

Verteilungen

3 5 5 8 10 16 17 17 17 18 22 23 90

Discrete Verteilungen:

Gleichverteilung :

$$Q_X(x) = \begin{cases} \frac{1}{m} & x = 1, 2, 3, \dots, m \\ 0 & \text{sonst} \end{cases}$$

mit Erwartungswert

$$E(X) = \frac{1}{m} \sum_{i=1}^m x_i$$

und Varianz

$$\text{Var}(X) = \frac{1}{m} \sum_{i=1}^m x_i^2 - \left(\frac{1}{m} \sum_{i=1}^m x_i \right)^2.$$

m = Trägerpunkte

Poissonverteilung:

$$Q_X(x, \lambda) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$$

wenn n groß und p klein

mit den Sprungstellen $x \in \mathbb{N} \cup \{0\}$, Erwartungswert $E(X) = \lambda$ und Varianz $\text{Var}(X) = \lambda$. Wird, ausgehend von der Binomialverteilung, p definiert als $p := \frac{\lambda}{n}$

Erwartungswert = Varianz = λ

Binomialverteilung:

$$Q_X(x, n, p) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

mit Erwartungswert

$$E(X) = n \cdot p$$

und Varianz

$$\text{Var}(X) = n \cdot p \cdot (1-p).$$

Hypergeometrische Verteilung:

$$P(x, W, S, n) = \frac{\binom{W}{x} \cdot \binom{S}{n-x}}{\binom{W+S}{n}},$$

mit $P(x, W, S, n) = \text{dhyper}(x, W, S, n)$

x = Anzahl gezogener weißen Kugeln,

W = Anzahl weißer Kugeln in der Urne,

S = Anzahl schwarzer Kugeln in der Urne,

n = Stichprobenumfang.

Der Erwartungswert ist

$$E(X) = n \cdot \frac{W}{N},$$

Ihre Varianz berechnet sich über

$$\text{Var}(X) = n \cdot \frac{W}{N} \cdot \left(1 - \frac{W}{N}\right) \frac{N-n}{N-1}.$$

Stetige Gleichverteilung

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{für } a \leq x \leq b \\ 0, & \text{sonst} \end{cases}$$

mit $a, b \in \mathbb{R}$ und $a < b$. Sie hat Erwartungswert

$$E(X) = \frac{a+b}{2}$$

und Varianz

$$\text{Var}(X) = \frac{(a+b)^2}{12}.$$

Normalverteilung:

$$f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$$

proxim

Exponentialverteilung

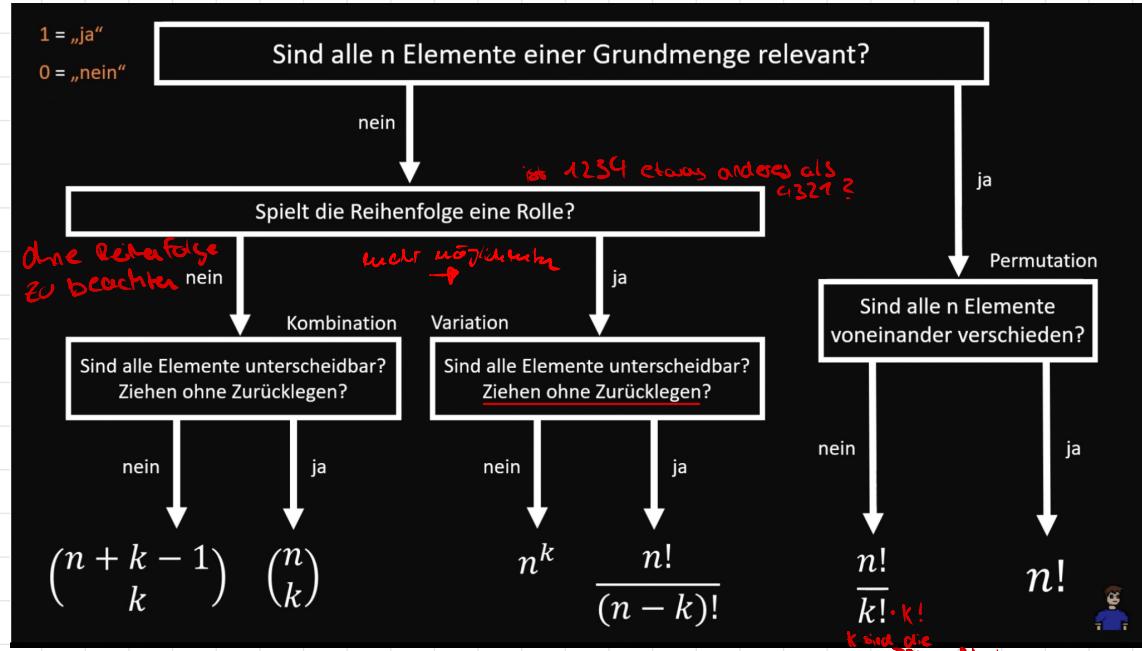
$$f_X(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{für } x \geq 0 \\ 0 & \text{sonst} \end{cases}$$

für $\lambda > 0$, mit Erwartungswert

$$E(X) = \frac{1}{\lambda}$$

und Varianz

$$\text{Var}(X) = \frac{1}{\lambda^2}.$$



Lagemaße

Empirische:

arithmetisches Mittel:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

gewichtetes Mittel:

$$\bar{x}_w = \sum_{i=1}^n x_i w_i, w_i > 0$$

Medien:

$$med(x) = \begin{cases} x_{(\frac{n+1}{2})}, n \text{ ungerade} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), n \text{ gerade} \end{cases} \quad n = \text{länge}$$

