

Mathematik für Informatiker II

Teil F: Stochastik

Prof. Dr. Joachim Weickert
L^AT_EX von Sarah Diehl und Christian Spaniol

Sommersemester 2004



Inhaltsverzeichnis

61 Grundbegriffe	6
61.1 Motivation	6
61.2 Gebietsabgrenzung	6
61.3 Definition (Wahrscheinlichkeit)	6
61.4 Definition (Laplace-Experiment)	7
61.5 Mengentheoretische Wahrscheinlichkeitsbeschreibung	7
62 Kombinatorik	8
62.1 Motivation	8
62.2 Zwei grundlegende Sprechweisen	8
62.3 Produktregel der Kombinatorik	9
62.4 Die vier kombinatorischen Grundsituationen	9
63 Erzeugende Funktionen	11
63.1 Motivation	11
63.2 Permutationen und Kombinationen	11
63.3 Beispiel einer erzeugenden Funktion	11
63.4 Definition (Erzeugende Funktion)	11
63.5 Beispiel	11
63.6 Beispiel	12
63.7 Beispiel	12
63.8 Satz (Kombinationen mit vorgegebenen Wiederholungen)	13
63.9 Satz (Kombinationen mit beliebigen Wiederholungen)	13
63.10 Bemerkungen	13
63.11 Definition (exponentiell erzeugende Funktion)	13
63.12 Bedeutung für Permutationen	14
63.13 Satz (Permutationen mit vorgegebenen Wiederholungen)	14
63.14 Satz (Permutationen mit beliebigen Wiederholungen)	14
64 Bedingte Wahrscheinlichkeiten	15
64.1 Motivation	15
64.2 Beispiel	15
64.3 Verallgemeinerung	15
64.4 Beispiel	16
64.5 Beispiel	16
64.6 Satz (Satz von der totalen Wahrscheinlichkeit)	16
64.7 Beispiel	17
64.8 Satz (Formel von Bayes)	17
64.9 Anwendungsbeispiele	18

65 Zufallsvariable, Erwartungswert, Varianz	19
65.1 Motivation	19
65.2 Definition (Zufallsvariable)	19
65.3 Beispiel	19
65.4 Definition (Verteilung)	19
65.5 Beispiel	19
65.6 Definition (Erwartungswert)	20
65.7 Beispiel	20
65.8 Satz (Linearität des Erwartungswerts)	20
65.9 Unabhängigkeit zweier Zufallsvariablen	21
65.10 Beispiele	21
65.11 Definition (Varianz)	22
65.12 Berechnung der Varianz (Verschiebungssatz)	22
65.13 Beispiel	22
65.14 Satz (Eigenschaften der Varianz)	23
65.15 Standardisierte Zufallsvariablen	23
65.16 Satz (Ungleichung von Jensen)	24
65.17 Beispiele	24
65.18 Satz (Gleichung von Wald)	24
65.19 Beispiel	25
65.20 Definition (Kovarianz, Korrelationskoeffizient)	25
65.21 Bemerkungen	25
65.22 Satz (Rechenregeln für die Korrelation)	25
65.23 Bemerkungen	26
65.24 Zuverlässigkeitsanalyse von Systemen	26
66 Abschätzungen für Abweichungen vom Erwartungswert	28
66.1 Motivation	28
66.2 Definition ((zentrales) k-tes Moment)	28
66.3 Bemerkungen	28
66.4 Definition (Die Momenten erzeugende Funktion)	28
66.5 Satz (Eigenschaften Momenten erzeugender Funktion)	29
66.6 Satz (Markow'sche Ungleichung)	29
66.7 Folgerungen (Markow- und Tschebyschew-Ungleichung, Chernoff-Schranke)	30
66.8 Beispiel	30
66.9 Schwaches Gesetz der großen Zahlen	31
67 Wichtige diskrete Verteilungen	32
67.1 Motivation	32
67.2 Die Gleichverteilung	32
67.3 Beispiel	32
67.4 Die Binomialverteilung	32
67.5 Beispiel: Zufallsabhängigkeit sportlicher Resultate	33
67.6 Die Poissonverteilung	34
67.7 Reales Beispiel: Der große Jubiläumstag	34
67.8 Die geometrische Verteilung	34
67.9 Beispiel	35
68 Wichtige kontinuierliche Verteilungen	36
68.1 Motivation	36
68.2 Definition (Dichte)	36
68.3 Veranschaulichung	36
68.4 Definition (Verteilungsfunktion)	37
68.5 Beispiel: Kontinuierliche Gleichverteilung	37

68.6	Die Standardnormalverteilung	37
68.7	Die allgemeine Normalverteilung	38
68.8	Approximation der Binomialverteilung durch die Gaußverteilung	39
68.9	Beispiel	39
68.10	Satz (Zentraler Grenzwertsatz)	40
68.11	Bemerkungen	40
68.12	Tabelle der Standardnormalverteilung	41
69	Multivariate Verteilungen und Summen von Zufallsvariablen	42
69.1	Motivation	42
69.2	Wichtige Definitionen (gemeinsame Dichte, Erwartungswertvektor, Kovarianzmatrix)	42
69.3	Beispiel	43
69.4	Satz (Gemeinsame Dichte von unabhängigen Zufallsvariablen)	43
69.5	Beispiel	43
69.6	Definition (Faltung)	44
69.7	Satz (Summe unabhängiger kontinuierlicher Zufallsvariabler)	44
69.8	Satz (Summe unabhängiger normalverteilter Zufallsvariabler)	44
69.9	Definition (diskrete Faltung)	44
69.10	Satz (Summe unabhängiger diskreter Zufallsvariabler)	44
69.11	Satz (Summe unabhängiger Poisson-verteilter Zufallsvariabler)	45
69.12	Beispiel	45
70	Parameterschätzung und Konfidenzintervalle	46
70.1	Motivation	46
70.2	Definition (Stichprobe, Stichprobenwerte)	46
70.3	Beispiel	46
70.4	Definition (Mittelwert, Varianz, Standardabweichung einer Stichprobe)	46
70.5	Bemerkungen	47
70.6	Beispiel	47
70.7	Konfidenzintervalle	47
70.8	Beispiel: Wahlumfrage aus 70.6	47
70.9	Beispiel: Überbuchung eines Flugzeugs	48
71	Hypothesentests	50
71.1	Motivation	50
71.2	Parameterstest am Beispiel eines Münzexperimentes	50
71.3	Bemerkungen	50
71.4	Der Chi-Quadrat-Test	51
71.5	Beispiel	52
71.6	Bemerkung	52
71.7	Schranken für Chi-Quadrat bei f Freiheitsgraden	53
72	Methode der kleinsten Quadrate	54
72.1	Problemstellung	54
72.2	Methode der kleinsten Quadrate	54
72.3	Minimierung des quadratischen Fehlers	55
72.4	Bemerkungen	55
72.5	Satz (Pseudolösung überbestimmter Gleichungssysteme)	56
72.6	Bemerkung	56
72.7	Beispiel	56

73 Robuste Statistik	58
73.1 Motivation	58
73.2 Median	58
73.3 Bemerkungen	58
73.4 Empirischer Median	59
73.5 Beispiele	59
73.6 M-Schätzer	59
73.7 Beispiele	59
74 Fehlerfortpflanzung	61
74.1 Motivation	61
74.2 Satz (Fehlerfortpflanzungsgesetz von Gauß)	61
74.3 Beispiel	61
75 Markowketten	62
75.1 Motivation	62
75.2 Definition (stochastischer Prozess, Markowkette)	62
75.3 Beispiel mit Definitionen (Übergangsmatrix)	62
75.4 Definition (homogen, Kette mit stationären Übergangsws.)	63
75.5 Bemerkungen	63
75.6 Zustandsbeschreibung endlicher homogener Markowketten	64
75.7 Satz (Eigenwerte stochastischer Matrizen)	64
75.8 Bedeutung des Eigenwerts $\lambda = 1$	65
75.9 Satz (Potenzen stochastischer Matrizen)	66
75.10 Gegenbeispiel	66
75.11 Definition (im Gleichgewicht)	67
75.12 Beispiel	67
75.13 Definition (transitiv, irreduzibel)	68
75.14 Praktisches Kriterium für Irreduzibilität	68
75.15 Beispiele	68
76 Verborgene Markowmodelle	70
76.1 Motivation	70
76.2 Beispiel: Spieler mit fairer und unfairer Münze	70
76.3 Schwierigeres Beispiel	70
76.4 Definition (verborgenes Markowmodell, HMM)	71
76.5 Beispiel	71
76.6 Problemstellung (Dekodierungsproblem)	71
76.7 Der Viterbi-Algorithmus (findet lokales Maximum)	72
77 Pseudozufallszahlen und Monte-Carlo-Simulation	73
77.1 Motivation	73
77.2 Erzeugung von gleichverteilten Pseudozufallszahlen	73
77.3 Erzeugung von $N(0,1)$ -verteilten Pseudozufallszahlen	74
77.4 Beispiel (Quicksort)	75
77.5 Beispiel (Buffon'sches Nadelexperiment)	75

61 Grundbegriffe

61.1 Motivation

Stochastik (aus Griechischen: vermuten, erwarten) ist die Mathematik des Zufalls. Sie ist von großer Bedeutung in der Informatik. Beispiele:

- Analyse der Auslastung von Datennetzen
- Modellierung von Antwortzeiten im Rechner
- Zuverlässigkeitsanalyse von Hardware
- Raytracing in der Computergrafik (Monte-Carlo-Simulationen)
- Optimierungsalgorithmen (genetische Algorithmen, simulated annealing)
- Analyse der Laufzeit von Algorithmen
- Kombinatorische Probleme in der Bioinformatik

61.2 Gebietsabgrenzung

Stochastik gliedert sich in zwei Gebiete:

a) *Wahrscheinlichkeitstheorie*

Ausgehend von einem stochastischen Modell werden Wahrscheinlichkeiten berechnet.

Beispiel: Modellannahme: beim Wurf eines Würfels habe jede der Augenzahlen $1, \dots, 6$ die Wahrscheinlichkeit $\frac{1}{6}$.

Folgerung: Wahrscheinlichkeit für Augenzahl 1 oder 3 ist $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

b) *Statistik*

Ausgehend von realen Daten/Messungen möchte man Schlussfolgerungen ziehen.

Beispiele:

- möglichst gute Approximationsgerade durch fehlerbehaftete Messwerte legen
- Hypothesentest: Ist ein neues Medikament wirksam?

61.3 Definition

Die *Wahrscheinlichkeit* eines Ereignisses beschreibt die zu erwartende relative Häufigkeit, mit der dieses Ereignis eintritt, wenn man den zu Grunde liegenden Prozess immer wieder unter den gleichen Bedingungen wiederholt.

Beispiel: Bei einer „fairen Münze“ beträgt die Wahrscheinlichkeit (Ws.), Zahl zu erhalten, $\frac{1}{2}$.

61.4 Definition

Ein Experiment heißt *Laplace-Experiment*, wenn es endlich viele, einander ausschließende Ausgänge hat, die alle gleich wahrscheinlich sind.

Beispiel:

- Der Wurf eines Würfels hat 6 gleich berechnigte Ausgänge \Rightarrow Laplace-Experiment.
- Fällt ein Marmeladenbrot zu Boden, gibt es zwei Ausgänge, die nach Murphy nicht gleich wahrscheinlich sind \Rightarrow kein Laplace-Experiment, falls dies stimmt.

61.5 Mengentheoretische Wahrscheinlichkeitsbeschreibung

Definition: Wir betrachten ein Zufallsexperiment mit endlich vielen Ausgängen und definieren drei Begriffe:

- Ergebnismenge* Ω (*Stichprobenraum, Grundraum*)
endliche, nicht leere Menge, deren Elemente ω_i die Versuchsausgänge beschreiben.
Beispiel: Würfelexperiment: $\Omega = \{1, 2, \dots, 6\}$.
- Ereignis*:
Teilmenge von Ω .
Beispiel: Ergebnisse, bei denen 2 oder 5 gewürfelt wird. \Rightarrow Ereignis $A = \{2, 5\}$.
- Wahrscheinlichkeitsverteilung, Wahrscheinlichkeitsmaß*
Abbildung P von der Potenzmenge $\wp(\Omega)$ nach $[0, 1]$ mit folgenden Eigenschaften:
 - Normiertheit: $P(\Omega) = 1$.
 - Nichtnegativität: $P(A) \geq 0 \quad \forall A \in \wp(\Omega)$
 - Additivität: $P(A \cup B) = P(A) + P(B)$ für alle *disjunkten* $A, B \in \wp(\Omega)$

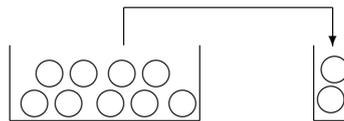
62 Kombinatorik

62.1 Motivation

Die Kombinatorik liefert wichtige Modelle zum Berechnen von Wahrscheinlichkeiten bei Laplace-Experimenten. Sie spielt eine grundlegende Rolle in der Informatik.

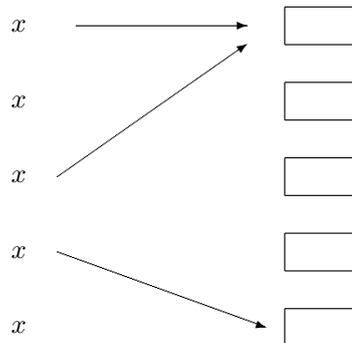
62.2 Zwei grundlegende Sprechweisen

a) *Stichprobensprechweise (Urnenmodell)*



Aus einer Urne mit n unterscheidbaren Kugeln werden k Kugeln gezogen. Dabei kann das Ziehen mit oder ohne Zurücklegen erfolgen und die Reihenfolge eine oder keine Rolle spielen.

b) *Zuordnungssprechweise (Schubladenmodell)*



k Objekte werden auf n Schubladen verteilt. Dabei sind die Objekte entweder unterscheidbar oder nicht unterscheidbar und die Schubladen dürfen einfach oder mehrfach besetzt werden.

Urnen- und Schubladenmodell sind äquivalent

Urnenmodell	Schubladenmodell
mit/ohne Zurücklegen	mit/ohne Mehrfachbesetzung
mit/ohne Berücksichtigung der Reihenfolge	unterscheidbare/ununterscheidbare Objekte

62.3 Produktregel der Kombinatorik

Bei einem k -stufigen Experiment habe der Ausgang einer Stufe keinen Einfluss auf die Anzahl der möglichen Ausgänge bei späteren Stufen. Haben die einzelnen Stufen n_1, \dots, n_k Ausgänge, so hat das Gesamtexperiment $n_1 \cdot \dots \cdot n_k$ Ausgänge.
 Die Produktregel ist wichtig bei der Beschreibung der vier kombinatorischen Grundsituationen.

62.4 Die vier kombinatorischen Grundsituationen

a) Geordnete Stichprobe mit Wiederholung

Bei k Ziehungen mit Zurücklegen aus einer Urne mit n Objekten gibt es bei Berücksichtigung der Reihenfolge n^k Möglichkeiten.

Beispiel (aus einem älteren Stochastikbuch):

Herr Meier will seinen ungezogenen Sohn mit 10 Ohrfeigen bestrafen. Auf wie viele Arten kann er das tun, wenn er bei jedem Schlag eine von 2 Möglichkeiten hat (links oder rechts)?
 Es gibt $2^{10} = 1024$ Möglichkeiten.

b) Geordnete Stichprobe ohne Wiederholung

Urnenmodell: k Ziehungen aus n Objekten ohne Zurücklegen, aber mit Berücksichtigung der Reihenfolge.

Möglichkeiten: $\underbrace{n}_{1. \text{ Ziehung}} \cdot (n-1) \cdot (n-2) \cdot \dots \cdot \underbrace{(n-k+1)}_{k\text{-te Ziehung}} = \frac{n!}{(n-k)!}$

Spezialfall: $k = n \Rightarrow n!$ Möglichkeiten

Für große n approximiert man $n!$ mit der *Stirling-Formel*

$$n! \approx \sqrt{2\pi n} n^n e^{-n}$$

Beispiel: Herr Meier will seine 5 Kinder in einer Reihe anordnen für eine Gruppenaufnahme.
 Es gibt $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ Möglichkeiten.

c) Ungeordnete Stichprobe ohne Wiederholung

Wie in (b), jedoch müssen die $k!$ Permutationen der k gezogenen Elemente miteinander identifiziert werden.

Daher gibt es nur $\frac{n!}{(n-k)! k!} = \binom{n}{k}$ Möglichkeiten.

Beispiel: Beim Lottospielen gibt es

$$\binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = 13.983.816 \text{ Möglichkeiten.}$$

d) Ungeordnete Stichprobe mit Wiederholung

Hier ist das Schubladenmodell sinnvoll: Es sollen k nicht unterscheidbare Objekte in n Schubladen verstaut werden, wobei Mehrfachbesetzung möglich ist.



$n - 1$ Trennungstriche

Gesamtzustand wird beschrieben durch die Reihenfolge von k Objekten und $n - 1$ Trennungsstrichen.

$$\Rightarrow \frac{(k+n-1)!}{k!(n-1)!} = \binom{k+n-1}{k} = \binom{k+n-1}{n-1} \text{ Möglichkeiten}$$

Beispiel: Auf wie viele Arten kann man 60 Parlamentssitze auf 3 Parteien verteilen:

$$k = 60, n = 3 \Rightarrow \binom{62}{60} = \binom{62}{2} = \frac{62 \cdot 61}{2 \cdot 1} = 1891 \text{ Möglichkeiten}$$

63 Erzeugende Funktionen

63.1 Motivation

Erzeugende Funktionen wirken auf den ersten Blick etwas abstrakt, aber sie sind ein wichtiges Werkzeug, um kombinatorische Probleme systematischer und eleganter zu lösen. Sie sind zudem in verschiedenen anderen Gebieten der Stochastik nützlich.

63.2 Permutationen und Kombinationen

In Kapitel 62 haben wir geordnete und ungeordnete k -elementige Stichproben einer n -elementigen Menge betrachtet. Im geordneten Fall nennt man eine solche Stichprobe auch k -Permutation, im ungeordneten Fall eine k -Kombination.

63.3 Beispiel einer erzeugenden Funktion

Nach dem Binomialsatz 19.7 gilt:

$$(x + 1)^n = \sum_{k=0}^n \binom{n}{k} x^k \cdot 1^{n-k} = \sum_{k=0}^n \binom{n}{k} x^k$$

Der Koeffizient $\binom{n}{k}$ von x^k beschreibt die Zahl der k -Kombinationen einer n -elementigen Menge ohne Wiederholungen (vgl. 62.4 (c)).

In den Koeffizienten der Funktion

$$f(x) = (x + 1)^n = \binom{n}{0} x^0 + \binom{n}{1} x^1 + \dots + \binom{n}{n} x^n$$

stecken somit alle Informationen über dieses kombinatorische Problem.

Dies motiviert folgende Definition:

63.4 Definition

Eine Funktion f mit

$$f(x) = \sum_{k=0}^n a_k x^k$$

heißt *erzeugende Funktion* für die Koeffizienten a_k ($k = 0, \dots, n$). Lassen sich die Zahlen a_k mit einem kombinatorischen Problem identifizieren, so nennen wir f die erzeugende Funktion für dieses Problem.

63.5 Beispiel

$f(x) = (1 + x)^n$ ist die erzeugende Funktion der Kombinationen ohne Wiederholung einer n -elementigen Menge.

Schubladeninterpretation

Jeder der n Faktoren $(1 + x)$ wird als Schublade aufgefasst, in der sich 0 oder 1 Element befinden.

Wählt man beim Ausmultiplizieren von $(1+x)$ den Faktor 1, bleibt die Schublade leer. Wählt man x , wird sie mit einem Element besetzt. Beispielsweise beschreibt

$$f(x) = (1+x)^3 = 1 + 3x + 3x^2 + x^3$$

alle Kombinationen, 0, 1, 2 oder 3 Elemente auf 3 Schubladen zu verteilen. Dabei bedeuten die einzelnen Summanden:

$$\begin{array}{rcl} 1 & = & 1 \cdot 1 \cdot 1 \quad 1 \text{ Möglichkeit bei 0 Elementen} \\ 3x & = & x \cdot 1 \cdot 1 + 1 \cdot x \cdot 1 + 1 \cdot 1 \cdot x \quad 3 \text{ Möglichkeiten bei 1 Element} \\ 3x^2 & = & x \cdot x \cdot 1 + x \cdot 1 \cdot x + 1 \cdot x \cdot x \quad 3 \text{ Möglichkeiten bei 2 Elementen} \\ x^3 & = & x \cdot x \cdot x \quad 1 \text{ Möglichkeit bei 3 Elementen} \end{array}$$

Wie verallgemeinert man dieses Prinzip, wenn die Schubladen mit mehreren Objekten besetzt werden sollen?

63.6 Beispiel

Lassen wir pro Schublade bis zu zwei Objekte zu (die sich wiederholen dürfen), lautet der Faktor $(1+x+x^2)$ statt $(1+x)$.

Z.B. können wir die Anzahl der Kombinationen mit Wiederholung aus einer 2-elementigen Menge bestimmen, wenn jedes Element 0-, 1- oder 2-mal ausgewählt werden kann.

