Stochastik für Informatiker, Vorlesung 21

Inhalt

- χ^2-Test
- Markov-Ketten
- Stochastische Matrizen

Lernziele

- Die Anwendungsbereiche des χ^2-Test kennen, den Test durchführen und interpretieren können
- Markov-Ketten und stochastische Matrizen kennen

Vorkenntnisse Verteilungen, Quantile, χ^2-Verteilung, bedingte Wahrscheinlichkeiten, Matrizenrechnung.

χ^2-Test

Der χ^2-Test kann in folgenden Situationen angewandt werden:

- Test ob die Daten einer bestimmten Verteilung folgen
- Test auf Unabhängigkeit

Grundprinzip (Test auf feste Verteilung):

- Gegeben: Daten $x_1, ..., x_n$, gruppiert in k Klassen mit Häufigkeiten $N_1, ..., N_k$.
- Grundannahme: Realisierungen unabhängiger Zufallsvariablen
- Nullhypothese: Die Daten folgen einer bestimmten Verteilung (Normalverteilung, uniforme Verteilung....)
- Falls die Klassengrößen alle hinreichend groß sind, so folgt die mittlere quadratische Abweichung der gemessenen Häufigkeiten von den theoretischen einer χ^2-Verteilung.

χ^2-Test auf Verteilung

1. Gruppieren der Daten in k Klassen $A_1, ..., A_k$, wobei die Klassen A_i Teilmengen von $\mathbb{R}$ sind, so dass $A_i \cap A_j = \emptyset$ gilt, und so dass jeder Messwert x_j in einer Klasse A_i enthalten ist. Bestimmung der Häufigkeiten

$$N_i := |\{j : x_j \in A_i\}|, i = 1, ..., k.$$

2. Aufstellen der Nullhypothese: H_0 : Die Daten folgen einer bestimmten Verteilung (z.B. Normalverteilung,...). Ev. Schätzung der Parameter der Verteilung.
3. Wahl des Fehlerniveaus $\alpha \in (0, 1)$
4. Berechnung des Freiheitsgrades: $f = k - m - 1$, wobei m die Anzahl geschätzert Parameter in der angenommenen Verteilung ist.
5. Bestimmung des $1 - \alpha-$Quantils $\chi^2_{1-\alpha,f}$ der χ^2-Verteilung mit Parameter $k - m - 1$ als Vergleichswert.

χ^2-Test auf Verteilung

6. Berechung der theoretischen Häufigkeiten unter der Nullhypothese:

$$F_i = n \cdot \mathbb{P}(X \in A_i \mid H_0), i = 1, ..., k$$

7. Berechnung des Testwerts

$$\chi^2 := \sum_{i=1}^{k} \frac{(N_i - F_i)^2}{F_i}.$$

8. Vergleich des Testwerts χ^2 mit dem Quantil $\chi^2_{1-\alpha,f}$. Entscheidungsregel:
   - Ist $\chi^2 > \chi^2_{1-\alpha,f}$ so wird H_0 verworfen.
   - Andernfalls wird H_0 angenommen.

Beispiel 10.6:

Gruppe	A	B	C	D	E	F	G	H
Häufigkeiten	2	4	2	3	3	7	5	4

Nullhypothese: Die Daten sind diskret gleichverteilt

Fehlerniveau $\alpha = 0.05$

Freiheitsgrad 8-1=7 (8 Klassen, keine geschätzen Parameter)

Theoretische Häufigkeiten: $F_i = 30/8 = 3.75, i = 1, ..., 8$
Testwert

$$\chi^2 = \frac{1}{3.75}\Big((2-3.75)^2 + (4-3.75)^2 + ... + (4-3.75)^2\Big) = 5.2$$

Tabelle: Quantile der χ^2–Verteilung

Tabelle der Quantile der Chiquadrat- und Gamma-Verteilung

α-Quantile $\chi^2_{n;\alpha}$ der Chiquadrat-Verteilungen $\boldsymbol{\chi}^2_n = \Gamma_{1/2,n/2}$ mit n Freiheitsgraden. $\chi^2_{n;\alpha}$ ist der Wert $c > 0$ mit $\boldsymbol{\chi}^2_n([0,c]) = \alpha$. Durch Skalierung erhält man die Quantile der Gamma-Verteilung $\Gamma_{\lambda,r}$ mit $\lambda > 0$ und $2r \in \mathbb{N}$. Notation: $^{-5}3.9 \equiv 3.9 \cdot 10^{-5}$.

