Angewandte Stochastik

Prof. Dr. Evgeny Spodarev | Vorlesungskurs | 5. Thema

Heutiges Thema

- Beschreibung von bivariaten Datensätzen

Einleitung

- Betrachten wir im Folgenden Datensätze bestehend aus 2 Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$, die als Realisierungen von stochastischen Stichproben $(X_1, \ldots, X_n)$ und $(Y_1, \ldots, Y_n)$ aufgefasst werden, wobei $X_1, \ldots, X_n$ unabhängige identisch verteilte Zufallsvariablen mit $X_i \stackrel{d}{=} X \sim F_X$, $Y_1, \ldots, Y_n$ unabhängige identisch verteilte Zufallsvariablen mit $Y_i \stackrel{d}{=} Y \sim F_Y$ sind. Wir betrachten hier ausschließlich quantitative Merkmale X und Y.

Einleitung

- Es wird ein Zusammenhang zwischen X und Y vermutet, der an Hand von (konkreten) Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ näher untersucht werden soll. Mit anderen Worten, wir interessieren uns für die Eigenschaften der bivariaten Verteilung $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ des Zufallsvektors $(X, Y)^T$.

Visualisierung

Um die Verteilung von $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ zu visualisieren, betrachten wir drei Möglichkeiten:

1. *Streudiagramme*
2. *Zweidimensionale Histogramme*
3. *Kerndichteschätzer* (im Falle eines absolut stetig verteilten Zufallsvektors $(X, Y)^T$)

1. *Streudiagramme* sind die erste sehr einfache und intuitive Visualisierungsmöglichkeit von bivariaten Daten.

- Um ein Streudiagramm zu erstellen, plottet man die "Punktwolke" $(x_i, y_i)_{i=1,\ldots,n}$ auf einer Koordinatenebene im $\mathbb{R}^2$.
- Dabei zeigt die Form der Punktwolke, ob ein linearer ($y = ax + b$) bzw. polynomialer ($y = P_d(x)$) Zusammenhang in den Daten zu erwarten ist.
- Später werden solche Zusammenhänge im Rahmen der Regressionstheorie untersucht (vgl. die einfache lineare Regression).

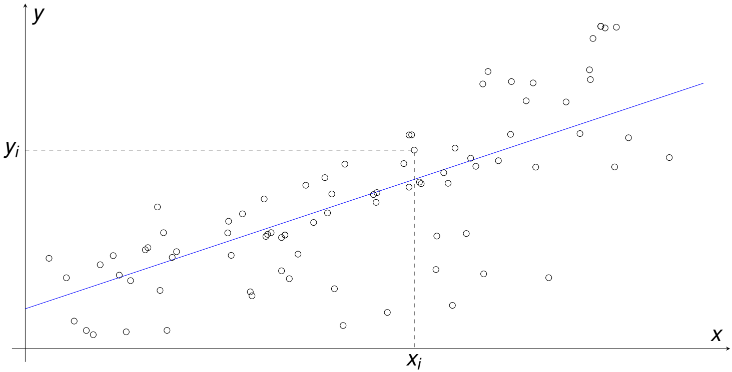

Figure: Punktwolke

2. *Zweidimensionale Histogramme* dienen der Darstellung der bivariaten Zähldichte $p(x, y)$ des Zufallsvektors (X, Y), falls er diskret verteilt ist, bzw. seiner Dichte $f(x, y)$ im Falle einer absolut stetigen Verteilung von (X, Y) aus den Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$. Dabei teilt man den Wertebereich von X in Intervalle

$$[c_{i-1}, c_i)\,, \quad i = 1, \ldots, k\,, \quad -\infty = c_0 < c_1 < \ldots < c_k = +\infty$$

und den Wertebereich von Y in Intervalle

$$[e_{i-1}, e_i)\,, \quad i = 1, \ldots, m\,, \quad -\infty = e_0 < e_1 < \ldots < e_m = +\infty\,.$$

Bezeichnen wir

$$h_{ij} = \#\{(x_k, y_k),\ k = 1, \ldots, n :\ x_k \in [c_{i-1}, c_i),\ y_k \in [e_{j-1}, e_j)\}$$

als die absolute Häufigkeit von (X, Y) in $[c_{i-1}, c_i) \times [e_{j-1}, e_j)$,

$f_{ij} = \frac{h_{ij}}{n}$ als die relative Häufigkeit.

