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

WiSe 18/19
06. Februar 2019

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 41 Seiten besteht.

c) Um Ihnen die Übersicht zu erleichtern, sind am Anfang der Klausur auf den Seiten 2 bis 7 alle Aufgaben einmal abgedruckt.

d) Hilfsmittel sind nicht zugelassen. Notationen und Hinweise finden Sie auf Seite 8.

e) Schreiben Sie Ihre Lösungen **ab Seite 10** jeweils direkt unter den Aufgabentext und auf die nächstfolgende Seite. Bei Bedarf stehen Ihnen am Ende der Klausur ab Seite 39 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.**

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

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

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

i) 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	30	10	12	11	11	17	91	7
erreicht								
Korrektur								

	Aufg. 7	Aufg. 8	Aufg. 9	Aufg. 10	Aufg. 11	Aufg. 12	$\sum$ 7–12	Gesamt	Note
Punkte	9	10	16	10	7	7	59	150 (+7)	
erreicht									
Korrektur									

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

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

a) Der MAD einer Stichprobe ist stets größer als Null.

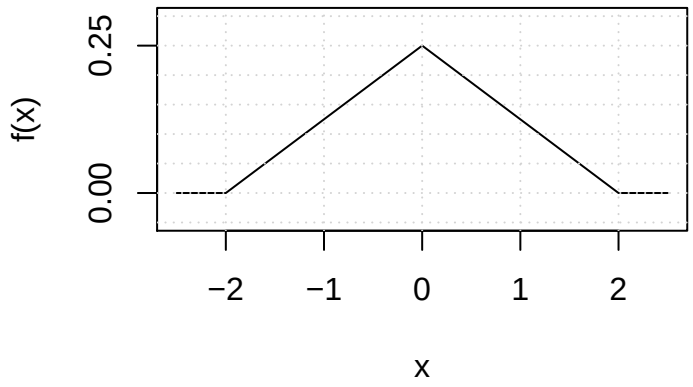
b) Für zwei Ereignisse A, B, mit $P(A), P(B) > 0$ gilt immer, dass

$$P(A|B^c) \geq P(A \setminus B).$$

NOTATION: Es ist $A \setminus B = A \cap B^c$.

c) Sei $n \in \mathbb{N}, n \geq 2$, die Länge einer gegebenen Stichprobe. Dann genügt es, eine Beobachtung zu manipulieren, um die empirische Varianz größer als jede Schranke zu treiben.

d) Die folgend graphisch dargestellte Funktion ist eine Dichte.



e) Ein Boxplot, der keine Antennen hat, hat auch keine Ausreißer.

f) Jede diskrete Dichte auf $\mathbb{N}$ lässt sich als Summe zweier diskreter Dichten auf $\mathbb{N}$ schreiben.

g) Folgt X einer $Bin(2, 0.3)-$Verteilung, so gilt für die Zähldichte p_Y von $Y := \max(X + 2, 3)$, dass $p_Y(3) = 0.91$.

h) Folgt die Prüfgröße S eines statistischen Tests einer absolutstetigen Verteilung, so lässt sich das Signifikanzniveau immer ausschöpfen.

i) Ein MSE-effizienter Schätzer ist immer unverzerrt.

j) Bei der multiplen Regression gelte

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2(X^T X)^{-1}).$$

Dann lässt sich für jeden Prädiktor anhand eines $t-$Tests für ein vorgegebenes Signifikanzniveau entscheiden, ob der zugehörige Koeffizient von 2 verschieden ist.

Aufgabe 2 Deskriptive Kennzahlen **(3+2+5=10 Punkte)**

a) Sie haben die Beobachtungen

$$x = (-12, 5, 8, 1, -1, -25, 0, -2)$$

vorliegen. Berechnen Sie das 30%− und das 80%−Quantil der Stichprobe mit der Konvention, dass das $\alpha-$Quantil gerade die $k-$te Beobachtung des geordneten Vektors ist mit

$$(k-1)/(n-1) = \alpha,$$

und bei $k = m + c, m \in \mathbb{N}, c \in]0, 1[$ wird zwischen der $m-$ten und $(m+1)-$ten Beobachtung linear interpoliert.

b) Berechnen Sie den MAD der obigen Stichprobe.

c) Zeichnen Sie nun den Boxplot der in a) vorliegenden Daten und erläutern Sie ausführlich Ihr Vorgehen.

Aufgabe 3 Konzentrationen **(3+9 Punkte)**

Benötigte Formeln finden Sie auf Seite 8.

a) **(M)** Zeigen Sie durch Betrachtung geeigneter extremer Konstellationen, dass der Gini-Index zwar den Wert Null, aber nicht den Wert 1 annehmen kann.

