Statistik I für WInf und WI

Prof. Dr. Wilhelm Stannat

Inhalt:

Das vorliegende Skript ist die Zusammenfassung der Vorlesung Statistik I für WInf und WI im Wintersemester 2008/09. Die Lektüre des Skriptes ist kein gleichwertiger Ersatz für den Besuch der Vorlesung.

Korrekturen bitte per Email an: `stannat@mathematik.tu-darmstadt.de`

I. Deskriptive Statistik

1. Grundbegriffe

Die deskriptive oder auch beschreibende Statistik beschäftigt sich mit der Erhebung und Aufbereitung von Daten, die im Rahmen von Erhebungen, wie zum Beispiel Volkszählungen und Umfragen, oder bei Messungen gewonnen werden.
Erhoben werden **Merkmale** wie zum Beispiel Alter, Geschlecht, Einkommen, Temperatur oder Druck. Unterschieden werden Merkmale nach **qualitativen Merkmalen**, wie Geschlecht, Nationalität oder Beruf, und **quantitativen Merkmale**, die man ihrerseits nochmals in **diskrete** Merkmale, etwa Alter und Einkommen, und **stetige** Merkmale, etwa Temperatur und Geschwindigkeit unterteilt.
Die **Merkmalsausprägungen** sind die Gesamtheit der möglichen Werte eines Merkmals, also:

Beispiele

Geschlecht: männlich, weiblich

Alter: $0, 1, 2, 3, \ldots$

Temperatur: die reellen Zahlen $\mathbb{R}$ oder Teilmengen der reellen Zahlen

Als **Merkmalsträger** bezeichnet man die für die Erhebung der Daten relevanten Objekte. Das sind also zum Beispiel bei einer Umfrage die Menge der relevanten Personen. Die Gesamtheit der für eine statistische Erhebung relevanten Merkmalsträger heißt **Grundgesamtheit**.

Bei Erhebungen unterscheidet man zwischen einer **Vollerhebung**, bei der alle Merkmalsträger der Grundgesamtheit erfasst werden (etwa Volkszählung) und einer **Teilerhebung** oder **Stichprobenerhebung**, bei der nur eine zufällig gewonnene Teilmenge der Grundgesamtheit erfasst wird, wie es bei Umfragen der Fall ist.

Merkmalstypen, Skalierung, Klassierung

Wir haben bereits die Unterscheidung zwischen quantitativen und qualitativen Merkmalen angesprochen. Durch **Quantifizierung** kann ein qualitatives Merkmal in ein quantitatives umgewandelt werden, z.B.:

$$\begin{array}{ll} \texttt{grün} & = 23 \\ \texttt{blau} & = 14 \end{array} \quad \text{oder} \quad \begin{array}{ll} \texttt{Europa} & = 3 \\ \texttt{Asien} & = 1 \end{array}$$

Skalierung

Bei quantitativen Merkmalen spielt die Skalierung eine wichtige Rolle. Man unterscheidet folgende Skalen:

Nominalskala: die zugeordneten Zahlen dienen lediglich zur Unterscheidung der Merkmalsausprägungen

Beispiel Steuerklassen I, II, ..., V.

Ordinalskala, Rangskala: die Merkmalsausprägungen werden zueinander in einer Rangfolge in Beziehung gesetzt

Beispiel Schadstoffklassen 1, 2, 3, 4.

Kardinalskala: zusätzlich zur Rangfolge spielt auch noch der Abstand zwischen zwei Merkmalsausprägungen eine Rolle

Beispiele Temperatur, Einkommen.

Klassierung
Ein stetig verteiltes Merkmal kann durch die **Aufteilung** der Merkmalsausprägungen **in Teilintervalle (Klassen)** in ein diskretes Merkmal überführt werden.

Beispiel

$$\begin{array}{rcll} & & & < 160\,\mathrm{cm} \quad 180\ldots189\,\mathrm{cm} \\ \texttt{Körpergröße in cm} & \longrightarrow & \texttt{Klassen} & 160\ldots169\,\mathrm{cm} \quad 190\ldots199\,\mathrm{cm} \\ & & & 170\ldots179\,\mathrm{cm} \quad \geq 200\,\mathrm{cm} \end{array}$$

Bei der Erhebung statistischer Daten unterscheidet man zwischen

- Befragung (z. B. Umfrage, Volkszählung)
- Beobachtung (z. B. Verkehrszählung, Messung,...)
- Experiment (Messung im "physikalischen" Experiment).

Bei der **Teilerhebung** statistischer Daten wird die **Stichprobenauswahl** entscheidend, d. h. von welchen Merkmalsträgern werden die Daten erhoben. Es gibt hierzu, neben **willkürlicher** Auswahl, Stichprobentechniken.

Beispiel Quotenauswahl

Bei der Auswahl achtet man darauf, dass bestimmte Merkmalsausprägungen in der Teilgesamtheit dieselbe relative Häufigkeit besitzen wie in der Grundgesamtheit. Man spricht dann von einer "repräsentativen" Auswahl, im Zusammenhang mit Umfragen etwa von einer repräsentativen Umfrage.

2. Auswertung eindimensionaler Datensätze

Die Gesamtheit der Daten aus der statistischen Erhebung bezeichnet man als **Urliste**. Wird nur ein Merkmal erhoben, so kann man die erhobenen Merkmalswerte als Folge aufschreiben:

$$x_1, x_2, x_3, \ldots, x_n$$

Auf diese Weise erhält man eine **Stichprobe der Länge** n. Alternativ spricht man auch von einer **Messreihe**, sowie statt von Merkmalswerten auch von **Messwerten** oder **Beobachtungen**.

Beispiel Jahreshöchsttemperaturen (in $^\circ C$) in Darmstadt in den Jahren 1996 - 2005

$$33.0 \quad 33.2 \quad 36.5 \quad 32.2 \quad 34.2 \quad 34.4 \quad 37.2 \quad 38.1 \quad 32.3 \quad 34.7$$

Absolute und relative Häufigkeiten

Es seien $a_1, a_2, \ldots, a_s$ die möglichen Merkmalsausprägungen. Die Anzahl der Merkmalswerte $x_1, \ldots x_n$, die mit a_j übereinstimmen, heißt **absolute Häufigkeit** von a_j und wird mit $h(a_j)$ bezeichnet $(j = 1, \ldots, s)$.

Der Anteil

$$f(a_j) := \frac{h(a_j)}{n} \qquad (j = 1, \ldots, s)$$

des Merkmalswertes a_j an der Gesamtzahl n der erhobenen Merkmalswerte heißt **relative Häufigkeit**. An den relativen Häufigkeiten kann man insbesondere sofort die Prozentanteile ablesen.

Offenbar gilt:

$$\sum_{j=1}^{s} h(a_j) = n \qquad \text{und} \qquad \sum_{j=1}^{s} f(a_j) = 1.$$

Graphische Darstellungen der Häufigkeitsverteilung

Die gängigen graphischen Darstellungen von Häufigkeitsverteilungen sind

- Tabellen
- Stabdiagramme und Histogramme
- Kreisdiagramme.

Beispiel Stimmenverteilung bei der Bundestagswahl 2005

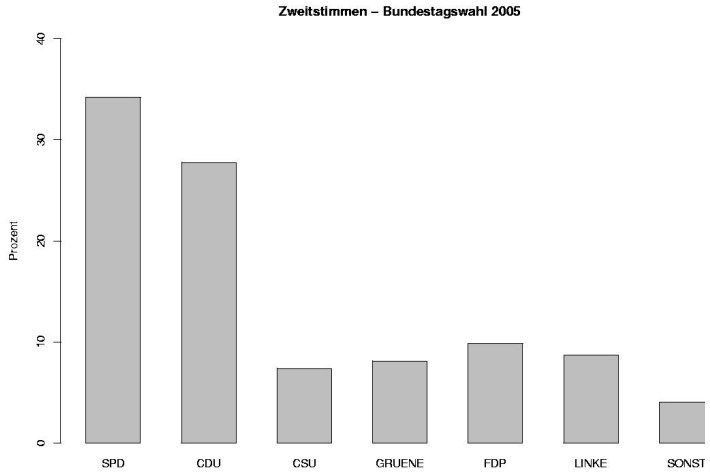
Das erhobene Merkmal ist in diesem Falle die mit der Zweitstimme gewählte Partei. Eine Beobachtungseinheit ist ein Stimmzettel. Die Gesamtheit der Merkmalswerte sind die zur Wahl stehenden Parteien, also SPD, CDU, CSU, usw. Um die Darstellung zu vereinfachen, sind die weniger häufig gewählten Parteien in der Klasse "Sonstige" zusammengefasst. Die Anzahl n der Merkmalswerte ist gleich der Anzahl der gültigen Zweitstimmen, in diesem Falle $n = 47\,287\,988$.

Häufigkeitstabelle

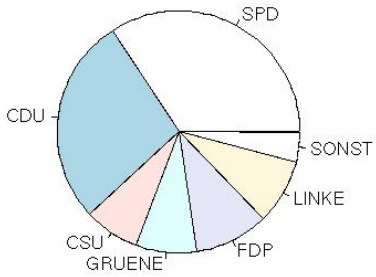
In der Häufigkeitstabelle werden die ermittelten absoluten und/oder relativen Häufigkeiten tabellarisch erfasst.

Partei	Zweitstimmen	Anteil in Prozent
SPD	16 194 665	34.2
CDU	13 136 740	27.8
CSU	3 494 309	7.4
Grüne	3 838 326	8.1
FDP	4 648 144	9.8
Die Linke	4 118 194	8.7
Sonstige	1 912 665	4.0

Stabdiagramm



Kreisdiagramm

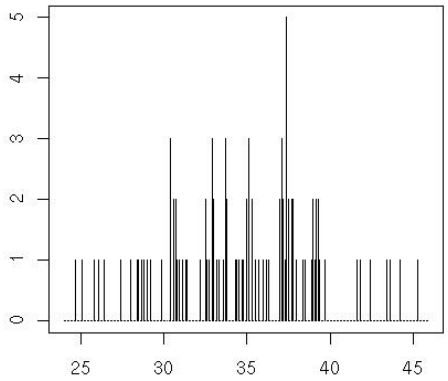


Bei stetigen oder quasistetigen Merkmalen ist die Aufstellung einer Häufigkeitstabelle oder eines Stabdiagramms sinnlos, denn die meisten Werte sind nur einfach oder gar nicht besetzt.

Beispiel

Jährliche Milchleistung von Kühen (in 100 Litern) (n=100).

37.4	37.8	29.0	35.1	30.9	28.5	38.4	34.7	36.3	30.4
39.1	37.3	45.3	32.2	27.4	37.0	25.1	30.7	37.1	37.7
26.4	39.7	33.0	32.5	24.7	35.1	33.2	42.4	37.4	37.2
37.5	44.2	39.2	39.4	43.6	28.0	30.6	38.5	31.4	29.9
34.5	34.3	35.0	35.5	32.6	33.7	37.7	35.3	37.0	37.8
32.5	32.9	38.0	36.0	35.3	31.3	39.3	34.4	37.2	39.0
41.8	32.7	33.6	43.4	30.4	25.8	28.7	31.1	33.0	39.0
37.1	36.2	28.4	37.1	37.4	30.8	41.6	33.8	35.0	37.4
33.7	33.8	30.4	37.4	39.3	30.7	30.6	35.1	33.7	32.9
35.7	32.9	39.2	37.5	26.1	29.2	34.8	33.3	28.8	38.9



Stabdiagramm

Ein Ausweg liefert hier die **Klassierung**. Bei der Wahl der Anzahl der Klassen ist allerdings zu beachten, dass

- bei zu großer Klassenanzahl viele Klassen unbesetzt bleiben,
- bei zu geringer Klassenanzahl Information verloren geht.

Als **Faustregel** gilt, dass die Anzahl der Klassen in etwa $\sqrt{n}$ entsprechen sollte, wobei n die Anzahl der Beobachtungen ist.
In obigem Beispiel erhalten wir bei der Wahl von 8 Klassen der Form

$$[a_1, a_2[,\ [a_2, a_3[,\ [a_3, a_4[,\ [a_4, a_5[,\ [a_5, a_6[,\ [a_6, a_7[,\ [a_7, a_8[,\ [a_8, a_9[$$

mit $a_1 = 24$, $a_2 = 27$, $a_3 = 29.6$, $a_n = 32$, $a_5 = 34.3$, $a_6 = 36.5$, $a_7 = 38.4$, $a_8 = 40.5$, $a_9 = 45.5$ die folgende Häufigkeitstabelle:

Milchleistung	[24, 27[	[27, 29.6[	[29.6, 32[	[32, 34.3[
Anzahl der Milchkühe	5	8	13	18
Milchleistung	[34.3, 36.5[	[36.5, 38.4[	[38.4, 40.5[	[40.5, 45.5[
Anzahl der Milchkühe	17	20	12	7

Im folgenden bezeichne K_j die Anzahl der Merkmalswerte in der Klasse $[a_j, a_{j+1}[$. K_j heißt **Klassenhäufigkeit** oder auch **Besetzungszahl**. Den zugehörigen relativen Anteil

$$k_j := \frac{K_j}{n}$$

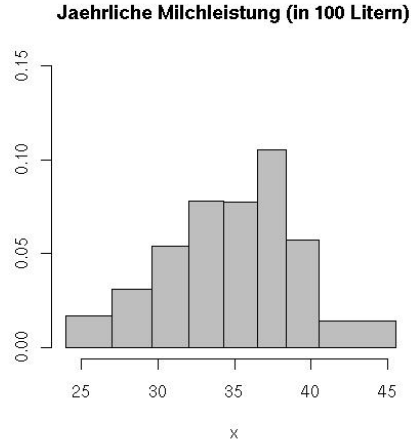
bezeichnet man als **relative Klassenhäufigkeit**.

Zur graphischen Darstellung klassierter Daten eignen sich **Histogramme**. Hierbei wird über jedem der Teilintervalle $[a_j, a_{j+1}[$ ein Rechteck mit der Fläche k_j errichtet. Die Höhe d_j des Rechtecks errechnet sich also gemäß der folgenden Gleichung

$$d_j(a_{j+1} - a_j) = k_j\,.$$

Man beachte, dass bei **gleicher Klassenbreite** nicht nur die Fläche, sondern **auch die Höhe** der Rechtecke proportional zur relativen Klassenhäufigkeit k_j ist.

Histogramm zu obigem Beispiel



Kumulierte Häufigkeitsverteilung

Die Funktion

$$H(x) := \sum_{a_j \leq x} h(a_j) \qquad \text{für } x \in \mathbb{R}$$

heißt **absolute kumulierte Häufigkeitsverteilung**. Sie zählt zu gegebenem $x \in \mathbb{R}$ die Anzahl der Beobachtungswerte die kleiner gleich x sind. Die Funktion

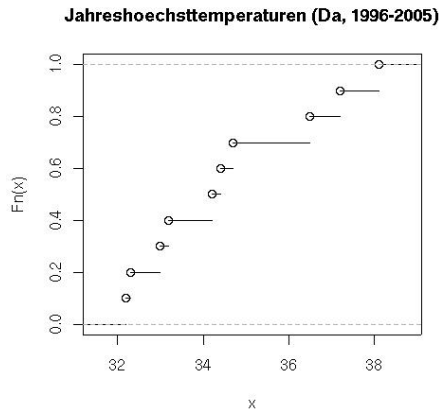
$$F(x) := \frac{1}{n} H(x) = \sum_{a_j \leq x} f(a_j)\,, \qquad x \in \mathbb{R}$$

heißt **relative kumulierte Häufigkeitsverteilung** oder **empirische Verteilungsfunktion**.

Eigenschaften der empirischen Verteilungsfunktion

- F ist eine monoton wachsende Treppenfunktion
- $0 \leq F \leq 1$
- F besitzt Sprünge an den Merkmalsausprägungen a_j

Als Beispiel für den typischen Verlauf einer empirischen Verteilungsfunktion im folgenden die Verteilungsfunktion zu den Jahreshöchsttemperaturen in Darmstadt aus den Jahren 1996-2005.



Lagemaße

Modalwert x_{Mod}

Diejenigen Ausprägungen a_j mit der größten Häufigkeit werden als **Modalwerte** bezeichnet. Die Verwendung des Modalwertes zur Beschreibung von Datensätzen sollte auf den Fall unimodaler Verteilungen beschränkt bleiben.

Median x_{Med}

Der **Median** oder auch **Zentralwert** ist derjenige Wert x_{Med}, für den mindestens 50 % aller Merkmalswerte kleiner gleich x_{Med} und mindestens 50 % aller Merkmalswerte größer gleich x_{Med} sind.
Zur Bestimmung des Medians ordnet man die Werte x_1, ..., x_n zunächst der Größe nach an,

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

und erhält auf diese Weise die sogannte **geordnete Urliste**. Dann definiert man

$$x_{Med} := \begin{cases} x_{(\frac{n+1}{2})} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{falls } n \text{ gerade} \end{cases} \tag{1.1}$$

Arithmetisches Mittel (Durchschnittswert)

Der bekannteste Lageparameter ist das arithmetische Mittel

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{s} a_j f(a_j)\,.$$

Beispiel Preise für Normal-Benzin an 20 örtlichen Tankstellen der Größe nach geordnet:

129.4	129.9	129.9	130.4	131.4
131.4	132.9	132.9	132.9	133.9
134.4	134.4	134.9	134.9	134.9
134.9	135.4	135.4	135.9	136.4

In diesem Beispiel ist $x_{Mod} = 134.9$, $x_{Med} = 134.15$, $\overline{x} = 133.325$. Würde eine Tankstelle als besondere Werbemaßnahme den Benzinpreis von 132.9 auf 125.9 senken, so würde dies den Durchschnittswert $\overline{x}$ von 133.325 auf 132.975 senken. Einen Einfluss auf den Median (oder auf den Modalwert) hätte die Senkung dagegen nicht.

Lagemaße, die nicht empfindlich auf Extremwerte oder Ausreißer reagieren heißen **robust**. Der Median ist also ein robustes Lagemaß.

Bemerkung

(i) Median und arithmetisches Mittel stimmen i.a. nicht mit einer der möglichen Merkmalsausprägungen überein.

Prominentes Beispiel: Durchschnittliche Anzahl der Kinder pro Familie.

(ii) **Äquivarianz unter linearer Transformation** Transformiert man die Daten gemäß einer affin linearen Transformation der Form

$$y_i = a + bx_i\,,$$

so gilt für das arithmetische Mittel

$$\overline{y} = a + b\overline{x}$$

und ebenso

$$y_{Mod} = a + bx_{Mod}\,, \quad y_{Med} = a + bx_{Med}\,.$$

(iii) **Optimalitätseigenschaften** Das arithmetische Mittel $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ **minimiert** die **Summe der quadratischen Abstände**, d.h. es gilt

$$\sum_{i=1}^n (x_i - \overline{x})^2 < \sum_{i=1}^n (x_i - r)^2 \text{ für alle } r \in \mathbb{R}, r \neq \overline{x}\,.$$

Beweis

$$\begin{aligned}\sum_{i=1}^n (x_i - r)^2 - \sum_{i=1}^n (x_i - \overline{x})^2 &= \sum_{i=1}^n \underbrace{(x_i - r)^2 - (x_i - \overline{x})^2}_{-2x_i r + r^2 + 2x_i\overline{x} - \overline{x}^2} \\ &= -2n\overline{x}r + nr^2 + 2n\overline{x}^2 - n\overline{x}^2 \\ &= n(r - \overline{x})^2 > 0 \text{ für } r \neq \overline{x}\,.\end{aligned}$$

Auch Median und Modalwert erfüllen entsprechende Optimalitätskriterien.

- Der Median x_{Med} minimiert die Summe der Abstände, d.h. es gilt

$$\sum_{i=1}^n |x_i - x_{Med}| < \sum_{i=1}^n |x_i - r| \text{ für alle } r \in \mathbb{R}\,, r \neq x_{Med}\,.$$

- Der Modalwert minimiert die Summe

$$\sum_{i=1}^n 1_{\{x_i \neq r\}} \text{ mit } 1_{\{x_i \neq r\}} = \begin{cases} 1 & \text{falls } x_i \neq r \\ 0 & \text{falls } x_i = r\,. \end{cases}$$

Weitere Lagemaße

Annahme: $x_1, \ldots, x_n > 0$

Geometrisches Mittel $\bar{x}_{geom}$

$$\bar{x}_{geom} := (x_1 \cdot \ldots \cdot x_n)^{\frac{1}{n}}$$

Findet Verwendung im Zusammenhang mit Wachstums- und Zinsmodellen. Sind etwa $x_1, \ldots, x_n$ die beobachteten Wachstumsfaktoren eines Portfolios mit Anfangsbestand K_0, so ist

$$K_n = K_0 \cdot x_1 \cdot \ldots \cdot x_n$$

der Bestand am Ende der Periode n. Schreibt man

$$K_n = K_0 \left(\underbrace{(x_1 \cdot \ldots \cdot x_n)^{\frac{1}{n}}}_{=\bar{x}_{geom}} \right)^n = K_0 \cdot \bar{x}_{geom}^n$$

so lässt sich $\bar{x}_{geom}$ als **mittlerer Wachstumsfaktor** über die n Perioden $1, \ldots, n$ interpretieren.

Beziehung zum arithmetischen Mittel

Logarithmiert man die Messwerte $y_i := \ln x_i$ so folgt

$$\ln \bar{x}_{geom} = \frac{1}{n} \ln(x_1 \cdot \ldots \cdot x_n) = \frac{1}{n} \sum_{i=1}^{n} \ln x_i = \frac{1}{n} \sum_{i=1}^{n} y_i$$

d.h., $\ln \bar{x}_{geom}$ stimmt mit dem arithmetischen Mittel der logarithmierten Messwerte $y_i = \ln x_i$ überein.

Harmonisches Mittel $\bar{x}_{harm}$

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

Typische Anwendung: Ermittlung von Gesamtdurchschnittswerten aus Durchschnitten über einzelne Teilbereiche.

Beispiel Der ICE von Frankfurt nach Berlin fährt

- 150 km mit durchschnittlich 100 km pro Stunde
- 450 km mit durchschnittlich 250 km pro Stunde

Es sei x_i die Durchschnittsgeschwindigkeit bei Kilometer i, $i = 1, \ldots 600$. Dann beträgt die Durchschnittsgeschwindigkeit über die gesamte Strecke

$$\frac{1}{\frac{1}{600} \left(\frac{150}{100} + \frac{450}{250} \right)} = 160 \left[\frac{km}{h} \right] .$$

Quantile und Box-Plots

Lagemaße alleine reichen zur Beschreibung der Daten einer Urliste nicht aus. Vergleicht man etwa eine Einkommenserhebung in zwei Ländern, so können die Durchschnittseinkommen gleich sein, jedoch in einem Land größere Einkommensunterschiede bestehen als im anderen Land. Daher benötigt man zusätzliche Kennzahlen, um die Lage der Daten möglichst effizient erfassen zu können. Eine wichtige Methode sind **Box-Plots**, die mit Hilfe von Quantilen definiert werden.

Definition Es sei $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ eine geordnete Urliste und $p \in]0,1]$. Jeder Wert x_p mit der Eigenschaft

$$\frac{1}{n}(\text{ Anzahl der Messwerte } \leq x_p) \geq p$$

und

$$\frac{1}{n}(\text{ Anzahl der Messwerte } \geq x_p) \geq 1-p.$$

heißt p-**Quantil**.

Damit folgt

$$x_p = x_{([np]+1)} \text{ falls } np \text{ nicht ganzzahlig}$$
$$x_p \in [x_{(np)}, x_{(np+1)}] \text{ falls } np \text{ ganzahlig.}$$

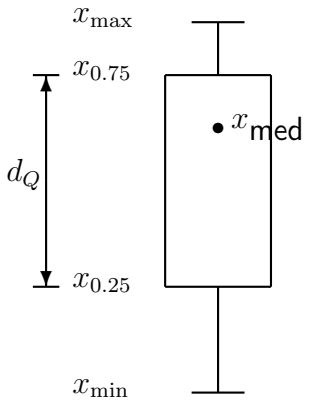
Der Median x_{Med} ist also insbesondere ein $\frac{1}{2}$-Quantil.

Spezialfälle

$x_{0.25} =$ Unteres Quartil $\qquad$ $x_{0.75} =$ Oberes Quartil

Die Distanz $d_Q = x_{0.75} - x_{0.25}$ heißt **Quartilsabstand**.

Aufbau eines zugehörigen **Box-Plots**



Modifikationen

Die Länge der Linien (engl. "whiskers", Barthaare) ober- bzw. unterhalb der Box können variieren. Eine gängige Variation besteht darin, die untere von

$$\max\{x_{0.25} - 1.5 * d_Q, x_{\min}\} \text{ bis } x_{0.25}$$

und die obere von

$$x_{0.75} \text{ bis } \min\{x_{0.75} + 1.5 * d_Q, x_{\max}\}$$

zu führen. Messwerte, die darunter bzw. darüber liegen, können gegebenenfalls als Ausreißer durch einzelne Punkte explizit kenntlich gemacht werden.

Streumaße

Neben der absoluten Lage der Messdaten ist auch ihre Streuung von großer Bedeutung. Die bekannteste Maßzahl für die Streuung einer Messreihe ist die **empirische Varianz** oder auch **mittlere quadratische Abweichung**:

$$s^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{j=1}^{s}(a_j - \overline{x})^2 f(a_j). \tag{1.2}$$

Sie ist also definiert als das arithmetische Mittel der quadratischen Abstände der einzelnen Messwerte zu ihrem Mittelwert. Die Wurzel hieraus

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

heißt **Standardabweichung**.

Der Zusammenhang zwischen der Standardabweichung s und der Streuung der Messwerte kann folgendermaßen präzisiert werden:

Für $k \geq 1$ liegen mindestens $100 \cdot \left(1 - \frac{1}{k^2}\right)$ Prozent der Messwerte $x_1, \ldots, x_n$ im Intervall $[\bar{x} - ks, \bar{x} + ks]$. Insbesondere:
im Intervall

- $[\overline{x} - \sqrt{2}s, \overline{x} + \sqrt{2}s]$ liegen mindestens 50 % der Daten
- $[\overline{x} - 2s, \overline{x} + 2s]$ liegen mindestens 75 % der Daten
- $[\overline{x} - 3s, \overline{x} + 3s]$ liegen mindestens 90 % der Daten.

