WS 2009/2010 *Name:*
Matrikel-Nr.:

KLAUSUR FÜR DAS UNABHÄNGIGE PRÜFUNGSMODUL STOCHASTIK II – STATISTIK I

16. DEZEMBER 2009
HILFSMITTEL: A4 BLATT MIT NOTIZEN, TR

Die Prüfung besteht aus zwei Teilen, die insgesamt 300 Punkte ergeben:
Teil 1: 10 Multiple Choice (MC) Aufgaben mit jeweils 5 Punkten
Teil 2: 5 von 8 Aufgaben mit jeweils 50 Punkten

Bitte ankreuzen, welche Aufgaben bewertet werden sollen (5 von 8)!

Aufgabe:	1	2	3	4	5	6	7	8	MC	Gesamt
Bewerten:										
Punkte:										

Schreiben Sie bitte auf jedes Blatt oben rechts Ihren Namen!

Viel Erfolg!

Teil 1. Multiple Choice

Zu jeder Frage ist genau eine richtige Antwortmöglichkeit vorgegeben. Tragen Sie Ihre Lösungen in die Kästchen auf der **übernächsten** Seite ein. Die Rückseiten der Blätter können Sie für Berechnungen sowie zu Anmerkungen und Erläuterungen Ihrer Lösungen verwenden.

(1) Die Resultate einer Studie in Thailand über eine Impfung für AIDS sind im Herbst veröffentlicht worden. 74 von etwa 8000 Personen, die das Placebo erhalten haben, bekamen AIDS. 51 von etwa 8000 Leute, die tatsächlich geimpft worden sind, bekamen AIDS. Um die Signifikanz zu überprüfen ist die folgende Analyse in R durchgeführt worden:

```
x<-matrix(c(74,51,7947,7970),ncol=2)
chisq.test(x,simulate.p.value = TRUE, B = 10000)
X-squared = 4.2652, df = NA, p-value = 0.04870
```

Was ist richtig?
(a) Die Analyse ist falsch, weil B die Stichprobengröße sein muss.
(b) Da eine 2×2 Tabelle analysiert worden ist, muss df=1 sein.
(c) Fast 500 der Simulationen erreichten einen höheren X^2 Wert als 4.2652.
(d) 487 der Simulationen hatten einen niedrigeren X^2 Wert als 4.2652.

(2) Welches Ziel kann mit einer Regressionsanalyse nicht erzielt werden?
(a) Schätzen der Parameter einer bekannten funktionalen Beziehung.

(b) Erkennen eines funktionalen Zusammenhangs.
(c) Interpolation fehlender Werte.
(d) Nachweis einer allgemeinen Unabhängigkeitsbeziehung.

(3) Eine χ^2-verteilte Zufallsvariable mit 200 Freiheitsgraden kann approximiert werden durch:
(a) Eine Normalverteilung mit $\mu = 0$ und $\sigma = 200$.
(b) Eine Normalverteilung mit $\mu = 200$ und $\sigma = 200$.
(c) Eine Normalverteilung mit $\mu = 200$ und $\sigma = 20$.
(d) Eine Normalverteilung mit $\mu = 20$ und $\sigma = 200$.

(4) An neun Patienten wurden zwei Schlafmittel erprobt, wobei folgende zusätzliche Schlafzeiten in Stunden beobachtet wurden:

Patient	1	2	3	4	5	6	7	8	9
X (Medikament1)	0,50	1,48	0,65	0,16	0,83	-0,24	0,63	0,00	0,40
Y (Medikament2)	0,27	0,20	0,25	0,14	-0,90	0,10	-0,80	0,32	-0,30

Man führt einen Wilcoxon-Rangsummen-Test mit $H_0 : F_X = F_Y$ durch, wobei F_X und F_Y die Verteilungsfunktionen von X und Y sind.

Welche der folgenden Aussagen ist nicht richtig?
(a) Es wird dabei geprüft, ob sich die Verteilungen der Grundgesamtheiten zweier Stichproben unterscheiden.
(b) Die Summe der Rangzahlen von X ist $w_X = 110$.
(c) Unter $H_1 : F_X \geq F_Y$ gilt: P-Wert $= P(W_X \leq x_X|H_0) = 0,0157$.
(d) Unter $H_1 : F_X \neq F_Y$ gilt: P-Wert $= 2P(W_X \geq x_X|H_0) = 0,0314$.

(5) **(Fortsetzung von (3)):** Man führt diesmal einen gepaarten T-Test durch mit $H_0 : \mu_X \leq \mu_Y$, wobei μ_X und μ_Y die Erwartungswerte von X und Y sind.

Welche der folgenden Aussagen ist nicht richtig?
(a) Es wird dabei angenommen, dass die beiden Stichproben paarweise abhängig sind.
(b) Es wird dabei vorausgesetzt, dass die Differenzen der beiden Stichproben unabhängig voneinander sind.
(c) Die Teststatistik ist t-verteilt mit 8 Freiheitsgraden.
(d) H_0 ist zum Signifikanzniveau $\alpha = 0,02$ abzulehnen.

(6) In einer Überprüfung der Unabhängigkeit von Beruf (4 Katgegorien) und Ausbildung (3 Kategorien) war der χ^2-Statistik $14,4$. Der p-Wert war
(a) $> 0,05$
(b) $< 0,05$ und $> 0,01$
(c) $< 0,05$ und $> 0,001$
(d) $< 0,001$

(7) Sei `myvar` ein asympthotisch erwartungstreuer Schätzer für die Varianz. Für welche Simulation `SIM` lässt der Ausdruck
`mean( apply(SIM,2,myvar) - apply(SIM,2,var) )`
das gerinste Ergebnis erwarten?
(a) SIM = replicate(100, runif(20,min=0,max=100))
(b) SIM = replicate(100, runif(100,min=0,max=100))
(c) SIM = replicate(1000, runif(20,min=0,max=100))
(d) SIM = replicate(100, runif(100,min=0,max=300))

(8) Der Bayes-Schätzer einer apriori Betaverteilung mit Parametern a und b und einer Stichprobe $X \sim Bin(N, p)$ ist:
(a) $\hat{p} = \frac{a+b}{a+x}$
(b) $\hat{p} = \frac{a+x}{b+N}$
(c) $\hat{p} = \frac{a+b}{a+b+N}$
(d) $\hat{p} = \frac{a+x}{a+b+N}$

(9) In den 17 Heimspielen der Bundesligasaison 08/09 waren beim FC Augsburg durchschnittlich 15571 Zuschauer mit einer Standardabweichung von 4986 im Rosenaustadion. In der neuen Impuls-Arena waren bei den ersten drei Heimspielen der Saison 09/10 im Schnitt 17767 Zuschauer bei einer Standardabweichung von 2223. Was lautet der Wert der Teststatistik eines 2-Stichproben-t-tests?
(a) -0.402
(b) -1.245
(c) -0.737
(d) -0.461

