Lo stimatore S^2 della varianza campionaria

Siano $x_1, x_2, \ldots, x_n$ n osservazioni provenienti da un campione casuale semplice estratto da una popolazione P *qualsiasi*. Sia μ la media vera di tutte le osservazioni appartenenti a P e sia σ^2 la loro varianza. Sappiamo che per stimare la varianza vera σ^2 a partire dai dati campionari $x_1, x_2, \ldots, x_n$ dobbiamo calcolare la somma dei quadrati degli scarti fra ciascun valore x_i e la media (aritmetica) $\overline{x}$ delle osservazioni (calcolare cioè la *devianza*) e dividerla per $n-1$ (il numero delle osservazioni meno 1, ovvero per i *gradi di libertà*):

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}.$$

Vogliamo far vedere che lo stimatore S^2 (di cui s^2 è una realizzazione) è *non distorto*, ovvero che $E(S^2) = \sigma^2$.
Per fare questo dobbiamo fare riferimento ad un modello teorico. Siano $X_1, X_2, \ldots, X_n$ n variabili casuali i.i.d. ciascuna con valore atteso μ e con varianza σ^2: $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$. Definiamo

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} \tag{1}$$

la variabile casuale *varianza campionaria* e troviamone il valore atteso.
Prima di iniziare, abbiamo bisogno di un risultato che ci servirà per eseguire la dimostrazione. Vogliamo trovare $E(\overline{X}^2)$, cioè il valore atteso del *quadrato* della media campionaria.
Sappiamo che $Var(\overline{X}) = \sigma^2/n$ e che, ovviamente, $E(\overline{X}) = \mu$. La relazione, che vale per qualsiasi variabile casuale X, $Var(X) = E(X^2) - (E(X))^2$ ci permette di dire che

$$E(\overline{X}^2) = Var(\overline{X}) + (E(X))^2 = \frac{\sigma^2}{n} + \mu^2.$$

Il primo passo della nostra dimostrazione consiste nel manipolare algebricamente la quantità al numeratore della (1) espandendone il quadrato.

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} \left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)$$

Applichiamo ora le proprietà dell'operatore sommatoria, ottenendo

$$\sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.$$

Applichiamo il valore atteso e le sue proprietà:

$$E\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) = \sum_{i=1}^{n} E(X_i)^2 - n\left(\frac{\sigma^2}{n} + \mu^2\right).$$

Nel passaggio precedente abbiamo sfruttato quanto detto sopra a proposito del valore atteso del quadrato della media campionaria. Ora sfruttiamo quanto sappiamo a proposito del valore atteso del quadrato di una variabile casuale X qualsiasi ottenendo

$$n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 = n\sigma^2 - \sigma^2 = (n-1)\sigma^2.$$

Siamo a questo punto arrivati alla fine della dimostrazione, perché, considerando anche il denominatore della (1) otteniamo:

$$E\left(S^2\right) = E\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2,$$

per cui lo stimatore S^2 della varianza campionaria è uno stimatore non distorto.
Questo risultato è valido per un campione casuale semplice estratto da una qualsiasi popolazione. se il campione è stato estratto da una distribuzione normale con media μ e varianza σ^2, allora siamo anche in grado di trovare la distribuzione di campionamento di S^2.
Anche in questo caso abbiamo bisogno di richiamare un risultato già noto:

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2,$$

cioè che la somma di n variabili casuali indipendenti normali standardizzate al quadrato segue la distribuzione *chi quadrato con n gradi di libertà*.

Nell'espressione precedente si può aggiungere e togliere la media campionaria

$$\sum_{i=1}^{n} \left(\frac{\left(X_i - \overline{X}\right) + \left(\overline{X} - \mu\right)}{\sigma} \right)^2 \sim \chi^2_n.$$

Espandendo il quadrato e applicato le proprietà della sommatoria troviamo che il doppio prodotto $\sum_{i=1}^{n} 2\left(X_i - \overline{X}\right)\left(\overline{X} - \mu\right)$ si annulla e, quindi, quello che rimane è

$$\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2} + \frac{n\left(\overline{X} - \mu\right)^2}{\sigma^2} \sim \chi^2_n.$$

Dal momento che il secondo addendo è distribuito come una variabile casuale chi quadrato con 1 grado di libertà, per la proprietà additiva della distribuzione chi quadrato, il primo addendo è distribuito come una variabile casuale chi quadrato con $n-1$ gradi di libertà:

$$\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Dal momento che

$$\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2} \times \frac{\sigma^2}{n-1} = S^2,$$

otteniamo il risultato seguente:

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

che è il risultato che cercavamo.
Ad esempio, possiamo dire che se $X \sim N(30, 25)$ ed estraiamo una campione casuale semplice di 20 osservazioni, allora il 95% "centrale" delle possibili varianze stimate sarà compreso fra 11.7 e 43.2. Infatti, i quantili 2.5% e 97.5% di una variabile casuale chi quadrato con 19 gradi di libertà sono

```
> (q <- qchisq(c(0.025,0.975),19))
[1]  8.906516 32.852327
```

