Problemi di stima puntuale

Una breve introduzione

Luca La Rocca[1]

Dipartimento di Scienze Fisiche, Informatiche e Matematiche
Università degli Studi di Modena e Reggio Emilia

Insegnamento di Analisi Statistica dei Dati
Corsi di Laurea Magistrale in Informatica e Matematica
Anno Accademico 2018/2019

[1] http://personale.unimore.it/rubrica/dettaglio/llarocca

Stimatori e stime puntuali

Dato il modello statistico definito da $f(z|\theta)$, $z \in \mathcal{Z}$, al variare di θ in H, si dice stimatore puntuale per il parametro di interesse $\xi = h(\theta)$ una qualsiasi statistica $T = g(Z)$ tale che l'immagine di $\mathcal{Z}$ attraverso g sia inclusa nell'immagine di H attraverso h; in simboli $g[\mathcal{Z}] \subseteq h[H]$.

Se $\theta = (\xi, \psi)$, eventualmente dopo un cambio di parametrizzazione, si dice che ψ è un parametro di disturbo.

Il valore $g(z)$ si dice stima puntuale di $h(\theta)$ con T osservando il dato z.

Supporremo senz'altro h e g a valori reali ($h : H \to \mathbb{R}$ e $g : \mathcal{Z} \to \mathbb{R}$).

Nel caso tipico in cui $z = x_{1:n}$ lo stimatore $T_n = g_n(X_{1:n})$ può essere considerato il generico elemento di una successione di stimatori.

Campione casuale da una popolazione normale

Sia $Z = X_{1:n}$ con $X_1, X_2, \ldots, X_n$ i.i.d. come $X \sim \mathrm{Norm}(\mu, \sigma^2)$, dove $\mu \in \mathbb{R}$ e $\sigma^2 \in \mathbb{R}_+^*$, quindi $\mathcal{Z} = \mathbb{R}^n$, $\theta = (\mu, \sigma^2)$ e $H = \mathbb{R} \times \mathbb{R}_+^*$.

Se ci interessa $\xi = \sigma$, quindi $h(\theta_1, \theta_2) = \sqrt{\theta_2}$, mentre $\psi = \mu$ ha un ruolo di disturbo, allora

$$D_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2},$$

dove $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, è uno stimatore per σ, secondo la definizione data, perché è una statistica a valori (quasi certamente) in $\mathbb{R}_+^*$.

In effetti D_n è lo stimatore di massima verosimiglianza per σ...

Espressione della verosimiglianza normale

...perché possiamo scrivere

$$\begin{aligned} f(x_{1:n}|\mu,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ &\propto \exp\left\{-\frac{n}{2}\log\sigma^2 - \frac{n}{2\sigma^2}\left(\overline{x_n^2} - 2\bar{x}_n\mu + \mu^2\right)\right\}, \end{aligned}$$

dove $\overline{x_n^2} = n^{-1}\sum_{i=1}^{n} x_i^2$, quindi

$$\ell(\mu,\sigma^2; x_{1:n}) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2\sigma^2}\left\{(\bar{x}_n-\mu)^2 + d_n^2(x_{1:n})\right\},$$

dove $d_n^2(x_{1:n}) = \overline{x_n^2} - \bar{x}_n^2 = n^{-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$...

Stimatori di massima verosimiglianza

. . . e la verosmiglianza sarà massimizzata prendendo

$$\begin{aligned} \hat{\mu}(x_{1:n}) &= \bar{x}_n \\ \hat{\sigma}^2(x_{1:n}) &= d_n^2(x_{1:n}); \end{aligned}$$

pertanto $D_n = \sqrt{d_n^2(X_{1:n})} = h(\hat{\Theta}_n)$ sarà lo stimatore di massima verosimiglianza per σ, dove $\hat{\Theta}_n = (\hat{\mu}(X_{1:n}), \hat{\sigma}^2(X_{1:n})) = (\bar{X}_n, D_n^2)$.

