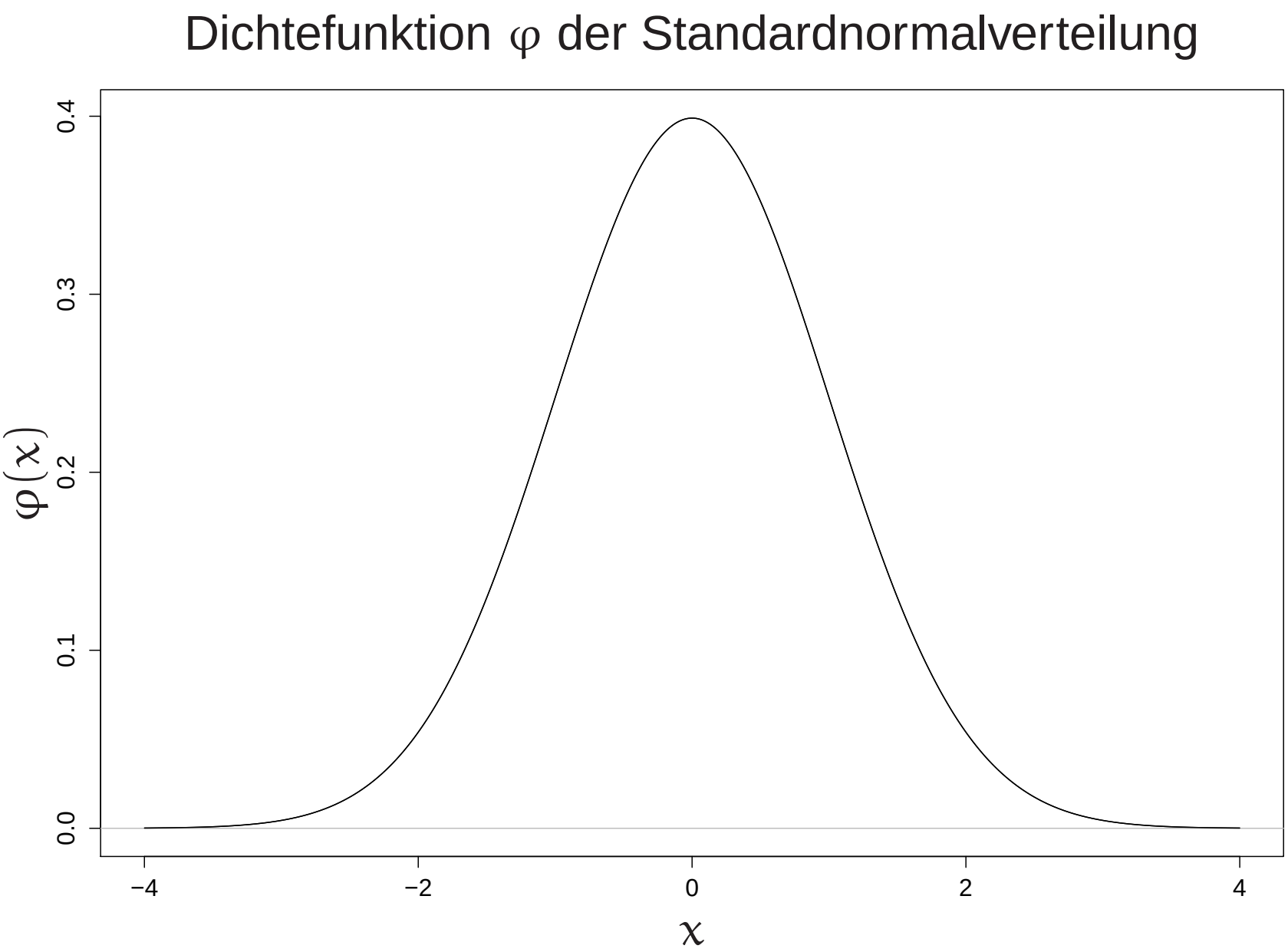
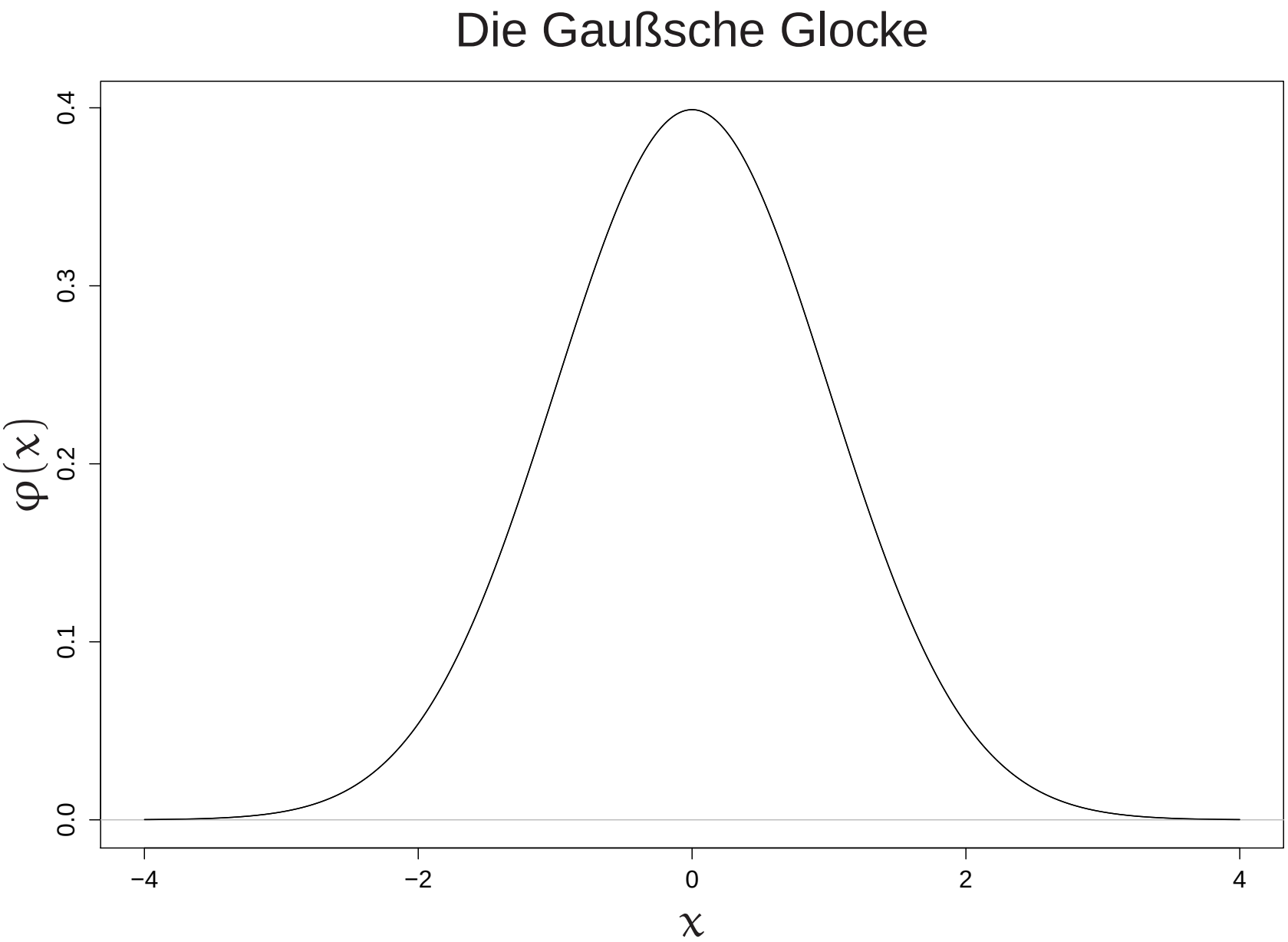
Vorlesung 6b

Der Zentrale Grenzwertsatz

Wiederholung:

Die Normalverteilung


Dichtefunktion φ der Standardnormalverteilung
φ(x)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


Die Gaußsche Glocke
φ(x)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

Die Standard-Normalverteilung

Dichtefunktion:

$$\varphi(x) := \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

Verteilungsfunktion:

$$\Phi(a) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$$

Z standard-normalverteilt:

$$\mathbf{P}(\ Z \leq a\ ) = \Phi(a)$$

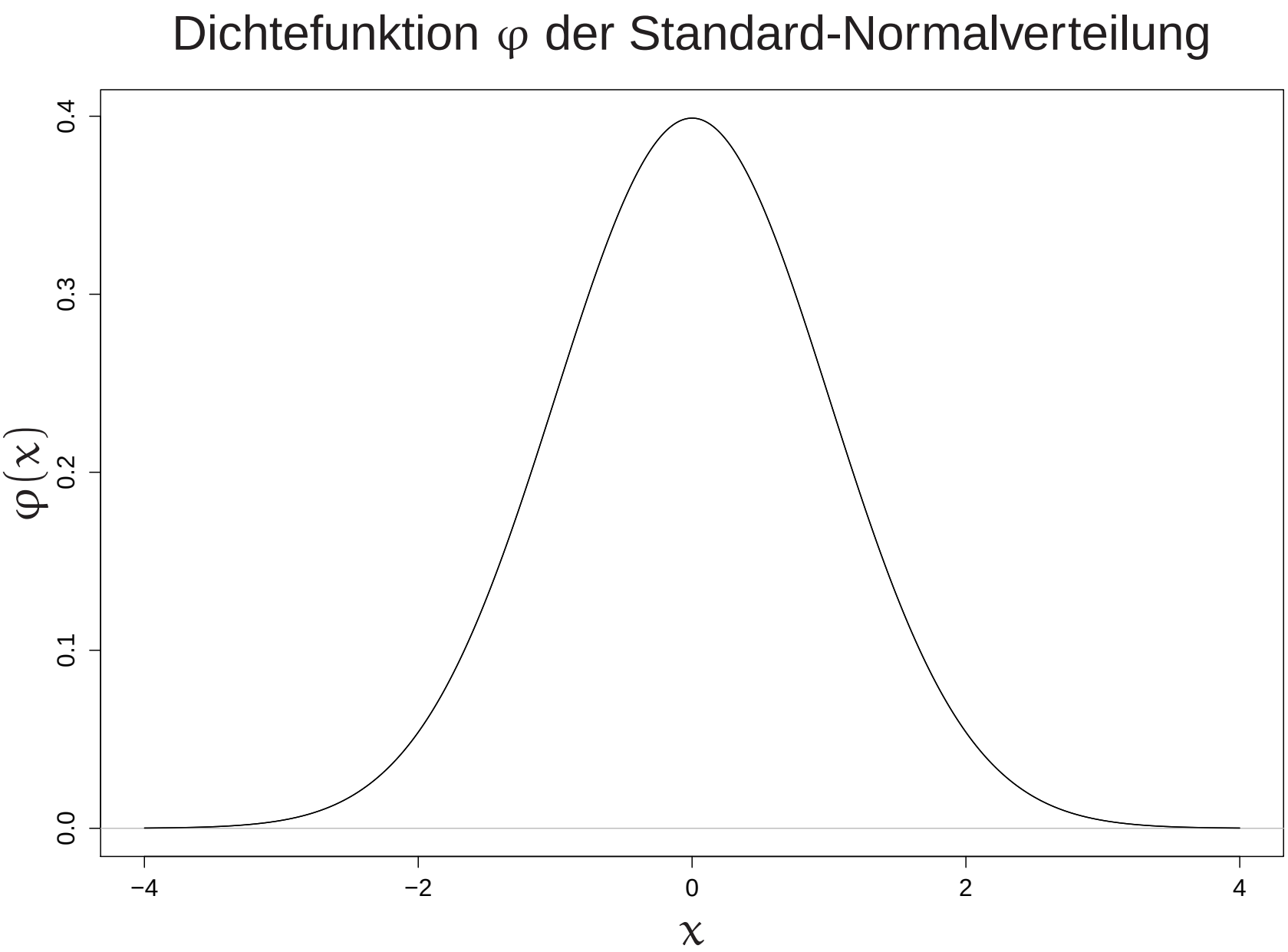
Die Standard-Normalverteilung

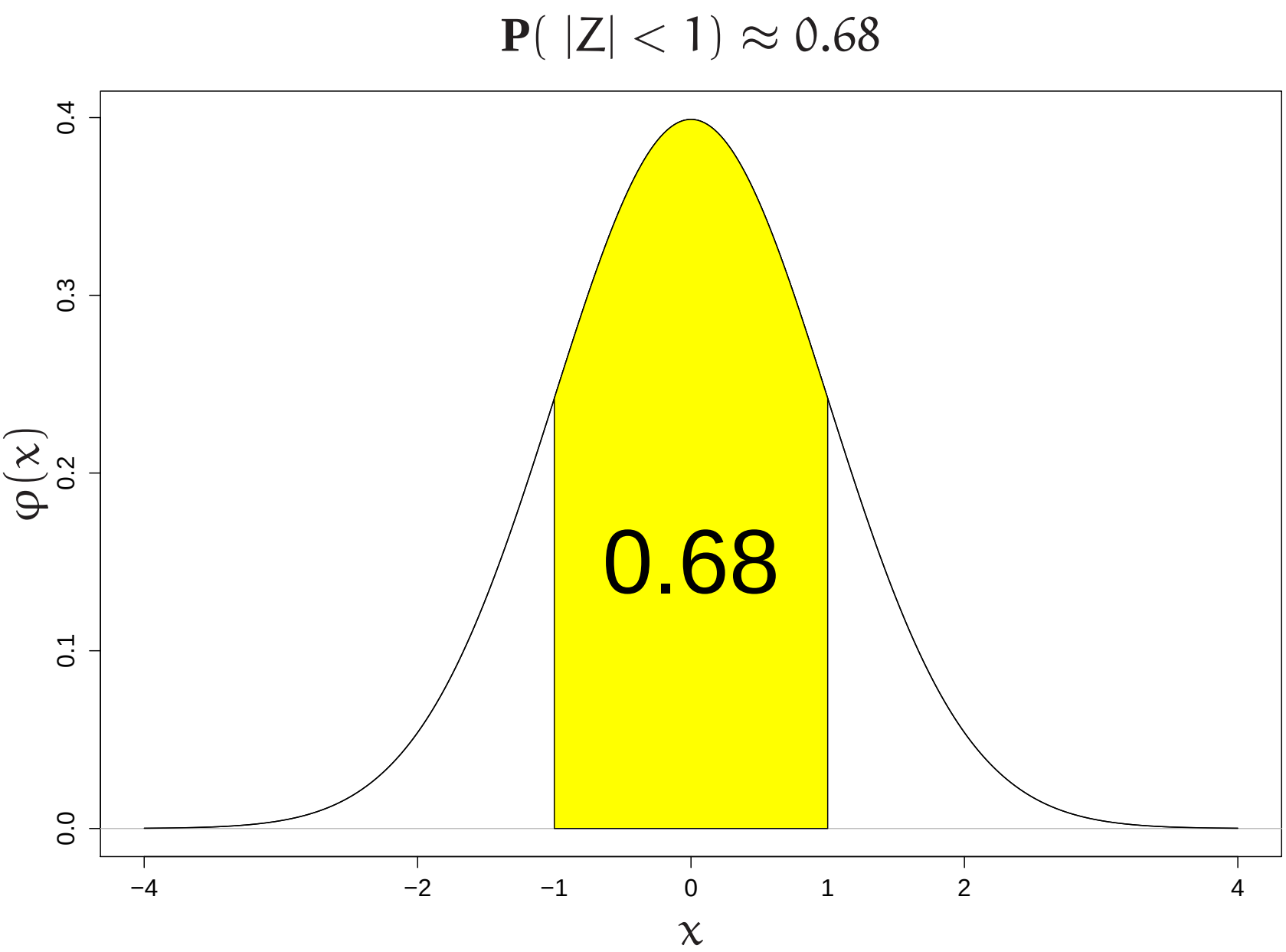
$$\mathbf{E}Z = 0 \qquad \sigma_Z = 1$$

Die allgemeine Normalverteilung

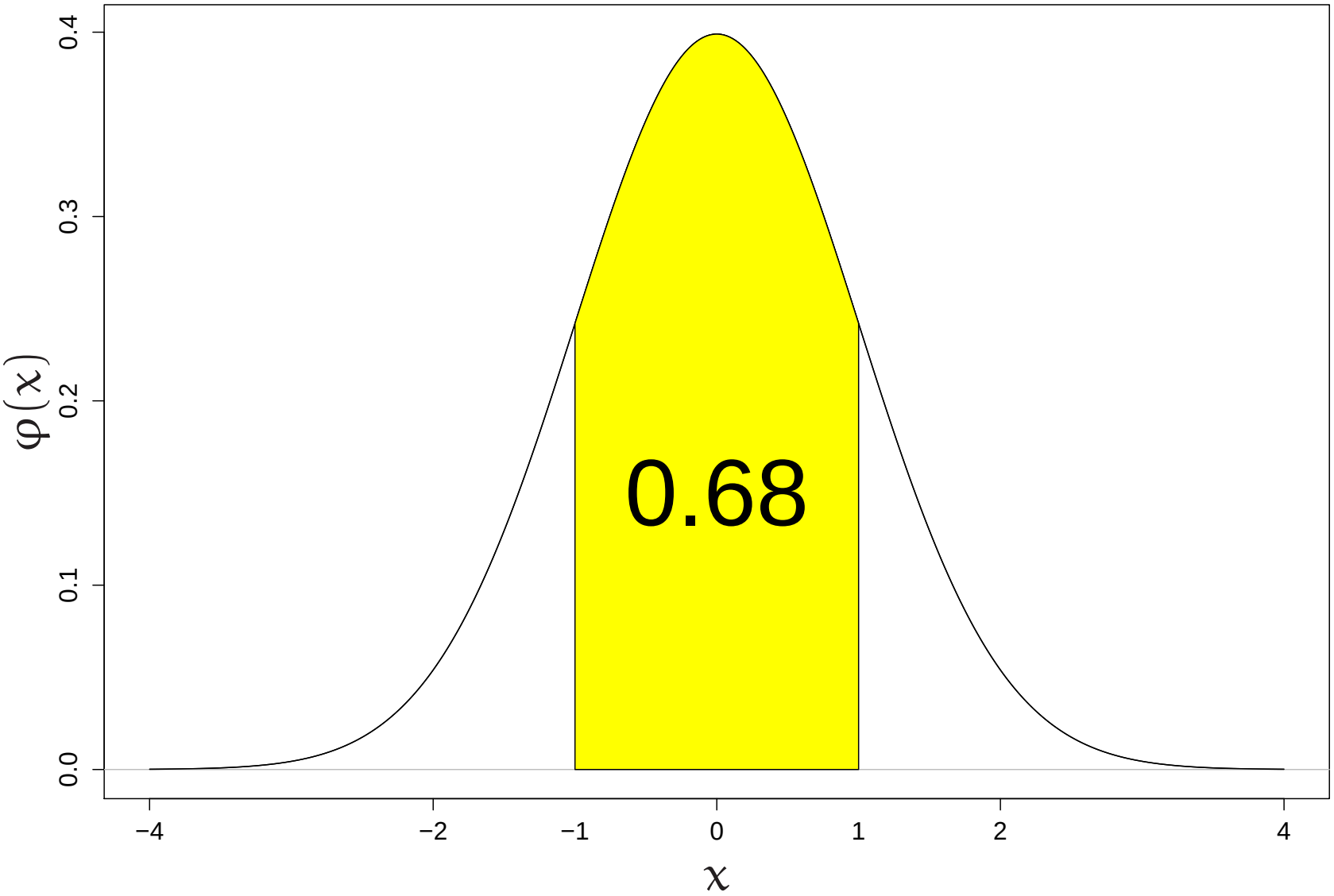
$$N = \mu + \tau Z$$

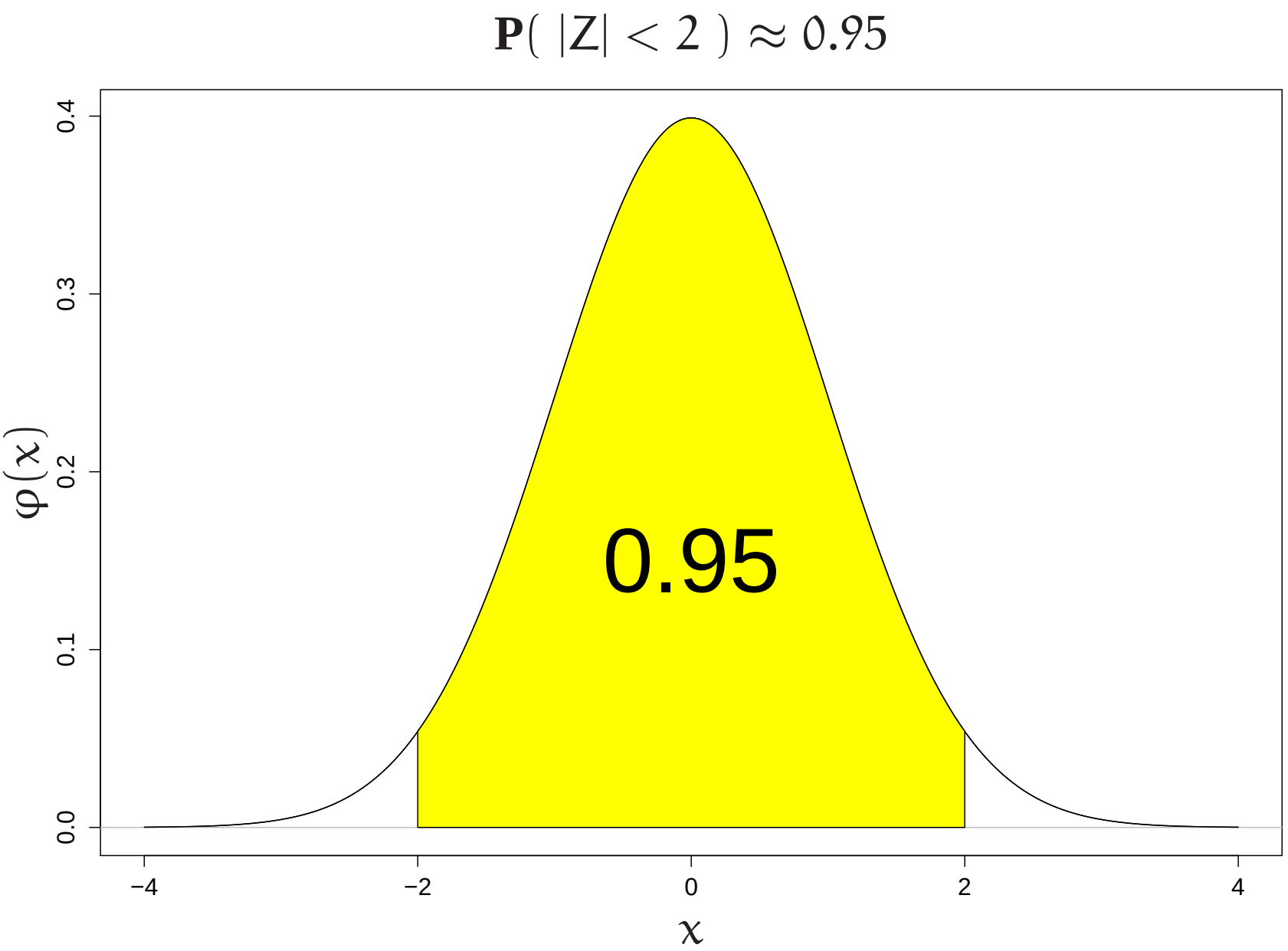
$$\mathbf{E}N = \mu \qquad \sigma_N = \tau$$


Dichtefunktion φ der Standard-Normalverteilung
φ(x)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

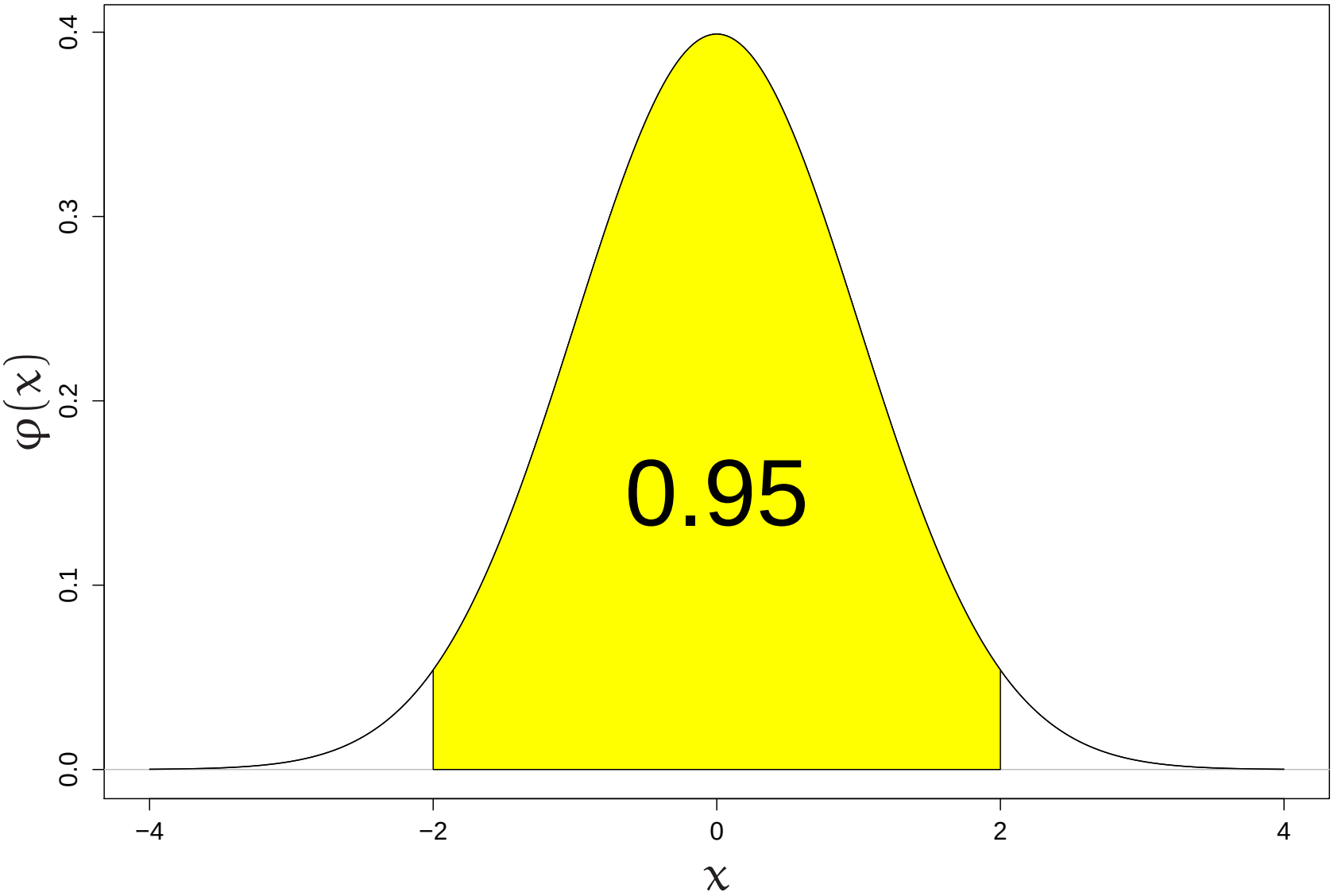

$\mathbf{P}(\ |Z| < 1) \approx 0.68$
0.68
$\varphi(x)$
0.0
0.1
0.2
0.3
0.4
−4
−2
−1
0
1
2
4
x

$$\Phi(1) - \Phi(-1) \approx 0.68$$

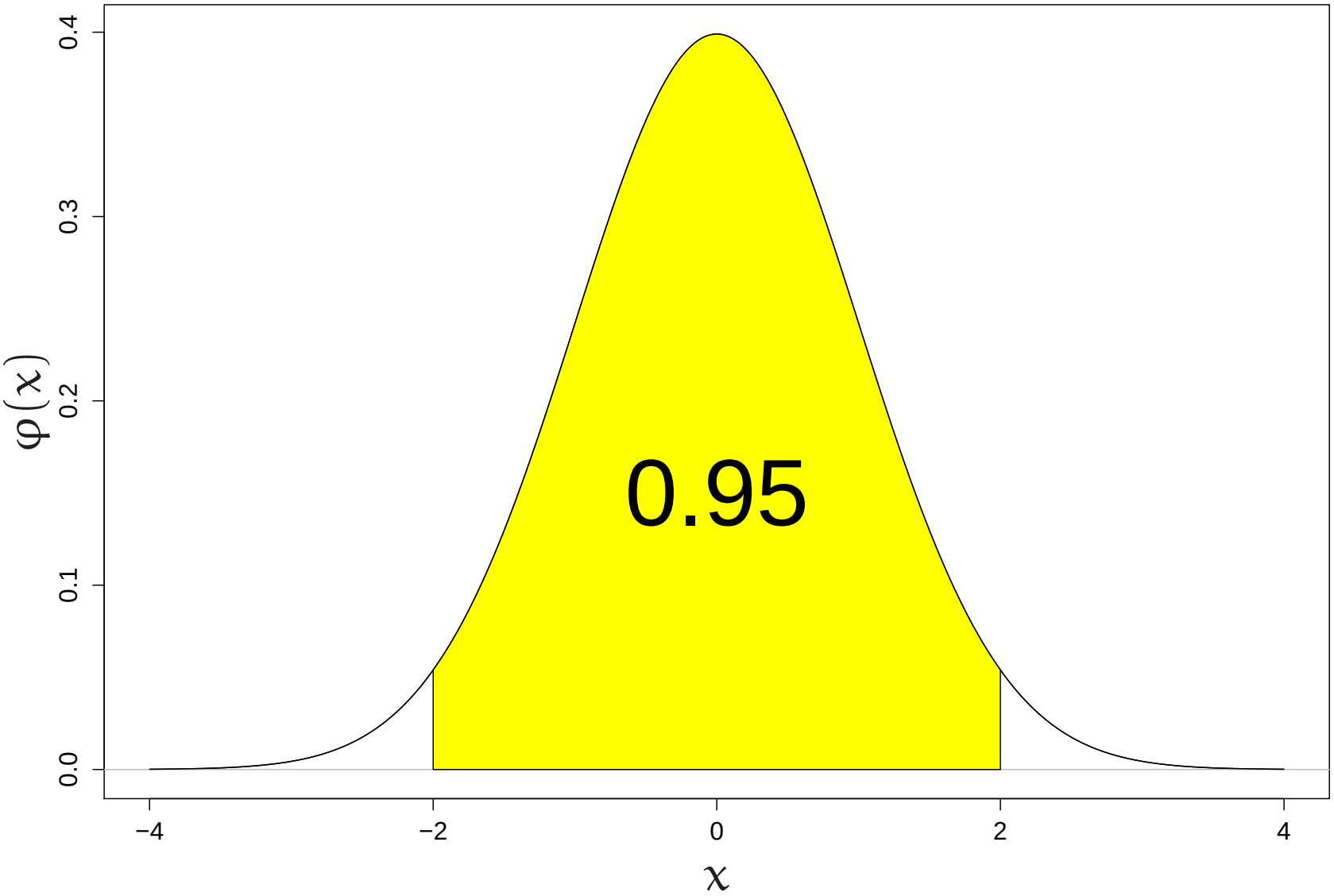


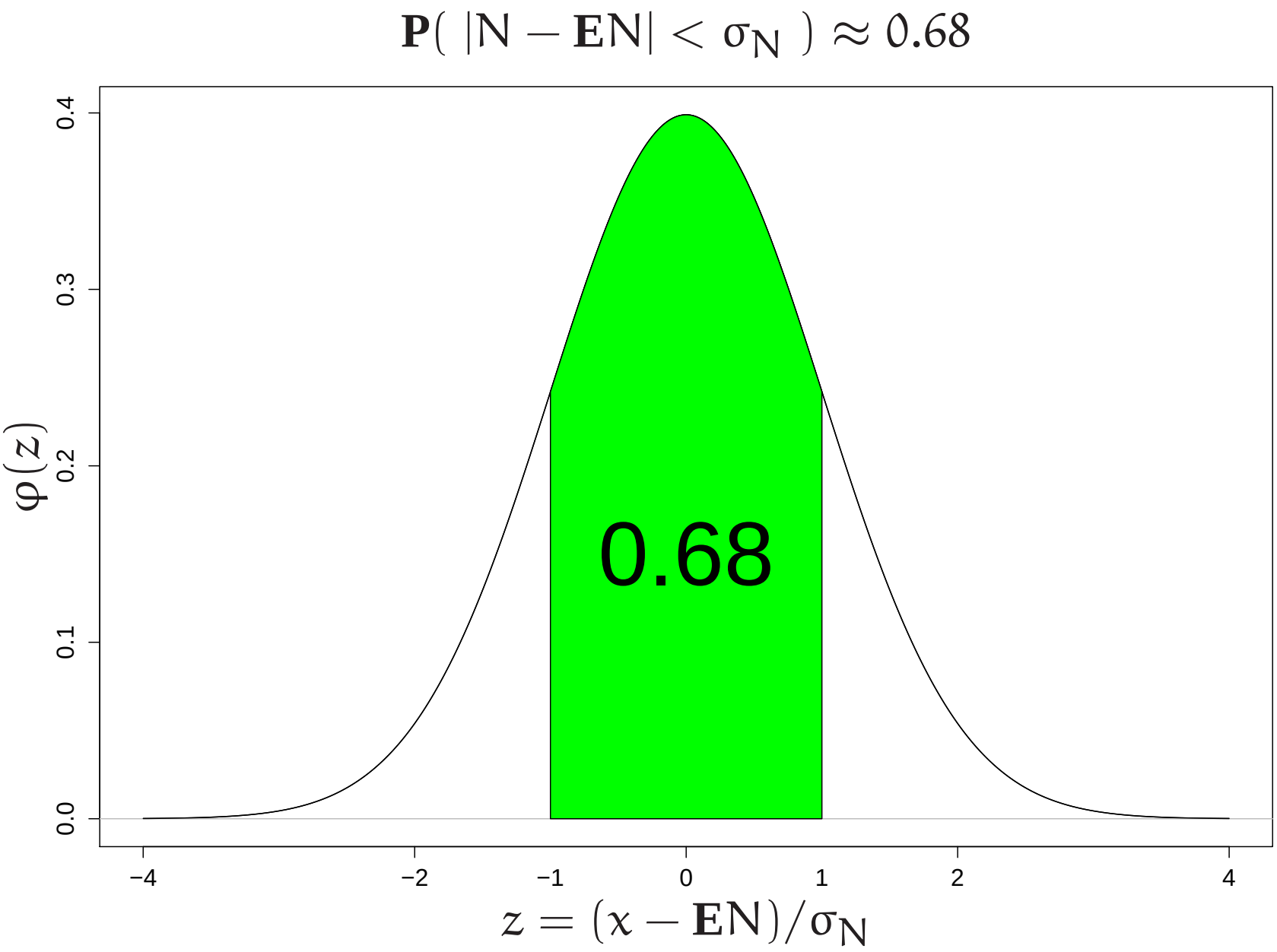

$\mathbf{P}(\ |Z| < 2\ ) \approx 0.95$
0.95
$\varphi(x)$
x
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4

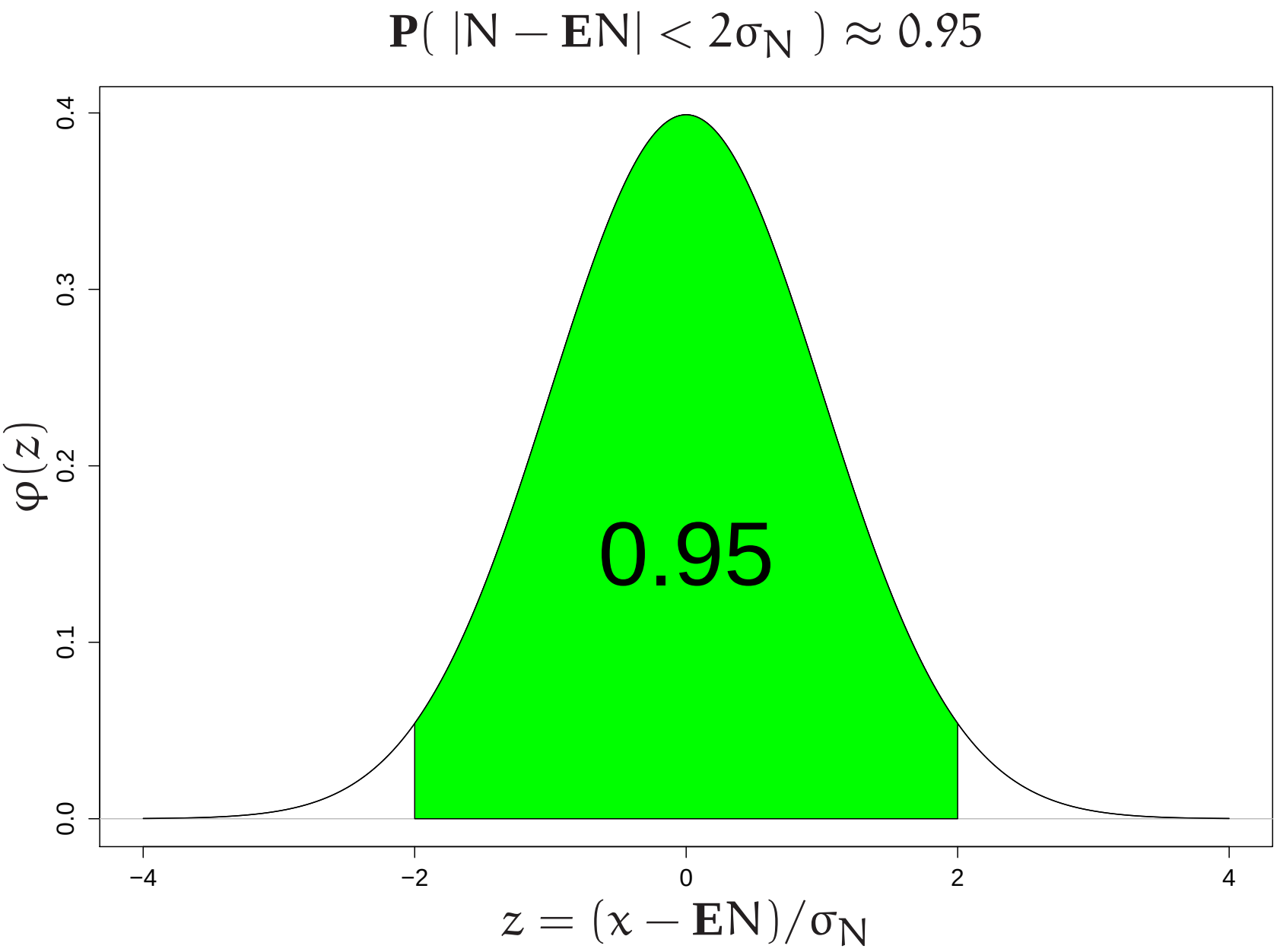
Und für allgemeine normalverteilte Zufallsgrößen N?



Dasselbe in grün.




$\mathbf{P}(\ |\mathrm{N} - \mathbf{E}\mathrm{N}| < \sigma_\mathrm{N}\ ) \approx 0.68$
0.68
$\varphi(z)$
0.0
0.1
0.2
0.3
0.4
−4
−2
−1
0
1
2
4
$z = (\mathrm{x} - \mathbf{E}\mathrm{N})/\sigma_\mathrm{N}$


$\mathbf{P}(\ |\mathrm{N} - \mathbf{E}\mathrm{N}| < 2\sigma_{\mathrm{N}}\ ) \approx 0.95$
0.95
$\varphi(z)$
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
$z = (\mathrm{x} - \mathbf{E}\mathrm{N})/\sigma_{\mathrm{N}}$

Der
Zentrale
Grenzwertsatz

Die Summe

von vielen kleinen,

unabhängigen Summanden

ist annähernd

normalverteilt.

DER ZENTRALE GRENZWERTSATZ

Seien $X_1, X_2, ...$

unabhängige, identisch verteilte Zufallsgrößen.

Sei

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

und

$$Z_n := (S_n - \mathbf{E}S_n) \,/\, \sigma_{S_n}.$$

$$(\, \mathbf{E}Z_n = 0, \quad \sigma_{Z_n} = 1. \,)$$

Dann gilt:

Z_n

ist

asymptotisch

standardnormalverteilt.

Das heißt:

$\mathbf{P}(\ Z_n \leq a\ ) \to \Phi(a)$.

