III. Wahrscheinlichkeiten, Zufallsvariablen, Verteilungen

3.1 Lernziele zu Wahrscheinlichkeiten, Zufallsvariablen, Verteilungen

- Grundbegriffe der elementaren Wahrscheinlichkeitsrechnung
- Axiome von Kolmogoroff, Wahrscheinlichkeit von Ereignissen
- Venn-Diagramme
- Bedingte Wahrscheinlichkeiten, Sensitivität und Spezifität
- Unabhängigkeit von Ereignissen
- Gesetz der großen Zahl
- diskrete und stetige Zufallsvariable
- Wahrscheinlichkeitsfunktion und Dichtefunktion
- Verteilungsfunktion
- Parameter einer Verteilungsfunktion
- Erwartungswert, Median
- Varianz
- Binomialverteilung, Poissonverteilung
- Verteilungsfunktion und Dichtefunktion der Normalverteilung
- Standardnormalverteilung
- Standardisierung
- Quantile der Normalverteilung
- zentraler Grenzwertsatz
- Konfidenzintervalle
- t-Verteilung

3.2 Elementare Wahrscheinlichkeitsrechnung

Grundlage der modernen Wahrscheinlichkeitsrechnung sind die Axiome von Kolmogoroff. Kolmogoroff geht von einem Zufallsexperiment aus, das im Prinzip unter gleichen Bedingungen beliebig oft wiederholt werden kann.

Trotz der gleichen Bedingungen ist das Ergebnis des Experiments von Wiederholung zu Wiederholung **nicht** vorhersehbar. Es gibt vielmehr eine ganze **Menge möglicher Ergebnisse**, die hier im Folgenden mit ***S*** bezeichnet wird. Die **einzelnen** Ergebnisse werden mit e_1, e_2, ... bezeichnet, d. h.

$$S = \{e_1, e_2, \ldots\}.$$

Beispiel 3.1
Typische Zufallsexperimente sind das Würfeln,

das Werfen von Münzen,

die Ziehung von nummerierten Kugeln aus einer Urne wie beim Lotto

oder das Austeilen von Karten beim Skat.

In all diesen Beispielen wird das für ein Zufallsexperiment Typische deutlich: es ist im Prinzip unter gleichen Bedingungen beliebig oft wiederholbar, das Ergebnis jeder einzelnen Wiederholung ist nicht vorhersehbar. Die Menge der möglichen Ergebnisse ist z. B. beim Münzwurf S={Wappen,Zahl}

beim Würfel S={1,2,3,4,5,6}.

3.2.1 Ereignisse, Wahrscheinlichkeit von Ereignissen

Teilmengen von S, der Menge der möglichen **Ergebnisse**, bezeichnet man als **Ereignisse**. Durch diese Definition ist es möglich geworden, mit Ereignissen wie mit Mengen zu rechnen. Dabei werden die aus der Mengenlehre vertrauten Symbole benutzt.

Ereignisse, die nur ein mögliches Ergebnis enthalten, nennt man Elementarereignisse. Die **Grundmenge** S heißt **sicheres** Ereignis, die Menge, die **kein mögliches Ergebnis** enthält, heißt **unmögliches** Ereignis und wird wie in der Mengenlehre üblich mit Ø bezeichnet.

Seien A und B zwei Ereignisse. $\overline{A}$, gelesen "A quer", bezeichnet das Komplement von A, ist also das Ereignis, dass A nicht eintritt.

$A \cup B$ ist die **Vereinigung** von A und B und bezeichnet das Ereignis, dass A **oder** B oder beide eintreten.

$A \cap B$ ist der **Durchschnitt** von A und B und bezeichnet das Ereignis, dass A **und** B **gemeinsam** eintreten.

Eine Menge von Ereignissen $A_1, A_2, ..., A_k$ ist eine **Zerlegung** von S, wenn die Ereignisse A_i **disjunkt** sind und ihre **Vereinigung** S ergibt, d. h.

$$A_i \cap A_j = \varnothing \quad \text{für } (i \neq j)$$
$$\cup A_i = S\,.$$

Vor diesem Hintergrund führte **Kolmogoroff** den Begriff der Wahrscheinlichkeit eines Ereignisses ein und formulierte die folgenden drei Axiome, die die Grundlage für das Rechnen mit Wahrscheinlichkeiten bilden:

Es sei

S die Menge der möglichen Ergebnisse eines Versuchs,

A, B seien Ereignisse, d. h. Teilmengen von S.

Mit $P(A)$ wird die Wahrscheinlichkeit eines Ereignisses A bezeichnet (P von engl.: probability).

Unter diesen Voraussetzungen gelten die **Axiome von Kolmogoroff**:

1. **Für jedes Ereignis A gilt: $P(A) \geq 0$,**
2. **Für das sichere Ereignis S gilt: $P(S) = 1$,**
3. **Sind die Ereignisse A und B disjunkt ($A \cap B = \emptyset$), dann gilt: $P(A \cup B) = P(A) + P(B)$.**

Aus den **Axiomen** folgt:

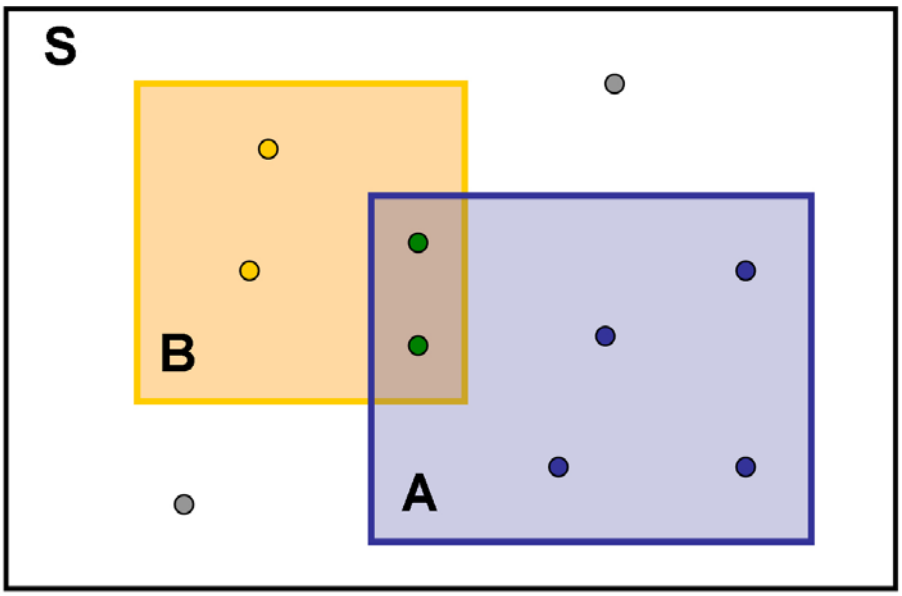
$$P(\overline{A}) = 1 - P(A) .$$

Für **beliebige** Ereignisse A und B gilt der **Additionssatz**:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) .$$

Das Rechnen mit Wahrscheinlichkeiten lässt sich gut anhand von sogenannten Venn-Diagrammen veranschaulichen. Abbildung 3.1 zeigt das Venn-Diagramm einer Menge S, die aus 10 **gleichwahrscheinlichen** Elementen besteht, sowie zwei Ereignisse A und B.

Abbildung 3.1: Venn-Diagramm



Beispiel 3.2
Für die Ereignisse A und B aus der Abbildung 3.1 gilt:

$$P(A) = 0.6,$$

$$P(B) = 0.4,$$

$$P(A \cup B) = 0.8,$$

$$P(A \cap B) = 0.2,$$

$$P(\overline{A}) = 0.4.$$

3.2.2 Bedingte Wahrscheinlichkeiten, Bayessche Formel

Beim Würfeln mit einem idealen Würfel ist die Wahrscheinlichkeit für eine 6 unter der **Bedingung**, dass eine gerade Zahl gewürfelt wird, offenbar 1/3, denn 6 ist eine der drei gleichwahrscheinlichen Möglichkeiten 2, 4, 6.

Dies ist ein Beispiel für die bedingte Wahrscheinlichkeit ***P(A|B)*** eines Ereignisses ***A*** - hier Würfeln einer 6 - unter der Bedingung des Ereignisses ***B*** - hier Würfeln einer geraden Zahl. Die allgemeine Definition lautet:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} .$$

Im Beispiel gilt $P(A) = 1 / 6$, $P(B) = 1 /2$ und $P(A \cap B) = 1 / 6$. Die formale Rechnung führt zum gleichen Ergebnis wie die intuitive Überlegung.

Im typischen medizinischen Anwendungsbeispiel ist ***A*** eine **Krankheit** - z. B. Masern - und ***B*** ein **Symptom** - z. B. roter Hautausschlag. In diesem Fall ist $P(A|B)$ die bedingte Wahrscheinlichkeit für Masern bei rotem Hautausschlag. Den Arzt, der den roten Ausschlag beim Patienten sieht, interessiert bei der Diagnosestellung diese bedingte Wahrscheinlichkeit und nicht etwa die unbedingte Wahrscheinlichkeit für Masern.

Wenn man die Rollen von A und B vertauscht, fragt man nach der bedingten Wahrscheinlichkeit von $P(B|A)$, d. h. nach der bedingten Wahrscheinlichkeit von rotem Hautausschlag bei Masern. Die **Formel von Bayes** stellt den Zusammenhang zwischen diesen beiden bedingten Wahrscheinlichkeiten her. Durch Auflösen der Definitionsgleichung nach $P(A \cap B)$ erhält man

$$P(A|B) \cdot P(B) = P(A \cap B) = P(B|A) \cdot P(A)$$

und damit

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Wegen

$$P(B) = P(B \cap A) + P(B \cap \overline{A}) = P(B \mid A) \cdot P(A) + P(B \mid \overline{A}) \cdot P(\overline{A})$$

erhält man hieraus schließlich die **Bayessche Formel**:

$$P(A \mid B) = \frac{P(B| A) \cdot P(A)}{P(B| A) \cdot P(A) + P(B| \overline{A}) \cdot P(\overline{A})}$$

*Im betrachteten Beispiel sagt diese Formel, dass man die bedingte Wahrscheinlichkeit **P(A|B)** von Masern bei rotem Ausschlag berechnen kann, wenn man die Wahrscheinlichkeit **P(A)** für Masern und die bedingten Wahrscheinlichkeiten des Symptoms 'roter Hautausschlag' für Patienten mit und Patienten ohne Masern kennt.*

*Man interpretiert dies als Berechnung der **a-posteriori** Wahrscheinlichkeit **P(A|B)** aus der **a-priori** Wahrscheinlichkeit **P(A)**.*

Beispiel 3.3
Ein Arzt in einer Rheuma-Ambulanz betreut im Jahr etwa 200 Patienten jüngeren und mittleren Lebensalters mit chronischen Rückenschmerzen, die nicht eindeutig auf einen Bandscheibenvorfall zurückzuführen sind. Aus langjähriger Erfahrung weiß der Arzt, dass bei etwa 60% der Patienten ein Morbus Bechterew (K^+) vorliegt. Um die Sicherheit bezüglich der Diagnose zu verbessern, könnte er eine HLA-Typisierung durchführen lassen. Er weiß, dass 95% aller Patienten mit einem Morbus Bechterew (K^+) das HLA-Antigen B 27 (T^+) haben, das in der übrigen Bevölkerung (K^-) nur mit einem Anteil von 8% vorkommt. Wie ändert sich die Sicherheit der Diagnose, nachdem der Arzt erfährt, dass das Ergebnis der Typisierung positiv ist?

Aus den obigen Angaben ergeben sich die folgenden Wahrscheinlichkeiten:

$$P(K^+) = 0.6, \qquad P(K^-) = 0.4$$

$$P(T^+ \mid K^+) = 0.95, \quad P(T^+ \mid K^-) = 0.08.$$

*Die gesuchte Wahrscheinlichkeit $P(K^+|T^+)$ ist dann nach der **Formel von Bayes***

$$P(K^+ \mid T^+) = (0.95 \cdot 0.6)/(0.95 \cdot 0.6 + 0.08 \cdot 0.4) = 0.947.$$

Dies bedeutet, dass in etwa 95% der Fälle mit einem positiven Testergebnis (HLA B 27-Antigen vorhanden) ein Morbus Bechterew vorliegen wird.

3.2.3 Sensitivität und Spezifität

Tabelle 3.1: Sensitivität und Spezifität beim klinischen Test

Testergebnis T	**Wirklichkeit W**		
	infiziert	**nicht infiziert**	**Summe**
positiv	900	9900	10800
negativ	100	89100	89200
Summe	1000	99000	100000

In einer fiktiven Grundgesamtheit von 100 000 Personen sind 1 000 Personen mit einem bestimmten Virus infiziert. Es gibt einen klinischen Test, mit dem man dies feststellen kann. Dieser Test ist allerdings nicht hundertprozentig sicher. Es werden nur 90 % der tatsächlich

infizierten Personen im Test als **positiv** erkannt (Sensitivität), und genauso sind nur 90 % der nicht infizierten Personen im Test **negativ** (Spezifität).

Die Verhältnisse sind in Tabelle 3.1 tabellarisch dargestellt.

Sensitivität und **Spezifität** sind in der Sprache der Wahrscheinlichkeitsrechnung nichts anderes als bedingte Wahrscheinlichkeiten:

Sensitivität: $P(T=+ \mid W=+) = 900 / 1000 = 0.9$,
Spezifität: $P(T=- \mid W=-) = 89100 / 99000 = 0.9$.

In der Praxis möchte man umgekehrt auch wissen, wie groß bei gegebener Sensitivität und Spezifität die Wahrscheinlichkeit ist, dass ein im Test positiver Patient tatsächlich infiziert ist, bzw. ein im Test negativer Patient tatsächlich nicht infiziert ist. Diese bedingten Wahrscheinlichkeiten nennt man "**positiven**" bzw. "**negativen**" prädiktiven Wert.

Für diese beiden bedingten Wahrscheinlichkeiten gilt:

Positiver prädiktiver Wert: $P(W= + \mid T= +) = 900 / 10800 = 0.0833$,

d. h. nur **8.3%** der Personen mit einem positiven Test sind tatsächlich auch mit dem Virus infiziert.

Negativer prädiktiver Wert: $P(W= - \mid T= -) = 89100 / 89200 = 0.9989$.

Aus der Bayesschen Formel folgt, dass der positive bzw. negative prädikative Wert außer von der Sensitivität und der Spezifität auch noch von der Wahrscheinlichkeit $P(W= +)$, der **Prävalenz** der Krankheit, abhängt.

Weitere Kennzahlen zur Beurteilung eines diagnostischen Verfahrens sind die Wahrscheinlichkeiten für einen falsch positiven bzw. falsch negativen Wert.

