Vorlesung 8a

Der zentrale Grenzwertsatz

und

Mittelwerte

1. Auftakt

1. Auftakt

Letztesmal hatten wir gesehen:

Die Summe von
unabhängigen, normalverteilten Zufallsvariablen
ist wieder normalverteilt.

Insbesondere ergibt sich:

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

Mit anderen Worten: Sind $N_1, N_2, \ldots, N_n$
unabhängig und $N(\mu, \sigma^2)$-verteilt, dann ist

$$\frac{N_1 + \cdots + N_n - n\mu}{\sqrt{n\sigma^2}}$$ standard-normalverteilt.

2. Der Zentrale Grenzwertsatz: Die Botschaft

Der Zentrale Grenzwertsatz liefert

eine gewaltige Weiterung der vorigen Aussage

(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 $c < d \in \mathbb{R}$

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

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.

3. Zentraler Grenzwertsatz:

Meilensteine in seiner Geschichte

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

$$\pi =$$

$$3.14159265358979323846264338327950288419716939937510\ldots$$

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?

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

Ein Beispiel:

$X_1, X_2, \ldots$ unabhängig

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

Empirische Verteilung von

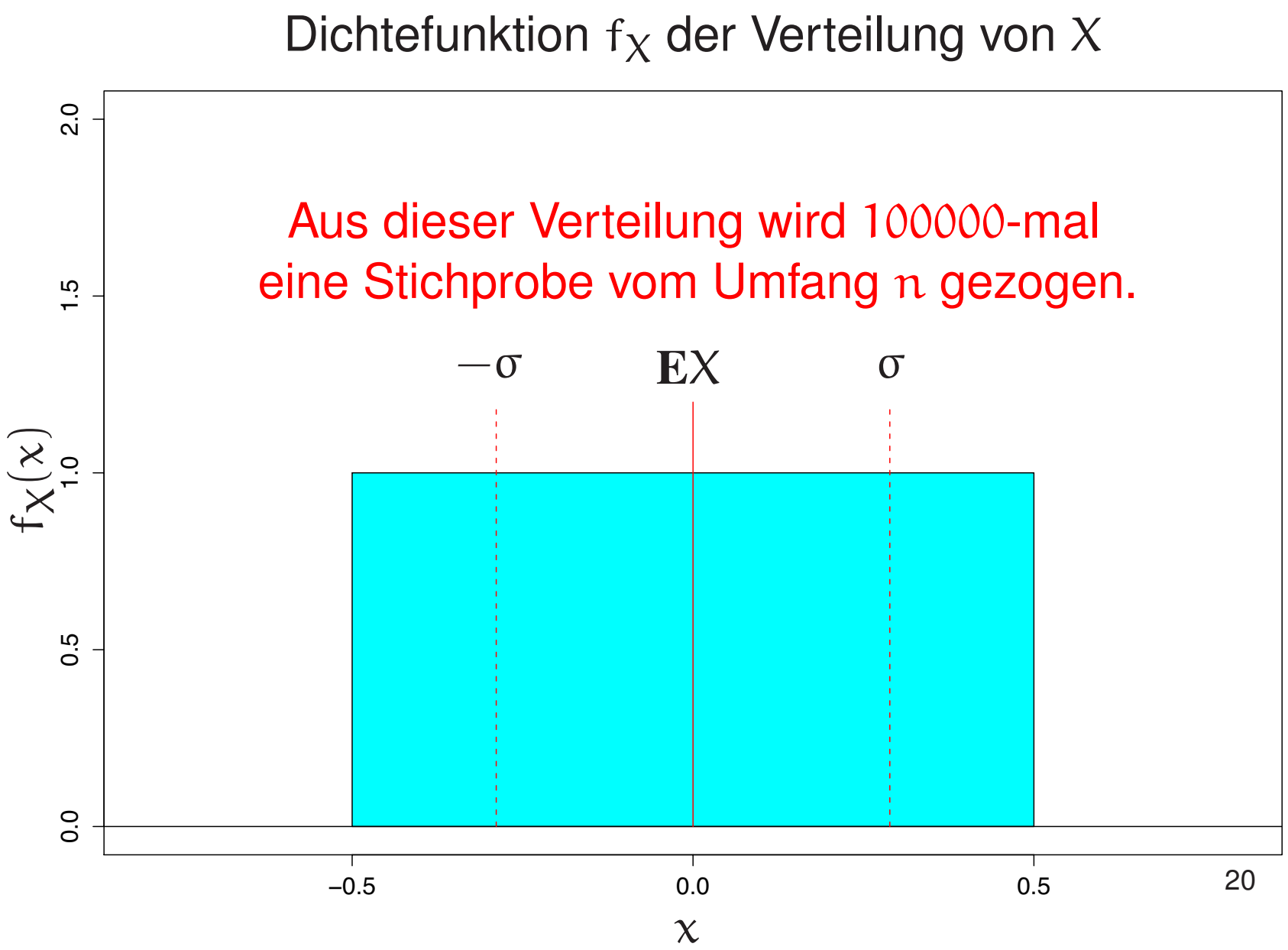
$$S_n := X_1 + ... + X_n$$

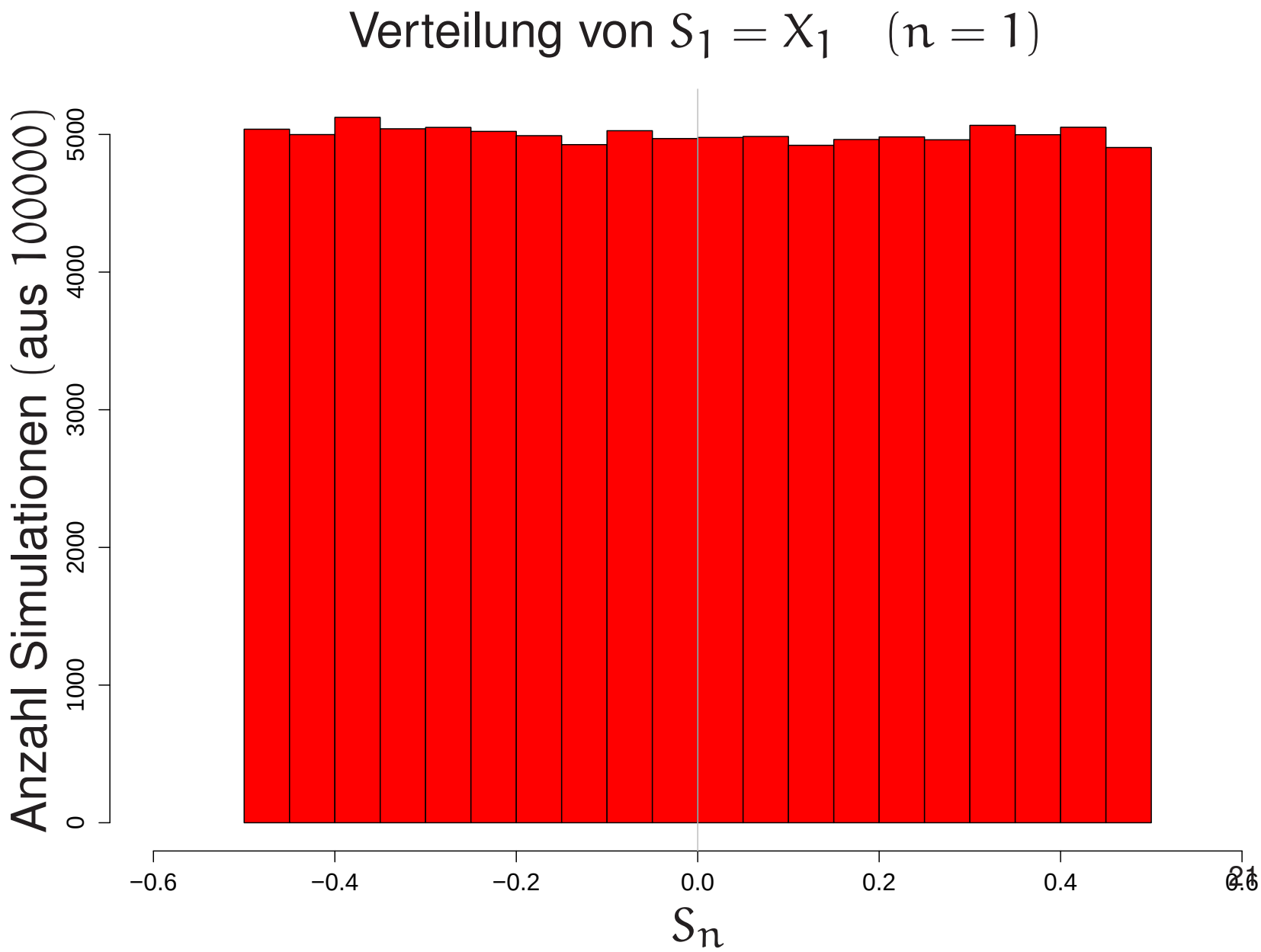
100000 Simulationen

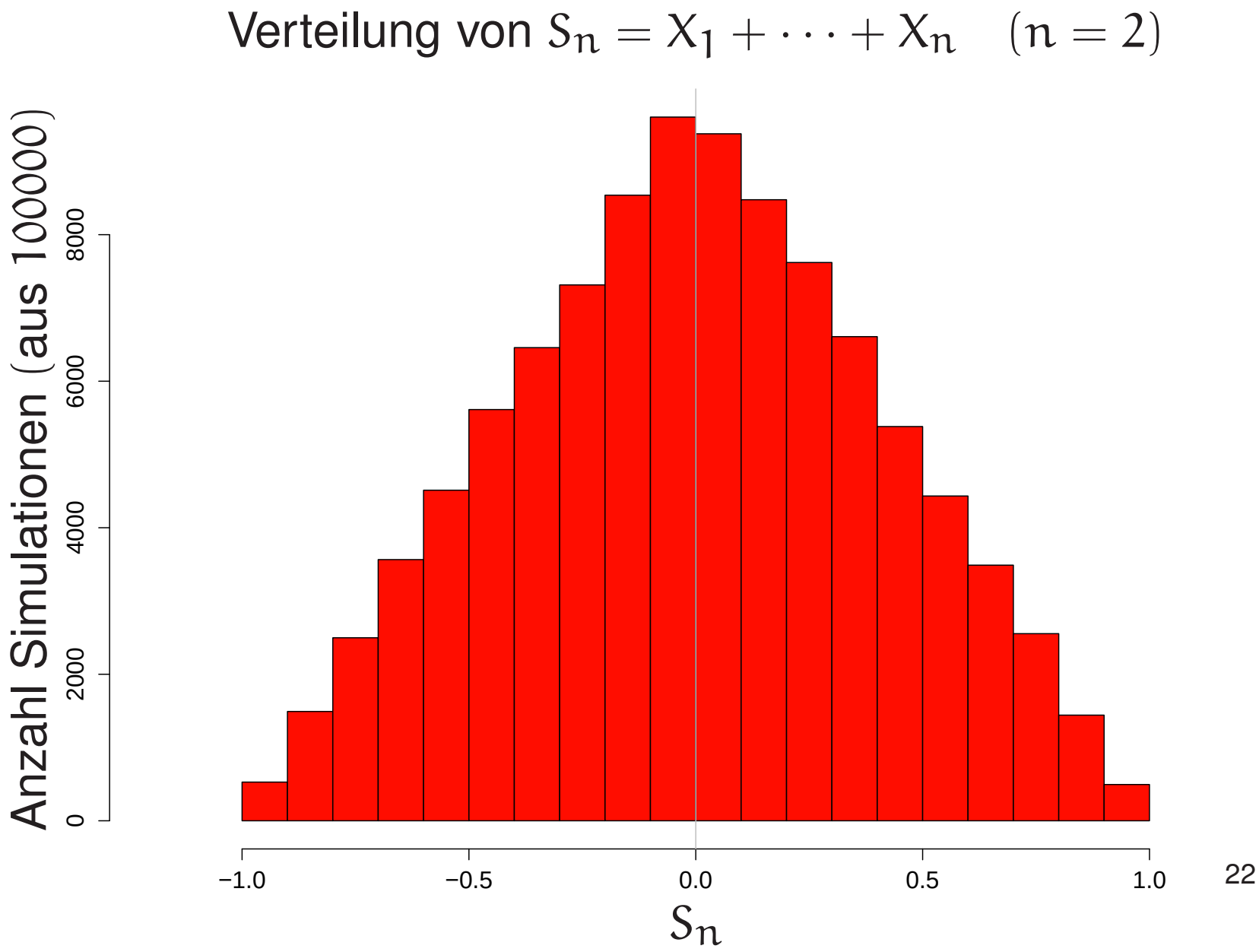
jeweils für

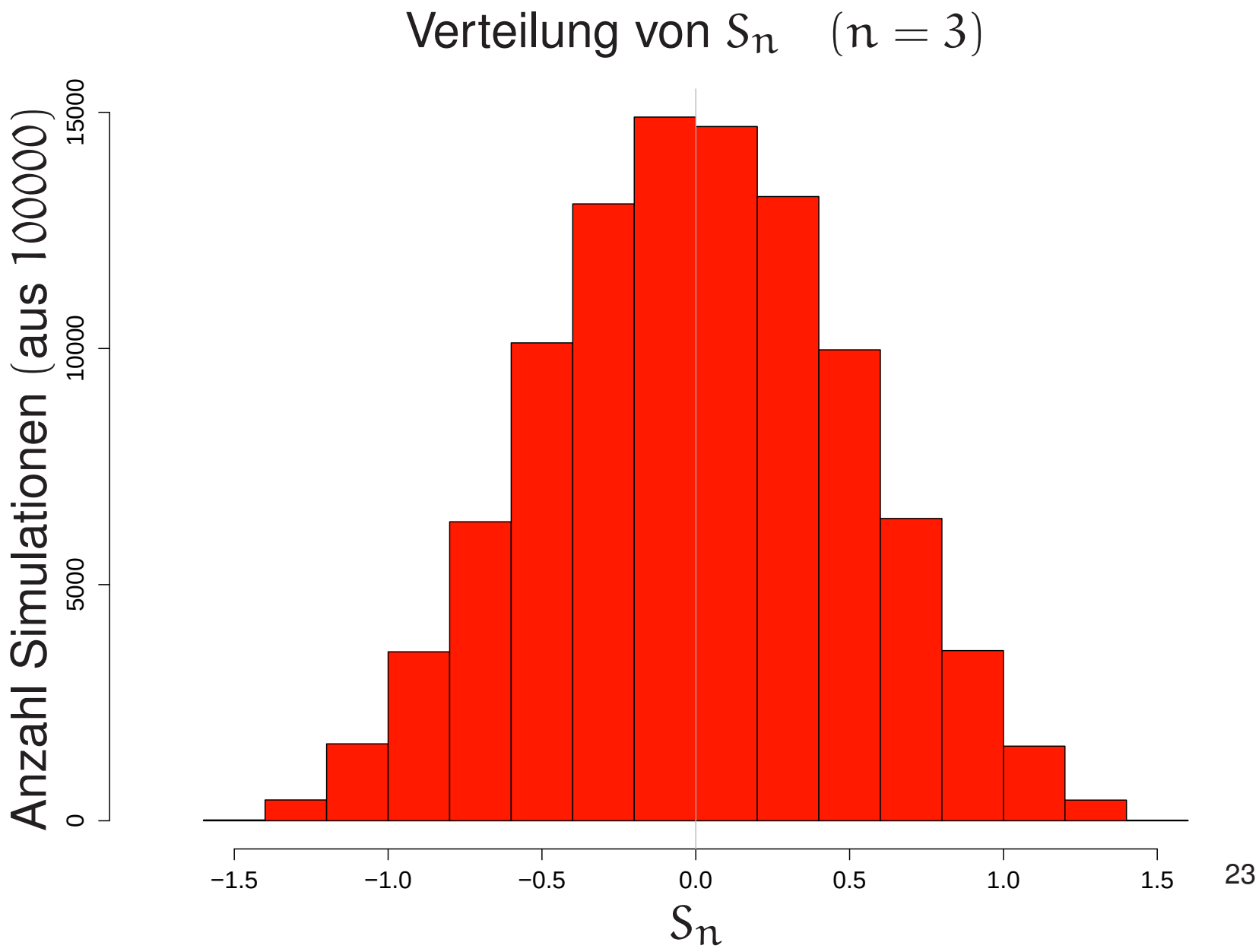
$$n = 1, 2, ..., 10$$

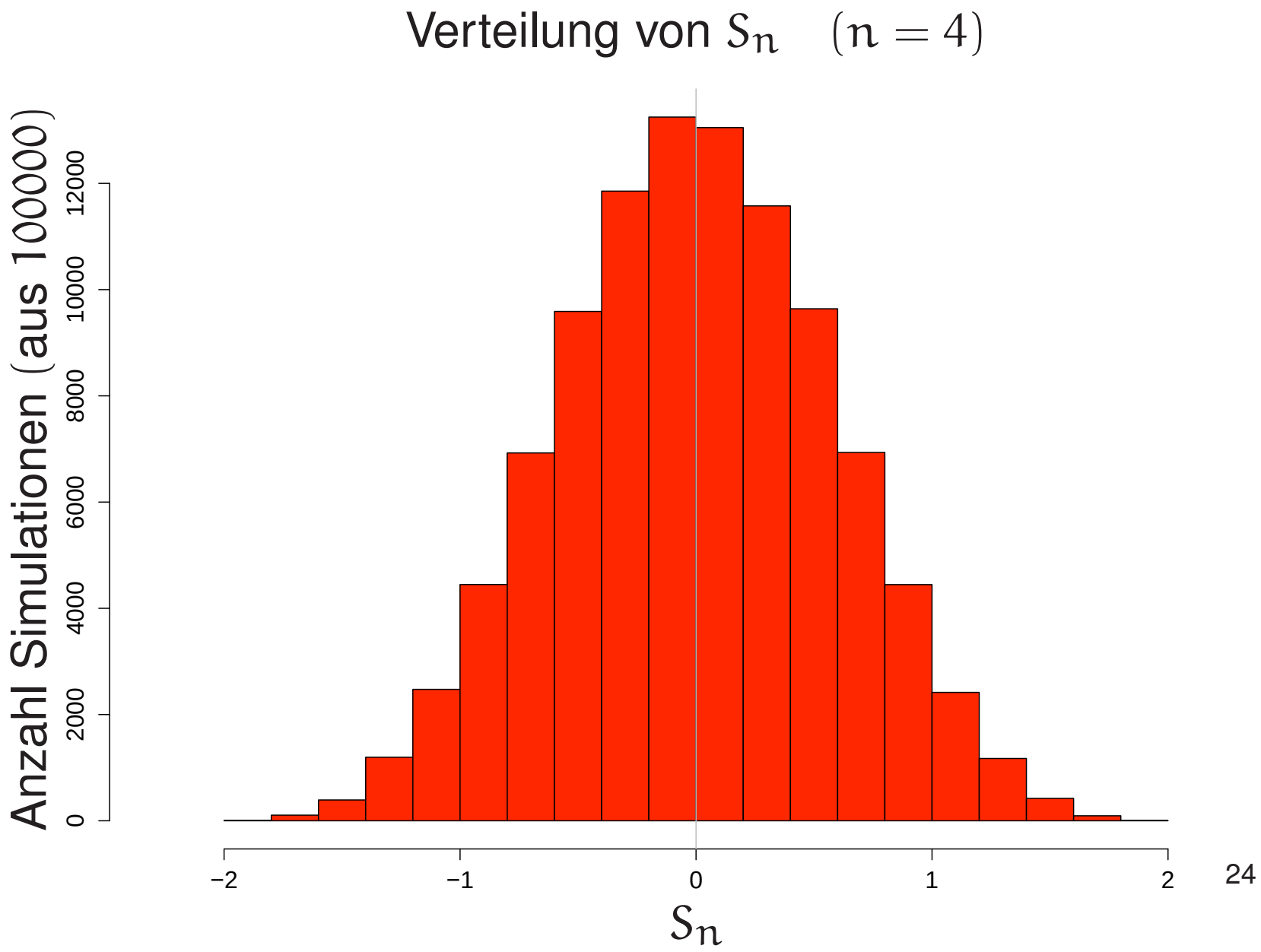
$$n = 15, 20, ..., 100$$

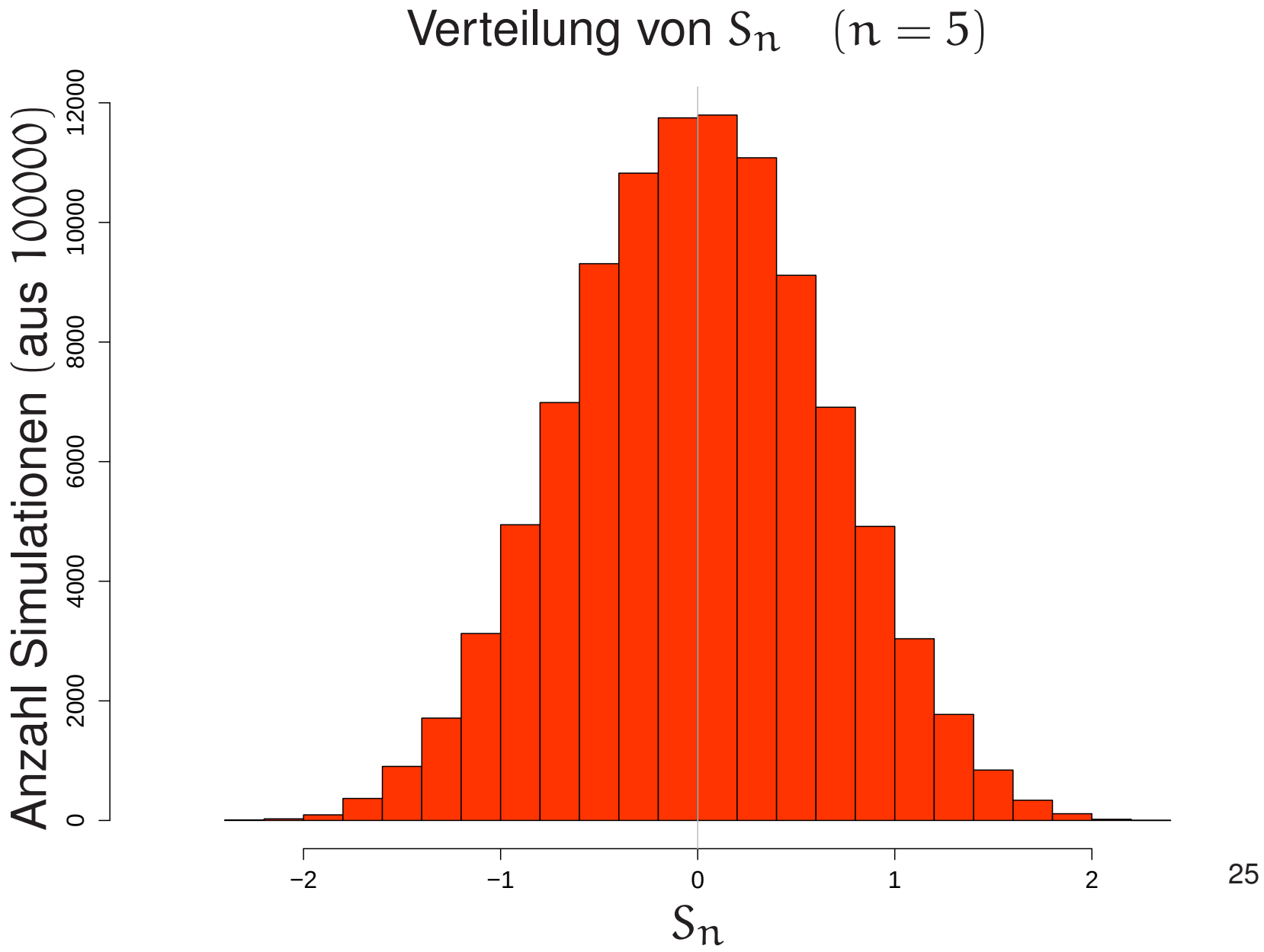

Dichtefunktion f_X der Verteilung von X
Aus dieser Verteilung wird 100000-mal
eine Stichprobe vom Umfang n gezogen.
$-\sigma$
EX
σ
$f_X(x)$
2.0
1.5
1.0
0.5
0.0
−0.5
0.0
0.5
20
x

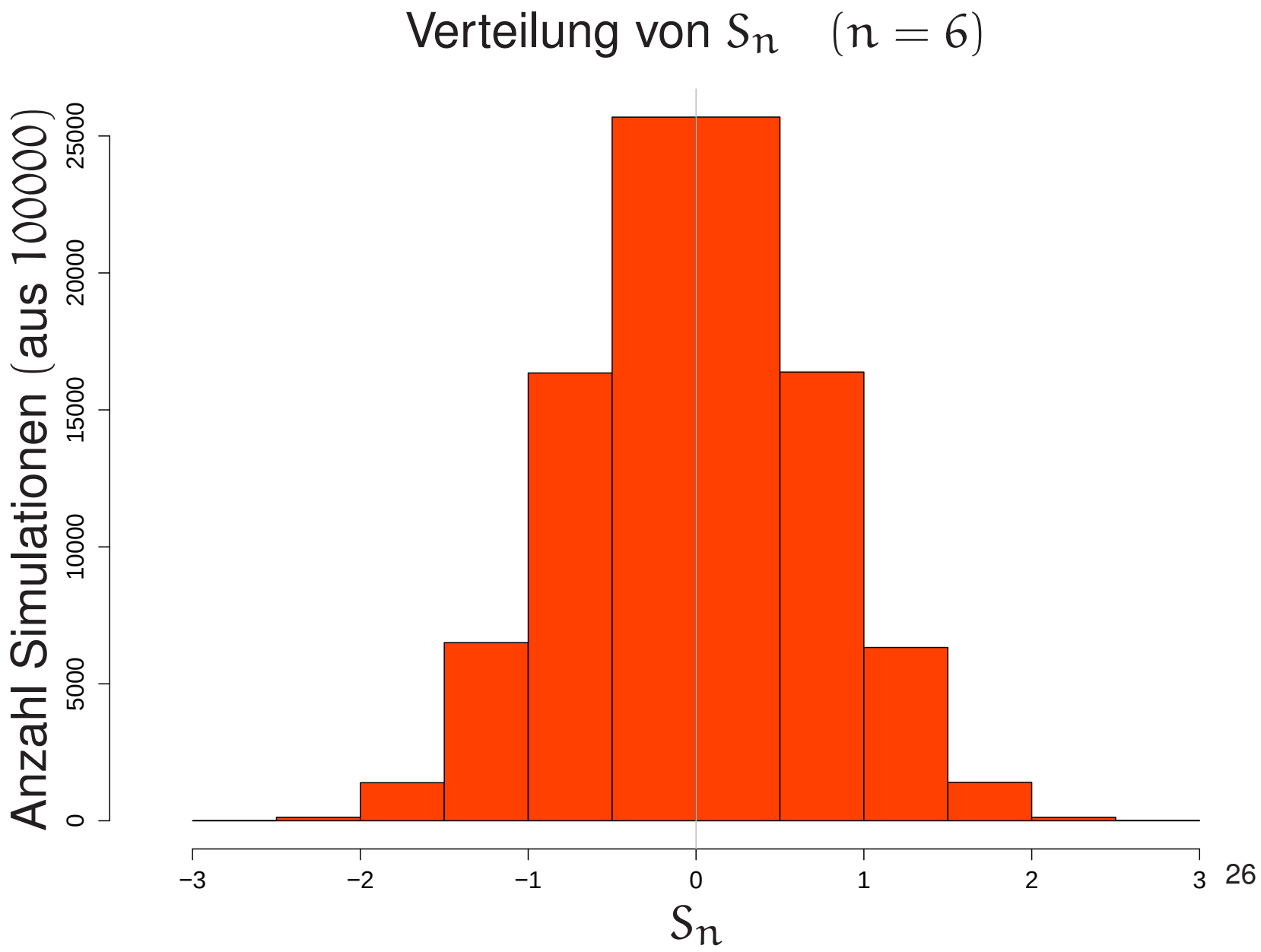

Verteilung von $S_1 = X_1 \quad (n = 1)$
Anzahl Simulationen (aus 100000)
0
1000
2000
3000
4000
5000
−0.6
−0.4
−0.2
0.0
0.2
0.4
S_n

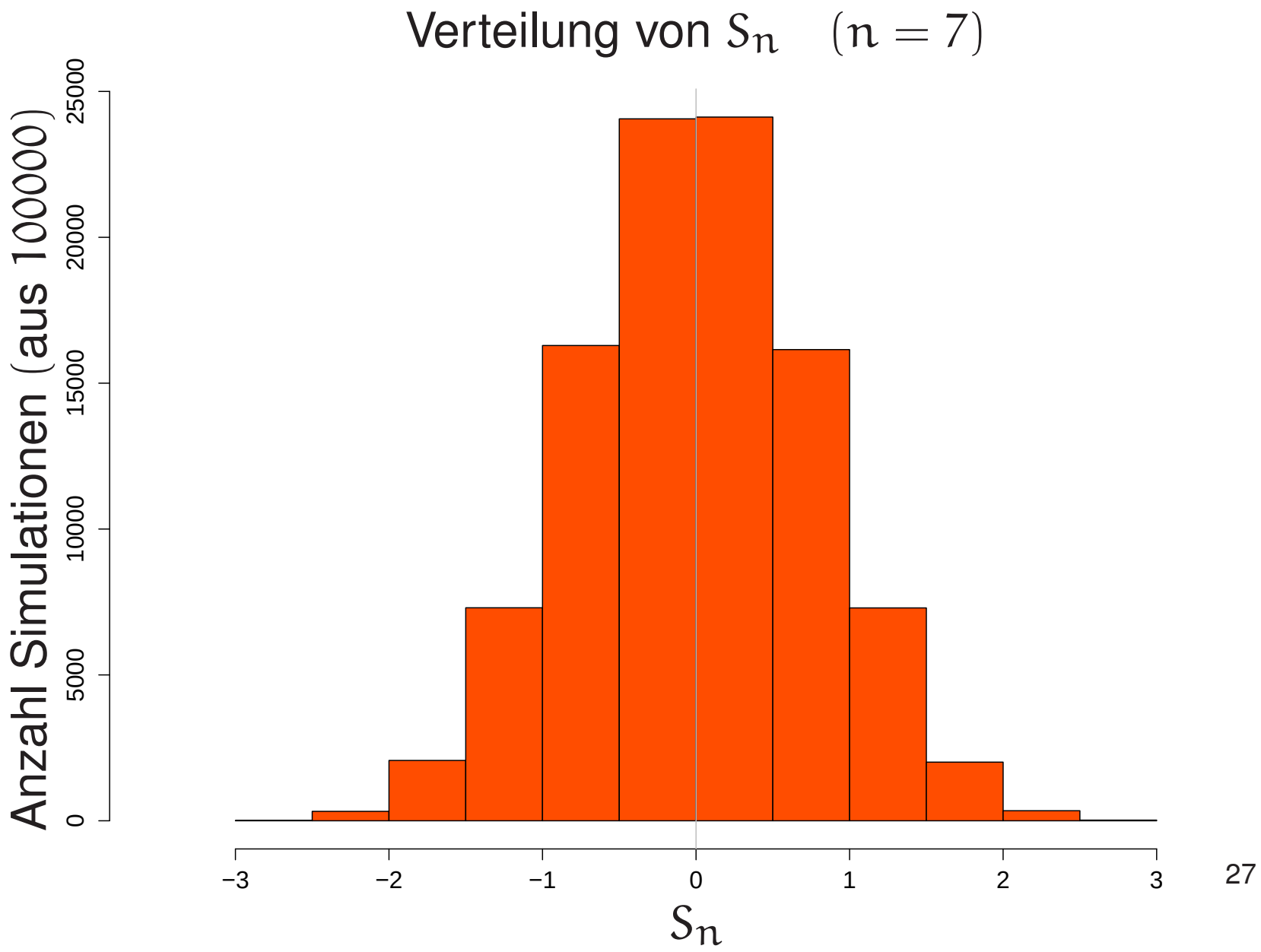

Verteilung von $S_n = X_1 + \cdots + X_n \quad (n = 2)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
−1.0
−0.5
0.0
0.5
1.0
S_n

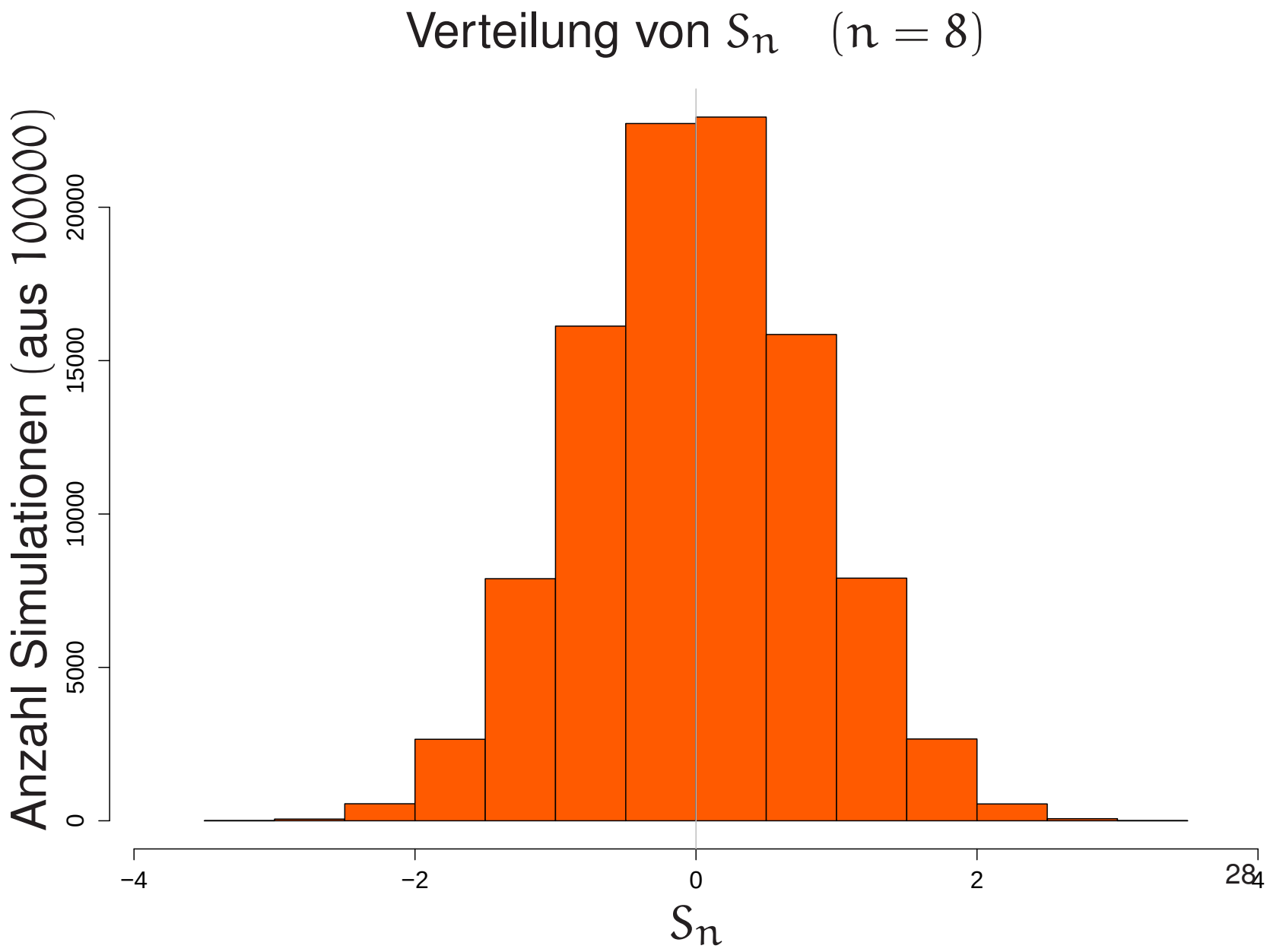

Verteilung von S_n $(n = 3)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
S_n

