL = 56.89469
```

```r
var(y)
```

```
## [1] 23.66205
```

```r
# Valore riportato nell'output
(sum(y^2)- n * mean(y)^2)/36
```

```
## [1] 7.230071
```

```r
pchisq(7.230071, 11)
```

```
## [1] 0.2198425
```

Cosa succede se in realtà i dati fossero generati da un valore più plausibile con $H_1 : \sigma^2 < 36$. Intuitivamente, ci aspetteremo che se il vero valore del parametro sia molto distante da quello ipotizzato sotto H_0, si propenderebbe per il rifiuto di H_0 in favore dell'alternativa. Verifichiamolo in questo semplice esempio.

```
set.seed(1)
n <- 12

mean <- 70
sigma <- 3

y <- rnorm(n, mean, sigma)
sum(y)
```

```
## [1] 849.671
```

```
sum(y^2)
```

```
## [1] 60226.8
```

```
varTest(y, sigma.squared = 36, alternative = "less")
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 36
##
## Alternative Hypothesis:          True variance is less than 36
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 5.915513
##
## Data:                            y
##
## Test Statistic:                  Chi-Squared = 1.807518
##
## Test Statistic Parameter:        df = 11
##
## P-value:                         0.0009336653
##
## 95% Confidence Interval:         LCL =  0.00000
##                                  UCL = 14.22367
```

```
var(y)
```

```
## [1] 5.915513
```

```
# Valore riportato nell'output
(sum(y^2)- n * mean(y)^2)/36
```

```
## [1] 1.807518
```

```
pchisq(1.807518, 11)
```

```
## [1] 0.0009336657
```