Quantile

$$Q_p = \begin{cases} x_{(\lceil n \cdot p \rceil)} & \text{aufgerundet } n \cdot p \text{ nicht ganzzahlig} \\ \frac{1}{2}(x_{(n \cdot p)} + x_{(n \cdot p + 1)}) & n \cdot p \text{ ganzzahlig} \end{cases}$$

Theoretisch:

Erwartungswert diskret:

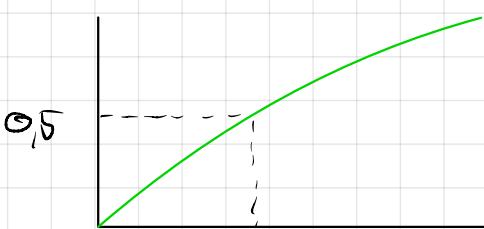
$$E(X) := \sum_{x \in T} x \cdot P(X = x)$$

Erwartungswert stetig:

$$E(X) := \int_{-\infty}^{+\infty} x \cdot f_X(x) dx$$

für jedes Intervall mit x multiplizieren!
Die Dichtefunktion! nicht aus Integral erst dann integrieren

Median / Quantile:



Bei Beurteilung der Dichtefunktion ist Verteilungsfunktion ≥ 1

Streuungsmaße

Empirisch:

Varianz: $\text{var}(x) := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ - "Summe der quadrierten Abweichungen vom Mittelwert"

Standardabweichung: $sd(x) := \sqrt{\text{var}(x)}$

Medianabweichung: $MA(x) := \frac{1}{n} \sum_{i=1}^n |x_i - \text{med}(x)|$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Theoretisch:

Steine: (Nur bei Varianz ! ! !)

| Der Erwartungswert der quadrierten Häufigkeiten minus den quadrierten Erwartungswert

$$\sigma_x^2 = \text{Var}(X) = E(X^2) - (E(X))^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x) dx \right) - (E(X))^2$$

$$\text{Var}(U) = E(U^2) - (E(U))^2$$



Varianz einer diskreten Zufallsvariable

↪ diskrete Zufallsvariable X mit Träger $\mathcal{T} = \{x_1, x_2, \dots\}$

↪ Varianz $\sigma_X^2 = \text{var}(X)$ ist definiert durch

$$\sigma_X^2 := E((X - \mu_X)^2) = \sum_{x \in \mathcal{T}} (x - \mu_X)^2 \cdot f_X(x)$$

↪ gewichtetes arithmetisches Mittel der quadrierten Abweichungen (der Elemente des Trägers) vom Erwartungswert μ_X

Varianz einer stetigen Zufallsvariable

↪ stetige Zufallsvariable X

↪ Varianz $\sigma_X^2 = \text{var}(X)$ ist definiert durch

$$\sigma_X^2 := E((X - \mu_X)^2) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 \cdot f_X(x) dx$$

steine

ux²

Schiefe

↪ sei X eine Zufallsvariable

↪ das dritte zentrale Moment $\mu_3 := E((X - \mu_X)^3)$ bezeichnet man als Schiefe von X

↪ der Quotient μ_3/σ_X^3 heißt Schiefekoeffizient

↪ eine Verteilung heißt rechtssteil / linksschief linkssteil / rechtsschief $\Leftrightarrow \mu_3 < 0$

↪ für diskrete bzw. stetige Zufallsvariablen X gilt:

X heißt symmetrisch $\Leftrightarrow \forall x \in \mathbb{R}: f_X(\mu_X - x) = f_X(\mu_X + x)$

$$\mu_3 = E((X - \mu_X)^3) = \int_{-\infty}^{\mu_X} \underbrace{(x - \mu_X)^3}_{\leq 0} \cdot f_X(x) dx + \int_{\mu_X}^{+\infty} \underbrace{(x - \mu_X)^3}_{\geq 0} \cdot f_X(x) dx = 0$$

wird hier eingesetzt

Wölbung

↪ sei X eine Zufallsvariable

↪ das vierte zentrale Moment $\mu_4 := E((X - \mu_X)^4)$ bezeichnet man als Wölbung bzw. Kurtosis von X

↪ der Quotient $(\mu_4/\sigma_X^4) - 3$ heißt Wölbungskoeffizient

↪ der Wölbungskoeffizient ist nur bei symmetrischen Verteilungen interpretierbar und wird stets mit der Verteilung der Standardnormalverteilung verglichen

↪ eine symmetrische Verteilung ist flacher spitzer $\Leftrightarrow [(\mu_4/\sigma_X^4) - 3] < 0$

wird eingesetzt

Gesetze großer Zahlen:

Mit Hilfe der Tschebyscheff-Ungleichung ist eine WS-Aussage zur Abweichung zwischen \bar{X}_n und μ möglich:

Obere Grenze ist gesucht!

Tschebyscheff:

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \text{var}(\bar{X}_n)/\epsilon^2 = \frac{\text{var}(X_1)}{n\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

$$P(|\bar{X}_n - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

In Worten: Die Wahrscheinlichkeit P , dass die absolute Differenz (nach oben und unten) zwischen der Zufallsvariablen X und deren Erwartungswert $E(X)$ größer als ein vorgegebener Wert t ist, ist kleiner gleich der Varianz geteilt durch t im Quadrat.

