Lennart Oelschläger (lennart.oelschlaeger@uni-bielefeld.de)
310214 Praktische Übung zu "Einführung in die Mikroökonometrie"
Wintersemester 2023/2024

Aufgabenblatt 0

Lösungen

Organisation

Die Praktische Übung zu "Einführung in die Mikroökonometrie" findet zweiwöchentlich statt. Die Termine und Räume finden Sie im eKVV. Ich lade vor den Treffen das aktuelle Aufgabenblatt in Moodle hoch. Während der Treffen haben Sie Zeit an den Aufgaben zu arbeiten, im Anschluss besteht die Möglichkeit die Lösungen gemeinsam zu besprechen.

Sie können eine Studienleistung für 31-M23 oder 31-SW-StatM erwerben. Dazu analysieren Sie einen von mir bereitgestellten Datensatz mit den Methoden aus der Vorlesung.

- Wenn Sie eine Studienleistung erwerben möchten, müssen Sie mir das bis zum 10.11.2023 durch eine Nachricht im Forum mitteilen.
- Zum 20.11.2023 erhalten Sie dann einen Datensatz mit Fragestellungen von mir. Die Fragestellungen umfassen eine Beschreibung der Daten, eine Formulierung geeigneter Methoden zur Analyse der Daten und ein konkretes Modellierungsziel.
- Ihre Bearbeitung der Fragestellungen ist in zwei Teile aufgeteilt, die Sie jeweils im Moodle als einzelne .pdf Datei hochladen.
  1. Bis zum 22.12.2023 geben Sie Ihre Beschreibung der Daten ab.
  2. Bis zum 12.01.2024 geben Sie Ihre Formulierung der Methoden und Modellierung ab.
- Am 29.01.2024 gibt es eine kleine Abschlussdiskussion der Abgaben.

Aufgaben

Die folgenden Aufgaben sind eine Wiederholung einiger Konzepte aus Mathematik, Wahrscheinlichkeitstheorie und Statistik, die in der Vorlesung eingesetzt werden. Außerdem wiederholen wir das Programmieren in `R` und das Erstellen von Berichten in `R Markdown`.

1. Die Dokumentation *simpleR* von John Verzani liefert eine gute Einführung in `R`, Sie finden das Dokument unter http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf. Lesen Sie *Section 2: Data*, um zu lernen, wie Datensätze mit der `c()` Funktion eingegeben werden können. Erzeugen Sie anschließend folgende zwei Vektoren in `R`:

$$x^\top = \begin{bmatrix} 10 & 8 & 13 & 9 & 11 & 14 & 6 & 4 & 12 & 7 & 5 \end{bmatrix}$$
$$y^\top = \begin{bmatrix} 8.1 & 6.9 & 7.5 & 8.8 & 8.3 & 9.9 & 7.2 & 4.2 & 10.8 & 4.8 & 5.6 \end{bmatrix}$$

```
R> x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
R> y <- c(8.1, 6.9, 7.5, 8.8, 8.3, 9.9, 7.2, 4.2, 10.8, 4.8, 5.6)
```

