STATISTISCHE METHODEN IN DER COMPUTERLINGUSITIK

ENRICO LIEBLANG

Zusammenfassung. Die Vorlesung Statistische Methoden in der Computerlinguistik behandelt die Grundlagen und gängigen Methoden zur statistischen Analyse von Texten und zur Verarbeitung von Sprache. Dieses Skript ist als vorlesungsbegleitende Unterlage für den ersten Teil (Lieblang) der Vorlesung Mathematische Methoden in der Computerlinguistik III gedacht. Weitere Diskussion und Erläuterungen werden in der Vorlesung gegeben.

1. Deskriptive Statistik

1.1. **Grundgesamtheit, Merkmale.** Deskriptive Statistik beschreibt die Daten einer Untersuchung. Bei der Untersuchung von Daten ist die zugrunde liegende Menge vorab zu definieren. Diese Menge heißt Grundgesamtheit oder Population. Sie entspricht der Menge von Objekten, über die man eine Aussagen machen will. Sie kann endlich oder unendlich sein. Eine Grundgesamtheit können z. B. die Menge aller Bürger eines Staates, die Menge aller Autos, die Menge aller möglichen Sätze einer Sprache. darstellen. Die Untersuchung kann auf Teilmengen der Grundgesamtheit eingeschränkt werden, etwa nur die weiblichen Einwohner einer Stadt, nur wisenschaftliche Texte etc. Man spricht dann von einer Teilgesamtheit. Ist die Grundgesamtheit zu groß, um vollständig untersucht zu werden, wird eine sog. Stichprobe gezogen. Die Elemente der Grundgesamtheit können bezüglich einer bestimmten Fragestellung bestimmte Merkmale besitzen wie Grösse, Gewicht, syntaktische Kategorie etc. Man versteht unter einem Merkmal einen speziellen Gesichtspunkt, unter dem die Elemente der Grundgesamtheit untersucht werden sollen. Denkbar ist die Größe von Personen, die Häufigkeit eines Wortes oder die Farbe von Autos. Das Feststellen eines Merkmals an einem konkreten Element heißt Merkmalsausprägung. Diese Merkmals-ausprägungen werden bei einer Untersuchung also an den einzelen Objekten bestimmt, also etwa die Farbe eines vorbeifahrenden Autos, die Häufigkeit des Wortes Essen in einem Corpus.

1.1.1. *Merkmalstypen.* Merkmale können wie folgt unterschieden werden (es besteht eine Rangfolge von Nominalskala zu Verhältnisskala):

(1) Nominalskala
Bei der Feststellung eines Merkmals erhalten Objekte mit gleicher Ausprägung dieselbe Zuordnung, verschiedene Ausprägungen erhalten verschiedene Zuordnungen.
Beispiel: Farbe von Autos: rot $\to$ 1, blau $\to$ 2 usw. Jedem blauen Auto wird dann die 2 zugeordnet.

(2) Ordinalskala
Es besteht eine Rangfolge der Merkmalsausprägungen. Ein Objekt mit der besseren Ausprägung erhält eine höhere Zahl als Zuordnung.
Beispiel: Schulnoten

(3) Intervallskala
Die Zuordnung von Zahlen gestattet die Interpretation von Abständen, es

	Zählen	Ordnen	Differenz	Quotient
nom	x	-	-	-
ord	x	x	-	-
interv	x	x	x	-
verh	x	x	x	x

fehlt ein absoluter Nullpunkt
Beispiel: Temperaturskala Celsius

(4) Verhältnisskala
Die Zuordnung gestattet die Berechnung von Quotienten, es existiert ein absoluter Nullpunkt.
Beispiel: Alter von Personen

Neben dieser Unterscheidung können Merkmale auch in qualitative und quantitative Merkmale unterschieden werden. Hierbei heißt qualitativ, dass das Merkmal höchstens endlich viele Ausprägungen besitzt und höchstens ordinalskaliert ist, quantitativ heißt, dass das Merkmal endlch oder unendlch viele Ausprägungen besitzt und und mindestens intervallskaliert ist. Entsprechend dem Merkmalstyp können nicht alle arithmetischen Operationen mit den Merkmalsausprägungen vorgenommen werden.

1.2. **Beschreibung von Daten.** Es seien n Untersuchungseinheiten $u_1, u_2, ..., u_n$ gegeben. An diesen werde das Merkmal X untersucht. Man stellt bei u_i die konkrete Merkmalsausprägung x_i fest. Die Menge dieser x_i heißt die Urliste der Untersuchung. Die möglichen Ausprägungen des Merkmals seien mit $a_1, ..., a_k$ bezeichnet. Jedes x_i besitzt dann einen Wert aus der Menge der a_i.
Die Anzahl der x_i mit Ausprägung a_i heißt absolute Häufigkeit der Merkmalsausprägung a_i. Man schreibt

$$h_j := h(a_j) = \text{Anzahl der} \quad x_i \quad \text{mit Merkmalsausprägung} \quad a_j.$$

Die relative Häufigkeit ist definiert wie folgt:

$$f(a_j) = \frac{h_j}{n} =: f_j$$

Es gilt:

$$\sum_{i=1}^{n} h_j = n, \quad \sum_{i=1}^{k} f_i = 1$$

Die Abbildung, die jeder Ausprägung eines Merkmals die relative Häufigkeit zuordnet, heißt Häufigkeitsverteilung des Merkmals. Bei sehr vielen Merkmalsausprägungen, etwa bei stetigen Merkmalen, wird eine sog. Klassenaufteilung vorgenommen, und man erstellt eine Häufigkeitsverteilung nach der Anzahl der Häufigkeiten einer Klasse.

1.3. **Lagemaße.** Oft stellt man bei einer Datenmenge die Fragen: Wo liegt das Datenzentrum? Wie stark streuen die Daten um das Zentrum? Gibt es Ausreißer? Sind die Daten symmetrisch um das Zentum verteilt?

Lagemaße beschreiben die Daten in Bezug auf diese Fragestellungen.

(1) Der Modus
Der häufigste Wert der vorliegenden Ausprägungen. Schon bei nominalem Niverau bildbar.

(2) Der Median
Der Wert, der die Datenmenge in 2 gleiche Hälften teilt, so dass die eine Hälfte oberhalb, die andere unterhalb des Medians liegt. Ist $x_1, ..., x_n$ eine Urliste (mindestens ordinal skaliert), so heißt die der Größe nach geordnete Liste $x_{(1)}, x_{(2)}, ..., x_{(n)}$ die geordnete Urliste. Man schreibt die Indizes in Klammern. Hiermit kann der Median wie folgt definiert werden:

$$x_{med} := \begin{cases} x_{(\frac{n+1}{2})} & : \quad n \quad ungerade \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & : \quad n \quad gerade \end{cases}$$

Man ordnet also die Daten der Größe nach an und bestimmt den Median nach dieser Formel. Eine Folge 122344555567 von Messwerten hat den Median $\frac{4+5}{2} = 4,5$. Jeder andere Wert zwischen 4 und 5 könnte ebenfalls als Median gewählt werden, man einigt sich jedoch auf den Mittelwert. Sind die Merkmalsausprägungen bei einem ordinalen Merkmal nicht als Zahlen gegeben (z.B. sehr gut, gut, befriedigend,...), so kann der Median nicht nach obiger Formel berechnet werden, sondern muss irgendwie sprachlich umschrieben werden.

(3) Das arithmetische Mittel
Liegen mindestens metrische Daten vor, so ist das arithmetische Mittel $\overline{x}$ folgender Wert:

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

Das arithmetische Mittel reagiert empfindlich auf Ausreißer (also sehr große oder kleine Werte). Der Median und der Modus sind hierfür unempfindlich.

1.4. **Streuungsmaße.** Auf die Frage, wie weit die Daten um ihren Datenmittelpunkt angesiedelt sind (streuen), geben die Streuungsmaße Antwort. Wichtigstes Streuungsmaß ist die Empirische Varianz und damit verbunden die Streuung von Daten.

(1) Empirische Varianz
Liegen die n Werte $x_1, ..., x_n$ einer Datenerhebung vor (mindestens intervallskaliert), so heißt

$$S^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

die empirische Vaianz der Daten. Der Wert

$$S := +\sqrt{S^2}$$

heißt Standadabweichung oder Streuung. Später wird für Zufallsvariablen ebenfalls eine Varianz eingeführt und zur Unterscheidung von der empirischen Varianz mit (σ^2) bezeichnet. Ein wichtiges Hilfsmittel zur Berechnung der empirischen Varianz ist die Varianzzerlegungsformel. Es gilt:

$$S^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i^2) - \overline{x}^2$$

(2) Die Spannweite
Ein weiteres Streuungsmaß ist die Spannweite:

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

also der Abstand zwischen dem größten un kleinsten Wert der Datenmenge. Auch hier ist mindestens intervallskaliertes Datenniveau notwendig.

Es ist oft schwierig, Daten mit verschiedenen Streuungen miteinander zu vergleichen. Dieser Vergleich wird ermöglicht durch den sogenannten z-Wert. Für jeden Datenwert x_i bildet man

$$z_i := \frac{x_i - \overline{x}}{S}$$

Die Abweichung vom Mittelwert wird in Beziehung gesetzt zur Streuung der Datenmenge. Der z-Wert gibt die Abweichung vom Mittelwert in Einheiten der Streuung an. Berechnet man für jeden Wert x_i des Stichprobenbefundes den z-Wert z_i, so hat die so entstandene Datenmenge den Mittelwert 0 und die Varianz 1.

2. Grundbegriffe der Wahrscheinlichkeitstheorie

Statistik und Wahrscheinlichkeitstheorie haben ihren Ursprung in den Glücksspielen des 16. Jahrhunderts, als man versuchte, Vorhersagen für den Ausgang der gängigen Glücksspiele zu erhalten. Viel verwendetes Anwendungsbeispiel ist das Würfeln mit einem Würfel. Es ist einsichtig, dass hier keine treffende Voraussage über den Ausgang gemacht werden kann, sondern nur eine Aussage, welche die ´´Wahrscheinlichkeit oder Chance´´ angibt, mit der eine bestimmte Zahl gewürfelt wird. In der Wahrscheinlichkeitstheorie wird dieser vage Begriff axiomatisch gefasst und in eine für weitere Untersuchungen brauchbare Fassung gebracht. Um ein Instrumentarium zu entwickeln, welches den Umgang mit solchen Aussagen erlaubt, ist es nötig, einige Begriffe zu definieren. Es werde von einer Grundgesamtheit GS ausgegangen, welche für eine gegebene Anwendung passend gewählt wurde. Wenn man z. B. Aussagen über das Würfeln mit einem Würfel machen will, wählt man als Grundgesamtheit die Zahlen von 1 bis 6. Folgende Begriffe sind nun wichtig:

(1) Die Elemente der Grundgesamtheit heißen Ergebnisse,

(2) Teilmengen der Grundgesamtheit heißen Ereignisse.

Es ist wichtig zu bemerken, dass Wahrscheinlichkeiten nur für Ereignisse definiert werden können. Einelementige Ereignisse heißen Elementarereignisse. Ein Ereignis ist aufgetreten, wenn ein Element dieses Ereignisses als Ergebnis beobachtet wird. Ein Ereignis, das immer auftritt, heißt sicheres Ereignis, ein Ereignis, das tritt niemals auftritt, heißt unmögliches Ereignis. Das Komplementäre Ereignis $\overline{A}$ zu einem Ereignis A ist das Ereignis, welches auftritt, wenn A nicht auftritt.

Die formale Fassung des Begriffs der Wahrscheinlichkeit ist auch unter dem Begriff Kolmogorovsche Axiome bekannt. Hiernach ist ein Wahrscheinlichkeitsmaß eine Abbildung P von der Potenzmenge F einer GS in das Intervall [0,1] mit den Eigenschaften:

(1) PA $\geq$ 0 für jedes A $\in$ F

(2) Für 2 Ereignisse A, B mit $A \bigcap B = \emptyset$ gilt: $P(A \bigcup B) = P(A) + P(B)$

(3) P(GS)=1

In dem schon genannten Würfelbeispiel entspricht die Grundgesamtheit GS der Ergebnismenge

$$GS = \{1, 2, 3, 4, 5, 6\}$$

In der WK-Theorie entspricht dem gemeinsamen Auftreten von A und B der Durchschnitt, der Vereinigung das Auftreten von A oder B (nicht ausschließend) und dem Komplement von A das Auftreten des Gegenteils von A. Es ergeben sich direkt einige leicht zu beweisende Tatsachen: Sind A und B Ereignisse, so gilt:

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

(2) $P\emptyset = 0$

(3) $PA \leq 1$

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

(5) $A \subseteq B \Rightarrow P(A) \leq P(B)$

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

(7) $P(A \setminus B) = P(A) - P(A \cap B)$

(8) Sind $A_1, ..., A_n$ paarweise disjunkte Ereignisse, so gilt $P(A_1 \cup A_2 ... \cup A_n) = \sum_{i=1}^{n} P(A_i)$

2.1. **Das Gesetz der großen Zahlen.** Die Axiome vom Kolmogorov geben kein Verfahren an, mit denen sich tatsächlich Wahrscheinlichkeiten berechnen lassen. Doch ist gerade das Bestimmen eines Wertes für eine Wahrscheinlichkeit eines Ereignisses ein wichtiger Vorgang. Es gibt mehrere Möglichkeiten, hier zu Werten zu kommen. Ein Zufallsexperment werde n mal wiederholt und es werde nur betrachtet, ob hierbei das Ereignis A aufgetreten ist oder nicht. Mit n_A sei die Häufigkeit des Auftretens von A beim n-ten Versuchen bezeichnet. Dann ist nach oben die Zahl $f(A) = \frac{n_A}{n}$ die relative Häufigkeit der Auftretens von A beim n-ten Versuch. Das Gesetz der großen Zahlen sagt nun aus, dass dieser Wert für den Grenzwert n gegen Unendlich gegen die wahre Wahrscheinlichkeit P(A) von A konvergiert, dass also gilt:

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

Diese Aussage heißt Gesetz der großen Zahlen:

Dies kann somit auch zur Bestimmung der Wahrscheinlichkeit eines Ereignisses benutzt werden und man spricht dann von der häufigkeitstheoretischen Wahrscheinlichkeit. Bei einer sehr Großen Anzahl von Würfen eines Würfels erhält man also eine immer bessere Näherung für die wahren Wahrscheinlichtkeiten der Augenzahlen. (Näherung nur deshalb, da ja nicht unendlich oft gewürfelt werden kann). Eine andere Art,Wahrscheinlichkeiten zu bestimmen, ist die der Laplacewahrscheinlichkeit, bei der die Anzahl der für ein Ereignis günstigen Fälle zu allen möglichen Fällen in Beziehung gesetzt wird. Hierbei ist es notwendig, dass alle Elementarwahrscheinlichkeiten dieselbe Wahrscheinlichkeit besitzen. Hierauf wird im Abschnitt Kombinatorik eingegangen.

2.2. **Bedingte Wahrscheinlichkeiten.** Wenn man die Wahrscheinlichkeit mehrerer Ereignisse untersucht (etwa A und B), kann es vorkommen, dass das Auftreten des Ereignisses A über Bedingungen an das Auftreten von B geknüpft ist, beispielsweise könnte A das Ziehen einer roten Karte und B das Ereignis Ziehen einer Dame sein. Eine Fragestellung, die auf bedingte Wahrscheinlichkeiten führt, wäre das Ziehen einer Dame unter der Bedingung, dass eine rote Karte gezogen wurde.

Es seien A und B zwei Ereignisse mit P(B)>0. Dann heißt

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

die bedingte Wahrscheinlichkeit von A unter (der Bedingung) B. Die bedingte Wahrscheinlichkeit ist die Wahrscheinlichkeit von A, nachdem B eingetreten ist, in obiger Motivation also die Wahrscheinlichkeit eine Dame zu ziehen, wenn man weiß,

dass die gezogene Karte schon rot ist. Wichtig bei der Anwendung ist zu erkennen, dass eine bedingte Wahrscheinlichkeit vorliegt und nicht das gemeinsame Auftreten (*cap*) vorliegt. Es gibt drei Sätze, welche das Behandeln und Rechnen mit bedingten Wahrscheinlichkeiten vereinfachen.

Satz 2.1. *Multiplikationssatz*
Es seien $A_1, A_2, \cdots, A_n$ Ereignisse. Dann gilt:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = PA_1 \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_2 \cap A_1) \cdots P(A_n \mid A_{n-1} \cap A_{n-2} \cap \cdots \cap A_2 \cap A_1)$$

Satz 2.2. *Satz von der totalen Wahrscheinlichkeit*
Es seien $A_1, A_2, \cdots, A_n$ Ereignisse einer Grundgesamtheit Ω mit $A_i \cap A_j = \emptyset$ für $i \neq j$ sowie $\bigcup_{i=1}^{n} A_i = \Omega$. Sei B ein beliebiges Ereignis über derselben Grundgesamtheit. Dann gilt:

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)$$

Ein letzter Satz, der eine gewisse Subjektivität in die Wahrscheinlichkeitstheorie bringt, ist der Satz von Bayes:

Satz 2.3. *Satz von Bayes*
Es seien
$A_1, A_2, \cdots, A_n$ Ereignisse in einer Grundgesamtheit Ω mit $A_i \cap A_j = \emptyset$ für $i \neq j$ sowie $\bigcup_{i=1}^{n} A_i = \Omega$. Sei B ein beliebiges Ereignis mit $P(B)>0$. Dann gilt:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)}$$

Die Wahrscheinlichkeit $P(A_i \mid B)$ heißt a posteriori Wahrscheinlichkeit für A_i. Sie ist nämlich die Wahrscheinlichkeit für A_i, nachdem B beobachtet wurde. Entsprechend heißt die Wahrscheinlichkeit $P(A_i)$ die a priori Wahrscheinlichkeit von A. Das Auftreten von B kann also als eine Zusatzinformation verstanden werden.

