Statistik Praktikum 2009
Handout

März 2009

Hajo Holzmann / Florian Ketterer

1 Why *R*?

First of all *R* is free. It is an open source project, which is similar to the S language environment which was developed at Bell Laboratories. Although there are some differences, much code written for S runs unaltered under *R*. Moreover *R* is growing at a rapid pace and there are many packages available covering a huge field of topics and methods.
You can download *R* at `http:\\www.r-project.org`.

2 Programming Basics

The goal of this section is to provide you with some basic programming guidelines. In the heat of programming all the code you produce may seem perfectly understandable and clear to you. Anyway, if you don't follow some basic rules, you will run into trouble trying to understand your own code some time later. (There is a saying, that programs older than two weeks might be written by someone else...) In the special case of this practical course, you have to keep in mind, that your code will be corrected by a third person, who must be able to make sense of what you did.

2.1 Guideline for writing code

2.1.1 Keep it simple

Think of the reader. Don't write just for yourself. Break down complexity into simpler chunks. Avoid implicit or obscure language features. Minimize scope, both logical and visual. Minimizing scope and breaking down complexity are not contradictory to each other. It may seem cool to do a very complex statement in on line, but if you ever tried to understand a program written by someone else, you surely have already cursed this particular programming style. When in doubt you should strive for clarity first, then efficiency.

2.1.2 Keep it readable

Use informative variable names. Good code can be read like a book. Make names clearly unique. Avoid abbreviations whenever possible. Name variables with noun or adjective noun combinations.
It is quite tempting to use short names for variables and functions. This, however, is one of the main problems of many programs. Stories of programmers who used the name of their girlfriend as the name of a time variable might be known to you in the context of the year 2000 problem. In writing statistical programs people tend to name their variables 'x', 'xx', 'y0' and so on. Use more meaningful names. Especially abbreviations often seem totally clear at the moment but tend to loose their clarity very fast. 'predicted.value' is much more understandable than 'prdVl'. The usual way of separating two words in *R* is the use of a point like in 'linear.fit'.

2.1.3 Comment your code

Clearly comment necessary complexity. Be clear and concise. Say what is happening and why. Do not restate code. Keep code and comments visually separate. Comments are the most important part of a program, if you try to understand it later. Comment coherent units of code. Make it easy to look for a special functionality of your code. A good indicator whether you should comment

or not is the amount of time you spend on producing the code. If you thought on some special lines of code for hours, they might be worth a short comment. Be aware, that comments may also decrease the readability of source code. They might even be misleading, if you fail to update them when changing your program. That is why you should make the code as clear as possible to reduce the need for comments. When using program packages like *R* it is sometimes very helpful to comment on the functions and the parameters of the functions you use. This is especially true for *R* since many names and parameters of functions do not meet the claim for clarity and intuitive comprehensibility.

2.2 The use of functions

During this practical course you will mainly use functions which are already defined in R, but at some points you will be asked to write simple functions yourself.

2.2.1 Avoid copy and paste

It may occur that you find yourself using the same code over and over again. (For example if you are analyzing different data in the same way.) In this case you might want to define a function in order to shorten your code. It also keeps you from using copy and past to do the same things over and over again.

2.2.2 Use a meaningful name

Let your function names tell the reader what the function does. Name functions with verb noun combinations. Name your functions in a way, that you can read your code easily. Choose names that are as self-documenting as possible. Use a short synonym instead of an abbreviation.

3 Aufgaben

Exercise 1

Der Datensatz `tax.dat` enthält Gesamteinkommen und Steueraufkommen verschiedener Einkommensklassen in den USA in den Jahren 1974 und 1978. Die Daten finden sich in der Tabelle 1.

Table 1: Taxes in the USA in the years 1974 and 1978.

Income Bracket	Income 1974	Tax Yield 1974	Income 1978	Taxes Yield 1978
<5.000$	41.651.643	2.244.467	19.879.622	689.318
5.000-9.999$	146.400.740	13.646.348	122.853.315	8.819.461
10.000-14.999$	192.688.922	21.449.597	171.858.024	17.155.758
15.000-99.999$	470.010.790	75.038.230	865.037.814	137.860.951
>100.000$	29.427.152	11.311.672	62.806.159	24.051.898

Simpsons Paradoxon. Falls man Anteile gruppenweise in zwei Stichproben vergleicht, kann es vorkommen, dass die Anteil in jeder Gruppe in einem Sample konsistent höher sind als in dem anderen Sample, aber dass insgesamt die Anteile geringer sind. Dies bezeichnet man als das Simpsonsche Paradoxon.

Exercise 2

ZNS data set `ZNS.dat`: Mentale Krankheiten sind Veränderungen der Funktion des zentralen Nervensystems. Die Basis von der vorliegenden Studie sind die Zellen der Ammonschen Horn Formation, welche eine zentrale Rolle in mehreren Lern - und Gedächtnismechanismen hat. Verschiedene Zelltypen kürzlich verstorbener Patienten in dieser Formation wurden gezählt. (A : Astrocyte, O: Oligodendrozyt). Die Variable AO ist der Quotient dieser Werte. Die untersuchten Patienten werden klassifiziert bezüglich ihrer mentalen Krankheiten. (Class 1: Alzheimer's Disease; Class 2: Pick's Disease; Class 3: Dementia; Class 4: Healthy, Alter 50-60; Class 5: Healthy, Alter 61-100.)
Die Daten sind aus [10], Auszüge sind in der folgenden Tabelle gezeigt.

	Class	AN	ON	MN	GN	AO	health
1	1.00	2.04	0.29	0.15	2.84	7.21	K
2	1.00	1.70	0.23	0.14	2.07	7.74	K
3	1.00	1.95	0.31	0.13	2.38	6.63	K
4	1.00	2.24	0.32	0.15	2.71	7.32	K
5	1.00	2.35	0.32	0.13	2.81	7.53	K
6	1.00	2.57	0.36	0.14	3.03	7.33	K
...	...	...	...	...	...	...	...
21	2.00	1.15	2.57	0.24	4.00	0.45	K
22	2.00	1.30	2.12	0.21	3.60	0.62	K
23	2.00	1.22	1.86	0.18	3.26	0.67	K
24	2.00	1.21	2.54	0.27	3.81	0.43	K
25	2.00	1.39	2.80	0.32	4.51	0.50	K
...	...	...	...	...	...	...	...
40	3.00	1.14	0.45	0.15	1.74	2.59	K
41	3.00	1.22	0.58	0.15	1.96	2.14	K
42	3.00	1.49	0.55	0.13	2.18	2.76	K
43	3.00	1.23	0.53	0.13	1.87	2.34	K
...	...	...	...	...	...	...	...

