Aufgabenblatt 6

Statistik für Wirtschaftsinformatiker, Übung, HTW Berlin

Martin Spott, Michael Heimann, Shirin Riazy

Stand: 11.05.2024

Wiederholung

- Was sind Varianz, Standardabweichung und Variationskoeffizient? Wie werden sie berechnet?
- Welche Fragestellung ist Gegenstand der Konzentrationsmessung einer Werteverteilung?
- Was wird mit einer Lorenzkurve dargestellt?
- Was ist der Gini-Koeffizient?

Aufgabe 6.1 (Streuungsmaße)

Berechne die mittlere absolute Abweichung vom arithmetischen Mittel, die Varianz und die Standardabweichung für das Merkmal *Schluss* aus dem Datensatz `bmw.csv` von Aufgabenblatt 5.

a) Berechne die drei Maße händisch in R
b) Benutze die R-Funktionen `var()` und `sd()` für Varianz und Standardabweichung.

Beachte: In R wird – wie bei den meisten anderen Statistikprogrammen – nicht die **empirische Varianz** (population variance) sondern die **Stichprobenvarianz** (sample variance) berechnet. Siehe in den Unterlagen zur Vorlesung nach, um den Unterschied herauszufinden.

Aufgabe 6.2 (Variationskoeffizient)

Der Aktienkurs der Volkswagenaktie wies in einem Zeitraum von 250 Handelstagen bei einem Mittelwert von 174,56€ eine Standardabweichung von 10,28€ auf. Für den gleichen Zeitraum ermittelt man für die Aktie der BMW AG eine Standardabweichung 4,68€ bei einem Mittelwert von 36,96€.

Berechne die Variationskoeffizienten für die Volkswagenaktie und für die BMW-Aktie. Was sagen uns die Werte? Vergleiche die Variationskoeffizienten der beiden Aktien.

Aufgabe 6.3 (Lorenzkurve)

Für das Jahr 2012 wurde folgende Statistik der Neuzulassungen für PKW veröffentlicht (Quelle: http://www.kfz-auskunft.de/kfz/pkw_neuzulassungen_hersteller_2012.html)

```
library(knitr)
setwd("C:/Users/rafaa/Desktop/HTW MODULE/Semester 3/Statistik/data")
zulassungen <- read.csv("neuzulassungen.csv")
kable(zulassungen)
```

Hersteller	Stueckzahl
VW	672921

Hersteller	Stueckzahl
BMW	284494
Mercedes	283006
Audi	266582
Opel	213627
Ford	206128
Renault	150740
Skoda	147197
Hyundai	100875
Toyota	83834

Erarbeite mit diesen Daten eine Lorenzkurve, ohne das R-Paket *ineq* zu nutzen. Füge dazu die folgenden Spalten als Hilfe zu, die wir auch in der Vorlesung im Beispiel mit den Einrichtungshäusern benutzt haben:

- i: Index, nach aufsteigenden Einheiten sortiert
- h_i: absolute Häufigkeit des Herstellers, also $h_i = 1$ für alle i
- f_i: relative Häufigkeit des Herstellers
- F_i: kumulierte relative Häufigkeit des Herstellers
- $h_i^* = \mathit{Stueckzahl}_i$: Stückzahl des Herstellers i
- $f_i^* = h_i^* / \sum_{j=1}^{10} h_j^*$: Einheiten jedes Herstellers relativ zur Summe aller Einheiten
- F_i^* sind die f_i^* kumuliert.

Welche beiden Spalten zeigt die Lorenzkurve? Zeichne die Lorenzkurve mit der Funktion `plot()`.

Aufgabe 6.4 (Lorenzkurve)

Erzeuge für die Daten aus Aufgabe 6.4 eine Lorenzkurve mit Hilfe des R-Paketes `ineq`. Betrachte dazu die Funktion `Lc()`, untersuche das Datenobjekt, das durch die Funktion erzeugt wird und plotte die Lorenzkurve.

Aufgabe 6.5 (Gini-Koeffizient)

a) Bestimme für die Daten aus Aufgabe 6.4 den Gini-Koeffizienten mit dem R-Paket `ineq` sowie den normierten Gini-Koeffizienten. Betrachte dazu die Funktionen `ineq()` und `Gini()`.
b) Warum ist es im Allgemeinen sinnvoll, neben dem Gini-Koeffizienten auch die Lorenzkurve zu betrachten?

Aufgabe 6.6 (Zusatzaufgabe)

Ergänze:

Besitzen alle Merkmalsträger denselben Merkmalswert, dann liegt eine _____ Konzentration vor. Auf 20% entfallen _____ % der Merkmalswertsummen, auf 40% entfallen _____ % der Merkmalswertsummen usw. Die Lorenzkurve und die Diagonale sind in diesem Fall _____ und die Fläche zwischen beidem ist gleich _____.

Vereinigt ein einziger Merkmalsträger die gesamte Merkmalswertsumme auf sich, so spricht man von _____ Konzentration. Je näher die Lorenzkurve zur Diagonalen liegt, desto _____ ist die Konzentration. Je weiter entfernt die Lorenzkurve zur Diagonalen liegt, desto _____ ist die Konzentration.

Das Ausmaß der Ungleichheit, also bildlich die Abweichung von Diagonale und Lorenzkurve, wird auch als _____ bezeichnet.

Aufgabe 6.7 (Zusatzaufgabe)

Bestimme für die Daten aus Aufgabe 4 den Gini-Koeffizienten ohne das R-Paket `ineq`, indem du die Fläche zwischen Lorenzkurve und Diagonale händisch in R berechnest.

Aufgabe 6.8 (Zusatzaufgabe, arithm. Mittel und Standardabw., klassifizierte Daten)

Wiederholung: Für eine Häufigkeitstabelle kann man das gewichtete arithmetische Mittel von Daten $x_1, x_2, \dots x_n$ wie folgt berechnen:

$$\bar{x} = \frac{1}{n}(h_1\, a_1 + h_2\, a_2 + \cdots + h_k\, a_k) = f_1\, a_1 + f_2\, a_2 + \cdots + f_k\, a_k$$

wobei die n Daten x_i nur die k verschiedenen Werte a_j mit absoluter bzw. relativer Häufigkeit h_j bzw. f_j annehmen.

Berechne näherungsweise das arithmetische Mittel und die Standardabweichung für die folgenden klassifizierten/gruppierten Daten. Benutze hierzu die obige Formel für das gewichtete arithmetische Mittel einer Häufigkeitstabelle. Ersetze die Werte a_j durch die Mittelpunkte der Intervalle und verwende die unten gelisteten Spalten als Hilfstabelle für die Zwischenschritte.

Die Daten beschreiben die Lebensdauer von Bauteilen in Stunden gruppiert in folgende Intervalle:

```
Intervall <- c("[300, 400)", "[400, 500)", "[500, 600)", "[600, 700)", "[700, 800)")
Haeufigkeit <- c(13, 25, 66, 58, 38)
lebensdauer <- data.frame(Intervall, Haeufigkeit)
kable(lebensdauer)
```

Intervall	Haeufigkeit
[300, 400)	13
[400, 500)	25
[500, 600)	66
[600, 700)	58
[700, 800)	38

Füge zur Berechnung folgende Spalten zu `lebensdauer` hinzu:

- j: der Index der Gruppe
- Intervallmitte a_j
- f_j: die relative Häufigkeit
- $f_j \cdot a_j$
- a_j^2
- $f_j \cdot a_j^2$

Warum kann man arithmetisches Mittel und Standardabweichung so nur näherungsweise berechnen?