Universität Duisburg-Essen
Lehrstuhl für Ökonometrie
Dr. Yannick Hoga
M.Sc. Martin Arnold

Wintersemester 2022/23

Übungsblatt 1 — Grundlagen in Statistik mit R

Zu diesem Übungsblatt empfehlen wir neben der Lektüre der Kapitel 2 und 3 des Lehrbuches *Introduction to Econometrics* von *Stock & Watson* eine Aufarbeitung mithilfe der Kapitel 2 und 3 in unserem Online-Companion *Introduction to Econometrics with R*.

Aufgabe 1 – Rechnen mit der Normalverteilung

Dem *Sozio-ökonomischen Panel* (SOEP) zufolge beträgt die Körpergröße von volljährigen, männlichen Bundesbürgern im Mittel 179cm bei einer Standardabweichung von etwa 7cm. Gehen Sie davon aus, dass die Körpergröße ein normalverteiltes Merkmal ist. Für diese Population gilt also:

$$\text{Körpergröße} \sim \mathcal{N}(\mu = 179, \sigma^2 = 49)$$

(d) Nutzen Sie `R`, um die in den Teilaufgaben (a) bis (c) gesuchten Ergebnisse zu berechnen. Hierfür ist die Funktion `pnorm()` hilfreich.

Lösung:

```
# (a)
pnorm(170, mean = 179, sd = 7)

# (b)
pnorm(190.5, mean = 179, sd = 7, lower.tail = F)

# (c)
pnorm(205, mean = 199, sd = 7.5) - pnorm(190, mean = 199, sd = 7.5)
```

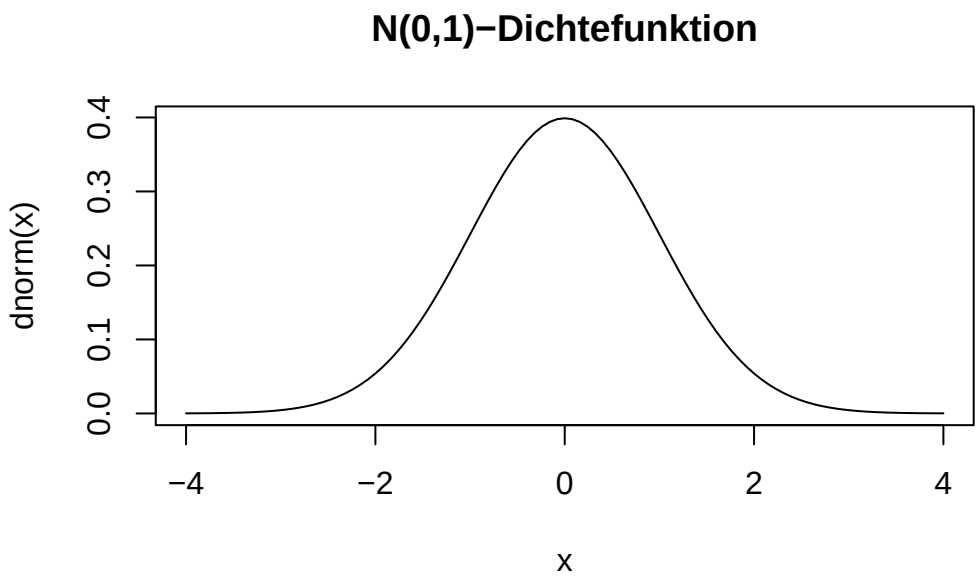
```
[1] 0.0992714
[1] 0.05020625
[1] 0.6730749
```

Aufgabe 2 – Wahrscheinlichkeitsverteilungen in R

(a) Plotten Sie die $\mathcal{N}(0,1)$-Dichte mit den Funktionen `dnorm()` und `curve()`.

Lösung:

```
# Wir plotten die N(0,1)-Dichte über dem Intervall [-4,4]
curve(dnorm(x),
      from = -4,
      to = 4,
      main = "N(0,1)-Dichtefunktion")
```



(b) Berechnen Sie die Dichte einer $\mathcal{N}(3, 4^2)$-verteilten Zufallsvariable an der Stelle 4 mit `dnorm()`.

