Anhang A

Appendix: Stetige Verteilungen

In diesem Appendix werden – der üblichen statistischen Praxis folgend – Zufallsvariablen mit Großbuchstaben und Realisationen mit Kleinbuchstaben bezeichnet.

A.1 Gleichverteilung

Die Gleichverteilung spielt für das Folgende zwar keine große Rolle, aber aufgrund ihrer Einfachheit ist sie aus didaktischen Gründen gut geeignet um die Grundkonzepte zu wiederholen.

Eine *diskrete Zufallsvariable* ist gleichverteilt, wenn jede ihrer möglichen m Ausprägungen mit gleicher Wahrscheinlichkeit eintritt:

$$f(x) = \Pr[X = x_i] = \frac{1}{m} \qquad (i = 1, \ldots, m)$$

Die Wahrscheinlichkeitsfunktion der Augenzahl eines Würfels ist z.B. $\Pr[X = x_i] = 1/6$ für $(i = 1, \ldots, 6)$.

Für eine *stetige Zufallsvariable* liegen die möglichen Realisationen zwischen den beiden endlichen Grenzen a und b. Da für jede Dichtefunktion gilt

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

folgt für die **Dichtefunktion einer gleichverteilten stetigen Zufallsvariable** X

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } a \leq b \\ 0 & \text{sonst} \end{cases}$$

Die Wahrscheinlichkeit, dass eine Realisation von X einen Wert zwischen x_1 und x_2 annimmt, ist

$$\int_{x_1}^{x_2} f(x)\,dx = \Pr[x_1 \leq x \leq x_2] = \frac{(x_2 - x_1)}{b - a}$$

Durch Integration der Dichtefunktion erhält man die **Verteilungsfunktion** $F(x)$

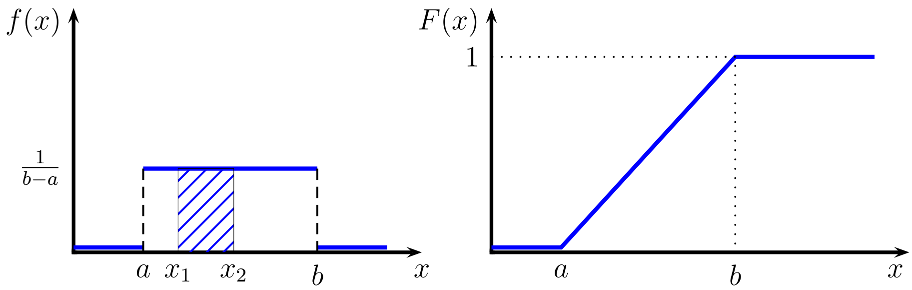


Abbildung A.1: Dichte- und Verteilungsfunktion einer stetigen gleichverteilten Zufallsvariable.

$$F(x) = \begin{cases} 0 & \text{für } x < a \\ \frac{x-a}{b-a} & \text{für } a \le x \le b \\ 1 & \text{für } x > b \end{cases}$$

Abbildung A.1 zeigt die Dichte- und Verteilungsfunktion einer stetigen gleichverteilten Zufallsvariable.

Der Erwartungswert ist

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{+\infty} x f(x) dx = \int_{-\infty}^{a} x \cdot 0 dx + \int_{a}^{b} x\, f(x) dx + \int_{b}^{+\infty} x \cdot 0 dx \\ &= \int_a^b x \frac{1}{b-a}\, dx \\ &= \left. \frac{x^2}{2(b-a)} \right|_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

weil $(b^2-a^2) = (a+b)(b-a)$.

Die Varianz $\mathrm{var}(X) = \mathrm{E}[X-\mathrm{E}(X)]^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$ kann ebenfalls berechnet werden:

$$\mathrm{E}(X^2) = \int_a^b x^2 \left(\frac{1}{b-a}\right) dx = \left. \frac{x^3}{3(b-a)} \right]_a^b = \frac{b^3-a^3}{3(b-a)}$$

Unter Verwendung von $b^3 - a^3 = (b-a)(a^2+ab+b^2)$ erhalten wir schließlich

$$\mathrm{var}(X) = \frac{(a^2+ab+b^2)}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(a^2+b^2-2ab)}{12} = \frac{(a-b)^2}{12}$$

Die Gleichverteilung ist also durch die zwei Parameter a und b vollständig bestimmt.