HINWEIS: Eine Zeichnung kann hilfreich sein, ist aber nicht gefordert.

b) Durch vulkanische Aktivität ist eine Insel aufgetaucht. Sie interessieren sich für die Verteilung der Biomasse der Tiere (ab einer gewissen Mindestgröße) nach einer gewissen Zeit nach dem Auftauchen der Insel. Es gibt dort bislang vier entsprechende Spezies, nämlich eine eine Möwenart, zwei Käferarten und eine Krabbenart. Sie schätzen durch Zählungen bzw. statistische Hochrechnungen die Anzahl der Möwen auf 30, die der Krabben auf 40, die der Käfer auf 2000 bzw. 6000. Sie wissen ferner, dass die Individuen der entsprechenden Möwenart ein durchschnittliches Gewicht von 900 Gramm auf die Waage bringen, die der Krabbenart von 300 Gramm und die der Käferarten von 2 Gramm bzw. 150 Milligramm.

Zeichnen Sie die Lorenzkurve und berechnen Sie den Gini-Index sowie den Herfindahl-Index.

Aufgabe 4 Definitionen, Sätze und Beweise **(3+2+2+4=7+4 Punkte)**

a) Formulieren und beweisen Sie den Satz von der totalen Wahrscheinlichkeit.

b) Formulieren Sie die Jensen-Ungleichung.

c) Beweisen Sie, dass für eine $\mathbb{N}_0-$wertige Zufallsvariable X immer gilt, dass

$$\mathbb{E}[X] \geq P(X \geq 1).$$

d) **(I)** Definieren Sie das Klumpenstichprobenziehungsverfahren und beschreiben Sie an einem selbst gewählten Beispiel, wie man es in einer Software realisieren würde.

Aufgabe 5 Diskrete Verteilungen **(3+3+2+3=11 Punkte)**

Ein Ikosaeder ist ein Körper mit 20 gleichen (und damit gleich wahrscheinlichen) Flächen. Diese sind mit den Zahlen von 1 bis 20 nummeriert.

a) Geben Sie den Ergebnisraum für den Wurf eines Ikosaeder-Paares an. Bestimmen Sie die Wahrscheinlichkeit, dass das Minimum beider gesehener Zahlen kleiner gleich 2 war.

HINWEIS: Stellen Sie keine $20 \times 20-$Tafel auf...

b) Sie werfen drei Ikosaeder und notieren sich das Produkt der Zahlen. Führen Sie eine Zufallsvariable X ein, die diese Produkte modelliert und geben Sie in Mengenschreibweise den Ergebnisraum an, in den X abbildet (diesen schreiben Sie NICHT explizit auf!). Berechnen Sie die Wahrscheinlichkeit $P(X \in \{1, 4\})$.

c) Ihr/e Übungspartner/in hat zwei Ikosaeder geworfen und teilt Ihnen mit, dass das Ergebnis jedes einzelnen der beiden kleiner als 13 war. Wie hoch ist mit diesem Vorwissen nun die Wahrscheinlichkeit, dass die Summe der beiden Zahlen 21 war?

d) Ihr/e Übungspartner/in hat zwei Ikosaeder geworfen und teilt Ihnen diesmal mit, dass die Summe größer als 37 ist. Sei Y die Zufallsvariable, die den Logarithmus des Produkts der Zahlen modelliert. Berechnen Sie den Erwartungswert von Y, bedingt auf die gegebene Vorinformation.

Aufgabe 6 Stetige Verteilungen **(1+2+4+2+4+4=9+4+4 Punkte)**

Gegeben sei folgende Funktion:

$$f_{XY}(x, y) = c(4 + 4x^2 - y - x^2 y) I_{[0,1]}(x) I_{[0,2]}(y).$$

a) Bestimmen Sie c, sodass die Funktion die Dichte eines Zufallsvektors (X, Y) ist.

HINWEIS: Prüfen Sie dazu, ob

$$f_{XY} \geq 0, \quad \int f_{X,Y}(x, y)\, dx\, dy = 1.$$

b) Bestimmen Sie die Randdichten f_X, f_Y.

c) Berechnen Sie $\mathrm{Var}(X), \mathrm{Var}(Y)$.

d) Sind X und Y unkorreliert?

e) **(M)** Berechnen Sie $\mathbb{E}[e^Y]$.

f) **(I)** Beschreiben Sie, wie Sie mit Hilfe einer Software vorgehen würden, um $\mathbb{E}[e^Y]$ anzunähern. Nehmen Sie an, dass Sie $U([0, 1])-$verteilte Zufallszahlen erzeugen können.

