Kapitel 2

Beschreibende Statistik eindimensionaler Daten

2.1 Datenerhebung und Darstellung

Aufgabenstellung: Die Aufgabenstellung besteht in der Aufnahme und Analyse von Daten mit zufälligen Schwankungen. Dazu gehört die Schätzung der Parameter einer Grundgesamtheit aus einer Stichprobe. Die schließende Statistik baut darauf auf und erlaubt zum Beispiel die Prognose von Über- und Unterschreitenswahrscheinlichkeiten, das Erstellen eines Risikobildes sowie das Testen der Adäquatheit von Erklärungsmodellen.

Wir führen einige der grundlegenden Sprechweisen ein.

Grundgesamtheit: Diese kann real oder fiktiv sein; zum Beispiel eine Menge von Individuen, wie alle österreichischen Arbeitnehmer (real), oder alle potentiell möglichen Messungen der Zugfestigkeit einer bestimmten Stahlsorte (fiktiv).

Merkmal, Ausprägung: Jedes Element der Grundgesamtheit besitzt ein oder mehrere Merkmale mit den zugehörigen Ausprägungen. Über die Merkmalsausprägung hinaus besitzen die Individuen keine Individualität.

Bsp. 2.1

Grundgesamtheit	Merkmal	Ausprägungen
Einwohner Österreichs	Geschlecht	m/w/d
	Parteipräferenz*	keine/ÖVP/SPÖ/Grüne/...
	Alter*	$0, 1, 2, 3, \ldots$
Erwerbstätige Österreichs	Einkommen*	$0, 1, 2, \ldots, 1000, \ldots$ Euro
Gemeinden Tirols	Einwohnerzahl*	$1, 2, 3, \ldots$
	Fläche	$1, 2, 3, \ldots$ km^2
Versuche einer Messreihe	Messwert	Zahlenwert der Messskala

Bei zeitabhängigen Merkmalen (*) ist der Zeitraum zusätzlich anzugeben, also etwa das "Einkommen im Jahre 2023".

Merkmaltypen:

1. **Qualitative Merkmale, auch kategoriale Merkmale:**
   Z. B. Geschlecht, Parteipräferenz. Messung erfolgt durch:

   (a) **Nominalskala** (Benennung), z. B. m/w/d; keine/ÖVP/SPÖ/Grüne/...

   (b) **Ordinalskala** (Anordnung), z. B. Noten: SGT1, GUT2, BEF3, GEN4, NGD5.

2. **Quantitative Merkmale:**
   Z. B. Fläche, Einwohnerzahl, Noten als Zahlen, Einkommen, physikalische Parameter. Die Ausprägung ist durch eine Zahl charakterisiert. Messung erfolgt durch:

   (c) **Kardinalskala** (Zahlenwerte), z. B. Fläche 1000, 2000 km^2; Einwohner 1, 2, 3, 4 Millionen; Noten aus der Menge $\{1, 2, 3, 4, 5\}$.

Zufälliger Versuch: Herausgreifen eines Individuums und Feststellung bzw. Messung der Ausprägung des Merkmals.

Stichprobe vom Umfang n: Die n-fache Wiederholung eines Versuches unter identischen Bedingungen, z. B. Erhebung unter 1000 Arbeitnehmern oder 10 Zugversuche an verschiedenen Stäben derselben Stahlsorte.

In der Folge beschränken wir uns auf quantitative Merkmale. Die Ausprägungen der Merkmale seien entweder

$x_1, \ldots, x_k$	diskret oder
$x \in \mathbb{R},\ x \in [a, b]$	kontinuierlich.

Bsp. 2.2 Merkmal Schulnoten (Mathematik) in einer Klasse mit Ausprägungen:

$$x_1 = 1,\ x_2 = 2,\ x_3 = 3,\ x_4 = 4,\ x_5 = 5.$$

Das Ergebnis einer Stichprobe ist zunächst die Urliste $\xi_1, \ldots, \xi_n$ = 3, 4, 4, 3, 1, 3, 3, 3, 2, 2, 1, 2, 4, 4, 3, 4, 2, 4, 3, 3, 5, 1, 4, 2, 5, 2, 4, 3, 3, 4, hier für $n = 30$ Schüler und Schülerinnen.

Bsp. 2.3 Mittlere Tagestemperaturen im Jänner in Innsbruck (siehe Tabelle 2.1 am Ende dieses Abschnitts). Die Temperatur besitzt die Ausprägung einer kontinuierlichen Größe, die Messdaten sind selbstverständlich diskret, werden aber zur besseren Darstellung und Auswertung in Klassen unterteilt, wie stets bei kontinuierlichen Größen. Beginnen wir mit dem betrachteten Merkmal X = mittlere Tagesgemperatur im Jänner 2001. Die Daten bilden die Urliste $\xi_1, \ldots, \xi_n$ mit $n = 31$ Tagen. Wir nehmen eine Klasseneinteilung der Temperatur vor, hier in Klassen von $-7°$ bis $7°$ in Schritten von zwei Grad und zählen die Zahl der gemessenen Temperaturwerte in jeder Klasse.