$\alpha =$	0.005	0.01	0.02	0.05	0.1	0.9	0.95	0.98	0.99	0.995
$n = 1$	$^{-5}3.9$	$^{-4}1.6$	$^{-4}6.3$	$^{-3}3.9$	0.0158	2.706	3.842	5.412	6.635	7.879
2	0.0100	0.0201	0.0404	0.1026	0.2107	4.605	5.991	7.824	9.210	10.60
3	0.0717	0.1148	0.1848	0.3518	0.5844	6.251	7.815	9.837	11.34	12.84
4	0.2070	0.2971	0.4294	0.7107	1.064	7.779	9.488	11.67	13.28	14.86
5	0.4117	0.5543	0.7519	1.145	1.610	9.236	11.07	13.39	15.09	16.75
6	0.6757	0.8721	1.134	1.635	2.204	10.65	12.59	15.03	16.81	18.55
7	0.9893	1.239	1.564	2.167	2.833	12.02	14.07	16.62	18.48	20.28
8	1.344	1.646	2.032	2.733	3.490	13.36	15.51	18.17	20.09	21.95
9	1.735	2.088	2.532	3.325	4.168	14.68	16.92	19.68	21.67	23.59
10	2.156	2.558	3.059	3.940	4.865	15.99	18.31	21.16	23.21	25.19
11	2.603	3.053	3.609	4.575	5.578	17.28	19.68	22.62	24.72	26.76
12	3.074	3.571	4.178	5.226	6.304	18.55	21.03	24.05	26.22	28.30
13	3.565	4.107	4.765	5.892	7.042	19.81	22.36	25.47	27.69	29.82
14	4.075	4.660	5.368	6.571	7.790	21.06	23.68	26.87	29.14	31.32
15	4.601	5.229	5.985	7.261	8.547	22.31	25.00	28.26	30.58	32.80
16	5.142	5.812	6.614	7.962	9.312	23.54	26.30	29.63	32.00	34.27
17	5.697	6.408	7.255	8.672	10.09	24.77	27.59	31.00	33.41	35.72

Beispiel 10.6:

Gruppe	A	B	C	D	E	F	G	H
Häufigkeiten	2	4	2	3	3	7	5	4

Nullhypothese: Die Daten sind diskret gleichverteilt

Fehlerniveau $\alpha = 0.05$

Freiheitsgrad $f = 8 - 1 = 7$ (8 Klassen, keine geschätzen Parameter)

Theoretische Häufigkeiten: $F_i = 30/8 = 3.75, i = 1, ..., 8$

Testwert

$$\chi^2 = \frac{1}{3.75}\Big((2 - 3.75)^2 + (4 - 3.75)^2 + ... + (4 - 3.75)^2\Big) = 5.2$$

Das 0.95-Quantil der χ^2-Verteilung mit Parameter $n = 7$ ist

$\chi^2_{0.95,7} = 14.07$

Die Nullhypothese wird also nicht abgelehnt.

Beispiel 10.6: χ^2-Test auf Normalverteilung

Daten: Umsätze von 197 börsennotierten Unternehmen (Quelle: Wikipedia)

Klasse	**Intervall**		**Beobachtete Häufigkeit**
j	**über**	**bis**	**n_j**
1	...	0	0
2	0	5000	148
3	5000	10000	17
4	10000	15000	5
5	15000	20000	8
6	20000	25000	4
7	25000	30000	3
8	30000	35000	3
9	35000	...	9
Summe			197

