Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Verteilung des Korrelationskoeffizienten r

für zwei unabhängige normalverteilte Merkmale

Studie

Autor: Helmut Vetter
Ort, Datum: Arlesheim, 23.07.2017

Diese Arbeit wurde mit TexLive erstellt. Als Berechnungstool wurde Excel und VBA-Basic verwendet.

Verteilung des Korrelationskoeffizienten r
für zwei unabhängige normalverteilte Merkmale

Autor
Vetter, Helmut
Schillerweg 2
CH-4144 Arlesheim
061 599 51 09
helmut.vetter@fhnw.ch

Auftraggeberschaft
Fachhochschule für Wirtschaft
Wombacher, Dr. Jörg

Arlesheim, Juli 2017

Ehrenwörtliche Erklärung
Ich versichere, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der im Literaturverzeichnis angegebenen Quellen und Hilfsmittel angefertigt habe.

Die wörtlich oder inhaltlich den im Literaturverzeichnis aufgeführten Quellen und Hilfsmitteln entnommenen Stellen sind in der Arbeit als Zitat bzw. Paraphrase kenntlich gemacht.

Diese Arbeit ist noch nicht veröffentlicht worden. Sie ist somit weder anderen Interessenten zugänglich gemacht noch einer anderen Prüfungsbehörde vorgelegt worden.

Arlesheim, 23.07.2017

Helmut Vetter

Helmut Vetter

Management Summary

Durch Simulation soll die Verteilung des Korrelationskoeffizienten r zweier unabhängiger normalverteilter Merkmale in Abhängigkeit vom Stichprobenumfang n ermittelt werden. Es wird sich zeigen, dass diese für $n \geq 25$ recht gut einer normalverteilten Zufallsvariable mit $\mu = 0$ und $\sigma = \frac{1}{\sqrt{n-1}}$ entspricht. ↪ Kapitel 1.

Analytisch ergibt sich, dass die Grösse $t = r \cdot \sqrt{\frac{n-2}{1-r^2}}$ t-verteilt ist mit $d = n - 2$ Freiheitsgraden. An den Simulationen in ↪ Kapitel 2 ist dies schön zu sehen.

Die Simulationen wurden in Excel und VBA-Basic programmiert. In ↪ Kapitel 3 ist das Programmcode abgedruckt.

Inhaltsverzeichnis

1 Simulationen für die Zufallsgrösse r bei wachsendem Stichprobenumfang n

Durch jeweils 100'000 Simulation wurde die empirische Wahrscheinlichkeitsdichtefunktion des Korrelationskoeffizienten r zweier unabhängiger normalverteilter Merkmale in Abhängigkeit vom Stichprobenumfang n ermittelt.

Zum Vergleich ist die Wahrscheinlichkeitsdichtefunktion einer normalverteilten Zufallsvariable mit $\mu = 0$ und $\sigma = \dfrac{1}{\sqrt{n-1}}$ eingezeichnet $\hookrightarrow$ gestichelte Linie.

Man erkennt, dass für Stichprobenumfänge $n \geq 25$ die empirische Wahrscheinlichkeitsdichte praktisch mit der Dichtefunktion der Normalverteilung übereinstimmt.

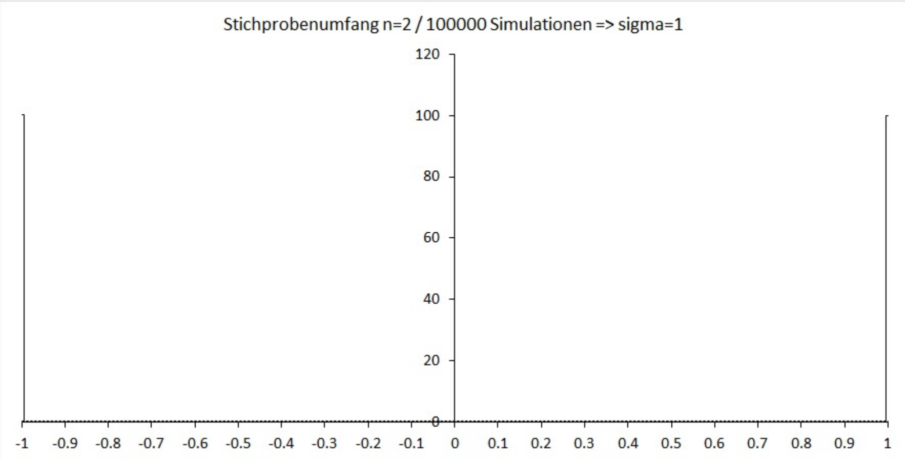

Abb. 1: Stichprobenumfang n=2, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 1)$

