SS 2012 *Name:*

Matrikel-Nr.:

KLAUSUR
STATISTIK I – (STOCHASTIK II)

20. JULI 2012
HILFSMITTEL: A4 BLATT MIT NOTIZEN, TR

Die Prüfung besteht aus zwei Teilen, die insgesamt 100 Punkte ergeben:
Teil 1: Multiple Choice (jede Aufgabe 2 Punkte)
Teil 2: 4 von 6 Aufgaben mit jeweils 20 Punkten

Bitte kreuzen Sie an, welche Aufgaben Sie abgegeben haben!
Es werden nur die besten 4 Aufgaben gezählt!

Aufgabe:	1	2	3	4	5	6	MC	Gesamt
Bearbeitet:								
Punkte:								

Schreiben Sie bitte auf jedes Blatt oben rechts Ihren Namen!

Viel Erfolg!

MC-Antworten:

1	2	3	4	5	6	7	8	9	10

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 Rückseite der Blätter können Sie für Berechnungen sowie zu Anmerkungen und Erläuterungen Ihrer Lösung verwenden.

(1) Welche Aussage über Graphiken ist **nicht** richtig?
   (a) Ein paralleler Koordinatenplot ist gut geeignet, um kategorielle Variablen und deren Zusammenhänge darzustellen.
   (b) Mit Hilfe eines Streudiagramms kann man zwei stetige Variablen und deren Zusammenhänge darstellen.
   (c) Ein Histogramm kann als Dichteschätzer angesehen werden.
   (d) Mit einem Mosaicplot lassen sich mehrere kategorielle Variablen und Assoziationen dieser Variablen darstellen.

(2) Welche Aussage über Tests ist richtig?
   (a) Nicht-parametrische Tests sind parametrischen Tests fast immer vorzuziehen, da die Anforderungen schwächer sind.
   (b) Parametrische Tests sind nicht-parametrischen Tests fast immer vorzuziehen, da die Anforderungen in der Praxis fast immer erfüllt sind.
   (c) Wird sowohl ein parametrischer als auch ein nicht parametrischer Test durchgeführt, so kann es passieren, dass nur der nicht parametrische Test signifikant ist.
   (d) Der Wilcoxon Rangsummentest und der Mann-Whitney U Test können zu unterschiedlichen Ergebnissen führen.

(3) Welche Aussage über Boxplots stimmt **nicht**?
   (a) Ein Boxplot basiert auf den Rängen der Daten.
   (b) In einem Boxplot lässt sich der Erwartungswert der Daten ablesen.
   (c) Ein Boxplot gibt Aufschluss über die Größe des Interquartilsabstandes.
   (d) Boxplots können zum Vergleich mehrerer stetiger Variablen verwendet werden.

(4) Es liegt eine Stichprobe $x_1, x_2, ..., x_n$ vor. Ein Kerndichteschätzer ist gegeben durch $\hat{f}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}{nh}$ Welche Aussage ist falsch?
   (a) $\int \sum_{i=1}^{n} K(\frac{x-x_i}{h}) dx = nh$
   (b) Die Bandbreite h sollte immer sehr klein gewählt werden, um eine möglichst gute Schätzung der Dichte zu erhalten.
   (c) Jede mögliche Kernfunktion ist eine Dichte.
   (d) Die Wahl der Kernfunktion hat, im Vergleich zur Wahl der Bandbreite h, einen geringen Einfluss auf die Dichteschätzung.

(5) Um den Problemen von Overplotting in einem Punktplot (stripchart() in R) entgegenzuwirken, sollte welche Methode **nicht** verwendet werden?
   (a) Jittering
   (b) α - Blending
   (c) Color-Brush, so dass jeder Punkt eine eigene Farbe bekommt.
   (d) Binning

(6) Welche Aussage über Glättungen ist falsch?
   (a) Eine loess-Glättung setzt die Unabhängigkeit der Beobachtungen voraus.
   (b) Eine loess-Glättung entspricht einem linearen Regressionsmodell, wenn man die Spannbreite auf 100% setzt.
   (c) Ein Kerndichteschätzer ist eine Glättung.
   (d) Eine gute Glättung ist robust gegenüber Ausreißern.

