Statistik

Prof. Dr. Ansgar Steland

EAS 2023

Ziele der Deskriptiven Statistik:

- Empirische Daten durch Tabellen, Grafiken und Kennzahlen übersichtlich darstellen und ordnen.
- Daten durch aussagekräftige Kennzahlen zahlenmäßig zu beschreiben und verdichten.
- Interpretation der aufbereiteten Daten.
- Generierung von Hypothesen.

Hierbei werden keine stochastischen Modelle verwendet, so dass getroffene Ausagen nicht durch Fehlerwahrscheinlichkeiten abgesichert sind. Dies ist Aufgabe der **Schließenden Statistik (Inferenzstatistik)**.

Vorgeschaltet ist die **Planung der statistischen Studie**:

- Was erheben? Wie erheben?
- Worüber sollen Aussagen getroffen? Welche Fragen sind zu beantworten?
- Definition der zu erhebenen Variablen
- Ein- und Ausschlusskriterien
- Sicherstellung der Datenqualität.
- Umgang mit fehlenden Daten.
- Festlegung von Verantwortlichkeiten, Zugriffsrechten.
- Datenspeicherung, Datenschutz.
- Planung der eigentlichen statistischen Analyse: Welche Methoden?

Vollständige Dokumentation.

Grundbegriffe:

Statistische Analyse von Daten:

1. Definition der relevanten **statistischen Einheiten (Untersuchungseinheiten, Merkmalsträger)**
2. Die **Grundgesamtheit** G ist die Menge aller statistischen Einheiten.
3. Erhebe Daten (**Merkmale**, Variablen) an allen (Totalerhebung) oder ausgewählten Einheiten.
4. Werden die Daten durch Experimente gewonnen, dann heißen die $g \in G$ auch **Versuchseinheiten** (experimental units). Werden die Daten durch Beobachtunge gewonnen, so spricht man von **Beobachtungseinheiten** (observational units).
5. Merkmale X nehmen gewissen **Merkmalsausprägungen** M. Formal:

$$X : G \to M, \qquad g \mapsto X(g)$$

(o.E. (durch Kodieren) $M \subset \mathbb{R}$)

6. Zufallsauswahl: Ziehe n Mal aus der 'Urne' G mit Zurücklegen:

$$\Omega = G \times \cdots \times G$$

7. Zufallsstichprobe: $X_1, \ldots, X_n : \Omega \to \mathbb{R}$, unabhängig und identisch verteilte Zufallsvariablen (Zufallsvektoren, wenn mehrere Variablen erhoben werden).

8. Bei Experimenten werden den gezogenen $g \in G$ gewisse Ausprägungen zugeordnet (z.B. Kontrollgruppe/Behandlungsgruppe). Diese haben i.d.R. nur wenige mögliche Ausprägungen (z.B. binär 0/1).

9. Deskriptive Statistik betrachtet Realisation $(x_1, \ldots, x_n)'$ als Input, die **Datenmatrix**.

statistische Einheit	Merkmal	Merkmalsausprägungen
Studierender	Studienfach	BWL/Informatik/WiIng/Biologie/...
	Geschlecht	M/W/D
	Alter	$\mathbb{R}^+$
IT-Unternehmen	Mitarbeiterzahl	$\mathbb{N}$
	Umsatz	$\mathbb{R}_0^+$
	Gewinn/Verlust	$\mathbb{R}$
Arbeitnehmer	Einkommen	$\mathbb{R}^+$
	Bildungsniveau	Abitur/Bachelor/Master/...
	Arbeitszeit	R_0^+
Regionen	Arbeitslosenquote	$[0, 1]$
	Wirtschaftskraft	$\mathbb{R}^+$
Ballungsräume	Populationsdichte	$\mathbb{N}$ oder $\mathbb{R}$
	politische Funktion	Mittelzentrum / Landeshauptstadt / Hauptstadt
Staaten	Bruttoinlandsprodukt	$\mathbb{R}^+$
	Verschuldung (in %)	$[0, 100]$

Skalenniveaus: $X : G \to M$

Diskrete Merkmale: M endlich oder abzählbar unendlich.