(10) Es wird ein Wilcoxon-Rangsummentest mit Teststatistik U_X für die Stichproben $X = X_1, ..., X_n$ und $Y = Y_1, ..., Y_m$ zum Signifikanzniveau α durchgeführt, wobei auf $X_{med} \leq Y_{med}$ getestet wird. Welches Quantil der Wilcoxon-Verteilung wird zur Testentscheidung benutzt?
(a) `qwilcox(alpha, n, m)`
(b) `qwilcox(alpha/2, n, m)` und `qwilcox(1 - alpha/2, n, m)`
(c) `qwilcox(1 - alpha, n, m)`
(d) `qwilcox(alpha/2, n, m)` oder `qwilcox(1 - alpha/2, n, m)`

(11) Welche Annahmen hat ein t-Test zum Vergleich der Erwartungswerte zweier Stichproben X und Y der Größen n und m?
(a) $X_1, ..., X_n, Y_1, ..., Y_m \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ und $n = m$
(b) $X_1, ..., X_n \overset{i.i.d.}{\sim} N(\mu_x, \sigma_x^2)$ und $Y_1, ..., Y_n \overset{i.i.d.}{\sim} N(\mu_y, \sigma_y^2)$
(c) $X_i - Y_j \overset{i.i.d.}{\sim} N(\mu_x - \mu_y, (\sigma_x + \sigma_y)^2)$ und σ_x, σ_y bekannt
(d) $X_1, ..., X_n$ i.i.d., $Y_1, ..., Y_m$ i.i.d. und $X_1, ..., X_n, Y_1, ..., Y_m$ unabhängig

(12) Welche Aussage über den Boxplot einer Zufallsstichprobe X ist korrekt?

(a) Die inneren Zäune sind
$Median(X) + 1.5 \cdot IQR(X)$ und $Median(X) - 1.5 \cdot IQR(X)$.
(b) Die äußeren Zäune sind das Minimum und Maximum der Stichprobe X.
(c) Er zeigt die Quantile von X zu 0%, 25%, 50%, 75% und 100%.
(d) Ist X normalverteilt, so ist der Boxplot symmetrisch und es gilt
$Median(X) = Mittelwert(X)$.

(13) Welche Aussage über Rangsummentests ist falsch?
(a) Der *Wilcoxon-Rangsummen*-Test ist invariant gegenüber Transformationen, die die Rangfolge innerhalb der beiden Stichproben nicht verändern.
(b) Für die *Wilcoxon-Rangsummen*-Statistik W gilt
$P(W \leq c + \frac{n(n+1)}{2}, m = a, n = b) = P(W \leq c + \frac{n(n+1)}{2}, m = b, n = a)$
(c) Die ersten beiden zentrierten Momente der *Mann-Whitney-U*-Statistik sind $M_1(U) = n \cdot m/2$ und $M_2(U) = n \cdot m(n + m + 1)/12$
(d) Die *Wilcoxon-Signed-Rank*-Statistik ist die Summe der Ränge aller positiven Differenzen zweier unabhängiger, kontinuierlicher Stichproben.

Antworten:

1	2	3	4	5	6	7	8	9	10

Verteilungstabelle der Standardnormalverteilung

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.000	0.5000	0.750	0.7734	1.500	0.9332	2.250	0.9878
0.025	0.5100	0.775	0.7808	1.525	0.9364	2.275	0.9885
0.050	0.5199	0.800	0.7881	1.550	0.9394	2.300	0.9893
0.075	0.5299	0.825	0.7953	1.575	0.9424	2.325	0.9900
0.100	0.5398	0.850	0.8023	1.600	0.9452	2.350	0.9906
0.125	0.5497	0.875	0.8092	1.625	0.9479	2.375	0.9912
0.150	0.5596	0.900	0.8159	1.650	0.9505	2.400	0.9918
0.175	0.5695	0.925	0.8225	1.675	0.9530	2.425	0.9923
0.200	0.5793	0.950	0.8289	1.700	0.9554	2.450	0.9929
0.225	0.5890	0.975	0.8352	1.725	0.9577	2.475	0.9933
0.250	0.5987	1.000	0.8413	1.750	0.9599	2.500	0.9938
0.275	0.6083	1.025	0.8473	1.775	0.9621	2.525	0.9942
0.300	0.6179	1.050	0.8531	1.800	0.9641	2.550	0.9946
0.325	0.6274	1.075	0.8588	1.825	0.9660	2.575	0.9950
0.350	0.6368	1.100	0.8643	1.850	0.9678	2.600	0.9953
0.375	0.6462	1.125	0.8697	1.875	0.9696	2.625	0.9957
0.400	0.6554	1.150	0.8749	1.900	0.9713	2.650	0.9960
0.425	0.6646	1.175	0.8800	1.925	0.9729	2.675	0.9963
0.450	0.6736	1.200	0.8849	1.950	0.9744	2.700	0.9965
0.475	0.6826	1.225	0.8897	1.975	0.9759	2.725	0.9968
0.500	0.6915	1.250	0.8944	2.000	0.9772	2.750	0.9970
0.525	0.7002	1.275	0.8988	2.025	0.9786	2.775	0.9972
0.550	0.7088	1.300	0.9032	2.050	0.9798	2.800	0.9974
0.575	0.7174	1.325	0.9074	2.075	0.9810	2.825	0.9976
0.600	0.7257	1.350	0.9115	2.100	0.9821	2.850	0.9978
0.625	0.7340	1.375	0.9154	2.125	0.9832	2.875	0.9980
0.650	0.7422	1.400	0.9192	2.150	0.9842	2.900	0.9981
0.675	0.7502	1.425	0.9229	2.175	0.9852	2.925	0.9983
0.700	0.7580	1.450	0.9265	2.200	0.9861	2.950	0.9984
0.725	0.7658	1.475	0.9299	2.225	0.9870	2.975	0.9985