Nr. j d. Klasse	Klasse	Häufigkeit H_j
1	$[-7°, -5°)$	0
2	$[-5°, -3°)$	4
3	$[-3°, -1°)$	6
4	$[-1°, +1°)$	6
5	$[+1°, +3°)$	7
6	$[+3°, +5°)$	4
7	$[+5°, +7°]$	4

Die Wahl der Klassenbreite bleibt den Bearbeitenden überlassen. Ist die Klassenbreite zu gering, zerflattert die Darstellung, ist sie zu groß, verschwimmen die Eigenschaften der Datenverteilung. Die Verwendung von Klassenintervallen vom Typ $[a, b)$ statt $(a, b]$ stimmt gerade mit der Funktionsweise des MATLAB-Befehls `histogram` überein.

Die Auswertung der Stichprobe liefert zunächst eine Häufigkeitsverteilung als Grundlage für alles Weitere.

Absolute Häufigkeit:

$$H_j = \begin{cases} H(X = x_j) & \text{Anzahl der Fälle, in denen } x_j \text{ aufgetreten ist,} \\ H(X \in \text{Klasse } j) & \text{Anzahl der Fälle, die in Klasse } j \text{ aufgetreten sind.} \end{cases}$$

Relative Häufigkeit:

$h_j = \frac{1}{n} H_j$ Anteil des Auftretens von x_j (bzw. Anteil der Elemente in Klasse j)

Bsp. 2.4 (Noten) Absolute und relative Häufigkeiten der Noten, $h_j\% = h_j \cdot 100$.

x_j	1	2	3	4	5
H_j	𝍸	𝍸 |	𝍸 𝍸	𝍸 ||||	||
H_j	3	6	10	9	2
$h_j\%$	10%	20%	33.33%	30%	6.67%

Graphische Darstellungen: Bei diskreten quantitativen oder kategorialen Merkmalen verwendet man meist ein Säulendiagramm für die Darstellung der absoluten bzw. relativen Häufigkeiten. Die Darstellung ist längentreu, d.h. die Höhe der Säule j entspricht der absoluten Häufigkeit H_j bzw. relativen Häufigkeit h_j.

Bsp. 2.5 (Noten) Absolute bzw. relative Häufigkeiten der Noten dargestellt als Säulen, siehe Abb. 2.1.

In MATLAB erhält man die Säulendiagramme mit `histogram(C,'BarWidth',.5)`, wobei `C` ein Kategorie-Objekt sein muss. Falls `x` die Noten als Zahlen enthält, wandelt

```
C = categorical(x) oder
C = categorical(x,1:5,'SGT1','GUT2','BEF3','GEN4','NGD5')
```

die Noten in die benötigten Kategorien um. Letzteres übersetzt die Zahlen in Bezeichnungen wie sie für die Zeugnisse verwendet werden.

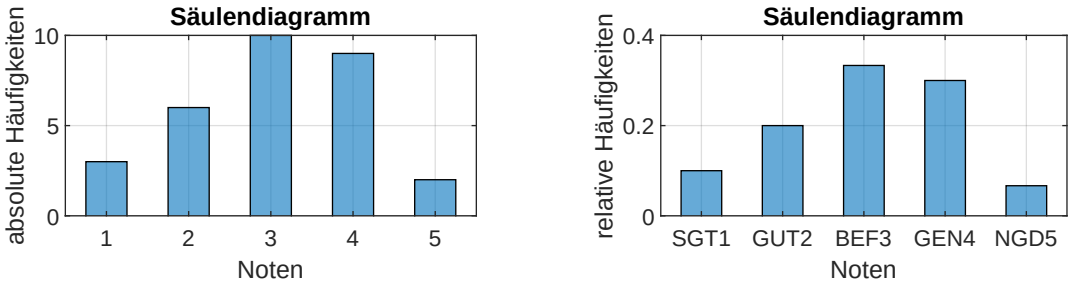

Abbildung 2.1: Säulendiagramme für absolute und relative Häufigkeiten.

Bei kontinuierlichen qualitativen Merkmalen verwendet man Histogramme. Die Höhe der Balken eines Histogramms ist so skaliert, dass nicht die Höhe, sondern die Fläche eines Balkens j der absoluten Häufigkeit H_j bzw. relativen Häufigkeit h_j der Elemente in einer Klasse j entspricht. Die Breite und Lage eines Balkens j entspricht genau der Breite und Lage der zugehörigen Klasse j. Die Höhe erhalten wir aus Höhe = Fläche / Breite, genauer aus Höhe = H_j / Breite bzw. Höhe = h_j / Breite. Die Gesamtfläche aller Balken ist dann entweder n (absolute Häufigkeit) oder 1 (relative Häufigkeit). Die Histogramme sind somit flächentreu.