Begründung der Abschätzung: Es reicht zu zeigen, dass

$$H := \text{Anzahl der } x_i \text{ mit } |x_i - \bar{x}| > k \cdot s$$

kleiner gleich $\frac{n}{k^2}$ ist. Zur Abschätzung von H beachte man, dass

$$H = \sum_{i=1}^{n} 1_{\{|x_i - \bar{x}| > k \cdot s\}} \text{ mit } 1_{\{|x_i - \bar{x}| > k \cdot s\}} = \begin{cases} 1 & \text{falls } |x_i - \bar{x}| > k \cdot s \\ 0 & \text{falls } |x_i - \bar{x}| \leq k \cdot s\,. \end{cases}$$

Offensichtlich gilt nun aber

$$\sum_{i=1}^{n} 1_{\{|x_i - \bar{x}| > k \cdot s\}} \leq \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{k \cdot s}\right)^2 = \frac{1}{k^2 \cdot s^2} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{= n \cdot s^2} = \frac{n}{k^2}\,.$$

Diese Abschätzung ist allgemein gültig und daher in vielen Fällen sehr ungenau. Wir werden später im Zusammenhang mit einem wahrscheinlichkeitstheoretischen Resultat sehen: Ist das Merkmal in etwa normalverteilt, so gilt:

im Intervall

- $[\overline{x} - s, \overline{x} + s]$ liegen etwa 68 % der Daten
- $[\overline{x} - 2s, \overline{x} + 2s]$ liegen etwa 95 % der Daten
- $[\overline{x} - 3s, \overline{x} + 3s]$ liegen etwa 99 % der Daten.

Diese Abschätzung ist also deutlich besser!

Bemerkung

In der induktiven Statistik verwendet man statt (1.2) die modifizierte Form

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

Sie heißt **Stichprobenvarianz** und ist in vielen Statistikprogrammpaketen voreingestellt. Für großen Stichprobenumfang n ist der Unterschied zwischen den beiden Normalisierungsfaktoren $\frac{1}{n}$ und $\frac{1}{n-1}$ vernachlässigbar.
Die Normierung mit $\frac{1}{n-1}$ statt mit $\frac{1}{n}$ liegt darin begründet, dass die Beziehung $\sum_{i=1}^{n} x_i - \overline{x} = 0$ eine der Abweichungen $x_i - \overline{x}$ bereits durch die übrigen $n-1$ eindeutig festlegt. Die Anzahl der Freiheitsgrade in der Summe $\sum_{i=1}^{n} (x_i - \overline{x})^2$ beträgt also $n-1$ und nicht n.

Eigenschaften der empirischen Varianz

(i) **Transformationsregel** Werden die Daten gemäß

$$y_i = a + bx_i$$

linear transformiert, so folgt für die empirische Varianz $s_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_1 - \overline{y})^2$ der transformierten Daten

$$s_y^2 = b^2 s_x^2 \,.$$

Beweis

$$s_y^2 = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(y_i - \overline{y})^2}_{(a+bx_i)-(a+b\overline{x})} = b^2 \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \square$$

Insbesondere folgt für die Standardabweichungen:

$$s_y = |b| \, s_x.$$

(ii) **Verschiebungssatz**

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

denn

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} \underbrace{(x_i - \overline{x})^2}_{=x_i^2 - 2x_i\overline{x} + \overline{x}^2} = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - 2\frac{1}{n} \sum_{i=1}^{n} x_i \overline{x} + \overline{x}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2 \,.$$

Konzentrationsmaße

Als Ausgangspunkt betrachten wir folgende aus [2] entnommene Statistik zu monatlichen Umsätzen der Möbelbranche in 1000 Euro in den drei Städten G, M und V:

Einrichtungshäuser	G	M	V
1	40	180	60
2	40	5	50
3	40	5	40
4	40	5	30
5	40	5	20

In der Stadt G ist der Umsatz unter den 5 Möbelhäusern also ausgeglichen, während in der Stadt M ein Möbelhaus quasi eine Monopolstellung besitzt. Zur Quantifizierung solcher Konzentrationen gibt es Konzentrationsmaße. Zur Diskussion solcher Maße betrachten wir folgende Ausgangsposition:

Gegeben sei ein kardinalskaliertes Merkmal mit nichtnegativen Merkmalsausprägungen. Weiterhin sei $x_1 \leq x_2 \leq ... \leq x_n$ eine bereits geordnete Stichprobe der Länge n mit positiver Merkmalssumme $\sum_{i=1}^{n} x_i > 0$.

Lorenzkurve

Es sei

$$v_k := \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{n} x_i} \qquad k = 0, 1, 2, \ldots, n$$

der Anteil der k kleinsten Merkmalsträger an der gesamten Merkmalssumme. Trägt man die Punkte

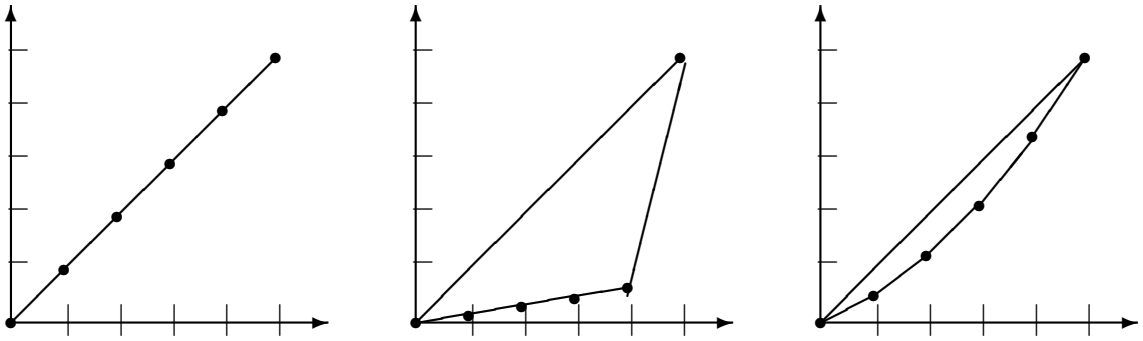
$$\left(\frac{k}{n}, v_k\right), k = 0, 1, 2, \ldots, n$$

in das Einheitsquadrat ein und verbindet sie durch einen Streckenzug, so erhält man die zugehörige **Lorenzkurve**.

In obigem Beispiel erhält man:

	Stadt G	Stadt M	Stadt V
k	v_k	v_k	v_k
1	0.2	0.025	0.10
2	0.4	0.050	0.25
3	0.6	0.075	0.45
4	0.8	0.100	0.70
5	1.0	1.0	1.0

Man erhält als zugehörige **Lorenzkurven**



Stadt G Stadt M Stadt V

Eigenschaften der Lorenzkurve

- Die Lorenzkurve ist immer monoton wachsend und konvex (d.h. nach unten gewölbt).
- Die Stärke der Wölbung, also ihre Abweichung von der Winkelhalbierenden, ist ein Maß für Konzentration. Verläuft die Kurve auf der Winkelhalbierenden, so liegt ein ausgewogener Markt vor.

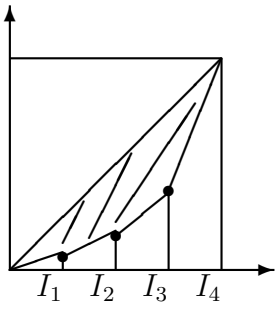
Der **Gini-Koeffizient** G ist definiert durch

$$\begin{aligned} G &= \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Fläche zwischen Diagonale und horizontaler Achse}} \\ &= 2 \cdot \text{ Fläche zwischen Diagonale und Lorenzkurve} \end{aligned}$$

Für die Berechnung des Gini-Koeffizienten gilt die folgende Formel:

$$G = \frac{2}{n} \frac{\sum_{i=1}^{n} i x_i}{\sum_{i=1}^{n} x_i} - \frac{n+1}{n}.$$

Beweis



Die Fläche der I_i beträgt gerade

$$I_i = \frac{1}{n} v_{i-1} + \frac{1}{2n}(v_i - v_{i-1})$$

also summiert sich die Gesamtfläche der I_i zu

$$\frac{1}{n}\sum_{i=1}^{n} v_{i-1} + \frac{1}{2n}\underbrace{\sum_{i=1}^{n}(v_i - v_{i-1})}_{=v_n - v_0 = 1} = \frac{1}{n}\sum_{i=1}^{n-1} v_i + \frac{1}{2n}\,.$$

Beachtet man noch, dass

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n-1} v_i &= \frac{1}{n}\frac{1}{\sum_{j=1}^{n} x_j}\left(\sum_{i=1}^{n-1}\sum_{k=1}^{i} x_k\right)\\ &= \frac{1}{n}\frac{1}{\sum_{j=1}^{n} x_j}\sum_{k=1}^{n}(n-k)x_k = 1 - \frac{1}{n}\frac{\sum_{k=1}^{n} k x_k}{\sum_{j=1}^{n} x_j}\end{aligned}$$

so erhält man nach Einsetzen in die obere Gleichung

$$G = 2\left(\frac{1}{2} - \left(1 - \frac{1}{n}\frac{\sum_{j=1}^{n} j x_j}{\sum_{j=1}^{n} x_j} + \frac{1}{2n}\right)\right) = \frac{2}{n}\frac{\sum_{j=1}^{n} j x_j}{\sum_{j=1}^{n} x_j} - \frac{n+1}{n}\,. \quad \square$$

3. Auswertung zwei- und mehrdimensionaler Messreihen

Zweidimensionale Messreihen

Werden bei einer Erhebung zwei Merkmale X und Y zugleich erhoben, so besteht die Urliste aus **Wertepaaren**

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Typische Fragestellungen im Zusammenhang zweier Merkmale sind die nach Abhängigkeiten/Unabhängigkeiten zwischen den beiden erhobenen Merkmalen. Zur Darstellung der zweidimensionalen Daten gibt es zunächst zwei Möglichkeiten:

- **Kontingenztabelle:** geeignet für nominalskalierte Merkmale
- **Streuungsdiagramm:** geeignet für kardinalskalierte Merkmale

(A) Kontingenztabelle

Bei diesem Verfahren werden die absoluten Häufigkeiten der möglichen Paare von Ausprägungen des Merkmals x und des Merkmals y tabellarisch aufgelistet:

	Ausprägungen von Y		
Ausprägungen von X	b_1	$\ldots$	b_l
a_1	h_{11}	$\ldots$	h_{1l}
$\vdots$	$\vdots$		$\vdots$
a_k	h_{k1}	$\ldots$	h_{kl}

Hierbei steht $h_{ij} = h(a_i, b_j)$ für die absolute Häufigkeit der Wertepaare (a_i, b_j).

Beispiel (entnommen aus [1])

Zur Untersuchung von Abhängigkeiten zwischen Berufsgruppen und sportlicher Betätigung werden 1000 Personen befragt. Es entstand dabei folgende **Kontingenztabelle:**

	sportl. Bet.		
	nie	gelegentlich	regelmäßig
Arbeiter	240	120	70
Angestellter	160	90	90
Beamter	30	30	30
Landwirt	37	7	6
sonst. freier Beruf	40	32	18

Die Einträge in der Kontigenztabelle heißen **gemeinsame Häufigkeiten**. Statt der absoluten, lassen sich hier natürlich auch die relativen Häufigkeiten betrachten:

$$f_{ij} = f(a_i, b_j) = \frac{h_{ij}}{n}.$$

Fragt man nach der absoluten Häufigkeit einer Merkmalsausprägung a_i (bzw.b_j) so hat man die gemeinsamen Häufigkeiten h_{ij} der entsprechenden Zeile (bzw. der entsprechenden Spalte) aufzusummieren:

$$h(a_i) = h_{i\cdot} := \sum_{j=1}^{l} h_{ij}$$

$$h(b_j) = h_{\cdot j} := \sum_{i=1}^{k} h_{ij}$$

Diese Häufigkeiten werden auch als **Randhäufigkeiten** bezeichnet.

In obigem Beispiel

	sportl. Bet.			
	nie	gelegentlich	regelmäßig	Randhäufigkeiten
Arbeiter	s.o.	s.o.	s.o.	430
Angesteller				340
Beamter				90
Landwirt				50
sonst. freier Beruf				90
Randhäufigkeiten	507	279	214	1000

Um nun die beiden Merkmale auf Abhängigkeit/Unabhängigkeit hin zu untersuchen, bildet man die **bedingten relativen Häufigkeiten**

$$f_X(a_i|b_j) := \frac{h_{ij}}{h_{\cdot j}} \text{ der Ausprägung } a_i \text{ gegeben die Ausprägung } b_j$$

und

$$f_Y(b_j|a_i) = \frac{h_{ij}}{h_{i\cdot}} \text{ der Ausprägung } b_j \text{ gegeben die Ausprägung } a_i\,.$$

Die bedingte relative Häufigkeit $f_X(a_i|b_j)$ gibt also die relative Häufigkeit der Ausprägung a_i an unter allen Merkmalsträgern, die bzgl. des anderen Merkmals die Ausprägung b_j besitzen. Sind die bedingten relativen Häufigkeiten

$$f_X(a_1|b_j), f_X(a_2|b_j), ..., f_X(a_k|b_j)$$

der Ausprägung $a_1, ..., a_k$ des ersten Merkmals unabhängig von b_j (also gleich für $j = 1, \ldots, l$), so beeinflussen sich die Merkmale nicht und man sagt, dass sie **unabhängig** sind.

Dieser Fall tritt genau dann ein, wenn auch die umgekehrten bedingten relativen Häufigkeiten

$$f_Y(b_1|a_i), f_Y(b_2|a_i), ..., f_Y(b_l|a_i)$$

unabhängig sind von a_i für $i = 1, \ldots, k$.

Im Falle der Unabhängigkeit gilt insbesondere

$$f_X(a_i|b_{j1}) = f_X(a_i|b_{j2})$$

und damit

$$h_{ij_1} \cdot h_{\cdot j_2} = h_{ij_2} \cdot h_{\cdot j_1}$$

Summation über $j_1 = 1, \ldots, l$ ergibt

$$h_{i\cdot} h_{\cdot j_2} = h_{ij_2} \cdot n$$

also

$$h_{ij_2} = \frac{h_{i\cdot} h_{\cdot j_2}}{n}$$

und somit - da j_2 beliebig:

$$h_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}\,. \tag{1.3}$$

Die **gemeinsamen Häufigkeiten** sind in diesem Falle über (1.3) also bereits durch die **Randhäufigkeiten** bestimmt.

Für die bedingten relativen Häufigkeiten folgt hieraus insbesondere

$$f_X(a_i|b_j) = \frac{h_{ij}}{h_{\cdot j}} = \frac{h_{i\cdot}}{n} \qquad \text{bzw. } f_Y(b_j|a_i) = \frac{h_{ij}}{h_{i\cdot}} = \frac{h_{\cdot j}}{n}\,,$$

sie sind also unabhängig von der Ausprägung des jeweils anderen Merkmals.

Der Kontingenzkoeffizient

Um die Abhängigkeit zwischen zwei Merkmalen X und Y quantitativ erfassen zu können, bildet man die folgende, als **Chi-Quadrat Koeffizient**, bezeichnete Größe:

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

Hierbei ist $\tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n}$.

χ^2 ist genau dann 0, wenn die Merkmale unabhängig sind, also wenn $h_{ij} = \tilde{h}_{ij}$ gilt. Je kleiner also der χ^2-Koeffizient, umso stärker spricht dies für die Unabhängigkeit der beiden Merkmale X und Y. Allerdings hängt die Größenordnung des χ^2-Koeffizienten von der Dimension der Kontingenztafel ab. Daher geht man vom χ^2-Koeffizienten über zum **Kontingenzkoeffizienten**

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}}.$$

Der Kontingenzkoeffizient K nimmt Werte an zwischen 0 und

$$K_{max} = \sqrt{\frac{M-1}{M}}, \text{ wobei } M = \min\{k, l\}\,.$$

Durch Normierung mit K_{max} erhält man hieraus schließlich den **normierten Kontingenzkoeffizienten**

$$K_* = \frac{K}{K_{max}}.$$

Beispiel (obiges Beispiel zum Zusammenhang zwischen Berufstätigkeit und sportlicher Betätigung)

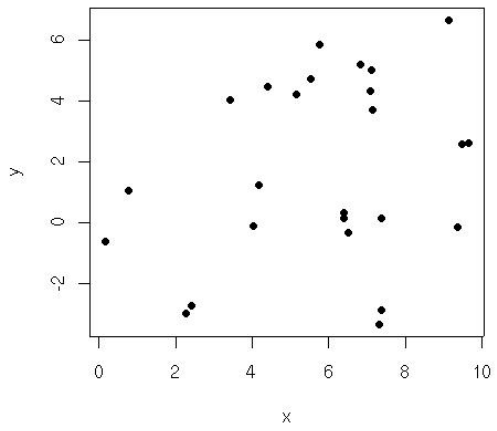
In diesem Falle ist $\chi^2 = 38.55412$ und wegen $n = 1000$ folgt für den Kontingenzkoeffizienten $K = 0.192673$ sowie wegen $k = 5$, $l = 3$, also $M = \min\{k, l\} = 3$, folgt für den normierten Kontingenzkoeffizienten $K_* = 0.2359753$.

(B) Streuungsdiagramm

Bei kardinalskalierten Merkmalen kann man die Wertepaare

$$(x_1, y_1), \ldots, (x_n, y_n)$$

der Urliste als Punkte der Ebene auffassen und somit ein zugehöriges **Streuungsdiagramm** erstellen:



Beispiel

In einem Krankenhaus wurden von 5 Neugeborenen Körperlänge X und Kopfumfang Y (in cm) gemessen. Es ergab sich folgende nach Köperlänge geordnete Messreihe:

(48.6, 35.1), (49.5, 34.1), (50.7, 36.8), (51.1, 35.7), (52.4, 37.4)

Zu den jeweiligen Messwerten bildet man zunächst die beiden Mittelwerte

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Im Beispiel $\overline{x} = \frac{1}{5}$ 252.3 = 50.46, $\overline{y} = \frac{1}{5}$ 179.1 = 35.82.

Liegt bei einem Wertepaar (x_i, y_i) der erste Wert um den Durchschnitt $x_i \sim \overline{x}$, aber der zweite Wert y_i deutlich über oder unter dem Durchschnitt $\overline{y}$, so spricht dies eher

für die Unkorreliertheit der beiden Merkmale Körperlänge X und Kopfumfang Y. Liegen jedoch bei diesem Wertepaar bei beiden Merkmalen deutliche Abweichungen vom Durchschnitt vor, so spricht dies für Korrelation. Folglich liefert das Produkt

$$(x_i - \overline{x})(y_i - \overline{y})$$

einen brauchbaren Ansatz für ein Korrelationsmaß.

Aufsummieren über die gesamte Stichprobe und Normierung ergibt die **empirische Kovarianz**

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Nach Normierung mit den jeweiligen Standardabweichungen

$$s_X = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{\frac{1}{2}} \text{ und } s_Y = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{\frac{1}{2}}$$

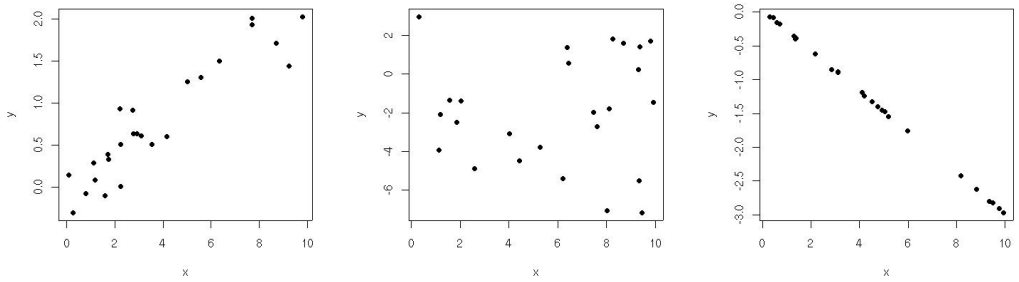
erhält man den **empirischen Korrelationskoeffizienten**

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Eigenschaften

- $-1 \leq r_{XY} \leq 1$
- $r_{XY} = -1$ (bzw. $r_{XY} = +1$) genau dann wenn die Wertepaare (x_i, y_i) auf einer Geraden mit negativer (bzw. positiver) Steigung liegen.
- $r_{XY} = 0$ spricht für die Unkorreliertheit der Merkmale X und Y. In diesem Falle sind die Wertepaare (x_i, y_i) "regellos" verteilt.
- Die Merkmale X und Y heißen
  - **positiv korreliert**, falls $r_{XY} > 0$
  - **negativ korreliert**, falls $r_{XY} < 0$.

$r_{XY} = 0.827$ $\qquad$ $r_{XY} = 0.046$ $\qquad$ $r_{XY} = -0.999$



- eine rechentechnisch günstigere Darstellung für den Korrelationskoeffizienten ist

$$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2)(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2)}}.$$

Regressionsrechnung

Liegen die Wertepaare der n Beobachtungen (x_i, y_i) annähernd auf einer Geraden, so kann man von einem **linearen Zusammenhang** der Form

$$y = a + bx \tag{1.4}$$

sprechen. Die Koeffizienten a und b wählt man dabei so, dass sich die zugehörige Gerade der gegebenen Punktwolke am besten anpasst. "Beste Anpassung" bedeutet dabei, dass die Summe der quadratischen Abstände

$$Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \,,$$

zwischen Messwert y_i und entsprechendem Punkt $a + bx_i$ auf der Geraden $y = a + bx$, minimal wird. ("Prinzip der kleinsten Quadrate" nach C.F. Gauß).

Diejenige Gerade, die sich der Punktwolke dabei am besten anpasst, heißt **Ausgleichsgerade** oder **Regressionsgerade**. Ihre Koeffizienten sind bestimmt durch

$$\hat{b} = \frac{s_{XY}}{s_X^2} \,, \quad \hat{a} = \bar{y} - \hat{b}\bar{x} \,. \tag{1.5}$$

Beispiel In obigem Beispiel ist

$$s_{XY} = \frac{1}{4}(9043.6 - 9037.386) \sim 1.55$$

und damit $r_{XY} \sim 0.8$ (d. h. Körpergröße und Kopfumfang sind (erwartungsgemäß) stark positiv korreliert). Die Koeffizienten der zugehörigen **Regressionsgeraden** sind gegeben durch

$$\hat{b} \sim 0.72 \text{ und } \hat{a} \sim -0.51$$

also hat die Regressionsgerade die Form

$$y = -0.51 + 0.72x \,.$$

Mit Hilfe der Regressionsgeraden können wir nun zum Beispiel einen Vorhersagewert ("Prognose") für den Kopfumfang eines Neugeborenen bei einer Körperlänge von 50 cm bestimmen: $y(50) = 35.49$.

Zu gegebenem Wertepaar (x_i, y_i) heißt die Differenz

$$u_i := y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i)$$

zwischen beobachtetem Wert y_i und dem durch die Regressionsgerade erklärten entsprechenden Wert $\hat{y}_i = \hat{a} + \hat{b}x_i$ **Residuum**. Den Quotienten

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r_{XY}^2$$

bezeichnet man als **Bestimmtheitsmaß**. Er ist ein Maß für die Güte der Approximation der Messwerte y_i durch die berechnete Ausgleichsgerade und stimmt mit dem Quadrat des Korrelationskoeffizienten überein.

Zur **Optimalität** der **Regressionsgeraden**

Satz Es sei $s_X^2 \neq 0$ und $\hat{a}$, $\hat{b}$ wie in (1.5). Dann gilt:

$$Q(a,b) > Q(\hat{a},\hat{b}) \qquad \text{für alle } (a,b) \neq (\hat{a},\hat{b})\,.$$

Beweis:

$$Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

ist ein Polynom vom Grad 2 mit Gradient

$$\begin{aligned}\text{grad } Q(a,b) &= \left(\frac{\partial Q}{\partial a}(a,b), \frac{\partial Q}{\partial b}(a,b)\right) \\ &= -2\left(\sum_{i=1}^{n}[y_i - (a+bx_i)], \sum_{i=1}^{n} x_i[y_i - (a+bx_i)]\right)\end{aligned}$$

und Hesse-Matrix

$$H_Q(a,b) = \begin{bmatrix} \frac{\partial^2 Q}{\partial a^2}(a,b) & \frac{\partial^2 Q}{\partial a \partial b}(a,b) \\ \frac{\partial^2 Q}{\partial a \partial b}(a,b) & \frac{\partial^2 Q}{\partial b^2}(a,b) \end{bmatrix} = 2\begin{bmatrix} n & n\overline{x} \\ n\overline{x} & \sum_{i=1}^{n} x_i^2 \end{bmatrix}.$$

Also

$$\det H_Q(a,b) = 4\left(n\sum_{i=1}^{n} x_i^2 - n^2\overline{x}^2\right) = 4n^2 s_X^2 > 0\,,$$

damit ist H_Q positiv definit und somit Q gleichmäßig strikt konvex.