Verteilungstabelle der Poissonverteilung

λ=15		λ=7		λ=6		λ=3	
x	$P(X \leq x)$	x	$P(X \leq x)$	x	$P(X \leq x)$	x	$P(X \leq x)$
1	0	0.5	0.0009	0.5	0.0025	0.2	0.0498
2	0	1	0.0073	1	0.0174	0.4	0.0498
3	0.0002	1.5	0.0073	1.5	0.0174	0.6	0.0498
4	0.0009	2	0.0296	2	0.062	0.8	0.0498
5	0.0028	2.5	0.0296	2.5	0.062	1	0.1991
6	0.0076	3	0.0818	3	0.1512	1.2	0.1991
7	0.018	3.5	0.0818	3.5	0.1512	1.4	0.1991
8	0.0374	4	0.173	4	0.2851	1.6	0.1991
9	0.0699	4.5	0.173	4.5	0.2851	1.8	0.1991
10	0.1185	5	0.3007	5	0.4457	2	0.4232
11	0.1848	5.5	0.3007	5.5	0.4457	2.2	0.4232
12	0.2676	6	0.4497	6	0.6063	2.4	0.4232
13	0.3632	6.5	0.4497	6.5	0.6063	2.6	0.4232
14	0.4657	7	0.5987	7	0.744	2.8	0.4232
15	0.5681	7.5	0.5987	7.5	0.744	3	0.6472
16	0.6641	8	0.7291	8	0.8472	3.2	0.6472
17	0.7489	8.5	0.7291	8.5	0.8472	3.4	0.6472
18	0.8195	9	0.8305	9	0.9161	3.6	0.6472
19	0.8752	9.5	0.8305	9.5	0.9161	3.8	0.6472
20	0.917	10	0.9015	10	0.9574	4	0.8153
21	0.9469	10.5	0.9015	10.5	0.9574	4.2	0.8153
22	0.9673	11	0.9467	11	0.9799	4.4	0.8153
23	0.9805	11.5	0.9467	11.5	0.9799	4.6	0.8153
24	0.9888	12	0.973	12	0.9912	4.8	0.8153
25	0.9938	12.5	0.973	12.5	0.9912	5	0.9161
26	0.9967	13	0.9872	13	0.9964	5.2	0.9161
27	0.9983	13.5	0.9872	13.5	0.9964	5.4	0.9161
28	0.9991	14	0.9943	14	0.9986	5.6	0.9161
29	0.9996	14.5	0.9943	14.5	0.9986	5.8	0.9161
30	0.9998	15	0.9976	15	0.9995	6	0.9665
31	0.9999	15.5	0.9976	15.5	0.9995	6.2	0.9665
32	1	16	0.999	16	0.9998	6.4	0.9665
33	1	16.5	0.999	16.5	0.9998	6.6	0.9665
34	1	17	0.9996	17	0.9999	6.8	0.9665
35	1	17.5	0.9996	17.5	0.9999	7	0.9881

Quantile $\chi^2_{df;1-\alpha}$ der χ^2-Verteilung

df ╲ α	0.100	0.050	0.025	0.020	0.010	0.005	0.0025	0.001	0.0005
1	2.705	3.841	5.023	5.411	6.634	7.879	9.140	10.827	12.115
2	4.605	5.991	7.377	7.824	9.210	10.596	11.982	13.815	15.201
3	6.251	7.814	9.348	9.837	11.344	12.838	14.320	16.266	17.730
4	7.779	9.487	11.143	11.667	13.276	14.860	16.423	18.466	19.997
5	9.236	11.070	12.832	13.388	15.086	16.749	18.385	20.515	22.105
6	10.644	12.591	14.449	15.033	16.811	18.547	20.249	22.457	24.102
7	12.017	14.067	16.012	16.622	18.475	20.277	22.040	24.321	26.017
8	13.361	15.507	17.534	18.168	20.090	21.954	23.774	26.124	27.868
9	14.683	16.918	19.022	19.679	21.665	23.589	25.462	27.877	29.665
10	15.987	18.307	20.483	21.160	23.209	25.188	27.112	29.588	31.419
11	17.275	19.675	21.920	22.617	24.724	26.756	28.729	31.264	33.136
12	18.549	21.026	23.336	24.053	26.216	28.299	30.318	32.909	34.821
13	19.811	22.362	24.735	25.471	27.688	29.819	31.883	34.528	36.477
14	21.064	23.684	26.118	26.872	29.141	31.319	33.426	36.123	38.109
15	22.307	24.995	27.488	28.259	30.577	32.801	34.949	37.697	39.718

Quantile $t_{n;1-\alpha}$ der t-Verteilung

n ╲ α	0.100	0.050	0.025	0.020	0.010	0.005	0.0025	0.001	0.0005
1	3.077	6.313	12.706	15.894	31.820	63.656	127.321	318.308	636.619
2	1.885	2.919	4.302	4.848	6.964	9.924	14.089	22.327	31.599
3	1.637	2.353	3.182	3.481	4.540	5.840	7.453	10.214	12.923
4	1.533	2.131	2.776	2.998	3.746	4.604	5.597	7.173	8.610
5	1.475	2.015	2.570	2.756	3.364	4.032	4.773	5.893	6.868
6	1.439	1.943	2.446	2.612	3.142	3.707	4.316	5.207	5.958
7	1.414	1.894	2.364	2.516	2.997	3.499	4.029	4.785	5.407
8	1.396	1.859	2.306	2.448	2.896	3.355	3.832	4.500	5.041
9	1.383	1.833	2.262	2.398	2.821	3.249	3.689	4.296	4.780
10	1.372	1.812	2.228	2.359	2.763	3.169	3.581	4.143	4.586
11	1.363	1.795	2.200	2.328	2.718	3.105	3.496	4.024	4.436
12	1.356	1.782	2.178	2.302	2.680	3.054	3.428	3.929	4.317
13	1.350	1.770	2.160	2.281	2.650	3.012	3.372	3.851	4.220
14	1.345	1.761	2.144	2.263	2.624	2.976	3.325	3.787	4.140
15	1.340	1.753	2.131	2.248	2.602	2.946	3.286	3.732	4.072
16	1.336	1.745	2.119	2.235	2.583	2.920	3.251	3.686	4.014
17	1.333	1.739	2.109	2.223	2.566	2.898	3.222	3.645	3.965
18	1.330	1.734	2.100	2.213	2.552	2.878	3.196	3.610	3.921
19	1.327	1.729	2.093	2.204	2.539	2.860	3.173	3.579	3.883
20	1.325	1.724	2.085	2.196	2.527	2.845	3.153	3.551	3.849
25	1.316	1.708	2.059	2.166	2.485	2.787	3.078	3.450	3.725
50	1.298	1.675	2.008	2.108	2.403	2.677	2.936	3.261	3.496
100	1.290	1.660	1.983	2.080	2.364	2.625	2.870	3.173	3.390

Verteilungsfunktion der Binomialverteilung $B_{n;p}(k)$ für $p = 0.5$.

