Grammatik von Grafiken

Grafik = Daten + geometrische Elemente + Ästhetische Zuordnung
+ Datentransformationen + Skalen + Koordinatensysteme
+ Facettierung + Theme + {Grafik}

Geometrische Elemente = Punkte | Linien | Rechtecke | Boxplots | Dichtefkt. | ...

Ästhetische Zuordnung = Position & Farbe & Größe & Form

Datentransformationen = id | Mittelwerte | Anteile | ...

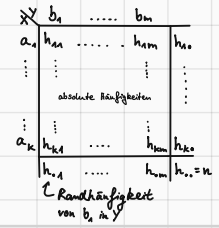
Skalen = Achsenabschnitte & Farbe & Legenden & Achsenbeschriftung & ...

Koordinatensysteme = kartesisch | logarithmisch | Polarkoord. | Kartenproj. | ...

Facettierung = small multiples | lattice plot | plot | ...

Theme = Font & Gitterlinien & Hintergrundfarben & Layout von Text & ...

Kontingenztafel und bedingte Häufigkeitsverteilung



Für relative Häufigkeiten ersetze h_{ij} durch $f_{ij} = \frac{h_{ij}}{n}$.

Bedingte Häufigkeitsverteilung: $f(Y=b_j \mid X=a_i)$

$f_Y(b_1 \mid a_i) = \frac{h_{ij}}{h_{i\bullet}}$, ..., $f_Y(b_m \mid a_i) = \frac{h_{im}}{h_{i\bullet}}$

Bedingte Odds / Odds ratio

Bedingte Odds: $y(1,2 \mid X=a_i) = \frac{h_{i1}}{h_{i2}}$

Relative Chancen (Odds ratio):

$y(1,2 \mid X=a_i, X=a_j) = \frac{y(1,2 \mid X=a_i)}{y(1,2 \mid X=a_j)} = \frac{h_{j2} \cdot h_{i1}}{h_{i2} \cdot h_{j1}}$

! Symmetrisches Maß und Risikofaktor

$y > 1$: Odds in $X=a_i$ höher als in $X=a_j$

χ^2-Koeffizient

(Geht auf allen Skalen, aber nur sinnvoll für nominal.)

Zusammenhangsmaß zum quantifizieren von "Abstand" zwischen beobachteten und erwarteten Häufigkeiten.

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}} = n \cdot \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(f_{ij} - \tilde{f}_{ij})^2}{\tilde{f}_{ij}}$$

Erwartete absolute / relative Häufigkeit

Wenn X, Y unabhängig sind, erwarten wir: absolute Häufigkeit: $\tilde{h}_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}$

relative Häufigkeit: $\tilde{f}_{ij} = f_{i\bullet} \cdot f_{\bullet j}$

Odds-Update

$$\underbrace{\frac{\mathbb{P}[B \mid A]}{\mathbb{P}[B^c \mid A]}}_{\text{a-posteriori-Verhältnis}} = \underbrace{\frac{\mathbb{P}[A \mid B]}{\mathbb{P}[A \mid B^c]}}_{\text{neue Information}} \cdot \underbrace{\frac{\mathbb{P}[B]}{\mathbb{P}[B^c]}}_{\text{a-priori Verhältnis}}$$

Satz

Für eine Kontingenztafel der Form

a	b
c	d

gilt $\chi^2 = \frac{n \cdot (a \cdot d - c \cdot b)^2}{(a+b)(a+c)(c+d)(b+d)}$

Lorenzkurve

Das Merkmal darf nur positive Werte annehmen.

$x_{(1)}, \dots, x_{(n)}$ sei die geordnete Stichprobe.

Die Lorenzkurve verbindet Punktepaare bestehend aus den Teilsummen von $x_{(i)}$ (d.h. $\sum_{i=0}^{k} x_{(i)}$) und dem relativen Anteil an Individuen, die diese Teilsumme besitzen.

