SS 2009 *Name:*

Matrikel-Nr.:

ERSTE KLAUSUR
STOCHASTIK II – STATISTIK I

23. JULI 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) Was gilt für die konjugierte Verteilung nicht?
  (a) Im Allgemeinen existiert nicht immer eine konjugierte Verteilung.
  (b) Die Betaverteilung ist eine konjugierte apriori Familie für die Binomialverteilung.
  (c) Die Betaverteilung ist eine konjugierte apriori Familie für die Poissonverteilung.
  (d) Die Normalverteilung ist eine konjugierte Familie für sich selbst (σ^2 bekannt).

(2) Welche der folgenden Aussagen stimmt nicht?
  (a) Bei nichtparametrischen Tests werden Statistiken benutzt, deren Verteilungen von der Grundgesamtheit unabhängig sind.
  (b) Der Mann-Whitney-U-Test dient nicht zur Überprüfung der Gleichheit der Mediane zweier Verteilungen.
  (c) Im Fall, dass beide Stichproben groß sind, ist die U Statistik approximativ normalverteilt.
  (d) Die Statistiken U im Mann-Whitney-U-Test und W im Wilcoxon-Rangsummen-Test führen zum gleichen Ergebnis.

(3) Was gilt für das Bestimmtheitsmaß R^2 bei der einfachen linearen Regression nicht?
  (a) Für steigende Anzahl Regressoren nimmt R^2 zu.
  (b) $R^2 = r^2$, wobei r der Stichproben-Korrelationskoeffizient ist.
  (c) R^2 liegt immer zwischen 0 und 1.
  (d) Es gibt Auskunft über das Verhältnis zwischen Konfidenz- und Vorhersage-Intervalle.

(4) Welche Aussage über den Gauß-Test ist richtig?
  (a) Der Gauß-Test testet, ob der Erwartungswert einer Exponentialverteilung gleich einem vorgegebenen Wert ist.
  (b) Der Gauß-Test testet, ob die Varianz einer Normalverteilung gleich einem vorgegebenen Wert ist.
  (c) Der Gauß-Test testet, ob der Erwartungswert einer Normalverteilung ungleich einem vorgegebenen Wert ist, bei bekannt vorausgesetzter Varianz.
  (d) Der Gauß-Test testet, ob der Erwartungswert einer Normalverteilung gleich einem vorgegebenen Wert ist, bei unbekannt vorausgesetzter Varianz.

(5) Was gilt ganz allgemein für einen Signifikanztest zum Testniveau α?
  (a) Der Fehler 2. Art bleibt quantifiziert, kleiner oder gleich α.
  (b) Der Fehler 1. Art bleibt quantifiziert, kleiner oder gleich α.
  (c) Der Fehler 1. Art bleibt quantifiziert, größer α.
  (d) Der Fehler 2. Art berechnet sich als '1 − Fehler 1. Art'.

(6) Der R-Befehl `chisq.test(stat1005, simulate.p.value=TRUE, B=5000)`
  (a) testet ob der Datensatz *stat1005* den Mittelwert 5000 hat.
  (b) führt einen Chiquadrat-Test der Nullhypothese keiner Assoziation zwischen den Spalten und Zeilen der Matrix *stat1005* durch.
  (c) erzeugt simulierte Daten aus der Nullhypothese keiner Assoziation zwischen den Spalten und Zeilen in der Matrix *stat1005* und vergleicht X^2 damit.
  (d) erzeugt simulierte Daten aus der Nullhypothese, dass der Korrelationskoeffizient zwischen den Spalten und Zeilen der Matrix *stat1005* den Wert 0 hat, und berechnet den entsprechenden p-Wert.

(7) Parallelkoordinatenplots sind
  (a) eine gute Graphik für multivariate kategorielle Daten.
  (b) eine geometrische Konstruktion, um den Satz von Pythagoras zu beweisen.
  (c) eine Möglichkeit, viele stetige Variablen gleichzeitig darzustellen.
  (d) eine gute Graphik, um Symmetrie in Verteilungen zu erkennen.