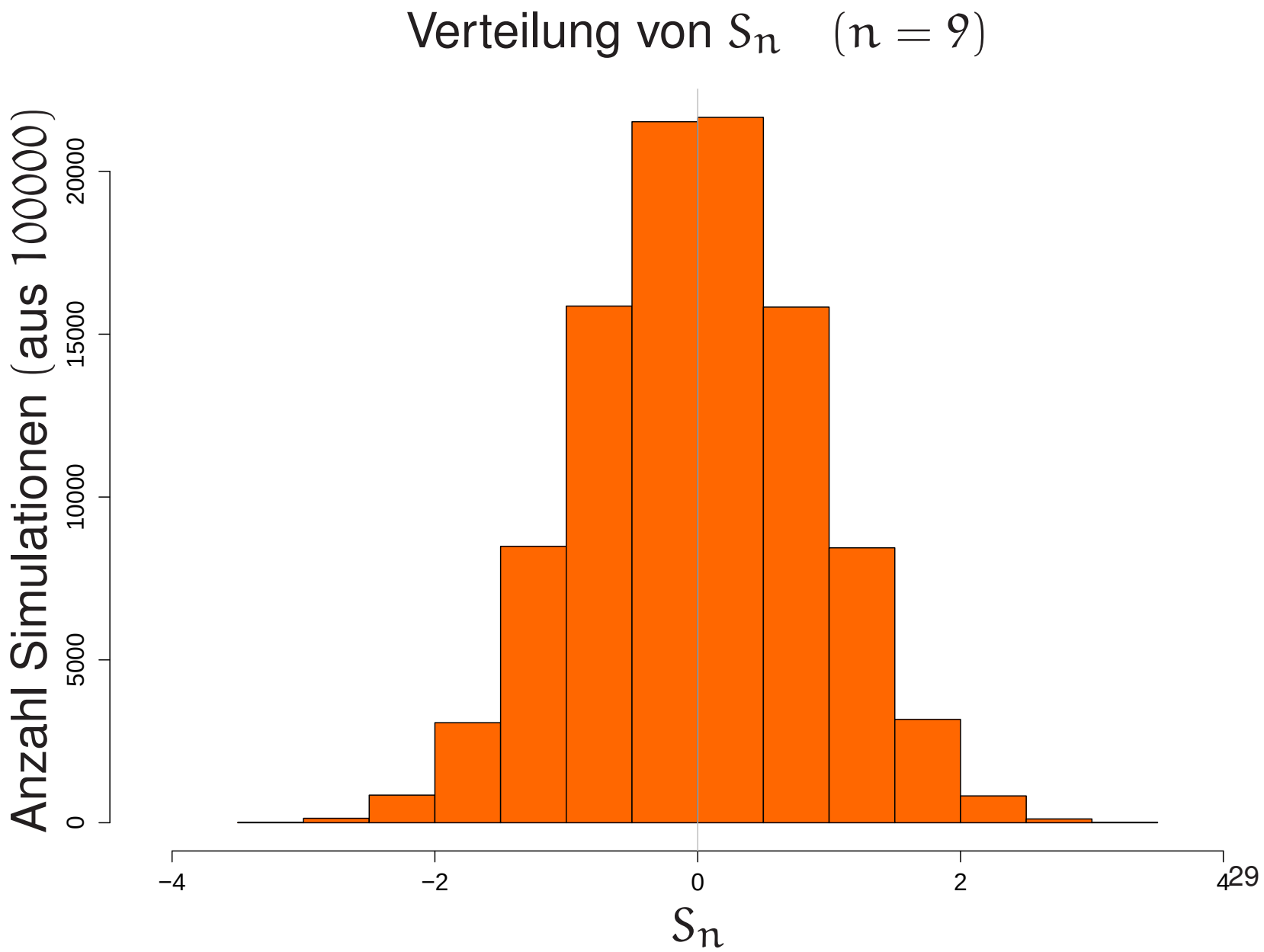

Verteilung von S_n $(n = 4)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
10000
12000
−2
−1
0
1
2
S_n

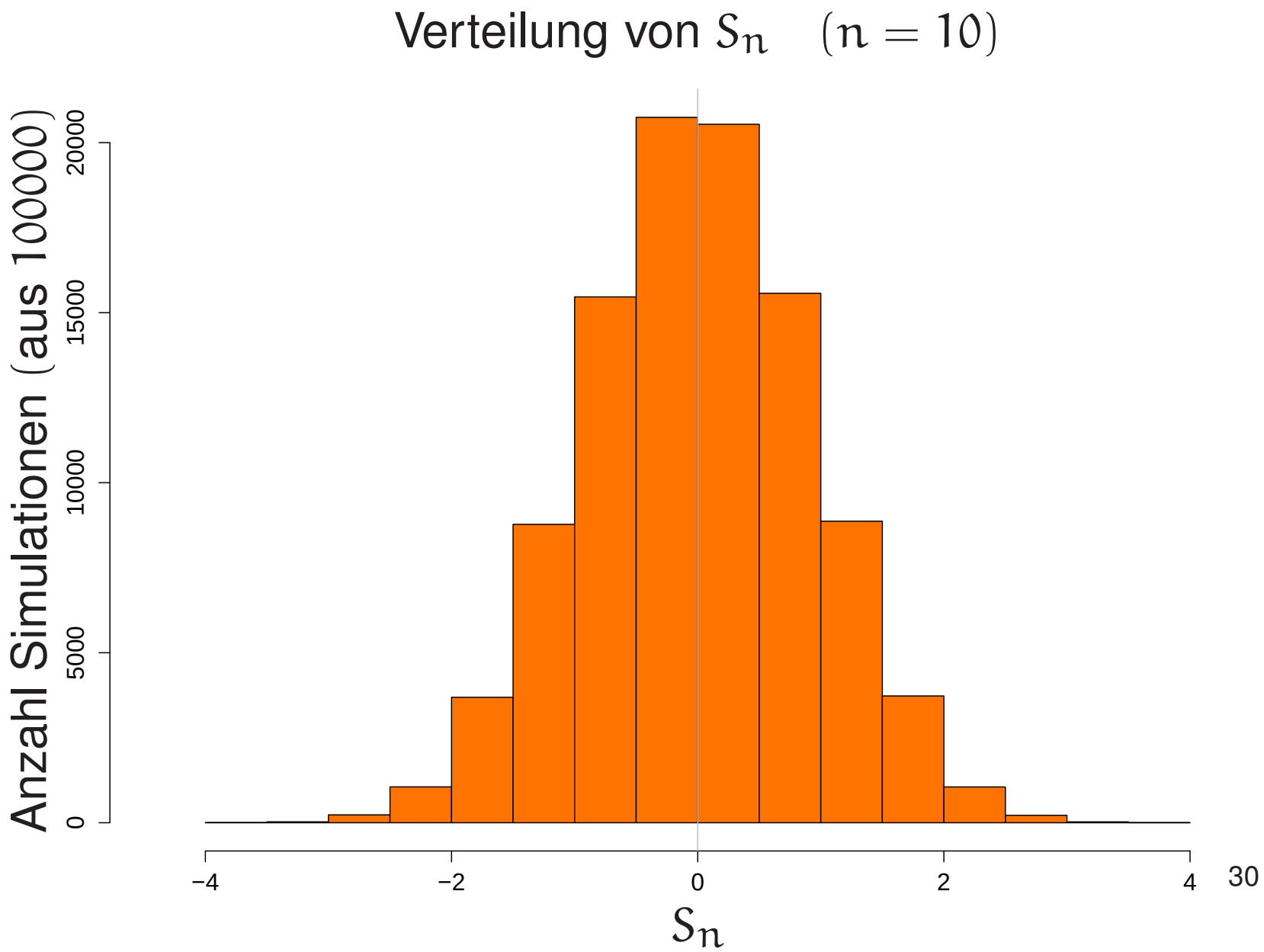

Verteilung von S_n $(n = 5)$
Anzahl Simulationen (aus 100000)
0
2000
4000
6000
8000
10000
12000
−2
−1
0
1
2
S_n

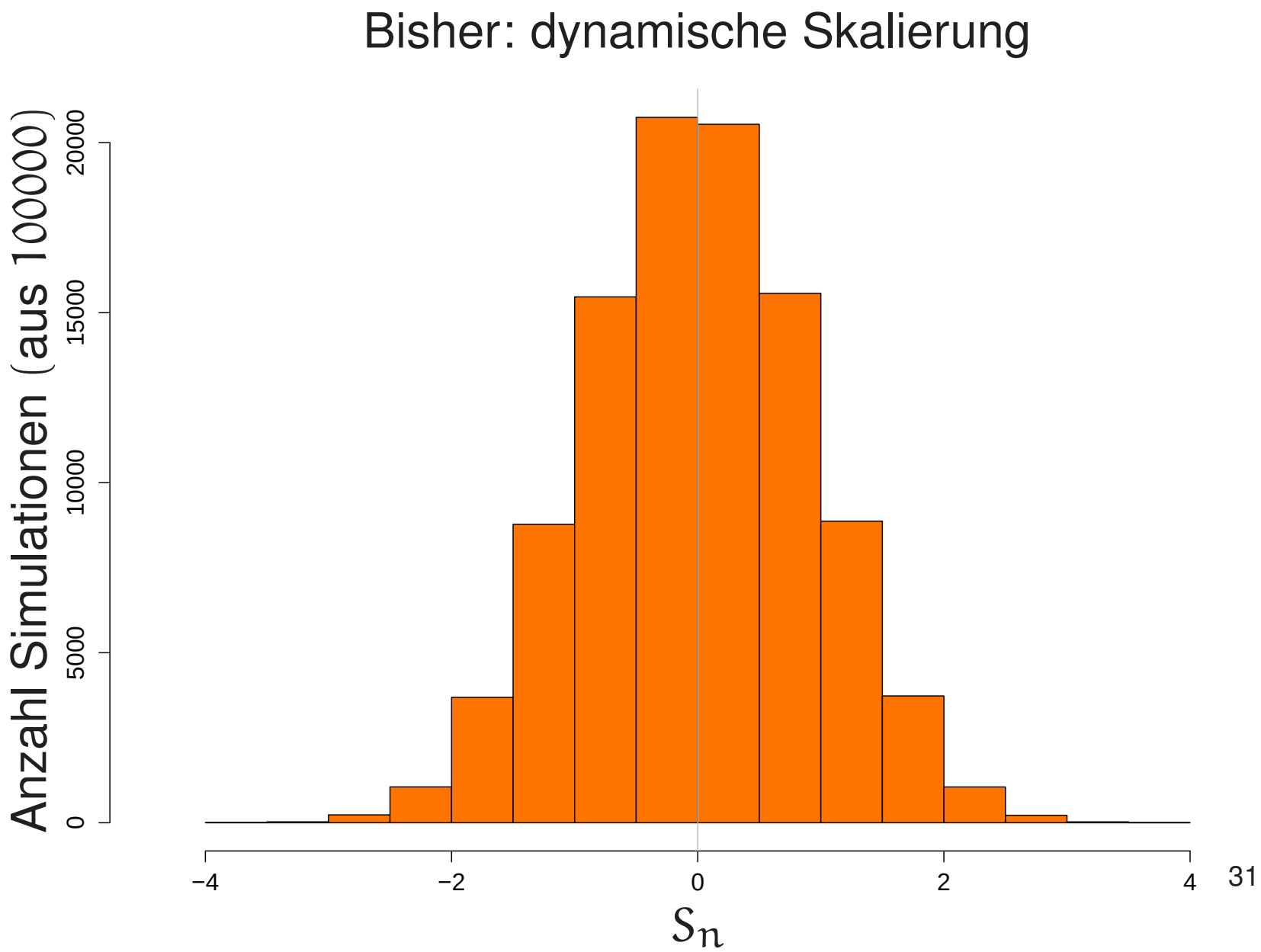

Verteilung von S_n $(n = 6)$
Anzahl Simulationen (aus 10000)
0
5000
10000
15000
20000
25000
−3
−2
−1
0
1
2
3
S_n

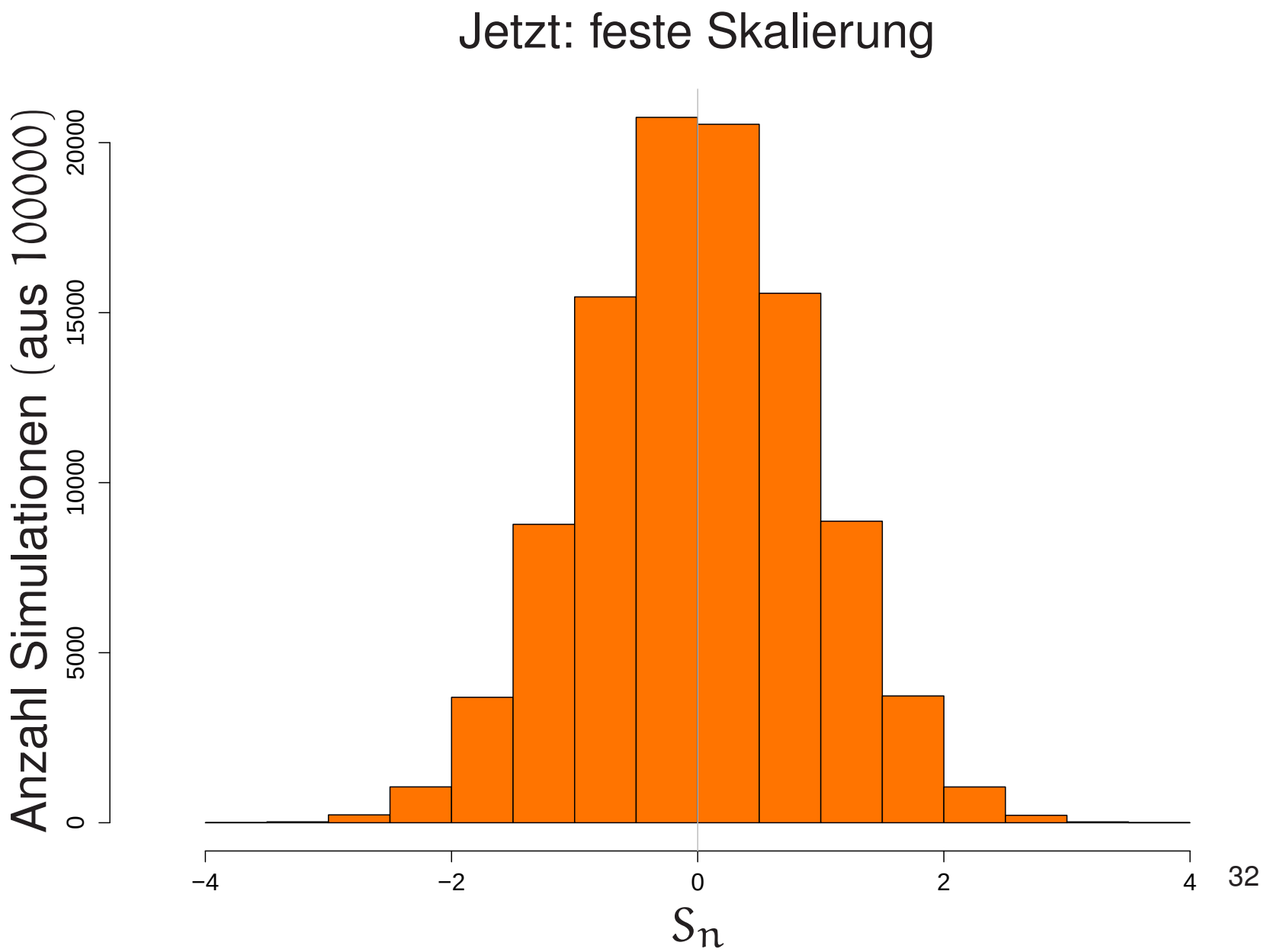

Verteilung von S_n $(n = 7)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
25000
−3
−2
−1
0
1
2
3
S_n

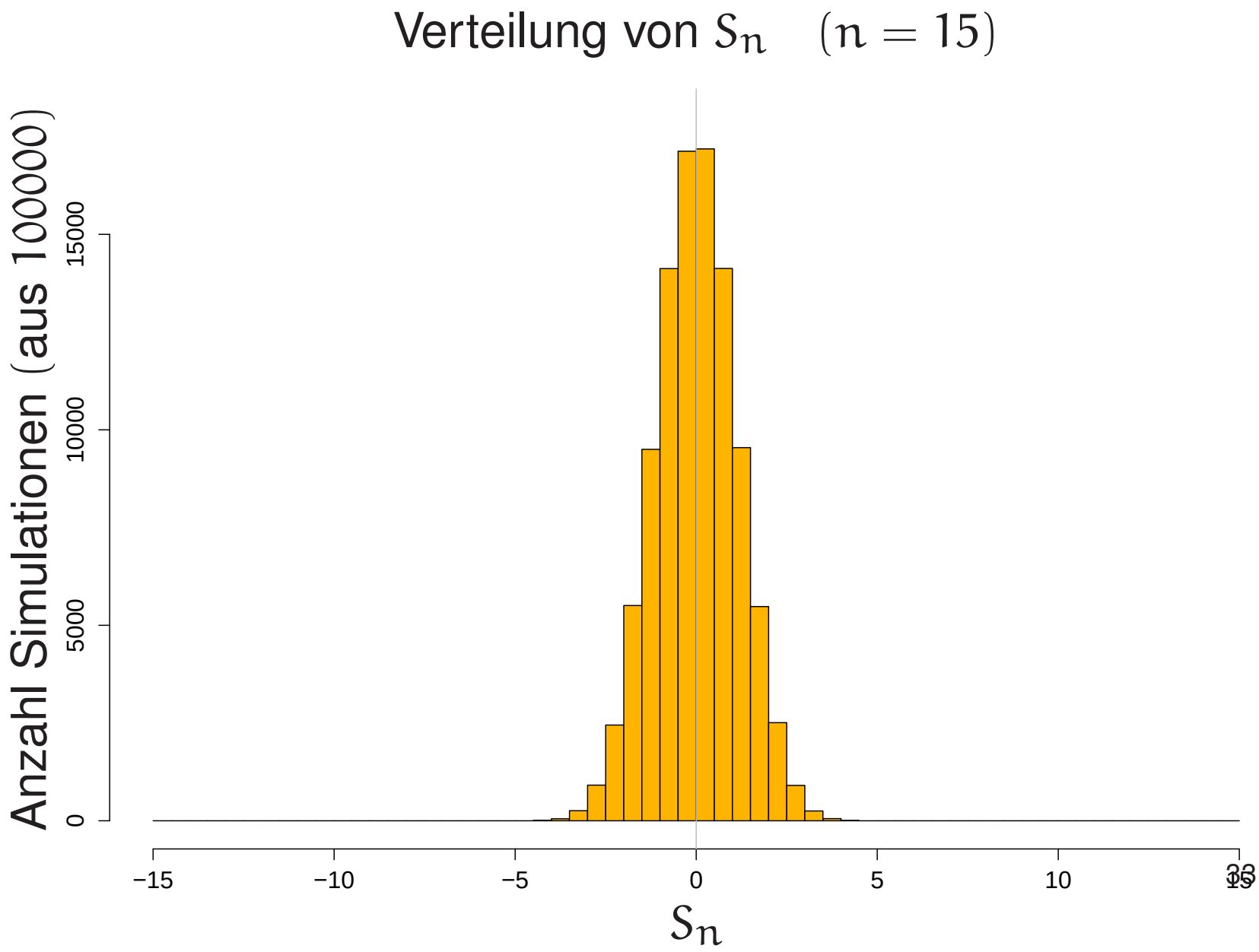

Verteilung von S_n $(n = 8)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

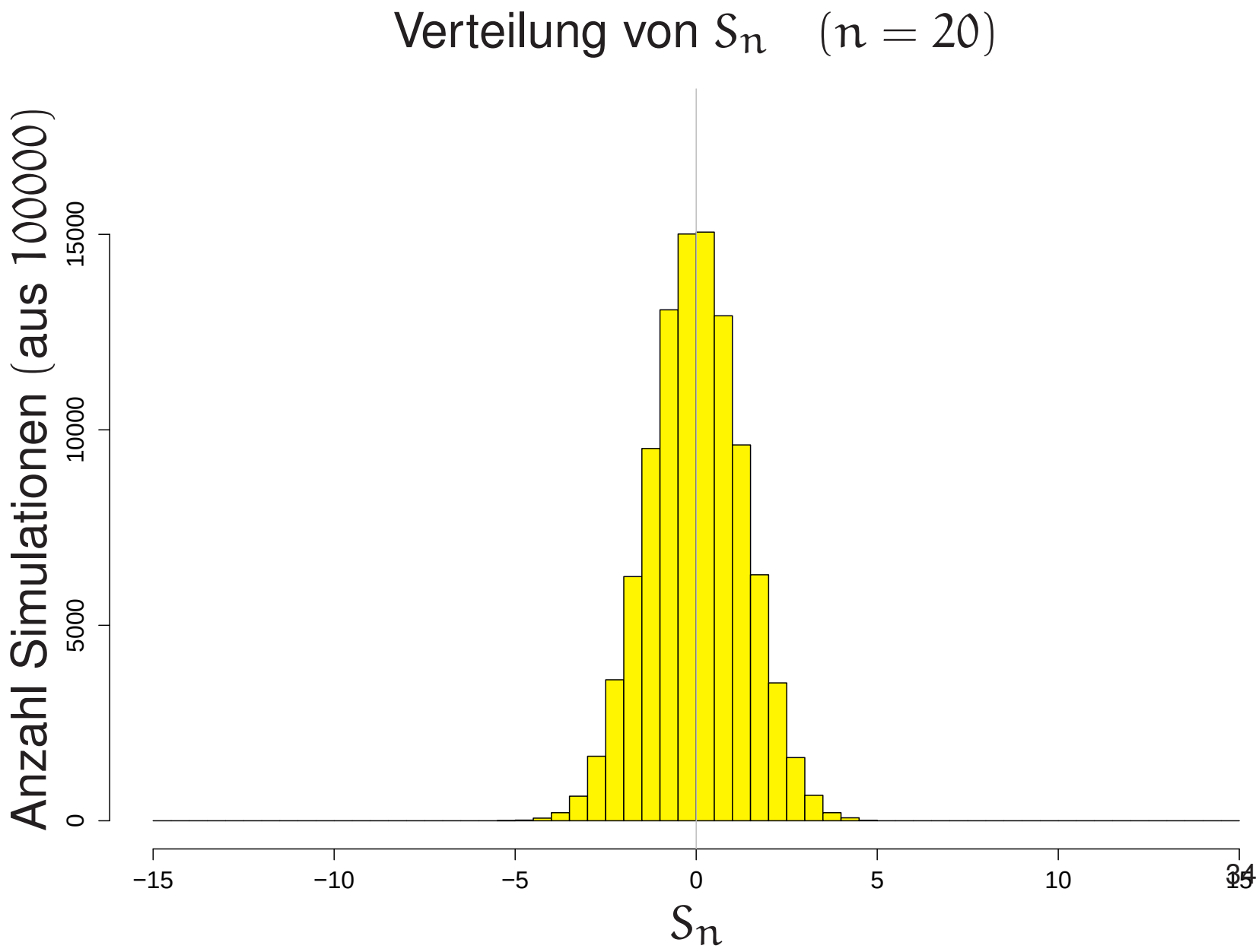

Verteilung von S_n $(n = 9)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

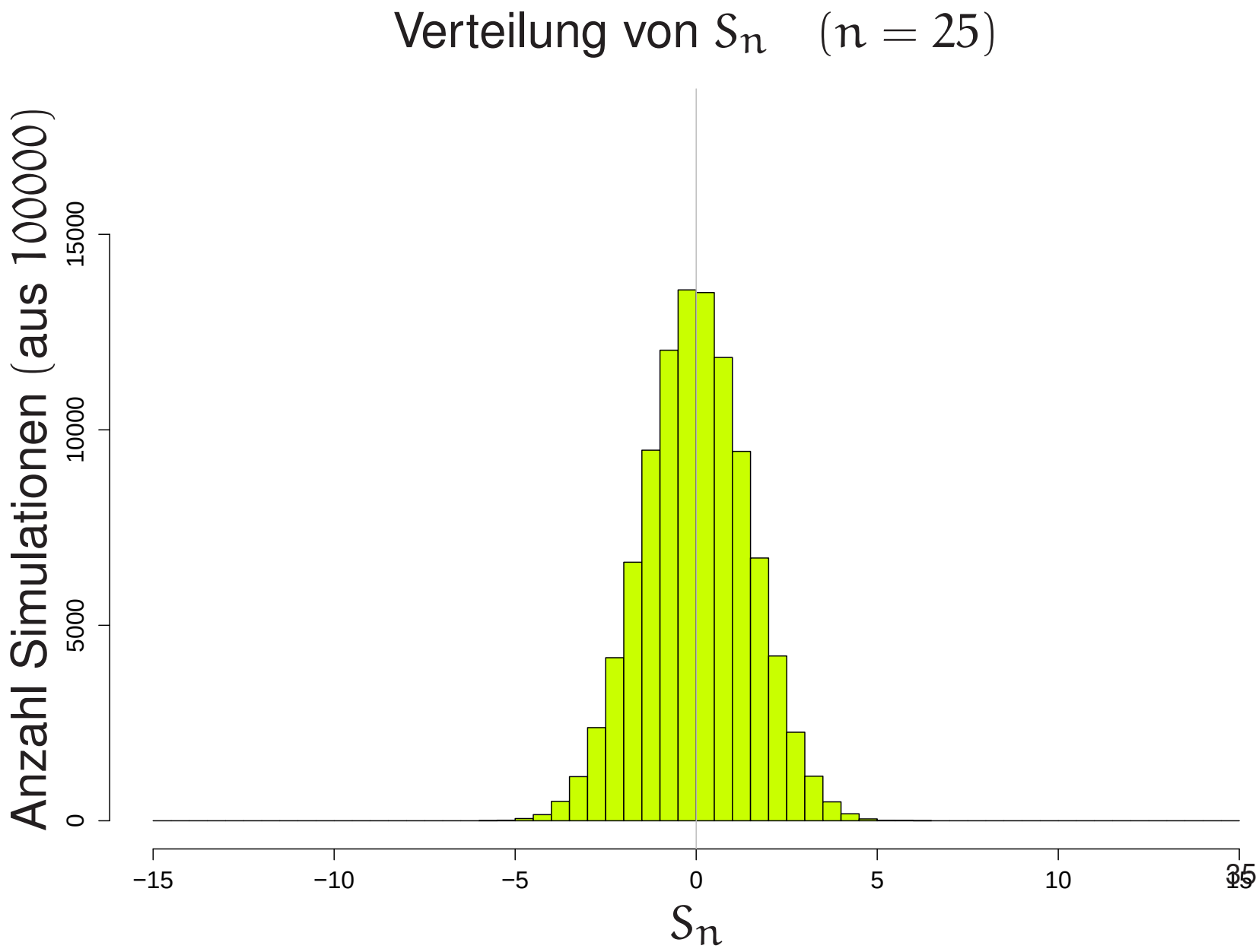

Verteilung von S_n $(n = 10)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
30
S_n

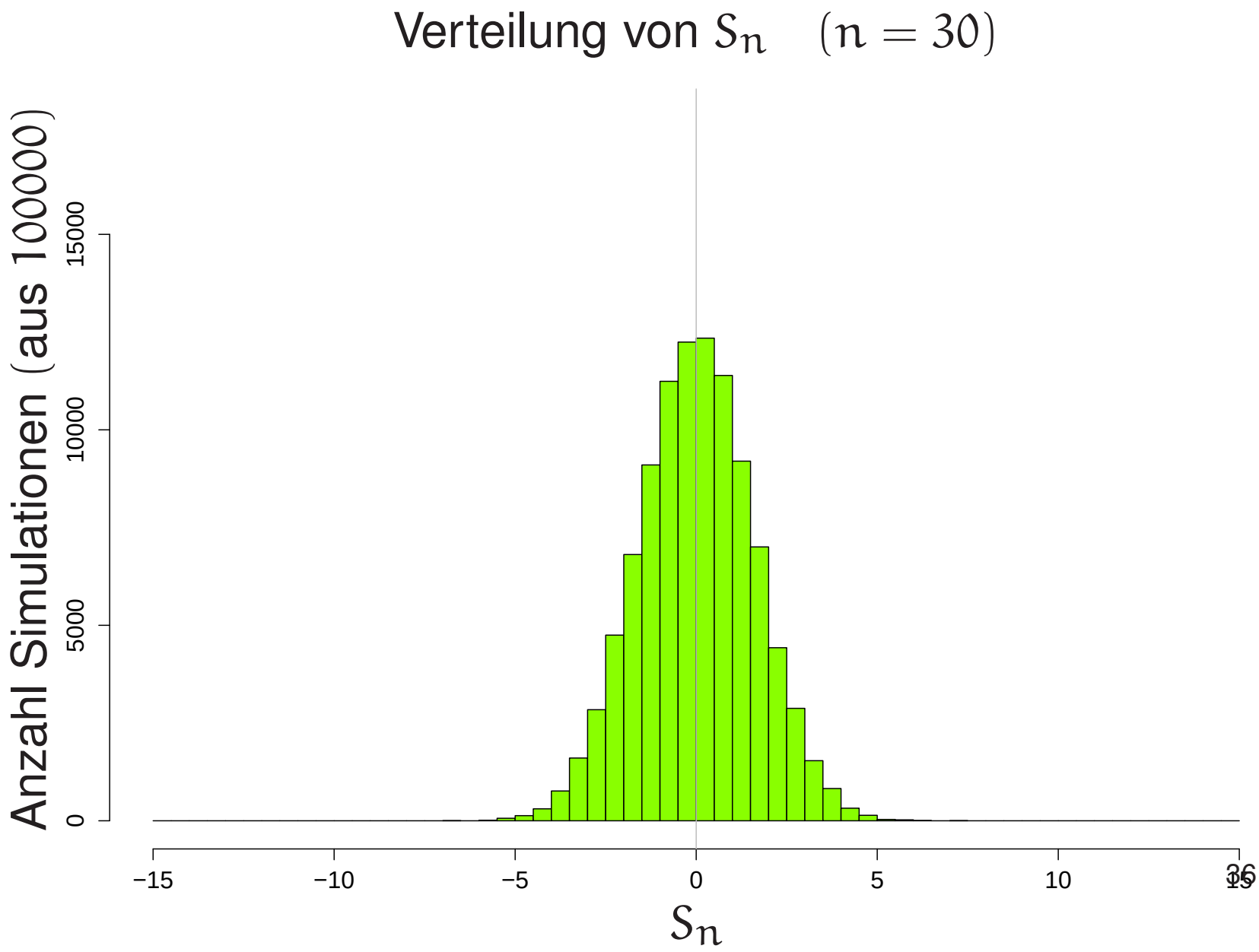

Bisher: dynamische Skalierung
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

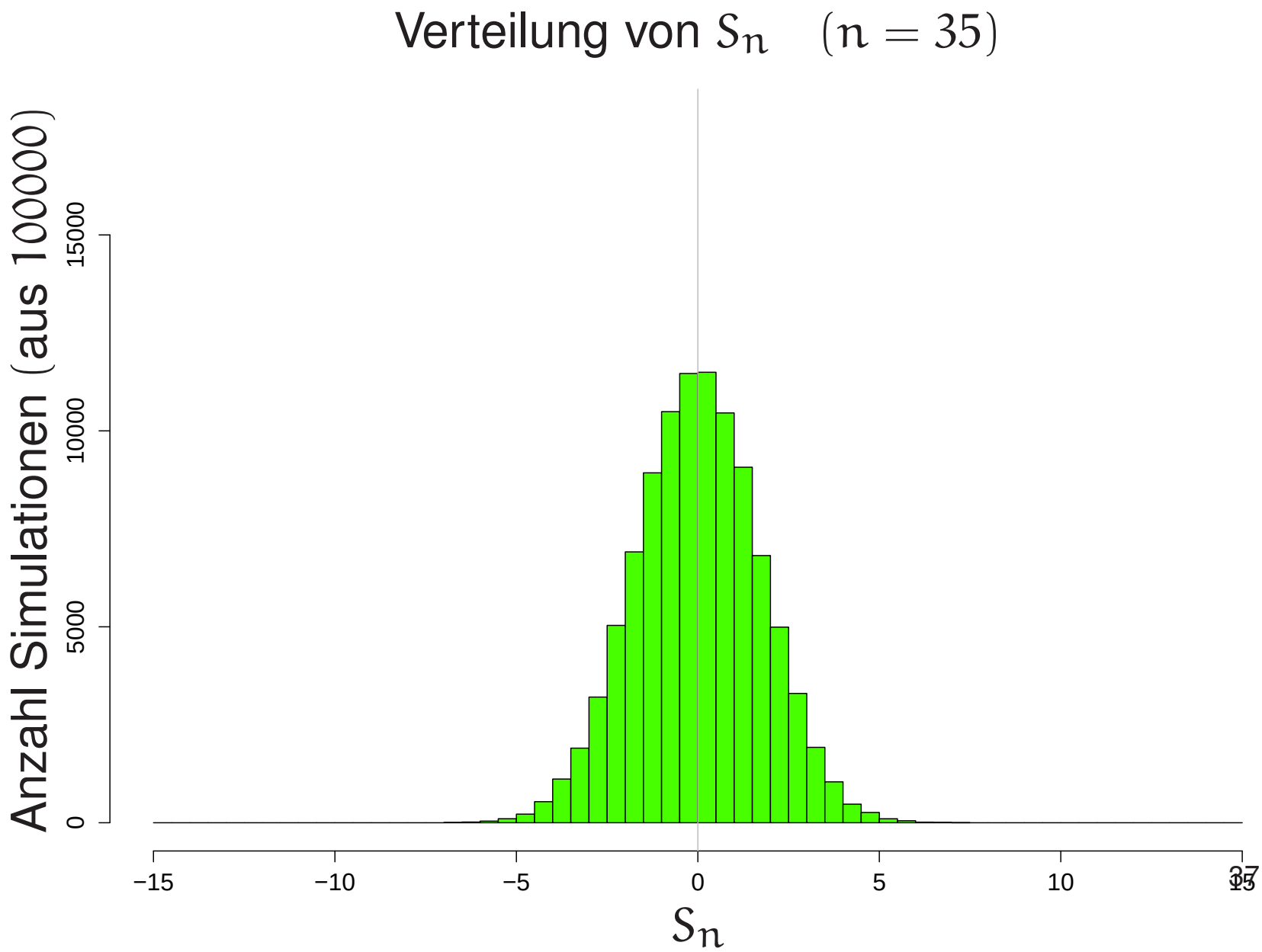

Jetzt: feste Skalierung
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
20000
−4
−2
0
2
4
S_n