Folglich besitzt Q genau ein eindeutig bestimmtes Minimum und dies wird an der "Nullstelle" (bzw. der kritischen Stelle) des Gradienten angenommen:

$$\begin{aligned}\text{grad } Q(a,b) = 0 &\Leftrightarrow \frac{\partial Q}{\partial a}(a,b) = 0 \text{ und } \frac{\partial Q}{\partial b}(a,b) = 0 \\ &\Leftrightarrow \overline{y} = a + b\overline{x} \text{ und} \\ &\qquad 0 = \sum_{i=1}^{n} x_i(y_i - (a+bx_i)) = \sum_{i=1}^{n} x_i(y_i - bx_i - (\overline{y} - b\overline{x})) \\ &\qquad = \sum_{i=1}^{n} x_i y_i - b\sum_{i=1}^{n} x_i^2 - n\overline{x}\overline{y} + nb\overline{x}^2 \\ &\Leftrightarrow a = \overline{y} - b\overline{x} \text{ und} \\ &\qquad b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{s_{XY}}{s_X^2} \quad \square\end{aligned}$$

Bemerkung (Nichtlineare Regression)

Bei vielen zweidimensionalen Messreihen ist von vorneherein klar, dass kein linearer Zusammenhang zwischen den beobachteten Messwerten erwartet werden kann, sondern ein funktionaler Zusammenhang der Form

$$y = f(x)$$

für eine geeignete **nichtlineare** Funktion f, z.B.

$$y = ae^{bx} \text{ für } b \in \mathbb{R}, a > 0.$$

Gesucht sind wieder diejenigen Parameter a und b, für die sich der zugehörige Funktionsgraph der gegebenen Punktwolke am besten anpasst. Häufig kann man durch geeignete Transformation der Daten das Problem auf einen linearen Zusammenhang zurückführen, wie etwa im Beispiel $y = ae^{bx}$

$$\log y = \log a + bx$$

und zu bestimmen ist die Regressionsgerade zu den transformierten Beobachtungswerten

$$(x_1, \log y_1), (x_2, \log y_2), ..., (x_n, \log y_n)\,.$$

Ausblick auf mehrdimensionale Messreihen

Bei einer statistischen Erhebung können natürlich mehr als zwei Merkmale zugleich erhoben werden. Als Urliste enstehen Tupel (d.h. geordnete Mengen) von Messwerten

$$(x_{11}, \ldots, x_{1m})\,, (x_{21}, \ldots, x_{2m})\,, \ldots (x_{n1}, \ldots, x_{nm})\,,$$

die man in einer **Datenmatrix** zusammenfasst:

$$\begin{bmatrix} x_{11} & \ldots & x_{1m} \\ x_{21} & \ldots & x_{2m} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nm} \end{bmatrix}$$

Die graphische Darstellung der Urliste als Streuungsdiagramm ist für $m \geq 4$ nicht mehr möglich. Zur Aufklärung von Abhängigkeiten zwischen den erhobenen Merkmalen könnte man zwar für jedes Paar von Merkmalen das zweidimensionale Streuungsdiagramm bzw. die zweidimensionale Kontingenztabelle aufstellen. Da aber die Anzahl der Merkmalspaare mit der Anzahl m der erhobenen Merkmale sehr schnell anwächst, ist dieser Ansatz sehr aufwändig. Effizientere Methoden sind Gegenstand weiterführender Veranstaltungen in der Statistik.

Teil II Wahrscheinlichkeitsrechnung

1. Zufallsexperimente und Wahrscheinlichkeitsräume

Unter einem **Zufallsexperiment** versteht man zunächst einmal einen zeitlich wie örtlich fest umrissenen Vorgang mit unbestimmtem Ausgang.

Beispiele

- Werfen eines Würfels oder Werfen einer Münze
- Wahlergebnis der nächsten Landtagswahl
- Temperatur oder Windgeschwindigkeit am Luisenplatz am 1. Dezember 2007, 12:00
- Körpergröße oder Kopfumfang eines Neugeborenen

Die Gesamtheit aller möglichen Ausgänge eines Zufallsexperiments heißt **Ergebnismenge** oder auch **Stichprobenraum** und wird mit Ω bezeichnet.
Ein Element $\omega \in \Omega$ heißt **Elementarereignis** oder **Stichprobe**. Es stellt einen möglichen Ausgang des zugrundeliegenden Zufallsexperiments dar.

Beispiele

(i) einmaliges Würfeln: $\Omega = \{1, 2, \ldots, 6\}$, $|\Omega| = 6$

(Hierbei bezeichnet $|\Omega|$ die **Mächtigkeit der Menge** Ω, also die Anzahl der Elemente in Ω.)

(ii) zweimaliges Würfeln:

$$\Omega = \{(i, j) : i, j \in \{1, \ldots, 6\}\} = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\} = \{1, 2, \ldots, 6\}^2$$

also $|\Omega| = 36$.

(iii) Münzwurf: $\Omega = \{$ Kopf, Zahl $\}$.

(iv) Autos am Darmstädter Kreuz am 25. August 2007: $\Omega = \{0, 1, 2, 3, \ldots\} = \mathbb{N} \cup \{0\}$

(v) Temperatur in Grad Kelvin am Luisenplatz am 1. Dezember 2007, 12 Uhr Mittags: $\Omega = [0, \infty[$ oder realistischer $[250, 290]$ $(O°C = 273.15°K)$

In den ersten vier Fällen sind die Ergebnisräume **endlich** oder **abzählbar unendlich**. Solche Ergebnisräume nennt man auch **diskret**. Im fünften Fall ist der Ergebnisraum nicht mehr abzählbar, sondern eine **kontinuierliche** Menge.
Die Wahrscheinlichkeitstheorie zu kontinuierlichen Ergebnisräumen ist mathematisch anspruchsvoller als die zu diskreten Ergebnisräumen. Daher betrachten wir **zunächst nur diskrete** Ergebnisräume Ω.

Ereignisse

Teilmengen $A \subset \Omega$ von Ω heißen **Ereignisse**. Die Gesamtheit aller Ereignisse ist somit nichts weiter als $\mathcal{P}(\Omega)$, also die **Potenzmenge** von Ω. Unter der Potenzmenge von Ω versteht man

die Gesamtheit aller Teilmengen von Ω einschließlich der leeren Menge $\emptyset$ und der Menge Ω selber.
Beachten Sie: Ereignisse sind Elemente der Potenzmenge $\mathcal{P}(\Omega)$ von Ω, also Teilmengen von Ω, während Elementarereignisse Elemente von Ω sind.

Beispiele

(i) $A = \{1, 3, 5\} =$ `Augenzahl ungerade`

(ii) $A = \{(5, 6), (6, 5), (6, 6)\} =$ `Augensumme > 10`

(iv) $A = \{22.000, 22.001, \ldots\} = \{n : n \geq 22.000\} =$ `ungewöhnlich hohes Verkehrsaufkommen`

Zwei Ereignisse sind besonders hervorzuheben:

- Ω = das sichere Ereignis
- $\emptyset$ = das unmögliche Ereignis.

Die bekannten Mengenoperationen lassen sich als **Operationen auf Ereignissen** interpretieren:

$A \cup B =$ A `oder` B `tritt ein`

$A_1 \cup A_2 \cup \ldots \cup A_n =: \bigcup_{k=1}^{n} A_k =$ `mind. eines der` A_k `tritt ein`

$A \cap B = A$ `und` B `treten ein`

$A_1 \cap A_2 \cap \ldots \cap A_n =: \bigcap_{k=1}^{n} A_k =$ `alle` A_k `treten ein`

$A^c := \Omega \backslash A := \{\omega \in \Omega : \omega \notin A\} = A$ `tritt nicht ein`

A^c heißt **Komplement** der Menge A (in Ω). Es gilt

$$\Omega^c = \emptyset \text{ und } \emptyset^c = \Omega\,.$$

Wahrscheinlichkeitsmaße

Für jedes Ereignis A legen wir im nächsten Schritt eine Wahrscheinlichkeit $P(A)$ zwischen 0 und 1 fest. $P(A)$ soll ein Maß dafür sein, dass das Ereignis A eintritt:

- tritt A niemals ein, so setzt man $P(A) = 0$. Insbesondere $P(\emptyset) = 0$.
- tritt A sicher ein, so setzt man $P(A) = 1$. Insbesondere $P(\Omega) = 1$.

Zusätzlich sollte gelten: Sind A und B disjunkte Ereignisse, d.h. A und B besitzen keine gemeinsamen Elementarereignisse, also $A \cap B = \emptyset$, so ist

$$P(A \cup B) = P(A) + P(B)\,. \tag{2.6}$$

Diese Eigenschaft von P bezeichnet man als **Additivität**.

Aus (2.6) folgt unmittelbar: sind $A_1, \ldots, A_n$ paarweise disjunkte Ereignisse, d.h. $A_k \cap A_l = \emptyset$ für $k \neq l$, so folgt:

$$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n)\,. \tag{2.7}$$

Gilt schließlich auch für jede **unendliche** Folge (A_n) paarweiser disjunkter Ereignisse

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k) \tag{2.8}$$

so spricht man von σ-**Additivität**.

Definition Ein **diskreter Wahrscheinlichkeitsraum** ist ein Paar (Ω, P), wobei

- Ω eine nichtleere, diskrete (d.h. endliche oder abzählbar unendliche) Menge
- P ein **diskretes Wahrscheinlichkeitsmaß auf** Ω, d.h. eine Abbildung

  $$P\,:\,\mathcal{P}(\Omega) \to \mathbb{R}$$

  mit folgenden Eigenschaften:
  - $P(A) \geq 0\ \forall A \in \mathcal{P}(\Omega)$ (Nichtnegativität)
  - $P(\Omega) = 1$ (Normiertheit)
  - $P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$ für jede Folge (A_k) paarweise disjunkter Ereignisse (σ-Additivität).

Rechenregeln für P

- P ist (insbesondere) **endlich additiv**, d.h. für $A_1, \ldots, A_n$ paarweise disjunkt, ist

  $$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n) = \sum_{k=1}^{n} P(A_k)\,.$$

- $P(A^c) = 1 - P(A)$, denn A und A^c sind disjunkt, $A \cup A^c = \Omega$, also

  $$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)\,.$$

- $P(\emptyset) = 0$, denn $\emptyset^c = \Omega$, also

  $$P(\emptyset) = 1 - P(\Omega) = 1 - 1 = 0\,.$$

- $A \subset B$ impliziert $P(A) \leq P(B)$
  denn $B = A \cup (B \cap A^c)$ und A und $B \cap A^c$ sind disjunkt, also

  $$P(B) = P(A) + P(B \cap A^c) \geq P(A)\,.$$

Konstruktion von Wahrscheinlichkeitsmaßen mit Hilfe von Wahrscheinlichkeitsfunktionen

Eine **Wahrscheinlichkeitsfunktion (auf Ω)** ist eine Funktion $p : \Omega \to [0, 1]$ mit

$$\sum_{\omega \in \Omega} p(\omega) = 1 \tag{2.9}$$

Bemerkung Beachten Sie, dass es sich bei (2.9) um eine unendliche Summe handelt, falls Ω unendlich viele Elemente enthält. Gemeint ist mit (2.9) also, dass die (möglicherweise unendliche) Reihe $\sum_{\omega \in \Omega} p(\omega)$ konvergiert und ihr Wert gleich 1 ist. Hierbei kommt es auf die **Reihenfolge**, in der die Wahrscheinlichkeiten $p(\omega)$ aufsummiert werden, **nicht** an, denn die Reihe ist wegen der Nichtnegativität der Summanden $p(\omega)$ absolut konvergent.

Zu gegebener Wahrscheinlichkeitsfunktion p definieren wir die Wahrscheinlichkeit $P(A)$ eines Ereignisses A durch

$$P(A) := \sum_{\omega \in A} p(\omega)\,. \tag{2.10}$$

Die Wahrscheinlichkeit von A ist also gleich der Summe der Wahrscheinlichkeiten aller Elementarereignisse ω die in A liegen. Die so definierte Abbildung P ist ein diskretes Wahrscheinlichkeitsmaß auf Ω, d.h. nichtnegativ, normiert und σ-additiv.

Umgekehrt können wir zu jedem diskreten Wahrscheinlichkeitsmaß P auf Ω durch

$$p(\omega) := P\left(\{\omega\}\right)\,, \omega \in \Omega \tag{2.11}$$

eine **Wahrscheinlichkeitsfunktion auf** Ω definieren.

Durch (2.10) und (2.11) ist also eine 1-1 Beziehung zwischen allen Wahrscheinlichkeitsmaßen über Ω und allen Wahrscheinlichkeitsfunktionen über Ω gegeben.

Beispiele

(i) Beim Würfeln mit einem fairen Würfel ist jede der sechs möglichen Augenzahlen gleichwahrscheinlich. Man setzt daher

$$p(\omega) = \frac{1}{6} \text{ für } \omega \in \Omega = \{1, 2, 3, 4, 5, 6\}\,.$$

Es folgt z.B.

$$P(\texttt{Augenzahl ungerade}) = P\left(\{1, 3, 5\}\right) = \frac{3}{6} = \frac{1}{2}\,.$$

(ii) Beim zweimaligen Würfeln mit einem fairen Würfel ist wiederum jedes der 36 Elementarereignisse aus $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ gleichwahrscheinlich, also $p(\omega) = \frac{1}{36}\ \forall \omega \in \Omega$. Es folgt z.B.

$$P(\texttt{Augensumme} > 10) = P\left(\{(5, 6), (6, 5), (6, 6)\}\right) = \frac{3}{36} = \frac{1}{12}\,.$$

Beide Beispiele sind Spezialfälle eines Laplaceschen Wahrscheinlichkeitsraumes.

Laplacescher Wahrscheinlichkeitsraum

Ist Ω eine endliche Menge, so definiert

$$p(\omega) := \frac{1}{|\Omega|}, \quad \omega \in \Omega$$

eine Wahrscheinlichkeitsfunktion auf Ω. Für die Wahrscheinlichkeit $P(A)$ eines beliebigen Ereignisse folgt hieraus sofort

$$P(A) = \sum_{\omega \in A} \frac{1}{|\Omega|} = \frac{|A|}{|\Omega|}. \tag{2.12}$$

$P(A)$ heißt **Laplace-Wahrscheinlichkeit von** A. Da jedes Elementarereignis gleichwahrscheinlich ist, spricht man von P auch als der **Gleichverteilung auf** Ω.

Die Berechnung der Wahrscheinlichkeit $P(A)$ in (2.12) führt auf das Problem der **Abzählung der Elemente in** A, also auf ein **Abzählproblem**. Die wichtigsten Abzählprobleme sollen im folgenden anhand von einfachen **Urnenmodellen** illustriert werden:

Eine Urne enthalte n unterscheidbare Kugeln $1, 2, \ldots, n$. Wir unterscheiden dann das k-malige Ziehen einer Kugel aus der Urne mit/ohne Zurücklegen, wobei es auf die Reihenfolge der gezogenen Kugeln ankommt/nicht ankommt:

1) **in Reihenfolge mit Zurücklegen**

$$\Omega = \{\omega = (x_1, \ldots, x_k) : x_i \in \{1, \ldots, n\}\}, |\Omega| = n^k$$

d.h., ein Elementarereignis $\omega = (x_1, \ldots, x_k)$ ist ein **k-Tupel**, d.h. eine geordnete Menge der Länge k, wobei x_i für die Nummer der i-ten gezogenen Kugel steht.

2) **in Reihenfolge ohne Zurücklegen**

$$\Omega = \{\omega = (x_1, \ldots, x_k) : x_i \in \{1, \ldots n\}, x_i \neq x_j \text{ für } i \neq j\}$$

$$|\Omega| = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}.$$

Zur Erinnerung: Fakultätsfunktion

$$m! := m(m-1) \cdot (m-2) \cdot \ldots \cdot 2 \cdot 1 = \Pi_{k=1}^{m} k, \text{ und } 0! := 1.$$

Insbesondere

$$n! = n \cdot (n-1)! = n \cdot (n-1) \cdot (n-2)! = \ldots = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) \cdot (n-k)!,$$

also

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

Für k = n erhält man als Spezialfall

$$|\Omega| = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$$

$n!$ ist also gleich der Anzahl aller möglichen Anordnungen (oder auch **Permutationen**) der n-elementigen Menge $\{1, \ldots, n\}$.

3) **ohne Reihenfolge ohne Zurücklegen**

$$\Omega = \{\omega = \{x_1, \ldots, x_k\} : x_i \in \{1, 2, \ldots, n\}, x_i \neq x_j \text{ für } i \neq j\}$$

Im Unterschied zum Ziehen in Reihenfolge werden nun alle k-Tupel $(x_1, \ldots, x_k)$, die zu derselben Menge der gezogenen Kugeln führen, zu einem Elementarereignis zusammengefasst. Insgesamt gibt es $k!$ solcher Tupel (das entspricht also gerade der Anzahl der Permutationen der Menge der k gezogenen Kugeln), also erhalten wir insgesamt

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

Elementargereignisse. Es gilt also

$$|\Omega| = \binom{n}{k}.$$

Insbesondere: $\binom{n}{k}$ ist gleich der Anzahl aller k-elementigen Teilmengen aus einer n-elementigen Grundmenge.

Alternative Darstellung von Ω**:** Unter allen k-Tupeln, die zur selben Menge $\{x_1, \ldots, x_k\}$ führen, gibt es genau ein Tupel $(x_{(1)}, \ldots, x_{(k)})$, in dem die Elemente ihrer Größe nach angeordnet sind:

$$x_{(1)} < x_{(2)} < \ldots < x_{(k)}\,.$$

Wir können daher auch schreiben

$$\Omega = \{(x_1, \ldots, x_k) \,:\, x_i \in \{1, \ldots, n\}\,, x_1 < x_2 < \ldots < x_k\}\,.$$

4) **ohne Reihenfolge mit Zurücklegen**

Analog zu 3) ordnen wir wieder die Nummern der gezogenen Kugeln der Größe nach an:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(k)} \tag{2.13}$$

wobei wegen des Zurücklegens Kugeln mehrfach gezogen werden können.

Durch Übergang von $x_{(i)}$ zu $x_{(i)} + i - 1$ erhält man aus (2.13) eine streng monoton aufsteigende Folge

$$x_{(1)} < x_{(2)} + 1 < x_{(3)} + 2 < \ldots < x_{(k)} + k - 1\,.$$

Wir erhalten als Stichprobenraum in diesem Falle also

$$\Omega = \{(x_1, \ldots, x_k) \,:\, x_i \in \{1, \ldots, n, n+1, \ldots, n+k-1\}, x_1 < x_2 < \ldots < x_k\}.$$

Für die Mächtigkeit $|\Omega|$ von Ω ergibt sich nach 3)

$$|\Omega| = \binom{n+k-1}{k}.$$

Bedingte Wahrscheinlichkeiten und Unabhängigkeit

Ist über den Ausgang eines Zufallsexperiments bereits eine Teilinformation verfügbar, ändern sich entsprechend die Wahrscheinlichkeiten der Elementarereignisse.

Beispiel

Zweimaliges Würfeln eines fairen Würfels

$$P(\texttt{Augensumme} \;\; > 10) = \frac{1}{12}.$$

Wie ändert sich diese Wahrscheinlichkeit, wenn bereits bekannt ist, dass beim ersten Würfeln eine 6 gewürfelt wurde? Unter dieser Annahme bleiben nur noch sechs gleichwahrscheinliche Möglichkeiten für die zweite Augenzahl übrig, von denen die Augenzahlen 5 und 6 insgesamt zu einer Augensumme größer als 10 führen. Für die Wahrscheinlichkeit des Ereignisses $\texttt{Augenzahl} \;\; > 10$ unter der Bedingung $\texttt{1.Augenzahl} \;\; 6$ ergibt sich somit

$$P(\texttt{Augensumme} \;\; > 10 \mid \texttt{1.Augenzahl} \;\; 6) = \frac{2}{6} = \frac{1}{3}.$$

Die bedingte Wahrscheinlichkeit ist also viermal höher als die ursprüngliche "a priori" Wahrscheinlichkeit.

Definition Für Ereignisse A, B mit $P(B) > 0$ heißt

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

die **bedingte Wahrscheinlichkeit von A unter der Bedingung B** (oder auch: **die bedingte Wahrscheinlichkeit von A gegeben B**). Im Falle $P(B) = 0$ setzen wir einfach $P(A \mid B) := 0$.

Eigenschaften der bedingten Wahrscheinlichkeit

- $P(A \,|\, B) \in [0, 1]$
- $P(\emptyset \,|\, B) = 0$
- Gilt $P(B) > 0$, so ist $P(\Omega \,|\, B) = 1$ und

  $$P(\cdot \,|\, B) \,:\, \mathcal{P}(\Omega) \to [0, 1], A \mapsto P(A \,|\, B)$$

  ist wieder eine diskrete Wahrscheinlichkeitsverteilung auf Ω. $P(\cdot \mid B)$ heißt **bedingte Wahrscheinlichkeitsverteilung unter der Bedingung B.**

Beispiel (Laplacescher Wahrscheinlichkeitsraum)

Ω endlich, $P(A) = \frac{|A|}{|\Omega|}$ sei die Gleichverteilung auf Ω. Dann folgt für $B \neq \emptyset$

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

Insbesondere: Die bedingte Wahrscheinlichkeitsverteilung ist im Falle des Laplaceschen Wahrscheinlichkeitsraumes gerade die Gleichverteilung auf B.

Beispiel

Bedingte Wahrscheinlichkeiten bilden die Grundlage für das Tarifsystem von Versicherungen. Verunglücken etwa mehr Männer als Frauen, sollten entsprechende Prämien einer Versicherung gegen Arbeitsunfälle für Männer höher als für Frauen sein, etwa:

$$\begin{aligned} P(\texttt{Unfall} \mid \texttt{V weiblich}) &= 0.002 \\ P(\texttt{Unfall} \mid \texttt{V männlich}) &= 0.005\,. \end{aligned}$$

Kennt man noch den Anteil der männlichen und weiblichen Versicherungsnehmer, etwa

$$P(\texttt{V weiblich}) = \frac{2}{5} = 1 - P(\texttt{V männlich})\,,$$

so kann man hieraus die totale Wahrscheinlichkeit eines Arbeitsunfalls errechnen:

$$\begin{aligned} P(\texttt{Unfall}) &= P(\texttt{Unfall und V weiblich}) + P(\texttt{Unfall und V männlich}) \\ &= P(\texttt{Unfall} \mid \texttt{V weiblich}) P(\texttt{V weiblich}) \\ &\quad + P(\texttt{Unfall} \mid \texttt{V männlich}) P(\texttt{V männlich}) \\ &= 0.002\frac{2}{5} + 0.005\frac{3}{5} = 0.0038\,. \end{aligned}$$

Die Berechnung der "totalen" Wahrscheinlichkeit für einen Arbeitsunfall ist ein Speziallfall des ersten Teils des folgenden Satzes.

Satz

Es seien $B_1, \ldots, B_n$ disjunkte Teilmengen von Ω und $A \subset B_1 \cup \ldots \cup B_n$. Dann folgt:

(i) **(Formel von der totalen Wahrscheinlichkeit)**

$$P(A) = \sum_{k=1}^{n} P(A \,|\, B_k) P(B_k)\,. \tag{2.14}$$

(ii) **(Formel von Bayes)** Für $P(A) > 0$ gilt

$$P(B_i \,|\, A) = \frac{P(A \,|\, B_i) P(B_i)}{\sum_{k=1}^{n} P(A \,|\, B_k) P(B_k)}\,. \tag{2.15}$$

Beispiel

In obigem Beispiel kennt man bereits die totale Wahrscheinlichkeit eines Arbeitsunfalls

$$P(\texttt{Arbeitsunfall}) = 0.0038\,.$$

Die Formel von Bayes liefert nun für die "umgekehrte" bedingte Wahrscheinlichkeit

$$\begin{aligned} P(&\texttt{V männlich} \mid \texttt{Arbeitsunfall}) \\ &= \frac{P(\texttt{Arbeitsunfall} \mid \texttt{V männlich})P(\texttt{V männlich})}{P(\texttt{Arbeitsunfall})} \\ &= \frac{0.003}{0.0038} = 0.789\,. \end{aligned}$$

Wie zu erwarten handelt es sich bei einer verunglückten Person in fast 80% aller Fälle um Männer. Dieses Verhältnis kann sich aber ins Gegenteil verkehren, wenn entweder der Anteil der weiblichen Versicherungsnehmer den Anteil der männlichen Versicherungsnehmer weit übersteigt oder die bedingte Wahrscheinlichkeit $P(\texttt{Arbeitsunfall} \mid \texttt{V weiblich})$ für einen Arbeitsunfall eines weiblichen Versicherungsnehmers die entsprechende Wahrscheinlichkeits eines Arbeitsunfalles eines männlichen Versicherungsnehmers weit übersteigt.

Beispiel

Mitunter liefert die Formel von Bayes scheinbar überraschende Aussagen wie im Falle des folgenden Tests auf eine seltene Krankheit.

Angenommen, 5 Promille der Bevölkerung haben eine seltene Krankheit K, d.h.

$$P(K) = 0.005\,.$$

Ein medizinischer Test zeigt bei 99% der Erkrankten eine positive Reaktion, d.h.

$$P(\texttt{Test positiv} \mid K) = 0.99\,.$$

Allerdings zeigt besagter Test auch bei 2% der Gesunden eine positive Reaktion, d.h.

$$P(\texttt{Test positiv} \mid K^c) = 0.02\,.$$

Von besonderem Interesse ist nun offenbar folgende

Frage: Angenommen, der Test ist positiv. Wie groß ist die Wahrscheinlichkeit, dass die getestete Person tatsächlich an K erkrankt ist? Wie groß ist also die bedingte Wahrscheinlichkeit

$$P(K \mid \texttt{Test positiv})\,?$$

Die Formel von Bayes liefert

$$\begin{aligned} P(K \mid \texttt{Test positiv}) &= \frac{P(\texttt{Test positiv} \mid K)P(K)}{P(\text{Test positiv} \mid K) \cdot P(K) + P(\texttt{Test positiv} \mid K^c)P(K^c)} \\ &= \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.02 \cdot 0.995} = \frac{495}{2485} \sim 0.2\,. \end{aligned}$$

Also: Nur in 2 von 10 Fällen mit postivem Testergebnis ist die getestete Person auch wirklich an K erkrankt.

Unabhängigkeit

Ist $P(A) = P(A|B)$, d.h. die Wahrscheinlichkeit von A **unabhängig** davon, ob das Ereignis B eingetreten ist oder nicht, so folgt:

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

und damit

$$P(A \cap B) = P(A) \cdot P(B)\,. \tag{2.16}$$

Zwei Ereignisse A und B mit (2.16) heißen **(stochastisch) unabhängig**.

