Manuskript

Mathematik III

Statistik und

Numerische Mathematik

Wirtschaftsingenieurwesen

DHBW Stuttgart

Campus Horb

Dozent

Dipl. Math. (FH) Roland Geiger

Inhaltsverzeichnis

Bisektions- bzw. Intervallhalbierungsverfahren **Fehler! Textmarke nicht definiert.**

Grundlagen

Beschreibende (Deskriptive) Statistik

Die beschreibende Statistik beschäftigt sich mit Methoden, die darauf zielen, bestimmte Aspekte der in den Daten enthaltenen Information möglichst prägnant wiederzugeben (durch Tabellen, Grafiken, Kennzahlen).

Statistiken sollen im besten Fall Phänomene aufdecken und erklären. Um Statistiken selbst besser zu verstehen, bedarf es einer Handvoll relevanter Begriffe, mit denen man sich vertraut machen sollte.

Ich stelle Ihnen hier die wichtigsten Begriffe vor die sie als Basis für das bessere Verständnis der Statistik brauchen.

Grundbegriffe

Grundgesamtheit

In der empirischen Forschung bezeichnet die Grundgesamtheit (auch Population, Zielpopulation oder target population) die Menge aller potentiellen Untersuchungsobjekte für eine bestimmte Fragestellung.

Definition 1:

Grundgesamtheit heißt die Menge der Merkmalsträger, über die eine Aussage getroffen werden soll, z.B. Tiere einer Herde, Menschen einer Region oder Stadt. Sie muss bei jeder Datenerhebung genau definiert werden.

Bemerkung 1:

- Die Grundgesamtheit kann aus einer endlichen Menge von Elementen bestehen, oder sie kann unendlich groß sein.
- Die Grundgesamtheit ist die Menge aller interessierender Daten
- Die Anzahl Elemente dieser Menge nennt man den Umfang der Grundgesamtheit
- Der Umfang kann endlich oder unendlich sein.

Empirische Forschung

Definition 2:

Empirische Forschung wissenschaftliche Methodik, welche Aussagen über die Realität durch Befragung, Beobachtung und Messung gewinnt.

Stichprobe

Aus pragmatischen Erwägungen wird normalerweise nicht die Grundgesamtheit, sondern eine Stichprobe untersucht, die für die Grundgesamtheit repräsentativ ist.

Definition 3:

Als Stichprobe bezeichnet man eine Teilmenge einer Grundgesamtheit, die unter bestimmten Gesichtspunkten ausgewählt wurde.

Bemerkung 2:

- Eine gesamte Untersuchung ist in der Regel nicht möglich, man wertet repräsentative Teilauswahlen oder Stichproben aus.
- Eine Möglichkeit, eine repräsentative Teilauswahl zu bekommen, ist die Zufallsstichprobe, in die jedes Element der Grundgesamtheit mit der gleichen Wahrscheinlichkeit aufgenommen wird.

Repräsentativ

Um die einzelnen Elemente einer Stichprobe zu erhalten, stehen verschiedene Auswahlverfahren zur Verfügung.

Die korrekte Wahl des Auswahlverfahrens ist wichtig, da die Stichprobe repräsentativ sein muss, um auf die Grundgesamtheit schließen zu können (siehe dazu z. B. Hochrechnung). Entscheidend ist eine vernünftige Probenahme, die über den Erfolg der Aussage entscheidet.

Definition 4:

Von Repräsentativität wird gesprochen, wenn sich aus einer Stichprobe zutreffende Rückschlüsse auf eine Grundgesamtheit ziehen lassen

Merkmal, Merkmalsträger und Merkmalsausprägungen

Wenn von einer statistischen Erhebung die Rede ist, so denken wir zunächst an das Befragen von Personen oder an das Zählen von Gegenständen.

Es braucht sich dabei jedoch nicht unbedingt um ein Befragen oder Zählen zu handeln, es kann sich auch um ein messen handeln

Wir sagen allgemein: Das Ergebnis wird durch Beobachten gefunden.

Die Beobachtung richtet sich auf ein bestimmtes Merkmal, das bei allen Objekten der Grundgesamtheit vorhanden ist; z. B. sind Alter, Geschlecht, Familienstand, Körpergröße, Blutgruppe, Zahl der Kinder, Monatseinkommen u. a. Merkmale von Personen.

Definition 5:

Merkmale sind jene Eigenschaften, die in einer Erhebung untersucht werden. Bei einer Befragung entspricht ein Merkmal einer gestellten Frage.

Definition 6:

Merkmale können verschiedene Werte annehmen, die Merkmalsausprägungen genannt werden. Bei Befragungen sind die Merkmalsausprägungen die Antwortmöglichkeiten, die der Befragte angeben kann.

Definition 7:

Als Merkmalsträger oder auch statistische Einheit bezeichnet man die untersuchten Einzelobjekte einer Erhebung.

Merkmalsträger sind zum Beispiel Personen, Produkte usw.

Bemerkung 3:

Auch die Ausprägungen eines Merkmals sind nicht zwangsläufig mit dem Merkmal gegeben, sondern müssen von uns - dem Ziel der statistischen Erhebung entsprechend - festgesetzt werden. Dies muss so geschehen, dass bei jeder Beobachtung klar ist, welche der vorgesehenen Ausprägungen vorliegt; die Liste der Ausprägungen muss also jeden möglicherweise auftretenden Fall enthalten und je zwei Ausprägungen müssen unterscheidbar und unvereinbar sein.

Beispiel 1:

Durch eine statistische Erhebung soll festgestellt werden, wie die Arbeitnehmer einer Stadt zu ihrer Arbeitsstätte gelangen.

Ist die Erhebung in Auftrag gegeben worden, um die Belastung der öffentlichen Verkehrsmittel generell zu untersuchen, so genügen die beiden Ausprägungen „mit öffentlichen Verkehrsmitteln; ohne öffentliche Verkehrsmittel".

Will man feststellen wie die Verkehrswege belastet werden, so kann man etwa die Liste „zu Fuß; Zweirad; Pkw; Bus; Straßenbahn; U-Bahn; Vorortzug; andere Verkehrsmittel" verwenden.

Beachten Sie: Ohne die zuletzt genannte Ausprägung wäre die Liste evtl. unvollständig (z. B. wenn Schifffahrtswege vorhanden sind); würde umgekehrt zusätzlich eine Ausprägung „Fahrrad" aufgenommen, so wäre zwischen „Fahrrad" und „Zweirad" keine eindeutige Entscheidung mehr möglich.

Definition 8:

Die einer statistischen Erhebung zugrunde liegende Menge von Merkmalsausprägungen wird mit S, ihre Elemente werden mit $a_1, \ldots, a_k$ bezeichnet. Es ist also

$S = \{a_1, \ldots, a_k\}$.

Skalenniveau

Nennen Sie Merkmale, die zahlenmäßige und solche, die keine zahlenmäßigen Ausprägungen haben. Gibt es Unterschiede in der Art und Weise wie die jeweiligen Ausprägungen festgestellt werden?

Wir vergleichen einige Merkmale:

Merkmal	**Merkmalsausprägungen**
a) Geschlecht	männlich, weiblich
b) Schulische Leistung	sehr gut, . . ., ungenügend
c) Geschwisterzahl	0, 1, 2, 3, . . .

In der Statistik werden je nach Art der erhobenen Daten der Merkmale verschiedene „Messlatten“ bzw. Skalen verwendet. Nicht jedes Merkmal lässt sich gleich gut in Zahlen darstellen. Während dies für die Körpergröße in Zentimetern sehr einfach ist, ist es für das Geschlecht gar nicht möglich, für die persönliche Zufriedenheit machbar aber schwierig.

Das Skalenniveau drückt aus, wie quantitativ ein Antwortwert ist, das heißt, inwieweit sinnvolle Rechenoperationen angewendet werden können.

Definition 9:

Variable sind oft nicht nur Zahlen, sie können auch Attribute einschließen. Daraus ergibt sich eine unterschiedliche Art der Skalierung, wobei vier Arten von Skalen unterschieden werden können:

Nominalskala (nicht-metrisch bzw. kategorial)

Ordinalskala (nicht-metrisch bzw. kategorial)

Intervallskala (metrisch)

Ratioskala/Verhältnisskala (metrisch)

Die Nominalskala bietet den geringsten statistischen Informationsgehalt, die Ratioskala den höchsten. Nominal- und Ordinalskala sind nicht-metrische bzw. kategoriale Skalen, das heißt, ihre Antwortwerte stehen nicht für einen direkt verwendbaren Zahlenwert. Intervall- und Ratioskala sind metrische Skalen, die verschiedene Rechenoperationen erlauben.

Nominalskala

Diese Skala basiert auf einem Satz von Attributen. Es existiert kein Kriterium, nach dem die Punkte einer nominal skalierten Variablen anzuordnen sind.

Beispiele: Tierarten, Geschlecht, die Nummern auf den Dressen der Fußballspieler.

Ordinalskala

Diese Skala bezieht sich auf Messungen, die in Termen wie "größer", "kleiner" oder "gleich" angeordnet werden können. Die Beobachtungen müssen nicht im gleichen Abstand erfolgen.

Beispiele: prozentuale Ränge, Reihenfolge der besten Rennläufer.

Intervallskala (metrisch)

Gleich unterteilte Einheiten entlang der Skala, ohne einen vordefinierten Nullpunkt.

Beispiele: Temperatur (in C, F oder R), Wasserpegel eines Flusses.

Ratioskala/Verhältnisskala (metrische)

Gleich unterteilte Einheiten entlang einer Skala, mit einem wahren Nullpunkt.

Beispiele: Temperatur in K, Gewicht, Geschwindigkeit

Definition 10:

Sind die Merkmalsausprägungen numerisch angegeben, so ist jeweils zu prüfen, ob es sich um eine Nominal-, eine Ordinal- oder eine metrische Skala handelt.

Der dadurch bedingte Unterschied muss bei der Verarbeitung von statistischem Material berücksichtigt werden.

Beispiel 2:

Nominalskala	Ordinalskala	Metrische Skala
Familienstand (led., verh., . . .)	Dienstgrad (Gefreiter, . . .)	Alter (in Jahren) (1, 2, . . .)
Berufsgruppe (Arbeiter, . . .)	Verhaltensnote (zufrieden stellend, . . .)	Körpergewicht (in kg) (3, 4, 5, . . .)

Nominalskala

Geschlecht (männlich, weiblich)

Augenfarbe (blau, gelb, grün, rot usw.)

Ordinalskala

Art des Wohnorts (Einzelhaus, Dorf, Kleinstadt, Großstadt)

Fahrzeugklasse (Kleinwagen, unterer Mittelklassewagen, oberer Mittelklassewagen, Oberklassewagen)

Intervallskala

Temperatur in Celsius

IQ-Skala

Ratioskala

Körpergröße

Monatseinkommen

Qualitative Merkmale

Definition 11:

Als qualitative Merkmale bezeichnet man Merkmale, bei denen sich die Merkmalsausprägungen (Antworten) zwar eindeutig in Kategorien unterscheiden lassen, diese Antworten jedoch keinen mathematischen Wert annehmen können.

Typische Beispiele für qualitative Daten sind Geschlecht, Religionszugehörigkeit oder Parteipräferenz. Für solche Merkmale kann lediglich ein Befragungsergebnis in Anteilen (x von 100%) wiedergegeben werden.

Streng genommen zählen auch ordinale Merkmale wie Bildungsgrad, gefahrene Fahrzeugklasse oder persönliche Zufriedenheit zu den qualitativen Merkmalen. Bei ordinalen Merkmalen kann eine Hierarchie erstellt werden, eine genaue numerische Skalierung ist aber nicht möglich. Ein Bildungsgrad ist nicht „50% besser“ als ein anderer, er kann lediglich mit „höherwertiger“ spezifiziert werden. Ordinalskalen sind nicht intervallskaliert.

Quantitative Merkmale

Definition 12:

Als quantitative Merkmale bezeichnet man Merkmale, deren Merkmalsausprägungen intervallskalierte metrische Werte annehmen.

Typische Beispiele sind Körpergewicht, Einkommen oder der IQ-Wert.

Für diese Merkmale können verschiedene mathematische Rechenoperationen durchgeführt werden, wie zum Beispiel die Errechnung eines Durchschnitts.

Diskrete Werte

Definition 13:

Diskret bedeutet, dass ein Merkmal nur bestimmte isolierte (z.B. ganzzahlige) Werte annehmen kann.

Stetige Werte

Definition 14:

Stetig dagegen bedeutet, dass es mit zwei Werten auch alle Werte dazwischen annehmen kann (Alle Werte aus einem Intervall annehmen kann). Dies wird nicht durch die Messgenauigkeit eingeschränkt. Diese könnte beliebig verfeinert werden.

Beispiel 3:

Kinderzahl und Einwohnerzahl sind diskrete,

Körpergröße und Fettgehalt von Milch stetige Merkmale.

Bemerkung 4:

Da jede Messung notwendig mit einer gewissen Messungenauigkeit behaftet ist, nimmt praktisch z. B. die Körpergröße nicht alle Zahlwerte eines Intervalls an, sondern nur gewisse durch Runden entstandene Werte. Das Merkmal Körpergröße tritt in diesem Sinne in der Praxis nicht als stetiges, sondern als diskretes Merkmal auf. Aus theoretischen Gründen ist es jedoch zweckmäßig, alle Zahlwerte eines Intervalls zugelassen zu denken, d. h. die Körpergröße als stetiges Merkmal anzusehen.

Tabellarische Aufbereitung von Stichprobenwerten

Urliste

Definition 15:

Die Urliste ist im Bereich der Statistik das direkte Ergebnis einer Datenerhebung, also die ursprüngliche Aufzeichnung der Beobachtungs- oder Messwerte.

Beispiel 4:

Anlässlich einer „Schulstatistik" wurde in einer Klasse das Alter der Schüler festgestellt. Von den 34 Schülern wurden folgende Zahlen genannt:

15, 14, 14, 15, 16, 15, 15, 14, 15, 15, 15, 16, 15, 15, 14, 15, 15, 16, 17, 15, 14, 14, 15, 15, 16, 15, 15, 15, 15, 14, 14, 15, 17, 15.

Bei einer statistischen Erhebung erhält man als Erstes eine solche Liste von Beobachtungswerten.

Definition 16:

Werden die Beobachtungswerte so notiert, wie sie sich bei einer statistischen Erhebung nacheinander ergeben, so nennt man das Ergebnis eine **Urliste**. Die einzelnen Beobachtungswerte der Urliste heißen **Stichprobenwerte** (Daten);

sie werden mit $x_1, \ldots, x_n$ bezeichnet.

Bemerkung 5:

Die Stichprobenwerte $x_1, \ldots, x_n$ sind von den Merkmalsausprägungen $a_1, \ldots, a_k$ wohl zu unterscheiden.

In der obigen Urliste handelt es sich um 4 Merkmalsausprägungen:

$a_1 = 14$, $a_2 = 15$, $a_3 = 16$, $a_4 = 17$ und um 34 Stichprobenwerte.

Jeder Stichprobenwert ist zwar eine der Merkmalsausprägungen $a_1, \ldots, a_k$; während jedoch die Merkmalsausprägungen $a_1, \ldots, a_k$ alle voneinander verschieden sind, kann bei den Stichprobenwerten $x_1, -, x_n$ wiederholt derselbe Wert auftreten.

Strichliste

Definition 17:

Eine Strichliste wird als Hilfsmittel verwendet, um die Häufigkeit des Auftretens bestimmter Merkmale oder Ereignisse zu ermitteln. Hierzu werden mögliche Merkmale oder Ereignisse vorab festgestellt und untereinander aufgetragen. Bei einer Datenerhebung können mittels einer Strichliste Ereignisse oder Merkmale gezählt werden.

Welche Möglichkeiten sehen Sie, das Notieren der Antworten bei der oben angeführten Schulstatistik einfacher zu gestalten?

Wir greifen nochmals auf die im vorherigen Abschnitt angegebene Urliste von Altersangaben zurück. Hier gewinnt man einen besseren Eindruck von der Altersstruktur der Klasse, wenn man nur die Merkmalsausprägungen (also die verschiedenen vorkommenden Altersangaben) der Größe nach aufschreibt und jeden beobachteten Wert durch einen bloßen Strich festhält:

Merkmalsausprägungen	Stichprobenwerte	Absolute Häufigkeit
14	IIIIIIII	8 mal
15	IIIIIIIIIIIIIIIIIII	20 mal
16	IIII	4 mal
17	II	2 mal

Beispiel 5:

Strichlisten werden oft bei Wahlen verwendet. Beim Auszählen der Stimmen werden die Namen der Kandidaten notiert und jede Stimme hinter dem Namen des gewünschten Kandidaten mit einem Strich vermerkt.

Oft wird (wie im Beispiel der Wahlen) bei der Erhebung anstelle der Urliste sofort eine Strichliste angelegt. In Fällen, wo die Urliste bereits vorliegt und die Stichprobenwerte nun übersichtlicher dargestellt werden sollen, zählt man meist, wie oft die verschiedenen Ausprägungen in der Urliste auftreten.

Absolute Häufigkeiten

Definition 18:

Kommt eine Merkmalsausprägung a_1 in der Urliste n_i-mal vor, so nennt man n_i die absolute Häufigkeit von a, in der Urliste. Eine Tabelle, die jeder Merkmalsausprägung ihre Häufigkeit zuordnet, heißt **Häufigkeitstabelle**.

Klassierung von Stichprobenwerten

In der Urliste für die Geburtsgröße von Säuglingen liegen offenbar die meisten Stichprobenwerte zwischen 50 und 55. Stellen Sie für die Merkmalsausprägungen unter 50, 50 bis 55, über 55 eine Häufigkeitstabelle auf. Welche Vor- und Nachteile hat eine solche Zusammenfassung von Stichprobenwerten?

Bereits in einem vorherigen Kapitel wurde darauf hingewiesen, dass bei stetigen Merkmalen die theoretisch möglichen Merkmalsausprägungen zu so genannten Merkmalsklassen zusammengefasst werden müssen. Diese Notwendigkeit ist praktisch oft auch schon dann gegeben, wenn die Urliste sehr viele Stichprobenwerte enthält.

Definition 19:

Werden in der Urliste verschiedene Merkmalsausprägungen zu einer neuen Ausprägung zusammengefasst, so spricht man von einer **Klassierung** der Stichprobenwerte.

Bemerkung 6:

- Durch die Klassierung werden die Stichprobenwerte der Urliste überschaubarer, man sollte deshalb die Anzahl der Klassen nicht zu groß wählen.
- Da jedoch durch die Klassierung notwendig ein Teil der in der Urliste enthaltenen Information verloren geht, sollte man andererseits die Anzahl der Klassen auch nicht zu klein wählen.
- In der Regel sind etwa 5 bis 15 Klassen zweckmäßig.

- Man wird es möglichst so einrichten, dass die Klassenmitten einfache Zahlen sind.
- Bei jeder Klassengrenze ist anzugeben, welcher Klasse ein auf sie entfallender Stichprobenwert zuzurechnen ist; dies kann z. B. durch eine Angabe wie von 50 einschließlich bis 60 ausschließlich.

Definition 20:

Die Häufigkeiten, mit welchen die Strichprobenwerte, auf die einzelnen Klassen entfallen, heißen **Klassenhäufigkeiten**.

Beispiel 6:

Schulnoten werden häufig auch in Punkten angegeben. Bei der Umrechnung der Punkte in die üblichen Noten (von 1 bis 6) werden die Punkte, wie die folgende Tabelle zeigt, klassiert.

Punkte	15;14;13	12;11;10	9;8;7	6;5;4	3;2;1	0
Note	1	2	3	4	5	6

Relative Häufigkeiten

Absolute Häufigkeiten können oftmals nicht verwendet werden, da es sich um eine unterschiedliche Anzahl von Stichprobenwerten handelt.

Beispiel 7:

In zwei Parallelklassen wurde das Alter der Schüler ermittelt; es ergaben sich die beiden folgenden Häufigkeitstabellen.

A-Klasse:

a_i	16	17	18	19
n_i	4	16	3	2

B-Klasse:

a_i	16	17	18	19
n_i	4	13	2	1

Wobei a_i Merkmalsausprägung kennzeichnet und n_i die absolute Häufigkeit in der entsprechenden Klasse.

Nun lautet die Frage:

In welcher Klasse ist der Anteil der 17jährigen größer?

Will man die Häufigkeiten einer Merkmalsausprägung in verschiedenen Urlisten vergleichen, so berechnet man jeweils den Anteil der Ausprägung an der Gesamtheit aller Stichprobenwerte der Urliste.

Definition 21:

Tritt die Merkmalsausprägung a_i in einer Urliste mit n Stichprobenwerten n_i mal auf, so nennt man $\frac{n_i}{n}$ die relative Häufigkeit von a_i in dieser Urliste. Die relative Häufigkeit von a_i wird mit $h(a_i)$ oder kurz h_i bezeichnet:

$$h(a_i) = h_i = \frac{n_i}{n}$$

Also ergibt sich für das obere Beispiel:

A-Klasse:

a_i	16	17	18	19
n_i	4	16	3	2
n_i / n	0,16	0,64	0,12	0,08

B-Klasse:

a_i	16	17	18	19
n_i	4	13	2	1
n_i / n	0,20	0,65	0,10	0,05

Daraus folgt: Die B-Klasse hat einen größeren Anteil von 17-jährigen.

Kumulierte Häufigkeit (Summenhäufigkeit)

Definition 22:

Die kumulierte Häufigkeit umfasst die bis zur betreffenden Ausprägung aufsummierten absoluten bzw. relativen Häufigkeiten.

Bei Merkmalen mit einer Ordinal- oder metrischen Skala $a_1, \ldots, a_k$ heißt die Summe der Häufigkeiten n_i bzw. $h(a_i)$ mit $a_i < c$ **Summenhäufigkeit**.

Beispiel 8:

In einem Betrieb mit 60 Beschäftigten sind

6 Mitarbeiter bis 20Jahre alt,

18 Mitarbeiter über 20 bis 30Jahre alt,

9 Mitarbeiter über 30 bis 40Jahre alt,

12 Mitarbeiter über 40 bis 50Jahre alt,

15 Mitarbeiter über 50 bis 65 Jahre alt.

Geben Sie die relative Häufigkeit der Beschäftigten an, die höchstens 20 (höchstens 30, 40, 50, 65) Jahre alt sind.

a_i	n_i	Absolute Summen-Häufigkeit	Relative Häufigkeit	Relative Summen-Häufigkeit
bis 20	6	6	0,10	0,10
20-30	18	24	0,30	0,40
30-40	9	33	0,15	0,55
40-50	12	45	0,20	0,75
50-65	15	60	0,25	1,00

Wie gezeigt, interessiert neben den Häufigkeiten der einzelnen Merkmalsausprägungen hin und wieder auch die so genannte Summenhäufigkeit.

Beispiel 9:

Eine Klassenarbeit in einer Klasse mit 40 Schülern brachte für 2 Schüler die Note 1, 8 Schüler eine 2, 15 Schüler eine 3, 10 Schüler eine 4, 4 Schüler eine 5, 1 Schüler eine 6. Stellen Sie das Ergebnis in der unten aufgeführten Tabelle dar.

Note a_i	abs. Häufigkeit n_i	rel. Häufigkeit h_i	Absolute Summenhäufigkeit	Relative Summenhäufigkeit
1	2	5,0%	2	5,0%
2	8	20,0%	10	25,0%
3	15	37,5%	25	62,5%
4	10	25,0%	35	87,5%
5	4	10,0%	39	97,5%
6	1	2,5%	40	100,0%

Graphische Darstellungen

„Das Auge ist noch lange aufnahmefähig, wenn der Verstand schon ermattet ist.“ (Ludwig Reiners)

Definition 23:

Ein Diagramm (v. griech.: diagramma = geometrische Figur, Umriss) ist eine grafische Darstellung von Daten, Sachverhalten oder Informationen. Je nach der Zielsetzung des Diagramms werden höchst unterschiedliche Typen eingesetzt. Die Bandbreite von bildhaften Elementen bis rein abstrakten Gebilden ist dabei sehr groß.

Mit Hilfe eines Diagramms wird vor allem versucht, einen Zusammenhang zu verdeutlichen. Diagramme sind zudem oft codiert, was bedeutet, dass man mit Hilfe seines Vorwissens ein Diagramm analysieren muss, um es verstehen zu können.

Die am häufigsten gewählten Darstellungsarten sind das **Kreisdiagramm** und das **Säulendiagramm**. Das Säulendiagramm wird oft auch als **Balkendiagramm**bezeichnet, wobei dieser Begriff den Querbalken vorbehalten sein.

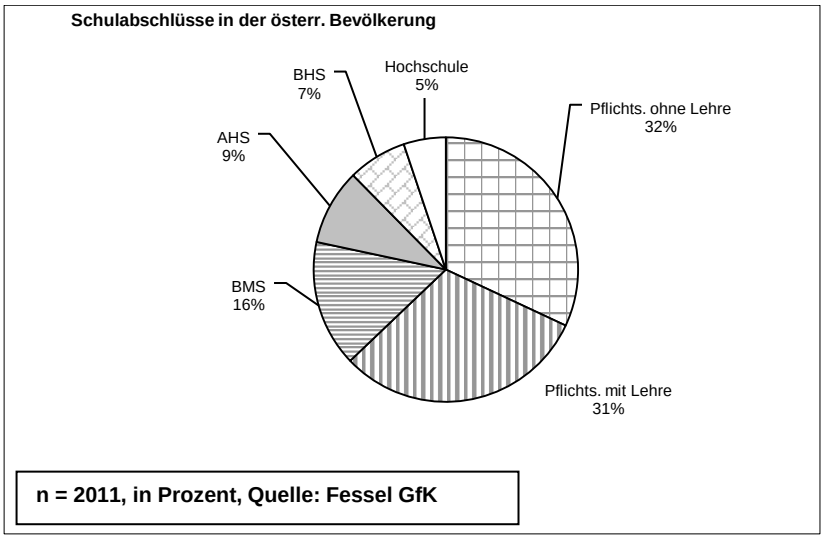


Abbildung: Kreisdiagramm (Tortendiagramm)

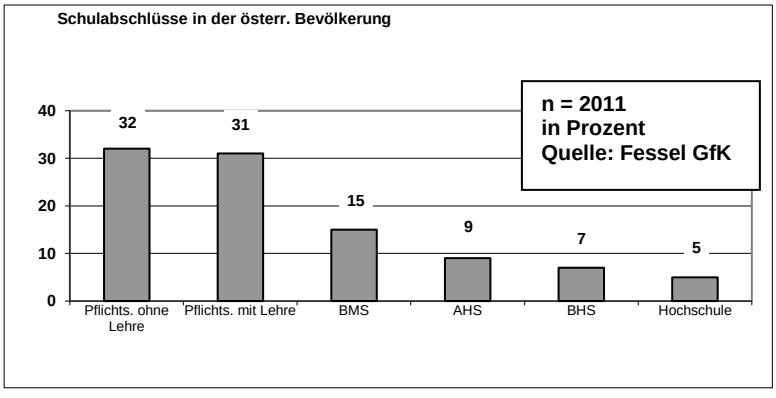


Abbildung: Säulendiagramm

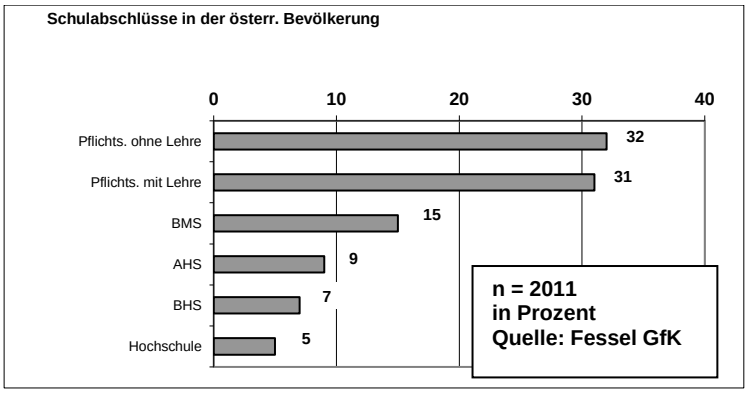


Abbildung: Balkendiagramm

Bemerkung 7:

- Die Funktion einer statistischen Grafik ist die schnelle Information über einen interessanten Sachverhalt, weshalb man auf **grafische Spielereien und Überladungen** verzichten sollte, um nicht von der wesentlichen Information abzulenken.
- Dabei ist eine ausreichende Beschriftung (der Balken, Achsen, Segmente etc.) zu gewährleisten. Zusätzlich erforderliche Angaben: Überschrift, Quelle, Stichprobengröße, Legende und ähnliches.
- Bei Balken- und Säulendiagrammen muss die Länge der Balken bzw. Säulen proportional zur darzustellenden Häufigkeit sein, z.B. muss der Balken, der eine Häufigkeit von 0,75 darstellt, dreimal so lang sein wie einer, der eine Häufigkeit von 0,25 darstellt.
- Beim Kreisdiagramm berechnet sich der Mittelpunktswinkel α_i des zur Darstellung der relativen Häufigkeit $h_n(x_i)$ der Merkmalsausprägung x_i gehörenden Kreissektors durch $\alpha_i = h_n(x_i) \cdot 2\pi$ (bzw. $\alpha_i = h_n(x_i) \cdot 360°$).

Lagemaße/Lageparameter/Maße der Zentraltendenz

Es gibt verschiedene Lagemaße, die alle jedoch ein Ziel verfolgen: Nämlich die Ermittlung einer **zentralen Tendenz**. Umgangssprachlich gestellte Fragestellungen wie

Welche Masse für eine erwachsene Frau "normal" sei

oder

Auf was sich das "durchschnittliche" Einkommen eines Managers in Deutschland beläuft

laufen auf die Ermittlung eines Lagemaßes (oder auch Lageparameters) hinaus.

Einleitung

In vorangegangenen Lektionen wurden Häufigkeitstabellen und Grafiken vorgestellt. Sie bieten Möglichkeiten, einen umfassenden Überblick der Verteilung einer Variablen zu erhalten.

Im Gegensatz dazu repräsentieren die nun Folgend dargestellten Maße der zentralen Tendenz das Typische einer Verteilung.

Sie informieren zusammenfassend über spezielle Eigenschaften der Merkmalsverteilung. Diese statistischen Kennwerte werden auch als Lagemaße bezeichnet, Modalwert, Median und das arithmetische Mittel sind die üblichen Maße der zentralen Tendenz.

Dabei muss berücksichtigt werden, dass unterschiedliche Voraussetzungen der Daten für die Anwendung der verschiedenen Lagemaße erfüllt sein müssen.

Arithmetisches Mittel

Das arithmetische Mittel ist das gebräuchlichste und wichtigste Maß der zentralen Tendenz. Umgangssprachlich kennen wir es als „Durchschnitt“.

Das arithmetische Mittel wird berechnet nach der Formel:

Definition 24:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Es wird berechnet als Summe der Werte, deren Mittelwert wir suchen, geteilt durch die Anzahl dieser Werte.

Aufgrund folgender zwei Eigenschaften besitzt das arithmetische Mittel eine hohe Bedeutung in der Statistik:

Bemerkung 8:

- Die Summe der Abweichungen der Einzelwerte vom arithmetischen Mittel ist Null; positive und negative Abweichungen gleichen sich gegenseitig aus.
- Alle Messwerte einer Variablen fließen in die Berechnung des arithmetischen Mittels ein. Somit liefert das arithmetische Mittel die meiste Information über die Verteilung der Werte einer Variablen.

- Vergleicht man die Mittelwerte von zwei Verteilungen, kann man z.B. Unterschiede oder auch Gemeinsamkeiten zwischen den Verteilungen feststellen.
- Aufgrund der Berücksichtigung aller Messwerte bei der Berechnung ist das arithmetische Mittel für Extremwerte bzw. Ausreißer anfällig. Insbesondere bei geringer Zahl der Einzelwerte können extreme Messwerte das arithmetische Mittel stark verzerren.

Bemerkung 9:

Voraussetzung zur Berechnung des arithmetischen Mittels:

- Der arithmetische Mittelwert soll nur dann berechnet werden, wenn die zu untersuchende Variable **metrisch skaliert** ist.
- Das arithmetische Mittel sollte nicht berechnet werden, wenn die Verteilung eindeutig mehrgipflig oder schief ist, und an den Enden offene Randklassen aufweist.

Beispiel 10:

Sie haben in der folgenden Tabelle die Daten einer Statistik-Vorlesung auf der DHBW Villingen-Schwenningen Fachrichtung Freizeitgestaltung.

Geschlecht	Größe in cm	Alter	Brille	Raucher/in	Augenfarbe
m	179	23	ja	nein	blau
w	164	22	ja	nein	blau
w	165	30	nein	nein	andere
m	176	28	ja	nein	graugrün
m	175	24	ja	nein	blau
m	180	32	ja	ja	braun
w	160	25	ja	nein	braun
w	164	23	ja	nein	graugrün
w	170	24	nein	nein	blau
m	182	30	ja	nein	andere

Berechnen Sie hier den Mittelwert der Größe.

Lösung:

Mittelwert: $\bar{x} = \frac{179 + 164 + 165 + 176 + 175 + 180 + 160 + 164 + 170 + 182}{10} = 171{,}50$

Median

Definition 25:

Der Median ist die Merkmalsausprägung des genau in der Mitte liegenden Einzelwertes.

Er teilt die der Größe nach geordneten Messergebnissen in zwei Hälften.

Er wird häufig auch als Zentralwert bezeichnet.

Bemerkung 10:

- Der Median eignet sich besonders, wenn das arithmetische Mittel nicht berechnet werden sollte, z.B. bei nicht metrischen Daten.
- Bei offenen Randklassen oder bei schiefen Verteilungen.

Bei der Berechnung sind 3 Fälle zu unterscheiden:

Definition 26:

1. Der Median kommt als Wert vor; unter ihm liegen gleich viele Werte wie über ihm. Die Anzahl der Werte ist ungerade.

$$x_{\text{Median}} = x_{\left(\frac{n+1}{2}\right)}$$

2. Der Median fällt in eine Lücke. Die Anzahl der Werte ist gerade.

$$x_{\text{Median}} = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right)$$

3. Wenn Daten in Klassen geordnet (gruppierte Daten) sind, und der Median liegt in einer besetzten Kategorie, kann es sein, dass darüber und darunter nicht gleich viele Fälle liegen. In diesem Fall muss interpoliert werden.

Eigenschaften des Medians

Bemerkung 11:

- Der Median ist gegenüber Extremwerten bzw. Ausreißern unempfindlich. Nur Veränderungen in den mittleren Bereichen beeinflussen ihn.
- Der Median kann bei mindestens ordinalskalierten Daten angegeben werden.

Nachteile des Median

- Der Vergleich von Medianen zwischen zwei Verteilungen zeigt nicht immer die Unterschiede, die der Vergleich der arithmetischen Mittelwerte ermöglicht.

Beispiel 11:

Sie haben in der folgenden Tabelle die Daten einer Statistik-Vorlesung auf der DHBW Villingen-Schwenningen Fachrichtung Freizeitgestaltung.

Geschlecht	Größe in cm	Alter	Brille	Raucher/in	Augenfarbe
m	179	23	ja	nein	blau
w	164	22	ja	nein	blau
w	165	30	nein	nein	andere
m	176	28	ja	nein	graugrün
m	175	24	ja	nein	blau
m	180	32	ja	ja	braun
w	160	25	ja	nein	braun
w	164	23	ja	nein	graugrün
w	170	24	nein	nein	blau
m	182	30	ja	nein	andere

a) Stellen Sie den Median für die gemessene Körpergröße fest.

Zuerst wird die Tabelle nach der Körpergröße sortiert

Geschlecht	Größe in cm	Alter	Brille	Raucher/in	Augenfarbe
w	160	25	ja	nein	braun
w	164	22	ja	nein	blau
w	164	23	ja	nein	graugrün
w	165	30	nein	nein	andere
w	170	24	nein	nein	blau
m	175	24	ja	nein	blau
m	176	28	ja	nein	graugrün
m	179	23	ja	nein	blau
m	180	32	ja	ja	braun
m	182	30	ja	nein	andere

Der Median fällt in eine Lücke. Die Anzahl der Werte ist gerade

Es wird nach der folgenden Formel die Berechnung durchgeführt:

$$x_{Median} = \frac{1}{2} \cdot \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)} \right) = \frac{1}{2} \cdot (170 + 175) = 172{,}5$$

b) Wie groß ist der Median, wenn dieser Datensatz noch hinzugefügt würde,

m	184	30	ja	nein	andere

Die Tabelle hat sich nun folgendermaßen erweitert und im gleichen Zuge auch noch sortiert:

Geschlecht	Größe in cm	Alter	Brille	Raucher/in	Augenfarbe
w	160	25	ja	nein	braun
w	164	22	ja	nein	blau
w	164	23	ja	nein	graugrün
w	165	30	nein	nein	andere
w	170	24	nein	nein	blau
m	175	24	ja	nein	blau
m	176	28	ja	nein	graugrün
m	179	23	ja	nein	blau
m	180	32	ja	ja	braun
m	182	30	ja	nein	andere
m	184	30	ja	nein	andere

Der Median kommt als Wert vor; unter ihm liegen gleich viele Werte wie über ihm. Die Anzahl der Werte ist ungerade.

Es wird nach der folgenden Formel die Berechnung durchgeführt:

$$x_{Median} = x_{\left(\frac{n+1}{2}\right)} = 175$$

Modus

Definition 27:

Der Modus ist derjenige Merkmalswert einer Verteilung, der am häufigsten vorkommt.

Bemerkung 12:

- In einer graphischen Darstellung ist er das Maximum einer Verteilung.
- So ist eine einfache Bestimmung möglich, da der Modus direkt aus der Häufigkeitstabelle oder aus der graphischen Darstellung entnommen werden kann.
- Hierbei ist zu beachten, dass der Modus entweder der einzelne Wert bei nicht gruppierten Daten oder eine Klasse bei gruppierten Daten ist, der/die am häufigsten vorkommt.
- Bei gruppierten Daten entspricht der Modus der Klassenmitte der Klasse mit der größten Häufigkeit.

Eigenschaften des Modus:

Bemerkung 13:

- Reale Merkmalsausprägung
- Der Modus ist bei metrisch skalierten, gruppierten Daten und Nominaldaten anwendbar. Er ist das einzige Maß der zentralen Tendenz, das auch auf Nominaldaten angewendet werden kann.
- Gegenüber Ausreißern ist der Modus unempfindlich.

Nachteile des Modus

- Der Modus unterliegt einer relativen Zufallsabhängigkeit. Durch geringe Änderungen der Daten in der Nähe der häufigsten Werte oder durch Änderung der Klassengrenzen kann es beim Modus zum Teil zu entscheidenden Veränderungen kommen, die dem objektiven Untersuchungsgegenstand und dessen Veränderungen nicht immer unbedingt entsprechen. Diese Zufallsabhängigkeit ist auch besonders augenfällig bei Verteilungen, die eher einer Rechtecksverteilung entsprechen (alle Werte haben die gleiche Häufigkeit). Kleine Veränderungen können dann den Modus von einem Ende der Häufigkeitsverteilung ans andere springen lassen.
- Gibt es zwei nebeneinander liegende x_i-Werte mit gleich großen Häufigkeiten, so ist das arithmetische Mittel dieser beiden x_i-Werte als Modalwert zu benennen. Gibt es aber zwei x_i-Werte mit gleich großen Häufigkeiten, die nicht nebeneinander liegen, so hat die Verteilung zwei Modalwerte, d.h. zwei „Gipfel“, sie ist bimodal.
- Die Berechnung des Modalwerts ist nur bei unimodalen Verteilungen sinnvoll. (Modus kommt nur einmal vor)

Die Schiefe

Definition 28:

In der mathematischen Statistik bezeichnet die Schiefe ein Maß für die Abweichung einer Zufallsvariablen von einer zum Mittelwert symmetrischen Verteilung.

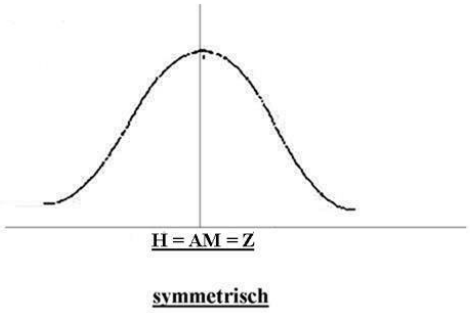
Bemerkung 14:

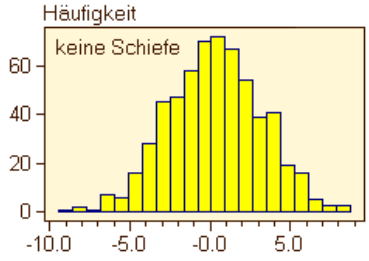
- Eine schiefe Verteilung ist ebenfalls asymmetrisch.
- Eine schiefe Verteilung ist gerade durch eine verschiedene Neigung zweier Kurvenäste gekennzeichnet.
- Auch in einer mehrgipfeligen Verteilung, die von mehr als zwei Kurvenästen gebildet wird, kann bei diesen Symmetrie oder Asymmetrie im Sinne von Schiefe entstehen.

Bemerkung 15:

- Die Schiefe nimmt Werte unter oder über Null an. Als Messkriterium wird die Normalverteilung angenommen, in ihr ist die Schiefe Null, das heißt, es befinden sich gleich viele Werte unter und ober dem arithmetischen Mittel der Verteilung.
- Je größer die Schiefe, desto weiter klaffen arithmetisches Mittel, Median und Modus auseinander.

Zusammenhang der Maße der zentralen Tendenz und Verteilungsform

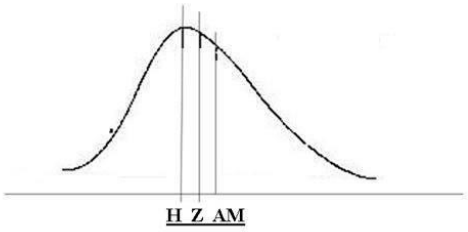


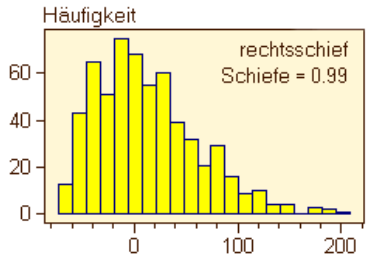


Bei symmetrischen Verteilungen fallen alle drei Maße der zentralen Tendenz zusammen:

Arithmetisches Mittel (AM) = Median (Z) = Modus (H)

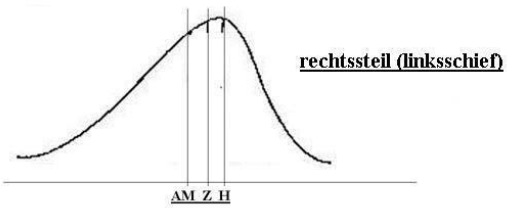
linkssteil (rechtsschief)

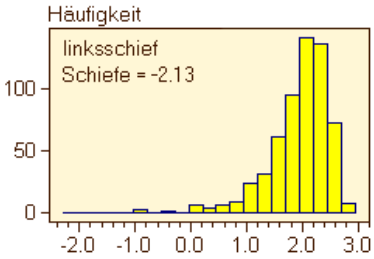




Bei rechtsschiefen Verteilungen verhalten sich die drei Maße folgendermaßen zueinander:

Modus (H) < Median (Z) < arithmetisches Mittel (AM)





Bei linksschiefen Verteilungen verhalten sich die drei Maße im Verhältnis zueinander:

arithmetisches Mittel (AM) < Median (Z) < Modus (H)

Die Schiefe ist ein Maß der Asymmetrie.

Definition 29:

Zur Berechnung der Schiefe einer empirischen Häufigkeitsverteilung wird die folgende Formel benutzt:

$$v = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

Bemerkung 16:

- Damit die Schiefe unabhängig von der Maßeinheit der Variablen ist, werden die Messwerte mit Hilfe des arithmetischen Mittelwertes $\bar{x}$ und der Standardabweichung der Beobachtungswerte x_i standardisiert.
- Deutung:
  Ist $v > 0$, so ist die Verteilung rechtsschief (auch genannt Linkssteil), ist $v < 0$, so ist die Verteilung linksschief (auch genannt rechtssteil). Gilt $v = 0$, so ist die Verteilung auf beiden Seiten ausgeglichen.

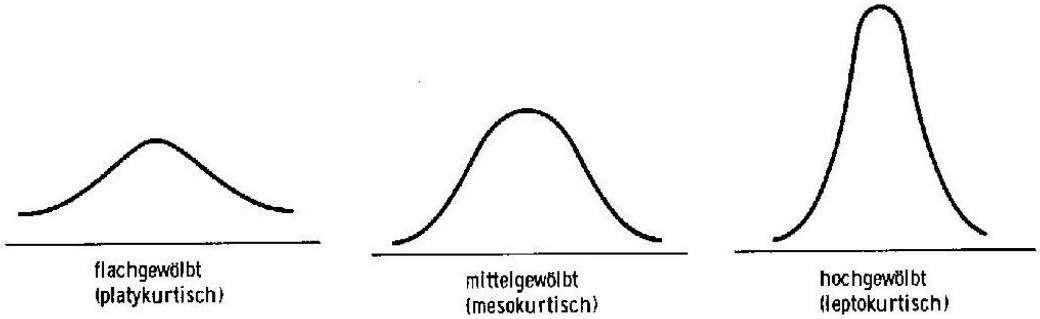
Die Wölbung

Neben der Schiefe ist auch die Wölbung kennzeichnend für eine Verteilung.

Definition 30:

Die **Wölbung** oder **Kurtosis** (griechisch: das Krümmen, Wölben) ist eine Maßzahl für die Steilheit bzw. „Spitzigkeit“ einer (eingipfligen) Wahrscheinlichkeitsfunktion, statistischen Dichtefunktion oder Häufigkeitsverteilung.

Eine Verteilung kann sehr schmalbrüstig oder sehr flach sein. Je nach dem, spricht man von einer **platykurtischen** (sehr flachen), **mesokurtischen** (mittelsteil) und **leptokurtischen** (sehr steilen) Verteilungskurve bzw. Verteilung.



Auch für die Messung der Wölbung ist die Normalverteilungskurve Kriterium. Wird nun die Verteilung nach unten flacher, so sinkt die Wölbung unter null und umgekehrt. Je steiler oder je flacher die Kurve, desto größer die Wölbung im positiven oder negativen Bereich.

Bemerkung 17:

- Verteilungen mit geringer Wölbung streuen relativ gleichmäßig; bei Verteilungen mit hoher Wölbung resultiert die Streuung mehr aus extremen, aber seltenen Ereignissen.
- Eine stark oder schwach gewölbte Kurve kann durchaus symmetrisch sein. Die Wölbung ist ein Maß für die Häufung von Werten.
- Entweder scharen sie sich um die Mitte der Verteilung oder sie verteilen sich gleichmäßig bis zu den Enden. Die flachste Kurve wäre eine, in der alle Werte gleich oft vorkommen (Gleichverteilung)

Definition 31:

Zur Berechnung der Wölbung einer *empirischen Häufigkeitsverteilung* $x_1, x_2, \cdots, x_n$ wird die folgende Formel benutzt:

$$w = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4$$

Damit die Wölbung unabhängig von der Maßeinheit der Variablen ist, werden die Beobachtungswerte x_i mit Hilfe des arithmetischen Mittelwertes $\bar{x}$ und der Standardabweichung s standardisiert.

Exzess

Um das Ausmaß der Wölbung besser einschätzen zu können, wird sie mit der Wölbung einer Normalverteilung verglichen, für die $\beta = 3$ gilt. Der Exzess (auch: Überkurtosis) ist daher definiert als

Definition 32:

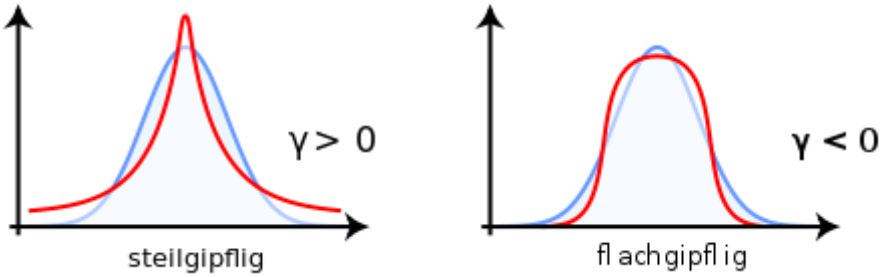
$$Exzess = Wölbung - 3$$

Nicht selten wird die Wölbung fälschlicherweise als Exzess bezeichnet.

Arten von Exzess

Verteilungen werden entsprechend ihres Exzesses eingeteilt in:

- **Exzess = 0**: normalgipflig oder mesokurtisch. Die Normalverteilung hat die Kurtosis $\beta = 3$ und entsprechend den Exzess 0.
- **Exzess > 0**: steilgipflig, supergaußförmig oder leptokurtisch. Es handelt sich hierbei um im Vergleich zur Normalverteilung spitzere Verteilungen, d.h. Verteilungen mit starken Peaks.
- **Exzess < 0**: flachgipflig, subgaußförmig oder platykurtisch. Man spricht von einer im Vergleich zur Normalverteilung abgeflachten Verteilung.

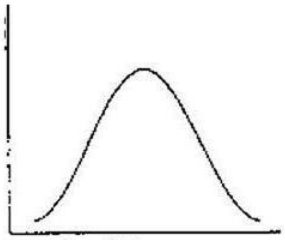


Die Modalität

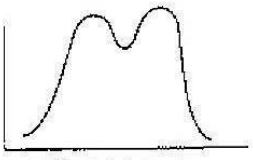
Definition 33:

Eine Verteilungskurve kann einen oder mehrere Höhepunkte aufweisen. Diese markieren die Spitzen der Verteilung.

Das sind entweder die häufigsten Werte einer Häufigkeitsverteilung oder die Häufigkeitsmaxima einer Dichteverteilung.



unimodal



bimodal

Gewogenes (gewichtetes) arithmetisches Mittel

Definition 34:

Das gewogene arithmetische Mittel wird verwendet wenn die Einzelwerte gehäuft vorkommen und wenn man ein arithmetisches Mittel aus Mittelwerten unterschiedlich großer Teilmengen berechnen möchte.

$$\bar{x}_g = \frac{\sum_{i=1}^{k} \bar{x}_i \cdot n_i}{n}$$

Bemerkung 18:

- Das kommt z.B. vor, wenn man eine Variable in mehreren Teilmengen gemessen und in jeder das Arithmetische Mittel berechnet hat. Mitunter müssen aber auch Werte gemittelt werden, die nicht von gleicher Wichtigkeit sind oder die gleiche Bedeutung haben.
- Beim gewogenen arithmetischen Mittel wird der unterschiedlichen Gruppenstärke durch Gewichtung der Gruppenmittelwerte mit der jeweiligen Gruppengröße oder anderen Gewichtungsfaktoren Rechnung getragen.

Beispiel 12:

Hundert Frauen sind durchschnittlich 168cm groß, 50 Männer durchschnittlich 180cm. Wie groß ist die Gesamtgruppe im Durchschnitt?

$$\bar{x}_g = \frac{168 \cdot 100 + 180 \cdot 50}{150} = 172 \text{ cm}$$

Geometrisches Mittel

Definition 35:

Das geometrische Mittel wird berechnet bei multiplikativ verknüpften Merkmalsreihen, wie z.B. Wachstumsraten.

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Bemerkung 19:

- Mit anderen Worten, bei Messwertreihen, deren Abstände sich untereinander proportional zur Größe der Messwerte verhalten, die einer geometrischen Reihe ähnlich sind, ist die Berechnung des geometrischen Mittels angemessen. Das geometrische Mittel setzt Daten auf Verhältnisskalenniveau und Werte größer Null voraus.
- Das geometrische Mittel wird immer dann verwendet, wenn eine Reihe von Einzelwerten vorliegt, die selbst nicht normalverteilt sind, während dies für ihre Logarithmen zutrifft.

Beispiel 13:

Eine Bakterienkultur wächst in pro Zeiteinheit durchschnittlich um 50%. Die Zuwachsrate schwankt zufällig. In fünf aufeinander folgenden Zeiteinheiten werden folgende Populationsbestände gemessen:

Bestand	1000	1800	2520	3276	4586
Veränderungsfaktor		1,8	1,4	1,3	1,4
Zuwachs		800	720	756	1310

Um den durchschnittlichen Veränderungsfaktor zu bestimmen, werden zum Vergleich das arithmetische und das geometrische Mittel berechnet.

Arithmetische Mittel:

$$\bar{x} = \frac{1{,}8 + 1{,}4 + 1{,}3 + 1{,}4}{4} = 1{,}4750$$

Geometrisches Mittel:

$$\bar{x}_g = \sqrt[4]{1{,}8 \cdot 1{,}4 \cdot 1{,}3 \cdot 1{,}4} = 1{,}4634$$

Multipliziert man den Bestand der Ausgangspopulation (n_o) viermal mit diesen Mittelwerten, so sollte sich der Bestand der Population nach dem vierten Zeitabschnitt ergeben.

$n_0 \cdot 1{,}475^4 = 4733$ und $n_0 \cdot 1{,}4634^4 = 4586$

Man sieht, dass bei Verwendung des geometrischen Mittels der Wert vorhergesagt wird, der tatsächlich aus den gemittelten Wachstumsraten resultiert.

Harmonisches Mittel

Definition 36:

Das harmonische Mittel ist ein Mittelwert einer Menge von Zahlen und wird typischerweise für die Mittelwertbildung von Anteilswerten oder Prozentzahlen genutzt.

Das harmonische Mittel $\bar{x}_h$ von n Merkmalswerten ist der Kehrwert des arithmetischen Mittels der Kehrwerte aller n Merkmalswerte

$$x_1, x_2, \cdots\cdots, x_n \;\rightarrow\; \bar{x}_h = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{\frac{a_1}{x_1} + \frac{a_2}{x_2} + \frac{a_3}{x_3} + \cdots\cdots\cdots + \frac{a_n}{x_n}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \cdots\cdots\cdots + \frac{1}{x_n}} = \frac{n}{\sum_1^n \frac{1}{x_i}}$$

Bemerkung 20:

- Das harmonische Mittel kommt zur Anwendung, wenn Indexzahlen (Kilometer pro Stunde oder Preis pro Liter etc.) zu mitteln sind und die Zählervariable in den Einzelwerten konstant ist.
- Es ist geeignet eine Reihe von Messwerten zu kennzeichnen, die z.B. Leistungslimits oder Zeitlimits darstellen.

Beispiel 14:

Ein Autofahrer fährt staubedingt 50 km mit einer Geschwindigkeit von 20 km/h und danach 50km mit 125 km/h. Wie lautet die Durchschnittsgeschwindigkeit für die Gesamtstrecke von 100 km?

Die spontane Antwort (20 km/h+125 km/h)/2=72,5 km/h ist falsch, denn die Durchschnittsgeschwindigkeit ergibt sich als Gesamtstrecke/Gesamtzeit.

Für die 2x50km benötigt der Fahrer 50/20+50/125=2.5+0.4=2.9 Stunden, so dass sich eine Durchschnittsgeschwindigkeit von 100km/2.9h=34.48km/h ergibt. Dieser Wert entspricht dem harmonischen Mittel der beiden Geschwindigkeiten.

$$\bar{x}_H = \frac{2 \cdot 50km}{\frac{50km}{20\frac{km}{h}} + \frac{50km}{125\frac{km}{h}}} = \frac{2}{\frac{1}{20\frac{km}{h}} + \frac{1}{125\frac{km}{h}}} = 34{,}48\frac{km}{h}$$

Auch das harmonische Mittel kann als gewogenes harmonisches Mittel berechnet werden.

Beispiel 15:

Erwin kauft auf den Großmarkt für 12 Euro Apfelsinen, die 0,50 Euro/Stück kosten.

Erwin kauft wieder für 12 Euro Apfelsinen, die jetzt nur noch 0,40 Euro/Stück kosten.

Erwin kauft noch einmal für 12 Euro Apfelsinen, die jetzt nur noch 0,30 Euro/Stück kosten.

Der Durchschnittspreis der Apfelsinen ist nicht 0,40 Euro/Stück.

Erwin hat für insgesamt 36 Euro Apfelsinen gekauft. Um den durchschnittlichen Preis zu ermitteln, müsste man die 36 Euro durch die Anzahl der gekauften Apfelsinen dividieren.

Erwin bekommt beim ersten Mal 24 Apfelsinen, beim zweiten Mal 30 Apfelsinen und beim dritten Mal 40 Apfelsinen, also insgesamt 94 Stück. Der Durchschnittspreis pro Apfelsine also:

$$\frac{36\ Euro}{94\ St\ddot{u}ck} = 0{,}3829787\ Euro/St\ddot{u}ck$$

Dieses Ergebnis erhält man auch dann, wenn das harmonische Mittel der Preise berechnet wird. Denn der Betrag des täglichen Kaufes ändert sich nicht.

$$\bar{x}_h = \frac{n}{\frac{1}{0{,}5} + \frac{1}{0{,}4} + \frac{1}{0{,}3}} = 0{,}3829787$$

Getrimmter Mittelwert

Der getrimmte Mittelwert verbindet die Vorteile des Medians mit denen des arithmetischen Mittelwerts.

Definition 37:

Er vermindert die Effekte von Ausreißern dadurch, dass er extreme Werte an den Enden der Verteilung unberücksichtigt lässt. Die verbleibenden Werte werden erst nach Ausschluss der Ausreißer gemittelt.

Bemerkung 21:

Es werden die Ausreißer bei der Berechnung nicht berücksichtigt.

Man sollte sich auf jeden Fall im Klaren sein, woher diese Ausreißer kommen und ob man sie "einfach" weglassen kann.