Exercise 3

Die Daten sind in R in dem Datensatz `faithful` enthalten, welcher nicht separat geladen werden muss. Information sind durch den Befehl `?faithful` zu erhalten.

(a) punktweise asymptotisches Konfidenzband: für $z \sim B(n,p)$, gilt

$$\frac{z-np}{\sqrt{z(1-z/n)}} \xrightarrow{d} N(0,1).$$

Man verwendet dies nun für die empirische Verteilungsfunktion, bei der ja $n\hat{F}_n(x) \sim B(n, F(x))$ verteilt ist. Somit ergibt sich als Konfidenzintervall bei x:

$$\Big[\hat{F}_n(x) - \frac{u_{1-\alpha/2}\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}{\sqrt{n}}, \hat{F}_n(x) + \frac{u_{1-\alpha/2}\sqrt{\hat{F}_n(x)(1-\hat{F}_n(x))}}{\sqrt{n}}\Big],$$

wobei u_β das β-Quantil der Standardnormalverteilung ist. Sind dabei $x_1, \ldots, x_n$ beobachtet, so sortiere sie $x_{(1)} < \ldots < x_{(n)}$. Dann ist für $x_{(i)} \le x < x_{(i+1)}$ die empirische Verteilungsfunktion $\hat{F}_n(x) = i/n$ konstant und das Konfidenzintervall ist hier stets

$$\Big[\frac{i}{n} - \frac{u_{1-\alpha/2}\sqrt{\frac{i}{n}(1-i/n)}}{\sqrt{n}}, \frac{i}{n} + \frac{u_{1-\alpha/2}\sqrt{\frac{i}{n}(1-i/n)}}{\sqrt{n}}\Big].$$

An den Rändern entsprechend. Dabei ist noch zu beachten, dass man stets im Intervall $[0,1]$ bleibt.

(b) Die Pearson-Clopper Grenzen: Für n Beobachtungen mit k Erfolgen berechnen sich diese für $1 \le k \le n-1$ durch

$$\begin{aligned} p_l &= \frac{k}{k+(n-k+1)F_{1-\alpha/2}(f_1,f_2)}, \\ p_u &= \frac{k+1}{k+1+(n-k)F_{\alpha/2}(f_1-2,f_2+2)}, \end{aligned}$$

wobei $f_1 = 2(n-k+1), f_2 = 2k$, und $F_\alpha(f_1,f_2)$ das α-Quantile der F Verteilung mit f_1 und f_2 Freiheitsgraden ist. Für $k=0$ ist die obere Grenze gegeben durch

$$p_u = \frac{F_{1-\alpha}(2,2n)}{n+F_{1-\alpha}(2,2n)}, \tag{1}$$

und für $k=n$ ist die untere Grenze durch $1-p_u$, p_u wie oben, gegeben.

Für die empirische Verteilungsfunktion berechne für $x_{(k)} \le x < x_{(k+1)}$ obige Grenzen für k Erfolge.

(c) Der Kern-Dichte Schätzer mit Kern K ($\int K = 1$) und Bandbreite $h > 0$ ist gegeben durch

$$\hat{f}_n(x;h) = \frac{1}{nh}\sum_{k=1}^{n} K\Big(\frac{x-X_k}{h}\Big).$$

Für K wählen wir die Dichte der Standardnormalverteilung. Die Bandbreite wird aus den Daten geschätzt, wobei hier verschiedene Verfahren zur Verfügung stehen. Für die asymptotische Verteilung gilt unter Bedingungen

$$\sqrt{nh}\big(\hat{f}_n(x;h) - f(x)\big) \to N\Big(0, f(x)\int K^2\Big).$$

Für den Normalverteilungskern gilt $\int K^2 = 1/(2\sqrt{\pi})$. Als Konfidenzintervall für $f(x)$ erhält man

$$\Big[\hat{f}_n(x;h) - \frac{u_{1-\alpha/2}\left(\hat{f}_n(x;h)\right)^{1/2}}{\sqrt{nh}\,\sqrt{2}\pi^{1/4}}, \hat{f}_n(x;h) + \frac{u_{1-\alpha/2}\left(\hat{f}_n(x;h)\right)^{1/2}}{\sqrt{nh}\,\sqrt{2}\pi^{1/4}}\Big].$$

(d) Sind $X_1, \ldots, X_n$ beobachtet und $X_{(1)} < \ldots < X_{(n)}$, so kann man zeigen, dass gilt

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \max_{i=0,\ldots,n} \max \Big(i/n - F(X_{(i)}), F\big(X_{(i+1)}\big) - i/n \Big),$$

wobei $X_{(0)} = -\infty$, $X_{(n+1)} = \infty$. Diese Verteilung hängt für stetiges F nicht von F ab, man kann daher ohne Einschränkung $F \sim U(0,1)$ annehmen. Dann ist für u.i.v. gleichverteilte Zufallsvariablen $U_1, \ldots, U_n$ die Verteilung gleich der von

$$\max_{i=0,\ldots,n} \max \Big(i/n - U_{(i)}, U_{(i+1)} - i/n \Big),$$

wobei $U_{(0)} = 0$, $U_{(n+1)} = 1$. Diese kann man mit beliebiger Genauigkeit simulieren, indem man immer wieder Stichproben der Größe n von u.i.v. gleichverteilten Zufallszahlen zieht, und obigen Ausdruck berechnet. Sei $q_{1-\alpha}$ das $1-\alpha$ Quantil dieser simulierten Verteilung. Das gleichmäßige Konfidenzband ist dann gegeben durch

$$\Big[\max(0, \hat{F}_n(x) - q_{1-\alpha}), \min(1, \hat{F}_n(x) + q_{1-\alpha}) \Big].$$

Die Gleichmäßigkeit bezieht sich *nicht* auf die gleichmäßige Breite, sondern darauf, dass das Niveau $1-\alpha$ gleichmäßig für alle x gilt (und nicht nur separat für jedes x).