Ab wievielen Würfen weicht die mittlere Augenzahl eines W6-Würfels mindestens mit WS 0,9 höchstens um 0,1 vom Erwartungswert 3,5 ab?

- Tschebytscheff-Ungleichung: $P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\text{var}(X_1)}{n\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$
- Gesucht: WS vom Gegenereignis: $P(|\bar{X}_n - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$ *n nur wenn angeben*
- Varianz Würfelwurf: $\sigma^2 = \text{var}(X_1) = \frac{1^2 + \dots + 6^2}{6} - 3,5^2 = \frac{35}{12}$ (Satz von Steiner)
- Daraus mit $\epsilon = 0,1$: $P(|\bar{X}_n - 3,5| \leq 0,1) \geq 1 - \frac{35/12}{0,01n}$
- Hinreichend damit: $1 - \frac{35/12}{0,01n} \geq 0,9$. Löse diese Ungleichung nach n auf

$$1 - \frac{35/12}{0,01n} \geq 0,9 \Leftrightarrow \frac{35/12}{0,01n} \leq 0,1 \Leftrightarrow n \geq \frac{35/12}{0,001} \approx 2916,67$$

- D.h. hinreichend sind mindestens 2917 Würfe.
- Für eine Verbesserung der Genauigkeit ϵ um eine Nachkommastelle muss der Stichprobenumfang ver-100-facht werden.

Zentraler Grenzwertsatz:

Es gilt immer dass wir eine \mathcal{E} haben.
Wir haben immer μ (Erwartungswert)
 σ (Standardabweichung)
Ein Ereignis A das wir untersuchen wollen
 $\rightarrow \mu = \sum_i x_i$ $\rightarrow \sum_i = \bar{x}_i$

Wir ziehen also \mathcal{E} Zufallszahlen mit μ und σ
Dann schreibe wir $P(A) \rightarrow P(\bar{x}_i \leq \frac{A}{n})$

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$$

$$\mu = 5 \quad S_{225} = \sum_{i=1}^{225} x_i \text{ mit identisch Verteilten } x \\ \sigma = 1$$

$$\begin{aligned} P(S_{225} \geq 1105) &= P\left(\sum_{i=1}^{225} x_i \geq 1105\right) = P\left(\bar{x} \geq \frac{1105}{225}\right) \\ &\leq P\left(\frac{\bar{x}_i - 5}{\sqrt{1/225}} \geq \frac{1105}{225} - 5\right) \\ &\leq P\left(\underbrace{\sqrt{1/225} \cdot \frac{\bar{x}_i - 5}{\sqrt{225}}}_{:= z} \geq \frac{1105}{225} - 5\right) \\ &\leq P\left(z \geq 1 - \phi\left(\frac{1105}{225} - 5\right)\right) \\ &\approx 0,09988 \end{aligned}$$

$$P(\bar{X}_n - 1,66) > 15 \leq \frac{642}{15}$$

$$P(\bar{X}_n - 1,66) \leq 0,1820$$

Die Wahrscheinlichkeit ist maximal 0,1820

$$\mu = 2 \text{ min} \quad \sigma = 0,8 \quad N(2, 0,8)$$

$$\geq P\left(\sum_i x_i \leq 180\right)$$

$$\leq P(x_i \leq 2)$$

$$\geq P\left(\frac{\bar{X}_n - 2}{\sqrt{0,8^2/25}} \leq \frac{180 - 2}{\sqrt{0,8^2/25}}\right)$$

$$\geq P\left(\sqrt{95} \frac{\bar{X}_n - 2}{0,8} \leq \frac{180 - 2}{\sqrt{0,8^2/25}}\right) \quad z$$

$$\geq P(z \leq \Phi^{-1}(0,998))$$

$$\geq P(z \leq 1 - (1 - \Phi^{-1}(0,998)))$$

$$S_{95} = \sum_{i=1}^{95} X_i, \quad \text{mit } X_i : \text{Minuten Karton } i$$

$$\begin{aligned} P(S_{95} \leq 180) &= P\left(\sum_{i=1}^{95} X_i \leq 180\right) = P\left(\bar{X}_{95} \leq \frac{180}{95}\right) \\ &= P\left(\frac{\bar{X}_{95} - 2}{\sqrt{(48/60)^2/95}} \leq \frac{180/95 - 2}{\sqrt{(48/60)^2/95}}\right) \\ &= P\left(\sqrt{95} \cdot \frac{\bar{X}_{95} - 2}{48/60} \leq \sqrt{95} \cdot \frac{180/95 - 2}{0,8}\right) \\ &\approx \Phi\left(\sqrt{95} \cdot \frac{180/95 - 2}{0,8}\right) = \Phi(-1,28) \\ &= 1 - \Phi(1,28) = 0,0998 \end{aligned}$$

Cluster

Distanzmaße
Kardinal

Ähnlichkeitsmaße
Ordinal

Ähnlichkeitsmaße
binär

Euklidische Distanz
Manhattan Distanz

Vergleichendes M-Koeffizient:

$$M(x, y) = \frac{u(x, y)}{n}$$

 Die Gleichen
 Alle Ausprägungen
 www, TU Dortmund Osnabrück...

ordinal

Ähnlichkeitsmaße für ordinale Merkmale

- betrachte die bivariaten Beobachtungen x und y mit den Merkmalen
 - Schulbildung („Hauptschulabschluss“, „mittlere Reife“ und „Abitur“)
 - Note im Leistungsspiegel (5 bis 1)

