Institut für Mathematik
Prof. Dr. P. Ruckdeschel
Dr. A. Mändle

WiSe 2016/17
08. Februar 2017

Statistik 1: Einführung in die Angewandte Statistik

Klausur (120 Minuten)

Name, Vorname ..... **Matr.-Nr.** .....

Studiengang ..... **Fachsemester:** .....

Zur Beachtung:

a) Die Bearbeitungszeit beträgt 120 Minuten. Davon verstehen sich 5 Minuten als Einlese- und Vorbereitungszeit und 5 Minuten als Endsortierzeit.

b) Bitte tragen Sie als erstes Ihre persönlichen Angaben ein und legen Sie Ihren Personalausweis/Lichtbildausweis bereit.

c) Lesen Sie die folgenden Punkte gründlich durch und unterschreiben Sie sodann.

d) Bitte prüfen Sie nach, ob Ihr Klausurexemplar vollständig ist und aus 35 Blatt besteht.

e) Um Ihnen die Übersicht zu erleichtern, sind am Anfang der Klausur auf den Seiten 2 bis 6 alle Aufgaben einmal abgedruckt.

f) Zugelassene Hilfsmittel sind: ein nicht-programmierbarer, nicht-graphikfähiger Taschenrechner. Weitere Hilfsmittel sind nicht zugelassen. Die benötigten Quantile finden Sie am Ende der Angabe tabelliert.

g) Täuschungsversuche, insbesondere mitgebrachte Zusatzmaterialien, wie z.B. Smartphone, führen zum Nichtbestehen der Klausur.

h) Es gibt 150 Punkte; maximal gewertet werden 100; treffen Sie daher eine Auswahl.

Zur Kenntnis genommen: .....
Unterschrift

Bewertung:

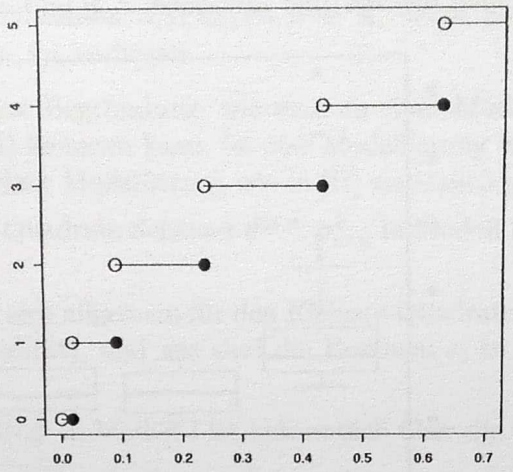
	Aufg. 1	Aufg. 2	Aufg. 3	Aufg. 4	Aufg. 5	Aufg. 6	$\sum$ 1–6	Bonus
Punkte	30	10	19	8	13	7	87	8
erreicht								
Korrektur								

	Aufg. 7	Aufg. 8	Aufg. 9	Aufg. 10	Aufg. 11	Aufg. 12	$\sum$ 7–12	Gesamt	Note
Punkte	4	10	7	6	17	19	63	150 (+8)	
erreicht									
Korrektur									

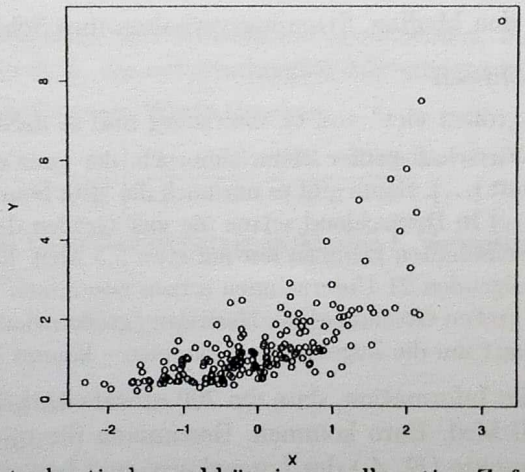
Aufgabe 1 Wahr oder falsch? **(30 Punkte)**