Erzeugende Funktion:

$$\begin{aligned} f(x) &= (1+x+x^2)^2 \\ &= (1+x+x^2)(1+x+x^2) \\ &= 1 \cdot 1 + 1 \cdot x + 1 \cdot x^2 + x \cdot 1 + x \cdot x + x \cdot x^2 + x^2 \cdot 1 + x^2 \cdot x + x^2 \cdot x^2 \\ &= 1 + 2x + 3x^2 + 2x^3 + x^4 \end{aligned}$$

Es gibt 1 Möglichkeit, 0 Objekte zu wählen.

Es gibt 2 Möglichkeiten, 1 Objekt zu wählen.

Es gibt 3 Möglichkeiten, 2 Objekte zu wählen.

Es gibt 2 Möglichkeiten, 3 Objekte zu wählen.

Es gibt 1 Möglichkeit, 4 Objekte zu wählen.

Man kann sogar für die unterschiedlichen Elemente verschiedene Beschränkungen einführen:

63.7 Beispiel

Bestimme die Kombinationen einer 4-elementige Menge $\{x_1, \dots, x_4\}$ mit den folgenden Beschränkungen:

Element	Beschränkung	Polynom
x_1	0-, 1- oder 3-mal	$1 + x + x^3$
x_2	1- oder 2-mal	$x + x^2$
x_3	1-mal	x
x_4	0- oder 4-mal	$1 + x^4$

erzeugende Funktion:

$$\begin{aligned} f(x) &= (1+x+x^3) \cdot (x+x^2) \cdot x \cdot (1+x^4) \\ &= \dots = x^2 + 2x^3 + x^4 + x^5 + 2x^6 + 2x^7 + x^8 + x^9 + x^{10} \end{aligned}$$

Es gibt also z.B. zwei 6-Kombinationen.

Diese Beispiele motivieren:

63.8 Satz (Kombinationen mit vorgegebenen Wiederholungen)

Die erzeugende Funktion der Kombinationen mit Wiederholung aus einer n -elementigen Menge $\{x_1, \dots, x_n\}$, in der x_i in den Anzahlen $v_1^{(i)}, \dots, v_{k_i}^{(i)}$ auftreten darf, lautet

$$f(x) = \prod_{i=1}^n (x^{v_1^{(i)}} + x^{v_2^{(i)}} + \dots + x^{v_{k_i}^{(i)}})$$

Bemerkung: Lässt man beliebige Wiederholungen zu, erhält man mit

$$1 + x + x^2 + x^3 + \dots = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \quad (\text{für } |x| < 1)$$

den folgenden Satz:

63.9 Satz (Kombinationen mit beliebigen Wiederholungen)

Die erzeugende Funktion der Kombination mit beliebigen Wiederholungen von n Elementen lautet:

$$f(x) = \frac{1}{(1-x)^n} = (1 + x + x^2 + \dots)^n$$

63.10 Bemerkungen

- Die gesuchten Koeffizienten vor x^k , $k \in \mathbb{N}_0$ ergeben sich durch eine formale Potenzreihenentwicklung. Dabei kann man Konvergenzbetrachtungen ignorieren, da man z.B. $|x| < 1$ annehmen darf.
- Muss jedes Element mindestens p -mal auftreten, ergibt sich wegen

$$x^p + x^{p+1} + x^{p+2} + \dots = x^p \sum_{k=0}^{\infty} x^k = \frac{x^p}{1-x}$$

die erzeugende Funktion

$$f(x) = \frac{x^{np}}{(1-x)^n}$$

Bisher haben wir nur Kombinationen betrachtet.

Sind erzeugende Funktionen auch bei Permutationen nützlich?

Hierzu müssen wir den Begriff der erzeugenden Funktion modifizieren.

63.11 Definition

Eine Funktion f mit $f(x) = \sum_{k=0}^n a_k \frac{x^k}{k!}$ heißt *exponentiell erzeugende Funktion* für die Koeffizienten a_k .

Bemerkung: Der Name wird motiviert durch die Exponentialreihe

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

63.12 Bedeutung für Permutationen

Nach 62.4 (b) gibt es bei einer n-elementigen Menge

$$n \cdot (n-1) \cdot \dots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

k-Permutationen ohne Wiederholung.

Wegen

$$(x+1)^n = \sum_{k=0}^n \binom{n}{k} x^k = \sum_{k=0}^n \frac{n!}{(n-k)!k!} x^k = \sum_{k=0}^n \frac{n!}{(n-k)!} \cdot \frac{x^k}{k!}$$

sind dies die Koeffizienten der exponentiell erzeugenden Funktion

$$f(x) = (1+x)^n.$$

Die exponentiell erzeugende Funktion spielt also bei Permutationen die gleiche Rolle wie die erzeugende Funktion bei Kombinationen.

Das Analogon zu Satz 63.8 lautet:

63.13 Satz (Permutationen mit vorgegebenen Wiederholungen)

Die exponentiell erzeugende Funktion der Permutationen mit Wiederholung aus einer n-elementigen Menge $\{x_1, \dots, x_n\}$, in der x_i in den Anzahlen $v_1^{(i)}, \dots, v_{k_i}^{(i)}$ auftreten darf, lautet

$$\prod_{i=1}^n \left(\frac{x^{v_1^{(i)}}}{v_1^{(i)}!} + \frac{x^{v_2^{(i)}}}{v_2^{(i)}!} + \dots + \frac{x^{v_{k_i}^{(i)}}}{v_{k_i}^{(i)}!} \right)$$

Lässt man beliebige Wiederholung zu, folgt mit

$$1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$$

das Analogon zu Satz 63.9:

63.14 Satz (Permutationen mit beliebigen Wiederholungen)

Die exponentiell erzeugende Funktion der Permutationen mit beliebigen Wiederholungen von n Elementen lautet:

$$f(x) = e^{nx} = (e^x)^n = \left(1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots\right)^n.$$

64 Bedingte Wahrscheinlichkeiten

64.1 Motivation

In Kapitel 61 haben wir den Begriff der Wahrscheinlichkeit kennengelernt. Oft hat man Zusatzinformationen, mit denen sich präzisere Wahrscheinlichkeitsaussagen machen lassen. Dies führt zu den sogenannten bedingten Wahrscheinlichkeiten.

64.2 Beispiel

Aufteilung der 1500 Angehörigen eines Betriebs nach Geschlecht und Rauchergewohnheiten:

	Frauen B	Männer \bar{B}
Raucher A	600	200
Nichtraucher \bar{A}	300	400

Ω : Menge der Betriebsangehörigen

A : Menge der Raucher ($\bar{A} = \Omega \setminus A$: Nichtraucher)

B : Menge der Frauen ($\bar{B} = \Omega \setminus B$: Männer)

Wir lösen eine Person zufällig aus. Dabei treffen folgende Wahrscheinlichkeiten zu:

($|A|$: Mächtigkeit von A)

$$P(A) = \frac{|A|}{|\Omega|} = \frac{800}{1500} = \frac{8}{15} \text{ Raucheranteil}$$

$$P(B) = \frac{|B|}{|\Omega|} = \frac{900}{1500} = \frac{3}{5} \text{ Frauenanteil}$$

$$P(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{600}{1500} = \frac{2}{5} \text{ Anteil der rauchenden Frauen}$$

Wie groß ist die Wahrscheinlichkeit $P(A|B)$, dass eine Person raucht, wenn es sich um eine Frau handelt?

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{\frac{|A \cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3} \quad (*)$$

Man nennt $P(A|B)$ die *bedingte Wahrscheinlichkeit* von A unter der Bedingung (Hypothese) B . Es gilt stets (vgl. (*)):

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

Zwei Ereignisse A, B heißen *unabhängig*, wenn gilt:

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

64.3 Verallgemeinerung

Seien $A_1, \dots, A_n \subset \Omega$ und $P(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$. Dann gilt:

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \dots \cdot P(A_n|A_1 \cap \dots \cap A_{n-1})$$

Beweis: Die rechte Seite lässt sich schreiben als

$$P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot \dots \cdot \frac{P(A_1 \cap \dots \cap A_n)}{P(A_1 \cap \dots \cap A_{n-1})} = P(A_1 \cap \dots \cap A_n). \quad \square$$

64.4 Beispiel

Wie groß ist die Wahrscheinlichkeit, mit 6 Würfeln 6 verschiedene Zahlen zu würfeln?

A_1 : irgendein Ergebnis für Würfel 1.

A_2 : ein vom 1. Ergebnis verschiedenes Ergebnis für Würfel 2.

\vdots

A_6 : ein von A_1, \dots, A_5 verschiedenes Ergebnis für Würfel 6.

Mit 64.3 gilt:

$$\begin{aligned} P(A_1 \cap \dots \cap A_6) &= P(A_1) \cdot P(A_2|A_1) \cdot \dots \cdot P(A_6|A_1 \cap \dots \cap A_5) \\ &= 1 \cdot \frac{5}{6} \cdot \frac{4}{6} \cdot \dots \cdot \frac{1}{6} = \frac{6!}{6^6} \approx 0,015. \end{aligned}$$

64.5 Beispiel

Ein Krebstest ist mit 96% iger Wahrscheinlichkeit positiv, falls der Patient Krebs hat, mit 94% iger Wahrscheinlichkeit negativ, falls er keinen Krebs hat. Bei einem Patienten, in dessen Altersgruppe 0,5% aller Personen Krebs haben, verläuft der Test positiv. Wie groß ist die Wahrscheinlichkeit, dass er tatsächlich Krebs hat?

Ereignis K: Patient hat Krebs

Ereignis T: Test ist positiv

$$P(K|T) = \frac{P(K \cap T)}{P(T)} = \frac{0,005 \cdot 0,96}{0,005 \cdot 0,96 + 0,995 \cdot 0,06} \approx 0,074$$

Der Patient kann also noch relativ beruhigt sein.

Fazit: Um eine seltene Krankheit zuverlässig zu erkennen, darf ein Test nur sehr wenige „false positives“ haben.

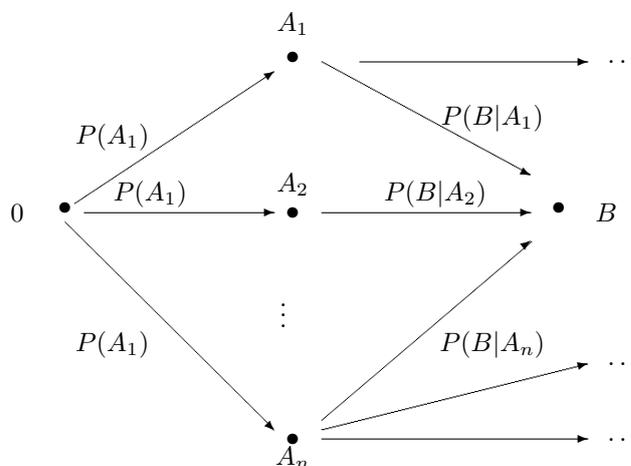
64.6 Satz (Satz von der totalen Wahrscheinlichkeit)

Sei $\Omega = \bigcup_{i=1}^n A_i$ eine Partition (disjunkte Vereinigung) von Ω in mögliche Ereignisse $A_i, i = 1, \dots, n$.

Ferner sei $P(A_i) > 0$ für alle i . Dann gilt für jedes Ereignis B :

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

Veranschaulichung:



Ein Fahrer startet bei 0 und fährt mit Wahrscheinlichkeit $P(A_1), \dots, P(A_n)$ zu A_1, \dots, A_n . Die Wahrscheinlichkeit, von dort nach B zu fahren, beträgt $P(B|A_1), \dots, P(B|A_n)$. Die Gesamtwahrscheinlichkeit, dass der Fahrer nach B gelangt, ist die Summe der Wahrscheinlichkeiten aller Pfade von 0 nach B :

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

Bemerkung: Im Beispiel 64.5 haben wir im Nenner bereits den Satz von der totalen Wahrscheinlichkeit verwendet.

64.7 Beispiel

Um einen binären Nachrichtenkanal robuster gegenüber Störungen zu machen, sendet man die Bitfolge 0000000 statt 0 und 1111111 statt 1. Störungen treten in 20% aller Fälle auf, und die Wahrscheinlichkeit für 0 sei 0,1.

Es wird 0100110 empfangen. Wie groß ist die Wahrscheinlichkeit, dass 0000000 gesendet wurde?

$$P(0000000|0100110) = \frac{0,1 \cdot 0,2^3 \cdot 0,8^4}{0,1 \cdot 0,2^3 \cdot 0,8^4 + 0,9 \cdot 0,2^4 \cdot 0,8^3} \approx 0,308$$

Man wird den Block also als 1 lesen, obwohl die Mehrzahl der Bits Nullen sind!

Eng verwandt mit dem Satz von der totalen Wahrscheinlichkeit ist

64.8 Satz (Formel von Bayes)

Sei $P(B) > 0$ und seien die Voraussetzungen von Satz 64.6 erfüllt. Dann gilt:

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

Beweis: Folgt aus $P(A_k|B) = \frac{P(A_k \cap B)}{P(B)}$, indem man

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i) \quad (64.6)$$

$$P(A_k \cap B) = P(A_k)P(B|A_k) \quad (64.2)$$

einsetzt.

Bemerkung: Im Satz 64.8 wird $P(A_k|B)$ aus $P(B|A_i), i = 1, \dots, n$ berechnet, d.h. Ursache und Wirkung kehren sich um. Eine typische Anwendung besteht darin, dass man weiß, welche Ursachen welche Auswirkungen haben. Hat man eine Auswirkung gemessen, fragt man nach der wahrscheinlichsten Ursache (*inverses Problem*).

64.9 Anwendungsbeispiele

- a) Ein Arzt beobachtet bei einem Patienten ein Symptom B . Es kann von n verschiedenen Krankheiten A_1, \dots, A_n herrühren. Um die wahrscheinlichste Ursache zu finden, muss man also $P(A_k|B)$ abschätzen.
- b) Aus einem verrauschten Bild will man das wahrscheinlichste unverrauschte Bild rekonstruieren (vgl. Beispiel 64.7).
- c) In der Computertomographie schickt man Röntgenstrahlung in verschiedenen Richtungen durch den Patienten und misst die durchgedrungene Intensität. Aus diesen Auswirkungen versucht man, Rückschlüsse auf die Ursache (Gewebe, Knochen, Tumor, ...) zu ziehen.

65 Zufallsvariable, Erwartungswert, Varianz

65.1 Motivation

Oft möchte man das Resultat eines Zufallsexperiments einer reellen Zahl zuordnen. Der Gewinn bei einem Glücksspiel ist ein Beispiel hierfür. In diesem Fall interessiert man sich auch für den zu erwartenden Gewinn und für ein Maß für die statistischen Schwankungen. Dies führt uns auf die Begriffe Zufallsvariable, Erwartungswert und Varianz. In der Informatik werden sie u.a. bei der Zuverlässigkeitsanalyse von Systemen benötigt.

65.2 Definition

Sei Ω ein Stichprobenraum. Eine Funktion X , die jedem Ergebnis $\omega \in \Omega$ eine reelle Zahl $X(\omega)$ zuordnet, heißt *Zufallsvariable*.

Bemerkung: Eine Zufallsvariable ist also weder zufällig noch eine Variable, sondern eine Funktion. Man kann sie stets als Gewinn bei einem Glücksspiel interpretieren.

65.3 Beispiel

Eine faire Münze mit Seiten 0 und 1 werde 3 Mal geworfen. Die Anzahl der Einsen sei der Gewinn. Man kann $\Omega = \{000, 001, \dots, 111\}$ als Stichprobenraum und den Gewinn als Zufallsvariable $X(\omega)$ auffassen:

Ergebnis ω	000	001	010	011	100	101	110	111
Gewinn $X(\omega)$	0	1	1	2	1	2	2	3
Wahrscheinlichkeit $P(\omega)$	$\frac{1}{8}$							

Verschiedene Ergebnisse ω_i, ω_j können zum selben Gewinn $X(\omega_i) = X(\omega_j)$ führen.

65.4 Definition

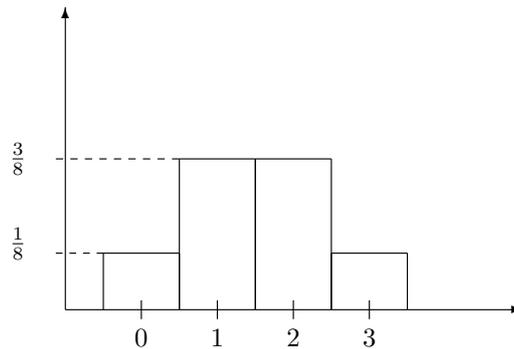
Die *Verteilung* P_X einer Zufallsvariablen X ordnet jedem Wert $x \in X$ eine Wahrscheinlichkeit $P_X(x)$ zu.

65.5 Beispiel

In Beispiel 65.3 können wir dem Gewinn folgende Wahrscheinlichkeiten zuordnen:

Gewinn x	0	1	2	3
Wahrscheinlichkeit $P_X(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Oft veranschaulicht man die Verteilung einer Zufallsvariablen durch ein *Histogramm*:



Da wir meistens diskrete Zufallsvariablen betrachten, sind die Verteilungen ebenfalls diskret.

Interessiert man sich für den Durchschnittsgewinn je Versuchswiederholung, gelangt man zum Begriff des Erwartungswerts:

65.6 Definition

Unter dem *Erwartungswert* $E(X)$ einer (diskreten) Zufallsvariablen X versteht man das gewichtete Mittel der Funktion X über Ω , wobei jeder Wert mit seiner Wahrscheinlichkeit gewichtet wird:

$$E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega)$$

65.7 Beispiel

Für die Zufallsvariable X aus 65.3 erhält man den Erwartungswert

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = \frac{12}{8} = \frac{3}{2}$$

Bequemer hätte man den Erwartungswert mit Hilfe der Verteilung P_X berechnet (vgl. Tabelle aus Beispiel 65.5):

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

Es gilt also auch:

$$E(X) = \sum_{x \in X(\Omega)} x P_X(x)$$

Bemerkung: Bei kontinuierlichen Zufallsvariablen verwendet man Integrale statt Summen.

Mit dem Erwartungswert kann man gut arbeiten, denn es gilt:

65.8 Satz (Linearität des Erwartungswerts)

Seien X, Y Zufallsvariablen und $\alpha, \beta \in \mathbb{R}$. Dann gilt:

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$$

Beweis:

$$\begin{aligned}
 E(\alpha X + \beta Y) &= \sum_{\omega \in \Omega} (\alpha X(\omega) + \beta Y(\omega)) P(\omega) \\
 &= \alpha \sum_{\omega \in \Omega} X(\omega) P(\omega) + \beta \sum_{\omega \in \Omega} Y(\omega) P(\omega) \\
 &= \alpha E(X) + \beta E(Y) \quad \square
 \end{aligned}$$

65.9 Unabhängigkeit zweier Zufallsvariabler

Mit 65.8 wissen wir: $E(X + Y) = E(X) + E(Y)$.

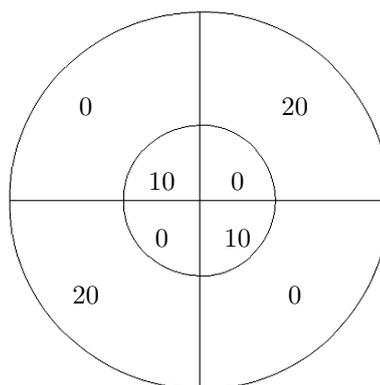
Gilt auch $E(X \cdot Y) = E(X) \cdot E(Y)$?

Man kann zeigen, dass dies zutrifft, wenn X und Y *unabhängig* sind, d.h. (vgl. 64.2):

$$P((X = a) \cap (Y = b)) = P(X = a) \cdot P(Y = b) \quad \forall a, b$$

65.10 Beispiele

- a) Glücksrad mit 4 Ergebnissen, auf denen die Zufallsvariable X (äußerer Ring) und Y (innerer Ring) definiert sind.



$$\left. \begin{aligned}
 E(X) &= \frac{1}{4} \cdot 20 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 20 + \frac{1}{4} \cdot 0 = 10 \\
 E(Y) &= \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 10 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 10 = 5
 \end{aligned} \right\} E(X)E(Y) = 50$$

$$E(XY) = \frac{1}{4} \cdot 20 \cdot 0 + \frac{1}{4} \cdot 0 \cdot 10 + \frac{1}{4} \cdot 20 \cdot 0 + \frac{1}{4} \cdot 0 \cdot 10 = 0 \neq E(X)E(Y)$$

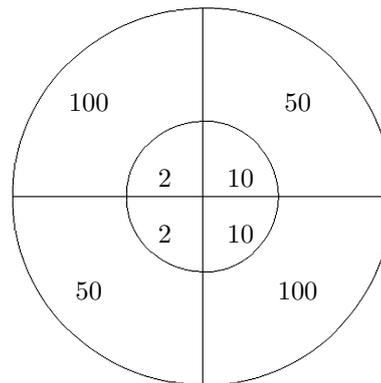
X und Y sind nicht unabhängig:

Das Ereignis $Y = 0$ hat die Wahrscheinlichkeit $\frac{1}{2}$. Weiß man jedoch, dass $X = 20$ eingetreten ist, hat $Y = 0$ die Wahrscheinlichkeit 1.

b)

$$\left. \begin{aligned}
 E(X) &= \frac{1}{4} \cdot 50 + \frac{1}{4} \cdot 100 + \frac{1}{4} \cdot 50 + \frac{1}{4} \cdot 100 = 75 \\
 E(Y) &= \frac{1}{4} \cdot 10 + \frac{1}{4} \cdot 10 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = 6
 \end{aligned} \right\} E(X)E(Y) = 450$$

$$E(XY) = \frac{1}{4} \cdot 50 \cdot 10 + \frac{1}{4} \cdot 100 \cdot 10 + \frac{1}{4} \cdot 50 \cdot 2 + \frac{1}{4} \cdot 100 \cdot 2 = 450 = E(X)E(Y)$$



X und Y sind unabhängig:

$Y = 2$ und $Y = 10$ sind gleich wahrscheinlich. Weiß man z.B., dass $X = 50$ eingetreten ist, so sind $Y = 2$ und $Y = 10$ noch stets gleich wahrscheinlich.

