2. Aufgabenblatt zum Stochastik–Praktikum

Deskriptive Statistik: Kenngrößen und Visualisierung von Daten

Aufgabe 1 („Methusalems machen Kasse“)

Im Rahmen einer Studie wurde eine Befragung in Unternehmen durchgeführt über die Einstiegsgehälter von Berufsanfängern. Bei der Befragung von 33 Unternehmen wurde das durchschnittliche Einstiegsgehalt in Bezug zur Studiendauer von Akademikern dokumentiert. Es ergaben sich die folgenden Werte:

Studiendauer	6	7	7	8	6	9	8	10	12	9	10
Durchschnittseinstiegsgehalt	30	27	26	21	31	20	27	18	20	23	21
Studiendauer	8	9	12	13	11	10	14	13	14	10	9
Durchschnittseinstiegsgehalt	37	36	28	26	33	35	31	23	28	37	39
Studiendauer	13	13	15	14	17	15	18	16	15	16	17
Durchschnittseinstiegsgehalt	42	41	38	40	37	36	30	35	35	33	36

Die Studiendauer ist in Semestern und das Einstiegsgehalt in 1000 Euro angegeben.

a) Definieren Sie 2 Vektoren mit jeweils den 33 angegebenen Werten für Studiendauer und Einstiegsgehalt. Erzeugen sie eine 33*2-Matrix aus den beiden Vektoren.

b) Bestimmen Sie die statistischen Kenngrößen Mittelwert, Median, Standardabweichung, Varianz, Minimum, Maximum und die 25- und 75-Prozent Quantile der Stichprobe für beide Variablen.

c) Definieren Sie eine Variable, die die Einstiegsgehälter in 3 Klassen ($<=$30, 30-35, $>$35) einteilt. Definieren sie außerdem eine Variable, die die Studiendauer in 4 Klassen ($<=$8, 8-10, 10-14, $>$14) einteilt.

d1) Stellen Sie den Zusammenhang von Einstiegsgehalt und Studiendauer anhand der diskreten Einteilungen in Klassen in einem bar-Plot (Typ: stack) dar und interpretieren sie diesen.

d2) Veranschaulichen Sie die relative Häufigkeitsverteilung der 3 Einstiegsgehälter–Klassen in einem pie–Diagramm.

e) Erstellen sie eine Grafik, in der die Werte der Einstiegsgehälter gegen die Studiendauern aufgetragen sind. Zeichnen sie eine Regressionsgerade ein. Analysieren sie das Ergebnis.

Eine ähnliche Studie wie diese (Erfundene) betitelte das Handelsblatt mit der Überschrift „Methusalems machen Kasse: Ein langes Studium zahlt sich in barer Münze aus“; Allerdings wurde bei dieser Schlussfolgerung eine Einteilung nach Studienfach einfach nicht berücksichtigt. Die Datei meth.dat enthält den vollständigen Datensatz für die Aufgabe mit der zusätzlichen Variablen des Studienfaches, nämlich Physik(PH), Chemie(CH) und Betriebswirtschaftslehre(BWL).

f) Lesen Sie die (gleichen) Daten als Datei diesmal direkt über Matlab ein und machen Sie sich mit den Eigenschaften und Bezeichnungen der Datei vertraut.

g) Berechnen Sie (zur Kontrolle nochmals) die statistischen Kenngrößen beider Variablen.

h) Plotten Sie einen Boxplot für beide Variablen

i) Erstellen Sie denselben Graphen wie oben mit Regressionsgerade für den gesamten Datensatz noch einmal.

j) Kennzeichnen sie die Punkte in dem Graphen dabei farblich unterschiedlich nach Studienfächern.

k) Zeichnen Sie in demselben Graphen zusätzlich Regressionsgeraden für die Teilmengen getrennt nach Studienfächern ein in unterschiedlichen Farben die mittels einer Legende bezeichnet werden. Analysieren sie das Resultat.

Aufgabe 2 (Grafische Darstellung von Quantilen/Konfidenzbereichen)

Das Ziel ist es, einen Plot anzufertigen, der die Dichte der Standardnormalverteilung in einem geeigneten Bereich mit den eingezeichneten $0, 50, 75, 90, 95, 99$–Prozent–Quantilen zeigt. Die Abschnitte unter der Dichtekurve zwischen den Quantilen sollen in unterschiedlichen Farben gekennzeichnet werden.

Zusatz* (freiwillig) Schreiben Sie eine Funktion, die von den vier Argumenten Dichte, (zugehöriger) Verteilungsfunktion, einem Vektor vorgegebener Quantile und einem Vektor von Gitterpunkten abhängt, und mit der sich ein derartiger Plot für beliebige andere Verteilungen erzeugen lässt.

Hinweise:
Beginnen Sie die Funktionswerte der Dichte auf einem diskreten Gitter zu plotten. Um einen Vektor an Farben vorab festzulegen benutzen sie z. Bsp. den colormap–Befehl. Definieren sie zunächst die Bereiche zwischen den Quantilen mit der Verteilungsfunktion normcdf und zugehörige Werte der Dichte mit normpdf. Um Bereiche einzufärben kann man die Funktion fill benutzen. Informieren sie sich (wie immer) über die Verwendung der Befehle in der Hilfe.

Aufgabe 3 (Simulation und Darstellung von multivariaten Daten mit Copulas)

a) Simulieren Sie $n = 500$ auf $[0, 1]$ gleichverteilte Pseudozufallsvariablen U_i und generieren Sie daraus mit der Inversionsmethode n (univariate) Pseudozufallsvariablen S_i die verteilt sind gemäß der Dichte $f(s) = \frac{1}{2}\exp(-|s|)$.

b) Geben Sie einen Boxplot der Stichprobe der S_i aus. Generieren Sie zudem eine Graphik, die den Graphen von f enthält gemeinsam mit einem Histogram der Stichprobe der S_i, welches Sie so skalieren, dass es einer Dichtefunktion entspricht (d.h. so dass die Fläche unter dem Graphen gleich Eins ist).

c) Simulieren Sie nun $n = 500$ zweidimensional verteilte Pseudozufallsvariablen $X_i = (X_i^1, X_i^2)$ derart, dass die Randverteilungen von X jeweils gemäß f aus a) verteilt sind und die gemeinsame Verteilung durch die Gauss-Copula einer multivariaten Normalverteilung gegeben ist, für welche beide Einzelkoordinaten standardnormalverteilt sind und die Korrelation der verschiedenen Koordinaten ρ ist. Illustrieren Sie die gemeinsame Verteilung und den Einfluß des Parameters ρ mit einem Scatterplot für X, welcher gemeinsam in jeweils verschiedenen Farben die Simulationsergebnisse für $\rho = -0.9, 0.0, 0.9$ zeigt.

d) Berechnen Sie den Mittelwert $\bar{M}$ von $M_i := min(X_i^1, X_i^2)$ auf Basis jeweils einer 500er Stichprobe als Funktion von ρ und plotten Sie den Graphen über $-1 \leq \rho \leq +1$, um die Abhängigkeit von ρ für $\bar{M} \approx E[M]$ zu illustrieren.