Zentraler Grenzwertsatz:

Die standardisierte Summe von n unabhängigen,

identisch verteilten $\mathbb{R}$-wertigen Zufallsvariablen

mit endlicher Varianz

konvergiert mit wachsendem n

in Verteilung

gegen eine standard-normalverteilte Zufallsvariable.

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 $\ell < r \in \mathbb{R}$

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

Dabei ist Z standard-normalverteilt.

Geschichte des Zentralen Grenzwertsatzes

Abraham de Moivre

Der faire Münzwurf (1733)

Pierre-Simon Laplace

Allgemeine binomiale Zufallsgrößen (1812)

Pafnuty Lvovich Chebyshev

$$\mathbf{P}(\ |X - \mathbf{E}X| > k\sigma\ ) \leq \frac{1}{k^2} \quad (1846)$$

Andrei Andreyevich Markov

Allgemeiner zentraler Grenzwertsatz

Aleksandr Mikhailovich Lyapunov

Noch allgemeiner (1906)

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

BEISPIEL

Rundungsfehler bei Addition

In Wirklichkeit

$$\pi =$$

$$3.14159265358979323846264338327950288419716939937510\ldots$$

Im Rechner

$$\pi \leftarrow 3.14159265358979$$

MODELL

Zahl = Rechnerdarstellung + Rundungsfehler.

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

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?

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

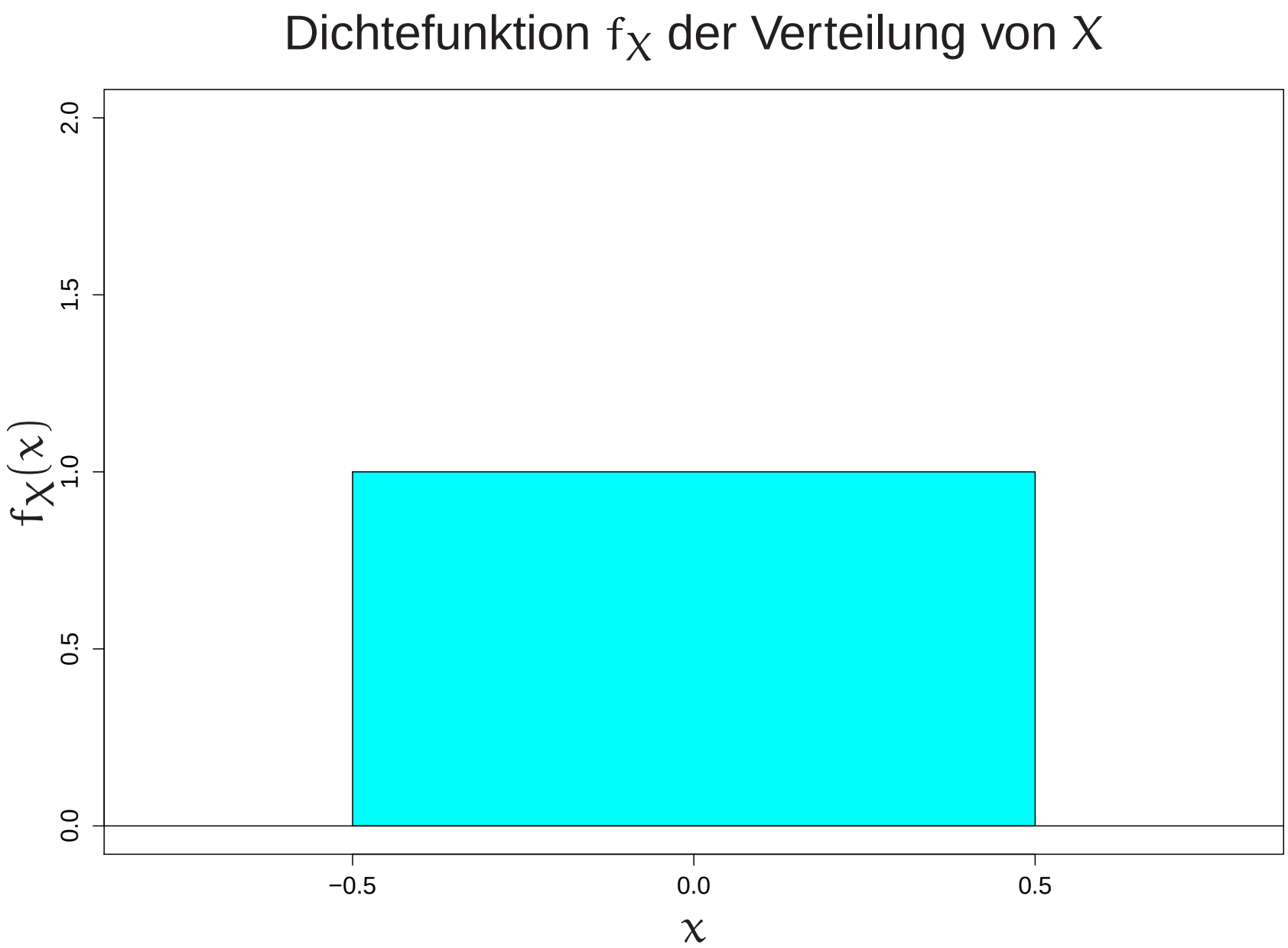
Empirische Verteilung von

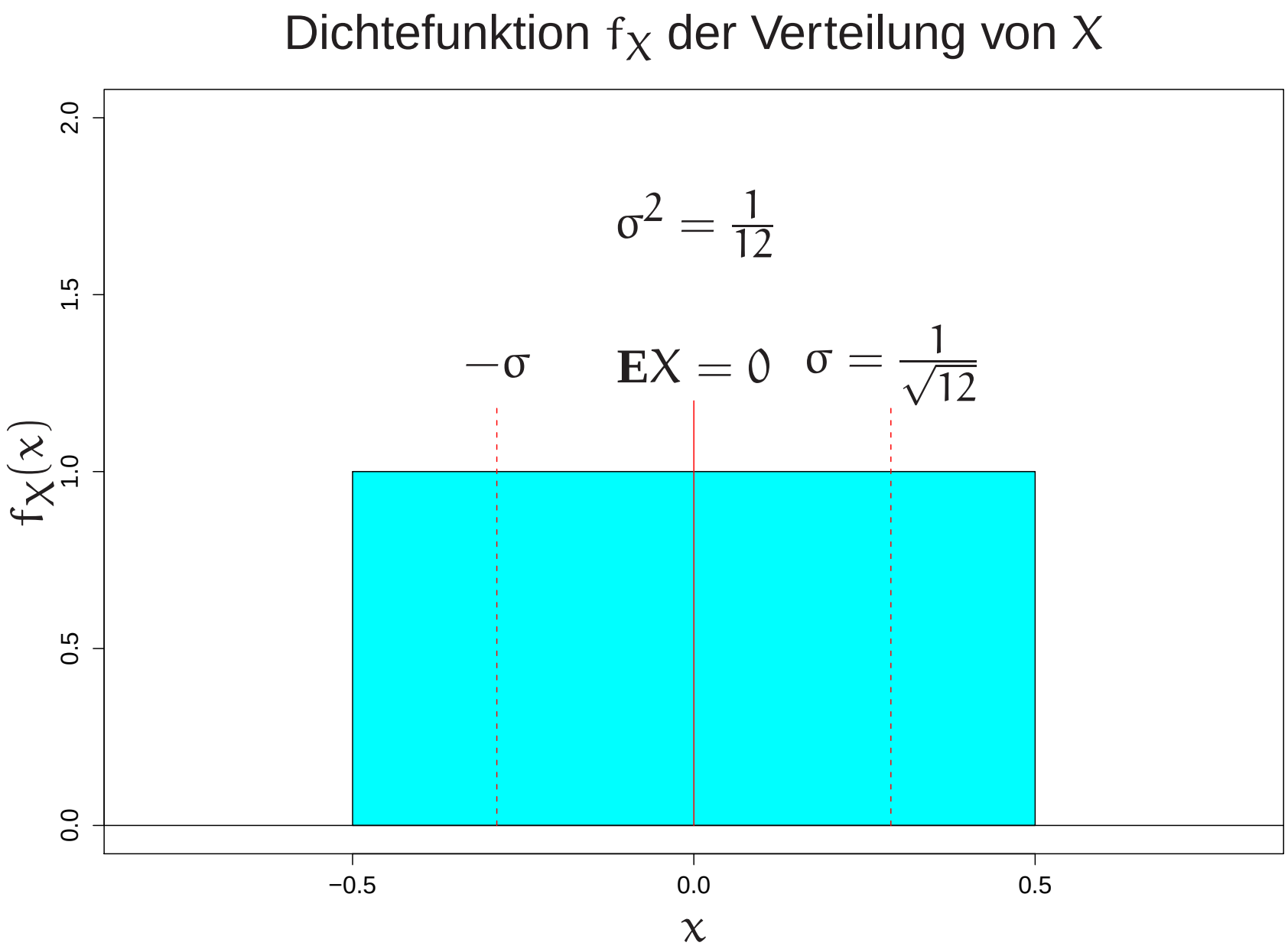
$$S_n := X_1 + ... + X_n$$

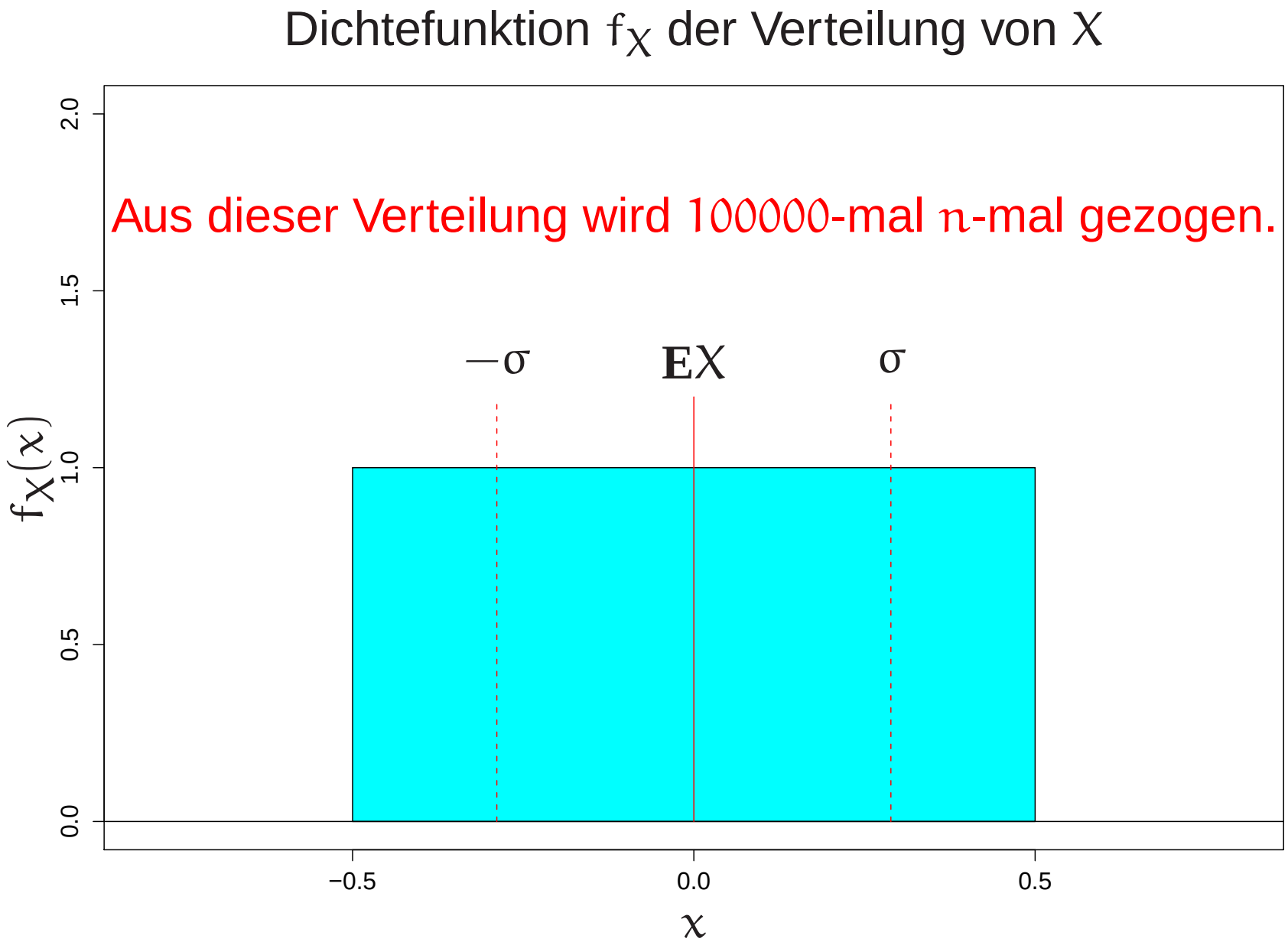
100000 Simulationen

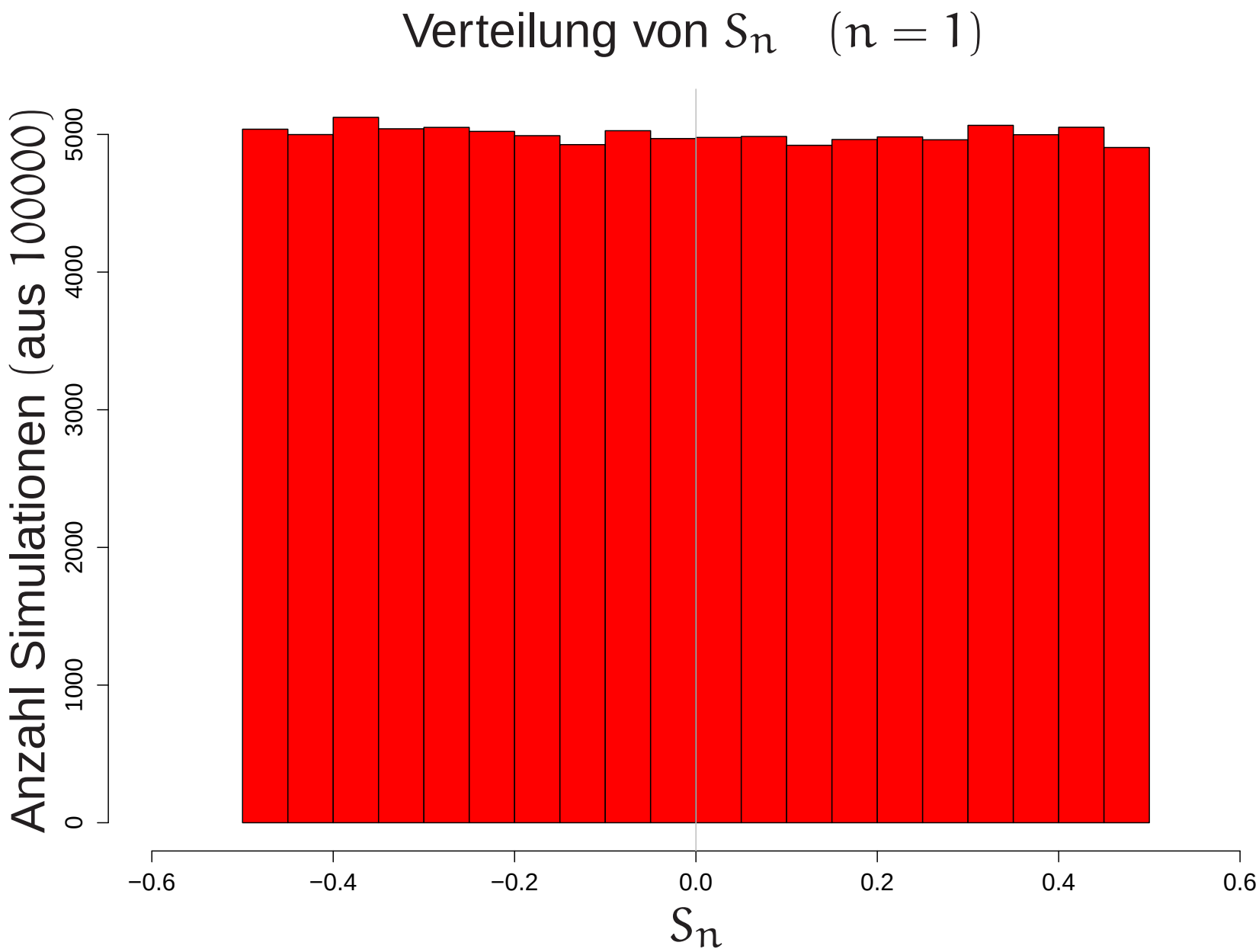
$$n = 1, 2, ..., 10$$

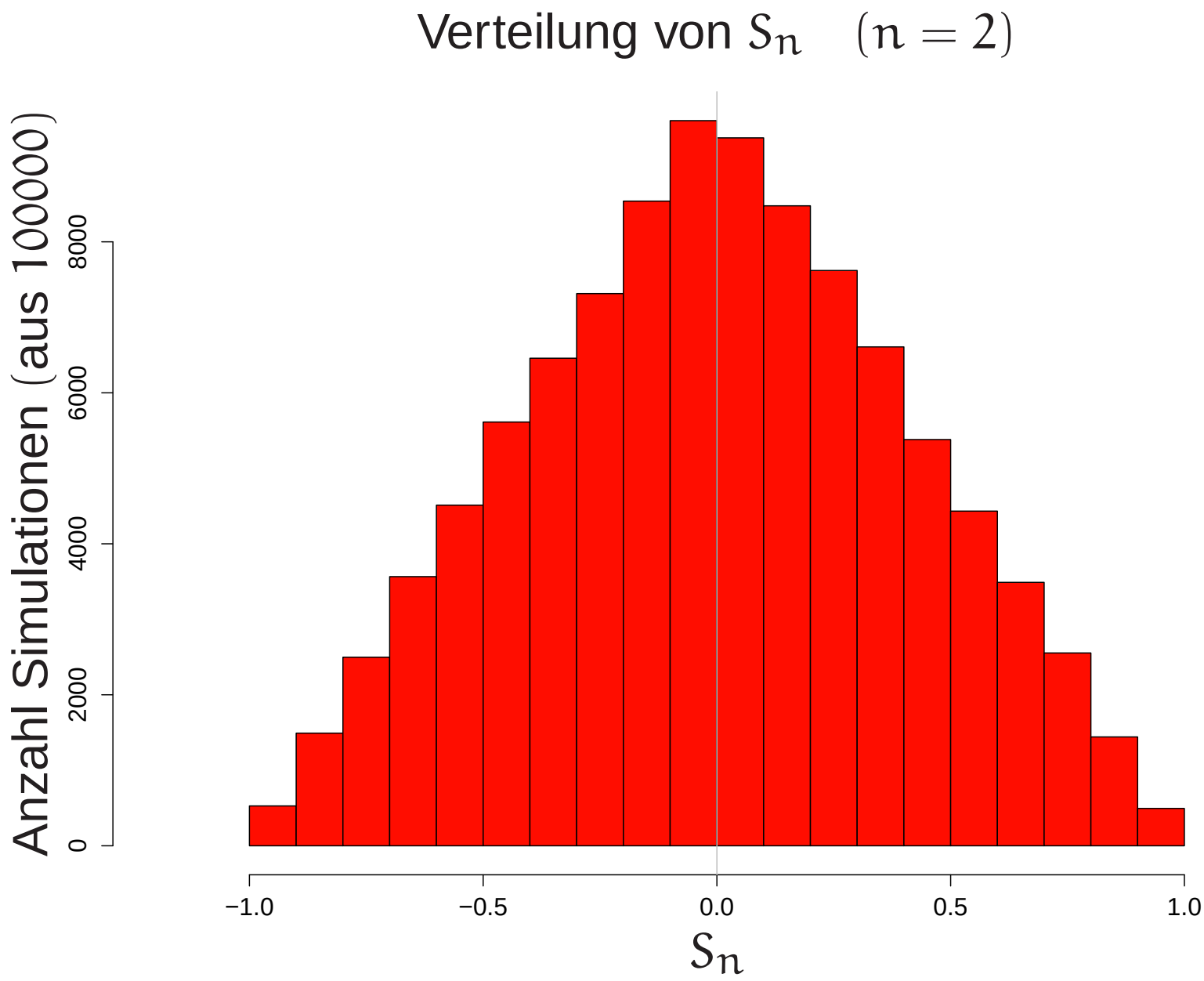
$$n = 15, 20, ..., 100$$

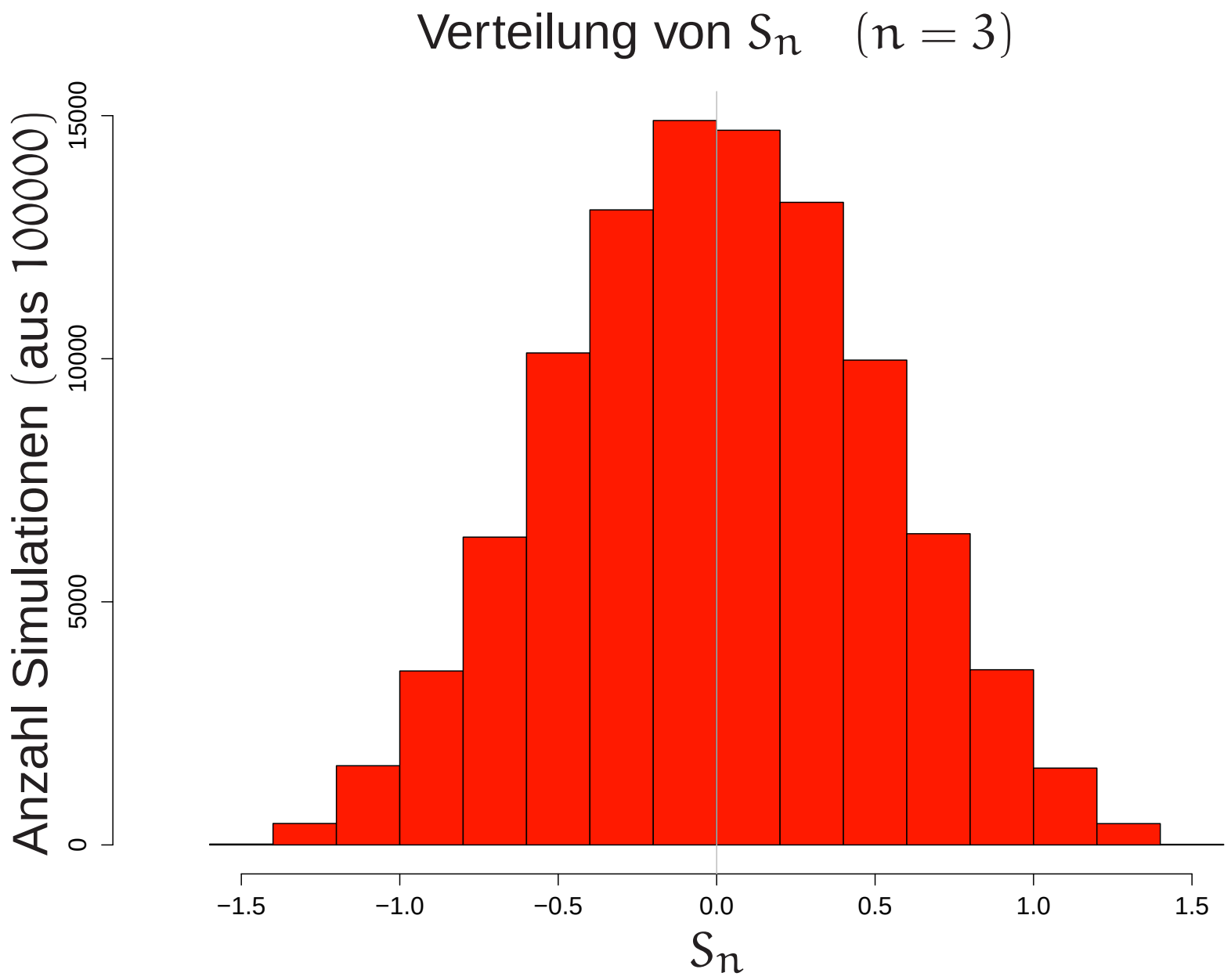

Dichtefunktion f_X der Verteilung von X
$f_X(x)$
2.0
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
x

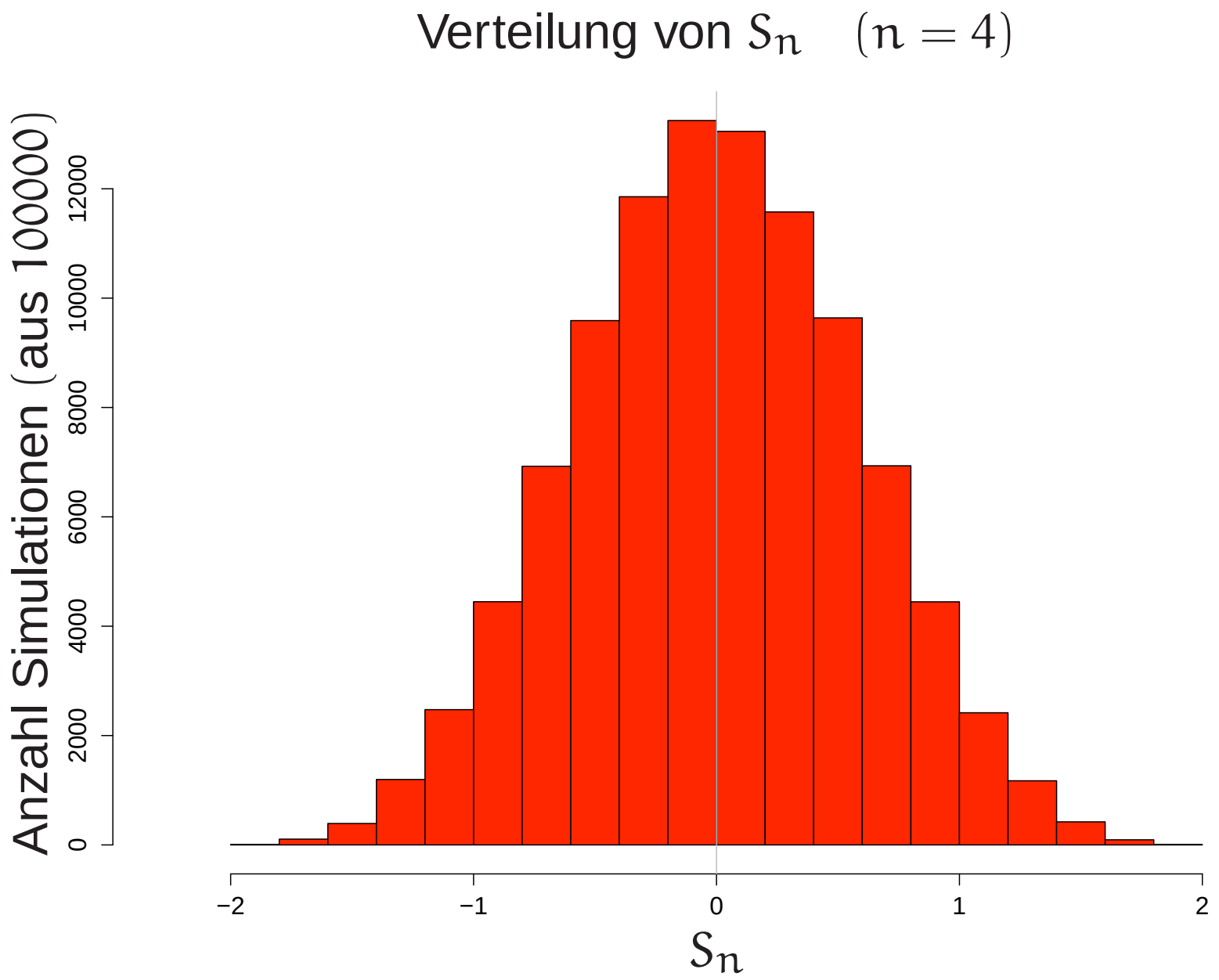

Dichtefunktion f_X der Verteilung von X
$\sigma^2 = \frac{1}{12}$
$-\sigma$
$\mathbf{E}X = 0$
$\sigma = \frac{1}{\sqrt{12}}$
$f_X(x)$
2.0
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
x

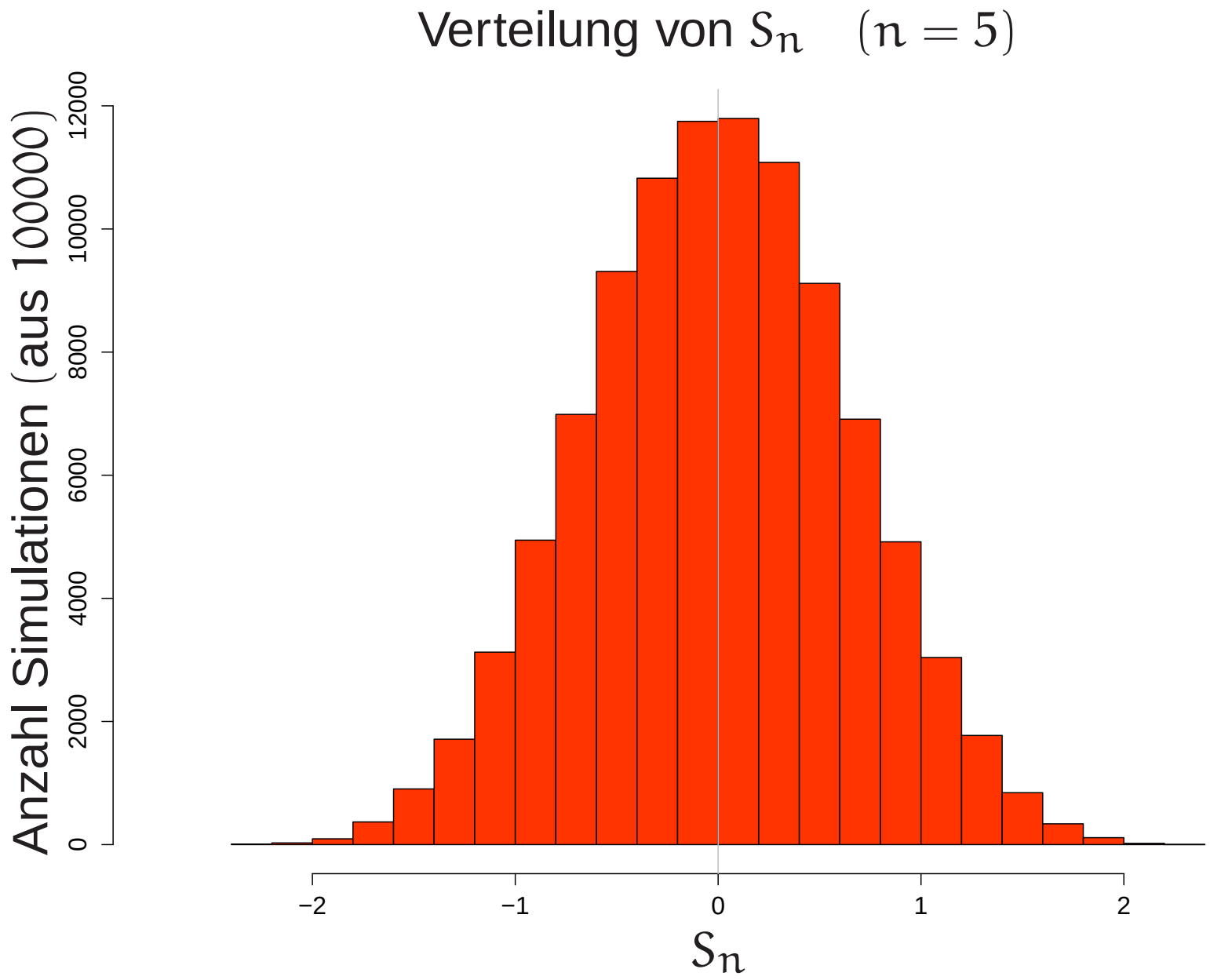

Dichtefunktion f_X der Verteilung von X
Aus dieser Verteilung wird 100000-mal n-mal gezogen.
$-\sigma$
EX
σ
$f_X(x)$
0.0
0.5
1.0
1.5
2.0
−0.5
0.0
0.5
x

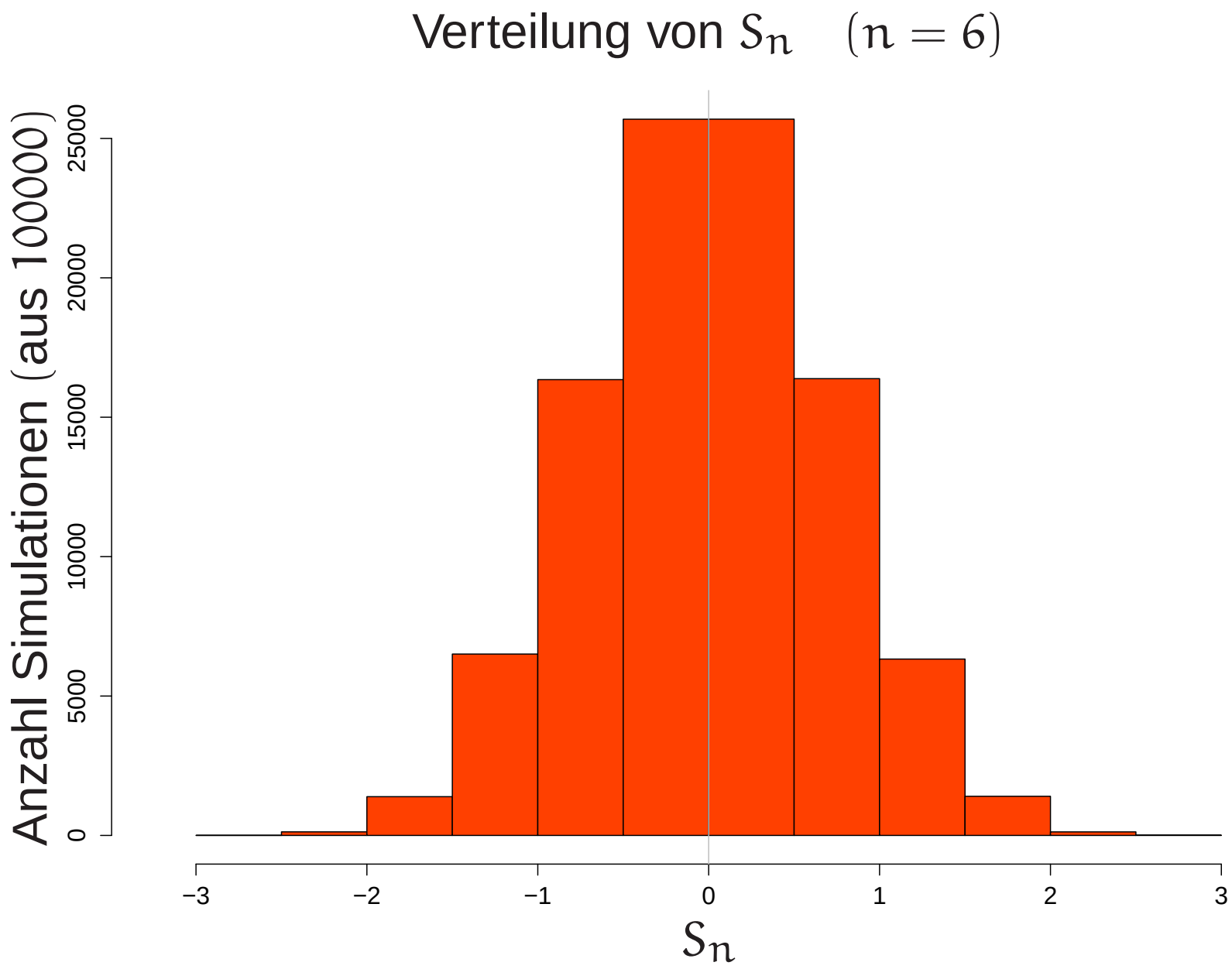

Verteilung von S_n $(n = 1)$
Anzahl Simulationen (aus 10000)
0
1000
2000
3000
4000
5000
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
S_n

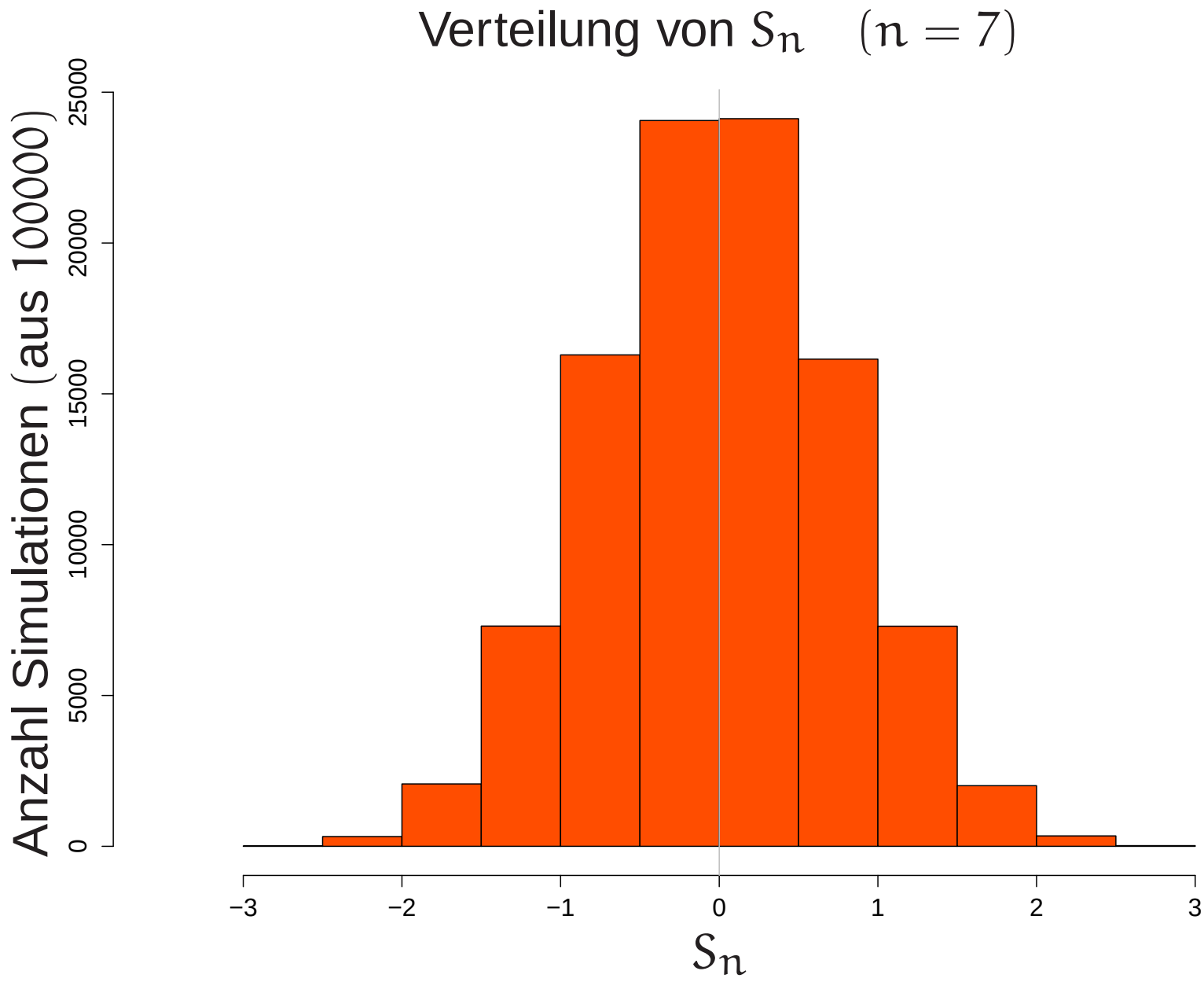

Verteilung von S_n $(n = 2)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−1.0
−0.5
0.0
0.5
1.0
S_n

