De bootstrap methode

Sytse Knypstra

2009

We introduceren de bootstrap methode voor het schatten van verdelingen en parameters door te vergelijken met de klassieke manier van schatten en met het idee van simulatie.

Schatten (klassiek)

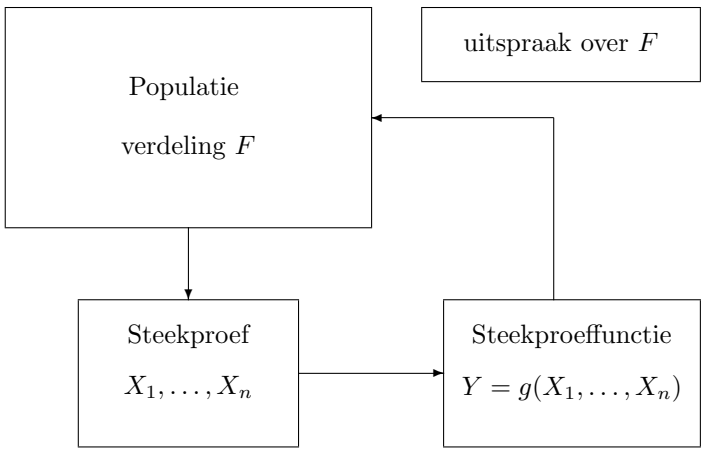
Populatie: verdeling F is niet volledig bekend. We willen een uitspraak doen over F of over een parameter θ die met F samenhangt.

Daartoe nemen we een aselecte steekproef $X_1, \ldots, X_n$ uit F.

We vatten deze samen in een steekproeffunctie $Y = g(X_1, \ldots, X_n)$.

Gegeven F, kunnen we analytisch de (asymptotische) verdeling van Y bepalen.

Op grond van de uitkomst van Y doen we een uitspraak over F of over θ.



Simulatie

De verdeling F is bekend.

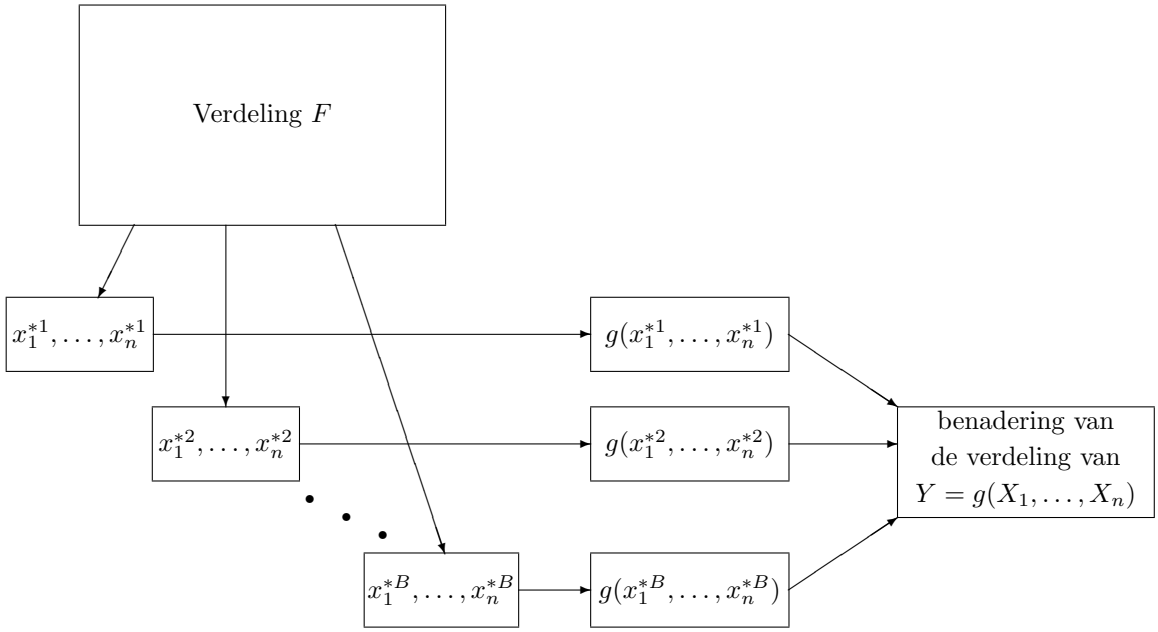
We nemen een aselecte steekproef $X_1, \ldots, X_n$ uit F.