Oft ist man nicht nur am Erwartungswert interessiert. Man möchte auch wissen, wie stark die Verteilung um den Erwartungswert streut. Hierzu dienen die Begriffe Varianz und Standardabweichung:

65.11 Definition

Sei X eine Zufallsvariable mit Erwartungswert $E(X) =: \mu$. Dann versteht man unter der *Varianz* $V(X) = \sigma^2$ den Erwartungswert von $(X - \mu)^2$:

$$\sigma^2 = V(X) = E((X - E(X))^2)$$

$\sigma = \sqrt{V(X)}$ nennt man die *Standardabweichung (Streuung)* von X .

65.12 Berechnung der Varianz

Wegen der Linearität des Erwartungswerts und wegen $E(\text{const.}) = \text{const.}$ gilt:

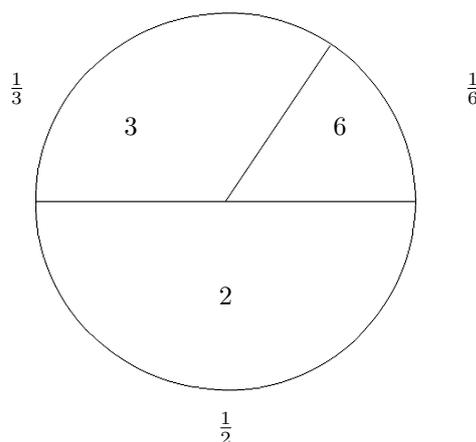
$$\begin{aligned} E((X - \mu)^2) &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu \underbrace{E(X)}_{\mu} + \mu^2 = E(X^2) - \mu^2 \end{aligned}$$

Wir erhalten somit eine wichtige Formel zur Berechnung der Varianz:

$$\sigma^2 = E(X^2) - \mu^2 \quad (\text{Verschiebungssatz})$$

65.13 Beispiel

Sei X der Gewinn auf dem Glücksrad



$$\begin{aligned} \mu &= \frac{1}{2} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{1}{6} \cdot 6 = 3 && \text{mittlerer Gewinn} \\ E(X^2) &= \frac{1}{2} \cdot 4 + \frac{1}{3} \cdot 9 + \frac{1}{6} \cdot 36 = 11 \\ \sigma^2 &= E(X^2) - \mu^2 = 11 - 3^2 = 2 && \text{Varianz} \\ \sigma &= \sqrt{2} && \text{Standardabweichung} \end{aligned}$$

65.14 Satz (Eigenschaften der Varianz)

Sei X eine Zufallsvariable und seien $\alpha, \beta \in \mathbb{R}$. Dann gilt:

- i) $V(\alpha X) = \alpha^2 V(X)$
- ii) $V(X + \beta) = V(X)$

Beweis:

- i) $V(\alpha X) \stackrel{65.8}{=} E((\alpha X - \alpha\mu)^2) = E(\alpha^2(X - \mu)^2) \stackrel{65.8}{=} \alpha^2 V(X).$
- ii) $V(X + \beta) = E((X + \beta - \mu - \beta)^2) = E((X - \mu)^2) = V(X) \quad \square$

65.15 Standardisierte Zufallsvariablen

Ist X eine Zufallsvariable mit Erwartungswert μ und Varianz σ^2 , nennt man

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

die *Standardisierte* von X . Es gilt:

$$\begin{aligned} E(X^*) &= \frac{1}{\sigma} \underbrace{E(X)}_{\mu} - \frac{\mu}{\sigma} = 0 \\ V(X^*) &\stackrel{65.14}{=} \frac{1}{\sigma^2} \underbrace{V(X)}_{\sigma^2} = 1. \end{aligned}$$

Eine solche Standardisierung ist nützlich, wenn man die Verteilung einer Zufallsvariablen mit einer tabellierten Verteilung vergleichen möchte, die meist nur in standardisierter Form vorliegt.

Wir wollen nun weitere wichtige Eigenschaften des Erwartungswerts studieren.
 Aus dem Verschiebungssatz 65.12 folgt wegen $\sigma^2 \geq 0$:

$$E(X^2) - (E(X))^2 \geq 0 \text{ und somit } E(X^2) \geq (E(X))^2$$

Dies ist der Spezialfall eines allgemeineren Resultats:

65.16 Satz (Ungleichung von Jensen)

Sei $r : \mathbb{R} \rightarrow \mathbb{R}$ eine *konvexe* (!) Funktion und X eine Zufallsvariable. Dann gilt:

$$E(r(X)) \geq r(E(X))$$

Beweis: Sei X zunächst eine diskrete Zufallsvariable.

Wegen $P(\omega) \geq 0 \quad \forall \omega \in \Omega$ und $\sum_{\omega \in \Omega} P(\omega) = 1$ können wir $E(X) = \sum_{\omega \in \Omega} P(\omega)X(\omega)$ als *Konvexkombination* der $X(\omega), \omega \in \Omega$ mit Gewichten $P(\omega)$ auffassen.

Aus der Konvexität von r folgt mit 25.5:

$$E(r(X)) = \sum_{\omega \in \Omega} P(\omega) r(X(\omega)) \geq r\left(\sum_{\omega \in \Omega} P(\omega)X(\omega)\right) = r(E(X))$$

Ist X eine kontinuierliche Zufallsvariable, ersetzt man Summen durch Integrale. □

65.17 Beispiele

a) Die Funktion $r(t) = t^2$ ist konvex, da $r''(t) = 2 > 0$. Daher gilt:

$$E(X^2) \geq (E(X))^2$$

b) Sei $\Theta > 0$. Dann ist $r(t) = e^{\Theta t}$ konvex und es gilt:

$$E(e^{\Theta X}) \geq e^{\Theta E(X)}$$

Die Funktion $E(e^{\Theta X})$ wird später eine wichtige Rolle spielen.

65.18 Gleichung von Wald

Seien X_1, X_2, \dots unabhängige Zufallsvariablen. Manchmal interessiert man sich für die erste Zeit (*Stoppzeit*) $N = n$, in der die Summe $S_N = \sum_{i=1}^n X_i$ einen vorgegebenen Wert y übersteigt, d.h.

$$\sum_{i=1}^{n-1} X_i \leq y \text{ und } \sum_{i=1}^n X_i > y.$$

Dann ist N selbst eine Zufallsvariable. Für ihren Erwartungswert kann man zeigen:

Satz (Gleichung von Wald)

Seien X_1, X_2, \dots unabhängige, *gleich verteilte* Zufallsvariablen (d.h. $P(\omega)$ ist gleich $\forall \omega \in \Omega$) mit endlichem Erwartungswert und sei N eine Stoppzeit für X_1, X_2, \dots . Ferner sei $S_N = \sum_{i=1}^N X_i$. Dann gilt:

$$E(S_N) = E(X) \cdot E(N)$$

65.19 Beispiel

Sei $X_i = 1$, falls beim i -ten Münzwurf Kopf eintritt, 0 sonst. Wie ist der Erwartungswert $E(N)$ für die Stoppzeit

$$N := \min\{n \mid X_1 + \dots + X_n \geq 10\}?$$

Es gilt: $E(X) = \frac{1}{2}$ und $E(S_N) = 10 \Rightarrow E(N) = 20$.

Eng verwandt mit dem Begriff der Varianz ist die Kovarianz. Sie ist ein Maß für die Unabhängigkeit zweier Zufallsvariablen.

65.20 Definition

Seien X, Y Zufallsvariablen mit Erwartungswert μ_X, μ_Y und Varianz σ_X^2, σ_Y^2 . Dann heißt

$$\text{Cov}(X, Y) := \sigma_{XY}^2 := E((X - \mu_X)(Y - \mu_Y))$$

die *Kovarianz* von X und Y , und

$$\varrho_{XY} := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

der *Korrelationskoeffizient* von X und Y .

Ist $\text{Cov}(X, Y) = 0$, so heißen X und Y *unkorreliert*.

65.21 Bemerkungen

a) $V(X) = \text{Cov}(X, X)$

b) Seien $X - \mu_X, Y - \mu_Y$ Funktionen auf einem endlichen Stichprobenraum $\Omega = \{\omega_1, \dots, \omega_n\}$. Dann kann man sie als Vektoren im \mathbb{R}^n mit der i -ten Komponente $X(\omega_i) - \mu_X$ bzw. $Y(\omega_i) - \mu_Y$ ansehen. Legt man das gewichtete Skalarprodukt

$$\langle u, v \rangle := \sum_{i=1}^n p_i u_i v_i \text{ mit } p_i := P(\omega_i)$$

zu Grunde, gelten folgende Entsprechungen (vgl. Kap. 41, Kap. 43):

- Standardabweichung $\hat{=}$ induzierte (euklidische) Norm
- Varianz $\hat{=}$ quadrierte Norm
- Kovarianz $\hat{=}$ Skalarprodukt
- Korrelationskoeffizient $\hat{=}$ Arcuscossinus des Winkels zwischen den Vektoren ($\Rightarrow -1 \leq \varrho_{XY} \leq 1$)
- Unkorreliertheit $\hat{=}$ Orthogonalität

Man kann folgende Regeln zeigen:

65.22 Satz (Rechenregeln für die Korrelation)

Seien X, Y Zufallsvariablen mit Erwartungswert μ_X, μ_Y . Dann gilt:

- a) $\text{Cov}(X, Y) = E(X \cdot Y) - \mu_X \mu_Y$ (vgl. 65.12)
- b) $\text{Cov}(\alpha X + \beta, \gamma Y + \delta) = \alpha \gamma \text{Cov}(X, Y)$ (vgl. 65.14)
- c) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

d) Für m Zufallsvariablen X_1, \dots, X_m gilt:

$$V(X_1 + \dots + X_m) = \sum_{i=1}^m V(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

e) Sind X, Y unabhängig, so sind sie auch unkorreliert.

f) Für *unabhängige* X_1, \dots, X_m gilt:

$$V\left(\sum_{i=1}^m X_i\right) = \sum_{i=1}^m V(X_i)$$

65.23 Bemerkungen

a) Die Umkehrung von 65.22 (e) gilt nicht! Beispiel:

Ergebnis ω	1	2	3	4
Zufallsvariable X	1	-1	2	-2
Zufallsvariable Y	-1	1	2	-2
Wahrscheinlichkeit $P(\omega)$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{1}{10}$

Dann ist $E(X) = 0 = E(Y)$ und

$$\text{Cov}(X, Y) = -1 \cdot \frac{2}{5} - 1 \cdot \frac{2}{5} + 4 \cdot \frac{1}{10} + 4 \cdot \frac{1}{10} = 0$$

Aber: X und Y sind nicht unabhängig, denn $X(\omega)$ bestimmte ω und $Y(\omega)$ eindeutig.

b) In der Informatik tauchen Erwartungswerte und Varianzen z.B. bei der Zuverlässigkeitsanalyse von Systemen auf oder bei Abschätzung von Wartezeiten bei Internetanfragen. Kovarianzen sind z.B. wichtig im Bereich des maschinellen Lernens.

65.24 Zuverlässigkeitsanalyse von Systemen

Ein System bestehe aus n Komponenten. Der Zustand der k -ten Komponente wird durch die Zufallsvariable (*Indikator*)

$$I_k := \begin{cases} 1 & \text{(k-te Komponente funktioniert)} \\ 0 & \text{(k-te Komponente funktioniert nicht)} \end{cases}$$

beschrieben. Ihr Erwartungswert beschreibt die *Zuverlässigkeit* der Komponente k . Wir setzen: $p_k = E(I_k)$ und bezeichnen mit $q_k = 1 - p_k$ die Ausfallwahrscheinlichkeit. Für die Zuverlässigkeit p des Gesamtsystems gilt dann:

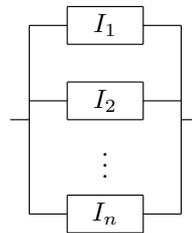
a) *Reihensysteme*



Ein Reihensystem arbeitet, wenn alle Komponenten arbeiten:

$$p = p_1 \cdot p_2 \cdot \dots \cdot p_n$$

b) *Parallelsysteme*

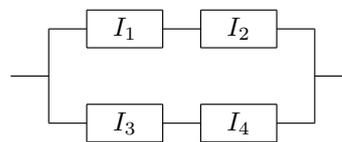


Ein Parallelsystem fällt aus, wenn alle Komponenten ausfallen:

$$q = q_1 \cdot q_2 \cdot \dots \cdot q_n$$

$$\Rightarrow p = 1 - q = 1 - (1 - p_1) \cdot \dots \cdot (1 - p_n)$$

c) *Gemischte Systeme*



Gemischte Systeme werden hierarchisch in Reihensysteme und Parallelsysteme zerlegt:

oberes Reihensystem: $p_1 \cdot p_2$

unteres Reihensystem: $p_3 \cdot p_4$

paralleles Gesamtsystem: $p = 1 - (1 - p_1 p_2)(1 - p_3 p_4)$

66 Abschätzungen für Abweichungen vom Erwartungswert

66.1 Motivation

Mit der Varianz bzw. Standardabweichung kennen wir bereits ein Maß für die Fluktuation einer Zufallsvariablen um ihren Erwartungswert.

Gibt es weitere nützliche Maßzahlen hierfür?

Ist es möglich, Abschätzungen zu finden, die wahrscheinlichkeitstheoretische Aussagen darüber liefern, wie häufig eine Zufallsvariable außerhalb eines Intervalls um den Erwartungswert liegt?

66.2 Definition

Sei X eine Zufallsvariable mit Erwartungswert $\mu = E(X)$.

Dann bezeichnen wir mit

$$E(X^k), k \in \mathbb{N}$$

das k -te *Moment* von X . Ferner definiert man das k -te *zentrale Moment* von X als

$$E((X - \mu)^k)$$

66.3 Bemerkungen

- a) Der Erwartungswert $\mu = E(X)$ ist das erste Moment.
- b) Die Varianz $\sigma^2 = E((X - \mu)^2)$ ist das zweite zentrale Moment.
- c) Mit dem 3. und 4. zentralen Moment lassen sich Aussagen über die Schiefe bzw. Flachheit der Verteilung einer Zufallsvariablen gewinnen.
- d) Höhere Momente sind anschaulich schwieriger zu interpretieren, liefern jedoch ebenfalls wichtige Aussagen.
- e) Momente haben eine große Bedeutung in der Mustererkennung bei der Analyse von Texturen und der Erkennung von Objekten unter Rotationen und Skalierungen.

Ähnlich wie wir die Menge aller k -Permutationen und k -Kombinationen einer n -elementigen Menge mit dem Begriff der erzeugenden Funktion kompakt beschreiben konnten, gibt es eine Funktion, die sämtliche Momente einer Zufallsvariablen beinhaltet.

66.4 Definition

Sei X eine Zufallsvariable.

Falls $M_X(\Theta) := E(e^{\Theta X})$ existiert, nennen wir $M_X(\Theta)$ die *Momente erzeugende Funktion* von X .

66.5 Satz (Eigenschaften Momenten erzeugender Funktion)

Die Momenten erzeugende Funktion $M_X(\Theta) = E(e^{\Theta X})$ einer Zufallsvariablen X hat folgende Eigenschaften:

- a) Die n -te Ableitung in $\Theta = 0$ liefert das n -te Moment:

$$M_X^{(n)}(0) = E(X^n)$$

- b) Skalierungsverhalten:

Sei $Y = aX + b$. Dann ist $M_Y(\Theta) = e^{b\Theta} M_X(a\Theta)$

- c) $M_{X+Y}(\Theta) = M_X(\Theta) \cdot M_Y(\Theta)$, falls X und Y unabhängige Zufallsvariablen sind.

Beweis von (a): Mit der Potenzreihenentwicklung der Exponentialfunktion und der Linearität der Erwartungswerts gilt:

$$M_X(\Theta) = E(e^{\Theta X}) = E\left(\sum_{k=0}^{\infty} \frac{\Theta^k X^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{\Theta^k}{k!} E(X^k)$$

Gliedweise Differentiation liefert:

$$M_X'(\Theta) = \sum_{k=1}^{\infty} \frac{\Theta^{k-1}}{(k-1)!} E(X^k) = \sum_{k=0}^{\infty} \frac{\Theta^k}{k!} E(X^{k+1})$$

⋮

$$M_X^{(n)}(\Theta) = \sum_{k=1}^{\infty} \frac{\Theta^k}{k!} E(X^{k+n})$$

und somit

$$M_X^{(n)}(0) = E(X^n)$$

□

Wir kommen nun zu den Abschätzungen für Fluktuationen jenseits eines vorgegebenen Abstands von dem Erwartungswert einer Zufallsvariablen. Grundlegend ist der folgende Satz:

66.6 Satz (Markow'sche Ungleichung)

Sei X eine Zufallsvariable und $h : \mathbb{R} \rightarrow \mathbb{R}$ eine nichtnegative und nichtfallende Funktion mit $h(t) > 0$ für ein t . Dann gilt:

$$P(X \geq t) \leq \frac{E(h(X))}{h(t)}$$

Beweis (für eine diskrete Zufallsvariable):
 Nach Definition 65.7 gilt:

$$\begin{aligned}
 E(h(X)) &= \sum_{z \in X(\Omega)} \underbrace{h(z)}_{>0} \underbrace{P_X(z)}_{\geq 0} \\
 &\geq \sum_{\substack{z \in X(\Omega) \\ z \geq t}} h(z) P_X(z) \quad (\text{Nichtnegativität}) \\
 &\geq h(t) \underbrace{\sum_{\substack{z \in X(\Omega) \\ z \geq t}} P_X(z)}_{P(X \geq t)} \quad (\text{Nichtfallend}) \\
 \Rightarrow P(X \geq t) &\leq \frac{E(h(X))}{h(t)}
 \end{aligned}$$

□

66.7 Folgerungen

a) Setzt man $h(x) = \begin{cases} x & (x > 0) \\ 0 & (\text{sonst}) \end{cases}$ folgt die *einfache Markow-Ungleichung*:

$$P(X \geq t) \leq \frac{E(X)}{t} \quad (t > 0)$$

b) Mit $Y := (X - E(X))^2$ und $h(x) = x$ für $x > 0$ kann man die *Tschebyschew-Ungleichung* für den Erwartungswert μ und die Standardabweichung σ von X beweisen:

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Alternativschreibweise:

$$P(|X - \mu| \geq c \cdot \sigma) \leq \frac{1}{c^2}$$

c) Mit $h(x) = e^{\Theta x}$ für $\Theta \geq 0$ ergibt sich:

$$P(X \geq t) \leq e^{-\Theta t} M_X(\Theta)$$

Da $\Theta \geq 0$ beliebig war, führt dies zu *Chernoff-Schranke*

$$P(X \geq t) \leq \inf_{\Theta \geq 0} e^{-\Theta t} M_X(\Theta)$$

Bemerkung: Für die einfache Markow-Ungleichung benötigt man das erste Moment (μ), für die Tschebyschew-Ungleichung die ersten beiden Momente μ, σ^2 , für die Chernoff-Ungleichung alle Momente (die Momenten erzeugende Funktion). Je mehr Momente man kennt, desto mehr weiß man über die Verteilung von X und desto schärfere Abschätzungen kann man erwarten.

