Lehrstuhl für Statistik und Ökonometrie
der Otto-Friedrich-Universität Bamberg
Prof. Dr. Susanne Rässler

PRÜFUNG: METHODEN DER STATISTIK I
Sommersemester 2018

Aufgabe 1

Das Kraftfahrt-Bundesamt hat 50 Dieselfahrzeuge von drei Automobilherstellern (Merkmal A) hinsichtlich überhöhter Werte beim CO2-Ausstoß (w_1: nicht überhöht, w_2: überhöht) überprüft und die Daten in einer Kontingenztabelle erfasst:

W \ A	a_1	a_2	a_3
w_1	3	4	8
w_2	7	26	2

a) Erstellen Sie eine Häufigkeitstabelle (inklusive Randhäufigkeiten!) der relativen Häufigkeiten.

b) Berechnen Sie:

   b1) $n(A_2)$

   b2) $f_{1\bullet}$

   b3) f_{12}

   b4) $f(a_3|w_2)$

c) Ermitteln Sie folgende Werte *unter Verwendung der korrekten Notation*:

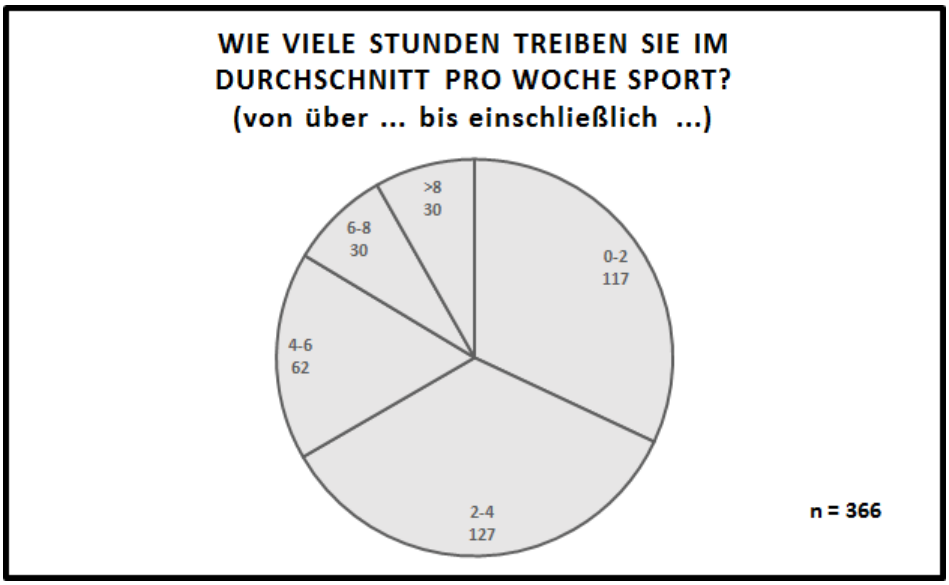
   c1) Wie viel Prozent der überhöhten Werte entfielen auf Automobilhersteller a_2?

   c2) Wie groß ist der Anteil an Dieselfahrzeugen, die keine überhöhten Werte hatten und von Automobilhersteller a_1 gebaut wurden?

d) Berechnen Sie ein geeignetes Maß für den Zusammenhang der beiden Merkmale W und A und interpretieren Sie das Ergebnis.

Aufgabe 2

Gerade im Sommer bietet Bamberg viele Möglichkeiten, sich sportlich zu betätigen. Ob joggen im Hain, Schwimmen im Stadionbad oder Volleyball-Spielen auf der Jahnwiese. Auch die Teilnehmer des Kurses Methoden der Statistik I zeigen Freude an sportlichen Aktivitäten, wie die Ergebnisse der Befragung zu Semesterbeginn ergaben. Der Student mit der maximalen Anzahl an Stunden treibt 15 Stunden Sport pro Woche. Im Folgenden sollen Sie das Sportverhalten von Ihnen und Ihren Kommilitonen nun genauer betrachten. Dazu liegt folgende Grafik vor:



a) Um welchen Merkmalstyp handelt es sich beim vorliegenden Merkmal X 'Durchschnittliche Anzahl Stunden Sport pro Woche'?

