.eps

Vorlesung Statistik für Wirtschaftsingenieure

-Sommer 2024-

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Inhaltsverzeichnis

1 Einleitung

Fragebogen

Geschlecht (m/w)	
Größe [cm]	
Schuhgröße	
Haarfarbe (blond, braun, schwarz, sonstige)	
Musikalität (gar nicht, etwas, mittel, sehr)	
Letzte Schultnote in Mathematik (0-15 Punkte)	
Stunden am Tag in WhatsApp	
Anzahl der Paare Schuhe im Schrank	
Fußballfan (ja/nein)	

Die Statistik befasst sich mit der Aufgabe eine große Gruppe von Objekten (z.B. 'Studierende an deutschen Hochschulen') auf ihre Merkmale hin zu untersuchen und mögliche Zusammenhänge zu erfassen - hierbei ist die Gruppe meist zu groß um alle Objekte zu erfassen, weshalb man von Stichproben ausgeht. Die Statistik lässt sich grob in drei Teilbereiche zusammenfassen:

Die Teilbereiche der Statistik

- Beschreibende Statistik: Sammeln, ordnen und validieren von Beobachtungsdaten, (zusammenfassende) Darstellung von Kenngrößen (z.B. 'Durchschnittliche Schuhgröße', Bearbeiten von Schuhgrößen wie 'Gr. 370').
- Explorative Statistik Erkennen von Mustern in den Daten (z.B. 'Die Merkmale Geschlecht und Schuhgröße korrelieren'.).
- Schließende Statistik Unter Anwendung der Wahrscheinlichkeitsrechnung werden allgemeine Schlussfolgerungen gezogen ('Männliche Studierende an deutschen Hochschulen haben größere Füße als weibliche Studierende an deutschen Hochschulen').

2 Beschreibende Statistik

2.1 Grundbegriffe

Wir wollen anhand des Fragebogens, den Sie am Anfang der Vorlesung bekommen haben einige wichtige Begriffe der Statistik klären.

Stichprobe und Grundgesamtheit

- Grundgesamtheit: Alle Objekte über die man eine Aussage gewinnen will, die man aber eventuell nicht vollständig erfassen kann ($\rightarrow$ Alle Studierende an deutschen Hochschulen).
- Stichprobe Eine Teilmenge der Grundgesamtheit, die tatsächlich befragt wurde ($\rightarrow$ Studierende einer Gruppe, die den Fragebogen ausgefüllt haben). Die Anzahl n der Objekte der Stichprobe heißt Stichprobenumfang.

2.2 Stichproben

Einige Methoden zur Durchführung von Stichprobenuntersuchungen

- Zufallsstichprobe
- Systematische Auswahl: Objektives Kriterium, z.B. jeder 100. Artikel
- Schichtenstichprobe: Die Grundgesamtheit wird auf Basis eines oder mehrerer Merkmale in Schichten eingeteilt. Die Schichten sollen bezüglich des Untersuchungsmerkmals möglichst homogen sein. Anschließend: ziehe aus jeder Schicht eine bestimmte Anzahl von Stichprobenstücken. Der Anteil der in die Stichprobe aufgenommenen Objekte kann von Schicht zu Schicht unterschiedlich sein.
- Klumpenstichprobe: Aus der Grundgesamtheit werden Gruppen (Klumpen) von statistischen Einheiten (oft geographisch definiert) zufällig ausgewählt. Innerhalb dieser Klumpen wird dann eine Vollerhebung durchgeführt.
- Quotenverfahren: Die Stichprobe soll die Werte gewisser Merkmale mit den gleichen Quoten bzw. Anteilen, wie in der Grundgesamtheit enthalten. ⇒ Repräsentative Stichprobe

2.3 Fehler, Ausreißer und fehlende Werte

Behandlung von Datenausreißern
Ein „Ausreißer“ ist ein Extremwert innerhalb einer Stichprobe, der so extrem ist, dass die Person, die die Stichprobe prüft, glaubt, dass er nicht stimmen kann. ⇒ Daten, die offenbar viel zu groß oder viel zu klein sind. Vorgehen:

1. Ausreißer identifizieren;
2. überprüfen, ggf. berichtigen;

Falls Ausreißer nicht berichtigt werden können:

1. Datensatz streichen oder
2. fehlerhafte Daten abändern (z. B. Ersetzen durch den Mittelwert der nicht fraglichen Daten) oder
3. Datensatz unverändert beibehalten.

Die Möglichkeiten 2) und 3) sollten nur mit größter Zurückhaltung angewendet werden. Im Zweifelsfall 1)!
Genauso behandelt man andere unmögliche oder unplausible Werte.
Behandlung von Fehlern: wie Ausreißer aber ohne 3).

2.4 Merkmale

Klassifizierung von Merkmalen
Messbare Merkmale können folgendermaßen klassifiziert werden:

- Qualitative Merkmale

  - Nominale Merkmale: Merkmale deren Ausprägungen keine bestimmte Reihenfolge haben (→ Haarfarbe).
  - Ordinale Merkmale: Merkmale deren Ausprägungen endliche Klassen sind, die man in eine sinnvolle Reihenfolge bringen kann (→ Musikalität, Schulnote in Geschichte).

- Quantitative Merkmale
  - Diskret nennt man ein Merkmal, wenn man es *zählen* kann (→ Anzahl der Freunde auf FaceBook).
  - Stetig nennt man ein Merkmal, wenn man es messen kann (→ Entfernung von Wohnort zu Hochschule, Größe).

Eindimensionale Merkmale

2.5 Definition: Urliste, relative und absolute Häufigkeit

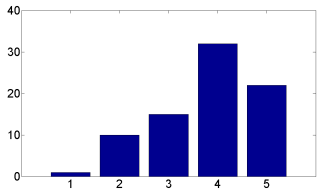
Urliste, relative und absolute Häufigkeit Hat man eine Stichprobe vom Umfang n erhoben, so kann man die gemessenen Merkmalsausprägungen ungeordnet aufschreiben. Die so entstandene Liste $x_1, \ldots, x_n$ nennt man Urliste (hier kommen auch Ausprägungen doppelt vor!). Listet man alle vorkommenden Merkmalsausprägungen $a_1, \ldots, a_m$ (hier ist kein Wert doppelt!) auf dann heißt die Anzahl $h_k = \sum_{i=1}^{n} 1_{\{x_i = a_k\}}$ absolute Häufigkeit der Ausprägung a_k und $h_k^* = \frac{h_k}{n}$ heißt relative Häufigkeit der Ausprägung a_k.

2.6 Tabellarische und graphische Darstellung - Häufigkeit

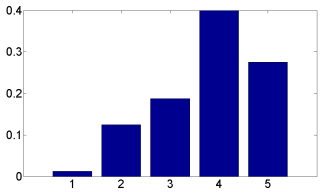
Betrachtet man zunächst ordinale bzw. nominale Merkmale, so kann man diese in Balkendiagrammen auftragen. Die Höhe eines Balkens entspricht hier der absoluten bzw. relativen Häufigkeit der entsprechenden Ausprägung. Häufig sieht man auch die Darstellung in einem Torten-Diagramm - diese Darstellung ist besonders dann sinnvoll, wenn es sich um nominale Merkmale handelt ($\rightarrow$ prozentualer Anteil der Parteien/ Anzahl der Sitze im Parlament).

Absolute und relative Häufigkeit: Balken- und Tortendiagramm

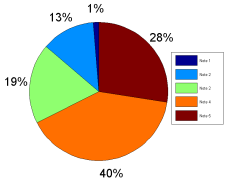
Absoluten Häufigkeiten



Relative Häufigkeiten



Tortendiagramm



Um auch stetige Merkmale oder Merkmale mit sehr vielen verschiedenen Ausprägungen in Häufigkeitsdiagrammen darstellen zu können, wird zunächst eine Unterteilung in Klassen vorgenommen - die relative Häufigkeit der einzelnen Klassen wird dann in einem Histogramm aufgetragen. Die **Flächen** unter den einzelnen Balken entsprechen hier der absoluten bzw. relativen Häufigkeit der entsprechenden Klasse.

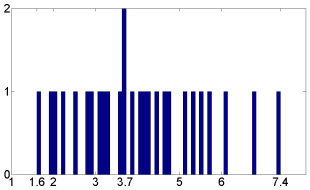
2.7 Klasseneinteilung

Ergebnis eines *Dosenstechens* unter 27 Studierenden [1]

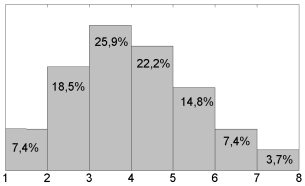
Zeiten Dosenstechen in Sekunden								
4.5	2.2	2.9	3.1	3.9	4.1	4.3	4.6	5.1
5.3	6.1	6.8	1.6	2.5	3.6	3.7	4.2	2.8
7.4	5.7	4.7	3.7	3.3	2.0	1.9	3.2	5.5

Absolute und relative Häufigkeit: Histogramm

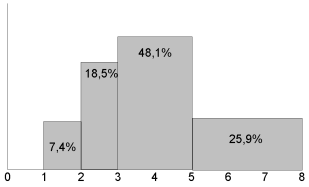
Balkendiagramm



Histogramm



Histogramm mit unterschiedlichen Breiten



Bemerkung: Im ersten Histogramm sieht es so aus, als könne man einfach die relative Häufigkeit einer Klasse als Höhe des Balkens ansetzen - dies gilt aber nur für die gewählte Klassenbreite von 1! Besonders wichtig wird die 'Flächenregel' dann, wenn die Klassen unterschiedlich breit sind (zweites Histogramm).

2.8 Regeln für die Wahl von Klassen

Sei $x_1, \ldots x_n$ eine Stichprobe vom Umfang n. Bei der Klassierung in k Klassen müssen folgende Regeln beachtet werden:

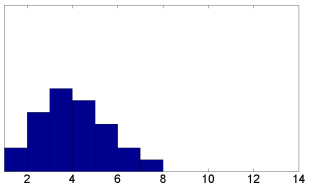
- Anzahl der Klassen $k \approx \sqrt{n}$ für $n \leq 400$.
- Jeder Stichprobenwert muss eindeutig zu einer Klasse gehören (Klassengrenzen!).
- Klassen müssen nicht gleich breit sein (Flächenregel!).

