WAHRSCHEINLICHKEITSRECHNUNG

WAHRSCHEINLICHKEITEN IN DER MEDIZIN 3
Aufgaben der Wahrscheinlichkeitsrechnung 3
Rechnen mit Wahrscheinlichkeiten 3
Zufallsexperimente und deren Beschreibung 3
Ermitteln einer Wahrscheinlichkeit 3
Verknüpfung zweier Ereignisse 4
Axiome von Kolmogoroff und deren Folgerungen 4
Abhängigkeit und bedingte Wahrscheinlichkeit 5
Bayes-Theorem 6
Rechenregeln für Wahrscheinlichkeiten 6
Wahrscheinlichkeiten in der Epidemiologie 6
Epidemiologie 6
Prävalenz, Punktprävalenz 6
Periodenprävalenz 7
Inzidenz 7
Krankheitsspezifische Mortalität 7
Letalität 7
Morbidität 7
Kontagionsindex 7
Manifestationsindex 7
Bevölkerungsstatistiken 7
Spezielle Wahrscheinlichkeiten 7
Sterbetafeln 8
Diagnostische Tests 8
Gütekriterien eines diagnostischen Tests 8
Vorhersagewerte 9
Kenngrößen diagnostischer Tests 9
DISKRETE VERTEILUNGEN 10
Diskrete Zufallsvariable 10
Bedeutung einer Zufallsvariablen 10
Wahrscheinlichkeiten 10
Lageparameter 10
Streuungsparameter 11
Binomialverteilung B(n,p) 12
Bernoulli-Experiment 12
Eigenschaften der Binomialverteilung 12
Symmetrische Binomialverteilung 13
Andere diskrete Verteilungen 14
Poissonverteilung $P(\lambda)$ 14
Polynomialverteilung $P(n_1,\ldots n_k|p_1,\ldots,p_k)$ 14
Negative Binomialverteilung NB(r,p) 15
Hypergeometrische Verteilung HG(n;N,M) 15
Diskrete Gleichverteilung DG(k) 16
Diskrete Verteilungen 17
STETIGE VERTEILUNGEN 18
Stetige Zufallsvariable 18
Funktionen 18
Lage- und Streuungsparameter 19
Zentrale Momente 19
Normalverteilung 19
Allgemeine Eigenschaften 19

Standardnormalverteilung N(0, 1) 20
σ -Bereiche und Referenzbereiche 21
Normalisierende Transformationen 21
Sätze der Wahrscheinlichkeitsrechnung 22
Tschebyscheff'sche Ungleichung 22
Gesetz der großen Zahlen 22
Zentraler Grenzwertsatz 23
Bedeutung der Normalverteilung 24
Verteilung von Überlebenszeiten 24
Wichtige Begriffe 24
Exponentialverteilung Exp(λ) 25
Weibull-Verteilung WB(λ, γ) 25
Prüfverteilungen 26
t-Verteilung 26
Chi²-Verteilung 27
F-Verteilung 27
Stetige Verteilungen 28

Wahrscheinlichkeiten in der Medizin

Aufgaben der Wahrscheinlichkeitsrechnung

- probabilistische Vorgänge: unterliegen dem Zufall
- deterministische Vorgänge: lassen sich exakt berechnen
- objektive Wahrscheinlichkeiten: Zahlenangaben
- subjektive Wahrscheinlichkeit: Vermutungen, die auf alltäglichen Erfahrungen basieren

Rechnen mit Wahrscheinlichkeiten

Zufallsexperimente und deren Beschreibung

- Zufallsexperimente: probabilistische Vorgänge mehrmals wiederholen, die Ergebnisse der einzelnen Experimente dokumentieren und auswerten
- Eigenschaften der Zufallsexperimente
  - nach bestimmter Vorschrift durchgeführt
  - beliebig oft wiederholbar
  - mehrere Ausgänge oder Ergebnisse sind möglich
  - Ergebnis eines einzelnen Experiments ist vorab ungewiss
- Beispiele: Würfeln (6 mögliche Ausgänge), Werfen einer Münze (2 mögliche Ausgänge)
- <u>Ereignisraum Ω:</u> Menge aller möglichen Ergebnisse
- <u>Ereignisse:</u> Teilmengen von Ω
- <u>Elementarereignisse:</u> 1-elementige Teilmengen
- sicheres Ereignis: Ereignisraum Ω
- unmögliches Ereignis: leere Menge Ø

<u>Beispiel „Würfeln"</u>
Ω = {1,2,3,4,5,6}
Ereignis „gerade Zahl" → Teilmenge A={2,4,6}
Elementarereignis von A = 2

Unterschied deskriptive Statistik und Wahrscheinlichkeitsrechnung
- deskriptive Statistik: befasst sich mit Stichproben und Merkmalen
- Wahrscheinlichkeitsrechnung: untersucht die mathematisch-theoretischen Eigenschaften von Grundgesamtheiten

Ermitteln einer Wahrscheinlichkeit

Theoretische Herleitung
<u>Pierre Simon Marquis de Laplace</u>
- befasste sich mit den Zufallsgesetzen bei Glücksspielen
- definierte basierend auf dem Begriff der Zufallsexperimente die Wahrscheinlichkeit, dass ein bestimmtes Ereignis A eintritt

$$P(A) = \frac{\text{Anzahl der günstigen Ereignisse}}{\text{Anzahl der möglichen Ereignisse}}$$

$$P(A) = \frac{\text{Anzahl der Elemente von A}}{\text{Anzahl der Elemente von } \Omega}$$

- Laplace'sche Definition: ordnet jedem Ereignis eine Zahl zwischen 0 und 1 zu
- Wahrscheinlichkeit eines Ereignisses ist vergleichbar mit der relativen Häufigkeit einer Merkmalsausprägung
- Voraussetzung: alle Elementarereignisse treten mit gleicher Wahrscheinlichkeit ein
  - z.B. Würfeln, jeder Augenzahl kann die Wahrscheinlichkeit 1/6 zugeordnet werden

Beispiel „Würfeln"
Wahrscheinlichkeit eine gerade Zahl zu würfeln:
6 mögliche Ereignisse, 3 „günstige" Ereignisse (2,4,6)
P(A)=3/6=1/2
für das unmögliche Ereignis „7" ergibt sich die leere Menge P(Ø)=0 (günstige Ereignisse=0)
für das sichere Ereignis eine Augenzahl zwischen 1 und 6 zu würfeln erhält man P(Ω)=1

Empirische Herleitung
- Untersuchung einer hinreichend großen Stichprobe bezüglich eines Merkmals
- Wert der relativen Häufigkeit einer Merkmalsausprägung wird dann als Näherungswert für die entsprechende Wahrscheinlichkeit zugrunde gelegt

Beispiel
100 Studenten
40 Frauen → P(w)=40%, 60 Männer → P(m)=60%

Computersimulation
Problem im Computer simulieren und mit Hilfe dieses Modells die Wahrscheinlichkeit für das Auftreten eines bestimmten Ereignisses ermitteln

Verknüpfung zweier Ereignisse

Verbindungen zwischen zwei oder mehreren Ereignissen lassen sich durch mengentheoretische Operationen beschreiben.

Vereinigungsmenge $A \cup B$	Ergebnis, dass A allein oder B allein oder beide Ereignisse gemeinsam eintreten (A vereinigt B)
Schnittmenge $A \cap B$	Ereignis, dass A und B gemeinsam eintreten (A Schnitt B)
Differenzmenge A - B	Ereignis, dass A aber nicht B eintritt (A minus B)

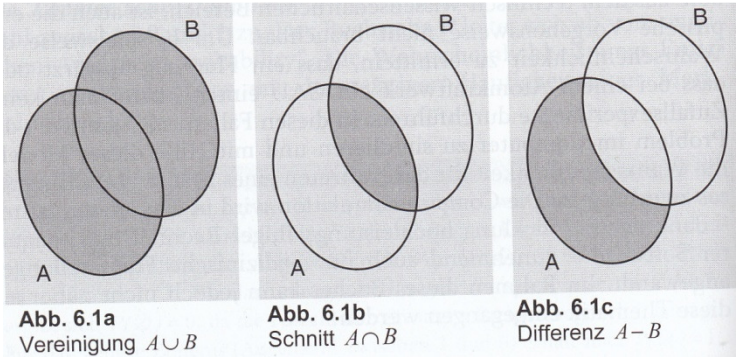


Abb. 6.1a Vereinigung $A \cup B$ **Abb. 6.1b** Schnitt $A \cap B$ **Abb. 6.1c** Differenz $A - B$

- disjunkt/unvereinbar: zwei Ereignisse A und B, deren Durchschnitt die leere menge bilden
  - z.B. männlichen Geschlecht und schwanger
- komplementär: zwei disjunkte Ereignisse, die sich zum Ereignisraum Ω ergänzen
  - $A \cup \bar{A} = \Omega$ (Ereignisse ergänzen sich)
  - $A \cap \bar{A} = \emptyset$ (Ereignisse sind disjunkt)