Berechnung

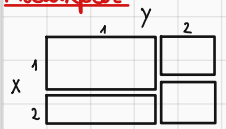
$u_{(0)} := 0$, $v_{(0)} := 0$, $j = 1, \dots, n$.

$u_{(j)} := \frac{j}{n}$ (Aufteilung x-Achse)

$v_{(j)} := \sum_{i=1}^{j} x_{(i)} \cdot \frac{1}{\sum_{i=1}^{n} x_{(i)}}$ ← (y-Werte)

$v_{(1)}, \dots, v_{(n)}$ ist monoton steigend

Mosaikplot



Breite vom Kasten $Y=j$ ist $\frac{h_{\bullet j}}{n}$

Höhe vom Kasten ij ist $\frac{h_{ij}}{h_{\bullet j}}$

(Korrigierter) Kontingenzkoeffizient

! Misst nur Stärke des Zusammenhangs von X und Y. Nicht die Richtung, wie bei γ

Normierung von χ^2: Kontingenzkoeffizient $K := \sqrt{\frac{\chi^2}{n + \chi^2}}$, $K \in [0, \sqrt{\frac{\min\{k,m\}-1}{\min\{k,m\}}}]$

Korrigierter Kontingenzkoeffizient $K^* := \frac{K}{\sqrt{\frac{\min\{k,m\}-1}{\min\{k,m\}}}}$, $K^* \in [0,1]$

Sensitiviät und Spezifität

$Y \in \{0,1\}$ (Zielgröße), X mindestens ordinalskaliert

$Y=1$: "positiver" Fall, $Y=0$: "negativer" Fall

Sei $\hat{y}_i$ die Prognose für y_i auf Basis von x_i. Es soll gelten

$\hat{y}_i = 1 \Leftrightarrow x_i \geq c$.

	$y_i = 0$	$y_i = 1$	
Vorhersage $\hat{y}_i = 0$	wahr negativ	falsch negativ	# negative Vorhersagen
Vorhersage $\hat{y}_i = 1$	falsch positiv	wahr positiv	# positive Vorhersagen
	# negative	# positive	

Sensitivität: $TPR(c) = f(\hat{y}=1 \mid y=1) = f(x \geq c \mid y=1) = \frac{\#\text{wahr pos.}}{\#\text{positiv}}$

(Wenn theoretisch: $\mathbb{P}[\hat{y}=1 \mid y=1]$ z.B. $\mathbb{P}[\text{Test pos.} \mid \text{Krank}]$)

Spezifität: $TNR(c) = f(\hat{y}=0 \mid y=0) = 1 - f(x \geq c \mid y=0) = \frac{\#\text{wahr neg.}}{\#\text{negativ}} = 1 - FPR(c)$

(Wenn theoretisch: $\mathbb{P}[\hat{y}=0 \mid y=0]$ z.B. $\mathbb{P}[\text{Test neg.} \mid \text{nicht krank}]$)

False pos. rate: $FPR(c) = f(\hat{y}=1 \mid y=0) = f(x \geq c \mid y=0) = \frac{\#\text{falsch pos.}}{\#\text{negativ}}$

Gini-Koeffizient

Maßzahl, die das Ausmaß der Konzentration beschreibt.

Definiert als $G = 2 \cdot F \in [0, \frac{n-1}{n}]$, wobei F die Fläche zwischen $y=x$ und der Lorenzkurve ist. $G = 1 - \frac{1}{n} \cdot \sum_{i=0}^{n} (v_{(i-1)} + v_{(i)})$

$G = \frac{n-1}{n} - \frac{2}{n} \sum_{i=1}^{n-1} v_{(i)}$. Normiert: $G^+ = \frac{n}{n-1} \cdot G \in [0,1]$

$G^+ = 0$ bedeutet keine Konzentration (Gleichverteilung)

$G^+ = 1$ bedeutet volle Konzentration (Monopol)

Merke

Lorenzkurve muss monoton steigend, stetig, kleiner $y=x$, und eine Funktion sein

Startet in (0,0), endet in (1,1)

ROC-Kurve

Die ROC-Kurve zeigt die Zuverlässigkeit der Vorhersagen für alle möglichen Schwellenwerte c an.

