Beschreibende Statistik

Absolute Häufigkeiten

$$H = \sum_{1}^{n} h_i$$

Relative Häufigkeiten

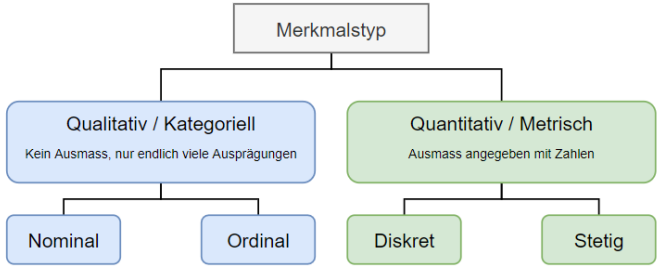
$$F = \sum_{1}^{m} f_i, \qquad F(x) = \frac{H(x)}{n}$$

Kennwerte (Lagemasse)

- Quantil $\quad i = \lceil n \cdot q \rceil,\ Q = x_i = x_{\lceil n \cdot q \rceil}$
- Interquartilsabstand $\quad IQR = Q_3 - Q_1$
- Modus $\quad x_{mod}$ = Häufigste Wert

Begriffe

- Ω = Grundgesamtheit
- n = Anzahl Objekte
- X = Stichprobenwerte
- a = Ausprägungen
- h = Absolute Häufigkeit
- f = Relative Häufigkeit
- H = Kumulative Absolute Häufigkeit
- F = Kumulative Relative Häufigkeit



Arithmetisches Mittel

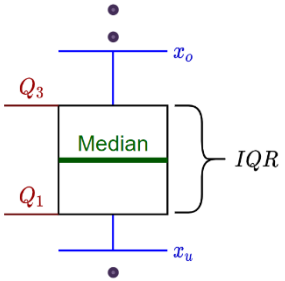
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} a_i \cdot f_i$$

Median

$$\begin{cases} x_{\lceil \frac{n+1}{2} \rceil} & n\ ungerade \\ 0.5 \cdot \left(x_{\lceil \frac{n}{2} \rceil} + x_{\lceil \frac{n}{2}+1 \rceil} \right) & n\ gerade \end{cases}$$

Boxplot

- $Q_1, Q_2 = x_{med}, Q_3$
- $IQR = Q_3 - Q_1$
- *Untere Antenne* x_u: $\quad u = \min[Q_1 - 1.5 \cdot IQR, Q_1]$
- *Obere Antenne* x_o: $\quad o = \max[Q_3 + 1.5 \cdot IQR, Q_3]$
- *Ausreisser:* $\quad x_i < x_u \vee x_i > x_o$



Stichprobenvarianz s^2 (Streumasse)

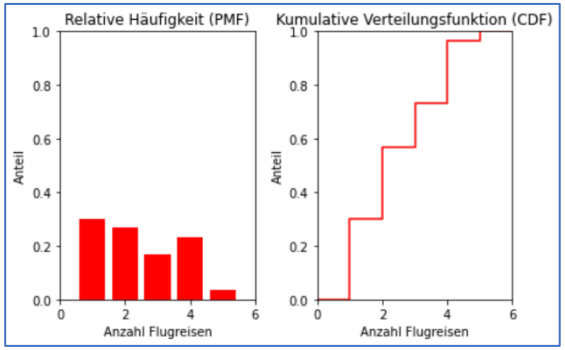
$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \qquad (s_{kor})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$(s_{kor})^2 = \frac{n}{n-1} \cdot s^2$$

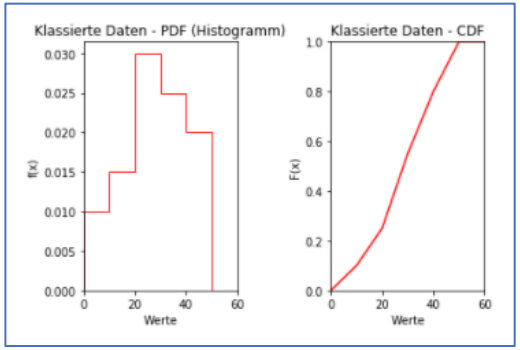
Standardabweichung s (Streumasse)

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

PMF + CDF



PDF + CDF



Pascal Isliker

Nicht klassierte Daten (*PMF* und CDF)

Die *absolute Häufigkeit* kann als Funktion $h: \mathbb{R} \to \mathbb{R}$ bezeichnet werden.

$$h_i$$

Die *relative Häufigkeit* kann als Funktion $f: \mathbb{R} \to \mathbb{R}$ bezeichnet werden.

