Statistik I **Übungsblatt Ferien**

für Studierende der Soziologie, des Nebenfachs Statistik, der Medieninformatik und der Cultural and Cognitive Linguistics

(Augustin, Schollmeyer, Brandt, Endres, Plaß) **Wintersemester 2016/17**

Aufgabe 1

Hier sehen Sie zehn zufällig ausgewählte Beobachtungen aus dem ALLBUS 2008, bereitgestellt von GESIS (Leibnitz-Institut für Sozialwissenschaften). In der Tabelle sind folgende Merkmale dargestellt:

- Geschlecht des Befragten (Geschlecht)
- Wohnort in Ost- oder Westdeutschland (Ostwest)
- Fernsehkonsum in Minuten (Fernsehen)
- Einkommen in Euro (Eink.)
- Gewicht in kg (Gewicht)
- Größe des Befragten in cm (Groesse)
- Body-Mass-Index (BMI)

Füllen Sie als Wiederholung die Tabelle auf der nächsten Seite aus.

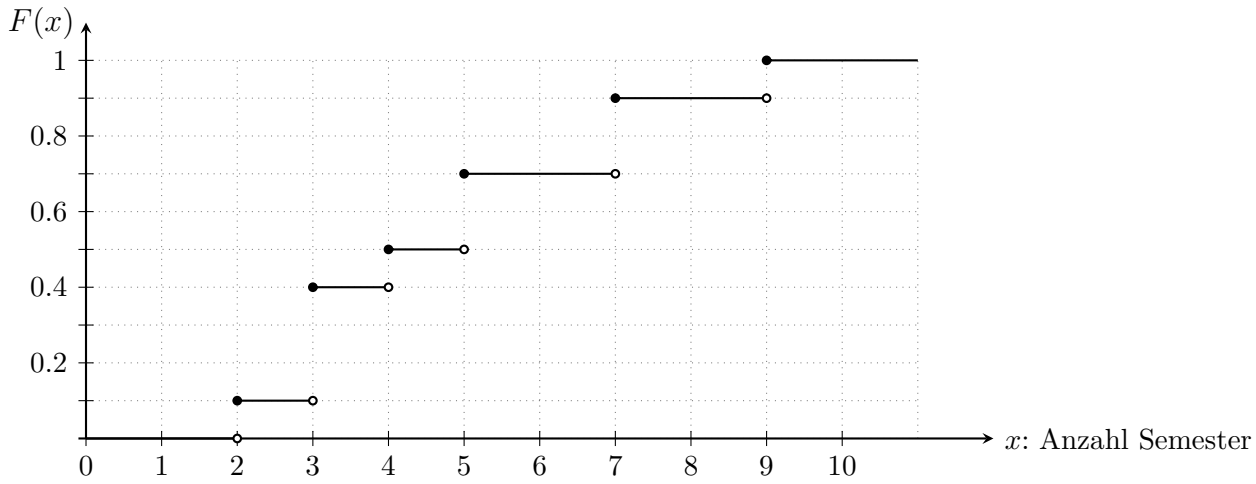
- Berechnen Sie nur diejenigen Größen, die sinnvoll sind.
- Verwenden Sie für die Berechnung von $\tilde{s}^2$, $\tilde{s}^2_{\text{zw}}$ und $\tilde{s}^2_{\text{in}}$ Geschlecht als Schichtungsvariable.
- Berechnen Sie das arithmetische Mittel direkt und über die Formel für geschichtete Daten. Welche Variable kann neben dem Geschlecht noch als Schichtungsvariable verwendet werden?

Geschlecht	Ostwest	Fernsehen	Alter	Eink.	Gewicht	Groesse	BMI
Frau	West	60	43	860	67	164	24.9
Frau	West	180	67	1500	65	164	24.2
Frau	West	240	20	250	55	170	19.0
Frau	Ost	360	51	1200	85	151	37.3
Mann	Ost	240	56	1300	83	171	28.4
Mann	West	240	82	1000	92	168	32.6
Frau	West	150	27	345	60	164	22.3
Frau	West	120	26	700	55	175	18.0
Mann	Ost	270	52	308	100	176	32.3
Mann	West	240	70	1000	80	170	27.7

Variable	Skala	Modus	Median	arith-metisches Mittel	empirische Varianz ($\tilde{s}^2$)	Varianz zwischen ($\tilde{s}^2_{\text{zw}}$)	Varianz innerhalb ($\tilde{s}^2_{\text{in}}$)	Anteil erklärter Varianz	Standard-abweichung
Geschlecht									
Ostwest									
Fernsehen									
Alter									
Einkommen									
Gewicht									
Groesse									
BMI									

Aufgabe 2

In einer Studenten-WG mit 10 Mitbewohnern ergab sich für die Frage nach der *Anzahl der bereits studierten Semester* folgende empirische Verteilungsfunktion:



a) Bestimmen Sie die durchschnittliche *Anzahl studierter Semester* in der WG.

Hinweis: Lesen Sie die Merkmalsausprägungen a_j und die zugehörigen Häufigkeiten h_j in der empirischen Verteilungsfunktion ab.

b) Zeichnen Sie den Boxplot für das Merkmal *Anzahl studierter Semester*.

Hinweis: Entweder Sie lesen die benötigten Quantile aus der empirischen Verteilungsfunktion ab, oder Sie erstellen Sich anhand der empirischen Verteilungsfunktion und der Angabe $n = 10$ eine Urliste der Daten.