(7) Welche Aussage über Schätzer ist richtig?
   (a) ML-Schätzer sind konsistent erwartungstreu.
   (b) ML-Schätzer maximieren die Wahrscheinlichkeit, die vorliegenden Daten zu beobachten.
   (c) Durch Maximierung des log-Likelihoods wird zwar nur eine approximative Lösung gefunden, doch diese ist viel leichter zu berechnen.
   (d) Mittlerer quadratischer Fehler und Varianz sind genau dann gleich, wenn der Schätzer erwartungstreu ist.

(8) Welche Aussage über Korrelation ist richtig?
   (a) Besteht zwischen zwei Variablen kein kausaler Zusammenhang, so ist der Stichprobenkorrelationskoeffizient gleich 0.
   (b) Ist der Stichprobenkorrelationskoeffizient für zwei Variablen gleich 0, so sind diese statistisch unabhängig.
   (c) Ist für zwei Messreihen der Korrelationskoeffizient größer 0.5, so kann man von einem kausalen Zusammenhang ausgehen.
   (d) Ist die Standardabweichung einer der beiden Variablen 0, kann kein Korrelationskoeffizient berechnet werden

(9) Für einen optimalen Test nach Neyman-Pearson gilt:
   (a) Die Wahrscheinlichkeitsmasse des Ablehnungsbereiches ist bei einem zweiseitigen Test an beiden Enden gleich $\alpha/2$.
   (b) Für die Gütefunktion des optimalen Tests gilt $g(\theta) \geq g^*(\theta)$ für alle $\theta \in H_A$ und alle möglichen Tests mit Gütefunktion $g^*(\theta)$.
   (c) Der Ablehnungsbereich R ist so gewählt, dass $f(x) \leq k \ \forall x \in R$ bei mminimalem Niveau k ist.
   (d) Der p-Wert eines zweiseitigen Tests mit Teststatistik s ist gleich $2 \cdot min(\int_{-\infty}^{s} f(x)dx, \int_{s}^{\infty} f(x)dx)$.

(10) In der Süddeutschen Zeitung vom 24. Juni 2012 wird über Stammtische berichtet: *Keine Statistik offenbart, wie viele Stammtische es in deutschen Wirtshäusern gibt. Nur so viel ist bekannt: 28 Prozent der ber 18-jährigen Deutschen treffen sich regelmäßig zu einer Runde, die als Stammtisch bezeichnet werden kann, so das Ergebnis einer von der Tageszeitung* Die Welt *allerdings schon 2005 in Auftrag gegebenen Umfrage. Fast exakt so viele übrigens, wie eine ähnliche Erhebung zehn Jahre zuvor ergab. 50 Jahre früher sollen es nur 17 Prozent gewesen sein.* Daraus können wir schliessen, dass
   (a) höchstens 28 Befragte haben im Jahr 2005 gesagt, dass Sie sich zu einer Runde regelmäßig treffen.
   (b) ein χ^2-Test zum Vergleich der Werte 28% und 17% ist signifikant zum Niveau $\alpha = 0.05$, falls beide Umfragen der Größe 120 waren.
   (c) es 2005 65% mehr Stammtische gab als zehn Jahre zuvor.
   (d) es 2005 11% mehr Stammtische gab als zehn Jahre zuvor.

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

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

df $\backslash\alpha$	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

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

$n\backslash k$	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

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

Teil 2.

Bearbeiten Sie 4 der 6 Aufgaben!

1. Aktienrenditen

Im Folgenden sind die logarithmischen Aktienrenditen R_i zweier Unternehmen A und B in einem Zeitraum von 10 Handelstagen aufgeführt. Hierbei berechnet sich die logarithmische Aktienrendite gemäß der Formel $R_i := log(\frac{S_i}{S_{i-1}})$, wobei S_i den Aktienpreis am Tag i bezeichnet:

	1	2	3	4	5	6	7	8	9	10
A	0.030	0.052	0.051	0.073	0.050	0.042	-0.008	0.099	0.053	0.113
B	0.076	0.035	-0.036	0.030	0.059	0.055	-0.050	-0.022	-0.066	0.047

(a) Wie würden Sie die gegebenen Daten graphisch darstellen? Erläutern und skizzieren Sie die entstehende Graphik kurz.