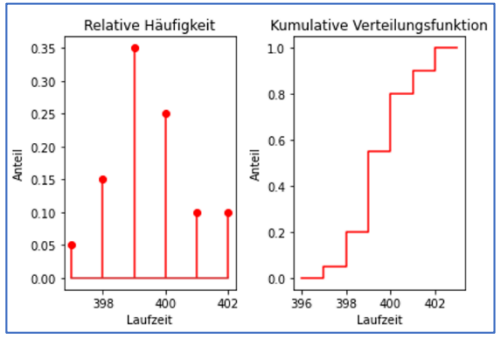
$$f_i = \frac{h_i}{n}$$

Klassenbildung (Faustregeln)

- Die Klassen sollten gleich breit gewählt werden
- Die Anzahl der Klassen sollte zwischen 5 und 20 liegen, jedoch $\sqrt{n}$ nicht überschreiben.

Beispiel:

a_i	397	398	399	400	Total
h_i	1	3	7	5	16
f_i	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{7}{16}$	$\frac{5}{16}$	1
H_i	1	4	11	16	
F_i	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{11}{16}$	$\frac{16}{16}$	



Klassierte Daten (*PDF* und CDF)

Die *absolute Häufigkeitsdichtefunktion* erhält man, indem der Wert der absoluten Häufigkeit h_i, durch die Klassenbreite (Säulenbreite) d_i geteilt wird.

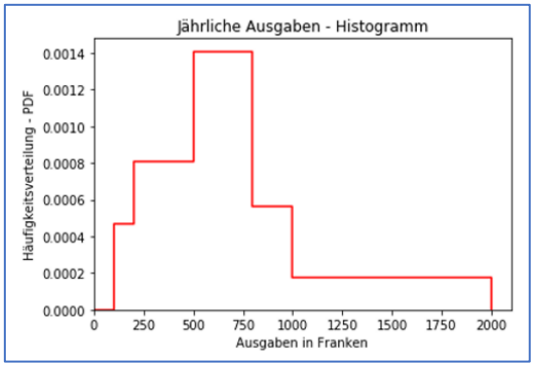
$$h(x) = \frac{h_i}{d_i}$$

Die *relative Häufigkeitsdichtefunktion* (PDF) $f: \mathbb{R} \to [0,1]$ erhält man aus der absoluten Häufigkeitsdichtefunktion, indem man den Wert durch die Stichprobengrösse n teilt.

$$PDF = f(x) = \frac{h(x)}{n}$$

Beispiel:

Klassen	$100-200$	$200-500$	$500-800$	$800-1000$	Total
h_i	35	182	317	84	618
f_i	$\frac{35}{618}$	$\frac{182}{618}$	$\frac{317}{618}$	$\frac{84}{618}$	$Area = 1$
d_i	100	300	300	200	
$h(x)$	$\frac{35}{100}$	$\frac{182}{300}$	$\frac{317}{300}$	$\frac{84}{200}$	
$f(x)$	$\frac{35}{100 \cdot 618}$	$\frac{182}{300 \cdot 618}$	$\frac{317}{300 \cdot 618}$	$\frac{84}{200 \cdot 618}$	



Pascal Isliker

Deskriptive Statistik

Varianz s_x^2, s_y^2

$$(s_x)^2 = \overline{x^2} - \bar{x}^2, \qquad (s_y)^2 = \overline{y^2} - \bar{y}^2$$