Streumaße (Dispersionsmaße)

Einleitung

Dispersionsmaße verdeutlichen, wie stark sich die Merkmalswerte voneinander unterscheiden. Sie beschreiben die Streuung in der Gesamtheit, einer Stichprobe oder einer Gruppe von Fällen bzw. Untersuchungseinheiten.

Maßzahlen, welche die Streuung in einer Verteilung ausdrücken, bilden wesentliche Ergänzungen zu den Lagemaßen. Erst beide Maßzahlen gemeinsam geben Aufschluss über die Form einer Verteilung und damit über die Variabilität eines Merkmals.

Ein Streuungsvergleich zwischen verschiedenen Stichproben lässt sich grob anhand der Kurve der Häufigkeitsverteilungen vornehmen.

Eine „schmale“ Kurve verweist auf eine eher geringe Streuung, eine „breite“ Kurve auf eine größere Streuung. Rechnerisch gibt es grundsätzlich zwei Wege, die Streuung mit einem Kennwert zu erfassen. Entweder man berechnet die Differenzen zwischen hohen und niedrigen Werten, oder man ermittelt die durchschnittlichen Abstände der Messwerte vom Mittelpunkt der Verteilung.

Zur ersten Gruppe gehören die Kennwerte Spannweite, Zentilabstand und Quartilsabstand.

Zur zweiten Gruppe, den Abstandsmaßen im engeren Sinn, in deren Berechnung der Mittelwert einfließt, zählen die durchschnittliche Abweichung, die Varianz, die Standardabweichung und der Variationskoeffizient.

Spannweite

Definition 38:

Die Spannweite gibt den gesamten Streuungsbereich der Messwerte eines Kollektivs bzw. einer Stichprobe an. Sie ergibt sich aus der Differenz des größten und kleinsten Werts der Verteilung. Die Formel lautet:

$$R = x_{max} - x_{min}$$

Bemerkung 22:

- Bei kontinuierlichen Merkmalen beschreibt die Spannweite die Größe des Intervalls, in welchem alle gemessenen Werte der Variablen liegen.
- Bei diskreten Variablen oder Klassenvariablen ist die Interpretation schwieriger. Sie gibt dann die Anzahl der Kategorien vermindert um eins an.
- Die Spannweite kann für alle Skalenniveaus, mit Ausnahme der Nominalskala, berechnet werden.
- Sie ist sehr einfach zu berechnen.

Diesem Vorteil stehen allerdings einige gewichtige Nachteile gegenüber.

- Da sie nur zwei Messwerte berücksichtigt, ist sie für die Verteilung der Werte insgesamt nicht sehr repräsentativ.

- Sie ist anfällig gegenüber einzelnen sehr extremen Werten, die ihr Ergebnis schnell verzerren.
- In der Praxis, besonders bei diskreten Variablen, wird die Spannweite oft nicht explizit berechnet, sondern lediglich der kleinste und der größte Wert genannt. Es ist zum Beispiel üblich, anzugeben, dass etwa das Alter aller Befragten zwischen 18 und 45 Jahren lag. Die Spannweite von 27 Jahren zu erwähnen, erübrigt sich.

Bedeutsam ist die Angabe der Spannweite, oder des niedrigsten und des höchsten Werts, vor allem bei numerischen Variablen ohne Antwortvorgabe.

Dies gilt ganz besonders bei Variablen, die Rahmenbedingungen für wissenschaftliche Untersuchungen darstellen.

Ein Beispiel ist die Variable Alter bei sozialwissenschaftlichen Untersuchungen. In medizinischen Studien sind dies auch Merkmale wie Gewicht oder Körpergröße aller untersuchten Fälle.

Bei Variablen mit Antwortvorgaben entlang einer vordefinierten Skala ist die Spannweite von untergeordneter Bedeutung. Es gibt niemals Werte die größer oder kleiner sind als die beiden Endpunkte der Skala.

Nur wenn bei den gemessenen Werten die oberen und / oder unteren Werte der Skala überhaupt nicht auftreten, könnte die Angabe der Spannweite von besonderem Interesse sein.

Beispiel 16:

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	
Note Mädchen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6	$\bar{x} = 3,0$
Note Jungen	1,0	1,0	2,0	2,5	3,2	2,8	3,5	2,0	6,0	6,0	$\bar{x} = 3,0$

Spannweite Mädchen: $R_M = 3,5 - 2,5 = 1$

Spannweite Jungen: $R_J = 6,0 - 1,0 = 5$

Quantil

Quantile sind ein Streuungsmaß in der Statistik. Quantile sind Punkte einer nach Rang oder Größe der Einzelwerte sortierten statistischen Verteilung.

Definition 39:

Wird die gesamte Verteilung in n gleich große Teile unterteilt, so gibt es n − 1 Quantile, also umgangssprachlich die Schnittstellen. Je nachdem wie groß n gewählt wird, spricht man z. B. von **Quartilen (n = 4)**, **Quintilen (n = 5)**, **Dezilen (n = 10)** und **Perzentilen (n = 100)**.

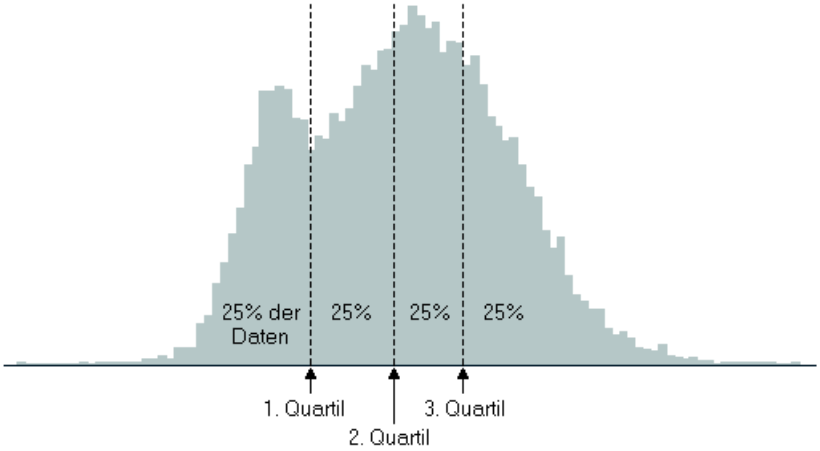
Quartil

Definition 40:

Mit Quartilen und Spannweiten lassen sich Messreihen miteinander vergleichen.

Bemerkung 23:

- Zur Berechnung dieser Streumaße muss aber eine Rangwertliste vorliegen, d. h. **die Werte müssen der Größe nach sortiert** werden.
- Ein Wert heißt oberes Quartil, wenn mindestens ein Viertel aller Werte größer (oder gleich) ist, als dieser Wert.
- Die Spannweite ist die Differenz aus dem größten und kleinsten Wert, der Quartilsabstand die Differenz aus dem oberen und unteren Quartil der Rangwertliste.
- Quartile teilen, wie der Name suggeriert, die zu Grunde liegende Verteilung in vier Viertel. Ein bestimmtes Quartil ist also die Grenze zwischen zwei bestimmten Vierteln der Verteilung.



Die Berechnung von Quartilen ist manchmal (vor allem bei Stichproben deren Umfang nicht durch vier teilbar ist) unklar. Darum im Folgenden eine exakte Anleitung zur Berechnung von Quartilen. Für eine Stichprobe von N Beobachtungen gilt ("round" steht für die "normale" Rundung):

Definition 41:

1. Quartil: jener Wert der sortierten Reihenfolge der an x-ter Stelle steht, wobei für x gilt: x = round(0.25*(N+1))

2. Quartil (Median): falls N gerade, ist Q_2 der Mittelwert der beiden Werte an den Stellen N/2 und N/2+1; falls N ungerade ist Q_2 der Wert an der Stelle (N+1)/2

3. Quartil: jener Wert der sortierten Reihenfolge der an x-ter Stelle steht, wobei für x gilt: x = round(0.75*(N+1))

Beispiel 17:

Angenommen man hat folgende 20 Beobachtungen gemacht:

2, 4, 7, -20, 22, -1, 0, -1, 7, 15, 8, 4, -4, 11, 11, 12, 3, 12, 18, 1

Zur Berechnung der Quartile ist die Liste der Beobachtungen zuerst zu sortieren:

-20, -4, -1, -1, 0, 1, 2, 3, 4, 4, 7, 7, 8, 11, 11, 12, 12, 15, 18, 22

Für das 1. Quartil gilt nun: x = round(0.25*(20+1)) = round(5.25) = 5.

Das heißt, Q_1 ist der Wert der 5. Stelle in der sortierten Reihenfolge, also $Q_1 = 0$.

Für Q_2 ergibt sich analog $Q_2 = 5.5$ und für das 3. Quartil $Q_3 = 12$.

Anmerkung zur Praxis:

Quartile gibt man üblicherweise erst ab 12 Beobachtungen an (besser wären aber mehr als 20). Eine etwas andere Rechenweise finden Sie hier.

Beispiel 18:

Die Liste enthält von 13 Schülern die Körpergröße.
Die Merkmalsausprägungen (Beobachtungswerte) wurden nach der Größe geordnet.

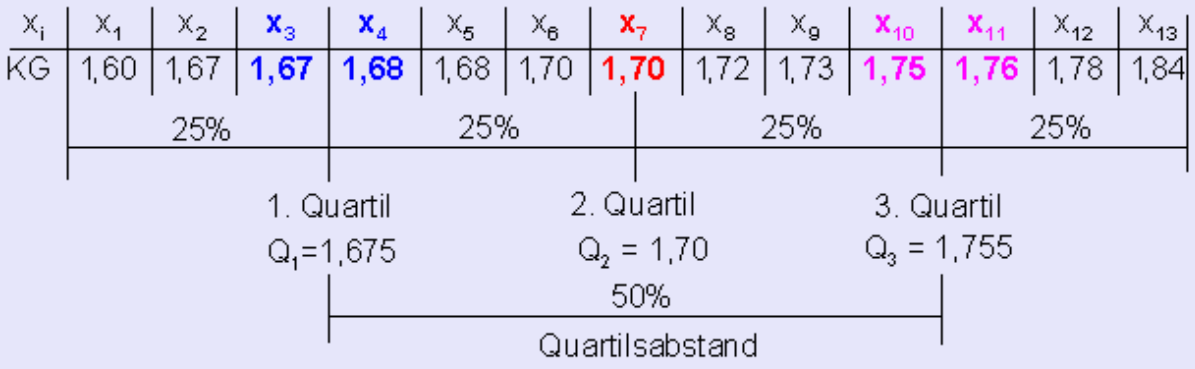
x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
KG	1,60	1,67	1,67	1,68	1,68	1,70	1,70	1,72	1,73	1,75	1,76	1,78	1,84

x_i = Beobachtungswert x_i ; KG = Körpergröße in m

Median / 2. Quartil: $Q_2 = x_7 = \underline{1{,}70}$

1. Quartil: $Q_1 = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(1{,}67 + 1{,}68) = \underline{1{,}675}$

3. Quartil: $Q_3 = \frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(1{,}75 + 1{,}76) = \underline{1{,}755}$



Etwa 25% aller geordneten Beobachtungswerte sind kleiner als das 1. Quartil.

Etwa 50% aller geordneten Beobachtungswerte sind kleiner als das 2. Quartil.

Etwa 75% aller geordneten Beobachtungswerte sind kleiner als das 3. Quartil.

Beispiel 19:

Ein Landwirt misst im Monat April jeweils mittags um 12 Uhr die Außentemperatur und trägt sie in eine Tabelle ein.

Berechnen Sie den Mittelwert, die Spannweite und den Median.

Berechnen Sie das 1. und 3. Quartil und den Quartilsabstand.

Tag	1	2	3	4	5	6	7	8	9	10
Temperatur	7	10	12	16	16	17	18	20	22	29
Tag	11	12	13	14	15	16	17	18	19	20
Temperatur	23	19	20	21	18	17	15	29	22	23
Tag	21	22	23	24	25	26	27	28	29	30
Temperatur	8	25	24	23	23	25	26	27	19	16

Mittelwert: $\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{1}{30}(7+10+\ldots+19+16) = \frac{590}{30} = \underline{\underline{19,\overline{6}}}$

0 7 8

1 0 2 5 6 6 6 7 7 8 8 9 9

2 0 0 1 2 2 3 3 3 3 4 5 5 6 7 9 9

Spannweite: $R = x_{max} - x_{min} = 29 - 7 = \underline{\underline{22}}$

Median: $x_{Med} = \frac{1}{2}(x_{15} + x_{16}) = \frac{1}{2}(20+20) = \underline{\underline{20}}$ (2. Quartil)

1. Quartil: $Q_1 = x_8 = \underline{\underline{16}}$

3. Quartil: $Q_3 = x_{23} = \underline{\underline{23}}$

Quartilsabstand: $Q_A = Q_3 - Q_1 = 23 - 16 = \underline{7}$

Quantile, Perzentile, Quartile, Dezile und Zentile

Für den gesamten Wertebereich eines Merkmals lässt sich an Hand der Tabelle der kumulierten prozentualen Häufigkeiten abschätzen, wie viel Prozent aller untersuchten Fälle unterhalb eines bestimmten Wertes liegen. Diesen Wert bezeichnet man allgemein als **Quantil** bzw. bei Verwendung von prozentualen Anteilen als **Perzentil**.

Die allgemeine Definition des p-Quantils für $0 < p < 1$ lautet:

Definition 42:

Das p-Quantil (Perzentil) x_p ist der Wert, für den gilt, dass mindestens p * 100% der Werte kleiner oder gleich und mindestens (1-p) * 100% größer oder gleich dem p-Quantil sind.

Beispiel 20:

Für p=0,55 sind 55% aller Messwerte kleiner oder gleich dem 55. Quantil und 45% größer oder gleich.

Das 50. Quantil ist die Grenze zwischen der unteren und oberen Hälfte aller Werte einer Stichprobe. Es ist der Median.

Bemerkung 24:

- Weitere häufig verwendete Werte sind das 25%- und 75%-Perzentil, die das untere und das obere Viertel der Verteilung markieren. Man bezeichnet sie daher auch als **untere und obere Quartile** bzw. als erstes und drittes Quartil (der Median ist das zweite Quartil). Als Schreibweise sind Q_1, Q_2 und Q_3 ebenso möglich wie Q_{25}, Q_{50} und Q_{75}.
- Von **Dezilen** spricht man, wenn die kumulierte Häufigkeitsverteilung in 10%-Abstände gegliedert wird. Das erste Dezil (D_1) markiert die Grenze zwischen den unteren 10% und den oberen 90% der Messwerte. Beim neunten Dezil (D_9) ist es genau umgekehrt. Das fünfte Dezil (D_5) ist der Median. Gebräuchlich sind für die 10%-Abstände auch die Bezeichnungen Centile oder Dezentile. Das neunte Centil wird mit C_{90} bezeichnet, dass fünfte als C_{50}.

Quartile und Dezile sind spezielle Quantile (Perzentile), die Aufteilung der kumulierten Häufigkeitsverteilung in gleich große Intervalle. Um den Einfluss einzelner Ausreißer an den Rändern des Wertespektrums auszuschalten, wird zumeist der **Quartilsabstand**, gelegentlich auch der Dezilabstand statt der Spannweite errechnet.

Quartilsabstand und Dezilabstand

Der Quartilsabstand (QA), auch als **Interquartilsabstand** oder Interquartilsbereich bezeichnet, ist die Differenz aus dem oberen (Q_3) und dem unteren Quartil (Q_1).

Der Quartilsabstand gibt somit das Ausmaß des Bereiches an, in dem die mittleren 50% der Beobachtungswerte liegen. Die Formel ist:

Definition 43:

Quartilsabstand = $Q_3 - Q_1$

Der weniger gebräuchliche Dezilabstand (DA) ist die Differenz zwischen dem 90%-Dezil (D_9) und dem 10%-Zentil (D_1). Er beinhaltet also die mittleren 80 Prozent der Werte einer Variablen. Als Formel:

Definition 44:

Dezilabstand = $D_9 - D_1$

Im Unterschied zur Spannweite sind Quartils- und Dezilabstand unabhängig von Extremwerten. Beide Werte können ab Ordinalskalenniveau berechnet werden.

In der Praxis wird gelegentlich auch mit der Hälfte von Quartils- und Zentilabstand operiert.

Weitere Auswertung des obigen Beispiels:

Quartilsabstand: $Q_A=Q_3-Q_1=1{,}755-1{,}675=0{,}08$

50% der Daten liegen in einem Bereich der Bandbreite von 0,08m bzw. 8cm.

Etwa 50% der Körpergrößen liegen zwischen 1,675m und 1,755m.

Vergleich zwischen Quartilsabstand und Spannweite

Bemerkung 25:

Quartilsabstand

- Von Ausreißern unabhängig.
- Gibt die Breite des mittleren Bereichs an, in dem ca. 50% aller Werte liegen. Vom kleinsten und größten Wert abhängig.

Spannweite

- Gibt die Gesamtbreite an in dem alle Werte liegen

Durchschnittliche Abweichung

Auch wenn in verschiedenen Stichproben eines Merkmals Lagewerte und Spannweite identisch sind, können sich die Verteilungen der Merkmale voneinander unterscheiden. Die Abweichung aller Messwerte vom Mittelwert der Verteilung muss berechnet werden, um die Streuung zu erfassen.

Definition 45:

Die durchschnittliche Abweichung ist der Mittelwert der in absoluten Beträgen gemessenen Abweichung aller Messwerte vom arithmetischen Mittel der Häufigkeitsverteilung einer Variablen.

$$X_D = \frac{\sum_{i=1}^{n} \left| \left(x_i - \bar{x} \right) \right|}{n}$$

Bemerkung 26:

- Man nimmt den absoluten Betrag der Differenz jedes Messwertes vom Mittelwert, da andernfalls die negativen Abweichungen unterhalb des arithmetischen Mittels die positiven Abweichungen aufheben würden, so dass die Summe aller Abweichungen bei Null-Läge.
- Wir könnten die durchschnittliche Abweichung aller Messwerte vom Mittelwert berechnen. Bei der schlichten Berechnung der Summe aller Abweichungen vom Mittelwert, dividiert durch n, heben sich die negativen und positiven Abweichungen gegenseitig auf.
- Die Summe der Beträge der Differenzen vom arithmetischen Mittel dividiert man durch die Anzahl der Messwerte.
- Zulässig ist die Berechnung der durchschnittlichen Abweichung für metrische Skalen, also Intervall- und Verhältnisskalen.
- Der wichtigste Unterschied der durchschnittlichen Abweichung zu den Quantilen ist, dass der Mittelwert der Häufigkeitsverteilung (und nicht die Ränder der Verteilung) Bezugspunkt für die Berechnung der Streuung ist.
- Außerdem geht jeder Wert in die Berechnung ein.

Beides gilt auch für die wichtigsten Streuungsmaße, Varianz und Standardabweichung. Sie haben gegenüber der durchschnittlichen Abweichung den Vorteil, dass sie in die Formeln zur Berechnung weiterer Kennwerte einfließen, insbesondere in die Formeln der Zusammenhangsmaße (Korrelation, Regression). In der Praxis wird heute statt der MAD die Standardabweichung einer Verteilung angegeben.

Die durchschnittliche Abweichung hat als Dispersionsmaß den Nachteil, dass sie nur eine geringe Stabilität aufweist und auf die tatsächliche Streuung in der Grundgesamtheit nur schwierig rückgeschlossen werden kann.

Die durchschnittliche Abweichung charakterisiert die Verteilung der Messwerte um das arithmetische Mittel.

Bemerkung 27:

- Da die Summe der Abweichungen der Messwerte von ihrem arithmetischen Mittel immer gleich Null ist, müssen die negativen Vorzeichen ausgeschaltet werden.

Beispiel 21:

Variable „Alter“				
x_i	f_i	$f_i x_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \cdot \lvert x_i - \bar{x} \rvert$
1	1	1	$\lvert 1-5 \rvert = 4$	$1 \cdot 4 = 4$
3	1	3	$\lvert 3-5 \rvert = 2$	$1 \cdot 2 = 2$
6	1	6	$\lvert 6-5 \rvert = 1$	$1 \cdot 1 = 1$
7	1	7	$\lvert 7-5 \rvert = 2$	$1 \cdot 2 = 2$
8	1	8	$\lvert 8-5 \rvert = 3$	$1 \cdot 3 = 3$
$\sum$	$N = 5$	25	12	12

$$\bar{x} = \frac{25}{5} = 5$$

$$AD = \frac{12}{5} = 2{,}4$$

Interpretation:

- Ein AD-Wert von 2,4 besagt, dass die Messwerte im Durchschnitt 2,4 Einheiten von ihrem arithmetischen Mittel abweichen.

Auf die Variable „Alter“ bezogen bedeutet dies:

- Die Messwerte weichen durchschnittlich um 2,4 Jahre vom Altersdurchschnitt (x = 5 Jahre) ab.

Varianz

Definition 46:

Die Varianz (s^2 oder σ^2) basiert auf den Quadraten der Abstände der Messwerte vom Mittelwert. Sie ist der Durchschnitt aller quadrierten Abweichungen der einzelnen Messwerte vom arithmetischen Mittel.

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

Liegen die Werte bereits als Häufigkeitstabelle vor, operiert man mit den Häufigkeiten in den beiden Formeln. Die ergänzten Formeln lauten:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} f_i \cdot (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} h_i \cdot (x_i - \bar{x})^2$$

Bemerkung 28:

- Man berechnet für jeden Messwert den Abstand vom Mittelwert, quadriert diesen und summiert die quadrierten Abstände. Diese Summe wird durch die Gesamtzahl der Messwerte geteilt. Mit den nicht gruppierten Werten, der Urliste als Grundlage, ergibt sich die oben genannte Formel:
- Die Berechnung der Varianz ist nur bei metrischen Skalen zulässig. Im Unterschied zur durchschnittlichen Abweichung fallen mit der Quadrierung größere Abweichungen stärker ins Gewicht als kleinere.
- Die Maßeinheit der Varianz ist schwer interpretierbar, da sie nicht mehr der Maßeinheit der gemessenen Variablen entspricht. Ermittelt man z.B. die Varianz der Variablen Körpergröße (gemessen in Metern) in einer Stichprobe, hat die Varianz die Maßeinheit Quadratmeter. Sie ist ein Flächenmaß. Zieht man jedoch die Quadratwurzel aus der Varianz, ergibt sich wieder die ursprüngliche Maßeinheit. Zugleich gelangt man zum wichtigsten Streuungsmaß, der Standardabweichung.

Beispiel 22:

Variable „Alter“

x_i	f_i	$f_i x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
1	1	1	1-5 = -4	16	1 · 16 = 16
3	1	3	3-5 = -2	4	1 · 4 = 4
6	1	6	6-5 = 1	1	1 · 1 = 1
7	1	7	7-5 = 2	4	1 · 4 = 4
8	1	8	8-5 = 3	9	1 · 9 = 9
$\sum$	N = 5	25	0	34	34

$$\bar{x} = \frac{25}{5} = 5$$

$$s^2 = \frac{34}{5} = 6{,}8$$

Interpretation:

- Ein s^2-Wert von 6,8 besagt, dass die Messwerte im Durchschnitt 6,8 Quadrat-Einheiten von ihrem arithmetischen Mittel abweichen.

Auf die Variable „Alter“ bezogen bedeutet dies:

- Die Messwerte weichen durchschnittlich um 6,8 Quadrat-Jahre vom Altersdurchschnitt (x = 5 Jahre) ab.

Beispiel 23:

Wir betrachten noch mal die Notenverteilung von Mädchen und Jungen.

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	
Note Mädchen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6	$\bar{x} = 3,0$
Note Jungen	1,0	1,0	2,0	2,5	3,2	2,8	3,5	2,0	6,0	6,0	$\bar{x} = 3,0$

Mädchen				Jungen			
i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	3,2	0,2	0,04	1	1,0	−2,0	4,0
2	3,5	0,5	0,25	2	1,0	−2,0	4,0
3	2,9	−0,1	0,01	3	2,0	−1,0	1,0
4	3,3	0,3	0,09	4	2,5	−0,5	0,25
5	3,4	0,4	0,16	5	3,2	0,2	0,04
6	2,5	−0,5	0,25	6	2,8	−0,2	0,04
7	2,7	−0,3	0,09	7	3,5	0,5	0,25
8	2,8	−0,2	0,04	8	2,0	−1,0	1,0
9	3,1	0,1	0,01	9	6,0	3,0	9,0
10	2,6	−0,4	0,16	10	6,0	3,0	9,0
$\sum$	30	0	1,10	$\sum$	30	0	28,58

Varianz Mädchen: $s_M^2 = \frac{1}{10} \cdot 1,1 = \underline{\underline{0,11}}$ Varianz Jungen: $s_J^2 = \frac{1}{10} \cdot 28,58 = \underline{\underline{2,858}}$

Viele Daten sind mit Einheiten behaftet, z.B. Meter (m) oder kg.

Die Einheit für die Varianz wäre in diesen Fällen m^2 bzw. $(kg)^2$.

Um wieder auf die ursprüngliche Einheit zu kommen, zieht man die Wurzel aus der Varianz.

Dieser Wert wird Standardabweichung genannt.

Standardabweichung

Definition 47:

Die Standardabweichung ist die positive Quadratwurzel der Varianz. Sie ist das gebräuchlichste Maß zur Kennzeichnung der Variabilität einer Verteilung. Ausgehend von der Urliste lautet die Formel:

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

Bemerkung 29:

- Selbstverständlich können auch die anderen Formeln der Varianz, ergänzt um das Wurzelzeichen, angewendet werden. Wie bei der Varianz ist mindestens das Intervallskalenniveau Voraussetzung für die Berechnung der Standardabweichung eines Merkmals.
- Die Standardabweichung ist inhaltlich unmittelbar interpretierbar, da ihre Maßeinheit der Maßeinheit der gemessenen Variablen entspricht.

Beispiel 24:

Note (x_i)	1	2	3	4	5	6
Anz. d. Schüler (n_i)	5	8	14	16	5	2

Berechnung der Gesamtzahl aller Schüler aus den absoluten Häufigkeiten n_i:

Note (x_i)	1	2	3	4	5	6
Anz. d. Schüler (n_i)	5	8	14	16	5	2

Schüler insgesamt: $n = \sum_{i=1}^{6} n_i = 50$

Berechnung der Varianz über die absolute Häufigkeit:

i	x_i	n_i	$x_i \cdot n_i$	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 \cdot n_i$
1	1	5	5	3,28	−2,28	25,992
2	2	8	16	3,28	−1,28	13,1072
3	3	14	42	3,28	−0,28	1,0976
4	4	16	64	3,28	0,72	8,2944
5	5	5	25	3,28	1,72	14,792
6	6	2	12	3,28	2,72	14,7968
$\sum$		50	164	$\bar{x} = \frac{164}{50} = 3,28$		78,08

Varianz: $s^2 = \frac{1}{50}\sum_{i=1}^{6}(x_i - \bar{x})^2 \cdot n_i = \frac{78,08}{50} = \underline{\underline{1,5616}}$

Standardabweichung: $s = \sqrt{s^2} = \sqrt{1,5616} = \underline{\underline{1,2496}}$

Beispiel 25:

Note (x_i)	1	2	3	4	5	6
Anz. d. Schüler (n_i)	5	8	14	16	5	2
rel. Häufigkeit $h_i = \frac{n_i}{n}$	0,1	0,16	0,28	0,32	0,1	0,04

Schüler insgesamt: $n = \sum_{i=1}^{6} n_i = 50$

Berechnung der Varianz über die relative Häufigkeit:

i	x_i	h_i	$x_i \cdot h_i$	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 \cdot h_i$
1	1	0,1	0,1	3,28	−2,28	0,51984
2	2	0,16	0,32	3,28	−1,28	0,262144
3	3	0,28	0,84	3,28	−0,28	0,021952
4	4	0,32	1,28	3,28	0,72	0,165888
5	5	0,1	0,50	3,28	1,72	0,29584
6	6	0,04	0,24	3,28	2,72	0,295936
$\sum$		1	$\bar{x} = 3,28$			$s^2 = 1,5616$

Varianz: $s^2 = \sum_{i=1}^{6} (x_i - \bar{x})^2 \cdot h_i = \underline{\underline{1,5616}}$

Standardabweichung: $s = \sqrt{s^2} = \sqrt{1,5616} = \underline{\underline{1,2496}}$

Beispiel 26:

Bestimmen Sie aus der klassierten Häufigkeitstabelle für die Körpergröße die Standardabweichung.

Klasse x_i	$150 \le x < 160$	$160 \le x < 170$	$170 \le x < 180$	$180 \le x < 190$
abs. Häufigkeit n_i	9	12	7	2
Klassenmitte m_i	155	165	175	185
rel. Häufigkeit $h_i = \frac{n_i}{n}$	0,3	0,4	$0,2\overline{3}$	$0,0\overline{6}$

$$\text{Klassenmitte} = \frac{\text{Klassenanfang} + \text{Klassenende}}{2} \quad \text{z.B.} \quad \frac{160+170}{2} = 165$$

Berechnung über die absolute Häufigkeit:

i	m_i	n_i	$m_i \cdot n_i$	$\bar{x}$	$m_i - \bar{x}$	$(m_i - \bar{x})^2 \cdot n_i$
1	155	9	1395	$165,\overline{6}$	$-10,\overline{6}$	$1023,\overline{9}$
2	165	12	1980	$165,\overline{6}$	$-0,\overline{6}$	$5,\overline{3}$
3	175	7	1225	$165,\overline{6}$	$9,\overline{3}$	$609,\overline{7}$
4	185	2	370	$165,\overline{6}$	$19,\overline{3}$	$747,\overline{5}$
$\sum$		n = 30	4970	$\frac{4970}{30} = 165,\overline{6}$		$2386,\overline{6}$

Varianz: $s^2 = \frac{2386,\overline{6}}{30} = \underline{\underline{79,\overline{5}}}$ Standardabweichung: $s = \sqrt{79,\overline{5}} \approx \underline{\underline{8,9194}}$

Berechnung über die relative Häufigkeit:

i	m_i	h_i	$m_i \cdot h_i$	$\bar{x}$	$m_i - \bar{x}$	$(m_i - \bar{x})^2 \cdot h_i$
1	155	0,3	46,5	$165,\overline{6}$	$-10,\overline{6}$	$34,1\overline{3}$
2	165	0,4	66,0	$165,\overline{6}$	$-0,\overline{6}$	$0,1\overline{7}$
3	175	$0,2\overline{3}$	$40,8\overline{3}$	$165,\overline{6}$	$9,\overline{3}$	20,3259
4	185	$0,0\overline{6}$	$12,\overline{3}$	$165,\overline{6}$	$19,\overline{3}$	24,91
$\sum$		1	$165,\overline{6}$			$79,\overline{5}$

Varianz: $s^2 = \underline{\underline{79,\overline{5}}}$ Standardabweichung: $s = \sqrt{79,\overline{5}} \approx \underline{\underline{8,9194}}$

Bemerkung 30:

- Standardabweichung und Varianz sind grundsätzlich als gleichwertige Streuungsmaße anzusehen, denn wenn die Varianz groß (klein) ist, ist auch die Standardabweichung groß(klein).
- Für deskriptive Zwecke ist allerdings die Standardabweichung vorzuziehen, weil sie ein Kennwert in der Einheit der zugrunde liegenden Messwerte ist.

Variationskoeffizient (Variabilitätskoeffizient)

Definition 48:

Der Variationskoeffizient misst die Variation im Vergleich zum Mittelwert.

Der Varianzkoeffizient relativiert die Standardabweichung am Mittelwert. Der Variationskoeffizient drückt die Standardabweichung in Mittelwertseinheiten aus.

$$V = \frac{s}{\bar{x}} \cdot 100\,\%$$

Bemerkung 31:

- Dieses Maß wird gelegentlich eingesetzt, wenn Streuungen von Verteilungen mit unterschiedlichen Mittelwerten zu vergleichen sind und Mittelwert und Streuung voneinander abhängen.
- Dazu wird die Standardabweichung durch das arithmetische Mittel geteilt. In der Forschungspraxis bevorzugt man die Angabe des Variabilitätskoeffizienten in Prozentanteilen des Mittelwertes.
- Das Ergebnis ist also noch mit 100% zu multiplizieren.
- Die Standardabweichung hängt ab vom Wert des arithmetischen Mittels.
- Der Variationskoeffizient relativiert diese Abhängigkeit. Er ermöglicht den **Vergleich von Streuungen** zwischen Gruppen, die sich im absoluten Wert von Mittelwert und Streuung unterscheiden.
- Variationskoeffizienten sind wenig aussagekräftig, wenn die Datenreihe etwa gleich viele negative wie positive Werte aufweist. Der Mittelwert liegt dann nahe bei null. Der Variationskoeffizient wird unangemessen hoch, da man in der Formel durch eine Dezimalzahl dividiert. Ist der Mittelwert gleich Null, ist die Berechnung mathematisch verboten. Durch Null darf niemals dividiert werden.

Erklärung:

Der Variationskoeffizient ist in erster Linie ein Streuungsmaß, wird aber auch als Konzentrationsmaß gewählt.

Man teilt die Standardabweichung durch das arithmetische Mittel der Verteilung. Durch dieses Teilen/Normieren erhält man eine dimensionslose Maßzahl. Die Standardabweichung gibt ja eine ungefähre absolute Abweichung vom arithmetischen Mittel an. Teilt man nochmal durch das arithmetische Mittel, erhält man so etwas wie die "relative Standardabweichung".

Hoher Variationskoeffizient: hohe Streuung

Niedriger Variationskoeffizient: geringe Streuung

Beispiel 27:

Haushaltjahreseinkommen in den Ländern A und B

Land A:	50.000	50.000	50.000
Land B:	1.000	1.000	148.000

$\bar{x}_A = \bar{x}_B = 50.000, da \frac{150.000}{3} = 50.000$

Land A				Land B			
x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i \cdot (x_i - \bar{x})^2$
50.000	0	0	0	1.000	-49.000	2.401.000.000	2.401.000.000
50.000	0	0	0	1.000	-49.000	2.401.000.000	2.401.000.000
50.000	0	0	0	148.000	98.000	9.604.000.000	9.604.000.000
$\sum$	0	0	0	$\sum$	0	14.406.000.000	14.406.000.000

$s_A = \sqrt{\frac{0}{3}} = \sqrt{0} = 0$ und $s_B = \sqrt{\frac{14.406.000.000}{3}} = \sqrt{4.802.000.000} = 69.296$

Interpretation:

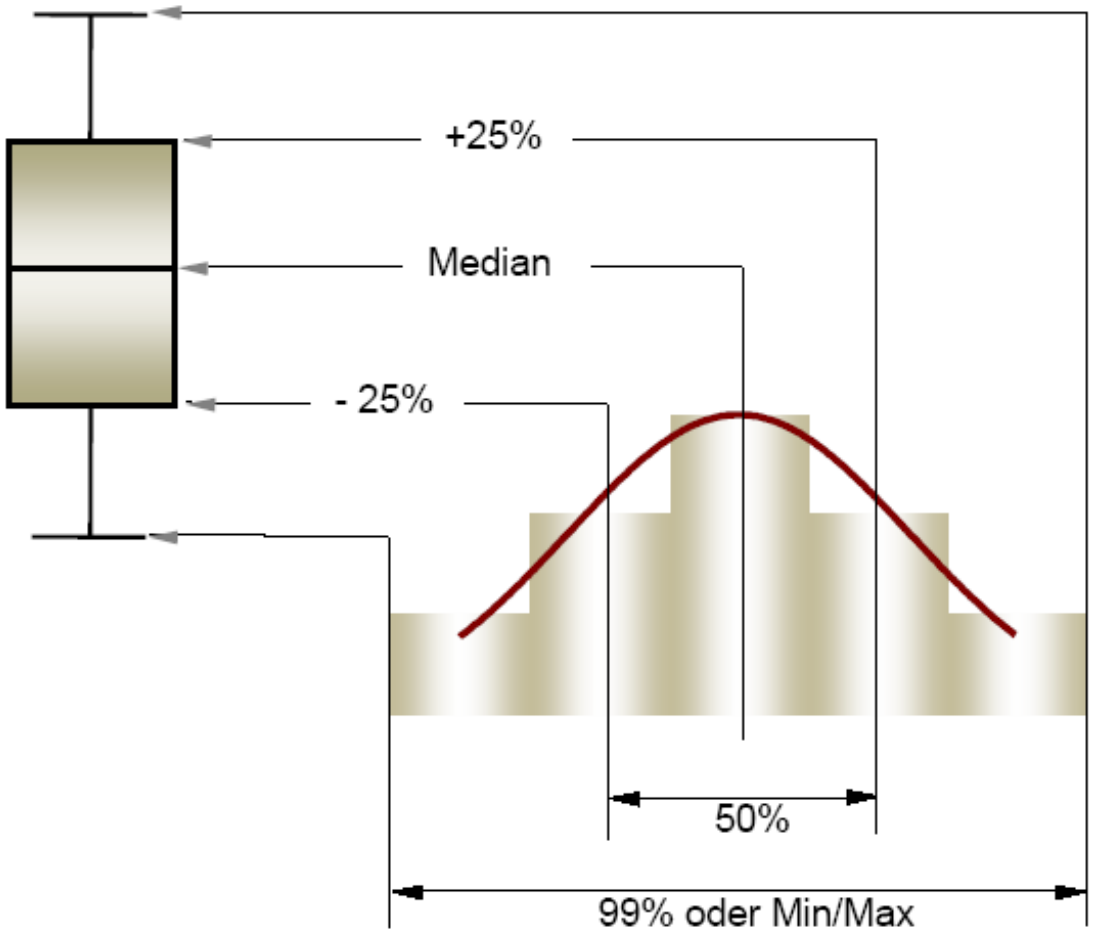
- In beiden Ländern streut, gemessen am Durchschnitt, das Haushaltseinkommen ungleich. Die relative Streuung ist für das Land B größer als für das Land A.

Boxplot als graphische Darstellung von Streuungsparametern

Definition 49:

Der Boxplot ist eine spezielle Art der Häufigkeitsverteilung. Bei ihm wird die Wertean-stelle der X-Achse über die Y-Achse dargestellt, wobei mehrere Boxplots neben einander in einem Diagramm möglich sind.

In der Mitte des Boxplots befindet sich eine Linie mit dem sogenannten Zentralwert bzw. Median. Optional kann auch der Mittelwertgewählt werden. Innerhalb des Bauches befinden sich 50% aller Werte. Innerhalb der äußeren Begrenzungs-Linien oben und unten befinden sich 99% aller Werte. Wahlweisekann auch der kleinste und größte vorkommenden Wert angezeigt werden (Min/Max-Werte). Sind zu wenige Datenwerte vorhanden, entsprechen die 99% Bereiche denen der Min/Max-Werte.



Man erhält hier einen schnellen Überblick über die einzelnen Werte.

Möchten Forscher in der grafischen Darstellung der Häufigkeitstabelle eines Merkmals Unterschiede in der Streuung zwischen verschiedenen Gruppen oder Stichproben hervorheben, bietet sich als Alternative zu den üblicherweise benutzten Diagrammen der sogenannte Box-and-Whiskers-Plot (kurz: Box-Plot) an.

Dies ist eine graphische Darstellung, die Verteilung und Streuungswerte miteinander verbindet.

Bemerkung 32:

- Der Boxplot besteht aus zwei umgekehrt T-förmigen Endpunkten und einem Kasten, der den Quartilsabstand, also die mittleren die 50% der Werte, umfasst.
- Die Linie in der Boxenmitte gibt die Lage des Medians bzw. des zweiten Quartils an.
- Die beiden gespiegelten T-Punkte oben und unten zeigen die untersten und obersten 25 Prozent der Werte an.
- Je länger die Box ist, desto stärker streuen die Beobachtungswerte im mittleren Bereich.
- Je länger die T-Punkte sind, desto stärker streuen die Ränder der Verteilungen.

Bemerkung 33:

Die Fünf-Punkte-Zusammenfassung einer Verteilung, bestehend aus

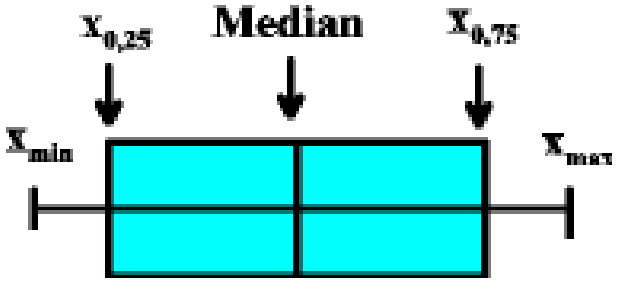
- x_{min}, $x_{0,25}$, x_{Median}, $x_{0,75}$, x_{max}

führt zu einer graphischen Darstellung der Verteilung als Box-Plot. In der zu konstruierenden Box entspricht

- $x_{0,25}$ dem Anfang,
- $x_{0,75}$ dem Ende und
- $x_{0,75} - x_{0,25}$, also der Interquartilabstand

die Länge der Box.

Der Median wird als Punkt oder Strich in der Box und x_{min} und x_{max} als Linien außerhalb der Box dargestellt:



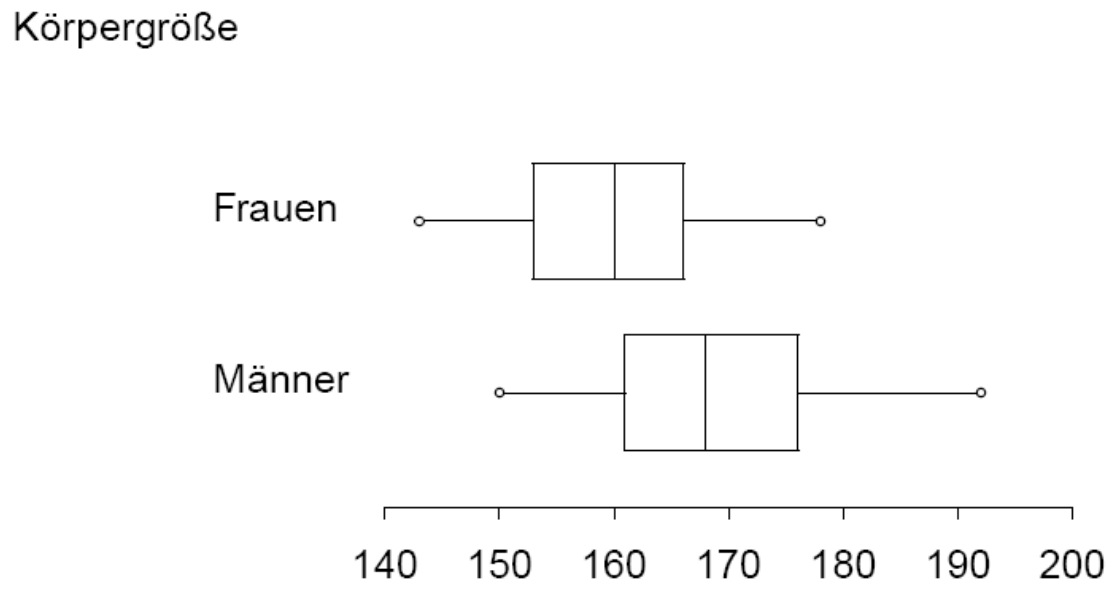
Ein Box-Plot zeigt die Lage und die Streuung einer Verteilung an.

Über den Box-Plot lassen sich verschiedene Verteilungen vergleichen und es kann sehr schnell ein visueller Eindruck gewonnen werden, ob

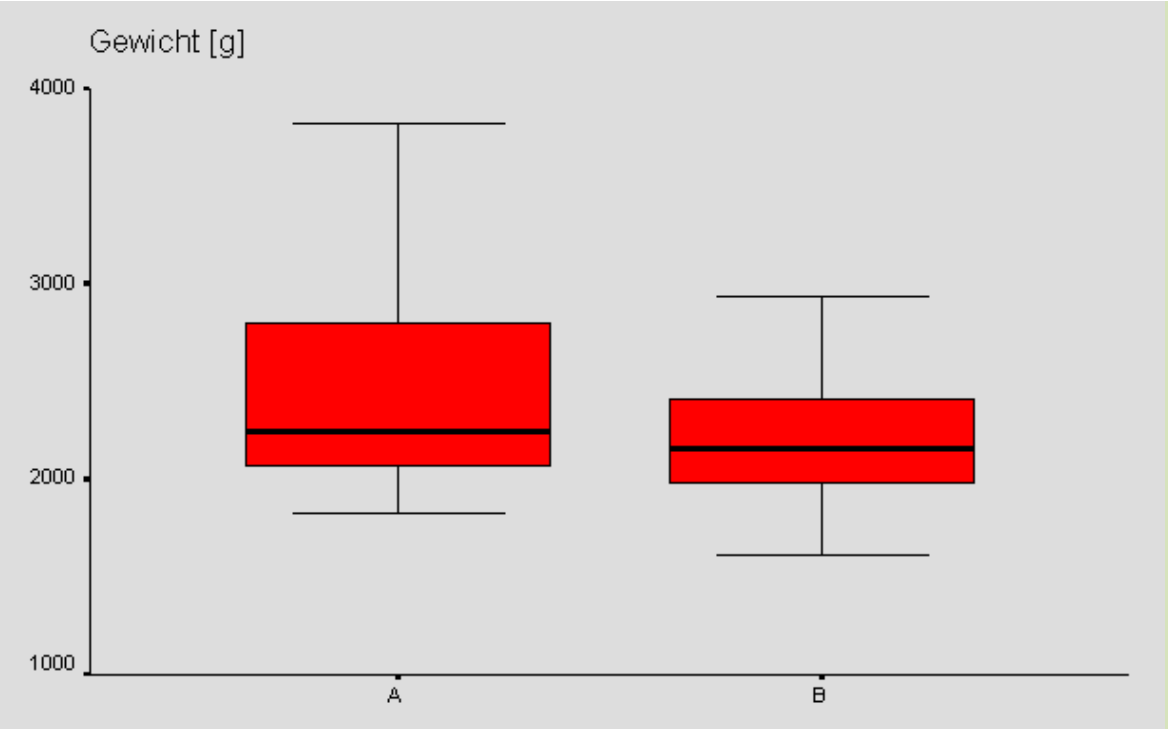
Bemerkung 34:

- Verteilung symmetrisch ist,
- oder ob Extremwerte vorliegen.

Beispiel 28:



Beispiel 29:



Am Beispiel wird dies deutlich. Während in der rechts dargestellten Gruppe B das hier untersuchte Gewicht nahezu normalverteilt ist, finden sich in der Gruppe A überproportional häufig groß Werte. Der Median in beiden Gruppen unterscheidet sich hingegen kaum.

Verteilungsformen

Im vorangegangenen Kapitel wurden Maße der zentralen Tendenz eingeführt, die Stellung der drei Maße Modus, Median, Arithmetisches Mittel ist abhängig von der Verteilungsform. Unterschiedliche Formen der Verteilung ergeben sich durch die Art der Häufigkeitsverteilung.

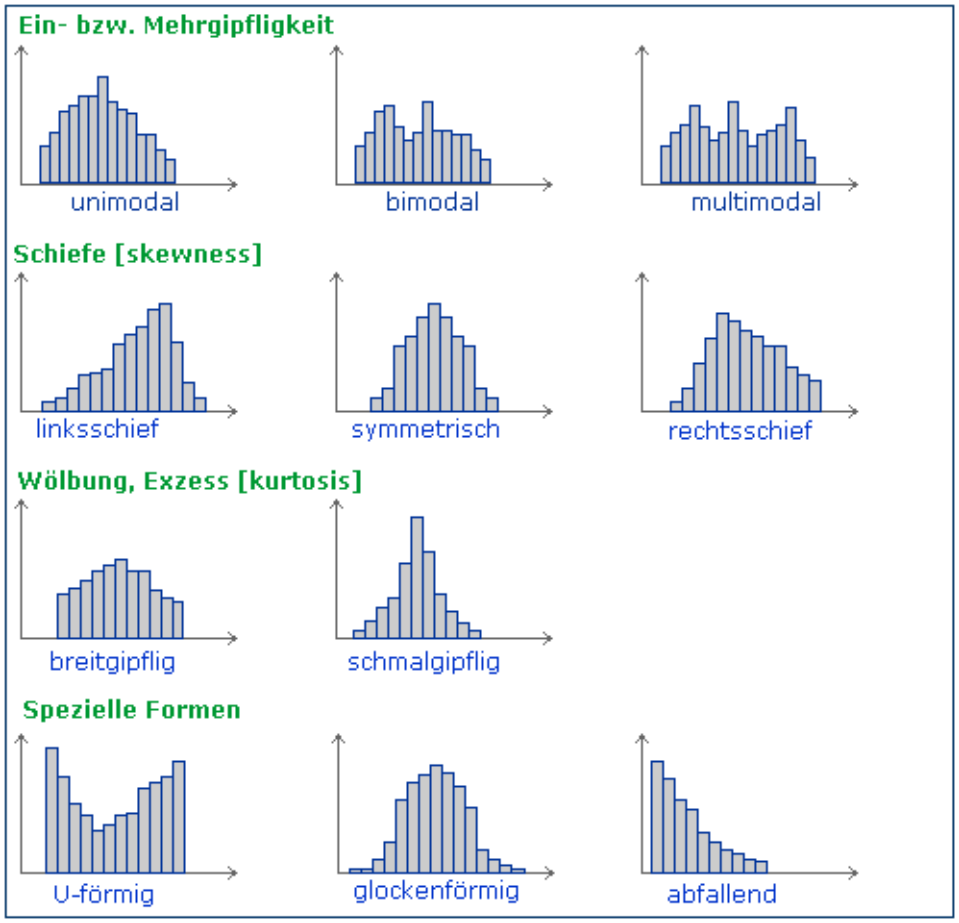
Oft findet man symmetrische, glockenförmige Verteilungen, die ihren Gipfel in der Verteilungsmitte haben. Hier kommen Messwerte im mittleren Bereich sehr häufig vor, hingegen extreme Messwerte eher selten.

Es treten jedoch auch asymmetrische Formen auf, deren Gipfel etwas nach links, hier liegen viele niedrige Messwerte und wenig hohe Messwerte vor, verschoben ist. Liegt der Gipfel eher rechts, handelt es sich um wenig niedrige und viele hohe Werte.

Eine weitere Form ist die bimodale Verteilung mit vielen hohen und niedrigen Extremwerten, bei denen darauf zu achten ist, dass einige statistische Kennwerte nicht zur Anwendung kommen dürfen.

Hier eine Aufzählung der verschiedenen Verteilungsformen:

- symmetrisch/asymmetrisch
- unimodal (eingipflig) / bimodal (zweigipflig)
- schmalgipflig / breitgipflig
- linkssteil / rechtssteil
- U-förmig / abfallend

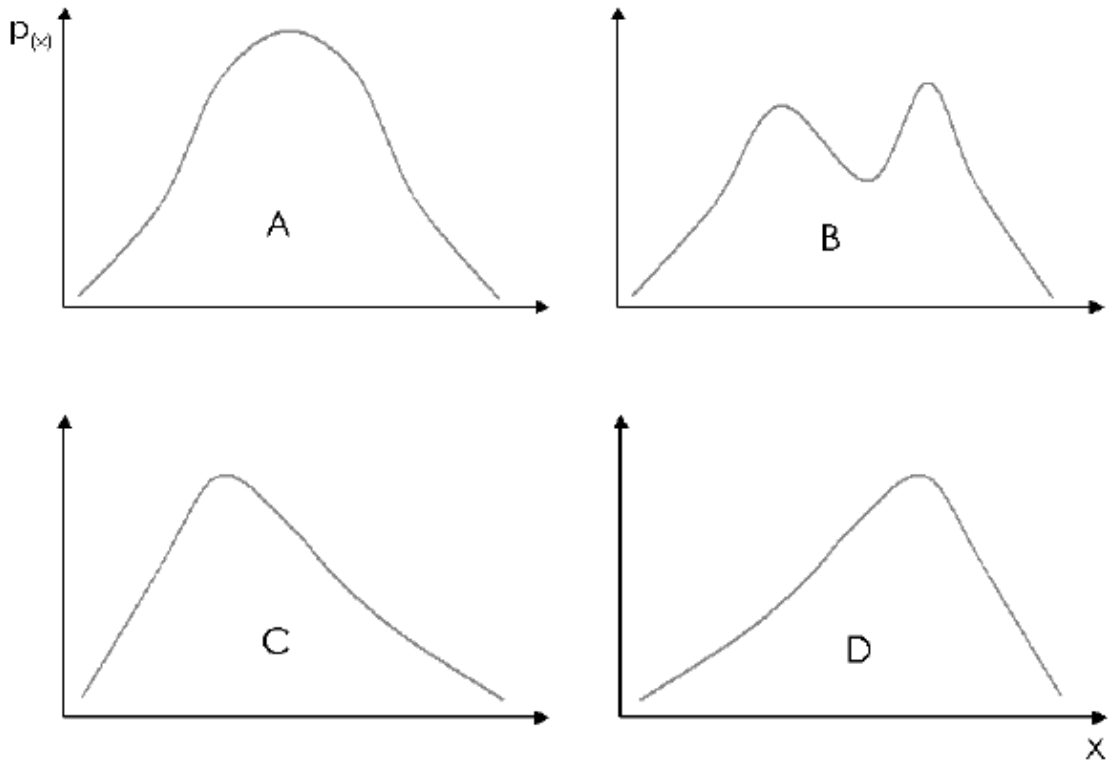


Bemerkung 35:

- Aus der Position der Lagemaße zueinander wird ersichtlich, ob eine Verteilung symmetrisch oder schief ist.

Beispiel 30:

Beschreiben Sie folgende Verteilungen:



Beschreibungen:

A: normalverteilt, unimodal (eingipflig), symmetrisch

B: asymmetrisch, bimodal, zweigipflig

C: asymmetrisch, unimodal, linkssteil, rechtsschief

D: asymmetrisch, unimodal, rechtssteil, linksschief

Schiefe

Definition 50:

Steigt eine Verteilung auf einer Seite steiler an, als auf der anderen Seite, wird sie als schief bezeichnet. Schiefe ist ein Maß für die Asymmetrie

$$\text{Schiefe} = \frac{\bar{x} - \text{Modalwert}}{s}$$

Die Schiefe einer Verteilung gibt an, ob sich die Werte normal verteilen oder in eine Richtung der Skala tendieren.

Bemerkung 36:

- Eine **linkssteile Verteilung** (**Schiefe < 0**) liegt vor, wenn der Modalwert kleiner ist als der Median einer Verteilung; die Schiefe ist in diesem Fall kleiner als 0.
- Eine **rechtssteile Verteilung** (**Schiefe > 0**) liegt vor, wenn der Modalwert größer ist als der Median einer Verteilung; die Schiefe ist in diesem Fall größer als 0.
- Eine **symmetrische Verteilung** (**Schiefe = 0**) liegt vor, wenn der Modalwert, der Mittelwert und der Median einer Verteilung gleich sind; die Schiefe ist in diesem Fall gleich 0.

Wahrscheinlichkeitsrechnung

Geschichte

Schon in der Antike tritt der Gedanke auf, dass die Naturgesetze durch eine sehr große Anzahl von zufälligen Ereignissen zur Geltung kommen, z.B. in dem Gedicht „De rerum natura“ von Lukrez.

Die Aufdeckung der Gesetzmäßigkeiten, auf deren Auftreten zahlreiche individuelle Einflüsse einwirken, die nicht, oder fast nicht miteinander verbunden sind, war auch das Ziel jener Gelehrten, die die Entstehung der Wahrscheinlichkeitsrechnung wesentlich beeinflussten. So schreib Huygens 1657, dass er sich nicht nur mit Spielen beschäftige, sondern dass er die Grundlagen einer „tiefsinnigen und hochinteressanten neuen Theorie“ vortrage.

Die mit Glücksspielen zusammenhängenden Probleme bildeten den Anlass dafür, dass sich bedeutende Gelehrte mit Fragen der Zufälligkeit von Ereignissen u.a. beschäftigen. Die eigentlichen Ursachen liegen jedoch in der Herausbildung frühkapitalistischer Wirtschaftsverhältnisse und den dabei auftretenden Fragestellungen z.B. im Versicherungswesen, der Bevölkerungsstatistik und der Auswertung von Beobachtungen.

Zufällige Erscheinungen

Zur Erzeugung von Stichproben

Statistische Erhebungen erstrecken sich - wie bereits erwähnt - meist nicht auf alle Merkmalsträger, sondern nur auf eine gewisse Auswahl derselben. Die gewonnenen Daten bilden dann eine Stichprobe aus der Grundgesamtheit. So wird man z. B., um die Altersstruktur der Bevölkerung eines Landes zu untersuchen, nicht das Alter sämtlicher Einwohner feststellen, sondern sich mit einer Stichprobe begnügen.

Definition 51:

Von einer solchen Stichprobe erwartet man, dass sie ein mehr oder weniger getreues Abbild der Grundgesamtheit darstellt.

Bemerkung 37:

- Im genannten Beispiel erwartet man also, dass die in der Stichprobe vorliegende Altersverteilung ungefähr mit der Altersverteilung in der Gesamtbevölkerung des Landes übereinstimmt.
- Diese Erwartung wäre gewiss nicht gerechtfertigt, wenn die statistische Erhebung z. B. ausschließlich in ländlichen Gegenden oder ausschließlich unter Straßenpassanten durchgeführt würde. Eine solche Erwartung ist allenfalls in den beiden folgenden Fällen gerechtfertigt.

Bemerkung 38:

Erster Fall:

Man versucht, durch gezielte Auswahl der Merkmalsträger für eine Stichprobe gewissermaßen ein verkleinertes Abbild der Grundgesamtheit zu konstruieren.