Nullhypothese: H_0 : Daten sind normalverteilt

χ^2−Test auf Normalverteilung

- $\alpha = 0.05$
- $f = k - m - 1 = 9 - 3$ (9 Klassen, 2 geschätzte Parameter)
- $\chi^2_{0.95,6} = 12.59$
- Theoretische Häufigkeiten:
  $F_1 = 63.59, F_2 = 25.02, F_3 = 26.08, F_4 = 24.35, F_5 = 20.36, F_6 = 15.25, F_7 = 10.23, F_8 = 6.14, F_9 = 5.98$
- Testwert

$$\chi^2 = \frac{(0 - 63.59)^2}{63.59} + \frac{(148 - 25.02)^2}{25.02} + \ldots + \frac{(9 - 5.98)^2}{5.98} = 710.79$$

- Da $\chi^2 > 12.59$ ist, wird die Nullhypothese verworfen
- (Anmerkung: Die Daten sind annähernd Log-Normalverteilt)

χ^2−Test auf Normalverteilung

- $\alpha = 0.05$
- $f = k - m - 1 = 9 - 3$ (9 Klassen, 2 geschätzte Parameter)
- $\chi^2_{0.95,6} = 12.59$
- Theoretische Häufigkeiten:
  $F_1 = 63.59, F_2 = 25.02, F_3 = 26.08, F_4 = 24.35, F_5 = 20.36, F_6 = 15.25, F_7 = 10.23, F_8 = 6.14, F_9 = 5.98$
- Testwert

$$\chi^2 = \frac{(0 - 63.59)^2}{63.59} + \frac{(148 - 25.02)^2}{25.02} + ... + \frac{(9 - 5.98)^2}{5.98} = 710.79$$

- Da $\chi^2 > 12.59$ ist, wird die Nullhypothese verworfen
- (Anmerkung: Die Daten sind annähernd Log-Normalverteilt)

χ^2-Test

Verwendung:

- Test ob die Daten einer bestimmten Verteilung folgen
- Test auf Unabhängigkeit

Situation (Test auf Unabhängigkeit):

- Daten von zwei gleichzeitig gemessenen Größen: Paare von Messwerten $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$.
- Grundannahmen: Realisierungen zweier Zufallsvariablen X und Y.

Nullhypothese H_0: X und Y sind unabhängig.

χ^2-Test auf Unabhängigkeit

1. Gruppieren der x-Werte in k verschiedene Gruppen, und die y-Werte in m verschiedene Gruppen. Bestimme die Häufigkeiten

   $$N_{i,j} := |\{l : x_l \text{ gehört zur Gruppe } i, y_l \text{ gehört zur Gruppe } j\}|$$

   Aufstellen der Kontingenztafel mit diesen Häufigkeiten. Berechnung der Randhäufigkeiten $N_{i,*} := \sum_{j=1}^{m} N_{i,j}$ und $N_{*,j} := \sum_{i=1}^{k} N_{i,j}$.
2. Aufstellen der Nullhypothese: H_0 : Die X_i sind unabhängig von den Y_i.
3. Wahl des Fehlerniveaus $\alpha \in (0, 1)$.
4. Berechnung des Freiheitsgrades $f = (k-1)(m-1)$
5. Bestimmung des $1-\alpha$-Quantils $\chi^2_{1-\alpha,f}$ der χ^2-Verteilung mit Parameter f (der Vergleichswert)

χ^2-Test auf Unabhängigkeit

Kontingenztabelle:

X \ Y	1	2	...	m	Total
1	$N_{1,1}$	$N_{1,2}$	...	$N_{1,m}$	$N_{1,*}$
2	$N_{2,1}$				
$\vdots$					
k	$N_{k,1}$			$N_{k,m}$	$N_{k,*}$
Total	$N_{*,1}$			$N_{*,m}$	n

$N_{i,j}$ die Anzahl der Paare mit $x-$Wert i und $y-$Wert j,
$N_{i,*}$, $N_{*,j}$ die Randhäufigkeiten.

χ^2-Test auf Unabhängigkeit

6. Berechnung der theoretischen Häufigkeiten unter der Nullhypothese:
$$F_{i,j} = \frac{N_{i,*} \cdot N_{*,j}}{n}$$
7. Berechnung des Testwerts $\chi^2 = \sum_{i,j} \frac{(F_{i,j} - N_{i,j})^2}{F_{i,j}}$
8. Vergleich des Testwerts χ^2 mit dem Quantil $\chi^2_{1-\alpha,f}$. Entscheidungsregel:
   - Ist $\chi^2 > \chi^2_{1-\alpha,f}$ so wird H_0 verworfen.
   - Andernfalls wird H_0 angenommen.