Kovarianz s_{xy}

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$$

Abkürzungen

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

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i$$

Varianz (Ränge) $(s_{rg(x)})^2, (s_{rg(y)})^2$

$$(s_{rg(x)})^2 = \overline{rg(x)^2} - (\overline{rg(x)})^2, \qquad (s_{rg(y)})^2 = \overline{rg(y)^2} - (\overline{rg(y)})^2$$

Kovarianz (Ränge) $s_{rg(xy)}$

$$s_{rg(xy)} = \overline{rg(xy)} - \overline{rg(x)} \cdot \overline{rg(y)} = \overline{rg(xy)} - \frac{(n+1)^2}{4}$$

Der **Korrelationskoeffizient (Pearson)** r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}}$$

Ist der Korrelationskoeffizient r_{xy}

- $r_{xy} \approx 1 \rightarrow$ starker positiver linearer Zusammenhang
- $r_{xy} \approx -1 \rightarrow$ starker negativer linearer Zusammenhang
- $r_{xy} \approx 0 \rightarrow$ Keine lineare Korrelation

<u>Bemerkungen</u>

Auch wenn zwischen zwei Grössen eine Korrelation besteht, so muss das noch lange nicht einen *kausalen Zusammenhang* bedeuten. Man spricht von *Scheinkorrelation*.

Graphische Darstellung

- Form — linear / gekrümmt
- Richtung — positiver / negativer Zusammenhang
- Stärke — starke / schwache Streuung

Korrelationskoeffizient (Spearman) r_{sp}

$$r_{sp} = \frac{s_{rg(xy)}}{s_{rg(x)} \cdot s_{rg(y)}} = \frac{\overline{rg(xy)} - \overline{rg(x)} \cdot \overline{rg(y)}}{\sqrt{\overline{rg(x)^2} - (\overline{rg(x)})^2} \cdot \sqrt{\overline{rg(y)^2} - (\overline{rg(y)})^2}}$$

Vereinfachte Formel, sofern *alle Ränge unterschiedlich* sind

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

Ränge

Der *Rang* $rg(x_i)$ des Stichprobenwertes x_i ist definiert als der Index von x_i in der nach der Grösse geordneten Stichprobe.

i	1	2	3	4	5	6
x_i	23	27	35	35	42	59
$rg(x_i)$	1	2	**3.5**	**3.5**	5	6

Bivariate Daten (Merkmale)

- 2x kategoriell — Kontingenztabelle + Mosaikplot
- 1x kategoriell + 1x metrisch — Boxplot oder Striptchart
- 2x metrisch — Streudiagramm

Pascal Isliker

Kombinatorik

Fakultät

$$n! = 1 \cdot 2 \cdot \ldots \cdot n = \prod_{k=1}^{n} k$$

Binomialkoeffizient

Wie viele Möglichkeiten gibt es k Objekte aus einer Gesamtheit von n Objekten auszuwählen.

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

Systematik

- k Anzahl Stellen
- n Anzahl Optionen pro Stelle

Variation (mit Reihenfolge)		Kombination (ohne Reihenfolge)	
Mit Wiederholung	Ohne Wiederholung	Mit Wiederholung	Ohne Wiederholung
n^k	$\frac{n!}{(n-k)!}$	$\binom{n+k-1}{k}$	$\binom{n}{k}$
Zahlenschloss	Schwimmwettkampf	Zahnarzt	Lotto

Variation mit Wiederholung (Zahlenschloss)

Wie viele Möglichkeiten gibt es bei einem Zahlenschloss (0 – 9) mit 6 Zahlenkränzen?

$$n = 10, \quad k = 6$$

$$n^k = 10^6$$

Variation ohne Wiederholung (Schimmwettkampf)

Bei einem Schwimmwettkampf starten 10 Teilnehmer. Wie viele mögliche Platzierungen der ersten drei Plätze (Podest) gibt es?

$$n = 10, \quad k = 3$$

$$\frac{n!}{(n-k)!} = \frac{10!}{(10-3)!} = \frac{10!}{(7)!}$$

Kombination mit Wiederholung (Zahnarzt)

3 Spielzeuge werden aus 5 Töpfen gezogen. Jeder Topf ist mit einer (unterschiedlichen) Art von Spielzeug befüllt.

Wie viele Möglichkeiten hat das Kind?

$$n = 5, \quad k = 3$$

$$\binom{n+k-1}{k} = \binom{5+3-1}{3} = \binom{7}{3}$$

Kombination ohne Wiederholung (Lotto)

Wie gross sind die Chancen beim Lotto 6 aus 49 Zahlen richtig zu ziehen?