b) Berechnen Sie das arithemtische Mittel von X für die Teilnehmer des Statistik I-Kurses und interpretieren Sie das Ergebnis. Legen Sie außerdem zur Berechnung eine geeignete Arbeitstabelle an, die Sie für die nachfolgenden Teilaufgaben noch erweitern können.

c) Berechnen Sie die Häufigkeitsdichte für alle Klassen (erweitern Sie die Arbeitstabelle aus b) entsprechend). Wie würden Sie die Häufigkeitsdichte geeignet grafisch darstellen (keine Zeichnung nötig)?

d) Vervollständigen Sie die in b) begonnene Arbeitstabelle und zeichnen Sie die approximierende Verteilungsfunktion. Welche Annahme liegt hierfür zu Grunde?

e) Berechnen und interpretieren Sie den Wert für die approximierende Verteilungsfunktion an der Stelle $x = 3$.

Aufgabe 3

Student Max interessiert sich für das Absatzvolumen der Bierbrauereien in Deutschland. Er vermutet, dass in Deutschland nur wenige Großbrauereien für den Großteil des Absatzvolumens verantwortlich sind, während es besonders viele Mikrobrauereien mit geringen Absatzmengen gibt. Aufgrund seiner Vermutung recherchiert er die folgenden Informationen für das Jahr 2017.

Absatzvolumen (in Tausend Hektoliter) von über ... bis einschließlich ...	**Anzahl Brauereien**
0 - 5	1065
5 - 50	273
50 - 200	86
200 - 1000	42
1000 - 2000	26

a) Berechnen Sie das einfache und das normierte Gini-Maß für die Konzentration des Absatzvolumens im Jahr 2017. Fertigen Sie hierzu eine geeignete Arbeitstabelle an.

b) Erklären Sie kurz (1-2 Sätze), warum das einfache und das normierte Gini-Maß unter Verwendung dieser Daten sehr ähnliche Werte annehmen.

c) Max denkt darüber nach, das Absatzvolumen statt in Tausend Hektoliter in Liter zu messen. Welche Auswirkungen hätte dies auf den Wert des normierten Gini-Maßes?

A: Das Gini-Maß steigt. B: Das Gini-Maß bleibt gleich. C: Das Gini-Maß sinkt.

Hinweis: Beantworten Sie diese Frage bitte auf Ihrem Bearbeitungsbogen.

Max interessiert sich zudem für die Anzahl der aktiven Brauereien in den drei Regierungsbezirken Frankens. Die nachstehende Tabelle zeigt die Ergebnisse seiner diesbezüglichen Recherche.

Regierungsbezirk	**Anzahl Brauereien**
Oberfranken	166
Mittelfranken	70
Unterfranken	57

d) Zeichnen Sie die Lorenzkurve für die Konzentration der Anzahl der Brauereien in Franken.

e) Skizzieren Sie die Lorenzkurve für den Fall einer maximalen Konzentration der Anzahl der Brauereien in den drei Regierungsbezirken.

f) Wie müssten sich die 293 Brauereien Frankens auf die drei Regierungsbezirke verteilen, damit das zugehörige Gini-Maß den kleinstmöglichen Wert annimmt?

g) Max erfährt, dass zehn der Brauereien aus Unterfranken Konkurs angemeldet haben. Welche Auswirkung hat dies auf die Konzentration der Anzahl der Brauereien?

A: Konzentration steigt. B: Konzentration bleibt gleich. C: Konzentration sinkt.

Hinweis: Beantworten Sie diese Frage bitte auf Ihrem Bearbeitungsbogen.