(8) Was gilt für den χ^2-Test nicht?
  (a) Er gibt nur gute Resultate, wenn die Stichprobe ausreichend groß ist.
  (b) χ^2-Tests können zur Überprüfung von stochastischer Abhängigkeit zweier Variablen verwendet werden.
  (c) Wenn die Stichprobe ausreichend groß ist, ist das Ergebnis eines χ^2-Tests signifikant.
  (d) Mit χ^2-Tests kann nur die Anpassung diskreter Verteilungen an die Daten untersucht werden.

(9) Das Interval

$$(\hat{a} + \hat{b}x_0) \pm t_{(n-2),0.995}\, s\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}\right)}$$

ist ein
  (a) 99% Vorhersage-Intervall an der Stelle $X = x_0$.
  (b) 95% Vorhersage-Intervall an der Stelle $X = x_0$.
  (c) 99% Konfidenz-Intervall an der Stelle $X = x_0$.
  (d) 99.5% Konfidenz-Intervall an der Stelle $X = x_0$.

(10) Was macht der `R`-Befehl `apply(air[, 1:4], 2, table)`, wobei *air* ein Datensatz aus dem Datenanalyse-Projekt ist?
  (a) Es wird eine 4-dimensionale Tabelle der ersten vier Variablen aus *air* erzeugt.
  (b) Es wird für die zweite Zeile der ersten vier Variablen in *air* eine Tabelle erzeugt.
  (c) Es werden für die ersten vier Variablen aus *air* alle 2-dimensionalen Tabellen erzeugt.
  (d) Es wird für jede der ersten vier Variablen aus *air* eine Tabelle erzeugt.

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 $\diagdown\alpha$	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

$n\diagdown\alpha$	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

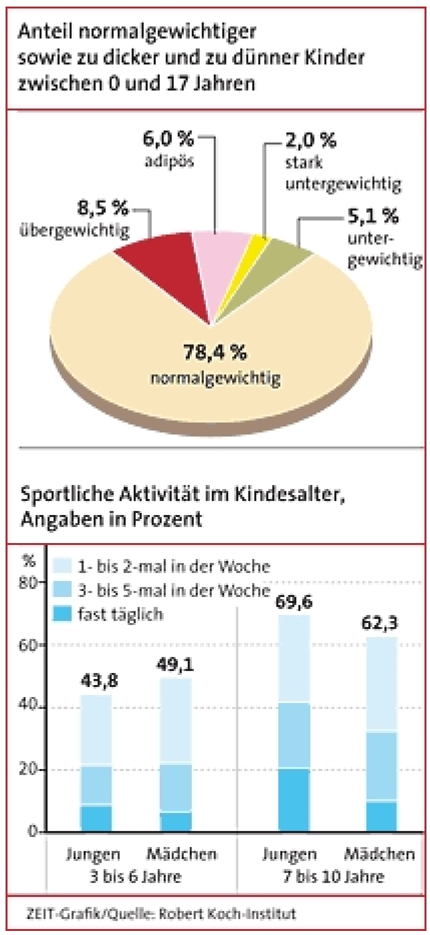
$\alpha = 0.05$	m					
n	10	11	12	13	14	15
10	72	78	85	92	98	105
11	78	86	93	100	107	114
12	85	93	101	108	116	124
13	92	100	108	117	125	133
14	98	107	116	125	134	143
15	105	114	124	133	143	152

$\alpha = 0.01$	m					
n	10	11	12	13	14	15
10	67	73	80	86	92	98
11	73	80	87	94	101	107
12	80	87	94	102	109	116
13	86	94	102	110	118	126
14	92	101	109	118	126	135
15	98	107	116	126	135	144

Teil 2. Bearbeiten Sie 5 der 8 Aufgaben!

1. Adipositas bei Kindern

Die Zunahme von Übergewicht und Fettsucht (Adipositas) bei Kindern wird seit einigen Jahren in der Presse diskutiert. Untere Abbildung zeigt zwei Grafiken zu diesem Thema. Die Ergebnisse beruhen auf Daten von 7186 Jungen und 6919 Mädchen im Alter von 3-17 Jahren.



(a) Was wird in den Grafiken dargestellt? Welche Informationen sollten zusätzlich bereitgestellt werden?

(b) Beurteilen Sie die Graphiken bezüglich Suggestivität, Übersichtlichkeit und Ästhetik.

(c) Welche Variablen besitzt der zugrundeliegende Datensatz und welchen Datentyp haben sie?