Ogni stimatore di massima verosimiglianza $T = h(\hat{\Theta}) = h(\hat{\theta}(Z))$ soddisfa la definizione data, essendo $T = g(Z)$ con $g = h \circ \hat{\theta}$.

Altri stimatori (e non)

Un altro stimatore per σ, nell'esempio del campione casuale da una popolazione normale, secondo la definizione data, è la statistica

$$T = |\bar{X}_n|,$$

perché, come D_n, prende valori (quasi certamente) in $\mathbb{R}^*_+$; si tratta di uno stimatore "scriteriato", nel senso che viene qui proposto senza un criterio che lo giustifichi e anzi, essendo $\bar{X}_n$ lo stimatore di massima verosimiglianza per μ, appare ragionevole solo ipotizzando $\sigma = |\mu|$.

La definizione di stimatore garantisce solo che le stime ottenute con T siano valori ammissibili per ξ: la statistica $\bar{X}_n$ non è uno stimatore per il parametro σ, perché assume (con probabilità positiva) valori negativi.

Un criterio di valutazione

Possiamo valutare le prestazioni frequentiste di uno stimatore puntuale $T = g(Z)$ per il parametro di interesse $\xi = h(\theta)$ mediante la funzione

$$\mathrm{MSE}_{T,\xi}(\theta) = \mathbb{E}_\theta(T - \xi)^2, \qquad \theta \in H,$$

detta errore quadratico medio (mean squared error) dello stimatore.

La sua radice quadrata (root mean squared error)

$$\mathrm{RMSE}_{T,\xi}(\theta) = \sqrt{\mathrm{MSE}_{T,\xi}(\theta)}, \qquad \theta \in H,$$

fornisce un valore tipico per l'errore di stima in funzione del parametro indice; possiamo stimarlo se disponiamo di uno stimatore per θ...

Una valutazione concreta

...per esempio, se $\hat{\Theta} = \hat{\theta}(Z)$ è lo stimatore di massima verosimiglianza per θ e osserviamo $Z = z$, possiamo affiancare alla stima $g(z)$ per ξ la quantità

$$\hat{\mathrm{rmse}}_{T,\xi}(z) = \mathrm{RMSE}_{T,\xi}(\hat{\theta}(z))$$

e affermare che essa costituisce un confine inferiore per la nostra incertezza residua su ξ (dopo l'osservazione del dato) visto che, se anche $\hat{\theta}(z)$ fosse esattamente il valore ignoto di θ, ripetendo l'osservazione e stimando ξ con T ci aspetteremmo, in media, un errore di stima pari a $\hat{\mathrm{rmse}}_{T,\xi}(z)$.

Un esempio specifico

Nell'esempio del campione casuale da una popolazione normale, se ci interessa $\xi = \mu$, mentre $\psi = \sigma^2$ ha un ruolo di disturbo, troviamo

$$\mathrm{MSE}_{\bar{X}_n,\mu}(\mu, \sigma^2) = \mathbb{E}_{\mu,\sigma^2}(\bar{X} - \mu)^2 = \frac{\sigma^2}{n}, \qquad \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}_+^*,$$

dove la seconda uguaglianza si ottiene osservando che

$$\begin{aligned}\mathbb{E}_{\mu,\sigma^2}(\bar{X} - \mu)^2 &= \mathbb{E}_{\mu,\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right\}^2 \\ &= \frac{1}{n^2}\,\mathbb{E}_{\mu,\sigma^2}\left\{\sum_{i=1}^{n}(X_i - \mu)^2 + 2\sum_{1\le i<j\le n}(X_i - \mu)(X_j - \mu)\right\}\end{aligned}$$

e $\mathbb{E}_{\mu,\sigma^2}(X_i - \mu)^2 = \sigma^2$, mentre $\mathbb{E}_{\mu,\sigma^2}(X_i - \mu)(X_j - \mu) = 0$.