Exercise 4

Bleidatensatz `lead.dat`: Der Bleigehalt in Blutproben von 33 Kindern deren Eltern in einer Fabrik arbeiten, die Batterien produziert, wurde gemessen. Es gab auch eine Kontrollgruppe mit 33 Kindern anderer Eltern, die jeweils ähnliches Alter und in der Nachbarschaft lebten wie ein entsprechendes Kind der Fallgruppe. Die Daten wurden [6] entnommen, im Folgenden ein Auszug.

	group 1	group 2
1	38	16
2	23	18
3	41	18
4	18	24
5	37	19
6	36	11
7	23	10
8	62	15
9	31	16
...	...	...

Details zu den relevanten Tests werden im Praktikum besprochen.

Exercise 5

Der Datensatz `wuermer.csv` enthalt Daten über Schwermetallbelastungen von Würmern, die als Parasiten in Fischen vorkommen. Die Variable `Ind.Masse` enthält das Gewicht der Würmer, die Variable `Pb` den Bleigehalt. Die Variable `Gruppe` numeriert die verschiedenen Fanggebiete. Im Folgenden ein Auszug des Datensatzes.

	Gruppe	Anzahl	Masse	Ind.Masse	Pb	Cd
1	A	11	6.10	0.60	10	7
2	A	24	13.30	0.60	2	6
3	A	14	4.20	0.30	6	5
4	A	14	5.80	0.40	7	4
5	A	9	4.80	0.50	9	9
6	A	9	6.70	0.70	7	8
7	A	19	13.30	0.70	20	12
8	A	11	6.50	0.60	3	11
9	A	5	3.20	0.60	18	13
10	A	5	6.10	1.20	3	4
11	B	6	7.80	1.30	20	40
12	B	11	4.40	0.40	4	2
13	B	4	3.60	0.90	9	13
14	B	3	8.30	2.80	28	59
15	B	5	6.80	1.40	28	16
16	B	7	12.30	1.80	72	29
17	B	17	5.60	0.30	57	8
18	B	2	3.00	1.50	44	38
...						

In einem Artikel aus der Zeitschrift "Stochastik in der Schule" wird der Zusammenhang zwischen der absoluten Anzahl in einem Land pro Jahr geborener Menschen (Merkmal `Geburtenrate`) und der absoluten Anzahl der in einem Land lebenden Storchenpaare (Merkmal `Stoerche`) mit Hilfe des Pearson-Korrelationskoeffizienten untersucht. Der Datensatz liegt in `storch.csv` vor, Auszüge im Folgenden.

	Land	Flaeche	Stoerche	Menschen	Geburtenrate
1	Albanien	28750	100	3.20	83
2	Belgien	30520	1	9.90	87
3	Bulgarien	111000	5000	9.00	117
4	Daenemark	43100	9	5.10	59
5	Deutschland	357000	3300	78.00	901
6	Frankreich	544000	140	56.00	774
7	Griechenland	132000	2500	10.00	106
8	Holland	41900	4	15.00	188
9	Italien	301280	5	57.00	551
10	Oesterreich	83860	300	7.60	87
11	Polen	312680	30000	38.00	610
12	Portugal	92390	1500	10.00	120
13	Rumaenien	237500	5000	23.00	23
14	Spanien	504750	8000	39.00	439
15	Schweiz	41290	150	6.70	82
16	Tuerkei	779450	25000	56.00	1576
17	Ungarn	93000	5000	11.00	124

Die Korrelationskoeffizienten und Tests werden im Praktikum besprochen.

Exercise 6

(a) Fishers exakter Test auf Homogenität dient dazu, die Erfolgswahrscheinlichkeiten zweier unabhängiger binomialverteilter Zufallsvariablen miteinander zu vergleichen. Hat man etwa zwei Patientengruppen, bei denen eine bestimmte Krankheit mit Medikament A in Gruppe 1

und Medikament B in Gruppe 2 behandelt wird, und jeweils untersucht wird, ob der Patient geheilt wurde, so kann man die Erfolgswahrscheinlichkeiten für Heilung durch Medikament A mit der von Medikament B vergleichen.

Gegeben seien also $X \sim B(n, p_X)$, $Y \sim B(m, p_Y)$, wobei X, Y unabhängig sind. Werden $X = x$, $Y = y$ beobachtet, dann fasst man diese Ergebnisse in einer 4-Felder-Tafel (2×2 Kontingenztafel) zusammen:

	X	Y	
0	$n - x$	$m - y$	$\sum = n + m - x - y$
1	x	y	$\sum = x + y$
	$\sum = n$	$\sum = m$	$\sum = n + m$

Wir testen die Hypothese

$$H: \ p_X = p_Y = p.$$

Unter H gilt

$$P(X = x, Y = y | X + Y = x + y) \quad = \quad \frac{\binom{n}{x}\binom{m}{y}}{\binom{n+m}{x+y}}. \tag{2}$$

Dies ist die Verteilung der hypergeometrischen Verteilung $Hyp(x + y, n, m)$.

Bei extremen Beobachtungen von x bzw. y gemäß der Verteilung (2) wird man die Hypothese H verwerfen. Dies geschieht etwa, falls der P-Wert kleiner als das gewünschte Niveau $\alpha > 0$ ist.

Wir bestimmen die p-Werte folgendermaßen:
Für die einseitige Alternative $K: \ p_X > p_Y$ bzw. $(K: \ p_X < p_Y)$ gilt:

$$PW = \sum_{k=x}^{\min\{n, x+y\}} P(N = k) \text{ (bzw. } PW = \sum_{k=0}^{x} P(N = k)),$$

dabei ist $N \sim Hyp(x + y, n, m)$.
Für die zweiseitige Alternative $K: \ p_X \neq p_Y$:
Man schätzt zunächst $\hat{p}_X = \frac{x}{n}$ und $\hat{p}_Y = \frac{y}{m}$. Ist etwa $\hat{p}_X > \hat{p}_Y$, so bilde

$$\tilde{P} = \sum_{k=x}^{\min\{n, x+y\}} P(N = k)$$

und

$$i_{\max} = \max\{i: \ \sum_{k=0}^{i} P(N = k) < \tilde{P}\}$$

Dann:

$$PW = \tilde{P} + \sum_{k=0}^{i_{\max}} P(N = k).$$

Für $\hat{p}_X < \hat{p}_Y$ verfahre analog.

(b) Wir beobachten ein Merkmal X mit Ausprägungen $1, 2$ zu aufeinanderfolgenden Zeitpunkten $T = 1, 2$, und wollen wissen, ob die Verteilung von X zu beiden Zeitpunkten gleich ist.