Allgemeiner

Definition Die Ereignisse $A_1, \ldots, A_n$ heißen (stochastisch) unabhängig, falls für jede nichtleere Teilmenge $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ gilt:

$$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_k}).$$

Man beachte, dass zum Nachweis der Unabhängigkeit dreier Ereignisse A, B und C, der Nachweis der **paarweisen Unabhängigkeit** je zweier Ereignisse nicht ausreicht. Als Beispiel betrachten wir beim zweimaligen Werfen einer fairen Münze die Ereignisse

$$\begin{aligned} A &= \texttt{1.Wurf Zahl} \\ B &= \texttt{2.Wurf Zahl} \\ C &= \texttt{1. und 2.Wurf gleich}\,. \end{aligned}$$

Diese sind paarweise unabhängig aber nicht unabhängig, denn $P(A) = P(B) = P(C) = \frac{1}{2}$, $P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4}$, aber

$$P(A \cap B \cap C) = \frac{1}{4} \neq P(A)P(B)P(C)\,.$$

Beispiel Beim zweimaligen Würfeln eines fairen Würfels ist die erste Augenzahl offenbar "unabhängig" von der zweiten Augenzahl, also jedes Ereignis A, das nur von der ersten Zahl abhängt, unabhängig von jedem Ereignis B, das nur von der zweiten Augenzahl abhängt, etwa:

$$\begin{aligned} A &= \texttt{1.Augenzahl gerade}\,, \qquad P(A) = \frac{1}{2} \\ B &= \texttt{2.Augenzahl} \geq 5\,, \qquad P(B) = \frac{1}{3}\,. \end{aligned}$$

Dann gilt

$$\begin{gathered} P(A \cap B) = P\left(\{(2,5),(2,6),(4,5),(4,6),(6,5),(6,6)\}\right) \\ \frac{6}{36} = \frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3} = P(A) \cdot P(B)\,. \end{gathered}$$

2. Zufallsvariablen und Verteilungen

Im ganzen Abschnitt sei (Ω, P) ein diskreter Wahrscheinlichkeitsraum. Eine Funktion

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

heißt **Zufallsvariable (auf Ω)**. Da Ω abzählbar, ist auch das Bild

$$X(\Omega) = \{X(\omega) : \omega \in \Omega\} \subset \mathbb{R}$$

abzählbar.

Für $x \in \mathbb{R}$ betrachten wir insbesondere das Ereignis

$$\{X = x\} := \{\omega \in \Omega : X(\omega) = x\} = X \texttt{ nimmt den Wert } x \texttt{ an}$$

Durch

$$p_X(x) := P(X = x)\,, \quad x \in X(\Omega)$$

wird dann eine neue Wahrscheinlichkeitsfunktion auf $X(\Omega)$ definiert. Das zugehörige diskrete Wahrscheinlichkeitsmaß P_X auf $\mathcal{P}(X(\Omega))$ heißt **Verteilung von X (unter P).**

Für beliebige Ereignisse $A \subset X(\Omega)$ gilt offenbar

$$\begin{aligned} P_X(A) &= \sum_{x\in A} p_X(x) = \sum_{x\in A} P(X = x) \\ &= P\Big(\underbrace{\bigcup_{x\in A} \{\omega : X(\omega) = x\}}_{=\{\omega : X(\omega)\in A\}} \Big) = P(X \in A)\,. \end{aligned}$$

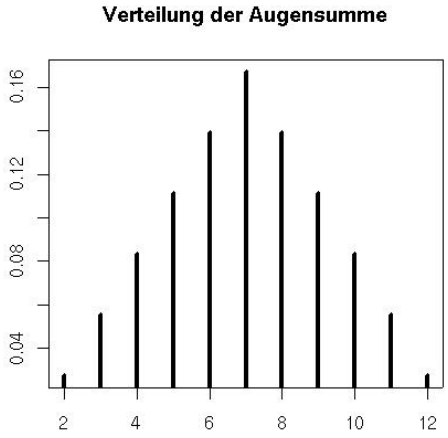
Beispiel Beim zweimaligen Würfel eines fairen Würfels sei X die Augensumme. X ist eine Zufallsvariable mit Werten in der Menge $\{2, 3, \ldots, 12\}$, von denen aber nicht alle Werte mit gleicher Wahrscheinlichkeit von X angenommen werden. Vielmehr gilt:

$$\begin{aligned} p_X(2) &= P(\{(k,l) \in \Omega : k + l = 2\}) = P(\{(1,1)\}) = \frac{1}{36} \\ p_X(12) &= P(\{6,6\}) = \frac{1}{36} \end{aligned}$$

und für die übrigen Werte

$$\begin{aligned} p_X(3) = p_X(11) = \frac{2}{36}\,, p_X(4) = p_X(10) = \frac{3}{36} \\ p_X(5) = p_X(9) = \frac{4}{36}\,, p_X(6) = p_X(8) = \frac{5}{36} \\ p_X(7) = \frac{6}{36}\,. \end{aligned}$$

Graphische Veranschaulichung der Verteilung von X mit Hilfe eines **Stabdiagramms**



Die Verteilungsfunktion einer Zufallsvariablen

Die Funktion

$$F(x) := P(X \leq x)\,, x \in \mathbb{R}$$

heißt **Verteilungsfunktion** von X. Sie besitzt wie jede empirische Verteilungsfunktion (siehe Abschnitt I.2) folgende Eigenschaften:

- F ist monoton wachsend
- $0 \leq F \leq 1$, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$
- F ist rechtsseitig stetig.

Unabhängigkeit von Zufallsvariablen

Definition Es seien $X_1, X_2, \ldots, X_n$ Zufallsvariablen auf dem Wahrscheinlichkeitsraum (Ω, P). $X_1, \ldots, X_n$ heißen **(stochastisch) unabhängig**, falls für alle Teilmengen $B_1, \ldots, B_n$ von $\mathbb{R}$ gilt:

$$P(X_1 \in B_1, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdot \ldots \cdot P(X_n \in B_n)\,. \tag{2.17}$$

Die Zufallsvariablen $X_1, \ldots, X_n$ sind also genau dann (stochastisch) unabhängig, wenn für beliebige Teilmengen $B_1, \ldots, B_n$ die Ereignisse

$$\{X_1 \in B_1\}\,, \ldots\,, \{X_n \in B_n\}$$

(stochastisch) unabhängig sind.

Äquivalent zu (2.17) ist folgende, in der Praxis einfacher zu überprüfende Bedingung: Für alle $x_1, \ldots, x_n \in \mathbb{R}$ ist

$$P\left(X_1 = x_1\,, \ldots\,, X_n = x_n\right) = P\left(X_1 = x_1\right) \cdot \ldots \cdot P\left(X_n = x_n\right)\,. \tag{2.18}$$

Beachten Sie, dass $P\left(X_k = x_k\right) = 0$ für die weitaus meisten Werte $x_k \in \mathbb{R}$, nämlich mindestens für alle $x_k \in \mathbb{R} \setminus X_k(\Omega)$.

Die Unabhängigkeit bleibt unter Transformationen erhalten, d.h., sind $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ stückweise stetige Abbildungen, so sind auch die Zufallsvariablen

$$f_1(X_1), \ldots, f_n(X_n)$$

unabhängig. Um dies einzusehen beachte man, dass $\{f_i(X_i) = x_i\} = \{X_i \in f_i^{-1}(x_i)\}$ und somit

$$\begin{aligned} P(f_1(X_1) = x_1, \ldots, f_n(X_n) = x_n) &= P(X_1 \in f_1^{-1}(x_1), \ldots, X_n \in f_n^{-1}(x_n)) \\ &= P(X_1 \in f_1^{-1}(x_1)) \cdot \ldots \cdot P(X_n \in f_n^{-1}(x_n)) \\ &= P(f_1(X_1) = x_1) \cdot \ldots \cdot P(f_n(X_n) = x_n)\,. \end{aligned}$$

Aufgrund des Kriteriums (2.18) folgt die Unabhängigkeit von $f_1(X_1), \ldots, f_n(X_n)$.

Beispiel Beim zweimaligen Würfeln sei X_1 die erste Augenzahl und X_2 die zweite. Mit (2.18) ist dann einfach zu sehen, dass X_1 und X_2 unabhängig sind. Ebenso sind auch die Zufallsvariablen $\sin(X_1)$ und X_2^2 unabhängig.

Spezielle Verteilungen

Bernoulli-Verteilung

Fixiere eine Teilmenge $A \subset \Omega$ und definiere

$$X(\omega) := \begin{cases} 1 & \text{für } \omega \in A \\ 0 & \text{für } \omega \in A^c\,. \end{cases}$$

Wir interpretieren das Ereignis $\{X = 1\} = A$ als "Erfolg". Dementsprechend bezeichnen wir

$$p := P(X = 1) = P(A)$$

als **Erfolgswahrscheinlichkeit**. Entsprechend gilt für die Wahrscheinlichkeit eines Mißerfolges

$$P(X = 0) = P(A^c) = 1 - P(A) = 1 - p\,.$$

Definition Es sei $p \in [0, 1]$. Das durch die Wahrscheinlichkeitsfunktion $p : \{0, 1\} \to [0, 1]$

$$p(1) = p\,, \text{ und } p(0) = 1 - p$$

definierte Wahrscheinlichkeitsmaß auf $\{0, 1\}$ heißt **Bernoulli-Verteilung zu p**. Zufallsexperimente, die nur zwei mögliche Ausgänge kennen, nennt man entsprechend **Bernoulli-Experimente**.

Beispiele für Bernoulli-Experimente

- Werfen einer fairen Münze: $P(\texttt{Kopf}) = P(\texttt{Zahl}) = \frac{1}{2}$
- Geschlecht eines Neugeborenen: $P(\texttt{weiblich}) = 0.47$, $P(\texttt{männlich}) = 0.53$
- Ziehen einer Kugel aus einer Urne mit s schwarzen und w weißen Kugeln:
  $P(\texttt{gez. Kugel schwarz}) = \frac{s}{s+w}$

Binomialverteilung

Es seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen, die alle Bernoulli-verteilt sind zu p.
Wir können X_i als Ausgang eines Bernoulli Experiments mit Erfolgswahrscheinlichkeit p interpretieren, wobei die Folge der n Experimente unabhängig ist. Dann zählt die Zufallsvariable

$$S_n := X_1 + \ldots + X_n \quad \in \{0, \ldots, n\}$$

die **Gesamtanzahl der Erfolge**.

Für die Verteilung P_{S_n} der Summe S_n gilt dann

$$p_{S_n}(k) = P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} =: b(k; n, p)$$

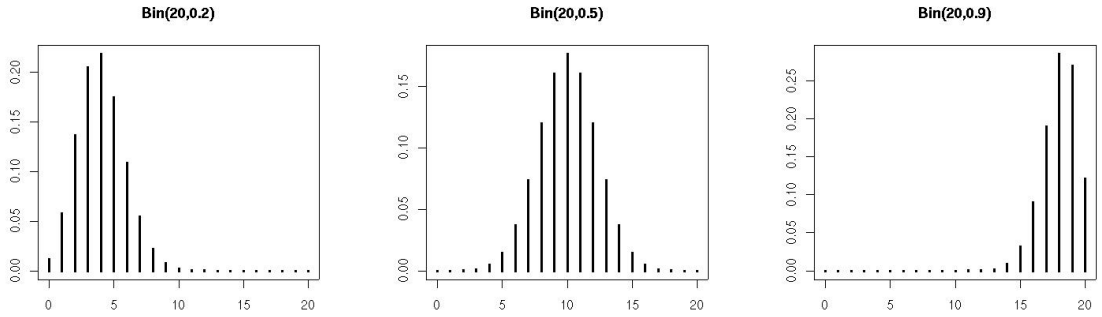
Hierbei ist $\binom{n}{k}$ gerade die Anzahl der n-Tupel mit genau k Einsen (und $n-k$ Nullen), p^k die Wahrscheinlichkeit für k Erfolge und $(1-p)^{n-k}$ die Wahrscheinlichkeit für $n-k$ Mißerfolge.

Definition Es sei $n \in \mathbb{N}$ und $p \in [0,1]$. Das durch die Wahrscheinlichkeitsfunktion

$$\begin{aligned} b(\cdot; n, p) \; : \; \{0, \ldots, n\} &\to [0,1] \\ k &\mapsto \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

definierte Wahrscheinlichkeitsmaß auf $\{0, \ldots, n\}$ heißt **Binominalverteilung zu n und p** und wird mit Bin (n, p) bezeichnet.

Wir haben insbesondere gesehen: Bei einer Folge von n unabhängigen Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit p ist die Summe der Erfolge binominalverteilt mit Parameter n und p.



Geometrische Verteilung

Wie groß ist die Wahrscheinlichkeit, dass man mit einem fairen Würfel genau k Versuche benötigt, bis zum ersten Mal eine 6 gewürfelt wird?

Für $k = 1$ ist die gesuchte Wahrscheinlichkeit offensichtlich $\frac{1}{6}$, für $k = 2$ ist sie gleich $\frac{5}{6} \cdot \frac{1}{6}$, denn die gesuchte Wahrscheinlichkeit ist aufgrund der Unabhängigkeit der beiden Würfe gleich dem Produkt aus der Wahrscheinlichkeit, beim ersten Würfeln keine 6 zu würfeln $(= \frac{5}{6})$, und der Wahrscheinlichkeit, beim zweiten Würfeln eine 6 zu würfeln $(= \frac{1}{6})$.

Für allgemeines k können wir wie folgt vorgehen: Wir definieren eine Folge von Zufallsvariablen $X_1, X_2, X_3, \ldots$ durch

$$X_k := 1 \text{ falls beim } k\text{-ten Wurf eine } 6 \text{ gewürfelt wird}$$

und $X_k := 0$ sonst. Offenbar sind die Zufallsvariablen $X_1, X_2, X_3, \ldots$ unabhängig Bernoulliverteilt mit Erfolgswahrscheinlichkeit $p = \frac{1}{6}$. Das Ereignis A_k, im k-ten Wurf zum ersten Mal eine 6 zu würfeln, kann mit Hilfe dieser Zufallsvariablen nun wie folgt beschrieben werden:

$$A_k = \{X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1\}.$$

Aufgrund der Unabhängigkeit der Zufallsvariablen ergibt sich für die gesuchte Wahrscheinlichkeit

$$\begin{aligned} P(A_k) &= P(X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1) \\ &= P(X_1 = 0) \cdot P(X_2 = 0) \cdot \ldots \cdot P(X_{k-1} = 0) \cdot P(X_k = 1) \\ &= \frac{5}{6} \cdot \frac{5}{6} \cdot \ldots \cdot \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6}. \end{aligned}$$

Allgemeiner Gegeben eine Folge von unabhängigen Zufallsvariablen $X_1, X_2, X_3, \ldots$, die alle Bernoulli-verteilt sind zu $p > 0$. Definiere die **Wartezeit auf den ersten Erfolg**

$$T := \min\{k \geq 1 \,:\, X_k = 1\}.$$

Wie in obigem Fall der Wartezeit auf die erste 6 beim Würfeln mit einem fairen Würfel, erhalten wir für die Verteilung von T

$$\begin{aligned} P(T = k) &= P(X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1) \\ &= P(X_1 = 0) \cdot P(X_2 = 0) \cdot \ldots \cdot P(X_{k-1} = 0) \cdot P(X_k = 1) \\ &= (1-p)^{k-1} \cdot p \end{aligned}$$

für $k = 1, 2, 3, \ldots$.

Definition Es sei $p \in]0, 1]$. Das durch die Wahrscheinlichkeitsfunktion

$$\begin{aligned} g_p \,:\, \mathbb{N} &\mapsto [0, 1] \\ k &\mapsto (1-p)^{k-1} p \end{aligned}$$

definierte Wahrscheinlichkeitsmaß auf $\mathbb{N}$ heißt **geometrische Verteilung zu p** und wird mit Geom (p) bezeichnet.

Poissonverteilung

Für $\lambda > 0$ definiert

$$\pi_\lambda(k) := e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0$$

eine Wahrscheinlichkeitsfunktion auf $\mathbb{N}_0$, denn aus der Reihenentwicklung der Exponentialfunktion

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \quad x \in \mathbb{R}$$

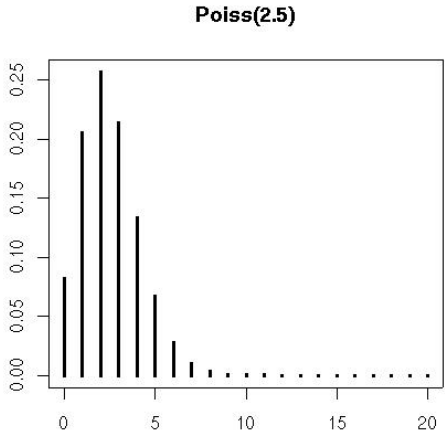
folgt

$$\sum_{k=0}^{\infty} \pi_\lambda(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = e^0 = 1\,.$$

Definition Es sei $\lambda > 0$. Das durch die Wahrscheinlichkeitsfunktion

$$\begin{aligned} \pi_\lambda \,:\, \mathbb{N}_0 &\to [0,1] \\ k &\mapsto e^{-\lambda}\frac{\lambda^k}{k!} \end{aligned}$$

definierte Wahrscheinlichkeitsmaß auf $\mathbb{N}_0$ heißt **Poissonverteilung zu** λ und wird mit Poiss (λ) bezeichnet.



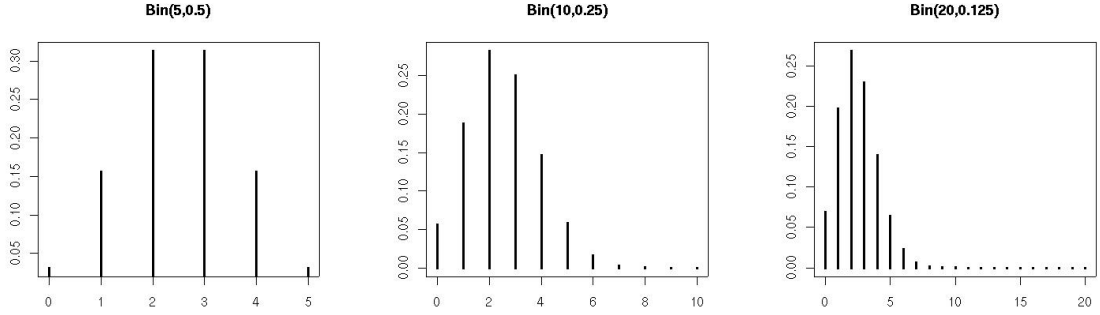
Die Poissonverteilung empfiehlt sich als Näherung der Binomialverteilung Bin (n,p) für große n und kleine p. Die Approximation ist umso besser, je kleiner der Wert np^2 ist. Diese Näherung wird gerechtfertigt durch die folgende Beobachtung:

Poissonscher Grenzwertsatz

Es sei $(p_n) \subset [0,1]$ eine Folge von Erfolgsparametern mit $\lim_{n\to\infty} np_n = \lambda > 0$. Dann folgt

$$\lim_{n\to\infty} b(k;n,p_n) = \pi_\lambda(k) \qquad \text{für alle } k \in \mathbb{N}_0\,.$$

Mit anderen Worten: Die Wahrscheinlichkeitsfunktion der Binomialverteilung Bin (n,p_n) konvergiert punktweise gegen die Wahrscheinlichkeitsfunktion der Poissonverteilung Poiss (λ). Im folgenden eine Illustration dieser Konvergenz für $\lambda = 2.5$.



Zum Beweis des Poissonschen Grenzwertsatzes beachte mann, dass unter der Annahme $\lim_{n\to\infty} np_n = \lambda$ folgt

$$\begin{aligned}\lim_{n\to\infty} b(k;n,p_n) &= \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \lim_{n\to\infty} \frac{1}{k!} \underbrace{\frac{n}{n}}_{\longrightarrow 1} \underbrace{\frac{(n-1)}{n}}_{\longrightarrow 1} \cdot \ldots \cdot \underbrace{\frac{(n-k+1)}{n}}_{\longrightarrow 1} \underbrace{(np_n)^k}_{\longrightarrow \lambda^k} \underbrace{\left(1-\frac{np_n}{n}\right)^{n-k}}_{\sim\left(1-\frac{\lambda}{n}\right)^n \longrightarrow e^{-\lambda}} \\ &= \frac{1}{k!}\lambda^k e^{-\lambda} = \pi_\lambda(k)\,.\end{aligned}$$

Eine näherungsweise Berechnung von Wahrscheinlichkeiten gewisser Ereignisse mit Hilfe einer Poissonverteilung ist immer dann gerechtfertigt, wenn es sich um seltene Ereignisse handelt.

Beispiel Bei der Herstellung von DVD-Scheiben ist ein Anteil von $p = 0.002$ bereits bei der Produktion defekt. Wie groß ist die Wahrscheinlichkeit, dass in einem Warenposten mit $n = 1.000$ DVD-Scheiben mindestens fünf Schreiben defekt sind?
Zur Beantwortung dieser Frage sei X die Anzahl der defekten DVD-Scheiben. Da es sich bei der Produktion einer defekten DVD-Scheibe (eher) um ein seltenes Ereignis handelt, empfiehlt sich eine Näherung der Verteilung von X mit Hilfe einer Poissonverteilung. Den Parameter λ wählt man gemäß der Regel

$$\lambda = np = 1000 \cdot 0.002 = 2\,.$$

Damit folgt für die gesuchte Wahrscheinlichkeit

$$\begin{aligned}P(X \geq 5) &= 1 - P(X \leq 4) = 1 - e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}\right) \\ &= 1 - e^{-2}\left(1 + 2 + 2 + \frac{4}{3} + \frac{2}{3}\right) \approx 0.05\,.\end{aligned}$$

Hypergeometrische Verteilung

Es sei eine Grundgesamtheit mit N Elementen gegeben, von denen K Elemente die Eigenschaft E besitzen. Aus dieser Grundgesamtheit werde n-mal ohne Zurücklegen gezogen. Wir sind interessiert an der Anzahl k der gezogenen Elemente, die die Eigenschaft E besitzen. Hierzu definieren wir

$$X = \text{Anzahl der gezogenen Elemente mit Eigenschaft } E.$$

Beispiel Hochrechnungen

Ein See enthalte eine (unbekannte) Anzahl N von Fischen. Um N zu schätzen, markiere man zunächst K Fische mit rot. Danach ziehe man n ($n \leq N$) Fische aus dem See. Dann ist X die Anzahl der markierten Fische aus dieser Stichprobe und

$$\hat{N} := \frac{n}{X}K$$

ist eine natürliche Schätzung für die unbekannte Gesamtanzahl N. Zur Begründung beachte man, dass der Anteil $\frac{X}{n}$ an rot markierten Fischen in der Stichprobe dem Anteil $\frac{K}{N}$ aller rot markierten Fische an der Gesamtpopulation entsprechen sollte, d.h.

$$\frac{X}{n} \sim \frac{K}{N} \qquad \text{und damit } N \sim \frac{n}{X}K = \hat{N}\,.$$

Ist $\frac{n}{N}$ klein, so gibt es keinen großen Unterschied zwischen dem Ziehen ohne Zurücklegen und dem Ziehen mit Zurücklegen. Daher empfiehlt sich in diesem Falle eine Approximation der Verteilung von X durch die Binomialverteilung Bin (n,p) mit $p = \frac{K}{N}$, also

$$P(X=k) \approx b(k;n,\frac{K}{N})\,. \tag{2.19}$$

Ist $\frac{n}{N}$ jedoch vergleichsweise groß, so muss die gesuchte Verteilung exakt berechnet werden:

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}\,, \quad k=0,\ldots,n\,. \tag{2.20}$$

Zur Herleitung der Formel (2.20) für die gesuchte Wahrscheinlichkeit beachte man, dass $\binom{K}{k}$ (bzw. $\binom{N-K}{n-k}$) gerade die Anzahl der k (bzw. $n-k$)-elementigen Teilmengen einer K (bzw. $N-K$)-elementigen Grundmenge ist, während $\binom{N}{n}$ die Anzahl aller n-elementigen Teilmengen der Grundgesamtheit aus N Elementen ist.

Definition Es sei $K \le N$, $n \le N$. Das durch die Wahrscheinlichkeitsfunktion

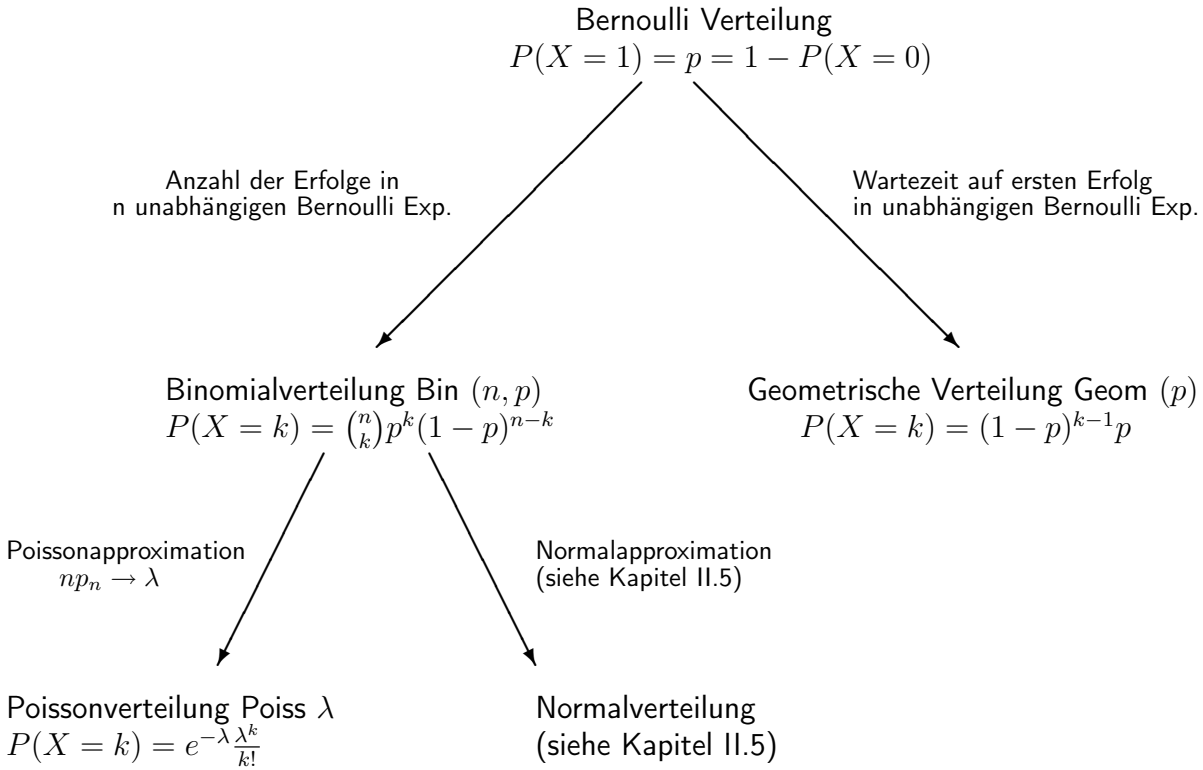
$$\begin{aligned} H(\cdot;n,N,K) \,:\, \{0,\ldots,n\} &\to [0,1] \\ k &\mapsto \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \end{aligned}$$

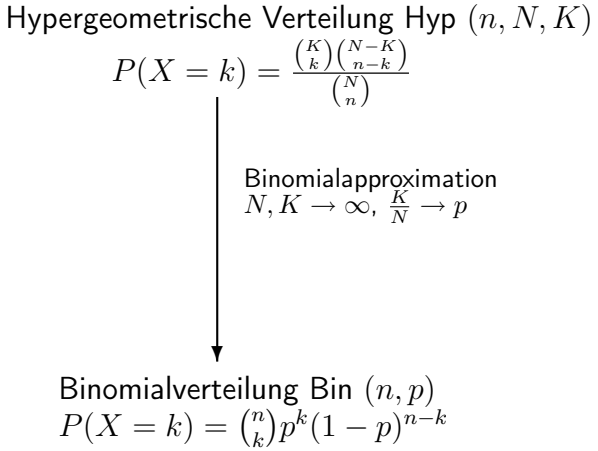
definierte Wahrscheinlichkeitsmaß auf $\{0,\ldots,n\}$ heißt **Hypergeometrische Verteilung zu** n,N **und** K und wird mit Hyp (n,N,K) bezeichnet.

