Institut für Mathematik
Prof. Dr. P. Ruckdeschel
M.Sc. T. Werner

WiSe 19/20
05. Februar 2020

Statistik I: Einführung in die Angewandte Statistik

Klausur (120 Minuten)

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

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

a) Tragen Sie zuerst Ihre persönlichen Angaben ein, unterschreiben Sie und legen Sie Ihren Personalausweis/Lichtbildausweis bereit.

b) Prüfen Sie nach, ob Ihr Klausurexemplar vollständig ist und aus 24 Seiten besteht.

c) Hilfsmittel sind nicht zugelassen. Hinweise und Verteilungstabellen finden Sie hinten.

d) Schreiben Sie Ihre Lösungen jeweils direkt unter den Aufgabentext und auf die nächstfolgenden Seiten. Bei Bedarf stehen Ihnen nach den Seiten mit den Tabellen noch zusätzliche Blätter zur Verfügung. Vermerken Sie bei der entsprechenden Aufgabe die Seitenzahl der Seite, die die Fortsetzung Ihrer Lösung enthält. **Die gehefteten Blätter dürfen nicht voneinander getrennt werden; die Benutzung eigenen Papiers ist nicht zulässig.**

e) Benutzen Sie einen dokumentenechten Stift (keinen Bleistift) und schreiben Sie nicht mit Rot. Mit Bleistift Geschriebenes und lose Blätter werden bei der Korrektur ignoriert.

f) Achten Sie darauf, den Lösungsweg zu jeder bearbeiteten Aufgabe nachvollziehbar aufzuschreiben und notwendige Zwischenschritte anzugeben.

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

h) Die durch die Übungen erworbenen Bonuspunkte werden nur bei Bestehen der Klausur berücksichtigt.

Zur Kenntnis genommen: .....
Unterschrift

Bewertung:

	Aufg. 1	Aufg. 2	Aufg. 3	Aufg. 4	Aufg. 5	Aufg. 6	$\sum$ 1–6	Bonus
Punkte	17	15	24	15	12	17	100	7
erreicht								
Korrektur								

Aufgabe 1 Deskriptive Statistik **([8+2]+[5+2]=17 Punkte)**

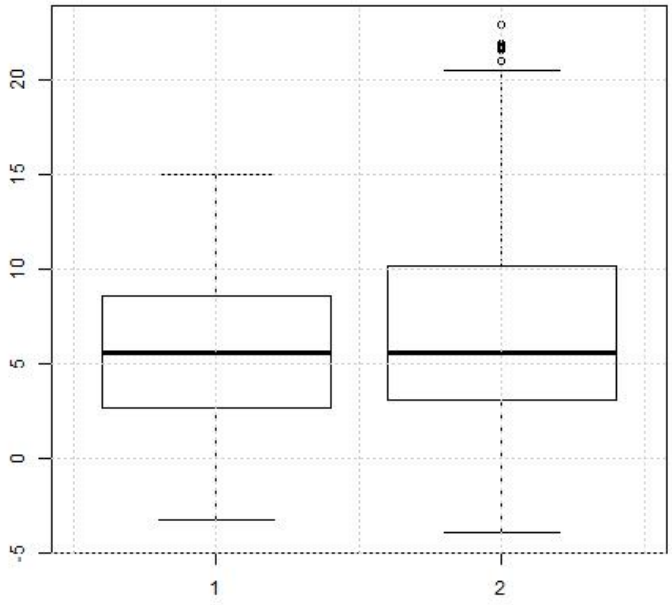
a) Eine Gruppe von 120 Personen teile sich ein Vermögen von 3 Millionen Euro. Es ist bekannt, dass die sechs reichsten Personen insgesamt 1,8 Millionen Euro besitzen. 24 Personen aus der Gruppe, die weniger reich sind als diese sechs, teilen sich 600000 Euro, 70 noch weniger reiche Personen teilen sich 490000 Euro und die 20 am wenigsten reichen Personen teilen sich 110000 Euro. Niemand hat Schulden.