(d) Welche Ihnen bekannten Grafiken würden Sie anstelle der beiden einzelnen Visualisierungen empfehlen? Wie könnte man interaktives *Highlighting* einsetzen?

(e) Die Information, die beide Grafiken verbindet, ist der Zusammenhang von Gewicht und sportlicher Aktivität. Dieser soll unter Berücksichtigung der anderen Variablen betrachtet werden. Welche Grafik kann zur Visualisierung eingesetzt werden? Beschreiben Sie diese genau.

2. ML-Schätzer und Regression

Sei $(x_1, y_1), \ldots, (x_n, y_n)$ eine Messreihe. Die Messwerte x_i seien ohne zufällige Fehler beobachtet. Die zufallsabhängigen Messwerte y_i seien Realisationen von unabhängigen, normalverteilten Zufallsvariablen $Y_1, \ldots, Y_n$, mit

$$E(Y_i) = a + bx_i \quad \text{und} \quad Var(Y_i) = \sigma^2,$$

wobei a, b und $\sigma^2 > 0$ unbekannt und zu schätzen sind.

Bearbeiten Sie folgende Teilaufgaben.

(a) Bestimmen Sie die zur Messreihe $(x_1, y_1), \ldots, (x_n, y_n)$ gehörende Likelihood-Funktion. Diese ist als Funktion welcher Größen aufzufassen?

(b) Wie sehen die Schätzer für a und b für einen fixen Wert von σ^2 aus?

(c) Wie verändert sich das Resultat, wenn σ^2 auch geschätzt werden muss?

(d) Was versteht man unter der Maximum-Likelihood-Methode ganz allgemein und wieso ist diese Methode so etabliert?

(e) Die Maximum-Likelihood-Schätzer für a und b stimmen mit den Schätzern überein, die nach welcher anderen Methode gewonnen werden können? Beschreiben Sie die Idee, die hinter dieser anderen Methode steckt. Bestimmen Sie nach dieser Methode, durch Untersuchung der Ableitungen, explizit die Schätzer für a und b.

3. Clairol

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

(b) In einer in diesem Jahr von der Firma Clairol gesponserten Studie, sind 4000 Frauen zwischen 25 und 65 befragt worden. Unter anderem heisst es, dass 56% der Frauen sich Sorgen machen, dass ihr Aussehen sich verschlechtert, wenn sie älter werden. Testen Sie die Nullhypothese, dass die tatsächliche Rate unter Frauen weniger als 50% ist.