[1] Oestreich, Romberg: *Keine Panik vor Statistik*, 2009

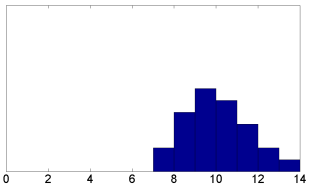
2.9 Maßzahlen einer eindimensionalen Stichprobe (Lage und Streuung)

Lage und Streuung

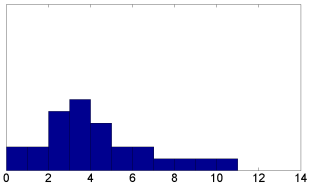
Dosenbeispiel



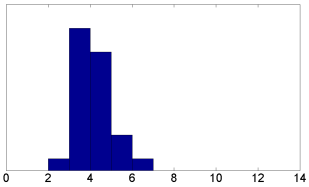
Verschobene Lage



Mehr Streuung



Weniger Streuung



2.10 Lagemaße

Arithmetisches Mittel (Mittelwert)
Sei $x_1, \ldots, x_n$ eine Stichprobe vom Umfang n, dann beschreibt das arithmetische Mittel $\bar{x}$ die Lage der Messwerte:

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Das *arithmetische Mittel* ist nur bei *quantitativen Merkmalen* sinnvoll und wird von *Ausreißern* stark beeinflusst!

Median
Sei $x_1, \ldots, x_n$ eine geordnete Stichprobe, d.h. $x_1 \leq x_2 \leq \ldots x_n$. Der Median $\tilde{x}$ der Stichprobe ist gerade der 'zentrale Wert' der Stichprobe, d.h.:

$$\tilde{x} = \begin{cases} x_{k+1}, & \text{falls } n = 2k+1 \\ \frac{x_k + x_{k+1}}{2}, & \text{falls } n = 2k. \end{cases}$$

Der *Median* ist auch bei *qualitativ ordinalen Merkmalen* sinnvoll und wird von *Ausreissern* weniger beeinflusst als das *arithmetische Mittel*!

Modalwert
Um auch bei *nominalen Merkmalen* ein Maß für die Lage der Daten zu bestimmen verwendet man den Modalwert - er entspricht derjenigen Ausprägung, die in der Stichprobe am häufigsten vorkommt. Der Modalwert ist nur dann sinnvoll, wenn eine Ausprägung deutlich häufiger vorkommt als andere.

Geometrisches Mittel
Sei $x_1, \ldots, x_n$ eine Stichprobe vom Umfang n, dann heißt

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

geometrisches Mittel der Stichprobe.

2.11 Streuungsmaße

Varianz und Standardabweichung
Sei $x_1, \ldots, x_n$ eine Stichprobe mit arithmetischem Mittel $\bar{x}$, dann heißt

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

Varianz der Stichprobe und die entsprechende positive Wurzel

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

Standardabweichung der Stichprobe.
Vereinfachte Formel für die empirische Varianz:

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

Beide Maße messen die Streuung der Stichprobe um den Mittelwert $\bar{x}$.

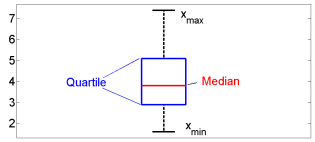
Quantile und Quartile
Sei $x_1, \ldots, x_n$ eine geordnete Stichprobe und sei $0 < p < 1$, dann heißt

$$\tilde{x}_p = \begin{cases} x_{[np]+1} & \text{falls } np \notin \mathbb{N} \\ \frac{x_{np} + x_{np+1}}{2} & \text{falls } np \in \mathbb{N} \end{cases}$$

das p-Quantil.
$[k]$ bezeichnet den ganzzahligen Anteil einer Zahl - d.h. die Stellen nach dem Komma werden abgeschnitten. Insbesondere ist $[k] = k$ falls $k \in \mathbb{N}$. Der Median entspricht gerade dem 0.5-Quantil, d.h. $\tilde{x} = \tilde{x}_{0.5}$. Die Quantile $\tilde{x}, \tilde{x}_{0.25}$ und $\tilde{x}_{0.75}$ heißen Quartile.

Boxplot
Ein Boxplot stellt den Median $\tilde{x}$, die Quartile $\tilde{x}_{0.25}$ und $\tilde{x}_{0.75}$, und die sogenannte Spannweite $x_{max} - x_{min}$ in einer Graphik dar:
Der Abstand zwischen oberem und unterem Quartil $\tilde{x}_{0.75} - \tilde{x}_{0.25}$ heißt Quartilsabstand (interquartile range).

Boxplot für das Dosenbeispiel



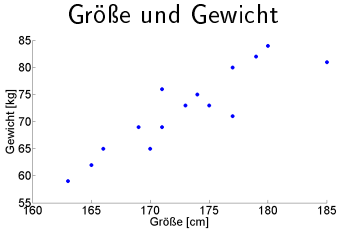
3 Explorative Statistik

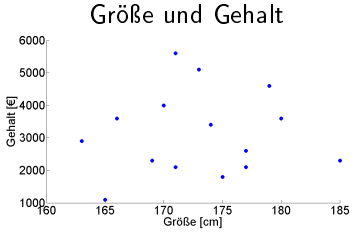
Zweidimensionale Merkmale

Bisher haben wir Merkmale isoliert betrachtet - allerdings stellt sich in der Praxis häufig die Frage, ob verschiedene Merkmale, die in der selben Population gemessen wurden in einem Zusammenhang zueinander stehen. Wichtig ist hierbei, dass zunächst nur Korrelationen beobachtet werden können, was alleine genommen keinen Rückschluss über einen kausalen Zusammenhang zulässt!
Betrachten wir also eine Stichprobe, bei der zu jedem Objekt jeweils zwei Merkmalsausprägungen vorliegen. Die Stichprobe kann man also schreiben als $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ - man spricht hier von einer zweidimensionalen Stichprobe. Gesucht wird nun ein Maß, dass eine mögliche Abhängigkeit der beiden Größen x_k und y_k quantifiziert.

3.1 Korrelation

Korrelation: Streudiagramm (scatter plot)
Sei $(x_1, y_1), \ldots, (x_n, y_n)$ eine zweidimensionale Stichprobe. Die Darstellung der Wertepaare (x_k, y_k) in der x-y-Ebene nennt man Streudiagramm (scatter plot).





Korrelationskoeffizient und Kovarianz
Sei $(x_1, y_1), \ldots, (x_n, y_n)$ eine zweidimensionale Stichprobe. Dann heißt

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{1}$$

Korrelationskoeffizient oder Pearson'scher Korrelationskoeffizient der Stichprobe. Die Zahl

$$s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$$

heißt Kovarianz der Stichprobe. Wobei $\bar{x}$ und $\bar{y}$ die jeweiligen arithmetischen Mittel und s_x und s_y die jeweiligen Standardabweichungen der eindimensionalen Stichproben $x_1, \ldots, x_n$ und $y_1, \ldots, y_n$ bezeichnen.

Eigenschaften des Korrelationskoeffizienten
Der Korrelationskoeffizient r_{xy} einer zweidimensionalen Stichprobe $(x_1, y_1), \dots (x_n, y_n)$ besitzt die folgenden Eigenschaften:

- $|r_{xy}| \leq 1$.
- $r_{xy} = r_{yx}$.
- $r_{xy} = 1 \Rightarrow$ positiver linearer Zusammenhang $y_k = mx_k + b$ mit Steigung $m > 0$.
- $r_{xy} = -1 \Rightarrow$ negativer linearer Zusammenhang $y_k = mx_k + b$ mit Steigung $m < 0$.
- $r_{xy} = 0 \Rightarrow$ keine **lineare** Korrelation.

3.2 Regression

Um den Zusammenhang zweier Merkmale einer zweidimensionalen Stichprobe besser zu beschreiben, versucht man eine Kurve bzw. Gerade zu finden, die der Darstellung in der x-y-Ebene von $(x_1, y_1), \dots, (x_n, y_n)$ möglichst ähnlich ist. Mathematisch gesprochen heißt das, das die Abstandsquadrate der Messpunkte (x_k, y_k) zur Kurve bzw. Geraden möglichst gering sind.

Lineare Regression
Sei $(x_1, y_1), \dots, (x_n, y_n)$ eine zweidimensionale Stichprobe. Die Gerade $f(x) = mx + b$, für die die Summe der Abstandsquadrate $\sum_{k=1}^{n} (y_k - f(x_k))^2$ minimal wird heißt Regressionsgerade. Die Steigung m und der Achsenabschnitt b lassen sich berechnen durch

$$m = r_{xy} \frac{s_y}{s_x}, \quad b = \bar{y} - m\bar{x},$$

wobei r_{xy} den Korrelationskoeffizienten, s_x und s_y die jeweiligen Standardabweichungen und $\bar{x}$ und $\bar{y}$ die arithmetischen Mittel der Stichproben $x_1, \dots, x_n$ und $y_1, \dots, y_n$ bezeichnen.

Nichtlineare Regression
Sei $(x_1, y_1), \dots, (x_n, y_n)$ eine zweidimensionale Stichprobe. Die Kurve, für die die Summe der Abstandsquadrate $\sum_{k=1}^{n} (y_k - f(x_k))^2$ minimal wird heißt Regressionskurve. Wie gut die Regressionskurve die Daten beschreibt, wird durch das Bestimmtheitsmaß $R^2 \in [0, 1]$ beschrieben. Wobei kleine Werte für eine schlechte Übereinstimmung sprechen.
Bemerkung: Für die lineare Regression gilt:$R^2 = r^2$, wobei r der Korrelationskoeffizient ist.

4 Wahrscheinlichkeitsrechnung

In den letzten Abschnitten haben wir gezeigt, wie man Daten einer Stichprobe zusammenfasst, darstellt und auf mögliche Korrelationen untersucht. Will man jetzt tatsächliche Aussagen über eine Grundgesamtheit treffen, so braucht man das Handwerkszeug der Wahrscheinlichkeitsrechnung, das Ihnen in den folgenden Abschnitten (in sehr verkürzter Form) gegeben werden soll.

4.1 Zufallsexperiment und Ereignisraum

Definition: Zufallsexperiment und Ereignisraum
Ein Zufallsexperiment ist ein Vorgang, der

- *beliebig oft* unter *gleichen Bedingungen* wiederholt werden kann.
- dessen Ergebnis man nicht sicher vorhersagen kann.

Die Menge aller möglichen Ergebnisse des Zufallsexperiments wird Ereignisraum Ω genannt.

