Angewandte Statistik I

Dr. Uli Wannek

Skript erstellt von Alina Renz

Wintersemester 2017/2018

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Inhaltsverzeichnis

1 Beschreibende Statistik

- auch deskriptive oder empirische Statistik
- Daten sind oft komplex $\rightarrow$ Reduktion der Daten auf einige Kenngrößen:
  - Lagemaß
  - Streuung
  - Verteilung

1.1 Lagemaße

- Statistik sucht Übersicht, Vereinfachung (Erklärung)
- Hier eine (wichtigste erste) Kennzahl: zentraler Wert **Vorsicht:** bei gleichem Lagemaß können sich die zugrundeliegenden Daten unterscheiden

Modus

- Wert, der am häufigsten Vorkommt
- Es kann mehrere Modi geben und der Modus kann atypisch sein
- Modus
$$x_i \;:\quad h_i \geq h_j \quad \forall j$$

Median

- Mindestens eine Hälfte der Werte kleiner gleich
- Mindestens eine Hälfte der Werte größer gleich
- in sortierter Liste aller Daten x_i mit $i \in 1...n$ und $x_{(1)} <= ... <= x_{(n)}$ ist der Median
$$\tilde{x} = \begin{cases} x_{(\frac{N+1}{2})} & N \text{ ungerade} \\ \frac{1}{2}(x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}) & N \text{ gerade} \end{cases}$$
- Minimale Betragsabweichung: Sei $f(t) = \sum_{i=1}^{N} |x_i - t|$
  dann gilt
$$f(t) = \sum_{i=1}^{N} |x_i - t| \geq \sum_{i=1}^{N} |x_i - \tilde{x}| = f(\tilde{x}) \quad \forall t \in \mathbb{R}$$
- Median unter linearer Transformation:
$$y = ax + b \qquad \Rightarrow \qquad \tilde{y} = a\tilde{x} + b$$
- in Python: numpy stellt Funktion `median()` bereit

Arithemtischer Mittelwert

- Gegeben seien $i = \{1 \dots N\}$ Werte x_i, dann ist deren arithmetisches Mittel

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

- Minimale quadratische Abweichung: Sei $f(t) = \sum_{i=1}^{N}(x_i - t)^2$
  dann gilt
  $$f(t) \geq f(\bar{x}) \quad \forall \quad t \in \mathbb{R}$$

- in Python:
  - numpy-ndarray hat die Methode `mean()`
  - numpy hat die Funktion `np.mean()`

- Lineare Transformation des arithmetischen Mittels

$$y_i = a x_i + b \qquad \Rightarrow \qquad \bar{y} = a\bar{x} + b$$

- Summe von Mittelwerten:
  - Gegeben seien N Werte x_i mit Mittelwert $\bar{x}$ und M Werte y_j mit Mittelwert $\bar{y}$.
  - Dann ist das gemeinsame arithmetische Mittel
    $$\bar{z} = \frac{1}{N+M}(N \cdot \bar{x} + M \cdot \bar{y})$$
    im allgemeinen $\neq \frac{1}{2}(\bar{x} + \bar{y})$

- Anwendung:
  - *herkömmlicher* Mittelwert $N \cdot \bar{x} = \sum x_i$
  - Sanfte Berücksichtigung der „Fehler“ um den Mittelwert
  - Starke Berücksichtigung der weit außen liegenden „Fehler“

Geometrisches Mittel

- Anwendungsbeispiel Preissteigerung: Inflation beschreibt jährliche Preissteigerung bei etwa gleichbleibendem Warenkorb. Ein Jahr wird als Referenz festgelegt (darauf auf 100% normiert).

- Mittlere Wachstumsrate als geometrisches Mittel der einzelnen Wachstumsraten

$$\bar{x}_{geom} = \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N} = \left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}}$$

Harmonisches Mittel

- Anwendungsbeispiel: Jeden Tag tanken für 20 € bei sich ändernden Preisen. Welches ist der durchschnittliche Preis?
- Harmonisches Mittel sinnvoll bei Verhältniszahlen

$$\bar{x}_{harm} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{x_i}}$$

- Kann durchaus extreme Unterschiede zeigen wie im Beispiel $x = \{1, 4, 4\}$ mit *arithmetischem* Mittel 3 und *harmonischen* Mittel 2
- Anwendung Elektronik: Parallelschaltung von Widerständen

Ergebnis Lagemaß

- Das Lagemaß (Mittelwert) liefert **eine** wichtige Kennzahl für die Daten
- Mittelwerte im Vergleich

$$x_{min} \leq \bar{x}_{harm} \leq \bar{x}_{geom} \leq \bar{x}_{arithm} \leq x_{max}$$

1.2 Streuung

Nach dem Lageparameter als erste (wichtigste) Kennzahl beschreiben weitere Kennzahl(en) die Verteilung der Werte.