Formal beobachten wir u.i.v. Zufallsvektoren $Z_1 = (X_{11}, X_{12}), \ldots, Z_n = (X_{n1}, X_{n2})$, wobei $X_{ij} \in \{1, 2\}$, und bilden die Kontingenztafel

		X_{i2}		
		1	2	
X_{i1}	1	n_{11}	n_{12}	$n_{1\cdot}$
	2	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	n

sowie die Tafel der Wahrscheinlichkeiten

		X_{i2}		
		1	2	
X_{i1}	1	p_{11}	p_{12}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$p_{2\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	

Wir testen die Hypothese: Die Randverteilungen zu beiden Zeitpunkten, also die Verteilungen von X_{i1} und X_{i2}, sind gleich, oder äquivalent

$$H: \ p_{\cdot 1} = p_{1\cdot}$$

welches sich zu $H: \ p_{21} = p_{12}$ reduziert. Sei

$$N_{ij} = \sum_{k=1}^{n} 1_i(X_{k1}) 1_j(X_{k2}), \ i,j = 1,2$$

Es kann gezeigt werden, dass unter H

$$N_{12} | N_{12} + N_{21} = n_{12} + n_{21} \sim B(n_{12} + n_{21}, 1/2).$$

gilt. Man verwirft nun die Hypothese H für extreme Werte von N_{12} unter dieser bedingten Wahrscheinlichkeitsverteilung.
Die P-Werte bestimmt man folgendermaßen:
P-Wert gegen einseitige Alternative $K: \ p_{12} > p_{21}$:

$$PW = \sum_{k=n_{12}}^{n_{12}+n_{21}} \binom{n_{12}+n_{21}}{n_{12}} (1/2)^{n_{12}+n_{21}}.$$

Zweiseitig: Ist $n_{12} \geq n_{21}$, so ist

$$PW = 2 \cdot \sum_{k=n_{12}}^{n_{12}+n_{21}} \binom{n_{12}+n_{21}}{n_{12}} (1/2)^{n_{12}+n_{21}}$$

ansonsten vertausche die Rollen.

(c) Angenommen, an n Versuchseinheiten werden zwei Merkmale, ein X-Merkmal mit Ausprägungen $1, \ldots, I$ sowie ein Y-Merkmal mit Ausprägungen $1, \ldots, J$, beobachtet. Es soll untersucht werden, ob X-Merkmal und Y-Merkmal unabhängig voneinander sind.
Wir beobachten $Z_k = (X_k, Y_k)$ unabhängig und identisch verteilt, $X_k \in \{1, \ldots, I\}, Y_k \in \{1, \ldots, J\}$,

$$N_{ij} = \sum_{k=1}^{n} 1_i(X_k) 1_j(Y_k), \quad i = 1, \ldots, I, \ j = 1, \ldots, J.$$

Für beobachtete $N_{ij} = n_{ij}$ bildet man wiederum die Kontingenztafel

		Y				
		1	2	$\cdots$	J	
X	1	n_{11}	n_{12}		n_{1J}	$n_{1\cdot}$
	2	n_{21}	n_{22}		n_{2J}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}		n_{IJ}	$n_{I\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot J}$	$n_{\cdot\cdot}$

sowie die zugehörige Tafel der Wahrscheinlichkeiten

		Y				
		1	2	$\cdots$	J	
X	1	p_{11}	p_{12}		p_{1J}	$p_{1\cdot}$
	2	p_{21}	p_{22}		p_{2J}	$p_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	p_{I1}	p_{I2}		p_{IJ}	$p_{I\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$		$p_{\cdot J}$	

, $\qquad P(X_1 = i, Y_1 = j) = p_{ij}.$

Man möchte nun die Hypothese testen

$$H: \ X_k \text{ und } Y_k \text{ sind unabhängig} \quad (k = 1, \ldots, n)$$

oder äquivalent $H: \ p_{ij} = p_{i\cdot} p_{\cdot j}, \ i = 1, \ldots, I, j = 1, \ldots, J.$ Setzte

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{\cdot\cdot}}, \qquad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n_{\cdot\cdot}}, \quad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n_{\cdot\cdot}}.$$

Zum Testen von H betrachtet man die Pearsonsche χ^2-Statistik

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{\cdot j} n_{i\cdot}/n)^2}{n_{\cdot j} n_{i\cdot}/n} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{\hat{p}_{\cdot j}\hat{p}_{i\cdot}}.$$

Man kann zeigen, dass unter H die Teststatistik X^2 asymptotisch $\chi^2_{(I-1)(J-1)}$-verteilt ist.

Exercise 7

Das Standardmodell der Urknalltheorie über den Ursprung des Universums besagt, dass sich dieses gleichmäßig und lokal gemäß dem Hubble'schen Gesetz

$$y = \beta x \tag{3}$$

ausdehnt, wobei y die relative Geschwindigkeit ("Fluchtgeschwindigkeit") von zwei beliebigen Galaxien, die den Abstand x voneinander haben, angibt und β die so genannte "Hubble-Konstante" bezeichnet (in der Astrophysik wird normalerweise die Notation $y \equiv v$, $x \equiv d$ und $\beta \equiv H_0$ verwendet). β^{-1} entspricht dabei dem approximativen Alter des Universums. Allerdings ist β unbekannt und muss irgendwie mit Hilfe der Beobachtungen von x und y für eine Vielzahl von Galaxien mit unterschiedlichem Abstand zu uns geschätzt werden.

Der unten in Auszügen gegebene Datensatz `hubble` aus dem Paket `gamair`, welches ggf. erst noch in R installiert werden muss, enthält Messungen für y (in km/s) und x (in Mpc=$3.09 \cdot 10^{19}$km) des Weltraumteleskops Hubble im Fall von 24 verschiedenen Galaxien.