Verbindet die Punkte $(FPR(c), TPR(c))$ $\forall c \in [x_{(1)}, x_{(n)}]$

Für $c < x_{(1)} \Rightarrow \hat{y}_i = 1 \ \forall i \Rightarrow (FPR(c), TPR(c)) = (1,1)$

Für $c > x_{(n)} \Rightarrow \hat{y}_i = 0 \ \forall i \Rightarrow (FPR(c), TPR(c)) = (0,0)$

AUC-Maß

Maß zur Bewertung der ROC Kurve.

AUC := Fläche unter der ROC-Kurve

> perfekte Trennbarkeit der Y-Gruppen durch c bedeutet AUC = 1

> X, Y unabhängig $\Rightarrow$ AUC $\approx$ 0.5

! Behandelt Spezifität & Sensitivität gleichwertig und beachtet ppV bzw. npV nicht.

Positiv / negativ prädikater Wert

$ppV := f(y=1 \mid \hat{y}=1) = \frac{\#\text{wahr pos.}}{\#\text{pos. Vorhersagen}}$

$npV := f(y=0 \mid \hat{y}=0) = \frac{\#\text{wahr neg.}}{\#\text{neg. Vorhersagen}}$

Faltungen

Sind X_1, X_2 unabhängig, so gilt $f_{X_1+X_2}(z) = \int_{\mathbb{R}} f_{X_1}(x_1) \cdot f_{X_2}(z - x_1) \, d\mu(x_1)$

! Grenzen bestimmen sich dadurch, dass $x_1 \in \mathcal{T}_{X_1}$ und $z - x_1 \in \mathcal{T}_{X_2}$ gelten muss.

Beispiel

Seien $X_1 \sim P(\lambda_1)$, $X_2 \sim P(\lambda_2)$ und unabhängig. Wir betrachten $\bar{Y} = X_1 + X_2$.

$$P(Y=n) = P(X_1 + X_2 = n) = \sum_{k=0}^{n} P(X_1 = k) \cdot P(X_2 = n-k) = \underbrace{\left(\sum_{k=0}^{n} \binom{n}{k} \cdot \lambda_1^k \cdot \lambda_2^{n-k} \right)}_{= (\lambda_1 + \lambda_2)^n} \cdot \frac{e^{-(\lambda_1+\lambda_2)}}{n!}$$

Eigenschaften von Kovarianz

(i) $Cov(X,X) = Var[X]$

(ii) $|Cov(X,Y)| \leq Var(X) \cdot Var(Y)$

(iii) $Cov(aX+b, cY+d) = a \cdot c \cdot Cov(X,Y)$

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

v) $Var(X+Y) = Var(X) + Var(Y) + 2 \cdot Cov(X,Y)$

vi) $\Rightarrow$ X, Y sto. unabhängig $\Rightarrow Var(X+Y) = Var(X) + Var(Y)$

Kovarianz / Korrelation

Seien $X, Y: \Omega \to \mathbb{R}$ zwei ZV mit $Var[X] > 0$ und $Var[Y] > 0$, dann ist die Kovarianz $Cov(X,Y) := E[(X - E(X)) \cdot (Y - E(Y))]$ und die Korrelation $\rho(X,Y) := \frac{Cov(X,Y)}{\sqrt{Var(X)} \sqrt{Var(Y)}}$

! Unabhängig $\not\Leftrightarrow$ Unkorreliert

$|\rho(X,Y)| = 1 \Leftrightarrow Y = a + bX$, $a, b \in \mathbb{R}$

Schiefe

Ist X^3 quasi-int.bar, dann heißt

$\gamma(X) := \frac{\mathbb{E}[(X-\mathbb{E}[X])^3]}{\sqrt{Var[X]^3}}$ Schiefe von X.