Axiome von Kolmogoroff und deren Folgerungen

Funktion P(A), die einem Ereignis A eine reelle Zahl zuordnet, heißt Wahrscheinlichkeit, falls folgende Axiome (einfache mathematische Aussagen, die nicht beweisbar sind) erfüllt sind

1. $0 \leq P(A) \leq 1$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ für disjunkte Ereignisse A und B

- Laplace
  - alle Elementarereignisse müssen mit gleicher Wahrscheinlichkeit eintreten
- Kolmogoroff
  - Wahrscheinlichkeit jedes Elementarereignisses muss einen Zahl zwischen 0 und 1 sein
  - deren Summe muss 1 ergeben
  - → diese Regeln gelten ebenso für relative Häufigkeiten

Rechenregeln, die sich aus den Axiomen herleiten lassen:
Wahrscheinlichkeit für das komplementäre Ereignis
Aus P(A) ergibt sich die Wahrscheinlichkeit für das Ereignis $\bar{A}$

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

unmögliches Ereignis:
P(∅)=0

Satz von der totalen Wahrscheinlichkeit
Ereignis A tritt entweder zusammen mit dem Ereignis B oder $\bar{B}$ auf

$$P(A) = P(A \cap B) + P(A \cap \bar{B})$$

Beispiel: eine Person mit Blutgruppe a kann entweder „Rhesusfaktor positiv" oder „Rhesusfaktor negativ" haben.

$A \cap \bar{B}$ ist identisch mit der Differenzmenge A-B

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

Additionssatz
Vereinigung zweier Ereignisse A und B

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

Ereignisse A und B sind disjunkt (A ∩ B=∅), dann ergibt sich folgende Form für den Additionssatz

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

Abhängigkeit und bedingte Wahrscheinlichkeit

- oft ist es nicht zweckmäßig Wahrscheinlichkeiten anzugeben, die sich auf die Grundgesamtheit beziehen
- viele Krankheiten stehen im Zusammenhang mit dem Geschlecht oder sind von bestimmten Risiken abhängig
- → bedingte Wahrscheinlichkeit: Wahrscheinlichkeiten für bestimmte Teilmengen der Grundgesamtheit getrennt berechnen

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

Diese Formel quantifiziert die Wahrscheinlichkeit für das Eintreten des Ereignisses A eingeschränkt auf die Menge, die dem Ereignis B entspricht.

Beispiel: Wahrscheinlichkeit, für einen Mann an Diabetes mellitus zu erkranken: P(D|M)=0,07

Multiplikationssatz
Wahrscheinlichkeit, dass zwei Ereignisse A und B gemeinsam eintreten

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

A und B unabhängig: Eintreten von B hat keinerlei Einfluss auf das Eintreten von A → P(A|B)=P(A)

Multiplikationssatz und Additionssatz für unabhängige Ereignisse

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

Bayes-Theorem

Berechnung der bedingten Wahrscheinlichkeit P(A|B), wenn folgendes bekannt ist

- P(A)
- P(B|A)
- P(B|$\bar{A}$)

$$P(A|B) = \frac{P(A) * P(B|A)}{P(A) * P(B|A) + P(\bar{A}) * P(B|\bar{A})}$$

Bayes-Theorem ermöglicht Rückschlüsse von der a-priori-Wahrscheinlichkeit P(A) auf die a-posteriori-Wahrscheinlichkeit P(A|B).

Beispiel
Ereignis A: Vorliegen einer bestimmten Krankheit
Ereignis B: Testergebnis positiv
Wahrscheinlichkeit P(A|B): Patient, der einen positiven Befund hat und tatsächlich erkrankt ist

Rechenregeln für Wahrscheinlichkeiten

Name des Satzes	Rechenregeln
Satz für das komplementäre Ereignis $\bar{A}$	$P(\bar{A}) = 1 - P(A)$
Satz von der totalen Wahrscheinlichkeit	$P(A) = P(A \cap B) + P(A \cap \bar{B})$
Additionssatz	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ A und B disjunkt → $P(A \cup B) = P(A) + P(B)$ A und B unabhängig → $P(A \cup B) = P(A) + P(B) - P(A) * P(B)$
Multiplikationssatz	$P(A \cap B) = P(A\|B) * P(B)$ A und B disjunkt → $P(A \cap B) = 0$ A und B unabhängig → $P(A \cap B) = P(A) * P(B)$

Wahrscheinlichkeiten in der Epidemiologie

Epidemiologie

… Befassung mit dem Auftreten von Krankheiten in einer größeren Population

Ziele

- Erkennen von Ursachen und Risikofaktoren von Krankheiten
- Bestimmen deren Verbreitung in der Bevölkerung
- Untersuchung des natürlichen Verlaufs und relevanter prognostischer Faktoren
- Evaluierung präventiver und therapeutischer Maßnahmen
- Schaffen von Grundlagen für politische Entscheidungen

Prävalenz, Punktprävalenz

- relativer Krankenbestand zu einem bestimmten Zeitpunkt
- Wahrscheinlichkeit P(K_t) für eine beliebige Person aus der Population zum Zeitpunkt t erkrankt zu sein
- Beispiele: Krankheit mit kurzer Dauer, chronische Krankheit, wiederholt auftretende Krankheit
- Querschnittsstudie

Periodenprävalenz

- relativer Krankenbestand in einem längeren Zeitraum
- Berücksichtigung aller Personen, die zu Beginn, während oder am Ende des Beobachtungszeitraums erkrankt waren
- Kohortenstudie

Lebenszeitprävalenz

- Wahrscheinlichkeit einer Person, krank geboren zu werden oder einmal im Laufe des Lebens zu erkranken

Inzidenz

- Neuerkrankungsrate
- Wahrscheinlichkeit P(K) für eine beliebige Person, während einer Beobachtungszeit zu erkranken
- Untersuchung einer Population, deren Mitglieder zu Beginn des Beobachtungszeitraums *nicht* erkrankt sind
- bei länger andauernden Krankheiten, kann die Prävalenz aus der Inzidenz berechnet werden:

Prävalenz = Inzidenz * durchschnittliche Dauer

- Inzidenz: wie groß ist das Erkrankungsrisiko für eine einzelne Person
- Prävalenz: informiert über die Auswirkungen einer Krankheit auf die Gesamtpopulation

Krankheitsspezifische Mortalität

- Todesrate
- Wahrscheinlichkeit P(K ∩ T), während der Beobachtungszeit an der Krankheit K zu erkranken und daran zu sterben

Letalität

- Tödlichkeitsrate der Erkrankten
- bedingte Wahrscheinlichkeit P(T|K)
- Beginn und Ende der Erkrankungen müssen innerhalb des Beobachtungszeitraums liegen

Mortalität = Inzidenz * Letalität

$$P(K \cap T) = P(K) * P(T|K)$$

Morbidität

- synonym für Prävalenz, teilweise synonym für Inzidenz

Kontagionsindex

- Maß für die Ansteckungsfähigkeit
- Index, der die Wahrscheinlichkeit angibt, dass sich eine nicht immune Person, die mit dem Erreger in Kontakt kommt, infiziert

Manifestationsindex

- Wahrscheinlichkeit, mit der eine infizierte Person manifest erkrankt (Krankheitsbereitschaft)
- je kleiner der Index, desto mehr Infektionsfälle verlaufen klinisch stumm

Bevölkerungsstatistiken

Spezielle Wahrscheinlichkeiten

Natalität

- Geburtenrate, Geburtenziffer
- Anteil lebend geborener Kinder im Verhältnis zur Gesamtpopulation während eines Beobachtungszeitraums

Fertilitätsziffer

- Fruchtbarkeitsrate, Fruchtbarkeitsziffer
- Wahrscheinlichkeit, das eine Frau im gebärfähigen Alter ein lebendes Kind zur Welt bringt (bezogen auf ein Jahr)

Pearl-Index
- Risikomaß bezüglich der Sicherheit einer Verhütungsmethode
- wie viele von 100 Frauen in einem Jahr ungewollt schwanger werden (wobei davon ausgegangen wird, dass eine nicht-schwangere Frau zwölf Zyklen pro Jahr hat)

Sterbeziffer
- Gesamtmortalität
- Anteil der im Beobachtungszeitraum Verstorbenen

Sterbetafeln

Sterbetafel: Verteilung von Lebensdauern

l_0 … Anzahl von Lebendgeborenen innerhalb eines Beobachtungszeitraums (z.B. ein Jahr)
l_x … Anzahl der Personen, die ihren x-ten Geburtstag erleben und danach noch unbestimmte Zeit leben
$\mathbf{d_x = l_x - l_{x+1}}$ … Anzahl der Lebendgeborenen, die zwischen ihrem x-ten und (x+1)-ten Geburtstag sterben

Lebensdauer wird als diskretes Merkmal aufgefasst mit den Ausprägungen x (Anzahl der erreichten Lebensjahre) und den absoluten Häufigkeiten d_x.