Elementarereignis und Ereignis
Sei ω ein Element des Ereignisraums Ω, d.h. $\omega \in \Omega$, dann heißt die einelementige Menge $\{\omega\}$ Elementarereignis. Geeignete Untermengen von $A \subseteq \Omega$ heißen Ereignisse. Insbesondere heißt das Ereignis $A = \Omega \subseteq \Omega$ sicheres Ereignis und das Ereignis $A = \emptyset \subseteq \Omega$ unmögliches Ereignis.

4.2 Ereignisalgebra

Interpretation von Mengenoperationen bei Ereignissen
Seien A und B Ereignisse aus einem Ereignisraum Ω, d.h. insbesondere $A, B \subseteq \Omega$, dann gilt:

- $A \cap B \mathrel{\widehat{=}}$ 'A und B'.
- $A \cup B \mathrel{\widehat{=}}$ 'A oder B'.
- $\bar{A} = \Omega \setminus A \mathrel{\widehat{=}}$ 'Nicht A'.
- $A \subset B \mathrel{\widehat{=}}$ 'Aus A folgt B'.
- $A \cap B = \emptyset \mathrel{\widehat{=}}$ 'A und B sind unvereinbar'.
- $A \setminus B \mathrel{\widehat{=}}$ 'A aber nicht B'.

Rechenregeln
Seien A, B und C Ereignisse aus einem Ereignisraum Ω, d.h. $A, B, C \subseteq \Omega$, dann gilt:

- $A \cap B = B \cap A$ (Kommutativgesetz).
- $A \cup (B \cup C) = (A \cup B) \cup C$ (Assoziativgesetz).
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ und $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (Distributivgesetz).

- $\overline{A \cup B} = \bar{A} \cap \bar{B}$ (de Morgan'sche Regel).
- $\bar{\bar{A}} = A$, $\bar{\Omega} = \emptyset$, $\bar{\emptyset} = \Omega$.

4.3 Laplace-Experiment und klassische Wahrscheinlichkeit

Definition Laplace-Experiment
Ein Laplace-Experiment ist ein Zufallsexperiment mit Ereignisraum Ω mit folgenden Eigenschaften:

- Es gibt nur endlich viele Elementarereignisse, d.h. $|\Omega| = n \in \mathbb{N}$.
- Jedes Elementarereignis ist gleichwahrscheinlich, d.h. man kann jedem Elementarereignis $\{\omega\} \subseteq \Omega$ die Wahrscheinlichkeit $P(\{\omega\}) = \frac{1}{n}$ zuordnen.
- Für jedes Ereignis $A \subseteq \Omega$ mit $|A| = k$ gilt:

$$P(A) = \frac{k}{n} = \frac{\text{'Anzahl der für A günstigen Fälle'}}{\text{'Anzahl der möglichen Fälle'}}.$$

Kombinatorik

Die Berechnung der Wahrscheinlichkeit im Laplace-Experiment wirkt zunächst einfach- muss man doch einfach die 'Anzahl günstigen Fälle' durch die 'Anzahl der möglichen Fälle' teilen. Manchmal ist es jedoch gar nicht so leicht, diese Anzahl festzustellen. Mit der Abzählung der möglichen Fälle beschäftigt sich die Kombinatorik.

Definition: Fakultat

- $n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n$ (lies: n Fakultat)
- $0! = 1$ (per Definition)
- Berechnung von n! mit TR mindestens bis 69! i. d. R. möglich für größere n näherungsweise mit der Formel von Stirling;

$$\lg(n!) \approx \frac{1}{2}\lg(2\pi n) + n\lg\left(\frac{n}{e}\right)$$

Urnenmodelle
Ein Urnenmodell ist ein Modell um verschiedene Arten von Zufallsexperimenten zu simulieren. In einer Urne befinden sich n Elemente aus denen Stichproben aus k Elementen entnommen werden können. Wir haben nun folgende Möglichkeiten dies zu tun:

- Wir ziehen die Elemente ohne zurücklegen.
- Wir ziehen die Elemente mit zurücklegen.
- Wir ziehen die Elemente geordnet (Reihenfolge ist wichtig).
- Wir ziehen die Elemente ungeordnet (Reihenfolge ist unwichtig).

Alle Kombinationen aus mit/ohne zurücklegen und geordnet/ungeordnet können vorkommen- wie viele Möglichkeiten zu ziehen ergeben sich also?

Übersicht der Möglichkeiten im Urnenmodell

Ziehen von k aus n	geordnet	ungeordnet
mit Zurücklegen	n^k	$\binom{n+k-1}{k}$
ohne Zurücklegen	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

4.4 Allgemeine Definition der Wahrscheinlichkeit

Die Definition der Wahrscheinlichkeit in 4.3 nennt man klassische Definition der Wahrscheinlichkeit - sie gilt nur für Laplace-Experimente. Der Begriff Wahrscheinlichkeit, lässt sich aber auch allgemeiner fassen.

Definition Wahrscheinlichkeit
Betrachte ein beliebiges Zufallsexperiment mit Ereignisraum Ω, dann wird jedem Ereignis $A \subseteq \Omega$ durch die Wahrscheinlichkeit $P(A)$ eine Zahl zugeordnet mit:

- $0 \leq P(A) \leq 1$.
- $P(\Omega) = 1$
- $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$.

Aus den Eigenschaften zur Definition der Wahrscheinlichkeit folgen direkt andere Eigenschaften, die zur Bestimmung der Wahrscheinlichkeit eines beliebigen Ereignisses nützlich sein können.

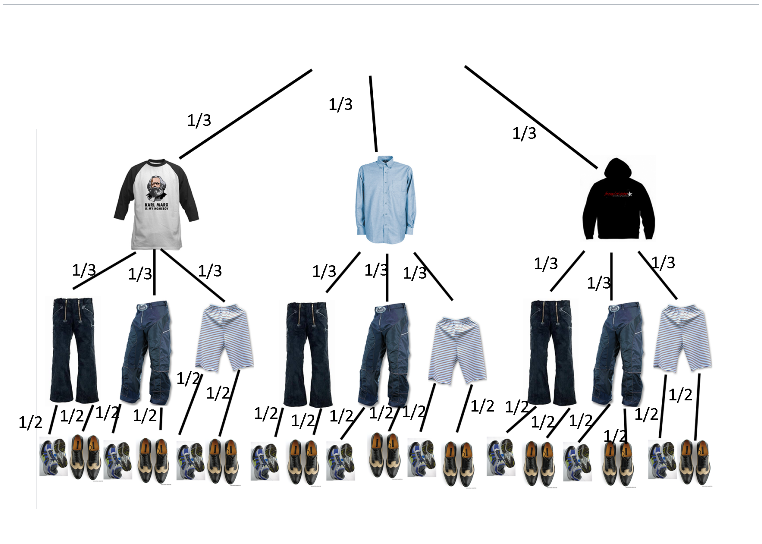
Eigenschaften der Wahrscheinlichkeit
Sei Ω ein Ereignisraum. Sei $A \subseteq \Omega$ ein beliebiges Ereignis und $P(A)$ die zugehörige Wahrscheinlichkeit:

- $P(\bar{A}) = 1 - P(A)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- Für paarweise unvereinbare Ereignisse $A_1, \ldots, A_n$ mit $A_i \cap A_j = \emptyset$ falls $i \neq j$ gilt:
$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n).$$
- $P(A) \leq P(B)$ falls $A \subseteq B$.

Die Definition der Wahrscheinlichkeit ist sehr allgemein gefasst - hier stellt sich die Frage, wie man auf die Wahrscheinlichkeiten von Ereignissen kommt. Hier hilft uns das Theorem von Bernoulli (folgt aus dem Gesetz der Großen Zahlen), das sinngemäß besagt, dass man ein Experiment nur häufig genug durchführen muss um die Wahrscheinlichkeit durch die relative Häufigkeit abschätzen zu können. Kennt man einmal die Wahrscheinlichkeit gewisser 'Grundereignisse' (das müssen keine Elementarereignisse sein!), so lassen sich die Wahrscheinlichkeiten für andere Ereignisse ableiten.

4.5 Rechnen mit Wahrscheinlichkeiten

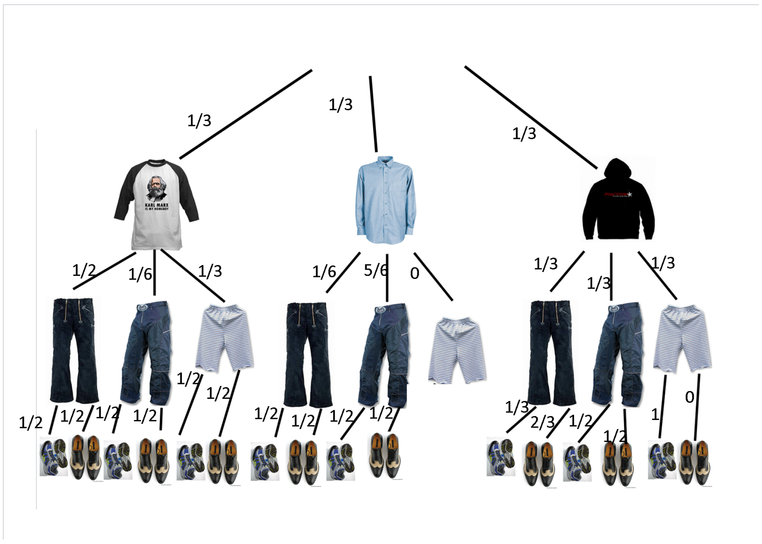
Was ziehe ich an?



Produktregel und Summenregel im Baumdiagramm
Bei einem mehrstufigen Zufallsexperiment ist die Wahrscheinlichkeit eines Pfades gleich dem Produkt der Wahrscheinlichkeiten entlang des Pfades (Produktregel). Die Wahrscheinlichkeit für ein Ereignis errechnet sich aus der Summe der Wahrscheinlichkeiten der Pfade, die zum Ereignis führen (Summenregel).

4.6 Bedingte Wahrscheinlichkeit

Was ziehe ich an (realistischer)?



Wir suchen nun einen Begriff, der die Wahrscheinlichkeit eines Ereignisses B angibt, wenn wir schon wissen, dass A eingetreten ist.