Spannweite

- Der Bereich vom kleinsten zum größten Wert

$$R = x_{max} - x_{min}$$

Empirische Varianz

$$\boxed{s^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

- Verschiebungssatz

$$s^2 = \overline{x^2} - \overline{x}^2$$

Stichprobenvarianz

- Die Stichprobenvarianz ist leicht unterschiedlich definiert

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

- Hintergrund
  - Die Anzahl der Freiheitsgrade ist $N - 1$, da Nebenbedingung $\sum_{i=1}^{N}(x_i - \bar{x}) = 0$
  - Für $N = 1$ ist die Varianz nicht definiert $\rightarrow$ anstatt 0 bei empirischer Varianz
  - Für große N ist sie asymptotisch gleich
- Eigenschaften
  - Unter der linearen Abbildung

    $$y_i = ax_i + b$$

    ergibt sich

    $$s_y^2 = a^2 s_x^2$$

Standardabweichung

$$s = +\sqrt{s^2}$$

- Unter der linearen Abbildung

  $$y_i = ax_i + b$$

  ergibt sich

  $$s_y = |a|\ s_x$$

Quantile

- Quantil = Grenzwert, vom dem der Anteil (das Quantum) an Datenwerten liegt
- Quartile
  - Unteres Quartil $x_{0.25} = 25\%$ der Werte
  - Oberes Quartil $x_{0.75} = 75\%$ der Werte
  - Median entspricht dem 50% Quartil
  - Der Interquartilsabstand ist ein Streuungsmaß, das robust gegen Ausreißert ist

$$d_Q = x_{0.75} - x_{0.25}$$

- Perzentile
  - 5% Perzentil: niedrigste 5% der Werte
  - 95% Perzentil: ohne oberste 5% der Werte
- Quantile sind einfach abzulesen aus der kumulativen Verteilungsfunktion

1.3 Graphische Darstellung

Ziel

- Datenreduktion durch Kenngrößen wie Mittelwert und Standardabweichung
- Struktur in den Daten erkennen durch Form der Verteilung und Form der Randverteilung
- Anschauliche Darstellung mit Feldertafeln oder 2D-Histogrammen

Histogramm

- viele Daten → von der Strichliste zum Histogramm
- Diskrete Ereignisse: Münze, Würfel, Karten, Körpergröße in 5 cm Schritten, ...
- Kontinuierliche Ereignisse: Körpergröße, Temperatur
- Klasseneinteilung (*binning*) führt wieder zurück auf den diskreten Fall
  - Sinnvoll für Informationsgehalt
  - Faustregel:
    - Bis zu 20 Klassen
    - Anzahl Klassen $n_{\text{bins}} \approx \sqrt{n}_{\text{data}}$ (für n_{data} bis 400)
    - Klassenbreite $w_{\text{bins}} = \frac{3.49 \cdot \sigma}{\sqrt[3]{n_{\text{data}}}}$ (Scotts Regel)
  - Dabei Breite Klassenbalken
    - möglichst konstant
    - sonst kodiere Anzahl in Fläche (nicht Höhe)
- Information
  - Häufigste Werte: „Modus“
  - Ausreißer
  - Mittelwert (=Schwerpunkt)
  - Wertebereich
  - Form der Verteilung

Empirische Kumulative Verteilungsfunktion F(x)

- Relative Häufigkeiten $h_i = \frac{n_i}{N}$ (beispielweise aus einem Histogramm)
- Der Größe nach sortieren
- Aufstapeln
- Schnelles Ablesen der Information:
  - Lage: Median, (Modus)
  - Bereich: Quartile, Perzentile
  - Form der Verteilung: S-Kurve (wenn Verteilung *ein* Maximum hat = „unimodal“)

- Kumulierter Mittelwert
  - Arithmetisches Mittel

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

$$\approx \bar{x'} = \frac{1}{\sum_{k=1}^{N_k} h_k}\sum_{k=1}^{N_k} h_k(x_k) \cdot x_k$$

