

Deskriptive Statistik

Daten werden über eine Stichprobe erhoben und ausgewertet.
Teilgebiete:

- 1. Deskriptive/beschreibende Statistik:** Übersichtliche Darstellung, die Ermittlung von Kenngrößen und Datenvalidierung.
- 2. Explorative Statistik:** Suche nach Strukturen/Besonderheiten.
- 3. Induktive/beurteilende Statistik:** Schlussfolgerungen mittels Wahrscheinlichkeitsrechnung.

Begriffe:

- Merkmalsträger/Statistische Einheiten:** Erfasste Objekte (z.B. Wohnung)
- Grundgesamtheit:** alle statistischen Einheiten (z.B. alle Mietwohnungen)
- Vollerhebung:** Eigenschaft bei jedem Individuum in der Gesamtheit wird abgefragt/gemessen/erhoben
- Stichprobe:** Untersuchte Teilmenge der Grundgesamtheit
- Stichprobengröße:** Anzahl der Einheiten in der Stichprobe.
- Merkmal (Variable):** Interessierte Größen (z. B. «Nettomiete» und «Baualter» für eine Wohnung ermittelt)
- (Merkmals-)Ausprägungen:** Werte, die jedes Merkmal annehmen kann (z.B. CHF 1234.– für Merkmal Nettomiete)
- Urliste:** Unbearbeitete notierten/beobachteten Stichprobenwerte

Merkmalstypen / Messniveaus

- Qualitativ/Kategorisch:** Angabe Ausprägungen aber *kein* Ausmass
 - 1. Nominal:** Kategorisierung ohne Ordnung (z.B. Geschlecht, Farben)
 - 2. Ordinal:** Kategorisierung mit Ordnung/Rangierung (z.B. Rangliste)
- Quantitativ/Metrisch:** Angabe Ausmass in Zahlen
 - 3. Diskret:** Abzählbar viele Ausprägungen (z.B. Anzahl Kinder) → CMF
 - 4. Stetig:** Ausprägungen in reellen Intervall (z.B. Körpergrösse) → CDF

Empirisch = nicht-klassiert (diskret oder stetig)

Diskrete Stichproben (PMF, CMF)

Absolute Häufigkeitsfunktion $h(x)$

$h(x)$ = Anzahl Elemente eines bestimmten Merkmals.

Absolute Summenhäufigkeit $H(x)$

$H(x) = \sum h_i$ = Summe aller Stichprobenwerte

Relative Häufigkeitsdichtefunktion $f(x)$ PMF

PMF (Probability Mass Function)

Wahrscheinlichkeitsverteilung. **Zähldichte** $P(\{\omega\})$

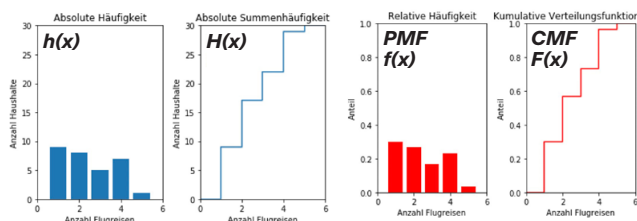
$$f(x) = \frac{h(x)}{n} \quad n = \text{Stichprobenumfang} = H(x) \quad f : \mathbb{R} \rightarrow [0, 1]$$

Kumulative Verteilungsfunktion $F(x)$ CMF

CMF (Cumulative Mass Function)

Auch Kumulative Summenhäufigkeit. Summe aller PMF $f(x)$

$$F(x) = \sum f_i = \frac{H(x)}{n} = \sum PDF_i \cdot b_i$$



Anteil aller Stichprobenwerte mit

- höchstens x : $F(x)$
- mindestens x : $1 - F(x)$
- genau x : $F(x_i) - F(x_{i-1})$
- zwischen x_i und x_j : $F(x_i) - F(x_{j-1})$ wobei $i > j$

Bsp:

$x = [-8, -8, -8, -8, -3, -1, \dots]$ mit $n = 18$

$F(-2) = 4/18 + 1/18 = 5/18$

Klassierte und stetige Stichproben (PDF)

Klassierte Häufigkeitsfunktion $g(x)$

Absolut

$h(x)$ = Anzahl Elemente einer bestimmten Klasse

Relativ

$$g(x) = f_i = \frac{h(x)}{n} \quad g : \mathbb{R} \rightarrow [0, 1]$$

n = Stichprobenumfang = $H(x)$

Klassierte Häufigkeitsdichtefunktion $f(x)$ PDF

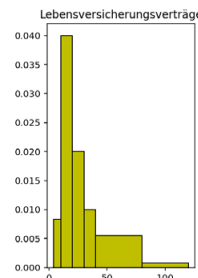
PDF (Probability Density Function) = **Histogramm**

PDF muss nicht zwingend klassiert sein:

PMF (diskret) = PDF (stetig) = h (absolut)

Bei Stichproben stetiger Merkmale teilt man die Stichprobenwerte in **Klassen** bzw. (nicht unbedingt gleich grosse) **Intervalle** ein. **Fläche** des Rechtecks ist ein Mass für die Anzahl der Messwerte.

Obere Intervallgrenzen zählen dabei immer zum darauffolgenden Intervall: $[100, 200[$ oder $[100, 200]$



Absolut

$$h = \frac{h_i}{d_i} \quad \begin{array}{l} h = \text{Höhe} \\ h_i = \text{Absolute Klassenhäufigkeit (= Flächeninhalt)} \\ d_i = \text{Klassenbreite} \end{array}$$

Relativ

$$f(x) = \frac{g(x)}{b_i} = \frac{h(x)}{n} \quad f : \mathbb{R} \rightarrow [0, 1]$$

f = PDF (= Höhe)

g_i = Relative Klassenhäufigkeit (= Flächeninhalt)

h_i = Absolute Klassenhäufigkeit

n = Stichprobengrösse

b_i = Klassenbreite

Faustregel

- Die Klassen sollten gleich breit gewählt werden. (Muss nicht immer gelten)
- Die Anzahl der Klassen sollte $\sqrt{n}$ nicht wesentlich überschreiten.
Die Klassen interessieren, nicht die einzelnen Werte.

Kumulative Verteilungsfunktion $F(x)$ CDF

CDF (Cumulative Distribution Function)

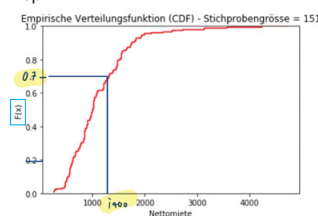
Auch kumulative Summenhäufigkeit. Summe aller $f(x)$ PDF

CMF (diskret, relativ) = CDF (stetig, relativ) = H (diskret, absolut)

$$F(x) = \int_{-\infty}^x f(x) dy \quad F : \mathbb{R} \rightarrow [0, 1]$$

In der letzten Klasse wird der **prozentuale Anteil** der relativen Häufigkeit der Klasse in der sich x befindet dazu gezählt!

Bsp:



Die PDF einer stetigen Zufallsvariablen X ist gegeben durch: $f(x) = \begin{cases} a(2x - x^2) & 0 \leq x \leq 2 \\ 0 & \text{sonst} \end{cases}$
Wie gross muss a sein, damit f eine PDF ist?

$$F(x) = \int_0^x f(x) dx = 1 \Rightarrow a \int_0^2 (2x - x^2) dx = 1$$

$$\Rightarrow a \left(x^2 - \frac{1}{3} x^3 \right) \Big|_0^2 = 1 \Rightarrow a = \frac{3}{4}$$

$$F(x) = \frac{3}{4} x^2 - \frac{1}{4} x^3$$

Quantile R_q

Teil die Stichprobe bzw. der Graph einer **CDF** in zwei Teile ein.

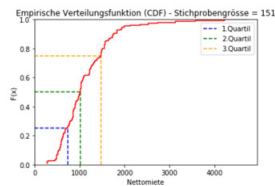
$q = F(x)$ = Wert (0 bis 1) auf Ordinate (y-Achse)

$R_q = x$ = Wert auf Abszisse (x-Achse)

1. Quantil: $R_{0.25}$, 2. Quantil: $R_{0.5}$, 3. Quantil: $R_{0.75}$

q -Quantil $R_{0.25} = q$ -Perzentil R_{25}

$R_{0.5} = Q_1 = \text{Median}$



R_q / Quantile (x-Werte) berechnen

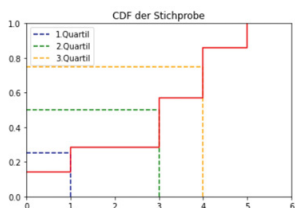
Stichprobenwerte müssen **geordnet** vorliegen.

2 Fälle:

– $n \cdot q$ gerade: $R_q = \frac{1}{2}(x_{n \cdot q} + x_{n \cdot q + 1})$

– $n \cdot q$ ungerade: $R_q = x_{[n \cdot q]}$
 n = Anzahl Werte oder Klassen

Bsp: Geordnete CDF der Stichprobenwerte [0, 1, 3, 3, 4, 4, 5]



Bsp. $q = 0.5$

• $[0, 1, 3, 3, 4, 4, 5]$

$n \cdot q = [3.5] = 4$ $R_{0.5} = x_4 = 3$

• $[-2, 0, 5, 7]$

$n \cdot q = 2$ $R_{0.5} = \frac{1}{2}(x_2 + x_3) = 2.5$

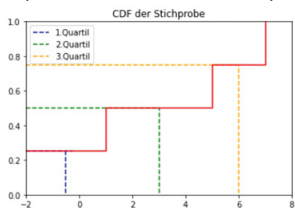
Bedeutung:

$q = 0.25$: $25\% \leq 1$,

$q = 0.5$: $50\% \leq 3$,

$q = 0.75$: $75\% \leq 4$

Bsp: Geordnet CDF der Stichprobenwerte [-2, 1, 5, 7]



Bedeutung:

$q = 0.25$: $25\% \leq -0.5$

$q = 0.5$: $50\% \leq 3$

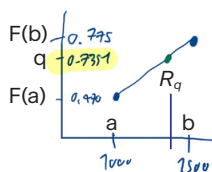
$q = 0.75$: $75\% \leq 6$

Linear interpoliert zwischen zwei q -Quantilen $F(a)$ und $F(b)$

a, b = Werte Abszisse (x-Achse) $F(x)$ = Werte Ordinate (y-Achse)
 = Klassengrenzen

$$R_q = a + \frac{(b - a)}{(F(b) - F(a))} \cdot (q - F(a))$$

$$q = F(a) + \frac{(F(b) - F(a))}{(b - a)} \cdot (R_q - a)$$



Bsp: Klassierte Stichproben

1. Relative Häufigkeiten $f(x)$ berechnen (sofern nicht gegeben)
2. Benötigte CDFs $F(x)$ berechnen (kumulative relative Häufigkeiten)

$q = 0.75$

Klassengrenzen	[8, 11[	[11, 56[	[56, 116[	[116, 165[	[165, 280[
Relative Häufigkeit $f(x)$	3/17	19/85	13/85	27/85	11/85
Summenhäufigkeit (CDF) $F(x)$	3/17 ≈ 0.18	$F_1 + 19/85$ ≈ 0.40	$F_2 + 13/85$ ≈ 0.55	$F_3 + 27/85$ ≈ 0.87	...