Aufgabe 4

Für die Fußball-WM-Kader von fünf Ländern liegen Informationen zum durchschnittlichen Alter der Spieler (X) und der durchschnittlichen Anzahl an Länderspielen (Y) vor:

Land	∅ Alter	∅ Länderspiele
Brasilien	28,6	30
England	26	20
Mexiko	29,4	62
Nigeria	26	24
Spanien	28,4	40

a) Berechnen Sie für beide Merkmale X und Y das arithmetische Mittel.

b) Übertragen Sie die Daten in ein Streudiagramm (wählen Sie geeignete Startpunkte für die Achsen) und zeichnen Sie für beide Merkmale auch das arithmetische Mittel ein.

c) Begründen Sie kurz, warum Sie anhand des Streudiagramms davon ausgehen, dass zwischen den beiden Merkmalen ein positiver (linearer) Zusammenhang besteht.

Wir unterstellen einen linearen Zusammenhang der Form $y_\nu = \beta_0 + \beta_1 x_\nu + u_\nu$, mit dem der 'Einfluss' des Durchschnittsalters auf die durchschnittliche Anzahl an Länderspielen untersucht werden soll.

d) Ermitteln Sie mit Hilfe der KQ-Methode die Schätzer für β_0 und β_1. Verwenden Sie optional hierzu folgende Information: $\overline{x^2} = 768,176$ und $\overline{xy} = 992,16$
(*Ersatzergebnis: Falls Sie Teilaufgabe d) nicht lösen konnten, verwenden Sie im Folgenden* $\hat{\beta}_0 = -221,82$ *und* $\hat{\beta}_1 = 8,49$)

e) Warum ist der Wert für $\hat{\beta}_0$ nicht plausibel?

f) Prognostizieren Sie die durchschnittliche Länderspielanzahl für einen WM-Kader, der durchschnittlich 30 Jahre alt ist.

Lösung zu Aufgabe 1

a)

W \ A	a_1	a_2	a_3	$\sum$
w_1	0,06	0,08	0,16	0,3
w_2	0,14	0,52	0,04	0,7
$\sum$	0,2	0,6	0,2	1

b) b1) $n(A_2) = 30$

b2) $f_{1\bullet} = 0,3$

b3) $f_{12} = 0,08$

b4) $f(a_3|w_2) = \frac{0,04}{0,7} = 0,0571$

c) c1) $f(a_2|w_2) = \frac{26}{35} = 0,7429 = 74,29\%$

c2) $f(w_1, a_1) = 0,06$

d)

W \ A	a_1	a_2	a_3
w_1	0.06	0.18	0.06
w_2	0,14	0,42	0,14

Berechnung ϕ_{WA}

$$\phi^2_{WA} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(f_{ij} - f^*_{ij})^2}{f^*_{ij}}$$

$$\Rightarrow \phi^2_{WA} = 0,3175$$

Normierung/ Interpretation

$$0 \leq V_{WA} = \sqrt{\frac{\phi^2_{WA}}{\min\{(k-1),(l-1)\}}} \leq 1$$

$$\Rightarrow V^2_{WA} = \sqrt{0,3175/1} = 0,5635$$

Mittelstarker Zusammenhang zwischen den beiden Merkmalen.

Lösung zu Aufgabe 2

a) Merkmalstyp: quantitativ/metrisch

b)

i	$(\tilde{x}_{i-1}; \tilde{x}_i]$	x_i	n_i	f_i	Δx_i	f_i^*	$F(\tilde{x}_i)$
1	0 - 2	1	117	0,3197	2	0,1599	0,3197
2	2 - 4	3	127	0,3470	2	0,1735	0,6667
3	4 - 6	5	62	0,1694	2	0,0847	0,8361
4	6 - 8	7	30	0,0820	2	0,0410	0,9181
5	8 - 15	11,5	30	0,0820	7	0,0117	1,0001
Σ	-	-	366	1,0001	-	-	-