2. Welche Operation wird durch `x + y` ausgeführt? Wenden Sie auch `-, *, /, %*%` an.

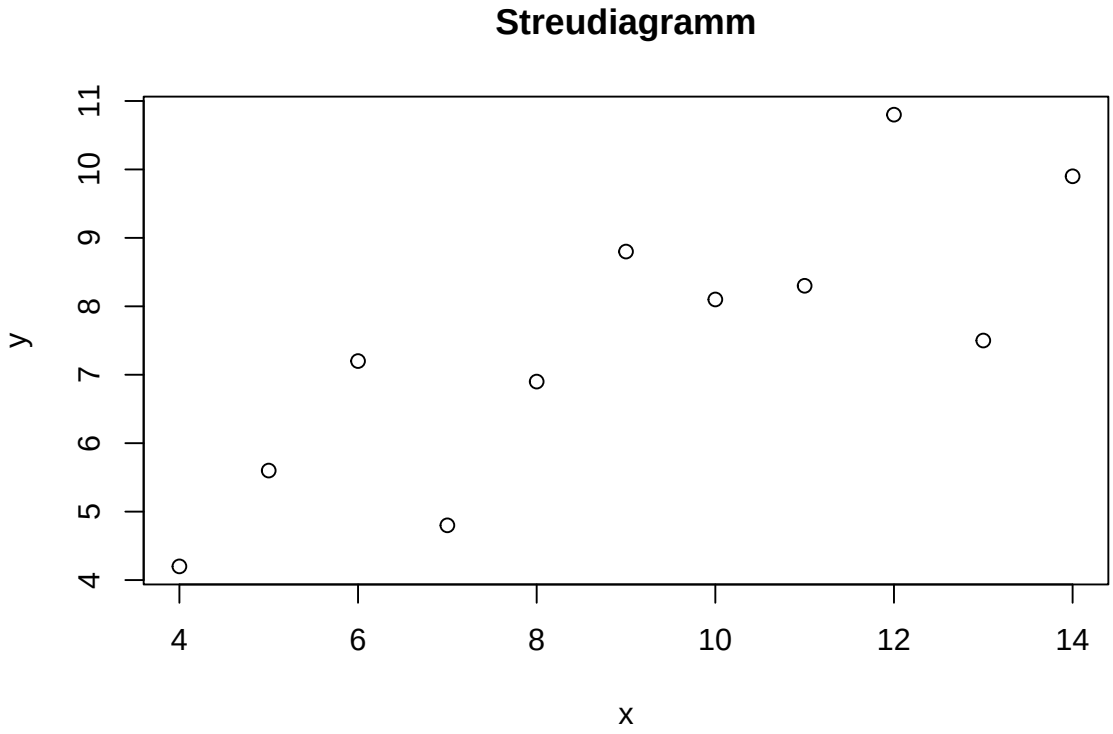
```
R> x + y
```

```
#  [1] 18.1 14.9 20.5 17.8 19.3 23.9 13.2  8.2 22.8 11.8 10.6
R> x - y
#  [1]  1.9  1.1  5.5  0.2  2.7  4.1 -1.2 -0.2  1.2  2.2 -0.6
R> x * y
#  [1]  81.0  55.2  97.5  79.2  91.3 138.6  43.2  16.8 129.6  33.6  28.0
R> x / y
#  [1] 1.2345679 1.1594203 1.7333333 1.0227273 1.3253012 1.4141414 0.8333333
#  [8] 0.9523810 1.1111111 1.4583333 0.8928571
R> x %*% y
#      [,1]
# [1,]  794
R> t(x) %*% y
#      [,1]
# [1,]  794
R> # x %*% t(y)
R> # t(x) %*% t(y)
```

`x + y` ist die elementweise Addition. Analog sind `-, *, /` die elementweise Subtraktion, Multiplikation und Division. `x %*% y` ist das Matrikprodukt $x^\top y$.

3. Die Einträge der beiden Vektoren können wir als Wertepaare $(x_1, y_1), \ldots, (x_{11}, y_{11})$ betrachten. Erzeugen Sie mit `plot()` ein Streudiagramm dieser Daten.

```
R> plot(x, y, main = "Streudiagramm")
```



4. Lesen Sie *Section 5: Multivariate Data* und lernen Sie den Datentyp `data.frame` kennen. Erstellen Sie einen `data.frame` aus `x` und `y`.

