Zufallsvariablen

1 Zusammenhang: Wirklichkeit–Modell

Wirklichkeit	Modell
Stichprobe	Population
Daten	Zufallsvariable
diskret	diskret
stetig	stetig
rel. Häufigkeit	Wahrscheinlichkeit
Häufigkeitstabelle	Wahrscheinlichkeitsverteilung
Stabdiagramm	Stabdiagramm
Histogramm	Dichte
empir. Verteilungsfunktion	Verteilungsfunktion
empir. Kennzahlen	theoret. Kennzahlen
Mittelwert $\bar{x}$	Erwartungswert $E(X)$
Varianz s^2	Varianz $Var(X)$

Eine *Zufallsvariable* (random variable) ist eine quantitative Variable, deren Wert durch das zufälllige Ergebnis von Experimenten oder Beobachtungen bestimmt wird. Zufallsvariablen bilden ein Modell für die beobachteten Grössen, die Daten.
Es gibt *diskrete* (discrete) und *stetige* (continuous) Zufallsvariablen. Diskrete Zufallsvariablen haben nur eine endliche oder abzählbare Anzahl möglicher Werte, stetige Zufallsvariablen können alle Werte innerhalb eines Intervalls der reellen Zahlen annehmen.

2 Diskrete Zufallsvariablen

Die *Wahrscheinlichkeitsverteilung* p (probability function) ist definiert durch:

X	x_1	x_2	...	x_k
p	$p(x_1)$	$p(x_2)$	...	$p(x_k)$

$p(x_i) := P(X = x_i), \quad \sum p(x_i) = 1.$

Die *kumulative Verteilungsfunktion* F (cumulative distribution function) ist definiert durch:

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

F ist monoton wachsend, $\lim_{x \to -\infty} F(x) = 0$ und $\lim_{x \to \infty} F(x) = 1$.

Uniforme Verteilung

Eine Verteilung, bei der alle Werte die gleiche Wahrscheinlichkeit haben, heisst *uniform.*

X	x_1	x_2	...	x_n
p	$1/n$	$1/n$	...	$1/n$

Bernoulli Verteilung

Ein Experiment habe zwei mögliche Ausgänge, “Erfolg” und “Misserfolg”, mit den Wahrscheinlichkeiten p und $1 - p$.

$$X = \begin{cases} 0 & : \text{“Misserfolg”} \\ 1 & : \text{“Erfolg”} \end{cases}$$

X	0	1
p(x)	1-p	p

Binomialverteilung
Es werden n voneinander unabhängige Versuche gemacht. Jeder einzelne Versuch hat zwei mögliche Ausgänge, „Erfolg" und „Misserfolg". Die Wahrscheinlichkeit p für einen Erfolg ist konstant. Die Anzahl Erfolge X hat dann eine Binomialverteilung $\mathcal{B}(n, p)$ und die Wahrscheinlichkeit für k Erfolge ist gegeben durch:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{für } k = 0, 1, \ldots, n.$$

Geometrische Verteilung
Es werden unabhängige Versuche durchgeführt. Jeder einzelne Versuch hat zwei mögliche Ausgänge, „Erfolg" und „Misserfolg". Die Wahrscheinlichkeit p für einen Erfolg ist konstant. X ist die Anzahl Versuche bis und mit dem ersten Erfolg.

$$P(X = k) = (1-p)^{k-1} p \qquad \text{für } k = 1, 2, 3, \ldots.$$

Negative Binomialverteilung
Es werden unabhängige Versuche durchgeführt. Jeder einzelne Versuch hat zwei mögliche Ausgänge, „Erfolg" und „Misserfolg". Die Wahrscheinlichkeit p für einen Erfolg ist konstant. X ist die Anzahl Misserfolge bis r Erfolge eingetreten sind.

$$P(X = k) = \binom{k + r - 1}{k} p^r (1-p)^k \qquad \text{für } k = 0, 1, 2, \ldots.$$

Poissonverteilung
Betrachte die absolute Häufigkeit, mit der ein bestimmtes Ereignis eintritt. Wenn die Ereignisse unabhängig voneinander mit einer konstanten Rate λ pro Zeiteinheit passieren, dann hat die Anzahl Ereignisse X eine Poissonverteilung $\mathcal{P}(\lambda)$. Die Wahrscheinlichkeit für k Ereignisse pro Zeiteinheit ist:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad k = 0, 1, 2, \ldots$$

Für n gross und p klein ist die Poissonverteilung eine Näherung für die Binomialverteilung mit $\lambda = np$.
Ein Prozess, der innerhalb eines festen zeitlichen oder räumlichen Intervalls eine Anzahl Ereignisse erzeugt, die einer Poissonverteilung folgt, heisst *Poissonprozess*.