Folgendes Beispiel zeigt eine Anwendung des Satzes von Bayes:
Bei der Diagnose einer Krankheit, an der 10 Prozent der Bevölkerung leiden, werden Gesunde mit Wahrscheinlichkeit 0,01 als krank eingestuft, Kranke mit Wahrscheinlichkeit 0,9 richtig als krank eingestuft. Man kann sich nun fragen, mit welcher Wahrscheinlichkeit wird sich eine positive Diagnose (krank) als falsch herausstellen? Wie geht man an die Aufgabe heran? Zunächst bezeichne man die Gesunden durch G, die Kranken durch K. Es gilt dann nach Voraussetzung: PG= 0,9, PK=0,1. B sei das Ereignis, dass die Krankheit diagnostiziert wird. Dann gilt wieder nach Voraussetzung:$P(B|G) = 0,01, P(B|K) = 0,9$. Die Fragestellung ist jetzt die nach der Wahrscheinlichkeit $P(G|B)$, also dass ein Gesunder als krank diagnostiziert wird. Mit dem Satz von Bayes erhält man:

$$P(G|B) = \frac{P(B|G)P(G)}{P(B|G)P(G) + P(B|K)P(K)} = \frac{0,01 \cdot 0,9}{0,099} = 0,1$$

Die Tatsache, dass sich Ereignisse gegenseitig bedingen oder beeinflussen können, führt zu der folgenden Definition, in der dies nämlich gerade nicht der Fall ist.

Definition 2.4. *Unabhängigkeit von Ereignissen*
Zwei Ereignisse A, B heißen unabhängig, genau dann wenn

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

Äquivalent zu dieser Formulierung der Unabhängigkeit ist die Formulierung, welche sich aus der Definition der bedingten Wahrscheinlichkeit ergibt: Zwei Ereignisse A, B heißen unabhängig, wenn $P(A \mid B) = P(A)$ bzw. $P(B \mid A) = P(B)$. Der Nachweis der Unabhängigkeit zweier (oder mehrerer Ereignisse) geschieht über den Nachweis, dass die in der Definition angegebene Gleichheit erfüllt ist. Anschauliche Argumente dürfen hier nicht angewendet werden.

2.3. **Kombinatorik.** In diesem Abschnitt werden Rechenregeln angegeben, mit denen es möglich ist, die Wahrscheinlichkeit spezieller Ereigniskombinationen zu bestimmen. Hierbei müssen alle einelementigen Ereignisse einer Grundgesamtheit dieselbe Wahrscheinlichkeit besitzen. Speziell bei Würfelspielen oder Lotterien ist dies gegeben (im Idealfall). Bei einem fairen Würfel besitzt jede Zahl dieselbe Wahrscheinlichkeit, auch beim Lotto besitzt jede Kugel dieselbe Wahrscheinlichkeit, gezogen zu werden. Die interessierenden Ereignisse sind dann meist aus diesen Elemetarereignissen zusammengesetzt. Für ein Ereignis A heißen alle Ereignisse, bei denen A auftritt, günstig für A. Folgende Vorgehensweise liegt der Bestimmung dieser Wahrscheinlichkeiten zugrunde:

Man bestimme zunächst die Anzahl aller möglichen Ereigniskombinationen, dann die Anzahl der günstigen Fälle. Der Quotient aus diesen beiden Zahlen ist die gesuchte Wahrscheinlichkeit. Diese Wahrscheinlichkeit heißt dann auch Laplace - Wahrscheinlichkeit. Zum Herleiten der Formeln bietet sich das sog. Urnenmodell an. Hierbei werden aus einer Urne mit n verschieden- oder gleichfarbigen Kugeln k gezogen. Wird eine gezogene Kugel wieder zurückgelegt, ist das eine Ziehung mit Zurücklegen, ansonsten ohne Zurücklegen. Wird die Reihenfolge der Kugeln (oder deren Anordnung) berücksichtigt, ist es eine Ziehung mit Berücksichtigung der Anordnung, ansonsten ohne Berücksichtigung der Anordnung. Entsprechen gibt es 4 verschiedene Fälle (und damit Formeln) zu betrachten:

Regel 2.5. *Mit Zurücklegen, mit Berücksichtigung der Anordnung*
Alle Kugeln sind unterscheidbar (also z. B. nummeriert oder von unterschiedlicher Farbe). Es werden aus n Kugeln k gezogen. Die Reihenfolge der Kugeln wird berücksichtigt. Dann gibt es n^k Möglichkeiten.

Beispiel 2.6. *Sind 3 verschiedene Kugeln a,b,c gegeben, so gibt es $3^2 = 9$ Möglichkeiten, jeweils 2 zu ziehen und anzuordnen, wenn jede Kugel auch mehrmals vorkommen kann. Also : aa,ab,ac,bb,ba,bc,cc,ca,cb*

Regel 2.7. *Ohne Zurücklegen, mit Anordnung*
Aus n Kugeln weden k ohne Zurücklegen, aber mit Berücksichtigung der Anordnung gezogen. Dann gibt es $n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}$ Möglichkeiten.

Beispiel 2.8. *Sind 3 verschiedene Kugeln a,b,c gegeben, so gibt es bei dem jetztigen Ziehungsmodell die Möglichkeiten ab, ac, bc, ba, ca, cb, also $6 = 3 \cdot 2$ Möglichkeiten.*

Regel 2.9. *Ohne Zurücklegen, ohne Berücksichtigung der Anordnung*
Werden aus n Kugeln k ohne Zurücklegen und ohne Berücksichtigung der Anordnung gezogen, gibt es $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ Möglichkeiten. Der Ausdruck heißt auch Binomialkoeffizient.

Beispiel 2.10. *Bei 3 Kugeln gibt es bei diesem Ziehungsmodell die Möglichkeiten ab, ac, bc, also 3 Möglichkeiten.*
Ein anderes Beispiel ist das Ziehen der Lottozahlen 6 aus 49: Wieviele Möglichkeiten gibt es, aus 49 Zahlen 6 ohne Zurücklegen zu ziehen? Antwort: $\binom{(49}{6)}$ Möglichkeiten.

Regel 2.11. *Mit Zurücklegen, ohne Berücksichtigung der Anordnung*
Werden aus n Kugeln k mit Zurücklegen ohne Berücksichtigung der Anordnung gezogen, gibt es $\binom{n+k-1}{k}$ Möglichkeiten.

Beispiel 2.12. *Würde beim Lotto die gezogene Kugel jeweils wieder zurückgelegt, so gibt es* $\binom{49+6-1}{6}$ *Möglichkeiten.*

Regel 2.13. *Alle n unterscheidbaren Kugeln sollen auf k verschiedene Gruppen der Größe*
$k_1, k_2, \cdots, k_k$ *verteilt werden. Dann gibt es hierfür* $\binom{n!}{k_1! \cdot k_2! \cdots k_k!}$ *Möglichkeiten.*

Beispiel 2.14. *Auf wie viele Möglichkeiten lassen sich 9 Personen auf ein Zimmer mit 4 Betten und auf ein Zimmer mit drei Betten und auf ein Zimmer mit 2 Betten verteilen? Antwort: Es gibt* $\frac{9!}{4!3!2!} = 1260$ *Möglichkeiten.*

3. Zufallsvariable und deren Eigenschaften

3.1. **Zufallsvariable.** Die in Kapitel 2 vorgestellten Methoden der Wahrscheinlichkeitsrechnung reichen prinzipiell aus, die Fragestellungen der Statistik zu beantworten. Jedoch ist es bei bestimmten Fragestellungen nicht einfach, einen das Experiment beschreibenden Wahrscheinlichkeitsraum anzugeben. Betrachtet man das Experiment, 100 mal einen Würfel zu werfen, so wird man sicherlich nicht nur an der Frage interessiert sein, mit welcher Häufigkeit die einzelnen Augen auftreten, sondern auch an einer Frage der Art: Mit welcher Wahrscheinlichkeit tritt die Summe 85 auf? Mathematisch ist dies eine Transformation auf den Ergebnissen des Experimentes, denn man addiert die Augenzahlen. Der Begriff der Zufallsvariable vereinfacht die Beantwortung dieser Fragestellung.

Definition 3.1. *Es sei* Ω *eine Grundgesamtheit. Dann heißt eine Abbildung*

$$X : \Omega \to \mathbb{R}$$

mit

$$\omega \mapsto X(\omega) \in \mathbb{R}$$

eine Zufallsvariable. Der Wert $x{:=}X(\omega)$ *heißt Realisation der Zufallsvariablen.*

Bemerkung 3.2. *Eine Zufallsvariable X ist im strengen mathematischen Sinne zusätzlich noch eine meßbare Abbildung, wobei Meßbarkeit bedeutet, dass das Urbild jeder Borelmenge in* $\mathbb{R}$ *ein Element der* σ*-Algebra ist. Im Falle, dass als* σ*-Algebra die Potenzmenge gewählt wird, ist dies immer der Fall, so dass sich die Meßbarkeit in diesem Falle auf die Abbildungseigenschaft der Zufallsvariable reduziert. Deshalb genügt für das Folgende die obige Definition.*
Eine Zufallsvariable wird oft auch Merkmal genannt (etwa Merkmal Größe), die Realisation oft dann Merkmalsausprägung.

Beim Umgang mit Zufallsvariablen haben sich folgende Notationen als sehr nützlich (da abkürzend) erwiesen:

Bemerkung 3.3. *Für die Menge* $\{\omega \in \Omega \mid X(\omega) = x\} = X^{-1}(\{x\})$ *wird* $\{X = x\}$ *geschrieben. Analog bildet man die Mengen* $\{X \leq x\}, \quad \{X \geq x\}, \quad \{X \neq x\}, \quad \{X > x\}, \quad \{X < x\}$

3.2. **Eigenschaften von Zufallsvariablen.**

Definition 3.4. *Eine Zufallsvariable heißt diskret, wenn sie nur endlich oder abzählbar viele Realisationen besitzt.*

Bemerkung 3.5. *Eine Zufallsvariable, welche z.B. das Würfeln mit einem Würfel beschreibt, ist diskret.*

Die Natur von Zufallsvariablen bringt es mit sich, dass man ihre Realisation nicht vorhersagen kann. Dennoch möchte man irgendeine Auskunft über ihr Verhalten gewinnen. Dies geschieht durch Wahrscheinlichkeitsaussagen über ihre Realisationen. Hierbei unterscheidet man zwischen diskreten und stetigen Zufallsvariablen. Bei einer diskreten ZV X macht es Sinn zu fragen, mit welcher Wahrscheinlichkeit sie einen Wert $a \in \Re$ annimmt, also $P(X = a) =?$. Analog kann man dann auch fragen $P(X \leq a) =?, \quad P(X > a) =?...$
Um diese Werte zu bestimmen, macht man sich zunutze, dass jede Zufallsvariable eine Abbildung darstellt und es daher eine eindeutige Zuordnung zwischen den Werten im Bildbereich (hier a) und den Werten im Definitionsbereich (hier die Grundgesamtheit) gibt. Man definiert:

$$P(X = a) = P(x \in \Omega \mid X(x) = a)$$

Die Wahrscheinlichkeit der GS Ω wird mittels X also in den Bildbereich transportiert. Man spricht in diesem Zusammenhang auch vom Bildmaß einer ZV X. Die Realisationen x_i mit $p_i := P\{X = x_i\} > 0$ (i=1, ..,n) heißen Trägerpunkte von X. Die Angabe der Trägerpunkte samt ihrer Wahrscheinlichkeit heißt Verteilung von X. Neben der Verteilung gibt es noch eine andere wichtige Charakterisierung einer Zufallsvariablen: ihre Verteilungsfunktion.

Definition 3.6. *Es sei X eine diskrete Zufallsvariable mit Realisationen $x_1, x_2, \ldots, x_n$. Die Funktion*

$$F_X(x) := P\{X \leq x\}$$

heißt Verteilungsfunktion von X.

Folgende Eigenschaften lassen sich hieraus ableiten:

Satz 3.7. *Für die Trägerpunkte einer Zufallsvariablen gilt stets:*

(1) $\sum_{i=1}^{\infty} p_i = 1$.

(2) *Für die Verteilungsfunktion gilt* $F_X(x) = P\{X \leq x\} = \sum_{x_i \leq x} p_i$.

Auf die Eigenschaften von Verteilungsfunktionen soll hier nicht eingegangen werden.

Beispiel 3.8. *Die Zufallsvariable X beschreibe die Summe der Augen beim Würfeln mit zwei Würfeln. Dann hat man die folgenden Realisationen:*

x	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Erzeuger	*(1,1)*	*(1,2)*	*(1,3)*	*(1,4)*	*(1,5)*	*(1,6)*	*(2,6)*	*(3,6)*	*(4,6)*	*(5,6)*	*(6,6)*
		(2,1)	*(2,2)*	*(2,3)*	*(2,4)*	*(2,5)*	*(3,5)*	*(4,5)*	*(5,5)*	*(6,5)*	
			(3,1)	*(3,2)*	*(3,3)*	*(3,4)*	*(4,4)*	*(5,4)*	*(6,4)*		
				(4,1)	*(4,2)*	*(4,3)*	*(5,3)*	*(6,3)*			
					(5,1)	*(5,2)*	*(6,2)*				
						(6,1)					
Punktmassen	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Man erhält für die Verteilungsfunktion

$$F_X(x) := P\{X \leq x\} = \begin{cases} 0 & : \quad x < 2 \\ \frac{1}{36} & : \quad 2 \leq x < 3 \\ \frac{3}{36} & : \quad 3 \leq x < 4 \\ \cdots & : \quad \cdots \\ 1 & : \quad 12 \leq x \end{cases}$$

Neben den diskreten Zufallsvariablen gibt es noch die stetigen Zufallsvariablen:

Definition 3.9. *Eine Zufallsvariable X heißt stetig, wenn es eine Abbildung*

$$f_X : \mathbb{R} \longrightarrow \mathbb{R}$$

gibt mit den folgenden 3 Eigenschaften:

(1) $f_X \leq 0 \quad \forall x \in \mathbb{R}$

(2) $\int_a^b f_X(x)dx$ *existiert für alle* $a, b \in \mathbb{R} \cup \{-\infty; \infty\}$

(3) $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$

Die Abbildung f_X heißt Dichte (-funktion) der Zufallsvariablen X.

Ist also eine Dichte f_X zu einer Zufallsvariablen X gegeben, so gilt für die Wahrscheinlichkeit und damit auch für die Verteilungsfunktion von X:

$$P\{X \leq x\} = \int_{-\infty}^{x} f_X(t)dt$$

Geometrisch ist dies gerade der Flächeninhalt unter dem Graphen von f_X von $-\infty$ bis x.

Bemerkung 3.10. *Bei stetigen Zufallsvariablen X ist die Wahrscheinlichtkeit für eine bestimmte Realisation x immer gleich 0, also $P(X = x) = 0$ für jedes x aus den reellen Zahlen.*

Beispiel 3.11. *Sei X eine Zufallsvariable mit Dichte*

$$f_X(t) := \begin{cases} 6 \cdot t \cdot (1-t) & : \quad 0 \leq t \leq 1 \\ 0 & : \quad sonst \end{cases}$$

Dann ist die Verteilungsfunktion von X gegeben durch

$$F_X(x) = \int_{-\infty}^{x} 6 \cdot t \cdot (1-t)dt = \begin{cases} 3 \cdot x^2 - 2 \cdot x^3 & : \quad 0 \leq x \leq 1 \\ 0 & : \quad sonst \end{cases}$$

Unter Berücksichtigung des Additionstheorems für Wahrscheinlichkeiten erhält man weiterhin eine Möglichkeit, die folgende Wahrscheinlichkeit zu berechnen:

$$P\{a \leq X \leq b\} = \int_a^b f_X(t)dt$$

Bemerkung 3.12. *Für die Standardnormalverteilung (die ja eine stetige Verteilung ist) sind die Werte der Verteilungsfunktion Φ tabelliert und müssen daher nicht berechnet werden.*

Bemerkung 3.13. *Die Angabe der Trägerpunkte einer Zufallsvariablen zusammen mit den Wahrscheinlichkeiten ihres Auftretens (bei diskreten Zufallsvariablen) bzw. die Angabe der Dichte (bei stetigen Zufallsvariablen) heißt Verteilung der Zufallsvariablen.*

Verteilungen werden durch die Verteilungsfunktion eindeutig charakterisiert, doch ist es wünschenswert, auch andere Kennzahlen einer Verteilung zu besitzen. Die wichtigsten Kennzahlen sind die, welche schon im Abschnitt Deskriptive Statistik genannt wurden.

Definition 3.14. *Es sei X eine Zufallsvariable.*

(1) *Die Realisation, welche am häufigsten vorkommt, heißt Modus oder Modalwert von X.*

(2) *Der Wert m mit*

$$P\{X \leq m\} \geq \frac{1}{2}$$

und

$$P\{X \geq m\} \geq \frac{1}{2}$$

heißt Median von X.

(3) *Die Zahl E(X) mit*

$$E(X) := \begin{cases} \sum_{i=1}^{\infty} x_i \cdot p_i & : \quad X \text{ diskret mit Realisationen } x_i \text{ und Punktmassen } p_i \\ & : \\ \int_{-\infty}^{\infty} x \cdot f_X dx & : \quad X \text{ stetig mit Dichte } f_x \end{cases}$$

heißt der Erwartungswert von X. Man schreibt statt EX auch oft μ.

(4) *Die Zahl $\sigma^2 := VarX := E(X - EX)^2$ heißt Varianz von X. Die Zahl $\sigma = \sqrt{VarX}$ heißt Standardabweichung von X.*

Bemerkung 3.15. *Die bei einer Zufallsvariablen sinnvoll bildbaren Kennzahlen sind vom sogenannten Skalenniveau der Zufallsvariablen abhängig (s.u.). Bei einer endlichen Menge ist der Median bei einer ungeraden Anzahl von Werten gerade der mittlere Wert, wenn man die Werte der Größe nach notiert (wobei mehrfach vorkommende Werte entsprechend ihrer Vielfachheit hingeschrieben werden), bei einer geraden Anzahl von Werten das arithmetische Mittel aus den beiden mittleren Werten. Die Varianz einer Zufallsvariablen ist stets > 0.*

Für das obige Beispiel des Würfeln mit zwei Würfeln und der Addition der Augenzahlen erhält man den Erwartungswert 7, beim Würfeln mit einem Würfel ist der Erwartungswert 3,5. Der Erwartungswert muss also nicht immer zur Menge der Realisationen einer Zufallsvariable X gehören.