  - Varianz

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

$$\approx {\sigma'}^2 = \frac{1}{\sum_{k=1}^{N_k} h_k - 1}\sum_{k=1}^{N_k} h_k(x_k) \cdot (x_k - \bar{x})^2$$

  - Standardabweichung

$$s = \sqrt{\sigma^2}$$

$$s' = \sqrt{\bar{\sigma'}^2}$$

Fehlerbalken

- Arithmetischer Mittelwert als Symbol
- Fehlerbalken $\pm$ Standardabweichung σ
- Ausreißer als Sterne

Boxplot

- Median als Strich
- Größe der Box: Qartile $Q_{1/4} \dots Q_{3/4}$
- Länge der Striche $Q_{1/4} - \frac{3}{2} \cdot (Q_{3/4} - Q_{1/4}) \ \dots \ Q_{3/4} + \frac{3}{2} \cdot (Q_{3/4} - Q_{1/4})$
  - Merkregel: Werte innerhalb $\pm 100\%$ - Quartile
- Ausreißer als Sterne +
- Python: `plt.boxplot`

Violinenplot

- Mehrere Daten in einer Liste
  - kein 2D numpy-array
  - daher Listen-„Krücke“
- `np.random.normal` gibt die Normalverteilung zurück (später mehr)
- Python: `plt.violinplot`

Vergleich zweier Stichproben

- Kennwerte, Histogramm-Stapel, kumulierte Verteilung
- Wertebereich - Beispiele
  - Nominale Daten: Münze Kopf/Zahl, gezogene Farbe
  - Ordinale Daten: Schadensklasse, Zufriedenheit (sehr/ja/nein/gar nicht) → Kategorien
  - Diskrete Daten: gewürfelte Zahl, Alter
  - Kontinuierliche Daten: Temperatur, Alter, prozentuale Werte
- Weitere Unterscheidung für nominale und ordinale Daten:
  - dichotom/binär/binomial
  - polytom/multinomial
  - Mittels Klasseneinteilung diskretisieren und Histogramm darstellen

Gute Praxis

- Achsen beschriften: `xlabel, ylabel`
- *ehrliche* Achsen
  - 0 mit einbeziehen, wenn absolute Zahlen
  - sonst deutlich kennzeichnen (Lücke)
  - entsprechende Zahlen bzw. tickmarks: `xaxis.set_ticklabels`
  - sinnvollen Bereich: `axis`
- Überschrift: `title`
  - in Veröffentlichungen Bildunterschrift mit Text; keine Überschrift
- Graphen beschriften: `label` und `legend`
- deutlich unterscheiden (auch für schwarz-weiß-Druck): Symbole, Linienstil, Farbe wenn nötig
- nicht überfrachten
- manchmal hilfreich:
  - Gitterlinien: `grid`
  - Spiegelstriche: `ax.tick_params(axis='y', direction='out')` und `ax.yaxis.tick_left()`

1.4 Mehrdimensionale Daten

Bivariate Stichproben

Beispiele

- Zwei Würfel
- Strom und Spannung
- Luftdruck und Höhe über Meer
- Regentropfen auf Blatt Papier (Koordinaten x und y)

Diskrete Verteilung

- Kontingenztafel der absoluten Häufigkeiten
- Vierfeldertafel: Spezialfall mit $n_1 = n_2 = 2$
- Randverteilung = Summe

Tabelle 1.1: Absolute Häufigkeiten mit Randverteilung

1 / 2	head	tail	sum
head	5	5	10
tail	4	6	10
sum	9	11	20

- Relative Häufigkeiten

$$f_{ij} = \frac{h_{ij}}{n} \qquad f \in [0, 1]$$

Tabelle 1.2: Relative Häufigkeiten mit Randverteilung

1 / 2	head	tail	sum
head	0.25	0.25	0.5
tail	0.2	0.3	0.5
sum	0.45	0.55	1

Kontinuierliche Verteilung

- Streudiagramm mit Randverteilung der einzelnen Komponenten (Dimensionen)
- Mittelwert: Randverteilung erlaubt Bestimmung des 2D-Mittelwerts
- Varianz: Randverteilung erlaubt Bestimmung der 2D-Streuung
- Dichteverteilung
  - Bei zu vielen Datenpunkten gibt es analog zum eindimensionalen Fall *Histogramm* die Möglichkeit in die dritte Dimension z die Häufigkeitsverteilung aufzutragen.
  - Matplotlib hat dazu `hist2d` farbkodiert
- `seaborn` als weitere Bibliothek zur Datenexploration

