10. Vorlesung: Kovarianz und Korrelation

Nikolas Tapia

23. Mai 2024, Stochastik für Informatik(er)

Tabelle der Verteilungen

Verteilung	Erwartungswert	Varianz
Uniform(n)	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	λ	λ
Zipf(a)	$\frac{\zeta(a-1)}{\zeta(a)}, \quad a>2$	$\frac{\zeta(2-a)\zeta(a)-\zeta(1-a)^2}{\zeta(a)^2}, \quad a>3$

Bedingter Erwartungswert und Voraussage

Aussage 10.1

Seien X, Y Zufallsvariablen und $f: \mathbb{R} \to \mathbb{R}$ eine Funktion. Dann gilt

$$\mathbb{E}[(Y - f(X))^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2]. \quad Y \approx f(X),$$

$$\sum_{x,y} (y - f(x))^2 \, \mathbb{P}(X = x, Y = y)$$

$$\sum_{x,y} (y - \underbrace{\mathbb{E}[Y \mid X = x]}_{\stackrel{\text{def.}}{=} \sum_{y' \in Y(\Omega)} y' \, \mathbb{P}(Y = y', X = x) =: g(x)})^2 \, \mathbb{P}(X = x, Y = y)$$

Leibniz Gemeinschaft

Bedingte Varianz

Definition 10.1

Seien X, Y diskrete Zufallsvariablen. Die **bedingte Varianz** von X gegeben $Y = y$ ist definiert als

$$\mathbb{V}(X|Y = y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y = y],$$

sofern $\mathbb{P}(Y = y) > 0$. $\mathbb{V}(X|Y) \to ZV$

Aussage 10.2

Formel der bedingten Varianz Seien X, Y diskrete Zufallsvariablen. Dann gilt

$$\mathbb{V}(X) = \mathbb{E}[\mathbb{V}(X|Y)] + \mathbb{V}(\mathbb{E}[X|Y]).$$

Bew: Schreibe $\mathbb{V}(X|Y=y) = \mathbb{E}[X^2|Y=y] - \mathbb{E}[X|Y=y]^2$

(zur Erinnerung $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$)

$$\mathbb{V}(X \mid Y=y) = \mathbb{E}[X^2 \mid Y=y] - \mathbb{E}[X \mid Y=y]^2$$

$$\mathbb{E}[\mathbb{V}(X \mid Y)] = \mathbb{E}[X^2] - \underline{\mathbb{E}[\mathbb{E}[X \mid Y]^2]} \quad (1)$$

Formel der totalen $\mathcal{E}$ ↑

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$$

$$\mathbb{V}(\mathbb{E}[X \mid Y]) = \underline{\mathbb{E}[\mathbb{E}[X \mid Y]^2]} - \underbrace{\mathbb{E}[\mathbb{E}[X \mid Y]]^2}_{= \mathbb{E}[X]^2} \quad (2)$$

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

$$= \mathbb{V}(X)$$

Ausgabe von Kunde i

$X_1, \dots, X_n, X_{n+1}, \dots$

Beim Spiel von einem Geschäft: Was trägt zur Streuung der Gesamtausgaben an einem bestimmten Tag bei?

$\approx 400€$

$$S = \sum_{i=1}^{N} X_i \quad \text{Gesamtausgabe}$$

$$\mathbb{E}[S] = \underset{50}{\mathbb{E}[N]}\,\underset{8}{\mathbb{E}[X]} = 400$$

$$\mathbb{E}[S|N] = N\,\mathbb{E}[X]$$

$$\mathbb{V}(S) = ? \quad \mathbb{V}(S) = \mathbb{E}[\mathbb{V}(S|N)] + \mathbb{V}(\mathbb{E}[S|N]) = \begin{cases} \mathbb{E}[N]\mathbb{V}(X) \\ + \mathbb{E}[X]^2\,\mathbb{V}(N) \end{cases}$$

$$\mathbb{V}(S|N) \underset{\text{unabhängigkeit der } X_i}{=} N\,\mathbb{V}(X), \quad \mathbb{V}(\mathbb{E}[S|N]) = \mathbb{E}[X]^2\,\mathbb{V}(N)$$