Aufgabe 7 Konvergenz, Grenzwertsätze **(4+2+3=5+4 Punkte)**

a) **(M)** Betrachten Sie den zentralen Grenzwertsatz von Lindeberg-Lévy: X_i u.i.v. mit $\mathbb{E}[X_1] = \mu$, $\mathrm{Var}(X_1) = \sigma^2$. Dann gilt

$$\sqrt(n)(\bar{X}_n - \mu)/\sigma \xrightarrow{w} \mathcal{N}(0,1).$$

Leiten Sie daraus den zentralen Grenzwertsatz von de-Moivre-Laplace ab, nämlich dass für $Y_n \sim \mathrm{Bin}(n,p)$ gilt, dass

$$\sqrt{n}(Y_n/n - p) \xrightarrow{w} \mathcal{N}(0, p(1-p)).$$

In welchem Sinn konvergiert dabei die Verteilung der linken Seite der Konvergenzaussage gegen die der rechte Seite?

b) Spezifizieren Sie eine statistische Anwendung des zentralen Grenzwertsatzes von de-Moivre-Laplace beim Schätzen oder Testen.

c) Seien $X_1, ..., X_n$ unabhängig identisch $\mathcal{N}(6, \sigma^2)-$verteilt mit $\sigma^2 > 0$. Wie groß muss n sein, damit die Wahrscheinlichkeit, dass der Abstand von $\bar{X}_n$ und $\mathrm{E}[\bar{X}_n]$ weniger als 4 Prozent von $\mathrm{E}[\bar{X}_n]$ ausmacht, mindestens 95 Prozent ist?

Aufgabe 8 Schätzen **(4+3+3=7+3 Punkte)**

a) Sie haben eine Stichprobe $x = (x_1, ..., x_m)$, deren Komponenten Sie alle i.id. als $\Gamma(n, \lambda)-$verteilt annehmen. Sie kennen n, aber nicht λ. Bestimmen Sie den Maximum-Likelihood-Schätzer für λ. In welchem Sinne ist dieser konsistent?

HINWEIS: Beachten Sie die Hinweise auf Seite ....

b) Sei haben Beobachtungen $x = (x_1, ..., x_{101})$ erhoben, deren Komponenten alle i.id. einer unbekannten Verteilung mit existierendem Erwartungswert und existierender, positiver Varianz, folgen. Sie haben folgende Größen berechnet:

$$\frac{1}{101}\sum_{i=1}^{101} x_i = 3.5, \quad \frac{1}{100}\sum_{i=1}^{101}(x_i - 3.5)^2 = 4.84.$$

Bestimmen Sie ein asymptotisches, zweiseitiges Konfidenzintervall für den Erwartungswert für $\alpha = 0.1$.

c) **(I)** Beschreiben Sie, wie man mit Hilfe des statistischen Bootstraps ein empirisches 95%−Konfidenzintervall für einen Punktschätzer bestimmen kann.

Aufgabe 9 Hypothesentests **(2+2+4+4+4=16 Punkte)**

a) Was versteht man im Kontext von statistischen Tests unter dem Fehler 1. und 2. Art?

b) Was ist die Macht eines Tests? Beschreiben Sie anschaulich, warum eine große Macht wünschenswert ist.

c) Jemand behauptet, in einer Klausur, in der man insgesamt 150 Punkte bekommen kann, würden im Mittel 60 erreicht, mit einer Standardabweichung von 12. Sie glauben den Mittelwert nicht (nehmen aber die Standardabweichung als richtig an) und erheben selbst Daten. Der Mittelwert Ihrer 191 Beobachtungen ist 57.5. Erstellen Sie mit Hilfe eines Grenzwertsatzes einen asymptotischen Test und treffen Sie die Testentscheidung anhand Ihrer Daten.

d) Sie haben Beobachtungen $x_1, ..., x_{150}$, die alle unabhängig aus einer $\mathcal{N}(-2, \sigma^2)-$Verteilung stammen. Sie haben

$$\frac{1}{150}\sum_{i=1}^{150}(x_i + 2)^2 = 3.93$$

berechnet. Entscheiden Sie für $\alpha = 0.01$, ob Sie die Nullhypothese, dass die wahre Standardabweichung 1.7 ist, ablehnen können.

e) Bestimmen Sie für Ihren Test aus d) die Macht. Wie groß ist diese für $\theta_1 := 1.8$?