Für diese beiden Wahrscheinlichkeiten gilt:

Falsch positiver Wert: $P(\{W= -\} \cap \{T= +\}) = 9900 / 100000 = 0.099$

Falsch negativer Wert: $P(\{W= +\} \cap \{T= -\}) = 100 / 100000 = 0.001$

Beispiel 3.4
Das Bakterium Helicobacter Pylori spielt eine wesentliche Rolle bei der Entstehung von Ulzera (Geschwüren) im Magen. In Deutschland sind etwa 20 % der Erwachsenen unter 40 Jahren mit diesem Bakterium infiziert.

Der sogenannte CLO-Test war der erste diagnostische Test zum Nachweis von Helicobacter Pylori. Der Test ist seit 1988 im Einsatz und hat eine Sensitivität von 98 % und eine Spezifität von 97 %.

Für ein Screening von 1000 Erwachsenen unter 40 Jahren ergibt sich aus den Angaben folgende Tabelle:

Testergebnis T	Wirklichkeit W		
	infiziert	**nicht infiziert**	**Summe**
positiv	196	24	220
negativ	4	776	780
Summe	200	800	1000

Damit erhält man die folgenden weiteren Kennzahlen:

***Positiver prädiktiver Wert**: **P(W=+ | T=+)** = 196 / 220 = 0.8909, d. h. 89.1 % der Personen mit einem positiven Test sind tatsächlich auch mit dem Bakterium infiziert.*

***Negativer prädiktiver Wert**: **P(W=- | T=-)** = 766 / 780 = 0.99487, d. h. 99.5 % der Personen mit einem negativen Test sind tatsächlich auch nicht mit dem Bakterium infiziert.*

***Falsch positiver Wert**: **P({W=-}∩{T=+})** = 24/1000 = 0.024*

***Falsch negativer Wert**: **P({W=+} ∩ {T=-})** = 4/1000 = 0.004*

3.2.4 Unabhängigkeit von Ereignissen

Der Begriff Unabhängigkeit von Ereignissen spielt in den Anwendungen eine wichtige Rolle. Zwei Ereignisse ***A*** und ***B*** heißen **unabhängig** voneinander, wenn

$$P(A|B) = P(A)$$

gilt.

Nach Definition der bedingten Wahrscheinlichkeit ist die Definition gleichwertig mit

$$P(A \cap B) = P(A) \cdot P(B),$$

d. h., die Wahrscheinlichkeit für das gemeinsame Eintreten zweier unabhängiger Ereignisse entspricht dem Produkt der beiden Einzelwahrscheinlichkeiten. Man nennt dies auch den **Multiplikationssatz** der Wahrscheinlichkeitsrechnung.

Wenn man allein die mathematische Seite des Problems betrachtet, ist die Frage nach der Unabhängigkeit zweier Ereignisse A und B leicht zu beantworten. Man hat nur zu überprüfen, ob

$$P(A \cap B) = P(A) \cdot \mathrm{P}(B)$$

gilt oder nicht.

Für den Biometriker besteht aber das Problem darin, ob er in einem der folgenden Beispiele zur Auswertung ein mathematisches Modell ansetzen darf, das Unabhängigkeit voraussetzt, oder ob er das nicht darf.

Beispiel 3.5

1. *Das Geschlecht von Kind 1 und Geschlecht von Kind 2 bei zwei Kindern des gleichen Elternpaares sind unabhängig (?).*
2. *Die Dauer der Geburt und die Art der Narkose sind nicht unabhängig.*
3. *Die Dauer der Geburt und das Alter der Mutter sind nicht unabhängig.*
4. *Die Dauer der Geburt und die Parität der Mutter sind nicht unabhängig.*
5. *Die Dauer der Geburt und das Alter des Vaters sind unabhängig (?).*
6. *Die Augenfarbe und die Haarfarbe einer Person sind nicht unabhängig (?).*
7. *Lebenserwartung und Geschlecht sind nicht unabhängig.*
8. *Studienfach und Geschlecht sind nicht unabhängig.*

3.2.5 Gesetz der großen Zahl

Unbekannte Wahrscheinlichkeiten schätzt man durch die relativen Häufigkeiten aus einer zufälligen Stichprobe. Theoretische Grundlage hierfür ist das **Gesetz der großen Zahl**.

A sei ein Ereignis mit der Wahrscheinlichkeit $P(A)$. n_A sei die absolute Häufigkeit mit der A in n unabhängigen Versuchswiederholungen eintritt.

Das **Gesetz der großen Zahl** besagt:

$$\frac{n_A}{n} \to P(A)(n \to \infty)$$

d. h., die **relative Häufigkeit**, mit der ein Ereignis A in n unabhängigen Versuchswiederholungen eintritt, strebt mit wachsendem n gegen die Wahrscheinlichkeit des Ereignisses A.

Beispiel 3.6
Die relative Häufigkeit, mit der man bei Neugeborenen ein Geburtsgewicht unter 3000 g beobachtet, strebt mit wachsender Anzahl n der ausgewerteten Geburten gegen die Wahrscheinlichkeit dieses Ereignisses.

3.3 Zufallsvariable

Bei vielen Zufallsexperimenten werden mögliche Ergebnisse oder Ereignisse durch **Zahlen** beschrieben.

Beispiel 3.7
Bei einem Wurf mit einem Würfel kann man die möglichen Ereignisse 'k geworfene Augen' einfacher durch die Zahl k (k=1,2,...,6) beschreiben.

Dies ist aber nur sinnvoll, wenn - wie in diesem Beispiel - jedem möglichen Ergebnis der Grundmenge eine Zahl zugeordnet wird und die Zuordnung eindeutig ist. Die Zuordnungs- oder Abbildungsvorschrift heißt **Zufallsvariable**. Der Begriff der Zufallsvariablen entspricht dem Begriff des Merkmals in der deskriptiven Statistik. In diesem Kapitel wird dargelegt, wie eine Merkmalsverteilung mittels mathematischer Funktionen und statistischer Kenngrößen beschrieben werden kann. Wie bei den quantitativen Merkmalen unterscheiden wir diskrete und stetige Zufallsvariable.

3.3.1 Diskrete Zufallsvariable

Der Wertebereich der Zufallsvariablen besteht aus diskret auf der Zahlengeraden liegenden Zahlen $x_1, x_2, \ldots$ (vgl.: quantitativ diskretes Merkmal).

Beispiel 3.7 (Fortsetzung)
Bei der Zufallsvariablen ***'gewürfelte Augenzahl'*** *besteht der Wertebereich aus den Zahlen* ***1,2,3,4,5,6****.*

Wahrscheinlichkeitsfunktion *f*:

$$f(x_j) = p_j = P(X = x_j) \qquad (j = 1,2,\ldots).$$

Die Wahrscheinlichkeitsfunktion *f* liefert zu jedem *x* die Wahrscheinlichkeit, mit der die Zufallsvariable *X* den Wert *x* annimmt. Sie ist das theoretische Analogon zur Häufigkeit**sverteilung** eines diskreten Merkmals in der deskriptiven Statistik.

Beispiel 3.7 (Fortsetzung)
Bei der Zufallsvariablen ***'gewürfelte Augenzahl'*** *gilt für die Wahrscheinlichkeitsfunktion*

$$f(x_j) = 1/6 \qquad (j = 1,2,3,4,5,6).$$

Verteilungsfunktion *F*:

$$F(x) = \sum_{x_j \le x} f(x_j).$$

Die Verteilungsfunktion *F* liefert zu jedem *x* die Wahrscheinlichkeit, mit der die Zufallsvariable *X* Werte kleiner oder gleich *x* annimmt. Sie ist das theoretische Analogon zur empirischen Verteilungsfunktion in der deskriptiven Statistik.

Beispiel 3.7 (Fortsetzung)
Für die Zufallsvariablen ***'gewürfelte Augenzahl'*** *ergibt sich die in Abbildung 3.3 dargestellte* ***Verteilungsfunktion F(x***

Abbildung 3.2: Wahrscheinlichkeitsfunktion der Zufallsvariable 'gewürfelte Augenzahl'

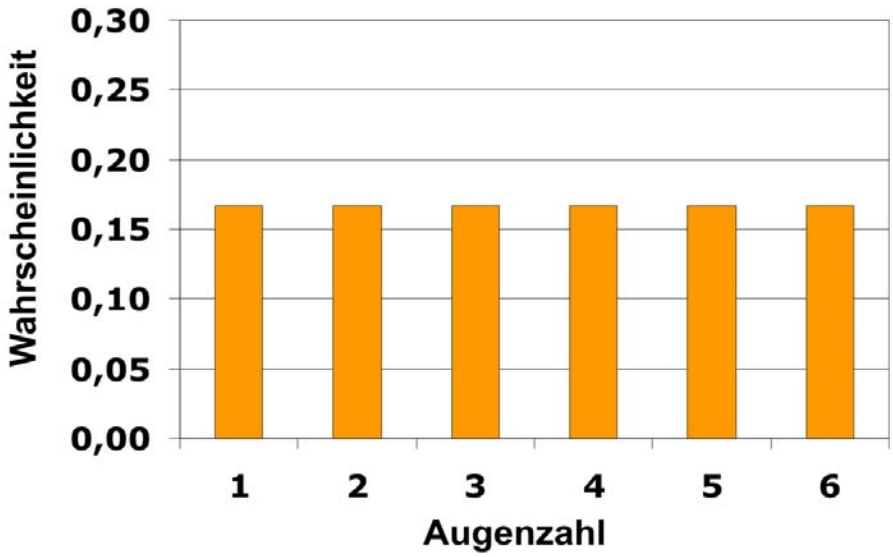
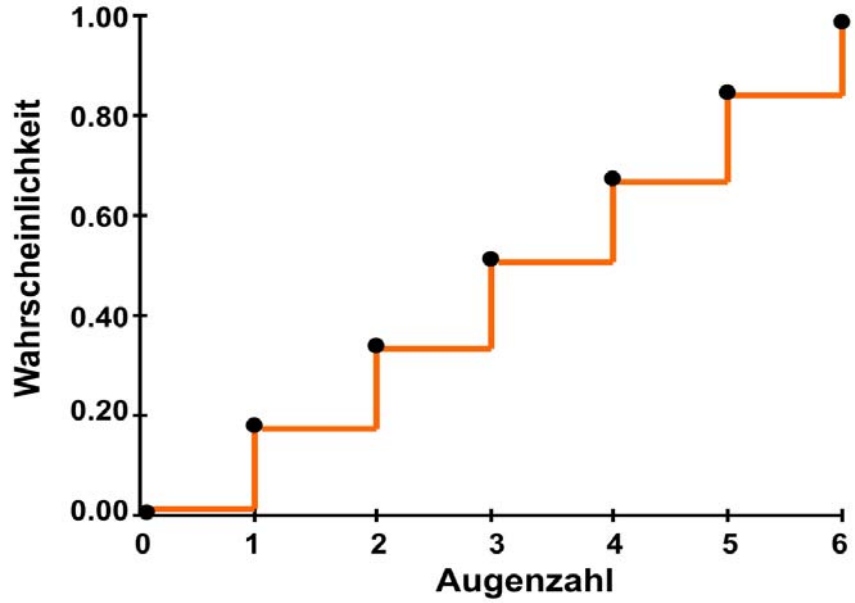


Abbildung 3.3: Verteilungsfunktion der Zufallsvariable 'gewürfelte Augenzahl'



Erwartungswert *E(X)*:

$$E(X) = \mu = \sum_j x_j \cdot p_j .$$

Der Erwartungswert *E(X)* einer Zufallsvariablen *X* ist formal ähnlich gebildet wie der **arithmetische** Mittelwert, nur sind die beim Mittelwert auftretenden relativen Häufigkeiten durch Wahrscheinlichkeiten ersetzt. Aus dem Gesetz der großen Zahl folgt daher, dass sich mit wachsendem Stichprobenumfang *n* der Mittelwert immer mehr dem Erwartungswert

nähert.

Beispiel 3.7 (Fortsetzung)
Bei der Zufallsvariablen 'gewürfelte Augenzahl' gilt für den Erwartungswert

$$E(X) = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 21/6 = 3.5 .$$

Die **Varianz *V(X)*** einer Zufallsvariable ist die erwartete quadratische Abweichung der Zufallsvariablen *X* von ihrem Erwartungswert *E(X)*.

$$V(X) = \sigma^2 = E[(X - \mu)^2] = \sum_j (x_j - \mu)^2 \cdot p_j .$$

Sie ist das theoretische Analogon zur **empirischen Varianz** der deskriptiven Statistik. Die positive Wurzel σ aus der Varianz heißt Standardabweichung.

Beispiel 3.7 (Fortsetzung)
*Bei der Zufallsvariablen **'gewürfelte Augenzahl'** gilt für die **Varianz***

$$V(X) = (1-3.5)^2 \cdot 1/6 + (2-3.5)^2 \cdot 1/6 + (3-3.5)^2 \cdot 1/6 + (4-3.5)^2 \cdot 1/6 + (5-3.5)^2 \cdot 1/6 + (6-3.5)^2 \cdot 1/6 = 17.5 \cdot 1/6 = 2.9167.$$

3.3.2 Stetige Zufallsvariable

Der Wertebereich einer stetigen Zufallsvariablen umfasst ein ganzes Intervall der Zahlengeraden (vgl.: quantitativ **stetiges** Merkmal).

Bei der Verteilungsfunktion ***F(x)*** einer **stetigen** Zufallsvariable wird das **Summenzeichen** in der Formel für eine diskrete Zufallsvariable durch ein **Integralzeichen** ersetzt:

$$F(X) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

f(x), die erste Ableitung der Verteilungsfunktion, ist die **Dichtefunktion**:

$$f(x) = F^t(x) = \frac{dF(x)}{dx}$$

Beim Erwartungswert ***E(x)*** und bei der Varianz ***V(x)*** einer stetigen Zufallsvariable wird analog zur Verteilungsfunktion das **Summenzeichen** in den Formeln für diskrete Zufallsvariable durch ein **Integralzeichen** ersetzt:

Der **Erwartungswert** *E(X)* ist dann:

$$E(X) = \mu = \int x \cdot f(x) dx.$$

Für die **Varianz** *V(X)* gilt:

$$V(X) = \sigma^2 = E[(X - \mu)^2] = \int (x - \mu)^2 \cdot f(x) dx.$$

Die bekannteste und für die Statistik wichtigste stetige Verteilung ist die Normalverteilung, die weiter unten noch ausführlich behandelt wird.

Allgemein kann man sagen, in der analytischen Statistik modelliert man stetige Merkmale durch stetige Zufallsvariable und diskrete Merkmale durch diskrete Zufallsvariable.