66.8 Beispiel

Eine faire Münze werde n Mal geworfen. Tritt beim k -ten Wurf Kopf auf, setzt man $Y_k := 1$, sonst $Y_k := 0$. Wir interessieren uns für die „Kopfhäufigkeit“ nach n Würfen:

$$X_n := Y_1 + \dots + Y_n$$

Y_1, \dots, Y_n sind unabhängige Zufallsvariablen. Für X_n gilt:

$$\begin{aligned}\mu &= E(X_n) = \frac{n}{2} \\ \sigma^2 &= E((X_n - \mu)^2) \\ &= \sum_{k=1}^n \left[\left(0 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} \right] = \frac{n}{4}\end{aligned}$$

und wegen

$$M_{Y_k}(\Theta) = E(e^{\Theta Y_k}) = e^{\Theta \cdot 0} \cdot \frac{1}{2} + e^{\Theta \cdot 1} \cdot \frac{1}{2} = \frac{1 + e^{\Theta}}{2}$$

folgt mit 66.5 (c):

$$M_{X_n}(\Theta) = M_{Y_1 + \dots + Y_n}(\Theta) = \left(\frac{1 + e^{\Theta}}{2}\right)^n$$

Wie groß ist die Wahrscheinlichkeit, nach $n = 100$ Würfeln $X_n \geq 80$ zu erhalten? Wir vergleichen die 3 Ungleichungen aus 66.7:

a) Einfache Markow-Ungleichung:

Mit $\mu = 0.5 \cdot 100 = 50$ und $t = 80$ ergibt sich

$$P(X_{100} \geq 80) \leq \frac{\mu}{t} = \frac{50}{80} \approx 0.625$$

b) Tschebyschew-Ungleichung:

Mit $\mu = 50$, $t = 30$ und $\sigma^2 = \frac{100}{4} = 25$ ergibt sich

$$P(X_{100} \geq 80) \leq P(|X_{100} - 50| \geq 30) \leq \frac{\sigma^2}{t^2} = \frac{25}{30^2} \approx 0.028$$

Obwohl Abweichungen nach beiden Seiten berücksichtigt werden, ist die Abschätzung schärfer als bei der einfachen Markow-Ungleichung.

c) Chernoff-Schranke:

$$P(X_{100} \geq 80) \leq \inf_{\Theta \geq 0} \underbrace{e^{-80 \cdot \Theta} \left(\frac{1 + e^{\Theta}}{2}\right)^{100}}_{=: f(\Theta)}$$

Ableiten ergibt: $f(\Theta)$ wird minimiert für $\Theta = \ln 4$ Daraus folgt:

$$P(X_{100} \geq 80) \leq 4^{-80} \cdot \left(\frac{1 + 4}{2}\right)^{100} \approx 4,26 \cdot 10^{-9}$$

Erwartungsgemäß ist dies eine wesentlich schärfere Schranke als (a) und (b).

66.9 Schwaches Gesetz der großen Zahlen

Mit Hilfe der Tschebyschew-Ungleichung kann man das schwache Gesetz der großen Zahlen beweisen:

Es werde ein Versuch n Mal wiederholt, bei dem das Ereignis A mit Wahrscheinlichkeit p eintritt. Dann strebt die Wahrscheinlichkeit, dass sich die relative Häufigkeit $h_n(A)$ um weniger als eine beliebig kleine Zahl $\epsilon > 0$ von p unterscheidet, gegen 1 für $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} P(|h_n(A) - p| < \epsilon) = 1$$

67 Wichtige diskrete Verteilungen

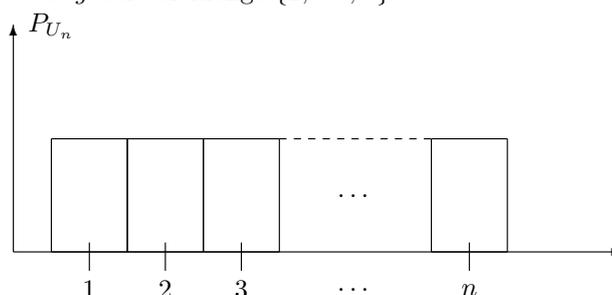
67.1 Motivation

Einige diskrete Verteilungen treten sehr häufig auf und tragen einen eigenen Namen. Wir wollen vier dieser Verteilungen genauer betrachten:

die Gleichverteilung, die Binomialverteilung, die Poisson-Verteilung, die geometrische Verteilung.

67.2 Die Gleichverteilung

Sei $\Omega = \{\omega_1, \dots, \omega_n\}$ ein Stichprobenraum mit $P(\omega_k) = \frac{1}{n}$ für $k = 1, \dots, n$ (Laplace-Experiment). Ferner sei U_n eine Zufallsvariable mit $U_n(\omega_k) = k$. Dann ist $P_{U_n}(k) = \frac{1}{n}$. Eine solche Verteilung heißt (diskrete) Gleichverteilung auf der Menge $\{1, \dots, n\}$.



Es gilt:

$$\begin{aligned}\mu &= E(U_n) = \sum_{k=1}^n k P_{U_n}(k) = \frac{1}{n} \sum_{k=1}^n k \\ \sigma^2 &= V(U_n) = E(U_n^2) - \mu^2 = \sum_{k=1}^n k^2 \cdot P_{U_n}(k) - \mu^2 = \frac{1}{n} \sum_{k=1}^n k^2 - \frac{1}{n^2} \left(\sum_{k=1}^n k \right)^2\end{aligned}$$

67.3 Beispiel

Beim Würfeln liegt eine diskrete Gleichverteilung auf der Menge $\{1, \dots, 6\}$ vor mit

$$\begin{aligned}\mu &= \frac{1}{6} \sum_{k=1}^6 k = \frac{21}{6} = 3,5 \\ \sigma^2 &= \frac{1}{6} \sum_{k=1}^6 k^2 - \mu^2 = \frac{1}{6} (1^2 + 2^2 + \dots + 6^2) - 3,5^2 \approx 2,917 \\ &\Rightarrow \sigma \approx 1,708\end{aligned}$$

67.4 Die Binomialverteilung

Wir betrachten ein Experiment, das aus einer n-fachen Wiederholung eines Einzelerperiments besteht, bei dem ein Ereignis A mit Wahrscheinlichkeit p eintritt. Ein solches Experiment heißt

Bernoulli-Experiment. Wir setzen

$$X_i(\omega) = \begin{cases} 1 & \text{falls } A \text{ beim } i\text{-ten Wurf eintritt} \\ 0 & \text{sonst} \end{cases}$$

A_k bezeichnet das Ereignis, dass im Gesamtexperiment gilt:

$$X(\omega) := \sum_{i=1}^n X_i(\omega) = k$$

Nach 62.4 (c) (ungeordnete Stichprobe ohne Wiederholung) hat A_k insgesamt $\binom{n}{k}$ Realisierungen. Jede solche Realisierung tritt mit Wahrscheinlichkeit $p^k(1-p)^{n-k}$ auf. Die entsprechende Verteilung

$$b_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}$$

heißt *Binomialverteilung* mit Parametern n und p . Es gilt:

$$\begin{aligned} E(X_i) &= 1 \cdot p + 0 \cdot (1-p) = p \quad \Rightarrow \quad E(X) = \sum_{i=1}^n E(X_i) = np \\ V(X_i) &= E(X_i^2) - (E(X_i))^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p) \end{aligned}$$

Da die Einzelexperimente unabhängig sind, gilt:

$$V(X) = \sum_{i=1}^n V(X_i) = np(1-p)$$

67.5 Beispiel: Zufallsabhängigkeit sportlicher Resultate

Andreas und Bernd tragen ein Tischtennisturnier mit $n = 1, 3, 5, \dots, 2m + 1$ Spielen aus. Sieger ist, wer die meisten Einzelspiele gewinnt. Andreas gewinnt ein Einzelspiel mit Wahrscheinlichkeit $p = 0,6$. Wie groß sind die Siegeschancen für den schlechteren Spieler Bernd?

Bernd siegt, wenn Andreas $S_n \leq m$ Erfolge erzielt. Die Wahrscheinlichkeit hierfür beträgt:

$$\begin{aligned} P(S_n \leq m) &= b_{n,p}(0) + b_{n,p}(1) + \dots + b_{n,p}(m) \\ &= \sum_{k=0}^m \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

$$n = 1 \quad : \quad P(S_1 \leq 0) = \binom{1}{0} 0,6^0 \cdot 0,4^1 = 0,4$$

$$n = 3 \quad : \quad P(S_3 \leq 1) = \binom{3}{0} 0,6^0 \cdot 0,4^3 + \binom{3}{1} 0,6^1 \cdot 0,4^2 \approx 0,352$$

$$n = 5 \quad : \quad P(S_5 \leq 2) = \binom{5}{0} 0,6^0 \cdot 0,4^5 + \binom{5}{1} 0,6^1 \cdot 0,4^4 + \binom{5}{2} 0,6^2 \cdot 0,4^3 \approx 0,317$$

$$n = 7 \quad : \quad P(S_7 \leq 3) \approx 0,290$$

$$n = 9 \quad : \quad P(S_9 \leq 4) \approx 0,267$$

Es ist also gar nicht so unwahrscheinlich, dass der schlechtere Spieler das Turnier gewinnt.

67.6 Die Poissonverteilung

Für große n wird das Arbeiten mit der Binomialverteilung unhandlich. Ist p klein ($0 \leq p \leq 0,1$), gibt es eine gute Approximation, die *Poisson-Verteilung* zum Parameter λ :

$$p(k) := \frac{\lambda^k}{k!} e^{-\lambda} \quad (\lambda > 0, k = 0, 1, 2, \dots)$$

Durch Umbenennen von „Erfolg“ und „Fehlschlag“ ist die Poissonverteilung auch für $0,9 \leq p \leq 1$ anwendbar.

Man kann zeigen, dass $p(k)$ für $\lambda = np$ die Binomialverteilung $b_{n,p}(k)$ approximiert. Ferner hat eine Poisson-verteilte Zufallsvariable den Erwartungswert λ und die Varianz λ . Generell beschreibt die Poisson-Verteilung Ereignisse, die im zeitlichen Verlauf zufällig und unabhängig von einander auftreten, z.B.:

- atomarer Zerfall
- das Eintreffen von Bedienwünschen an einem Server
- Anrufe in einem Call-Center
- das Auftreten von Softwarefehlern in einem Programmsystem

67.7 Reales Beispiel: Der große Jubiläumstag

Genau in einem Jahr feiert ein großer Betrieb seinen 100. Geburtstag. Die Direktion beschließt, allen Kindern von Betriebsangehörigen, die an diesem Tag geboren werden, ein Sparkonto von 3000€ anzulegen. Da rund 730 Kinder pro Jahr geboren werden, erwartet man Auslagen von 6000€. Um Zufallsschwankungen vorzubeugen, plant man 15.000€ ein. Wie groß ist die Wahrscheinlichkeit, dass das Geld nicht reicht?

$n = 730$ Kinder/Jahr

$p = \frac{1}{365}$ Wahrscheinlichkeit, dass Geburtstag auf Jubiläumstag fällt.

$\Rightarrow \lambda = p \cdot n = 2$

Das Geld reicht nicht, wenn $k \geq \frac{15.000}{3.000} = 6$ Kinder geboren werden.

$$\begin{aligned} p(k \geq 6) &= 1 - p(k \leq 5) = 1 - p(0) - p(1) - \dots - p(5) \\ &= 1 - \frac{2^0}{0!} e^{-2} - \frac{2^1}{1!} e^{-2} - \dots - \frac{2^5}{5!} e^{-2} \approx 0,0168 \end{aligned}$$

Die Wahrscheinlichkeit einer unangenehmen Zufallsüberraschung ist also gering. Man rechnet nicht damit.

Anmerkung: Am Jubiläumstag wurden 36 Kinder geboren! Die Direktion hat es also verstanden, ihre Betriebsangehörigen auch für außerbetriebliche Aktivitäten zu begeistern.

67.8 Die geometrische Verteilung

Eine diskrete Zufallsvariable X , die in einem Bernoulli-Experiment angibt, wann ein Ereignis A mit Wahrscheinlichkeit $p = p(A)$ zum ersten Mal eintritt, die heißt *geometrisch verteilt* mit Parameter p . Die entsprechende Verteilung lautet:

$$P_X(k) = p(1-p)^{k-1}$$

Man kann zeigen:

$$E(X) = \frac{1}{p}, \quad V(X) = \frac{1-p}{p^2}$$

Die geometrische Verteilung ist wichtig in der Informatik bei der Modellierung von Wartezeiten.

67.9 Beispiel

Wie lang muss man beim Würfeln warten, bis die Augenzahl 4 erstmalig auftritt?
Mit $p = \frac{1}{6}$ beträgt der Erwartungswert

$$E(X) = \frac{1}{p} = 6$$

und die Varianz

$$V(X) = \frac{1-p}{p^2} = \frac{\frac{5}{6}}{\frac{1}{6^2}} = 30.$$

68 Wichtige kontinuierliche Verteilungen

68.1 Motivation

Zufallsvariablen sind nicht immer diskret, sie können oft auch jede beliebige reelle Zahl in einem Intervall annehmen. Beispiele für solche „kontinuierlichen“ Zufallsvariablen sind Größe, Gewicht oder Zeit. In diesen Fällen macht es wenig Sinn, die Wahrscheinlichkeit anzugeben, dass eine Zufallsvariable einen bestimmten Wert annimmt (diese Wahrscheinlichkeit ist 0). Wir müssen Wahrscheinlichkeiten für Intervalle betrachten. Hierzu sind Begriffe wie Dichten notwendig.

68.2 Definition

Sei X eine kontinuierliche Zufallsvariable. Existiert eine Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ mit

a) $f(x) \geq 0 \quad \forall x \in \mathbb{R}$

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

und

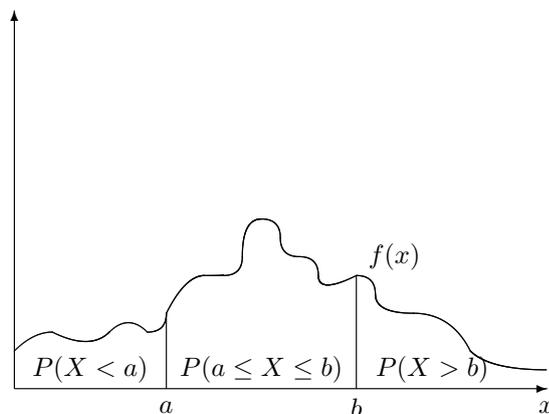
$$P(a \leq X \leq b) = \int_a^b f(x) \, dx$$

so nennt man $f(x)$ die *Dichte* von X .

Erwartungswert und Varianz von X sind definiert durch

$$\mu = E(X) := \int_{-\infty}^{\infty} x f(x) \, dx, \quad \sigma^2 = V(X) := \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx.$$

68.3 Veranschaulichung



68.4 Definition

Sei X eine kontinuierliche Zufallsvariable mit Dichte f . Dann nennt man die Stammfunktion

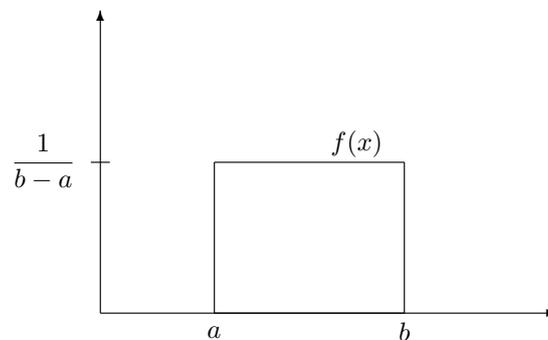
$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) \, du$$

die *Verteilungsfunktion* von X .

68.5 Beispiel: Kontinuierliche Gleichverteilung

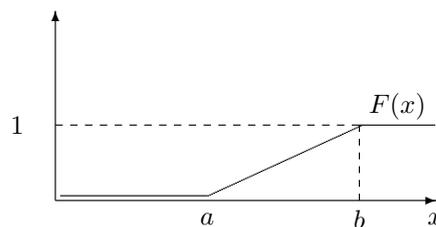
Eine kontinuierliche Zufallsvariable mit Gleichverteilung auf $[a, b]$ hat die Dichte

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



und die Verteilungsfunktion

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



Wir kommen nun zur wichtigsten kontinuierlichen Verteilung:

68.6 Die Standardnormalverteilung

Ist X eine kontinuierliche Zufallsvariable mit der Dichte

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

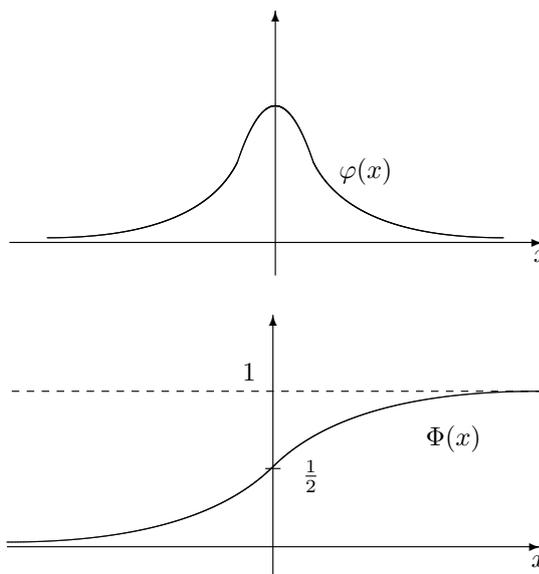
so kann man $P(a < X \leq b)$ mit der Stammfunktion

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$

berechnen:

$$P(a < X \leq b) = \Phi(b) - \Phi(a).$$

φ nennt man *normale Dichte* und Φ ist die *Standardnormalverteilung*.



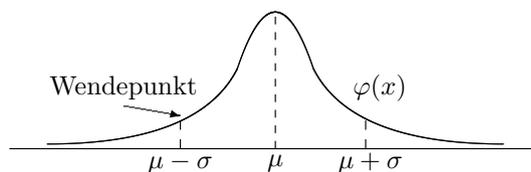
$\Phi(x)$ ist nicht analytisch auswertbar, liegt aber tabelliert vor.

Eine standardnormalverteilte Zufallsvariable hat Erwartungswert 0 und Varianz 1. Daher heißt sie auch $N(0,1)$ -verteilt.

68.7 Die allgemeine Normalverteilung

Eine kontinuierliche Zufallsvariable X genügt einer *allgemeinen Normalverteilung* ($N(\mu, \sigma^2)$ -Verteilung, *Gaußverteilung*) mit Erwartungswert μ und Varianz σ^2 , wenn ihre Dichte gegeben ist durch

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



Es gilt:

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 68\% \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 95,5\% \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 99,7\% \end{aligned}$$

Ist $X \sim N(\mu, \sigma^2)$ -verteilt, so ist $Y := \frac{X - \mu}{\sigma} \sim N(0,1)$ -verteilt. Somit ist die Tabelle der Standardnormalverteilung ausreichend.

68.8 Approximation der Binomialverteilung durch die Gaußverteilung

Eine Binomialverteilung mit n Einzelexperimenten mit Wahrscheinlichkeit p kann man durch eine Gaußverteilung mit Erwartungswert np und Standardabweichung $\sigma = \sqrt{np(1-p)}$ approximieren:

$$b_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \underbrace{\Phi\left(\frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)}_{\text{Fläche über dem Intervall } [k - \frac{1}{2}, k + \frac{1}{2}]}$$

Diese Approximation ist gut für $np > 5$ und $n(1-p) > 5$, d.h. insbesondere für große n oder $p \approx 0,5$.

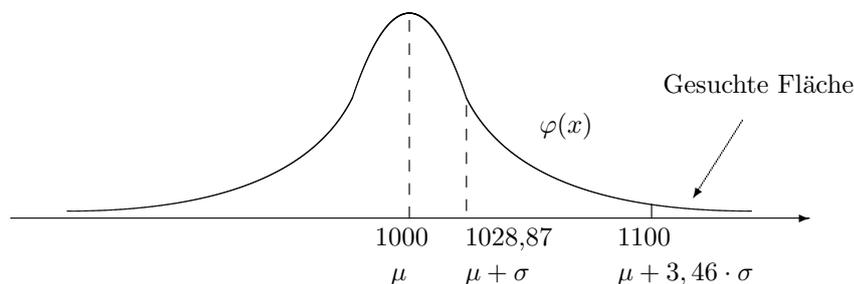
68.9 Beispiel

Wie groß ist die Wahrscheinlichkeit, dass in 6000 Würfeln eine fairen Würfels die Sechs mindestens 1100 Mal auftritt?

$$n = 6000, \quad p = \frac{1}{6}$$

Wegen $np = 1000 > 5$ und $n(1-p) = 5000 > 5$ ist die Approximation durch die Gaußverteilung sinnvoll.

$$\begin{aligned} \mu &= np = 6000 \cdot \frac{1}{6} = 1000 \\ \sigma &= \sqrt{np(1-p)} = \sqrt{6000 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx 28,87 \\ 1100 &= \underbrace{1000}_{\mu} + 3,46 \cdot \underbrace{28,87}_{\sigma} \end{aligned}$$



Die Wahrscheinlichkeit, mehr als 1100 Sechsen zu würfeln, ist

$$1 - \underbrace{\Phi(3,46)}_{\substack{\text{in Tabelle} \\ \text{nachschlagen}}} \approx 0,00028$$

68.10 Satz (Zentraler Grenzwertsatz)

Seien X_1, X_2, \dots unabhängige Zufallsvariablen, die alle die gleiche Verteilung und somit alle den gleichen Erwartungswert μ und die gleiche Varianz σ^2 besitzen. Ferner sei $Y_n := X_1 + \dots + X_n$ und

$$Z_n := \frac{Y_n - n\mu}{\sqrt{n}\sigma}$$

bezeichne die Standardisierte von Y_n (vgl. 65.15).

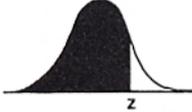
Dann konvergiert die Verteilungsfunktion $F_n(x)$ von Z_n für $n \rightarrow \infty$ gegen die Standardnormalverteilung $\Phi(x)$.

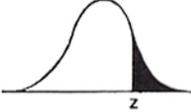
68.11 Bemerkungen

- a) Der Beweis von Satz 68.10 ist relativ aufwändig. Siehe z.B. R. Nelson: Probability, Stochastic Processes and Queueing Theory; Springer, New York 1995.
- b) Beachte, dass die einzelnen Zufallsvariablen X_i nicht normalverteilt sein müssen. Ihre Verteilung kann *beliebig* sein!
- c) Die Normalverteilung ist also eine sinnvolle Approximation in allen Fällen, in denen sich eine Zufallsvariable aus vielen gleichartigen Einzeleinflüssen zusammensetzt.
Beispiel: Eine Messung wird oft wiederholt. Dann approximieren die Ergebnisse eine Gaußverteilung.

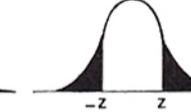
68.12 Tabelle der Standardnormalverteilung

Wahrscheinlichkeiten der folgenden vier Ereignisse:

$S_n - np \leq z \sigma$


$S_n - np \geq z \sigma$


$|S_n - np| \leq z \sigma$


$|S_n - np| \geq z \sigma$


z	$\Phi(z)$	$1 - \Phi(z)$	$2\Phi(z) - 1$	$2 - 2\Phi(z)$
0.0	.500	.500	.0000	1.0000
0.1	.540	.460	.0797	.9203
0.2	.579	.421	.159	.841
0.3	.618	.382	.236	.764
0.4	.655	.345	.311	.689
0.5	.691	.309	.383	.617
0.6	.726	.274	.451	.549
0.7	.758	.242	.516	.484
0.8	.788	.212	.576	.424
0.9	.816	.184	.632	.368
1.0	.841	.159	.683	.317
1.1	.864	.136	.729	.271
1.2	.885	.115	.770	.230
1.3	.9032	.0968	.806	.194
1.4	.9192	.0808	.838	.162
1.5	.9332	.0668	.866	.134
1.6	.9452	.0548	.890	.110
1.7	.9554	.0446	.9109	.0891
1.8	.9641	.0359	.9281	.0719
1.9	.9713	.0287	.9425	.0575
2.0	.9772	.0228	.9545	.0455
2.1	.9821	.0179	.9643	.0357
2.2	.9861	.0139	.9722	.0278
2.3	.9893	.0107	.9786	.0214
2.4	.99180	.00820	.9836	.0164
2.5	.99379	.00621	.9876	.0124
2.6	.99534	.00466	.99068	.00932
2.7	.99653	.00347	.99307	.00693
2.8	.99744	.00256	.99489	.00511
2.9	.99813	.00187	.99627	.00373
3.0	.99865	.00135	.99730	.00270
3.1	.999032	.000968	.99806	.00194
3.2	.999313	.000687	.99863	.00137
3.3	.999517	.000483	.999033	.000967
3.4	.999663	.000337	.999326	.000674
3.5	.999767	.000233	.999535	.000465
3.6	.999841	.000159	.999682	.000318
3.7	.999892	.000108	.999784	.000216
3.8	.9999277	.0000723	.999855	.000145
3.9	.9999519	.0000481	.9999038	.0000962
4.0	.9999683	.0000317	.9999367	.0000633

Die Stetigkeitskorrektur kann man berücksichtigen, indem man die Grenzen des schwarzen Gebiets um $\frac{1}{2\sigma}$ ins weiße Gebiet verlegt. Wahrscheinlichkeiten für $z < 0$ erhält man durch Symmetrie.

69 Multivariate Verteilungen und Summen von Zufallsvariablen

69.1 Motivation

Manchmal möchte man das Zusammenwirken mehrerer Zufallsvariablen X_1, \dots, X_n studieren. Gibt es in diesem „multivariaten“ Fall Aussagen über die gemeinsame Verteilung? Lassen sich Aussagen über die Verteilung der Summe von Zufallsvariablen treffen?

69.2 Wichtige Definitionen

Mehrere Zufallsvariablen X_1, \dots, X_n lassen sich in einem Vektor $X = (X_1, \dots, X_n)^\top$ zusammenfassen. Im kontinuierlichen Fall ist die resultierende Dichte eine Funktion mehrerer Variablen. Für diese *gemeinsame Dichte* gilt:

$$f(x_1, \dots, x_n) \geq 0 \quad \forall x_1, \dots, x_n \in \mathbb{R}$$

$$\int_{\mathbb{R}^n} f(x_1, \dots, x_n) dx_1 \cdot \dots \cdot dx_n = 1$$

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \dots, a_n \leq X_n \leq b_n) = \int_{a_n}^{b_n} \dots \int_{a_1}^{b_1} f(x_1, \dots, x_n) dx_1 \cdot \dots \cdot dx_n$$

Ferner betrachtet man die *multivariate Verteilungsfunktion*

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f(x_1, \dots, x_n) dx_1 \dots dx_n$$

Statt des Erwartungswerts betrachtet man den *Erwartungswertvektor*

$$\mu = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$$

und fasst Varianzen und Kovarianzen zusammen zur *Kovarianzmatrix*

$$\Sigma = \begin{pmatrix} V(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & V(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \dots & \dots & V(X_n) \end{pmatrix}$$

Sie ist symmetrisch und positiv definit.

Die wichtigste multivariate Verteilung ist die *multivariate Normalverteilung* ($N_n(\mu, \Sigma)$ -Verteilung): Ist $X = (X_1, \dots, X_n)^\top$ ein Vektor von normalverteilten Zufallsvariablen mit Erwartungswertvektor $\mu = (E(X_1), \dots, E(X_n))^\top$ und Kovarianzmatrix Σ , dann besitzt die multivariate Normalverteilung die Dichte

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \quad (x \in \mathbb{R}^n)$$

Sie spielt z.B. eine große Rolle im Bereich des maschinellen Lernens.

69.3 Beispiel

Eine Apfelbaumplantage mit gleich alten Bäumen werde durch drei normalverteilte Zufallsvariablen charakterisiert:

X_1 :	Höhe eines Baums [m]	$N(4,1)$ -verteilt
X_2 :	Ertrag [1 kg]	$N(20,100)$ -verteilt
X_3 :	Zahl der Blätter [1000 Stück]	$N(20,225)$ -verteilt

Diese Zufallsvariablen seien korreliert mit

$$\begin{aligned} \text{Cov}(X_1, X_2) &= 9 \\ \text{Cov}(X_1, X_3) &= 12,75 \\ \text{Cov}(X_2, X_3) &= 120 \end{aligned}$$

Dann liegt eine $N_3(\mu, \Sigma)$ -Verteilung vor mit

$$\mu = \begin{pmatrix} 4 \\ 20 \\ 20 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 9 & 12,75 \\ 9 & 100 & 120 \\ 12,75 & 120 & 225 \end{pmatrix}$$

Kann man unter geeigneten Voraussetzungen die gemeinsame Dichte $f(x_1, \dots, x_n)$ aus den einzelnen Dichten $f_1(x_1), \dots, f_n(x_n)$ berechnen? Man kann zeigen:

69.4 Satz (Gemeinsame Dichte von unabhängigen Zufallsvariablen)

Seien X_1, \dots, X_n unabhängige (!) Zufallsvariablen mit Dichten $f_1(x_1), \dots, f_n(x_n)$. Dann hat $X = (X_1, \dots, X_n)^\top$ die gemeinsame Dichte

$$f(x_1, \dots, x_n) = f_1(x_1) \cdot \dots \cdot f_n(x_n) \quad (*)$$

Hat umgekehrt $X = (X_1, \dots, X_n)^\top$ eine gemeinsame Dichte mit Produktdarstellung (*), so sind X_1, \dots, X_n unabhängig.

69.5 Beispiel

Zwei unabhängige Zufallsvariablen X_1, X_2 seien $N(\mu_1, \sigma_1^2)$ - bzw. $N(\mu_2, \sigma_2^2)$ -verteilt. Da Unabhängigkeit nach 65.22 (e) Unkorreliertheit impliziert, hat die Kovarianzmatrix Diagonalgestalt:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

mit

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$$

und $\det \Sigma = \sigma_1^2 \sigma_2^2$ liegt eine multivariate Normalverteilung für $X = (X_1, X_2)^\top$ vor mit Dichte

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}} e^{-\frac{1}{2}\left[(x_1-\mu_1)^2 \cdot \frac{1}{\sigma_1^2} + (x_2-\mu_2)^2 \cdot \frac{1}{\sigma_2^2}\right]} \quad (\text{vgl. 69.2}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}} \end{aligned}$$

Dies ist gerade das Produkt der zwei Einzeldichten $f_1(x_1)$ und $f_2(x_2)$.

Gibt es Aussagen über die Dichte, wenn man die Summe zweier Zufallsvariablen betrachtet? Hierzu benötigen wir

69.6 Definition

Falls für die Funktionen $f, g : \mathbb{R} \rightarrow \mathbb{R}$ das Integral

$$(f * g)(x) := \int_{-\infty}^{\infty} f(x-y)g(y) \, dy$$

existiert, so nennen wir $f * g$ die *Faltung* von f und g (engl.: *convolution*).

69.7 Satz (Summe unabhängiger kontinuierlicher Zufallsvariabler)

Seien X_1, X_2 unabhängige kontinuierliche Zufallsvariablen mit Dichten f_1, f_2 . Dann hat $X_1 + X_2$ die Dichte $f_1 * f_2$.

Beweis: Mit $B := \{(x_1, x_2) \mid x_1 + x_2 \leq s\}$ ergibt sich für die Verteilung von $X_1 + X_2$:

$$P(X_1 + X_2 \leq s) = \iint_B \underbrace{f_1(x_1) \cdot f_2(x_2)}_{\text{unabhängig}} \, dx_1 \, dx_2$$

Mit der Substitution $u := x_1 + x_2$ folgt:

$$P(X_1 + X_2 \leq s) = \int_{-\infty}^s \underbrace{\left(\int_{-\infty}^{\infty} f_1(u-x_2)f_2(x_2) \, dx_2 \right)}_{(f_1 * f_2)(u)} \, du$$

□

Hiermit lässt sich beweisen:

69.8 Satz (Summe unabhängiger normalverteilter Zufallsvariabler)

Seien X_1, X_2 unabhängige kontinuierliche Zufallsvariablen mit $N(\mu_1, \sigma_1^2)$ - bzw. $N(\mu_2, \sigma_2^2)$ -Verteilung. Dann ist $X = X_1 + X_2$ $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ -verteilt.

Auch im Fall diskreter Zufallsvariabler gibt es vergleichbare Aussagen zu 69.6 - 69.8:

69.9 Definition

Für $f = (f_i)_{i \in \mathbb{Z}}, g = (g_i)_{i \in \mathbb{Z}}$ definiert man die *diskrete Faltung* von f und g durch

$$(f * g)_i := \sum_{j \in \mathbb{Z}} f_{i-j} g_j$$

69.10 Satz (Summe unabhängiger diskreter Zufallsvariabler)

Seien X_1, X_2 unabhängige diskrete Zufallsvariablen mit Verteilungen P_{x_1}, P_{x_2} . Dann hat $X_1 + X_2$ die Verteilung $P_{x_1} * P_{x_2}$, wobei $*$ die diskrete Faltung bezeichnet.

69.11 Satz (Summe unabhängiger Poisson-verteilter Zufallsvariabler)

Seien X_1, X_2 unabhängige diskrete Zufallsvariablen, die einer Poisson-Verteilung mit Parameter λ_1 bzw. λ_2 genügen (kurz: $P(\lambda_1)$ -, $P(\lambda_2)$ -verteilt). Dann ist $X_1 + X_2$ $P(\lambda_1 + \lambda_2)$ -verteilt.

69.12 Beispiel

Beim radioaktiven Zerfall einer Substanz werden ionisierende Teilchen frei. Mit einem Geiger-Müller-Zählrohr zählt man die innerhalb einer Minute eintreffenden Teilchen. Sie sind Poisson-verteilt. Hat man zwei radioaktive Substanzen mit Poisson-Verteilung $P(\lambda_1)$ bzw. $P(\lambda_2)$, so genügt die Gesamtheit der pro Zeitintervall produzierten Teilchen einer $P(\lambda_1 + \lambda_2)$ -Verteilung.

70 Parameterschätzung und Konfidenzintervalle

70.1 Motivation

Bisher sind wir stets von theoretischen Modellen (z.B. „fairer Würfel“) ausgegangen, die erlauben, Parameter wie Erwartungswert oder Varianz einer Verteilung exakt zu berechnen. In vielen realen Situationen kennt man jedoch nur den Verteilungstyp und muss auf Grund von Stichproben die Parameter schätzen. Wie geht man dabei vor?

Die geschätzten Parameter sind im Allgemeinen Fehler behaftet. Lässt sich ein Vertrauensintervall angeben, innerhalb dessen ein Parameter mit einer vorgegebenen Sicherheit liegt?

70.2 Definition

Gegeben seien n Beobachtungswerte x_1, \dots, x_n eines Zufallsexperiments.

Dann nennen wir $(x_1, \dots, x_n)^\top$ *Stichprobe vom Umfang n* . Die einzelnen x_i heißen *Stichprobenwerte*.

70.3 Beispiel

In einer Kiste befinden sich 10.000 Schrauben. Ein Teil davon ist fehlerhaft. Für eine Stichprobe werden 100 Schrauben entnommen. Die Zufallsvariable X_i beschreibt den Zustand der i -ten entnommenen Schraube:

$$X_i(\omega) = \begin{cases} 0 & \text{falls } i\text{-te Schraube in Ordnung} \\ 1 & \text{falls } i\text{-te Schraube fehlerhaft} \end{cases}$$

Eine konkrete Realisierung des Zufallsvektors $(X_1, \dots, X_{100})^\top$ liefere eine Stichprobe $(x_1, \dots, x_{100})^\top = (0, 1, 0, 0, 1, 0, \dots, 1, 0)^\top$.

So wie wir bei den Zufallsvariablen Parameter wie Erwartungswert oder Varianz bestimmt haben, können wir auch für Stichproben entsprechende Kerngrößen definieren:

70.4 Definition

Für eine Stichprobe $(x_1, \dots, x_n)^\top$ definiert man:

- den *Mittelwert (arithmetisches Mittel)* durch:

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

- die *Varianz* durch:

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

- die *Standardabweichung* durch:

$$s := \sqrt{s^2}$$

70.5 Bemerkungen

- Man kann zeigen, dass der Mittelwert \bar{x} und die Varianz s^2 einer Stichprobe geeignete Approximationen an den Erwartungswert μ und die Varianz σ^2 einer Zufallsvariablen sind.
- Die Tatsache, dass im Nenner von s^2 die Größe $n - 1$ statt n steht, hat tiefere theoretische Gründe, auf die wir hier nicht eingehen (siehe z.B. Hartmann, Satz 21.9).
- Ähnlich zum Verschiebungssatz 65.12 gibt es eine häufig benutzte Formel zum Berechnen der Varianz einer Stichprobe:

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

Beweis:

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum_{i=1}^n x_i^2 - 2\bar{x} \underbrace{\sum_{i=1}^n x_i}_{n\bar{x}} + n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \end{aligned}$$

□

70.6 Beispiel

Bei einer Wahlumfrage geben 400 von 1000 Personen an, Partei A wählen zu wollen. Das Umfrageninstitut prognostiziert auf Grund dieser Stichprobe einen Wahlausgang mit 40% aller Stimmen für Partei A.

70.7 Konfidenzintervalle

In Beispiel 70.6 werden verschiedene Stichproben zu unterschiedlichen Resultaten führen, die wiederum i.A. vom tatsächlichen Wahlausgang abweichen.

Können wir statt eines einzelnen Werts $p = 0,4$ ein Vertrauensintervall (*Konfidenzintervall*) $[p_n, p_0]$ angeben, innerhalb dessen das Endresultat mit einer angegebenen Wahrscheinlichkeit (*Konfidenzniveau*) von z.B. 95% liegt?

70.8 Beispiel: Wahlumfrage aus 70.6

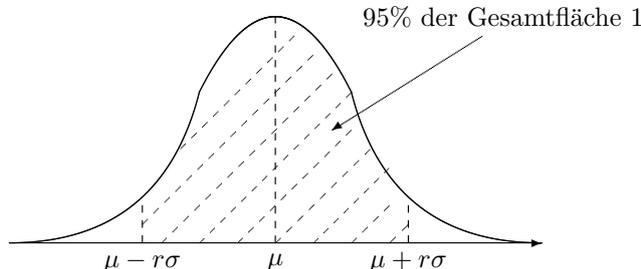
Wir gehen von einer Binomialverteilung aus und schätzen p durch $p = \frac{400}{1000} = 0,4$ ab. Sei

$$X_i = \begin{cases} 1 & \text{Befragte/r } i \text{ wählt A} \\ 0 & \text{Befragte/r } i \text{ wählt A nicht} \end{cases} \quad \text{und } X := \sum_{i=1}^{1000} X_i$$

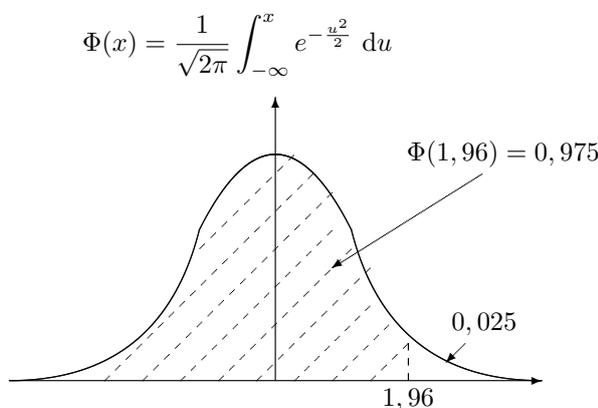
Dann gilt nach 67.4 mit $n = 1000$ und $p = 0,4$:

$$\begin{aligned} \mu &= E(X) = n \cdot p = 400 \\ \sigma^2 &= V(X) = np(1-p) = 240 \Rightarrow \sigma \approx 15,49 \end{aligned}$$

Wegen $n \cdot p = 400 > 5$ und $n(1 - p) = 600 > 5$ können wir für X auch eine Normalverteilung mit $\mu = 400$ und $\sigma = 15,49$ annehmen. Wir suchen ein Intervall $[\mu - r\sigma, \mu + r\sigma]$, innerhalb dessen das Integral über die Dichtefunktion den Wert 0,95 annimmt:



Tabelliert ist die Standardnormalverteilung ($\mu = 0, \sigma = 1$)



Man findet:

$$\Phi(1,96) \approx 0,975$$

Somit ist aus Symmetriegründen

$$\frac{1}{\sqrt{2\pi}} \int_{-1,96}^{1,96} e^{-\frac{u^2}{2}} du \approx 0,95$$

Somit ist $r = 1,96$.

Damit ergibt sich das Konfidenzintervall

$$[\mu - r\sigma, \mu + r\sigma] = [400 - 1,96 \cdot 15,49; 400 + 1,96 \cdot 15,49] \approx [369,6; 430,4]$$

Bei einem Konfidenzniveau von 95% erzielt Partei A also zwischen 36,96% und 43,04% der Stimmen. Möchte man ein kleineres Konfidenzintervall, muss man mehr Personen befragen.

70.9 Beispiel: Überbuchung eines Flugzeugs

Ein Flugzeug hat 200 Sitze. Wie viele Reservierungen dürfen angenommen werden, wenn erfahrungsgemäß 5% aller Passagiere nicht erscheinen? Die Flugesellschaft ist bereit, in 1 von 50 Fällen in Verlegenheit zu geraten.

Sei n die Anzahl der Reservierungen und X die Anzahl der tatsächlich erscheinenden Passagiere. Legt man eine Binomialverteilung zu Grunde mit $p = 0,95$, gilt:

$$\begin{aligned} \mu &= E(X) = n \cdot p = 0,95 \cdot n \\ \sigma^2 &= np(1 - p) \approx 0,0475n \Rightarrow \sigma = 0,2179\sqrt{n} \end{aligned}$$

Der Tabelle der Standardnormalverteilung entnimmt man:

$$\Phi(2,05) \approx 0,98 = 1 - \frac{1}{50}$$

Fordert man

$$\mu + 2,05 \cdot \sigma \stackrel{!}{\leq} 200$$

ergibt sich:

$$0,95n + 2,05 \cdot 0,2179\sqrt{n} \leq 200$$

Substituiert man $y = \sqrt{n}$ erhält man eine quadratische Gleichung in y . Als Lösung für n ergibt sich $n = 217,5$. D.h. man darf 217 Reservierungen annehmen.

Bemerkung: Die Approximation durch die Normalverteilung war gerechtfertigt wegen $np > 5$ und $n(1-p) > 5$.

71 Hypothesentests

71.1 Motivation

Bei Hypothesentests möchte man eine gewisse Annahme über eine Zufallsvariable darauf hin überprüfen, ob sie korrekt ist. Beispiele:

- Ist eine Münze fair ($p = \frac{1}{2}$)?
- Sind die Rechner von Hersteller A zuverlässiger als von Hersteller B?

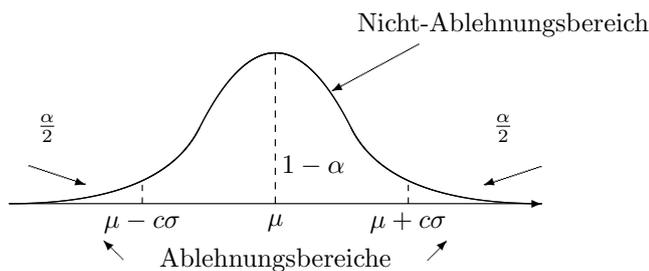
Ein statistisches Experiment soll uns dabei eine Entscheidung mit einer vorgegebenen Irrtumswahrscheinlichkeit ermöglichen. Gegenüber den Verfahren aus Kapitel 70 kann man in den Rechnungen die Hypothese mit verwenden, hat also mehr in der Hand.