(i) Berechnen Sie alle notwendigen Größen, um die Lorenzkurve zu zeichnen. Erklären Sie, warum Sie die wahre Lorenzkurve eigentlich nur linear approximieren.

Zeichnen Sie die Lorenzkurve (das muss nicht millimetergenau sein, nur qualitativ richtig) und berechnen Sie den Gini-Index.

(ii) Wie lautet die maximale Anzahl an Personen, die Null Euro besitzen können?

b) Betrachten Sie folgende Boxplots.



(i) Beschreiben Sie ausführlich die beiden Boxplots.

(ii) Vergleichen Sie die beiden korrespondierenden Stichproben hinsichtlich Streuung und Schiefe.

Aufgabe 2 Erzeugung von Zufallszahlen und Stichproben (4+2+[1+2+6]=15 Punkte)

a) Angenommen, Sie haben nur einen Zufallszahlengenerator zur Verfügung, der Ihnen für ein festes $N \in \mathbb{N}$ uniform verteilte Zufallszahlen auf der Menge $\{0, 1, ..., N\}$ erzeugt. Entwickeln Sie eine Idee, wie Sie damit näherungsweise auf $[0, 1]$ gleichverteilte Zufallszahlen erzeugen können, sodass der maximale Abstand Ihrer Zahl zu einer tatsächlich aus einer $U([0, 1])-$Verteilung stammenden Zufallszahl maximal ϵ für ein festes $\epsilon > 0$ ist.

Schreiben Sie anschließend Ihren Algorithmus als Pseudocode auf (es werden keine R−Befehle erwartet!).

b) Sie wollen eine Stichprobe erzeugen, die einen Mittelwert von 20 und eine Standardabweichung von 16 hat. Zu diesem Zwecke haben Sie bereits eine Stichprobe x aus einer $\mathcal{N}(5, 4)-$Verteilung erzeugt. Wie können Sie durch Modifikation von x nun eine Stichprobe mit den gewünschten Eigenschaften bekommen? Begründen Sie Ihren Vorschlag mathematisch exakt.

c) X folge der Verteilung mit Dichte

$$f_X(x) = cx^2 I_{[0,3]}(x)$$

(i) Bestimmen Sie c, sodass f_X eine Dichte ist.

(ii) Sei $Y := -ln(2\sqrt{X}) - 1$. Bestimmen Sie zunächst den Bereich, in dem Y Werte annehmen kann und berechnen Sie anschließend den Wert der Verteilungsfunktion $F_Y(t)$ von Y für alle $t \in \mathbb{R}$.

(iii) Wir haben zwei Zufallszahlengeneratoren zur Verfügung, nämlich einen zur Erzeugung von Realisationen von X und einen zur Erzeugung von Realisationen einer $U([0, 1])-$verteilten Zufallsvariable. Schreiben Sie den Pseudocode (keine konkreten R−Befehle!) für zwei Varianten auf, wie man Realisationen von Y erzeugen könnte.

Aufgabe 3 Schätzen und Normalapproximation ([2+6+3]+[2+7+4]=24 Punkte)

a) Betrachten Sie für ein $\theta \in \mathbb{R}$ die Zufallsvariablen $X_1, ..., X_n$, die i.id. $U([5-\theta, 5+\theta])-$verteilt sind.

(i) Bestimmen Sie Erwartungswert und Varianz von X_1.

(ii) Bestimmen Sie einen Momentenschätzer für θ basierend auf dem zweiten Moment. Prüfen Sie Ihren Schätzer auf Erwartungstreue, schwache und starke Konsistenz.