Bsp. 2.6 (Mittlere Tagestemperatur Jänner 2001) Wir verwenden die weiter oben definierten Klassen und zeichnen die Häufigkeiten H_j bzw. h_j als flächentreues Histogramm. Für Klasse 5, d.h. Intervall $[1, 3)$ mit Breite 2, und Häufigkeit 7, erhalten wir als Höhe $\frac{7}{2} = 3.5$ (absolut) bzw. $\frac{1}{31} \cdot \frac{7}{2} = 0.1129$ (relativ), siehe Abb. 2.2.

In MATLAB erhält man Histogramme für die Daten `x` und Klassen `K` mit

`histogram(x,K,'Normalization','countdensity')` (absolut, flächentreu) oder
`histogram(x,K,'Normalization','pdf')` (relativ, flächentreu).

Hier für die Temperaturwerte in `x` und Klassen `K = -7:2:7`.

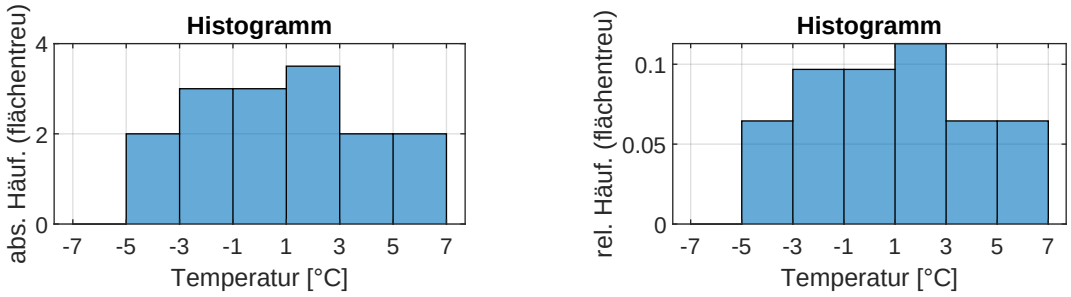

Abbildung 2.2: Histogramme (absolut/relativ) zur mittleren Tagestemperatur im Jänner 2001.

Für die empirische Verteilungsfunktion/Summenhäufigkeit brauchen wir das Konzept des Ereignisses X.

Elementarereignisse: Auftreten eines einzelnen Merkmales, also $X = x_j$. Die H_j sind dann die Auftrittshäufigkeiten der Elementarereignisse $X = x_j$.

Zusammengesetzte Ereignisse: Vereinigung von mehreren Elementarereignissen.

Bsp. 2.7 (Noten zwischen 2 und 4) Die absolute Häufigkeit davon ist

$$H(2 \leq X \leq 4) = H(X = 2 \text{ oder } X = 3 \text{ oder } X = 4) = 6 + 10 + 9 = 25$$

und die relative Häufigkeit $h(2 \leq X \leq 4) = 25/30 = 0.8333$.

Empirische Verteilungsfunktion, kumulierte- oder Summenhäufigkeit: Dies ist die für alle $x \in \mathbb{R}$ definierte (oberhalbstetige) Stufenfunktion

$$F_{\text{emp}}(x) = h(X \leq x) = \sum_{x_j \leq x} h_j = \frac{1}{n} \sum_{x_j \leq x} H_j.$$

Die Funktion F_{emp} ist wachsend und nimmt Werte in $[0, 1]$ an. Zur Berechnung von $F_{\text{emp}}(x)$ summiert man einfach die relativen Häufigkeiten $h_j = h(x_j)$ von jedem $x_j \leq x$. An der Stelle x_j macht F_{emp} einen Sprung der Höhe h_j. Die x_j sind entweder alle mit Vielfachheit h_j gemessenen Ausprägungen x_j, oder, falls bereits eine Klasseneinteilung vorliegt, üblicherweise die Klassenmitten. Mann könnte aber auch die unteren/oberen Klassengrenzen nehmen und würde dann die obere/untere empirische Verteilungsfunktion erhalten.

Bsp. 2.8 Empirische Verteilungsfunktionen bzw. Summenhäufigkeiten, siehe Abb. 2.3. In MATLAB erhält man die empirische Verteilungsfunktion mit dem Befehl `Femp`, der aus dem OLAT heruntergeladen werden kann.

Aufruf: `Femp(x,[],a,b)` (direkt aus den Daten) oder `Femp(x,K,a,b)` (über ein Histogramm mit Klassen `K`) für die Daten `x` und den Bereich `[a,b]`. Achtung: Der Bereich `[a,b]` muss alle Daten enthalten.

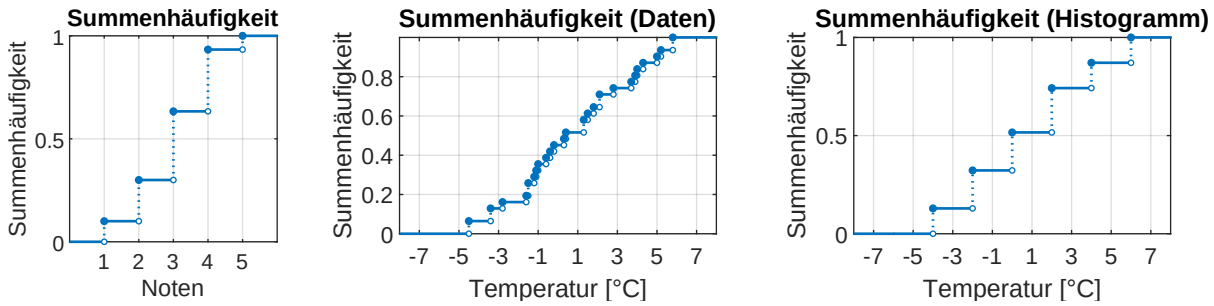

Abbildung 2.3: Empirische Verteilungsfunktion bzw. Summenhäufigkeit für die Noten (links) und die mittlere Tagestemperatur im Jänner 2001 basierend auf die Daten (Mitte) bzw. auf das Histogramm (rechts).