$n \diagdown k$	0	1	2	3	4	5	6	7	8	9	10
1	0.500	1.000									
2	0.250	0.750	1.000								
3	0.125	0.500	0.875	1.000							
4	0.062	0.313	0.687	0.938	1.000						
5	0.031	0.188	0.500	0.812	0.969	1.000					
6	0.016	0.109	0.344	0.656	0.891	0.984	1.000				
7	0.008	0.063	0.227	0.500	0.773	0.938	0.992	1.000			
8	0.004	0.035	0.145	0.363	0.637	0.855	0.965	0.996	1.000		
9	0.002	0.020	0.090	0.254	0.500	0.746	0.910	0.980	0.998	1.000	
10	0.001	0.011	0.055	0.172	0.377	0.623	0.828	0.945	0.989	0.999	1.000

Kritische Werte $W_{n;\alpha}$ des Vorzeichenrangtests von Wilcoxon

$n \diagdown \alpha$	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
4	0	0	0	1	9	10	10	10
5	0	0	1	3	12	14	15	15
6	0	1	3	4	17	18	20	21
7	1	3	4	6	22	24	25	27
8	2	4	6	9	27	30	32	34
9	4	6	9	11	34	36	39	41
10	6	9	11	15	40	44	46	49
11	8	11	14	18	48	52	55	58
12	10	14	18	22	56	60	64	68
13	13	18	22	27	64	69	73	78
14	16	22	26	32	73	79	83	89
15	20	26	31	37	83	89	94	100
16	24	30	36	43	93	100	106	112
17	28	35	42	49	104	111	118	125
18	33	41	48	56	115	123	130	138
19	38	47	54	63	127	136	143	152
20	44	53	61	70	140	149	157	166

Kritische Werte $W_{n;m;\alpha}$ des Rangsummentests von Wilcoxon

$\alpha = 0.05$	m																	
n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	7	8	9	9	10	10	11	12	12	13	14	14	15	16	16	17	18
4	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29
5	17	18	20	21	22	24	25	27	28	29	31	32	34	35	36	38	39	41
6	24	25	27	29	30	32	34	36	38	39	41	43	45	47	48	50	52	54
7	31	33	35	37	40	42	44	46	48	50	53	55	57	59	62	64	66	68
8	40	42	45	47	50	52	55	57	60	63	65	68	70	73	76	78	81	84
9	49	52	55	58	61	64	67	70	73	76	79	82	85	88	91	94	97	100
10	60	63	67	70	73	76	80	83	87	90	93	97	100	104	107	111	114	118
11	72	75	79	83	86	90	94	98	101	105	109	113	117	121	124	128	132	136
12	84	88	92	96	100	105	109	113	117	121	126	130	134	139	143	147	151	156
13	98	102	107	111	116	120	125	129	134	139	143	148	153	157	162	167	172	176
14	113	117	122	127	132	137	142	147	152	157	162	167	172	177	183	188	193	198
15	128	133	139	144	149	154	160	165	171	176	182	187	193	198	204	209	215	221
16	145	151	156	162	167	173	179	185	191	197	202	208	214	220	226	232	238	244
17	163	169	174	180	187	193	199	205	211	218	224	231	237	243	250	256	263	269
18	181	188	194	200	207	213	220	227	233	240	247	254	260	267	274	281	288	295
19	201	208	214	221	228	235	242	249	256	263	271	278	285	292	300	307	314	321
20	222	229	236	243	250	258	265	273	280	288	295	303	311	318	326	334	341	349

$\alpha = 0.01$	m																	
n	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	6	6	6	6	7	7	8	8	8	9	9	9	10	10	11	11	11	12
4	10	10	11	12	12	13	14	14	15	16	16	17	18	18	19	20	20	21
5	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
6	21	23	24	25	26	28	29	30	31	33	34	35	37	38	40	41	42	44
7	29	30	32	33	35	36	38	40	41	43	45	46	48	50	52	53	55	57
8	37	39	41	43	44	46	48	50	52	54	57	59	61	63	65	67	69	71
9	47	49	51	53	55	57	60	62	64	67	69	72	74	77	79	82	84	86
10	57	59	62	64	67	69	72	75	78	80	83	86	89	92	94	97	100	103
11	68	71	74	76	79	82	85	89	92	95	98	101	104	108	111	114	117	120
12	81	84	87	90	93	96	100	103	107	110	114	117	121	125	128	132	135	139
13	94	97	101	104	108	112	115	119	123	127	131	135	139	143	147	151	155	159
14	108	112	116	119	123	128	132	136	140	144	149	153	157	162	166	171	175	179
15	124	128	132	136	140	145	149	154	158	163	168	172	177	182	187	191	196	201
16	140	144	149	153	158	163	168	173	178	183	188	193	198	203	208	213	219	224
17	158	162	167	172	177	182	187	192	198	203	209	214	220	225	231	236	242	247
18	176	181	186	191	196	202	208	213	219	225	231	237	242	248	254	260	266	272
19	195	200	206	211	217	223	229	235	241	247	254	260	266	273	279	285	292	298
20	216	221	227	233	239	245	251	258	264	271	278	284	291	298	304	311	318	325

Teil 2. **Bearbeiten Sie** 5 **der** 8 **Aufgaben!**

1. Blitzhäufigkeit

Die Blitzhäufigkeit wird anhand der Einschläge pro Quadratkilometer gemessen. Folgende Werte liegen für Deutschland, Österreich und die Schweiz vor.

Deutschland: 6.5 Einschläge pro km^2 und Jahr
Österreich: 2.4 Einschläge pro km^2 und Jahr
Schweiz: 9.1 Einschläge pro km^2 und Jahr

(a) Es wurden an einem bestimmten Quadratkilometer von 2004 - 2008 (fünf Jahre) 29 Blitze gezählt. Bestimmen Sie mittels der Maximum-Likelihood Methode das Land, indem am ehesten dieser Ort liegt

(b) Erklären Sie die Güteeigenschaften Erwartungstreue, Konsistenz und Effizienz. Welche davon erfüllen Maximum-Likelihood Schätzer.

(c) Folgende Tabelle gibt die Anzahl Blitze pro km^2 von 1992-2007 für die österreichischen Bundesländer Vorarlberg und Wien wieder:

V	0.9	1.3	1.2	0.8	0.7	0.4	0.5	0.5	1.4	0.7	1.5	1.3	1.1	0.6	1.3	0.8
W	0.5	0.7	0.6	1.0	0.9	0.6	1.6	0.6	1.7	0.9	1.5	2.9	1.1	0.8	1.2	2.8

Verwenden Sie einen geeigneten Test (mit Begründung!), um zu untersuchen ob in beiden Bundesländern die Anzahl Blitze pro km^2 verschieden ist.

(d) Wie könnte man die Daten aus (c) grafisch darstellen?

2. Testtheorie

(a) Erklären Sie die folgenden Begriffe: Fehler 1. Art und Fehler 2. Art.

(b) Ist ein Einstichproben-Gaußtest mit $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$ (μ_0 gegeben) unverfälscht? Begründen Sie Ihre Aussage ausführlich.

(c) Für einen approximativen Binomialtest (Approximation durch Normalverteilung) lautet die Nullhypothese $H_0 : p = 0,3$ gegen $H_1 : p \neq 0,3$. Skizzieren Sie die Gütefunktion des Tests unter dem Signifikanzniveau von $0,05$ für $n = 30$, und $n = 100$.