Stetige Merkmalee: $M \subset \mathbb{R}$ Intervall (oder ganz $\mathbb{R}$).

In der Praxis werden stetige Merkmale oft vergröbert (komprimiert) durch **Gruppierung**.
Bsp: Einkommensklassen $[0, 500], (500, 1000], (1000, 5000], (5000, \infty)$.

Klassifikation von Merkmalen aufgrund des Skalenniveaus:
• **Nominalskala**: Ausprägungen nur unterscheidbar (Labels)
• **Ordinalskala**: Ausprägungen können verglichen werden (Schulnoten, Grad der Zustimmung 1-5, ...).
• **Metrische Skala (Kardinalskala, Intervallskala, Ratioskala)**:
Kardinalskala: Messe Vielfache einer Grundeinheit (analog Messtab).
Intervallskala: Nullpunkt willkürlich. Dann können Quotienten nicht interpretiert werden (Temperatur).
Verhältnis-, Quotienten- o. Ratioskala: Nullpunkt physikalisch zwingend (Längen, Gewichte, Geld, Anzahlen)

ACHTUNG:

• Daten sind oft durch Zahlen kodiert. Dies heißt noch lange nicht, dass Rechenoperationen sinnvoll sind.

• Welche Rechenoperationen und statistischen Verfahren sinnvoll angewendet werden können, hängt oft vom Skalenniveau der Daten ab.

Ziele: Tabellarische und grafische Aufbereitung von Zahlenmaterial.
Ausgangspunkt: **Rohdaten** (**Primärdaten**, **Urliste**) nach der Erhebung.
Allgemeine Situation: Erhebe p Merkmale an n statistischen Einheiten.

Darstellung der Daten in der **Datenmatrix** (Tabelle):

stat. Einheit Nr.	Geschlecht	Alter	Größe	Messwert
1	M	18	72.6	10.2
2	W	21	18.7	9.5
⋮				⋮
n	W	19	15.6	5.6

i-te Zeile: Werte der p Variablen für die i-ten statistischen Einheit (Beob.)
j-te Spalte: Stichprobe der n beobachteten Werte des j-ten Merkmals.

Zeilen = Beobachtungen, Spalten = Variablen

Selektiere Spalte:
$\rightarrow$ Stichprobe $x_1, \ldots, x_n$,
$\rightarrow$ Datenvektor $\mathbf{x} = (x_1, \ldots, x_n)'$.

Aufgabe: Visualisierung von Zahlenmaterial:

Prinzip der Flächentreue: Sollen Zahlen grafisch durch Flächenelemente visualisiert werden, so müssen die Flächen proportional zu den Zahlen gewählt werden.
Grund: Gehirn spricht auf Fläche an, nicht auf Höhe oder Breite eines grafischen Elements.

Beispiel: Visualisierung durch Kreisflächen.

$$F = \pi r^2$$

r : Radius, F : Fläche.
Man muss die Radii proportional zur Wurzel der Zahlen wählen.

Nominale/ordinale Daten:
Zähle aus, wie oft die Ausprägungen im Datensatz vorkommen.
Nominales Merkmal mit den Ausprägungen $a_1, \ldots, a_k$
Die **absoluten Häufigkeiten** (engl.: *frequencies, counts*) $h_1, \ldots, h_k$, sind durch

$$\begin{aligned} h_j &= \text{Anzahl der } x_i \text{ mit } x_i = a_j \\ &= \sum_{i=1}^{n} \mathbf{1}(x_i = a_j), \end{aligned}$$

$j = 1, \ldots, k$ gegeben. Die (tabellarische) Zusammenstellung der absoluten Häufigkeiten $h_1, \ldots, h_k$ heißt **absolute Häufigkeitsverteilung.**
Es gilt:

$$n = h_1 + \cdots + h_k.$$

Dividiert man die absoluten Häufigkeiten durch den Stichprobenumfang n, so erhält man die **relativen Häufigkeiten** $f_1, \ldots, f_k$. Für $j = 1, \ldots, k$ berechnet sich f_j durch

$$f_j = \frac{h_j}{n}.$$

f_j ist der Anteil der Beobachtungen, die den Wert a_j haben.