Geben Sie jeweils Gegenbeispiel oder einen Beweis/ Nachweis/ eine Begründung. Dabei dürfen Sie sich gegebenenfalls auf Sätze der Vorlesung berufen.

a) Folgende Graphik ist eine Verteilungsfunktion



b) Die im folgenden Scatterplot dargestellten Variablen X und Y sind positiv korreliert.



c) Seien X, Y beschränkte, stochastisch unabhängige, reellwertige Zufallsvariablen. Dann sind X^2 und Y^2 positiv miteinander korreliert.

d) Seien $1 \leq X, Y \leq 10$ reellwertige, stochastisch unabhängige Zufallsvariablen. Dann gilt $\mathrm{E}[X/Y] = \mathrm{E}[X]/\mathrm{E}[Y]$.

e) Sei P ein Wahrscheinlichkeitsmaß auf der σ-Algebra $\mathcal{A}$ und $A, B \in \mathcal{A}$. Dann gilt $P(A \cup B) \leq 2\max(P(A), P(B))$.

f) Seien P, A, B wie in e) und $P(B) > 0$. Dann gilt $P(A|B) \geq P(A \cap B)$.

g) Seien X, Y stochastisch unabhängige reellwertige Zufallsvariablen mit Dichten f^X und f^Y; sei $a \in (0,1)$ und $Z = aX + (1-a)Y$. Dann besitzt Z die Dichte $f^Z = af^X + (1-a)f^Y$.

h) Sei X eine reellwertige Zufallsvariable. Dann gilt $(\mathrm{E}|X|^4)^3 \geq (\mathrm{E}|X|^3)^4$.

i) Seien X, Y reellwertige Zufallsvariablen und $\mathrm{E}[X^2 + Y^2] < \infty$. Dann existiert $\mathrm{E}[XY]$.

j) Sei X eine reellwertige Zufallsvariable mit $\mathrm{E}\,X = 0$ und $\mathrm{Var}\,X = \sigma^2 < \infty$. Dann ist das 95%-Quantil kleiner als 5σ.

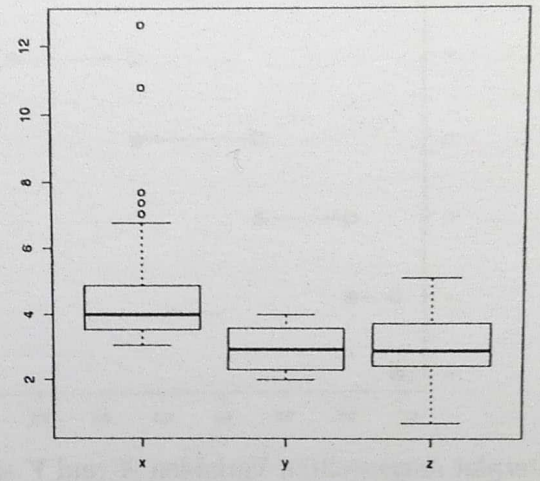
Aufgabe 2 Deskriptive Kennzahlen **(4+6 Punkte)**

a) Sie haben die (geordnete) Messreihe

$$0.7, 0.9, 1.4, 3.3, 5.4, 8.9, 11.1, 16.4, 27.7$$

Bestimmen Sie Median und Interquantilsabstand, sowie den MAD (den Median der absoluten Abweichungen vom Median).

b) Sie sehen drei Messreihen reellwertiger Merkmale x, y und z als Boxplots visualisiert.



Was können Sie über den Median, Streuungsverhalten und Schiefe der Variablen sagen?