3.3.3 Binomial- und Poissonverteilung

Beispiel 3.8
In einem Versuch sollen 10 Versuchstiere auf zwei Behandlungen A bzw. B verteilt werden. Jedem Behandlungsarm sollen 5 Versuchstiere zufällig zugeteilt werden. Der Versuchsleiter nummeriert die Tiere von 1 bis 10 durch. Den Tieren mit den ungeraden Nummern 1, 3, 5, 7, 9 teilt er die Behandlung per Münzwurf zu. Bei "Zahl" wird Therapie A zugeordnet, bei "Wappen" wird Therapie B zugeordnet.

Den Tieren mit den geraden Nummern 2, 4, 6, 8, 10 teilt er die jeweils andere Behandlung zu. Auf diese Weise ist die geforderte 5:5 Verteilung sichergestellt. Die Anzahl der Tiere mit ungerader Versuchsnummer, die der Behandlung A zugeteilt werden, ist eine diskrete Zufallsvariable X.

X, die Anzahl der Versuchstiere mit ungerader Nummer, denen die Behandlung A zugeteilt wird, kann die Werte 0, 1, 2, 3, 4 oder 5 annehmen.

Die Menge S der möglichen Ergebnisse des fünfmaligen Münzwurfs, lässt sich als die Menge aller fünfstelligen Folgen (x_1, x_2, x_3, x_4, x_5) schreiben, wobei x_i entweder 0 oder 1 ist, je nachdem, ob im i-ten Versuch Wappen oder Zahl geworfen wurde (i=1,2,3,4,5).

Das so definierte S enthält 2^5=32 Elemente (= mögliche Ergebnisse).

Wenn man davon ausgeht, dass Wappen bzw. Zahl mit der gleichen Wahrscheinlichkeit geworfen werden, sind alle Ergebnisse in S gleichwahrscheinlich, und es gilt für jedes

Ereignis {e} aus der Menge S:

$$P(e) = 1/32$$

Binomialkoeffizient

Die Anzahl der Möglichkeiten, ohne Berücksichtigung der Reihenfolge, aus ***n*** Dingen ***k*** auszuwählen, wird mit $\binom{n}{k}$ - gelesen *n* über *k* - bezeichnet und heißt Binomialkoeffizient.

Beim **Zahlenlotto** werden 6 aus 49 Kugeln ausgewählt. Es gibt also $\binom{49}{6}$ Möglichkeiten.

Dies ist nur die Schreibweise. Wenn man die Anzahl wirklich ausrechnen will, benötigt man die Formel

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot k}$$

Eine Herleitung der Formel findet man in den einschlägigen Lehrbüchern. Beim Zahlenlotto wird daraus z. B.:

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = 13983816 \quad .$$

Es ist allgemein üblich, die Abkürzung

$$k! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot k$$

(*k*! wird "*k* Fakultät" gelesen) zu verwenden. Mit dieser Abkürzung gilt

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} = \frac{n!}{k!(n-k)!} \quad .$$

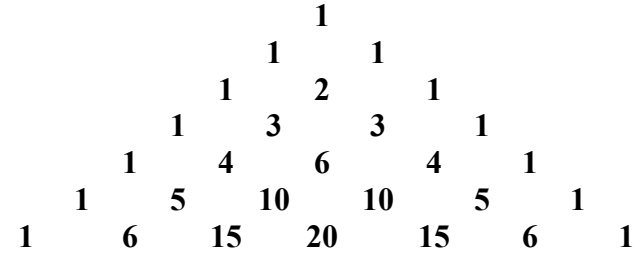
Aufgrund der obigen Definition gilt

$$\binom{n}{k} = \binom{n}{n-k},$$
$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \quad .$$

Mit Hilfe dieser Formeln und der Festlegung

$$\binom{n}{0} = 1$$

lassen sich die Binomialkoeffizienten rekursiv berechnen (**Pascalsches Dreieck**).



Binomial- und Bernoulliverteilung sind zwei spezielle diskrete Wahrscheinlichkeits-verteilungen, die in den Anwendungen häufig auftreten. Deshalb sollen sie hier etwas ausführlicher betrachtet werden.

Es wird 12-mal gewürfelt und gefragt, wie oft dabei eine 6 erscheint. Diese Anzahl X ist ein Beispiel für eine binomialverteilte Zufallsvariable mit den speziellen Parametern n=12 für die Anzahl der Würfe und p=1/6 für die Wahrscheinlichkeit einer 6 in jedem einzelnen der 12 Würfe. Man schreibt abkürzend:

$$X : B(12;\ 1/6).$$

Allgemein ergibt sich eine binomialverteilte Zufallsvariable ***X:B(n,p)*** unter folgenden Bedingungen:

In **einem** Versuch tritt ein Ereignis A mit der Wahrscheinlichkeit $P(A) = p$ ein. (Beim Würfeln ist das die 6, die mit der Wahrscheinlichkeit 1/6 gewürfelt wird). Dieser Versuch wird ***n*-mal** unter identischen Bedingungen **wiederholt**. (Im Beispiel wird 12-mal gewürfelt).

Die binomialverteilte Zufallsvariable X gibt an, wie oft A bei den n Versuchswiederholungen eintritt. X kann offenbar die Werte $0,1,...,n$ annehmen. Z. B. $X = n$ bedeutet, dass in allen n Versuchswiederholungen das Ereignis A eingetreten ist, als z. B. dass in allen 12 Würfen die 6 gekommen ist. Das wäre zwar recht überraschend, ist aber keineswegs ausgeschlossen.

Man interessiert sich für die Wahrscheinlichkeit

$P(X=k)$ $\quad (k=0, 1, ..., n)$, mit der X den möglichen Wert k annimmt.

Man findet

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)}.$$

Der Beweis für diese Formel soll am Würfelbeispiel $X:B(12\ ;\ 1/6)$ kurz skizziert werden. Wir fragen beispielsweise nach $P(X=3)$.

Jedes Ergebnis des 12-maligen Würfelns lässt sich durch eine 12-stellige Folge von 6 und $\overline{6}$ beschreiben. Dabei steht '6' für das Würfeln einer 6 und '$\overline{6}$' für das Würfeln einer von 6 verschiedenen Zahl.

Ein mögliches Ergebnis wäre z. B.

$$e = (\overline{6},6,\overline{6},\overline{6},\overline{6},6,\overline{6},\overline{6},\overline{6},\overline{6},6,\overline{6}) .$$

X(e)=3 bedeutet dann, dass in dieser Folge genau dreimal eine 6 und 9-mal eine $\overline{6}$ steht. Es gibt $\binom{12}{3}$ verschiedene solcher Folgen, denn jede Folge entspricht genau einer Möglichkeit, aus den 12 Plätzen 3 auszuwählen und sie mit einer 6 zu besetzen.

Jede einzelne dieser $\binom{12}{3}$ Folgen hat wegen der Unabhängigkeit der Versuchswiederholungen die Wahrscheinlichkeit $(\frac{1}{6})^3(\frac{5}{6})^9$. Da das Ereignis $\{X = 3\}$ aus genau $\binom{12}{3}$ solcher Folgen besteht, ergibt sich insgesamt

$$P(X = 3) = \binom{12}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^9 ,$$

genau wie es die Formel behauptet.

Das Ergebnis einer einzelnen der *n* Versuchswiederholungen lässt sich ebenfalls durch eine Zufallsvariable beschreiben:

$$Y_i = \begin{cases} 1 & A \text{ tritt in der i - ten Wiederholung ein} \\ 0 & \text{sonst} \end{cases} \qquad (i = 1,2,\ldots,n) .$$

Offenbar ist Y_i*:B(1,p)* der Spezialfall einer Binomialverteilung für *n=1*. Diesen Spezialfall nennt man **Bernoulliverteilung**. Aus der Definition folgt unmittelbar

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

d. h., jede Binomialverteilung ist als Summe von *n* unabhängigen Bernoulliverteilungen darstellbar.

Diese Tatsache hilft bei der Berechnung von *E(X)* und *V(X)*. Für die Bernoulliverteilung Y_i folgt unmittelbar aus der Definition von Erwartungswert und Varianz

$$\begin{aligned} E(Y_i) &= 1\cdot p + 0\cdot(1-p) = p \qquad (i = 1,\ldots,n) \quad , \\ V(Y_i) = E((Y_i - p)^2) &= (1-p)^2\cdot p + (0-p)^2\cdot(1-p) \\ &= (1-p)^2\cdot p + p^2(1-p) \\ &= p\cdot(1-p)\cdot[(1-p)+p] \\ &= p\cdot(1-p) \quad . \end{aligned}$$

Für den Erwartungswert *E(X)* folgt nun

$$
\begin{aligned}
E(X) &= E\left(\sum_{i=1}^{n}(Y_i)\right) \\
&= \sum_{i=1}^{n} E(Y_i) \\
&= n \cdot p \ .
\end{aligned}
$$

Bei der Berechnung der Varianz *V(X)* macht man sich die - hier nicht bewiesene - Tatsache zunutze, dass sich die Varianzen unabhängiger Zufallsvariablen addieren, daher also

$$
\begin{aligned}
V(X) &= V\left(\sum_{i=1}^{n}(Y_i)\right) \\
&= \sum_{i=1}^{n} V(Y_i) \\
&= n \cdot p(1-p).
\end{aligned}
$$

Die Abbildungen 3.4 bis 3.6 zeigen die Wahrscheinlichkeitsfunktionen der Binomialverteilungen *B(10,0.3)*, *B(10,0.5)* und *B(10,0.7)*. Man erkennt, dass sich der Gipfel der Verteilung mit wachsendem *p* nach rechts verlagert.

Abbildung 3.4: Wahrscheinlichkeitsfunktion der Binomialverteilung *B(10,0.3)*

f(x)

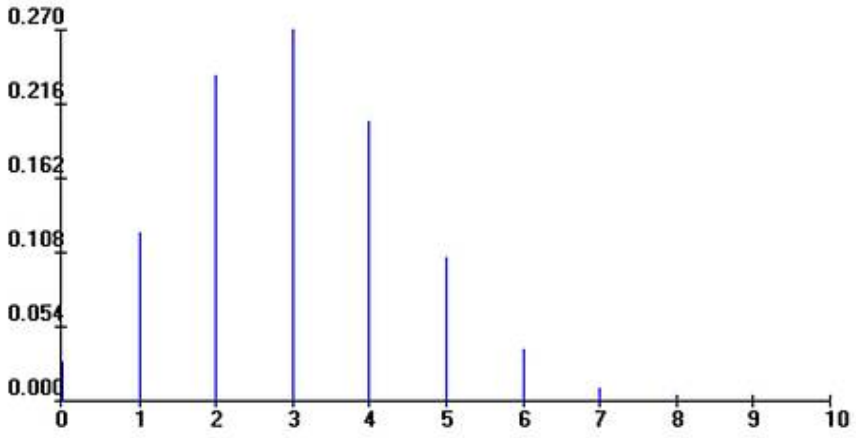


Abbildung 3.5: Wahrscheinlichkeitsfunktion der Binomialverteilung *B(10,0.5)*

f(x)

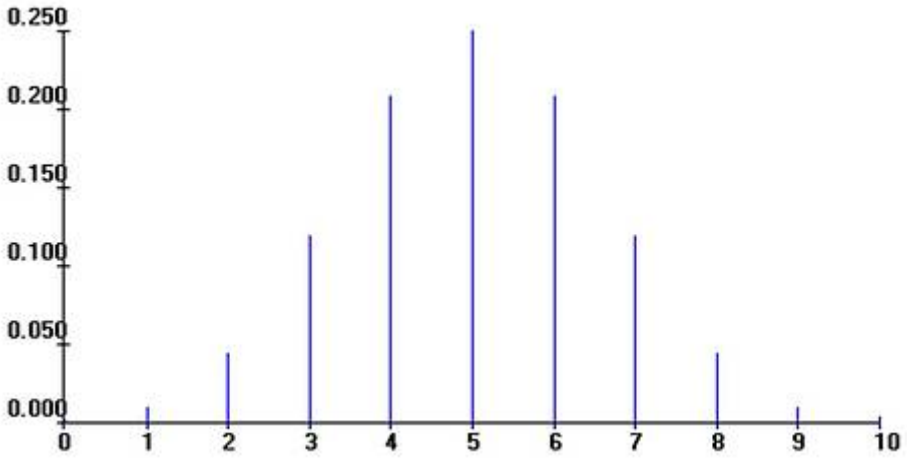
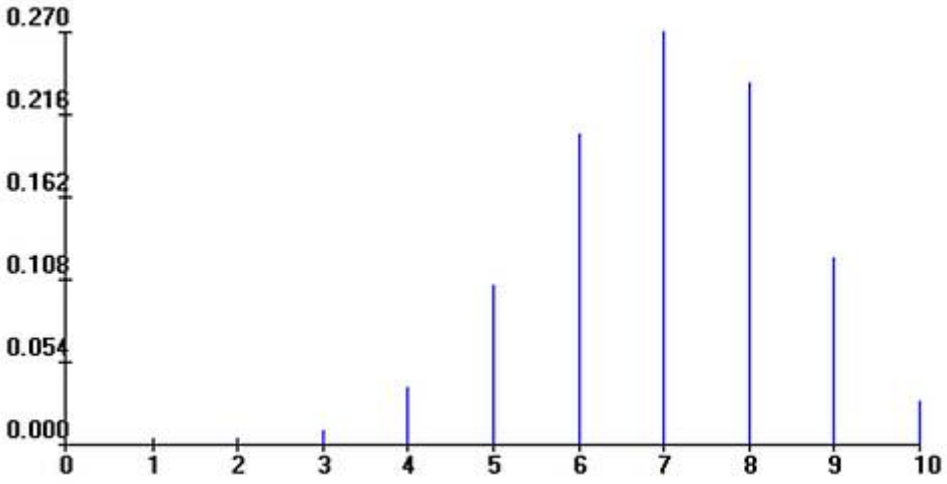


Abbildung 3.6: Wahrscheinlichkeitsfunktion der Binomialverteilung *B(10,0.7)*

f(x)



Beispiel 3.9
Ein Kinderarzt weiß aus Erfahrung, dass 20 % aller Neugeborenen nach unauffälliger Schwangerschaft weniger als 2500 g wiegen. Mehrlingsgeburten sind dabei ausgeschlossen. Aus den entsprechenden Geburtsprotokollen des vergangenen Jahres entnimmt er eine zufällige Stichprobe von n = 10 Protokollen. Die Anzahl der Protokolle, in denen ein Geburtsgewicht von weniger als 2500 g dokumentiert ist, ist eine diskrete ***binomialverteilte*** *Zufallsvariable X. X kann die Werte 0,1,2,3,4,5,6,7,8,9,10 annehmen. Die Wahrscheinlichkeitsfunktion von X ist in Tabelle 3.2 zu sehen, der Graf in Abbildung 3.7.*

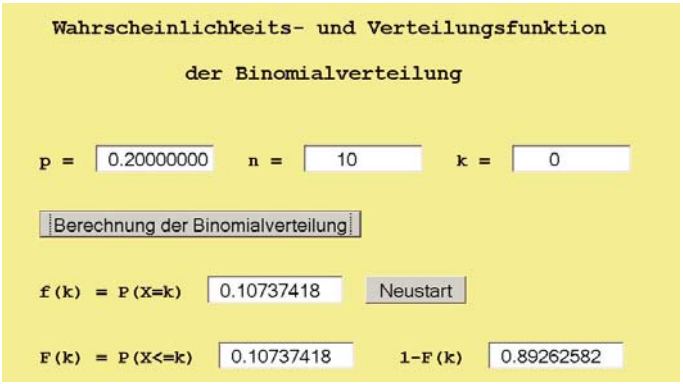
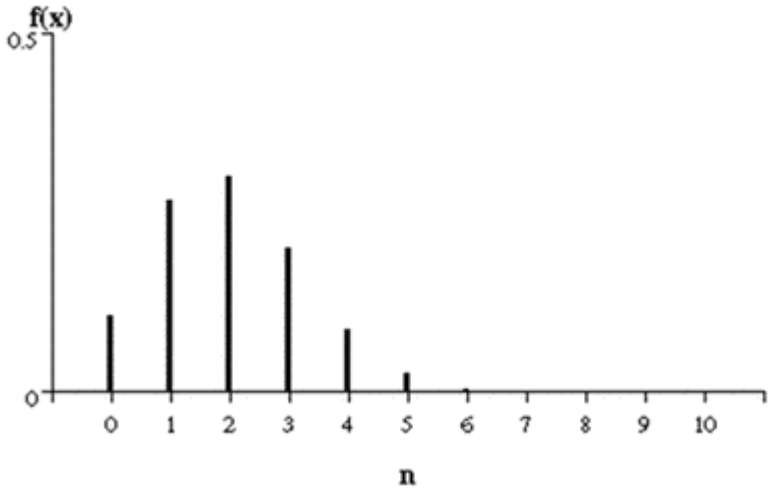


Tabelle 3.2: Wahrscheinlichkeitsfunktion von X

k=	0	1	2	3	4	5	6	7	8	9	10
f(k)=	.10737	.26844	.30199	.20133	.08808	.02642	.00551	.00079	.00007	.00000	.00000

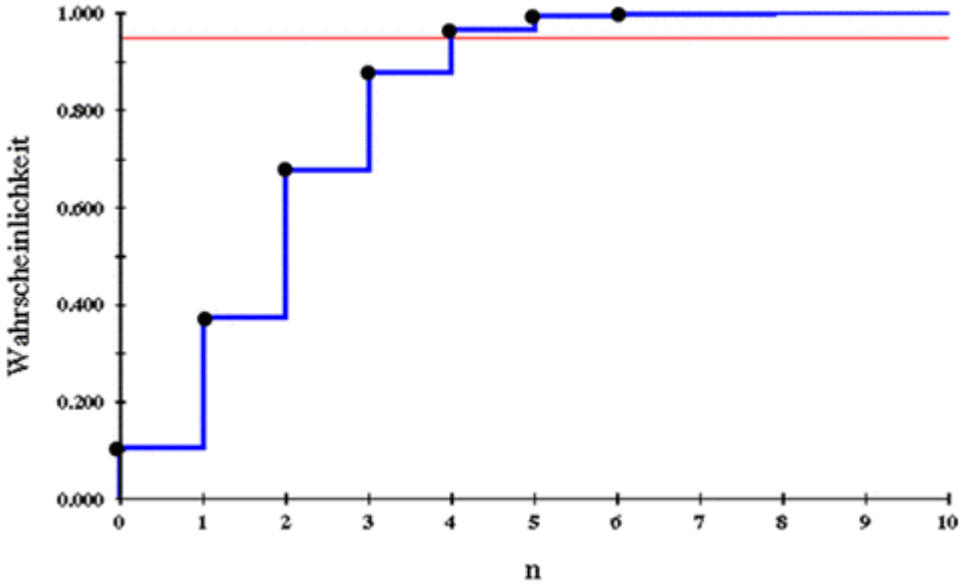
Abbildung 3.7: Wahrscheinlichkeitsfunktion der Binomialverteilung *B(10,0.2)*



Der Arzt muss im Mittel mit $n \cdot p = 10 \cdot 0{,}2 = 2$ Geburtsgewichten unter 2500 g rechnen. Wann man an der Richtigkeit des Erfahrungswerts p=0.2 zu zweifeln beginnt, dürfte individuell verschieden sein.

In der Statistik ist die Konvention weit verbreitet, dann eine Theorie zu verwerfen, wenn sie dem tatsächlich beobachteten Ereignis eine Wahrscheinlichkeit von $p \leq 0.05$ zuweist. Das ist hier der Fall, wenn in der Stichprobe von 10 Geburtsprotokollen mehr als 4 mit einem Geburtsgewicht unter 2500 g vorkommen.

Abbildung 3.8: Verteilungsfunktion der Binomialverteilung *B(10,0.2)*



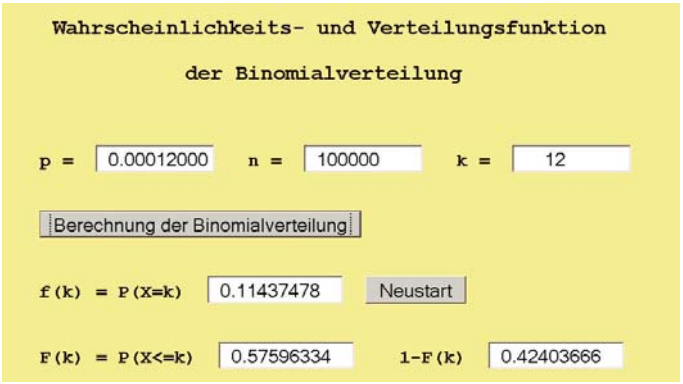
Die soeben vorgestellte Binomialverteilung ist durch die beiden Parameter ***n* und *p*** definiert, wobei *n* die Anzahl der unabhängigen Wiederholungen eines Zufallsexperiments und *p* die Erfolgswahrscheinlichkeit für jedes einzelne Experiment ist. Die Poissonverteilung hat nur einen **Parameter *λ***. Dabei ist sowohl der Erwartungswert als auch die Varianz gleich diesem Parameter.

Die Poissonverteilung gilt für **seltene** Ereignisse, die unabhängig voneinander auftreten. In einem solchen Fall entspricht die Binomialverteilung mit großem *n* und kleinem *p* näherungsweise einer Poissonverteilung mit Parameter $\lambda = n \cdot p$, dem Erwartungswert der Binomialverteilung.

Die Wahrscheinlichkeitsfunktion der Poissonverteilung lautet:

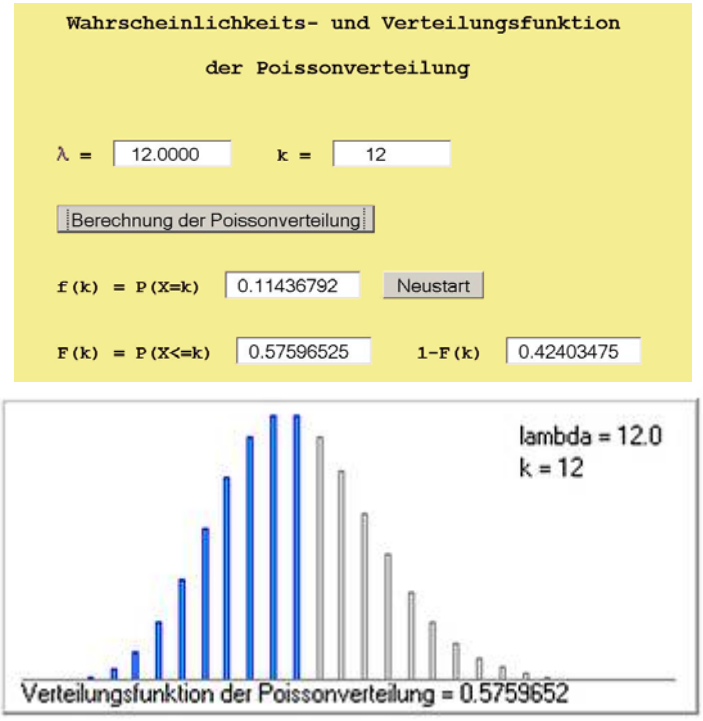
$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad k = 0,1,2,3,\ldots$$

Beispiel 3.10
Die mittlere Anzahl dem Kinderkrebsregister in Mainz gemeldeter Malignome betrug in den letzten zehn Jahren etwa 12 Fälle pro Jahr auf 100000 Kinder. Die Binomialverteilung mit n=100000 und p=12/100000 gibt an, wie groß die Wahrscheinlichkeit ist, dass z. B. im kommenden Jahr k=0,1,2,... Fälle pro 100000 Kinder gemeldet werden. Die Wahrscheinlichkeit für z. B. k=12 Fälle beträgt nach der Formel für die ***Binomialverteilung 0.11437478****.*



*Für die **Poissonverteilung** mit dem Parameter λ=(12/100000) · 100000=12 ergibt sich mit **0.11436792** nahezu der gleiche Wert.*

Abbildung 3.9: Wahrscheinlichkeitsfunktion der Poissonverteilungmit *λ = 12*



3.3.4 Schätzen von p

Bei den bisherigen Überlegungen ging man davon aus, dass *p*, die Wahrscheinlichkeit für das Eintreten des Ereignisses *A*, **bekannt** ist. In den meisten Anwendungen ist das aber **nicht** der Fall, und die beschriebenen Versuche werden durchgeführt, um *p* aufgrund des Versuchsergebnisses zu schätzen. Wenn das Ereignis *A* bei den *n* Versuchswiederholungen *k*-mal eingetreten ist, schätzt man *p* durch

$$\hat{p} = \frac{k}{n}.$$

Die Bezeichnung $\hat{p}$ (gelesen "p - Dach") für den **Schätzwert** ist allgemein üblich. Bei genauem Hinsehen erkennt man, dass der Schätzwert Realisation der Zufallsvariablen

$$H = \frac{1}{n} \cdot X$$

ist, wobei X die oben bereits eingeführte Zufallsvariable ist, die angibt, wie oft A bei den n Versuchswiederholungen eintritt. H ist nichts anderes als die relative Häufigkeit, mit der A in den n Versuchswiederholungen eintritt. Nach den Rechenregeln für Erwartungswert und Varianzwert bestimmt man

$$\begin{aligned} E(H) &= E\left(\frac{1}{n} \cdot X\right) = \frac{1}{n} \cdot E(X) = \frac{1}{n} \cdot n \cdot p = p \quad , \\ V(H) &= V\left(\frac{1}{n} \cdot X\right) = \frac{1}{n^2} \cdot V(X) = \frac{1}{n^2} \cdot n \cdot p \cdot (1-p) = \frac{p \cdot (1-p)}{n}. \end{aligned}$$

Das sind zwei erfreuliche und starke Ergebnisse. Das erste besagt, dass der Erwartungswert der Zufallsvariablen H, die man zum Schätzen von p verwendet, gleich dem Wert ist, den man schätzen will. Eine **Schätzung**, die diese Eigenschaft besitzt, nennt man **erwartungstreu** oder auch **unverzerrt** (engl.: unbiased).

Das zweite Ergebnis besagt, dass die erwartete quadratische Abweichung der Schätzung von ihrem Erwartungswert mit wachsendem n gegen Null geht. Das ist eine weitere Version des Gesetzes der großen Zahl.

Beispiel 3.11
Der Erfahrungswert von p=0.2 für die Geburtsgewichte unter 2500 g erscheint Ihnen nicht richtig. Sie wollen ihn aus Ihren Daten selber schätzen. Wenn Sie bei n=10 Protokollen k=3 Geburtsgewichte unter 2500 g gefunden haben, erhalten Sie mit der relativen Häufigkeit einen Schätzwert für p die Wahrscheinlichkeit:

$$\hat{p} = \frac{3}{10} = 0.3 .$$

Um die Genauigkeit der Schätzung zu verbessern, muss man den Stichprobenumfang vergrößern. Dies ist aus den obigen Formeln für den Erwartungswert und die Varianz ersichtlich. Der Erwartungswert der Schätzung entspricht dem wahren Wert (unverzerrte Schätzung), und die Varianz der Schätzung konvergiert mit wachsendem n gegen Null.

3.4 Normalverteilung

3.4.1 Verteilungsfunktion und Dichte

Eine stetige Zufallsvariable ***X*** heißt mit Erwartungswert μ und Varianz σ^2 normalverteilt, wenn die Wahrscheinlichkeit dafür, dass *X* höchstens gleich *x* ist, durch das Integral der **Gaußschen Fehlerfunktion** gegeben ist, in Formeln:

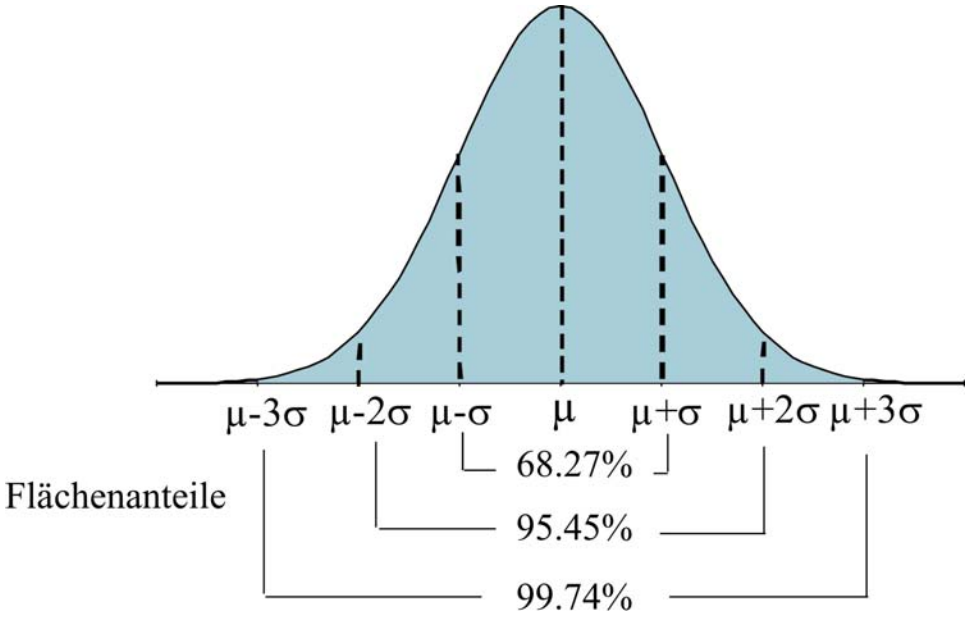
$$P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

Hierfür schreibt man abkürzend $X{:}N(\mu,\sigma^2)$. $F(x) = P(X \le x)$ ist die Verteilungsfunktion der Normalverteilung. Deren **erste Ableitung**

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

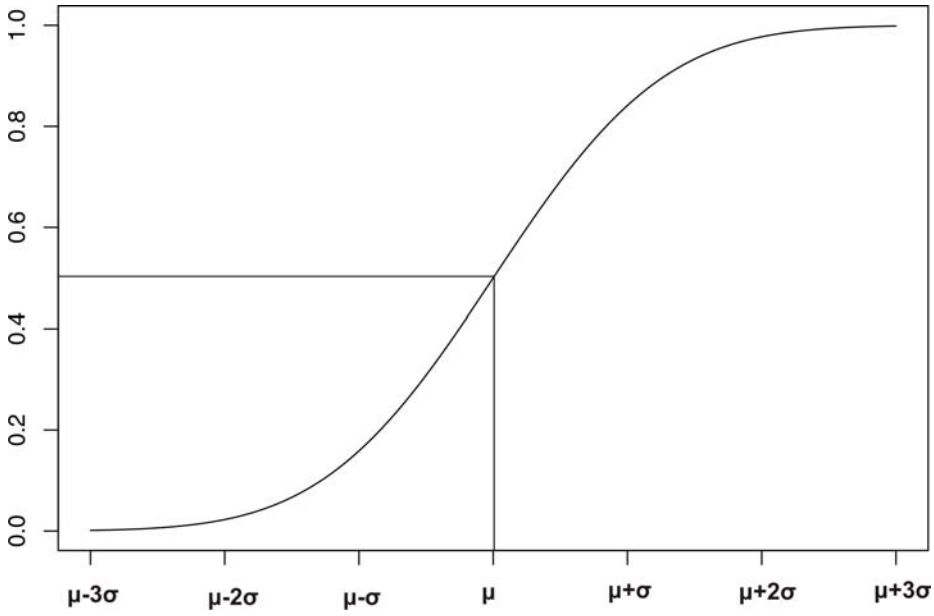
ist die Dichtefunktion der Normalverteilung. Das Bild der Dichte ist die bekannte **Glockenkurve** (Abbildung 3.10):

Abbildung 3.10: Dichtefunktion der Normalverteilung $N(\mu,\sigma^2)$



Die Verteilungsfunktion der Normalverteilung hat einen **sigmoiden** (s-förmigen) Kurvenverlauf (Abbildung 3.11).

Abbildung 3.11: Verteilungsfunktion der Normalverteilung $N(\mu,\sigma^2)$



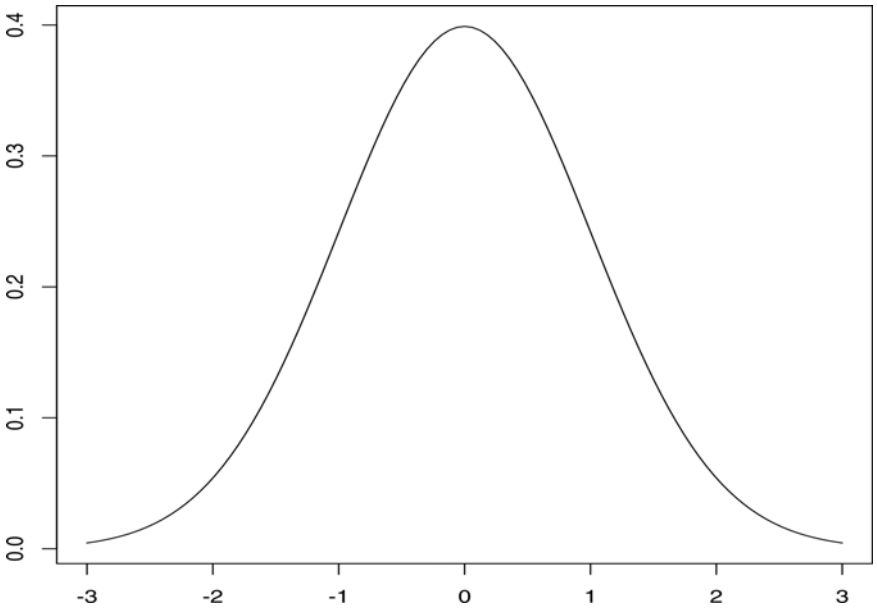
Es gibt die Normalverteilung für **jedes μ** und **jedes positives** σ.

Aus den Formeln und den Abbildungen werden die folgenden **Eigenschaften** der Normalverteilung deutlich:

- Die **Dichtefunktion** ist **symmetrisch** um den **Erwartungswert** μ.
- Sie hat zwei **Wendepunkte** bei $x = \mu-\sigma$ und $x = \mu+\sigma$.
- Sie erreicht ihr **Maximum** an der Stelle $x = \mu$.
- Der **Erwartungswert** und der **Median** stimmen überein
- Die **Dichtefunktion** *f(x)* ist für **jede** reelle Zahl definiert und **immer** größer als 0.
- Für $x \rightarrow \pm\infty$ nähert sie sich **asymptotisch** der x-Achse.

N(0,1), die Normalverteilung mit **Erwartungswert 0** und **Varianz 1**, nennt man Standardnormalverteilung.

Abbildung 3.12: Dichtefunktion der Standardnormalverteilung *N(0,1)*



3.4.2 Standardisierung

Mit der **Standardisierungsformel**

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

transformiert man eine normalverteilte Zufallsvariable $X{:}N(\mu,\sigma^2)$ in eine Standardnormalverteilung. Die Verteilungsfunktion der Standardnormalverteilung wird allgemein mit Φ bezeichnet. Mit Hilfe einer **Tabelle von Φ** (Tabelle 3.3) kann man den Wert der Verteilungsfunktion F jeder beliebigen Normalverteilung über die **Formel**

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

berechnen. Die Quantile z_p der **Standardnormalverteilung** bzw. x_p einer **beliebigen Normalverteilung** kann man mit Hilfe der Formeln

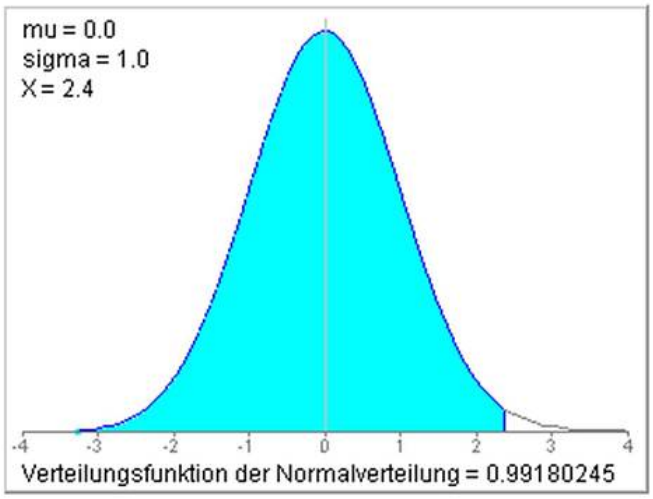
$$z_p = \frac{x_p - \mu}{\sigma} \quad \text{bzw.} \quad x_p = \mu + \sigma \cdot z_p$$

umrechnen.

Tabelle 3.3: Verteilungsfunktion und Quantile der Standardnormalverteilung

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

p =	**0.5000**	**0.7500**	**0.8000**	**0.9000**	**0.9500**	**0.9750**	**0.9900**	**0.9950**	**0.9975**	**0.9990**
z_p =	0.0000	0.6745	0.8416	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902



Verteilungsfunktion der Standardnormalverteilung

Neustart | **z** 2.4000 | Berechne F(z)

F(z) 0.991802 **1-F(z)** 0.008198

Beispiel 3.12

Das Geburtsgewicht von Neugeborenen nach unauffälliger Schwangerschaft sei mit Erwartungswert μ = 3500 g und Standardabweichung σ = 500 g normalverteilt.

Die Wahrscheinlichkeit, dass ein Neugeborenes aus dieser Grundgesamtheit nicht mehr als 4700 g wiegt, ist dann

$$F(4700) = \Phi\left(\frac{4700 - 3500}{500}\right) = \Phi(2.4) = 0.9918$$

D. h. die Wahrscheinlichkeit ist 0.9918, in der genannten Grundgesamtheit wiegen damit 99.18 % aller Neugeborenen nicht mehr als 4700 g.

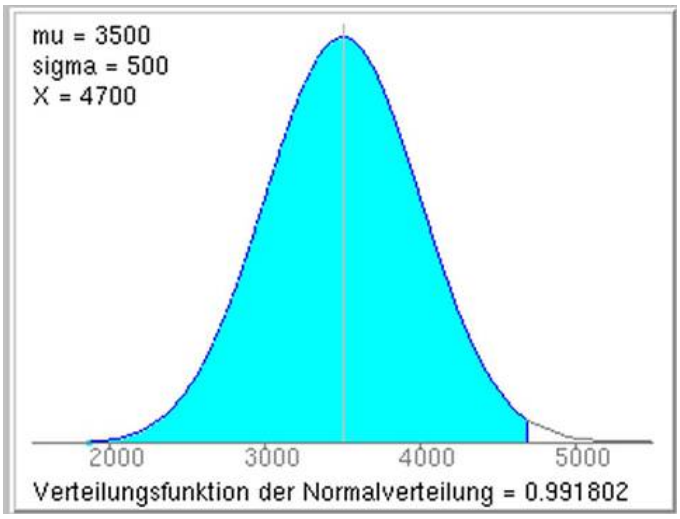
Normalverteilung - Standardisierung und Verteilungsfunktion

Neustart

x 4700 μ : 3500 σ : 500

Berechne

z 2.4000 F(z) 0.991802 1-F(z) 0.008198

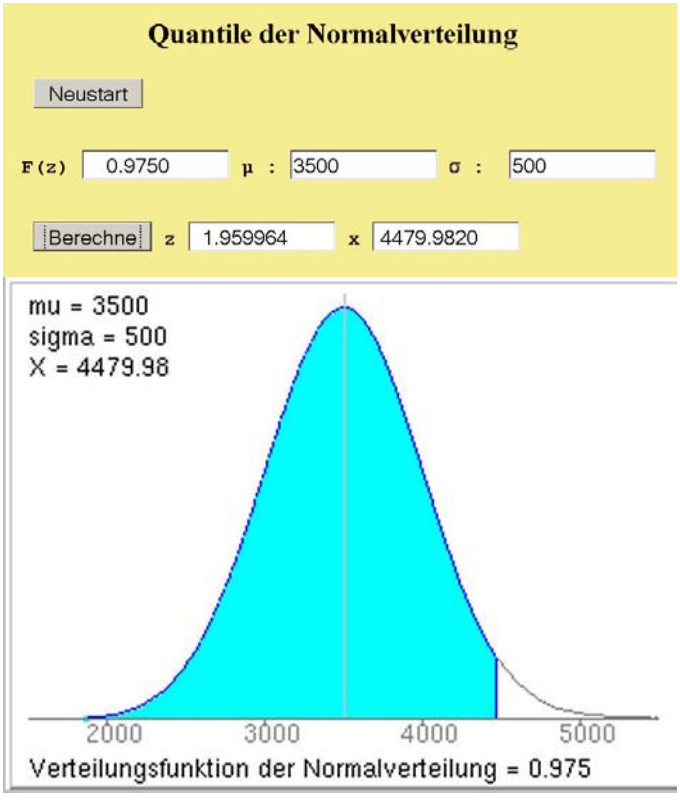


Es soll eine Grenze für das Geburtsgewicht angegeben werden, die nur vom 2.5 % aller Neugeborenen übertroffen wird.

Es gilt:

$$x_{0.975} = \mu + \sigma \cdot z_{0.975} = 3500 + 500 \cdot 1.96 = 4480.$$

D. h. die gesuchte Schranke beträgt 4480 g. In der genannten Grundgesamtheit wiegen also 97.5 % aller Neugeborenen nicht mehr als 4480 g. 2.5 % wiegen mehr als 4480 g.



Die Normalverteilung wird häufig verwendet, um quantitative**, symmetrisch verteilte, eingipflige** Merkmale zu beschreiben. Zum Standardisieren einer Normalverteilung benötigt

man deren Erwartungswert und Varianz bzw. Standardabweichung. In der Praxis sind diese häufig nicht bekannt, und man muss sie aus einer Stichprobe durch den **arithmetischen** Mittelwert und die **empirische Varianz bzw. Standardabweichung** schätzen. Diese Schätzung aus empirischen Daten ist aber nur dann sinnvoll, wenn es sich um **quantitative, symmetrisch verteilte, eingipflige Merkmale** handelt. Ein erster Hinweis auf **Symmetrie** liegt dann vor, wenn der Median und der Mittelwert annähernd **gleich** sind. Eine **optische** Überprüfung ist durch ein Histogramm mit einer angepassten Normalverteilungsdichte möglich. Die **sigmoide** Form der empirischen Verteilungsfunktion ist ebenfalls ein Hinweis auf annähernd normalverteilte Merkmale. Noch geeigneter ist der sogenannte **Normalverteilungsplot**, wo mit Hilfe der Normalverteilung die **empirische Verteilungsfunktion** so **transformiert** wird, dass bei **normalverteilten** Merkmalen eine **Gerade** entsteht. Diese Möglichkeiten veranschaulicht das folgende Beispiel.

Beispiel 3.13
Mit Hilfe des Applets- Exploration und Tests erhält man mit dem Datensatz "AML92_99" für das Merkmal "Körpergröße" die folgende Basisstatistik.

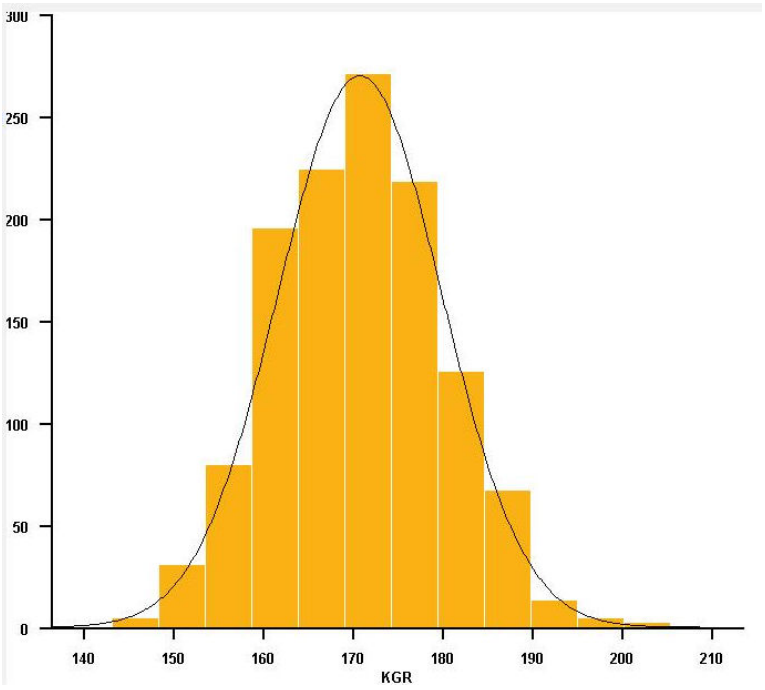
Abbildung 3.13: Basisstatistik des Merkmals "*Körpergröße*"

Basisstatistik der Variablen

	Mittelwert	Varianz	Standardabw.	Minimum	1. Quartil	Median	3. Quartil	Maximum	n
KGR	170.8027	82.80184	9.099551	146	164.5	170.5	177	203	1232

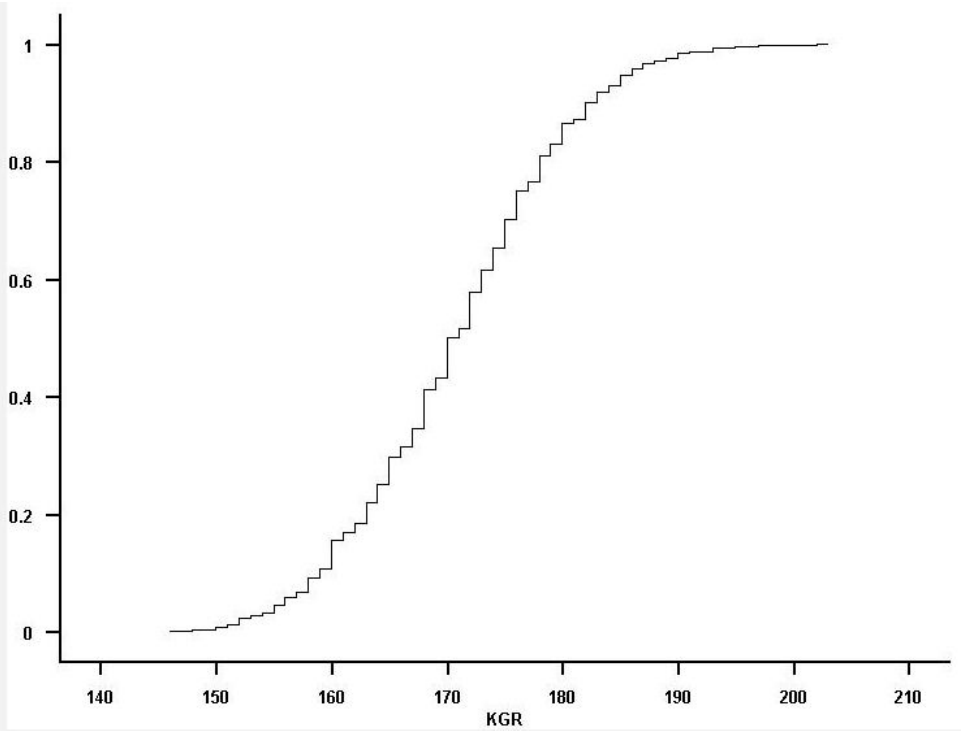
Median und Mittelwert sind annähernd gleich. Auch das Histogramm zeigt eine gute Übereinstimmung mit der angepassten Normalverteilungsdichte.

Abbildung 3.14: Histogramm mit angepasster Normalverteilungsdichte des Merkmals "*Körpergröße*"



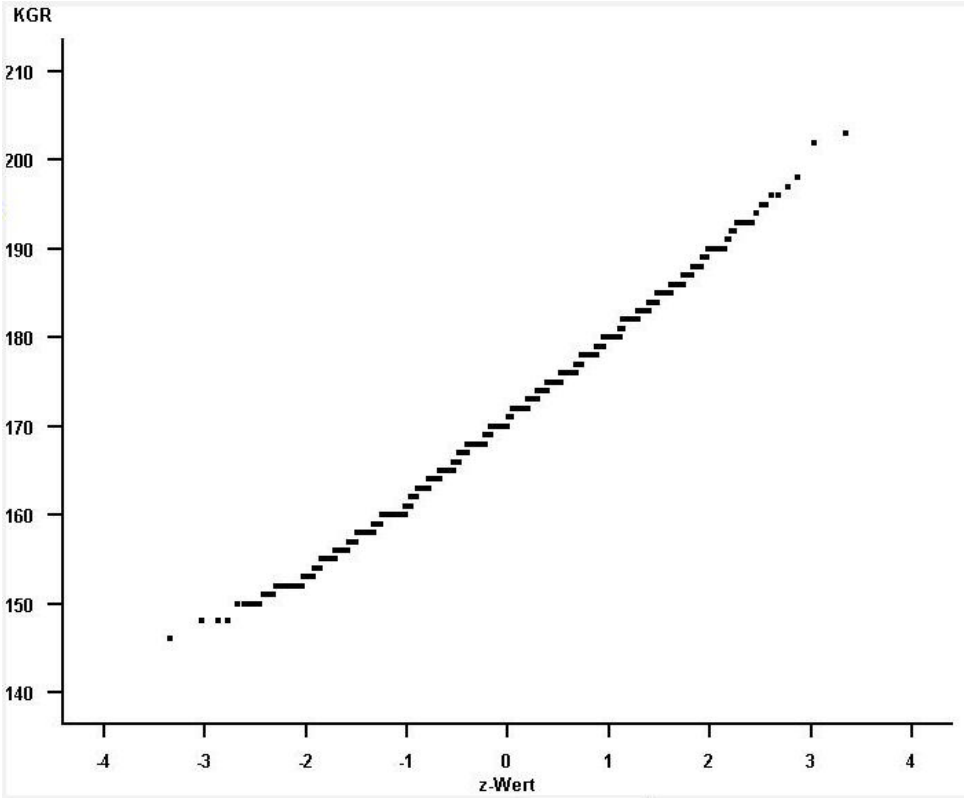
Die empirische Verteilungsfunktion zeigt ebenfalls die für die Normalverteilung typische sigmoide Form.

Abbildung 3.15: Empirische Verteilungsfunktion des Merkmals "*Körpergröße*"



Der Normalverteilungsplot gibt annähernd eine Gerade wieder. Das Merkmal "Größe" kann als normalverteilt mit Erwartungswert 170 cm und Standardabweichung 9.4 cm angenommen werden.

Abbildung 3.16: Normalverteilungsplot des Merkmals "*Körpergröße*"



Leider sind in der Medizin viele Merkmale **nicht normalverteilt** sondern **rechtsschief** verteilt; d. h. die Dichtefunktion ist **nicht symmetrisch** sondern hat einen **Gipfel** am **linken Rand** und einen langen **Auslauf** an der **rechten Seite**. Ein erster Hinweis auf **Nichtsymmetrie** liegt dann vor, wenn der **Median** und der **Mittelwert** stark **unterschiedlich** sind. Eine **optische** Überprüfung ist durch ein **Histogramm** mit einer **schlecht angepassten Normalverteilungsdichte** möglich. Die **sigmoide** Form der **empirischen Verteilungsfunktion** liegt bei solchen nichtnormalverteilten Merkmalen nicht vor. Noch geeigneter ist der **Normalverteilungsplot**, der **keine Gerade** zeigt. Viele solcher Merkmale folgen einer **logarithmischen Normalverteilung**, d. h. ihre logarithmierten Werte sind normalverteilt. Durch eine **Logarithmustransformation** erhält man dann ein normalverteiltes Merkmal. Diese Möglichkeiten veranschaulicht das folgende Beispiel.

Beispiel 3.14
Mit Hilfe des Applets- Exploration und Tests erhält man mit dem sonstigen Datensatz "aml1.html" für das Merkmal "Lactatdehydrogenese (LDH)" und "log (LDH)" die folgende Basisstatistik

Abbildung 3.17: Basisstatistik des Merkmals "*LDH und log(LDH)*"

Basisstatistik der Variablen

	Mittelwert	Varianz	Standardabw.	Minimum	1. Quartil	Median	3. Quartil	Maximum	n
LDH_0	636.0969	618607.4	786.5160	88	265.5	432	729.5	14332	1248
log(LDH_0)	6.135366	0.537677	0.733264	4.477336	5.581613	6.068425	6.592357	9.570250	1248

Median und Mittelwert unterscheiden sich bei "LDH" stark, sind aber annähernd gleich bei "log(LDH)". Auch das Histogramm zeigt beim "LDH" eine schlechte Übereinstimmung mit der angepassten Normalverteilungsdichte und eine gute Übereinstimmung bei den logarithmierten Werten.

Abbildung 3.18: Histogramm mit angepasster Normalverteilungsdichte des Merkmals "*LDH*"

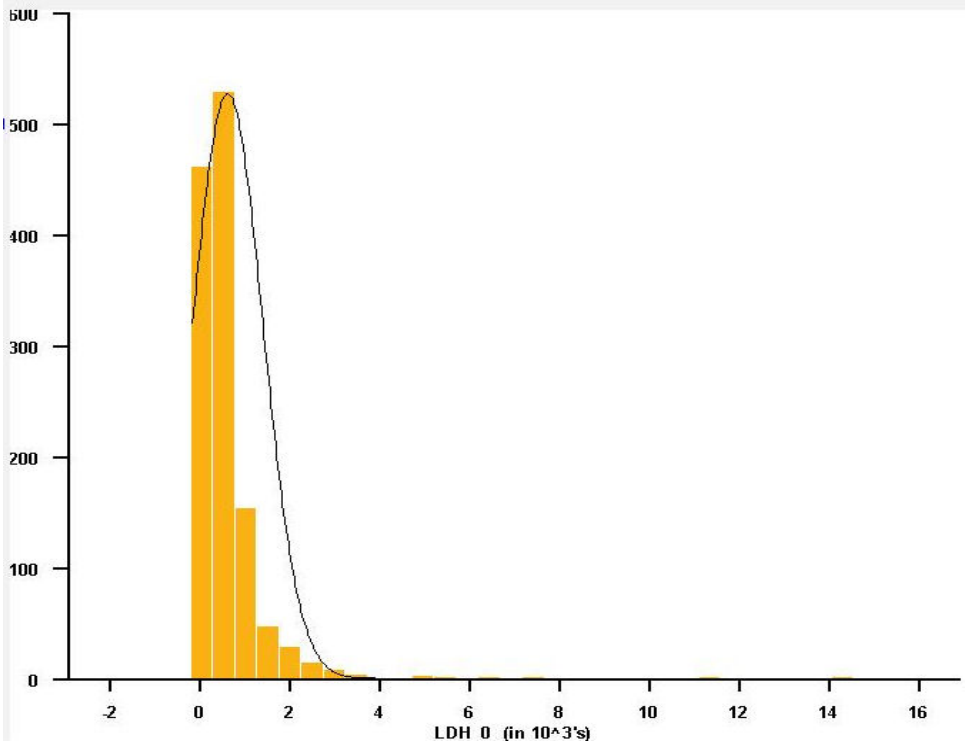
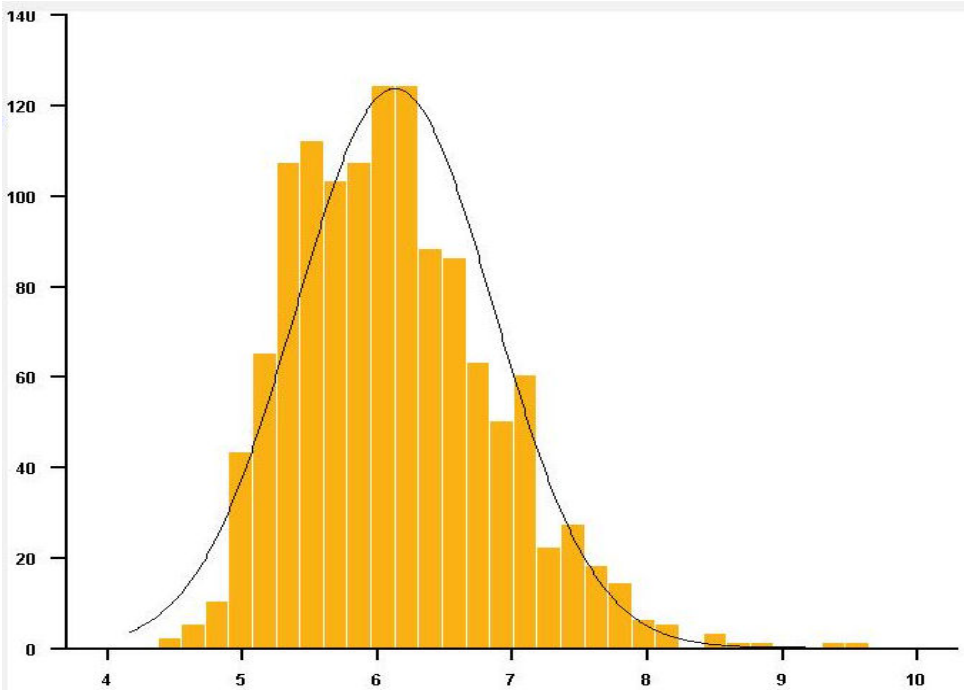


Abbildung 3.19: Histogramm mit angepasster Normalverteilungsdichte des Merkmals "*log(LDH)*"



Die empirische Verteilungsfunktion zeigt für "LDH" ebenfalls nicht die für die Normalverteilung typische sigmoide Form. Bei "log(LDH)" ist diese Form zu sehen.