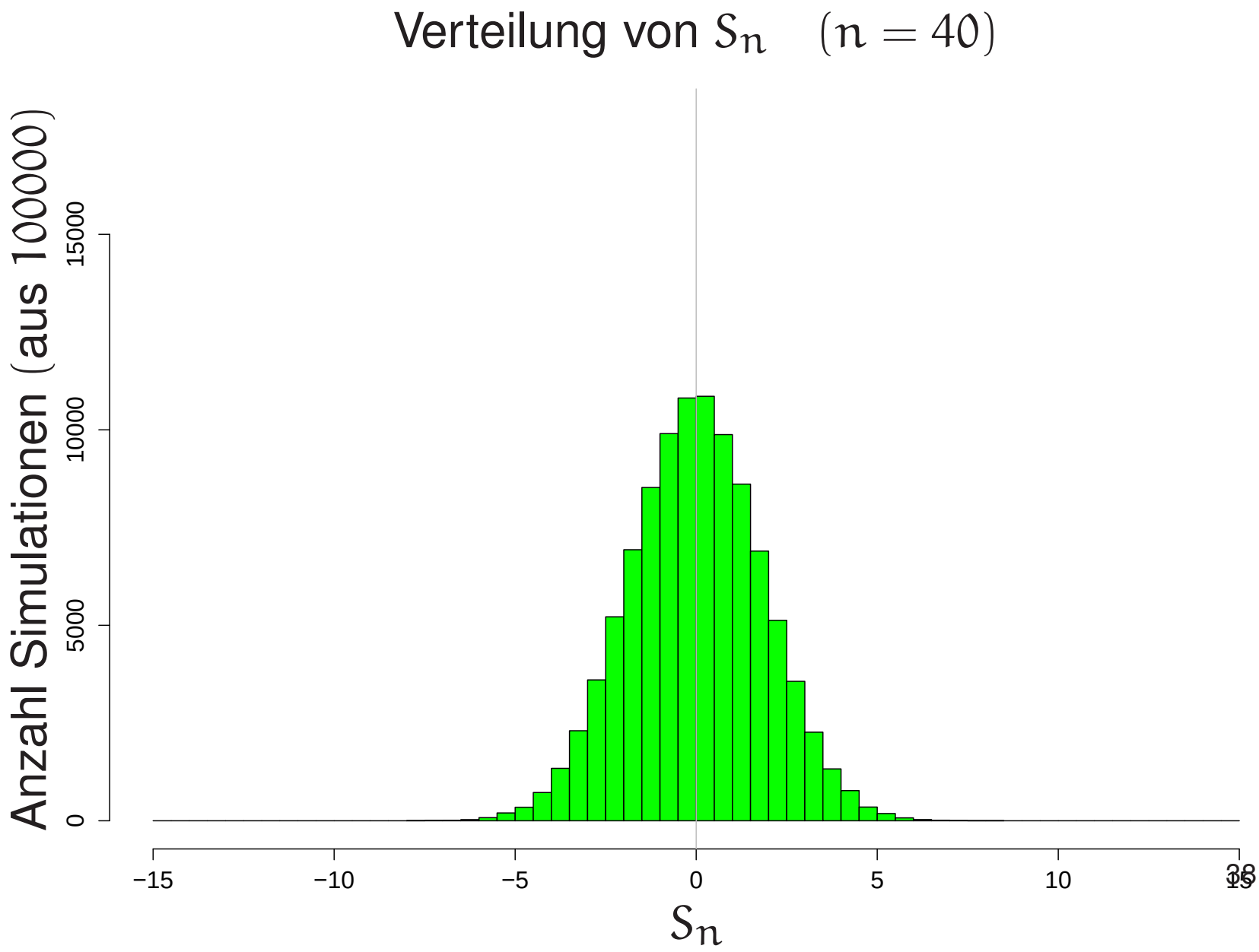

Verteilung von S_n $(n = 15)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

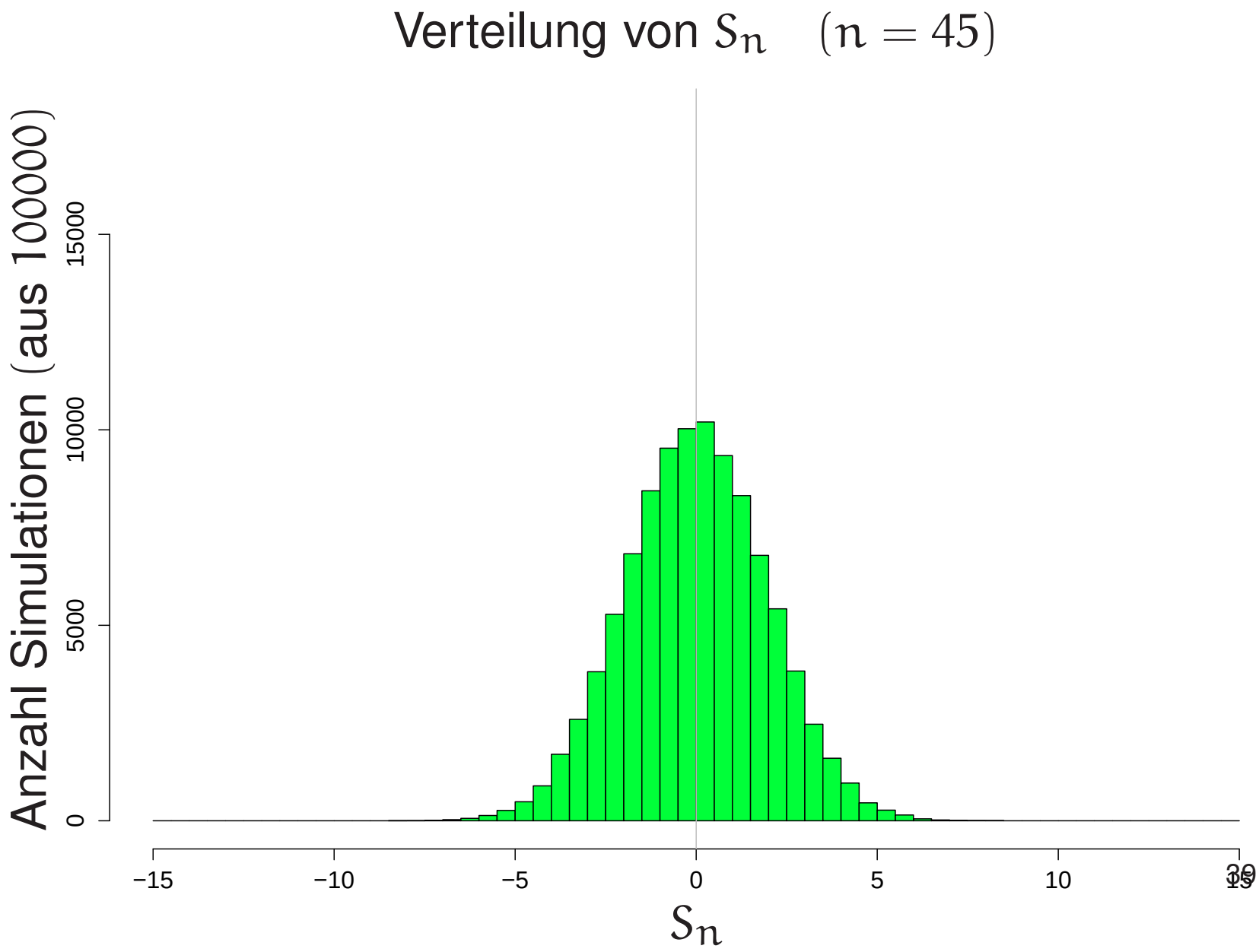

Verteilung von S_n $(n = 20)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

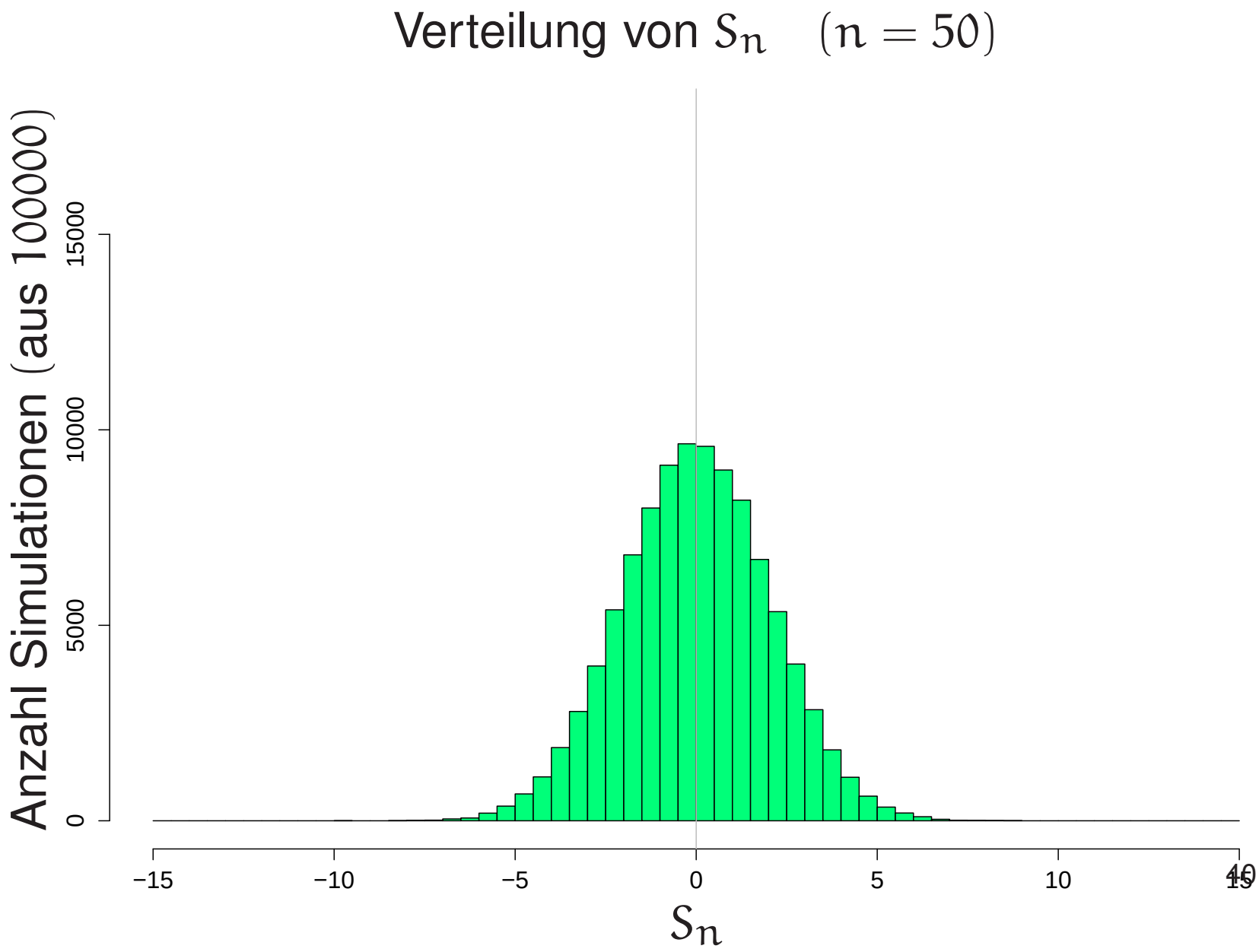

Verteilung von S_n $(n = 25)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

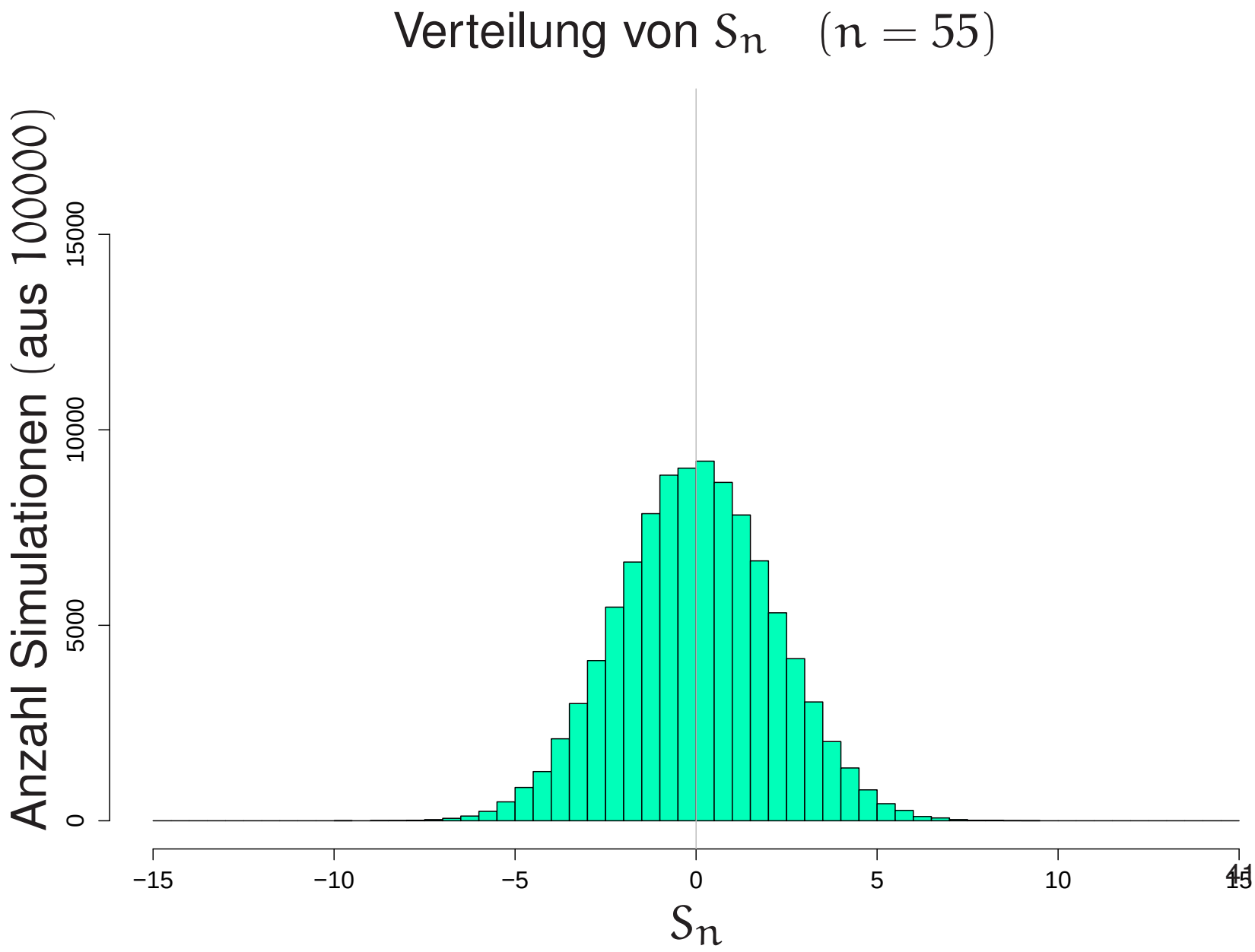

Verteilung von S_n $(n = 30)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

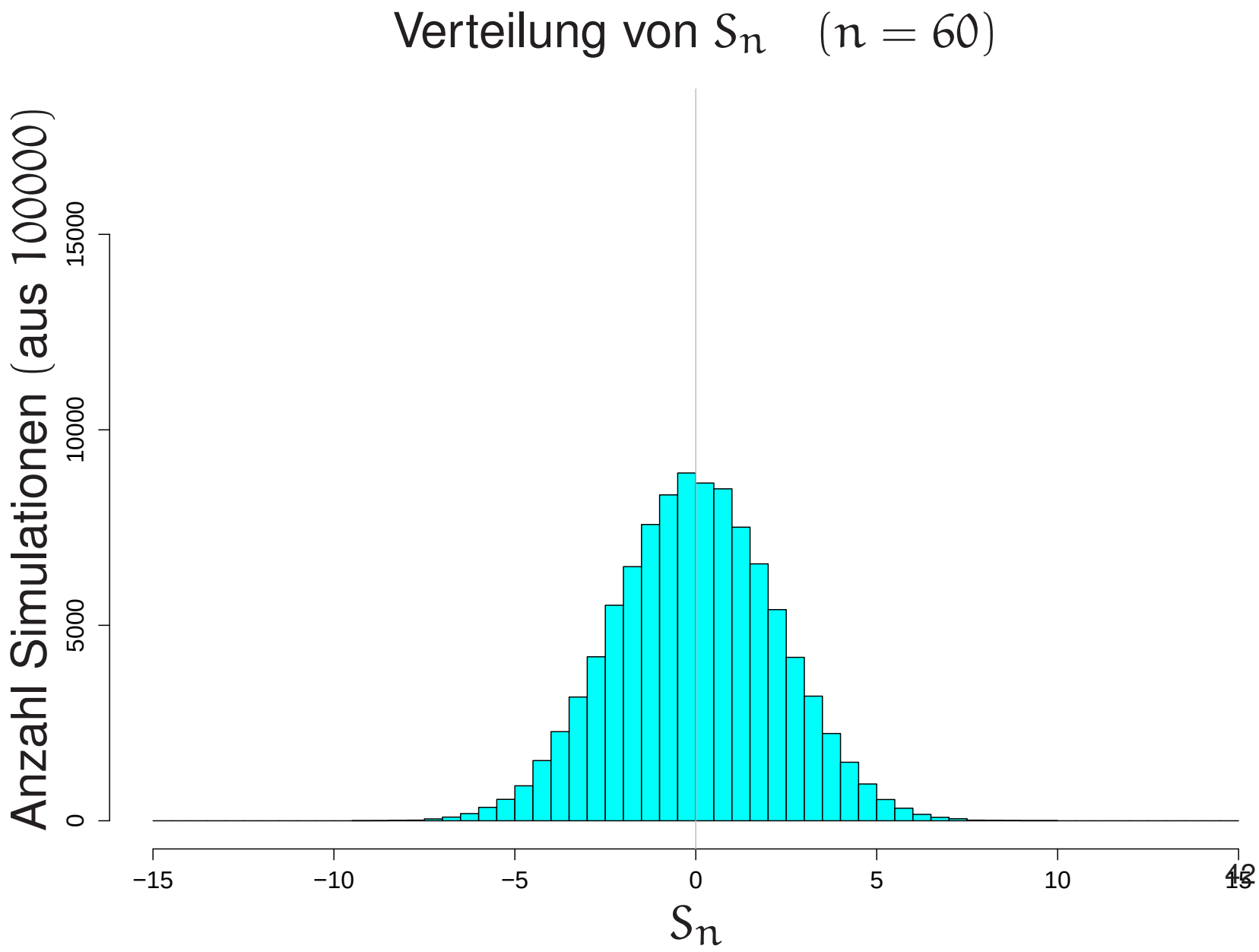

Verteilung von S_n $(n = 35)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

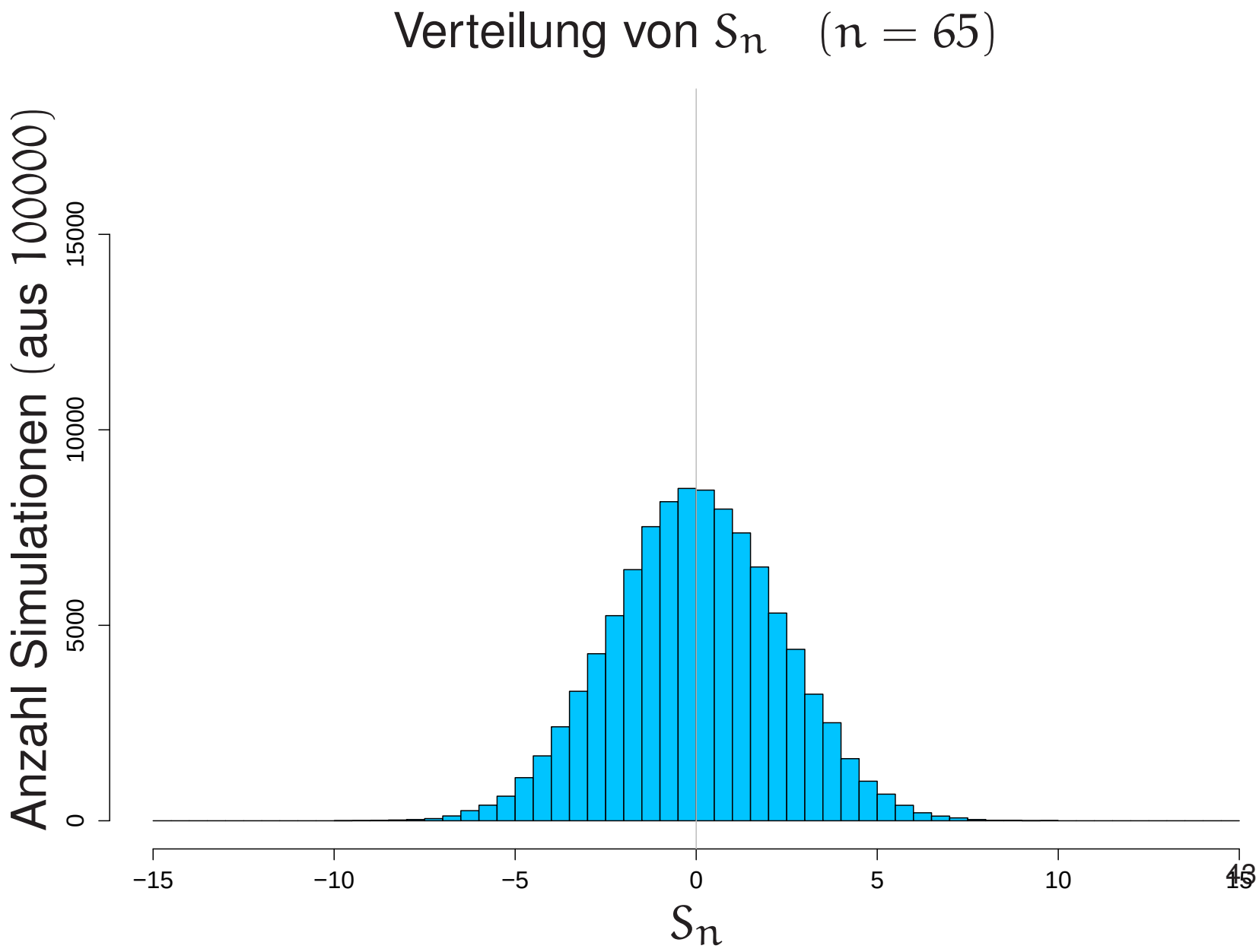

Verteilung von S_n $(n = 40)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

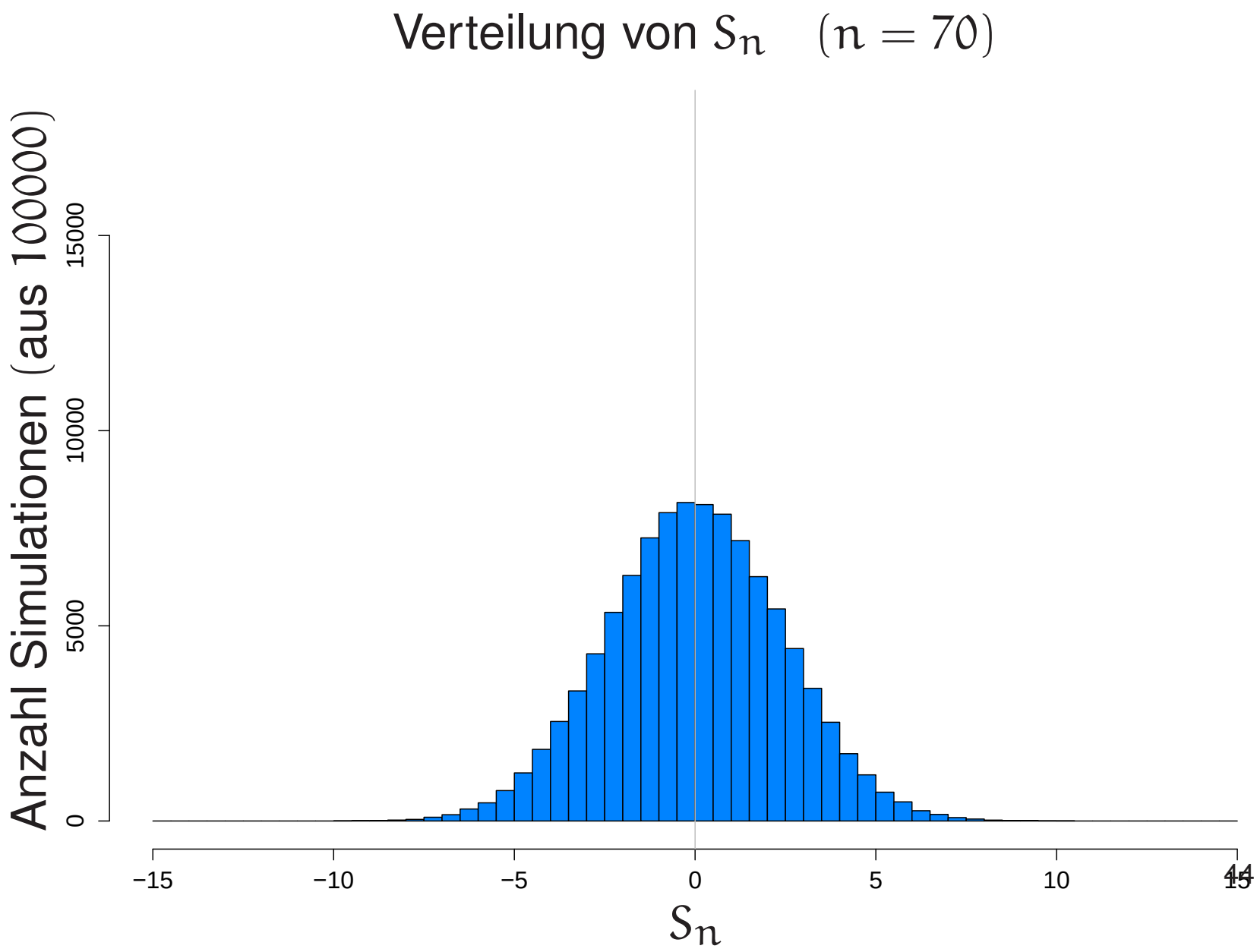

Verteilung von S_n $(n = 45)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

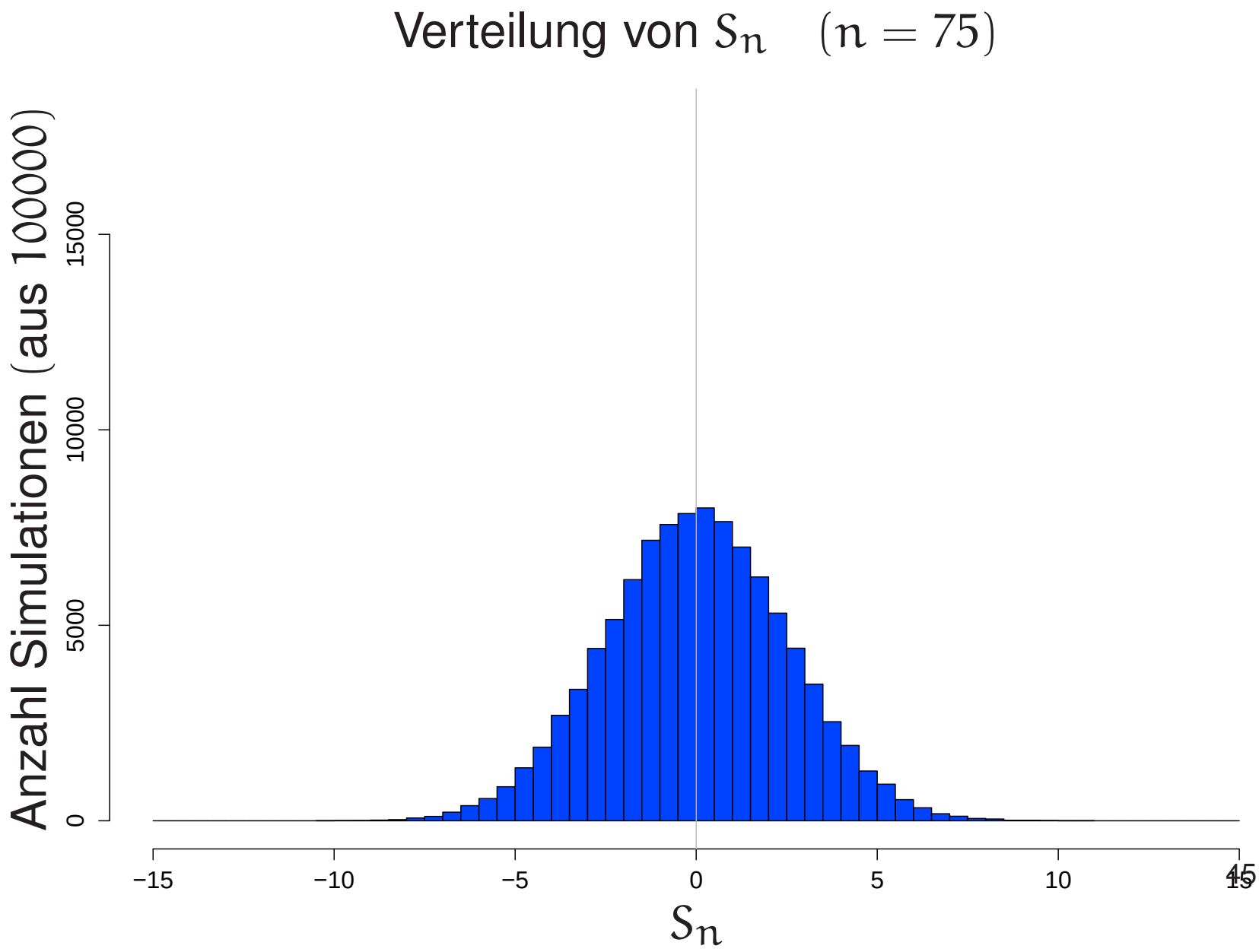

Verteilung von $S_n \quad (n = 50)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

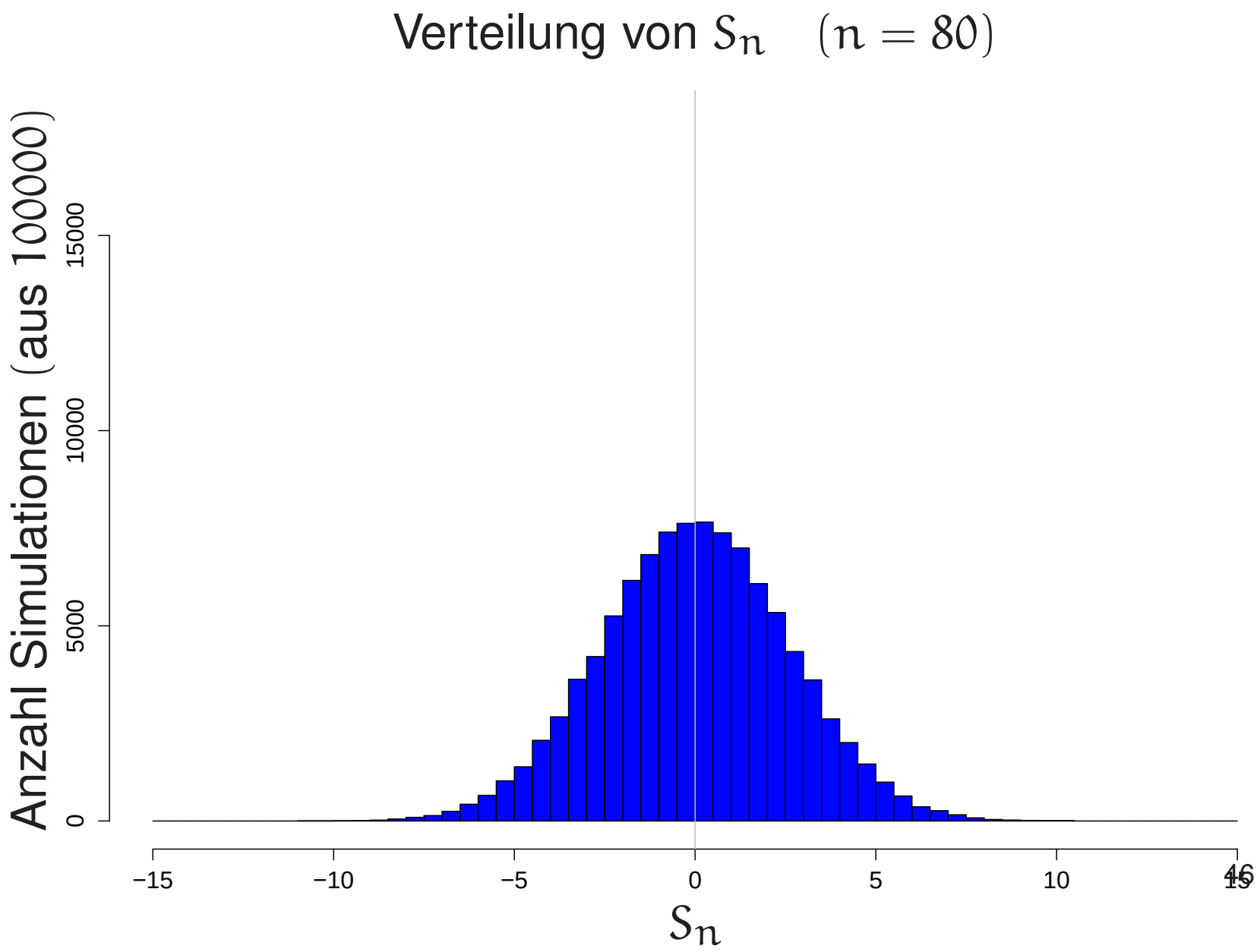

Verteilung von S_n $(n = 55)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