→ Realisierungen für zwei Personen (ordinal):

$$x = (\text{mittlere Reife}, 3)^T, \text{ sowie } y = (\text{Abitur}, 4)^T$$

→ daraus resultieren die neuen (binären) Beobachtungsvektoren:

$$\tilde{x} = \begin{pmatrix} 1 & 1 & 0 \\ \text{mittl. R.} & = \text{Note 3} \end{pmatrix}, \text{ sowie } \tilde{y} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ \text{Abitur} & = \text{Note 4} \end{pmatrix}$$

→ nun z.B. Berechnung des S-Koeffizienten $s(\tilde{x}, \tilde{y})$

$$s(\tilde{x}, \tilde{y}) = \frac{4}{6} = \frac{2}{3}$$

Ähnlichkeitsmaße für binäre Merkmale

Von sieben Kunden eines Versandhauses ist bekannt, ob sie im vergangenen Jahr Einkäufe in einer von acht Sparten getätigt haben (1 = ja, 0 = nein).

Kunde	kleidung	dkleidung	schuhe	mobel	elektro	garten	sport	musik
1	1	0	1	1	0	0	0	0
2	0	0	0	0	1	0	1	0
3	1	1	0	1	0	1	0	0
4	1	1	1	1	1	1	1	1
5	0	0	0	0	1	1	0	0
6	1	0	0	1	0	1	0	0
7	1	0	1	0	1	0	1	0

Steuerschlüsselkenn

n-grammatische Obj

$$\text{Ergänzung der Ähnlichkeitsmatrix der S-Koeffizienten } s = \frac{a}{a+b+c} ?$$

M-Koeffizient:

$$s = \frac{a+c}{n}$$

Übereinstimmungen!

und sind Übereinstimmungen!
n = Spaltalänge

$$s(x_2, x_2) = \frac{2}{(2+6)}$$

$$s(x_5, x_7) = \frac{1}{(1+1+4)}$$

$x_2 \setminus x_3$	1	0
1	2	1
0	1	3
8	0	8

$$s(x_2, x_3) = \frac{2+3}{7} = \frac{5}{7}$$

$$s(x_5 \setminus x_7) = \frac{1}{(1+1+4)} = \frac{1}{6}$$

$$s(x_5, x_7) = \frac{1}{(1+1+4)} = \frac{1}{6}$$

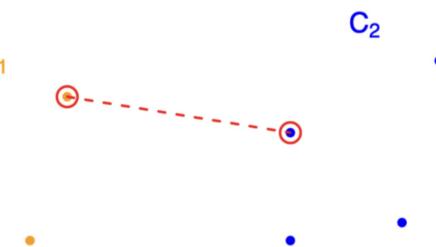
Bei ordinalen, binären oder nominalen Ähnlichkeitmatrizen muss zuvor $1 - s(x_i, y_j)$ gerechnet werden.
Es muss wieder 0 auf die Hauptdiagonalen stehen.

Single Linkage:

$$D(C_i, C_j) := \min_{x \in C_i, y \in C_j} d(x, y)$$

*	{1}	{2}	{3}	{4}	{5}
{1}	0	17.2	29.1	14.3	15.7
{2}	17.2	0	13	28.9	3
{3}	29.1	13	0	38.1	15.8
{4}	14.3	28.9	38.1	0	28.3
{5}	15.7	3	15.8	28.3	0

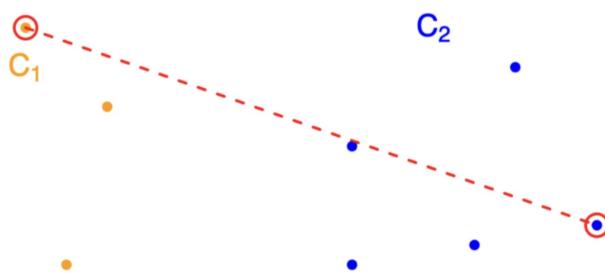
*	{1}	{2, 5}	{3}	{4}
{1}	0	15.7	29.1	14.3
{2, 5}	15.7	0	13	28.3
{3}	29.1	13	0	38.1
{4}	14.3	28.3	38.1	0



Complete Linkage

$$D(C_i, C_j) := \max_{x \in C_i, y \in C_j} d(x, y)$$

	{1}	{2}	{3}	{4}	{5}	{6}		{1}	{2}	{3}	{4,5}	{6}
{1}	0	16	20	17	15	11	{1}	0	16	20	17	11
{2}	16	0	4	9	11	27	{2}	16	0	4	11	27
{3}	20	4	0	9	11	27	{3}	20	4	0	11	27
{4}	17	9	9	0	2	18	{4,5}	17	11	11	0	18
{5}	15	11	11	2	0	16	{6}	11	27	27	18	0
{6}	11	27	27	18	16	0						



$$\begin{aligned} B &\approx CFD \quad 0,75 + 0,25 + 0,2 \\ E &\approx CFD \quad 0,28 + 0,39 + 0,28 \end{aligned}$$

$\alpha =$ Durchschnitt der Durchschnittlichen Strecke von A zu {CDE} und der Durchschnittlichen Strecke von B zu {CDE}

$$\alpha = \frac{1}{2} \cdot \left(\frac{1}{3} (\bar{AC} + \bar{AD} + \bar{AE}) + \frac{1}{3} (\bar{BC} + \bar{BD} + \bar{BE}) \right)$$