(b) Gehen Sie von normalverteilten logarithmischen Renditen aus. Testen Sie mit einem geeigneten parametrischen Test, ob der wahre unbekannte Erwartungswert μ_A für Unternehmen A signifikant von 0 verschieden ist. Gehen Sie dabei auch auf evtl. getroffene Annahmen ein und erläutern Sie kurz, ob diese hier gegeben sind. (Die Stichproben-Standardabweichung beträgt $\hat{\sigma}_A = 0.0341$)

(c) Übungsleiter C. schaut sich die Renditen an und meint dazu: „Es ist egal welche Aktien man kauft, die sind beide gleich gut!“ Testen Sie dies mit einem geeigneten nichtparametrischen Test. Geben Sie alle getroffenen Annahmen an!

(d) Erläutern Sie folgenden R-Code: Welcher Test wurde hier durchgeführt? Welche Annahmen wurden getroffen und wie ist der Output zu deuten?

```
t.test(A,B,paired=TRUE)

Paired t-test

data:  A and B
t = 2.3886, df = 9, p-value = 0.04065
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
0.002259938 0.083140062
sample estimates:
mean of the differences  0.0427
```

(e) Warum wird hier mit logarithmischen Renditen und nicht mit den Aktienkursen oder den Renditen gearbeitet?

2. Atemwegserkrankung von Schulkindern

Zur Untersuchung des Einflusses der industriell bedingten Umweltbelastung auf die Gesundheit wurde von der Universität Innsbruck 1989 eine Studie über Lungenfunktion und Atemwegserkrankungen durchgeführt. Die Querschnittstudie enthält diverse Informationen über insgesamt 1549 Pflichtschulkinder aus der größeren Umgebung von Brixlegg in Tirol. Neben der Umweltbelastung wurden im Rahmen der Studie auch andere Einflussfaktoren, wie z.B. das elterliche Rauchen berücksichtigt.
Quelle: http://www.wi.hs-wismar.de/ cleve/vorl/dmdaten/daten/atemwege.htm

Im Folgenden ist die Tabelle der Umweltbelastung am Wohnort V_Z und der bivariaten Variable V_K, ob eine Krankheit der Lungen oder Atemwege vorhanden ist, abgebildet.

	V_K=nein	V_K=ja	Total
stark belastet	106	80	186
eher wenig belastet	364	256	620
erhöhte Ozonbelastung durch Hochlage	470	273	743
Total	940	609	1549

(a) Testen Sie mit einem χ^2-Unabhängigkeitstest, ob die beiden Variablen unabhängig voneinander sind. Geben sie hierzu neben der exakt formulierten Null- und Alternativhypothese auch die Teststatistik und deren Verteilung an. Kann die Nullhypothese zu einem Signifikanzniveau von $\alpha = 0.05$ verworfen werden?

(b) Skizzieren Sie die Verteilung der Teststatistik. Zeichnen Sie anschließend die Prüfgröße und den p-Wert ein.

(c) Wie groß müsste die Stichprobe sein, damit die Nullhypothese zu einem Signifikanzniveau von $\alpha = 0.01$ verworfen werden könnte? Macht es demnach Sinn, mehr Schulkinder zu befragen?

(d) Wie kann die Unabhängigkeitsannahme beider Variablen grafisch überprüft werden? Skizzieren Sie ihren Vorschlag.

(e) Betrachtet wird nun nicht mehr die erste Kategorie $V_Z =$ *"stark belastet"*.

	V_K=nein	V_K=ja	Total
eher wenig belastet	364	256	620
erhöhte Ozonbelastung durch Hochlage	470	273	743
Total	834	529	1363

Testen Sie nun mit einem Binomialtest, ob sich die Krankheitsquoten voneinander unterscheiden. Geben Sie hierzu die Null- und Alternativhypothese, die Teststatistik und deren Verteilung sowie den p-Wert an. Ist dieser Test äquivalent zu einem entsprechenden χ^2-Test?

Hinweis: Als Approximation kann eine Normalverteilung mit entsprechenden Parameter gewählt werden.

3. Schätzen

Die geometrische Verteilung beschreibt die Anzahl X von Bernoulli-Versuchen, die nötig sind, um einen Erfolg zu haben. Diese Wahrscheinlichkeitsverteilung ist gegeben durch

$$P(X = k) = p(1-p)^{k-1}, k \in \mathbb{N}, 0 < p \leq 1.$$

(a) Bestimmen sie den Maximum-Likelihood-Schätzer für den Parameter p, wenn sie die Daten $k_1, k_2, ..., k_n$ als Ausprägungen von X beobachtet haben.