Klassenobergrenze ($\tilde{x}_k = 15$) letzte Klasse
Arithmetisches Mittel für klassierte Daten:

$$\begin{aligned} \bar{x} &= \sum_{i=1}^{k} x_i f_i \\ &= 1 \cdot 0{,}3197 + 3 \cdot 0{,}3470 + 5 \cdot 0{,}1694 + 7 \cdot 0{,}0820 + 11{,}5 \cdot 0{,}0820 \\ &= 3{,}72475 \end{aligned}$$

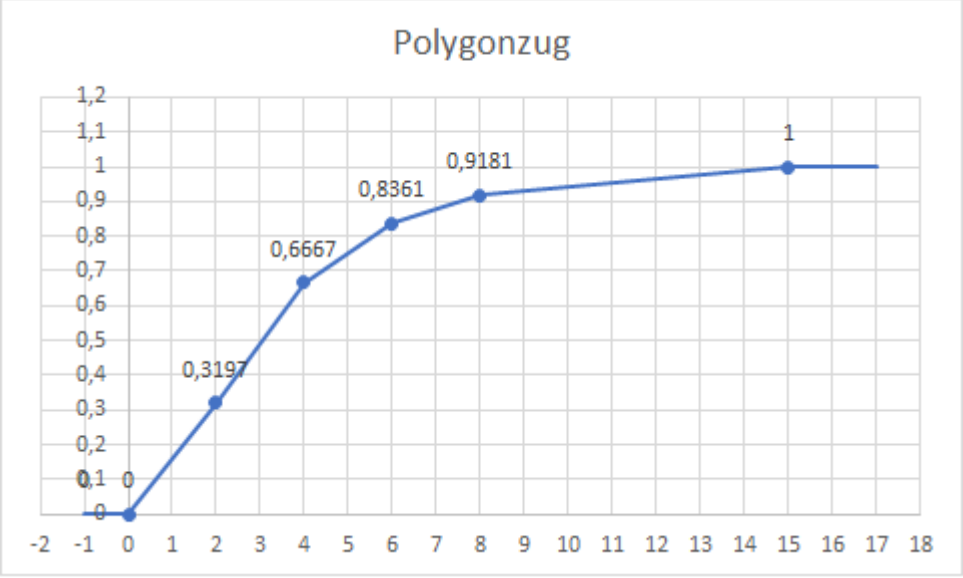
Approximativ treiben die Studierenden im Mittel 3,8846 Stunden Sport pro Woche.

c) Häufigkeitsdichte s. Tabelle

$$f^*(x) = \frac{f_i}{\Delta x_i}$$

Grafische Darstellung: Histogramm

d) Verteilungsfunktion für das Merkmal X $F(\tilde{x}_i)$ s. Tabelle
Annahme: Gleichverteilung in den Klassen/ Polygonzug



e) Approximierende Verteilungsfunktion an der Stelle x = 3:

$$\begin{aligned} F^*(x) &= F(\tilde{x}_{i-1}) + \frac{f_i}{\Delta x_i}(x - \tilde{x}_{i-1}) \\ F^*(3) &= 0{,}3197 + 0{,}1735 \cdot (3 - 2) \\ &= 0,4932, \end{aligned}$$

d.h. approximativ treiben 49,32 % der Teilnehmer des Statistik I-Kurses bis zu 3 Stunden Sport pro Woche.