Jede Zahl ist nur einmal vorhanden und die Zahlen werden nicht zurückgelegt. Die Reihenfolge in der gezogen wird spielt keine Rolle.

$$n = 49, \quad k = 6$$

$$\binom{n}{k} = \binom{49}{6}$$

Pascal Isliker

Ideen

- Berechnung durch Aufteilung in mehrere Kombinationen
- Berechnung über Inverse
- Prozente = Wahrscheinlichkeit / Gesamt-Wahrscheinlichkeit

Wahrscheinlichkeit, dass ein Ereignis *x*-Mal Auftritt

Beim Rommé spielt man mit *110 Karten*: *sechs* davon sind *Joker*. Zu Beginn eines Spiels erhält jeder Spieler genau *12 Karten*.

In wieviel Prozent aller möglichen Fälle sind darunter ***genau zwei*** *Joker*?

$$\frac{\binom{6}{2} \cdot \binom{104}{10}}{\binom{110}{12}}$$

In wieviel Prozent aller möglichen Fälle ist darunter ***mindestens ein*** *Joker*?

$$1 - \frac{\binom{104}{12}}{\binom{110}{12}}$$

Von *100 Glühbirnen* sind genau *drei defekt*. Es werden nun *6 Glühbirnen* zufällig ausgewählt.

Wie viele Möglichkeiten gibt es, wenn sich ***mindestens eine*** *defekte* Glühbirne in der Auswahl befinden soll?

$$\binom{100}{6} - \binom{97}{6} = 203'880'032$$

Mit wie viel Prozent Chancen ist bei einer Auswahl von 6 Glühbirnen ***keine*** *defekt*?

$$\frac{\binom{97}{6}}{\binom{100}{6}}$$

Sind in mehr als 60% aller Fälle von vier (nicht gleichaltrigen) Geschwistern mindestens zwei im gleichen Monat geboren?

$$1 - \frac{12 \cdot 11 \cdot 10 \cdot 9}{12^4}$$

Auf wie viele Arten lassen sich 10 Bücher in ein Regal reihen?

$$n = 10, \qquad k = 10$$

$$\frac{n!}{(n-k)!} = 10!$$

Aufteilung in mehrere Kombinationen

Wie viele Worte lassen sich aus den Buchstaben des Wortes ABRAKADABRA bilden? (Nur Worte in denen alle Buchstaben vorkommen!)

$$A = 5x, \qquad B = 2x, \qquad R = 2x, \qquad D = 1x, \qquad K = 1x$$

$$\binom{11}{5} \cdot \binom{6}{2} \cdot \binom{4}{2} \cdot \binom{2}{1} \cdot \binom{1}{1} = 83160$$

Pascal Isliker

Elementare Wahrscheinlichkeitsrechnung

Ergebnisraum Ω: Menge aller möglichen Ergebnisse des *Zufallsexperiments*. *Zähldichte* $\rho: \Omega \to [0,1]$, die jedem Ereignis seine Wahrscheinlichkeit zuordnet.

Für jedes Ereignis aus Ω gleichwahrschenlich ist, wird (Ω, P) *Laplace-Raum* genannt.

$$P(M) = \frac{|M|}{|\Omega|}$$

Zwei Ereignisse A und B heissen ***stochastisch unabhängig***, falls

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

Zwei Zufallsvariablen $X: \Omega \to \mathbb{R}$ und $Y: \Omega \to \mathbb{R}$ heissen ***stochastisch abhängig***, falls

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y), \qquad für\ alle\ x, y \in \mathbb{R}$$