(b) Sei nun $n = 1$ und nur der Wert k sei für X beobachtet worden. Der Maximum-Likelihood-Schätzer ist in diesem Fall $\widehat{p} = \frac{1}{k}$. Der Erwartungswert dieses Schätzers ist dann $E[\widehat{p}] = -\frac{p}{1-p} ln(p)$. Ist dieser Schätzer erwartungstreu?

Ein Rentner geht in Pattaya auf die Suche nach einer Partnerin. Die Anzahl der Versuche, die nötig sind, dass er eine Frau findet, die bereit ist sich mit ihm zu unterhalten (über griechische Philosophie oder ähnliches), sei pro Abend annähernd geometrisch verteilt. Er erzählt stolz: Am ersten Abend habe er vier Anläufe benötigt, am zweiten sechs und am dritten nur drei.

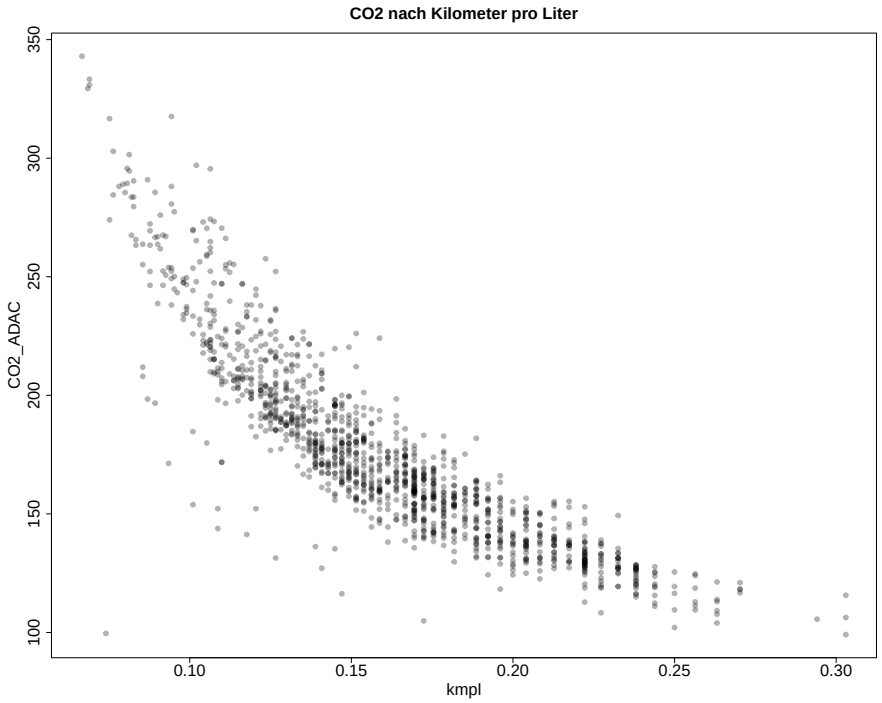
(c) Schätzen sie die Quote der Frauen, die sich mit ihm unterhalten, als überdurchschnittlich ($p > 0.75$) mittelmäßig ($0.75 \geq p > 0.25$) oder unterdurchschnittlich ($p \leq 0.25$) ein? Sind Annahmen, die sie für ihre Schätzung verwendet haben, eventuell verletzt?

(d) Um seine Quote zu verbessern, sagt er sich: "Übung macht den Meister" und spricht am vierten Abend zehn Frauen an. Drei von ihnen unterhalten sich tatsächlich mit ihm. Wie würden sie seine Erfolgswahrscheinlichkeit p ausgehend von diesem Abend schätzen?

(e) Bestimmen sie ein normal-approximatives 95%-Konfidenzintervall für diesen Schätzer. Wie ist dieses zu interpretieren?

(f) Der Rentner ist mit der Qualität der Schätzung nicht zufrieden und schlägt folgendes Verfahren vor: Gehen sie apriori von einer Betaverteilung mit Parameter $a = 1$ und Erwartungswert identisch $\tilde{p}$ aus, wobei $\tilde{p}$ der Wert des Schätzers aus Aufgabe (c) ist. Bestimmen Sie den Bayes-Schätzer mit den neuen Daten aus Aufgabe (d). Geben Sie sowohl einen Vorteil als auch einen Nachteil dieser Methode an!

4. EcoTest



Obige Grafik zeigt ein Streudiagramm des jeweils vom ADAC gemessenen CO_2-Ausstoßes gegen den inversen Verbrauch *KMPL* (Kilometer pro Liter Treibstoff) laut Hersteller.