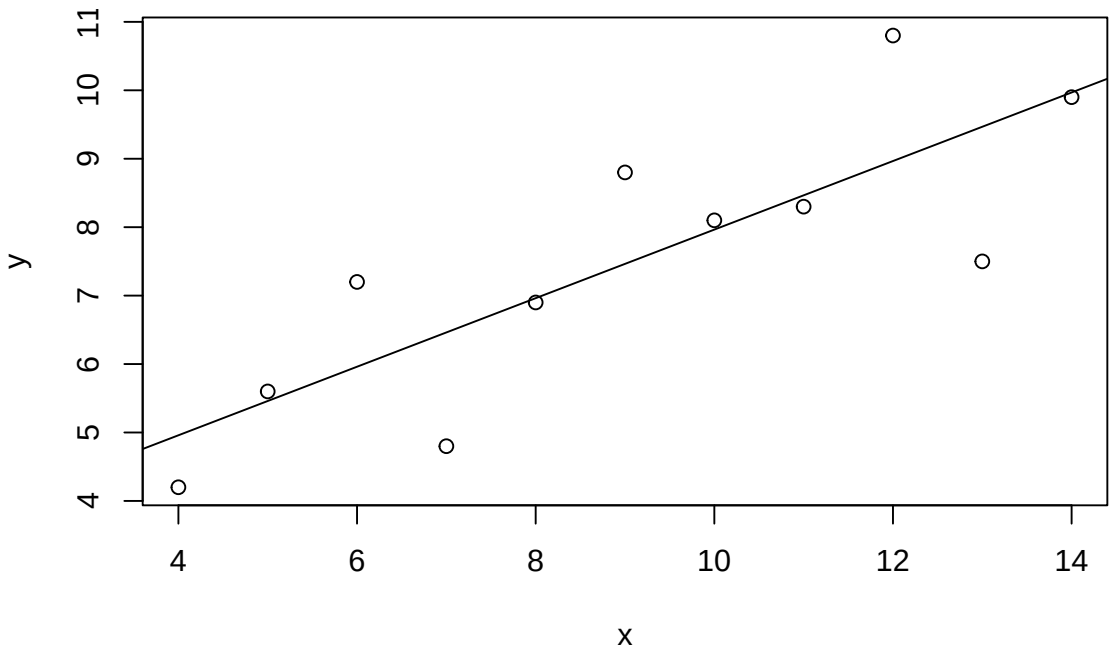
```
R> data <- data.frame(x = x, y = y)
R> data
```

```
#     x    y
# 1  10  8.1
# 2   8  6.9
# 3  13  7.5
# 4   9  8.8
# 5  11  8.3
# 6  14  9.9
# 7   6  7.2
# 8   4  4.2
# 9  12 10.8
# 10  7  4.8
# 11  5  5.6
```

5. Schätzen Sie die Koeffizienten β_0 und β_1 des linearen Modells $y = \beta_0 + \beta_1 x + u$ mithilfe der Funktion `lm()` (steht für *lineares Modell*). Erklärungen finden Sie in *Section 13: Regression Analysis*. Zeichnen Sie dann mithilfe von `abline()` die angepasste Regressionsgerade in Ihr Streudiagramm ein.

```
R> model <- lm(formula = y ~ x, data = data)
R> plot(data)
```

```
R> abline(model)
```



6. Erklären Sie den Output von `summary()`, angewendet auf Ihr geschätztes Modell.

```
R> summary(model)

#
# Call:
# lm(formula = y ~ x, data = data)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.96727 -0.46227 -0.06273  0.68955  1.83364
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   2.9555     1.1330   2.609  0.02834 *
# x             0.5009     0.1188   4.217  0.00225 **
# ---
# Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
#
# Residual standard error: 1.246 on 9 degrees of freedom
# Multiple R-squared:  0.664,   Adjusted R-squared:  0.6267
# F-statistic: 17.79 on 1 and 9 DF,  p-value: 0.002248
```

Eine Erklärung des Outputs liefert https://www.youtube.com/watch?v=NEfjirpOj7s.

7. `R Markdown` ist eine Kombination aus `R`, unserer Programmiersprache für Datenanalyse, und `Markdown`, einer einfachen Auszeichnungssprache für die Erstellung von Berichten. Durch die Kombination können wir `R` Code direkt in Textdokumente integrieren und mit minimalem Aufwand dynamische Analyseberichte generieren. Entwickelt wurde `R Markdown` 2014 von Yihui Xie, der einen einfachen Weg gesucht hat, seine `R` Hausaufgaben aufzuschreiben. Das Buch *R for Data Science* von Hadley Wickham bietet unter anderem eine gute Einführung in `R Markdown`. Das Buch ist kostenlos online unter https://r4ds.had.co.nz/ verfügbar. Lesen Sie *Chapter 27: R Markdown*, um zu lernen, wie ein `R Markdown` Dokument erstellt werden kann. Verwenden Sie die Vorlage aus Abschnitt 27.2 und erstellen Sie selbst ein `R Markdown` Dokument im `.html`-Format.

