# Computergestützte Statistik

Prof. Dr. Andreas Behr

## Übungsblatt 1

Einführung

1. Daten erzeugen.

   (a) Erzeugen Sie eine Sequenz von 1 bis 9 und weisen Sie die Sequenz der Variablen $a$ zu.

   (b) Erzeugen Sie den Vektor $b = \{1, 1, 1, 2, 2, 2, 3, 3, 3\}$.

   (c) Erzeugen Sie den Vektor $c = \{1, 2, 3, 1, 2, 3, 1, 2, 3\}$.

   (d) Erzeugen Sie den Vektor $d = \{1, 2, 2, 3, 3, 3, 4, 4, 4\}$.

   (e) Erzeugen Sie aus den vier Vektoren $a, b, c, d$ eine Matrix.

   (f) Erzeugen Sie aus den vier Vektoren $a, b, c, d$ einen Dataframe.

   (g) Ändern Sie die Namen der vier Variablen des Dataframes in $''x1'','' x2'','' x3'','' x4''$.

   (h) Erzeugen Sie einen Vektor mit logischen Werten, der angibt, ob $x2 = 2$ ist.

2. Berechnen Sie:

   (a) $2 + 3 * (4 + 5 * (6 + 7 * (8 + 9)))$

   (b) $2^{11}$

   (c) $log_e(2)$

   (d) $log_{10}(2)$

   (e) $\sqrt{333}$.

3. Berechnen Sie:

   (a) $\sum_{i=1}^{20} 0.1i^3$

   (b) $\prod_{i=1}^{20} i^{0.4}$.

   (c) $\frac{1335}{4}y^6 + x^2\left(11x^2y^2 - y^6 - 121y^4 - 2\right) + \frac{11}{2}y^8$ für $x = 77617$ und $y = 33096$. (Hinweis: Die korrekte Lösung lautet $-2$)

# Übungsblatt 2

Daten

1. Erzeugen Sie den folgenden data.frame und weisen Sie ihn dem Objekt *dat* zu:

| pid | name | sex | age |
|---|---|---|---|
| 1 | Susi | 1 | 22 |
| 2 | Carmen | 1 | 24 |
| 3 | Herbert | 0 | 21 |
| 4 | Karl | 0 | 27 |

   (a) Erzeugen Sie aus Ihrem data.frame *dat* einen neuen data.frame *dat*2, der nur die Variablen *name* und *age* enthält, indem Sie den subset-Befehl (alternativ: den Selektionsbefehl `[]`) verwenden.

   (b) Erzeugen Sie aus Ihrem data.frame *dat* einen weiteren neuen data.frame *dat*3, der nur die Beobachtungen von Susi und Karl enthält, indem Sie den subset-Befehl (alternativ: den Seletionsbefehl `[]`) verwenden.

   (c) Speichern Sie Ihren data.frame *dat* als R-Datendatei, entfernen Sie mit dem Befehl `rm()` (von remove) den Datensatz aus dem Arbeitsspeicher und lesen Sie das gespeicherte Datenfile wieder in den Arbeitsspeicher.

   (d) Speichern Sie Ihren data.frame *dat* als CSV-Datei, entfernen Sie den Datensatz aus dem Arbeitsspeicher und lesen Sie das gespeicherte Datenfile wieder in den Arbeitsspeicher.

   (e) Sortieren Sie das Datenfile dat aufsteigend (absteigend) nach *age*.

2. Laden Sie den Datensatz *usa.csv* in den Arbeitsspeicher und weisen Sie ihn dem Objekt *d* zu. (Die Sektorkodierung lautet: 0=Missing, 1=Agriculture, 2=Energy, 3=Mining, 4=Manufacturing, 5=Construction, 6=Trade, 7=Transport, 8=Bank/Insurance, 9=Services, sex: Männer sind die 0en)

   (a) Ermitteln Sie die Dimension des Datensatzes.

   (b) Welche Variablen enthält der Datensatz?

   (c) Ergänzen Sie den Datensatz um die Variable *wagerate*, die den Stundenlohn enthält.

   (d) Ergänzen Sie den Datensatz um die Variable *logwagerate*, die den Logarithmus zur Basis 10 des Stundenlohns enthält.

   (e) Speichern Sie den Datensatz *usa*2.*csv* in Ihrem Arbeitsverzeichnis unter dem Filenamen *usa*2.*csv* mit einem Punkt als Dezimaltrennzeichen und einem Komma als Trennzeichen der Einträge.

   (f) Löschen Sie ihn anschließend aus dem Arbeitsspeicher und laden Sie den Datensatz wieder ein.

   (g) Ermitteln Sie die Dimension des Datensatzes und lassen Sie sich die Variablennamen ausgeben.

# Übungsblatt 3

Grafiken

1. Geben Sie bitte die folgenden Befehle ein:

   ```
   >set.seed(1)
   >x <- c(rnorm(500),rnorm(500,mean=5,sd=4))
   >y <- 2*x+rnorm(1000)
   ```

   (a) Stellen Sie die Häufigkeitsverteilung von $x$ als Histogramm dar. Vergleichen Sie dabei die Auswirkungen unterschiedlicher Berechnungsvorschriften der "optimalen Klassenzahl".

