ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

Informatik
& Security

fh
st.pölten

Stetige Verteilungen

Dipl.-Ing. Hubert Schölnast, BSc
Stand: 05. Juli 2022

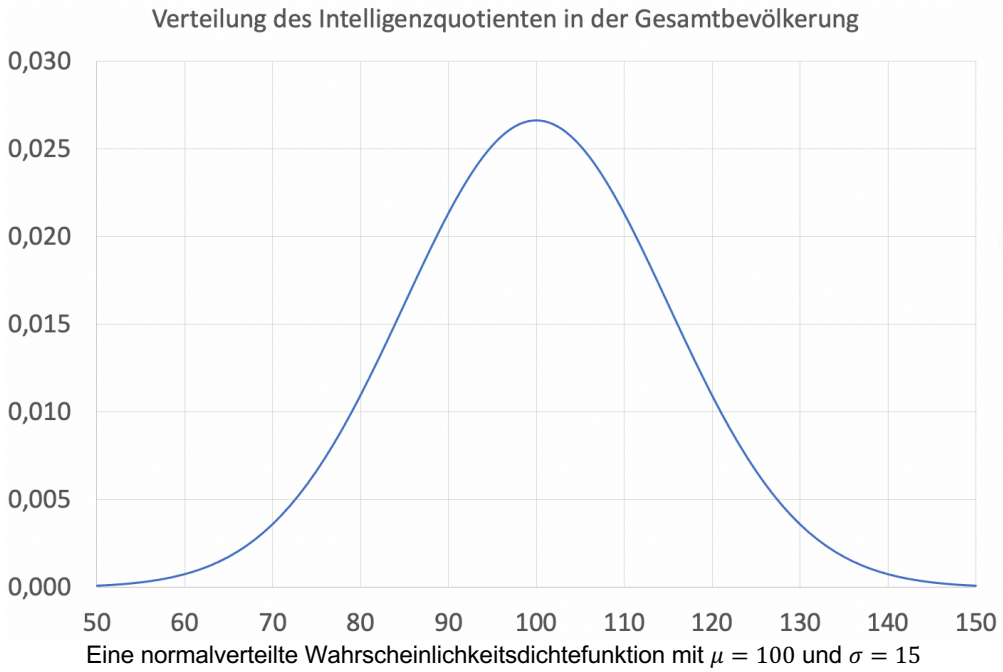
Inhaltsverzeichnis

1 Normalverteilung

Die Normalverteilung ist die am häufigsten auftretende Verteilung. Viele technische oder wirtschaftliche Zusammenhänge lassen sich sehr gut durch diese Form der Verteilung beschreiben.

Die Normalverteilung entsteht überall da »von selbst«, wenn sehr viele sehr kleine zufällige Störungen das Zustandekommen eines konstanten Wertes überlagern. Wenn beispielsweise in einer Brauerei Bier in Flaschen abgefüllt wird, könnte man glauben, dass in allen Flaschen exakt dieselbe Menge Bier landet, aber in Wahrheit schwanken die Füllmengen. Sie schwanken einerseits aus systematischen Gründen, also etwa dadurch, dass sich durch Verschleiß oder ähnliche Einflüsse etwas an der Anlage verändert, so dass die Füllmenge monoton größer oder kleiner wird. Andererseits passiert es aber auch, dass kleinste zufällige Störungen dazu führen, dass die Füllmengen von Flaschen, die unmittelbar hintereinander aus der Abfüllanlage kommen, geringfügig schwanken.

Im Fall von abgefüllten Flüssigkeiten wird es vermutlich gelingen, diese Schwankungen so klein zu halten, dass sie im Verhältnis zur Gesamtfüllmenge vernachlässigbar sind und von den Kunden gar nicht bemerkt werden. Aber wenn z.B. in einer Kiesgrube Schotter auf LKWs geladen wird, wird es hier zu klar messbaren Schwankungen kommen, die jeweils auf mehrere zufällige Einflüsse zurückzuführen sind.



Eine normalverteilte Wahrscheinlichkeitsdichtefunktion mit $\mu = 100$ und $\sigma = 15$

Die Wichtigkeit der Normalverteilung hat ihre Ursache aber auch darin, dass viele andere Verteilungen, egal ob diskret oder stetig, sich beim Anwachsen bestimmter Parameter immer stärker an eine Normalverteilung annähern. Die Normalverteilung ist also so etwas wie eine universelle Näherung für eigentlich alle Verteilungen.

Erstmals ausführlich beschrieben wurde die Normalverteilung von Carl Friedrich Gauß[1], weswegen man sie auch »Gauß-Verteilung« nennt. Die charakteristische Form des Graphen dieser Funktion ist schuld am Namen »Glockenkurve«.

1.1 Dichte- und Verteilungsfunktion

1.1.1 Allgemeine Formel

Die Dichtefunktion der Normalverteilung wird durch zwei Parameter vollständig beschrieben:

- μ Erwartungswert
- σ Standardabweichung

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Etwas weniger furchteinflößend:

$$f(x) = A \cdot e^{-B\cdot(x-\mu)^2}$$

mit

$$A = \frac{1}{\sigma\sqrt{2\pi}} \qquad \mathrm{B} = \frac{1}{2\sigma^2}$$

Die Variable x ist das Argument der Funktion. Der Wert von x gibt an, für welchen Wert der Zufallsvariable X man die Wahrscheinlichkeitsdichte berechnen will.

1.1.2 Wichtige Eigenschaften

Die Dichtefunktion der Normalverteilung ist symmetrisch um den Wert μ

$$f(x) = f(\mu - x)$$

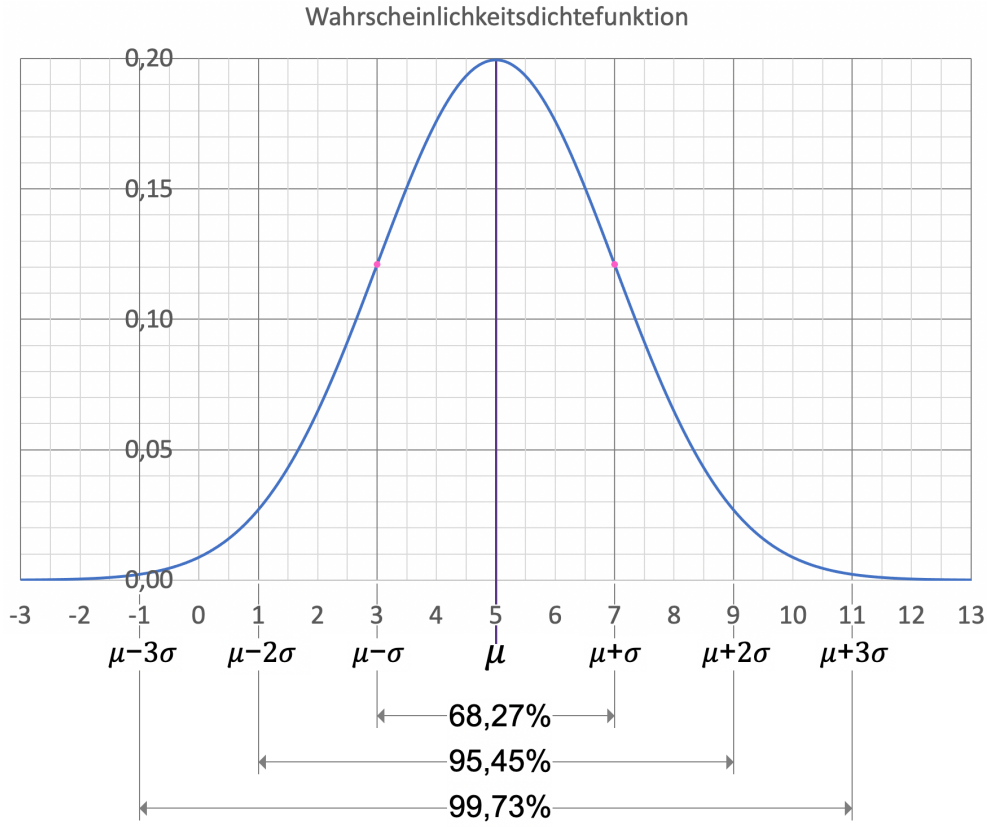
Die Dichtefunktion ist überall positiv, hat ein Maximum bei $x = \mu$, aber sie hat nirgendwo ein Minimum. Stattdessen nähert sie sich asymptotisch dem Wert 0 wenn der Betrag von x gegen unendlich strebt.