3 Stetige Zufallsvariablen

Die *Dichte* f (density) ist eine stückweise stetige Funktion mit $f(x) \geq 0$ und $\int_{-\infty}^{\infty} f(x)\,dx$. Wenn X eine stetige Zufallsvariable mit Dichte f ist, dann gilt:

$$P(a < X < b) = \int_a^b f(x)\,dx \qquad \text{für } a < b.$$

Die *kumulative Verteilungsfunktion* F (cumulative distribution function) ist definiert durch:

$$F(x) = P(X \leq x), \qquad -\infty < x < \infty.$$

Es gilt:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Das $\alpha-$*Quantil* x_α ist definiert durch: $F(x_\alpha) = \alpha$. Für $\alpha = 1/2$ erhält man den *Median*, für $\alpha = 1/4$ und $\alpha = 3/4$ das 1. und das 3. *Quartil.*

Uniforme Verteilung

Die Dichte einer uniformverteilten Zufallsvariablen ist:

$$f(x) = \frac{1}{b-a} \qquad \text{für } a \leq x \leq b.$$

Die Verteilungsfunktion ist:

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b. \end{cases}$$

Exponentialverteilung

Modell für Warte- oder Ueberlebenszeiten. Wird in einem Poissonprozess mit Paramter λ statt der Anzahl Ereignisse in einem bestimmten Zeitintervall die Dauer bis zum Eintreten des nächsten Ereignisses betrachtet, so ist diese Dauer exponentialverteilt, $Exp(\lambda)$. Die Exponentialverteilung ist *gedächtnislos* (memoryless).
Die Dichte einer exponentialverteilten Zufallsvariablen ist:

$$f(x) = \lambda e^{-\lambda x} \qquad \text{für } x \geq 0, \qquad \lambda > 0.$$

Die Verteilungsfunktion ist:

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Gammaverteilung

Die Dichte einer gammaverteilten Zufallsvariablen ist:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \qquad \text{für } x \geq 0, \qquad \alpha > 0, \lambda > 0.$$

Die Gammafunktion $\Gamma(x)$ ist definiert durch: $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\,du \quad x > 0$.
Es gilt: $\Gamma(1) = 1, \quad \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ für $\alpha > 1, \quad \Gamma(\alpha) = (\alpha-1)!$, wenn α eine ganze Zahl grösser als 1 ist.

Normalverteilung

Die Normalverteilung ist das weitaus häufigste Modell für Messdaten. Entwickelt wurde sie als Modell für Messfehler, sie passt aber oft auch in andern Situationen recht gut. Das hat sich empirisch gezeigt und ein mathematisches Resultat, der *Zentrale Grenzwertsatz*, bestätigt das. Ein grosser Teil der statistischen Methoden setzt Normalverteilung voraus.
Die Dichte einer Zufallsvariablen mit Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ ist gegeben durch:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty < x < +\infty, \qquad -\infty < \mu < +\infty, \quad \sigma > 0.$$

Die spezielle Normalverteilung $\mathcal{N}(0,1)$ heisst *Standardnormalverteilung*. Für Dichte und kumulative Verteilungsfunktion verwendet man in diesem Fall die Bezeichnungen ϕ und Φ.

Chiquadrat-Verteilung
Seien $Z_1, \ldots, Z_n \sim \mathcal{N}(0,1)$, iid. Dann hat $X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ eine Chiquadrat-Verteilung mit n Freiheitsgraden, $X \sim \chi_n^2$.

***t*-Verteilung**
Seien $Z \sim \mathcal{N}(0,1)$ und $X \sim \chi_n^2$ unabhängige Zufallsvariablen. Dann hat $T = \dfrac{Z}{\sqrt{X/n}}$ eine t-Verteilung mit n Freiheitsgraden, $T \sim t_n$.