Dazu in R Studio links oben `File > New File > R Markdown` wählen. Es öffnet sich ein Fenster, dort links unten auf `Create Empty Document` klicken. Es öffnet sich eine leere Datei, dort die Vorlage einfügen. Anschließend auf `Knit` klicken.

8. Fügen Sie in das `R Markdown` Dokument Ihre Lösungen aus Aufgabe 1 ein.

````
---
title: "Aufgabenblatt 0"
date: "13.10.2023"
author: "Lennart"
output: html_document
---
Aufgabe 1

```{ r }
x <- c(10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5)
y <- c(8.1, 6.9, 7.5, 8.8, 8.3, 9.9, 7.2, 4.2, 10.8, 4.8, 5.6)
```
````

9. Wie können Sie ein `.pdf`-Dokument erstellen?

Indem `output: html_document` durch `output: pdf_document` ersetzt wird.

10. Wie wird das multiple lineare Regressionsmodell definiert, und wie können die Modellparameter geschätzt werden?

Eine Variable y wird mit k Regressoren x_k durch $y = \sum_{j=1}^{k} x_j \beta_j + u$ erklärt. Der u Term beinhaltet den Modellfehler, die β_j's sind Modellparameter. Es sollen konkrete Werte für die Parameter gefunden werden, die am besten zu gegebenen Daten passen. Die populärste Methode ist der Kleinste-Quadrate-Schätzer: Er bestimmt solche Schätzwerte, sodass die Summe der quadrierten Residuen minimal ist.

11. Welche Eigenschaften hat der Kleinste-Quadrate Schätzer, und welche Voraussetzungen müssen dafür erfüllt sein?

Unter den Annahmen MLR.1 (lineares Modell), MLR.2 (Zufallsstichprobe), MLR.3 (Information in den Regressoren, keine Multikollinearität) und MLR.4 (bedingte Fehlererwartung ist Null) ist der KQ-Schätzer erwartungstreu. Gilt zusätzlich MLR.5 (Homoskedastie), so ist die Varianz des KQ-Schätzers die minimale Varianz unter allen linearen und erwartungstreuen Schätzern (Gauss-Markov Theorem). Gilt auch MLR.6 (Fehler sind normalverteilt), so ist der KQ-Schätzer normalverteilt. Unter einer technischen Annahme an die Daten ist er konsistent.

12. Bitte prognostizieren Sie den y Wert für $x_1 = 6$ und $x_2 = -1$, gegeben die Daten

$$\begin{aligned} x_1 &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 1 & 2 & 3 & 4 & 5 \end{pmatrix}, \\ x_2 &= \begin{pmatrix} 1 & 1 & 2 & 2 & 3 & 3 & 4 & 4 & 5 & 5 \end{pmatrix}, \\ y &= \begin{pmatrix} 1 & 3 & 1 & 5 & 2 & -3 & -3 & -1 & -2 & -2 \end{pmatrix}. \end{aligned}$$

```
R> x_1 <- c(1, 2, 3, 4, 5, 1, 2, 3, 4, 5)
R> x_2 <- c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5)
R> y <- c(1, 3, 1, 5, 2, -3, -3, -1, -2, -2)
R> data <- data.frame(y, x_1, x_2)
R> model <- lm(y ~ x_1 + x_2, data)
R> model$coefficients %*% c(1, 6, -1)
```

```
#          [,1]
# [1,] 11.96667
```

13. Die Schlusskurse von Bitcoin und Ethereum aus 2022 haben eine Kovarianz von 7718820. Kann daraus ein starker, positiver Zusammenhang geschlossen werden?

Eine positive Kovarianz zeigt einen positiven Zusammenhang an. Sie ist aber nicht normiert, daher bleibt die relative Stärke des Zusammenhanges unbekannt.

```
R> btc <- fHMM::download_data("BTC-USD", from = "2022-01-01", to = "2022-12-31")
R> eth <- fHMM::download_data("ETH-USD", from = "2022-01-01", to = "2022-12-31")
R> cov(btc$Close, eth$Close)
```

```
# [1] 7718820
```

```
R> cor(btc$Close, eth$Close)
```

```
# [1] 0.9759673
```

14. Finden Sie ein Beispiel, dass im Allgemeinen $\mathrm{Var}(X + Y) \neq \mathrm{Var}(X) + \mathrm{Var}(Y)$ für zwei Zufallsvariablen X und Y gilt.