$0.55 < 0.75 < 0.87$

→ $a = 116$, $b = 165$, $F(a) = 0.55$, $F(b) = 0.87$

0.75 -Quantil $R_{0.75} = 116 + (165 - 116) / (0.87 - 0.55) \cdot (0.75 - 0.55) \approx 146.63$

Lagemass

Median x_{med}

Zentralwert. Zentrum der Verteilung.

Teilt eine **sortierte** Stichprobe in zwei gleich grosse Hälften ein.

2 Fälle:

– n ist ungerade: $x_{\lceil \frac{n+1}{2} \rceil}$ – n ist gerade: $\frac{1}{2}(x_{\lceil \frac{n}{2} \rceil} + x_{\lceil \frac{n}{2} \rceil + 1})$

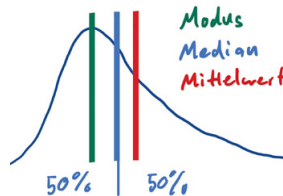
Bsp: $\text{Median}(x_1, \dots, x_7) = x_{[4]}$ $\text{Median}(x_1, \dots, x_6) = \frac{1}{2}(x_{[3]} + x_{[4]})$

Bei **klassierten** Daten muss das **2. Quantil** ($R_{0.5}$) berechnet werden.

Modus (Modalwert) x_{mod}

Klassenmitte. Wo h_i (absolute Häufigkeit) am grössten ist.

Gibt das **Maximum der Verteilung** an, d.h. den **häufigsten Wert**.



Arithmetische Mittel $\bar{x}$

Empirische Mittelwert / Erwartungswert

Ohne Häufigkeitsangaben

Mit Häufigkeitsangaben

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^m a_i \cdot h_i = \sum_{i=1}^m a_i \cdot f_i$$

n = Anzahl Stichproben

m = Anzahl Merkmale / Klassen

a = Wert oder **Klassenmitte**

h = Absolute Häufigkeit

f = Relative Häufigkeit $= h_i / n$

Streuungsmass

Varianz σ^2

Durchschnittliche quadratische Abweichung vom Mittelwert der Verteilung.

Empirische Standardabweichung (ohne Häufigkeitsangaben)

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

Klassierte Standardabweichung (mit Häufigkeitsangaben)

$$= \frac{1}{n} \sum_{i=1}^m h_i (a_i - \bar{x})^2 = \left[\frac{1}{n} \sum_{i=1}^m a_i^2 \cdot h_i \right] - \bar{x}^2 = \left[\sum_{i=1}^m a_i^2 \cdot f_i \right] - \bar{x}^2$$

n = Stichprobenwerte

m = Anzahl Merkmale / Klassen

a = Wert oder **Klassenmitte**

h = Absolute Häufigkeit

f = Relative Häufigkeit $= h_i / n$

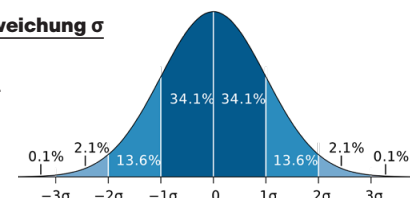
Empirische Standardabweichung σ

(engl. Population).

Wahrscheinlichkeitsdichte.

Normalverteilung

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



Korrigierte Varianz s^2 / Standardabweichung s

(engl. Sample)

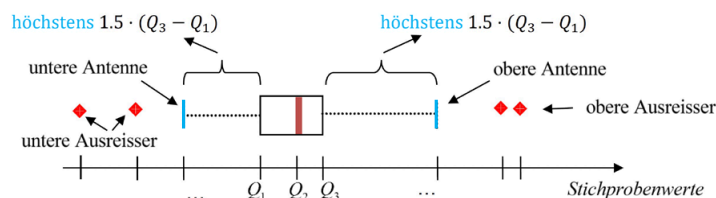
Statt mit n wird mit $n - 1$ gemittelt.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad s^2 = \frac{n}{n-1} \cdot \tilde{s}^2 \quad s = \sqrt{\frac{n}{n-1}} \cdot \tilde{s}$$

Boxplot

Nur für unklassierte Daten sinnvoll.

$Q_1 = R_{0.25}$, $Q_2 = R_{0.5}$, $Q_3 = R_{0.75}$

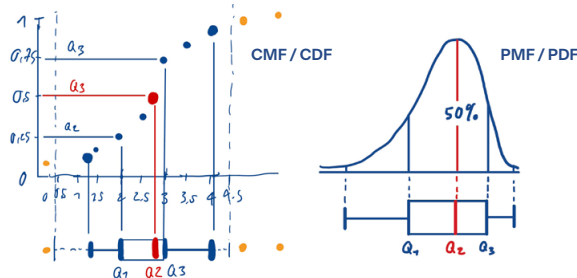


$Q_3 - Q_1 = \text{Interquartilsabstand (IQA)}$

Untere Antenne (Whisker): **Minimaler Wert** innerhalb $Q_1 - 1.5 \cdot (Q_3 - Q_1)$

Obere Antenne (Whisker): **Maximaler Wert** innerhalb $Q_3 + 1.5 \cdot (Q_3 - Q_1)$

Ausreisser: Werte ausserhalb Interquartilsabstand (IQA)



Korrelation

Korrelationskoeffizient (Bravais-Pearson) r_{xy}

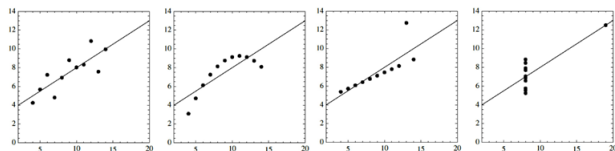
(engl. Correlation coefficient)

Kennwert, der den Grad einer linearen Abhängigkeit bzw. die Stärke (= Streuung) und Richtung eines **linearen Zusammenhangs** zwischen **zwei metrischen Merkmalen** oder **zwei Zeitreihen** misst.

Bivariate Daten: Zwei Merkmale (Wertepaare) werden miteinander verglichen

Multivariate Daten: Mehrere Merkmale werden miteinander verglichen

Der Pearson-Korrelationskoeffizient **ist nicht robust** gegen Ausreisser:



Korrelation bedeutet nicht unbedingt einen kausalen Zusammenhang.

Ist der Zusammenhang **nicht kausal**, spricht man von einem

Scheinkorrelation.

Starke Korrelation wenn $|r_{xy}|$ nahe 1:

Nahe 1: positiv korreliert (= positive Steigung)

Nahe -1: negativ korreliert (= negative Steigung)

Nahe 0: unkorreliert

Weil es ein Verhältnis ist, kommt es nicht darauf an, ob die korrigierten Kovarianz und Standardabweichung verwendet wird oder nicht.

$$r_{xy} = \frac{\tilde{s}_{xy}}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} \quad [-1, 1]$$

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Standardabweichung

$\tilde{s}_x$ und $\tilde{s}_y$

Varianz (empirisch)

$$\tilde{s}_x^2 = \overline{x^2} - \bar{x}^2 \quad \tilde{s}_y^2 = \overline{y^2} - \bar{y}^2$$

Mittel der quadrierten x-/y-Werte.
Mittel der x-/y-Werte quadriert.

Varianz vs. Kovarianz

Die **Varianz** misst, wie weit die einzelnen Werte in einem Datensatz vom **Mittelwert** dieses Datensatzes abweichen.

Die **Kovarianz** misst, wie zwei Variablen gemeinsam variieren oder **zusammenhängen**.

Kovarianz $\tilde{s}_{xy}$ (empirisch)

(engl. Covariance) Misst, in welche Richtung sich y bewegt,

wenn sich x weg vom Mittelwert bewegt → misst absolute Werte $[-\infty, \infty]$

$$\tilde{s}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y}$$

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y))$$

Rangkorrelationskoeffizient (Spearman) r_{sp}

Misst Stärke und Richtung des **monotonen** Zusammenhangs,

d.h. wie die Punkte um eine Kurve streuen. **Robust** gegen Ausreisser.

$$r_{sp} = \frac{\sum_{i=1}^n (\text{rg}(x_i) - \overline{\text{rg}(x)}) (\text{rg}(y_i) - \overline{\text{rg}(y)})}{\sqrt{\sum_{i=1}^n (\text{rg}(x_i) - \overline{\text{rg}(x)})^2} \cdot \sqrt{\sum_{i=1}^n (\text{rg}(y_i) - \overline{\text{rg}(y)})^2}}$$

Alternativ, **nur** wenn **alle Ränge verschieden** sind:

$$r_{sp} = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)} \quad d_i = \text{rg}(x_i) - \text{rg}(y_i)$$

$\text{rg}(x_i)$ = Rang der geordneten x-Werte.

Verbundene Ränge

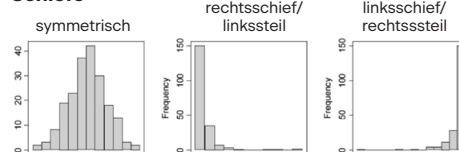
Wenn unterschiedliche Klassen gleiche Werte haben.

i	1	2	3	4	5	6
x_i	23	27	35	35	35	59
$\text{rg}(x_i)$	1	2	4	4	4	6

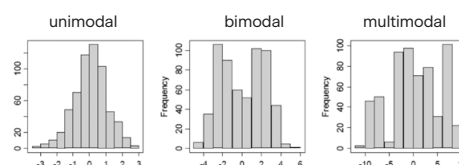
$$4 = 1/3 \cdot (3 + 4 + 5)$$

Verteilungsformen

Schiefe



Modalität (lokale Maxima)



Covariance Matrix

The covariance matrix captures the pairwise covariances between variables.

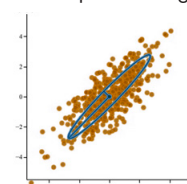
$$\Sigma = \begin{bmatrix} \text{Cov}(x_1, x_1) & \text{Cov}(x_1, x_2) & \cdots \\ \text{Cov}(x_2, x_1) & \text{Cov}(x_2, x_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$\text{Cov}(A, B) = \text{Cov}(B, A)$

$\text{Cov}(A, A) = \text{Var}(A)$ (variance)

The Covariance Matrix is symmetric.

To compute the eigenvectors from the covariance matrix:



$$\Sigma v = \lambda v$$

v is an eigenvector s

λ is the corresponding eigenvalues

These eigenvectors represent principal directions of variance in your data.

Mögliche Anzahl Gruppen: Wie viele verschiedene Gruppen können aus 10 Elementen gebildet werden? → **Erstazproblem**
 Jede Gruppe kann man mit einem Bitmuster identifizieren.
 z.B: 1 0 0 0 0 1 1 1 0
 Abzahl Möglichkeiten: $2^{10} - 1 = 1023$ (= Potenzmenge $\setminus \{\}$)

In wieviele Arten können 5 Paare in 4er Gruppen unterteilt werden, so dass die Gruppen keine Paare enthält → **Erstazproblem**

Paar als Werkpaar 0 und 1 betrachten.

$$\{(0,1), (0,1), (0,1), (0,1), (0,1)\} \quad \binom{5}{4} \cdot 2^4 = 80$$

0 1 1 0 → Bsp. Gruppe

Wie viele 6-stellige positive ganze Zahlen mit mind. einer Ziffer 2 gibt es?

$$9 \cdot 10^5 - 8 \cdot 9^5 = 427'608$$

- Gesamtanzahl der 6-stelligen, positiven ganzen Zahlen:
 Ziffer 1: 1-9 = 9 Ziffern 2 bis 6: 0-9 = 10^5
- Anzahl der 6-stelligen, positiven ganzen Zahlen ohne Ziffer 2:
 Ziffer 1: 1, 3-9 = 8 Ziffern 2 bis 9: 0, 1, 3-9 = 9^5

Möglichkeiten 6 Personen in 3 gleichgrosse Gruppen einzuteilen

$$\frac{1}{3} \binom{6}{2} \binom{4}{2} \binom{2}{2} = 30 \quad \frac{1}{n} \prod_{i=1}^n \binom{k \cdot i}{k} \quad \begin{matrix} n = \text{Anzahl Gruppen} \\ k = \text{Gruppengrösse} \\ (\text{Gesamtanzahl} / n) \end{matrix}$$

20 Karten, 4 davon sind Joker. Jeder Spieler erhält 5 Karten.
 Kombinationen mit **mindestens einem** Joker?

mind. 1 Joker = 1 - kein Joker = **Komb. alle Karten** - **Komb. ohne Joker**

$$20C5 - 16C5 = 11135$$

$\{0,0,0,0,0,0,0,0\}$ 4er-Gruppe mit mind. 2 Blauen.

$$\binom{4}{2} \binom{5}{2} + \binom{4}{3} \binom{5}{1} + \binom{4}{4}$$

Diskrete Wahrscheinlichkeitsräume

Zähldichte $p(\omega) = (PMF)$

Ergebnis-Wahrscheinlichkeit $P(\{\omega\})$

$$p: \Omega \rightarrow [0,1], p(\omega), \omega \in \Omega, \sum_{\omega \in \Omega} p(\omega) = 1$$

Ordnet jedem **Ergebnis** ω seine Wahrscheinlichkeit zu (= relative Häufigkeit f)

Bsp: Ω Ergebnisraum = $\{1,2,3,4,5,6\}$, $p(3) = 1/6$

Sind Alle Ereignisse gleichwahrscheinlich, dann handelt es sich um eine

Gleichverteilung und Ω ist ein **Laplace-Raum** (= gleichförmige Ereignisgrösse)

Ergebnisraum Ω

Menge aller sich gegenseitig ausschliessenden **Ergebnisse** eines Zufallsexperiments. z.B. $\Omega = \{1,2,3\}$

Diskret falls **endlich** oder **abzählbar unendliche** Anzahl Ergebnissen.

Die leere Menge ist ein unmögliches Ereignis: $P(\emptyset) = 0$

Ω trifft immer ein: $P(\Omega) = 1$

Ereignis

Teilmenge von Ω bzw. Zusammenfassung von gewissen Ergebnissen. $A = \{1,3\}$

Ereignisraum 2^Ω

Menge aller möglichen, sich gegenseitig **ausschliessenden Ereignisse**.

= Menge aller Teilmengen von Ω ($2^{|\Omega|} = \text{Potenzmenge von } \Omega$)

z.B. $\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$

Wahrscheinlichkeitsmass / Wahrscheinlichkeitsverteilung P

$$P: 2^\Omega \rightarrow [0,1], P(M) = \sum_{\omega \in M} p(\omega) \quad \text{mit } M \subseteq \Omega.$$

Gibt die Wahrscheinlichkeit eines Ereignisses an.

Bsp: $A = \{2,4,6\}$, $P(A) = P(2) + P(4) + P(6)$

Komplementäres Ereignis $P(\bar{A})$

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

Sigma-Additivität

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

falls die Ereignisse **paarweise disjunkt** sind.

Vereinigung $P(A \text{ oder } B \text{ oder beide})$

$$P(A \cup B) = P(B) + P(A) - P(A \cap B) = \frac{|B| + |A| - |A \cap B|}{|\Omega|}$$

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

Zufallsexperiment

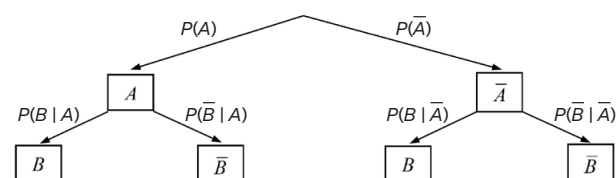
Definition

- Der Vorgang lässt sich unter den gleichen äusseren Bedingungen beliebig oft wiederholen.
- Es sind mehrere sich gegenseitig ausschliessende Ergebnisse möglich, diese sind vor der Durchführung des Zufallsexperiments bekannt.
- Das Ergebnis lässt sich nicht mit Sicherheit voraussagen, sondern ist zufallsbedingt.

Stochastische Abhängigkeit

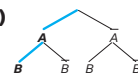
Stochastisch abhängig (Bedingte Wahrscheinlichkeit)

Ereignisbaum: **Nacheinander Ziehen ohne zurücklegen**



Multiplikationssatz (Pfadwahrscheinlichkeit) $P(A \text{ und } B)$

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



Satz der Totalen Wahrscheinlichkeit $P(B)$

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



Satz von Bayes $P(B \text{ gegeben } A)$

B gegeben A = B ist abhängig von A.

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{|A \cap B|}{|A|} = 1 - P(\bar{B}|A)$$

Stochastisch unabhängig

Ereignisse, die keinen Einfluss aufeinander haben oder nicht zur selben Zeit passieren können. Oder: Elemente einer Stichprobe werden zufällig aber mit der selben Wahrscheinlichkeit ausgewählt.

Definition

$$P(X = x \wedge Y = y) = P(X = x) \cdot P(Y = y)$$

Multiplikationssatz $P(A \text{ und } B)$

$$P(A \cap B) = P(A) \cdot P(B) = \frac{|A|}{|\Omega|} \cdot \frac{|B|}{|\Omega|} = \frac{|A \cdot B|}{|\Omega|^2}$$

$$P(A \cap (\Omega \setminus B)) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A) \cdot (1 - P(B)) = P(A) \cdot P(\Omega \setminus B)$$

Wahrscheinlichkeit eines Ereignisses $P(A)$ $0 \leq P(A) \leq 1$

$$P(A) = \frac{|A|}{|\Omega|} \quad P(A|B) = P(A) \quad \text{falls } P(B) \neq 0$$

$$P(B|A) = P(B) \quad \text{falls } P(A) \neq 0$$

Diskrete Zufallsvariable X

Eine Zufallsvariable X ist eine Funktion, die jedem Ergebnis aus Ω eine reelle Zahl zuweist.

X_i = Merkmalsausprägung des i-ten Elements in der Stichprobe.

x_i = Stichprobenwert (Realisierung der Zufallsvariable)

Dichtefunktion / PMF bzw. PDF	$f(x) = P(X = x)$
Kumulative Verteilungsfunktion/ CDF	$F(x) = P(X \leq x) = \sum_{x \leq X} f(x)$
Wahrscheinlichkeiten	$P(a \leq X \leq b) = \sum_{a \leq x \leq b} f(x)$
Graphische Darstellung von f	Stabdiagramm
Erwartungswert	$E(X) = \sum_{x \in \mathbb{R}} f(x) \cdot x$
Varianz	$V(X) = \sum_{x \in \mathbb{R}} f(x) \cdot (x - E(X))^2$

$$P(X = 2) = f(2) = P(X \leq 2) - P(X \leq 1) = F(2) - F(1)$$

$$P(X \neq 2) = 1 - P(X = 2) = 1 - f(2)$$

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - F(3)$$

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2)$$

$$P(X < 2) = P(X \leq 1) = F(1)$$

$$P(X \leq 2) = P(0 \vee 1 \vee 2) = P(0) + P(1) + P(2)$$

$$P(X \leq 3.6) = P(X \leq 3)$$

$$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) \quad X = \{\dots, 5, 6\}$$

$$P(1 \leq X < 3) = F(2) - F(0)$$

$$P(1 < X \leq 3) = F(3) - F(1)$$

$$P(4 \leq X \leq 6) = P(X \leq 6) - P(X \leq 3) = P(X \leq 6) - (1 - P(X \geq 4))$$

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

$$f(-3) = 0$$

$$f(2.2) = 0$$

Erwartungswert E(X)

= arithmetisches Mittel / Mittelwert

$$E(X) = \sum_{x \in \mathbb{R}} P(X = x) \cdot x$$

Stochastisch unabhängig:

