Vorlesung 7a

Der zentrale Grenzwertsatz

1. Auftakt

Am Ende von V6b haben wir festgestellt:

Die standardisierte Summe von unabhängigen, identisch normalverteilten Zufallsvariablen ist standard-normalverteilt

Mit anderen Worten: Sind $Y_1, Y_2, \ldots, Y_n$ unabhängig und $N(\mu, \sigma^2)$-verteilt, dann ist

$$\frac{Y_1 + \cdots + Y_n - n\mu}{\sqrt{n\sigma^2}} \quad \text{standard-normalverteilt.}$$

2. Der Zentrale Grenzwertsatz: Die Botschaft

Der Zentrale Grenzwertsatz liefert

eine gewaltige Weiterung der Aussage aus Abschnitt 1

(asymptotisch für große n):

Zentraler Grenzwertsatz:

“Die standardisierte Summe von VIELEN

unabhängigen, identisch verteilten

nicht notwendig normalverteilten

$\mathbb{R}$-wertigen Zufallsvariablen

mit endlicher Varianz

ist annähernd standard-normalverteilt”

Formal:

Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariable mit endlichem Erwartungswert μ und endlicher Varianz $\sigma^2 > 0$. Dann gilt für alle $c < d \in \mathbb{R}$

$$\mathbf{P}\Big(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} \in [c, d]\Big) \underset{n\to\infty}{\longrightarrow} \mathbf{P}(Z \in [c, d]).$$

Dabei ist Z standard-normalverteilt.

In Worten:

Die standardisierte Summe von n unabhängigen,
identisch verteilten $\mathbb{R}$-wertigen Zufallsvariablen
mit endlicher Varianz
konvergiert für $n \to \infty$ in Verteilung
gegen eine standard-normalverteilte Zufallsvariable.

Abraham de Moivre:

Der faire Münzwurf (1733)

Pierre-Simon Laplace:

Allgemeine binomiale Zufallsgrößen (1812)

Pafnuty Lvovich Chebyshev:

Skizze eines Beweises für den allgemeinen Fall (1887)

Aleksandr Mikhailovich Lyapunov:

Noch allgemeiner (1906)

Allgemeiner zentraler Grenzwertsatz (1901)

Andrei Andreyevich Markov:

weitere Verallgemeinerungen (~ 1910)

Nehmen wir an,

diese Herren hätten sich

auf ihre vielen anderen Interessen

beschränkt.

ZENTRALER GRENZWERTSATZ

Nehmen wir an,

diese Herren hätten sich

auf ihre vielen anderen Interessen

beschränkt.

ZENTRALER GRENZWERTSATZ

Unbekannt.

Könnten wir ihn entdecken?

Wie kämen wir auf φ?

Warum gerade $e^{-x^2/2}$?

4. Ein Beispiel: Summen von unabhängigen uniform verteilten Zufallsvariablen

Wir denken an

Rundungsfehler bei Addition

In Wirklichkeit

$$\pi =$$

$$3.14159265358979323846264338327950288419716939937510 5\ldots$$

Im Rechner

$$\pi \leftarrow 3.14159265358979$$

MODELL

Zahl = Rechnerdarstellung + Rundungsfehler.

$$A = a^{[R]} + \varepsilon X \qquad \varepsilon = 10^{-15}$$

Annahme: X uniform verteilt auf $[-0.5, 0.5]$.

$$\sum_{i=1}^{n} A_i = ?$$

$$\sum_{i=1}^{n} A_i = \sum_{i=1}^{n} a_i^{[R]} + \varepsilon \sum_{i=1}^{n} X_i$$

Wie groß ist der Fehler?

$$\sum_{i=1}^{n} X_i \approx ?$$

Der Zentrale Grenzwertsatz gibt die Auskunft:

$$\sum_{i=1}^{n} X_i$$

ist - zumindest für große n -

approximativ N(0, $\sqrt{n}\,\sigma_X$)-verteilt.

Davon machen wir uns ein Bild durch

Simulationsexperimente:

$X_1, X_2, \ldots$ unabhängig

und uniform auf $[-0.5, 0.5]$ verteilt

Empirische Verteilung von

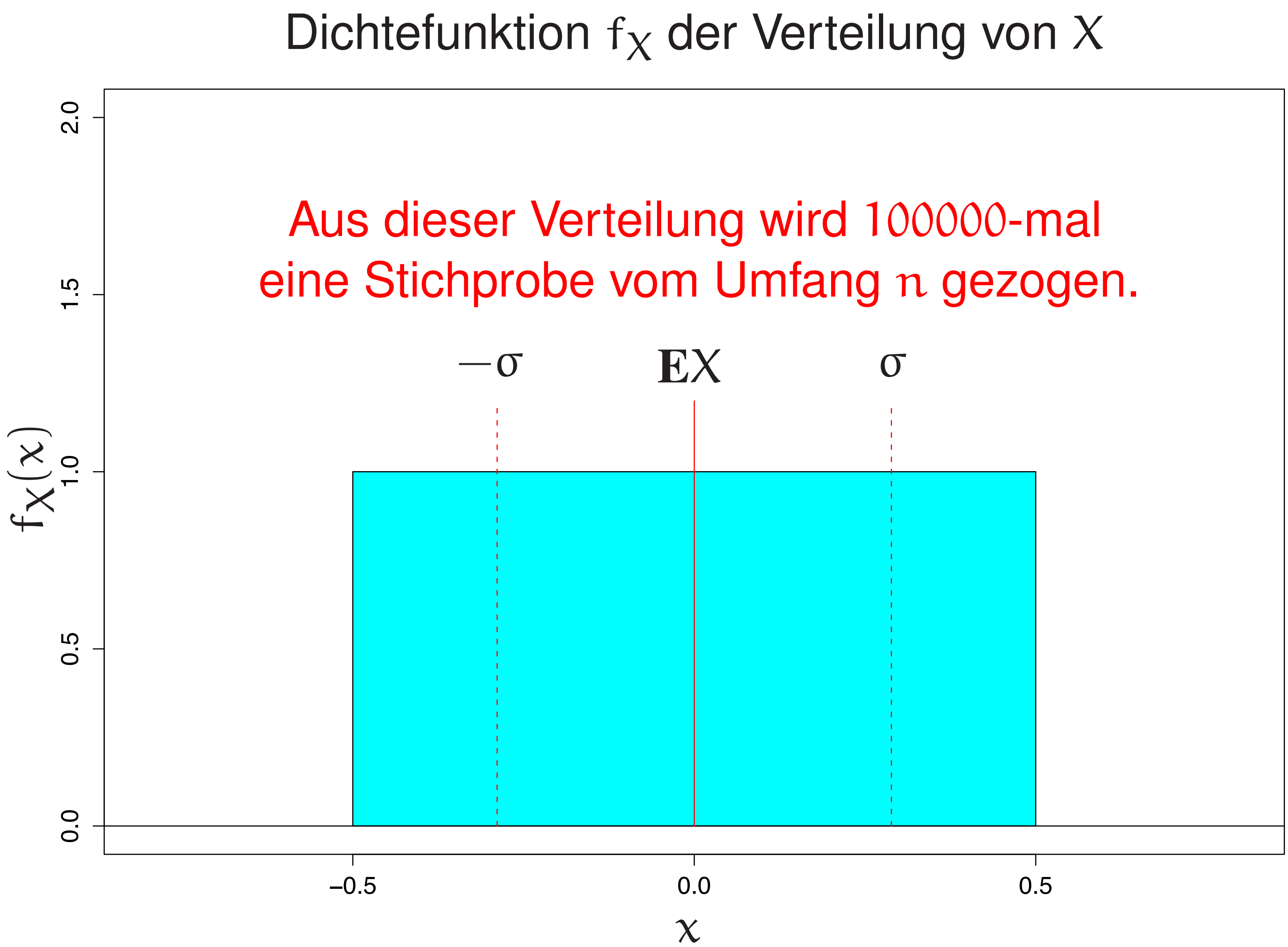
$$S_n := X_1 + ... + X_n$$

100000 Simulationen

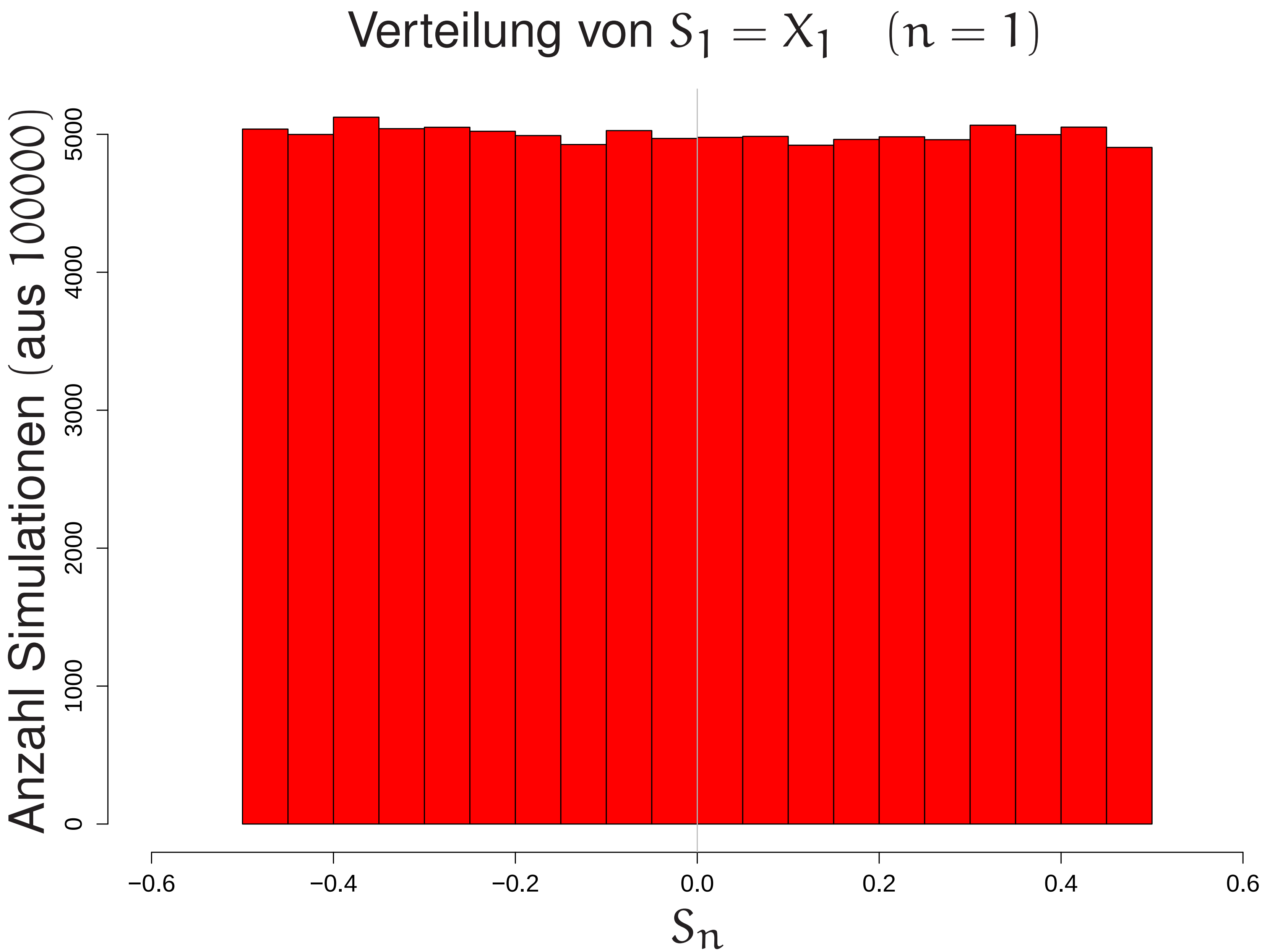
jeweils für

$$n = 1, 2, ..., 10$$

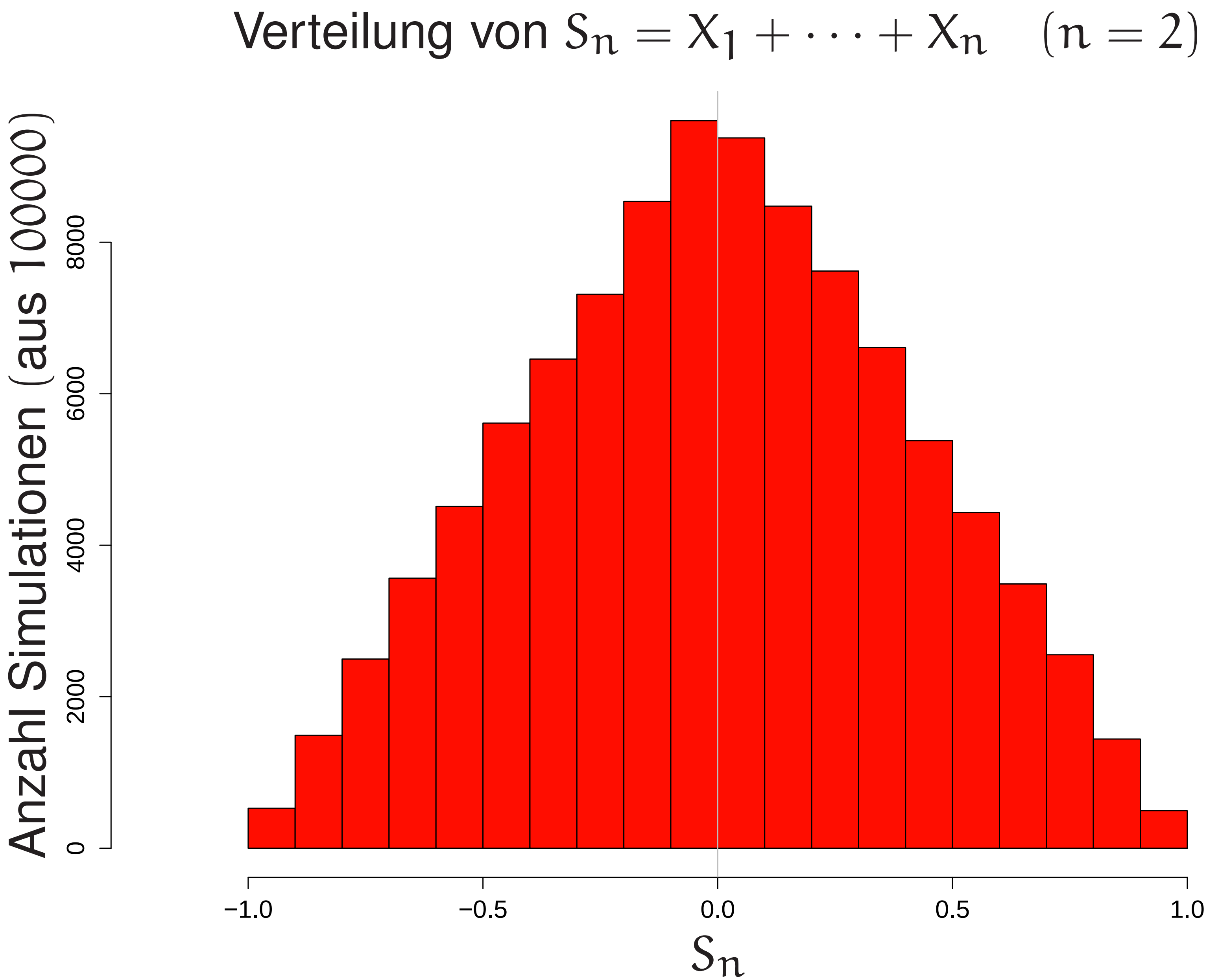
$$n = 15, 20, 30, ..., 100$$



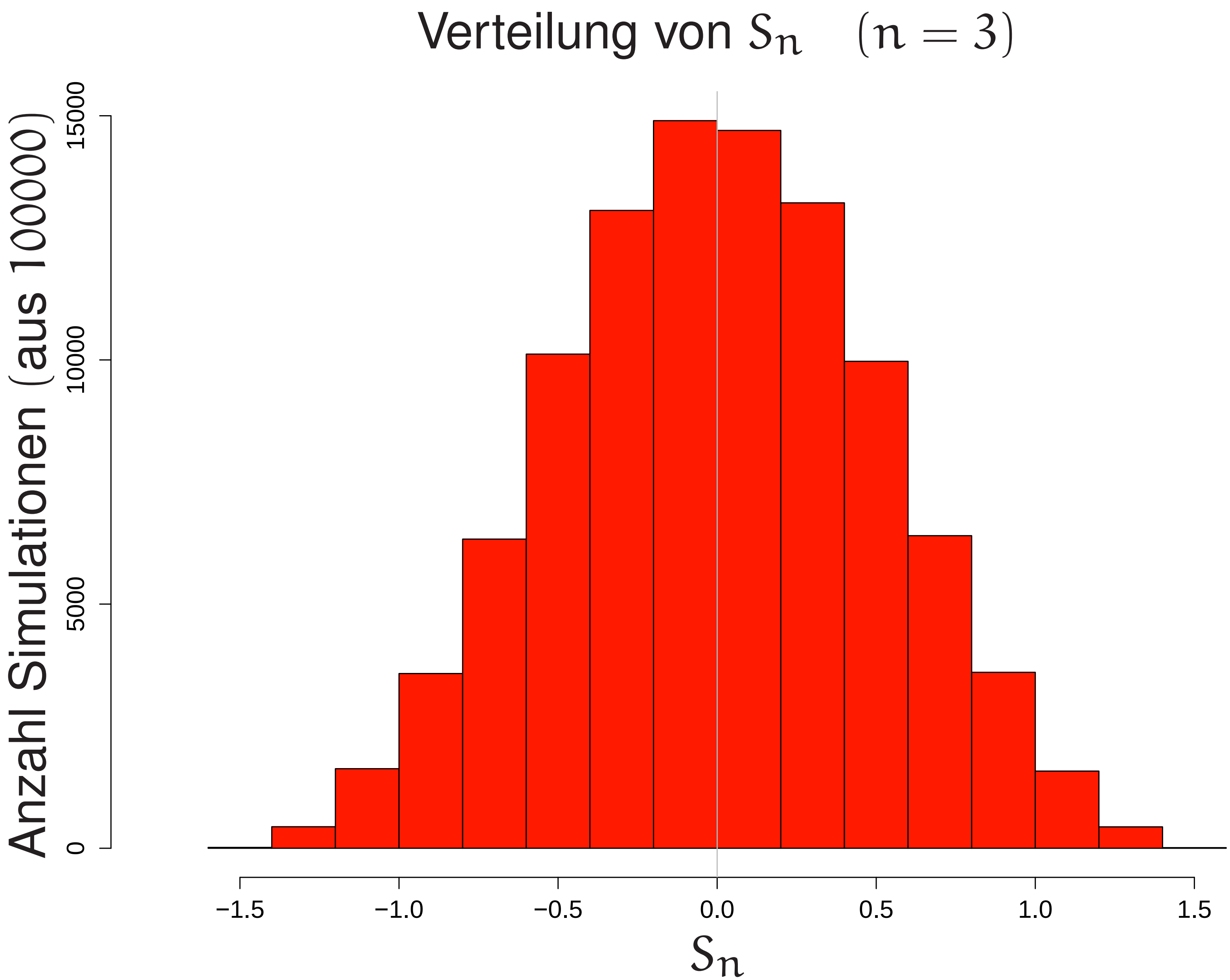
Dichtefunktion f_X der Verteilung von X
Aus dieser Verteilung wird 100000-mal
eine Stichprobe vom Umfang n gezogen.
$-\sigma$
$\mathbf{E}X$
σ
$f_X(x)$
x
0.0
0.5
1.0
1.5
2.0
−0.5
0.0
0.5



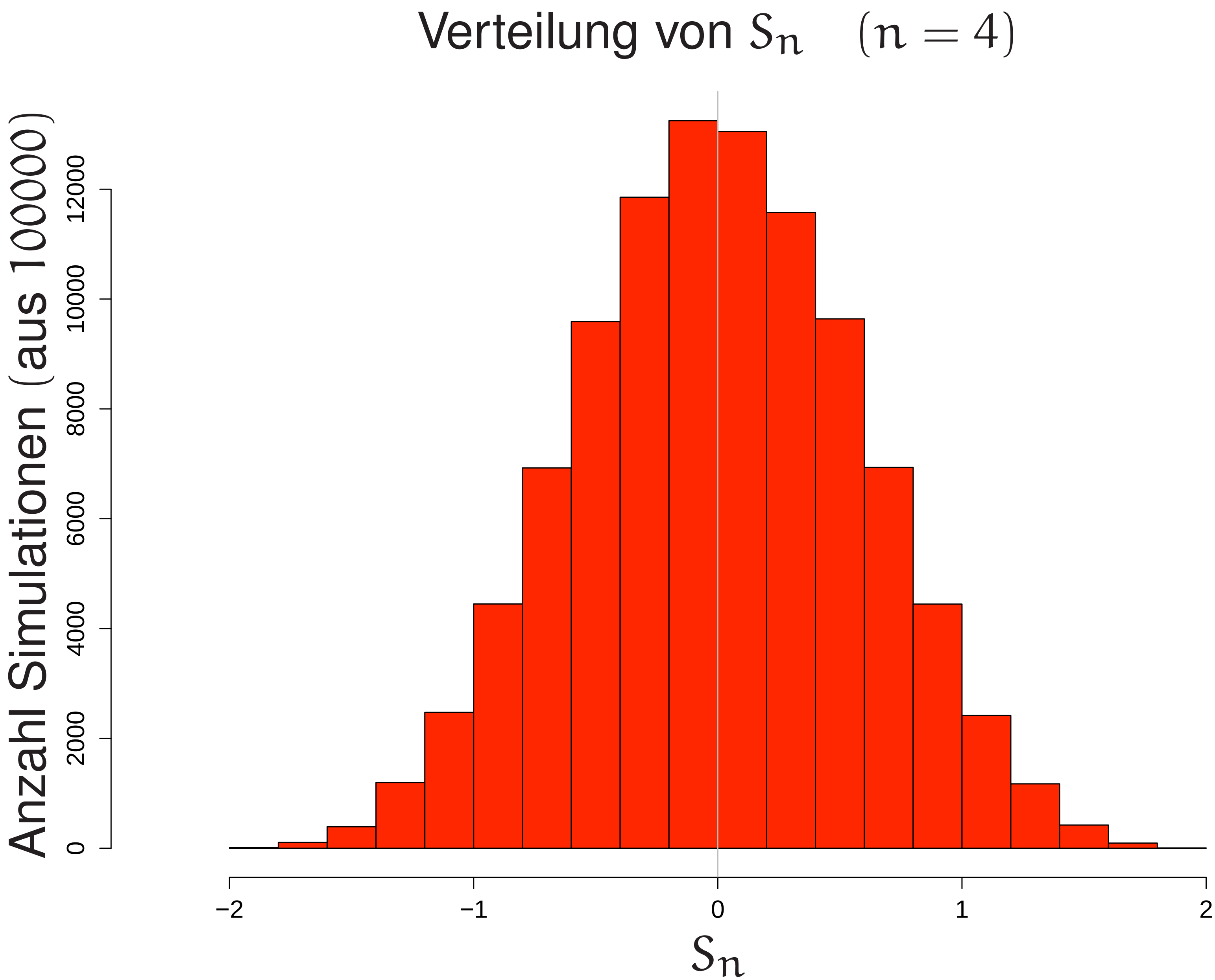
Verteilung von $S_1 = X_1 \quad (n = 1)$
Anzahl Simulationen (aus 100000)
0
1000
2000
3000
4000
5000
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
S_n

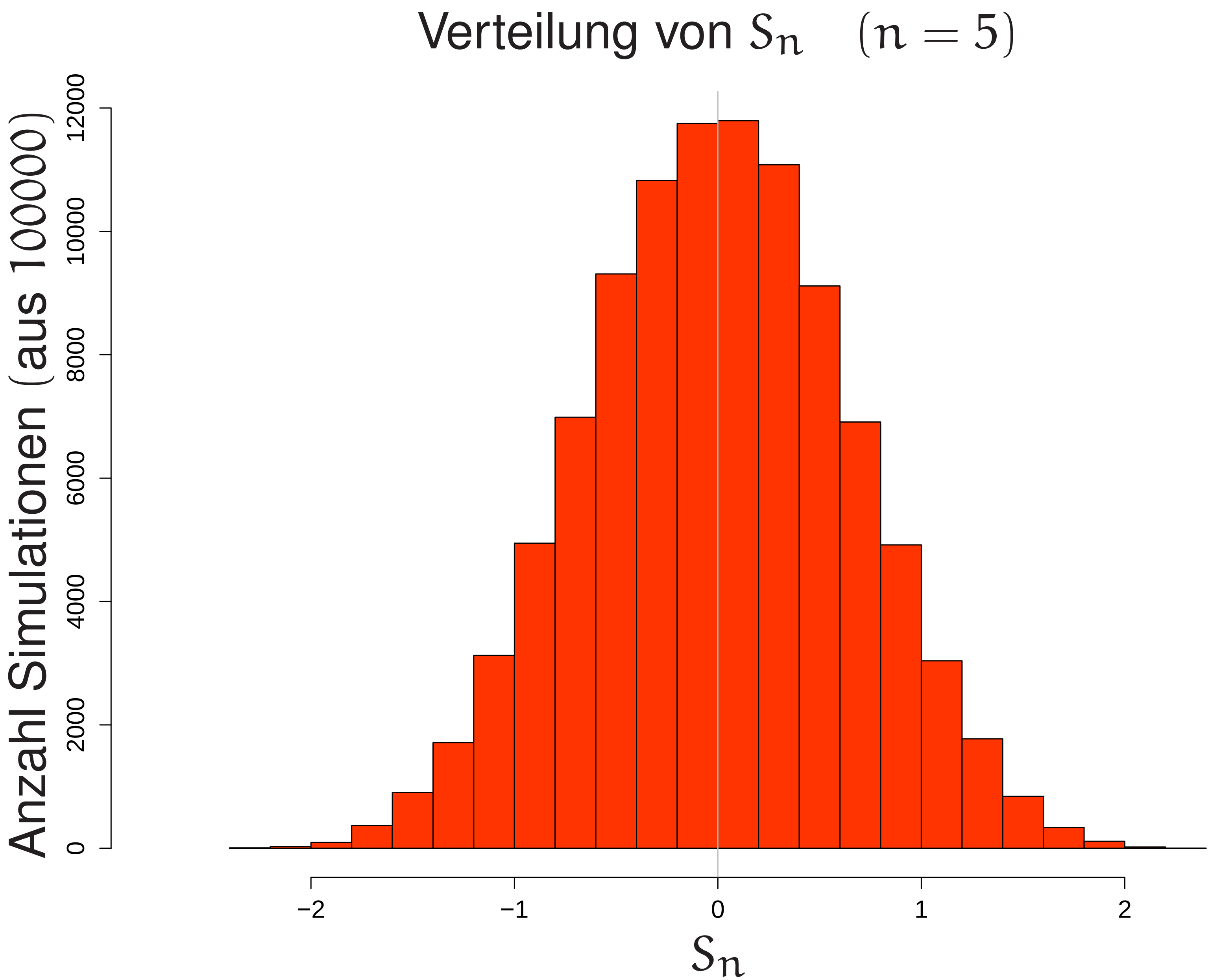


Verteilung von $S_n = X_1 + \cdots + X_n \quad (n = 2)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−1.0
−0.5
0.0
0.5
1.0
S_n