Aufgabe 3 Konzentrationen **(8+3+3+5 Punkte)**

Im Artikel "Im Sog der großen vier" von G. Giersberg und J. Jahn, FAZ, 27.11.2011, heißt es

> "Die Branche der Wirtschaftsprüfer ist im Umbruch, der ganz eindeutig auf zunehmende Konzentration hinausläuft [...]. Heute gibt es nur noch die "Big Four": Deloitte, KPMG, PWC und Ernst & Young. [...] In Deutschland setzen die vier Großen der Branche 4,5 Mrd. Euro um. Die 25 größten Gesellschaften kommen nur auf etwa 5,5 Mrd. Euro. Das heißt, die diesen vier vom Umsatz her folgenden 21 Unternehmen setzen zusammen gerade einmal so viel um wie einer (sic!) der vier großen Gesellschaften. Nach den "großen vier" – und Deloitte muss mit 600 Millionen Umsatz hart um die Zugehörigkeit kämpfen – kommt lange nichts."

a) Sie bekommen noch die Information, dass die 200 umsatzstärksten Gesellschaften auf einen Gesamtumsatz von 10 Mrd. Euro kommen. Bestimmen Sie mit denen Ihnen vorliegenden Angaben die Knotenpunkte (F_i, Λ_i) der Lorenzkurve und berechnen Sie den Gini-Index. (Hierfür ist es nicht erforderlich, der Unsicherheit Rechnung zu tragen, dass Sie nicht alle Einzelumsätze kennen!)

b) Welchen Wert auf der vertikalen Achse hat die Lorenzkurve aus a) an der Stelle $92,75\%$?

c) Geben Sie Schranken nach oben und nach unten für den Umsatz des zweitgrößten Wirtschaftsprüfers an.

d) Vergleichen Sie Mittelwert und Median der Umsätze. Wie sind diese im konkreten Fall angeordnet und was kann man generell über deren Anordnung in diesem Kontext sagen?

Aufgabe 4 Definitionen **(3+5 Punkte)**

a) Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Was ist dabei $\mathcal{A}$ für ein Objekt und welche (definierenden) Eigenschaften hat es?

b) Sei $X = (X_1, \ldots, X_k)$ eine $\mathbb{R}^k$-wertige Zufallsvariable und $\mathrm{E}|X|^2 < \infty$. Außerdem sei $\sigma_i^2 := \mathrm{Var}\, X_i > 0$ für $i = 1, \ldots, k$. Definieren Sie die Korrelationsmatrix von X, indem Sie deren Einträge in Ausdrücken in Erwartungswerten darstellen.

Aufgabe 5 Stadtratswahl **(6+3+4 Punkte)**

Zur Stadtratswahl einer mittelgroßen Kommune treten 6 Listen an. In fünf dieser Listen treten jeweils 33 Kandidaten an, auf Liste 6 nur 10. Insgesamt also treten 175 Kandidaten an. Es sollen auf faire Art und Weise 3 Kandidaten für ein Interview gezogen werden.

(A) es werden blind drei Kandidaten zufällig ohne Zurücklegen aus den 175 gezogen

(B) Die Kandidaten seien über die Listen nummeriert. P.R. wird um eine zufällige Zahl z zwischen 1 und 175 gebeten, anschließend werden (modulo 175) die Kandidaten mit Schrittweite 85 ohne Zurücklegen gezogen, d.h. bei Überschreiten der 175 wird "vorne" wieder angefangen, sodass etwa der Wert 130 nach einem Schritt der Schrittweite 85 zu $130 + 85 - 175 = 40$ wird (mathematisch also $z + 85 \bmod 175$); gezogen werden also die Personen mit Nummer z, z+85 mod 175 und z+170 mod 175.

(C) zufälliges Ziehen einer Liste (mit gleicher Wahrscheinlichkeit für Listen 1–6) und dann zufälliges Ziehen einer Person aus der Liste

a) Kann man die Verfahren (A)–(C) als Zufallsauswahlen auffassen? Welches Ihnen bekannte Stichprobenziehungsverfahren aus der Vorlesung kommt jeweils einer Ziehung gemäß (A)–(C) am nächsten?

b) Kandidat K.R. steht auf Liste 3 (mit 33 Kandidaten). Wie hoch ist seine Ziehungswahrscheinlichkeit für das Interview in (A) und (C) bei der ersten Ziehung?

c) A.M. steht auf Liste 6. Wie wahrscheinlich ist er bei (C) unter den gezogenen 3 Kandidaten für das Interview?