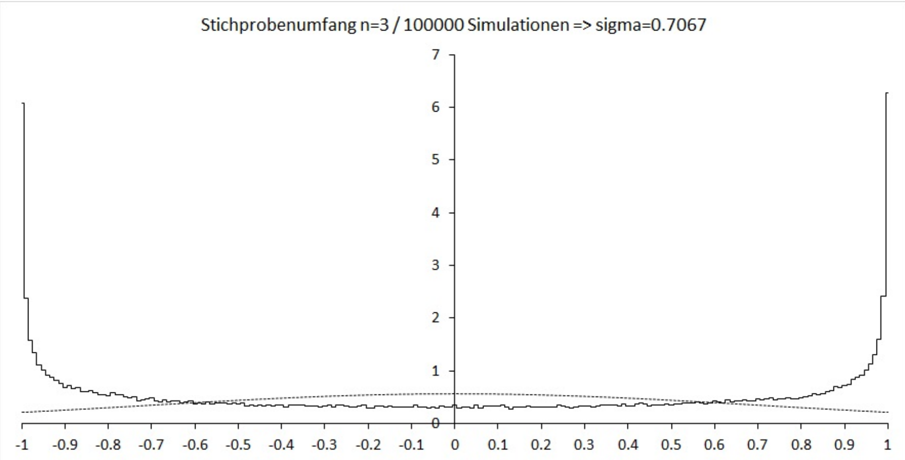

Abb. 2: Stichprobenumfang n=3, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.7071)$

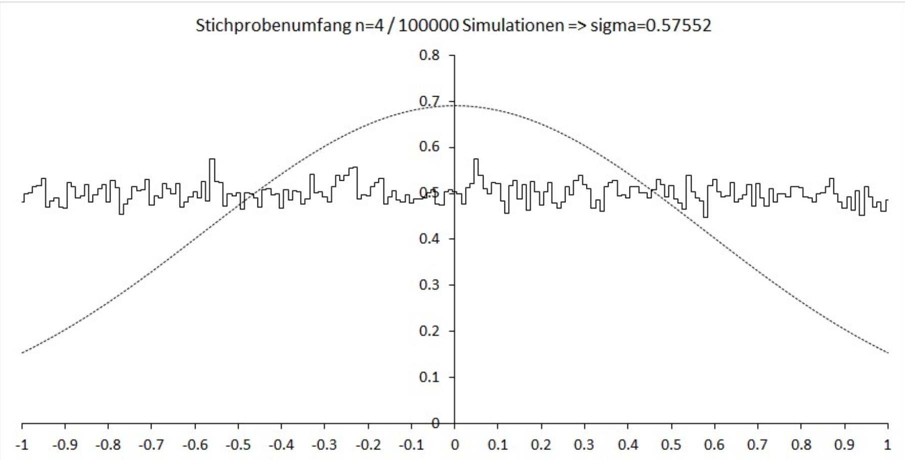

Abb. 3: Stichprobenumfang n=4, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.5774)$

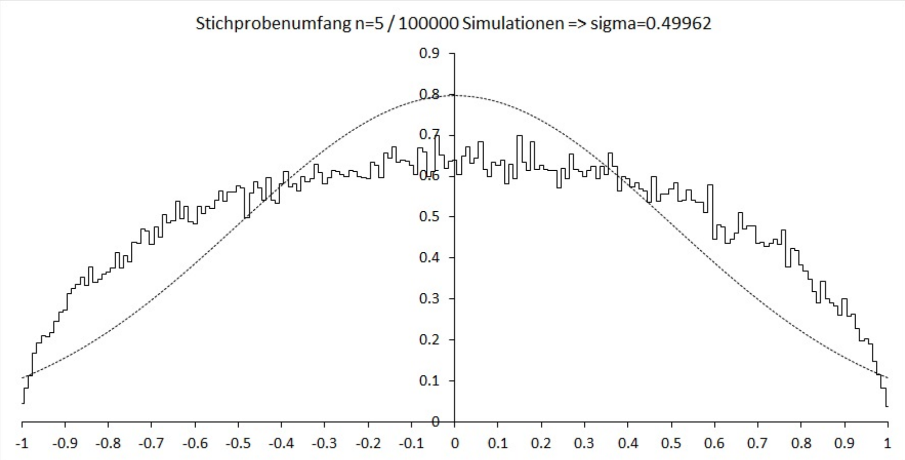

Abb. 4: Stichprobenumfang n=5, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.5)$

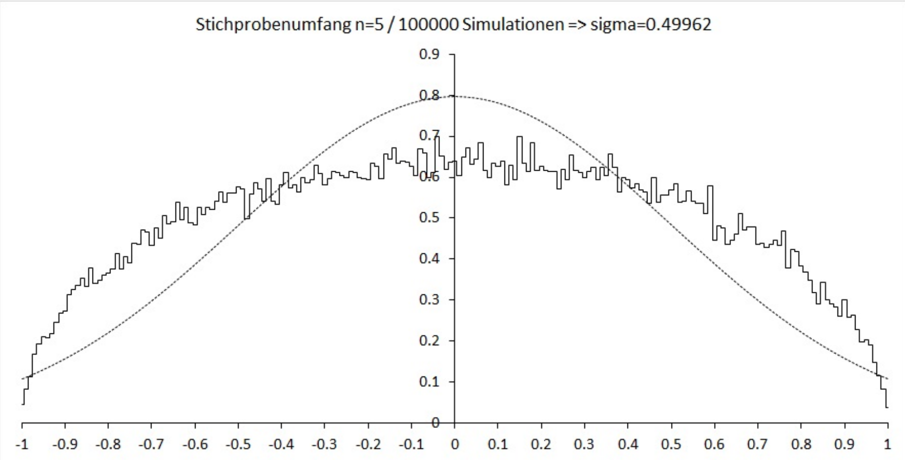

Abb. 5: Stichprobenumfang n=6, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.4472)$