Sterbeziffern sind die altersspezifischen Mortalitätsraten.

$\mathbf{q_x = d_x / l_x}$ $(x = 0, \dots, \omega)$

qx … Wahrscheinlichkeit, dass jemand der seinen x-ten Geburtstag erlebt hat, vor seinem (x+1)-ten Geburtstag stirbt
ω … das letzte in der Sterbetafel berücksichtigte Alter
$l_{\omega+1} = 0$ (oft wird ω = 100 gesetzt)

durchschnittliche Lebenszeit (Lebenserwartung) eines Neugeborenen

$$e_0 = \frac{1}{2} + \frac{1}{l_0}\sum_{x=1}^{\omega} l_x$$

Lebenserwartung eines x-jährigen

$$e_x = \frac{1}{2} + 1/l_x \sum_{y=x+1}^{\omega} l_y$$

Verteilungsfunktion F(x) gibt den relativen Anteil der Lebendgeborenen an, deren Sterbealter kleiner als x ist:

$$F(x) = 1 - \frac{l_x}{l_0} \quad \text{für } 0 \le x \le \omega$$

Diagnostische Tests

Gütekriterien eines diagnostischen Tests

Diagnostischer Test
- größere Sicherheit erlangen über den Krankheitsstatus eines Patienten (z.B. HIV-Test)
- zwei Testergebnisse sind möglich
  - positives Ergebnis bei einer erkrankten Person
  - negatives Ergebnis bei einer nicht-erkrankten Person

T_+ … Ereignis, dass das Testergebnis positiv ist
T_- … Ereignis, dass das Testergebnis negativ ist
K … zu untersuchende Krankheit liegt vor
$\bar{K}$ … zu untersuchende Krankheit liegt nicht vor

<u>Sensitivität</u>
… bedingte Wahrscheinlichkeit P (T_+|K), dass der Test bei einer kranken Person richtig (positiv) ist

<u>Spezifität</u>
… bedingte Wahrscheinlichkeit P(T_-|$\bar{K}$), dass eine nicht-erkrankte Person ein richtiges (falsches) Testergebnis erhält

<u>falsch negatives Ergebnis</u>
… Test übersieht die Krankheit eines Patienten
… ergibt sich aus der Sensitivität (T_- und T_+ → komplementäre Ereignisse)

$$P(T_-|K) = 1 - P(T_+|K)$$

<u>falsch positives Ergebnis</u>
… ergibt sich aus der Spezifität

$$P(T_+|\bar{K}) = 1 - P(T_-|\bar{K})$$

Vorhersagewerte
(prädiktiven Werte)

- Wahrscheinlichkeiten, dass das Testergebnis den richtigen Krankheitsstatus anzeigt
- Vorhersagewerte sind abhängig von der Prävalenz
- Prävalenz P(K)
  - *„a-priori-Wahrscheinlichkeit"*
  - Wahrscheinlichkeit, erkrankt zu sein, <u>bevor</u> das Testergebnis bekannt ist
- positiver Vorhersagewert
  - *„a-posteriori-Wahrscheinlichkeit"*
  - bedingte Wahrscheinlichkeit P(K|T_+)
  - Wahrscheinlichkeit erkrankt zu sein, <u>nachdem</u> das positive Ergebnis vorliegt
  - kann bei geringer Prävalenz sehr niedrig sein, auch wenn Sensitivität und Spezifität sehr hoch sind
- negativer Vorhersagewert
  - bedingte Wahrscheinlichkeit P($\bar{K}$|T_-)

$$P(K|T_+) = \frac{P(K) * P(T_+|K)}{P(K) * P(T_+|K) + P(\bar{K}) * P(T_+|\bar{K})}$$

$$P(\bar{K}|T_-) = \frac{P(\bar{K}) * P(T_-|\bar{K})}{P(\bar{K}) * P(T_-|\bar{K}) + P(K) * P(T_-|K)}$$

Kenngrößen diagnostischer Tests

Ereignis	Bezeichnung der Wahrscheinlichkeit	formelle Schreibweise
Krankheit liegt vor	Prävalenz (a-priori-Wahrscheinlichkeit)	P(K)
Testergebnis richtig positiv	Sensitivität	P(T_+|K)
Testergebnis falsch negativ		P(T_-|K)
Testergebnis richtig negativ	Spezifität	P(T_-|$\bar{K}$)
Testergebnis falsch positiv		P(T_+|$\bar{K}$)
Krankheit liegt vor, falls Testergebnis positiv	positiver Vorhersagewert (a-posteriori-Wahrscheinlichkeit)	P(K|T_+)
Krankheit liegt nicht vor, falls Testergebnis negativ	negativer Vorhersagewert	P($\bar{K}$|T_-)

DISKRETE VERTEILUNGEN

Diskrete Zufallsvariable

Bedeutung einer Zufallsvariablen

Deskriptive Statistik

- Beschreibung einer Stichprobe beruht
  - auf den Häufigkeiten der Merkmalsausprägungen und
  - auf statistischen Kenngrößen (z.B. Mittelwert, Standardabweichung)

Wahrscheinlichkeitsrechnung

Zufallsvariable

- = Merkmal
- Funktion, die jedem möglichen Ergebnis eines Zufallsexperiments eine reelle Zahl zuordnet
  - Zahlenwerte entsprechen den Merkmalsausprägungen
- Realisationen (Realisierungen): x_i der Zufallsvariablen X
  - quantitative Merkmale: x_i … Mess- oder Zählwerte
  - qualitative Merkmale: x_i … numerische Codierungen der einzelnen Ausprägungen
- Zufallsvariable kann einem bestimmten Skalenniveau zugeordnet werden
- diskrete und stetige Zufallsvariablen

Wahrscheinlichkeiten

- diskrete Zufallsvariable – bei der Beobachtung von Zufallsexperimenten, bei denen abzählbar viele Ergebnisse möglich sind
  - z.B. Münzwurf, Blutgruppe
- Elementarereignis A – Zufallsvariable X nimmt den Wert xi an

$P(A) = P(X=x_i)=P(x_i)=p_i$

Wahrscheinlichkeiten aller Elementarereignisse summieren sich:
→ (wie relative Häufigkeiten)
→ 2. Axiom von Kolmogoroff (Wahrscheinlichkeit des Ereignisraums gleich 1)

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} f(x_i) = 1$$

Wahrscheinlichkeitsfunktion f(x) ordnet jedem Wert x_i dessen Wahrscheinlichkeit p_i zu:

$$f(x) = \begin{cases} p_i & f\ddot{u}r\ x = x_i\ \ (i = 1, \dots, k) \\ 0 & sonst \end{cases}$$

Verteilungsfunktion $F(x)=P(X \leq x)$ einer diskreten Zufallsvariablen
→ muss mindestens ordinal skaliert sein
→ Wahrscheinlichkeit, dass die Zufallsvariable X einen Wert annimmt, der kleiner oder gleich x ist
→ Funktionswerte F(x) durch Aufaddieren der Wahrscheinlichkeiten p_i

Lageparameter

Erwartungswert μ

- = Mittelwert
- Charakterisierung der Grundgesamtheit
- Erwartungswert von X: E(X), EX, μ_x

diskrete Zufallsvariable mit k möglichen Realisationen:

$$\mu = \sum_{i=1}^{k} x_i * p_i$$

Additivität der Erwartungswerte:

$$E(aX + b) = a * EX + b$$
$$E(X_1 + \ldots + X_n) = \sum_{i=1}^{n} EX_i$$

Median und Quantile

- Median $\tilde{\mu}$ einer Grundgesamtheit
  - bestimmt durch die Verteilungsfunktion
  - bei diskreten Zufallsvariablen: Median ist die kleinste Zahl zwischen 0 und k
  - $F(\tilde{\mu}) \geq 0{,}5$
- α-Quantil $\tilde{\mu}_\alpha$
  - $0 < \alpha < 1$
  - kleinste Zahl mit $F(\tilde{\mu}_\alpha) \geq \alpha$

Modus

- Modus der Grundgesamtheit
  - Wert mit der größten Wahrscheinlichkeit
  - bei bi- oder multimodalen Verteilungen existieren eventuell mehrere Modalwerte

deskriptive Statistik	Wahrscheinlichkeitsrechnung
Merkmal	Zufallsvariable X
Ausprägungsliste	Ereignisraum Ω
Merkmalsausprägung	Elementarereignis A
ermittelter Merkmalswert der Beobachtungseinheit i	Realisation x_i der Zufallsvariablen
relative Häufigkeit h_i	Wahrscheinlichkeit p_i
empirische Verteilungsfunktion ^F(x)	Verteilungsfunktion F(x)
Mittelwert $\bar{x}$	Erwartungswert μ