Das zweidimensionale Histogramm setzt sich aus den Säulen mit Grundriss

$$[c_{i-1}, c_i) \times [e_{j-1}, e_j)$$

und Höhe

$$\frac{h_{ij}}{(c_i - c_{i-1})(e_j - e_{j-1})}$$

für das Histogramm absoluter Häufigkeiten bzw.

$$\frac{f_{ij}}{(c_i - c_{i-1})(e_j - e_{j-1})}$$

für das Histogramm relativer Häufigkeiten zusammen, damit das Volumen dieser Säulen h_{ij} bzw. f_{ij} ist.

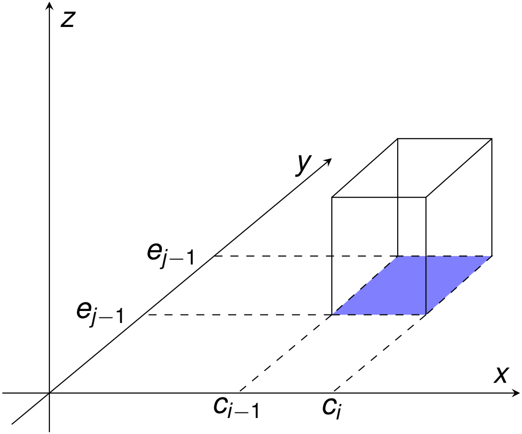

Figure: Zweidimensionales Histogramm

- Dabei hat solch ein Histogramm dieselben Vor- bzw. Nachteile wie ein eindimensionales, wenn es um die grafische Darstellung einer bivariaten Dichte $f(x, y)$ geht.
- Deshalb benutzt man oft *Kerndichteschätzer*, um eine glatte Darstellung zu bekommen.

3. *Zweidimensionale Kerndichteschätzer* haben die Form

$$\hat{f}(x, y) = \frac{1}{nh_1 h_2} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_1}\right) K\left(\frac{y - y_i}{h_2}\right)$$

für die Bandbreiten $h_1, h_2 > 0$, die Glättungsparameter sind.

- Dabei ist $K(\cdot)$ eine Kernfunktion (vgl. Video 4).
- Seine Eigenschaften übertragen sich aus dem eindimensionalen Fall.

Zusammenhangsmaße

Jetzt wird uns die Frage beschäftigen, in welchem Maße die Merkmale X und Y voneinander abhängig sind.

1. Um die $\operatorname{Cov}(X, Y) = \mathrm{E}(X - \mathrm{E}X)(Y - \mathrm{E}Y)$ aus den Daten zu schätzen, setzt man die sogenannte *empirische Kovarianz*

$$S_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)$$

ein. Dabei ist S_{xy}^2 jedoch von den Skalen von X und Y abhängig.

Zusammenhangsmaße

2. Um eine skaleninvariantes Zusammenhangsmaß zu bekommen, betrachtet man die empirische Variante des Korrelationskoeffizienten

$$\varrho(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sqrt{\text{Var}\ X} \cdot \sqrt{\text{Var}\ Y}}\,,$$

den sogenannten

Bravais-Pearson-Korrelationskoeffizienten

$$\varrho_{xy} = \frac{S_{xy}^2}{\sqrt{S_{xx}^2 \cdot S_{yy}^2}}\,,$$

wobei

$$S_{xx}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2\,, \qquad S_{yy}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

die Stichprobenvarianzen der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ sind.