(c) Es sei das (hier allgemein multiple) lineare Modell $Y = X\beta + \epsilon$ mit

- $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ Realisierungen der beobachtbaren Zielvariable
- $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}$ Designmatrix

	Galaxy	y	x
1	NGC0300	133	2.00
2	NGC0925	664	9.16
3	NGC1326A	1794	16.14
4	NGC1365	1594	17.95
5	NGC1425	1473	21.88
6	NGC2403	278	3.22
7	NGC2541	714	11.22
8	NGC2090	882	11.75
9	NGC3031	80	3.63
⋮	⋮	⋮	⋮
24	NGC7331	999	14.72

- $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \in \mathbb{R}^p$ unbekannter Parametervektor
- $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n$ (nicht beobachtbarer) zufälliger Fehler mit $\epsilon_1, \ldots, \epsilon_n$ iid, $E\epsilon_i = 0$, $E\epsilon_i^2 = \mathrm{Var}(\epsilon_i) = \sigma^2$. (In Matrixschreibweise: $\mathrm{Cov}(\epsilon) = \sigma^2 \cdot I_n$)

gegeben.
Gesucht: Eine "möglichst gute" Schätzung von β.
Der Kleinste-Quadrate Schätzer $\widehat{\beta}$ minimiert $\|y - X\beta\|^2 = \sum_{i=1}^n (y_i - x_i^T\beta)^2$:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|^2$$

Partielles Ableiten von $\|y - X\beta\|^2$ nach β und Nullsetzen liefert unter der Voraussetzung, dass X vollen Rang p hat, die (eindeutige) Lösung

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

In R liefert Ihnen der Befehl `lm` diesen Kleinste-Quadrate Schätzer.
Für den Kleinste-Quadrate Schätzer $\widehat{\beta}$ gilt:

- $E(\widehat{\beta}) = \beta$
- $Cov(\widehat{\beta}) = \sigma^2 (X^T X)^{-1}$

(d) Die Regressionsgerade bei der einfachen linearen Regression ist durch $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ gegeben. In ein vorhandenes Schaubild kann man diese in R leicht mit `abline` einzeichnen.

(f) Die Modellannahmen bei der linearen Regressionsanalyse sind:

1) Die Beziehung zwischen der Zielvariablen y und den erklärenden Variablen ist (zumindest approximativ) linear
2) Die (zufälligen) Fehler ϵ_i haben Erwartungswert Null
3) Die Fehler ϵ_i haben konstante Varianz σ^2
4) Die Fehler ϵ_i sind unkorreliert
5) Die Fehler ϵ_i sind normalverteilt

Die Annahme 5) benötigt man beim Testen sowie der Bestimmung von Konfidenz- und Prognoseintervallen.
Die Annahmen 4) und 5) ergeben zusammen, dass die Fehler unabhängige Zufallsvariablen sind.

Residuen, die man in R u.a. mit `residuals` erhält, sind folgendermaßen definiert:

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i, \qquad \widehat{\epsilon} = (\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n)^T,$$

wobei $\widehat{y} = X\widehat{\beta}$ die angepassten Werte (fitted values) sind. Diese erhält man in R mit dem Befehl `fitted`.

Unter einer Residuenanalyse versteht man die graphische Überprüfung, ob die Modellannahmen der linearen Regression erfüllt sind. Die wichtigsten Plots sind dabei (unter anderem):

- Residuen gegen die angepassten Werte (fitted values) $\widehat{y}_i$ der Zielvariablen y.
- QQ-Plot (Normalplot) der Residuen gegen die Normalverteilung (in R: `qqnorm` und `qqline`)

Diese Plots und ihr gewünschtes Aussehen werden im Praktikum genauer besprochen.

(j) Falls die zufälligen Fehler normalverteilt sind, gilt unter der Hypothese $H_0: \beta = \beta_0$ (bei der einfachen linearen Regression ohne Intercept):

$$T = \frac{\widehat{\beta} - \beta_0}{\widehat{\sigma}_{\widehat{\beta}}} \sim t_{n-1},$$

wobei n die Anzahl der Beobachtungen angibt und t_{n-1} für die $t-$Verteilung mit $n-1$ Freiheitsgraden steht.

Exercise 8

(d) Sei $\hat{\beta}$ der KQ-Schätzer im linearen Modell. Als Vorhersage für y_{n+1} bei x_{n+1} betrachten wir

$$\hat{y}_{n+1} = x_{n+1}^T \hat{\beta}.$$

(e) Das Konfidenzintervall für $x_{n+1}^T\beta$ zum Niveau α ist gegeben durch

$$[x_{n+1}^T\hat{\beta} - \hat{\sigma}\sqrt{(x_{n+1}^T(X^TX)^{-1}x_{n+1})}t_{1-\alpha/2}(n-p), x_{n+1}^T\hat{\beta} + \hat{\sigma}\sqrt{(x_{n+1}^T(X^TX)^{-1}x_{n+1})}t_{1-\alpha/2}(n-p)].$$

(f) Das Prognoseintervall für den nicht beobachteten Werte $y_{n+1} = x_{n+1}^T\beta + \epsilon_{n+1}$ für eine weitere Realisierung der erklärenden Variablen x_{n+1} zum Niveau $\alpha > 0$ ist gegeben durch

$$[\hat{y}_{n+1} - \hat{\sigma}\sqrt{(1 + x_{n+1}^T(X^TX)^{-1}x_{n+1})}t_{1-\alpha/2}(n-p), \hat{y}_{n+1} + \hat{\sigma}\sqrt{(1 + x_{n+1}^T(X^TX)^{-1}x_{n+1})}t_{1-\alpha/2}(n-p)].$$

Bemerkung: Das Prognoseintervall für y_{n+1} ist breiter als das Konfidenzintervalll für $x_{n+1}^T\beta$, da der zusätzliche Fehler ϵ_{n+1} in y_{n+1} mit berücksichtigt werden muss.

Exercise 9

In dieser Aufgabe geht es um zwei Datensätze von R. R. Baker und M. A. Bellis, die man in [1] findet, welche diese als Beleg für die so genannte Theorie der Spermienkonkurrenz (theory of sperm competition) beim Menschen anführen. Als Spermienkonkurrenz bezeichnet man allgemein die Konkurrenz von Spermien eines oder mehrerer Männchen um die Chance zur Befruchtung einer Eizelle (siehe u.a. Wikipedia). Die zugrunde liegende Idee bei der Untersuchung von Baker und

Bellis war, dass es für Männchen aus evolutionärer Sicht von Vorteil ist, ihre Spermienanzahl (unterbewusst), entsprechend den Möglichkeiten ihrer Weibchen einen "Seitensprung" zu begehen, zu erhöhen. Ein solches Verhalten kann bei einer Vielzahl von anderen Tieren beobachtet werden und daher wollten Baker und Bellis an einer Auswahl von Studenten und freiwilligen Mitarbeitern der Universität von Manchester untersuchen, ob es Anhaltspunkte für ein ähnliches Verhalten beim Menschen gibt.