Una valutazione specifica

Per i dati presentati nella sezione 1.1.7 del testo di riferimento[2], supponendo di avere a che fare con un campione casuale da una popolazione normale e senza fare distinzione di genere, si ha

$$\begin{aligned} \bar{x}_n &= 2449.2 \\ d_n(x_{1:n}) &= 237.8, \end{aligned}$$

con $n = 185$, quindi troviamo

$$\hat{\mathrm{rmse}}_{\bar{X}_n,\mu}(x_{1:n}) = \frac{d_n(x_{1:n})}{\sqrt{n}} = \frac{237.8}{\sqrt{185}} \simeq 17.5$$

come confine inferiore della nostra incertezza residua su μ; scriveremo $\mu = 2449.2\,(17.5)$ per riassumere il nostro risultato.

[2] L. Held & D. Sabanés Bové. Applied Statistical Inference. Springer, 2014.

Altri criteri di valutazione

Una possibile alternativa all'errore quadratico medio è la funzione

$$\mathrm{MAE}_{T,\xi}(\theta) = \mathbb{E}_\theta |T - \xi|, \qquad \theta \in H,$$

detta errore assoluto medio (mean absolute error) dello stimatore.

Più in generale, data una funzione di perdita $\mathcal{L}(t, \xi)$, $t \in \mathbb{R}$, $\xi \in \Re$, a valori in $\mathbb{R}_+$, dove $\mathcal{L}(t, \xi)$ rappresenta la perdita che subiamo usando t come stima di ξ, possiamo valutare le prestazioni frequentiste di T come stimatore di ξ mediante la funzione

$$R_{T,\xi}(\theta) = \mathbb{E}_\theta \mathcal{L}(T, \xi), \qquad \theta \in H,$$

detta funzione di rischio; anche se idealmente si dovrebbe scegliere $\mathcal{L}$ in funzione dell'applicazione, nel seguito, per semplicità e concretezza, prenderemo senz'altro $\mathcal{L}(t, \xi) = (t - \xi)^2$ e cioè $R_{T,\xi}(\theta) = \mathrm{MSE}_{T,\xi}(\theta)$.

Approccio decisionale

Non è possibile scegliere T minimizzando $\mathrm{MSE}_{T,\xi}(\theta)$ uniformemente rispetto a θ, se esistono θ' e θ'' in H tali che $\xi' = h(\theta') \neq h(\theta'') = \xi''$, perché $T' \equiv h(\theta')$ è ottimo per $\xi = \xi'$ e $T'' \equiv h(\theta'')$ è ottimo per $\xi = \xi''$; in caso contrario $\xi = h(\theta)$, per ogni θ in H, è noto a priori.

Lo Statistico può provare a determinare uno stimatore ottimo $T^\star$ adottando una delle due seguenti strategie:

1. $T^\star = \operatorname{argmin}_T \mathbb{E}\left\{\mathrm{MSE}_{T,\xi}(\Theta)\right\}$, dove Θ è una variabile aleatoria la cui distribuzione rappresenta le informazioni che abbiamo a priori sul parametro θ (stimatore bayesiano).
2. $T^\star = \operatorname{argmin}_T \sup_{\theta \in H} \mathrm{MSE}_{T,\xi}(\theta)$, dove consideriamo, per ogni scelta di T, il caso più sfavorevole a T (stimatore minimax).

La strategia 1 immagina θ scelto da una Natura indifferente,
la strategia 2 immagina θ scelto da un Avversario ostile.

Distorsione e varianza

Vale in generale la seguente decomposizione:

$$\mathrm{MSE}_{T,\xi}(\theta) = B_\theta^2(T;\xi) + \mathrm{Var}_\theta(T),$$

dove

$$B_\theta(T;\xi) = \mathbb{E}_\theta\{T - \xi\} = \mathbb{E}_\theta(T) - \xi$$

è detta distorsione (bias) dello stimatore.

Per esempio, nel caso di un campione casuale da una popolazione normale, volendo stimare σ^2 con D_n^2, troviamo

$$B_{\mu,\sigma^2}(D_n^2;\sigma^2) = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n} < 0$$

perché...