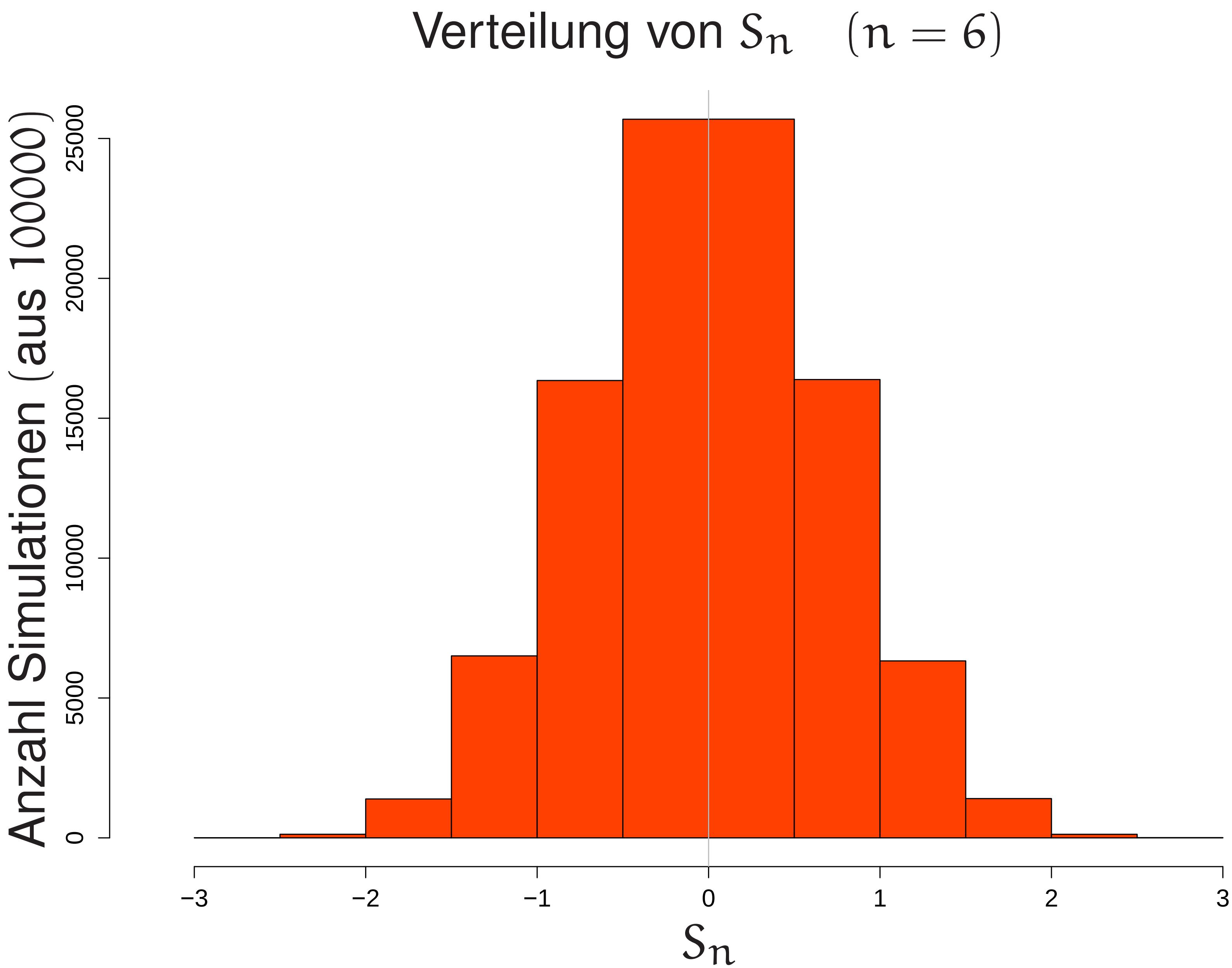


Verteilung von S_n $(n = 3)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
S_n

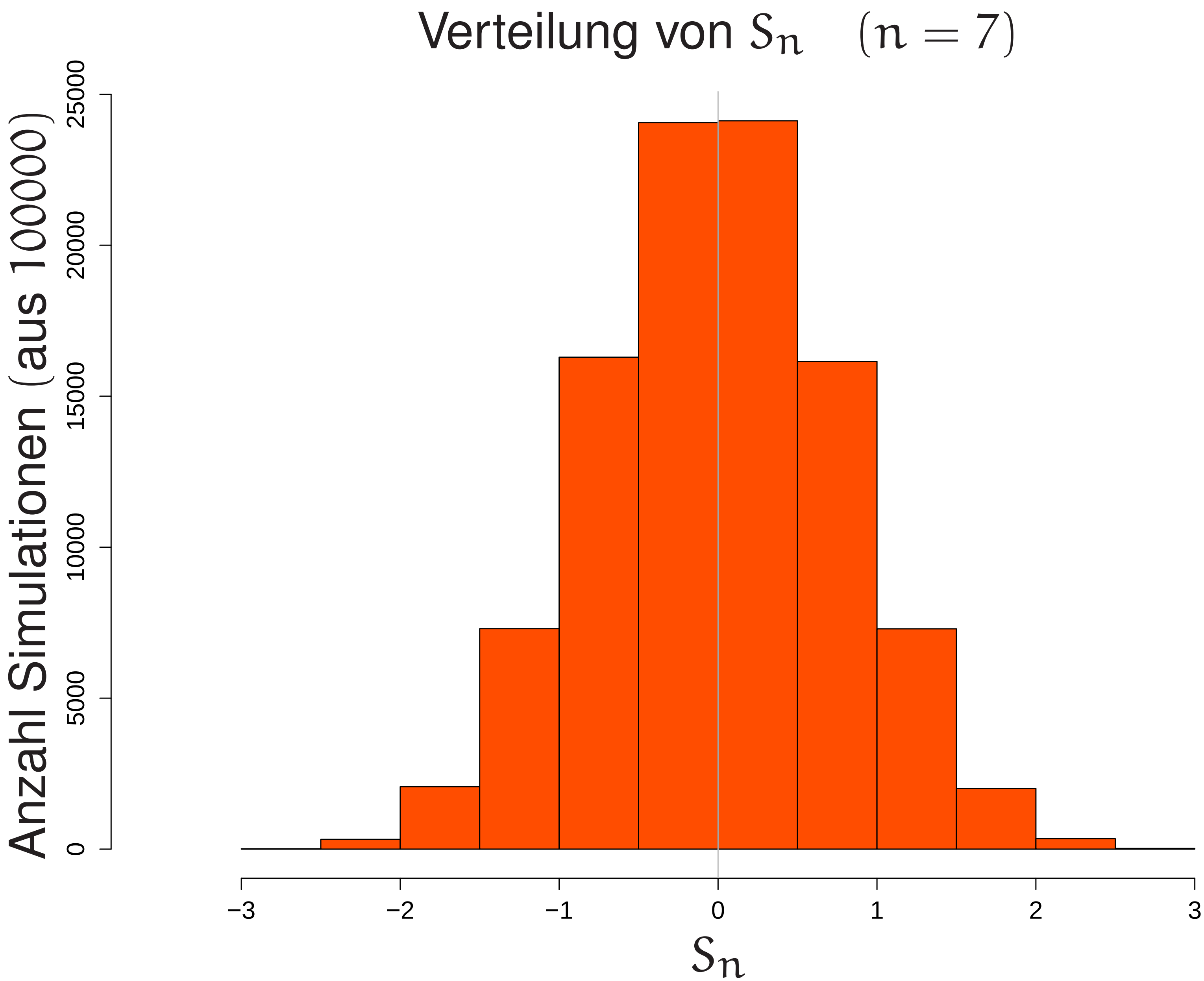

Verteilung von S_n $(n = 4)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
10000
12000
−2
−1
0
1
2
S_n



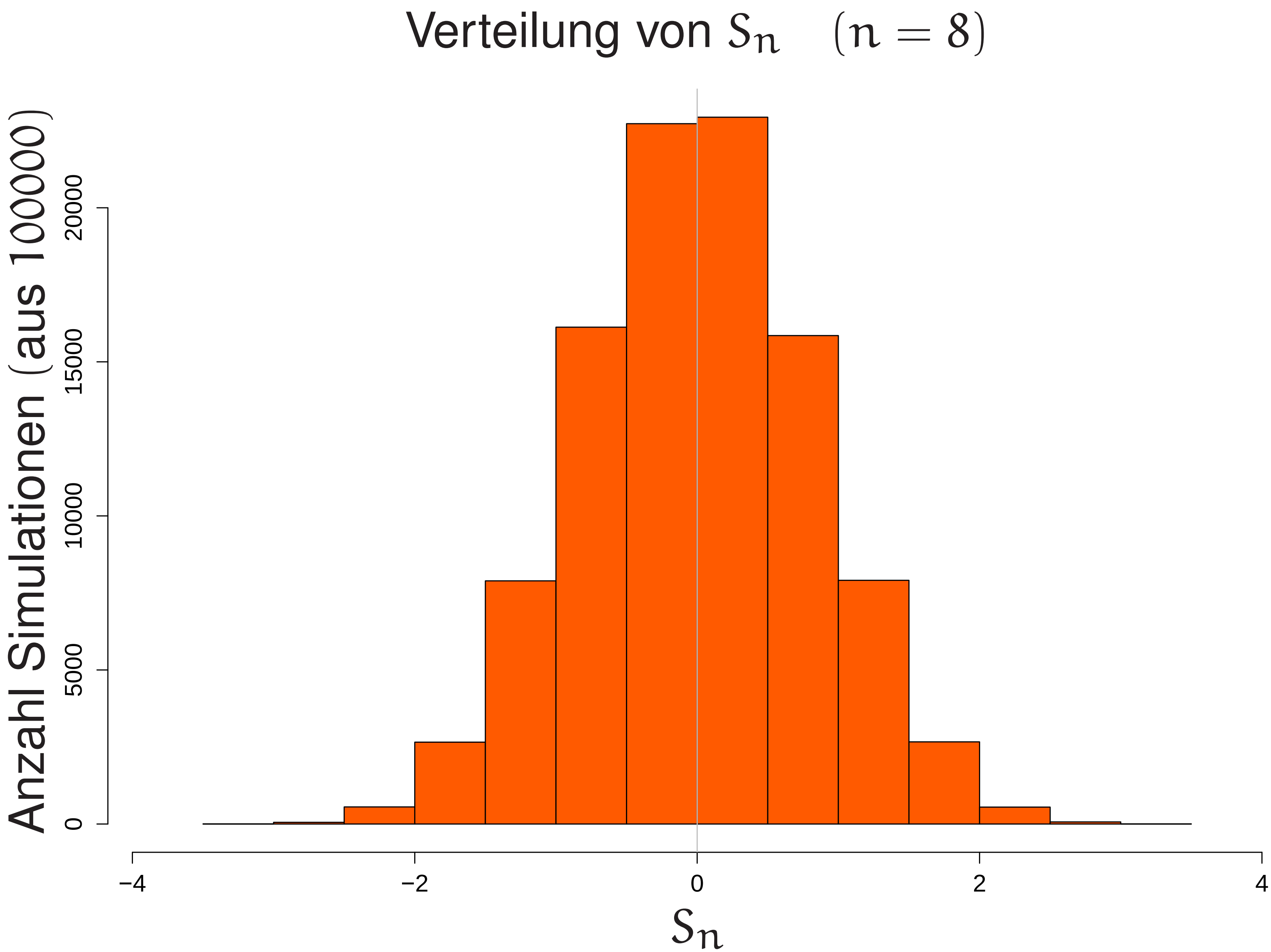
Verteilung von S_n $(n = 5)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
10000
12000
−2
−1
0
1
2
S_n



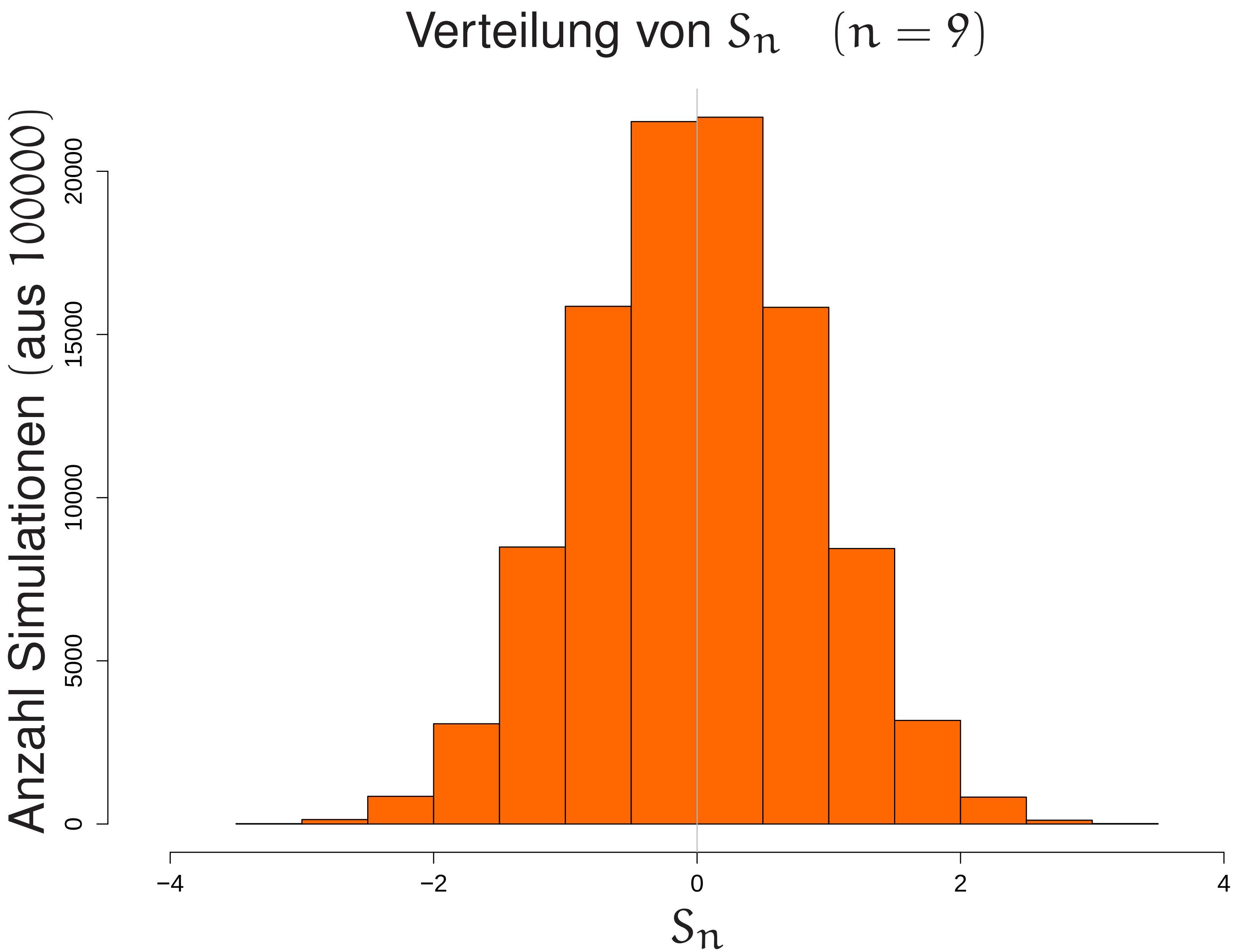
Verteilung von S_n $(n = 6)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
25000
−3
−2
−1
0
1
2
3
S_n



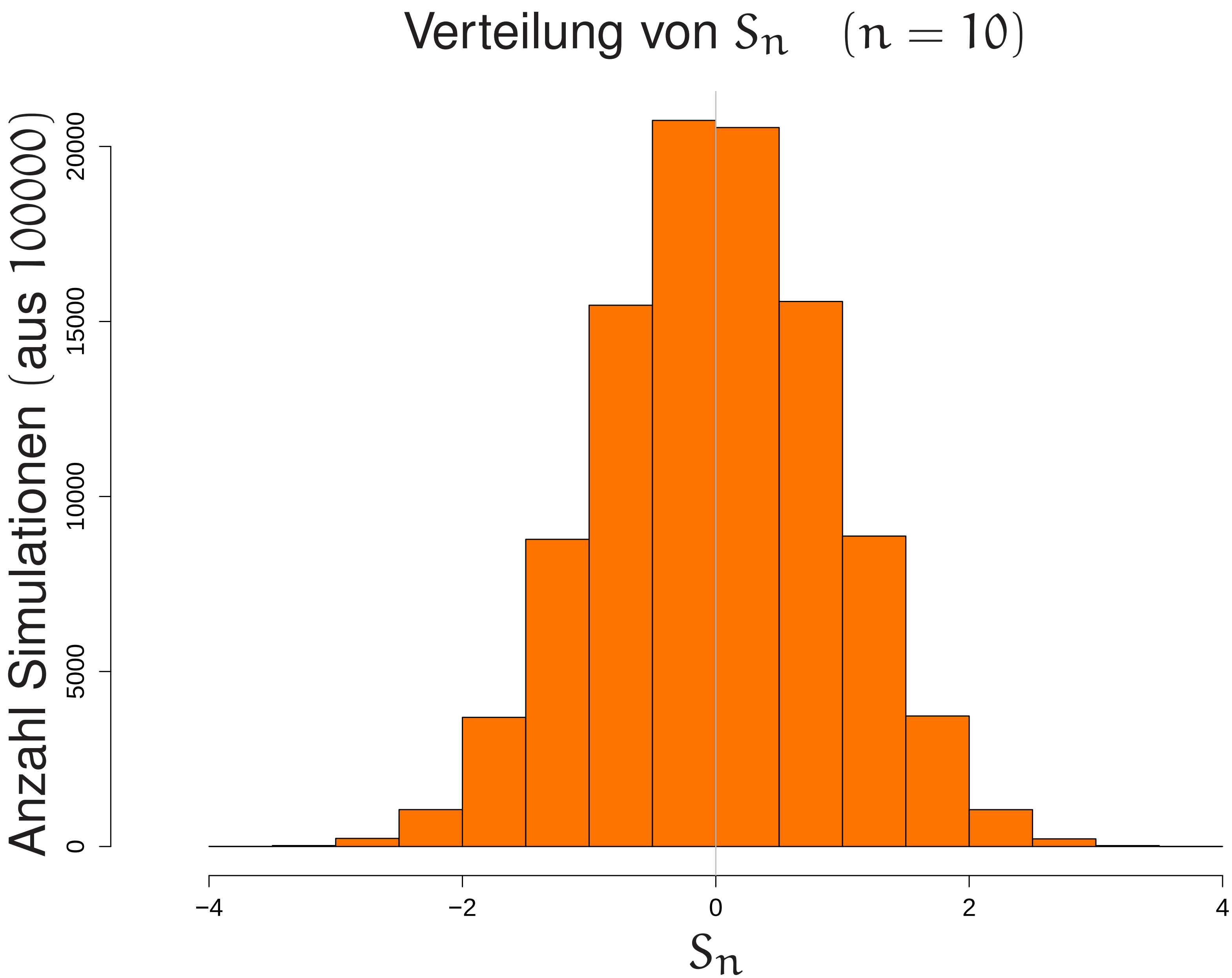
Verteilung von S_n $(n = 7)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
25000
−3
−2
−1
0
1
2
3
S_n

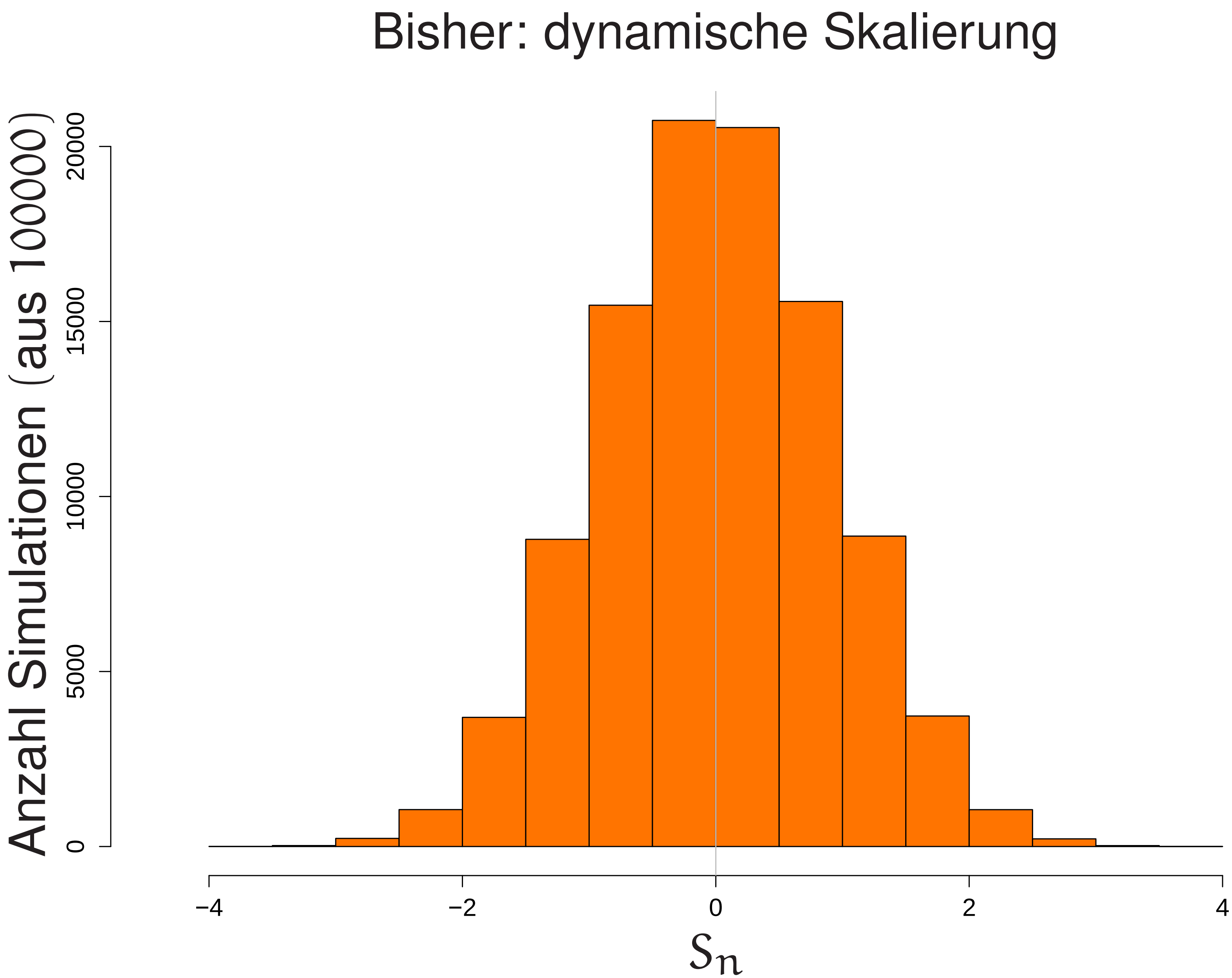


Verteilung von S_n $(n = 8)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n



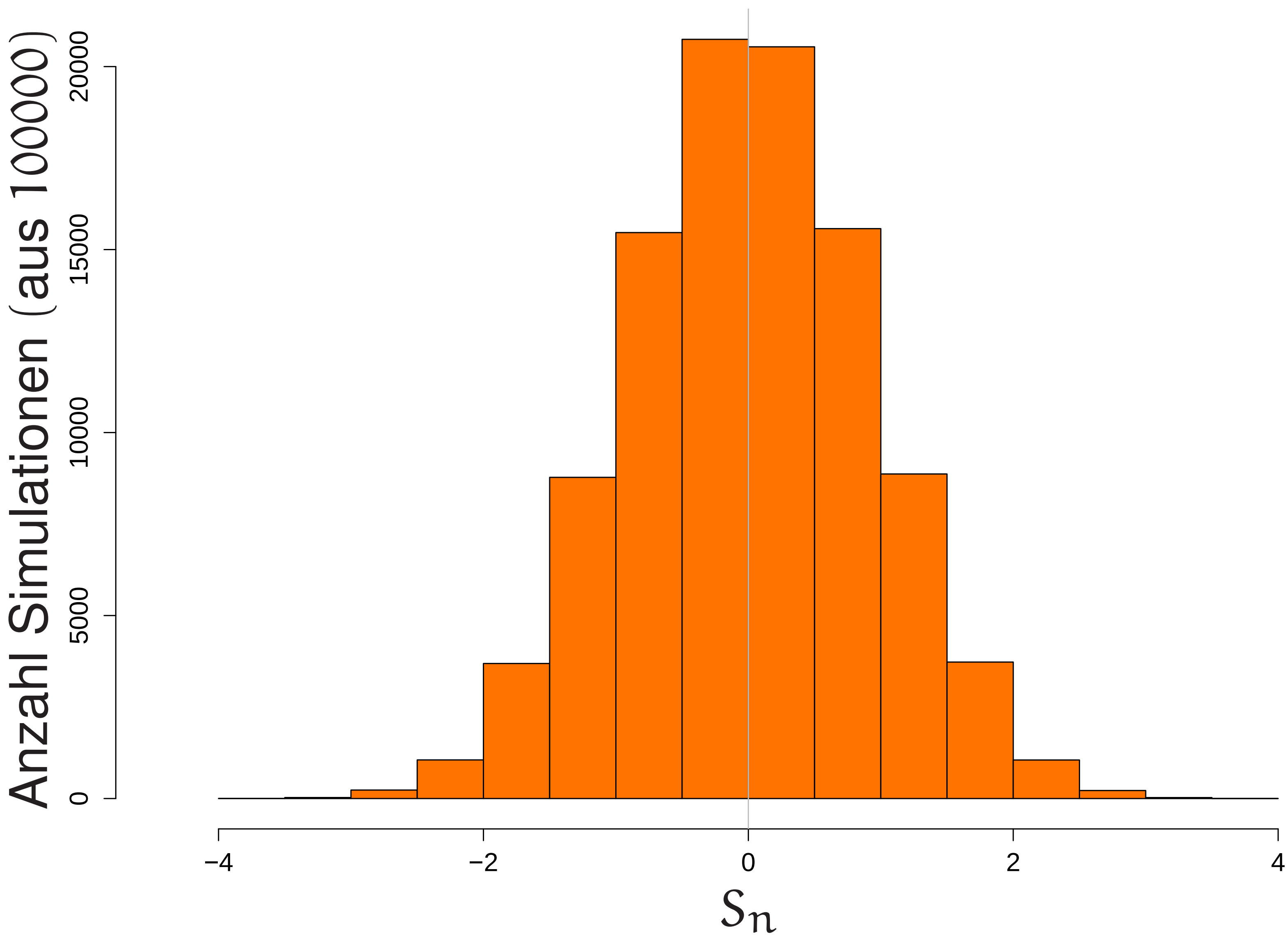
Verteilung von S_n $(n = 9)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