Die Daten zu dieser Untersuchung sind in den beiden Datensätzen `sperm.comp1` und `sperm.comp2` im Paket `gamair` gegeben. Der Datensatz `sperm.comp1` enthält die Variablen Spermienanzahl (`count`) in Millionen, Zeit seit dem letzten Geschlechtsverkehr (`time.ipc`) in Stunden sowie den Anteil der Zeit seit dem letzten Geschlechtsverkehr, den die Paare miteinander verbracht haben (`prop.partner`) für 15 heterosexuelle Paare. Der Datensatz `sperm.comp2` enthält Daten über die durchschnittliche Spermienanzahl über mehrere (`n`) Kopulationen von 24 heterosexuellen Paaren zusammen mit dem Gewicht, Größe und Alter der jeweiligen Partner sowie das Volumen eines Hoden des Mannes. Im Folgenden sehen Sie Auszüge der beiden Datensätze.

	subject	time.ipc	prop.partner	count
1	A	60	0.20	570
2	B	149	0.98	219
3	D	70	0.50	485
4	F	168	0.50	516
5	K	48	0.20	448
6	L	32	1.00	60
7	M	48	0.02	282
8	N	56	0.37	455
9	P	31	0.30	76
10	Q	38	0.45	228
⋮	⋮	⋮	⋮	⋮
15	Y	44	0.75	225

	pair	n	count	f.age	f.height	f.weight	m.age	m.height	m.weight	m.vol
1	A	4	514	25	170	64	25	188	95	28
2	B	27	393	24	175	58	44	180	79	20
3	C	2	305	22	180	57	20	185	76	NA
4	D	1	485	27	183	66	30	183	67	NA
5	E	1	422	26	163	54	25	183	67	14
6	F	1	516	30	166	59	32	175	73	18
7	G	1	244	20	173	63	31	180	75	31
8	H	2	65	21	164	57	24	175	70	15
9	I	1	525	20	155	56	18	178	75	23
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
23	Y	1	225	NA	NA	NA	20	NA	NA	NA
24	AB	2	242	20	180	57	20	180	70	12

(b) Mit den Notationen aus Exercise 6 gilt:
Ein (unbiased) Schätzer für die Varianz der zufälligen Fehler im linearen Regressionsmodell ist durch

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\widehat{\epsilon}_i^2$$

gegeben. Damit erhält man durch $\widehat{\sigma}\left(\left(X^TX\right)^{-1}_{ii}\right)^{1/2}$ die Standardabweichung $\widehat{\sigma}_{\widehat{\beta}_i}$ von $\widehat{\beta}_i$. Weiter ist das Konfidenzintervall für β_i zum Niveau $\alpha > 0$:

$$[\widehat{\beta}_i - \widehat{\sigma}_{\widehat{\beta}_i} \cdot t_{n-p,1-\frac{\alpha}{2}} \ ; \ \widehat{\beta}_i + \widehat{\sigma}_{\widehat{\beta}_i} \cdot t_{n-p,1-\frac{\alpha}{2}}]$$

Bestimmtheitsmaß: $R^2 = 1 - \frac{\sum_{i=1}^n \widehat{\epsilon}_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$, da gilt

$$\|y - \overline{y}1_n\|^2 = \|\widehat{y} - \overline{y}1_n\|^2 + \|y - \widehat{y}\|^2$$

gewichtetes (adjusted) Bestimmtheitsmaß:

$$R_a^2 = 1 - \frac{\sum_{i=1}^n \widehat{\epsilon}_i^2/(n-p)}{\sum_{i=1}^n (y_i - \overline{y})^2/(n-1)}$$

F-Test: Falls die zufälligen Fehler normalverteilt sind, gilt unter der Hypothese $H_0 : \ \beta_1 = \cdots = \beta_{p-1} = 0$:

$$F = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2/(p-1)}{\sum_{i=1}^n (y_i - \widehat{y}_i)^2/(n-p)} \sim F_{p-1,n-p},$$

wobei $F_{p-1,n-p}$ für die $F-$Verteilung mit $p-1$ und $n-p$ Freiheitsgraden steht.

(f) Der Leverage h_{ii} der Beobachtung i wird gerade durch das $i-$te Diagonalelement der orthogonalen Projektion (Hat-Matrix) $H = X(X^TX)^{-1}X^T$, wobei X wieder für die Designmatrix der Beobachtungen steht, gegeben.

(h) Modellwahl auf Basis von so genannten Informationskriterien. Darunter versteht man Zahlen, die die Güte der Anpassung eines Modells relativ zu dessen Komplexität erfassen.

AIC: Wähle Modell mit kleinstem

$$\mathrm{AIC} = -2\mathcal{L}_n(\widehat{\beta}, \widehat{\sigma}^2_{\mathrm{ML}}) + 2(p+1) = n + n\log(\mathrm{RSS}_p/n) + n\log(2\pi) + 2(p+1)$$

beziehungsweise ohne Konstanten mit kleinstem $\overline{\mathrm{AIC}} = n\log(\mathrm{RSS}_p/n) + 2p$, wobei RSS_p die Residuenquadratsumme des betrachteten Modells bezeichnet.

BIC: Wähle Modell mit kleinstem

$$\mathrm{BIC} = -2\mathcal{L}_n(\widehat{\beta}, \widehat{\sigma}^2_{\mathrm{ML}}) + \log(n) \cdot (p+1) = n + n\log(\mathrm{RSS}_p/n) + n\log(2\pi) + \log(n) \cdot (p+1)$$

bzw. wieder ohne Konstanten mit kleinstem $\overline{\mathrm{BIC}} = n\log(\mathrm{RSS}_p/n) + \log(n) \cdot p$.

(m) Das Vorgehen bei der Backward Elimination bzw. der Forward Selection wird im Praktikum erklärt.

Exercise 10

Die frühe Erkennung eines Herzinfarkts ist für die optimale Versorgung des Patienten sehr wichtig. Ein vorgeschlagenes diagnostisches Hilfsmittel ist dabei der Anteil des Enzyms Creatin-Kinase (CK, auch als Creatin-Phosphokinase oder als Kreatinkinase bezeichnet) im Blutkreislauf. In einer Studie (Smith, 1967) wurde der Wert dieses Enzyms bei 360 Personen, bei denen der Verdacht auf einen Herzinfarkt bestand, bestimmt. Später wurde nach weiteren medizinischen Untersuchen festgestellt, ob der Patient tatsächlich einen Herzinfarkt erlitten hatte. Die Ergebnisse dieser Studie, welche man in [8] findet, sind in der folgenden Tabelle gegeben.