3-D Darstellung und höherdimensionale Daten

- zweidimensionale Basis: $y = f(x_1, x_2)$
- Graphische Darstellung
  - diskret: mit *Nadeln*
  - kontinuierlich: 3D-Graphik
- Höherdimensional: Streumatrix

1.5 Abhängigkeit

Zusammenhang von zusammengehörigen, mehrdimensionalen Daten - Beispiele:

- zwei Würfel: sollten unabhängig sein
- Strom und Spannung: festes Verhältnis gemäß Ohmschen Gesetz
- Luftdruck und Höhe über Meer: Barometrische Höhenformel
- Frage: Hängen x und y zusammen?

Korrelation (nach Pearson)

- Frage: Je größer x desto größer y?
  $\rightarrow$ Antwort: Produkt-Moment-Korrelation nach Pearson
- Vorgehen
  - Standardisieren
    * Abziehen des Mittelwerts
      $$x_i'' = x_i - \bar{x}, \qquad y_i'' = y_i - \bar{y}$$
    * Strecken auf Standardabweichung
      $$x_i' = \frac{x_i''}{\sigma_x}, \qquad y_i' = \frac{y_i''}{\sigma_y}$$
  - $\rightarrow$ Neue Verteilung x' und y'
    * Mittelwerte $\bar{x'} = 0$ bzw. $\bar{y'} = 0$
    * Standardabweichung $\sigma_x' = 1$ bzw. $\sigma_y' = 1$.
- Zusammenhang
  - Quadranten I und III
  - Quadranten II und IV
  - ABER: haben unterschiedliche Vorzeichen, daher sinnvoller:
- Definition Korrelationskoeffizient
  $$r_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} x_i' \cdot y_i'$$
- Eigenschaften des Korrelationskoeffizienten und der Kovarianz
  - $-1 \leq r_{XY} \leq +1$
  - $y = x \quad \Rightarrow \quad r_{XY} = +1$
  - $y = -x \quad \Rightarrow \quad r_{XY} = -1$
  - In ursprünglichen Koordinaten
    $$r_{XY} = \frac{\mathrm{Cov}_{XY}}{\sigma_X \sigma_Y}$$

- mit der Kovarianz

$$\mathrm{Cov}_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- Symmetrie: $\mathrm{Cov}_{XY} = \mathrm{Cov}_{YX}$ sowie $r_{XY} = r_{YX}$
- Varianz: $\mathrm{Var}(X) = \mathrm{Cov}_{XX}$

• Python: beachte den Faktor $\frac{1}{N-1}$,
  - der in der Kovarianz-Matrix berücksichtigt ist
  - jedoch nicht per default in `numpy.var()`
  - deshalb dann `ddof=1` (*Delta Degrees Of Freedom*) wählen

• Ergebnis der Korrelation nach Pearson
  - Bei guter Übereinstimmung: $y \simeq x$ $\qquad r \to 1$
  - Bei Betragsgleichheit mit umgekehrtem Vorzeichen: $y \simeq -x$ $\qquad r \to -1$
  - Bei grober Entsprechung: $\qquad 0 \leq r \leq 1$
  - Sind x und y unkorreliert, dann geht $\qquad r \to 0$

• Stärke (willkürlich)
  - bis 0.2 → sehr geringe Korrelation
  - bis 0.5 → geringe Korrelation
  - bis 0.7 → mittlere Korrelation
  - bis 0.9 → hohe Korrelation
  - über 0.9 → sehr hohe Korrelation

• **Obacht**: Korrelationskoeffizient trifft nur Aussage über **linearen** Zusammenhang
  - Nicht-linearer Zusammenhang kann eine klare Abhängigkeit besitzen
  - Korrelation beschreibt *nur* lineare Abhängigkeit
  - Lösungsansätze für Korrelationen
    * Jede Abhängigkeit kann stückweise als linear angesehen werden
    * Linearisierung durch Abbilden der Daten oder Bilden einer Modellfunktion aus der Theorie