71.2 Parametertest am Beispiel eines Münzexperimentes

Wir beobachten das Ereignis $A =$ „Münze zeigt Kopf“ und wollen die Hypothese $p_0 = p(A) = \frac{1}{2}$ überprüfen. Hierzu führen wir 200 Münzwürfe durch:

$$X_i := \begin{cases} 1 & \text{(Münze zeigt Kopf beim } i\text{-ten Wurf)} \\ 0 & \text{(Münze zeigt Zahl beim } i\text{-ten Wurf)} \end{cases}$$

Wie weit darf $S_{200} := \sum_{i=1}^{200} X_i$ sich vom Erwartungswert 100 unterscheiden, damit wir mit einer Irrtumswahrscheinlichkeit von $\alpha = 0,05$ (*Signifikanzniveau* von $1 - \alpha = 0,95$) die Hypothese $p_0 = \frac{1}{2}$ nicht ablehnen?



Wir legen eine Binomialverteilung mit $n = 200$, $p = 0,5$ zu Grunde, die wir durch eine Normalverteilung mit $\mu = np = 100$, $\sigma = \sqrt{np(1-p)} = \sqrt{50} \approx 7,07$ approximieren.

Wegen $\Phi(1,96) \approx 0,975$ ist $c = 1,96$ für $\alpha = 0,05$. Tritt bei 200 Würfeln eine Kopffzahl S_n außerhalb des Intervalls $[\mu - c\sigma, \mu + c\sigma] \approx [86, 1; 113, 6]$ auf, lehnt man die Hypothese $p_0 = \frac{1}{2}$ auf einem Signifikanzniveau von 0,95 ab. Andernfalls lehnt man sie nicht ab.

71.3 Bemerkungen

- Eine Hypothese an einen Parameter (etwa $p_0 = \frac{1}{2}$) nennt man auch *Nullhypothese* H_0 , die Gegenannahme (etwa $p \neq \frac{1}{2}$) ist die *Gegenhypothese* H_1 .
- Bei Hypothesentests können zwei Arten von Fehlern auftreten:

- *Fehler 1. Art:* Hypothese wird abgelehnt, obwohl sie richtig ist (wird durch die Irrtumswahrscheinlichkeit α beschrieben).
- *Fehler 2. Art:* Hypothese wird nicht abgelehnt, obwohl sie falsch ist. Dieser Fehler kann insbesondere für kleines α sehr groß werden.

71.4 Der χ^2 -Test (Chi-Quadrat-Test)

Der χ^2 -Test ist einer der wichtigsten Tests. Er wird bei folgendem Problem angewandt: Ein Versuch habe m mögliche Ausgänge. Wir testen die Hypothese H_0 , dass die Resultate mit vorgegebenen Wahrscheinlichkeiten p_1, \dots, p_m auftreten. Trifft H_0 zu, erwarten wir bei n Versuchen als Häufigkeit für die einzelnen Ausgänge: np_1, \dots, np_m . In Wirklichkeit werden die Häufigkeiten X_1, \dots, X_m beobachtet. Als Maß für die Abweichung zwischen X_i und np_i , $i = 1, \dots, m$ verwendet man

$$\chi^2 := \sum_{i=1}^m \frac{(X_i - np_i)^2}{np_i}$$

Ist χ^2 „zu groß“, wird man H_0 ablehnen.

Um zu beurteilen, was „zu groß“ bedeutet, ist es sinnvoll den Erwartungswert von χ^2 zu kennen. Ist jedes X_i $b_{n,p}$ -verteilt, gilt

$$V(X_i) = E((X_i - np_i)^2) = np_i(1 - p_i)$$

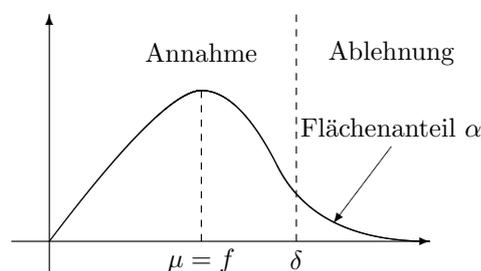
und aus der Linearität des Erwartungswerts folgt:

$$\begin{aligned} E(\chi^2) &= \sum_{i=1}^m \frac{1}{np_i} E((X_i - np_i)^2) = \sum_{i=1}^m \frac{1}{np_i} np_i(1 - p_i) \\ &= \sum_{i=1}^m 1 - \sum_{i=1}^m p_i = m - 1 \end{aligned}$$

$f := m - 1$ bezeichnet die *Freiheitsgrade* der χ^2 -Verteilung, d.h. $m - 1$ der p_i , $i = 1, \dots, m$ sind frei wählbar. Es gilt also

$$\mu = E(\chi^2) = f$$

Der typische Verlauf einer χ^2 -Verteilung sieht folgendermaßen aus:



Für einen gegebenen Freiheitsgrad f und eine Irrtumswahrscheinlichkeit α ist die χ_f^2 -Verteilung tabelliert.

Man nimmt H_0 an, falls der berechnete χ^2 -Wert $\leq \delta$ ist.

Man lehnt H_0 ab, falls der berechnete χ^2 -Wert $> \delta$ ist.

71.5 Beispiel

Wir wollen mit 120 Würfeln nachprüfen, ob ein Würfel „fair“ ist, d.h. alle Augenzahlen sind gleich wahrscheinlich sind:

$$p_1 = \dots = p_6 = \frac{1}{6}$$

Als Ergebnis erhalten wir:

Augenzahl i	1	2	3	4	5	6
beobachtete Häufigkeit X_i	15	21	25	19	14	26
erwartete Häufigkeit np_i	20	20	20	20	20	20

Wir erhalten:

$$\chi^2 = \frac{(15 - 20)^2}{20} + \frac{(21 - 20)^2}{20} + \dots + \frac{(26 - 20)^2}{20} \approx 6,2$$

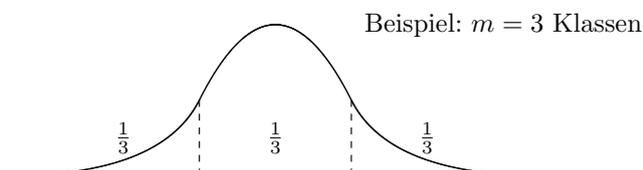
Geben wir eine Irrtumswahrscheinlichkeit von $\alpha = 0,1$ vor, so findet man in einer Tabelle für $f = 6 - 1 = 5$ Freiheitsgrade:

$$p(\chi_5^2 \leq \underbrace{9,24}_{\delta}) = \underbrace{0,9}_{1-\alpha}$$

Wegen $\chi^2 = 6,2 \leq 9,24 = \delta$ akzeptieren wir die Hypothese H_0 , dass alle Augenzahlen gleich wahrscheinlich sind.

71.6 Bemerkung

Möchte man eine $N(\mu, \sigma^2)$ -Verteilung mit dem χ^2 -Test für ein gegebenes μ, σ^2 verifizieren, teilt man in m (z.B. gleich wahrscheinliche) Klassen ein:



Dann überprüft man, ob die experimentell ermittelten Häufigkeiten in jeder Klasse das χ^2 -Kriterium zu einer vorgegebenen Irrtumswahrscheinlichkeit α und $f = m - 1$ Freiheitsgraden erfüllen.

71.7 Schranken für χ^2 bei f Freiheitsgraden

$f \backslash P$	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01
1	0,00016	0,00098	0,00393	0,01579	2,70554	3,84146	5,02389	6,63490
2	0,00201	0,00506	0,10259	0,21072	4,60517	5,99147	7,37776	9,21034
3	0,11483	0,21580	0,35185	0,58438	6,25139	7,81473	9,34840	11,3449
4	0,29711	0,48442	0,71072	1,06362	7,77944	9,48773	11,1433	13,2767
5	0,55430	0,83121	1,14548	1,61031	9,23635	11,0705	12,8325	15,0863
6	0,87209	1,23735	1,63539	2,20413	10,6446	12,5916	14,4494	16,8119
7	1,23904	1,68987	2,16735	2,83311	12,0170	14,0671	16,0128	18,4753
8	1,64648	2,17973	2,73264	3,48954	13,3616	15,5073	17,5346	20,0902
9	2,08781	2,70039	3,32511	4,16816	14,6837	16,9190	19,0228	21,6660
10	2,55821	3,24697	3,94030	4,86518	15,9871	18,3070	20,4831	23,2093
11	3,0535	3,8158	4,5748	5,5778	17,275	19,675	21,920	24,725
12	3,5706	4,4038	5,2260	6,3038	18,549	21,026	23,337	26,217
13	4,1069	5,0087	5,8919	7,0415	19,812	22,362	24,736	27,688
14	4,6604	5,6287	6,5706	7,7895	21,064	23,685	26,119	29,143
15	5,2294	6,2621	7,2604	8,5468	22,307	24,996	27,488	30,578
16	5,812	6,908	7,962	9,312	23,54	26,30	28,85	32,00
17	6,408	7,564	8,672	10,09	24,77	27,59	30,19	33,41
18	7,015	8,231	9,390	10,86	25,99	28,87	31,53	34,81
19	7,633	8,907	10,12	11,65	27,20	30,14	32,85	36,19
20	8,260	9,591	10,85	12,44	28,41	31,41	34,17	37,57
21	8,897	10,28	11,59	13,24	29,62	32,67	35,48	38,93
22	9,542	10,98	12,34	14,04	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	14,85	32,00	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96
28	13,56	15,31	16,93	18,94	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89

Erläuterung der Tafel. Z.B. die 5. Zeile der Tafel bedeutet: Bei 5 Freiheitsgraden ist $P(\chi^2 \geq 0,5543) = 0,99$, $P(\chi^2 \geq 0,83121) = 0,975, \dots, P(\chi^2 \geq 15,0863) = 0,01$.

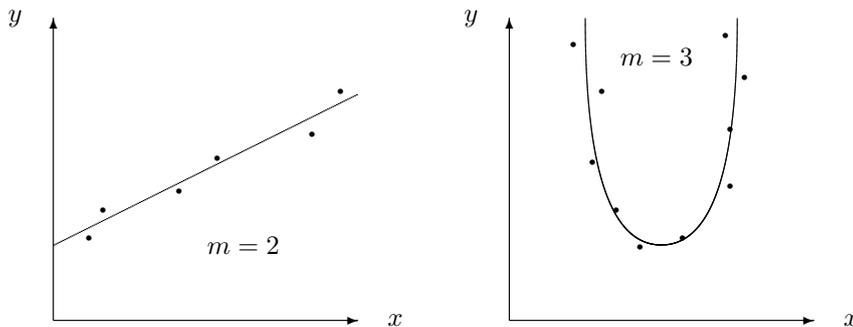
72 Methode der kleinsten Quadrate

72.1 Problemstellung

In einem Experiment interessiert man sich für die Beziehung zwischen zwei Variablen x und y . Hierzu hat man viele Wertepaare $(x_1, y_1), \dots, (x_n, y_n)$ gemessen. Die Messungen können Fehler in Form von statistischen Fluktuationen enthalten. Man möchte nun die Beziehung zwischen x und y durch ein einfaches Polynom

$$y = f(x) = \sum_{k=1}^m a_k x^{k-1}$$

approximieren (z.B. $m = 2$: Gerade, $m = 3$: Parabel) und sucht die „optimalen“ Koeffizienten a_1, \dots, a_m .



72.2 Methode der kleinsten Quadrate

Jede der n Messungen (x_i, y_i) beschreibt eine lineare Gleichung für die Unbekannten a_1, \dots, a_m :

$$\begin{aligned} a_1 + a_2 x_1 + \dots + a_m x_1^{m-1} &= y_1 \\ &\vdots \\ a_1 + a_2 x_n + \dots + a_m x_n^{m-1} &= y_n \end{aligned}$$

Im Allgemeinen hat man sehr viel mehr Gleichungen als Unbekannte ($n \gg m$), und das lineare Gleichungssystem

$$\underbrace{\begin{pmatrix} 1 & x_1 & \dots & x_1^{m-1} \\ \vdots & & & \vdots \\ 1 & x_n & \dots & x_n^{m-1} \end{pmatrix}}_{M \in \mathbb{R}^{n \times m}} \underbrace{\begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}}_{a \in \mathbb{R}^m} = \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{y \in \mathbb{R}^n}$$

ist inkonsistent, d.h. einzelne Gleichungen widersprechen sich. Beispielsweise kann man nicht erwarten, dass 50 Messwerte exakt auf einer Geraden liegen ($n = 50, m = 2$).

Da $Ma = y$ nicht exakt lösbar ist, sucht man statt dessen eine „Lösung“ a^* , die den *quadratischen Fehler*

$$|Ma - y|^2 = \left(\sum_{k=1}^m a_k x_1^{k-1} - y_1 \right)^2 + \dots + \left(\sum_{k=1}^m a_k x_n^{k-1} - y_n \right)^2$$

minimiert.

Die Koeffizienten a_1, \dots, a_m bestimmen dann die *Ausgleichskurve* (*Regressionskurve*)

$$y = \sum_{k=1}^m a_k x^{k-1}.$$

72.3 Minimierung des quadratischen Fehlers

Wir suchen das Minimum der Funktion

$$\begin{aligned} f(a_1, \dots, a_m) &= |Ma - y|^2 = (Ma - y)^\top (Ma - y) \\ &= (a^\top M^\top - y^\top)(Ma - y) \\ &= a^\top M^\top Ma - a^\top M^\top y - \underbrace{y^\top Ma}_{a^\top M^\top y} + y^\top y \\ &= a^\top M^\top Ma - 2a^\top M^\top y + y^\top y \end{aligned}$$

Notwendige Bedingung:

$$\begin{aligned} 0 \stackrel{!}{=} \nabla_a f &:= \begin{pmatrix} \frac{\partial f}{\partial a_1} \\ \vdots \\ \frac{\partial f}{\partial a_m} \end{pmatrix} = M^\top Ma + \underbrace{a^\top M^\top M}_{\substack{M^\top Ma, \\ \text{da } M^\top M \text{ symmetrisch}}} - 2M^\top y \\ &= 2M^\top Ma - 2M^\top y \end{aligned}$$

Die Lösung a^* löst also die so genannte *Normalengleichung*

$$\boxed{M^\top Ma = M^\top y}$$

Dies ist ein System aus m Gleichungen mit m Unbekannten a_1, \dots, a_m . Ist $M^\top M$ invertierbar, gilt:

$$\boxed{a^* = (M^\top M)^{-1} M^\top y}$$

Man nennt

$$M^+ := (M^\top M)^{-1} M^\top$$

die *Pseudoinverse* (*Moore-Penrose-Inverse*) der (nicht invertierbaren!) $n \times m$ -Matrix M .

72.4 Bemerkungen

a) a^* ist tatsächlich ein Minimum:

Die Hesse-Matrix $H_a f = 2M^\top M$ ist positiv semidefinit:

$$x^\top 2M^\top M x = 2(Mx)^\top Mx = 2|Mx|^2 \geq 0 \quad \forall x \in \mathbb{R}^m$$

Da $M^\top M$ invertierbar sein soll, ist $H_a f$ sogar positiv definit. Nach 57.5 folgt also: a^* ist Minimum.

b) Man kann zeigen, dass $M^\top M$ invertierbar ist, falls $\text{rang}(M) = m$, d.h. es gibt m der n Gleichungen des Systems $Ma = y$, die linear unabhängig sind.

c) Wir haben also am Beispiel der Ausgleichsrechnung ein allgemeines Verfahren hergeleitet, um ein überbestimmtes (und i.A. inkonsistentes) Gleichungssystem zu „lösen“:

72.5 Satz (Pseudolösung überbestimmter Gleichungssysteme)

Sei $n > m$, $A \in \mathbb{R}^{n \times m}$, $\text{rang}(A) = m$ und $b \in \mathbb{R}^n$. Falls das überbestimmte lineare Gleichungssystem

$$Ax = b$$

inkonsistent ist, ist es nicht exakt lösbar. Es gibt jedoch eine eindeutig bestimmte *Pseudolösung* x^* , die den quadratischen Fehler $|Ax - b|^2$ minimiert:

$$x^* = A^+b$$

mit der Pseudoinversen $A^+ = (A^T A)^{-1} A^T$.

72.6 Bemerkung

Pseudolösungen spielen auch in der Informatik eine große Rolle. Überbestimmte Gleichungssysteme treten z.B. bei der Suche in Internetdatenbanken auf (\rightarrow Prof. Weikum: Informationssysteme).

72.7 Beispiel

Bestimme mit der Methode der kleinsten Quadrate die Regressionsgerade durch die 4 Punkte $(0, 1), (1, 3), (2, 4), (3, 4)$.

Lösung: Wir suchen die Koeffizienten a_1, a_2 der Geradengleichung

$$y = a_1 + a_2x$$

Hierzu haben wir 4 Bestimmungsgleichungen:

$$\begin{aligned} a_1 + 0 \cdot a_2 &= 1 \\ a_1 + 1 \cdot a_2 &= 3 \\ a_1 + 2 \cdot a_2 &= 4 \\ a_1 + 3 \cdot a_2 &= 4 \end{aligned}$$

Das überbestimmte Gleichungssystem lautet: $Ma = y$ mit

$$M = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad a = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 4 \end{pmatrix}$$

$$\Rightarrow M^T M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} 4 & 6 \\ 6 & 14 \end{pmatrix}$$

$(M^T M)^{-1}$ existiert, da $\det(M^T M) = 4 \cdot 14 - 6 \cdot 6 = 20 \neq 0$.

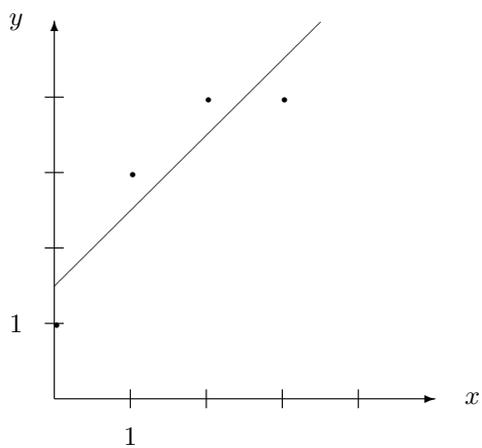
Nach kurzer Rechnung erhält man:

$$(M^T M)^{-1} = \frac{1}{10} \begin{pmatrix} 7 & -3 \\ -3 & 2 \end{pmatrix}$$

Damit lautet die Pseudolösung von $Ma = y$:

$$\begin{aligned} a^* = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} &= (M^\top M)^{-1} M^\top y \\ &= \frac{1}{10} \begin{pmatrix} 7 & -3 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 4 \\ 4 \end{pmatrix} \\ &= \frac{1}{10} \begin{pmatrix} 7 & -3 \\ -3 & 2 \end{pmatrix} \begin{pmatrix} 12 \\ 23 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 15 \\ 10 \end{pmatrix} = \begin{pmatrix} 1,5 \\ 1 \end{pmatrix} \end{aligned}$$

Die Ausgleichsgerade ist somit $y = 1,5 + x$.



73 Robuste Statistik

73.1 Motivation

Will man aus realen Daten statistische Parameter schätzen (z.B. arithmetisches Mittel als Schätzer für Erwartungswert; Parameter einer Regressionskurve), kann es sein, dass das Ergebnis auf Grund von Ausreißern stark verfälscht wird.

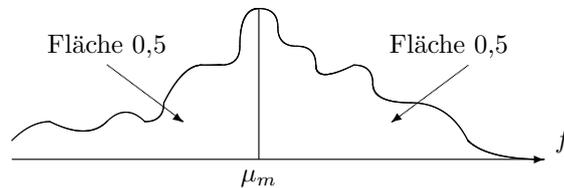
Beispiel: 9 Studierende benötigen 10 Semester für ihr Studium, einer benötigt 40 Semester. Das arithmetische Mittel liefert eine mittlere Studiendauer von 13 Semestern. Sie ist jedoch nicht repräsentativ für die Mehrzahl der Studierenden.

Gibt es statistische Verfahren, die robuster gegenüber Ausreißern sind?

73.2 Median

Sei X eine Zufallsvariable. Dann nennt man jede Zahl μ_m mit $P(X \geq \mu_m) \geq \frac{1}{2}$ und $P(X \leq \mu_m) \geq \frac{1}{2}$ einen *Median* von X .

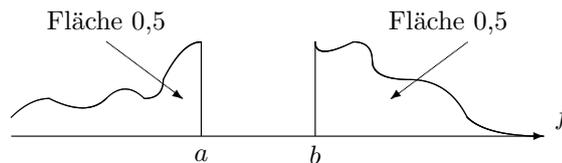
Veranschaulichung für kontinuierliche Zufallsvariable mit Dichte f :



Für die Verteilungsfunktion F gilt: $F(\mu_m) = \frac{1}{2}$.

73.3 Bemerkungen

- a) Nicht immer gibt es einen eindeutigen Median. Gibt es ein Intervall $[a, b]$ mit $P(X \leq a) = \frac{1}{2}$ und $P(X \geq b) = \frac{1}{2}$, ist jede Zahl aus $[a, b]$ ein Median:



- b) I.A. stimmen Median und Erwartungswert nicht überein.