Definition: Bedingte Wahrscheinlichkeit
Die Wahrscheinlichkeit eines Ereignisses B unter der Bedingung, dass A schon eingetreten ist heißt bedingte Wahrscheinlichkeit $P(B|A)$ und ist definiert durch:

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

4.7 Multiplikationssatz

Aus der Definition der bedingten Wahrscheinlichkeit folgt direkt durch Umformen der Multiplikationssatz.
Seien A und B Ereignisse mit einer Wahrscheinlichkeit ungleich Null. Dann gilt:

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

4.8 Unabhängigkeit

Definition: Unabhängigkeit
Zwei Ereignisse A und B heißen unabhängig voneinander, wenn eine der folgenden Bedingungen erfüllt sind:

- Falls $P(A) > 0$: $P(B|A) = P(B)$.
- Falls $P(B) > 0$: $P(A|B) = P(A)$.
- $P(A \cap B) = P(A) \cdot P(B)$ (Multiplikationssatz für unabhängige Ereignisse).

Falls eine der drei Bedingungen erfüllt ist, sind alle erfüllt!

4.9 Satz von der totalen Wahrscheinlichkeit

Sei $E_1, \ldots, E_n$ eine Partition eines Ereignisraumes Ω, d.h. $E_1 \cup \cdots \cup E_n = \Omega$ und $E_i \cap E_j = \emptyset$ für $i \neq j$. Sei $A \subseteq \Omega$ ein beliebiges Ereignis, dann gilt:

$$P(A) = \sum_{k=1}^{n} P(A \cap E_k) = \sum_{k=1}^{n} P(E_k)P(A|E_k).$$

4.10 Satz von Bayes

Satz von Bayes für zwei Ereignisse
Seien A und B zwei Ereignisse, dann gilt:

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

5 Zufallsvariablen und Wahrscheinlichkeitsverteilung

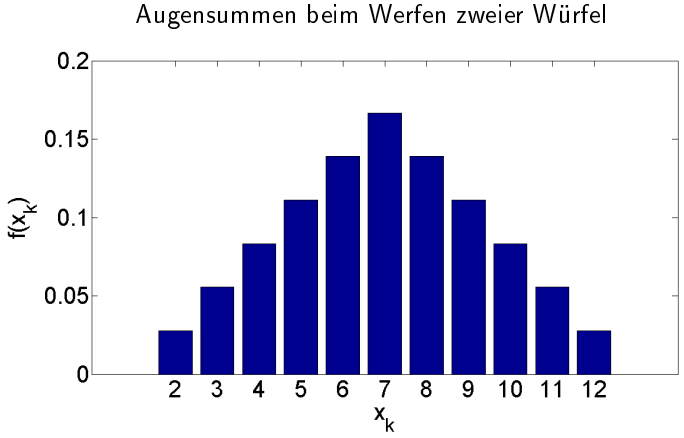
5.1 Definition: Zufallsvariable

Eine Zufallsvariable X zu einem Zufallsexperiment ist eine Funktion, die jedem Elementarereignis $\omega \in \Omega$ einen reellen Zahlenwert $X(\omega)$ zuordnet. Die Werte die eine Zufallsvariable X annehmen kann, heißen Realisierungen der Zufallsvariable und werden mit Kleinbuchstaben x bezeichnet.

5.2 Definition: Diskrete Zufallsvariable und Wahrscheinlichkeitsfunktion

Eine Zufallsvariable X heißt diskret, wenn sie nur endlich viele oder *abzählbar unendlich* viele Werte $x_1, x_2, \ldots$ annehmen kann. Jeder möglichen Realisierung x_i kann eindeutig eine Wahrscheinlichkeit $f(x_i) = P(X = x_i)$ zugeordnet werden. Diese Zuordnung nennt man Wahrscheinlichkeitsfunktion oder auch diskrete Dichte der Zufallsvariablen.

Darstellung von Wahrscheinlichkeitsfunktionen
Wie relative Häufigkeiten lassen sich auch Wahrscheinlichkeitsfunktionen als Balkendiagramme darstellen.



Bei einem Zufallsexperiment interessieren uns häufig nicht nur die Wahrscheinlichkeiten für eine Realisierung einer Zufallsvariablen $f(x_i) = P(X = x_i)$ sondern wir wollen wissen, wie wahrscheinlich es ist, dass X in einem bestimmten Bereich liegt: $P(a \leq X \leq b)$ bzw. dass X unterhalb einer Schranke liegt (z.B. wie wahrscheinlich ist es, dass die Augensumme kleiner oder gleich 6 ist?). Hierfür gibt es eine eigene Funktion.

5.3 Definition: Verteilungsfunktion diskreter Zufallsvariablen

Sei X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion $f(x_i)$, dann heißt die Funktion

$$F(x) = P(X \leq x) = \sum_{i:x_i \leq x} f(x_i), \quad x \in \mathbb{R}$$

Verteilungsfunktion von X.

5.4 Eigenschaften von Wahrscheinlichkeits- und Verteilungsfunktion diskreter Zufallsvariablen

Die Wahrscheinlichkeitsfunktion $f(x_i)$ einer diskreten Zufallsvariable X hat folgende Eigenschaften:

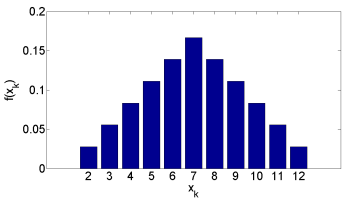
- $0 \leq f(x_i) \leq 1$.
- $\sum_i f(x_i) = 1$.

Aus der Definition 5.3 der Verteilungsfunktion $F(x)$ folgt:

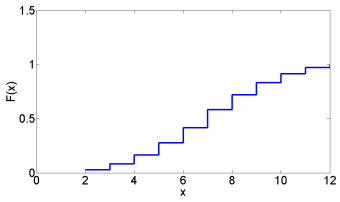
- $F(x)$ ist monoton wachsend und besitzt den Wertebereich $[0, 1]$.
- $F(x)$ ist eine *rechtsseitig stetige* Treppenfunktion mit Stufenhöhe $f(x_i)$.

Wahrscheinlichkeits- und Verteilungsfunktion für die Augensumme zweier Würfel

Wahrscheinlichkeitsfunktion $f(x_k)$.



Verteilungsfunktion $F(x)$



5.5 Erwartungswert und Varianz diskreter Zufallsvariablen

In der beschreibenden Statistik haben wir den Mittelwert als Lagemaß für die Verteilung einer Stichprobe kennen gelernt. Natürlich möchte man auch die Verteilungen von diskreten Zufallsvariablen durch wenige Parameter beschreiben. Hier muss allerdings die Wahrscheinlichkeitsfunktion berücksichtigt werden, da Realisierungen x_i mit größerer Wahrscheinlichkeitsfunktion $f(x_k)$ auch stärker berücksichtigt werden müssen. Daher verwendet man zur Berechnung von Erwartungswert und Varianz gewichtete Mittelwerte der Realisierungen x_k bzw. der entsprechenden Abstandsquadrate.

Definition: Erwartungswert einer diskreten Zufallsvariable
Sei X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion $f(x_k)$. Dann wird der Erwartungswert μ von X definiert durch:

$$\mu = E(X) = \sum_k x_k f(x_k). \tag{2}$$

Definition: Varianz einer diskreten Zufallsvariable
Sei X eine diskrete Zufallsvariable mit Wahrscheinlichkeitsfunktion $f(x_k)$ und Erwartungswert μ. Dann wird die Varianz σ^2 von X definiert durch:

$$\sigma^2 = Var(X) = E\left((X-\mu)^2\right) = \sum_k (x_k - \mu)^2 f(x_k). \tag{3}$$

Einfacher zu berechnen:

$$Var(X) = \left(\sum_k x_k^2 f(x_k)\right) - \mu^2$$

$\sigma = \sqrt{Var(X)}$ heißt Standardabweichung.

5.6 Definition: Stetige Zufallsvariable, Verteilungsfunktion und Dichte

Eine Zufallsvariable X heißt stetig, wenn sie jeden Wert in einem Intervall annehmen kann. Sie besitzt also eine stetige Verteilungsfunktion $F(x) = P(X \leq x)$. Ist die Verteilungsfunktion sogar stetig differenzierbar, d.h. $F'(x)$ ist wieder eine stetige Funktion, so heißt die Ableitung $f(x) = F'(x)$ die Dichte von X. Wir werden sehen, dass die Dichte bei einer stetigen Zufallsvariablen die entsprechende Größe zur Wahrscheinlichkeitsfunktion $f(x_k)$ bei einer diskreten Zufallsvariablen ist. Die Wahrscheinlichkeit eines einzelnen Wertes $P(X = x)$ ist bei einer stetigen Zufallsvariablen immer Null. Statt wie bei der diskreten Zufallsvariable die Einzelwahrscheinlichkeiten aufzusummieren um $F(x)$ zu erhalten, muss im stetigen Fall über die Dichte **integriert** werden.

5.7 Eigenschaften stetiger Zufallsvariablen

Sei X eine stetige Zufallsvariable mit Dichte $f(x)$ und Verteilungsfunktion $F(x)$, dann gilt:

- $F(x) = \int_{-\infty}^{x} f(u)\,du.$
- $F(\infty) = \int_{-\infty}^{\infty} f(u)\,du = 1.$
- $P(a \leq x \leq b) = \int_{a}^{b} f(u)\,du = F(b) - F(a).$
- $F(a \leq X) = \int_{a}^{\infty} f(u)\,du = 1 - F(a).$
- $P(X = x) = 0 \;\; \forall x \in \mathbb{R}$

5.8 Erwartungswert und Varianz stetiger Zufallsvariablen

Völlig analog zum diskreten Fall lassen sich auch für stetige Zufallsvariablen (normalerweise) Erwartungswert und Varianz berechnen. Da wir uns hier im Kontinuierlichen bewegen treten an die Stelle der Summen in der Definition 5.5 Integrale.

Definition: Erwartungswert einer stetigen Zufallsvariable
Sei X eine stetige Zufallsvariable mit Dichte $f(x)$. Dann wird der Erwartungswert μ von X definiert durch:

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{4}$$

Definition: Varianz einer stetigen Zufallsvariable
Sei X eine stetige Zufallsvariable mit Dichte $f(x)$ und Erwartungswert μ. Dann wird die Varianz σ^2 von X definiert durch:

$$\sigma^2 = Var(X) = E\left((X-\mu)^2\right) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx. \tag{5}$$

$\sigma = \sqrt{Var(X)}$ heißt Standardabweichung.