(d) Schneckenzüchter A liefert einem Restaurant große graue Schnecken, die durchschnittlich mindestens $15g$ wiegen sollen. Es sei bekannt, dass das Gewicht seiner Schnecken normalverteilt ist mit der Standardabweichung $3g$ und einem unbekannten Erwartungswert μ. Als Kontrolleur dieses Restaurants nehmen Sie eine Stichprobe vom Umfang 25 und testen mit einem Gaußtest zum Signifikanzniveau von $0,05$, ob man $\mu \geq 15$ ablehnen darf (falls ja, wird die Lieferung abgelehnt). Skizzieren Sie die Gütefunktion Ihres Tests. Welcher Fehler 2. Art ergibt sich, wenn μ in Wirklichkeit $13g$ beträgt?

3. Wählerwanderungen

In einer Studie über Wählerwanderungen zwischen den letzten zwei Bundestagswahlen sind 1500 Personen befragt worden. Zweidrittel behaupten in beiden Wahlen, ihre Stimme abgegeben zu haben und können die Parteien nennen, wofür sie gestimmt haben. Unter den 10% dieser Personen, die sagen, dass sie 2005 für die Grünen gestimmt haben, haben 2009 10 für die FDP gestimmt. Berechnen Sie ein 95% Konfidenzintervall für die Wahrscheinlichkeit, dass eine Person, die 2005 für die Grünen gestimmt hat, 2009 für die FDP stimmte. Welche Annahmen haben Sie gemacht?

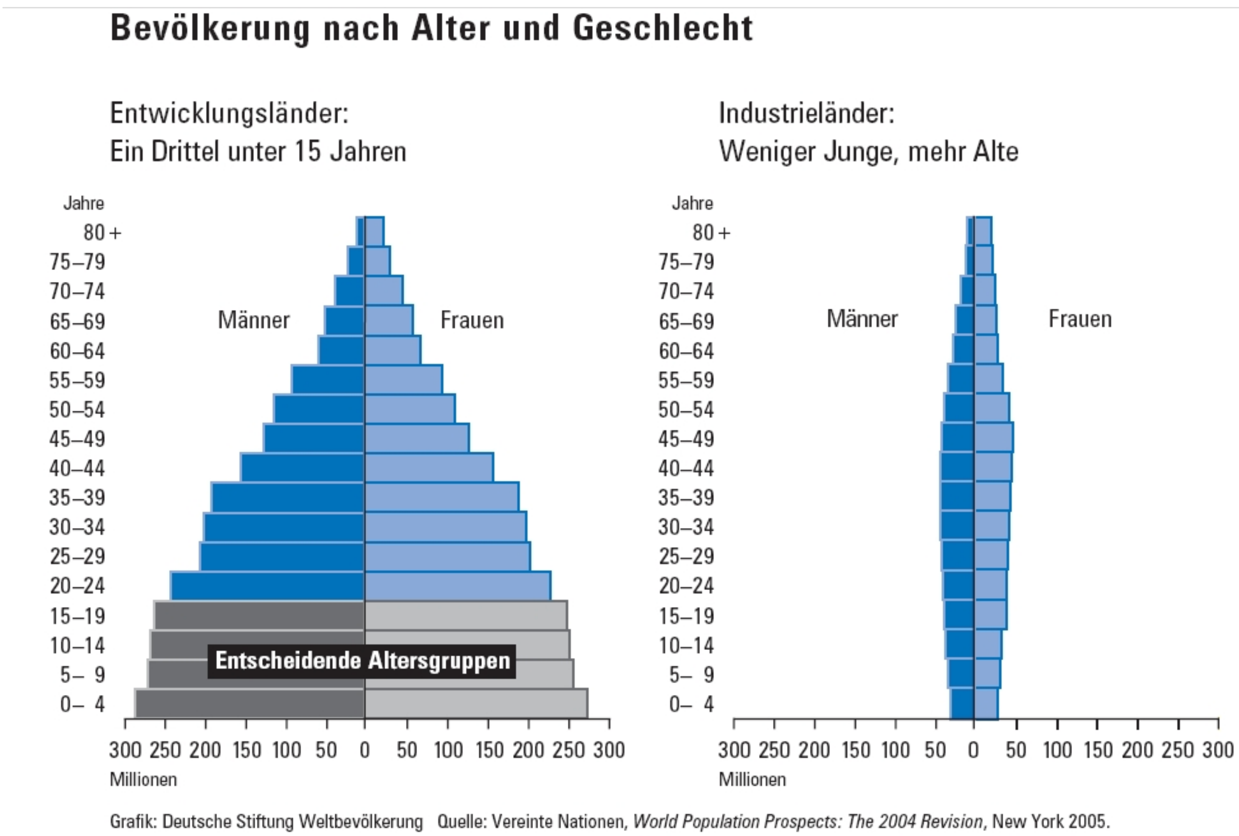
Ein Bayesianer meint, dass seine apriori Verteilung für diese Wahrscheinlichkeit, p,

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$$

war. Beschreiben Sie, wie Sie seine aposteriori Verteilung berechnen würden. Wie soll ein 95% Intervall für die Wahrscheinlichkeit p mit diesem Modell berechnet werden?

Ein Kritiker beanstandet an beiden Verfahren, dass 2005 die Grünen $8,1\%$ der Stimmen erhalten haben, so dass die Stichprobe nicht gültig sein kann. Weiterhin kritisiert er, dass die Anzahl Leute in der Stichprobe, die 2005 für die Grünen gestimmt haben, eine Zufallsvariable ist, und das wird weder in dem einen, noch in dem anderen Verfahren berücksichtigt. Was sagen Sie zu diesen beiden Kritikpunkten?