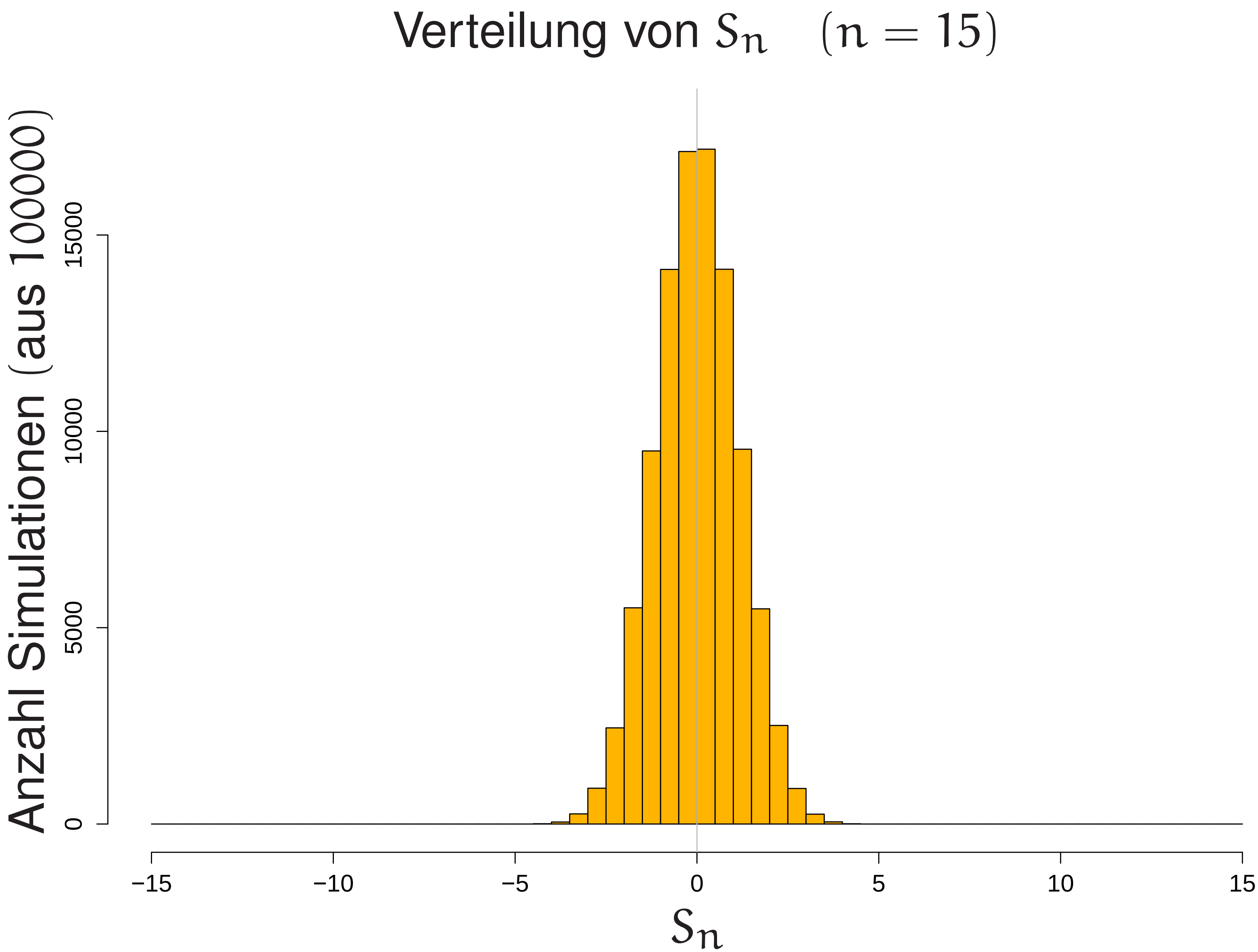

Verteilung von S_n $(n = 10)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n



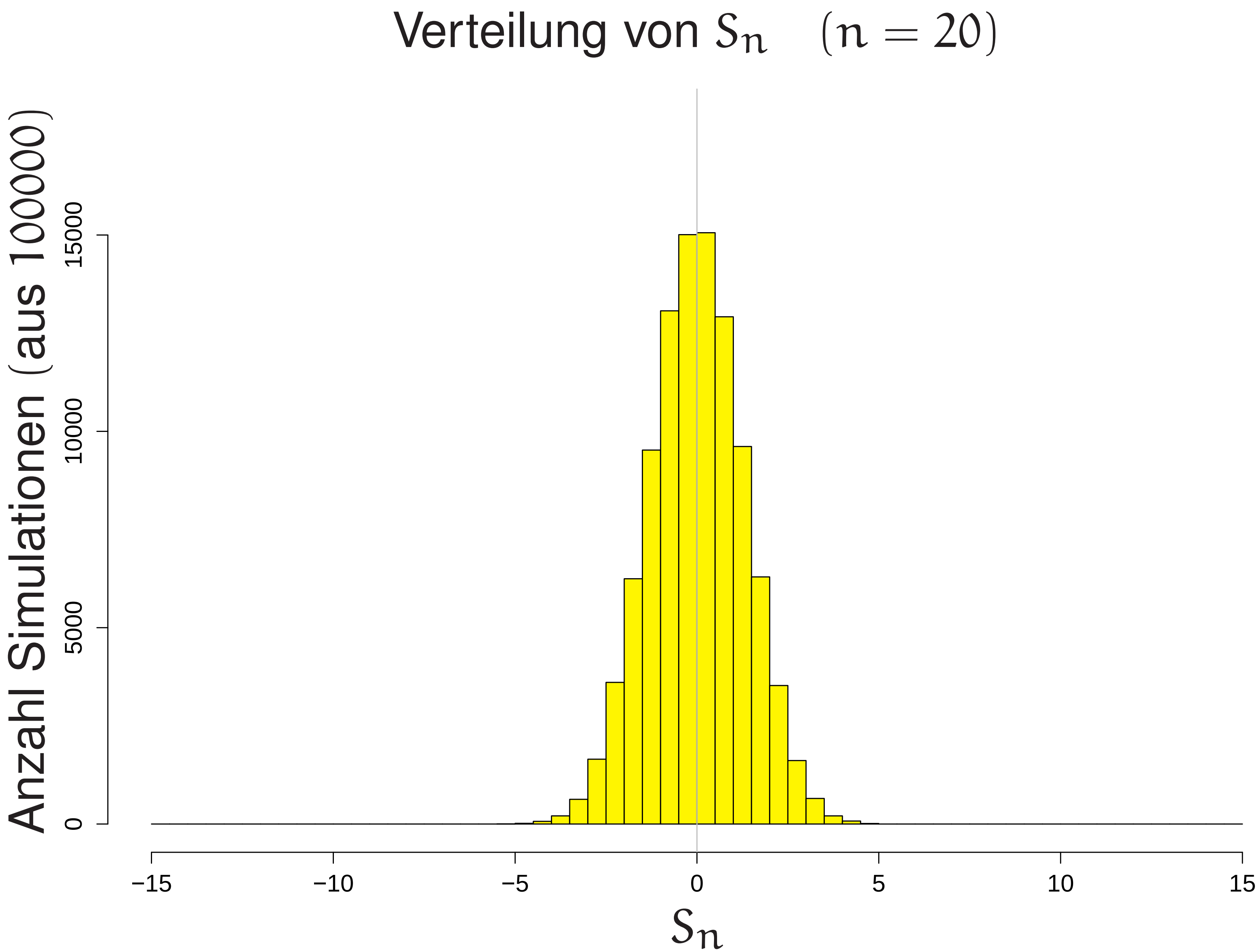
Bisher: dynamische Skalierung
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

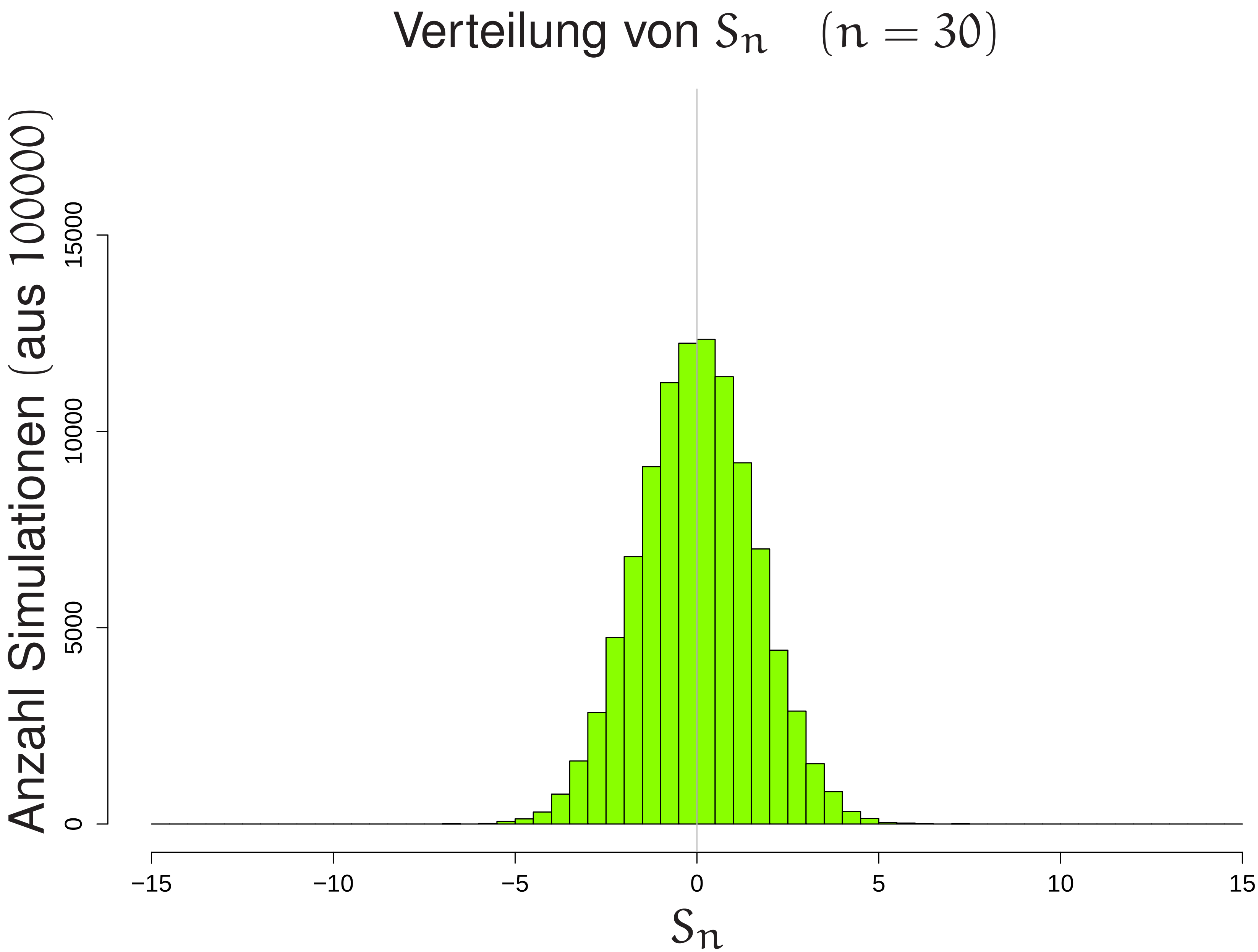
Jetzt: feste Skalierung



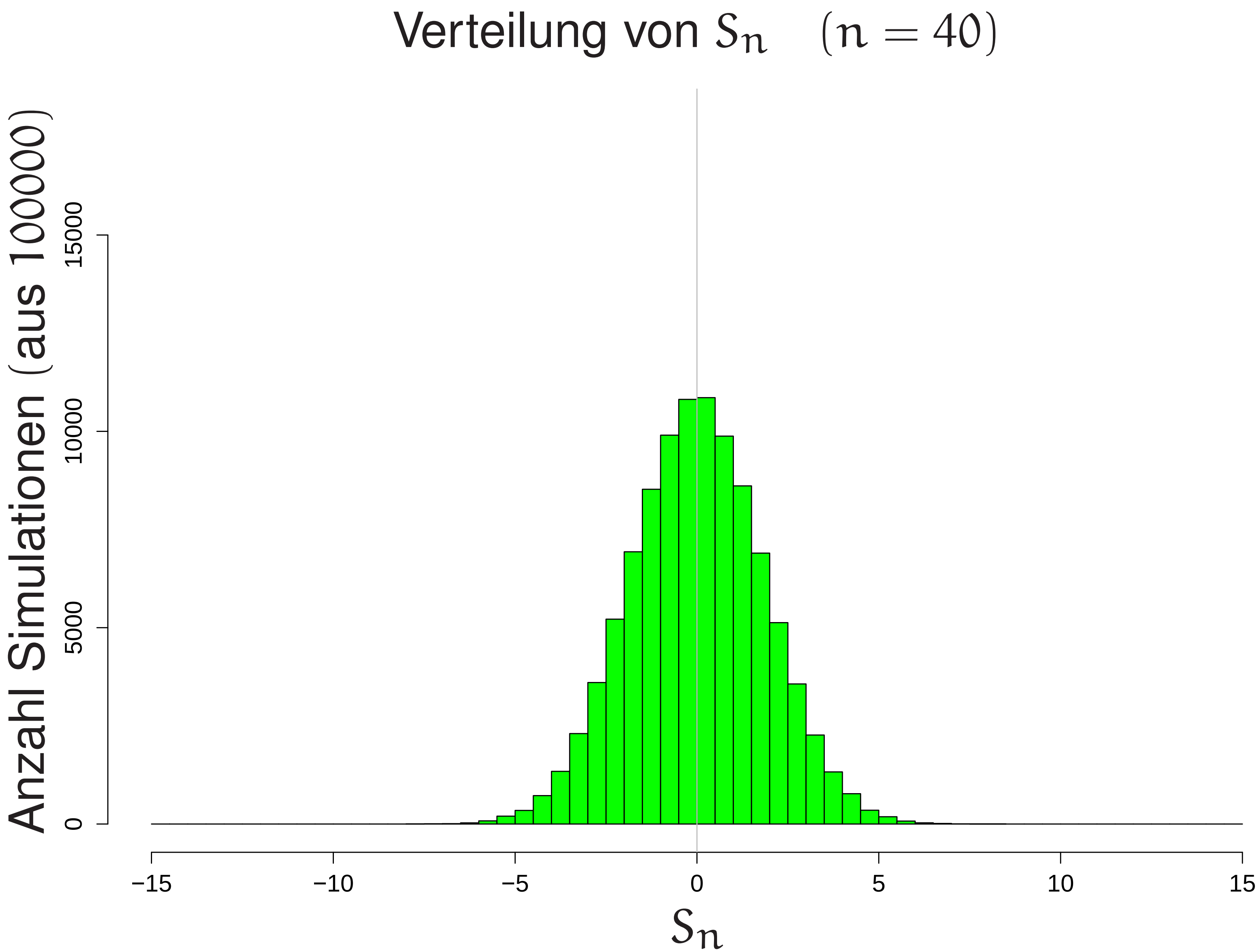


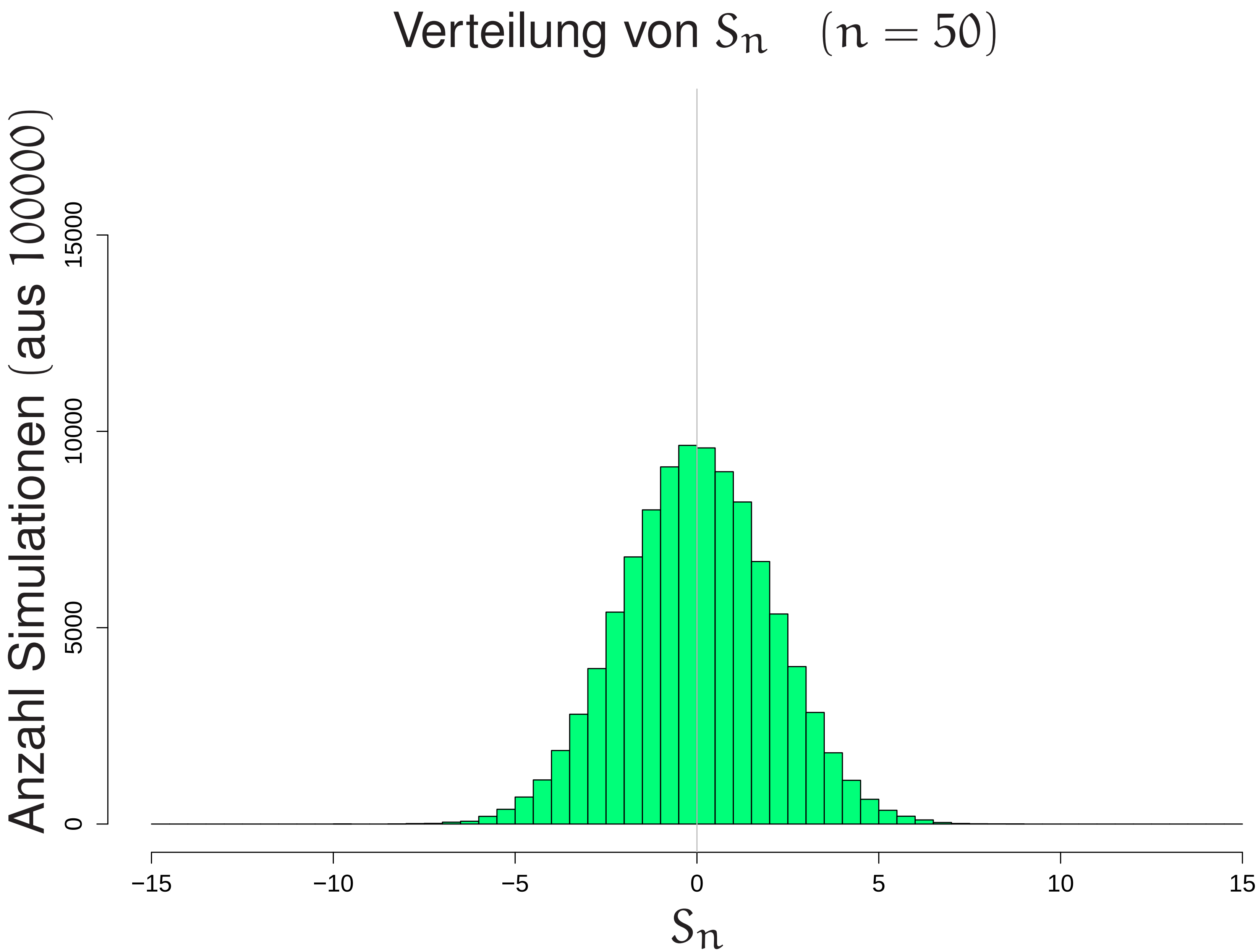
Verteilung von S_n $(n = 15)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


Verteilung von S_n $(n = 20)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

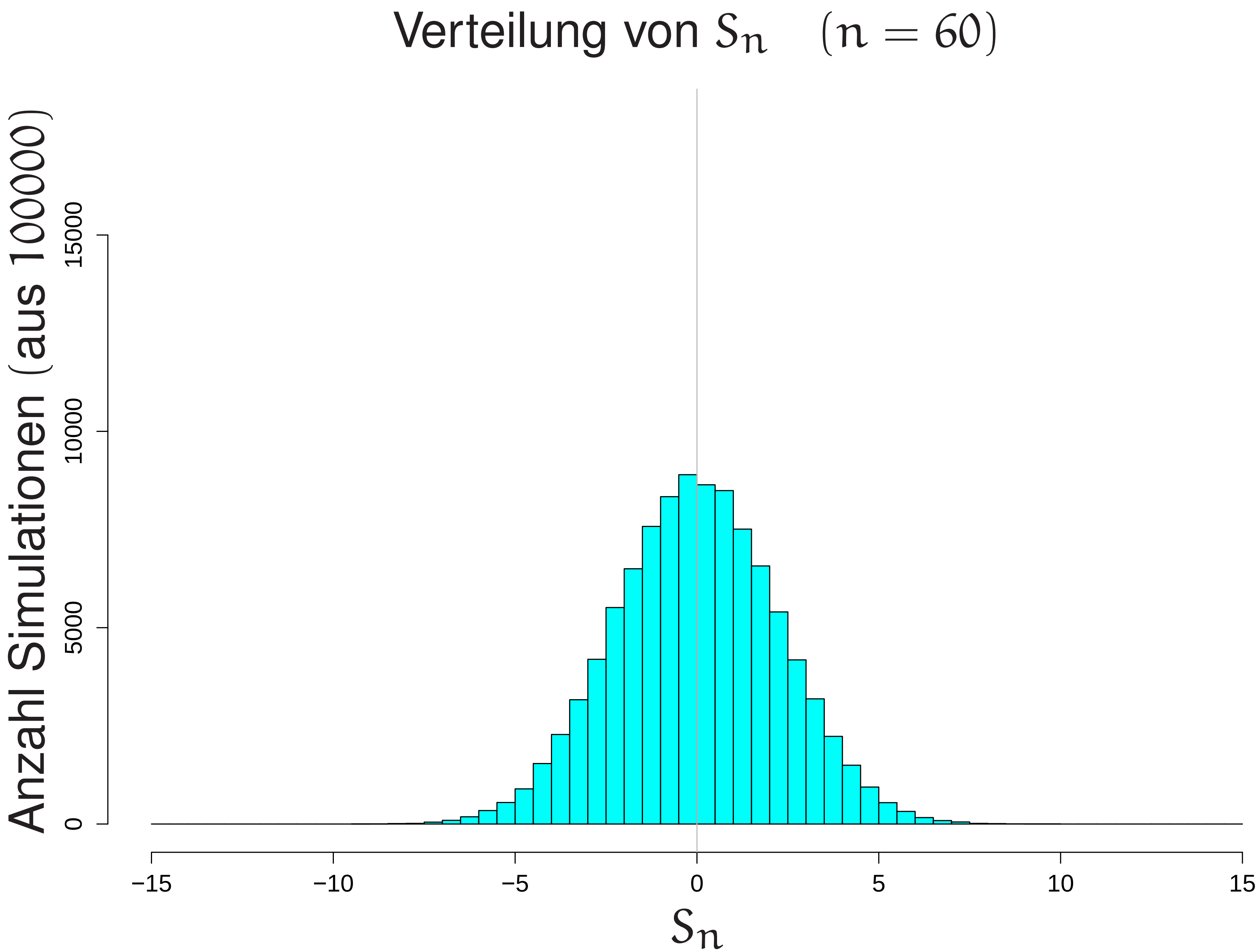


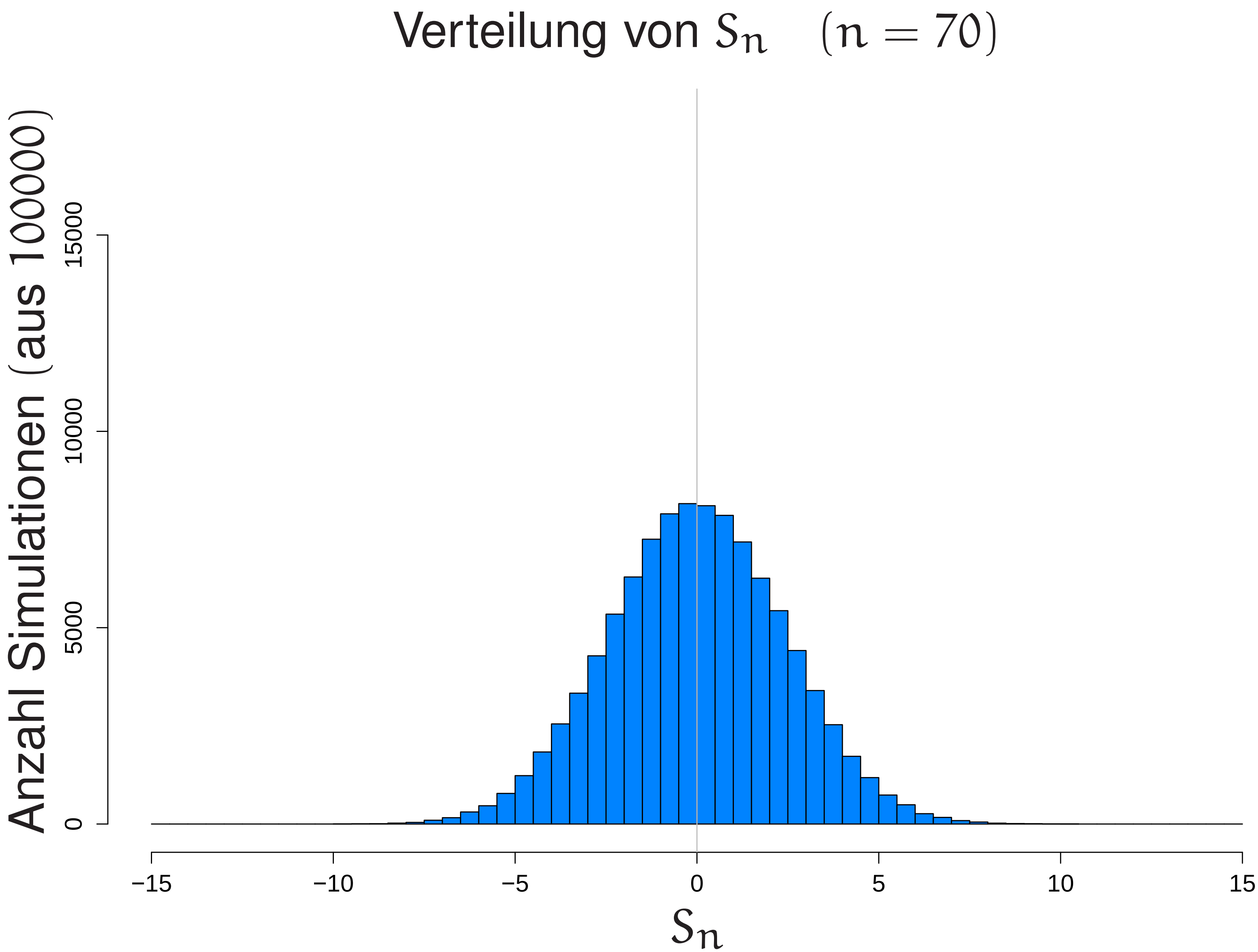
Verteilung von S_n $(n = 30)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