5.9 Quantile stetiger Zufallsvariablen

Definition: Quantil
Sei X eine stetige Zufallsvariable mit Verteilungsfunktion $F(x)$ und sei $p \in (0,1)$ derjenige Wert x_p für den gilt:

$$F(x_p) = p$$

heißt p-Quantil von X. Das Quantil zu $p = 0.5$ heißt Median.

5.10 Lineare Transformation von Zufallsvariablen

Unabhängig davon ob eine Zufallsvariable stetig oder diskret ist besitzen Erwartungswert und Varianz spezifische Eigenschaften, die hier vorgestellt werden sollen.

Erwartungswert und Varianz von Zufallsvariablen unter lineare Transformationen
Sei X eine Zufallsvariable und $a, b \in \mathbb{R}$ beliebig. Dann gilt für den Erwartungswert E und die Varianz Var:

$$E(aX+b) = aE(X) + b, \quad Var(aX+b) = a^2 Var(X)$$

Standardisierung von Zufallsvariablen
Sei X eine Zufallsvariable mit Erwartungswert $E(X) = \mu$ und Varianz $Var(X) = \sigma^2$. Dann wird durch

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

eine Zufallsvariable definiert mit $E(Z) = 0$ und $Var(Z) = 1$. Z heißt normierte Zufallsvariable.

5.11 Summen von Zufallsvariablen

Bei der Summenbildung von Zufallsvariablen muss man zunächst vorsichtig sein - selbst wenn man identisch verteilte Zufallsvariablen addiert ist das Ergebnis nicht die Verteilung von $2X$! - Summation ändert etwas an der Verteilung. Man kann zeigen, dass die Summation von Zufallsvariablen der Faltung von Wahrscheinlichkeitsfunktionen (diskret) bzw. von Dichtefunktionen (stetig) entspricht.

Erwartungswert und Summe von Zufallsvariablen
Seien X und Y Zufallsvariablen, dann gilt für den Erwartungswert E:

$$E(X + Y) = E(X) + E(Y).$$

Durch Summenbildung wird zwar die Verteilung verändert, aber der Erwartungswert addiert sich einfach. Um jetzt Aussagen über die Varianz der Summe zweier Zufallsvariablen zu machen muss man sich nun mehr Gedanken machen. Es leuchtet ein, dass die Varianz einer Summe von zwei Zufallsvariablen davon abhängt, ob die Zufallsvariablen voneinander abhängig sind. Deshalb soll hier Unabhängigkeit vorausgesetzt werden.

Varianz der Summe von unabhängigen Zufallsvariablen
Seien X und Y unabhängige Zufallsvariablen, dann gilt für die Varianz Var:

$$Var(X + Y) = Var(X) + Var(Y).$$

Wichtige diskrete Verteilungen

5.12 Binomialverteilung

Bernoulli-Experiment
Ein Bernoulli-Experiment ist ein Zufallsexperiment bei dem es nur zwei mögliche Ausgänge gibt (A und $\bar{A}$). Führt man ein Bernoulli-Experiment n mal hintereinander unter den selben Bedingungen aus, so spricht man von einer Bernoulli-Kette der Länge n.

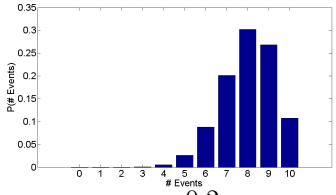
Binomialverteilung
Bei einer Bernoulli-Kette der Länge n sei $P(A) = p$ und $P(\bar{A}) = q = 1 - p$. Betrachte die Zufallsvariable $X :=$ *'Anzahl der Versuchsdurchführungen, bei denen A eintritt'* mit möglichen Realisierungen $k = 0, 1, \ldots, n$. Dann gilt:

$$f(k) = P(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}.$$

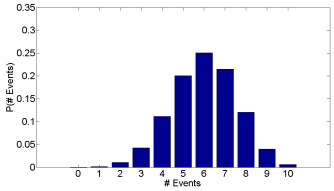
Die durch die Wahrscheinlichkeitsfunktion $f(k)$ definierte Verteilung nennt man Binomialverteilung und X heißt binomialverteilt mit Parametern n und p. Man schreibt $X \sim B(n, p)$.

Beispiel: Butterbrot mit 10 Versuchen

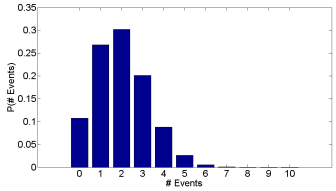
$p = 0.8.$



$p = 0.6$



$p = 0.2$



Eigenschaften der Binomialverteilung
Seien X und Y Binomialverteilte Zufallsvariablen mit Erfolgswahrscheinlichkeit p, d.h. $X \sim B(n,p)$ und $Y \sim B(m,p)$. Dann gilt:

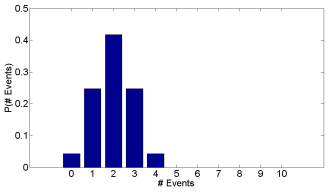
- Erwartungswert: $E(X) = \mu = np$.
- Varianz: $Var(X) = \sigma^2 = np(1-p) = npq$.
- Symmetrie: Die Zufallsvariable $Z := n - X$ ist auch Binomialverteilt: $Z \sim B(n, 1-p)$.
- Addition: Falls X und Y unabhängig sind gilt: $X + Y \sim B(n+m, p)$.

5.13 Hypergeometrische Verteilung

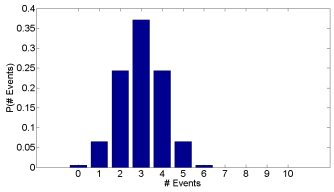
Die zweite diskrete Verteilung die wir kennen lernen ist die hypergeometrische Verteilung. Wir erinnern uns an das Urnenmodell und betrachten das ungeordnete Ziehen ohne zurücklegen und legen jetzt mehrere Kugeln von jeder Farbe in die Urne. Die Frage ist jetzt, wie wahrscheinlich es ist, bei n-maligem Ziehen eine bestimmte Anzahl von Kugeln einer bestimmten Farbe zu ziehen. Diese Frage spielt vor allem in Situationen eine Rolle, in denen man die Zusammensetzung einer Grundgesamtheit nur stichprobenweise ermitteln kann (Anteil defekter Glühbirnen, verschiedene Fischarten im See).

Beispiel: Ziehen von $n = 10$ aus $N = 20$ Kugeln unter welchen M blaue Kugeln sind

$M = 4$.



$M = 6$



Hypergeometrische Verteilung
Gegeben sei eine Grundgesamtheit aus N Elementen unter denen $M \leq N$ Elemente die Eigenschaft A besitzen. Zieht man nun ohne zurücklegen n Elemente aus N, dann wird durch die Zuordnung $X =$*'Anzahl der Elemente in der Stichprobe mit der Eigenschaft A'* eine Zufallsvariable mit möglichen Realisierungen $k = 0, 1, \ldots, \min(M, n)$ definiert. Für die Wahrscheinlichkeitsfunktion $f(k) = P(X = k)$ gilt:

$$f(k) = \frac{\left(\begin{array}{c} M \\ k \end{array}\right)\left(\begin{array}{c} N-M \\ n-k \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)}.$$

Die durch die Wahrscheinlichkeitsfunktion $f(k)$ definierte Verteilung heißt hypergeometrische Verteilung und X heißt hypergeometrisch verteilt mit Parametern n, M und N. Man schreibt $X \sim H(n, M, N)$.

Erwartungswert und Varianz der hypergeometrisch verteilter Zufallsvariablen
Sei X hypergeometrisch Verteilt mit Parametern n, M und N $X \sim H(n, M, N)$. Dann gilt für den Erwartungswert und die Varianz von X:

$$E(X) = \mu = n\frac{M}{N},$$

$$Var(X) = \sigma^2 = n\frac{M}{N}(1 - \frac{M}{N})\frac{N-n}{N-1}.$$

5.14 Näherung der hypergeometrischen Verteilung durch die Binomialverteilung

Wir haben zum Verständnis der hypergeometrischen Verteilung das Beispiel des ungeordneten Ziehens ohne zurücklegen von n Kugeln aus einer Urne mit N Kugeln von denen M blau sind betrachtet. Die Anzahl der blauen Kugeln ist hypergeometrisch verteilt mit Parametern n, M und N. Jetzt stellen wir uns eine sehr große Menge N von Kugeln vor. Wenn wir nun eine Kugel entnehmen, dann ändert das den Anteil der blauen Kugeln an allen kaum, denn $\frac{M-1}{N-1} \approx \frac{M}{N} \approx \frac{M}{N-1}$ für sehr große N. D.h. wir können, wenn wir nicht zu viele n entnehmen, näherungsweise davon ausgehen, dass bei jedem Zug die Wahrscheinlichkeit eine blaue Kugel zu ziehen immer ungefähr $\frac{M}{N}$ ist. Das heißt wir können das Experiment näherungsweise als Bernoulli-Kette der Länge n betrachten! Die Frage ist nun, wann man N gegenüber n als 'groß genug' betrachten kann.

Faustregel für die Näherung der hypergeometrischen Verteilung durch die Binomialverteilung
Falls gilt:

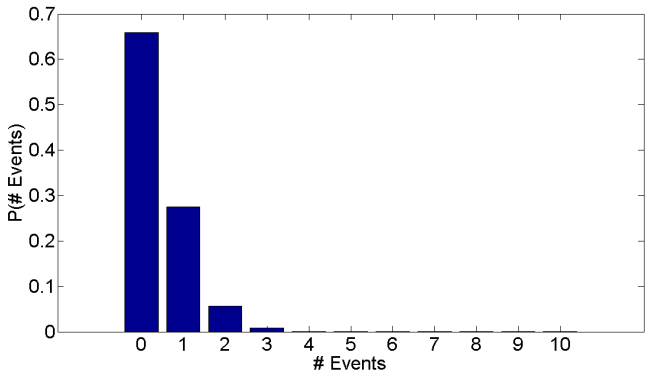
$$n \leq \frac{N}{20},$$

dann kann man die hypergeometrische Verteilung mit Parametern n, M und N durch die Binomialverteilung mit Parametern n und $p = \frac{M}{N}$ annähern ohne einen 'zu großen Fehler' zu machen.

5.15 Poissonverteilung

Beispiel: Elchjagd

Anzahl der Elche in 5 Stunden $E(X) = \frac{5}{12}$.