Folgende Rechenregeln gelten für Erwartungswert und Varianz einer Zufallsvariablen:

Satz 3.16. *Es seien X und Y Zufalssvariablen, a eine reelle Zahl und g eine stetige Funktion. Dann gilt:*

(1) *E(X+Y)= EX + EY*

(2) *E(aX)= aE(X)*

(3) *E(a)=a*

(4) *Var(a+X)= Var (X)*

(5) *Var (aX)= a^2 Var(X)*

(6) *Var(X)= $E(X - EX)^2 = EX^2 - (EX)^2$*

(7) *E(gX)= $\sum_{x_i} g(x_i) \cdot p_i$ für diskretes X und E(gX)= $\int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$*

Beispiel 3.17. *Man betrachte die Funktion g(x)=x^2. Beschreibt X das Würfeln mit einem fairen Würfel (mit Erwartungswert 3,5), so gilt $E(gX) = \sum_{i=1}^{6} i^2 \frac{1}{6} =$ 15,166. Weiterhin ermöglicht .7 auch das Berechnen der Varianz einer Zufallsvariablen durch das Betrachten der Funktion $g(x) = (x - EX)^2$.*

Ist X eine Zufallsvariable mit $EX = \mu$ und Varianz σ^2, so kann die standardisierte Zufallsvariable Z berechnet werden:

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

Z besitzt dann den Erwartungswert 0 und die Varianz 1. Da bei vielen Zufallsvariablen X die zugehörige standardisierte Zufallsvariable Z dieselbe Verteilung wie X besitzt, kann sich bei der noch zu behandelnden Tabellierung von Zufallsvariablen auf eine Tabelle beschränkt werden.

3.3. **Skalenniveaus von Zufallsvariablen.** Durch Zufallsvariablen können viele Erscheinungen des täglichen Lebens beschrieben werden. So etwa die Bestimmung der Größe einer beliebig ausgewählten Person, die Temperatur an einem beliebigen Tag, die Farbe eines vorbeifahrenden Autos, etc. Betrachtet man jedoch die Realisationen der einzelnen Zufallsvariablen (Größe, Geschwindigkeit, Temperatur, Farbe), so weisen diese qualitativ Unterschiede folgender Art auf. Sind die Farben vorbeifahrender Autos durch Zahlen 1,2,3,.. codiert, so stellen diese nur eine Umbenennung dar, keinesfalls können mit diesen Werten irgendwelche arithmetischen Operationen vorgenommen werden. Mir der Temperatur verhält es sich so, dass man nicht sagen kann, dass an irgendeinem Tag die doppelte Temperatur geherrscht habe. Diese qualitativen Unterschiede werden in den sogenannten Skalenniveaus von Zufallsvariablen beschrieben. Das Skalenniveau hat Auswirkungen auf die möglichen erlaubten arithmetischen Operationen mit Zufallsvariablen und damit auf die möglichen statistischen Analysemethoden. Skalenniveaus wurden in der deskriptiven Statistik für Merkmale definiert. Für Zusallsvariable gilt dieselbe Art der Klassifizierung.

Definition 3.18. *Es sei X eine Zufallsvariable mit Realisationen x. Dann heißt X nach einem der nachfolgend genannten Skalenniveaus skaliert, wenn die Realisationen die zum jeweiligen Skalenniveau gehörenden Eigenschaften besitzen.*

(1) *nominal: Objekten mit gleicher Mermalsausprägung werden gleiche Zahlen zugeordnet, Objekten mit unterschiedlicher Merkmalsausprägung verschiedene Zahlen, ein Beispiel ist die Wahl einer Farbe*

(2) *ordinal: die Zuordnung von Zahlen geschieht so, dass das Objekt mit der größeren (besseren) Merkmalsausprägung die größere Zahl erhält, ein Beispiel sind Schulnoten*

(3) *intervallskaliert: die Unterschiede zwischen zwei Merkmalsausprägungen entsprechen genau der Differenz der Zuordnung von Zahlen zu diesen Merkmalsausprägungen, ein Beispiel ist die Temperatur*

(4) *verhältnisskaliert: den Merkmalsausprägungen werden Zahlen so zugeordnet, dass das Verhältnis dieser Zahlen gerade dem Verhältnis der Merkmalsausprägungen entspricht, ein Beispiel ist das Alter einer Person.*

Intervall- und Verhältnisskalen werden auch als metrische Skalen bezeichnet. Folgende Tabelle zeigt die Operationen bzw. Kennzahlen, welche mit den jeweiligen Skalenniveaus möglich bzw. bildbar sind:

Skalenniveau	*arithm. Operation*	*Kennzahl*
nominal	*Gleichheit*	*Modus*
ordinal	*größer-kleiner Relation*	*Modus*
intervall	*Differenzen, Summen*	*Mittelwert*
verhältnis	*Quotienten*	*alle*

Natürlich sind alle Kennzahlen, die auf einem niedrigeren Skalenniveau möglich sind, auch auf den höheren bildbar. Liegt bei einer mindestens ordinalskalierten Zufallsvariable eine ungerade Anzahl von Realisationen vor, so erhält man den Median als den nach Sortierung der Größe nach in der Mitte stehenden Wert. Bei einer geraden Anzahl von Realisationen ist der Median eine beliebige zwischen den beiden

mittleren Werten stehende Zahl, in der Regel nimmt man das arithmetische Mittel dieser beiden Werte.

3.4. **Unabhängigkeit von Zufallsvariablen.** Bei der Modellierung von Experimenten mittels mehrerer Zufallsvariablen steht die Frage der Beziehung zwischen den einzelnen Zufallsvariablen im Vordergrund. Die wichtigste zu klärende Frage ist die, ob die Zufallsvariablen unabhängig oder nicht unabhängig sind. Das Konzept der Unabhängigkeit von Ereignissen liegt dem Konzept der Unabhängigkiet bei Zufallsvariablen zugrunde. Es genügt das Konzept für den Fall zweier Zufallsvariablen X und Y zu erläutern. Da man es mit Realisationen zweier Zufallsvariablen zu tun hat, benötigt man eine Notation für die Realisation X nimmt den Wert x und Y den Wert y an. Hierfür schreibt man:

$$P\{X = x, Y = y\}$$

Im Falle des Würfelns (ZV X, Realisationen 1,..,6) und gleichzeitigen Müntewerfens (ZV Y, Realisationen K oder Z) heißt das in obiger Notation, wenn X den Wert 4 annimmt und Y Zahl zeigt

$$P\{X = 4, Y = Z\}$$

Wie sich diese wahrscheinlichkiet berechnen lässt, klärt folgende Definition:

Definition 3.19. *Seien X und Y diskrete Zufallsvariablen mit Trägerpunkten* $\{x_1, x_2, ..., x_n\}$ *bzw.* $\{y_1, y_2, ..., y_m\}$.
X und Y heißen unabhängig genau dann,
wenn für jedes i und j: $P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$

Analog definiert man dies für n Zufallsvariable. Die Prüfung auf Unabhängigkeit erfordert wie bei Ereignissen das Nachprüfung der Definition. Im Falle stetiger Zufallsvariablen ersetzt man das Produkt der Wahrscheinlichkeiten für die Tragerpunkte durch das Produkt der Dichten.

Definition 3.20. *Zwei Zufallsvariablen X und Y mit Dichten* f_X *bzw.* f_Y *heißen unabhängig genau dann, wenn gilt:*
$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} f_X(t)dt \cdot \int_{-\infty}^{y} f_Y(t)dt$

Beispiel 3.21. *Die Annahme der Unabhängigkeit von Zufallsvariablen ermöglicht in vielen Fällen erst die Berechnung von Wahrscheinlichkeiten. Ein fairer Würfel werde n mal geworfen und es soll die Wahrscheinlichkeit für m-mal die Sechs berechnet werden. Man modelliert wie folgt: Der i-te Wurf werde durch eine Zufallsvariable* X_i *modelliert mit zwei Realisationen X=1, wenn 6 fällt, X=0 wenn keine 6 fällt. Jede einzelen Zufallsvariable ist dann wie folgt verteilt:*
$P(X =_i = 1) = \frac{1}{6}$ *und* $P(X_i = 0) = \frac{5}{6}$
Jedes X_i *ist also alternativ verteilt.*
Nimmt man nun an, dass die X_i *unabhängig voneinander sind, so beträgt die Wahrscheinlichkeit für m-mal 6 gerade* $\frac{1}{6}^m \cdot \frac{5}{6}^{n-m}$. *Fragt man sich nach der Anzahl der 6-en bei n Würfen, so wird dies durch Bildung der ZX* $X = \sum X_i$ *beantwortet. Die Zufallsvariable X ist dann Binomial-verteilt. Dies ist nur dann der Fall, wenn die einzelnen* X_i *unabhängig voneinander sind.*

Für die Berechnung des Erwartungswertes zweier Zufallsvariablen X und Y gilt im Falle der Unabhängigkeit

$$E(X \cdot Y) = EX \cdot EY$$

4. Verteilungen

Es werden nun einige wichtige Verteilungen über die oben genannten Kennzahlen, Trägerpunkte oder Dichten definiert. Beispiele zu einigen dieser Verteilungen werden in der Vorlesung behandelt.

Definition 4.1. *Diskrete Verteilungen*

(1) *Diskrete Gleichverteilung*
Trägerpunkte: $x_i \in \mathbb{R}, i = 1, 2, 3, \ldots, x_i \neq x_j \quad$ *für* $\quad i \neq j.$

Punktwahrscheinlichkeiten: $p_i = P(X = x_i) = \frac{1}{n}$
$EX = \frac{1}{n}\sum_{i=1}^{n} x_i; VarX = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$
Speziell für n=1 liegt eine Einpunktverteilung auf x_1 *vor.*

(2) *Alternativverteilung (Zweipunktverteilung)*
Trägerpunkte: $i \in \{0, 1\}$
Punktwahrscheinlichkeiten:
$p = P(X = 1); 1 - p = P(X = 0); 0 < p < 1.$
$EX = p; VarX = p(1 - p)$

(3) *Binomialverteilung*
Trägerpunkte: $i \in \{0, 1, \ldots, n\}$
$p_i = P(X = i) = \binom{n}{i} \cdot p^i(1 - p)^{n-i} =: b(i \mid n, p), 0 < p < 1.$
$EX = np; VarX = np(1 - p)$

(4) *Hypergeometrische Verteilung*
Trägerpunkte:
$i \in \mathbb{N} \cup \{0\}$ *mit n,M,N* $\in \mathbb{N} \cup \{0\}$

$max\{0, n + M - N\} \leq i \leq min\{n, M\}, n \leq N, M \leq N$

$p_i = P(X = i) = \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}}$

$EX = n \cdot \frac{M}{N}; VarX = n \cdot \frac{M}{N}(1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$

(5) *Geometrische Verteilung*
Trägerpunkte: $i \in \mathbb{N} \cup \{0\}$
$p_i = P(X = i) = p(1 - p)^i; 0 < p < 1$
$EX = \frac{p}{1-p}; VarX = \frac{p}{(1-p)^2}$

(6) *Poisson-Verteilung Trägerpunkte:* $i \in \mathbb{N} \cup \{0\}$
$p_i = P(X = i) = \frac{\lambda^i}{i!} \cdot e^{-\lambda}; \lambda > 0.$
$EX = \lambda; VarX = \lambda$

Definition 4.2. *Stetige Verteilungen*

(1) *Stetige Gleichverteilung*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & : \quad a \leq x \leq b; a, b \in \mathbb{R}; a < b \\ 0 & : \quad sonst \end{cases}$$

$EX = \frac{a+b}{2}; VarX = \frac{(b-a)^2}{12}$

(2) *Gaußverteilung N(0,1); Standardnormalverteilung*

$$\varphi_X(x) = f_X(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}} \quad \textit{für} \quad -\infty < x < \infty$$

Für die Verteilungsfunktion gilt: $\Phi_X(x) = \int_{-\infty}^{x} \varphi_X(y)dy$

$EX = 0; VarX = 1.$

(3) *Normalverteilung* $N(\mu, \sigma^2)$ *mit* $\sigma > 0$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot(\frac{x-\mu}{\sigma})^2} \quad \textit{für} \quad -\infty < x < \infty.$$

$EX = \mu; VarX = \sigma^2.$

(4) *Exponentialverteilung*
$a>0$

$$f_X(x) = \begin{cases} a \cdot e^{-ax} & : \quad x \geq 0 \\ 0 & : \quad sonst \end{cases}$$

$EX = \frac{1}{a}; VarX = \frac{1}{a^2}$

(5) *Chi-Quadrat-Verteilung*
Es seien $X_1, \cdots, X_n$ *paarweise unabhängige N(0,1) verteilte Zufallsvariablen. Dann heißt die Zufallsvariable*

$$\chi_n^2 := X_1^2 + \cdots + X_n^2$$

χ^2*-verteilt mit n Freiheitsgraden.*

(6) *t-Verteilung Es seien eine N(0,1)-verteilte Zufallsvariable X und eine* χ_n^2 *Zufallsvariable* χ_n^2 *gegeben.Dann heißt die Zufallsvariable*

$$t_n := \frac{X}{\sqrt{\frac{\chi_n^2}{n}}}$$

t-verteilt mit n Freiheitsgraden.

5. Mehrdimensionale Zufallsvariablen

Am Ende des Abschnitts über Zufallsvariablen wurde die Unabhängigkeit von Zufallsvariablen definiert. Hierbei trat sowohl der Fall mehrerer Zufallsvariablen über verschiedenen Grundgesamtheiten (Würfeln und Münze werfen) auf wie auch der Fall, dass alle Zufallsvariablen über derselben Grundgesamtheit definiert waren. Dieser Fall ist der nun interessierende Fall. In vielen Experimenten ist man nicht nur an einer Merkmalsausprägung einer Person (etwa Gewicht) interessiert, sondern auch noch an anderen Merkmalsausprägungen wie etwa Größe, Haarfarbe etc. Man stellt an einem Untersuchungsgegenstand gleich mehrere Merkmale fest. Das Ergebnis ist somit eine bestimmte Anzahl von Meßwerten pro Untersuchungsgegenstand. Auch beim Würfeln mit zwei Würfeln benötigt man ein Paar von Werten, um das Experiment zu beschreiben, somit benötigt man auch eine Zufallsvariable, welche Werte in einem zweidimensionalen Raum annimmt. Diese Tatsache führt zum Begriff der mehrdimensionalen Zufallsvariablen.

Definition 5.1. *Es sei eine Grundgesamtheit Ω gegeben.Es seien*

$$X_1, \cdots, X_n : \Omega \longrightarrow \mathbb{R}$$

n Zufallsvariablen.
Dann heißt der Vektor $X := (X_1, X_2, \cdots, X_n)$ ein n-dimensionaler Zufallsvektor oder n-dimensionale Zufallsvariable. Der Vektor aus den Realisationen der einzelnen Zufallsvariablen heißt Realisation der Zufallsvariable $X = (X_1, X_2, \cdots, X_n)$.

Im einführend genannten gleichzeitigen Bestimmen von Gewicht, Größe und Haarfarbe einer Person ist die Grundgesamtheit Ω durch die Menge aller Personen, welche untersucht werden könnten bestimmt. X_1 bestimmt dann das Gewicht, X_2 die Größe und X_3 die Haarfarbe.

Analog zum eindimensionalen Fall gibt es die Begriffe Verteilungsfunktion und Dichte. Zur besseren Darstellung wird im weiteren Verlauf dieses Kapitels n=2 gesetzt, d.h. es werden nur zweidimensionale Zufallsvariable betrachtet. Auch wird sich hier auf den Fall von diskreten Zufallsvariablen beschränkt.

Definition 5.2. *Diskrete Zufallsvariable, Trägerpunkte, Verteilungsfunktion*
Ist der Ergebnisraum einer Zufallsvariable (X_1, X_2) abzählbar, so heißt die Zufallsvariable diskret. Die Werte $p_{ij} := P\{X_1 = x_i, X_2 = x_j\}$ heißen Punktmassen und die Punkte (x_i, x_j) mit $p_{ij} > 0$ heißen Trägerpunkte der Zufallsvariable. Die Funktion

$$F_X : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

mit

$$F_X(x_1, x_2) := P\{X_2 \leq x_1, X_2 \leq x_2\}$$

heißt Verteilungsfunktion der Zufallsvariablen.

Zweidimesnionale Zufallsvariablen lassen sich gut in Form einer Tabelle, einer sog. Kontingenztabelle, beschreiben. Zeilen- und Spaltenüberschriften beinhalten die möglichen Realisationen der Zufallsvariablen, die Einträge innerhalb der Tabelle stellen die Wahrscheinlichkeiten für das Auftreten einer Realisation P(X=x, Y=y) dar.

X	0	1	2
Y			
1	0,1	0,2	0,05
2	0,1	0,3	0,1
3	0,05	0,03	0,07

Bei mehrdimensionalen Zufallsvariablen besteht die Möglichkeit, eine oder mehrere der Variablen konstant zu halten und die anderen variabel zu lassen. Dieses Vorgehen entspricht einer Betrachtung der Verteilungen der einzelnen Komponenten einer mehrdimensionalen Zufallsvariablen Die Verteilung der Komponenten wird als Randverteilung bezeichnet.

Definition 5.3. *Es sei $X = (X_1, X_2)$ eine zweidimensionale Zufallsvariable. Ist X diskret, so seien die Punktmassen gegeben durch: $p_{ij} = P\{X_1 = x_i, X_2 = x_j\}$ (i,j jeweils aus einer abzählbaren Indexmenge I bzw. J) Dann heißt für $i \in I$ die Summe $\sum_j p_{ij} =: p_{i\cdot}$ die Randverteilung von X_1 und analog für $j \in J$ die Summe $\sum_i p_{ij} =: p_{\cdot j}$ die Randverteilung von X_2.*

Das heißt für obige Zufallsvariable (X,Y):

X	0	1	2	
Y				
1	0,1	0,2	0,05	$0{,}35 = p_{1\cdot}$
2	0,1	0,3	0,1	$0{,}5 = p_{2\cdot}$
3	0,05	0,03	0,07	$0{,}15 = p_{3\cdot}$
	0,25	0,53	0,22	
	$= p_{\cdot 1}$	$= p_{\cdot 2}$	$= p_{\cdot 3}$	

Der Wert $p_{1\cdot}$ gibt $P(Y = 1)$ an, entsprechend $p_{\cdot 2}$ die Wahrscheinlichkeit $P(X = 1)$ an. Klar, dass Zeilen- bzw. Spaltensummen sich jeweils zu 1 addieren.