4. Altersstruktur

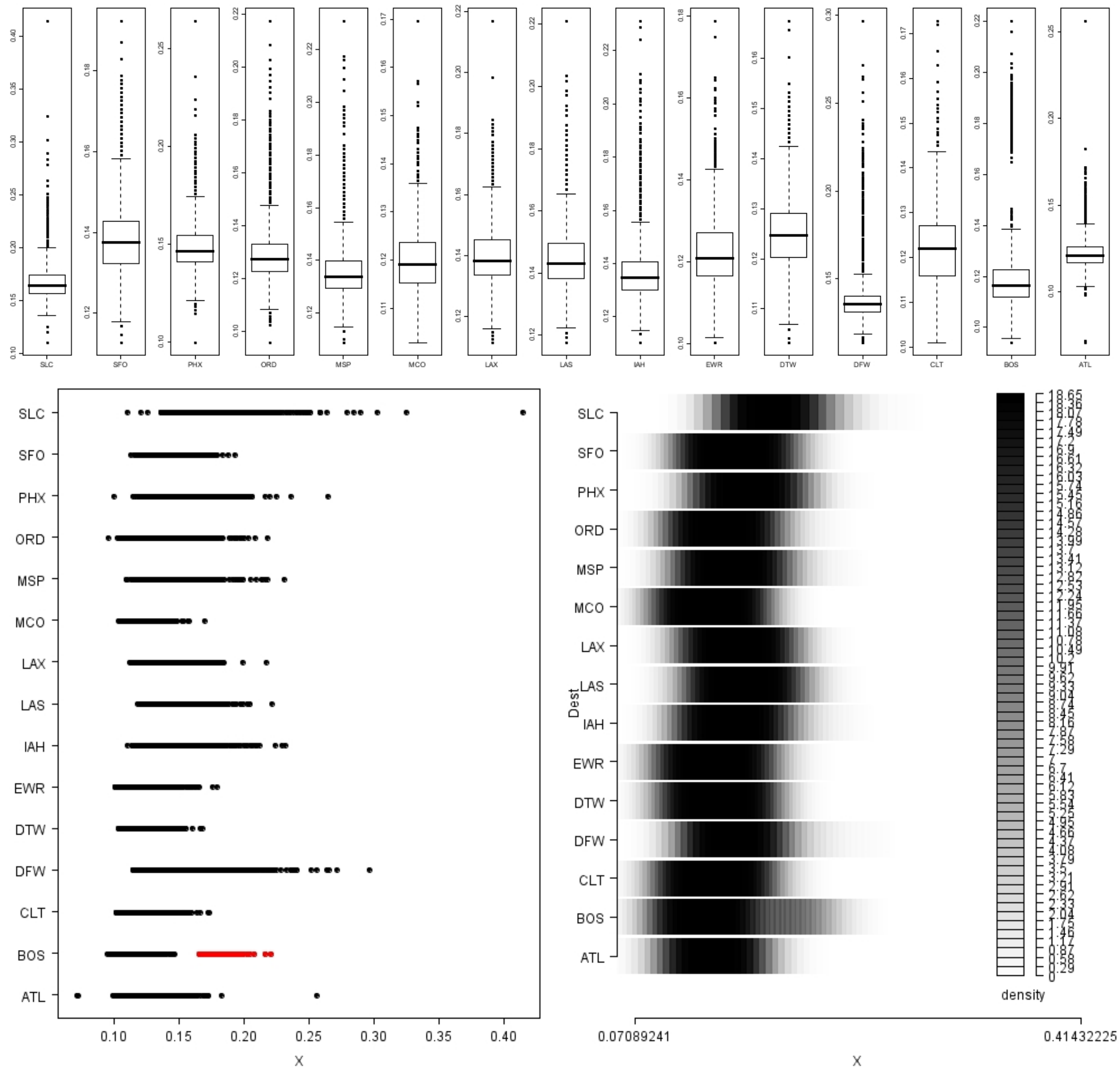


(a) Erläutern Sie das Thema der Grafik und beschreiben Sie ihren Aufbau.
(b) Beurteilen Sie die Grafik in Bezug auf Suggestivität, Übersichtlichkeit und Ästhetik.
(c) Nennen Sie die drei Ihrer Meinung nach wichtigsten Kritikpunkte. Mit welcher aus der Vorlesung bekannten Grafik könnten die beabsichtigten Vergleiche besser vorgenommen werden? Beschreiben Sie diese.

Aufgrund der starken Ähnlichkeit der Altersverteilungen der beiden Geschlechter sollen diese nun nicht weiter getrennt berücksichtigt werden.

d) Ursprünglich liegen die Werte für jedes Land einzeln vor. Schlagen Sie eine Grafik vor, um Industrieländer und Entwicklungsländer anhand dieser Einzelwerte zu vergleichen ohne diese (wie in obiger Grafik) zu aggregieren? Beschreiben Sie diese kurz.

5. Geschwindigkeiten



Die drei obenstehenden Grafiken visualisieren auf unterschiedliche Weise die durchschnittliche Geschwindigkeit (X) einzelner Flüge auf den Routen von Denver (DEN) zu 15 Zielflughäfen (Dest) innerhalb der USA. Bearbeiten Sie dazu folgende Aufgaben:

(a) Beschreiben Sie möglichst präzise den Aufbau der drei Grafiken.

(b) Beschreiben Sie zu jeder der Grafiken wesentliche Kritikpunkte und schlagen Sie zu jeder der Grafiken Verbesserungen vor.

(c) Unter den Flügen nach Boston (BOS) zeichnen sich zwei getrennte Gruppen ab. Testen Sie mit dem Wilcoxon-Rangsummen-Test, ob die Geschwindigkeit dieser Gruppen signifikant verschieden ist, wenn die beiden Gruppen 100 bzw 30 Flüge umfassen.

6. Regression

Abb.1 links zeigt ein Streudiagramm des Kopfumfangs (head) und des Gewichtes (weight) von Neugeborenen im Klinikum Augsburg.

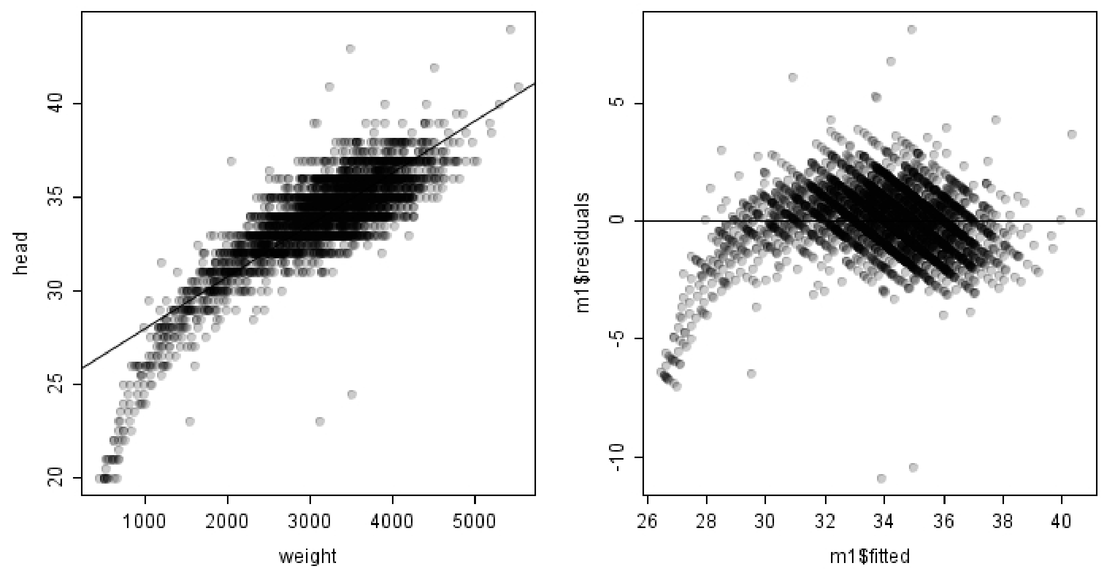


Abbildung 1. Streudiagramme:*head(y)* vs. *weight(x)* und Residuen vs. Modellwerte

```
> summary(lm(head~weight))
    Call: lm(formula = head ~ weight)
    Residuals:    Min     1Q   Median   3Q     Max
            -10.88  -0.72   0.06   0.78   8.11
    Coefficients:   Estimate    Std.Error   t.value Pr(>|t|)
    (Intercept)     2.517e+01   8.755e-02   287.5   <2e-16 ***
    weight          2.791e-03   2.685e-05   103.9   <2e-16 ***

    Residual standard error: 1.29 on 4525 df
    (44 obs. deleted due to missingness)
    Multiple R-squared: 0.7048, Adjusted R-squared: 0.7048
    F-statistic: 10810 on 1 and 4525 df, p-value: <2.2e-16
```