We vormen de steekproeffunctie $Y = g(X_1, \ldots, X_n)$.

Analytisch kunnen we niets over de verdeling van Y afleiden.

Wel kunnen we door simulatie de verdeling van Y benaderen:

- Trek uit F een aselecte steekproef van n stuks met uitkomst $x_1^{*1}, \ldots, x_n^{*1}$ en bepaal $g(x_1^{*1}, \ldots, x_n^{*1})$.
- Voer ditzelfde proces B keer (onafhankelijk van elkaar) uit. We krijgen uitkomsten $g(x_1^{*1}, \ldots, x_n^{*1}), \ldots, g(x_1^{*B}, \ldots, x_n^{*B})$.
- Deze uitkomsten samen leveren een benadering op van de verdeling van $Y = g(X_1, \ldots, X_n)$.



Bootstrap

De verdeling F is onbekend.

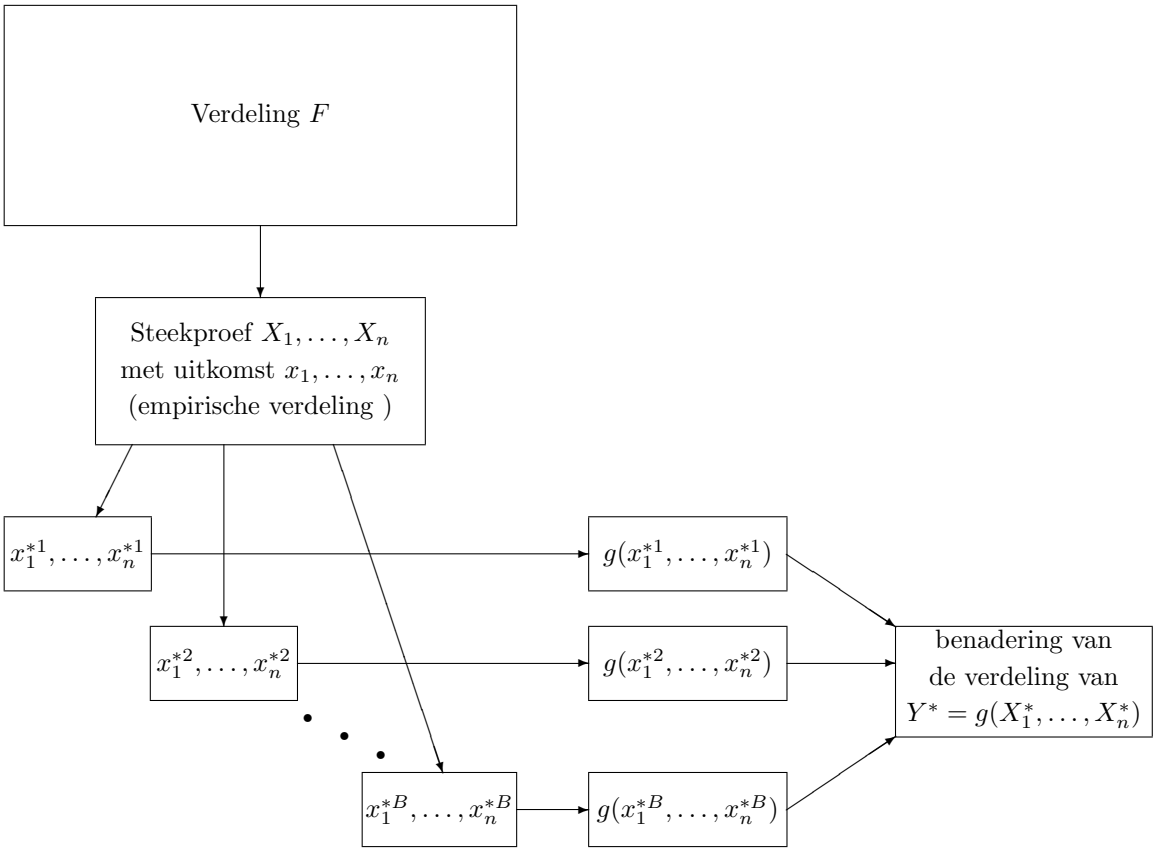
We nemen een aselecte steekproef $X_1, \ldots, X_n$ uit F met uitkomst $x_1, \ldots, x_n$.
Deze bepaalt de empirische verdeling .