Solche repräsentativen Stichproben werden z. B. in der Meinungsforschung verwendet; stellvertretend für die Gesamtheit des interessierenden Bevölkerungskreises werden etwa 1000 bis 3000 Personen befragt, die sich nach Geschlecht, Familienstand, Kinderzahl, Religionszugehörigkeit, Beruf, Wohnverhältnissen usw. entsprechend zusammensetzen wie die Gesamtbevölkerung.

Die Erzeugung einer brauchbaren repräsentativen Stichprobe ist meist mit erheblichen zeitlichen, technischen und finanziellen Aufwendungen verbunden.

Zweiter Fall:

Hier wird eine in gewissem Sinne entgegen gesetzte Methode angewendet; sie besteht darin, die Merkmalsträger für eine Stichprobe nach dem Zufallsprinzip auszuwählen.

Um beispielsweise aus der 500 Schüler umfassenden Sekundarstufe I einer Schule 20 Schüler nach dem Zufallsprinzip auszuwählen, könnte man vorgehen wie bei einer Verlosung: Von 500 Zetteln werden 20 angekreuzt, danach werden die Zettel gerollt, gut gemischt und den Schülern zum Ziehen angeboten; diejenigen Schüler, welche angekreuzte Zettel ziehen, werden als Merkmalsträger der Stichprobe ausgewählt.

Hinter diesem Vorgehen steht die Erwartung, dass bei Anwendung des Zufallsprinzips - wenn also jedes gezielte Auswählen unsererseits unterbleibt - die einzelnen Merkmalsausprägungen in der Stichprobe sozusagen von selbst mit etwa denselben relativen Häufigkeiten vertreten sein werden wie in der Grundgesamtheit. Stichproben, die nach dem Zufallsprinzip gewonnen wurden, heißen zufällige Stichproben.

Bemerkung 39:

- Im Folgenden gehen wir davon aus, dass die von uns betrachteten Stichproben nach dem Zufallsprinzip entstanden sind. Wir werden deshalb auf den Zusatz „zufällig" meist verzichten und kurz von Stichproben reden.
- Aus dem Vorangehenden erkennt man, dass die wichtige Aufgabe der Statistik, von einer Stichprobe auf die Grundgesamtheit zu schließen, Anlass gibt zur Betrachtung von Erscheinungen, die dem Zufall unterliegen.

Zufallsexperimente

Erscheinungen nennen wir zufällig, wenn sie nicht mit absoluter Sicherheit eintreten und insofern also nicht voraussagbar sind.

Beispiel 31:

Wenn beispielsweise in einer Klinik ein Kind zur Welt kommt, so können wir nicht voraussagen, welche Körpergröße es haben wird; wir nennen die Körpergröße des Kindes eine zufällige Erscheinung.

Dagegen können wir mit Sicherheit sagen, dass sich die Mutter des Kindes zur Zeit der Geburt ebenfalls in jener Klinik befindet; der Aufenthaltsort der Mutter ist nicht zufällig.

Definition 52:

Das Beobachten eines zufälligen Merkmals nennt man auch Durchführen eines Zufallsexperiments. Die möglichen Ausprägungen $a_i, \ldots, a_k$ des Merkmals heißen Ausgänge (Ergebnisse), die Menge S aller möglichen Ausgänge heißt Ausgangsmenge des Zufallsexperiments.

Beispiel 32:

Bei Fertigungsmaschinen werden die Produkte untersucht, ob sie brauchbar sind. Die Untersuchung jedes Produkts ist ein Zufallsexperiment mit der Ausgangsmenge S = {brauchbar; nicht brauchbar}.

Solche Fragen lassen sich natürlich nicht beantworten, wenn ein Zufallsexperiment nur ein einziges Mal durchgeführt wird. Es ist also wichtig, dass ein Zufallsexperiment wiederholt durchgeführt werden kann.

Definition 53:

Von Interesse sind solche Zufallsexperimente, die wiederholt (theoretisch sogar beliebig oft) durchgeführt werden können.

Beispiel 33:

Um die Brenndauer einer Glühbirne zu ermitteln, muss diese so lange in Betrieb genommen werden, bis der Glühfaden durchbrennt.

Dieses Zufallsexperiment kann also mit derselben Glühbirne nicht wiederholt werden. Wählt man jedoch statt ihr eine andere Glühbirne, die nach demselben Verfahren hergestellt wurde, und stellt deren Brenndauer fest, so kann dies als Wiederholung des Zufallsexperiments aufgefasst werden.

Das letzte Beispiel eröffnet einen weiteren wichtigen Aspekt. Um für eine Produktionsserie von Glühbirnen die mittlere Brenndauer zu bestimmen, entnimmt man der Serie eine Anzahl von Glühbirnen und stellt bei jeder von ihnen die Brenndauer fest. Auf diese Weise erhält man eine Stichprobe aus der Grundgesamtheit der Brenndauern aller Birnen der Serie. Fasst man nun das Feststellen der Brenndauer einer Birne als Zufallsexperiment auf, so kann man sagen:

Definition 54:

Das Entnehmen einer Stichprobe vom Umfang n aus einer Grundgesamtheit kann aufgefasst werden als n-maliges Durchführen eines Zufallsexperiments.

Bemerkung 40:

Zusammenfassung:

Ein Zufallsexperiment ist ein Experiment mit folgenden Eigenschaften:

- Unter gleichen Bedingungen beliebig oft wiederholbar.
- Es gibt mindestens zwei mögliche Ergebnisse.
- Das Ergebnis ist nicht vorhersagbar.

Modelle für Zufallsexperimente

Wir betrachten einige Zufallsexperimente, die besonders typisch und gleichzeitig einfach sind; sie dienen deshalb bei späteren Überlegungen oft als Modelle für Zufallsexperimente.

Definition 55:

(Münzmodell) Die einfachsten Zufallsexperimente sind solche mit nur 2 Ausgängen; sie heißen **Bernoulli-Experimente**.

Solche Zufallsexperimente treten in der Praxis häufig auf. (Feststellen, ob ein Medikament wirkt oder nicht, ob ein Fertigungsartikel brauchbar ist oder nicht, ob ein Tier männlichen oder weiblichen Geschlechts ist, ob eine Telefonzelle belegt ist oder nicht, usw.)

Als Standardbeispiel für Bernoulli-Experimente kann das Werfen einer Münze dienen. Die Münze fällt entweder so, dass „Bild" (kurz: B) oder so, dass „Wappen" (kurz: W) oben liegt.

Das Zufallsexperiment hat also die Ausgangsmenge S = {B;W}. es ist beliebig oft wiederholbar.

Ausgangsmengen von Zufallsexperimenten

Zur Bestimmung einer Ausgangsmenge

Beim Brettspiel „Mensch ärgere dich nicht" bestimmt sich die Zahl der Felder, um die ein Spieler vorrücken darf, nach der mit einem Würfel geworfenen Augenzahl. Für den Spielbeginn gilt eine Sonderregel: erstmals vorrücken darf nur, wer zuvor eine Sechs geworfen hat. - Welche Ausgänge interessieren einen Spieler zu Beginn bzw. im Verlauf des Spiels?

Wie das Beispiel „Würfeln mit einem Spielwürfel" ein Zufallsexperiment noch nicht ausreichend beschrieben. Es kommt wesentlich darauf an, welche Ausgänge in Betracht gezogen werden. Dies wiederum richtet sich nach den besonderen Interessen, die wir mit der Durchführung des Experiments verfolgen. Man sollte daher nicht z. B. von dem Zufallsexperiment „Würfeln" sprechen, sondern genauer von „Würfeln mit Feststellen der Augenzahl".

Definition 56:

Ein Zufallsexperiment ist erst dann ausreichend bestimmt, wenn eine Menge S von möglichen Ausgängen (Ausgangsmenge) so festgelegt ist, dass bei jeder Durchführung des Zufallsexperiments genau einer der zu S gehörenden Ausgänge eintreten muss.

Beispiel 34:

Die Triebwerke eines 3motorigen Flugzeuges werden getestet.

a) Zunächst interessiert, ob alle Triebwerke einwandfrei laufen, es ist S_1 = {alle Triebwerke einwandfrei; nicht alle Triebwerke einwandfrei}.

b) Falls es Beanstandungen gibt, wird man weiter fragen, wie viele der Triebwerke nicht einwandfrei laufen. Hierfür erweist sich S_1 als nicht ausreichend; man wird S_1 ersetzen durch S_2 = {0; 1; 2; 3}.

c) Mit S_2 erhält man keine Auskunft darüber, welche der drei Triebwerke schadhaft sind. Um auch dies zu erfassen, wird man die Triebwerke z. B. mit A,B,C kennzeichnen, jeweils die schadhaften Triebwerke angeben und etwa 0 schreiben, wenn alle einwandfrei laufen. In diesem Fall ist die Menge S_3 = {0; A; B; C; AB; AC; BC; ABC} Ausgangsmenge.

Besondere Ausgangsmengen, Baumdiagramme

In vorherigen Kapitel haben wir den Münzwurf als Modell für Bernoulli-Experimente betrachtet; die Ausgangsmenge ist hier S = {BW}. Wir beschäftigen uns nun mit der Wiederholung solcher Experimente.

Wird eine Münze 3-mal nacheinander geworfen so kann diese 3malige Durchführung eines Bernoulli-Experiments als ein einziges, neues Zufallsexperiment aufgefasst werden. Jeder Ausgang dieses neuen Experiments lässt sich durch ein Tripel beschreiben;

z. B. kennzeichnet (B;W;B) den Ausgang, bei dem im 1. Wurf.

Die Menge {(B;B;B), ... mit allen Möglichkeiten...} ist Ausgangsmenge. Diese Menge wird bekanntlich mit $S \times S \times S$ oder auch S^3 bezeichnet, wobei S={B;W} die Ausgangsmenge des Bernoulli-Experiments ist.

Alle Elemente von S^3 sind durch ein so genanntes Baumdiagramm veranschaulicht. Jedem Ausgang entspricht ein Weg (von links nach rechts) durch den Baum. Solche Baumdarstellungen erleichtern in Fällen wie dem angegebenen das Auffinden der möglichen Ausgänge.

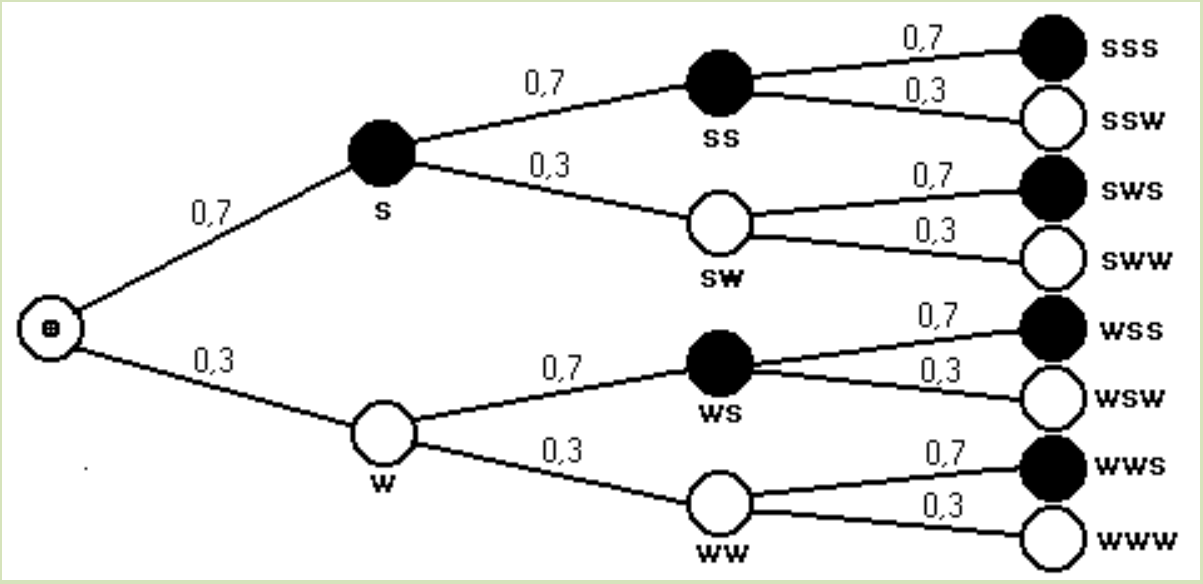
Definition 57:

n Durchführungen eines Bernoulli-Experiments mit der Ausgangsmenge S kann man auffassen als neues Zufallsexperiment mit der Ausgangsmenge S^n.

Beispiel 35:

Nacheinander sollen drei Kugeln (mit Zurücklegen; ohne zurücklegen) aus einer Urne mit 10 Kugeln (7 schwarze und drei weiße) entnommen werden.

Für das Urnenbeispiel erhält man den folgenden Ereignisbaum:



$$P(sss) = 0{,}7 \cdot 0{,}7 \cdot 0{,}7 = 0{,}343 \quad \rightarrow \quad P(3s) = P(0w) = 0{,}343$$

$$P(ssw) = 0{,}7 \cdot 0{,}7 \cdot 0{,}3 = 0{,}147$$
$$P(sws) = 0{,}7 \cdot 0{,}3 \cdot 0{,}7 = 0{,}147 \quad \rightarrow \quad P(2s) = P(1w) = 0{,}441$$
$$P(wss) = 0{,}3 \cdot 0{,}7 \cdot 0{,}7 = 0{,}147$$

$$P(sww) = 0{,}7 \cdot 0{,}3 \cdot 0{,}3 = 0{,}063$$
$$P(wsw) = 0{,}3 \cdot 0{,}7 \cdot 0{,}3 = 0{,}063 \quad \rightarrow \quad P(1s) = P(2w) = 0{,}189$$
$$P(wws) = 0{,}3 \cdot 0{,}3 \cdot 0{,}7 = 0{,}063$$

$$P(www) = 0{,}3 \cdot 0{,}3 \cdot 0{,}3 = 0{,}027 \quad \rightarrow \quad \frac{P(0s) = P(3w) = 0{,}027}{\sum P = 1}$$

Die Wahrscheinlichkeit dafür, dass bei drei Entnahmen mindestens zwei schwarze Kugeln dabei sind : P=0,343+0,441=0,784

Die Wahrscheinlichkeit dafür, dass bei drei Entnahmen mindestens zwei weiße Kugeln dabei sind : P=0,189+0,027=0,216

Pfadregel

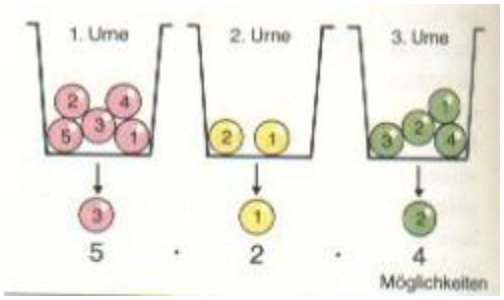
Definition 58:

Pfadregel (Produktregel):

Die Wahrscheinlichkeit eines Pfades in einem mehrstufigen Baumdiagramm ist gleich dem Produkt der Wahrscheinlichkeiten entlang dieses Pfades im Baumdiagramm.

Wird aus n Urnen nacheinander je eine Kugel gezogen, so ist die Anzahl der Möglichkeiten das Produkt aus den Anzahlen der Kugeln n den einzelnen Urnen.

(Produktregel)



Definition 59:

Pfadregel (Summenregel):

Die Wahrscheinlichkeit, vom Startpunkt zum Ziel zu gelangen, ist gleich der Summe der Wahrscheinlichkeiten aller Pfade, die vom Start zum Ziel führen.

Ereignisse

Bei dem bekannten Würfelspiel „Mensch ärgere dich nicht" interessiert bei Spielbeginn die Ausgangsmenge $S_1 = \{6; \text{nicht } 6\}$, im Verlauf des Spiels dagegen $S_2 = \{1;2;3;4;5;6\}$. Für die mathematische Behandlung der mit dem Würfeln zusammenhängenden Fragen ist ein solcher Wechsel der Ausgangsmenge hinderlich. Man versucht daher, mit einer einzigen Ausgangsmenge auszukommen. Als solche bietet sich hier S_2 an: ist ein bestimmter Ausgang von S_2 eingetreten, so weiß man auch welcher Ausgang von S_1 damit eingetreten ist; ist dagegen z. B. der Ausgang „nicht 6" von S_1 eingetreten, so lässt sich daraus nicht ersehen, welcher Ausgang von S_2 eingetreten ist.

Allgemein wird man versuchen, bei Zufallsexperimenten die Ausgangsmenge S so zu wählen, dass nachträglich möglichst alle im Zusammenhang mit dem Zufallsexperiment interessierenden Fragen beantwortet werden können. Dazu muss man die Ausgänge hinreichend fein unterscheiden. Dies gibt dann andererseits Veranlassung, auch allgemeinere Ausgangsmöglichkeiten (wie z. B. beim Würfeln „nicht 6" oder „gerade Augenzahl") in Betracht zu ziehen.

Definition 60:

Ein Zufallsexperiment habe die Ausgangsmenge $S = \{a_1, \ldots, a_k\}$. Dann nennt man jede Teilmenge von S ein zu diesem Zufallsexperiment gehöriges **zufälliges Ereignis** (kurz: Ereignis). Endet eine Durchführung des Zufallsexperiments mit dem Ausgang a_i und ist A ein Ereignis mit $a_i \in A$, so sagt man: das Ereignis A ist eingetreten.

Bemerkung 41:

- Im bisher betrachteten Beispiel war jeder Ausgang (d. h. jedes Element von S) eine Ausprägung des Merkmals Nummer.
- Interessiert ein anderes Merkmal wie z. B. Farbe oder Größe, so wird man auf Ereignisse geführt: jeder Ausprägung des neuen Merkmals entspricht eindeutig eine Teilmenge von S.
- Umgekehrt entspricht nun aber einer Teilmenge von S nicht wiederum eindeutig eine Merkmalsausprägung; es kann durchaus Ausprägungen verschiedener Merkmale geben. denen in S dieselbe Teilmenge zugeordnet ist (so können z. B. in einer Urne die kleinen Kugeln mit den schwarzen Kugeln identisch sein).
- Eine umkehrbar eindeutige Zuordnung erhält man erst, wenn man alle Merkmalsausprägungen, denen in S dieselbe Teilmenge entspricht, zu einem Ganzen zusammenfasst. Dies zeigt, dass ein Ereignis auch als Zusammenfassung einer Vielzahl von Merkmalsausprägungen und die angegebene Definition als mathematische Präzisierung dieses Sachverhaltes aufgefasst werden kann.

Besondere Ereignisse, Ereignisraum

Nach obiger Definition sind auch die Ausgangsmenge S und die leere Menge Ø Ereignisse.

Wann treten diese Ereignisse ein?

Unter den zu einer Ausgangsmenge S gehörenden Ereignissen gibt es besonders einfache, aus denen sich alle andern (außer Ø) durch Mengenvereinigung erzeugen lassen. Welche Ereignisse sind das?

Definition 61:

S heißt das sichere Ereignis,

Ø heißt das unmögliche Ereignis,

die 1-elementigen Ereignisse heißen Elementarereignisse

Bemerkung 42:

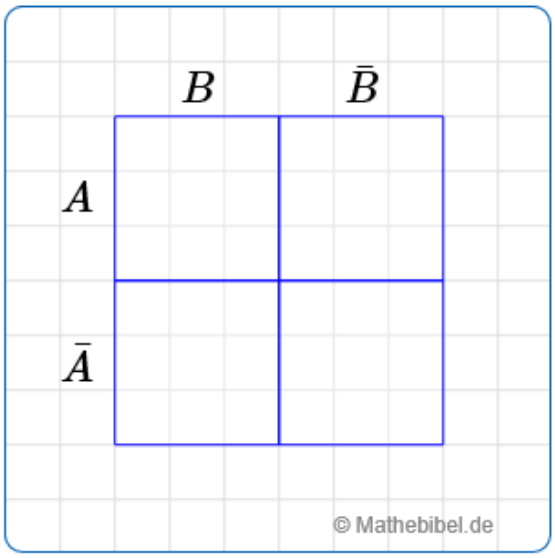
- Die Elementarereignisse erhält man wenn aus den Ausgängen, eines Zufallsexperiments 1-elementige Mengen bildet. Es gibt also jeweils ebenso viele Elementarereignisse wie Ausgänge.

Bei den Erörterungen dieses Abschnitts sind wir davon ausgegangen, dass die betrachteten Zufallsexperimente nur endlich viele Ausgänge haben. Trifft dies nicht zu, hat also S unendlich viele Elemente, so braucht der Ereignisraum des Zufallsexperiments nicht mit der Potenzmenge von S übereinzustimmen, evtl. ist der Ereignisraum dann eine echte Teilmenge der Potenzmenge von S.

Unsere Definition des Begriffs Ereignis lässt sich also nicht unmittelbar auf den Fall unendlicher Zufallsexperimente übertragen; in diesem Fall sind zur Definition weitergehende mathematische Hilfsmittel notwendig.

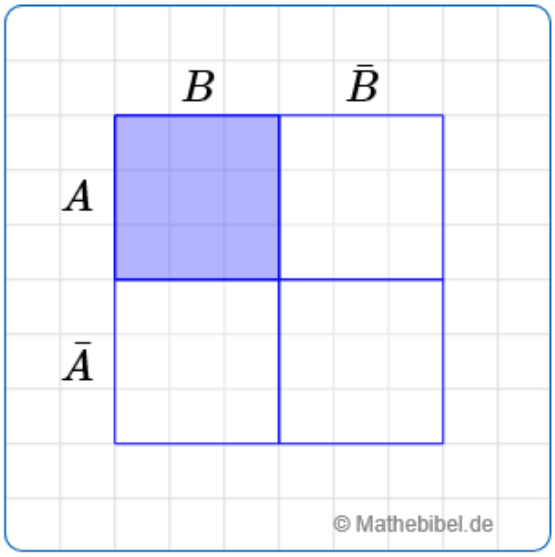
Vierfeldertafel

Die Vierfeldertafel eignet sich hervorragend, um die Verknüpfungen von Ereignissen zu visualisieren. An dieser Stelle geht es vor allem darum, die Vierfeldertafel zu verstehen und richtig zu interpretieren.



Wie der Name bereits vermuten lässt, besteht die Vierfeldertafel aus vier Feldern. Wichtig ist, dass man sich die richtige Beschriftung merkt.

Alle vier Felder zusammen entsprechen dem Ergebnisraum Ω.

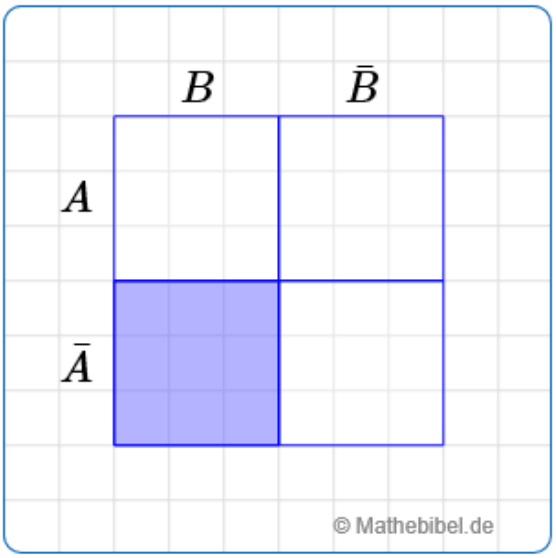


Das farblich hervorgehobene Feld beschreibt diejenigen Elemente des Ergebnisraums Ω, die sowohl in dem Ereignis A als auch in dem Ereignis B vorkommen: $A \cap B$.

Beispiel

$\Omega = \{1, 2, 3, 4, 5, 6\}$;
$A = \{2, 4, 6\}$; $B = \{2, 3, 5\}$;

$A \cap B = \{2\}$

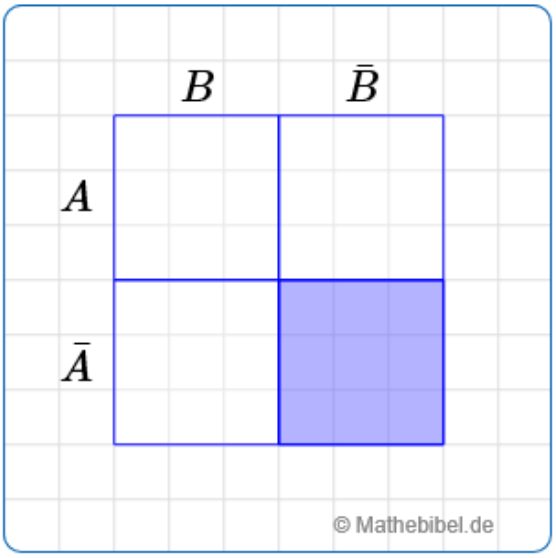


Das farblich hervorgehobene Feld beschreibt diejenigen Elemente des Ergebnisraums Ω, die in dem Ereignis B, aber nicht zugleich in dem Ereignis A, vorkommen: $\bar{A} \cap B$.

Beispiel

$\Omega = \{1,2,3,4,5,6\}$;
$A = \{2,4,6\}$; $B = \{2,3,5\}$;

$\bar{A} \cap B = \{3,5\}$

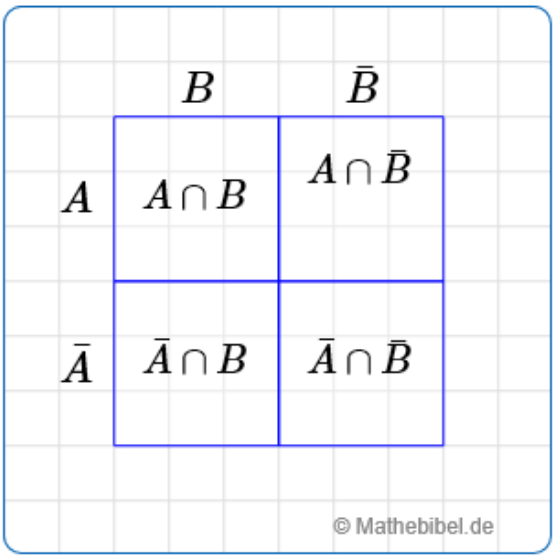


Das farblich hervorgehobene Feld beschreibt diejenigen Elemente, die weder in dem Ereignis A noch in dem Ereignis B vorkommen: $\bar{A} \cap \bar{B}$.

Beispiel

$\Omega = \{1,2,3,4,5,6\}$;
$A = \{2,4,6\}$; $B = \{2,3,5\}$;

$\bar{A} \cap \bar{B} = \{1\}$



Zusammenfassung

$\Omega = \{1,2,3,4,5,6\}$;
$A = \{2,4,6\}$; $B = \{2,3,5\}$;

$A \cap B = \{2\}$
$A \cap \bar{B} = \{4,6\}$
$\bar{A} \cap B = \{3,5\}$
$\bar{A} \cap \bar{B} = \{1\}$

Beispiel 36:

Die 16 Jungen und 14 Mädchen einer Schulklasse nehmen an einem Mathematik-Test teil. 13 Jungen bestehen. Insgesamt bestehen 20 Schüler den Test.
Mit welcher Wahrscheinlichkeit hat ein Schüler den Test nicht bestanden und ist gelichzeitig ein Mädchen?

Lösung

Die beiden Ereignisse sind

1. Schüler ist ein Junge J, oder ein Mädchen (= "nicht Junge) $\overline{\mathrm{J}}$
2. Test bestanden B, Test nicht bestanden $\overline{\mathrm{B}}$

Aus dem Text lassen sich die Wahrscheinlichkeiten $P(J \cap B)$, $P(B)$, $P(J)$ sowie $P(\overline{B})$ bestimmen.

Insgesamt befinden sich 30 SchülerInnen in der Klasse. Die Wahrscheinlichkeit für jedes Ereignis ist dessen relative Häufigkeit.

Die zugehörige Vierfeldertafel wird nun aufgestellt:

Zuerst fertigt man eine Vierfeldertafel an, beschriftet Zeilen und Spalten und trägt die Wahrscheinlichkeiten aus dem Text ein.

	J	$\overline{\mathrm{J}}$	
B	$\frac{13}{30}$		$\frac{20}{30}$
$\overline{\mathrm{B}}$			
	$\frac{16}{30}$	$\frac{14}{30}$	1

Um die fehlenden Werte zu bestimmen benutzt man die Eigenschaften und Rechenregeln von oben.

Den fehlenden Wert in der zweiten Zeile zum Beispiel wird berechnet als: $\frac{20}{30} - \frac{13}{30} = \frac{7}{30}$

	J	$\overline{\mathrm{J}}$	
B	$\frac{13}{30}$	$\frac{7}{30}$	$\frac{2}{3}$
$\overline{\mathrm{B}}$	$\frac{1}{10}$	$\frac{7}{30}$	$\frac{1}{3}$
	$\frac{8}{15}$	$\frac{7}{15}$	1

Die Werte in der dritten Zeile ergeben sich dann durch ähnliche Rechnungen.

Daraus lässt sich leicht die Wahrscheinlichkeit $P(\text{nicht bestanden und nicht Junge})$ $= P(\overline{B} \cap \overline{J}) = \frac{7}{30}$ (3. Zeile, 3. Spalte) ermitteln. Diese ist die gesuchte Lösung.

Mit absoluten Häufigkeiten würde diese Vierfeldertafel so aussehen:

	J	$\overline{\mathrm{J}}$	
B	13	7	20
$\overline{\mathrm{B}}$	3	7	10
	16	14	30

Beispiel 37:

125 Schüler werden nach ihrer Teilnahme am Wahlunterricht Französisch (F) und Informatik (I) befragt: 30 Schüler besuchen Französisch und 20 Informatik. die relative Häufigkeit der Schüler, die nur Französisch besuchen beträgt 11,2%.

a) Erstellen Sie eine Vierfeldertafel.

b) Bestimmen Sie die relative Häufigkeit der Schüler, die nur am Informatikkurs teilnehmen.

c) Bestimmen Sie die Anzahl der Schüler, die keinen der beiden Wahlkurse besuchen.

d) Bestimmen Sie die relative Häufigkeit der Schüler, die mindestens einen Wahlkurs besuchen.

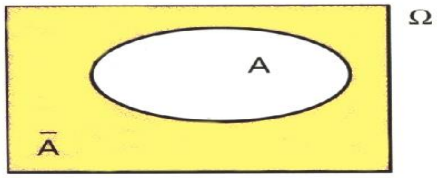
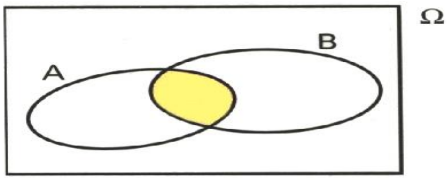
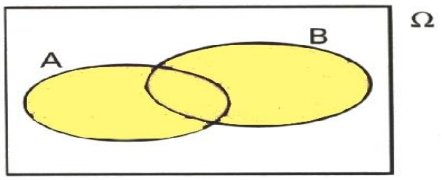
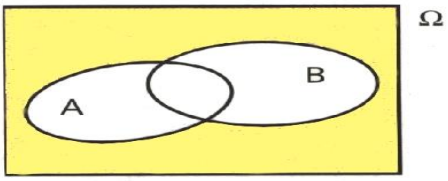
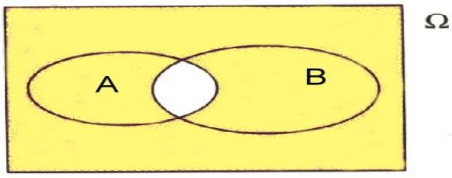
Lösung:

	F	$\bar{F}$	Summe
I	16	4	20
$\bar{I}$	14	91	105
Summe	30	95	125

	F	$\bar{F}$	Summe
I	12,8%	3,2%	16%
$\bar{I}$	11,2%	72,8%	84%
Summe	24%	76%	100%

Mengenalgebra (Ereignisalgebra)

Basis-Verknüpfungen

Symbol	Sprechweise	Veranschaulichung (Venn-Diagramm)
$\bar{A}$	Gegenereignis von A	
$A \cap B$	Ereignis A und Ereignis B	
$A \cup B$	Ereignis A oder Ereignis B	
$\overline{A \cap B} = \overline{A} \cup \overline{B}$	weder A noch B	
$\overline{A \cup B} = \overline{A} \cap \overline{B}$	höchstens eines der Ereignisse	

Oder-Verknüpfung (Additionsgesetz)

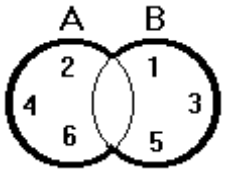
Die Zusammenhänge werden am Würfelbeispiel erläutert. Als bekannt vorausgesetzt wird, dass die Wahrscheinlichkeit für das Ereignis 'Gerade Zahl' 0.5, für das Ereignis 'Ungerade Zahl' ebenfalls 0.5 und für das Ereignis 'Durch drei teilbare Zahl' 1/3 beträgt.

Würfel	A : Wurf einer geraden Zahl ;	A={2,4,6}
	B : Wurf einer ungeraden Zahl ;	B={1,3,5}
	C : Wurf einer durch 3 teilb. Zahl ;	C={3,6}

Additionsgesetz für unvereinbare Ereignisse (Oder-Verknüpfung)

Beim Würfeln will man zum Beispiel die Wahrscheinlichkeit für das Ereignis 'Gerade Zahl' **oder** 'Ungerade Zahl' wissen.

Die Wahrscheinlichkeit dafür, dass das Ereignis **A oder B** eintritt, lässt sich bildlich als Vereinigungsmenge wie folgt darstellen.

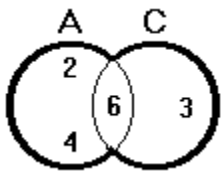
$P(A \cup B) = P(A) + P(B)$ $P(A \cup B) = 0.5 + 0.5 = 1$	

Der Wurf mit dem Ereignis 'gerade Zahl‘ oder 'ungerade Zahl‘ ist ein sicheres Ereignis ($P(A \cup B)=1$).

Die Unvereinbarkeit der Ereignisse findet in der freibleibenden Vereinigungsmenge ihren Ausdruck.

Additionsgesetz für vereinbare Ereignisse (Oder-Verknüpfung)

Die Wahrscheinlichkeit dafür, dass das Ereignis **A oder C** eintritt lässt sich bildlich als Vereinigungsmenge wie folgt darstellen.

$P(A \cup C) = P(A) + P(C) - P(A \cap C)$ $P(A \cup C) = 1/2 + 1/3 - 1/6 = 2/3$	

Der Wurf mit den Ereignissen 'gerade Zahl‘ oder 'durch 3 teilbare Zahl' tritt mit 66,7-prozentiger Wahrscheinlichkeit ein ($P(A \cup C)=2/3$). Weil der Wurf der 6 sowohl dem Ereignis A als auch dem Ereignis C zugerechnet wird, muss die Subtraktion von $P(A \cap C)$ dafür sorgen, dass das Ereignis '6' nur einmal berücksichtigt wird. Die Vereinbarkeit der Ereignisse findet in der belegten Vereinigungsmenge ihren Ausdruck.

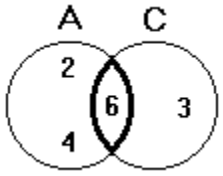
UND-Verknüpfung (Multiplikationsgesetz)

Auch hier werden die Zusammenhänge am Würfelbeispiel erläutert. Als bekannt vorausgesetzt wird, dass die Wahrscheinlichkeit für das Ereignis 'Gerade Zahl' 0.5, für das Ereignis 'Ungerade Zahl' ebenfalls 0.5 und für das Ereignis 'Durch drei teilbare Zahl' 1/3 beträgt.

Würfel	A : Wurf einer geraden Zahl ; A={2,4,6} B : Wurf einer ungeraden Zahl ; B={1,3,5} C : Wurf einer durch 3 teilb. Zahl ; C={3,6}

Multiplikationsgesetz für vereinbare Ereignisse (Und-Verknüpfung)

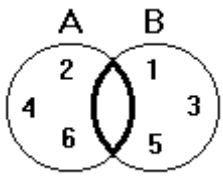
Die Wahrscheinlichkeit dafür, dass das Ereignis **A und C** eintritt, lässt sich bildlich als Schnittmenge wie folgt darstellen.

$P(A) \cap P(C) = P(A) \cdot P(C)$ $P(A) \cap P(C) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$	

Die Wahrscheinlichkeit dafür, dass der Wurf mit dem Ereignis 'gerade Zahl' **und** dem Ereignis 'durch 3 teilbare Zahl' eintritt, beträgt 1/6. Es ist die Wahrscheinlichkeit dafür, dass eine '6' gewürfelt wird.

Multiplikationsgesetz für unvereinbare Ereignisse (Und-Verknüpfung)

Die Wahrscheinlichkeit dafür, dass das Ereignis **A und B** eintritt, lässt sich bildlich als leere Schnittmenge wie folgt darstellen.

$P(A \cap B) = P(A) \cap P(B) = 0$ Schnittmenge = Leere Menge	

Unvereinbare Ereignisse können nicht gleichzeitig auftreten. Die Schnittmenge ist leer.

Beispiel 38:

Unabhängigkeit von Ereignissen

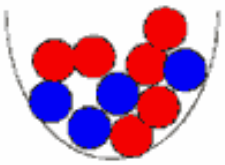
Zunächst eher etwas unscharf an die Umgangssprache angelehnt:

Unabhängig sind Ereignisse, wenn sie sich nicht beeinflussen.

Wenn zwei Ergebnisse unabhängig sind, gilt $P(A \cap B) = P(A) \cdot P(B)$.
Dies kann an einem Beispiel erarbeitet werden.

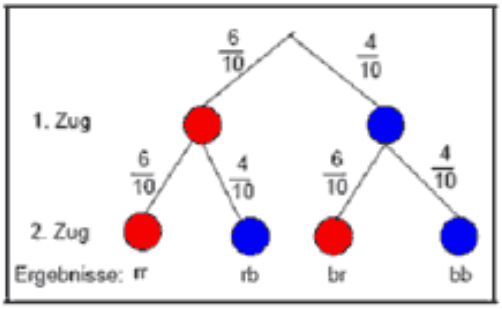
Ausgangspunkt:
Für $P(A \cup B)$ haben wir einen Zusammenhang gefunden (Additionssatz).
Gibt es auch für $P(A \cap B)$ einen?

Beispiel *Eine Schale enthält sechs rote und vier blaue Kugeln, und es werden zufällig zwei davon gezogen.*
Die Ereignisse
*A :"**im ersten Zug rot**" und B: "**im zweiten Zug rot**" werden untersucht.*
Vergleiche $P(A \cap B)$, $P(A)$ und $P(B)$..

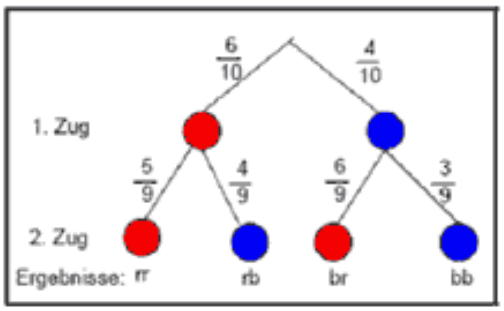


Es ist A = {rr, rb} und B = {rr, br} sowie $A \cap B$ = {rr} ("in beiden Zügen rot").

Ziehen **mit** Zurücklegen



Ziehen **ohne** Zurücklegen



Nach der Pfadregel und der Summenregel ergibt sich:

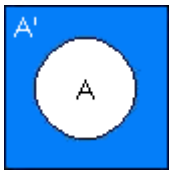
$$P(A) = \frac{6}{10} \cdot \frac{6}{10} + \frac{6}{10} \cdot \frac{4}{10} = \frac{6}{10}$$
$$P(B) = \frac{6}{10} \cdot \frac{6}{10} + \frac{4}{10} \cdot \frac{6}{10} = \frac{6}{10}$$
$$P(A \cap B) = \frac{6}{10} \cdot \frac{6}{10} = \frac{36}{100}$$
also: $P(A \cap B) = P(A) \cdot P(B)$

$$P(A) = \frac{6}{10} \cdot \frac{5}{9} + \frac{6}{10} \cdot \frac{4}{9} = \frac{6}{10}$$
$$P(B) = \frac{6}{10} \cdot \frac{5}{9} + \frac{4}{10} \cdot \frac{6}{9} = \frac{6}{10}$$
$$P(A \cap B) = \frac{6}{10} \cdot \frac{5}{9} = \frac{30}{90}$$
also: $P(A \cap B) \neq P(A) \cdot P(B)$

Komplementärmenge

Definition 62:

Dies Komplementärmenge A' ist die Menge aller Elemente die nicht zur Menge A gehören. Es wird oft durch A' oder $\overline{A}$symbolisiert. Alle Werte einer Grundgesamtheit sind entweder Elemente der Menge A oder der Menge A', es gibt keine Werte die sowohl in A als auch A' enthalten sind.



Die Summe der Wahrscheinlichkeiten des Ereignisses A und seines Komplementärereignisses A' ist eins.

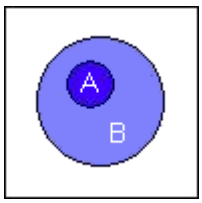
$P(A) + P(A') = 1$

In manchen Fällen ist es einfacher P(A') zu berechnen. In diesen Fällen kann P(A) über P(A) = 1 - P(A') berechnet werden.

Untermengen

Definition 63:

Wenn die Werte des Ereignisses A eine Untermenge der Werte von B sind, dann wird von A gesagt, dass es in B enthalten ist, und wird geschrieben als $A \subset B$. Deshalb ergibt sich, wenn $A \subset B$, aus dem Auftreten von A notwendigerweise auch das Auftreten von B. Man kann leicht sehen, dass $P(A) \subset P(B)$.



Gleichverteilung

Der Begriff Gleichverteilung stammt aus der Wahrscheinlichkeitstheorie und beschreibt eine Wahrscheinlichkeitsverteilung mit bestimmten Eigenschaften. Im diskreten Fall tritt jeder mögliche Zustand mit der gleichen Wahrscheinlichkeit ein, im stetigen Fall ist die Dichte konstant. Der Grundgedanke einer Gleichverteilung ist, dass es keine Präferenz gibt.

Definition 64:

Eine Wahrscheinlichkeitsverteilung, die allen Elementarereignissen die gleiche Wahrscheinlichkeit zuordnet, heißt Gleichverteilung (gleichverteilte Wahrscheinlichkeitsfunktion).

Beispiel 39:

Ein idealer Würfel lässt sich mittels Wahrscheinlichkeiten dadurch kennzeichnen, dass bei ihm jeder der 6 Augenzahlen die gleiche Wahrscheinlichkeit zugeordnet ist, nämlich 1/6. Entsprechend ist bei einer idealen Münze die Wahrscheinlichkeit für jedes Elementarereignis 1/2.

Ist eine Wahrscheinlichkeitsverteilung eine Gleichverteilung und handelt es sich um k Elementarereignisse, so ordnet sie jedem Elementarereignis die Wahrscheinlichkeit k zu. Daraus folgt:

Definition 65:

Hat ein Ereignisraum mit k Elementarereignissen eine gleichverteilte Wahrscheinlichkeitsfunktion, so gilt für ein Ereignis A mit r Ausgängen

$$P(A) = \frac{r}{k}$$

Zufallsexperimente mit gleichverteilter Wahrscheinlichkeitsfunktion heißen auch **Laplace-Experimente**.

Beispiel 40:

Von den 1450 Schülern einer Schule spielen 580 ein Streichinstrument.

Wie groß ist die Wahrscheinlichkeit, dass ein zufällig herausgegriffener Schüler dieser Schule ein Streichinstrument spielt?

Lösung:

Wir denken uns die Schüler nummeriert von 1 bis 1450. Dann haben wir es mit 1.450 Elementarereignissen E_i zu tun mit E_i: der Schüler Nr. i wird herausgegriffen.

Setzen wir eine gleichverteilte Wahrscheinlichkeitsfunktion voraus, so ist also $P(E_i) = 1/1450$ also für jedes i. Für das interessierende Ereignis A (der herausgegriffene Schüler spielt ein Streichinstrument) sind von den 1450 möglichen Ausgängen 580 günstig; wir erhalten damit

$$P(A) = \frac{580}{1450} = 0{,}40$$

Hilfsmittel aus der Kombinatorik

Um eine gleichverteilte Wahrscheinlichkeitsfunktion zu bestimmen, braucht man die Anzahl der Ausgänge bzw. Elementarereignisse. Diese Anzahl kann sehr groß sein. Sie kann dann nicht mehr durch direktes Abzählen ermittelt werden; man muss sie dann berechnen. Hierbei leisten die folgenden Sätze gute Dienste.

Beispiel 41:

Max besitzt 3 Hemden und 2 Krawatten. Welche (wie viele) Möglichkeiten hat er, jeweils ein Hemd mit einer Krawatte zu kombinieren?

Lösung:

Kennzeichnen Sie die drei Hemden durch die Ziffern 1, 2, 3 und die beiden Krawatten durch 1 und 2. Geben Sie nun die Kombinationsmöglichkeiten als geordnete Paare an. Zeichnen Sie ein Baumdiagramm.

Beispiel 42:

Wie viele Autokennzeichen gibt es, die aus einem der 26 Buchstaben des Alphabets und einer der Ziffern 1, . . ., 9 bestehen?

Lösung:

$26 \cdot 9 = 234$

Geordnete Stichproben mit Zurücklegen (Variationen mit Wiederholungen)

Innerhalb der Gruppen dürfen jedoch Elemente zwei- oder mehrfach auftreten.

Der Urne mit n unterschiedlichen Kugeln werden geordnete Stichproben von k Kugeln entnommen. Dabei wird jede gezogene Kugel vor der nächsten Ziehung in die Urne zurückgelegt.

n=3 ;	M={ 1, 2, 3 }	
k=1	V_W = { 1, 2, 3 }	$V_W(3,1) = 3^1 = 3$
k=2	V_W={ 11 21 31 12 22 32 13 23 33	$V_W(3,2) = 3^2 = 9$
k=3	V_W= 111 211 311 112 212 312 113 213 313 121 221 321 122 222 322 123 223 323 131 231 331 132 232 332 133 233 333	$V_W(3,3) = 3^3 = 27$
Allgemein :	$V_W(n,k) = n^k$	

Bemerkung 43:

Voraussetzungen

- Alle (n) Elemente der Ausgangsmenge unterscheiden sich voneinander.
- Es werden einige (k) Elemente ausgewählt.
- Ein Element kann mehrmals ausgewählt werden.

Definition 66:

Einer Gesamtheit von n verschiedenen Elementen kann man

$$\mathbf{n^k}$$

geordnete Stichproben mit Zurücklegen vom Umfang k entnehmen.

Beispiel 43:

Bei einer Fußballwette soll hinter 11 auf dem Wettschein angegebenen Spielen jeweils eine der Ziffern 1 (1. Verein gewinnt), 2 (2. Verein gewinnt) oder 0 (unentschieden) gesetzt werden.

Wie groß ist die Wahrscheinlichkeit, dass jemand zufällig bei allen 11 Spielen richtig tippt?

Lösung:

Den Elementen 1, 2, 0 (n = 3) ist eine geordnete Stichprobe mit Zurücklegen vom Umfang 11 zu entnehmen. Hierfür gibt es (nach der obigen Definition)

3^{11} = 177147 Möglichkeiten.

Setzt man eine gleichverteilte Wahrscheinlichkeitsfunktion voraus, so beträgt die Wahrscheinlichkeit, mit einer einzigen Wette 11 „Richtige" zu tippen, 1:177 147 (0,000 006); d. h. im Durchschnitt werden von einer Million Wetten etwa sechs gewinnen.

Beispiel 44:

Wie viel unterschiedliche (auch sinnlose) Wörter mit drei Buchstaben lassen sich aus den 6 Buchstaben a, b, c, d, e, und f bilden, wenn jeder Buchstabe auch mehrmals verwendet werden darf?

Lösung:

$$V_w(6,3) = 6^3 = \underline{\underline{216}}$$

Beispiel 45:

Kfz-Schilder verwenden 2 Buchstaben und vier Ziffern (in dieser Reihenfolge). Die erste Ziffer darf keine Null sein. Wie viel verschiedene Kennzeichen sind möglich, wenn jeder Buchstabe und jede Ziffer mehrmals verwendet werden dürfen?

Lösung:

Zwei Buchstaben : $V_w(2,26) = n^k = 26^2 = \underline{\underline{676}}$

Erste Ziffer 1..9 : $\underline{\underline{9 \text{ Möglichkeiten}}}$

Letzte drei Ziffern : $V_w(10.3) = n^k = 10^3 = \underline{\underline{1000}}$

Zusammen : $676 \cdot 9 \cdot 1000 = \underline{\underline{6084000}}$ Möglichkeiten

Beispiel 46:

Wie viele 4stellige Nummern lassen sich mit den Ziffern 1, . . ., 9 bilden?

Lösung:

$$9^4 = 6.561$$

Geordnete Stichproben ohne Zurücklegen (Variationen ohne Wiederholung)

Innerhalb der Gruppen wird die Reihenfolge der Elemente berücksichtigt.

Der Urne mit n unterschiedlichen Kugeln werden geordnete Stichproben von k Kugeln (ohne Zurücklegen) entnommen.

n=3 ;	M={ 1, 2, 3 }	
k=1	V={ 1, 2, 3 }	$V(3,1)=\frac{3!}{2!}=3$
k=2	V={ 12 21 31 13 23 32	$V(3,2)=\frac{3!}{1!}=6$
k=3	V= 123 213 312 132 231 321	$V(3,3)=\frac{3!}{0!}=6$
Allgemein :	$\boxed{V(n,k)=\frac{n!}{(n-k)!}}$	

Definition 67:

Eine Gesamtheit von n verschiedenen Elementen kann man

$$n\cdot(n-1)\cdot\ldots\ldots\cdot(n-k+1) \text{ oder } \frac{n!}{(n-k)!}$$

geordnete Stichproben ohne Zurücklegen vom Umfang k entnehmen.

Beispiel 47:

Wir denken uns neun Schilder mit je einer der Ziffern 1 bis 9 gegeben.

Wie viele 3stellige Nummern kann man mit diesen Schildern zusammenstellen?

Im Unterschied zum vorherigen Beispiel kann jetzt bei den zu bildenden Nummern jede Ziffer nur ein einziges Mal vorkommen.

Wir haben also bei der Hunderterziffer unter 9, bei der Zehnerziffer unter 8 und bei der Einerziffer unter 7 Möglichkeiten die Wahl. Nach der Produktregel lassen sich also $9 \cdot 8 \cdot 7 = 504$ 3-stellige Nummern bilden.

Beispiel 48:

Eine Urne enthält 10 Kugeln, die durch die Zahlen 0, 1, 2, ...9 unterschieden werden. Wie viel verschiedene geordnete Stichproben vom Umfang k=3 können der Urne ohne Zurücklegen entnommen werden?

Lösung:

$$V(10,3) = \frac{10!}{(10-3)!} = \frac{10!}{7!} = 10 \cdot 9 \cdot 8 = \underline{\underline{720}}$$

Beispiel 49:

Auf einem Tisch liegen verdeckt 26 Kärtchen mit den Buchstaben des Alphabets. Sie ziehen nacheinander 10 Kärtchen.

Wie groß ist die Wahrscheinlichkeit, dass sich das Wort „Stichprobe" ergibt?

Nach dem Satz von oben können

26 · 25 · 24 · 23 · 22 · 21 · 20 · 19 · 18 · 17 = 19.275.223.968.000)

verschiedene Wörter entstehen.

Nimmt man eine gleichverteilte Wahrscheinlichkeitsfunktion an, so entfällt auf jedes dieser Wörter somit eine Wahrscheinlichkeit von

1:19.275.223.968.000 (~ 0,000 000 000 000 05),

d. h. es wird durchschnittlich unter 20 Billionen Versuchen 1-mal das Wort „Stichprobe" auftreten.

Beispiel 50:

Aus einer Urne mit sechs unterschiedlich gefärbten Kugeln sollen vier Kugeln entnommen werden (ohne Zurücklegen). Wie viel Möglichkeiten gibt es, wenn die Reihenfolge beachtet werden muss?

Lösung:

$$V(n,k) = \frac{n!}{(n-k)!} \qquad \underline{\underline{V(6,4)}} = \frac{6!}{2!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 2} = \underline{\underline{360}}$$

Beispiel 51:

Wie viel unterschiedliche (auch sinnlose) Wörter mit drei Buchstaben lassen sich aus den 6 Buchstaben a, b, c, d, e, und f bilden, wenn jeder Buchstabe nur einmal verwendet werden darf?

Lösung :

$$V(6,3) = \frac{6!}{(6-3)!} = \frac{6!}{3!} = 6 \cdot 5 \cdot 4 = \underline{\underline{120}}$$

Beispiel 52:

Beim Pferdetoto muss in der sog. Dreierwette der Zieleinlauf der ersten drei Pferde in der richtigen Reihenfolge vorhergesagt werden. Wie viel verschiedene Dreierwetten sind möglich, wenn 10 Pferde starten.

Lösung:

$$V(n,k)=\frac{n!}{(n-k)!} \qquad \begin{matrix} n=10 \\ k=3 \end{matrix} \rightarrow V(10,3)=\frac{10!}{7!}=10\cdot 9\cdot 8=\underline{\underline{720}}$$

Beispiel 53:

An einem Autorennen nehmen 16 Wagen teil. Wie viele Möglichkeiten gibt es, die drei ersten Plätze zu tippen?

Lösung:

$$\frac{16!}{(16-3)!}=16\cdot 15\cdot 14=3.360$$

Beispiel 54:

Wie viele 6stellige Nummern kann man aus den Ziffern .1, . . ., 9 bilden, wenn jede Ziffer höchstens 1mal vorkommen darf?

Lösung:

$$\frac{9!}{(9-6)!}=60.480$$

Geordnete Vollerhebungen

Alle Elemente der Menge {1, 2, ..., n} werden in eine bestimmte Anordnung gebracht. Die Reihenfolge wird berücksichtigt.

Z.B. werden für jeweils 2, 3 oder 4 gleichgroße Lotto-Kugeln mit den Ziffern 1..n alle möglichen Anordnungen gefunden.

n=2 M={1, 2} 	$P_O(2)$= 12 21	$P_O(2)=2!$
n=3 M={1, 2, 3} 	$P_O(3)$= 123 213 312 132 231 321	$P_O(3)=3!$
n=4 M={1,2,3,4} 	$P_O(4)$=1234 2134 3124 4123 1243 2143 3142 4132 1324 2314 3214 4213 1342 2341 3241 4231 1423 2413 3412 4312 1432 2431 3421 4321	$P_O(4)=4!$
Allgemein :	$P_O(n) = n!$	

Definition 68:

Bei einer Gesamtheit von n verschiedenen Elementen gibt es

1 ·2 ·.....· n oder n! (lies: n Fakultät)

geordnete Vollerhebungen.

Beispiel 55:

Es sei M = {1;2;3;4}. Wählen Sie eine der Ziffern aus M als Tausenderziffer, danach eine der restlichen Ziffern als Hunderterziffer, schließlich eine der noch verbliebenen Ziffern als Zehnerziffer und die andere als Einerziffer.

Wie viele 4stellige Nummern lassen sich auf diese Weise bilden? Inwiefern handelt es sich um einen Sonderfall von dem Satz aus dem vorherigen Kapitel?

Sehr oft wird diesem Satz im Sonderfall k = n benötigt. In diesem Fall werden also einer Gesamtheit von n Elementen nacheinander sämtliche Elemente entnommen; die Stichprobe hat dann also den Umfang n.

Da die Elemente nacheinander entnommen werden, treten die Elemente überdies in einer bestimmten Reihenfolge auf (Geordnete Vollerhebung).

Bemerkung 44:

- Eine geordnete Vollerhebung von n Elementen ist zu verstehen als n-Tupel, bei dem alle Komponenten verschieden sind. Ein solches n-Tupel heißt auch eine Permutation der gegebenen n Elemente.

Beispiel 56:

Bei einem Festakt sind für 8 Ehrengäste namentlich gekennzeichnete Plätze reserviert. Kurz vor Eintreffen der Gäste werden versehentlich die Namensschilder entfernt. Wie groß ist die Wahrscheinlichkeit, dass die 8 Ehrengäste zufällig in der vorgesehenen Reihenfolge Platz nehmen, wenn alle Möglichkeiten gleich wahrscheinlich sind?

Nach Satz von oben können die Ehrengäste auf 8! (= 40 320) verschiedene Arten Platz nehmen, wovon eine die vorgesehene ist. Die gesuchte Wahrscheinlichkeit beträgt also 1 : (8!) (~ 0,000 025); d. h. bei 1000 000 derartigen Fällen wird sich durchschnittlich etwa 25mal zufällig die richtige Reihenfolge einstellen.

Beispiel 57:

Wie viel Anordnungen gibt es für 6 verschiedene Kugeln?

Lösung:

P(6)=6! = 720

Mit 6 unterschiedlichen Ziffern lassen sich 720 unterschiedliche. Zahlen anordnen.

Beispiel 58:

Wie viele Wörter lassen sich aus den Buchstaben a, m, o bilden, wenn jeder der Buchstaben in jedem Wort genau 1mal vorkommt? (6)

Lösung:

$$3! = 6$$

Beispiel 59:

Auf einem Tisch liegen 5 adressierte Briefkuverts und 5 dazugehörige Briefe an verschiedene Personen. Es wird blind jeder Brief in ein Kuvert gesteckt.

Wie groß ist die Wahrscheinlichkeit, dass zufällig jeder Brief im richtigen Kuvert ist?