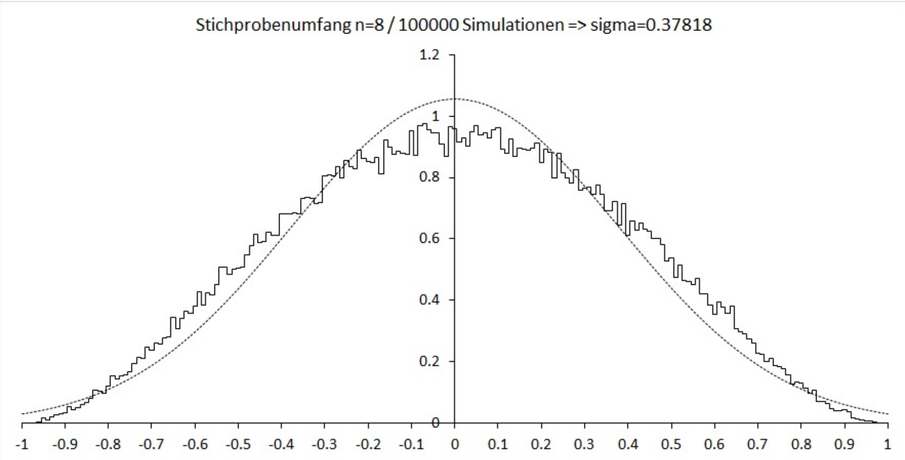

Abb. 6: Stichprobenumfang n=8, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.378)$

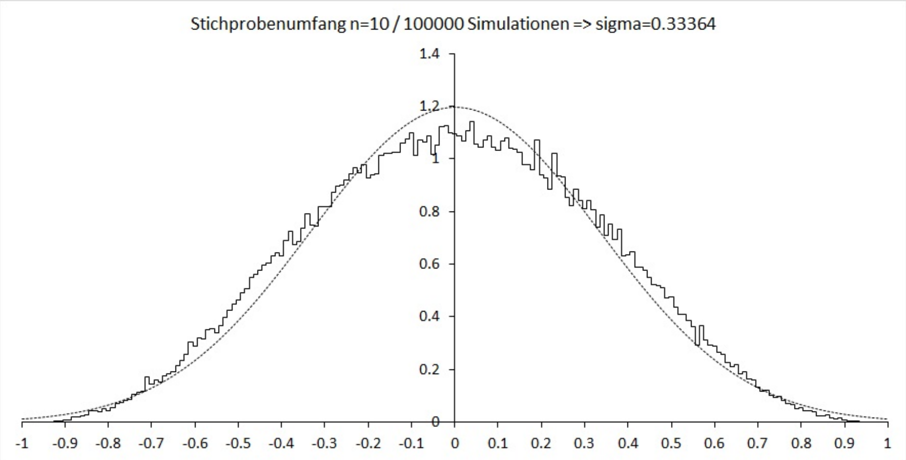

Abb. 7: Stichprobenumfang n=10, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.3333)$

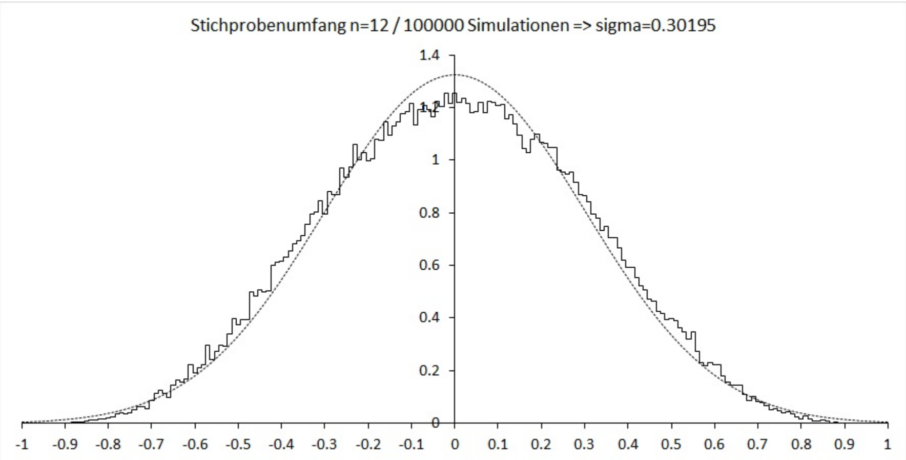

Abb. 8: Stichprobenumfang n=12, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.3015)$