Die (tabellarische) Zusammenstellung der $f_1, \ldots, f_k$ heißt **relative Häufigkeitsverteilung.**

Die relativen Häufigkeiten summieren sich zu 1 auf: $f_1 + \cdots + f_k = 1$.

Darstellung durch Stab-, Balken- oder Kreisdiagramme.

Kreisdiagramm (Kuchendiagramm): Die Winkelsumme von 360° (Gradmaß) bzw. 2π (Bogenmaß) wird entsprechend den absoluten oder relativen Häufigkeiten aufgeteilt.

Zu einer relativen Häufigkeit f_i gehört also der Winkel

$$\varphi_i = \frac{h_i}{n} \cdot 360^\circ = 2\pi f_i [\text{rad}].$$

$\rightarrow$ Ordinales Merkmal: Ordne die Stäbe, Balken oder Kreissegmente entsprechend der Anordnung der Ausprägungen an.

Tipp: Zum Erkennen von Zusammenhängen mit einem anderen Merkmal Y ordne die Stäbe, Balken oder Kreissegmente nach dem anderen Merkmal Y an! (s. Beispiel zu Öleinnahmen und BIP im Buch).

Die sortierten Beobachtungen werden mit $x_{(1)}, \ldots, x_{(n)}$ bezeichnet. Die Klammer um den Index deutet somit den Sortiervorgang an. Es gilt:

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

$x_{(i)}$ heißt *i*-**te Ordnungsstatistik**,
$(x_{(1)}, \ldots, x_{(n)})$ heißt **Ordnungsstatistik** der Stichprobe $x_1, \ldots, x_n$.
Das **Minimum** $x_{(1)}$ wird auch mit $x_{\min}$ bezeichnet, das **Maximum** $x_{(n)}$ entsprechend mit $x_{\max}$.

Tipp: Für wenig Daten: Markiere die Beobachtungen x_i auf der reellen Zahlenachse und schreibe jeweils x_i drüber. Dann hat man von links nach rechts die Ordnungsstatistik und zugleich die Zuordnung zu den Ausgangsdaten $x_1, \ldots, x_n$. Zudem erkennt man, in welchen Bereichen sich die Daten häufen.

Messbereich (range): $[x_{\min}, x_{\max}]$ (kleinste Intervall, das alle Daten enthält).

Gruppierung (Klassierung) von Daten:
Lege k Intervalle

$$I_1 = [g_1, g_2],\ I_2 = (g_2, g_3],\ \ldots,\ I_k = (g_k, g_{k+1}],$$

fest, welche den Messbereich überdecken.
I_j heißt j-te **Gruppe** oder **Klasse** und ist für $j = 2, \ldots, k$ gegeben durch $I_j = (g_j, g_{j+1}]$. Die Zahlen $g_1, \ldots, g_{k+1}$ heißen **Gruppengrenzen**. Des Weiteren führen wir noch die k **Gruppenbreiten**

$$b_j = g_{j+1} - g_j, \qquad j = 1, \ldots, k,$$

und die k **Gruppenmitten**

$$m_j = \frac{g_{j+1} + g_j}{2}, \qquad j = 1, \ldots, k,$$

ein.

Histogramm:
Das Histogramm ist eine grafische Darstellung der relativen Häufigkeitsverteilung, die dem Prinzip der Flächentreue folgt.

1. Gruppiere in k Klassen mit Gruppengrenze $g_1 < \cdots < g_{k+1}$.
2. Berechne zugehörige relative Häufigkeiten $f_1, \ldots, f_k$.
3. Zeichne über Gruppe j ein Rechteck der Fläche f_j

Hierzu bestimmen wir die Höhe l_j des j-ten Rechtecks so, dass die Fläche $F_j = b_j l_j$ des Rechtecks der relativen Häufigkeit f_j entspricht:

$$F_j = b_j l_j \stackrel{!}{=} f_j \qquad \Rightarrow \qquad l_j = \frac{f_j}{b_j}, \qquad j = 1, \ldots, k.$$

Beispiel: Histogramm von $n = 30$ Leistungsdaten der Solarmodule.