(a) Es wurde ein lineares Regressionsmodell für die Variablen berechnet und die Gerade in das Streudiagramm eingezeichnet.
   i. Erläutern Sie die *F-Statistik*, R^2 und die Signifikanztests der Parameter.
   ii. Beurteilen Sie die Güte des Modells anhand der beiden Grafiken und des nachstehenden R-Outputs. Sind alle Annahmen erfüllt?

(b) Zeigen Sie, dass der Korrelationskoeffizient $cor(X, Y)$ zweier Variablen X und Y dem Koeffizienten β_1 des linearen Regressionsmodell der standardisierten Variablen

$$\tilde{Y} \sim \beta_0 + \beta_1 \tilde{X} \quad \text{mit} \quad \tilde{Y} = \frac{Y - E[Y]}{\sigma_Y} \quad \text{und } \tilde{X} = \frac{X - E[X]}{\sigma_X}$$

entspricht.
Lassen sich Rückschlüsse vom Korrelationskoeffizienten auf die Güte des Modells im Sinne des R^2-Wertes ziehen?

(c) Die Grafik zeigt keinen linearen Verlauf zwischen den beiden Variablen. Auf welche Weise könnte trotzdem ein lineares Regressionsmodell zum Einsatz kommen?

7. Bundestagswahl

(a) Nachfolgende Tabelle zeigt die Bundesergebnisse der Parteien bei den Bundestagswahlen 2009 und 2005. Die Gesamtzahl der Stimmen sei in beiden Jahren als identisch angenommen.

Partei	P09	P05	Diff
CDU/CSU	33.8	35.2	-1.4
GRüNE	10.7	8.1	+2.6
SPD	23.0	34.2	-11.2
DIE LINKE	11.9	8.7	+3.2
FDP	14.6	9.8	+4.8
andere	6.0	4.0	+2.0

i. Die Daten sollen mit Hilfe eines χ^2-Tests untersucht werden. Formulieren Sie die Nullhypothese und berechnen Sie die Statistik aus den relativen Häufigkeiten in Abhängigkeit von der Gesamtzahl der Wähler n.
Was lässt sich aus dieser Abhängigkeit in Bezug auf die Eignung des Tests sagen?

ii. Welche Grafik würden Sie zur Visualisierung der Daten im Hinblick auf Ihre Hypothese verwenden? Beschreiben Sie den Aufbau dieser Grafik präzise.

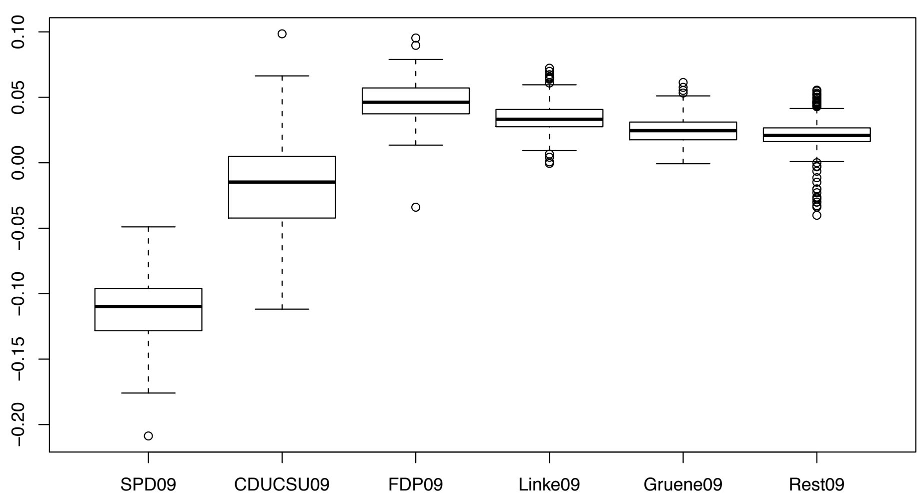
iii. Beurteilen Sie, ob ein T-Test ein geeignetes Mittel zum Vergleich der Werte 2005 und 2009 ist.

8. R und Mondrian

In Deutschland gibt es bei Bundestagswahlen 299 Wahlkreise. In dem Datensatz *wahl* sind Angaben zu allen Wahlkreisen über die Wahlbeteiligung und die Ergebnisse der Parteien bei den Bundestagswahlen 2005 und 2009 enthalten.

(a) Erläutern und interpretieren Sie den folgenden Code und Output in R ab dem Aufruf von *boxplot*().

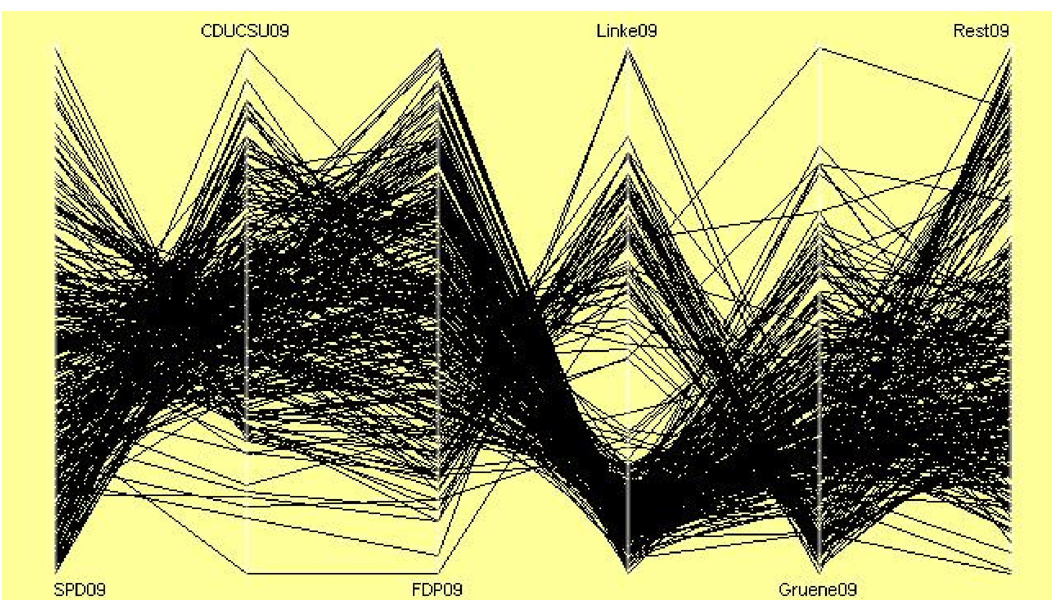
```
> wahl <- read.table("D:/wahl.txt",header=T,sep="\t", quote="")
>
> colnames(wahl)
 [1] "SPD09"          "CDUCSU09"       "FDP09"          "Linke09"
 [5] "Gruene09"       "Rest09"         "SPD05"          "CDUCSU05"
 [9] "FDP05"          "Linke05"        "Gruene05"       "Rest05"
[13] "Name"           "Bundesland"     "beteiligung09" "beteiligung05"
> attach(wahl)
>
> boxplot(wahl[,1:6]-wahl[,7:12])
```