Aufgabe 6 Grenzwertsätze **(2+2+3 Punkte)**

Seien $X_i \colon (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathbb{B})$, $i \in \mathbb{N}$ eine Folge u.i.v. Zufallsvariablen mit endlicher Varianz.

a) Definieren Sie stochastische Konvergenz von reellwertigen Zufallsvariablen.

b) Formulieren Sie den zentralen Grenzwertsatz für diese Situation.

c) Beweisen Sie das schwache Gesetz der großen Zahlen mit den Sätzen von Bienaymé und Chebyshev.

Aufgabe 7 Eine Verteilungsfunktion? **(4 Punkte)**

Wann ist eine Funktion $G \colon \mathbb{R} \to \mathbb{R}$ eine Verteilungsfunktion eines Wahrscheinlichkeitsmaßes auf $(\mathbb{R}, \mathbb{B})$, d.h. welche Eigenschaften charakterisieren eine Verteilungsfunktion?

Aufgabe 8 ML-Schätzung am Aktienmarkt **(4+4+2 Punkte)**

Sie sammeln die letzten 7 Jahresendstände des Kurses einer Aktie X_t, $t = 0, \ldots, 6$ und erhalten die Werte

$$100, 95, 94, 123, 128, 134, 179$$

Jemand berichtet Ihnen, dass ein gutes Modell für den Aktienmarkt darin besteht, für einen Startwert X_0 (hier 100) anzunehmen, dass $X_t = X_{t-1} R_t$, wobei die R_t unabhängig identisch lognormalverteilt sind, also $\log R_t \overset{\text{uiv}}{\sim} \mathcal{N}(\mu, \sigma^2)$ für unbekannte Parameter $\mu \in \mathbb{R}$ und $\sigma > 0$.

a) Wie lauten die ML-Schätzer für μ und σ^2 (als Schätzer, ohne Verwendung der Zahlenwerte)? Sind diese erwartungstreu und konsistent?

b) Werten Sie die ML-Schätzer auf den Daten aus.

c) In welchem Sinn ist der ML-Schätzer für μ optimal hinsichtlich Präzision?

Aufgabe 9 Konfidenzintervall in der Situation von Aufgabe 8 (2+3+2 Punkte)

a) Konstruieren Sie ein zweiseitiges 95% Konfidenzintervall für μ basierend auf einer t-Verteilung mit 5 Freiheitsgraden.

b) Wieviele Beobachtungen bräuchte man bei diesem σ, damit μ dadurch auf $\pm 0,5\%$ festgelegt wird? (Ignorieren Sie – für Klausurzwecke – die dadurch eintretende Variation der Freiheitsgrade.)

c) Wie wahrscheinlich ist es, dass der Kurs beim nächsten Mal den Wert 200 überschreitet, wenn Sie für μ und σ Ihre Schätzwerte aus Aufgabe 8 (bzw. die Fallbacks) verwenden?

Verwenden Sie ggf. als Fallback: $\hat{\mu} = 0.06$, $\hat{\sigma} = 0.10$.

Aufgabe 10 Testen eines Mittelwertes (1+1+2+2 Punkte)

In der Situation von Aufgabe 8 behauptet jemand: Der Mittelwert μ sei nicht größer als 1%.

a) Sie glauben die Behauptung nicht und wollen das testen; formulieren Sie dazu entsprechende Hypothesen.

b) Welchen Test zum Niveau $\alpha \in (0, 1)$ würden Sie zum Testen dieser Hypothesen verwenden?

c) Stellen Sie den Fehler 1. und 2. Art, sowie die Macht des Tests in Termen von Quantils- und Verteilungsfunktionen und ohne Rückgriff auf die numerischen Werte dar.

d) Können Sie die Nullhypothese zum Niveau 10% ablehnen?