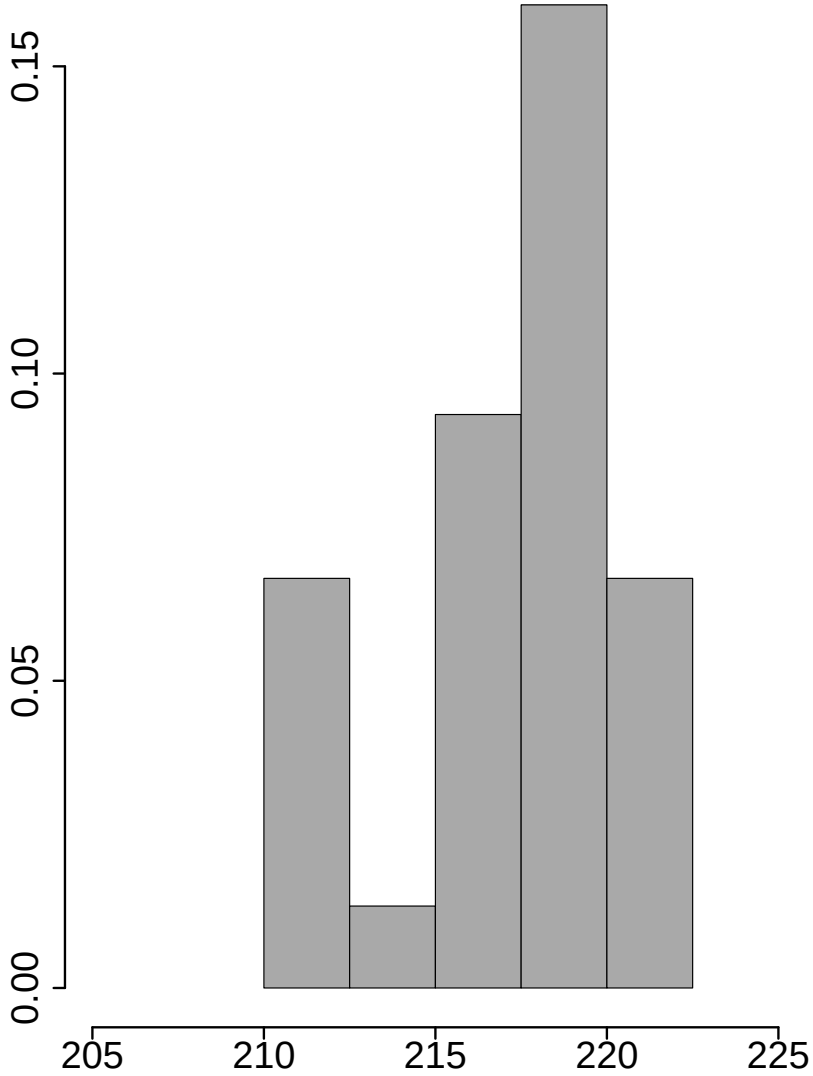
214.50	210.07	219.75	210.48	217.93	217.97	217.07	219.05
218.43	217.69	217.19	220.42	217.60	222.01	219.58	217.87
212.38	222.44	219.72	217.99	217.87	221.96	210.42	217.48
211.61	217.40	216.78	216.11	217.03	222.08		

Wir wählen 5 äquidistante Gruppen der Breite 2.5.
Mit den $k = 6$ Gruppengrenzen

$$g_1 = 210,\ g_2 = 212.5, \ldots,\ g_6 = 222.5$$

erhält man folgende Arbeitstabelle:

j	I_j	h_j	f_j	l_j
1	[210.0,212.5]	5	0.167	0.067
2	(212.5,215.0]	1	0.033	0.013
3	(215.0,217.5]	7	0.233	0.093
4	(217.5,220.0]	12	0.400	0.160
5	(220.0,222.5]	5	0.167	0.067


0.00
0.05
0.10
0.15
205
210
215
220
225

Der obere Rand des Histogramms definiert eine Treppenfunktion $\widehat{f}(x)$, die über dem j-ten Intervall I_j der Gruppeneinteilung den konstanten Funktionswert l_j annimmt. Außerhalb der Gruppeneinteilung setzt man $\widehat{f}(x)$ auf 0.

$$\widehat{f}(x) = \begin{cases} 0, & x < g_1, \\ l_1, & x \in [g_1, g_2], \\ l_j, & x \in (g_j, g_{j+1}],\ j = 2, \ldots, k, \\ 0, & x > g_{k+1}. \end{cases}$$

$\widehat{f}(x)$ heißt **Häufigkeitsdiche** oder auch **Dichteschätzer**.
$\rightarrow$ Die aus dem Histogramm abgeleitete Häufigkeitsdichte ist ein Schätzer für die Wahrscheinlichkeitsdichte $f(x)$ des Merkmals.