Aufgabe 10 Kontingenztabellen **(2+2+6=10 Punkte)**

Sie interessiert die Frage, ob Personaler sich von einem Foto beeinflussen lassen, d.h., ob die Reaktion auf die Bewerbung von der Art des Fotos bzw. dessen Vorhandenseins abhängt. Dazu schicken drei Personen mit sehr ähnlicher (guter) Qualifikation und sehr ähnlichen Lebensläufen jeweils Bewerbungen für jeweils die gleiche Stelle ab. Jedes Mal hat eine zufällig ausgewählte Person kein Foto angehängt, eine andere hat ein professionelles Bewerbungsfoto und die dritte ein Passbild. Wir nehmen für diese Aufgabe an, dass die marginalen Unterschiede, die nicht das Foto betreffen, keinen Einfluss auf die Entscheidung des Personalers haben.

Dieses Vorgehen wurde für 120 Stellen durchgeführt. Von den Bewerbungen ohne Foto führten 54 zu einer positiven Reaktion, was bei den Bewerbungen mit Passbild nur 30 Mal der Fall war. Bei den übrigen Bewerbungen fiel die Reaktion in zwei Dritteln der Fälle positiv aus. Eine Reaktion gab es immer, d.h., die Alternative zu einer positiven Reaktion ist lediglich die Absage.

a) Stellen Sie aus den obigen Informationen die Kontingenztabelle auf, einmal für die absoluten, einmal für die relativen Häufigkeiten.

b) Berechnen Sie die Tabellen der bedingten Häufigkeiten.

c) Beantworten Sie durch einen geeigneten statistischen Test, ob Foto und Reaktion abhängig sind. Es sei $\alpha = 0.01$.

Aufgabe 11 Lineare Regression **(2+[1+1+1+1+1]=7 Punkte)**

a) Stellen Sie das multiple lineare Regressionsmodell auf (für eine allgemeine Anzahl p an erklärenden Variablen). Schreiben Sie das Optimierungsproblem auf, welches durch den Regressionsschätzer gelöst wird.

b) Betrachten Sie folgenden R−Output:

```
Call: lm(formula = Y ~ ., data = D)

Residuals:     Min      1Q  Median      3Q     Max
            -2.7240 -0.9957 -0.0331  1.0083  3.3947

Coefficients: Estimate     Std.     Error    t value     Pr(>|t|)
(Intercept)   -0.0986     0.2452    -0.402    0.6896
X1             1.3088     0.2508     5.219    4.67e-06    ***
X2            -0.5275     0.2769    -1.905    0.0633      .
X3             1.8363     0.2483     7.395    3.02e-09    ***
X4             1.2781     0.2766     4.621    3.34e-05    ***
X5             0.2833     0.2127     1.332    0.1897
— Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.574 on 44 degrees of freedom
Multiple R-squared: 0.764, Adjusted R-squared: 0.7371
F-statistic: 28.48 on 5 and 44 DF, p-value: 9.228e-13
```

(a) Welche Koeffizienten sind signifikant? Begründung?

(b) Was bedeutet anschaulich der Koeffizient von X2?

(c) Zeigen Sie anhand der anderen vorliegenden Informationen, wie Sie auf den t-Wert für X4 kommen. Welche Hypothese steckt dahinter?

(d) Wie bekommen Sie für die Variable X3 den angegebenen p-Wert?

(e) Begründen Sie die Anzahl der Freiheitsgrade.

Aufgabe 12 Generalisierte lineare Modelle **(2+3+1+1=7 Punkte)**

a) Definieren Sie ein allgemeines generalisiertes lineares Modell. Was versteht man unter einer Linkfunktion? Wie lautet diese für das gewöhnliche lineare Regressionsmodell (Aufgabe 11)?

b) Nehmen Sie an, Ihre abhängige Variable Y nimmt die Werte ,,grün" und ,,rot" an. Sie haben p erklärende Variablen $X_1, ..., X_p, p \geq 2$, zur Verfügung. Stellen Sie für diesen konkreten Fall Ihr Logit-Modell mit allen erklärenden Variablen auf.

c) Sie haben nun die Koeffizienten für Ihr Modell aus b) vorliegen. Der Koeffizient β_2 von X_2 ist 0.663. Wie interpretieren Sie diesen Koeffizienten anschaulich?

d) Angenommen, für drei neue Beobachtungen (d.h., alle X-Werte sind bekannt, der Y-Wert jeweils unbekannt) sind die berechneten Log-Odds 0.36, 0.71 und 0.88. Für welche der Beobachtungen sagen Sie rot bzw. grün voraus?