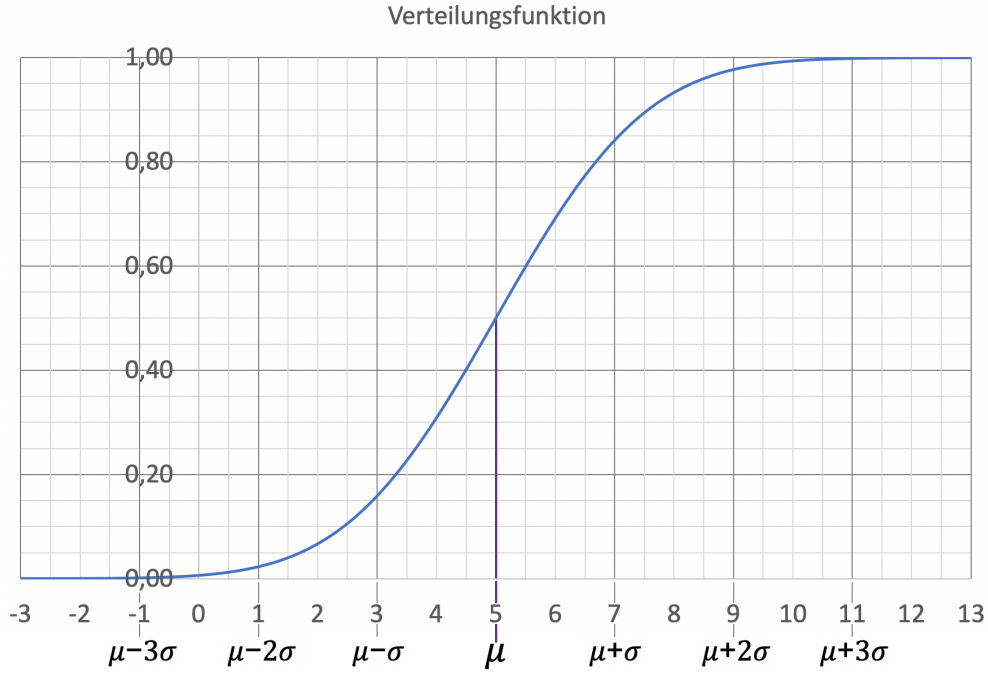
Die Funktion hat 2 Wendepunkte, sie liegen genau bei $x = \mu - \sigma$ und $x = \mu + \sigma$. Die Fläche unter der Kurve, zwischen diesen beiden Wendepunkten beträgt 68,27% der Gesamtfläche.

Die Verteilungsfunktion ist, wie bei allen stetigen Funktionen, das Integral der Dichtefunktion von $-\infty$ bis x.

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

[1] Carl Friedrich Gauß (1777-1855) war ein deutscher Mathematiker, Statistiker, Astronom und Physiker. Man darf Gauß getrost als einen der wichtigsten Mathematiker aller Zeiten nennen. Auf der Wikipedia-Seite über C.F.Gauß sind mehr als 30 Verfahren, Methoden und Ideen aufgelistet, die Gauß entwickelt hat, und die heute mit seinem Namen verbunden sind. Die Gauß'sche Glockenkurve gehört da mit dazu.


Wahrscheinlichkeitsdichtefunktion
0,20
0,15
0,10
0,05
0,00
-3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ
68,27%
95,45%
99,73%


Verteilungsfunktion
1,00
0,80
0,60
0,40
0,20
0,00
-3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ

1.1.3 Standardnormalverteilung

Eine Normalverteilung mit dem Erwartungswert $\mu_z = 0$ und einer Standardabweichung $\sigma_z = 1$ wird als Standardnormalverteilung bezeichnet. Eine normalverteilte und nicht entartete Zufallsvariable X kann mittels der Standardisierungsformel

$$Z = \frac{X - \mu_x}{\sigma_x}$$

in eine Standardnormalverteilung Z transformiert werden.

Der Vorteil einer standardisierten Funktion ist, dass sie leicht in jede beliebige Verteilung umgerechnet werden kann

$$X = \sigma_x \cdot Z + \mu_x$$

Daher reicht es aus, die Standard-Version genauer zu untersuchen.

Die Dichtefunktion der Standardnormalverteilung benötigt keine Parameter, sie hängt nur von x ab:

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

Das gilt auch für die Verteilungsfunktion:

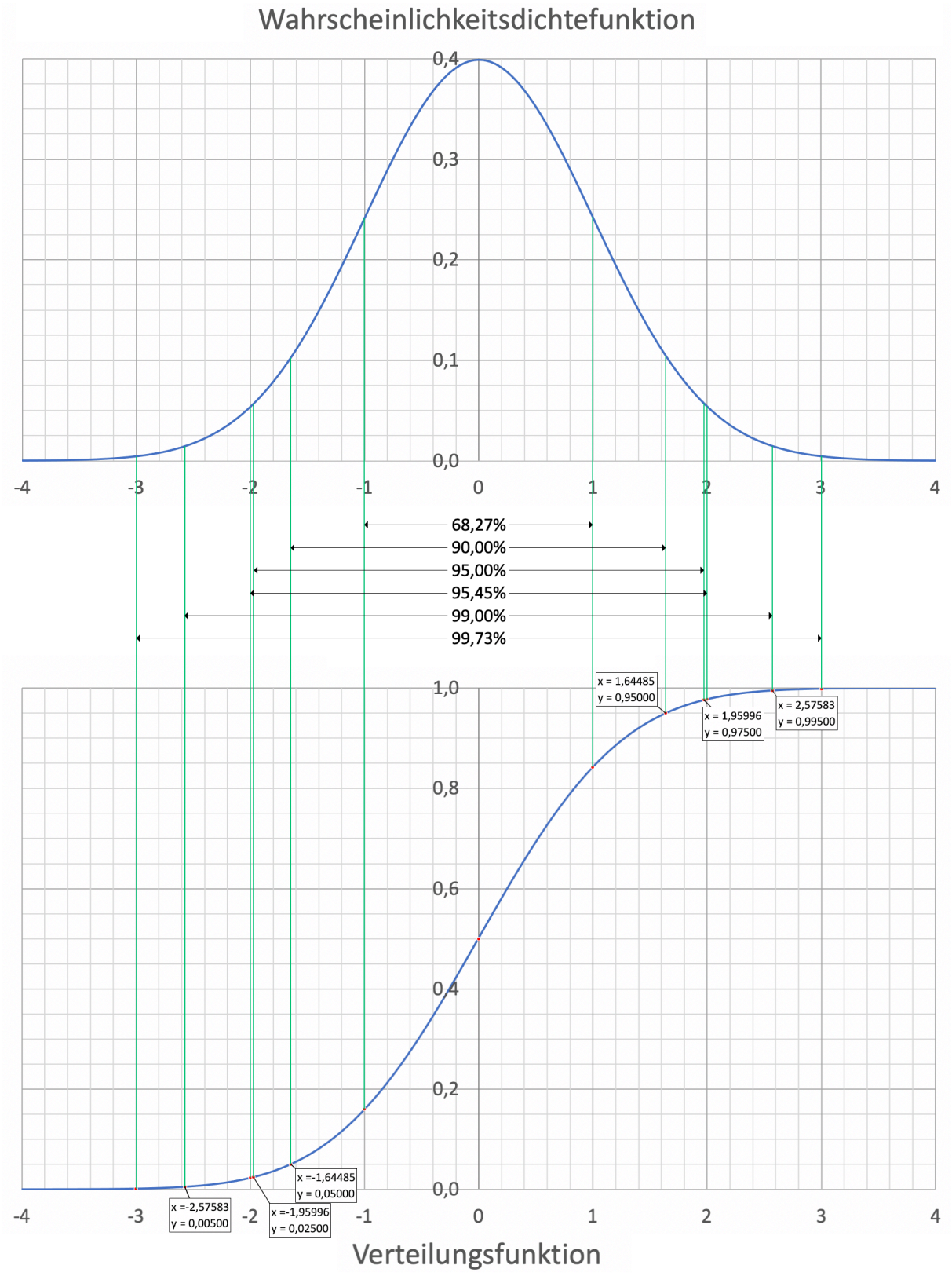
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$

Bei der Standardnormalverteilung liegen ...

- 68,269 % der Fläche unter der Kurve zwischen -1 und 1. (ca. $\frac{2}{3}$, ca. 70%)
- 95,450 % der Fläche unter der Kurve zwischen -2 und 2. (ca. 95%)
- 99,730 % der Fläche unter der Kurve zwischen -3 und 3. (ca. 99,7%)
- 99,994 % der Fläche unter der Kurve zwischen -4 und 4. (4 Neuner)
- 99,99994 % der Fläche unter der Kurve zwischen -5 und 5. (6 Neuner)
- 99,9999998 % der Fläche unter der Kurve zwischen -6 und 6. (8 Neuner)
- 99,9999999997 % der Fläche unter der Kurve zwischen -7 und 7. (11 Neuner)

1.1.4 Wichtige Quantile

- 10,00 % der Fläche unter der Kurve liegen zwischen -0,1257 und 0,1257. (ca. $\frac{1}{8}$)
- 25,00 % der Fläche unter der Kurve liegen zwischen -0,3186 und 0,3186. (ca. $\frac{1}{3}$)
- 50,00 % der Fläche unter der Kurve liegen zwischen -0,6745 und 0,6745. (ca. $\frac{2}{3}$)
- 75,00 % der Fläche unter der Kurve liegen zwischen -1,1503 und 1,1503. (ca. $\frac{7}{6}$)
- 90,00 % der Fläche unter der Kurve liegen zwischen -1,6449 und 1,6449. (ca. $\frac{5}{3}$)



Standardnormalverteilung

1.1.5 Beispiel Körpergröße

Quelle des Beispiels: https://de.wikipedia.org/wiki/Normalverteilung

Die Körpergröße des Menschen ist näherungsweise normalverteilt. Bei einer Stichprobe von 1.284 Mädchen und 1.063 Jungen zwischen 14 und 18 Jahren wurde bei den Mädchen eine durchschnittliche Körpergröße von 166,3 cm (Standardabweichung 6,39 cm) und bei den Jungen eine durchschnittliche Körpergröße von 176,8 cm (Standardabweichung 7,46 cm) gemessen.

Demnach lässt obige Schwankungsbreite erwarten, dass

- 68,3 % der Mädchen eine Körpergröße im Bereich 166,3 cm ± 6,39 cm und
- 95,4 % im Bereich 166,3 cm ± 12,8 cm haben,
- 16 % [≈ (100 % − 68,3 %) / 2] der Mädchen kleiner als 160 cm (und 16 % entsprechend größer als 173 cm) sind und
- 2,5 % [≈ (100 % − 95,4 %) / 2] der Mädchen kleiner als 154 cm (und 2,5 % entsprechend größer als 179 cm) sind.