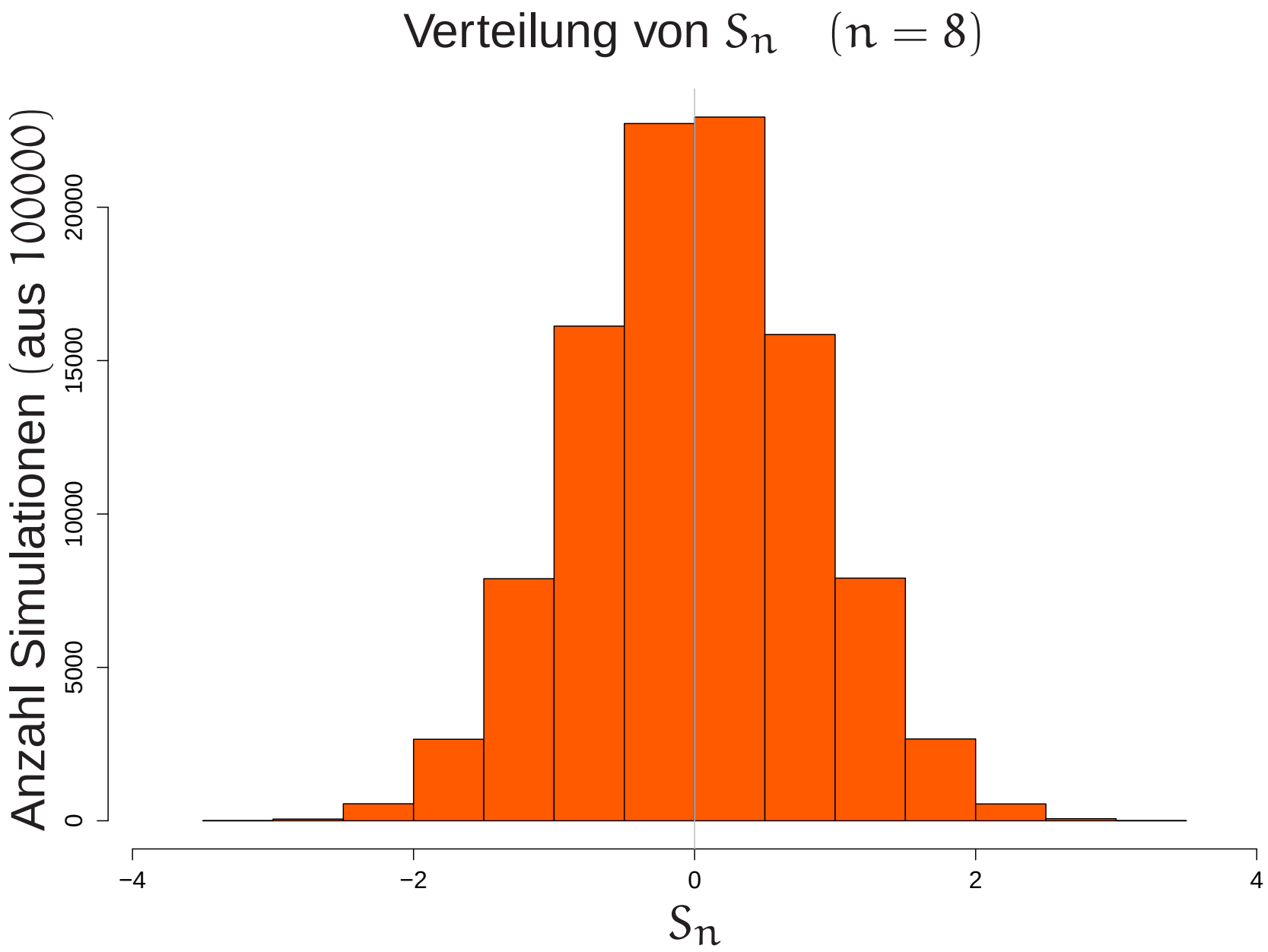

Verteilung von S_n $(n = 3)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
S_n

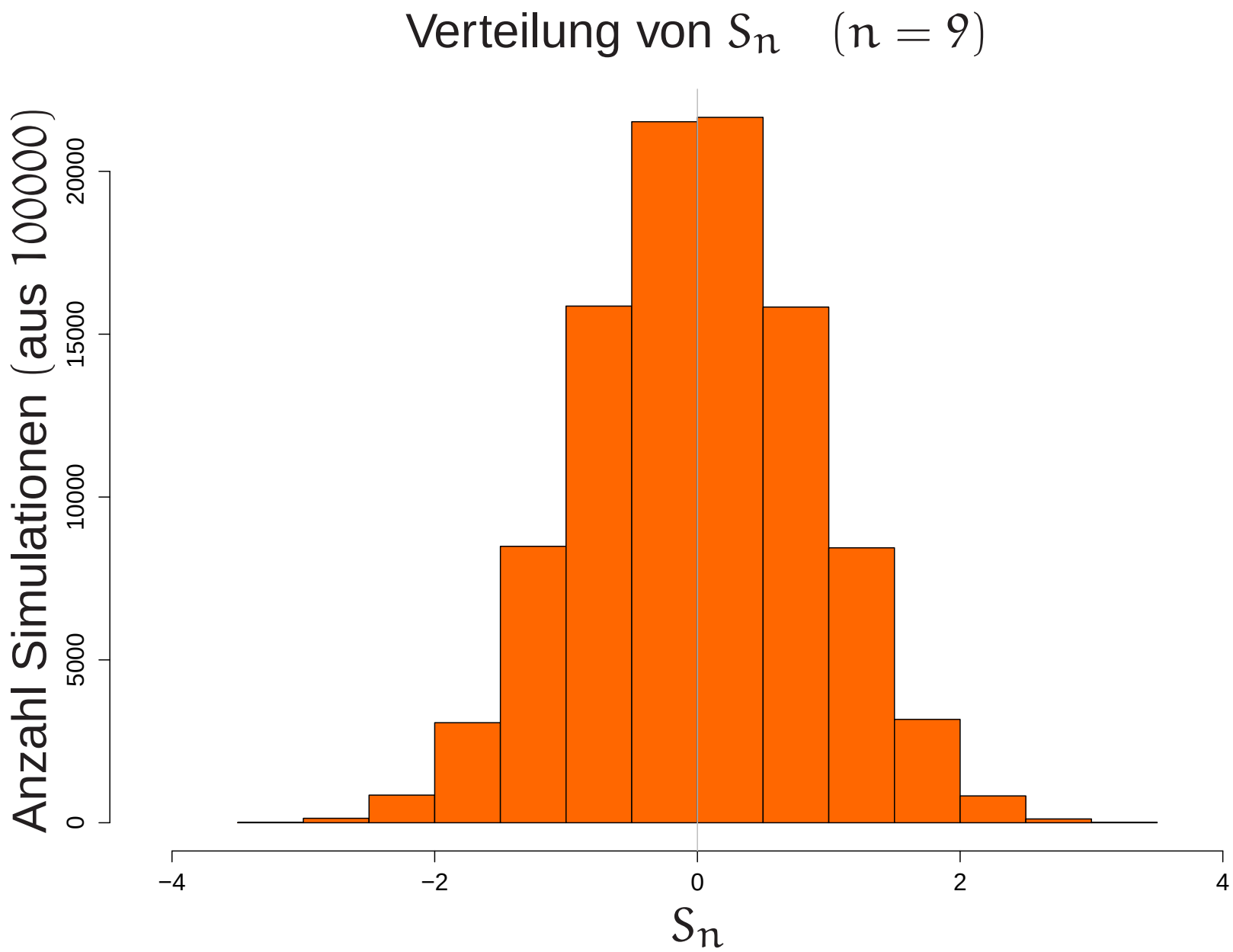

Verteilung von S_n $(n = 4)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
10000
12000
−2
−1
0
1
2
S_n

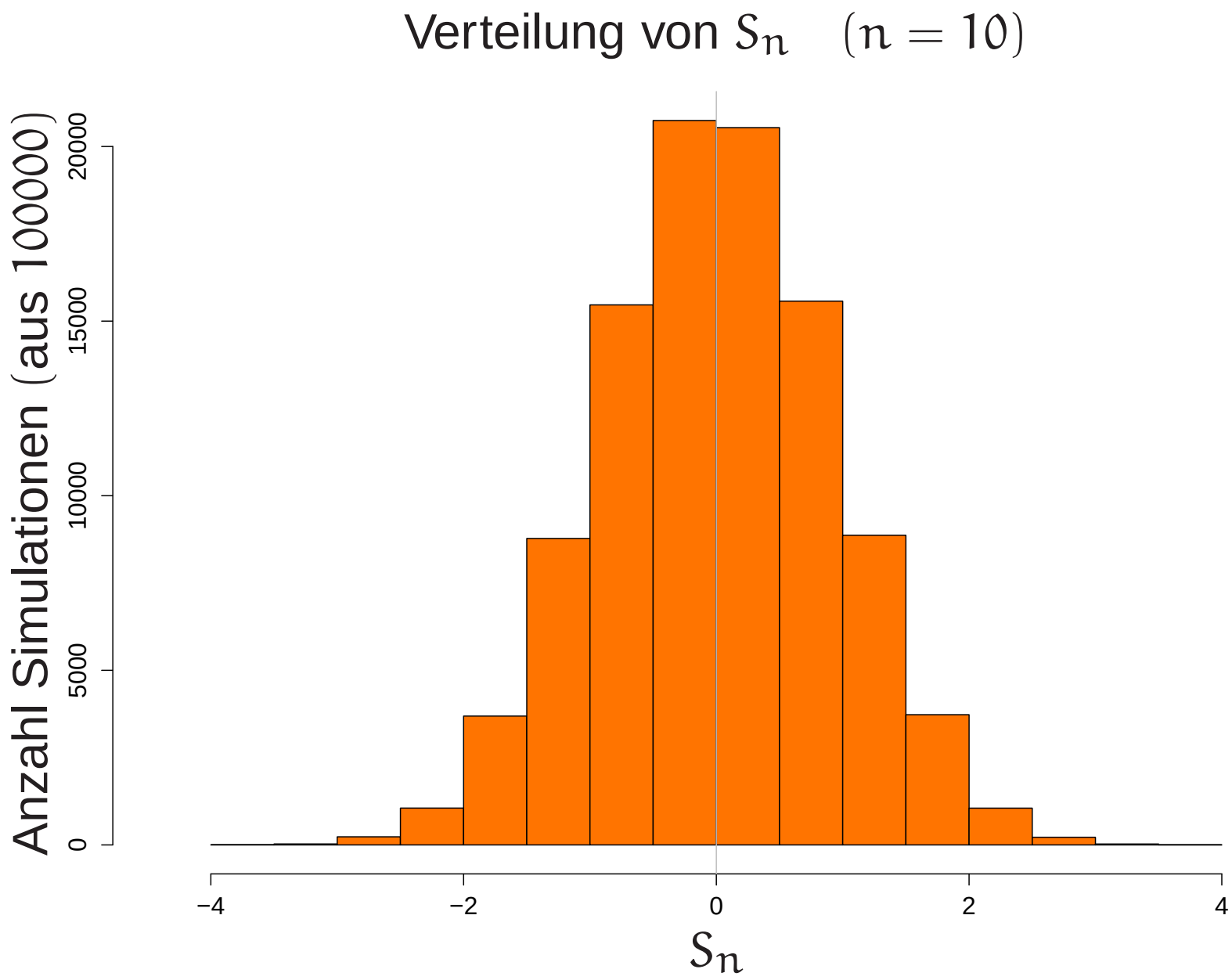

Verteilung von S_n $(n = 5)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
10000
12000
−2
−1
0
1
2
S_n

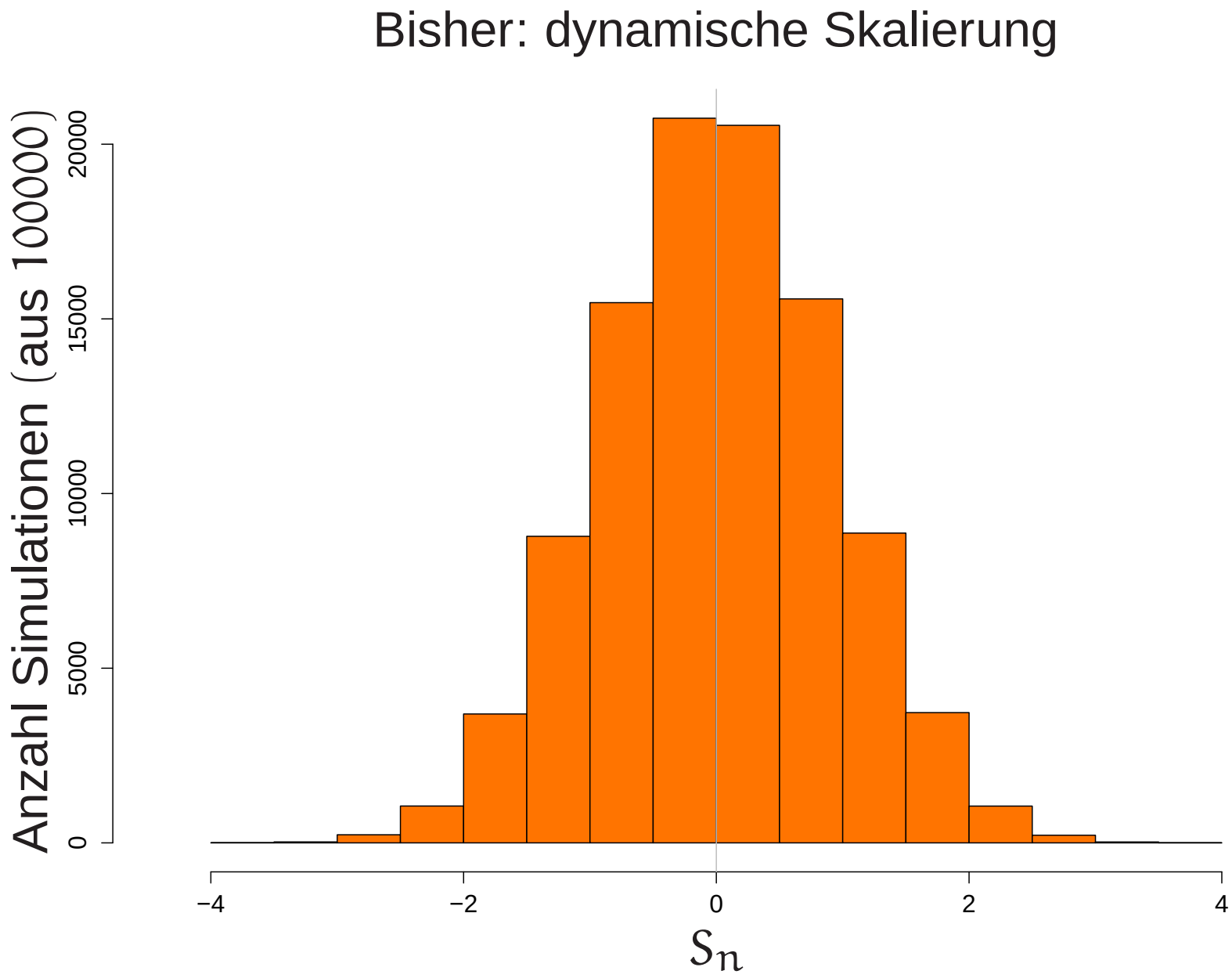

Verteilung von S_n $(n = 6)$
Anzahl Simulationen (aus 10000)
0
5000
10000
15000
20000
25000
−3
−2
−1
0
1
2
3
S_n

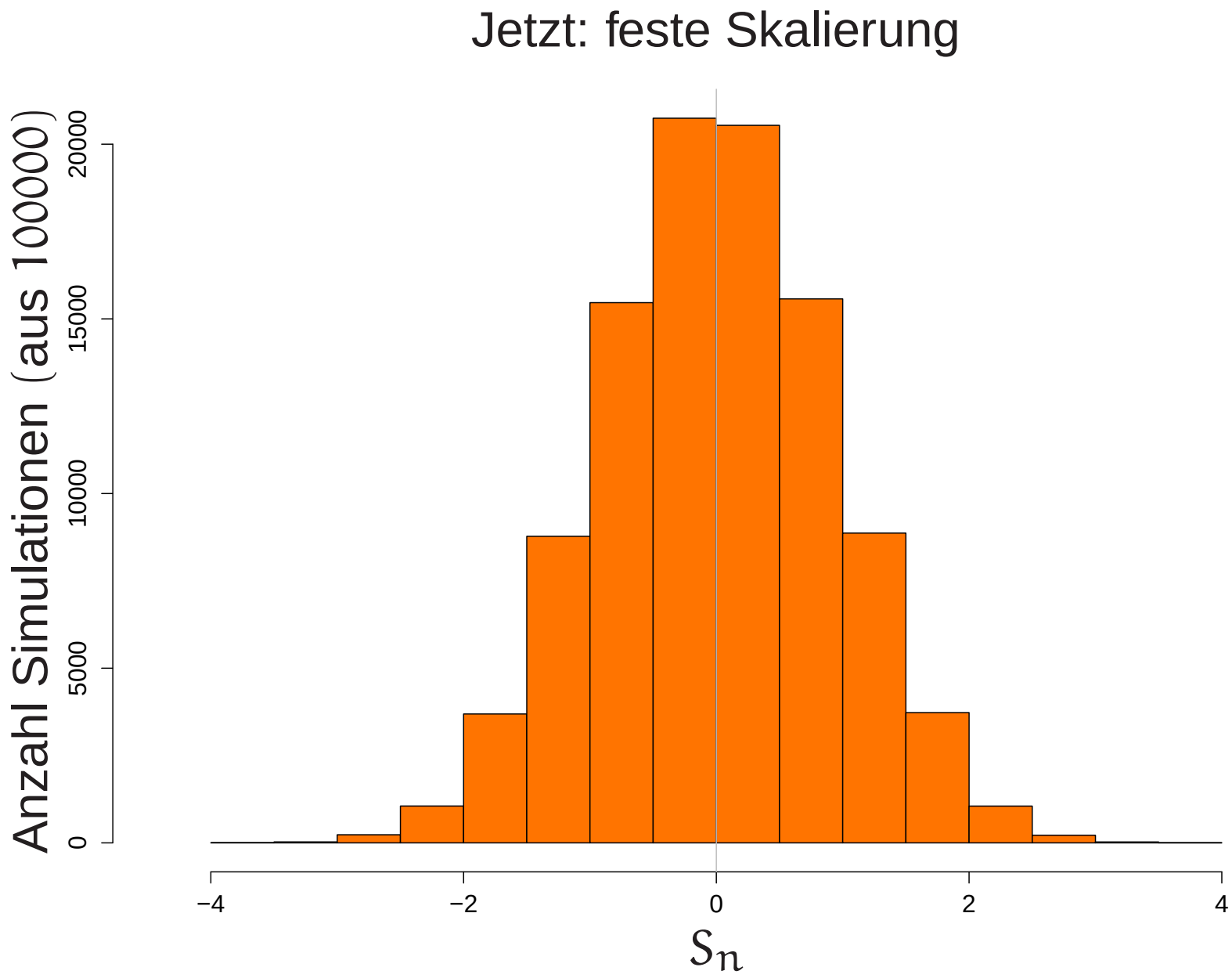

Verteilung von S_n $(n = 7)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
25000
−3
−2
−1
0
1
2
3
S_n

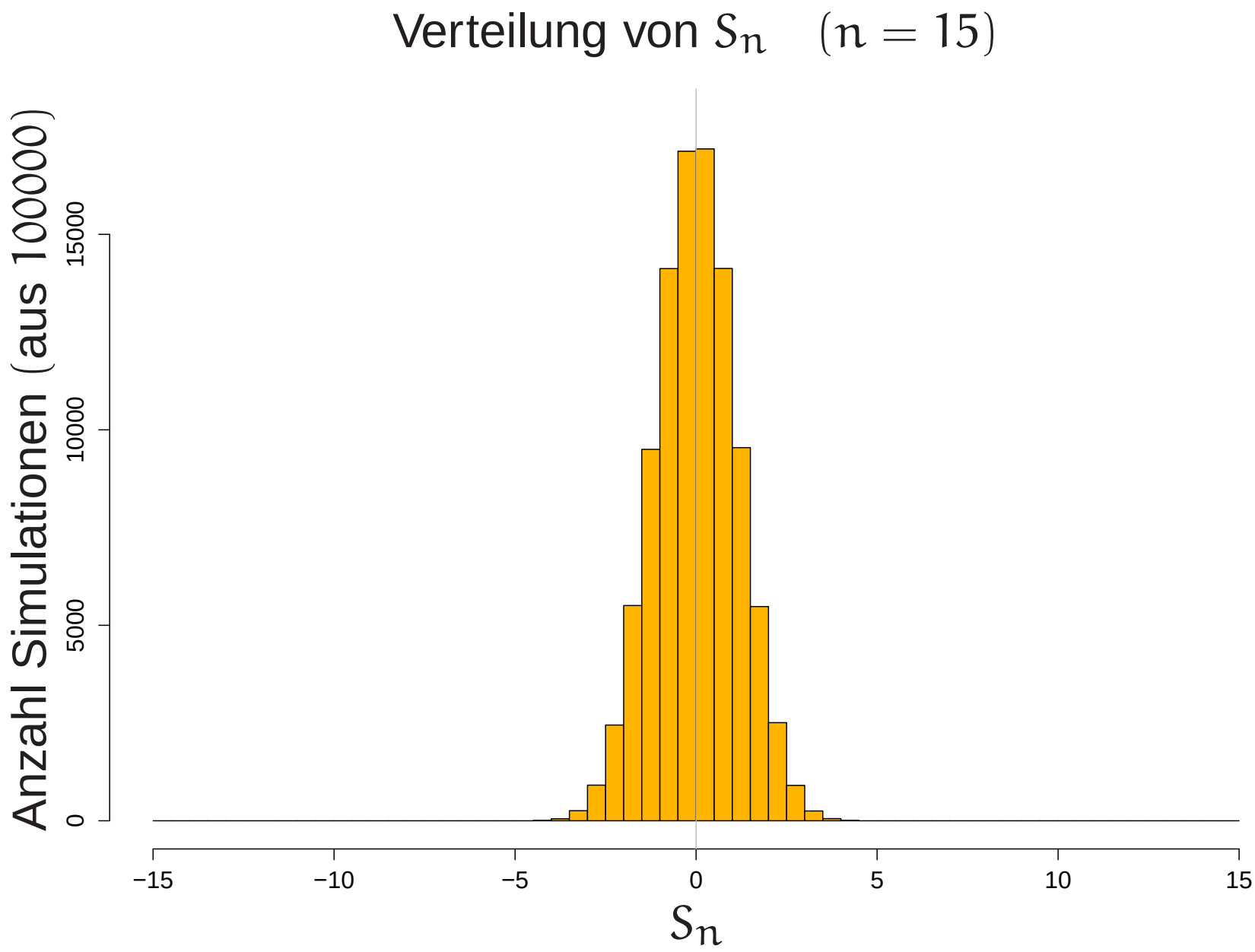

Verteilung von S_n $(n = 8)$
Anzahl Simulationen (aus 10000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

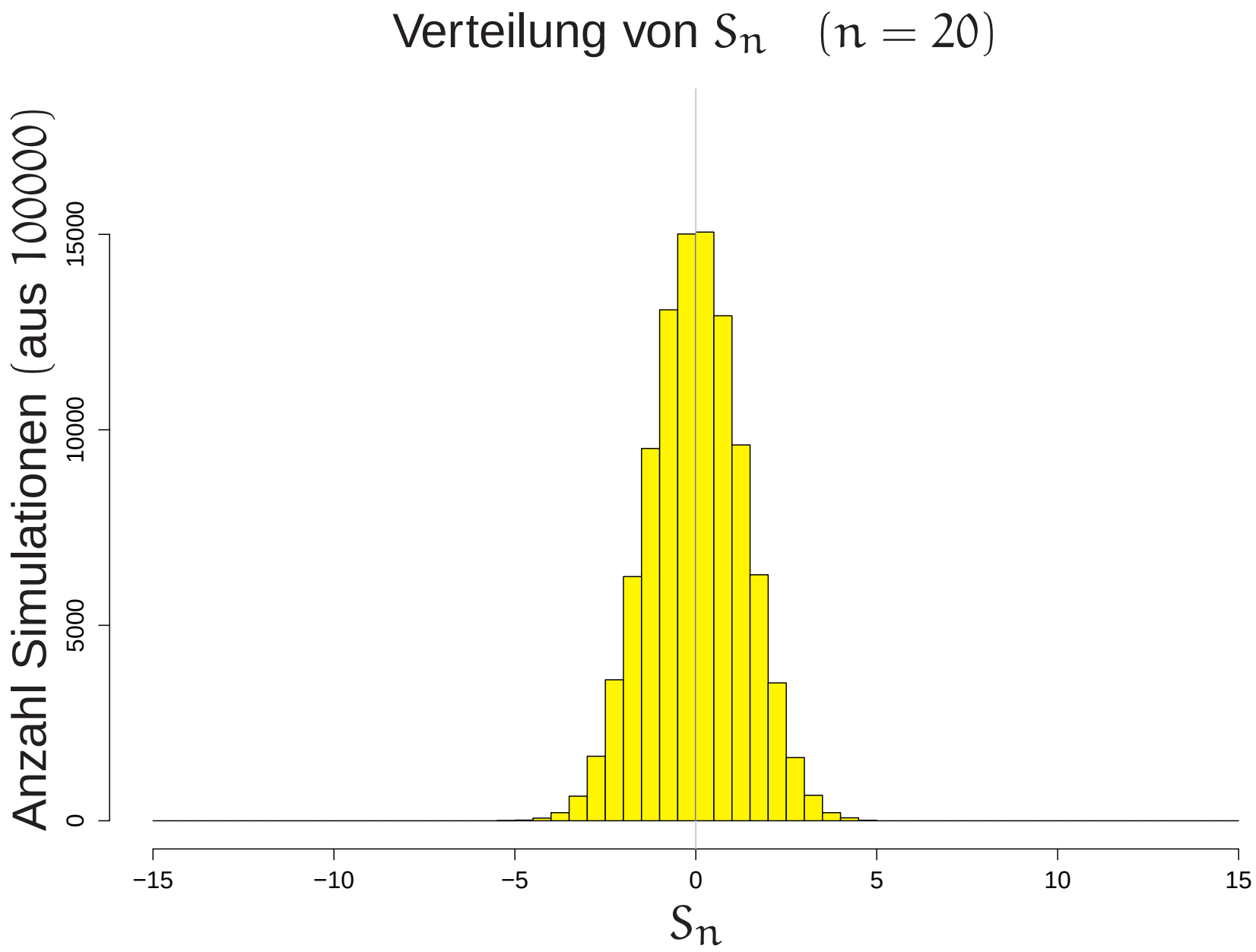

Verteilung von S_n $(n = 9)$
Anzahl Simulationen (aus 10000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

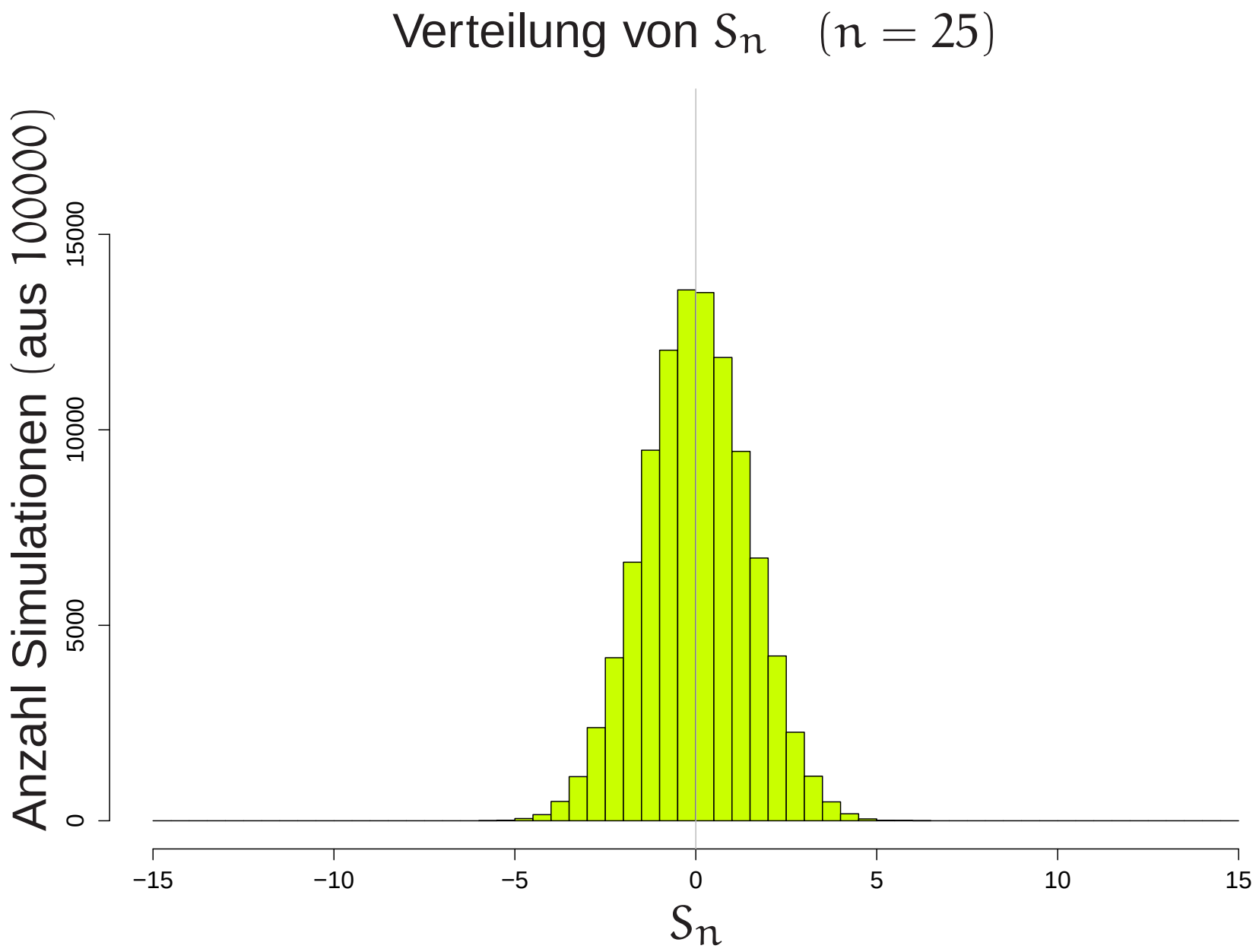

Verteilung von S_n $(n = 10)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

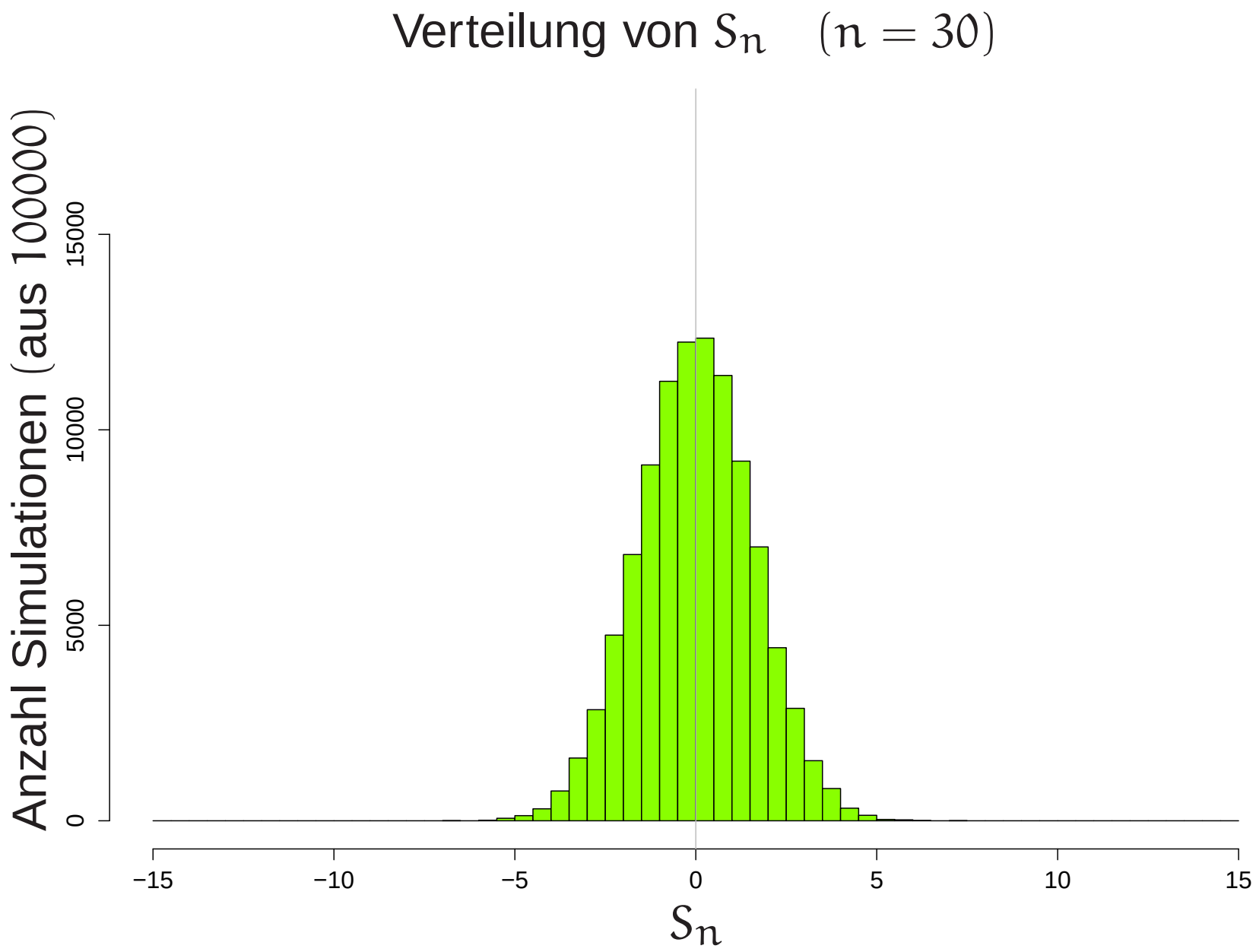

Bisher: dynamische Skalierung
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

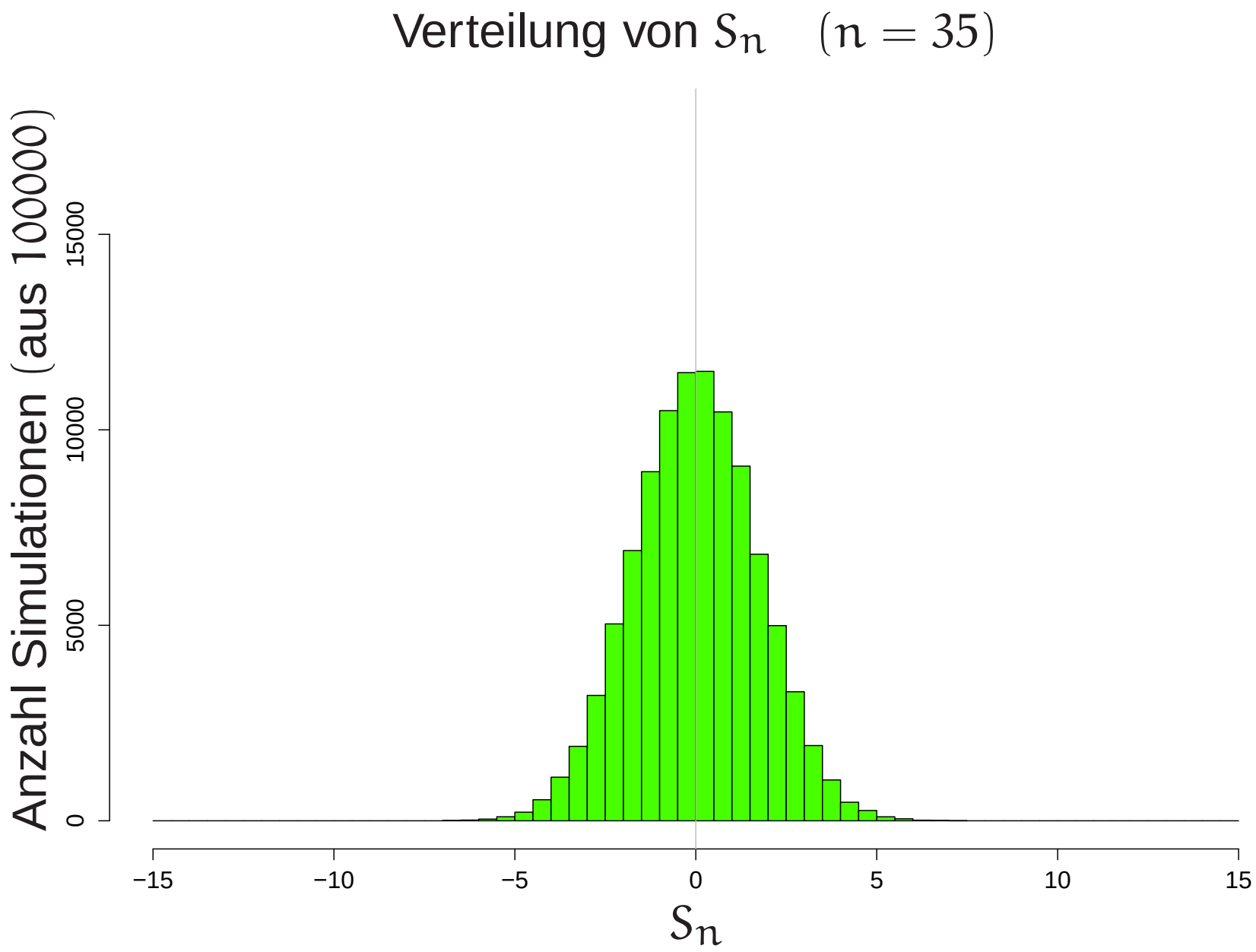

Jetzt: feste Skalierung
Anzahl Simulationen (aus 10000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

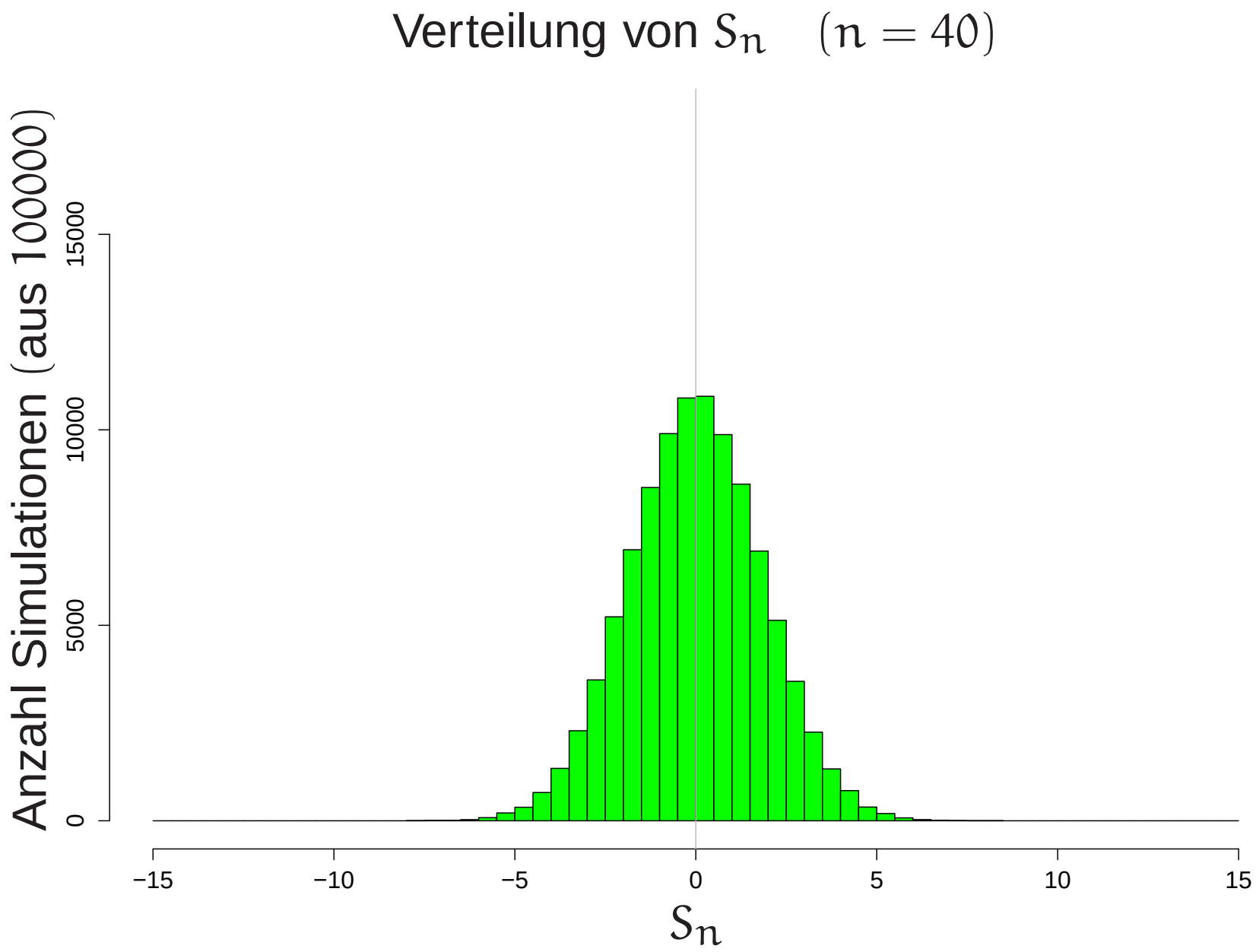

Verteilung von S_n $(n = 15)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

