Bias und Biaskorrektur

Bei einem erwartungstreuen Parameterschätzer $\hat{\theta}$ ist der Erwartungswert der Schätzfunktion $E(\hat{\Theta})$ gleich dem wirklichen Parameter θ. Somit ist die Differenz $E(\hat{\Theta}) - \theta = 0$.

BIAS

Ist ein Schätzer nicht erwartungstreu, so ist entweder $E(\hat{\Theta}) > \theta$ oder aber $E(\hat{\Theta}) < \theta$.

Ist $E(\hat{\Theta}) > \theta$, also $E(\hat{\Theta}) - \theta > 0$, dann wird der Parameter tendenziell *überschätzt*, man spricht von einem *positiven Bias*. *(Vorsicht: Einzelne Schätzwerte können auch in diesem Fall zu klein geraten, im Durchschnitt kommt es aber zur Überschätzung.)*

Analog bedeutet ein *negativer Bias*, dass der Parameter systematisch *unterschätzt* wird ($E(\hat{\Theta}) < \theta \quad \Leftrightarrow \quad E(\hat{\Theta}) - \theta < 0$).

Ein Beispiel zur Veranschaulichung

$\eta = 100$ sei ein bestimmter Parameter der Gesamtpopulation, $\hat{\lambda}$, $\hat{\varepsilon}$ und $\hat{\varphi}$ seien drei verschiedene Schätzfunktionen für η.

Ich ziehe nun 10 unabhängige Stichproben und ermittle für jede von ihnen mit allen drei Methoden einen Schätzwert für η:

i	$\hat{\lambda}_i$	$\hat{\varepsilon}_i$	$\hat{\varphi}_i$
1	104	81	134
2	97	102	142
3	93	74	120
4	99	88	151
5	108	78	138
6	101	112	144
7	102	93	98
8	98	72	123
9	100	79	139
10	104	94	128
$\sum_i$	1006	873	1317

Es ergeben sich folgende durchschnittliche Schätzwerte:

$$\bar{\hat{\lambda}} = 100.6$$
$$\bar{\hat{\varepsilon}} = 87.3$$
$$\bar{\hat{\varphi}} = 131.7$$

Die Schätzfunktion $\hat{\lambda}$ schätzt den Parameter η offenbar am besten. Die Abweichungen sind vermutlich nicht überzufällig, also nicht groß genug, um statistisch aussagekräftig („signifikant“) zu sein.

Im Gegensatz dazu wirft $\hat{\varepsilon}$ tendenziell Werte aus, die η eindeutig (signifikant) unterschätzen ($\rightarrow$ negativer Bias), während $\hat{\varphi}$ vermutlich einen positiven Bias aufweist.

Bias der Stichprobenvarianz:

Ein Beispiel für einen nicht erwartungstreuen Schätzer ist die Stichprobenvarianz s_X^2. Dies kann man zeigen, indem man ihren Erwartungswert umformt und mit der Populationsvarianz σ_X^2 vergleicht.

Wir ziehen immer wieder unabhängige Stichproben und berechnen jeweils $\bar{x}$ und s_X^2, woraus sich die Zufallsvariablen $\bar{X}$ und S_X^2 ergeben.

Wir wissen:

$$s_X^2 = \frac{1}{n} \cdot \sum_i x_i^2 - \bar{x}^2$$

Für die entsprechenden Zufallsvariablen gilt demnach:

$$S_X^2 = \frac{1}{n} \cdot \sum_i X_i^2 - \bar{X}^2$$

Nun berechnen wir auf beiden Seiten den Erwartungswert:

$$E(S_X^2) = E(\frac{1}{n} \cdot \sum_i X_i^2 - \bar{X}^2)$$

$$E(S_X^2) = \frac{1}{n} \cdot \sum_i E(X_i^2) - E(\bar{X}^2)$$

Einschub 1:

Da alle X_i dieselbe Verteilung wie die Variable X haben, gilt:

$$\begin{aligned}\frac{1}{n}\cdot\sum_i E(X_i^2) &= \frac{1}{n}\cdot\sum_i E(X^2) = \\ &= \frac{n}{n}\cdot E(X^2) = \\ &= E(X^2)\end{aligned}$$

Ende Einschub 1

Einschub 2:

$$\begin{aligned}E(\bar{X}^2) &= E[(\frac{1}{n}\cdot\sum_i X_i)^2] = \\ &= E[(\frac{1}{n})^2\cdot(\sum_i X_i)^2] = \\ &= \frac{1}{n^2}\cdot E[(\sum_i X_i)^2] = \\ &= \frac{1}{n^2}\cdot E[(\sum_j X_j)\cdot(\sum_l X_l)] = \\ &= \frac{1}{n^2}\cdot E[\sum_j\sum_l (X_j\cdot X_l)] =\end{aligned}$$