Abbildung 3.20: Empirische Verteilungsfunktion des Merkmals "*LDH*"

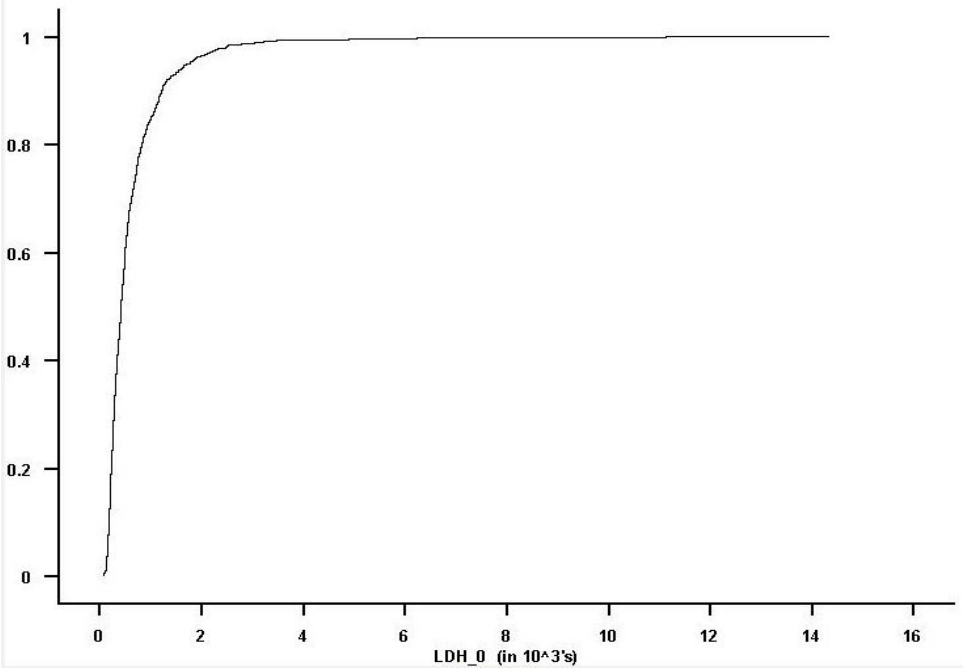
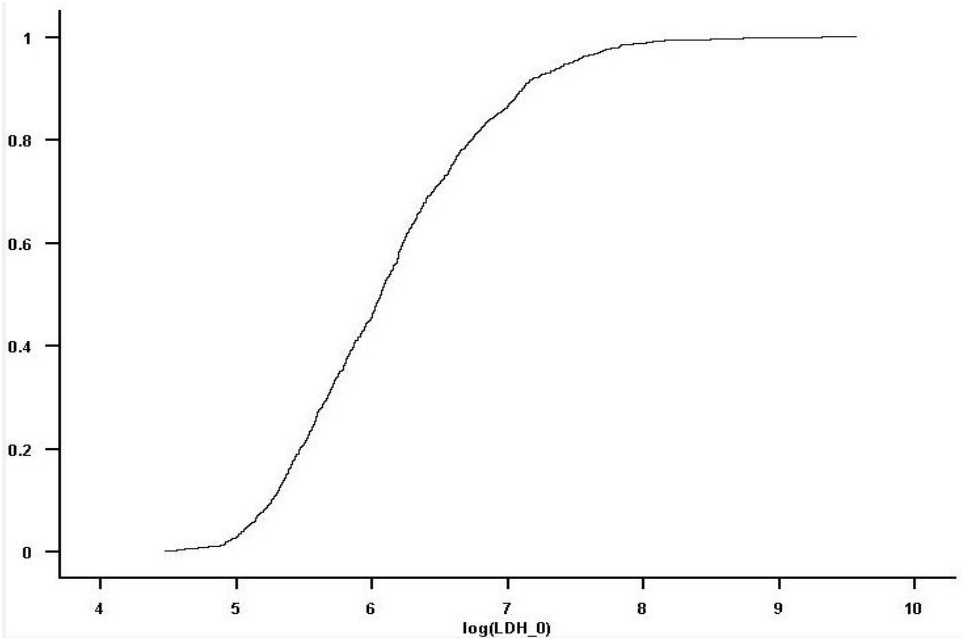


Abbildung 3.21: Empirische Verteilungsfunktion des Merkmals "*log(LDH)*"



Der Normalverteilungsplot für "LDH" weicht stark von einer Geraden ab, gibt aber annähernd für das "log(LDH)" eine Gerade wieder.

Abbildung 3.22: Normalverteilungsplot des Merkmals "*LDH*"

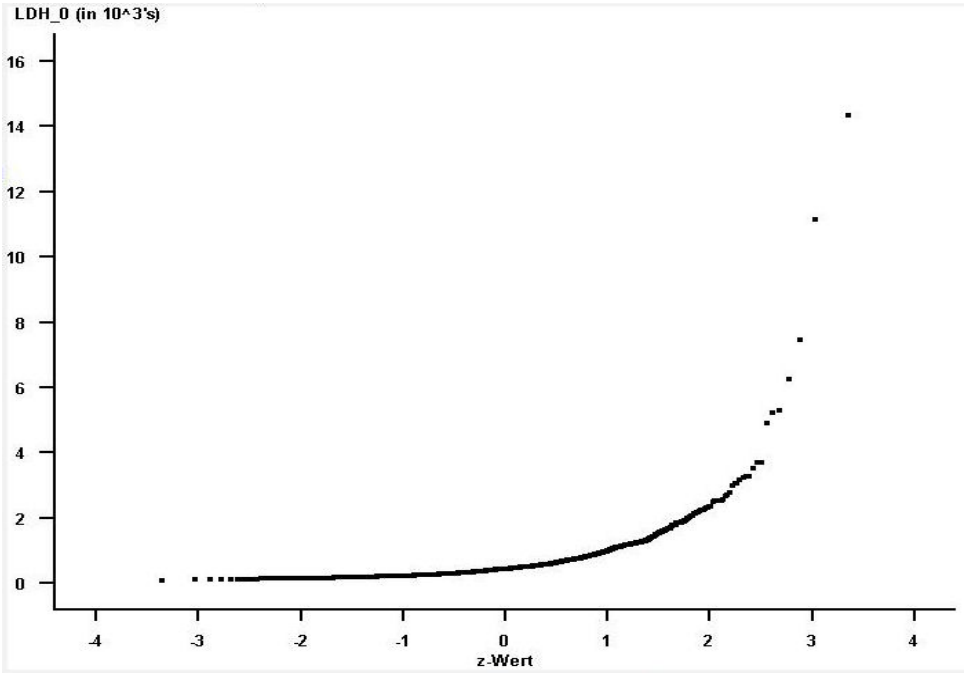
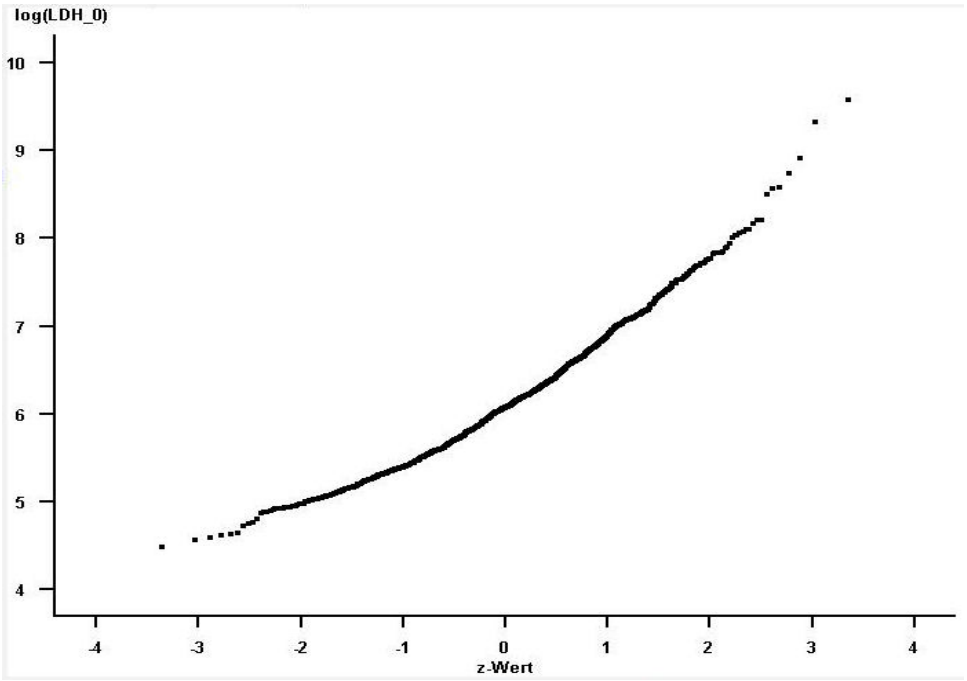


Abbildung 3.23: Normalverteilungsplot des Merkmals "*log(LDH)*"



Arithmetischer Mittelwert und empirische Standardabweichung sind die **Schätzwerte** für die **Standardisierung**. Die **Subtraktion** des Mittelwertes bei der **Standardisierung** ist **unproblematisch**, man erhält eine Normalverteilung mit Erwartungswert 0. Beim **Dividieren**

durch die empirische Standardabweichung ergibt sich aber das **Problem**, dass die Verteilung des Quotienten **keine** Normalverteilung mehr ist. **W. Gosset** hat 1903 die resultierende Verteilung berechnet und ihr den Namen *t*-Verteilung gegeben. Er hat gezeigt, dass ihre **Dichtefunktion** der Gleichung

$$f(t) = c_{n-1} \cdot \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} \qquad (-\infty < t < +\infty)$$

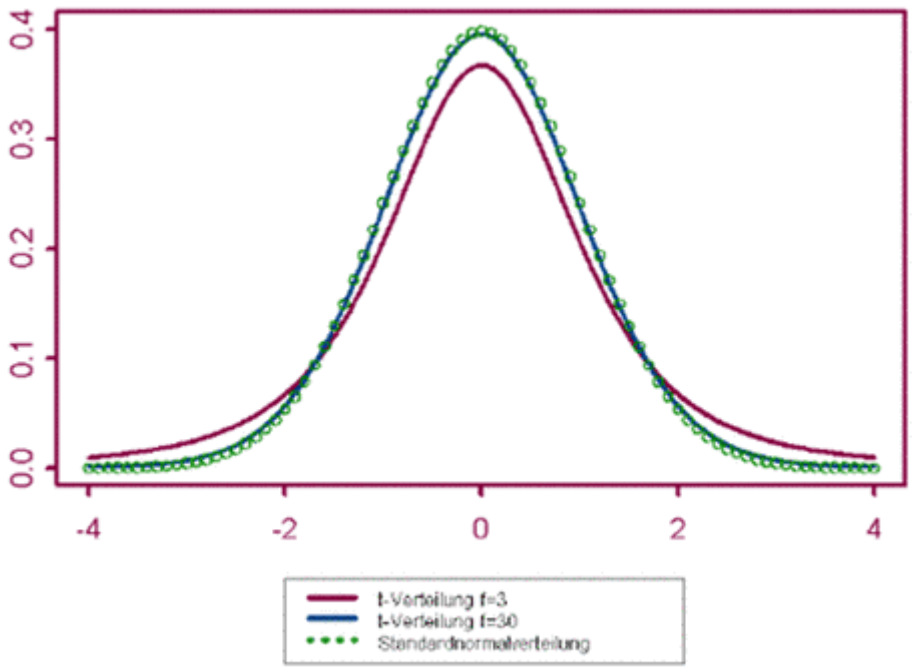
genügt. Hierin ist c_{n-1} eine **Konstante**, die sich aus der Gleichung

$$\int_{-\infty}^{+\infty} f(t)dt = 1$$

bestimmen lässt. Der **Graph von *f*** ähnelt dem der Dichte der **Standardnormalverteilung**. *f* hat sein Maximum bei *t*=0 und nähert sich symmetrisch zur *y*-Achse asymptotisch der *t*-Achse. Die Form der Verteilung hängt noch vom **Umfang *n*** der Stichprobe ab, aus der die empirische Standardabweichung berechnet wurde. Je größer *n* ist, desto mehr nähert sich die *t*-Verteilung der Standardnormalverteilung an.

Historisch hat es sich eingebürgert, die verschiedenen t-Verteilungen nicht mit n sondern mit ***f*=*n*-1**, der sogenannten Zahl der **Freiheitsgrade** (engl. degrees of freedom (*df*)) durchzunummerieren.

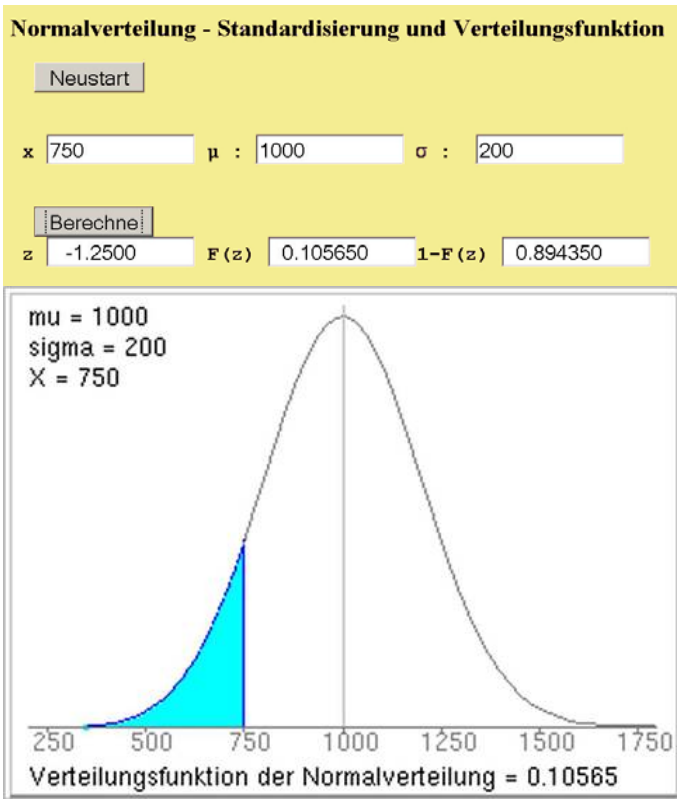
Abbildung 3.24: Dichtefunktion der t-Verteilung (*f*=3 und *f*=30) und der Standardnormalverteilung



Die ***t*-Verteilung** braucht man insbesondere dann, wenn man Hypothesen über den **Erwartungswert einer Normalverteilung** prüfen will, deren Standardabweichung **nicht** bekannt ist (***t*-Test, Kapitel 4**).

Beispiel 3.15
Eine Klinikapotheke benötigt täglich im Durchschnitt etwa 1000 g einer bestimmten Substanz X. Angenommen, der tägliche Verbrauch sei mit Erwartungswert μ = 1000 g und Standardabweichung σ = 200 g normalverteilt.