Für die Jungen lässt sich erwarten, dass

- 68 % eine Körpergröße im Bereich 176,8 cm ± 7,46 cm und
- 95 % im Bereich 176,8 cm ± 14,92 cm haben,
- 16 % der Jungen kleiner als 169 cm (und 16 % größer als 184 cm) und
- 2,5 % der Jungen kleiner als 162 cm (und 2,5 % größer als 192 cm) sind.

1.1.6 Additionssatz der Normalverteilung

Additionssatz der Normalverteilung

- Wenn X eine normalverteilte Zufallsvariable mit dem Erwartungswert μ_x und der Varianz σ_x^2 ist, und
- wenn Y eine davon unabhängige normalverteilte Zufallsvariable mit dem Erwartungswert μ_y und der Varianz σ_y^2 ist, und
- wenn Z die Summe der beiden Zufallsvariablen ist,

$$Z = X + Y$$

dann ist auch Z normalverteilt, mit diesem Parametern:

$$\mu_z = \mu_x + \mu_y$$

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

1.2 Zentraler Grenzwertsatz

Der Zentrale Grenzwertsatz heißt auch Satz von Lindberg-Lévy

Der Additionssatz sagt aus, dass sich die Erwartungswerte und die Varianzen ganz einfach addieren lassen. Das ist aber keine Spezialität der Normalverteilung, sondern gilt für alle Verteilungen solange die Einzelexperimente voneinander unabhängig sind.

Zentraler Grenzwertsatz

Wenn man die Ergebnisse beliebiger voneinander unabhängiger Zufallsexperimente $X_1, X_2, \cdots, X_n$ mit den Erwartungswerten $\mu_1, \mu_2, \cdots, \mu_n$ und den Varianzen $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$ addiert, hat diese Summe den Erwartungswert

$$\mu = \sum_{i=1}^{n} \mu_i$$

und die Varianz

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$$

Aus der dazugehörenden Zufallsvariablen

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

und den bereits erhaltenen Werten μ und σ^2 bzw. σ kann man mit der bereits bekannten Formel

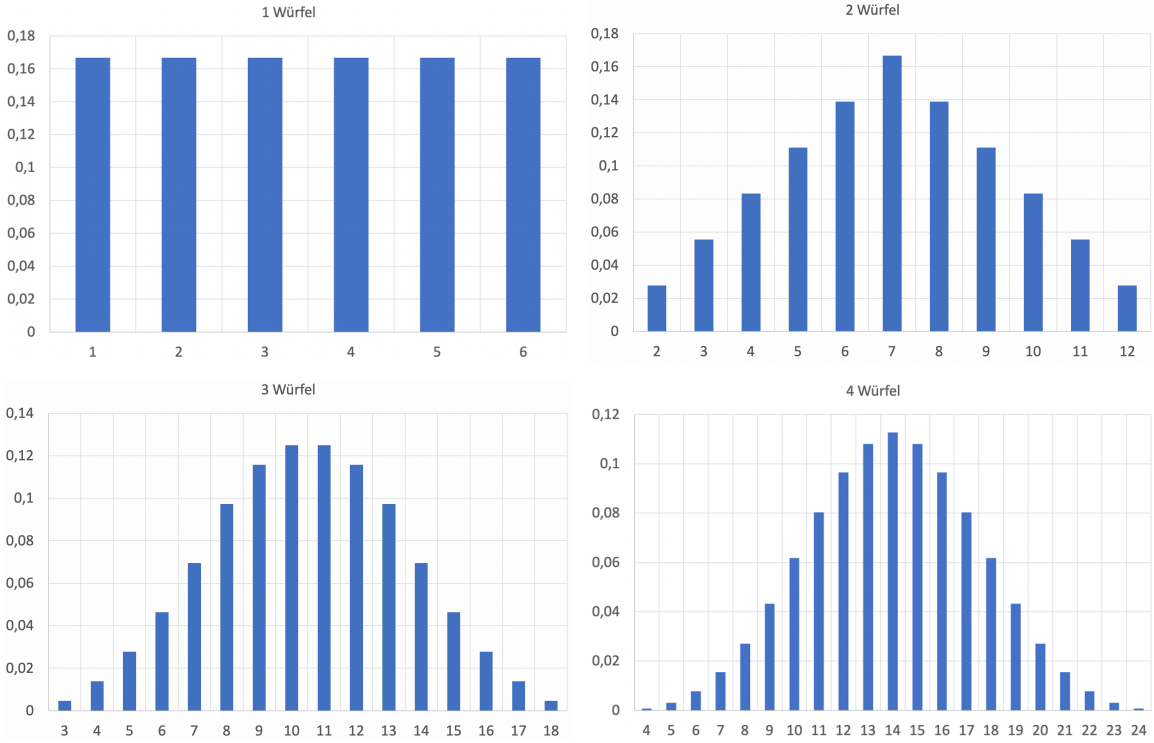
$$Z = \frac{X - \mu}{\sigma}$$

eine standardisierte Zufallsvariable machen, für die folgendes gilt:

$$\lim_{n \to \infty} P(Z \leq z) = \Phi(z)$$

Wenn die Anzahl der Zufallsexperimente ausreichend groß ist, hat die Zufallsvariable, welche die Summe der Ereignisse beschreibt, eine Verteilung, die von einer Normalverteilung nicht mehr zu unterscheiden ist.

1.2.1 Beispiel Augensumme mehrerer Würfel



Würfel	μ	σ^2
1	3,5	$2{,}91667\cdots$
2	7,0	$5{,}83333\cdots$
3	10,5	8,75
4	14,0	$11{,}66667\cdots$

Der Zentrale Grenzwertsatz erlegt den Wahrscheinlichkeits- bzw. Dichte-Funktionen der einzelnen Zufallsexperimenten keinerlei Beschränkung auf. Es ist völlig egal, wie schief oder wie stark gewölbt diese Funktionen sind. Sogar mehrgipfelige Funktionen zeigen dasselbe Verhalten: Wenn man mehrere Experimente mit diesen Wahrscheinlichkeits- bzw. Dichte-Funktionen durchführt, wird die Summe (und damit auch der Mittelwert) der resultierenden Verteilung umso mehr einer Normalverteilung ähneln, je mehr Wiederholungen man macht.

Es ist auch egal, ob die ursprüngliche Verteilung diskret oder stetig war. Wie man am Beispiel des diskret-gleichverteilten Würfel-Wurfs sieht, zeigt bereits die Wahrscheinlichkeitsfunktion für den Wurf von 3 Würfeln eine verblüffende Ähnlichkeit mit einer Normalverteilung.

Das ergibt sich direkt aus dem zentralen Grenzwertsatz.

1.3 Normalverteilung als Näherung

Eine weitere Folge ist, dass auch so gut wie alle Verteilungen beim Anwachsen der Werte für bestimmte Parameter nicht mehr von einer Normalverteilung zu unterscheiden sind.

1.3.1 Normalverteilung als Näherung der Binomialverteilung

Satz von Moivre-Laplace

X sei eine binomialverteilte Zufallsvariable mit den Parametern n und p.

$\Phi(x)$ sei die Verteilungsfunktion der Standardnormalverteilung. Dann gilt

$$\lim_{n\to\infty} P\left(\frac{X - n\cdot p}{n\cdot p\cdot(1-p)} \le x\right) = \Phi(x)$$

Die Anwendung dieses Satzes ist die Approximation der Binomialverteilung durch eine Normalverteilung:

X sei eine binomialverteilte Zufallsvariable mit den Parametern n und p. Wenn die beiden Produkte $n\cdot p$ und $n\cdot(1-p)$ ausreichend groß sind, kann die Verteilungsfunktion der Binomialfunktion $F_B(x)$ durch die Verteilungsfunktion einer Normalverteilung $F_N(x)$ mit folgenden Parametern ersetzt werden:

$$\mu = n\cdot p$$

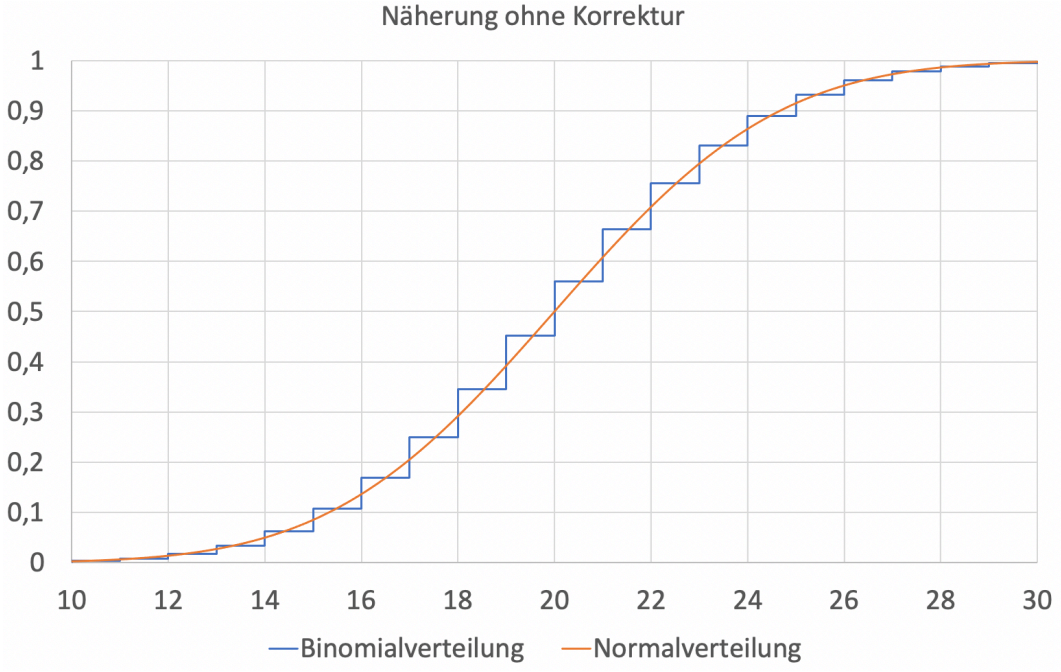
$$\sigma^2 = n\cdot p\cdot(1-p)$$

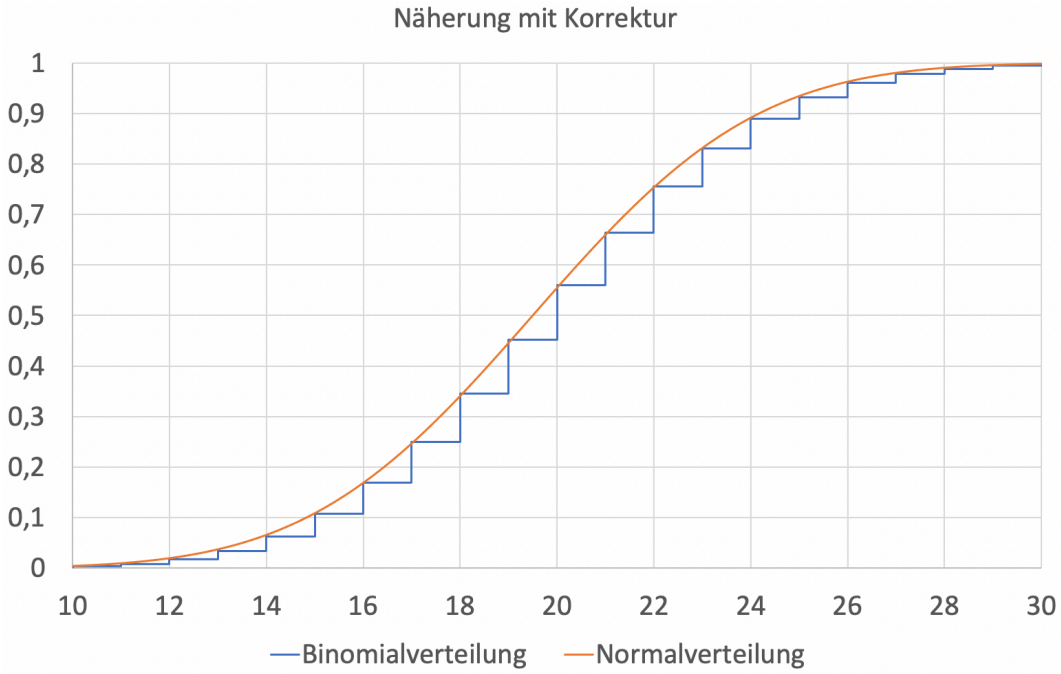
Es gilt dann

$$F_B(x) \cong F_N\left(x + \frac{1}{2}\right)$$

Als Faustregel gilt, dass man diese Näherung anwenden kann, wenn $n\cdot p\cdot(1-p) \ge 9$.

Der Wert $\frac{1}{2}$, der als Stetigkeits-Korrektur zum Wert von x hinzuaddiert wird, bewirkt ein Verschieben der Normalverteilung um 0,5 nach links. Damit wird erreicht, dass die Näherung durch die stetigen Verteilungsfunktion besonders gut dort mit der Binomialverteilung übereinstimmt, wo die Sprungstellen der diskreten Verteilungsfunktion sind, also genau dort, wo die diskrete Verteilung ihre Werte hat.





Normalverteilung als Näherung einer Binomialverteilung mit und ohne Korrektur

1.3.2 **Beispiel:**

Es wird 1000-mal mit einem fairen Spielwürfel gewürfelt und man will die Wahrscheinlichkeit ausrechnen, dass die Augenzahl ⚅ mindestens 100-mal und höchstens 150-mal kommt. (Die beiden Randwerte eingeschlossen.)

Exakte Berechnung:

$$p = \frac{1}{6} \qquad n = 1000$$

$$P_B(100 \le X \le 150) = \sum_{i=100}^{150} \binom{1000}{i} \cdot p^i \cdot (1-p)^{1000-i} \cong 0{,}08369$$

Bedenken Sie dabei, dass z.B.

$$\binom{1000}{150} = \frac{1000 \cdot 999 \cdot 998 \cdot 997 \cdot 996 \cdot 995 \cdot \quad \cdots \quad \cdot 855 \cdot 854 \cdot 853 \cdot 852 \cdot 851}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot \quad \cdots \quad \cdot 146 \cdot 147 \cdot 148 \cdot 149 \cdot 150}$$

Über dem Bruchstrich steht eine Zahl mit 445 Stellen, der Nenner des Bruchs hat 263 Stellen und der Wert des Bruches selbst ist eine Zahl mit 183 Stellen.

$\left(\frac{1}{6}\right)^{150}$ ist eine Dezimalzahl, die mit »0,« beginnt, dann folgen 116 Nullen, bevor die die erste von 0 verschiedene Ziffer kommt. Und $\left(1-\frac{1}{6}\right)^{850}$ hat immerhin noch 67 Nullen nach dem Komma. Da können kleine Rechen- und Rundungsfehler sehr schnell zu einem völlig falschen Ergebnis führen.

Approximation:

Nachdem die Bedingung $n \cdot p \cdot (1-p) \cong 138{,}889 > 9$ erfüllt ist, kann die Approximation angewandt werden.

Berechnen der Parameter

$$\mu = n \cdot p = 1000 \cdot \frac{1}{6} = \frac{1000}{6} \cong 166{,}6666667$$

$$\sigma = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{1000 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \sqrt{\frac{5000}{36}} = \sqrt{\frac{1250}{9}} \cong 11{,}78511302$$

Ausführen der Näherungsrechnung

$$P_N(100 \le X \le 150) = \frac{1}{\sigma\sqrt{2\pi}} \int_{100}^{150} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \cong 0{,}08506$$

Relativer Fehler

$$\Delta_r = \frac{0{,}08506 - 0{,}08369}{0{,}08369} = \frac{0{,}001376}{0{,}08369} = 0{,}01644 \cong 1{,}644\%$$

Der relative Fehler der Näherung beträgt ca. 1,6%.

1.3.3 Normalverteilung als Näherung der Poissonverteilung

Nachdem die Poissonverteilung bereits als Näherung für eine Binomialverteilung verwendet werden kann, letztere aber auch durch eine Normalverteilung approximiert werden kann, liegt es auf der Hand, auch für die Poissonverteilung ein normalverteiltes Näherungsverfahren anzugeben.

X sei eine poissonverteilte Zufallsvariable mit dem Parameter λ. Wenn λ ausreichend groß sind, kann die Verteilungsfunktion der Poissonverteilung $F_P(x)$ durch die Verteilungsfunktion einer Normalverteilung $F_N(x)$ mit folgenden Parametern ersetzt werden:

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

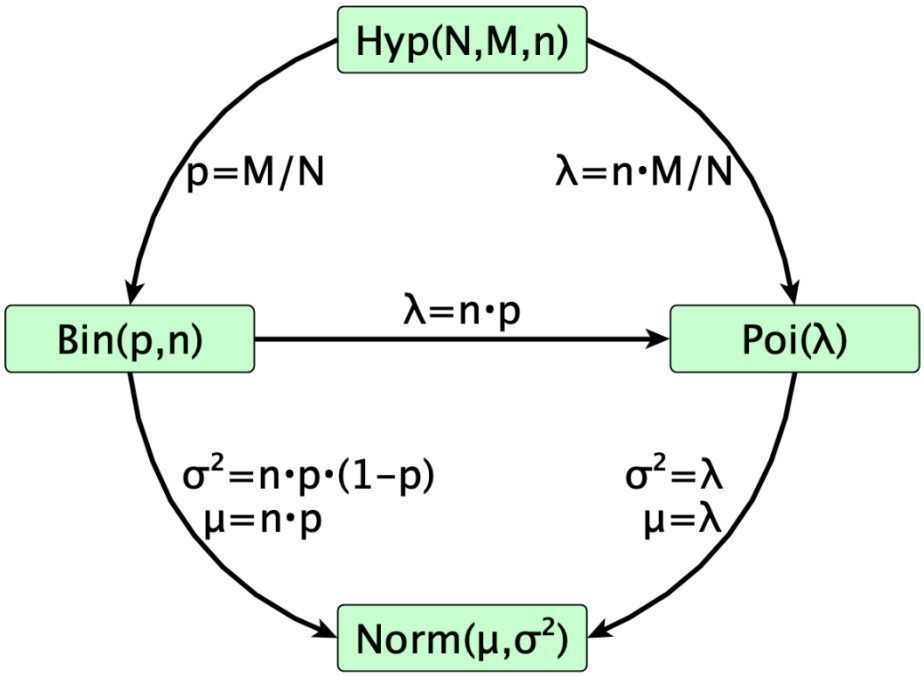
Es gilt dann

$$F_P(x) \cong F_N\left(x + \frac{1}{2}\right)$$

Als Faustregel gilt, dass man diese Näherung anwenden kann, wenn $\lambda \geq 9$.

Der Wert $\frac{1}{2}$, der als Stetigkeits-Korrektur zum Wert von x hinzuaddiert wird, hat dieselbe Begründung wie zuvor bei der Näherung der Binomialfunktion.

