Stochastik

Serie 2

Aufgabe 2.1

Der Geysir Old Faithful im Yellowstone National Park ist eine der bekanntesten heissen Quellen. Für die Zuschauer und den Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse.

Auf Ilias sind die Messungen in der Datei `geysir.dat` vom 1.8.1978 - 8.8.1978 in 3 Spalten abgelegt: `Tag`, `Zeitspanne` und `Eruptionsdauer`.

a) Zeichnen Sie Histogramme von der Zeitspanne zwischen zwei Ausbrüchen:

```r
# Datensatz einlesen
geysir <- read.table("./Daten/geysir.dat", header = TRUE)
# 4 Graphiken im Graphikfenster
par(mfrow = c(2, 2))
hist(geysir[, "Zeitspanne"])
hist(geysir[, "Zeitspanne"], breaks = 20)
hist(geysir[, "Zeitspanne"], breaks = seq(41, 96, by = 11))
```

Was fällt auf? Was ist der Unterschied zwischen den drei Histogrammen?

Bemerkung: Wenn man die Anzahl Klassen mit `breaks=20` vorgibt, so wird dies nur als „Vorschlag“ interpretiert und intern unter Umständen abgeändert.

b) Zeichnen Sie Histogramme (Anzahl Klassen variieren) von der Eruptionsdauer:

```r
hist(geysir[, "Eruptionsdauer"])
```

Was fällt auf? Vergleichen Sie mit der ersten Teilaufgabe.

c) Zeichnen Sie die empirische kumulative Verteilungsfunktion von der Eruptionsdauer von Old Faithful Geysir. Untersuchen Sie, wie viel Prozent der Eruptionen höchstens 2 Minuten gedauert haben, sowie welche Eruptionsdauer der 60 % Eruptionen, die am längsten gedauert haben, mindestens gedauert haben.

```r
eruptionsdauern <- geysir[, "Eruptionsdauer"]
n <- length(eruptionsdauern)
# Bei jedem Datenpunkt springt der Wert der
```

```r
# kumulativen Verteilungsfunktion um 1/n
plot(sort(eruptionsdauern), ...., type = "s", ylim = c(0,
    1), ylab = "...", xlab = "...", main = "")
```

Aufgabe 2.2

In einer Klasse wurden in einer Statistik-Prüfung folgende Noten geschrieben:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

a) Ändern Sie drei Noten im Datensatz so ab, dass der Median gleich bleibt, aber der Mittelwert sich stark ändert.

b) Erstellen Sie zu den beiden Datensätzen je ein Histogramm und einen Boxplot.

Aufgabe 2.3

21 Labors bestimmten den Kupfergehalt von 9 verschiedenen Klärschlammproben. Die Daten stehen in der auf Ilias abgelegten Datei `klaerschlamm.dat` zur Verfügung. Die erste Spalte bezeichnet das Labor, die restlichen 9 Spalten sind die verschiedenen Klärschlammproben. Die Daten (in mg/kg) können mit dem Befehl

```r
schlamm.all <- read.table(file = "./Daten/klaerschlamm.dat",
    header = TRUE)
schlamm <- schlamm.all[, -1]  # Labor-Spalte entfernen
```

eingelesen werden.

a) Erstellen Sie für jede Probe einen Boxplot, und berechnen Sie jeweils das arithmetische Mittel und den Median. Bei welchen Proben gibt es Ausreisser, und wo unterscheiden sich arithmetisches Mittel und Median wesentlich? Bei welchen der 9 Proben ist es plausibel, dass die wahre Konzentration unter 400 mg/kg liegt?

R-Hinweise:

```r
summary(schlamm)
boxplot(schlamm)
```

b) Erstellen Sie für jedes Labor einen Boxplot der Messfehler. Unter dem Messfehler eines Labors bei einer Probe verstehen wir den gemessenen Wert minus den Median über alle Labors. Welche der 21 Labors haben systematische Fehler in