$\gamma(X)=0$: symmetrisch $\gamma(X)<0$: rechtsteil

$\gamma(X)>0$: linkssteil

Kurtosis

Ist X^4 quasi-int.bar, dann heißt

$K(X) := \frac{\mathbb{E}[(X-\mathbb{E}(X))^4]}{Var[X]^2}$ Kurtosis von X

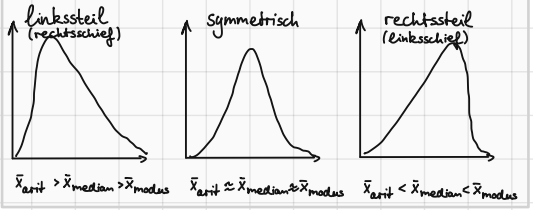
$K^*(X) = K(X) - 3$ heißt Exzess-Kurtosis.

$K^*(X) = 0$: mesokurtisch

$K^*(X) > 0$: leptokurtisch (viele extreme Werte)

$K^*(X) < 0$: platykurtisch (Wenig extreme Werte)

Verteilung charakterisieren



Datenformen

Querschnittdaten: Eine Beobachtung pro UE.

Zeitreihe: Mehrere Beobachtungen einer UE.

Längsschnittdaten: Mehrere Beobachtungen mehrerer UE.

Wahrscheinlichkeitsmaß

$(\Omega, \mathcal{F})$ sei ein Meßraum und $\mu: \mathcal{F} \to \overline{\mathbb{R}}$

i) $\mu(\emptyset) = 0$ ii) $\forall A \in \mathcal{F}: \mu(A) \geq 0$

iii) σ-Additivität: $\mu\left(\dot{\bigcup_{i=1}^{\infty}} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$

iv) Normiertheit: $\mu(\Omega) = 1$

σ-Algebra

$\mathcal{F}$ heißt σ-Algebra, wenn

1) $\Omega \in \mathcal{F}$ 2) $A \in \mathcal{F} \Rightarrow \bar{A} \in \mathcal{F}$

3) $A_i \in \mathcal{F}, i \in I \Rightarrow \bigcup_{i \in I} A_i \in \mathcal{F}$

Bedingte Momente

$$\mathbb{E}(X \mid Z=z) := \begin{cases} \sum_{x \in T_X} x \cdot \mathbb{P}(X=x \mid Z=z) & X \text{ diskret} \\ \int_{\mathbb{R}} x \cdot f_{X|Z}(x \mid Z=z)\, dx & X \text{ stetig} \end{cases}$$

! $\mathbb{E}(X|Z)$ ist ein ZV

$\mathbb{E}(X \mid Z=z) = \mathbb{E}(X|z)(z)$ ist ein Zahlenwert

$$Var(X \mid Z=z) := \mathbb{E}\left(X - \mathbb{E}(X \mid Z=z) \mid Z=z\right) = \begin{cases} \sum_{x \in T_X} (x - \mathbb{E}(X|Z=z))^2 \cdot \mathbb{P}(X=x \mid Z=z) & X \text{ diskret} \\ \int_{\mathbb{R}} (x - \mathbb{E}(X|Z=z))^2 f_{X|Z}(x \mid Z=z)\, dx & X \text{ stetig} \end{cases}$$

Satz vom iterierten Erwartungswert

Für beliebige ZV X, Y und Funktion f gilt:

$\mathbb{E}[\mathbb{E}(f(X) \mid Z)] = \mathbb{E}(f(X))$

Satz von der totalen Varianz

Für beliebige ZV X, Y gilt:

$Var(X) = \mathbb{E}(Var(X|Z)) + Var(\mathbb{E}(X|Z))$

↑ erwartete bedingte Varianz ↑ Varianz des bedingten Erwartungswertes

Markov- und Chebyshev-Ungleichungen

Sei $X: \Omega \to \mathbb{R}$ eine reelle ZV. Dann gilt

$\forall \varepsilon > 0: \mathbb{P}(|X| > \varepsilon) \leq \frac{1}{\varepsilon^n} \cdot \mathbb{E}[|X|^n]$