Streuungsparameter

Varianz

<u>deskriptive Statistik</u>

empirische Varianz: mittlere quadratische Abweichung der Stichproben-Daten vom Mittelwert

<u>Wahrscheinlichkeitsrechnung</u>

Erwartungswert der quadratischen Abweichung der Zufallsvariablen X vom Erwartungswert μ

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

Varianz für diskrete Zufallsvariablen:

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 p_i$$

$Var(aX+b) = a^2 * Var(X)$

für a=0: Var(b)=0

→ eine Konstante hat keine Varianz

Standardabweichung σ

… Wurzel der Varianz

Variationskoeffizient

- für verhältnisskalierte Zufallsvariablen
- Quotient σ / μ

Summe zweier Zufallsvariablen:
Var(X+Y) = Var(X) + Var(Y) + 2 * Cov(X, Y)

Kovarianz
Kovarianz = 0, wenn X und Y unabhängige Variablen

Cov(X, Y) = E((X - μ_x) * (Y - μ_y)) = E(XY) - μ_x * μ_y

Summe mehrerer unabhängiger Zufallsvariablen:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i)$$

Binomialverteilung B(n,p)

…Zufallsexperiment, bei dem nur zwei Ausgänge möglich sind

Bernoulli-Experiment

- Bernoulli-Experiment – alle Experimente, bei denen ein Alternativmerkmal beobachtet wird
  - z.B. Münzwurf, Geschlecht einer Person
- = Binomialverteilung
- anwendbar bei allen qualitativen und quantitativen Merkmalen deren Ausprägungen in zwei Gruppen oder Klassen eingeteilt sind
  - z.B. Blutgruppe A oder andere Blutgruppe

- zwei komplementäre Ereignisse A und $\bar{A}$
- Zufallsvariable X, welche die Werte 1 (A tritt ein) und 0 ($\bar{A}$ tritt ein)
- dazugehörigen Wahrscheinlichkeiten

$$P(A) = P(X = 1) = p$$
$$P(\bar{A}) = P(X = 0) = q$$

- Wahrscheinlichkeit des komplementären Ereignisses $\bar{A}$:

$$q = 1 - p$$

- Wahrscheinlichkeit p kann empirisch geschätzt werden
  - Bernoulli-Experiment hinreichend oft wiederholen
  - relative Häufigkeit des Ereignisses A als Schätzwert für p verwenden

Eigenschaften der Binomialverteilung

- *Bernoulli-Prozess* – Beobachtungsserien: Bernoulli-Experiment wird mehrfach wiederholt, Wiederholungen sind unabhängig voneinander
  - z.B. Wurfserie – Anzahl der 6er beim Würfeln wird gezählt

Eigenschaften des Prozesses
- Durchführung **n** unabhängiger Bernoulli-Experimente, die durch gleich verteilte Zufallsvariable X_i (i=1,…,n) beschrieben werden
- jedes X_i nimmt mit
  - Wahrscheinlichkeit p den Wert 1 (A tritt ein),
  - Wahrscheinlichkeit q=1-p den Wert 0 ($\bar{A}$ tritt ein) an
- Zufallsvariable $X=X_1+X_2+\ldots+X_n$ quantifiziert wie häufig bei n Experimenten das Ereignis A eintritt
- X wird durch Binomialverteilung beschrieben

Binomialverteilte Zufallsvariable ***X ~ B(n, p)***
- ist durch die Parameter n und p eindeutig festgelegt
- kann jede Zahl zwischen 0 und n annehmen
- Zahl gibt an, wie oft bei n Zufallsexperimenten das Ereignis A eingetreten ist

Erwartungswert

$$E(X) = \sum_{i=1}^{n} EX_i = n * p$$

Varianz

$$Var(X) = \sum_{i=1}^{n} Var(X_i) = n * p * q$$

Wahrscheinlichkeit, wie oft Ereignis A eingetreten ist

$$P(X = k) = \binom{n}{k} * p^k * q^{n-k} \quad für\ k = 0, \dots, n$$

$\binom{n}{k}$ … Binomialkoeffizient - quantifiziert die Anzahl der Möglichkeiten, aus einer Menge von n Elementen genau k Elemente auszuwählen

$$\binom{n}{k} = \frac{n!}{k! * (n-k)!} = \frac{1 * 2 * \dots * n}{(1 * \dots * k) * (1 * \dots * (n-k))}$$

- jeder Binomialkoeffizient ist eine natürliche Zahl
- für alle p gilt: $p^0=1$ $p^1=p$
- für alle k=0,…,n gilt: $\binom{n}{k} = \binom{n}{n-k}$
- $\binom{n}{0} = \binom{n}{n} = 1$

Schiefe einer Binomialverteilung

$$\gamma_1 = \frac{q-p}{\sigma}$$

→ Verteilung ist symmetrisch, wenn p=q

<u>Beispiel</u>
Wahrscheinlichkeit, dass 3 von 10 Antworten in der Klausur richtig sind.
$p^3 * q^7 = 0{,}2^3 * 0{,}8^7 \approx 0{,}0017$
$\binom{10}{3} = 120\ Möglichkeiten$
$P(X = 3) = 120 * 0{,}2^3 * 0{,}8^7 = 0{,}2013$

k	P(X=k)	F(k)=P(X ≤ k)
0	$1*0{,}2^0*0{,}8^{10}=0{,}8^{10}=0{,}1074$	0,1074
1	$10*0{,}2^1*0{,}8^9=0{,}2684$	0,3758 (0,1074+0,2684)
2	$45*0{,}2^2*0{,}8^8=0{,}3020$	0,6778 (0,3758+0,3020)
3	$120*0{,}2^3*0{,}8^7=0{,}2013$	0,8791 (0,6778+0,2013)
…	…	…
10	$1*0{,}2^{10}*0{,}8^0=0{,}210=10^{-7}$	1 ($0{,}9999999+10^{-7}$)

P(X ≤ 3) … 87,91%-ige Wahrscheinlichkeit, weniger als 4 Punkte zu erreichen

Symmetrische Binomialverteilung
… symmetrische Binomialverteilung (p=q=0,5)

Erwartungswert

$$E(X) = 0{,}5 * n$$

Varianz

$$Var(X) = 0{,}25 * n$$

Wahrscheinlichkeit

$$P(X = k) = P(X = n - k) = \binom{n}{k} * 0{,}5^n$$

Andere diskrete Verteilungen

Poissonverteilung P(λ)

- „Verteilung der seltenen Ereignisse"
- Poissonverteilung kann auch angewandt werden, wenn Grundgesamtheit nicht konkret angegeben ist oder die Wahrscheinlichkeit p nicht explizit quantifizierbar ist
- Binomialverteilung wurde untersucht für den Fall, dass die Anzahl der Wiederholungen n groß und die Wahrscheinlichkeit p für das Eintreten des Ereignisses A klein ist
  - Beispiele:
    Untersuchung einer großen Population, bei einer kleinen Wahrscheinlichkeit, dass ein Individuum erkrankt;
    radioaktiver Zerfall
- Approximation der Binomialverteilung für n ≥ 30 und p ≤ 0,1 durch folgende Grenzverteilung

$$P(X = k) = \frac{\lambda^k}{k!} * e^{-\lambda}$$

Euler'sche Zahl e = 2,718
λ … Erwartungswert der Verteilung

Erwartungswert

$$E(X) = \lambda = n * p$$

Approximieren der Varianz

$$Var(X) = n * p * q = n * \frac{\lambda}{n} * \left(1 - \frac{\lambda}{n}\right) \xrightarrow[n \to \infty]{} \lambda$$

→ Erwartungswert und Varianz stimmen überein

Poissonverteilte Zufallsvariable ***X ~ P(λ)***

- eindeutig durch den Parameter λ festgelegt

Schiefe

Poissonverteilung ist immer rechtsschief (oder linksgipfelig), da für die Schiefe gilt

$$\gamma_1 = \frac{q - p}{\sigma} \xrightarrow[n \to \infty]{} \frac{1 - 0}{\sqrt{\lambda}} = \frac{1}{\sqrt{\lambda}} > 0$$

<u>Beispiel</u>
n=2000, p=0,001, λ=n*p=2000*0,001=2

k	P(X=k)	F(k)=P(X ≤ k)
0	e^{-2} = 0,135	0,134
1	2 * e^{-2} = 0,271	0,406 (0,134+0,271)
2	$^4/_2$ * e^{-2} = 0,271	0,677 (0,406+0,271)
3	$^8/_6$ * e^{-2} = 0,180	0,857 (0,677+0,180)

Polynomialverteilung $P(n_1,\ldots n_k \mid p_1,\ldots,p_k)$

(Multinomialverteilung)