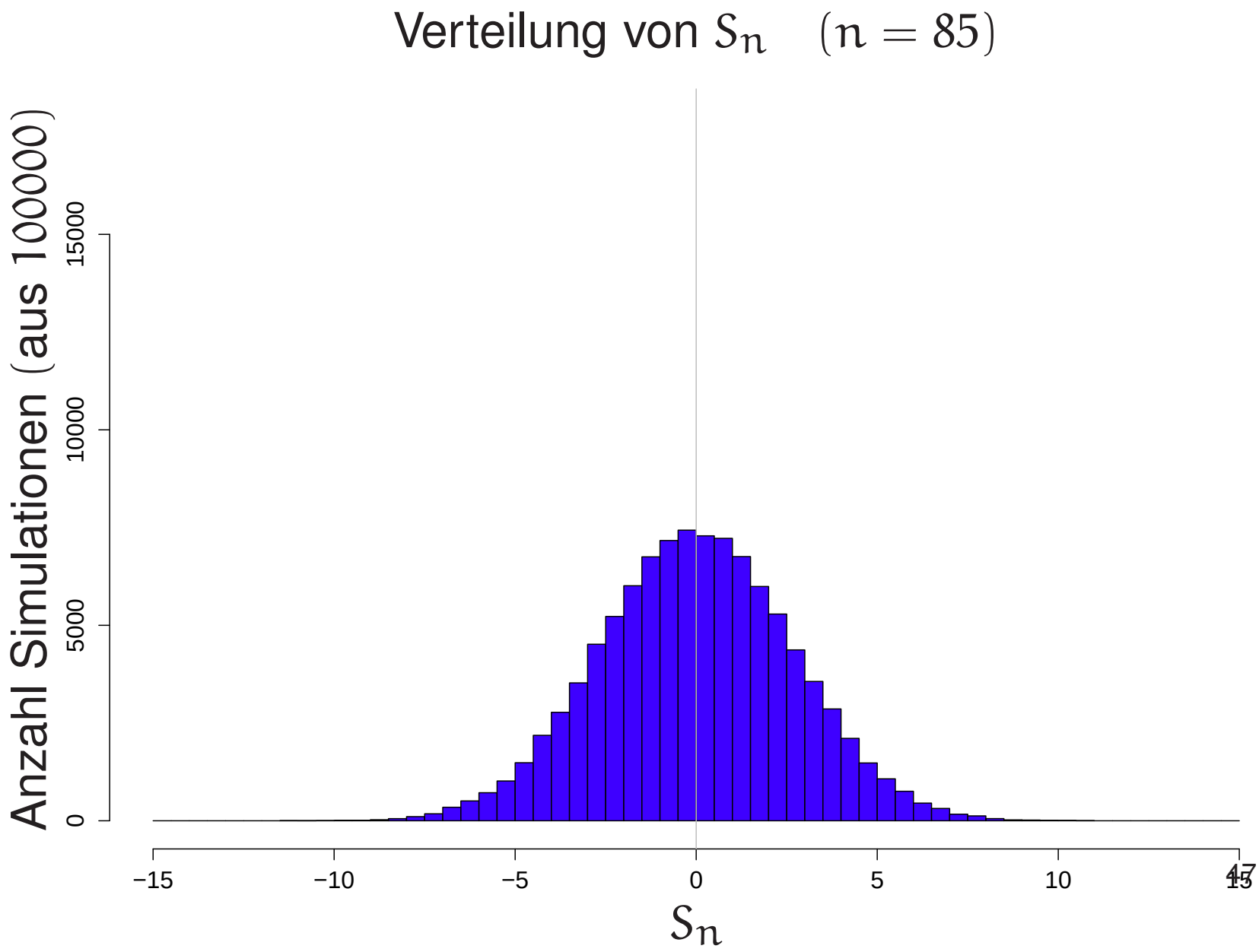

Verteilung von S_n $(n = 60)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

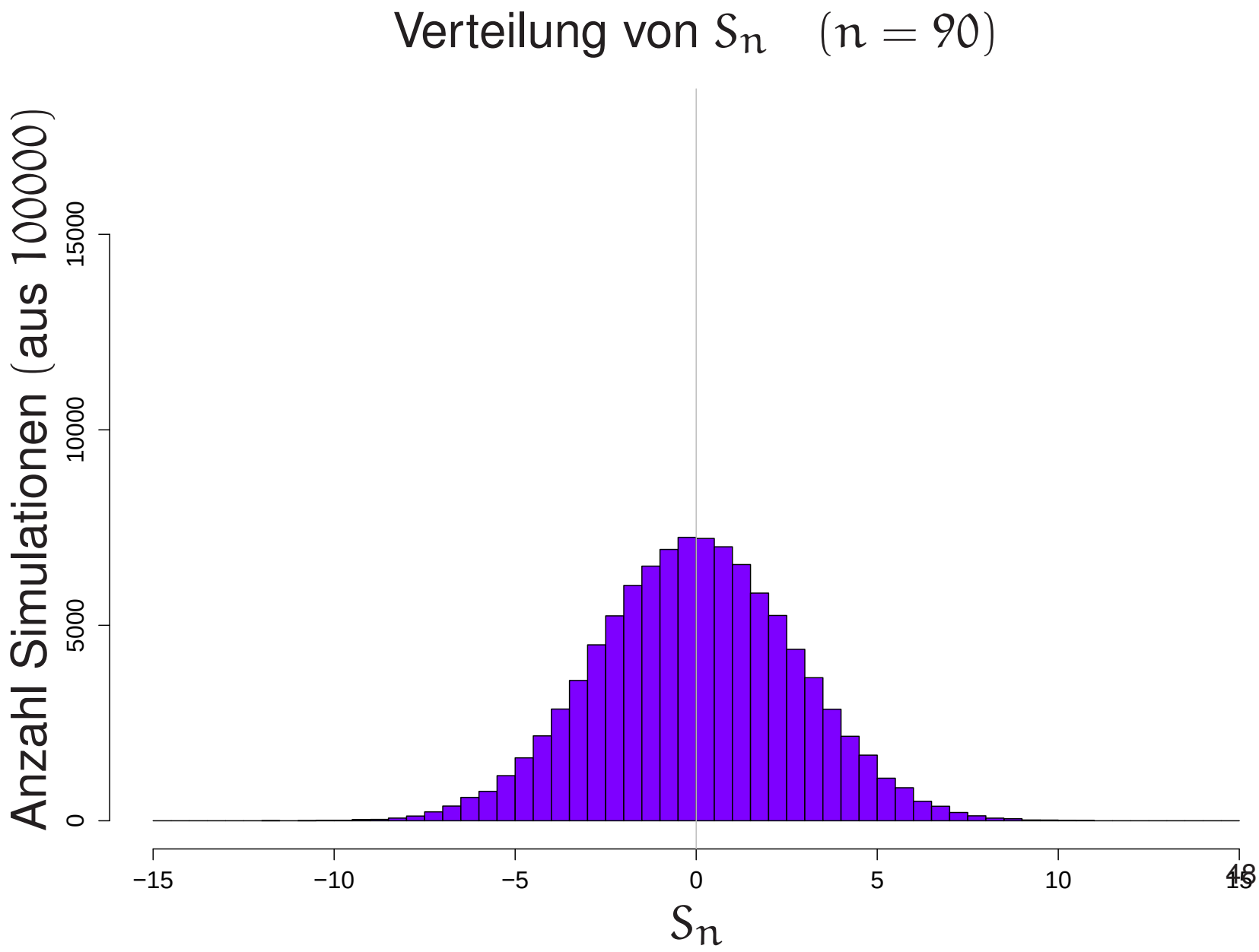

Verteilung von S_n $(n = 65)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

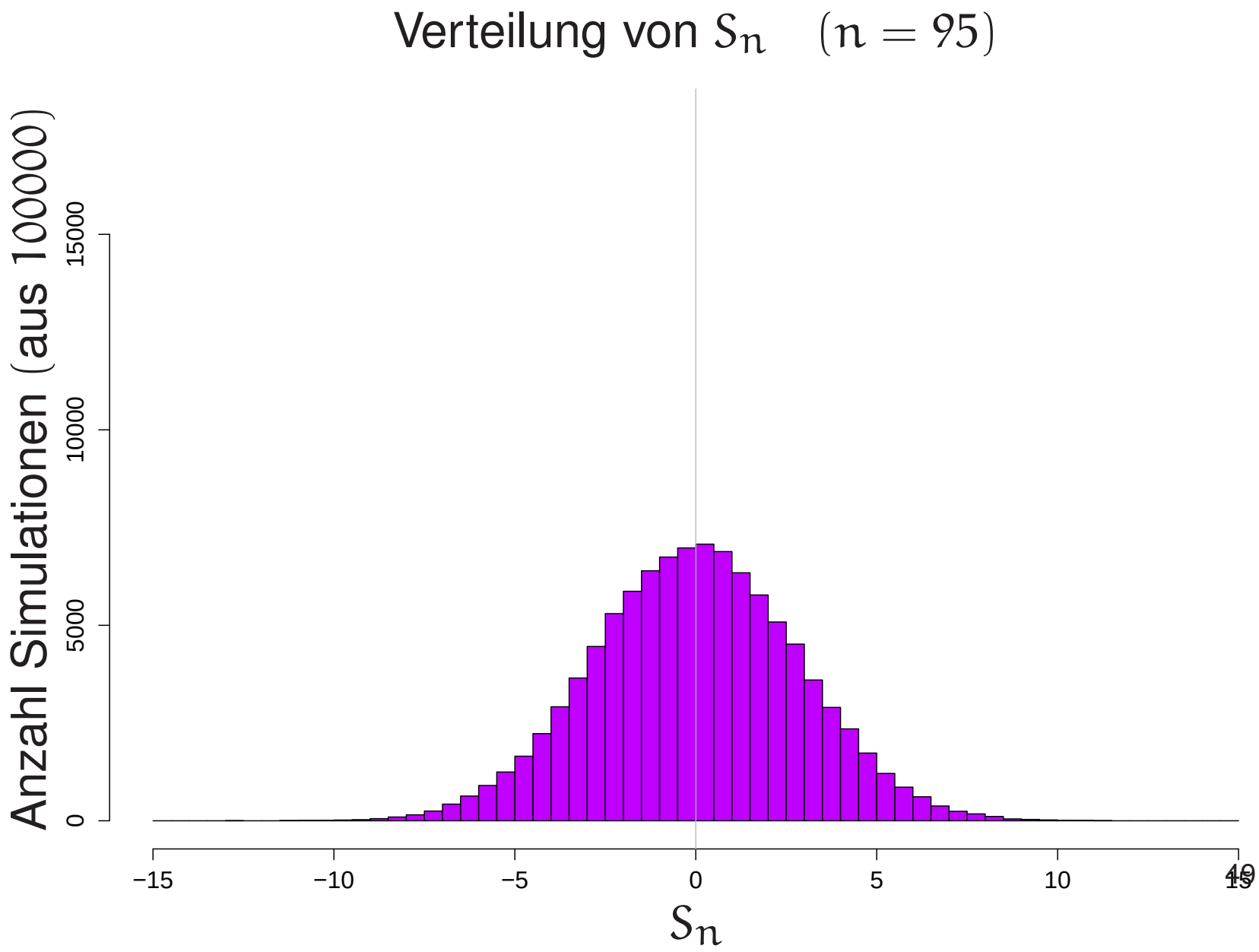

Verteilung von S_n $(n = 70)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

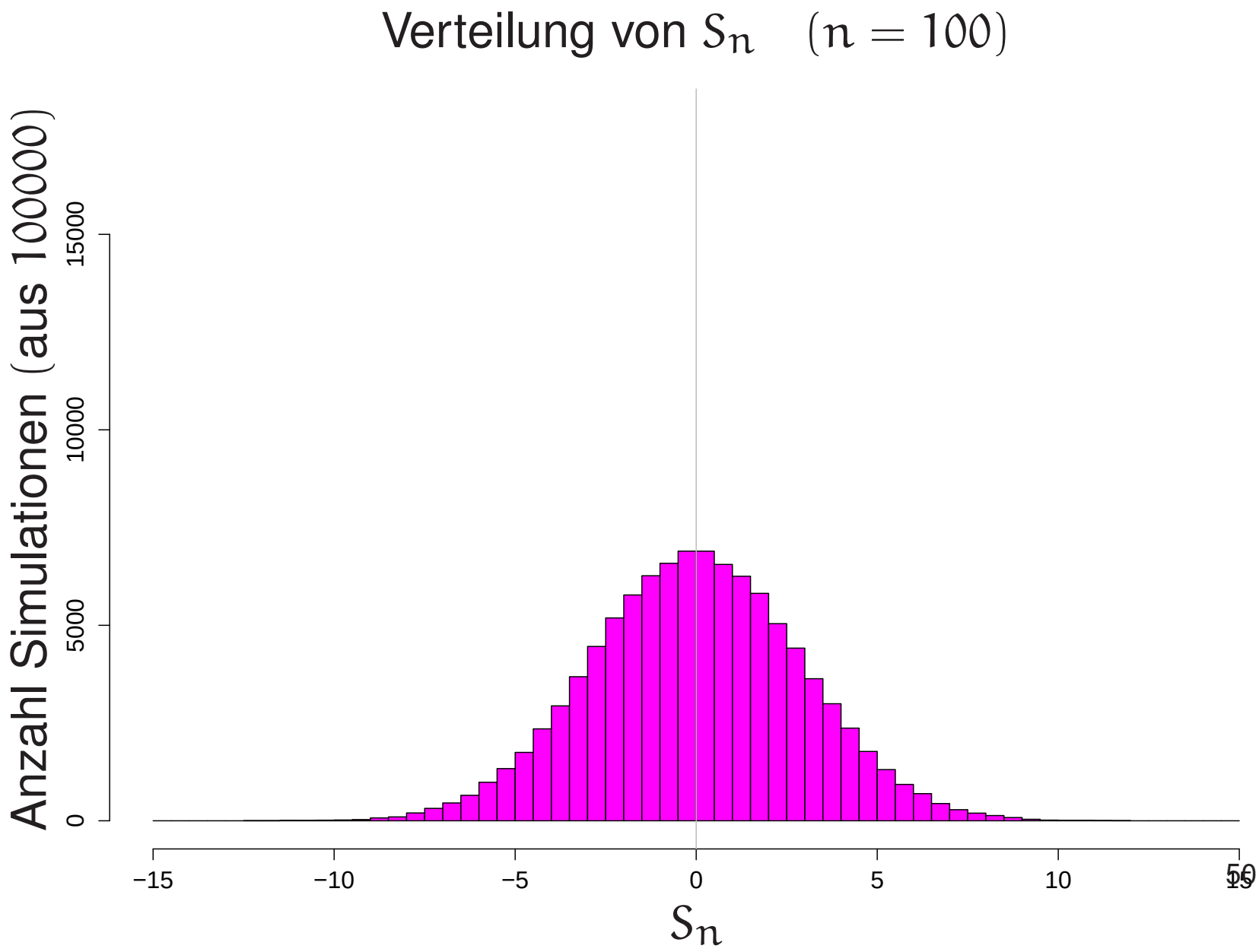

Verteilung von S_n $(n = 75)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


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

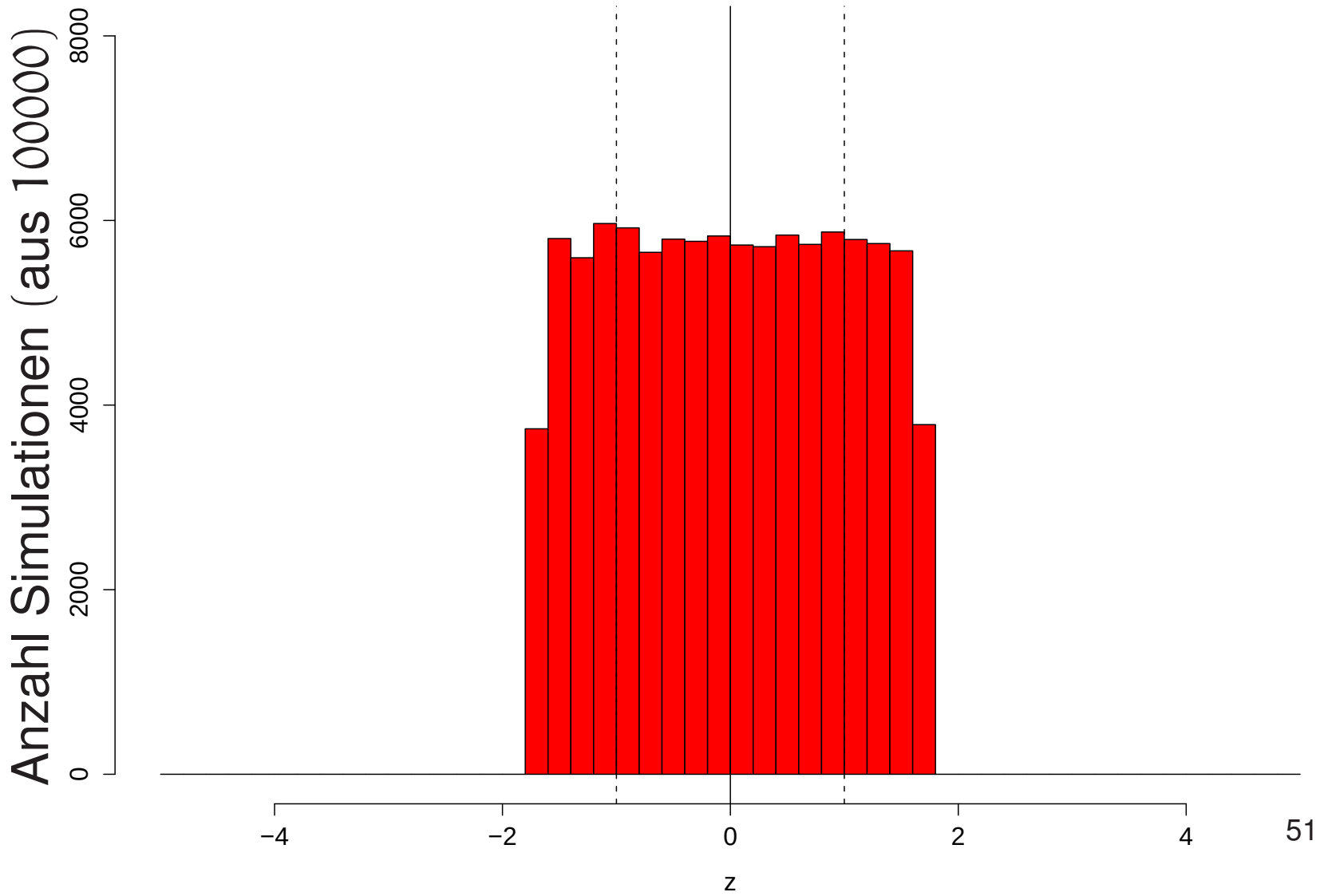

Verteilung von S_n $(n = 85)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n


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

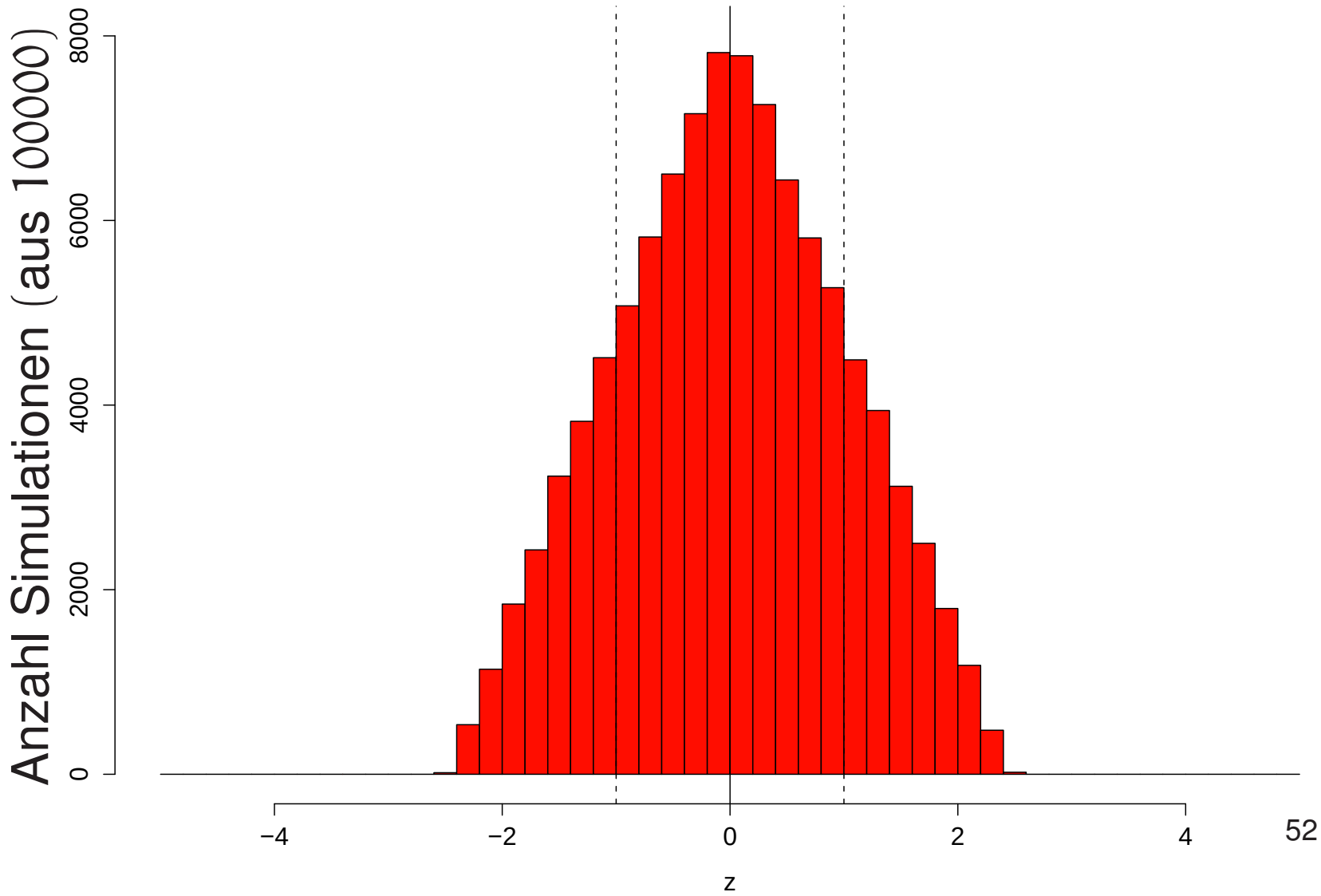

Verteilung von S_n $(n = 100)$
Anzahl Simulationen (aus 100000)
0
5000
10000
15000
−15
−10
−5
0
5
10
15
S_n

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

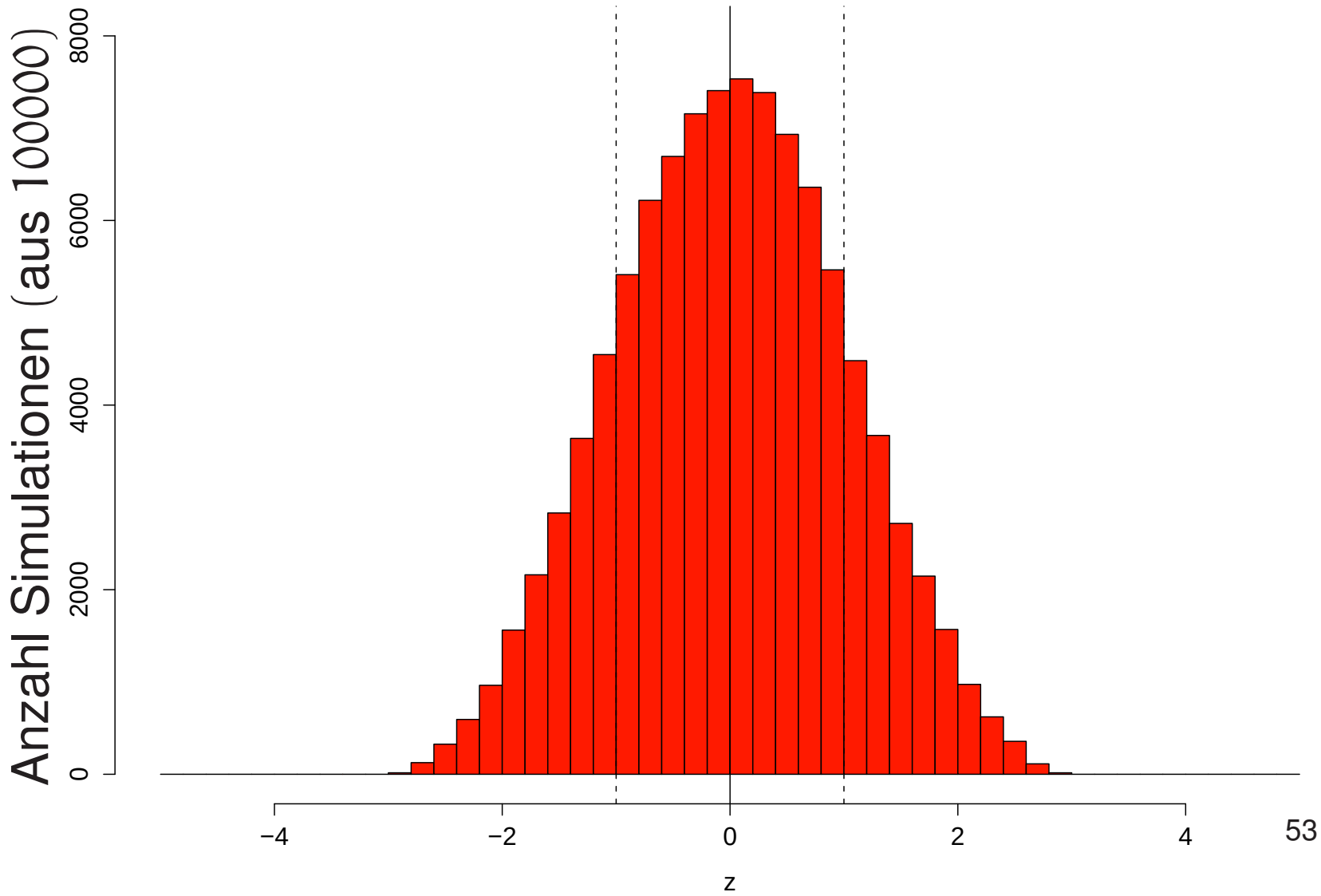


51

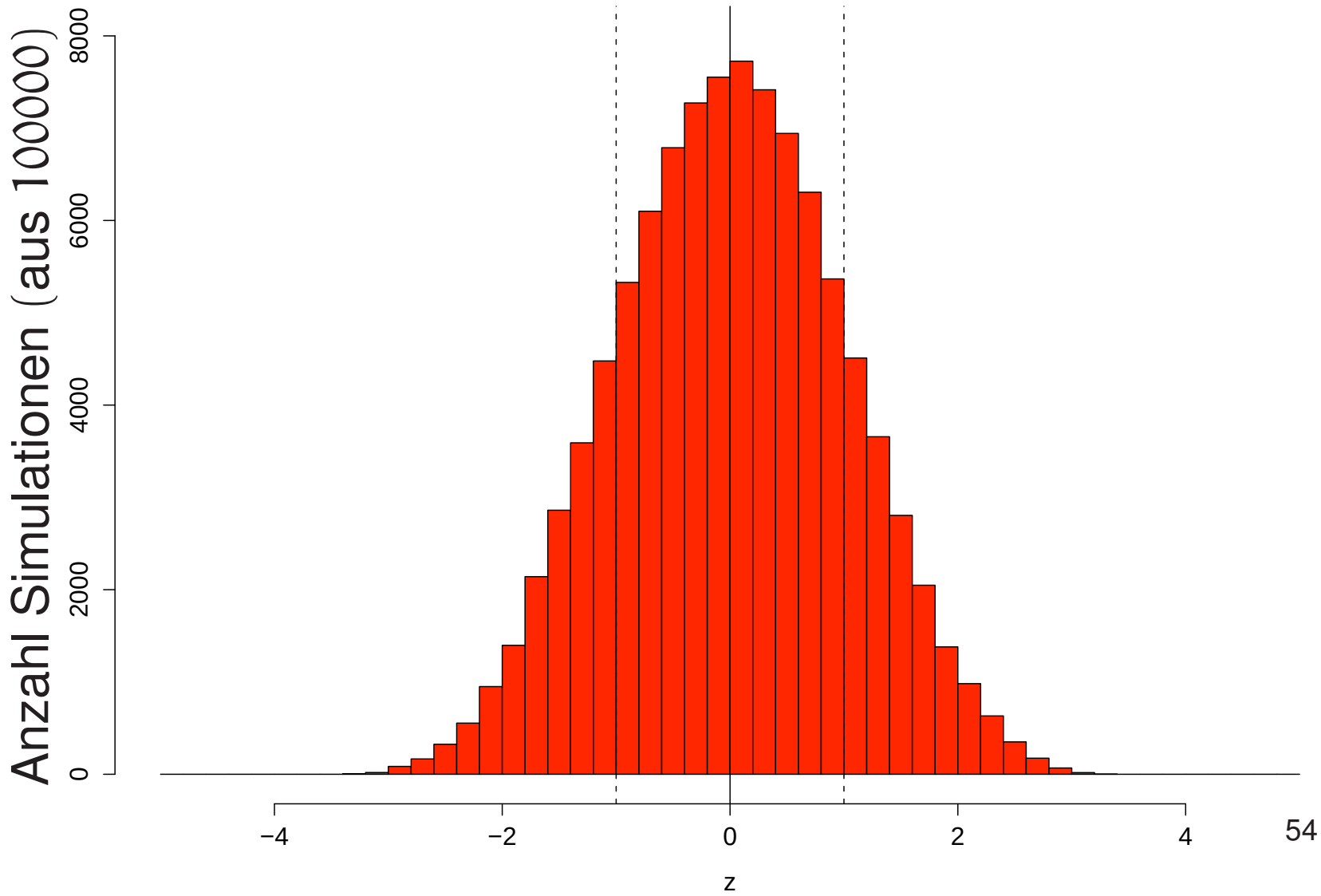
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 2)$



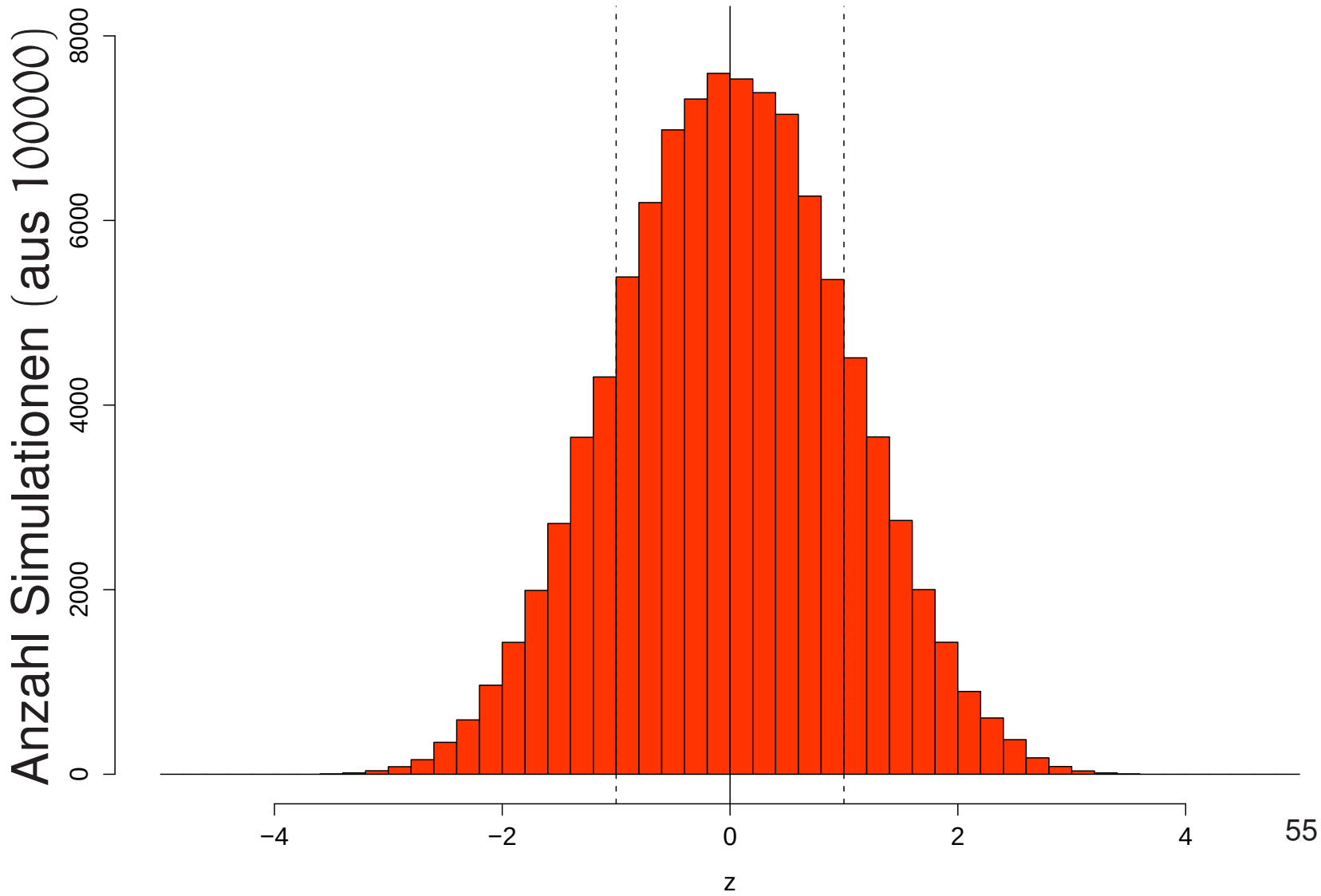
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 3)$