Verteilung von S_n $(n = 40)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

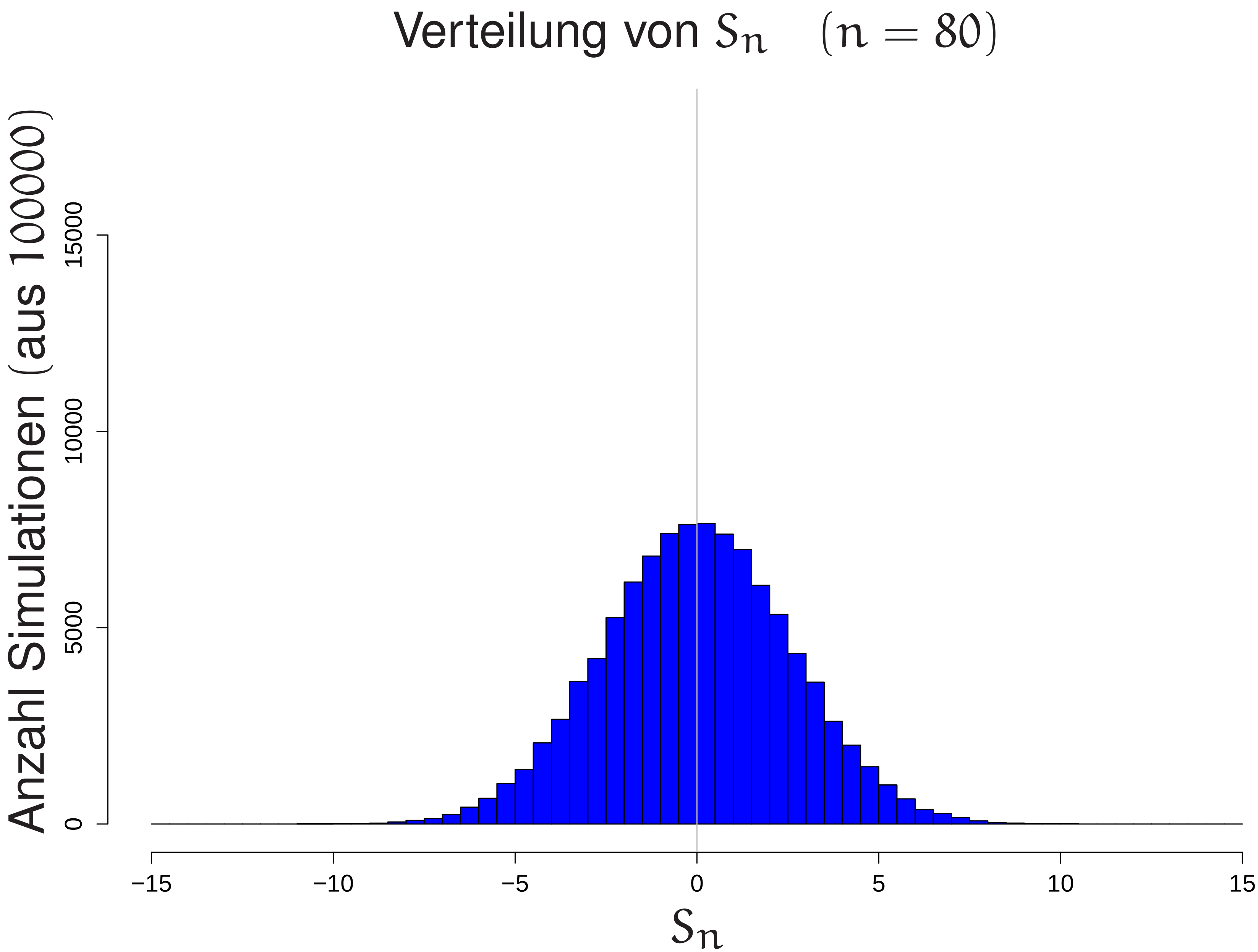


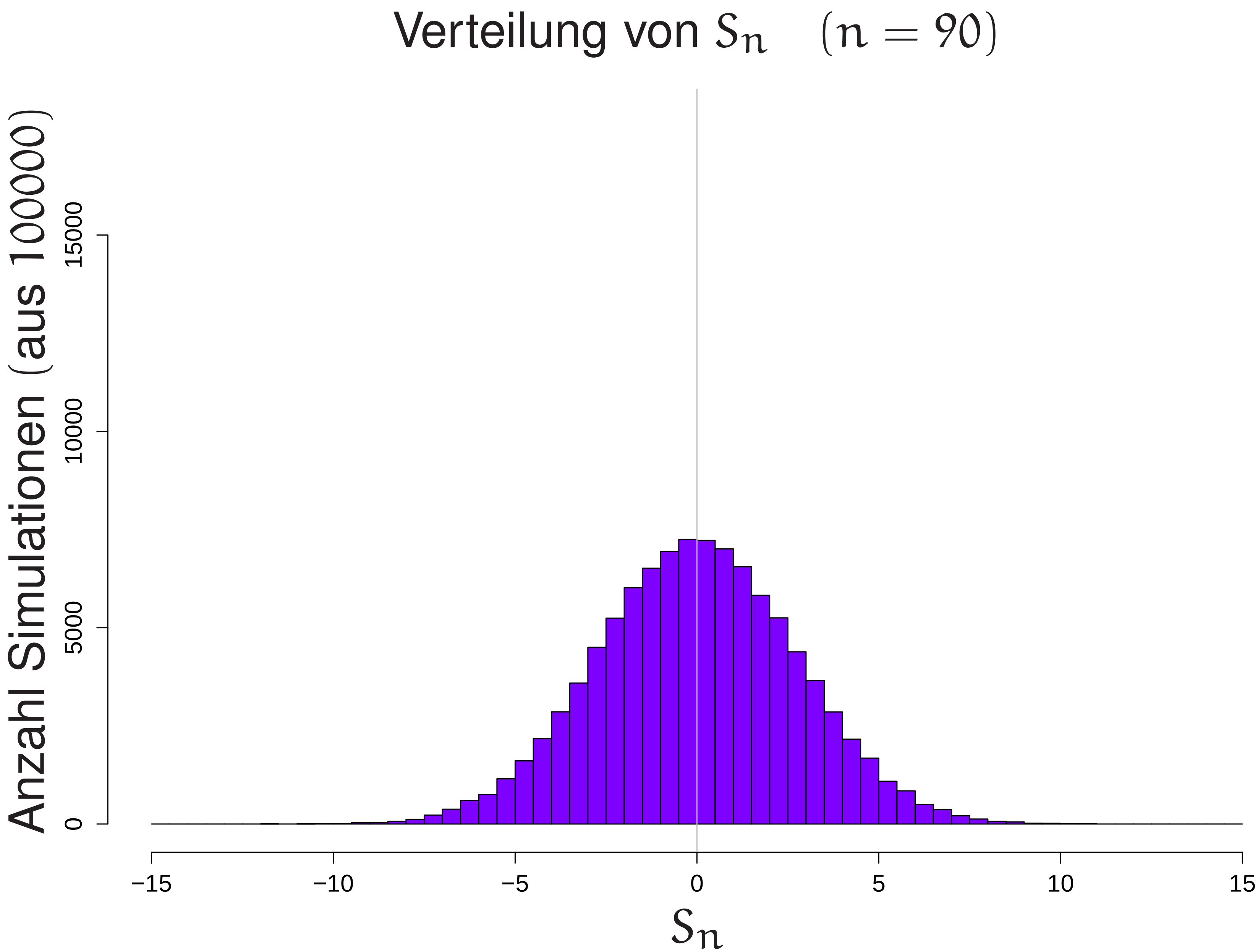
Verteilung von $S_n \quad (n = 50)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


Verteilung von S_n $(n = 60)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n



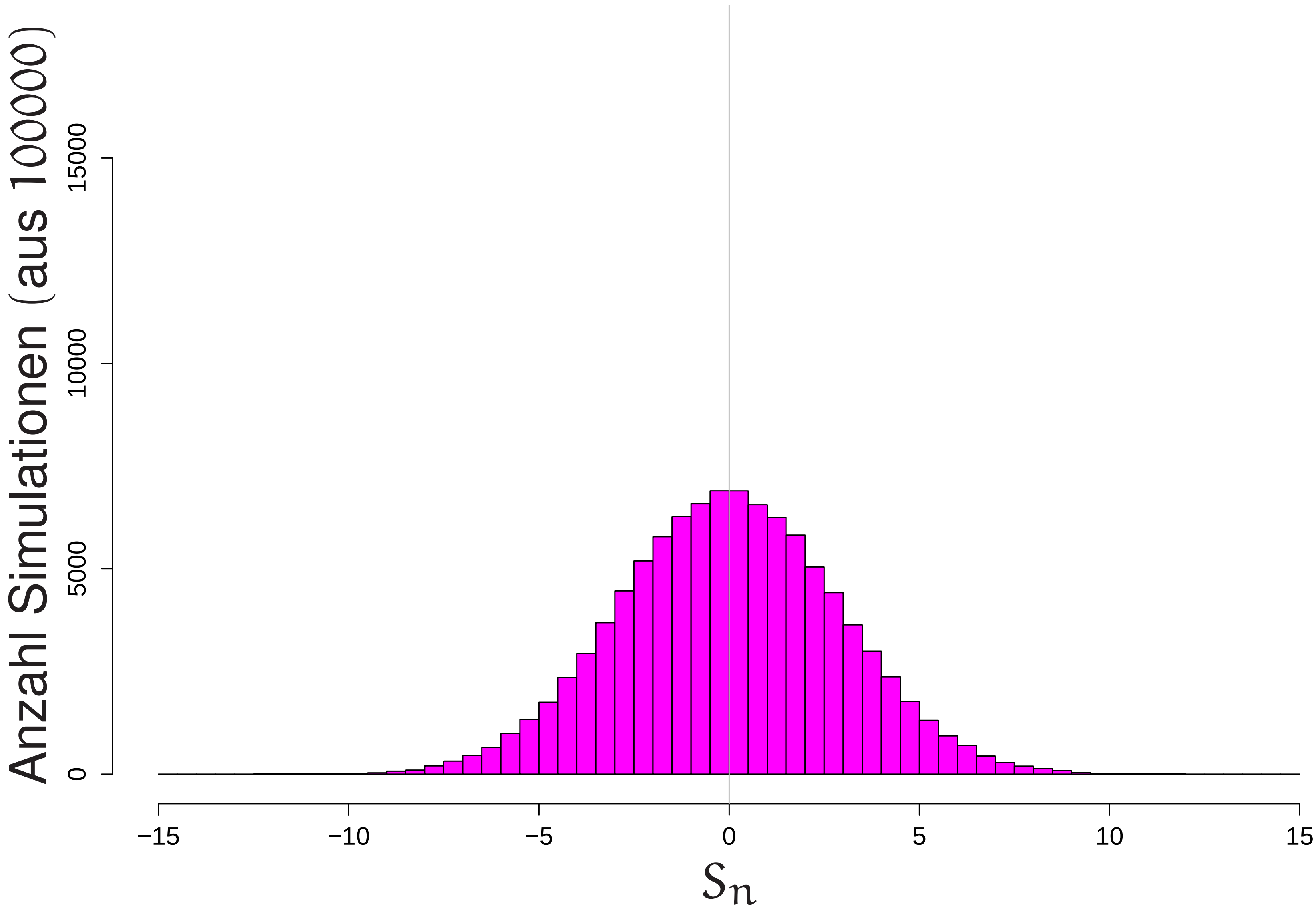
Verteilung von $S_n \quad (n = 70)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

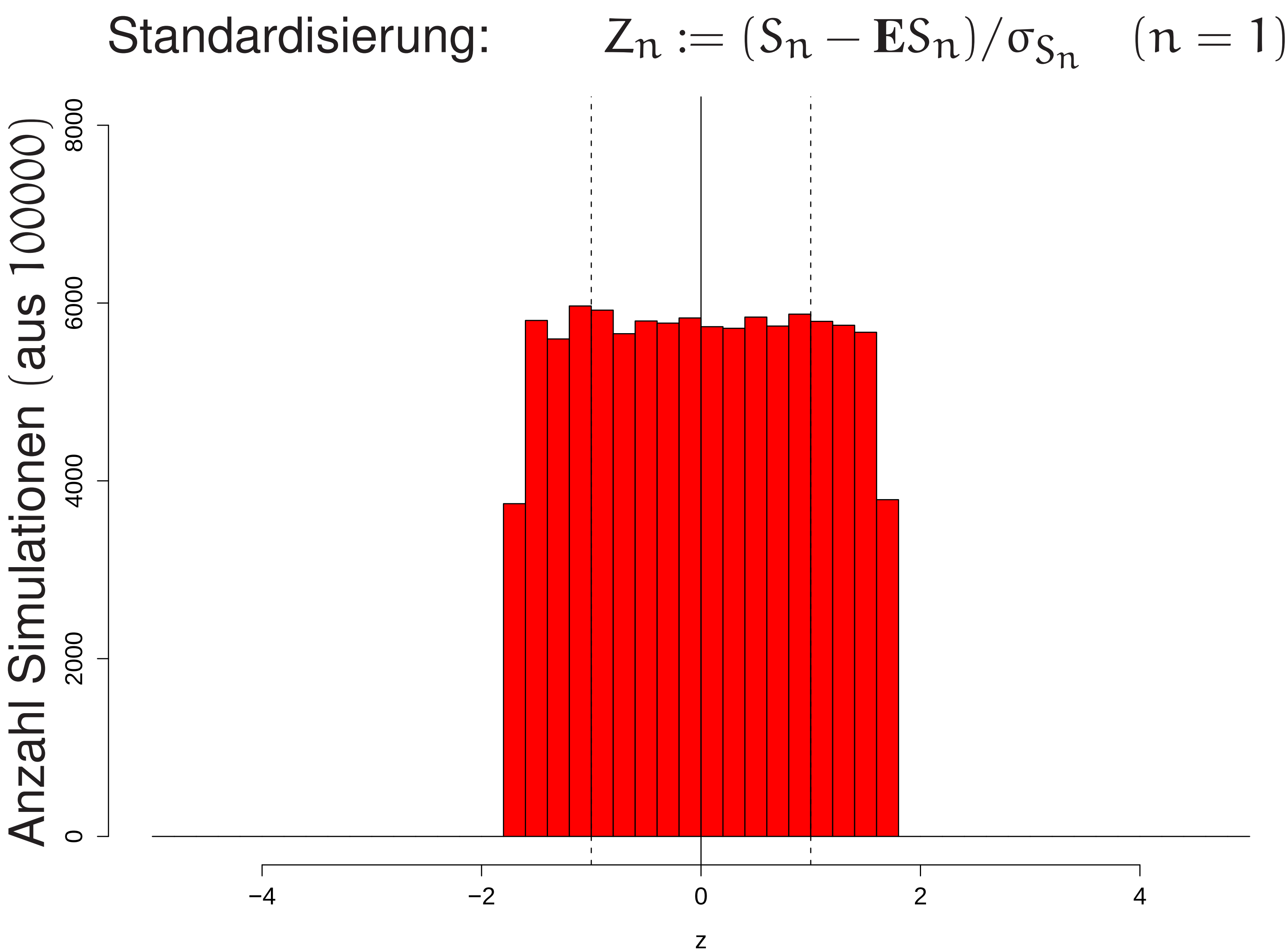

Verteilung von S_n $(n = 80)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

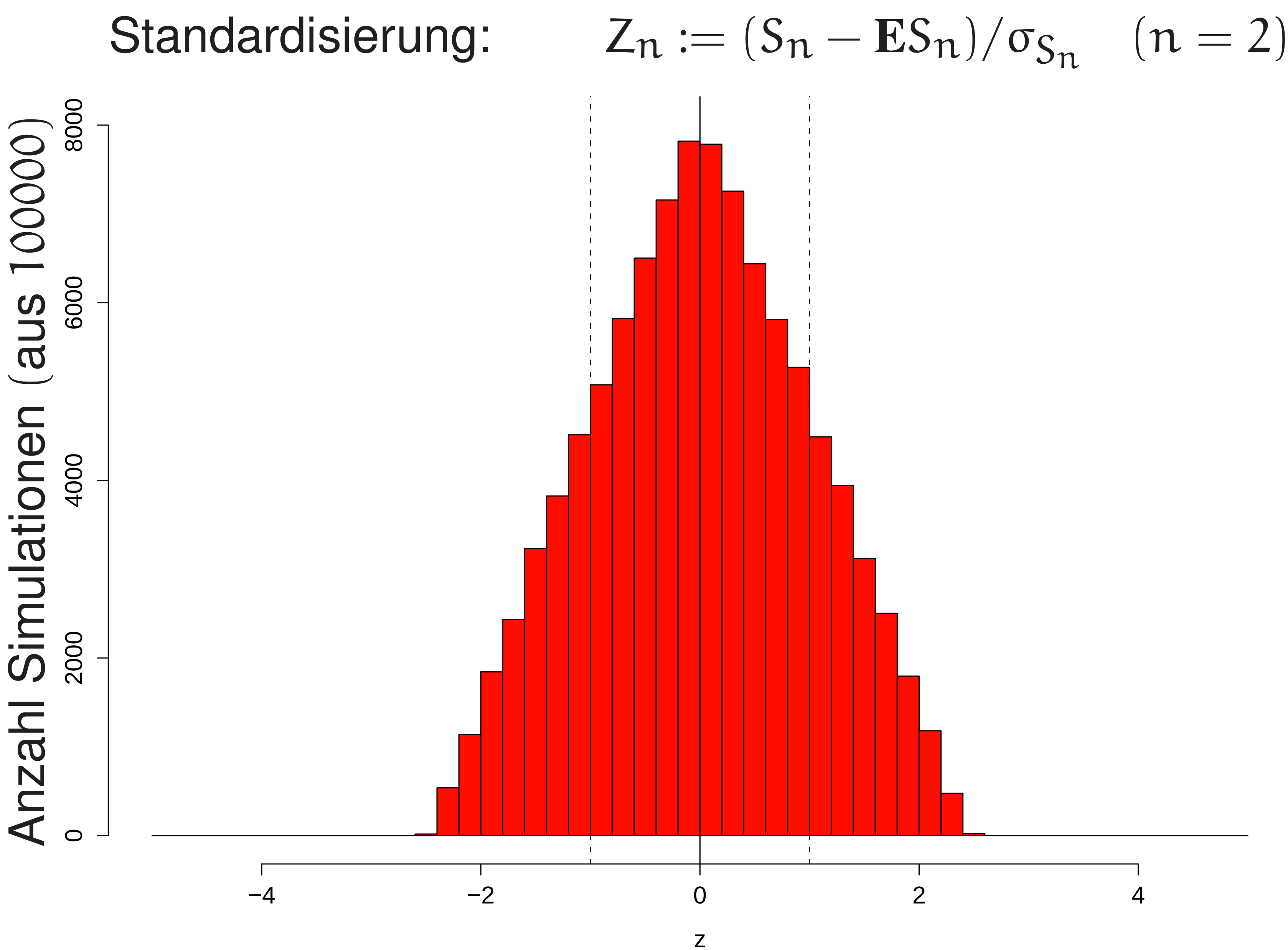


Verteilung von S_n $(n = 90)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

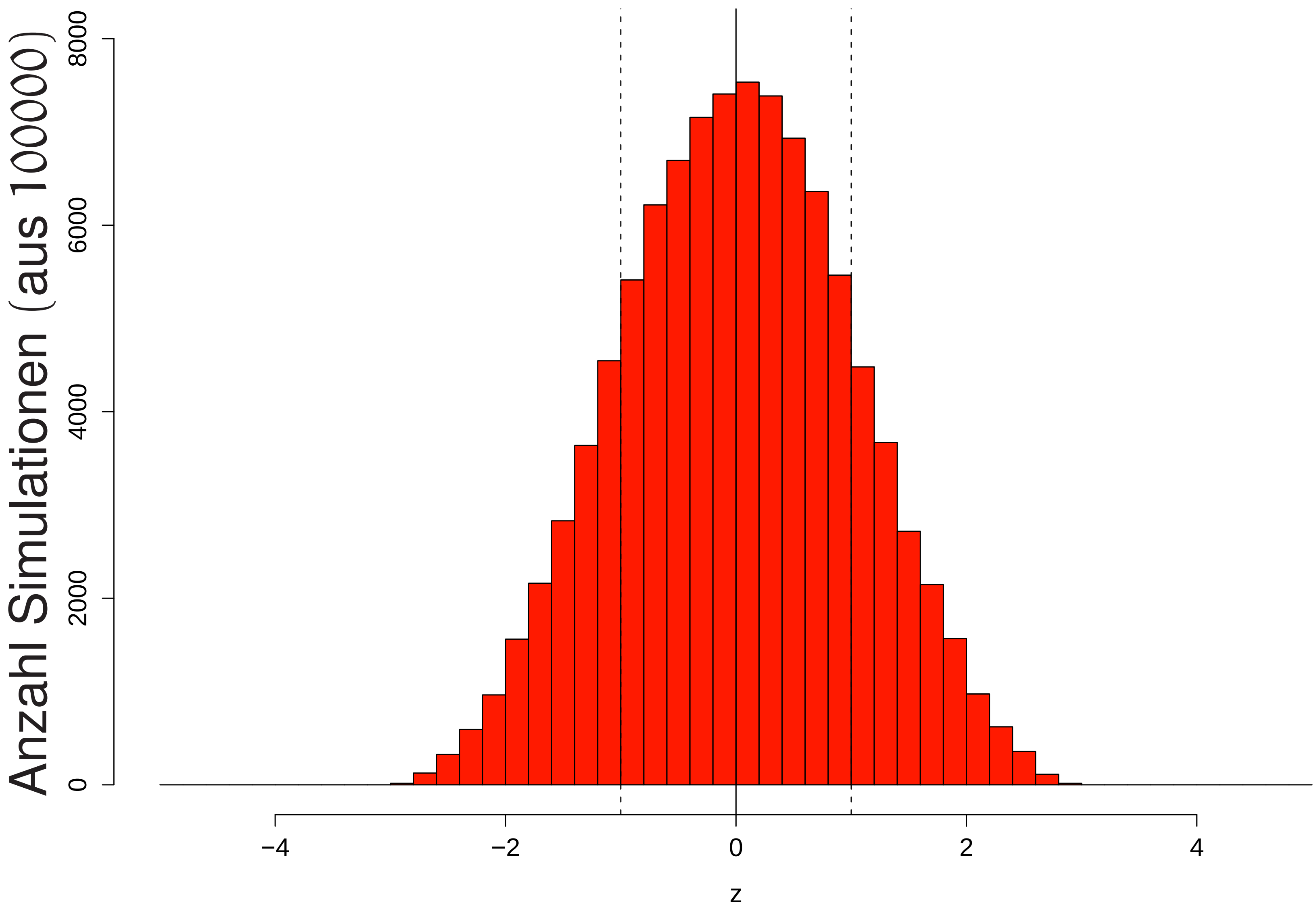
Verteilung von $S_n \quad (n = 100)$



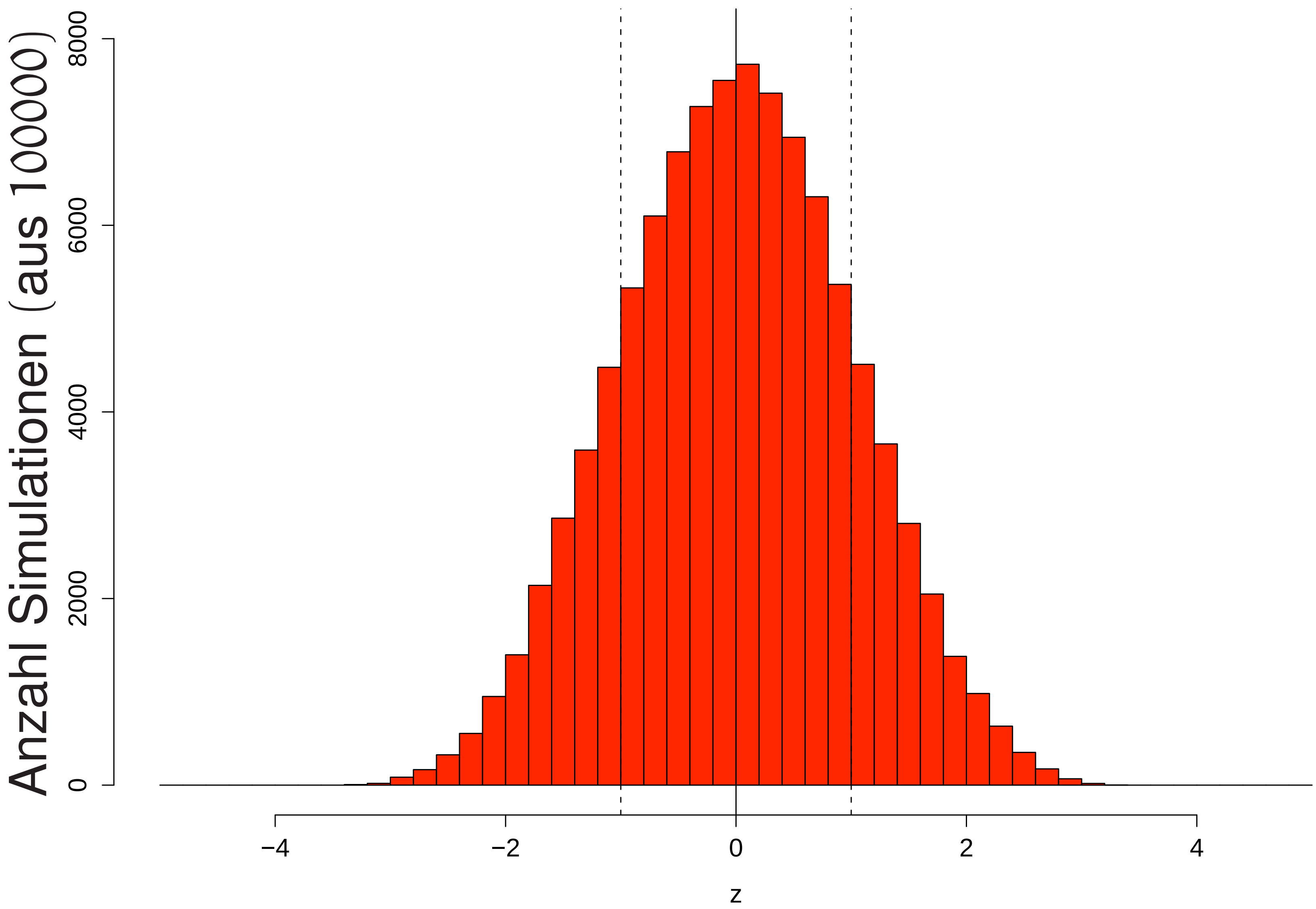

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 1)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