Wie bei Ereignissen, welche sich gegenseitig bedingen können, können auch die Ergebnisse von Zufallsvariablen voneinander abhängen. Während man sich bei den Randverteilung nicht für die Werte einer Komponente der Zufallsvariablen interessiert, spielen diese bei der Bildung bedingter Wahrscheinlichkeiten durchaus eine Rolle, da man sich für die Wahrscheinlichkeit der Zufallsvariable unter der Bedingung, dass eine Komponente einen bestimmten Wert annimmt, interessiert. Hierbei spielen die Randverteilungen eine wichtige Rolle.

Definition 5.4. *Es sei (X, Y) eine zweidimensionale Zufallsvariable mit Punktmassen $p_{ij} := P\{X = x_i, Y = y_j\}$ (i=1, . . . , n; j= 1, . . . , m). Für ein festes j_0 gelte weiterhin $p_{\cdot j_0} = \sum_{i=1}^{n} p_{ij_0} > 0$. Dann heißt*

$$P\{X = x_i \mid Y = y_{j_0}\} := \frac{P\{X = x_i, Y = y_{j_0}\}}{P\{Y = y_{j_0}\}} (i = 1, \ldots n)$$

die bedingte Verteilung von X. Analog wird die bedingte Verteilung von Y definiert.

Ein weiterer wichtiger Begriff im Zusammenhang mehrdimensionaler Zufallsvariablen ist die Unabhängigkeit von Zufallsvariablen. Die Definition entspricht der im Abschnitt Zufallsvariablen formulierten Unabhängigkeit, jetzt aber auf den speziellen Fall der hier betrachteten Zufallsvariablen bezogen.

Definition 5.5. *Es sei Z:=(X,Y) ein Zufallsvektor . Ist Z diskret, so heißen X und Y unabhängig genau dann, wenn*

$$P\{X = x_i, Y = y_j\} = P\{X = x_i\}P\{Y = y_j\} \text{für alle Paare} (x_i, y_j).$$

Die hier behandelten bedingten Verteilungen treten bei Stochastischen Prozessen wieder auf. Insbesondere liegt dann aber der Fall vor, dass die Zufallsvariablen nicht unabhängig sind. Hier bedient man sich dann der folgenden Gleichheit, welche im Fall der Anwnedung auf Sachverhalte der Sprachmodellierung (Bigramme) eine Auflösung ermöglicht. Es gilt:

$$P(X = x_i, Y = y_j) = P(X = x_i \mid Y = y_j) \cdot P(Y = y_j)$$

In Kurzform schreibt man: $P(X,Y) = P(X \mid Y) \cdot P(Y)$ ohne auf spezielle Realisationen einzugehen.

6. Maximum Likelihood Schätzung

Während in den vorausgegangenen Abschnitten die Verteilung der Grundgesamtheit oder der Zufallsvariablen als bekannt vorausgestzt wurden, wird in diesem Abschnitt eine unbekannte Verteilung angenommen, über die man sich durch geeignete Verfahren Informationen beschaffen will, um dann konkrete Aussagen über diese Verteilung machen zu können. Das Mittel hierzu sind Stichproben. Eine Stichprobe stellt im Prinzip eine Auswahl von Elementen der Grungesamtheit dar. An den gezogenen Elementen stellt man die interessierenden Eigenschaften fest. Man schreibt $X_1, \cdots, X_n$ für eine Stichprobe vom Umfang n. Da Stichprobenergebnisse Zufallsergebnisse sind schreibt man hierfür also groše X. Ein X_i bezeichnet das Ziehen des i-ten Elementes. Die Realisationen werden dann durch $x_1, \cdots, x_n$ bezeichnet. Eine Stichprobe wird immer in Bezug zu einem interessierenden Merkmal Y gezogen (Y ist ebenfalls Zufallsvariable). Daher spreicht man oft auch von einer einfachen Stichprobe X zu einer Zufallsvariable Y. Mit Hilfe der Stichprobe sollen Eigenschaften der Grundgesamtheit bestimmt werden. Stichproben sollten nicht beliebig gezogen werden, sondern nach festen Regeln. Die wichtigste Art einer Stichprobe ist die sog. Einfache Stichprobe. Hierbei wird verlangt, dass die einzelnen Ziehungen der X_i undabhängig sind und jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit besitzt, in die Stichprobe zu gelangen. Die $X_1, \cdots, X_n$ können als Zufallsvariablen addiert werden oder es kann auch ein arithmetisches Mittel berechnet werden. Diese Zahlen dienen dann dazu, Aussagen über die Parameter einer Verteilung zu ermöglichen. Viele Verteilungen besitzen Parameter (Nomalverteilung $N(\mu;\sigma^2)$, Exponentialverteilung $\exp(\lambda)$, etc. Aus den Informationen der Stichprobenergebnisse soll hier jetzt eine Aussage über den Wert des Parameters getroffen werden. Dies heišt auch Schätzung des Parameters. Eine Möglichkeit, hier zu einer Aussage zu gelangen, ist die Maximum-Likelihood-Methode. Das Grundprinzip besteht darin, den Parameter so zu schätzen, dass die Wahrscheinlchkeit für einen vorleiegenden Stichprobenbefund maximal wird. Es liegt somit ein Extremwertproblem vor. Dies wird analytisch dadurch gelöst, dass eine passende Funktion maximiert wird, indem die erste Ableitung dieser Funktion =0 gesetzt wird. Da dies in dem beschriebenen Fall sehr aufwendig wird, naximiert man statt der Funktion die logarithmierte Funktion.

Zunächst soll eine diskrete Zufallsvariable betrachtet werden.

Definition 6.1. *Es sei Y eine diskrete Zufallsvariable mit Trägerpunkten y_i und Punktmassen $P\{Y = y_i, \theta\} =: p(i \mid \theta)(i = 1, 2, \ldots)$. Es sei $X = (X_1, X_2, \ldots, X_n)$ eine einfache Stichprobe zu Y mit Realisationen $(x_1, \ldots, x_n)$. Dann heißt die Funktion*

$$L(\theta \mid x_1, \ldots, x_n) := p(x_1 \mid \theta)p(x_2 \mid \theta) \cdots p(x_n \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$$

die Likelihood Funktion (LF) von X. Ein $\widehat{\theta} \in \Theta$ heißt Maximum-Likelihood- Schätzwert (ML-Schätzwert) für θ, wenn

$$L(\widehat{\theta} \mid x_1, \ldots, x_n) = \max_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n)$$

Die Berechnung einer ML-Schätzung wird sehr vereinfacht, wenn man folgenden Satz benutzt:

Satz 6.2. *Es sei $L(\cdot \mid x_1, \ldots, x_n)$ die Likelihood-Funktion zu einer Stichprobe. Ein $\widehat{\theta}$ ist genau dann ML-Schätzwert, wenn*

$$lnL(\widehat{\theta} \mid x_1, \ldots, x_n) = \max_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n)$$

Die Berechnung des Maximums des Produktes der Punktmassen wird hierdurch in der Regel auf die Berechnung des Maximums einer Summe zurückgeführt.

Ist Y nun eine stetige Zufallsvariable, so gilt die obige Definition entsprechend, nur muss statt der Punktmassen die Dichtefunktion von Y genommen werden.

Definition 6.3. *Es sei Y eine stetige Zufallsvariable mit Dichte $f_Y(\cdot \mid \theta)$. Es sei $X = (X_1, X_2, \ldots, X_n)$ eine einfache Stichprobe zu Y mit Realisationen $(x_1, \ldots, x_n)$. Dann heißt die Funktion*

$$L(\theta \mid x_1, \ldots, x_n) := f_Y(x_1 \mid \theta) f_Y(x_2 \mid \theta) \cdots f_Y(x_n \mid \theta) = \prod_{i=1}^{n} f_Y(x_i \mid \theta)$$

die Likelihood Funktion (LF) von X. Ein $\widehat{\theta} \in \Theta$ heißt Maximum-Likelihood- Schätzwert (ML-Schätzwert) für θ, wenn

$$L(\widehat{\theta} \mid x_1, \ldots, x_n) = \max_{\theta \in \Theta} L(\theta \mid x_1, \ldots, x_n)$$

Zwei Beispiele mögen die Vorgehensweise bei der Berechnung einer ML-Schätzung verdeutlichen:

Beispiel 6.4. *Stets sei $X = (X_1, X_2, \ldots, X_n)$ eine einfache Stichprobe zu einer Zufallsvariablen Y mit Realisationen $(x_1, \ldots, x_n)$.*

Y sei zunächst alternativ verteilt mit Parameter $p \in]0;1[$. Dann erhält man die Likelihood-Funktion :

$$L(p \mid x_1, \ldots, x_n) = p^{k_0} \cdot (1-p)^{n-k_0}$$

wobei bei der Stichprobe k_0 mal der Wert 1 aufgetreten sei. Weiter erhält man:

$$lnL(p \mid x_1, \ldots, x_n) = k_0 \cdot ln(p) + (n-k) \cdot ln(1-p)$$

Dies muss jetzt noch bezüglich p maximiert werden.

$$\frac{dlnL}{dp}(p \mid x_1, \ldots, x_n) =$$

$$\frac{k_0}{p} + \frac{(n-k)(-1)}{1-p} = \frac{k_0 - np}{p(1-p)}$$

Nullsetzen der ersten Ableitung liefert

$$\widehat{p} = \frac{k_0}{n}$$

also genau den Anteil der Einsen in der Stichprobe. Da die zweite Ableitung > 0, ist dieser Wert ein Maximum.

Ist Y nun eine stetige Zufallsvariable mit Dichte

$$f_X(x) := \begin{cases} \lambda \cdot e^{-\lambda y} & : \quad y > 0 \\ 0 & : \quad sonst \end{cases}$$

so erhält man für die Likelihood Funktion:

$$L(\lambda \mid x_1, \ldots, x_n) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

und für die logarithmierte Funktion

$$lnL(\lambda \mid x_1, \ldots, x_n) = n \cdot ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

Ableiten nach Lambda und Nullsetzen liefert

$$\widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} x - i}$$

Dies ist gerade 1/Stichprobenmittelwert.

Es gibt noch andere Schätzmethoden, doch ist die ML-Methode die in der Computerlinguistik am meisten angewandte, so dass weitere Schätzmethoden (etwa die Momentenmethode oder die Kleinste Quadrat-Schätzung), nicht behandelt werden.

7. Stochastische Prozesse

Die meisten Erscheinungen in der Natur sind keine Einzelereignisse, sondern Vorgänge, welche vom Zufall abhängen, aber prinzipiell in der Zeit unendlich lange weiterlaufen. Man denke an die Bewegung eines Moleküls, welche für jeden Bewegungsschritt durch eine Zufallsvariable beschrieben werden kann, vollständig aber erst durch viele von der Zeit abhängige Zufallsvariable zu beschreiben ist. Solche, durch die Zeit indizierte Zufallsvariablen bezeichnet man als stochastischen Prozess. Weitere Beispiele sind Aktienkurse, Benzinpreisentwicklungen oder für die Computerlinguistik die Folge der Tags in einem Satz. Zunächst soll jetzt die Definition eines Stochastischen Prozesses gegeben werden.

Definition 7.1 (Stochastischer Prozess)**.** *Ein stochastischer Prozess (SP) ist eine Menge (Familie)$\{q_t \mid t \in T\}$ (T Indexmenge) von Zufallsvariablen über derselben Grundgesamtheit und mit gemeinsamem Wertebereich S. T heißt Parameterraum und S Zustandsraum (States) des stochastischen Prozesses. Eine Realisation des SP ist die Folge der Realisationen der einzelnen Zufallsvariablen für die verschiedenen Werte aus der Indexmenge T(entspricht also einem Zufallsvektor). Man sagt, dass sich der Stochastische Prozess zu einem Zeitpunkt t Prozess im Zustand s befindet, wenn $q_t = s$.*

Bemerkung 7.2. *In der Computerlinguistik ist T stets endlich, auch die Zustandsmenge S der stochastischen Prozesse ist endlich. Für das Folgende sei die Zustandsmenge stets $S := \{s_1, s_2, \cdots, s_n\}$. Befindet sich der Stochastische Prozesse im Zustand s_j, so schreibt man $q_t = s_j$.*

Beispiel 7.3. *Betrachtet man den Satz 'Ich studiere Computerlinguistik', so lässt sich dieser als SP $q_1q_2q_3$ der Länge 3 ansehen. Als Zustandsraum definiert man die Menge der möglichen Tags. Dann lässt sich der Satz als Realisation eines stochastischen Prozesses ansehen mit den Zuständen q_1 = Personalpronomen, q_2 = finites Verb, q_3 = Nomen. Zum Zeitpunkt t=2 befindet sich der Prozess im Zustand Finites Verb.*

Eine wesentliche Fragestellung im Zusammenhang mit stochastischen Prozessen ist die nach der Wahrscheinlichkeit einer bestimmten Realisation, also

$$P(q_1, q_2, \cdots, q_T) = P(q_1 = s_{i_1}, q_2 = s_{i_2}, \cdots, q_T = s_{i_T})$$

Die Zustände werden bei den meisten Berechnung weggelassen. Bezüglich des obigen Beispieks ist das die Frage nach der Wahrscheinlichkeit des Satzes 'Ich studiere Computerlinguistik'. Da die einzelnen Zufallsvariablen nicht als unabhängig vorausgesetzt werden, ist die Bestimmung dieser Wahrscheinlichkeit nicht trivial.

Definition 7.4. *Anfangsverteilung Ist ein stochastischer Prozess $\{q_t \mid t \in T\}$ mit Zustandsraum $\{s_1, ..., s_n\}$ gegeben, so heißt die Wahrscheinlichkeit*

$$\pi_j = P(q_1 = s_j); \quad j = 1, \cdots, n$$

die Anfangsverteilung des stochastischen Prozesses.

Die Anfangsverteilung gibt also an, mit welcher Wahrscheinlichkeit ein Zustand aus dem Zustandsraum zum Zeitpunkt t=1 angenommen wird. Mit Hilfe der bedingten Verteilungen kann nun die Wahrscheinlichkeit einer Realisation eines stochastischen Prozesses berechnet werden:

$$P(q_1, \cdots, q_T) = P(q_T \mid q_1, q_2, \cdots, q_{T-1}) \cdot P(q_1, \cdots, q_{T-1}) = ...$$
$$P(q_T \mid q_1, \cdots, q_{T-1}) \cdot P(q_{T-1} \mid q_1, \cdots, q_{T-2}) \cdots P(q_2 \mid q_1) \cdot P(q_1)$$

Dies sieht zwar kompliziert aus, hilft aber nicht unbedingt bei der Berechnung der Wahrscheinlichkeit, da auch die reduzierten Wahrscheinlichkeiten bis auf den letzten faktor (der Anfangsverteilung) nur schwer zu berechnen sind. Deshalb wird eine neue Klasse von Stochstischen Prozessen definiert, die Markov-Ketten.

7.1. **Markovketten.**

Definition 7.5. *Ein stochastischer Prozess $\{q_t \mid t \in T\}$ heißt Markovkette (MK) erster Ordnung, wenn gilt*

$$P(q_T \mid q_1, \cdots, q_{T-1}) = P(q_T \mid q_{T-1})$$

Der Zustand des stochastischen Prozesses zum Zeitpunkt t hängt somit also nur vom Vorgängerzustand ab. Je weiter die Abhängigkeit von Vorgängerzustaänden besteht, um so höher wird die Ordnung der Markovkette. Die einzelnen Zufallsvariablen des Stochastischen Prozesses hängen also wie Kettenglieder aneinander. Ein Stochastischer Prozess mit der obigen Eigenschaft besitzt die Markov-Eigenschaft. Trifft man aus Modellierungsgünden die Annahme, dass ein eine Markovkette vorliegen soll, so trifft man die Markovannahme. Obige Berechnung der Wahrscheinlichkeit wir dann einfach zu

$$P(q_1, \cdots, q_T) = P(q_T \mid q_1, q_2, \cdots, q_{T-1}) \cdot P(q_1, \cdots, q_{T-1}) =$$
$$P(q_T \mid q_1, \cdots, q_{T-1}) \cdot P(q_{T-1} \mid q_1, \cdots, q_{T-2}) \cdots P(q_2 \mid q_1) \cdot P(q_1) =$$
$$P(q_T \mid q_{T-1}) \cdot P(q_{T-2} \mid q_{T-3}) \cdots P(q_2 \mid q_1) \cdot P(q_1)$$

Zur Berechnung benötigt man also nur die Anfangsverteilung sowie die bedingten Wahrscheinlichkeiten $P(q_t \mid q_{t-1})$, die Übergangswahrscheinlichkeiten. Genauer hingeschrieben benötigt man also
$P(q_t = s_j \mid q_{t-1} = s_i) \quad i, j \in \{1, ..., n\}$
Da dies nur endlich viele Werte sind, definiert man

$$a_{ij} := P(q_t = s_j \mid q_{t-1} = s_i) \qquad i, j \in \{1, ..., n\}$$

Eine Markovkette erster Ordnung ist durch die Anfangsverteilung und die gerade definierten Übergangswahrscheinlichkeiten somit vollständig charakterisiert. Übergangswahrscheinlichkeiten werden in einer Übergangsmatrix zusammengefasst. Für alle Zustandsfolgen $s_i \longrightarrow s_j$ lassen sich diese Wahrscheinlichkeiten in einer $n \times n$ Matrix A schreiben:

$$A := (a_{ij}) := \begin{pmatrix} a_{11} & a_{12} & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & \dots & a_{nn} \end{pmatrix}$$

heißt die Übergangsmatrix des stochastischen Prozesses.