Begründung von (2.19) Für $N,K \to \infty$ mit $p := \frac{N}{K}$ konstant, gilt

$$\begin{aligned} P(X=k) &= \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \\ &= \binom{n}{k}\frac{K!}{(K-k)!}\frac{(N-K)!}{((N-K)-(n-k))!}\frac{(N-n)!}{N!} \\ &= \binom{n}{k}\frac{K}{N}\frac{K-1}{N}\cdots\frac{K-k+1}{N}\,\frac{N-K}{N}\frac{N-K-1}{N}\cdots\frac{N-K-n-k+1}{N} \\ &\qquad \frac{N}{N}\frac{N}{N-1}\cdots\frac{N}{N-n+1} \quad \to \binom{n}{k}p^k(1-p)^{n-k}\,. \end{aligned}$$

Die wichtigsten diskreten Verteilungen im Überblick





3. Erwartungswert und Varianz

Erwartungswert und Varianz sind die beiden wichtigsten Kennzahlen einer Zufallsvariablen. Im ganzen Abschnitt sei (Ω, P) ein diskreter Wahrscheinlichkeitsraum, p die zugehörige Wahrscheinlichkeitsfunktion.

Der **Erwartungswert** $E(X)$ einer Zufallsvariablen X wird definiert als der Mittelwert

$$E(X) := \sum_{\omega\in\Omega} X(\omega)p(\omega) \tag{2.21}$$

der Funktionswerte $X(\omega)$ gewichtet mit den Einzelwahrscheinlichkeiten $p(\omega)$.

Ist Ω endlich, so bereitet diese Definition keine Probleme. Im Falle Ω unendlich muss man noch Sorge tragen, dass die Reihe (2.21) absolut konvergiert. Dies ist dann der Fall, wenn die Reihe

$$\sum_{\omega\in\Omega} |X(\omega)|p(\omega)$$

konvergiert, und man sagt in diesem Fall, dass der Erwartungswert $E(X)$ von X existiert.

Beispiel X sei die Augenzahl beim Würfeln eines fairen Würfels
Dann gilt

$$E(X) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{7}{2}\,.$$

Der Erwartungswert stimmt also in diesem Falle mit dem arithmetischen Mittel der Funktionswerte überein.

Es sei X eine Zufallsvariable, deren Erwartungswert existiert. Ist $x_1, x_2, \ldots$ eine Aufzählung des Bildes $X(\Omega)$ von X, so folgt

$$\begin{aligned} E(X) &= \sum_{\omega\in\Omega} X(\omega)p(\omega) = \sum_k \sum_{\omega\in\Omega\,:\,X(\omega)=x_k} X(\omega)p(\omega) \\ &= \sum_k x_k P\left(X = x_k\right) = \sum_k x_k p_X(x_k)\,. \end{aligned}$$

Insbesondere gilt also, dass der Erwartungswert einer Zufallsvariablen X **nur von ihrer Verteilung P_X abhängt**!

Rechenregeln für Erwartungswerte

Es seien X, Y Zufallsvariablen, deren Erwartungswerte existieren. Dann gilt:

- **Linearität** $E(aX + bY) = aE(X) + bE(Y)$ für alle a,$b \in \mathbb{R}$.
- **Nichtnegativität** $X \geq 0$ (d.h. $X(\omega) \geq 0$ für alle $\omega \in \Omega$)

$$\Longrightarrow \qquad E(X) \geq 0\,.$$

- **Monotonie** $X \leq Y$ (d.h. $Y - X \geq 0$)

$$\Longrightarrow \qquad E(X) \leq E(Y)\,.$$

- Ist X konstant, also $X = c$ für eine Konstante c (d.h. $X(\omega) = c$ für alle $\omega \in \Omega$), so folgt
$$E(X) = c\,.$$

- **Transformationssatz** Ist $h : \mathbb{R} \to \mathbb{R}$ eine stückweise stetige Funktion und ist $x_1, x_2, x_3, \ldots$ eine Aufzählung des Bildes $X(\Omega)$, so gilt: Der Erwartungswert der Zufallsvariablen $h(X)$ existiert, genau dann wenn die Summe $\sum_k |h(x_k)| p_X(x_k) < \infty$ konvergiert und in diesem Fall ist
$$E\left(h(X)\right) = \sum_k h(x_k) p_X(x_k) \tag{2.22}$$

- Sind X,Y unabhängig, so existiert auch der Erwartungswert von XY, und es gilt
$$E(XY) = E(X)\,E(Y)\,.$$

Beispiele

(i) Sind $X_1, \ldots, X_n$ unabhängig Bernoulli-verteilt mit Erfolgswahrscheinlichkeit p, so folgt
$$E(X_k) = 0 \cdot P(X_k = 0) + 1 \cdot P(X_k = 1) = p\,.$$
Insbesondere gilt für den Erwartungswert der Summe
$$S_n = X_1 + \ldots + X_n$$
$$E(S_n) = E(X_1) + \ldots + E(X_n) = p + \ldots + p = np\,.$$
Da S_n binomialverteilt ist mit Parameter n und p, folgt insbesondere: Für den Erwartungswert einer binomialverteilten Zufallsvariablen S_n mit Parametern n und p gilt:
$$E(S_n) = np\,.$$

Die Anwendung des Transfomrationssatzes ergibt weiterhin, dass für $\alpha \in \mathbb{R}$
$$\begin{aligned} E\left(e^{\alpha X_1}\right) &= e^{\alpha 0} P(X_1 = 0) + e^{\alpha 1} P(X_1 = 1) \\ &= e^{\alpha 0}(1-p) + e^{\alpha 1} p = (1-p) + pe^{\alpha}\,, \end{aligned}$$
also
$$E\left(e^{\alpha X_i}\right) = (1-p) + pe^{\alpha} \qquad \text{für } i = 1, \ldots, n\,,$$
und damit folgt
$$\begin{aligned} E\left(e^{\alpha S_n}\right) &= E\left(e^{\alpha \sum_{i=1}^n X_i}\right) = E\left(e^{\alpha X_1} e^{\alpha X_2} \ldots e^{\alpha X_n}\right) \\ &= E\left(e^{\alpha X_1}\right) E\left(e^{\alpha X_2}\right) \ldots E\left(e^{\alpha X_n}\right) = (1 - p + pe^{\alpha})^n\,. \end{aligned}$$

(ii) Ist X Poiss (λ)-verteilt, so folgt

$$E(X) = \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!}$$
$$= \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda\,\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda\,.$$

Weiterhin folgt mit dem Transformationssatz

$$E\left(e^{\alpha X}\right) = \sum_{k=0}^{\infty} e^{\alpha k}P(X=k) = \sum_{k=0}^{\infty} e^{\alpha k}e^{-\lambda}\frac{\lambda^k}{k!}$$
$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\left(e^{\alpha}\lambda\right)^k}{k!} = e^{-\lambda}e^{e^{\alpha}\lambda} = e^{-\lambda(1-e^{\alpha})}\,.$$

Ein Maß für die Streuung der Funktionswerte $X(\omega)$ um ihren Erwartungswert $E(X)$ ist die mittlere quadratische Abweichung

$$\mathrm{Var}(X) := E\left((X-E(X))^2\right) = \sum_{\omega\in\Omega}\left(X(\omega)-E(X)\right)^2 p(\omega)\,. \tag{2.23}$$

Sie heißt **Varianz** von X.

Damit der Ausdruck (2.23) wohldefiniert ist, müssen die Erwartungswerte $E(X)$ und $E\left((X-E(X))^2\right)$ existieren. Man kann zeigen, dass beide existieren, falls der Erwartungswert $E(X^2)$ von X^2 existiert.

Unter der **Standardabweichung** von X versteht man die Größe

$$s_X := \sqrt{\mathrm{Var}(X)}\,.$$

Wie der Erwartungswert, so hängt auch die Varianz (und damit auch die Standardabweichung) nur von der Verteilung P_X von X unter P ab. Ist nämlich $x_1, x_2, x_3 \ldots$ eine Aufzählung der Werte von X, so folgt aus dem Transformationssatz

$$\mathrm{Var}(X) = E\left((X-E(X))^2\right) = \sum_{k}\left(x_k-E(X)\right)^2 p_X(x_k)\,.$$

Beispiel X sei die Augenzahl beim Würfeln eines fairen Würfels. Dann folgt

$$\mathrm{Var}(X) = \left(1-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \left(2-\frac{7}{2}\right)^2\cdot\frac{1}{6} + \ldots + \left(6-\frac{7}{2}\right)^2\cdot\frac{1}{6} = \frac{35}{12}\,.$$

Rechenregeln für Varianzen

Es seien $X, Y, X_1, \ldots, X_n$ Zufallsvariablen, für die die Erwartungswerte $E(X^2)$, $E(Y^2)$, $E(X_1^2), \ldots, E(X_n^2)$ existieren. Dann gilt:

- $\mathsf{Var}(aX+b) = a^2\mathsf{Var}(X)$ für alle $a, b \in \mathbb{R}$.
  Denn aus $E(aX+b) = aE(X)+b$ folgt

$$\mathsf{Var}(aX+b) = E\left((aX+b-E(aX+b))^2\right) = E\left((aX-aE(X))^2\right) = a^2\,\mathsf{Var}(X)\,.$$

- Verschiebungssatz $\mathsf{Var}(X) = E(X^2) - (E(X))^2$.
  Denn

$$\begin{aligned}\mathsf{Var}(X) &= E\left((X-E(X))^2\right) = E\left(X^2 - 2XE(X) + (E(X))^2\right)\\ &= E(X^2) - 2(E(X))^2 + (E(X))^2 = E(X^2) - (E(X))^2\,.\end{aligned}$$

- X, Y unabhängig $\Rightarrow \mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.
  Denn

$$\begin{aligned}\mathsf{Var}(X+Y) &= E\left((X+Y)^2\right) - (E(X+Y))^2\\ &= E(X^2+2XY+Y^2) - (E(X)+E(Y))^2\\ &= E(X^2) + 2E(XY) + E(Y^2) - \left(E(X)^2 + 2E(X)E(Y) + E(Y)^2\right)\\ &= E(X^2) - (E(X))^2 + E(Y^2) - ((Y))^2 + 2\left(E(XY) - E(X)E(Y)\right)\\ &= \mathsf{Var}(X) + \mathsf{Var}(Y) - 2\left(E(XY) - E(X)E(Y)\right)\,.\end{aligned}$$

  Da X und Y unabhängig, folgt $E(XY) = E(X)E(Y)$, und damit verschwindet der dritte Term auf der rechten Seite.

Allgemeiner gilt die **Identität von Bienaymé**

Sind $X_1 \ldots, X_n$ unabhängig, so folgt

$$\mathsf{Var}(X_1 + \ldots + X_n) = \mathsf{Var}(X_1) + \ldots + \mathsf{Var}(X_n)\,.$$

Beispiele

(i) Sind $X_1, \ldots, X_n$ unabhängig Bernoulli-verteilt mit Erfolgswahrscheinlichkeit p, so folgt für die Varianz der Summe $S_n = X_1 + \ldots + X_n$

$$\mathsf{Var}(S_n) = \mathsf{Var}(X_1 + \ldots + X_n) = \mathsf{Var}(X_1) + \ldots + \mathsf{Var}(X_n)\,.$$

Für die Varianz der Bernoulli-verteilten Zufallsvariablen X_k errechnet man sofort

$$\mathsf{Var}(X_k) = E\left(X_k^2\right) - \left(E(X_k)\right)^2 = p - p^2 = p(1-p)\,,$$

so dass

$$\mathsf{Var}(S_n) = np(1-p)\,.$$

Da S_n binomialverteilt ist mit Parameter n und p, folgt insbesondere: Für die Varianz einer binomialverteilten Zufallsvariablen S_n mit Parameter n und p gilt

$$\mathsf{Var}(S_n) = np(1-p)\,.$$

(ii) Ist X Poiss (λ)-verteilt, so folgt

$$\begin{aligned} E\left(X^2\right) &= \sum_{k=0}^{\infty} k^2 P(X=k) = \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=0}(k-1+1)e^{-\lambda}\frac{\lambda\cdot\lambda^{k-1}}{(k-1)!} \\ &= \lambda\sum_{k=1}^{\infty}(k-1)e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} + \lambda\sum_{k=1}^{\infty}e^{-\lambda}\frac{\lambda\cdot\lambda^{k-1}}{(k-1)!} \\ &= \lambda\sum_{k=0}^{\infty}ke^{-\lambda}\frac{\lambda^k}{k!} + \lambda = \lambda^2+\lambda\,, \end{aligned}$$

also

$$\mathsf{Var}(X) = E\left(X^2\right) - (E(X))^2 = \lambda^2+\lambda-\lambda^2 = \lambda\,.$$

Kovarianz

Sind X und Y zwei Zufallsvariablen, deren Varianzen existieren, so ist die **Kovarianz**

$$\mathsf{Cov}(X,Y) := E\left((X-E(X))(Y-E(Y))\right)$$

wohldefiniert. Sie ist das Analogon zur empirischen Kovarianz einer zweidimensionalen Messreihe. Die Größe

$$\varrho(X,Y) := \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}}$$

heißt dementsprechend der **Korrelationskoeffizient** von X und Y. Ist $\varrho(X,Y)=0$, so heißen X und Y **unkorreliert**.

Die Kovarianz hängt nur von der **gemeinsamen Verteilung** P_{XY} der Zufallsvariablen X und Y unter P ab. Hierunter versteht man die diskrete Wahrscheinlichkeitsverteilung zur Wahrscheinlichkeitsfunktion

$$p_{XY}(x,y) := P(X=x, Y=y)\,, \quad x\in X(\Omega), y\in Y(\Omega)$$

auf dem Produktraum $X(\Omega)\times Y(\Omega) := \{(x,y)\,:\,x\in X(\Omega), y\in Y(\Omega)\}\subset\mathbb{R}^2$.

Ist nämlich $x_1,x_2,x_3,\ldots$ eine Aufzählung der Werte von X und $y_1,y_2,y_3,\ldots$ eine Aufzählung der Werte von Y, so folgt

$$\begin{aligned} \mathsf{Cov}(X,Y) &= \sum_{\omega\in\Omega}\left(X(\omega)-E(X)\right)\left(Y(\omega)-E(Y)\right) \\ &= \sum_k\sum_l \sum_{\omega\in\Omega\,:\,X(\omega)=x_k, Y(\omega)=y_l} (x_k-E(X))\,(y_l-E(Y)) \\ &= \sum_k\sum_l (x_k-E(X))\,(y_l-E(Y))\,p_{XY}(x_k,y_l)\,. \end{aligned}$$

Rechenregeln für Kovarianzen

- $\mathrm{Cov}((aX+b),(cY+d)) = ac\ \mathrm{Cov}(X,Y)$ für alle $a,b,c,d \in \mathbb{R}$.

  Denn

$$\begin{aligned}\mathrm{Cov}(aX+b,cY+d) &= E\left((aX+b-E(aX+b))(cY+d-E(cY+d))\right)\\ &= E\left(a(X-E(X))c(Y-E(Y))\right) = ac\ \mathrm{Cov}(X,Y)\,.\end{aligned}$$

- Verschiebungssatz $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$.

  Denn

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\left((X-E(X))(Y-E(Y))\right)\\ &= E\left(XY - XE(Y) - E(X)Y) + E(X)E(Y)\right)\\ &= E(XY) - 2E(X)E(Y) + E(X)E(Y) = E(XY) - E(X)E(Y)\end{aligned}$$

- Insbesondere: X, Y unabhängig $\Rightarrow \mathrm{Cov}(X,Y) = 0$. Die Umkehrung gilt im allgemeinen nicht.

4. Stetige Verteilungen

In vielen Fällen kann der Wertebreich einer Zufallsvariablen X nicht diskret gewählt werden (z.B. Wartezeiten, Laufzeiten, Körpergröße, Lufttemperatur,...) sondern muss als Intervall $[a, b]$ oder gleich ganz $\mathbb{R}$ gewählt werden. Eine solche Zufallsvariable kann natürlich nicht auf einem diskreten Wahrscheinlichkeitsraum (Ω, P) definiert sein. Es bedarf hierzu also einer Erweiterung des Begriffes des Wahrscheinlichkeitsraumes auf überabzählbare Ergebnismengen Ω. Die mathematische Theorie zur rigorosen Durchführung dieser Erweiterung sprengt eindeutig den Rahmen dieser Vorlesung, man findet sie in Büchern zur Wahrscheinlichkeitstheorie.

Im folgenden betrachten wir nur den für die Anwendungen enorm wichtigen Fall stetig verteilter Zufallsvariablen X. Dabei heißt X **stetig verteilt mit Dichte f**, falls gilt

$$P(X \leq b) = \int_{-\infty}^{b} f(x)\,dx \qquad \text{für alle } b \in \mathbb{R}\,. \tag{2.24}$$

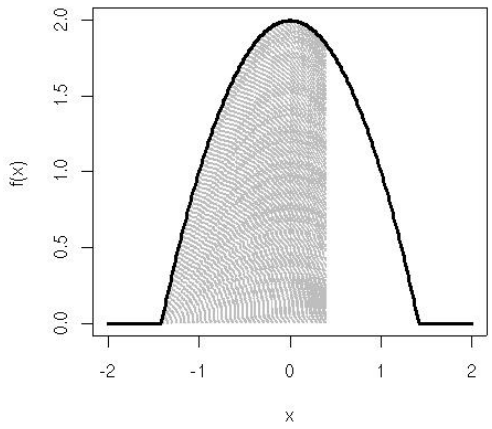
Hierbei ist $f : \mathbb{R} \to \mathbb{R}$ eine uneigentlich Riemann-integrierbare Funktion mit

- $f(x) \geq 0$ für alle $x \in \mathbb{R}$,
- $\int_{-\infty}^{+\infty} f(x)\,dx = 1$.

Für eine mit Dichte f stetig verteilte Zufallsvariable X wird also die Wahrscheinlichkeit der Ereignisse

$$\{\omega : X(\omega) \leq b\}$$

durch die schraffierte Fläche angegeben.



Wie im Falle diskreter Zufallsvariablen heißt die Funktion

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

die Verteilungsfunktion von X. Sie besitzt genau dieselben Eigenschaften wie im diskreten Fall, d.h.

- F ist monoton wachsend

- $0 \leq F \leq 1$, $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$
- F ist (rechtsseitig) stetig.

Ist X stetig verteilt mit Verteilungsfunktion F und ist $p \in (0,1)$, so heißt jede Zahl $x_p \in \mathbb{R}$ mit

$$F(x_p) = p$$

p-Quantil der Verteilung von X. Ist F streng monoton steigend, d.h., $F(x) < F(y)$ für $x < y$, so ist $x_p = F^{-1}(p)$ eindeutig bestimmt durch den Wert der Umkehrfunktion F^{-1} von F in p.

Wie im Falle empirischer Verteilungen bezeichnet man

- $x_{0.5}$ als Median,
- $x_{0.25}$ als unteres Quartil,
- $x_{0.75}$ als oberes Quartil.

Mit Hilfe von (2.24) können wir dann auch sofort die Wahrscheinlichkeit des Ereignisses $\{\omega : a < X(\omega) \leq b\}$ berechnen, denn

$$\begin{aligned} P(a < X \leq b) &= P(X \leq b) - P(X \leq a) = F(b) - F(a) \\ &= \int_{\infty}^{b} f(x)\,dx - \int_{\infty}^{a} f(x)\,dx = \int_{a}^{b} f(x)\,dx\,. \end{aligned} \tag{2.25}$$

Für eine stetig verteilte Zufallsvariable X gilt

$$P(X = x) = 0 \qquad \forall x \in \mathbb{R}\,,$$

d.h. X nimmt einen **bestimmten** Wert x nur mit Wahrscheinlichkeit 0 an. Dies ist ein fundamentaler Unterschied zu diskreten Zufallsvariablen. Damit gilt insbesondere

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) \qquad \forall\, a, b \in \mathbb{R}\,. \tag{2.26}$$

Stochastische Unabhängigkeit

Der Begriff der stochastischen Unabhängigkeit lässt sich unmittelbar auf stetig verteilte Zufallsvariablen übertragen: zwei (stetig verteilte) Zufallsvariablen X und Y heißen **stochastisch unabhängig**, falls

$$P(X \leq x\,, Y \leq y) = P(X \leq x)P(Y \leq y) \qquad \forall\, x, y \in \mathbb{R}\,.$$

Allgemeiner: Die (stetig verteilten) Zufallsvariablen $X_1, \ldots, X_n$ heißen **stochastisch unabhängig**, falls

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = P(X_1 \leq x_1)\cdot P(X_2 \leq x_2)\cdot \ldots \cdot P(X_n \leq x_n) \tag{2.27}$$

für alle $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Die Analogie zum diskreten Fall erkennt man wie folgt: Ist $B_i :=]-\infty, x_i]$, so kann man (2.27) in der Form

$$P(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = P(X_1 \in B_1) \cdot P(X_2 \in B_2) \cdot \ldots \cdot P(X_n \in B_n)$$

schreiben.

Erwartungswert, Varianz und Kovarianz

Ist X stetig verteilt mit Dichte f, so sagen wir, dass der Erwartungswert $E(X)$ von X existiert, falls die Funktion $|x|f(x)$ uneigentlich Riemann-integrierbar ist (dann ist auch $xf(x)$ uneigentlich Riemann-integrierbar) und man setzt in diesem Falle

$$E(X) := \int_{-\infty}^{+\infty} xf(x)\,dx\,.$$

Ist zusätzlich auch die Funktion $(x - E(X))^2 f(x)$ uneigentlich Riemann-integrierbar, so definiert man die Varianz $\mathsf{Var}(X)$ durch

$$\mathsf{Var}(X) := \int_{-\infty}^{+\infty} (x - E(x))^2 f(x)\,dx$$

und die Standardabweichung wie im diskreten Fall durch

$$s_X := \sqrt{\mathsf{Var}(X)}\,.$$

Die Rechenregeln für Erwartungswerte und Varianz diskret verteilter Zufallsvariablen (siehe Abschnitt II.3) übertragen sich unmittelbar auf den Fall stetig verteilter Zufallsvariablen. Der Transformationssatz überträgt sich dabei wie folgt: Ist $h : \mathbb{R} \to \mathbb{R}$ eine stückweise stetige Funktion so gilt: Der Erwartungswert der Zufallsvariablen $h(X)$ existiert genau dann wenn die Funktion $|h(x)|f(x)$ uneigentlich Riemann-integrierbar ist und in diesem Fall ist

$$E\left(h(X)\right) = \int_{-\infty}^{+\infty} h(x)f(x)\,dx\,. \tag{2.28}$$

Zwei Zufallsvariablen X und Y heißen **gemeinsam stetig verteilt** mit **gemeinsamer stetiger Dichte** f_{XY}, falls gilt

$$P(X \le a, Y \le b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{XY}(x,y)\,dy\,dx \qquad \forall a, b \in \mathbb{R}$$

für eine integrierbare Funktion $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ mit

- $f_{XY}(x,y) \ge 0$ für alle $x, y \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dx\,dy = 1$.

Die Berechnung der Kovarianz $\mathrm{Cov}(X,Y)$ erfolgt dann über die gemeinsame Dichte mit Hilfe der Formel

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(X))(y-E(Y))f_{XY}(x,y)\,dx\,dy\,.$$

Die Rechenregeln für die Kovarianzen für diskret verteilte Zufallsvariablen übertragen sich Wort für Wort auf den gemeinsam stetig verteilten Fall.

Wichtige stetige Verteilungen

Gleichverteilung

Für $a<b$ heißt eine Zufallsvariable mit Dichte

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x\in[a,b] \\ 0 & \text{sonst} \end{cases}$$

(stetig) gleichverteilt auf $[a,b]$. Die zugehörige Verteilungsfunktion ist

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

Für alle Teilintervalle $[c,d]$ folgt aus (2.25) und (2.26)

$$P(c\le X\le d) = F(d)-F(c) = \frac{d-a}{b-a}-\frac{c-a}{b-a} = \frac{d-c}{b-a}\,.$$

Mit anderen Worten: X überdeckt Teilintervalle derselben Länge $d-c$ mit jeweils derselben Wahrscheinlichkeit. Dies erklärt die Bezeichnung Gleichverteilung.

