



## zu 25. Warum ist bei der empirischen Stichprobenvarianz $(n - 1)$ und nicht $n$ im Nenner?

Bekanntlich ist die empirische Stichprobenvarianz durch

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (1)$$

mit

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

gegeben. Häufig wird argumentiert, der Nenner  $(n - 1)$  berücksichtigt die Tatsache, dass man sonst eine empirische Varianz aus einer Stichprobe vom Umfang  $n = 1$  mit dem Ergebnis  $s = 0$  berechnen könnte. Dieses “Handwaving”-Argument stellen wir nun auf eine solide statistische Basis, indem wir fordern, dass unsere Schätzgröße  $S^2$  **erwartungstreu** zur tatsächlichen Varianz  $\sigma^2$  sein soll, also

$$E(S^2) \stackrel{!}{=} \sigma^2.$$

Zunächst ”erweitern” wir die Summanden mit dem “echten” Erwartungswert  $\mu = E(X_i)$  (*warum* dies sinnvoll sein wird, können Sie vielleicht noch nicht sehen, diese “Weitsicht” wird aber in der Klausur auch nicht verlangt ..;-), dann multiplizieren wir aus, benutzen die Linearität des Erwartungswertes und schreiben alle in  $S^2$  vorkommenden Summen, *einschließlich der in  $\bar{X}$  vorkommenden Summen*, explizit hin:

$$\begin{aligned}
(n-1)E(S^2) &= E\left(\sum_{i=1}^n (X_i - \bar{X})^2\right) \\
&= E\left(\sum_{i=1}^n ((X_i - \mu) + (\mu - \bar{X}))^2\right) \\
&= \sum_{i=1}^n E(X_i - \mu)^2 + \sum_{i=1}^n E(\mu - \bar{X})^2 + 2E\left(\sum_{i=1}^n (X_i - \mu)(\mu - \bar{X})\right) \\
&= n\sigma^2 + n\sigma_{\bar{X}}^2 - 2E\left(\sum_{i=1}^n (X_i - \mu)\left(\frac{1}{n}\sum_{j=1}^n (X_j - \mu)\right)\right) \\
&= n\sigma^2 + n\frac{\sigma^2}{n} - \frac{2}{n}\sum_{i=1}^n \sum_{j=1}^n E((X_i - \mu)(X_j - \mu)) \\
&= (n+1)\sigma^2 - \frac{2}{n}n\sigma^2 \\
&= \underline{\underline{(n-1)\sigma^2}}.
\end{aligned}$$

Die Varianz des Mittelwertes

$$\sigma_{\bar{X}}^2 = E(\bar{X} - E(\bar{X}))^2 = E(\bar{X} - \mu)^2 = \frac{1}{n^2}E(X_i - \mu)^2 = \frac{1}{n}\sigma^2$$

wird nicht nur hier, sondern auch an vielen anderen Stellen der induktiven Statistik gebraucht.

Zur Auswertung des letzten Terms berücksichtigten wir, dass die Stichprobenvariablen  $X_i$  nicht nur *identisch verteilt* (mit Varianz  $\sigma_i^2 = \sigma^2$ ), sondern auch *unabhängig* sind und damit ihr Korrelationskoeffizient

$$r_{ij} = r(X_i, X_j) = \frac{E(X_i - \mu)(X_j - \mu)}{\sigma_i \sigma_j} = \frac{E(X_i - \mu)(X_j - \mu)}{\sigma^2} = 0$$

für  $i \neq j$ . Damit hat die obige Doppelsumme nur Beiträge für  $i = j$  und wird damit zur Einfachsumme:

$$\frac{2}{n}\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu)(X_j - \mu) = \frac{2}{n}\sum_{i=1}^n ((X_i - \mu)^2) = \frac{2n}{n}\sigma^2.$$

Damit ist gezeigt, dass die durch (1) definierte Stichprobenvarianz, *nicht* jedoch die empirische Varianz der deskriptiven Statistik mit  $n$  im Nenner erwartungstreu ist!