1.3.4 Zusammenfassung Näherungen



2 Prüfverteilungen

2.1 Prüfgröße

Eine Prüfgröße ist das Ergebnis eine Rechenvorschrift, mit der die Daten einer Stichprobe zu einem einzelnen Zahlenwert zusammenfasst werden. Dieser Wert kann verwendet werden, um eine Aussage darüber zu treffen, ob die Nullhypothese eines statistischen Tests zutrifft oder nicht.

Man definiert eine bestimmte Irrtumswahrscheinlichkeit, die in die Berechnung der Prüfgröße einfließen. Die Entscheidung für oder gegen die Nullhypothese wird dann durch einen Vergleich der Prüfgröße mit einem Schwellenwert getroffen.

Beispiele für Prüfgrößen sind unter anderen ein Stichprobenmittelwert oder auch das Verhältnis der Varianzen zweier Stichproben.

Eine Prüfgröße ist selbst ebenfalls eine Zufallsvariable mit einer bestimmten Wahrscheinlichkeitsverteilung. Die Wahrscheinlichkeitsverteilungen von Prüfgrößen nennt man Prüfverteilungen. Die drei wichtigsten sind:

- Chi-Quadrat-Verteilung
- t-Verteilung (anderer Name: Studentsche Verteilung)
- F-Verteilung (Fisher-Verteilung)

2.2 Chi-Quadrat-Verteilung

Andere Schreibweise (bei gleicher Aussprache): χ^2-Verteilung

Anderer Name: Helmert-Pearson-Verteilung

Man verwendet diese Verteilung für die Durchführung von Chi-Quadrat-Anpassungs-, Unabhängigkeits- oder Homogenitätstest sowie für die Konstruktion eines Konfidenzintervalls für die Varianz einer normalverteilten Zufallsvariable. Außerdem bildet sie die Grundlage für die t-Verteilung und die F-Verteilung.

Die Chi-Quadrat Verteilung ist eine Testverteilung, also eine Verteilung, die konstruiert wurde, um Hypothesentests durchführen zu können. Sie ist die Verteilung der Quadratsumme standardnormalverteilter Zufallsvariablen.

Im Gegensatz zur Normalverteilung, die über ganz $\mathbb{R}$ definiert ist, ist die Chi-Quadrat-Verteilung nur für nichtnegative reelle Zahlen definiert. (Es gibt keine Funktionswerte für negative Argumente.)

Chi-Quadrat-Verteilung

Es seien $Z_1, Z_2, \cdots, Z_k$ standardnormalverteilte Zufallsvariablen (ihre Anzahl ist k), die paarweise unabhängig sind und es sei

$$X = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

die Summe der Quadrate dieser Zufallsvariablen. Dann nennt man die Verteilung der Zufallsvariablen X eine Chi-Quadrat-Verteilung mit k Freiheitsgraden.

Der Erwartungswert ist

$$E(X) = k$$

Die Varianz ist

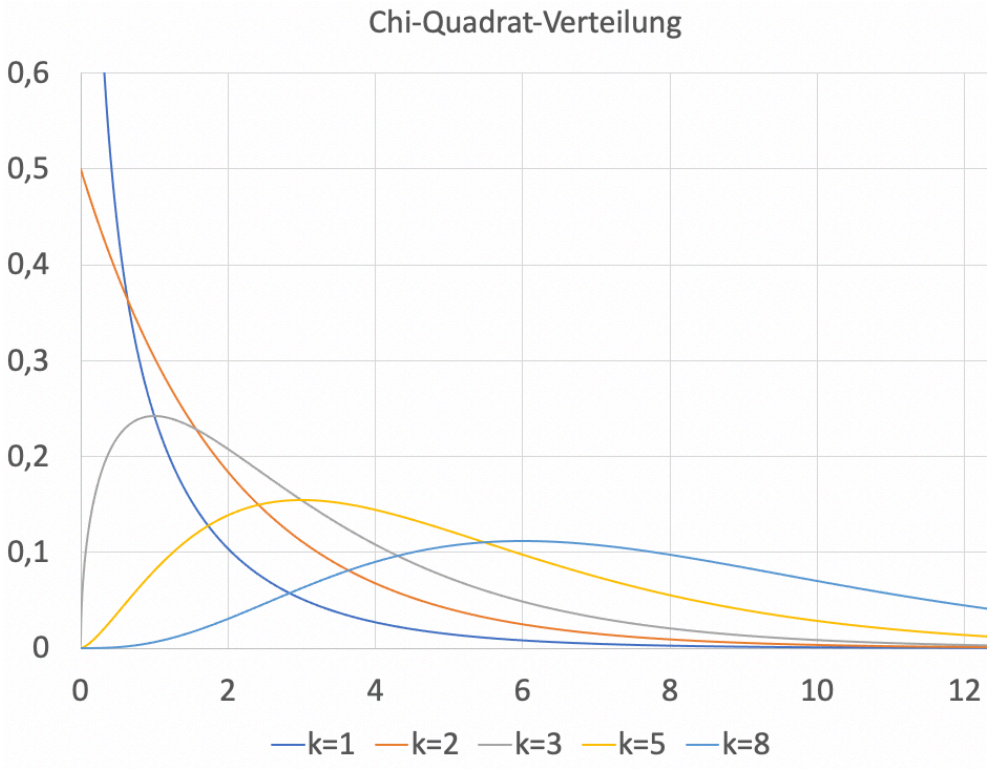
$$\mathrm{VAR}(X) = 2k$$

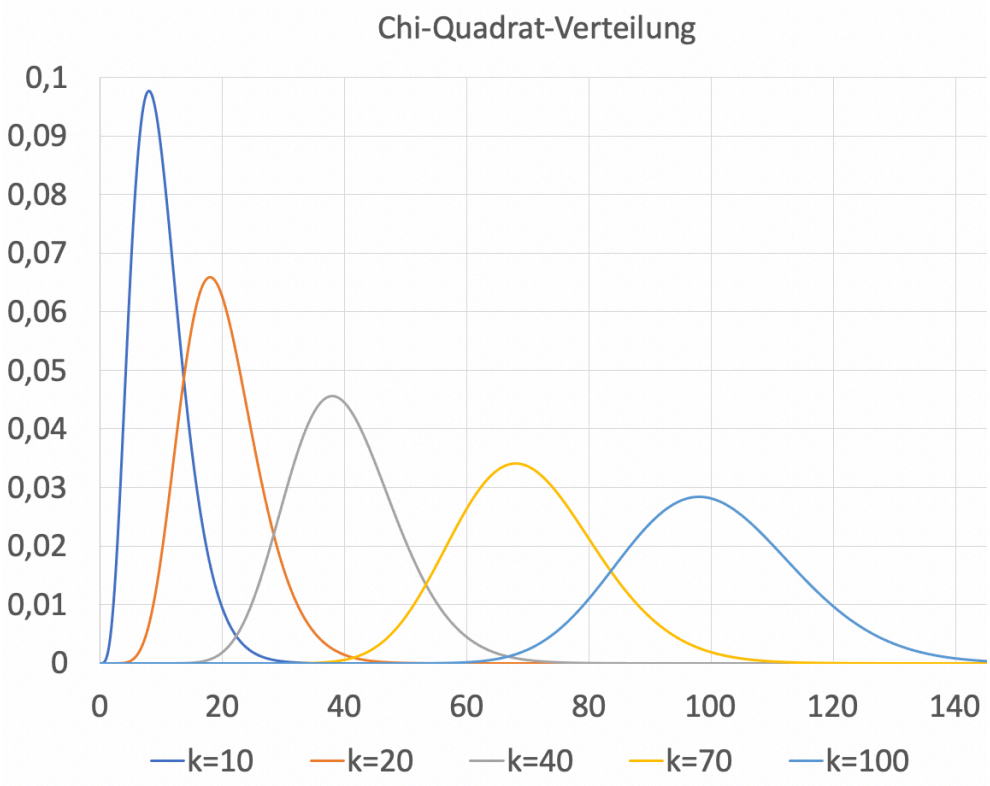
Die Anzahl der Freiheitsgrade kann man sich als die Anzahl der »frei verfügbaren« Beobachtungen vorstellen. Wenn der Umfang einer Stichprobe n ist, und man aus dieser Stichprobe p verschiedene Parameter schätzt, dann ist der Freiheitsgrad $k = n - p$.

Beispiel:

Man kennt zwei Parameter ($p = 2$), bei einer Stichprobe, die aus 5 Messungen besteht ($n = 5$). Die Paramater sind beispielsweise der Mittelwert und die Standardabweichung. Wenn man die Messwerte verloren hat, und sie rekonstruieren will, kann man nur 3 der 5 Messwerte frei wählen. Die beiden letzten Messwerte müssen dann ganz bestimmte Werte annehmen, damit die Parameter zu allen 5 Messwerten passen. Weil man nur 3 Messwerte frei wählen kann, beträgt der Freiheitsgrad in diesem Beispiel 3.

2.2.1 Wahrscheinlichkeitsdichtefunktion





Der Funktionswert bei $x = 0$ strebt im Fall von $k = 1$ gegen unendlich, ist bei $k = 2$ gleich $\frac{1}{2}$ und bei allen anderen Freiheitsgraden gleich 0. Die Kurve hat bei allen Freiheitsgraden ab $k = 3$ genau 1 Maximum, das ein klein wenig links vom Erwartungswert liegt. Je größer der Freiheitsgrad ist, desto ähnlicher wird die Chi-Quadrat-Verteilung einer Normalverteilung.

Die Wahrscheinlichkeitsdichtefunktion der Chi-Quadrat-Verteilung lässt sich mit dieser Formel für $x > 0$ berechnen:

$$f(x) = \frac{x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)}$$

Dabei ist k der Freiheitsgrad und $\Gamma(n)$ ist die Gamma-Funktion, das ist eine Verallgemeinerung der Fakultät. Es gilt

$$\Gamma(n) = (n-1)!$$

bzw.

$$\Gamma(n) = \Gamma(n-1) \cdot n$$

Spezielle Werte:

$$\Gamma(1) = 0! = 1$$

$$\Gamma(2) = 1! = 1$$

$$\Gamma(3) = 2! = 2$$

$$\Gamma(4) = 3! = 6$$

$$\Gamma(5) = 4! = 24$$

aber auch

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \cong 1{,}77245$$

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2} \cdot \sqrt{\pi} \cong 0{,}88623$$

$$\Gamma\left(\frac{5}{2}\right) = \frac{1 \cdot 3}{2 \cdot 2} \cdot \sqrt{\pi} \cong 1{,}32934$$

$$\Gamma\left(\frac{7}{2}\right) = \frac{1 \cdot 3 \cdot 5}{2 \cdot 2 \cdot 2} \cdot \sqrt{\pi} \cong 3{,}32335$$

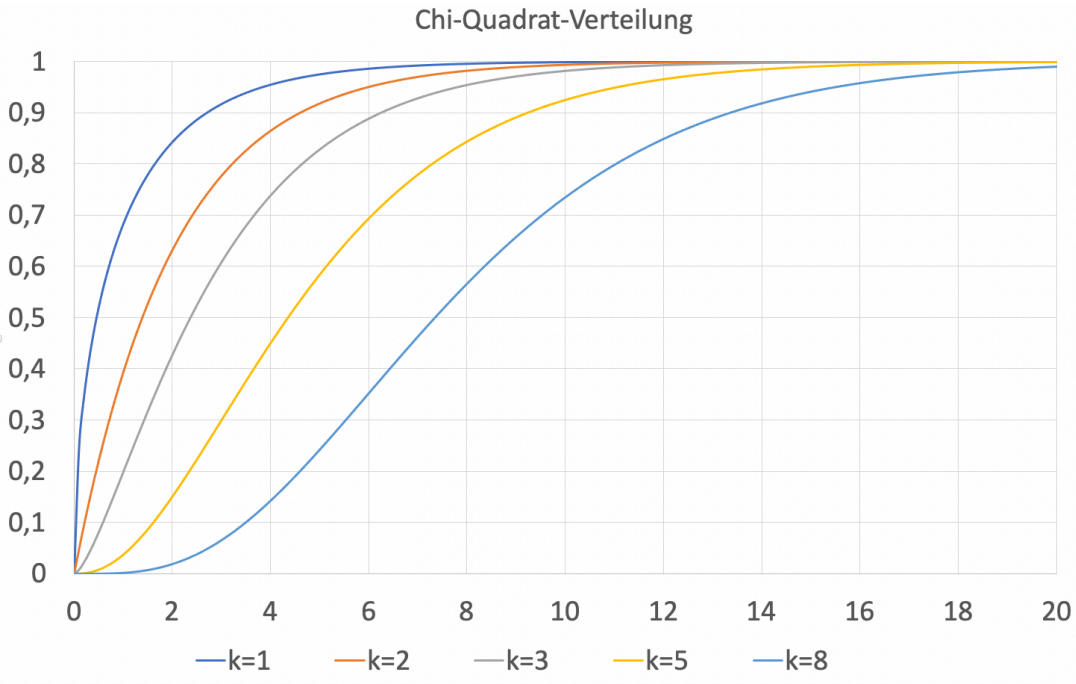
Allgemein ist die Gamma-Funktion durch folgendes Integral definiert:

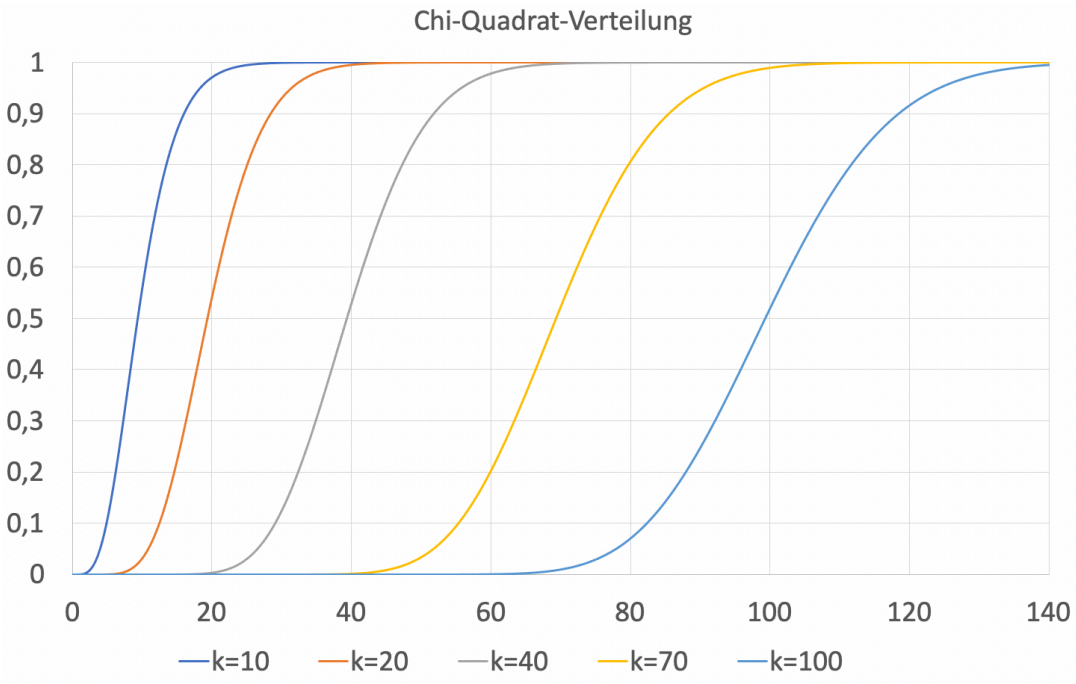
$$\Gamma(n) = \int_0^\infty x^{n-1} \cdot e^{-x} dx$$

2.2.2 **Verteilungsfunktion**

Wie immer erhält man die Verteilungsfunktion, indem man die Dichtefunktion von minus unendlich bis x integriert.

$$F(x) = \int_{-\infty}^{x} f(u)du = \frac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot \int_{-\infty}^{x} u^{\frac{k}{2}-1} \cdot e^{-\frac{u}{2}} du$$





2.2.3 **Beispiel**

Es wurden k Stichproben gezogen $(Z_1, Z_2, \cdots, Z_k)$ und wir nehmen an, dass alle Werte der Grundgesamtheit normalverteilt mit den Parametern μ und σ sind. Daraus kann die empirische Varianz berechnen:

$$s^2 = \frac{1}{k-1} \cdot \sum_{i=1}^{k} \left(Z_i - \overline{Z}\right)^2$$

wobei

$$\overline{Z} = \frac{1}{k} \cdot \sum_{i=1}^{k} Z_i$$

Hier werden also Zufallsvariablen quadriert und diese Quadrate werden aufsummiert. Das ist genau das Kochrezept für eine Chi-Quadrat-Verteilung. Um die Sache noch perfekt zu machen (standardisieren) muss man mit der Anzahl der Freiheitgrade multiplizieren, und durch die Varianz dividieren:

$$X = \frac{(k-1) \cdot s^2}{\sigma^2}$$

Diese Zufallsvariable X ist nun genau Chi-Quadrat-verteilt mit $k-1$ Freiheitsgraden.

2.3 **Studentsche Verteilung (t-Verteilung)**

Der englische Statistiker und Chemiker William Gosset, der diese Verteilung eingeführt hat, arbeitete zu diesem Zeitpunkt bei einer Brauerei (bei Guinnes). Gossets Vorgesetzte waren nicht damit einverstanden, dass er wissenschaftliche Arbeiten publizierte. Daher wählte er ein Pseudonym und veröffentlichte seine Arbeite unter einem anderen Namen. Er wählte dafür den Namen »Student«. Daher ist diese Verteilung auch unter dem Namen »Studentsche Verteilung« (englisch: »Student's distribution«) bekannt. Ein Zusammenhang mit Studenten besteht nicht.

2.3.1 **Wofür braucht man diese Verteilung?**

Sie kennen den Erwartungswert einer Normverteilung, haben aber keine Information über die Varianz dieser Verteilung. Sie ziehen einige wenige Stichproben und sollen nun abschätzen, ob ihre Stichproben zu dieser Verteilung gehören oder nicht.

Oder

Sie ziehen Stichproben von 2 Verteilungen (Sie kaufen Semmeln in 2 verschiedenen Filialen und wiegen sie ab) und wollen dann feststellen, ob die Mittelwerte der Stichproben zusammenpassen (Ist es möglich, dass diese Semmeln in derselben Bäckerei hergestellt wurden?)

Gosset hat in seinen Untersuchungen bemerkt, dass die Schätzfunktion des Stichproben-Mittelwerts normalverteilter Daten nicht normalverteilt ist, wenn die Varianz des Merkmals unbekannt ist. Er bemerkte, dass die Schätzfunktion vor allem bei kleinen Stichprobengrößen eine andere Verteilung annimmt, die der Normalverteilung zwar ähnlich sieht, aber links und rechts vom Maximum weniger schnell gegen 0 strebt. Nachdem die Fläche unter der Kurve vorgegeben ist (sie ist 1), führt das dazu, dass die neue Verteilung bei ihrem Maximalwert niedriger ist als die Normalverteilung.

Die t-Verteilung macht es insbesondere bei kleinen Stichprobenumfängen möglich die Verteilung der Differenz vom Mittelwert der Stichprobe zum wahren Mittelwert der Grundgesamtheit zu berechnen.

2.3.2 **Herleitung**

Man geht von k voneinander unabhängigen und identisch verteilten Zufallsgrößen $X_1, X_2, \cdots, X_k$ aus, die alle den gleichen Erwartungswert μ und die gleiche Varianz σ^2 haben. Dass die Parameter alle gleich sind, ist in der Praxis oftmals bestens bekannt, aber welche genauen Werte μ und σ^2 haben, weiß man nur in den seltensten Fällen. μ und σ^2 sind also unbekannt. Man kann diese Zufallsvariablen $X_1, X_2, \cdots, X_k$ als Stichproben interpretieren, für die man den Mittelwert berechnen kann:

$$\overline{X} = \frac{1}{k} \cdot \sum_{i=1}^{k} X_i$$

Dieser Mittelwert $\overline{X}$ ist selbst wieder eine Zufallsvariable, die ebenfalls normalverteilt ist. Das folgt aus dem Additionssatz der Normalverteilung (1.1.6 auf Seite 8).

Um diese Zufallsvariable mit anderen Verteilungen vergleichen zu können, möchte man sie gerne standardisieren. Man möchte also eine Verteilung haben, die den Erwartungswert 0 und die Varianz 1 hat:

$$\overline{X}^* = \frac{\overline{X} - \mu}{\sigma}$$

Das geht aber nicht, weil man μ und σ^2 (und somit σ) nicht kennt. Für μ lässt sich vielleicht aus früheren Untersuchungen ein Wert finden, oder man nimmt einfach einen passenden Wert an.

Für σ kann man die empirische Standardabweichung (auch: Standardabweichung einer Stichprobe) berechnen:

$$s = \sqrt{\frac{1}{k-1} \cdot \sum_{i=1}^{k} \left(X_i - \overline{X}\right)^2}$$

Aber s und σ sind nicht dasselbe. Und aus diesem Grund ist auch $\overline{X}^*$ nicht normalverteilt. Die Verteilung von $\overline{X}^*$ entspricht der t-Verteilung mit $k-1$ Freiheitsgraden. (Es gibt k Stichproben, daraus wird 1 Parameter berechnet, bleiben $k-1$ Freiheitsgrade).

2.3.3 **Formel**

Gosset (»Student«) hat folgendes gezeigt: Wenn X eine chi-quadrat-verteilte Zufallsvariable mit k Freiheitsgraden ist und wenn Z eine standardnormalverteilte Zufallsvariable ist, dann ist

$$T = \frac{Z}{\sqrt{\frac{X}{k}}}$$

eine t-verteilte Zufallsvariable mit k Freiheitsgraden. Die explizite Formel für die Wahrscheinlichkeitsdichtefunktion sieht so aus:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

Im Spezialfall $k = 1$ vereinfacht sich die Gleichung zu

$$f(x) = \frac{\Gamma(1)}{\sqrt{\pi} \cdot \Gamma\left(\frac{1}{2}\right)} \cdot (1 + x^2)^{-1} = \frac{1}{\sqrt{\pi} \cdot \sqrt{\pi}} \cdot \frac{1}{1 + x^2} = \frac{1}{\pi \cdot (1 + x^2)}$$

und heißt dann »Standard-Cauchy-Verteilung«.

Für $k = 2$ ergibt sich ein ähnlich einfacher Ausdruck:

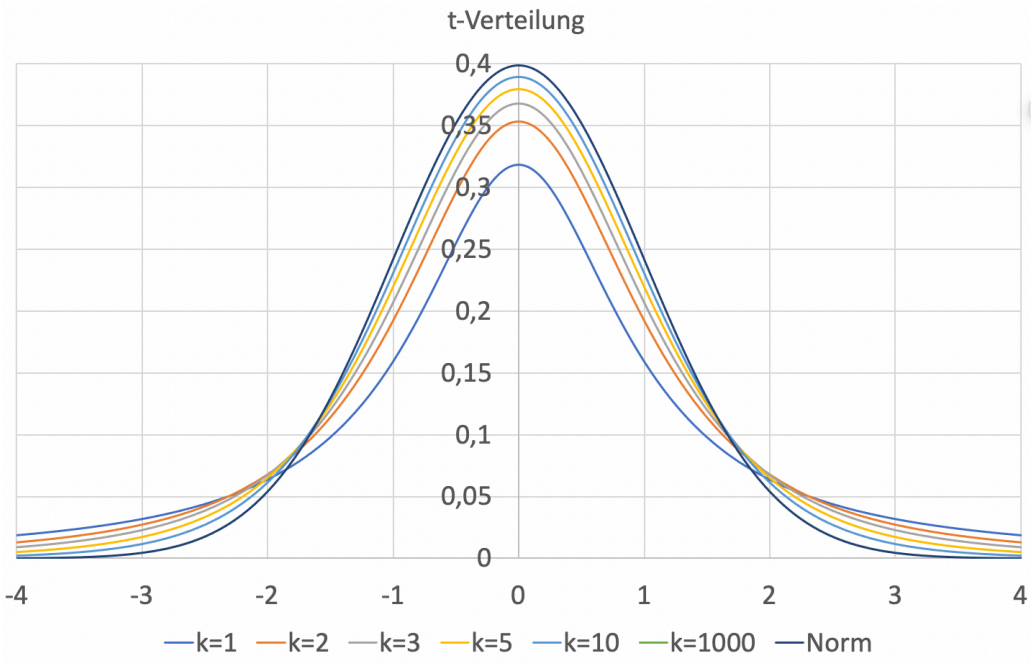
$$f(x) = \frac{1}{(2 + x^2)^{\frac{3}{2}}}$$

Wenn $k > 1$ ist der Erwartungswert der Verteilung

$$E(T) = 0$$

Wenn $k > 2$ ist die Varianz der Verteilung

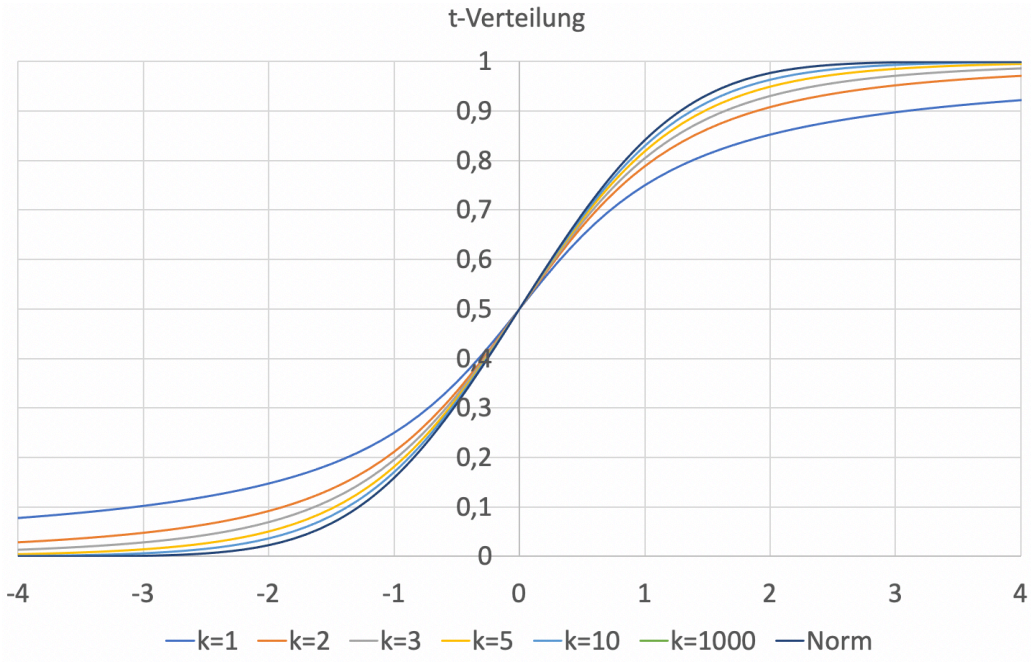
$$VAR(T) = \frac{k}{k - 2}$$



Wahrscheinlichkeitsdichtefunktion der t-Verteilung

Mit zunehmendem Freiheitsgrad wird die t-Verteilung der Standardnormalverteilung immer ähnlicher. Im Bild verdeckt die Kurve für die Normalverteilung vollständig die Kurve für eine t-Verteilung mit $k = 1000$. Die Dichtefunktion ist symmetrisch um $x = 0$.

$$f(x) = f(-x)$$



Verteilungsfunktion einer t-Verteilung

Mittelwert, Median und Modus einer t-Verteilung sind immer 0.

Je kleiner der Freiheitsgrad ist, desto weiter nach außen verteilt sich die Kurve. Die Varianz nimmt also zu, je kleiner der Freiheitsgrad wird. Das geht so weit, dass beim Freiheitsgrad $k = 2$ die Varianz sogar unendlich groß wird, und dass die Varianz für $k = 1$ in einem gewissen Sinn sogar noch größer ist.

2.4 Fisher-Verteilung (F-Verteilung)

Die Fisher-Verteilung ist nach dem britischen Biologen und Statistiker Ronald A. Fisher benannt, der sie gemeinsam mit seinem Kollegen George W. Snedecor entwickelt hat (sie heißt daher auch Snedecor-Verteilung). Die entsteht, wenn man zwei Chi-Quadrat-Verteilungen durcheinander dividiert.

