Stochastik und Statistik

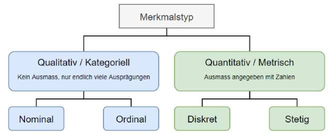
Jil Zerndt, Lucien Perret
December 2024

Intro

Begriffe

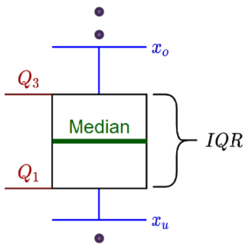
Grundlegende Begriffe

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



Boxplot

- $Q_1, Q_2 = x_{\text{med}}, Q_3$
- $IQR = Q_3 - Q_1$
- Untere Antenne x_u :
  $u = \min\left[Q_1 - 1.5 \cdot IQR, Q_1\right]$
- Obere Antenne x_0 :
  $o = \max\left[Q_3 + 1.5 \cdot IQR, Q_3\right]$
- Ausreisser: $x_i < x_u \vee x_i > x_0$



Deskriptive Statistik

Bivariate Daten (Merkmale)

- 2x kategoriell → Kontingenztabelle + Mosaikplot
- 1x kategoriell + 1x metrisch → Boxplot oder Stripchart
- 2x metrisch → Streudiagramm

Absolute Häufigkeiten

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

H: Absolute Häufigkeit,
h_i: Einzelhäufigkeit der i-ten Beobachtung,
n: Anzahl der Beobachtungen.

Relative Häufigkeiten

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

F: Relative Häufigkeit,
f_i: Einzelrelative Häufigkeit der i-ten Beobachtung,
$H(x)$: Absolute Häufigkeit eines Wertes x,
n: Anzahl der Beobachtungen.

Kennwerte (Lagemasse)

Quantil

$$i = \lceil n \cdot q \rceil, \quad Q = x_i = x_{\lceil n \cdot q \rceil}$$

i: Position des Quantils,
n: Anzahl der Beobachtungen,
q: Quantilswert (z. B. 0.25 für das erste Quartil),
x_i: Beobachtung an Position i.

Interquartilsabstand

$$IQR = Q_3 - Q_1$$

IQR: Interquartilsabstand,
Q_3: Oberes Quartil (75. Perzentil),
Q_1: Unteres Quartil (25. Perzentil).

Modus

$$x_{\text{mod}} = \text{Häufigste Wert}$$

Arithmetisches Mittel

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

$\bar{x}$: Arithmetisches Mittel,
n: Anzahl der Beobachtungen,
x_i: Einzelbeobachtung,
a_i: Klassenmitte,
f_i: Relative Häufigkeit der Klasse i.

Median

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

n: Anzahl der Beobachtungen,
$x_{[k]}$: Beobachtung an der k-ten Position.

Stichprobenvarianz s^2 (Streumasse)

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

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

s^2: Stichprobenvarianz,
s^2_{kor}: Korrigierte Stichprobenvarianz,
x_i: Einzelbeobachtung,
$\bar{x}$: Arithmetisches Mittel,
n: Anzahl der Beobachtungen.

Standardabweichung s (Streumasse)

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

s: Standardabweichung,
s_{kor}: Korrigierte Standardabweichung,
x_i: Einzelbeobachtung,
$\bar{x}$: Arithmetisches Mittel,
n: Anzahl der Beobachtungen.

PDF + CDF

Nicht klassierte Daten (PMF und CDF)

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

$$h_i$$

h_i: Absolute Häufigkeit der i-ten Beobachtung.

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

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

f_i: Relative Häufigkeit der i-ten Beobachtung,
h_i: Absolute Häufigkeit der i-ten Beobachtung,
n: Anzahl der Beobachtungen.

PMF und CDF für diskrete und stetige Daten

Diskrete Verteilungsfunktionen

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:

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

Diskrete Häufigkeitsverteilung

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	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{11}{16}$	$\frac{16}{16}$	

Klassenbildung (Faustregeln)

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

Stetige Verteilungsfunktionen

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:

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

Stetige Häufigkeitsverteilung

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	
$h(x)$	$\frac{35}{100}$	$\frac{182}{300}$	$\frac{317}{300}$	$\frac{84}{200}$	
$f(x)$	$\frac{35}{100 \cdot 618}$	$\frac{182}{300 \cdot 618}$	$\frac{317}{300 \cdot 618}$	$\frac{84}{200 \cdot 618}$	

Varianz und Kovarianz

Varianz s_x^2, s_y^2:

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

Kovarianz s_{xy}:

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

Abkürzungen

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

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

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

Rang-Varianz und Kovarianz

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

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

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}:

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

Bemerkungen

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:

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

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

Kombinatorik

Fakultät

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

n = Die positive ganze Zahl, für die die Fakultät berechnet wird
k = Laufvariable in der Produktnotation
$\prod$ = Produkt aller Terme von $k = 1$ bis n

Binomialkoeffizient

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

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

n = Gesamtanzahl der Objekte in der Menge
k = Anzahl der auszuwählenden Objekte
$n!$ = Fakultät von n
$(n-k)!$ = Fakultät von $(n-k)$
$k!$ = Fakultät von k

Systematik

Grundbegriffe

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.