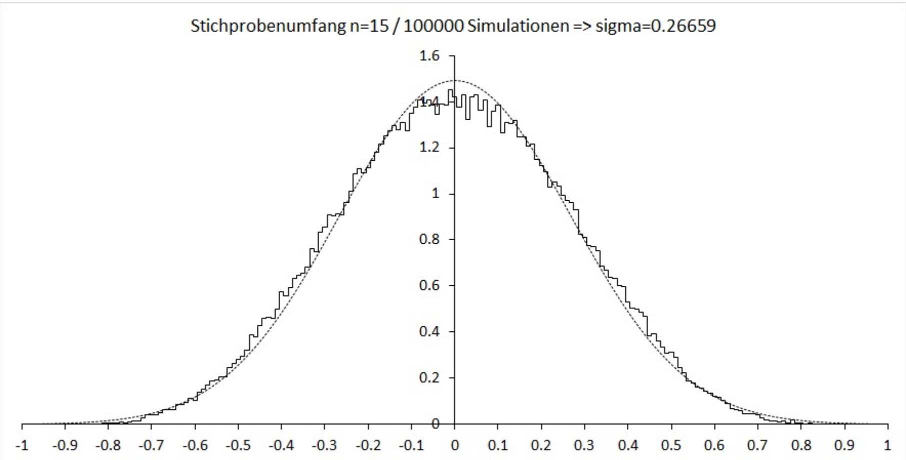

Abb. 9: Stichprobenumfang n=15, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.2673)$

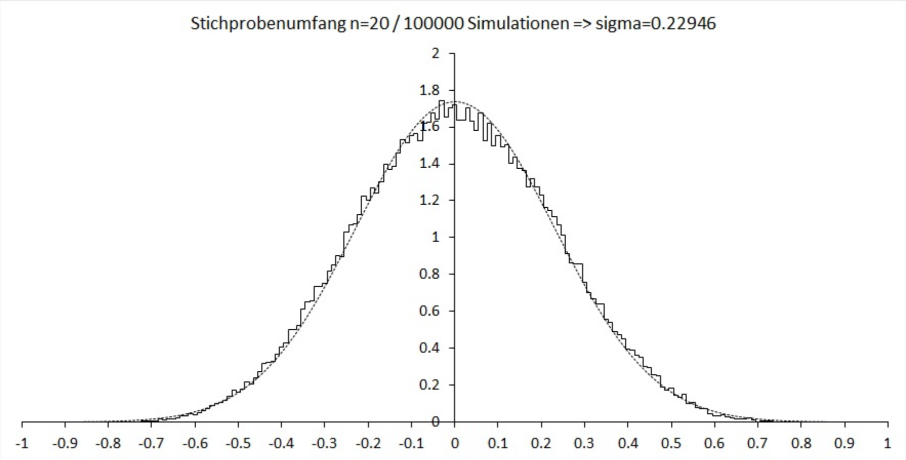

Abb. 10: Stichprobenumfang n=20, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.2294)$

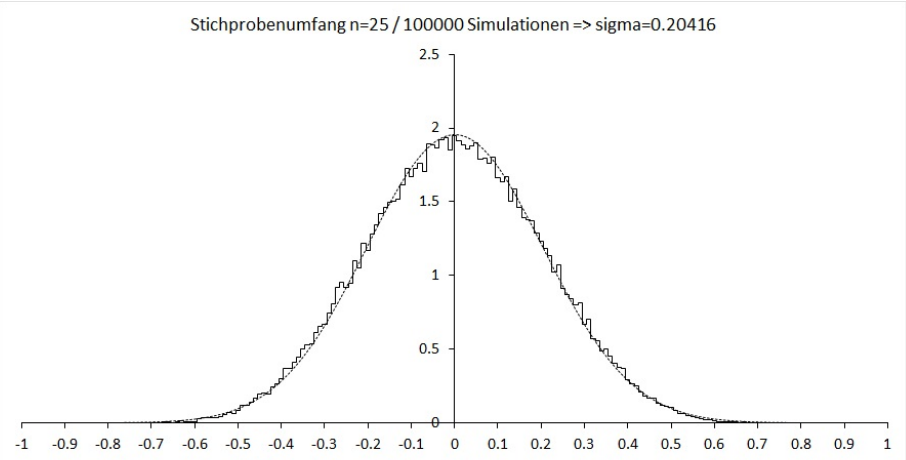

Abb. 11: Stichprobenumfang n=25, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.2041)$