(a) Die Grafik zeigt eine Gruppe von Ausreißern unterhalb des "Hauptfeldes" sowie einen krassen Ausreißer links unten. Erläutern Sie, wie Sie mit interaktiven Grafiken in Mondrian (oder alternativ mit R) herausfinden, um welche Fahrzeuge es sich handelt.

(b) Welche Eigenschaften könnten die Fahrzeuge eventuell haben? Wie können Sie dies in Mondrian überprüfen?

(c) Nach Entfernung der Ausreißer wird zur Modellierung das folgende Modell vorgeschlagen:

$$CO2_i = \beta_0 + \beta_1 KMPL_i + \beta_2 KMPL_i^2 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Ist das Modell sinnvoll? Kann es eingesetzt werden, um den CO_2-Ausstoß

- des neuen Mazda mit 20km/Liter
- des neuen UAB NEGIE EcoSprit mit fantatsischen 40km/Liter

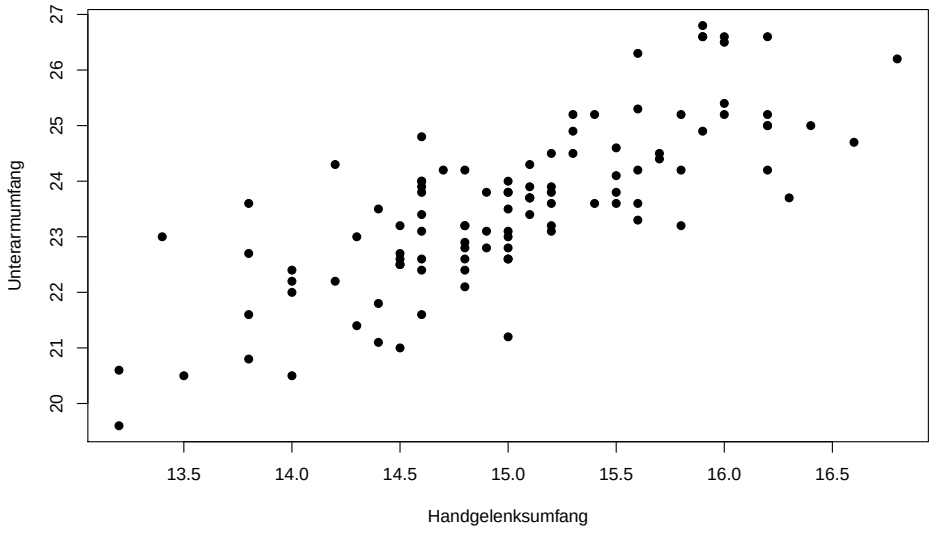
vorherzusagen? Begründen Sie Ihre Antwort!

(d) Wie sind die Parameter β_0 und β_1 zu interpretieren? Ändern Sie sich, wenn man *KMPL* zuvor zentriert bzw. standardisiert?

(e) Der R-Output eines linearen Modells (vgl. Aufgabe 5) enthält Tests für die Parameter $\beta_0, \beta_1, ..., \beta_k$. Erläutern Sie präzise den Test für β_1. Was können Sie folgern, wenn dieser signifikant bzw. nicht signifikant ist?

5. Körpermessungen

An 105 Frauen im Alter zwischen 18 und 25 Jahren wurden verschiedene Körpermessungen durchgeführt. Zwischen einigen der Variablen wird ein linearer Zusammenhang vermutet, so auch zwischen dem Unterarmumfang und dem Handgelenksumfang. Um dies weiter zu überprüfen wurde zunächst ein Streudiagramm erstellt.



(a) Schildern Sie kurz, wodurch die linienartige Struktur im Streudiagramm zu erklären ist und ob Sie die Annahme eines linearen Zusammenhangs für gerechtfertigt halten!

Weiter wurde die Korrelation in *R* berechnet.

```
> cor(Handgelenksumfang,Unterarmumfang)
[1] 0.7753674
```

Herr Meier, der mit der Auswertung der Daten beauftragt wurde, möchte anstatt der Korrelation lieber lineare Modelle verwenden und schlägt folgende vor:

```
> model1 <- lm(Unterarmumfang~Handgelenksumfang)
> model2 <- lm(Handgelenksumfang~Unterarmumfang)
```

Den `summary`-Output finden Sie auf dem nächsten Blatt.

(b) Wie hängen der Korrelationskoeffizient und die linearen Modelle zusammen? Worin unterschieden Sie sich?