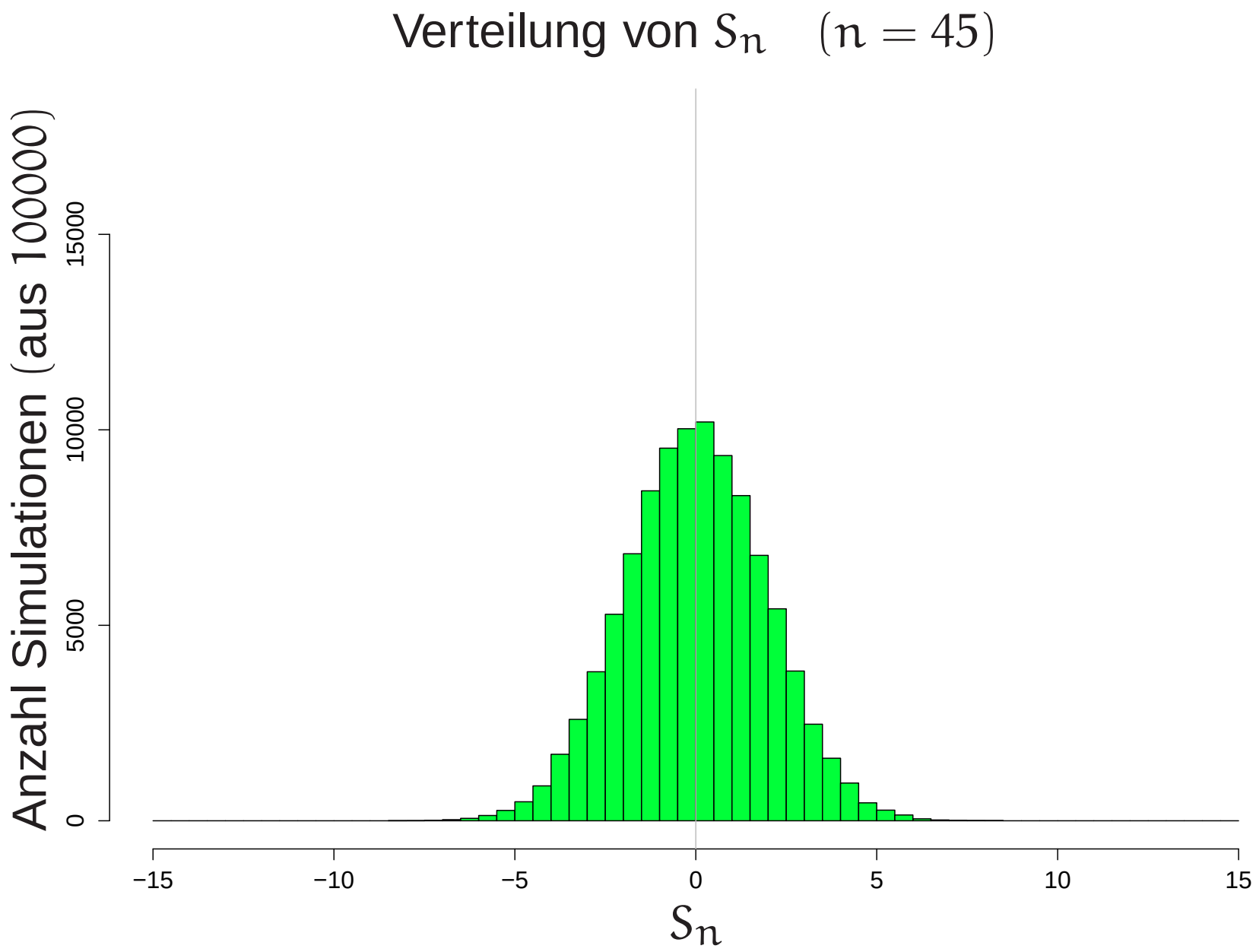

Verteilung von S_n $(n = 20)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

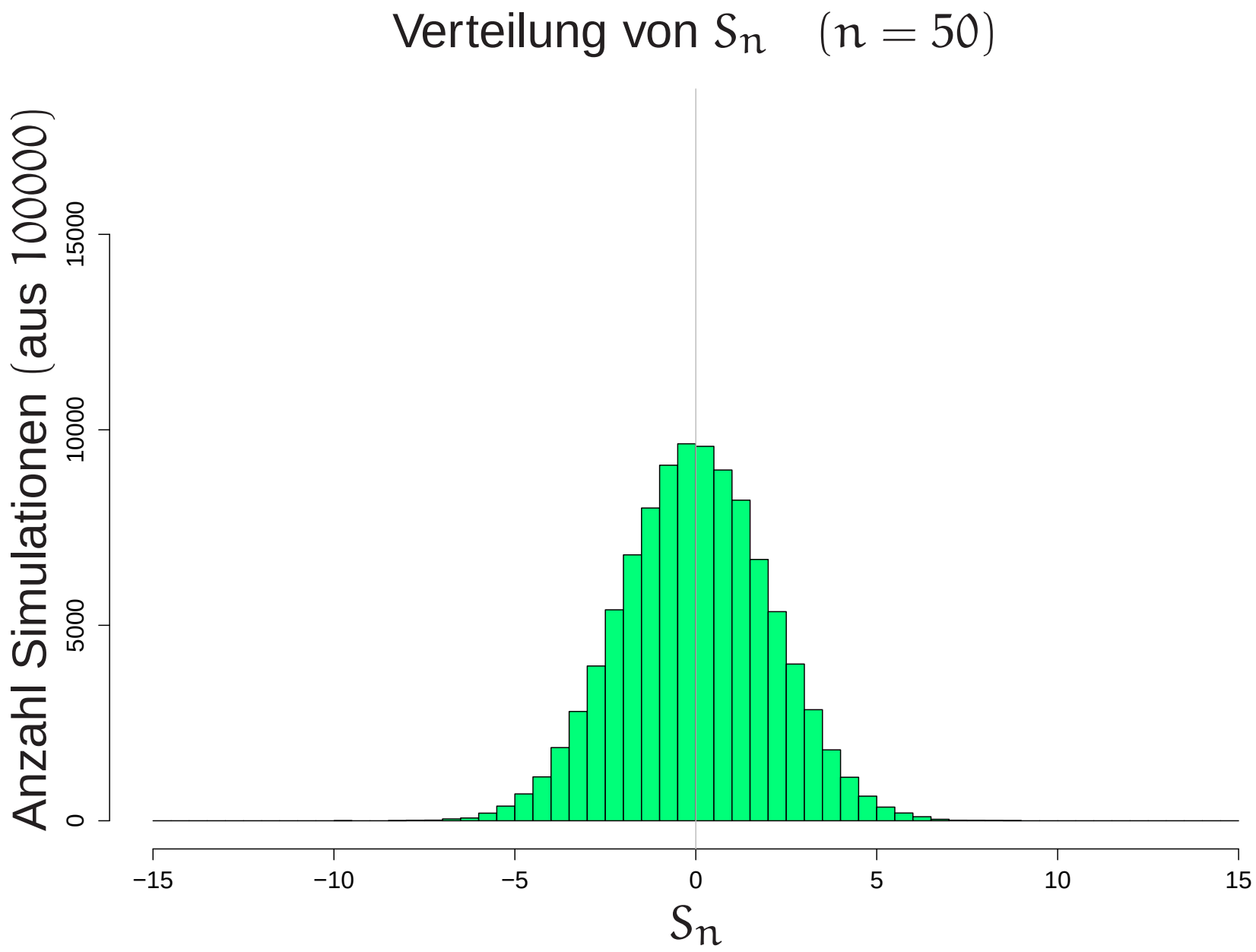

Verteilung von S_n $(n = 25)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

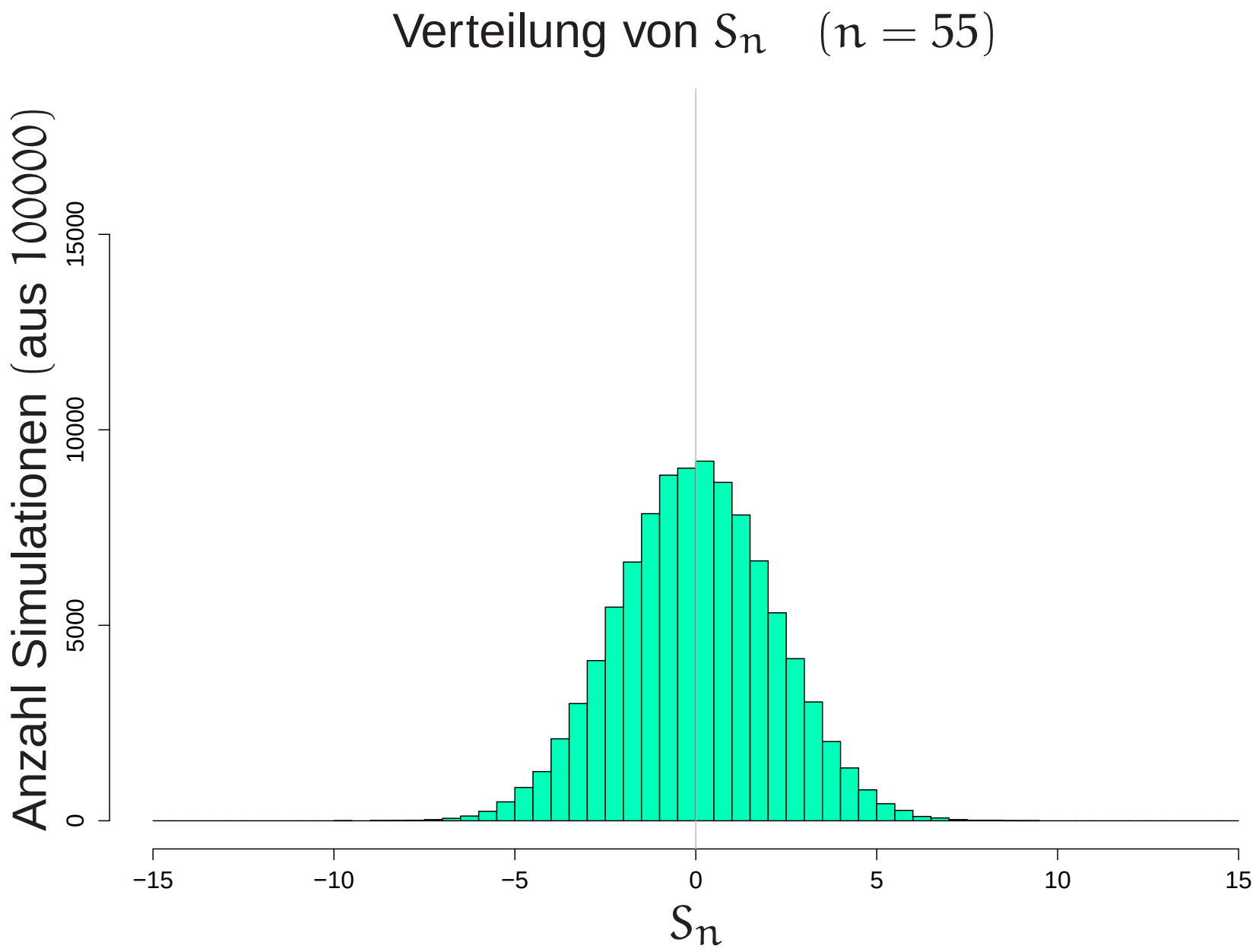

Verteilung von S_n $(n = 30)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

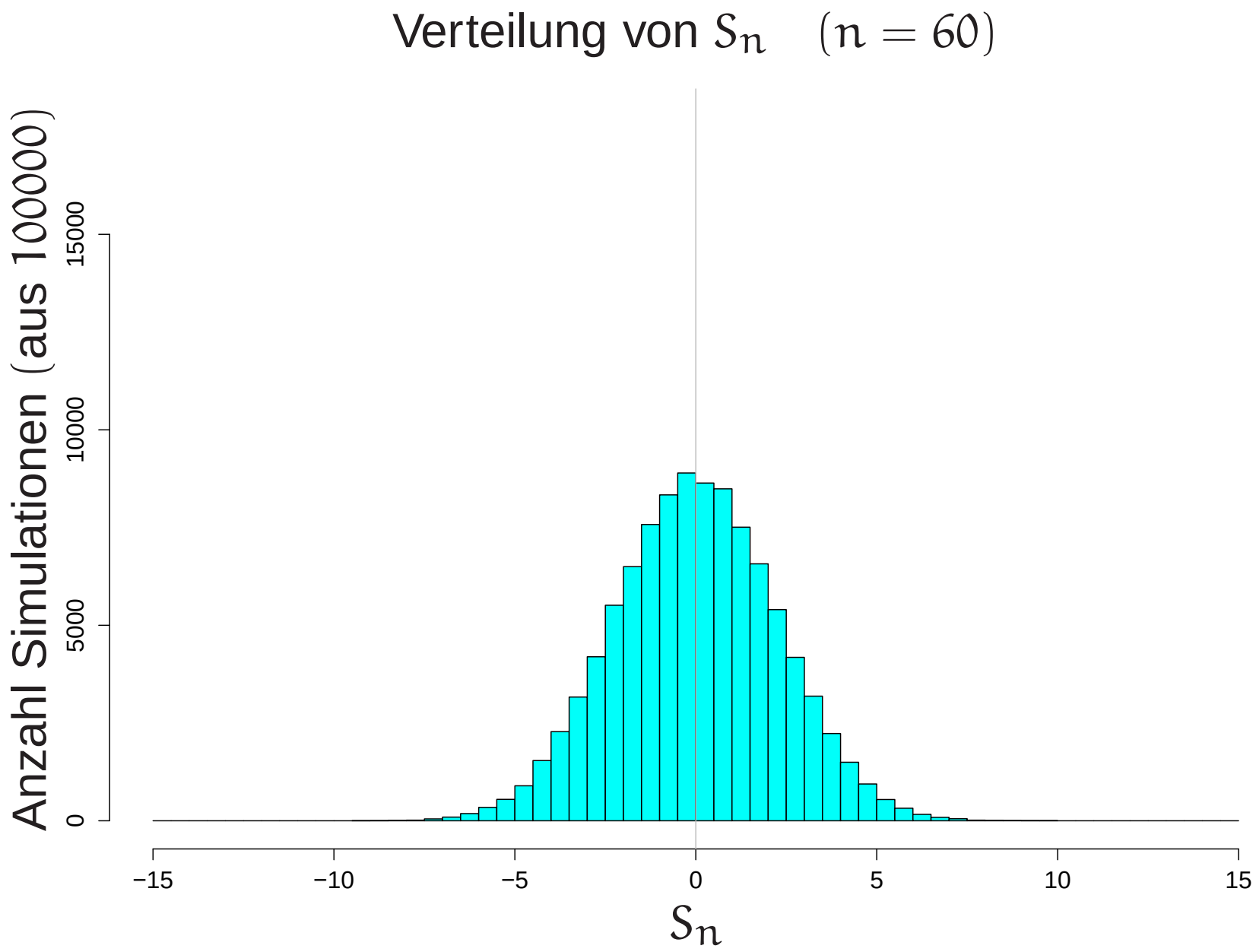

Verteilung von S_n $(n = 35)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

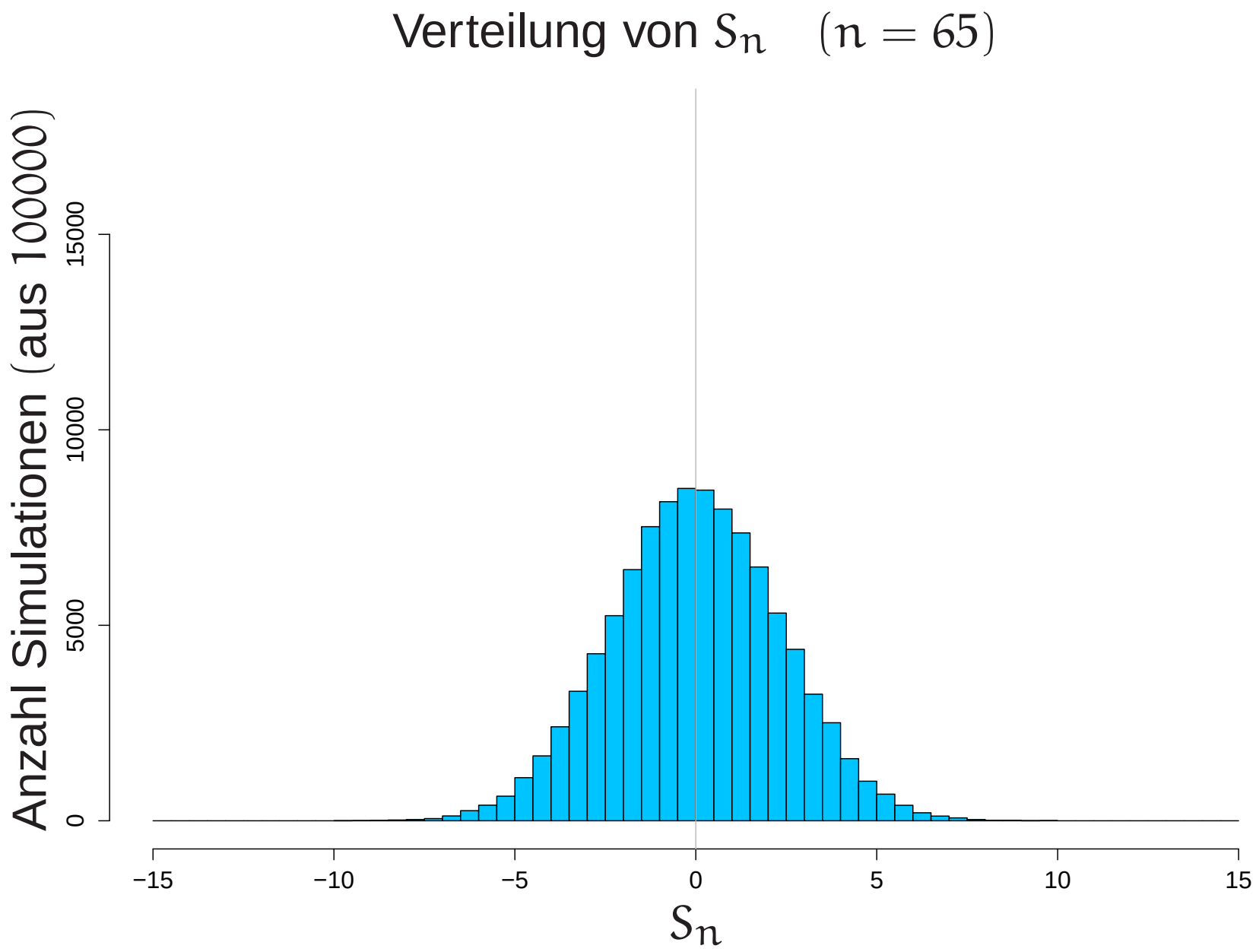

Verteilung von S_n $(n = 40)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

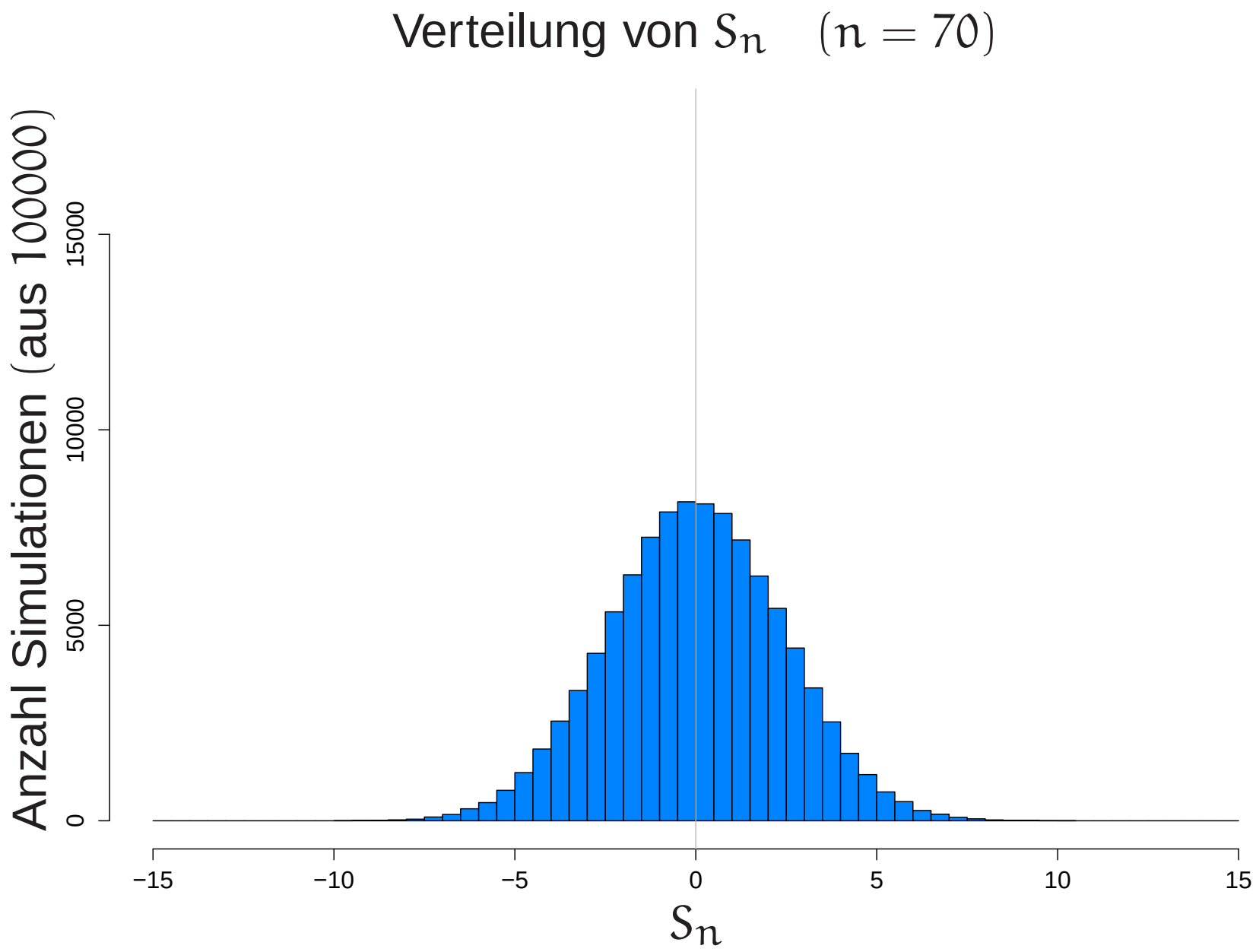

Verteilung von S_n $(n = 45)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

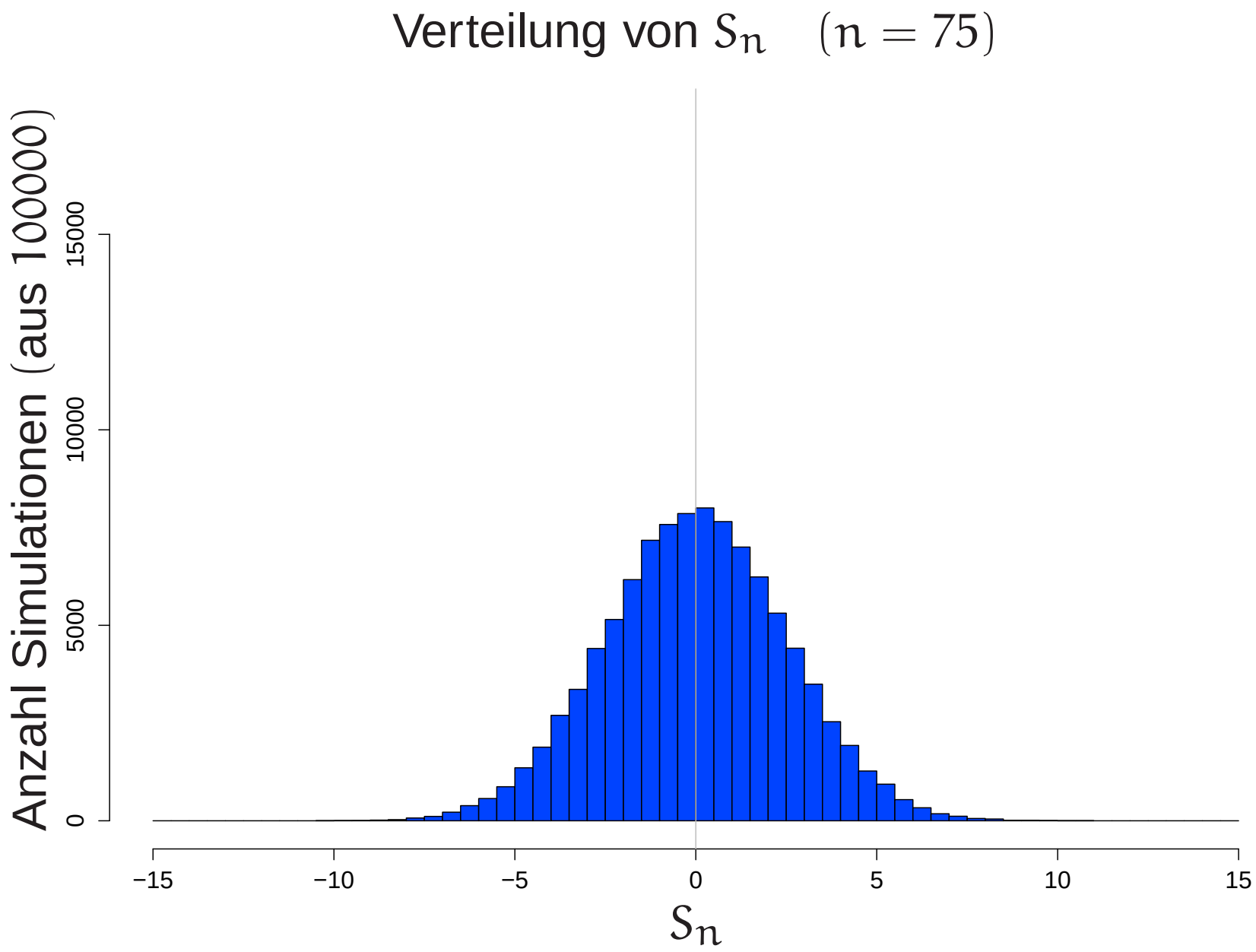

Verteilung von S_n $(n = 50)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

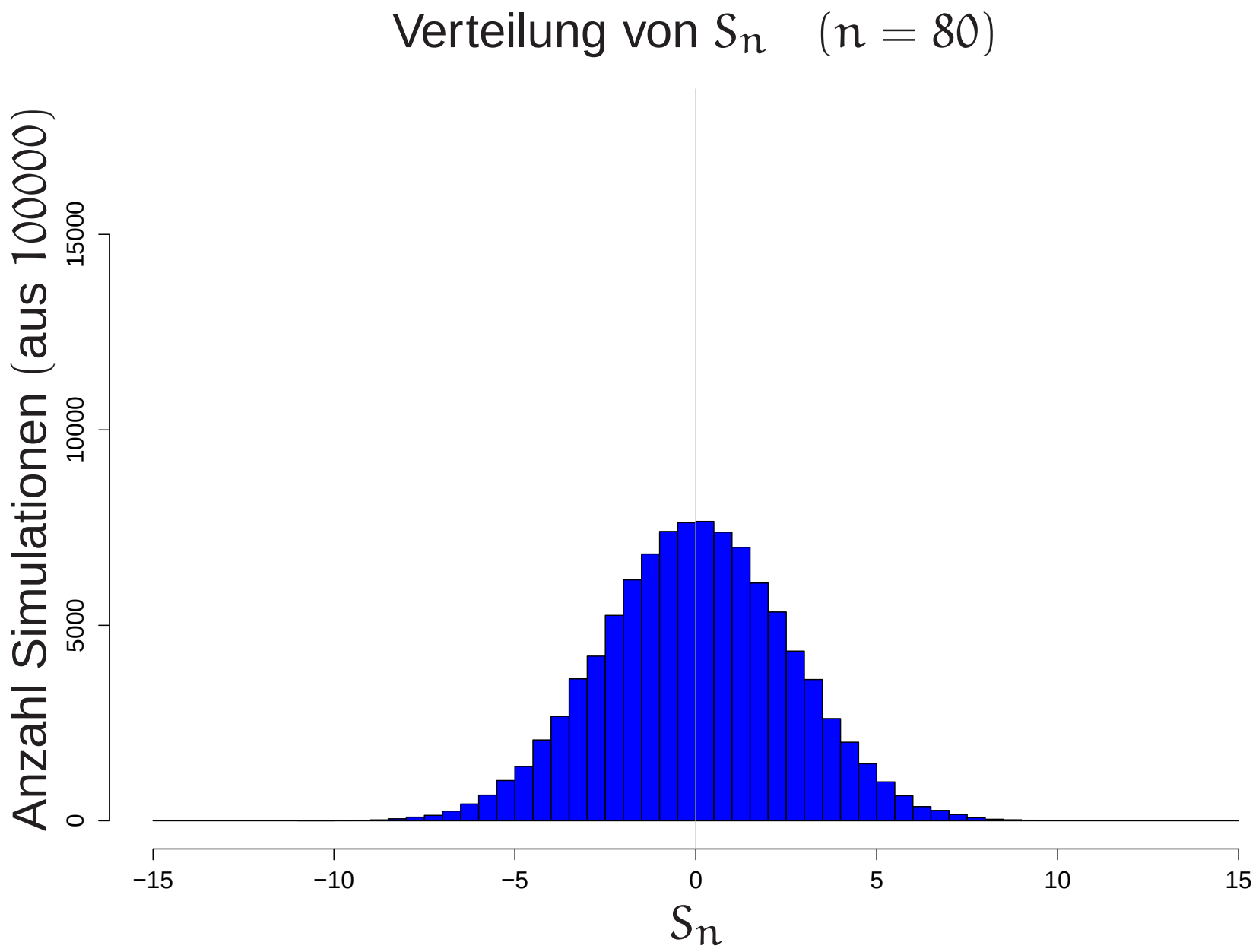

Verteilung von S_n $(n = 55)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

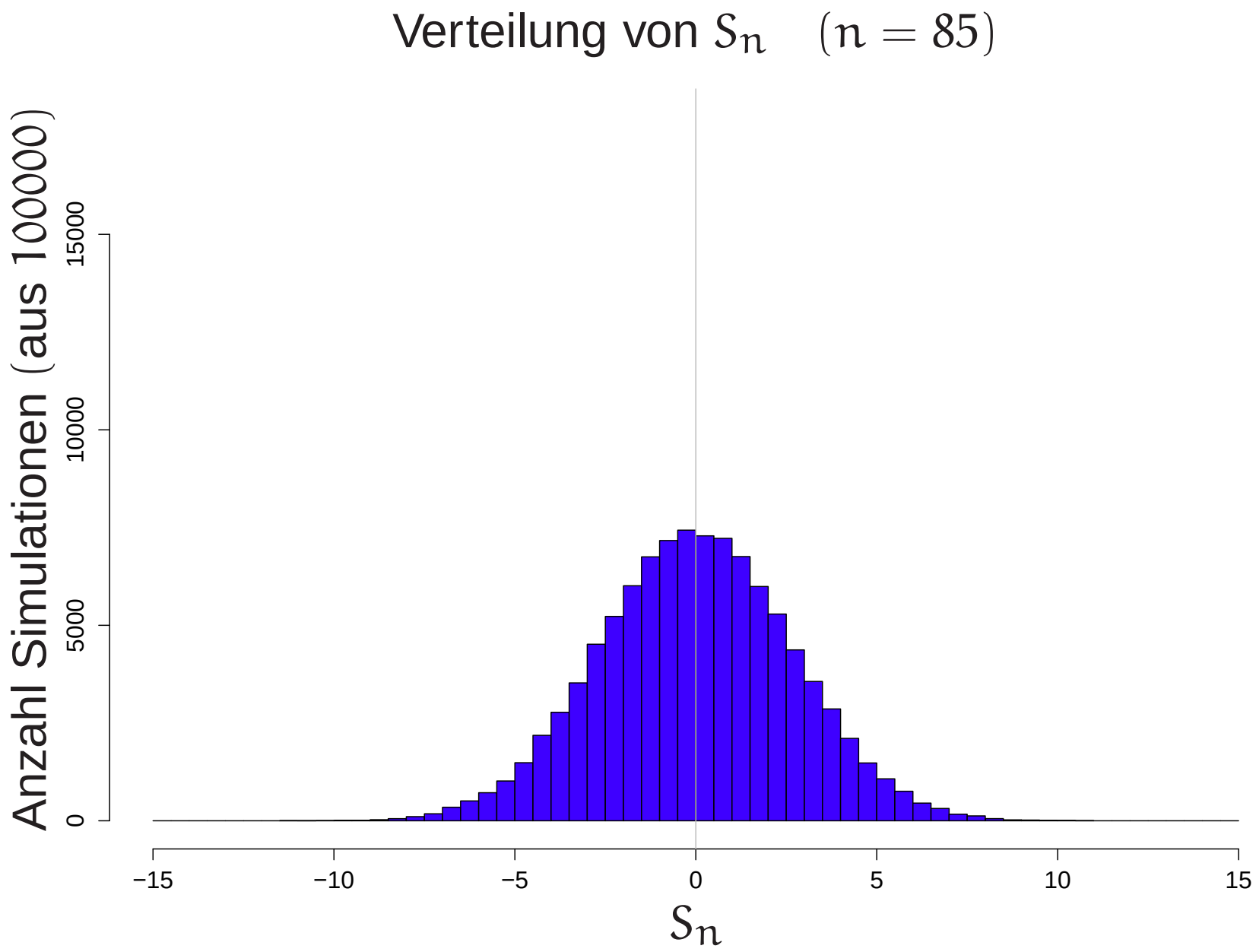

Verteilung von S_n $(n = 60)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

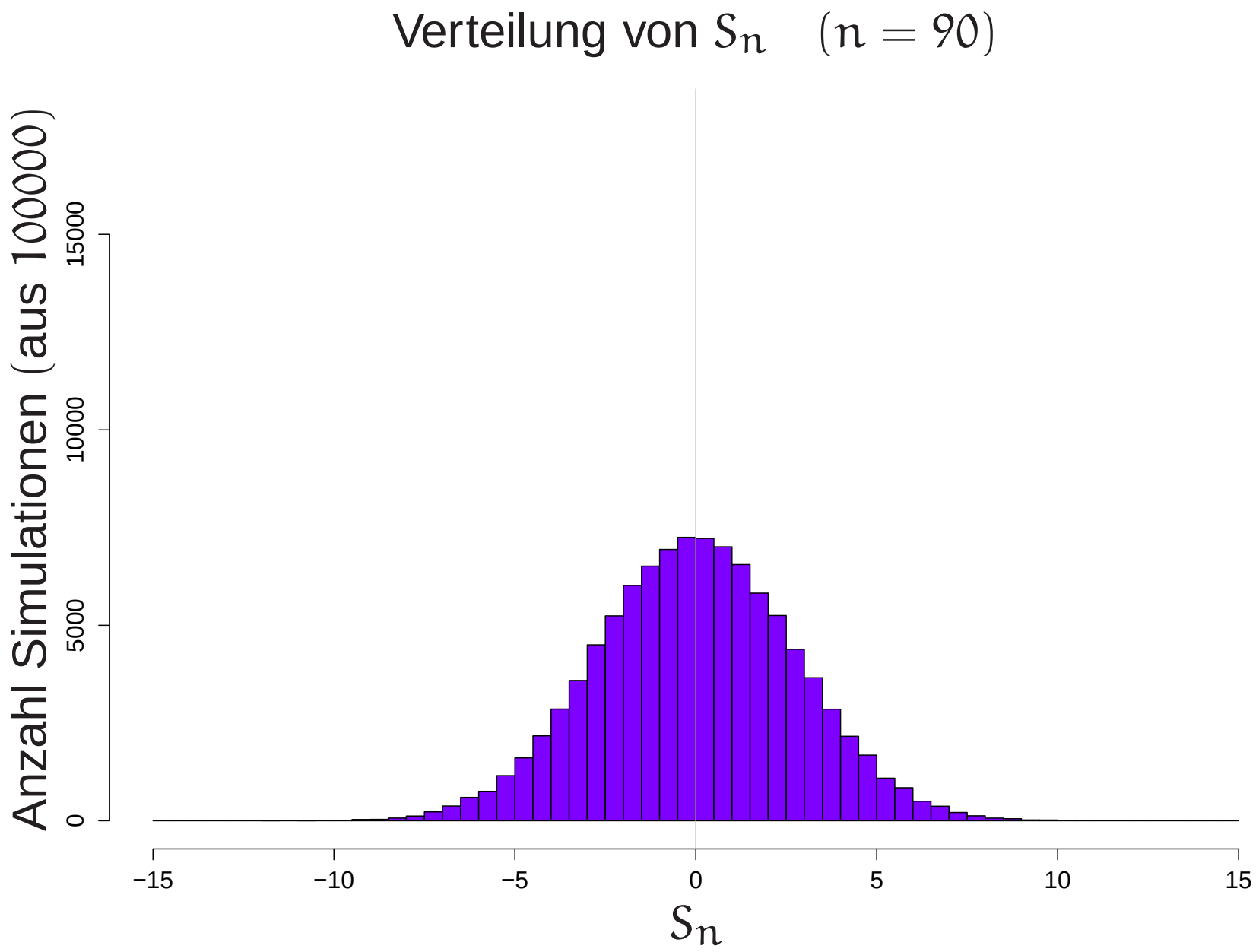

Verteilung von S_n $(n = 65)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

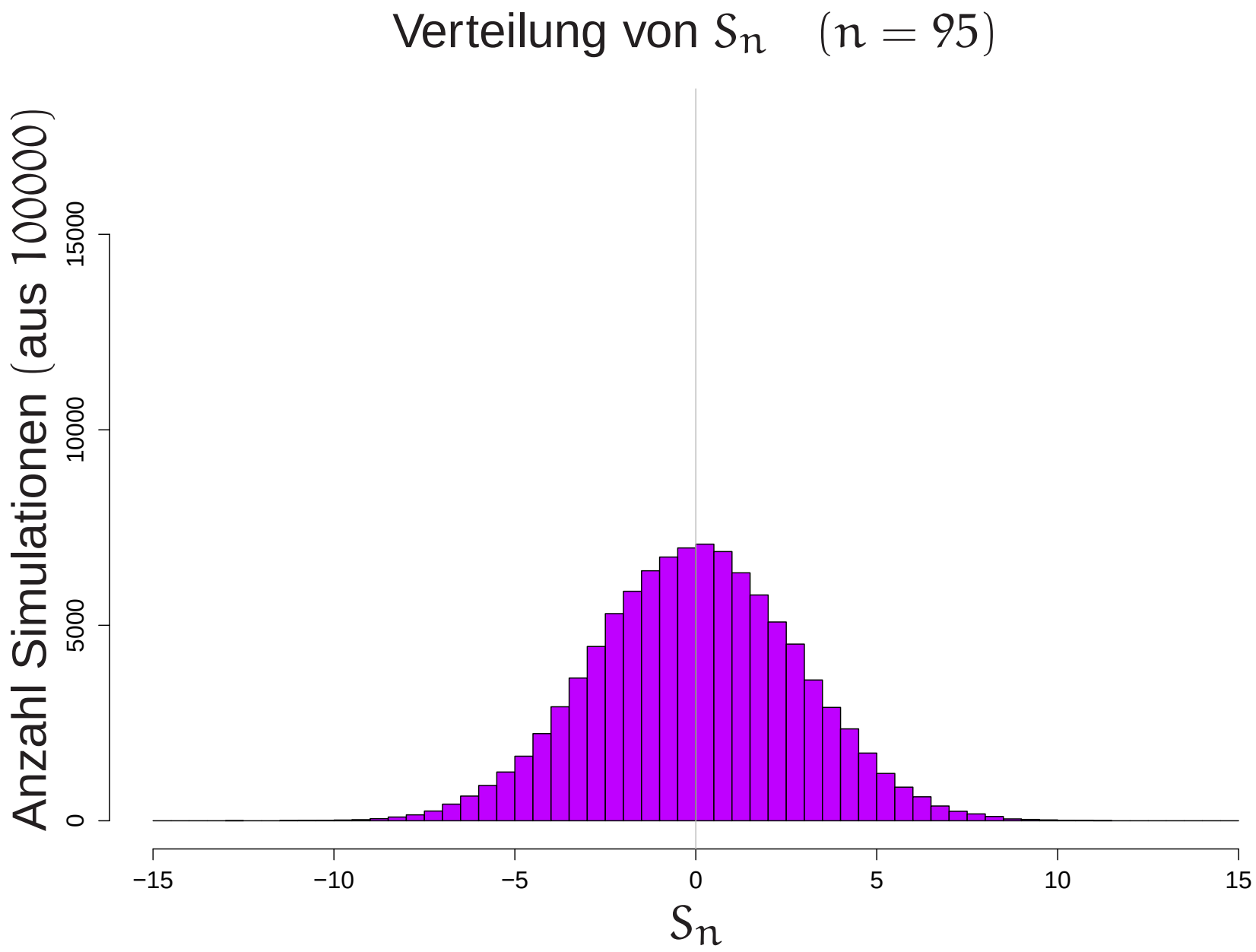

Verteilung von S_n $(n = 70)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