(iii) Bestimmen Sie den Maximum-Likelihood-Schätzer für θ.

b) Man habe 500 Personen, die sich zu einer Klausur angemeldet haben. Allerdings hat der Hörsaal nur 420 Sitzplätze. Aus den Erfahrungen der letzten Jahre leitet man ab, dass eine angemeldete Person mit einer Wahrscheinlichkeit von 12 Prozent nicht erscheint.

(i) Verwenden Sie geeignet einen zentralen Grenzwertsatz, um die Wahrscheinlichkeit zu berechnen, dass die Anzahl der Sitzplätze nicht ausreichen wird.

HINWEIS: Ausdrücke wie $\Phi(...)$ bleiben so stehen.

(ii) Man lernt aus Erfahrungen: Im nächsten Durchgang (gleiche Anzahl der Sitzplätze, gleiche Wahrscheinlichkeit des Erscheinens) wird eine Obergrenze für die Anzahl der Klausurteilnehmer festgelegt. Bestimmen Sie diese, basierend auf einer Normalapproximation, so, dass die Sitzplätze mit einer Wahrscheinlichkeit von 95 Prozent ausreichen werden.

(iii) Angenommen, Sie hätten keine Punktschätzung für die Wahrscheinlichkeit, dass eine zufällig zur Klausur angemeldete Person nicht erscheint, sondern hätten diese Wahrscheinlichkeit durch einen Prior geschätzt, der eine uniforme Verteilung auf $[0.1, 0.14]$ ist. Es seien in a) nun von den 500 angemeldeten Personen tatsächlich 454 erschienen. Schreiben Sie die Posterior-Wahrscheinlichkeit, dass die Wahrscheinlichkeit p, dass eine zufällig ausgewählte angemeldete Person erscheint, auf. Verwenden Sie dafür die exakten Dichten und keine Normalapproximation!

HINWEIS: Versuchen Sie nicht, die Integrale zu lösen, sondern vereinfachen Sie die Ausdrücke so weit wie möglich. Am Ende bleibt ein Quotient von zwei Integralen stehen.

Aufgabe 4 Testen I **(6+2+6+1=15 Punkte)**

a) Sei die Situation von Aufgabe 3 gegeben, d.h., 454 der 500 Personen sind erschienen. Führen Sie den exakten zweiseitigen Binomialtest durch, um zum Niveau $\alpha = 0.05$ zu testen, ob die ursprünglich geschätzte Quote von 12 Prozent abgelehnt werden kann. Bestimmen Sie die Macht des Binomialtests für eine wahre Wahrscheinlichkeit von 0.15 für das Nicht-Erscheinen und beschreiben Sie, was Sie damit über den Fehler 2. Art aussagen können.

b) Begründen Sie, welchen Vorteil eine Randomisierung des Binomialtests hat und bestimmen Sie den randomisierten Test. Wie lautet nun Ihre Testentscheidung?

c) Es seien $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ Zufallsvariablen, die die Länge von Schrauben in Millimetern modellieren. Wir wollen testen, ob die Schraubenlänge im Mittel 5.2 Zentimeter beträgt. Für eine Stichprobe $(x_1, ..., x_{100})$ von 100 Schrauben haben wir die Größen

$$\sum_i x_i = 5016.68, \quad \sum_i x_i^2 = 263182.8$$

berechnet. Bestimmen Sie ein geeignetes Schwankungsintervall für den Mittelwert und treffen Sie die Testentscheidung. Erklären Sie, ohne ihn konkret auszurechnen, was hier den p-Wert ist und treffen Sie eine Abschätzung, in welchem Intervall sich der p-Wert hier aufhält.

d) Begründen Sie, ob die Aussage, dass bei Vorliegen eines p-Wertes von 0.02 die Alternative mit einer Wahrscheinlichkeit von 98 Prozent wahr ist, korrekt ist oder nicht.

