Reelle Zufalsvariablen

DEFINITION: Sei X eine Zufallsvariable. Die Funktion

$$F(x) = P\left(\{\omega \mid X(\omega) \leq x\}\right),$$

heißt *Verteilungsfunktion* von X. Meist schreibt man kurz $P(X \leq x)$ anstelle von $P\left(\{\omega \mid X(\omega) \leq x\}\right)$. Für Verteilungsfunktionen gilt:

1. Sie sind *monoton steigend*. d.h. wenn $x \leq y$, dann ist auch $F(x) \leq F(y)$.
2. Für kleine Werte von x nähert sich $F(x)$ an Null an: $\lim_{x \to -\infty} F(x) = 0$.
3. Für große Werte von x nähert sich $F(x)$ an Eins an: $\lim_{x \to -\infty} F(x) = 0$.

DEFINITION: Zufallsvariablen, die nur Werte aus $\mathbb{R}$ annehmen, werden *reelle Zufallsvariablen* genannt. Eine reelle Zufallsvariable besitzt eine *Wahrscheinlichkeitsdichte* f, falls für die Verteilungsfunktion F eine entsprechende Funktion f existiert, für die gilt:

$$F(x) = \int_{-\infty}^{\infty} f(y)dy. \tag{1}$$

Die Funktion f muß insbesondere stetig sein. Aus der Gleichung 1 folgt

$$P(a \leq X \leq b) = \int_{a}^{b} f(y)dy,$$

$$\int_{-\infty}^{\infty} f(y)dy = 1.$$

Erwartungswert

Definition: Der *Erwartungswert* einer Zufallsvariable ist im diskreten Fall folgendermaßen definiert:

$$\mathcal{E}(X) = \sum_x xp(x)$$

Im stetigen Fall lautet die Definition

$$\mathcal{E}(X) = \int_{-\infty}^{\infty} xf(x)\ d(x)$$

wobei $\int_{-\infty}^{\infty} f(x)d(x) = 1$.

Seien X und Y Zufallsvariablen über dem selben Wahrscheinlichkeitsraum mit den Erwartungswerten $\mathcal{E}(X)$ und $\mathcal{E}(Y)$ und sei $a \in \mathbb{R}$. Dann gilt *Homogenität* und *Additivität*:

$$\mathcal{E}(aX) = a\mathcal{E}(X)$$

$$\mathcal{E}(X+Y) = \mathcal{E}(X) + \mathcal{E}(Y)$$

Rechnen mit Erwartungswerten

1. $\mathcal{E}(a) = a$
2. $\mathcal{E}(aX) = a\mathcal{E}(X)$.

3. $\mathcal{E}(X+a) = \mathcal{E}(X) + a$.

4. $\mathcal{E}(X+Y) = \mathcal{E}(X) + \mathcal{E}(Y)$.

5. $\mathcal{E}(X+Y+Z) = \mathcal{E}(X) + \mathcal{E}(Y) + \mathcal{E}(Z)$.

6. Wenn X und Y stochastisch unabhängig sind, gilt

$$\mathcal{E}(X \cdot Y) = \mathcal{E}(X) \cdot \mathcal{E}(Y).$$

   Ist $\mathcal{E}(X \cdot Y) \neq \mathcal{E}(X) \cdot \mathcal{E}(Y)$, dann sind die Variablen X und Y nicht unabhängig.

7. Für paarweise unabhängige Zufallsvariablen gilt:

$$\mathcal{E}(X \cdot Y \cdot Z) = \mathcal{E}(X) \cdot \mathcal{E}(Y) \cdot \mathcal{E}(Z).$$

Varianz

Die Varianz bezeichnet die mittlere quadrierte Abweichung vom Erwartungswert; somit gilt im diskreten Fall:

$$\sigma^2(X) = \mathrm{Var}(X) \quad = \quad \mathcal{E}\left[(X - \mathcal{E}(X))^2\right]$$

und im stetigen Fall gilt

$$\mathrm{Var}(X) \quad = \quad \int_{-\infty}^{\infty} (x - \mathcal{E}(X))^2 f(x) dx,$$

wobei $f(x)$ die Wahrscheinlichkeitsdichte der Zufallsvariable X bezeichnet. Die *Standardabweichung* von X ist die positive Wurzel aus der Varianz.