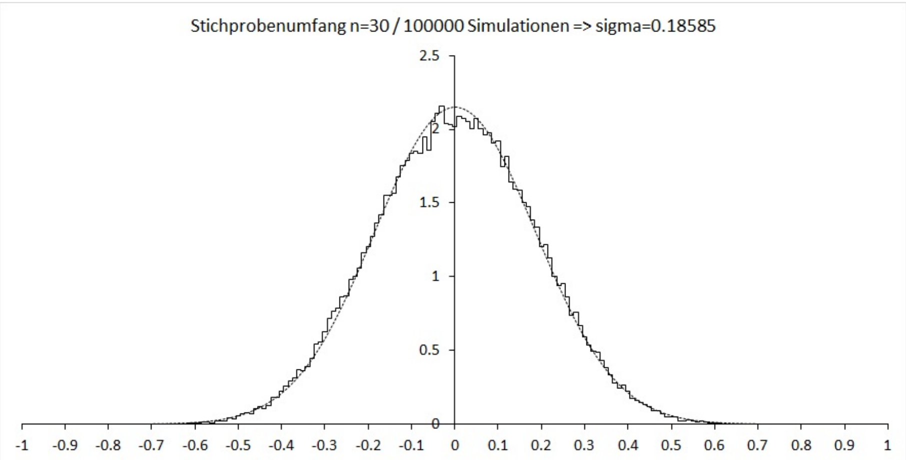

Abb. 12: Stichprobenumfang n=30, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.1857)$

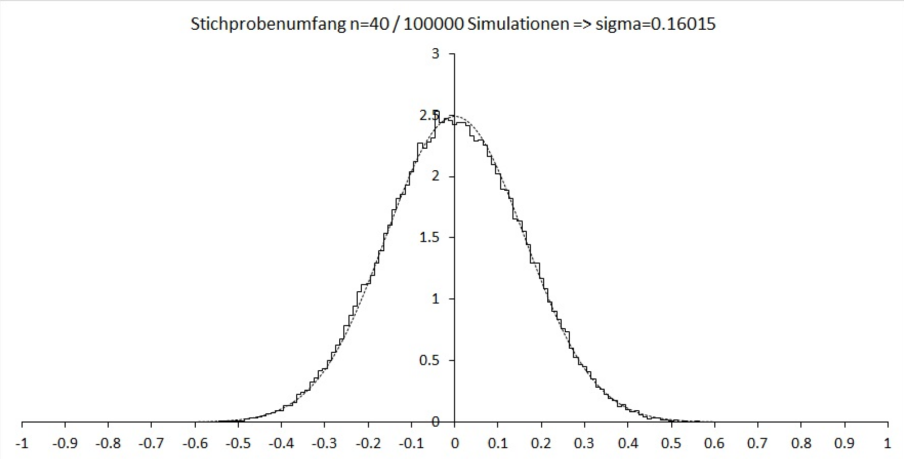

Abb. 13: Stichprobenumfang n=40, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.1601)$

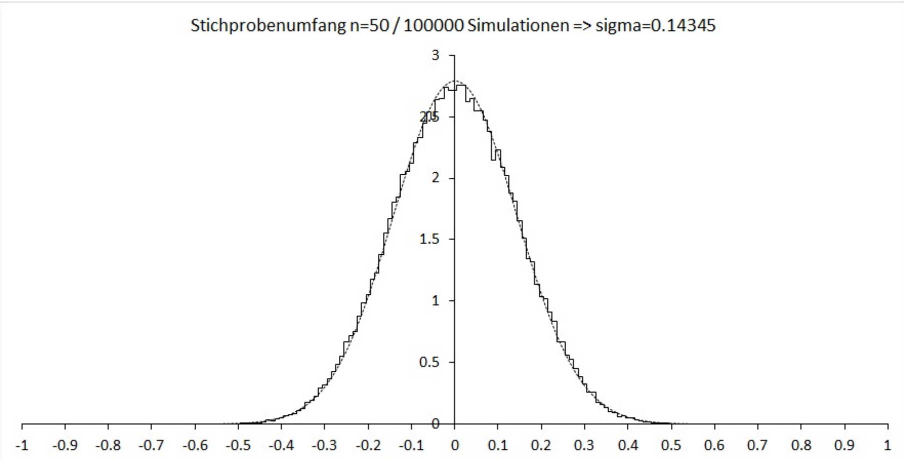

Abb. 14: Stichprobenumfang n=50, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.1429)$

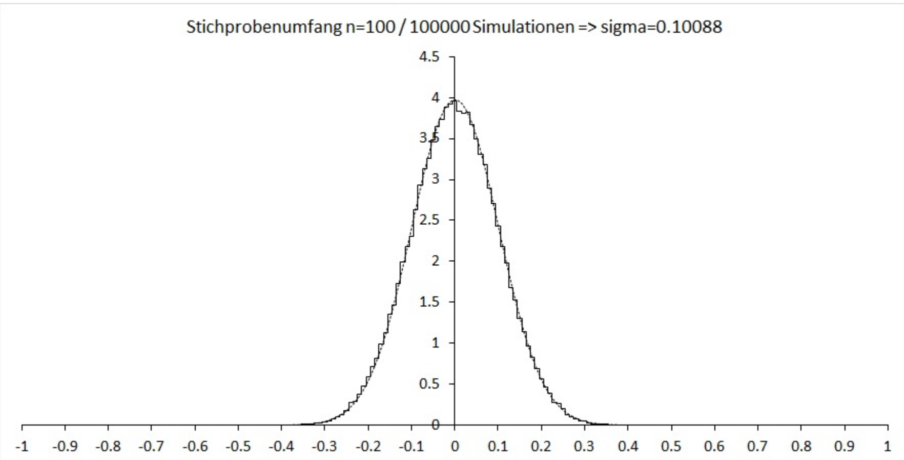

Abb. 15: Stichprobenumfang n=100, 100'000 Simulationen, Vergleich N$(\mu = 0, \sigma = 0.1005)$