Aufgabe 5 Testen II **(4+2+3+3=12 Punkte)**

a) Seien $X_1, ..., X_8 \sim P$ i.id. und $Y_1, ..., Y_6 \sim Q$ i.id. für stetige Verteilungen P und Q. Wir haben folgende Beobachtungen:

$$x = (8, 7, 10, 12, 19, 28, 4, -1), \quad y = (-2, 5, 11, 26, 25, 17).$$

Testen Sie zum Niveau $\alpha = 0.01$ die Nullhypothese, dass $P = Q$ gegen die Alternative, ob P stochastisch größer ist als Q. Sie dürfen die Normalapproximation der Teststatistik verwenden.

b) Sie haben eine Stichprobe $(x_1, ..., x_n)$ von $X_1, ..., X_n \sim P$ i.id. und wollen mit Hilfe eines Kolmogorov-Smirnov-Tests prüfen, ob $P = P_0$ für Verteilungen P_0 und P. Skizzieren Sie die Idee hinter dem Test.

c) Für zwei binäre Merkmale X und Y, die ohne Einschränkung die Ausprägungen 0 und 1 besitzen, haben Sie 200 Beobachtungen des Paares (X, Y) erhoben und folgende absolute Häufigkeitstabelle bekommen:

X Y	0	1
0	55	23
1	101	21

Testen Sie auf Unabhängigkeit zum Niveau $\alpha = 0.025$. Was können Sie über den p-Wert sagen?

d) Sie haben woanders noch einmal 300 Beobachtungen des Paares (X, Y) erhoben und kommen auf die absolute Häufigkeitstabelle

X Y	0	1
0	77	31
1	144	48

Skizzieren Sie Ihr Vorgehen, um einen Homogenitätstest zum Niveau $\alpha = 0.05$ durchzuführen. Sie brauchen nichts explizit auszurechnen! Geben Sie aber am Ende den kritischen Wert für die Teststatistik an.

Aufgabe 6 Regression **(2+[4+1+2+1+2]+3+2=17 Punkte)**

a) Sie haben folgende Regressormatrix X für die vier Variablen (Spalten) $X_1, ..., X_4$ gegeben:

$$\begin{pmatrix} 1 & 4 & -2 & A \\ -1 & 3 & -5 & B \\ 2 & 0 & 4 & B \\ -2 & 3 & -4 & B \end{pmatrix}.$$

Stellen Sie die Regressormatrix auf, die Sie für die Regression für das Modell

$$Y_i = \beta_0 + \beta_1 X_{i,1}^2 + \beta_2 X_{i,3} + \beta_3 X_{i,3}^3 + \beta_4 X_{i,4}$$

brauchen.

b) Betrachten Sie folgenden reduzierten Output einer linearen Regression:

Call: lm(formula = y ∼ ., data = D)

Residuals:	Min	1Q	Median	3Q	Max
	-25.361	-6.7933	-1.013	6.351	25.002

Coefficients:	Estimate	Std. Error
(Intercept)	3.101	13.794
V1	1.744	1.254
V2	-2.722	1.195
V3	1.254	1.329
V4B	19.335	3.127
V4C	-22.907	2.867

Residual standard error: 10.73 on 76 degrees of freedom
Multiple R-squared: 0.74, Adjusted R-squared: 0.7229
F-statistic: 43.26 on 5 and 76 DF, p-value: ¡ 2.2e-16

Bearbeiten Sie folgende Punkte dazu:

(i) Testen Sie die Variablen V_1 bis V_3 auf Signifikanz. Nehmen Sie $\alpha = 0.05$ als Signifikanzniveau.

(ii) Welchen Effekt hat eine Erhöhung der Variable $V2$ um eine Einheit?

(iii) Wie interpretieren Sie die Koeffizienten $V4B$ und $V4C$?

(iv) Wie viele Beobachtungen liegen vor?