Esempio di distorsione negativa

...possiamo scrivere

$$
\begin{aligned}
nD_n^2 &= \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X}_n)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 + \sum_{i=1}^{n}(\bar{X}_n - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\bar{X}_n - \mu) \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 + n(\bar{X}_n - \mu)^2 - 2n(\bar{X}_n - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X}_n - \mu)^2
\end{aligned}
$$

e quindi $\mathbb{E}_{\mu,\sigma^2}(nD_n^2) = n\sigma^2 - n\sigma^2/n = (n-1)\sigma^2$.

Correttezza

Lo stimatore puntuale T per $\xi = g(\theta)$ si dice corretto quando la sua distorsione è identicamente nulla:

$$B_\theta(T;\xi) \equiv 0 \qquad \Leftrightarrow \qquad \mathbb{E}_\theta(T) \equiv g(\theta).$$

Per esempio, nel caso di un campione casuale da una popolazione normale, volendo stimare μ con $\bar{X}_n$, troviamo

$$\mathbb{E}_{\mu,\sigma^2}(\bar{X}_n) = \mathbb{E}_{\mu,\sigma^2}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mu,\sigma^2}X_i = \frac{n\mu}{n} = \mu$$

e quindi $\bar{X}_n$ è uno stimatore corretto per μ.

Invece, volendo stimare σ^2 con D_n^2, visto che $B_{\mu,\sigma^2}(D_n^2;\sigma^2) < 0$, D_n^2 è uno stimatore distorto per σ^2; in media D_n^2 sottostima σ^2.

Vincolo di correttezza

La distorsione è un problema in sé, perché uno stimatore distorto erra (sovrastima o sottostima) anche in media.

Questa considerazione suggerisce, per stimare σ^2, di correggere D_n^2:

$$S_n^2 = \frac{n}{n-1} D_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.$$

Suggerisce inoltre di cercare uno stimatore ottimo tra quelli corretti:

$$T^\star = \underset{T | \mathbb{E}_\theta(T) = \xi}{\operatorname{argmin}} \operatorname{Var}_\theta(T)$$

uniformemente rispetto a θ (stimatore corretto di varianza uniformemente minima[3]) tenendo conto che $B_\theta^2(T; \xi) \equiv 0$.

[3] In inglese: Uniformly Mininum Variance Unbiased Estimator (UMVUE).

Problemi con la correttezza

Se stimiamo σ con $S_n = \sqrt{S_n^2}$ la correttezza si perde per strada:

$$\mathbb{E}_{\mu,\sigma^2}(S_n) = \mathbb{E}_{\mu,\sigma^2}\left(\sqrt{S_n^2}\right) < \sqrt{\mathbb{E}_{\mu,\sigma^2}(S_n^2)} = \sqrt{\sigma^2} = \sigma$$

in virtù della disuguaglianza di Jensen.

La correttezza è dunque, in generale, incompatibile con l'equivarianza per cambi di parametrizzazione.

Inoltre, in generale, il vincolo di correttezza risulta essere eccessivamente rigido...

Correttezza impossibile

...per esempio, preso $Z \sim \mathrm{Geom}(\pi) + 1$, con $\pi \in]0, 1]$,
dove escludiamo l'ipotesi $\pi = 0$, perché risulterebbe in $Z = \infty$,
uno stimatore corretto $T = g(Z)$ per π sarebbe tale che

$$\sum_{z=1}^{\infty} g(z)(1-\pi)^{z-1} = 1, \qquad 0 < \pi < 1,$$

visto che $f(z|\pi) = \pi(1-\pi)^{z-1}$, $z = 1, 2, \ldots$; quindi necessariamente avremmo $g(1) = 1$ e $g(z) = 0$ per $z \neq 1$ (in virtù dell'unicità dello sviluppo in serie di potenze).

La statistica $T = g(Z)$ così definita non soddisfa la definizione di stimatore, perché assume il valore 0 (escluso a priori); inoltre...