Wir benötigen die empirische Verteilungsfunktion einerseits zur graphischen Kurvenanpassung, andererseits für das Konzept der Überschreitenswahrscheinlichkeiten.

2.2 Datenbeschreibung: Lageparameter

Die wichtigsten Lageparameter sind der Mittelwert (Stichprobenmittel), Maximal- und Minimalwert, der Median und die Quantile.

Mittelwert: Arithmetisches Mittel der aufgetretenen Werte,

$$\begin{aligned}\overline{x} &= \frac{1}{n}\big(\xi_1+\xi_2+\cdots+\xi_n\big) \quad \text{(aus Urliste)}\\ &= \frac{1}{n}\big(x_1H_1+x_2H_2+\cdots+x_kH_k\big) \quad \text{(aus Ausprägungen und Häufigkeit)}\\ &= \frac{1}{n}\sum_{j=1}^{k} x_jH_j \;=\; \sum_{j=1}^{k} x_jh_j.\end{aligned}$$

Bsp. 2.9 (Noten) $\overline{x} = \frac{1}{30}\big(1\cdot 3+2\cdot 6+3\cdot 10+4\cdot 9+5\cdot 2\big) = 3.0333.$

Der Mittelwert ist i. A. kein Wert mehr aus den möglichen Ausprägungen $1, 2, 3, 4, 5$.

Maximalwert und Minimalwert: $x_{\max}\,, x_{\min}$.

Median (Zentralwert): Dieser gibt die Mitte der Urliste bei größenmäßig geordneter Anordnung an. Zur Ermittlung stellt man die Variationsreihe (geordnete Urliste) auf,

$$\xi_1 \le \xi_1 \le \cdots \le \xi_n\,.$$

Falls $n = 2\ell+1$ (also ungerade) ist, so ist der Median $\widetilde{x}$ gleich $\xi_{\ell+1}$; falls $n = 2\ell$ (also gerade) ist, so ist $\widetilde{x} = \big(\xi_\ell + \xi_{\ell+1}\big)/2\,.$

In MATLAB erhält man Mittelwert und Median direkt aus einer Urliste `x` mit `mean(x)` und `median(x)`.

Bsp. 2.10 (Noten) $x_{\max} = 5$ und $x_{\min} = 1$.

Variationsliste der Noten:

$$\underbrace{1,1,1,2,2,2,2,2,2,2,3,3,3,3,\mathbf{3}}_{50\%} \mid \underbrace{\mathbf{3},3,3,3,4,4,4,4,4,4,4,4,4,5,5}_{50\%}.$$

Hier ist $n = 30$ gerade. Daher befindet sich der Median $\widetilde{x} = (3+3)/2 = 3$ "zwischen den beiden fettgeschriebenen Dreiern".

Bsp. 2.11 (Temperaturen) Für die mittleren Tagestemperaturen im Jänner 2001 ergibt sich $\overline{x} = 0.76$, $x_{\max} = 5.80$ und $x_{\min} = -4.50$. Die Variationsreihe dazu ist

-4.5, -4.5, -3.4, -3.4, -2.8, -1.6, -1.5, -1.5, -1.2, -1.1, -1.0, -0.6, -0.4, -0.2, 0.3, **0.4**, 1.3, 1.3, 1.5, 1.8, 2.1, 2.1, 2.8, 3.7, 3.9, 4.0, 4.3, 5.0, 5.2, 5.8, 5.8

Hier ist $n = 31$ ungerade. Genau in der Mitte an 16. Stelle findet sich der Median $\widetilde{x} = 0.4$.

Quantile: Das q%-Quantil Q ist dadurch gekennzeichnet, das q% der Werte "unterhalb von Q" und $(100-q)$% "oberhalb von Q" liegen. Der Median ist somit das 50%-Quantil.

Für die Berechnung der Quantile gibt es leider unterschiedliche Varianten mit oder ohne Interpolation, allerdings erhält man für das 50%-Quantil immer den klassischen Median.

Klassische Definition (Stufenfunktion):

$$Q_q = \begin{cases} \frac{1}{2}(\xi_{n\cdot q} + \xi_{n\cdot q+1}) & n \cdot q \text{ ganzzahlig}, \\ \xi_{\lceil n\cdot q \rceil} & \text{sonst}. \end{cases}$$

Diese Formel lässt sich später mit der Definition von Quantilen für Wahrscheinlichkeitsverteilungen leicht erklären. MATLAB und Excel verwenden stückweise lineare Interpolation, deren Ergebnisse sich für q nahe 0 bzw. 1 leicht unterscheiden, siehe Abb. 2.4.