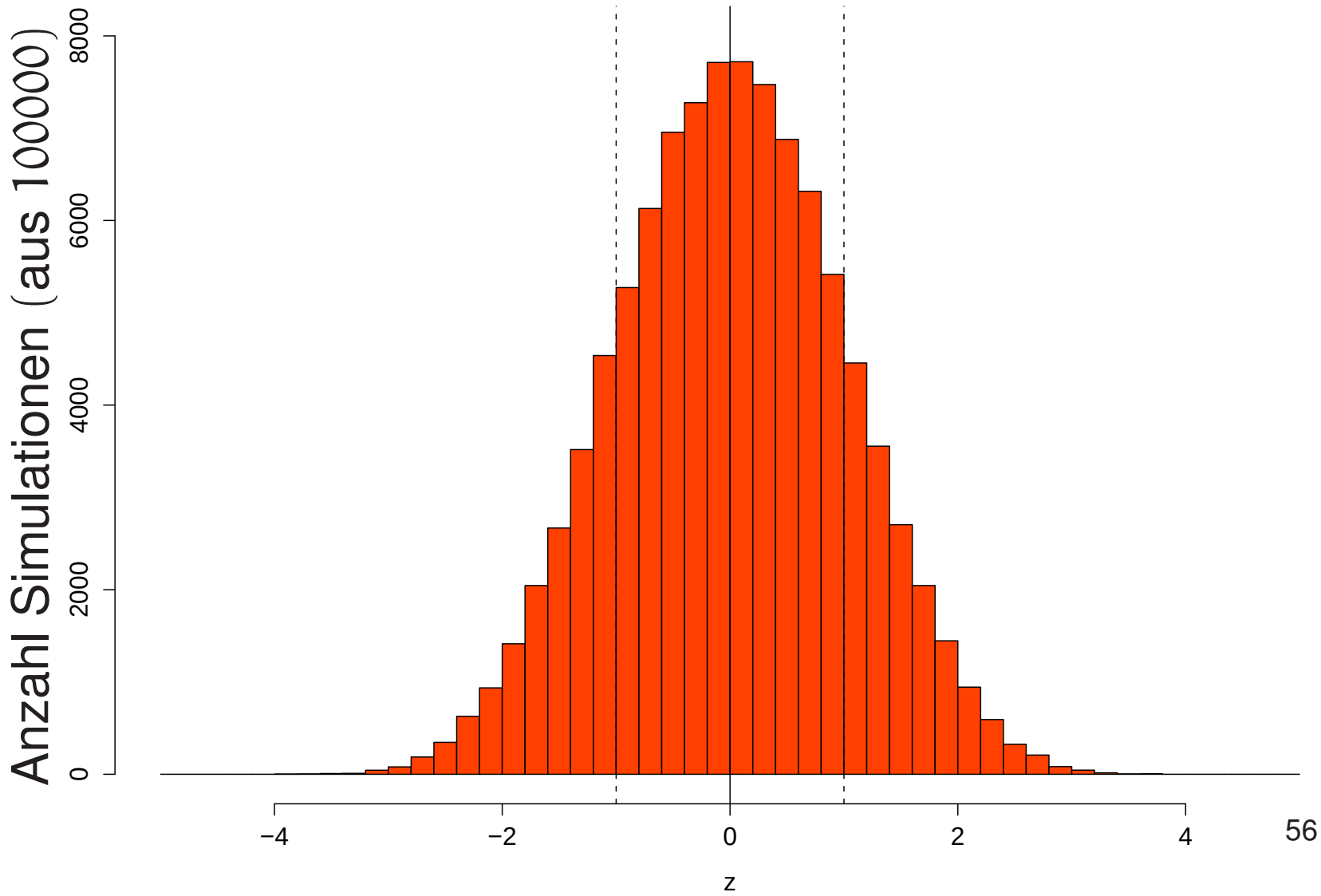
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 4)$



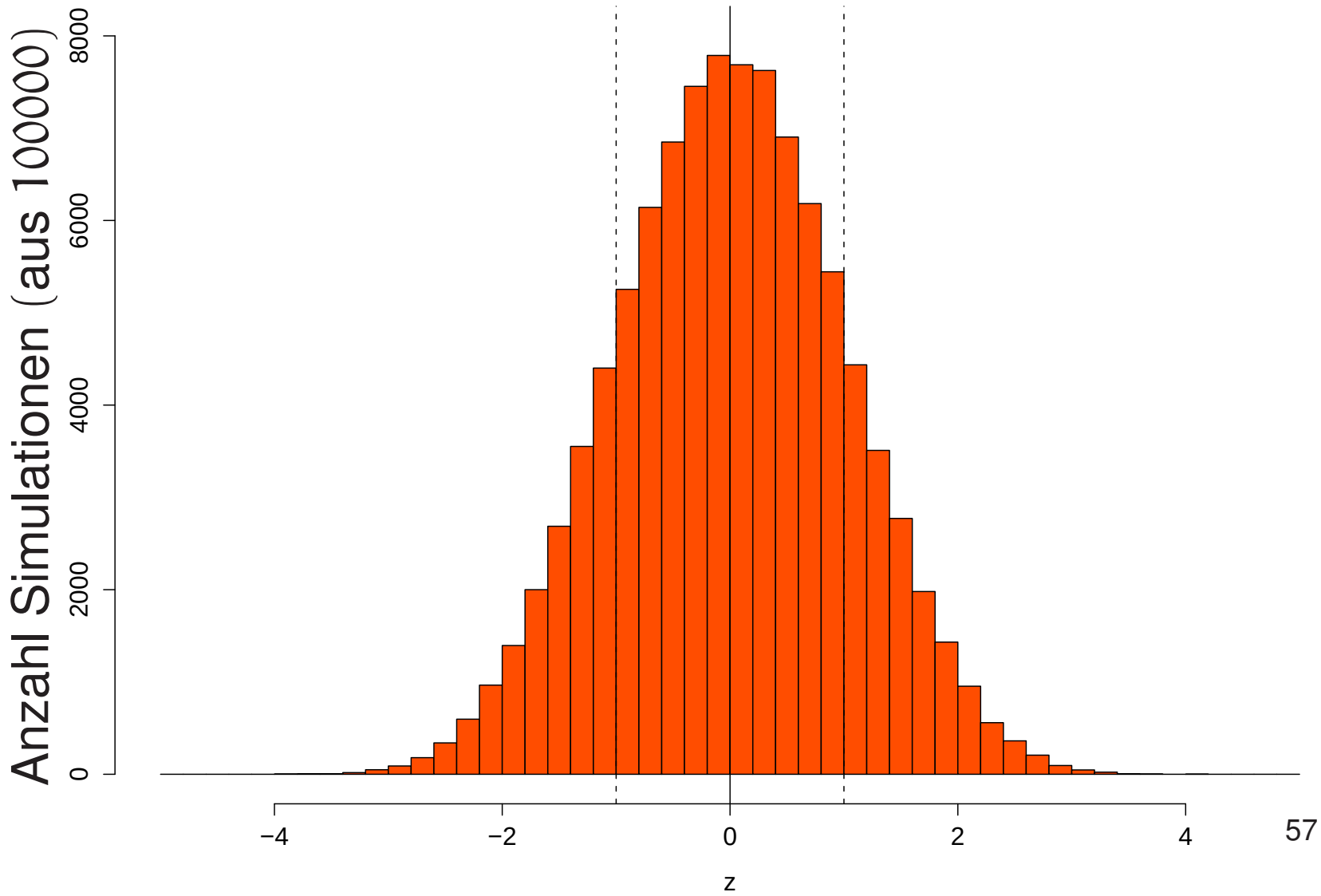
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 5)$



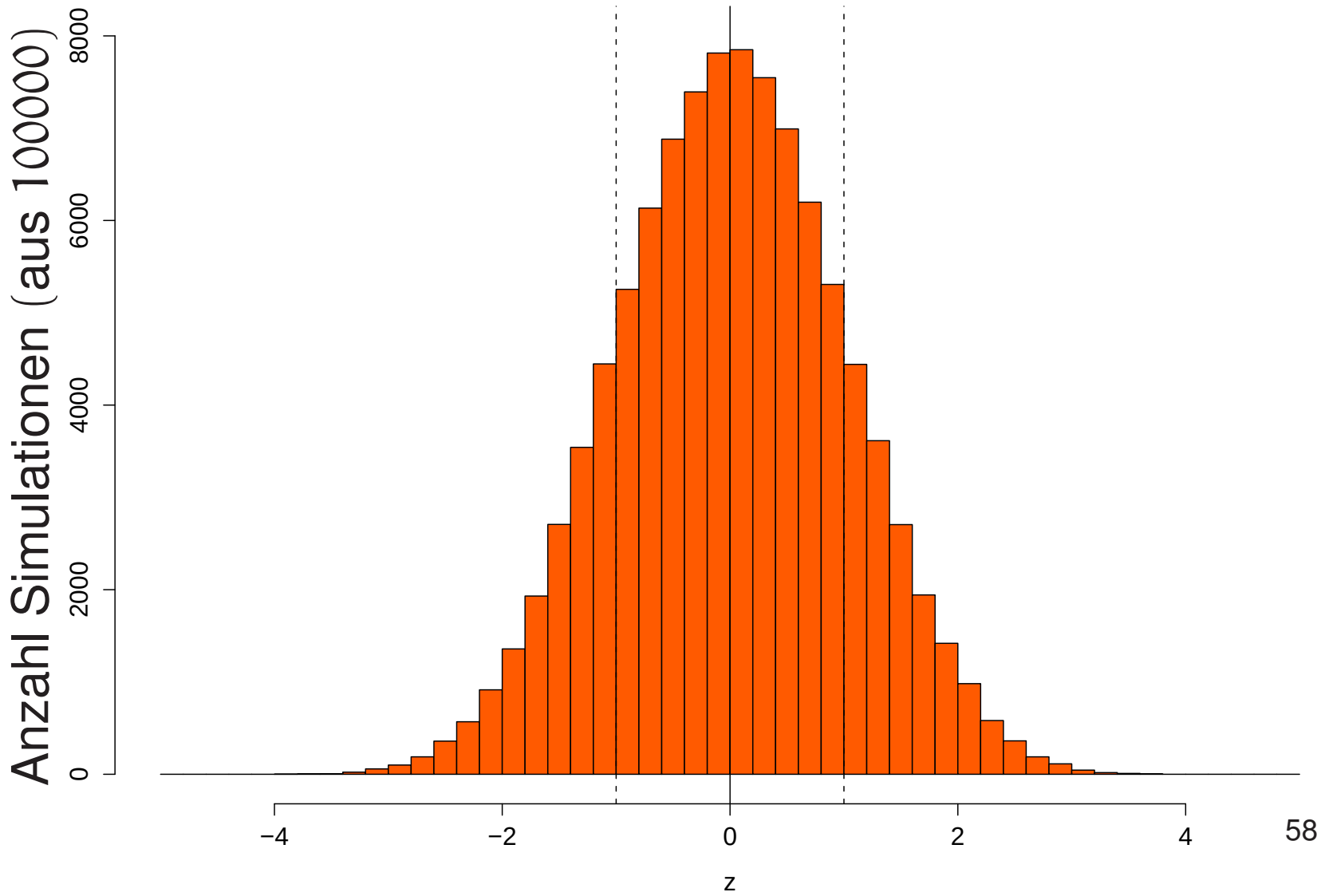
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 6)$



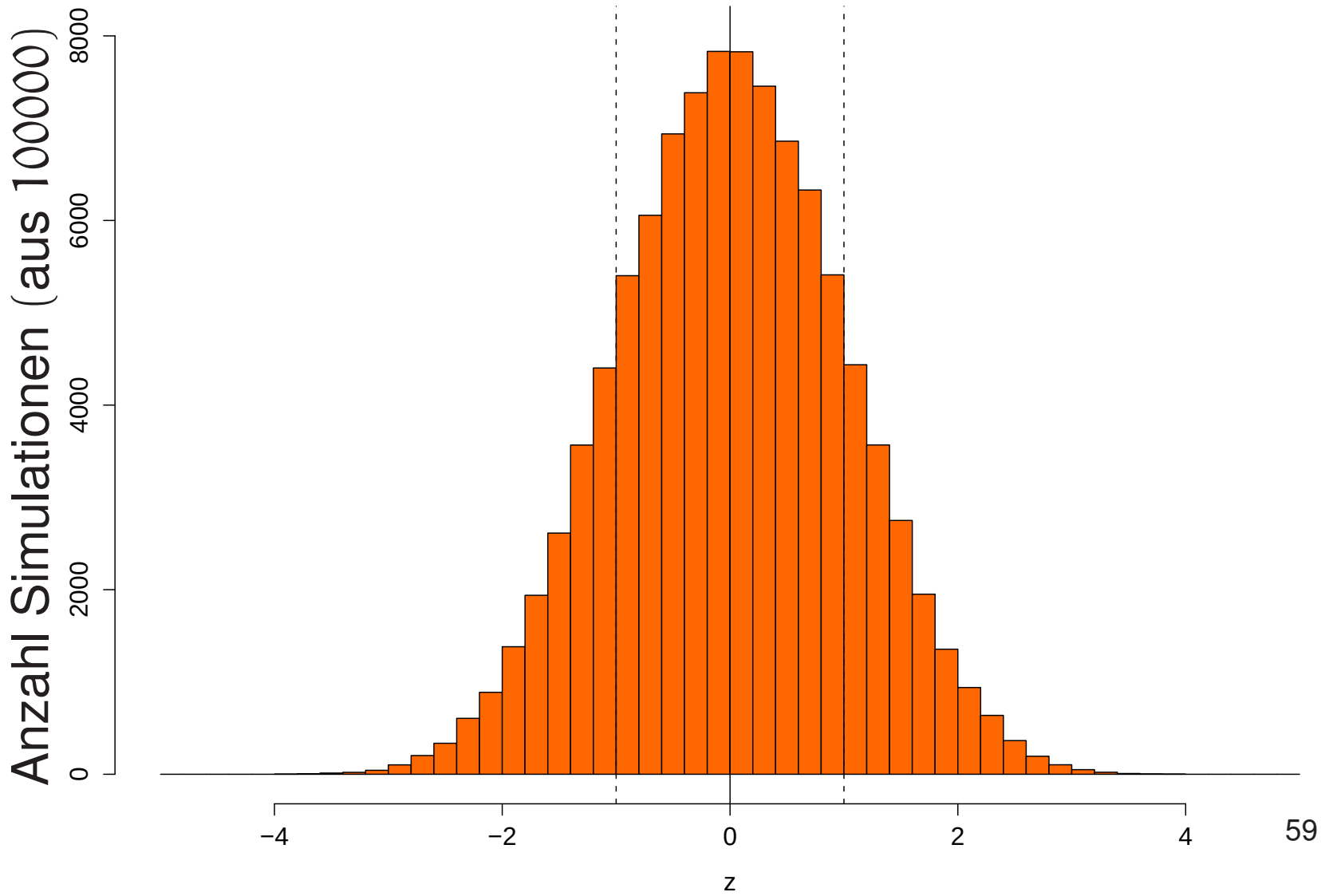
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 7)$



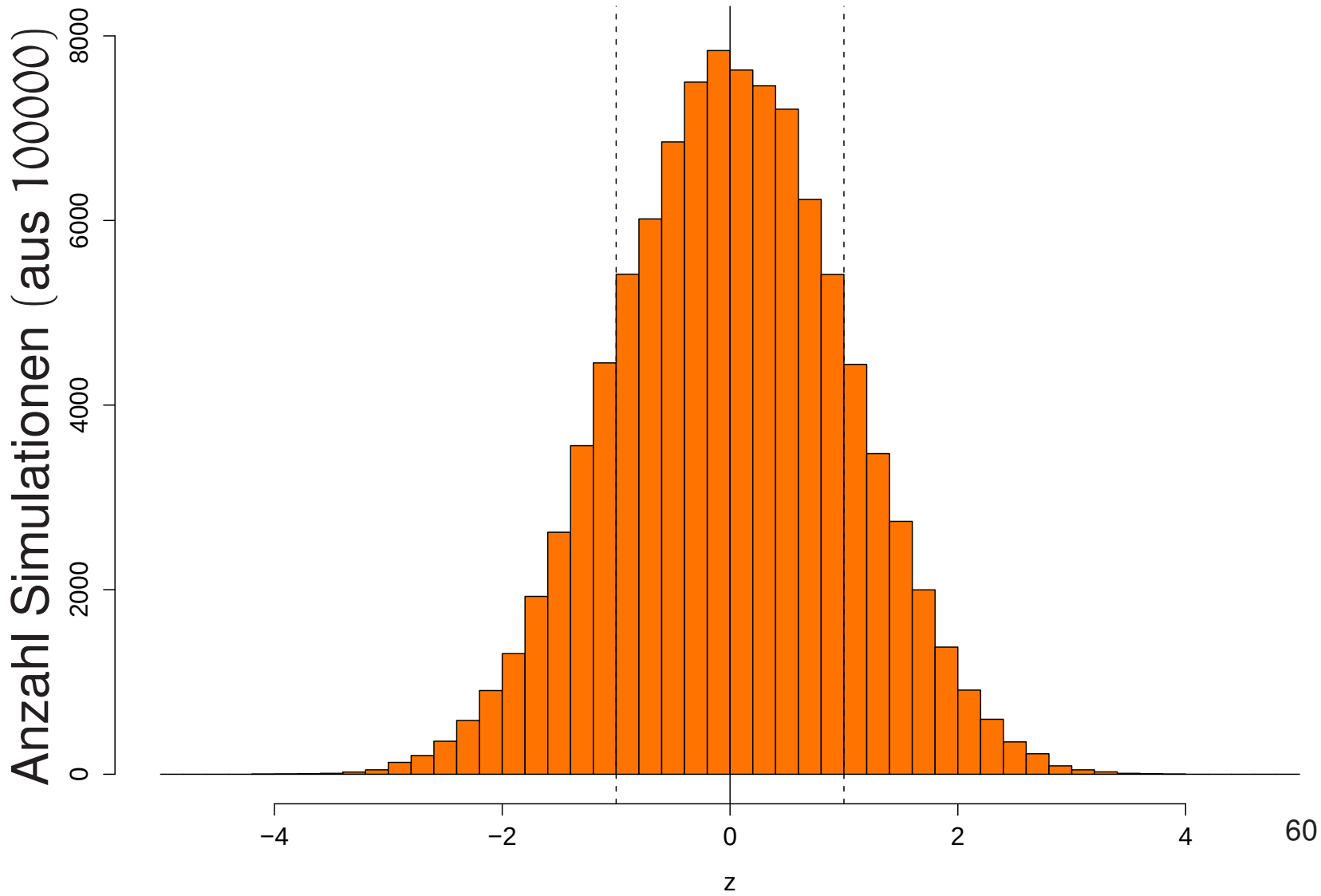
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 8)$



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

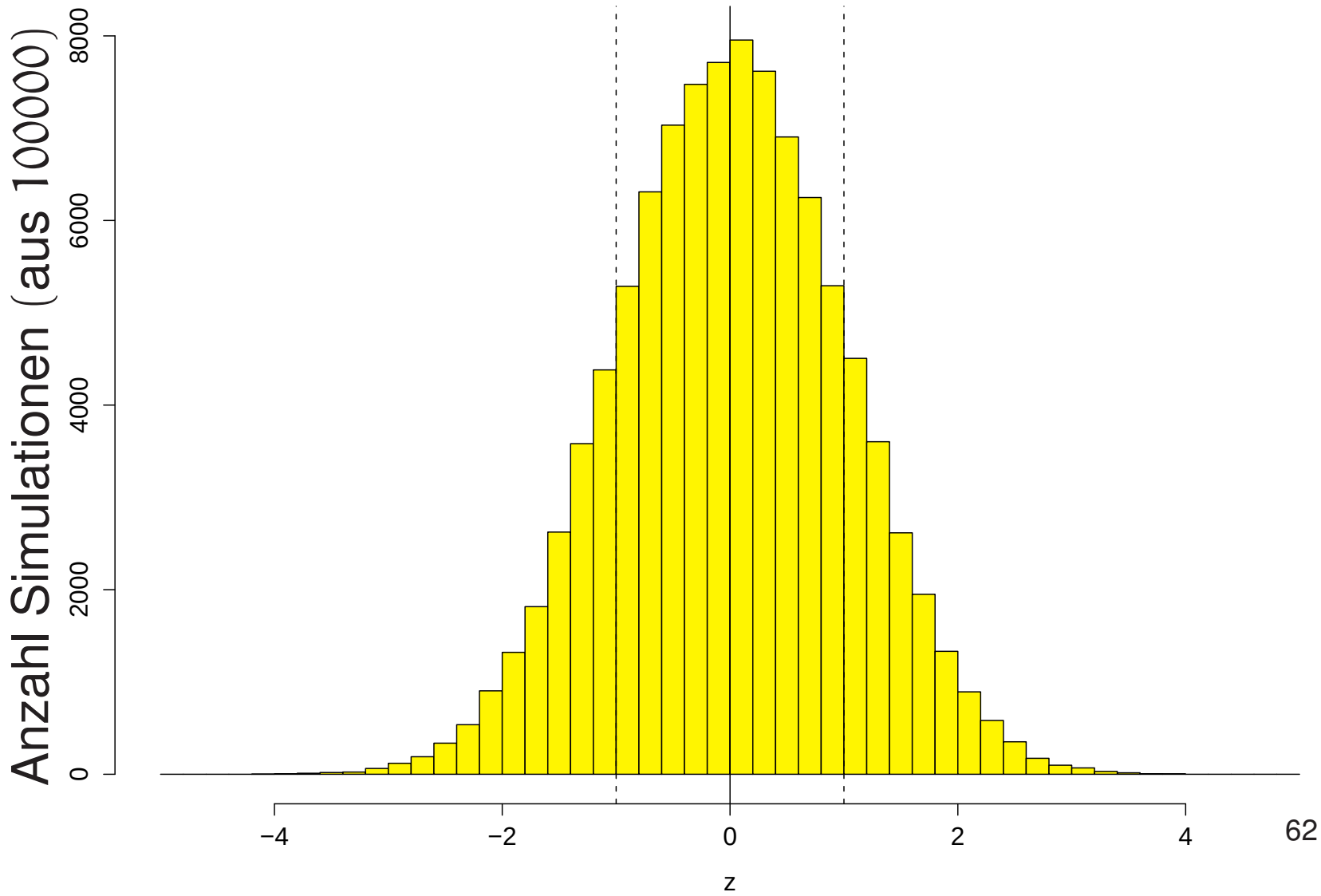


Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 10)$

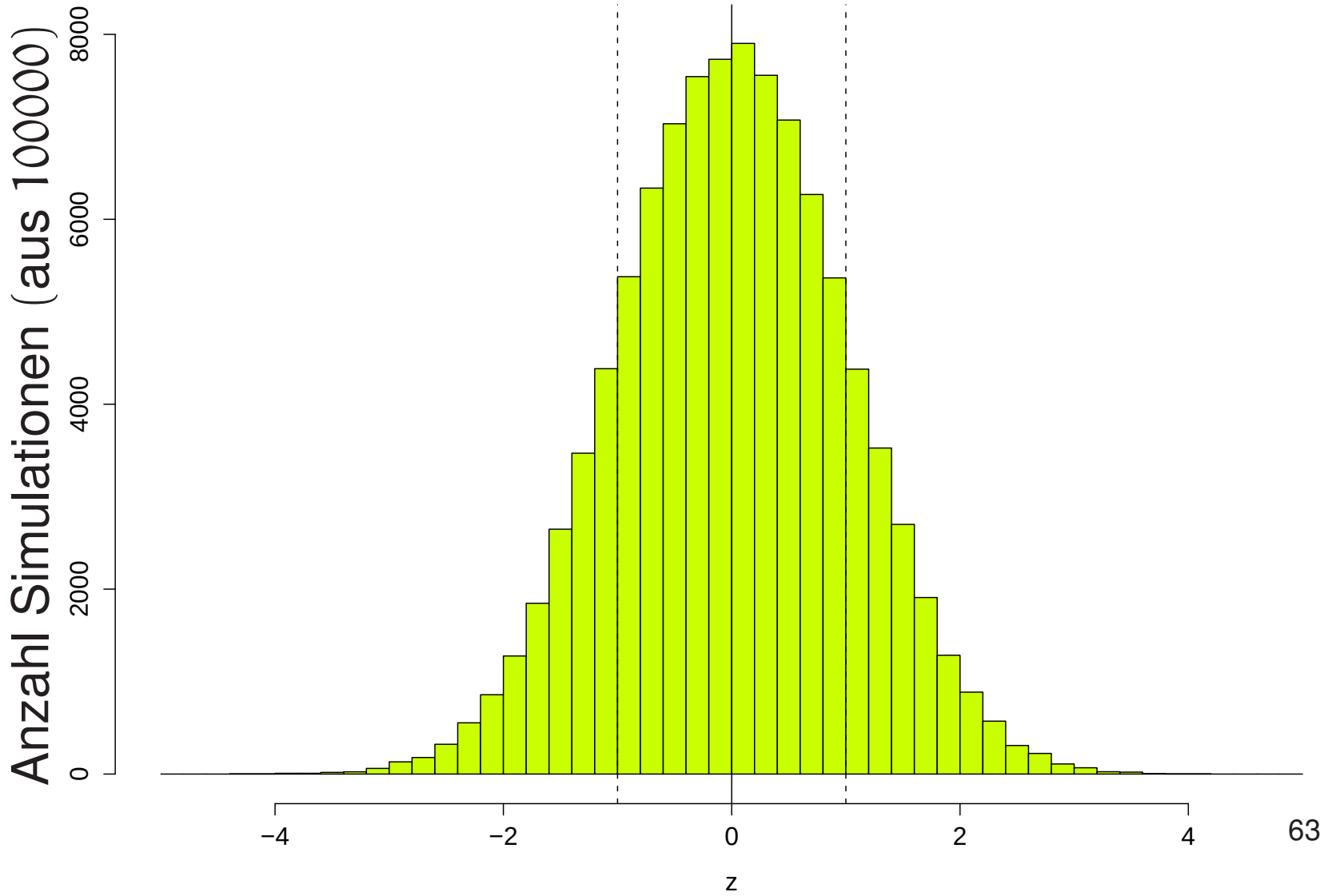


Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 15)$

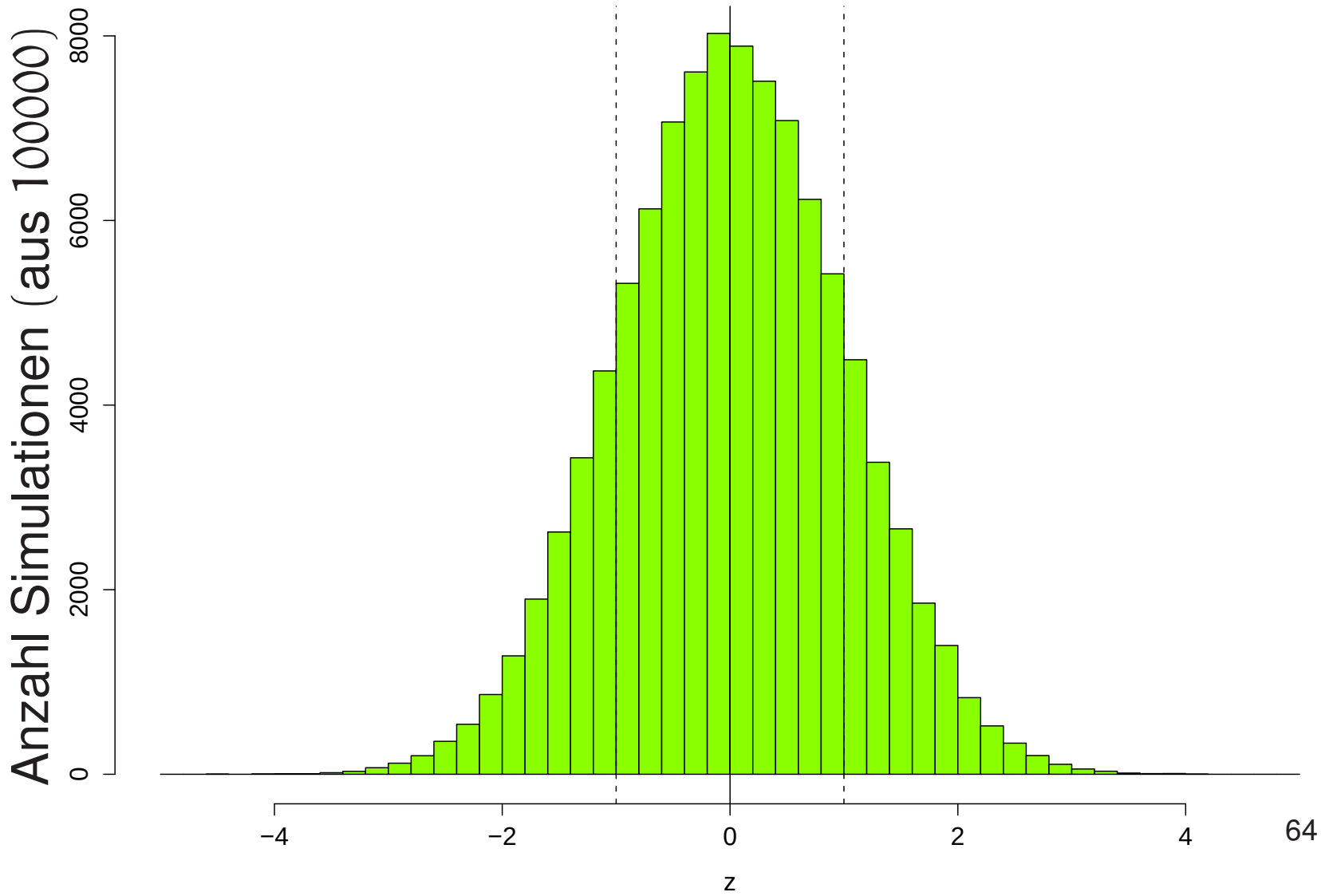
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 20)$



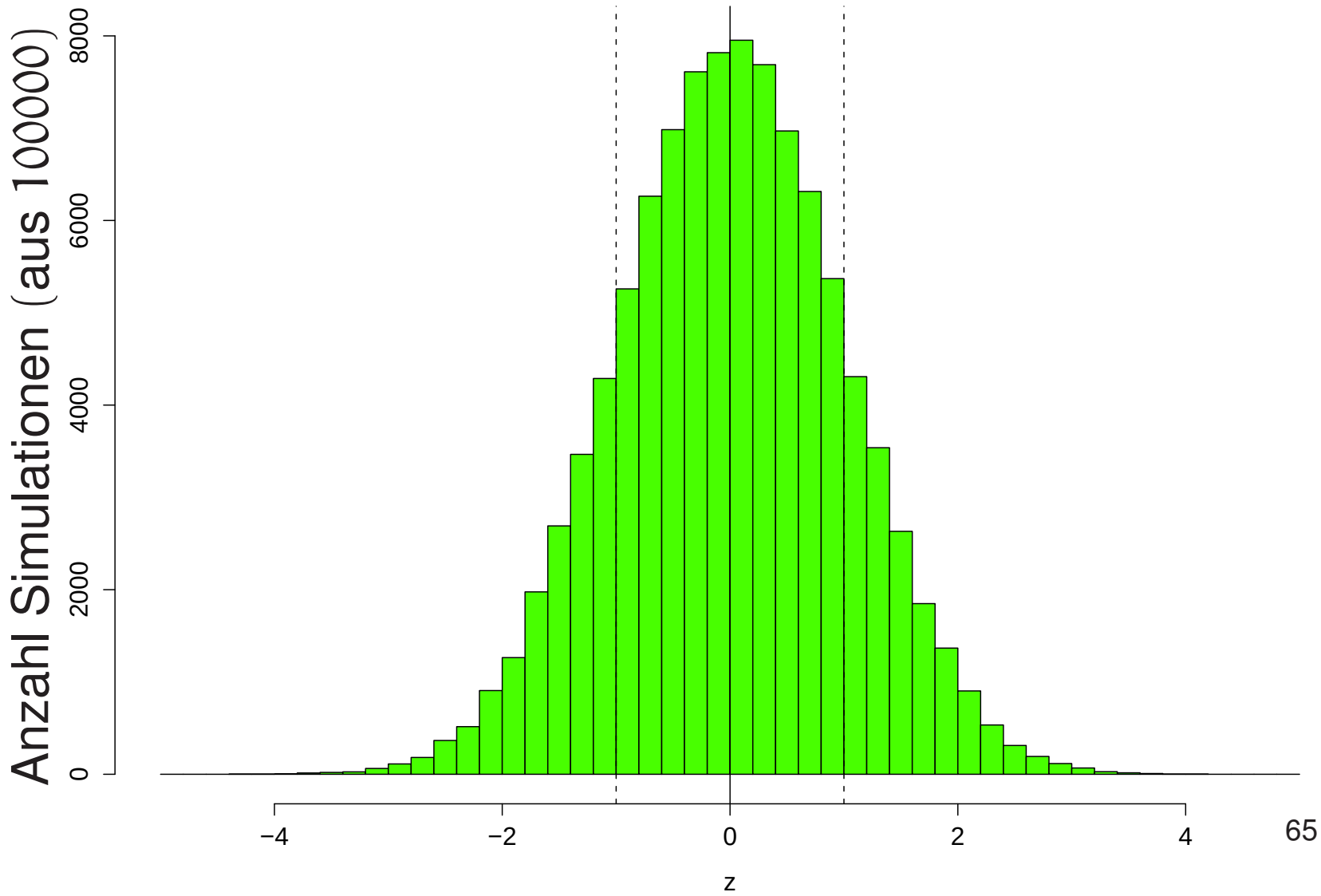
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 25)$