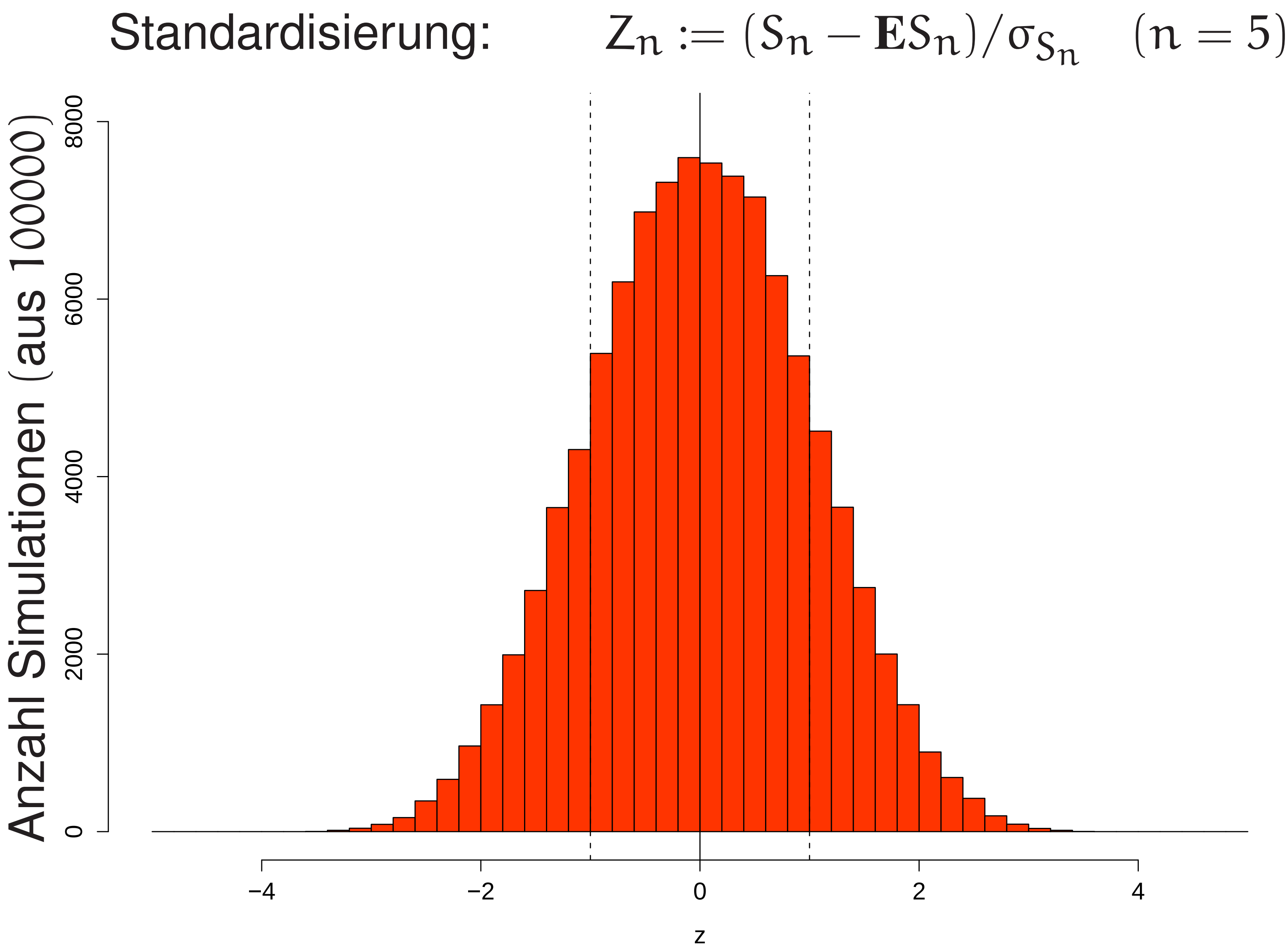

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 2)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 3)$

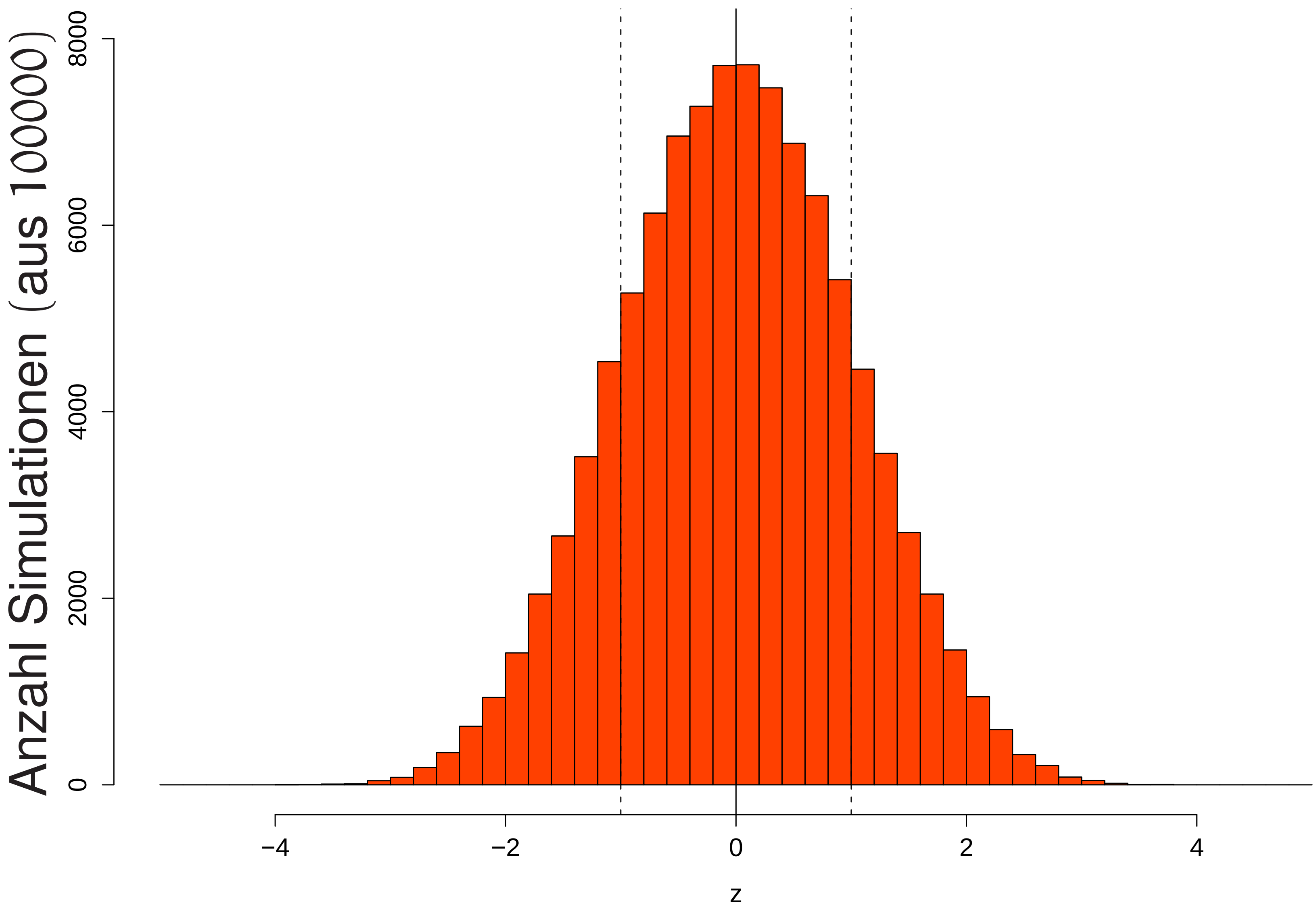


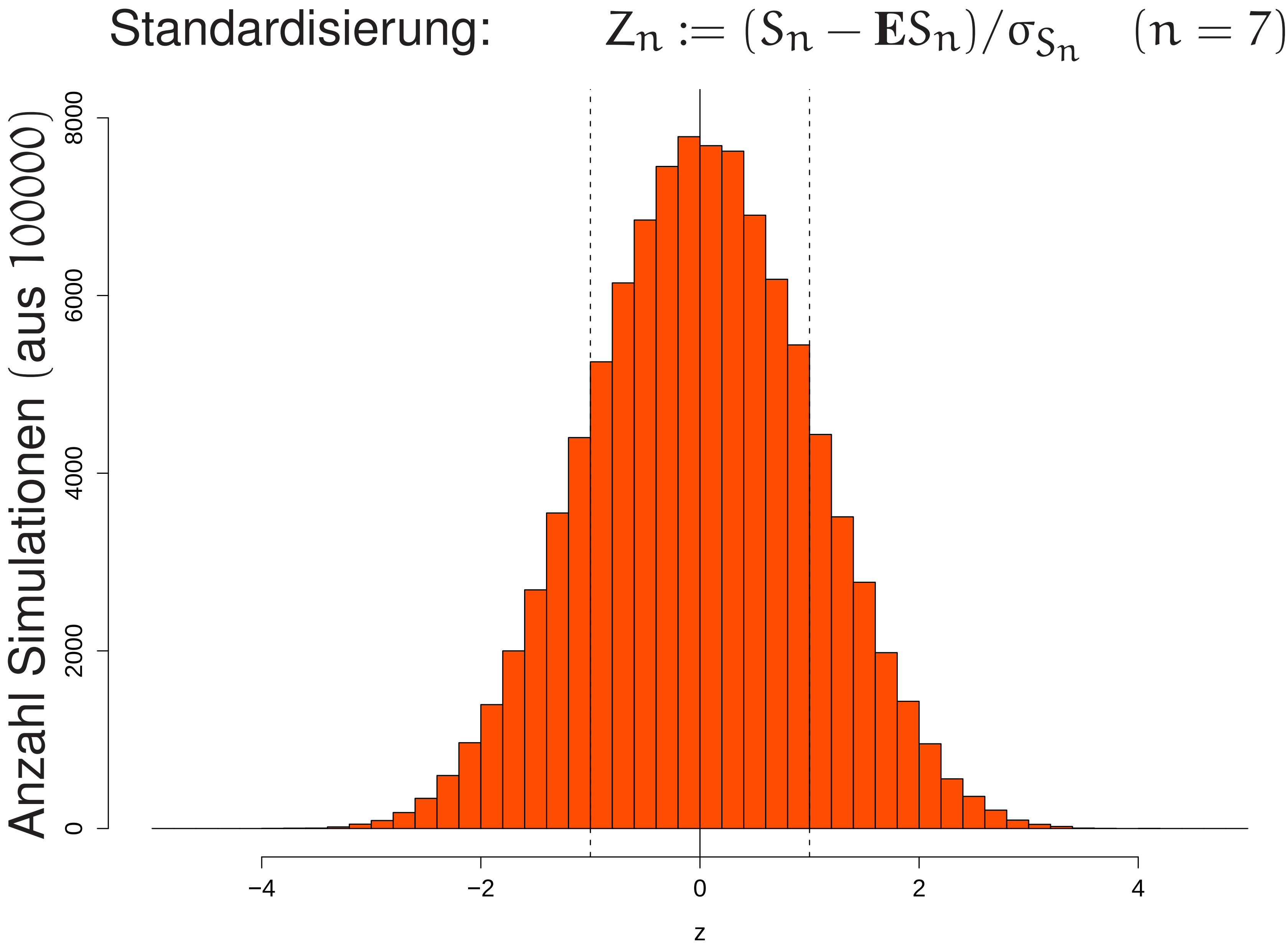
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 4)$



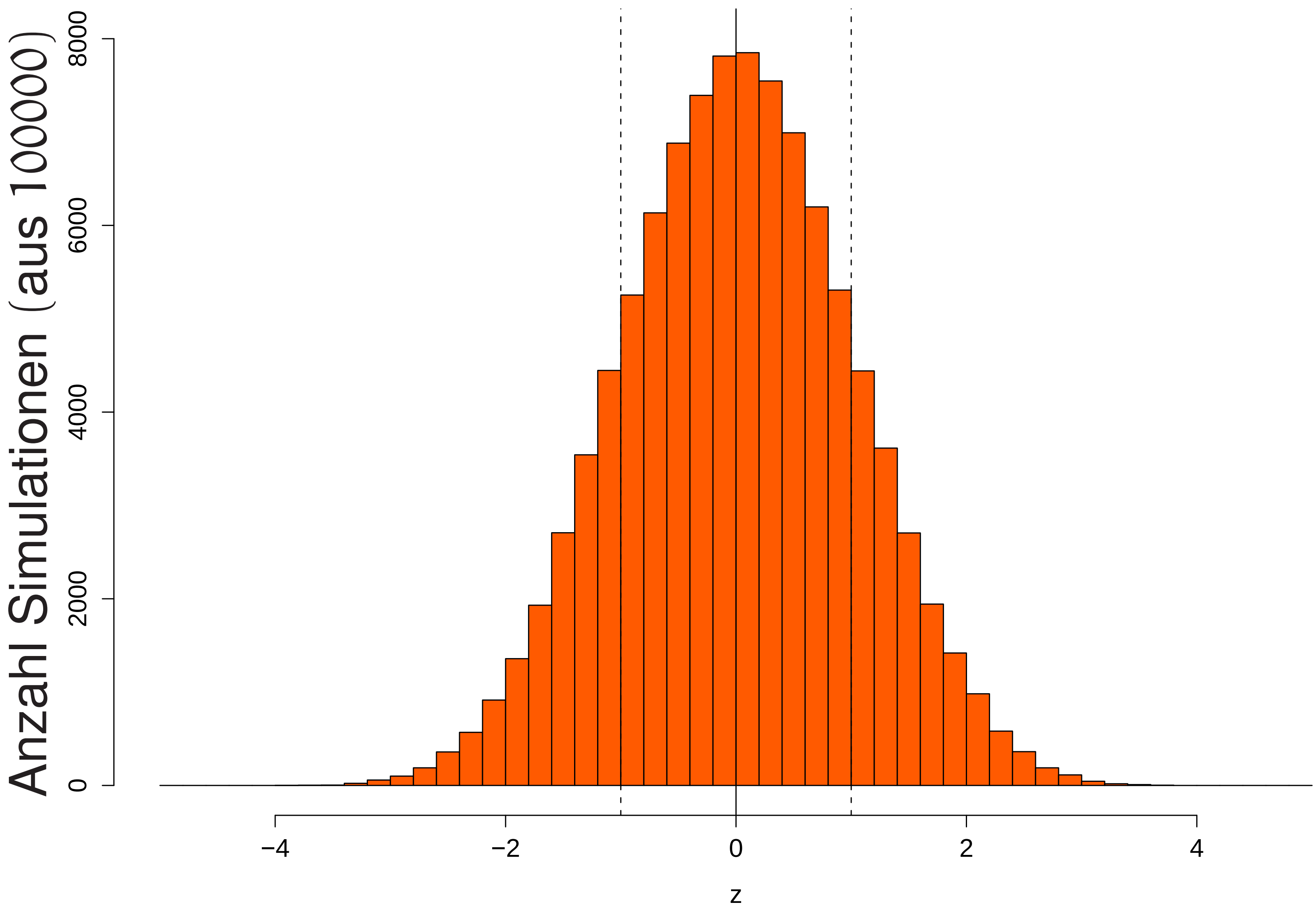

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 5)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 6)$



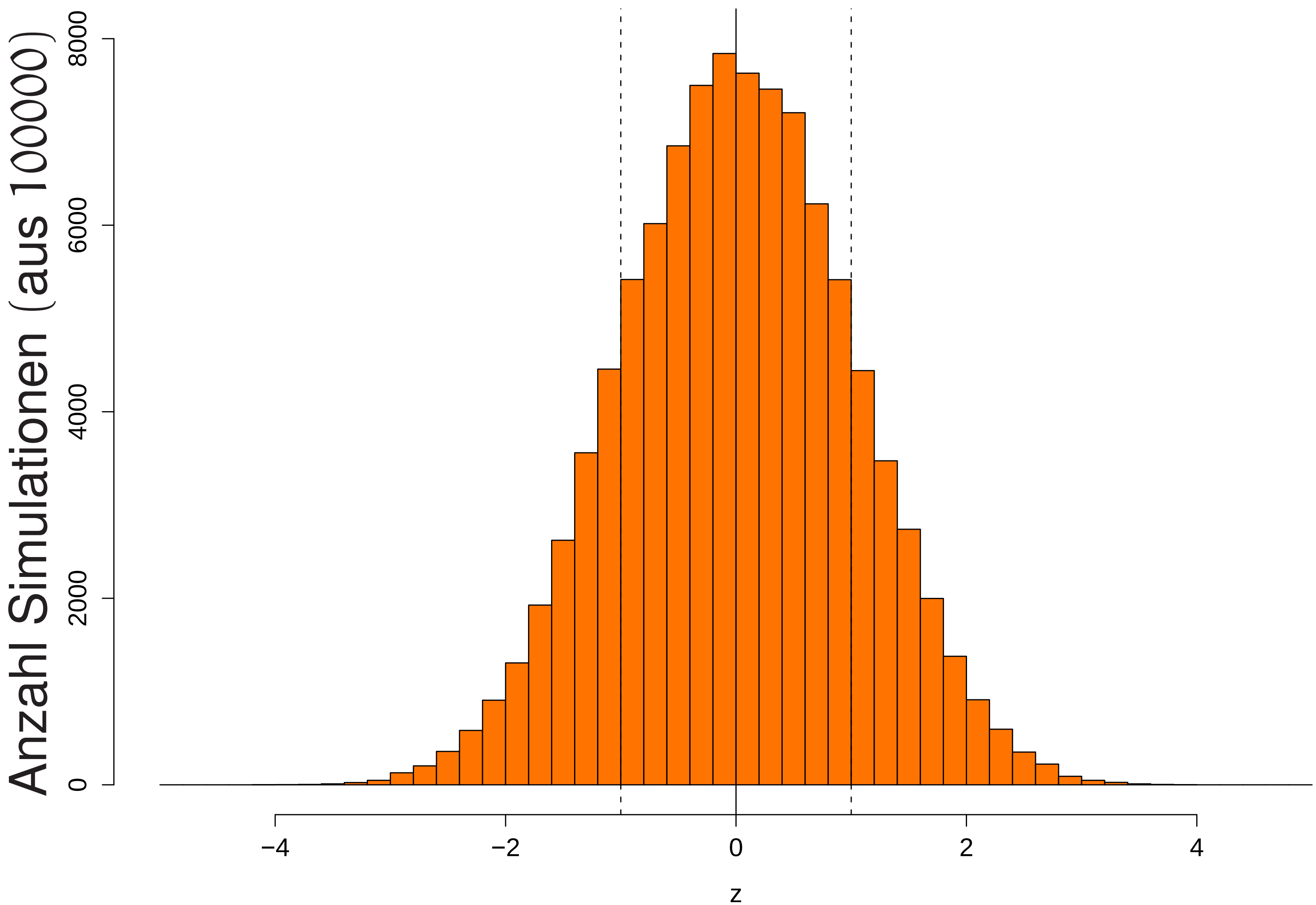

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 7)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 8)$



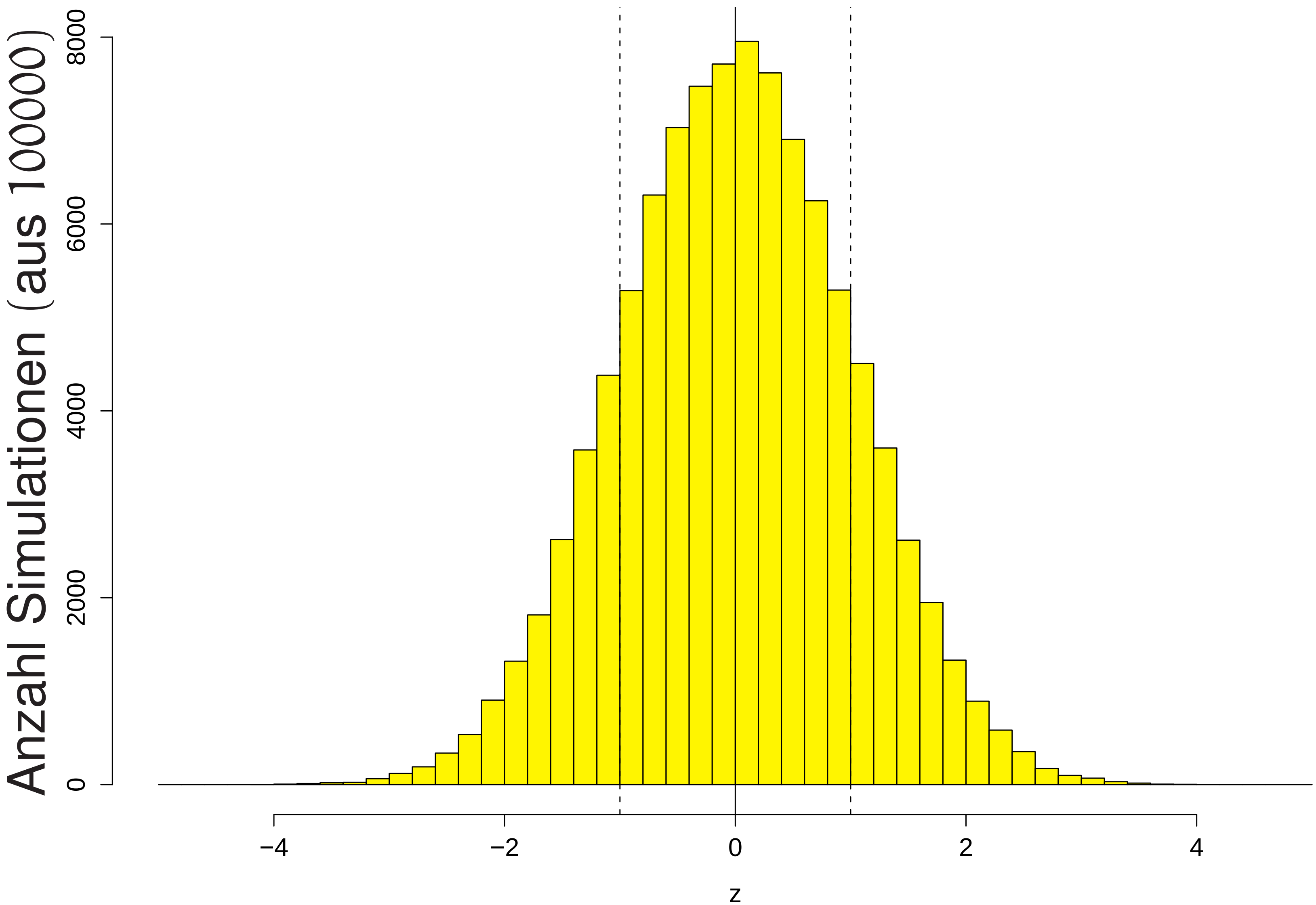
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 9)$