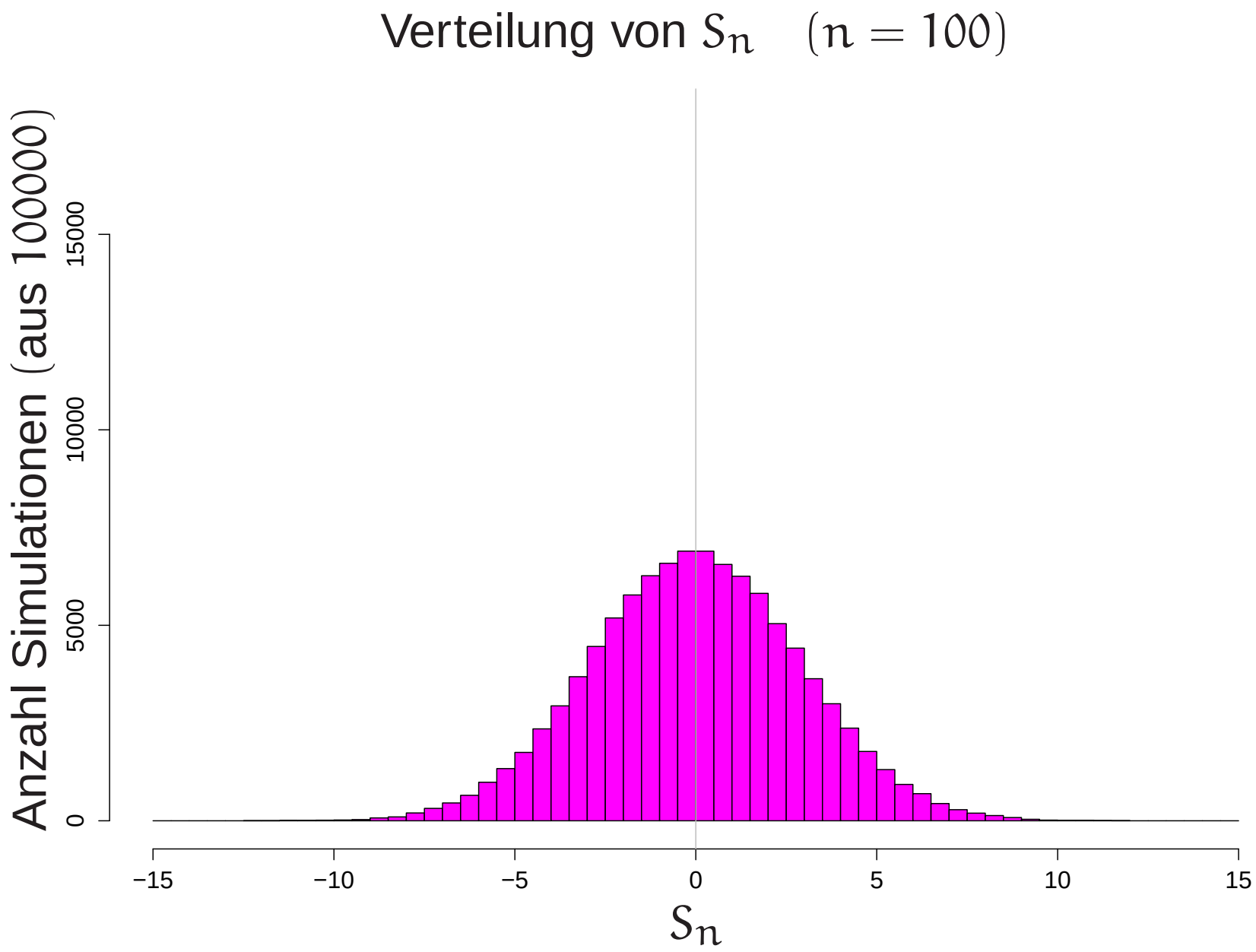

Verteilung von S_n $(n = 75)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


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

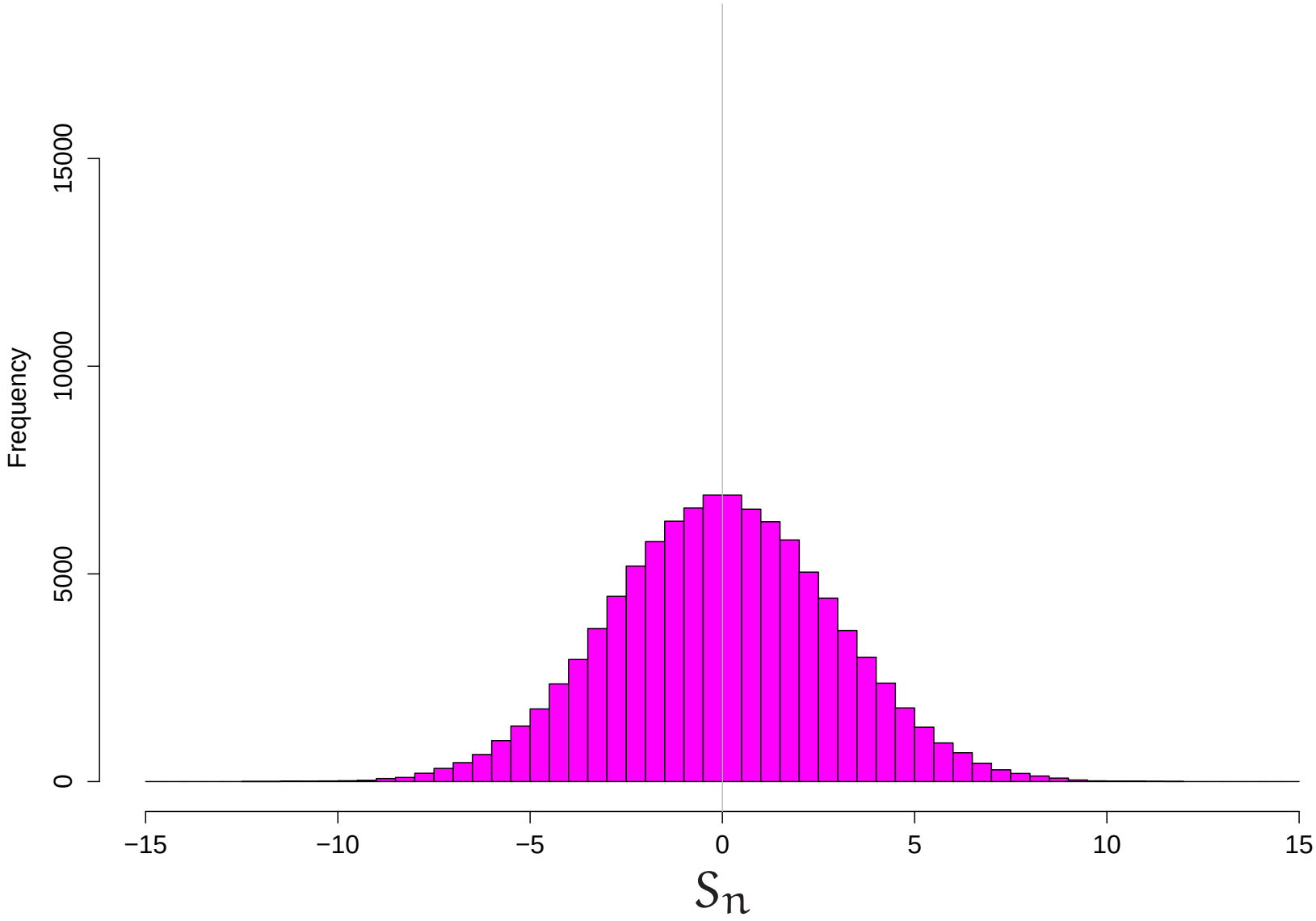

Verteilung von S_n $(n = 85)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


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

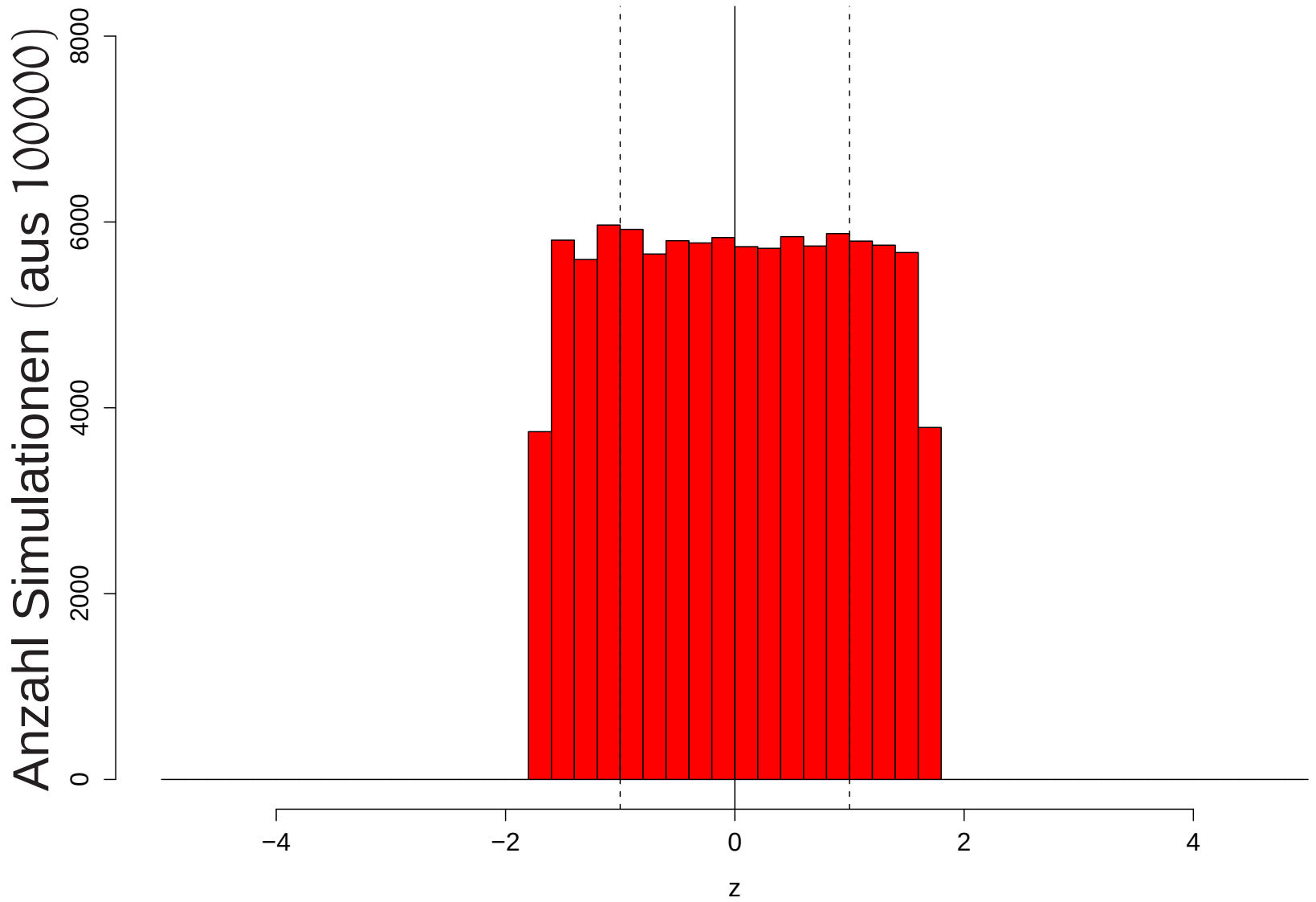

Verteilung von S_n $(n = 95)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


Verteilung von S_n $(n = 100)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

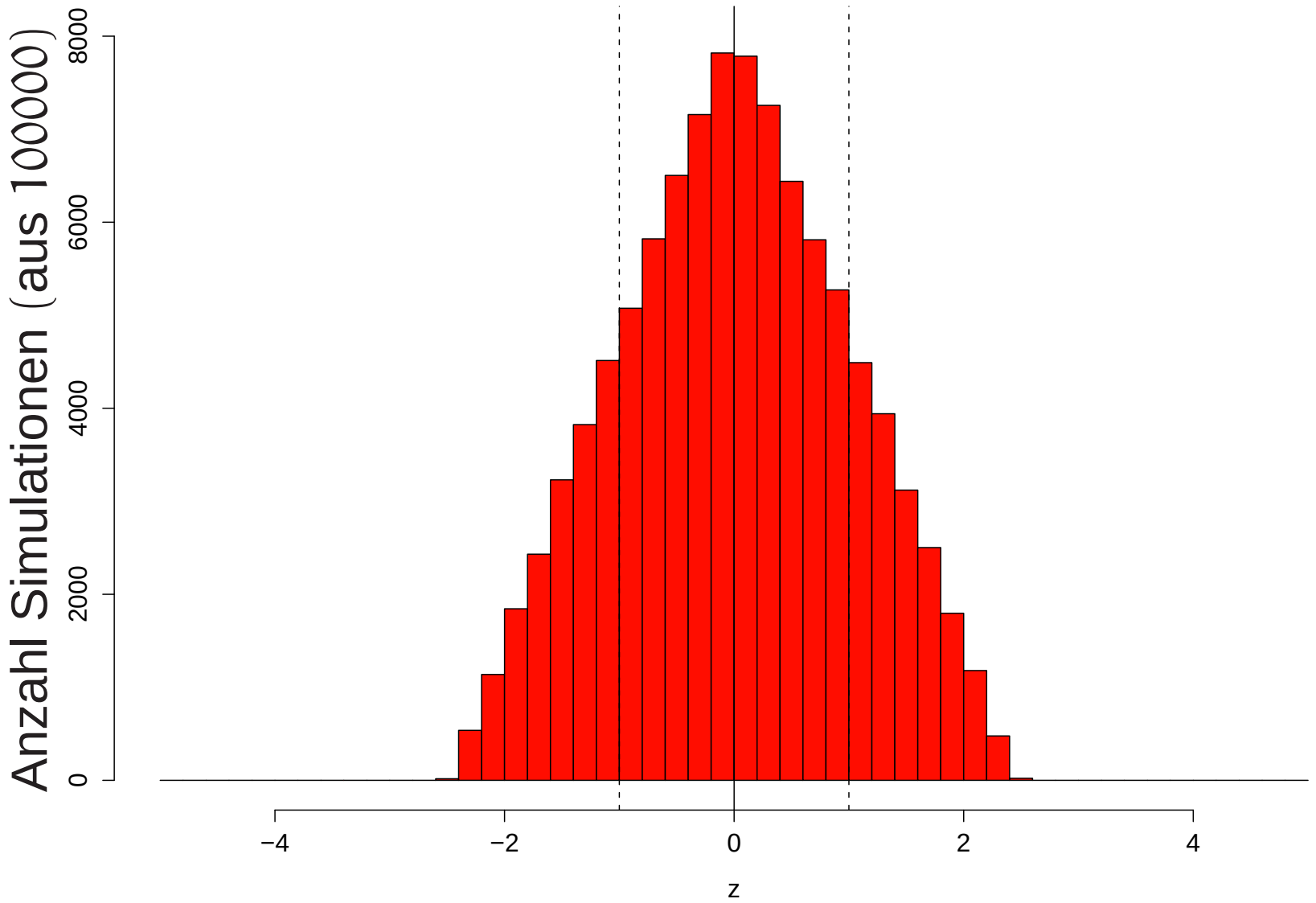
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n}$



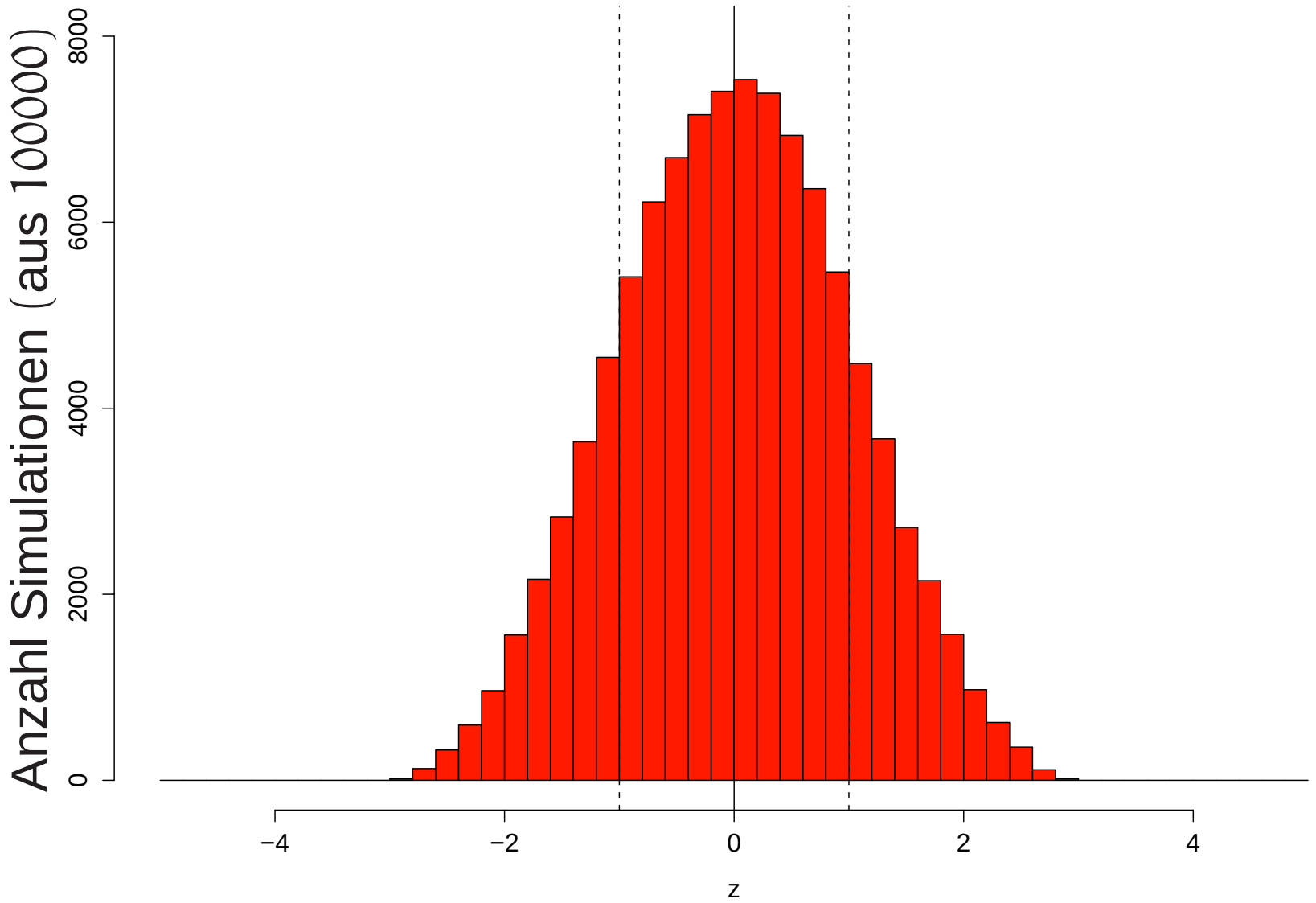
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 1)$



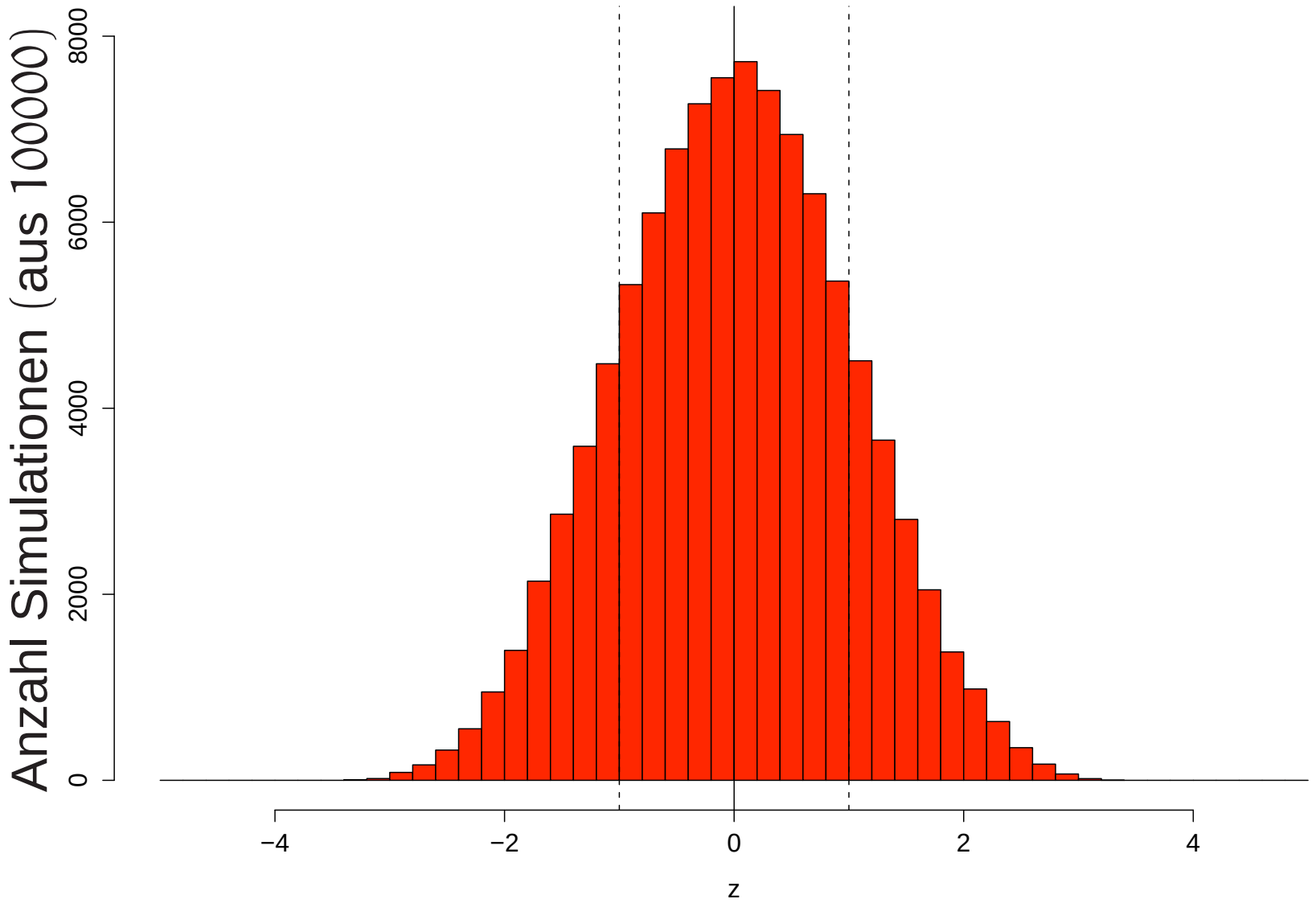
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 2)$



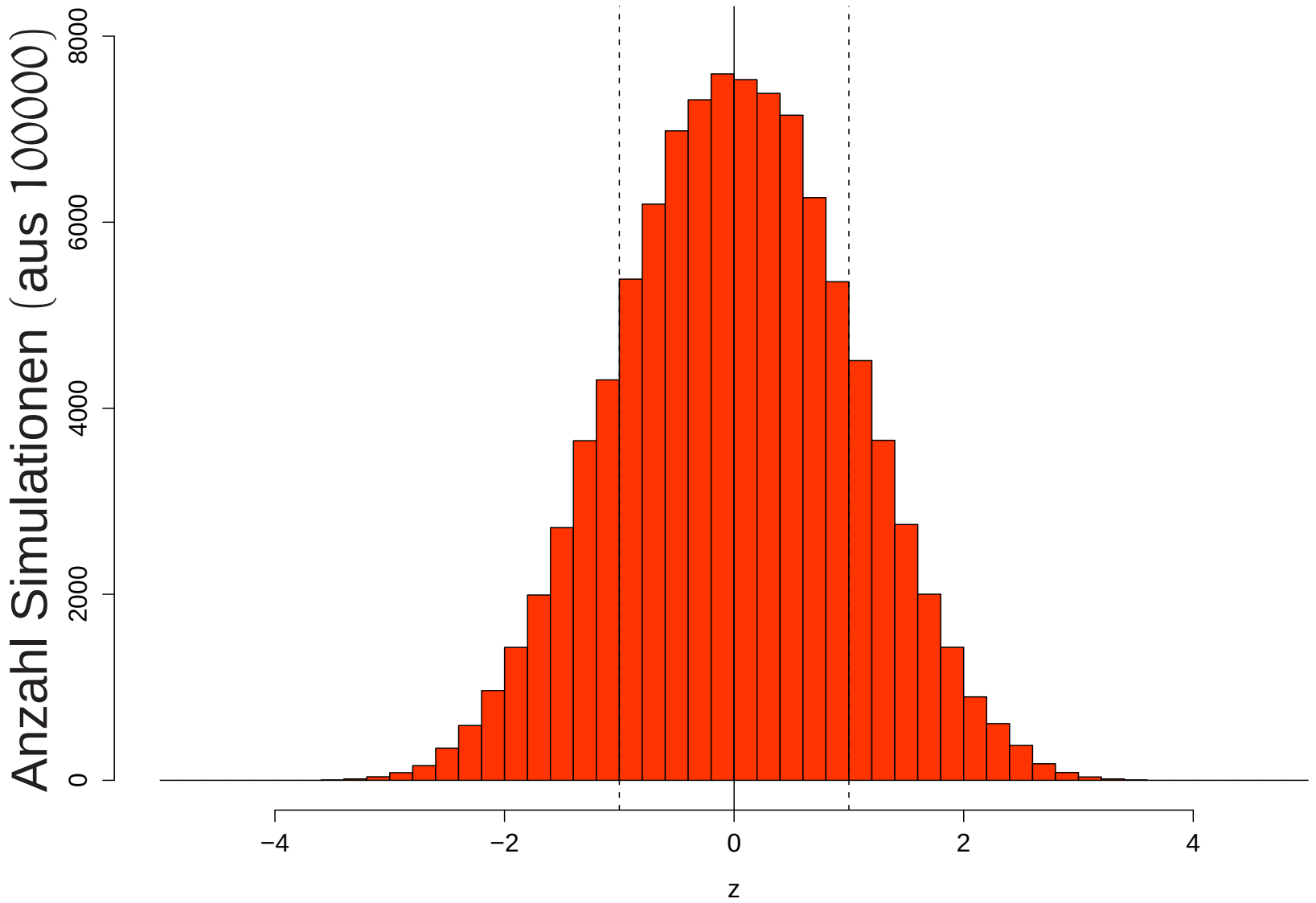
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 3)$



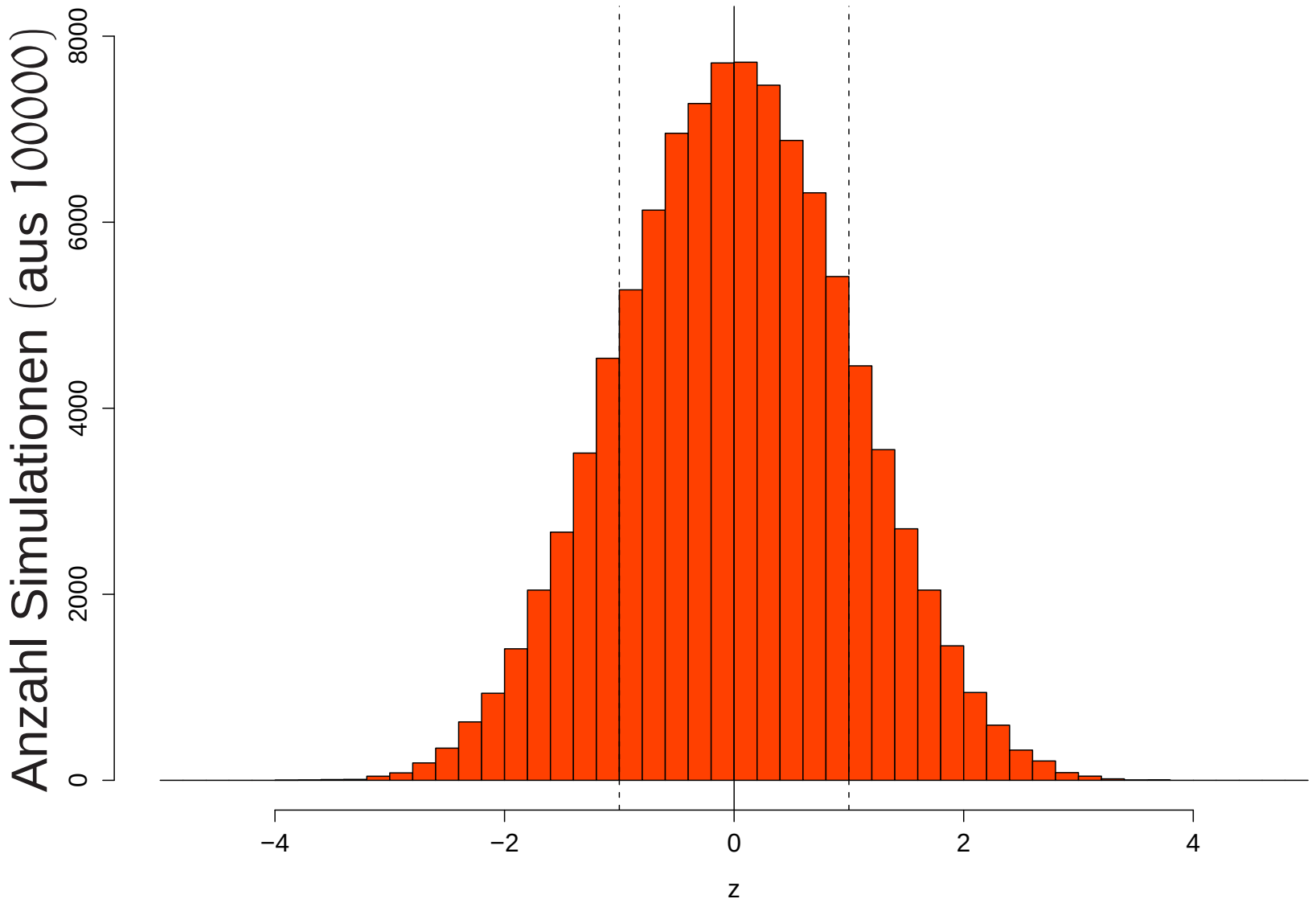
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 4)$



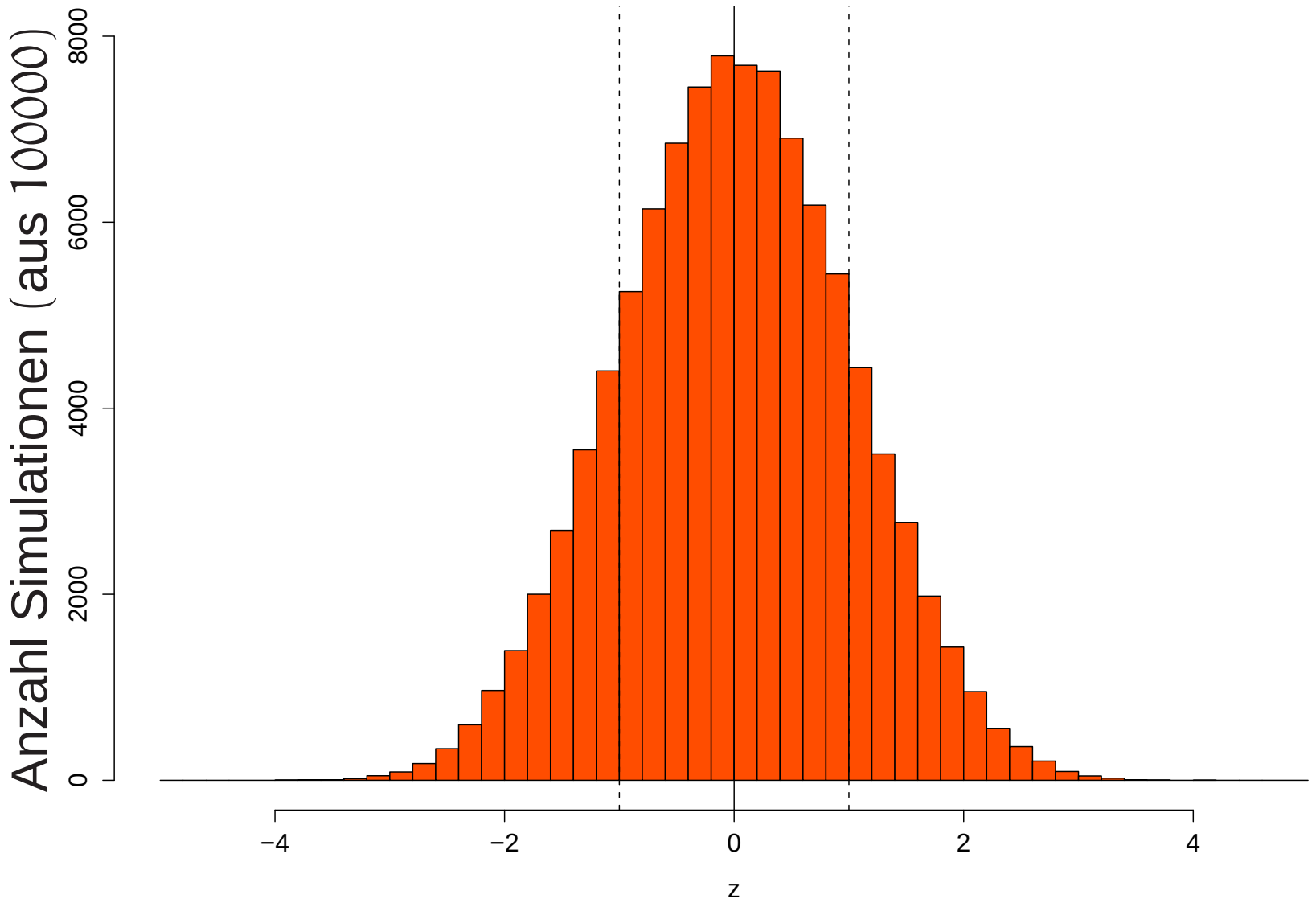
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 5)$



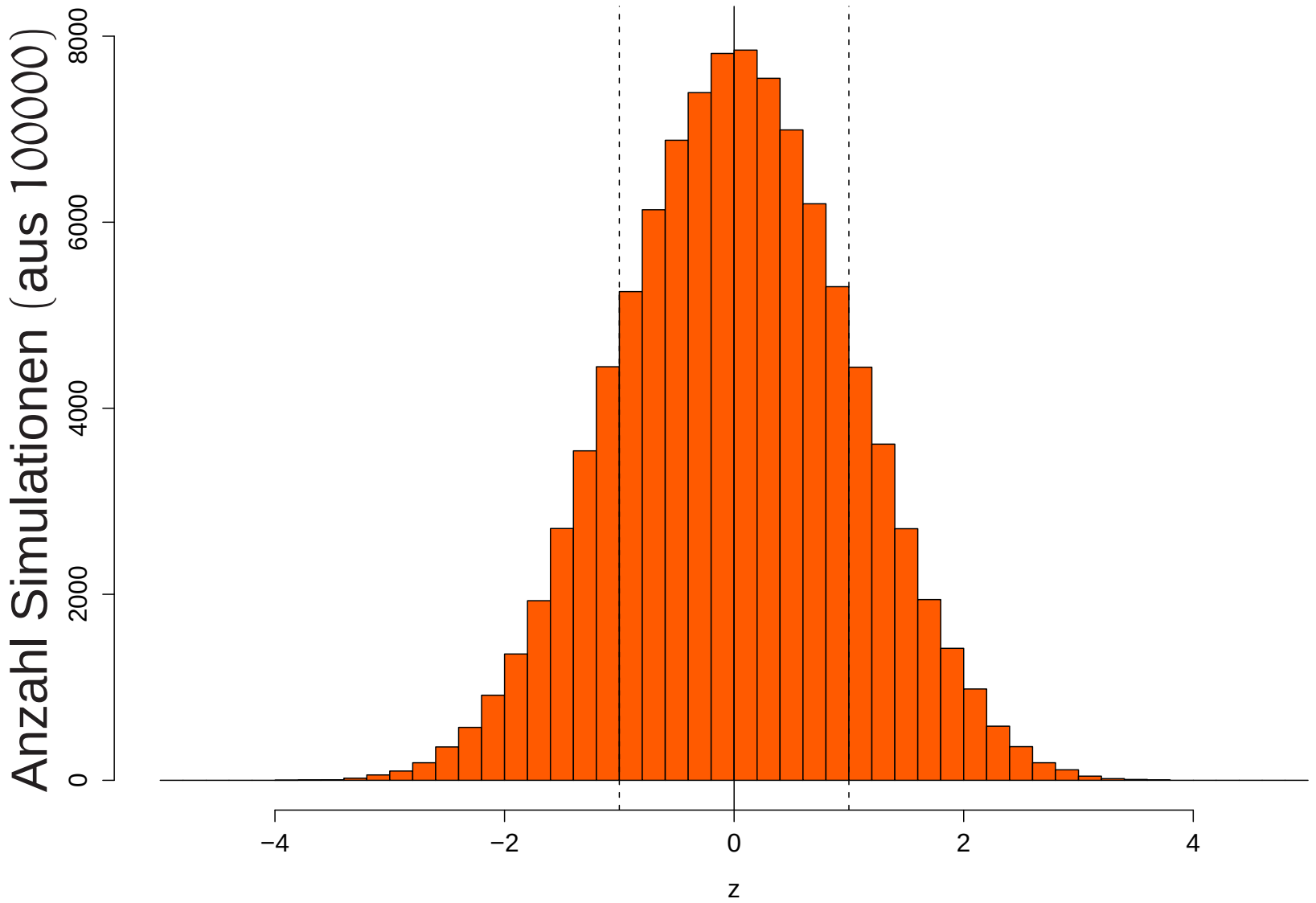
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 6)$