Hinweise (für alle Aufgaben):

a) Der normierte Herfindahl-Index ist

$$K_H^{norm}(X) = \frac{k\left[\sum_{i=1}^k f_i^2\right] - 1}{k-1}.$$

b) Der Gini-Index berechnet sich durch

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

c) Die Dichte der $\Gamma(n,\lambda)-$Verteilung ist

$$f_{n,\lambda}(x) = \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x} I(x > 0)$$

für $n \in \mathbb{N}, \lambda > 0$.

d) Für $X \sim Exp(\lambda)$ gilt $\mathbb{E}[X] = \lambda^{-1}$ und $Var(X) = \lambda^{-2}$.

Tabelle einiger Quantile der Standardnormalverteilung:

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

Tabelle von $(\chi_d^2)^{-1}(\alpha)-$Werten:

d \ α	0.01	0.025	0.05	0.95	0.975	0.99
1	0.00016	0.00098	0.00393	3.84	5.02	6.63
2	0.02	0.05	0.1	5.99	7.38	9.21
3	0.11	0.22	0.35	7.81	9.35	11.34
6	0.87	1.24	1.64	12.59	14.45	16.81

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

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

a) Der MAD einer Stichprobe ist stets größer als Null.

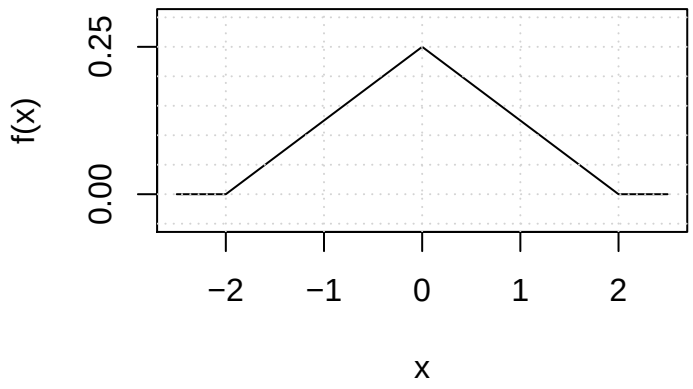
b) Für zwei Ereignisse A, B, mit $P(A), P(B) > 0$ gilt immer, dass

$$P(A|B^c) \geq P(A \setminus B).$$

NOTATION: Es ist $A \setminus B = A \cap B^c$.

c) Sei $n \in \mathbb{N}, n \geq 2$, die Länge einer gegebenen Stichprobe. Dann genügt es, eine Beobachtung zu manipulieren, um die empirische Varianz größer als jede Schranke zu treiben.

d) Die folgend graphisch dargestellte Funktion ist eine Dichte.



e) Ein Boxplot, der keine Antennen hat, hat auch keine Ausreißer.

f) Jede diskrete Dichte auf $\mathbb{N}$ lässt sich als Summe zweier diskreter Dichten auf $\mathbb{N}$ schreiben.

g) Folgt X einer $Bin(2, 0.3)-$Verteilung, so gilt für die Zähldichte p_Y von $Y := \max(X + 2, 3)$, dass $p_Y(3) = 0.91$.

h) Folgt die Prüfgröße S eines statistischen Tests einer absolutstetigen Verteilung, so lässt sich das Signifikanzniveau immer ausschöpfen.

i) Ein MSE-effizienter Schätzer ist immer unverzerrt.

j) Bei der multiplen Regression gelte

$$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2(X^T X)^{-1}).$$

Dann lässt sich für jeden Prädiktor anhand eines $t-$Tests für ein vorgegebenes Signifikanzniveau entscheiden, ob der zugehörige Koeffizient von 2 verschieden ist.

Aufgabe 2 Deskriptive Kennzahlen **(3+2+5=10 Punkte)**

a) Sie haben die Beobachtungen

$$x = (-12, 5, 8, 1, -1, -25, 0, -2)$$

vorliegen. Berechnen Sie das $30\%-$ und das $80\%-$Quantil der Stichprobe mit der Konvention, dass das $\alpha-$Quantil gerade die $k-$te Beobachtung des geordneten Vektors ist mit

$$(k-1)/(n-1) = \alpha,$$

und bei $k = m + c, m \in \mathbb{N}, c \in]0, 1[$ wird zwischen der $m-$ten und $(m+1)-$ten Beobachtung linear interpoliert.

b) Berechnen Sie den MAD der obigen Stichprobe.

c) Zeichnen Sie nun den Boxplot der in a) vorliegenden Daten und erläutern Sie ausführlich Ihr Vorgehen.