Zusammenhangsmaße

Dabei erbt ϱ_{xy} alle Eigenschaften des Korrelationskoeffizienten $\varrho(X,Y)$:

(a) $|\varrho_{xy}| \leq 1$

(b) $\varrho_{xy} = \pm 1$, falls ein linearer Zusammenhang in den Daten $(x_i, y_i)_{i=1,\dots,n}$ vorliegt, d.h. alle Punkte (x_i, y_i), $i = 1, \dots, n$ liegen auf einer Gerade mit positivem (bei $\varrho_{xy} = 1$) bzw. negativem (bei $\varrho_{xy} = -1$) Anstieg.

Zusammenhangsmaße

(c) Wenn $|\varrho_{xy}|$ klein ist ($\varrho_{xy} \approx 0$), so sind die Datensätze unkorreliert. Dabei wird oft folgende grobe Einteilung vorgenommen:
Merkmale X und Y sind

- *"schwach korreliert"*, falls $|\varrho_{xy}| < 0.5$,
- *"stark korreliert"*, falls $|\varrho_{xy}| \geq 0.8$.

Ansonsten liegt ein mittlerer Zusammenhang zwischen X und Y vor.

Lemma

Für ϱ_{xy} gilt die alternative rechengünstige Darstellung

$$\varrho_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}_n^2\right)}} . \tag{1}$$

Zusammenhangsmaße

3.

- Einen alternativen Korrelationskoeffizienten erhält man durch den *Spearmans Korrelationskoeffizient*, wenn man die Stichprobenwerte x_i bzw. y_i in ϱ_{xy} durch ihre *Ränge* $\mathrm{rg}(x_i)$ bzw. $\mathrm{rg}(y_i)$ ersetzt, die als Position dieser Werte in den ansteigend geordneten Stichproben zu verstehen sind: $\mathrm{rg}(x_i) = j$, falls $x_i = x_{(j)}$ für ein $j \in \{1, \dots, n\}$, $\forall i = 1, \dots, n$. Es bedeutet, dass $\mathrm{rg}(x_{(i)}) = i \; \forall i = 1, \dots, n$, falls $x_i \neq x_j$ für $i \neq j$.

Zusammenhangsmaße

- Falls die Stichprobe $(x_1, \ldots, x_n)$ k identische Werte x_i (die sogenannten *Bindungen*) enthält, so wird diesen Werten der sogenannte Durchschnittsrang $\mathrm{rg}(x_i)$ zugewiesen, der als arithmetisches Mittel der k in Frage kommenden Ränge errechnet wird.
- Zum Beispiel findet folgende Zuordnung statt:

x_i	$(3, 1, 7, 5, 3, 3)$
$\mathrm{rg}(x_i)$	$(a, 1, 6, 5, a, a)$

wobei der Durchschnittsrang a von Stichprobeneintrag 3 gleich $a = {}^1\!/_3(2 + 3 + 4) = 3$ ist.

Zusammenhangsmaße

- Somit wird der sogenannte *Spearmans Korrelationskoeffizient* (Rangkorrelationskoeffizient) der Stichproben

$$(x_1, \dots, x_n) \quad \text{und} \quad (y_1, \dots, y_n)$$

als der *Bravais-Pearson-Koeffizient* der Stichproben ihrer Ränge

$$\big(\operatorname{rg}(x_1), \dots, \operatorname{rg}(x_n)\big) \quad \text{und} \quad \big(\operatorname{rg}(y_1), \dots, \operatorname{rg}(y_n)\big)$$

definiert:

$$\varrho_{sp} = \frac{\sum_{i=1}^n \big(\operatorname{rg}(x_i) - \overline{\operatorname{rg}}_x\big)\big(\operatorname{rg}(y_i) - \overline{\operatorname{rg}}_y\big)}{\sqrt{\sum_{i=1}^n \big(\operatorname{rg}(x_i) - \overline{\operatorname{rg}}_x\big)^2 \sum_{i=1}^n \big(\operatorname{rg}(y_i) - \overline{\operatorname{rg}}_y\big)^2}},$$

Zusammenhangsmaße

wobei

$$\overline{\mathrm{rg}}_x = \frac{1}{n}\sum_{i=1}^{n} \mathrm{rg}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{rg}(x_{(i)}) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n(n+1)}{2n} = \frac{n+1}{2},$$

$$\overline{\mathrm{rg}}_y = \frac{1}{n}\sum_{i=1}^{n} \mathrm{rg}(y_i) = \frac{n+1}{2}.$$

- Dieselbe Darstellung $\overline{\mathrm{rg}}_y$ gilt auch, wenn Bindungen vorhanden sind.
- Dieser Koeffizient misst monotone Zusammenhänge in den Daten. Aus den Eigenschaften der Bravais-Pearson-Koeffizienten folgt $|\varrho_{sp}| \leq 1$.