Lösung zu Aufgabe 3

a) Einfaches und normiertes Gini-Maß
Arbeitstabelle:

i	Absatz	x_i	n_i	$x_i n_i$	f_i	F_i	$g_i = \frac{x_i n_i}{S^*}$	G_i
1	0-5	2.5	1065	2662.5	0.7138	0.7138	0.0313	0.0313
2	5-50	27.5	273	7507.5	0.1830	0.8968	0.0882	0.1195
3	50-200	125	86	10750	0.0576	0.9544	0.1263	0.2458
4	200-1000	600	42	25200	0.0282	0.9826	0.2961	0.5419
5	1000-2000	1500	26	39000	0.0174	1	0.4582	1
			1492	85120	1		1	

Einfaches Gini-Maß:

$$M_G \doteq \frac{K}{1/2} = 2 \cdot K = 1 - \sum_{i=1}^{k} (G_{i-1} + G_i) f_i$$

$$\begin{aligned} M_G &\doteq 1 - [(0 + 0{,}0313) \cdot 0{,}7138 + (0{,}0313 + 0{,}1195) \cdot 0{,}1830 \\ &+ (0{,}1195 + 0{,}2458) \cdot 0{,}0576 + (0{,}2458 + 0{,}5419) \cdot 0{,}0282 \\ &+ (0{,}5419 + 1) \cdot 0{,}0174] = 1 - 0.1200 = 0.88 \end{aligned}$$

Normiertes Gini-Maß:

$$M_G = \frac{1 - \sum_{i=1}^{k} (G_{-1} + G_i) f_i}{1 - f_k} = \frac{1 - 0{,}1200}{1 - 0{,}0174} = 0.8956$$

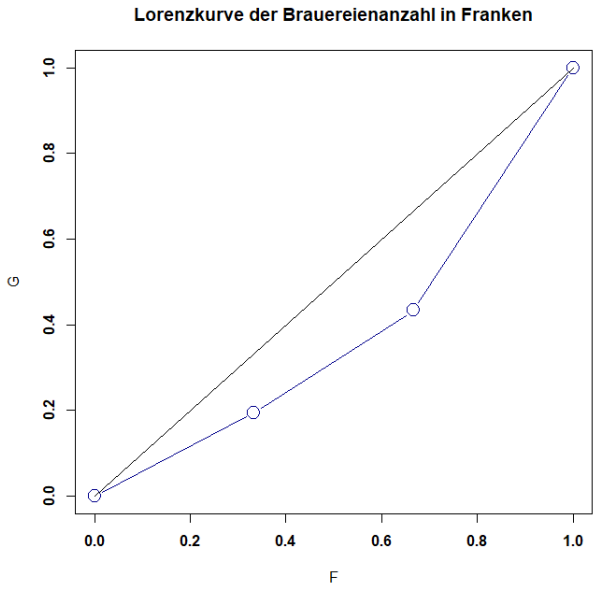
b) Exaktes vs. normiertes Gini-Maß
Das einfache und das exakte Gini-Maß nehmen unter Verwendung dieser Daten sehr ähnliche Werte an, da hier eine sehr hohe Anzahl an Beobachtungen ($n =$ 1492 Brauereien) vorliegt.
Bei der Berechnung des einfachen Gini-Maßes wird als Approximation für die Fläche bei maximaler Konzentration der Wert 0.5 angenommen. Die Fläche bei maximaler Konzentration nähert sich mit steigendem n dem Wert 0.5 an, damit nähern sich auch das einfache und das exakte Gini-Maß an.

c) B: Das Gini-Maß bleibt gleich.