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

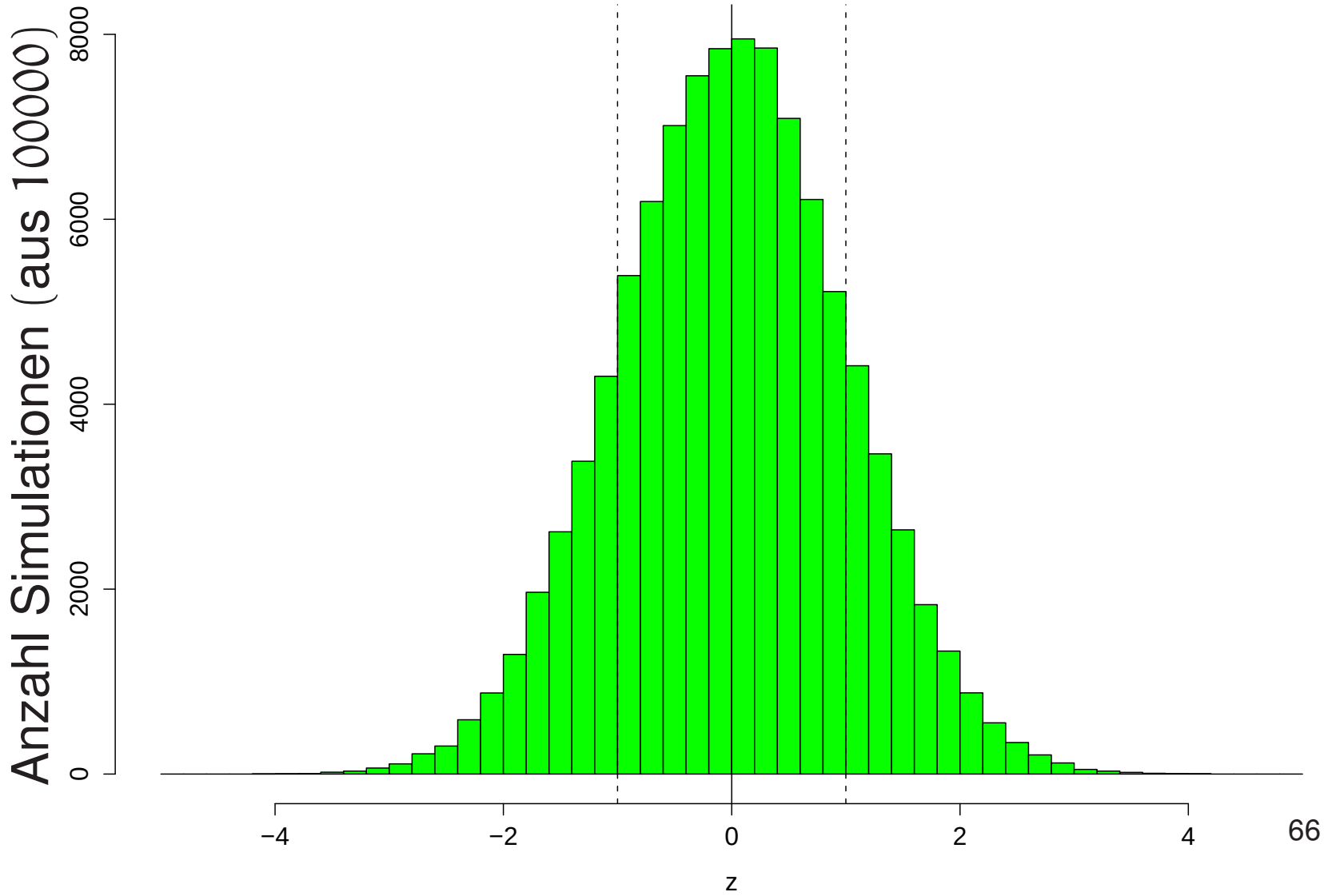


Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 35)$

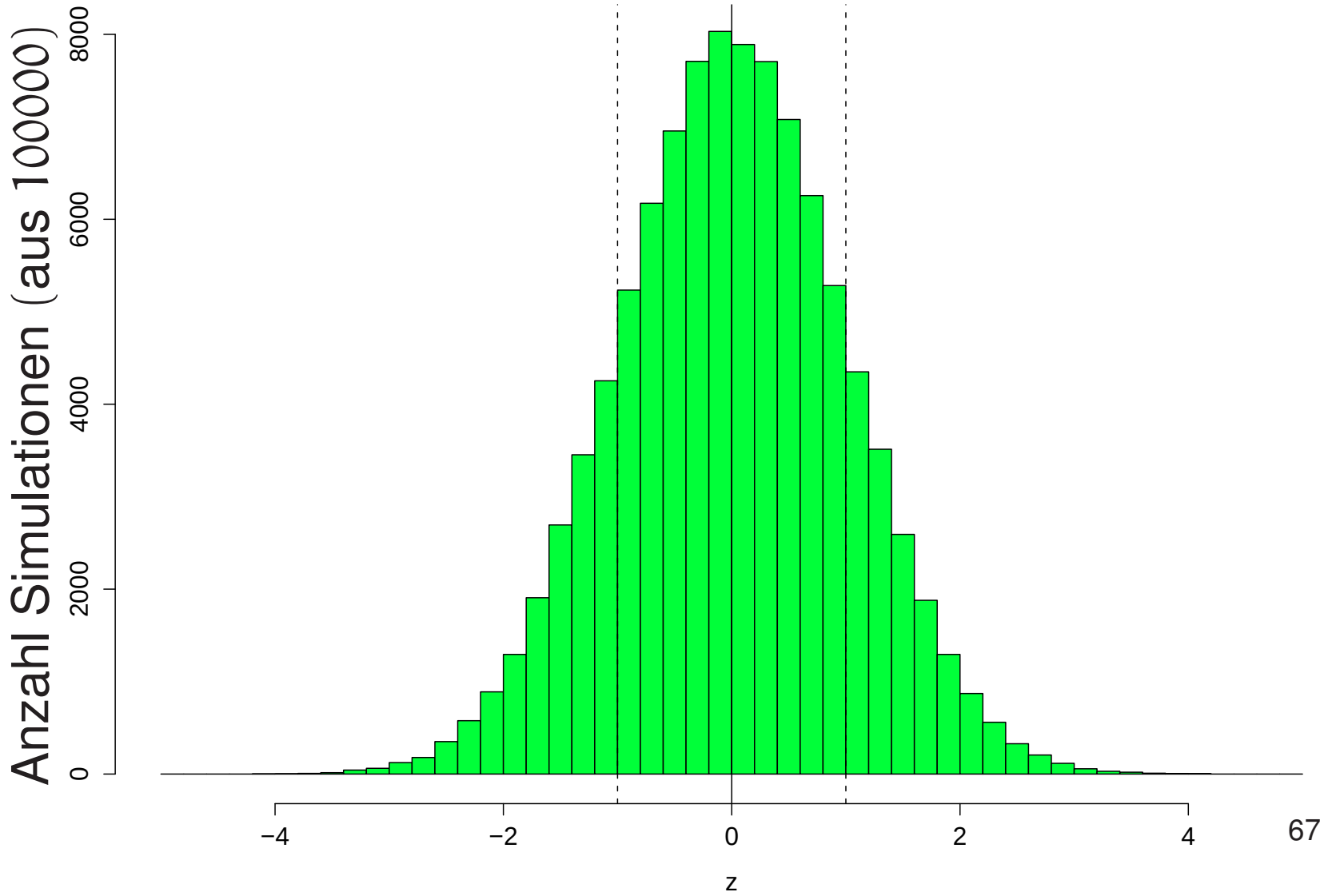


65

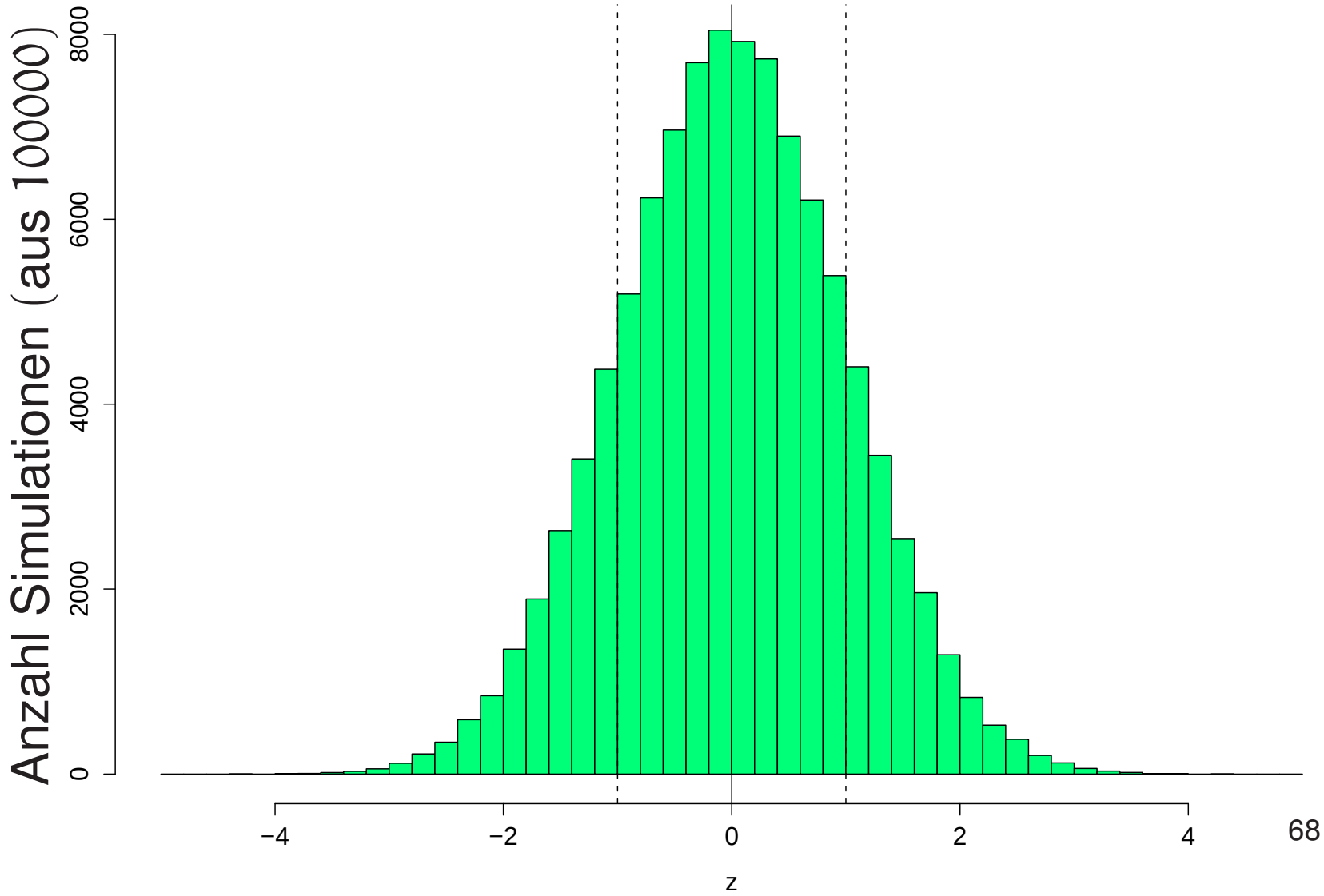
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 40)$



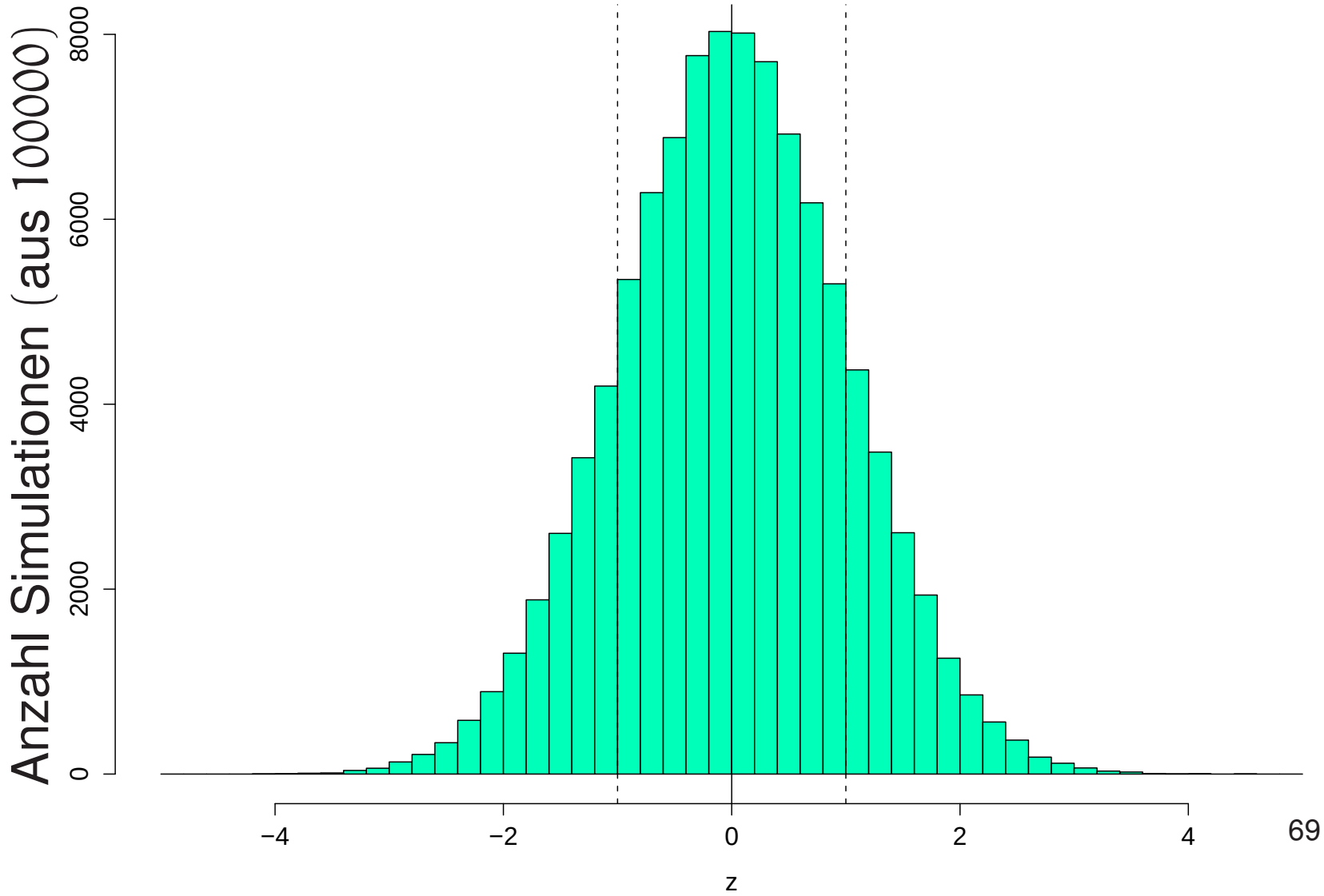
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 45)$



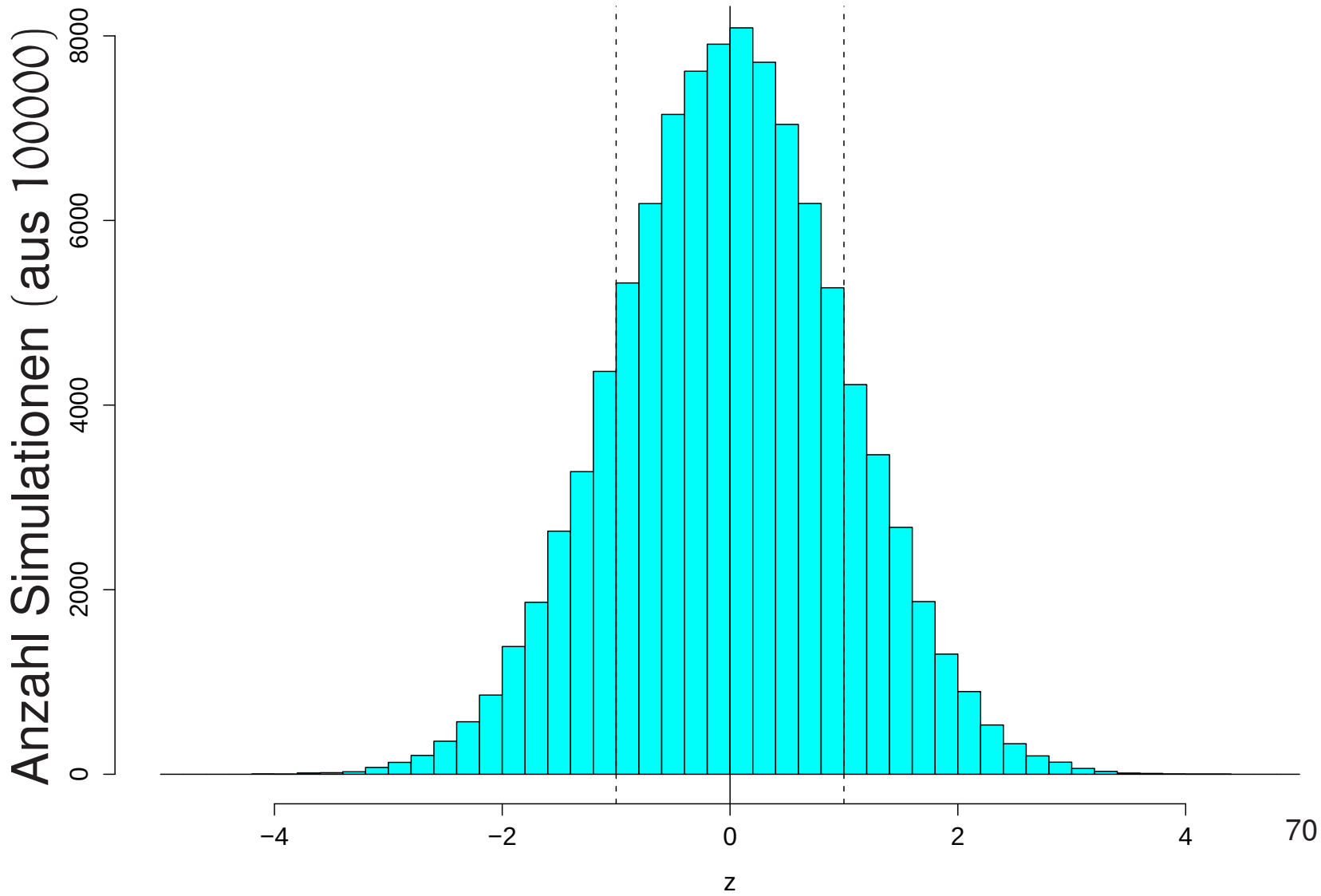
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 50)$



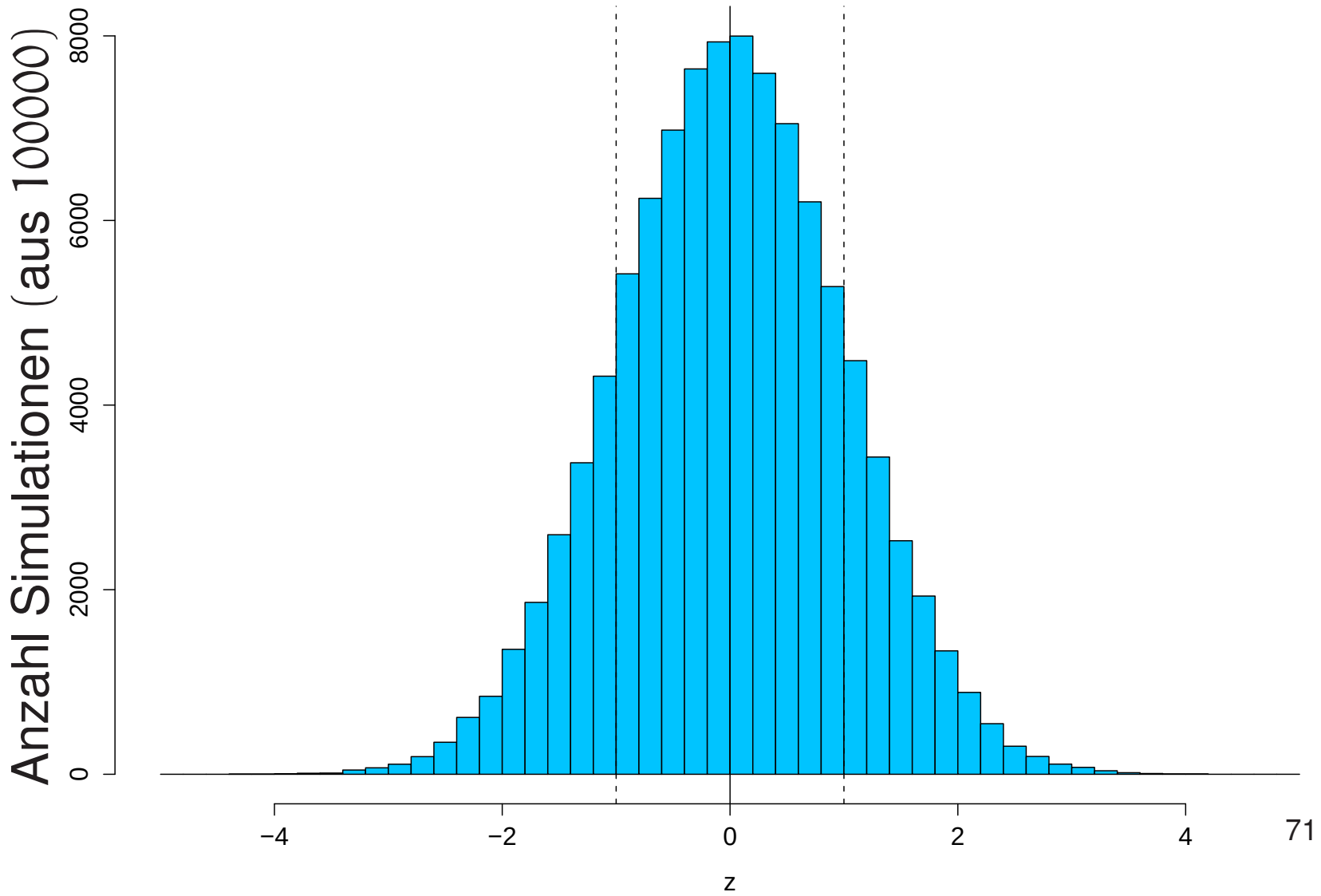
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 55)$



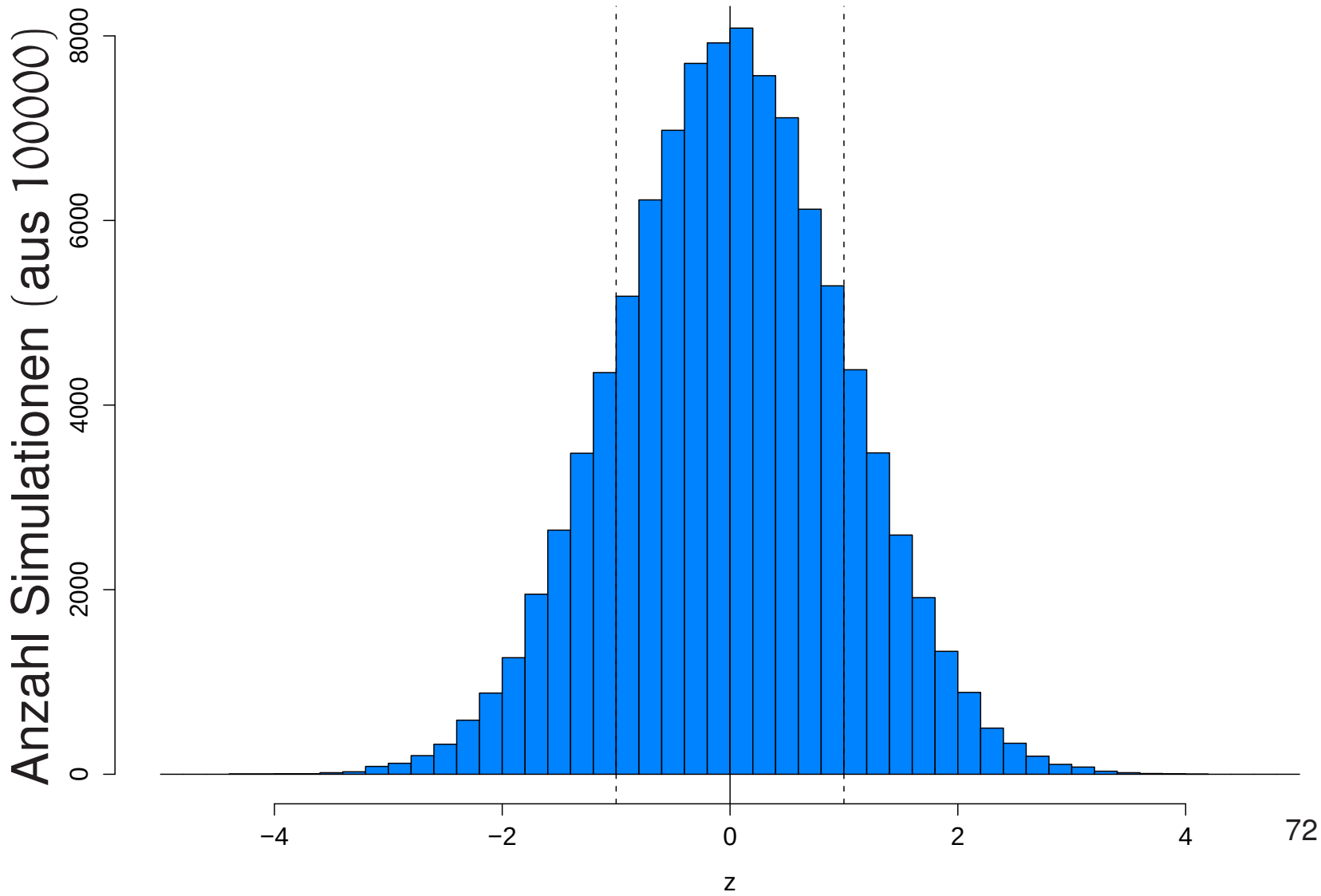
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 60)$



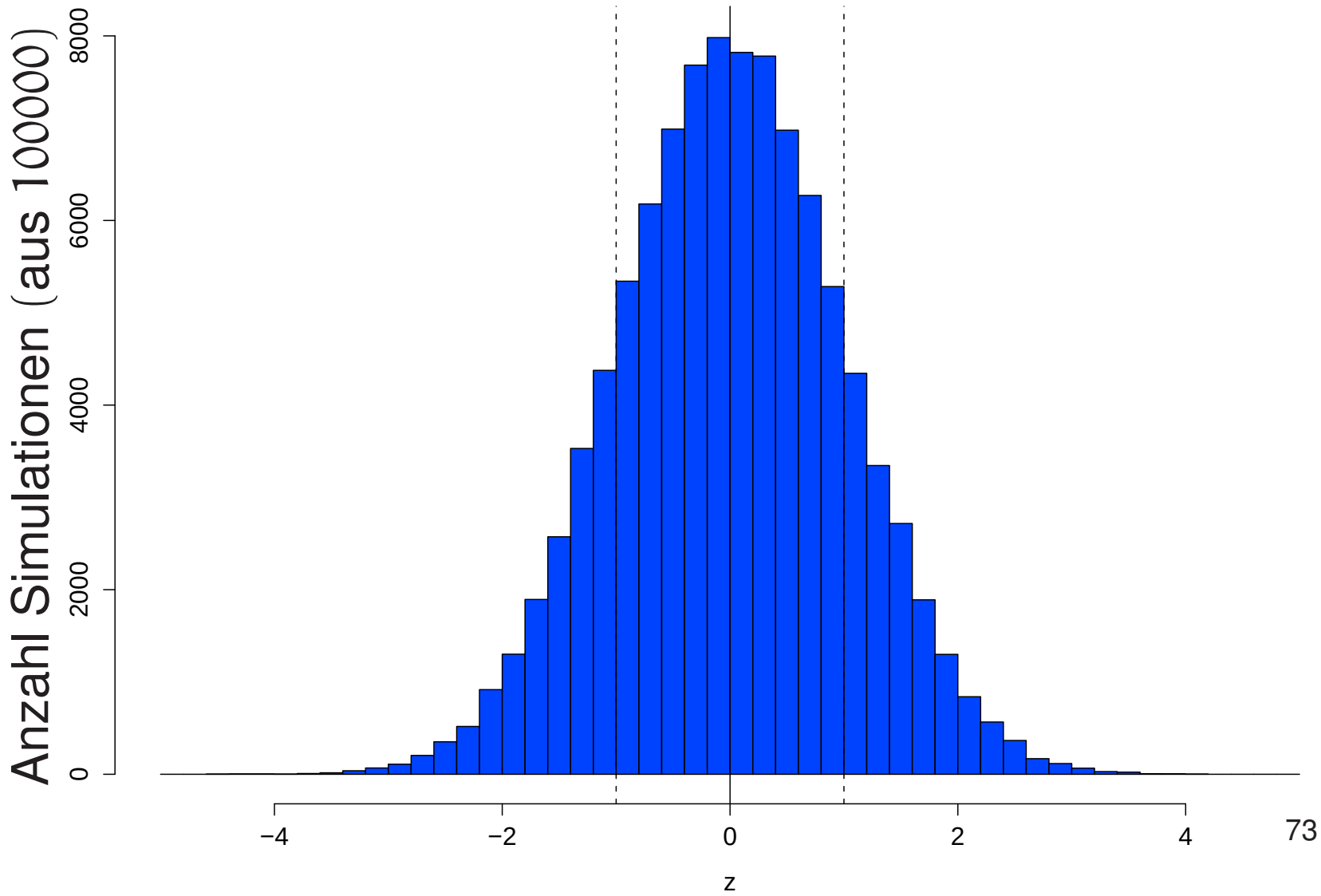
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 65)$



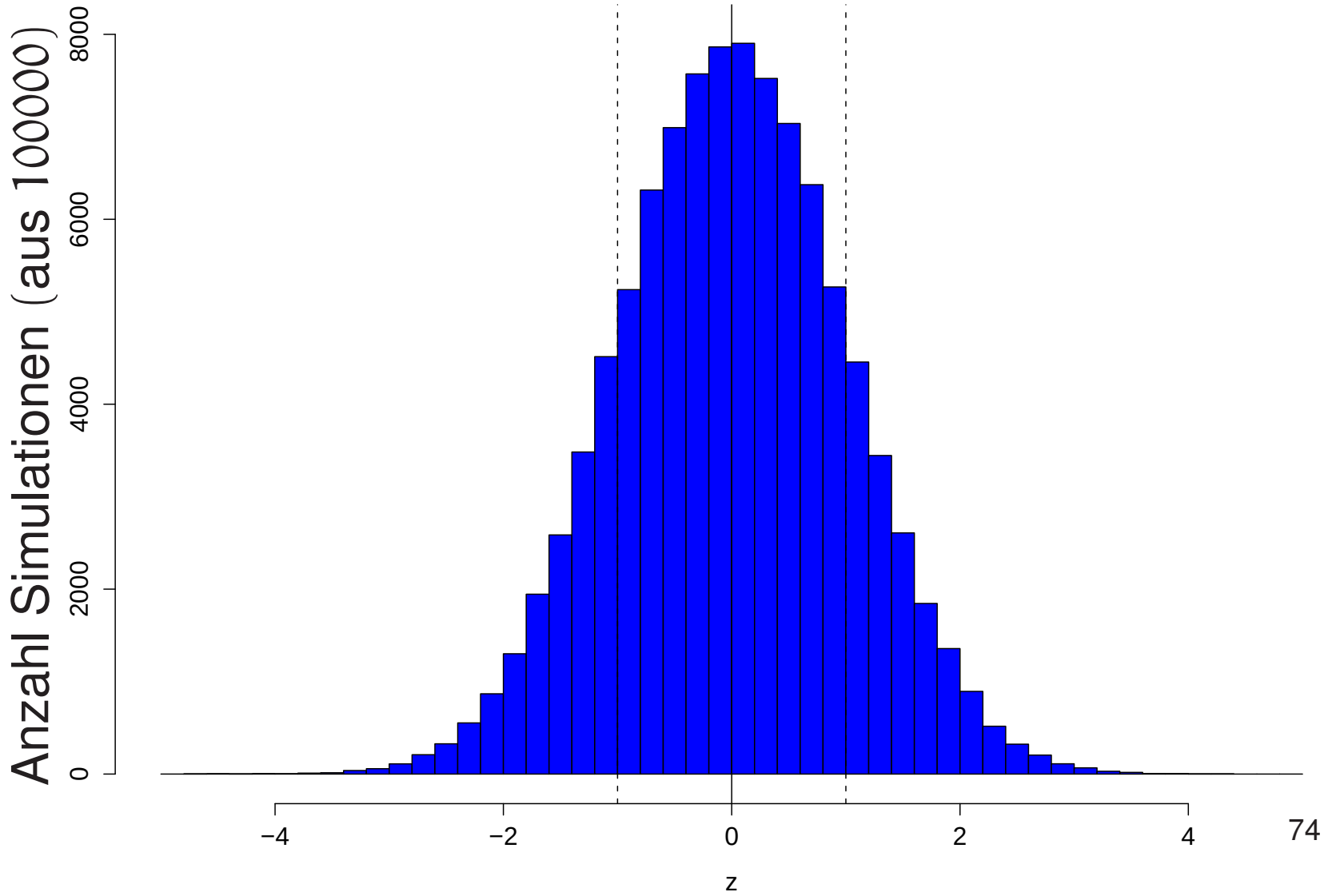
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 70)$