Definition: Poissonverteilung
Sei X eine Zufallsvariable die typischerweise die Anzahl des Vorkommens eines relativ seltenen Ereignisses beschreibt, $k = 0, 1, 2, \ldots$. Gibt es nun einen Parameter $\lambda > 0$, so dass die Wahrscheinlichkeitsfunktion die Form

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

besitzt, dann heißt X poissonverteilt mit Parameter λ. Man schreibt $X \sim Po(\lambda)$.

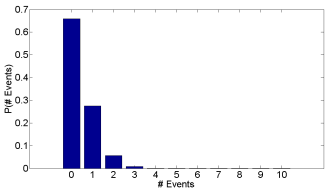
Erwartungswert, Varianz und Summenbildung bei Poissonverteilung
Seien X, X_1 und X_2 poissonverteilte Zufallsvariablen mit Parameter λ, λ_1 und λ_2 dann gilt :

- $E(X) = \mu = \lambda$
- $Var(X) = \sigma^2 = \lambda$
- $X_1 + X_2$ ist wieder eine poissonverteilte Zufallsvariable mit Parameter $\lambda_1 + \lambda_2$.

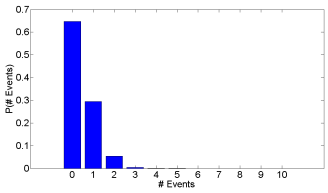
5.16 Poissonverteilung und Binomialverteilung

Elchjagd II

Anzahl der Elche $\sim Po(\frac{5}{12})$.



Anzahl der Elche $\sim B(5, \frac{1}{12})$.



Faustregel zur Annäherung der Binomialverteilung durch die Poissonverteilung
Sei X Binomialverteilt mit Parametern n und p. Falls $n \geq 50$ und $p \leq 0.1$, dann kann man die Verteilung von X näherungsweise durch die Poissonverteilung mit Parameter $\lambda = n \cdot p$ annähern ohne 'zu große Fehler' zu machen.

Wichtige stetige Verteilungen

5.17 Gleichverteilung

Die Gleichverteilung ist die einfachste stetige Verteilung. Die Zufallsvariable nimmt hier nur Werte aus einem Intervall $[a, b]$ an. Die Dichte $f(x)$ ist jetzt für alle $x \in [a, b]$ einfach gleich.

Definition: Gleichverteilung
Eine stetige Zufallsvariable X heißt gleichverteilt auf einem Intervall $[a, b]$, wenn für ihre Dichte $f(x)$ gilt:

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

Für die Verteilungsfunktion folgt dann:

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

5.18 Normalverteilung

Die Normalverteilung ist die wichtigste Verteilung überhaupt! Einfach weil sie in der Anwendung sehr häufig vorkommt - nicht zuletzt, weil hinreichend große Summen von identisch verteilten Zufallsvariablen immer annähernd normalverteilt sind (zentraler Grenzwertsatz). Wenn man also mit der Normalverteilung umgehen kann, dann ist das in der Statistik schon 'die halbe Miete'.

Definition: Normalverteilung
Eine stetige Zufallsvariable X heißt normalverteilt mit Parametern μ und σ, wenn sie die Dichte

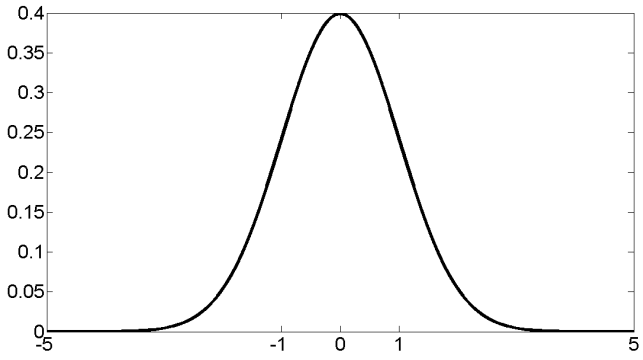
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

besitzt. Man schreibt $X \sim N(\mu, \sigma^2)$. Die Parameter μ und σ sind gerade Erwartungswert und Standardabweichung von X.

Verteilungsfunktion der Normalverteilung
Sei X eine Normalverteilte Zufallsvariable mit Parametern μ und σ, dann berechnet sich ihre Verteilungsfunktion folgendermaßen:

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

Die Gauß'sche Glockenkurve



5.19 Standardnormalverteilung

Aus Abschnitt 5.10 wissen wir, dass sich Zufallsvariablen durch lineare Transformationen standardisieren lassen. Um Werte der Dichte und Verteilungsfunktion der Normalverteilung aus Tabellen zu entnehmen, kann man sich also ausschließlich auf die standardisierte Form der Normalverteilung, die sogenannte Standardnormalverteilung konzentrieren.

Standardnormalverteilung
Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ, dann ist die Zufallsvariable $Z = \frac{X-\mu}{\sigma}$ eine standardnormalverteilte Zufallsvariable mit Parametern $\mu = 0$ und $\sigma = 1$. Die Dichte der Standardnormalverteilung $\varphi(z)$ ist

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

und die Verteilungsfunktion $\Phi(z)$ ist

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

5.20 Quantile der Standardnormalverteilung

Wir haben im Abschnitt 5.9 den Begriff Quantil für stetige Zufallsvariablen kennen gelernt. Wenn wir also zu einem vorgegebenen Wert p das Quantil x_p mit $\Phi(x_p) = p$ suchen, dann heißt das nichts anderes als die Suche nach dem Wert der Umkehrfunktion Φ^{-1} an der Stelle p, $x_p = \Phi^{-1}(p)$. Diese Werte sind in der Regel auch in Tabellen zur Standardnormalverteilung aufgelistet.

Standardnormalverteilte und normalverteilte Größen
Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ dann gelten folgende Zusammenhänge zwischen der Verteilungsfunktion F und Dichte f von X und den entsprechenden Größen der Standardnormalverteilung Φ und φ:

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \quad f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).$$

Für die p-Quantile gilt:

$$x_p = \sigma\Phi^{-1}(p) + \mu, \quad \Phi^{-1}(p) = -\Phi^{-1}(1-p).$$

p-Quantile in der Gauß'schen Glockenkurve

Zufallsstreubereiche
Sei X eine normalverteilte Zufallsvariable mit Parametern μ und σ dann gilt:

- $P(\mu - \sigma < X < \mu + \sigma) \approx 68.3\%$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 95.5\%$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 99.7\%$

Exponentialverteilung

Definition 5.20.1 (Exponentialverteilung). *Eine Zufallsvariable X mit Dichte*

$$f(x) = \alpha \cdot e^{-\alpha x} \cdot 1_{[0,\infty)}(x)$$

heißt ***exponentialverteilt mit Parameter*** *α. Man schreibt $X \sim Exp(\alpha)$.*
Für Erwartungswert und Varianz von X gilt:

$$E(X) = \frac{1}{\alpha}, \;\; Var(X) = \frac{1}{\alpha^2}$$

χ^2-Verteilung

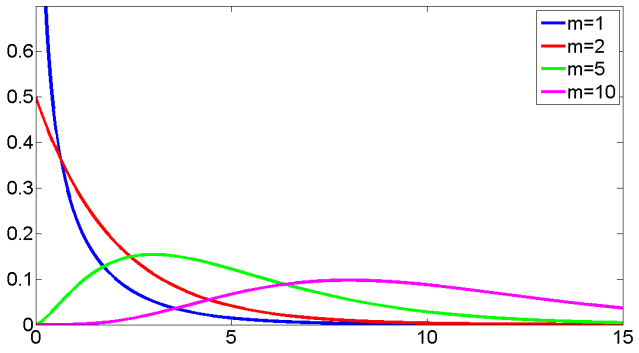
Definition 5.20.2 (χ^2-Verteilung mit m Freiheitsgraden). *Seien $Z_1, \ldots, Z_m$ unabhängige standardnormalverteilte Zufallsvariablen. Dann ist die Zufallsvariable*

$$X = Z_1^2 + Z_2^2 + \cdots + Z_m^2$$

*χ^2-**verteilt mit** m **Freiheitsgraden**. Man schreibt $X \sim \chi^2(m)$.*
Für Erwartungswert und Varianz von X gilt:

$$E(X) = m, \;\; Var(X) = 2m$$

χ^2-Verteilung mit m Freiheitsgraden



Satz 5.20.3 (Additionseigenschaft). *Seien X_1 und X_2 χ^2-verteilte Zufallsvariablen mit Freiheitsgraden m_1 und m_2, dann ist*

$$X = X_1 + X_2$$

χ^2-verteilt mit $m = m_1 + m_2$ Freiheitsgraden.

Satz 5.20.4. *Seien S^2 und $\bar{X}$ empirische Varianz und Mittelwert einer Stichprobe $X_1, \ldots, X_n$, aus einer normalverteilten Grundgesamtheit mit Varianz σ^2. Dann ist die Größe*

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

χ^2-verteilt mit $n-1$ Freiheitsgraden

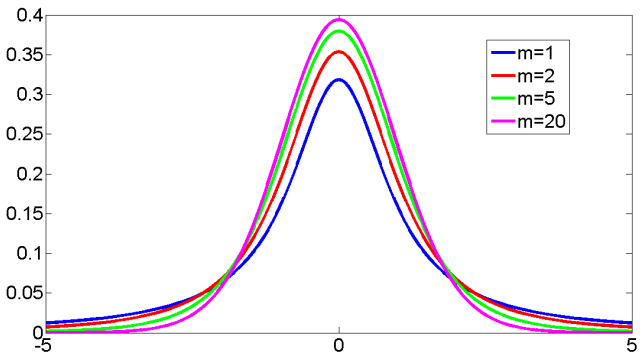
Student t-Verteilung

Definition 5.20.5 (Student t-Verteilung). *Sei X eine χ^2-verteilte Zufallsvariable mit m Freiheitsgraden und Z eine standardnormalverteilte Zufallsvariable, dann heißt die Verteilung der Zufallsvariable*

$$T = \frac{Z}{\sqrt{X/m}}$$

***t-Verteilung mit m Freiheitsgraden**. Man schreibt $T \sim t(m)$.*
Erwartungswert und Varianz sind gegeben durch:
$E(T) = 0$ für $m > 1$ und $Var(T) = \frac{m}{m-2}$ für $m > 2$.

t-Verteilung mit m Freiheitsgraden



5.21 Der zentrale Grenzwertsatz

Der zentrale Grenzwertsatz ist der Hauptgrund für die enorme Bedeutung der Normalverteilung.

Zentraler Grenzwertsatz

Seien $X_1, \dots X_n$ unabhängige identisch (aber beliebig!) verteilte Zufallsvariablen mit Erwartungswert $E(X_k) = \mu$ und Varianz $Var(X_k) = \sigma^2$. Betrachte die Summe $S = X_1 + \cdots + X_n$. S ist eine Zufallsvariable mit Erwartungswert $E(S) = n\mu$ und Varianz $Var(S) = n\sigma^2$. Falls n groß genug ist so ist S näherungsweise normalverteilt mit Parameter $n\mu$ und $n\sigma^2$. Man schreibt $S \overset{a}{\sim} N(n\mu, n\sigma^2)$.

5.22 Näherung der Binomialverteilung durch die Normalverteilung

Näherung der Binomialverteilung

Sei X eine binomialverteilte Zufallsvariable mit Parametern n und p. Falls np und $n(1-p) = nq$ 'groß genug' ($npq \geq 9$) sind, so lässt sich die Verteilung von X durch die Normalverteilung mit Parametern $\mu = np$ und $\sigma^2 = np(1-p) = npq$ annähern, d.h:

$$F(x) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{npq}}\right).$$

6 Beurteilende Statistik

Nun werden wir die Begriffe des Abschnittes über die beschreibende Statistik und die Begriffe der Wahrscheinlichkeitsrechnung verbinden, um mit Hilfe der Wahrscheinlichkeitsrechnung aus einer Stichprobe Rückschlüsse auf eine Grundgesamtheit zu ziehen.

6.1 Bezeichnungen

Wir betrachten eine Grundgesamtheit der Größe N (z.B 'Studierende an deutschen Hochschulen'). Wir wollen nun ein Merkmal X (z.B: 'Größe in [cm]') näher untersuchen. Hierfür betrachten wir eine zufällige Stichprobe $X_1, \ldots, X_n$ vom Umfang n, wir können die Stichprobe als Kette von n unabhängigen identisch verteilten Zufallsvariablen interpretieren. Die Möglichen Ausgänge $x_1, \ldots, x_n$ des Experimentes sind nun die Realisierungen von $X_1, \ldots, X_n$. Um jetzt aus dem Ergebnis der Zufallsstichprobe auf die Grundgesamtheit zu schließen muss man voraussetzen, dass die Stichprobe tatsächlich die Grundgesamtheit wiederspiegelt - sie also wirklich zufällig ist.

Konfidenzintervalle bei Normalverteilung

Zunächst reduzieren wir das Problem auf eine einfache Fragestellung: Wir wissen, dass ein Merkmal X normalverteilt ist und wir wollen aus einer Stichprobe $X_1, \ldots, X_n$ die Parameter μ und/oder σ schätzen. Wir können diesen Wert nicht exakt bestimmen, weil wir nur eine Stichprobe zur Verfügung haben, deshalb bestimmen wir einen Schätzwert (Punktschätzer) zusammen mit einem Intervall, in welchem der wahre Wert zu einer vorgegebenen Wahrscheinlichkeit liegt (Konfidenzintervall).

Definition: Konfidenzintervall
Gesucht ist ein Parameter Θ der Verteilung einer Zufallsvariablen X. Der Parameter Θ soll aus einer Stichprobe $X_1, \ldots, X_n$ geschätzt werden. Ein Intervall

$$[a(X_1, \ldots, X_n), b(X_2, \ldots, X_n)]$$

dessen Grenzen a und b sich aus der Stichprobe $X_1, \ldots, X_n$ berechnen lassen und das den gesuchten Parameter Θ zu einer vorgegebenen Wahrscheinlichkleit $1 - \alpha$ überdeckt, d.h.

$$P(\Theta \in [a(X_1, \ldots, X_n), b(X_2, \ldots, X_n)]) = 1 - \alpha$$

heißt Konfidenzintervall zum Konfidenzniveau $1-\alpha$. α heißt Irrtumswahrscheinlichkeit.

Geeignete Punktschätzer für μ und σ bei Normalverteilung
Ist ein Merkmal X normalverteilt und $X_1, \ldots, X_n$ eine Stichprobe, so stellt der Mittelwert $\bar{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$ einen guten Schätzer für den Parameter μ und die empirische Varianz der Stichprobe $S^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2$ einen *guten* Schätzer für den Parameter σ^2 der Normalverteilung dar.

Guter Schätzer heißt streng genommen, dass der Schätzer erwartungstreu und konsistent ist - zur detaillierten Definition siehe Teschl S.33.

6.2 Konfidenzintervall für den Erwartungswert bei bekannter Varianz

Sei $x_1, \ldots, x_n$ das Ergebnis einer Stichprobe zu einem normalverteilten Merkmal $X \sim N(\mu, \sigma^2)$. Wir gehen davon aus, dass der Parameter σ^2 bekannt ist und wollen ein Konfidenzintervall für den Parameter μ zum Niveau $1-\alpha$ bestimmen.

- Bestimme den Mittelwert $\bar{x}$.
- Bestimme das $1-\frac{\alpha}{2}$-Quantil $z_{1-\frac{\alpha}{2}}$ der Standardnormalverteilung (Tabelle).
- Das gesuchte Konfidenzintervall ist

$$[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$$

6.3 Konfidenzintervall für den Erwartungswert bei unbekannter Varianz

Sei $x_1, \ldots, x_n$ das Ergebnis einer Stichprobe zu einem normalverteilten Merkmal $X \sim N(\mu, \sigma^2)$. Wir gehen nun davon aus, dass der Parameter σ^2 unbekannt ist und wollen ein Konfidenzintervall für den Parameter μ zum Niveau $1-\alpha$ bestimmen.

- Bestimme den Mittelwert $\bar{x}$ und die empirische Standardabweichung s.
- Bestimme das $1-\frac{\alpha}{2}$-Quantil $t_{n-1;1-\frac{\alpha}{2}}$ der t-Verteilung mit $n-1$ Freiheitsgraden (Tabelle).
- Das gesuchte Konfidenzintervall ist

$$[\bar{x} - t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1;1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$$

6.4 Konfidenzintervall für die Differenz zweier Erwartungswerte bei unbekannter aber gleicher Varianz

Seien $x_1, \ldots, x_n$ und $y_1, \ldots, y_m$ Ergebnisse von Stichproben zu zwei normalverteilten Merkmalen $X \sim N(\mu_1, \sigma^2)$ und $Y \sim N(\mu_2, \sigma^2)$. Wir gehen nun davon aus, dass der Parameter σ^2 unbekannt aber bei beiden Merkmalen gleich ist und wollen ein Konfidenzintervall für die Differenz der Parameter $\mu_1 - \mu_2$ zum Niveau $1-\alpha$ bestimmen.

- Bestimme die Mittelwerte $\bar{x}$ und $\bar{y}$ bzw. die empirischen Standardabweichungen s_x und s_y.
- Bestimme das $1-\frac{\alpha}{2}$-Quantil $t_{n+m-2;1-\frac{\alpha}{2}}$ der t-Verteilung mit $n+m-2$ Freiheitsgraden (Tabelle).
- Das gesuchte Konfidenzintervall ist

$$[\bar{x} - \bar{y} - t_{n+m-2;1-\frac{\alpha}{2}} s_d, \bar{x} - \bar{y} + t_{n+m-2;1-\frac{\alpha}{2}} s_d]$$

mit $s_d = \sqrt{(n-1)s_x^2 + (m-1)s_y^2} \cdot \sqrt{\frac{m+n}{n \cdot m(m+n-2)}}$.

Konfidenzintervalle bei Binomialverteilung

6.5 Konfidenzintervall für p einer Binomialverteilung

In Abschnitt 5.22 haben wir gesehen, dass man für 'hinreichend große' Stichprobenumfänge n die Binomialverteilung $B(n,p)$ durch die Normalverteilung $N(np, npq)$ annähern kann. Diese Eigenschaft kann man auch dafür nutzen ein Konfidenzintervall für den Anteilswert p einer Binomialverteilung zu bestimmen.

Approximatives Konfidenzintervall für einen Anteilswert p bei großem Stichprobenumfang
Sei $x_1, \ldots, x_n$ das Ergebnis einer Stichprobe zu einem Binomialverteilten Merkmal $X \sim B(n,p)$ mit $n \geq 20$. Wir wollen ein Konfidenzintervall für den Anteilswert p zum Niveau $1-\alpha$ bestimmen:

- Bestimme den empirischen Anteil $\bar{p}$ aus der Stichprobe.
- Bestimme das Quartil $z_{1-\frac{\alpha}{2}}$ der Standardnormalverteilung.
- Falls $n\bar{p}(1-\bar{p}) \geq 9$, so ist das gesuchte Konfidenzintervall:

$$\left[\bar{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} - \frac{0.5}{n}, \bar{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} + \frac{0.5}{n}\right].$$

7 Testen von Hypothesen

7.1 Statistisches Testproblem

Ein statistisches Testproblem besteht aus einer Nullhypothese H_0 und einer Alternativhypothese H_1 über ein bestimmtes Merkmal Θ. Aus einer Zufallsstichprobe $X_1, \ldots, X_n$ lässt sich eine Teststatistik $T(X_1, \ldots, X_n)$ ermitteln, deren Verteilung unter H_0 bekannt ist. Der Ablehnungsbereich A umfasst Werte für T, die unter H_0 sehr unwahrscheinlich sind, d.h. $P(T \in A|H_0) \leq \alpha$. Die Wahrscheinlichkeit α heißt das Signifikanzniveau des Tests und liegt üblicherweise bei $\alpha = 0.1$ oder $\alpha = 0.05$. Fällt das Testergebnis T in den Ablehnungsbereich, so wird H_0 zugunsten von H_1 abgelehnt. Ansonsten wird H_0 beibehalten.

7.2 Fehler 1. und 2. Art

Übersicht über mögliche Entscheidungen: Fehler 1. und 2. Art können nicht gleich-

	Testentscheidung für	
	H_0	H_1
H_0 wahr	richtig (1-α) *Sensitivität*	Fehler 1. Art (α-Fehler)
H_1 wahr	Fehler 2. Art (β-Fehler)	richtig (1-β) *Power*

zeitig kontrolliert werden, deshalb wird immer der Fehler 1. Art kontrolliert. Das Signifikanzniveau entspricht einer oberen Grenze für den Fehler 1. Art.
Deshalb formuliert man einen Test immer so, dass der Fehler 1. Art der 'schlimmere' Fehler wäre!.
Auch bei kontrolliertem Fehler 1. Art versucht man mit möglichst hoher 'Power' zu testen - dies erreicht man häufig durch einen größeren Stichprobenumfang.