Die Summe der Zeilen dieser Matrix sind stets gleich 1. Die Werte der i-ten Zeile stellen die Übergangswahrscheinlichkeiten vom zustand i in den Zustand des jeweiligen Spaltenindex dar.

Das Werfen einer fairen Münze kann als Markovkette angesehen werden. Es gibt die zwei Zustände Kopf (K) und Zahl (Z). Damit ergibt sich die Anfangsverteilung $\pi = (0,5; 0,5)$ und die Übergangsmatrix

$$A = \begin{pmatrix} 0,5 & 0,5 \\ 0,5 & 0,5 \end{pmatrix}$$

Markovketten können sehr gut grafisch dargestellt werden, in dem die Zustände als Kreise und die Übergänge als Pfeile zu den Zuständen gezeichnet werden.

7.2. **Markovketten in der Computerlinguistik.** Beim Part of Speech Tagging interessieren die Wortkategorien der Wörter eines Satzes. Das Auftreten einer bestimmten Wortkategorie wird als Realiation einer Zufallsvariablen angesehen, so dass die Folge der Wortkategorien einen stochastischen Prozess darstellt. Wird nun die Annahme getroffen, dass das Auftreten einer bestimmten Kategorie nur von der Vorgängerkategorie abhängt, so erhält man eine Markovkette. Es ergeben sich Bigramwahrscheinlichkeiten. Je nachdem, wie weit man auf Vorgängertags zurückgreift, erhält man Trigramme etc. Ein Unigramm ist ein Modell, bei dem auf kein Vorgängertag zurückgegriffen wird. Die einzelnen Übergangswahrscheinlichkeiten können dann in einer Übergangsmatrix zusammengefasst werden.

Es werde der Satz 'Ich komme morgen' betrachtet. Mögliche Kategorien sind Personalpronomen (PP), finites Verb (VF), Nomen (NN) und Adverb (Adv), da das letzte Wort bei einer ersten Analyse beide Kategorien besitzen könnte. So könnte beispielsweise durch Auszählen von Häufigkeiten in einem Corpus folgende Übergangsmatrix vorliegen:

$$A = \begin{pmatrix} & PP & VF & NN & Adv \\ PP & 0 & 1 & 0 & 0 \\ VF & 0 & 0,13 & 0,43 & 0,44 \\ NN & 0,65 & 35 & 0 & 0 \\ Adv & 0,73 & 0,26 & 0 & 0 \end{pmatrix}$$

Die Lesart ist die folgende: In der ersten Zeile der Matrix stehen die Wahrscheinlichkeiten $P(PP \mid PP), P(VF \mid PP), P(Adv \mid PP), P(NN \mid PP)$. Normalerweise werden die zugrunde liegenden Kategorien nicht mehr aufgeführt. Die Spalten beziehen sich auf die bedingte Kategorie, die Zeilen auf die bedingende Kategorie. Als Anfangsverteilung könnte sich ergeben haben, dass $P(PP) = 0,71, \quad P(NN) = 0,29)$ gilt. Damit wäre die Markovkette vollständig spezifiziert. Die Folge PP VF Adv besitzt dann die folgende Wahrscheinlichkeit:

$$P(PP) \cdot P(VF \mid PP) \cdot P(Adv \mid VF) = 0,71 \cdot 1 \cdot 0,43 = 0,3053$$

Die Einträge der Übergangsmatrx werden normalerweise durch Auszählen eines Corpus erhalten. Diese eine Berechnung für eine mögliche Kategorienfolge beantwortet noch nicht die Frage, welche Wahrscheinlichkeit der Satz 'Ich komme mogen' in dem durch die Übergangsmatrix spezifizierten Modell besitzt. Da 'morgen' hier sowohl Nomen als auch Adverb sein kann, müssen die Wahrscheinlichkeiten für beide Alternativen addiert werden. Somit

$$P(Ich \quad komme \quad morgen) =$$

$$P(PP) \cdot P(VF \mid PP) \cdot P(Adv \mid VF) + P(PP) \cdot P(VF \mid PP) \cdot P(NN \mid VF) =$$

$$0,3053 + 0,3124 = 0,6177$$

7.3. **Markov-Modelle.** Das Modell der Markovmodelle zeigt für die Anwendung in der Computerlinguistik eine Schwäche, nämlich die, dass zur Brechnung der Wahrscheinlichkeit einer Kategoienfolge die entsprechenden bedingten Wahrscheinlichleiten bekannt sein müssen. Dies erfordert jedoch, dass die Kategorien im Satz bekannt sind. In den meisten Anwendungfällen ist jedoch nur der Satz bekannt (also die Wortfolge), die Folge der Kategorien jedoch, also gerade das, was mittels Berechnung herausgefunden werden soll, ist nicht bkannt. Ziel ist also jetzt, ein Modell zu definieren. welche diese Situation widerspiegelt. Hierzu wird das Modell der Markovketten erweitert um den Umstand erweitert wird, dass zu jedem Zeitpunkt t und einem Zustand s_i noch ein Signal w_j mit einer bestimmten Wahrscheinlichkeit, die von s_i und w_j abhängt, ausgegeben wird. Ein Markovmodell besteht dann aus folgenden Komponenten:

(1) Eine Markovkette $\{q_t \mid t \in T\}$ mit einer endlichen Folge von Zuständen $\{s_1, \cdots, s_n\}$

(2) einer Übergangsmatrix $A = a_{ij}$ wobei $a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$

(3) einer Anfangsverteilung $\Pi = (\pi_1, \cdots, \pi_n)$ wobei $\pi_i = P(q_1 = s_i)$.

(4) einer Signalmenge $\Sigma = (w_1, \cdots, w_k)$

(5) Einer Signalmatrix, die wie folgt erklärt ist:
Seien $o_1, \cdots, o_T$ Zufallsvariablen, die zu jedem t und jedem Zustand $q_t = s_i$ ein Signal $o_t = w_j$ ausgeben. Mit b_{ij} werde die Wahrscheinlichkeit bezeichnet, dass ein Signal w_j ausgegegeben wird, wenn sich sich die Markovkette im Zustand $q_t = s_i$ befindet, also:
$b_{ij} = P(o_t = w_j \mid q_t = s_i)$
Die Signalmatrix ist die Matrix dieser b_{ij}, also $B_{n\times k} = (b_{ij})$
$i = 1, \cdots n, \quad j = 1, \cdots, k$

Eine neben der Markovannahme weitere wichtige Annahme ist nun die, dass die Ausgabe eines Signals $o_t = w_j$ nur vom Zustand $q_t = s_i$ abhängt, nicht aber vom Vorgängerzustand. Diese Annahme heißt auch zweite Markovannahme.
Folgende Notation bietet sich für bestimmte Rechnungen an: Statt b_{ij} wird $b_{q_i}(o_j)$ geschrieben, um zu verdeutlichen, dass sich die Markovkette im Zustand s_i befindet und das Signal $o_j = w_j$ ausgegeben wird. Bei dieser Definition eines Markovmodells sind alle Parameter bekannt.

Den Signalen entsprechen später die Wörter eines Satzes, den Kategorien die bei der Ausgabe eines Signals zugrunde liegenden Zustände, also die Wortkategorien. Bezüglich der später zu berechnenden Wahrscheinlichkeiten bieten beide Markovannahmen erhebliche Vereinfachungen: Es gilt nämlich:

$$P(q_1 = s_{i_1}, q_2 = s_{i_2}, \cdots, q_t = s_{i_T}, o_1 = w_{j_1}, o_2 = w_{j_2}, \cdots, o_t = w_{j_T}) =$$

$$P(q_T = s_{i_T} \mid q_{T-1} = s_{i_{T-1}}) \cdots P(q_2 = s_{i_2} \cdot \mid q_1 = s_{i_1}) \cdot \pi_1 \cdot$$

$$P(o_1 = w_{j_1} \mid q_1 = s_{i_1}) \cdots P(o_T = w_T \mid q_T = s_{i_T})$$

Ohne die Zustände und Signale zu nennen schreibt man kürzer:

$$P(q_1, \cdots, q_T, o_1, \cdots, o_T) =$$

$$P(q_T \mid q_{T-1}) \cdots P(q_2 \mid q_1) \cdot \pi_1 \cdot P(o_t \mid q_T) \cdots P(o_1 \mid q_1)$$

Der erste Term kann also in einen Faktor aus dem Produkt der Übergangswahrscheinlichkeiten und einem Faktor aus dem Produkt der Ausgabewahrscheinlichkeiten aufgespalten werden (s.u.). Markovmodelle können ebenfalls sehr einfach grafisch dargestellt werden.

7.4. **Markovmodelle in der Computerlinguistik.** Betrachtet man obige Formel für das Berechnen der Wahrscheinlichkeit einer vorliegenden Folge von Kategorien und Signalen, so lässt sich das Ergebnis der Umformung in 2 Teile trennen, einen Teil, der nur ein Produkt von Übergangswahrscheinlichkeiten darstellt und einen Teil, der die bedingten Wahrscheinlichkeiten der Signale multipliziert. Mit $S := q_1, \cdots, q_T$ und $O = w_1, \cdots, w_T$ kann dann obige Wahrscheinlichkeit wie folgt geschrieben werden:
$P(S, O) = P(O \mid S) \cdot P(S)$
$P(S)$ heißt Sprachmodell, $P(O \mid S)$ heißt Signalmodell. Die Umformung ergibt sich aus der Definition der bedingten Wahrscheinlichkeiten.
Betrachtet man wieder den Satz 'Ich komme morgen', so liefertdie Modellierung mit einem Markovmodell folgende Interpretation: Die Kategorienfolge sei PP VF Adv. Dann befindet sich das Modell zu t=1 im Zustand PP und gibt in diesem Zustand das Wort Ich aus (also das Signal Ich), zu t=2 befindet sich die Kette in Zustand VF und gibt in diesem Zustand das Wort komme aus, zu t=3 liegt Zustand Adv vor und es wird das Signal morgen ausgegeben. Somit werden in Markovmodellen beobachtete Wortfolgen und Kategorien in einen Zusammenhang gebracht.

Interessant ist nun die Bstimmung der Wahrscheinlichkeit einer Beobachtung $P(o_1, \cdots, o_T)$. Hierzu muss überlegt werden, welche Zustandsfolgen zu dieser Beobachtung geführt haben. Da man dies nicht weiß, müssen alle möglichen Zustände betrachtet werden und die Wahrscheinlichkeiten hierüber addiert werden. Das obige Ergebnis $P(S, O) = P(O \mid S) \cdot P(S)$ liefert also einen Summanden der jetzt zu bildenden Summe. Ausführlich hingeschrieben gilt:

$$P(O) = \sum_S P(O \mid S) \cdot P(S) =$$
$$\sum_{q_1, \cdots, q_T} P(q_1) \cdot P(q_2 \mid q_1) \cdots P(q_T \mid q_{T-1}) \cdot P(o_1 \mid q_1) \cdots P(o_T \mid q_T)$$

Bei n^T Summanden sind dies $n^T - 1$ Additionen, sowie pro Summand 2T-1 Multiplikationen, also insgesamt für n^T Summanden $2Tn^T - 1$ Operationen. Dies ist auf keinem Rechner auch bei nur wenigen Zuständen durchführbar. Die Summe eribt sich direkt aus den Rechenregeln für Wahrscheinlichkeiten, da alle vorkommenden Urbild-Mengen disjunkt sind. Dies ist der Fall, da die Zufallsvariablen q und o wohldefinierte Abbildungen sind. Um die gesuchte Wahrscheinlichkeit einer Beobachtung zu ermitteln, muss eine andere Möglichkeit gefunden werden.

7.5. **Hidden Markov Modelle.** Kennt man in einem Markov-Modell nur die in einem Zustand ausgesendeten Signale, nicht aber die Zustände, spricht man von einem Hidden Markov Modell (HMM). Man definiert ein HMM als ein Tripel $\lambda = (A, B, \pi)$ mit

(1) A ist eine Übergangsmatrix (verborgen, nicht bekannt), Zustandsraum sei $S = \{s_1, \cdots, s_N\}$ als Menge der möglichen Realisationen der Zufallsvariablen $q_1, \cdots, q_T$.

(2) B ist Signalmatrix, offen, k sichtbare Symbole $w_1, \cdots, w_k$ als Menge der Realistionen der T Zufallsvariablen $o_1, \cdots, o_T$.

(3) π ist die Anfangsverteilung der Markovkette, verborgen

Die drei Komponenten des Tripels heißen Modellparameter. Ziel ist es nun, auf die zugrunde liegenden Zustände zu schließen. Bei der Spracherkennung bedeutet dies, man sieht zwar ein Wort, kennt aber nicht die Kategorie (Zustand), unter der das Wort ausgesendet wurde. Grundansatz der folgenden Algorithmen ist, wenn

ein Wort von mehreren Kategorien ausgesendet sein könnte, die Kategorie mit der größten Wahrscheinlichkeit zu wählen. Dazu muss diese bestimmt werden. Es sind zur Lösung des eingangs genannten Problems 3 Fragen zu beantworten.

(1) Die Wahrscheinlichkeit $P(O \mid \lambda)$ in einem gegebenen HMM für eine Beobachtungsfolge $O = o_1, \cdots, o_T$

(2) Die Wahrscheinlichkeit $\max_S P(S \mid O)$, die Bestimmung der wahrscheinlichsten Zustandsfolge bei gegebener Beobachtungsfolge.

(3) Die Bestimmung der Wahrscheinlichkeit $\max_\lambda P(O \mid \lambda)$, also die Bestimmung der Modellparameter so, dass O mit maximaler Wahrscheinlichkeit erklärt wird. Dies nennt man Training des HMM´s.

Hier werden nur 1 und 2 betrachtet.

7.5.1. *Vorwärtsalgorithmus.* Die Bestimmung von P(O): Oben wurde gezeigt, dass folgendes gilt:

$$P(O) = \sum_S P(O \mid S) \cdot P(S) =$$

$$\sum_{q_1,\cdots,q_T} P(q_1) \cdot P(q_2 \mid q_1) \cdots P(q_T \mid q_{T-1}) \cdot P(o_1 \mid q_1) \cdots P(o_T \mid q_T)$$

Der Rechenaufwand ist beträchtlich. Die Summe über alle Zustandskombinationen ist notwendig, da unbekannt ist, welcher Zustand zu einem Signal geführt hat. Ziel des Algorithmus ist es, den Rechenaufwand dadurch zu reduzieren, dass nicht für jede Zustandskombination ein Wert P(O) berechnet wird, sondern dass Zwischenwerte für die Zeitpunkte 1, 2, ...,t, ..,T gebildet werden, die die jeweiligen bis dahin berechneten Werte enthalten. Es liege ein HMM im obigen Sinne vor. Der Algorithmus besteht aus 3 Schritten:

(1) Initialisierung
Definiere für $1 \leq j \leq n$:
$\alpha_1(j) = \pi_j \cdot b_{q_1=s_j}(o_1)$
Dies gibt die Wahrscheinlichkeit an, zu $t = 1$ im Zustand $q_1 = s_j$ zu sein und das Signal o_1 auszusenden.

(2) Iterationsschritt
Definiere für $1 \leq j \leq T - 1$
$\alpha_{t+1}(j) = (\sum_{i=1}^{n} \alpha_t(i) \cdot a_{ij}) \cdot b_{q_{t+1}=s_j}(o_{t+1})$
Dies gibt die Wahrscheinlichkeit an, zu Zeitpunkt $t+1$ im Zustand $q_{t+1} = s_j$ zu sein und bis dahin $o_1, o_2, \cdots, o_t$ gesehen zu haben und zu $t+1$ das Signal o_{t+1} auszusenden.

(3) Terminierung
$P(O \mid \lambda) = \sum_{j=1}^{n} \alpha_T(j)$
Die einzelnen $\alpha_T(j)$ geben die Wahrscheinlichkeit an, zu $t = T$ im Zustand $q_T = j$ zu sein und die Folge $o_1, \cdots, o_T$ gesehen zu haben. Summiert man über alle $\alpha_T(j)$, so ergibt das gerade $P(O \mid \lambda)$.

Die Variablen $\alpha_t(j)$ heißen Vorwärtsvariablen. Betrachtet man den Rechenaufwand, so sind dies n Multiplikationen für die Initialisierung, n(n+1) Multiplikationen sowie n(n-1) Additionen bei der Iteration und n Additionen bei der Terminierung, insgesamt also $2n + 2n^2(T-1) \sim O(n^2T)$ Rechenoperationen.

Folgendes Beispiel werde betrachtet:

(1) Zustandsraum $S = \{s_1, s_2, s_3, s_4\}$

(2) Signalmenge $O =\mid w_1, w_2, \cdots, w_5 \mid$

(3) Anfangsverteilung $\pi_1 = 0,1, \pi_2 = 0,3, \pi_3 = 0,4, \pi_4 = 0,2$

(4) Übergangsmatrix

$$A = \begin{pmatrix} 0,1 & 0,2 & 0,5 & 0,2 \\ 0,3 & 0,1 & 0,6 & 0 \\ 0 & 0 & 0,2 & 0,8 \\ 0,3 & 0 & 0,1 & 0,6 \end{pmatrix}$$

(5) Signalmatrix

$$B = \begin{pmatrix} 0,2 & 0,2 & 0 & 0,3 & 0,3 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0,1 & 0,1 & 0,2 & 0,6 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

(6) Es werde die Folge $o_1 o_2 = w_1 w_2$ beobachtet.

Der Eintrag $a_{32} = 0,2$ der Beobachtungsmatrix sagt, dass von Zustand s_3 in den Zustand s_2 mit Wahrscheinlichkeit 0,2 gewechselt wird.
Der Eintrag $b_{34} = 0,2$ in der Signalmatrix gibt an, dass im Zustand s_3 mit Wahrscheinlichkeit 0,2 das Signal w_4 ausgegeben wird.
Zum Algorithmus:

(1) Initialisierung t=1: $\alpha_1(j) = \pi_j \cdot b_j(o_1)$
   (a) $\alpha_1(1) = 0,1 \cdot 0,2 = 0,02$
   (b) $\alpha_1(2) = 0,3 \cdot 0 = 0$
   (c) $\alpha_1(3) = 0,4 \cdot 0,1 = 0,04$
   (d) $\alpha_1(4) = 0,2 \cdot 0 = 0$

(2) Iteration, t=2, $\alpha_2(j) = (\sum_{i=1}^{4} \alpha_1(i) \cdot a_{i1}) \cdot b_1(o_2)$
   (a) $\alpha_2(1) = (0,02 \cdot 0,1 + 0 \cdot 0,3 + 0,04 \cdot 0 + 0 \cdot 0,3) \cdot 0 = 0$
   (b) $\alpha_2(2) = 0,004$
   (c) $\alpha_2(3) = 0,0018$
   (d) $\alpha_2(4) = 0$

(3) Terminierung, T=2
   $P(w_1 w_2) = 0,004 + 0,0018 = 0,0022$

Die Wahrscheinlichkeit der Beobachtung $w_1 w_2$ im oben spezifizierten HMM ist also 0,0022.