Verwenden Sie ggf. als Fallback: $\hat{\mu} = 0.06$, $\hat{\sigma} = 0.10$.

Aufgabe 11 Kontingenztabelle (1+4+2+2+2+3+3 Punkte)

In einem internationalen Studiengang nehmen 450 Hörer an einer Klausur teil. Von diesen stammen 40 nicht aus der EU (und müssen Studiengebühren bezahlen). 340 Klausurteilnehmer bestehen die Klausur, von denen 33 nicht aus der EU stammen.

a) Erstellen Sie eine Kontingenztabelle mit den absoluten Häufigkeiten.

X \ Y	"EU"	"nicht EU"
"bestanden"	·	·
"nicht bestanden"	·	·

Seien dazu die Zellen mit (i, j), $i, j = 1, 2$ und i der Zeilen-, j der Spaltenindex bezeichnet.

b) Bestimmen Sie die relativen Häufigkeiten $f_{i,j}$ der einzelnen Ereignisse, sowie die Randhäufigkeiten $f_{i,\cdot}$ und $f_{\cdot,j}$ von X und Y.

c) Bestimmen Sie die bedingte relative Häufigkeit $\hat{P}_n(X = \text{"bestanden"} | Y = \text{"EU"})$.

d) Bestimmen Sie den Maximum-Likelihood-Schätzer für $P(X = \text{"bestanden"}, Y = \text{"nicht EU"})$.

e) Sei H_0 die Hypothese, dass die Wahrscheinlichkeit zu bestehen p unabhängig von der Herkunft ist. Wie lauten dann die Wahrscheinlichkeiten für die einzelnen Zellen der Tabelle (in Abhängigkeit von p und der Wahrscheinlichkeit q, aus der EU zu kommen)? Bezeichnen Sie dazu die Wahrscheinlichkeit unter H_0 für Zelle (i, j) mit $p_{i,j}$.

f) Bestimmen Sie unter H_0 den Maximum-Likelihood-Schätzer für p und für q.
Fallback: Wenn Sie nicht auf die Lösung von e) kommen, verwenden Sie folgende (falschen(!)) Zellwahrscheinlichkeiten $p_{1,1} = 0.1$, $p_{1,2} = p$, $p_{2,1} = q$, $p_{2,2} = 1 - p - q$.

g) Testen Sie mit einem χ^2-Test zum Niveau $\alpha = 5\%$, ob H_0 zutrifft – Ihre Teststatistik lautet

$$D_{n,p} = n[(f_{1,1} - p_{1,1})^2/p_{1,1} + (f_{1,2} - p_{1,2})^2/p_{1,2} + (f_{2,1} - p_{2,1})^2/p_{2,1} + (f_{2,2} - p_{2,2})^2/p_{2,2}]$$

und ist (asymptotisch) χ^2_1 verteilt. Wie lautet Ihr Test? Und kommen Sie bei den vorliegenden Zahlen zu einer Ablehnung?
Fallback: Verwenden Sie ggf. die Werte $p_{1,1} = 0.1$, $p_{1,2} = 0.6$, $p_{2,1} = 0.2$, $p_{2,2} = 0.1$.