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

Wahrscheinlichkeitsrechnung

Ideen

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

Wahrscheinlichkeit bei Rommé
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?

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

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

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

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

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

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

Glühbirnen auswählen
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**?

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

Buchstabenkombinationen
Wie viele Worte lassen sich aus den Buchstaben des Wortes ABRAKADABRA bilden? (Nur Worte in denen alle Buchstaben vorkommen!)
$A = 5x, \quad B = 2x, \quad R = 2x, \quad D = 1x, \quad K = 1x$

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

Wahrscheinlichkeitstheorie

Ergebnisraum

Ergebnisraum Ω ist die Menge aller möglichen Ergebnisse des Zufallsexperiments. Zähldichte $\rho : \Omega \to [0,1]$ ordnet jedem Ereignis seine Wahrscheinlichkeit zu.
Für ein Laplace-Raum (Ω, P) gilt:

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

Ω = Ergebnisraum (Menge aller möglichen Ergebnisse)
$P(M)$ = Wahrscheinlichkeit des Ereignisses M
$|M|$ = Anzahl der für M günstigen Ergebnisse
$|\Omega|$ = Anzahl aller möglichen Ergebnisse

Stochastische Unabhängigkeit

Zwei Ereignisse A und B heissen stochastisch unabhängig, falls:

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

$P(A \cap B)$ = Wahrscheinlichkeit dass beide Ereignisse eintreten
$P(A)$ = Wahrscheinlichkeit von Ereignis A
$P(B)$ = Wahrscheinlichkeit von Ereignis B

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

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

$P(X = x, Y = y)$ = Wahrscheinlichkeit dass X den Wert x und Y den Wert y annimmt
$P(X = x)$ = Wahrscheinlichkeit dass X den Wert x annimmt
$P(Y = y)$ = Wahrscheinlichkeit dass Y den Wert y annimmt

Bedingte Wahrscheinlichkeit

Bedingte Wahrscheinlichkeit

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

$P(B|A)$ = Wahrscheinlichkeit von B unter der Bedingung dass A eingetreten ist
$P(B \cap A)$ = Wahrscheinlichkeit dass beide Ereignisse eintreten
$P(A)$ = Wahrscheinlichkeit von Ereignis A

Multiplikationssatz

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

$P(A \cap B)$ = Wahrscheinlichkeit dass beide Ereignisse eintreten
$P(A)$ = Wahrscheinlichkeit von Ereignis A
$P(B|A)$ = Wahrscheinlichkeit von B unter der Bedingung dass A eingetreten ist
$P(A|B)$ = Wahrscheinlichkeit von A unter der Bedingung dass B eingetreten ist

Kenngrössen (Varianz und Erwartungswert)

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

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

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

$E(X)$ = Erwartungswert der Zufallsvariable X
$V(X)$ = Varianz der Zufallsvariable X
$S(X)$ = Standardabweichung der Zufallsvariable X
α, β = Konstanten
$P(X = x)$ = Wahrscheinlichkeit, dass X den Wert x annimmt
$\sum_{x \in \mathbb{R}}$ = Summe über alle möglichen Werte von x in den reellen Zahlen

Verteilungen und Erwartungswerte

Für diskrete Verteilungen:

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

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

$E(X)$ = Erwartungswert der Zufallsvariable X
$V(X)$ = Varianz der Zufallsvariable X
$f(x)$ = Wahrscheinlichkeitsfunktion
x = Mögliche Werte der Zufallsvariable

Für stetige Verteilungen:

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

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

$E(X)$ = Erwartungswert der Zufallsvariable X
$V(X)$ = Varianz der Zufallsvariable X
$f(x)$ = Dichtefunktion
x = Mögliche Werte der Zufallsvariable