Average Linkage

$$D(C_i, C_j) := \frac{1}{n_i \cdot n_j} \cdot \sum_{x \in C_i} \sum_{y \in C_j} d(x, y)$$

	P1	P2	P3	P4	P5	P6		P1	P2	P3,P6	P4	P5
P1	0						P1	0				
P2	0.23	0					P2	0.23	0			
P3	0.22	0.15	0				P3,P6	0.23	0.2	0		
P4	0.37	0.20	0.15	0			P4	0.37	0.20	0.19	0	
P5	0.34	0.14	0.28	0.29	0		P5	0.34	0.14	0.34	0.29	0
P6	0.23	0.25	0.11	0.22	0.39	0						

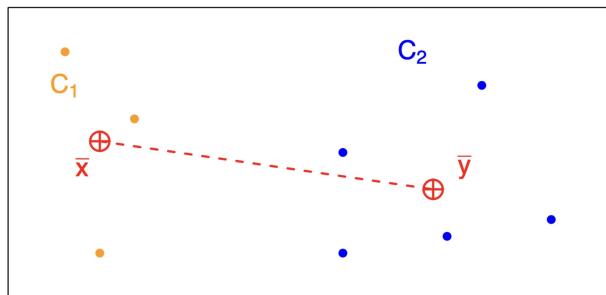
	P1	P2,P5	P3,P6	P4
P1	0			
P2,P5	0.29	0		
P3,P6	0.23	0.27	0	
P4	0.37	0.25	0.19	0

	P1	P2,P5	P3,P6,P4
P1	0		
P2,P5	0.29	0	
P3,P6,P4	0.3	0.26	0

$$\begin{aligned} P2 &\rightarrow \frac{1}{3} (0,15 + 0,25 + 0,20) = \frac{1}{3} \cdot \frac{13}{25} \cdot \frac{1}{2} = 0,26 \\ P5 &\rightarrow \frac{1}{3} (0,28 + 0,39 + 0,28) = \frac{1}{3} \cdot \frac{13}{25} \cdot \frac{1}{2} = 0,26 \end{aligned}$$

Zentroid-Verfahren

$$D(C_i, C_j) := \|\bar{x} - \bar{y}\|, \quad \text{mit } \bar{x} := \frac{1}{n_i} \sum_{x \in C_i} x \text{ und } \bar{y} := \frac{1}{n_j} \sum_{y \in C_j} y$$



Durchschnittspunkte von einzelnen Clustern berechnen und davon dann die Distanzen
→ bei einzelnen Punkten ist das dann der Schnitt

Wort

Minimiere pro Iteration die Heterogenität über alle Cluster

Bivariate Merkmale

Diskretes Fall:

		deskriptiv		gemeinsames WS-Modell	
Urliste		Aufbereitung		(theoretisch)	
		diskret: Kontingenztafel		Dichte	
Träger	X Y	$B_1 \dots B_L$	$H_{11} \dots H_{1L}$	$B_1 \dots B_L$	$f(X)$
1	$x_1 y_1$		$H_{11} \dots H_{1L}$	$p_{11} \dots p_{1L}$	$p_{1\bullet}$
2	$x_2 y_2$		$\dots \dots \dots$	$\dots \dots \dots$	$\dots \dots \dots$
\vdots	$\vdots \vdots$		$H_{K1} \dots H_{KL}$	$p_{K1} \dots p_{KL}$	$p_{K\bullet}$
n	$x_n y_n$		$H_{\bullet 1} \dots H_{\bullet L}$	$p_{\bullet 1} \dots p_{\bullet L}$	1
		Zeil-/Spaltsumme			
		stetig: Streudiagramm:			

Aufsummieren von WS!

Wartung einer Maschine:

		Defekt II			f_{X_1}
		0	1	2	
Defekt I	0	0.1	$+ \varepsilon$	$0.3 - \varepsilon$	0.05
	1	0.05	$- \varepsilon$	$0.1 + \varepsilon$	0.1
	2	0.05	0.05	0.2	0.3
	f_{X_2}	0.2	0.45	0.35	1

Wartung einer Maschine:

X : Anzahl Defekte vom Typ I im Wartungszeitraum

Y : Anzahl Defekte vom Typ II im Wartungszeitraum

Die Einträge der Verteilungsfunktion (links) ergeben sich durch Addition des Eintrags der Dichtefunktion summiert zu den Zellwerten, die links und oberhalb liegen:

$F_{X,Y}(x,y)$			$f_{X,Y}(x,y)$				
		Defekt II			Defekt II		
		0	1	2	0	1	2
Defekt I	0	0.1	0.4	0.45	Defekt I	0	0.1
	1	0.15	0.55	0.7	1	0.05	0.1
	2	0.2	0.65	1	2	0.05	0.2
		$F_{X,Y}(x,y)$		$f_{X,Y}(x,y)$		Randverteilungsfunktion von X	
		0	0.1	0.4	0.45	Randverteilungsfunktion von Y	
		1	0.15	0.55	0.7		
		2	0.2	0.65	1		

Stetiger Fall:

Für einen bivariate Zufallsvektor $(X Y)$ werden die Dichtefunktionen

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy \text{ und}$$

$$f_Y(y) := \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$$

stetige Randdichtefunktionen von X bzw. Y genannt.

Unabhängigkeit: Stetig + diskret
 $f(x,y) = f_X(x) \cdot f_Y(y)$.

Randdichten multiplizieren!

Bei diskret: Tabellensummen müssen inneren Wert ergeben!