73.4 Empirischer Median

Hat man $2k + 1$ der Größe nach geordnete Messwerte

$$x_1 \leq x_2 \leq \dots \leq x_{2k+1}$$

dann nennt man $\hat{\mu}_m := x_{k+1}$ den (*empirischen*) *Median*.

Es sind 50% der Daten $\geq \hat{\mu}_m$ und 50% sind $\leq \hat{\mu}_m$.

Bei einer geraden Anzahl von Messungen

$$x_1 \leq x_2 \leq \dots \leq x_{2k}$$

gilt für jedes $\hat{\mu} \in [x_k, x_{k+1}]$: $\geq 50\%$ der Daten sind $\geq \hat{\mu}$ und $\geq 50\%$ sind $\leq \hat{\mu}$. Man definiert in diesem Fall:

$$\hat{\mu}_m := \frac{1}{2}(x_k + x_{k+1})$$

als „den“ *empirischen Median*.

73.5 Beispiele

- Der Median der Studiendauer aus 73.1 beträgt 10 Semester. Der Ausreißer mit 40 Semestern hat somit keinen Einfluss auf den Median.
- In der Bildverarbeitung ersetzt der Medianfilter einen Grauwert durch seinen Median innerhalb einer $(2k + 1) \times (2k + 1)$ -Maske:

32	17	24
35	251	21
12	24	25

Ordnen der Grauwerte:

$$12 \leq 17 \leq 21 \leq 24 \leq \underset{\substack{\uparrow \\ \text{Median}}}{24} \leq 25 \leq 32 \leq 35 \leq 251$$

Der Grauwert 251 (Ausreißer) wird durch den Median 24 ersetzt. Medianfilter sind robust gegenüber Ausreißern (Impulsrauschen) und erhalten Kanten.

73.6 M-Schätzer

Seien x_1, \dots, x_n Messwerte und $\Psi : [0, \infty) \rightarrow \mathbb{R}$ eine monoton wachsende *Strafffunktion* (engl.: *penalizer*). Dann nennt man dasjenige μ , das

$$E(X) = \sum_{i=1}^n \Psi(|x - x_i|)$$

minimiert, den *M-Schätzer* von x_1, \dots, x_n .

73.7 Beispiele

- Beliebt ist die Familie $\Psi(s) = s^p$ mit $p \geq 0$.
Man kann zeigen:

i) $p = 2$ liefert das arithmetische Mittel \bar{x} . Es minimiert die quadratische Abstandssumme

$$E(X) = \sum_{i=1}^n (x_i - x)^2$$

ii) $p = 1$ liefert den Median $\hat{\mu}$. Er minimiert die Abstandssumme

$$E(X) = \sum_{i=1}^n |x_i - x|$$

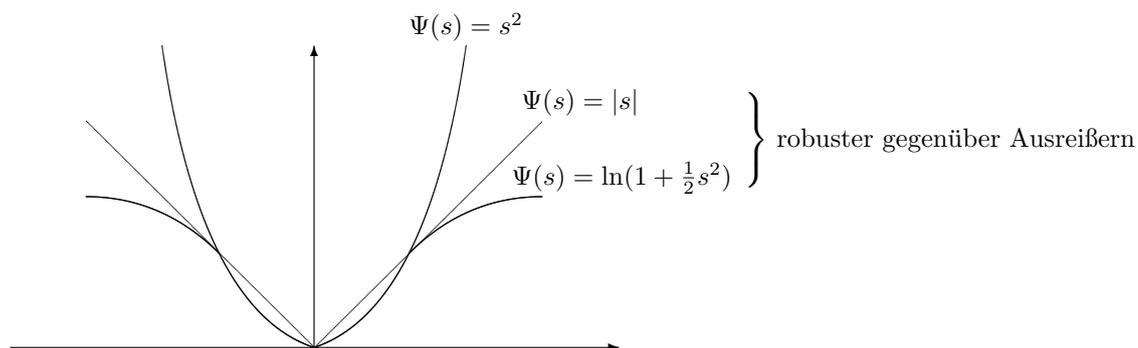
iii) $p \rightarrow 0$ liefert als Minimierer die *Modalwerte* (Maxima des Histogramms).

iv) $p \rightarrow \infty$ ergibt den *Midrange* $\frac{\max\{x_i\} + \min\{x_i\}}{2}$.

Kleinere Werte für p liefern robustere M-Schätzer, da Ausreißer x_i , für die $\Psi(|x - x_i|) = |x - x_i|^p$ groß wird, weniger stark eingehen.

b) Eine andere Straffunktion, die robuster als die übliche quadratische Straffunktion $\Psi(s) = s^2$ ist, ist z.B. die *Lorentz-Strafffunktion*

$$\Psi(s) = \ln\left(1 + \frac{1}{2}s^2\right)$$



74 Fehlerfortpflanzung

74.1 Motivation

Es werden zwei physikalische Größen x und y mehrmals gemessen. Man erhält als Mittelwert \bar{x}, \bar{y} und als (empirische) Standardabweichung s_x, s_y (vgl. 70.4). Nun will man hieraus eine neue Größe $z = f(x, y)$ berechnen. Als Schätzer für den Erwartungswert verwendet man $\bar{z} := f(\bar{x}, \bar{y})$. Wie gehen jedoch die Messfehler s_x, s_y in die Standardabweichung s_z ein? Man kann zeigen:

74.2 Satz (Fehlerfortpflanzungsgesetz von Gauß)

Schätzt man aus zwei fehlerbehafteten Größen x, y mit Mittelwert \bar{x}, \bar{y} und Standardabweichung s_x, s_y eine neue Größe $z = f(x, y)$, so ist der Schätzer für ihren Erwartungswert gegeben durch

$$\bar{z} = f(\bar{x}, \bar{y}).$$

Für die Varianz von z gilt:

$$s_z^2 = \left(\frac{\partial f}{\partial x} \right)^2 \Big|_{\bar{z}} s_x^2 + \left(\frac{\partial f}{\partial y} \right)^2 \Big|_{\bar{z}} s_y^2$$

Beweisidee: Taylorentwicklung.

74.3 Beispiel

Zwei Widerstände R_1, R_2 werden mehrmals gemessen:

$$\begin{aligned} \bar{R}_1 &= 100(\text{Ohm}) & s_{R_1} &= 0,8 & (\text{Schreibweise: } R_1 = 100 \pm 0,8) \\ \bar{R}_2 &= 200 & s_{R_2} &= 1 & (R_2 = 200 \pm 1) \end{aligned}$$

Wie groß sind Gesamt Widerstand und Fehler in Parallelschaltung?

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_2 + R_1}{R_1 R_2} \Rightarrow R = \frac{R_1 R_2}{R_1 + R_2}$$

$$\bar{R} = \frac{\bar{R}_1 \cdot \bar{R}_2}{\bar{R}_1 + \bar{R}_2} = \frac{100 \cdot 200}{100 + 200} \approx 66,67$$

$$\frac{\partial R}{\partial R_1} = \frac{(R_1 + R_2)R_2 - R_1 R_2}{(R_1 + R_2)^2} = \frac{R_2^2}{(R_1 + R_2)^2} \quad \frac{\partial R}{\partial R_1} \Big|_{\bar{R}} \approx 0,44$$

$$\frac{\partial R}{\partial R_2} = \frac{R_1^2}{(R_1 + R_2)^2} \quad \frac{\partial R}{\partial R_2} \Big|_{\bar{R}} \approx 0,11$$

$$s_R = \sqrt{\left(\frac{\partial R}{\partial R_1} \right)^2 \Big|_{\bar{R}} s_{R_1}^2 + \left(\frac{\partial R}{\partial R_2} \right)^2 \Big|_{\bar{R}} s_{R_2}^2}$$

$$= \sqrt{0,44^2 \cdot 0,8^2 + 0,11^3 \cdot 1^2} \approx 0,37$$

$$R = \bar{R} \pm s_R \approx 66,67 \pm 0,37$$

75 Markowketten

75.1 Motivation

Der Zustand eines Systems zur Zeit $n \in \mathbb{N}$ werde durch eine Zufallsvariable X_n beschrieben und soll nur von X_{n-1} abhängen (nicht jedoch von früheren Zuständen X_{n-2}, X_{n-3}, \dots). Wir möchten das zeitliche Verhalten dieses Systems studieren, insbesondere das Langzeitverhalten für $n \rightarrow \infty$. Prozesse dieser Art sind in der Informatik z.B. bei der Untersuchung der Auslastung von Servern wichtig (Warteschlangenmodelle).

75.2 Definition

Ein *stochastischer Prozess* ist eine Familie (X_t) von Zufallsvariablen mit $t \in \mathbb{R}$ oder $t \in \mathbb{N}$. Wir denken dabei an t als Zeitparameter, der kontinuierlich oder diskret ist.

Ein diskreter stochastischer Prozess $(X_n), n \in \mathbb{N}$ heißt *Markowkette*, wenn die Verteilung von X_n bei gegebenem X_{n-1} nicht von den früheren Verteilungen $X_k, k < n - 1$ abhängt:

$$P(X_n = i_n \mid X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \dots) = P(X_n = i_n \mid X_{n-1} = i_{n-1}).$$

Bemerkung: Wichtig sind insbesondere Markowketten, die nur endlich viele Zustände annehmen:

$$P(X_n \in \{1, \dots, k\}) = 1.$$

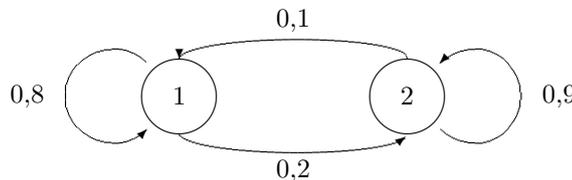
75.3 Beispiel mit Definitionen

Jedes Jahr ziehen 10 % der Bevölkerung außerhalb Kaliforniens nach Kalifornien und 20% der Bevölkerung Kaliforniens zieht aus.

Eine Person befindet sich im Jahr $n - 1$ in einem von zwei Zuständen:

$$X_{n-1} := \begin{cases} 1 & \text{(Person wohnt in Kalifornien)} \\ 2 & \text{(Person wohnt nicht in Kalifornien)} \end{cases}$$

Der Zustand im Jahr n lässt sich dann durch ein graphisches Modell mit Übergangswahrscheinlichkeiten beschreiben:



Sei p_{ij}^n die Wahrscheinlichkeit, dass ein Zustand j zur Zeit $n - 1$ in den Zustand i zur Zeit n übergeht (z.B. $p_{12}^n = 0,1$)

$$p_{ij}^n = P(X_n = i \mid X_{n-1} = j)$$

Wir definieren die *Matrix der Übergangswahrscheinlichkeiten* (*Übergangsmatrix*) durch

$$M_n := (p_{ij}^n) \in \mathbb{R}^{k \times k} \text{ (bei } k \text{ Zuständen)}$$

Im Beispiel:

$$M_n = \begin{pmatrix} 0,8 & 0,1 \\ 0,2 & 0,9 \end{pmatrix}$$

Die Verteilung von X_n auf die Zustände $1, \dots, k$ werde durch einen Vektor $u_n \in \mathbb{R}^k$ beschrieben:

$$u_n = \begin{pmatrix} u_{n_1} \\ \vdots \\ u_{n_k} \end{pmatrix} \in \mathbb{R}^k$$

Dann berechnet sich u_n aus u_{n-1} durch:

$$u_n = M_n u_{n-1}$$

Sind im Beispiel zur Zeit $n - 1$ 60% der Bevölkerung außerhalb Kaliforniens, gilt:

$$\begin{aligned} u_{n-1} &= \begin{pmatrix} 0,4 \\ 0,6 \end{pmatrix} \\ \Rightarrow u_n &= \begin{pmatrix} 0,8 & 0,1 \\ 0,2 & 0,9 \end{pmatrix} \begin{pmatrix} 0,4 \\ 0,6 \end{pmatrix} = \begin{pmatrix} 0,38 \\ 0,62 \end{pmatrix} \end{aligned}$$

Im Jahr n leben also 62% außerhalb Kaliforniens.

75.4 Definition

Eine Markowkette (X_n) heißt *homogen* oder *Kette mit stationären Übergangswahrscheinlichkeiten*, wenn die Übergangsmatrix M_n unabhängig von der Zeit n ist:

$$M_n = M \quad \forall n \in \mathbb{N}.$$

75.5 Bemerkungen

- Beispiel 75.3 beschreibt eine homogene Markowkette.
- Wir nennen eine Matrix $A = (a_{ij}) \in \mathbb{R}^{k \times k}$ eine *stochastische Matrix*, wenn alle Einträge nichtnegativ sind und alle Spaltensummen 1 sind:

$$a_{ij} \geq 0 \quad \forall i, j \in \{1, \dots, k\}$$

$$\sum_{i=1}^n a_{ij} = 1 \quad \forall j \in \{1, \dots, k\}$$

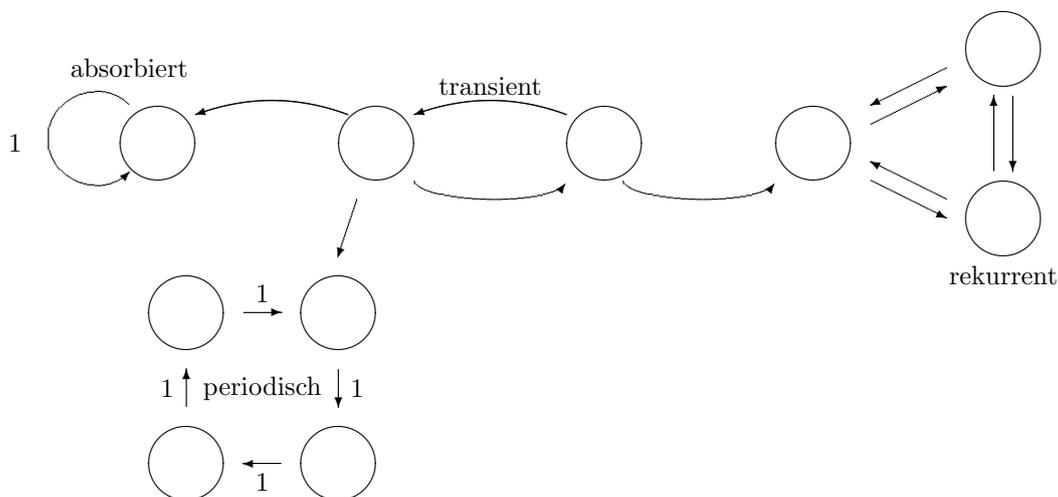
Übergangsmatrizen sind Beispiele für stochastische Matrizen.

- In der Stochastikliteratur werden oft *Zeilenvektoren* u_n betrachtet und man definiert p_{ij}^n als die Übergangswahrscheinlichkeit von Zustand i nach Zustand j . Dann ist

$$u_n = u_{n-1} M_n$$

und stochastische Matrizen haben *Zeilensumme* 1.

75.6 Zustandsbeschreibung endlicher homogener Markowketten



Definition: Sei (X_n) eine endliche homogene Markowkette. Ein Zustand i heißt

- transient*, wenn $P(X_m \neq i \forall m > n \mid X_n = i) > 0$
(kann verlassen werden).
- rekurrent*, wenn $P(X_m \neq i \forall m > n \mid X_n = i) = 0$
(mit Wahrscheinlichkeit 1 kehren wir zurück).
- periodisch* mit Periode l , wenn $P(X_{n+l} = i \mid X_n = i) = 1$.

Eine Menge I von Zuständen heißt *absorbierend*, wenn

$$P(X_{n+1} \in I \mid X_n \in I) = 1$$

Um das Zeitverhalten von Markowketten zu verstehen, müssen wir stochastische Matrizen näher untersuchen.

75.7 Satz (Eigenwerte stochastischer Matrizen)

Sei $M = (p_{ij}) \in \mathbb{R}^{k \times k}$ eine stochastische Matrix. Dann gilt:

- $\lambda = 1$ ist Eigenwert von M^\top .
- $|\lambda| \leq 1$ für alle Eigenwerte λ von M und M^\top .
- $\lambda = 1$ ist einziger Eigenwert von M^\top mit $|\lambda| = 1$, falls $\min_j p_{jj} > 0$.

Bemerkung: Aussage 75.7 (c) gilt auch für M statt M^\top .

Beweis:

- Ist M stochastisch, so hat $A = M^\top$ Zeilensumme 1.

$$\Rightarrow A \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \sum_j a_{1j} \\ \vdots \\ \sum_j a_{kj} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Somit ist $\lambda = 1$ Eigenwert von A zum Eigenvektor $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$.

b) Betrachte die Spaltensummennorm

$$\|M\|_S := \max_j \left(\sum_{i=1}^k |p_{ik}| \right) = 1.$$

Dann gilt nach Satz 50.8 für jeden Eigenwert λ von M :

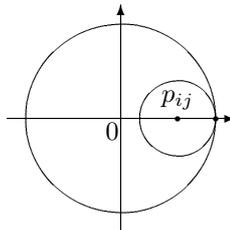
$$|\lambda| \leq \|M\|_S = 1.$$

Für M^T betrachtet man die Zeilensummennorm.

c) Nach dem Satz von Gerschgorin (50.10) gibt es zu jedem Eigenwert λ von $M^T = (a_{ij})$.

$$|\lambda - p_{ii}| < \sum_{\substack{j=1 \\ j \neq i}}^k |a_{ij}| = 1 - p_{ii}$$

Also liegt λ in dem Kreis mit Mittelpunkt p_{ii} und Radius $1 - p_{ii}$. Er berührt den Einheitskreis von innen in $(1, 0)$:



Aus $|\lambda| = 1$ folgt somit: $\lambda = 1$.

75.8 Bedeutung des Eigenwerts $\lambda = 1$

Wir interessieren uns für das Verhalten einer endlichen homogenen Markowkette (X_n) für $n \rightarrow \infty$. Für einen Anfangszustand u_0 und eine Übergangsmatrix $M = (p_{ij})$ gilt:

$$\begin{aligned} u_1 &= Mu_0 \\ u_2 &= Mu_1 = M^2u_0 \\ &\vdots \\ u_n &= M^n u_0 \end{aligned}$$

Ist u_0 Eigenvektor von M zum Eigenwert $\lambda = 1$ gilt:

$$u_n = M^n u_0 = \lambda^n u_0 = u_0 \quad \forall n \in \mathbb{N}$$

d.h. der Zustand u_0 ist stabil.

Darüber hinaus kann man Folgendes zeigen:

75.9 Satz (Potenzen stochastischer Matrizen)

Sei M eine stochastische Matrix. Genau dann existiert $\lim_{n \rightarrow \infty} M^n$, wenn 1 der einzige Eigenwert von M mit Betrag 1 ist.

Beweis einer Richtung: Wir zeigen:

Sei λ ein Eigenwert von M mit $|\lambda| = 1$. Falls $\lim_{n \rightarrow \infty} M^n$ existiert, ist 1 einziger Eigenwert von M mit Betrag 1.

Sei $v \neq 0$ ein Eigenvektor von M zum Eigenwert λ mit $|\lambda| = 1$:

$$Mv = \lambda v \Rightarrow M^n v = \lambda^n v \quad \forall n \in \mathbb{N}$$

Würde $\lim_{n \rightarrow \infty} M^n$ existieren, hätten wir

$$\left(\lim_{n \rightarrow \infty} M^n \right) v = \lim_{n \rightarrow \infty} (M^n v) = \lim_{n \rightarrow \infty} (\lambda^n v) = \underbrace{\left(\lim_{n \rightarrow \infty} \lambda^n \right)}_{=: \mu} v$$

Also würde auch $\mu := \lim_{n \rightarrow \infty} \lambda^n$ existieren. Dann wäre auch

$$\mu = \lim_{n \rightarrow \infty} \lambda^{n+1} = \lambda \lim_{n \rightarrow \infty} \lambda^n = \lambda \cdot \mu$$

Wegen $|\lambda| = 1$ ist natürlich auch $|\lambda^n| = 1$ und somit auch $|\mu| = 1 \neq 0$.
 Aus $\mu = \lambda\mu$ folgt damit $\lambda = 1$.

□

75.10 Gegenbeispiel

Wir betrachten eine stochastische Matrix, die Eigenwert λ mit $|\lambda| = 1$ besitzt, aber $\lambda \neq 1$ sein kann:

$$M = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 1 & & & 0 \end{pmatrix} \in \mathbb{R}^{k \times k}$$

beschreibt eine zyklische Vertauschung.

Sei $\alpha := e^{2\pi i/k} = \cos \frac{2\pi}{k} + i \sin \frac{2\pi}{k}$ ($\Rightarrow \alpha^k = e^{2\pi i} = 1$).

Setzen wir

$$v_j := \begin{pmatrix} 1 \\ \alpha^j \\ \alpha^{2j} \\ \vdots \\ \alpha^{(k-1)j} \end{pmatrix}, \quad 0 \leq j \leq k-1$$

so folgt wegen $\alpha^k = 1$:

$$Mv_j = \begin{pmatrix} \alpha^j \\ \alpha^{2j} \\ \vdots \\ \alpha^{kj} \end{pmatrix} = \alpha^j v_j$$

Also sind $1, \alpha^1, \alpha^2, \dots, \alpha^{k-1} \in \mathbb{C}$ sämtliche Eigenwerte von M . Alle haben Betrag 1.