In MATLAB erhält man die Quantile mit `quantile(x,q)` mit z.B. `q = 0.25` für das 25%-Quantil. Mit dem Befehl `quantile(x,q,'inverse')` aus dem OLAT erhält man die Werte der klassischen Definition.

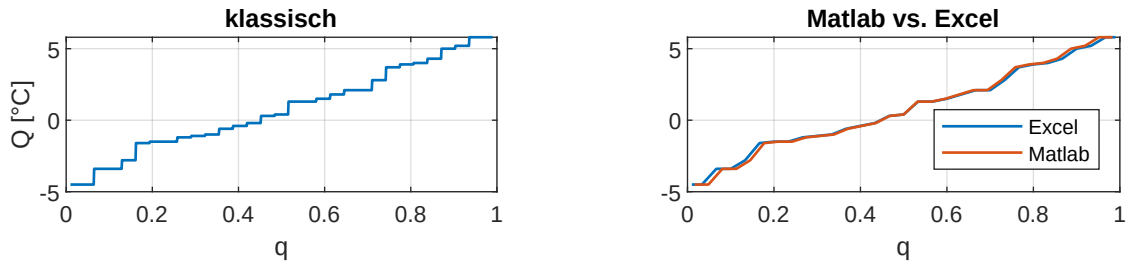

Abbildung 2.4: Quantile als Funktion von $q \in (0,1)$ für die mittleren Tagestemperaturen im Jänner 2001.

Viertelwerte:
V_u = unterer Viertelwert = 25%-Quantil und V_o = oberer Viertelwert = 75%-Quantil.

Die folgenden Größen sind eigentlich Streuungsmaße, leiten sich aber unmittelbar aus den Lageparametern her:

Spannweite: $\Delta = x_{\max} - x_{\min}$. **Viertelweite:** $d_V = V_o - V_u$.

Bsp. 2.12 (Notenbeispiel) $V_u = 2,\ V_o = 4,\ \Delta = 4,\ d_V = 2.$

Bsp. 2.13 (Mittlere Tagestemperatur Jänner 2001)

$$V_u = -1.4250,\ V_o = 3.4750,\ \Delta = 10.30,\ d_V = 4.90.$$

Ausreißer: Dies sind sind Messwerte, die augenfällig deutlich abseits der Mehrzahl der übrigen Messwerte liegen. Der Begriff bleibt vage, ein Ausreißer kann auf einen Mess- oder Aufnahmefehler hinweisen oder in der breiten Streuung der Daten begründet sein. Dies muss in jedem Einzelfall untersucht werden, die statistische Ausreißertheorie geht über den Rahmen der Vorlesung hinaus, siehe etwa Barnett et al. (1994).

In der Datenanalyse behilft man sich mit einer empirischen Definition von Ausreißern, indiziert durch das überschreiten gewisser Grenzen im Vergleich zu den übrigen Lageparametern. Üblicherweise legt man diese Grenzen wie folgt fest.

Ausreißergrenzen:

Obere Ausreißergrenze: $A_o = V_o + 1.5 \cdot d_V$. Untere Ausreißergrenze: $A_u = V_u - 1.5 \cdot d_V$.

Vergleich Median-Mittelwert:

Der Median ist stabil gegenüber Ausreißern, der Mittelwert nicht. Einzelne große/kleine Messwerte können den Mittelwert deutlich nach oben/unten verschieben.

Bsp. 2.14

x_j	1	2	3	4	5	20	21
H_j	3	7	10	9	0	1	1

$$\widetilde{x} = 3, \quad \overline{x} = \frac{1}{31}\big(1 \cdot 3 + 2 \cdot 7 + 3 \cdot 10 + 4 \cdot 9 + 20 \cdot 1 + 21 \cdot 1\big) = 4\,.$$

Boxplot: Der Boxplot besteht aus einem Rechteck, begrenzt durch den unteren Viertelwert V_u und dem oberen Viertelwert V_o. Der Median wird durch einen Strich im Rechteck gekennzeichnet. Links und recht schließen sich die "Whiskers" (Katzenhaare) an, die $x_{\min}$ und $x_{\max}$ andeuten.

Im Falle von Ausreißern geben die Whiskers den minimalen bzw. maximalen Wert unter Entfernung der Ausreißer an. Die Ausreißer werden durch zusätzliche Ringe, Sterne oder hier Kreuzchen gekennzeichnet, siehe Abb. 2.5.

In MATLAB erhält man für die Daten im Vektor `x` einen Boxplot mit `boxplot(x)`.

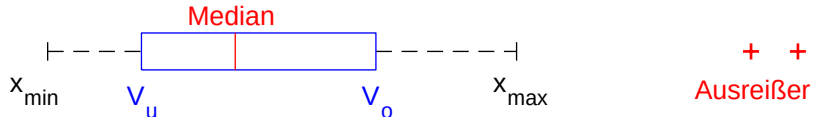

Abbildung 2.5: Schema eines Boxplots.