A.2 Normalverteilung

Die Normalverteilung spielt in der Ökonometrie eine besondere Rolle, und dies aus mehreren Gründen. Erstens sind viele Variablen in ihren Grundgesamtheiten

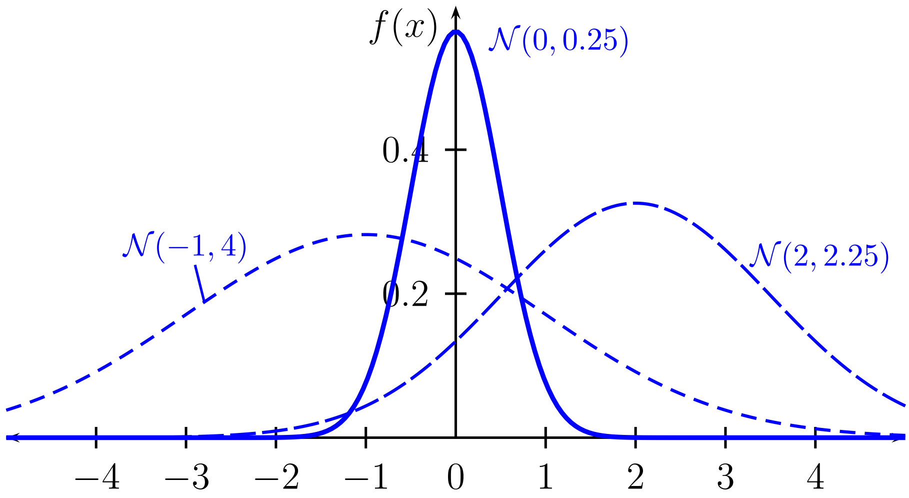


Abbildung A.2: Dichtefunktionen von normalverteilten Zufallsvariablen mit unterschiedlichem μ und σ^2.

'annähernd' normalverteilt (z.B. Körpergröße, Intelligenz, Meßfehler, etc.), zweitens können viele andere Verteilungen aus ihr hergeleitet werden (z.B. die χ^2-, t- oder F-Verteilung), und drittens nähern sich bei sehr großen Beobachtungszahlen andere Verteilungen manchmal der Normalverteilung an (z.B. die Binomialverteilung). Außerdem werden wir später sehen, sind für große Beobachtungszahlen die *Stichprobenkennwertverteilungen* unabhängig von der Verteilung der Grundgesamtheit annähernd normalverteilt (*zentraler Grenzwertsatz*).

Die Normalverteilung ist eine stetige Verteilung mit der Dichte

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1/2)[(x_i - \mu)/\sigma]^2}$$

mit μ = Erwartungswert, σ = Standardabweichung von X, $\pi = 3.14159\ldots, e = 2.71828\ldots$.

Wenn eine Zufallsvariable X normalverteilt ist mit Mittelwert μ und Varianz σ^2, d.h.

$$\begin{aligned} \mathrm{E}(X) &= \mu \\ \mathrm{var}(X) &= \sigma^2 \end{aligned}$$

wird dies üblicherweise geschrieben als

$$X \sim \mathrm{N}(\mu, \sigma^2)$$

Abbildung A.2 zeigt drei Dichtefunktionen von normalverteilten Zufallsvariablen mit unterschiedlichem μ und σ^2.

Eigenschaften der Normalverteilung:

1. Die Verteilung ist stetig und symmetrisch um den Erwartungswert μ.

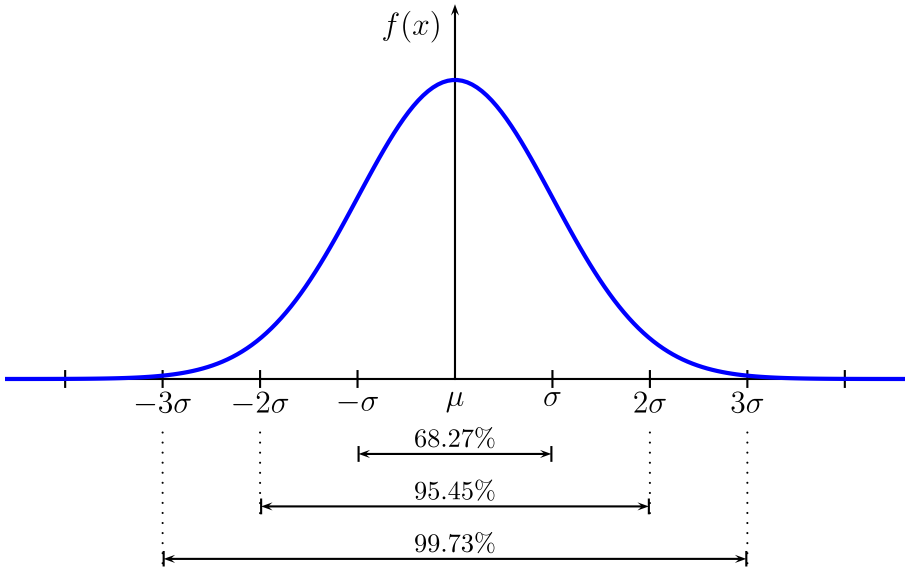


Abbildung A.3: Dichtefunktion einer normalverteilten Zufallsvariable und % der Fläche darunter.