Die Häufigkeitsdichte ist selbst eine Wahrscheinlichkeitsdichte:
a) $\widehat{f}(x) \geq 0$ für alle $x \in \mathbb{R}$.
b) Für $x \in (g_j, g_{j+1}]$ ist sie konstant mit Wert

$$\widehat{f}(x) = l_j = \frac{f_j}{g_{j+1} - g_j}$$

so dass

$$\int_{g_j}^{g_{j+1}} \widehat{f}(x)dx = (g_{j+1} - g_j)\widehat{f}(x) = f_j.$$

Summation über j liefert daher den Wert 1 und somit

$$\begin{aligned}\int_{-\infty}^{\infty} \widehat{f}(x)\,dx &= \int_{g_1}^{g_{k+1}} \widehat{f}(x)\,dx \\ &= \int_{g_1}^{g_2} \widehat{f}(x)\,dx + \cdots + \int_{g_k}^{g_{k+1}} \widehat{f}(x)\,dx \\ &= \sum_{j=1}^{k} f_j = 1.\end{aligned}$$

Quantifizierung der Gestalt empirischer Verteilungen

Ziel

Beschreibe Zentrum der Daten, um das die Zahlen streuen.

Beispiel-Datensatz: Ozonkonzentration in 1000 [ppm]

i	1	2	3	4	5	6	7	8	9	10	11	12	13
x_i	66	52	49	64	68	26	86	52	43	75	87	188	118

Ordinal skalierte Daten

Definition

x_{med} heißt **Median** von $x_1, \ldots, x_n$, wenn

- mind. 50 % der Daten kleiner oder gleich x_{med} sind *und*
- mind. 50 % der Daten großer oder gleich x_{med} sind.

Median

Berechnung

- n ungerade: $x_{\text{med}} = x_{(k)}$, $k = \frac{n+1}{2}$.
- n gerade: Jede Zahl des Intervalls $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$.

Median

Konvention (metrisch skalierte Daten)

$$x_{\text{med}} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n \text{ ungerade}, \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & n \text{ gerade}. \end{cases}$$

Beispiel: Median

Beispiel

Sortiere die Daten...

$$26 \quad 43 \quad 49 \quad 52 \quad 52 \quad 64 \quad 66 \quad 68 \quad 75 \quad 86 \quad 87 \quad 118 \quad 188$$

Der Median dieser 13 *Messungen ist der* 7*-te Wert,* $x_{(7)} = 66$*, der sortierten Messungen.*

Median: Eigenschaften

Eigenschaften

- Vollzieht affin-lineare Transf. nach (Umrechnung von Einheiten!)

$$y_i = a + b \cdot x_i, \qquad i = 1, \ldots, n.$$

  Dann: $y_{\text{med}} = a + b \cdot x_{\text{med}}$.

- Vollzieht monotone Transformationen $f(x)$ nach:

$$y_i = f(x_i), \qquad i = 1, \ldots, n.$$

  Dann gilt: $y_{\text{med}} = f(x_{\text{med}})$.

- x_{med} minimiert $Q(m) = \sum_{i=1}^{n} |x_i - m|$.

Metrische skalierte Daten

Definition

Die Kennzahl

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

heißt **arithmetisches Mittel** oder **arithmetischer Mittelwert**.

Gruppierte Daten:

- $f_1, \ldots, f_k$: rel. Hf.
- $m_1, \ldots, m_k$: Gruppenmitten

Dann verwendet man:

$$\overline{x}_g = f_1 m_1 + \cdots + f_k m_k$$

Beispiel: Ozondaten

Beispiel

Für die Ozondaten erhält man:

$$\sum_{i=1}^{n} x_i = 66 + 52 + 49 + 64 + 68 + 26 + 86 + 52 + 43 + 75 + 87 + 188 + 118$$
$$= 974$$

und hieraus $\overline{x} = \frac{974}{13} = 74.923$.