Bsp. 2.15 Boxplots für das Notenbeispiel und die mittleren Tagestemperaturen im Jänner 2001, siehe Abb. 2.6

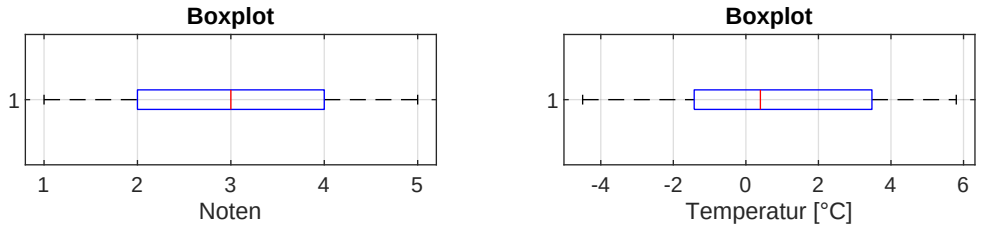

Abbildung 2.6: Boxplot für das Notenbeispiel (links) und die Temperaturen für Jänner 2001 (rechts).

Boxplots sind auch zum Vergleich verschiedener Gesamtheiten von Interesse.

Bsp. 2.16 Wir führen den Vergleich der Temperaturen im Jänner 2001 und im Jänner 2002 mit Hilfe verschiedener Darstellungsmethoden vor, siehe Abb. 2.7

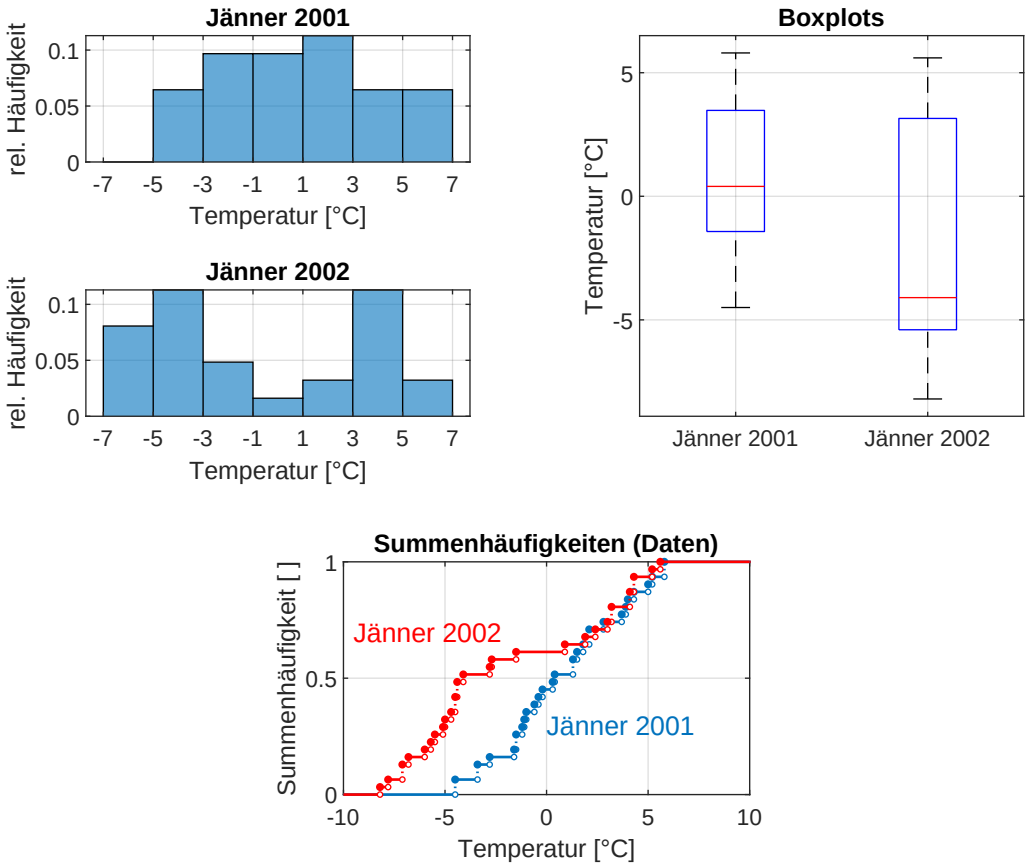

Abbildung 2.7: Vergleich der Tagesmittelwerte der Jännertemperaturen in den Jahre 2001 und 2002: Histogramme, Boxplots und empirische Verteilungsfunktionen.

