# Blatt 4

Data Science 2

Sommersemester 2023

*Hinweis:* Halten Sie Ihre Lösungen (insbesondere R-Code) hier und auch bei allen künftigen Aufgaben kurz, verständlich und übersichtlich.

## Aufgabe 1:

Sei $X \sim Exp(100)$. Sei $f(x) := 10x^2 + 3$ für $x \in \mathbb{R}$.

a) Ermitteln Sie den Erwartungswert sowie die Varianz der Zufallsvariablen $f(X)$ mittels Monte-Carlo-Simulation, indem Sie 10 000 voneinander unabhängige Realisationen der Zufallsvariable $X$ generieren.

b) Ermitteln Sie den Erwarungswert von $f(X)$ analytisch. Wie groß ist die Abweichung zum in a) ermittelten Erwartungswert?

c) Stellen Sie die Dichte von $f(X)$ mittels Monte-Carlo-Simulation anhand von 10 000 voneinander unabhängigen Realisationen der Zufallsvariable $X$ in einem Histogramm dar.

d) Bestimmen Sie die Dichte von $f(X)$ analytisch. Berechnen Sie den Unterschied zur simulierten Verteilungsunktion. Zur Berechnung des Unterschieds nutzen Sie bitte die Formel der mittleren absoluten Abweichung: $\frac{1}{n}\sum_{k=1}^{n}|g(x_k) - h(x_k)|$ für zwei Wahrscheinlichkeitsfunktionen $g$ und $h$.

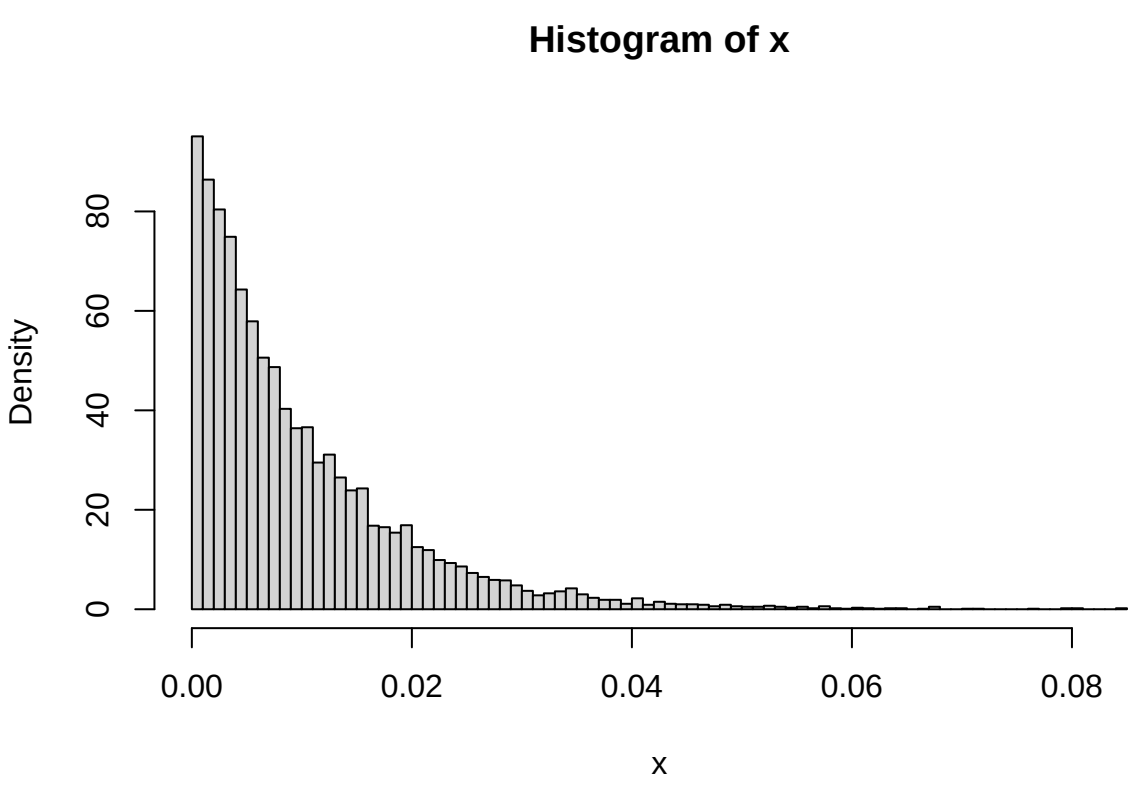
*Hinweis:* -

### Lösung

### zu a)

Wir erzeugen zunächst die benötigten Zufallsvariablen und vergewissern uns der korrekten Erzeugung mittels eines Histogramms.

```
n <- 10^4
lambda <- 100
x <- rexp(n, lambda)
hist(x,breaks=100,probability=TRUE)
```




Wir erkennen die korrekte Verteilung und fahren fort. Nun berechnen wir Erwartungswert und Varianz von $f(X)$ auf Basis der erzeuten Zufallsvariablen:

```
f <- function(x){
  10*x^2+3
}
print(paste("Es gelten E(f(X))=", mean(f(x)), "und Var(f(X)) =", var(f(x))))
```

```
## [1] "Es gelten E(f(X))= 3.00192035112541 und Var(f(X)) = 1.85944883103751e-05"
```

### zu b)

Wir setzen die Formel für $f$ ein und nutzen die Linearität des Erwarungswerts aus.

$$E(f(X)) = E(10X^2 + 3) = 10E(X^2) + 3$$

Wir wissen, dass $Var(X) = E(X^2) - (E(X))^2$ gilt. Da wir Varianz und Erwarungswert der Exponentialverteilung bereits kennen, nutzen wir diese Gleichung in Umgestellter Form:

$$E(X^2) = Var(X) + (E(X))^2 = \frac{1}{\lambda^2} + \left(\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda^2}$$

für $X \sim Exp(\lambda)$

Wir schlussfolgern

$$E(f(X)) = 10 \cdot \frac{2}{\lambda^2} + 3 = 20\lambda^2 + 3$$

Mit $\lambda = 100$ ergibt sich:

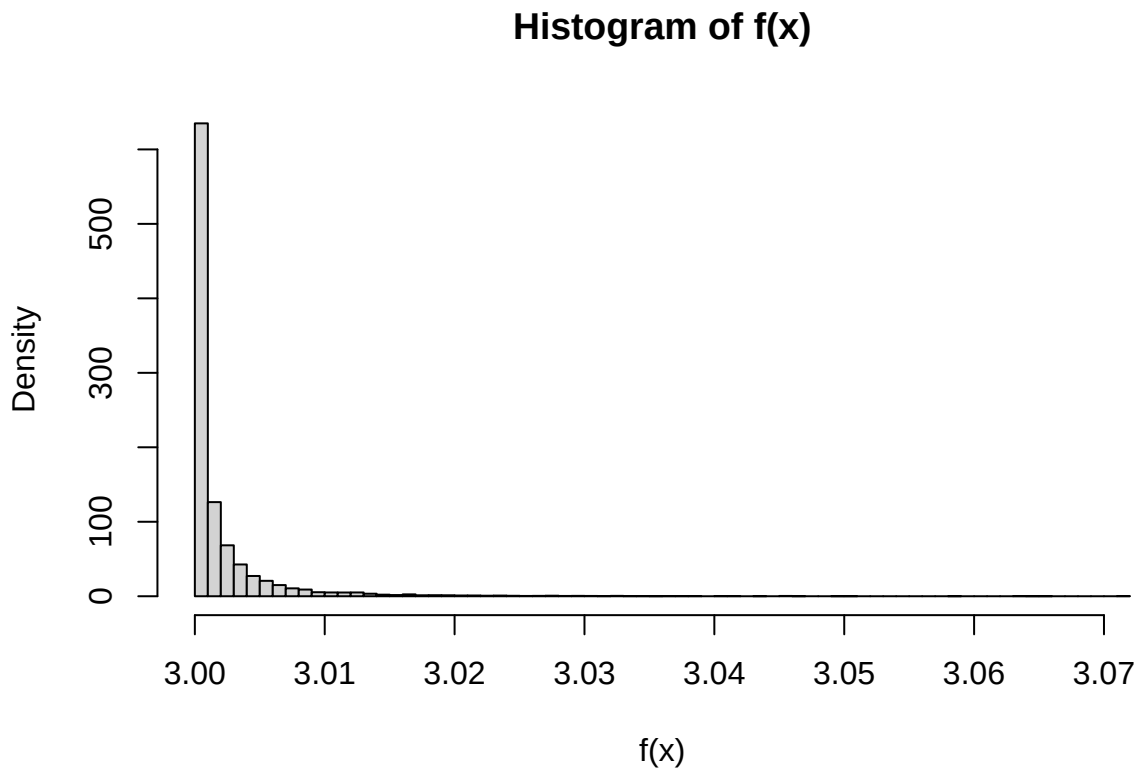
```
result <- 20*lambda^(-2)+3
print(paste("E(f(X)) =", result))
```

```
## [1] "E(f(X)) = 3.002"
```

### zu c)

Wir stellen die Wahrscheinlichkeitsfunktion von $f(X)$ mittels Monte-Carlo-Integration in einem Histogramm dar.

```
hist(f(x), breaks=100, probability=TRUE)
```




### zu d) <- ueberarbeiten!

Sei $ x$ eine reelle Zahl. Es gilt:

$$P(F(X) = x) = P(10X^2 + 3 = x) = P(X = \sqrt{\frac{x-3}{10}})$$

In der letzten Gleichheit haben wir ausgenutzt, dass $P(X \geq 0) = 1$ gilt, da $X \sim Po(\lambda)$ ist. Könnte $X$ auch negative Werte annehmen, so müssten wir die negative Wurzel entsprechend ergänzen.

Nun gilt entsprechend der Poissonverteilung falls

$$x \in M := \{10 \cdot z^2 + 3 : z \in \mathbb{N}_0\}$$

liegt, so gilt

$$P(F(X) = x) = \exp(-\lambda)\lambda^{\sqrt{(x-3)/10}} \cdot \frac{1}{(\sqrt{\frac{x-3}{10}})!}$$

. Hierbei sei erwähnt, dass $x \in M$ bedeutet, dass $\sqrt{\frac{x-3}{10}} \in \mathbb{N}_0$ liegt.

Falls $x \notin M$ liegt, gilt $P(F(X) = x) = 0$

# Aufgabe 2:

Sei $X_n \sim B(n, p)$ mit $n = 10^4$ und $p = 10^{-4}$.

a) Bestimmen Sie die Verteilung von $\lim_{n\to\infty} \frac{X_n - np}{\sqrt{np(1-p)}}$ analytisch. Durch welche Verteilung könnte man die Verteilung von $X_n$ also approximieren?
b) Ermitteln Sie die Wahrscheinlichkeitsfunktion der Zufallsvariablen $X_n$ mittels Monte-Carlo-Simulation unter Zuhilfenahme von 10 000 Realisationen.
c) Erstellen Sie ein Histogramm zur simulierten Wahrscheinlichkeitsfunktion und ergänzen Sie die Dichte einer normalverteilten Zufallsvariable mit entsprechend angepassten Paramteren.
d) Erstellen Sie ein Histogramm zur simulierten Wahrscheinlichkeitsfunktion und ergänzen Sie die Wahrscheinlichkeitsfunktion einer poissonverteilten Zufallsvariable mit entsprechend angepasstem Parameter.