X nimmt mit Wahrscheinlichkeit 1 nur Werte in $[a,b]$ an, denn

$$P(X\in[a,b]) = P(a\le X\le b) = \frac{b-a}{b-a} = 1\,.$$

Für Erwartungswert und Varianz einer auf $[a,b]$ gleichverteilten Zufallsvariablen gilt

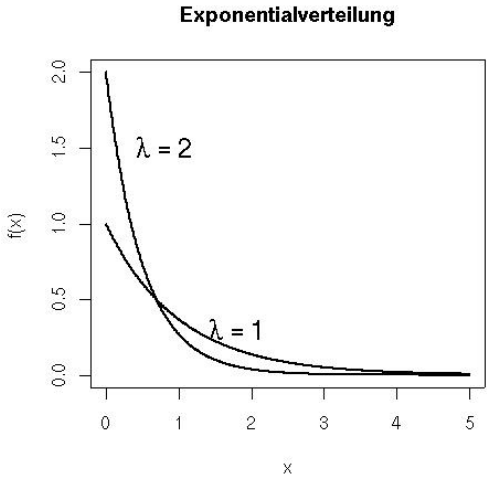
$$\begin{aligned} E(X) &= \int_{-\infty}^{+\infty} xf(x)\,dx = \int_a^b x\frac{1}{b-a}\,dx = \frac{1}{2}\frac{x^2}{b-a}\Big|_a^b = \frac{1}{2}(a+b) \\ \mathrm{Var}(X) &= \int_{-\infty}^{+\infty}\left(x-\frac{1}{2}(a+b)\right)^2 f(x)\,dx \\ &= \int_a^b\left(x-\frac{1}{2}(a+b)\right)^2\frac{1}{b-a}\,dx = \frac{1}{12}(b-a)^2\,. \end{aligned}$$

Exponentialverteilung

Für $\lambda>0$ ist

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & \text{für } x\ge 0 \\ 0 & \text{sonst} \end{cases}$$

eine Dichte. Die zugehörige Verteilung heißt **Exponentialverteilung** zum Parameter λ. Sie wird mit Exp(λ) bezeichnet.



Die zugehörige Verteilungsfunktion ist

$$F(x) = \begin{cases} 0 & \text{für } x < 0 \\ 1 - e^{-\lambda x} & \text{für } x \geq 0\,. \end{cases}$$

Die Exponentialverteilung ist das stetige Analogon der geometrischen Verteilung, die ja die Verteilung von Wartezeiten auf den ersten Erfolg in einer Folge von unabhängigen Bernoulli Experimenten beschreibt. Dementsprechend verwendet man die Exponentialverteilung zur Modellierung von stetig verteilten Wartezeiten.

Ist X Exp(λ) verteilt, so gilt

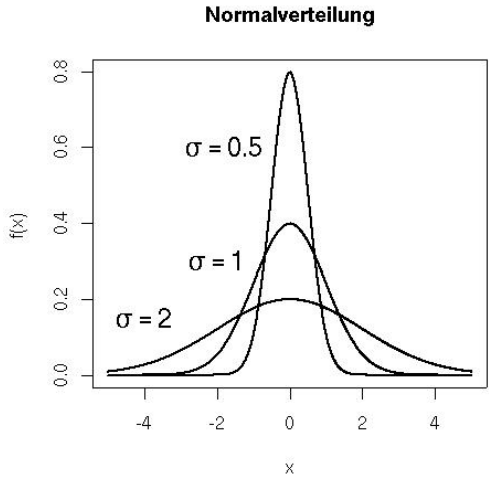
$$E(X) = \lambda \int_0^{+\infty} x e^{-\lambda x}\, dx = \frac{1}{\lambda}$$
$$\text{Var}(X) = \lambda \int_0^{+\infty} \left(x - \frac{1}{\lambda}\right)^2 e^{-\lambda x}\, dx = \frac{1}{\lambda^2}\,.$$

Normalverteilung

Für $m \in \mathbb{R}$ und $\sigma > 0$ ist

$$f_{m,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}$$

eine Dichte. Die zugehörige Verteilung heißt **Normalverteilung** mit Mittel m und Varianz σ^2. Sie wird mit $N(m, \sigma^2)$ bezeichnet. Im Falle $m = 0$ und $\sigma^2 = 1$ spricht man von der **Standardnormalverteilung**.



f_{m,σ^2} besitzt ein absolutes Maximum in $x = m$ und Wendepunkte in $m \pm \sigma$. Wegen ihrer Form wird f auch als **Gaußsche Glockenkurve** bezeichnet. σ bestimmt **Breite** und **Höhe** der Glockenkurve.

Eine Zufallsvariable X mit Dichte f_{m,σ^2} heißt normalverteilt mit Mittel m und Varianz σ^2, denn es gilt

$$E(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}\, dx = m$$
$$\mathrm{Var}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-m)^2 e^{-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}}\, dx = \sigma^2\,.$$

Eigenschaften normalverteilter Zufallsvariablen

- Die Werte der Verteilungsfunktion der Standardnormalverteilung

$$\Phi(x) := P(Y \le x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt \text{ für } x \ge 0$$

  findet man tabelliert in Formelsammlungen und in jeder guten Programmbibliothek. Da die Dichte $f_{0,1}$ der Standardnormalverteilung eine gerade Funktion ist (also $f_{0,1}(x) = f_{0,1}(x)$), ergibt sich

$$\Phi(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-\frac{t^2}{2}}\, dt = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{t^2}{2}}\, dt = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt = 1 - \Phi(x)\,,$$

  also

$$\Phi(-x) = 1 - \Phi(x) \qquad \text{für alle } x \in \mathbb{R}\,, \tag{2.29}$$

  woraus sich dann auch die Werte $\Phi(x)$ für $x \le 0$ berechnen lassen. Für die p-Quantile z_p der Standardnormalverteilung gilt wegen (2.29)

$$z_p = -z_{1-p}\,.$$

- Ist X eine $N(m, \sigma^2)$-verteilte Zufallsvariable, so ist

$$Y = \frac{X - m}{\sigma}$$

eine $N(0,1)$-verteilte, also standardnormalverteilte, Zufallsvariable. Man kann also die Berechnung der Wahrscheinlichkeiten $P(X \leq b)$ zurückführen auf die Berechnung entsprechender Wahrscheinlichkeiten einer standardnormalverteilten Zufallsvariablen

$$P(X \leq b) = P\left(\frac{X-m}{\sigma} \leq \frac{b-m}{\sigma}\right) = P\left(Y \leq \frac{b-m}{\sigma}\right). \tag{2.30}$$

Mit Hilfe der Verteilungsfunktion Φ der Standardnormalverteilung berechnet man dann

$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-m}{\sigma} \leq Y \leq \frac{b-m}{\sigma}\right) \\ &= \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right). \end{aligned} \tag{2.31}$$

- Sind X_i, $i = 1 \ldots n$, unabhängig normalverteilt mit Mittel m_i und Varianz σ_i^2, so ist die Summe $S_n = X_1 + \ldots + X_n$ wieder normalverteilt mit Mittel $\sum_{i=1}^n m_i$ und Varianz $\sum_{i=1}^n \sigma_i^2$.

Anwendung: Konfidenzschätzungen

Im Vorgriff auf das nächste Kapitel wollen wir im folgenden eine der wichtigsten Anwendungen der Normalverteilung in der Statistik diskutieren.

Eine Messreihe $X_1, \ldots, X_n$ unterliegt in der Regel zufälligen Mess- oder Beobachtunsfehlern. Daher können $X_1, \ldots, X_n$ auch als Zufallsvariablen angesehen werden. Als Verteilung empfiehlt sich in der Regel eine Normalverteilung $N(m, \sigma^2)$ für unbekannte m und σ^2. Als Schätzungen für m und σ^2 wählt man naheliegenderweise das

- empirische Mittel $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ für m und die
- Stichprobenvarianz $s_X^2 := \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ für σ^2.

Aussagen über **Genauigkeit** und **Sicherheit** dieser Schätzung liefern **Konfidenzschätzungen:** Von zentraler Bedeutung ist die Wahrscheinlichkeit

$$P\left(\left|\overline{X} - m\right| \leq t \cdot \frac{s_X}{\sqrt{n}}\right) \tag{2.32}$$

dafür, dass das Mittel m in einem vorgegebenen **Vertrauensbereich** (bzw. **Konfidenzintervall**) der Form

$$\left[\overline{X} - t\frac{s_X}{\sqrt{n}}, \overline{X} + t\frac{s_X}{\sqrt{n}}\right]$$

liegt. Für große Stichproben $(n \geq 30)$ wird die gesuchte Wahrscheinlichkeit angenähert durch die Standardnormalverteilung

$$P\left(\left|\overline{X} - m\right| \leq t\frac{s_X}{\sqrt{n}}\right) = P\left(\left|\sqrt{n}\left(\frac{\overline{X}-m}{s_X}\right)\right| \leq t\right) \sim 2\Phi(t) - 1\,.$$

Man spricht in diesem Zusammenhang auch von einer Normalapproximation.

In der Praxis geht man von einem Vertrauensniveau γ aus (z. B. $\gamma = 95\%$) und fragt nach dem **Vertrauensbereich** für m. Zum Beispiel für $\gamma = 95\%$ ist $t = 1.96$. Mit einer Sicherheit von 95% liegt also der unbekannte Erwartungswert m im Intervall

$$\left[\overline{X} - 1.96\frac{s_X}{\sqrt{n}}, \overline{X} + 1.96\frac{s_X}{\sqrt{n}}\right] .$$

Für $n < 30$ muss obige Wahrscheinlichkeit (2.32) mit Hilfe der t-Verteilung approximiert werden (s.u.). Man erhält z.B. für $\gamma = 95\%$ und $n = 10$ den Wert $t = 2.26$. Mit einer Sicherheit von 95% liegt der unbekannte Erwartungswert m im Intervall $\left[\overline{X} - \frac{2.26}{\sqrt{10}}s_X, \overline{X} + \frac{2.26}{\sqrt{10}}s_X\right]$.

Zum Abschluss dieses Abschnitts noch einige weitere für die induktive Statistik wichtige stetige Verteilungen in einer Übersicht.

χ^2-Verteilung

Es seien $X_1, \ldots, X_n$ unabhängig $N(0,1)$-verteilte Zufallsvariablen. Dann heißt die Verteilung der Zufallsvariablen

$$Z_n = X_1^2 + \ldots + X_n^2$$

χ^2_n-Verteilung (oder **χ^2-Verteilung mit n Freiheitsgraden**).

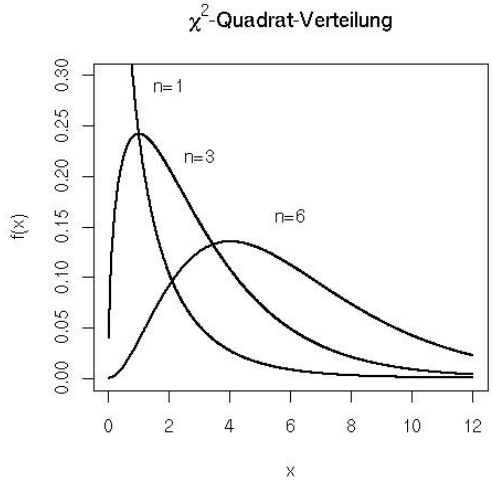
Aus den Rechenregeln für Erwartungswert und Varianz folgt sofort

$$E(Z_n) = n\,, \mathsf{Var}(Z_n) = \underbrace{\mathsf{Var}(X_1^2)}_{=2} + \cdots + \underbrace{\mathsf{Var}(X_1^2)}_{=2} = 2n\,.$$

Die Dichte g_n der χ^2_n-Verteilung hat die Form

$$g_n(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{für } x > 0 \\ 0 & \text{sonst} \end{cases}$$

Für wachsendes n nähern sich die Dichten g_n der Gaußschen Glockenkurve an, weshalb man ab $n > 30$ eine Normalverteilungsapproximation wählt.



Hinweis zur Normalapproximation für $n > 30$: Die naheliegende Approximation der χ^2_n-Verteilung durch $N(n, 2n)$ legt eine Approximation der p-Quantile $\chi^2_{n;p}$ der χ^2_n-Verteilung

durch die entsprechenden p-Quantile der Normalverteilung $N(n, 2n)$ nahe. Eine bessere Approximation ist aber

$$\chi^2_{n;p} \sim \frac{1}{2}\left(z_p + \sqrt{2n-1}\right)^2$$

siehe [2] (Seite 303).

t-Verteilung

Es seien X und Z_n unabhängig, X $N(0,1)$-verteilt und Z_n χ^2_n-verteilt. Dann heißt die Verteilung der Zufallsvariablen

$$T_n := \frac{X}{\sqrt{Z_n/n}}$$

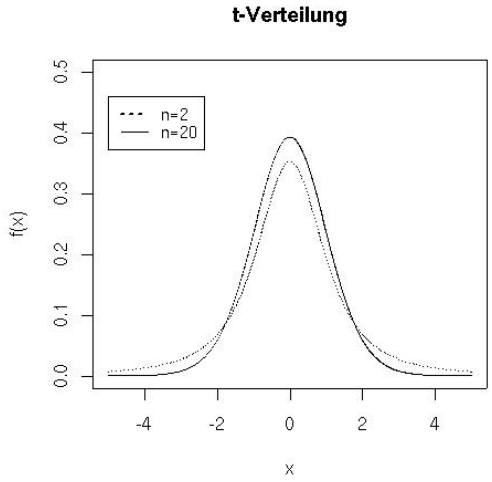
t_n-Verteilung (oder **t-Verteilung mit n Freiheitsgraden**).
Es gilt

$$E(T_n) = 0\,, \mathrm{Var}(T_n) = \frac{n}{n-2} \text{ für } n \geq 3\,.$$

Die Dichte h_n der t_n-Verteilung ist gegeben durch

$$h_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\sqrt{n}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}\,.$$

Die Dichte h_n hat eine ähnliche Form wie die Gaußsche Glockenkurve, jedoch für kleine n breitere Enden als die Standardnormalverteilung. Für $n > 30$ ist jedoch eine Approximation durch die Standardnormalverteilung bereits sehr gut.



Wie für die Quantile der Standardnormalverteilung gilt auch für die Quantile $t_{n;p}$ der t_n-Verteilung

$$t_{n;p} = -t_{n;1-p}\,.$$

F-Verteilung (Fisher-Verteilung)

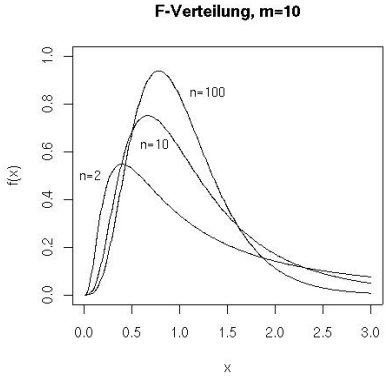
Es seien Z_m und $\tilde{Z}_n$ unabhängig, Z_m χ^2_m-verteilt, $\tilde{Z}_n$ χ^2_n-verteilt. Dann heißt die Verteilung der Zufallsvariablen

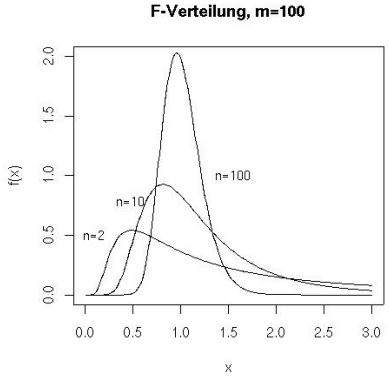
$$Z_{m,n} := (Z_m/m) \,/\, \left(\tilde{Z}_n/n\right)$$

$F_{m,n}$-**Verteilung** (oder F-**Verteilung mit** m **und** n **Freiheitsgraden**).

Es gilt

$$E\left(Z_{m,n}\right) = \frac{n}{n-2} \quad \text{für } n \geq 3$$
$$\text{Var}\left(Z_{m,n}\right) = \frac{2n^2(n+m-2)}{m(n-4)(n-2)^2} \quad \text{für } n \geq 5\,.$$





Für die Quantile $F_{m,n;p}$ der $F_{m,n}$-Verteilung gilt

$$F_{m,n;p} = \frac{1}{F_{n,m;1-p}}\,,$$

denn

$$p = P\left(Z_{m,n} \leq F_{m,n;p}\right) = P\left(\frac{Z_m}{m}\Big/\frac{\tilde{Z}_n}{n} \leq F_{m,n;p}\right)$$
$$= P\left(\frac{\tilde{Z}_n}{n}\Big/\frac{Z_m}{m} \geq \frac{1}{F_{m,n;p}}\right) = 1 - P\left(\frac{\tilde{Z}_n}{n}\Big/\frac{Z_m}{m} \leq \frac{1}{F_{m,n;p}}\right)\,.$$

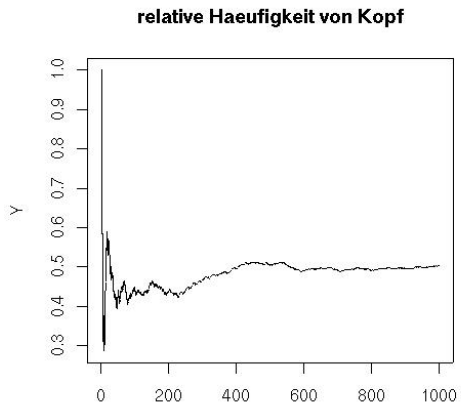
5. Grenzwertsätze

(A) Gesetz der großen Zahlen und der Hauptsatz der Statistik

Werfen wir eine faire Münze n mal und setzen wir $X_k = 1$ (bzw. $X_k = 0$) falls beim k-ten Münzwurf Kopf (bzw. Zahl) oben liegt, so nähert sich die relative Häufigkeit für Kopf

$$\frac{1}{n}\sum_{k=1}^{n} X_k(\omega)$$

für wachsendes n mit großer Wahrscheinlichkeit der *theoretischen* Wahrscheinlichkeit $\frac{1}{2}$ für Kopf. Man bezeichnet $\frac{1}{n}\sum_{k=1}^{n} X_k(\omega)$ auch als *empirisches Mittel* und $m = E(X_k) = \frac{1}{2}$ als *theoretisches Mittel*. Bei vielfacher Wiederholung des Münzwurfes stellt man fest, dass sich das empirische Mittel für wachsende n dem theoretischen Mittel annähert.
In der folgenden Grafik ist als Illustration die Folge der empirischen Mittel für insgesamt 1000 Münzwürfe aufgetragen.



Diese Beobachtung gilt ganz allgemein für die relativen Häufigkeiten eines beliebigen Ereignisses in einer unabhängigen Wiederholung ein und desselben Zufallsexperimentes. Sie wird als Gesetz der großen Zahlen bezeichnet.

Satz (Gesetz der großen Zahlen) Es sei $X_1, X_2, \ldots$ eine Folge unabhängiger Zufallsvariablen mit gemeinsamem Erwartungswert $E(X_k) = m$ und gemeinsamer Varianz $\mathrm{Var}(X_k) = \sigma^2$. Dann folgt für alle $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left\{\omega \,:\, \left|\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - m\right| \geq \varepsilon\right\}\right) = 0\,.$$

Die obige Aussage zur Asymptotik der relativen Häufigkeiten eines Ereignisses A leitet sich aus dem Satz wie folgt ab: Es sei

$$X_k(\omega) := \begin{cases} 1 & \text{falls } A \text{ in der k-ten Wiederholung eintritt} \\ 0 & \text{sonst.} \end{cases}$$

Dann sind die $X_1, X_2, \ldots$ eine Folge unabhängig Bernoulli-verteilter Zufallsvariablen mit Parameter $p := P(A) = E(X_k)$. Für die relativen Häufigkeiten $f_{n,\omega}(A) := \frac{1}{n}\sum_{k=1}^{n} X_k(\omega)$

des Ereignisses A in n Wiederholungen gilt dann die Aussage des Gesetzes der großen Zahlen:

$$\lim_{n\to\infty} P\left(\{\omega \,:\, |f_{n,\omega}(A) - P(A)| \geq \varepsilon\}\right) = 0 \qquad \forall \varepsilon > 0\,.$$

Exkurs: Tschebychev-Ungleichung

Der Beweis des Gesetzes beruht im wesentlichen auf folgender Ungleichung:

Satz (Tschebychevsche Ungleichung)

Es sei X eine Zufallsvariable für die der Erwartungswert von X^2 existiert. Dann gilt für alle $c > 0$

$$P\left(\{\omega \,:\, |X(\omega) - E(X)| \geq c\}\right) \leq \frac{1}{c^2}\mathsf{Var}(X)\,. \tag{2.33}$$

Beweis Wir geben den Beweis nur im Falle eines diskreten Wahrscheinlichkeitsraumes. Offenbar gilt

$$\begin{aligned} P\left(\{\omega \,:\, |X(\omega) - E(X)| \geq c\}\right) &= \sum_{\omega\in\Omega\,:\,|X(\omega)-E(X)|\geq c} p(\omega) \\ &\leq \sum_{\omega\in\Omega\,:\,|X(\omega)-E(X)|\geq c} \left(\frac{X(\omega)-E(X)}{c}\right)^2 p(\omega) \\ &\leq \sum_{\omega\in\Omega} \left(\frac{X(\omega)-E(X)}{c}\right)^2 p(\omega) = E\left(\left(\frac{X-E(X)}{c}\right)^2\right) \\ &= \frac{1}{c^2}\mathsf{Var}(X)\,. \end{aligned}$$

Dabei haben wir in der zweiten Ungleichung verwandt, dass $\left(\frac{X(\omega)-E(X)}{c}\right)^2 \geq 0$ für alle $\omega \in \Omega$, und damit die Summe von $\left(\frac{X(\omega)-E(X)}{c}\right)^2 p(\omega)$ über alle $\omega \in \Omega$ nicht kleiner sein kann als die Teilsumme. □

Analog zur empirischen Tschebychev-Ungleichung (siehe Kapitel I.2) quantifizert die Tschebychev-Ungleichung (2.33) der Streuung der Funktionswerte von X um den Erwartungswert $m = E(X)$: Für $k > 0$ kann die Wahrscheinlichkeit, dass X einen Wert annimmt im Intervall

$$[m - ks_X, m + ks_X] \tag{2.34}$$

mit Hilfe der Tschebychev Ungleichung abgeschätzt werden durch

$$P\left(m - ks_X \leq X \leq m + ks_X\right) \geq 1 - \frac{1}{k^2}\,.$$

Insbesondere:

- $P\left(m - \sqrt{2}s_X \leq X \leq m + \sqrt{2}s_X\right) \geq \frac{1}{2}$
- $P\left(m - 2s_X \leq X \leq m + 2s_X\right) \geq \frac{3}{4}$
- $P\left(m - 3s_X \leq X \leq m + 3s_X\right) \geq \frac{9}{10}$.

Begründung Aus der Tschebychevschen Ungleichung (2.33) folgt

$$\begin{aligned} P\left(m - ks_X \leq X \leq m + ks_X\right) &= P\left(|X - m| \leq ks_X\right) \\ &\geq P\left(|X - m| < ks_X\right) \\ &= 1 - P\left(|X - m| \geq ks_X\right) \\ &\geq 1 - \frac{1}{\left(ks_X\right)^2}\,\mathsf{Var}(X) = 1 - \frac{1}{k^2}\,. \end{aligned}$$

Bemerkung Da man die Standardabweichung s_X einer Zufallsvariablen häufig mit σ bezeichnet, bekommt das Intervall in (2.34) die Bezeichnung $[m-k\sigma, m+k\sigma]$ und man spricht aus diesem Grund auch von den $k\,\sigma$-Bereichen der Zufallsvariablen X.

Beweis des Gesetzes der großen Zahlen

Es sei n fest gewählt. Wir definieren die Zufallsvariable

$$Y := \frac{1}{n}\sum_{k=1}^{n} X_k\,.$$

Da $E(X_k) = m$ für alle k, folgt aus der Linearität des Erwartungswertes

$$E(Y) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{k=1}^{n} E(X_k) = m\,.$$

Die Zufallsvariablen $X_1, \ldots, X_n$ sind nach Annahme unabhängig, also besagt die Identität von Bienaymé, dass

$$\mathsf{Var}(Y) = \mathsf{Var}\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathsf{Var}\left(X_k\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}\,.$$

Die Tschebychevsche Ungleichung, angewandt auf Y, ergibt die Abschätzung

$$\begin{aligned} P\left(\left\{\omega\,:\,\left|\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - m\right| \geq \varepsilon\right\}\right) &\leq P\left(\{\omega\,:\,|Y(\omega) - E(Y)| \geq \varepsilon\}\right) \\ &\leq \frac{\mathsf{Var}(Y)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n}\,. \end{aligned}$$

Da $\frac{\sigma^2}{\varepsilon^2 n} \to 0$ für $n \to \infty$, folgt schließlich auch

$$\lim_{n\to\infty} P\left(\left\{\omega\,:\,\left|\frac{1}{n}\sum_{k=1}^{n} X_k(\omega) - m\right| \geq \varepsilon\right\}\right) = 0\,. \quad \square$$

Der Hauptsatz der Statistik

Satz Es sei X eine Zufallsvariable mit Verteilungsfunktion F und es seien $X_1, X_2, \ldots$ eine Folge von unabhängig und identisch verteilten Zufallsvariablen mit derselben Verteilungsfunktion F. Dann gilt für die Folge der zugehörigen **empirischen Verteilungsfunktionen**

$$\begin{aligned} F_{n,\omega}(x) &:= \frac{1}{n}\#\{i \in \{1,\ldots,n\}\,:\,X_i(\omega) \leq x\} \\ &= \text{ relative Häufigkeit } f_{n,\omega}(A) \text{ des Ereignisses } A = \{X \leq x\}\,, \end{aligned}$$

der ersten n Realisierungen $X_1(\omega), \ldots, X_n(\omega)$, dass

$$\lim_{n\to\infty} P\left(\left\{\omega : \sup_{x\in\mathbb{R}} |F_{n,\omega}(x) - F(x)| \geq \varepsilon\right\}\right) = 0 \qquad \forall \varepsilon > 0\,,$$

d.h. die maximale Abweichung zwischen empirischer Verteilungsfunktion $F_{n,\omega}$ und theoretischer Verteilungsfunktion F konvergiert mit wachsendem n mit großer Wahrscheinlichkeit gegen 0.