Aufgabe 12 Regression **(4+6+4+3+2 Punkte)**

Sie wollen den Einfluss von Schokoladenkonsum auf Diätpläne betrachten. Dazu sei Y die Gewichtabnahme während des betrachteten Zeitraums in kg und zusätzlich kennen sie die im Zeitraum verspeiste Schokolade in Gramm X. Sie machen zwei Ansätze:

$$(I) \quad Y = \alpha X + \beta + U, \qquad (II) \quad Y = \alpha_1 \, \mathrm{I}(X \in (10; 300]) + \alpha_2 \, \mathrm{I}(X \in (300; \infty)) + \tilde{\beta} + \tilde{U}$$

für Fehler U bzw. $\tilde{U}$ und Koeffizienten $\alpha, \beta, \alpha_1, \alpha_2, \tilde{\beta} \in \mathbb{R}$. Dazu haben Sie n stochastisch unabhängige Beobachtungen (y_i, x_i) vorliegen.

a) Überlegen Sie sich jeweils eine Begründung, wie man zu einer Modellierung des Schokoladeneffektes wie in (I) und (II) kommen kann. Ist eine Modellierung wie in (II) grundsätzlich gröber? Was kann man mit einer Modellierung wie in (II) zusätzlich gegenüber (I) abbilden?

b) Bestimmen Sie den Kleinste-Quadrate Schätzer $\theta^{\text{LS.II}}$, $\sigma^2_{\text{LS.II}}$ in Modell II (unabhängig von den Zahlenwerten).

c) Welche Eigenschaften lassen sich allgemein für den Kleinste-Quadrate Schätzer zeigen? (Verzerrtheit, Linearität, Optimalität), und wie sind die Residuen r_i in einem linearen Modell definiert?

d) Der Kleinste-Quadrate Schätzer in Modell I ist bekanntlich über die Geradengleichung

$$y - \bar{y} = \hat{\rho}_{X,Y} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}(x - \bar{x})$$

bestimmt, für $\hat{\rho}_{X,Y}$ den (empirischen) Korrelationskoeffizienten, $\hat{\sigma}_X$, $\hat{\sigma}_Y$, die empirischen Standardabweichungen und $\bar{x}$, $\bar{y}$, die entsprechenden Mittelwerte, d.h. $\hat{\alpha}^{\text{LS.I}} = \hat{\rho}_{X,Y} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$ und $\hat{\beta}^{\text{LS.I}} = \bar{y} + \hat{\alpha}^{\text{LS.I}} \bar{x}$. Seien nun folgende Werte gegeben:

x_i	0	100	400	1000
y_i	3.5	3.3	0.5	−3

Berechnen Sie die kleinsten Quadrate-Schätzer in Modell I. Verwenden Sie dazu, dass $\hat{\sigma}_X = 450$ und (in erster Näherung) $\hat{\sigma}_Y = 3$.

e) Wie könnte man die beiden Modelle graphisch vergleichen?

Hinweise (für alle Aufgaben):

Verwenden Sie, dass für die $\mathcal{N}(0,1)$-, die t_m- und die χ^2_m-Verteilung folgende Quantile gelten:

α	1%	2.5%	5%	10%	90%	95%	97.5%	99%
$\Phi^{-1}(\alpha)$	-2.33	-1.96	-1.64	-1.28	1.28	1.64	1.96	2.33
$t_5^{-1}(\alpha)$	-3.36	-2.57	-2.02	-1.48	1.48	2.02	2.57	3.36
$t_6^{-1}(\alpha)$	-3.14	-2.45	-1.94	-1.44	1.44	1.94	2.45	3.14
$(\chi^2_1)^{-1}(\alpha)$	0	0.001	0.004	0.02	2.71	3.84	5.02	6.63
$(\chi^2_3)^{-1}(\alpha)$	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
$(\chi^2_5)^{-1}(\alpha)$	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
$(\chi^2_6)^{-1}(\alpha)$	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81

Verwenden Sie weiter folgende Werte der Verteilungsfunktion Φ von $\mathcal{N}(0,1)$

t	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$\Phi(t)$	0.50	0.54	0.58	0.62	0.66	0.69	0.73	0.76	0.79	0.82	0.84

Weiter gilt für den standardisierten Gini-Index (mit der Notation der Vorlesung):
$g = \frac{n}{n-1}[\sum_{i=1}^{k}(F_i + F_{i-1})(A_i - A_{i-1}) - 1]$