Zur praktischen Berechnung der Varianz wird häufig die folgende einfachere Formel $\mathcal{E}(X^2) - \mathcal{E}(X)^2$ verwendet; sie läßt sich folgendermaßen herleiten:

$$\begin{aligned}
\sigma^2(X) &= \mathcal{E}[(X - \mathcal{E}(X))^2] \\
&= \mathcal{E}[X^2 - 2X\mathcal{E}(X) + (\mathcal{E}(X))^2] \\
&= \mathcal{E}(X^2) - \mathcal{E}(2X\mathcal{E}(X)) + \mathcal{E}[(\mathcal{E}(X))^2] \\
&= \mathcal{E}(Y^2) - 2(\mathcal{E}(X))(\mathcal{E}(X)) + \mathcal{E}(X))^2 \\
&= \mathcal{E}(X^2) - (\mathcal{E}(X))^2.
\end{aligned}$$

Rechnen mit Varianzen

1. $\sigma^2(X + a) = \sigma^2_X$. Deshalb muß auch gelten $\sigma^2(a) = 0$.
2. $\sigma^2(aX) = a^2\sigma^2(X)$.

Diese Formeln wurden folgendem Buch entnommen:

Hays, W. L. (1973). *Statistics for the social sciences*, 2nd ed. London et al.:Holt, Rinehart and Wilston.

Kovarianz und Korrelation

Definition: Seien X und Y zwei Zufallsvariablen. Der

Erwartungswert

$$\sigma(X,Y) = \mathcal{E}\left[(X - \mathcal{E}(X)) \cdot (Y - \mathcal{E}(Y))\right] \tag{2}$$

heißt *Kovarianz von X und Y*. Anstelle von $\sigma(X,Y)$ schreibt man auch σ_{XY} oder $\mathrm{Cov}(X,Y)$. Es läßt sich zeigen, daß gilt

$$\begin{aligned} \sigma(X,Y) &= \mathcal{E}\left[(X - \mathcal{E}(X)) \cdot (Y - \mathcal{E}(Y))\right] \\ &= \mathcal{E}(XY) - \mathcal{E}(X)\mathcal{E}(Y). \end{aligned}$$

DEFINITION: Sind X und Y Zufallsvariablen, dann heißt

$$\rho(X,Y) = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)}$$

der *Korrelationskoeffizient von X und Y*; Statt $\rho(X,Y)$ schreibt man auch ρ_{XY}.

Rechnen mit Kovarianzen und Korrelationen

Für zwei Zufallsvariablen X und Y und reelle Konstanten a, b gelten folgende Rechenregeln:

1. $\sigma^2(aX + bY) = a^2\sigma^2(X) + b^2\sigma^2(Y) + 2ab\sigma(X,Y)$.

2. Sind X und Y stochastisch unabhängig, dann ist die Varianz $\sigma^2(X+Y)$ der Summe gleich der Summe der Varianzen:

$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y);$$

ist dagegen die Kovarianz $\sigma(X,Y)$ der beiden Zufallsvariablen von Null verschieden, dann lautet der Zusammenhang

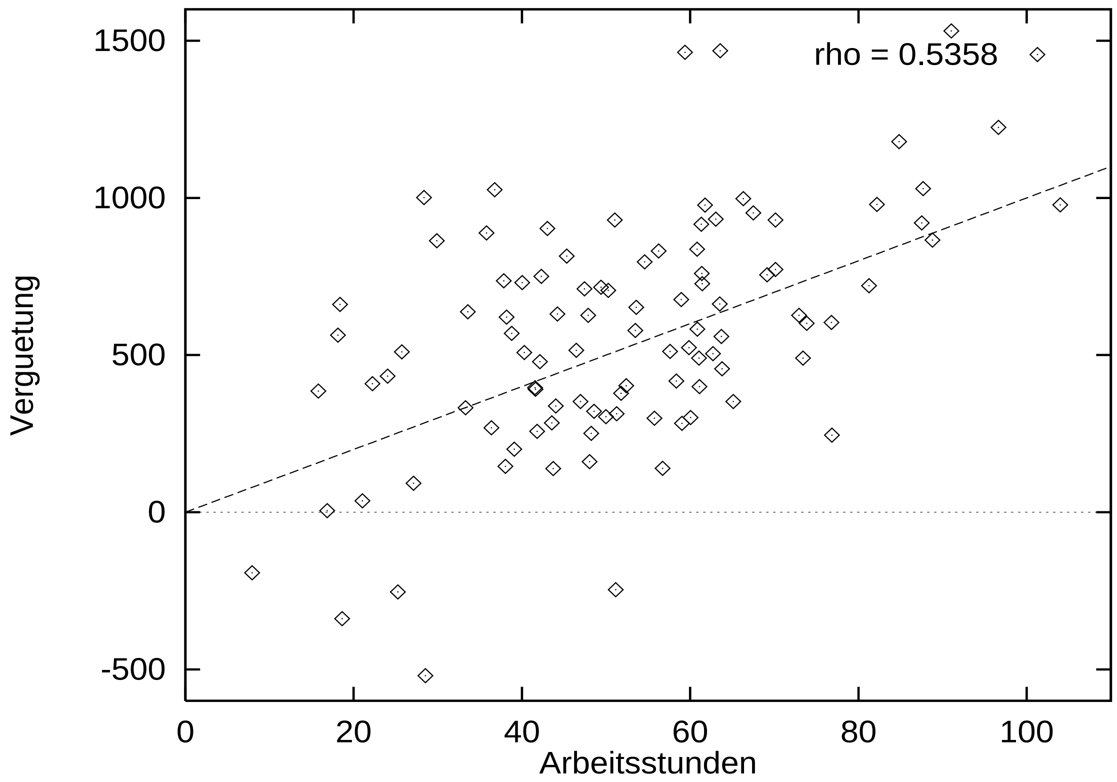
$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y) + 2\sigma(X,Y).$$