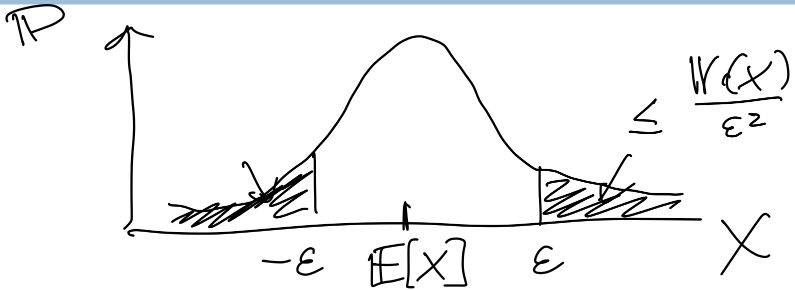
Beispiel

Chebyshev-Ungleichung

Aussage 10.3

Sei X eine Zufallsvariable, deren Erwartungswert und Varianz existieren. Dann gilt für jedes $\varepsilon > 0$

$$\mathbb{P}(|X-\mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}(X)}{\varepsilon^2}. \xrightarrow[\varepsilon \to \infty]{} 0$$



Markov-Ungleichung

Aussage 10.4

Sei X eine Zufallsvariable, und sei $f\colon \mathbb{R} \to [0,\infty)$ eine monoton erwachsende Funktion. Falls $\mathbb{E}[f(X)]$ existiert, so gilt

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[f(X)]}{f(a)}.$$

Die Chebyshev-Ungleichung ist ein Spezialfall.

$$f(x) = |x - \mathbb{E}[X]|^2$$

Coupon Collector

Ein Sammler sammelt n verschiedene Coupons. Diese werden in Verpackungen verkauft, die jeweils einen zufälligen Coupon enthalten. Wie viele Verpackungen muss der Sammler im Durchschnitt kaufen, um alle n Coupons zu erhalten?

1. Y_k = # Packung, um ein neues Bild zu erhalten, falls wir k verschiedene Bilder besitzen.

$$Y_k \sim \text{Geom}\left(\frac{n-k}{n}\right), \quad k=0,\dots,n-1$$

2. X = # Packung, um alle Bilder zu erhalten.

$$X = Y_0 + Y_1 + \dots + Y_{n-1}$$

3. $$\mathbb{E}[X] = \mathbb{E}[Y_0] + \mathbb{E}[Y_1] + \dots + \mathbb{E}[Y_{n-1}]$$

$$= \frac{n}{n-0} + \frac{n}{n-1} + \dots \frac{n}{n-(n-1)} = \sum_{k=0}^{n-1} \frac{n}{n-k}$$

Coupon Collector

$$\mathbb{E}[X] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \sum_{k=1}^{n} \frac{1}{k} \approx n \int_1^n \frac{1}{t}\,dt = n \log n$$

$$\mathbb{V}(X) = \sum_{k=0}^{n-1} \mathbb{V}(Y_k) = \sum_{k=0}^{n-1} \frac{1-\left(\frac{n-k}{n}\right)}{\left(\frac{n-k}{n}\right)^2} \rightarrow 1-\frac{n-k}{n} = \frac{k}{n}$$

$$= \sum_{k=0}^{n-1} \frac{nk}{(n-k)^2}$$

Analysis

$k \leq n$

$$\leq n^2 \sum_{k=1}^{n-1} \frac{1}{k^2} \approx n^2 \frac{\pi^2}{6}$$

Coupon Collector

Wie wahrscheinlich ist es, dass wir mehr als $2n\log n$ Packung kaufen müssen? D.h.

$$\mathbb{P}(X \geq 2\mathbb{E}[X]) = \mathbb{P}(X - \mathbb{E}[X] \geq \overset{n \log n}{\mathbb{E}[X]})$$

$$\cong \mathbb{P}(X - \mathbb{E}[X] \geq n\log n)$$

$$\leq \frac{\mathbb{V}(X)}{(n\log n)^2} = \frac{n^2\pi^2/6}{n^2\log n} = c\,\frac{1}{\log n} \to 0$$

Kovarianz

Definition 10.2

Seien X, Y Zufallsvariablen. Die **Kovarianz** von X und Y ist definiert als

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

$$\left(\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]\right)$$

Aussage 10.5

Seien X, Y Zufallsvariablen. Dann gilt

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Bew: $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$

$= \mathbb{E}\left[XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\right)$