Aufgabe 3 Konzentrationen **(3+9 Punkte)**

Benötigte Formeln finden Sie auf Seite 8.

a) **(M)** Zeigen Sie durch Betrachtung geeigneter extremer Konstellationen, dass der Gini-Index zwar den Wert Null, aber nicht den Wert 1 annehmen kann.

HINWEIS: Eine Zeichnung kann hilfreich sein, ist aber nicht gefordert.

b) Durch vulkanische Aktivität ist eine Insel aufgetaucht. Sie interessieren sich für die Verteilung der Biomasse der Tiere (ab einer gewissen Mindestgröße) nach einer gewissen Zeit nach dem Auftauchen der Insel. Es gibt dort bislang vier entsprechende Spezies, nämlich eine eine Möwenart, zwei Käferarten und eine Krabbenart. Sie schätzen durch Zählungen bzw. statistische Hochrechnungen die Anzahl der Möwen auf 30, die der Krabben auf 40, die der Käfer auf 2000 bzw. 6000. Sie wissen ferner, dass die Individuen der entsprechenden Möwenart ein durchschnittliches Gewicht von 900 Gramm auf die Waage bringen, die der Krabbenart von 300 Gramm und die der Käferarten von 2 Gramm bzw. 150 Milligramm.

Zeichnen Sie die Lorenzkurve und berechnen Sie den Gini-Index sowie den Herfindahl-Index.

Aufgabe 4 Definitionen, Sätze und Beweise (3+2+2+4=7+4 Punkte)

a) Formulieren und beweisen Sie den Satz von der totalen Wahrscheinlichkeit.

b) Formulieren Sie die Jensen-Ungleichung.

c) Beweisen Sie, dass für eine $\mathbb{N}_0-$wertige Zufallsvariable X immer gilt, dass

$$\mathbb{E}[X] \geq P(X \geq 1).$$

d) **(I)** Definieren Sie das Klumpenstichprobenziehungsverfahren und beschreiben Sie an einem selbst gewählten Beispiel, wie man es in einer Software realisieren würde.

Aufgabe 5 Diskrete Verteilungen **(3+3+2+3=11 Punkte)**

Ein Ikosaeder ist ein Körper mit 20 gleichen (und damit gleich wahrscheinlichen) Flächen. Diese sind mit den Zahlen von 1 bis 20 nummeriert.

a) Geben Sie den Ergebnisraum für den Wurf eines Ikosaeder-Paares an. Bestimmen Sie die Wahrscheinlichkeit, dass das Minimum beider gesehener Zahlen kleiner gleich 2 war.

HINWEIS: Stellen Sie keine $20 \times 20-$Tafel auf...

b) Sie werfen drei Ikosaeder und notieren sich das Produkt der Zahlen. Führen Sie eine Zufallsvariable X ein, die diese Produkte modelliert und geben Sie in Mengenschreibweise den Ergebnisraum an, in den X abbildet (diesen schreiben Sie NICHT explizit auf!). Berechnen Sie die Wahrscheinlichkeit $P(X \in \{1, 4\})$.

c) Ihr/e Übungspartner/in hat zwei Ikosaeder geworfen und teilt Ihnen mit, dass das Ergebnis jedes einzelnen der beiden kleiner als 13 war. Wie hoch ist mit diesem Vorwissen nun die Wahrscheinlichkeit, dass die Summe der beiden Zahlen 21 war?

d) Ihr/e Übungspartner/in hat zwei Ikosaeder geworfen und teilt Ihnen diesmal mit, dass die Summe größer als 37 ist. Sei Y die Zufallsvariable, die den Logarithmus des Produkts der Zahlen modelliert. Berechnen Sie den Erwartungswert von Y, bedingt auf die gegebene Vorinformation.

Aufgabe 6 Stetige Verteilungen **(1+2+4+2+4+4=9+4+4 Punkte)**

Gegeben sei folgende Funktion:

$$f_{XY}(x,y) = c(4 + 4x^2 - y - x^2 y) I_{[0,1]}(x) I_{[0,2]}(y).$$

a) Bestimmen Sie c, sodass die Funktion die Dichte eines Zufallsvektors (X, Y) ist.

HINWEIS: Prüfen Sie dazu, ob

$$f_{XY} \geq 0, \quad \int f_{X,Y}(x,y)\, dx\, dy = 1.$$

b) Bestimmen Sie die Randdichten f_X, f_Y.

c) Berechnen Sie $\mathrm{Var}(X), \mathrm{Var}(Y)$.

d) Sind X und Y unkorreliert?

e) **(M)** Berechnen Sie $\mathbb{E}[e^Y]$.

f) **(I)** Beschreiben Sie, wie Sie mit Hilfe einer Software vorgehen würden, um $\mathbb{E}[e^Y]$ anzunähern. Nehmen Sie an, dass Sie $U([0,1])-$verteilte Zufallszahlen erzeugen können.