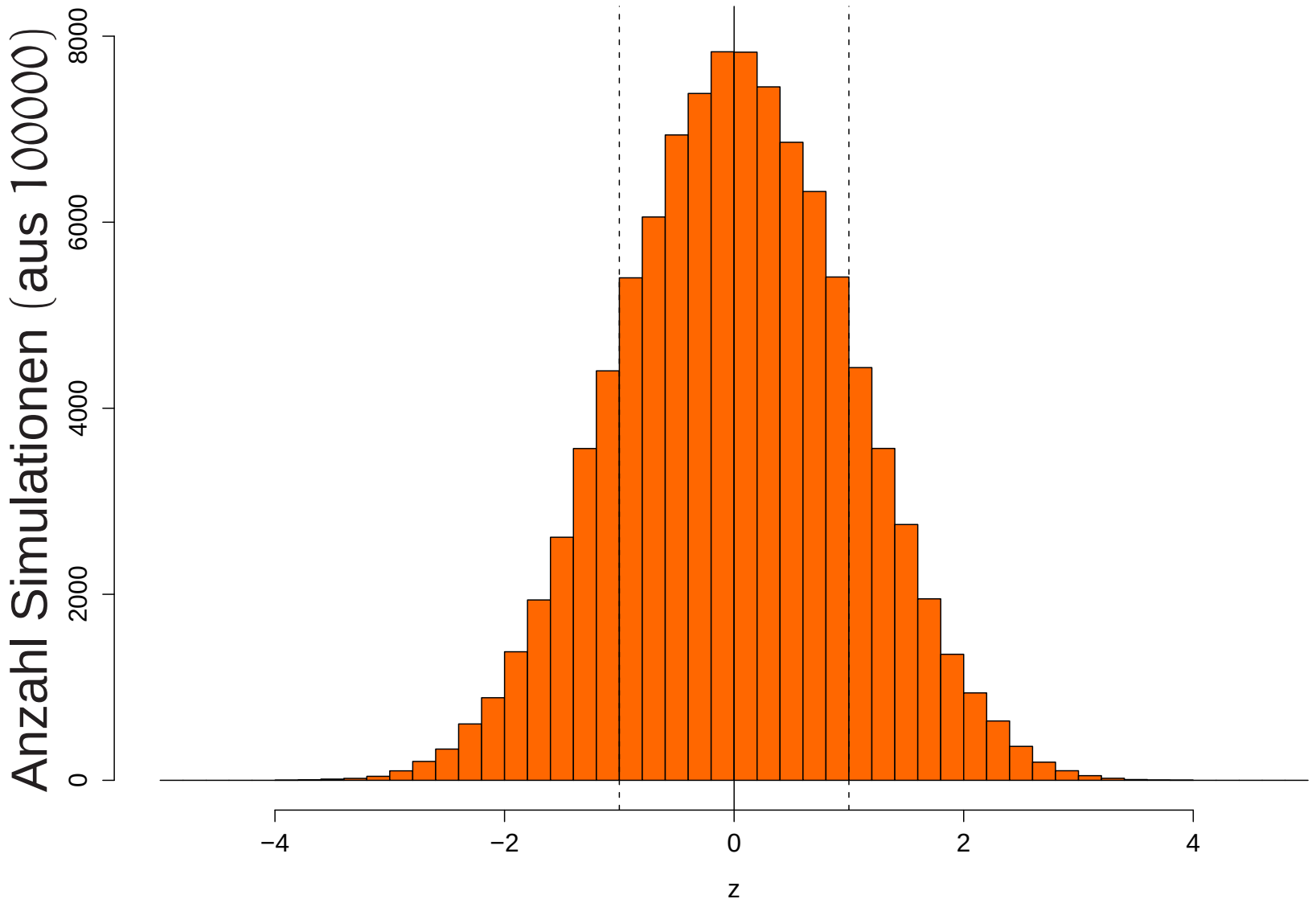
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 7)$



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



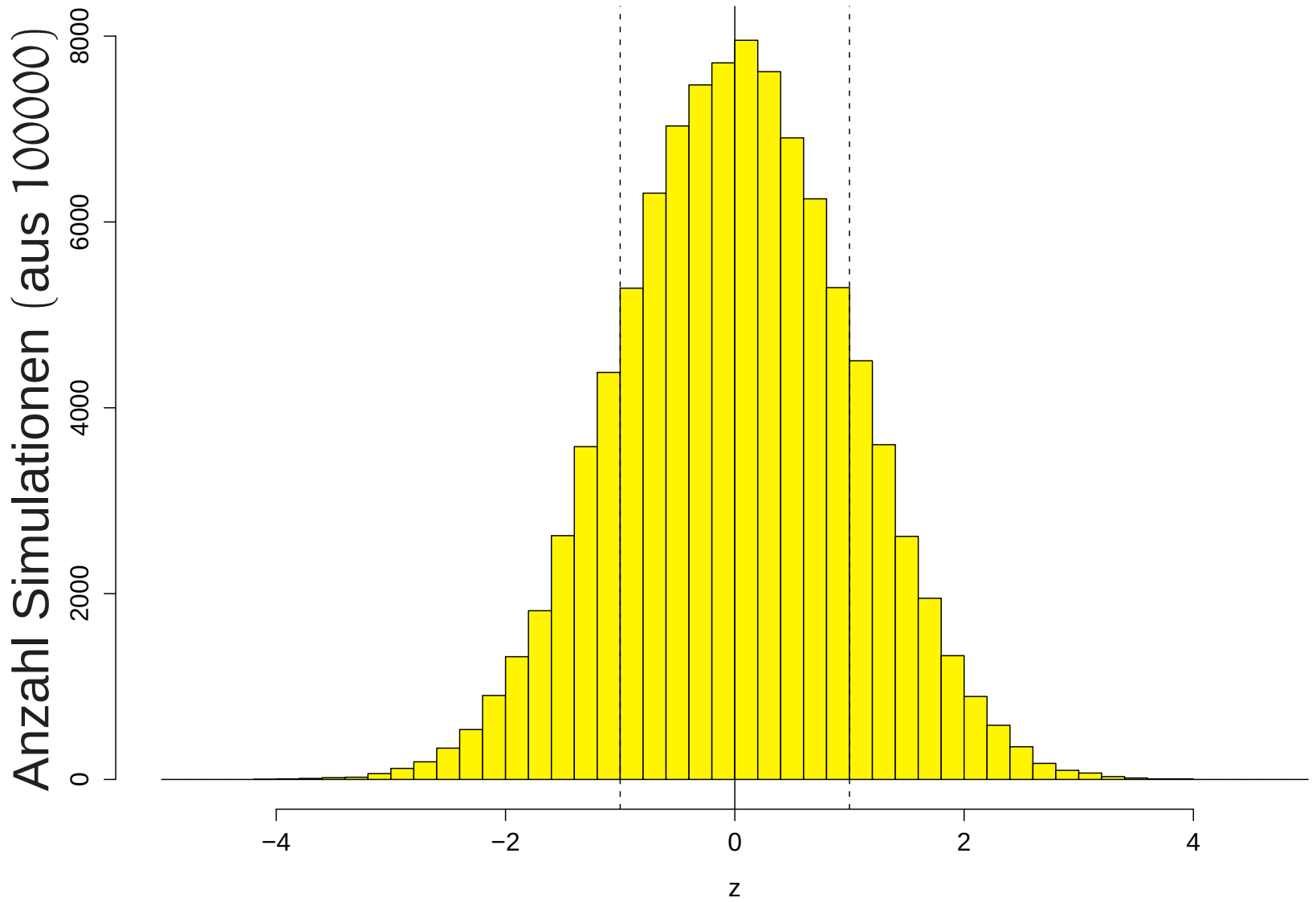
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 9)$



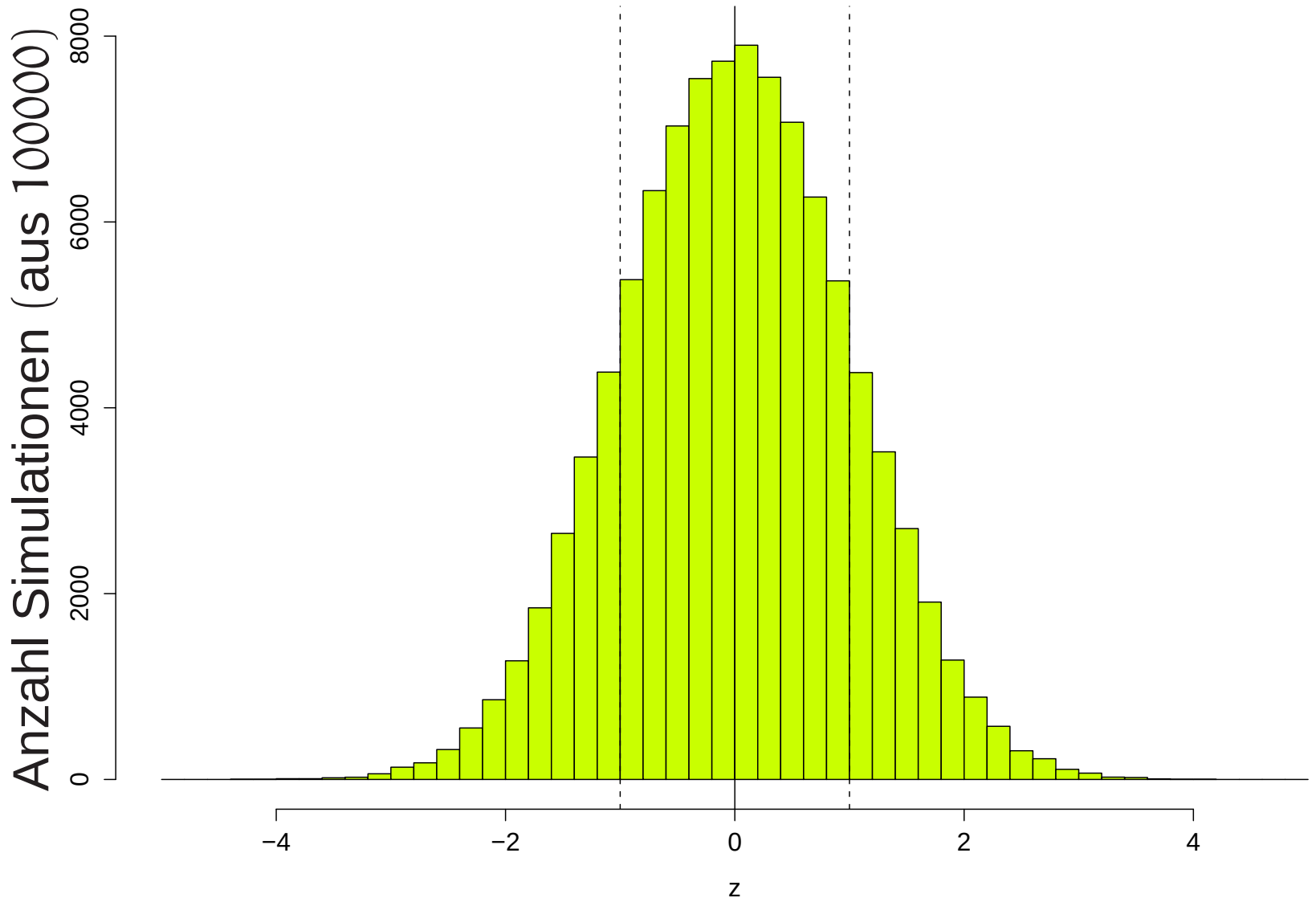
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 10)$

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

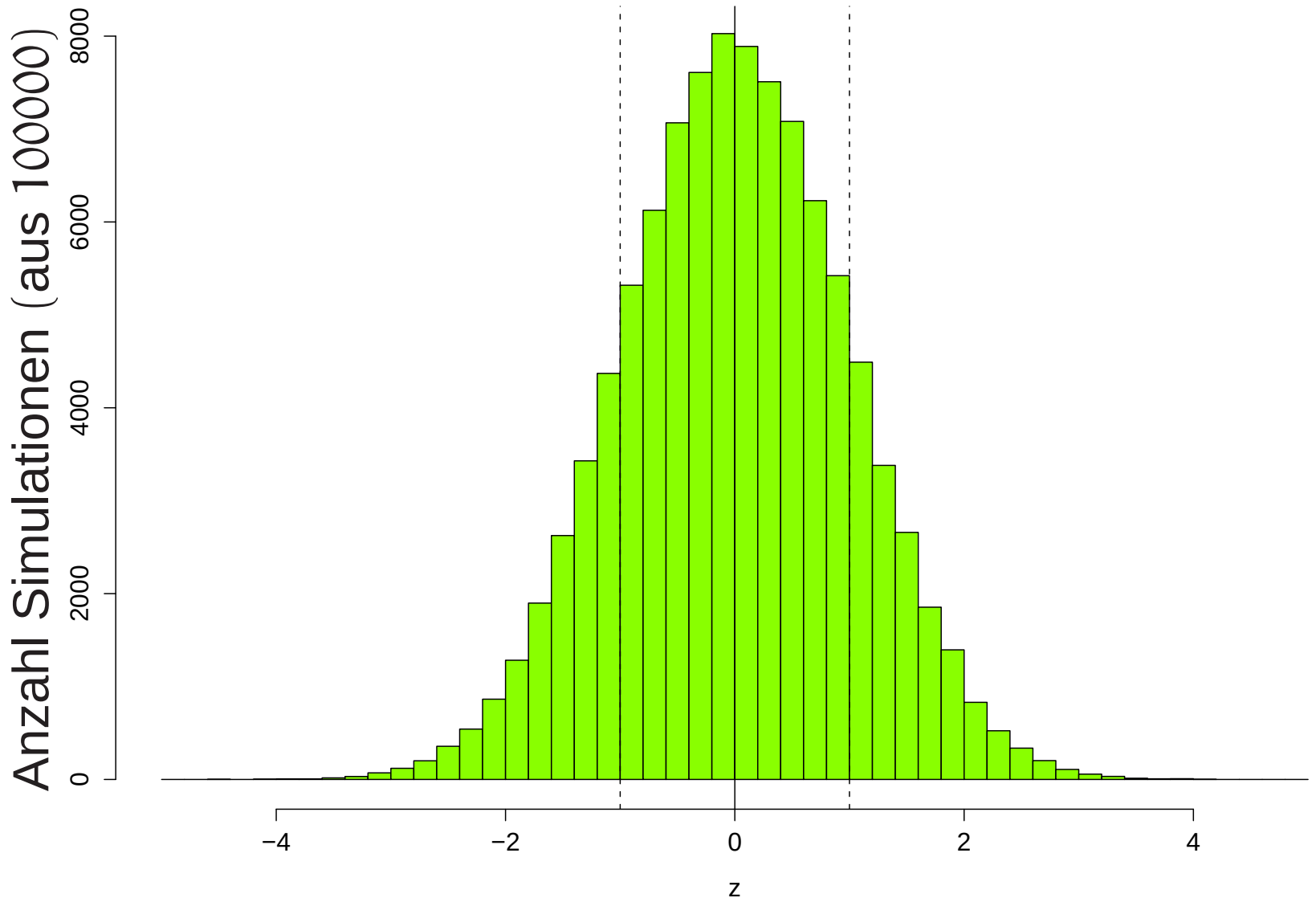
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 20)$



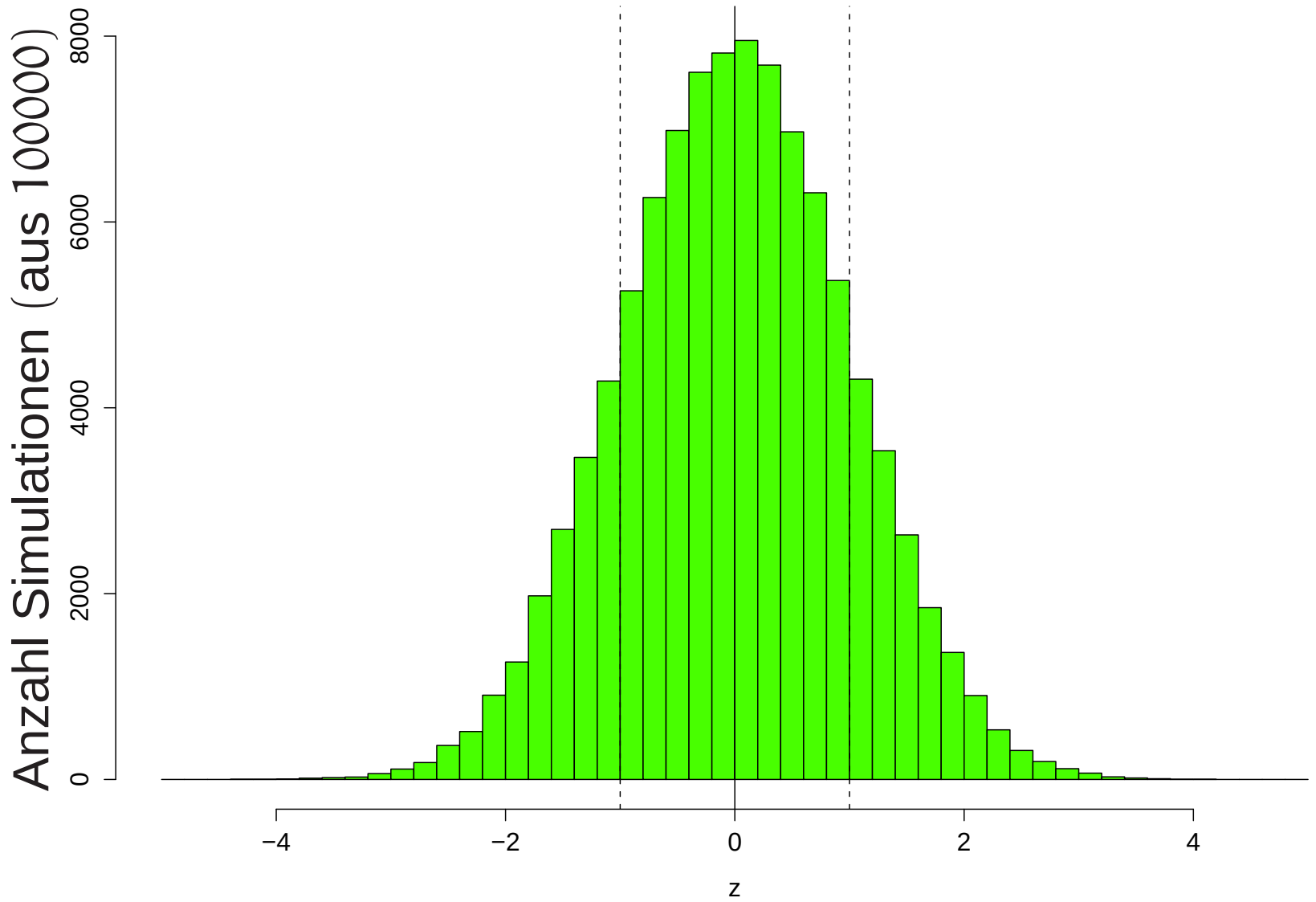
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 25)$



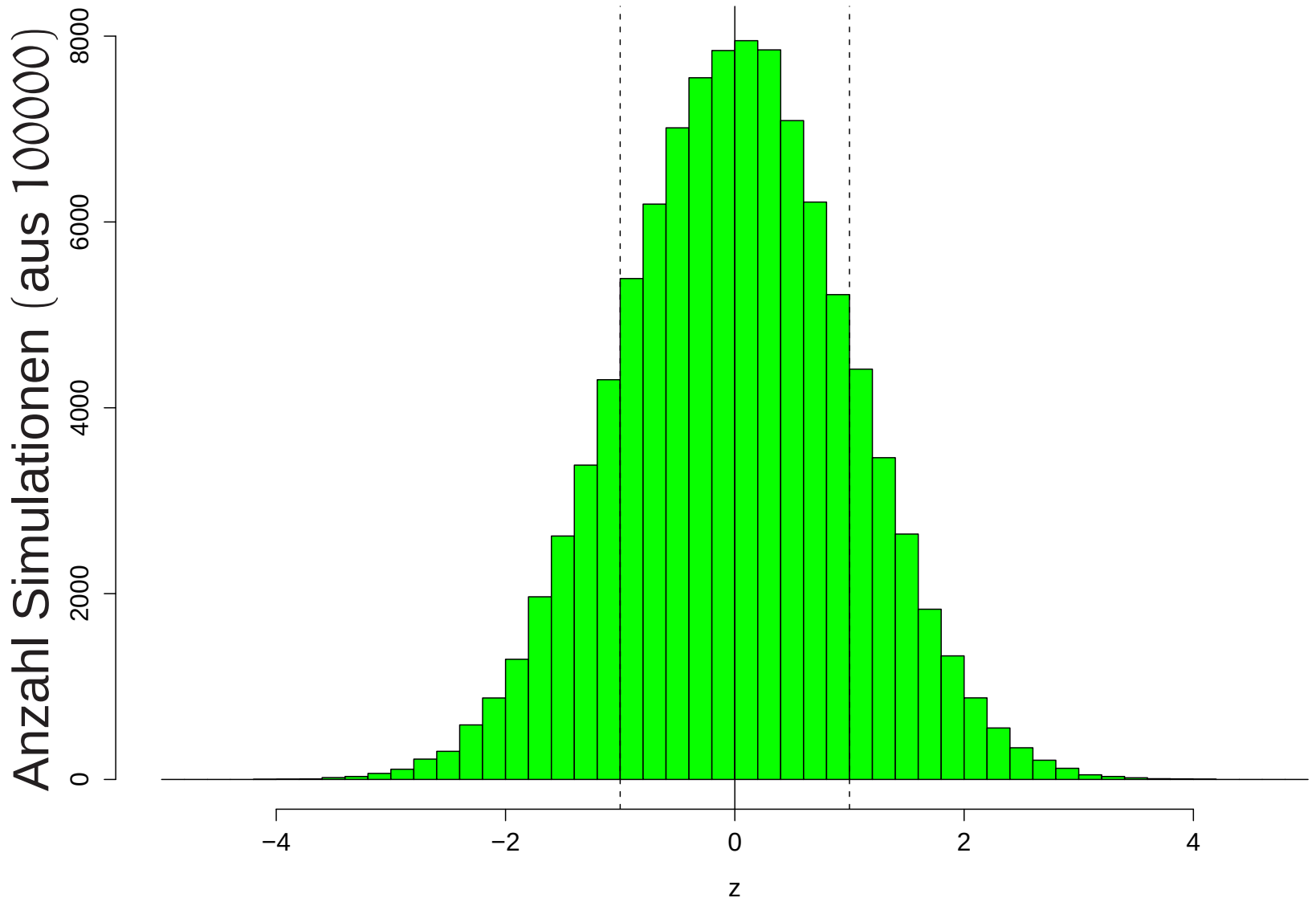
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 30)$



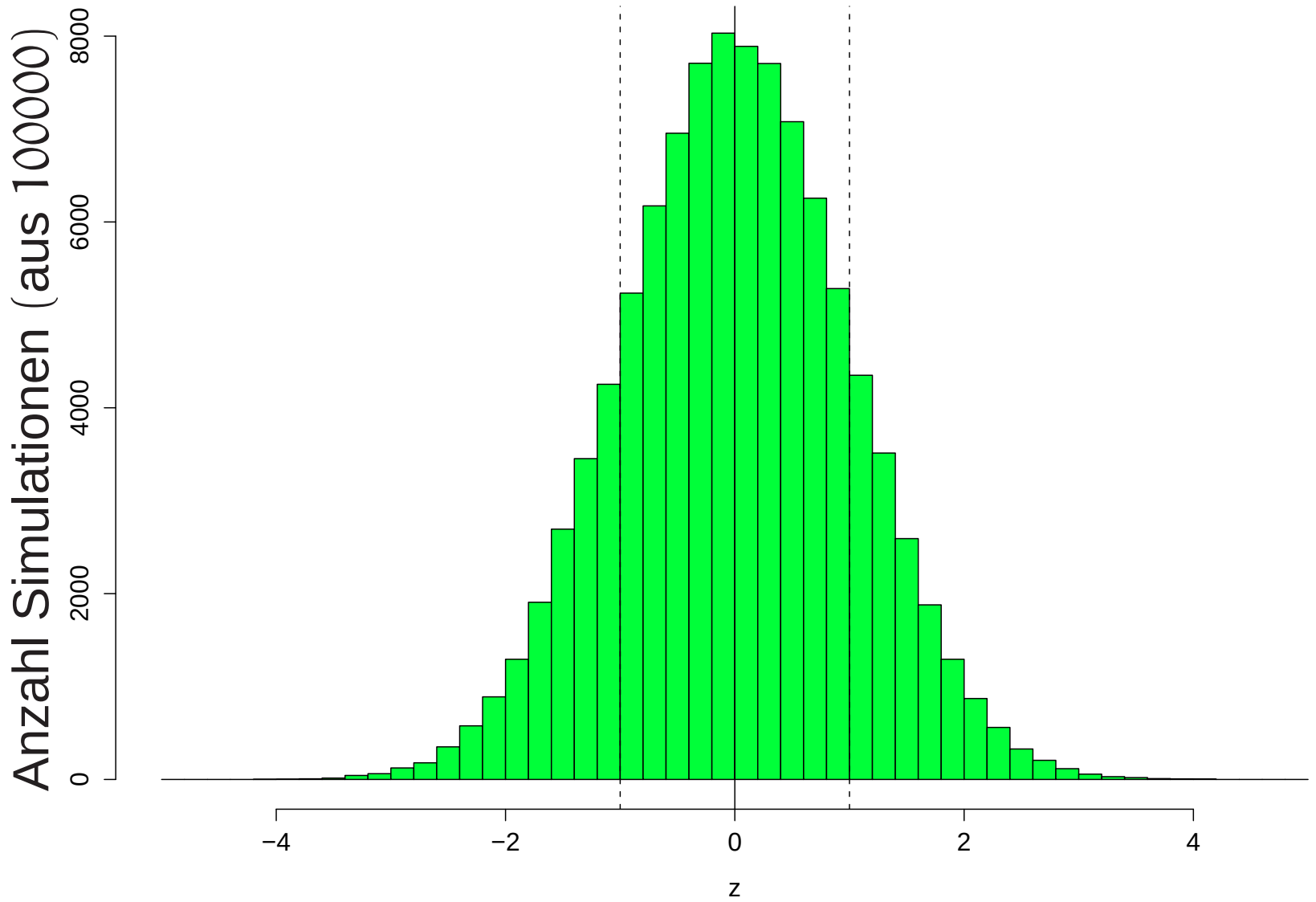
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 35)$



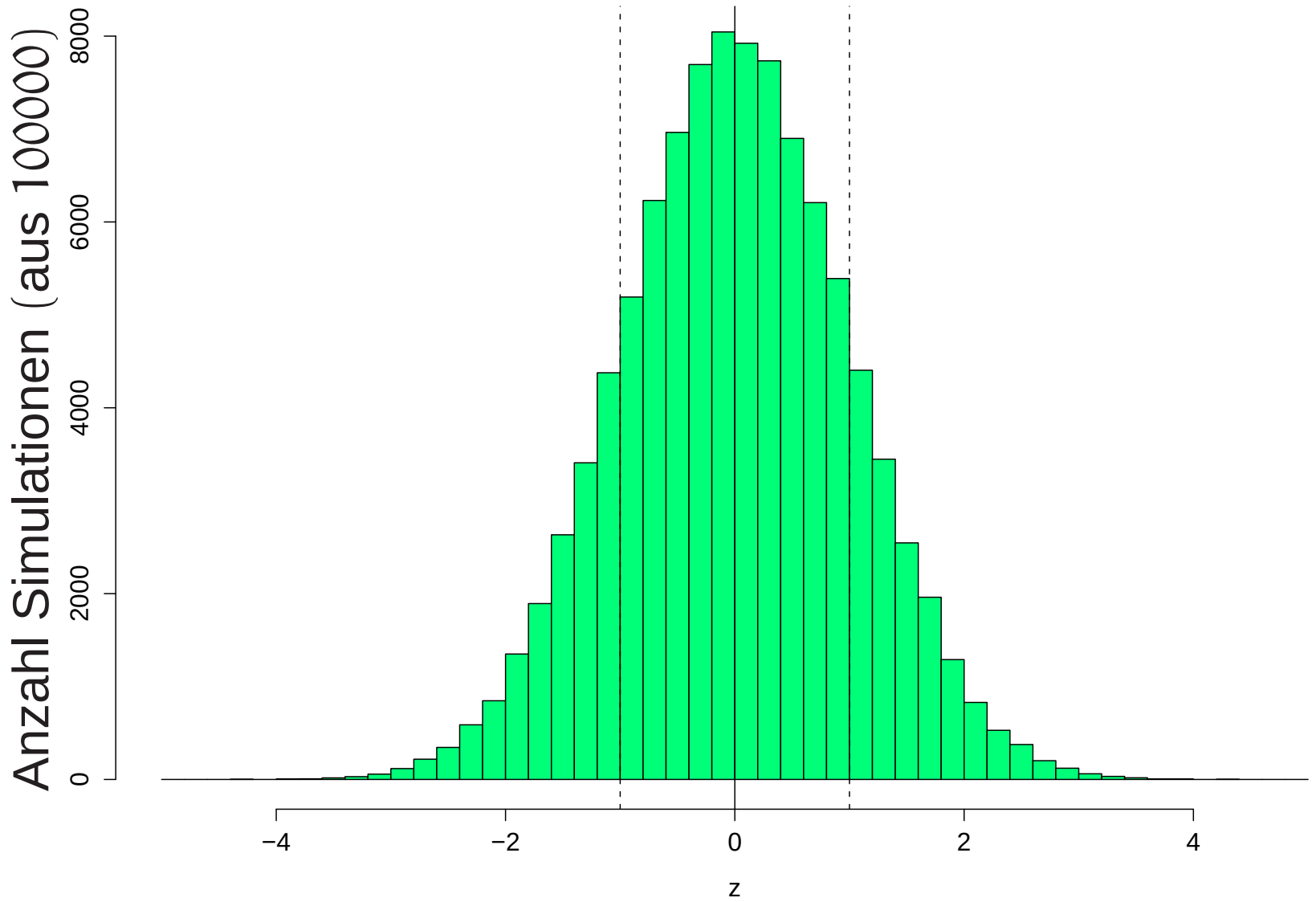
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 40)$



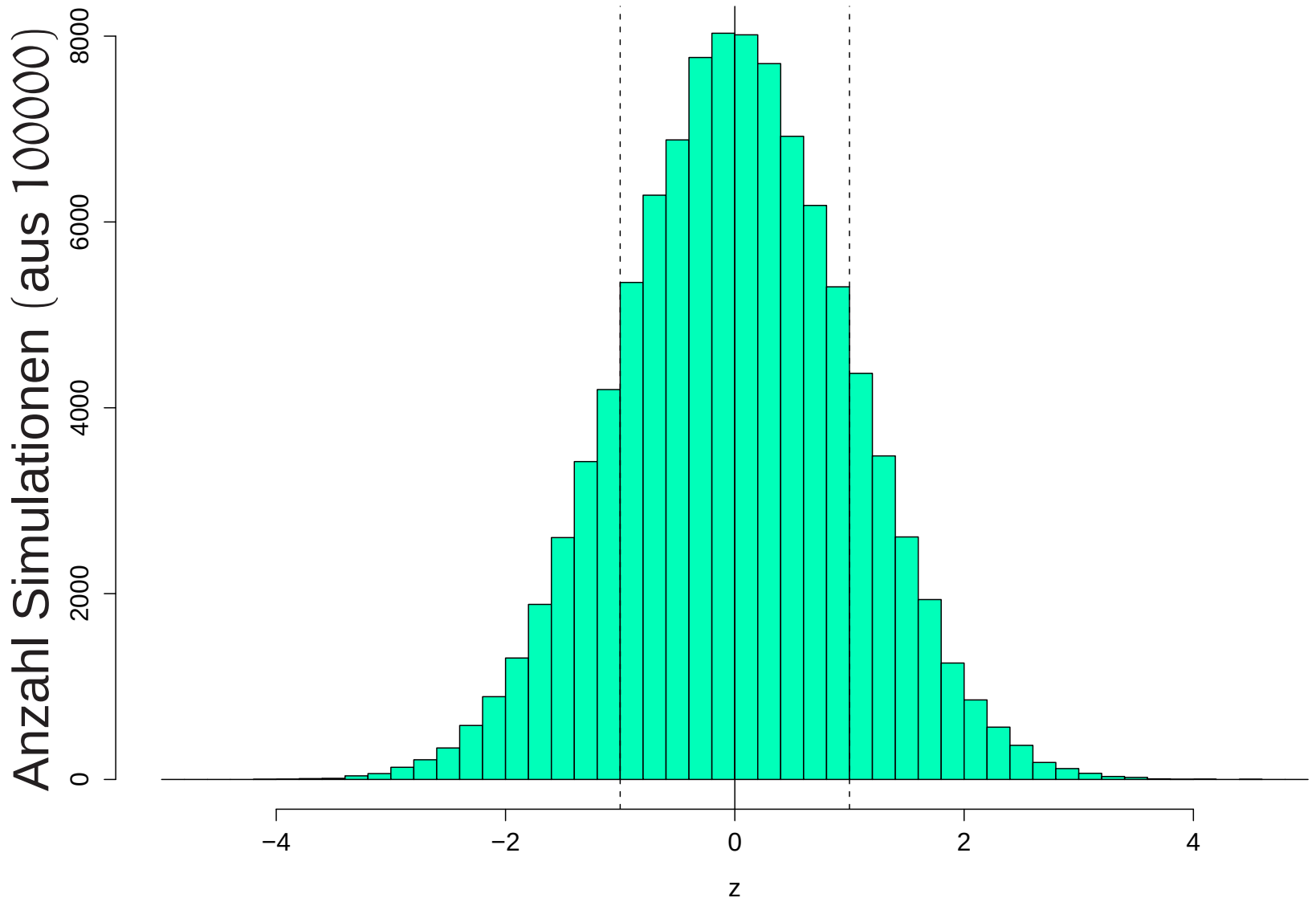
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 45)$



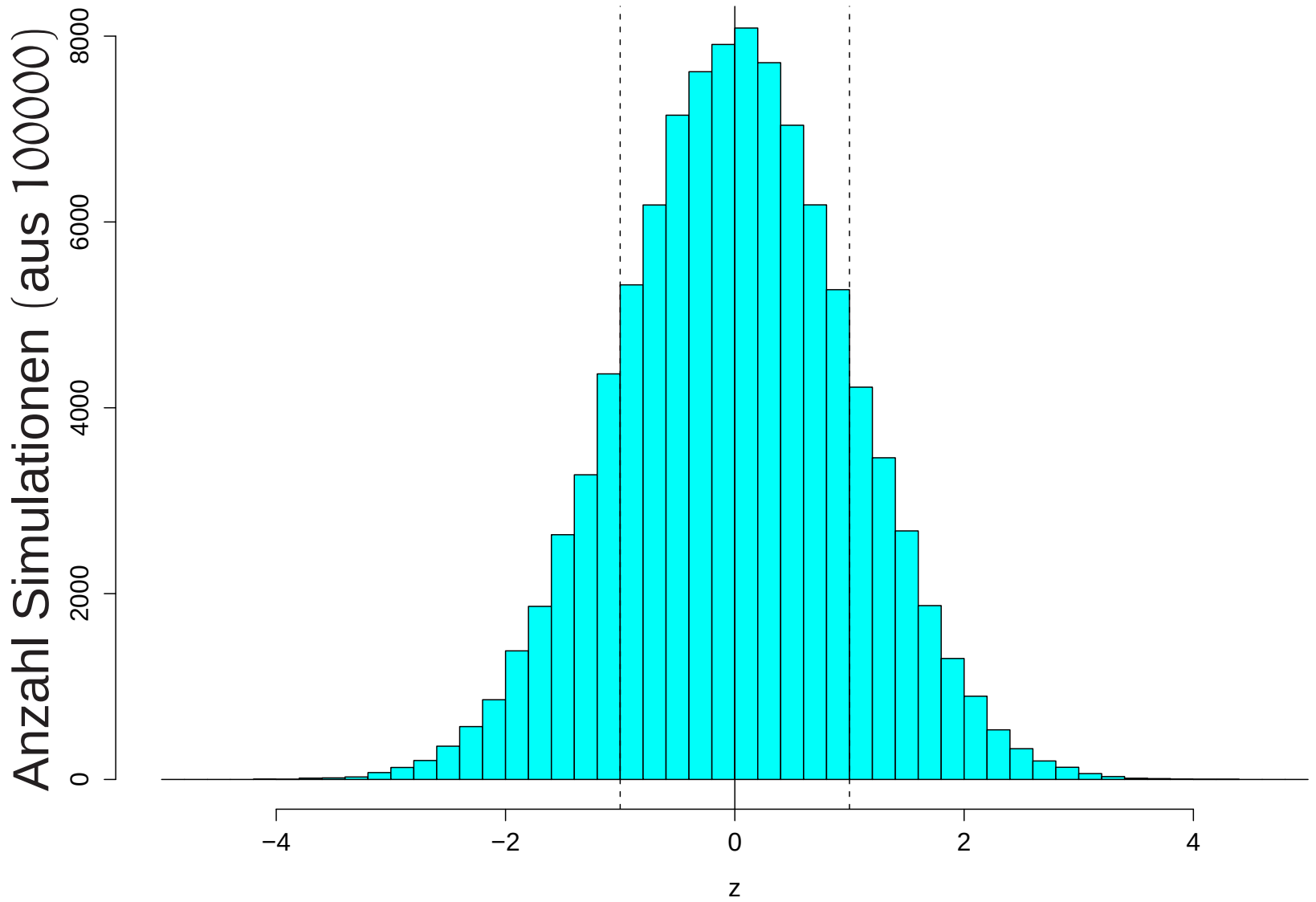
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 50)$



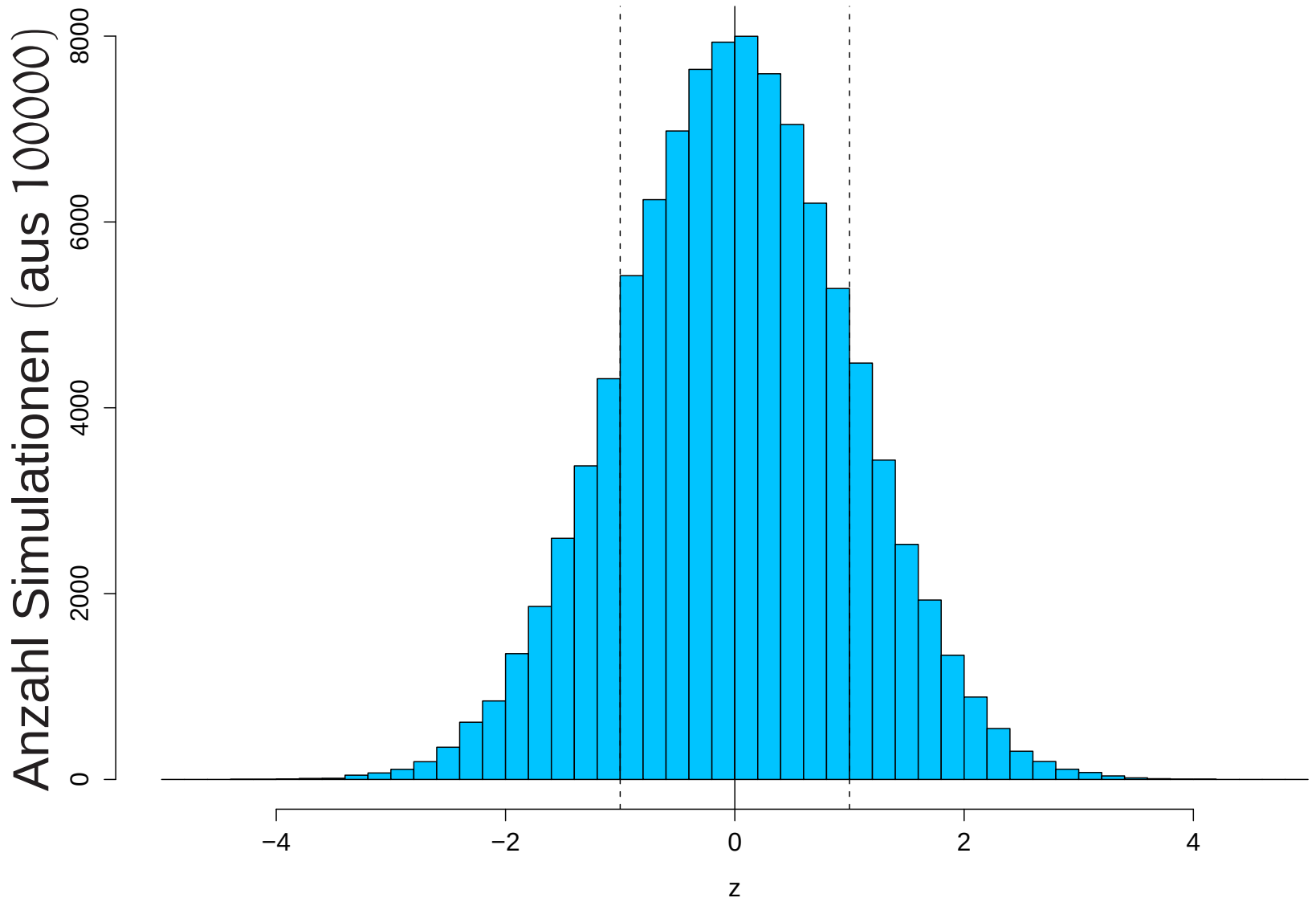
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 55)$