*Hinweis:* -

### Lösung

---

# Aufgabe 3

Sie verfolgen täglich den Gewinn der Deutschen Lotteriegesellschaft. Dieser hat aktuell den Wert 0. Entwerfen Sie R-Code um einen beispielhaften Verlauf des Gewinns für die nächsten 10 Tage entsprechend der folgenden Bedingungen aufzuzeigen:

a. Die Änderungen des Gewinns seien standardnormalverteilt und voneinander unabhängig.

b. Die erste Änderung $X_1$ sei standardnormalverteilt. Für die übrigen Änderungen gelte: $X_{i+1}$ ist $N(x_i, 25)$-verteilt. Hierbei entspreche $x_i$ dem Wert, den die Zufallsvariable $X_i$ angenommen hat.

c. Die erste Änderung $X_1$ sei standardnormalverteilt. Für die übrigen Änderungen gilt: $X_{i+1}$ ist $N(x_1 + x_2 + \ldots + x_i, 25)$-verteilt.

*Hinweis:* -

### Lösung

--

# Aufgabe 4: Pi raten

Eine Möglichkeit die Kreiszahl $\pi$ zu schätzen besteht darin sehr viele (hier $n$) Kugeln unabhängig voneinander uniformverteilt in einen Sandkasten mit quadratischer Grundfläche fallen zu lassen. Wir notieren die Koordinaten $(x, y)$ der Aufprallpunkte der Kugeln und ermitteln, wie groß der Anteil der in den Innenkreis des Quadrats gefallenen Kugeln an den insgesamt fallen gelassenen Kugeln ist. Eine Kugel ist in den Innenkreis des Quadrats gefallen, falls ihr Aufprallpunkt der Kreisvorschrift $(x - m_1)^2 + (y - m_2)^2 <= (d/2)^2$ genügt. Hierbei entspricht $(m_1, m_2)$ dem Mittelpunkt des Sandkastens und $d$ der Seitenlänge des Sandkastens. Die Idee hinter dem Vorgehen besteht darin auszunutzen, dass $P(K) = \frac{\pi(d/2)^2}{d^2}$ mit $K$="Der erste Aufprallpunkt erfüllt die Kreisvorschrift" gilt. Außerdem nutzen wir, dass für sehr viele Kugeln $P(K)$ dem oben ermittelten Anteil näherungsweise entspricht.

a. Modellieren Sie die Situation mathematisch.

b. Approximieren Sie $\pi$, indem Sie den oben erläuterten Algorithmus implementieren. Erläutern Sie Ihr Vorgehen und geben Sie Zwischenergebnisse aus. Fixieren Sie dazu $n = 10^4, m_1 = 0.5, m_2 = 0.5$ und $d = 1$.

c. Ändern Sie den Algorithmus dahingehend, dass so lange Kugeln fallen gelassen werden bis die Abweichung vom in R integrierten Wert für $\pi$ kleiner als $10^{-4}$ ist.

d. Wo und wie wurde in dieser Aufgabe das Gesetz der großen Zahl angewendet?

*Hinweis:* -

## Lösung

### zu a)

Seien $X_1, X_2, \ldots \sim U[-\frac{d}{2} + m_1, m_1 + \frac{d}{2}]$ sowie $Y_1, Y_2, \ldots \sim U[-\frac{d}{2} + m_2, m_2 + \frac{d}{2}]$ allesamt voneinander unabhängige Zufallsvariablen. Dann entspricht $(X_i, Y_i)$ dem Aufprallpunkt der $i$-ten Kugel. Somit gilt

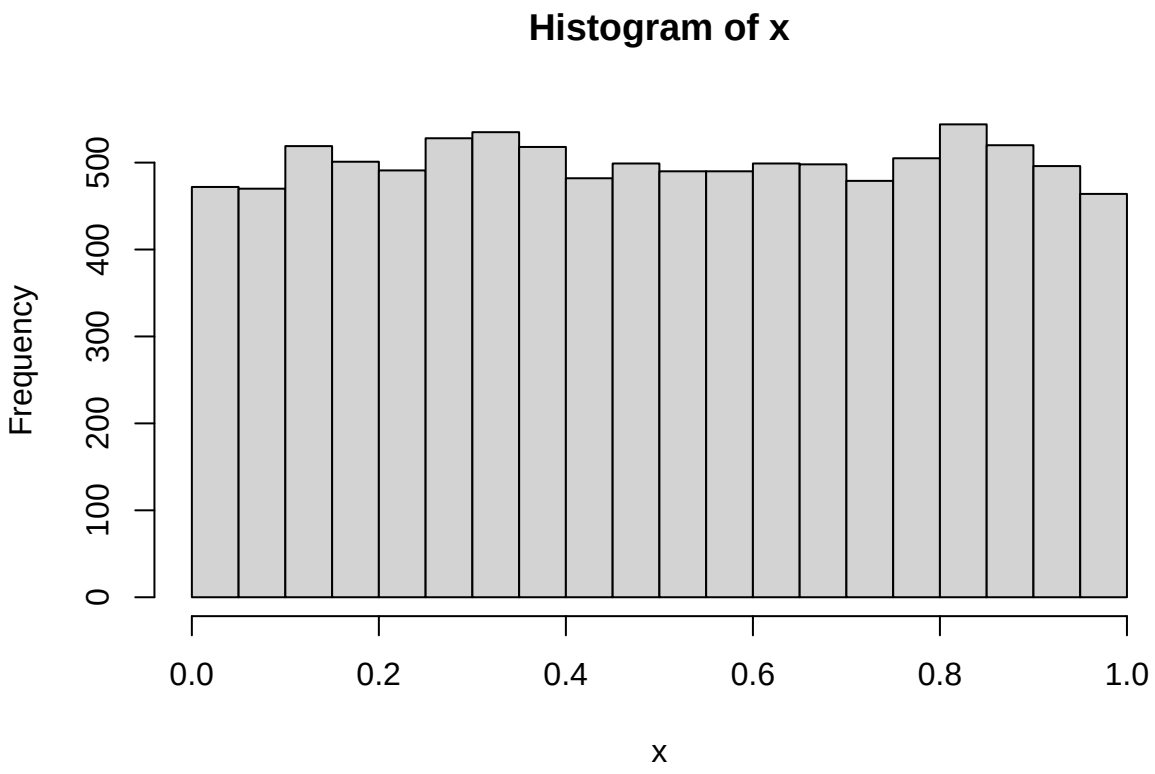
$$K_i := \text{"i-te Kugel fällt in den Innenkreis"} = \{(X_i - m_1)^2 + (Y_i - m_2)^2 \leq \left(\frac{d}{2}\right)^2\}$$

Die Zufallsvariable $\frac{1}{n}(1_{K_1} + 1_{K_2} + \ldots + 1_{K_n})$ misst also das gesuchte Verhältnis.