Zusammenhangsmaße

- Betrachten wir die Fälle $\varrho_{sp} = \pm 1$ gesondert:
  - $\varrho_{sp} = 1$ bedeutet, dass die Punkte $(\mathrm{rg}(x_i), \mathrm{rg}(y_i))$, $i = 1, \ldots, n$ auf einer Geraden mit positiver Steigung liegen. Da aber $\mathrm{rg}(x_i), \mathrm{rg}(y_i) \in \mathbb{N}$, kann diese Steigung nur 1 sein. Es bedeutet, dass dem kleinsten Wert in der Stichprobe $(x_1, \ldots, x_n)$ der kleinste Wert in $(y_1, \ldots, y_n)$ entspricht, usw., d.h., für wachsende x_i wachsen auch die y_i streng monoton: $x_i < x_j \Longrightarrow y_i < y_j \quad \forall i \neq j.$
  - Analog gilt dann für $\varrho_{sp} = -1$, dass $x_i < x_j \Longrightarrow y_i > y_j \quad \forall i \neq j.$

Zusammenhangsmaße

- Dies kann folgendermaßen zusammengefasst werden:
  - $\varrho_{sp} > 0$: gleichsinniger monotoner Zusammenhang (x_i groß $\Longleftrightarrow$ y_i groß)
  - $\varrho_{sp} < 0$: gegensinniger monotoner Zusammenhang (x_i groß $\Longleftrightarrow$ y_i klein)
  - $\varrho_{sp} \approx 0$: kein monotoner Zusammenhang.
- Da der Spearmans Korrelationskoeffizient nur Ränge von x_i und y_i betrachtet, eignet er sich auch für ordinale (und nicht nur quantitative) Daten.

Lemma

Falls die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ keine Bindung enthalten ($x_i \neq x_j,\ y_i \neq y_j\ \ \forall i \neq j$), dann gilt

$$\varrho_{sp} = 1 - \frac{6}{(n^2-1)n} \sum_{i=1}^{n} d_i^2 \,,$$

wobei $d_i = rg(x_i) - rg(y_i) \quad \forall i = 1, \ldots, n$.

Satz

1. Wenn die Merkmale X und Y linear transformiert werden:

$$f(X) = a_x X + b_x\,, \quad a_x \neq 0,\ b_x \in \mathbb{R},$$
$$g(Y) = a_y Y + b_y\,, \quad a_y \neq 0,\ b_y \in \mathbb{R},$$

   dann gilt $\varrho_{f(x)g(y)} = \operatorname{sgn}(a_x a_y) \cdot \varrho_{xy}$.

2. Falls Funktionen $f : \mathbb{R} \to \mathbb{R}$ und $g : \mathbb{R} \to \mathbb{R}$ beide monoton wachsend oder beide monoton fallend sind, dann gilt

$$\varrho_{sp}(f(x), g(y)) = \varrho_{sp}(x, y)\,.$$

   Falls f monoton wachsend und g monoton fallend (oder umgekehrt) sind, dann gilt $\varrho_{sp}(f(x), g(y)) = -\varrho_{sp}(x, y)$.

Bemerkung

1. Da lineare Transformationen monoton sind, gilt Aussage 1) auch für Spearmans Korrelationskoeffizienten ϱ_{sp}.
2. Der Koeffizient ϱ_{xy} erfasst lineare Zusammenhänge, während ϱ_{sp} monotone Zusammenhänge aufspürt.