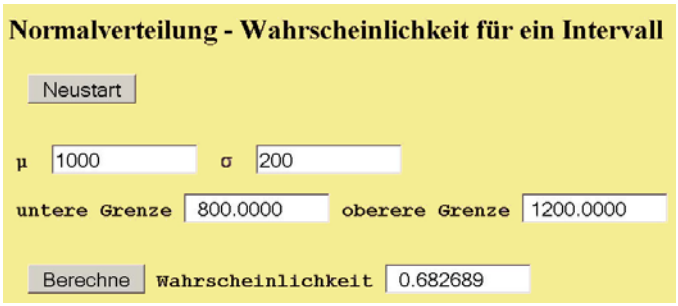
Wie groß ist die Wahrscheinlichkeit, dass an einem Tag weniger als 750 g benötigt werden?



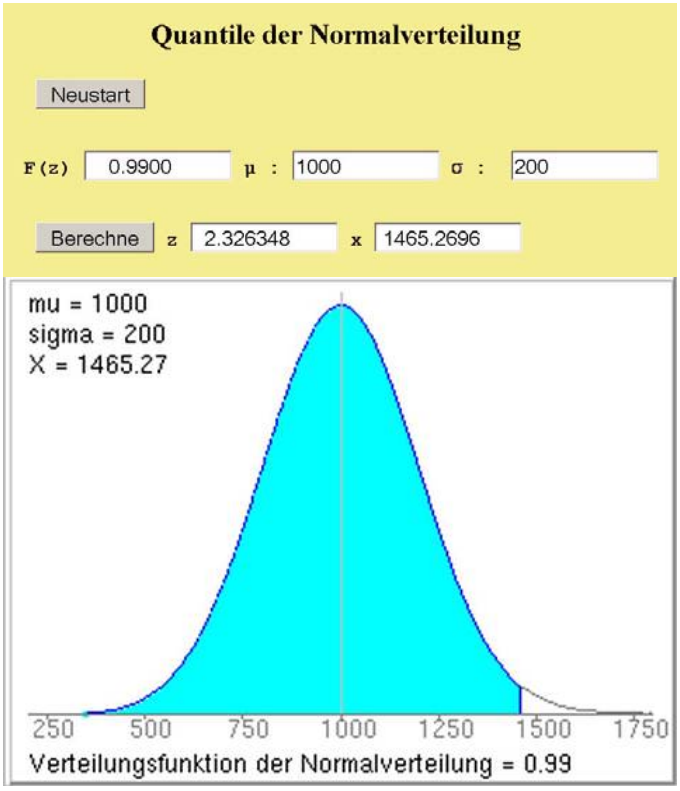
Wie groß ist die Wahrscheinlichkeit, dass der Bedarf an einem Tag

a) zwischen 800 und 1200 g
b) zwischen 600 und 1400 g
c) zwischen 400 und 1600 g

liegt?



Wie groß muss der Vorrat der Apotheke mindestens sein, damit der tägliche Bedarf ohne Nachbestellung mit 99 % (99.9 %) Sicherheit gedeckt werden kann?



3.4.3 Zentraler Grenzwertsatz

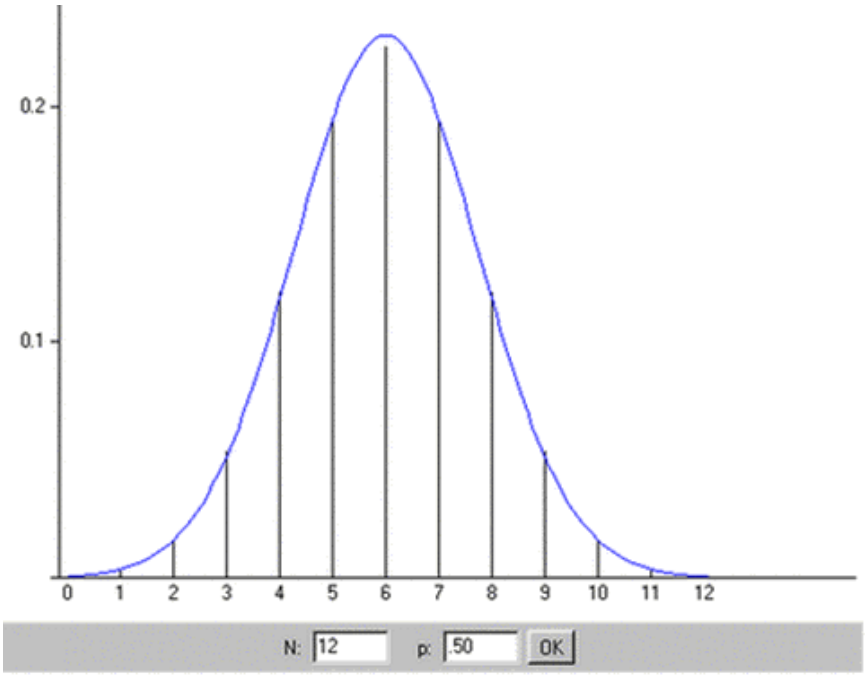
Der zentrale Grenzwertsatz besagt, dass die **Summe unabhängiger Zufallsvariablen**, die alle die **gleiche** Verteilungsfunktion besitzen, näherungsweise **normalverteilt** ist. Die Annäherung ist umso besser, je größer die Anzahl der Summanden ist.

Beispiel: Eine binomialverteilte Zufallsvariable X ist eine Summe von n unabhängigen bernoulliverteilten Zufallsvariablen $Y_1, Y_2, Y_3, ..., Y_n$:

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

Nach dem **Zentralen Grenzwertsatz** lässt sich die **Binomialverteilung** mit dem Erwartungswert np und der Varianz $np(1-p)$ näherungsweise durch die entsprechende **Normalverteilung** mit dem Erwartungswert np und der Varianz $np(1-p)$ ersetzen.

Abbildung 3.25: Anpassung der Binomialverteilung durch die Normalverteilung

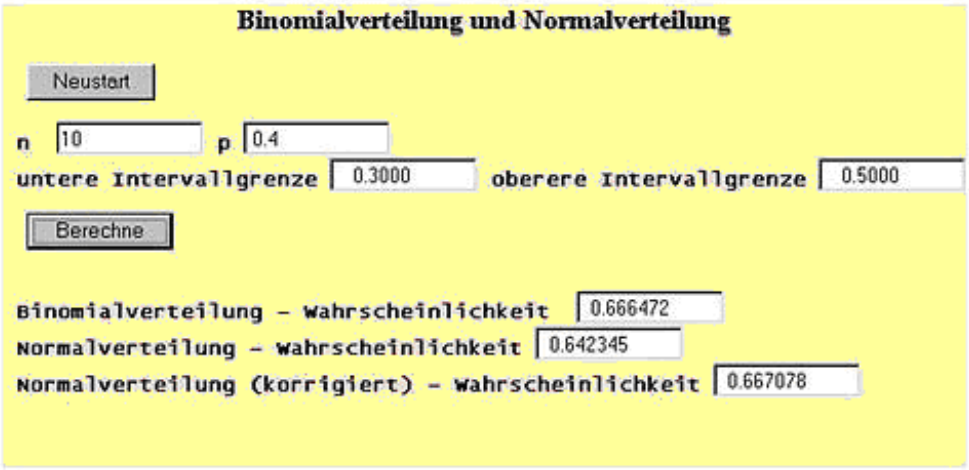


An einer Skizze kann man sich klarmachen, dass man die **Wahrscheinlichkeit** $P(k_1 \le X \le k_2)$ der Binomialverteilung **nicht** durch $F(k_2) - F(k_1 - 1)$ der entsprechenden Normalverteilung, sondern **besser** durch $F(k_2 + {}^1/_2) - F(k_1 - {}^1/_2)$ approximiert. Diese Korrektur nennt man **Stetigkeitskorrektur**.

Beispiel 3.16
In einer Grundgesamtheit haben 40 % aller Personen die Blutgruppe 0. Wie groß ist die Wahrscheinlichkeit, dass in einer zufälligen Stichprobe vom Umfang n=10, 50, 100 aus dieser Grundgesamtheit der Anteil der Personen mit Blutgruppe 0 zwischen 30 % und 50 % liegt? Die folgende Tabelle enthält die gefragten Wahrscheinlichkeiten sowohl über die Binomialverteilung als auch näherungsweise über die entsprechende Normalverteilung mit und ohne Stetigkeitskorrektur.

Tabelle 3.4: Approximation der Binomialverteilung durch die Normalverteilung

n	Binomialverteilung	Normalverteilung	Normalverteilung (korrigiert)
10	0.66647	0.64234	0.66708
50	0.88870	0.88391	0.88765
100	0.96846	0.96701	0.96791



3.4.4 Konfidenzintervall

Der **unbekannte Erwartungswert μ** einer Normalverteilung $N(\mu,\sigma^2)$ wird durch den **Mittelwert** aus einer zufälligen Stichprobe geschätzt. Zu dem Mittelwert lässt sich ein **Intervall**, das sogenannte **Konfidenzintervall** angeben, das den unbekannten Erwartungswert μ mit einer vorgegebenen **Konfidenzwahrscheinlichkeit 1-α** enthält. Die **Intervallgrenzen** t_u bzw. t_o berechnet man aus den Formeln

$$t_u = \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2} \quad ,$$

$$t_o = \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha/2} \quad .$$

Dabei ist σ die Standardabweichung der betrachteten Normalverteilung, n ist der Stichprobenumfang und $z_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Standardnormalverteilung.

Wenn die Standardabweichung σ nicht bekannt ist, muss sie ebenfalls aus der Stichprobe geschätzt werden. Als Schätzwert benutzt man die empirische Standardabweichung s. In den Formeln für die Intervallgrenzen muss dann aber auch das Quantil $z_{1-\alpha/2}$ der Standardnormalverteilung durch das **Quantil** $t_{n-1;1-\alpha/2}$ der t_{n-1}**-Verteilung** ersetzt werden. Man erhält

$$t_u = \bar{x} - \frac{s}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2} \quad ,$$

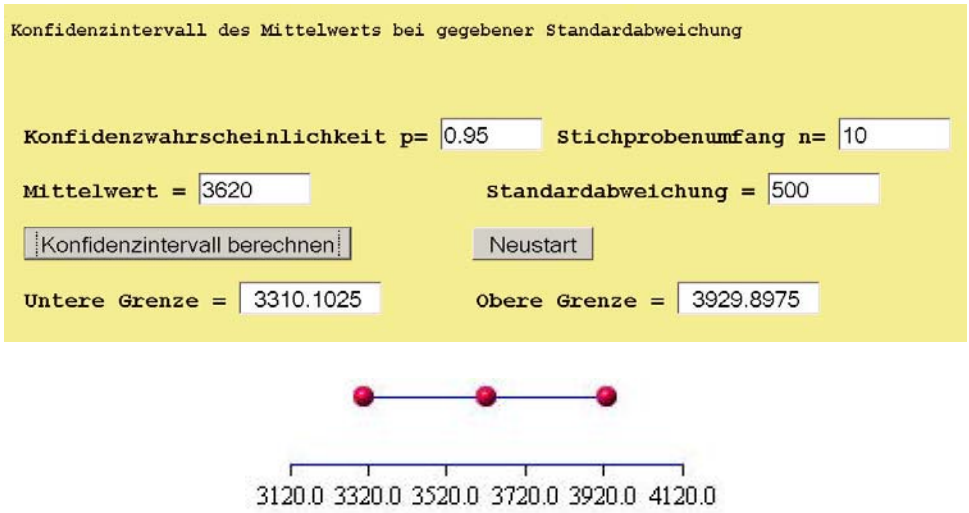
$$t_o = \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{n-1;1-\alpha/2} \quad .$$

Beispiel 3.17
Es wird vorausgesetzt, dass das Körpergewicht von Neugeborenen nach unauffälliger Schwangerschaft und unter Ausschluss von Mehrlingsgeburten einer Normalverteilung $N(\mu,\sigma^2)$ folgt.

Geht man von der Standardabweichung σ= 500 g aus und wählt die Konfidenzwahrscheinlichkeit 1-α= 0.95 (d. h. Irrtumswahrscheinlichkeit α= 0.05), dann ergeben sich die in Tabelle 3.5 wiedergegebenen zweiseitigen Konfidenzintervalle für den unbekannten Erwartungswert μ.

Tabelle 3.5: Konfidenzintervall bei gegebener Standardabweichung

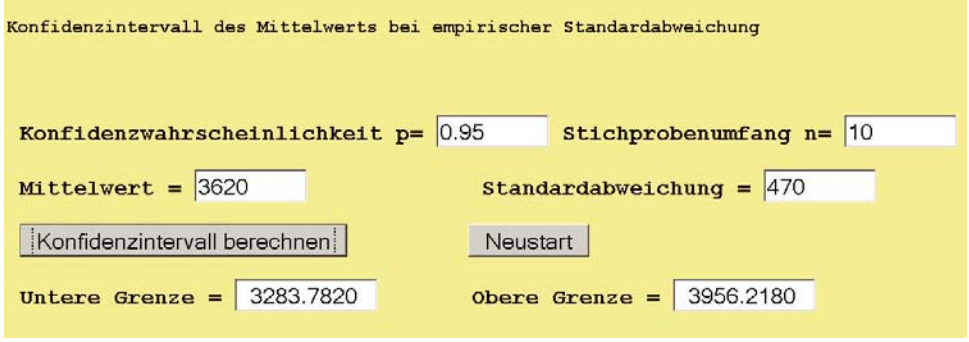
Stichprobenumfang	Mittelwert	untere Grenze	Obere Grenze	Intervalllänge
10	3620	3310.1	3929.9	619.8
20	3490	3270.9	3709.1	438.2
40	3570	3415.1	3724.9	309.8



Wird die Standardabweichung wie angegeben aus der Stichprobe geschätzt, so muss man statt der Quantile der Standardnormalverteilung die Quantile der entsprechenden t-Verteilung benutzen und erhält die Ergebnisse in Tabelle 3.6. Die benötigten Quantilwerte der t-Verteilung sind in Tabelle 3.7 enthalten.

Tabelle 3.6: Konfidenzintervall bei empirischer Standardabweichung (a= 0.05)

Stichprobenumfang	Mittelwert	emp. Standardabw.	untere Grenze	Obere Grenze	Intervalllänge
10	3620	470	3283.8	3956.2	672.4
20	3490	560	3227.9	3752.1	524.2
40	3570	510	3406.9	3733.1	326.2



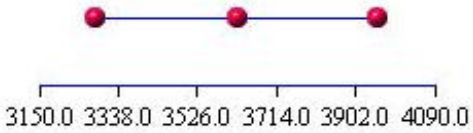


Tabelle 3.7: Ausgewählte Quantile der t_f-Verteilung

f	9	19	39	∞
$t_{f;0.975}$	2.262	2.093	2.023	1.96