Per ottenere i corrispondenti valori delle varianze campionarie relative al nostro esempio, dobbiamo moltiplicare ciascuno di questi due valori per la costante di proporzionalità data da $\sigma^2/(n-1)$, cioè per (circa) 1.316:

```
> k <- 25/19
> q*k
[1] 11.71910 43.22675
```

Verifichiamo questo risultato con una simulazione.

```
> m <- 30; s <- 5; n <- 20; nrep <- 100000
> set.seed(123456)
> x <- rnorm(n*nrep,m,s)
> X <- matrix(x,ncol=n)
> v <- apply(X,1,var)
> sum((11.71910 <= v) & (v <= 43.22675))/nrep
[1] 0.95108
```

Nella nostra simulazione il numero di campioni la cui varianza stimata è caduta entro i limiti specificati in precedenza è sostanzialmente uguale al 95%. Per di più, questo 95% è davvero "centrale", nel senso che il numero di campioni per i quali le stime di σ^2 sono caduti al di fuori dell'intervallo considerato sono dividi quasi esattamente a metà (2.5% al di sotto dell'estremo inferiore e 2.5% al di sopra dell'estremo superiore).

```
> sum(v < 11.71910)/nrep
[1] 0.02468
> sum(v > 43.22675)/nrep
[1] 0.02424
```

La figura 1 mostra nel pannello di sinistra il *Q-Q plot* che confronta i centili campionari con quelli attesi di una distribuzione di probabilità proporzionale a un chi quadrato con 19 gradi di libertà (con una costante di proporzionalità pari a 25/19). Esiste una corrispondenza quasi perfetta fra quantili campionari e quantili teorici, come evidenziato dall'allineamento dei punti lungo la bisettrice.

```
> q <- c(1:99)/100
> qo <- quantile(v,q)
```

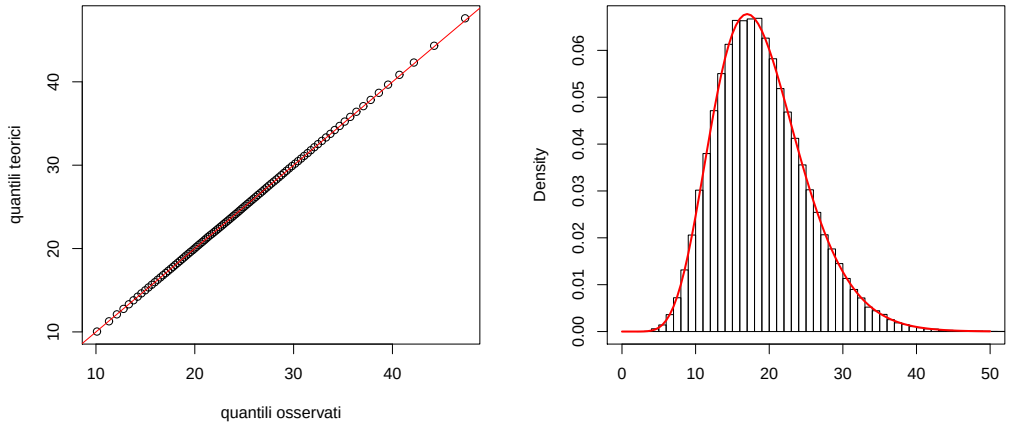

Figura 1: A sinistra: *Q-Q plot* che confronta i quantili empirici della simulazione con quelli teorici (cioè quelli di una variabile casuale chi quadrato con 19 gradi di libertà moltiplicati per 25/19). A destra: La curva in rosso è la funzione di densità di probabilità di una variabile casuale chi quadrato con 19 gradi di libertà. L'istogramma è relativo ai valori delle varianze campionarie ottenute nella simulazioni divisi per 25/19.

```
> k <- s^2/(n-1)
> qt <- qchisq(q,n-1)*k
> qqplot(qo,qt,xlab="quantili osservati",
+ ylab="quantili teorici")
> abline(0,1, col="red"
```

Nel pannello di destra della figura 1 è riportato l'istogramma dei valori ottenuti facendo il rapporto fra le varianze campionarie e la costante di proporzionalità 25/19. Questi valori seguono una distribuzione chi quadrato con 19 gradi di libertà, come evidenziato dalla funzione di densità di probabilità sovraimposta.

```
> k <- s^2/(n-1)
> cq <- v/k
> hist(cq, breaks=50, prob=TRUE, xlim=c(0,50), xlab="",
+ main=""); box()
> curve(dchisq(x,(n-1)), add=TRUE, col="red", lwd=2)
```

È inoltre immediato calcolare la varianza di S^2 (ricordando sia le proprietà dell'operatore varianza, sia che la varianza di una variabile casuale chi quadrato con ν gradi di libertà è uguale a 2ν):

$$Var(S^2) = Var\left(\frac{\sigma^2}{n-1}\chi^2_{n-1}\right) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}.$$

Questo risultato è confermato dai risultati della nostra simulazione:

```
> c(var(v),2*s^4/(n-1))
[1] 65.25941 65.78947
```