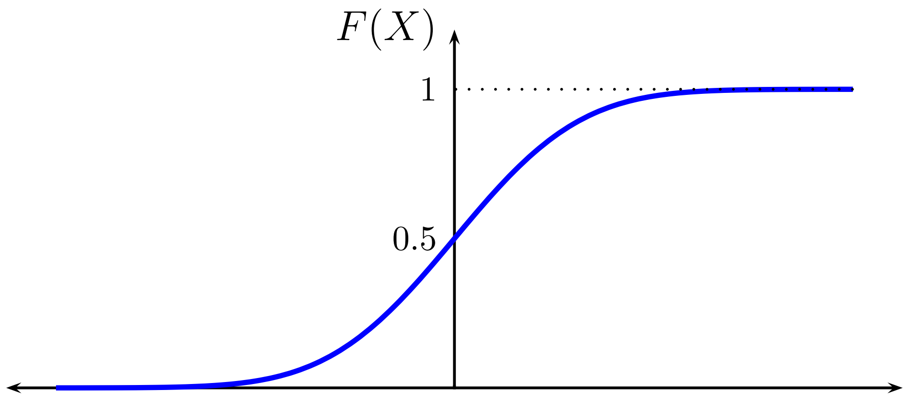


Abbildung A.4: Verteilungsfunktion einer normalverteilten Zufallsvariable.

2. Die Verteilung ist unbeschränkt und erstreckt sich von $-\infty$ bis $+\infty$.
3. Das Maximum der Verteilung liegt beim Mittelwert μ und der Wendepunkt bei $x = \mu \pm \sigma$. Außerdem gilt:

$$\begin{aligned} \Pr(\mu - 1.96\sigma < x_i < \mu + 1.96\sigma) &\approx 0.95 \\ \Pr(\mu - 2.57\sigma < x_i < \mu + 2.57\sigma) &\approx 0.99 \end{aligned}$$

4. Die Normalverteilung ist durch die zwei Parameter μ und σ^2 vollständig spezifiziert, wir schreiben $X \sim \mathrm{N}(\mu, \sigma^2)$.

Abbildung A.3 zeigt die Dichtefunktion einer Normalverteilung.

Die Verteilungsfunktion einer normalverteilten Zufallsvariable ist in Abbildung A.4 dargestellt.

Theorem 1 *Wenn $X, Y, \dots, Z$ normal und unabhängig verteilte Zufallsvariablen und $a, b, \dots, c$ beliebige Konstanten sind, dann ist die Linearkombination $aX + bY + \cdots + cZ$ auch normalverteilt.*

Beispiel: Wenn $X_i \sim \mathrm{N}(\mu, \sigma^2)$ mit $i = 1,\ 2,\ 3$ drei *unabhängig* verteilte Zufallsvariablen sind, und dann sind Erwartungswert und Varianz der Zufallsvariable $W = 2X_1 + X_2 - 4X_3$

$$\begin{aligned}
\mathrm{E}(W) &= 2\,\mathrm{E}(X_1) + \mathrm{E}(X_2) - 4\,\mathrm{E}(X_3) = 2\mu + \mu - 4\mu = -1\mu \\
\mathrm{var}(W) &= 2^2\,\mathrm{var}(X_1) + 1^2\,\mathrm{var}(X_2) + 4^2\,\mathrm{var}(X_3) = 21\sigma^2
\end{aligned}$$

d.h. $W \sim \mathrm{N}(-\mu, 21\sigma^2)$.

Oder, wenn $X \sim \mathrm{N}(\mu, \sigma^2)$ dann ist $Y = a + bX \sim \mathrm{N}(a + b\mu, b^2\sigma^2)$, z.B. $X \sim \mathrm{N}(3, 4)$ und $Y = 2X - 5$ dann ist $Y \sim \mathrm{N}(1, 16)$.

Dieses Theorem gestattet eine Standardisierung der Normalverteilungen auf einen Mittelwert von Null und eine Standardabweichung von Eins (z–Transformation).

$$z_i = \frac{(x_i - \mu)}{\sigma} \quad \text{bzw.} \quad Z = \frac{(X - \mu)}{\sigma}$$

mit $Z \sim \mathrm{N}(0, 1)$, man sagt Z ist *standardnormalverteilt.* Die Tabelle der Standardnormalverteilung findet sich in jedem Statistiklehrbuch.

Wenn X und Y gemeinsam normalverteilt sind, dann sind sie unabhängig wenn und nur wenn $\mathrm{cov}(X, Y) = 0$. Achtung, dies gilt nur für die Normalverteilung! Sind zwei Zufallsvariablen nicht gemeinsam normalverteilt ist $\mathrm{cov}(X, Y) = 0$ in der Regel nicht hinreichend um Unabhängigkeit zu garantieren.

Bivariate Normalverteilung: $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, $|\rho| \leq 1$ oder in Matrixschreibweise

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right]$$

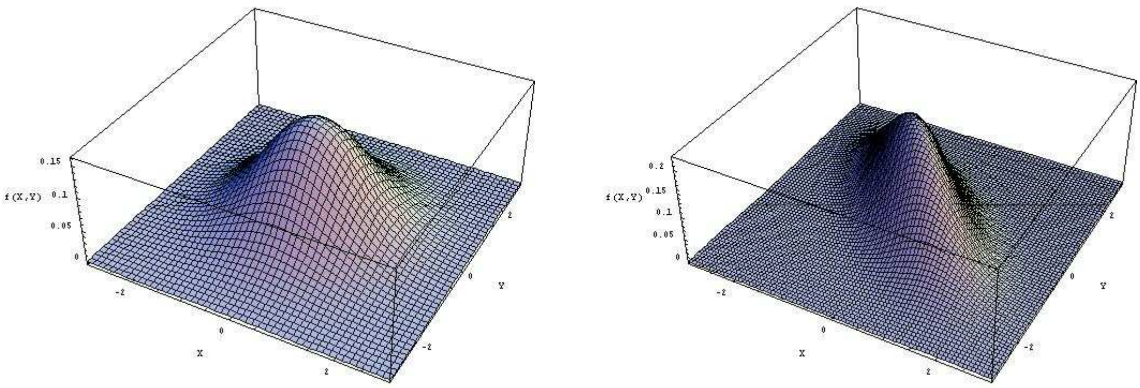
mit der Dichtefunktion

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \exp\left\{-\frac{1}{2(1 - \rho^2)}\left[z_X^2 - z_X z_Y + z_Y^2\right]\right\}$$

mit

$$z_X = \frac{x - \mu_X}{\sigma_X}, \qquad z_Y = \frac{y - \mu_Y}{\sigma_Y}$$

siehe Abbildung A.5.



$\mu_X = \mu_Y = 0,\ \sigma_X^2 = \sigma_Y^2 = 1;\ \rho = 0$ $\qquad$ $\mu_X = \mu_Y = 0,\ \sigma_X^2 = \sigma_Y^2 = 1;\ \rho = -0.7$

Abbildung A.5: Bivariate Normalverteilungen

Übungsbeispiele:

1. Zwei Zufallsvariablen X und Y seien bivariat normalverteilt mit $X \sim \mathrm{N}(2,3)$, $Y \sim \mathrm{N}(1,2)$, und $\mathrm{Cov}(X,Y) = 0.5$. bzw. in Matrixschreibweise

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right]$$

Wie ist dann eine Zufallsvariable $V = 2X + 3Y$ verteilt?

V ist univariat normalverteilt mit

$$\begin{aligned} \mathrm{E}(V) &= \mathrm{E}(2X + 3Y) = 2\,\mathrm{E}(X) + 3\,\mathrm{E}(Y) \\ &= 2 \times 2 + 3 \times 1 \\ &= 7 \end{aligned}$$

und

$$\begin{aligned} \mathrm{var}(V) &= \mathrm{var}(2X + 3Y) \\ &= \mathrm{var}(2X) + \mathrm{var}(3Y) + 2\,\mathrm{cov}(2X, 3Y) \\ &= 2^2 \underbrace{\mathrm{var}(X)}_{3} + 3^2 \underbrace{\mathrm{var}(Y)}_{2} + 2 \times 2 \times 3 \underbrace{\mathrm{cov}(X,Y)}_{0.5} \\ &= 12 + 18 + 6 = 36 \end{aligned}$$

Also ist $V \sim \mathrm{N}(7, 36)$.

Dies kann einfach simuliert werden, z.B. mit R . Der Befehl `mvrnorm` aus dem Packet MASS erzeugt Zufallszahlen (Realisationen) aus einer multivariaten Normalverteilung. Der folgende Code demonstriert das obige Resultat anhand von 10000 Replikationen.

```
library(MASS)
mu = c(2,1)
Sigma <- matrix(c(3,0.5,0.5,2), nrow=2, ncol=2)
set.seed(1234567)
XY <- mvrnorm(n=10000, mu, Sigma)
V <- 2*XY[,1] + 3*XY[,2]
mean(V)  #  -> 7.05138
var(V)   #  -> 36.11783
```

2. Erzeugen Sie in einem Computerprogramm zwei unabhängige Zufallsvariablen $X \sim \mathrm{N}(2,3)$, $Y \sim \mathrm{N}(1,2)$, und $\mathrm{cov}(X,Y) = 0$, d.h.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}\right]$$

mit je 100,000 Beobachtungen, und erzeugen Sie eine Grafik mit den Histogrammen.

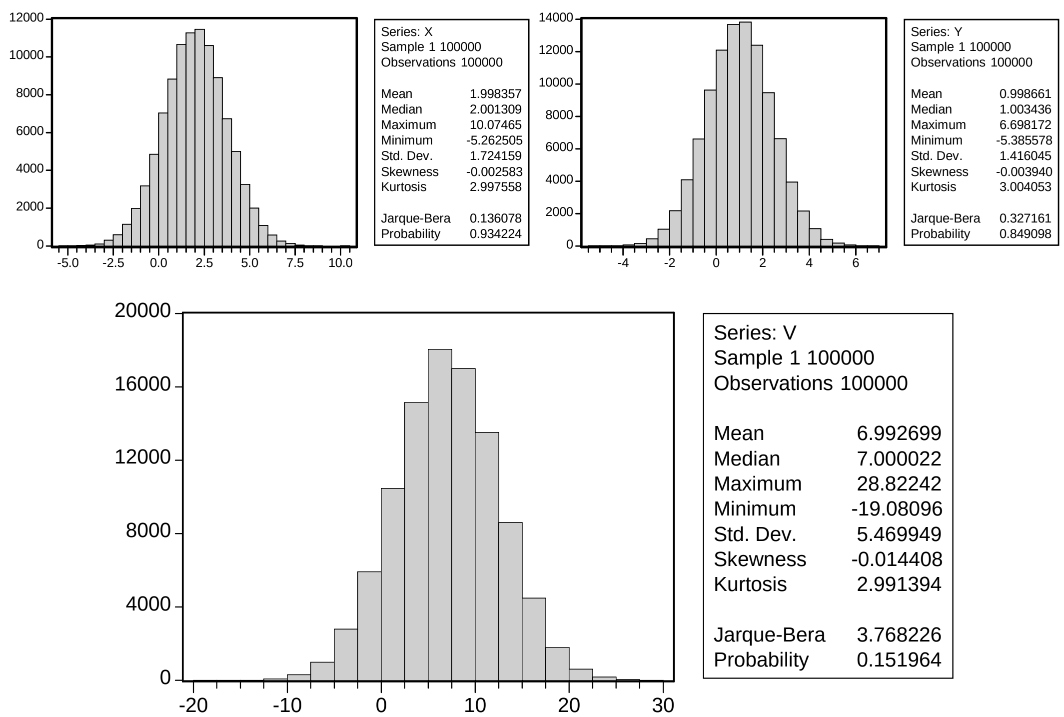
Berechnen Sie $V = 2X + 3Y$ und zeigen Sie das Histogramm. Welchen Mittelwert und welche Varianz von V erwarten Sie? Stimmen die erwarteten Werte näherungsweise mit den empirischen Werten überein?

Wenn $Z \sim \mathrm{N}(0,1)$ ist $X = 2 + \sqrt{3}\,Z \sim \mathrm{N}(2,3)$, da $\mathrm{var}(X) = \mathrm{var}(2 + \sqrt{3}\,Z) = \sqrt{3}^2\,\mathrm{var}(Z) = 3$.

In EViews wird eine standardnormalverteilte Zufallszahl z.B. mit der Funktion `@rnorm` erzeugt.

Der folgende EViews Code legt einen Workfile mit dem Namen `TEST` für 100000 Querschnitts-Beobachtungen (Option `u` für '*undated*') an und erzeugt die drei Zufallsvariablen (Befehle `series`). Die `freeze`-Befehle erzeugen die Grafiken mit den drei einzelnen Histogrammen, die mit dem `graph` und `merge` Befehl zu einer Grafik zusammengefaßt und mit dem Befehl `show` angezeigt werden.

```
wfcreate(wf=TEST) u 100000
series X = 2 + @sqrt(3)*@rnorm
series Y = 1 + @sqrt(2)*@rnorm
series V = 2*X + 3*Y
freeze(HX)  X.hist
freeze(HY)  Y.hist
freeze(HV)  V.hist
graph GRAFIK.merge HX HY HV
show GRAFIK
```



$$
\begin{aligned}
\mathrm{var}(V) &= \mathrm{var}(2X+3Y) \\
&= \mathrm{var}(2X) + \mathrm{var}(3Y) + 2\,\mathrm{cov}(2X,3Y) \\
&= 2^2 \underbrace{\mathrm{var}(X)}_{3} + 3^2 \underbrace{\mathrm{var}(Y)}_{2} + 2 \times 2 \times 3 \underbrace{\mathrm{cov}(X,Y)}_{0} \\
&= 12 + 18 + 0 = 30 \\
\mathrm{Std.Dev.}(V) = \sqrt{\mathrm{var}(V)} &= \sqrt{30} = 5.477 \approx 5.469949
\end{aligned}
$$

A.3 Die Chi–Quadrat Verteilung:

Die Chi–Quadrat oder χ^2 Verteilung geht auf den Astronomen F.R. Helmert (1875) zurück.

Theorem 2 *Sind $Z_1, Z_2, \ldots, Z_\nu$ unabhängig standardnormalverteilte Zufallsvariablen (d.h. normalverteilte Zufallsvariable mit Mittelwert 0 und Varianz 1), so folgt die Quadratsumme $\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$ einer χ^2 Verteilung mit ν Freiheitsgraden.*

$$Z_i \sim \mathrm{N}(0,1) \quad \Rightarrow \quad \sum_{i=1}^{\nu} Z_i^2 \sim \chi_\nu^2$$

Die χ^2 Verteilung ist nicht symmetrisch und abhängig von der Anzahl der Freiheitsgrade ν. Da sie eine Quadratsumme ist kann eine χ^2-verteilte Zufallsvariable natürlich nie negativ sein. Erwartungswert und Varianz sind

$$
\begin{aligned}
\mathrm{E}(\chi^2) &= \nu \\
\mathrm{var}(\chi^2) &= 2\nu
\end{aligned}
$$

Wenn X_1 und X_2 zwei unabhängig χ^2-verteilte Zufallsvariablen mit ν_1 bzw. ν_2 Freiheitsgraden sind, so ist die Summe $X_1 + X_2$ auch χ^2-verteilt mit $\nu_1 + \nu_2$ Freiheitsgraden.

Diese Verteilung wird v.a. für Tests benötigt, die Varianzen von Zufallsvariablen betreffen. Abbildung A.6 zeigt Dichtefunktionen von χ^2-verteilten Zufallsvariablen mit unterschiedlichen Freiheitsgraden.

Um dies zu verdeutlichen haben wir in EViews das Quadrat von 100 standardnormalverteilten Variablen mit je 1000 Beobachtungen erzeugt. Abbildung A.7 zeigt ein Histogramm von Z_1^2, von $\sum_{i=1}^{5} Z_i^2$, $\sum_{i=1}^{10} Z_i^2$ und $\sum_{i=1}^{100} Z_i^2$.

Auch wenn Sie das Programm, das Abbildung A.7 erzeugt, im Moment noch nicht verstehen können, so sei es hier doch für eine spätere Referenz wiedergegeben.

```
' Erzeugung von Chi2 - verteilten Variablen.
wfcreate temp u 1000

for !i = 1  to 100
  series x{!i} = @rnorm
```

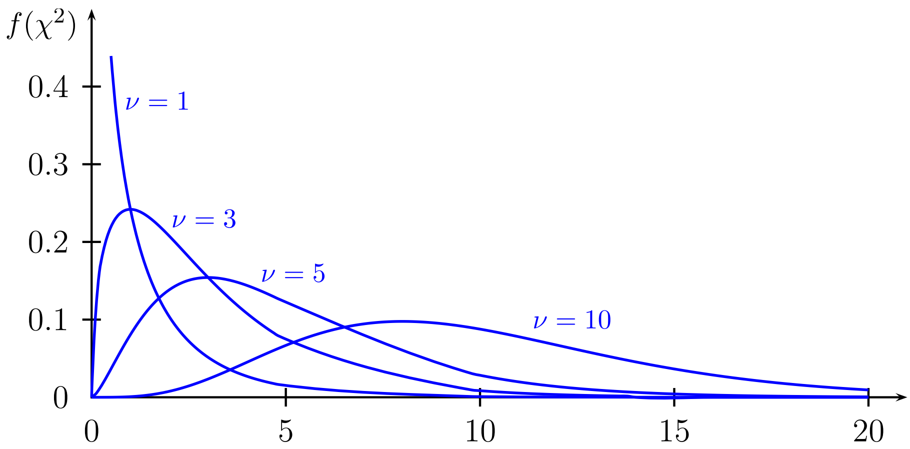


Abbildung A.6: Dichtefunktionen von χ^2 verteilten Zufallsvariablen mit $\nu = 1, 3, 5$ und 10 Freiheitsgraden.

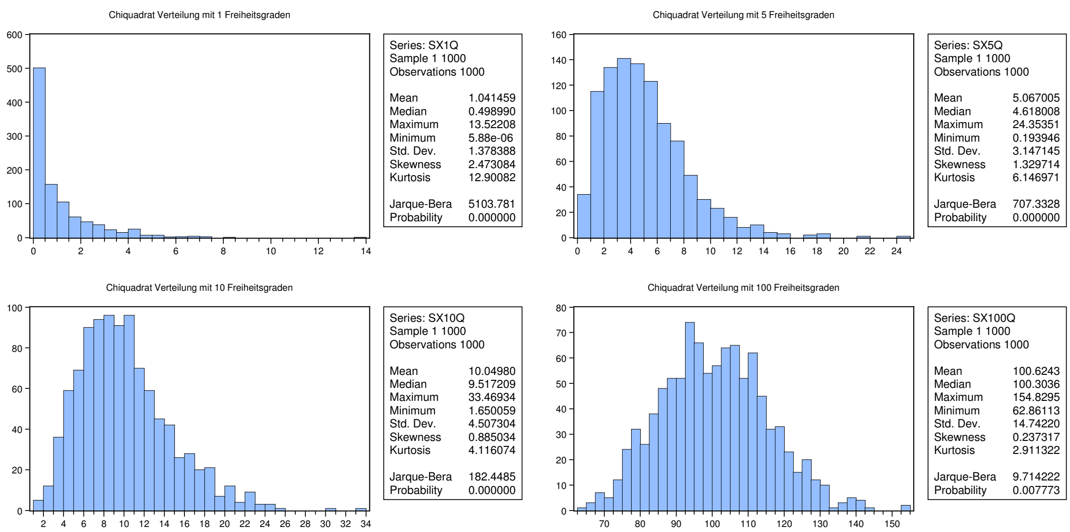


Abbildung A.7: Histogramme der Summe von quadrierten standardnormalverteilten Zufallsvariablen.

```
  series x{!i}q = x{!i}^2
  if !i = 1 then
    series sx1q = x1q
  else
    !j = !i-1
    series sx{!i}q = x{!i}q + sx{!j}q
  endif
  if !i = 1 or !i = 5 or !i = 10 or !i = 100 then
    freeze(graph{!i}) sx{!i}q.hist
    graph{!i}.addtext(t)  Chiquadrat Verteilung mit !i Freiheitsgraden
  endif
next

graph gr.merge graph1 graph5 graph10 graph100
show gr.align(2,3,1)
```

A.4 Die t-Verteilung:

Die t- bzw. Studentverteilung verdankt ihren Namen W.S. Gosset, der deren Ableitung 1908 unter dem Pseudonym 'Student' veröffentlichte.

Theorem 3 *Sei Z eine standardnormalverteilte Zufallsvariable [$Z \sim \mathrm{N}(0,1)$] und V eine χ^2-verteilte Zufallsvariable mit ν Freiheitsgraden [$V \sim \chi^2(\nu)$], wobei Z und V unabhängig voneinander verteilt sind, dann ist die Zufallsvariable T*

$$t = \frac{Z}{\sqrt{(V/\nu)}} \sim \mathrm{t}_\nu$$

t-*verteilt mit ν Freiheitsgraden.*

Wie aus Abbildung A.8 ersichtlich ist die t Verteilung symmetrisch und 'flacher' als die Standardnormalverteilung. Sie nähert sich mit steigender Zahl von Freiheitsgraden der Standardnormalverteilung an und ist für $\nu \geq 120$ nahezu mit dieser identisch.

Weiters ist

$$\begin{aligned} \mathrm{E}(t) &= 0 \qquad \text{für } \nu > 1 \\ \mathrm{var}(t) &= \frac{\nu}{\nu - 2} \quad \text{für } \nu > 2 \end{aligned}$$

Wir werden die t-Verteilung v.a. benötigen, wenn die Varianz der Grundgesamtheit nicht bekannt ist, sondern aus der Stichprobe geschätzt werden muss.

A.5 Die F-Verteilung:

Die F Verteilung werden wir später häufig benötigen, z.B. für Tests, die mehrere Regressionskoeffizienten betreffen, oder um die Gleichheit zweier Varianzen zu testen. Sie ist nach R.A. Fisher benannt.