Una stima più verosimile

...sembra assai più ragionevole optare per lo stimatore di massima verosimiglianza $\hat{\Pi} = \hat{\pi}(Z) = 1/Z$:

- se $z = 1$, allora $f(z|\pi) = \pi$ è massimizzata da $\hat{\pi}(1) = 1$;
- se $z \neq 1$, allora $f(z|1) = 0$, mentre

$$\ell(\pi; z) = \log(\pi) + (z-1)\log(1-\pi), \qquad 0 < \pi < 1,$$
$$s(\pi; z) = \frac{1}{\pi} - \frac{z-1}{(1-\pi)}, \qquad 0 < \pi < 1,$$

quindi l'equazione del punteggio fornisce

$$\frac{\hat{\pi}(z)}{1-\hat{\pi}(z)} = \frac{1}{z-1},$$

da cui segue $\hat{\pi}(z) = 1/\{1 + (z-1)\} = 1/z$.

Un'utile formula alternativa

Se $X_1, X_2, \ldots, X_n$ sono i.i.d. come X, con $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathrm{Var}(X)$ e $\gamma_4 = \mathbb{E}[\{X - \mathbb{E}(X)\}^4]$ finiti, sfruttando la riscrittura

$$S_n^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2,$$

si trova

$$\mathrm{Var}(S_n^2) = \frac{1}{n} \left(\gamma_4 - \frac{n-3}{n-1} \sigma^4 \right)$$

e naturalmente

$$\mathrm{Var}(D_n^2) = \left(\frac{n-1}{n} \right)^2 \mathrm{Var}(S_n^2).$$

Prestazioni a confronto

Nell'esempio del campione casuale da una popolazione normale, tenendo presente che $\gamma_4 = 3\sigma^4$, troviamo

$$\mathrm{Var}_{\mu,\sigma^2}(S_n^2) = \frac{1}{n}\left(3\sigma^4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{2}{n-1}\sigma^4$$

$$\mathrm{Var}_{\mu,\sigma^2}(D_n^2) = \frac{(n-1)^2}{n^2}\frac{2}{n-1}\sigma^4 = \frac{2(n-1)}{n^2}\sigma^4$$

e quindi

$$\mathrm{MSE}_{S_n^2,\sigma^2}(\mu,\sigma^2) = 0 + \frac{2}{(n-1)}\sigma^4 = \frac{2}{n-1}\sigma^4$$

$$\mathrm{MSE}_{D_n^2,\sigma^2}(\mu,\sigma^2) = \frac{1}{n^2}\sigma^4 + \frac{2(n-1)}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4$$