- stellt eine Verallgemeinerung der Binomialverteilung dar
- beschreibt eine Serie von n Zufallsexperimenten, bei denen pro Beobachtung eines von k möglichen Ereignissen A_1, A_2, …, A_k mit den Wahrscheinlichkeiten p_1, p_2,…,p_k auftreten kann
- Wahrscheinlichkeit, dass bei n Beobachtungen das Ereignis A_1 mit der Häufigkeit n_1, das Ereignis A_2 mit der Häufigkeit n_2, etc. eintritt:

$$P(n_1, n_2, \dots, n_k | p_1, p_2, \dots, p_k) = \frac{(p_1)^{n_1} * \dots * (p_k)^{n_k}}{n_1! * \dots * n_k!} * n!$$

Dabei ist:

$$\sum_{i=1}^{k} p_i = 1 \qquad und \qquad \sum_{i=1}^{k} n_i = n$$

<u>Beispiel</u>
Wahrscheinlichkeiten für das Auftreten der Blutgruppen:
P(0)=p_1=0,39 P(A)=p_2=0,44 P(B)=p_3=0,13 P(AB)=p_4=0,04
n=10 Personen; 4x 0 und A, 1x B und AB

$$P(4,4,1,1|039; 044; 013; 0,04) = \frac{0{,}39^4 * 0{,}44^4 * 0{,}13^1 * 0{,}04^1}{4! * 4! * 1! * 1!} * 10! = 0{,}0284$$

Negative Binomialverteilung NB(r,p)

- Binomialverteilung – mit welcher Wahrscheinlichkeit das Ereignis A bei n unabhängigen Beobachtungen k-mal gezählt wird
- negative Binomialverteilung – mit welcher Wahrscheinlichkeit das Ereignis A gerade bei der **j**-ten Beobachtung zum **r**-ten Mal eintritt
  - z.B. zur Analyse von Wartezeiten
- Spezialfall: **geometrische Verteilung NB(1,p)** – mit welcher Wahrscheinlichkeit das Ereignis A bei der j-ten Beobachtung erstmals eintritt

$$P(X = j) = q^{j-1} * p$$

… bei den ersten j-1 Beobachtungen ist jeweils das Ereignis $\bar{A}$ (q=1-p) und bei der j. Beobachtung das Ereignis A (p) eingetreten

Wahrscheinlichkeit, dass bei der j-ten Beobachtung das Ereignis A zum r-ten Mal eintritt
… unter den vorangegangenen (j-1) Beobachtungen ist das Ereignis A genau (r-1)-mal gezählt worden

$$P(X = j) = \binom{j-1}{r-1} * q^{j-r} * p^r \quad f\ddot{u}r\ j \geq r$$

<u>Beispiel</u>
Wie groß ist die Wahrscheinlichkeit, dass man nach der Blutentnahme bei maximal 14 Personen 10 positive Konserven hat? X ~ NB(10;0,85)
r = 10 p=0,85

$P(X = 10) = \binom{9}{9} * 0{,}85^{10} = 0{,}1969$
$P(X = 11) = \binom{10}{9} * 0{,}15 * 0{,}85^{10}$ =0,2953
$P(X = 12) = \binom{11}{9} * 0{,}15^2 * 0{,}85^{10} = 0{,}2436$
$P(X = 13) = \binom{12}{9} * 0{,}15^3 * 0{,}85^{10} = 0{,}1462$
$P(X = 14) = \binom{13}{9} * 0{,}15^4 * 0{,}85^{10} = 0{,}0713$

P(X ≤ 14) = 0,1969+0,2953+0,2436+0,1462+0,0713=0,9533

Hypergeometrische Verteilung HG(n;N,M)

- beschreibt n Beobachtungen, bei denen jeweils alternativ die Ereignisse A und $\bar{A}$ eintreten können
- Beobachtungen sind <u>nicht</u> unabhängig voneinander (im Gegensatz zur Binomialverteilung)
  - Auftreten eines bestimmten Ereignisses beeinflusst die Wahrscheinlichkeit aller nachfolgenden Ereignisse
- Approximation durch die Binomialverteilung, falls N im Vergleich zu n sehr groß ist

Binomial- und hypergeometrische Verteilung im Vergleich

- Urne mit roten und weißen Kugeln
- Anteil roter Kugeln beträgt p
- Binomialverteilung
  - zieht man aus der Urne nacheinander n Kugeln und nach jeder Ziehung wird die Kugel zurückgelegt, sind die Ziehungen unabhängig voneinander
  - Wahrscheinlichkeit eine rote Kugel zu ziehen beträgt bei jedem Zug p
- hypergeometrische Verteilung
  - gezogenen Kugeln werden nicht zurückgelegt
  - Wahrscheinlichkeiten ändern sich bei jedem Zug
  - Ziehungen sind voneinander abhängig

Hypergeometrische Verteilung

- insgesamt **N** Objekte stehen zur Verfügung
  - **M** haben die Eigenschaft A
  - **N-M** die Eigenschaft $\bar{A}$
- von N Objekten werden **n** zufällig gewählt

Zufallsvariable X ~ ***HG(n; N,M)***

- gibt an, wie häufig das Merkmal A bei n Beobachtungen auftritt

Wahrscheinlichkeiten

$$P(X = k) = \frac{\binom{M}{k} * \binom{N-M}{n-k}}{\binom{N}{n}}$$

k=0,…,n
p=M/N … *Anteilswert*

Erwartungswert

$$E(X) = np = n * \frac{M}{N}$$

Varianz

$$Var(X) = \frac{N-n}{N-1} * n * p * (1-p)$$

(N-n)/(N-1) … *Endlichkeitskorrektur*

<u>Beispiel</u>
71 Studenten, davon 23 männlich
Wahrscheinlichkeit, dass von 5 zufällig ausgewählten Studenten 2 männlich sind
→ N=71 M=23 n=5 k=2

$$P(X = 2) = \frac{\binom{23}{2} * \binom{48}{3}}{\binom{71}{5}} = \frac{253 * 17.296}{13.019.909} = 0{,}336$$

Diskrete Gleichverteilung DG(k)

- beschreibt ein Zufallsexperiment, bei dem k Ereignisse $A_1, A_2, \ldots A_k$ mit jeweils derselben Wahrscheinlichkeit eintreten können
  - z.B. Würfel – Wahrscheinlichkeit jeder Augenzahl p=1/6
  - z.B. Zufallszahlen: Ziffernfolgen (zw. 0 und 99), bei denen jede Ziffer mit derselben Wahrscheinlichkeit p=0,1 auftritt

Diskrete Verteilungen

Name und Bezeichnung der Verteilung	Anzahl und Art der Beobachtungen	mögliche Einzelereignisse
Binomialverteilung B(n,p)	n unabhängige	A und $\bar{A}$
Poisson-Verteilung P(λ)	n unabhängige $n \geq 30$, $p \leq 0{,}1$	A und $\bar{A}$
Polynomial-Verteilung	n unabhängige	$A_1,\ldots,A_k$
geometrische Verteilung NB(1,p)	bis A zum 1. Mal eintritt	A und $\bar{A}$
negative Binomialvertelung NB(r,p)	bis A zum r. Mal eintritt	A und $\bar{A}$
hypergeometrische Verteilung HG(n; N,M)	n abhängige	A und $\bar{A}$
Gleichverteilung DG(k)	1	$A_1,\ldots,A_k$

Stetige Verteilungen

Stetige Zufallsvariable

Funktionen

- stetige Zufallsvariable X kann alle Zahlenwerte innerhalb eines bestimmten Intervalls annehmen
  - z.B. Körpergewicht, Körpergröße
- Wahrscheinlichkeitsverteilung wird durch die Dichtefunktion beschrieben
  - Funktion ordnet jedem Wert xi der Zufallsvariablen einen Funktionswert f(xi) > 0 zu
  - Gesamtflächte unter der Kurve f(x) ist gleich 1

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

Verteilungsfunktion einer stetigen Zufallsvariablen ist das Integral über der Dichte:

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

Komplementäres Ereignis:

$$P(X > x) = \int_{x}^{+\infty} f(t)dt = 1 - F(x)$$

Eigenschaften der Verteilungsfunktion F(x)

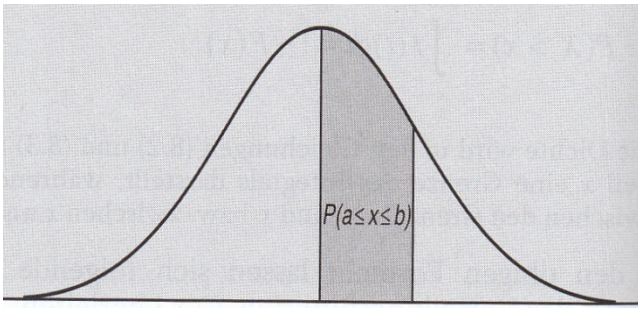
- F(x) ist eine monoton wachsende Funktion
- F(x) hat die Grenzwerte F(-∞)=0 und F(+∞)=1
- Dichte f(x) ist die Ableitung der Verteilungsfunktion

$$f(x) = F'(x)$$

Wahrscheinlichkeit, dass X einen Wert zwischen a und b annimmt:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$$