Beispiel 10.7: Unabhängigkeit von Genveränderungen

100 Personen werde auf zwei verschiedene Genveränderungen (A und B) getetstet. Dabei findet man folgende Häufigkeiten:

$N_{i,j}$	A vorhanden	A nicht vorhanden	Total
B vorhanden	16	14	30
B nicht vorhanden	24	46	70
Total	40	60	100

Fragestellung: Ist das Auftreten der beiden Genveränderungen unabhängig voneinander?

Kapitel 11: Markov-Ketten

Beispiel: Erkunden einer Netzwerkstruktur

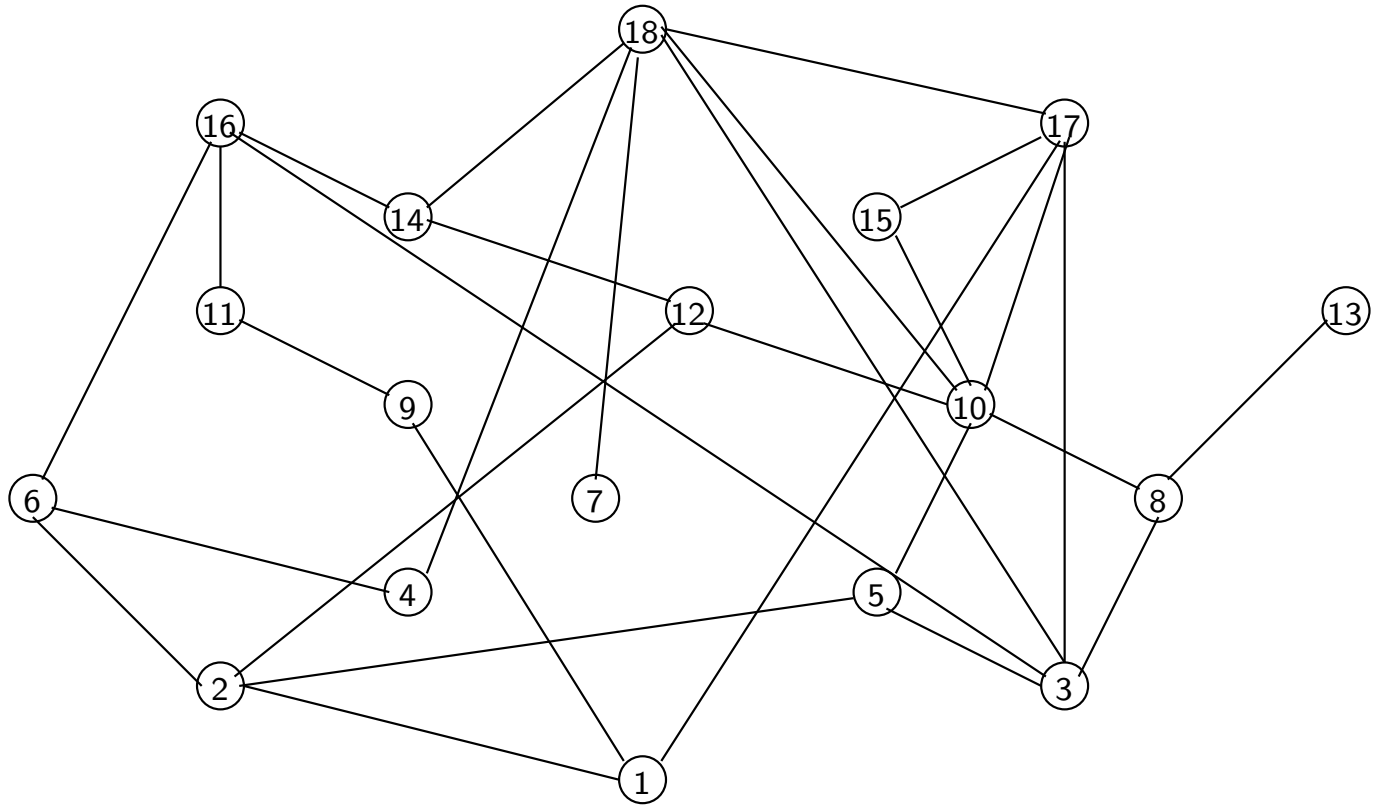


Abbildung: Netzwerk aus Knoten und Kanten

- Von innerhalb des Netzwerks sieht man nur seine Nachbarn
- Erkunden mittels Springen zu einem zufälligen Nachbarknoten