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & 0 \leq x_1, x_2 \leq 1, \\ 0 & \text{sonst} \end{cases}$$

$f_{X_1, X_2}(x_1, x_2)$ ist eine Dichte:

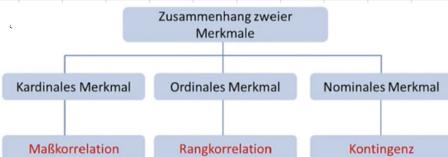
$$\begin{aligned} & \int_0^1 \int_0^1 f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \\ &= \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_2 \\ &= \frac{1}{4} \cdot \left[x_2 \right]_0^1 + \frac{3}{2} \cdot \left[\frac{x_2^2}{2} \right]_0^1 \\ &= \frac{1}{4} + \frac{3}{4} = 1 \end{aligned}$$

$$f_{X_1}(x_1) = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_2$$

$$= \frac{1}{2}x_1 + \frac{3}{2} \left[\frac{x_2^2}{2} \right]_0^1 = \frac{1}{2}x_1 + \frac{3}{4}$$

$$f_{X_2}(x_2) = \int_0^1 \left(\frac{1}{2}x_1 + \frac{3}{2}x_2 \right) dx_1 = \frac{3}{2}x_2 + \frac{1}{4}$$

Korrelation



Kontingenzkoeffizient:

$$\chi^2 := \sum_{i=1}^K \sum_{j=1}^L \frac{(H_{ij} - E_{ij})^2}{E_{ij}}$$

$$= n \cdot \left(\left(\sum_{i=1}^K \sum_{j=1}^L \frac{H_{ij}^2}{Z_i \cdot S_j} \right) - 1 \right)$$

Pearsons (korrigierter) Kontingenzzindex

$$K_P := \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \text{mit } 0 \leq K_P < 1$$

$$K_{P^*} := \sqrt{\frac{\chi^2}{\chi^2 + n}} / K_P^{\max} \quad \text{mit } 0 \leq K_{P^*} \leq 1$$

$$K_P^{\max} = K_P((\chi^2)^{\max}) = \sqrt{\frac{\min(K, L) - 1}{\min(K, L)}} \quad \text{Minimum aus Zeilen/Spalten}$$

Abschluss	ja	vielleicht	nein	Σ
Realschule	4	12	8	24
Gymnasium	8	24	16	48
Gesamtschule	2	6	4	12
Σ	14	42	28	84

E_M aus Kontingenztafel:

	Σ	Z_i
Hau	8	7
	11	7
	1	17
Σ	13	31
	$\frac{13}{44}$	$\frac{31}{44}$
Z_i	8	

$$E_M = 8 \cdot \frac{13}{44} \cdot \frac{15}{44}$$

Kovarianz (empirisch)

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \leftrightarrow \quad \text{cov}(x, y) = \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y}$$

Braais-Pearson (Kardinale Merkmale)

$$r_{xy} = \frac{\text{cov}(x, y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{S_x \cdot S_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Bei Braais-Pearson Varianz und Kovarianz benutzen dann einsetzen!

misst Linearität

Wie sehr sind die Punkte von der Linie weg

Spearmann (Rangkorrelationskoeffizient)

Student	A	B	C	D	E	F	G	H	I	J	K
Mathe x _i	38	47	44	51	35	29	22	14	12	19	9
R(x _i)	8	10	9	11	7	6	5	3	2	4	1
Statistik y _i	39	34	31	48	46	23	17	12	16	28	10
R(y _i)	9	8	7	11	10	5	4	2	3	6	1

Durchschnittliche Rangzahlen:

$$\bar{R}_x = \bar{R}_y = 6$$

Bei Spearmann muss Tabelle gerecht werden!

$$(x_{(1)}, \dots, x_{(10)}) = (1, 1, 2, 2, 3, 4, 4, 4, 5)$$

$$R(x_{(1)}) = R(x_{(2)}) = \frac{1+2}{2} = 1.5$$

$$R(x_{(3)}) = \dots = R(x_{(5)}) = \frac{3+4+5}{3} = 4$$

$$R(x_{(7)}) = \dots = R(x_{(9)}) = \frac{7+8+9}{3} = 8$$

Stud.	$R(x_i) - \bar{R}_x = M_i$	M_i^2	$R(y_i) - \bar{R}_y = S_i$	S_i^2	$M_i \cdot S_i$
A	2	4	3	9	6
B	4	16	2	4	8
C	3	9	1	1	3
D	5	25	5	25	25
E	1	1	4	16	4
F	0	0	-1	1	0
G	-1	1	-2	4	2
H	-3	9	-4	16	12
I	-4	16	-3	9	12
J	-2	4	0	0	0
K	-5	25	-5	25	25
Σ	0	110	0	110	97

$$r_{xy}^R = \frac{\sum_{i=1}^n (R(x_i) - \bar{R}_x)(R(y_i) - \bar{R}_y)}{\sqrt{\sum_{i=1}^n (R(x_i) - \bar{R}_x)^2 \cdot \sum_{i=1}^n (R(y_i) - \bar{R}_y)^2}}$$

Bei einzigartigen Rängen:

$$r_{xy}^R = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad \text{mit}$$

$$d_i := R(x_i) - R(y_i)$$

Bei mehrfachen: Durchschnittsrang

$$r_{xy}^R = \frac{97}{\sqrt{110 \cdot 110}} = \frac{97}{110} = 0.88$$

Misst die Monotonie → Wie stark das in eine Richtung geht

Theorie der Zweiwurtszahlen

Seien X, Y Zufallsvariablen mit Dichtefunktion $f_{X,Y}$ und existierenden Varianzen.