Eigenschaften

Eigenschaften

- Schwerpunkteigenschaft
- Hochrechnung
- Verhalten unter affin-linearen Transformationen
- $\overline{x}$ minimiert $Q(m) = \sum_{i=1}^{n}(x_i - m)^2$.

Minimierungseigenschaft

$\overline{x}$ minimiert $Q(m) = \sum_{i=1}^{n}(x_i - m)^2$, $m \in \mathbb{R}$:

Ableitung von $(x_i - m)^2$ nach m: $2(x_i - m) \cdot (-1) = -2(x_i - m)$.

Ableitungen von $Q(m)$: Für alle $m \in \mathbb{R}$ gilt:

$$\begin{aligned} Q'(m) &= -2\sum_{i=1}^{n}(x_i - m) \\ &= -2\sum_{i=1}^{n} x_i + 2 \cdot n \cdot m, \\ Q''(m) &= 2n > 0 \end{aligned}$$

Nullsetzen der 1. Ableitung:

$$Q'(\widehat{m}) \stackrel{!}{=} 0 \quad \Leftrightarrow \quad 2n\widehat{m} = 2\sum_{i=1}^{n} x_i \quad \Leftrightarrow \quad \underline{\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Robustheit

Median oder arithmetisches Mittel?

- 9 arme Bauern: (Einkommen (in Euro): 1000) und 1 Reicher: (20000)
- $\overline{x} = (9 \cdot 1000 + 20000)/10 = 2900.$

Der Reiche ist ein **Ausreißer**. $\overline{x}$ reagiert sehr empfindlich auf solche Ausreißer!

- $x_{\text{med}} = 1000$ (Median-Einkommen)

Nominale/ordinale Daten

Streuung kategorialer Daten

Ausgangspunkt: relative Häufigkeitsverteilung

1	2	$\cdots$	k
f_1	f_2	$\cdots$	f_k

Nominale/ordinale Daten

Streuung kategorialer Daten

Keine Streuung:

1	2	$\cdots$	k
?	?	$\cdots$	?

Streuung kategorialer Daten

Keine Streuung: Nur eine Kategorie besetzt, z.B.:

1	2	$\cdots$	k
1	0	$\cdots$	0

Nominale/ordinale Daten

Streuung kategorialer Daten

Maximale Streuung:

1	2	$\cdots$	k
?	?	$\ldots$	?

Streuung kategorialer Daten

Maximale Streuung: Alle Kategorien gleich stark besetzt, d.h.:

1	2	$\cdots$	k
$\frac{1}{k}$	$\frac{1}{k}$	$\ldots$	$\frac{1}{k}$

Entropie

Betrachte: Gleichverteilung auf $r \leq k$ Kategorien $\rightarrow f_j = 1/r$

Anzahl r misst Streuung. In Binärdarstellung $001, 010, \ldots$ benötigte Bits:

$$b = \log_2(r) = -\log_2(1/r) = -\log_2(f_j)$$

Umlegen auf r Kategorien:

$$-\frac{1}{r}\log_2\left(\frac{1}{r}\right) = -f_j \log_2(f_j)$$

Erinnerung: Umrechnung Logarithmen: $\log_a(x) = \log_b(x) \cdot \log_a(b)$

Entropie

Definition

Die Kennzahl

$$H = -\sum_{j=1}^{k} f_j \log(f_j)$$

heißt **Shannon-Wieder-Index** oder **Shannon-Entropie**.

$$J = \frac{H}{\log(k)}$$

heißt **relative Entropie**

Eigenschaften

Eigenschaften

- $0 \leq H \leq \log(k)$
- $0 \leq J \leq 1$
- Minimalwert: 1-Punkt-Verteilung
- Maximalwert: Gleichverteilung auf k Kategorien

Metrisch skalierte Daten

Datenvektor: $\mathbf{x} = (x_1, \ldots, x_n)$

Definition

Stichprobenvarianz (empirische Varianz):