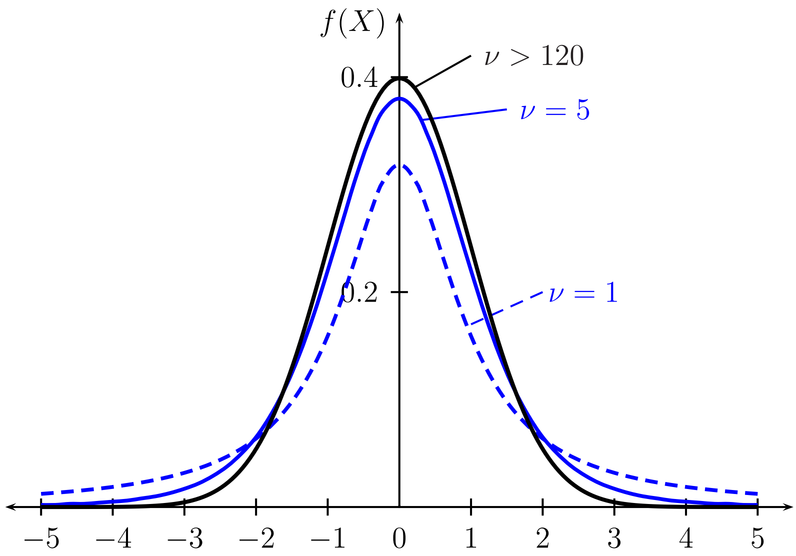


Abbildung A.8: Dichtefunktionen für t-verteilte Zufallsvariablen mit $\nu = 1,\ 5$ und ∞ Freiheitsgraden.

Theorem 4 *Wenn V_1 und V_2 zwei unabhängig χ^2-verteilte Zufallsvariablen mit ν_1 bzw. ν_2 Freiheitsgraden sind, dann ist*

$$F = \frac{V_1/\nu_1}{V_2/\nu_2} \sim \mathrm{F}_{\nu_1,\nu_2}$$

F*-verteilt mit ν_1 Zähler- und ν_2 Nennerfreiheitsgraden.*

Wie aus Abbildung A.9 ersichtlich ist die F Verteilung schief. Für große ν_1 und ν_2 nähert sie sich einer Normalverteilung an.

Man kann zeigen, dass

$$\begin{aligned}
\mathrm{E}(F_{\nu_1,\nu_2}) &= \frac{\nu_2}{\nu_2 - 2} && \text{für } \nu_2 > 2 \\
\mathrm{var}(F_{\nu_1,\nu_2}) &= \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} && \text{für } \nu_2 > 4
\end{aligned}$$

Das Quadrat einer t-verteilten Zufallsvariable mit ν Freiheitsgraden folgt einer F Verteilung mit 1 Zähler- und ν Nennerfreiheitsgraden, d.h.

$$\mathrm{t}_\nu^2 \sim \mathrm{F}_{1,\nu}$$

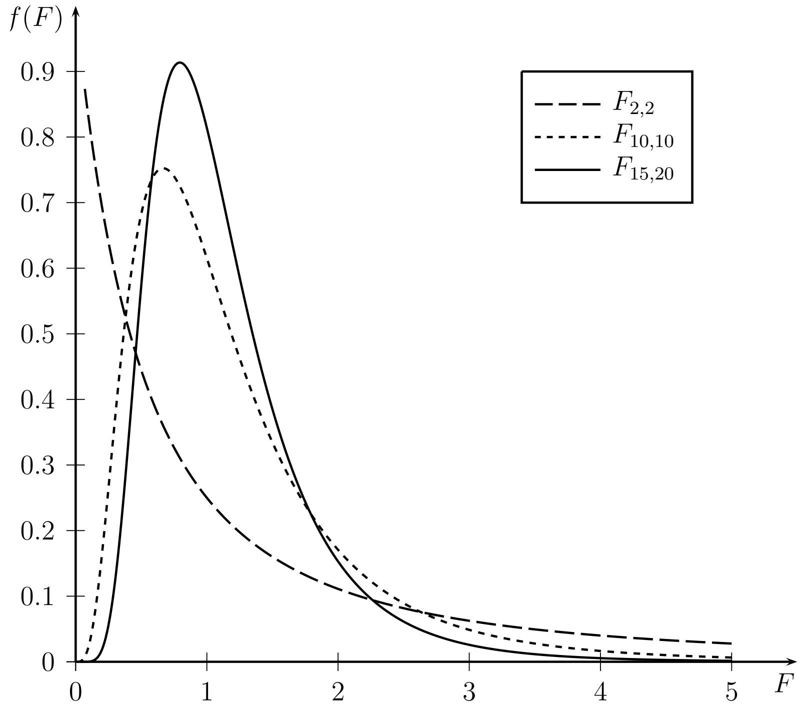


Abbildung A.9: Dichtefunktionen für F-verteilte Zufallsvariablen mit unterschiedlichen Freiheitsgraden.