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 10)$

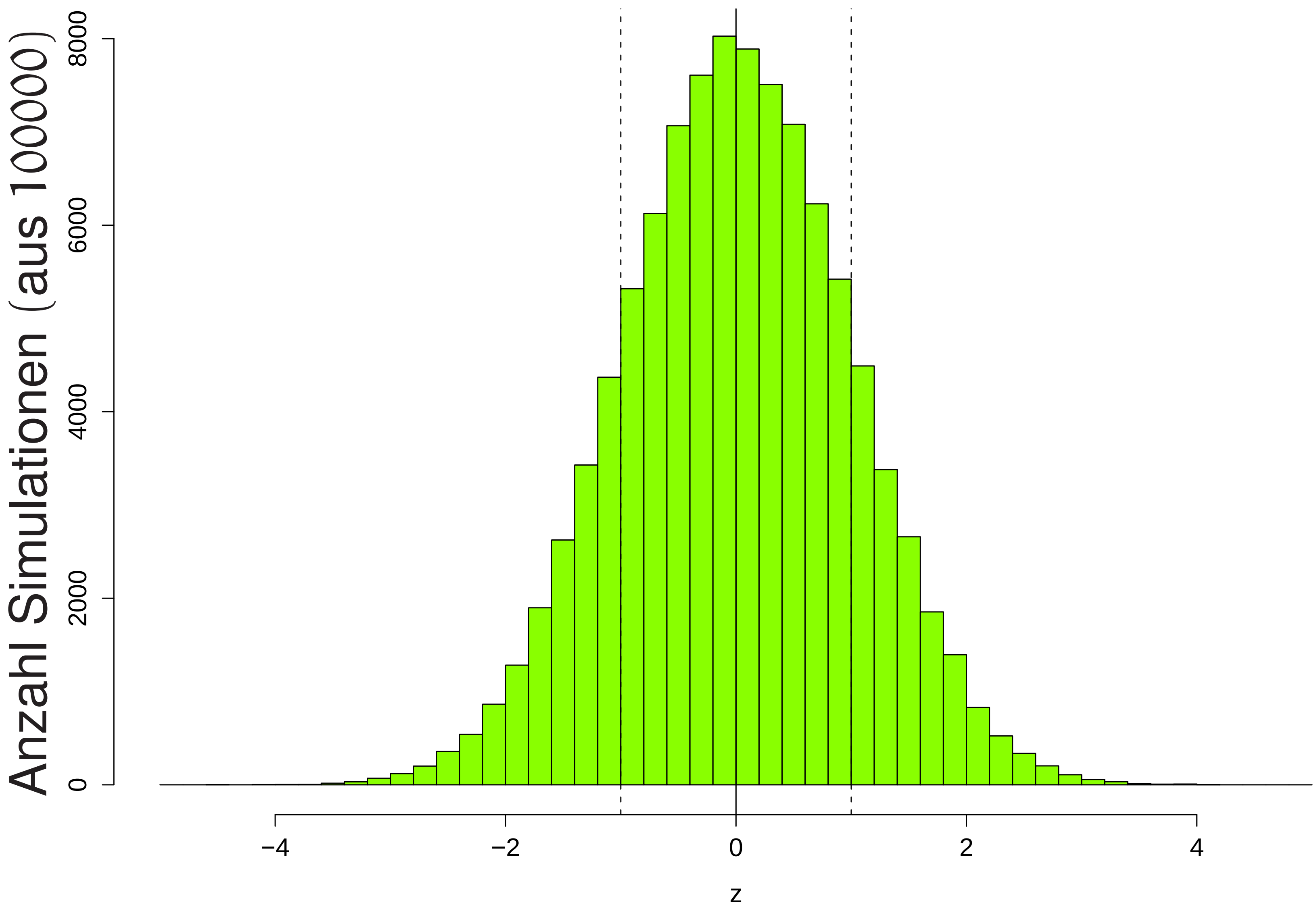


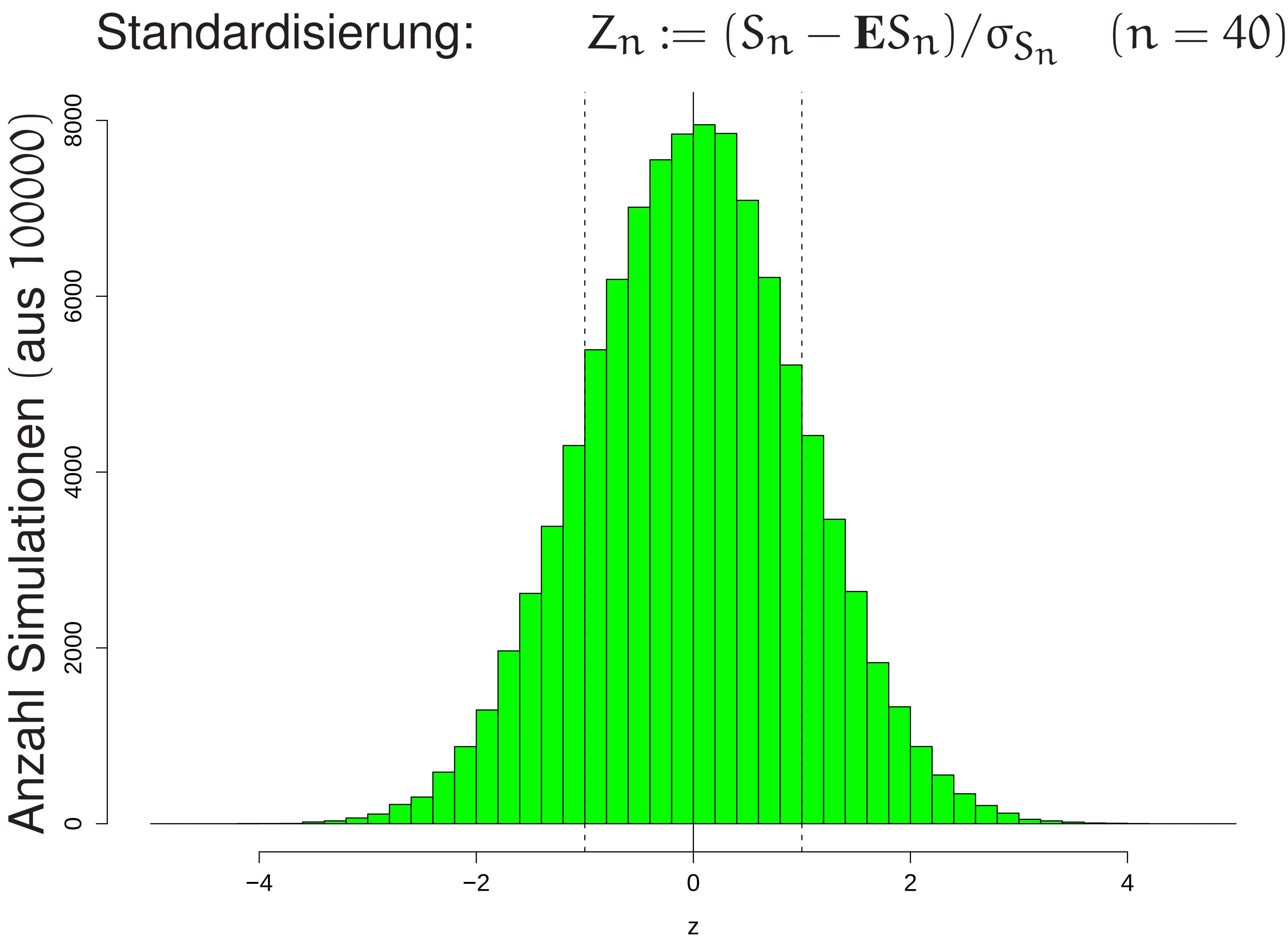
Standardisierung: $\quad Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 15)$

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 20)$



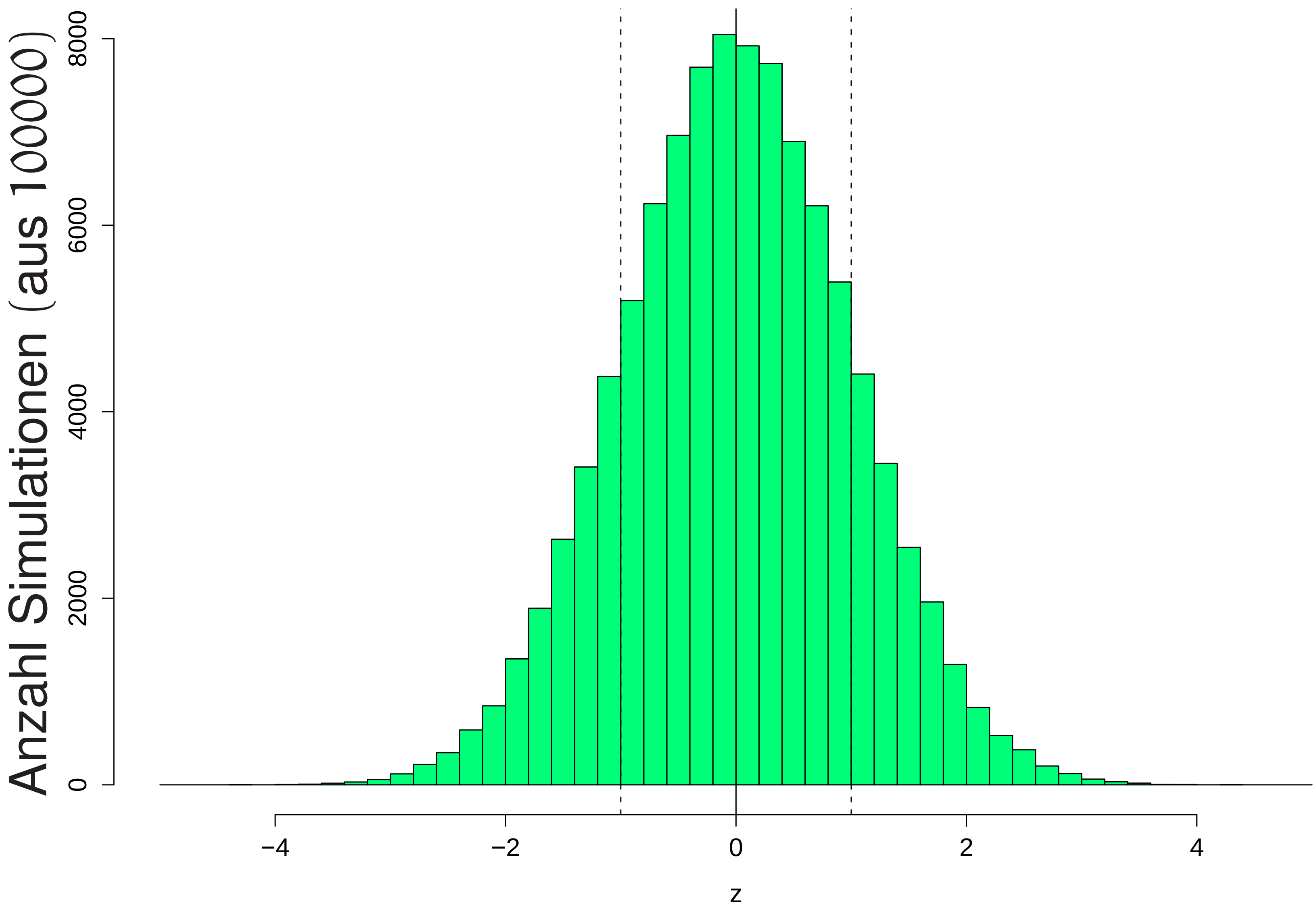
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 30)$

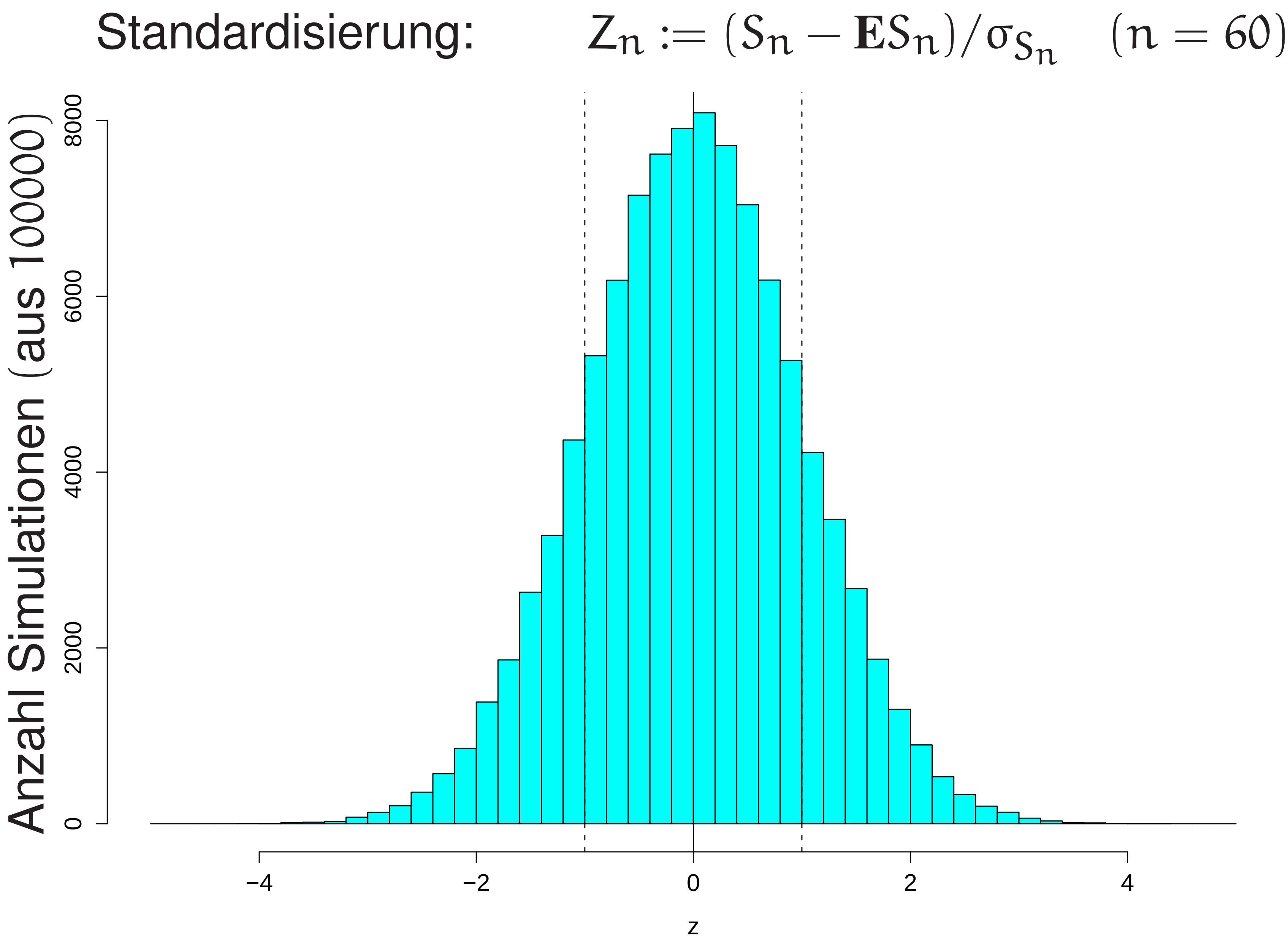




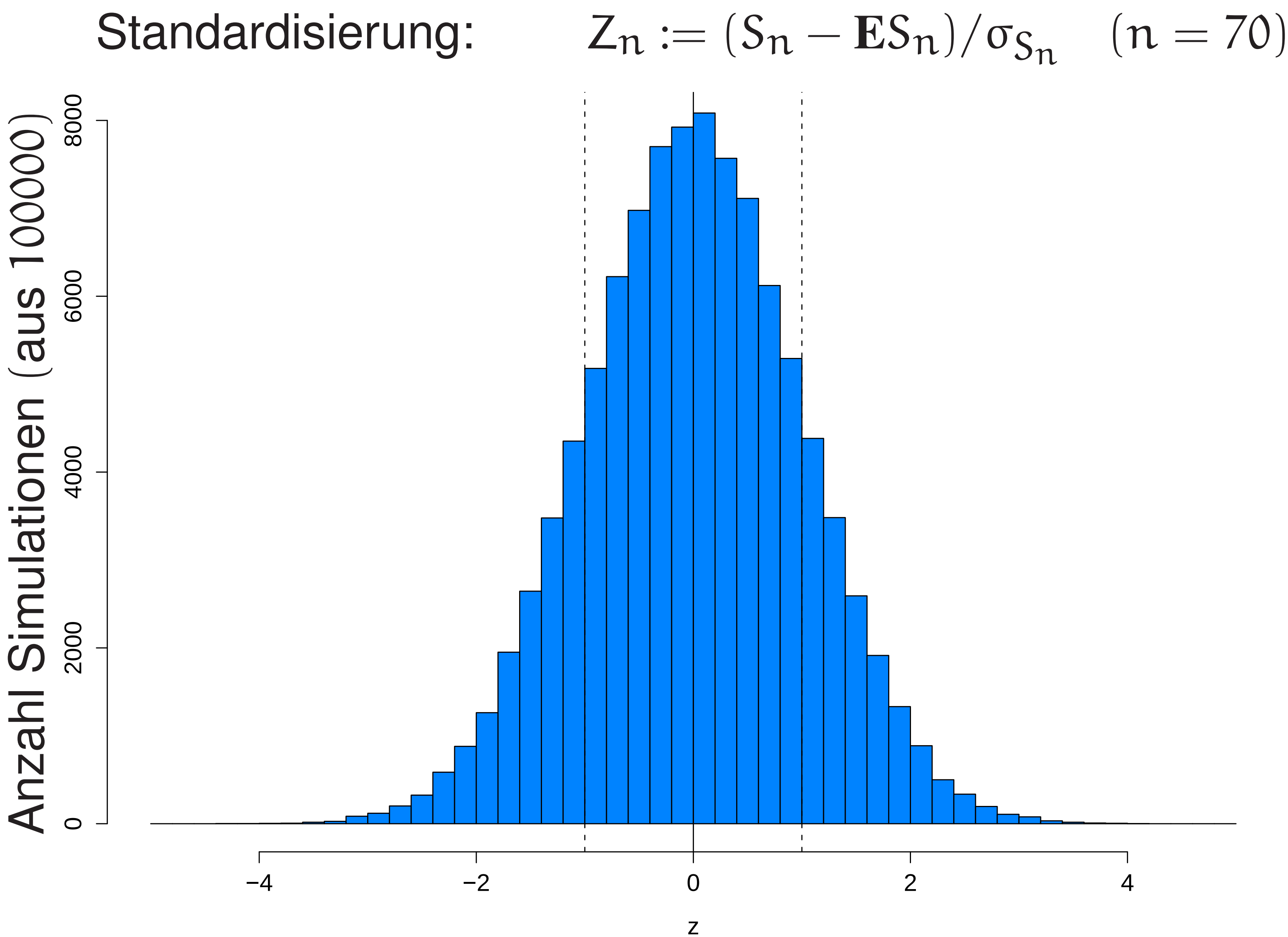
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 40)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 50)$

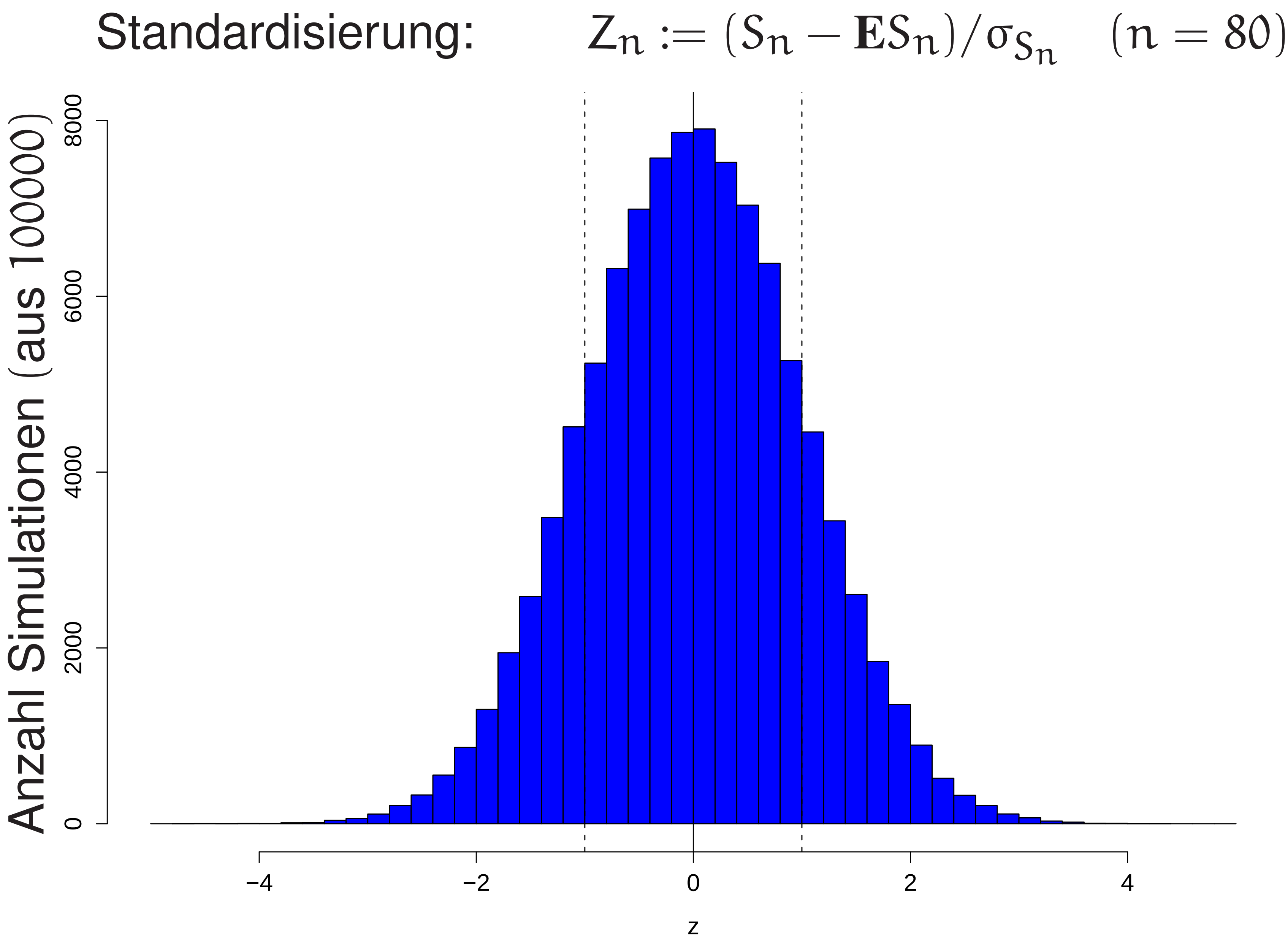




Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 60)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

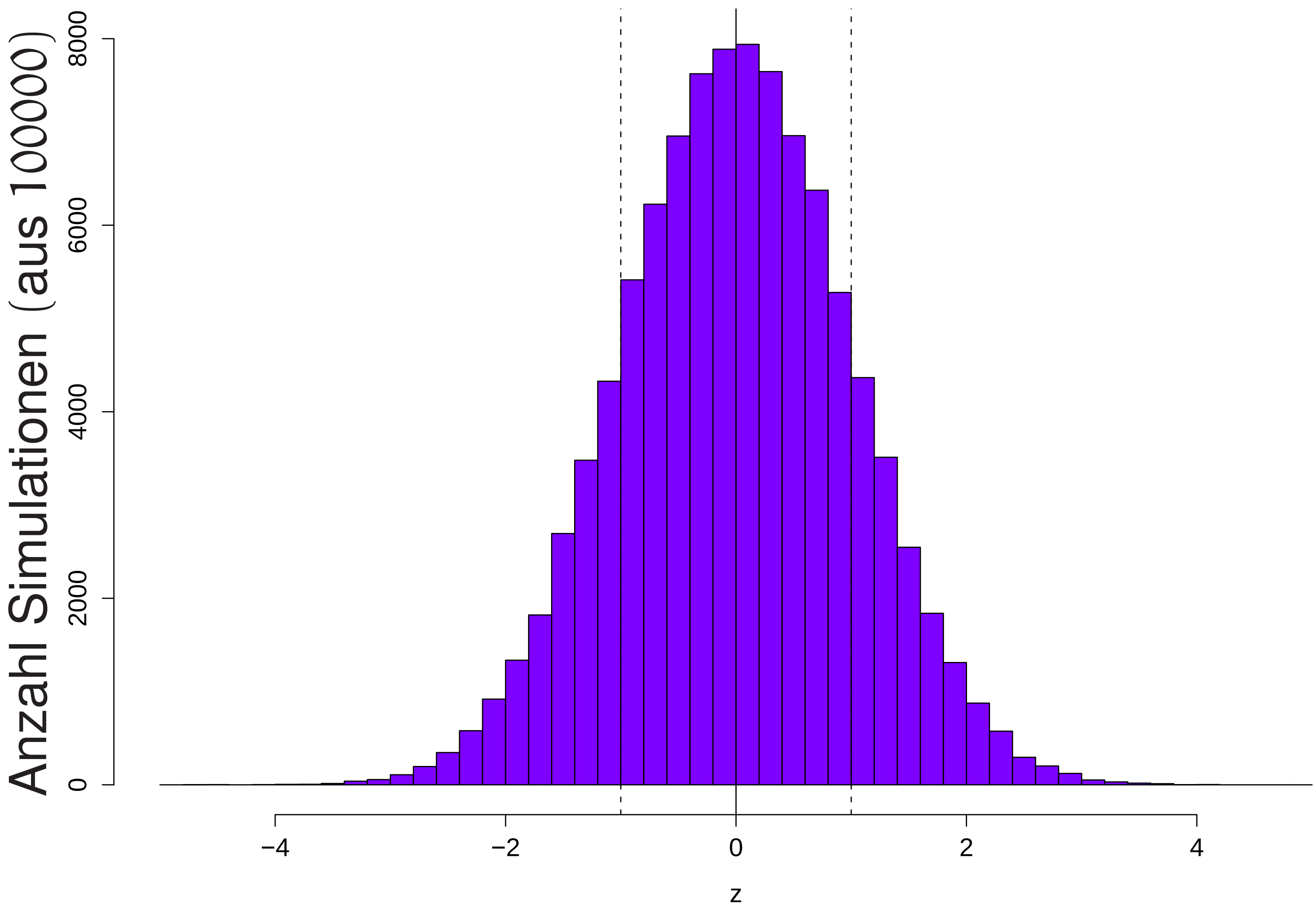


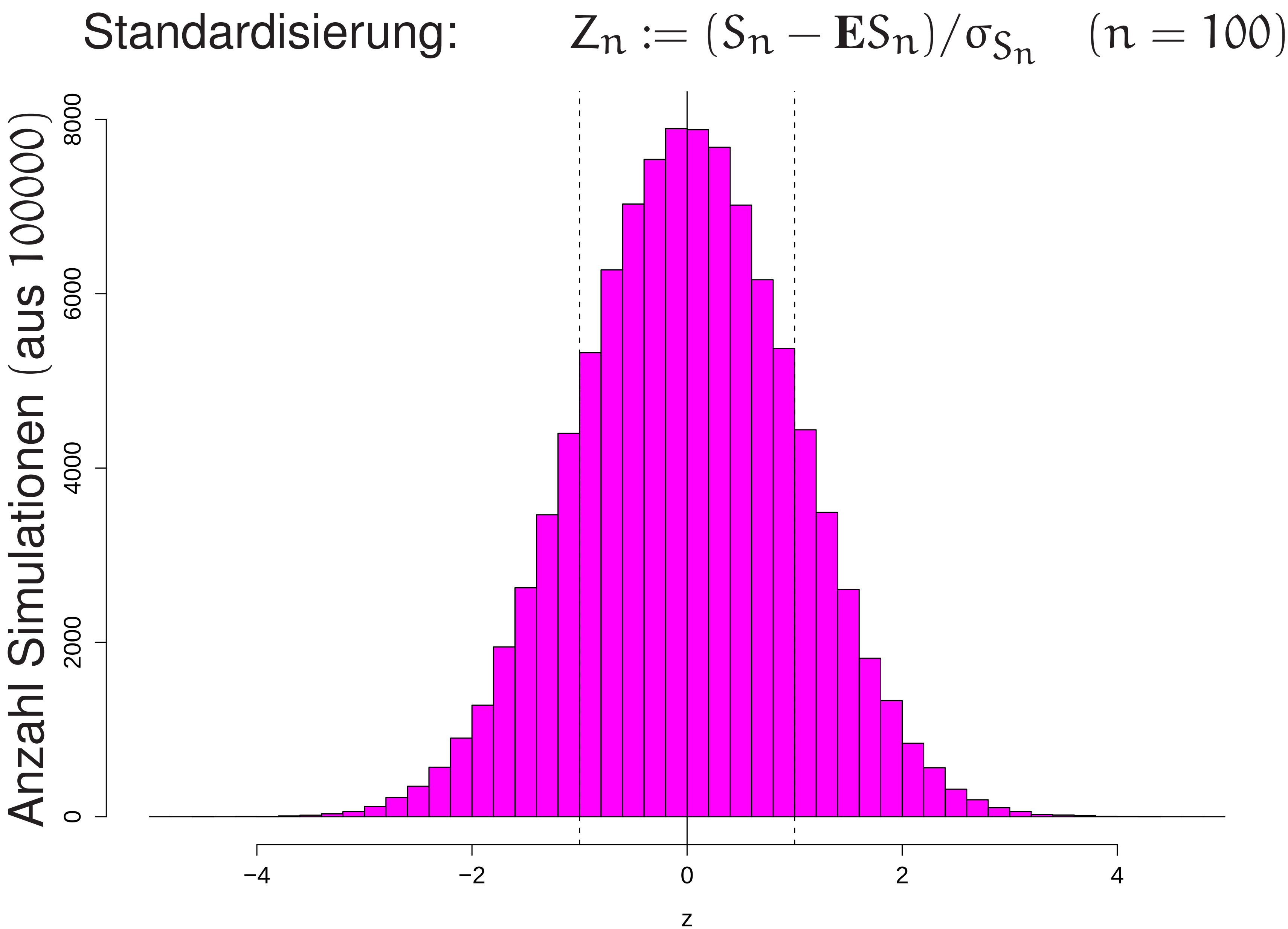
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 70)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z



Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 80)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

Standardisierung: $\quad Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 90)$





Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 100)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−4
−2
0
2
4
z

Die Verteilung von Z_n

scheint zu konvergieren.