(c) Berechnen Sie den Wert von R^2!

(d) Was unterscheidet *model1* und *model2*? Wie lautet jeweils die Modellgleichung?

(e) Ist die Verwendung von *model1* problematisch, weil der Signifikanztest des Intercepts einen sehr hohen p-Wert hat? Was wird hier getestet?

(f) Nehmen Sie Stellung zu folgender Aussage:

"*Die Hinzunahme weiterer erklärender Variablen (z.B. Körpergröße) ist immer dann sinnvoll, wenn sich R^2 dadurch erhöht.*"

(g) Welche Annahmen werden bei Verwendung linearer Modelle gemacht und wie würden Sie überprüfen, ob Sie erfüllt sind?

(h) Welchen Unterarmumfang erwarten Sie bei einem Mann mit einem Handgelenksumfang von 18 cm? Für wie vertrauenswürdig halten Sie den Wert?

```
> summary(model1)
Call: lm(formula = Unterarmumfang ~ Handgelenksumfang)

Residuals:
     Min       1Q   Median       3Q      Max
-2.33827 -0.58700 -0.08194  0.50793  1.97186

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.8484     1.8260   0.465    0.643
Handgelenksumfang   1.5127     0.1214  12.461   <2e-16 ***

Residual standard error: 0.9273 on 103 degrees of freedom
Multiple R-squared: -----,Adjusted R-squared: 0.5973
F-statistic: 155.3 on 1 and 103 DF,  p-value: < 2.2e-16

> summary(model2)
Call: lm(formula = Handgelenksumfang ~ Unterarmumfang)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39556 -0.27761  0.02623  0.32750  1.22623

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.6544     0.7533   7.506 2.26e-11 ***
Unterarmumfang   0.3974     0.0319  12.461  < 2e-16 ***

Residual standard error: 0.4753 on 103 degrees of freedom
Multiple R-squared: -----,Adjusted R-squared: 0.5973
F-statistic: 155.3 on 1 and 103 DF,  p-value: < 2.2e-16
```

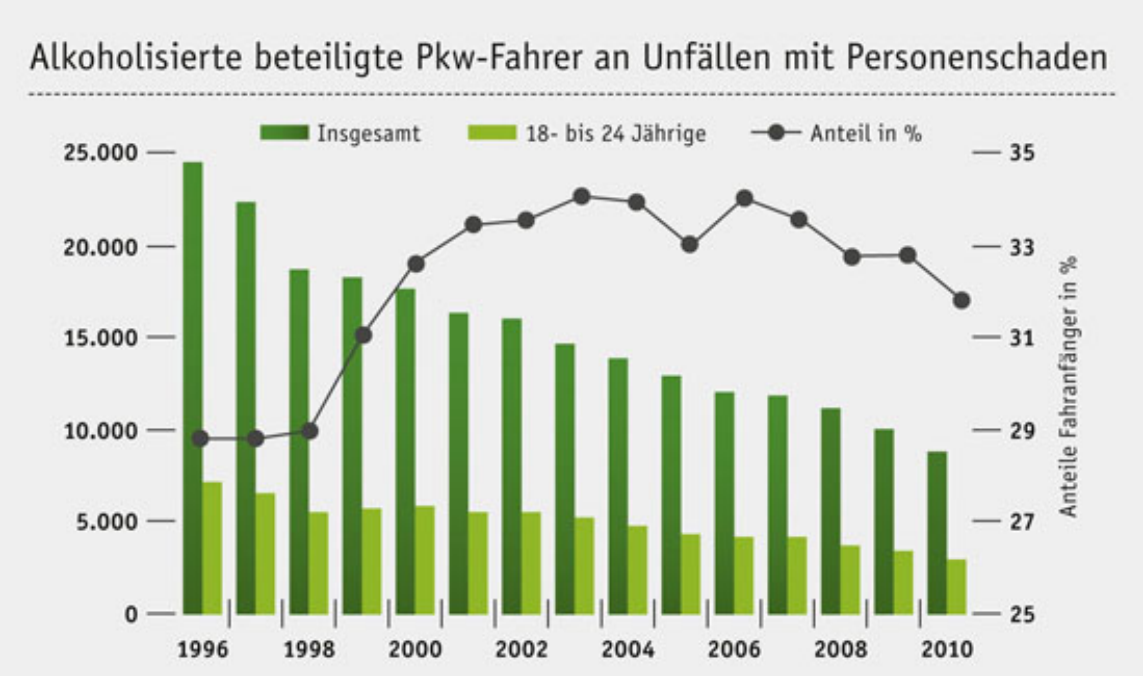
6. Alkoholunfälle

Auf der Webseite

`http://riskiernichts.de/riskier-nichts/unfaelle-in-zahlen/`

wird über Unfälle im Zusammenhang berichtet:

Die Anzahl der an Unfällen mit Personenschaden beteiligten alkoholisierten Autofahrer ist seit 1997 um 62 Prozent zurückgegangen. Eine nachweisbare Wirkung zeigt das absolute Alkoholverbot für Fahranfänger in der Probezeit und für alle unter 21 Jahre: Die Zahl der unter Alkoholeinfluss stehenden Autofahrer bei Unfällen mit Personenschaden ist von 2006 bis 2010 in der Altersgruppe von 18 bis 24 Jahren um 32 Prozent zurckgegangen.



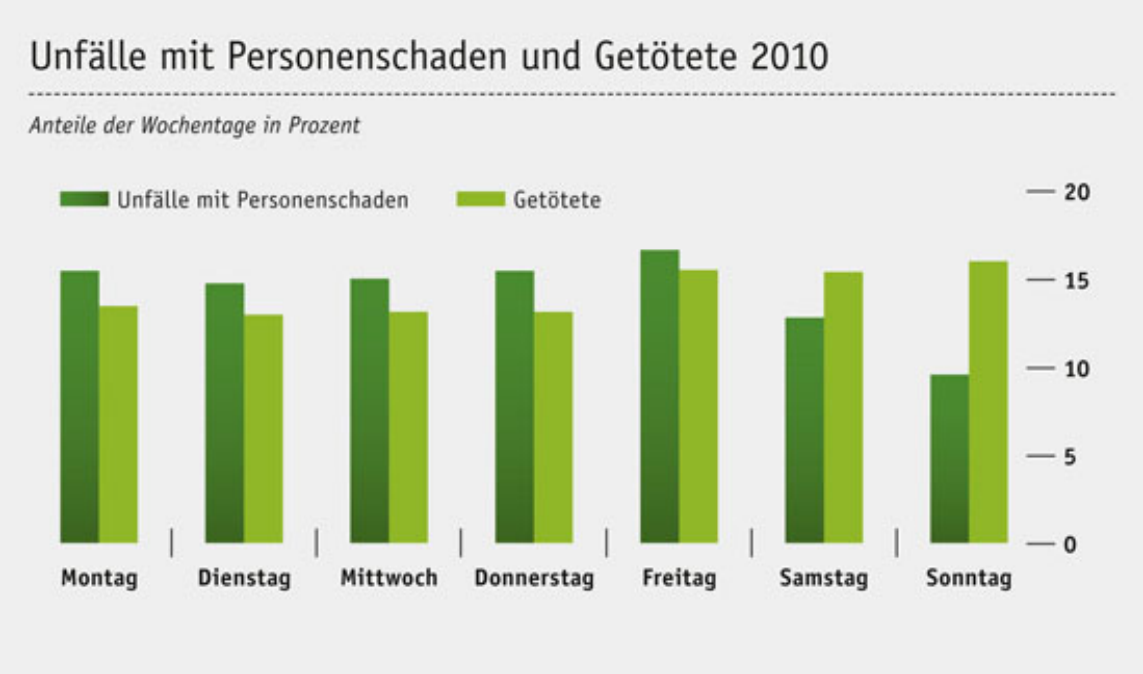
(a) Beschreiben Sie die Grafik: Was wird dargestellt und auf welche Weise?

(b) Diskutieren Sie die Darstellungsweise anhand positiver und negativer Gesichtspunkte!

(c) Ist die Hauptaussage (letzter Satz) des Textes geeignet dargestellt? Handelt es sich Ihrer Meinung nach um die wichtigste Aussage? Was hätten Sie selbst aus der Darstellung geschlossen?

(d) Mit welcher Ihnen bekannten Visualisierung hätten Sie die Daten dargestellt? Beschreiben Sie diese kurz!

Im Bericht ist auch die folgende Grafik zu den Unfällen nach Wochentagen gegeben:



(e) Wie könnten die Wochentage in einer grafischen Darstellung der Alkohol-Daten mit einbezogen werden? Beschreiben Sie die Darstellung kurz.