Sei $X = Y \sim N(0, 1)$. Dann ist $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, aber $\mathrm{Var}(X + Y) = 4$.

15. Finden Sie zwei Zufallsvariablen, die unkorreliert, aber nicht unabhängig sind.

Wähle X, Y unabhängig mit $\Pr(X = 0) = \Pr(X = 1) = 0.5$, $\Pr(Y = -1) = \Pr(Y = 1) = 0.5$ und definiere $Z := XY$. Dann ist $\mathrm{Cov}(Z, X) = \mathrm{E}(Z(X - 0.5)) = \mathrm{E}(X^2 - 0.5X)\mathrm{E}(Y) = 0$ aber

$\Pr(Z = 1 \mid X = 0) = 0 \neq 0.5 = \Pr(Z = 1 \mid X = 1)$.

16. Finden Sie zwei reelle Vektoren x und y, jeweils der Länge 10, sodass die empirische Korrelation $\widehat{\mathrm{Cor}}(x, y)$ exakt -1 bzw. 0 bzw. $+1$ beträgt.

```
R> x <- 1:10
R> y <- x
R> cor(x, y)
# [1] 1
R> y <- -x
R> cor(x, y)
# [1] -1
R> y <- (x - mean(x))^2
R> cor(x, y)
# [1] 0
```

17. Es seien X, Y zwei unabhängige Zufallsvariablen, die mit gleicher Wahrscheinlichkeit die Werte in $\{1, 2, 3\}$ annehmen. Berechnen Sie $\mathrm{E}(1 + 4X + 2Y \mid X = 2)$.

$\mathrm{E}(1 + 4X + 2Y \mid X = 2) = 9 + 2\mathrm{E}(Y) = 13$

18. Ein Losverkäufer behauptet, dass mindestens 20% seiner Lose Gewinne seien. Die Käufer aber vermuten, dass der Anteil geringer ist. Es werden $n = 100$ Lose überprüft. Führen Sie einen statistischen Test zum Signifikanzniveau $\alpha = 5\%$ zur Streitschlichtung durch, wobei Sie die Aussage des Losverkäufers als Nullhypothese wählen.

Der Anteil an Gewinnlosen sei binomialverteilt zu $n = 100$ und $p = 20\%$. Die Wahrscheinlichkeit, höchstens 14 Erfolge zu haben, beträgt somit 8%, und höchstens 13 Erfolge zu haben, 4.7%. Zu $\alpha = 5\%$ geben wir also dem Losverkäufer recht, falls in der Stichprobe mindestens 14 Gewinne sind, ansonsten den Käufern. Siehe `https://www.youtube.com/watch?v=MBf9Iin6bpg` für mehr Details.

```
R> pbinom(13, size = 100, prob = 0.2)
# [1] 0.04691224
R> pbinom(14, size = 100, prob = 0.2)
# [1] 0.08044372
R> binom.test(13, n = 100, p = 0.2, alternative = "less")
#
#   Exact binomial test
#
# data:  13 and 100
```

```
# number of successes = 13, number of trials = 100, p-value = 0.04691
# alternative hypothesis: true probability of success is less than 0.2
# 95 percent confidence interval:
#  0.0000000 0.1987198
# sample estimates:
# probability of success
#                   0.13
```