- Integral beschreibt eine Fläche, die einen Teil der Gesamtfläche unter der Dichtefunktion darstellt
  - Wert zwischen 0 und 1



Wahrscheinlichkeit, dass X einen bestimmten Wert a annimmt:

$$P(X = a) = F(a) - F(a) = 0$$

Lage- und Streuungsparameter

Erwartungswert einer stetigen Zufallsvariablen:

$$\mu = \int_{-\infty}^{+\infty} x * f(x)\, dx$$

Median, α-Quantile und Modus

- Median F($\tilde{\mu}$)=0,5
- Quantil F($\tilde{\mu}_\alpha$)= α
- Modus: Wert, an dem die Dichtefunktion f(x) ein Maximum aufweist
  - bei multimodalen Verteilungen gibt es mehrere relative Maxima

Varianz

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)\, dx$$

Zentrale Momente

Weitere Charakterisierungen einer quantitativen Zufallsvariablen gestatten

- die Momente EX^k und
- die zentralen Momente $E(X-EX)^k$

(k ist eine natürliche Zahl)

1. Moment EX
Erwartungswert μ

2. zentrales Moment E(X-EX)²
Varianz σ^2

3. zentrales Moment
Schiefe γ_1

$$\gamma_1 = \frac{E(X-EX)^3}{\sigma^3}$$

- $\gamma_1 = 0$
  - symmetrische Verteilung
- $\gamma_1 > 0$
  - rechtsschiefe Verteilung
- $\gamma_1 < 0$
  - linksschiefe Verteilung

4. zentrales Moment
Wölbung γ_2

$$\gamma_2 = \frac{E(X-EX)^4}{\sigma^4 - 3}$$

4. Moment der Normalverteilung $3\sigma^4$
Wölbung einer normalverteilten Zufallsvariablen ist gleich 0

Normalverteilung

Allgemeine Eigenschaften

- Dichte durch die Gauß'sche Glockenkurve dargestellt

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

normalverteilte Zufallsvariable ***X ~ N(μ, σ²)***

- durch den Erwartungswert μ und durch die Standardabweichung σ eindeutig charakterisiert

Eigenschaften der Normalverteilung

- Glockenkurve ist symmetrisch um den Erwartungswert μ
$$f(\mu + x) = f(\mu - x)$$
- zwei Wendepunkte bei x=μ-σ und x=μ+σ
- Maximum x= μ
- Erwartungswert=Median=Modalwert
- Dichte f(x) ist für jede reelle Zahl definiert und größer als 0.
  x → ±∞: Dichte nähert sich asymptotisch der x-Achse
  - „asymptotisch": Glockenkurve für hinreichend große x-Beträge kommt beliebig nahe an die x-Achse heran, ohne diese jedoch zu erreichen

- Form der Glockenkurve hängt von der Standardabweichung σ ab
  - kleines σ – schmal und hoch
  - großes σ – breit und niedrig
- Schiefe $\gamma_1 = 0$
- Wölbung $\gamma_2 = 0$

Wahrscheinlichkeit, dass eine normalverteilte Zufallsvariable X einen Wert zwischen zwei Grenzwerten a und b annimmt:

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi} * \sigma} * \int_a^b e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = F(b) - F(a)$$

Standardnormalverteilung N(0, 1)

- spezielle Normalverteilung mit dem Erwartungswert μ=0 und der Varianz σ²=1
- jede normalverteilte Zufallsvariable X ~ N(μ, σ²) lässt sich in die Standardnormalverteilung Z ~ N(0, 1) transformieren
  - Glockenkurve wird entlang der x-Achse so verschoben, dass der Erwartungswert μ=0
  - Form wird so angepasst, dass die Standardabweichung σ=1

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

Dichte der Standardnormalverteilung

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

Verteilungsfunktion der Standardnormalverteilung

$$\phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \varphi(t)\, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt$$

<u>Beispiel</u>
Körpergröße von Männern X sei normalverteilt: μ=180cm, σ=10cm
Wahrscheinlichkeit P(170≤X≤190)
Grenzen der standardisierten Variablen:

$$z_1 = \frac{170 - 180}{10} = -1 \qquad z_2 = \frac{190 - 180}{10} = +1$$

z_1 … Körpergröße 170cm hat eine Standardabweichung unter dem Erwartungswert
z_2 … Körpergröße 190cm hat eine Standardabweichung über dem Erwartungswert

P(-1 ≤ Z ≤ +1) = P(Z ≤ 1) – P(Z ≤ -1) P(Z ≤ 1) = ϕ(1)

Wegen der Symmetrie der Glockenkurve gilt:

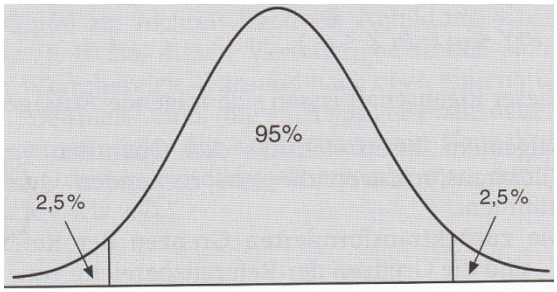
$P(Z \leq -1) = P(Z \geq 1) = 1 - P(Z \leq 1) = 1 - \phi(1)$

$P(-1 \leq Z \leq +1) = \phi(1) - (1 - \phi(1)) = 2\,\phi(1) - 1 = 2 * 0{,}81 - 1 = 0{,}68$

σ -Bereiche und Referenzbereiche

- Normalverteilung ist für alle x zwischen -∞ und +∞ definiert
- Konzentrierung der Werte in unmittelbarer Umgebung des Erwartungswertes μ
  - etwa **2/3** aller Messwerte liegen innerhalb der Grenzen **μ±σ**
  - Wahrscheinlichkeit, einen Wert außerhalb des **3σ**-Bereichs zu finden, beträgt nahezu **0**
  - Normbereiche (Referenzbereiche) – 95% oder 99% aller Werte sind dort enthalten

Intervallgrenzen für X:N(μ, σ²)	Intervallgrenzen für Z:N(0, 1)	Bezeichnung des Intervalls	Wahrscheinlichkeit p
$\mu-\sigma \leq X \leq \mu+\sigma$	$-1 \leq Z \leq 1$	1σ-Bereich	0,6827
$\mu-2\sigma \leq X \leq \mu+2\sigma$	$-2 \leq Z \leq 2$	2σ-Bereich	0,9545
$\mu-3\sigma \leq X \leq \mu+3\sigma$	$-3 \leq Z \leq 3$	3σ-Bereich	0,9973
$\mu-1{,}96\sigma \leq X \leq \mu+1{,}96\sigma$	$-1{,}96 \leq Z \leq 1{,}96$	95%-Referenzbereich	0,95
$\mu-2{,}58\sigma \leq X \leq \mu+2{,}58\sigma$	$-2{,}58 \leq Z \leq 2{,}58$	99%-Referenzbereich	0,99



95%-Referenzbereich einer Normalverteilung

Normalisierende Transformationen

Rechtsschiefe Verteilungen

- viele Verfahren setzen normalverteilte Daten voraus
- etliche Merkmale in der Medizin sind rechtsschief verteilt $\gamma_1 > 0$
  - Dichtefunktion hat einen Gipfel am linken Rand und langen Auslauf auf der rechten Seite
  - Merkmal weist nach unten eine natürliche Grenze auf, während im oberen Wertebereich die Einflussgrößen multiplikativ zusammen wirken
  - z.B. Körpergewicht der erwachsenen Bevölkerung
- durch logarithmische Transformation der Originaldaten kann eine Normalverteilung angenähert werden
- Betrachtung der transformierten Y-Variablen anstelle der X-Variablen
  - Y … normalverteilt
  - X … logarithmisch normalverteilt (lognormalverteilt)

$$Y = \ln X$$

- $X \sim LN(\mu, \sigma^2)$
  - Erwartungswert μ und Varianz σ² von Y
- lognormalverteile Zufallsvariable X muss positiv sein, ansonsten ist Transformation X → ln X nicht möglich
  - kleine x-Werte zwischen 0 und 1 in negative y-Werte abgebildet
  - große x-Werte am rechten Rand der Verteilung gestaucht
- Rücktransformation

$$X = e^Y$$

- e-Funktion ist monoton wachsend
  - → für jede Zahl c > 0: $Y \leq c$
  - gleichbedeutend mit $X = e^Y \leq e^c$

$$P\,(Y \leq c) = P(X \leq e^c)$$

- aus Quantile von Y=ln X lassen sich nach der Rücktransformation Quantile von X bestimmen