   (b) Wählen Sie unterschiedliche Klassenzahlen und vergleichen Sie die resultierenden Histogramme.

   (c) Erstellen Sie ein $XY$-Streudiagramm mit einem Diagrammtitel. Vergleichen Sie die Wirkung verschiedener Symbole und Farben für die Darstellung der Datenpunkte.

   (d) Erstellen Sie eine Liniengraphik, die die Entwicklung von $x$ darstellt.

   (e) Erzeugen Sie einen Boxplot der beiden Variablen $x$ und $y$ und lassen Sie sich die in dem Plot enthaltenen Informationen ausgeben.

   (f) Vervollständigen Sie die erstellten Graphiken durch weitere Spezifikationen (Legende, Titel, Untertitel, Achsenbeschriftungen, etc.).

   (g) Erzeugen Sie ein pdf-File der erstellten Grafik.

2. Lesen Sie den Datensatz *usa2.csv* ein und weisen Sie ihn dem Objekt $d$ zu.

   (a) Stellen Sie die Verteilung der Jahreslöhne mit Hilfe eines Histogramms dar. Wie beurteilen Sie die Aussagekraft Ihres Histogramms?

   (b) Möglicherweise verdienen einige Menschen sehr viel mehr als andere, weil sie sehr viel mehr arbeiten. Versuchen Sie, ob die Verwendung der Stundenlöhne anstelle der Jahreslöhne das Problem löst.

   (c) Stellen Sie die logarithmierten Stundenlöhne mittels eines Histogramms dar.

   (d) Was vermuten Sie ist der Medianstundenlohn?

   (e) Erzeugen Sie ein Streudiagramm von Alter und logarithmierten Stundenlöhnen. Können Sie einen Zusammenhang erkennen? Vermuten Sie ein Problem bei dieser Darstellung? Haben Sie eine Idee, wie Sie dieses Problem beheben können?

# Übungsblatt 4

Deskription

1. Geben Sie bitte die folgenden Befehle ein:

   `set.seed(123); x <- round(rlnorm(1000)*1000)`

   (a) Stellen Sie die Häufigkeitsverteilung mittels eines Histogramms dar.

   (b) Beschreiben Sie die Verteilung mittels der mit dem Befehl `summary(x)` berechneten Maßzahlen.

   (c) Berechnen Sie die Dezile.

   (d) Vergleichen Sie die Ergebnisse der drei Streuungsmaße Quartilsabstand, Standardabweichung und mittlere absolute Abweichung vom Zentralwert.

   (e) Berechnen Sie die aus den Zentralmomenten abgeleiteten Maße für die Schiefe und die Kurtosis.

2. Erzeugen Sie die folgenden drei Variablen: `x <- rep(1:3,4); y <- rep(1:2,6); z <- c(1:2,1:3,1:4,3:5)`.

   (a) Betrachten Sie die drei paarweisen Häufigkeitstabellen.

   (b) Fügen Sie an die Häufigkeitstabelle der Variablen $X$ und $Y$ die Randsummen an.

   (c) Erzeugen Sie die Randverteilung von $X$ und $Y$ aus einem Objekt der Funktion `table(x,y,z)`.

   (d) Erzeugen Sie mit Hilfe des Befehls `table()` Häufigkeitstabellen für alle drei Variablen und ändern Sie dabei die Reihenfolge der Variablen.

   (e) Erzeugen Sie eine 'flache Tabelle' und betrachten Sie auch hier verschiedene Anordnungen der drei Variablen.

   (f) Ermitteln Sie für $Z$ die relativen Häufigkeiten unter der Bedingung, dass $Y = y$.

3. Lesen Sie den Datensatz *usa2.csv* ein und weisen Sie ihn dem Objekt $d$ zu.

   (a) Ermitteln Sie das mittlere Jahreseinkommen.

   (b) Um wieviel Prozent liegt das mittlere Jahreseinkommen der Männer über dem der Frauen?

   (c) Vermuten Sie, dass der prozentuale Unterschied in den Stundenlöhnen größer oder kleiner ist? Überprüfen Sie Ihre Vermutung.

   (d) Ermitteln Sie die mittleren Stundenlöhne in den verschiedenen Sektoren. In welchem Sektor wird der höchste, in welchem der niedriegste Stundenlohn gezahlt?

   (e) Informieren Sie sich mit der Hilfe über den Befehl `cut()`. Erzeugen Sie eine Variable *cwage*, die angibt, ob eine Person im unteren, mittleren oder oberen Drittel der Einkommensverteilung liegt. Überprüfen Sie Ihre Operation mit Hilfe der Befehle `length()` und `table()`.

(f) Erzeugen Sie eine Häufigkeitstabelle der drei Variablen *sex*, *cwage* und *sector*.

i. Wieviel Prozent der Frauen liegen im untersten Einkommensdrittel?

ii. Wieviel Prozent der Männer liegen im obersten Einkommensdrittel?

iii. Wie hoch ist der Anteil der Männer im obersten Einkommensdrittel, wie hoch insgesamt?

iv. Vergleichen Sie die sektorale Beschäftigungsstruktur der Personen im obersten Einkommensdrittel mit der im untersten Einkommensdrittel. Was fällt Ihnen besoners auf?