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

Varianz V(X)

$$V(X) = \sum_{x \in \mathbb{R}} P(X = x) \cdot (x - E(X))^2$$

Verschiebungssatz (gilt auch für stetige)

$$V(X) = E(X^2) - (E(X))^2 = \left(\sum P(X = x) \cdot x^2 \right) - E(X)^2$$

Linearität für stetige und diskrete Zufallsvariablen

$$E(\alpha X + \beta) = \alpha E(X) + \beta \quad \alpha \in \mathbb{R}$$

$$E(\alpha X) = \alpha E(X)$$

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

$$V(\alpha X + \beta) = \alpha^2 \cdot V(X)$$

$$V(X + Y) = V(X) + V(Y)$$

Für unabhängige / unkorrelierte X und Y

Stetige Zufallsvariable X

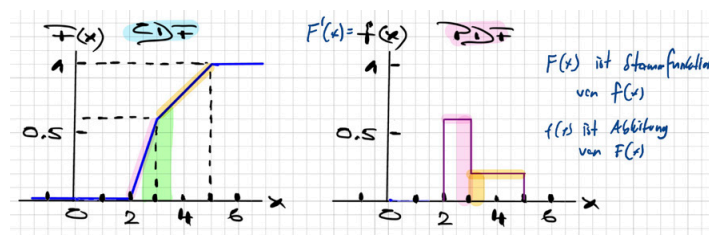
Dichtefunktion / PMF bzw. PDF	$f(x) = F'(x) \neq P(X = x)!!!$
Kumulative Verteilungsfunktion/ CDF	$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$
Wahrscheinlichkeiten	$P(a \leq X \leq b) = \int_a^b f(x) dx$
Graphische Darstellung von f	Graph
Erwartungswert	$E(X) = \int_{-\infty}^{\infty} f(x) \cdot x dx$
Varianz	$V(X) = \int_{-\infty}^{\infty} f(x) \cdot (x - E(X))^2 dx$

$$\left. \begin{aligned} P(a \leq X \leq b) \\ P(a < X \leq b) \\ P(a < X < b) \end{aligned} \right\} = \int_a^b f(x) dx = F(b) - F(a)$$

$$F(a) = \int_{-\infty}^a f(x) dx \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$P(a < X) = 1 - F(a) \quad P(X = x) = 0$$

Wahrscheinlichkeit P aus PDF oder CDF bestimmen



Bestimme P aus CDF

$$P(2.5 \leq X \leq 3.5) = P(X \leq 3.5) - P(X \leq 2.5) = F(3.5) - F(2.5) = 0.7 - 0.3 = 0.4$$

Bestimme P aus PDF

$$P(2.5 \leq X \leq 3.5) = \int_{2.5}^{3.5} f(x) dx = 0.6 \cdot 0.5 + 0.2 \cdot 0.3 = 0.4$$

Stetige Zufallsvariable

$$f(x) = \begin{cases} a(2x - x^2) & 0 \leq x \leq 2 \\ 0 & \text{sonst} \end{cases}$$

Bestimme die CDF.

$$F(x) = \int_0^x f(x) dx = 1 \Rightarrow a \int_0^2 2x - x^2 dx = 1$$

$$\Rightarrow a \left(x^2 - \frac{1}{3} x^3 \right) \Big|_0^2 = 1 \Rightarrow a = \frac{3}{4}$$

$$F(x) = \frac{3}{4} x^2 - \frac{1}{4} x^3$$

Bestimme Erwartungswert sowie Varianz.

$$E(x) = \int_0^2 f(x) \cdot x dx = \frac{3}{4} \int_0^2 2x^2 - x^3 dx = \frac{3}{4} \left(\frac{2}{3} x^3 - \frac{1}{4} x^4 \right) \Big|_0^2 = 1$$

$$E(x^2) = \frac{3}{4} \int_0^2 2x^3 - x^4 dx = \frac{3}{4} \left(\frac{1}{2} x^4 - \frac{1}{5} x^5 \right) \Big|_0^2 = 1.2$$

$$V(x) = E(x^2) - E(x)^2 = 1.2 - 1^2 = 0.2$$

Bedingte Wahrscheinlichkeit

	F (Frauen)	M (Männer)	Summe
R (Raucher/in)	$P(F \cap R) = \frac{100}{1400}$	$P(M \cap R) = \frac{400}{1400}$	$P(R) = \frac{500}{1400}$
N (Nichtraucher/in)	$P(F \cap N) = \frac{200}{1400}$	$P(M \cap N) = \frac{700}{1400}$	$P(N) = \frac{900}{1400}$
Summe	$P(F) = \frac{300}{1400}$	$P(M) = \frac{1100}{1400}$	$P(\Omega) = \frac{1400}{1400} = 1$

Wie gross ist die Wahrscheinlichkeit, eine Raucherin anzutreffen?

$$P(F \cap R) = 100 / 1400 = 1/14$$

Sie treffen eine Frau an. Mit welcher Wahrscheinlichkeit ist sie Raucherin?

$$P(R | F) = 100 / 300 = 1/3$$

Mit welcher Wahrscheinlichkeit ist eine Person, die raucht, ein Mann?

$$P(M | R) = 400 / 500 = 4/5$$

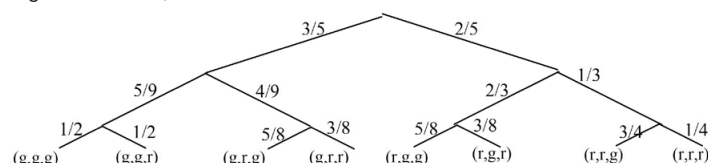
Ziehen nacheinander ohne zurücklegen

$$\Omega = \{ r, r, r, r, r, g, g, g, g, b, b, b \}, |\Omega| = 13$$

$$P(\{ r, r, g, g, b \}) = 6/13 \cdot 5/12 \cdot 4/11 \cdot 3/10 \cdot 3/9$$

$$\Omega = \{ g, g, g, g, g, r, r, r, r \}, |\Omega| = 10$$

$$P(g) = 6/10 = 3/5, P(r) = 4/10 = 2/5$$



$$P(\{ g, g, r \}) = 3/5 \cdot 5/9 \cdot 1/2 = 1/6$$

Stochastische Unabhängigkeit

y \ x	0 (schwarz)	1 (rot)	P(Y = y)
0 (schwarz)	1/4	1/4	1/2
1 (rot)	1/4	1/4	1/2
P(X = x)	1/2	1/2	1 (Spalten- bzw. Zeilensumme)

$$P(X = x \wedge Y = y) = P(X = x) \cdot P(Y = y)$$

Beispiele Wahrscheinlichkeit

Bsp. Zufallsvariable (stochastisch unabhängig)

$$\text{geg. zwei Würfel, } A=6, P(A) = \frac{1}{6} \\ \rightarrow \text{Zielgruppe} = 2$$

$$P(2\text{-mal } 6) = P(A \wedge A) = P(A) \cdot P(A) = \frac{1}{36} \\ \{6, 6\}$$

$$P(\text{genau eine } 6) = P(A) \cdot P(\bar{A}) + P(\bar{A}) \cdot P(A) = 2 \cdot P(A) \cdot P(\bar{A}) = \frac{5}{18} \\ \{6, \bar{6}\} \vee \{\bar{6}, 6\}$$

$$P(\text{keine } 6) = P(\bar{A}) \cdot P(\bar{A}) = \frac{25}{36} \\ \{\bar{6}, \bar{6}\}$$

$$P(\text{mindestens eine } 6) = P(\geq 1) \rightarrow \text{Komplementär} \\ = 1 - P(=0) = 1 - P(\text{keine } 6) = 1 - P(\bar{A}) \cdot P(\bar{A}) = \frac{11}{18} \\ 1 - \{\bar{6}, \bar{6}\}$$

$$P(\text{höchstens eine } 6) = P(0 \vee 1) = P(\leq 1) = P(\text{keine } 6) + P(\text{genau eine } 6) \\ \{\bar{6}, \bar{6}\} \vee \{6, \bar{6}\} \vee \{\bar{6}, 6\} = P(\bar{A}) \cdot P(\bar{A}) + P(A) \cdot P(\bar{A}) + P(\bar{A}) \cdot P(A) = \frac{25}{36}$$

$$P(\text{Summe } 11) = P(5 \text{ und } 6 \text{ oder } 6 \text{ und } 5) = P(5) \cdot P(6) + P(6) \cdot P(5) = \frac{1}{18} \\ \{5, 6\} \vee \{6, 5\}$$

Multiplikationssatz P(A und B)

4-seitiger Würfel und eine Münze.

Was ist die Wahrsch. für eine ungerade Zahl und Kopf?

$$\Omega_W = \{1, 2, 3, 4\} \quad \Omega_U = \{1, 3\}$$

$$\Omega_M = \{\text{Kopf, Zahl}\} \quad \Omega_K = \{\text{Kopf}\}$$

$$\Omega = \Omega_W \times \Omega_M = \{(1,K), (2,K), (3,K), (4,K), (1,Z), (2,Z), (3,Z), (4,Z)\}$$

$$P(\text{Ungerade und Kopf}) = \Omega_U / \Omega_W \cdot \Omega_K / \Omega_M = 2/4 \cdot 1/2 = 2/8 = 1/4$$

Wahrscheinlichkeit für mindestens eine sechs bei 4 Würfeln:

$$1 - P(\text{keine } 6) = 1 - P(\{\bar{6}, \bar{6}, \bar{6}, \bar{6}\}) \\ = 1 - P(\bar{6})^4 = 1 - \left(\frac{5}{6}\right)^4 \approx 0.518$$

Wie viele unabhängige parallelgeschaltete Komponenten von 35% braucht es, damit das System mit einer Zuverlässigkeit von mehr als 99% läuft?