7.5.2. *Viterbialgorithmus.* Neben der Frage, wie wahrscheinlch eine Beobachtung ist, ist die Frage interessant, welche Zustandsfolge diese Beobachtung erzeugt hat. Dies ist nicht eindeutig zu beantworten, da in der Regel viele Zustandsfolgen eine Beobachtung erzeugen können. Die beste Antwort, die gegeben werden kann, ist die Folge von Zuständen anzugeben, welche die wahrscheinlichste bei der gegebenen Beobachtung ist. Der Viterbialgorithmus bietet eine Möglichkeit, diese Folge von Zuständen zu berechnen. Formal hingeschrieben heißt dies: Es ist die Folge $q_1^*, \cdots, q_T^*$ von Zuständen gesucht mit

$$P(q_1^*, \cdots, q_T^* \mid o_1, \cdots, o_T) = max_{q_1, \cdots, q_T}(P(q_1, \cdots, q_T \mid o_1, \cdots, o_T)$$

In der hier verwendeten Schreibweise wurde auf die konkrete Angabe von Zuständen $q_1 = s_{i_1}, \cdots q_T = s_{i_T}$ verzichtet. Formal korrekt wird natürlich eine Folge von Zuständen $s_1^*, \cdots s_T^*$ gesucht. Kurz geschrieben heißt das Ziel dieses Kapitels: Suche ein S^* mit

$$P(S^* \mid O) = max_S P(S \mid O)$$

Im rechten Teil dieser Gleichung liegt der Ansatz für den Viterbialgorithmus. Es gilt nämlich nach dem Satz von Bayes:

$$P(S \mid O) = \frac{P(O \mid S) \cdot P(S)}{P(O)}$$

Bezüglich der Maximumsbildung über alle Zustände ist der Nenner nicht wichtig, so dass gilt:

$$max_S P(S \mid O) = max_S P(O \mid S) \cdot P(S)$$

Andererseits gilt auch noch nach der Definition der bedingten Wahrscheinlichkeit für die linke Seite:

$$max_S P(S \mid O) = max_S \frac{P(S,O)}{P(O)} = max_S P(S,O)$$

Also gilt die Gleichheit:

$$max_S P(S \mid O) = max_S P(S,O) = max_S P(O \mid S) \cdot P(S)$$

Die Berechnung der maximalen Wahrscheinlichkeit wird im Viterbialgorithmus durchgeführt und diese liefert somit das gesuchte Maximum. Es wird der mittlere Term berechnet. Für das weitere Vorgehen liege wieder das obige HMM vor. Der Viterbialgorithmus besteht ebenfalls aus 3 Schritten, wobei die Variablen jetzt mit δ bezeichnet werden. Auch diese werden zunächst für alle Zustände getrennt definiert.

(1) Initialisierung, t=1, definiere für $1 \leq j \leq n$: $\delta_1(j) = \pi_j \cdot b_{q_1=s_j}(o_1)$
In $\delta_1(j)$ steht die Wahrscheinlichkeit, zu t=1 im Zustand s_j zu sein und o_1 zu sehen. $\delta_1(j)$ entspricht dem $\alpha_1(j)$ beim Vorwärtsalgorithmus.

(2) Iteration, $t = 2, \cdots, T, \quad j = 1, \cdots n$
$\delta_t(j) = (\max_{i=1,..,n} \delta_{t-1}(i) \cdot a_{ij}) \cdot b_{q_t=s_j}(o_t)$
In $\delta_t(j)$ steht die größte der Wahrscheinlichkeiten, $o_1, \cdots, o_n$ gesehen zu haben, im Zustand $q_t = s_j$ zu sein und dort das Signal o_t auszugeben.
Da in dieser Variable nur eine Zahl steht, aber kein Zustand, an dem man eigentlich interessiert ist, muss sich irgendwo noch gemerkt werden, von welchem Zustand q_{t-1} man zu $q_t = s_j$ gekommen ist, also welcher Zustand zur maximalen Wahrscheinlichkeit geführt hat. Man muss sich also den Vorgängerzustand von $q_t = s_j$ merken. Es ist daher eine sog. Pfadrückverfolgung durchzuführen. Der Pfad, von dem man gekommen ist, lässt sich formal wie folgt angeben (wobei nur der Index j des Zustandes geschrieben wird; formal müsste $\psi_t(q_t = s_j)$ geschrieben werden.)
$\psi_t(j) = arg \max_i \delta_{t-1}(i) \cdot a_{ij}$ für $t = 2, \cdots, T, \quad j = 1, \cdots n$
Gilt also z. B. $\psi_t(j) = s_{i_0}$, so heißt dies
$\max_i \delta_{t-1}(i) \cdot a_{ij} = \delta_{t-1}(i_0) \cdot a_{i_0 j}$,
man ist also über s_{i_0} nach s_j mit der größten Wahrscheinlicheit gelangt.

(3) Terminierung, t=T
Für t=T stehen in den $\delta_T(j)$ die (für jeden Endzustand s_j) jeweils maximalen Wahrscheinlichkeiten, die Folge $o_1 \cdots o_T$ zu sehen und im Endzustand $q_T = s_j$ zu sein. Die größte Wahrscheinlichkeit ist dann einfach der größte dieser Werte, also
$\max_q P(q,o) = \max_{j=1,..,n} \delta_T(j) =: P^*(O)$
Der Zustand q_T^* mit $\delta_T(q_T^* = s_{j^*}) = max_j \delta_T(j)$ ist auch der Zustand, an dem die Zustandskette, die mittels Pfadrückverfolgung zu bestimmen ist,

endet. Von diesem Endzutand aus wird nun die Pfadrückverfolgung vorgenommen. Mit obiger Notation gilt zunächst:
$q_T^* = arg\max_j \delta_T(j)$
Da in $\psi_T(q_T^*)$ der Vorgängerzustand abgespeichert ist, gilt also sukzessiv:
$q_{T-1}^* = \psi_T(q_T^*), \cdots, \quad q_1^* = \psi_2(q_2^*)$
so dass in dieser Notation

$$q_1^* q_2^* \cdots q_T^*$$

die gesuchte Zustandsfolge ist, die mit der größten Wahrscheinlichkeit zu der Beobachtung geführt hat.

Dies werde jetzt an einem konkreten HMM als Beispiel durchgeführt:

(1) Es seien 3 Zustände $S = \{s_1, s_2, s_3\}$ gegeben.

(2) Die Signalmenge bestehe aus 4 Signalen, also $\Sigma = \{w_1, w_2, w_3, w_4\}$

(3) Die Anfangsverteilung sei $\Pi = (0,3; 0,3; 0,4)$

(4) Die Übergangsmatrix sei

$$A = \begin{pmatrix} 0 & 0,8 & 0,2 \\ 0,5 & 0 & 0,5 \\ 0,2 & 0,8 & 0 \end{pmatrix}$$

(5) Die Signalmatrix sei

$$B = \begin{pmatrix} 0 & 0,5 & 0 & 0,5 \\ 0,1 & 0,5 & 0,4 & 0 \\ 0,2 & 0,1 & 0,5 & 0,2 \end{pmatrix}$$

(6) Die beobachtete Folge sei $w_2 w_3 w_1$.

Die Schritte des Viterbialgorithmus ergeben dann folgende Werte:

(1) Initialisierung t=1, statt $\delta_i(s_j)$ werde einfach $\delta_i(j)$ geschrieben.
   (a) $\delta_1(1) = 0,3 \cdot 0,5 = 0,15$
   (b) $\delta_1(2) = 0,3 \cdot 0,5 = 0,15$
   (c) $\delta_1(3) = 0,4 \cdot 0,1 = 0,04$

(2) Iteration, t=2, $\delta_2(j) = \max_i(\delta_1(i) \cdot a_{ij}) \cdot b_{q_2=s_j}(o_2 = w_3)$
   (a) $\delta_2(1) = \max(0,15 \cdot a_{11}; 0,15 \cdot a_{21}; 0,04 \cdot a_{31}) \cdot b_{q_2=s_1}(w_3) = 0$
   (b) $\delta_2(2) = \max(0,15 \cdot a_{12}; 0,15 \cdot a_{22}; 0,04 \cdot a_{32}) \cdot b_{q_2=s_2}(w_3) = 0,15 \cdot 0,8 \cdot 0,4 = 0,048$
   (c) $\delta_2(3) = \max(0,15 \cdot a_{13}; 0,15 \cdot a_{23}; 0,04 \cdot a_{33}) \cdot b_{q_2=s_3}(w_3) = 0,075 \cdot 0,5 = 0,0375$
   (d) Pfadrückverfolgung: es muss sich gemerkt werden, welcher Zustand jeweils Ausgangspunkt für das Maximum war. $\psi_t(j) = arg\max_i \delta_{t-1} \cdot a_{ij}$
   Man erhält
   $\psi_2(1) = arg\max_i(0,5; 0,075; 0,008) = s_2$
   $\psi_2(2) = arg\max_i(0,12; 0; 0,032) = s_1$
   $\psi_2(3) = arg\max_i(0,03; 0,075; 0) = s_2$

(3) Iteration, t=3, $\delta_3(j) = \max_i(\delta_2(i) \cdot a_{ij}) \cdot b_{q_3=s_j}(o_3 = w_1)$
Aus dem vorherigen Schritt hat man: $\delta_1(1) = 0, \quad \delta_2(2) = 0,048, \quad \delta_2(3) = 0,0375$

(a) $\delta_3(1) = \max(0 \cdot a_{11}; 0,048 \cdot a_{21}, 0,0375 \cdot a_{31}) \cdot b_{q_3=s_1}(o_3 = w_1) = 0$

(b) $\delta_3(2) = \max(0 \cdot a_{12}; 0,048 \cdot a_{22}, 0,0375 \cdot a_{32}) \cdot b_{q_3=s_2}(o_3 = w_1) = 0,003$

(c) $\delta_3(3) = \max(0 \cdot a_{13}; 0,048 \cdot a_{23}, 0,0375 \cdot a_{33}) \cdot b_{q_3=s_3}(o_3 = w_1) = 0,0048$
Das Maximum dieser 3 Werte ist $\delta_3(3) = 0,0048$. Dies ist also die maximale Wahrscheinlichkeit von $w_2w_3w_1$ im Modell. Es fehlt jetzt noch die Zustandsfolge, die diese Wahrscheinlichkeit liefert. Diese wird über die Pfadrückverfolgung erhalten.

(d) Pfadrückverfolgung
$\psi_3(1) = arg\max_i(0; 0,024; 0,0075) = s_2$
$\psi_3(2) = arg\max_i(0; 0; 0,03) = s_3$
$\psi_3(3) = arg\max_i(0; 0,024; 0) = s_2$

Es muss noch der Zielzustand für t=3 bestimmt werden. Es gilt

$$q_3^* = arg\max_j(\delta_3(j)) = arg\max(0; 0,003; 0,0048) = s_3$$

Der Vorgängerzustand von s_3 ist in $\psi_3(3)$ abgespeichert und ist $q_2^* = \psi_3(3) = s_2$. Der Vorgängerzustand von s_2 liegt in $\psi_2(2) = s_1$. Insgesamt ergibt dies also die Zustandsfolge

$$s_1s_2s_3,$$

welche die maximale Wahrscheinlichkeit der Beobachtung $w_2w_3w_1$ liefert.

8. Statistische Tests

Statistische Tests dienen dazu, Hypothesen über Verteilungen oder Parameter zu prüfen, sie abzulehnen oder anzunehmen. Entscheidungsgrundlage ist in der Regel ein Stichprobenbefund, d.h. es wird eine Stichprobe gezogen, anhand derer eine Entscheidung für oder gegen eine Hypothese vorgenommen wird. Wie kann man den Begriff einer Stichprobe formal fassen? Intuitiv wird beim Ziehen einer Stichprobe angegeben, wie viele Elemente der Grundgesamtheit gezogen werden sollen. Dies ist der sog. Stichprobenumfang. Ist X ein Merkmal, etwa Größe einer Person, und zieht man n Elemente der Grundgesamtheit, so kann das Ziehen des i-ten Elementes durch eine Zufallsvariable X_i beschrieben werden. Insgesamt erhält man bei einer Stichprobe vom Umfang n also einen Verktor $X_1, X_2, .., X_n$). Dies ist dann eine Stichprobe vom Umfang n zur Zufallsvariable X. Da die einzelnen X_i Zufallsvariablen darstellen, besitzen sie eine Verteilung und können voneinander unabhängig oder abhängig sein. Ein wichtiger, auch hier dann vorausgesetzter Stichprobentyp ist der einer einfachen Stichprobe. Man spricht von einer einfachen Stichprode vom Umfang n zu einer Zufallsvariablen X, wenn alle X_i stochastisch unabhängig sind und alle X_i identisch wie X verteilt sind. Die Menge $\mathbf{X}$ aller möglichen Stichprobenbefunde heißt Stichprobenraum und mit x wird ein konkreter Stichprobenbefund bezeichnet. Es werden bei den sog. Parametertests Werte der Parameter formuliert und dann durch den Test entweder bestätigt(Hypothese annehmen) oder widerlegt (Hypothese ablehnen). Das Problem besteht jetzt darin zu klären, wie weit man ein zufällig zustande gekommenes Ergebnis (der Stichprobenbefund) als für oder gegen die Hypothese sprechendes Ergebnis zu akzeptieren bereit ist.

Hierzu wird zunächst die Menge der möglichen Parameter W in zwei disjunkte Teilmengen W_0, W_1 zerlegt mit $W = W_0 \cup W_1$. Die Nullhypothese H_0 lautet dann: der Parameter (oder die Verteilung) liegt in W_0, die sog. Gegenhypothese oder Alternativhypothese H_1 lautet entsprechend, dass der interessierende Parameter (bzw. die Verteilung)in W_1 liegt. Ist W_0 einelementig, so heißt die Hypothese einfache Hypothese, sonst zusammengesetzte Hypothese.

Natürlich lassen sich nicht nur Parameter von Verteilungen mit Hypothesentests überprüfen, sondern auch Aussagen über die Art einer Verteilung oder Unabhängigkeit von Zuzfallsvariablen. Wichtig ist hier ebenfalls, dass die Menge der möglichen Hypothesen in zwei disjunkte Entscheidungsmengen geteilt werden kann.

Zunachst sollen Tests über Parameter einer Verteilung betrachtet werden. Wenn man z. B. vermutet, dass eine bestimmte Buchstabenfolge etwa in jedem 10ten Wort oder öfter einer Sprache vorkommt und soll dies durch einen Test abgesichert werden, so würde man H_0: der Anteil der Buhstabenfolge ist ≥ 10 Prozent und die Hypothese H_1: der Anteil ist $\neq 10$ Prozent formulieren. Die Menge der möglichen Parameter ist dann das Intervall [0;100].

Wie geht man weiter vor? Es wird Stichprobenergebnisse geben, bei denen man H_0 annimmt, aber auch welche, bei denen man H_0 verwirft. Dementsprechend spricht man von Annahmebereich A und Ablehnungs- oder kritischem Bereich K. Man kann schreiben:

$$A := \{x \in \mathbf{X} \mid H_0 \text{wird angenommen}\} K := \{\in \mathbf{X} \mid H_0 \text{wird abgelehnt}\}$$

Es geht also bei Tests um die Nullhypothese, diese steht in der Regel auf dem Prüfstand und soll eigentlich durch den Stichprobenbefund widerlegt werden, also abgelehnt werden. Die Nullhypothese stellt meist eine bekannte Tatsache dar oder wird aus Erfahrungswerten gewonnen. Z.B. kennt man die mittlere Körpergröße von Populationsmitgliedern oder man weiß aufgrund längerer Beobachtungen, dass die mittlere Abfüllmenge einer Abfüllanlage 0,98 Liter beträgt. Wie ihre Definition im konkreten Fall an statistische Eigenschaften geknüpft wird, folgt noch. Als Gegenhypothese ist demnach immer eine Aussage zu wählen, die man eigentlich gerne statistisch bestätigt hätte. Grundlegend ist der Ansatz, dass man gegen die Nullhypothese entscheidet, wenn ein Stichprobenbefund vorliegt, der, wenn man die Nullhypothese als wahr ansieht, sehr unwahrscheinlich ist. Dies kann man wie folgt formal beschreiben: Man kennt den wahren Parameter oder die wahre Verteilung der Grundgesamtheit nicht. Es ist zu unterscheiden zwischen der Realität (die man nicht kennt, die aber einer der beiden Hypothesen entspricht) und dem Stichprobenbefund, der ein Abbild der Reealität sein soll. Aufgrund des Stichprobenbefundes ist eine Entscheidung für oder gegen H_0 zu treffen. Das führt zu folgenden 4 möglichen Entcheidungen: Da das Stichprobenergebnis zufällig ist, kann folgendes passieren, was zu falschen Schlüssen führt: Obwohl in Wirklichkeit H_0, d.h. das, was durch die Nullhypothese formuliert wird, gilt, deutet der Stichprobenbefund auf H_1 hin (Fehler erster Art) bzw. obwohl H_1 gilt, deutet der Stichprobenbefund auf H_0 hin (Fehler zweiter Art). Folgendes Diagramm beschreibt diese Fälle:

	wahrer Zustand	
	H_0 gilt	H_1 gilt
Entscheidung		
H_0 annehmen	o.k.	Fehler 2. Art
H_1 annehmen	Fehler 1. Art	o.k.