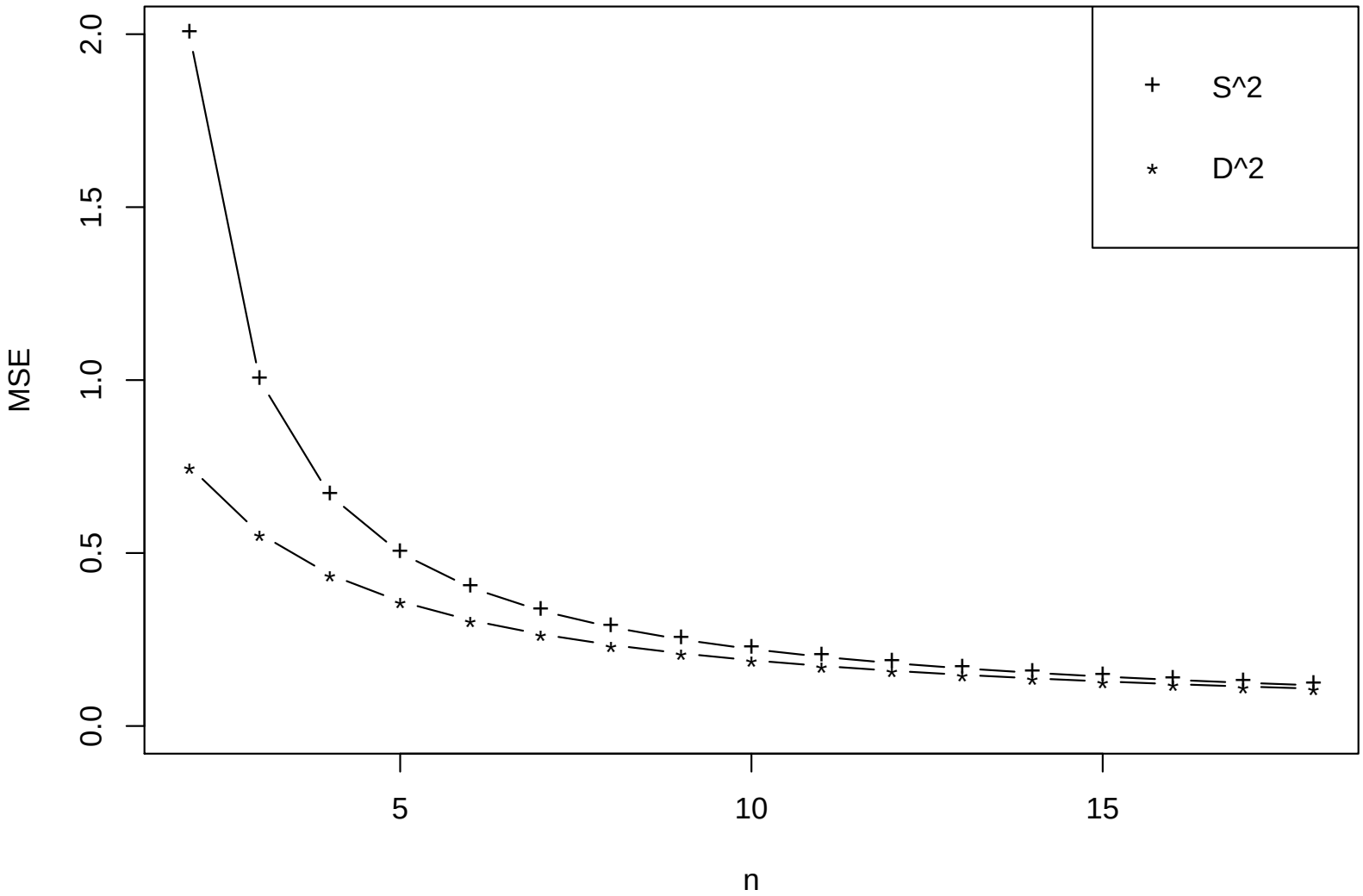
dopo di che...

Correttezza sconveniente

. . . possiamo confrontare graficamente le prestazioni di D_n^2 e S_n^2, in termini di $\mathrm{MSE}(\mu, \sigma^2)/\sigma^4$, usando il software R:

```
n = 2:18
MSE = 2/(n-1)
MSEwithBias = (2*n-1)/(n^2)
plot(n,MSE,type="b",pch="+",ylim=c(0,max(MSE)))
title(main="Variance estimation in the normal model")
lines(n,MSEwithBias,type="b",pch="*")
legend("topright",legend=c("S^2","D^2"),pch=c("+","*"))
```

Il grafico seguente mostra come la correzione di D_n^2 in S_n^2 risulti sconveniente in termini di errore quadratico medio.


Variance estimation in the normal model
MSE
0.0
0.5
1.0
1.5
2.0
5
10
15
n
S^2
D^2

In conclusione

Visto che la varianza può essere utilmente scambiata con la distorsione ai fini della riduzione dell'errore quadratico medio (bias-variance tradeoff) non conviene considerare irrinunciabile il vincolo di correttezza (per altro incompatibile con l'equivarianza per cambi di parametrizzazione e non sempre soddisfacibile).

Resta però auspicabile che

$$B_\theta^2(T;\xi) \ll \mathrm{Var}_\theta(T),$$

cioè che la distorsione sia trascurabile rispetto alla varianza, essendo la distorsione un problema in sé.

In effetti, se $|B_\theta(T;\xi)| \ll \mathrm{Sd}_\theta(T)$, la distorsione contribuisce in modo praticamente nullo all'incertezza residua su ξ (stimato con T).