Welche Form

hat die Grenzverteilung?

Die Verteilung von

$$Z_{100}$$

ist glockenförmig.

Um welche Glockenkurve handelt es sich genau?

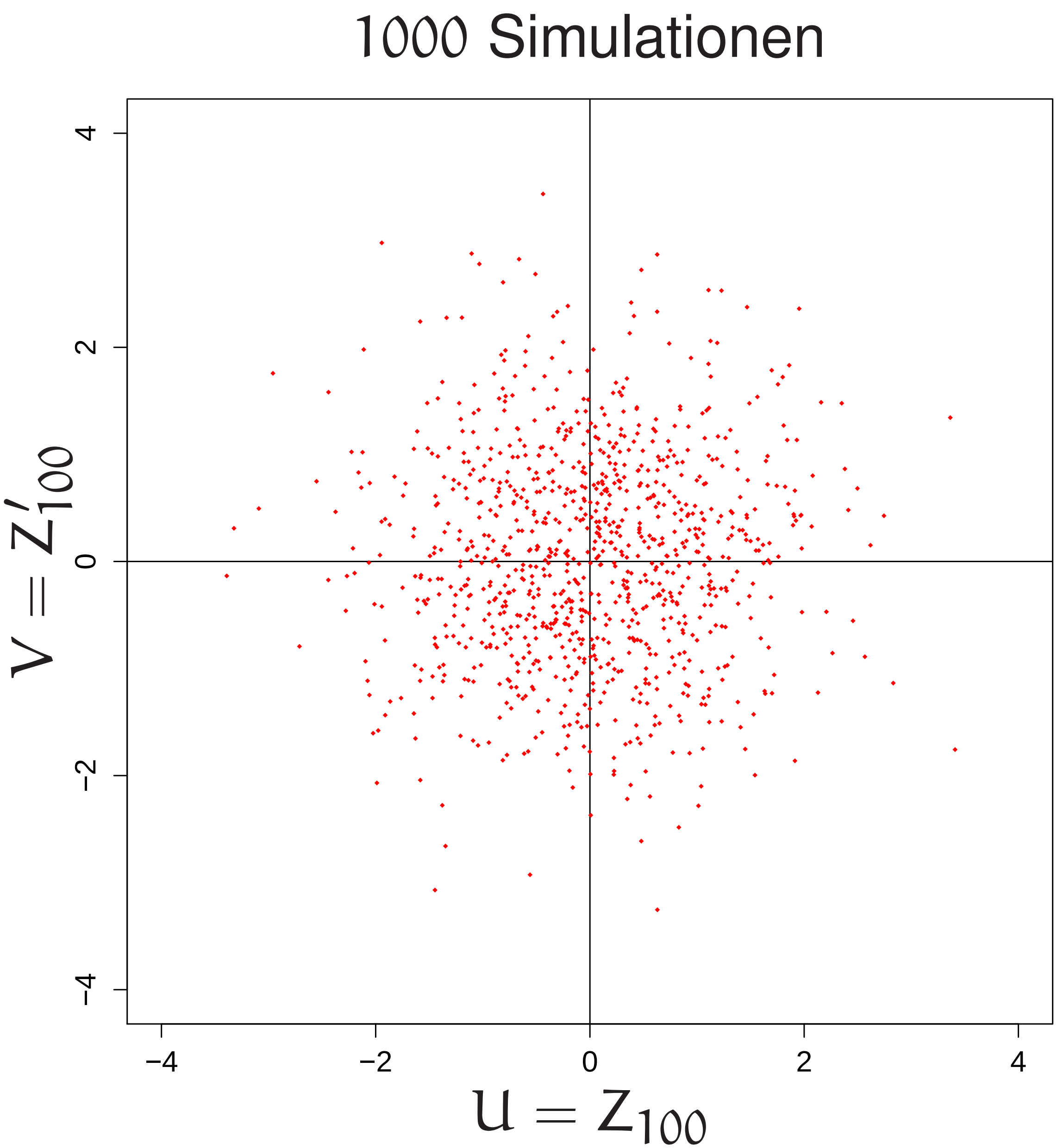
Glücklicher Einfall:

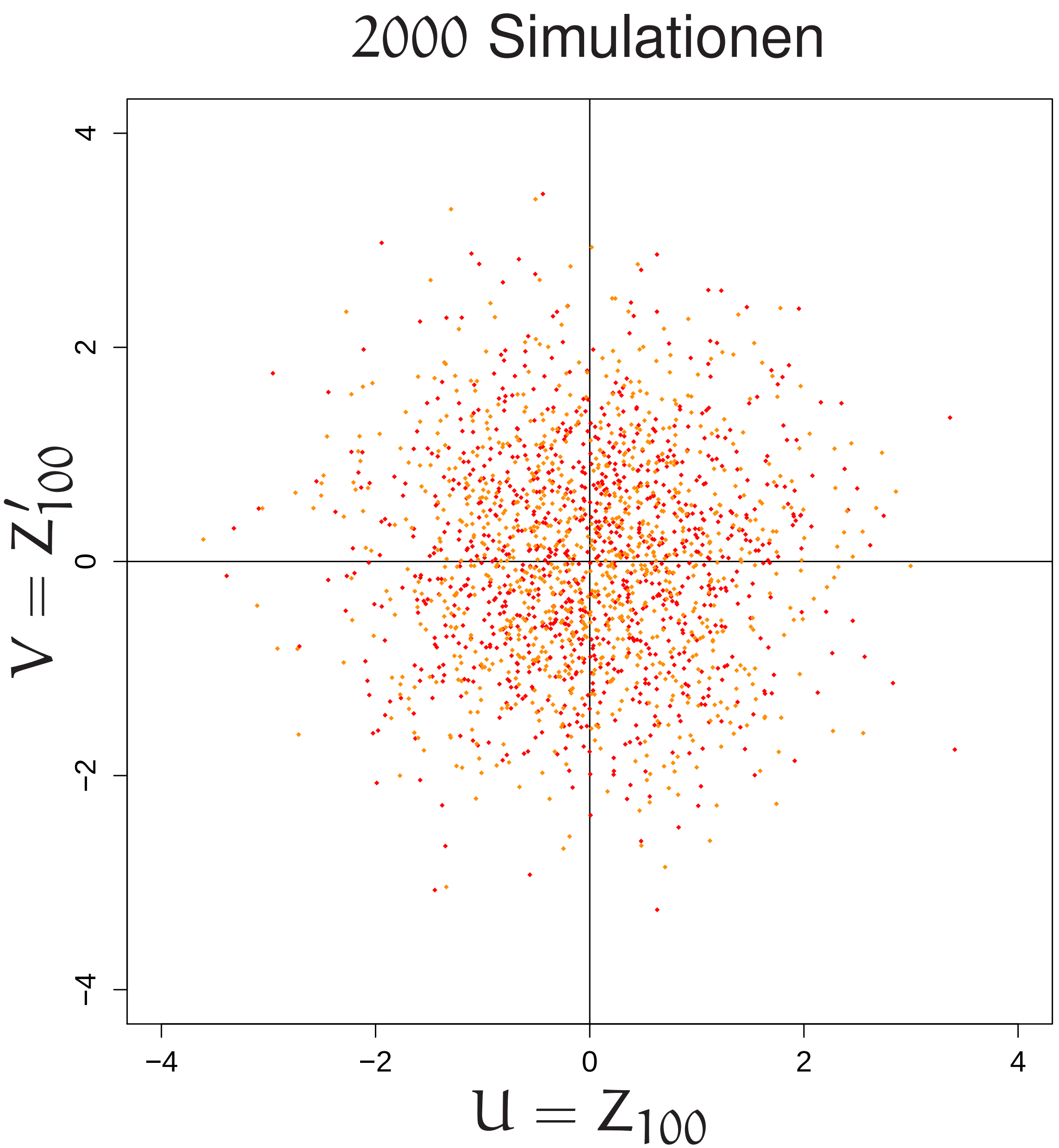
Nimm zwei unabhängige Kopien

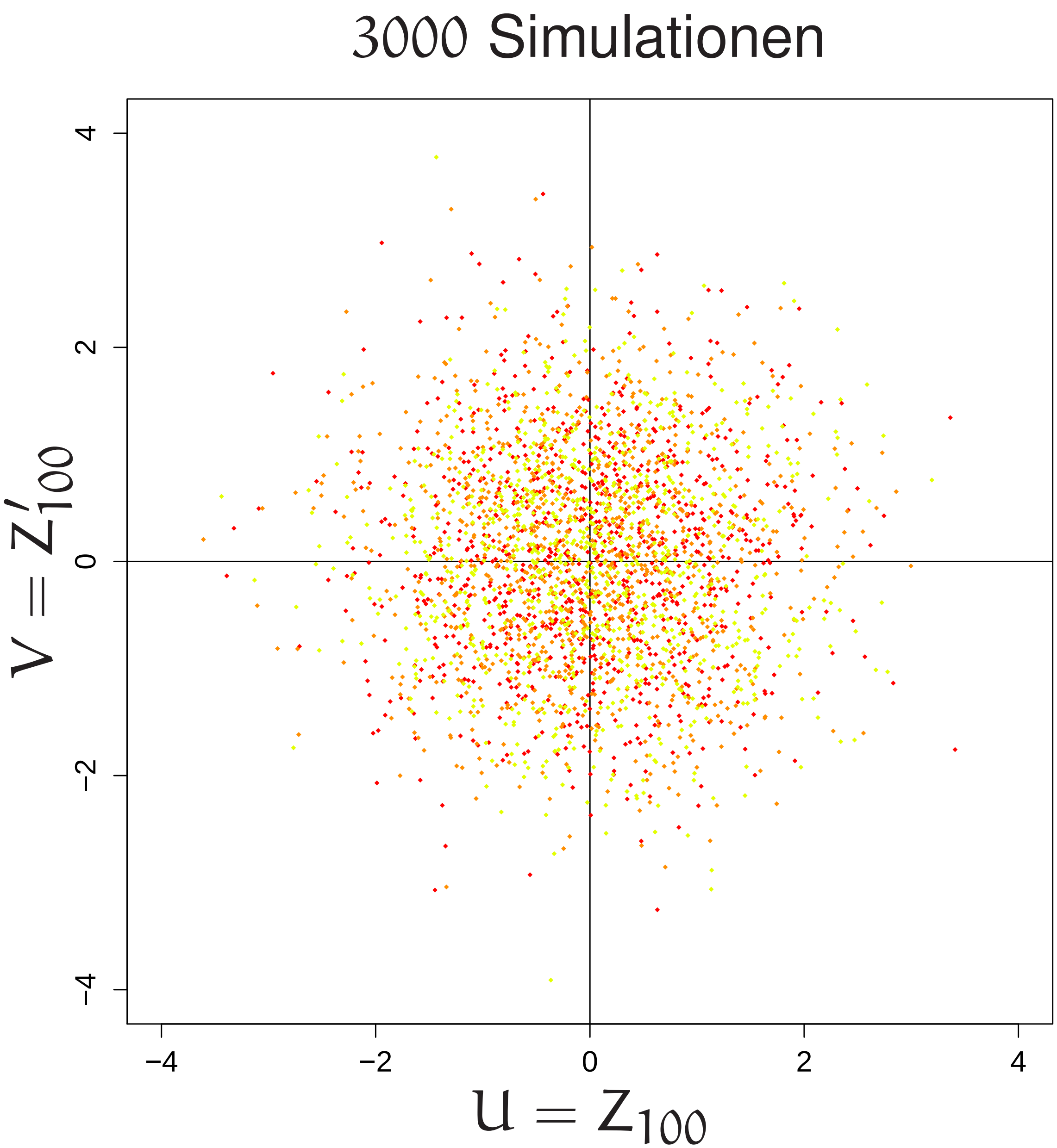
$$(U, V) := (Z_{100}, Z'_{100})$$

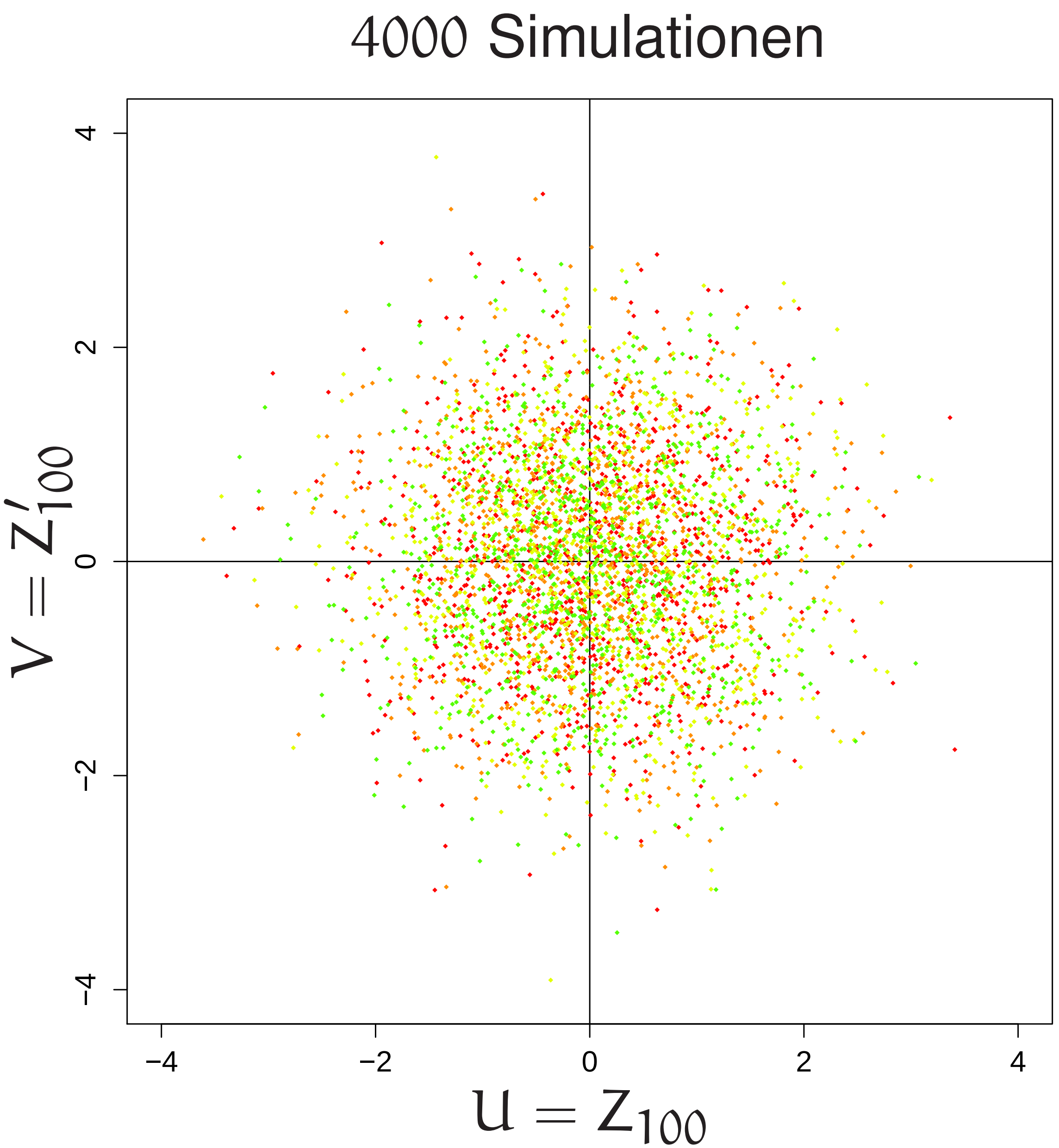
Wie sieht die gemeinsame Verteilung

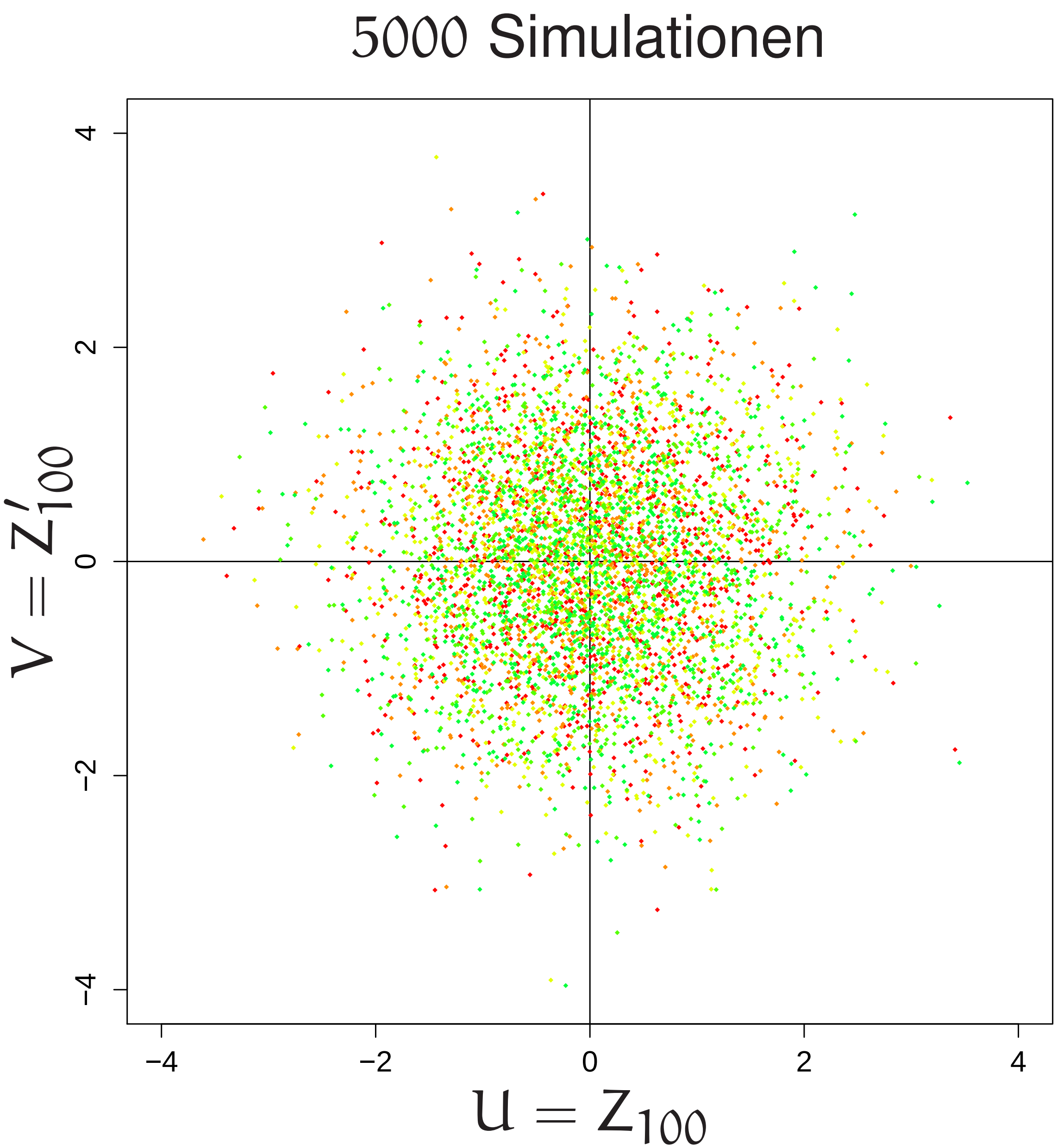
von U und V aus?