(B) Der zentrale Grenzwertsatz

Betrachtet man in der Situation des Gesetzes der großen Zahlen mit $m := E(X_k)$ und $\sigma^2 = \mathsf{Var}(X_k)$ die **standardisierten Summen**

$$S_n^* := \frac{S_n - E(S_n)}{\sqrt{\mathsf{Var}(S_n)}} = \frac{\sum_{k=1}^n X_k - nm}{\sqrt{n\sigma^2}} = \sqrt{n}\frac{\overline{X}_n - m}{\sigma}$$

so stellt man fest, dass die zugehörigen Verteilungsfunktionen

$$F_{S_n^*}(x) := P\left(S_n^* \leq x\right) = P\left(\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq x\right)$$

punktweise für alle x gegen die Verteilungsfunktion der Standardnormalverteilung konvergieren, d.h. es gilt

$$\begin{aligned}\lim_{n\to\infty} F_{S_n^*}(x) &= \lim_{n\to\infty} P\left(\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq x\right) = \lim_{n\to\infty} P\left(\frac{S_n - E(S_n)}{\sqrt{\mathsf{Var}(S_n)}} \leq x\right)\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}\,dt = \Phi(x) \qquad \forall x \in \mathbb{R}\,.\end{aligned}$$

Man sagt auch, dass die standardisierten Summen **asymptotisch normalverteilt** sind und bezeichnet die Aussage als **zentralen Grenzwertsatz** .

Man kann dieses Resultat wiederum insbesondere auf die n-fache unabhängige Wiederholung ein und desselben Zufallsexperimentes anwenden:

Ist A ein Ereignis mit Wahrscheinlichkeit $p := P(A)$ und

$$X_k(\omega) := \begin{cases} 1 & \text{falls } A \text{ in der k-ten Wiederholung eintritt}\\ 0 & \text{sonst,}\end{cases}$$

so sind $X_1, X_2, \ldots$ unabhängig Bernoulli verteilt mit $m = E(X_k) = p$ und $\sigma^2 = \mathsf{Var}(X_k) = p(1-p)$ und die standardisierten Häufigkeiten

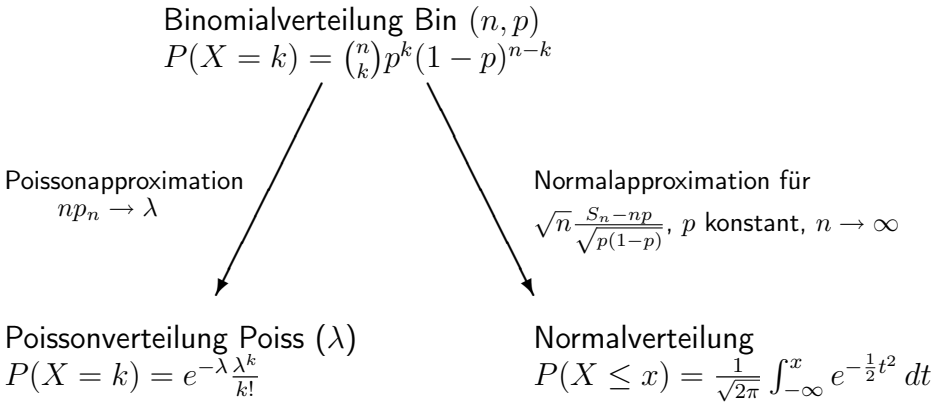
$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}$$

sind asymptotisch normalverteilt, d.h.

$$\lim_{n\to\infty} P\left(S_n^* \leq x\right) = \Phi(x) \quad \text{ für alle } x \in \mathbb{R}\,.$$

Die Bedeutung des zentralen Grenzwertsatzes für die induktive Statistik besteht vor allem darin, dass man aufgrund der Aussage dieses Satzes die Verteilung einer standardisierten Summe S_n^* von unabhängig und identisch verteilten Zufallsvariablen (in der induktiven Statistik: die Stichprobenvariablen) mit wachsendem n (in der induktiven Statistik: mit wachsender Stichprobenlänge) zunehmend besser durch eine Standardnormalverteilung approximieren kann. Diese Approximation heißt **Normalapproximation**.

Zum Abschluss des Kapitels die beiden Approximationen der Binomialverteilung im Überblick:



III Induktive Statistik

Im Gegensatz zur deskriptiven Statistik, die sich auf die Beschreibung von Daten anhand von Kennzahlen und Grafiken beschränkt, versucht die induktive (d.h. die schließende) Statistik von beobachteten Daten auf deren Verteilungen (oder Eigenschaften ihrer Verteilungen) zu schließen. Dies kann zum Beispiel dann notwendig sein, wenn eine vollständige Datenerhebung unmöglich, zu zeitaufwendig oder zu kostspielig ist, wie es etwa bei Umfragen der Fall ist.

In der schließenden Statistik gibt es im Wesentlichen drei zu bearbeitende Problemstellungen:

1. Konstruktion eines **Schätzers** für einen Parameter der unbekannten Verteilung
2. Berechnung von **Konfidenzintervallen**, d.h. von Schranken, die einen unbekannten Parameter mit vorgegebener Wahrscheinlichkeit einfangen.
3. Entwicklung **statistischer Tests**, mit denen vorgegebene Parameter auf Verträglichkeit mit Beobachtungen überprüft werden können.

1. Schätzen

Ausgangspunkt ist wieder eine Grundgesamtheit G von Merkmalsträgern. Unter einer **Stichprobenerhebung** versteht man eine zufällige Entnahme von endlich vielen Objekten aus G. Dabei bedeutet zufällig, dass für jedes Objekt die Wahrscheinlichkeit der Entnahme gleich ist.
In der Sprache der Wahrscheinlichkeitstheorie handelt es sich bei der Stichprobenerhebung um Zufallsexperimente, deren Ausgang man durch Zufallsvariablen

$$X_1, X_2, \ldots, X_n$$

beschreiben kann. In diesem Zusammenhang nennt man die X_i auch **Stichprobenvariablen**. In der Regel betrachtet man nur unabhängige Wiederholungen desselben Zufallsexperiments, d.h. also, dass die Stichprobenvariablen $X_1, \ldots, X_n$ stochastisch unabhängig und identisch verteilt sind.

Unter dem **Stichprobenergebnis** oder der **Stichprobenrealisation** versteht man dann das n-Tupel $(x_1, \ldots, x_n)$ der Realisierung von $X_1, \ldots, X_n$.
Eine **Punktschätzung** ist eine Funktion $g : \mathbb{R}^n \to \mathbb{R}$. Sie ordnet der Stichprobenrealisation $x_1, \ldots, x_n$ den Schätzwert $g(x_1, \ldots, x_n)$ zu.
Die zugehörige **Schätzfunktion** (oder auch **Statistik**) $g(X_1, \ldots, X_n)$ ist diejenige Zufallsvariable, die man durch Einsetzen der Stichprobenvariablen X_i für x_i in die Funktion g erhält.

Beispiele $X_1, \ldots, X_n$ mit Mittel m und Varianz σ^2

Schätzfunktion	Bezeichnung	Erwartungswert	Varianz
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	Stichprobenmittel	m	$\frac{\sigma^2}{n}$
$\sqrt{n}\frac{\overline{X}-m}{\sigma}$	Gauß-Statistik	0	1
$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$	mittlere quadratische Abweichung	$\frac{n-1}{n}\sigma^2$	
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$	Stichprobenvarianz	σ^2	
$S = \sqrt{S^2}$	Stichprobenstandard - abweichung		
$\sqrt{n}\frac{\overline{X}-m}{S}$	t-Statistik		

Im Folgenden wollen wir annehmen, dass die (unbekannte) Verteilung der **Stichprobenvariablen** aus einer Menge möglicher Verteilungen stammt, die über einen Parameter $\theta \in \Theta$ parametrisiert sind.

Beispiel X_i seien $N(m, \sigma^2)$-verteilt mit unbekanntem Mittel m und unbekannter Varianz σ^2. In diesem Falle ist also $\theta = (m, \sigma^2)$ aus $\Theta = \mathbb{R} \times]0, \infty[$ eine (mögliche) Parametrisierung der zugrundeliegenden Verteilungen.

Ist nun $T = g(X_1, \ldots X_n)$ ein Schätzer, so wird der Erwartungswert $E(T)$ abhängen von der Verteilung der Zufallsvariablen X_i. Um diese Abhängigkeit im folgenden kenntlich zu machen, schreiben wir $E_\theta(T)$ für $E(T)$, wenn die zu θ gehörende Verteilung die tatsächliche Verteilung der X_i ist.

Ein zu schätzender Parameter aus der Menge der zugrundeliegenden Verteilungen kann nun realisiert werden als Abbildung

$$\tau : \Theta \to \mathbb{R}.$$

Eigenschaften von Schätzern

Erwartungstreue

Ein Schätzer $T = g(X_1, \ldots, X_n)$ heißt **erwartungstreu** für den Parameter τ, falls

$$E_\theta(T) = E_\theta\left(g(X_1, \ldots, X_n)\right) = \tau(\theta)$$

für jedes $\theta \in \Theta$.

Mit anderen Worten: Bestimmt man den Erwartungswert von T unter der Voraussetzung, dass der Parameter θ zugrundeliegt, ergibt sich $\tau(\theta)$ als Erwartungswert.

Beispiele

(i) Das Stichprobenmittel $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist ein erwartungstreuer Schätzer für das Mittel $m = E_\theta(X)$, denn

$$E_\theta(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E_\theta(X_i) = m\,.$$

(ii) Die mittlere quadratische Abweichung

$$T = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

ist kein erwartungstreuer Schätzer für die Varianz $\sigma^2 = E_\theta\left((X - E_\theta(X))^2\right)$, denn

$$\begin{aligned} E_\theta(T) &= \frac{1}{n}\sum_{i=1}^{n} E_\theta(X_i^2) - 2E_\theta(X_i\overline{X}) + E_\theta(\overline{X}^2) \\ &= E_\theta(X^2) - E_\theta\left(\overline{X}^2\right) = \frac{n-1}{n}\sigma^2\,, \end{aligned}$$

denn

$$E_\theta\left(\overline{X}^2\right) = \frac{1}{n^2}\sum_{i,j=1}^{n} E_\theta(X_iX_j) = \frac{n-1}{n}E_\theta(X)^2 + \frac{1}{n}E_\theta(X^2)\,.$$

Im Gegensatz hierzu ist die Stichprobenvarianz $S^2 = \frac{n}{n-1}T$ ein erwartungstreuer Schätzer für σ^2, denn

$$E_\theta\left(S^2\right) = \frac{n}{n-1}E_\theta(T) = \sigma^2\,.$$

Als Abschwächung der Erwartungstreue betrachtet man **asymptotische Erwartungstreue** bei wachsender Stichprobenlänge. Dazu nimmt man an, dass zu jeder Stichprobenlänge n ein Schätzer $T_n = g_n(X_1, \ldots, X_n)$ für $\tau(\theta)$ gegeben ist. Die Folge $T_1, T_2 \ldots$ heißt **asymptotisch erwartungstreu** (für τ), falls

$$\lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

für jedes $\theta \in \Theta$.

Beispiel Die mittlere quadratische Abweichung

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

ist asymptotisch erwartungstreu für die Varianz, denn

$$E_\theta(T_n) = \frac{n-1}{n}\,\mathsf{Var}_\theta(X) \to_{n\to\infty} \mathsf{Var}_\theta(X)\,.$$

Für einen nicht erwartungstreuen Schätzer T bezeichnet man die Abweichung

$$\mathsf{Bias}_\theta(T) := E_\theta(T) - \tau(\theta)$$

als **Bias** (oder **Verzerrung**).

Der **mittlere quadratische Fehler**

$$MSE(T) := E_\theta\left((T-\tau(\theta))^2\right)$$

ist ein Maß für die Schätzgüte. MSE steht dabei für **mean squared error**.

Struktur des mittleren quadratischen Fehlers

$$MSE(T) = E_\theta((T-\tau(\theta))^2) = \mathsf{Var}_\theta(T) + \mathsf{Bias}\,(T)^2.$$

Beweis

$$\begin{aligned} E_\theta((T-\tau(\theta))^2) &= E_\theta(T^2) - 2E_\theta(T)\tau(\theta) + \tau(\theta)^2 \\ &= E_\theta(T^2) - (E_\theta(T))^2 + (E_\theta(T))^2 - 2E_\theta(T)\tau(\theta) + \tau(\theta)^2 \\ &= \mathsf{Var}_\theta(T) + (E_\theta(T) - \tau(\theta))^2\,. \end{aligned}$$

Es sei $T_1, T_2, \ldots$ wieder eine Folge von Schätzern für $\tau(\theta)$. Dann heißt diese Folge

- **konsistent im quadratischen Mittel**, falls

$$\lim_{n\to\infty} E_\theta\left((T_n-\tau(\theta))^2\right) = 0$$

- **schwach konsistent**, falls

$$\lim_{n\to\infty} P_\theta\left(|T_n-\tau(\theta)| \geq \varepsilon\right) = 0 \qquad \forall\, \varepsilon > 0\,.$$

Aufgrund der Ungleichung von Tschebychev ist klar, dass aus Konsistenz im quadratischen Mittel immer schwache Konsistenz folgt, denn

$$P_\theta\left(|T_n-\tau(\theta)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} E_\theta\left((T_n-\tau(\theta))^2\right) \to 0 \qquad \text{für } n\to\infty\,.$$

Beispiel Das Stichprobenmittel $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ist konsistent im quadratischen Mittel für das Mittel $m = E_\theta(X)$ (und damit auch schwach konsistent), denn

$$E_\theta((\overline{X}-m)^2) = \frac{1}{n}\,\mathsf{Var}_\theta(X) \to 0 \text{ für } n\to\infty\,.$$

Effizienz von Schätzern

Der mittlere quadratische Fehler eines Schätzers liefert ein Vergleichskriterium zwischen den verschiedenen Schätzern für τ. Offensichtlich ist von zwei Schätzern T_1, T_2 mit

$$MSE\,(T_1) \leq MSE(T_2)$$

der Schätzer T_1 mit dem kleineren mittleren quadratischen Fehler **wirksamer** für die Schätzung von $\tau(\theta)$.

Beschränkt man sich beim Vergleich zweier Schätzer auf erwartungstreue Schätzer, also Schätzer mit

$$\text{Bias}\,(T_i) = 0 \text{ und damit } MSE(T_i) = \text{Var}(T_i)\,,$$

so reduziert sich der Vergleich der mittleren quadratischen Fehler auf den Vergleich der Varianzen.

Sind T_1, T_2 zwei erwartungstreue Schätzer für $\tau(\theta)$, so heißt T_1 **effizienter** (bzw. **wirksamer**), falls

$$\text{Var}(T_1) \leq \text{Var}(T_2).$$

Bemerkung (Cramér-Rao Schranke) Die Varianz eines erwartungstreuen Schätzers kann nicht beliebig klein werden, sondern wird nach unten beschränkt durch die "Cramér-Rao Schranke". Wir wollen diese Schranke hier nicht angeben, sondern nur bemerken, dass sie von der Variation der Verteilungen in Abhängigkeit von θ abhängt. Ein erwartungstreuer Schätzer, dessen Varianz diese untere Schranke annimmt, heißt **effizient** (oder **wirksamst**).

Beispiele für effiziente Schätzer

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ für den Erwartungswert, wenn man

- alle Verteilungen mit endlicher Varianz zulässt
- alle Normalverteilungen zulässt.

Prinzipien zur Konstruktion von Schätzern

(a) Maximum Likelihood Schätzer

Es seien $X_1, \ldots, X_n$ zunächst diskret verteilt und

$$f(x_1, \ldots, x_n \mid \theta) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$$

die Wahrscheinlichkeitsfunktion zur gemeinsamen Verteilung der Stichprobenvariablen bei zugrundeliegender Verteilung zum Parameter θ. Zu gegebener Stichprobe $x_1, \ldots, x_n$ heißt die Funktion

$$L \,:\, \theta \longmapsto f(x_1, \ldots, x_n \mid \theta)\,,\ \theta \in \Theta,$$

die **Likelihoodfunktion**, denn sie gibt an, wie wahrscheinlich die gewonnene Stichprobe $x_1, \ldots, x_n$ bei angenommener zugrundeliegender Verteilung zum Parameter θ ist.

Die Grundidee der **Maximum Likelihood Schätzung** besteht darin, als Schätzer für θ gerade denjenigen Parameter $\widehat{\theta}$ zu wählen, für den die gewonnene Stichprobe am **wahrscheinlichsten** ist, also $\widehat{\theta}$ mit

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

Der Einfachheit halber betrachten wir im folgenden nur unabhängig und identisch verteilte Stichprobenvariablen. Dann bekommt die Likelihoodfunktion die Produktgestalt

$$L(\theta) = f(x_1, \ldots, x_n \,|\, \theta) = f(x_1 \,|\, \theta) \cdot \ldots \cdot f(x_n \,|\, \theta) \tag{3.35}$$

mit $f(x \,|\, \theta) = P_\theta(X = x)$.

Sind die Stichprobenvariablen zu $\theta \in \Theta$ stetig verteilt mit Dichte $f(x \mid \theta)$, so ersetzt man in der Likelihoodfunktion (3.35) die Wahrscheinlichkeitsfunktion der Verteilung durch die entsprechende Dichte.

Bemerkung Über Existenz (und Eindeutigkeit) des Maximums der Likelihoodfunktion wird hier keine Aussage gemacht! Insbesondere muss i.a. der Maximum-Likelihood Schätzer nicht existieren, oder er muss nicht eindeutig bestimmt sein.

Die Bestimmung der Maximum-Likelihood Schätzung erfolgt in der Regel durch Nullsetzen der Ableitung der Likelihoodfunktion L. Wegen der Produktgestalt von L in (3.35) ist es zweckmäßig, L zunächst zu logarithmieren:

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta) \tag{3.36}$$

und dann zu maximieren. $\ln L$ heißt **Log-Likelihood Funktion**.

Beispiele

(a) **Bernoulli-Experiment**

$$S_n = X_1 + \ldots + X_n$$

sei die Anzahl der Erfolge in einem Bernoulli-Experiment der Länge n bei unbekanntem Erfolgsparameter p. Die Likelihoodfunktion hat die Form

$$L(p) = \binom{n}{S_n} p^{S_n}(1-p)^{n-S_n}, p \in [0,1],$$

wobei S_n die beobachtete Anzahl der Erfolge ist. In diesem Falle ist $\widehat{p} = \frac{S_n}{n}$ das eindeutig bestimmte Maximum, also

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

die (eindeutig bestimmte) Maximum-Likelihood Schätzung für p.

Insbesondere Die Maximum-Likelihood Schätzung $\widehat{p} = \frac{S_n}{n} = \frac{1}{n}(X_1 + \ldots + X_n)$ ist gerade das Stichprobenmittel!

(b) **Normalverteilung**

$X_1, \ldots, X_n$ unabhängig $N(m, \sigma^2)$-verteilt, also $\theta = (m, \sigma)$, und die zugehörige Likelihoodfunktion hat die Gestalt.

$$L(m,\sigma) = \Pi_{i=1}^{n} f_{m,\sigma^2}(x_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2}\right)$$

Logarithmieren ergibt

$$\ln L(m,\sigma) = -n\ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2}$$

mit partiellen Ableitungen

$$\frac{\partial \ln L}{\partial m}(m,\sigma) = \sum_{i=1}^{n} \frac{x_i - m}{\sigma^2}$$
$$\frac{\partial \ln L}{\partial \sigma}(m,\sigma) = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - m)^2}{\sigma^3}\,.$$

Nullsetzen der partiellen Ableitungen liefert die Maximum-Likelihood Schätzung

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{und} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{m})^2\,. \tag{3.37}$$

Hier hat man schließlich noch zu überprüfen, dass (3.37) tatsächlich (eindeutig bestimmtes) Maximum der Likelihoodfunktion ist.

Insbesondere Die Maximum-Likelihood Schätzung $\widehat{m}$ für m entspricht dem Stichprobenmittel, diejenige für σ^2 der mittleren quadratischen Abweichung.

(b) Kleinste Quadrate Schätzung

Ein weiteres Prinzip der Parameterschätzung besteht in der Minimierung der Summe der quadratischen Abweichungen zwischen Beobachtungswert und geschätztem Wert. Dies haben wir bereits bei der Regression kennengelernt.

Beispiel Arithmetisches Mittel

$$\min_{m\in\mathbb{R}} \sum_{i=1}^{n}(x_i - m)^2$$

führt wieder auf das Stichprobenmittel

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Intervallschätzungen

Die bisher konstruierten Schätzer liefern zu gegebenen Beobachtungen $x_1, \ldots, x_n$ eine Schätzung $g(x_1, \ldots, x_n)$ für den unbekannten Parameter $\tau(\theta)$. Daher spricht man auch von Punktschätzungen. In den seltensten Fällen wird die Schätzung exakt mit $\tau(\theta)$ übereinstimmen, sondern bestenfalls "in der Nähe" liegen.

Daher ist es zweckmäßiger, zu gegebener Beobachtung ein Intervall

$$I(x_1, \ldots, x_n) = [U(x_1, \ldots, x_n),\, O(x_1, \ldots, x_n)]$$

anzugeben, in dem der wahre Parameter $\tau(\theta)$ mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ liegt, also:

$$P_\theta\left(\tau(\theta) \in [U(X_1, \ldots, X_n),\, O(X_1, \ldots, X_n)]\right) \geq 1-\alpha \text{ für alle } \theta \in \Theta\,.$$

$1-\alpha$ heißt **Konfidenzwahrscheinlichkeit**, das Intervall $I(X_1, \ldots, X_n)$ **Konfidenzintervall** für $\tau(\theta)$ (zur Konfidenzwahrscheinlichkeit $1-\alpha$).

Konfidenzintervalle für unabhängige normalverteilte Stichprobenvariablen

Es seien $(X_1, \ldots, X_n)$ unabhängig $N(m, \sigma^2)$-verteilt.

(i) **Konfidenzintervall für m bei bekannter Varianz $\sigma^2 = \sigma_0^2$**

$$\Theta = \{(m, \sigma_0) : m \in \mathbb{R}\},\ \tau(m, \sigma_0) = m\,.$$

Eine Punktschätzung für τ ist das Stichprobenmittel

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

Die zugehörige Schätzfunktion

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

ist bei zugrundeliegendem Parameter $\theta = (m, \sigma_0)$ $N(m, \frac{\sigma_0^2}{n})$-verteilt und damit ist die zugehörige Gauß-Statistik

$$\overline{Y} = \sqrt{n}\frac{\overline{X} - m}{\sigma_0} \qquad N(0,1) - \text{verteilt.} \tag{3.38}$$

Zu gegebener Konfidenzwahrscheinlichkeit $1 - \alpha$ ist also

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \overline{Y} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha\,,$$

wobei z_p das p-Quantil der Standard-Normalverteilung bezeichnet, denn

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \overline{Y} \leq z_{1-\frac{\alpha}{2}}\right) = \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi(-z_{1-\frac{\alpha}{2}}) = 2\underbrace{\Phi\left(z_{1-\frac{\alpha}{2}}\right)}_{=1-\frac{\alpha}{2}} - 1 = 1 - \alpha\,.$$

Das zugehörige Konfidenzintervall hat also die Form

$$I(X_1, \ldots, X_n) = \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\,,\ \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right]\,.$$

(ii) **Konfidenzintervall für m bei unbekannter Varianz σ^2**

$$\Theta = \{(m, \sigma) : m \in \mathbb{R}, \sigma > 0\}\,, \tau(m, \sigma) = m\,.$$

Bei unbekannter Varianz σ^2 muss diese erst anhand der Stichprobe $x_1, \ldots, x_n$ geschätzt werden. Dafür bietet sich die Stichprobenvarianz an:

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

mit zugehöriger Schätzfunktion

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

Einsetzen in (3.38) liefert als Schätzfunktion

$$\sqrt{n}\frac{\overline{X} - m}{S}$$

und diese ist t_{n-1}-verteilt. Zu gegebener Konfidenzwahrscheinlichkeit $1-\alpha$ ist also

$$P\left(-t_{n-1,1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\overline{X} - m}{S} \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha\,,$$

wobei $t_{n-1,p}$ das p-Quantil der t-Verteilung mit $n-1$-Freiheitsgraden bezeichnet. Das zugehörige Konfidenzintervall hat somit die Form

$$I(X_1,\ldots,X_n) = \left[\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\,,\ \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]\,.$$

(iii) **Konfidenzintervall für σ^2**

$$\Theta = \{(m,\sigma) : \sigma > 0\}\,,\ \tau(m,\sigma) = \sigma^2.$$

Eine Punktschätzung für die Varianz σ^2 ist die Stichprobenvarianz

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

mit zugehöriger Schätzfunktion

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

Da $X_1,\ldots,X_n$ unabhängig $N(m,\sigma^2)$-verteilt, also $Y_i = \frac{X_i - m}{\sigma}$ unabhängig $N(0,1)$-verteilt, folgt, dass

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

χ^2_{n-1}-verteilt ist. Zu gegebener Konfidenzwahrscheinlichkeit $1-\alpha$ ist also

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha\,,$$

wobei $\chi^2_{n-1,p}$ das p-Quantil der χ^2_{n-1}-Verteilung bezeichnet, denn

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = F_{\chi^2_{n-1}}\left(\chi^2_{n-1,1-\frac{\alpha}{2}}\right) - F_{\chi^2_{n-1}}\left(\chi^2_{n-1,\frac{\alpha}{2}}\right) = 1-\alpha\,.$$

Es ergibt sich als Konfidenzintervall zur Konfidenzwahrscheinlichkeit $1-\alpha$

$$I(X_1,\ldots,X_n) = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\,,\,\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right].$$

Einschub (zur χ^2-Verteilung)

Bemerkung Ihre Bedeutung erhält die χ^2-Verteilung in der induktiven Statistik durch folgende Beobachtung: Sind $X_1,\ldots,X_n$ unabhängig $N(0,1)$-verteilt, so gilt für die Stichprobenvarianz

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

(mit $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$)), dass $(n-1)S^2$ χ^2_{n-1}-verteilt ist.

