

Klausur zur Vorlesung Statistik I für Bachelor I, SS 2009

Lösungsvorschlag

Aufgabe 1

(30 Punkte)

Um den Verkehr zu lenken, führten seit wenigen Jahren verschiedene Städte wie London, Stockholm oder San Francisco ein sog. Mobility Pricing ein.

Ziele dieser Maßnahmen können z.B. Verkehrsvermeidung oder eine räumliche bzw. zeitliche Verkehrsverlagerung (Entzerrung der Spitzenstunden) sein. Auch eine Verschiebung des Modal-Splits (hin zu ÖV) ist denkbar. Schließlich kann die Motivation rein finanzieller Natur sein (mehr Geld in die Stadtkasse). Die Gebühren können sowohl für den fließenden als auch für den ruhenden Verkehr (Parkgebühren) erhoben werden, und zwar für PKW, LKW, Taxis, ÖV usw. Ferner hängt die Höhe der Gebühren i.A. von Nutzungszeit und -raum, der Schadstoffklasse, der Nutzergruppe und des Besetzungsgrads ab.

Die Erhebung der Gebühren kann manuell oder automatisch (Durchfahren von Kontrollpunkten mit Nummernschilderkennung oder GPS gestützt) erfolgen. Je nach System sind unterschiedlich viele Teilnehmer betroffen. Die Gesamteinnahmen können den Verkehrssystemen zugute kommen oder aber nicht zweckgebunden sein.

- (a) *Geben Sie die statistische Gesamtheit und den Merkmalsträger an. Grenzen Sie die statistische Gesamtheit ab.*

Gesamtheit: Alle Städte mit Mobility Pricing; Statistische Einheit bzw. Merkmalsträger: Eine solche Stadt;

Abgrenzung:

- räumlich: Auf der ganzen Erde
- zeitlich: “Seit einigen Jahren”
- sachlich: Keine besondere Einschränkung

- (b) *Geben Sie mindestens drei häufbare Merkmale und die entsprechenden häufbaren Merkmalsausprägungen an.*

Häufbar: Bei nominalskalierten Merkmalen können mehrere Merkmalsausprägungen gleichzeitig zutreffen. Hier also z.B.

- Merkmal “Ziel”; Ausprägungen Verkehrsvermeidung, Verkehrsverlagerung, ...
- Merkmal “Art der Erhebung”; Ausprägungen manuell, automatisch (auch Toll-Collect nutzt beide Arten der Erhebung)
- Merkmal “betroffene Nutzergruppen”; Ausprägungen PKW, LKW, Taxis, ÖV usw.

- (c) *Begründen Sie, warum häufbare Merkmale immer nominalskaliert sein müssen*

Bei den anderen Kategorien ordinalskaliert und kardinalskaliert gibt es bei den möglichen Ausprägungen eine natürliche Reihenfolge, welche durch die Vergleichsoperatoren “>” bzw. “<” definiert wird. Zwei unterschiedliche Ausprägungen sind durch unterschiedliche Werte $a_1 > a_2$ bzw. $a_1 < a_2$, also auf jeden Fall $a_1 \neq a_2$ definiert.

Es können also nicht beide Werte gleichzeitig zutreffen und ordinal- sowie kardinalskalierte Merkmale sind nie häufbar.

(d) *Geben Sie mindestens drei Bestandsmassen und eine Bewegungsmasse an.*

Bestandsmassen: Ausmaß der Verkehrsvermeidung bzw. -verlagerung; Zahl der betroffenen Teilnehmer; Höhe der Gebühren.

Bewegungsmassen: Eingenommene Gelder, Reduktion der gesamten CO₂-Emissionen.

(e) *Geben Sie je ein dichotomes (binäres) häufbares und ein dichotomes nichthäufbares Merkmal an.*

Dichotom, häufbar: Betroffene Verkehrskategorien mit den Ausprägungen fließender und ruhender Verkehr; Erhebungsart mit den Ausprägungen manuell und automatisch.

Dichotom, nichthäufbar: Zweckgebundenheit der Einnahmen: Entweder ist diese gegeben oder aber nicht.