1000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

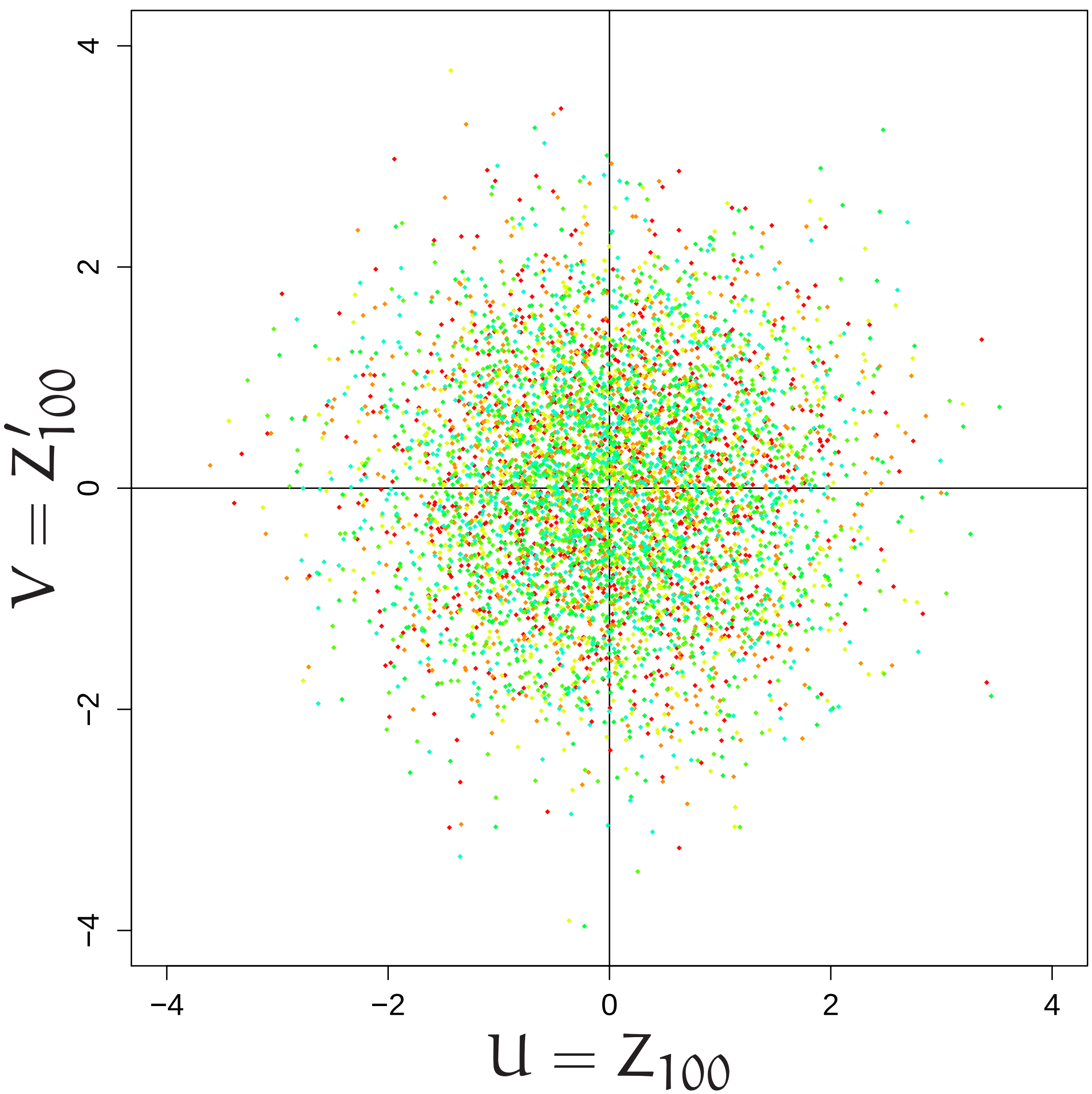

2000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$


3000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

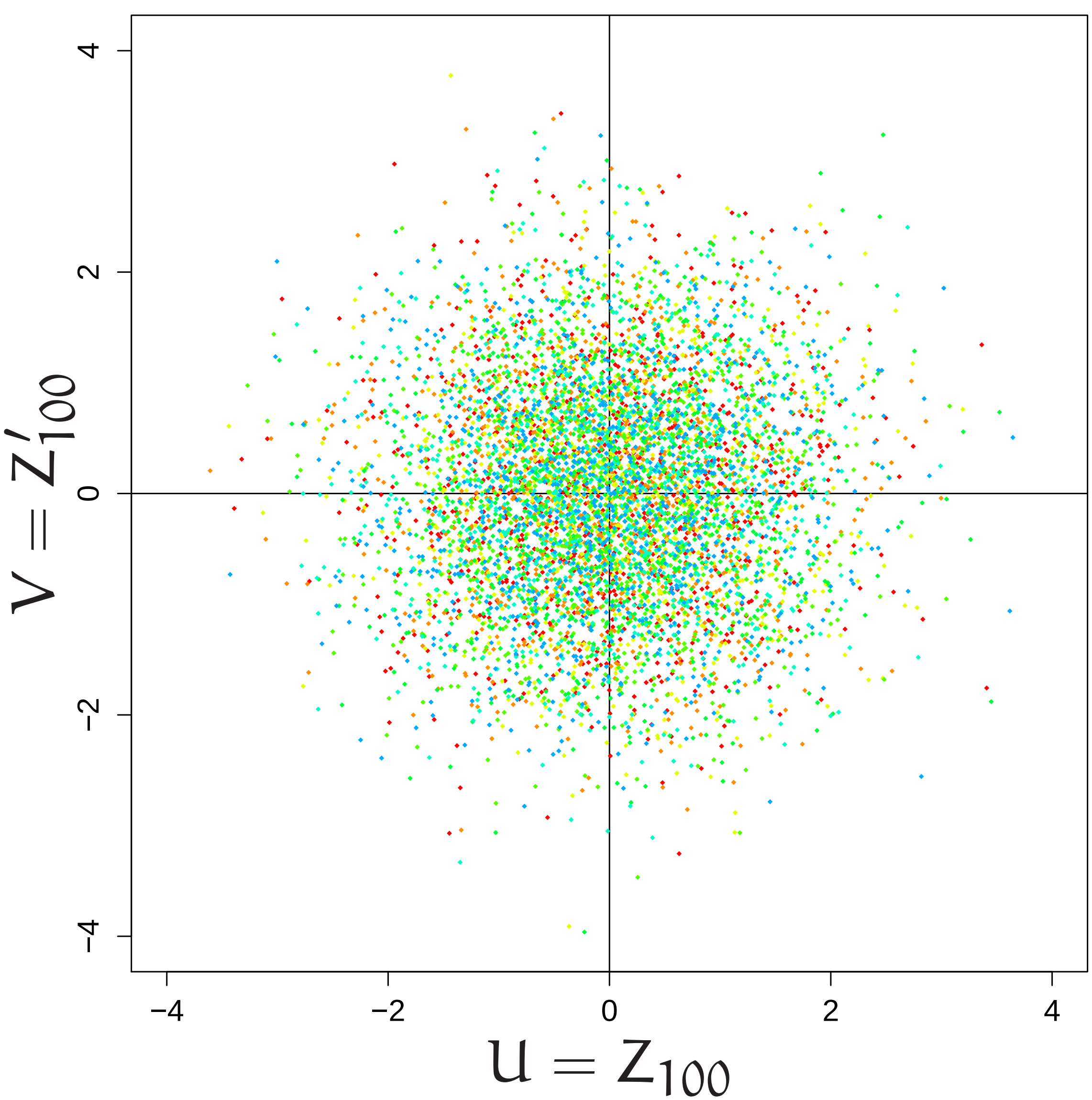

4000 Simulationen
4
2
0
−2
−4
$V = Z'_{100}$
−4
−2
0
2
4
$U = Z_{100}$


5000 Simulationen
$U = Z_{100}$
$V = Z'_{100}$
−4
−2
0
2
4

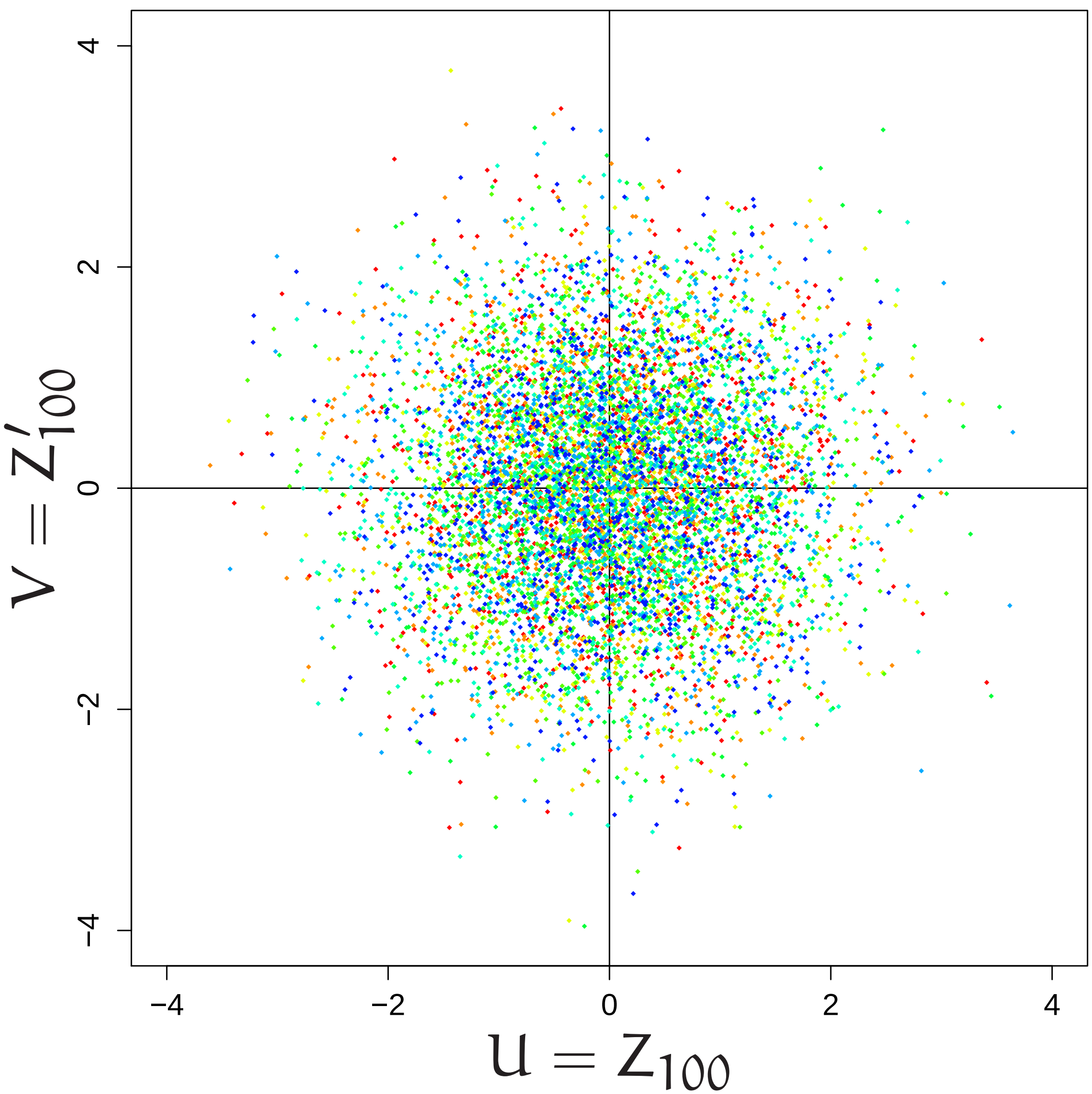
6000 Simulationen

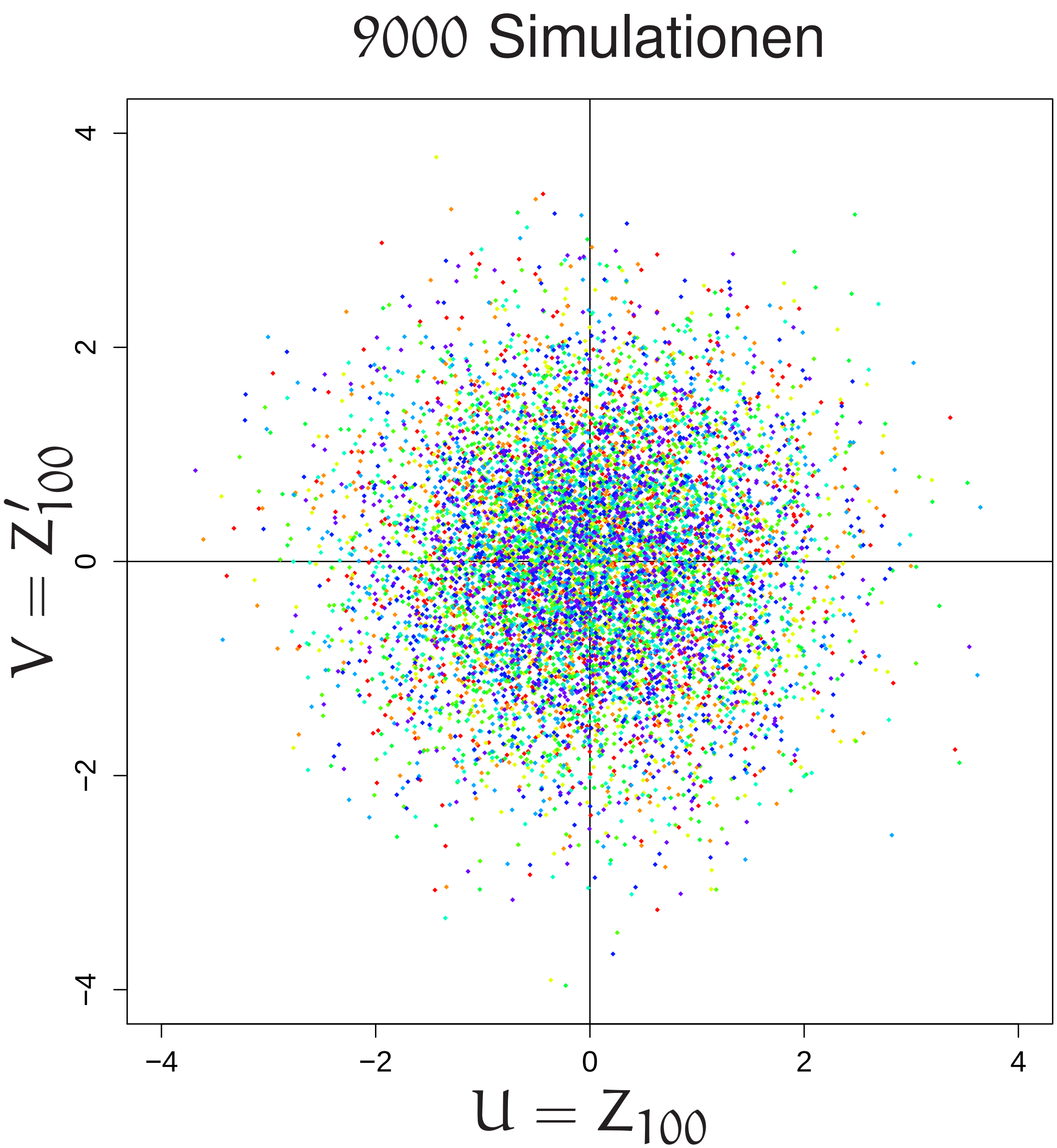


7000 Simulationen

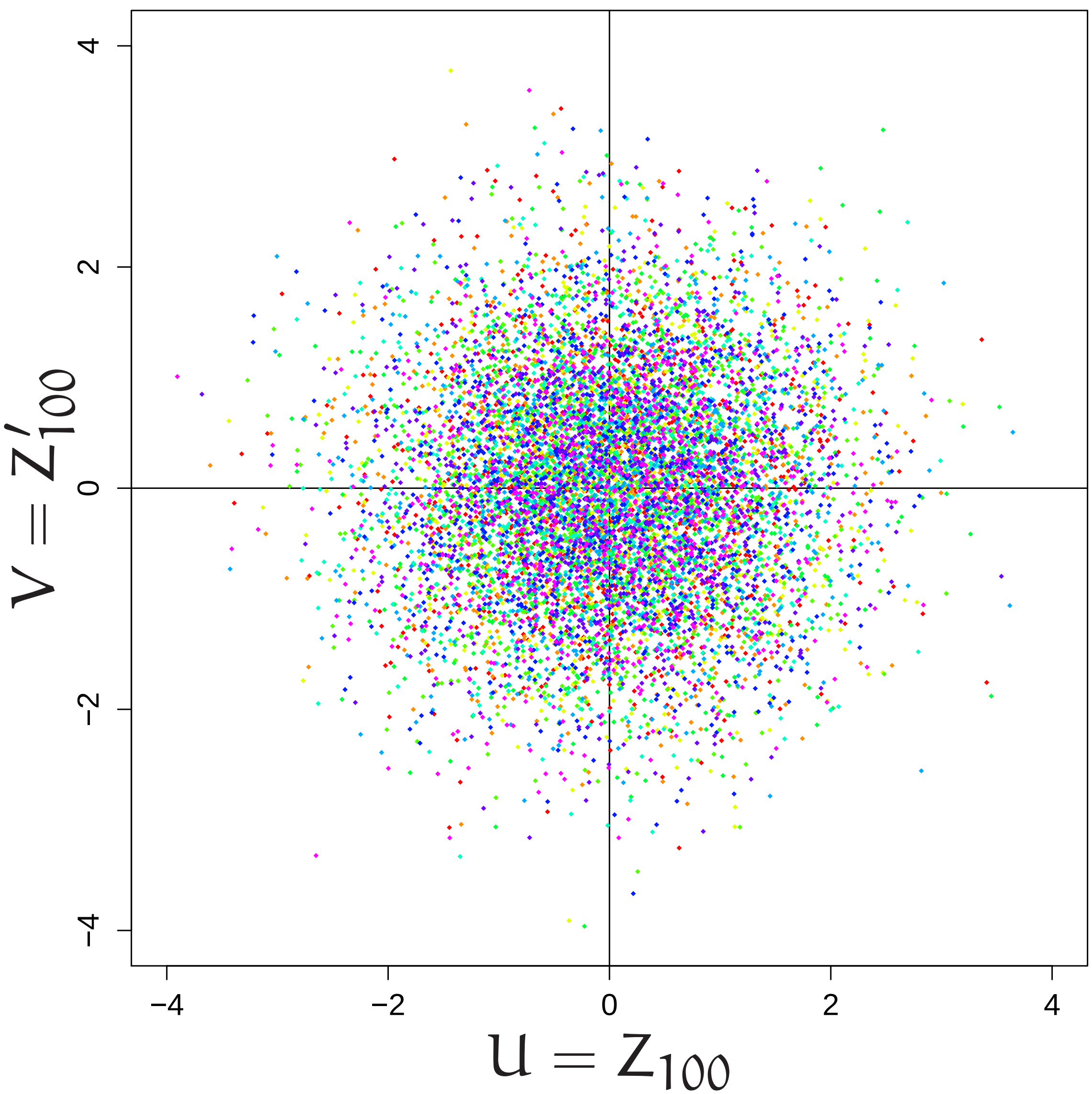


8000 Simulationen

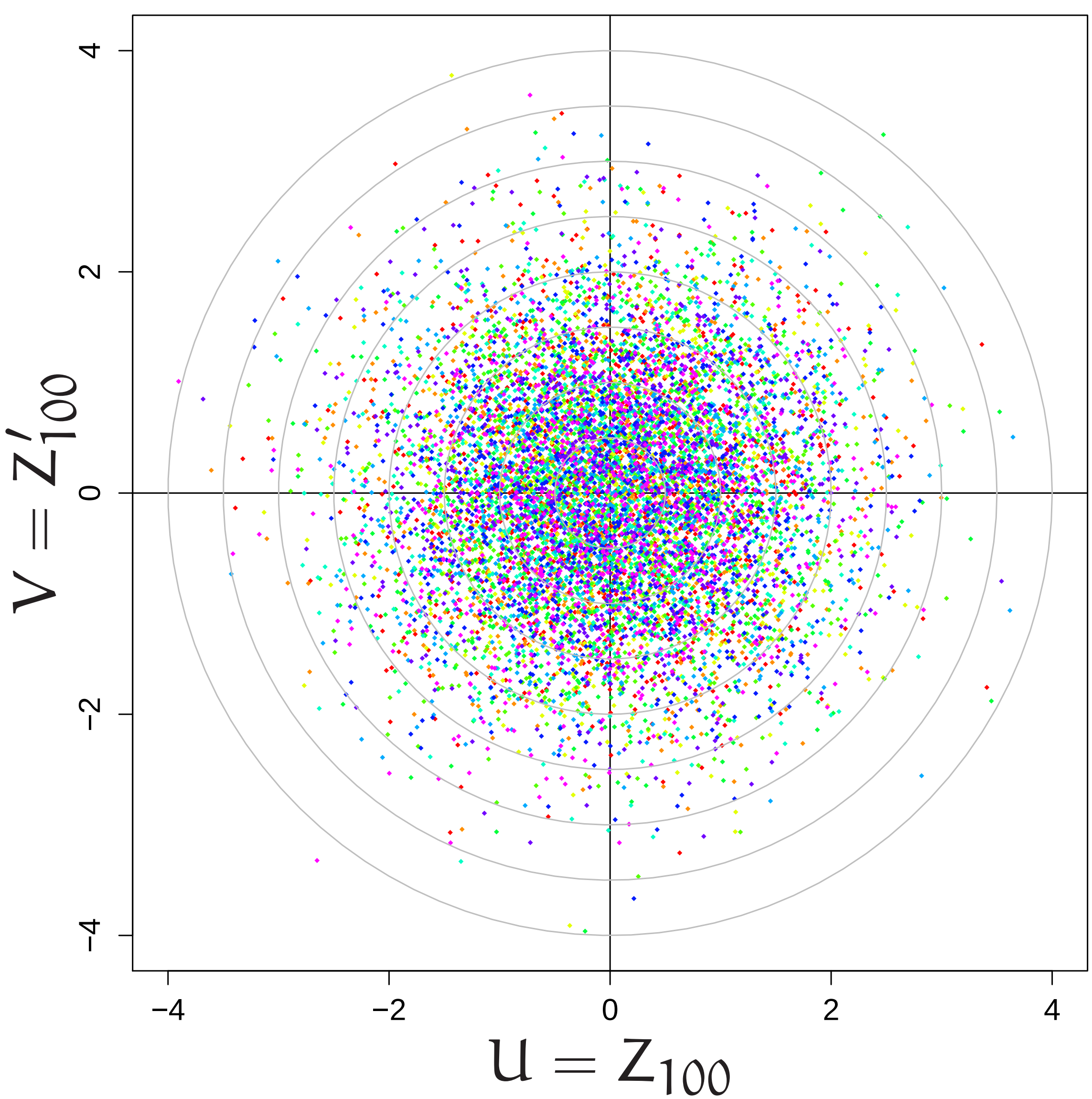



9000 Simulationen
4
2
0
−2
−4
$V = Z'_{100}$
−4
−2
0
2
4
$U = Z_{100}$

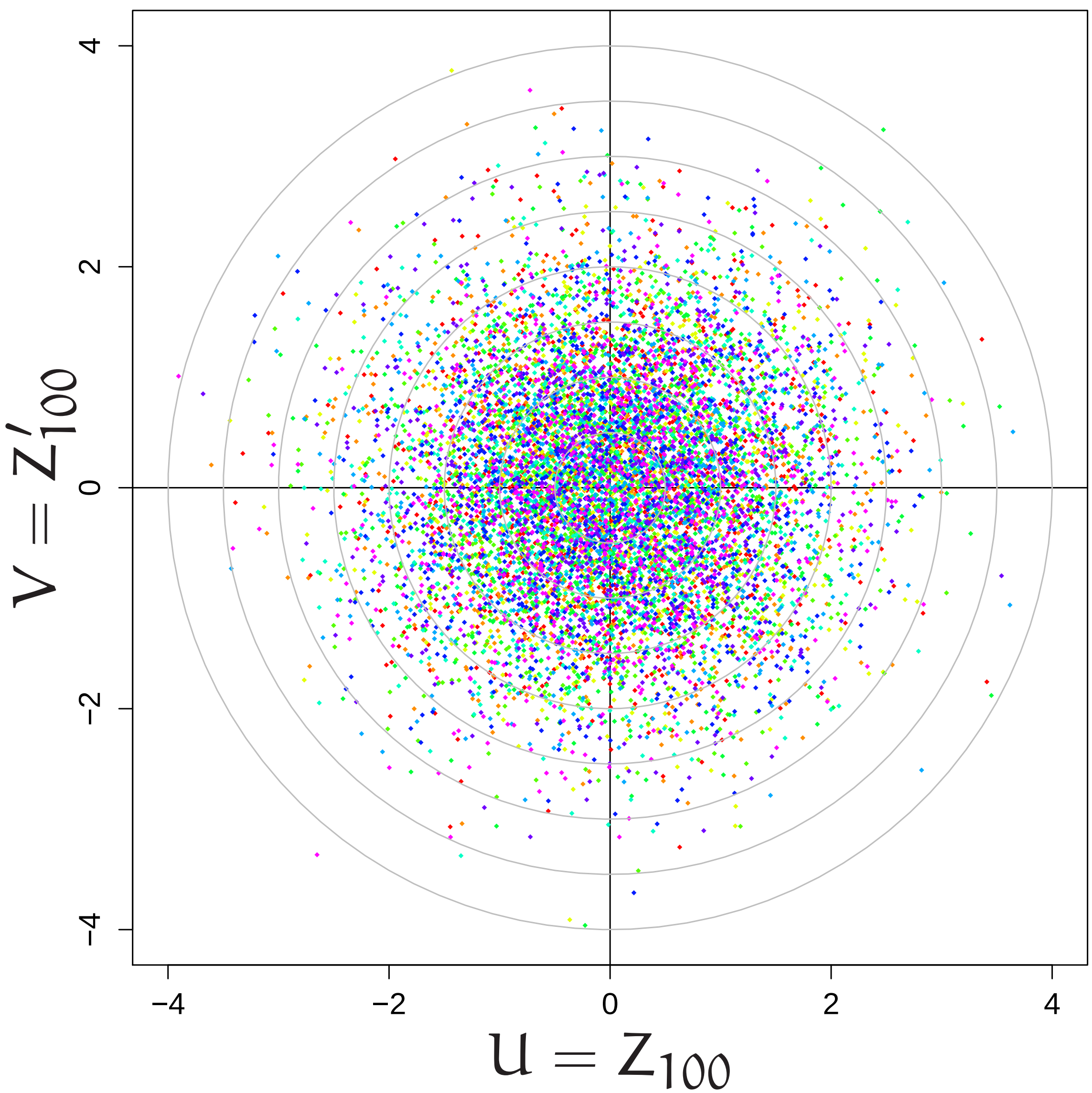
10000 Simulationen



10000 Simulationen



Die Verteilung von (U, V) ist annähernd rotationssymmetrisch!



5. Eine Charakterisierieung der zweidimensionalen Standardnormalverteilung

Behauptung:

Aus “U und V unabhängig und identisch verteilt’

und

“Verteilung von (U, V) rotationssymmetrisch”

folgt,

dass U und V normalverteilt sind:

$$f_U(x) = f_V(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/(2\sigma^2)}$$

Denn:

U, V unabhängig bedeutet:

$$f_{(U,V)}(a, b) = f_U(a) f_V(b)$$

$f_{(U,V)}$ *rotationssymmetrisch* heißt: es existiert ein g mit

$$f_{(U,V)}(a, b) = g(r), \qquad r := \sqrt{a^2 + b^2}.$$

Mit $f_U = f_V =: h$ folgt

$$h(a)h(b) = g(r)$$

$$h(a)\,h(b) = g(r), \qquad r := \sqrt{a^2 + b^2}$$

Die zwei Paare (a, b) und $(0, \sqrt{a^2 + b^2})$ haben dasselbe r.

Also:

$$h(a)\,h(b) = h(0)\,h(\sqrt{a^2 + b^2})$$

Das ist eine Gleichung für h. Eine Lösung hiervon ist:

$$h(x) = e^{-x^2}$$

Denn

$$e^{-a^2} e^{-b^2} = 1 \cdot e^{-(a^2+b^2)}$$

$$h(a)\,h(b) = h(0)\,h(\sqrt{a^2+b^2})$$

Wie sieht die allgemeine Lösung aus?

$w(u) := h(\sqrt{u})$, $u \geq 0$, erfüllt

$$w(a^2)w(b^2) = w(0)w(a^2+b^2), \quad a, b \in \mathbb{R}.$$

$$w(u)\,w(v) \;=\; k_0\,w(u+v), \qquad u, v \geq 0$$

hat als allgemeine Lösung

$$w(u) = k_0 e^{-k_1 u},$$

Daraus folgt:

$$h(a) = w(a^2) = k_0\,e^{-k_1 a^2}.$$

FAZIT

Der Zentrale Grenzwertsatz

lässt sich erraten

(in konkreten Fällen,

mit etwas Glück).

Hier ist noch einmal die (im ZGS präzisierte) Botschaft der Stunde:

Summen (und Mittelwerte) von vielen unabhängigen,
identisch verteilten ZV mit endlicher Varianz
sind annähernd normalverteilt.

Diese Aussage bleibt übrigens auch
unter schwächeren Bedingungen bestehen,
sowohl was die Unabhängigkeit,
als auch was die identische Verteiltheit betrifft.

Eine Botschaft zum Mitnehmen ins Leben

(salopp formuliert):

“Die Summe von vielen

annähernd unabhängigen Zufallsvariablen,

die nicht notwendig identisch verteilt, aber

ungefähr von derselben Größenordnung sind,

ist annähernd normalverteilt.”

Dazu sei (als Ausblick) angemerkt der

Satz von Lindeberg-Feller:

Für jedes n seien $X_{n,m}$, $1 \le m \le n$, unabhängige ZV'e mit

$$\mathbf{E}X_{n,m} = 0 \quad \text{und} \quad \sum_{m=1}^{n} \mathbf{E}X_{n,m}^2 \stackrel{n\to\infty}{\to} \sigma^2 > 0.$$

Für alle $\varepsilon > 0$ gelte $\displaystyle\sum_{m=1}^{n} \mathbf{E}[X_{n,m}^2 I_{\{|X_{n,m}|>\varepsilon\}}] \stackrel{n\to\infty}{\to} 0.$

Dann konvergiert $X_{n,1} + \cdots + X_{n,n}$ für $n \to \infty$

in Verteilung gegen eine $N(0, \sigma^2)$-verteilte Zufallsvariable.

6. Münzwurf und Zentraler Grenzwertsatz

Der Münzwurf passt in den Zentralen Grenzwertsatz:

Sei $X_1, X_2, \ldots,$ ein fortgesetzer p-Münzwurf. Dann ergibt sich aus dem Zentralen Grenzwertsatz der (alte) Satz von de Moivre und Laplace:

Für Binomial-(n, p)-verteilte Zufallsvariable B_n (mit festem p) gilt für alle $c < d \in \mathbb{R}$:

$$\mathbf{P}\left(\frac{B_n - np}{\sqrt{npq}} \in [c, d]\right) \underset{n\to\infty}{\longrightarrow} \mathbf{P}(Z \in [c, d]).$$

Dabei ist Z standard-normalverteilt.