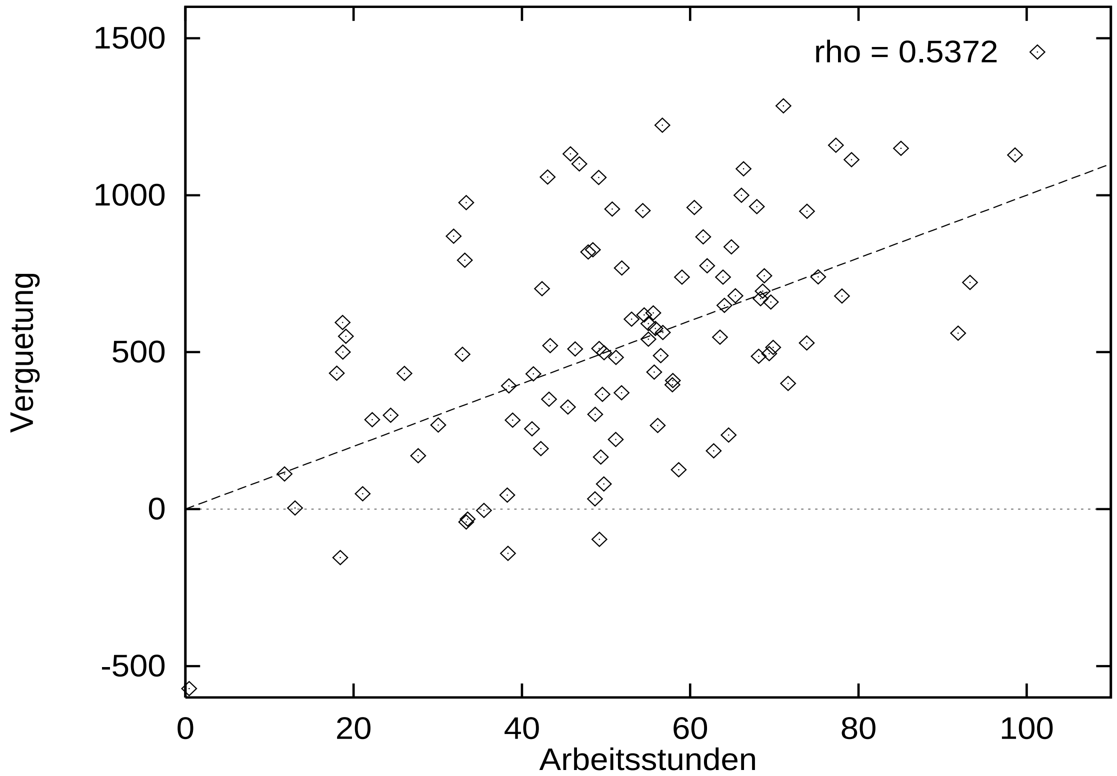
3. $\sigma(X,Y) = \sigma(Y,X)$.

Ein Beispiel für die Korrelation

In diesem Beispiel ist eine mittlere Korrelation von 0.53 veranschaulicht: Der Zusammenhang zwischen Anzahl abgearbeiteter Stunden und der dafür bezahlten Vergütung sei nicht deterministisch (nämlich 10 DM pro Stunde), sondern schwanke zufällig, so daß insgesamt eine Korrelation von 0.53 besteht, wie sie für sehr gute psychologische Test als maximale Validität erreicht werden kann.