Satz von der Totalen Wahrscheinlichkeit

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

$P(B)$ = Wahrscheinlichkeit von Ereignis B
$P(A)$ = Wahrscheinlichkeit von Ereignis A
$P(\bar{A})$ = Wahrscheinlichkeit des Gegenereignisses von A
$P(B|A)$ = Wahrscheinlichkeit von B unter der Bedingung dass A eingetreten ist
$P(B|\bar{A})$ = Wahrscheinlichkeit von B unter der Bedingung dass A nicht eingetreten ist

Satz von Bayes

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

$P(A|B)$ = Wahrscheinlichkeit von A unter der Bedingung dass B eingetreten ist
$P(A)$ = Wahrscheinlichkeit von Ereignis A
$P(B|A)$ = Wahrscheinlichkeit von B unter der Bedingung dass A eingetreten ist
$P(B)$ = Wahrscheinlichkeit von Ereignis B

Spezielle Verteilungen

Verteilungen und Erwartungswerte

Für diskrete Verteilungen:

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

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

$E(X)$ = Erwartungswert der Zufallsvariable X
$V(X)$ = Varianz der Zufallsvariable X
$f(x)$ = Wahrscheinlichkeitsfunktion
x = Mögliche Werte der Zufallsvariable

Für stetige Verteilungen:

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

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

$E(X)$ = Erwartungswert der Zufallsvariable X
$V(X)$ = Varianz der Zufallsvariable X
$f(x)$ = Dichtefunktion
x = Mögliche Werte der Zufallsvariable

Bernoulliverteilung

Bernoulli-Experimente sind Zufallsexperimente mit nur zwei möglichen Ergebnissen (1 und 0):

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

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

$E(X)$ = Erwartungswert
$V(X)$ = Varianz
$P(X = 1)$ = Wahrscheinlichkeit für Erfolg
p = Erfolgswahrscheinlichkeit
q = Gegenwahrscheinlichkeit $(1 - p)$

Normalverteilung

Gauss-Verteilung

Die stetige Zufallsvariable X folgt der Normalverteilung mit den Parametern $\mu, \sigma \in \mathbb{R}, \sigma > 0$:

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

Standardnormalverteilung ($\mu = 0$ und $\sigma = 1$):

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

$\varphi_{\mu,\sigma}(x)$ = Dichtefunktion der Normalverteilung
$\varphi(x)$ = Dichtefunktion der Standardnormalverteilung

μ = Erwartungswert
σ = Standardabweichung
e = Eulersche Zahl
π = Kreiszahl Pi

Die Verteilungsfunktion der Normalverteilung

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

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

$\phi_{\mu,\sigma}(x)$ = Verteilungsfunktion der Normalverteilung
$\varphi_{\mu,\sigma}(x)$ = Dichtefunktion der Normalverteilung
$P(X \leq x)$ = Wahrscheinlichkeit dass X kleiner oder gleich x ist
μ = Erwartungswert
σ = Standardabweichung
π = Kreiszahl Pi
e = Eulersche Zahl

Approximation durch die Normalverteilung

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

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

$P(a \leq X \leq b)$ = Wahrscheinlichkeit dass X zwischen a und b liegt
$\phi_{\mu,\sigma}$ = Verteilungsfunktion der Normalverteilung
a, b = Untere und obere Grenze

Standardisierung der Normalverteilung

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:

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

$F(x)$ = Verteilungsfunktion
$P(X \leq x)$ = Wahrscheinlichkeit dass X kleiner oder gleich x ist
$f(u)$ = Dichtefunktion
U = Standardisierte Zufallsvariable
X = Ursprüngliche Zufallsvariable
μ = Erwartungswert
σ = Standardabweichung

Erwartungswert und Varianz der Normalverteilung

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

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

$E(X)$ = Erwartungswert der Zufallsvariable X
$V(X)$ = Varianz der Zufallsvariable X
μ = Erwartungsparameter
σ^2 = Varianzparameter

Zentraler Grenzwertsatz
Für eine Folge von Zufallsvariablen $X_1, X_2, \dots, X_n$ mit gleichem Erwartungswert μ und gleicher Varianz σ^2 gilt:

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

S_n = Summe der Zufallsvariablen
$\bar{X}_n$ = Arithmetisches Mittel der Zufallsvariablen
n = Anzahl der Zufallsvariablen
μ = Erwartungswert der einzelnen Zufallsvariablen
σ^2 = Varianz der einzelnen Zufallsvariablen

Die standardisierte Zufallsvariable:

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

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

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

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

Faustregeln für Approximationen
- Die Approximation (Binomialverteilung) kann verwendet werden, wenn $npq > 9$
- Für grosses $n (n \geq 50)$ und kleines $p (p \leq 0.1)$ kann die Binomial- durch die Poisson-Verteilung approximiert werden:

$$B(n, p) \approx \text{Poi}(n\cdot p)$$

  $B(n, p)$ = Binomialverteilung
  $\text{Poi}(\lambda)$ = Poissonverteilung mit Parameter $\lambda = n \cdot p$