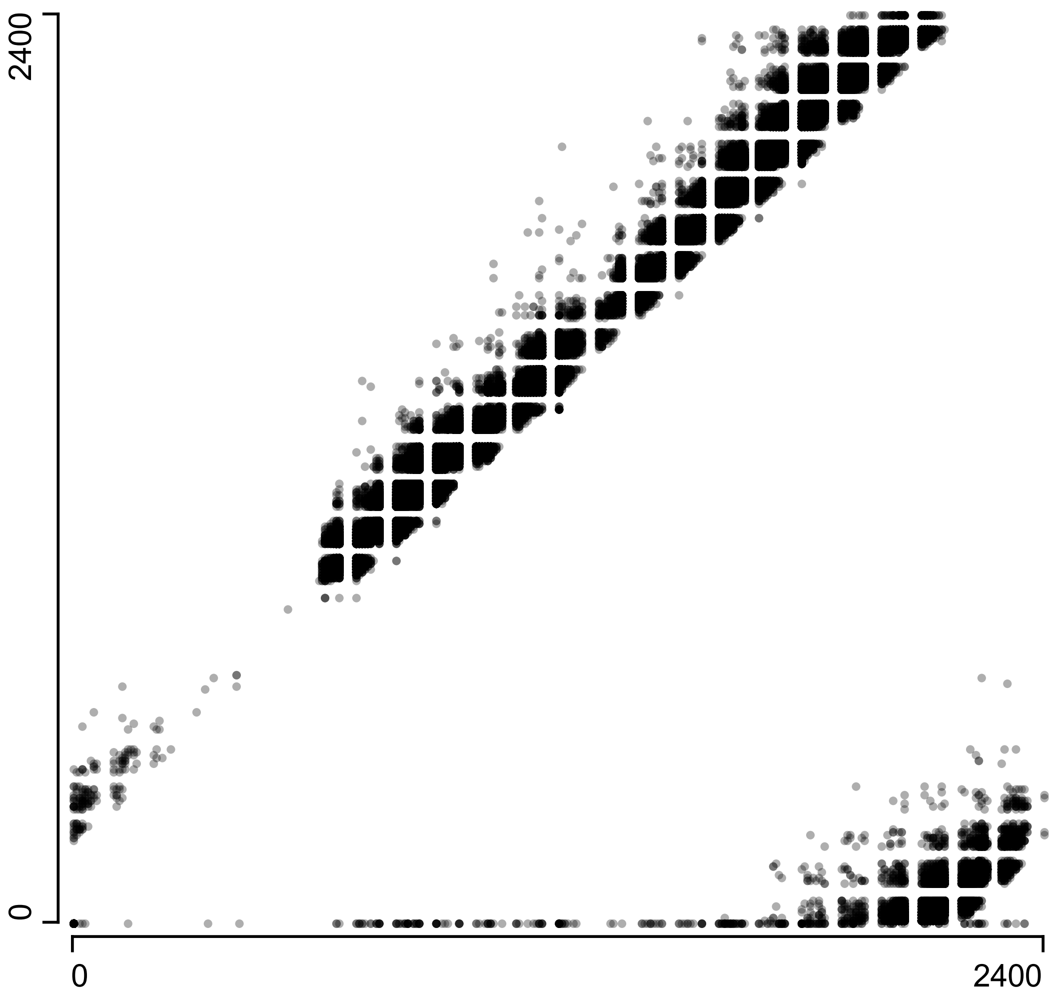
(c) Wie würde die Gütefunktion für Alternativhypothesen zwischen 50% und 90% aussehen? Wie berechnet man die Gütefunktion in diesem Fall?

(d) Die Frauen werden in vier Altersgruppen aufgeteilt: $25-35, 35-45, 45-55, 55-65$ und die Raten in diesen Gruppen verglichen. Wie würden Sie solche Daten grafisch darstellen?

(e) Auf die Studie in (b) bezogen wird weiter berichtet, dass Frauen im Durchschnitt sechs "bad hair days" (Frisurdebakeltage) pro Monat haben. Berechnen Sie ein 95% Konfidenzintervall für diesen Parameter. Welche Annahmen haben Sie gemacht?

4. Flüge

(a) Das Streudiagramm in der unteren Abbildung zeigt die tatsächlichen Abflugs- und Ankunftszeiten aller Flüge zwischen JFK (New York) und LAX (Los Angeles) über zwanzig Jahre. Warum gibt es Blöcke von Daten? Warum gibt es Daten unten rechts in der Graphik? Was könnten die 0-Werte für Ankunftszeiten bedeuten? Was muss unternommen werden, um einen sinnvollen Plot zu bekommen?



Abflugszeiten (X) und Ankunftszeiten (Y) für Flüge von JFK nach LAX zwischen 1987 und 2008.

(b) Von Interesse in der Analyse der Daten waren hauptsächlich die Flugverspätungen. Wie könnte man die Verspätungen nach Fluglinie (auf dieser Strecke gab es sieben Fluglinien) graphisch vergleichen? 2008 waren zwei Fluglinien für die meisten Flüge verantwortlich. Welchen Test würden Sie empfehlen, um zu überprüfen, ob die durchschnittlichen Verspätungen der zwei Linien sich signifikant voneinander unterschieden? Welche Annahmen müssen hier gemacht werden und wie realistisch sind diese? Welche anderen Faktoren könnten Einfluss auf die Verspätungen haben?

5. Produktionsprozess

In einem Produktionsprozess ist ein Teil mit Wahrscheinlichkeit $0 < p < 1$ defekt. Für eine Stichprobe $x_1, ..., x_n$ der Größe n , nehmen Sie an, dass $X_i \sim B(1, p)$. Die Wahrscheinlichkeit, dass die ersten zwei Teile defekt sind, soll geschätzt werden.

(a) Zeigen Sie, dass der Schätzer $\bar{X}^2$ asymptotisch erwartungstreu für p^2 ist.

(b) Leiten Sie einen erwartungstreuen Schätzer für p^2 für obigen Sachverhalt her.