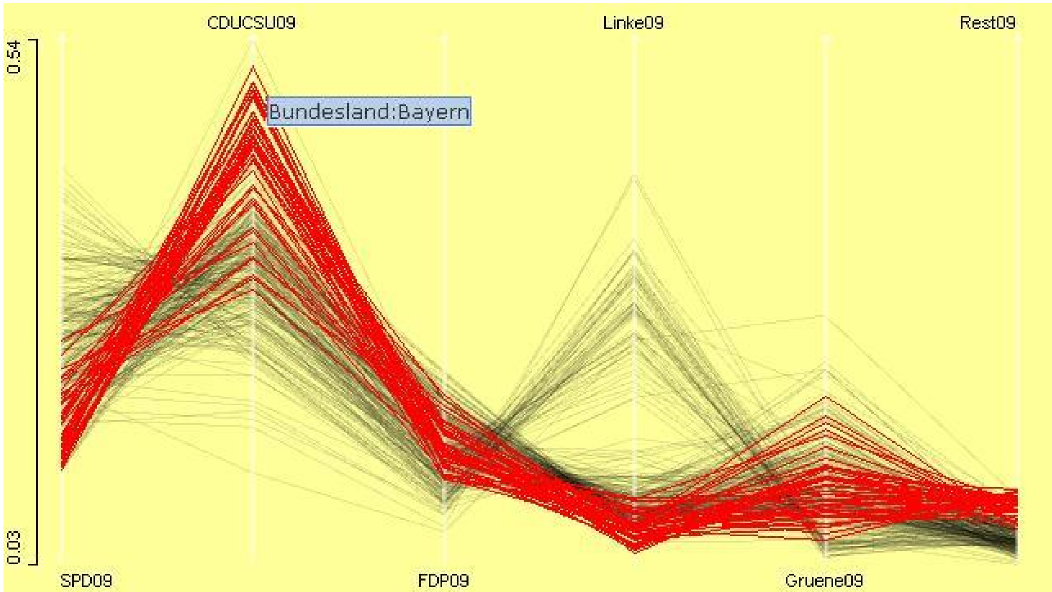



```
> wahl[c(which.min(FDP09-FDP05),which.max(FDP09-FDP05)),13:14]
                         Name        Bundesland
160                 Dresden I           Sachsen
295 Zollernalb - Sigmaringen  Baden-Wrttemberg
```

```
> summary(lm(CDUCSU09 ~ CDUCSU05 + SPD05 + FDP05 + Gruene05))

Call:
lm(formula = CDUCSU09 ~ CDUCSU05 + SPD05 + FDP05 + Gruene05)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0461459 -0.0114430 -0.0004903  0.0112098  0.0468128

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.201782   0.009495  21.251  < 2e-16 ***
CDUCSU05     0.576804   0.012835  44.939  < 2e-16 ***
SPD05       -0.120223   0.015103  -7.960 3.76e-14 ***
FDP05        0.119503   0.050097   2.385   0.0177 *
Gruene05    -0.463359   0.028988 -15.984  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01616 on 294 degrees of freedom
Multiple R-squared: 0.9435,Adjusted R-squared: 0.9427
F-statistic:  1227 on 4 and 294 DF,  p-value: < 2.2e-16
```

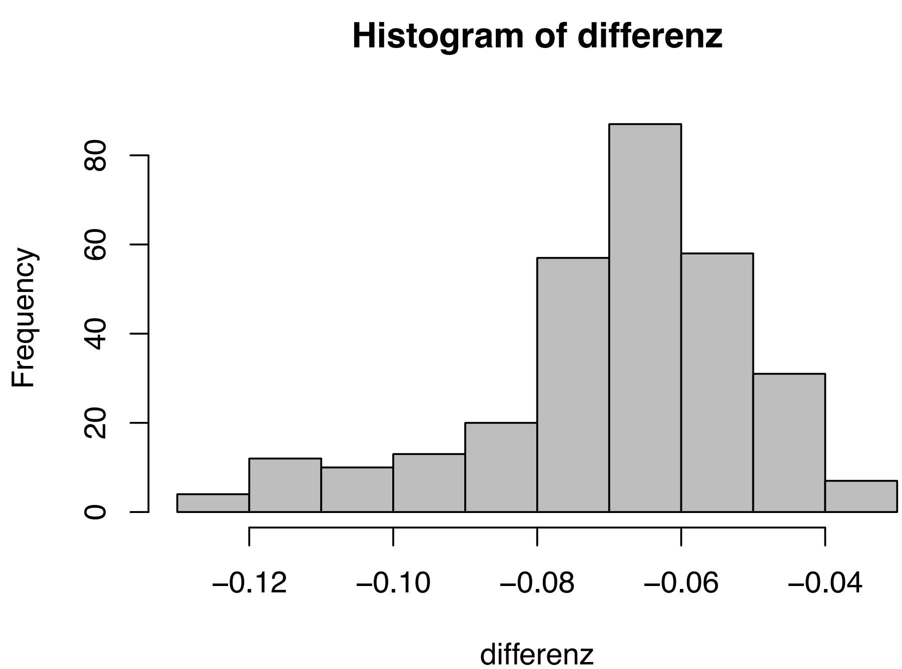
(b) Geben Sie an, um welche Art von Plot es sich bei den zwei Graphiken handelt und was von der ersten zur zweiten Graphik geändert wurde? Interpretieren und beschreiben Sie außerdem die zweite Graphik.





(c) Erläutern und interpretieren Sie folgenden R-Code. Gehen Sie bei dem durchgeführten Test insbesondere auf die getroffenen Annahmen ein.

```
> differenz <- beteiligung09-beteiligung05
> hist(differenz, col="grey")
```



```
> t.test(beteiligung09, beteiligung05, paired=T)

Paired t-test

data:  beteiligung09 and beteiligung05
t = -63.1776, df = 298, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.07066105 -0.06639190
sample estimates:
mean of the differences
            -0.06852647
```