We willen de verdeling van de steekproeffunctie $Y = g(X_1, \ldots, X_n)$
bestuderen.

We kunnen de verwachting en de variantie van Y als volgt schatten:

- Trek uit $\{x_1, \ldots, x_n\}$ (met teruglegging) een aselecte steekproef van n stuks met uitkomst $x_1^{*1}, \ldots, x_n^{*1}$ en bepaal $g(x_1^{*1}, \ldots, x_n^{*1})$.
- Voer ditzelfde proces B keer (onafhankelijk van elkaar) uit.
  We krijgen uitkomsten $g(x_1^{*1}, \ldots, x_n^{*1}), \ldots, g(x_1^{*B}, \ldots, x_n^{*B})$.
- Deze uitkomsten samen geven een benadering van de verdeling van $Y^* = g(X_1^*, \ldots, X_n^*)$, waarbij $X_1^*, \ldots, X_n^*$ een aselecte steekproef is uit de empirische verdeling .

Schatter voor Y is: $\bar{Y}^* = \frac{1}{B} \sum Y^{*b}$ met standaardfout $[\frac{1}{B-1} \sum (Y^{*b} - \bar{Y}^*)^2]^{\frac{1}{2}}$.



Stel je wilt de parameter $\theta = \theta(F)$ schatten. Daartoe gebruik je de functie $t(X_1, \ldots, X_n)$. Vaak zul je als schatting de overeenkomstige functie van de empirische verdelingsfunctie nemen: $t(x_1, \ldots, x_n) = \hat{\theta} = \theta(\hat{F})$. Efron noemt dit het 'plug-in principle'.
Welke schatting je ook neemt, je kunt de verdeling van $t(X_1, \ldots, X_n)$ benaderen door B steekproeven $x_1^{*b}, \ldots, x_n^{*b}$ te nemen uit $x_1, \ldots, x_n$ en de empirische verdeling van $t(x_1^*, \ldots, x_n^*)$ te bepalen. Analoog kun je de verdeling van $t(X_1, \ldots, X_n) - \theta$ benaderen met behulp van de empirische verdeling van $t(x_1^*, \ldots, x_n^*) - \hat{\theta}$. Laat a het 0.025-kwantiel van deze verdeling zijn en b het 0.0975-kwantiel. Dan geldt dus:

$$P(a < t(x_1^*, \ldots, x_n^*) - \hat{\theta} < b) = 0.95$$

en dus ook:

$$P(t(x_1^*, \ldots, x_n^*) - b < \hat{\theta} < t(x_1^*, \ldots, x_n^*) - a) = 0.95.$$

Het interval $[t(x_1, \ldots, x_n) - b, t(x_1, \ldots, x_n) - a] = [\hat{\theta} - b, \hat{\theta} - a]$ beschouwen we nu als een 95%-betrouwbaarheidsinterval voor θ.

Literatuur: Bradley Efron, Robert J. Tibshirani [1993]: An Introduction to the Bootstrap, Chapman & Hall.