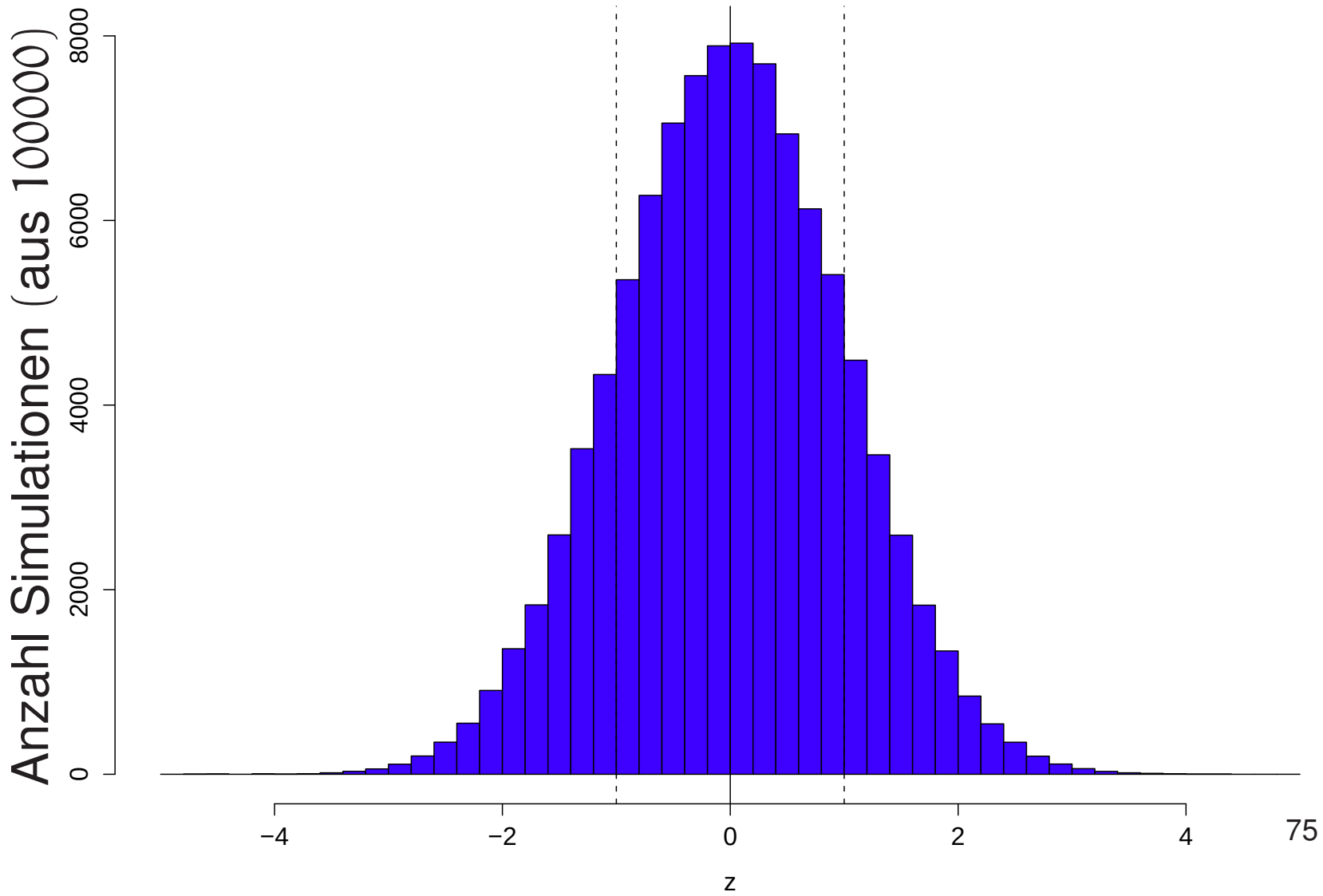
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 75)$



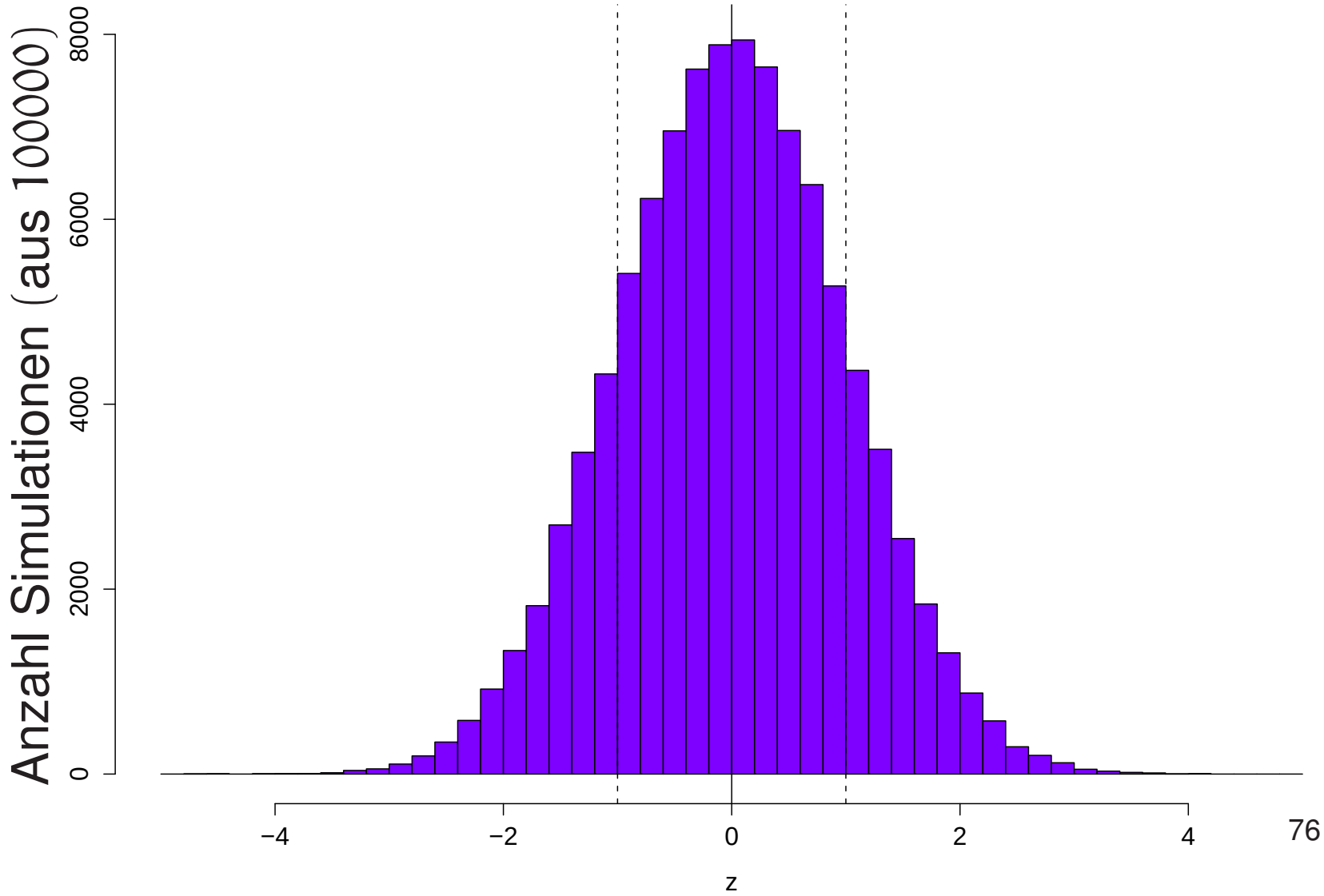
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 80)$



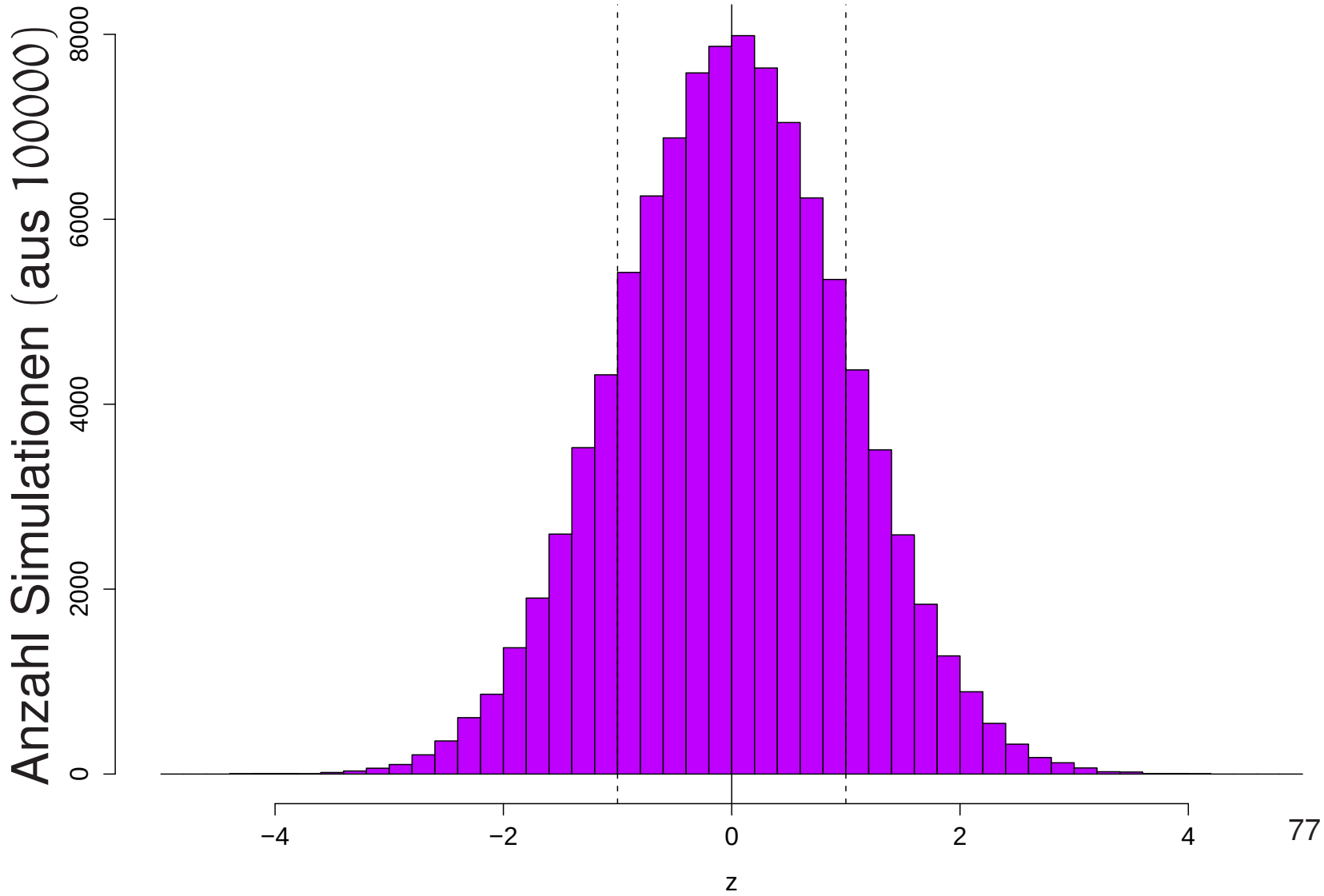
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 85)$



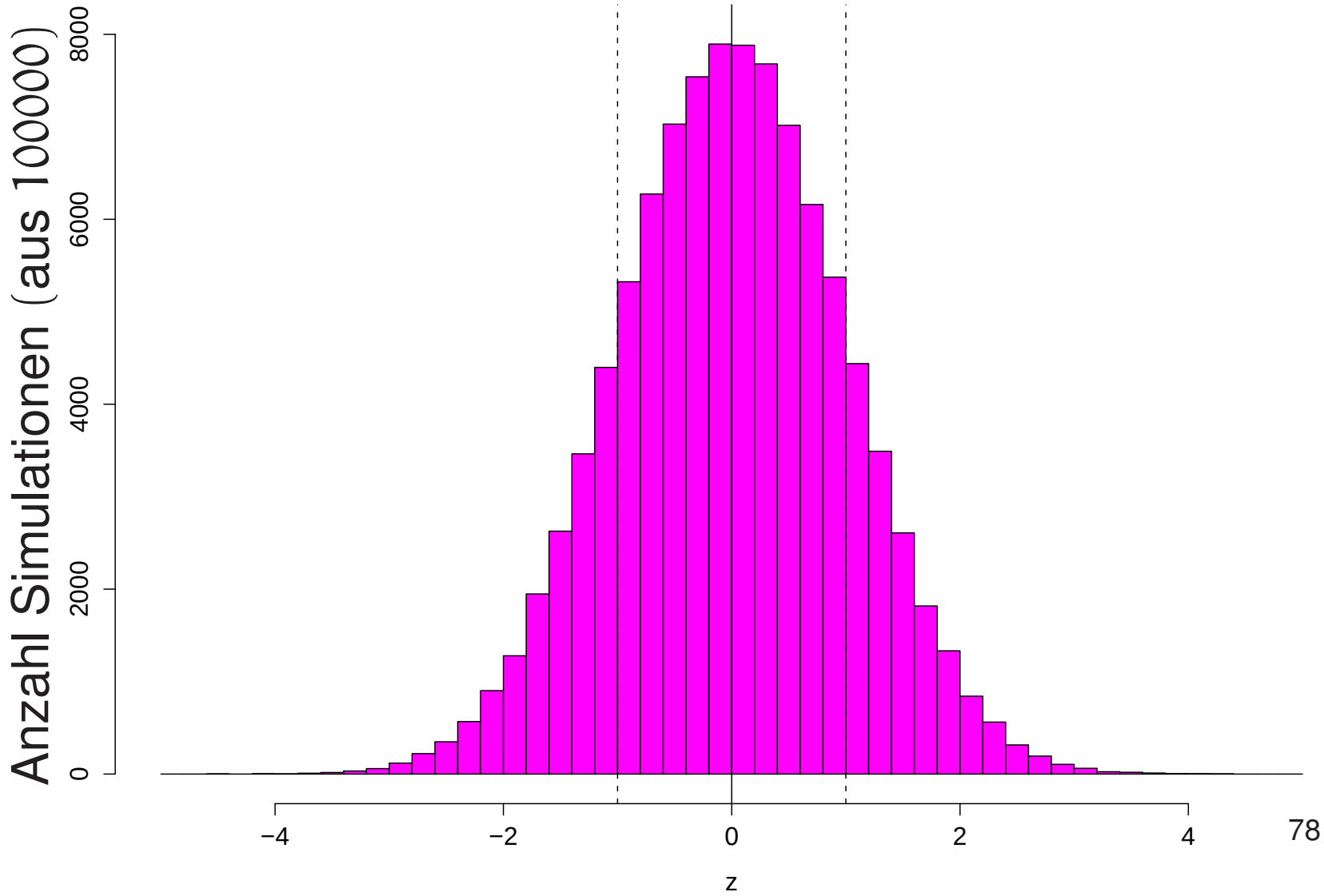
Standardisierung: $Z_n := (S_n - \mathbf{E}S_n)/\sigma_{S_n} \quad (n = 90)$



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.

Um welche Glockenkurve handelt es sich genau?

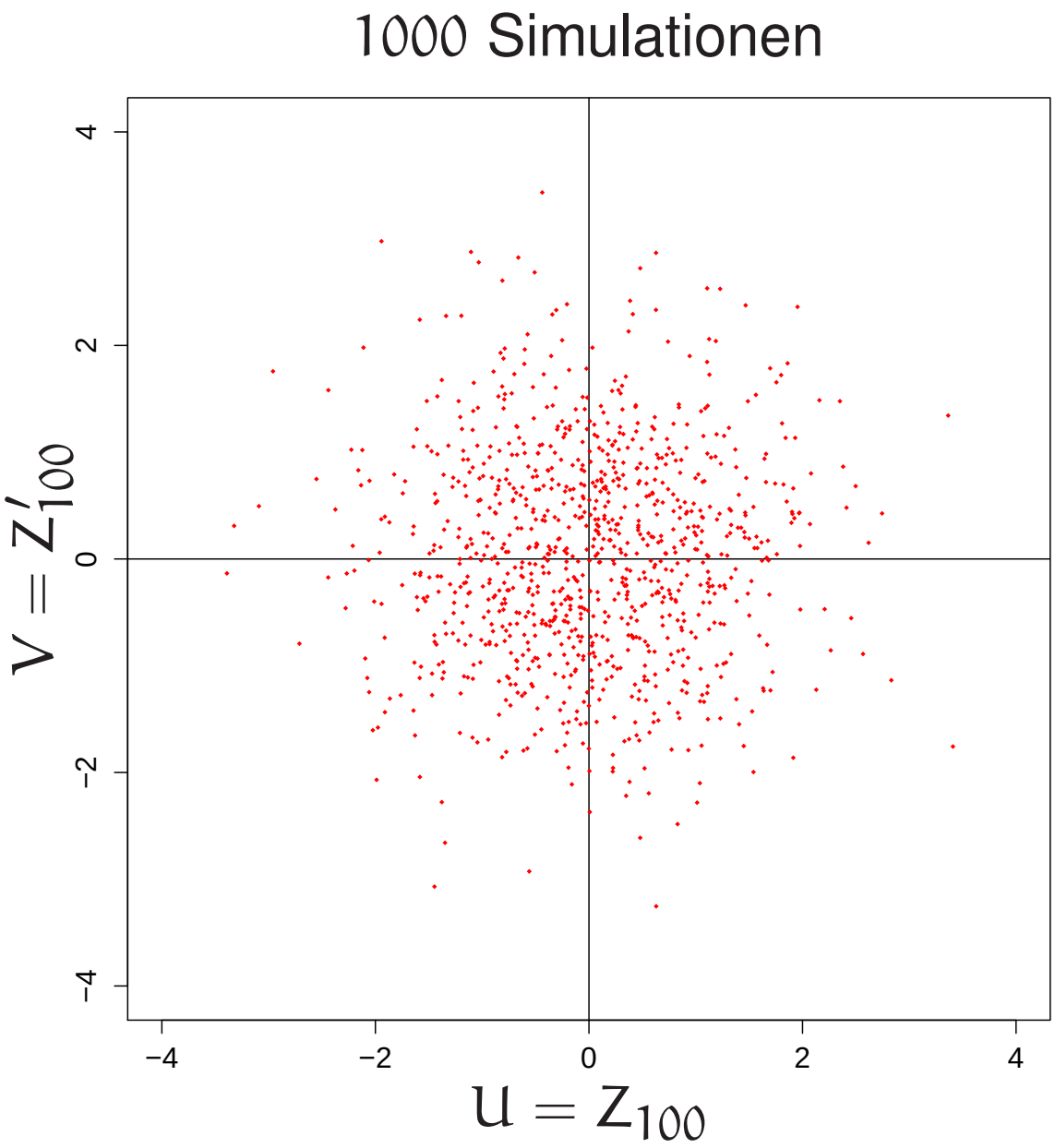
Glücklicher Einfall:

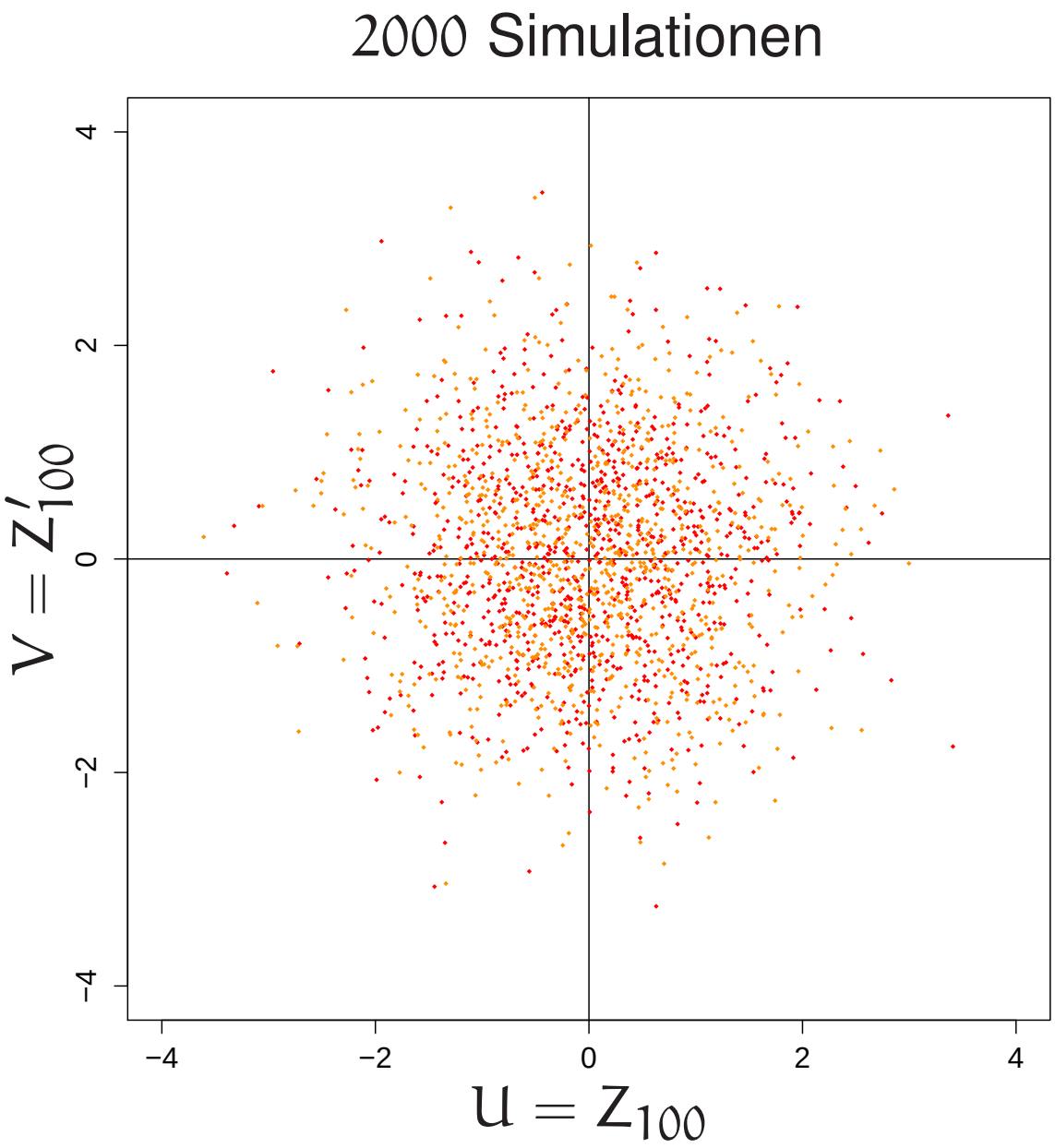
Nimm zwei unabhängige Kopien

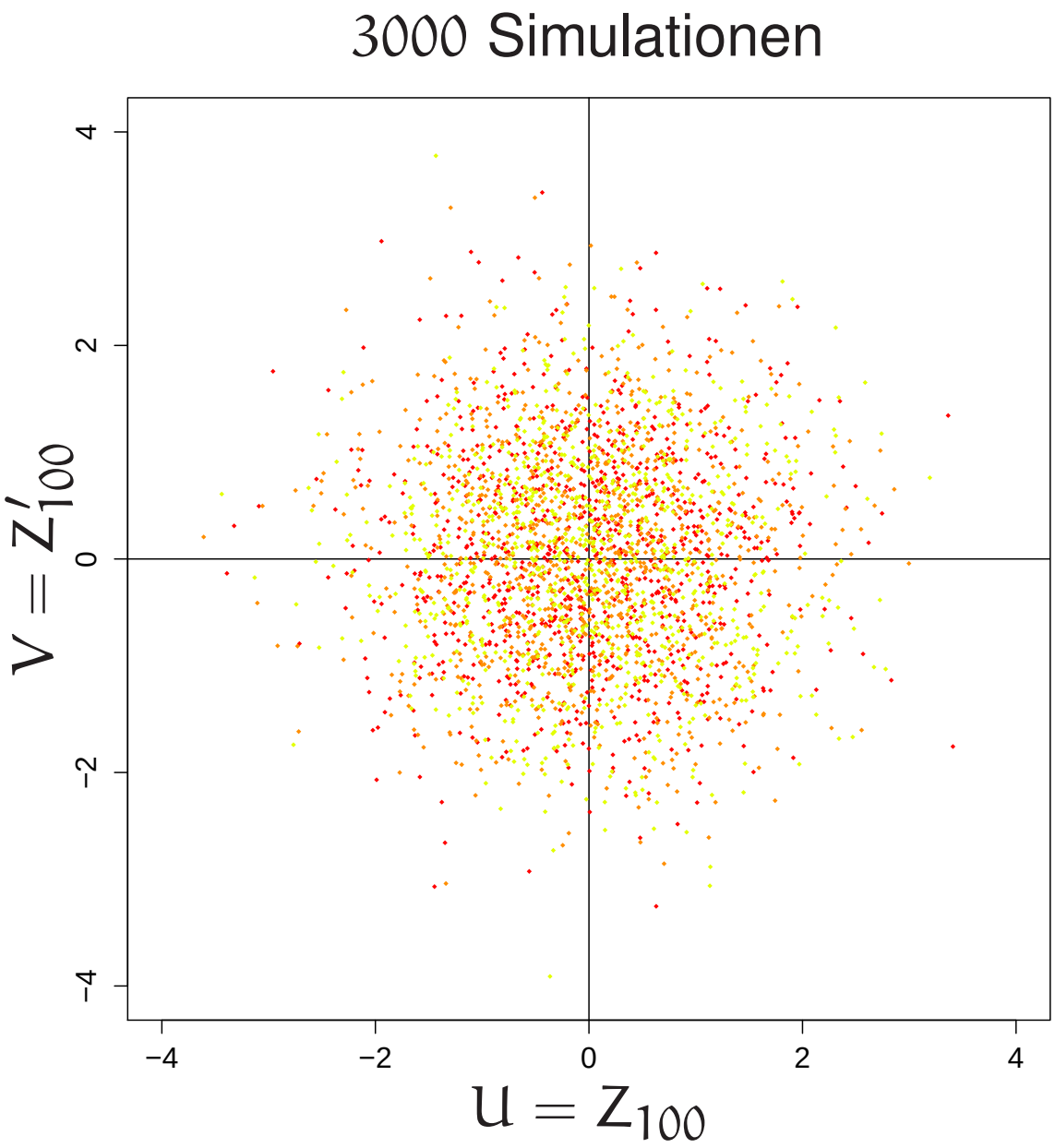
$$(U, V) := (Z_{100}, Z'_{100})$$

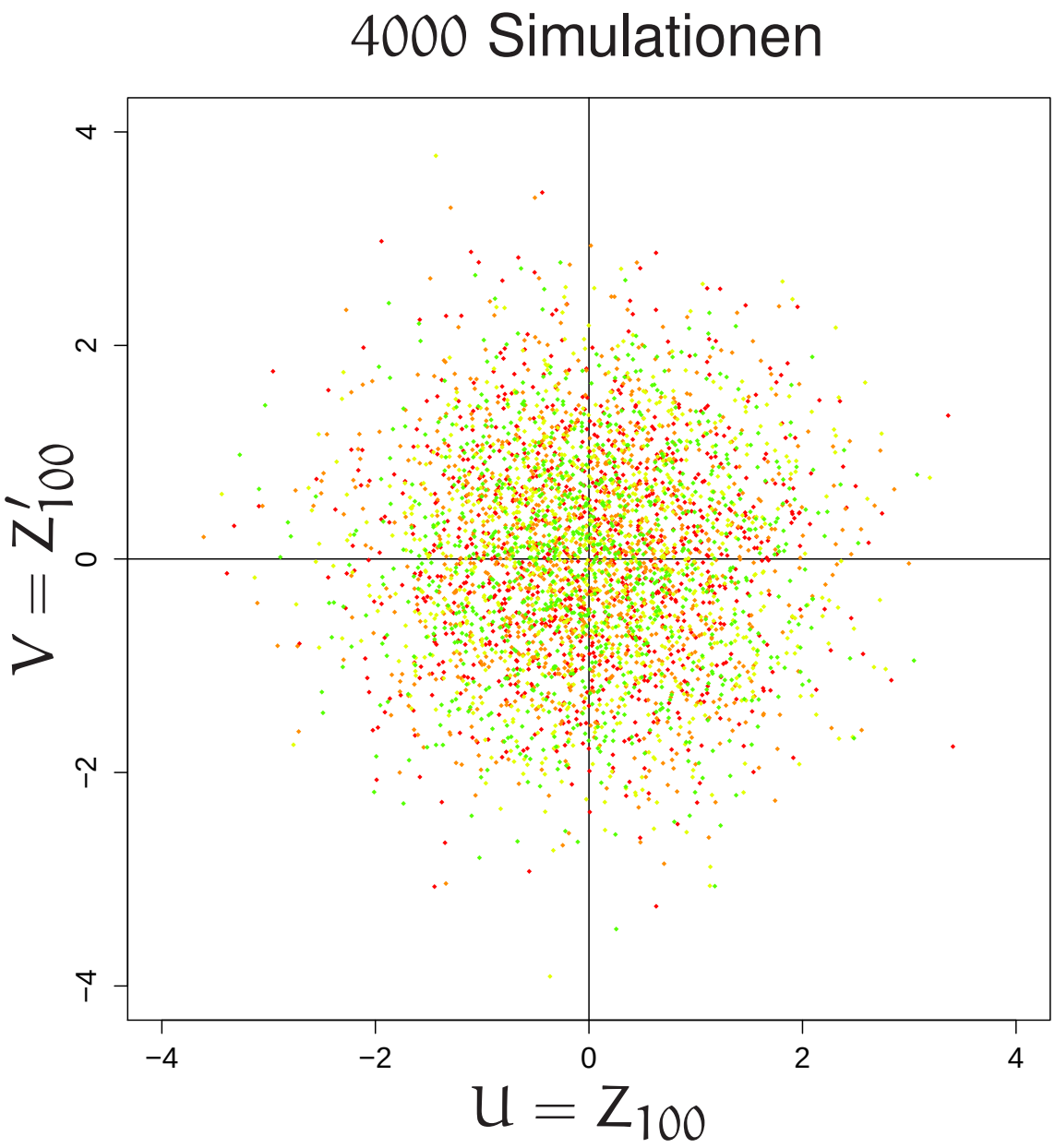
Wie sieht die gemeinsame Verteilung

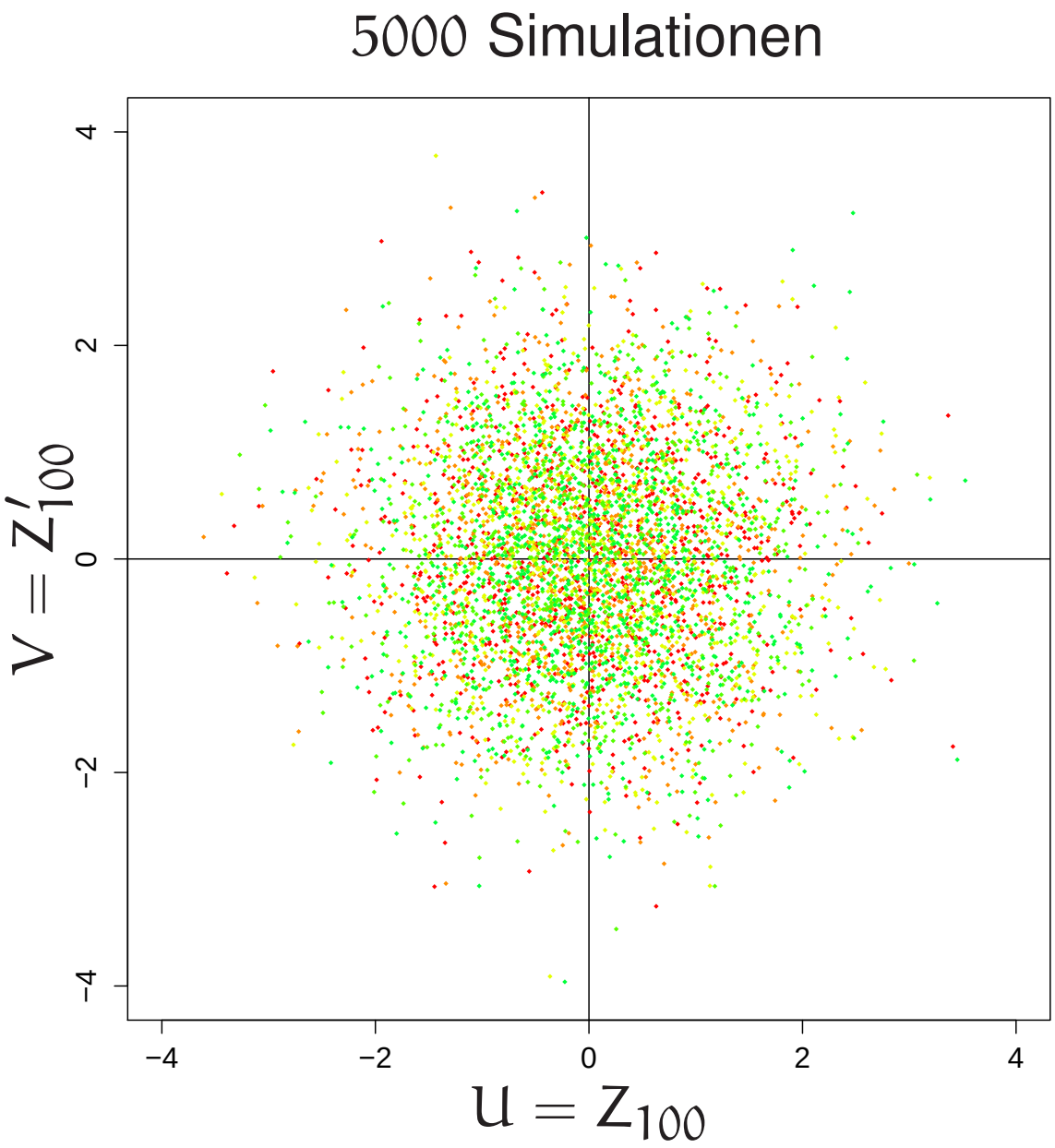
von U und V aus?