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

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

  $H(N, M, n)$ = Hypergeometrische Verteilung
  $B(n, p)$ = Binomialverteilung
  N = Grundgesamtheit
  M = Anzahl der Erfolge in der Grundgesamtheit
  n = Stichprobengröße

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. $\sigma^2 = V(X) = n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$ 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
- λ = Rate

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

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

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 Regressionsgerade ist diejenige Gerade, für die die Summe der quadrierten Residuen $\sum_{i=1}^{n} \epsilon_i^2$ am kleinsten ist.
(x_i, y_i) = Datenpunkte
ϵ_i = Residuum (Abweichung) des i-ten Datenpunkts
$g(x_i)$ = Wert der Regressionsgerade an der Stelle x_i
n = Anzahl der Datenpunkte

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

m = Steigung der Regressionsgerade
d = y-Achsenabschnitt
s_{xy} = Kovarianz von x und y
s_x^2 = Varianz der x-Werte
s_y^2 = Varianz der y-Werte
$\bar{x}$ = Arithmetisches Mittel der x-Werte
$\bar{y}$ = Arithmetisches Mittel der y-Werte
s_ϵ^2 = Residualvarianz

Bestimmtheitsmass

Varianzaufspaltung
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

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

s_y^2 = Totale Varianz der beobachteten y-Werte
s_ϵ^2 = Varianz der Residuen
$s_{\hat{y}}^2$ = Varianz der durch die Regression geschätzten Werte

Bestimmtheitsmass
Das Bestimmtheitsmass 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:

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

R^2 = Bestimmtheitsmass (zwischen 0 und 1)
$s_{\hat{y}}^2$ = Varianz der prognostizierten Werte
s_y^2 = Totale Varianz

Das Bestimmtheitsmass R^2 entspricht dem Quadrat des Korrelationskoeffizienten:

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

s_{xy} = Kovarianz von x und y
s_x^2 = Varianz der x-Werte
s_y^2 = Varianz der y-Werte
r_{xy} = Korrelationskoeffizient

Linearisierungsfunktionen

Transformationen

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

y = Abhängige Variable
x = Unabhängige Variable
q, m = Parameter der Funktion
e = Eulersche Zahl
$\ln$ = Natürlicher Logarithmus
$\log$ = Logarithmus zur Basis 10

Methode der kleinsten Quadrate

Matrix-Darstellung

Die Parameter m und q der Regressionsgeraden werden mit der Matrix A berechnet:

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

Residuenberechnung

Die Residuen ϵ_i ergeben sich als:

$$\epsilon_i = y_i - \hat{y}_i = y_i - (mx_i + q)$$

Die Summe der quadrierten Residuen wird minimiert:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (mx_i + q))^2 \rightarrow \min$$

Schliessende Statistik

Erwartungstreue Schätzfunktion

Eine Schätzfunktion Θ eines Parameters θ heisst erwartungstreu, wenn:

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

Effizienz einer 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 einer Schätzfunktion

Eine Schätzfunktion Θ heisst konsistent, wenn:

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

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

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

Θ_1 = Schätzfunktion
X_1, X_2 = Zufallsvariablen aus der Stichprobe

Ist diese Schätzfunktion erwartungstreu (Parameter: μ)?

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

$E(\Theta_1)$ = Erwartungswert der Schätzfunktion
$E(X_1), E(X_2)$ = Erwartungswerte der einzelnen Zufallsvariablen
μ = Wahrer Parameterwert

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

Effizienz einer Schätzfunktion
Grundgesamtheit mit Erwartungswert μ, Varianz σ^2 und Zufallsstichprobe X_1, X_2, X_3. Gegeben ist die Schätzfunktion:

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

Berechnung der Effizienz:

$$\begin{aligned} V(\Theta_1) &= V(\frac{1}{3} \cdot (2X_1 + X_2)) \\ &= \frac{1}{9} \cdot V(2X_1 + X_2) \\ &= \frac{1}{9} \cdot (V(2X_1) + V(X_2)) \\ &= \frac{1}{9} \cdot (4V(X_1) + V(X_2)) \\ &= \frac{1}{9} \cdot (4\sigma^2 + \sigma^2) \\ &= \frac{5\sigma^2}{9} \end{aligned}$$

$V(\Theta_1)$ = Varianz der Schätzfunktion
$V(X_1), V(X_2)$ = Varianzen der einzelnen Zufallsvariablen
σ^2 = Varianz der Grundgesamtheit

Die Effizienz der Schätzfunktion ist also $\frac{5\sigma^2}{9}$.