7.3 p-Wert

Der p-Wert für ein Testergebnis entspricht der Wahrscheinlichkeit unter H_0 das Ergebnis oder noch ein extremeres Ergebnis in Richtung H_1 zu erhalten. Ist der p-Wert kleiner als das Signifikanzniveau α so wird H_0 abgelehnt.

7.4 Zweiseitige und einseitige Tests

Ein Testproblem heißt

- zweiseitig, wenn es von der Form

  $$H_0 : \mu = \mu_0 \text{ gegen } H_1 : \mu \neq \mu_0$$

  ist.

- einseitig, wenn es von der Form

  $$H_0 : \mu \leq \mu_0 \text{ gegen } H_1 : \mu > \mu_0$$

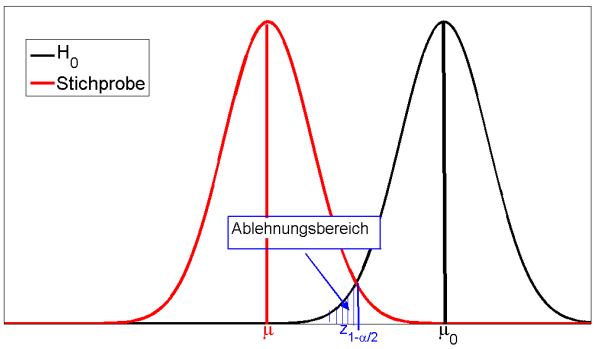
  bzw.

  $$H_= : \mu \geq \mu_0 \text{ gegen } H_1 : \mu < \mu_0$$

  ist.

7.5 Gauß-Test

Zweiseitiger Gauß-Test (z-Test)



Der Gauß Test ist ein Test für μ eines normalverteilten Merkmals $X \sim N(\mu, \sigma^2)$ bei bekanntem σ^2. Die Hypothesen sind $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$. Das Signifikanzniveau sei α.

- Ziehe eine Stichprobe vom Umfang n, berechne den Mittelwert $\bar{x}$.
- Bestimme das $1 - \frac{\alpha}{2}$-Quantil der Standardnormalverteilung $z_{1-\frac{\alpha}{2}}$ (Tabelle).
- Bestimme den Zufallsstreubereich

$$I = [\mu_0 - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$$

- Verwerfe H_0 falls $\bar{x} \notin I$.

Einseitiger Gauß-Test (z-Test)

Einseitig werden die Hypothesen zu $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$ bzw. $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$. Das Signifikanzniveau sei α.

- Ziehe eine Stichprobe vom Umfang n, berechne den Mittelwert $\bar{x}$.
- Bestimme das $1 - \alpha$-Quantil der Standardnormalverteilung $z_{1-\alpha}$ (Tabelle).
- Bestimme den Zufallsstreubereich

$$I = [\mu_0 - z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}]$$

- Verwerfe H_0 falls $\bar{x} \notin I$.

Da das Prinzip des einseitigen Testens völlig Analog auf andere Tests bei Normalverteilung übertragen werden kann, wird im Folgenden nur jeweils der zweiseitige Fall betrachtet.

7.6 Test für den Parameter p bei Binomialvertleilung

Um Hypothesen zum Parameter p einer Binomialverteilung zu testen wird die Approximation der Binomialerteilung durch die Normalverteilung genutzt. Die Hypothesen sind $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$. Das Signifikanzniveau sei α.

- Ziehe eine Stichprobe vom Umfang n, berechne den Anteilswert $\bar{p}$:
- Bestimme das $1 - \frac{\alpha}{2}$-Quantil der Standardnormalverteilung $z_{1-\frac{\alpha}{2}}$ (Tabelle).
- Bestimme den Zufallsstreubereich

$$I = [p_0 - z_{1-\alpha} \cdot \sqrt{\frac{p_0(1-p_0)}{n}} - \frac{0.5}{n}, p_0 + z_{1-\alpha} \cdot \sqrt{\frac{p_0(1-p_0)}{n}} + \frac{0.5}{n}]$$

- Verwerfe H_0 falls $\bar{p} \notin I$.

7.7 t-Test

Der t-Test ist ein Test für den Erwartungswert μ eines normalverteilten Merkmals $X \sim N(\mu, \sigma^2)$ bei unbekanntem σ^2. Die Hypothesen sind $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$. Das Signifikanzniveau sei α.

- Ziehe eine Stichprobe vom Umfang n, berechne den Mittelwert $\bar{x}$ und die empirische Standardardabweichung s.
- Bestimme das $1 - \frac{\alpha}{2}$-Quantil der t-Verteilung mit $n - 1$ Freiheitsgraden $t_{n-1;1-\frac{\alpha}{2}}$ (Tabelle).
- Bestimme den Zufallsstreubereich

$$I = [\mu_0 - t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \mu_0 + t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}]$$

- Verwerfe H_0 falls $\bar{x} \notin I$.

7.8 Zweistichproben t-Test

Der Zweistichproben t-Test ist ein Test für die Differenz der Erwartungswerte $\mu_1 - \mu_2$ zweier normalverteilter Merkmale $X \sim N(\mu_1, \sigma^2)$ und $Y \sim N(\mu_2, \sigma^2)$ bei unbekanntem aber gleichem σ^2. Die Hypothesen sind $H_0 : \mu_1 - \mu_2 = 0$ gegen $H_1 : \mu_1 - \mu_2 \neq 0$. Das Signifikanzniveau sei α.

- Ziehe zwei Stichproben vom Umfang n und m, berechne die Mittelwerte $\bar{x}$ und $\bar{y}$, die empirischen Standarbweichungen s_x und s_y und die Größe $s_d = \sqrt{(n-1)s_x^2 + (m-1)s_y^2} \cdot \sqrt{\frac{m+n}{n \cdot m(m+n-2)}}$
- Bestimme das $1 - \frac{\alpha}{2}$-Quantil der t-Verteilung mit $m + n - 2$ Freiheitsgraden $t_{m+n-2;1-\frac{\alpha}{2}}$ (Tabelle).
- Bestimme den Zufallsstreubereich

$$I = [-t_{m+n-2;1-\frac{\alpha}{2}} \cdot s_d, t_{m+n-2;1-\frac{\alpha}{2}} \cdot s_d]$$

- Verwerfe H_0 falls $(\bar{x} - \bar{y}) \notin I$.

7.8.1 χ^2-Unabhängigkeits-Test

Ziel des χ^2-Unabhängigkeits-Tests ist es zu prüfen ob zwei nominal skalierte Zufallsvariablen unabhängig voneinander sind oder ob die eine Variable die andere eventuell beeinflusst.
Wir beschränken uns zunächst auf eine Kontigenztafel aus zwei Variablen mit jeweils zwei Kategorien bezeichnet mit A und $\bar{A}$ bzw. B und $\bar{B}$. In der Tafel sieht man jeweils die Häufigkeiten einer Stichprobe vom Umfang N zusammen mit den jeweiligen Summen über die jeweilige Zeile bzw. Spalte (**Randsummen**). Die Häu-

	A	$\bar{A}$	$\sum$
B	N_{11}	N_{12}	$N_{11}+N_{12}$
$\bar{B}$	N_{21}	N_{22}	$N_{21}+N_{22}$
$\sum$	$N_{11}+N_{21}$	$N_{12}+N_{22}$	$N=N_{11}+N_{12}+N_{21}+N_{22}$

figkeiten $N_{11},\ldots,N_{22}$ sollen jetzt mit den sogenannten 'erwarteten Häufigkeiten' $E_{11},\ldots,E_{22}$ verglichen werden. Wie wir aus der Wahrscheinlichkeitsrechnung bereits wissen, lassen sich Wahrscheinlichkeiten durch relative Häufigkeiten annähern. Zusätzlich verwenden wir, das für unabhängige Ereignisse A und B gilt:

$$P(A\cap B)=P(A)\cdot P(B)\approx\frac{(N_{11}+N_{21})}{N}\cdot\frac{(N_{11}+N_{12})}{N}.$$

Um die erwartete Häufigkeit zu ermitteln muss dieser Wert nur noch mit N multipliziert werden:

$$E_{11}=\frac{(N_{11}+N_{21})\cdot(N_{11}+N_{12})}{N}.$$

für die anderen Werte gilt analog:

$$E_{ij}=\frac{(N_{1j}+N_{2j})\cdot(N_{i1}+N_{i2})}{N}.$$

Besitzt die erste Variable n Kategorien und die zweite Variable m Kategorien, so verallgemeinern sich die erwarteten Häufigkeiten zu:

$$E_{ij}=\frac{\sum\limits_{j=1}^{m}N_{ij}\cdot\sum\limits_{i=1}^{n}N_{ij}}{N}.$$

Als Teststatistik dient die Summe der relativen quadratischen Abweichung der gemessenen von der erwarteten Häufigkeit:

$$T=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{E_{ij}-N_{ij}}{E_ij}.$$

Diese Größe ist bei ausreichend großer Stichprobe ($\star$) χ^2-verteilt mit $(n-1)\cdot(m-1)$ Freiheitsgraden. Als Nullhypothese dient hier die stochastische nabhängigkeit der beiden Variablen.
Achtung: Der χ^2-Test ist nur für große Stichproben sinnvoll ($E_{ij}>5$), da sonst die χ^2-Verteilungsannahme nicht mehr erfüllt ist.

7.8.2 Exakter Fisher Test

Wie bereits erwähnt müssen in der Kontingenztafel die Zellen ausreichend belegt sein, damit der χ^2-Test sinvoll ist. Bei 'dünnerer' Zellenbelegung verwendet man alternativ den **exakten Fisher Test**. Hier geht man von der Nullhypothese aus, dass die Häufigkeiten von A und $\bar{A}$ in der Gruppe B und $\bar{B}$ gleich sind. Nun werden alle möglichen Variationen der Tabelle 7.8.1 betrachtet, deren Zellwerte mindestens genauso extrem sind, die aber gleichzeitig exakt dieselben Randhäufigkeiten besitzen. Deren Wahrscheinlichkeiten unter der Nullhypothese entsprechen einer hypergeometrischen Verteilung - Sie werden aufsummiert und können direkt als p-Wert interpretiert werden.