Vorlesung 6a

Varianz und Kovarianz

Teil 5

Die Varianz der hypergeometrischen Verteilung

(Buch S. 32 und S. 61)

Ein Beispiel für die Anwendung der Formel

$$\mathbf{Var}[Z_1 + \cdots + Z_n] =$$

$$\mathbf{Var}\, Z_1 + \cdots + \mathbf{Var}\, Z_n + 2 \sum_{i<j} \mathbf{Cov}[Z_i, Z_j]$$

liefert das Zählen der “Erfolge”

beim Ziehen ohne Zurücklegen.

In einer Urne sind r rote und b blaue Kugeln.

Es wird n-mal ohne Zurücklegen gezogen.

$X :=$ Anzahl der gezogenen roten Kugeln.

$$\mathbf{Var}[X] =?$$

Zur Erinnerung:

Mit $g := r + b$ ist

$$\mathbf{P}(X = k) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{g}{n}}, \quad k = 0, \ldots, r.$$

X heißt hypergeometrisch verteilt mit Parametern n, g und r.

Erwartungswert und Varianz kann man direkt

über die Verteilungsgewichte ausrechnen (siehe Buch S. 32).

Es geht auch eleganter (vgl Buch S. 50/51):

Wir betrachten dazu die Zufallsvariable Z_i, die

. . . den Wert 1 annimmt, falls die i-te gezogene Kugel rot ist,

. . . und sonst den Wert 0.

Man sagt dafür auch:

Z_i ist die *Indikatorvariable*

(kurz: der *Indikator*)

des Ereignisses $\{i$-te gezogene Kugel rot$\}$.

$$X := Z_1 + \cdots + Z_n$$

$$\mathbf{E}[Z_i] = \mathbf{P}(Z_i = 1) = p, \quad \text{mit}$$

$p := \frac{r}{g}$ der Anteil der roten Kugeln in der Urne.

Also: $\mathbf{E}[X] = np$

(vgl. V3a4)

Und wie stehts mit der Varianz von X?

$$X := Z_1 + \cdots + Z_n$$

$$\mathbf{Var}\, X = \mathbf{Var}\, Z_1 + \cdots + \mathbf{Var}\, Z_n + 2 \sum_{1 \le i < j \le n} \mathbf{Cov}[Z_i, Z_j]$$

Sei $g = r + b$ die Gesamtanzahl der Kugeln,

$p := \frac{r}{g}$ der Anteil der roten Kugeln in der Urne,

$$q := 1 - p.$$

$$\mathbf{Var}\, Z_i = pq.$$

$$\mathbf{Cov}[Z_i, Z_j] = ?$$

Ein eleganter Weg zur Berechnung von $\mathbf{Cov}[Z_i, Z_j]$:

Wir ziehen in Gedanken, bis die Urne leer ist

(d.h. wir setzen $n = g$.)

Wir ziehen in Gedanken, bis die Urne leer ist.

Dann ist

$$Z_1 + \cdots + Z_g = r,$$

also

$$\mathbf{Var}[Z_1 + \cdots + Z_g] = 0.$$

$$0 = \mathbf{Var}\, Z_1 + \cdots + \mathbf{Var}\, Z_g + 2 \sum_{1 \le i < j \le g} \mathbf{Cov}[Z_i, Z_j], \quad \text{d.h.}$$

$$0 = gpq + g(g-1)\mathbf{Cov}[Z_1, Z_2], \quad \text{d.h.}$$

$$\mathbf{Cov}[Z_1, Z_2] = -\frac{1}{g-1} pq$$

$$X = Z_1 + \cdots + Z_n$$

$$\mathbf{Var}\, X = \mathbf{Var}\, Z_1 + \cdots + \mathbf{Var}\, Z_n + 2 \sum_{1 \le i < j \le n} \mathbf{Cov}[Z_i, Z_j]$$

$$= n\mathbf{Var}\, Z_1 + n(n-1)\mathbf{Cov}[Z_1, Z_2]$$

$$= npq - n(n-1)\frac{1}{g-1}pq$$

$$= npq\left(1 - \frac{n-1}{g-1}\right) = npq\,\frac{g-n}{g-1}. \quad \square$$

Fazit:

Die Varianz von Hyp(n, g, pg) ist

$$npq\,\frac{g-n}{g-1}.$$

Zusammenfassung

des Wichtigsten aus V6a

$$\mathbf{Var}[X] := \mathbf{E}[(X - \mu)^2]$$

$$\mathbf{Var}[X + Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}[X, Y]$$

Die Varianz einer Summe von unkorrelierten ZV'en

ist gleich der Summe der Varianzen,

die Varianz einer Summe von negativ korrelierten ZV'en

ist kleiner als die Summe der Varianzen.

Die Varianz von Bin(n, p) ist npq.

Die Varianz von Hyp (n, g, pg) ist $npq\,\dfrac{g-n}{g-1}$. Die Varianz einer Poisson(λ)-verteilten Zufallsvariablen

ist so groß wie ihr Erwartungswert,

nämlich λ.

Ungleichung von Chebyshev:

$$\mathbf{P}(|Y-\mu| \geq \varepsilon\,\sigma_Y)) \leq \frac{1}{\varepsilon^2}$$

$$\begin{aligned}\mathbf{Cov}[X, Y] &:= \mathbf{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]\end{aligned}$$

Speziell für Indikatorvariable:

$$\begin{aligned}&\mathbf{Cov}[I_{E_1}, I_{E_2}] \\ &= \mathbf{P}(E_1 \cap E_2) - \mathbf{P}(E_1)\mathbf{P}(E_2).\end{aligned}$$