Lösung:

```
# Beachte, dass sd die Standardabweichung ist
dnorm(x = 4, mean = 3, sd = 4)
```

```
[1] 0.09666703
```

(c) Berechnen Sie $P(X \leq 4)$ für $X \sim N(3, 4^2)$ mit `pnorm()`.

Lösung:

```
# Gesucht ist P(X<=4) für X ~ N(3,16)
pnorm(q = 4, mean = 3, sd = 4)
```

```
[1] 0.5987063
```

(d) Berechnen Sie das 0.95-Quantil einer $\mathcal{N}(3, 4^2)$-verteilten Zufallsvariable mit `qnorm()`.

Lösung:

```
# Quantile können mit qnorm() berechnet werden
qnorm(p=0.95, mean = 3, sd = 4)
```

```
[1] 9.579415
```

(e) Erstellen Sie eine 100-elementige Zufallsstichprobe der χ^2_k-Verteilung, wobei die Anzahl der Freiheitsgerade $k = 10$ beträgt. Nutzen Sie die Funktion `rchisq()`. Berechnen Sie das Stichprobenmittel.

Lösung:

```
# Wir nutzen rchisq() um 100 Beobachtungen von X ~ Chi^2_10 zu erzeugen
# df ist die Anzahl der Freiheitsgrade
X <- rchisq(n = 100, df = 10)

# Stichprobenmittel berechnen mit mean()
mean(X)
```

```
[1] 10.20466
```

(f) Sei X binomialverteilt mit $n = 50$ und $p = 1/3$. Berechnen Sie $P(10 \leq X \leq 30)$.

Lösung:

```
# Siehe '?Binomial'. Wir nutzen pbinom() umd Wahrscheinlichkeiten für
# binomialverteilte Zufallsvariablen zu berechnen
pbinom(q = 30, size = 100, prob = 1/3) - pbinom(q = 9, size = 100, prob = 1/3)
```

```
[1] 0.2765538
```

Beachte, dass die Binomialverteilung eine diskrete Verteilung ist, d.h. im Algemeinen gilt

$$P(X < x) \neq P(X \leq x) \text{ bzw. } P(X > x) \neq P(X \geq x).$$

Aufgabe 3 – Daten einlesen und t-Test mit R

(a) Die Datei "`Daten.csv`" enthält 1000 Beobachtungen einer Variable X. Lesen Sie den Datensatz in `R` ein. *Hinweis:* Nutzen Sie hierfür die Funktion `read.csv2()`.

```
# Arbeitsverzeichnis abfragen
getwd()

# Arbeitsverzeichnis setzen
# (Beispielpfad)
setwd("Z:/RUebung")

# Daten anzeigen/einlassen
read.csv2("Daten.csv")

# Daten einlesen
Daten <- read.csv2("Daten.csv")
```

(b) Berechnen Sie mit `R` deskriptive Statistiken für den Datensatz und stellen Sie die Verteilung der Daten graphisch dar. Was fällt Ihnen auf?

```
# Übersicht über die Struktur mit str(), siehe ?str
str(Daten)
```