Die Berechnung der Wahrscheinlichkeit des Fehlers erster Art ist nun der Schlüssel zur Definition von Annahmebereich und kritischem Bereich. wie entscheidet man, für oder gegen eine Hypothese, also, ob ein Stichprobenbefund in A oder in K

liegt. Dies führt zum Signifikanzniveau des Tests: Der Stichprobenbefund ist in der Regel eine Kennzahl. Man wählt den Annahmebereich so, dass bei Gültigkeit der Nullhypothese ein Stichprobenbefund im K sehr unwahrscheinlich ist. Wenn also der Stichprobenbefund bei Annahme der Gültigkeit von H_0 in den kritischen Bereich fällt, sagt man, das sei so unwahrscheinlich (bei Gültigkeit der Hullhypothese), dass man sie verwerfen müsse, also dass die Alternativhypothese gelten müsse. H_1 wird somit angenommen. Man wählt sich daher in der Praxis ein sog. Signifikanzniveau (eine kleine Wahrscheinlichkeit $\alpha = 0,05$ *oder* $\alpha = 0,01$) und bestimmt den Annahmebereich A unter der Bedingung, dass H_0 gilt, so, dass

$$P\{x \in \mathbf{X} \mid x \in A\} \geq 1 - \alpha$$

wenn H_0 gilt. Entsprechend gilt dann für den kritischen Bereich K

$$P\{x \in \mathbf{X} \mid x \in K\} \leq \alpha$$

wenn H_0 gilt. Dies bedeutet, dass das Signifikanzniveau gerade die Wahrscheinlichkeit des Fehlers erster Art ist. Bei statistischen Tests dient die Wahrscheinlichkeit des Fehlers zweiter Art, unter mehreren Tests mit gleichem Fehler erster Art den günstigsten, also mit dem geringsten Fehler zweiter Art, herauszufinden. Hierauf wird hier nicht eingegangen. Hier wird die Rolle der Nullhypothese nochmals sichtbar. Es muss die Verteilung der Grundgesamtheit unter H_0 bekannt sein; nur dann lässt sich ein Fehler erster Art berechnen.

Ein Beispiel soll das nochmals verdeutlichen, insbesondere, welche Annahmen formuliert und bei jeder Anwendung überprüft bzw. als erfüllt angesehen werden. Es werde behauptet, dass bei einer neuen Sprachlernmethode weniger Fehler gemacht werden als bei einer herkömmlichen Methode, bei der im Mittel μ=40 Punkte erreicht wurden. Dies ist als Nullhypothese zu wählen. Als Gegenhypothese wählt man dann, dass die neue Methode besser ist, also dass mh als 40 Punkte erreicht werden. Die Annahmen über die Standardabweichung sind erforderlich, um eine Verteilung unter H_0 definieren zu können. Definiert man μ_0 als den Erwartungswert dieser Verteilung. so lautet die Nullhypothese, dass μ_0 gleich 40 ist.

Die Standardabweichung von diesem Wert 40 Punkte betrage $\overline{\sigma} = 8$ Punkte. Aus der Gruppe der nach der neuen Methode unterrichteten Schüler wird ein einfache Stichprobe vom Umfang n=100 erhoben und hierfür ein Stichprobenmittel von $\overline{X}$=42 Punkten bestimmt. Da die alte Lernmethode widerlegt werden soll, wählt man ihr Ergebnis für die Nullhypothese , also $H_0 : \mu < 40$ und $H_1 \mu \geq 40$. Denn man will ja, wenn H_0 gilt, die Wahrscheinlichkeit für einen Gegenbefund (das Ergebnis der neuen Lernmethode) möglichst klein halten. Wenn dann trotzdem der Gegenbefund (wenn man H_0 annimmt) eintritt, ist dessen Ergebnis so unwahrscheinlich (entsprechend dem gewählten Signifikanzniveau), dass man H_0 ablehnen muss.

Setzt man die Ergebnisse nach der alten Lernmethode als normalverteilt voraus, ist das Stichprobenmittel $\overline{X}$ jeweils N(40, $\frac{64}{100}$) verteilt (kann man beweisen). Man fragt nun nach der Wahrscheinlichkeit für 42 Punkte nach der alten Lernmethode (also wenn H_0 gilt). Intuitiv ist klar, dass die Abweichung der Punktzahl nach der neuen Methode von den 40 Punkten der alten Methode ein gutes Maß ist, allerdings muss dieses noch statistisch betrachtet werden. Eine hohe Abweichung heißt noch lange nicht, dass die neue Methode tatsächlich besser ist.

Deshalb wird ide Sache statistisch betrachtet:

Es ist $N_0 := \frac{\overline{X}-40}{8} \cdot 10$ somit N(0,1) verteilt und mit $\overline{X}$=42 erhält man N_0=2,5. Die Wahrscheinlichkeit für diesen Wert beträgt 0,0062 unter der Annahme einer Normalverteilung. Dies ist aber gerade die Wahrscheinlichkeit für 42 Punkte nach der alten Lernmethode. Diese Punktzahl ist aber so unwahrscheinlich, dass man H_1

annimmt und die neue Lernmethode vorzieht, wenn man das Signifikanzniveau größer als diese Wahrscheinlichkeit gewählt hat (also etwa 1 Prozent). Natürlich muss das Signifikanzniveau vor der Durchführung eines Tests gewählt werden. Statistisch nicht korrekt ist es, erst nach der berechnung der Wahrscheinlichleit des Stichprobenbefundes das Signifikanzniveau zu bestimmen oder gar zu ändern! Eine andere wichtige Bemerkung ist die, dass durch Ablehnen oder Annehmen einer der beiden Hypothesen nichts bewiesen wird! Die beiden möglichen Fehlerarten werden nicht ausgeschaltet, sondern nur in ihrer Wahrscheinlichleit kontrolliert.

Wie führt man nun allgemein solche Tests durch? Üblicherweise werden Tests in Rezepten formuliert, welche für jede Art von Stichprobenbefund in Abhängigkeit vom Signifikanzniveau angeben, wie man entscheiden muss.

Beispiel 8.1. *Es wird behauptet, dass das mittlere Einkommen der Studierenden 1000 DM im Monat betrage. Dieses Einkommen sei N($\mu, 120^2$) verteilt. Eine einfache Stichprobe vom Umfang 100 zum Einkommen ergab einen mittleren Betrag von 1060 DM. Es soll überprüft werden, ob die 1000 DM noch korrekt sind oder ob der Stichprobenbefund darauf hinweist, dass das mittlere Einkommen gestiegen ist.*

Vorgehensweise: Wahl von H_0 und H_1.

$H_0: \quad \mu_0 \leq 1000 \qquad H_1: \quad \mu_0 > 1000$

Wahl eines Signifikanzniveaus $\alpha = 0,05$ (in diesem Fall oder auch 0,01). Für das Signifikanzniveau werden in der Regel Standardwerte 0,1, 0,01 oder 0,05 gewählt.

Aus einer N(0,1) Tabelle ist der Wert$\lambda_{1-\alpha}$ so zu bestimmen, dass

$$P\{N(0,1) \leq \lambda_{1-\alpha}\} = 1 - \alpha$$

In diesem Fall ergibt sich $\lambda_{1-\alpha}$= 1,65 (sog. kritischer Wert, Schwellenwert). Die Werte $\lambda_{1-\alpha}$ heißen auch Quantile. Die N(0,1) Tabelle kann gewählt werden, da für die Verteilungen eine Normalverteilung angenommen wurde.

Berechnung von $N_0 = \frac{\overline{X}-\mu_0}{\sigma_0} \cdot \sqrt{n} = \frac{1060-1000}{120} \cdot 10 = 5$.

Entscheiden zugunsten einer der Hypothesen durch Vergleich mit dem kritischen Wert, hier H_0 ablehnen, wenn N_0 den kritischen Wert übersteigt. Dies ist der Fall beim Signfikanzniveau $\alpha = 0,05$), somit wird H_0 abgelehnt.

In diesem Beispiel liegt eine einseitige Fragestellung vor, denn μ wurde nur nach oben hin überprüft. Es gibt auch zweiseitige Fragestellungen $H_0 \quad : \mu = \mu_0 \qquad H_1 :$ $\mu \neq \mu_0$. Die möglichen Tests hängen vom Skalenniveau der Zufallsvariablen ab.

Im Folgenden werden für einige mögliche Tests die Testrezepte angegeben. Die auf ein Probleme anzuwendenden Tests hängen von der zugrude liegenden Verteilung ab (wird jeweils zu Beginn eines Rezeptes angegeben) und natürlich vom Skalenniveau der Zufallsvariablen.

8.0.3. *Einstichprobentests.*

Testrezept 8.2. *(Tests über μ bei bekanntem σ_0^2 einer N(μ, σ_0^2) verteilten Zufallsvariablen (Gauß-Test))*

	Fall 0	*Fall 1*	*Fall 2*
$H_0:$	$\mu = \mu_0$	$\mu \leq \mu_0$	$\mu \geq \mu_0$
$H_1:$	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$

Berechne aus N(0,1) -Tabelle Werte $\lambda_{1-\frac{\alpha}{2}}, \lambda_{1-\alpha}$ *mit:*

$$P\{N(0,1) \leq \lambda_{1-\alpha}\} = 1-\alpha$$
$$P\{N(0,1) \leq \lambda_{1-\frac{\alpha}{2}}\} = 1-\frac{\alpha}{2}$$

Berechne die Testgröße

$$N_0 := \frac{\overline{X}-\mu_0}{\sigma_0} \cdot \sqrt{n}$$

Entscheide wie folgt:

	Fall 0	*Fall 1*	*Fall 2*
H_0 *annehmen*	$-\lambda_{1-\frac{\alpha}{2}} \leq N_0 \leq \lambda_{1-\frac{\alpha}{2}}$	$N_0 \leq \lambda_{1-\alpha}$	$N_0 \geq -\lambda_{1-\alpha}$
H_1 *annehmen*	$\mid N_0 \mid > \lambda_{1-\frac{\alpha}{2}}$	$N_0 > \lambda_{1-\alpha}$	$N_0 < -\lambda_{1-\alpha}$

Dieser Test wurde in den zwei vorangegangenen Beispielen angewendet. Ist nun die Varianz der zugrunde liegenden Zufallsvariablen unbekannt, so kann der Gauß-Test nicht mehr verwendet werden. Stattdessen muss der folgende Test benutzt werden:

Testrezept 8.3. *(Test über* μ *bei unbekanntem* σ^2*)*

	Fall 0	*Fall 1*	*Fall 2*
$H_0:$	$\mu = \mu_0$	$\mu \leq \mu_0$	$\mu \geq \mu_0$
$H_1:$	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$

Berechne aus t_{n-1}*-Tabelle Werte* $\lambda_{1-\frac{\alpha}{2}}, \lambda_{1-\alpha}$.
Berechne die Testgröße

$$t_0 := \frac{\overline{X}-\mu_0}{S} \cdot \sqrt{n}$$

mit $S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Entscheide wie folgt:

	Fall 0	*Fall 1*	*Fall 2*
H_0 *annehmen*	$\lvert t_0 \rvert \leq \lambda_{1-\frac{\alpha}{2}}$	$t_0 \leq \lambda_{1-\alpha}$	$t_0 \geq \lambda_{1-\alpha}$
H_0 *annehmen*	$\lvert t_0 \rvert > \lambda_{1-\frac{\alpha}{2}}$	$t_0 > \lambda_{1-\alpha}$	$t_0 < -\lambda_{1-\alpha}$

Liegt ein Test über den Parameter einer binomialverteilten Zufallsvariable (etwa die Häufigkeit eines Merkmals) vor, so kann man folgenden Test benutzen:

Testrezept 8.4 (Binomialtest mit Normalverteilungsapproximation)**.** *Gegeben sei eine B(n,p) verteilte Zufallsvariable Y. Es sei* $X = (X_1, \ldots, X_n)$ *eine einfache Stichprobe zu Y vom Umfang n. Sei* $\alpha \in]0,1[$ *das Signifikanzniveau und* $p_0 \in]0,1[$ *mit* $n > \frac{9}{p_0(1-p_0)}$*. Folgende Fälle werden unterschieden:*

	Fall 0	*Fall 1*	*Fall 2*
H_0	$p = p_0$	$p \leq p_0$	$p \geq p_0$
H_1	$p \neq p_0$	$p > p_0$	$p < p_0$

Berechne aus N(0,1) Tabelle $\lambda_{1-\frac{\alpha}{2}}, \lambda_{1-\alpha}$ *und*

$$b_0 := \frac{\sum_{i=1}^{n} x_i - np_0}{\sqrt{np_0(1-p_0)}}$$

Entscheide wie folgt:

	Fall 0	*Fall 1*	*Fall 2*
H_0 *annehmen*	$-\lambda_{1-\frac{\alpha}{2}} \leq b_0 \leq \lambda_{1-\frac{\alpha}{2}}$	$b_0 \leq \lambda_{1-\alpha}$	$b_0 \geq -\lambda_{1-\alpha}$
H_1 *annehmen*	$\lvert b_0\rvert > \lambda_{1-\frac{\alpha}{2}}$	$b_0 > \lambda_{1-\alpha}$	$b_0 < -\lambda_{1-\alpha}$

Der Ausdruck Normalverteilungsapproximation sagt aus, dass nicht mit den Quantilen einer Binomial-verteilten Zufallsvariablen gerechnet wird, sondern dass bei den genannten Voraussetzungen die Binomialverteilung hinreichend gut durch eine Normalverteilung approximiert werden kann. Hierzu ein Beispiel:

Beispiel 8.5. *Es wird vermutet, dass 50 Prozent aller Studierenden der Computerlinguistik das Fach gut finden. Eine Umfrage ergibt bei 900 Studierenden der Computerlinguistik eine Zahl von 540 Zustimmungen zum Fach. Kann auf einem Signifikanzniveau von $\alpha = 0{,}05$ die erste Meinung noch gehalten werden?*

Test:

$$H_0: \quad p \leq 0,5 \qquad H_1: \quad p > 0,5$$

Es gilt: n=900 $> \frac{9}{0{,}25} = 36$, also ist die Voraussetzung erfüllt. Für b_0 ergibt sich ein Wert von 6, für das Quantil (einseitiger Test) $\lambda_{1-\alpha}$ ein Wert von 1,645. Somit liegt für den Fall 1 b_0 im kritischen Bereich, also wird H_0 abgelehnt.

8.0.4. *Zweistichprobentests.* Mit den vorgestellten Tests kann man Hypothesen über eine vorliegende Grundgesamtheit testen. Was fehlt, sind noch Tests über Hypothesen, die zu Zufallsvariablen X und Y aus verschiedenen Grundgesamtheiten gehören. Hier unterscheidet man, ob die Stichproben verbunden sind, dass heißt, die Elemente der Stichprobe sind sich paarweise einander zugeordnet, oder ob sie unabhängig voneinander sind. Paarweise einander zugeordnet sind Stichproben zum Beispiel, wenn dieselben Personen vor und nach einem Experiment befragt werden. Weiterhin ist wichtig, wie die Varianzen der beiden Zufallsvariablen zueinander stehen. Man unterscheidet:

(1) $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 beide bekannt

(2) $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 beide unbekannt

(3) $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ σ_X^2, σ_Y^2 beide unbekannt, aber gleich

(4) X, Y beliebig verteilt, aber beide Stichprobenumfänge ≥ 30.

Hier werden nur der erste und letzte Fall vorgestellt.

Testrezept 8.6. *Tests über die Erwartungswerte zweier gemeinsam normalverteilter Zufallsvariablen mit bekannten Varianzen und unverbundenen Stichproben*

Es seien $X \sim N(\mu_X, \sigma_X^2)$ und $Y \sim N(\mu_Y, \sigma_Y^2)$

Es sei $(X_1, X_2), \cdots, X_n)$ eine einfache Stichprobe zu X und es sei $(Y_1, Y_2), \cdots, X_m)$ eine einfache Stichprobe zu Y.

Man wähle ein Signifikanzniveau $\alpha \in]0, 1[$.
Man unterscheidet die 3 Fälle:

	Fall 0	*Fall 1*	*Fall 2*
H_0:	$\mu_X = \mu_Y$	$\mu_1 \geq \mu_2$	$\mu_1 \leq \mu_2$
H_1:	$\mu_1 \neq \mu_2$	$\mu_1 - \mu_2 > \theta$	$\mu_1 - \mu_2 < \theta$

Man berechne aus N(0,1)-Tabelle die Quantile $\lambda_{1-\alpha}$ sowie $\lambda_{1-\frac{\alpha}{2}}$.
Die Testgröße ist

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Unter H_0 ist die Testgröße N(0,1)-verteilt.

Man entscheide für die Ablehnung von H_0, wenn:

(1) *Fall 0:* $\mid Z \mid > \lambda_{1-\frac{\alpha}{2}}$
(2) *Fall 1:* $Z < -\lambda_{1-\alpha}$
(3) *Fall 2:* $Z > \lambda 1 - \alpha$

Da im folgenden Test keine Verteilungsannahme zu treffen ist, kann er auch für Sprachanalysen genutzt werden.

Testrezept 8.7. *X und Y seien Zufallsvariable mir $EX = \mu_X, EY = \mu_Y . X_1, \cdots, X_n$ sei einfache Stichprobe zu X und $Y_1, \cdots, Y_m$ sei einfache Stichprobe zu Y. Zusätzlich seien $n, m \geq 30$. Folgende drei Fälle werden unterschieden:*

	Fall 0	*Fall 1*	*Fall 2*
$H_0:$	$\mu_X = \mu_Y$	$\mu_1 \geq \mu_2$	$\mu_1 \leq \mu_2$
$H_1:$	$\mu_1 \neq \mu_2$	$\mu_1 - \mu_2 > \theta$	$\mu_1 - \mu_2 < \theta$

Die Teststatistik ist

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{\S_Y^2}{m}}}$$

Unter H_0 ist Z approximativ N(0,1) verteilt. H_0 wird abgelehnt, wenn

(1) *Fall 0:* $\mid Z \mid > \lambda_{1-\frac{\alpha}{2}}$
(2) *Fall 1:* $Z < -\lambda_{1-\alpha}$
(3) *Fall 2:* $Z > \lambda 1 - \alpha$

Ein Beispiel hierzu wird in der Vorlesung gegeben.