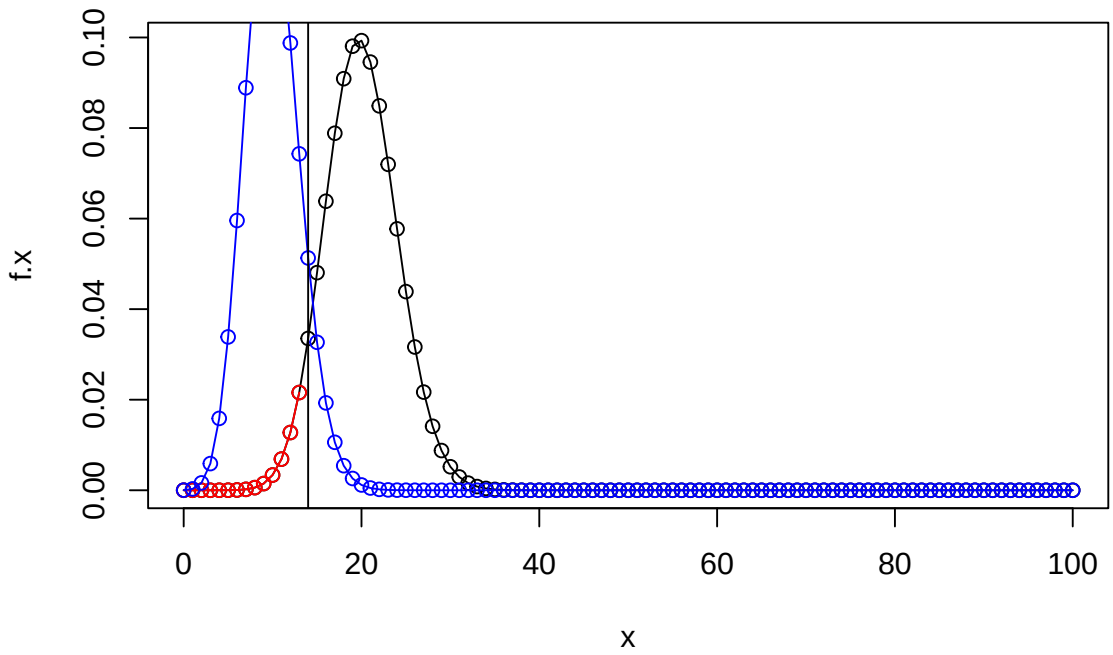
```
R> binom.test(14, n = 100, p = 0.2, alternative = "less")
```

```
#
#   Exact binomial test
#
# data:  14 and 100
# number of successes = 14, number of trials = 100, p-value = 0.08044
# alternative hypothesis: true probability of success is less than 0.2
# 95 percent confidence interval:
#  0.0000000 0.2101739
# sample estimates:
# probability of success
#                   0.14
```

19. Angenommen, der wahre Anteil an Gewinnlosen beträgt nur 10%. Berechnen Sie die Wahrscheinlichkeit, mit der Ihr Test dem Losverkäufer fälschlicherweise recht gibt.

Das ist die Wahrscheinlichkeit, unter der Binomialverteilung mit $n = 100$ und $p = 10\%$ mindestens 14 Erfolge zu beobachten. Sie beträgt 12.4%.

```
R> f.x <- function(x, p) dbinom(x, size = 100, prob = p)
R> plot(0:100, f.x(x = 0:100, p = 0.2), type = "o", xlab = "x", ylab = "f.x")
R> lines(0:13, f.x(x = 0:13, p = 0.2), type = "o", col = "red")
R> lines(0:100, f.x(x = 0:100, p = 0.1), type = "o", col = "blue")
R> abline(v = 14)
```




```
R> 1 - pbinom(13, size = 100, prob = 0.1)

# [1] 0.1238768
```

20. Sie möchten die in b) berechnete Wahrscheinlichkeit auf unter 5% reduzieren. Auf welche Werte müssen Sie dafür entweder α oder n verändern?