Kovarianz / Pearson-Korrelation:

$$\begin{aligned} \text{cov}(X, Y) &:= E[(X - EX)(Y - EY)] = E[XY] - E[X]E[Y] \\ E[XY] &= \sum_{k,i} x_k \cdot y_i \cdot f_{X,Y}(x_k, y_i) \quad (\text{diskreter Fall}) \\ E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot f_{X,Y}(x, y) dx dy \quad (\text{stetiger Fall}) \end{aligned}$$

$$\text{cor}(X, Y) := \rho = \text{cov}(X, Y) / \sqrt{\text{var}(X)\text{var}(Y)} \quad -1 \leq \rho \leq 1$$

Beispiel direkt:

Beispiel: Wartung einer Maschine

		Defekt II			f_{X_1}
		0	1	2	
Defekt I	0	0.1	0.3	0.05	0.45
	1	0.05	0.1	0.1	0.25
	2	0.05	0.05	0.2	0.3
		0.2	0.45	0.35	1

$$E[X_1] = 0 \cdot 0.45 + 1 \cdot 0.25 + 2 \cdot 0.3 = 0.85$$

$$E[X_2] = 0 \cdot 0.2 + 1 \cdot 0.45 + 2 \cdot 0.35 = 1.15$$

$$E[X_1 X_2] = 0 \cdot (0.1 + \dots + 0.05) + 1 \cdot (0.1) + 2 \cdot (0.05 + 0.1) + 4 \cdot (0.2) = 1.2$$

$$\text{cov}(X_1, X_2) = E[X_1 X_2] - E[X_1] \cdot E[X_2] = 1.2 - 0.85 \cdot 1.15 = 0.2225$$

$$E[X_1^2] = 0 \cdot 0.45 + 1 \cdot 0.25 + 2^2 \cdot 0.3 = 1.45, \quad \text{var}(X_1) = 1.45 - 0.85^2 = 0.7275$$

$$E[X_2^2] = 0 \cdot 0.2 + 1 \cdot 0.45 + 2^2 \cdot 0.35 = 1.85, \quad \text{var}(X_2) = 1.85 - 1.15^2 = 0.5275$$

$$\text{cor}(X_1, X_2) = \text{cov}(X_1, X_2) / \sqrt{\text{var}(X_1) \cdot \text{var}(X_2)} = 0.22 / \sqrt{0.728 \cdot 0.528} \approx 0.359$$

wichtig: Für Erwartungswerte:

für X_1 hier Zeilensumme $\Rightarrow f_{X_1}$

für X_2 hier Spaltensumme $\Rightarrow f_{X_2}$

Mengen

Potenzmenge: bei $\{1, 2, 3\}$ wären es $\{\emptyset\}, \{1\}, \{2\}, \dots, \{1, 2\}, \dots, \{1, 2, 3\}, \emptyset$
keine Kombinationen oder etwas enthalten, nur die Teilmengen, die Menge selber und die leere Menge.

Grundgesamtheit: Menge über die wir nachdenken also $\{1, 2, 3\}$

Ergebnis σ -Algebra: Ist eine Teilmenge der Potenzmenge, da sie nicht alle Teilmengen enthält aber auf jeden Fall \emptyset und Ω

Es muss immer das Gegenereignis auch vorhanden sein.

Es muss immer das Vereinigungseignis auch enthalten sein $\{A\}, \{B\}$ und $\{A \cup B\}$

$$P(A) * P(B) = P(A \cap B)$$

$$P(A) + P(B) - P(A) * P(B) = P(A \cup B)$$

sicheres/unmögliches Ereignis, Komplementärereignisse, endliche Vereinigungen und Durchschnitte von Ereignissen liegen in der Ergebnisalgebra.

Empirische Dichte / Verteilungsfunktion

Vorgehen:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
k_i	41	36	25	36	41	32	41	36	32	25	36	36	41	25	36	25

1. n bestimmen: $n = 16$
2. Auftretstabellen machen (wie oft kommt was vor?)

j	k_j	$H(k_i)$
1	25	4
2	32	2
3	36	6
4	41	4

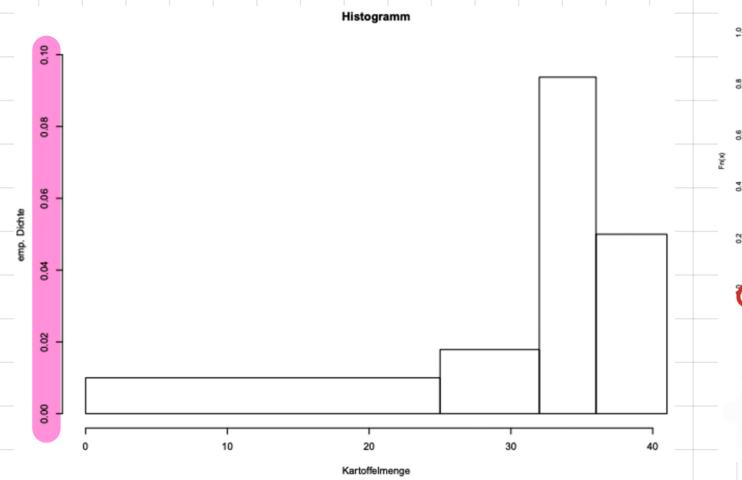
3. Werte für grafische Darstellung ermitteln:

b_i	25	7	4	5
H_i	4	2	6	4
$h_i = H_i/N$	0.25	0.125	0.375	0.25
r_i	0.01	0.0179	0.09375	0.05