2.3 Datenbeschreibung: Varianz

Die Stichprobenvarianz s^2 ist ein Streuungsmaß und gibt an, wie sehr die Messwerte im Quadratmittel um den Mittelwert schwanken.

$$(x_j - \overline{x})^2 = \text{ quadratische Abweichung des Messwertes } x_j \text{ vom Mittelwert } \overline{x}.$$

Mittelwert der Abweichungsquadrate, Stichprobenvarianz:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\Big((x_1-\overline{x})^2 H_1 + (x_2-\overline{x})^2 H_2 + \cdots + (x_k-\overline{x})^2 H_k\Big) \\ &= \frac{1}{n-1}\sum_{j=1}^{k}(x_j-\overline{x})^2 H_j \;=\; \frac{n}{n-1}\sum_{j=1}^{k}(x_j-\overline{x})^2 h_j. \end{aligned}$$

Der Faktor $(n-1)$ statt n wird in der Schätztheorie erklärt (Erwartungstreue des Schätzers s^2 für die Varianz einer Zufallsgröße).

Standardabweichung: Dies ist Quadratwurzel s der Stichprobenvarianz.

In MATLAB erhält man aus einer Urliste `x` die Varianz und Standardabweichung mit `var(x)` und `std(x)`.

Variationskoeffizient: Dieser gibt die relative Größe der Streuung in Bezug auf den Mittelwert an und ist definiert durch

$$v = s/\overline{x} \;\cdot\; 100\,\%.$$

Vorsicht: In der Ingenieurliteratur findet man auch die Bezeichung CoV, was aber zur Verwechslung mit der Kovarianz (COV) führen kann.

Bsp. 2.17 Standardabweichung, Notenverteilungen vom Umfang n:

x_j	1	2	3	4	5
H_j	2	2	2	2	2

$\overline{x} = \widetilde{x} = 3,\ s \approx 1.5$ (große Standardabweichung)

x_j	1	2	3	4	5
H_j	0	1	8	1	0

$\overline{x} = \widetilde{x} = 3,\ s \approx 0.67$ (kleine Standardabweichung)

Die Wichtigkeit der Stichprobenvarianz ergibt sich erst bei der Parameterschätzung und Anpassung von Verteilungen. Mit der Stichprobe allein kann man noch keine schließende Statistik machen. Dies geht erst nach Anpassung eines wahrscheinlichkeitstheoretischen Modells.

Vorschau: Wie die Anpassung einer Wahrscheinlichkeitsverteilung an die Daten mit Hilfe des Stichprobenmittels und der Stichprobenvarianz vorgenommen wird, können wir beispielhaft anhand des Modells einer Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ vorführen. Die Normalverteilung wird durch die Parameter μ (Erwartungswert) und σ^2 (Varianz) gesteuert, siehe Abb. 2.8.

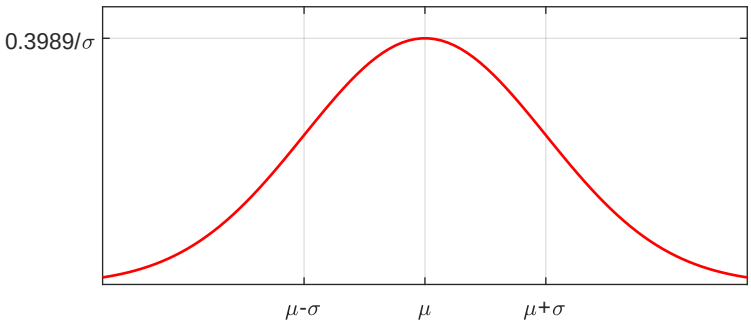

Abbildung 2.8: Dichte der Normalverteilung und Bedeutung der Parameter: Bei μ ist das Maximum der Dichtefunktion und bei $\mu \pm \sigma$ sind die Wendepunkte.

Nach der Momentenmethode erfolgt die Anpassung durch Schätzung aus den entsprechenden Stichprobenparametern, also $\mu \approx \overline{x}$, $\sigma \approx s$. Im Falle der mittleren Tagestemperaturen im Jänner 2001 erhält man die angepasste Normalverteilung $\mathcal{N}(0.76, 9.1085)$. Die Anpassungsgüte ist zufriedenstellend, wie an Hand der empirischen Verteilungsfunktion ablesbar ist, siehe Abb. 2.9.

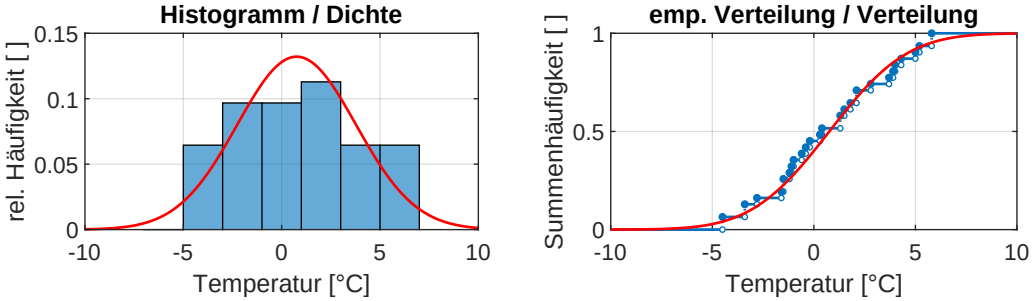

Abbildung 2.9: Mit der Momentenmethode an die Tagesmittelwerte im Jänner 2001 angepasste Normalverteilung, Jänner 2001. Histogramm und angepasste Dichte (links) sowie empirische Verteilungsfunktion und angepasste Verteilungsfunktion.