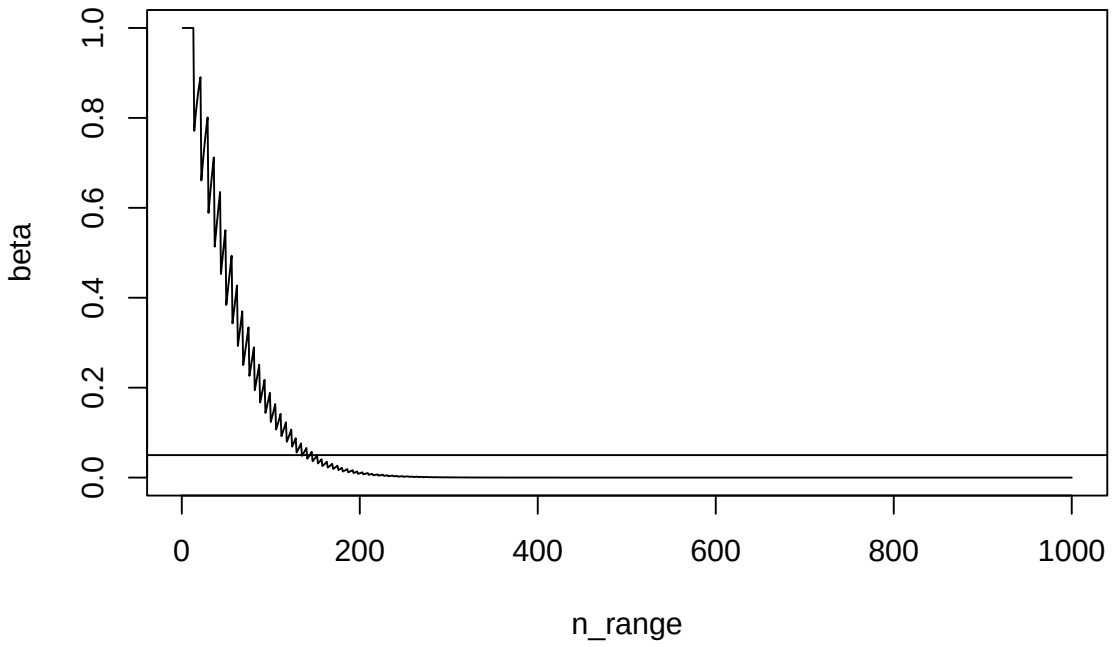
```
R> pbinom(15, size = 100, prob = 0.2)

# [1] 0.1285055

R> 1 - pbinom(15, size = 100, prob = 0.1)

# [1] 0.03989053

R> n_range <- 1:1000
R> crit <- qbinom(0.05, size = n_range, prob = 0.2)
R> beta <- 1 - pbinom(crit - 1, size = n_range, prob = 0.1)
R> plot(n_range, beta, type = "l")
R> abline(h = 0.05)
```




```
R> min(which(beta < 0.05))
# [1] 135
```

Entweder α auf mindestens 13% oder n auf mindestens 135 erhöhen.

21. Es seien X_1 und X_2 zwei unabhängige Zufallsvariablen mit Varianz 1. Wie lautet die Kovarianzmatrix von $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ X_2 \end{pmatrix}$?

Es gilt

$$\operatorname{Cov} \begin{pmatrix} X_1 + X_2 \\ X_2 \end{pmatrix} = \begin{pmatrix} \operatorname{Var}(X_1 + X_2) & \operatorname{Cov}(X_1 + X_2, X_2) \\ \operatorname{Cov}(X_2, X_1 + X_2) & \operatorname{Var}(X_2) \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

da $\operatorname{Var}(X_1 + X_2) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2)$ und $\operatorname{Cov}(X_1 + X_2, X_2) = \operatorname{Cov}(X_2, X_1 + X_2) = \operatorname{Cov}(X_2, X_1) + \operatorname{Cov}(X_2, X_2) = 0 + 1$ (wegen Unabhängigkeit) ist.

Alternativ können wir die Rechenregel $\operatorname{Cov}(AX) = A\operatorname{Cov}(X)A^\top$ verwenden:

$$\operatorname{Cov} \left(\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \operatorname{Cov} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

Hier ist wieder wegen Unabhängigkeit $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

22. Betrachten Sie die folgende Kontingenztabelle, die drei Alterskategorien und die Präferenz für Tee oder Kaffee in einer Gruppe von 30 Personen zeigt:

	Tee	Kaffee
jung	14	1
mittelalt	2	8
älter	3	2

- Erstellen Sie die Randverteilung der Alterskategorien.

  $\Pr(\text{jung}) = 15/30$, $\Pr(\text{mittelalt}) = 10/30$, $\Pr(\text{älter}) = 5/30$

- Finden Sie die bedingte Verteilung der Getränkepräferenz gegeben das Alter.

  $\Pr(\text{Tee} \mid \text{jung}) = 14/15$, $\Pr(\text{Tee} \mid \text{mittelalt}) = 2/10$, $\Pr(\text{Tee} \mid \text{älter}) = 3/5$ (Kaffee analog)

- Berechnen Sie die gemeinsame Verteilung von Alter und Getränkepräferenz.

  Dividiere die Tabellenelemente durch 30, zum Beispiel $\Pr(\text{Tee}, \text{jung}) = 14/30$.