(f) *Geben Sie schließlich noch an:*

- *ein absolutskaliertes Merkmal:* Zahl der betroffenen Teilnehmer
- *drei weitere kardinalskalierte Merkmale:* räumliche Verkehrsverlagerung, zeitliche Verkehrsverlagerung, Modal-Split, Einnahmen, Höhe der Gebühren, ...
- *ein ordinalskaliertes Merkmal:* Schadstoffklasse

Aufgabe 2

(55 Punkte)

Im Rahmen einer Mobilitätsbefragung (Umfang: 1000 Personen) wurden für einen Stichtag die Reiseweiten aller durchgeführten Wege erfasst:

Reiseweitenklasse (km)	0-1	1-2	2-3	3-5	5-10	10-25	25-50
Häufigkeiten	32	165	247	488	770	1365	670

- (a) Wieviel Wege wurden insgesamt erfasst? Welcher Mobilitätskennziffer (Wege pro Person und pro Tag) entspricht dies?

Bezeichnet man die absoluten Häufigkeiten einer Reiseweitenklasse k mit h_k , gilt

- Gesamtzahl der Wege:

$$n = \sum_k h_k = \underline{\underline{3737}}$$

- Mobilitätskennziffer:

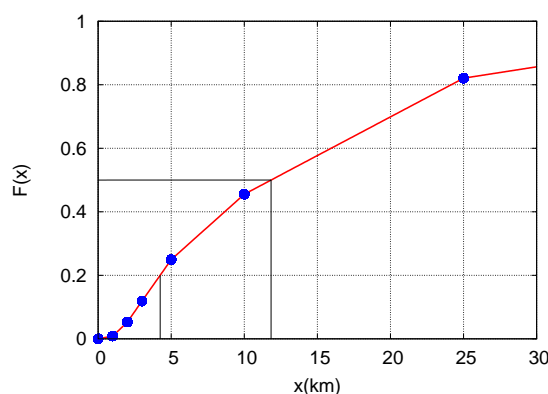
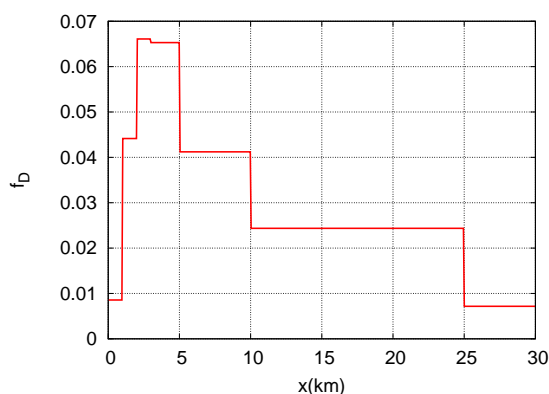
$$\mu = \frac{n}{n_{\text{Pers}}} = \underline{\underline{3.737 \text{ Wege/Person/Tag}}}$$

- (b) Bestimmen Sie die relativen Häufigkeiten, Summenhäufigkeiten und die empirische Dichtefunktion für jede Klasse.

Arbeitstabelle:

Klasse	x_k^*	f_k	F_k	f_k^D
0-1 km	0.5	0.0086	0.0086	0.0086
1-2 km	1.5	0.0442	0.0527	0.0442
2-3 km	2.5	0.0661	0.1188	0.0661
3-5 km	4.0	0.1306	0.2494	0.0653
5-10 km	7.5	0.2060	0.4554	0.0412
10-20 km	17.5	0.3653	0.8207	0.0244
20-50 km	37.5	0.1793	1.0000	0.0072

- (c) Zeichnen Sie die Dichte- und Verteilungsfunktion in die beiden folgenden Diagramme dieses Aufgabenblatts ein.



Warum ist die Häufigkeitsdichte der kürzesten Wege eher gering, obwohl diese doch sicherlich die attraktivsten sind?

Da es in sehr geringen Entfernungen einfach zu wenig mögliche Ziele gibt. Geht man von einer konstanten flächenbezogenen "Zieldichte" an möglichen Einkaufsgelegenheiten, Arbeitsstätten etc. aus, wächst die *entfernungsbezogene* Zieldichte linear mit der Entfernung

(d) Arithmetische Mittel (in km):

$$\bar{x} = \sum_k x_k^* f_k = \underline{\underline{15.42}}$$

Der Median ist in der 6. Klasse (siehe auch die Grafik):

$$x_{0.5} = x_5^u + \Delta x_5 \frac{F_5 - 0.5}{f_5} = \underline{\underline{11.82}}$$

Der Modus ist in der 3. Klasse (vgl. Grafik der Dichtefunktion). Ohne Feinberechnung gilt

$$\bar{x}_D = x_3^* = \underline{\underline{2.5}}$$

(mit Feinberechnung, aber nicht verlangt, 2.93). 60. Perzentil: Aus der Grafik der Verteilungsfunktion liest man $\underline{\underline{x_{0.6} = 16}}$ ab.