Klar: M^n konvergiert nicht für $n \rightarrow \infty$.

75.11 Definition

Eine endliche homogene Markowkette (X_n) ist *im Gleichgewicht*, falls zu ihrer Übergangsmatrix $M = (p_{ij})$ eine Zustand u existiert mit

$$Mu = u,$$

d.h. u ist Eigenvektor von M zum Eigenwert 1.

Zusätzlich muss gelten:

$$\sum_{i=1}^k u_i = 1, \quad u_i \geq 0 \quad \forall i \in \{1, \dots, k\}.$$

75.12 Beispiel

Im Beispiel 75.3 war die Übergangsmatrix gegeben durch

$$M = \begin{pmatrix} 0,8 & 0,1 \\ 0,2 & 0,9 \end{pmatrix}$$

Berechnung der Eigenwerte:

$$\begin{aligned} 0 &\stackrel{!}{=} \begin{vmatrix} 0,8 - \lambda & 0,1 \\ 0,2 & 0,9 - \lambda \end{vmatrix} = (0,8 - \lambda)(0,9 - \lambda) - 0,02 \\ &= 0,72 - 0,8\lambda - 0,9\lambda + \lambda^2 - 0,02 \\ &= \lambda^2 - 1,7\lambda + 0,7 \\ \lambda_{1/2} &= \frac{1,7 \pm \sqrt{1,7^2 - 4 \cdot 0,7}}{2} = \frac{1,7 \pm \sqrt{2,89 - 2,8}}{2} = \frac{1,7 \pm 0,3}{2} \\ &\lambda_1 = 1, \quad \lambda_2 = 0,7 \end{aligned}$$

Eigenvektor zu $\lambda_1 = 1$:

$$\begin{pmatrix} -0,2 & 0,1 \\ 0,2 & -0,1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$0,2x = 0,1y \Rightarrow x = \frac{y}{2}$$
$$v = \begin{pmatrix} \alpha \\ 2\alpha \end{pmatrix}, \quad \alpha \neq 0$$

Da v eine Verteilung auf die Zustände beschreiben soll, muss gelten:

$$v_i \geq 0, \quad \sum_{i=1}^n v_i = 1$$

Zusätzlich muss gelten: $\alpha \geq 0, \alpha + 2\alpha = 1 \Rightarrow \alpha = \frac{1}{3}$.

$$\Rightarrow v = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}$$

Dies beschreibt den Gleichgewichtszustand.

$\frac{1}{3}$ der Personen wohnt in Kalifornien, $\frac{2}{3}$ außerhalb.

Wann ist es möglich, bei einer Markowkette von einem Zustand in einen beliebigen anderen zu gelangen?

75.13 Definition

Eine stochastische Matrix $M = (p_{ij}) \in \mathbb{R}^{k \times k}$ heißt *transitiv (irreduzibel)*, wenn man von jedem Zustand n mit positiver Wahrscheinlichkeit in jeden Zustand m in endlich vielen Schritten gelangen kann:

Es gibt ein r , so dass für $M^r =: B = (b_{ij})$ gilt:

$$b_{mn} > 0.$$

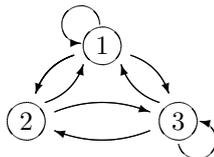
75.14 Praktisches Kriterium für Irreduzibilität

Man zeichnet zur Übergangsmatrix $M = (p_{ij})$ einen gerichteten Graphen: Ist $p_{ij} > 0$, zeichnet man einen Pfeil von j nach i . Falls man von jedem Zustand längs solcher Pfeile zu jedem anderen Zustand gelangt, ist M irreduzibel, andernfalls nicht.

75.15 Beispiele

a)

$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

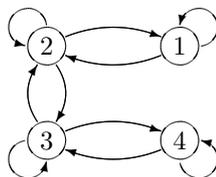


irreduzibel: Es gibt von jedem Punkt einen Weg zu jedem anderen Punkt.

Beachte: Ein Weg darf auch mehrere Pfeile umfassen. Daher führt ein Weg von 2 nach 2.

b)

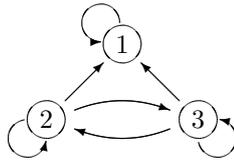
$$M = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$



irreduzibel.

c)

$$M = \begin{pmatrix} 1 & \frac{2}{5} & \frac{1}{4} \\ 0 & \frac{1}{5} & \frac{1}{4} \\ 0 & \frac{2}{5} & \frac{1}{2} \end{pmatrix}$$



reduzibel: Es führt z.B. kein Weg von 1 nach 2.

76 Verborgene Markowmodelle

76.1 Motivation

Verborgene Markowmodelle (hidden Markov models, HMMs) spielen eine große Rolle in der Bioinformatik, der Sprach- und der Mustererkennung. Mit Hilfe der aus dem Bereich Markowketten bekannten Übergangsmatrizen (Kapitel 75) sucht man nach der wahrscheinlichsten Zustandsfolge, die eine gegebene Beobachtung erzeugt haben könnte.

76.2 Beispiel: Spieler mit fairer und unfairer Münze

Ein Spieler besitzt eine faire Münze, bei der Kopf und Zahl gleichwahrscheinlich sind ($p^+(0) = p^+(1) = \frac{1}{2}$) und eine gezinkte Münze, bei der Zahl wahrscheinlicher ist ($p^-(0) = \frac{1}{4}$, $p^-(1) = \frac{3}{4}$). Wir beobachten n Münzwürfe und wollen entscheiden, ob er die faire oder die unfaire Münze genommen hat.

Sei k die Zahl der Ergebnisse, bei denen Zahl beobachtet wurde. Die einzelnen Ergebnisse seien x_1, \dots, x_n .

Fall 1: Münze war fair.

Wahrscheinlichkeit, dass die Beobachtung $x = (x_1, \dots, x_n)$ eintritt, ist

$$P(x \mid \text{faire Münze}) = \prod_{i=1}^n p^+(x_i) = \frac{1}{2^n}$$

Fall 2: Münze war unfair.

$$\underbrace{P(x \mid \text{unfaire Münze})}_{\prod_{i=1}^n p^-(x_i)} = \underbrace{\left(\frac{1}{4}\right)^{n-k}}_{n-k \text{ Mal Kopf}} \cdot \underbrace{\left(\frac{3}{4}\right)^k}_{k \text{ Mal Zahl}} = \frac{3^k}{4^n}$$

Wir vermuten, dass die Münze fair war, falls

$$\begin{aligned} P(x \mid \text{faire Münze}) &> P(x \mid \text{unfaire Münze}) \\ \frac{1}{2^n} &> \frac{3^k}{4^n} && | \cdot 4^n \\ 2^n &> 3^k && | \log_2 \\ n &> k \log_2(3) \\ \frac{k}{n} &< \frac{1}{\log_2(3)} && (*) \end{aligned}$$

76.3 Schwierigeres Beispiel

Wir nehmen an, dass der Spieler mit einer geringen Wahrscheinlichkeit von 0,1 die Münze innerhalb des Spiels austauscht. Wir wollen wissen, wann er welche Münze genommen hat.

Naiver Ansatz:

Wir betrachten die Beobachtungen innerhalb eines Fensters und überprüfen, ob (*) zutrifft oder nicht.

Problem:

Lösung hängt von Fenstergröße ab und wir wissen nicht, wann der Spieler die Münze wechselt.

Verborgene Markowmodelle bieten einen alternativen stochastischen Zugang, um dieses Problem zu lösen.

76.4 Definition

Ein *verborgenes Markowmodell* (*hidden Markov model, HMM*) ist ein Tupel $\mathcal{M} = (\Sigma, Q, M, E)$. Dabei bezeichnen:

- Σ : Alphabet von Symbolen (Annahme: $|\Sigma| = r$)
- Q : Menge von Zuständen mit Symbolen aus Σ (Annahme: $|Q| = k$)
- $M = (p_{ij})$: stochastische Matrix mit Übergangswahrscheinlichkeiten, $M \in \mathbb{R}^{k \times k}$
- $E = (e_i(b))$: Matrix der Emissionswahrscheinlichkeiten, $E \in \mathbb{R}^{k \times r}$

76.5 Beispiel

In Beispiel 76.3 bezeichnen:

- $\Sigma = \{0, 1\}$: Kopf (0) oder Zahl (1)
- $Q = \{f, b\}$: faire (f) oder unfaire (b) Münze
- $M = \begin{pmatrix} 0,9 & 0,1 \\ 0,1 & 0,9 \end{pmatrix}$: Übergangswahrscheinlichkeit für Münzwechsel
- $e_f(0) = \frac{1}{2}$, $e_f(1) = \frac{1}{2}$ Emissionswahrscheinlichkeit für faire Münze

$e_b(0) = \frac{1}{4}$, $e_b(1) = \frac{3}{4}$ Emissionswahrscheinlichkeit für unfaire Münze

Die vom Spieler tatsächlich verwendete Zustandsfolge ist $\pi = (\pi_1, \dots, \pi_n)$ mit $\pi_i \in Q$. Sie bleibt uns „verborgen“. Wir beobachten nur den Ergebnisvektor $x = (x_1, \dots, x_n)$ mit $x_i \in \Sigma$.

Die Wahrscheinlichkeit, dass die beobachtete Sequenz x durch einen Pfad π erzeugt wurde, ist

$$P(x | \pi) = \prod_{i=1}^n P(x_i | \pi_i) \underbrace{P(\pi_i | \pi_{i+1})}_{\text{Übergangsws.}} \quad \text{mit } P(\pi_n | \pi_{n+1}) := 1$$

76.6 Problemstellung

Gelöst werden soll nun das *Dekodierungsproblem*:

Finde zu einem gegebenem HMM $\mathcal{M} = (\Sigma, Q, M, E)$ und einer Beobachtung $x = (x_1, \dots, x_n)$ einen optimalen Pfad $\pi^* = (\pi_1^*, \dots, \pi_n^*)$, der $P(x | \pi)$ maximiert:

$$\pi^* := \operatorname{argmax}_{\pi} P(x | \pi)$$

76.7 Der Viterbi-Algorithmus (findet lokales Maximum)

Grundidee: optimaler Pfad $(\pi_1^*, \dots, \pi_{i+1}^*)$ zur Beobachtung (x_1, \dots, x_{i+1}) ist optimal unter allen Pfaden, die in den (unbekannten) Zustand $\pi_i^* = m \in Q$ enden.

$s_m(i)$: Wahrscheinlichkeit, dass der optimale Pfad $(\pi_1^*, \dots, \pi_i^*)$ zur Beobachtung (x_1, \dots, x_i) in m endet ($1 \leq i \leq n$).

$$\Rightarrow \quad s_l(i+1) = \underbrace{e_l(x_{i+1})}_{\text{Emissionsws. für } l} \cdot \underbrace{\max_{m \in Q} \{s_m(i) \cdot p_{lm}\}}_{\text{Übergang } m \rightarrow l} \quad (**)$$

Initialisiere:

$$\begin{aligned} s_0(0) &= 1 \\ s_m(0) &= 0 \quad m \in \{1, \dots, n\} \end{aligned}$$

Im optimalen π^* gilt:

$$P(x \mid \pi^*) = \max_{m \in Q} \{s_m(i) \cdot p_{nm}\}$$

Das Viterbi-Verfahren benötigt $\mathcal{O}(nk)$ Rechenoperationen.

77 Pseudozufallszahlen und Monte-Carlo-Simulation

77.1 Motivation

Es gibt Anwendungsgebiete (z.B. Ray Tracing in der Computergrafik, Strömungssimulation im Bereich Computational Fluid Dynamics, Berechnung hochdimensionaler Integrale), bei denen stochastische Simulationen einfache oder effiziente Alternativen zu deterministischen Algorithmen sind. Solche Simulationen nennt man auch *Monte-Carlo-Verfahren*. Sie benötigen die Erzeugung von Zufallszahlen auf dem Rechner. Da funktionierende Computer jedoch deterministisch arbeiten, verwendet man statt dessen Algorithmen, die Zahlen liefern, die echten Zufallszahlen ähneln. Solche Zahlen nennt man *Pseudozufallszahlen*.

77.2 Erzeugung von gleichverteilten Pseudozufallszahlen

Ziel: Erzeugung einer Sequenz zu Z_n von gleichverteilten Pseudozufallszahlen aus $[0, 1]$.

Beliebte Vorgehensweise: Lineare Kongruenzmethoden (in vielen Compilern verwendet).

Geg: $m \in \mathbb{N}$	Modus
$a \in \{1, \dots, m-1\}$	Multiplikator
$b \in \{0, \dots, m-1\}$	Inkrement
$x_0 \in \{1, \dots, m-1\}$	Startwert

Verfahren:

$$\begin{aligned}x_n &:= a \cdot x_{n-1} + b && (\text{modulo } m) \\Z_n &:= x_n/m\end{aligned}$$

Dann approximiert die Sequenz (Z_n) eine Folge von stochastisch unabhängigen Zufallsvariablen, die auf $[0, 1]$ gleichverteilt sind.

Die Approximationsgüte hängt von der Parameterzahl ab.

Klar: Nach spätestens m Schritten wiederholt sich die Folge.

Häufig verwendet, aber schlecht:

$$\begin{aligned}m &= 2^{16} = 65536 \\a &= 25173 \\b &= 13849\end{aligned}$$

Besser (Minimalstandard):

$$\begin{aligned}m &= 2^{31} - 1 = 2.147.483.647 \\a &= 7^5 = 16807 \\b &= 0\end{aligned}$$

77.3 Erzeugung von $N(0,1)$ -verteilten Pseudozufallszahlen

Die Standardnormalverteilung im \mathbb{R}^2 hat die Dichte

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

Für eine Kreisscheibe $K_r(0)$ um 0 mit Radius t und zwei stochastisch unabhängig $N(0,1)$ -verteilte Variablen X, Y gilt:

$$P(X^2 + Y^2 \leq t^2) = \int_{K_r(0)} \frac{1}{2\pi} \cdot e^{-\frac{x^2+y^2}{2}} dx dy$$

Geht man zu Polarkoordinaten (r, φ) über, ergibt sich mit

$$x = r \cdot \cos \varphi =: f_1(r, \varphi)$$

$$y = r \cdot \sin \varphi =: f_2(r, \varphi)$$

$$Jf(r, \varphi) = \begin{pmatrix} \frac{\partial f_1}{\partial r} & \frac{\partial f_1}{\partial \varphi} \\ \frac{\partial f_2}{\partial r} & \frac{\partial f_2}{\partial \varphi} \end{pmatrix} = \begin{pmatrix} \cos \varphi & -r \cdot \sin \varphi \\ \sin \varphi & r \cos \varphi \end{pmatrix}$$

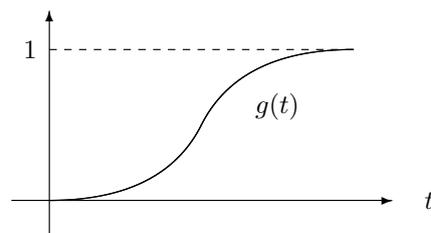
$$\det(Jf(r, \varphi)) = r \cdot \cos^2 \varphi + r \cdot \sin^2 \varphi = r$$

und der Transformationsregel (Kapitel 60):

$$\begin{aligned} P(X^2 + Y^2 \leq t^2) &= \int_{\varphi=0}^{2\pi} \int_{r=0}^t \frac{1}{2\pi} \cdot e^{-\frac{r^2}{2}} r dr d\varphi \\ &= 2\pi \cdot \int_{r=0}^t \frac{1}{2\pi} \cdot e^{-\frac{r^2}{2}} r dr \\ &= [e^{-\frac{r^2}{2}}]_0^t = 1 - e^{-\frac{t^2}{2}} \end{aligned}$$

Der Radius R des Zufallszahlenvektors (X, Y) erfüllt also

$$P(R \leq t) = 1 - e^{-\frac{t^2}{2}} =: g(t)$$



Wertebereich von $g : [0, 1)$.

Umkehrfunktion zu $g(t)$: $g^{-1}(z) = \sqrt{-2 \cdot \ln(1-z)}$

Hat man eine in $[0,1]$ gleichverteilte Zufallsvariable Z_1 , ergibt sich als Zufallsvariable R für den Radius der 2D-Standardnormalverteilung:

$$R = \sqrt{-2 \cdot \ln(1 - Z_1)}$$

Aus einer zweiten, auf $[0, 1]$ gleichverteilten Zufallsvariablen Z_2 erhält man als Zufallsvariable für einen auf $[0, 2\pi]$ gleichverteilten Winkel T :

$$T = 2\pi \cdot Z_2$$

Dies motiviert folgenden Algorithmus (*Box-Müller-Verfahren*) zur Erzeugung zweier $N(0,1)$ -verteilter Zufallszahlen X, Y aus zwei auf $[0, 1]$ gleichverteilten Zufallszahlen Z_1, Z_2 :

$$\begin{aligned}
R &:= \sqrt{-2 \cdot \ln(1 - Z_1)} \\
T &:= 2\pi \cdot Z_2 \\
X &:= R \cdot \cos T \\
Y &:= R \cdot \sin T
\end{aligned}$$

Bemerkung: Benötigt man $N(\mu, \sigma^2)$ -verteilte Zufallszahlen \tilde{X}, \tilde{Y} , setzt man (vgl. 65.15):

$$\begin{aligned}
\tilde{X} &= \mu + \sigma X \\
\tilde{Y} &= \mu + \sigma Y
\end{aligned}$$

(Pseudo-) Zufallszahlen benötigt man z.B. bei probabilistischen Algorithmen.

77.4 Beispiel: Quicksort

Sortieren einer Liste mit Quicksort:

- Suche ein zufälliges Element z der Liste (mit gleichverteilter Pseudozufallszahl aus 77.2)
- Sortiere Teilliste mit Elementen $\leq z_1$ mit Quicksort
- Sortiere Teilliste mit Elementen $> z_1$ mit Quicksort
- usw.

Quicksort hat eine mittlere Laufzeit von $\mathcal{O}(n \log n)$, ist also nicht schlecht.

77.5 Beispiel: Buffon'sches Nadelexperiment

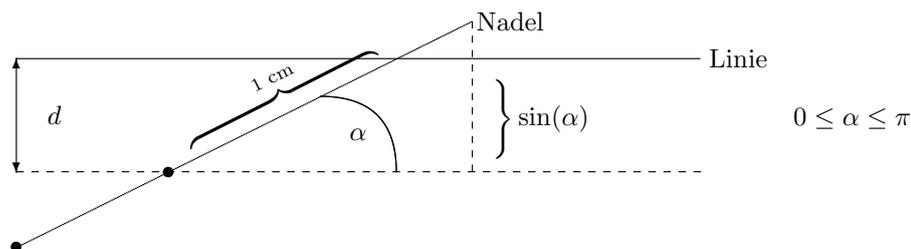
Ein probabilistischer Algorithmus zur Bestimmung von π :

- Zeichne auf einem Blatt Papier parallele Linien im Abstand einer Stecknadellänge.
- Lasse die Stecknadel auf das Papier fallen und überprüfe, ob sie eine der Linien trifft. Denke dabei fest an π .
- Zähle die Zahl N der Versuche und die Zahl T der Treffer. Dann gilt:

$$\frac{N}{T} \rightarrow \frac{\pi}{2} \text{ für } N \rightarrow \infty$$

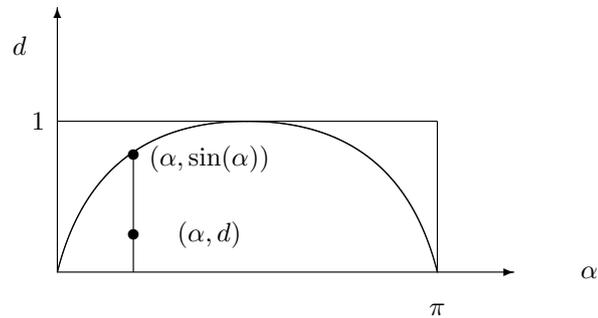
Warum funktioniert dies?

Annahme: Nadellänge und Linienabstand seien 2 [cm]. Nadel kann nur nächstgelegene Linie schneiden und nur dann ist ihr Abstand d zwischen Nadelmitte und Linie $d < 1$ erfüllt:



Nadel schneidet Linie $\Leftrightarrow d < \sin \alpha$

Zu jedem Wurf gehört ein Wertepaar $(\alpha, d) \in [0, \pi] \times [0, 1]$.
 $d < \sin \alpha \Leftrightarrow (\alpha, d)$ liegt unter Graphen von $\sin x$



Im Zufallsexperiment sind die Punkte (α, d) gleichverteilt auf $[0, \pi] \times [0, 1]$

$$\Rightarrow \frac{T}{N} = \frac{\int_0^\pi \sin x \, dx}{\pi \cdot 1} = \frac{[-\cos x]_0^\pi}{\pi} = \frac{2}{\pi}$$

Unser Experiment war also ein stochastisches Integrationsverfahren für $\int_0^\pi \sin x \, dx$. Im 1D Fall sind solche Verfahren ineffizient, aber bei hochdimensionalen Integrationsproblemen haben sie eine bessere Komplexität als deterministische numerische Ansätze.