Aufgabe 7 Konvergenz, Grenzwertsätze **(4+2+3=5+4 Punkte)**

a) **(M)** Betrachten Sie den zentralen Grenzwertsatz von Lindeberg-Lévy: X_i u.i.v. mit $\mathbb{E}[X_1] = \mu$, $\mathrm{Var}(X_1) = \sigma^2$. Dann gilt

$$\sqrt(n)(\bar{X}_n - \mu)/\sigma \xrightarrow{w} \mathcal{N}(0,1).$$

Leiten Sie daraus den zentralen Grenzwertsatz von de-Moivre-Laplace ab, nämlich dass für $Y_n \sim \mathrm{Bin}(n,p)$ gilt, dass

$$\sqrt{n}(Y_n/n - p) \xrightarrow{w} \mathcal{N}(0, p(1-p)).$$

In welchem Sinn konvergiert dabei die Verteilung der linken Seite der Konvergenzaussage gegen die der rechte Seite?

b) Spezifizieren Sie eine statistische Anwendung des zentralen Grenzwertsatzes von de-Moivre-Laplace beim Schätzen oder Testen.

c) Seien $X_1, ..., X_n$ unabhängig identisch $\mathcal{N}(6, \sigma^2)-$verteilt mit $\sigma^2 > 0$. Wie groß muss n sein, damit die Wahrscheinlichkeit, dass der Abstand von $\bar{X}_n$ und $\mathrm{E}[\bar{X}_n]$ weniger als 4 Prozent von $\mathrm{E}[\bar{X}_n]$ ausmacht, mindestens 95 Prozent ist?

Aufgabe 8 Schätzen **(4+3+3=7+3 Punkte)**

a) Sie haben eine Stichprobe $x = (x_1, ..., x_m)$, deren Komponenten Sie alle i.id. als $\Gamma(n, \lambda)-$verteilt annehmen. Sie kennen n, aber nicht λ. Bestimmen Sie den Maximum-Likelihood-Schätzer für λ. In welchem Sinne ist dieser konsistent?

HINWEIS: Beachten Sie die Hinweise auf Seite ....

b) Sei haben Beobachtungen $x = (x_1, ..., x_{101})$ erhoben, deren Komponenten alle i.id. einer unbekannten Verteilung mit existierendem Erwartungswert und existierender, positiver Varianz, folgen. Sie haben folgende Größen berechnet:

$$\frac{1}{101}\sum_{i=1}^{101} x_i = 3.5, \quad \frac{1}{100}\sum_{i=1}^{101}(x_i - 3.5)^2 = 4.84.$$

Bestimmen Sie ein asymptotisches, zweiseitiges Konfidenzintervall für den Erwartungswert für $\alpha = 0.1$.

c) **(I)** Beschreiben Sie, wie man mit Hilfe des statistischen Bootstraps ein empirisches $95\%-$Konfidenzintervall für einen Punktschätzer bestimmen kann.

Aufgabe 9 Hypothesentests **(2+2+4+4+4=16 Punkte)**

a) Was versteht man im Kontext von statistischen Tests unter dem Fehler 1. und 2. Art?

b) Was ist die Macht eines Tests? Beschreiben Sie anschaulich, warum eine große Macht wünschenswert ist.

c) Jemand behauptet, in einer Klausur, in der man insgesamt 150 Punkte bekommen kann, würden im Mittel 60 erreicht, mit einer Standardabweichung von 12. Sie glauben den Mittelwert nicht (nehmen aber die Standardabweichung als richtig an) und erheben selbst Daten. Der Mittelwert Ihrer 191 Beobachtungen ist 57.5. Erstellen Sie mit Hilfe eines Grenzwertsatzes einen asymptotischen Test und treffen Sie die Testentscheidung anhand Ihrer Daten.

d) Sie haben Beobachtungen $x_1, ..., x_{150}$, die alle unabhängig aus einer $\mathcal{N}(-2, \sigma^2)-$Verteilung stammen. Sie haben

$$\frac{1}{150}\sum_{i=1}^{150}(x_i + 2)^2 = 3.93$$

berechnet. Entscheiden Sie für $\alpha = 0.01$, ob Sie die Nullhypothese, dass die wahre Standardabweichung 1.7 ist, ablehnen können.

e) Bestimmen Sie für Ihren Test aus d) die Macht. Wie groß ist diese für $\theta_1 := 1.8$?