2 Simulationen für die Zufallsgrösse $r \cdot \sqrt{\frac{n-2}{1-r^2}}$ bei wachsendem Stichprobenumfang n

Durch jeweils 100'000 Simulation wurde die empirische Wahrscheinlichkeitsdichtefunktion der Zufallsvariablen $r \cdot \sqrt{\dfrac{n-2}{1-r^2}}$ zweier unabhängiger normalverteilter Merkmale in Abhängigkeit vom Stichprobenumfang n ermittelt.

Zum Vergleich ist die (theoretisch korrekte) Wahrscheinlichkeitsdichtefunktion einer t-verteilten Zufallsvariable mit $d = n-2$ Freiheitsgraden eingezeichnet $\hookrightarrow$ gestichelte Linie.

Für $n = 2$ ist die Zufallsvariable $r \cdot \sqrt{\dfrac{n-2}{1-r^2}}$ nicht definiert, da sich in diesem Fall stets $r = \pm 1$ und also $1 - r^2 = 0$ ergeben.

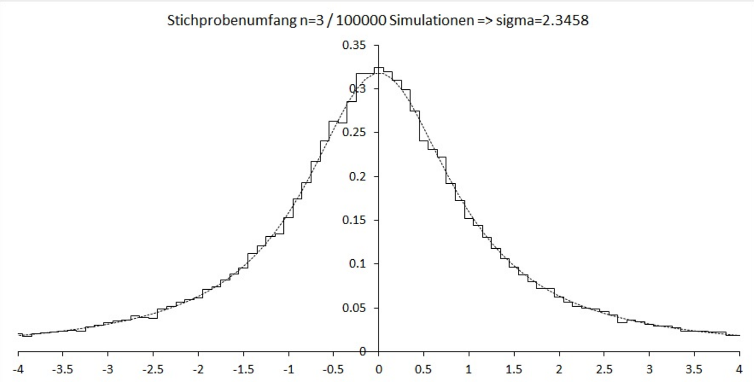

Abb. 16: Stichprobenumfang n=3, 100'000 Simulationen, Vergleich T$(d = 1)$

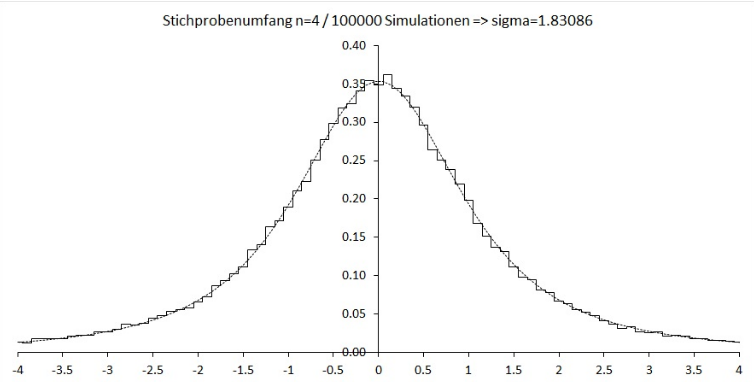

Abb. 17: Stichprobenumfang n=4, 100'000 Simulationen, Vergleich T$(d = 2)$

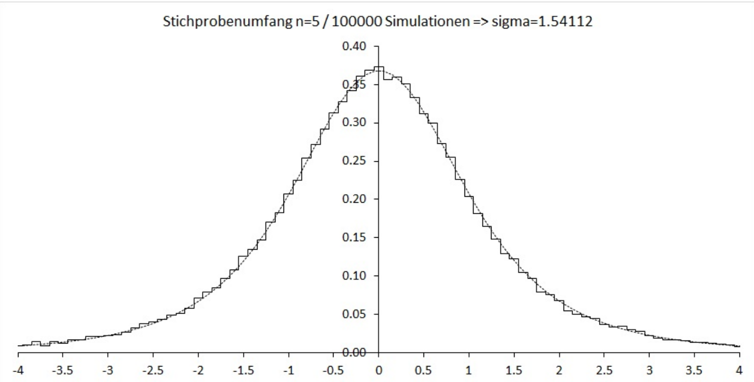

Abb. 18: Stichprobenumfang n=5, 100'000 Simulationen, Vergleich T$(d = 3)$

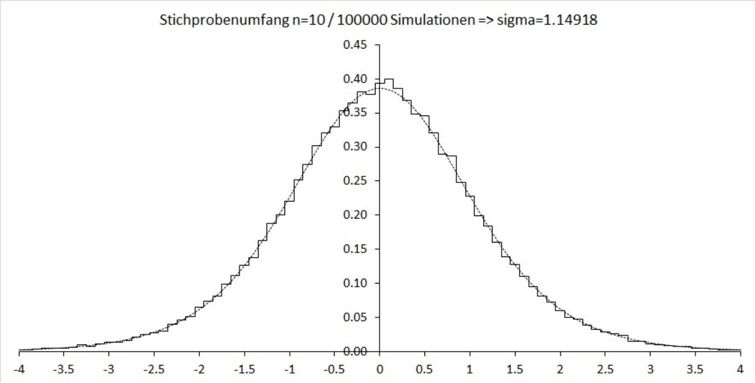

Abb. 19: Stichprobenumfang n=10, 100'000 Simulationen, Vergleich T$(d = 8)$

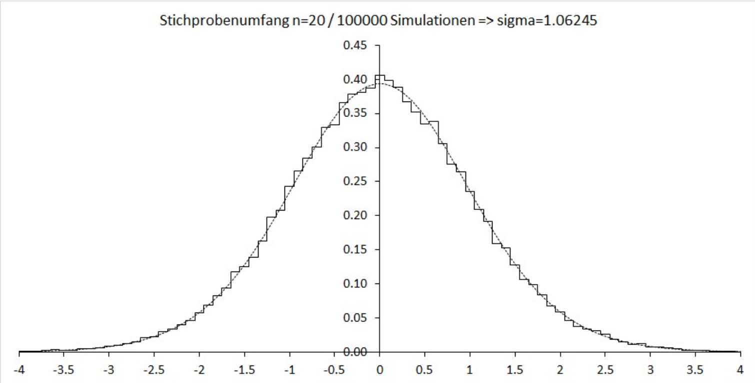

Abb. 20: Stichprobenumfang n=20, 100'000 Simulationen, Vergleich T$(d = 18)$

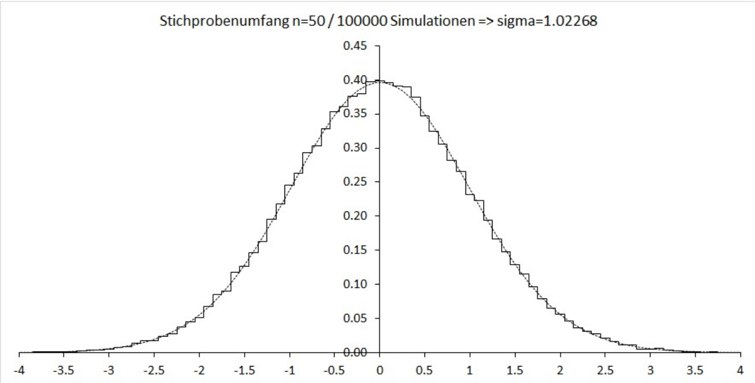

Abb. 21: Stichprobenumfang n=50, 100'000 Simulationen, Vergleich T$(d = 48)$

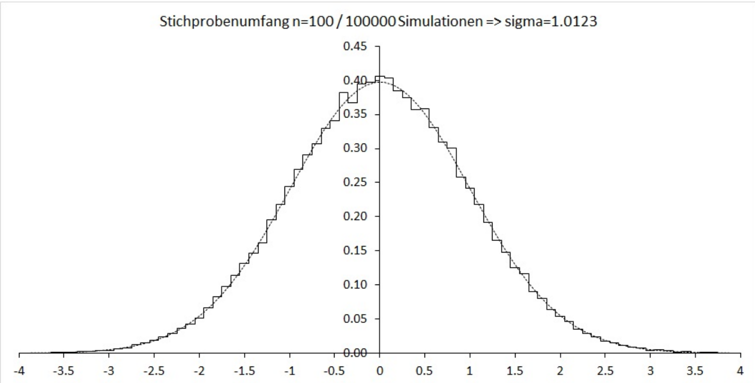

Abb. 22: Stichprobenumfang n=100, 100'000 Simulationen, Vergleich T$(d = 98)$

3 Programmcode

```
Option Explicit
Const KoernungP = 100
Const KoernungQ = 10
Dim P(-KoernungP To KoernungP) As Double
Dim Q(-11 * KoernungQ To 11 * KoernungQ) As Double
```

```
'****************************************
Sub VerteilungKorr(ByVal n As Integer, ByVal Simulationen As Long)        ...simuliert Verteilung von r.
Dim k As Long, Wert As Double, NummerP As Integer, NummerQ As Integer
Dim ra As Range, sum2 As Double
Dim x1(-KoernungP To KoernungP) As Double