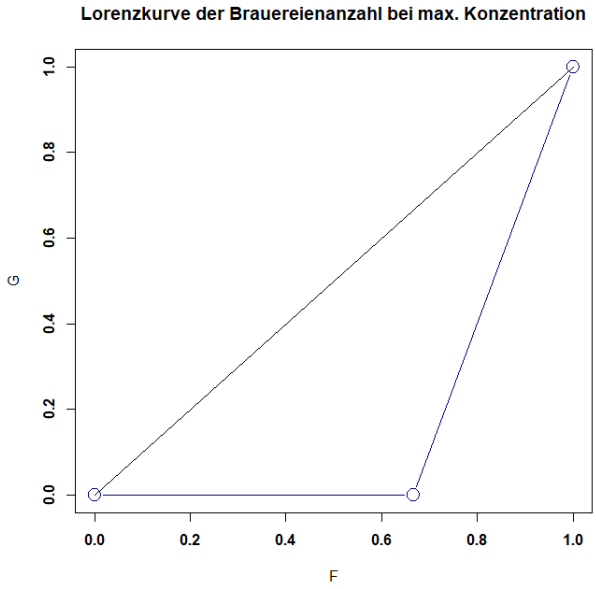
d) Zeichnen der Lorenzkurve

Regierungsbezirk	ν	x_ν	F_ν	$g_\nu = \frac{x_{[\nu]}}{S^*}$	G_ν
Unterfranken	1	57	0,3333	0.1945	0.1945
Mittelfranken	2	70	0,6667	0.2389	0.4334
Oberfranken	3	166	1	0.5666	1
		$S^* = 293$			

Skizze:



e) Lorenzkurve für den Fall einer max. Konzentration



f) kleinstmögliche Konzentration
Damit das zugehörige Gini-Maß den kleinstmöglichen Wert annimmt, müssten sich die 293 Brauereien in Franken möglichst gleichmäßig auf die drei Regierungsbezirke verteilen. Dies könnte z.B. so aussehen:
Oberfranken: 98 Brauereien
Mittelfranken: 98 Brauereien
Unterfranken: 97 Brauereien

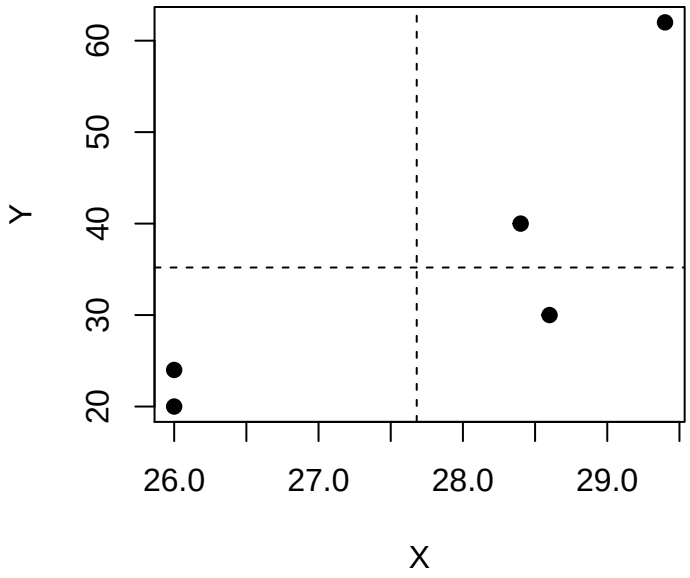
g) A: Konzentration steigt

Lösung zu Aufgabe 4

a) X: ∅ Alter
Y: ∅ Länderspiele

$\bar{x} = 27,68$
$\bar{y} = 35,2$

b)



c) Vier von fünf Beobachtungen sind in den Quadranten links unten (Werte für X und Y kleiner als MW) oder rechts oben (Werte für X und Y größer als MW) zu finden, die positiv auf die Kovarianz 'einzahlen'.

d)

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \qquad \text{und} \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\Rightarrow \widehat{\beta}_1 = \frac{992,16 - 27,68 \cdot 35,2}{768,176 - 27,68^2} = 8,9406 \quad \text{und} \quad \widehat{\beta}_0 = 35,2 - 8,9401 \cdot 27,68 = -212,2758$$

e) Der Achsenabschnittsparameter gibt an, wie viele Länderspiele bei einem Durchschnittsalter von 0 Jahren zu erwarten sind. Eine negative Anzahl an Länderspielen ist nicht möglich.

f) $\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x \Rightarrow \widehat{y}(30) = -212,2758 + 8,9406 \cdot 30 = 55,9422$
mit Ersatzergebnis:
$\Rightarrow \widehat{y}(30) = -221,82 + 8,49 \cdot 30 = 32,88$