***F*-Verteilung**
Seien $X_1 \sim \chi_n^2$ und $X_2 \sim \chi_m^2$ unabhängige Zufallsvariablen. Dann hat $F = \dfrac{X_1/n}{X_2/m}$ eine F-Verteilung mit n und m Freiheitsgraden, $F \sim F_{n,m}$.

4 Erwartungswert und Varianz

Sei X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion p. Der *Erwartungswert* von X (expected value, mean) ist definiert als:

$$E(X) = \sum_i x_i p(x_i),$$

falls die Summe existiert. Oft wird der Erwartungswert mit μ bezeichnet.
Sei X eine stetige Zufallsvariable mit Dichte f. Der *Erwartungswert* $E(X)$ von X ist definiert als:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx,$$

falls das Integral existiert.

Sei X eine Zufallsvariable mit Erwartungswert $E(X)$. Dann ist die *Varianz* von X (variance) gegeben durch:

$$Var(X) = E\{[X - E(X)]^2\},$$

falls der Erwartungswert existiert. Die *Standardabweichung* von X (standard deviation) ist die Wurzel aus der Varianz. Oft wird die Varianz mit σ^2 und die Standardabweichung mit σ bezeichnet.
Wenn X diskret ist, gilt mit $E(X) = \mu$:

$$Var(X) = \sum_i (x_i - \mu)^2 p(x_i),$$

für X stetig:

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx.$$

Rechenregeln

Seien X_1 und X_2 Zufallsvariablen und a, b konstante Zahlen. Dann gilt:

$$\begin{aligned} E(X_1 + X_2) &= E(X_1) + E(X_2) \\ E(a + bX_1) &= a + bE(X_1) \\ Var(X_1 + X_2) &= Var(X_1) + Var(X_2) \text{ , falls } X_1 \text{ und } X_2 \text{ unabhängig sind} \\ Var(a + bX_1) &= b^2 Var(X_1). \end{aligned}$$

Zusammenstellung der wichtigsten Verteilungen

Verteilung	$P(X = k)$ **bzw.** $f(x)$	Wertebereich	$E(X)$	$Var(X)$
Binomial	$\binom{n}{k} p^k (1-p)^{n-k}$	$k = 0, 1, \ldots, n$	np	$np(1-p)$
Neg. Binomial	$\binom{k+r-1}{k} p^r (1-p)^k$	$k = 0, 1, \ldots$	$r\frac{1-p}{p}$	$r\frac{1-p}{p^2}$
Geometrisch	$(1-p)^{k-1} p$	$k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	$\frac{\lambda^k e^{-\lambda}}{k!}$	$k = 0, 1, \ldots$	λ	λ
Uniform	$\frac{1}{b-a}$	$a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$-\infty < x < \infty$	μ	σ^2
Gamma	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$	$x > 0$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Exponential	$\lambda e^{-\lambda x}$	$x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Chiquadrat	$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2}$	$x > 0$	n	$2n$
t	$f(x) = \frac{\Gamma[\frac{n+1}{2}]}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	$-\infty < x < \infty$	0	$\frac{n}{n-2}$
F	$f(x) = \frac{\Gamma[\frac{n+m}{2}]}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} (\frac{n}{m})^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m}x\right)^{-\frac{n+m}{2}}$	$x > 0$	$\frac{n}{n-2}$	

5 Kovarianz und Korrelation

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Erwartungswerten μ_X und μ_Y. Dann ist die *Kovarianz* von X und Y (covariance) gegeben durch:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

falls der Erwartungswert existiert.

Seien X_1, X_2, Y_1 und Y_2 Zufallsvariablen und a, b, c und d konstante Zahlen, dann gilt:

$$\begin{aligned} Cov(X_1 + X_2, Y_1 + Y_2) &= Cov(X_1, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_1) + Cov(X_2, Y_2) \\ Cov(a + bX_1, c + dY_1) &= bdCov(X_1, Y_1). \end{aligned}$$

Daraus folgt:

$$Var(aX_1 + bX_2) = a^2 Var(X_1) + b^2 Var(X_2) + 2abCov(X_1, X_2).$$

Seien X und Y gemeinsam verteilte Zufallsvariablen mit Varianzen verschieden von Null. Dann ist die *Korrelation* von X und Y (correlation) gegeben durch:

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}}.$$

Es gilt: $-1 \leq \rho \leq 1$ und $\rho = \pm 1$ für $Y = a + bX$.