Dim x2(-KoernungP To KoernungP) As Double
Dim y(-KoernungP To KoernungP) As Double
'****************************************
For k = 1 To Simulationen
  Wert = Korr(n)
  NummerP = Int(Wert * KoernungP + 0.5)
  P(NummerP) = P(NummerP) + 1
Next k
Set ra = Sheets("tabelle1").Range("A1")
sum2 = 0
For k = -KoernungP To KoernungP
  If k = -KoernungP Then
    x1(k) = -1
    x2(k) = x1(k) + 1 / (2 * KoernungP)
    y(k) = 2 * P(k) * KoernungP / Simulationen
  ElseIf k = KoernungP Then
    x2(k) = 1
    x1(k) = x2(k) - 1 / (2 * KoernungP)
    y(k) = 2 * P(k) * KoernungP / Simulationen
  Else
    x1(k) = (k - 0.5) / KoernungP
    x2(k) = (k + 0.5) / KoernungP
    y(k) = P(k) * KoernungP / Simulationen
  End If
  ra(2 * (KoernungP + k) + 1, 1) = x1(k)
  ra(2 * (KoernungP + k) + 2, 1) = x2(k)
  ra(2 * (KoernungP + k) + 1, 2) = y(k)
  ra(2 * (KoernungP + k) + 2, 2) = y(k)
  sum2 = sum2 + (k / KoernungP) ^ 2 * P(k) / Simulationen
  P(k) = 0
Next k
Sheets("Tabelle1").Range("sigma") = Sqr(sum2 + 1 / (KoernungP ^ 2 * 12))
End Sub
```

```
'****************************************
Sub VerteilungKorrT(ByVal n As Integer, ByVal Simulationen As Long)
```

. . . simuliert Verteilung von $r \cdot \frac{n-2}{1-r^2}$.

```
Dim k As Long, Wert As Double, NummerP As Long, NummerQ As Long
Dim ra As Range, sum2 As Double
Dim x1(-10 * KoernungQ To 10 * KoernungQ) As Double
Dim x2(-10 * KoernungQ To 10 * KoernungQ) As Double
Dim y(-10 * KoernungQ To 10 * KoernungQ) As Double
'****************************************
For k = 1 To Simulationen
  Wert = Korr(n)
  If Wert <> -1 And Wert <> 1 Then
    NummerQ = mymax(-11 * KoernungQ, _
      mymin(11 * KoernungQ, Int(Wert * Sqr((n - 2) / (1 - Wert ^ 2)) * KoernungQ + 0.5)))
    Q(NummerQ) = Q(NummerQ) + 1
  End If
Next k
Set ra = Sheets("tabelle2").Range("A1")
sum2 = 0
For k = -10 * KoernungQ To 10 * KoernungQ
  x1(k) = (k - 1 / 2) / KoernungQ
  x2(k) = (k + 1 / 2) / KoernungQ
  y(k) = Q(k) * KoernungQ / Simulationen
  ra(2 * (10 * KoernungQ + k) + 1, 1) = x1(k)
  ra(2 * (10 * KoernungQ + k) + 2, 1) = x2(k)