Wir ziehen nun aus der Doppelsumme jene Produkte heraus, bei denen der Index gleich ist *(der Einfachheit halber schreiben wir für die Doppelsumme nur noch ein einzelnes Summenzeichen)*:

$$\begin{aligned}&= \frac{1}{n^2}\cdot E[\sum_i X_i^2 + \sum_{j\neq l}(X_j\cdot X_l)] = \\ &= \frac{1}{n^2}\cdot E(\sum_i X_i^2) + \frac{1}{n^2}\cdot E[\sum_{j\neq l}(X_j\cdot X_l)] = \\ &= \frac{1}{n^2}\cdot\sum_i E(X_i^2) + \frac{1}{n^2}\cdot[\sum_{j\neq l} E(X_j\cdot X_l)] =\end{aligned}$$

Da je zwei Variablen X_j und X_l unabhängig sind, gilt für die Erwartungswerte $E(X_j \cdot X_l) = E(X_j) \cdot E(X_l)$:

$$= \frac{1}{n^2} \cdot [n \cdot E(X^2)] + \frac{1}{n^2} \cdot \sum_{j \neq l} [E(X_j) \cdot E(X_l)] =$$

Alle X_j und X_l sind wie X verteilt, somit gilt $E(X_j) = E(X_l) = E(X)$:

$$= \frac{1}{n} \cdot E(X^2) + \frac{1}{n^2} \cdot \sum_{j \neq l} [E(X)]^2 =$$

Insgesamt gäbe es $n \cdot n$ Kombinationen von j und l, wir haben aber alle n gleichen Index-Paare herausgezogen, somit bleiben $n \cdot n - n = n \cdot (n-1)$.

$$= \frac{1}{n} \cdot E(X^2) + \frac{1}{n^2} \cdot n \cdot (n-1) \cdot [E(X)]^2$$

$$E(\overline{X}^2) = \frac{1}{n} \cdot E(X^2) + \frac{n-1}{n} \cdot [E(X)]^2$$

Ende Einschub 2

Zurück zur Gleichung:

$$E(S_X^2) = \frac{1}{n} \cdot \sum_i E(X_i^2) - E(\overline{X}^2)$$

Wir setzen unsere Ergebnisse aus *Einschub 1* und *Einschub 2* ein:

$$E(S_X^2) = E(X^2) - \frac{1}{n} \cdot E(X^2) - \frac{n-1}{n} \cdot [E(X)]^2$$

$$E(S_X^2) = \frac{n}{n} \cdot E(X^2) - \frac{1}{n} \cdot E(X^2) - \frac{n-1}{n} \cdot [E(X)]^2$$

$$E(S_X^2) = \frac{n-1}{n} \cdot E(X^2) - \frac{n-1}{n} \cdot [E(X)]^2$$

$$E(S_X^2) = \frac{n-1}{n} \cdot [E(X^2) - [E(X)]^2]$$

Zum Vergleich:

$$\sigma_X^2 = E(X^2) - [E(X)]^2$$

also:

$$E(S_X^2) = \frac{n-1}{n} \cdot \sigma_X^2$$

Wir sehen, dass die Schätzer durchschnittlich kleiner sind als der zu schätzende Parameter. Die Stichprobenvarianz als Schätzfunktion der Populationsvarianz hat also einen negativen Bias.

BIASKORREKTUR

Natürlich ist man daran interessiert, derartige systematische Fehler auszumerzen. Ein Schätzer, der stets zu kleine oder zu große Werte liefert, ist wenig bis gar nicht brauchbar. Man versucht, eine derartige Verzerrung formal in den Griff zu bekommen, um die Schätzfunktion korrigieren zu können.

Dies lässt sich sehr gut am (oben errechneten) Bias der Stichprobenvarianz demonstrieren.

Wir haben gezeigt:

$$E(S_X^2) = \frac{n-1}{n} \cdot \sigma_X^2$$

Um auf die gesuchte Populationsvarianz zu kommen, genügt eine kleine Umformung:

$$\sigma_X^2 = \frac{n}{n-1} \cdot E(S_X^2)$$

Wir können also die verzerrte Schätzung korrigieren, indem wir $s_X^2 = \frac{1}{n} \cdot \sum_i (x_i - \bar{x})^2$ mit $\frac{n}{n-1}$ multiplizieren. So erhalten wir folgende erwartungstreue Schätzfunktion für σ_X^2:

$$s_{korr}^2 = \frac{n}{n-1} \cdot s_X^2$$

$$s_{korr}^2 = \frac{1}{n-1} \cdot \sum_i (x_i - \bar{x})^2$$