Es seien X_1 und X_2 zwei voneinander unabhängige Zufallsvariablen, die beide chi-quadrat-verteilt sind, mit den Freiheitsgraden k_1 und k_2. Dann ist

$$X = \frac{\frac{X_1}{k_1}}{\frac{X_2}{k_2}}$$

ebenfalls eine Zufallsvariable. Sie folgt aber einer neuen Verteilung, nämlich der Fisher-Verteilung. Die Parameter der Fisher-Verteilung sind die beiden Freiheitsgrade k_1 und k_2.

Auch das Quadrat einer t-Verteilung ist Fisher-verteilt, der erste Freiheitsgrad k_1 ist dann 1 und k_2 ist gleich dem Freiheitsgrad der t-Verteilung.

Der Erwartungswert einer Fisher-Verteilung ist

$$E(X) = \frac{k_2}{k_2 - 2}$$

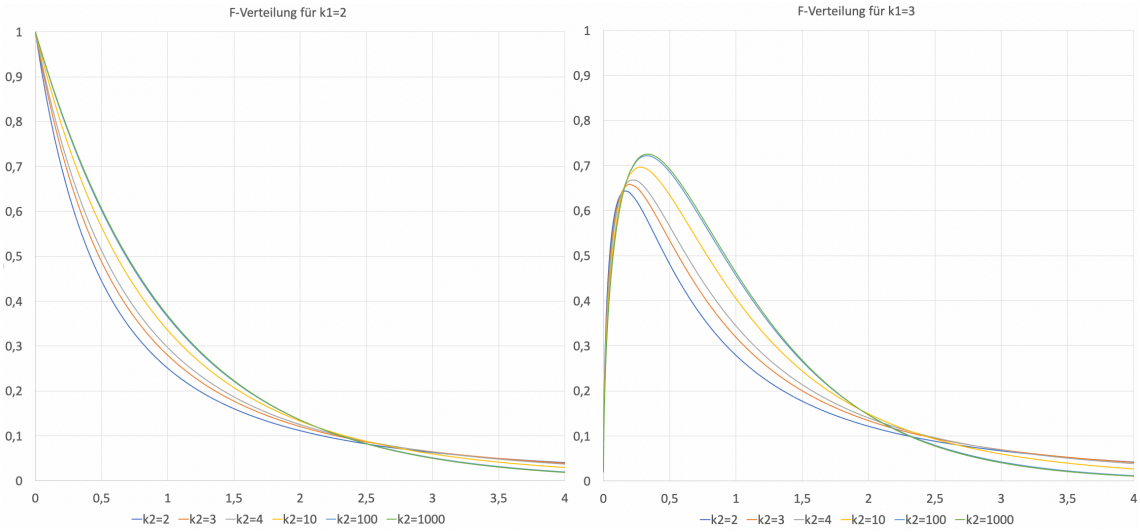
und die Varianz

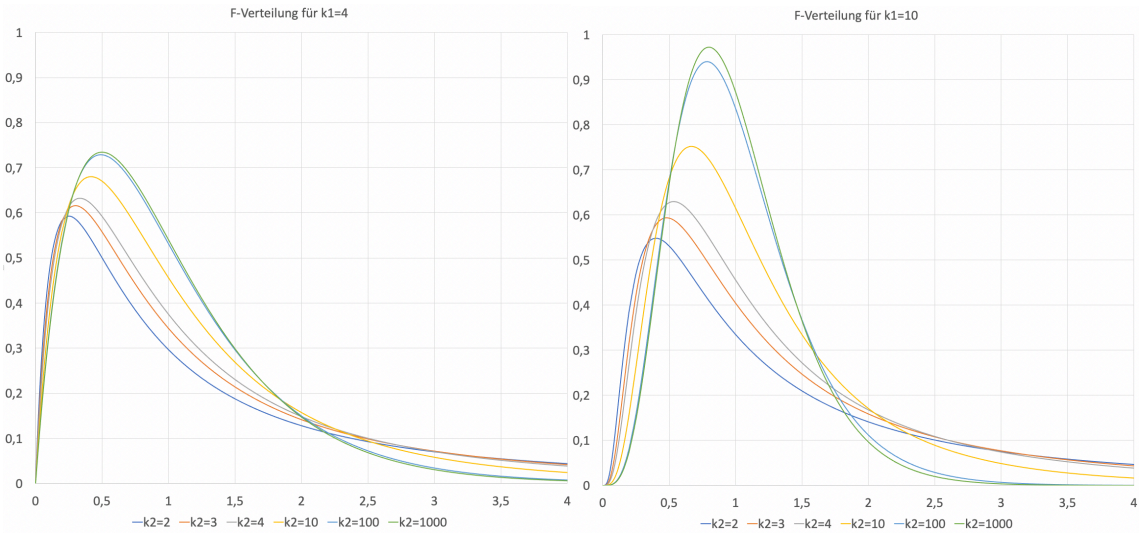
$$VAR(X) = \frac{2k_2^2 \cdot (k_1 + k_2 - 2)}{k_1 \cdot (k_2 - 4) \cdot (k_2 - 2)^2}$$

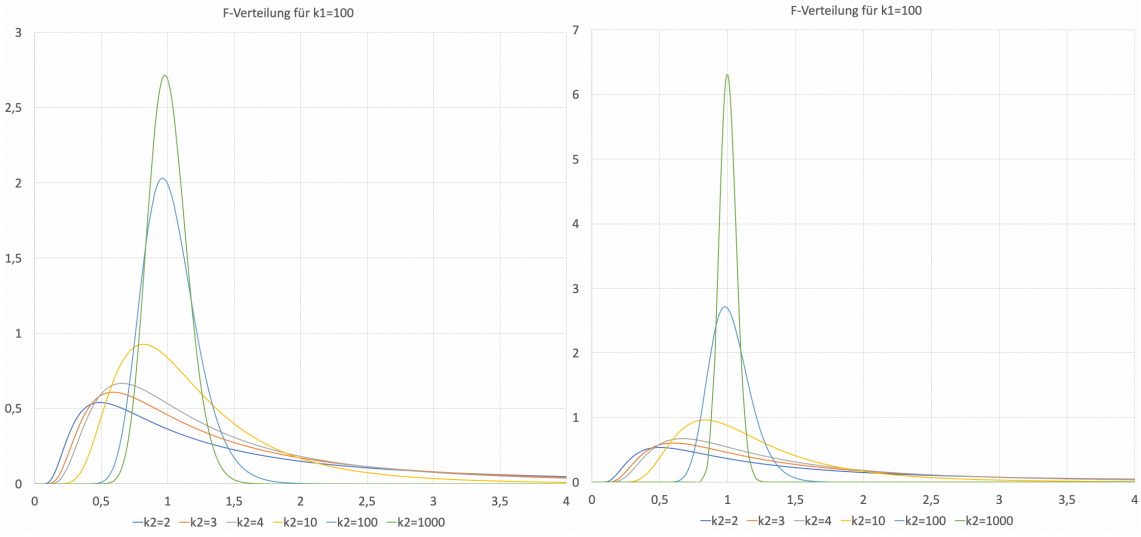
Der Erwartungswert ist nur für $k_2 > 2$ definiert, die Varianz nur für $k_2 > 4$.

Die Dichtefunktion sieht folgendermaßen aus:

$$f(x) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right) \cdot \Gamma\left(\frac{k_2}{2}\right)} \cdot \left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}} \cdot \frac{x^{\frac{k_1}{2}-1}}{\left(1 + \frac{k_1}{k_2} \cdot x\right)^{\frac{k_1+k_2}{2}}}$$


F-Verteilung für k1=2
F-Verteilung für k1=3
k2=2 k2=3 k2=4 k2=10 k2=100 k2=1000


F-Verteilung für k1=4
F-Verteilung für k1=10
k2=2 k2=3 k2=4 k2=10 k2=100 k2=1000


F-Verteilung für k1=100
F-Verteilung für k1=100
k2=2 k2=3 k2=4 k2=10 k2=100 k2=1000

2.4.1 Interpretation und Verwendung

Gegeben sind die normalverteilten Zufallsgrößen $X_1, X_2, \cdots, X_m$ mit μ_x und σ_x^2 und $Y_1, Y_2, \cdots, Y_n$ mit μ_y und σ_y^2. Diese kann man zum Beispiel als Werte einer Stichprobe ansehen. Damit kann man die Stichprobenvarianzen berechnen:

$$s_x^2 = \frac{1}{m-1} \cdot \sum_{i=1}^{m} \left(X_i - \overline{X}\right)^2$$

$$s_y^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

Dabei ist

$$\overline{X} = \frac{1}{m} \cdot \sum_{i=1}^{m} X_i \qquad \overline{Y} = \frac{1}{n} \cdot \sum_{i=1}^{n} Y_i$$

Dann sind die Zufallsvariablen

$$\frac{s_x^2}{\sigma_x^2} \quad und \quad \frac{s_y^2}{\sigma_y^2}$$

jeweils Chi-Quadrat verteilt mit den Freiheitsgraden $m-1$ bzw. $n-1$.

Der Quotient Z dieser beiden Zufallsvariablen ist nun Fisher-verteilt mit den Freiheitsgraden $m-1$ und $n-1$.

$$Z = \frac{\frac{s_x^2}{\sigma_x^2}}{\frac{s_y^2}{\sigma_y^2}}$$

Beim F-Test nimmt man an, dass die beiden Stichproben dieselbe Varianz haben. Also $\sigma_x^2 = \sigma_y^2$. Dann kürzen sich die beiden Varianzen weg und es bleibt

$$Z = \frac{s_x^2}{s_y^2}$$

Mit Hilfe der Stichprobenvarianzen bzw. der F-Verteilung kann man dann also berechnen, wie wahrscheinlich der Wert von Z ist und so bestimmen, ob die Annahme von gleichen Varianzen zulässig ist.