(c) In obigem Prozess wurden bei 700 Teilen 21 defekte entdeckt. Bestimmen Sie mittels des plug-in Schätzers ein 95%-Konfidenzintervall.

(d) Führen Sie einen geeigneten Test zum Niveau 0.99 durch, um zu überprüfen, ob man in dem vorliegenden Prozess einen durchschnittlichen Ausschuß von 1% hat.

(e) Wenn Sie zu einem Niveau von 0.99 150 Hypothesentests durchführen, wie viele signifikante Ergebnisse würden Sie erwarten, wenn immer die Nullhypothese richtig wäre? Diskutieren Sie dies kritisch.

6. Wahlen

Folgende Tabelle gibt die Abgeordneten zur Bundespräsidentenwahl 2009 wieder:

Partei	Anzahl
CDU/CSU	497
SPD	418
FDP	107
Die Grünen	95
Die Linke	90
Rest	17

(a) *Die Linke* hat 91 Stimmen erhalten; nehmen sie an, dass 90 davon aus der eigenen Partei sind. Bestimmen Sie mittels der Maximum-Likelihood Methode von welcher Partei *Die Linke* die zusätzliche Stimme am ehesten bekommen hat. Gehen Sie hierbei von folgenden Wahrscheinlichkeiten aus:
$\mathbb{P}(\text{Die Linke wählen}|\text{CDU/CSU Abgeordneter}) = 0.0005$,
$\mathbb{P}(\text{Die Linke wählen}|\text{SPD Abgeordneter}) = 0.0025$,
$\mathbb{P}(\text{Die Linke wählen}|\text{FDP Abgeordneter}) = 0.0011$,
$\mathbb{P}(\text{Die Linke wählen}|\text{Die Grünen Abgeordneter}) = 0.0045$ und
$\mathbb{P}(\text{Die Linke wählen}|\text{Rest Abgeordneter}) = 0.0$.

(b) Nach der Wahl zum Bundestag 2005 wurde folgendes lineare Modell berechnet. Beschreiben Sie den Output. (Hinweis: `CDUCSU` gibt den Anteil der Stimmen der CDU/CSU, `unemp04` die Arbeitslosigkeit 2004, `manufact` den Anteil der Arbeitnehmer im produzierenden Gewerbe und `popdensity` die Bevölkerungsdichte in einer Region wieder.)

```
Call:
lm(formula = CDUCSU ~ unemp04 + manufact + popdensity)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12745 -0.04797 -0.01504  0.03361  0.19105

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.960e-01  2.220e-02  17.838  < 2e-16 ***
unemp04          -9.007e-03  7.935e-04 -11.351  < 2e-16 ***
manufact          2.239e-03  4.365e-04   5.129 5.28e-07 ***
popdensity       -1.358e-05  3.040e-06  -4.467 1.13e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06402 on 295 degrees of freedom
Multiple R-squared: 0.5412,Adjusted R-squared: 0.5365
F-statistic:   116 on 3 and 295 DF,  p-value: < 2.2e-16
```

(c) Berechnen Sie den Anteil der CDU/CSU laut des Modells in einer Region mit 15% Arbeitslosigkeit, 20% der Arbeitnehmer im produzierenden Gewerbe und einer Bevölkerungsdichte von 800. Wie kann man die Güte eines linearen Modells beurteilen?

7. Stochastik - Klausur

(a) An der Klausur der Vorlesung Stochastik I im Wintersemester 05/06 nahmen 34 Studentinnen und 36 Studenten teil. Insgesamt haben 40 Studenten die Klausur bestanden, wovon 16 weiblich waren.

(i) Man prüfe anhand eines χ^2-Unabhängigkeitstests zum Signifikanzniveau von 5%, ob es Assoziationen zwischen Geschlecht und Bestehen einer Klausur gibt.

(ii) Geben Sie den p-Wert approximativ an und entscheiden Sie damit den Test zu einem Signifikanzniveau von 0.1.

(b) Die 16 Studentinnen, welche die Klausur bestanden haben, erzielten folgende Punktwerte:

225 205 180 166 150 140 133 130 124 120 116 109 107 106 106 104

Bestimmen Sie mittels eines z-Tests zu einem Test-Niveau von $\alpha = 0.05$, ob die Studenten aus einer Population mit Erwartungswert 150 stammen, wenn eine Standardabweichung von 30 vorausgesetzt wird.