# Übungsblatt 5

Wahrscheinlichkeitsverteilungen

1. Erzeugen Sie 120 Realisationen einer normalverteilten Zufallsvariable mit $\mu = 20$ und $\sigma^2 = 16$ (`set.seed(123)`), weisen Sie die Realisationen der Variable $x$ zu und stellen Sie die Verteilung mittels eines Histogramms dar. Achten Sie darauf, dass Sie die Dichte und nicht die absolute Häufigkeit abtragen.

   (a) Berechnen Sie die Dichte der Normalverteilung mit den Parametern $\mu = 20, \sigma^2 = 16$ an der Stelle $x = 18$.

   (b) Wie lautet der Wert der Verteilungsfunktion an der Stelle $x = 18$?

   (c) Wieviel Prozent Ihrer Beobachtungen haben einen Wert $\leq 18$?

   (d) Kennzeichnen Sie in Ihrem Histogramm das symmetrische Intervall um den Erwartungswert, in das zufällige Realisationen mit einer Wahrscheinlichkeit von 80% zu liegen kommen.

   (e) Wieviel Prozent Ihrer Realisationen liegen tatsächlich in dem Intervall?

   (f) Erzeugen Sie (mehrfach) neue Realisationen und ermitteln Sie jeweils den Anteil der Beobachtungen, die in dem in d) berechneten Intervall liegen.

2. Lesen Sie den Datensatz *usa2.csv* ein und weisen Sie ihn dem Objekt $d$ zu.

   (a) Stellen Sie die logarithmierten Stundenlöhne mit Hilfe eines Histogramms dar (Dichten).

   (b) Ermitteln Sie für den logarithmierten Stundenlohn das arithmetische Mittel und die Varianz.

   (c) Tragen Sie in Ihr Histogramm eine 'angepasste' Normalverteilung ein. D.h. eine Normalverteilung mit den von Ihnen berechneten Maßzahlen.

   (d) Zeichnen Sie die empirische Verteilungsfunktion und tragen Sie zusätzlich die Verteilungsfunktion der angepassten Normalverteilung in die Grafik ein. (Hilfe: Informieren Sie sich mit der Hilfe über die beiden Funktionen `ecdf()` und `plot.ecdf()`.)

# Übungsblatt 6

Regression

1. Stochastisches Regressionsmodell

   (a) Erzeugen Sie 20 Realisationen einer standardnormalverteilten Zufallsvariablen (`set.seed(1)`) und weisen Sie die Realisationen der Variablen $x$ zu. Berechnen Sie Regressionswerte $xb$ mittels einer linearen Regression mit dem Achsenabschnitt 1 und der Steigung 0.5. Überlagern Sie die Regressionswerte $xb$ mit Realisationen (`set.seed(3)`) einer standardnormalverteilten Zufallsvariablen ($u$) und weisen Sie die so entstandenen Werte dem Vektor $y$ zu.

   (b) Stellen Sie $x$ und $y$ in einem Streudiagramm dar und zeichnen Sie den den "wahren" linearen Zusammenhang ein.

   (c) Ermitteln Sie mittels der Funktion `lm()` die Regressionswerte einer nach der Methode der kleinsten Quadrate geschätzten Regressionsgeraden und tragen Sie die Gerade in Ihr Diagramm ein.

   (d) Erzeugen Sie einen neuen Vektor $v$ mit 20 Realisationen einer standardnormalverteilten Zufallsvariablen (`set.seed(5)`) und überlagern Sie die deterministische Komponente $xb$ mit diesen Störtermen. Tragen Sie die neuen Realisationen farblich unterschieden in Ihr Diagramm ein.

   (e) Ermitteln Sie für die neuen Realisationen Ihrer Zufallsvariablen $Y$ (z.B. $y2$) eine Regressionsgerade und tragen Sie diese zusätzlich in Ihr Diagramm ein.

   (f) Testen Sie für Ihre beiden Realisationen $(y, y2)$ jeweils die Hypothese, dass der Steigungsparameter im Modell 0.5 beträgt.

2. Lesen Sie den Datensatz *usa2.csv* ein und weisen Sie ihn dem Objekt $d$ zu.

   (a) Erzeugen Sie ein Streudiagramm des Alters und der logarithmierten Stundenlöhne.

   (b) Berechnen Sie mit Hilfe der Funktion $lm()$ eine lineare Regressionsfunktion.

   (c) Was besagt der von Ihnen ermittelte Steigungsskoeffizient?

   (d) Welche Informationen liefert Ihnen die Funktion $summary()$ über die Ihre Regressionsfunktion?

   (e) Tragen Sie die berechnete Regression in das Streudiagramm ein.

   (f) Erzeugen Sie ein Histogramm der ermittelten Residuen. Ähnelt es einer Normalverteilung?