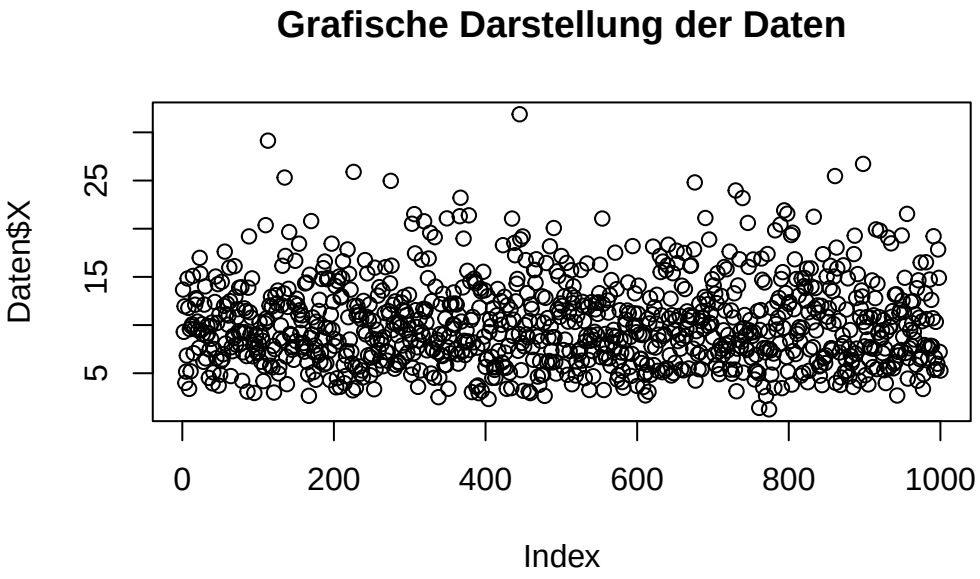
```
'data.frame':   1000 obs. of  2 variables:
 $ ID: int  1 2 3 4 5 6 7 8 9 10 ...
 $ X : num  13.68 9.3 11.95 4.01 5.17 ...
```

```
# Deskriptive Statistiken
summary(Daten$X)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.253   6.874   9.446  10.088  12.515  31.880
```

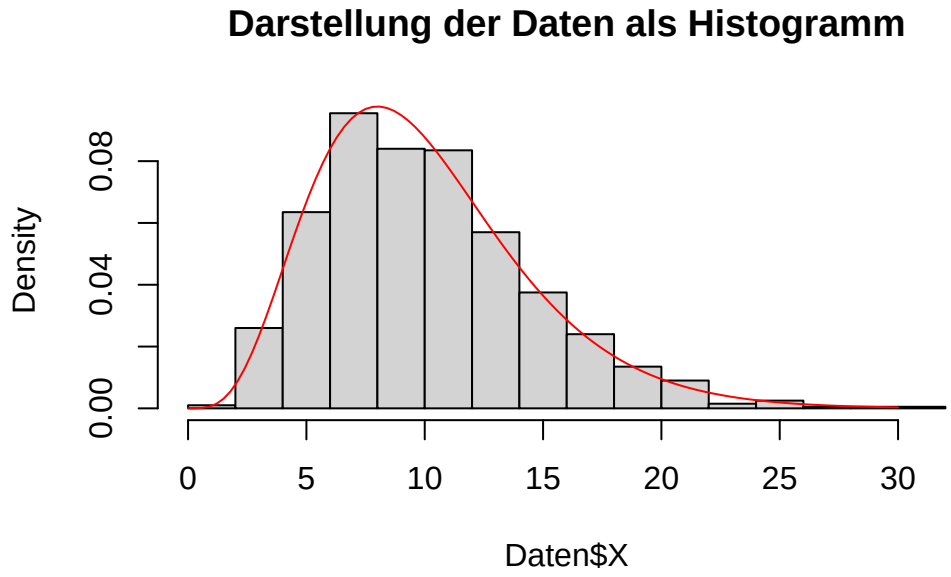
```
### Grafische Darstellung(en)

# Plot trägt X gegen den Index ab
plot(Daten$X,
     main = "Grafische Darstellung der Daten")
```




```r
# Besser: Histogramm. Mit freq=TRUE werden absolute Häufigkeiten an der
# Y-Achse abgetragen, bei freq=FALSE die (geschätzte) Dichte.
hist(Daten$X,
     main = "Darstellung der Daten als Histogramm",
     freq = FALSE
     )

# Das Histogramm lässt vermuten, dass X ~ Chi^2 verteilt ist.
# Wir sehen, dass eine Chi^2_10 Verteilung gut passt.
curve(dchisq(x, df = 10),
      from = 0,
      to = 30,
      add = T,
      col = "red"
      )
```