(c) Betrachten Sie folgenden `R`-Code:

```
> # Punktwerte der Studentinnen, welche die Klausur bestanden haben
> x <- c(225, 205, 180, 166, 150, 140, 133, 130, 124, 120, 116, 109,
+ 107, 106, 106, 104)
>
> # Punktwerte der Studenten, welche die Klausur bestanden haben
> y <- c(245, 220, 218, 200, 180, 165, 162, 154, 145, 144, 144, 142,
+ 140, 132, 130, 128, 128, 122, 120, 111, 109, 107, 103, 101)
>
> # Kennzahlen der beiden Variablen
> data.frame(Length = c(length(x), length(y)), Mean = c(NA, mean(y)),
+ Sd = c(sd(x), sd(y)))
  Length     Mean       Sd
1     16       NA 37.27594
2     24 147.9167 39.30252
```

(i) Nutzen Sie den Output des `R`-Codes zur Durchführung eines zwei Stichproben t-Tests mit gleichen, aber unbekannten Varianzen. Testen Sie hierbei zu einem Signifkanzniveau von $\alpha = 0.01$ auf Gleichheit der Erwartungswerte von Studentinnen und Studenten.

(ii) Geben Sie, aufbauend auf obigem `R`-Code, den `R`-Befehl an, welcher diesen Test realisiert.

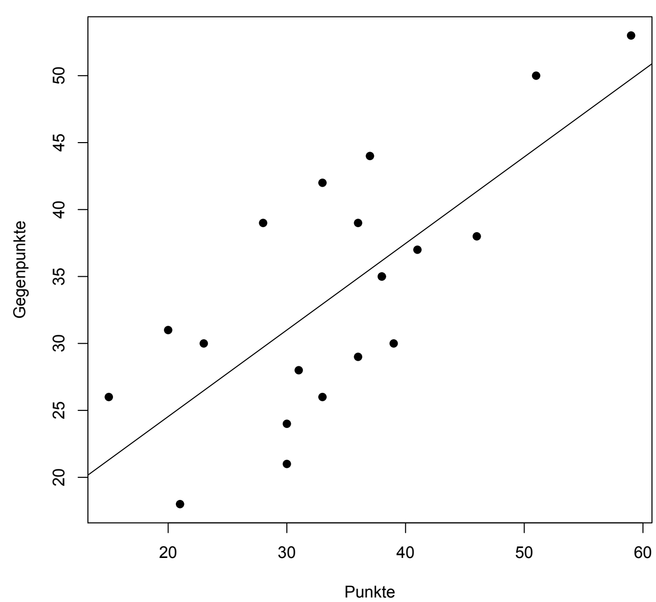
8. Wimbledon

Thomas Haas hat bei Wimbledon 20 Sätze Tennis gespielt. Zu jedem Satz sind Informationen zur Satzdauer in Minuten, Ausgang des Satzes, Anzahl Asse, Anzahl Doppelfehler, relative Häufigkeit gewonnener Punkte beim ersten bzw. zweiten Aufschlag, und Anzahl gewonnener bzw. verlorener Punkte gegeben.

Erläutern Sie den folgenden `R` Code und interpretieren Sie diesen. Gehen sie dabei auch auf die Annahmen, den Sinn und die Verwendbarkeit der enthaltenen Tests ein.