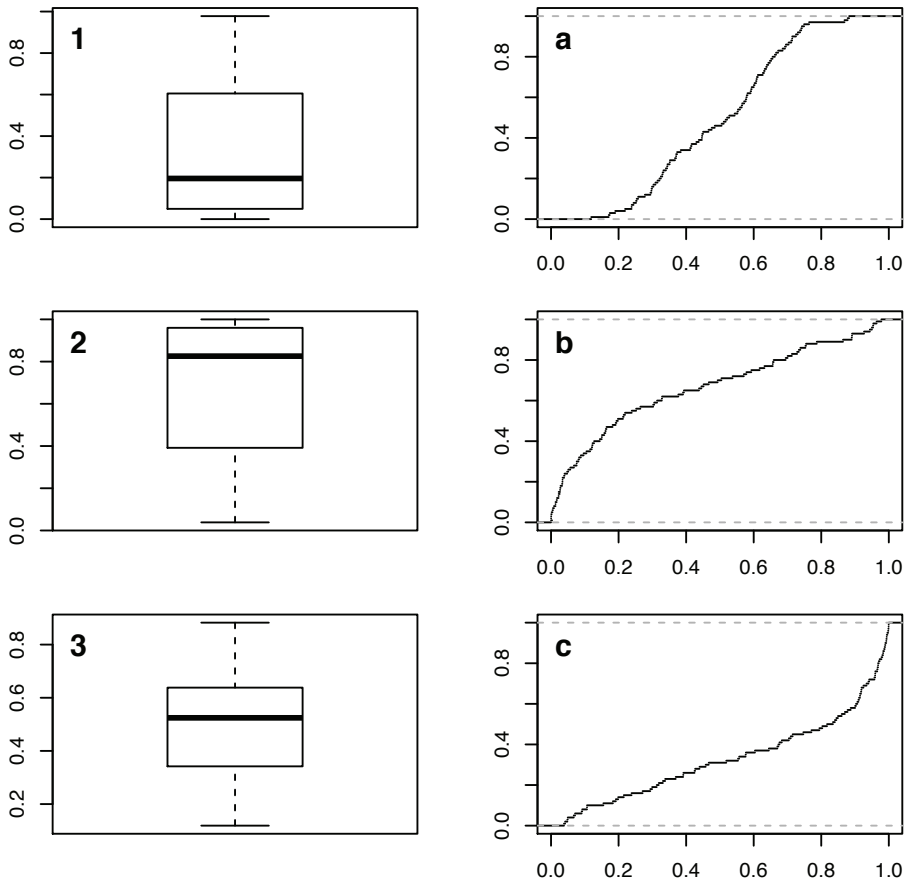
ihrem Analyseverfahren? Welche haben grosse Zufallsfehler, und bei welchen Labors ist die Qualität der Analysen besonders gut?

R-Hinweise:

```
# Fuer jede Spalte Median berechnen
med <- apply(schlamm, 2, median)
# Median von jeder *Spalte* abziehen
schlamm.centered <- scale(schlamm, scale = FALSE, center = med)
# Boxplot zeichnen. Dazu zuerst data-frame transponieren
boxplot(data.frame(t(schlamm.centered)))
```

Aufgabe 2.4

Für drei Stichproben vom Umfang $n = 100$ wurden je ein Boxplot und die empirische Verteilungsfunktion gezeichnet. Ordnen Sie die drei Boxplots den entsprechenden empirischen Verteilungsfunktionen zu:



Aufgabe 2.5

Edwin Hubble untersuchte seit 1920 am Mount Wilson Observatory die Eigenschaften von Galaxien ausserhalb der Milchstrasse. Mit Überraschung bemerkte er einen Zusammenhang zwischen der Distanz einer Galaxie zur Erde und dessen Geschwindigkeit, sich von der Erde fortzubewegen (Fluchtgeschwindigkeit). Hubbles ursprüngliche Daten von 24 galaktischen Nebeln (E. Hubble, „Proceedings of the National Academy of Science 15 (1929): 168-73“) sind in Tabelle 1 gezeigt. Die Fluchtgeschwindigkeit ist in Kilometer pro Sekunde angegeben und konnte aufgrund der Rotverschiebung im Lichtspektrum der Galaxien mit grosser Genauigkeit bestimmt werden. Die Distanz einer Galaxie zur Erde wird in Megaparsec (Mpc) gemessen: ein Megaparsec entspricht etwa 3.09×10^{10} m. Die Distanzen werden durch Vergleich der mittleren Luminosität von Galaxien mit der Luminosität von bestimmten bekannten Sternen bestimmt, wobei diese Methode relativ ungenau ist.

a) Erstellen Sie von den Daten in Tabelle 1 ein Streudiagramm, in dem Sie die Distanz versus Fluchtgeschwindigkeit aufzeichnen.

b) Schätzen Sie aus den Daten die Parameter β_0 und β_1 für die Regressionsgerade

$$y = \beta_0 + \beta_1 x \,,$$

wobei y die Distanz und x die Fluchtgeschwindigkeit bezeichnet. Benützen Sie dabei die aufgrund der Methode der kleinsten Quadrate ermittelten Parameterschätzungen.

c) Identifizieren Sie in der R-Ausgabe von

```
lm(y ~ x)
```

die Koeffizienten β_0 und β_1.

Aufgabe 2.6

Wir betrachten eine Studie, die 1979 in den Vereinigten Staaten durchgeführt wurde (National Longitudinal Study of Youth, NLSY79): von 2584 Amerikanern im Jahr 1981 wurde der Intelligenzquotient (gemäss AFQT - armed forces qualifying test score) gemessen; 2006 wurden dieselben Personen nach ihrem jährlichen Einkommen im Jahr 2005 und der Anzahl Jahre Schulbildung befragt. Uns interessiert hier natürlich, ob ein hoher IQ oder eine lange Schulbildung zu einem höheren Einkommen führen.