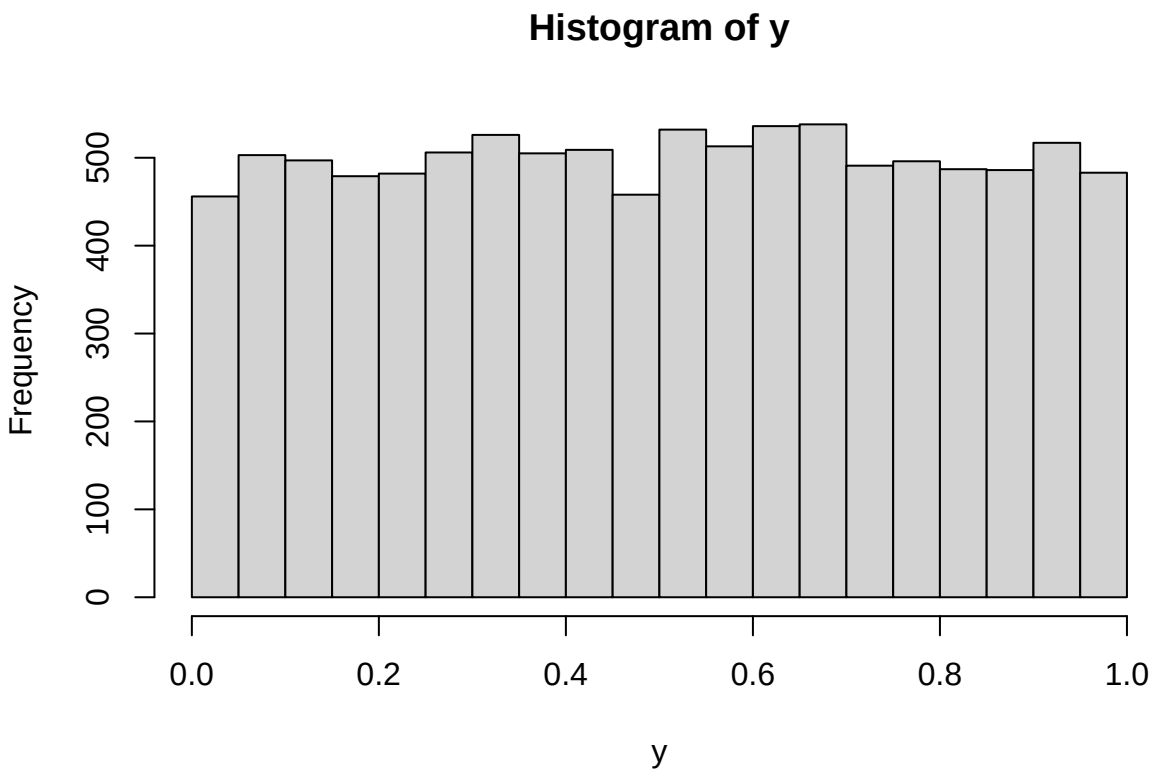
### zu b)

Wir erzeugen zunächst die benötigten Zufallsvariablen und vergewissern uns ihrer korrekten Erzeugung mittels eines Histogramms.

```r
n <- 10^4
d <- 1
m_1 <- 0.5
m_2 <- 0.5
x <- runif(n, min=-d/2+m_1,max=m_1+d/2) # x-koordinaten der Aufprallpunkte
hist(x)
```




```r
y <- runif(n, min=-d/2+m_2, max=m_2+d/2)
hist(y)
```




Wir erstellen zunächst eine Funktion, die den quadratischen Abstand eines Aufprallpunkts (x,y) zum Mittelpunkt (m_1,m_2) misst.

```
radius_sq <- function(x,y,m_1,m_2){
  (x-m_1)^2 + (y-m_2)^2
}
```

Dann ermitteln wir den Anteil der in den Kreis gefallenen Kugeln, indem wir 1. prüfen, ob der quadratische Abstand zum Mittelpunkt <= (d/2)^2 ist. 2. die im Vektor enthaltenen TRUE-Werte aufsummieren und durch die Länge des Vektors teilen.

```
ratio <- mean( radius_sq(x,y,m_1,m_2) <= (d/2)^2)
ratio
```

```
## [1] 0.7949
```

Zu guter letzt bestimmen wir unseren approximativen Wert für $\pi$:

```
ratio*4
```

```
## [1] 3.1796
```

**zu c)**

```
approx_pi <- function(n,m_1,m_2,d){
  x <- runif(n, min=-d/2+m_1, max=m_1+d/2)
  y <- runif(n, min=-d/2+m_2, max=m_2+d/2)
  mean( radius_sq(x,y,m_1,m_2) <= (d/2)^2)*4
}
```

```r
tolerance <- 10^(-4)
error <- 1
n <- 0
my_pi <- NA
while( error > tolerance ){
  n <- n + 1000
  my_pi <- approx_pi(n,1/2,1/2,1)
  error <- abs(pi - my_pi )
}
print(paste("Nach", n , "gefallenen Kugeln, liegt der Fehler für pi mit", error, "innerhalb des Toleranz
```

```
## [1] "Nach 290000 gefallenen Kugeln, liegt der Fehler für pi mit 9.01050308965701e-05 innerhalb des To
```

**zu d)**

Es gilt:

$$plim_{n\to\infty}\frac{1}{n}(1_{K_1}+1_{K_2}+\ldots+1_{K_n}) = E(1_{K_1}) = 0\cdot P(1_{K_1}=0)+1\cdot P(1_{K_1}=1) = P(K_1)$$

wobei wir zuerst das Gesetz der großen Zahlen und dann die Definition für den Erwartungswert von diskret verteilten Zufallsvariablen genutzt haben.

# Aufgabe 5:

Seien $Y \sim U[-5,5]$, $Z \sim U[20,40]$ und $B \sim B(1,0.5)$ drei unabhängige Zufallsvariablen. $X$ sei eine weitere Zufallsvariable. Sie nimmt den Wert von $Y$ an, falls $B=1$ und den Wert von $Z$, falls $B=0$. Seien $X_1, X_2, X_3, ...$ untereinander unabhängige Kopien von $X$. $X, X_1, X_2, X_3, ...$ sind also voneinander unabhängige identisch verteilte Zufallsvariablen.

a) Wie lassen sich die Verteilung von $(Y,Z)$ aus zwei unabhängigen Zufallsvariablen $U_1, U_2 \sim U[0,1]$ möglichst einfach gewinnen? Was ist der Erwarungswert und was ist die Varianz von $X$?

b) Bestimmen Sie die Dichte der Zufallsvariable $X$ mittels Monte-Carlo-Simulation anhand von 10 000 Realisationen und plotten Sie das zugehörige Histogramm.

c) Ergänzen Sie Konstanten auf der linken Seite der folgenden Gleichung, um deren Richtigkeit zu gewähleisten.

$$\lim_{m\to\infty} P(X_1+X_2+\ldots+X_m \le u) = \Phi(u)$$

Hierbei ist $\Phi$ die Verteilungsfunktion der Standardnormalverteilung. Wählen Sie die Konstanten so, dass die Korrektheit der Gleichung auf Basis des Wissens aus der Vorlesung offensichtlich ist.

d) Bestimmen Sie die Dichte der Zufallsvariable $X_1+X_2+\ldots+X_n$ mittels Monte-Carlo-Simulation anhand von 10 000 Realisationen und plotten Sie das zugehörige Histogramm für $m = 1,2,3,4,5,10$ und 100.

## Lösung

**zu a) <- überarbeiten!**

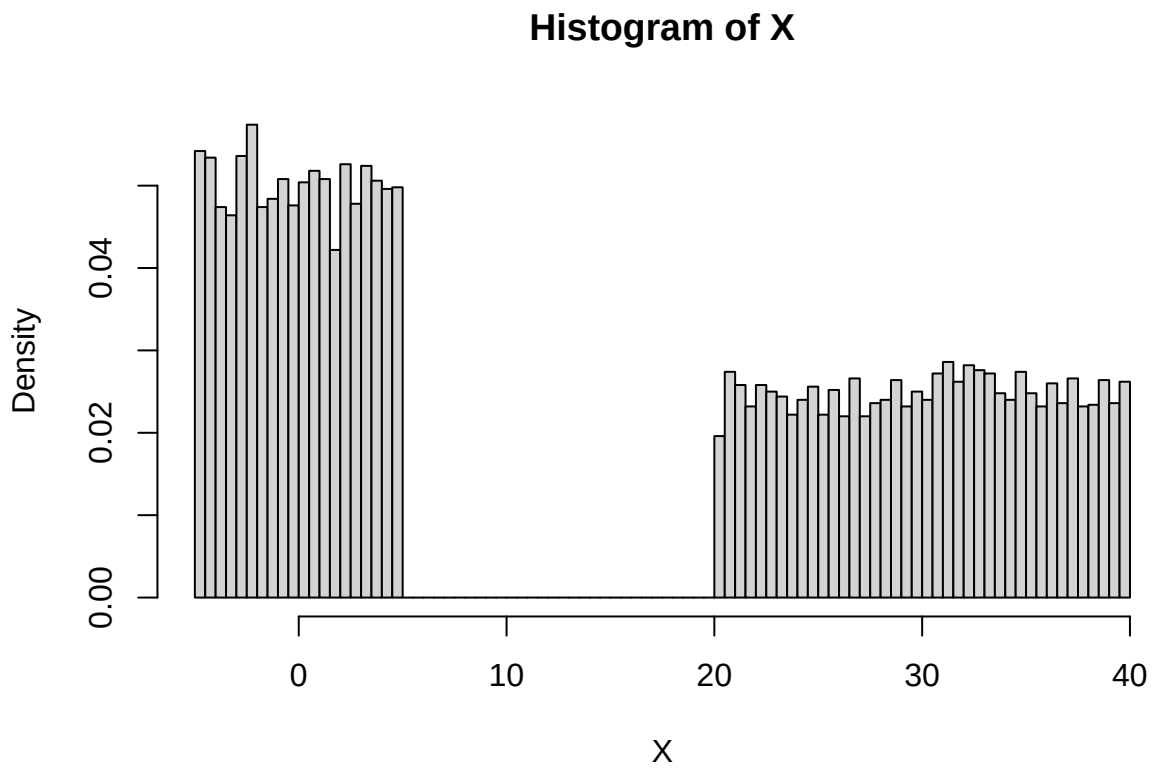
Seien $N_1, N_2 \sim N(0,1)$ unabhängig. Dann gilt

$$P(Y \le s, Z \le t) = P(N_1 - 5 \le s, 3N_2 + 5 \le t) \text{ für alle } t \in \mathbb{R}$$

Somit haben die Zufallsvektoren $(Y,Z)$ und $(N_1-5, 3N_2+5)$ die gleiche Verteilung.

**zu b)**

```
n <- 10^4
p <- 0.5
B <- rbinom(n,1,p)
Y <- runif(n, -5, 5)
Z <- runif(n, 20, 40)
X <- B*Y+(1-B)*Z
hist(X, breaks=100, probability=TRUE)
```




### zu c)

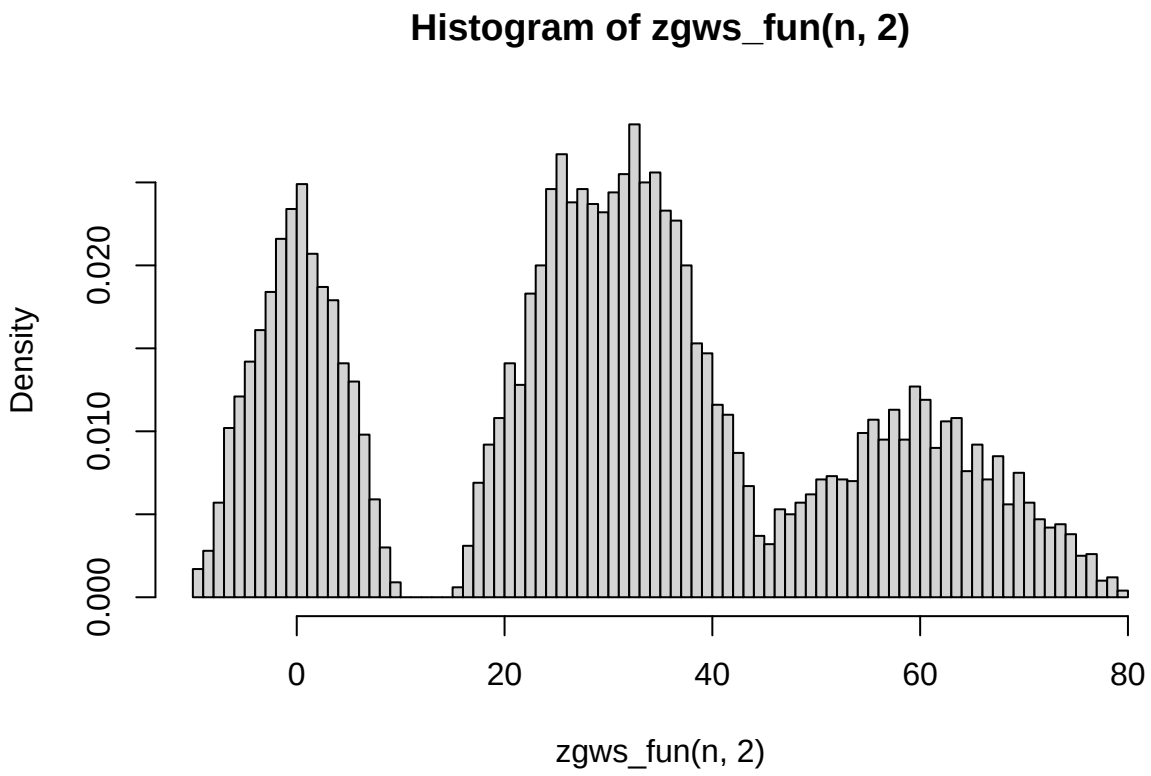
**zu d)**

```
zgws_fun <- function(realisations, summands){
  k <- 1
  limit_variable <- c(1:realisations)*0
  while(k <= realisations){
    B <- rbinom(summands,1,p)
    # Y <- rnorm(n, -5, 1)
    # Z <- rnorm(n, +5, 3)
    Y <- runif(summands, -5, 5)
    Z <- runif(summands, 20, 40)
    X <- B*Y+(1-B)*Z
    limit_variable[k] <- sum(X)
    k <- k+1
  }
```

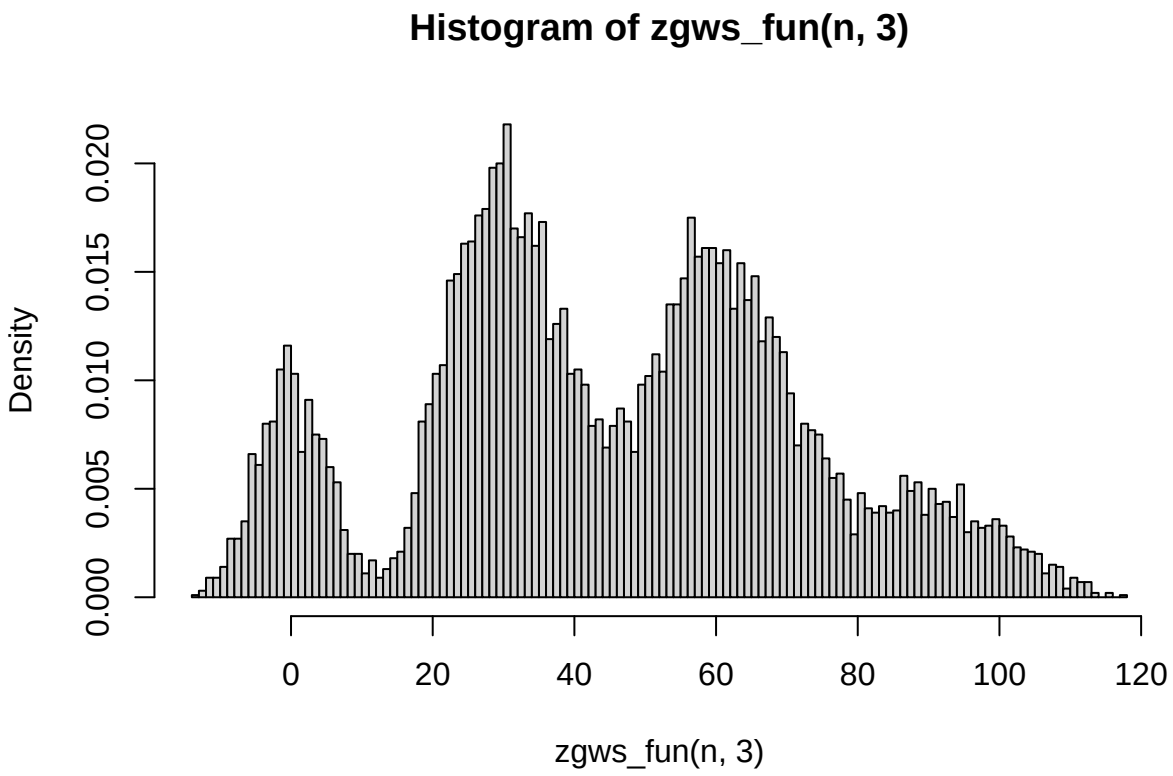
```
  limit_variable


}
hist(zgws_fun(n, 1), breaks=100, probability=TRUE)
```

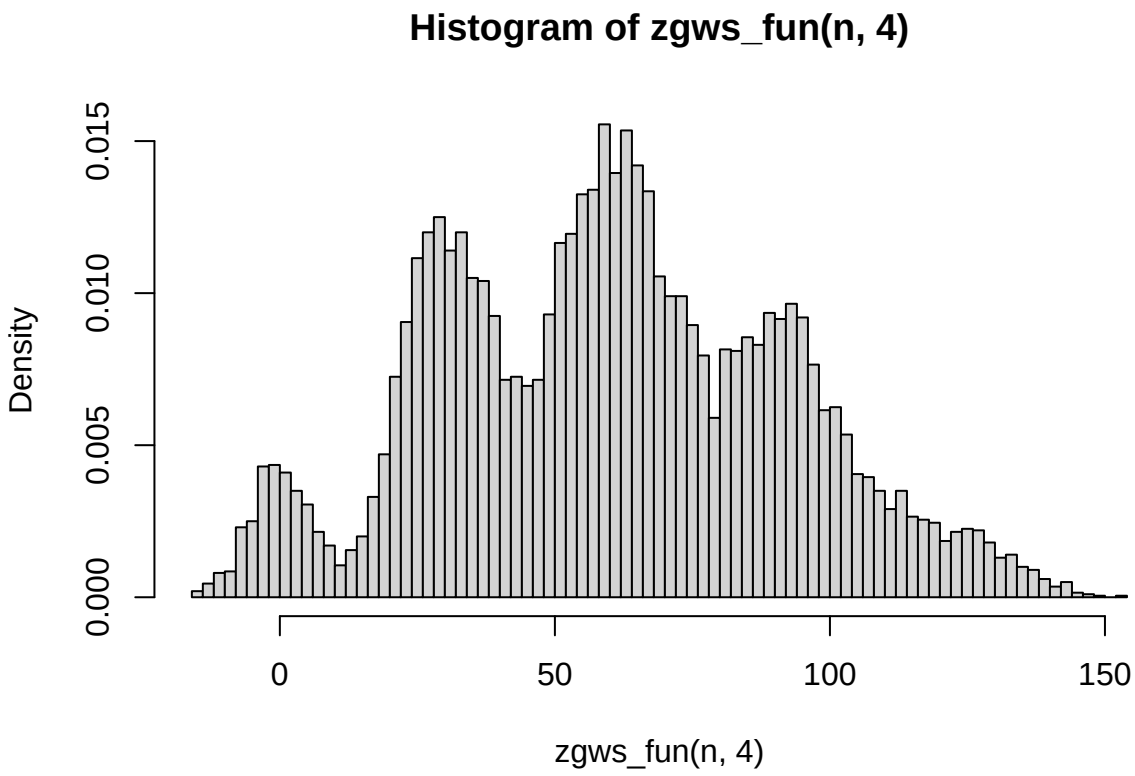

```
hist(zgws_fun(n, 2), breaks=100, probability=TRUE)
```



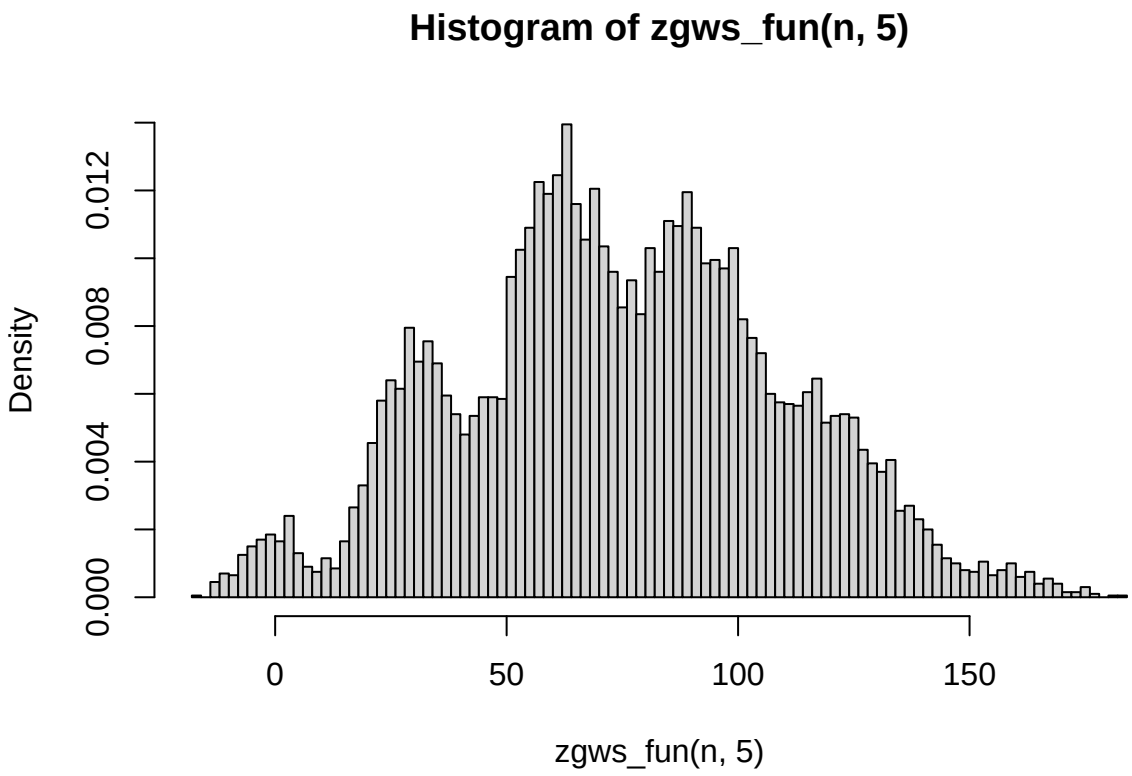

```
hist(zgws_fun(n, 3), breaks=100, probability=TRUE)
```



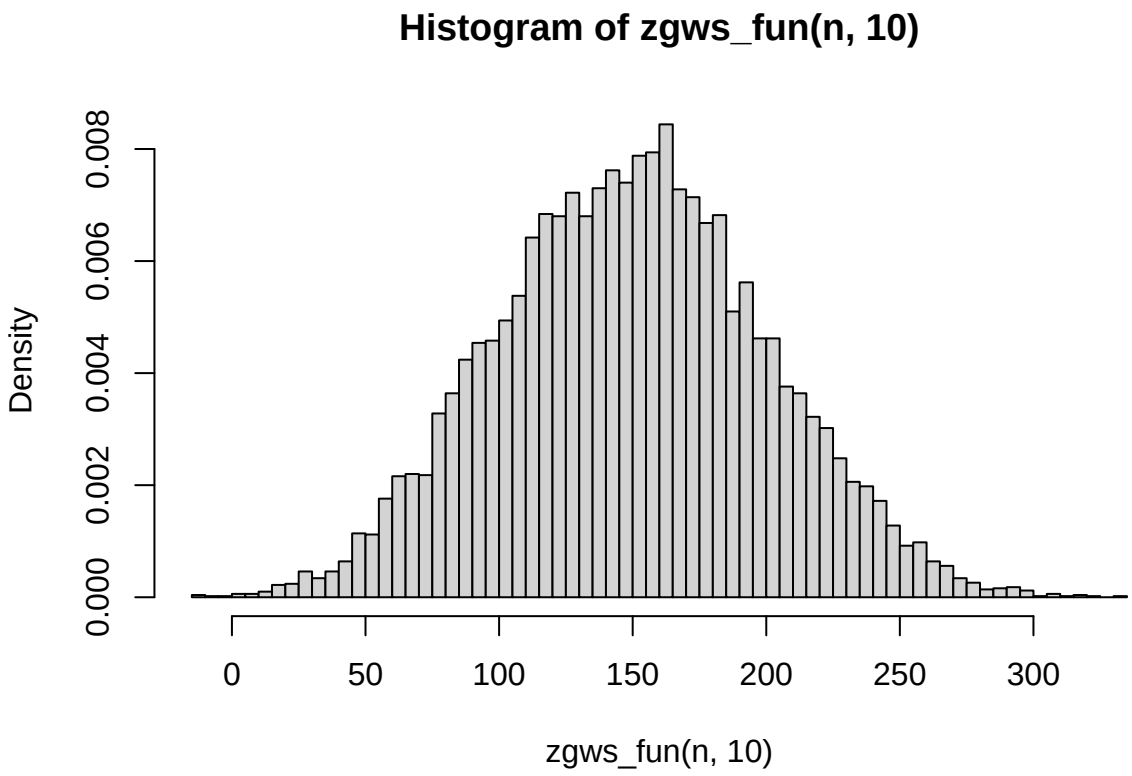

```
hist(zgws_fun(n, 4), breaks=100, probability=TRUE)
```



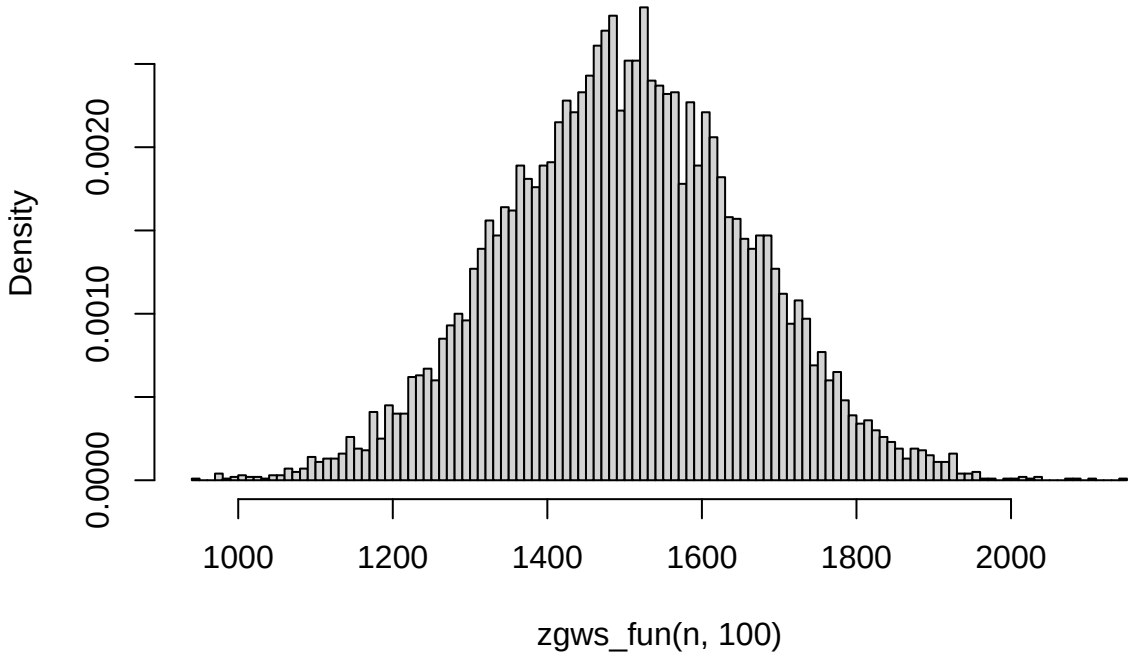

```
hist(zgws_fun(n, 5), breaks=100, probability=TRUE)
```




```
hist(zgws_fun(n, 10), breaks=100, probability=TRUE)
```




```
hist(zgws_fun(n, 100), breaks=100, probability=TRUE)
```


Histogram of zgws_fun(n, 100)
Density
0.0000
0.0010
0.0020
1000
1200
1400
1600
1800
2000
zgws_fun(n, 100)