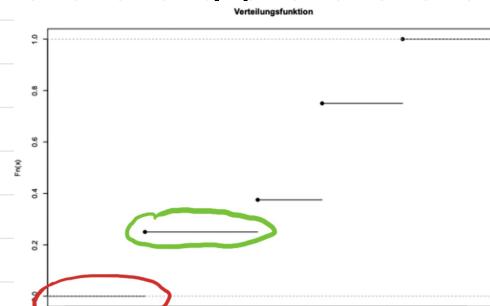
$$h_i = H_i/N$$

$$r_i = H_i/b_i$$

4. Histogramm zeichnen:



5. Verteilungsfunktion zeichnen:

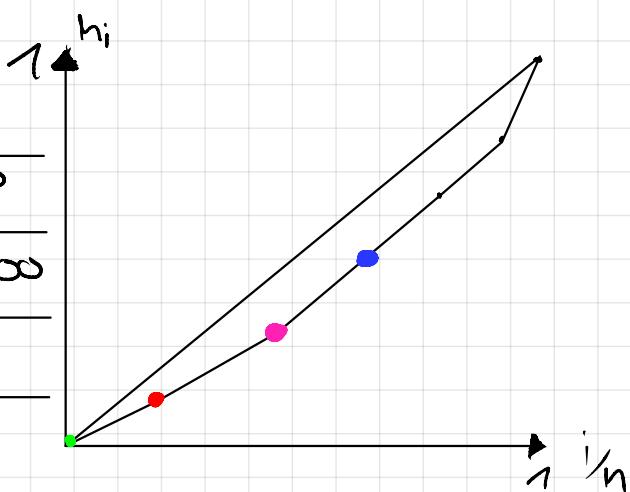


Kumulierte h_i 's

$\sum h_i$ bis zum ersten Auftreten = 0

Normz-Kurve

Tafelvortrag	T_1	T_2	T_3	T_4	T_5	T_6
Bratföldin	200	300	600	1300	2000	4000
Partikularmenne	2000	500	1100	2900	4400	8400
i/n	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1
h_i	$\frac{200}{8400}$	$\frac{500}{8400}$	$\frac{1100}{8400}$	$\frac{2900}{8400}$	$\frac{4400}{8400}$	1



$$L(x_{\text{vor}}) = 1 - \frac{2}{n} \cdot \sum_{i=1}^n \left(\frac{v_{(i)} - v_{(i-1)}}{2} + v_{(i-1)} \right)$$

$$\begin{aligned} V(0) &= 0 & V(2) &= \frac{500}{8900} \\ V(1) &= \frac{200}{8900} & V(3) &= \frac{1100}{8900} \\ V(6) &= 1 \end{aligned}$$

Dann aufsummieren!

Satz von Bayes und totale Wahrscheinlichkeit

$P(A)$ $P(B|A)$ bestimmen

Immer Baum zeichnen!

Formel von Bayes

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

$$P(\bar{A}|B) = \frac{P(B|\bar{A}) \cdot P(\bar{A})}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

↪ oftmals nützlich:

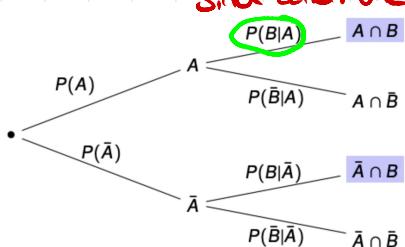
$$\begin{aligned} P(A|B) + P(\bar{A}|B) &= 1 \\ P(A|\bar{B}) + P(\bar{A}|\bar{B}) &= 1 \end{aligned}$$

↪ ! es gilt nicht:

$$\begin{aligned} P(A|B) + P(A|\bar{B}) &= 1 \\ P(\bar{A}|B) + P(\bar{A}|\bar{B}) &= 1 \end{aligned}$$

Totale Wahrscheinlichkeit

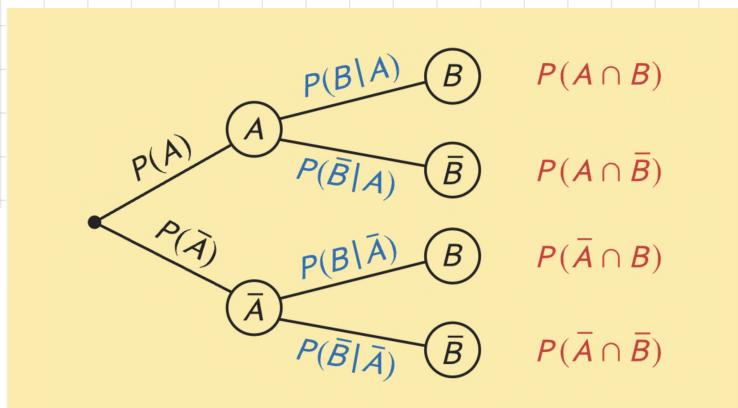
Sind schon die Bedingten



↪ Die totale WS entspricht der Addition aller Pfad-WS, die zu Blättern führen, in denen B eintritt.

↪ Die Pfad-Wahrscheinlichkeiten werden mit der Multiplikationsregel bestimmt (s.o.).

$$P(B) = P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})$$



Bei Satz von Bayes nur einfache WS nehmen