Lösung:

Anzahl der Möglichkeiten: $5! = 120$

Wahrscheinlichkeit: $\frac{1}{120} = 0{,}008$

Geordnete Vollerhebung mit p, g,… gleichen Elementen

Enthält eine Menge mit n Elementen gleiche Elemente, so verringert sich die Anzahl der Permutationen

M={1, 1, 2, 3 } 	P_W= 1123 2113 1132 2131 1213 2311 1231 3112 1312 3121 1321 3211	$P_W(4,2) = \frac{4!}{2!} = 12$
M={ 1, 1, 1, 2 } ; 	P_W= 1112 1121 1211 2111	$P_W(4,3) = \frac{4!}{3!} = 4$

Definition 69:

Permutationen von n Elementen, von denen jeweils p oder q Elemente gleich sind:

$$P_W(n,p,q) = \frac{n!}{p! \cdot q!}$$

Beispiel 60:

Wie viel Anordnungen gibt es bei 6 verschiedenen Kugeln, von denen	
a) 2 Einser	
b) drei Einser	
c) 3 Einser und 2 Zweier	

Lösung	:	M={1, 1, 2, 3, 4, 5}	P_W(6,2) = (6! / 2!)	= 360
		M={1, 1, 1, 2, 3, 4}	P_W(6,3) = (6! / 3!)	= 120
		M={1, 1, 1, 2, 2, 3}	P_W(6,3,2)= (6! / (3! * 2!))	= 60

Ungeordnete Stichproben ohne Zurücklegen

Aus n Elementen werden ungeordnete Gruppen mit jeweils k Elementen gebildet (Reihenfolge beliebig).

Einer Urne mit n unterschiedlichen Kugeln werden nacheinander jeweils k Kugeln ohne Zurücklegen entnommen. Die Reihenfolge ist dabei beliebig (ungeordnete Untermenge).

n=4 ; M={ 1, 2, 3, 4 }		
k=1	C={ 1, 2, 3, 4, }	$C(4,1)=\binom{4}{1}=4$
k=2	C={ 12 23 34 13 24 14 }	$C(4,2)=\binom{4}{2}=\frac{4\cdot 3}{1\cdot 2}=6$
k=3	C={ 123 234 124 134 }	$C(4,3)=\binom{4}{3}=\frac{4\cdot 3\cdot 2}{1\cdot 2\cdot 3}=4$
k=4	C= 1234	$C(4,4)=\binom{4}{4}=\frac{4\cdot 3\cdot 2\cdot 1}{1\cdot 2\cdot 3\cdot 4}=1$
Allgemein :	$C(n,k)=\binom{n}{k}=\frac{n!}{k!\cdot (n-k)!}$	

Definition 70:

Eine Gesamtheit von n verschiedenen Elementen kann man

$$\frac{n(n-1)\cdot\ldots\cdot(n-k+1)}{1\cdot 2\ldots\cdot k} \text{ oder } \frac{n!}{k!\,(n-k)!} \text{ oder } \binom{n}{k}$$

ungeordnete Stichproben ohne Zurücklegen vom Umfang k entnehmen.

Beispiel 61:

Berechnen Sie

$\binom{4}{2}=6$	$\binom{8}{5}=56$	$\binom{9}{3}=84$	$\binom{6}{6}=1$	$\binom{6}{1}=6$

Beispiel 62:

Wie viel Zweier-Mannschaften lassen sich aus 5 Spielern für ein Doppel zusammenstellen; das gegen einen anderen Verein spielen soll?

Lösung :

K=(5 über 2)=10 mögliche Mannschaften

(EH, EL, EM, EW, HL, HM, HW, LM, LW, MW)

Beispiel 63:

Wie viele unterschiedliche Möglichkeiten gibt es zwei Karten aus einem Skatspiel zu ziehen?

Lösung:

Kombinationen ohne Wiederholung: $C(32,2) = \binom{32}{2} = \frac{32 \cdot 31}{1 \cdot 2} = \underline{\underline{496}}$

Beispiel 64:

Ein Elektrogeschäft erhält eine Sendung Glühbirnen mit 60 Stück, in der sich 2 defekte Birnen befinden.

a) Sie kaufen in dem Geschäft 2 Glühbirnen und werden aus dieser Sendung bedient. Wie groß ist die Wahrscheinlichkeit, dass zufällig die beiden defekten Glühbirnen ausgewählt werden?

b) Wie groß ist die Wahrscheinlichkeit, dass sich unter 3 zufällig ausgewählten Glühbirnen der Sendung keine der beiden defekten Glühbirnen befindet?

Lösung:

a) $P(A) = \frac{\binom{2}{2}}{\binom{60}{2}} = \frac{1}{1.770} = 0{,}00056$

b) $P(B) = \frac{\binom{58}{3}}{\binom{60}{3}} = \frac{30.856}{34.220} = 0{,}9017$

Ungeordnete Stichproben mit Zurücklegen (Kombinationen mit Wiederholung)

Innerhalb der Gruppen dürfen Elemente mehrmals vorkommen. Die Anzahl der Kombinationen ist größer als wenn alle Elemente verschieden sind.

Einer Urne mit n unterschiedlichen Kugeln werden nacheinander jeweils k Kugeln mit Zurücklegen entnommen. Die Reihenfolge ist dabei nicht von Interesse (ungeordnete Untermenge).

Herleitung der allgemeingültigen Formel am Beispiel einer Dreier-Menge.

M = {1, 2, 3 }		
k=1	C_W = 1 2 3	$C_w(3,1)=\binom{3}{1}=3$
k=2	C_W = 11 22 33 12 13 23	$C_w(4,2)=\binom{4}{2}=\frac{4\cdot 3}{1\cdot 2}=6$
k=3	C_W = 111 222 333 112 223 113 233 122 123 133	$C_w(5,3)=\binom{5}{3}=\frac{5\cdot 4}{1\cdot 2}=10$
Allgemein: $\boxed{C_w(n,k)=\binom{n+k-1}{k}}$		

Definition 71:

Eine Gesamtheit von n verschiedenen Elementen kann man

$$C_w(n,k)=\binom{n+k-1}{k}$$

ungeordnete Stichproben mit Zurücklegen vom Umfang k entnehmen.

Beispiel 65:

Für eine Parallelschaltung aus drei Widerständen stehen fünf Widerstände R_1, R_2, ...R_5 zur Verfügung.

Wie viel Widerstandskombinationen gibt es, wenn jeder der fünf Widerstände auch mehrmals verwendet werden darf?

Lösung:

$$C_W(n,k) = \binom{n+k-1}{k} \text{ mit } \begin{matrix} n=5 \\ k=3 \end{matrix} \rightarrow \underline{\underline{C_W(5,3)}} = \binom{7}{3} = \frac{7 \cdot 6 \cdot 5}{1 \cdot 2 \cdot 3} = \underline{\underline{35}}$$

Beispiel 66:

Gesucht ist bei einem Wurf mit zwei Würfeln:

a) Ergebnismenge Ω

b) Teilmengen A : Augensumme ist vier

B : Augensumme ist höchstens fünf

C : Beide Augenzahlen sind ungerade

D : Augensumme ist ungerade

E : Augenprodukt ist geradzahlig

Lösung:

Möglichkeiten : $C_W(6,2) = \binom{6+2-1}{2} = \frac{7 \cdot 6}{1 \cdot 2} = \underline{\underline{21}}$

11 12 13 14 15 16

21 **22 23 24 25 26** Die nicht-fettgedruckten

31 32 **33 34 35 36** Kombinationen sind

41 42 43 **44 45 46** doppelt.

51 52 53 54 **55 56**

61 62 63 64 65 **66**

Ω = {11 12 13 14 15 16 22 23 24 25 26 33 34 35 36 44 45 46 55 56 66}

A = {13 22}

B = {11 12 13 14 22 23}

C = {11 13 15 33 35 55}

D = {12 14 16 23 25 34 36 45 56}

E = {12 14 16 22 23 24 25 26 34 36 44 45 46 56 66}

Beispiel 67:

Eine Urne enthält 100 Kugeln.

70 Kugeln bestehen aus dem Material Holz und 30 Kugeln sind aus Kunststoff.

25 der Holzkugeln sind mit der Farbe Rot gestrichen und 45 sind grün.

10 der Kunststoffkugeln sind rot und 20 sind grün.

Folgende Ereignisse werden definiert:

A: Die Kugel ist aus Holz	$\overline{A}$: Die Kugel ist aus Kunststoff
B: Die Kugel ist rot	$\overline{B}$: Die Kugel ist grün

Die Kugeln tragen **zwei Merkmale** mit jeweils **zwei Ausprägungen**:

Merkmal I	Ausprägung	Merkmal II	Ausprägung
Material	A: Holz $\overline{A}$: Kunststoff	Farbe	B: rot $\overline{B}$: grün

Dieser Sachverhalt kann in einer Vierfeldtafel dargestellt werden:

		Merkmal II (Farbe)		
		B: rot	$\overline{B}$: grün	Summe
Merkmal I Material	A: Holz	25	45	70
	$\overline{A}$: Kunststoff	10	20	30
	Summe	35	65	100

Aus der Urne wird eine Kugel zufällig gezogen.

Mit den Daten der Tafel lassen sich direkt folgende Wahrscheinlichkeiten berechnen:

$$P(A) = \frac{70}{100} = \frac{7}{10} = 0,7 \qquad P(\overline{A}) = \frac{30}{100} = \frac{3}{10} = 0,3$$

$$P(B) = \frac{35}{100} = \frac{7}{20} = 0,35 \qquad P(\overline{B}) = \frac{65}{100} = \frac{13}{20} = 0,65$$

$$P(A \cap B) = \frac{25}{100} = \frac{1}{4} = 0,25 \qquad P(A \cap \overline{B}) = \frac{45}{100} = \frac{9}{20} = 0,45$$

$$P(\overline{A} \cap B) = \frac{10}{100} = \frac{1}{10} = 0,1 \qquad P(\overline{A} \cap \overline{B}) = \frac{20}{100} = \frac{2}{10} = 0,2$$

Die zugehörige Vierfeld - Tafel:

	B	$\overline{B}$	Summe
A	$P(A \cap B) = 0{,}25$	$P(A \cap \overline{B}) = 0{,}45$	$P(A) = 0{,}7$
$\overline{A}$	$P(\overline{A} \cap B) = 0{,}1$	$P(\overline{A} \cap \overline{B}) = 0{,}2$	$P(\overline{A}) = 0{,}3$
Summe	$P(B) = 0{,}35$	$P(\overline{B}) = 0{,}65$	1

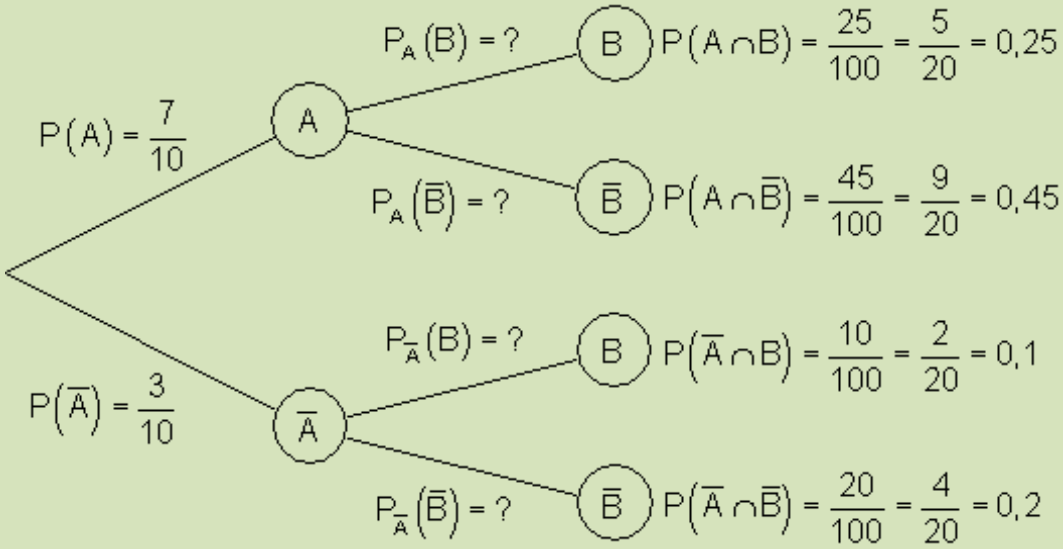
Jemand zieht eine Kugel und spürt mit der Hand, dass es sich um eine Kunststoffkugel handelt.

Wie groß ist nun die Wahrscheinlichkeit dafür, dass die Kugel in seiner Hand grün ist?

Das ist nicht die Wahrscheinlichkeit, mit der man eine grüne Kunststoffkugel zieht.

Aus der Vierfeld - Tafel lässt sich die gesuchte Wahrscheinlichkeit nicht ablesen.

Mit einem Ereignisbaum soll diese Frage nun geklärt werden.



Die Bezeichnung $P_A(B)$ bedeutet:

Die Wahrscheinlichkeit von B unter der Bedingung, dass A bereits eingetreten ist.

Diese Wahrscheinlichkeit nennen wir bedingte Wahrscheinlichkeit.

In Bezug auf die Fragestellung wird also $P_{\overline{A}}(\overline{B})$ gesucht.

In Worten:

Wie groß ist die Wahrscheinlichkeit dafür eine grüne Kugel gezogen zu haben, wenn man weiß, dass die gezogene Kugel aus Kunststoff ist.

Es wird nach einer Wahrscheinlichkeit gesucht, die von einer Bedingung abhängt.

In diesem Fall lautet die Bedingung: Die gezogene Kugel ist aus Kunststoff.

Um die im Baumdiagramm noch fehlenden Wahrscheinlichkeiten auszurechnen, verwendet man die Pfadmultiplikationsregel:

$$P(A) \cdot P_A(B) = P(A \cap B) \Leftrightarrow P_A(B) = \frac{P(A \cap B)}{P(A)}$$

Die Regel, nach der die bedingte Wahrscheinlichkeit berechnet wird, geht auf den englischen Mathematiker Thomas Bayes (1702 - 1761) zurück und wird daher auch Bayes'sche Regel oder auch Satz von Bayes genannt.

Sind A und B Ereignisse mit $P(A) \neq 0$ dann gilt: $P_A(B) = \frac{P(A \cap B)}{P(A)}$

$$P_A(B) = \frac{P(A \cap B)}{P(A)} = \frac{5}{20} : \frac{7}{10} = \frac{5 \cdot 10}{7 \cdot 20} = \frac{5}{14} \approx 0{,}36$$

$$P_A(\overline{B}) = \frac{P(A \cap \overline{B})}{P(A)} = \frac{9}{20} : \frac{7}{10} = \frac{9 \cdot 10}{7 \cdot 20} = \frac{9}{14} \approx 0{,}64$$

$$P_{\overline{A}}(B) = \frac{P(\overline{A} \cap B)}{P(\overline{A})} = \frac{2}{20} : \frac{3}{10} = \frac{2 \cdot 10}{3 \cdot 20} = \frac{1}{3} = 0{,}\overline{3}$$

$$P_{\overline{A}}(\overline{B}) = \frac{P(\overline{A} \cap \overline{B})}{P(\overline{A})} = \frac{4}{20} : \frac{3}{10} = \frac{4 \cdot 10}{3 \cdot 20} = \frac{2}{3} = 0{,}\overline{6}$$

Wenn man also weiß, dass die gezogene Kugel aus Kunststoff besteht, dann ist die Wahrscheinlichkeit dafür, dass sie Farbe grün hat: 2/3. Die Wahrscheinlichkeit eine grüne Kunststoffkugel zu ziehen ist hingegen 0,2.

Beispiel 68:

Eine Urne enthält 3 grüne und 2 rote Kugeln. Zwei Kugeln werden nacheinander ohne Zurücklegen gezogen.

Es werden vier Ereignisse definiert:

A: Grün wird im 1. Zug gezogen.

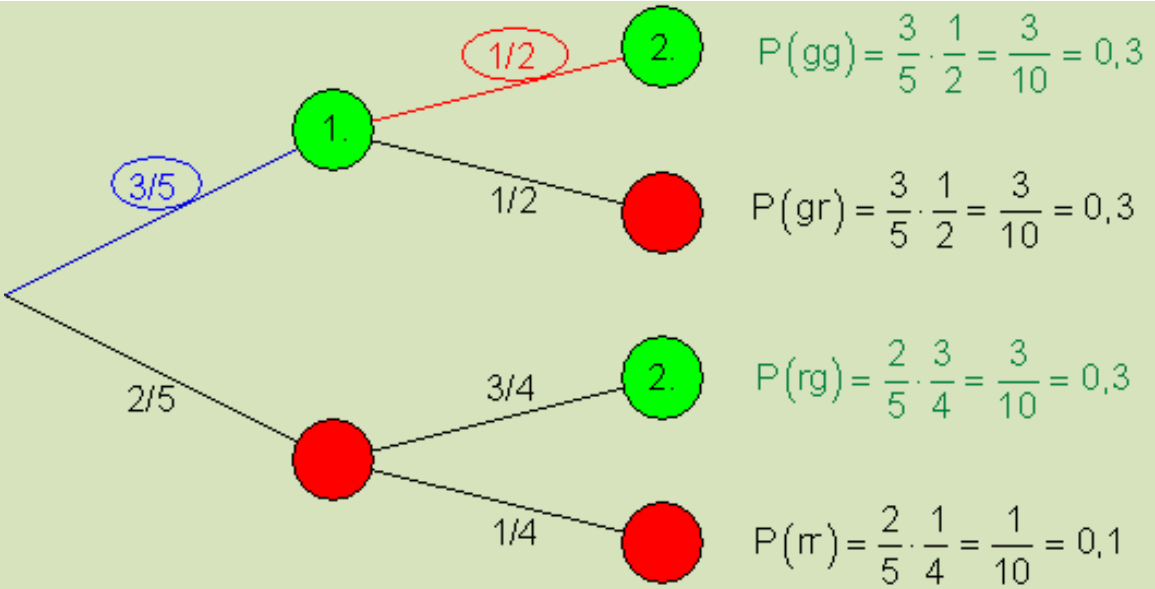
B: Grün wird im 2. Zug gezogen.

C: Grün wird im ersten und zweiten Zug gezogen.

D: Grün im zweiten Zug unter der Bedingung, dass grün bereits im ersten Zug gezogen wurde.

Zu bestimmen sind die Wahrscheinlichkeiten aller Ereignisse.

Ein Baumdiagramm mit den Pfadwahrscheinlichkeiten veranschaulicht den Zusammenhang.



Dem Baumdiagramm sind folgende Ergebnisse zu entnehmen:

Grün im 1. Zug: $P(A) = 3/5 = 0,6$ und

Grün im 2. Zug: $P(B) = \frac{3}{10} + \frac{3}{10} = \frac{6}{10} = \frac{3}{5} = 0,6$

Für grün im 1. Zug und grün im 2. Zug erhält man mit der Pfadmultiplikationsregel $P(C) = P(A \cap B) = \frac{3}{5} \cdot \frac{1}{2} = \underline{\underline{\frac{3}{10}}}$

$P(D) = \frac{1}{2}$ wird abgelesen.

Der Wert von P(D) wurde wie folgt ermittelt:

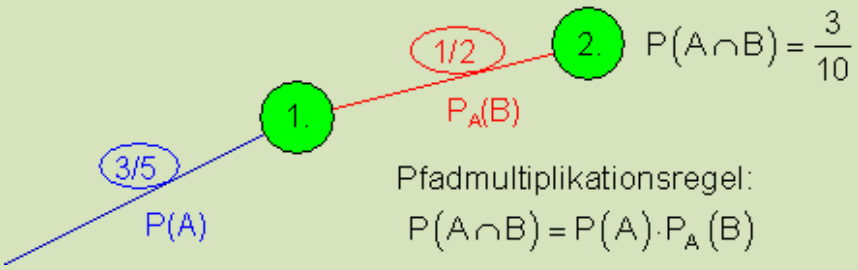
Unter der Voraussetzung (Bedingung) dass im 1. Zug grün gezogen wurde weiß man, dass noch 2 grüne und 2 rote Kugeln in der Urne sind.

Die Wahrscheinlichkeit für grün im 2. Zug ist dann 1/2.

Für die Wahrscheinlichkeit von D (grün im 2. Zug) unter der Voraussetzung dass A (grün im 1. Zug) schon eingetreten ist, wählt man die Bezeichnung $P(D) = P_A(B)$.

Im dargestellten Fall gilt: $P_A(B) = \frac{1}{2}$ $\left(\neq P(B) = \frac{3}{5}\right)$

Für eine weitere Untersuchung dient der Ausschnitt aus dem Pfaddiagramm, in dem $P_A(B)$ vorkommt.



Ist nach der Wahrscheinlichkeit $P_A(B)$ gefragt, so kann obige Gleichung wie folgt umgeformt werden:

$$P_A(B) = \frac{P(A \cap B)}{P(A)} \qquad \text{für } P(A) \neq 0$$

$P_A(B)$ ist die Wahrscheinlichkeit von B unter der Bedingung, dass A bereits eingetreten ist.

Wir überprüfen dieses Gesetz mit den vorliegenden Ergebnissen:

$$\begin{aligned} P(A \cap B) &= \frac{3}{10} \\ P(A) &= \frac{3}{5} \end{aligned} \quad \Rightarrow \quad P_A(B) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{3}{10}}{\frac{3}{5}} = \frac{3 \cdot 5}{3 \cdot 10} = \underline{\underline{\frac{1}{2}}}$$

Aus dem Urnenversuch (mehrfaches ziehen ohne zurücklegen) geht klar hervor, das die Wahrscheinlichkeit für die jeweils nächste Ziehung von der vorigen abhängt.

In einem solchen Fall sagt man, die Ereignisse sind voneinander abhängig.

Unabhängigkeit von Ereignissen

Bei einem Urnenversuch (mehrfaches ziehen mit Zurücklegen), wird die Anfangsbedingung immer wieder hergestellt, so dass die Wahrscheinlichkeit für die jeweils nächste Ziehung gleich ist, wie bei der ersten.

In einem solchen Fall sagt man, die Ereignisse sind voneinander unabhängig.

Eine Urne enthält 3 grüne und 2 rote Kugeln.

Zwei Kugeln werden nacheinander mit Zurücklegen gezogen.

Es werden vier Ereignisse definiert:

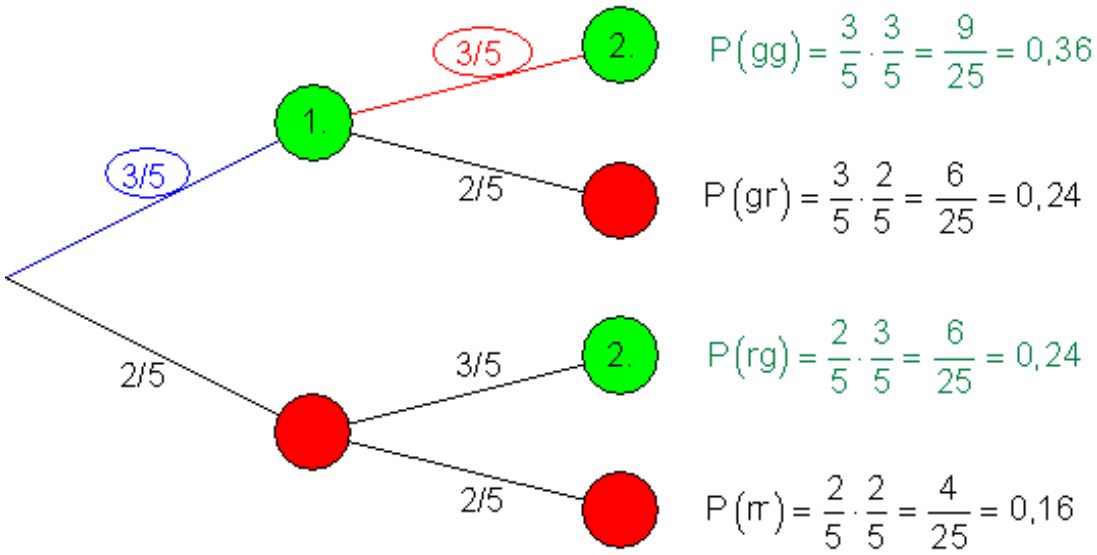
A: Grün wird im 1. Zug gezogen.

B: Grün wird im 2. Zug gezogen.

C: Grün wird im ersten und zweiten Zug gezogen.

D: Grün im zweiten Zug unter der Bedingung, dass grün bereits im ersten Zug gezogen wurde.

Das Baumdiagramm mit den zugehörigen Pfadwahrscheinlichkeiten:



Dem Baumdiagramm ist zu entnehmen:

Grün im 1. Zug: $P(A) = 3/5 = 0{,}6$

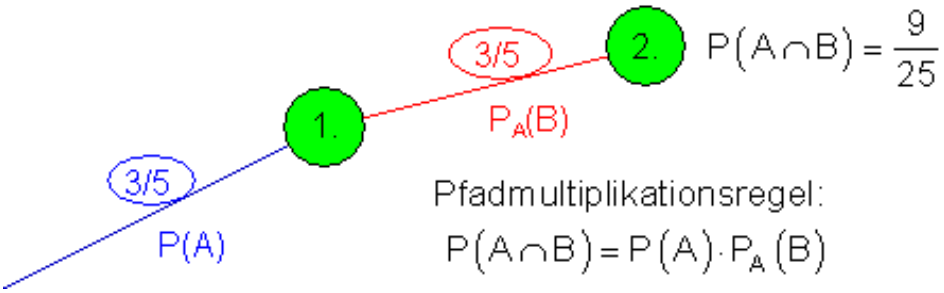
Grün im 2. Zug: $P(B) = \frac{9}{25} + \frac{6}{25} = \frac{15}{25} = \frac{3}{5} = 0{,}6$

Für Grün im 1. Zug und grün im 2. Zug erhält man mit der Pfadmultiplikationsregel $P(C) = P(A \cap B) = \frac{3}{5} \cdot \frac{3}{5} = \underline{\underline{\frac{9}{25}}}$

$P(D) = \frac{3}{5}$ wird abgelesen.

Die Wahrscheinlichkeit eine grüne Kugel zu ziehen bleibt immer gleich, da nach jedem Zug durch Zurücklegen der Kugel, die Ausgangssituation wieder hergestellt wird. Die Wahrscheinlichkeit für grün im 2. Zug unter der Bedingung, das grün im 1. Zug bereits gezogen wurde ist $P(D) = P_A(B)$.

Ein Ausschnitt aus dem Baumdiagramm:



Pfadmultiplikationsregel:

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

Eine Auflistung der Ergebnisse ergibt:

$$P(A) = \frac{3}{5} \qquad P_A(B) = \frac{3}{5} \qquad P(B) = \frac{3}{5}$$

es ist also $P_A(B) = P(B)$

Damit gilt für die Pfadmultiplikationsregel:

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

Gilt $P_A(B) = P(B)$, so beeinflusst das Eintreten des Ereignisses A die Wahrscheinlichkeit von B nicht.

Man sagt, die Ereignisse A und B sind unabhängig voneinander.

Unabhängige Ereignisse

Das Ereignis B heißt unabhängig vom Ereignis A, wenn das Eintreten von A die Wahrscheinlichkeit für das Eintreten von B nicht beeinflusst.

Es gilt: $P(A \cap B) = P(A) \cdot P(B)$

Beispiel: Urnenziehung mit zurücklegen.

Merke:

Für den Nachweis der Unabhängigkeit zweier Ereignisse A und B geht man wie folgt vor:
Man berechnet $P(A); P(B)$ und $P(A \cap B)$.
Gilt $P(A \cap B) = P(A) \cdot P(B)$,
so sind die Ereignisse A und B voneinander unabhängig.

Beispiel 69:

Eine Umfrage an Schulen über die Essgewohnheiten der Schüler hat ergeben, dass 45% aller Schüler gerne Schokolade essen.

55% aller Schüler ziehen andere Süßigkeiten vor.

60% aller Schüler gaben an Geschwister zu haben.

27% der Schüler mit Geschwistern essen gerne Schokolade.

Eine Schokoladefabrik interessiert sich dafür, ob Schüler mit Geschwister eine besondere Vorliebe für Schokolade haben.

Anders ausgedrückt:

Hat die Tatsache, dass ein Schüler Geschwister hat, einen Einfluss auf seine Vorliebe für Schokolade?

Die Erhebungsdaten lassen sich in einer Vierfeld - Tafel darstellen:

	B	$\overline{B}$	Summe
A	$P(A \cap B) = 0,27$	$P(A \cap \overline{B}) = 0,33$	$P(A) = 0,6$
$\overline{A}$	$P(\overline{A} \cap B) = 0,18$	$P(\overline{A} \cap \overline{B}) = 0,22$	$P(\overline{A}) = 0,4$
Summe	$P(B) = 0,45$	$P(\overline{B}) = 0,55$	1

Die zugehörigen Ereignisse sind:

A: Der Schüler hat Geschwister.

B: Der Schüler isst gerne Schokolade.

Überprüfung auf Abhängigkeit:

$$\left.\begin{array}{l} P(A) = 0,6 \\ P(B) = 0,45 \\ P(A \cap B) = 0,27 \end{array}\right| \Rightarrow P_A(B) = \frac{P(A \cap B)}{P(A)} = \frac{0,27}{0,6} = 0,45 = P(B)$$

Die Ereignisse sind unabhängig voneinander.

Das Bedeutet, ob ein Schüler Geschwister hat oder nicht, hat keinen Einfluss auf seine Vorliebe für Schokolade.

Allgemeines zu Verteilungen

Vergleich der verschiedenen Verteilungen

Wann benutze ich welche Verteilung?

Diskrete Verteilung

Bernoulli- oder Binomialverteilung

- Zufallsexperiment ist durch zwei mögliche Versuchsausgänge gekennzeichnet.
- Ziehen mit zurücklegen

Hypergeometrische Verteilung

- Ziehen ohne zurücklegen
- Ansonsten Binomialverteilung

Poisson-Verteilung

- Größere Anzahl von n (Stichprobenumfang)
- Wahrscheinlichkeit für das Auftreten eines Ereignisses sehr gering
- Ansonsten wie die Binomial-Verteilung

Kontinuierliche Verteilungen

Exponential-Verteilung

- Bei physikalischen Problemen, Wachstums oder Zerfallsprozessen

Weibull-Verteilung

- Lebensdauer von Systemen und Festigkeiten von Materialien in einem bestimmten Zeitraum

Gauß'sche Normalverteilung

- Erwartungsverteilung für eine unendlich große Grundgesamtheit, die real aber immer nur durch eine oder mehrere Stichproben charakterisiert werden.

Diskrete Verteilungen

Im Folgenden sollen die wichtigsten diskreten Verteilungen besprochen werden. Dazu zählen vor allem die Binomialverteilung, die hypergeometrische- und die Poisson – Verteilung.

Binomialverteilung

Ausgangspunkt sind Verteilungen von Experimenten mit zwei alternativen Ausgangsmöglichkeiten, welche aber gleiche Wahrscheinlichkeiten besitzen.

Die Binomialverteilung ist die Wahrscheinlichkeitsfunktion für die Zufallsvariable „Häufigkeiten des Auftretens von einem Ereignis bei n Bernoulli Experimenten".

Die Einzelwahrscheinlichkeiten für die beiden möglichen Ereignisse addiert sich zu 1.

Gesucht ist also die jeweilige Wahrscheinlichkeit für die möglichen Kombinationen der Alternativereignisse. Es ist einsichtig, dass 7mal die Zahl zu werfen bei 10 Münzwürfen, weniger wahrscheinlich ist, als 6/4 oder 5/5, die Kombi mit der höchsten Wahrscheinlichkeit. Diese hängt natürlich auch von der Anzahl der Bernoulli Versuchen ab.

Man kann also der Binomialverteilung entnehmen, wie oft ein Ereignis zu erwarten ist, wenn ich die Wahrscheinlichkeit für die Ereignisse und die Anzahl der Versuche kenne.

Beispiel 70:

Ein Ereignis A mit der Wahrscheinlichkeit von 0,25 tritt mit einer Wahrscheinlichkeit von 0,0186 genau 7-mal auf, wenn wir den Versuch 13mal durchführen. Dabei ist die Auftretenswahrscheinlichkeit der Wert der Binomialverteilung.

Auch die Summe der Auftretenswahrscheinlichkeiten ergeben 1. Genauso kann natürlich berechnet werden, wie hoch die Wahrscheinlichkeiten sind, dass A höchstens bzw. mindesten k- mal auftritt. (Summe der Wahrscheinlichkeiten von k bis n, bzw. von 0 bis k. Die zweite Möglichkeit entspricht also auch der Verteilungsfunktion einer Binomialverteilung.

Allgemein:

Diese Funktion definiert die Wahrscheinlichkeit der Häufigkeiten für das Auftreten eines Alternativereignisses A in n Versuchen, wenn A mit der Wahrscheinlichkeit von p eintritt. Diese Wahrscheinlichkeitsfunktion heißt Binomialverteilung mit den Parametern n und p.

Hypergeometrische Verteilungen

Bei der Binomialverteilung wird vorausgesetzt, dass die Wahrscheinlichkeiten der einzelnen Ereignisse stets gleich bleiben. Bei einem Münzwurf ist dies auch gegeben. Stellen wir uns allerdings eine Urne vor, in der sich 5 rote und 5 schwarze Kugeln befinden und wir ziehen einzelne Kugeln, so verändern sich die Wahrscheinlichkeiten von Zug zu Zug.

Die Binomialverteilung könnten wir in diesem Falle nur anwenden, wenn wir die Kugeln wieder zurücklegen. Tun wir dies nicht müssen wir die hypergeometrische Verteilung anwenden.

Anders formuliert gibt die Wahrscheinlichkeitsfunktion der hypergeometrischen Verteilung an, mit welcher Wahrscheinlichkeit die Zufallsvariable „A oder alternative zu A“ einen bestimmten Wert annimmt. Diese ist deshalb eine Zufallsvariable, da es vom Zufall abhängt, welche Kugel z. B. gezogen wird, und wie sich die „neuen Wahrscheinlichkeiten gestalten.

Beschreibende Parameter:

N: Gesamtzahl der Objekte

K: Anzahl der Alternative A (und N-K Objekte für A-quer)

n: Größe der Stichprobe (Bsp.: man will 4 rote aus 10 Kugeln n=4, N=10)

k: Häufigkeit der Alternative A (entsprechend n-k für A-quer)

Die Wahrscheinlichkeiten werden auch hier über die Regel „Anzahl der günstigen Fälle durch Anzahl der Möglichen Fälle berechnet, dies muss aber nach jedem Durchgang erneut getan werden. Diese Anzahl wird durch die 2. Kombinationsregel N über n berechnet.

Die Hypergeometrische Verteilung beschreibt also genau wie die Binomialverteilung eine Auftretenswahrscheinlichkeit, jedoch ist die Wahrscheinlichkeit p des Eintretens von A zufällig und nicht gleich bleibend.

Poisson – Verteilung

Diese ist die Verteilung seltener Ereignisse. Ist die Anzahl n der möglichen Ereignisse sehr groß und die Wahrscheinlichkeit p des Auftretens von A sehr gering, wird die Berechnung über die Binomialverteilung sehr umständlich.

Deshalb lässt sich dieser Wert über die Poisson – Verteilung approximieren. Für unendlich großes n und p = 0 gehen diese beiden Verteilungen ineinander über. Mittelwert und Varianz dieser Verteilungen sind identisch: n*p.

Stetige Verteilungen

Stetige Verteilungen sind theoretische Verteilungen, die einen „optimalen“ Zustand repräsentieren. Die wohl wichtigste ist die Normalverteilung.

Normalverteilung

Die Normalverteilung beschreibt ähnlich wie die bislang kannengelernten diskreten Verteilungen eine Klasse von Verteilungen. Diese haben bestimmte Eigenschaften, die nun aufgezeigt werden sollen:

- Glockenkurve
- Symmetrisch
- Modalwert, Mittelwert und Median fallen zusammen
- Die Kurve nähert sich asymptotisch der x- Achse
- 2/3 aller Fälle befinden sich zwischen den Wendepunkten der Kurve

Unterschiede in der Form einer Normalverteilung sind auf Unterschiedliche Streuungen und Erwartungswerte zurückzuführen.

Da zwei Normalverteilungen mit gleichem Mittelwert und Standardabweichung identisch sind, werden sie durch diese beiden Parameter eindeutig beschrieben. Diese sind daher auch maßgeblich bei der Bestimmung der Dichtefunktion (Wahrscheinlichkeitsfunktion) einer NV.

Unter diesen Normalverteilungen gibt es eine mit dem Mittelwert 0 und einer Standartabweichung von 1. Diese wird dann als Standartnormalverteilung bezeichnet. Sie ist von größter Bedeutung, da sämtliche NV in diese durch die z- Transformation transformiert werden können. Über diese Transformation lässt sich für beliebige Bereiche unter der Kurve die Verteilungsfunktion errechnen.

Also, mit welcher Wahrscheinlichkeit ein bestimmter Wert auftritt oder nicht. So ermitteln wir für den Bereich zwischen –1z und 1z eine Wahrscheinlichkeit von 68,26%. Entsprechend also der Bedingung, dass im Bereich von +/- 1 68% aller Fälle liegen müssen.

Weibull-Verteilung

Die Weibull-Verteilung ist eine statistische Verteilung, die beispielsweise zur Untersuchung von Lebensdauern in der Qualitätssicherung verwendet wird. Man verwendet sie vor allem bei Fragestellungen wie Materialermüdungen von spröden Werkstoffen oder Ausfällen von elektronischen Bauteilen, ebenso bei statistischen Untersuchungen von Windgeschwindigkeiten. Benannt ist sie nach dem Schweden Waloddi Weibull (1887-1979).

Ein anschauliches Beispiel für die Anwendung der Weibull-Statistik ist die Ausfallwahrscheinlichkeit einer Kette. Das Versagen eines Glieds führt zum Festigkeitsverlust der ganzen Kette. Spröde Werkstoffe zeigen ein ähnliches Bruchverhalten. Es genügt ein Riss, der die kritische Risslänge überschreitet, um das Bauteil zu zerstören.

So erklärt sich auch die Abhängigkeit der Festigkeit spröder Werkstoffe von der Geometrie. Das Verlängern einer Kette (bzw. eines spröden Bauteils) reduziert die Festigkeit, eine Verstärkung der Kettenglieder (bzw. Vergrößerung des Bauteil-Querschnitts) erhöht sie.

Die Weibull-Verteilung kann zur Beschreibung steigender, konstanter und fallender Ausfallraten technischer Systeme verwendet werden.

In der Praxis ist die Weibull-Verteilung neben der Exponentialverteilung die am häufigsten verwendete Lebensdauerverteilung

Exponentialverteilung

Die Exponentialverteilung ist die Wahrscheinlichkeitsverteilung, mit der die zeitlichen Abstände eines ungestörten, poissonverteilten Verkehrsstroms beschrieben werden können. Daraus werden Formeln zur Leistungsfähigkeit von Verkehrsknoten abgeleitet. Sie wird auch bei Lebensdauer- und Zuverlässigkeitstests eingesetzt.

Anwendungen der Exponentialverteilung

Zufallsvariablen, bei denen die Zeit eine entscheidende Rolle spielt, sind häufig exponential verteilt. Beispiele dafür sind

- Dauer von Telefongesprächen
- Lebensdauer des radioaktiven Zerfalls
- Arbeitszeit einer Maschine zwischen zwei Stillständen
- Lebensdauer von Bauteilen oder Lebewesen

Binomialverteilung

Ausführliche Vorbetrachtung

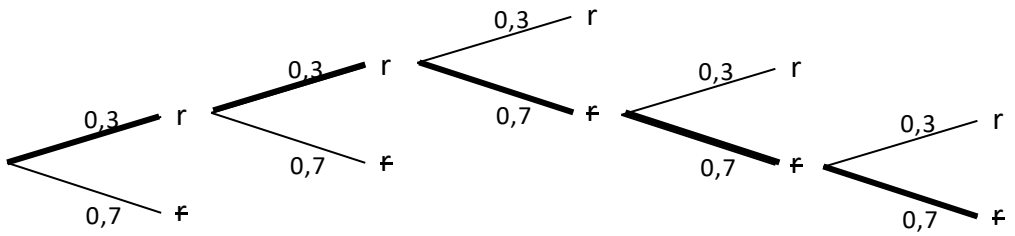
Beispiel 71:

In einer Urne befinden sich 10 Kugeln, davon 3 rote.

Wie ziehen 5-mal mit Zurücklegen und notieren das Ergebnis mit Beachtung der Reihenfolge.

Wir untersuchen nun verschiedene Ereignisse:

A: Genau die ersten beiden gezogenen Kugeln sind rot. Berechne die Wahrscheinlichkeit.

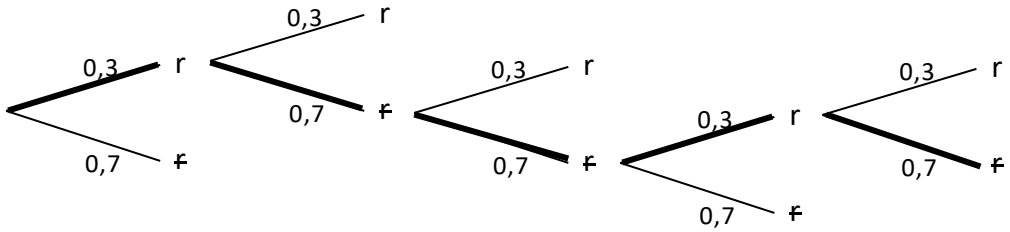


Mit Hilfe eines Baumdiagramms und der Pfadregel errechnen wir:

P(A) = P (r r~~fff~~) = $0{,}3 \cdot 0{,}3 \cdot 0{,}7 \cdot 0{,}7 \cdot 0{,}7 = 0{,}3^2 \cdot 0{,}7^3$ = 0,03087 $\approx$ 3,1%

B: Genau die 1. und die 4. gezogene Kugel sind rot. Berechne die Wahrscheinlichkeit.

Mit Hilfe eines Baumdiagramms und der Pfadregel erhalten wir dieselbe Wahrschein-



lichkeit. wie bei A:

P(B) = P (r ~~fff~~r) = $0{,}3 \cdot 0{,}7 \cdot 0{,}7 \cdot 0{,}3 \cdot 0{,}7 = 0{,}3^2 \cdot 0{,}7^3$ = 0,03087 $\approx$ 3,1%

C: Genau 2 der 5 gezogenen Kugeln sind rot. Berechne die Wahrscheinlichkeit.

Wir finden nun mit Hilfe des Baumdiagramms genau 10 verschiedene mögliche Pfade.

P(C) = P (r r~~fff~~) + P (r ~~f~~rf~~f~~) + P (r ~~ff~~rf~~f~~) + P (r ~~fff~~r)

\+ P (~~f~~rr~~ff~~) + P (~~f~~r~~f~~r~~f~~) + P (~~f~~r~~ff~~r)

\+ P (~~ff~~rr~~f~~) + P (~~ff~~r~~f~~r) + + P (~~fff~~r)

Also ist

P(C) = $10 \cdot 0{,}3^2 \cdot 0{,}7^3 = \binom{5}{2} \cdot 0{,}3^2 \cdot 0{,}7^3$ = 0,3087 $\approx$ 30,9%

Bernoulli-Experiment, Bernoulli-Kette

Definition 72:

Ein Zufallsexperiment mit nur 2 möglichen Ergebnissen (welche oftmals z.B. als „Treffer“ und „Niete“ bezeichnet werden) heißt **Bernoulli-Experiment**.

Beispiele von Bernoulliexperimenten:

(1) Ziehen von Kugeln aus einer Urne, die genau zwei Sorten enthält.

(2) Werfen einer Münze (Wappen - Zahl)

(3) Auswahl von Schüler (männlich - weiblich)

(4) Würfeln - wenn man etwa 6 oder nicht 6 unterscheidet bzw. 1 oder nicht 1.

(5) Testen eines Gerätes: defekt - gut

Beispiel 72:

Einmaliges Ziehen aus der o. g. Urne und überprüfen, ob man r gezogen hat, ist ein Bernoulli-Experiment.

Die Wahrscheinlichkeit für „Treffer“ (r) ist $p = 0{,}3$.

Die Wahrscheinlichkeit für „Niete“ ($\not{r}$) ist $q = 1 - p = 0{,}7$

Definition 73:

Ein n-stufiges Bernoulli-Experiment heißt **Bernoulli-Kette der Länge *n***.

Beispiel 73:

Fünfmaliges Ziehen aus der o. g. Urne und überprüfen, ob man r gezogen hat, ist eine Bernoulli-Kette der Länge 5.

Die Formel von Bernoulli, Binomialverteilung

Definition 74:

Bei einer Bernoulli-Kette der Länge n mit Trefferwahrscheinlichkeit p kann man die Wahrscheinlichkeit für die Anzahl k der Treffer nach der **Bernoulli-Formel** berechnen:

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

(X ist die Zufallsvariable für die Anzahl der Treffer)

Beispiel 74:

Man muss an dieser formellen Darstellung der Bernoulli-Formel nicht erschrecken; wir haben mit ihr bereits gerechnet, nämlich bei der Berechnung der Wahrscheinlichkeit des Ereignisses C (s. o.).

Hierbei war einfach $n = 5$, $k = 2$, $p = 0{,}3$:

$$B_{5;0,3}(2) = P(X = 2) = \binom{5}{2} \cdot 0{,}3^2 \cdot 0{,}7^3 \approx 30{,}9\%$$

Definition 75:

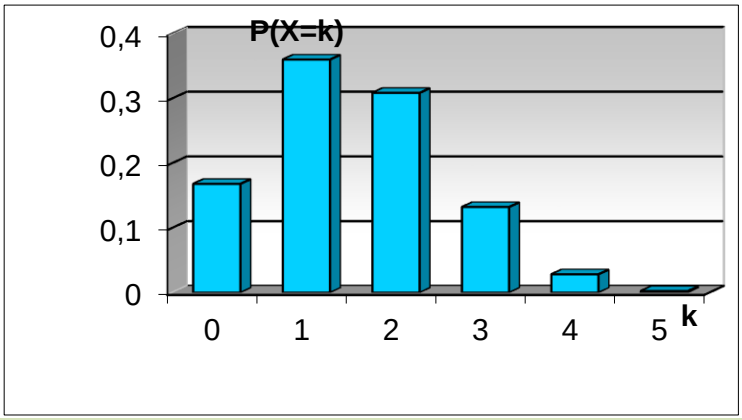
Die nach der Bernoulli-Formel berechnete Wahrscheinlichkeitsverteilung P($X = k$) heißt Binomialverteilung $B_{n;p}(k) = b(k;n;p)$.

Beispiel 75:

Wir berechnen genauso wie für Ereignis C jeweils die Wahrscheinlichkeit, dass genau 0, 1, 2, 3, 4 oder 5 der gezogenen Kugeln rot sind und erhalten folgende Wahrscheinlichkeitsverteilung:

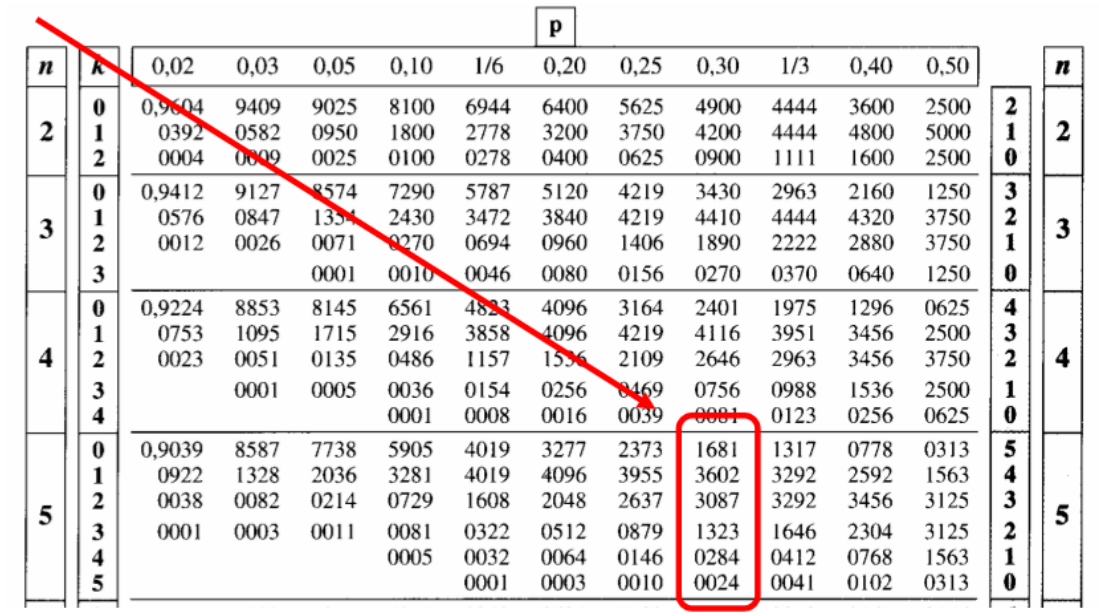
k	0	1	2	3	4	5
P($X = k$)	$0{,}7^5$ = 0,16807 ≈16,8%	$5 \cdot 0{,}3 \cdot 0{,}7^4$ = 0,36015 ≈36,0%	$10 \cdot 0{,}3^2 \cdot 0{,}7^3$ = 0,3087 ≈30,9%	$10 \cdot 0{,}3^3 \cdot 0{,}7^2$ = 0,1323 ≈13,2%	$5 \cdot 0{,}3^4 \cdot 0{,}7$ = 0,02835 ≈2,8%	$0{,}3^5$ = 0,00243 ≈0,2%

Anschaulicher als eine Tabelle ist die grafische Darstellung in Form eines Stabdiagrammes:



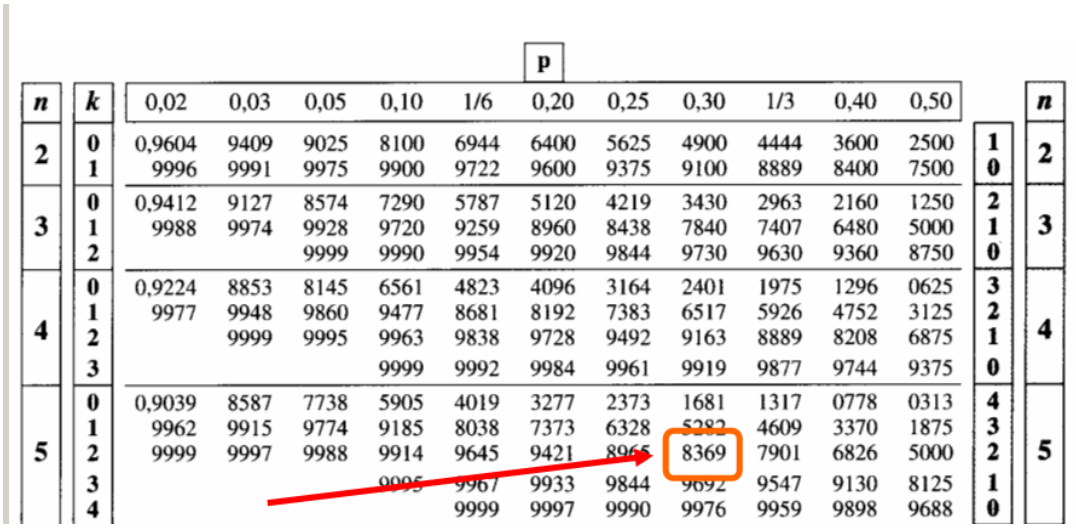
Praxis der Binomialverteilung

In der Praxis muss man solche binomialverteilten Wahrscheinlichkeiten nicht jedes Mal von Hand berechnen, sondern man verwendet tabellierte Werte.



		p											
n	k	0,02	0,03	0,05	0,10	1/6	0,20	0,25	0,30	1/3	0,40	0,50	n
2	0	0,9604	9409	9025	8100	6944	6400	5625	4900	4444	3600	2500	2
	1	0392	0582	0950	1800	2778	3200	3750	4200	4444	4800	5000	1
	2	0004	0009	0025	0100	0278	0400	0625	0900	1111	1600	2500	0
3	0	0,9412	9127	8574	7290	5787	5120	4219	3430	2963	2160	1250	3
	1	0576	0847	1354	2430	3472	3840	4219	4410	4444	4320	3750	2
	2	0012	0026	0071	0270	0694	0960	1406	1890	2222	2880	3750	1
	3			0001	0010	0046	0080	0156	0270	0370	0640	1250	0
4	0	0,9224	8853	8145	6561	4823	4096	3164	2401	1975	1296	0625	4
	1	0753	1095	1715	2916	3858	4096	4219	4116	3951	3456	2500	3
	2	0023	0051	0135	0486	1157	1536	2109	2646	2963	3456	3750	2
	3		0001	0005	0036	0154	0256	0469	0756	0988	1536	2500	1
	4				0001	0008	0016	0039	0081	0123	0256	0625	0
5	0	0,9039	8587	7738	5905	4019	3277	2373	1681	1317	0778	0313	5
	1	0922	1328	2036	3281	4019	4096	3955	3602	3292	2592	1563	4
	2	0038	0082	0214	0729	1608	2048	2637	3087	3292	3456	3125	3
	3	0001	0003	0011	0081	0322	0512	0879	1323	1646	2304	3125	2
	4				0005	0032	0064	0146	0284	0412	0768	1563	1
	5					0001	0003	0010	0024	0041	0102	0313	0

hier:



		p											
n	k	0,02	0,03	0,05	0,10	1/6	0,20	0,25	0,30	1/3	0,40	0,50	n
2	0	0,9604	9409	9025	8100	6944	6400	5625	4900	4444	3600	2500	2
	1	9996	9991	9975	9900	9722	9600	9375	9100	8889	8400	7500	1
3	0	0,9412	9127	8574	7290	5787	5120	4219	3430	2963	2160	1250	3
	1	9988	9974	9928	9720	9259	8960	8438	7840	7407	6480	5000	2
	2			9999	9990	9954	9920	9844	9730	9630	9360	8750	1
4	0	0,9224	8853	8145	6561	4823	4096	3164	2401	1975	1296	0625	4
	1	9977	9948	9860	9477	8681	8192	7383	6517	5926	4752	3125	3
	2		9999	9995	9963	9838	9728	9492	9163	8889	8208	6875	2
	3				9999	9992	9984	9961	9919	9877	9744	9375	1
5	0	0,9039	8587	7738	5905	4019	3277	2373	1681	1317	0778	0313	5
	1	9962	9915	9774	9185	8038	7373	6328	5282	4609	3370	1875	4
	2	9999	9997	9988	9914	9645	9421	8965	8369	7901	6826	5000	3
	3				9995	9967	9933	9844	9692	9547	9130	8125	2
	4					9999	9997	9990	9976	9959	9898	9688	1

Die hier eingerahmte Zahl 0,8369 gibt z.B. an, wie hoch die Wahrscheinlichkeit ist, dass bei 5-maligem Ziehen einer Kugel aus der Urne höchstens 2-mal eine rote Kugel gezogen wird:

Wir können die Zahl auch direkt aus unserer Wahrscheinlichkeitsverteilung berechnen:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

=0,16807+0,36015+0,3087=0,83692 $\approx 83{,}7\%$

Erwartungswert, Varianz, Standardabweichung einer Binomialverteilung

Betrachten wir nun ein neues Beispiel, an dem wir uns zunächst den Begriff „Erwartungswert“ veranschaulichen:

Beispiel 76:

„20-maliges Würfeln und feststellen, wie viele Einsen geworfen wurden.“

Es erscheint uns seit Kindesalter als äußerst unwahrscheinlich, dass von 20 Würfen z.B. 15-mal die Eins auftritt, dagegen erscheint es realistisch, dass die Eins z.B. 3-mal auftritt.

Je öfter wir würfeln, umso mehr „erwarten“ wir etwa in einem Sechstel der Würfe die Eins.

Bei 20 Würfen erwarten wir also $\frac{1}{6} \cdot 20 = 3,\overline{3}$ mal die Eins.

Formeller ausgedrückt:

Der Erwartungswert der $B_{20;0,3}$-verteilten Zufallsvariable *X* ist

$E(X) = 20 \cdot \frac{1}{6} = 3,\overline{3}.$

Definition 76:

Eine binomialverteilte Zufallsvariable *X* hat den **Erwartungswert** $E(X) = \mu = n \cdot p$.

In unserem o. g. Urnenbeispiel wäre also der Erwartungswert

$E(X) = 5 \cdot 0,3 = 1,5.$

Beispiel 77:

In einer Urne befinden sich Kugeln, und zwar 30% rote und 70% blaue.
Wir entnehmen daraus 20 Kugeln und legen sie jeweils wieder zurück.
Dann liegt eine 20-stufige Bernoullikette vor, und die Wahrscheinlichkeit wird mit der Binomialverteilung berechnet.
Es sei X die Zufallsvariable "Zahl der roten Kugeln". Dann gilt z.B:

a) $P(X = 5) = \binom{20}{5} \cdot 0,3^5 \cdot 0,7^{15} = 0,1789$. Oder aus einer Tafel:

$f_B(5;20;0,3) = 0,1789$
Die Wahrscheinlichkeit, unter diesen 20 Kugeln 5 rote zu haben, ist also 0,1789.

b) $P(X = 6) = \binom{20}{6} \cdot 0,3^6 \cdot 0,7^{14} = 0,1916$. Oder aus einer Tafel:

$f_B(6;20;0,3) = 0,1916$
Die Wahrscheinlichkeit, unter diesen 20 Kugeln 6 rote zu haben, ist also 0,1916.

c) $P(X=7)=\binom{20}{7}\cdot 0{,}3^7\cdot 0{,}7^{13}=0{,}1643$. Oder aus einer Tafel:

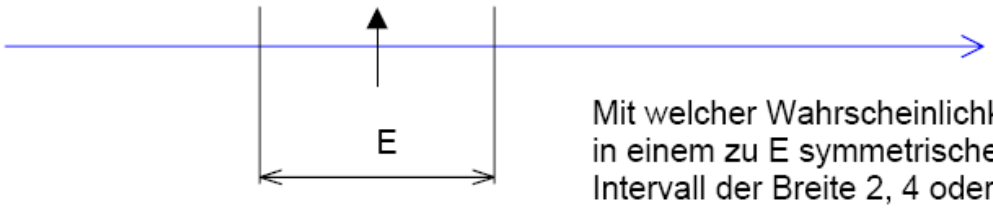
$f_B(7;20;0{,}3)=0{,}1643$

Die Wahrscheinlichkeit, unter diesen 20 Kugeln 7 rote zu haben, ist also 0,1643.