  ra(2 * (10 * KoernungQ + k) + 1, 2) = y(k)
  ra(2 * (10 * KoernungQ + k) + 2, 2) = y(k)
  sum2 = sum2 + (k / KoernungQ) ^ 2 * Q(k) / Simulationen
  Q(k) = 0
Next k
Sheets("Tabelle2").Range("sigma") = Sqr(sum2 + 1 / (KoernungQ ^ 2 * 12))
End Sub
```

```
'****************************************
Function Korr(ByVal n As Integer) As Double
```

. . . simuliert Korrelationskoeffizienten r.

```
Dim sumx As Double, sumy As Double, sumx2 As Double
Dim sumy2 As Double, sumxy As Double, rx As Double, ry As Double, i As Integer
'****************************************
sumx = 0
sumy = 0
sumx2 = 0
sumxy = 0
sumy2 = 0
For i = 1 To n
  rx = ZufallNormal()
  ry = ZufallNormal()
  sumx = sumx + rx
  sumy = sumy + ry
  sumx2 = sumx2 + rx ^ 2
  sumxy = sumxy + rx * ry
  sumy2 = sumy2 + ry ^ 2
Next i
Korr = (sumxy * n - sumx * sumy) / Sqr((sumx2 * n - sumx ^ 2) * (sumy2 * n - sumy ^ 2))
```

```
End Function
'**************************************
Function ZufallNormal() As Double                    ...standardnormalverteilte ZV r.
'**************************************
ZufallNormal = ErrInv(Rnd())
End Function
'**************************************
Function ErrInv(ByVal x As Double, Optional ByVal mu As Double = 0, Optional ByVal sigma As Double = 1) As Double
                                                                       ...Hilfsfunktion.
'**************************************
If x = 0 Then
  ErrInv = -10
ElseIf x = 1 Then
  ErrInv = 10
Else
  ErrInv = WorksheetFunction.NormInv(x, mu, sigma)
End If
End Function
'**************************************
Function mymin(ByVal x As Long, ByVal y As Long) As Long               ...Hilfsfunktion.
'**************************************
If x >= y Then
  mymin = y
Else
  mymin = x
End If
End Function
'**************************************
Function mymax(ByVal x As Long, ByVal y As Long) As Long               ...Hilfsfunktion.
'**************************************
If x >= y Then
  mymax = x
Else
  mymax = y
End If
End Function
```