Bemerkung Ist die Normalverteilungsannahme an die Stichprobenvariablen $X_1,\ldots,X_n$ nicht gerechtfertigt, so kann man unter Ausnutzung des zentralen Grenzwertsatzes eine Normalapproximation für die standardisierte Summe $\sqrt{n}\frac{\overline{X}-m}{\sigma}$ betrachten (siehe Bemerkung zu Konfidenzintervallen in Abschnitt 2.4).

Zum Abschluss noch der wichtige Spezialfall von unabhängig Bernoulli-verteilten Stichprobenvariablen

$$X_1,\ldots,X_n\,.$$

In diesem Falle ist die Summe

$$S_n = X_1+\ldots+X_n$$

Bin (n,p)-verteilt, bei unbekannter Erfolgswahrscheinlichkeit p. Nach dem zentralen Grenzwertsatz ist

$$S_n^* := \frac{S_n-np}{\sqrt{np(1-p)}} = \sqrt{n}\frac{\overline{X}-p}{\sqrt{p(1-p)}}$$

näherungsweise $N(0,1)$-verteilt, also

$$P\left(-z_{1-\frac{\alpha}{2}} \le S_n^* \le z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha\,.$$

Auflösen der Ungleichungen

$$-z_{1-\frac{\alpha}{2}} \le \frac{S_n-np}{\sqrt{np(1-p)}} \le z_{1-\frac{\alpha}{2}}$$

nach p liefert

$$\overline{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \le p \le \overline{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\,,$$

also ist

$$I(X_1,\ldots,X_n) = \left[\overline{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\,,\,\overline{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right]$$

ein (approximatives) Konfidenzintervall für p zur Konfidenzwahrscheinlichkeit $1 - \alpha$.

Beispiel zur Illustration In einem Warenposten aus DVD-Scheiben soll der Anteil der defekten Scheiben geschätzt werden. Dazu wird eine Stichprobe von 200 DVD-Scheiben überprüft. Angenommen, es werden dabei 6 defekte Scheiben gefunden, so ergibt sich für die Ausschusswahrscheinlichkeit bei Konfidenzwahrscheinlichkeit 0.95, also $\alpha = 0.05$, das approximative Konfidenzintervall

$$[0.0063, 0.0537]\,.$$

Mit einer Wahrscheinlichkeit von 95% liegt also der tatsächliche Anteil der defekten DVD-Scheiben im getesteten Warenposten zwischen 0.6 Prozent und 5.37 Prozent.

2. Testen

Ein zentrales Problem der Statistik ist die Frage, wie eine Vermutung über eine Eigenschaft der Verteilung einer Grundgesamtheit anhand einer Stichprobe überprüft werden kann.
Eine solche Vermutung bezeichnet man als **Nullhypothese** H_0. Ein **statistischer Test** ist dann zunächst einmal eine **Entscheidungsregel**

$$\varphi(x_1, \ldots, x_n) \in \{0, 1\}$$

die als Funktion der n Beobachtungen $x_1, \ldots, x_n$ die Nullhypothese H_0 annimmt ($\varphi(x_1, \ldots, x_n) = 0$) oder verwirft ($\varphi(x_1, \ldots, x_n) = 1$).
Demnach ist ein Test durch seinen **Verwerfungsbereich** (oder auch **kritischer Bereich**), also durch die Menge

$$K = \{(x_1, \ldots, x_n) : \varphi(x_1, \ldots, x_n) = 1\}$$

eindeutig bestimmt.

Beispiel Wir betrachten wieder das Beispiel der Warenposten aus DVD-Scheiben. Als Vermutung über den Anteil der defekten DVD-Scheiben soll die Nullhypothese

$$H_0\text{: Anteil der defekten DVD-Scheiben beträgt } 10\%$$

mit Hilfe eines statistischen Tests anhand einer Stichprobe von $n = 100$ DVD-Scheiben überprüft werden.
In diesem Fall wird man den Verwerfungsbereich mit Hilfe einer kritischen Schranke c definieren, ab der man sagt: Ist die beobachtete Anzahl $S_{100} = \sum_{i=1}^{100} X_i > c$, so wird die Nullhypothese verworfen.
Es kann nun allerdings vorkommen, dass die Hypothese in Wahrheit zutrifft, aber aufgrund der getroffenen Entscheidungsregel verworfen wird, da die beobachtete Anzahl s_n der defekten DVD-Scheiben die kritische Schranke übersteigt **(Fehler 1. Art)**. Die Wahrscheinlichkeit für eine solche fälschliche Ablehnung von H_0 soll möglichst klein sein. Dazu gibt man sich ein **Niveau** α vor (etwa $\alpha = 0.05$) und bestimmt die kritische Schranke c so, dass die Wahrscheinlichkeit für eine fälschliche Ablehnung der Hypothese maximal α ist.
Jetzt könnte man natürlich c so wählen, dass die Wahrscheinlichkeit eines Fehlers 1. Art Null ist (einfach: Hypothese immer annehmen!). Dann wird der statistische Test aber sinnlos, da nicht mehr zwischen "guter" und "schlechter" Warenprobe unterschieden wird. Deshalb wählt man c minimal, um damit die Wahrscheinlichkeit dafür, die Nullhypothese zu verwerfen, wenn sie tatsächlich nicht zutrifft, zu maximieren. Diese Wahrscheinlichkeit nennt man die **Macht** des statistischen Tests. Das Komplementärereignis hierzu, d.h. die Nullhypothese zu akzeptieren, obwohl sie in Wahrheit nicht zutrifft, heißt **Fehler 2. Art**.

Die möglichen Ausgänge eines statistischen Tests im Überblick:

	Entscheidung	
	für H_0	gegen H_0
H_0 wahr	richtig	falsch Fehler 1.Art
H_0 falsch	falsch Fehler 2.Art	richtig

- **Niveau** = Wahrscheinlichkeit für einen Fehler 1. Art

- **Macht** = Komplementärwahrscheinlichkeit für einen Fehler 2. Art

Ein **Signifikanztest** zum **Signifikanzniveau** α, $0 < \alpha < 1$, ist ein statistischer Test zum Niveau α, d.h. ein Test mit

$$P\,(\text{ Fehler 1. Art }) \leq \alpha\,.$$

Im Beispiel geht man also wie folgt vor: Zu α wähle c minimal mit

$$P_{0.1}\,(S_{100} > c) \;\leq\; \alpha\,.$$

Hierbei deutet der Index 0.1 an, dass S_{100} unter P Bin $(100, 0.1)$-verteilt ist. Normalapproximation für

$$S^*_{100} = \frac{S_{100} - 100 \cdot 0.1}{\sqrt{100 \cdot 0.1(1-0.1)}} = \frac{S_{100} - 10}{3}$$

ergibt

$$P_{0.1}\,(S_{100} > c) = P_{0.1}\left(S^*_{100} > \frac{c-10}{3}\right) \approx 1 - \Phi\left(\frac{c-10}{3}\right)\,.$$

Also ist c minimal zu wählen mit

$$\Phi\left(\frac{c-10}{3}\right) = 1 - \alpha \text{ und das liefert } c = 3z_{1-\alpha} + 10\,.$$

Allgemein: Approximativer Binomialtest ("Gut - Schlecht" Prüfung)

Gegeben sei die Summe

$$S_n = X_1 + \ldots + X_n$$

von n unabhängig Bernoulli-verteilten Zufallsvariablen X_i mit unbekanntem Parameter p und die Nullhypothese

$$H_0 \,:\, p = p_0$$

zu fest gewähltem Parameter $p_0 \in [0, 1]$.

Zu gegebenem Niveau α bestimme man dann die kritische Schranke

$$c = \sqrt{np_0(1-p_0)}z_{1-\alpha} + np_0\,.$$

Dann ist die Hypothese zu verwerfen, falls die Stichprobensumme $s_n = \sum_{i=1}^n x_i$ größer als c ist.

Bemerkung (zweiseitiger approximativer Binomialtest) In Wahrheit haben wir bei obigem Test nur getestet, ob der Anteil der defekten DVD-Scheiben gleich 10% ist, wenn der unbekannte Parameter p aus der Menge $[0.1, 1]$ stammt. Ist allerdings auch $p < 0.1$ möglich, so setzt sich der Verwerfungsbereich aus einer unteren kritischen Schranke c_u und einer oberen kritischen Schranke c_o zusammen:

$$K = \{(x_1, \ldots, x_n) \,:\, s_u < c_u\} \cup \{(x_1, \ldots, x_n) \,:\, s_n > c_o\}$$

und man spricht von einem **zweiseitigen Ablehnungsbereich**.

Zweckmäßigerweise wählt man dann zu gegebenem Niveau α

$$c_u = -\sqrt{np_0(1-p_0)}z_{1-\frac{\alpha}{2}} + np_0$$
$$c_o = \sqrt{np_0(1-p_0)}z_{1-\frac{\alpha}{2}} + np_0\,,$$

d.h., die Nullhypothese wird verworfen, wenn die Stichprobensumme kleiner c_u oder größer c_o ist, oder in Größen der standardisierten Summe

$$S_n^* = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}}\,,$$

falls $|S_n^*| > z_{1-\frac{\alpha}{2}}$.

Der **approximative Binomialtest** im Überblick:

Test auf den Parameter p einer Binomialverteilung

Annahme $X_1, \ldots, X_n$ unabhängig Bernoulli-verteilt, also $S_n = X_1 + \ldots + X_n$ binomialverteilt.

Hypothese

(a) $H_0 : p = p_0$ (b) $H_0 : p \leq p_0$ (c) $H_0 : p \geq p_0$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} = \sqrt{n}\frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}} \qquad \text{(approx. N(0,1)-verteilt, falls } p = p_0)$$

Hierbei ist $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ das Stichprobenmittel.
Ablehnung, falls

(a) $|T| > z_{1-\frac{\alpha}{2}}$ (b) $T > z_{1-\alpha}$ (c) $T < -z_{1-\alpha}$.

Will man die Annahme an die Verteilung der Stichprobenvariablen fallenlassen, muss man sich im allgemeinen auf das Testen einiger weniger Kennzahlen beschränken.

Gauß-Test

Test auf das Mittel m einer Verteilung bei bekannter Varianz

Annahme $X_1, \ldots, X_n$ unabhängig, identisch verteilt mit bekannter Varianz $\text{Var}(X_i) = \sigma_0^2$ und $X_i \sim N(m, \sigma_0^2)$ oder bei $n \geq 30$ X_i beliebig verteilt mit $E(X_i) = m$

Hypothese

(a) $H_0 : m = m_0$ (b) $H_0 : m \leq m_0$ (c) $H_0 : m \geq m_0$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X} - m_0}{\sigma_0} \qquad \text{(approx. N(0,1)-verteilt, falls } m = m_0)$$

wobei

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

das Stichprobenmittel bezeichnet.

Ablehnung, falls

(a) $|T| > z_{1-\frac{\alpha}{2}}$ (b) $T > z_{1-\alpha}$ (c) $T < -z_{1-\alpha}$.

Der Rest dieses Abschnittes dient der Übersicht einiger wichtiger Testprobleme. Dabei wird danach unterschieden, ob es sich um Ein-Stichproben oder Mehr-Stichproben Tests handelt.

Ein-Stichproben Tests

t-Test

Test auf das Mittel m einer Verteilung mit σ^2 unbekannt

Annahme $X_1, \ldots, X_n$ unabhängig, identisch verteilt mit
$X_i \sim N(m, \sigma^2)$ bzw. bei $n \geq 30$ beliebig verteilt mit $E(X_i) = m$ und Var$(X_i) = \sigma^2$

Hypothese
(a) $H_0 : m = m_0$ (b) $H_0 : m \leq m_0$ (c) $H_0 : m \geq m_0$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X} - m_0}{S} \quad (\text{ approx. } t_{n-1} - \text{verteilt, falls } m = m_0)\,.$$

Hierbei ist

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ die Stichprobenvarianz.}$$

Ablehnung, falls

(a) $|T| > t_{n-1,1-\frac{\alpha}{2}}$ (b) $T > t_{n-1,1-\alpha}$ (c) $T < -t_{n-1,1-\alpha}$.

Für $n \geq 30$ kann man die Quantile der t-Verteilung durch die entsprechenden Quantile der Standardnormalverteilung ersetzen.

χ^2-Test für die Varianz

Annahme $X_1, \ldots, X_n$ unabhängig $N(m, \sigma^2)$-verteilt, m unbekannt

Hypothese

(a) $H_0 : \sigma^2 = \sigma_0^2$ (b) $H_0 : \sigma^2 \leq \sigma_0^2$ (c) $H_0 : \sigma^2 \geq \sigma_0^2$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \frac{n-1}{\sigma_0^2}S^2 \quad (\,\chi^2_{n-1} - \text{verteilt, falls } \sigma^2 = \sigma_0^2).$$

Ablehnung, falls

(a) $T < \chi^2_{n-1,\frac{\alpha}{2}}$ oder $T > \chi^2_{n-1,1-\frac{\alpha}{2}}$ (b) $T > \chi^2_{n-1,1-\alpha}$ (c) $T < \chi^2_{n-1,\alpha}$.

Mehr-Stichproben Tests

Bei mehr-Stichproben Tests sollen Zusammenhänge mehrerer unabhängiger Stichproben mit möglicherweise verschiedenen Längen

$$\begin{matrix} X_1^{(1)}, \ldots, X_{n_1}^{(1)} \\ X_1^{(2)}, \ldots, X_{n_2}^{(2)} \\ \vdots \quad \vdots \quad \vdots \\ X_1^{(k)}, \ldots, X_{n_k}^{(k)} \end{matrix}$$

getestet werden. Die zentrale Frage in diesem Zusammenhang ist dann die nach der Gleichheit der zugrundeliegenden Verteilungen bzw. nach der Gleichheit gewisser Kennzahlen der zugrundeliegenden Verteilungen.

Zwei-Stichproben Gauß-Test

Test auf Gleichheit der Mittel m_X und m_Y zweier Verteilungen bei bekannten Varianzen

Annahme $X_1, \ldots, X_m$ unabhängig, identisch verteilt
$Y_1, \ldots, Y_n$ unabhängig, identisch verteilt mit
$X_i \sim N(m_X, \sigma_X^2)$, $Y_i \sim N(m_Y, \sigma_Y^2)$-verteilt
oder
X_i, Y_j mit beliebiger (stetiger) Verteilung

$$E(X_i) = m_X\,, \mathsf{Var}(X_i) = \sigma_X^2\,, E(Y_j) = m_Y\,, \mathsf{Var}(Y_j) = \sigma_Y^2 \qquad \text{und } m, n \geq 30\,.$$

In beiden Fällen seien σ_X^2 und σ_Y^2 bekannt.

(a) $H_0 : m_X = m_Y$ (b) $H_0 : m_X \leq m_Y$ (c) $H_0 : m_X \geq m_Y$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \qquad \text{approx. N(0,1)-verteilt, falls } m_X = m_Y\,.$$

Ablehnung, falls
(a) $|T| > z_{1-\frac{\alpha}{2}}$ (b) $T > z_{1-\alpha}$ (c) $T < -z_{1-\alpha}$.

Bei unbekannten Varianzen verwendet man

Zwei-Stichproben t-Test

Test auf Gleichheit der Mittel m_X und m_Y zweier Verteilungen bei unbekannten Varianzen

Annahme $X_1, \ldots, X_m$ unabhängig $N(m_X, \sigma_X^2)$-verteilt
$Y_1, \ldots, Y_n$ unabhängig $N(m_Y, \sigma_Y^2)$-verteilt
$\sigma_X^2 = \sigma_Y^2$ unbekannt.

Hypothesen wie im zwei-Stichproben Gauß-Test

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \sqrt{\frac{mn(m+n-2)}{m+n}} \frac{\bar{X} - \bar{Y}}{\sqrt{(m-1)S_X^2 + (n-1)S_Y^2}}$$

mit

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \text{ und } S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 .$$

(t_{m+n-2}-verteilt, falls $m_X = m_Y$.)

Ablehnung, falls

(a) $|T| > t_{m+n-2,1-\frac{\alpha}{2}}$ (b) $T > t_{m+n-2,1-\alpha}$ (c) $T < -t_{m+n-2,1-\alpha}$.

Für eine Erweiterung des zwei-Stichproben t-Tests auf den Fall ungleicher Varianzen siehe [2].

F-Test

Test auf Gleichheit der Varianzen zweier Normalverteilungen

Annahme $X_1, \ldots, X_m$ unabhängig $N(m_X, \sigma_X^2)$-verteilt
$Y_1, \ldots, Y_n$ unabhängig $N(m_Y, \sigma_Y^2)$-verteilt
Mittel unbekannt

Hypothese
(a) $H_0 : \sigma_X^2 = \sigma_Y^2$ (b) $H_0 : \sigma_X^2 \leq \sigma_Y^2$ (c) $H_0 : \sigma_X^2 \geq \sigma_Y^2$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) = \frac{S_X^2}{S_Y^2} \quad (\, F_{m-1,n-1} - \text{verteilt, falls } \sigma_X^2 = \sigma_Y^2)$$

Ablehnung, falls

(a) $T < F_{m-1,n-1,\frac{\alpha}{2}}$ oder $T > F_{m-1,n-1,1-\frac{\alpha}{2}}$
(b) $T > F_{m-1,n-1,1-\alpha}$ (c) $T < F_{m-1,n-1,\alpha}$.

Statt Mittel und Varianzen auf Gleichheit zu überprüfen kann man schließlich auch zwei (oder mehr) Verteilungen auf Gleichheit überprüfen.

χ^2-Homogenitätstest

Annahme

$X_1^{(1)}, \ldots, X_{n_1}^{(1)}$	unabhängig und identisch verteilt mit Verteilungsfunktion F_1
$X_1^{(2)}, \ldots, X_{n_2}^{(2)}$	unabhängig und identisch verteilt mit Verteilungsfunktion F_2
$\vdots$	$\vdots$ $\vdots$
$X_1^{(k)}, \ldots, X_{n_k}^{(k)}$	unabhängig und identisch verteilt mit Verteilungsfunktion F_k

Hypothese
$H_0 : F_1 = F_2 = \ldots = F_k$

Um die Hypothese zu testen, unterteilen wir zunächst die x-Achse in $m \geq 2$ disjunkte Intervalle

$$A_1 =]-\infty, z_1], A_2 =]z_1, z_2], \ldots, A_m =]z_{m-1}, \infty[$$

und bestimmen für jedes Intervall die Häufigkeiten

$$h_{ij} = \#\{X_l^{(i)} \,:\, X_l^{(i)} \in A_j\}\,, i = 1, \ldots, k,\; j = 1, \ldots, m$$

und bilde hierzu die Spaltensummen

$$h_{\cdot j} = h_{1j} + \ldots + h_{kj}\,, j = 1, \ldots, m$$

Begründung Unter der Hypothese sind die Stichprobenvariablen identisch verteilt und damit sollten für alle j die relativen Häufigkeiten

$$\frac{h_{ij}}{n_i} \qquad i = 1, \ldots, k$$

nahezu übereinstimmen, d.h.

$$\frac{h_{ij}}{n_i} \sim \frac{h_{\cdot j}}{n} \qquad \text{oder } h_{ij} - \frac{n_i h_{\cdot j}}{n} \sim 0\,.$$

Hierbei ist $n = n_1 + \ldots + n_k$.

Entscheidungsregel Betrachte als Testgröße

$$T(X_1^{(1)}, \ldots, X_{n_k}^{(k)}) = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{n_i h_{\cdot j}}{n}\right)^2}{\frac{n_i h_{\cdot j}}{n}}\,.$$

Ablehnung, falls

$T > \chi^2_{(k-1)(m-1),1-\alpha}$.

χ^2-Anpassungstest

Häufig ist man daran interessiert, ob die unbekannte Verteilung einer Grundgesamtheit gleich einer gegebenen hypothetischen Verteilung ist.
Dazu stellen wir uns vor, dass die Stichprobenvariablen $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind mit einer Verteilungsfunktion F und wir stellen zu gegebener Verteilungsfunktion F_0 die

Hypothese $H_0 \,:\, F = F_0$

auf.

Im nächsten Schritt unterteilen wir die x-Achse in $k \geqslant 2$ disjunkte Intervalle

$$A_1 =]-\infty, z_1],\; A_2 =]z_1, z_2], \ldots, A_k =]z_{k-1}, \infty[$$

und bestimmen für jedes Intervall A_j

- die Anzahl h_j der in A_j liegenden Stichprobenwerte

$$h_j = \#\{x_i : x_i \in A_j\}$$

- die theoretische Wahrscheinlichkeit p_j, dass eine Stichprobenvariable X mit Verteilungsfunktion F_0 einen Wert in A_j annimmt

$$p_j = P(X \in A_j) = F_0(z_j) - F_0(z_{j-1})\,.$$

  Hierbei setzt man $F_0(z_0) = 0$ und $F_0(z_k) = 1$.

Hinweis: Ist der Wertebereich der Stichprobenvariablen endlich, etwa $\{a_1, \ldots, a_k\}$, so kann man auf die Klassifizierung verzichten und h_j, p_j definieren durch

$$h_j = \#\{x_i : x_i = a_j\} \qquad p_j = P(X = a_j)\,.$$

Entscheidungsregel Betrachte die Testgröße

$$T(X_1, \ldots, X_n) = \sum_{j=1}^{k} \frac{(h_j - np_j)^2}{np_j} \quad \left(= \frac{1}{n} \sum_{j=1}^{n} \frac{h_j^2}{p_j} - n \right).$$

Ablehnung, falls

$$T > \chi^2_{k-1,1-\alpha}\,.$$

Dieser Test hat dann das approximative Niveau α.

Hinweis Die Anzahl der Beobachtungen sollte mindestens so groß sein, dass $np_j \geq 5$ gilt für $j = 1, \ldots, k$.

χ^2-Test auf Unabhängigkeit (Kontingenztest)

Ausgangspunkt des Tests auf Unabhängigkeit ist die Frage, ob zwei Merkmale X und Y in einer gegebenen Grundgesamtheit voneinander unabhängig sind oder nicht. Es ist also ein statistischer Test zu konstruieren, der aufgrund einer zweidimensionalen Stichprobe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

entscheidet, ob die folgende

Hypothese H_0: X und Y sind unabhängig

angenommen werden kann oder nicht.
Wie beim χ^2-Anpassungstest unterteilen wir die x-Achse in $k \geq 2$ disjunkte Intervalle

$$A_1 =]-\infty, z_1],\ A_2 =]z_1, z_2], \ldots, A_k =]z_{k-1}, \infty[$$

und die y-Achse in $l \geq 2$ disjunkte Intervalle

$$B_1 =]-\infty, \tilde{z}_1],\ B_2 =]\tilde{z}_1, \tilde{z}_2], \ldots, B_l =]\tilde{z}_{l-1}, \infty[.$$

Hierzu stellen wir dann die zughörige Kontingenztabelle mit Randhäufigkeiten auf

	y				
x	B_1	B_2	$\ldots$	B_l	
A_1	h_{11}	h_{12}	$\ldots$	h_{1l}	$h_{1\cdot}$
A_2	h_{21}	h_{22}	$\ldots$	h_{2l}	$h_{2\cdot}$
$\vdots$	$\vdots$			$\vdots$	$\vdots$
A_k	h_{k1}	h_{k2}	$\cdots$	h_{kl}	$h_{k\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$\ldots$	$h_{\cdot l}$	

und bilden die Größe

$$\tilde{h}_{ij} := \frac{h_{i\cdot}h_{\cdot j}}{n};\ 1 \le i \le k,\ 1 \le j \le l\,.$$

Begründung Unter der Hypothese sind die Merkmale X und Y unabhängig und damit

$$P(X \in A_i, Y \in B_j) = P(X \in A_i)\,P(Y \in B_j)\,. \tag{3.39}$$

Bei großer Stichprobenlänge n sollte zudem die relative Häufigkeit in der Nähe der theoretischen Wahrscheinlichkeit liegen, also

$$\frac{h_{ij}}{n} \sim P(X \in A_i, Y \in B_j)\,,$$

$$\frac{h_{i\cdot}}{n} \sim P(X \in A_i) \quad \text{und} \quad \frac{h_{\cdot j}}{n} \sim P(Y \in B_j)\,.$$

Ingesamt sollte also gelten

$$\frac{h_{ij}}{n} \sim P(X \in A_i, Y \in A_j) = P(X \in A_i)P(Y \in A_j) \sim \frac{h_{i\cdot}}{n} \cdot \frac{h_{\cdot j}}{n}\,,$$

also $h_{ij} \sim \tilde{h}_{ij}$.

Folglich führen wir nun einen χ^2-Anpassungstest gegen die Produktverteilung $\tilde{h}_{ij}/n$ durch und bilden dementsprechend die Testgröße

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{h_{ij}^2}{\tilde{h}_{ij}} - n\,.$$

Entscheidungsregel

Ablehnung, falls $T > \chi^2_{(k-1)(l-1),1-\alpha}$.

Bemerkung (zur Anzahl der Freiheitsgrade)

Da pro Zeile (bzw. Spalte) eine der Häufigkeiten h_{ij} von den übrigen $l-1$ (bzw $k-1$) Häufigkeiten über die entsprechende Randhäufigkeit $h_{i\cdot}$ (bzw. $h_{\cdot j}$) abhängt, ergibt sich als Anzahl der Freiheitsgrade in der Testgröße

$$kl - l - k + 1 = (k-1)(l-1)\,.$$

Literatur

[1] G. Bamberg, F. Baur, M. Krapp, Statistik, 13. Auflage, R. Oldenbourg Verlag, 2007.

[2] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, Statistik, 6. Auflage, Springer Verlag, 2007.

Weitere Literatur

[3] J. Bleymüller, G. Gehlert, H. Gülicher, Statistik für Wirtschaftswissenschaftler, 14. Auflage, Verlag Vahlen, 2004.

[4] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, A. Caputo, S. Lang, Arbeitsbuch Statistik, 4. Auflage, Springer Verlag, 2004.

[5] J. Schira, Statistische Methoden der VWL und BWL: Theorie und Praxis, 2. Auflage, Pearson Studium, 2005.