Die Zufallsvariable *Arbeitsstunden* sei normalverteilt mit einem Mittelwert von 50 und einer Standardabweichung von 20. Es werden 100 derartige Arbeitszeiten zufällig bestimmt. Zu jeder dieser Zeiten wird der Stundenlohn berechnet, der sich ergibt wenn genau 10 DM pro Stunde bezahlt werden. Dieser "deterministische" Lohn wird dann mit einer zufälligen normalverteilten Schwankung (durch Ausprobieren ergab sich, daß eine Standardabweichung von 300 zur gewünschten Korrelation führt) versehen, so daß sich eine Korrelation von 0.53 zwischen den beiden zufallsvariablen ergibt. Die so bestimmten datenpunkte sind in der folgenden Abbildung zu sehen.





Dieses Beispiel soll demonstrieern, daß eine Korrelation von 0.53, wie sie maximal bei psychologischen Tests erreicht werden kann, sehr starke Schwankungen zuläßt. Die Überprüfung der Homogenität der Probanden mit gleichem Rohwert durch die Korrelation der Testergebnisse mit einem Außenkriterium ist deshalb sehr ungenau.

Kombinatorik

Es sei ein Test mit k Aufgaben gegeben, zu denen es jeweils m Antwortalternativen gibt. Dann gibt es insgesamt m^k verschiedene Lösungsmuster (wird nur zwischen richtigen und falschen Lösungen unterschieden, reduziert sich diese Anzahl auf 2^k.

Wird nur die Summe der richtig gelösten Aufgaben betrachtet, dann spielt es keine Rolle, welche Aufgaben im Einzelnen gelöst gelöst wurden. Wird nur zwichen richtigen und falschen Lösungen unterschieden, dann berechnet sich die Anzahl N der verschiednen Möglichkeiten, bei denen genau k der insgesamt n Aufgaben richtig gelöst sind, folgendermaßen:

$$N = \binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Erschöpfende Statistik

Eine andere Möglichkeit zur Überprüfung, ob sich Personen mit gleicher Anzahl gelöster Aufgaben in ihrer tatsächlichen

Merkmalsausprägung unterscheiden, führt über die Betrachtung von *erschöpfenden (suffizienten) Statistiken.*

Definition Statistik: Eine Statistik ist eine Abbildung

$$S : \mathbb{R}^n \to \mathbb{R}^r;$$

eine Statistik ist also eine Abbildung eines n-elementigen Zufallsvektors (Daten) auf ein r-dimensionales Tupel von (geschätzten) Parametern.

Nun werden die Ergebnisse des Test des Probanden h, der die wahre Fähigkeit θ_h besitzt, untersucht; Das Ergebnis des Tests liege als Antwortvektor $\mathbf{u}_h$ vor; die Statistik s_h bezeichne die Anzahl der in diesem Antwortvektor vorkommenden richtig gelösten Aufgaben.

Nach der Definition der Bedingten Wahrscheinlichkeit gilt dann:

$$\begin{aligned} P(\mathbf{u}_h \mid \theta_h) &= \frac{P(\mathbf{u}_h, \theta_h)}{P(\theta_h)} \\ &= \frac{P(\mathbf{u}_h, s_h, \theta_h)}{P(\theta_h)}, \end{aligned}$$

weil die Statistik s_h eine Funktion der Daten $\mathbf{u}_h$ ist und somit die Wahrscheinlichkeit $P(\mathbf{u}_h, \theta_h)$ nicht verändert. Nun wird wieder die Formel der bedingten Wahrscheinlichkeit $P(\mathbf{u}_h, s_h, \theta_h) = P(\mathbf{u}_h \mid s_h, \theta_h)P(s_h, \theta_h)$ eingesetzt:

$$\begin{aligned} P(\mathbf{u}_h \mid \theta_h) &= \frac{P(\mathbf{u}_h \mid s_h, \theta_h)P(s_h, \theta_h)}{P(\theta_h)} \\ &= P(\mathbf{u}_h \mid s_h, \theta_h)\frac{P(s_h, \theta_h)}{P(\theta_h)} \end{aligned}$$

$$= \quad P(\mathbf{u}_h \mid s_h, \theta_h) P(s_h \mid \theta_h). \qquad (3)$$

Die Gleichung 3 gilt für jede Statistik; demgegenüber wird für *erschöpfende Statistiken* folgende Einschränkung gefordert:

$$P(\mathbf{u}_h \mid \theta_h) \quad = \quad P(\mathbf{u}_h \mid s_h) P(s_h \mid \theta_h).$$