• Bedeutung der Korrelation nach Pearson
  - $r = +1$ für Gleichheit $y = a \cdot x$
  - $r > 0$ für mittleren linearen Zusammenhang $y \sim x$
  - $r < 0$ für mittleren linearen Zusammenhang $y \sim -x$
  - $r = -1$ für Antikorrelation $y = -a \cdot x$
  - $r = 0$ wenn kein (linearer) Zusammenhang besteht
    * Eine Abhängigkeit y von x kann trotz $r = 0$ bestehen
    * Die Form der Abhängigkeit kann stark variieren

• Verhalten unter linearer Transformation

  - Unter der linearen Transformation

$$\begin{aligned} x' &= a \cdot x &+ b \\ y' &= c \cdot y &+ d \end{aligned}$$

  - bleibt betragsmäßig

$$r_{x'y'} = r_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} = \frac{\text{Cov}(x',y')}{\sigma_{x'} \sigma_{y'}}$$

    das Vorzeichen ändert sich, wenn a und c unterschiedliche Vorzeichen haben.

- Korrelation und Kausalität
  - Keine Aussage über die Ursache:

$$y(x)? \qquad x(y)? \qquad x(z) \text{ und } y(z) \qquad Zufall?$$

  - Schein-Korrelation durch verstecke Variable
    * Schulkinder: Geschicklichkeit korreliert mit Körpergröße
      → versteckte Variable „Alter“
  - Inhomogenitäts-Korrelation
    * Schuhgröße vs. Einkommen → Männer, Frauen
  - Schein-Korrelation durch Ausreißer/Standardisierung
    * zufällige Variation innerhalb Streuung
  - Oft bei Zeitreihen

Regression

- Voraussetzung: Zusammenhang von zusammengehörigen Daten nachgewiesen
  → Korrelationskoeffizient $\neq 0$
- Beispiele:
  - Strom und Spannung: festes Verhältnis gemäß Ohmschen Gesetz
  - Luftdruck und Höhe über Meer: barometrische Höhenformel
- Frage: Wie *stark* hängt y von x ab?
  - Beispiel: Fahrstrecke x und Fahrzeit t:
  - Zusammenhang: je mehr Zeit t, desto weiter die Strecke s
    * linear $x \propto t$
    * Proportionalitätsfaktor *Geschwindigkeit* v: $x = v \cdot t$

Lineare Regression

- Anwendung
  - Bei linearer Abhängigkeit $x = v \cdot t$ erlaubt der Proportionalitätsfaktor Vorhersagen
  - Mehrere Messungen: Mehrere (n) Messungen y_i zu verschiedenen Werten der x-Variable, hier Schrittperiodendauer (Zeiten, Längen, Alter, Höhe, ...)
  - Korrelation? → lineare Abhängigkeit: mehrere x-Werte (Periodendauern) und mehrere Messungen (Geschwindigkeit) → Ausgleichsgerade
- Methode der kleinsten Quadratischen Abweichungen

  $$\operatorname*{argmin}_{a,b} \sum_i (y_i - f_{a,b}(x_i))^2$$

  des linearen Zusammenhangs

  $$y = f_{a,b}(x) = a \cdot x + b$$

  führt zu

  $$a = \frac{\sum_i (y_i - \bar{y}) x_i}{\sum_i (x_i - \bar{x}) x_i} \qquad b = \bar{y} - a \cdot \bar{x}$$

  und entspricht

  $$a = r_{XY} \cdot \frac{\sigma_Y}{\sigma_X}$$

  Im Allgemeinen:

  $$r_{XY} = r_{YX}, \qquad a_{XY} \overset{\text{i. A.}}{\neq} a_{YX}, \qquad a_{XY} \overset{\text{i. A.}}{\neq} \frac{1}{a_{YX}}$$

Polynomiale Regression

- Nicht-lineares Beispiel: Wurfparabel
- Ergebnis (für Parabel 2. Grades):
  - Parabel wird relativ gut gefittet
  - die gefitteten Parameter stimmen gut mit den ursprünglichen überein
    * die Gravitationskonstante ist jedoch eine solche
    * Startgeschwindigkeit
    * Start-Zeit sollte 0 sein
- Ergebnis (für Parabel 14. Grades):

  Obwohl der Restfehler weniger wird gegenüber dem (korrekten) parabolischen Modell, ist das Ergebnis zweifelhaft
  - Ein Ball steigt nicht wieder
  - der quadratische Koeffizient ist unsinnig hoch
- Python: `np.polyfit` oder `scipy.optimize.curve_fit`
- Höherdimensional: Streumatrix