CK-Wert	Patienten mit Herzinfarkt	Patienten ohne Herzinfarkt
20	2	88
60	13	26
100	30	8
140	30	5
180	21	0
220	19	1
260	18	1
300	13	1
340	19	1
380	15	0
420	7	0
460	8	0

Im Originalpaper wurden die Patienten in verschiedene Bereiche für den CK-Wert eingeteilt, in der obigen Tabelle sind allerdings nur die Mittelpunkte dieser Bereiche angegeben.

Theorie zur logistischen Regression
Wir betrachten Regression auf binomial-verteilte Daten. Dabei liegen also binmialverteilte abhängige Daten vor, $Y_i \sim Bin(n_i, p_i)$, $i = 1, \ldots, n$ sowie Kovariablen $x_i = (x_{i1,\ldots,x_{ir}})^T$, $i = 1, \ldots, n$, die die Erfolgswahrscheinlichkeit beeinflussen, d.h. $p_i = p_i(x_i)$. Wir gehen davon aus, dass die p_i von einem linearen Prädiktor $\eta_i = (1, x_i^T)^T \beta$ abhängen. Hierbei ist $\beta \in \mathbb{R}^p$ wie bei der linearen Regression ein unbekannter Parameterverktor. Um sicherzustellen, dass $0 < p_i < 1$ gilt, wählen wir eine strikt monoton wachsende, stetige Linkfunktion $g : (0, 1) \to \mathbb{R}$, so dass $\eta_i = g(p_i)$ bzw. $p_i = h(\eta_i)$ mit $h = g^{-1}$ gilt. Ein wichtiges Beispiel ist der sog. Logit-Link $g(p) = \log\left(\frac{p}{1-p}\right)$ bzw. $h(\eta) = \frac{e^\eta}{1+e^\eta}$. Dies führt uns zum logistischen Regressionsmodell.
Um in R ein logistisches Regressionsmodell anzupassen, nutzen wir die Funktion `glm` mit der Option `family='binomial'`.
Falls das logistische Regressionsmodell korrekt ist, gilt für festes n und für $n_i \to \infty$, dass D asymptotische χ^2_{n-p}-verteilt ist.
Wir können die Devianz auch nutzen, um zwei ineinander geschachtelte logistische Regressionsmodelle miteinander zu vergleichen. Dazu seien
L: das große logistische Regressionsmodell mit $\beta \in \mathbb{R}^l$,
S: das klein Teilmodell von L mit $\beta \in \mathbb{R}^s$, $s < l$.
Dann kann die Likelihood Quotienten Statistik von Modell S gegen Modell L durch Differenzbildung der zugehörigen Devianzen$D_S - D_L$ berechnen. Falls S korrekt ist, gilt

$$D_S - D_L \xrightarrow{d} \chi^2_{l-s},$$

falls $n_i \to \infty$ oder auch $n \to \infty$.

Exercise 11

Im frühen Stadium einer Krankheitsepidemie wächst die Rate, mit der Neuerkrankungen auftreten oft exponentiell schnell in der Zeit an. Ein Modell der Form

$$\mu_i = \gamma \exp(\delta t_i) \tag{4}$$

mit unbekannten Parametern γ und δ für die erwartete Anzahl von Neuerkrankungen μ_i zum Zeitpunkt t_i erscheint daher recht vernünftig. In [14] findet man eine Zeitreihe der neuen Aids-Fälle pro Jahr in Belgien für die Jahre 1981 bis 1993.

Jahr	1981	1982	1983	1984	1985	1986	1987	1988
Neue Aids-Fälle	12	14	33	50	67	74	123	141

Jahr	1989	1990	1991	1992	1993
Neue Aids-Fälle	165	204	253	246	240

Poisson-Regression
Falls die Zielvariable Y_i eine Zählvariable ist, also Werte in $\mathbb{N}_0$ hat, ist das Standard Regressionsmodell die Poisson-Regression. Dabei liegen also poissonverteilte abhängige Daten vor, $Y_i \sim Po(\lambda_i)$, $i = 1, \ldots, n$ sowie Kovariablen $x_i = (x_{i1,\ldots,x_{ir}})^T$, $i = 1, \ldots, n$, die den Parameter der Poisson-Verteilung beeinflussen, d.h. $\lambda_i = \lambda_i(x_i)$. Wir gehen davon aus, dass die λ_i von einem linearen Prädiktor $\eta_i = (1, x_i^T)^T \beta$ abhängen. Hierbei ist $\beta \in \mathbb{R}^p$. Um sicherzustellen, dass $\lambda_i > 0$ gilt, wählen wir eine strikt monoton wachsende, stetige Linkfunktion $g : \mathbb{R}_{>0} \to \mathbb{R}$, so dass $\eta_i = g(\mu_i)$ bzw. $\lambda_i = h(\eta_i)$ mit $h = g^{-1}$ gilt.

Exercise 12

Informationen zu dem Datensatz `InsectSprays` können in R aufgerufen werden (`?InsectSprays`). Grundlagen zur einfaktoriellen Varianzanalyse werden in dem Praktikum besprochen.

Exercise 13

Die Datensätze `pvc`, `oatvar` und `abrasion` sind in der Library `faraway` (laden mit `library(faraway)`) enthalten. Informationen zu diesen Datensätzen können dann erhalten werden mit `pvc` etc. Grundlagen zur mehrfaktoriellen Varianzanalyse und dem Design von Studien werden in dem Praktikum besprochen.

Exercise 14

Eine Zufallsvariable X hat als Verteilung eine Mischung von zwei Normalverteilungen, falls X eine Dichte der Form

$$f_X(x; p, \mu_1, \mu_2, \sigma_1, \sigma_2) = p\phi(x; \mu_1, \sigma_1) + (1-p)\phi(x; \mu_2, \sigma_2) \tag{5}$$

hat. Dabei ist $\phi(x; \mu, \sigma)$ die Dichte von $N(\mu, \sigma)$ (σ die Standardabweichung) und $0 < p < 1$. Falls in (5) gilt $\sigma_1 = \sigma_2 = \sigma$, so schreiben wir $f_X(x; p, \mu_1, \mu_2, \sigma)$.

Man kann eine Zufallszahl mit der Dichte (5) wie folgt erzeugen. Ziehe eine Hilfszufallszahl $Z \sim U(0,1)$. Falls $Z \leq p$, ziehe $X \sim N(\mu_1, \sigma_1)$, ansonsten ziehe $X \sim N(\mu_2, \sigma_2)$.