Markov-Ketten

Stochastische Modellierung dieses Beispiels:

- Graph bestehend aus Knoten und Kanten, Knoten nummeriert $\{1, 2, ..., k\}$.
- Start in (zufällig ausgewähltem) Knoten Nr. X_0.
- Wähle zufällig (gleichverteilt) einen Nachbarknoten aus, und springe dorthin: Knoten X_1.
- Iterativ: X_n die Nummer des Knoten, die man im $n-$ten Schritt besucht.
- X_n hängt von X_{n-1} ab, aber nicht von $X_{n-2}, X_{n-3}...$, d.h. für $a_0, ..., a_n \in \{1, ..., k\}$,

$$\mathbb{P}(X_n = a_n \,|X_0 = a_0, ...X_{n-1} = a_{n-1}) = \mathbb{P}(X_n = a_n \,|\, X_{n-1} = a_{n-1}).$$

- (Rechenbeispiel)

Markov-Ketten

Mögliche Fragestellungen:

- Wie oft besucht man (auf lange Sicht) einen bestimmten Knoten/einen durchschnittlichen Knoten? Besucht man alle gleich oft?
- Wie lange dauert es (im Mittel), bis man wieder beim Ausgangspunkt ist?
- Wie lange dauert es (im Mittel), bis man jeden Knoten einmal besucht hat?
- Welche Rolle spielt die Wahl des Startpunktes?
- Was kann ich aus solchen Informationen über die Struktur des Netzwerks sagen?
- Anwendung: PageRank (später)
- Begriff der Markov-Kette: Allgemeinere Formalisierung der wichtigsten Eigenschaften dieses Beispiels.

Markov-Ketten

(Def. 11.1) Sei S eine (höchstens abzählbare) Menge. Eine (homogene) Markov-Kette auf S ist eine Folge von Zufallsvariablen $X_0, X_1, X_2, \ldots$ auf $(\Omega, \mathbb{P})$ mit Werten in S so dass für alle $n \in \mathbb{N}_0$ und für alle $a_0, a_1, \ldots, a_n \in S$ gilt:

$$\begin{aligned}\mathbb{P}(X_n = a_n \,|\, X_0 = a_0, \ldots, X_{n-1} = a_{n-1}) &= \mathbb{P}(X_n = a_n \,|\, X_{n-1} = a_{n-1}) \\ &= \mathbb{P}(X_1 = a_n \,|\, X_0 = a_{n-1}).\end{aligned}$$

In Worten: Der Zustand der Kette im Schritt n hängt nur vom Zustand im Schritt $n-1$ ab, und nicht von der weiter zurückliegenden Vergangenheit (und auch nicht von n).

Die Menge S heißt Zustandsraum der Markov-Kette, ein Element $a \in S$ heißt Zustand.

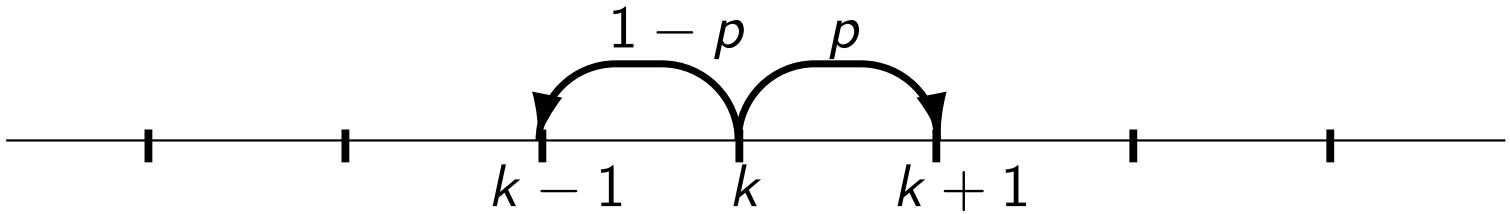
Beispiel 11.1: Irrfahrt auf $\mathbb{Z}^d$