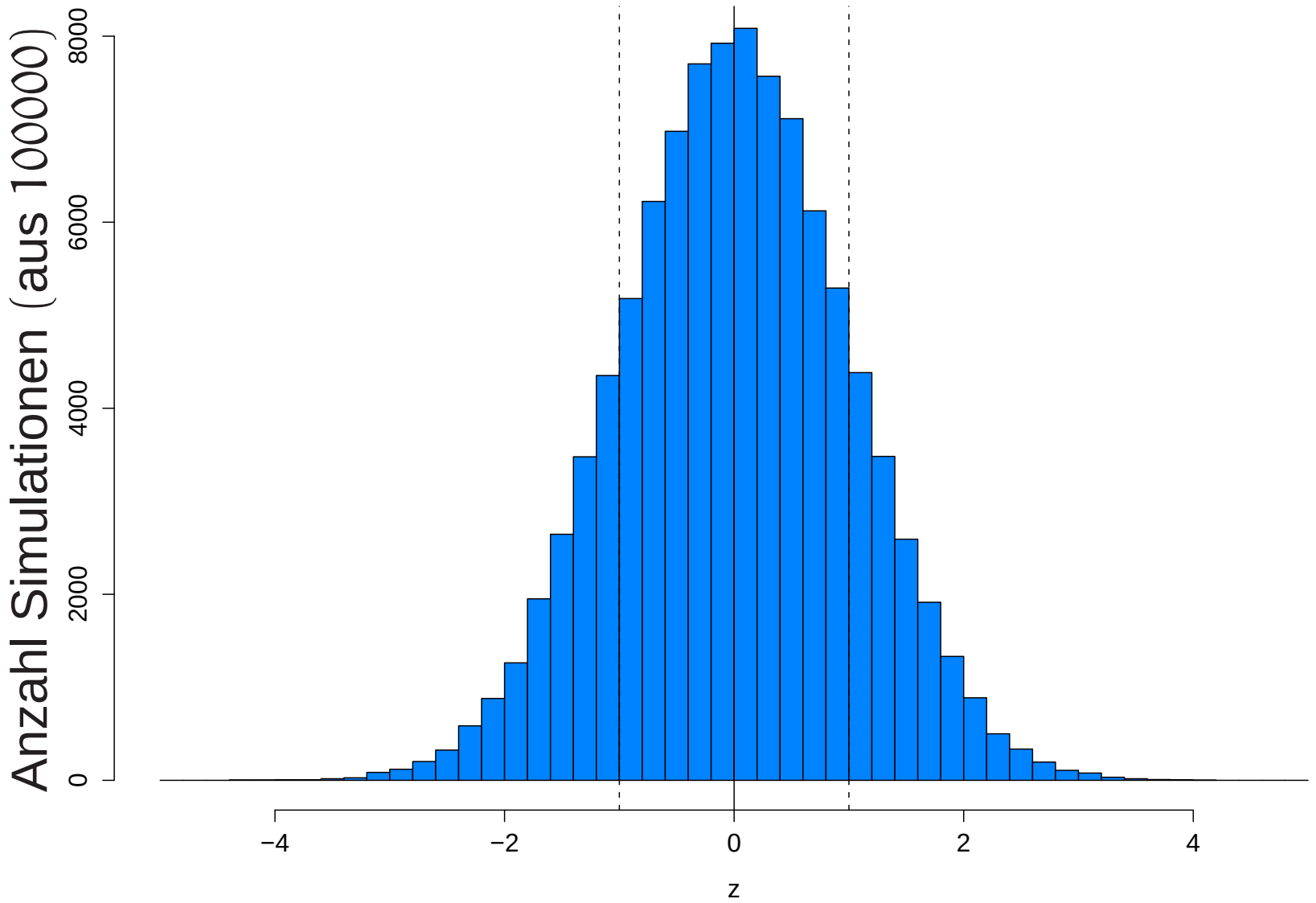
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 60)$



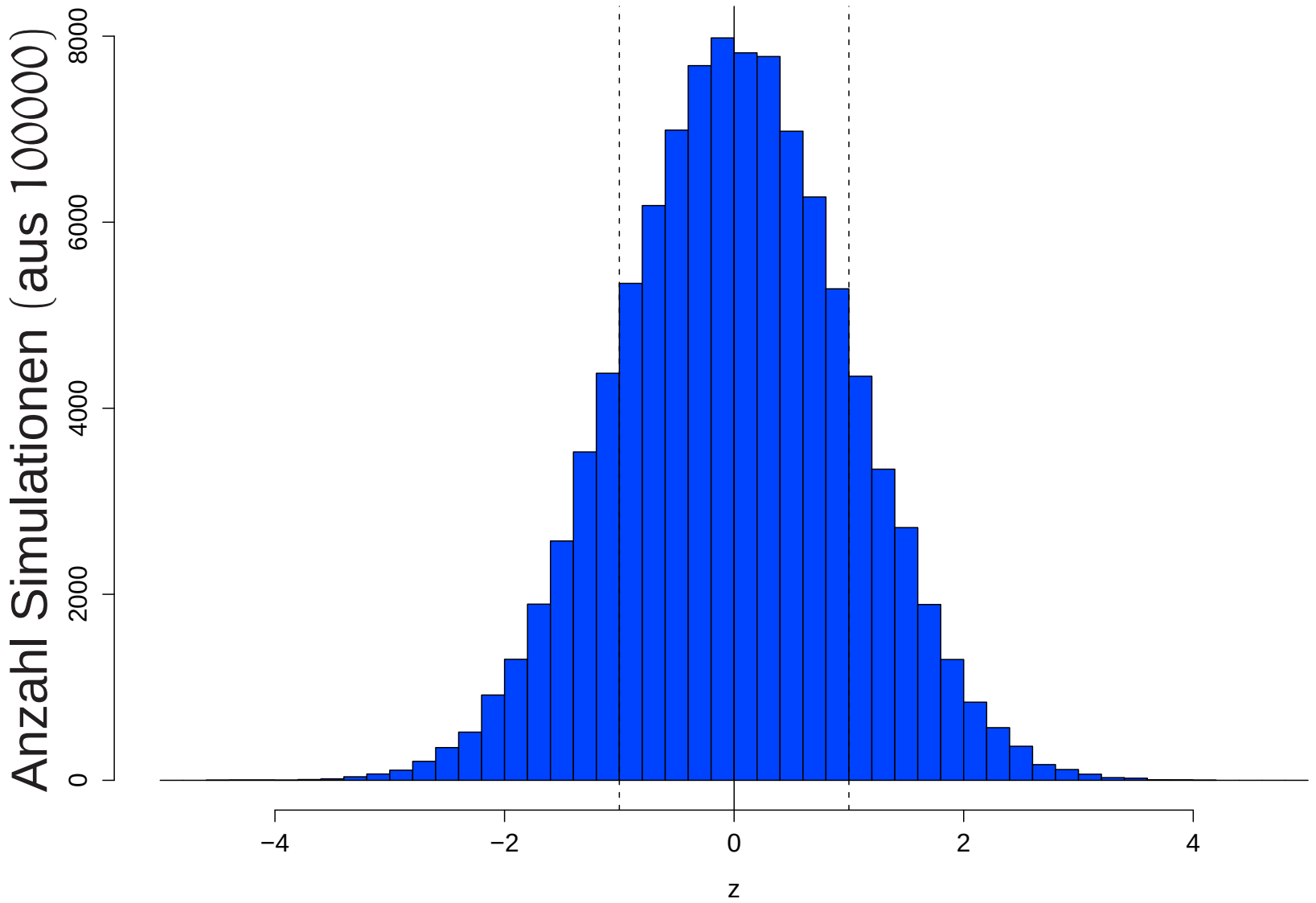
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 65)$



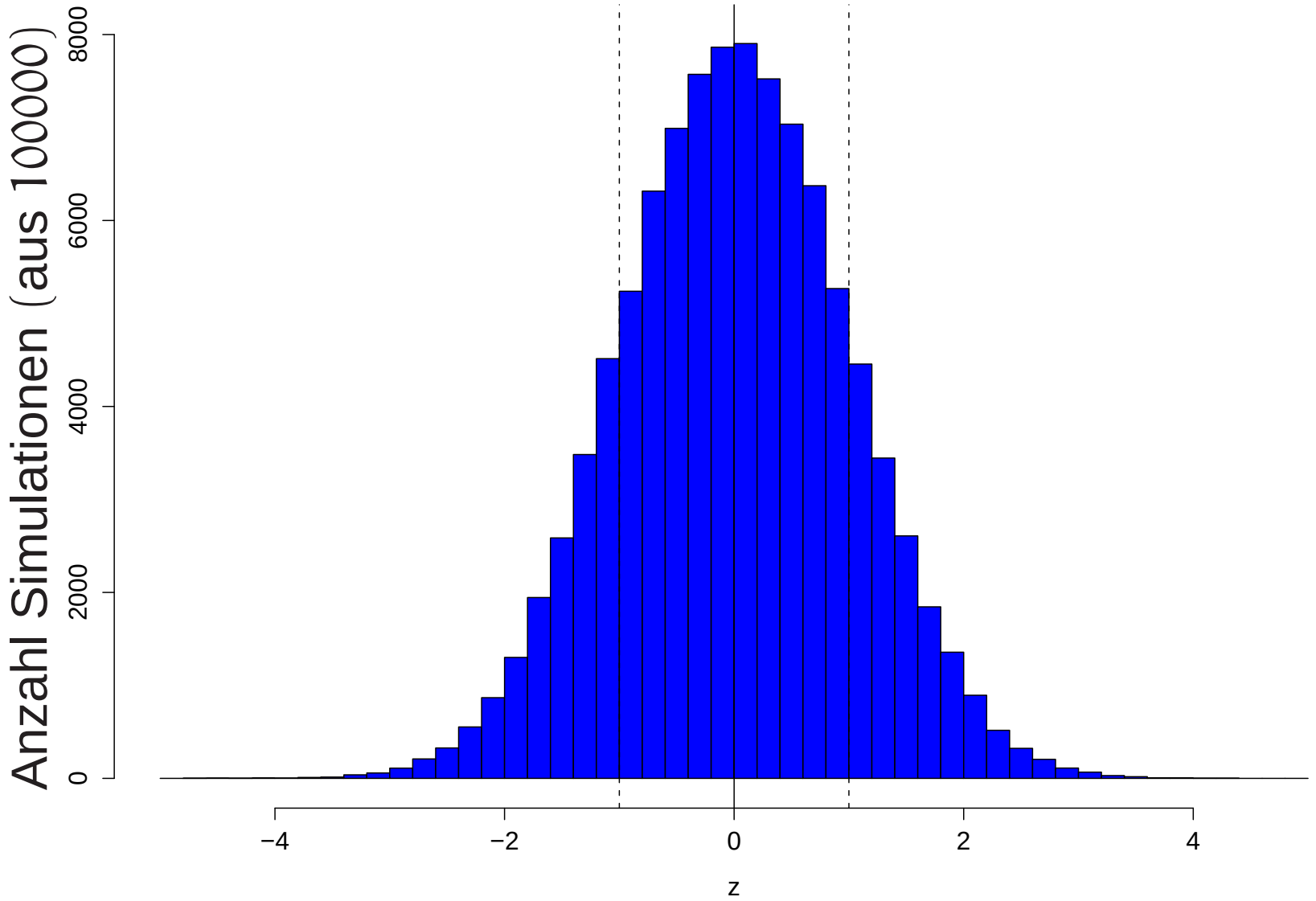
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 70)$



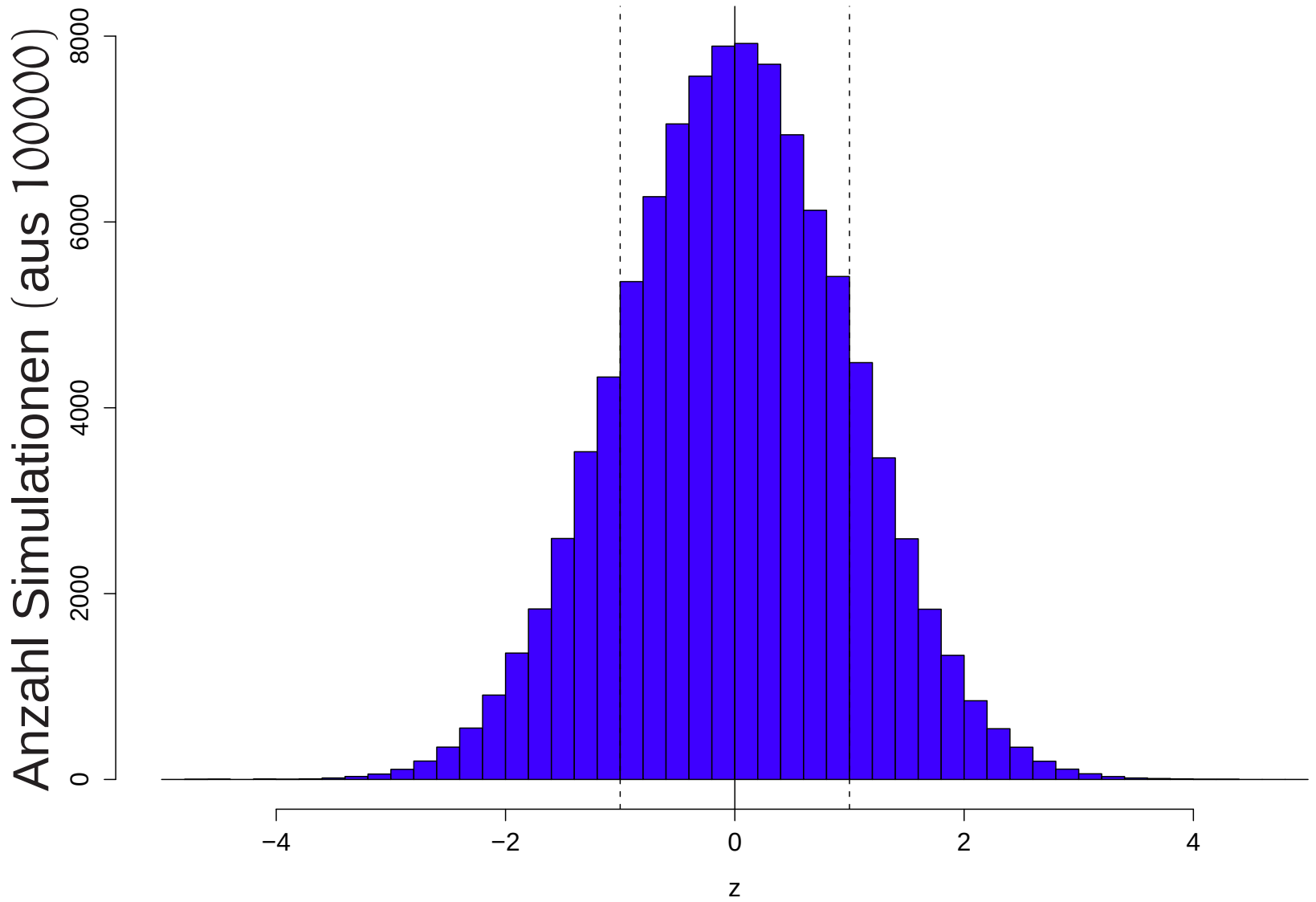
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 75)$



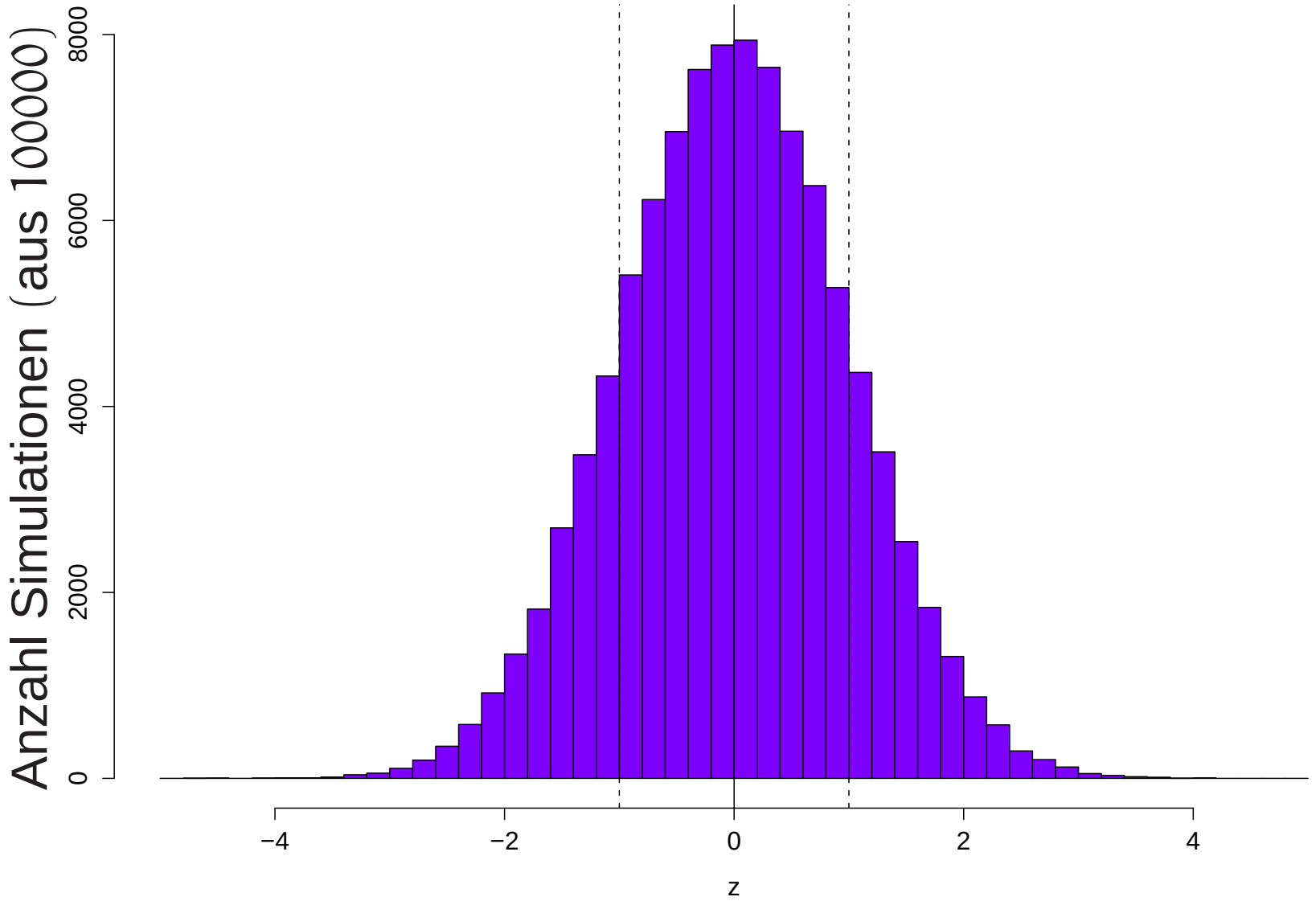
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 80)$



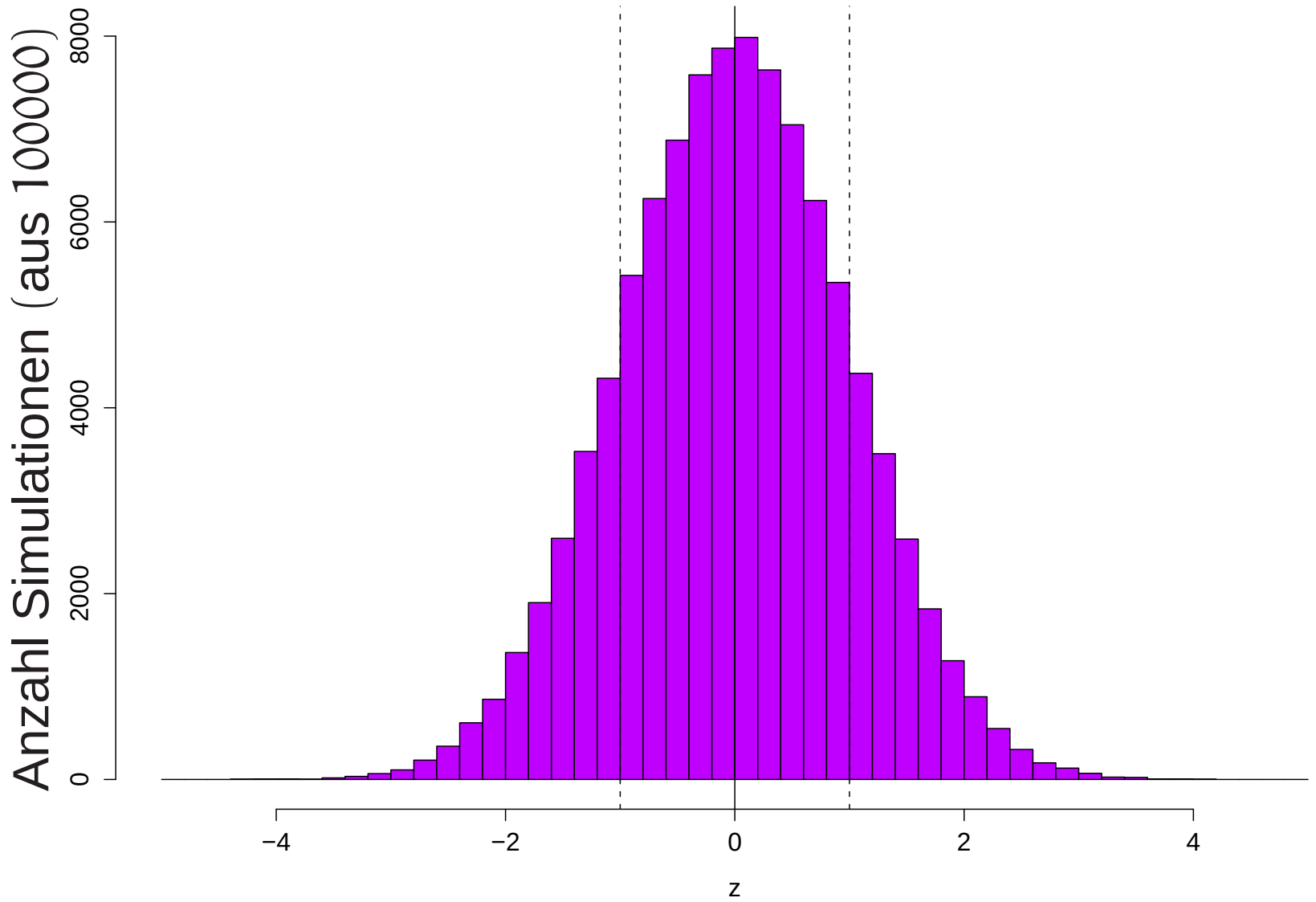
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 85)$



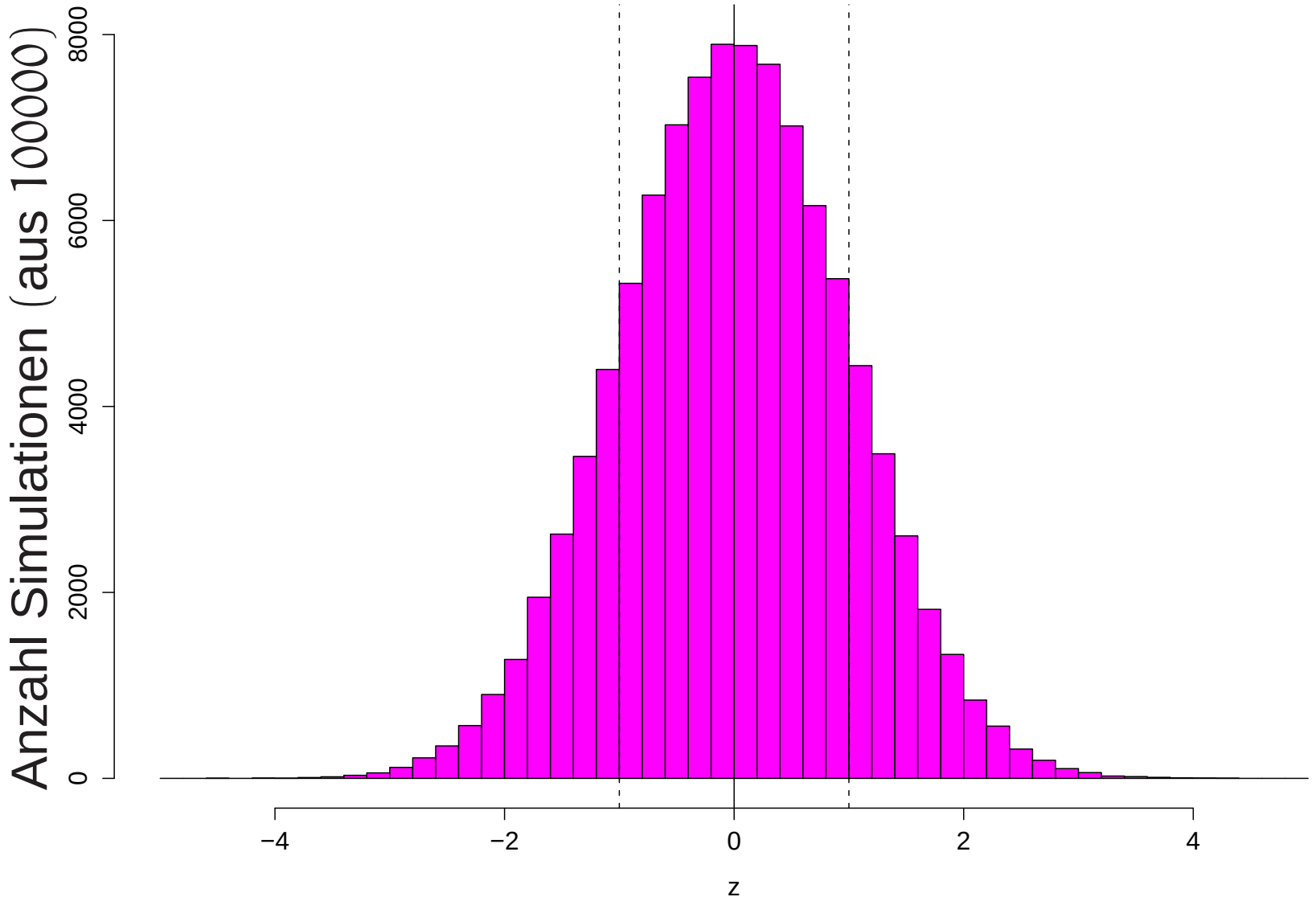
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 90)$



Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 95)$



Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 100)$



Die Verteilung von Z_n

scheint zu konvergieren.

Welche Form

hat die Grenzverteilung?

Die Verteilung von

Z_{100}

ist glockenförmig.

Welche Glocke?

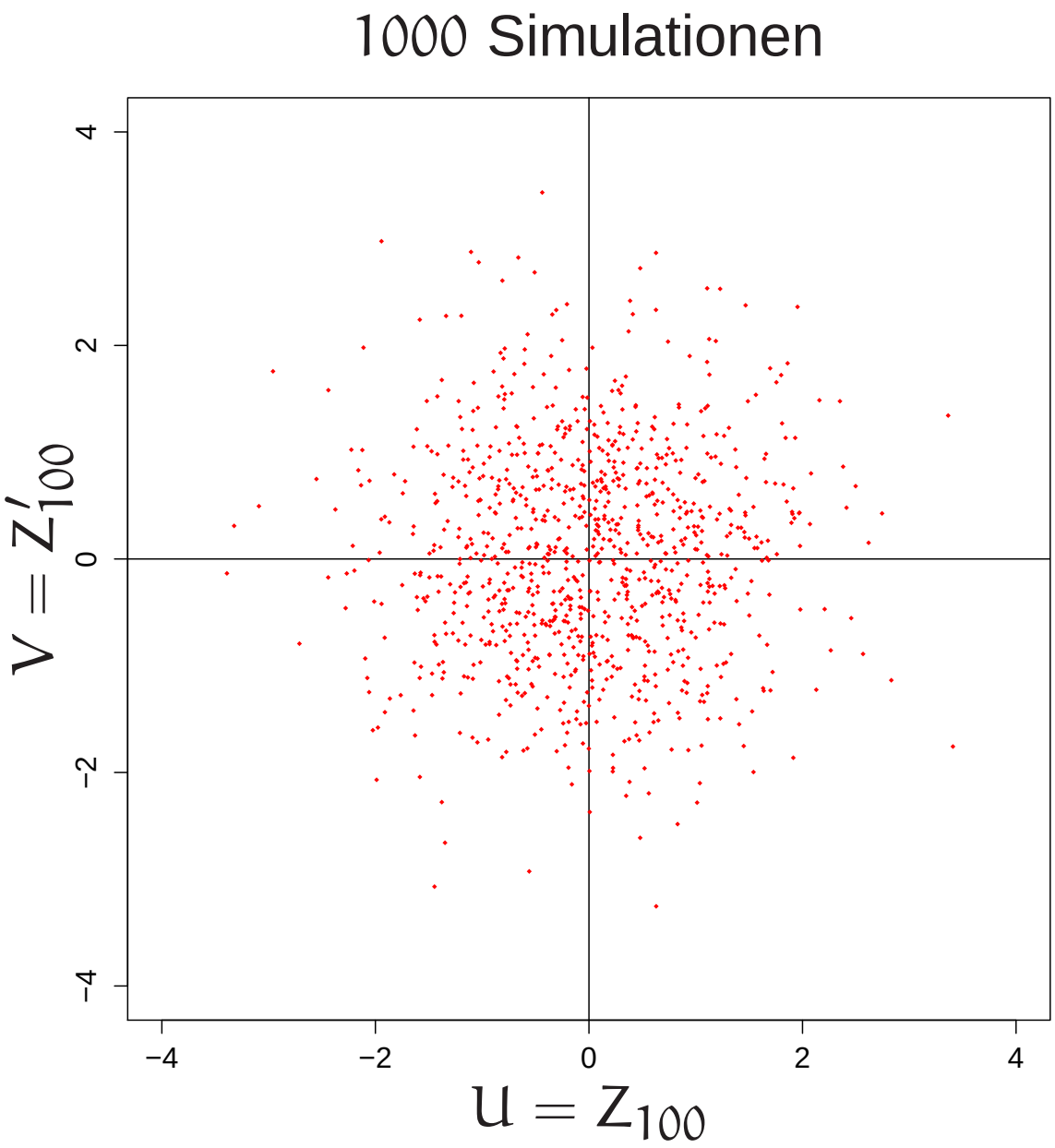
Glücklicher Einfall:

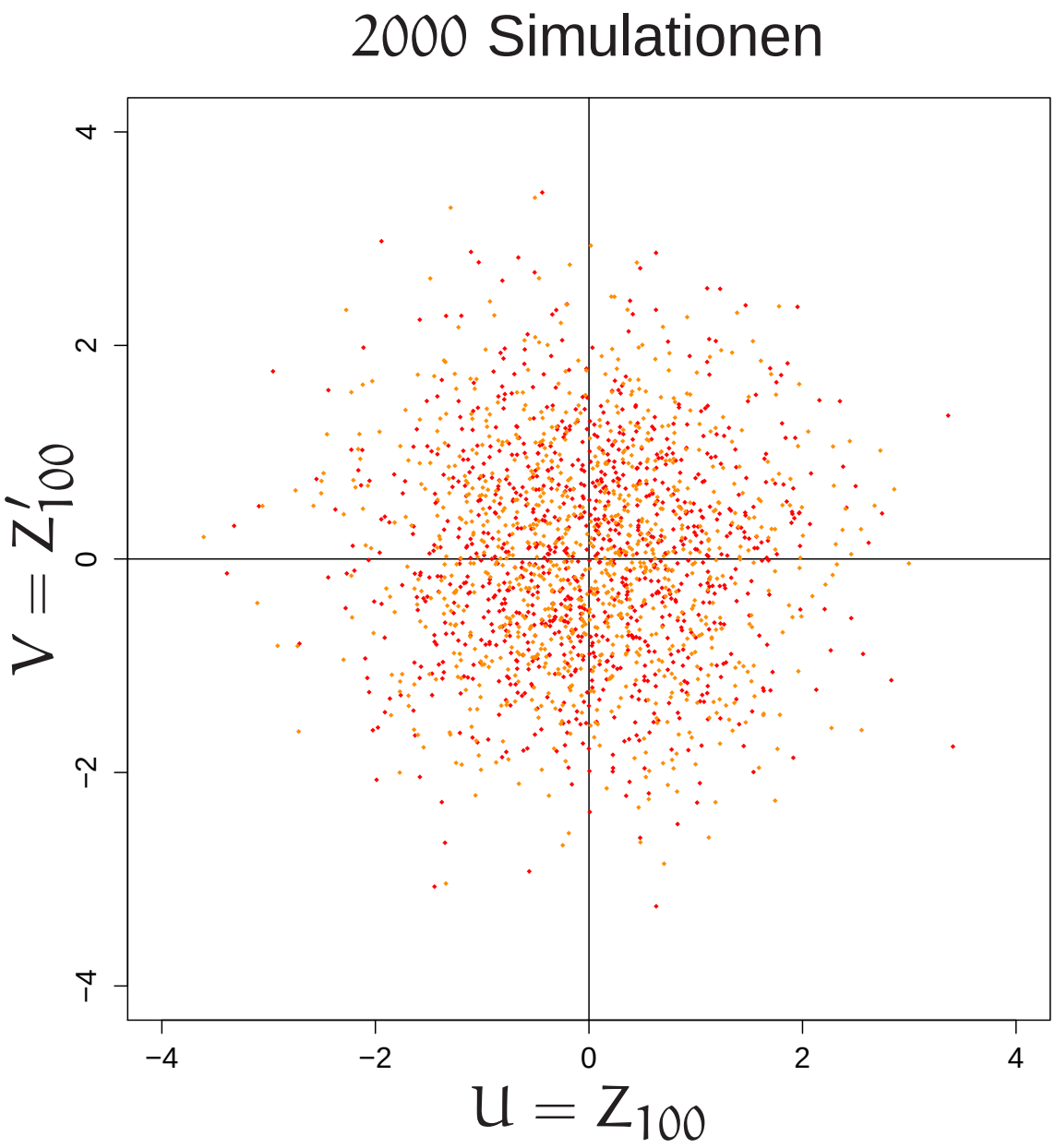
Zwei unabhängige Kopien

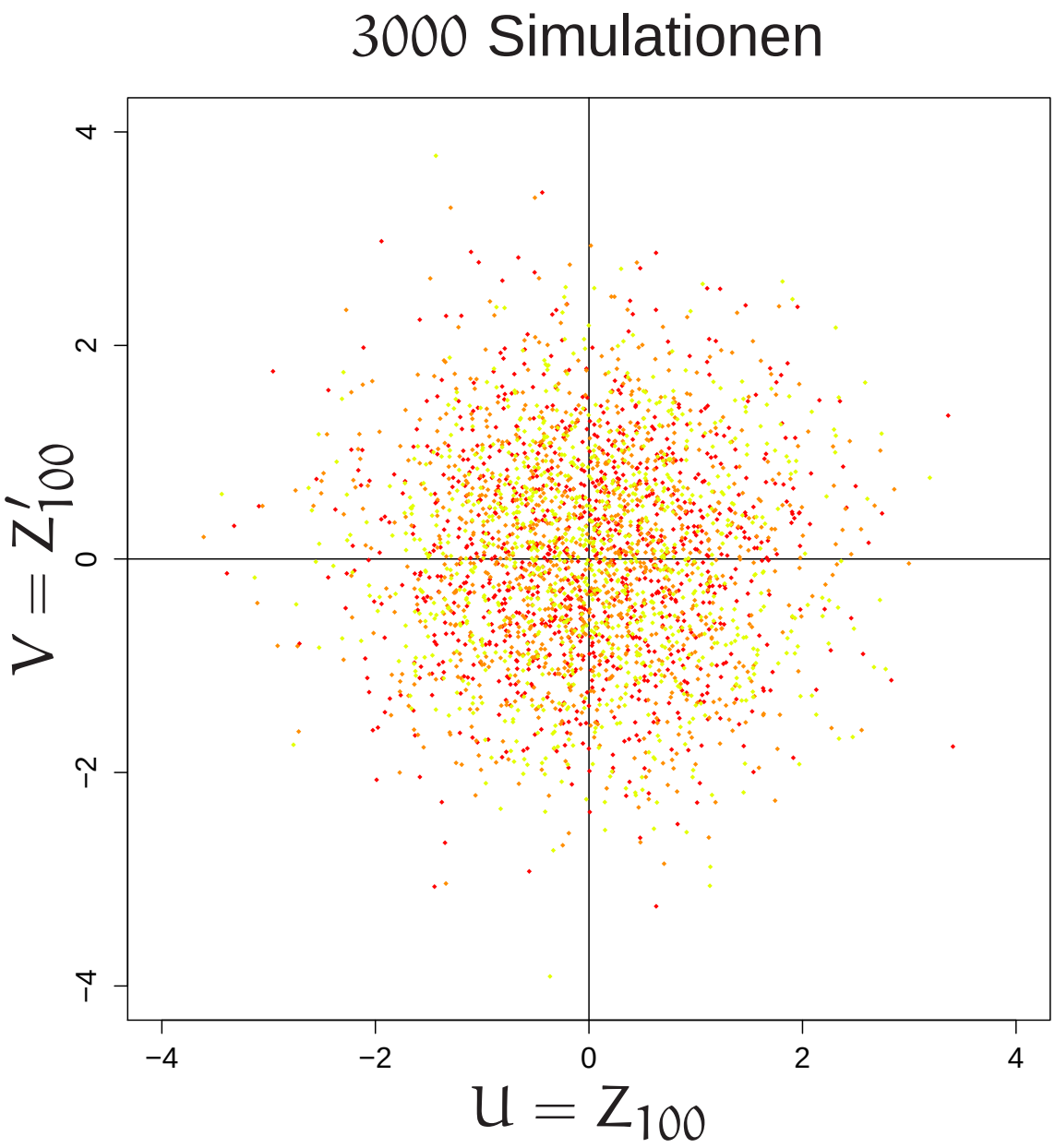
$$(U, V) = (Z_{100}, Z'_{100})$$

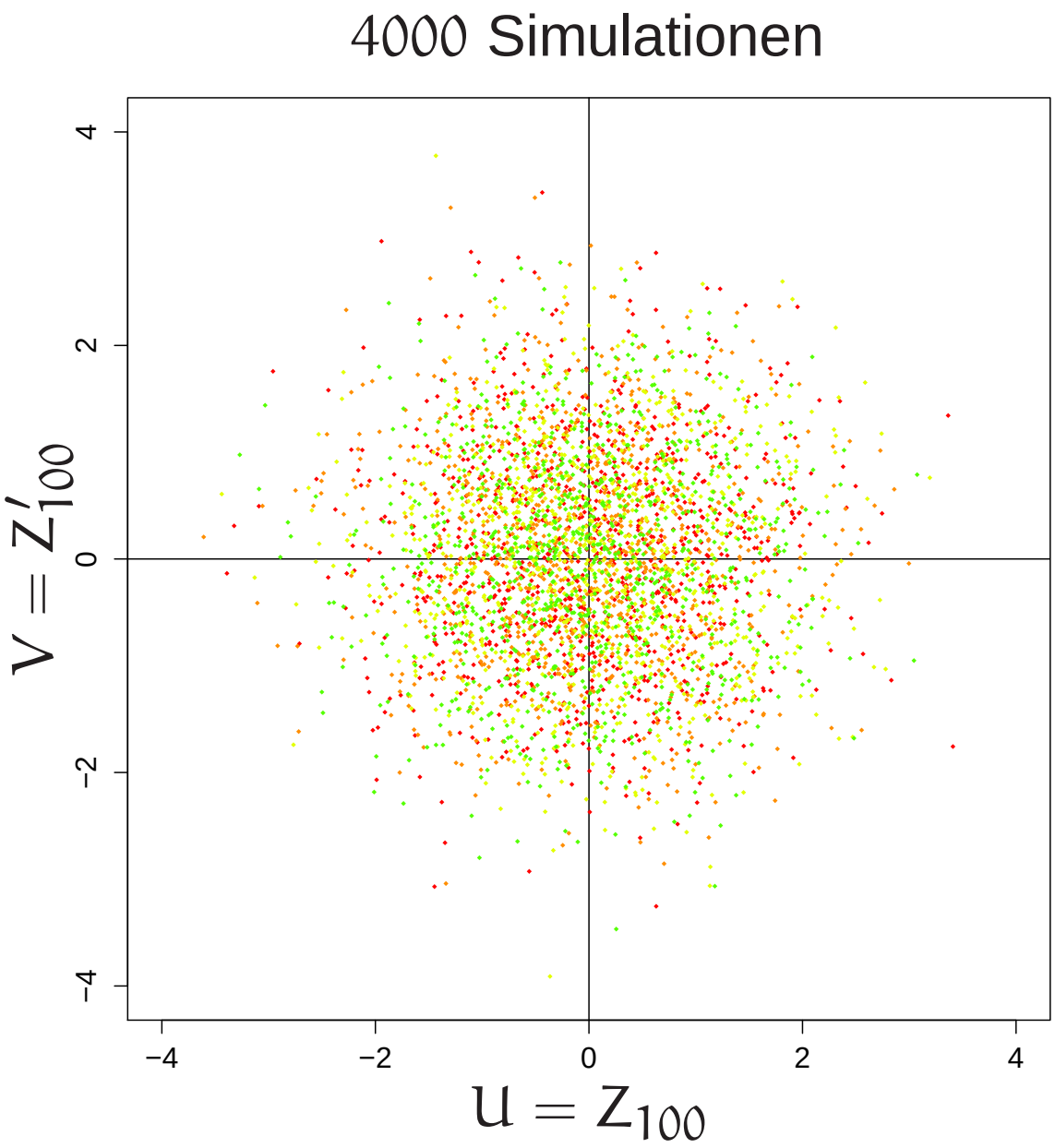
Wie sieht die gemeinsame Verteilung

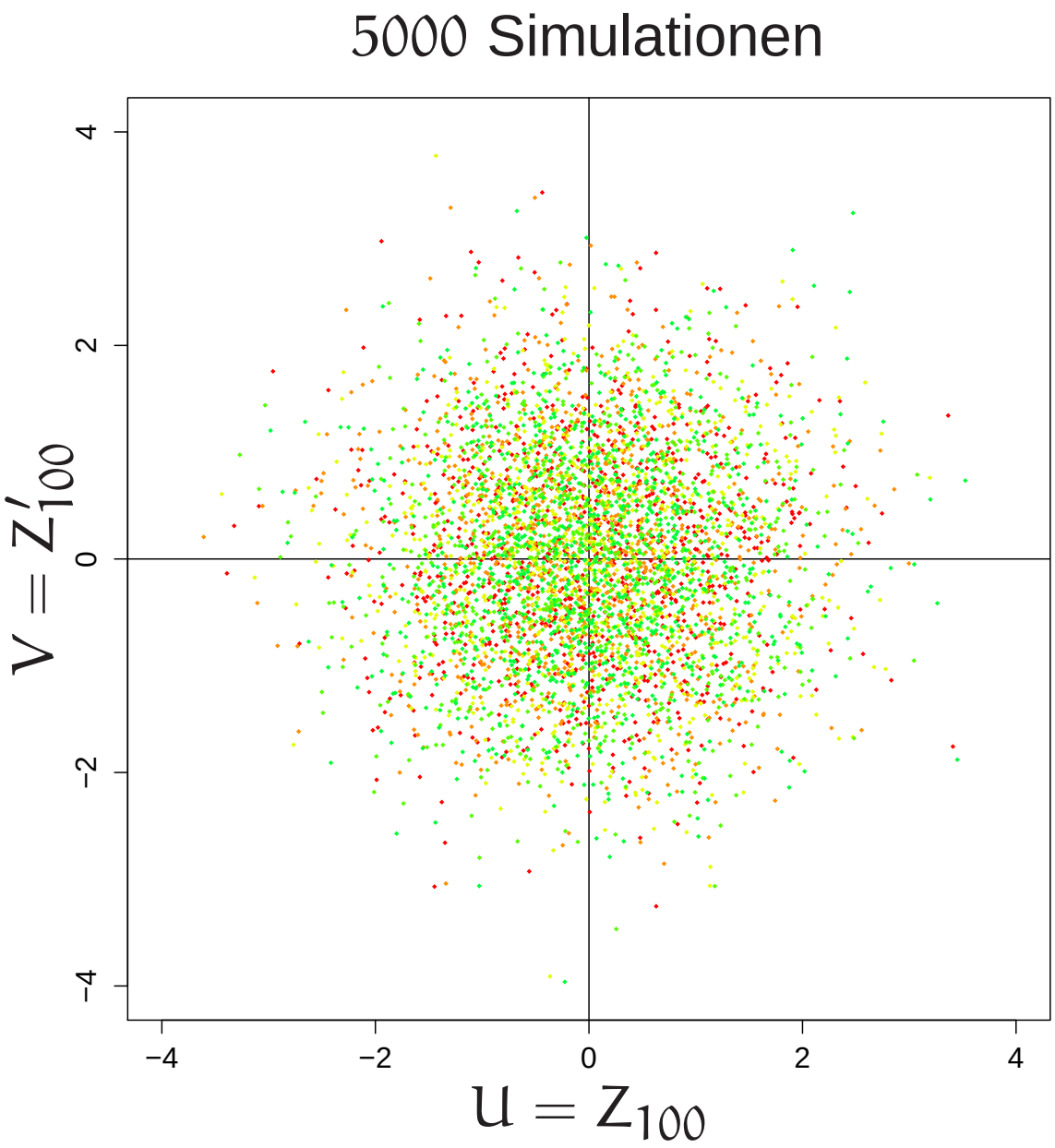
von U und V aus?