Aufgabe 10 Kontingenztabellen **(2+2+6=10 Punkte)**

Sie interessiert die Frage, ob Personaler sich von einem Foto beeinflussen lassen, d.h., ob die Reaktion auf die Bewerbung von der Art des Fotos bzw. dessen Vorhandenseins abhängt. Dazu schicken drei Personen mit sehr ähnlicher (guter) Qualifikation und sehr ähnlichen Lebensläufen jeweils Bewerbungen für jeweils die gleiche Stelle ab. Jedes Mal hat eine zufällig ausgewählte Person kein Foto angehängt, eine andere hat ein professionelles Bewerbungsfoto und die dritte ein Passbild. Wir nehmen für diese Aufgabe an, dass die marginalen Unterschiede, die nicht das Foto betreffen, keinen Einfluss auf die Entscheidung des Personalers haben.

Dieses Vorgehen wurde für 120 Stellen durchgeführt. Von den Bewerbungen ohne Foto führten 54 zu einer positiven Reaktion, was bei den Bewerbungen mit Passbild nur 30 Mal der Fall war. Bei den übrigen Bewerbungen fiel die Reaktion in zwei Dritteln der Fälle positiv aus. Eine Reaktion gab es immer, d.h., die Alternative zu einer positiven Reaktion ist lediglich die Absage.

a) Stellen Sie aus den obigen Informationen die Kontingenztabelle auf, einmal für die absoluten, einmal für die relativen Häufigkeiten.

b) Berechnen Sie die Tabellen der bedingten Häufigkeiten.

c) Beantworten Sie durch einen geeigneten statistischen Test, ob Foto und Reaktion abhängig sind. Es sei $\alpha = 0.01$.

Aufgabe 11 Lineare Regression **(2+[1+1+1+1+1]=7 Punkte)**

a) Stellen Sie das multiple lineare Regressionsmodell auf (für eine allgemeine Anzahl p an erklärenden Variablen). Schreiben Sie das Optimierungsproblem auf, welches durch den Regressionsschätzer gelöst wird.

b) Betrachten Sie folgenden R−Output:

```
Call: lm(formula = Y ~ ., data = D)

Residuals:      Min       1Q   Median       3Q      Max
              -2.7240  -0.9957  -0.0331   1.0083   3.3947

Coefficients: Estimate      Std.      Error     t value      Pr(> |t|)
(Intercept)   -0.0986       0.2452    -0.402     0.6896
X1             1.3088       0.2508     5.219     4.67e-06     ***
X2            -0.5275       0.2769    -1.905     0.0633       .
X3             1.8363       0.2483     7.395     3.02e-09     ***
X4             1.2781       0.2766     4.621     3.34e-05     ***
X5             0.2833       0.2127     1.332     0.1897
— Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 1.574 on 44 degrees of freedom
Multiple R-squared: 0.764, Adjusted R-squared: 0.7371
F-statistic: 28.48 on 5 and 44 DF, p-value: 9.228e-13
```

(a) Welche Koeffizienten sind signifikant? Begründung?

(b) Was bedeutet anschaulich der Koeffizient von X2?

(c) Zeigen Sie anhand der anderen vorliegenden Informationen, wie Sie auf den t-Wert für X4 kommen. Welche Hypothese steckt dahinter?

(d) Wie bekommen Sie für die Variable X3 den angegebenen p-Wert?

(e) Begründen Sie die Anzahl der Freiheitsgrade.

Aufgabe 12 Generalisierte lineare Modelle (2+3+1+1=7 Punkte)

a) Definieren Sie ein allgemeines generalisiertes lineares Modell. Was versteht man unter einer Linkfunktion? Wie lautet diese für das gewöhnliche lineare Regressionsmodell (Aufgabe 11)?

b) Nehmen Sie an, Ihre abhängige Variable Y nimmt die Werte ,,grün" und ,,rot" an. Sie haben p erklärende Variablen $X_1, ..., X_p, p \geq 2$, zur Verfügung. Stellen Sie für diesen konkreten Fall Ihr Logit-Modell mit allen erklärenden Variablen auf.

c) Sie haben nun die Koeffizienten für Ihr Modell aus b) vorliegen. Der Koeffizient β_2 von X_2 ist 0.663. Wie interpretieren Sie diesen Koeffizienten anschaulich?

d) Angenommen, für drei neue Beobachtungen (d.h., alle X-Werte sind bekannt, der Y-Wert jeweils unbekannt) sind die berechneten Log-Odds 0.36, 0.71 und 0.88. Für welche der Beobachtungen sagen Sie rot bzw. grün voraus?