Für erschöpfende Statistiken gilt somit

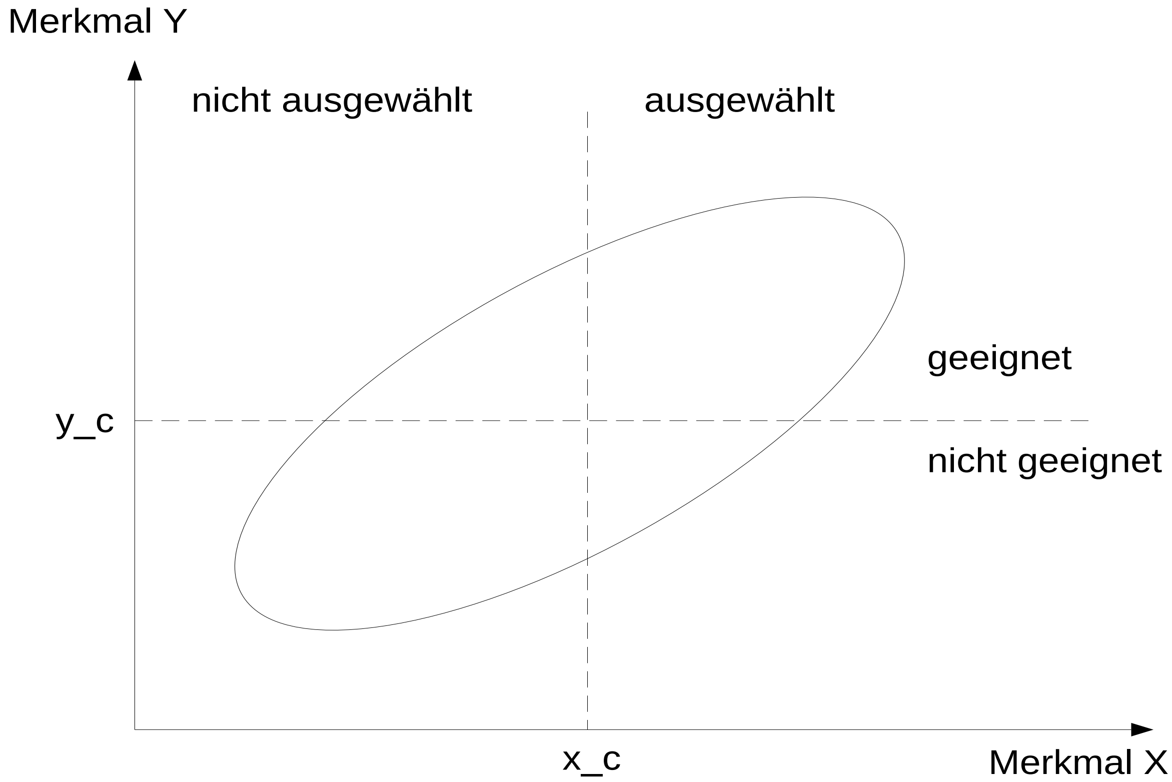
$$P(\mathbf{u}_h \mid s_h, \theta_h) \quad = \quad P(\mathbf{u}_h \mid s_h);$$

dies bedeutet, daß die Wahrscheinlichkeit für einen bestimmten Antwortvektor $\mathbf{u}_h$ bei gegebener Anzahl gelöster Aufgaben nicht weiter von der Personenfähigkeit θ_h abhängt, als durch die Statistsik s_h bereits erfaßt ist.

Entscheidungstheoretische Überlegungen

Angenommen, man will aufgrund eines Testergebnisses die Eignung eines Probanden für eine bestimmte Tätigkeit vorhersagen. Probanden ab einem Testwert von x_c werden für die Tätigkeit akzeptiert; tatsächlich geeignet sind diejenigen, deren Bewährungskriterium Y über der Schwelle y_c liegt. Bei einer binären Selektionsentscheidung sind dann immer vier Fälle zu betrachten:

H_1 **wahr,** H_1 **akzeptiert:** richtige Entscheidung, *true negatives*. $P(\text{richtig} \mid H_1 \text{akzeptiert}) = 1 - \alpha$; *hit*



H_1 wahr, H_1 abgelehnt: falsche Entscheidung, *false positives*. Fehler 2. Art, β-Fehler; $P(\text{falsch} \mid H_1\text{abgelehnt}) = \beta$; *miss*

H_1 falsch, H_1 akzeptiert: falsche Entscheidung, *false negatives*. Fehler 1. Art, α-Fehler; $P(\text{falsch} \mid H_1\text{akzeptiert}) = \alpha$; *false alarm*

H_1 falsch, H_1 abgelehnt: richtige Entscheidung, *true positives*; $P(\text{richtig} \mid H_1\text{abgelehnt}) = 1 - \beta$ (Power eines Tests); *correct rejection*