Das System funktioniert, wenn mind. eine Komponente läuft.

$$P(\text{System ok}) = P(\text{Komponente ok}) = 1 - P(\text{Komponente ausgefallen}) = 0.35 \\ \Rightarrow 1 - 0.65^n > 0.99 \Rightarrow n > \log_{0.65}(0.01) = 10.6...7 \\ \Rightarrow n = 11$$

Komplementäreignis

Mindestens 3 Würfel lang erscheint kein Kopf:

$$A = \{ZZZK, ZZZZK, ZZZZK, \dots\}$$

mindestens $\rightarrow$ von Komplementäreignis ausgehen:

$$\bar{A} = \{K, ZK, ZK, \dots\}$$

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

$$= 1 - [P(K) + P(ZK) + P(ZZK)] = 1 - [1/2 + 1/4 + 1/8] = 1/8$$

Diskrete Wahrscheinlichkeitsverteilungen (PMF von X)

Dichtefunktionen / Zähldichte $P(\{\omega\})$

Für Zufallsexperimente unabhängiger gleichartiger Ereignisse in einem bestimmten Zeitfenster:

- gleichzeitiges Ziehen → Hypergeom. Vert.
- (nacheinander) Ziehen ohne Zurücklegen → Hypergeom. Vert.
- (nacheinander) Ziehen mit Zurücklegen → Binomialverteilung

Übersicht

	Hypergeometrische Vert.	Binomialverteilung	Poissonverteilung
Dichtefunktion (Zähldichte) (PMF)	$P(X=x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$	$P(X=x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$	$P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda}$
Erwartungswert	$n \cdot \frac{M}{N}$	$n \cdot p$	λ
Varianz	$n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$	$n \cdot p \cdot (1-p)$	λ
Standardabweichung	$\sqrt{n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}}$	$\sqrt{n \cdot p \cdot (1-p)}$	$\sqrt{\lambda}$
Parameter	n, N, M	n, p	λ

Hypergeometrische Verteilung

Gleichzeitiges Ziehen oder (nacheinander) Ziehen ohne Zurücklegen

$$X \sim H(N, M, n)$$

- N Gesamtanzahl Objekte
- n Anzahl Experimente / Ziehungen bzw. Objekte in einer Stichprobe
- $N-M$ Anzahl andersartige Objekte
- M Anzahl Merkmalsträger
- x Anzahl Merkmalsträger in einer Stichprobe

Gleichzeitiges Ziehen (Reihenfolge spielt keine Rolle)

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

Kombinationen mit x-Merkmalsträger
Kombinationen ohne Merkmalsträger
Total Anzahl Kombinationen

(Nacheinander) Ziehen ohne Zurücklegen (Reihenfolge beachten)

$$P(X=x) = \frac{M P_x \cdot (N-M) P_{(n-x)}}{N P_n}$$

Erwartungswert und Varianz

$$\mu = E(X) = n \cdot \frac{M}{N} \quad \sigma^2 = V(X) = n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}$$

$$\sigma = S(X) = \sqrt{n \cdot \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}}$$

Approximation durch Binomialverteilung

rechnerisch einfacher

$$\text{Wenn } n \leq \frac{N}{20} \text{ dann } H(N, M, n) \approx B(n, \frac{M}{N})$$

Bernoulli-Verteilung

Wenn Zufallsvariable nur zwei mögliche Werte annehmen kann.

$$p = P(X=1) \quad q = 1 - p = P(X=0)$$

$$E(X) = E(X^2) = p$$

$$V(X) = pq = p \cdot (1-p)$$

$$S(X) = \sqrt{pq}$$

Binomialverteilung

(Nacheinander) Ziehen mit Zurücklegen (Reihenfolge beachten)

Bernoulli-Verteilung n -mal hintereinander. Wiederholungen sind stochastisch unabhängig. X zählt, wie oft das Ereignis 1 bei der n -fachen Wiederholung eines Bernoulli-Experiments eintritt.

$$X \sim B(n, p) \quad X = X_1 + \dots + X_n \quad n = \text{Anzahl Experimente}$$

$$P(X=x) = \binom{n}{x} \cdot p^x \cdot q^{n-x}$$

$$P(X \leq m) = \sum_{x=0}^m \binom{n}{x} \cdot p^x \cdot q^{n-x}$$

$$P(a \leq X \leq b) = \sum_{x=a}^b \binom{n}{x} \cdot p^x \cdot q^{n-x}$$

$$\binom{n}{x} \cdot p^n \quad \text{Wenn } p = q$$

$$\mu = E(X) = np$$

$$\sigma^2 = V(X) = npq$$

$$\sigma = S(X) = \sqrt{npq}$$

Approximation an hypergeometrische Verteilung

$$\text{Wenn } n \leq \frac{N}{20} \text{ dann } H(N, M, n) \approx B(n, \frac{M}{N})$$

Approximation durch Poissonverteilung

$$\text{Wenn } n \geq 50 \text{ und } p \leq 0.1 \text{ dann } B(n, p) \approx \text{Poi}(n \cdot p)$$

Beispiele

Eine Münze wird vier mal geworfen.
 X sei Anzahl Kopf.

$$P(X=3) = \binom{4}{3} \cdot \frac{1}{16} = \frac{4}{16}$$

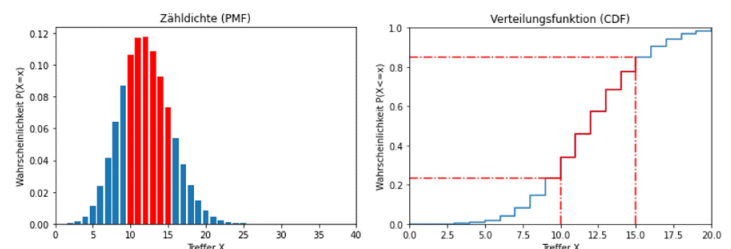
oder Binomialverteilung:

$$= \binom{4}{3} \cdot P(K)^3 \cdot P(\bar{K})^1$$

$$= \binom{4}{3} \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^1 = \frac{4}{16}$$

Bsp

$$P(10 \leq X \leq 15) = P(X \leq 15) - P(X \leq 9) = F(15) - F(9) \\ = \sum_{k=10}^{15} \binom{200}{k} \cdot 0.06^k \cdot 0.94^{200-k} \approx 0.6169$$



Poissonverteilung

Für gleichartige unabhängige Ereignisse.

Durchschn. **Anzahl Ereignisse pro Zeitintervall / räumlichen Bereich**, z.B:

- Radioaktiver Zerfall von Atomen **pro** Sekunde
- Autounfall **pro** Wochenende
- Anzahl Lackfehler **pro** cm²

$$X \sim \text{Poi}(\lambda) \quad \lambda \text{ Werte } 0, 1, 2, \dots \text{ an}$$

Einheit von λ muss evtl. angepasst werden, so dass ganzzahlig
z.B: 0.05 cm² $\rightarrow \lambda = 5$. λ ist somit als 1/100 cm² zu verstehen.

$$P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad P(X \leq m) = e^{-\lambda} \sum_{x=0}^m \frac{\lambda^x}{x!}$$

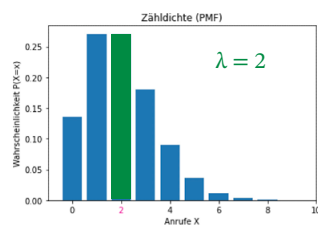
$$\mu = E(X) = \lambda$$

$$\sigma^2 = V(X) = \lambda$$

$$\sigma = S(X) = \sqrt{\lambda}$$

Approximation an Binomialverteilung

Wenn $n \geq 50$ und $p \leq 0.1$ dann $B(n, p) \approx \text{Poi}(n \cdot p)$



Verteilung wird annähernd symmetrisch für $\lambda \geq 10$.

Beispiele Verteilung

$$\{w, w, w, s, s\} \quad N=5, \quad M=3, \quad N-M=2, \quad n=3$$

Gleichzeitiges ziehen: $\rightarrow$ Hypergeometrische Verteilung $X \sim H(5, 3, 2)$

$$P(X=2) = P(w, w, s) = \frac{1}{\binom{5}{3}} \binom{3}{2} \binom{2}{1} = 0.6$$

Ziehen mit Zurücklegen: $\rightarrow$ Binomialverteilung $X \sim B(2, 3/5)$

$$P(X=2) = P(w, w, s) \vee (w, s, w) \vee (s, w, w) \neq P(w, w, s) \\ = \binom{3}{2} \cdot \left(\frac{3}{5}\right)^2 \cdot \left(1 - \frac{3}{5}\right)^1 = \binom{3}{2} \cdot \frac{3}{5} \cdot \frac{3}{5} \cdot \frac{2}{5} = 0.432$$

Ziehen ohne Zurücklegen: $\rightarrow$ Hypergeometrische Verteilung $X \sim H(5, 3, 2)$

$$P(X=2) = P(w, w, s) = P(w, s, w) = P(s, w, w) \\ = \frac{1}{5P_3} \cdot 3P_2 \cdot 2P_1 = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = 0.2$$

Auflösen nach $n \rightarrow$ Frage wie oft?

Wie oft muss mit zwei Würfeln gewürfelt werden, um mit Wahrscheinlichkeit 0.9 oder mehr mindestens einmal die Augensumme 12 zu erhalten? Vergleichen Sie die exakte Lösung mit der Abschätzung, die sich aus der Poissonapproximation ergibt.