Aufgabe 3

Sind die folgenden Aussagen richtig?

a) Die absolute kumulierte Häufigkeitsverteilung beschreibt die Verteilung der Daten vollständig.

b) Der Mittelwert und die Varianz einer Verteilung beschreiben diese vollständig.

c) Aus der empirischen Verteilungsfunktion ist der Mittelwert ableitbar.

d) Aus der Lorenzkurve ist der Gini-Koeffizient ableitbar.

e) Aus der Lorenzkurve ist die empirische Verteilungsfunktion ableitbar.

f) Der Interquartilsabstand ist ein Lagemaß.

g) Für die Varianzzerlegung gilt, dass die Gesamtvarianz immer größer oder gleich der Varianz innerhalb der Schichten ist.

h) Für die Varianzzerlegung gilt, dass die Gesamtvarianz immer größer oder gleich der Varianz zwischen den Schichten ist.

i) Aus dem Boxplot ist das 0%-Quantil ablesbar.

j) Aus der empirischen Verteilungsfunktion ist die Varianz ableitbar.

k) Aus der empirischen Verteilungsfunktion kann man die Stichprobengröße n ermitteln.

l) Aus der empirischen Verteilungsfunktion kann man eine untere Schranke für die Stichprobengrößen ableiten. **raus?**

m) Der Median eines verhältnisskalierten Merkmals ist sinnvoll interpretierbar.

n) Der Mittelwert eines nur ordinalskalierten Merkmals ist sinnvoll interpretierbar.

o) Der Variationskoeffizient eines lediglich intervallskalierten Merkmals ist sinnvoll interpretierbar.

p) $\left(\sum_{i=1}^{n} a_i\right) \cdot \left(\sum_{i=1}^{n} b_i\right) = \sum_{i=1}^{n} a_i \cdot b_i.$

q) $\left(\sum_{i=1}^{n} a_i\right) \cdot \left(\sum_{j=1}^{n} b_j\right) = \sum_{i,j \in \{1,\ldots,n\}}^{n} a_i \cdot b_j.$

r) $\sum_{i=1}^{n} c \cdot a_i = c \cdot \sum_{i=1}^{n} a_i.$

s) $a \leq b \quad \Longrightarrow \quad a + c \leq b + c.$

t) $a \leq b \quad \Longrightarrow \quad a \cdot c \leq b \cdot c.$

u) Der Median der aus der Größe X abgeleiteten Größe $a \cdot X + b$ hat (, falls er eindeutig bestimmt ist,) den Wert $a \cdot Median(X) + b$.

v) Das 75% - Quantil der aus der Größe X abgeleiteten Größe $Y := a \cdot X + b$ hat (, falls es eindeutig bestimmt ist,) den Wert $a \cdot x_{0.75} + b$. **raus?** Stattdessen: Der Mittelwert der aus der Größe X abgeleiteten Größe $a \cdot X + b$ hat den Wert $a \cdot Median(X) + b$.

w) Das 75% - Quantil der aus der Größe X abgeleiteten Größe $Y := a \cdot X + b$ hat (, falls es eindeutig bestimmt ist,) für negatives a den Wert $a \cdot x_{0.25} + b$. **raus?**

x) Das Histogramm ist längentreu.

y) Die Summe aller kumulierten relativen Häufigkeiten ist 1.

z) Zwei verschiedene Lorenzkurven führen zu verschiedenen Gini-Koeffizienten.

Aufgabe 4

a) Betrachten Sie folgende 12 Beobachtungen:

Beobachtung	1	2	3	4	5	6	7	8	9	10	11	12
Ausprägung	13	10	5	32	21	9	4	11	1	19	8	6

(Sie können sich vorstellen, dass es sich um die Anzahl (in Tausend) von grünen Talern der 12 Marskönige handelt. Wir wollen die Verteilung der Taler dieser 12 Marskönige beschreiben.)

(i) Zeichnen Sie die Lorenzkurve.

(ii) Berechnen Sie den Gini-Koeffizienten und den normierten Gini-Koeffizienten.

(iii) Interpretieren Sie Ihre Ergebnisse.

b) Bestimmen Sie aus den Daten von a) die zugehörigen Quartilsdaten (das heißt $0 < \alpha_{0.25} < \alpha_{0.5} < \alpha_{0.75} < 1$, also $q = 4$).

(i) Zeichnen Sie die induzierte Lorenzkurve.

(ii) Berechnen Sie den induzierte Gini-Koeffizienten.

(iii) Vergleichen Sie Ihre Ergebnisse mit denen aus a) und interpretieren Sie diese.

c) Bestimmen Sie den Herfindahl-Index, sowie die Konzentrationsrate zum Grad 3. Was unterscheidet diese Maße vom Gini-Koeffizienten?

Aufgabe 5 (Zusatzaufgabe für Medieninformatikstudierende)

a) Lösen Sie **Aufgabe 1**, **Aufgabe 2**, und **Aufgabe 4** (soweit möglich) mit `R`.

b) Auf der Veranstaltungshomepage finden Sie die R-Skript-Datei ferien.R. Beschreiben Sie, was in den Teilaufgaben (i) bis (vi) ausgeführt wird. Interpretieren Sie beispielhaft jeweils einen der berechneten Werte in den Teilaufgaben (i), (ii), (iv), (v) und (vi).