Nebel	Geschwindigkeit (km/s)	Distanz (Mpc)
S. Mag.	170	0.032
L. Mag. 2	290	0.034
NGC 6822	-130	0.214
NGC 598	-70	0.263
NGC 221	-185	0.275
NGC 224	-220	0.275
NGC 5457	200	0.450
NGC 4736	290	0.500
NGC 5194	270	0.500
NGC 4449	200	0.630
NGC 4214	300	0.800
NGC 3031	-30	0.900
NGC 3627	650	0.900
NGC 4626	150	0.900
NGC 5236	500	0.900
NGC 1068	920	1.000
NGC 5055	450	1.100
NGC 7331	500	1.100
NGC 4258	500	1.400
NGC 4151	960	1.700
NGC 4382	500	2.000
NGC 4472	850	2.000
NGC 4486	800	2.000
NGC 4649	1090	2.000

Tabelle 1: Zusammenhang zwischen Distanz und Fluchtgeschwindigkeit von Galaxien.

In der auf Ilias abgelegten Datei `income.dat` finden Sie den Datensatz mit dem Einkommen, der Anzahl Jahre abgeschlossener Schulbildung und den ermittelten Intelligenzquotienten von 2584 Amerikanern.

a) Lesen Sie den Datensatz `income.dat` ein und generieren Sie Streudiagramme, in welchen das Einkommen versus Anzahl Jahre Schulbildung und Einkommen versus Intelligenzquotient aufgetragen sind.

b) Bestimmen Sie die Parameter a und b des linearen Modells $y = a + bx$, wobei y das Einkommen bezeichnet und x die Anzahl Jahre Schulbildung. Zeichnen Sie die Regressionsgerade mit der R-Funktion

```
plot(..., ..., type = "l")
```

Wie interpretieren Sie die Parameter a und b?

c) Berechnen Sie die Korrelation zwischen Einkommen und Anzahl Jahre Schulbildung. Wie angebracht ist ein Regressionsmodell für diesen Datensatz?

Aufgabe 2.7

a) Erzeugen Sie den Vektor `t.x` mit den Werten $-10, -9, \ldots, 9, 10$ und den Vektor `t.x1` mit den Werten $0, 1, \ldots, 9, 10$. Erzeugen Sie dann die Vektoren `t.y` und `t.y1`, deren Elemente die Quadratwerte der entsprechenden Elemente von `t.x` bzw. `t.x1` enthalten.

b) Zeichnen Sie die Streudiagramme `t.y` vs. `t.x` und `t.y1` vs `t.x1`. Benützen Sie die R-Funktion

```
plot()
```

c) Berechnen Sie die Korrelationskoeffizienten zwischen `t.x` und `t.y` bzw. zwischen `t.x1` und `t.y1`. Benützen Sie die R-Funktion

```
cor()
```

Warum sind die beiden Korrelationen so verschieden?

Aufgabe 2.8

In dieser Aufgabe betrachten wir 4 Datensätze, die von Anscombe konstruiert wurden. In jedem der Datensätze gibt es eine Zielvariable y und eine erklärende Variable x.

a) Stellen Sie jeden der 4 Datensätze als Streudiagramm dar, zeichnen Sie die Regressionsgerade ein und kommentieren Sie die Ergebnisse.

b) Vergleichen Sie die Schätzungen von β_0 und β_1, wobei $y = \beta_0 + \beta_1 x$.

R-Hinweise:

```
data(anscombe)   ## Einlesen des Datensatzes
```

Die Schätzungen für die Koeffizienten β_0 und β_1 des linearen Regressionsmodells kann man mit

```
lm(y1 ~ x1, data = anscombe)   # oder
lm(anscombe$y1 ~ anscombe$x1)
```

berechnen und numerisch auswerten. Mit `par(mfrow=c(2,2))` wird das Grafikfenster so eingeteilt, dass alle 4 Bilder nebeneinander passen. Den Scatterplot und die Regressionsgerade erhält man mit

```
plot(anscombe$x1, anscombe$y1)
reg <- lm(anscombe$y1 ~ anscombe$x1)
abline(reg)
```

Kurzlösungen vereinzelter Aufgaben

A 2.6:

b) $a = -40200$ und $b = 6451$

c) $r = 0.346$

A 2.7:

c) $r_{t.x,t.y} = 0$ und $r_{t.x1,t.y1} = 0.963$

A 2.8:

b) $\widehat{\beta_0} \approx 3.00$ und $\widehat{\beta_1} \approx 0.500$