- Symmetrische Irrfahrt auf $\mathbb{Z}$:

$$\mathbb{P}(X_n = k+1 \,|\, X_{n-1} = k) = \mathbb{P}(X_n = k-1 \,|\, X_{n-1} = k) = \frac{1}{2}$$

- Asymmetrische Irrfahrt auf $\mathbb{Z}$:

$$\mathbb{P}(X_n = k+1 \,|\, X_{n-1} = k) = p$$
$$\mathbb{P}(X_n = k-1 \,|\, X_{n-1} = k) = 1-p$$



- Symmetrische Irrfahrt auf $\mathbb{Z}^2$:

$$\mathbb{P}(X_n = y \,|\, X_{n-1} = x) = \frac{1}{4}$$

falls x und y Nachbarn sind.

Übergangswahrscheinlichkeiten

(Def. 11.2) Sei $X_0, X_1, \ldots$ eine Markov-Kette auf einem Zustandsraum S. Seien $a, b \in S$. Dann heißt

$$p_{a,b} := \mathbb{P}(X_n = b \,|\, X_{n-1} = a)$$

Übergangswahrscheinlichkeit von a nach b.

Notation: Wir verwenden äquivalent die verschiedenen Schreibweisen

$$p_{a,b} = p_{ab} = p(a, b)$$

für die Übergangswahrscheinlichkeiten.

- Beispiel 11.2 Symmetrische Irrfahrt auf $\mathbb{Z}^d$
- Beispiel 11.3 Irrfahrt auf Graph
- Übergangsgraph

Übergangsmatrix

(Def. 11.3) Die Übergangsmatrix einer Markovkette auf einem endlichen Zustandsraum $S = \{a_1, ..., a_K\}$ ist gegeben durch

$$P := \left(p_{a_n,a_m}\right)_{n,m=1,...,K} = \begin{bmatrix} p_{a_1,a_1} & p_{a_1,a_2} & \cdots & p_{a_1,a_K} \\ p_{a_2,a_1} & p_{a_2,a_2} & \cdots & p_{a_2,a_K} \\ \vdots & \ddots & & \vdots \\ p_{a_K,a_1} & p_{a_K,a_2} & \cdots & p_{a_K,a_K} \end{bmatrix}$$

- Beispiel 7.4: Zusatandsraum $S = \{1, 2, 3\}$, Übergangsmatrix

$$P := \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

(Übergangsgraph zeichnen)

Übergangsmatrix

(Def. 11.3) Die Übergangsmatrix einer Markovkette auf einem endlichen Zustandsraum $S = \{a_1, ..., a_K\}$ ist gegeben durch

$$P := \left(p_{a_n,a_m}\right)_{n,m=1,...,K} = \begin{bmatrix} p_{a_1,a_1} & p_{a_1,a_2} & \cdots & p_{a_1,a_K} \\ p_{a_2,a_1} & p_{a_2,a_2} & \cdots & p_{a_2,a_K} \\ \vdots & \ddots & & \vdots \\ p_{a_K,a_1} & p_{a_K,a_2} & \cdots & p_{a_K,a_K} \end{bmatrix}$$

- Beispiel 7.4: Zusatandsraum $S = \{1, 2, 3\}$, Übergangsmatrix

$$P := \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}.$$

(Übergangsgraph zeichnen)

Stochastische Matrizen

(Satz 11.1) Eine Übergangsmatrix P hat die Eigenschaften

- $0 \leq p_{a,b} \leq 1$ für alle $a, b \in S$
- $\sum_{b \in S} p_{ab} = 1 \quad \forall a \in S$.

Eine Matrix mit diesen zwei Eigenschaften heißt stochastische Matrix. Jede Übergangsmatrix einer Markov-Kette ist eine stochastische Matrix, und umgekehrt existiert zu jeder stochastischen Matrix eine Markov-Kette.