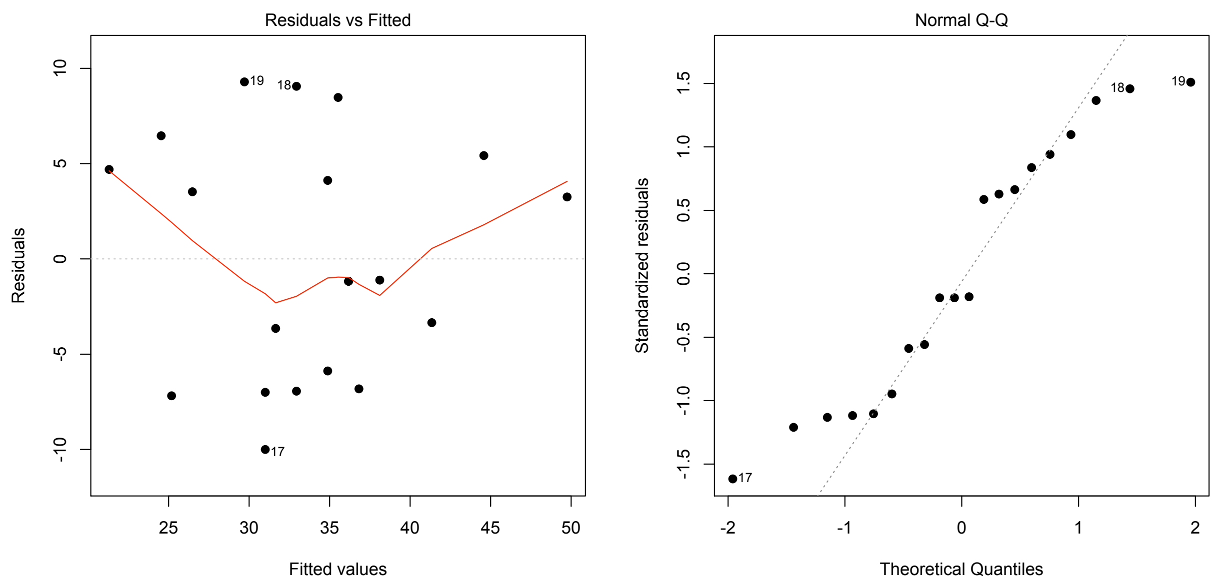
```
> th <- read.table(".../TommyHaas.txt",header=T,sep="\t",quote="")
> attach(th)
> names(th)
[1] "Gegner"       "Satzdauer"    "Gewonnen"     "Asse"
[5] "Doppelfehler" "serve1"       "serve2"       "Punkte"
[9] "Gegenpunkte"
>
> t.test(serve1, serve2, paired = T)

Paired t-test

data:  serve1 and serve2
t = 7.4148, df = 19, p-value = 5.082e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 20.95752 37.44248
sample estimates:
mean of the differences
                   29.2

>
> t.test(serve1 ~ Gewonnen)

Welch Two Sample t-test

data:  serve1 by Gewonnen
t = -1.8842, df = 11.659, p-value = 0.0847
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.031610   1.559082
sample estimates:
mean in group FALSE  mean in group TRUE
           72.57143            82.30769

>
> wilcox.test(Punkte,Gegenpunkte,conf.level=0.99,correct=F)

Wilcoxon rank sum test
```

```
data:  Punkte and Gegenpunkte
W = 207, p-value = 0.8496
alternative hypothesis: true location shift is not equal to 0

Warnmeldung:
In wilcox.test.default(Punkte,Gegenpunkte,conf.level=0.99,correct=F):
  kann bei Bindungen keinen exakten p-Wert Berechnen
>
> cor(Punkte, Gegenpunkte)
[1] 0.7389
> plot(Punkte, Gegenpunkte, pch = 19)
>
> mod<-lm(Gegenpunkte ~ Punkte)
> abline(mod)
```




```
> summary(mod)

Call:
lm(formula = Gegenpunkte ~ Punkte)

Residuals:
    Min      1Q  Median      3Q     Max
-10.003  -6.116  -1.144   4.876   9.290

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.6093     4.9679   2.337 0.031205 *
Punkte        0.6464     0.1389   4.652 0.000198 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.376 on 18 degrees of freedom
Multiple R-squared: 0.546,Adjusted R-squared: 0.5207
F-statistic: 21.65 on 1 and 18 DF,  p-value: 0.0001981

>
> par(mfrow = c(1,2))
> plot(mod, which = 1:2, pch = 19)
```




```
> mu <- mean(Satzdauer)
> sigma <- sd(Satzdauer)
>
> (quant<-qnorm(seq(0,1,0.25), mu, sigma))
[1]      -Inf 35.49719 44.50000 53.50281      Inf
>
> (klassen<-table(cut(Satzdauer, breaks = quant)))

(-Inf,35.5] (35.5,44.5] (44.5,53.5] (53.5, Inf]
          7           3           6           4
>
> chisq.test(klassen, correct = F)

Chi-squared test for given probabilities

data:  klassen
X-squared = 2, df = 3, p-value = 0.5724
```