Sind nun $X_1, \ldots, X_n$ u.i.v. nach einer Zweikomponenten Mischung von Normalverteilungen verteilt, so ist die log-Likelihood definiert durch

$$\mathcal{L}_n(p, \mu_1, \mu_2, \sigma_1, \sigma_2) = \sum_{i=1}^{n} \log\big(f_X(X_i; p, \mu_1, \mu_2, \sigma_1, \sigma_2)\big).$$

Ein Maximum Likelihood Schätzer $(\hat{p}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$ von $(p, \mu_1, \mu_2, \sigma_1, \sigma_2)$ ist ein Argmax von $\mathcal{L}_n(p, \mu_1, \mu_2, \sigma_1, \sigma_2)$. Dies liefert auch eine Schätzung für die Dichte der X_i, nämlich $f_X(x; \hat{p}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$. Analog ist die log-Likelihood Funktion unter Annahme gleicher Varianzen definiert als

$$\mathcal{L}_n(p, \mu_1, \mu_2, \sigma) = \sum_{i=1}^{n} \log\big(f_X(X_i; p, \mu_1, \mu_2, \sigma)\big).$$

Maximum Likelihood Schätzer und geschätzte Dichte ergeben sich analog.
Die Likelihood Quotienten Statistik zum Test von $H : \sigma_1 = \sigma_2$ ist definiert durch

$$LRTS = 2\Big(\max_{(p,\mu_1,\mu_2,\sigma_1,\sigma_2)} \mathcal{L}_n(p, \mu_1, \mu_2, \sigma_1, \sigma_2) \; - \max_{(p,\mu_1,\mu_2,\sigma)} \mathcal{L}_n(p, \mu_1, \mu_2, \sigma)\Big).$$

Diese ist unter der Hypothese H und falls $\mu_1 \neq \mu_2$ verteilt nach χ^2_1.

Exercise 15

Eine Markov Kette mit endlichem Zustandsraum und stationären Übergängen ist ein Prozess $(X_n)_{n\geq 1}$, wobei die X_i Werte in einer endlichen Menge $\mathcal{M} = \{1, \ldots, m\}$ annehmen, und

$$P\Big(X_n = j | X_{n-1} = i, X_{n-2} = i_{n-2}, \ldots X_1 = i_1\Big) = P\Big(X_n = j | X_{n-1} = i\Big) = p_{ij},$$

wobei $j, i, i_1, \ldots, i_{n-2} \in \mathcal{M}$, und $n \geq 2$. Die Matrix

$$P = (p_{ij})_{i,j=1,\ldots,m}$$

bezeichnet man als Übergangsmatrix der Markov Kette. Die Verteilung der Markov Kette $(X_n)_{n\geq 1}$ ist somit eindeutig bestimmt durch die Übergangsmatrix P und die Startverteilung $\nu_i = P(X_1 = i)$. Unter geeigneten Bedingungen existiert genau eine stationäre Startverteilung π (Zeilenvektor), diese ist bestimmt durch

$$\pi P = \pi.$$

Der Einfluss der Startverteilung ist gering, in Simulationen benutzte stets die stationäre Startverteilung π.

Für eine Stichprobe $X_1, \ldots, X_n$ setzte

$$f_{ij} = \sum_{k=2}^{n} 1_{X_{k-1}=i, X_k=j}, \qquad f_i = \sum_{k=1}^{n} 1_{X_k=i}.$$

Der Maximum Likelihood Schätzer für p_{ij}, bedingt auf die erste Beobachtung X_1, ist

$$\hat{p}_{ij} = \frac{f_{ij}}{f_i}.$$

Die Maximum Likelihood Schätzer $\hat{p}_{ij}$ sind asymptotisch normalverteilt. Darüber hinaus gilt: a. für jedes i ist die asymptotische Kovarianz von $(\hat{p}_{i1}, \ldots, \hat{p}_{im})$ die gleiche wie für die (unabhängige) Multinomialverteilung mit Parametern $(p_{i1}, \ldots, p_{im})$, b. für $i_1 \neq i_2$ sind die Vektoren $(\hat{p}_{i_1 1}, \ldots, \hat{p}_{i_1 m})$ und $(\hat{p}_{i_2 1}, \ldots, \hat{p}_{i_2 m})$ asymptotisch unabhängig.

Hieraus folgt, dass man Hypothesen an die p_{ij} oder auch einen Vergleich der Parameter zweier Markov Ketten wie für die Multinomialverteilung (pro Zeile von P) durch χ^2-Tests ausführen kann, und diese für mehrere Spalten gemeinsam (wegen der asymptotischen Unabhängigkeit) einfach addieren kann.

References

[1] R. R. Baker und M. A. Bellis (1993). Human sperm competition: ejaculate adjustment by males and the function of masturbation, *Animal behaviour* **46**, 861–885.

[2] P. Billingsley(1961).Statistical methods in Markov chains. *Ann. Math. Statist.* **32** , 12–40

[3] H. Chen, J. Chen and J. Kalbfleisch (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Statist. Soc. B*, **63**, 19-29.

[4] D. I. Cherny, G. Striker, V. Subramaniam, S. D. Jett, E. Palecek and T. M. Jovin (1999). DNA Bending Due to Specific p53 and p53 Core Domain-DNA Interactions Visualized by Electron Microscopy. *J. Mol. Biol.*, **294**, 1015-1026.

[5] E. Brunner (1997). *Statistik Teil 1*, Skript Universität Göttingen.

[6] M. Falk, R. Becker and F. Marohn (1995). *Angewandte Statistik mit SAS*. Springer.

[7] C. P. Farrington, G. Manning (1990. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in medicine*, **9**, 1447-1454.

[8] D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway und E. Ostrowski (1994). *A Handbook of Small Data Sets*. Chapman & Hall.

[9] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

[10] H. Läuter and R. Pincus (1989). *Mathematisch-Statistische Datenanalyse*. Oldenbourg Verlag, München-Wien.

[11] A.J. Lee and A.J. Scott (1986). Ultrasound in ante-natal diagnosis. *In* R.J. Brook, G.C. Arnold, T.H. Hassard and R.M. Pringle (Eds.), *The Fascination of Statistics*. Marcel Dekker, New York.

[12] K. V. Mardia, P. E. Jupp (2000). *Directional statistics*. John Wiley & Sons.

[13] F. H. Ruymgaart (2002). *A short introduction to inverse statistical inference*. Skript University Göttingen.

[14] W. N. Venables und B.D Ripley (2003). *Modern Applied Statistics with S* 4. Auflage. Springer.