(v) Was sagen anschaulich (d.h., ohne Formeln) die Werte für ,,Multiple R-squared" und ,,Adjusted R-squared" aus?

c) Betrachten Sie folgendes Konkurrenz-Modell:

```
Call: lm(formula = y ~ ., data = D[,-3])

Residuals:      Min       1Q   Median       3Q      Max
              -24.7037  -6.633  -0.7329   6.3637  25.8074

Coefficients: Estimate   Std. Error
(Intercept)   11.705      10.342
V1             1.548      1.236
V2            -2.470      1.164
V4B           19.291      3.125
V4C          -22.929      2.864


Residual standard error: 10.72 on 77 degrees of freedom
Multiple R-squared: 0.737, Adjusted R-squared: 0.7233
F-statistic: 53.93 on 4 and 77 DF, p-value: ¡ 2.2e-16
```

Welches Modell würden Sie aufgrund der zur Verfügung stehenden Statistiken vorziehen?

d) Stellen Sie das generalisierte lineare Modell für die Logit-Regression auf.

Hinweise (für alle Aufgaben):

a) Der Gini-Index berechnet sich durch

$$G(X) = \sum_{i=1}^{k} (F_{i-1} + F_i) a_i - 1.$$

b) Im Zusammenhang von Wilcoxon-Tests haben wir

$$W = \sum_i R_i, \quad U = \sum_i \sum_j I(X_i > Y_j)$$

und es gilt, dass

$$\frac{U - 0.5mn}{\sqrt{\frac{mn(m+n+1)}{12}}} \to \mathcal{N}(0,1)$$

schwach für $m, n \to \infty$.

c) Tabelle einiger Quantile der Standardnormalverteilung:

α	0.005	0.01	0.025	0.05
	-2.58	-2.33	-1.96	-1.64

d) Tabelle von $(\chi^2_d)^{-1}(\alpha)-$Werten:

d \ α	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995
1	0.000039	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.1	5.99	7.38	9.21	10.6
3	0.072	0.11	0.22	0.35	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	9.488	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	11.07	12.83	15.09	16.75
6	0.678	0.87	1.24	1.64	12.59	14.45	16.81	18.55
199	151.37	155.55	161.83	167.36	232.91	239.96	248.33	254.14
299	239.77	245.07	252.99	259.95	340.33	348.79	358.81	365.74

e) Tabelle von $t_\nu^{-1}-$Werten:

ν \ α	0.005	0.01	0.025	0.05
75	-2.6430	-2.3771	-1.9921	-1.6654
76	-2.6421	-2.3764	-1.9917	-1.6652
77	-2.6412	-2.3758	-1.9913	-1.6649
78	-2.6403	-2.3751	-1.9908	-1.6646
99	-2.6264	-2.3646	-1.9842	-1.6604
100	-2.6259	-2.3642	-1.9840	-1.6602
101	-2.6254	-2.3638	-1.9837	-1.6601
190	-2.6019	-2.3461	-1.9725	-1.6529
191	-2.6018	-2.346	-1.9725	-1.6529
192	-2.6017	-2.3459	-1.9724	-1.6528

f) Tabelle von Werten von $P(X \leq k)$ für $X \sim Bin(n, p)$ (auf drei bzw. vier Nachkommastellen gerundet):

$k \setminus (n,p)$	(500,0.12)	(500,0.85)	(499,0.88)	(499,0.15)
45	0.020	0		
46	0.028	0		
47	0.039	0		
48	0.053	0		
59		0		0.024
60		0		0.033
61		0		0.044
62		0		0.058
71	0.941	0		
72	0.954	0		
73	0.966	0		
74	0.974	0		
75	0.981	0		
87		0		0.941
88		0		0.954
90		0		0.973
91		0		0.979
408		0.022		
409		0.028		
411		0.048		
412		0.061		
424		0.469	0.025	
425		0.519	0.033	
426		0.569	0.044	
427		0.618	0.057	
437		0.944		
438		0.957		
439		0.968		
440		0.977		
450		0.9996	0.945	
451		0.9998	0.959	
452		0.9999	0.971	
453		0.9999	0.979	