Markov-Ungleichung ($n=1$): $\mathbb{P}(|X| > \varepsilon) \leq \frac{1}{\varepsilon} \cdot \mathbb{E}[|X|]$

Chebyshev-Ungleichung ($n=2$): $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \cdot Var(X)$

Jensen-Ungleichung

Sei X eine int.bare ZV und $g: \mathbb{R} \to \mathbb{R}$ konvex. Dann gilt

$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$. Ist $f: \mathbb{R} \to \mathbb{R}$ konkav, so gilt $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$

Zentraler Grenzwertsatz

Seien $(X_n)_{n \in \mathbb{N}}$ i.i.d. ZVen mit $\mathbb{E}[X_n] = \mu$ und $Var[X_n] = \sigma^2 < \infty$. Sei $S_n = \sum_{i=1}^{n} X_i$. Dann gilt:

1) $\lim_{n\to\infty} \mathbb{P}\left[\frac{S_n - n\mu}{\sqrt{n \cdot \sigma^2}} \leq x\right] = \mathcal{N}(0,1)$ bzw. $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{\mathcal{D}} N(0,1)$ bzw. $S_n \dot{\sim} \mathcal{N}(n\mu, n\sigma^2)$ $\mathbb{P}(S_n \leq x) = \Phi\left(\frac{x - n\cdot\mu}{\sqrt{n\cdot\sigma^2}}\right)$

2) $\lim_{n\to\infty} \mathbb{P}\left[\sqrt{n} \cdot \frac{(\bar{X}_n - \mu)}{\sigma} \leq x\right] = \mathcal{N}(0,1)$ bzw. $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1)$ bzw. $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$ $\mathbb{P}(\bar{X}_n \leq x) = \Phi\left(\frac{x-\mu}{\sqrt{\frac{\sigma^2}{n}}}\right)$

Verteilungen

Verteilung	Verteilungsfunktion
$\mathcal{U}(a,b)$	$F_X(x) = \frac{x-a}{b-a} \cdot \mathbb{1}_{[a,b)}(x) + \mathbb{1}_{[b,\infty)}(x)$
$\mathcal{N}(\mu,\sigma^2)$	—
Cauchy	$F_X(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}$
Gamma $Ga(a,b)$ $\Gamma(a,b)$	$F_X(x) = \frac{1}{\Gamma(a)} \cdot \int_{-\infty}^{bx} t^{a-1} \cdot e^{-t} \cdot \mathbb{1}_{[0,\infty)}(t)\, dt$ $\quad \Gamma(a) := \int_0^\infty x^{a-1} e^{-x}\, dx$
	! $Exp(\lambda) = \Gamma(a=1, b=\lambda)$. $\quad \chi^2(d) = \Gamma(a=\frac{d}{2}, b=\frac{1}{2})$
Beta (α,β)	$\frac{1}{B(a,b)} \cdot \int_0^x t^{a-1} \cdot (1-t)^{b-1}\, dt \cdot \mathbb{1}_{[0,1]}(x)$ $\quad B(a,b) := \frac{\Gamma(a)\cdot\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx$
$Exp(\lambda)$	$F_X(x) = (1 - e^{-\lambda x}) \cdot \mathbb{1}_{[0,\infty)}$: $\mathbb{P}$(Wartezeit zwischen zwei Erfolgen einer $Poi(\lambda)$ verteilten ZV)
$Bin(n,\pi)$	$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} \cdot \pi^k \cdot (1-\pi)^{n-k} \cdot \mathbb{1}_{[0,n]}(x) + \mathbb{1}_{(n,\infty)}(x)$
$Poi(\lambda)$	$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \mathbb{P}(X=k) = \frac{\Gamma(\lfloor x+1 \rfloor, \lambda)}{\lfloor x \rfloor !}$: $\mathbb{P}$(x Erfolge in bestimmter Zeit) λ ist die Rate, mit der Erfolge in der bestimmten Zeit auftreten
$Geom(\pi)$	$F_X(x) = \left(1 - (1-\pi)^{\lfloor x \rfloor}\right) \cdot \mathbb{1}_{[1,\infty)}$: $\mathbb{P}$(genau x Versuche für ersten Erfolg) einer Bernoulli-ZV
$Geom_0(\pi)$	$F_X(x) = \left(1 - (1-\pi)^{\lfloor x \rfloor + 1}\right) \cdot \mathbb{1}_{[0,\infty)}$: $\mathbb{P}$(genau x Fehlversuche vor ersten Erfolg) einer Bernoulli-ZV