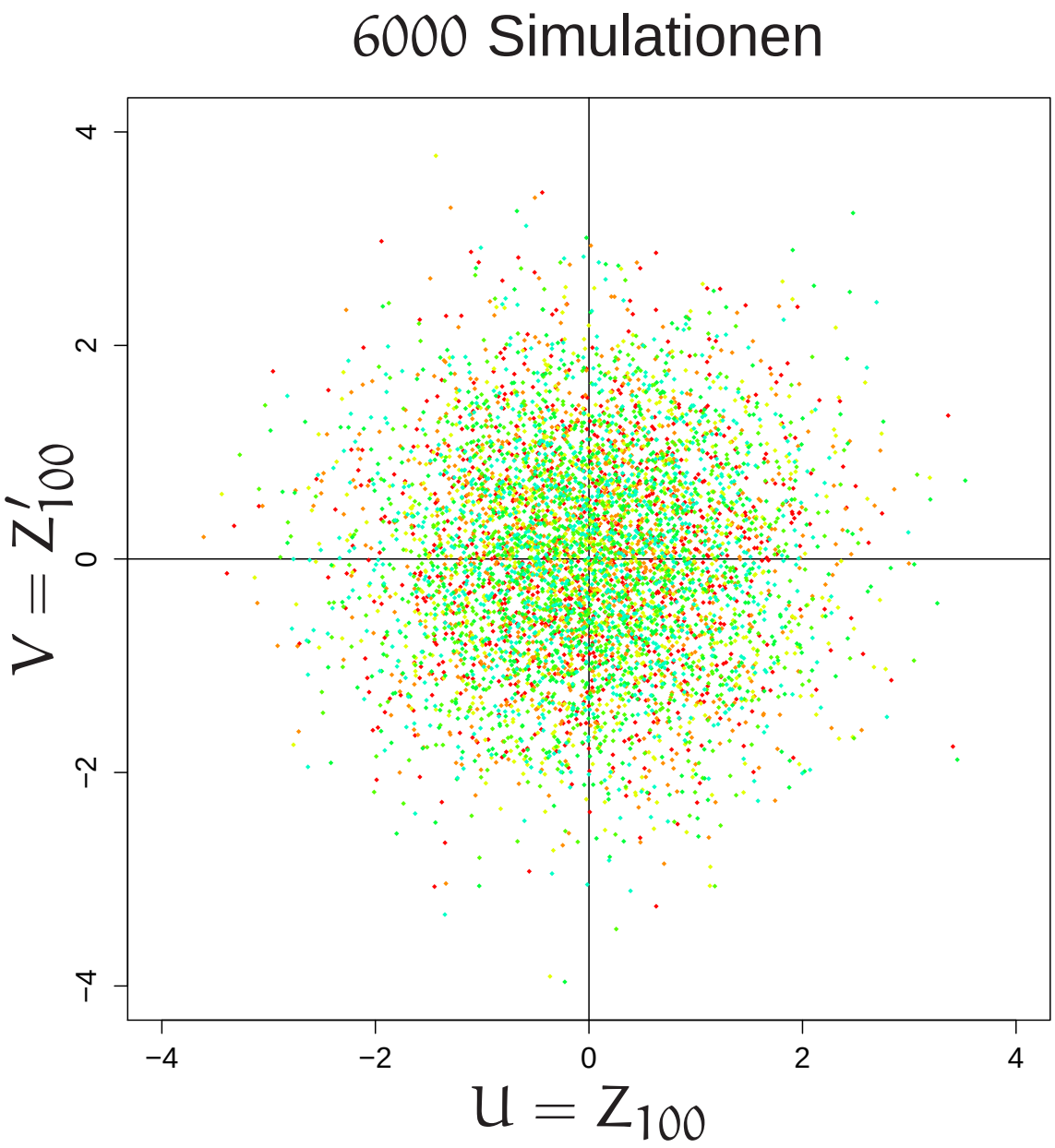

1000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

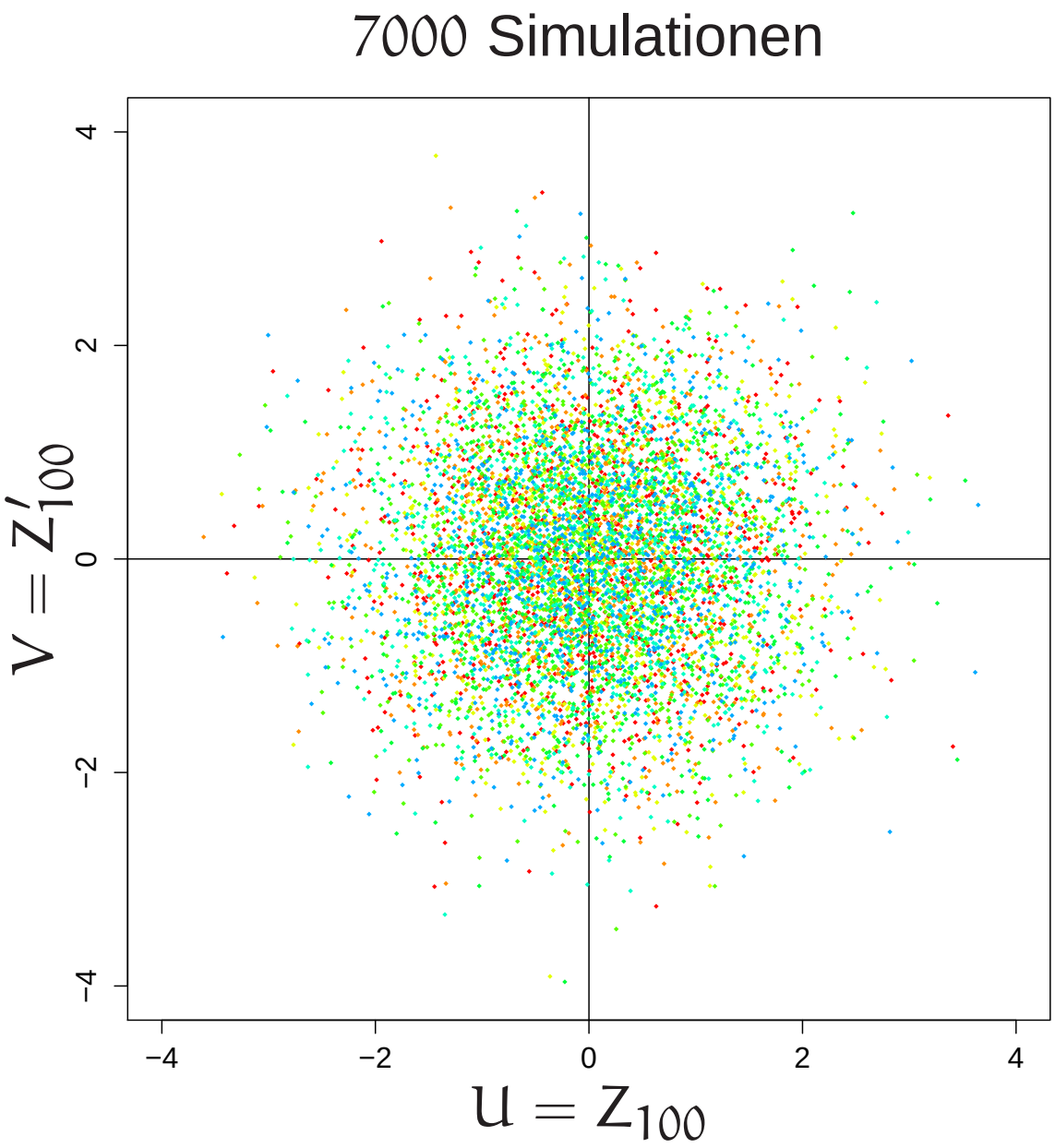

2000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

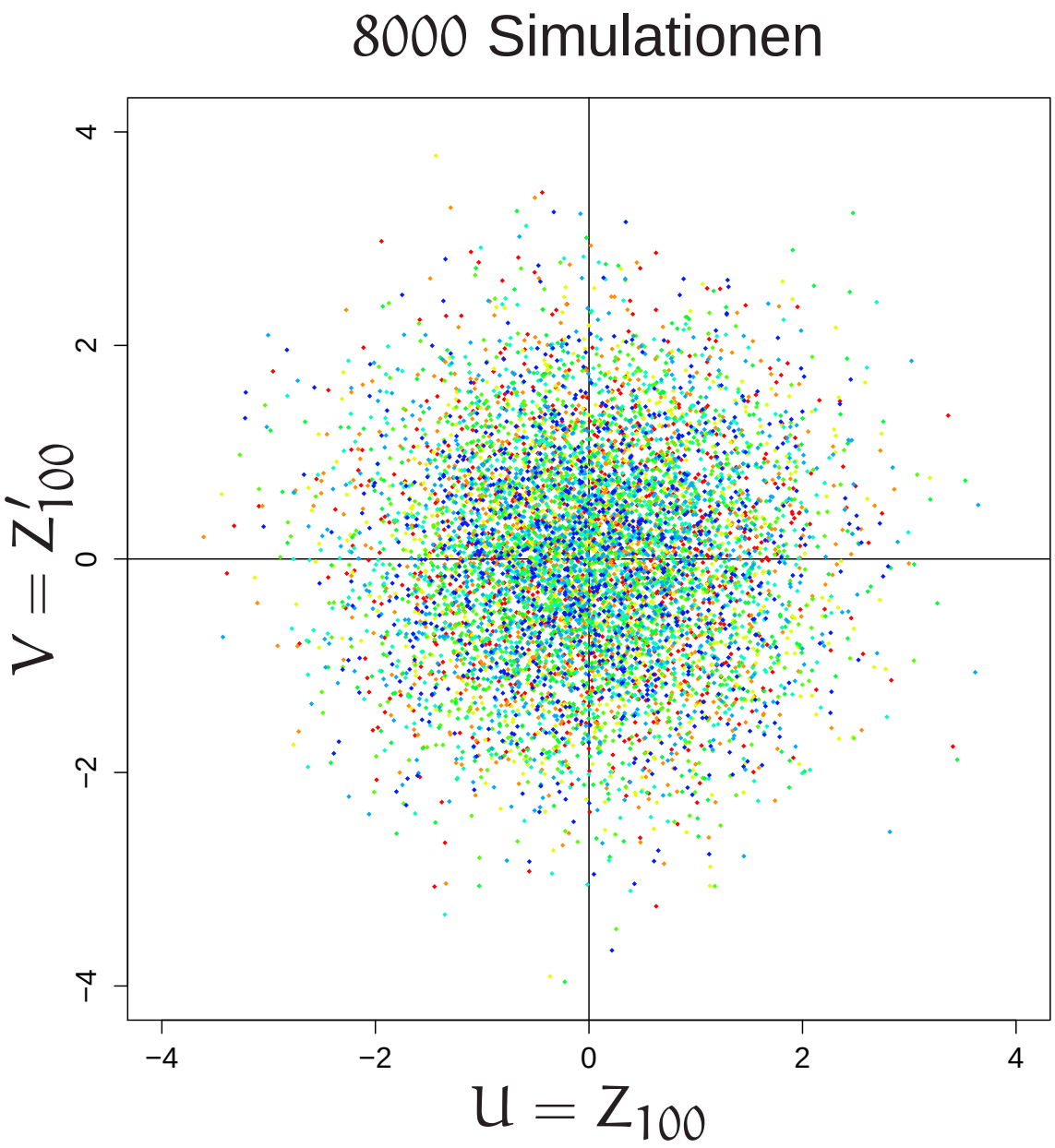

3000 Simulationen
4
2
0
−2
−4
$V = Z'_{100}$
−4
−2
0
2
4
$U = Z_{100}$

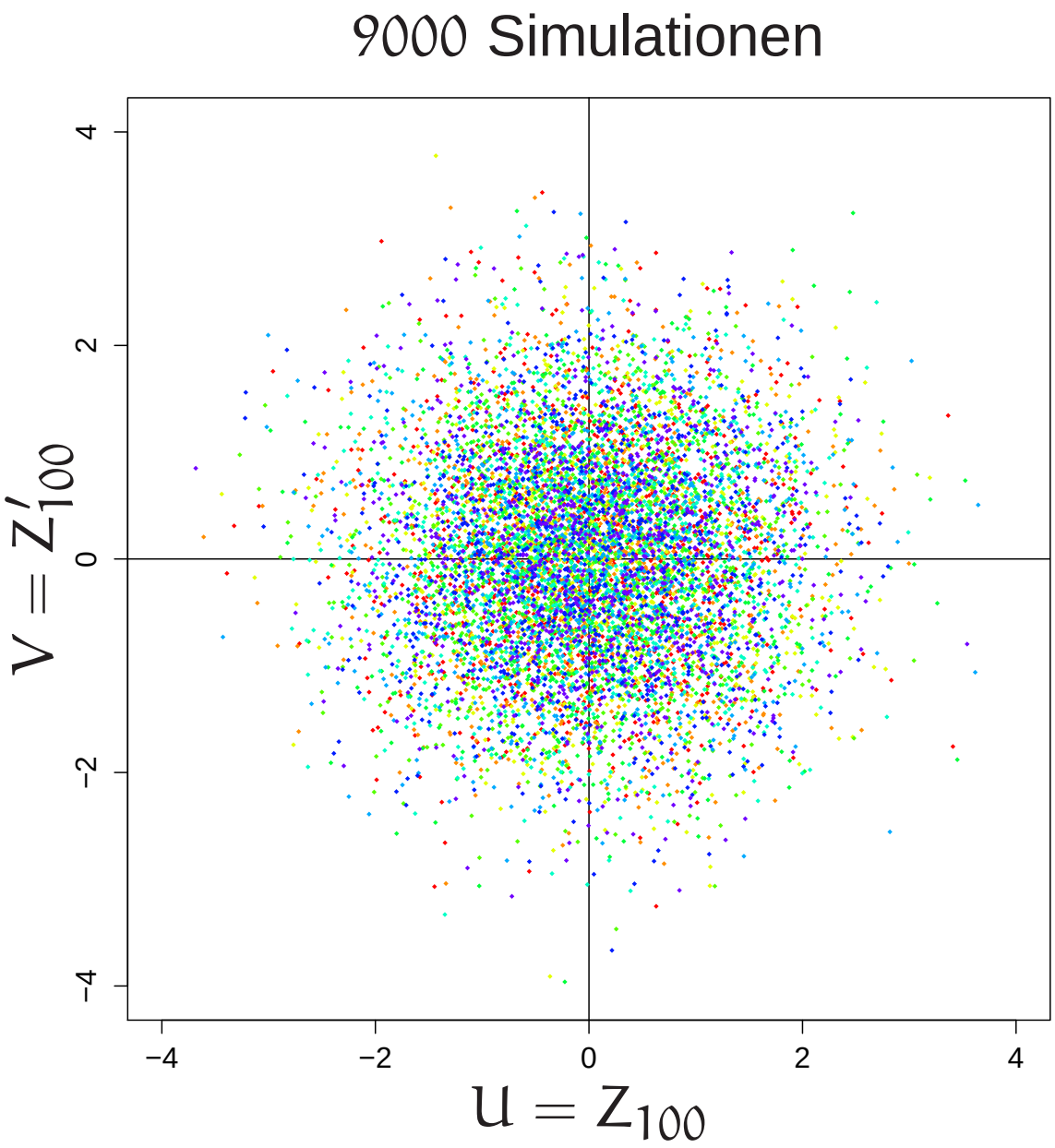

4000 Simulationen
4
2
0
−2
−4
$V = Z'_{100}$
−4
−2
0
2
4
$U = Z_{100}$

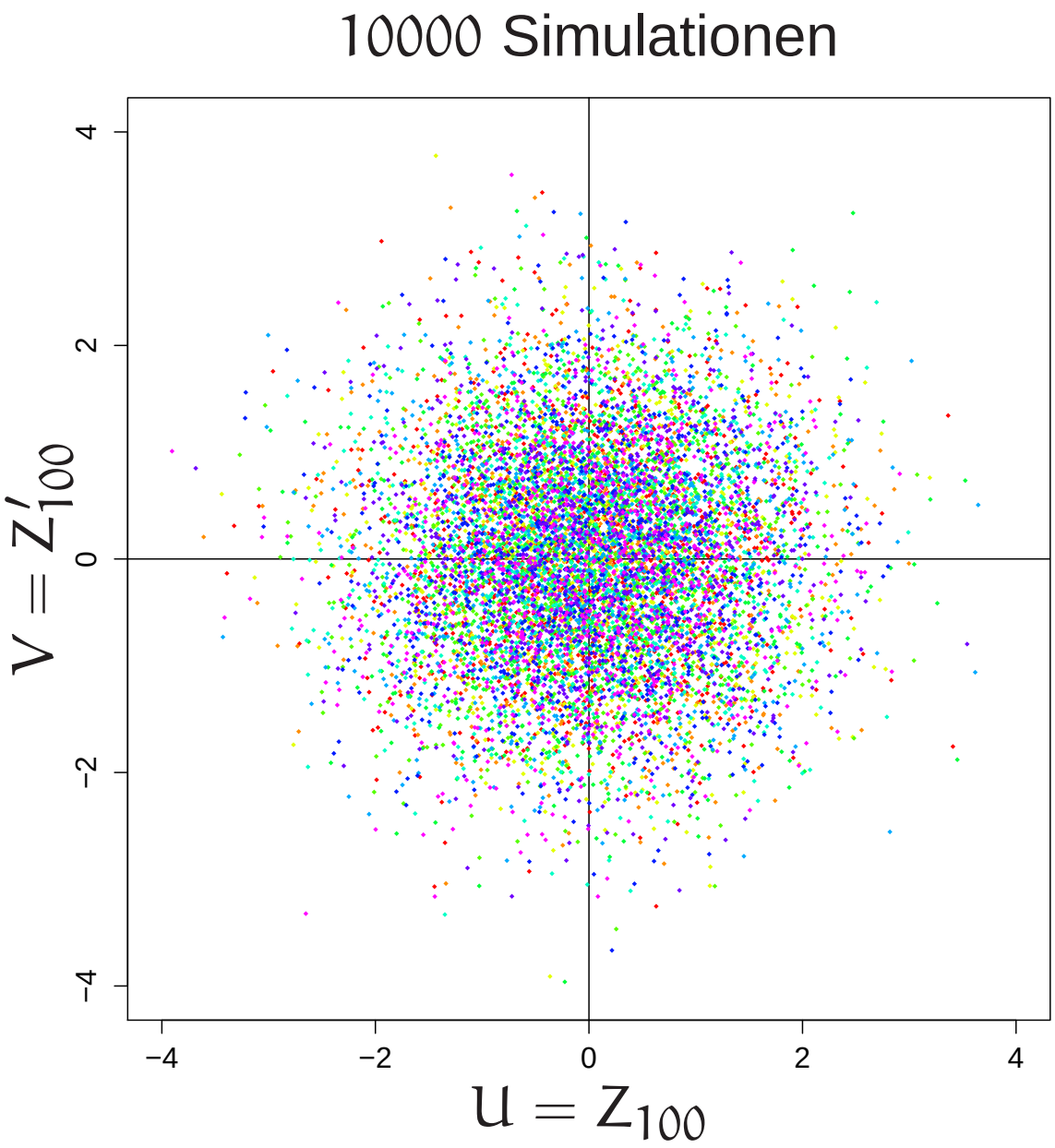

5000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

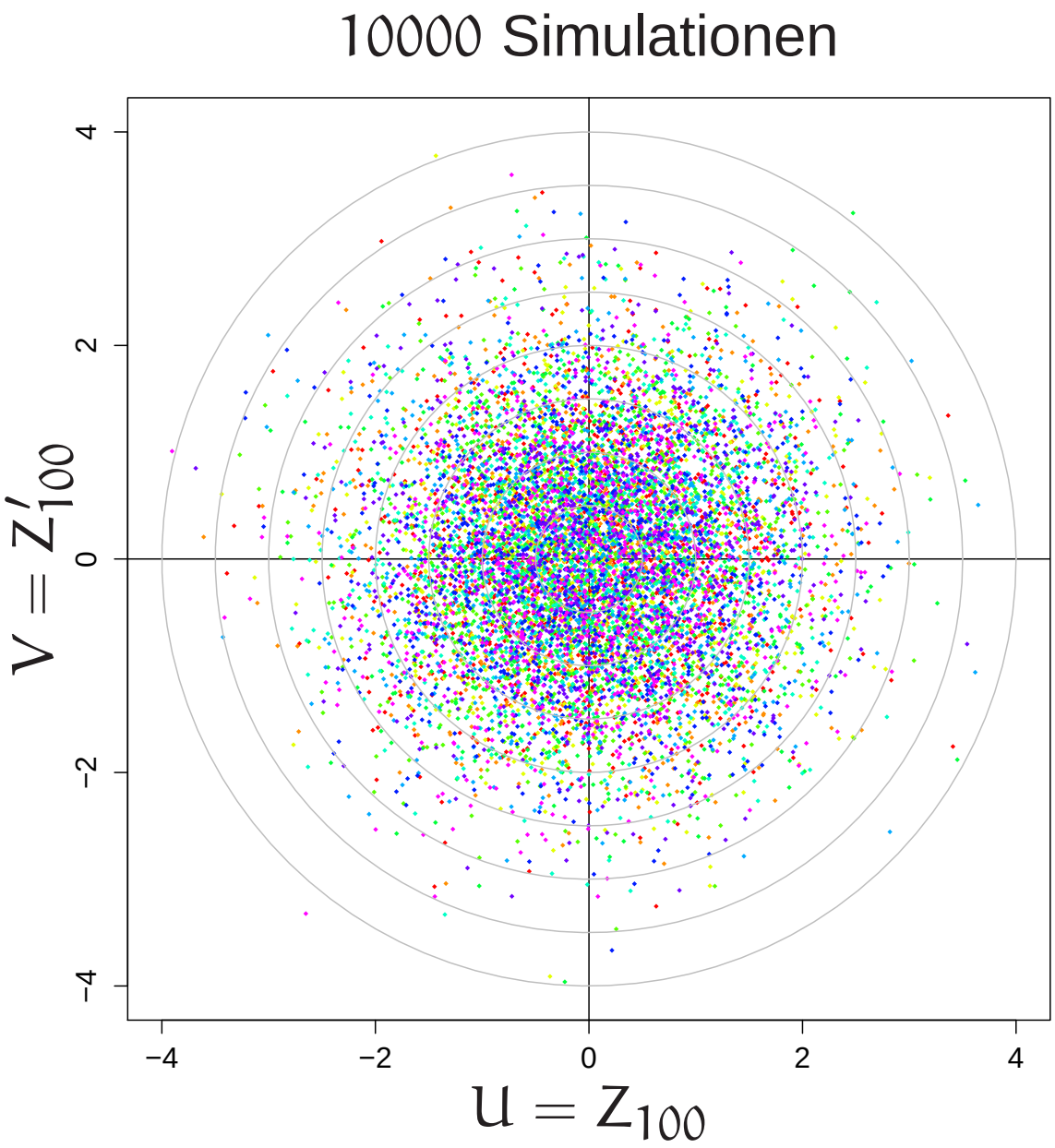

6000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$


7000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

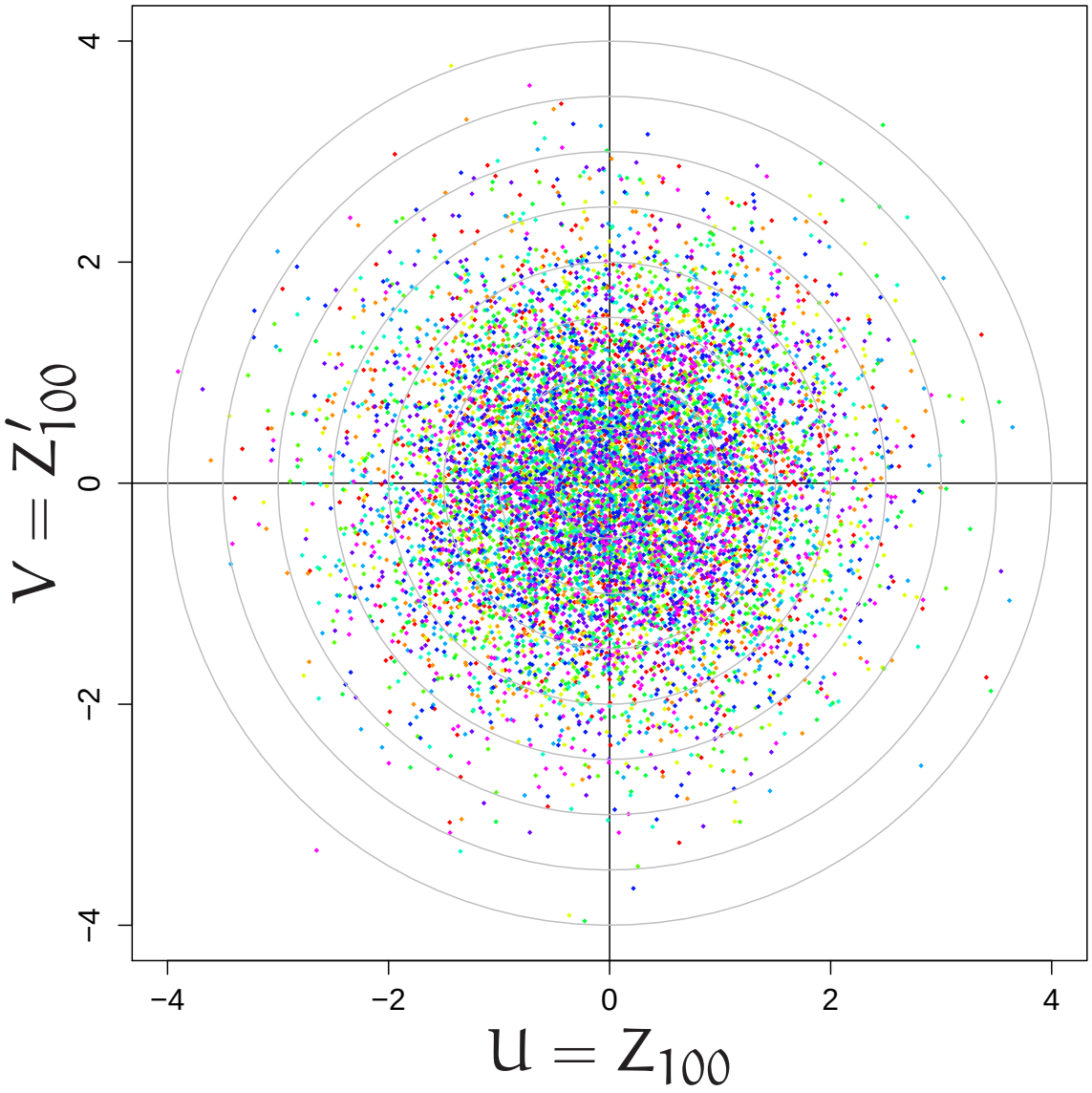

8000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$


9000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$v = z'_{100}$
$u = z_{100}$


10000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$


10000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$v = z'_{100}$
$u = z_{100}$

Die Verteilung von (U, V) ist rotationssymmetrisch!



Behauptung:

Aus U und V unabhängig

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

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

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

$$g(r) = c \cdot h(r) \qquad (c := h(0))$$

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

Eine Lösung:

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

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

Allgemeine Lösung

(logarithmieren; nach a ableiten)

$$h(x) = k_1 e^{-k_2 x^2}$$

Nebenbedingungen:

$$\mathbf{P}(\ U \in \mathbb{R}\ ) = \int h(x)\,dx = 1$$

$$\sigma_U^2 = \int x^2 h(x)\,dx = 1$$

$$\boxed{h(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}}$$

FAZIT

Der Zentrale Grenzwertsatz

lässt sich erraten

(in konkreten Fällen,

mit etwas Glück).

Der Münzwurf passt in den Zentralen Grenzwertsatz:

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 $\ell < r \in \mathbb{R}$

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

Dabei ist Z standard-normalverteilt.

Mit $\mu = p$ und $\sigma^2 = pq$ ergibt sich der alte Satz von de Moivre und Laplace.

Das passt zur Approximation der Binomialgewichte

$$\binom{n}{k} p^k q^{n-k} \quad \approx \quad \frac{1}{\sigma}\varphi\left(\frac{k-\mu}{\sigma}\right) \quad \approx \quad \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\, dx,$$

mit $\mu := np, \;\; \sigma := \sqrt{npq}$

$$\mathbf{P}(k_1 \le X \le k_2) = \int_{k_1-\frac{1}{2}}^{k_2+\frac{1}{2}} \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)\, dx = \int_{(k_1-\frac{1}{2}-\mu)/\sigma}^{(k_2+\frac{1}{2}-\mu)/\sigma} \varphi(z)\, dz$$

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.

Für unabhängige, identisch verteilte ZV'e $X_1, X_2, \ldots$
mit Erwartungswert μ und endlicher Varianz σ^2
bekommen wir aus dem Zentralen Grenzwertsatz:

Für große n ist der Mittelwert $\frac{X_1+\cdots+X_n}{n}$
annähernd normalverteilt
mit Erwartungswert μ und Standardabweichung $\frac{\sigma}{\sqrt{n}}$.

Daraus folgt insbesondere, dass für große n
die Verteilung von $\frac{X_1+\cdots+X_n}{n}$
in der Nähe von μ konzentriert ist:

$$\mathbf{P}\left(\left|\frac{X_1+\cdots+X_n}{n}-\mu\right| \geq \varepsilon\right) \to 0 \quad \text{für } n \to \infty$$

Das Schwache Gesetz der großen Zahlen

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \to 0 \quad \text{für } n \to \infty$$

gilt sogar für jede Folge von
paarweise unkorrelierten Zufallsvariablen
mit ein- und demselben Erwartungswert μ
und ein-und derselben Varianz σ^2.
Dahinter steckt,
dass der Erwartungswert von $\frac{X_1+\cdots+X_n}{n}$ gleich μ ist
und seine Standardabweichung klein wird, nämlich $\frac{\sigma}{\sqrt{n}}$.

Den (einfachen) Beweis des

Schwachen Gesetzes der Großen Zahlen

bereiten wir vor durch zwei einfache Ungleichungen:

die Markov-Ungleichung und die Chebyshev-Ungleichung.

Markov-Ungleichung:

Für jede Zufallsvariable $X \geq 0$ und jedes $\varepsilon > 0$ gilt

$$\boxed{\mathbf{P}(X \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbf{E}[X].}$$

Beweis: Es gilt

$$\varepsilon I_{\{X \geq \varepsilon\}} = \varepsilon \mathbf{1}_{[\varepsilon,\infty]}(X) \leq X.$$

Mit der Monotonie des Erwartungswertes

folgt die Behauptung. $\square$

Chebyshev-Ungleichung:

Für eine reellwertige Zufallsvariable X mit endlichem Erwartungswert gilt für beliebiges $\varepsilon > 0$

$$\boxed{\mathbf{P}\big(|X - \mathbf{E}[X]| \geq \varepsilon\big) \leq \frac{1}{\varepsilon^2} \cdot \mathbf{Var}[X]}$$

Beweis:

Markov-Ungleichung angewandt auf $Y = (X - \mathbf{E}[X])^2$:

$$\begin{aligned} \mathbf{P}\big(|X - \mathbf{E}[X]| \geq \varepsilon\big) &= \mathbf{P}(Y \geq \varepsilon^2) \\ &\leq \varepsilon^{-2}\mathbf{E}[Y] = \varepsilon^{-2}\mathbf{Var}[X] \,. \end{aligned}$$

□

Schwaches Gesetz der Großen Zahlen

(Buch S. 74)

Die Zufallsvariablen $X_1, X_2, \ldots$ seien reellwertig, identisch verteilt mit endlichem Erwartungswert μ und endlicher Varianz, und sie seien unkorreliert.

Dann gilt für alle $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0 \,.$$

Beweis: Gemäß Voraussetzung gilt

$$\mathbf{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \mu\,,$$
$$\mathbf{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{\mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n]}{n^2} = \frac{\mathbf{Var}[X_1]}{n}\,.$$

Die Chebyshev-Ungleichung, angewandt auf

$(X_1 + \cdots + X_n)/n$ ergibt

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\mathbf{Var}[X_1]}{\varepsilon^2 n}\,.$$

Die Behauptung folgt nun mit $n \to \infty$. $\square$

Das Schwache Gesetz der Großen Zahlen wurde von Jacob Bernoulli im Münzwurfmodell endeckt.

Ergänzung

Zum Schätzen von Mittelwerten

Eine große Population von Werten

$$w_1, \ldots, w_g$$

ist auf der Zahlengeraden verteilt.

Man interessiert sich für den Populationsmittelwert:

$$\mu := \frac{1}{g} \sum_{j=1}^{g} w_j.$$

Zur Verfügung stehen die Werte einer

rein zufällig aus der Population gezogene Stichprobe

$$x_1, \ldots, x_n$$

$$m := \frac{1}{n}(x_1 + \ldots + x_n)$$

ist ein *Schätzwert* für μ.

Wie zuverlässig ist diese Schätzung?

Das hängt ab

– von der “Streuung” der Werte w_j in der Population

und

– von der Größe n der Stichprobe.

Goldene Idee der Statistik:

Man fasst $x_1, \ldots, x_n$ auf als Ergebnis eines (rein) zufälligen Ziehens aus der Population:

$$X_1 := w_{J_1},\ X_2 := w_{J_2}, \ldots$$

mit $J_1, J_2, \ldots$ rein zufällige Wahl aus $\{1, \ldots, g\}$.

Ein Maß für die Variabilität in der Population ist

$$\sigma^2 := \frac{1}{g} \sum_{j=1}^{g} (w_j - \mu)^2$$

Stellen wir uns vor:

$$X_1, X_2, \ldots$$

unabhängig und identisch verteilt

mit Erwartungswert μ und Varianz σ^2

$$m = \frac{1}{n}(x_1 + \ldots + x_n)$$

fasst man auf

als eine Realisierung (einen Ausgang)

der Zufallsvariable

$$\bar{X} := \frac{1}{n}(X_1 + \ldots + X_n)$$

Wie ist $\bar{X}$ verteilt?

Wie ist $\bar{X}$ verteilt?

Der Zentrale Grenzwertsatz gibt eine Antwort:

$\bar{X}$ ist approximativ $N(\mu, \frac{\sigma^2}{n})$-verteilt.

Ein Problem in der Praxis: Man kennt σ^2 nicht.

Auch σ^2 muss man schätzen.

Ein Vorschlag:

$$s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2$$

Warum im Nenner $n-1$ und nicht n?

Eine Antwort dazu gibt's auf dem nächsten Übungsblatt.

Eine anschauliche Präsentation der Thematik

“Wie genau ist eine Schätzung des Mittelwertes?”

vom Standpunkt der Anwendung aus

finden Sie auf den Folien der Vorlesung

“Statistik für Biologen”

http://ismi.math.uni-frankfurt.de/wakolbinger/teaching/statbio10/statbio.html

Folien 2, “Der Standardfehler”