$= \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]$

Kovarianz

Definition 10.2

Seien X, Y Zufallsvariablen. Die **Kovarianz** von X und Y ist definiert als

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

Aussage 10.5

Seien X, Y Zufallsvariablen. Dann gilt

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Es gibt 3 Fälle:

- $\mathrm{Cov}(X,Y) > 0$: positiv korreliert.
- $\mathrm{Cov}(X,Y) = 0$: unkorreliert
- $\mathrm{Cov}(X,Y) < 0$: negativ korreliert

Aussage 10.6

Seien X, Y, Z Zufallsvariablen und seien $a, b \in \mathbb{R}$. Dann gilt

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$,
2. $\mathrm{Cov}(X, X) = \mathbb{V}(X)$,
3. $\mathrm{Cov}(aX + bY, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$,
4. Falls X und Y unabhängig sind, dann ist $\mathrm{Cov}(X, Y) = 0$.

Unabhängig $\Rightarrow$ unkorreliert

$\not\Leftarrow$

Unabhängigkeit und Unkorreliertheit

Beispiel

Seien X, Y Zufallsvariablen mit folgende gemeinsame Verteilung:

Y\X	-1	0	1	$\mathbb{P}(Y = y)$
0	1/4	0	1/4	1/2
1	0	1/2	0	1/2
$\mathbb{P}(X = x)$	1/4	1/2	1/4	1

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \underbrace{\cancel{\mathbb{E}[X]\mathbb{E}[Y]}}_{=0}$$

$$\mathbb{E}[X] = 0 = -1 \times \frac{1}{4} + 1 \times \frac{1}{4}$$

$$\mathbb{E}[XY] = 0 \Rightarrow \mathrm{Cov}(X,Y) = 0.$$

Aber $\mathbb{P}(X=0, Y=0) = 0$

$$\neq \mathbb{P}(X=0)\mathbb{P}(Y=0) = \frac{1}{4}$$

$\Rightarrow X, Y$ nicht unabhängig

Aussage 10.7

Seien X, Y Zufallsvariablen. Dann gilt $\left(\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) \quad X, Y \text{ unab.}\right)$

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathrm{Cov}(X,Y).$$

Bew:

$$\begin{aligned}\mathbb{V}(X+Y) &= \mathbb{E}\left[(X+Y)^2\right] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}\left[X^2 + 2XY + Y^2\right] - \left(\mathbb{E}[X] + \mathbb{E}[Y]\right)^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \mathbb{V}(X) + \mathbb{V}(Y) + 2\,\mathrm{Cov}(X,Y).\end{aligned}$$

Korrelation

Definition 10.3

Seien X, Y Zufallsvariablen mit positiver Varianz. Die **Korrelation** von X und Y ist definiert als

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}.$$

Schreibt man auch $\rho(X, Y) = \mathrm{Corr}(X, Y)$.

Aussage 10.8

Seien X, Y Zufallsvariablen mit positiver Varianz. Dann gilt

1. $-1 \leq \mathrm{Corr}(X, Y) \leq 1$,
2. Das Vorzeichen der Korrelation stimmt mit dem Vorzeichen der Kovarianz überein.

Korrelation und Kausalität

“(mittellat. correlatio für ‘Wechselbeziehung’) beschreibt eine Beziehung zwischen zwei oder mehreren Merkmalen, Ereignissen, Zuständen oder Funktionen” (Wikipedia). **Das eine steht in Beziehung zum anderen, bedingt es aber nicht zwingend.**

“(lat. causa ‘Ursache’) ist die Beziehung zwischen Ursache und Wirkung oder ‘Aktion’ und ‘Reaktion’, betrifft also die Abfolge aufeinander bezogener Ereignisse und Zustände” (Wikipedia). **Das eine verursacht das andere.**

Korrelation zwischen dem Rückgang der Storchenpopulation und der Abnahme der Geburtenzahl in Baden-Württemberg

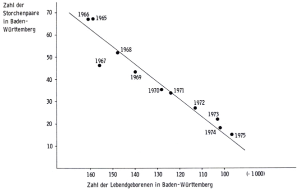


Abbildung aus der Monographie „Kontrazeption mit Hormonen“
Von Prof. Dr. Hans-Dieter Taubert und Prof. Dr. Herbert Kuhl (Georg Thieme Verlag, Stuttgart 1981)