(e) *Als sog. "Pendlerpauschale" wurde in den Jahren 2007 und 2008 ein Satz von 30 Cent/km für den einfachen Arbeitsweg ab Beginn des 21. Kilometers auf die Steuer angerechnet. Wie viele Prozent kommen in den Genuss dieser Pauschale falls sich die Arbeitswege statistisch nicht von den anderen Wegen unterscheiden?*

Der Anteil der Reiseweiten ≥ 20 km (der 21. Kilometer beginnt bei $x = 20.000\dots1$!) ist gegeben durch

$$f_{\text{Pendlerpauschale}} = 1 - F(20) = 0.30$$

(direkt aus der Grafik abgelesen; Berechnung mit $F(20) = F_5 + f_6^D(20 - 10)$ ist auch möglich)

Falls die Grenze bei $x = 21$ gezogen wurde (ergibt auch volle Punktzahl): $f_{\text{Pendlerpauschale}}^I = 1 - F(21) = 0.27$.

(f) *Ab 2009 (und vor 2007) gibt es die Pendlerpauschale für Arbeitswege beliebiger Länge. Es wurde aber auch über eine komplette Abschaffung der Pendlerpauschale nachgedacht. Wie hoch sind die steuerlichen Mindereinnahmen pro Erwerbstätigen und Jahr durch (i) die Pendlerpauschale ab Beginn des 21. Kilometers, (ii) für alle Wege im Vergleich zu ihrer Abschaffung? Nehmen Sie dabei 240 Arbeitstage und einen mittleren Grenzsteuersatz von 30% an.*

Fall (i): Alle Kilometer ab Beginn des 21. ten werden vergütet. Dies betrifft also nur den Anteil $1 - F(20)$ von Leuten und auch bei diesen werden von der Reiseweite 20 Kilometer abgezogen. Berechnung des Mittelwertes (bezüglich aller Arbeitnehmer) der angerechneten Kilometer unter Annahme einer Gleichverteilung in den beiden obersten relevanten Klassen:

$$\bar{x}_{\text{anger}} = f_6^D(x_6^o - 20) \left(\frac{x_6^o + 20}{2} - 20 \right) + f_7(x_7^* - 20) = \underline{\underline{3.44 \text{ km}}}$$

(Hierbei ist $(x_6^o + 20)/2$ die mittlere Kilometerzahl der betroffenen Leute in Klasse 6.)

Der Steuerausfall S pro Jahr und Person ergibt sich nun, in dem man diese Zahl mit dem Satz 0.3 Euro/km, dem Grenzsteuersatz 0.3 und mit 240 Tagen multipliziert:

$$S = \bar{x}_{\text{anger}} * 0.3 \text{ Eur/km} * 0.3 * 240 = \underline{\underline{74 \text{ Eur/Person/Jahr}}}.$$

Fall (ii): Alle Kilometer von allen Personen werden berechnet. Dann gilt einfach

$$\bar{x}_{\text{anger}} = \bar{x} = \underline{\underline{15.4 \text{ km}}}$$

und

$$S = \bar{x} * 0.3 \text{ Eur/km} * 0.3 * 240 = \underline{\underline{333 \text{ Eur/Person/Jahr}}}.$$

Hinweis: Implizit muss man in diesem Aufgabenteil annehmen, dass die Weglängenverteilung der *Arbeitswege* identisch zu der aller Wege und damit auch identisch zu der der übrigen Wege ist. Das muss hinterfragt werden und ist i.A. nicht erfüllt (eine entsprechende Bemerkung ergab 3 Extrapunkte!)