Vertrauensintervalle

Vertrauensintervall

Wir legen eine grosse Wahrscheinlichkeit γ fest (z.B. $\gamma = 95\%$). γ heisst statistische Sicherheit oder Vertrauensniveau. $\alpha = 1 - \gamma$ ist die Irrtumswahrscheinlichkeit.
Dann bestimmen wir zwei Zufallsvariablen Θ_u und Θ_o so, dass sie den wahren Parameterwert Θ mit der Wahrscheinlichkeit γ einschliessen:

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

Intervallschätzung

Verteilungstypen und zugehörige Quantile:

Verteilung	Parameter	Quantile
Normalverteilung (σ^2 bekannt)	μ	$c = u_p, p = \frac{1+\gamma}{2}$
t-Verteilung (σ^2 unbekannt)	μ	$c = t_{(p;f=n-1)}, p = \frac{1+\gamma}{2}$
Chi-Quadrat-Verteilung	σ^2	$c_1 = \chi^2_{(\frac{1-\gamma}{2};n-1)}, c_2 = \chi^2_{(\frac{1+\gamma}{2};n-1)}$

Berechnung eines Vertrauensintervalls
Geben Sie das Vertrauensintervall für μ an (σ^2 unbekannt). Gegeben sind:

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

1. Verteilungstyp mit Param μ und σ^2 unbekannt $\rightarrow$ T-Verteilung
2. $f = n - 1 = 9$, $p = \frac{1+\gamma}{2} = 0.995$, $c = t_{(p;f)} = t_{(0.995;9)} = 3.25$
3. $e = c \cdot \frac{S}{\sqrt{n}} = 4.111$, $\Theta_u = \bar{X} - e = 97.89$, $\Theta_o = \bar{X} + e = 106.11$

Likelyhood-Funktion

Likelyhood-Funktion

Wir betrachten eine Zufallsvariable X und ihre Dichte (PDF) $f_x(x|\theta)$, welche von x und einem oder mehreren Parameten θ abhängig sind.
Für eine Stichprobe vom Umfang n mit $x_1, \ldots, x_n$ nennen wir die vom Parameter θ abhängige Funktion die Likelyhood-Funktion der Stichprobe:

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

Vorgehen bei Maximum-Likelihood-Schätzung

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

Erwartungswert und Varianz (Funktion und Wert)

Erwartungswert:

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

$\bar{X}$ = Arithmetischer Mittelwert (Zufallsvariable)
$\hat{\mu} = \bar{x}$ = Arithmetischer Mittelwert (Stichprobenwert)
n = Stichprobenumfang
X_i = i-te Zufallsvariable
x_i = i-ter Stichprobenwert

Varianz:

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

S^2 = Stichprobenvarianz (Zufallsvariable)
$\hat{\sigma}^2 = s^2$ = Stichprobenvarianz (Stichprobenwert)
$\bar{X}$ = Arithmetischer Mittelwert (Zufallsvariable)
$\bar{x}$ = Arithmetischer Mittelwert (Stichprobenwert)

Detaillierte Vertrauensintervalle

Verteilungstypen und Quantile

Verteilung	Parameter	Standardisierung	Quantile
Normalverteilung (σ^2 bekannt)	μ	$U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$c = u_p, p = \frac{1+\gamma}{2}$
t-Verteilung (σ^2 unbekannt)	μ	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$c = t_{(p;f=n-1)}, p = \frac{1+\gamma}{2}$
Chi-Quadrat	σ^2	$Z = (n-1)\frac{s^2}{\sigma^2}$	$c_1 = \chi^2_{(\frac{1-\gamma}{2};n-1)}$ $c_2 = \chi^2_{(\frac{1+\gamma}{2};n-1)}$

Konfidenzintervalle

Für verschiedene Verteilungen ergeben sich folgende Intervallgrenzen:

1. Normalverteilung (σ^2 bekannt):

$$\Theta_u = \bar{X} - c\frac{\sigma}{\sqrt{n}}, \quad \Theta_o = \bar{X} + c\frac{\sigma}{\sqrt{n}}$$

2. t-Verteilung (σ^2 unbekannt):

$$\Theta_u = \bar{X} - c\frac{S}{\sqrt{n}}, \quad \Theta_o = \bar{X} + c\frac{S}{\sqrt{n}}$$

3. Chi-Quadrat-Verteilung:

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