(c) Nutzen Sie die Funktion `t.test()` für den Hypothesentest von

$$H_0 : \mu_X = 10 \text{ vs. } H_1 : \mu_X \neq 10.$$

Welche weiteren Informationen finden Sie im Output von `t.test()`? Erläutern Sie kurz.

```
# Aufgrund der Plots vermuten wir, dass mu = 10 ist.
# Wir nutzen die Funktion t.test() um zu testen, ob mu = 10 ist.
?t.test

# t.test berechnet standardmäßig den p-Wert für einen Test von
# H_0 gegen eine beidseitige Alternativhypothese.
t.test(Daten$X, mu = 10)
```

```
    One Sample t-test

data:  Daten$X
t = 0.62433, df = 999, p-value = 0.5326
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
  9.811313 10.364774
sample estimates:
mean of x
 10.08804
```

Wir vergleichen den p-Wert mit dem 5% Signifikanzniveau: Da `p-value` > 0.05 wird $H_0 : \mu_X = 10$ zum 5%-Niveau beibehalten: Wir haben nicht genug statistische Evidenz in den Daten, um H_0 abzulehnen.

Der Output enthält außerdem ein 95%-Konfidenzintervall für μ_X sowie den Stichprobenmittelwert.

Aufgabe 4 – Normalverteilungsdichte als R-Funktion definierem

Die Normalverteilung besitzt die Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

(a) Implementieren Sie diese Funktion als R-Funktion mit dem Namen `fnorm`.

Lösung:

```
fnorm <- function(x, mean, sd) {
  1/(sqrt(2 * pi) * sd) * exp(-0.5 * (x-mean)^2/sd^2)
}
```

(b) Überprüfen Sie, ob Ihre Funktion für $\mu = 0$ und $\sigma = 1$ die Dichte an den Stellen $-1.96, 0$ und 1.96 korrekt berechnet. Benutzen Sie hierzu eine geeignete Funktion der Familie `Normal`, siehe `?Normal`. Welche Quantile sind $-1.96, 0$ und 1.96?

Lösung:

Wir berechnen die gesuchten Dichte-Werte mit der in (a) definierten Funktion `fnorm()` und vergleichen mit dem Ergebnissen von `dnorm()`

Anhand von `fnorm()`:

```r
fnorm(x = 0, mean = 0, sd = 1)
fnorm(x = -1.96, mean = 0, sd = 1)
fnorm(x = 1.96, mean = 0, sd = 1)
```

```
[1] 0.3989423
[1] 0.05844094
[1] 0.05844094
```

Anhand von `dnorm()`:

```r
dnorm(x = 0, mean = 0, sd = 1)
dnorm(x = -1.96, mean = 0, sd = 1)
dnorm(x = 1.96, mean = 0, sd = 1)
```

```
[1] 0.3989423
[1] 0.05844094
[1] 0.05844094
```

Offenbar stimmen die Werte überein.

Wir können mit `pnorm()` überprüfen welche Quantile -1.96, 0 und 1.96 sind.

Hinweis: Für das α-Quantil Q_α der Verteilung von X gilt $P(X \leq Q_\alpha) = \alpha$.

```r
pnorm(-1.96)
pnorm(0)
pnorm(1.96)
```

```
[1] 0.0249979
[1] 0.5
[1] 0.9750021
```

Aufgrund der Symmetrie der N(0,1)-Verteilung ist `pnorm(-1.96) = 1 - pnorm(1.96)`. Außerdem liegen links und rechts von 0 jeweils 50% Wahrscheinlichkeitsmasse, d.h. 0 ist das 50%-Quantil.

(c) Ziehen Sie 1000 Zufallszahlen aus der Standardnormalverteilung und berechnen Sie geläufige deskriptive Statistiken für Ihre Stichprobe.

Hinweis zu (a): Der folgende Code definiert in `R` die Dichtefunktion der *Standardnormalverteilung* als Funktion `f`.

```r
f <- function(x) {
  1/(sqrt(2 * pi)) * exp(-0.5 * x^2)
}
```

Lösung:

Wir generieren Zufallszahlen mit `rnorm()` und berechnen zusammenfassende Statistiken mit `summary()`.

```r
# Zufallsstichprobe generieren
X <- rnorm(n = 1000)
summary(X)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.46932 -0.70474  0.01047 -0.02879  0.59586  2.86786
```