Ziehen mit zurücklegen $\rightarrow X \sim B(n, p)$

$$p = P(\text{Augensumme } 12) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}, \quad q = \frac{35}{36}$$

$n = \text{gesucht}$

$X = \text{Anzahl Würfe mit Augenzahl } 12 \text{ bei } n \text{ Würfeln}$

Exakt

$$X \sim B\left(n, \frac{1}{36}\right) \geq 0.9$$

Es muss mind. einmal gewürfelt werden:

$$P(X > 0) \geq 0.9 \Rightarrow 1 - P(X=0) \leq 0.9$$

$$\Rightarrow 1 - \binom{n}{0} \left(\frac{1}{36}\right)^0 \cdot \left(\frac{35}{36}\right)^{n-0} \leq 0.9$$

$$\Rightarrow 1 - 1 \cdot 1 \cdot \left(\frac{35}{36}\right)^n \leq 0.9$$

$$\Rightarrow n \geq \log_{\frac{35}{36}}(0.1) = \lceil 81.7 \rceil \Rightarrow n = 82$$

Approximation

$$X \sim \text{Poi}(\lambda) \quad \lambda = n \cdot p = \frac{n}{36}$$

$$1 - P(X=0) \geq 0.9 \Rightarrow 1 - \frac{\left(\frac{n}{36}\right)^0}{0!} e^{-\frac{n}{36}} \geq 0.9$$

$$\Rightarrow e^{-\frac{n}{36}} \leq 0.1 \Rightarrow -\frac{n}{36} \leq \ln(0.1)$$

$$\Rightarrow n \geq -36 \cdot \ln(0.1) = \lceil 82.8 \rceil \Rightarrow n = 83$$

Beispiele Normalverteilung

Anstatt 400 Rechnungsposten genau zu addieren, addiert eine Kauffrau diese nur auf ganze Franken gerundet. Wir nehmen an, dass der Fehler pro Posten auf dem Intervall $(-0.5; 0.5]$ gleichverteilt ist. Wie gross ist dann (approximativ) die Wahrscheinlichkeit, dass ihre Rechnung insgesamt um weniger als 10 Franken vom tatsächlichen Wert abweicht?

$$X \sim N(n\mu, \sqrt{n}\sigma)$$

$$E(S_{400}) = n \cdot \mu = 400 \cdot E(X_i) = 400 \cdot \frac{-0.5 + 0.5}{2} = 0$$

$$V(X_i) = \int_{-0.5}^{0.5} (x - E(X))^2 dx = \int_{-0.5}^{0.5} x^2 dx = \frac{1}{3} x^3 \Big|_{-0.5}^{0.5} = 0.08\bar{3}$$

$$V(S_{400}) = n \cdot \sigma^2 = 400 \cdot V(X_i) = 400 \cdot 0.08\bar{3} = 33.\bar{3}$$

$$\sigma = \sqrt{0.08\bar{3}} = 0.2887$$

$$P(|S_{400}| < 10) = 2 \cdot \Phi\left(\frac{10 - n\mu}{\sqrt{n} \cdot \sigma}\right) - 1 = 0.9168$$

0.9584 (interpoliert)

Eine faire Münze wird 1000 Mal geworfen. Berechnen Sie die Wahrscheinlichkeit, dass 540 Mal Zahl geworfen wird. Approximieren Sie dann diese Wahrscheinlichkeit mit Hilfe einer geeigneten Normalverteilung.

$$n = 1000 \quad X = \text{Anzahl Zahl} \quad p = q = \frac{1}{2}$$

Exakt mit Binomialverteilung:

$$P(X=540) = \binom{1000}{540} \left(\frac{1}{2}\right)^{1000} = 0.001$$

Python

Approximation mit Normalverteilung

$$X \sim N(\mu = np, \sigma = \sqrt{npq}) = N(\mu = 500, \sigma = 15.8114)$$

$$P(X=540) = \frac{1}{\sqrt{2\pi} \cdot 15.8114} \cdot e^{-\frac{1}{2} \left(\frac{540-500}{15.8114}\right)^2} = 0.001$$

Normalverteilung PMF $f(x)$

Dichtefunktion für **Stetige Wahrscheinlichkeitsverteilung**.
Auch Gauss-Verteilung genannt.

$$X \sim N(\mu; \sigma) \quad \mu, \sigma \in \mathbb{R}, \sigma > 0$$

μ = Erwartungswert (= Kurvenhöhenpunkt) (mü)
 σ = Standardabweichung (sigma)

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

Varianz

$$\sigma^2 = \int_a^b (x - \mu)^2 dx \quad \text{Siehe Stetige Zufallsvariable X}$$

Standardnormalverteilung

$$N(0; 1) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$$

Eigenschaften

- Wendepunkte $\mu - \sigma$ und $\mu + \sigma$
- Symmetrisch
- Normiert, d.h. $\int_{-\infty}^{\infty} \varphi_{\mu, \sigma}(x) dx = 1$

$$E(X) = \mu$$

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

Bei einer Zufallsvariable X , die der Normalverteilung $N(\mu; \sigma)$ folgt, liegen

- ca. 68 % der beobachteten Werte zwischen $\mu - \sigma$ und $\mu + \sigma$,
- ca. 95 % der beobachteten Werte zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$,
- ca. 99.7 % der beobachteten Werte zwischen $\mu - 3\sigma$ und $\mu + 3\sigma$.

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$$

Kumulative Verteilfunktion CDF $\Phi(x)$ $F(x)$

Kann **nicht** auf elementare Weise **berechnet werden**

→ **Ablesen von Tabelle** (Papula S. 514)

$$\Phi_{\mu, \sigma}(x) = P(X \leq x) = \int_{-\infty}^x \varphi_{\mu, \sigma}(t) dt$$

Standardisierung (Z-Score, Standard score)

Wenn eine **beliebige** Normalverteilung vorliegt muss standardisiert werden.

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

Eigenschaften (ohne Standardisierung)

$$P(X \leq c) = \Phi(c)$$

$$P(X \leq -c) = 1 - \Phi(c)$$

$$P(X \geq c) = 1 - P(X \leq c) = 1 - \Phi(c)$$

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

$$P(-a \leq X \leq b) = \Phi(b) - (1 - \Phi(a))$$

$$\begin{aligned} P(\mu - c \leq X \leq \mu + c) &= P(|X - \mu| \leq c) \\ &= 2 \cdot \Phi\left(\frac{c}{\sigma}\right) - 1 = 1 - 2 \cdot \Phi\left(-\frac{c}{\sigma}\right) \end{aligned}$$

Standardisieren

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$\begin{aligned} P(\mu - c \leq X \leq \mu + c) &= P\left(-\frac{c}{\sigma} \leq U \leq \frac{c}{\sigma}\right) \\ &= 2 \cdot \Phi\left(\frac{c}{\sigma}\right) - 1 = 1 - 2 \cdot \Phi\left(-\frac{c}{\sigma}\right) \end{aligned}$$

$$P(-e \leq U \leq e) = P(|U| \leq e) = 2 \cdot \Phi(e) - 1 = 1 - 2 \cdot \Phi(-e)$$

Bsp. mit $N(\mu = 3, \sigma = 2)$

$$\begin{aligned} P(1.26 \leq X \leq 5.12) &= P\left(\frac{1.26-3}{2} \leq U \leq \frac{5.12-3}{2}\right) \\ &= \Phi(1.06) - (1 - \Phi(0.87)) \end{aligned}$$

Interpolieren

Wenn Grenze mehr als zwei Nachkommastellen hat → **linear interpolieren**:

$$P(X \leq 0.037) = \Phi(0.03) + 0.007 \cdot \frac{\Phi(0.04) - \Phi(0.03)}{0.01}$$

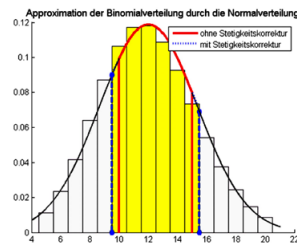
Stetigkeitskorrektur bei diskreten Werten!

Keine Stetigkeitskorrektur nötig wenn x -Werte bereits **stetig** sind.

$$P(a \leq X \leq b) = P(a - 0.5 \leq X \leq b + 0.5)$$

$$P(a < X < b) = P(a + 0.5 \leq X \leq b - 0.5)$$

$$P(X \geq a) = P(X \geq a + 0.5)$$



Zentraler Grenzwertsatz

Wenn alle Zufallsvariablen X_i identisch verteilt und stochastisch unabhängig sind (und denselben Erwartungswert und dieselben Varianz haben).

Verteilfunktion $F(x)$ der standardisierten Zufallsvariable

Wenn die **Summe** aller X_i gefragt ist:

$$X \sim N(n\mu, \sqrt{n} \cdot \sigma = \sqrt{n\sigma^2}) \quad U_n = \frac{(X_1 + X_2 + \dots + X_n) - n\mu}{\sqrt{n} \cdot \sigma}$$

$$E(S_n) = E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n) = n \cdot \mu$$

$$V(S_n) = V(X_1 + \dots + X_n) = V(X_1) + \dots + V(X_n) = n \cdot \sigma^2$$

Wenn das **Mittel** $\bar{X}$ aller X_i gefragt ist:

$$X \sim N\left(\mu, \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}\right) \quad U_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$E(\bar{X}_n) = E\left(\frac{1}{n} [X_1 + \dots + X_n]\right) = \frac{1}{n} [E(X_1) + \dots + E(X_n)] = \mu$$

$$V(\bar{X}_n) = V\left(\frac{1}{n} [X_1 + \dots + X_n]\right) = \frac{1}{n^2} [V(X_1) + \dots + V(X_n)] = \frac{\sigma^2}{n}$$

Bsp. **nach n auflösen**

$$\mu = 0, \sigma = 4 \quad X \sim N\left(0, \frac{4}{\sqrt{n}}\right) \quad \text{gesucht ist } n$$

$$2 \cdot \Phi\left(\frac{2}{\frac{4}{\sqrt{n}}}\right) - 1 \geq 0.95 \Rightarrow \Phi\left(\frac{2}{\frac{4}{\sqrt{n}}}\right) \geq \frac{1 + 0.95}{2}$$

$\underbrace{\hspace{1cm}}_{0.975} \quad \underbrace{\hspace{1cm}}_{0.975}$

Tabelle → 1.96

$$\frac{2}{\frac{4}{\sqrt{n}}} \geq 1.96 \Rightarrow n \geq \lceil 15.3 \rceil \Rightarrow n = 16$$

Approximation an Binomialverteilung

Wenn $npq > 9$. X ist näherungsweise normalverteilt mit:

$$\mu = np \quad \sigma^2 = npq \quad \sigma = \sqrt{npq}$$

$$P(a \leq X \leq b) = \sum_{x=a}^b P(X = x) \approx \Phi_{\mu, \sigma}\left(b + \frac{1}{2}\right) - \Phi_{\mu, \sigma}\left(a - \frac{1}{2}\right)$$

$$P(X = x) = \varphi_{\mu, \sigma}(x)$$

Approximation an Poissonverteilung

Wenn λ gross genug ist. X ist näherungsweise normalverteilt mit:

$$\mu = \lambda \quad \sigma^2 = \lambda \quad \sigma = \sqrt{\lambda}$$

Gleiche Formeln

Lineare Regression

Methode der kleinsten Quadrate (KQM).

Die Regressionsgerade verläuft immer durch **Schwerpunkt** $(\bar{x}, \bar{y})$.

Anwendung: **Prognose (Erklärung)** von Merkmal Y durch Merkmal X.