8.0.5. *Verteilungsfreie Tests.* Bisher wurden Parameter von Verteilungen getestet, jetzt soll die Verteilung der Grundgesamtheit, also die Verteilung der wahren Zufallsvariable Y getestet werden. Dies geschieht z. B. in Chi-Quadrat-Tests. Man bestimmt die empirische Verteilungsfunktion der Stichprobe und prüft, ob diese die wahre Verteilungsfunktion genügend gut annähert. Hierzu vergleicht man die durch die Nullhypothese zu erwartenden Häufigkeiten in Intervallen mit denen durch die Realisationen der Stichprobe beobachteten Häufigkeiten. Man unterteilt die x-Achse (die Menge der Realisationen) in Teilintervalle $I_1, \ldots, I_k$.

Testrezept 8.8 (Chi-Quadrat- Anpassungstest bei kategorialem Merkmal)**.** *$X_1, \cdots, X_n$ seien eine einfache Stichprobe zu X. X nehme nur Werte aus einer Menge $\{1, \cdots, k\}$ an.*

Das Signifikanzniveau sei $\alpha \in]0,1[$.
Hypothesen:

$H_0 : P(X = i) = p_1 \quad i = 1, ..., k \qquad H_1 : P(X = i) \neq p_i$ *für mindestens ein i*

Die Teststatisik ist nun

$$\chi_0^2 := \sum_{i=1}^{k} \frac{(h_i - n \cdot p_i)^2}{np_i}$$

Unter H_0 ist diese approximativ $\chi^2(k-1)$. Diese Approximation ist anwendbar, wenn $np_i \geq 1$ für alle i und $np_i \geq 5$ für mindestens 80 Prozent der Zellen. Entscheide wie folgt:

H_0 annehmen: $\chi_0^2 \leq \lambda_{1-\alpha}$
H_1 annehmen: $\chi_0^2 > \lambda_{1-\alpha}$

Beispiel 8.9. *Es wird vermutet, dass sich in einer Sprache die Häufigkeit von Verben zu Adjektiven wie 3:1 verhält. In einem Text werden 355 Verben und 123 Adjektive gezählt. Kann die Vermutung bei diesem Stichprobenbefund auf einem 5 Prozent Signifikanzniveau aufrecht erhalten werden?*

Der Umfang der Stichprobe beträgt 478. X bezeichne die Anzahl der Verben in der Stichprobe. Wenn ein Verb an der i-ten Stelle auftritt, sei dies mit X=1 bezeichnet. Dann lauten die Hypothesen:

$H_0 : P(X = 1) = 0,75 \quad P(X = 0) = 0,25$
$H_1 : P(X = 1) \neq 0,75 \quad \text{oder} \quad P(X = 0) \neq 0,25$

Für die zu erwartende Anzahl (also die theoretische Häufigkeit unter H_0) ergibt sich: $np_1^0 = \frac{3 \cdot 478}{4} = 358,5$ für die Verben und $np_2^0 = \frac{478}{4} = 119{,}5$ für die Adjektive. Für die Teststatistik ergibt sich:

$$\chi_0^2 = 0,1367$$

und aus der Tafel für die Chi-Quadrat-Verteilung (1 Freiheitsgrad) entnimmt man: $\lambda_{1-0,05}^2$=3,84. Somit kannH_0 nicht verworfen werden.

Mit einem Chi-Quadrat-Test kann man auch testen, ob 2 gegebene Zufallsvariablen unabhängig sind. Z.B. kann man bei einer Stichprobe nicht nur die Größe, sondern auch das Gewicht ermitteln und dann testen, ob Gewicht und Größe voneinander unabhängig sind. Insbesondere kann hiermit das Vorliegen einer Kollokation überprüft werden.

Dieser Test heißt dann χ^2-Unabhängigkeitstest.

Testrezept 8.10. *χ^2- Unabhängigkeitstest*
Zwei intessierende Merkmale X, Y, in einer Kontingenztabelle mit k Zeilen und m Spalten gruppiert(k Ausprägungen für X, m für Y). Eine Stichprobe vom Umfang n zu (X,Y) erhoben. h_{ij} absolute Häufigkeit des Auftretens der Merkmalskombination (X=i, Y=j).

Y / X	y_1	y_2	$\ldots$	y_m	
x_1	h_{11}	h_{12}	$\ldots$	h_{1m}	$h_{1\cdot}$
x_2	h_{21}	h_{22}	$\ldots$	h_{2m}	$h_{2\cdot}$
....					
x_k	h_{k1}	h_{k2}	$\ldots$	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$		$h_{\cdot m}$	n

Signifikanzniveau $\alpha \in]0, 1[$, Hypothesen:
H_0 : X und Y sind unabhängig gegen H_1 : X und Y sind abhängig
Teststatistik:

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

mit $\tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$
Unter H_0 ist χ_0^2 approximativ χ^2-verteilt mit (k-1)(m-1) Freiheitsgraden.
Entscheidung: H_0 ablehnen, wenn $\chi_0^2 > \chi_{1-\alpha}^2(k-1)(m-1)$

Beispiel 8.11. *Es soll geprüft werden, ob zwischen dem Einkommen Y von Personen und der Wahl einer bestimmten Marke M ein Zusammenhang besteht. Die Daten seien in folgender Tabelle zusammengestellt:*

jährliches Einkommen	*Marke 1*	*Marke 2*	*Summe*
$Y_1 :< 30.000\ DM$	*150*	*50*	*200*
$Y_2 :\geq 30.000\ DM$	*40*	*60*	*100*
Summe	*190*	*110*	*300*

Die Hypothesen lauten:

H_0 : Markenwahl und Einkommen sind voneinander unabhängig

H_1 : Markenwahl und Einkommen sind nicht unabhängig voneinander

Zur Berechnung der Teststatistik:
Es wird unterschieden zwischen den (unter Gültigkeit von H_0) zu erwartenden Häufigkeiten und den beobachteten Häufigkeiten. Die zu erwartenden Häufigkeiten werden wie folgt berechnet: Wenn H_0 gilt, also die Unabhängigkeit der beiden Merkmale, so ist die gemeinsame Verteilung das Produkt der Randverteilungen. Für die Randverteilungen erhält man:

$$Y_{1\cdot} = \frac{200}{300} \quad Y_{2\cdot} = \frac{100}{300} \quad M_{\cdot 1} = \frac{190}{300} \quad M_{\cdot 2} = \frac{110}{300}$$

Hieraus sind jetzt noch die zu erwartenden absoluten Häufigkeiten $p_{ij} = 300 \cdot Y_{i\cdot} \cdot M_{\cdot j}$ zu errechnen: Diese sind gerade die Anteile der Randverteilungen an dem Gesamtumfang der Stichprobe, also:

$$p_{11} = 126,67 \quad p_{12} = 73,33 \quad p_{21} = 63,33 \quad p_{22} = 36,67$$

Jetzt wird die Teststatistik berechnet:

$$\chi_0^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(\nu_{ij} - p_{ij})^2}{p_{ij}} = 35,22.$$

Diese Teststatistik ist unter H_0 asymptotisch Chi-Quadrat-verteilt mit (2-1)(2-1) Freiheitsgraden. Bei einem Signifikanzniveau von 1 Prozent ergibt sich ein Wert für das Quantil von 6,63. Man lehnt H_0 ab, wenn der Wert der Teststatistik größer als das Quantil ist, was hier der Fall ist.

Insbesonere ist dieser Test dazu geeignet, Kollokationen herauszufinden. Das Auftreten der beiden Wörter einer Kolllokation ist nicht zufällig, sondern es hängt sehr stark von den beiden Wörtern ab. Dies kann mit Hilfe des obigen Tests überprüft werden.

Es folgt jetzt noch eine Übersicht über die verschiedenen Tests.

Testrezepte

n bzw. m bezeichnen stets den Stichprobenumfang.

(1) Tests über μ bei bekanntem σ_0^2 einer $N(\mu, \sigma_0^2)$ verteilten Zufallsvariablen (Gauß-Test)

	Fall 0	Fall 1	Fall 2
H_0 :	$\mu = \mu_0$	$\mu \leq \mu_0$	$\mu \geq \mu_0$
H_1 :	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$

Berechne aus N(0,1) -Tabelle Werte $\lambda_{1-\frac{\alpha}{2}}, \lambda_{1-\alpha}$ mit:

$$P\{N(0,1) \leq \lambda_{1-\alpha}\} = 1-\alpha$$
$$P\{N(0,1) \leq \lambda_{1-\frac{\alpha}{2}}\} = 1-\frac{\alpha}{2}$$

Testgröße

$$N_0 := \frac{\overline{X}-\mu_0}{\sigma_0} \cdot \sqrt{n}$$

$\sim N(0,1)$unter H_0 Entscheide:

	Fall 0	Fall 1	Fall 2
H_0 annehmen	$-\lambda_{1-\frac{\alpha}{2}} \leq N_0 \leq \lambda_{1-\frac{\alpha}{2}}$	$N_0 \leq \lambda_{1-\alpha}$	$N_0 \geq -\lambda_{1-\alpha}$
H_1 annehmen	$\mid N_0 \mid > \lambda_{1-\frac{\alpha}{2}}$	$N_0 > \lambda_{1-\alpha}$	$N_0 < -\lambda_{1-\alpha}$

(2) Binomialtest mit Normalverteilungsapproximation (Approximativer Binomialtest)
Test über Parameter p einer Binomialverteilung; Voraussetzung: $np \geq 5$ und $n(1-p) \geq 5$

	Fall 0	Fall 1	Fall 2
$H_0:$	$p = p_0$	$p \leq p_0$	$p \geq p_0$
$H_1:$	$p \neq p_0$	$p > p_0$	$p < p_0$

Teststatistik:

$$Z = \frac{X-np_0}{\sqrt{np_0(1-p_0)}}$$

X ist die Anzahl des Auftretens des interessierenden Ereignisses. Berechne zu $\alpha \in]0,1[$ aus N(0,1) Tabelle Werte $\lambda_{1-\alpha}$ bzw $\lambda_{1-\frac{\alpha}{2}}$ mit

$$P\{N(0,1) \leq \lambda_{1-\alpha}\} = 1-\alpha$$
$$P\{N(0,1) \leq \lambda_{1-\frac{\alpha}{2}}\} = 1-\frac{\alpha}{2}$$

Entscheide wie folgt:
Fall 0: H_0 annehmen, wenn $|Z| \leq \lambda_{1-\frac{\alpha}{2}}$, sonst H_1 annehmen
Fall 1: H_0 annehmen, wenn $Z \leq \lambda_{1-\alpha}$, sonst H_1 annehmen
Fall 2: H_0 annehmen, wenn $Z \geq -\lambda_{1-\alpha}$, sonst H_1 annehmen

(3) t-Test: Test über μ bei unbekanntem σ^2, Grundgesamtheit $N(\mu_X, \sigma^2)$ verteilt mit unbekanntem σ^2.

	Fall 0	Fall 1	Fall 2
$H_0:$	$\mu = \mu_0$	$\mu \leq \mu_0$	$\mu \geq \mu_0$
$H_1:$	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$

Berechne aus t_{n-1}-Tabelle Werte $\lambda_{1-\frac{\alpha}{2}}, \lambda_{1-\alpha}$.
Testgröße

$$t_0 := \frac{\overline{X}-\mu_0}{\sqrt{S^2}} \cdot \sqrt{n}$$

mit $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$. Entscheide wie folgt:

	Fall 0	Fall 1	Fall 2
H_0 annehmen	$\lvert t_0 \rvert \leq \lambda_{1-\frac{\alpha}{2}}$	$t_0 \leq \lambda_{1-\alpha}$	$t_0 \geq -\lambda_{1-\alpha}$
H_1 annehmen	$\lvert t_0 \rvert > \lambda_{1-\frac{\alpha}{2}}$	$t_0 > \lambda_{1-\alpha}$	$t_0 < -\lambda_{1-\alpha}$

(4) t-Test auf Gleichheit zweier Mittelwerte bei unverbundenen Stichproben:
2 unverbundene Stichproben $X = X_1, X_2, ..., X_n$ (Umfang n) und $Y = Y_1, ..., Y_m$ (Umfang m), wobei jedes $X_i \sim N(\mu_X, \sigma_X^2)$ und jedes $Y_j \sim N(\mu_Y, \sigma_Y^2)$ und $\sigma_X^2 = \sigma_Y^2$ unbekannt.
Hypothesen: $H_0 : \mu_X = \mu_Y$ gegen $H_1 : \mu_x \neq \mu_Y$
Teststatistik:

$$t_0 = \sqrt{\frac{nm(m+n-2)}{n+m}} \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2(n-1) + S_Y^2(m-1)}}$$

Die Teststatistik ist unter H_0 t-verteilt mit (m+n-2) Freiheitsgraden.
Bestimme aus t-Tabelle mit n+m-2 Freiheitsgraden zu gegebenem Signifikanzniveau α ein $\lambda_{1-\frac{\alpha}{2}}$ mit

$$P\{t_{n+m-2} \leq \lambda_{1-\frac{\alpha}{2}}\} = 1 - \frac{\alpha}{2}$$

Entscheidung: H_0 ablehnen wenn $\mid t_0 \mid > t_{1-\frac{\alpha}{2}}$
Sind n und m sehr groß ($\geq$30), können die Quantile der N(0,1)-Verteilung benutzt werden, da die Teststatistik dann approximativ normalverteilt ist.

(5) χ^2-Test bei kategorialem Merkmal
Es liege eine Stichprobe vom Umfang n vor. Trägerpunkte von X $\in \{1, ..., k\}$
Hypothesen:
$H_0 : P(X = i) = p_i, \qquad i = 1, 2, \ldots, k$
$H_1 : P(X = i) \neq p_i, \qquad$ für mindestens ein i $\in \{1, 2, \ldots, k\}$
Teststatistik:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(h_i - np_i)^2}{np_i}$$

Unter H_0 ist die Teststatistik approximativ χ^2-verteilt mit (k-1) Freiheitsgraden, wenn $np_1 \geq 1$ für alle i und $np_i \geq 5$ für mindestens 80 Prozent der Zellen.
Entscheidung: H_0 ablehnen, wenn $\chi_0^2 > \chi_{1-\alpha}^2(k-1)$

(6) χ^2- Unabhängigkeitstest
Zwei intessierende Merkmale X, Y, in einer Kontingenztabelle mit k Zeilen und m Spalten gruppiert(k Ausprägungen für X, m für Y). Eine Stichprobe vom Umfang n zu (X,Y) erhoben. h_{ij} absolute Häufigkeit des Auftretens der Merkmalskombination (X=i, Y=j).

Y / X	y_1	y_2	...	y_m	
x_1	h_{11}	h_{12}	...	h_{1m}	$h_{1\cdot}$
x_2	h_{21}	h_{22}	...	h_{2m}	$h_{2\cdot}$
....					
x_k	h_{k1}	h_{k2}	...	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$		$h_{\cdot m}$	n

Signifikanzniveau $\alpha \in]0, 1[$, Hypothesen:
H_0 : X und Y sind unabhängig gegen H_1 : X und Y sind abhängig
Teststatistik:

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

mit $\tilde{h}_{ij} = \frac{h_{i.}h_{.j}}{n}$
Unter H_0 ist χ_0^2 approximativ χ^2-verteilt mit (k-1)(m-1) Freiheitsgraden.
Entscheidung: H_0 ablehnen, wenn $\chi_0^2 > \chi_{1-\alpha}^2(k-1)(m-1)$

(7) χ^2- Homogenitätstest
Überprüft, ob k Verteilungen identisch sind. Ein Merkmal X mit k Kategorien wird in m Populationen (evtl. verschiedene Stichpobenumfänge) erhoben. Ergebnisse in Kontingenztabelle:

Ausprägungen / Population	1	2	...	m	
x_1	h_{11}	h_{12}	...	h_{1m}	n_1
x_2	h_{21}	h_{22}	...	h_{2m}	n_2
....					
x_k	h_{k1}	h_{k2}	...	h_{km}	n_k
	$h_{.1}$	$h_{.2}$		$h_{.m}$	n

Bedeutung der h_{ij} wie im Unabhängigkeitstest, $n = \sum_{i=1}^{k} n_i$, Signifikanzniveau $\alpha \in]0,1[$.
Hypothesen:
H_0: $P\{X_1 = j\} = P\{X_2 = j\} = \cdots = P\{X_k = j\}, \qquad j = 1, \cdots, m$
H_1: $P\{X_{i1} = j\} \neq P\{X_{i2} = j\}$, für mindestens ein Tupel $(i1, i2, j)$
Teststatistik:

$$\chi_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

mit $\tilde{h}_{ij} = \frac{n_i h_{.j}}{n}$
Unter H_0 ist χ_0^2 approximativ χ^2-verteilt mit (k-1)(m-1) Freiheitsgraden.
Entscheidung: H_0 ablehnen, wenn $\chi_0^2 > \chi_{1-\alpha}^2(k-1)(m-1)$

Literaturliste

(1) Manning-Schütze: Foundations of statistical language processing, MIT Press, Cambridge 1999

(2) Jurafsky-Martin: An Introduction to Natural Language Processing, Computational Lingistics, ans Speech Recognition, Prentice Hall Series 2000

(3) Carstensen-Ebert-Endriss-Jekat-Klabunde-Langer (Hrg.): Computerlinguistik und Sprachtechnologie, Spektrum Akademischer Verlag Heidelberg-Berlin, 2001

(4) Krenn-Samuelsson: The Linguist´s Guide to Statistics, Homepage der Coli Skript, Dezember 1997

(5) Fahrmeir-Künstler-Pigeot-Tutz: Statistik, der Weg zur Datenanalyse, Springer Berlin 1997

(6) Kreyszig: Statistische Methoden und ihre Anwendungen, Vandenhoeck-Ruprecht Göttingen, 1999

(7) Bortz: Statistik für Sozialwissenchaftler, Springer Berlin 1999

(8) Bamberg- Baur: Statistik Arbeitsbuch, Oldenbourg München Wien 1994

Prof. Dr. Enrico Lieblang, HTW des Saarlandes

E-mail address: enrico.lieblang@htw-saarland.de