Aufgabe 3

(35 Punkte)

Aus einer Mobilitätsuntersuchung wurde der Modal Split (Verkehrsmittelwahl) in Abhängigkeit der Weglänge festgestellt. Insbesondere ergaben sich die Häufigkeiten der Wahl "zu Fuß" in Abhängigkeit der Entfernungsklassen gemäß folgender Tabelle:

Reiseweitenklasse (km)	0-0.5	0.5-1	1-1.5	1.5-2	2-3	3-5
Häufigkeiten Wege	30	50	120	140	50	40
Häufigkeiten Wege zu Fuß	28	44	64	62	11	1

- (a) Bestimmen Sie zunächst für jede Entfernungsklasse die Anteilswerte der Fußwege. Anteile $\{A_1 = 0.933, A_2 = 0.880, A_3 = 0.533, A_4 = 0.443, A_5 = 0.220, A_6 = 0.025\}$
- (b) Gemäß eines Verkehrsmittelwahlmodells ist der Fußweganteil y in Abhängigkeit der Weglänge x (in km) gegeben durch

$$\hat{y}(x) = \frac{1}{1 + e^{a+bx}}.$$

Zeigen Sie, dass durch die Transformation

$$y \rightarrow v = \ln\left(\frac{1}{y} - 1\right)$$

das Modell in den transformierten Variablen linear ist.

Es gilt

$$\frac{1}{\hat{y}} - 1 = e^{a+bx}$$

und damit

$$\hat{v} = \ln\left(\frac{1}{\hat{y}} - 1\right) = a + bx$$

- (c) Wurde die unabhängige oder die abhängige Variable transformiert? Ist demnach die Regression des transformierten Modells äquivalent zu einer direkten Regression des nichtlinearen Modells oder nicht?

Es wurde die abhängige Variable transformiert. Damit ist das Ergebnis nicht äquivalent zu einer direkten nichtlinearen Regression.

- (d) Führen Sie die lineare Regression in den transformierten Variablen durch.

Arbeitstabelle mit den Klassenmitten $x_k^* = x_k$ als unabhängige und den Anteilen $A_k = y_k$ als abhängige Variable:

Klasse	x_k	y_k	v_k
0-0.5 km	0.25	0.933	-2.64
0.5-1 km	0.75	0.880	-1.99
1-1.5 km	1.25	0.533	-0.13
1.5-2 km	1.75	0.443	0.23
2-3 km	2.50	0.220	1.27
3-5 km	4.00	0.025	3.66

Damit die normale lineare Regression:

$$\bar{x} = \sum_{k=1}^6 f_k x_k = 1.686, \quad \bar{v} = \sum_{k=1}^6 f_k v_k = 0.110, \quad \sum_{k=1}^6 f_k x_k v_k = 686.1, \quad \sum_{k=1}^6 f_k x_k^2 = 1599,$$

also

$$b = \frac{\sum_{k=1}^6 f_k x_k v_k - \bar{x}\bar{v}}{\sum_{k=1}^6 f_k x_k^2 - \bar{x}^2} = \underline{\underline{1.61}}, \quad a = \bar{v} - b\bar{x} = \underline{\underline{-2.61}}.$$

Hinweis: Nach Aufgabenstellung ist die Kovarianz ist hier, ebenso wie die Mittelwerte und Varianzen, eine gewichtete Einfachsumme. Da mit den relativen Häufigkeiten gearbeitet wurde, fallen die Faktoren n bei $\bar{x}\bar{v}$ und \bar{x}^2 weg.

- (e) *Das Modell kann als Grenzfall einer logistischen Regression aufgefasst werden (vgl. die Vorlesungsunterlagen):*

$$\hat{y}(x) = \frac{y_s}{1 + e^{-\tilde{b}(x-x_0)}}$$

Geben Sie allgemein die Bedeutung der drei Parameter y_s , \tilde{b} und x_0 und für die vorliegende Situation die Zahlenwerte an. Gehen Sie dabei vom Regressionsergebnis $a = -2.61$ und $b = 1.61$ aus. Was würde ein negativer Wert von \tilde{b} ausdrücken?

Vergleich $y_s/(1 + e^{\tilde{b}x_0 - \tilde{b}x})$ mit $1/(1 + e^{a+bx})$:

- $y_s = 1$,
- $\tilde{b}x_0 = a = -2.61$, und
- $\tilde{b} = -b = -1.61$.

Damit

- $y_s = 1$ Sättigung,
- $\tilde{b} = -b = -1.61$ Geschwindigkeit des Anstiegs (da negativ: Geschwindigkeit des Abfalls) bzw. typische Intervallbreite $|1/b| = 0.62$ km.
- $x_0 = a/\tilde{b} = 1.61$ x -Wert, also hier Entfernung in km, bei stärkster Änderung des Modal Splits und gleichzeitig die Entfernung, bei der der Modal Split 1:1 beträgt.