Steigung

$$m = \frac{S_{xy}}{S_x^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}}{\frac{\sum(x_i - \bar{x})^2}{n}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$\tilde{s}_{xy}$ = Kovarianz (empirisch), $\tilde{s}_x^2$ = Varianz (empirisch)

y-Achsenabschnitt

$$d = \bar{y} - m\bar{x}$$

Residuen

$$\epsilon_i = y_i - \hat{y}_i$$

Residualvarianz (minimale)

$$\tilde{s}_\epsilon^2 = \tilde{s}_y^2 - \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$$

Prognostizierte y-Werte $\hat{y}$

= Punkte auf der Regressionslinie $g(x)$

$$\hat{y}_i = g(x_i) = mx_i + d$$

Prognostizierte (erklärte) Varianz

$$\tilde{s}_{\hat{y}}^2 = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (mx_i + d - \bar{y})^2 = m^2 \tilde{s}_x^2 = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2}$$

(Gesamt-)Varianz (x oder y)

Arithmetische Mittelwerte

$$\tilde{s}_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{und} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Kovarianz

$$\tilde{s}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - \bar{x} \cdot \bar{y} = \overline{xy} - \bar{x} \cdot \bar{y}$$

Bestimmtheitsmass R^2

Beurteilt die globale Anpassungsgüte einer Regression: [0, 1].

0 = schlecht, 1 = perfekt (alle Datenpaare liegen auf der Regressionslinie)

$$R^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2} = \frac{\tilde{s}_{xy}^2}{\tilde{s}_x^2 \tilde{s}_y^2} = \frac{(\overline{xy} - \bar{x} \cdot \bar{y})^2}{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)} = r_{xy}^2$$

R^2 stimmt überein mit dem Quadrat des **Korrelations-Koeffizienten** r_{xy} (nach Bravais-Pearson).

Interpretation

R^2 = Prozentmass. $R^2 = 0.75$ ist 1.5-mal so gut wie $R^2 = 0.5$.

Die x/y-Beziehung weist eine R^2 (= Prozent) Variation auf. Oder:

R^2 (= Prozent) der Gesamtvarianz in den y-Daten kann durch die Regressionsgerade erklärt werden

Nichtlineare Regression

Ausgangsfunktion	Transformation
$y = q \cdot x^m$	$\log(y) = \log(q) + m \cdot \log(x)$
$y = q \cdot m^x$	$\log(y) = \log(q) + \log(m) \cdot x$
$y = q \cdot e^{m \cdot x}$	$\ln(y) = \ln(q) + m \cdot x$
$y = \frac{1}{q + m \cdot x}$	$V = q + m \cdot x; \quad V = \frac{1}{y}$
$y = q + m \cdot \ln(x)$	$y = q + m \cdot U; \quad U = \ln(x)$
$y = \frac{1}{q \cdot m^x}$	$\log\left(\frac{1}{y}\right) = \log(q) + \log(m) \cdot x$

$q = d$

Vorgehen

1. y-Werte logarithmieren: $\log(y_i)$ oder $\ln(y_i)$
2. m und d berechnen (mit neuen logarithmierten y-Werten)
3. m und d Rücktransformieren: 10^m und 10^d resp. e^m und e^d
4. Rücktransformierte m und d in Ausgangsfunktion einsetzen

Beispiele nichtlineare Regression

$$y = d \cdot m^x$$

$$\begin{array}{ccccc} x & 2 & 5 & 8 & 12 & 15 \\ \log y & 0.6 & 0.7 & 1.1 & 1.3 & 1.5 \end{array} \quad \overline{\log y} = 1.0763, \quad \overline{x^2} = 92.4$$
$$\bar{x} = 8.4, \quad \bar{y} = 1.0845$$

$$m = \frac{\overline{\log y} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = 0.0737 \xrightarrow{10^{0.0737} \text{ Rücktransformation}} 1.18$$

$$d = \bar{y} - m\bar{x} = 1.0845 - 0.0737 \cdot 8.4 = 0.4705 \xrightarrow{10^{0.4705}} 2.9543$$

$$y = 2.9543 \cdot 1.18^x$$

$$T - 300 = C \cdot a^b = d \cdot m^x \Rightarrow \log(T - 300) = \log(d) + \log(m) \cdot x$$

$$\begin{array}{ccccc} t & 5 & 10 & 15 & 20 & 30 \\ C & 720 & 540 & 430 & 380 & 335 \end{array}$$

$$y_i = \ln(T_i - 300) = 6.1, 5.5, 4.9, 4.4, 3.6$$

$\underbrace{720, 540, \dots}$

$$\overline{\log y} = 70.44, \quad \bar{x} = 16, \quad \overline{x^2} = 330, \quad \bar{y} = \overline{\ln(T - 300)} = 4.87$$

$$m = \frac{\overline{\log y} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = -0.0997 \xrightarrow{e^{-0.0997}} 0.9057$$

$$d = \bar{y} - m\bar{x} = 4.87 - (-0.0997) \cdot 16 = 6.46 \xrightarrow{e^{6.46}} 639.06$$

$$T = d \cdot m^x + 300 = 639.06 \cdot 0.9057^x + 300$$

Mehrere Variablen

Prognostizierte y-Werte $\hat{y}$

= Punkte auf der Regressionslinie $g(x)$.

$$g(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n \rightarrow A = \begin{pmatrix} 1 & x_{11} & \dots & x_{1n} \\ 1 & x_{21} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \dots & x_{in} \end{pmatrix}$$

$$g(x) = \beta_1 x_1 + \dots + \beta_n x_n + \beta_{n+1} \rightarrow A = \begin{pmatrix} x_{11} & \dots & x_{1n} & 1 \\ x_{21} & \dots & x_{2n} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \dots & x_{in} & 1 \end{pmatrix}$$

Normalengleichung

p sind die Koeffizienten.

$$X^T X p = X^T y \quad p = (X^T X)^{-1} X^T y \quad p = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

Residualvektor

$$\epsilon = y - Xp$$

Wegen der orthogonalen Zerlegung gilt die Quadratsummenzerlegung:

$$y^2 = (Xp)^2 + \epsilon^2$$
$$y = \sqrt{(Xp)^2 + \epsilon^2}$$

More on the Normal equation in the appendix.

Preis soll durch den Jahrgang und den Kilometerstand modelliert werden.

Preis	Jahrgang	Kilometerstand
28500	2017	28100
31500	2019	28200
...	...	...
22500	2016	8400

$$y = \begin{pmatrix} 28500 \\ 31500 \\ \vdots \\ 22500 \end{pmatrix} \quad X = \begin{pmatrix} 2017 & 28100 & 1 \\ 2019 & 28200 & 1 \\ \vdots & \vdots & \vdots \\ 2016 & 8400 & 1 \end{pmatrix} \quad p = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

$p = \text{np.linalg.inv}(X.T @ X) @ X.T @ y$

oder

$p = \text{np.linalg.solve}(X, y)$

Preis = a * Jahrgang + b * Kilometerbestand + c

Interpretation der Steigung einer linearen Regression:

Die Zunahme von x um 1 Einheit bedeutet die Zunahme (oder Abnahme) um m (Steigung) Einheiten von y.

Parameterschätzung und Vertrauensintervall

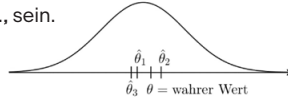
Deskriptiven Statistik: Werte einer Stichprobe/Grundgesamtheit übersichtlich darstellen → Tabellen/Graphiken

Schliessenden Statistik: Versucht Aussagen über die Grundgesamtheit zu machen.

Schätzfunktion $\Theta = g(X_1, X_2, \dots, X_n)$

Parameter θ = wahrer Wert: kann Mittelwert, Median, relative Häufigkeit, Standardabweichung, etc., sein.
 θ soll geschätzt werden.

Schätzwert $\hat{\theta} = g(x_1, x_2, \dots, x_n)$

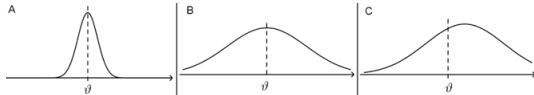


Kriterien

Erwartungstreue: $E(\Theta) = \theta$ bzw. $E(\bar{X}) = E(X_i) = \mu$

Effizienz: $V(\Theta_1) < V(\Theta_2)$

Konsistenz: $E(\Theta) \rightarrow \theta$ und $V(\Theta) \rightarrow 0$ für $n \rightarrow \infty$ bzw. $V(\bar{X}) = \frac{V(X_i)}{n} \xrightarrow{n \rightarrow \infty} 0$



A: Erwartungstreue und effizienter als B

B: Erwartungstreue und weniger effizient als A

C: Nicht erwartungstreu

$$E(\Theta_1) = E\left(\frac{1}{3} \cdot (X_1 + X_2 + X_3)\right) = \frac{1}{3} \cdot (E(X_1) + E(X_2) + E(X_3)) = \frac{1}{3} \cdot (\mu + \mu + \mu) = \mu$$

$$V(\Theta_1) = V\left(\frac{1}{3} \cdot (X_1 + X_2 + X_3)\right) = \frac{1}{9} \cdot V(X_1 + X_2 + X_3) = \frac{1}{9} \cdot (V(X_1) + V(X_2) + V(X_3)) \\ = \frac{1}{9} \cdot (\sigma^2 + \sigma^2 + \sigma^2) = \frac{\sigma^2}{3}$$

Wahl der Verteilung ↓

Siehe Tabelle unten.

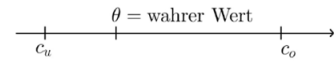
Normalverteilten Grundgesamtheit:

- Vertrauensintervall für den Erwartungswert μ :
 - Varianz σ^2 unbekannt und $n \leq 30 \rightarrow$ **t-Verteilung**
 - Varianz σ^2 bekannt $\rightarrow$ **Standardnormalverteilung**
- Vertrauensintervall für die Varianz $\sigma^2 \rightarrow$ **Chi-Quadrat-Verteilung**

Bernoulli-Verteilte Grundgesamtheit:

- Vertrauensintervall für den Anteilswert p mit $n\hat{p}(1-\hat{p}) > 9$ $\rightarrow$ **Standardnormalverteilung**

Vertrauensintervall $[c_u, c_o]$



$$P(\Theta_u \leq \theta \leq \Theta_o) = \gamma$$

$$P(\bar{X} - e \leq \mu \leq \bar{X} + e) = \gamma \quad P(|\bar{X} - \mu| \leq e) = \gamma \quad P\left(|U| \leq \frac{e}{\sigma/\sqrt{n}}\right) = \gamma$$

e ist gesucht.

γ = **Vertrauensniveau**: 0.90, 0.95 oder 0.99 (%)

α = **Irrtumswahrscheinlichkeit**: $\alpha = 1 - \gamma$

Bedeutung: γ (Prozent) der Stichproben liegen innerhalb des Intervalls.

Beantwortet die Fragestellung: **Wie gut ist eine Schätzung?**

- Vertrauensintervall grösser $\rightarrow$ Genauigkeit kleiner.
- Vertrauensniveau γ (= Sicherheit) grösser $\rightarrow$ Vertrauensintervall grösser.
- n (Stichprobengrösse) grösser $\rightarrow$ Vertrauensintervall kürzer.
- Varianz der Grundgesamtheit bekannt $\rightarrow$ Vertrauensintervall kleiner.
- Grenzen des Vertrauensintervall liegen symmetrisch um $\bar{X}$.

Vertrauensintervall bestimmen

Fehlt die Varianz der Grundgesamtheit, nimmt man die Varianz der Stichprobe (falls bekannt) = **Approximation**

$$n = 10, \bar{x} = 102, s^2 = 16, \gamma = 0.99$$

$$P(\bar{x} - e \leq \theta \leq \bar{x} + e) = 0.99 \Rightarrow P\left(\frac{|\bar{x} - \theta|}{\frac{s}{\sqrt{n}}} \leq \frac{e}{\frac{s}{\sqrt{n}}}\right) = 0.99$$

$\theta = \mu \rightarrow$ **t-Verteilung**

$$p = \frac{1+\gamma}{2} = 0.995, f = n-1 = 9, c = t(p, f) \stackrel{\text{Tabelle}}{=} 3.250$$

$$e = c \cdot \frac{s}{\sqrt{n}} = 3.25 \cdot \frac{\sqrt{16}}{\sqrt{10}} = 4.111$$

$$99\% \text{-Vertrauensintervall für } \mu: [\bar{x} - e, \bar{x} + e] \text{ bzw. } [\Theta_u, \Theta_o] = [97.89, 106.11]$$

$\theta = \sigma^2 \rightarrow$ **Chi-Quadrat-Verteilung**

$$f = n-1 = 9, p_1 = \frac{1-\gamma}{2} = 0.005, p_2 = \frac{1+\gamma}{2} = 0.995$$

$$c_1 = z(p_1, f) \stackrel{\text{Tabelle}}{=} 1.73, c_2 = z(p_2, f) \stackrel{\text{Tabelle}}{=} 2.359$$

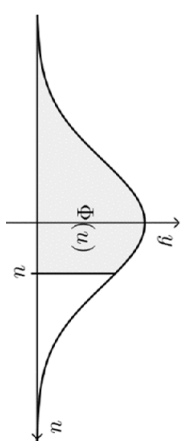
$$\Theta_u = \frac{(n-1) \cdot s^2}{c_2} = 6.10, \Theta_o = \frac{(n-1) \cdot s^2}{c_1} = 83.24$$

$$99\% \text{-Vertrauensintervall für } \sigma^2: [\Theta_u, \Theta_o] = [6.10, 83.24]$$

	(1) Verteilung der Grundgesamtheit	(2) Param.	(3) Schätzfunktionen	(4) zugehörige standardisierte Zufallsvariable	(5) Verteilung und benötigte Quantile	(6) Zufallsvariablen für Intervallgrenzen
1	Normalverteilung (Varianz σ^2 bekannt)	μ	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$	$U = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	Standardnormalverteilung (Tabelle 2) $c = u_p$ mit $p = \frac{1+\gamma}{2}$	$\Theta_u = \bar{X} - c \cdot \frac{\sigma}{\sqrt{n}}$ $\Theta_o = \bar{X} + c \cdot \frac{\sigma}{\sqrt{n}}$
2	Normalverteilung (Varianz σ^2 unbekannt und $n \leq 30$; sonst Fall 1 mit s als Schätzwert für σ)	μ	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	t-Verteilung (Tabelle 4) mit $f = n-1$ $c = t(p; f)$ mit $p = \frac{1+\gamma}{2}$	$\Theta_u = \bar{X} - c \cdot \frac{S}{\sqrt{n}}$ $\Theta_o = \bar{X} + c \cdot \frac{S}{\sqrt{n}}$
3	Normalverteilung	σ^2	$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^n X_i$ $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$	$Z = (n-1) \frac{S^2}{\sigma^2}$	Chi-Quadrat-Verteilung (Tabelle 3) mit $f = n-1$ $c_1 = z(p_1; f)$ mit $p_1 = \frac{1-\gamma}{2}$ $c_2 = z(p_2; f)$ mit $p_2 = \frac{1+\gamma}{2}$	$\Theta_u = \frac{(n-1) \cdot S^2}{c_2}$ $\Theta_o = \frac{(n-1) \cdot S^2}{c_1}$
4	Bernoulli-Verteilung Anteilsschätzung (mit $n\hat{p}(1-\hat{p}) > 9$)	p	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = p$ X_i 0/1-wertig mit $P(X_i = 1) = p$	$U = \frac{\bar{X} - p}{\sqrt{p(1-p)/n}}$	Standardnormalverteilung (näherungsweise), Tabelle 2 $c = u_q$ mit $q = \frac{1+\gamma}{2}$	$\Theta_u = \bar{X} - c \cdot \sqrt{\frac{\bar{X} \cdot (1-\bar{X})}{n}}$ $\Theta_o = \bar{X} + c \cdot \sqrt{\frac{\bar{X} \cdot (1-\bar{X})}{n}}$
5	beliebig mit $n > 30$	μ, σ^2	wie im Fall 1 (gegebenenfalls mit s als Schätzwert für σ) bzw. im Fall 3			

Fehlt die Varianz der Grundgesamtheit, nimmt man die Varianz der Stichprobe (falls bekannt) = **Approximation**

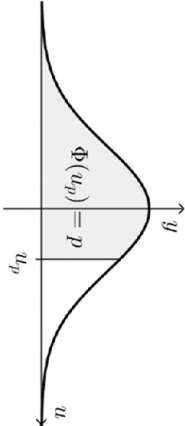
Tabelle 1: CDF $\Phi(u)$ der Standardnormalverteilung



$$P(U \leq u) = \Phi(u)$$
$$P(U \geq u) = 1 - \Phi(u)$$
$$P(-u \leq U \leq u) = 2 \cdot \Phi(u) - 1$$
$$\Phi(-u) = 1 - \Phi(u)$$

<i>u</i>	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9993	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabelle 2: Quantile der Standardnormalverteilung

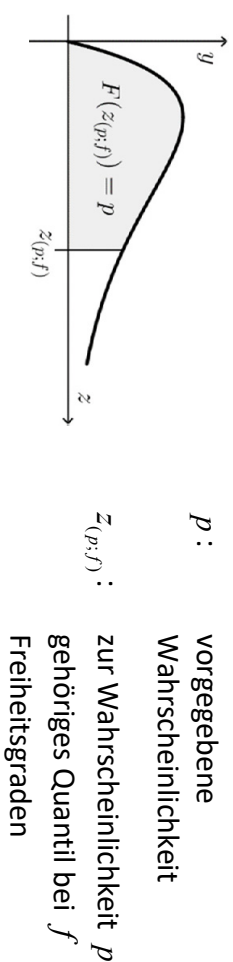


p : vorgegebene
Wahrscheinlichkeit

u_p : zur Wahrscheinlichkeit *p*
gehöriges Quantil

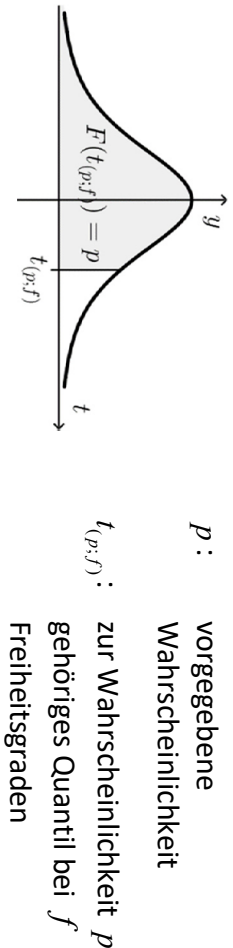
<i>p</i>	<i>u_p</i>	<i>p</i>	<i>u_p</i>
0.90	1.282	0.10	−1.282
0.95	1.645	0.05	−1.645
0.975	1.960	0.025	−1.960
0.99	2.326	0.01	−2.326
0.995	2.576	0.005	−2.576
0.999	3.090	0.001	−3.090

Tabelle 3: Quantile der Chi-Quadrat-Verteilung



	p									
f	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995
1	0.000	0.000	0.001	0.004	0.016	2.71	3.84	5.02	6.63	7.88
2	0.010	0.020	0.051	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
22	8.6	9.5	11.0	12.3	14.0	30.8	33.9	36.8	40.3	42.8
24	9.9	10.9	12.4	13.8	15.7	33.2	36.4	39.4	43.0	45.6
26	11.2	12.2	13.8	15.4	17.3	35.6	38.9	41.9	45.6	48.3
28	12.5	13.6	15.3	16.9	18.9	37.9	41.3	44.5	48.3	51.0
30	13.8	15.0	16.8	18.5	20.6	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	55.3	85.5	90.5	95.0	100.4	104.2
80	51.2	53.5	57.2	60.4	64.3	96.6	101.9	106.6	112.3	116.3
90	59.2	61.8	65.6	69.1	73.3	107.6	113.1	118.1	124.1	128.3
100	67.3	70.1	74.2	77.9	82.4	118.5	124.3	129.6	135.8	140.2

Tabelle 4: Quantile der t-Verteilung von «Student»



$t_{(1-p;f)} = -t_{(p;f)}$

	p				
f	0.9	0.95	0.975	0.99	0.995
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
22	1.321	1.717	2.074	2.508	2.819
24	1.318	1.711	2.064	2.492	2.797
26	1.315	1.706	2.056	2.479	2.779
28	1.313	1.701	2.048	2.467	2.763
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
100	1.290	1.660	1.984	2.364	2.626
200	1.286	1.653	1.972	2.345	2.601
500	1.283	1.648	1.965	2.334	2.586
∞	1.282	1.645	1.960	2.326	2.576