Der Erwartungswert für X ist $E(X) = np = 6$.
Man kann also durchschnittlich mit 6 roten Kugeln rechnen.
Dieser Erwartungswert setzt eigentlich unendlich viele Experimente voraus, und davon ist es der Mittelwert. Wir sehen, daß für diese Zahl auch die größte Wahrscheinlichkeit vorliegt. Die anderen Werte liegen darunter.

Die Ergebnismenge für ein solches Experiment ist S:

S = { 0 ; 1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9 ; 10 ; 11 ; 12 ; ... ; 19 ; 20 }



Mit welcher Wahrscheinlichkeit liegt X in einem zu E symmetrischen Intervall der Breite 2, 4 oder 6?

Intervalle, die zu E=6 symmetrisch sind:

{ 5 ; E=6; 7 } hat die Breite 2 (7 minus 5 = 2), also den Radius 1, wir haben eine Zahl links von E und eine rechts von E.

$P(5\le X\le 7)=F_B(7;20;0{,}3)-F_B(4;20;0{,}3)=0{,}7723-0{,}2375=0{,}5348$

{ 4 ; 5 ; 6 ; 7 ; 8 } hat die Breite 4 d.h. den Radius 2, denn dieses Intervall reicht von E aus um 2 nach links und um 2 nach rechts.

$P(4\le X\le 8)=F_B(8;20;0{,}3)-F_B(3;20;0{,}3)=0{,}8867-0{,}1071=0{,}7796$

{ 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 9} hat die Breite 6, also den Radius 3:

$P(3\le X\le 9)=F_B(9;20;0{,}3)-F_B(2;20;0{,}3)=0{,}9520-0{,}0355=0{,}9165$

Wir können diese Ergebnisse beispielsweise für Tests verwenden. Wollen wir durch Entnahme von 20 Kugeln testen, ob der Anteil der roten mit 30% glaubhaft ist. Dann können wir uns darauf festlegen, daß wir dies glauben wollen, wenn wir 3 bis 9 rote Kugeln ziehen. Die Wahrscheinlichkeit dafür, daß dies eintritt, ist 92,65 %. Der Zufall, daß wir ein anderes Ergebnis bekommen, ist also mit 7,35 % sehr klein.

Definition 77:

Die **Standardabweichung** σ ist ein Maß dafür, wie „breit“ die Binomialverteilung „gestreut“ ist.

Man berechnet die **Varianz** V(*X*) einer binomialverteilten Zufallsvariable X nach $V(X)=\sigma^2=n\cdot p\cdot(1-p)$.

Die Standardabweichung σ ist dabei einfach die Wurzel aus der Varianz σ^2.

Die Mathematiker haben für die Theorie der Binomialverteilung eine Größe ermittelt, die sehr hilfreich für solche Untersuchungen sind. Er ist die sogenannte Standardabweichung. Man verwendet dafür den griechischen Buchstaben Sigma: σ.

Für die Binomialverteilung gilt:

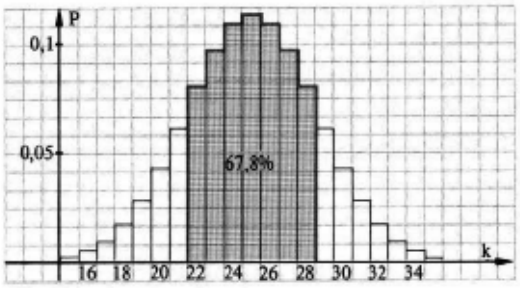
Standardabweichung: $\sigma = \sqrt{n \cdot p \cdot (1-p)}$

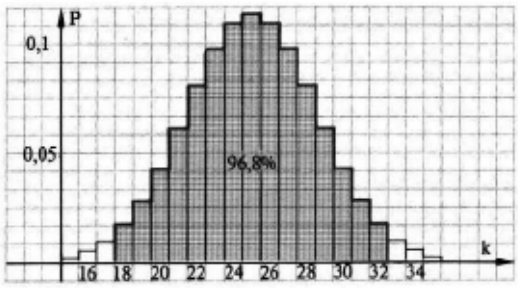
Varianz $V(X) = np(1-p) = \sigma^2$

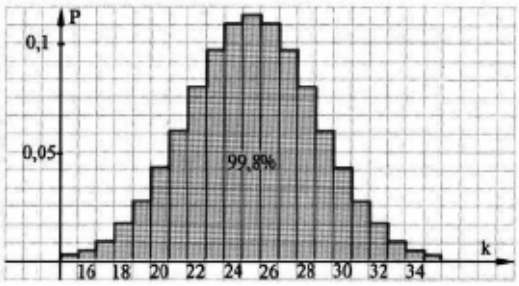
Mit dieser Standardabweichung kann man drei sogenannte Sigma-Umgebungen definieren.

Für eine n-stufige Bernoullikette mit der Zufallsvariable X und wenn $\sigma > 3$ ist gilt näherungsweise:

1. Die Wahrscheinlichkeit, daß X in der 1σ-Umgebung liegt, d.h.
   $E - \sigma \le X \le E + \sigma$, ist ca. 68 %.
   d.h. $P(E - \sigma \le X \le E + \sigma) \approx 0{,}68$
   d.h. $P(|X - E| \le \sigma) \approx 0{,}68$
2. Die Wahrscheinlichkeit, daß X in der 2σ-Umgebung liegt, d.h.
   $E - 2\sigma \le X \le E + 2\sigma$, ist ca. 95,5 %.
   d.h. $P(E - 2\sigma \le X \le E + 2\sigma) \approx 0{,}955$
   d.h. $P(|X - E| \le 2\sigma) \approx 0{,}955$
3. Die Wahrscheinlichkeit, daß X in der 3σ-Umgebung liegt, d.h.
   $E - 3\sigma \le X \le E + 3\sigma$, ist ca. 99,7 %.
   d.h. $P(E - 3\sigma \le X \le E + 3\sigma) \approx 0{,}997$
   d.h. $P(|X - E| \le 3\sigma) \approx 0{,}997$

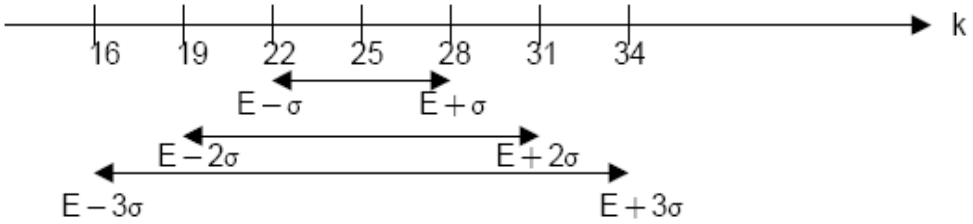






Beispiel 78:

Bei einem Bernoulli-Experiment sei p = 0,5 und n = 50. Dann folgt E(X) = np = 25 und $\sigma = \sqrt{np(1-p)} = \sqrt{50 \cdot 0{,}5 \cdot 0{,}5} = 3{,}5 > 3$. Also gilt näherungsweise:



Mit 95,5 % Wahrscheinlichkeit liegt also X im Intervall { 19; 20; ...; 31 }.
Die drei Histogramme rechts oben zeigen diese Situationen.

Hypergeometrische Verteilung

Die Hypergeometrische Verteilung ist eine diskrete Wahrscheinlichkeitsverteilung in der Stochastik. Umgangssprachlich werden Fragestellungen, die von der hypergeometrischen Verteilung erfasst werden auch als **Ziehen ohne Zurücklegen** bezeichnet.

Sie wird verwendet, um Vorgänge zu modellieren bei denen aus einer Menge zufällig eine Stichprobe entnommen und auf eine bestimmte Eigenschaft geprüft wird. Die zu Grunde liegende Menge kann daher als Grundgesamtheit bezeichnet werden.

Die hypergeometrische Verteilung gibt dann Auskunft darüber mit welcher Wahrscheinlichkeit in der Stichprobe eine bestimmte Anzahl von Elementen vorkommt, die die gewünschte Eigenschaft haben. Bedeutung kommt dieser Verteilung daher etwa bei Qualitätskontrollen zu.

Beispiel 79:

In einer Urne befinden sich 45 Kugeln, 20 davon sind gelb. Wie hoch ist die Wahrscheinlichkeit in einer 10-elementigen Stichprobe 4 gelbe Kugeln zu ziehen? - Das Beispiel wird unten durchgerechnet.

Definition 78:

Mit Hilfe der hypergeometrischen Verteilung wird die Frage “Wie groß ist die Wahrscheinlichkeit in der Stichprobe (Stichprobenumfang) genau x fehlerhafte Einheiten/Objekte vorzufinden?” beantwortet.

Bemerkung 45:

Da aus dem Losumfang (Grundgesamtheit) eine zufällig gezogene Stichprobe ohne zurückzulegen, was in der realen Welt der Probenentnahme oft auch nicht möglich oder erwünscht (Rückstellmuster) ist, genommen wird, kann hier nicht die Binomialverteilung angewendet werden.

Formalisierung

Definition 79:

Die hypergeometrische Verteilung ist abhängig von drei Parametern:

- Der Elementzahl einer Grundgesamtheit (im Folgenden als **N** bezeichnet).
- Der Zahl der Elemente mit einer bestimmten Eigenschaft in dieser Grundmenge (bezeichnet als **M**).
- Der Zahl der Elemente in einer Stichprobe die gezogen werden. (bezeichnet als **n**).
- Die Verteilung gibt nun Auskunft darüber, wie wahrscheinlich es ist, dass sich **x** Elemente mit der zu prüfenden Eigenschaft in der Stichprobe befinden, häufig geschrieben als $h(x|N;M;n)$ oder $H(N;M;n)(\{x\})$. Der Ergebnisraum S ist daher {0,1,...,n}.

Wenn man die Zufallsvariable "Zahl der Kugeln erster Sorte in der Stichprobe" als X bezeichnet, kann man die Wahrscheinlichkeit dafür angeben als

Definition 80:

$$h(x|N;M;n) = P(X = x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

Die Verteilungsfunktion H(x|N;M;n) gibt dann die Wahrscheinlichkeit an, dass **höchstens** x viele Kugeln erster Sorte in der Stichprobe sind.

Beispiel 80:

In einem Behälter befinden sich 45 Kugeln, davon sind 20 gelb. Es werden 10 Kugeln ohne Zurücklegen entnommen.

Die Hypergeometrische Verteilung gibt die Wahrscheinlichkeit h(x|45;20;10) an, dass genau x = 0, 1, 2, 3, ..., 10 der entnommenen Kugeln gelb sind.

Beim Zahlenlotto gibt es 49 nummerierte Kugeln; davon werden bei der Auslosung 6 gezogen; auf dem Lottoschein werden 6 Zahlen angekreuzt.

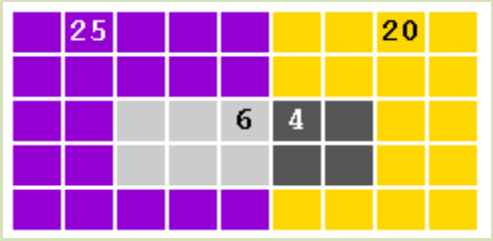
h(x|49;6;6) gibt die Wahrscheinlichkeit dafür an, genau x = 0, 1, 2, 3, ..., 6 "Treffer" zu erzielen.

Lösung:

$$h(x|N;M;n) = h(6|49;6;6) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{6}{6} \cdot \binom{49-6}{6-6}}{\binom{49}{6}}$$

Lösung: für Beispiel 1:

h(4|45,20,10)



Zu dem oben aufgeführten Beispiel der farbigen Kugeln soll die Wahrscheinlichkeit ermittelt werden, dass genau 4 gelbe Kugeln resultieren. Die Wahrscheinlichkeit ergibt sich aus:

Anzahl der Möglichkeiten, genau 4 gelbe (und damit genau 6 violette) Kugeln auszuwählen geteilt durch

Anzahl der Möglichkeiten, genau 10 Kugeln beliebiger Farbe auszuwählen

Es gibt $\binom{20}{4} = 4845$ Möglichkeiten, genau 4 gelbe Kugeln auszuwählen.

Es gibt $\binom{25}{6} = 177.100$ Möglichkeiten, genau 6 violette Kugeln auszuwählen.

Da jede "gelbe Möglichkeit" mit jeder "violetten Möglichkeit" kombiniert werden kann, ergeben sich

$4845 \cdot 177.100 = 858.049.500$

Möglichkeiten für genau 4 gelbe und 6 violette Kugeln. Wir erhalten also die Wahrscheinlichkeit

$$P(X = 4) = h(4|45,20,10) = \frac{\binom{20}{4} \cdot \binom{25}{6}}{\binom{45}{10}} = 0{,}2690$$

das heißt, in rund 27 Prozent der Fälle werden genau 4 gelbe (und 6 violette) Kugeln entnommen

Beispiel 81:

In einer Produktionsserie vom Umfang N=20 seien M=10 Produkteinheiten fehlerhaft. Wie groß ist die Wahrscheinlichkeit, in einer Zufallsstichprobe (Ziehungen ohne Zurücklegen) vom Umfang n=5 zwei fehlerhafte Erzeugnisse zu finden?

Lösung:

Da $p = \frac{M}{N}$, gilt auch

$$f(x/n;N;M) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{10}{2}\binom{20-10}{5-2}}{\binom{20}{5}} = \frac{45*120}{15504} \approx 0{,}348$$

Definition 81:

Erwartungswert

$$E(x) = \mu = n \cdot \frac{M}{N}$$

Varianz

$$V(x) = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \cdot \left(\frac{N-n}{N-1}\right)$$

Standardabweichung

$$S(x) = \sqrt{V(x)}$$

Beispiel 82:

Nehmen wir an, wir haben 20 Kugeln in einem Sack, von denen 8 blau sind und 12 rot. Nun mischen wir die Kugeln und ziehen 15 Kugeln. Wie groß ist die Wahrscheinlichkeit, genau 5 blaue und 10 rote Kugeln zu ziehen?

Lösung:

N = 20

n = 15

p = 0.4 (8 der 20 Kugeln sind blau)

k = 5

Die Wahrscheinlichkeit genau 5 blaue Kugeln zu ziehen ist 0.238

Beispiel 83:

Unter 10 Losen befinden sich 2 Gewinnlose. Bestimmen Sie die Wahrscheinlichkeit, dass sich unter fünf willkürlich ausgewählten Losen

a) genau ein Gewinnlos befindet,

b) beide Gewinnlose befinden,

c) höchstens ein Gewinnlos befindet.

Lösung:

a) $$p = \frac{\binom{2}{1} \cdot \binom{8}{4}}{\binom{10}{5}} = \frac{2 \cdot 70}{252} = \frac{5}{9} \approx 55.5\%$$

b) $$p = \frac{\binom{2}{2} \cdot \binom{8}{3}}{\binom{10}{5}} = \frac{1 \cdot 56}{252} = \frac{2}{9} \approx 22.2\%$$

c) "höchstens ein Gewinnlos" heisst: "0 oder 1 Gewinnlos"

oder (einfacher): "nicht 2 Gewinnlose": $p = 1 - \frac{2}{9} = \frac{7}{9} = 77.8\%$

Poisson-Verteilung

Siméon Denis Poisson (1781-1840) veröffentlichte 1837 diese Verteilung zusammen mit seiner Wahrscheinlichkeitstheorie in dem Werk „Recherchessur la probabilité des jugements en matièrescriminelles et en matièrecivile“. („Forschungsarbeiten zur Wahrscheinlichkeit von Urteilen im verbrecherischen Bereich und im Zivilbereich“).

Es handelt sich um eine diskrete Wahrscheinlichkeitsverteilung, die beim mehrmaligen Durchführen eines Bernoulli-Experiments entsteht. Letzteres ist ein Zufallsexperiment, das nur zwei mögliche Ergebnisse besitzt (z.B. „Erfolg“ und „Misserfolg“). Führt man ein solches Experiment sehr oft durch und ist die Erfolgswahrscheinlichkeit gering, so ist die Poisson-Verteilung eine gute Näherung für die entsprechende Wahrscheinlichkeitsverteilung. Die Poisson-Verteilung wird deshalb manchmal als die **Verteilung der seltenen Ereignisse** bezeichnet (siehe auch Gesetz der kleinen Zahlen).

Definition 82:

Die Poissonverteilung ist eine Wahrscheinlichkeitsverteilung für diskrete Ja-Nein-Verteilungen.

Es handelt sich um eine diskrete Wahrscheinlichkeitsverteilung, die beim mehrmaligen Durchführen eines Bernoulli-Experiments entsteht.

Letzteres ist ein Zufallsexperiment, das nur zwei mögliche Ergebnisse besitzt (z.B. „Erfolg“ und „Misserfolg“).

Führt man ein solches Experiment **sehr oft** durch und ist die **Erfolgswahrscheinlichkeit gering**, so ist die Poisson-Verteilung eine gute Näherung für die entsprechende Wahrscheinlichkeitsverteilung.

Die Poisson-Verteilung wird deshalb manchmal als die **Verteilung der seltenen Ereignisse** bezeichnet (siehe auch Gesetz der kleinen Zahlen).

Sie ersetzt die Binomialverteilung bei im Verhältnis zu den Beobachtungen sehr seltenen Ereignissen. Das bedeutet, dass sie ab einer Wahrscheinlichkeit p kleiner oder gleich 0,01 hinreichend genau ist.

Berechnung

Für große n (n >= 100) und kleine p (p <= 0,1) fand Poisson eine gute Näherung für die Binomialverteilung. Betrachten wir für eine Bernoulli-Kette die Wahrscheinlichkeit für das Auftreten eines Treffers:

$$\begin{aligned} P(X=1) &= \binom{n}{1} \cdot p \cdot (1-p)^{n-1} \qquad \mu = n \cdot p \text{ sei der (kleine) Erwartungswert, d.h. } p = \frac{\mu}{n}. \\ &= \binom{n}{1} \cdot \frac{\mu}{n} \cdot \left(1-\frac{\mu}{n}\right)^{n-1} \\ &= \mu \cdot \underbrace{\left(1-\frac{\mu}{n}\right)^{n}}_{\approx e^{-\mu}} \cdot \underbrace{\left(1-\frac{\mu}{n}\right)^{-1}}_{\approx 1} = \mu \cdot e^{-\mu} \end{aligned}$$

Für ein beliebiges k gilt:

$$\begin{aligned} P(X=k) &= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \\ &= \frac{n(n-1)\ldots(n-(k-1))}{k!} \cdot \frac{\mu^k}{n^k} \cdot \left(1-\frac{\mu}{n}\right)^{n-k} \\ &= \frac{n(n-1)\ldots(n-(k-1))}{n^k} \cdot \frac{\mu^k}{k!} \cdot \left(1-\frac{\mu}{n}\right)^{n} \cdot \left(1-\frac{\mu}{n}\right)^{-k} \\ &= \underbrace{1 \cdot \left(1-\frac{1}{n}\right) \ldots \left(1-\frac{k-1}{n}\right)}_{\approx 1} \cdot \ldots \cdot \frac{\mu^k}{k!} \cdot \underbrace{\left(1-\frac{\mu}{n}\right)^{n}}_{\approx e^{-\mu}} \cdot \underbrace{\left(1-\frac{\mu}{n}\right)^{-k}}_{\approx 1} \\ &= \frac{\mu^k}{k!} \cdot e^{-\mu} \end{aligned}$$

Definition 83:

Eine Zufallsgröße X heißt poissonverteilt, falls die Wahrscheinlichkeiten für k = 0, 1, 2, . . . Treffer mit

$P(X=k) = \frac{\mu^k}{k!} \cdot e^{-\mu}$

berechnet werden.

Definition 84:

Erwartungswert:

$E(x) = \mu$

Varianz

$V(x) = \mu$

Standardabweichung:

$S(x) = \sqrt{V(x)}$

Bemerkung 46:

Die Varianz ist also genauso groß wie der Erwartungswert.

Beispiel 84:

Von 100 Personen ist durchschnittlich eine Person farbenblind.

Mit welcher Wahrscheinlichkeit befinden sich unter 100 zufällig ausgewählten Personen mindestens zwei farbenblinde Personen?

Berechnung mit der Binomialverteilung:

$$P(x \geq 2) = 1 - P(x \leq 1) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \binom{100}{0} 0{,}01^0 \cdot 0{,}99^{100} - \binom{100}{1} 0{,}01^1 \cdot 0{,}99^{99}$$

$$= 1 - 0{,}3660 - 0{,}3697 = 0{,}2643$$

Rechnen wir dieses jetzt mit nur 50 Personen durch.

$$P(x \geq 2) = 1 - P(x \leq 1) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \binom{50}{0} 0{,}01^0 \cdot 0{,}99^{50} - \binom{50}{1} 0{,}01^1 \cdot 0{,}99^{49}$$

$$= 1 - 0{,}6050 - 0{,}3056 = 0{,}0894$$

Rechnen wir dieses jetzt mit nur 200 Personen durch.

$$P(x \geq 2) = 1 - P(x \leq 1) = 1 - P(X = 0) - P(X = 1)$$

$$= 1 - \binom{200}{0} 0{,}01^0 \cdot 0{,}99^{200} - \binom{200}{1} 0{,}01^1 \cdot 0{,}99^{199}$$

$$= 1 - 0{,}1340 - 0{,}2707 = 0{,}5953$$

Beispiel 85:

Durchschnittlich ist eine Person farbenblind. Wie groß ist die Wahrscheinlichkeit mindestens zwei farbenblinde Personen zu erhalten?

Berechnung mit der Poisson-Verteilung:

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{1^0}{0!} \cdot e^{-1} - \frac{1^1}{1!} \cdot e^{-1}$$

$$P(X \geq 2) = 1 - 0{,}3679 \; - 0{,}3679 = 0{,}2642$$

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{0,5^0}{0!} \cdot e^{-0{,}5} - \frac{0{,}5^1}{1!} \cdot e^{-0{,}5}$$

$$P(X \geq 2) = 1 - 0{,}6065 \; - 0{,}3033 = 0{,}0902$$

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{2^0}{0!} \cdot e^{-2} - \frac{2^1}{1!} \cdot e^{-2}$$

$$P(X \geq 2) = 1 - 0{,}1353 - 0{,}2707 = 0{,}5940$$

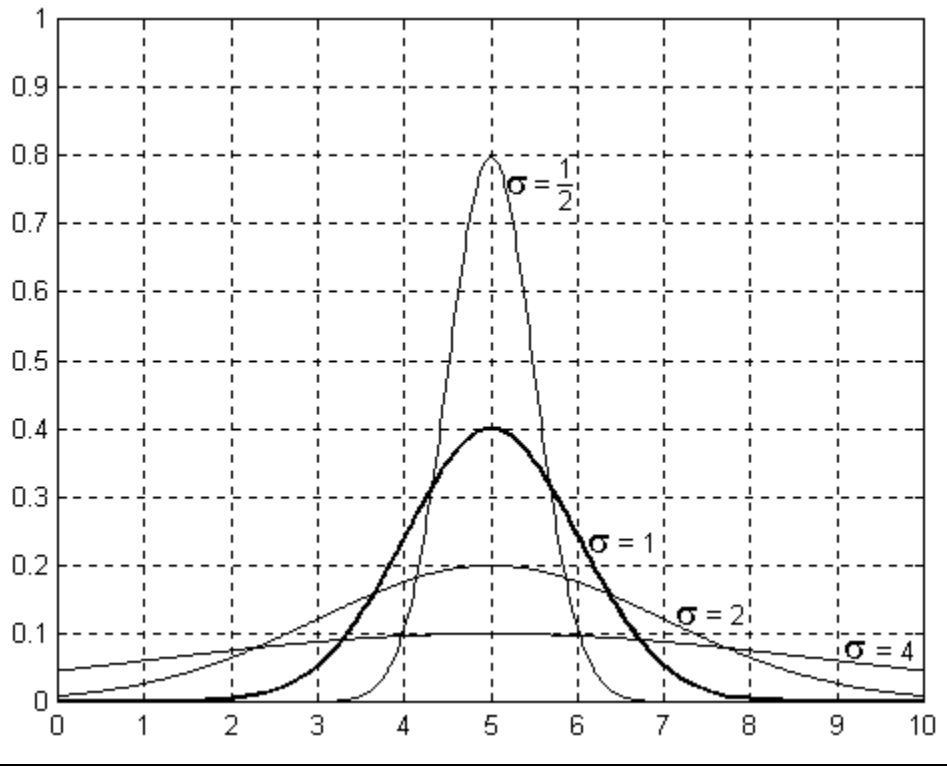
Normalverteilung

Die Normalverteilung ist eine der wichtigsten theoretischen Verteilungen der analytischen Statistik.

Es handelt sich um eine stetige, symmetrische, eingipflige Verteilung, die sich asymptotisch der x-Achse nähert. Ihre Bedeutung leitet sich über den zentralen Grenzwertsatz daraus her, dass sie für viele andere Wahrscheinlichkeitsverteilungen, und insbesondere für Stichprobenverteilungen, eine Grenzverteilung darstellt, der sich diese Verteilungen asymptotisch nähern.

Sie hat die Dichtefunktion

Die **Dichtefunktion** nach Gauss ist auf die Fläche A=1 normiert. $-\infty < x < \infty$ Mittelwert μ Standardabweichung σ	$$f(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot \exp\left(-\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma} \right)^2 \right)$$



Die Verteilung ist durch ihre Parameter μ (Mittelwert) und σ (Standardabweichung) genau bestimmt. Arithmetisches Mittel, Median und häufigster Wert (Modus) fallen zusammen.

Ihre Wendepunkte sind durch $\mu \pm \sigma$ gegeben.

Definition 85:

Die Verteilungsfunktion einer normal verteilten Zufallsvariablen x hat die Form

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

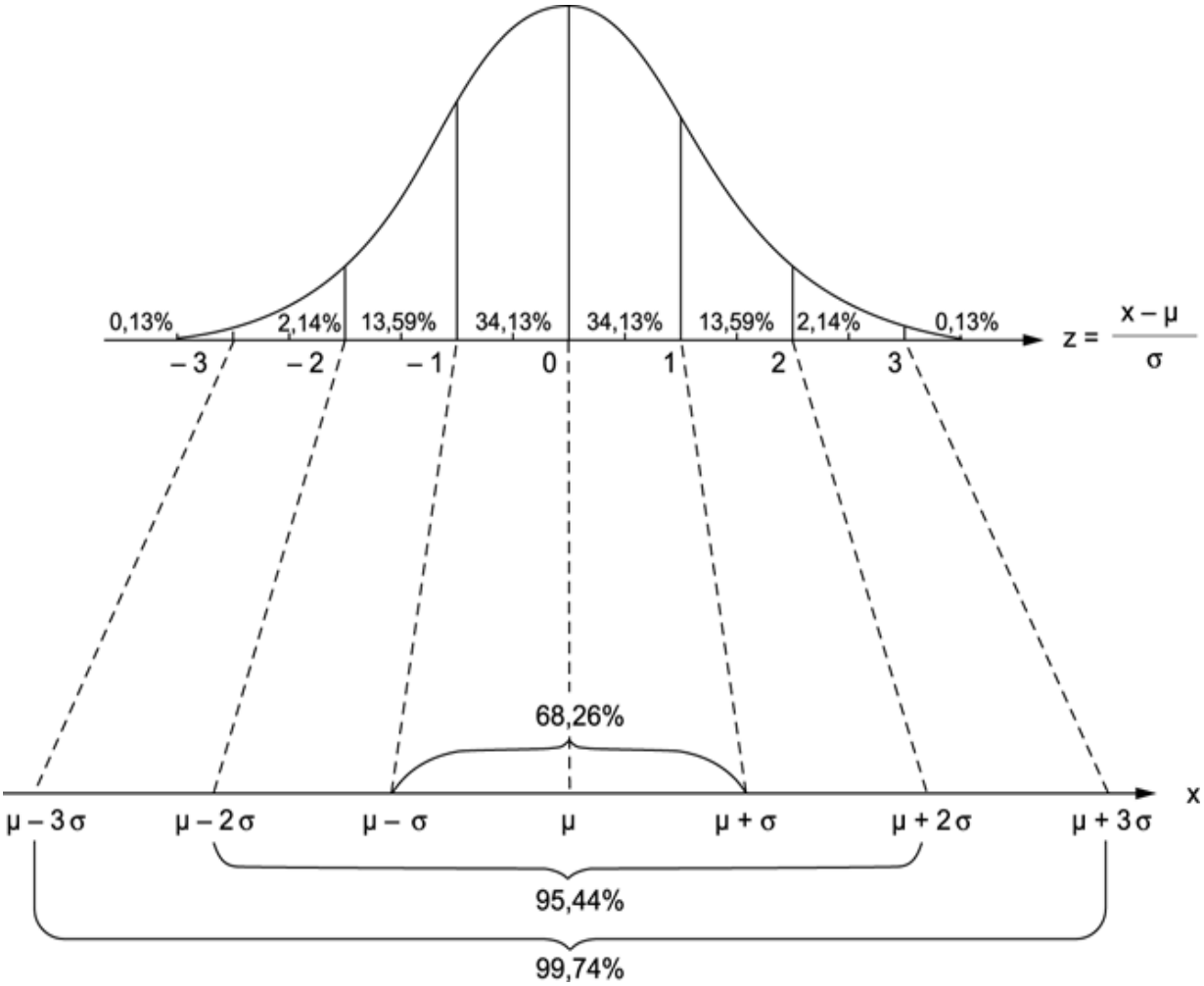
Um die Wahrscheinlichkeit für ein bestimmtes Intervall zu berechnen, kann man jede Normalverteilung so transformieren, dass sie das betreffende μ und σ hat.

Definition 86:

Dafür bietet sich die Standardnormalverteilung an, d.h. die Verteilung, bei der $\mu = 0$ und $\sigma = 1$ ist. Die Standardisierung erfolgt mit Hilfe der sog. z-Transformation

$$z = \frac{x - \bar{x}}{s}$$

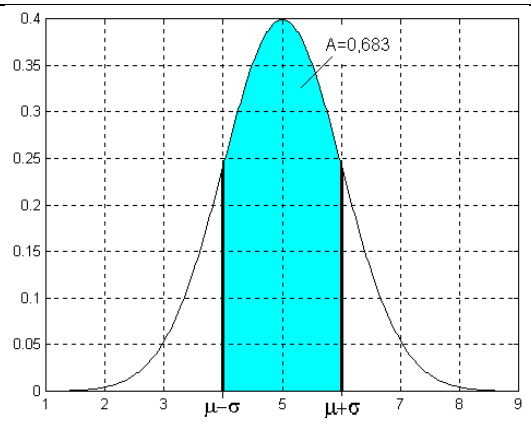
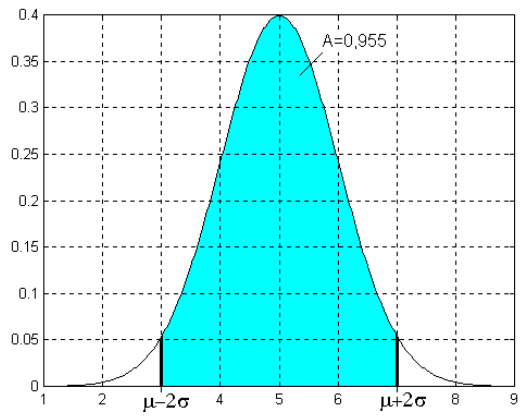
Durch Rücktransformation lassen sich dann die Werte für jede beliebige Normalverteilung berechnen. Die grafische Darstellung der Normalverteilung ist die Gauß'sche Glockenkurve.

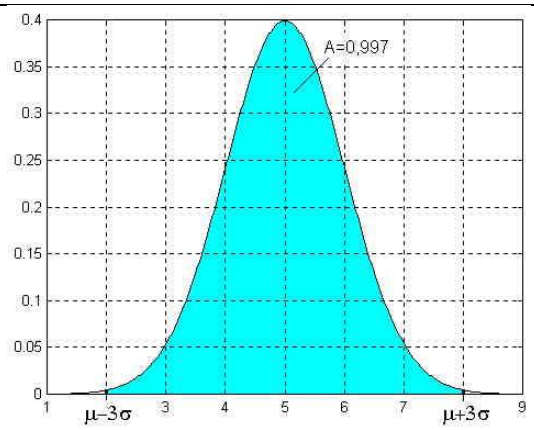


Bemerkung 47:

- In der Normalverteilung ist die als der Abstand zwischen den Wendepunkten definierte Standardabweichung σ die Größe, der eine Normalkurve ihre konkrete Gestalt verdankt.
- Das Grundprinzip der Berechnung von Vertrauensbereichen bzw. Fehlerspannen besteht darin, dass die beiden durch die Wendepunkte der Normalkurve abgegrenzten Flächen mit 68,26 % ungefähr zwei Drittel der Gesamtfläche unterhalb der Normalkurve ausmachen.
- Für die Stichprobentheorie bedeutet das, dass rund zwei Drittel aller Stichproben, die sich aus einer Grundgesamtheit bilden lassen, einen Anteil von Merkmalen haben, der innerhalb des durch die doppelte Standardabweichung gegebenen Vertrauensbereichs liegt und mithin nicht mehr als $\pm 1 \cdot \sigma$ vom wahren Wert der Grundgesamtheit entfernt liegen kann.
- Anders ausgedrückt: Die Wahrscheinlichkeit, dass ein konkreter Wert einer konkreten Stichprobe einen maximal um $\pm 1 \cdot \sigma$ vom wahren Wert abweichenden Wert liefert, ist gleich 0,6826.

Den Zusammenhang zwischen der Standardabweichung σ und der Fläche A geht aus der folgenden Tabelle hervor.

Dichtefunktion $\sigma = 1$ $\mu = 5$	
Dichtefunktion $\sigma = 1$ $\mu = 5$	

Dichtefunktion $\sigma = 1$ $\mu = 5$	

Die Grafik zeigt, dass die Fläche unterhalb der Normalkurve auf 95,45 bzw. 99,73 % anwächst, wenn der **Vertrauensbereich** auf zwei bzw. 3 ausgedehnt wird.

Die Unterschiede zwischen den verschiedenen Fehleraussagen werden meist dadurch charakterisiert, dass man die Wahrscheinlichkeitsaussagen auf Stichprobenbasis mit einem **Signifikanzniveau** von $1 \cdot \sigma$ (einem Sicherheitsfaktor von z = 1) oder einem Sicherheitsgrad von 68,26 % bzw. einer **Irrtumswahrscheinlichkeit** von 31,74 % trifft.

Die folgende Übersicht zeigt die Zusammenhänge mit auf eine Stelle nach dem Komma gerundeten Zahlen:

Signifikanzniveau		
Sicherheits-faktor	Sicherheits-grad	Irrtums-wahrschein-lichkeit
z = 1,00	68,3%	31,7%
z = 1,64	90,0%	10,0%
z = 1,96	95,0%	5,0%
z = 2,00	95,5%	4,5%
z = 2,58	99,0%	1,0%
z = 3,00	99,7%	0,3%
z = 3,29	99,9%	0,1%

Das Prinzip der Normalverteilung wurde im Ansatz von dem Franzosen Abraham de Moivre, einem nach England emigrierten Hugenotten, 1756 erkannt und von Carl Friedrich Gauß zur mathematischen Vollendung geführt.

Durch lineare Transformation lassen sich hieraus auch beliebige normalverteilte Zufallszahlen erstellen.

Mittelwert und Standardabweichung für eine normalverteilte Messreihe

Als Messwert wird das arithmetische Mittel der aus n Einzelmessungen gebildet. Er gilt als bester Schätzwert für den unbekannten "Wahren" Wert.

Definition 87:

Erwartungswert oder Mittelwert

$$\bar{x} = \frac{1}{n} \cdot (x_1 + x_2 + ... + x_n)$$

Der Begriff "Standardabweichung" wird bei Messreihen durch den Begriff "Streuung" ersetzt und berechnet sich wie folgt :

Definition 88:

Streuung oder Standardabweichung

$$S = \sqrt{\frac{1}{n-1} \cdot \sum_{K=1}^{n} (x_K - \bar{x})^2}$$

Beispiel 86:

In Mathematanien wurde die Körpergröße aller Studenten gemessen. Es stellte sich heraus, dass die Größe normalverteilt ist, mit dem Erwartungswert μ = 175 cm und der Standardabweichung σ = 7,5 cm.

Wie groß ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Student

a) kleiner als 160 cm (2,28%)

b) größer als 180 cm (25,14%)

c) zwischen 170 und 182 cm groß ist? (57,24%)

Lösung:

a) 0,0228

b) 0,2514

c) 0,5724

Beispiel 87:

Die Abgabemenge X (in cm^3) eines Getränkeautomaten sei normalverteilt mit dem Erwartungswert μ = 250 cm^3 und der Standardabweichung σ = 2 cm^3. Wie groß ist die Wahrscheinlichkeit, dass die Abgabemenge

(a) mehr als 253 cm^3 (6,68%)

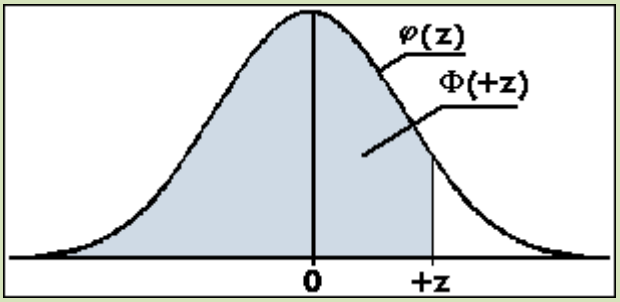
(b) mindestens 249 cm^3 bis höchstens 251 cm^3) beträgt? (38,29%)

Lösung:

Mittelwert=250

Standardabweichung=2

(a)

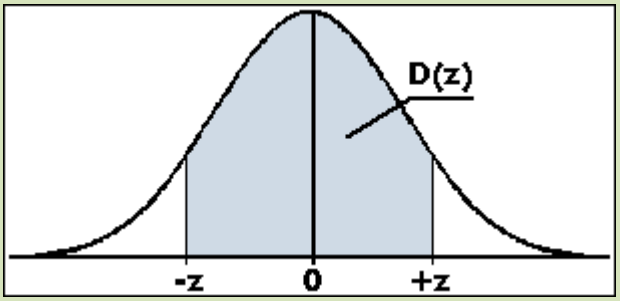


Transformation(+z): $\frac{253-250}{2}=\frac{3}{2}=1{,}5 \Rightarrow z=1{,}5$

Aus Tabelle F(+z): 0,9332

Mehr als 253: 1-0,9332=0,0668

(b)



Transformation (+z): $\frac{249-250}{2}=-0{,}5 \Rightarrow z=0{,}5$

Transformation (-z): $\frac{251-250}{2}=0{,}5 \Rightarrow z=0{,}5$

Aus Tabelle F(-z)=f(z)=0,3829

Beispiel 88:

Eine Reifenfirma untersucht die Lebensdauer eines neu entwickelten Reifens. Dabei zeigt sich, dass die ermittelte Lebensdauer der Reifen gut durch eine Normalverteilung mit den Parametern $\mu = 36.000\text{km}$ und $\sigma = 4.000\text{km}$ angenähert werden kann.

a) Welche Lebensdauer wird von 95 % der Reifen nicht überschritten? (42.600)

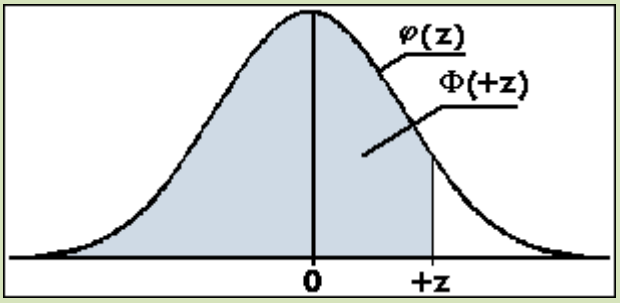
b) Wie groß ist die Wahrscheinlichkeit dafür, dass ein zufällig ausgewählter Reifen mehr als 28000 km hält? (97,72%)

c) Berechnen Sie das kürzeste symmetrische Schwankungsintervall, in das 95 % der Reifen fallen. (±7.840)

d) Die Firma ist in der Lage, den Herstellungsprozess der Reifen so zu steuern, dass $\mu = 36.000\text{km}$ konstant bleibt, aber die Standardabweichung σ veränderbar ist. Die Firma will den Abnehmern eine Lebensdauer von mindestens 30000 km garantieren; Reifen von geringerer Lebensdauer will sie kostenlos umtauschen. Die Firma hat sich ausgerechnet, dass es für sie tragbar ist, wenn im Durchschnitt 2.28% der Reifen diese, Mindestlebensdauer` unterschreiten. Mit welcher Standardabweichung σ muss der Produktionsprozess ablaufen, damit nicht höhere Umtauschforderungen an die Firma herangetragen werden?

Lösung:

a)Mittelwert=36.000 Standardabweichung=4.000

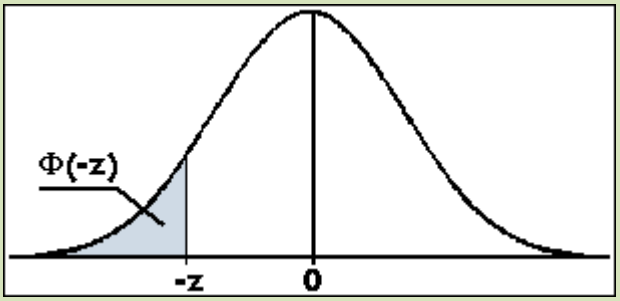


Gesuchtes Intervall: 0,9500

Dieses aus der Tabelle ergibt 1,65

$$z = \frac{x - \mu}{\sigma} \Leftrightarrow x = z \cdot \sigma + \mu = 1{,}65 \cdot 4.000 + 36.000 = 42.600$$

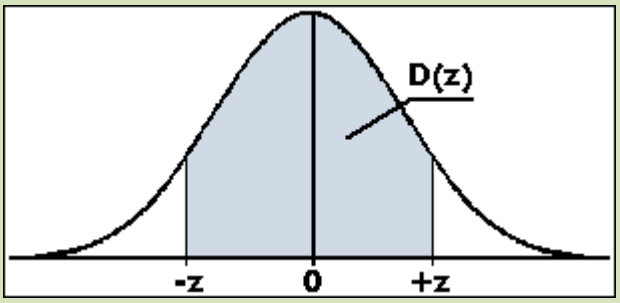
b)



Transformation: $z = \frac{28.000 - 36.000}{4.000} = 2 \quad \Rightarrow 0{,}0228$

Gegenereignis: 1-0,0228=0,9772

c)

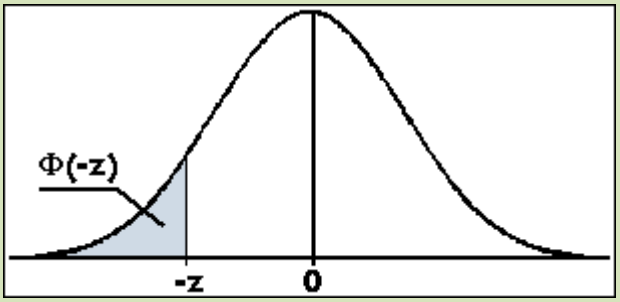


Gesuchtes Intervall: 0,9500

Dieses aus der Tabelle ergibt 1,96

$$z = \frac{x - \bar{x}}{\sigma} \Leftrightarrow x = z \cdot \sigma \pm \bar{x} = 1{,}96 \cdot 4.000 + 36.000 = 7.840$$

d)



Suchen von 0,0228 in der Tabelle: 2

$$2 = \frac{30.000 - 36.000}{\sigma} \Leftrightarrow \sigma = \frac{30.000 - 36.000}{2} = 3.000$$

Beispiel 89:

Wie groß muss ein Student in Mathematanien sein (s. Bsp. 1), damit er

a) zu den 20% kleinsten (<169)

b) zu den 10% größten Studenten gehört? (>184)

c) In welchem symmetrischen Bereich [μ-ε, μ+ε] liegen die Größen von 95% aller Studenten? (160-190)

Lösung:

a) < 169 cm

b) > 1845 cm

c) [160 cm, 190 cm]

Zufallsvariablen

Eine **Zufallsvariable** oder **Zufallsgröße** ist ein Begriff aus dem mathematischen Teilgebiet Stochastik. Man bezeichnet damit eine Funktion, die den Ergebnissen eines Zufallsexperiments Werte zuordnet. Diese Werte werden als **Realisationen** der Zufallsvariable bezeichnet.

Betrachtet man ein Zufallsexperiment Münzwurf, so kann man beispielsweise eine Zufallsvariable *X* definieren, indem man dem Ergebnis Ω = „die Münze zeigt Kopf“ den Wert 0 und dem Ergebnis Ω = „die Münze zeigt Zahl“ den Wert 1 als Realisation zuordnet.

$$X(\omega) = \begin{cases} 0 & \text{die Münze zeigt Kopf} \\ 1 & \text{die Münze zeigt Zahl} \end{cases}$$

Die Zufallsvariable selbst wird üblicherweise mit einem Großbuchstaben bezeichnet (hier *X*), während man für die Realisationen die entsprechenden Kleinbuchstaben verwendet (hier beispielsweise $x = 1$).

Während früher der Begriff Zufallsgröße (manchmal auch Zufallsveränderliche) der übliche deutsche Begriff war, hat sich heute (ausgehend vom englischen *random variable*) der etwas irreführende Begriff Zufallsvariable durchgesetzt.

Zufallsvariablen sind jedoch Funktionen und dürfen nicht mit den Variablen verwechselt werden, die üblicherweise in der Mathematik eingesetzt werden.

In sehr vielen Fällen kann der Ausgang eines Zufallsexperimentes durch einen Zahlenwert gekennzeichnet werden. Beispielsweise können wir beim Würfeln direkt die gewürfelte Augenzahl als Kennzeichen für den Ausgang des Experimentes verwenden. Würfelt man mit einem Würfel so oft, bis man zum ersten Mal die Sechs erhält, so kann die Zahl der erforderlichen Würfe als Kennzeichen für den Ausgang des Experimentes herangezogen werden. Bei solchen Zufallsexperimenten erscheint es daher umständlich, den gesamten Ereignisraum und die einzelnen Ereignisse aufzuzählen, um das Experiment und dessen konkreten Ausgang im Einzelfall zu kennzeichnen. Vereinfacht könnte man den Ereignisraum einfach durch einen bestimmten Ausschnitt aus einer Zahlenskala und den konkreten Ausgang durch einen bestimmten Wert aus diesem Wertebereich beschreiben. Wir kommen auf diese Weise zum Begriff der *Zufallsvariable*, *stochastischen Variable* oder *zufälligen Variable*.

Definition 89:

Zufallsvariable heißt eine Abbildung (Funktion) *X*, die den Ergebnisraum Ω eines Zufallsexperimentes auf der Ereignisalgebra *A* in eine Teilmenge *X* der reellen Zahlen abbildet.

Beispiel 90:

Die fränkische Druckerei Prinzing nennt 10 multifunktionelle Hochleistungsdrucker ihr Eigen. Drei Drucker sind von der Firma Alpha, zwei sind von Beta, vier von Gamma und einer stammt von der Firma Delta.

Da die Drucker auch von Kunden bedient werden, fallen sie aufgrund unsachgemäßer Handhabung häufig aus. Man hat festgestellt, dass alle Drucker in gleichem Maße anfällig sind. Wegen der Gewährleistung wird bei jedem Ausfall ein Wartungstechniker der betreffenden Firma geholt. Die Kosten für die Wiederherstellung eines Druckers hängen vom Hersteller ab, wobei die Drucker der Firma Gamma in der Reparatur am billigsten sind.

Am liebsten ist es natürlich Herrn Prinzing, wenn ein Drucker mit den geringsten Reparaturkosten ausfällt.

Überlegen wir:

Welche Ergebnismenge gehört zu dem Zufallsvorgang:

Ein Drucker fällt zufällig aus?

Mit welcher Wahrscheinlichkeit entstehen Herrn Prinzing die geringsten Kosten?

Wir erhalten die Ergebnismenge

$\Omega = \{A_1, A_2, A_3, B_1, B_2, G_1, G_2, G_3, G_4, D_1\}$,

wobei z.B. B_2 Drucker Nr. 2 der Firma Beta bedeutet. G sei das Ereignis, die geringsten Reparaturkosten zu haben. Jeder Drucker hat die gleiche Wahrscheinlichkeit, auszufallen. Dann ist nach dem Symmetrieprinzip

$$P(G) = \frac{\text{Zahl der G-Drucker}}{\text{Zahl aller Drucker}} = \frac{|G|}{|\Omega|} = \frac{4}{10} = 0,4$$

Die Kosten für die Reparatur eines Druckers betragen je nach Hersteller wie folgt:

Hersteller	Alpha	Beta	Gamma	Delta
Kosten (Euro)	50	60	30	100

Überlegen wir: Wie viel muss Herr Prinzing pro Ausfall im Durchschnitt bezahlen?

Ordnen wir nun der Ergebnismenge die entsprechenden Kosten zu:

A_1	A_2	A_3	B_1	B_2	G_1	G_2	G_3	G_4	D_1
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
50	50	50	60	60	30	30	30	30	100

Ω hat 10 Ergebnisse und jedes Elementarereignis hat die Wahrscheinlichkeit 1/10. Jeder Drucker fällt dann auch mit der Wahrscheinlichkeit 1/10 aus. Die durchschnittlichen Reparaturkosten sind also

$$50 \cdot \frac{1}{10} + 50 \cdot \frac{1}{10} + 50 \cdot \frac{1}{10} + 60 \cdot \frac{1}{10} + 60 \cdot \frac{1}{10} + \ldots + 100 \cdot \frac{1}{10}$$

$$= 50 \cdot \frac{3}{10} + 60 \cdot \frac{2}{10} + 30 \cdot \frac{4}{10} + 100 \cdot \frac{1}{10}$$

$$= \frac{150}{10} + \frac{120}{10} + \frac{120}{10} + \frac{100}{10} = \frac{490}{10} = 49 Euro$$

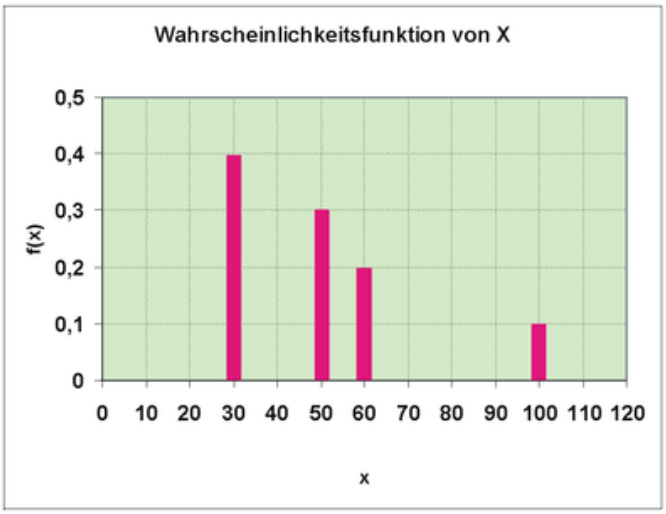
Wir haben soeben eine Zufallsvariable konstruiert und zwar, indem wir allen Ergebnissen von Ω eine Zahl zugeordnet haben.

Bemerkung 48:

- Den Durchschnitt konnten wir erst berechnen, nachdem wir die Drucker mit einer Zahl versehen hatten. Man kann je nach Interesse den Elementarereignissen beliebige Zahlen zuordnen. So könnten für die laufende Wartung wieder ganz andere Kosten gelten. Nur die Ergebnismenge ist festgelegt. Man könnte nun die Wahrscheinlichkeit berechnen, dass bei einem Ausfall 60 Euro fällig werden: Es gibt 10 Elementarereignisse und zwei davon entsprechen 60 Euro. Also beträgt diese Wahrscheinlichkeit 2/10.
- Wir bezeichnen eine Zufallsvariable mit einem großen Buchstaben. Die Werte, die eine Zufallsvariable annehmen kann, nennt man Ausprägung. Eine bestimmte Ausprägung kennzeichnen wir mit einem Kleinbuchstaben. Nennen wir unsere Zufallsvariable "Reparaturkosten" X. Wir fassen jetzt die verschiedenen Wahrscheinlichkeiten der Zufallsvariablen X in einer Wahrscheinlichkeitstabelle zusammen. Herr Prinzing hat 4 mal die "Chance", 30 Euro zu bezahlen, also ist die Wahrscheinlichkeit, dass X = 30 ist, gleich 4/10, usw.

Wahrscheinlichkeitstabelle:

	x_1	x_2	x_3	x_4
Ausprägung x_i	30	50	60	100
Wahrscheinlichkeit $f(x_i)$	0,4	0,3	0,2	0,1



Wahrscheinlichkeitsfunktion von X: Reparaturkosten

f(x) bezeichnet die zur bestimmten Ausprägung x gehörende Wahrscheinlichkeit. Es ist beispielsweise

$P(X = 60) = f(x_3) = f(60) = 0{,}2$,

aber

$P(X = 70) = f(70) = 0$,

denn für X = 70 existiert kein Ergebnis.

Definition 90:

Die Summe aller Wahrscheinlichkeiten ist

$$\sum_{i=1}^{m} f(x) = 1$$

Bemerkung 49:

- Man kann diese Wahrscheinlichkeiten auch grafisch als Stabdiagramm darstellen.
- Man sieht, dass an den x-Stellen 30, 50, 60 und 100 die Wahrscheinlichkeitsfunktion die Werte 0,4, 0,3, 0,2 und 0,1 annimmt, aber an allen sonstigen Werten von x Null ist.

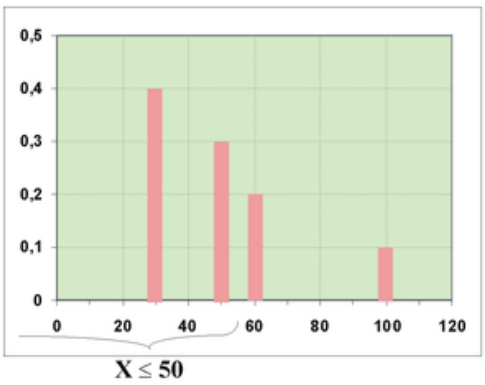
Wie groß ist nun aber die Wahrscheinlichkeit, dass Herr Prinzing höchstens 50 Euro bezahlen muss?

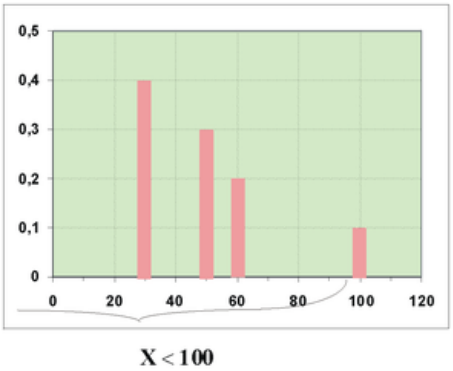
$P(X \leq 50) = P(X = 30) + P(X = 50) = 0{,}4 + 0{,}3 = 0{,}7$.

Das kann man auch aus der Graphik ersehen: Es ist die Summe der "Stäbchen" für $x \leq 50$.

Mit welcher Wahrscheinlichkeit muss Herr Prinzing weniger als 100 Euro zahlen? Gefragt ist hier nach P(X < 100). Ein Blick auf die Grafik verrät uns, dass gilt

$P(X < 100) = P(X \leq 60) = P(X = 30) + P(X = 50) + P(X = 60) = 0{,}4 + 0{,}3 + 0{,}2 = 0{,}9$.





Wie viel ist nun $P(30 < X \leq 60)$?

Man kann hier wieder die "Stäbchenmethode" anwenden:

$P(30 < X \leq 60) = 0{,}3 + 0{,}2 = 0{,}5$.

Definition 91:

Es gibt aber auch eine Rechenregel, die man mit Hilfe der Grafik leicht erkennt:

$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$,

also

$P(30 < X \leq 60) = P(X \leq 60) - P(X \leq 30) = 0{,}9 - 0{,}4 = 0{,}5$.

Definition 92:

Die Wahrscheinlichkeiten $P(X \leq a)$ einer bestimmten Ausprägung a von X bilden die Verteilungsfunktion von X, die die Wahrscheinlichkeitsverteilung von X in eindeutiger Weise beschreibt. Das ist eine Festlegung, die die Statistiker als sinnvoll erachten.

Bemerkung 50:

- Die Verteilungsfunktionen werden mit Großbuchstaben als F(a) bezeichnet.

Meist wird statt a das Symbol x verwendet. Wir wollen die Verteilungsfunktion konstruieren, indem wir die obige Graphik zu Hilfe nehmen und für einzelne Stützwerte x die Verteilungsfunktion berechnen.

Wie groß ist z.B. $P(X \leq 10)$? Es ist $P(X \leq 10) = F(10) = 0$.

Ebenso sind $P(X \leq 15) = 0$ und $P(X \leq 20) = 0$.

Es ist also $F(a) = 0$ für alle Werte von a mit $-\infty < a < 30$.

Als nächstes untersuchen wir $P(X \leq 30)$:

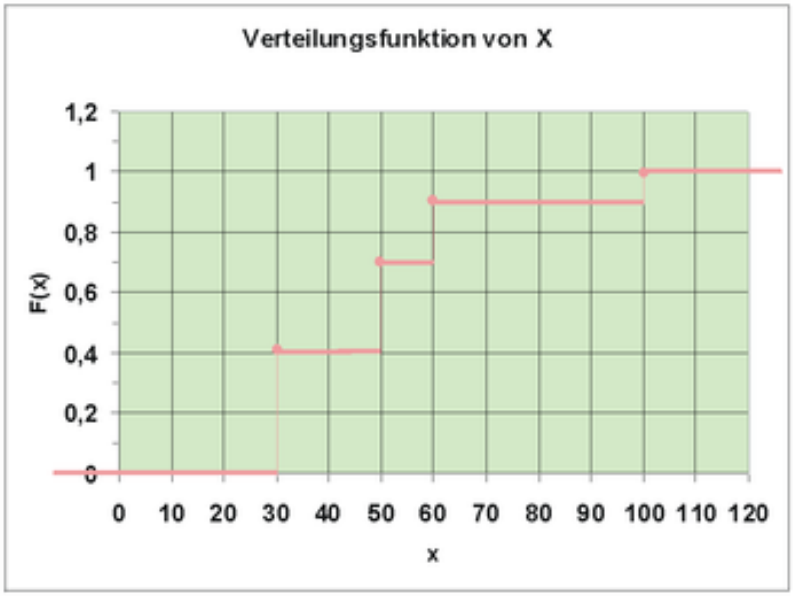
$P(X \leq 30) = F(30) = 0{,}4$. Ebenso sind $P(X \leq 30{,}1) = 0{,}4$ und $P(X \leq 49{,}99999) = 0{,}4$.

Die Verteilungsfunktion hat also den Wert $F(a) = 0{,}4$ für $30 \leq a < 50$.

Es gilt weiter: $P(X \leq 50)$, $P(X \leq 59)$, ... $P(X< 60)$ sind, siehe Graphik: 0,4 + 0,3 = 0,7.

Schließlich ist die Wahrscheinlichkeit $P(X \leq 100)$ oder auch $P(X \leq 110)$, $P(X \leq 1000)$ usw. ... gleich 1.

Wir können die Wahrscheinlichkeiten zusammenfassen in der Verteilungsfunktion



Bemerkung 51:

Verteilungsfunktion von X: Reparaturkosten

$$P(X \leq a) = F(a) = \begin{cases} 0 & \text{für } a < 30 \\ 0,4 & \text{für } 30 \leq a < 50 \\ 0,7 & \text{für } 50 \leq a < 60 \\ 0,9 & \text{für } 60 \leq a < 100 \\ 1 & \text{für } a \geq 100 \end{cases}$$

Bemerkung 52:

- Man sieht, dass diese Verteilungsfunktion grafisch eine Treppenfunktion darstellt. Die Punkte links an den Stufen zeigen an, dass der Funktionswert dieser Stufe genau zum Punkt a gehört.
- Man kann hier auch die Wahrscheinlichkeiten der Grafik entnehmen, z.B. ist P(X ≤ 70) = 0,9.

Besonders interessiert man sich bei einer Zufallsvariable für zwei Kennwerte, Parameter genannt, die die Zufallsvariable genauer beschreiben.

- Einer ist der durchschnittliche Wert, den die Zufallsvariable „auf lange Sicht" annimmt, wenn der Zufallsvorgang „sehr oft" durchgeführt wird. Dieser Parameter wird Erwartungswert E(X) genannt, also der Wert, den man langfristig erwarten kann. Wir hatten ihn schon oben ermittelt als

$$EX = 50 \cdot \frac{3}{10} + 60 \cdot \frac{2}{10} + 30 \cdot \frac{4}{10} + 100 \cdot \frac{1}{10} = 49$$

die durchschnittlichen Reparaturkosten.

Ein weiterer Parameter ist die Streuung der X, ein Maß, wie stark die einzelnen Werte von X von E(X) abweichen, also 30-49, 50-49, 60-49, 100-49. Da z.B. 100 viel seltener auftritt als 30, gewichtet man auch diese Abweichungen mit ihrer Wahrscheinlichkeit.

- Eine Quadrierung sorgt dann einerseits dafür, dass sich positive und negative Abweichungen nicht aufheben, andererseits für eine überproportionale Berücksichtigung von besonders starken Abweichungen. Man erhält im Ergebnis als durchschnittliche quadratische Abweichung der X-Werte von E(X) die Varianz

$$var\,X = (30-49)^2 \cdot 0,4 + (50-49)^2 \cdot 0,3 + (60-49)^2 \cdot 0,2 + (100-49)^2 \cdot 0,1$$
$$= 361 \cdot 0,4 + 1 \cdot 0,3 + 121 \cdot 0,2 + 2601 \cdot 0,1 = 429$$

wobei zu beachten ist, dass sich hier als Einheit Euro^2 ergibt.

- Die Wurzel der Varianz ist die Standardabweichung; man könnte sie salopp als mittlere Abweichung der Ausprägungen vom Durchschnitt bezeichnen. Sie beträgt in unserem Beispiel etwa 20,71.

Diskrete Zufallsvariablen

Definition 93:

Eine Zufallsvariable ist diskret, wenn sie in jedem beschränkten Intervall nur endlich viele Ausprägungen annehmen kann. Die diskrete Zufallsvariable kann endlich oder abzählbar unendlich viele Werte x_i (i = 1,2,..., m bzw. i = 1,2,...) annehmen.

Beispiel 91:

Zahl der Schadensleistungen, die in einem Jahr bei einer Versicherung auftreten

Kinderzahl von Konsumenten

Zahl der defekten Kondensatoren in einem Fertigungslos

Definition 94:

Ihre Wahrscheinlichkeitsfunktion ist

$$P(X = x) = f(x) = \begin{cases} f(x_i) & \text{für } x = x_i \\ 0 & \text{sonst} \end{cases}$$

Es gilt

$$\sum_{i=1}^{m} f(x_i) = 1 \, .$$

Definition 95:

Die Verteilungsfunktion $P(X \leq a) = F(a)$ ist die Summe aller Wahrscheinlichkeiten $f(x_i)$ für $x_i \leq a$.

Beispiel 92:

Zwei Würfel (ein blauer und ein grüner) werden 400-mal zusammen geworfen. Die Häufigkeiten für die einzelnen Ergebnisse werden in einer Tabelle aufgelistet. Jedem der Zahlenpaare (1 | 1) ... (6 | 6) kann deren Augensumme zugeordnet werden.

Die relativen Häufigkeiten der Augensummen sollen mit der Wahrscheinlichkeit ihres Auftretens verglichen werden. Dieser Sachverhalt soll in einer Tabelle dargestellt werden.

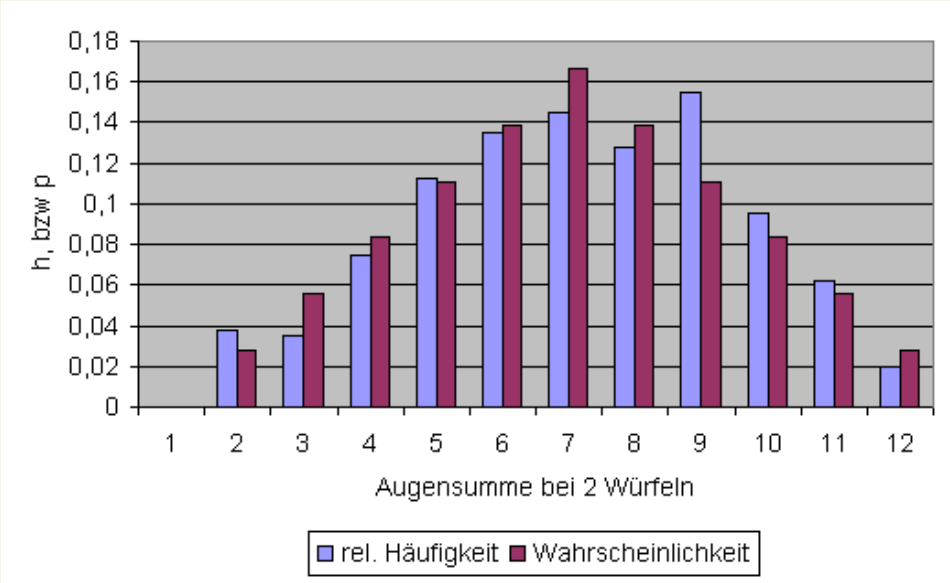
Ergebnisse des Würfelns:

grüner Würfel \ blauer Würfel	⚀	⚁	⚂	⚃	⚄	⚅
⚀	15	9	12	13	12	11
⚁	5	8	11	11	9	15
⚂	10	9	14	10	10	17
⚃	12	7	9	7	17	15
⚄	10	14	7	13	9	14
⚅	5	12	15	14	11	8

Tabelle:

Augensumme	zugehöriges Ergebnis	abs. H	rel. h	P(X)
2	(1|1)	15	0,0375	$\frac{1}{36} \approx 0{,}028$
3	(1|2);(2|1)	14	0,035	$\frac{2}{36} \approx 0{,}056$
4	(1|3);(2|2);(3|1)	30	0,075	$\frac{3}{36} \approx 0{,}083$
5	(1|4);(2|3);(3|2);(4|1)	45	0,1125	$\frac{4}{36} \approx 0{,}111$
6	(1|5);(2|4);(3|3);(4|2);(5|1)	54	0,135	$\frac{5}{36} \approx 0{,}139$
7	(1|6);(2|5);(3|4);(4|3);(5|2);(6|1)	58	0,145	$\frac{6}{36} \approx 0{,}167$
8	(2|6);(3|5);(4|4);(5|3);(6|2)	51	0,1275	$\frac{5}{36} \approx 0{,}139$
9	(3|6);(4|5);(5|4);(6|3)	62	0,155	$\frac{4}{36} \approx 0{,}111$
10	(4|6);(5|5);(6|4)	38	0,095	$\frac{3}{36} \approx 0{,}083$
11	(5|6);(6|5)	25	0,0625	$\frac{2}{36} \approx 0{,}056$
12	(6|6)	8	0,02	$\frac{1}{36} \approx 0{,}028$

Säulendiagramm:



Zufallsvariable:

Werden zwei Würfel gleichzeitig geworfen, so ist die Ergebnismenge:

$E = \{(1|1);(1|2);(1|3);....;(6|6)\}$

Ordnet man jedem Ergebnis die Augensumme zu, dann erhalten wir eine Zufallsvariable in der Form:

$X((1|1)) = 2 \qquad X((1|3)) = 4 \qquad X((6|3)) = 9$

$X = 4$ steht für das Ergebnis: Augensumme gleich 4, also für $\{(1|3);(2|2);(3|1)\}$

$X = 2$ steht für das Ergebnis: Augensumme gleich 2, also für $\{(1|1)\}$

Unter einer Zufallsvariablen X eines Zufallsexperimentes versteht man eine Funktion, die jedem Ergebnis e_i der Ergebnismenge E dieses Experimentes eine Zahl zuordnet.

$X: e_i \rightarrow X(e_i)$ in Analogie zur Funktion f mit $f: x \rightarrow f(x)$

Wertetabelle einer Zufallsvariablen für den Wurf zweier Würfel, deren Augenzahl addiert wird.

Ergebnis	(1|1)	(1|2)	(2|1)	(1|3)	(2|2)	(3|1)	....	(5|6)	(6|5)	(6|6)
$X(e_i) = x_1$	2	3	3	4	4	4	....	11	11	12

Wahrscheinlichkeitsverteilung

Wird beim Werfen mit zwei Würfeln jedem Ergebnis die Augensumme zugeordnet, so entsteht die Zufallsvariable X.

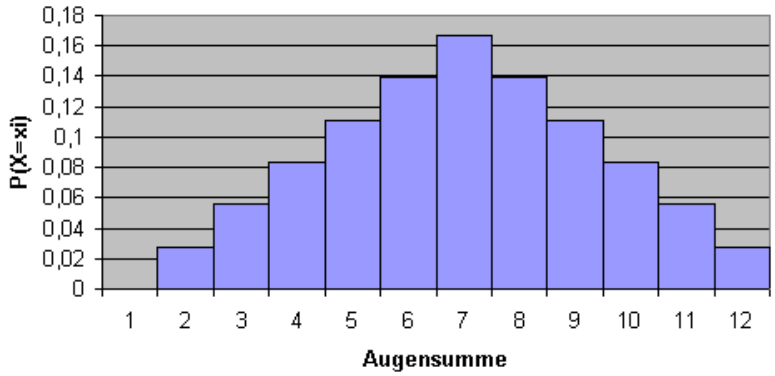
Definition 96:

Ordnet man nun jedem Wert dieser Zufallsvariablen ihre Wahrscheinlichkeit zu, so entsteht eine Wahrscheinlichkeitsverteilung (Wahrscheinlichkeitsfunktion). Die Wahrscheinlichkeitsverteilung oder Verteilung der Zufallsgröße kann man durch eine Tabelle und ein Histogramm darstellen.

Tabelle:

x_i	2	3	4	5	6	7	8	9	10	11	12
$P(X = x_i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Histogramm

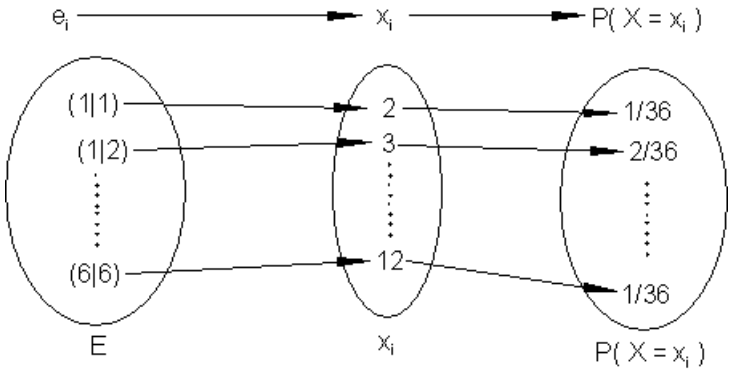


Unter einer Wahrscheinlichkeitsverteilung (Wahrscheinlichkeitsfunktion) f der Zufallsvariablen X versteht man die Funktion f mit

$$f : x_i \rightarrow P(X = x_i)$$

Der Funktionswert $f(x) = P(X = x_i)$ gibt die Wahrscheinlichkeit dafür an, dass X den Wert x_i annimmt.

Funktionsdarstellung zum Beispiel werfen zweier Würfel, deren Augensumme gebildet wird.



Erwartungswert einer Wahrscheinlichkeitsverteilung

Mit Hilfe der Wahrscheinlichkeit möchte man z. B. bei Glücksspielen Aussagen über den zu erwartenden Gewinn bzw. Verlust machen. Es stellt sich die Frage: Welchen Gewinn pro Spiel kann man bei häufiger Durchführung erwarten?

Zur Veranschaulichung betrachten wir wieder die Augensumme der zwei Würfel.

Beispiel 93:

Man könnte ein Glücksspiel daraus machen, indem man folgende Regel aufstellt:

Regel: Die in einem Wurf erreichte Augensumme wird in € ausgezahlt.

Der Betreiber des Spiels muss sich natürlich Gedanken darüber machen, wie hoch der Einsatz pro Spiel sein muss, damit er keinen Verlust erleidet.

Dazu muss er wissen, welchen Betrag er im Mittel pro Spiel bei sehr vielen Spielen auszuzahlen hat. So hoch muss auch mindestens der Einsatz ein.

Ähnlich wie bei der Mittelwertbildung aus einer Häufigkeitsverteilung in der beschreibenden Statistik kann man durch Multiplikation der Auszahlungsbeträge mit ihren Wahrscheinlichkeiten einen Wert bilden.

Diesen Wert nennen wir Erwartungswert.

Für unser Beispiel bedeutet der Wert 7, dass bei einer hohen Anzahl von Spielen im Mittel 7 € pro Spiel auszuzahlen sind.

x_i	$P(X = x_i)$	$x_i \cdot P(X = x_i)$
2	$\frac{1}{36}$	$2 \cdot \frac{1}{36} = \frac{2}{36}$
3	$\frac{2}{36}$	$3 \cdot \frac{2}{36} = \frac{6}{36}$
4	$\frac{3}{36}$	$4 \cdot \frac{3}{36} = \frac{12}{36}$
5	$\frac{4}{36}$	$5 \cdot \frac{4}{36} = \frac{20}{36}$
6	$\frac{5}{36}$	$6 \cdot \frac{5}{36} = \frac{30}{36}$
7	$\frac{6}{36}$	$7 \cdot \frac{6}{36} = \frac{42}{36}$
8	$\frac{5}{36}$	$8 \cdot \frac{5}{36} = \frac{40}{36}$
9	$\frac{4}{36}$	$9 \cdot \frac{4}{36} = \frac{36}{36}$
10	$\frac{3}{36}$	$10 \cdot \frac{3}{36} = \frac{30}{36}$
11	$\frac{2}{36}$	$11 \cdot \frac{2}{36} = \frac{22}{36}$
12	$\frac{1}{36}$	$12 \cdot \frac{1}{36} = \frac{12}{36}$
	Erwartungswert $E(X)$	$\frac{252}{36} = 7$

Der Betreiber des Spiels muss also mindestens einen Einsatz von 7 € pro Spiel verlangen, damit er keinen Verlust erleidet.

Die Auszahlungsbeträge oder auch Ausspielungen entsprechen der Zufallsvariablen

X mit den Werten: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

Nun betrachten wir das Spiel aus der Sicht eines Spielers, der pro Spiel 7 € Einsatz zahlen muss.

Für ihn berechnet sich der Gewinn aus:

Gewinn = Ausspielung - Einsatz.

Der Gewinn entspricht nun einer Zufallsvariablen, die wir Y nennen, also

Y mit den Werten: -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5

Damit lässt sich nun der Erwartungswert für den Gewinn ermitteln.

$$E(Y) = -5\cdot\frac{1}{36} - 4\cdot\frac{2}{36} - 3\cdot\frac{3}{36} - 2\cdot\frac{4}{36} - 1\cdot\frac{5}{36} + 1\cdot\frac{5}{36} + 2\cdot\frac{4}{36} + 3\cdot\frac{3}{36} + 4\cdot\frac{2}{36} + 5\cdot\frac{1}{36} = 0$$

Der Erwartungswert für einen Gewinn ist 0. Das bedeutet, auf lange Sicht gewinnt der Spieler nichts. Aber er verliert auch nichts. Die Chancen sind ausgeglichen.

Erwartungswert von X

Hat eine Zufallsvariable X die Werte $x_1; x_2; \ldots; x_n$ dann heißt:

$$E(X) = x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots + x_n \cdot P(X = x_n) = \sum_{i=1}^{n} x_i \cdot P(X = x_i)$$

Erwartungswert von X

Definition 97:

Ist E(X) > 0, so nennt man das Spiel günstig für den Spieler.

Ist E(X) = 0, so nennt man das Spiel fair.

Ist E(X) < 0, so nennt man das Spiel ungünstig (unfair) für den Spieler.

Bemerkung 53:

- Der Erwartungswert ist der zu erwartende Mittelwert von X in einer Reihe von Zufallsversuchen.
- Während sich der Mittelwert - eine Größe aus der beschreibenden Statistik - auf die Vergangenheit bezieht, also auf Werte, die in einer Stichprobe tatsächlich aufgetreten sind, beschreibt der Erwartungswert eine Größe, die sich auf die Zukunft bezieht, also auf eine Größe, mit der auf lange Sicht zu rechnen ist.
- Statt $E(X)$ schreibt man auch μ_x oder kürzer μ.
  Statt $P(X = x_i)$ schreibt man auch p_i.
- Wie beim Mittelwert gehört auch der Erwartungswert in vielen Fällen nicht zu den Werten die die Zufallsvariable X annehmen kann.

Indexberechnung

Der Preisindex für die Lebenshaltung

Der Preisindex für die Lebenshaltung ist ein wichtiger Bestandteil des preisstatistischen Berichtssystems für die Bundesrepublik Deutschland.

Dieses System umfasst u.a. den Index der **Erzeugerpreise** gewerblicher Produkte, Preisindizes für **Bauwerke**, den Index der Erzeugerpreise **land- und forstwirtschaftlicher Produkte**, den Index der **Großhandelsverkaufspreise**, den Index der **Einzelhandelspreise** und den Index der **Ausfuhrpreise**.

Der Preisindex für die Lebenshaltung misst die durchschnittliche Preisveränderung aller Waren und Dienstleistungen, die von privaten Haushalten gekauft bzw. verbraucht werden.

Mit dem Preisindex für die Lebenshaltung wird somit die Veränderung der Verbraucherpreise umfassend abgebildet (Verbraucherpreisindex). Er wird deshalb oft zur **Messung der Geldwertentwicklung** verwendet

Der Preisindex für die Lebenshaltung misst die durchschnittliche Preisveränderung aller Waren und Dienstleistungen, die von privaten Haushalten gekauft bzw. verbraucht werden.

Mit dem Preisindex für die Lebenshaltung wird somit die Veränderung der Verbraucherpreise umfassend abgebildet (Verbraucherpreisindex). Er wird deshalb oft zur Messung der Geldwertentwicklung verwendet.

Der Preisindex für die Lebenshaltung bezieht sich in der weitesten Abgrenzung auf alle privaten Haushalte der Bundesrepublik Deutschland. Daneben werden getrennte Preisindizes für das frühere Bundesgebiet und die neuen Länder einschl. Berlin-Ost sowie für spezielle Haushaltstypen ausgewiesen.

Da diese speziellen Haushaltstypen die Zusammensetzung der privaten Haushalte in der Bundesrepublik Deutschland hinsichtlich ihrer sozialen und ökonomischen Merkmale nicht mehr repräsentativ abbilden, wird das Statistische Bundesamt deren Veröffentlichung mit der nächsten Umstellung auf das Basisjahr 2000 einstellen.

Die Preisveränderungen werden gemäß der Verbrauchsbedeutung, die den Waren und Dienstleistungen im Budget der privaten Haushalte zukommt, im Preisindex berücksichtigt. Hierzu wird eine Verbrauchsstruktur auf der Grundlage der Ausgaben der privaten Haushalte für die Käufe von Waren und Dienstleistungen bestimmt. Die Ausgaben der privaten Haushalte für Waren und Dienstleistungen werden auf Stichprobenbasis in regelmäßigen Haushaltsbefragungen ermittelt.

Berechnet wird der Preisindex für die Lebenshaltung als sog. Laspeyres-Preisindex mit festem Basisjahr, d.h. die Indexwerte beziehen sich auf die Verbrauchsstrukturen desselben Jahres, das als Basisjahr festgelegt wird. Die Verbrauchsstrukturen werden bis zur Einführung eines neuen Basisjahres konstant gehalten.

Zur Monatsmitte werden Preise in 190 Berichtsgemeinden im ganzen Bundesgebiet erhoben. Die Berichtsgemeinden sind regional über die gesamte Bundesrepublik Deutschland verteilt (118 Gemeinden im Westen, 72 Gemeinden im Osten). Großstädte

werden ebenso abgedeckt wie mittlere und kleine Gemeinden (bis zu einer Einwohnerzahl von mindestens 5.000).

Im Preisindex für die Lebenshaltung werden die Preisveränderungen von etwa 750 genau beschriebenen Waren und Dienstleistungen zusammengefasst. Die Waren und Dienstleistungen werden mit dem Ziel ausgewählt, den Verbrauch der privaten Haushalte hinreichend genau zu repräsentieren.

Insgesamt werden etwa 350.000 Preisreihen für das gesamte Bundesgebiet ermittelt.

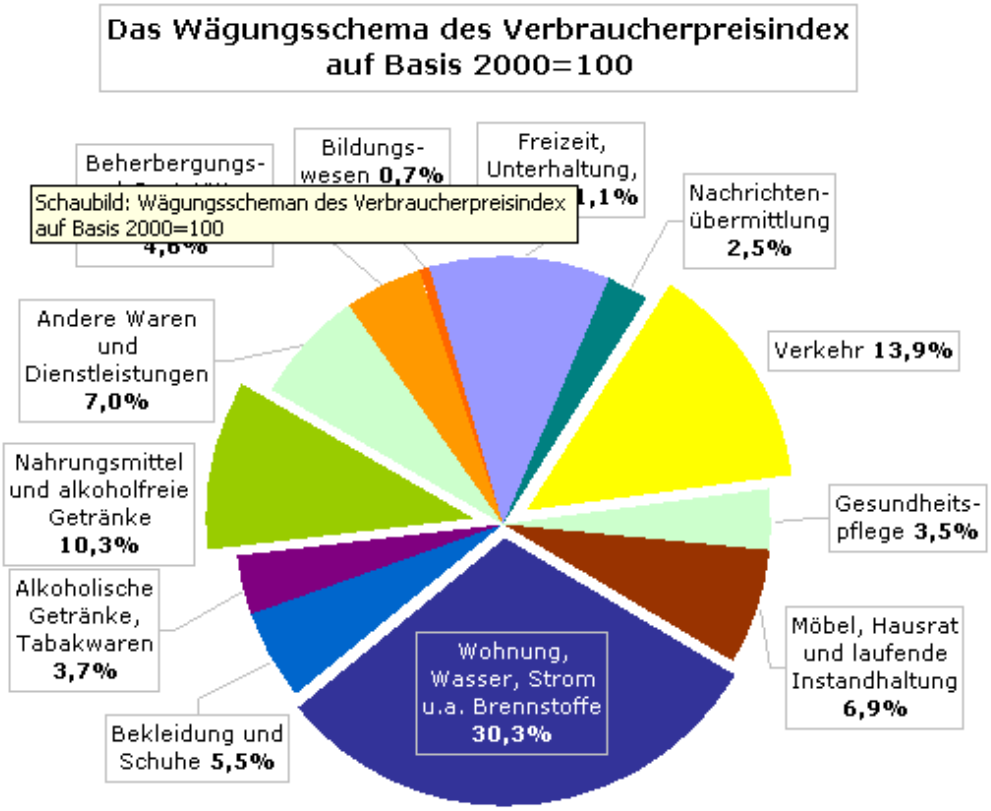
Aus diesen Preisreihen berechnen die 16 Statistischen Landesämter und das Statistische Bundesamt Verbraucherpreisindizes. Monatlich werden etwa 150 Positionen veröffentlicht. Noch tiefer gegliederte Angaben für die Bundesrepublik Deutschland, das frühere Bundesgebiet und die neuen Länder einschl. Berlin-Ost sind auf Datenträger verfügbar.

Daraus erstellt das Statistische Bundesamt elf Verbraucherpreisindizes, jeweils vier für das frühere Bundesgebiet und die neuen Länder (jeweils einen umfassenden für alle privaten Haushalte und drei für spezielle Haushaltstypen) und drei für die Bundesrepublik Deutschland insgesamt (Preisindex für die Lebenshaltung aller privaten Haushalte, Preisindex für den Einzelhandel, Gastgewerbepreisindex). Zusätzlich veröffentlicht das Statistische Bundesamt seit 1997 für Deutschland einen innerhalb der Mitgliedstaaten der Europäischen Union harmonisierten Verbraucherpreisindex (sog. HVPI). Darüber hinaus stellen auch einige Statistische Landesämter Verbraucherpreisindizes bereit, die auf das jeweilige Bundesland bezogen sind.

An aktuellen Ergebnissen des Preisindex für die Lebenshaltung besteht großes Interesse. Das Statistische Bundesamt veröffentlicht deshalb zum Ende eines jeden Berichtsmonats vorläufige Ergebnisse des Preisindex für die Lebenshaltung aller privaten Haushalte. Diese Schätzungen beruhen auf den Ergebnissen von sechs Bundesländern (Baden-Württemberg, Bayern, Brandenburg, Hessen, Nordrhein-Westfalen, Sachsen). Endgültige Ergebnisse werden ca. zwei Wochen später veröffentlicht.

Der Preisindex für die Lebenshaltung ist ein wichtiger Indikator für die Beurteilung der Geldwertstabilität, eines der herausragenden wirtschaftspolitischen Ziele. Der Index dient auch zur Absicherung des Wertes von Forderungen in längerfristigen Vertragsbeziehungen. Nutzer der Ergebnisse sind Öffentlichkeit, Bundesregierung und Bundesbank, Tarifparteien, Banken und Finanzdienstleister, Mieter und Vermieter von Wohnungen und Geschäften, ehemalige Betriebsinhaber, die ihren Betrieb auf Rentenbasis verkauft haben

Wägungsschemata für den Preisindex für die Lebenshaltung aller privaten Haushalte in Deutschland Angaben in Promille



COICOP-VPI[1)]	Gewichte 1991	Gewichte 1995
01 Nahrungsmittel und alkoholfreie Getränke	144,81	131,26
02 Alkoholische Getränke und Tabakwaren	45,19	41,67
03 Bekleidung und Schuhe	76,89	68,76
04 Wohnung, Wasser, Elektrizität, Gas und andere Brennstoffe	240,46	274,77
05 Hausrat und laufende Instandhaltung des Hauses	72,87	70,56
06 Gesundheitspflege	30,56	34,39
07 Verkehr	156,77	138,82
08 Nachrichtenübermittlung	17,92	22,66
09 Freizeit und Kultur	99,59	103,57

10 Bildungswesen	5,42	6,51
11 Hotels, Cafés und Restaurants	58,44	46,08
12 Verschiedene Waren und Dienstleistungen	51,08	60,95
Insgesamt	1000,00	1000,00

[1)]Classification of Individual Consumption by Purpose in einer für Zwecke der Verbraucherpreisstatistik leicht abgewandelten Form.

Eigenschaften von Indexzahlen

Bemerkung 54:

- Indexzahlen sind besondere Messzahlen
- Vergleich von Größen in verschiedenen Perioden (Zeitpunkten oder Zeitabschnitten)
  Preisindex
  Baukostenindex
  Mengenindex
  Umsatzindex
  Aktienindex
- Meist auf eine Basis bezogen ("1990=100")

Beispiel 94:

Die Großhandelspreise sind im Mai um vier Prozent gestiegen. Im April waren es 4,1 Prozent

Einfache Indexzahlen

Eine Ware, die am 1.7.2003 143,50 Euro gekostet hat, kostet am 1.7.2004 148,30 Euro

$$\text{Index} = \frac{148{,}30}{143{,}50} = 1{,}0334 = 103{,}34\%$$

Daraus folgt: Die Ware ist um 3,34 % teurer geworden.

Bei der Bezeichnung müssen die beiden Jahre angegeben werden.

Wie ist die Vorgehensweise, wenn man die Preise eines "Warenkorbes" aus n Waren vergleichen will?

Es ergeben sich hieraus die folgenden Bezeichnungen:

Definition 98:

t: Laufende Nummer der Periode (Jahr, Monat, Stichtag)

t=0 Basisperiode

$p_t^{(i)}$: Preis des Gutes i in der Periode t

$m_t^{(i)}$: Menge des Gutes i in der Periode t

Preisindex des Gutes i von der Periode 0 zur Periode t

$$P_{0,t}^{(i)} = \frac{p_t^{(i)}}{p_0^{(i)}}$$

Beispiel 95:

Artikel	Preise 1995	Preis 2000	Preisindex des Artikels i
Schnellzuglok	250	320	$P_{1995,2000}^{(1)} = \frac{320}{250} = 1{,}28$
Güterwagen	40	48	$P_{1995,2000}^{(2)} = \frac{48}{40} = 1{,}20$
Gleis	4	5	$P_{1995,2000}^{(3)} = \frac{5}{4} = 1{,}25$
Weiche	60	69	$P_{1995,2000}^{(4)} = \frac{69}{60} = 1{,}15$

Durchschnittliche Preissteigerung

Ansatz 1:

$$\frac{1{,}28 + 1{,}20 + 1{,}25 + 1{,}15}{4} = 1{,}22$$

Kritik:

Die Preissteigerungen eines billigen Artikels (Schiene) ist hier genauso gewichtig wie die eines teuren Artikels (Weiche).

Ansatz 2:

$$\frac{320 + 48 + 5 + 69}{250 + 40 + 4 + 60} = 1{,}25$$

Kritik:

Es bleibt unberücksichtigt, dass man mehr Schienen als Weichen und mehr Wagen als Loks braucht.

Ansatz 3:

Man berücksichtigt einen "Warenkorb":

Wie viel Stück von jedem Artikel benötigt man für eine haushaltsübliche Eisenbahnanlage?

Erweiterte Tabelle:

Artikel	Preise 1995	Preis 2000	Menge	Kosten 1995	Kosten 2000
Schnellzuglok	250	320	1	250	320
Güterwagen	40	48	4	160	192
Gleis	4	5	20	80	100
Weiche	60	69	2	120	138
Summe:				610	750

Berechnung: $P^{F}_{1995,2000} = \frac{750}{610} = 1{,}229$

Änderung des Warenkorbes

Wir sind stillschweigend davon ausgegangen, dass der Warenkorb unverändert geblieben ist.

Tatsächlich ändert sich das Konsumverhalten.

Artikel	Menge 1995	Menge 2000
Schnellzuglok	1	1
Güterwagen	4	4
Gleis	20	30
Weiche	2	4

Folgende Frage stellt sich:

Soll man die Preise mit dem Warenkorb 1995 (Basisjahr) oder 2000 (Berichtsjahr gewichten?

Artikel	Preise 1995	Preis 2000	Menge	Kosten 1995	Kosten 2000
Schnellzuglok	250	320	1	250	320
Güterwagen	40	48	4	160	192
Gleis	4	5	30	120	150
Weiche	60	69	4	240	276
Summe:				770	938

Berechnung: $P^{E}_{1995,2000} = \frac{938}{770} = 1{,}218$

Preisindizes

Definition 99:

Ein Preisindex ist ein statistisches Konstrukt, das eine Aussage über die **Höhe der Inflation in einem volkswirtschaftlichen Bereich** machen soll.

Dazu wird ermittelt, wie sich die Preise der Güter eines für diesen Wirtschaftsbereich repräsentativen Warenkorbes im Durchschnitt über die Zeit geändert haben.

Bemerkung 55:

- Auch Aussagen über regionale Preisniveau-Unterschiede können mit einem Preisindex ausgedrückt werden, der dann in analoger Weise wie der zeitliche Preisindex aufgebaut ist. Derartige Preisindizes werden jedoch selten ermittelt.
- In der Preisstatistik wird ein ganzes Bündel von Preisindizes ermittelt. Die folgenden Betrachtungen werden der Einfachheit halber nur für einen Einkaufs-Preisindex (z. B. Verbraucherpreisindex) angestellt.

In der Preisstatistik sind zwei Konzepte der Bildung von Preisindizes weit verbreitet:

Der Preisindex nach Laspeyres

In der Volkswirtschaftlichen Gesamtrechnung wird dagegen entsprechend internationalen Konventionen - in Deutschland ab 2005 - eine Preisbereinigung mit sog. Kettenindizes vorgenommen.

Laspeyres-Index

Definition 100:

Preisindex nach Laspeyres	$P_{0,t}^{L} = \frac{\sum_{i=1}^{n} p_t^i \cdot m_0^i}{\sum_{i=1}^{n} p_0^i \cdot m_0^i}$
	m_0^i *Menge des Basisjahres*
	p_0^i *Preis des Basisjahres*
	p_t^i *Preis des Berichtsjahres*

Wichtige Eigenschaften:

- In der amtlichen Statistik des In- und Auslandes hat diese Indexzahl eine hohe Bedeutung.

Der Preisindex nach Laspeyres antwortet auf folgende Frage:

"Was kostet der Warenkorb der Basisperiode zu Preisen der Berichtsperiode im Vergleich zum Preis in der Basisperiode?"

Vorteile des Preisindex nach Laspeyres:

- plausible ökonomische Aussagekraft
- Konstanz des Warenkorbes
- geringer Rechenaufwand
- geringer Erhebungsaufwand (Wir müssen nur die Preise beobachten.)

Nachteile des Preisindex nach Laspeyres:

- Konstanz des Warenkorbes (Verbrauchsstrukturänderungen, neue Güter, Änderung der Produktqualität, veraltete Produkte)
- Der Warenkorb muss regelmäßig auf seine Tauglichkeit geprüft werden

Definition 101:

Der Laspeyres-Preisindex (benannt nach Etienne Laspeyres) untersucht, was der Kauf eines Warenkorbes in der Zusammensetzung der Periode 0 (Basisjahr) in der Periode t kostet im Vergleich zum Kauf des gleichen Warenkorbes in der Periode 0.

Bei der Ermittlung werden – formal gesehen – die aktuellen Kosten des Warenkorbes, wie er sich im Basisjahr zusammensetzte (Summe über die Mengen q der Güter i zum Zeitpunkt 0, multipliziert mit ihren aktuellen Preisen p), auf die Kosten dieses Warenkorbes zum Zeitpunkt 0 bezogen. In der Praxis der amtlichen Statistik wird der Laspeyres-Preisindex jedoch als gewogener Mittelwert des Verhältnisses der aktuellen Güterpreise bezogen auf die Preise des Basisjahres ("Messzahl") ermittelt. Die Gewichte sind dabei im Falle eines Verbraucherpreisindex die Ausgaben der privaten Haushalte für die einzelnen Güter des Warenkorbes.

Der Laspeyres-Preisindex stellt vor allem auf die Ermittlung "**reiner Preisänderungen**" ab.

Die Reaktion der Käufer auf Preisänderungen, nämlich der Wechsel von teurer zu billiger gewordenen Gütern ("**Substitutionseffekt**"), wirkt sich auf den Laspeyres-Index nicht aus. Preiserhöhungen wirken sich daher weniger stark auf das Verbraucherbudget aus, als es dieser Index ausweist.

Der praktische Vorteil von Laspeyres-Indizes besteht darin, dass die Gewichte nur für das Basisjahr ermittelt werden müssen und dann unverändert bleiben. Damit sie trotzdem als repräsentativ für das aktuelle Preisgeschehen gelten können, werden sie in der amtlichen Statistik - ebenso wie die Zusammensetzung des Warenkorbes - regelmäßig (in der Regel alle 5 Jahre) aktualisiert.

Die Bestimmung des Verbraucherpreisindex erfolgt in Deutschland mit Hilfe eines Laspeyres-Index.

Paasche-Index

Definition 102:

Preisindex nach Paasche	$P_{0,t}^{P} = \frac{\sum_{i=1}^{n} p_t^i \cdot m_t^i}{\sum_{i=1}^{n} p_0^i \cdot m_t^i}$
	m_t^i *Menge des Berichtsjahres*
	p_0^i Pr*eis des Basisjahres*
	p_t^i Pr*eis des Berichtsjahres*

Wichtige Eigenschaften:

Der einzige Unterschied zwischen den Preisindizes von Laspeyres und Paasche besteht darin, dass die Warenkörbe mit denen die Gewichte ermittelt werden, zu unterschiedlichen Zeitpunkten zusammengestellt werden.

Vorteile des Preisindex nach Paasche:

- Wir können stets einen neuen, aktuellen Warenkorb zusammenstellen, der den Bedürfnissen der Nachfrager entspricht.

Nachteile des Preisindex nach Paasche:

- höherer Berechnungs- und Erhebungsaufwand
- Theoretisch können wir heute Produkte in den Warenkorb legen, die es zwei oder drei Perioden vorher noch gar nicht gegeben hat.

Definition 103:

Der Paasche-Preisindex (benannt nach Hermann Paasche) untersucht, was der Kauf eines Warenkorbes in der Zusammensetzung der Periode t in der Periode t kostet im Vergleich zum Kauf des gleichen Warenkorbes in der Periode 0 (Basisjahr).

Mit anderen Worten: **die Preise für ein zum Zeitpunkt t gekauftes Güterbündel werden damit verglichen, was für das gleiche Güterbündel zum Zeitpunkt 0 hätte bezahlt werden müssen**.

Bei der Ermittlung eines Paasche-Preisindex variieren also die Gewichte von Periode zu Periode.

Der Paasche-Preisindex misst die Preisentwicklung mit den Gewichten der aktuellen Periode, das heißt nachdem die Ausweichreaktion der Verbraucher auf veränderte Preise, nämlich der Wechsel von teurer zu billiger gewordenen Gütern ("Substitutionseffekt"), stattgefunden hat. Die "tatsächliche" Preiserhöhung ist daher höher, als es vom Paasche-Index ausgewiesen wird.

Die Alternativ-Darstellung weist den Paasche-Index als Ausgabengewichteten harmonischen Mittelwert der n Preisverhältnisse aus. Wegen des Substitutionseffektes, aber auch weil ein harmonischer Mittelwert kleiner ist als der entsprechende arithmetische Mittelwert (siehe auch Mittelwert), ist der Paasche-Index bei einem Einkaufs-Preisindex im Allgemeinen kleiner als der Lasyperes-Index.

Der übliche Maßstab für die Höhe einer Inflation in einem volkswirtschaftlichen Bereich ist die Veränderungsrate eines Preisindex für den Bereich.

Im Falle des Paasche-Index besteht das Problem, dass in diese Veränderungsrate nicht nur die Veränderung der Preise von $p_{i,t-1}$ zu $p_{i,t}$ eingeht, sondern auch die Veränderung der Mengen von $q_{i,t-1}$ zu $q_{i,t}$.

Ein (reiner) Paasche-Preisindex wird von der Amtlichen Statistik selten berechnet, da er durch die notwendigen regelmäßigen Aktualisierungen der Gewichte ressourcen- und zeitaufwendig ist. Er wird aber bei der Deflationierung von Umsatzentwicklungen benötigt, um "echte" Mengenentwicklungen als Laspeyres-Mengenindizes zu erhalten.

Vergleich zwischen den Preisindizes nach Laspeyres und Paasche

- Bei ‘normalen’ Nachfragereaktionen wird der Preisindex nach Laspeyres höher sein als der nach Paasche. Das kommt daher, dass die Preise für Güter in der Regel steigen, die Nachfrage aber sinkt.
- Die steigenden Preise werden von beiden Verfahren berücksichtigt. Nur aber Paasche berücksichtigt auch die Änderung der Nachfrage innerhalb des Warenkorbs.
- Trotzdem wird dem Preisindex nach Laspeyres in der Praxis häufig der Vorzug gewährt, weil er den geringeren Erhebungs- und Berechnungsaufwand erfordert

Beispiel 96:

Eisenbahnbeispiel

Laspeyres-Index

$$P^{L}_{1995,2000} = \frac{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{1995}}{\sum_{i=1}^{n} p^{i}_{1995} \cdot m^{i}_{1995}} = \frac{320 \cdot 1 + 48 \cdot 4 + 5 \cdot 20 + 69 \cdot 2}{250 \cdot 1 + 40 \cdot 4 + 4 \cdot 20 + 60 \cdot 2} = \frac{750}{610} = 1{,}229$$

Paasche-Index

$$P^{P}_{1995,2000} = \frac{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{2000}}{\sum_{i=1}^{n} p^{i}_{1995} \cdot m^{i}_{2000}} = \frac{320 \cdot 1 + 48 \cdot 4 + 5 \cdot 30 + 69 \cdot 4}{250 \cdot 1 + 40 \cdot 4 + 4 \cdot 30 + 60 \cdot 4} = \frac{938}{770} = 1{,}218$$

Beispiel 97:

Jahr	Preis Schrauben	Menge Schrauben	Preis Nagel	Menge Nagel
2000	10	10	20	5
2001	11	11	24	4
2002	12	12	28	3

Berechnen Sie nach folgender Tabelle die folgenden Preisindizes

$$P^{L}_{2000,2001} = \frac{\sum_{i=1}^{n} p^{i}_{2001} \cdot m^{i}_{2000}}{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{2000}} = \frac{11 \cdot 10 + 24 \cdot 5}{10 \cdot 10 + 20 \cdot 5} = \frac{230}{200} = 1{,}15$$

$$P^{L}_{2000,2002} = \frac{\sum_{i=1}^{n} p^{i}_{2002} \cdot m^{i}_{2000}}{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{2000}} = \frac{12 \cdot 10 + 28 \cdot 5}{10 \cdot 10 + 20 \cdot 5} = \frac{260}{200} = 1{,}3$$

$$P^{L}_{2001,2002} = \frac{\sum_{i=1}^{n} p^{i}_{2002} \cdot m^{i}_{2001}}{\sum_{i=1}^{n} p^{i}_{2001} \cdot m^{i}_{2001}} = \frac{12 \cdot 11 + 28 \cdot 4}{11 \cdot 11 + 24 \cdot 4} = \frac{244}{217} = 1{,}12$$

$$P^{P}_{2000,2001} = \frac{\sum_{i=1}^{n} p^{i}_{2001} \cdot m^{i}_{2001}}{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{2001}} = \frac{11 \cdot 11 + 24 \cdot 4}{10 \cdot 11 + 20 \cdot 4} = \frac{217}{190} = 1{,}14$$

$$P^P_{2000,2002} = \frac{\sum_{i=1}^{n} p^i_{2002} \cdot m^i_{2002}}{\sum_{i=1}^{n} p^i_{2000} \cdot m^i_{2002}} = \frac{12 \cdot 12 + 28 \cdot 3}{10 \cdot 12 + 20 \cdot 3} = \frac{228}{180} = 1{,}27$$

$$P^P_{2001,2002} = \frac{\sum_{i=1}^{n} p^i_{2002} \cdot m^i_{2002}}{\sum_{i=1}^{n} p^i_{2001} \cdot m^i_{2002}} = \frac{12 \cdot 12 + 28 \cdot 3}{11 \cdot 12 + 24 \cdot 3} = \frac{228}{204} = 1{,}12$$

Fisher-Preisindex

Definition 104:

Der Fisher-Preisindex (benannt nach Irving Fisher) ist das geometrische Mittel der Preisindizes nach Paasche und Laspeyres. Der Fisher-Preisindex wird in der Statistik auch "Fishers idealer Preisindex" genannt.

$P^F(t) = \sqrt{P^L(t) \cdot P^P(t)}$

Bemerkung 56:

- Der Preisindex nach Fisher versucht die Neigung des Laspeyres-Preisindex zur Überschätzung des Preisanstiegs und die Neigung des Paasche-Preisindex zur Unterschätzung des Preisanstiegs durch Mittelung auszugleichen. Da in seine Berechnung jedoch der Paasche-Index eingeht, wird er in der amtlichen Statistik selten berechnet.

Beispiel 98:

Berechnung mit unserem Eisenbahnbeispiel:

$$P^F(t) = \sqrt{P^L(t) \cdot P^P(t)} = \sqrt{\mathrm{P}^{\mathrm{L}}_{2000,2001} \cdot \mathrm{P}^{\mathrm{P}}_{2000,2001}} = \sqrt{1{,}229 \cdot 1{,}218} = 1{,}223$$

Mengenindizes

- Umgekehrte Fragestellung
- statt: "Wie haben sich die Preise entwickelt" (gewichtet mit Mengen)
- jetzt: "Wie haben sich die Mengen entwickelt" (gewichtet mit Preisen)
- Gewichtung mit den Preisen des Basisjahres: Mengenindex nach Laspeyres
- Gewichtung mit den Preisen des Berichtsjahres: Mengenindex nach Paasche

Analog zur Berechnung der Preisindizes kann man auch **Mengenindexzahlen** ausrechnen.

Definition 105:

Hierbei wird einfach der Preis konstant gehalten und die Mengenänderungen betrachtet. Je nachdem, ob wir die Preise der Basis- oder der Berichtsperiode wählen, sprechen wir dann von Mengenindizes nach Laspeyres oder Paasche.

Mengenindex nach Laspeyres

Definition 106:

$$M_{0,t}^{L} = \frac{\sum_{i=1}^{n} p_0^i \cdot m_t^i}{\sum_{i=1}^{n} p_0^i \cdot m_0^i}$$

m_0^i *Menge des Basisjahres*

m_t^i *Menge des Berichtsjahres*

p_0^i *Preis des Basisjahres*

p_t^i *Preis des Berichtsjahres*

Mengenindex nach Paasche

Definition 107:

$$M_{0,t}^{P} = \frac{\sum_{i=1}^{n} p_t^i \cdot m_t^i}{\sum_{i=1}^{n} p_t^i \cdot m_0^i}$$

m_0^i *Menge des Basisjahres*

m_t^i *Menge des Berichtsjahres*

p_0^i *Preis des Basisjahres*

p_t^i *Preis des Berichtsjahres*

Beispiel 99:

Eisenbahnanlage

$$M^{L}_{1995,2000} = \frac{\sum_{i=1}^{n} p^{i}_{1995} \cdot m^{i}_{2000}}{\sum_{i=1}^{n} p^{i}_{1995} \cdot m^{i}_{1995}} = \frac{250 \cdot 1 + 40 \cdot 4 + 4 \cdot 30 + 60 \cdot 4}{250 \cdot 1 + 40 \cdot 4 + 4 \cdot 20 + 60 \cdot 2} = \frac{770}{610} = 1{,}262$$

$$M^{P}_{1995,2000} = \frac{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{2000}}{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{1995}} = \frac{320 \cdot 1 + 48 \cdot 4 + 5 \cdot 30 + 69 \cdot 4}{320 \cdot 1 + 48 \cdot 4 + 5 \cdot 20 + 69 \cdot 2} = \frac{938}{750} = 1{,}251$$

Wert- oder Umsatzindizes

Definition 108:

Umsatzindexzahlen U verwenden wir immer dann, wenn wir sowohl die Veränderung von Preisen als auch die Veränderungen von Mengen in der Berichts- zur Basisperiode betrachten wollen.

Will man wissen, wie sich der Wert des Warenkorbs entwickelt hat, rechnet man folgendermaßen:

Umsatzindex

Definition 109:

$$U_{0,t} = \frac{\sum_{i=1}^{n} p^{i}_{t} \cdot m^{i}_{t}}{\sum_{i=1}^{n} p^{i}_{0} \cdot m^{i}_{0}}$$

Beispiel 100:

Eisenbahnanlage

$$U_{1995,2000} = \frac{\sum_{i=1}^{n} p^{i}_{2000} \cdot m^{i}_{2000}}{\sum_{i=1}^{n} p^{i}_{1995} \cdot m^{i}_{1995}} = \frac{320 \cdot 1 + 48 \cdot 4 + 5 \cdot 30 + 69 \cdot 4}{250 \cdot 1 + 40 \cdot 4 + 4 \cdot 20 + 60 \cdot 2} = \frac{938}{610} = 1{,}538$$

Daraus folgt:

Für eine Modelleisenbahn nach dem jeweiligen Standard muss man 2000 53,8% mehr bezahlen als 1995.

Kettenpreisindex

Definition 110:

Kettenpreisindizes (chainprices) ermitteln für jedes Jahr, wie viel die im Vorjahr gekaufte Waren im aktuellen Jahr kosten (in der Laspeyresform) bzw. wie viel die im aktuellen Jahr gekauften Waren im Vorjahr gekostet haben (in der Paascheform).

$$P^{K/L}(t) = \frac{\sum_{i=1}^{n} p_i(t) \cdot q_i(t-1)}{\sum_{i=1}^{n} p_i(t-1) \cdot q_i(t-1)} \cdot 100\%$$

$$P^{K/P}(t) = \frac{\sum_{i=1}^{n} p_i(t) \cdot q_i(t)}{\sum_{i=1}^{n} p_i(t-1) \cdot q_i(t)} \cdot 100\%$$

Bemerkung 57:

- Dadurch wird für jedes Jahr ein anderer Warenkorb zu Grunde gelegt und so bei der Ermittlung der Preisänderungen die jeweils aktuellsten Verbrauchsgewohnheiten berücksichtigt werden.
- Nachteil des Verfahrens ist, dass die Ergebnisse von Jahr zu Jahr nicht direkt vergleichbar sind - wegen des sich wandelnden Warenkorbes - und die längerfristige Betrachtung nur durch Verkettung (daher der Name des Index) der Jahresergebnisse möglich ist.

Der Harmonisierte Verbraucherpreisindex wird als Kettenindex (Laspeyresform) berechnet.

Harmonisierter Verbraucherpreisindex

Der harmonisierte Verbraucherpreisindex (HVPI) ist ein in der Europäischen Union erhobener Verbraucherpreisindex, dem kein EU-weit einheitlicher Warenkorb zugrunde liegt. Der HVPI ist die Kennzahl, mit der in der Europäischen Wirtschafts- und Währungsunion (EWWU) die Preisniveauentwicklung gemessen wird.

Die Berechnung eines HVPI ist erforderlich, da sich die nationalen Verbraucherpreisindizes auf Grund historischer Besonderheiten, unterschiedlichen gesellschaftlichen Rahmenbedingungen sowie abweichender Struktur der statistischen Systeme unterscheiden.

Zusätzlich zu den nationalen Verbraucherpreisindizes werden daher in den EWWU-Staaten (in Deutschland seit 1997) auch nationale HVPIs berechnet. Das Statistische Amt der EU (Eurostat) überwacht die Einhaltung der Regeln zur Ermittlung der nationalen HVPIs und berechnet auf dieser Grundlage Verbraucherpreisindizes für die EU und für den europäischen Wirtschaftsraum insgesamt.

Die monatlichen Werte für den HVPI werden von Eurostat immer drei Wochen nach Monatsende veröffentlicht. Die Konzeption des HVPI versucht die zuvor beschriebenen Messfehler zu berücksichtigen. Zu den Verordnungen, mit denen Seitens der Europäischen Kommission die traditionelle Inflationsmessung des Laspeyres-Preisindex zukünftig verbessert werden soll, gehören z. B.:

- die Überprüfung der Güterauswahl: Gelangt ein Gut in einem Land der EWWU zur Marktbedeutung, müssen auch die anderen Länder die Einbeziehung in den Warenkorb überprüfen.

- die Überprüfung der Qualität von Gütern und ihrer Gewichte während der Laufzeit des Index: Damit wird ein deutlicher Druck in Richtung auf eine nahezu jährliche Überprüfung des Index ausgeübt.

Kettenvolumenindex oder Kettenmengenindex

Zu Kettenpreisindizes gibt es - wie bei den traditionellen Indizes mit festem Preis- bzw. Mengenbezugsjahr - korrespondierenden Kettenvolumenindizes. Der Laspeyres-Volumenindex berechnet sich dabei analog zum Kettenpreisindex (Laspeyres-Form) indem die mit Vorjahrespreisen bewerteten Mengen des aktuellen Jahres durch die nominellen Angaben des Vorjahres geteilt werden. Entsprechend ergibt sich der Paasche-Volumenindex aus dem Verhältnis der nominellen Angaben des laufenden Jahres zu den mit Preisen des aktuellen Jahres bewerteten Mengen des Vorjahres.

Zusammenhang zwischen Kettenpreis- und Kettenvolumenindizes

Ebenso wie bei der traditionellen Methode mit festem Preisbezugsjahr gilt, dass nominelle Angaben zu einem Kettenvolumenindex führen, wenn durch den entsprechenden Paasche-Index geteilt wird (so z.B. in den VGR) oder dass die Division der nominellen Angaben durch eine Laspeyres-Preisindex in einem Paasche-Volumenindex resultiert (theoretische beim HVPI möglich, jedoch in der amtlichen deutschen Statistik nicht praktiziert).

Eigenschaften von Kettenindizes

Da Kettenindizes nicht mehr aus einem einfachen Bruch, sondern aus einer (steigenden) Anzahl von Faktoren bestehen, sind Teilkomponenten nicht mehr ohne weiteres addierbar oder mit den relativen Anteilen eines anderen Jahres als dem des unmittelbaren Vorjahres zusammenwiegbar. Daher muss zur Aggregation von Zeitreihen erst die zu aggregierenden Zeitreihen entkettet, dann mit den jeweiligen Vorjahresanteilen zusammengewichtet und anschließend wieder verkettet werden (Rechenweise eines Excel-Makros, KIX-Makro, das von der Deutschen Bundesbank zur Aggregation von VGR-Größen auf Anfrage zur Verfügung gestellt wird).

Kettenindizes in der deutschen VGR

In den Volkswirtschaftlichen Gesamtrechnungen in der EU werden entsprechend den europäischen Vorgaben - in Deutschland seit April 2005 - die „realen“ Größen als Kettenindizes berechnet.

Aufgrund der mangelnden Additivität beziehungsweise der Komplexität der Aggregation normiert das Statistische Bundesamt die in Kettenindizes ausgewiesenen preisbereinigten Größen der VGR in einem bestimmten Basisjahr auf 100 % (normiert mit dem Jahr 2000 als 100 %). So hatte das BIP in Deutschland im Jahr 2001 einen Betrag von 2.113,06 Mrd. € und im Jahr 2005 einen Betrag von 2.245,50 Mrd. €. Der preisbereinigte Kettenindex dazu—normiert auf das Jahr 2000 mit 100 %—hatte für das Jahr 2001 den Wert 101,24 % und für das Jahr 2005 den Wert 103,67 %. Unter Zugrundelegung dieser Kettenindizes lag das BIP vom Jahre 2001 real um 1,24 % über seinem Wert von 2000 und im Jahre 2005 um 3,67 % über seinem Wert vom Jahre 2000.

Beispiel 101:

Nachfolgende Tabelle stellt die Berechnung eines **Laspeyres-Index** schematisch dar.

Die Spalten 2 und 3 enthalten die Preise für zwei Güter in den Jahren 0,1 und 2.

In den Spalten 4 und 5 stehen die jeweils gekauften Mengen, wobei für die Berechnung hier nur die Angaben des Jahres 0 relevant sind.

In den Spalten 6 und 7 werden die Preise mit den Mengen des Basisjahres multipliziert, anschließend addiert (Spalte 8) und so umbasiert, dass im Jahr 0 der Wert 100 beträgt. Spalte 10 gibt die aus dem Index abgeleiteten Inflationsraten an, die 15% im Jahr 1 und 30% im Jahr 2 betragen.

Laspeyres-Index

Jahr	$p_1(t)$	$p_2(t)$	$q_1(t)$	$q_2(t)$	$p_1(t)\ q_1(t_0)$	$p_2(t)\ q_2(t_0)$	$\Sigma(t)$	$\Sigma(t)\ \Sigma(t_0)^{-1}$	$\Delta\Sigma$
0	10	20	10	5	100	100	200	100 %	—
1	11	24	10	4	110	120	230	115 %	15 %
2	12	28	12	3	120	140	260	130 %	30 %
1	2	3	4	5	6	7	8	9	10
Spaltennummer									

Im Vergleich dazu ist die Berechnung eines **Paasche-Index** etwas aufwändiger.

Die Angaben in den Spalten 2 bis 5 sind die gleichen wie im vorhergehenden Beispiel.

In den Spalten 6 und 7 werden die Preise eines jeden Jahres mit den Mengen des gleichen Jahres multipliziert und die Ergebnisse anschließend addiert (Spalte 8).

In den Spalten 9 und 10 werden die Preise des Basisjahres mit den Mengen des jeweils laufenden Jahres multipliziert und anschließend addiert (Spalte 11).

In Spalte 12 werden die Ergebnisse in Spalte 8 und 11 ins Verhältnis zueinander gesetzt, anschließend wird der Wert wieder so umgerechnet, dass das Basisjahr = 100 ist (Spalte 13).

Die in Spalte 14 ausgewiesenen Inflationsraten sind, obwohl gleiche Preise unterstellt wurden, niedriger als im vorhergehenden Beispiel, das Gut 2, dessen Preise rascher steigen, weniger stark nachgefragt wird, also an Gewicht im Preisindex verliert.

Paasche-Index

Jahr	$p_1(t)$	$p_2(t)$	$q_1(t)$	$q_2(t)$	$p_1(t)$ $q_1(t)$	$p_2(t)$ $q_2(t)$	Σ_1	$p_1(t_0)$ $q_1(t)$	$p_2(t_0)$ $q_2(t)$	Σ_2	$\Sigma_1\Sigma_2^{-1}$		$\Delta\Sigma$
0	10	20	10	5	100	100	200	100	100	200	1	100 %	—
1	11	24	11	4	121	96	217	110	80	190	1,142	114,2 %	14,2 %
2	12	28	12	3	144	84	228	120	60	180	1,267	126,7 %	10,9 %
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spaltennummer													

Kettenindizes werden ähnlich ermittelt wie ein Paasche-Index, mit dem Unterschied, dass hier nur von Jahr zu Jahr gerechnet wird.

Man beachte aber: Die beiden letzten Spalten der Tabelle stehen in umgekehrter Reihenfolge.

Da stets der Preisindex des Vorjahres = 100 gesetzt ist, kann man aus Spalte 12 unmittelbar die Veränderungsraten ablesen, aus denen dann in Spalte 14 die Indizes errechnet werden.

Kettenindizes

Jahr	$p_1(t)$	$p_2(t)$	$q_1(t)$	$q_2(t)$	$p_1(t)$ $q_1(t)$	$p_2(t)$ $q_2(t)$	Σ_1	$p_1(t\text{-}1a)$ $q_1(t)$	$p_2(t\text{-}1a)$ $q_2(t)$	Σ_2	$\Sigma_1\Sigma_2^{-1}$	$\Delta\Sigma$	$\Sigma_1\Sigma_2^{-1}$
0	10	20	10	5	—	—	—	—	—	—	—	—	—
1	11	24	11	4	121	96	217	110	80	190	1,142	14,2%	114,2%
2	12	28	12	3	144	84	228	132	72	204	1,118	11,8%	127,7%
1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spaltennummer													

Indexreihen

Indizes werden meist jährlich neu berechnet und bilden dann Indexreihen

z.B. $P_{0,1}$, $P_{1,2}$, $P_{2,3}$ ... Steigerung gegenüber der Vorperiode

z.B. $P_{0,1}$, $P_{0,2}$, $P_{0,3}$ ... Steigerung gegenüber der Basisperiode

Bei Indizes, die sich aus Warenkörben errechnen, gilt im Allgemeinen nicht die "Rundprobe", d.h.

$$P_{0,t} \neq P_{0,1} \cdot P_{1,2} \cdot \ldots \cdot P_{t-1,t}$$

Man legt den Wert der Basisperiode meist zu "100" fest

Beispiel 102:

Lebenshaltungskosten

	Deutschland	Österreich
	1980=100	1966=100
1975	82,6	163,5
1976	86,3	175,5
1977	89,3	185,1
1978	91,6	191,7
1979	95,0	198,8
1980	100,0	211,4
1981	106,3	225,8
1982	112,0	238,1
1983	115,6	246,0
1984	118,4	260,0
1985	120,9	268,3
1986	120,7	272,8

Problem:

Wegen unterschiedlicher Basisjahre sind die Indizes nicht miteinander vergleichbar.

Lösung:

Man sucht ein gemeinsames Basisjahr (z. B. 1975)

Die Zahlen jeder Indexreihe werden durch den Wert des neuen Basisjahres geteilt.

Umbasierung

	Deutschland	**Österreich**	**Deutschland**	**Österreich**
	1980=100	**1966=100**	**1975=100**	**1975=100**
1975	82,6	163,5	100,0	100,0
1976	86,3	175,5	104,5	107,3
1977	89,3	185,1	108,1	113,2
1978	91,6	191,7	110,9	117,2
1979	95,0	198,8	115,0	121,6
1980	100,0	211,4	121,1	129,3
1981	106,3	225,8	128,7	138,1
1982	112,0	238,1	135,6	145,6
1983	115,6	246,0	140,0	150,5
1984	118,4	260,0	143,3	159,0
1985	120,9	268,3	146,4	164,1
1986	120,7	272,8	146,1	166,9

Beispiele für die Berechnung:

1976 (D): $\frac{86,3}{82,6} = 104,5$

1976 (A): $\frac{175,5}{163,5} = 107,3$

1984 (D): $\frac{118,4}{82,6} = 143,3$

Verknüpfung von Indizes

Bemerkung 58:

- Von Zeit zu Zeit wird es notwendig, Indizes mit einer neuen Reihe beginnen zu lassen, wenn z. B. ein Warenkorb umgestellt werden muss.
- Alter und neuer Index müssen in mindestens einer Periode parallel erhoben werden.
- Der alte und der neue Index können miteinander verknüpft werden, indem man eine gemeinsame Basisperiode zu "100" setzt.
- Auch eine zeitliche Vorwärts- oder Rückwärtsergänzung der Indizes ist möglich.

Beispiel 103:

	Alter Index	**Neuer Index**
	1962=100	**1970=100**
1962	100,0	
1963	103,0	
1964	105,4	
1965	109,0	
1966	112,8	
1967	114,4	
1968	116,1	
1969	119,3	
1970	123,7	100,0
1971		105,1
1972		110,7
1973		118,7
1974		126,3

Beispielsberechnungen:

Alter Index 1971: $\frac{123,7}{100} \cdot 105,1 = 130,0$

Neuer Index 1969: $\frac{119,3}{123,7} \cdot 100 = 96,4$

	Alter Index	Neuer Index
	1962=100	**1970=100**
1962	100,0	80,8
1963	103,0	83,3
1964	105,4	85,2
1965	109,0	88,1
1966	112,8	91,2
1967	114,4	92,5
1968	116,1	93,9
1969	119,3	96,4
1970	123,7	100,0
1971	130,0	105,1
1972	136,9	110,7
1973	146,8	118,7
1974	156,2	126,3

Beispielsberechnungen:

Alter Index 1971: $\frac{123,7}{100} \cdot 105,1 = 130,0$

Neuer Index 1969: $\frac{119,3}{123,7} \cdot 100 = 96,4$

Lorenz-Kurve

Die **Lorenz-Kurve** (auch Lorenzkurve) wurde von Max Otto Lorenz zur grafischen Darstellung von statistischen Verteilungen und der Veranschaulichung des Ausmaßes an Ungleichheit entwickelt. Sie wird insbesondere zur Analyse der Einkommensverteilung verwendet.

Eigenschaften der Lorenz-Kurve

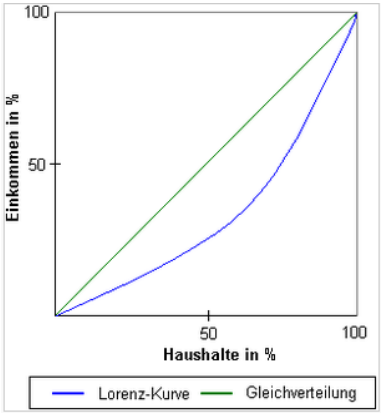
Die Lorenzkurve hat drei charakteristische Eigenschaften:

sie beginnt bei (0%|0%) und endet bei (100%|100%)

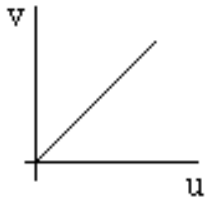
sie ist konvex

sie ist stetig

Anders als manchmal angenommen ist sie weder immer monoton steigend (Gegenbeispiel: negatives Einkommen bei einem Teil der Haushalte, aber positives Gesamteinkommen aller Haushalte) noch immer glatt (Gegenbeispiel: Ein Teil der Haushalte verdient genau 1000 € im Monat, der andere Teil genau 10000 € ohne Zwischenwerte. Dann ergibt sich ein Graph aus zwei Geraden verschiedener Steigung, und die zugehörige Funktion ist im Knickpunkt nicht differenzierbar)

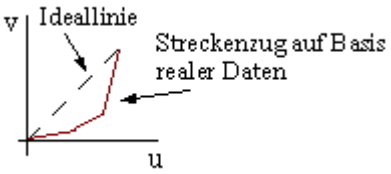


Liegt der Idealfall vor, wobei dieser sicher vom Standpunkt abhängt, also bei einer "Konzentrationsgleichverteilung", sieht die Lorenzkurve wie folgt aus:



Hier liegt der 45°-Fall, das heißt Gleichverteilung, vor.

Liegt eine Abweichung von der Gleichverteilung vor, d. h., um beim Umsatzbeispiel zu bleiben, gibt es unter den Unternehmen einen "Umsatzriesen", weicht die Lorenzkurve wie im nächsten Bild von der Ideallinie (45°) ab:



Die Konzentration befasst sich mit der Intensität, mit der sich ein Objekt auf eine vorgegebene Menge verteilt. Eine typische Aussage der Konzentrationsmessung wäre etwa: 20% der Menschen eines bestimmten Staates besitzen 90% des Vermögens. Demnach teilen sich die anderen 80% die restlichen 10%. Hier kann man von einer starken Konzentration sprechen.

Beispiel 104:

Im Rahmen einer Controlling-Analyse eines Kinos wurden die Besucherzahlen (Merkmal x) für die 5 angebotenen Spielfilme an einem Tag erfasst. Man erhielt die Tabelle.

Besucherzahlen im Kinopalast	
Filmtitel	Zahl der Besucher x
Rotkäppchen	25
Verliebt ins Abendrot	75
Leif Erikson	125
Söhne der Alhambra	250
Galaxy-Fighter	525
Summe	1000

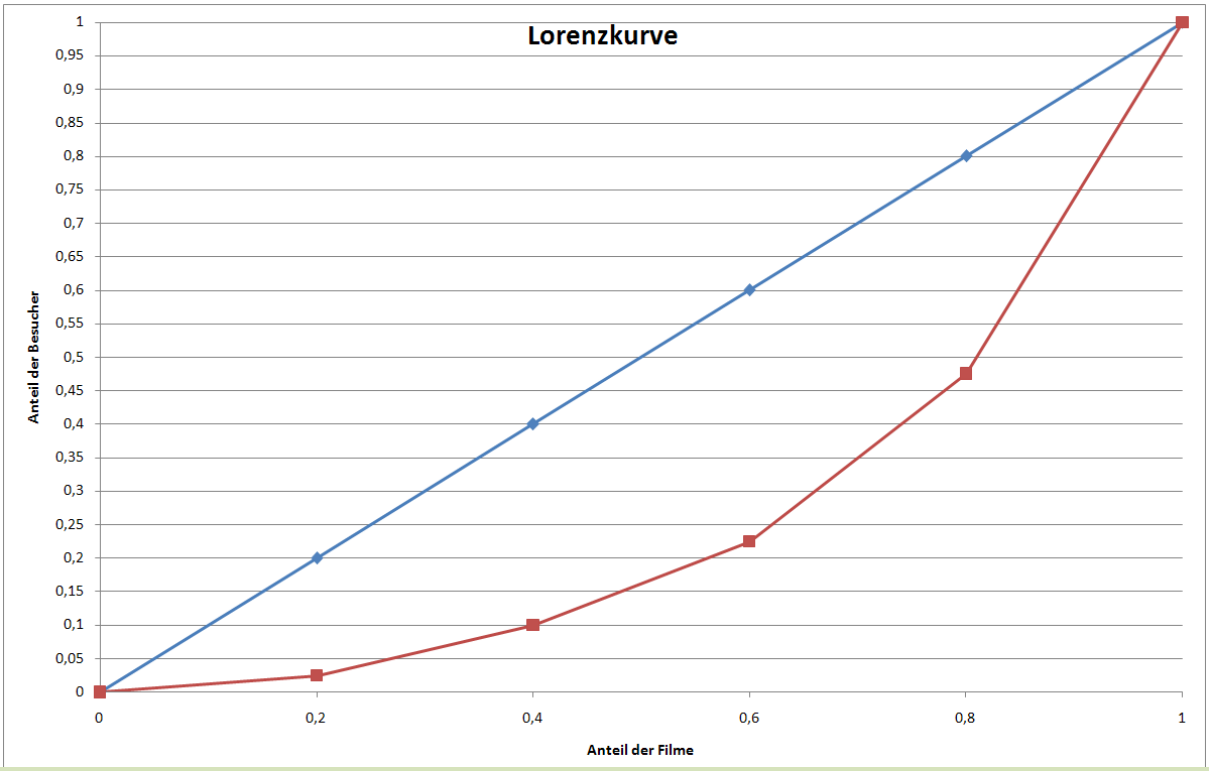
Es gibt verschiedene Verfahren zur Konzentrationsmessung. Man kann die Konzentration grafisch darstellen oder Kennwerte berechnen. Die Merkmalsbeträge x müssen aufsteigend geordnet vorliegen.

Dies ist in unserem Fall gegeben.

Die für die Lorenzkurve benötigten Zwischenwerte welche in der folgenden Tabelle aufgeführt.

Besucherzahlen im Kinopalast						
Filmtitel	Zahl der Besucher x	q_i	p_i	p_i^*	S_i	S_i^*
-	0	0	0	0	0	0,0
Rotkäppchen	25	25	0,025	0,025	1	0,2
Verliebt ins Abendrot	75	100	0,075	0,100	2	0,4
Leif Erikson	125	225	0,125	0,225	3	6,0
Söhne der Alhambra	250	475	0,250	0,475	4	0,8
Galaxy-Fighter	525	1000	0,525	1,000	5	1,0
Summe	1000		1			
q_i	kummulierte absolute Häufigkeit					
p_i	relative Häufigkeit					
p_i^*	Kummulierte relative Häufigkeit					
S_i	absolute Summenhäufigkeiten als Zahl der Filme					
S_i^*	relative Summenhäufigkeit der Filme					

Daraus ergibt sich die folgende Konzentrationskurve:



Die Lorenzkurve ist ein grafisches Maß für das Ausmaß einer Konzentration. Je weiter die Kurve "durchhängt", desto größer ist die Konzentration.

Ginikoeffizient

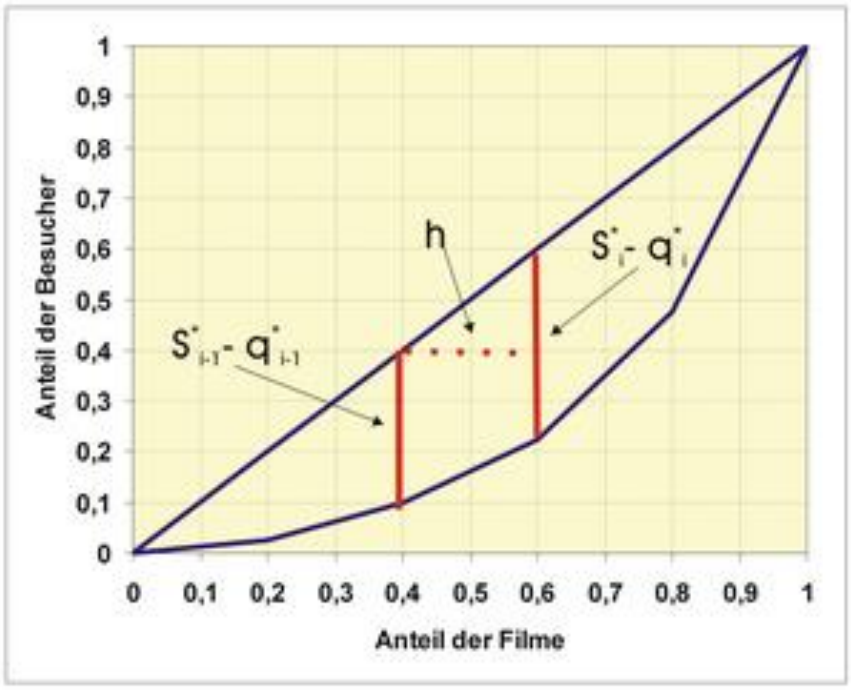
Der **Ginikoeffizient** oder auch Gini-Index ist ein statistisches Maß für Verteilungsgleichheit, entwickelt vom italienischen Statistiker Corrado Gini. Er wird besonders in der Wohlfahrtsökonomie verwendet.

Der Wert kann beliebige Größen zwischen 0 und 1 (bzw. 0 und 100 Prozent) annehmen. Je näher an 1 der Ginikoeffizient ist, desto größer ist die Ungleichheit (zum Beispiel einer Einkommensverteilung).

Als Ginikoeffizient G wird bezeichnet der Anteil der Fläche, die durch die Winkelhalbierende und die Lorenzkurve gebildet wird, an der Gesamtfläche unter der Winkelhalbierenden. Wenn vollkommene Konzentration besteht, ist die Fläche über der Lorenzkurve deckungsgleich mit dem Dreieck unter der Winkelhalbierenden. G ist dann 1. Bei fehlender Konzentration ist dann G=0.

Verbindet man die Punkte auf der Lorenzkurve mit den entsprechenden Punkten auf der Winkelhalbierenden, wird klar, dass wir es mit n vielen Trapezen zu tun haben, deren Flächen wir einzeln bestimmen und dann aufsummieren. Die Fläche eines Trapezes, wie in der Grafik angegeben, ermittelt man als

$$F = \frac{1}{2} \cdot (a + c) \cdot h$$



Wir wollen die Fläche F3 des Trapezes zwischen den Abszissenwerten (x-Achse) 0,4 und 0,6 ermitteln. Man sieht, dass das Trapez im Vergleich zur obigen Grafik gekippt vorliegt. Die Höhe h ist also die Differenz

$S^*_3 - S^*_3 = 0{,}6 - 0{,}4 = 0{,}2$.

Wir fassen a als linke Senkrechte von F3 als a auf: Dann ist

$a = 0{,}4 - 0{,}1 = 0{,}3$.

Entsprechend beträgt die rechte Seite c

$c = 0{,}6 - 0{,}225 = 0{,}375$

und wir erhalten als Fläche

$$F_2 = (0,3 + 0,375) \cdot 0,5 \cdot 0,2 = 0,0675.$$

Allgemein: Die obige Fläche ergibt sich dann als

$$\sum_{i=1}^{n} (S*_i - S*_{i-1}) \cdot \frac{1}{2}((S*_i - q*_i) + (S*_{i-1} - q*_{i-1}))$$

Es folgt beispielhaft die Berechnung des Gini in der Tabelle. Mit Tabellenkalkulation kann der Ginikoeffizient leicht ermittelt werden. Wir erhalten schließlich für den Ginikoeffizienten

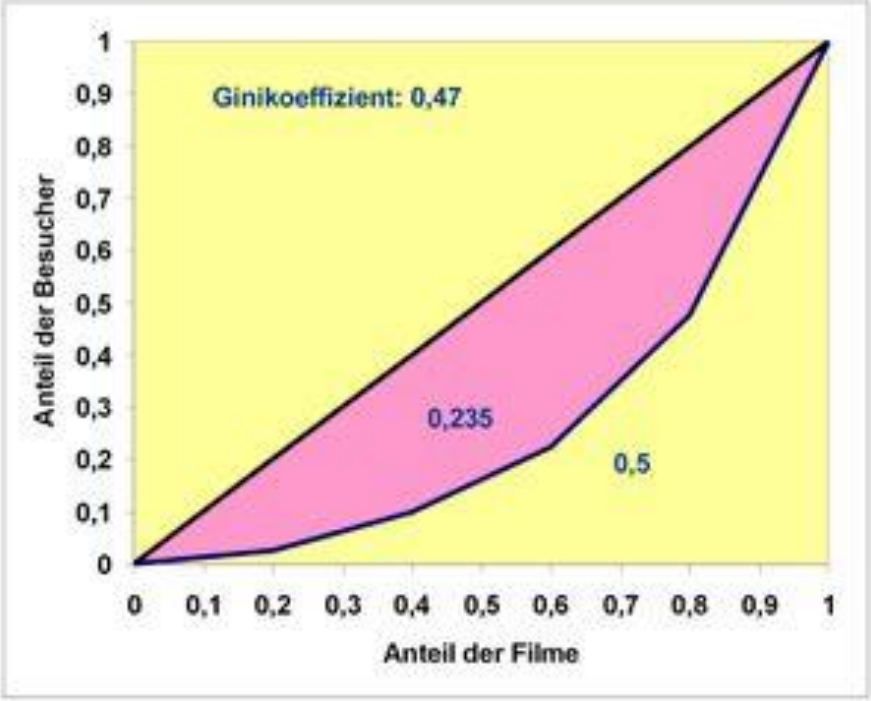
Gini-Koeffizient									
	Zahl der Besucher x	p_i	p_i^*	S_i^*	$h_i=s_i^*-s_{i-1}^*$	$a_i=s^*_i-p^*_i$	$c_i=s^*_{i-1}-p^*_{i-1}$	$0{,}5(a_i+c_i)$	$0{,}5(a_i+c_i)h_i$
-			0	0					
1	25	0,025	0,025	0,2	0,2	0,175	0,000	0,0875	0,0175
2	75	0,075	0,100	0,4	0,2	0,300	0,175	0,2375	0,0475
3	125	0,125	0,225	0,6	0,2	0,375	0,300	0,3375	0,0675
4	250	0,25	0,475	0,8	0,2	0,325	0,375	0,35	0,07
5	525	0,525	1,000	1,0	0,2	0,000	0,325	0,1625	0,0325
Summe	1000								**0,235**

Den **Gini-Ungleichverteilungskoeffizient** (GUK) enthält man durch Auswertung einer Lorenz-Kurve.

$$GUK = \frac{A - B}{A} = \frac{0{,}5 - B}{0{,}5}$$

Graphisch betrachtet ist der Ginikoeffizient das Verhältnis der Fläche zwischen Gleichverteilungslinie und Lorenzkurve (A-B) zur Fläche unterhalb der Gleichverteilungslinie (A).

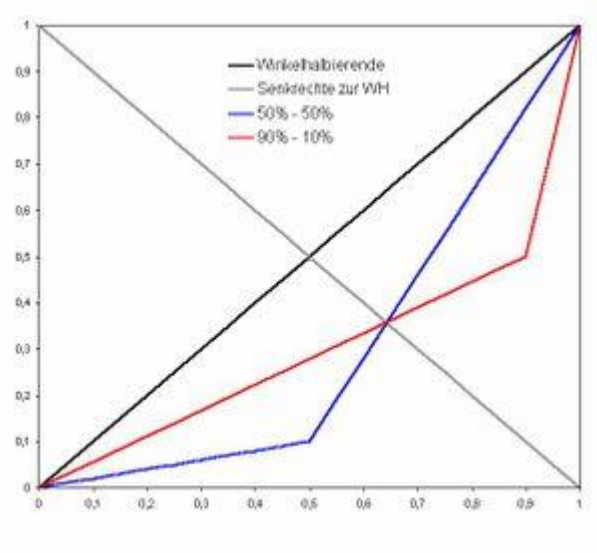
$$G = \frac{0,235}{0,5} = 0,47$$



Kritik

Der Gini-Koeffizient ist ein sehr einfaches Maß zur Berechnung einer (Un-)Gleichheitsverteilung. Daher liefert er auch Ergebnisse, die zu Missinterpretationen führen können. Grundsätzlich gibt es zu jeder Lorenzkurve eine andere Lorenzkurve mit exakt dem gleichen Gini-Wert. Diese erhält man durch Spiegelung der ursprünglichen Lorenzkurve an der Senkrechten zur Winkelhalbierenden, die durch die Punkte (0,1) und (1,0) verläuft.

Ein Beispiel soll die Kritik verdeutlichen: In einer Volkswirtschaft befindet sich 10% des Eigentums in den Händen von 50% der Bevölkerung, die restlichen 50% besitzen die restlichen 90% (jeweils in den Gruppen gleichverteilt). In einer anderen Volkswirtschaft besitzen 90% der Bevölkerung 50% des Eigentums, während eine Minderheit von 10% die andere Hälfte des Eigentums beansprucht. Die beiden Lorenzkurven sind in der Abbildung verdeutlicht.

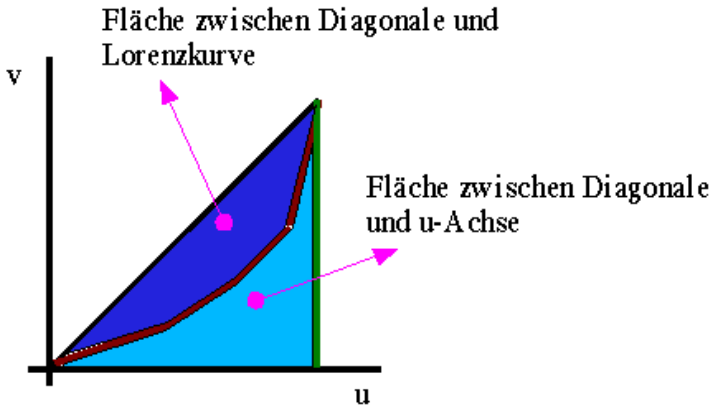


Für beide Kurven ergibt sich ein Gini-Koeffizient von 0.3. Dies liegt daran, dass ein Repräsentant des reicheren Teils der Bevölkerung in beiden Fällen das 9-fache Eigentum eines Repräsentanten des ärmeren Teils der Bevölkerung besitzt.

Interpretation

Die Stärke der Konzentration, also die Abweichung von der Gleichverteilung, drückt sich durch die Entfernung von der 45°-Linie aus.

Der Gini-Koeffizient G ist das Verhältnis aus der Fläche zwischen der 45°-Linie (Diagonale) und Lorenzkurve und der Gesamtfläche (= Fläche zwischen Diagonale und u-Achse):



$$G = \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Fläche zwischen Diagonale und u-Achse}}$$

Da der Gini-Koeffizient nur im Zusammenhang mit der Lorenzkurve zu betrachten ist, gelten hier die gleichen Voraussetzungen für die Merkmalsausprägungen.

Der Gini-Koeffizient beträgt bei Gleichverteilung 0 und bei maximaler Konzentration (n-1)/n:

$G_{min} = 0$ Gleichverteilung

$G_{max} = \frac{n-1}{n}$ maximale Konzentration

Aus G_{max} ist ersichtlich, dass die maximale Ausprägung des Koeffizienten von der Merkmalsanzahl abhängt. Um diese Abhängigkeit zu vermeiden, wird der G-Koeffizient normiert:

$$G^* = \frac{G}{G_{max}} = \frac{n}{n-1} \cdot G$$

G^* Normierter G-Koeffizient

$G^* \in [0,1]$

Der G-Koeffizient ist *immer* mit der Lorenzkurve zu betrachten.

Regressionsanalyse und Korrelationsanalyse

Betrachtung von Zusammenhängen, also von Ursache -> Wirkung.

Regression: Besteht ein Zusammenhang (positiv oder negativ)?

Korrelation: Wie stark ist der Zusammenhang?

Beispiele:

Werbung -> Umsatz

Investition -> Gewinn

Zinsen -> Investition

Regressionsrechnung

Die einfache lineare Regressionsanalyse sucht nach einer linearen Gleichung, die den Zusammenhang zwischen x_i und y_i zum Ausdruck bringt.

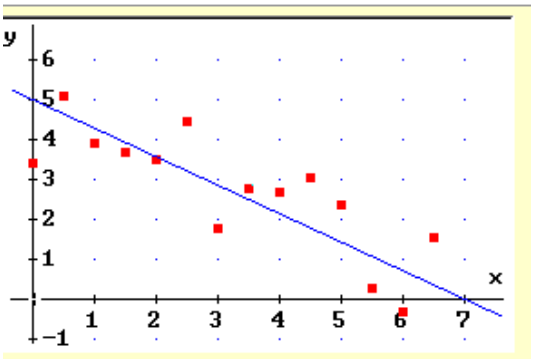
Voraussetzung: x_i und y_i sind mindestens intervall-, d.h. metrisch skaliert.

Begriffe:

X: exogene Variable = Einflussfaktor = erklärende Variable = Regressor = unabhängige Variable

Y: endogene Variable = Zielvariable = abhängige Variable = Regressand

Schätzgleichung: Gleichung, die exakt die Stichprobe beschreibt



Folgende Probleme lassen sich mit der linearen Einfachregression lösen:

1. Man will wissen, welche Grundrichtung der Beziehung zwischen X und Y besteht. Wie groß ist die prop. Veränderung in Y, wenn X_i um eine Einheit erhöht/vermindert wird?
   Bsp.: Pro Jahr zusätzlicher Schulbildung erhöht sich das Einkommen um *b* Einheiten.
2. Man will einen Schätzwert von Y für einen X-Wert ermitteln, der außerhalb der Reihe der Beobachtungswerte liegt (→ Extrapolation). Man prognostiziert also.
3. Man will einen Schätzwert von Y wissen, wobei der X-Wert zwischen zwei bekannten X-Werten liegt, selbst aber nicht realisiert ist (→ Interpolation).

Bei Zeitreihen wird ein Entwicklungstrend berechnet und als Prognose in die Zukunft fortgeschrieben

Es gibt zwei verschiedene Problemansätze:

Die Frage nach der

a) mathematischen Art der Beziehung zwischen *x* und *y* liefert die *Regressionsgleichung*

b) Stärke der Beziehung liefert den Korrelationskoeffizienten *r* (Bravais-Pearson)

Das Modell der einfachen linearen Regression

Ein reales Problem kann in die folgende angemessene formale Form übersetzt werden. Zwischen X und Y besteht ein Zusammenhang, der durch die Gleichung

$$\hat{y}_i = \alpha + \beta x_i + u_1$$

zum Ausdruck gebracht werden kann.

Jeder Wert von Y_i lässt sich aus zwei Komponenten zusammengesetzt auffassen:

$\alpha + \beta x_i$: Wert, den y_i annehmen würde, falls der Zusammenhang zwischen X und Y streng deterministisch (sprich linear) wäre.

u_i : Wert, um den y_i von seiner deterministischen Komponente $\alpha+\beta x_i$ abweicht (Abweichung zwischen dem realen Wert und der später zu berechnenden Regressionsgerade), u_i ist der Wert der Störgröße u_i. u_i spezifiziert den stochastischen Teil des Zusammenhangs.

u_i lässt sich als Zufallsvariable auffassen, da oft nicht angegeben werden kann, welchen Wert u_i bei vorgegebenem Wert x_i annimmt. u_i lässt sich aber auch als Störvariable auffassen, da die u_i die Abweichungen von einer linearen Regressionsfunktion darstellen.

Die Regressionsgleichung

Die Regressionsgleichung der Stichprobe ergibt sich durch die Gleichung:

$$y_i = a_i + bx_i + d_i,$$

wobei d_i die Summe der Schätzfehler, d.h. die Summe der Differenzen zwischen y_i und $a + bx_i$, ist. Der **Schätzfehler** heißt auch **Residuum**, die Summe Residuen.

Diese Gleichung zur exakten Beschreibung ist (leider) nicht linear, daher benötigt man als exakte Beschreibung die Gleichung der Regressionsgerade $\hat{y}_i$:

Die Gleichung der Schätzgerade $\hat{y}_i$ lautet:

$$\hat{y}_i = a_i + bx_i$$

Um die beste Regressionsgerade zu bestimmen

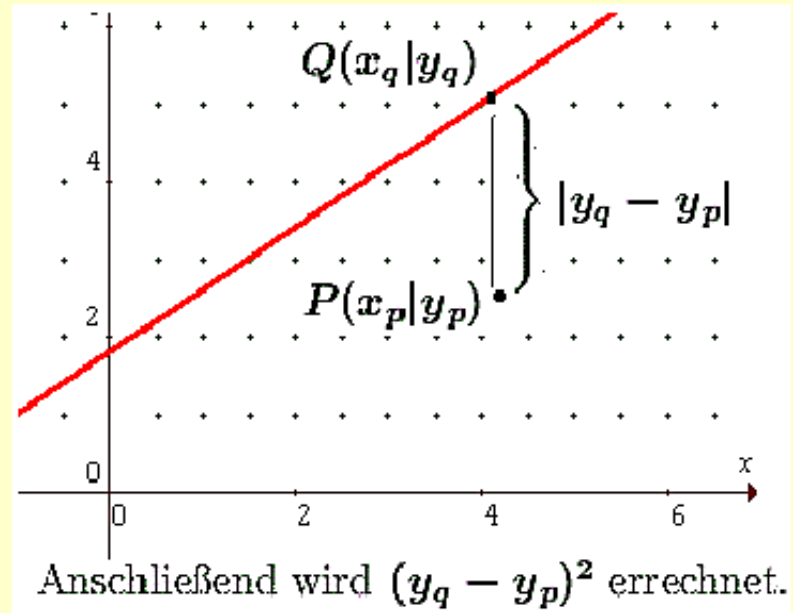
a) soll die Summe der Schätzfehler 0 sein, d.h. die einzelnen Fehler sollen sich aufheben, d.h. die Gerade muss durch $\bar{x}$ und $\bar{y}$ laufen

b) die Zahl der Schätzfehler muss minimal sein

Methode der kleinsten Quadrate für eine einfache Regressionsgleichung

Um die Parameter a und b einer Regressionsgraden so zu bestimmen, dass die Gerade den beobachteten Wertepaaren optimal angepasst ist, muss die Summe der quadrierten Abweichungen der beobachteten Y_i von den rechnerischen Y_t ein Minimum ergeben. D.h. die Regressionsgerade ist dann optimal berechnet, wenn die Summe der Abweichungsquadrate minimal ist.

$$z(a,b) = \sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y})^2 \rightarrow \text{Minimum}$$



Durch partielle Ableitung und Nullsetzen dieser Ableitungen ergeben sich die Normalgleichungen zur Bestimmung der Koeffizienten einer linearen Kleinste-Quadrate-Reressionsfunktion. Löst man das System der Normalgleichungen nach a und b auf, erhält man die Regressionskoeffizienten *a* und *b:*

Für eine **einfache Regressionsgleichung** ergeben sich die Regressionskoeffizienten:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Bedeutung der Regressionsfunktionsbestandteile

Eine univariate Regressionsfunktion hat die Funktion $y_i + a + bx_i$

Dabei bedeuten:

x_i Werte auf der X-Achse

y_i Werte auf der Regressionsfunktion

Man nennt diese Werte auch zu erwartende oder theoretische Werte, weil diese Y-Werte in Abhängigkeit von Veränderungen der Variablen X zu erwarten wären,

wenn die Regressionslinie den Zusammenhang zwischen X und Y korrekt widerspiegelt.

Insoweit kommt in der Regressionsfunktion selbst eine Hypothese über den vermuteten Zusammenhang zwischen X und Y zum Ausdruck.

a Ordinatenabschnitt der linearen Funktion

b Steigung (= Tangens des Steigungswinkels) der Funktion

Die Koeffizienten a und b spezifizieren den deterministischen Teil des Zusammenhangs und stellen die wahren Parameter für die gesamte Population her.

Beispiel 105:

x_i	y_i	$x_i - \hat{x}$	$y_i - \hat{y}$	$(x_i - \hat{x})*(y_i - \hat{y})$	$(x_i - \hat{x})^2$	$(y_i - \hat{y})^2$
1,9	5,1	-6,1	-2,8	17,08	37,21	7,84
3,0	5,6	-5,0	-2,3	11,50	25,00	5,29
4,2	6,1	-3,8	-1,8	6,84	14,44	3,24
5,5	6,3	-2,5	-1,6	4,00	6,25	2,56
7,0	7,0	-1,0	-0,9	0,90	1,00	0,81
8,9	8,2	0,9	0,3	0,27	0,81	0,09
10,0	9,0	2,0	1,1	2,20	4,00	1,21
11,5	9,8	3,5	1,9	6,65	12,25	3,61
13,0	10,6	5,0	2,7	13,50	25,00	7,29
15,0	11,3	7,0	3,4	23,80	49,00	11,56
$\bar{x} = 8,0$	$\bar{y} = 7,9$			$\sigma_{xy} =$	$\sigma_{xx} =$	$\sigma_{yy} =$
80,0	79,0			86,74	174,96	43,5

Lösung:

$$b = \frac{\sigma_{xy}}{\sigma_{xx}} = \frac{86,74}{174,96} = 0,4958$$

$$a = \bar{y} - b * \bar{x} = 7,9 - 0,4958 * 8$$

$$a = 3,9336$$

$$y = a + b \cdot x$$

$$\bar{y} = 3,9336 + 0,4958 \cdot x$$

$$r = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx} \cdot \sigma_{yy}}} = \frac{86,74}{\sqrt{174,96 \cdot 43,50}}$$

$$= \frac{7523,8276}{7610,7600} = 0,9942$$

Korrelationskoeffizient nach Bravais-Pearson

Die Korrelationsrechnung dient dazu, die Stärke des Zusammenhangs zwischen zwei Untersuchungsvariablen in einer einzigen statistischen Maßzahl zum Ausdruck zu bringen. *r* ist eine dimensionslose Größe

Voraussetzung für die Anwendung des Korrelationskoeffizienten von Bravais-Pearson sind mindestens *intervallskalierte Daten.*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad -1 \le r \le 1$$

$\bar{x}$	arithmetisches Mittel von X
$\bar{y}$	arithmetisches Mittel von Y
n	Anzahl von (yi, xi); Anzahl der statistischen Einheiten

Interpretation von r

Der Korrelationskoeffizient von Bravais-Pearson nimmt nur Werte zwischen -1 und +1 an. Wertebereich von -1 bis +1:

r=-1 maximaler reziproker Zusammenhang, d.h. mit sehr hoher Wahrscheinlichkeit nehmen die Y-Werte tendenziell ab, wenn die Werte der Variablen X zunehmen

r=0 kein Zusammenhang zwischen X und Y

r=+1 maximaler gleichgerichteter Zusammenhang, d.h. mit sehr hoher Wahrscheinlichkeit nehmen die Werte der Variablen Y tendenziell zu, wenn die X-Werte zunehmen.

Anmerkungen zum Korrelationskoeffizienten *r*

- in der Praxis taucht ein Wert für r größer 0,5 nur selten auf, man betrachtet ein *r* zwischen 0,3 und 0,5 als ein Indiz für einen starken Zusammenhang
- je größer die Zahl der Merkmalsträger, desto aussagekräftiger ist *r*
- die Treffsicherheit von *Prognosen* ist umso höher, je größer *r* ist, d.h. je stärker der Zusammenhang zwischen zwei Variablen X und Y ist und je größer N ist.
- die Interpretation des Korrelationskoeffizienten muss immer auf dem Hintergrund einer *linearen Regressionsfunktion* erfolgen. Wäre in einem konkreten Fall eine nichtlineare Funktion angemessen, dann könnte sich beispielsweise ein r-Wert nahe bei 0 ergeben, weil gleichwohl eine lineare Funktion unterstellt wird.
- Die Prüfung, ob eine nichtlineare Funktion zugrunde gelegt werden muss, kann z.B. graphisch oder durch eine Clusteranalyse erfolgen.

Hypothesentest

Einführung

Wer Entscheidungen zu treffen hat, weiß oft erst im Nachhinein ob seine Entscheidung richtig war.

Die Unsicherheit eine Entscheidung zu treffen, beinhaltet immer eine gewisse Fehlerwahrscheinlichkeit.

Der Hypothesentest gibt uns eine Richtlinie für die Wahl einer Alternativentscheidung.

Wir treffen unsere Entscheidung auf der Grundlage dessen, was wir für richtig erachten.

Das nennen wir die **Nullhypothese**.

Eine Alternativentscheidung nennen wir **alternative Hypothese**.

Dieses wird nach der jeweiligen Aufgabenstellung festgelegt.

Das Testen von Hypothesen ist immer ein Vorgang, den man in mehrere Schritte unterteilen kann:

- Formulierung der Nullhypothese H_0 und der Alternativhypothese H_1
- Festlegung des Signifikanzniveaus
- Bestimmung des Annahme- und Ablehnungsbereichs der Nullhypothese
- Ziehung der Stichprobe
- Treffen der Testentscheidung und Interpretation:

Liegt das Ergebnis der Stichprobe innerhalb des Annahmebereichs, wird H_0 angenommen, anderenfalls abgelehnt.

Für das richtige Aufstellen der Hypothesen gibt es ein paar Regeln:

- Was ich zeigen oder beweisen will, gehört in die Alternativhypothese
- Das Gleichheitszeichen gehört immer in die Nullhypothese
- Beim Aufstellen der Nullhypothese geht man davon aus, "Alles bleibt beim alten, nichts hat sich geändert"

Die Annahme der Nullhypothese führt immer zur Ablehnung der Alternativhypothese, ist aber kein Beweis dafür, dass die Nullhypothese stimmt.

Die Ablehnung der Nullhypothese führt zur Annahme der Alternativhypothese.

Beispiel 106:

Die Befragung aller Studenten einer Fachhochschule ergab im letzten Jahr, dass 10% der befragten Studenten mit dem Mensaessen unzufrieden waren. Es wird vermutet, dass die Unzufriedenheit im laufenden Semester sogar noch zugenommen hat. Das Mensa-Organisationsteam steht nun vor der Entscheidung, ob Maßnahmen zur Verbesserung der Qualität des Essens ergriffen werden müssen. Um eine Entscheidung treffen zu können, werden in einer Umfrage 100 Studenten befragt.

a) Sind mehr als 10 Studenten mit dem Essen unzufrieden, so soll die Qualität des Essens verbessert werden. In der ersten Umfrage erklärten 12 Studenten, mit dem Mensaessen unzufrieden zu sein.

b) Das Mensateam ist sich der Zufälligkeit von Stichprobenergebnissen bewusst und lässt in einer 2. Umfrage wieder 100 Studenten befragen. Dabei gibt sich das Team mit einer Sicherheit von 95% mit dem Befragungsergebnis zufrieden. Es erklären 13 Studenten, mit dem Essen nicht zufrieden zu sein.

Wie wird das Team in beiden Fällen entscheiden?

Die Entscheidung soll über einen Hypothesentest gefunden werden.

Lösung:

Zu zeigen ist p > 0,1

Das heißt: Mehr als 10% aller Studenten sind mit dem Essen unzufrieden.

Damit liegen die Null- und die Alternativhypothese fest.

$H_0: p \leq 0{,}1 \; und \; H_1: p > 0{,}1$

Es wird die H_0-Hypothese getestet. Sie wird angenommen, bzw. beibehalten, wenn die Zahl der Studenten im Annahmebereich und sie wird abgelehnt, wenn die Zahl der Studenten im Ablehnungsgsbereich liegt.

$H_0: p \leq 0{,}1$

Annahmebereich $\quad A = \{0,1,2, \ldots\ldots,10\}$

Ablehnungsbereich $\quad \overline{A} = \{11,12,13, \ldots\ldots,100\}$

Beim 1. Test erklärten 12 Studenten, mit dem Essen unzufrieden zu sein.

Die H_0- Hypothese würde damit abgelehnt, aber dies könnte wegen der Zufälligkeit der Stichprobe falsch sein, wenn der tatsächliche Anteil der unzufriedenen in der Grundgesamtheit (Menge aller Studenten der Fachhochschule) tatsächlich 10% ist. Man begeht also bei der Ablehnung der H_0-Hypothese mit einer gewissen Wahrscheinlichkeit einen Fehler.

Dieser Fehler, auch als **Irrtumswahrscheinlichkeit** des Tests bezeichnet, berechnet sich aus der Ablehnungswahrscheinlichkeit.

$P(X \geq 11) = 1 - P(X \leq 10) = 1 - 0{,}583 = 0{,}417$ $(siehe\ Tabelle)$

Tabelle 1:
Kumulierte Binomialverteilung für n = 100 und p = 0,1

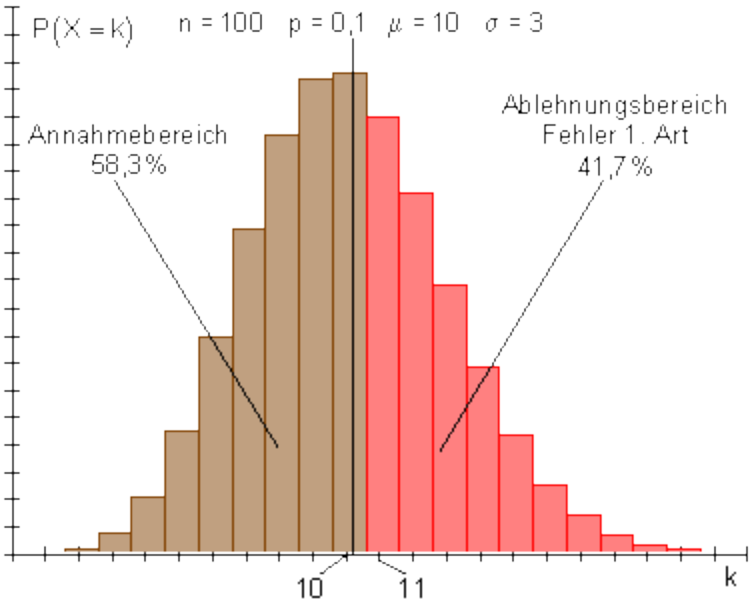
k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)
0	0,000	4	0,024	8	0,321	12	0,802	16	0,979	20	0,999
1	0,000	5	0,058	9	0,451	13	0,876	17	0,990	21	1,000
2	0,002	6	0,117	10	0,583	14	0,927	18	0,995	22	1,000
3	0,008	7	0,206	11	0,703	15	0,960	19	0,998	23	1,000

Das heißt, mindestens 11 Studenten sind mit dem Essen unzufrieden.

Man muss also sagen:

"Unter der Annahme, dass tatsächlich 10% aller Studenten unzufrieden sind, kommt es bei der angegebenen Befragung mit einer Wahrscheinlichkeit von ca. 41,7% zu einem solchen Ergebnis und damit zu fälschlichen Ablehnung der Nullhypothese".

Eine Verteilungsfunktion soll das verdeutlichen:



b) Beim 2. Test wird ein Fehler von 5% zugestanden (Sicherheit von 95%).

Dadurch ändern sich Annahme- und Ablehnungsbereich.

$P(x \geq k) \leq 0{,}05$

$1 - P(X \leq k-1) \leq 0{,}05 \quad |-1$

$-P(X \leq k-1) \leq -0{,}95 \quad |\cdot(-1)$

$P(X \leq k-1) \geq 0{,}95$

Tabelle 1:
Kumulierte Binomialverteilung für n = 100 und p = 0,1

k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)
0	0,000	4	0,024	8	0,321	12	0,802	16	0,979	20	0,999
1	0,000	5	0,058	9	0,451	13	0,876	17	0,990	21	1,000
2	0,002	6	0,117	10	0,583	14	0,927	18	0,995	22	1,000
3	0,008	7	0,206	11	0,703	15	0,960	19	0,998	23	1,000

$P(x \leq 15) \approx 0{,}96 > 0{,}95$

$k - 1 = 15$

$k = 16$

$k - 1 = 15$

$k = 16$

Annahmebereich $\quad A = \{0,1,2, \ldots\ldots,15\}$

Ablehnungsbereich $\quad \overline{A} = \{16,17,18, \ldots\ldots,100\}$

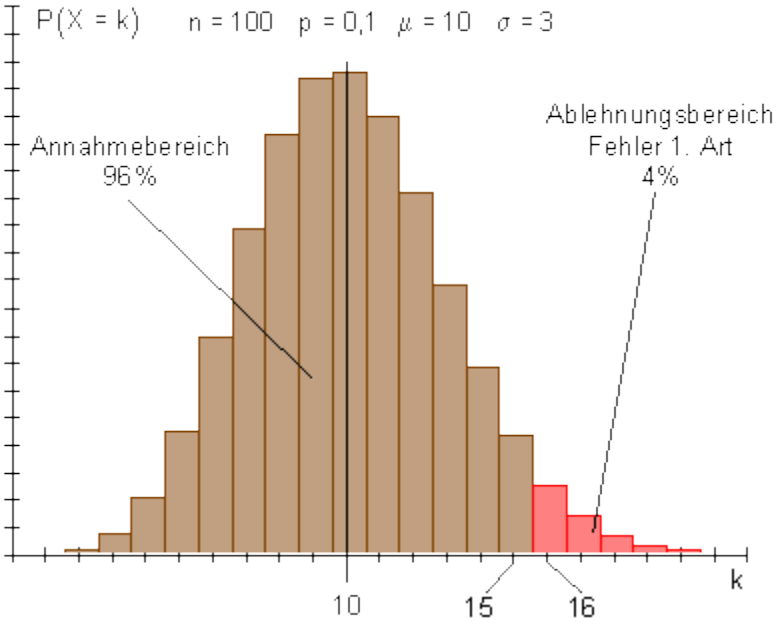
Während der Annahmebereich größer wird, wird der Ablehnungsbereich kleiner.

Da das 2. Testergebnis mit 13 unzufriedenen Studenten in den neuen Annahmebereich fällt, wird die Nullhypothese angenommen.

Das Mensateam sieht keine Veranlassung, das Essen zu verbessern.

Erst wenn mehr als 15 Studenten mit dem Essen nicht zufrieden wären, würde die Nullhypothese abgelehnt, die Alternativhypothese angenommen und das Essen verbessert werden.

Eine Verteilungsfunktion soll das verdeutlichen:



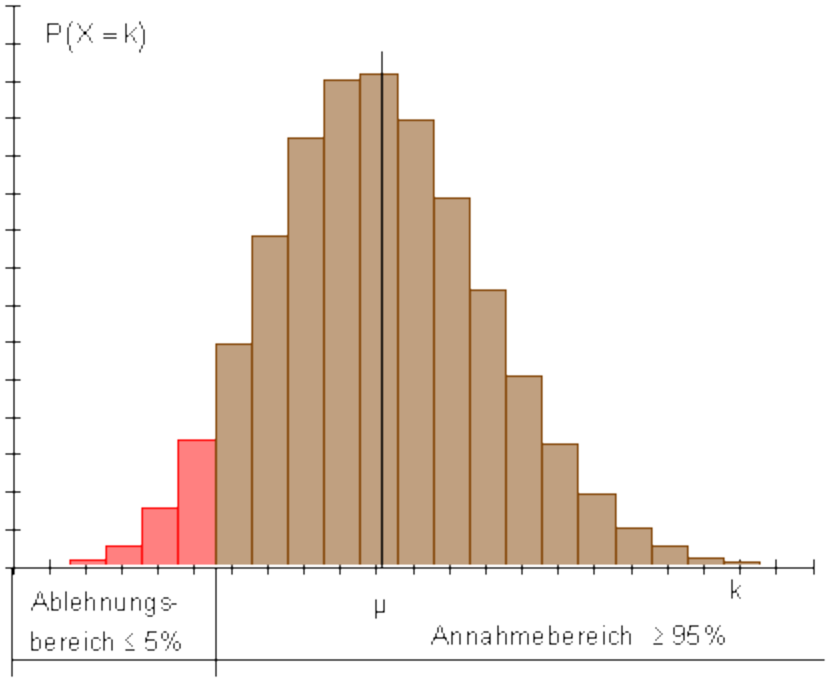
Die Wahrscheinlichkeit beim Testen einen gewissen Fehler zuzulassen, heißt Irrtumswahrscheinlichkeit.

Sie wird in der Regel vor Durchführung des Zufallsexperimentes festgelegt. Dabei sind 1% und 5% übliche Werte.

Sie ist die größte Wahrscheinlichkeit für eine irrtümliche Ablehnung der H_0- Hypothese. Statt Irrtumswahrscheinlichkeit sagt man auch **Signifikanzniveau**.

Ein Test, bei dem der Ablehnungsbereich unterhalb, also links vom Erwartungswert liegt, heißt Linksseitiger Hypothesentest. Vielfach wird dieses Verfahren dann benutzt, wenn die Alternativhypothese $H_1: p < a$ lautet, und $H_0: p \geq a$ zu testen ist.

Linksseitiger Hypothesentest:



Fehler beim Testen von Hypothesen

Die Entscheidung, die aufgrund eines Versuchsergebnisses (Test, Umfrage, ...) getroffen wird kann falsch sein. Die zu testende Hypothese H_0 (höchstens 10% aller Studenten sind mit dem Essen unzufrieden) kann wahr oder falsch sein.

Man unterscheidet zwei Arten von Fehlern:

Fehler 1. Art: Die Nullhypothese wird verworfen, obwohl sie richtig ist.

Fehler 2. Art: Die Nullhypothese wird angenommen, obwohl sie falsch ist.

Der Fehler 2. Art lässt sich nur berechnen, wenn man für die Alternativhypothese eine andere Wahrscheinlichkeit, als für H_0 annimmt.

Beispiel 107:

Ein Babybasar verkauft gebrauchte Kinderschuhe. Etwa 60% der Schuhe befinden sich in einem einwandfreien Zustand. Der Rest weist kleine Schäden auf.

Ein neuer Lieferant behauptet, er könne gebrauchte Kinderschuhe liefern, von denen sich etwa 80% in einem einwandfreien Zustand befinden.

Der Ladeninhaber möchte keine falsche Kaufentscheidung treffen und will die Behauptung des Lieferanten überprüfen.

Dazu testet er 20 Paar Kinderschuhe aus dem Sortiment des Anbieters.

Fall I:

Angenommen, die Behauptung des Lieferanten ist richtig, d.h. die Wahrscheinlichkeit für einwandfreie Schuhe ist p = 0,8.

Der Ladeninhaber bezweifelt die Aussage des Lieferanten, er geht von p < 0,8 aus.

Er stellt folgende Hypothesen auf:

Nullhypothese H_0: p ≥ 0,8 und die Alternativhypothese H_1: p < 0,8

Im Versuch mit n = 20 Paar Schuhen erwartet man

$E(x) = n \cdot p = 20 \cdot 0{,}8 = 16$

einwandfreie Paare.

Wenn mindestens 16 Paar Schuhe einwandfrei sind, dann spricht das für die Behauptung des Lieferanten, dann soll H_0 angenommen werden.

Zufällig kann es auch zu weniger als 16 Paar einwandfreien Schuhen kommen, obwohl p = 0,8 ist.

Die Nullhypothese soll abgelehnt werden, wenn weniger als 16 Paar Schuhe einwandfrei sind.

Tabelle 3:
Kumulierte Binomialverteilung für n = 20 und p = 0,8

k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)
3	0,000	6	0,000	9	0,001	12	0,032	15	0,370	18	0,931
4	0,000	7	0,000	10	0,003	13	0,087	16	0,589	19	0,988
5	0,000	8	0,000	11	0,010	14	0,196	17	0,794	20	1,000

$P(X \leq 15) = 0{,}370$

Mit welcher Wahrscheinlichkeit ist dies der Fall?

Mit einer Wahrscheinlichkeit von 37% kann es vorkommen, dass bei dem Test weniger als 16 Paar Schuhe einwandfrei sind, obwohl 80% der Schuhe einwandfrei sind. Man würde also mit einer Wahrscheinlichkeit von 37% irrtümlicher Weise die Nullhypothese ablehnen.

Fall II:

Angenommen, die Schuhe des Lieferanten sind auch nur zu 60% einwandfrei.

Mit dieser Annahme stellt der Ladenbesitzer folgende Hypothesen auf:

Nullhypothese H0: p ≤ 0,6 und die Alternativhypothese H1: p > 0,6

Im Versuch mit n = 20 Paar Schuhen erwartet man

$E(x) = n \cdot p = 20 \cdot 0{,}6 = 12$

einwandfreie Paare.

Wenn mehr als 12 einwandfreie Paar Schuhe gefunden werden, spricht das eher gegen die Vermutung des Ladenbesitzers (p = 0,6).

Zufällig kann es aber auch zu mehr als 12 einwandfreien Paaren kommen, obwohl p = 0,6 ist.

Die Nullhypothese soll abgelehnt werden, wenn mehr als 12 Paar Schuhe einwandfrei sind.

Mit welcher Wahrscheinlichkeit ist dies der Fall?

Tabelle 2:
Kumulierte Binomialverteilung für n = 20 und p = 0,6

k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)
3	0,000	6	0,006	9	0,128	12	0,584	15	0,949	18	0,999
4	0,000	7	0,021	10	0,245	13	0,750	16	0,984	19	1,000
5	0,002	8	0,057	11	0,404	14	0,874	17	0,996	20	1,000

$P(X > 12) = P(X \leq 20) - P(x \leq 12) = 1 - 0{,}584 = 0{,}416$

Mit einer Wahrscheinlichkeit von 41,6% kann es vorkommen, dass bei dem Test mehr als 12 Paar Schuhe einwandfrei sind, obwohl nur 60% der Schuhe einwandfrei sind. Man würde also mit einer Wahrscheinlichkeit von 41,6% irrtümlicher Weise die Nullhypothese ablehnen.

In beiden Fällen ist die Wahrscheinlichkeit dafür, eine Fehlentscheidung zu treffen ziemlich groß (Fall I 37%, Fall II 41,6%).

Bevor der Test durchgeführt wird, ist es sinnvoll sich dafür zu entscheiden, bei welcher Anzahl von einwandfreien Schuhen man p = 0,8 oder p = 0,6 für richtig halten will. Eine solche Entscheidung ist willkürlich. Dabei sollte man nicht zu nah am Erwartungswert liegen, damit die Wahrscheinlichkeit für eine Fehlentscheidung nicht zu groß wird.

Es wird folgende **Entscheidungsregel** aufgestellt:

Falls mindestens 15 Paar einwandfrei sind, wird p = 0,8 als richtig angesehen, sonst soll p = 0,6 gelten.

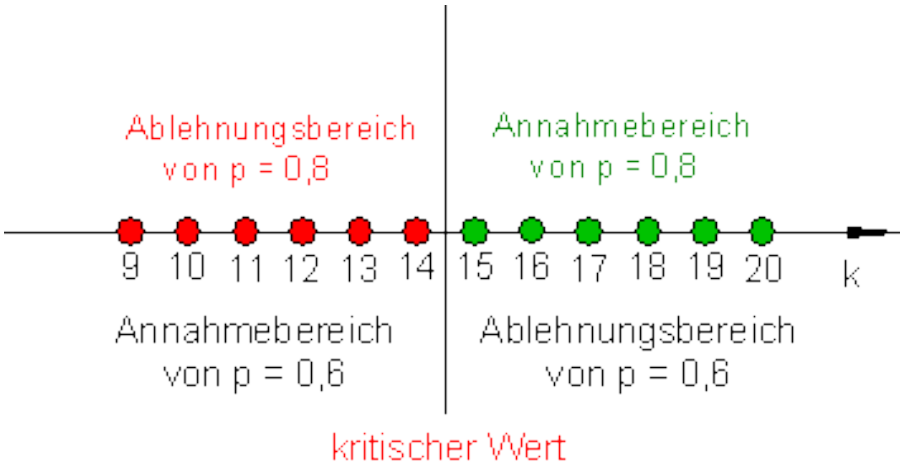
Die Hypothesen lauten: H_0: $p \geq 0{,}8$ und H_1: $p < 0{,}8$.

Aus der Vorgabe folgen Annahme- und Ablehnungsbereich für H_0.

Annahmebereich $A = \{15,16, \ldots 20\}$

Ablehnungsbereich $\overline{A} = \{0,1,2, \ldots ,14\}$

Falls H_0 abgelehnt werden muss, soll H_1: $p < 0{,}8 = 0{,}6$ gelten.



Fehlermöglichkeiten dieser Entscheidung:

1) p = 0,8 ist richtig, das bedeutet, der neue Lieferant kann wirklich Schuhe höherer Qualität liefern. Zufällig kann es vorkommen, das weniger als 15 Paar Schuhe einwandfrei sind. Dann würde man dem Lieferanten nicht glauben.

Die Wahrscheinlichkeit einen solchen Fehler zu begehen beträgt

$$P_{80}(X \leq 14) = 0{,}193$$

Tabelle 3:
Kumulierte Binomialverteilung für n = 20 und p = 0,8

k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)	k	P(X ≤ k)
3	0,000	6	0,000	9	0,001	12	0,032	15	0,370	18	0,931
4	0,000	7	0,000	10	0,003	13	0,087	16	0,589	19	0,988
5	0,000	8	0,000	11	0,010	14	0,196	17	0,794	20	1,000

Das bedeutet, wenn man einen solchen Zufallsversuch mit 20 Paar Schuhen sehr oft durchführen würde, könnte man in 19,6% der Fälle ein Ergebnis erwarten, das gegen die tatsächliche Qualität der Schuhe spricht.

Fehler 1. Art:

In 19,6% aller Fälle würde die wahre Hypothese, (die Schuhe des neuen Lieferanten sind besser) verworfen werden.

2) p = 0,6 ist richtig, das bedeutet, der neue Lieferant kann auch keine besseren Schuhe liefern, als die, die man bereits hat. Zufällig kann es aber vorkommen, dass trotzdem 15 oder mehr Paar Schuhe einwandfrei sind.

Man würde in diesem Fall fälschlicherweise die Schuhe des neuen Lieferanten für besser halten.

Die Wahrscheinlichkeit einen solchen Fehler zu begehen beträgt

$$P_{60} = P(x \leq 20) - P(X \leq 14) = 1 - 0{,}874 = 0{,}126$$

Das bedeutet, wenn man einen solchen Zufallsversuch mit 20 Paar Schuhen sehr oft durchführen würde, könnte man in 12,6% der Fälle ein Ergebnis erwarten, dass die Qualität der Schuhe höher angesehen wird, als sie tatsächlich ist.

Fehler 2. Art:

In 12,6% aller Fälle würde die falsche Hypothese, (die Schuhe des neuen Lieferanten sind besser) nicht verworfen werden.

Zusammenfassung der Fehlerarten:

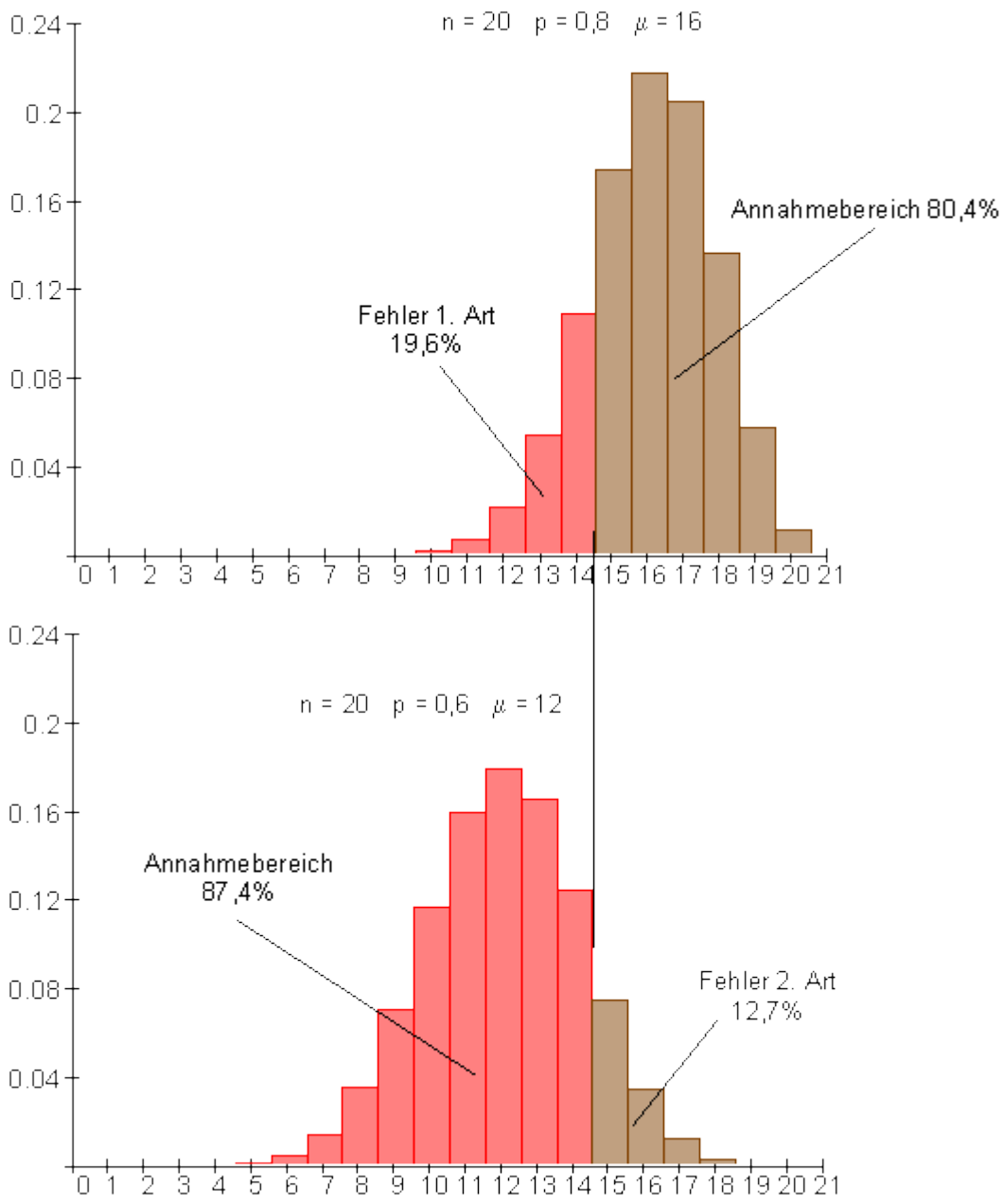
	Versuchsergebnis liegt im Annahmebereich von $H_0 : p \geq 0{,}8 \quad A = \{15 \ldots 20\}$	Versuchsergebnis liegt im Ablehnungsbereich von $H_0 : p \geq 0{,}8 \quad \overline{A} = \{0 \ldots 14\}$
Hypothese $H_0 : p \geq 0{,}8$ ist wahr	Entscheidung ist richtig, Hypothese $H_0 : p \geq 0{,}8$ wird angenommen $P_{80}(X \geq 15) = 0{,}804 \triangleq 80{,}4\%$	Entscheidung ist falsch, Hypothese $H_0 : p \geq 0{,}8$ wird zu unrecht abgelehnt $P_{80}(X \leq 14) = 0{,}196 \triangleq 19{,}6\%$ (Fehler 1. Art)
Hypothese $H_1 : p < 0{,}8 = 0{,}6$ ist wahr	Entscheidung ist falsch, Hypothese $H_1 : p < 0{,}8 = 0{,}6$ wird abgelehnt $P_{60}(X \geq 15) = 0{,}126 \triangleq 12{,}6\%$ (Fehler 2. Art)	Entscheidung ist richtig, Hypothese $H_1 : p < 0{,}8 = 0{,}6$ wird angenommen $P_{60}(X \leq 14) = 0{,}874 \triangleq 87{,}4\%$

Die Wahrscheinlichkeit einen Fehler 1. Art zu machen wird mit α bezeichnet (α=0,196)

Die Wahrscheinlichkeit einen Fehler 2. Art zu machen wird mit β bezeichnet (β=0,196)

Um den Fehler 2. Art zu berechnen, betrachtet man den Annahmebereich der Nullhypothese unter der Voraussetzung dass die Alternativhypothese gilt.

Der Fehler 2. Art ist die Wahrscheinlichkeit dafür, dass ein Testergebnis in den Annahmebereich der Nullhypothese fällt, obwohl die Alternativhypothese gilt.



Irrtumswahrscheinlichkeit wird vorgegeben.

Wird eine Irrtumswahrscheinlichkeit vorgegeben, dann ergibt sich daraus der Annahme und der Ablehnungsbereich.

Vorgabe der Irrtumswahrscheinlichkeit α ≤ 0,1

$P_{0,8}(X \leq k) \leq 0{,}1$

$k = 13$ aus der Tabelle ablesen, denn

$P_{0,8}(X \leq 13) \leq 0{,}087$

$\alpha = 0{,}087 < 0{,}1$ ist die Wahrscheinlichkeit für einen Fehler 1. Art

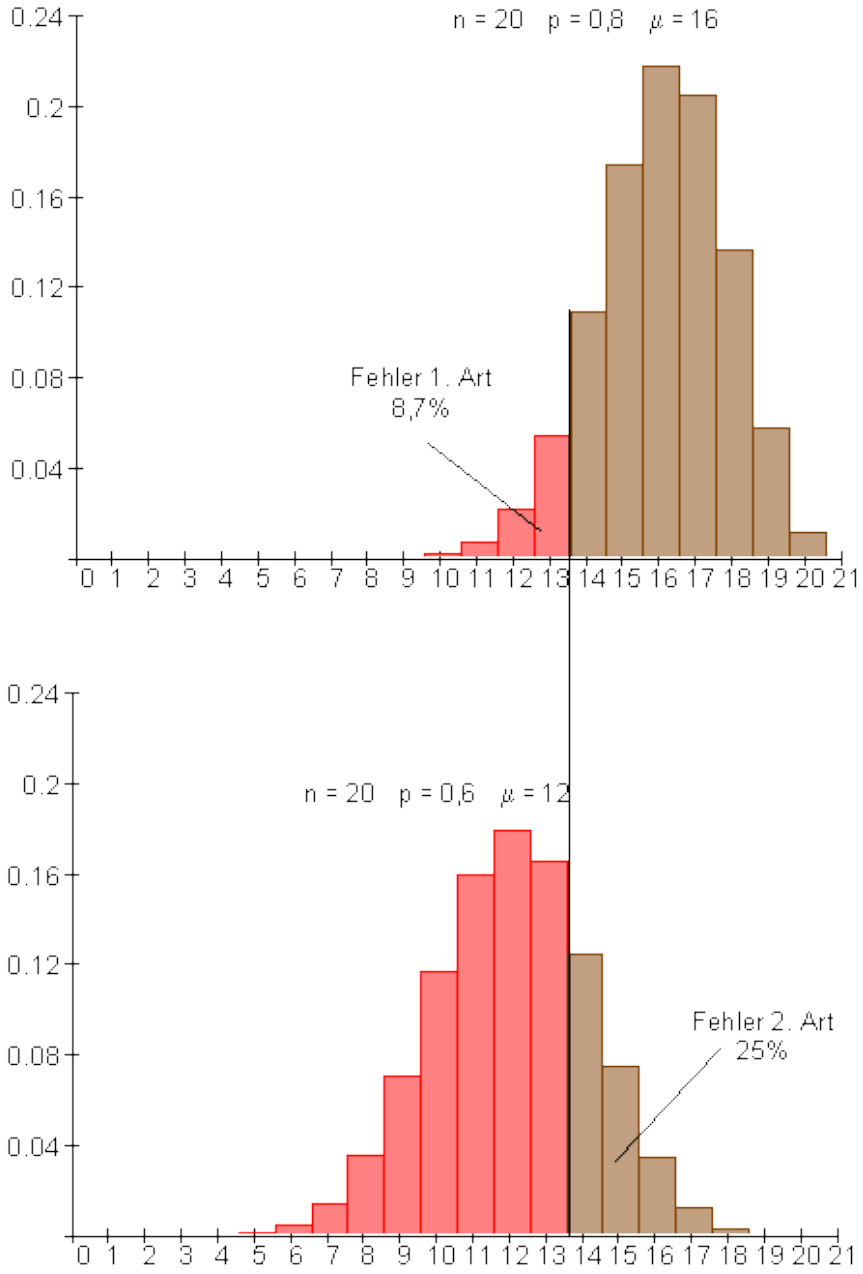
Annahmebereich $\quad A = \{14{,}16, \ldots 20\}$

Ablehnungsbereich $\quad \overline{A} = \{0{,}1{,}2, \ldots, 13\}$

$P_{0,6}(X \geq 14) = P_{0,6}(X \leq 20) - P_{0,6}(X \leq 13) = 1 - 0{,}750 = 0{,}25$

$\beta = 0{,}25$ ist die Wahrscheinlichkeit für einen Fehler 2. Art

Dadurch, dass die Wahrscheinlichkeit für einen Fehler 1. Art auf mehr als die Hälfte verringert wurde, hat sich der Fehler 2. Art etwa verdoppelt.



Falls die Hypothese p = 0,8 wahr ist, ist die Wahrscheinlichkeit dafür, dass sie aufgrund eines Testergebnisses fälschlicherweise abgelehnt wird 8,7%.

Denn in 8,7% aller Fälle liegt das Testergebnis im Ablehnungsbereich von p = 0,8.

Falls die Hypothese p = 0,6 wahr ist, ist die Wahrscheinlichkeit dafür, dass sie aufgrund eines Testergebnisses fälschlicherweise abgelehnt wird 25%.

Denn in 25% aller Fälle liegt das Testergebnis im Annahmebereich von p = 0,8.

Numerische Mathematik

Iterationsverfahren

Nullstellen von Funktionen bzw. Lösungen von Gleichungen werden mittels Computer sehr oft unter Nutzung verschiedener Näherungs- bzw. Iterationsverfahren ermittelt.

Definition 111:

Ein Iterationsverfahren wird immer dann verwendet, wenn für die Lösung einer Aufgabenstellung kein exakter analytischer Lösungsalgorithmus existiert.

Dies ist in der Praxis der typische Fall und nicht die stets exakt lösbare Aufgabe wie in der klassischen Schulmathematik.

Definition 112:

Die Grundprinzipien dieser Verfahren sind dabei Iteration und Intervallschachtelung, d.h., durch wiederholtes Anwenden einer Berechnungsvorschrift wird der Bereich, in dem die gesuchte Lösung liegt, immer weiter eingeschränkt.

Ergibt ein Verfahren bei bestimmten Anfangswerten eine Lösung, so nennt man das Verfahren **konvergent**, andernfalls **divergent**.

Bemerkung 59:

Die Güte eines Verfahrens wird durch seine Konvergenzgeschwindigkeit charakterisiert.

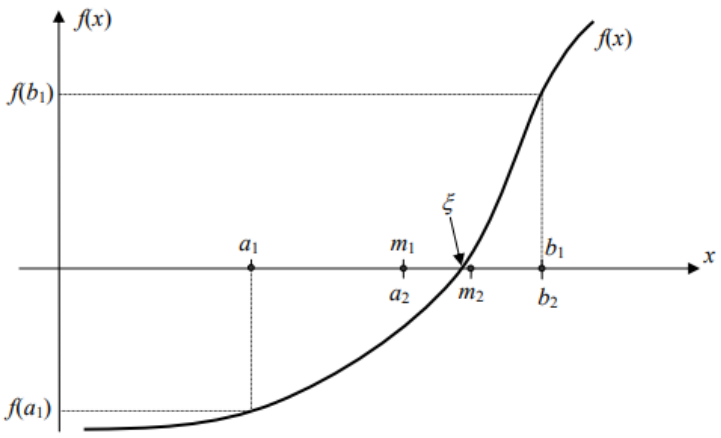
Bisektionsverfahren

Definition 113:

Dieses Verfahren ermöglicht es, Nullstellen numerisch zu berechnen, indem das Intervall, in dem sie auftreten können, immer weiter eingegrenzt wird, bis es kleiner als die geforderte Rechengenauigkeit ist.

Definition 114:

Sei f auf dem abgeschlossenen Intervall [a,b] setig mit f(a)<= und f(b)>0, dann hat f in]a,b[eine Nullstelle.[



Verfahren

- Setze die beiden Grenzen a und b. Bedingung: In diesem Intervall liegt eine Nullstelle.
- Testen ob eine gewünschte Genauigkeit vorhanden ist $|b-a| < \varepsilon$. Wenn ja, ist das Lösungsintervall (Nullstelle) gefunden.
- Sonst teile das Intervall [a,b] in der Mitte und tausche das Ergebnis mit der Zahl aus dem Intervall, welches das gleiche Vorzeichen besitzt.
- Wähle das Teilintervall, in dem wieder f das Vorzeichen wechselt.

Beispiel 108:

Berechnung von $\sqrt{2}$ als Nullstelle von $f(x) = x^2 - 2$ im Intervall [1, 2] mittels Bisektionsverfahren. Der Konvergenzradius soll $\varepsilon < 0{,}04$ sein.

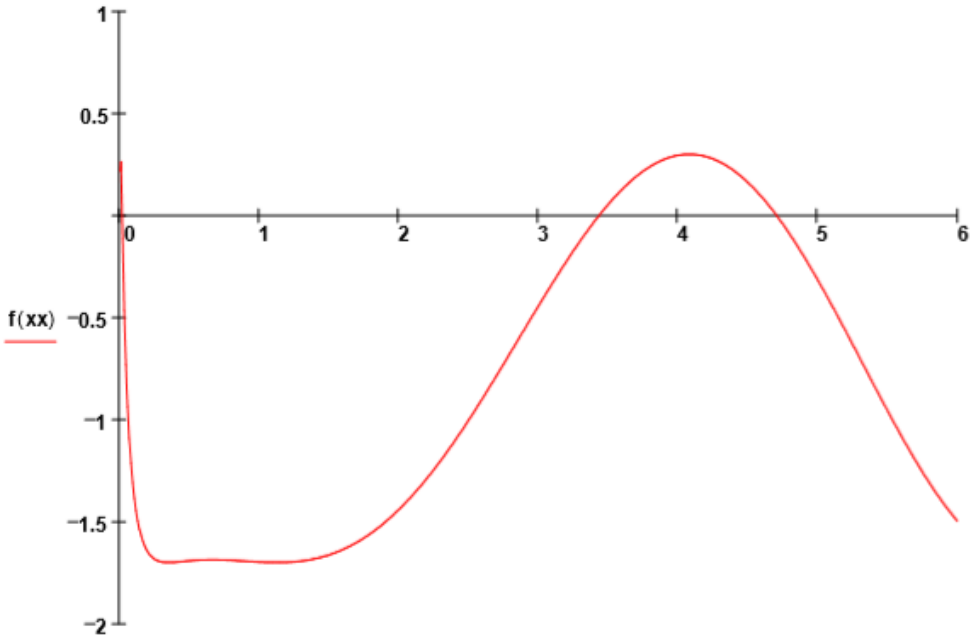
Start:	$f(1) = -1 < 0,\ f(2) = 2 > 0;$	$a_1 = 1,\ b_1 = 2$
Schritt 1:	$f(1.5) = 0.25 > 0;$	$a_2 = 1,\ b_2 = 1.5$
Schritt 2:	$f(1.25) = -0.4375 < 0;$	$a_3 = 1.25,\ b_3 = 1.5$
Schritt 3:	$f(1.375) = -0.109375 < 0;$	$a_4 = 1.375,\ b_4 = 1.5$
Schritt 4:	$f(1.4375) = 0.066406... > 0;$	$a_5 = 1.375,\ b_5 = 1.4375$
Schritt 5:	$f(1.40625) = -0.022461... < 0;$	$a_6 = 1.40625,\ b_6 = 1.4375$
usw.		

Nach 5 Schritten ist somit die erste Nachkommastelle ermittelt:

$$1.40625 < \sqrt{2} < 1.4375.$$

Beispiel 109:

$f(x) := \sin\left(\ln(x) - \frac{3}{2} \cdot x\right) - 0.7$ D = [3 ; 4]



Startwerte: xl := 3 xr := 4

f(xl) = −0.443 f(xr) = 0.295 $x := \frac{xl + xr}{2}$ x = 3.5 f(x) = 0.055

xl := wenn(f(x) · f(xl) ≥ 0 , x , xl) xr := wenn(f(x) · f(xr) ≥ 0 , x , xr)

xl = 3 xr = 3.5

f(xl) = −0.443 f(xr) = 0.055 $x := \frac{xl + xr}{2}$ x = 3.25 f(x) = −0.173

xl := wenn(f(x) · f(xl) ≥ 0 , x , xl) xr := wenn(f(x) · f(xr) ≥ 0 , x , xr)

xl = 3.25 xr = 3.5

f(xl) = −0.173 f(xr) = 0.055 $x := \frac{xl + xr}{2}$ x = 3.375 f(x) = −0.052

xl := wenn(f(x) · f(xl) ≥ 0 , x , xl) xr := wenn(f(x) · f(xr) ≥ 0 , x , xr)

xl = 3.375 xr = 3.5

f(xl) = −0.052 f(xr) = 0.055 $x := \frac{xl + xr}{2}$ x = 3.4375 $f(x) = 3.217 \times 10^{-3}$

xl := wenn(f(x) · f(xl) ≥ 0 , x , xl) xr := wenn(f(x) · f(xr) ≥ 0 , x , xr)

xl = 3.375 xr = 3.4375

$f(xl) = -0.052$ $f(xr) = 3.217 \times 10^{-3}$ $x := \frac{xl + xr}{2}$ $x = 3.40625$ $f(x) = -0.024$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.40625$ $xr = 3.4375$

$f(xl) = -0.024$ $f(xr) = 3.217 \times 10^{-3}$ $x := \frac{xl + xr}{2}$ $x = 3.421875$ $f(x) = -0.01$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.421875$ $xr = 3.4375$

$f(xl) = -0.01$ $f(xr) = 3.217 \times 10^{-3}$ $x := \frac{xl + xr}{2}$ $x = 3.4296875$ $f(x) = -3.528 \times 10^{-3}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.4296875$ $xr = 3.4375$

$f(xl) = -3.528 \times 10^{-3}$ $f(xr) = 3.217 \times 10^{-3}$ $x := \frac{xl + xr}{2}$ $x = 3.43359375$ $f(x) = -1.478 \times 10^{-4}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.43359375$ $xr = 3.4375$

$f(xl) = -1.478 \times 10^{-4}$ $f(xr) = 3.217 \times 10^{-3}$ $x := \frac{xl + xr}{2}$ $x = 3.435546875$ $f(x) = 1.537 \times 10^{-3}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.43359375$ $xr = 3.435546875$

$f(xl) = -1.478 \times 10^{-4}$ $f(xr) = 1.537 \times 10^{-3}$ $x := \frac{xl + xr}{2}$ $x = 3.4345703125$ $f(x) = 6.949 \times 10^{-4}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.43359375$ $xr = 3.4345703125$

$f(xl) = -1.478 \times 10^{-4}$ $f(xr) = 6.949 \times 10^{-4}$ $x := \frac{xl + xr}{2}$ $x = 3.43408203125$ $f(x) = 2.736 \times 10^{-4}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.43359375$ $xr = 3.43408203125$

$f(xl) = -1.478 \times 10^{-4}$ $f(xr) = 2.736 \times 10^{-4}$ $x := \frac{xl + xr}{2}$ $x = 3.433837890625$ $f(x) = 6.294 \times 10^{-5}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.43359375$ $xr = 3.433837890625$

$f(xl) = -1.478 \times 10^{-4}$ $f(xr) = 6.294 \times 10^{-5}$ $x := \frac{xl + xr}{2}$ $x = 3.4337158203125$ $f(x) = -4.244 \times 10^{-5}$

$xl := wenn(f(x) \cdot f(xl) \geq 0, x, xl)$ $xr := wenn(f(x) \cdot f(xr) \geq 0, x, xr)$

$xl = 3.4337158203125$ $xr = 3.433837890625$

$f(xl) = -4.244 \times 10^{-5}$ $f(xr) = 6.294 \times 10^{-5}$ $x := \frac{xl + xr}{2}$ $x = 3.43377685546875$ $f(x) = 1.025 \times 10^{-5}$

Newton-Verfahren

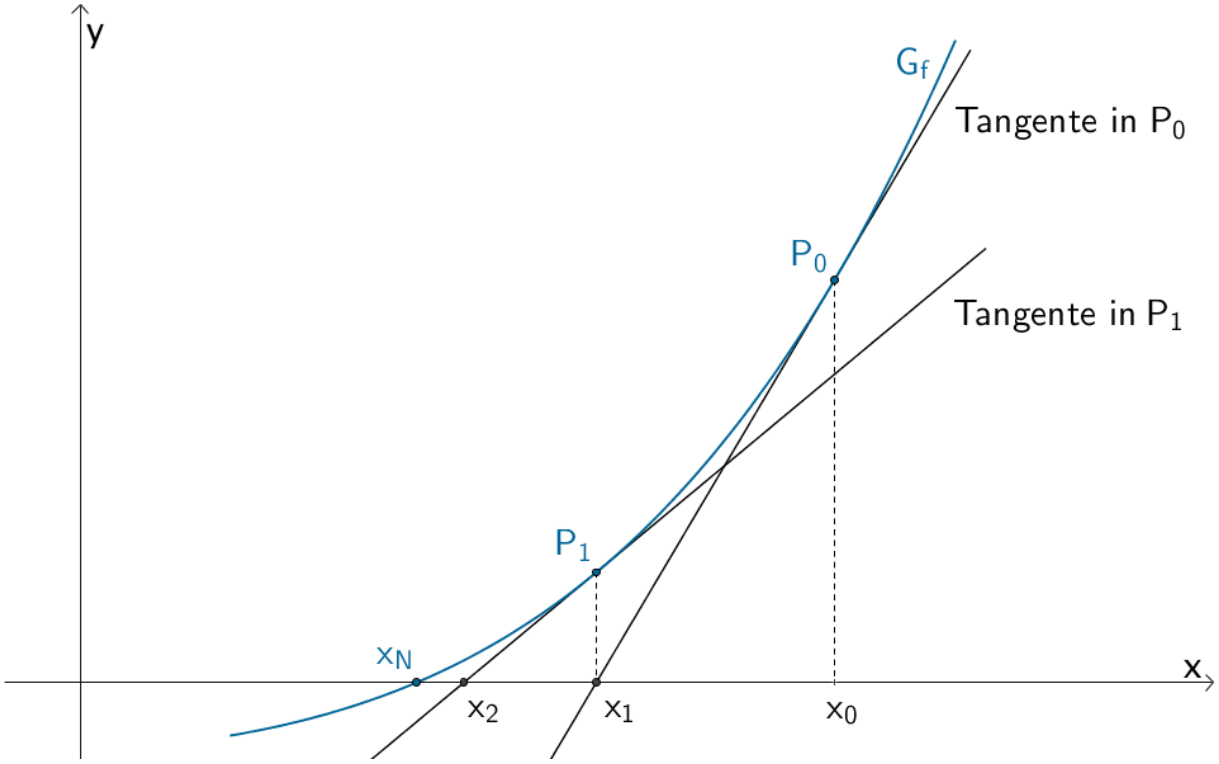
Das Newton-Verfahren, auch Tangentenverfahren genannt, ist ein iteratives Verfahren zur approximativen Bestimmung von Nullstellen. Es geht also um die Näherungen von Nullstellen.

Die Grundidee: Wir legen eine Tangente im Punkt

$}P(x_0|y_0 = f(x_0))$

an die nichtlineare Funktion an.

X_0 wird als Startwert bezeichnet. Dann wird die Nullstelle x_1 dieser Tangenten bestimmt. Im Koordinatensystem ist dieser erste Schritt abgebildet.



Dann wiederholen wir diesen Schritt (daher heißt es ja auch ein iteratives Verfahren). Wir legen diesmal die Tangente im Punkt $P(x_1|y_1)$ an die nichtlineare Funktion an. Die Nullstelle dieser Tangente wird dann x_2 genannt.

Herleitung der Iterationsvorschrift

Schauen wir uns einmal, ausgehend von dem Startwert x_0, die Bestimmung von x_1 an.

Aufstellen der Tangentengleichung

Allgemein hat eine Tangentengleichung die Form $t(x) = m \cdot x + n$.

- Dabei ist $m = f'(x_0)$ die Steigung des Tangenten (die erste Ableitung von $f(x)$ an der Stelle x_0).
- n ist der y-Achsenabschnitt. Dieser muss noch bestimmt werden.

Da der Punkt $P(x_0|f(x_0))$ auf der Tangente liegt, erhältst du diese Gleichung:

$$f(x_0) = f'(x_0) \cdot x_0 + n.$$

Diese Gleichung kann nach n umgeformt werden: $n = f(x_0) - f'(x_0) \cdot x_0$. Dies führt zu dieser Tangentengleichung:

$$t(x) = f'(x_0) \cdot x + f(x_0) - f'(x_0) \cdot x_0.$$

Bestimmung der Nullstelle der Tangente

Wir benennen nun die Nullstelle der Tangenten mit x_1, dann gilt $t(x_1) = 0$ oder

$$f'(x_0) \cdot x_1 + f(x_0) - f'(x_0) \cdot x_0 = 0 \,.$$

Division durch $f'(x_0)$ führt zu $x_1 + \frac{f(x_0)}{f'(x_0)} - x_0 = 0$. Nun kann x_0 addiert und $\frac{f(x_0)}{f'(x_0)}$ subtrahiert werden:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Wenn du nun in dieser Gleichung auf der rechten Seite überall x_0 durch x_1 ersetzt, erhältst du x_2 und so weiter.

Die Iterationsvorschrift des Newton-Verfahrens

Damit kannst du die **Iterationsvorschrift** des Newton-Verfahrens angeben:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Beispiel 110:

$f(x) = x^3 + 4x - 4$

-4	-3	-2	-1	0	1	2	3	4
-84	-43	-20	-9	-4	1	12	35	76

Vorzeichenwechsel im Intervall $x \in [0;1] \Rightarrow$ wähle z.B. $x_0 = 0,5$

So erhältst du deine angenäherte Lösung:

Je länger du das Verfahren anwendest desto näher kommst du an die Nullstelle. Ein Ziel deiner Näherung könnte sein, die ersten drei Nachkommastellen korrekt zu bestimmen. Wenn sich nach mehreren Iterationsschritten deine drei Nachkommastellen nicht mehr ändern, kannst du davon ausgehen, dass du am Ziel bist.

Beispiel:

$x_2 \approx 0,8486187342$

$x_3 \approx 0,8477079411$

$x_4 \approx 0,8477075981$

$x_5 \approx 0,8477075981 \Rightarrow$ Die Nullstelle liegt bei ca. $0,8477075981$.

Beispiel 111:

Du benötigst		Ergebnis	Erhältst du durch
$f(x)$	$=$	$\frac{1}{3}x^3 - x^2 - \frac{1}{3}$	
$f'(x)$	$=$	$x^2 - 2x$	Berechnen
x_0	$=$	3	Berechnen

$f'(x)$

- **Die Ableitung von $\frac{1}{3}x^3 - x^2 - \frac{1}{3} = x^2 - 2x$**
- **Übersicht zu den Rechenregel zur Ableitung von Polynomen.**

Lösungsweg Ableitung (hier Klicken)

x_0

Wertetabelle:

Setze verschiedene Werte als x ein um jeweils nach dem y-Wert aufzulösen. Trage dies anschließend in eine Wertetabelle ein und finde den Übergang vom Positiven/Negativen, diese zwei Punkte stellen dann dein Intervall dar. Beim Wählen beachte, dass x_0 keine Extremstelle darstellen darf.

Beispiel:

$f(1) = \frac{1}{3} \cdot 1^3 - 1^2 - \frac{1}{3}$

$f(1) = -1$

-4	-3	-2	-1	0	1	2	3	4
$-\frac{53}{3}$	$-\frac{19}{3}$	-1	$-\frac{5}{3}$	$-\frac{1}{3}$	-1	$-\frac{5}{3}$	$-\frac{1}{3}$	5

Vorzeichenwechsel im Intervall $x \in [3;4] \Rightarrow$ wähle z.B. $x_0 = 3,5$.

Berechnung	Erklärung
$x_m = x_n - \frac{f(x_n)}{f'(x_n)}$	
$x_1 = 3,5 - \frac{\frac{1}{3}\cdot 3,5^3 - 3,5^2 - \frac{1}{3}}{3,5^2 - 2\cdot 3,5}$	Setze $f(x), f'(x)$ und x_0 in die Formel ein. Und löse nach x_1 auf.
$x_1 = 3,5 - \frac{\frac{41}{24}}{\frac{287}{8}}$	
$x_1 = 3,452380952 = \frac{145}{42}$	
$x_2 = \frac{145}{42} - \frac{\frac{1}{3}\cdot(\frac{145}{42})^3 - (\frac{145}{42})^2 - \frac{1}{3}}{(\frac{145}{42})^2 - 2\cdot\frac{145}{42}}$	Setze $f(x), f'(x)$ und x_1 in die Formel ein. Und löse nach x_2 auf.
$x_2 = \frac{145}{42} - \frac{1,463966274}{5,014172336}$	
$x_2 = 3,160415266$	
$x_3 = 3,160415266 - \frac{\frac{1}{3}\cdot 3,160415266^3 - 3,160415266^2 - \frac{1}{3}}{3,160415266^2 - 2\cdot 3,160415266}$	Setze $f(x), f'(x)$ und x_2 in die Formel ein. Und löse nach x_3 auf.
$x_3 = 3,160415266 - \frac{0,2007545716}{3,667394122}$	
$x_3 = 3,10567488$	
$x_4 = 3,10567488 - \frac{\frac{1}{3}\cdot 3,10567488^3 - 3,10567488^2 - \frac{1}{3}}{3,10567488^2 - 2\cdot 3,10567488}$	Setze $f(x), f'(x)$ und x_3 in die Formel ein. Und löse nach x_4 auf.
$x_4 = 3,10567488 - \frac{0,006419030671}{3,4338667}$	
$x_4 = 3,10380555$	
$x_5 = 3,10380555 - \frac{\frac{1}{3}\cdot 3,10380555^3 - 3,10380555^2 - \frac{1}{3}}{3,10380555^2 - 2\cdot 3,10380555}$	Setze $f(x), f'(x)$ und x_4 in die Formel ein. Und löse nach x_5 auf.
$x_5 = 3,10380555 - \frac{0,000007,356513587}{3,425897892}$	
$x_5 = 3,103853353$	

$x_5 = 3,1038$ **ist die Annäherung der Nullstelle bis zur 4. Nachkommastelle von** $f(x) = \frac{1}{3}x^3 - x^2 - \frac{1}{3}$

Beispiel 112:

$f(x) = \ln(x^4 + 5x^3 - 5)$

Lösung ausblenden

$f(x) = \ln(x^4 + 5x^3 - 5)$

Der natürliche Logarithmus $\ln(x)$ ist nur auf den positiven, reellen Zahlen definiert.

Um die Nullstellen der Funktion $f(x)$ zu bestimmen macht man zuvor eine kurze Vorüberlegung.

Man betrachtet die Nullstellen des natürlichen Logarithmus und stellt folgendes fest:

Sei $g(x) = \ln(x)$:

$g(x) = 0 \Leftrightarrow \ln(x) = 0 \quad |e$

$e^{\ln(x)} = e^0$ Umformen.

$\Rightarrow x = 1$

Man sieht, dass $x = 1$ die einzige Nullstelle von $\ln(x)$ ist.

Um die Nullstellen von $f(x)$ zu approximieren, kann man also die "Einsstellen" der Funktion $h(x) = x^4 + 5x^3 - 5$ approximieren, d.h. man sucht die Lösung für die Gleichung $h(x) = x^4 + 5x^3 - 5 = 1$.

Da das Newtonverfahren Nullstellen approximiert macht man eine kleine Umformung und erhält:

$h(x) = x^4 + 5x^3 - 5 = 1 \quad |-1$

$x^4 + 5x^3 - 6 = 0$

Wir approximieren also die Nullstellen der Funktion $\tilde{h}(x) = x^4 + 5x^3 - 6$ um die Nullstellen von $f(x)$ zu finden.

Wertetabelle

x	-6	-5	-4	-3	-2	-1	0	1	2
$\tilde{h}(x)$	210	-6	-70	-60	-30	-10	-6	0	50

Bestimmen der Intervalle

Eine Nullstelle kann direkt aus der Tabelle abgelesen werden:

$\tilde{x}_1 = 1$

Man sieht außerdem, dass die Funktion $\tilde{h}(x)$ im Intervallen $]-6; -5[$ ihr Vorzeichen ändert.

Daraus folgt für die Nullstellen $\tilde{x}_2$:

$\Rightarrow \tilde{x}_2 \in]-6; -5[$

Um das Intervall weiter zu verkleinern und so einen besseren Anfangswert für das Newton-Verfahren zu bekommen, berechnet man den Funktionswert der Mittelwerte der ausgewählten Intervalle:

x	**-6**	**-5,5**	**-5**
f(x)	210	77,1875	-6

Man sieht nun, dass die Funktion $\tilde{h}(x)$ in den Intervallen $]-5,5; -5[$ ihr Vorzeichen ändert.

Daraus folgt für die Nullstellen $\tilde{x}_2$:

$\Rightarrow \tilde{x}_2 \in]-5,5; -5[$

Anwenden des Newton-Verfahrens

$\tilde{h}(x) = x^4 + 5x^3 - 6$

$\tilde{h}'(x) = 4x^3 + 15x^2$

$$\Rightarrow x_{n+1} = x_n - \frac{x_n^4 + 5x_n^3 - 6}{4x_n^3 + 15x_n^2}$$

Bestimmen der Nullstellen

Man wählt einen beliebigen Wert x_0 aus dem Intervall $]-5,5; -5[$, z.B. $x_0 = -5,25$.

$$\Rightarrow x_1 = x_0 - \frac{x_0^4 + 5x_0^3 - 6}{4x_0^3 + 15x_0^2}$$

Man berechnet jetzt x_1 mit der oben angegebenen Rekursionsformel.

$$= (-5,25) - \frac{(-5,25)^4 + 5(-5,25)^3 - 6}{4(-5,25)^3 + 15(-5,25)^2}$$

$$= -5,067531179 \approx -5,06753$$

$$x_2 = x_1 - \frac{x_1^4 + 5x_1^3 - 6}{4x_1^3 + 15x_1^2}$$

Dann berechnet man x_2 mit dem gerade berechneten x_1 und der oben angegebenen Rekursionsformel.

$$= (-5,067531179) - \frac{(-5,067531179)^4 + 5(-5,067531179)^3 - 6}{4(-5,067531179)^3 + 15(-5,067531179)^2}$$

$$= -5,046930085 \approx -5,04693$$

$$x_3 = x_2 - \frac{x_2^4 + 5x_2^3 - 6}{4x_2^3 + 15x_2^2}$$

Dann berechnet man x_3 mit dem gerade berechneten x_2 und der oben angegebenen Rekursionsformel.

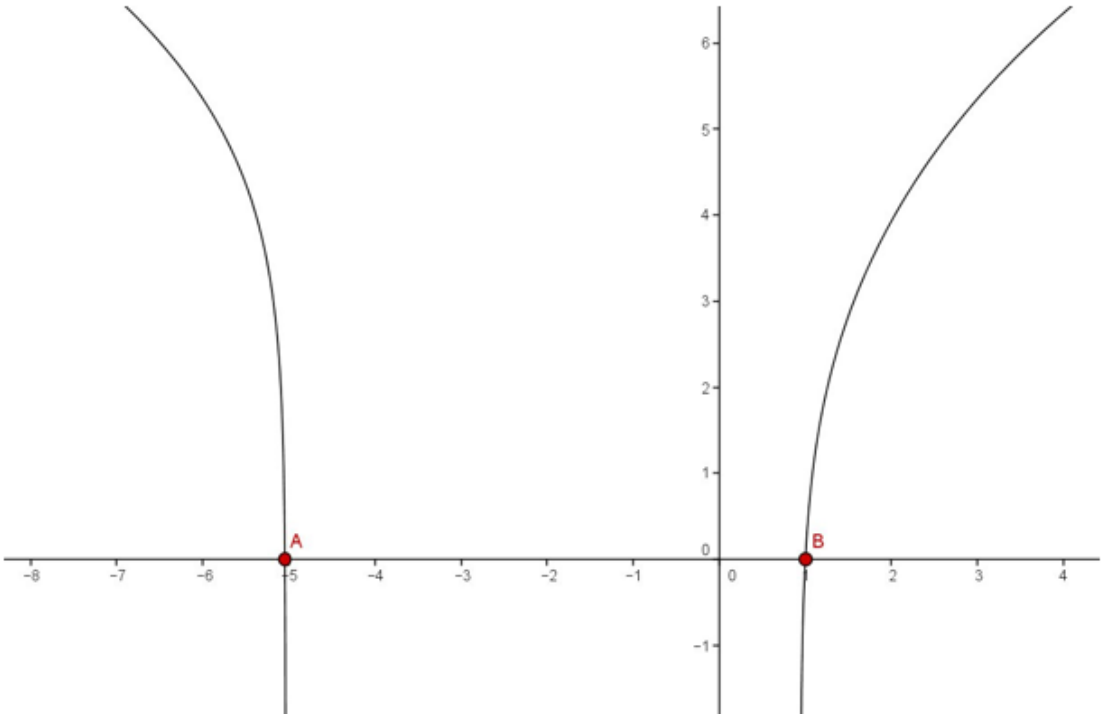
$$= (-5,046930085) - \frac{(-5,046930085)^4 + 5(-5,046930085)^3 - 6}{4(-5,046930085)^3 + 15(-5,046930085)^2}$$

$$= -5,046680361 \approx -5,04668$$

Man erkennt jetzt, dass sich die Genauigkeit der Lösung im letzten Schritt nur noch in der vierten Nachkommastelle verbessert.

Da nur eine Angabe bis auf zwei Nachkommastellen gefordert war, ist man in diesem Schritt fertig und das Ergebnis lautet:

$$\tilde{x}_2 = -5,05$$



Interpolationsverfahren

Definition 115:

In der numerischen Mathematik bezeichnet der Begriff Interpolation (aus lateinisch inter = dazwischen und polire = glätten, schleifen) eine Klasse von Problemen und Verfahren. Zu gegebenen diskreten Daten (z. B. Messwerten) soll eine stetige Funktion (die sogenannte **Interpolante** oder Interpolierende) gefunden werden, die diese Daten abbildet. Man sagt dann, die Funktion interpoliert die Daten.

Manchmal sind von einer Funktion nur einzelne Punkte bekannt, aber keine analytische Beschreibung der Funktion, durch die sie an beliebigen Stellen ausgewertet werden könnte.

Ein Beispiel sind Punkte als Resultat einer physikalischen Messung. Könnte man die Punkte durch eine (eventuell glatte) Kurve verbinden, so wäre es möglich, die unbekannte Funktion an den dazwischenliegenden Stellen zu schätzen.

In anderen Fällen soll eine schwierig handhabbare Funktion näherungsweise durch eine einfachere dargestellt werden.

Eine Interpolationsfunktion kann diese Anforderung der Einfachheit erfüllen. Diese Aufgabe bezeichnet man als Interpolationsproblem.

Es gibt für das Problem mehrere Lösungen, der Anwender muss zunächst geeignete Ansatzfunktionen wählen. Je nach Ansatzfunktionen erhalten wir eine andere Interpolante.

Bemerkung 60:

Interpolation ist die Kunst, zwischen den Zeilen einer Tabelle zu lesen (Rutishauser).

Lagrange-Interpolation oder Polynominterpolation

Definition 116:

Von einer Funktion f(x) seien Funktionswerte

$(x_i, f(x_i))$ mit $i = 0,1, \dots, n$

bekannt. Diese heißen Stützstellen, die xi heißen Knoten.

Gesucht ist eine Funktion $p(x_i = f(x_i).\, i = 0,1, \dots, n$

Dabei ergeben sich Polynome mit einem bestimmten Höchstgrad.

Dabei werden Lagrange-Polynome wie folgt definiert.

$$L_k(x_i) = \prod_{i=0, i\neq k}^{n} \left(\frac{x - x_i}{x_k - x_i}\right)$$

$$= \frac{(x - x_0) \cdot \dots \cdot (x - x_{k-1}) \cdot (x - x_{k+1}) \cdot \dots \cdot (x - x_n)}{(x - x\_0\,) \cdot \dots \cdot (x - x_{k-1}) \cdot (x - x_{k+1}) \cdot \dots \cdot (x - x_n)}$$

Das endgültige Polynomergibt sich dann aus:

$$p_n(x) = \sum_{k=0}^{n} y_k \cdot L_k(x)$$

Beispiel 113:

Erstellen sie ein Interpolationspolynom, welches durch die folgenden Stützstellen geht.

Beispiel: Gegeben seien (wiederum) die Punkte
$P_0(1;2), P_1(2;3), P_2(3;1) und P_3(4;3)$.

Dann sind zunächst die lagrangeschen Polynome 3. Grades $L_0; L_1; L_2 und L_3$ zu berechnen, und es ist:

$$L_0 = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}$$
$$= \frac{(x-2)(x-3)(x-4)}{(1-2)(1-3)(1-4)} = -\frac{1}{6}\left(x^3 - 9x^2 + 26x - 24\right)$$

$$L_1 = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}$$
$$= \frac{(x-1)(x-3)(x-4)}{(2-1)(2-3)(2-4)} = \frac{1}{2}\left(x^3 - 8x^2 + 19x - 12\right)$$

$$L_2 = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}$$
$$= \frac{(x-1)(x-2)(x-4)}{(3-1)(3-2)(3-4)} = -\frac{1}{2}\left(x^3 - 7x^2 + 14x - 8\right)$$

$$L_3 = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}$$
$$= \frac{(x-1)(x-2)(x-3)}{(4-1)(4-2)(4-3)} = \frac{1}{6}\left(x^3 - 6x^2 + 11x + 6\right)$$

Dann ergibt sich das lagrangesche Interpolationspolynom in der Form
$L(x) = 2 \cdot L_0 + 3 \cdot L_1 + 1 \cdot L_2 + 3 \cdot L_3,$
woraus nach Einsetzen und Umformen
$L(x) = \frac{1}{6}\left(7x^3 - 51x^2 + 110x - 54\right)$
folgt. Erwartungsgemäß ist dieses Ergebnis identisch mit dem des vorangehenden Beispiels.

Beispiel 114:

Erstellen sie ein Interpolationspolynom, welches durch die folgenden Stützstellen geht.

$x_0 = 2\ ,\ y_0 = 3$
$x_1 = 7\ ,\ y_1 = 2$
$x_2 = 10,\ y_2 = 4$

$$L_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 7)(x - 10)}{(2 - 7)(2 - 10)} = \frac{(x - 7)(x - 10)}{40}$$

$$L_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 2)(x - 10)}{(7 - 2)(7 - 10)} = \frac{(x - 2)(x - 10)}{-15}$$

$$L_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 2)(x - 7)}{(10 - 2)(10 - 7)} = \frac{(x - 2)(x - 7)}{24}$$

$$p_2(x) = 3 \cdot \frac{(x - 7)(x - 10)}{40} + 2 \cdot \frac{(x - 2)(x - 10)}{-15} + 4 \cdot \frac{(x - 2)(x - 7)}{24}$$

$$p_2(x) = \frac{270\,(x^2 - 17x + 70)}{3600} - \frac{480\,(x^2 - 12x + 20)}{3600} + \frac{600\,(x^2 - 9x + 14)}{3600}$$

$$p_2(x) = \frac{270x^2 - 480x^2 + 600x^2 - 4590x + 5760x - 5400x + 18900 - 9600 + 8400}{3600}$$

$$p_2(x) = \frac{390}{3600}x^2 - \frac{4230}{3600}x + \frac{17700}{3600}$$

$$\mathbf{p_2(x) = \frac{13}{120}x^2 - \frac{47}{40}x + \frac{59}{12}}$$

Beispiel 115:

Erstellen sie ein Interpolationspolynom, welches durch die folgenden Stützstellen geht.

Beispiel 1: Bestimmung der Lagrange-Darstellung des Interpolationspolynoms zu den Daten

i	0	1	2	3
x_i	-3	-1	1	3
y_i	0	16	32	0

Da nur y_1 und y_2 von Null verschieden sind, braucht man nur L_1 und L_2 zu berechnen.

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x+3)(x-1)(x-3)}{(-1+3)(-1-1)(-1-3)}$$

Der Zähler ist $(x-1)(x^2-9) = x^3-x^2-9x+9$, der Nenner ist der Zähler bei $x=-1$, also $-1-1+9+9=16$. Also ist $L_1(x) = \frac{1}{16}(x^3-x^2-9x+9)$.

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x+3)(x+1)(x-3)}{(1+3)(1+1)(1-3)}$$

$$= \frac{(x^2-9)(x+1)}{-16} = -\frac{1}{16}(x^3+x^2-9x-9)$$

Damit ist das Interpolationspolynom

$$p(x) = 16 \cdot L_1(x) + 32 \cdot L_2(x) = 1 \cdot (x^3-x^2-9x+9) - 2 \cdot (x^3+x^2-9x-9)$$

$$= -x^3-3x^2+9x+27$$

Beispiel 116:

Beispiel 2: Wie sieht das allgemeine Interpolationspolynom zu den Knoten $a-h$, a und $a+h$, $h \neq 0$, in der Lagrange-Darstellung aus?

Zunächst wird das Problem für $a = 0$ gelöst: Mit $x_0 = -h$, $x_1 = x$ und $x_2 = +h$ erhält man

$$L_0(x) = \frac{x(x-h)}{(-h)(-2h)} = \frac{1}{2h^2}(x^2 - hx)$$

$$L_1(x) = \frac{(x+h)(x-h)}{h(-h)} = \frac{-1}{h^2}(x^2 - h^2)$$

$$L_2(x) = \frac{(x+h)x}{h \cdot 2} = \frac{1}{2h^2}(x^2 + hx)$$

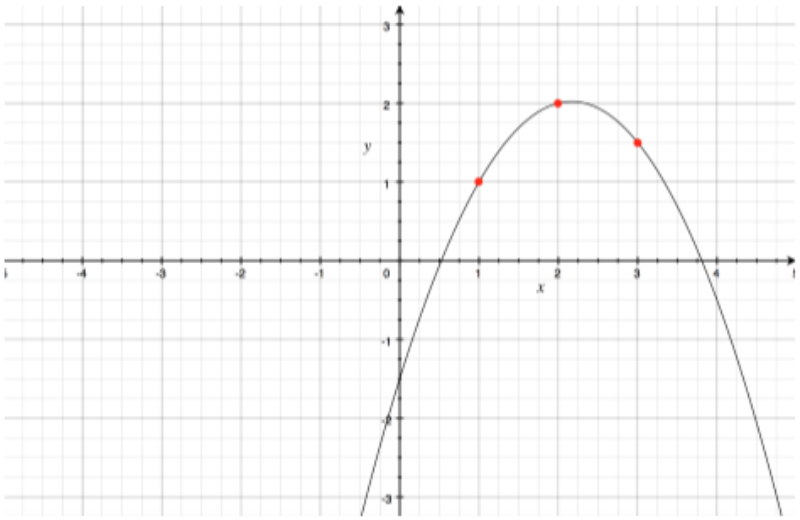
Damit ist

$$p(x) = \frac{1}{2h^2}(y_0 \cdot (x^2 - hx) - 2y_1 \cdot (x^2 - h^2) + y_2 \cdot (x^2 + hx).$$

Die Lösung des allgemeinen Problems erhält man jetzt, indem man x durch $(x-a)$ ersetzt:

$$p(x) = \frac{1}{2h^2}\Big(y_0 \cdot ((x-a)^2 - h(x-a)) - 2y_1 \cdot ((x-a)^2 - h^2) + y_2 \cdot ((x-a)^2 + h(x-a))\Big)$$

Beispiel 117:



$$f(x) = \sum_{j=0}^{k} y_j l_j(x)$$

$$l_j(x) = \prod_{m=0\ldots k;\ m\neq j} \frac{x - x_m}{x_j - x_m}$$

Wenn man eine gewisse Anzahl Punkte gegeben hat, aber keine Funktion kennt, die durch alle Punkte geht, benutzt man eine Interpolation. Dabei verändert man eine bekannte Funktion so, dass sie durch alle Punkte geht. Die Lagrange Interpolation ist dabei eine spezielle Methode, Polynome so anzupassen, dass sie durch alle Punkte gehen. Wir betrachten jetzt ein Beispiel mit drei Punkten, also k=2.
x_i= 1, 2, 3 und y_i= 1, 2, 1.5

$$l_0 = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}$$

$$l_1 = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)}$$

$$l_2 = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

Nun setzt man die Punkte ein:

$$l_0 = \frac{(x-2)(x-3)}{(-1)(-2)} = \frac{1}{2}(x^2 - 5x + 6)$$

$$l_1 = \frac{(x-1)(x-3)}{(1)(-1)} = -(x^2 - 4x + 3)$$

$$l_2 = \frac{(x-1)(x-2)}{(2)(1)} = \frac{1}{2}(x^2 - 3x + 2)$$

Jetzt setzt man die y_i in die Formel ein und erhält:

$$f(x) = (1)\frac{1}{2}(x^2 - 5x + 6) - (2)(x^2 - 4x + 3) + (1.5)\frac{1}{2}(x^2 - 3x + 2)$$

$$f(x) = -0.75x^2 + 3.25x - 1.5$$

Wie man sieht, ist der Grad des Polynoms immer um eins kleiner als die Anzahl Punkte die man verwendet. Betrachtet man den Graphen der gefundenen Funktion, sieht man, dass sie dort genau wie erwünscht durch den Punkt geht.

Beispiel 118:

Wir betrachten also die Stützpunkte

$$(2, 0.5)\ ,\ (2.5, 0.4)\ ,\ (4, 0.25)$$

$$L_{2,0}(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-2.5)(x-4)}{(2-2.5)(2-4)} = (x-6.5)x + 10$$

$$L_{2,1}(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-2)(x-4)}{(2.5-2)(2.5-4)} = \frac{1}{3}((-4x+24)x - 32)$$

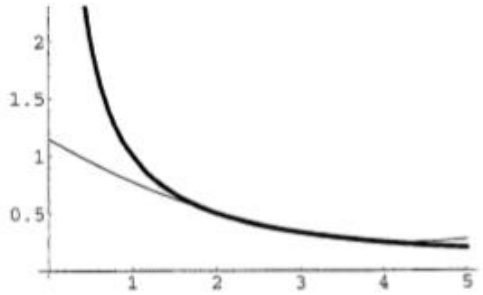
$$L_{2,2}(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-2)(x-2.5)}{(4-2)(4-2.5)} = \frac{1}{3}((x-4.5)x + 5)$$

Damit erhalten wir

$$P_2(x) = \sum_{k=0}^{2} y_k L_{2,k}(x) =$$

$$= 0.5 \cdot [(x-6.5)x + 10] + \frac{0.4}{3} \cdot [(-4x+24)x - 32] + \frac{0.25}{3} \cdot [(x-4.5)x + 5] =$$

$$= (0.05x - 0.425)x + 1.15$$



Wir können jetzt etwa $f(3) = \frac{1}{3}$ durch $P_2(3) = 0.325$ approximieren.