Konvergenzen

$X_n \xrightarrow{f.s.} X$: $\mathbb{P}(\{\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$

$X_n \xrightarrow{\mathbb{P}} X$: $\forall \varepsilon > 0: \lim_{n\to\infty} \mathbb{P}[|X_n - X| > \varepsilon] = 0$

$X_n \xrightarrow{r} X$: $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0$

$X_n \xrightarrow{\mathcal{D}} X$: $\forall x \in \{x \in \mathbb{R} \mid F \text{ stetig in } x\}: F_{X_n}(x) \xrightarrow{n\to\infty} F_X(x)$

Hypergeometrische Verteilung

Wahrscheinlichkeit für x Erfolge beim "ziehen" von n Elementen aus einer Stichprobe der Größe N, wobei es insgesamt M günstige Elemente gibt.

X ist hypergeometrisch verteilt $X \sim H(N, M, n)$, wenn $p_{N,M,n}(X=x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$, für

$x \in \{\max\{0, n-(N-M)\}, \ldots, \min\{n, M\}\}$

Dann ist $F_X(x) = \sum_{K=0}^{\lfloor x \rfloor} p_{N,M,n}(X=k) \cdot \mathbb{1}_{[0,\infty)}$

Farbskalentypen

- Qualitativ: (eher) nur für nominales Skalenniveau.
- Sequentiell: mindestens ordinales Skalenniveau.
- Divergent: mindestens ordinales Skalenniveau mit "neutralem" mittleren Wert.

Sinnvolle Zusammenhangsmaße

X, Y nominal → χ^2, Odds ratio, Kontingenzkoeffizient

X, Y ordinal → Rangkorr. nach Spearman r_{SP}

X, Y metrisch → Korrelation Pearson r_{BP}

Komb.	mit Wdh. mit Zurückl.	ohne Wdh. ohne Zurückl.
ohne Rhf.	$\binom{n}{m}$	$\binom{n+m-1}{m}$
mit Rhf.	$\frac{n!}{(n-m)!}$	n^m
Permutation	$n!$	$\frac{n!}{n_1! \cdot \ldots \cdot n_s!}$

Charakteristische / Momenterzeugende Funktion

Sei X eine ZV. Die Funktion $\varphi_X: \mathbb{R} \to \mathbb{C}$ mit

$\varphi_X(t) := \mathbb{E}[\exp(itX)] = \int \exp(itX)\, d\mathbb{P} = \int f_X(x) \cdot \exp(itx)\, d\mu$

heißt charakteristische Fkt. von X. Die Funktion $M: \mathbb{D} \to \mathbb{R}$

$M(t) := \mathbb{E}[\exp(tX)] = \int f_X(x) \cdot \exp(tx)\, d\mu$ heißt momenterzeug. Fkt.

Dichtetransformationssatz

Sei $X: \Omega \to \mathbb{R}$ eine ZV mit stetiger Verteilungsfkt. F_X und Dichte $f_X(x) = \frac{\partial F_X(x)}{\partial x}$ bzgl. des λ-Maßes. Sei $g: \mathbb{R} \to \mathbb{R}$ bijektiv und stetig diff.bar mit $g'(x) \neq 0$. Dann hat $g \circ X$ die Dichte $f_{g\circ X}(y) = f_X(g^{-1}(y)) \cdot |(g^{-1})'(y)|$