- zurücktransformierte Grenzen der Referenzbereiche von Y sind die Grenzen der Referenzbereiche von X
- Median der transformierten Variablen Y ist gleich deren Erwartungswert μ
  - Median der log-normalen Verteilung X ist gleich e^{μ}

$$P(X \leq e^{\mu}) = P(Y \leq \mu) = 0{,}5$$

- Erwartungswert μ von X ist nicht einfach zu bestimmen → geometrische Mittel ist sinnvollere Lagemaß
- geometrisches Mittel der x-Werte entspricht dem Median e^{μ}

Linksschiefe Verteilungen

- linksschiefe Verteilungen sind seltener $\gamma_1 < 0$
  - langer Anlauf links und ein Gipfel am rechten Rand
  - Daten am unteren Wertebereich, nach oben hin natürliche Grenze
  - z.B. Schwangerschaftsdauer, Tragezeit von Säugetieren
- Normalisierung dieser Verteilung durch Potenztransformation
  - Gipfel am rechten Rand wird in die Breite gezogen
  - Rechtsgipfeligkeit potenziert man mit einem höheren Wert

$$Y = X^{1,5}$$

Sätze der Wahrscheinlichkeitsrechnung

Tschebyscheff'sche Ungleichung

- erlaubt eine Abschätzung der Wahrscheinlichkeit, mit der die Zufallsvariable X um mehr als eine feste Zahl vom Erwartungswert μ abweicht
- setzt keine besondere Verteilungsform voraus
  - gilt für alle Verteilungen, also symmetrische sowie schiefe
  - hergeleitete Abschätzungen recht grob

$$P(|X-\mu| > k\sigma) \leq \frac{1}{k^2} \qquad \text{für alle } k > 0$$

$$P(|X-\mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2} \qquad \text{für alle } \varepsilon > 0$$

<u>k=1</u>

$P(|X-\mu| > \sigma) \leq 1$

<u>k=2</u>

$P(|X-\mu| > 2\sigma) \leq \frac{1}{4}$

<u>k=3</u>

$P(|X-\mu| > 3\sigma) \leq \frac{1}{9}$

→ mind. 8/9 aller Werte liegen innerhalb der Grenzen μ±3σ

Ungleichung von Gauß

- für symmetrische, eingipfelige Verteilungen
- bessere Abschätzung

$$P(|X-\mu| > k\sigma) \leq \frac{4}{9k^2} \qquad \text{für alle } k > \frac{2}{\sqrt{3}} \approx 1{,}155$$

<u>k=2</u>

$P(|X-\mu| > 2\sigma) \leq \frac{1}{9} \approx 0{,}111$

<u>k=3</u>

$P(|X-\mu| > 3\sigma) \leq \frac{4}{81} \approx 0{,}049$

Gesetz der großen Zahlen

- Erwartungswert einer Grundgesamtheit lässt sich durch einen Mittelwert umso genauer schätzen, je größer der zugrunde liegende Stichprobenumfang ist
  - Mittelwert wird aus n Werten berechnet, die zufällig in die Stichprobe gelangen
  - würde man aus derselben Grundgesamtheit n andere Stichprobenwerte ziehen, erhielte man einen anderen Mittelwert
  - bei einer großen Grundgesamtheit ergeben sich eine Vielzahl von Stichproben und somit viele Mittelwerte

- Mittelwert ist vom Zufall abhängig und lässt sich als Realisation einer Zufallsvariablen auffassen

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- alle Variablen X_i haben den Erwartungswert μ und die Varianz σ^2
- Erwartungswert

$$E(\bar{X}) = \mu$$

- Varianz

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

- Standardabweichung

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Verteilung der Mittelwerte

- Mittelwert ist zufällig zustande gekommen – Vielzahl von Mittelwerten ist möglich
- Variabilität dieser möglichen Mittelwerte wird durch die Standardabweichung $\sigma_{\bar{X}}$ quantifiziert
- „Standardfehler des Mittelwerts" = Standardabweichung
  - dieser ist umso geringer,
    je kleiner die Standardabweichung der Grundgesamtheit σ,
    je größer der Stichprobenumfang n
  - homogene Grundgesamtheiten mit kleinem σ ermöglichen bessere Schätzungen des Erwartungswerts als heterogene Populationen mit großem σ
- Verteilung der Mittelwerte entspricht einer Normalverteilung

Schwaches Gesetz der großen Zahlen

- Mittelwert $\bar{x}$ mit wachsendem Stichprobenumfang nähert sich dem Erwartungswert μ
- Mittelwert konvergiert gegen den Erwartungswert

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} \mu$$

Starkes Gesetz der großen Zahlen

- diese Annäherung des Mittelwerts an den Erwartungswert erfolgt mit einer Wahrscheinlichkeit von nahezu 1
- Differenz ε zwischen Mittelwert und Erwartungswert kann beliebig klein gehalten werden, wenn n entsprechend groß ist
  - beliebige positive Zahl - ε > 0
  - ab einer gewissen Größe ist der Unterschied zwischen Mittelwert und Erwartungswert so gering, dass eine Erhöhung des Stichprobenumfangs nicht mehr sinnvoll ist

$$P(|\bar{X} - \mu| < \varepsilon) \xrightarrow[n\to\infty]{} 1$$

Zentraler Grenzwertsatz

- Summe einer großen Anzahl von Zufallsvariablen ist normalverteilt
  - X_i (i=1,..,n) - unabhängige, identisch verteilte Zufallsvariablen mit dem Erwartungswert μ und der Varianz σ^2
  - Summe der X_i ist asymptotisch normalverteilt mit dem Erwartungswert n*μ und der Varianz n* σ^2
  - Variable Z_n ist asymptotisch normalverteilt

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n * \mu}{\sqrt{n} * \sigma} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Verteilung von Zufallsvariablen
- Zufallsvariable ist normalverteilt, wenn zahlreiche Einflüsse additiv und unabhängig voneinander zusammenwirken
  - aus diesem Grund sind z.B. Messfehler normalverteilt

Verteilung von Mittelwerten
- „Gesetz der großen Zahlen": Gesamtheit aller theoretisch denkbaren Mittelwerte, die aus Stichproben des Umfangs n derselben Grundgesamtheit berechnet werden, hat den Erwartungswert μ und die Varianz σ^2/n
- „zentraler Grenzwertsatz": falls Stichprobenumfang n hinreichend groß ($n \geq 25$) – sind diese Mittelwerte normalverteilt
  - auch wenn die Grundgesamtheit nicht normalverteilt ist

Binomialverteilung
- binomialverteilte Zufallsvariable $X \sim B(n, p)$ lässt sich auffassen als Summe von n identisch verteilten, unabhängigen Variablen X_i
  - mit jeweils Wert 1 oder 0 und Wahrscheinlichkeiten p oder q=1-p
- „zentraler Grenzwertsatz": Binomialverteilung für hinreichend großes n kann durch eine Normalverteilung X mit dem Erwartungswert $\mu=np$ und der Varianz $\sigma^2=npq$ approximiert werden
  - Ungleichung $npq \geq 9$ muss erfüllt sein

Bedeutung der Normalverteilung

Empirische Verteilung
- einige medizinisch relevante Merkmale sind angenähert normalverteilt
- andere wichtige Verteilungen sind jedoch nicht symmetrisch

Approximative Verteilung
- schiefe Verteilungen lassen sich eventuell in Normalverteilung transformieren
- Binomial- und Poissonverteilung lassen sich durch Normalverteilung approximieren (unter gewissen Bedingungen)

Verteilung für statistische Kennwerte
- zentraler Grenzwertsatz: Mittelwerte aus Stichproben des Umfangs n beliebiger Verteilungen sind normalverteilt
- bei normalverteilten Grundgesamtheiten sind auch andere Kenngrößen (z.B. Median, Varianz) normalverteilt

Basisverteilung für Prüfverteilungen
- Normalverteilung bildet Grundlage für wichtige Prüfverteilungen in der induktiven Statistik

Verteilung von Überlebenszeiten

Wichtige Begriffe

Überlebenszeit
- Dauer (Zeitspanne), die zwischen einem definierten Anfangsereignis und dem Eintritt eines zufallsbedingten Endereignisses vergeht
  - Anfangsereignisse – z.B. Geburt eines Individuums, Beginn einer therapeutischen Maßnahme
  - Endereignisse – z.B. Tod eines Patienten, eingetretener Heilerfolg
- Lebensdauer – Beobachtung eines Lebewesens vom Zeitpunkt der Geburt bis zu seinem Tod