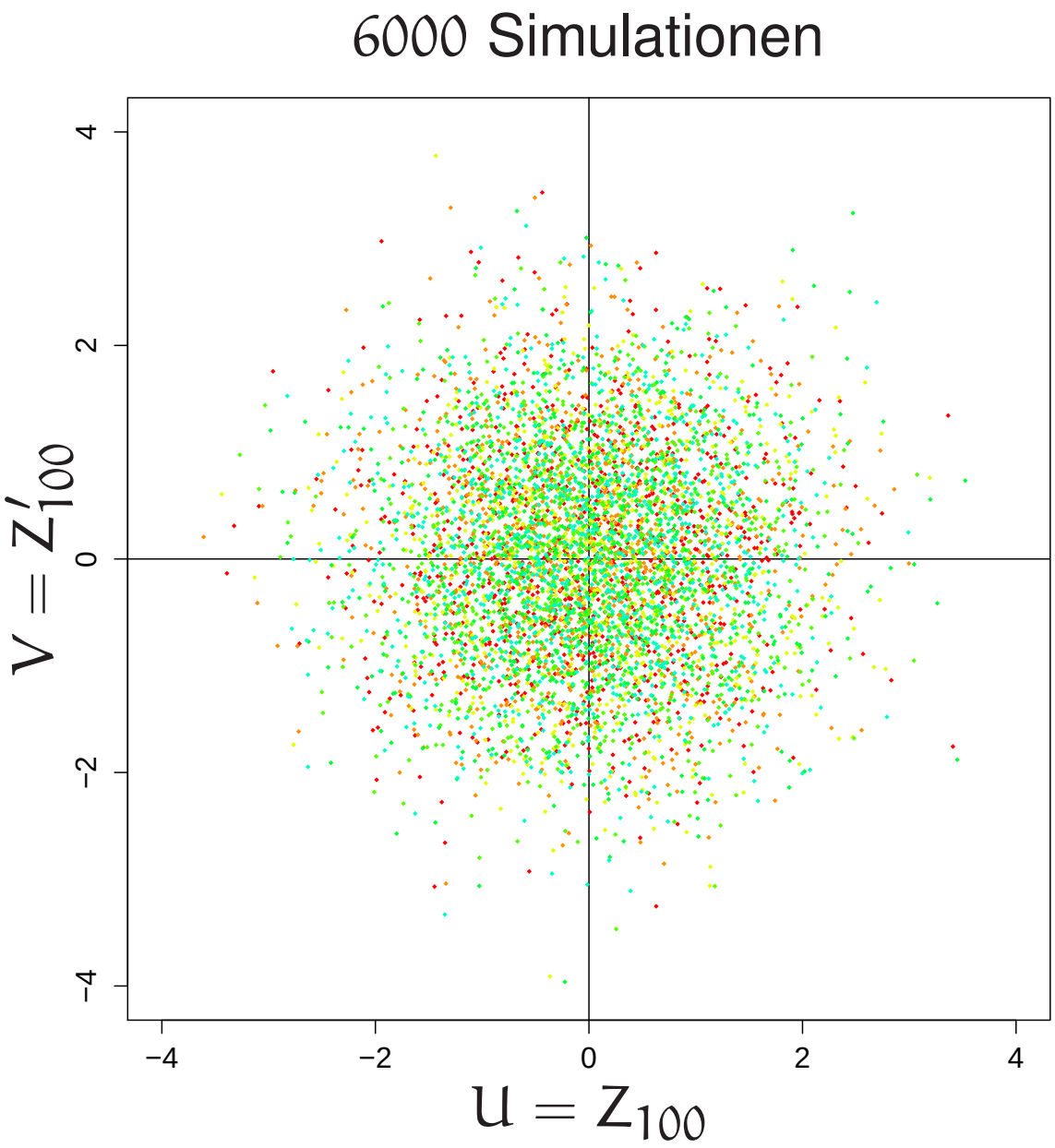

1000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

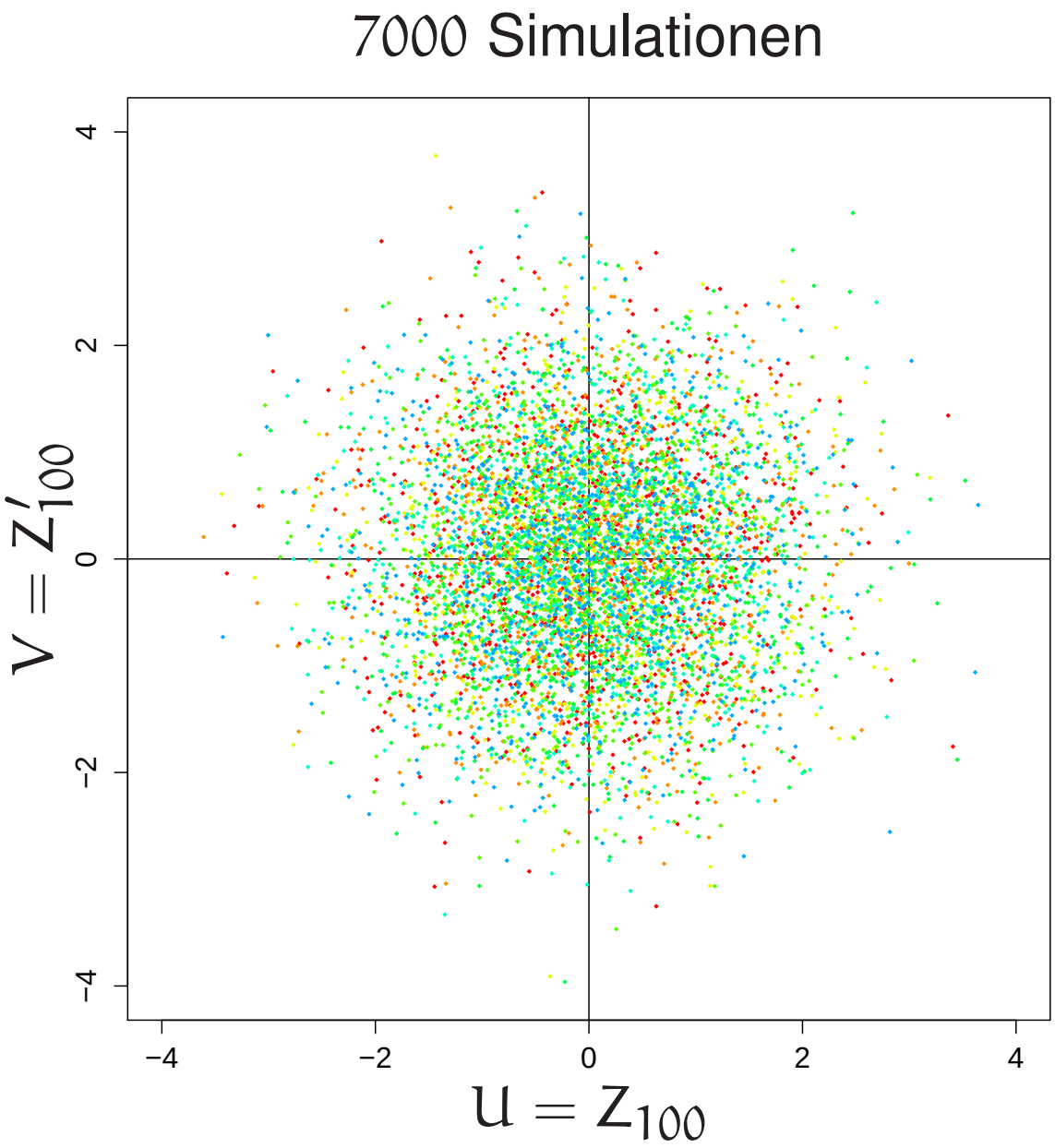

2000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

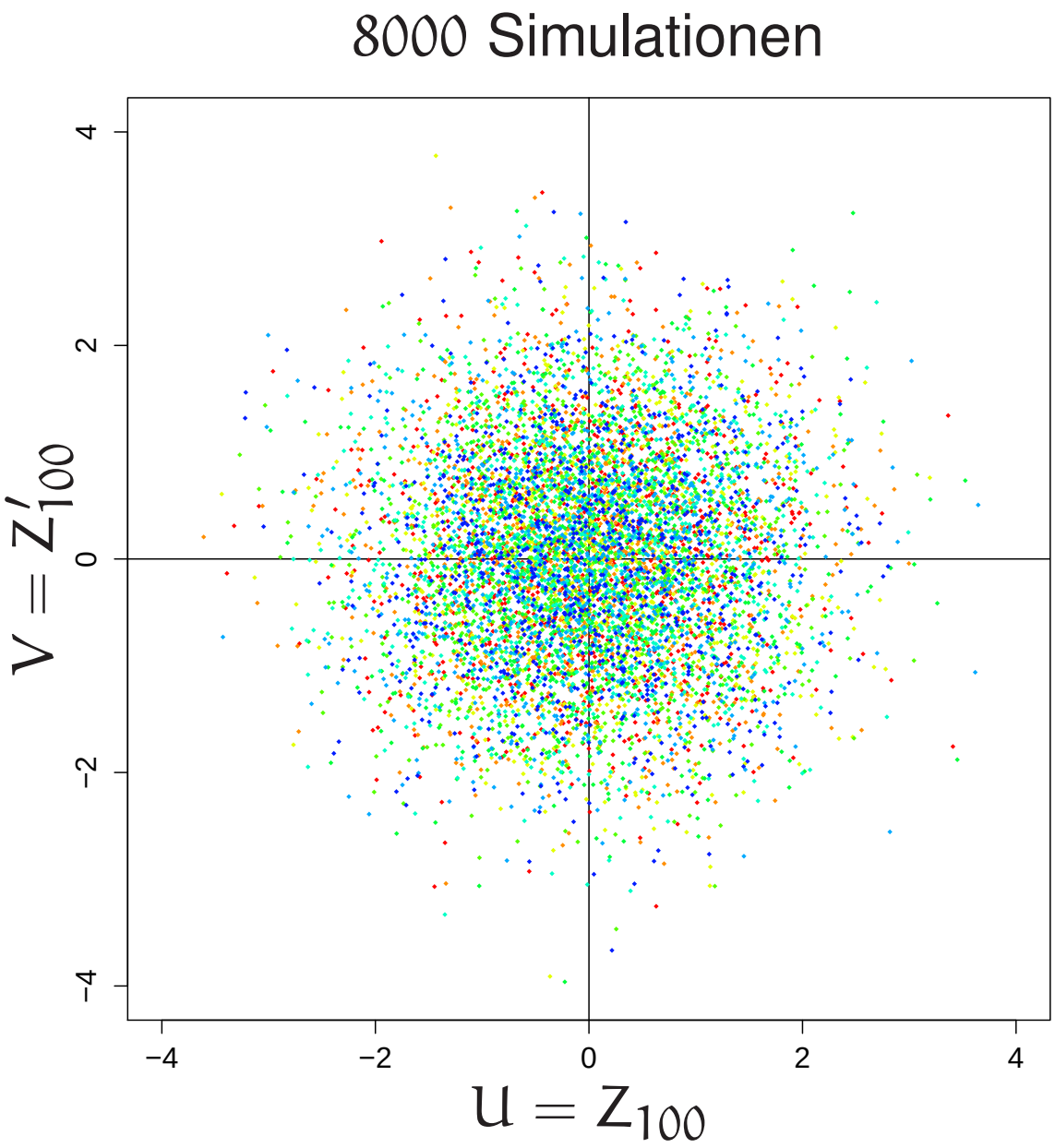

3000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

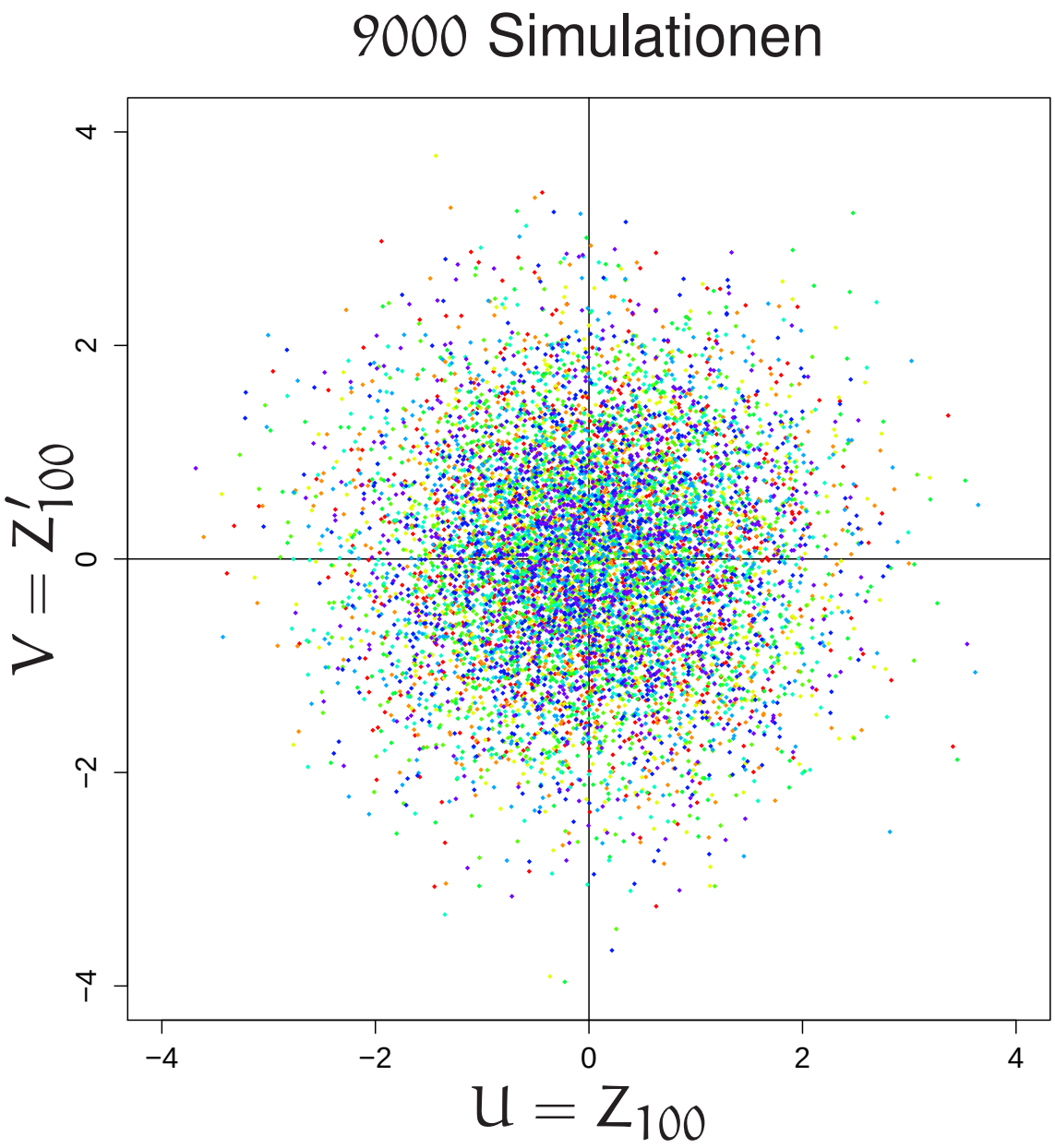

4000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

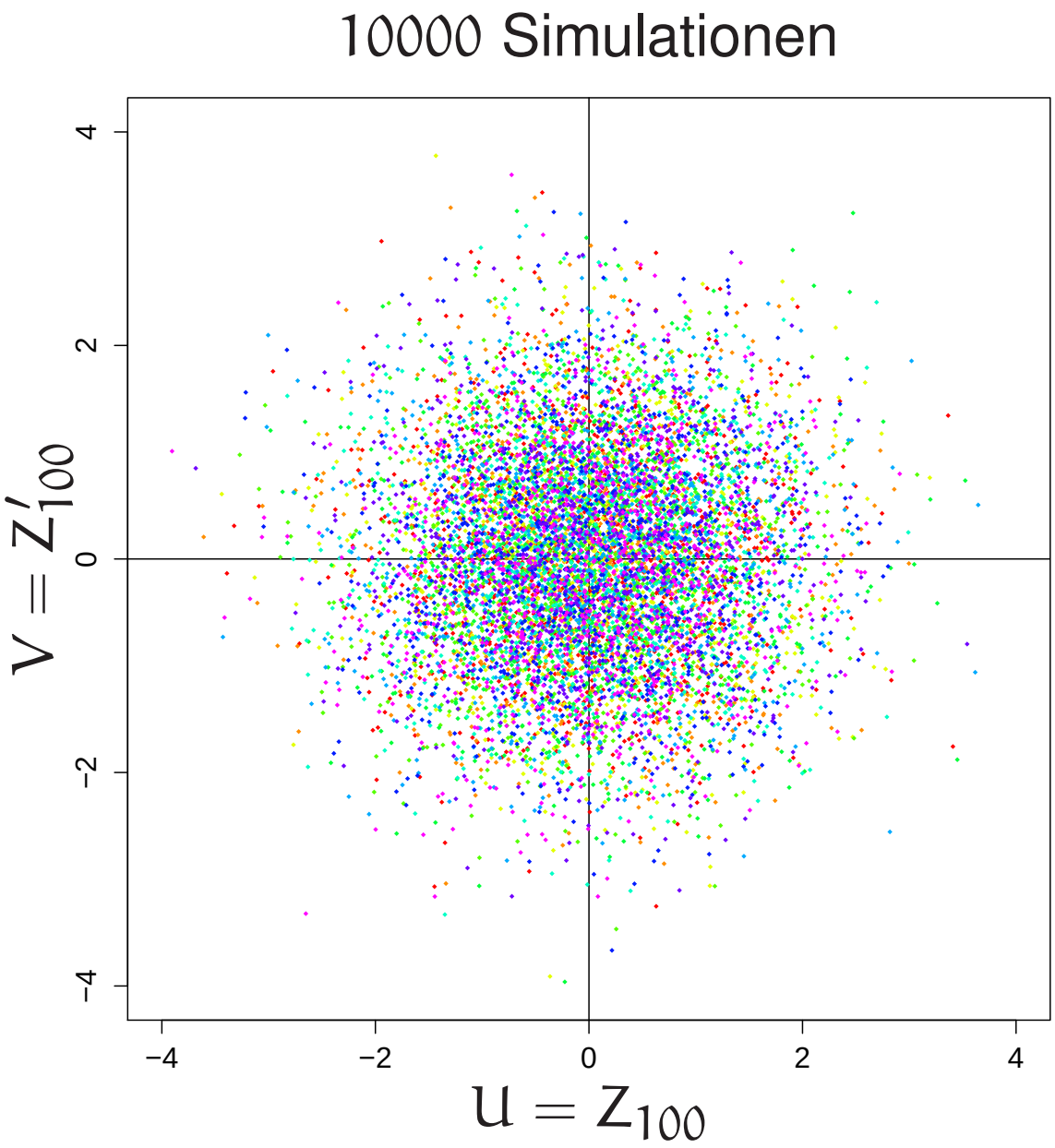

5000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$

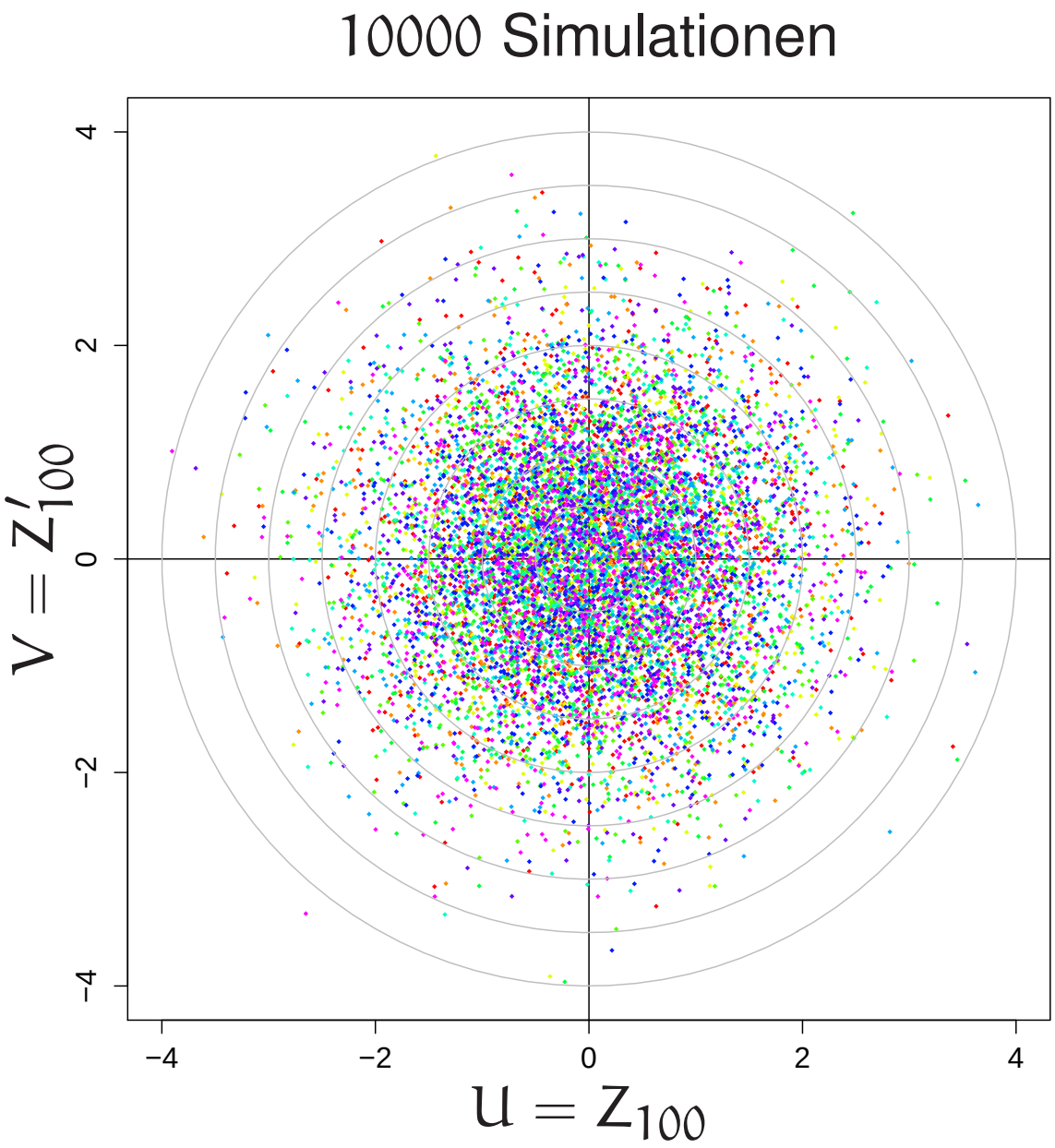

6000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$V = Z'_{100}$
$U = Z_{100}$


7000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$v = z'_{100}$
$u = z_{100}$

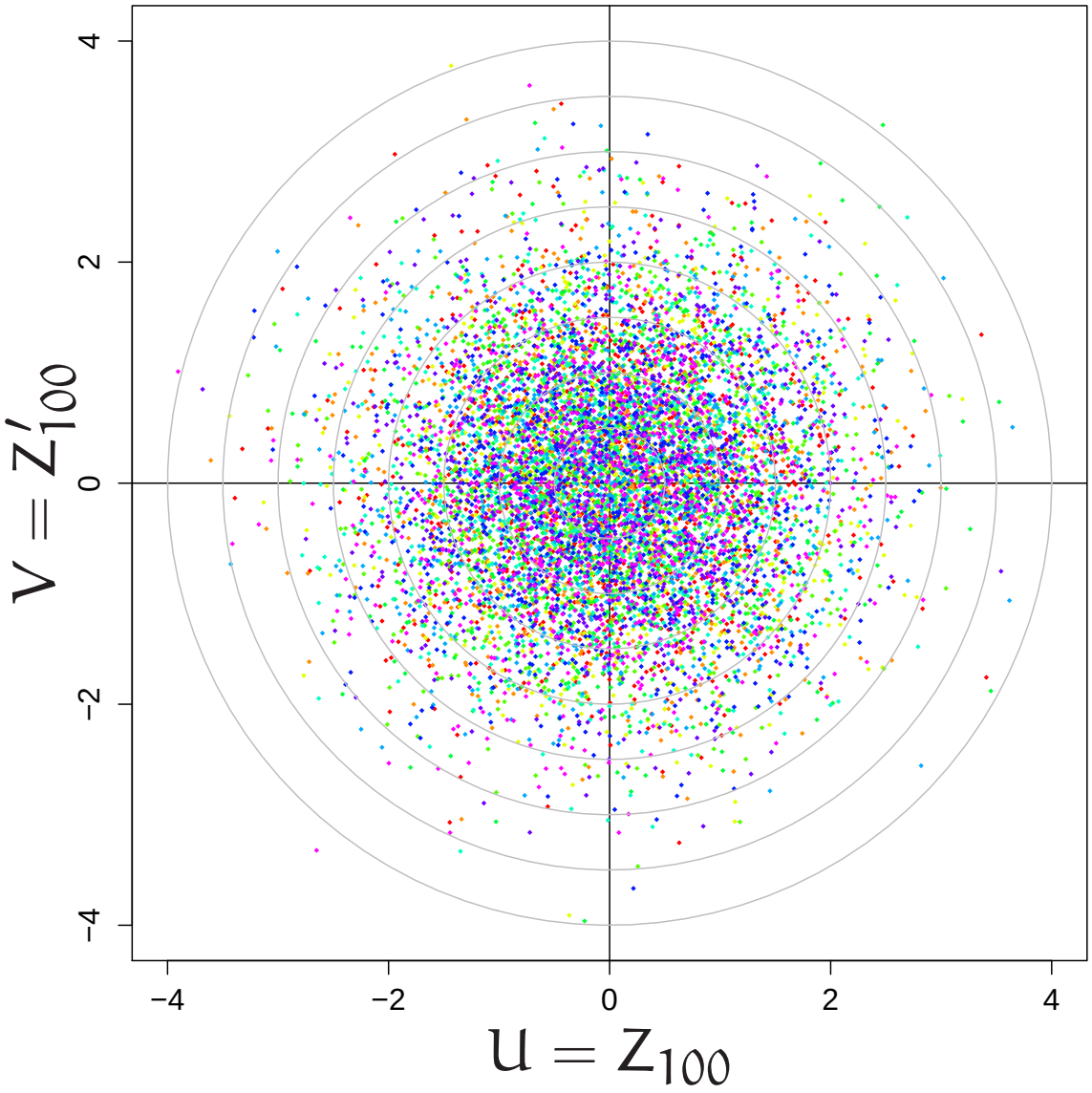

8000 Simulationen
4
2
0
−2
−4
−4
−2
0
2
4
$v = z'_{100}$
$u = z_{100}$


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 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:

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

Eine Lösung hiervon:

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

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

$$w(a^2)w(b^2) = w(0)w(a^2+b^2)$$

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

hat als allgemeine Lösung

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

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

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.”

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.

7. Populationsmittelwert und Stichprobenmittelwert

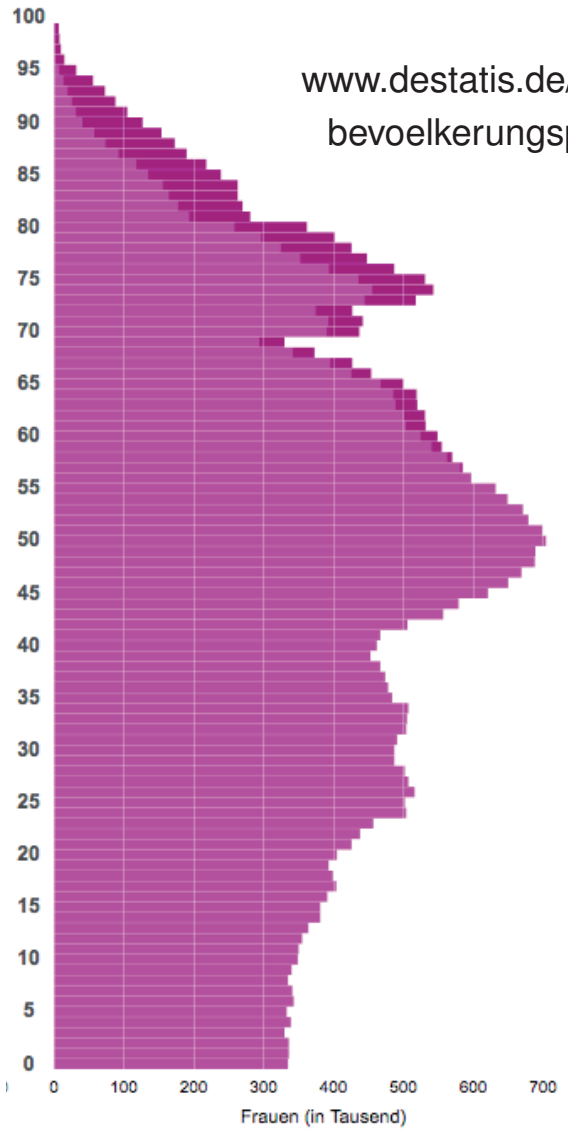
Denken wir an eine Liste (eine “Population”)

von reellen Daten

$$w_1, \ldots, w_g$$

z. B. die Lebensalter aller Frauen

in der deutschen Bevölkerung 2014



www.destatis.de/
bevoelkerungspyramide/

Angenommen man möchte den Populationsmittelwert

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

schätzen,

und zwar aus den Werten einer

aus der Population gezogenen Stichprobe

$$x_1, \ldots, x_n.$$

Als *Schätzwert* für μ bietet sich an:

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

Wie zuverlässig ist diese Schätzung?

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 $\{1, \ldots, g\}$ ("Ziehen mit Zurücklegen").

Wir setzen hier g als (sehr) groß gegenüber n voraus,

damit entstehen *auch*

beim n-maligen Ziehen *mit* Zurücklegen

Kollisionen nur mit (verschwindend) kleiner W'keit.

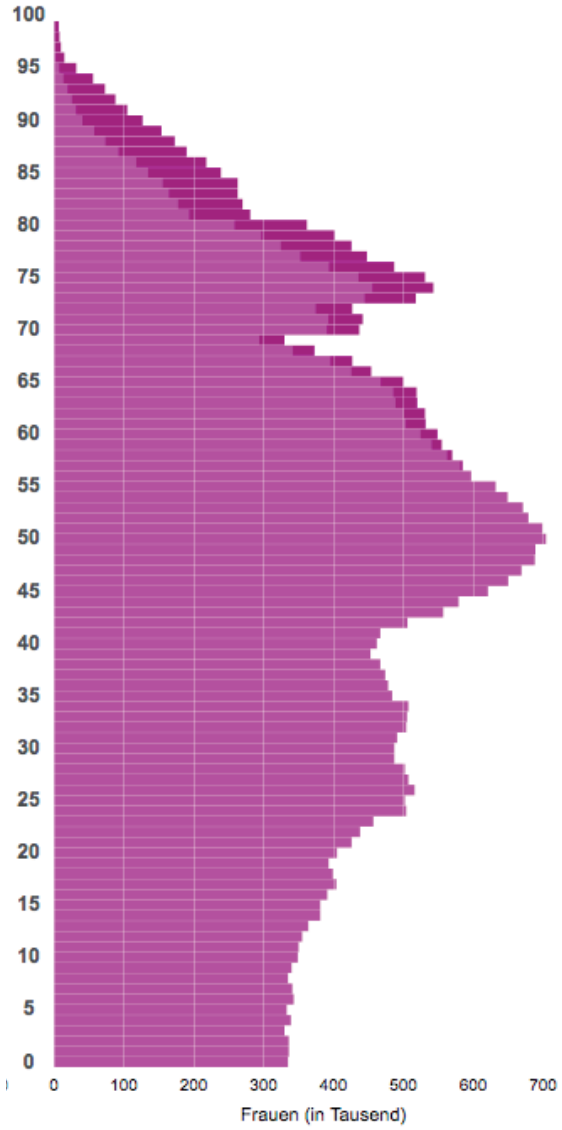
Das führt auf die Vorstellung:

$x_1, \ldots, x_n$ sind entstanden

durch n-maliges unabhängiges Ziehen $X_1, \ldots, X_n$

aus der Verteilung ρ auf $\mathbb{R}$

mit $\rho([c, d]) := \frac{1}{g}\#\{i : w_i \in [c, d]\}$


100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
100
200
300
400
500
600
700
Frauen (in Tausend)

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

fasst man also auf

als *eine* Realisierung (*einen* Ausgang)

der **Zufallsvariable**

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

(Mittelwert der zufälligen Stichprobe $(X_1, \ldots, X_n)$)

Es gilt jedenfalls:

$$\mathbf{E}[M] = \mu$$

Der Erwartungswert des Stichprobenmittelwertes

ist gleich dem Populationsmittelwert.

8. Populationsvarianz
und Varianz des Stichprobenmittelwertes

Der Populationsmittelwert war gleich $\mathbf{E}[X_1] = \mu$.

Wir haben

$$\mathbf{Var}[X_1] = \frac{1}{g} \sum_{j=1}^{g} (w_j - \mu)^2 =: \sigma^2$$

Diese Zahl σ^2 heißt auch die **Populationsvarianz**.

$$M = \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\mathbf{E}[M] = \mu$$

$$M = \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\mathbf{Var}[M] = ?$$

Wird mit Zurücklegen gezogen, dann sind die X_i unabhängig,

und es ergibt sich

$$\mathbf{Var}[M] = \frac{\sigma^2}{n}$$

Wird ohne Zurücklegen gezogen

und wäre die Populationsgröße g nicht sehr groß

gegenüber der Stichprobengröße n, dann hätte es Sinn,

die *Korrektur für endliche Populationen* zu berücksichtigen

(vgl. Aufgabe 23):

$$\mathbf{Var}[M] = \frac{\sigma^2}{n} \cdot \frac{g-n}{g-1}$$

Wird ohne Zurücklegen gezogen

und wäre die Populationsgröße g nicht sehr groß

gegenüber der Stichprobengröße n, dann hätte es Sinn,

die *Korrektur für endliche Populationen* zu berücksichtigen

(vgl. Aufgabe 23):

$$\mathbf{Var}[M] = \frac{\sigma^2}{n} \cdot \frac{g-n}{g-1}$$

Diese Korrektur werden wir

für den Rest dieser Vorlesung vernachlässigen (wir denken

an großes g, oder auch an ein

wiederholtes unabhängiges Ziehen aus einer Verteilung).

Dann gilt:

$$\mathbf{Var}[M] = \frac{\sigma^2}{n};$$

die Standardabweichung des Stichprobenmittelwertes M

ist also $\frac{\sigma}{\sqrt{n}}$.

9. Approximative Verteilung des Stichprobenmittelwertes

Wie ist (für nicht zu kleines n)

der Stichprobenmittelwert M verteilt?

Der Zentrale Grenzwertsatz gibt eine Antwort:

In der oben beschriebenen Situation gilt

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

10. Die Stichprobenvarianz als Schätzung für die Populationsvarianz

Ein Problem in der Praxis: Im Allgem. kenn man σ^2 nicht.

Auch σ^2 muss man dann schätzen.

Zwei Vorschläge für die

(aus der Stichprobe) **geschätzte** (Populations-)**Varianz**:

(i) die *Stichprobenvarianz*

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

(ii) die *modifizierte Stichprobenvarianz*

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

Es gibt theoretische Begründung für beide Vorschläge

(vgl. Buch S. 124, S. 138).

Wir kommen darauf später zurück

und halten uns erst einmal an den Vorschlag (ii):

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

Die Standardabweichung des Stichprobenmittelwertes $\bar{X}$ ist

$$\frac{\sigma}{\sqrt{n}}.$$

Die geschätzte Standardabweichung

des Stichprobenmittelwertes $\bar{X}$ ist

$$\boxed{s/\sqrt{n} =: f}$$

Diese Größe nennen wir auch den *Standardfehler*.

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

Und (gut für die Praxis):

M ist approximativ $N(\mu, f^2)$-verteilt.