Als erste Anwendung können wir damit Wahrscheinlichkeiten für Temperaturbereiche näherungsweise berechnen, welche aus dem Histogramm nicht ablesbar sind (unter der Annahme der Gültigkeit des Normalverteilungsmodells).

Bsp. 2.18 Die Wahrscheinlichkeit, dass im Jänner eine Temperatur größer als 5° auftritt, ist nach diesem Modell gleich 0.08, also 8%, und kann als Flächeninhalt der rot gefärbten Fläche $\int_5^\infty f(\xi)\,\mathrm{d}\xi \approx 0.08$ unter der Dichtefunktion f der Normalverteilung $\mathcal{N}(0.76, 9.1085)$ abgelesen werden. Eine robuste Aussage würde allerdings verlangen, Temperaturdaten im Jänner von einer größeren Zahl von Jahren zu verwenden. Alternativ und einfacher kann die Überschreitenwahrscheinlichkeit mit Hilfe der Verteilungsfunktion $F(x) = \int_{-\infty}^x f(\xi)\,\mathrm{d}\xi$ abgelesen werden. Die Überschreitenswahrscheinlichkeit ist dann $1 - F(5) \approx 0.08$. Die Unterschreitenswahrscheinlichkeit wäre einfach $F(5) \approx 0.92$. Siehe Abb. 2.10.

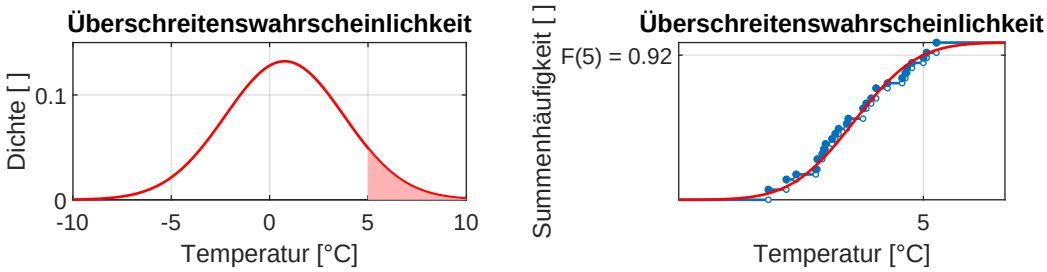

Abbildung 2.10: Darstellung einer Überschreitenswahrscheinlichkeit als Fläche unter der Dichte Funktion. Empirische Verteilungsfunktion (blau) und angepasste Normalverteilung (rot) für die mittleren Tagestemperaturen im Jänner 2002 in Innsbruck.

Ein Blick auf die empirische Verteilungsfunktionen in Abb. 2.11 zeigt übrigens, dass das Normalverteilungsmodell für Jänner 2002 nicht adäquat ist.

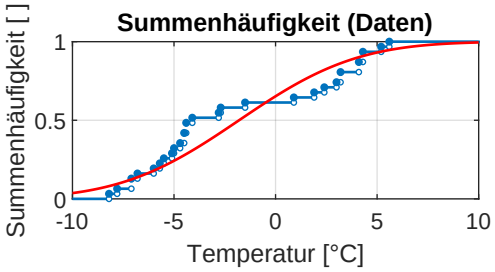

Abbildung 2.11: Versuch einer Anpassung einer Normalverteilung an die Tagesmittelwerte im Jänner 2002.

Tag	Temperatur Jänner 2001	Temperatur Jänner 2002	Bedeckung Jänner 2002	Tag	Temperatur Jänner 2001	Temperatur Jänner 2002	Bedeckung Jänner 2002
1	-3.4	-6.8	13	17	-4.5	-2.8	7
2	3.7	-7.1	0	18	-1.5	-4.4	90
3	2.8	-7.8	0	19	-1.1	-4.7	73
4	1.5	-8.2	10	20	-0.2	0.9	70
5	4.0	-7.1	27	21	-0.6	2.4	53
6	5.8	-5.7	10	22	2.1	1.9	57
7	5.0	-5.0	27	23	4.3	3.0	97
8	1.3	-4.5	10	24	5.8	5.2	53
9	0.4	-4.4	7	25	5.2	5.6	50
10	-0.4	-5.1	27	26	3.9	4.1	97
11	2.1	-6.0	73	27	1.3	4.3	100
12	1.8	-4.1	53	28	-1.2	4.3	30
13	0.3	-2.7	33	29	-1.6	4.1	40
14	-2.8	-1.5	10	30	-1.5	3.2	73
15	-3.4	-5.5	10	31	-1.0	3.2	40
16	-4.5	-4.5	33				

Tabelle 2.1: Mittlere Tagestemperaturen [°C] für Jänner 2001 und 2002 sowie der Bewölkungsgrad [%] für Jänner 2002 in Innsbruck, (aus: http://www.zamg.ac.at).