Überlebensfunktion
- Zufallsvariable T zur Beschreibung einer Überlebenszeit
  - nimmt nur positive Werte an
  - es handelt sich um Zeiten t
- Verteilungsfunktion F(t) gibt die Wahrscheinlichkeit an, mit der ein Individuum vor dem Zeitpunkt t stirbt
- Überlebenswahrscheinlichkeit, Überlebensfunktion S(t)
  - Wahrscheinlichkeit, dass ein Individuum den Zeitpunkt t überlebt

$$S(t) = P(T > t) = 1 - F(t)$$

Bedingte Überlebenswahrscheinlichkeit

- Wahrscheinlichkeit für ein Individuum, dass den Zeitpunkt t erreich hat, eine weitere Zeitspanne Δt zu überleben

$$P(T > t + \Delta t | T > t) = \frac{P(T > t + \Delta t)}{P(T > t)}$$

Momentane Sterberate

- momentane Sterberate r(t) ist unabhängig vom Beobachtungszeitraum und kann für jeden Zeitpunkt t angegeben werden
  - Dichtefunktion f(t) der Variablen T

$$r(t) = \frac{f(t)}{S(t)}$$

Exponentialverteilung Exp(λ)

- Überlebenswahrscheinlichkeit (λ > 0)

$$S(t) = P(T > t) = e^{-\lambda t}$$

- Wahrscheinlichkeit, mit der ein Individuum vor dem Zeitpunkt t stirbt
  - Zufallsvariable T mit dieser Verteilungsfunktion nennt man exponentialverteilt T ~ Exp(λ)

$$F(t) = 1 - S(t) = P(T \leq t) = 1 - e^{-\lambda t}$$

- Dichtefunktion

$$f(t) = F'(t) = \lambda e^{-\lambda t}$$

- bedingte Überlebenswahrscheinlichkeit
  - Wahrscheinlichkeit, noch eine Zeitspanne Δt zu leben, ist unabhängig vom Alter
  - Exponentialverteilung = „gedächtnislose Verteilung"

$$P(T > t + \Delta t | T > t) = \frac{e^{-\lambda(t+\Delta t)}}{e^{-\lambda t}} = e^{-\lambda * \Delta t}$$

- Sterberate ist über die Zeit konstant

$$r(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

- Exponentialverteilung eignet sich zur Beschreibung von
  - Lebensdauern nicht alternder Objekte
    z.B. Lebensdauer radioaktiver Teilchen
  - Überlebenszeiten bei Individuen, deren Tod unabhängig vom aktuellen Alter eintritt
    z.B. Überleben nach einer schweren Erkrankung mit kurzer Lebenserwartung
- mediane Überlebenszeit
  - = Median

$$\tilde{\mu} = \frac{1}{\lambda} * ln\, 2$$

- mittlere Lebensdauer
  - = Erwartungswert

$$\mu = \frac{1}{\lambda}$$

- Varianz

$$\sigma^2 = \frac{1}{\lambda^2}$$

- diese Maßzahlen sind umso größer, je kleiner die momentane Sterberate λ
- Exponentialverteilung ist rechtsschief – Schiefe = 2

Weibull-Verteilung WB(λ, γ)

- zur Analyse von Überlebenszeiten
- Weibull-Verteilung stellt eine Verallgemeinerung der Exponentialverteilung dar

- Zufallsvariable T heißt Weibull-verteilt mit den Parametern λ > 0 und γ > 0 mit Verteilungsfunktion

$$F(t) = 1 - e^{-\lambda * t^{\gamma}} \qquad f\ddot{u}r\ t > 0$$

- Verteilung ist durch die Parameter λ und γ eindeutig festgelegt – T ~ WB(λ, γ)
  - durch den Parameter γ wesentlich flexibler, aber komplizierter zu berechnen
- Überlebenswahrscheinlichkeit

$$S(t) = P(X > t) = 1 - F(t) = e^{-\lambda * t^{\gamma}}$$

- Dichtefunktion

$$f(t) = F'(t) = \lambda\gamma * t^{(\gamma-1)} * e^{-\lambda * t^{\gamma}}$$

- momentane Sterberate

$$r(t) = \frac{f(t)}{S(t)} = \lambda\gamma * t^{\gamma-1}$$

- Sterberate konstant (γ =1)
  - → Exponentialverteilung
- Sterberate monoton wachsend (γ > 1)
  - Weibullverteilung mit γ > 1 ist geeignet, ein Überleben mit Altern zu beschreiben
- Sterberate monoton fallend (0 < γ < 1)
  - Verteilung beschreibt ein Überleben mit Regeneration, bei dem mit wachsendem Alter die Sterberate abnimmt

- Median der Weibullverteilung

$$\tilde{\mu} = (\frac{ln\,2}{\lambda})^{\frac{1}{\gamma}}$$

Prüfverteilungen

- nicht nur einzelne Messwerte x_i, sondern auch statistische Kennwerte (Mittelwert $\bar{x}$, empirische Standardabweichung s) unterliegen dem Zufall → Realisationen einer Zufallsvariablen $\bar{X}$ bzw. S
- Prüfverteilungen
  - beschreiben die Verteilung von statistischen Kenngrößen
  - Grundlage für Schätz- und Testmethoden der induktiven Statistik

t-Verteilung

- William Sealy Gosset befasste sich mit der Schätzung von Mittelwerten, deren Verteilung nach dem zentralen Grenzwertsatz durch die standardnormalverteilte Zufallsvariable Z beschrieben wird

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- t-Verteilung: Parameter σ in der Praxis meist unbekannt → Ersetzung durch die Standardabweichung s und Betrachtung der Variable T
  - für alle n ≥ 2
  - Einzelbeobachtungen X_i, aus denen $\bar{X}$ und S berechnet werden, müssen normalverteilt sein mit dem Erwartungswert μ und der Varianz σ^2

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Eigenschaften der t-Verteilung

- ähnlich wie die der Standardnormalverteilung
- ist symmetrisch um 0, stetig und glockenförmig
- kann Werte zwischen –∞ und +∞ annehmen
- Erwartungswert ist 0

- ist nicht direkt abhängig von σ sondern von s
- ist abhängig vom Parameter f, der die Anzahl der Freiheitsgrade angibt

- f=n-1 Freiheitsgrade

- für jeden Freiheitsgrad f existiert eine t-Verteilung
- Varianz beträgt $\frac{f}{f-2}$ für alle f ≥ 3 und somit größer als 1
  - t-Verteilung hat für kleine Freiheitsgrade einen flacheren Verlauf als die Standard-Normalverteilung

Chi²-Verteilung

- beschreibt in ihrer einfachsten Form die Verteilung des Quadrats einer standardnormalverteilten Zufallsvariablen Z ~ N(0,1)
- Erwartungswert von $\chi_1^2 = Z^2$

$$EZ^2 = VarZ + (EZ)^2 = 1$$

- falls mehrere Variablen Z1,…,Zn unabhängig voneinander nach N(0,1) verteilt sind, ist deren Quadratsumme $\sum Z_i^2$ χ^2-verteilt mit n Freiheitsgraden → χ_n^2-verteilt
- Erwartungswert dieser Zufallsvariablen ist gleich n
- Varianz beträgt 2n
- Schiefe $\gamma_1 = \sqrt{\frac{8}{n}}$ → rechtsschief
  - mit wachsendem n nähert sie sich einer Normalverteilung
- Betrachtung von n unabhängigen, normalverteilten Variablen X_i ~ N(μ, σ²)
  - $\frac{X_i-\mu}{\sigma}$ sind standardnormalverteilt, demnach gilt für deren Quadratsumme:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi_n^2$$

- durch Ersetzen des Erwartungswerts μ durch die Variable $\bar{X}$ → χ^2-Verteilung mit n-1 Freiheitsgraden
  - X_i ist wegen des Mittelwerts $\bar{X}$ einer einschränkenden Bedingung unterlegen

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = \frac{(n-1)*S^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Erwartungswert dieser Variablen n-1
- Varianz 2(n-1)

F-Verteilung

- Varianzen S_1^2 und S_2^2 zweier unabhängiger Stichproben aus zwei normalverteilten Grundgesamtheiten mit derselben Varianz σ²
- Variable $F_{m,n}$ einer F-Verteilung mit m und n Freiheitsgraden
  - m und n entsprechen den um 1 reduzierten Stichprobenumfängen
- F-Verteilung: Anwendung bei der Varianzanalyse

$$F_{m,n} = \frac{S_1^2}{S_2^2}$$

$$F_{1,n} = t_n^2$$

$$F_{m,n} = \frac{\chi_m^2}{\chi_n^2} * \frac{n}{m}$$

Stetige Verteilungen

Name und Bezeichnung der Verteilung	X beschreibt	Beispiele
Normalverteilung $N(\mu, \sigma^2)$	symmetrisch verteilte Daten, Dichte glockenförmig	Messfehler, Körpergröße
logarithmische Normalverteilung $LN(\mu, \sigma^2)$	rechtsschief verteilte Daten	Körpergewicht, Blutdruck
Exponentialverteilung $Exp(\lambda)$	Lebensdauern mit konstanter Sterberate	Zerfall radioaktiver Teilchen
Weibullverteilung $WB(\lambda,\gamma)$	Lebensdauern mit nicht konstanter Sterberate	Überleben mit Altern, Überleben mit Regeneration