Für ***stochastisch unabhängige*** Zufallsvariablen X und Y gilt

$$E(X \cdot Y) = E(X) \cdot E(Y), \qquad V(X + Y) = V(X) + V(Y)$$

Kenngrössen (Varianz und Erwartungswert)

$$E(X + Y) = E(X) + E(Y), \qquad E(\alpha X) = \alpha E(X)$$

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

$$V(\alpha X + \beta) = \alpha^2 \cdot V(X), \qquad S(X) = \sqrt{V(X)}$$

Wahrscheinlichkeit eines *Ereignisses* B mit *Vorbedingung* A

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

Multiplikationssatz

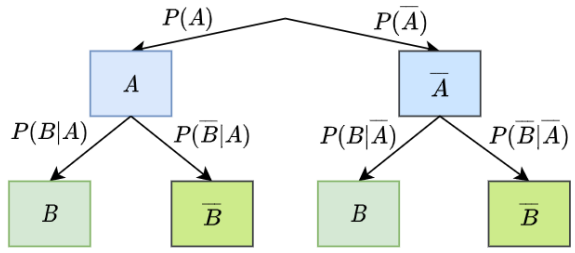
$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

Satz von der *Totalen Wahrscheinlichkeit*

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

Satz von *Bayes*

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



	A	$\bar{A}$	Σ
B	$P(A \cap B)$	$P(\bar{A} \cap B)$	$P(B)$
$\bar{B}$	$P(A \cap \bar{B})$	$P(\bar{A} \cap \bar{B})$	$P(\bar{B})$
Σ	$P(A)$	$P(\bar{A})$	$P(\Omega)$

Spezielle Verteilung (Summary)

Diskret

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

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

Stetig

$$E(X) = \int_{-\infty}^{\infty} f(x) \cdot x\, dx$$

$$V(X) = \int_{-\infty}^{\infty} f(x) \cdot \big(x - E(X)\big)^2$$

Pascal Isliker

Spezielle Verteilungen

Bernoulliverteilung (Einmaliges zurücklegen)

Bernoulli-Experimente sind Zufallsexperimente mit nur zwei möglichen Ergebnissen. Wir bezeichnen diese Ergebnisse mit 1 und 0.

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

1. $E(X)=E(X^2)=p$
2. $V(X)=p\cdot(1-p)$

Eine *Hypergeometrische* Verteilung kann durch eine *Binomialverteilung* angenähert werden, wenn $n \leq \frac{N}{20}$

$$H(N,M,N) \approx B(n,\frac{M}{N})$$

Approximation durch die *Normalverteilung*

- *Binomialverteilung*: $\mu=np,\ \sigma^2=npq$
- *Poissonverteilung*: $\mu=\lambda,\ \sigma^2=\lambda$

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

Faustregel

Die Approximation (*Binomialverteilung*) kann verwendet werden, wenn $npq>9$

Für grosses n $(n\geq 50)$ und kleiner p $(p\leq 0.1)$ kann *Binomial*- durch die *Poisson*-Verteilung approximiert werden

$$B(n,p)\approx Poi(n\cdot p)$$

Hypergeometrische Verteilung (Ohne zurücklegen)

- $N=$ Objekte gesamthaft
- $M=$ Objekte einer bestimmten Sorte
- $n=$ Stichprobengrösse
- $x=$ Merkmalsträger

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

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

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

Binomialverteilung (Mit zurücklegen)

- $n=$ Anzahl Wiederholungen
- $p=$ Wahrscheinlichkeit für ein Ergebnis 1
- $q=1-p$

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

Schreibweise: $X\sim B(n;p)$

1. $\mu = E(X)=np$
2. $\sigma^2=V(X)=npq$
3. $\sigma = S(X)=\sqrt{npq}$

Poisson Verteilung

- $\lambda=$ Rate

$$P(X=x)=\frac{\lambda^x}{x!}\cdot e^{-\lambda}, \qquad \lambda>0$$

Schreibweise $X\sim Poi(\lambda)$

1. $\mu = E(X)=\lambda$
2. $\sigma^2=V(X)=\lambda$
3. $\sigma = S(X)=\sqrt{\lambda}$

Pascal Isliker

Spezielle Verteilungen

Bei einer *stetigen* Zufallsvariable X lässt sich die Verteilungsfunktion als Integral einer Funktion f darstellen

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

Liegt eine beliebige Normalverteilung $N(\mu, \sigma)$ vor, muss *standardisiert* werden. Statt ursprünglichen Zufallsvariablen X betrachtet man die Zufallsvariable

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

Gauss-Verteilung oder Normalverteilung

Die stetige Zufallsvariable X folgt der *Normalverteilung* mit den Parametern, $\mu, \sigma \in \mathbb{R}, \sigma > 0$, wenn sie folgende Dichtefunktion hat:

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

Schreibweise: $X \sim N(\mu; \sigma)$

Ist $\mu = 0$ und $\sigma = 1$, so spricht man von der *Standardnormalverteilung*. Ihre Dichtefunktion wird einfach mit φ bezeichnet; sie ist gegeben durch

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

Die Verteilungsfunktion der Normalverteilung

Die kumulative Verteilungsfunktion (CDF) von $\varphi_{\mu,\sigma}(x)$ wird mit $\phi_{\mu,\sigma}(x)$ bezeichnet. Sie ist definiert durch

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

Erwartungswert und Varianz der Normalverteilung

Für eine Zufallsvariable $X \sim N(\mu; \sigma)$ gilt

$$E(X) = \mu, \qquad V(X) = \sigma^2$$

Zentraler Grenzwertsatz

Für eine Folge $X_1, X_2, \ldots, X_n$ von Zufallsvariablen definieren wird die *n-te Summe* S_n und das *arithmetische Mittel* $\bar{X}_n$.

Haben alle Zufallsvariablen denselben Erwartungswert $E(X_i) = \mu$ und dieselbe Varianz $V(X_i) = \sigma^2$ so folgt

$$E(S_n) = n \cdot \mu, \qquad V(S_n) = n \cdot \sigma^2, \qquad E(\bar{X}_n) = \mu, \qquad V(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{1}{n^2} \cdot V(S_n)$$

Sind die Zufallsvariablen alle identisch $N(\mu, \sigma)$ verteilt, so sind die Summe S_n und das arithmetische Mittel $\bar{X}_n$ wieder normalverteilt mit…

- S_n: $N\left(n \cdot \mu, \sqrt{n} \cdot \sigma\right)$
- $\bar{X}_n$: $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

Verteilungsfunktion $F_n(u)$ der dazugehörigen standardisierten Zufallsvariable

$$U_n = \frac{\left((X_1 + X_2 + \cdots + X_n) - n\mu\right)}{\sqrt{n} \cdot \sigma} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Konvergiert für $n \to \infty$ gegen die Verteilungsfunktion $\phi(u)$ der Standardnormalverteilung:

$$\lim_{n\to\infty} F_n(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{u} e^{-\frac{1}{2}t^2} dt$$

Pascal Isliker

Methode der kleinsten Quadrate

Lineare Regression

Gegeben sind Datenpunkte $(x_i; y_i)$ mit $1 \leq i \leq n$. Die *Residuen / Fehler* $\epsilon_i = g(x_i) - y_i$ dieser Datenpunkte sind Abstände in y −Richtung zwischen y_i und der Geraden g. Die *Ausgleichs-* oder *Regressiongerade*, sei diejenige Gerade für die, die Summe der quadrierten Residuen $\sum_{i=1}^{n} \epsilon_i^2$ am kleinsten ist.

Regressionsgerade

Die *Regressionsgerade* $g(x) = mx + d$ mit den Parametern m und d ist die Gerade, für welche die *Residualvarianz* s_ϵ^2 minimal ist.

$$Steigung: m = \frac{s_{xy}}{s_x^2}, \qquad y - Achsenabschnitt: d = \bar{y} - m\bar{x}, \qquad s_\epsilon^2 = s_y^2 - \frac{s_{xy}^2}{s_x^2}$$

Bestimmtheitsmass

Die Totale Varianz setzt sich zusammen aus der Residualvarianz und der Varianz der prognostizierten Werte

- s_y^2 Totale Varianz
- $s_{\hat{y}}^2$ prognostizierte (erklärte) Varianz
- s_ϵ^2 Residualvarianz

$$s_y^2 = s_\epsilon^2 + s_{\hat{y}}^2$$

Das *Bestimmheitsmass* R^2 beurteilt die globale Anpassungsgüte einer Regression über den Anteil der prognostizierten Varianz $s_{\hat{y}}^2$ an der totalen Varianz s_y^2

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

Das *Bestimmheitsmass* R^2 entspricht dem Quadrat des *Korrelationskoeffizienten*

$$R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} = (r_{xy})^2$$

Residuenquadrate

- $y_i: beobachtete\ y - Werte$
- $\hat{y}_i: prognostizierte\ bzw.\ erklärte\ y - Werte$
- $\epsilon: Residuen\ (oder\ auch\ Fehler)$

$$\sum_{i=1}^{n} \left(\underbrace{y_i - g(x_i)}_{\epsilon_i} \right)^2 = \sum_{i=1}^{n} \left(\underbrace{y_i - \hat{y}_i}_{\epsilon_i} \right)^2$$

Kleinste Quadrate (KQM)

Die Parameter m und q werden mit der Matrix A berechnet

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \qquad A^T \cdot A \cdot \begin{pmatrix} m \\ q \end{pmatrix} = A^T \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Linearisierungsfunktionen

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

Pascal Isliker

Schliessende Statistik

Erwartungstreue Schätzfunktion

Eine Schätzfunktion Θ eines Parameters θ heisst *erwartungstreu*, wenn

$$E(\Theta) = \theta$$

Effizienz Schätzfunktion

Gegeben sind zwei *erwartungstreue* Schätzfunktionen Θ_1 und Θ_2 desselben Parameters θ. Man nennt Θ_1 *effizienter als* Θ_2, falls

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

Konsistenz Schätzfunktion

Eine Schätzfunktion Θ heisst *konsistent*, wenn

$$E(\Theta) \to \theta \text{ und } V(\Theta) \to 0 \text{ für } n \to \infty$$

Grundgesamtheit mit Erwartungswert μ, Varianz σ^2 und Zufallsstichprobe X_1, X_2, X_3. Die folgende Schätzfunktion ist gegeben.

$$\Theta_1 = \frac{1}{3} \cdot (2X_1 + X_2)$$

Ist diese Schätzfunktion *erwartungstreu* (Parameter: μ)?

$$E(\Theta_1) = E\left(\frac{1}{3} \cdot (2X_1 + X_2)\right) = \frac{1}{3} \cdot \big(2E(X_1) + E(X_2)\big)$$

$$E(\Theta_1) = \frac{1}{3} \cdot (2\mu + \mu) = \frac{3\mu}{3} = \mu$$

Da $E(\Theta_1) = \mu$ ist die Funktion erwartungstreu.

Berechne die *Effizienz* der Schätzfunktion (Parameter: σ^2):

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

$$V(\Theta_1) == \frac{1}{9} \cdot \big(4 \cdot V(X_1) + V(X_2)\big) = \frac{1}{9} \cdot (4\sigma^2 + \sigma^2) = \frac{5\sigma^2}{9}$$

Likelyhood-Funktion

Wir betrachten eine Zufallsvariable X und ihre Dichte (PDF)

$$f_x(x|\theta)$$

Welche von x und einem oder mehreren Parametern θ abhängig sind. Für eine Stichprobe vom Umfang n mit $x_1, \dots, x_n$ nennen wir die vom Parameter θ abhängige Funktion … die Likelyhood-Funktion der Stichprobe.

$$L(\theta) = f_x(x_1|\theta) \cdot f_x(x_2|\theta) \cdot \dots \cdot f_x(x_n|\theta)$$

Vorgehen – Likelyhood Funktion

1. Likelyhood-Funktion bestimmen
2. Maximalstelle der Funktion bestimmen
   - (Partielle) Ableitung $L'(\theta) = 0$

Erwartungswert (Funktion, Wert)

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i\,, \qquad \hat{\mu} = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Varianz (Funktion, Wert)

$$S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2\,, \qquad \hat{\sigma}^2 = s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Pascal Isliker

Vertrauensintervalle

Wir legen eine grosse Wahrscheinlichkeit γ fest (z.B. $\gamma = 95\%$). γ heisst *statistische Sicherheit* oder *Vertrauensniveau*. $\alpha = 1 - \gamma$ ist die sogenannte *Irrtumswahrscheinlichkeit*.

Dann bestimmen wir zwei Zufallsvariablen Θ_u und Θ_o so, dass sie den wahren Parameterwert Θ mit der Wahrscheinlichkeit γ einschliessen:

$$P(\Theta_u \leq \Theta \leq \Theta_u) = \gamma$$

Spezialfall: Anteilswert p einer Bernoulli-Verteilung (Funktion, Wert)

$$\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i\,, \qquad \hat{p} = \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Intervallschätzung

Geben Sie das Vertrauensintervall für μ an (σ^2 unbekannt). Gegeben sind…

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

1. Verteilungstyp bestimmen

   Verteilungstyp mit Param μ und σ^2 unbekannt → *T-Verteilung*

2. Verteilung und Quantile berechnen

$$f = n - 1 = 9, \qquad p = \frac{1+\gamma}{2} = 0.995, \qquad c = t_{(p;f)} = t_{(0.995;9)} = 3.25$$

3. Vertrauensintervall bestimmen

$$e = c \cdot \frac{S}{\sqrt{n}} = 4.111, \qquad \Theta_u = \bar{X} - e = 97.89, \qquad \Theta_o = \bar{X} + e = 106.11$$

	Verteilung der Grundgesamtheit	Param	Schätzfunktionen	Standardisierte Zufallsvariable	Verteilung / Quantile	Intervallgrenzen
1	Normalverteilung (Varianz σ^2 bekannt)	μ	$\bar{X}$	$U = \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}$	*Standardnormalverteilung* $c = u_p,\ p = \frac{1+\gamma}{2}$	$e = c \cdot \frac{\sigma}{\sqrt{n}}$ $\Theta_u = \bar{X} - e, \quad \Theta_o = \bar{X} + e$
2	Normalverteilung (Varianz σ^2 unbekannt und $n \leq 30$; sonst Fall 1 mit s als Schätzwert für σ)	μ	$\bar{X}, \quad S^2$	$T = \frac{(\bar{X}-\mu)}{S/\sqrt{n}}$	*t-Verteilung* $c = t_{(p;f=n-1)},\ p = \frac{1+\gamma}{2}$	$e = c \cdot \frac{S}{\sqrt{n}}$ $\Theta_u = \bar{X} - e, \quad \Theta_o = \bar{X} + e$
3	Normalverteilung	σ^2	$\bar{X}, \quad S^2$	$Z = (n-1)\frac{s^2}{\sigma^2}$	*Chi-Quadrat-Verteilung* $c_1 = z_{(p_1;f=n-1)},\ p_1 = \frac{1-\gamma}{2}$ $c_2 = z_{(p_2;f=n-1)},\ p_2 = \frac{1+\gamma}{2}$	$e = (n-1) \cdot S^2$ $\Theta_u = \frac{e}{c_2}, \quad \Theta_o = \frac{e}{c_1}$
4	Bernoulli-Verteilung mit $n\hat{p}(1-p) > 9$	p	$\bar{X}$ $P(X_i = 1) = p$	$U = \frac{\bar{X}-p}{\sqrt{p(1-p)/n}}$	*Standardnormalverteilung* näherungsweise $c = u_q,\ q = \frac{1+\gamma}{2}$	$e = c \cdot \sqrt{\frac{\bar{X} \cdot (1-\bar{X})}{n}}$ $\Theta_u = \bar{X} - e, \quad \Theta_o = \bar{X} + e$
5	Beliebig mit $n > 30$	μ, σ^2	Wie im Fall 1 (gegebenenfalls mit s als Schätzwert für σ) bzw. wie im Fall 3			

Pascal Isliker