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

Bei gruppierten Daten:

$$s_g^2 = \sum_{j=1}^{k} f_j (m_j - \overline{x}_g)^2$$

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

Eigenschaften

Maßstabsänderung von Datenvektoren $\mathbf{x} = (x_1, \ldots, x_n)$

$$b \cdot \mathbf{x} = (b \cdot x_1, \ldots, b \cdot x_n)$$

Lageänderung

$$\mathbf{x} + a = (x_1 + a, \ldots, x_n + a)$$

Rechenregeln

- Invarianz unter Lageänderung

$$\text{var}(a + \mathbf{x}) = \text{var}(\mathbf{x})$$

- Quadratische Reaktion auf Maßstabsänderung

$$\text{var}(b \cdot \mathbf{x}) = b^2 \cdot \text{var}(\mathbf{x})$$

Verschiebungssatz

Verschiebungssatz

Es gilt:

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

sowie

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

Was macht die Praxis?

Praxis:

In der Praxis wird die Formel

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

verwendet. (Begründung in der LV *Statistik*).

Quantile

Beispiel

PC-Händler bestellt monatlich TFT-Monitore. In 9 von 10 Fällen soll die Lieferung bis zum Monatsende reichen.

Ansatz: Daten $x_{(1)} \leq \cdots \leq x_{(9)} \leq x_{(10)}$.
Für jede Zahl $x \in [x_{(9)}, x_{(10)}]$ gilt:

- Mindestens $9/10$ der x_i sind $\leq x$ und
- mindestens $1/10$ der x_i sind $\geq x$

(Für jedes $x \in (x_{(9)}, x_{(10)})$ gilt: Genau $9/10$ sind $\leq x$ und genau $1/10$ sind $\geq x$).

Quantile

Definition

Ein **empirisches** p-**Quantil**, $p \in (0,1)$, von $x_1, \ldots, x_n$ ist jede Zahl $\widetilde{x}_p$, so dass

- mindestens $100 \cdot p\%$ der Datenpunkte sind $\leq \widetilde{x}_p$ *und*
- mindestens $100 \cdot (1-p)\%$ der Datenpunkte sind $\geq \widetilde{x}_p$ *und*

Quantile

Berechnung

- np ganzzahlig: Jede Zahl aus $[x_{(np)}, x_{(np+1)}]$.
  Nicht immer ist das Merkmal metrisch skaliert. Dann sind mitunter nur bestimmte x-Werte interpretierbar, nicht jedoch 'Zwischenwerte'. Dann sind (nur) $x_{(np)}$ und $x_{(np+1)}$ Quantile.
- sonst: $\widetilde{x}_p = x_{(\lfloor np \rfloor + 1)}$.

Hierbei ist

$$\lfloor x \rfloor$$

die **Abrundung** einer Zahl $x \in \mathbb{R}$.

Quantile

Für metrische skalierte Daten gibt es verschiedene Konventionen, um die Defintion des Quantils eindeutig zu machen. Zum Beispiel:

Konvention: Intervallmitte: $\frac{1}{2}(x_{(np)} + x_{(np+1)})$

Quartile

Quartile:

$Q_1 = \widetilde{x}_{0,25}$: unteres Quartile (grenzt das untere Viertel ab)
$Q_2 = \widetilde{x}_{0,5}$: Median (grenzt die untere Hälfte ab, teilt die Verteilung)
$Q_3 = \widetilde{x}_{0,75}$: oberes Quartil (grenzt das obere Viertel ab).

Zwischen Q_1 und Q_3 liegen die zentralen 50% der Datenpunkte (die Mitte)!

$Q_3 - Q_1$ heißt **Interquartilabstand (IQR)** und ist ein robustes Streuungsmaß.

Beispiel

Fünfpunkte-Zusammenfassung und Boxplot
Die 5 Statistiken (Kennzahlen) $x_{\min}, Q_1, \widetilde{x}_{0.5} = x_{med}, Q_3, x_{\max}$ heißt **Fünfpunkte-Zusammenfassung**.

Boxplot: Grafische Darstellung der 5-Punkte-Zusammenfassung: