STS
Stochastik und Statistik

Die vorliegenden Notizen stammen aus dem Kurs STS (Stochastik und Statistik) für den Studiengang Aviatik, Informatik und Verkehrssysteme.
Es handelt sich um «work in progress», nicht um eine polierte finale Fassung.
Das Material wurde aus den Vorlagen von Ines Stassen Böhlen, Karl Lermer, Martin Frey und Marcel Dettling zusammengestellt. Als weitere Quellen dienten das Buch Mathematik für Informatiker von Gerald und Susanne Teschl, Statistik von Ludwig Fahrmeir, Grundlagen der Wahrscheinlichkeitsrechnung und Statistik von Erhard Cramer sowie diverse weitere Beispiele aus dann zitierten Quellen.
Rückmeldungen zu Druck- und anderen Fehlern werden sehr gerne entgegengenommen.

1 Deskriptive Statistik

Eine grundlegende Aufgabe der Statistik besteht darin, Informationen über bestimmte Objekte zu gewinnen, ohne dass dabei alle Objekte untersucht werden müssen. Es werden also Daten über eine Stichprobe erhoben und in der Folge ausgewertet, um daraus Schlussfolgerungen ziehen zu können. Man unterscheidet die folgenden drei Teilbereiche:

- *Deskriptive Statistik* (auch beschreibende Statistik): Ihre Aufgabe ist die Beschreibung («Deskription») und übersichtliche Darstellung von Daten, die Ermittlung von Kenngrössen und die Datenvalidierung (d.h., das Erkennen und Beheben von Fehlern im Datensatz).
- *Explorative Statistik*: Sie ist eine Weiterführung und Verfeinerung der beschreibenden Statistik. Ihre Aufgabe ist insbesondere die Suche («Exploration») nach Strukturen und Besonderheiten in den Daten.
- *Induktive Statistik* (auch schliessende Statistik, inferentielle Statistik, beurteilende Statistik): Sie versucht mithilfe der Wahrscheinlichkeitsrechnung über die erhobenen Daten hinaus allgemeinere Schlussfolgerungen zu ziehen.

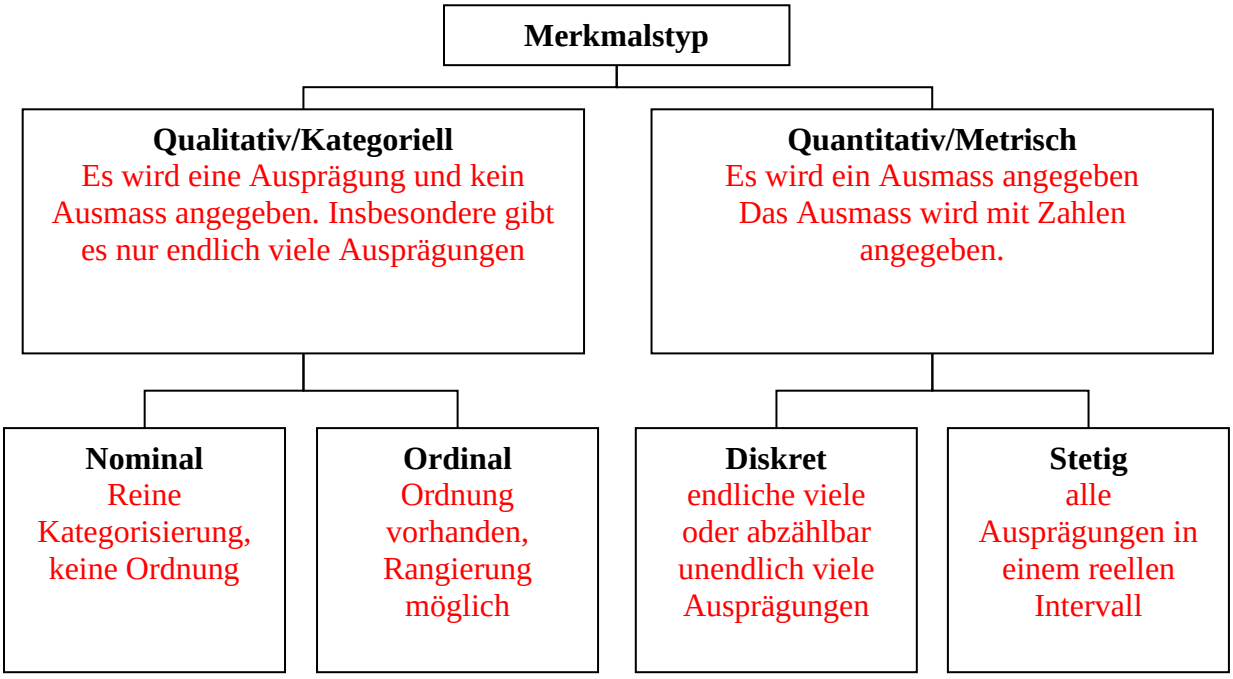
1.1 Begriffe

Wir erklären zunächst einige der Begriffe, die wir in diesem Modul verwenden werden:

- *Merkmalsträger* bzw. Statistische Einheiten: Objekte, an denen interessierende Grössen beobachtet und erfasst werden (z. B. Wohnungen, Menschen, Unternehmen, . . .).
- *Grundgesamtheit*: alle statistischen Einheiten, über die man Aussagen gewinnen möchte (z. B. alle Mietwohnungen in Zürich, alle Wahlberechtigten, alle im Vorjahr gegründeten Unternehmen, . . .). Eine Grundgesamtheit kann aus endlich vielen oder aus unendlich vielen Elementen bestehen. Sie kann real oder hypothetisch (z. B. alle potentiellen Kundinnen und Kunden) sein.
- *Vollerhebung*: Bei einer Vollerhebung werden bei jedem Individuum in der Grundgesamtheit die Eigenschaften erhoben/abgefragt/gemessen. Sobald die Grundgesamtheit eine gewisse Grösse hat, ist eine Vollerhebung meist weder nötig, noch mit vernünftigem Aufwand durchzuführen.
- *Stichprobe*: tatsächlich untersuchte Teilmenge der Grundgesamtheit. Sie soll ein möglichst getreues Abbild der Grundgesamtheit sein, diese also möglichst genau repräsentieren (repräsentative Stichprobe). Meist handelt es sich daher um eine Zufallsstichprobe. Diese ist dadurch gekennzeichnet, dass jede statistische Einheit dieselbe Chance hat, in die Stichprobe zu gelangen.
- *Stichprobengrösse*: Anzahl der Einheiten in der Stichprobe.
- *Merkmal* (auch Variable): interessierende Grösse, die an den statistischen Einheiten in der Stichprobe beobachtet (gemessen, erhoben) wird. Statistische Einheiten heissen in diesem Zusammenhang auch Merkmalsträger. In der Regel wird an einem Objekt mehr als ein Merkmal erhoben (z. B. werden die Merkmale «Nettomiete», «Baualter», «Grösse» für eine Wohnung ermittelt).

- (Merkmals)*Ausprägungen*: die verschiedenen Werte, die jedes Merkmal annehmen kann (z. B. hat das Merkmal «Geschlecht» die Ausprägungen «männlich»/«weiblich»/»divers»).

Bei statistischen Daten ist zu berücksichtigen, welche Information die Merkmale, die man beobachtet, beinhalten. Kann man sie der Grösse nach ordnen oder ist dies nicht sinnvoll? Oder sollen arithmetische Operationen möglich sein, wie z.B. die Berechnung des Durchschnitts?
Entsprechend solchen Anforderungen unterscheidet man im Wesentlichen vier Kategorien von Merkmalsarten. Man spricht auch vom Merkmalstyp oder Messniveaus.
Je nach Merkmalstyp sind andere Darstellungen geeignet, um die darin enthaltene Information zu vermitteln, und es gibt unterschiedliche Kennwerte für die Beschreibung. Das folgende Diagramm gibt einen Überblick über die merkmalstypen:



Bemerkung
Die Ausprägung der Merkmale ist nicht entscheidend für den Merkmalstyp bzw. das Messniveau. So haben z. B. Telefon-Nummern nominales Messniveau.

Beispiel 1: Merkmalstyp

Fragestellung	Ein Würfel mit den Zahlen 1 bis 6 wird viermal geworfen
Merkmalsausprägungen	Zahlen aus dem Bereich 1 bis 6.
Messniveau	Metrisch diskret

Fragestellung	Bei einer Umfrage werden 100 Menschen gefragt, welche der Parteien sie wählen.
Merkmalsausprägungen	BDP, CVP, FDP, GLP, Grüne, SP, SVP,…
Messniveau	Nominal

Fragestellung	Es werden 100 Programme auf ihre Robustheit getestet.
Merkmalsausprägungen	Prüfresultate schlecht, mittel und sehr gut
Messniveau	Ordinal

Fragestellung	Es werden 100 Programme auf ihre Geschwindigkeit bei einem bestimmten Input getestet.
Merkmalsausprägungen	Prüfresultate Laufzeiten
Messniveau	Metrisch stetig

1.2 Häufigkeiten und Verteilungsfunktion

Im folgenden Abschnitt werden wir uns zunächst einmal mit Merkmalträgern mit jeweils nur einem Merkmal befassen. Diese werden dann auch als univariate Daten bezeichnet.

1.2.1 Graphische Darstellung von Häufigkeiten

Bei der graphischen Darstellung ist zwischen kategoriellen und metrischen Daten zu unterscheiden:

- Kategoriell:…… Säulendiagramm, Stabdiagramm
- Metrisch:..…… Säulendiagramm, Stabdiagramm, Histogramm (Klassenbildung)

……

1.2.2 Absolute und relative Häufigkeiten

Das Entnehmen einer **Stichprobe** aus einer gegebenen Menge von Objekten kann mathematisch so formuliert werden: Aus einer Grundgesamtheit Ω von bestimmten Objekten werden n Objekte $\omega_1, \ldots, \omega_n$ ausgewählt und die Werte $X(\omega_1), \ldots, X(\omega_n)$ eines bestimmten Merkmals betrachtet. Die Funktion X ordnet jedem Objekt ω_i aus der Grundgesamtheit Ω seinen Merkmalswert $X(\omega_i)$ zu. Für die Stichprobenwerte $X(\omega_1), \ldots, X(\omega_n)$ schreiben wir auch häufig $x_1, \ldots, x_n$.

Beispiel 2: Geburtsmonate
Aus einer Klasse Ω werden 3 Studierende $\omega_1, \omega_2, \omega_3$ ausgewählt und deren Geburtsmonate $X(\omega_1), X(\omega_2), X(\omega_3)$ notiert.
Die Funktion $X: \Omega \rightarrow \{1,2,3,\ldots 12\}$ ordnet jedem Studierenden den Geburtsmonat zu.
……

……

Bemerkung
Jede Stichprobe ist durch die darin vorkommenden unterschiedlichen Merkmalsausprägungen und deren Häufigkeit charakterisiert. Wir unterscheiden *absolute und relative Häufigkeiten* der Merkmalsausprägungen.

Zunächst werden die n beobachteten Stichprobenwerte (Messwerte) nacheinander notiert. Die so entstehende *Liste* bezeichnet man auch als *Urliste*. Im Allgemeinen werden dabei gewisse Werte mehrmals auftreten. Bezeichnen wir diese verschiedenen Werte mit $a_1, a_2, \ldots, a_m$ und zählen wir, wie oft jeder dieser Werte in der Stichprobe auftritt.

absolute Häufigkeit

Diese Anzahl h_i nennt man…….. von a_i $(i = 1, \ldots, m)$.

Definition
Die relative Häufigkeit f_i von a_i erhält man, wenn man die absolute Häufigkeit durch den Stichprobenumfang dividiert:

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

Die Summe der absoluten Häufigkeiten ergibt den Umfang der Stichprobe, d.h. es ist
$h_1 + h_2 + .. + h_m = n$

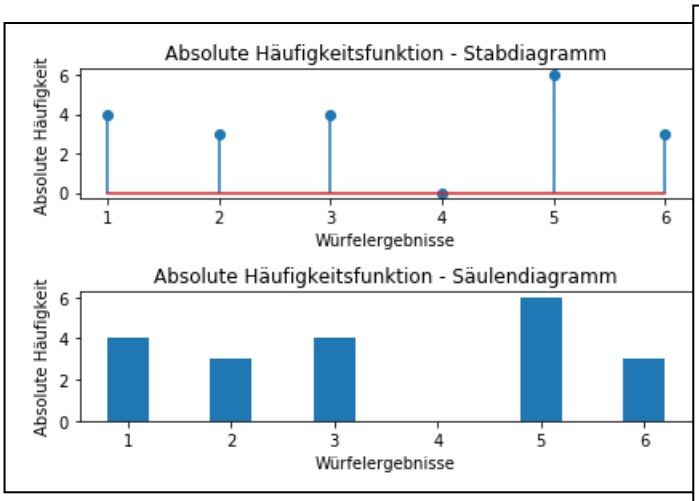
…….
Die Summe der relativen Häufigkeiten ist
$f_1 + f_2 + ... + f_m = 1$
…….

Beispiel 3: Würfel
Ein Würfel wird 20 Mal geworfen. Die folgende Tabelle gibt die Resultate, deren absolute und relative Häufigkeiten an:

a_i	1	2	3	4	5	6	Total
Absolute Häufigkeit h_i	4	3	4	0	6	3	20
Relative Häufigkeit f_i	4/20	3/20	4/20	0	6/20	3/20	1

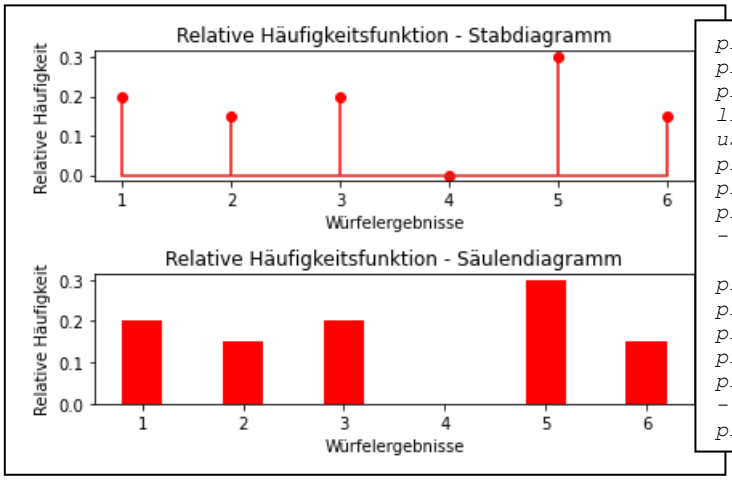
Die absolute Häufigkeit kann als Funktion $h: \mathbb{R} \to \mathbb{N}_0$ auf die reellen Zahlen fortgesetzt werden, indem diese für alle nicht in der Tabelle vorkommenden Merkmalsausprägungen gleich Null gesetzt wird. Entsprechend verfährt man mit der relativen Häufigkeit und fasst diese als Funktion $f: \mathbb{R} \to [0,1]$ auf den reellen Zahlen mit Werten zwischen 0 und 1 auf. Die so resultierende relative *Häufigkeitsfunktion* wird auch (empirische) *Dichtefunktion* (diskrete Merkmale: *PMF*, probability mass function, stetige Merkmale: *PDF*, probability density function) genannt.

Das folgende *Stabdiagramm* bzw. *Säulendiagramm* stellt die absolute Häufigkeitsfunktion des Beispiels graphisch dar. Rechts ist der Python Code zu sehen.
Um mit den Daten in Python arbeiten zu können bietet sich bei univariaten Daten das Modul «NumPy» an. Für die grafische Darstellung dient das Modul «MatPlotLib». Mehr zu den verwendeten Modulen finden Sie in den Moodle-Kurzdokumentationen.




```
Würfeln
import numpy as np
import matplotlib.pyplot as plt

A=np.arange(1,7,1) #Mögliche Ergebnisse
h=np.array([4, 3, 4, 0, 6, 3])
#Absolute Häufigkeiten
f=h/20 #relative Häufigkeiten

plt.figure(1)
plt.subplot(2,1,1)
plt.stem(A,h, use_line_collection=True)
plt.ylabel('Absolute Häufigkeit')
plt.xlabel('Würfelergebnisse')
plt.title('Absolute Häufigkeitsfunktion
- Stabdiagramm')

plt.subplot(2,1,2)
plt.bar(A,h,0.4)
plt.ylabel('Absolute Häufigkeit')
plt.xlabel('Würfelergebnisse')
plt.title('Absolute Häufigkeitsfunktion
- Säulendiagramm')
plt.tight layout()
```

Entsprechend ist im Folgenden die relative Häufigkeitsfunktion (*PMF*) im Stabdiagramm bzw. Säulendiagramm zu sehen. Lediglich die Skala auf der y-Achse hat sich im Vergleich zu vorher verändert.




```
plt.figure(2)
plt.subplot(2,1,1)
plt.stem(A,f,
linefmt='red',markerfmt='ro',
use_line_collection=True)
plt.ylabel('Relative Häufigkeit')
plt.xlabel('Würfelergebnisse')
plt.title('Relative Häufigkeitsfunktion
- Stabdiagramm')

plt.subplot(2,1,2)
plt.bar(A,f, 0.4,color="red")
plt.ylabel('Relative Häufigkeit')
plt.xlabel('Würfelergebnisse')
plt.title('Relative Häufigkeitsfunktion
- Säulendiagramm')
plt.tight_layout()
```

Beispiel 4: Programmlaufzeiten

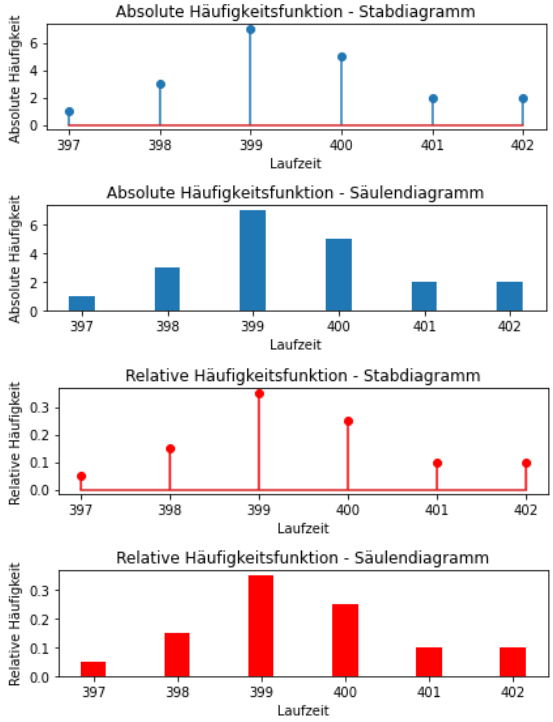
Ein Programm wird auf 20 Rechnern ausgeführt und jeweils die Laufzeit ermittelt. Auf diese Art erhalten wir eine Stichprobe vom Umfang $n = 20$.
Folgende Laufzeiten (in ms) werden dabei gemessen:

400, 399, 398, 400, 398, 399, 397, 400, 402, 399,
401, 399, 400, 402, 398, 400, 399, 401, 399, 399

Geben Sie die absoluten und die relativen Häufigkeiten der Messwerte an. Fertigen Sie Bilder der absoluten und relativen (Stabdiagramm, Säulendiagramm) Häufigkeitsfunktion an.

a_i	397	398	399	400	401	402	Total
Absolute Häufigkeit h_i	1	3	7	5	2	2	20
Relative Häufigkeit f_i	1/20	3/20	7/20	5/20	2/20	2/20	1

Skizze machen lassen, die Lösung ist in den Python-Dateien.
Je nach Wunsch können die Studierenden das natürlich auch mit Python machen.



1.2.3 Kumulative Verteilungsfunktion

Bei metrischen Messniveaus wird ausgehend von der Häufigkeitsfunktion eine weitere Funktion bestimmt, die sogenannte (empirische) *kumulierende* oder *kumulative Verteilungsfunktion* (*CDF*, cumulative distribution function) mit deren Hilfe man die Häufigkeiten beliebiger Intervalle im Messbereiche bestimmen kann.

Beispiel 5: Flugreisen
30 Haushalte (repräsentativ) werden nach der Anzahl Flugreisen im Jahr befragt mit dem folgenden Ergebnis.

Anzahl Flugreisen a_i	1	2	3	4	5	Total
Absolute Häufigkeit h_i	9	8	5	7	1	30
Relative Häufigkeit f_i	9/30	8/30	5/30	7/30	1/30	1
Kumulative abs. Häufigkeit H_i	9	17	22	29	30	
Kumulative rel. Häufigkeit F_i	9/30	17/30	22/30	29/30	1	

Ähnlich wie bei der Häufigkeitsfunktion kann die kumulative Häufigkeit als Funktion auf den reellen Zahlen aufgefasst werden. Die (empirische) *absolute Summenhäufigkeit* $H: \mathbb{R} \to \mathbb{N}_0$ ist definiert durch

$H(x) = Anzahl\ aller\ Stichprobenwerte\ \leq x$ bzw.
$H(x) = \sum_{i:a_i \leq x} h_i$

Die (empirische) *relative Summenhäufigkeit* $F: \mathbb{R} \to \mathbb{R}$ ist definiert durch

$F(x) = H(x)/Stichprobengrösse$ bzw.
$F(x) = \sum_{i:a_i \leq x} f_i$

Diese *CDF* ist eine Stufenfunktion, die monoton von 0 bis 1 wächst, und an den Stellen a_i genau um f_i springt. Dazwischen ist sie konstant.

Im Beispiel oben sind

$H(-23) = 0$ $H(4.6) = 29$ $H(2) = 17$ $H(6) = 30$

und entsprechend

$F(-23) = 0$ $F(4.6) = 29/30$ $F(2) = 17/30$ $F(6) = 1$

Der Anteil an Haushalten mit zwischen 2 und 4 Flugreisen berechnet sich mit der *CDF* als die Differenz:

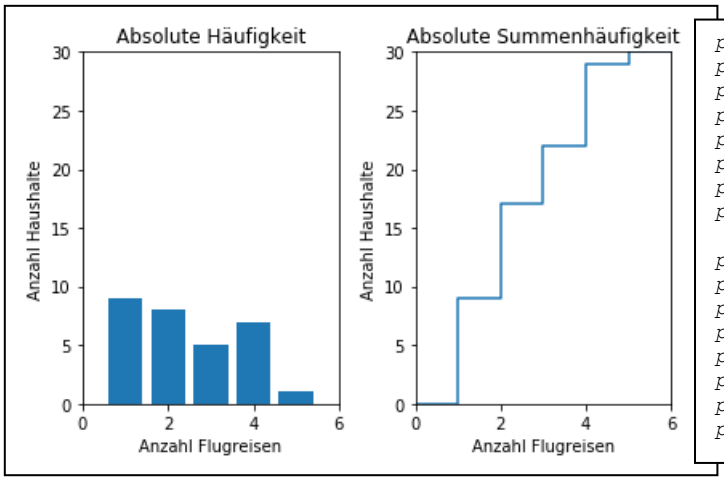
$F(4) - F(1) = 29/30 - 9/30 = 20/30 = 2/3$

Der Anteil an Haushalten mit über 3 Flugreisen berechnet sich als:

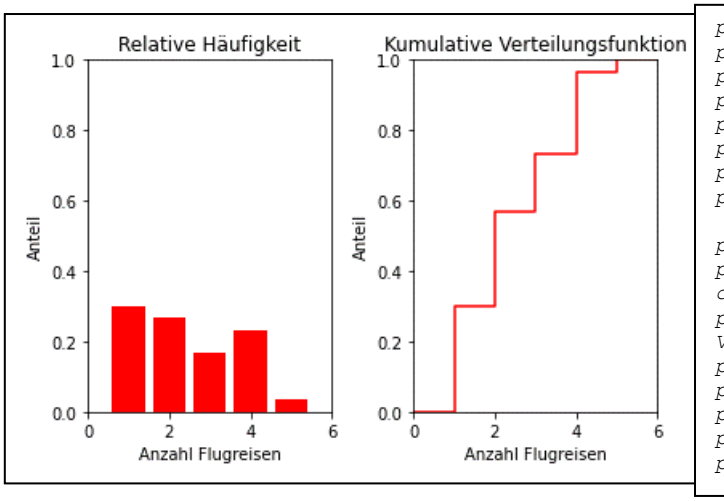
$1 - F(3) = 1 - 22/30 = 8/30 = 4/15$

```
#Anzahl Flugreisen
#Stichprobe bei 30 Haushalten
import numpy as np
import matplotlib.pyplot as plt

A=np.array([1, 2, 3, 4, 5]) #Anzahl der Flugreisen
h=np.array([9, 8, 5, 7, 1]) #Absolute Häufigkeit
n=np.sum(h) #Grösse der Stichprobe
f=h/n #relative Häufigkeit
H=np.cumsum(h) #absolute Summenhäufigkeit
F=np.cumsum(f) #relative Summenhäufigkeit, kumulative
Verteilungsfunktion(CDF)

#Anpassung der Arrays, damit der Plot schöner wird
A_p=np.concatenate((np.array([0]),A,np.array([6])),axis=0)
h_p=np.concatenate((np.array([0]),h,np.array([0])),axis=0)
H_p=np.concatenate((np.array([0]),H,np.array([n])),axis=0)
f_p=h_p/n
F_p=np.cumsum(f_p)
```

Graphisch wird die *CDF* als Stufenfunktion (oder auch Treppenfunktion) dargestellt. Im ersten nachfolgenden Diagramm sind die absolute Häufigkeit als Säulendiagramm und die absolute Summenhäufigkeit als Treppenfunktion zu sehen. Das zweite Diagramm zeigt die relative Häufigkeit und die kumulative Verteilungsfunktion (*CDF*).




```
plt.figure(1)
plt.subplot(1,2,1)
plt.bar(A_p, h_p)
plt.title('Absolute Häufigkeit')
plt.xlabel('Anzahl Flugreisen')
plt.ylabel('Anzahl Haushalte')
plt.xlim(0,6)
plt.ylim(0,30)

plt.subplot(1,2,2)
plt.step(A_p, H_p,where='post')
plt.title('Absolute Summenhäufigkeit')
plt.xlabel('Anzahl Flugreisen')
plt.ylabel('Anzahl Haushalte')
plt.xlim(0,6)
plt.ylim(0,30)
plt.tight_layout()
```

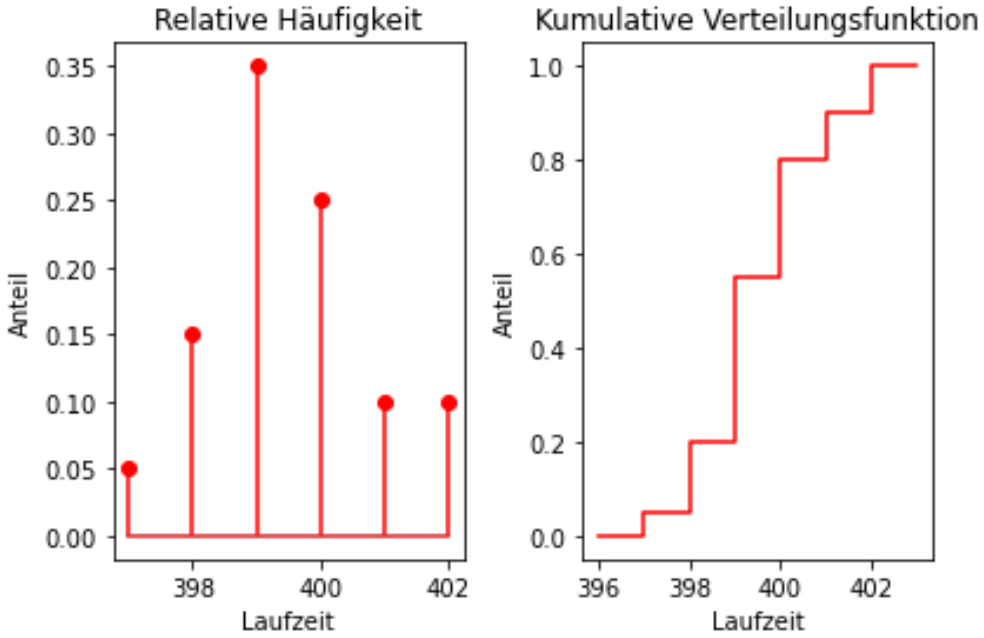



```
plt.figure(2)
plt.subplot(1,2,1)
plt.bar(A_p, f_p, color='red')
plt.title('Relative Häufigkeit')
plt.xlabel('Anzahl Flugreisen')
plt.ylabel('Anteil')
plt.xlim(0,6)
plt.ylim(0,1)

plt.subplot(1,2,2)
plt.step(A_p, F_p,where='post',
color='red')
plt.title('Kumulative
Verteilungsfunktion')
plt.xlabel('Anzahl Flugreisen')
plt.ylabel('Anteil')
plt.xlim(0,6)
plt.ylim(0,1)
plt.tight_layout()
```

Beispiel 6: Programmlaufzeiten (Fortsetzung)
Bestimmen Sie die kumulative absolute Häufigkeitsfunktion und die kumulative relative Häufigkeitsfunktion (Verteilungsfunktion *CDF*) für die Laufzeiten aus Bsp. 4 und fertigen Sie den Plot der *CDF* an.

a_i	397	398	399	400	401	402	Total
Absolute Häufigkeit h_i	1	3	7	5	2	2	20
Relative Häufigkeit f_i	1/20	3/20	7/20	5/20	2/20	2/20	1
Kumulative abs. Häufigkeit H_i	1	4	11	16	18	20	
Kumulative rel. Häufigkeit F_i	1/20	4/20	11/20	16/20	18/20	1	

Skizze machen lassen, die Lösung ist in den Python-Dateien.
Je nach Wunsch können die Studierenden das natürlich auch mit Python machen.



Zusammenfassung
Zu einer Stichprobe $x_1, \ldots, x_n$ von reellen Zahlen, definiert man die reellen Funktionen:

- (empirische) absolute Summenhäufigkeit
  $H: \mathbb{R} \to \mathbb{N}_0$, $H(x) = (\text{Anzahl aller Stichprobenwerte } x_i \leq x)$.
- (empirische) kumulative Verteilungsfunktion (*CDF*)
  $F: \mathbb{R} \to [0,1]$, $F(x) = H(x)/n$.

Jede empirische kumulative Verteilungsfunktion $F: \mathbb{R} \to \mathbb{R}$ hat die folgenden Eigenschaften:

- $F(x) = \sum_{r \leq x} f(r)$, mit der relativen Häufigkeitsfunktion (*PMF*)
  $f: \mathbb{R} \to \mathbb{R}$, $f(r) = (\text{Anzahl aller Stichprobenwerte } x_i = r)/n$
- Für alle reellen Zahlen x gilt $0 \leq F(x) \leq 1$.
- $F(x)$ ist monoton wachsend, d.h. $F(x) \leq F(y)$ für $x \leq y$.
- Der Graph von $F(x)$ ist eine rechtsseitig stetige Treppenfunktion.
- Es gibt eine reelle Zahl x mit $F(x) = 0$.
- Es gibt eine reelle Zahl y mit $F(y) = 1$.
- Der Anteil aller Stichprobenwerte x_i im Bereich $a < x_i \leq b$ berechnet sich als
  $F(b) - F(a)$

1.3 Klassierte Stichproben

Bei grossen Stichproben metrisch stetiger Merkmale teilt man die Stichprobenwerte in Klassen ein. Dazu wird der Merkmalsbereich in aneinander angrenzende (nicht notwendig gleich grosse) Intervalle aufgeteilt. Jeder Stichprobenwert wird so genau einem Intervall zugeordnet. Obere Intervallgrenzen zählen dabei immer zum darauffolgenden Intervall.
Die Anzahl der in einem Intervall enthaltenen Stichprobenwerte ergibt die absolute Häufigkeit bzw. nach Teilen durch die Stichprobengrösse, die relative Häufigkeit des betreffenden Intervalls. So werden absolute und relative Häufigkeiten den gebildeten Intervallen zugeordnet. Die Stichprobe wird *klassiert*.

Bei nichtklassierten Daten ergibt sich die relative Häufigkeit eines Wertes grafisch
aus der Höhe des Stabes/der Säule im Diagramm der PMF
……

……

Bei klassierten Daten treten die einzelnen Stichprobenwerte in den Hintergrund, es gibt nur noch Intervalle, welche nach der Häufigkeit des Auftretens der darin enthaltenen Stichprobenwerte gewichtet werden. Die relative Häufigkeit eines bestimmten Intervalls entspricht der Anzahl der darin enthaltenen Stichprobenwerte dividiert durch die Stichprobengrösse.
In der Graphik, d.h. im Histogramm, wird dies als Rechteck über dem jeweiligen Intervall veranschaulicht.

Bei klassierten Daten ergibt sich die relative Häufigkeit eines Intervalls
aus dem Flächeninhalt des darüber liegenden Rechtecks im Histogramm
……

……

Bemerkung
Graphische Darstellung: Die Häufigkeitsdichtefunktion h wird also nicht als Stab über der Klasse c_i dargestellt, sondern man stellt die Häufigkeit als *Rechtecksfläche* über der *Klassenbreite* d_i dar.
Die Höhe des Rechtecks ist die absolute (Häufigkeits-)*Dichte* h bzw. relative (Häufigkeits-) *Dichte* f. Dafür wird die Häufigkeit h_i bzw. f_i durch die Klassenbreite d_i geteilt.

$$h = \frac{h_i}{d_i} \text{ und } f = \frac{f_i}{d_i}$$

Dies wird genauer anhand des folgenden Beispiels erläutert.

Beispiel 7: Ausgaben für Transport
Die folgende Tabelle gibt die jährlichen Ausgaben für Transport im ÖV von 750 Personen wieder. Dabei wurden die Stichprobenwerte in fünf verschiedene Klassen eingeteilt. Die Klassengrenzen sowie die absoluten und relativen Häufigkeiten der Klassen sind ersichtlich. Die exakten Stichprobenergebnisse sind aus dieser Tabelle nicht mehr zu ergründen.

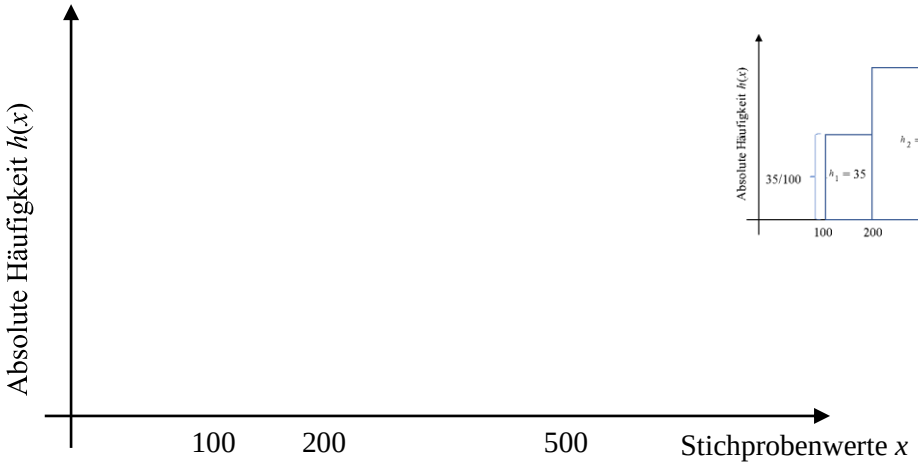
Klasse c_i Von … bis weniger als …	[100,200[	[200,500[	[500,800[	[800,1000[	[1000,2000[	Total
Absolute Häufigkeit h_i des Intervalls	35	182	317	84	132	750
Relative Häufigkeit f_i des Intervalls	35/750	182/750	317/750	84/750	132/750	1
Säulenbreite im Histogramm d_i	100	300	300	200	1000	

Die absolute Häufigkeitsdichtefunktion $h: \mathbb{R} \to \mathbb{R}$ ist bei klassierten Merkmalen eine Treppenfunktion (hier dargestellt als rechteckige Säulen) über den Klassen c_i.

Die Breiten der Säulen entsprechen den jeweiligen Intervallbreiten und die Flächen dieser Säulen den absoluten Häufigkeiten h_i im kompletten Intervall c_i.
Die Höhen der Säulen ergeben sich dann, indem man den Wert der absoluten Häufigkeit h_i, durch die Klassenbreite d_i teilt.

$$h = \frac{h_i}{d_i}$$

In der folgenden Grafik sind die beiden ersten Säulen des *Histogramms* der absoluten Häufigkeitsdichtefunktion $h: \mathbb{R} \to \mathbb{R}$ dargestellt:



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

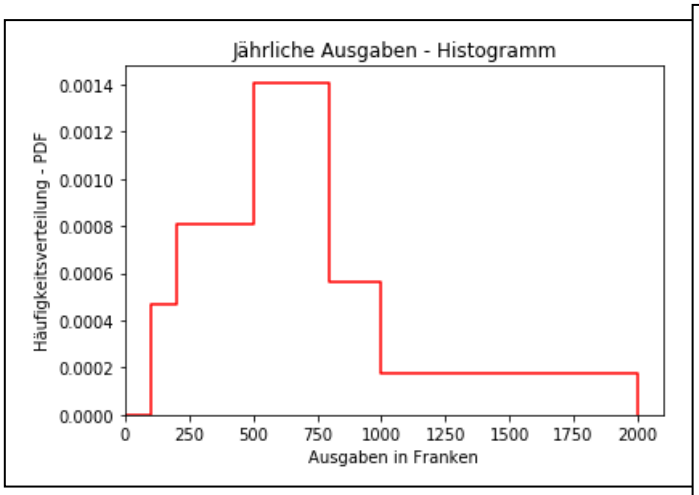
$$f(x) = h(x)/750$$

In der Tabellenübersicht ergibt sich damit:

Klasse c_i Von … bis weniger als …	[100,200[	[200,500[	[500,800[	[800,1000[	[1000,2000[	Total
Absolute Häufigkeit h_i des Intervalls	35	182	317	84	132	750
Relative Häufigkeit f_i des Intervalls	35/750	182/750	317/750	84/750	132/750	1
Säulenbreite im Histogramm d_i	100	300	300	200	1000	
Säulenhöhe im Histogramm (abs. Häufigkeitsdichte h: $\mathbb{R} \to \mathbb{R}$)	35/100	182/300	317/300	84/200	132/1000	
Säulenhöhe im Histogramm (rel. Häufigkeitsdichte f: $\mathbb{R} \to [0,1]$)	(35/750) /100	(182/750) /300	(317/750) /300	(84/750) /200	(132/750) /1000	

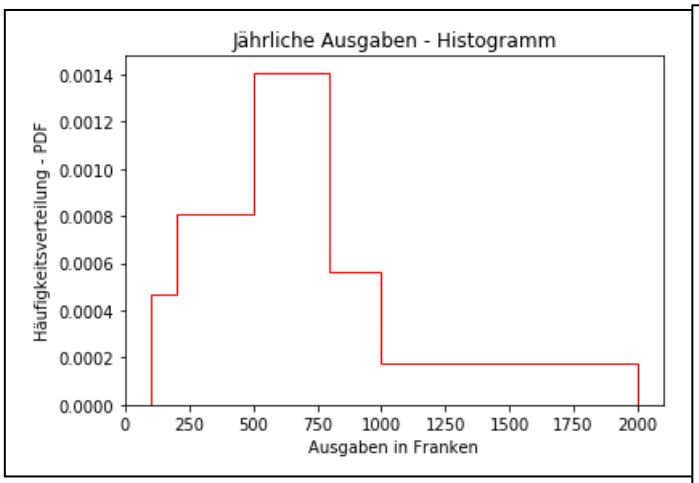
Dass die Fläche des Rechtecks ein Mass für die Anzahl der Messwerte ist, die in die Klasse fallen, hat den Vorteil, dass sich die graphische Darstellung kaum ändert, wenn Klassen zusammengelegt werden.

Das Histogramm kann grafisch wieder als Stufenfunktion dargestellt werden.




```
#Ausgaben für Transport
#stetig metrisches Merkmal und
Klassierung der Daten

import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

K=np.array([0, 100, 200, 500, 800,
1000, 2000]) #Klassengrenzen
h=np.array([0, 35, 182, 317, 84, 132,
0]) #absolute Klassenhäufigkeiten
APDF=np.array([0, 35/100, 182/300,
317/300, 84/200, 132/1000, 0])
#absolute Säulenhöhen im Histogramm
PDF=APDF/np.sum(h) #Säulenhöhen der PDF
im Histogramm

plt.figure(1)
plt.step(K,PDF,color='red',where='post'
);
plt.xlabel('Ausgaben in Franken')
plt.ylabel('Häufigkeitsverteilung -
PDF')
plt.title('Jährliche Ausgaben -
Histogramm')
plt.xlim(left=0)
plt.ylim(ymin=0)
```

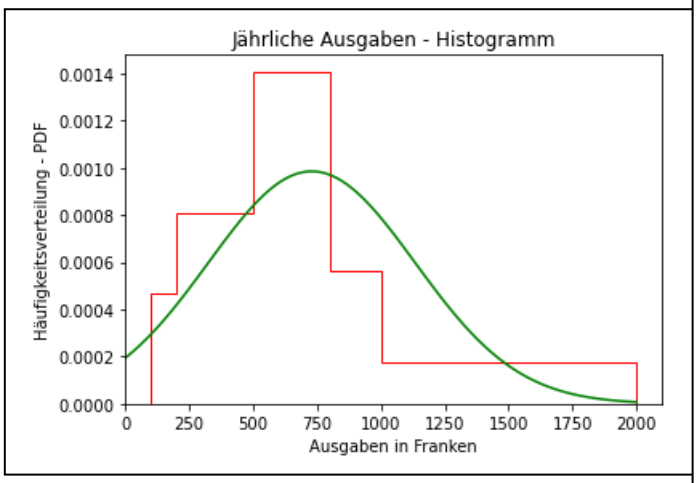
Alternativ dazu kann auch der Histogrammbefehl genutzt werden. Dafür werden allerdings die Rohdaten (nicht-klassiert) benötigt. Diese muss man sich in diesem Beispiel wieder generieren.




```
#Alternativ mit dem histogram Befehl
K=np.array([0, 100, 200, 500, 800, 1000,
2000]) #Klassengrenzen
#Stichprobe der Ausgaben
#mittlere Werte (für das Plotten des
Histogramms)
St=np.concatenate((np.ones(h[1])*((K[0]+
K[2])/2) , np.ones(h[2])*((K[2]+K[3])/2)
,np.ones(h[3])*((K[3]+K[4])/2) ,
np.ones(h[4])* ((K[4]+K[5])/2) ,
np.ones(h[5])*((K[5]+K[6])/2)))
plt.figure(2)
#Histogram der Häufigkeitsfunktion (PDF)
plt.hist(St,K,density='True',histtype='s
tep',color='red')
plt.xlabel('Ausgaben in Franken')
plt.ylabel('Häufigkeitsverteilung -
PDF')
plt.title('Jährliche Ausgaben -
Histogramm')
plt.xlim(left=0)
```

Bei stetigen Merkmalen, welche durch eine (möglichst repräsentative) Stichprobe beschrieben werden sollen, versucht man eine (möglichst gute) Approximation der Häufigkeitsfunktion mit einer stetigen Funktion zu erstellen. Zu dieser Funktionsgleichung zu kommen, ist Thema der schliessenden Statistik (wird später betrachtet).

Unten im Histogramm ist eine solche stetige Funktion (Normalverteilung) in grün eingezeichnet, welche die Häufigkeitsfunktion approximiert. Die Fläche unter dem Graphen dieser Dichte kann als Näherung in beliebigen Bereichen verwendet werden, um Prognosen über die Ausgaben für Transport zu erstellen.




```
plt.figure(3)
#Histogram der Häufigkeitsfunktion (PDF)
plt.hist(St,K,density='True',histtype='s
tep',color='red')
plt.xlabel('Ausgaben in Franken')
plt.ylabel('Häufigkeitsverteilung -
PDF')
plt.title('Jährliche Ausgaben -
Histogramm')
plt.xlim(left=0)

#Angenäherte Normalverteilung
mu= np.mean(St) #Erwartungswert einer
approx Normalverteilung
sig= np.std(St) #Standardabweichung
einer approx Normalverteilung
XN=np.linspace(0,2000,num=100)
#Definition der approx. Normalverteilung
YN=norm.pdf(XN,mu,sig)
plt.plot(XN,norm.pdf(XN,mu,sig),color='g
reen')
plt.xlim(left=0)
```

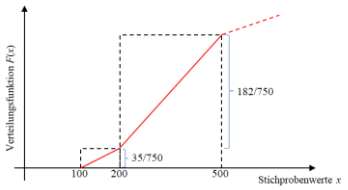
Definition

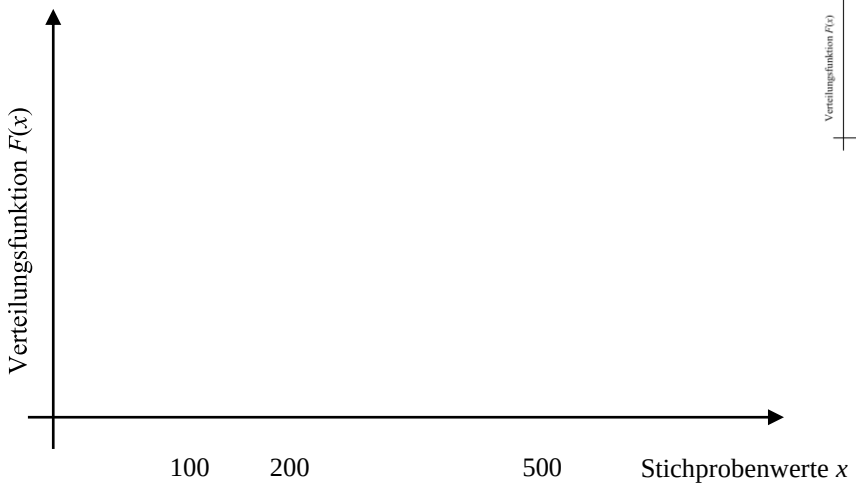
Durch Integration der relativen Häufigkeitsfunktion (*PDF*) $f(x)$ erhält man die *kumulative Verteilungsfunktion (CDF)*

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

der klassierten Daten.

Dabei gilt, dass $F(x)$ stetig und sogar stückweise differenzierbar ist. Die Werte von $F(x)$ an den rechten Klassengrenzen erhält man durch Kumulieren der relativen Häufigkeiten f_i im kompletten Intervall.



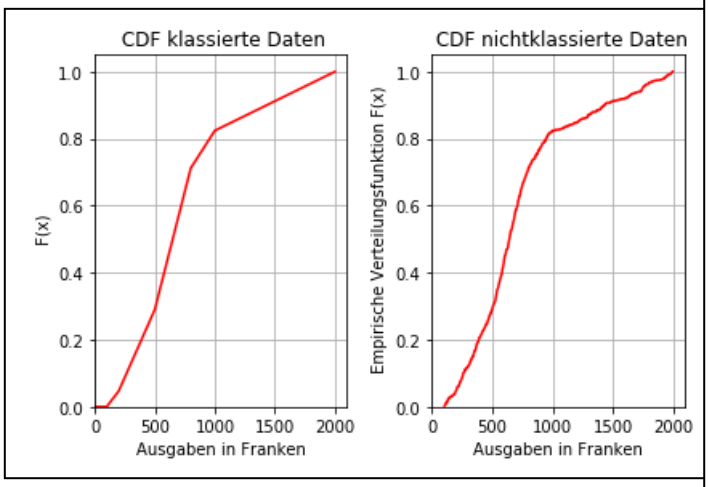


Insgesamt ergibt sich dann für die rechten Intervallgrenzen:

Klasse a_i Von … bis weniger als …	[100,200[	[200,500[	[500,800[	[800,1000[	[1000,2000[	Total
Relative Häufigkeit f_i des Intervalls	35/750	182/750	317/750	84/750	132/750	1
Werte der kum. Verteilungsfunktion (*CDF*) $F(x)$ an der rechten Intervallgrenze	35/750	217/750	534/750	618/750	1	

Im nachfolgenden Plot ist links die Verteilungsfunktion der klassierten Daten (Polygonzug) und rechts die empirische Verteilungsfunktion der Rohdaten (Treppenfunktion) zu sehen. Bei grossen Stichproben ist der Unterschied vernachlässigbar.

```
##Stichprobe generieren für die klassierten Daten
#Klassierte Daten
St_s=np.concatenate((np.random.randint(100,
199,size=35),np.random.randint(200, 499,size=182), np.random.randint(500,
799,size=317), np.random.randint(800,999,size=84), np.random.randint(1000,
2000,size=132)))
K=np.array([0, 100, 200, 500, 800, 1000, 2000])  #Klassengrenzen
h_s,K_s=np.histogram(St,K) #h=[35 182 317 84 132]; #abs. Klassenhäufigk.
Flaechen=h_s/750 #Saeulenflächen im Histogramm der PDF
CDF=np.append([0],np.cumsum(Flaechen)) #Werte der kum. Vert.funktion CDF

#Ohne Klassierung bzw. mit 750 Klassen
K_num = 750
h_s_m,K_m = np.histogram (St_s, bins=K_num, density=True)
CDF_m = np.cumsum (h_s_m)
```




```
plt.figure(4)
plt.subplot(1,2,1)
plt.plot(K,CDF,color='red') #
Verteilungsfunktion CDF
plt.xlabel('Ausgaben in Franken')
plt.ylabel(' F(x)')
plt.grid()
plt.title(' CDF klassierte Daten')
plt.xlim(left=0)
plt.ylim(ymin=0)

plt.subplot(1,2,2)
plt.plot (K_m[1:], CDF_m/CDF_m[-
1],color='red') # Empirische
Verteilungsfunktion CDF
plt.xlabel('Ausgaben in Franken')
plt.ylabel(' Empirische
Verteilungsfunktion F(x)')
plt.grid()
plt.title(' CDF nichtklassierte Daten')
plt.xlim(left=0)
plt.ylim(ymin=0)
plt.tight_layout()
```

Bemerkung

Für die Klassenbildung gibt es gewisse häufig angewandte Faustregeln, unter anderem:

- Die Klassen sollten gleich breit gewählt werden
- Die Anzahl der Klassen sollte etwa zwischen 5 und 20 liegen, jedoch $\sqrt{n}$ nicht wesentlich überschreiten (wobei n der Umfang der Stichprobe ist). Das stellt bis zu einem gewissen Grad sicher, dass alle Klassen «gut gefüllt» sind.

Beispiel 8: Klassierung von Daten

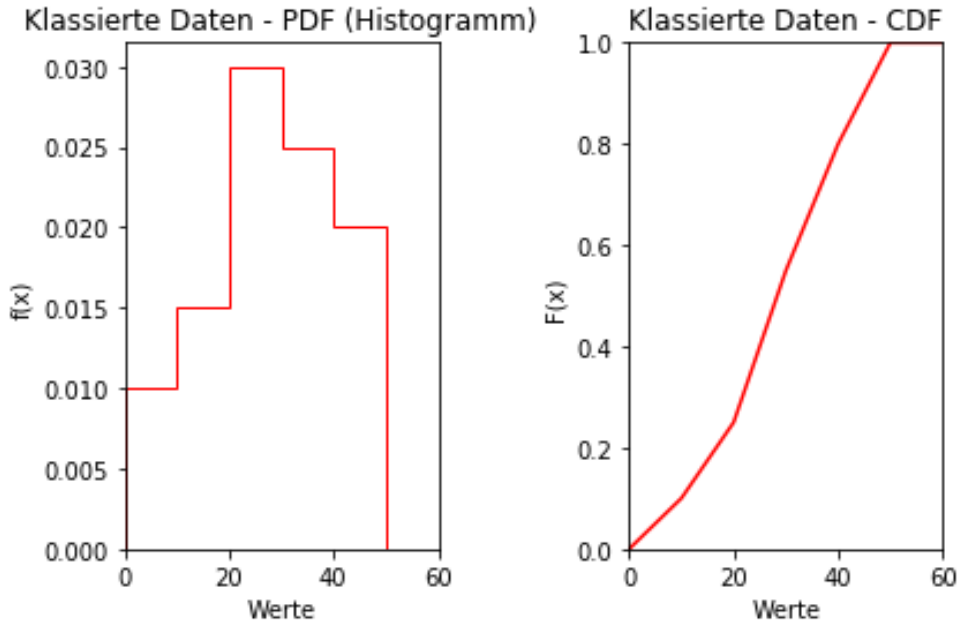
Gegeben ist folgende (bereits geordnete) Liste einer Stichprobe vom Umfang $n = 20$:

3, 7, 12, 18, 19, 20, 25, 25, 27, 28, 29, 31, 32, 34, 37, 38, 40, 41, 45, 47.

Gruppieren Sie die Stichprobenwerte in geeignete Klassen und bestimmen Sie die absoluten und die relativen Häufigkeiten der Klassen. Fertigen Sie Plots der *PDF* und *CDF* an.

Klasse c_i Von … bis weniger als …	[0,10[	[10,20[	[20,30[	[30,40[	[40,50[	Total
Absolute Häufigkeit h_i des Intervalls	2	3	6	5	4	20
Relative Häufigkeit f_i des Intervalls	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{6}{20}$	$\frac{5}{20}$	$\frac{4}{20}$	1
Säulenhöhe im Histogramm (*PDF*) $f(x)$	$\frac{2}{20 \cdot 10}$	$\frac{3}{20 \cdot 10}$	$\frac{6}{20 \cdot 10}$	$\frac{5}{20 \cdot 10}$	$\frac{4}{20 \cdot 10}$	
Werte der kumulativen Verteilungsfunktion (*CDF*) $F(x)$ an der rechten Intervallgrenze	2/20=0.1	5/20=0.25	11/20=0.55	16/20=0.8	1	

Plots skizzieren lassen oder mit Python machen lassen. Die Lösungen sind in den Python-Files. Der Plot würde so aussehen:



1.4 Kenngrössen

Bei Stichproben zu einem bestimmten Merkmal ist man meist nicht an den einzelnen Werten interessiert. Durch Analyse sollen charakterisierende Eigenschaften, Gemeinsamkeiten, Tendenzen und Abweichungen gefunden werden, welche zum Beispiel Rückschlüsse auf die Gesamtheit aller Merkmale, aus denen die Stichprobe stammt, erlauben oder den Vergleich zu anderen Verteilungen ermöglichen. Um empirische Verteilungen metrischer Merkmale studieren und miteinander vergleichen zu können bedient man sich sogenannter Kenngrössen, wie
Lage-, Streuungs- und Schiefemasse

.....

Lagemasse beschreiben das Zentrum der Verteilung

.....

Streuungsmasse charakterisieren die Abweichung vom Zentrum und

.....

Schiefemasse die Form der Verteilung.

.....

Beispiel 9: Mieten
Gegeben sind die Nettomieten von 151 Wohnungen (keine Klassenbildung, F_i in % angegeben) (auch als Excel-File).

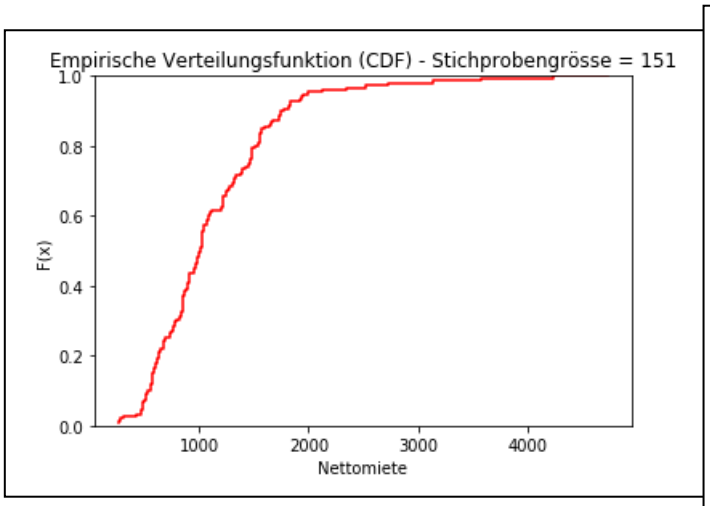
a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i
274	1	1	0.7	672	1	33	21.9	955	1	67	44.4	1250	1	100	66.2	1595	1	128	84.8
281	1	2	1.3	675	1	34	22.5	970	1	68	45	1260	1	101	66.9	1645	1	129	85.4
310	1	3	2	677	1	35	23.2	974	1	69	45.7	1275	1	102	67.5	1660	1	130	86.1
435	1	4	2.6	695	2	37	24.5	980	1	70	46.4	1301	1	103	68.2	1679	1	131	86.8
475	1	5	3.3	731	1	38	25.2	990	1	71	47	1305	1	104	68.9	1725	1	132	87.4
480	1	6	4	749	2	40	26.5	1000	2	73	48.3	1320	1	105	69.5	1730	1	133	88.1
483	1	7	4.6	772	1	41	27.2	1008	1	74	49	1325	1	106	70.9	1745	1	134	88.7
490	1	8	5.3	775	2	43	28.5	1020	1	75	49.7	1343	1	107	70.9	1750	1	135	89.4
510	2	10	6.6	800	2	45	29.8	1025	2	77	51	1375	1	108	71.5	1770	1	136	90.1
515	1	11	7.3	825	1	46	30.5	1035	2	79	52.3	1395	1	109	72.2	1825	1	137	90.7
520	1	12	7.9	840	1	47	31.1	1037	1	80	53	1398	1	110	72.8	1830	1	138	91.4
530	2	14	9.3	850	2	49	32.5	1040	4	84	55.6	1430	1	111	73.5	1835	1	139	92.1
560	1	15	9.9	855	1	50	33.1	1042	1	85	56.3	1460	1	112	74.2	1920	1	140	92.7
565	1	16	10.6	857	1	51	33.8	1050	1	86	57	1470	1	113	74.8	1930	1	141	93.4
570	2	18	11.9	860	1	52	34.4	1070	1	87	57.6	1480	2	115	76.2	1945	1	142	94
580	2	20	13.2	870	4	56	37.1	1080	1	88	58.3	1481	1	116	76.8	2000	1	143	94.7
584	1	21	13.9	876	1	57	37.7	1085	1	89	58.9	1485	1	117	77.5	2130	1	144	95.4
589	2	23	15.2	880	1	58	38.4	1095	1	90	59.6	1510	3	120	79.5	2345	1	145	96
600	2	25	16.6	895	1	59	39.1	1096	1	91	60.3	1547	1	121	80.1	2520	1	146	96.7
605	1	26	17.2	896	1	60	39.7	1120	1	92	60.9	1552	1	122	80.8	2730	1	147	97.4
614	1	27	17.9	900	1	61	40.4	1204	1	93	61.6	1554	1	123	81.5	3130	1	148	98
625	1	28	18.5	907	1	62	41.1	1220	1	94	62.3	1555	1	124	82.1	3575	1	149	98.7
628	1	29	19.2	914	1	63	41.7	1225	2	96	63.6	1560	1	125	82.8	4230	1	150	99.3
639	1	30	19.9	920	2	65	43	1226	1	97	64.2	1565	1	126	83.4	4725	1	151	100
650	2	32	21.2	950	1	66	43.7	1245	2	99	65.6	1570	1	127	84.1	Total	151		

Nun lesen wir die Daten in Python ein und plotten die empirische Verteilungsfunktion (*CDF*). Für das Einlesen verwenden wir das Modul «pandas», das wir auch noch bei den multivariaten Daten weiterverwenden werden.

Bereits aus dieser Darstellung können wir einige Informationen entnehmen:

- aus der Summenkurve sieht man, dass ca. 70 % der Nettomieten ≤ 1400 sind (in der Tabelle kann man genauer ablesen: 72.8 % aller Werte sind ≤ 1398)
- ca. 50% der Mieten sind kleiner als 1020.-
- ca. 20 % der Wohnungen sind teurer als 1547.-
- die 10% günstigsten Wohnungen liegen unter 565.-

Eine solche Verteilung kann mit Hilfe von statistischen Kennwerten beschrieben werden, welche die Lage und Form der Verteilung beschreiben. Eine Auswahl dieser Kennwerte wird im Folgenden erklärt.




```
#Mieten
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#Daten einlesen
df = pd.read_excel('Bsp_1_14.xlsx')
print (df.columns) #Bezeichnung der
Datenreihen
Miete=df.values[:,0]
hi=df.values[:,1]

n=np.sum(hi)  #Stichprobengrösse 151
Hi=np.cumsum(hi)  #kumulierte absolute
Häufigkeiten
Fi=Hi/n #emp. kum. rel. Häufigkeit

#Nichtklassierte Daten plotten
plt.figure(1)
plt.step(Miete, Fi,color='red')
plt.xlabel('Nettomiete')
plt.ylabel('F(x)')
plt.ylim(0,1)
plt.title('Empirische
Verteilungsfunktion (CDF) -
Stichprobengrösse = {}'.format(n))
```

1.4.1 Quantile

Ist eine Stichprobe $x_1, x_2, \dots, x_n$ eines metrischen Merkmals gegeben, so kann man die Werte der Stichprobe der Grösse nach ordnen und in verschiedene Bereiche unterteilen. Nach eventueller Umordnung gilt $x_{[1]} \leq x_{[2]} \leq \cdots \leq x_{[n]}$.

Die Position in dieser geordneten Liste bezeichnen wir auch als *Rang* der Beobachtung. Dabei kann es vorkommen, dass zwei Beobachtungen denselben Wert haben. Verschiedene Verfahren können benutzt werden, um auch hier eine eindeutige Rangzuordnung zu finden. Für die Bestimmung der Quantile ist es am einfachsten, den ranggleichen Beobachtungswerten zufällig einen der Ränge derselben Werte zuzuordnen.

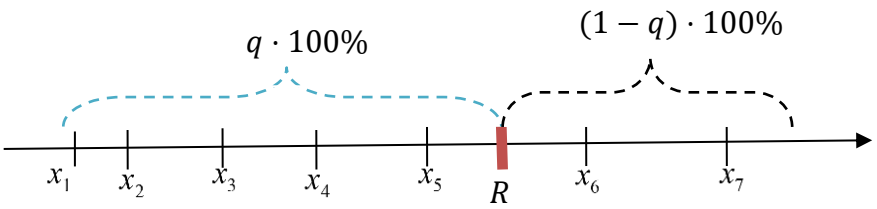
Definition

Für eine reelle Zahl $0 \leq q \leq 1$ heisst eine Zahl R ein *q-Quantil* (bzw. $q \cdot 100$tes Perzentil) der Stichprobe $x_1, x_2, \dots, x_n$, falls der Anteil der Stichprobenwerte $x_i \leq R$ mindestens q und der Anteil der Stichprobenwerte $x_i \geq R$ mindestens $1 - q$ ist.

1., 2. und 3. Quartil

Die 0.25, 0.5, und 0.75- Quantile werden auch……genannt.

Eine Zahl R ist genau dann ein q-Quantil, falls sie, wie im Bild unten angedeutet, die Stichprobe in zwei Teile aufteilt, mit einem unteren Teil mit mindestens $q \cdot 100\%$ und einem oberen Teil mit mindestens $(1 - q) \cdot 100\%$ aller Werte.



Da die so definierten q-Quantile im Allgemeinen nicht eindeutig bestimmt sind, wird bei (geordneten) Stichproben häufig folgende Definition verwendet:

Bemerkung
Um ein (empirisches) q-Quantil der geordneten Stichprobe zu bestimmen wird die Stichprobengrösse mit q multipliziert. Die Zahl bestimmt den Anteil für den unteren Teil der Stichprobenwerte.
Dann können die folgenden Fälle auftreten:

- $n \cdot q$ ist eine ganze Zahl: So ist das q-Quantil:
  $R_q = \frac{1}{2}(x_{n \cdot q} + x_{n \cdot q+1})$
  .....
- $n \cdot q$ ist keine ganze Zahl. So ist das q-Quantil:
  $R_q = x_{\lceil n \cdot q \rceil}$ mit $\lceil n \cdot q \rceil$ die nächstgrössere ganze Zahl
  .....

Beispiel 10: Quantile
Wir betrachten die Stichprobe: 4,4,0,3,5,3,1
Der Grösse nach geordnet lauten die Werte: 0,1,3,3,4,4,5
Die Stichprobengrösse ist $n = 7$.
Wir suchen nun die drei folgenden Quantile:
0.25-Quantil:
$n \cdot q = 7 \cdot 0.25 = 1.75$ mit $\lceil n \cdot q \rceil = 2$
$x_{\lceil n \cdot q \rceil} = x_{[2]} = 1$ (2. Stichprobenwert)
$R_{0.25} = 1$

0.5-Quantil:
$n \cdot q = 7 \cdot 0.5 = 3.5$ mit der nächstgrösseren ganzen Zahl $\lceil n \cdot q \rceil = 4$
$x_{\lceil n \cdot q \rceil} = x_{[4]} = 3$ (4. Stichprobenwert)
$R_{0.5} = 3$

0.75-Quantil:
$n \cdot q = 7 \cdot 0.75 = 5.25$ mit $\lceil n \cdot q \rceil = 6$
$x_{\lceil n \cdot q \rceil} = x_{[6]} = 4$ (6. Stichprobenwert)
$R_{0.75} = 4$

Man beachte dabei, dass sich in Python (wie auch bei anderen Statistik Programmen) mit dem quantile(..) Befehl unter Umständen abweichende Werte ergeben können. Solche Abweichungen sind bei grossen Stichproben unerheblich! So erhalten wir im Beispiel beim 0.25-Quantil einen abweichenden Wert:

```
#Quantile
import numpy as np
import matplotlib.pyplot as plt

X=np.array([4, 4, 0, 3, 5, 3, 1])
Q1=np.quantile(X,0.25)          → 2.0
Q2=np.quantile(X,0.5)           → 3.0
Q3=np.quantile(X,0.75)          → 4.0
```

Beispiel 11: Quantile
Wir betrachten die Stichprobe: 5,-2,1,7
Geordnet lauten die Stichprobenwerte: -2,1,5,7
Die Stichprobengrösse ist $n = 4$.
Wir suchen nun die drei folgenden Quantile:
0.25-Qantil:

$n \cdot q = 4 \cdot 0.25 = 1$

$$R_{0.5} = \frac{1}{2}(x_{[4\cdot 0.25]} + x_{[4\cdot 0.25+1]}) = \frac{1}{2}(x_{[1]} + x_{[2]}) = -1/2$$

0.5-Quantil:

$n \cdot q = 4 \cdot 0.5 = 2$

$$R_{0.25} = \frac{1}{2}(x_{[4\cdot 0.5]} + x_{[4\cdot 0.5+1]}) = \frac{1}{2}(x_{[2]} + x_{[3]}) = 3$$

0.75-Quantil:

$n \cdot q = 4 \cdot 0.75 = 3$

$$R_{0.75} = \frac{1}{2}(x_{[4\cdot 0.75]} + x_{[4\cdot 0.75+1]}) = \frac{1}{2}(x_{[3]} + x_{[4]}) = 6$$

Der quantile(..) Befehl in Python liefert das folgende Ergebnis:

```
#Quantile
import numpy as np
import matplotlib.pyplot as plt

X=np.array([5, -2, 1, 7])
Q1=np.quantile(X,0.25)          → 0.25
Q2=np.quantile(X,0.5)           → 3.0
Q3=np.quantile(X,0.75)          → 5.5
```

Das 2. Quartil bzw. 0.5-Quantil besitzt zusätzlich die Bedeutung eines Mittelwerts.

Definition
Das 0.5-Quantil (2.Quartil), die Mitte der Stichprobe, wird auch *Median* oder auch *Zentralwert* genannt.

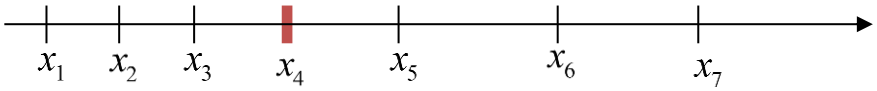
Allgemein teilt ein Median einen Datensatz, eine Stichprobe oder eine Verteilung so in zwei (gleich grosse) Hälften, dass die Werte in der einen Hälfte nicht grösser als der Medianwert sind, und in der anderen nicht kleiner.

Bemerkung
Damit der Median als wichtiges *Lagemass* für Stichproben eindeutig bestimmt ist, wird bei (geordneten) Stichproben $x_1, x_2, \ldots, x_n$ häufig folgende Definition verwendet:

$$Median(x_1, \ldots, x_n) = x_{med} = \begin{cases} x_{\left[\frac{n+1}{2}\right]} & \text{falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right) & \text{falls } n \text{ gerade} \end{cases}$$ (als Wiederholung)

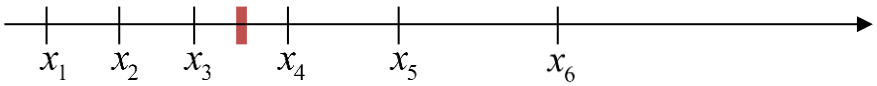
Das heisst, wenn eine ungerade Anzahl n von Stichprobenwerten vorliegt, so wird als Median, der Stichprobenwert in der Mitte $x_{\left[\frac{n+1}{2}\right]}$ genommen.

$$Median(x_1, \ldots, x_7) = x_{[4]}$$



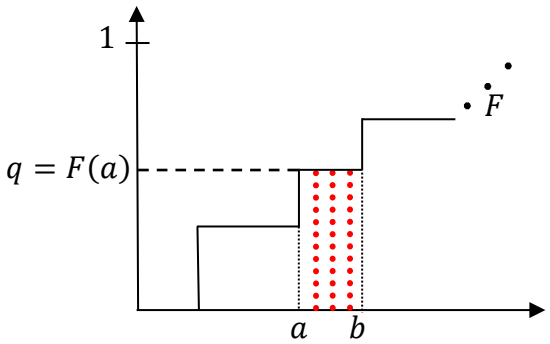
Falls die Anzahl der Stichprobenwerte n gerade ist, so wird das arithmetische Mittel der beiden angrenzenden Stichprobenwerte in der Mitte genommen: $\frac{1}{2}\left(x_{\left[\frac{n}{2}\right]} + x_{\left[\frac{n}{2}+1\right]}\right)$

$$Median(x_1, \ldots, x_6) = \frac{1}{2}(x_{[3]} + x_{[4]})$$



Bemerkung
Die Differenz zwischen dem 3. und 1. Quartil heisst *Interquartilsabstand* (*IQR* von interquartile range). Dies ist ein zum Median gehörendes Streuungsmass der Verteilung.

Mit Hilfe der *kumulativen Verteilungsfunktion* (*CDF*) $F(x)$ der Stichprobe $x_1, x_2, \ldots, x_n$ lassen sich die q-Quantile, also insbesondere die Quartile und der Median, alternativ bestimmen.
Zur Erinnerung: Die Verteilungsfunktion (*CDF*) $F(x)$ ist eine (rechtsseitig stetige) Treppenfunktion, welche die kumulativen relativen Häufigkeiten der Stichprobe angibt. Für eine Zahl $0 \leq q \leq 1$ können genau zwei Fälle eintreten, wie unten in den beiden Diagrammen zu sehen ist:



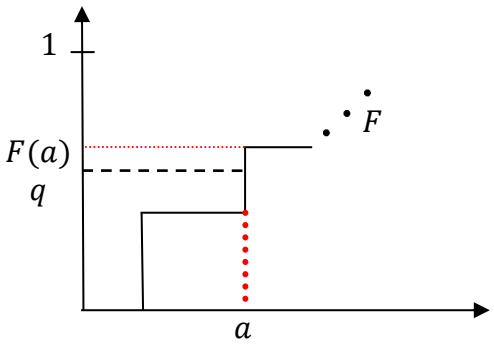


Diagramm (links)

Der q-Wert trifft auf ein Plateau der Treppenfunktion $F(x)$. In diesem Fall ist $R_q = \frac{1}{2}(a+b)$ das q-Quantil, mit dem auf a folgenden, nächstgrösseren Samplewert b.

In diesem Fall ist $n \cdot q$ eine ganze Zahl

.....

Diagramm (rechts)

Der q-Wert trifft auf einen Sprung der Treppenfunktion $F(x)$. In diesem Fall ist der Samplewert $R_q = a$ das q-Quantil. Man beachte, dass a auch ein r-Quantil ist, für alle $q \leq r < F(a)$.

In diesem Fall ist $n \cdot q$ keine ganze Zahl oder

.....

der auf a folgende Samplewert in der Stichprobe ist nochmals a.

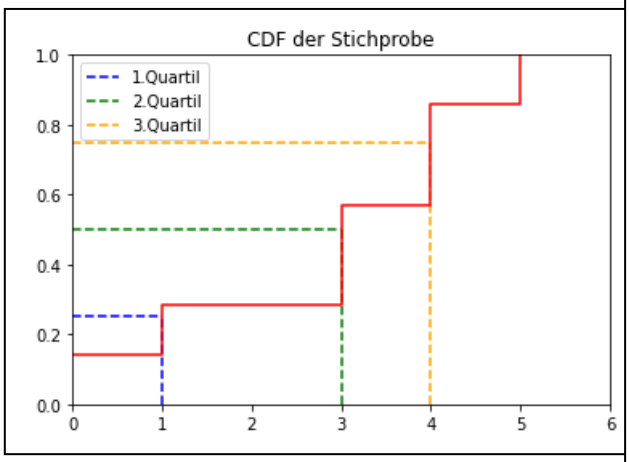
.....

Beispiel 12: Quantile (aus *CDF*)

Wir betrachten die Stichprobe: 4,4,0,3,5,3,1

Geordnet lauten die Stichprobenwerte: 0,1,3,3,4,4,5

Wir suchen nun die Quartile, indem wir die *CDF* betrachten.



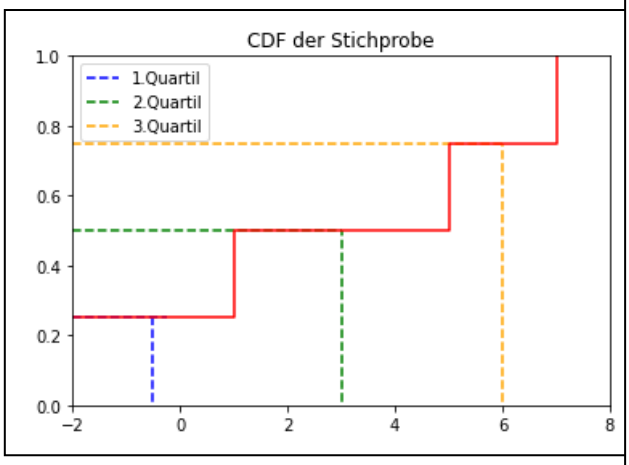

```
#Quantile (aus CDF)
import numpy as np
import matplotlib.pyplot as plt

X=np.array([4, 4, 0, 3, 5, 3, 1])
Xnum,absH =np.unique(X, return_counts=True)
relH=absH/np.size(X)
kumrelH =np.cumsum(relH)

plt.figure()
plt.step(Xnum,kumrelH,where='post',
color='red')
plt.xlim(0, np.max(X)+1)
plt.ylim(0, 1)
plt.vlines(1,0,0.25,color='blue',linestyles='d
ashed',label='1.Quartil') #1.Quartil
plt.hlines(0.25,0,1,color='blue',linestyles='d
ashed')
plt.vlines(3,0,0.5,color='green',linestyles='d
ashed',label='2.Quartil') #Median
plt.hlines(0.5,0,3,color='green',linestyles='d
ashed')
plt.vlines(4,0,0.75,color='orange',linestyles=
'dashed',label='3.Quartil') #3.Quartil
plt.hlines(0.75,0,4,color='orange',linestyles=
'dashed')
plt.legend()
plt.title('CDF der Stichprobe')
```

Das 1.Quartil lautet: 1

Der Median lautet: 3

Das 3. Quartil lautet: 4

Beispiel 13: Quantile (über *CDF*)
Wir betrachten die Stichprobe: $-2,1,5,7$
Wir suchen nun die Quartile,
indem wir die *CDF* betrachten.




```
#Quantile (aus CDF)
import numpy as np
import matplotlib.pyplot as plt

X=np.array([-2,1,5,7])
Xnum,absH =np.unique(X, return_counts=True)
relH=absH/np.size(X)
kumrelH =np.cumsum(relH)

plt.figure()
plt.step(Xnum,kumrelH,where='post',color='red')
plt.xlim(np.min(X), np.max(X)+1)
plt.ylim(0, 1)
plt.vlines(-0.5, np.min(X),0.25,
color='blue',linestyles='dashed',label='1.Quart
il') #1.Quartil
plt.hlines(0.25,np.min(X),-0.25,color='blue',
linestyles='dashed')
plt.vlines(3,np.min(X),0.5,color='green',linest
yles='dashed',label='2.Quartil') #Median
plt.hlines(0.5,np.min(X),3,color='green',linest
yles='dashed')
plt.vlines(6,np.min(X),0.75,color='orange',line
styles='dashed',label='3.Quartil') #3.Quartil
plt.hlines(0.75,np.min(X),6,color='orange',line
styles='dashed')
plt.legend()
plt.title('CDF der Stichprobe')
```

Das 1.Quartil lautet: -0.5

Der Median lautet: 3

Das 3. Quartil lautet: 6

Beispiel 14: Quantile (aus *CDF*)
Zurück zu **Beispiel 9:** Gegeben sind die Nettomieten von 151 Wohnungen (die Tabelle aus Bsp. 9 ist auf der folgenden Seite nochmals zu finden).
Lesen Sie anhand der Tabelle die Quartile und das $\frac{111}{151}$-Quantil sowie den Interquartilsabstand ab

Quartile: Q1 = 731, Q2 = 1025, Q3 = 1480

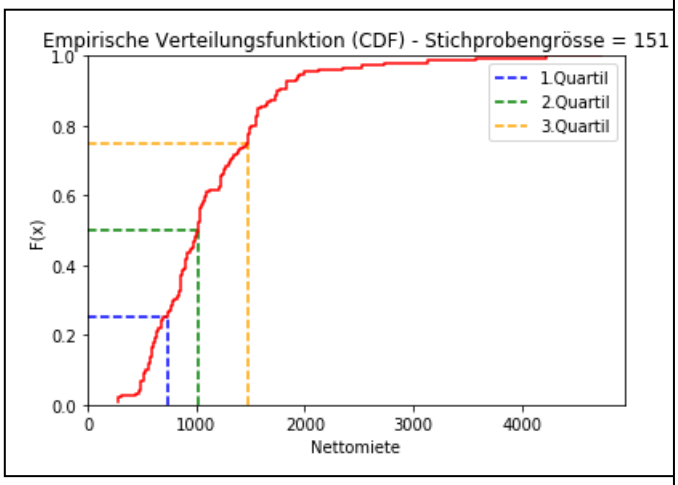
Interquartilsabstand: Q3 – Q1 = 1480 – 731 = 749

0.7351-Quantil: $R_{0.7351}$ = ½ (1430 + 1460) = 1445

a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i	a_i	h_i	H_i	F_i
274	1	1	0.7	672	1	33	21.9	955	1	67	44.4	1250	1	100	66.2	1595	1	128	84.8
281	1	2	1.3	675	1	34	22.5	970	1	68	45	1260	1	101	66.9	1645	1	129	85.4
310	1	3	2	677	1	35	23.2	974	1	69	45.7	1275	1	102	67.5	1660	1	130	86.1
435	1	4	2.6	695	2	37	24.5	980	1	70	46.4	1301	1	103	68.2	1679	1	131	86.8
475	1	5	3.3	731	1	38	25.2	990	1	71	47	1305	1	104	68.9	1725	1	132	87.4
480	1	6	4	749	2	40	26.5	1000	2	73	48.3	1320	1	105	69.5	1730	1	133	88.1
483	1	7	4.6	772	1	41	27.2	1008	1	74	49	1325	1	106	70.2	1745	1	134	88.7
490	1	8	5.3	775	2	43	28.5	1020	1	75	49.7	1343	1	107	70.9	1750	1	135	89.4
510	2	10	6.6	800	2	45	29.8	1025	2	77	51	1375	1	108	71.5	1770	1	136	90.1
515	1	11	7.3	825	1	46	30.5	1035	2	79	52.3	1395	1	109	72.2	1825	1	137	90.7
520	1	12	7.9	840	1	47	31.1	1037	1	80	53	1398	1	110	72.8	1830	1	138	91.4
530	2	14	9.3	850	2	49	32.5	1040	4	84	55.6	1430	1	111	73.5	1835	1	139	92.1
560	1	15	9.9	855	1	50	33.1	1042	1	85	56.3	1460	1	112	74.2	1920	1	140	92.7
565	1	16	10.6	857	1	51	33.8	1050	1	86	57	1470	1	113	74.8	1930	1	141	93.4
570	2	18	11.9	860	1	52	34.4	1070	1	87	57.6	1480	2	115	76.2	1945	1	142	94
580	2	20	13.2	870	4	56	37.1	1080	1	88	58.3	1481	1	116	76.8	2000	1	143	94.7
584	1	21	13.9	876	1	57	37.7	1085	1	89	58.9	1485	1	117	77.5	2130	1	144	95.4
589	2	23	15.2	880	1	58	38.4	1095	1	90	59.6	1510	3	120	79.5	2345	1	145	96
600	2	25	16.6	895	1	59	39.1	1096	1	91	60.3	1547	1	121	80.1	2520	1	146	96.7
605	1	26	17.2	896	1	60	39.7	1120	1	92	60.9	1552	1	122	80.8	2730	1	147	97.4
614	1	27	17.9	900	1	61	40.4	1204	1	93	61.6	1554	1	123	81.5	3130	1	148	98
625	1	28	18.5	907	1	62	41.1	1220	1	94	62.3	1555	1	124	82.1	3575	1	149	98.7
628	1	29	19.2	914	1	63	41.7	1225	2	96	63.6	1560	1	125	82.8	4230	1	150	99.3
639	1	30	19.9	920	2	65	43	1226	1	97	64.2	1565	1	126	83.4	4725	1	151	100
650	2	32	21.2	950	1	66	43.7	1245	2	99	65.6	1570	1	127	84.1	Total		151	

Um die Quantilsfunktion von «numpy» zu nutzen, müssen wir zunächst unsere Liste der Rohdaten generieren:

```
#Mieten
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

#Daten einlesen
df = pd.read_excel('Bsp_1_14.xlsx')
print (df.columns) #Bezeichnung der Datenreihen
Miete=df.values[:,0]
h=df.values[:,1]

#Rohdaten erzeugen
Stichprobe=[]
for k in range (0,Miete.size):
    Stichprobe=np.append([Stichprobe], np.ones(h[k])*[Miete[k]]) #Rohdaten werden generiert

#Quartile
Q1=np.quantile(Stichprobe,0.25)
Q2=np.quantile(Stichprobe,0.5)
Q3=np.quantile(Stichprobe,0.75)
IQR=Q3-Q1
print(IQR)
```



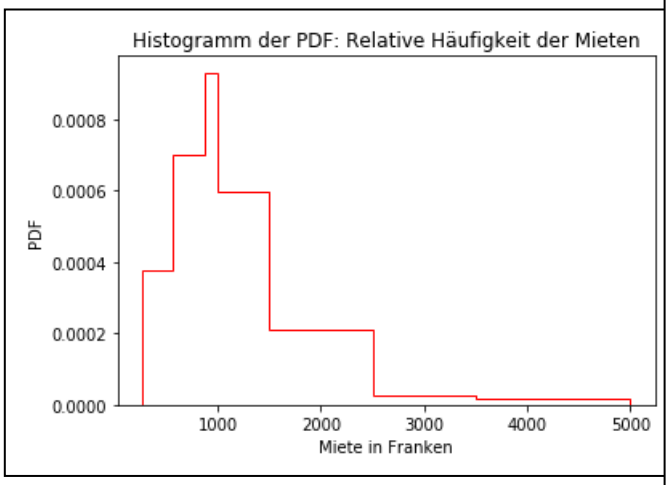

```
n=np.sum(h)  #Stichprobengrösse 151
H=np.cumsum(h)  #kum. abs. Häufigkeit
F=H/n # kum. rel. Häufigkeit

#Nichtklassierte Daten plotten
plt.figure(1)
plt.step(Miete, F,color='red')
plt.xlabel('Nettomiete')
plt.ylabel('F(x)')
plt.ylim(0,1)
plt.xlim(left=0)
plt.vlines(731,0,0.25,color='blue',linesty
les='dashed',label='1.Quartil') #Q1
plt.hlines(0.25,0,731,color='blue',linesty
les='dashed')
plt.vlines(1025,0,0.5,color='green',linest
yles='dashed',label='2.Quartil') #Median
plt.hlines(0.5,0,1025,color='green',linest
yles='dashed')
plt.vlines(1480,0,0.75,color='orange',line
styles='dashed',label='3.Quartil') #Q3
plt.hlines(0.75,0,1480,color='orange',line
styles='dashed')
plt.legend()
plt.title('Empirische Verteilungsfunktion
(CDF) - Stichprobengrösse = {}'.format(n))
```

Nun sollen die Daten klassiert werden mit den Klassengrenzen 270, 570, 870, 1000, 1500, 2500, 3500, 5000. Das Histogramm der *PDF* und die kumulierte Verteilungsfunktion (*CDF*) $F(x)$ der klassierten Daten soll geplottet werden.
Wir generieren per Hand (durch Abzählen) oder mit Python die Tabelle für die klassierten Daten mit den Klassen, absoluten Klassenhäufigkeiten, kumulierten absoluten Klassenhäufigkeiten, relativen Klassenhäufigkeiten und kumulierten relativen Klassenhäufigkeiten.
Diese Punkte werden zu einem Polygon ergänzt um den Graphen von $F(x)$ zu erhalten.

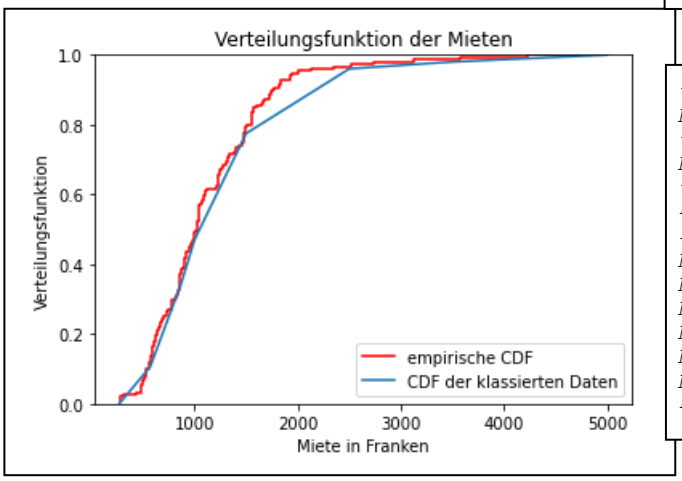
Klasse	AbsH	SummeAbsH	RelH	*CDF*
270-570	16	16	0.106	0.106
570-870	36	52	0.238	0.344
870-1000	19	71	0.126	0.470
1000-1500	46	117	0.305	0.775
1500-2500	28	145	0.185	0.960
2500-3500	3	148	0.020	0.980
3500-5000	3	151	0.020	1

Achtung: Der *CDF* Wert in der rechten Spalte der Tabelle entspricht dem Wert der Verteilungsfunktion $F(x)$ der klassierten Daten am rechten Randpunkt der jeweiligen Klasse. Zum Beispiel gilt

$F(870) = 0.34437$ und $F(270) = 0$




```
#Tabelle der Klassenhäufigkeiten
K=pd.Series(['270-570', '570-870', '870-
1000', '1000-1500', '1500-2500', '2500-
3500', '3500-5000'])
K_gr=[270, 570, 870, 1000, 1500, 2500,
3500, 5000] #Klassengrenzen
#Klassierte Daten als Tabelle ausgeben
hk=np.histogram(Stichprobe,bins=K_gr)[0]
Hk=np.cumsum(hk)
Fk=Hk/n
dfk=pd.DataFrame((hk,Hk,Fk),index=['abs.
Häufigkeit', 'abs. Summenhäufigkeit','rel.
Summenhäufigkeit'], columns=K)
print(dfk)

#Histogramm plotten
plt.figure(2)
plt.hist(Miete,K_gr,density='True',histtype
='step',color='red')
plt.xlabel('Miete in Franken')
plt.ylabel('PDF')
plt.title('Histogramm der PDF: Relative
Häufigkeit der Mieten')
```




```
#Verteilungsfunktion
plt.figure(3)
#empirisch Verteilungsfunktion
plt.step(Miete, F,color='red')
#empirische Verteilungsfunktion
klassierten Daten
Fk_p=np.append(0,Fk)
plt.plot(K_gr,Fk_p)
plt.ylim(0,1)
plt.xlabel('Miete in Franken')
plt.ylabel('Verteilungsfunktion')
plt.title('Verteilungsfunktion der Mieten')
plt.legend(('empirische CDF', 'CDF der
klassierten Daten'),loc='best')
```

Für die klassierten Daten soll nun das $\frac{111}{151}$-Quantil berechnet werden. Man beachte dazu, dass das q-Quantil R_q bei klassierten Daten mit der Umkehrfunktion der Verteilungsfunktion (*CDF*) $F(x)$ als $R_q = F^{-1}(q)$ bestimmt wird.

```
#Berechnung des 0.7351 Quantils der Rohdaten
Q=np.quantile(Stichprobe,0.7351)
print(Q)
```

Aus dem Plot der Verteilungsfunktion von $F(x)$ bzw. aus der generierten Tabelle, genauer der Spalte mit den relativen Häufigkeiten der klassierten Daten, ist ersichtlich, dass das 0.7351-Quantil zwischen 1000 und 1500 liegen muss.

Klasse	AbsH	SummeAbsH	RelH	*CDF*
270-570	16	16	0.106	0.106
570-870	36	52	0.238	0.344
870-1000	19	71	0.126	0.470
1000-1500	46	117	0.305	0.775
1500-2500	28	145	0.185	0.960
2500-3500	3	148	0.020	0.980
3500-5000	3	151	0.020	1

Die Funktionsgleichung von $F(x)$ lautet in diesem Bereich:

$$F(x) = q = \underbrace{F(1000)}_{71/151} + \underbrace{(F(1500) - F(1000))}_{46/151} \cdot \frac{x-1000}{500}, \text{ mit } 1000 \leq x \leq 1500.$$

Die Umkehrfunktion lautet

$$\frac{500}{F(1500) - F(1000)}(q - F(1000)) + 1000 = x$$

also

$$F^{-1}(q) = \frac{500}{F(1500) - F(1000)}(q - F(1000)) + 1000 = \frac{500 \cdot 151}{46}(q - \frac{71}{151}) + 1000$$

Damit berechnen wir das 0.7351-Quantil zu

$$F^{-1}(0.7351) = \frac{500 \cdot 151}{46}(0.7351 - \frac{71}{151}) + 1000 \approx 1434.784.$$

Allgemein verwendet man zur Berechnung der q-Quantile bei klassierten Daten die folgende Formel:

Definition
Bei klassierten Daten mit einer Verteilungsfunktion $F(x)$ ist das q-Quantil R_q definiert als:

$$R_q = F^{-1}(q)$$

Zur Berechnung von R_q sucht man diejenige Klasse $[a, b[$ mit $F(a) \leq q \leq F(b)$.
Dann gilt:

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

Und damit:

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

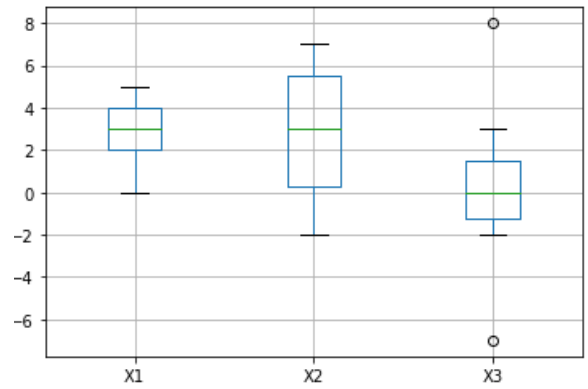
1.4.2 Boxplot

Mit dem Median haben wir ein Lagemass für eine Stichprobe $x_1, x_2, \ldots, x_n$ von reellen Zahlen kennengelernt. Das dazu verwendete Mass für die Streuung (Streuungsmass bzw. *Streumass*) der Stichprobenwerte um den Median, ist der sogenannte *Interquartilsabstand* $Q_3 - Q_1$, mit dem 1. Quartil Q_1 und dem 3. Quartil Q_3. Im Intervall $[Q_1, Q_3]$ befinden sich die Hälfte der Stichprobenwerte mit dem Median innerhalb dieses Intervalls.

Beispiel 15: Boxplot
Geben Sie für die folgenden Stichproben jeweils den Median, das erste Quartil, das dritte Quartil und den Interquartilsabstand an:

a) Stichprobe 1: 4,4,0,3,5,3,1
b) Stichprobe 2: –2,1,5,7
c) Stichprobe 3: –7,–2,–1,–1,1,1,3,8

Im **Boxplot** werden die Kenngrössen Median (2. Quartil) Q_2, erstes Quartil Q_1, drittes Quartil Q_3, das Minimum und das Maximum der Stichprobe graphisch dargestellt:
Für die Stichproben der drei Beispiele sind unten die mit Python erstellten Boxplots im Vergleich zu sehen. Man beachte die kleinen Unterschiede zu den oben berechneten Werten der Quartile. Bei grösseren (realistischen) Stichproben sind solche Abweichungen, die aufgrund der Implementation der Quantile entstehen, unerheblich.

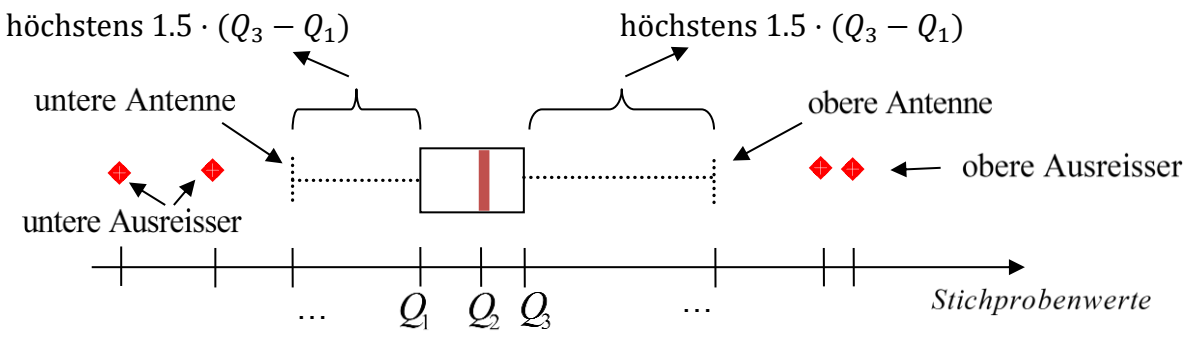



```
#Boxplots mehrerer Stichproben zum Vergleich
import numpy as np
import pandas as pd

X1=np.array([4,4,0,3,5,3,1]) #Stichprobe 1
X2=np.array([-2,1,5,7]) #Stichprobe 2
X3=np.array([-7,-2,-1,-1,1,1,3,8]) #Stichprobe 3

#Arbeiten mit Pandas, da mehrere Datenreihen
verwendet werden
#Fehlende Daten werden mit NaN aufgefüllt
df=pd.DataFrame((X1,X2,X3),index=['X1',     'X2',
'X3'])
df=df.transpose()
df.boxplot()
```

	Q1	Q2	Q3
X1	2.00	3.0	4.0
X2	0.25	3.0	5.5
X3	-1.25	0.0	1.5

```
#Darstellung der Quartile
Q11=np.quantile(X1,0.25)
Q12=np.quantile(X1,0.5)
Q13=np.quantile(X1,0.75)
Q21=np.quantile(X2,0.25)
Q22=np.quantile(X2,0.5)
Q23=np.quantile(X2,0.75)
Q31=np.quantile(X3,0.25)
Q32=np.quantile(X3,0.5)
Q33=np.quantile(X3,0.75)
df_q=pd.DataFrame(([Q11,Q12,Q13],[Q21,Q22,Q23],[Q31,Q32,Q33]),index=['X
1', 'X2', 'X3'],columns=['Q1', 'Q2', 'Q3'])
print(df_q)
```

Die *Box* markiert
den Median, das erste und dritte Quartil (unterer und oberer Rand)
……
Die *untere Antenne* (unterer Whisker) zeigt das Minimum derjenigen Stichprobenwerte, welche
1.5 · (Q3 – Q1) von Q1
nicht mehr als Abstand……..haben.
Die *obere Antenne* (oberer Whisker) zeigt das Maximum derjenigen Stichprobenwerte, welche
1.5 · (Q3 – Q1) von Q3
nicht mehr als Abstand……..haben
Ausreisser sind alle Stichprobenwerte, welche sich ausserhalb des durch die Box und die Antennen markierten Bereichs befinden.



1.4.3 Lagekennwerte

Die Definition und Bedeutung dieser Kennwerte von Stichproben zeigen wir am folgenden Gewinnspiel.

Beispiel 16: Gewinnspiel – empirisch
Ein Spieler zahlt den Einsatz von 1 Franken und würfelt mit 5 Würfeln. Der Gewinn (abzüglich des Einsatzes) wird nach der folgenden Tabelle bestimmt. Die Gewinnwahrscheinlichkeit, respektive der erwartete Gewinn pro Spiel sollen empirisch ermittelt werden. Dazu wird das Spiel 100 Mal durchgeführt. Die Ergebnisse sind in der Tabelle eingetragen.

Resultat	Sonstiges	3er Pasch	Full House	Grosse Strasse	4er Pasch	5er Pasch	Total
Gewinn a_i	-1	0	4	8	10	20	
Absolute Häufigkeit h_i	74	13	3	5	4	1	100
Relative Häufigkeit f_i	0.74	0.13	0.03	0.05	0.04	0.01	1
Kum. abs. Häufigkeit H_i	74	87	90	95	99	100	
Kum. rel. Häufigkeit F_i	0.74	0.87	0.90	0.95	0.99	1	

In der vierten Zeile der Tabelle stehen die ermittelten relativen Häufigkeiten. Diese entsprechen den empirischen Wahrscheinlichkeiten der möglichen Gewinne bei einem Spiel.

Das *arithmetische Mittel* $\bar{x}$ aller Stichprobenwerte, der sogenannte *Stichprobenmittelwert* ist der (empirisch ermittelte) erwartete Gewinn für ein Spiel:

$$\bar{x} = \frac{1}{100}(-1 \cdot 74 + 0 \cdot 13 + 4 \cdot 3 + 8 \cdot 5 + 10 \cdot 4 + 20 \cdot 1) = \frac{38}{100} = 0.38\ [\text{CHF}]$$

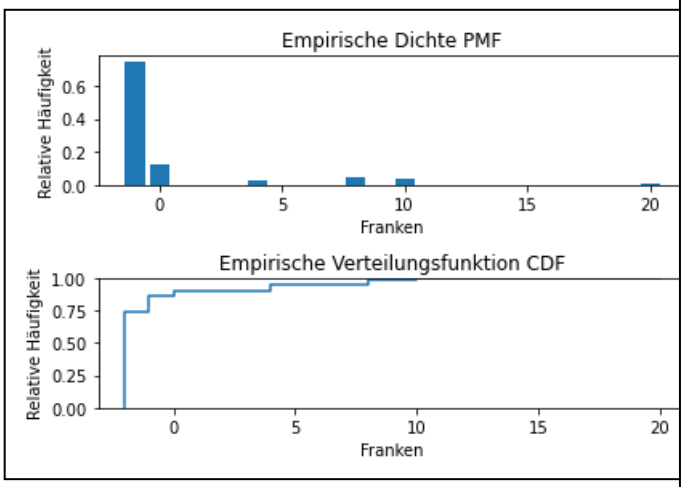
Pro Spiel erwartet man also einen mittleren Gewinn von 0.38 Franken. Bei 100 Spielen ergibt das einen mittleren Gewinn von 38 Franken. Ob diesem empirischen Wert zu trauen ist, ergründen wir später in der Stochastik (s. Abschnitt 4.3).

Wir betrachten noch die weiteren Lagemasse: Der Zentralwert oder *Median* x_{med} als der Wert, der die Stichprobe in 2 gleich grosse Teile teilt, wurde bereits besprochen. Für die vorliegende Stichprobe gilt:

$$Median(x_1, \dots x_n) = x_{med} = -1$$

Ein weiteres Lagemass wäre der sogenannte *Modus* oder *Modalwert* x_{mod}, der das Maximum der Verteilung (*PMF* bzw. *PDF*), d.h. den häufigsten Stichprobenwert angibt. Hier gilt

$$Modus(x_1, \dots x_n) = x_{mod} = -1$$




```
#Gewinnspiel
import numpy as np
import matplotlib.pyplot as plt

G=np.array([-1, 0, 4, 8, 10, 20]) #Gewinne
h=np.array([74, 13, 3, 5, 4, 1]) #absolute
Häufigkeiten
n=np.sum(h) #Stichprobengrösse 100
f=h/n #relative Häufigkeiten (PMF)
H=np.cumsum(h) #kumulierte absolute
Häufigkeiten
F=np.cumsum(f) #kumulierte relative
Häufigkeiten (CDF)

plt.figure()
plt.subplot(2,1,1)
plt.bar(G,f)
plt.title('Empirische Dichte PMF')
plt.xlabel('Franken')
plt.ylabel('Relative Häufigkeit')
plt.subplot(2,1,2)
plt.step(np.append(-2,G),np.append(0,F))
plt.title('Empirische Verteilungsfunktion
CDF')
plt.ylim(0,1)
plt.xlabel('Franken')
plt.ylabel('Relative Häufigkeit')
plt.tight_layout()
```

Definition Lagemasse
Gegeben ist eine Stichprobe $x_1, x_2, \ldots, x_n$ eines metrischen Merkmals mit den verschiedenen Merkmalsausprägungen $a_1, a_2, \ldots, a_m$, den absoluten Häufigkeiten h_i und den relativen Häufigkeiten f_i. Zudem gelte $n = \sum_{i=1}^{m} h_i$. Dann folgt:

Der *Modus* **(**auch *Modalwert***)** x_{mod} ist
der häufigste Wert, der in der Stichprobe vorkommt.
……
Jedoch ist er im Allgemeinen nicht eindeutig.

Das *arithmetische Mittel* (auch *empirischer Mittelwert*) $\bar{x}$ ist derjenige Mittelwert, der als Quotient aus der Summe der betrachteten Zahlen und ihrer Anzahl berechnet ist.

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

mit m = Anzahl der verschiedenen Merkmale und $n = \sum_{i=1}^{m} h_i$

Der *Median* (auch *Zentralwert*) x_{med} teilt eine Stichprobe oder eine Verteilung
so in zwei (gleich grosse) Hälften,
……
dass die Werte in der einen Hälfte nicht grösser als der Medianwert sind, und in der anderen nicht kleiner.

1.4.4 Streuungskennwerte

Weitere Masse charakterisieren die Abweichung vom Zentrum der Verteilung, sogenannte *Streuungsmasse* (bzw. *Streumasse*). Die durchschnittliche quadratische Abweichung vom Mittelwert wird die *empirische Varianz* oder *Stichprobenvarianz* $\tilde{s}^2$ genannt.

Für die Werte aus Bsp. 16 ergibt sich:

$$\tilde{s}^2 = \frac{1}{100}[(-1-\bar{x})^2 \cdot 74 + (0-\bar{x})^2 \cdot 13 + (4-\bar{x})^2 \cdot 3 + (8-\bar{x})^2 \cdot 5 + (10-\bar{x})^2 \cdot 4 + (20-\bar{x})^2 \cdot 1$$
$$= 30689/2500 \approx 12.2756 \text{ [CHF^2]}$$

Die aufwendige Berechnung wird vereinfacht, wenn man die Binome ausmultipliziert und die Mittelwerte miteinander verrechnet. So erhält man folgende, den Rechenaufwand vereinfachende Formel für die (empirische) Stichprobenvarianz:

$$\tilde{s}^2 = \frac{1}{100}[(-1)^2 \cdot 74 + (0)^2 \cdot 13 + (4)^2 \cdot 3 + (8)^2 \cdot 5 + (10)^2 \cdot 4 + (20)^2 \cdot 1] - \bar{x}^2$$
$$= 30689/2500 \approx 12.2756 \text{ [CHF}^2\text{]}$$

Die Stichprobenwerte haben die Einheit Fr, der Mittelwert ebenfalls und die Varianz allerdings das Quadrat der Einheit, also Fr^2. Zieht man die Wurzel aus der Stichprobenvarianz so gelangt man zur *Standardabweichung* $\tilde{s}$ mit der gleichen Einheit Fr wie die Stichprobenwerte:

$$\tilde{s} = \sqrt{\tilde{s}^2} \approx 3.5 \text{ [CHF]}.$$

In der Schätztheorie werden die sogenannte *korrigierte Stichprobenvarianz* s^2 und *korrigierte Standardabweichung* s verwendet. In den Formeln wird statt mit dem Faktor n, mit dem Faktor $n-1$ gemittelt (s. Abschnitt 7.2.4).

Für die *korrigierte Stichprobenvarianz* s^2 und *korrigierte Standardabweichung* s ergibt sich:

$$s^2 = \frac{1}{99}[(-1-\bar{x})^2 \cdot 74 + (0-\bar{x})^2 \cdot 13 + (4-\bar{x})^2 \cdot 3 + (8-\bar{x})^2 \cdot 5 + (10-\bar{x})^2 \cdot 4 + (20-\bar{x})^2 \cdot 1$$
$$= 30689/2475 \approx 12.3996 \text{ [CHF}^2\text{]}$$

$$s = \sqrt{s^2} \approx 3.52 \text{ [CHF]}$$

Bei grossen Stichproben sind die Unterschiede der Stichprobenvarianz und der korrigierten Varianz unerheblich.

```
#Kenngrössen
Mean=1/n*np.sum(G*h) #Mittelwert
Varianz1=1/n*np.sum((G-Mean)**2*h) #Stichprobenvarianz
Varianz2=1/n*np.sum(G**2*h)-Mean**2 #Stichprobenvarianz alternative Formel
StAbw=np.sqrt(Varianz1) #Standardabweichung
kVarianz=1/(n-1)*np.sum((G-Mean)**2*h) #korr. Stichprobenvarianz
kStAbw=np.sqrt(kVarianz) #korr. Standardabweichung
```

Zur Berechnung der Kennwerte per Hand ist es hilfreich die obige Tabelle mit den quadrierten Stichprobenwerten a_i^2 zu erweitern.

Resultat	Sonstiges	3er Pasch	Full House	Grosse Strasse	4er Pasch	5er Pasch	Total
Gewinn a_i	-1	0	4	8	10	20	
a_i^2	1	0	16	64	100	400	
Relative Häufigkeit f_i	0.74	0.13	0.03	0.05	0.04	0.01	1

Der Mittelwert $\bar{x}$ ergibt sich an Hand der erweiterten Tabelle durch Multiplikation der zweiten Zeile mit der vierten Zeile und anschliessender Summation:

$$\bar{x} = \sum_{i=1}^{m} a_i \cdot f_i \qquad \text{mit } m = \text{Anzahl der verschiedenen Merkmale und } n = \sum_{i=1}^{m} h_i$$

$$\bar{x} = (-1 \cdot 0.74 + 0 \cdot 0.13 + 4 \cdot 0.03 + 8 \cdot 0.05 + 10 \cdot 0.04 + 20 \cdot 0.01) = 0.38[\text{CHF}]$$

Die Varianz ergibt sich durch Multiplikation der dritten Zeile mit der vierten Zeile und anschliessender Summation gefolgt von der Subtraktion des Quadrates $\bar{x}^2$des Mittelwerts:

$$\tilde{s}^2 = \left(\sum_{i=1}^{m} {a_i}^2 \cdot f_i\right) - \bar{x}^2$$

$$\tilde{s}^2 = (1 \cdot 0.74 + 0 \cdot 0.13 + 16 \cdot 0.03 + 64 \cdot 0.05 + 100 \cdot 0.04 + 400 \cdot 0.01) - 0.38^2 = 12.2756 \ [\text{CHF}\text{^}2]$$

Definition Streuungsmasse

Gegeben ist eine Stichprobe $x_1, x_2, \ldots, x_n$ eines metrischen Merkmals mit den verschiedenen Merkmalsausprägungen $a_1, a_2, \ldots, a_m$, den absoluten Häufigkeiten h_i und den relativen Häufigkeiten f_i. Zudem gelte $n = \sum_{i=1}^{m} h_i$.

Die (*Stichproben-*)*Varianz* (auch *empirische Varianz*) $\tilde{s}^2$ ist ein Mass dafür, wie sehr die Stichprobenwerte x_i um den Mittelwert $\overline{x}$ streuen. Sie beschreibt die erwartete quadratische Abweichung der Zufallsvariablen von ihrem Erwartungswert.

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

Die Quadratwurzel der Varianz wird *Standardabweichung* $\tilde{s}$ genannt

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

In der Schätztheorie werden die folgenden Masse verwendet: *korrigierte Stichprobenvarianz* s^2 und *korrigierte Standardabweichung* s (s. Abschnitt 7.2.4).

In den Formeln wird statt mit dem Faktor n, mit dem Faktor $n-1$ gemittelt:

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

$$s^2 = \frac{n}{n-1}\tilde{s}^2 \text{ bzw. s} = \sqrt{\frac{n}{n-1}}\tilde{s}$$

Für die Verteilung einer metrischen Stichprobe $x_1, x_2, \ldots, x_n$ gilt nach der sogenannten *Tschebyscheff'schen Ungleichung* (s. Kap.4, S.13), dass mindestens 75% der Werte im Bereich $\bar{x} \pm 2 \cdot s$ sind. In der Praxis sind Stichproben häufig annähernd normalverteilt. Dann sind etwa 68% der Werte im Bereich $\bar{x} \pm s$ und etwa 95% im Bereich $\bar{x} \pm 2s$.
Bei klassierten Daten geht man anders vor, um den Mittelwert und die Varianz zu berechnen. Am Beispiel der Transportausgaben aus Beispiel 7 soll das gezeigt werden.

Beispiel 17: Ausgaben für Transport
Die folgende Tabelle gibt die jährlichen Transportausgaben von 750 Personen wieder. Dabei wurden die Stichprobenwerte in fünf verschiedene Klassen eingeteilt. Zur Berechnung von Mittelwert und Varianz der klassierten Daten werden die Klassenmitten verwendet.

Klasse a_i Von … bis weniger als …	[100,200[	[200,500[	[500,800[	[800,1000[	[1000,2000[	Total
Absolute Häufigkeit h_i des Intervalls	35	182	317	84	132	750
Relative Häufigkeit f_i des Intervalls	35/750	182/750	317/750	84/750	132/750	1
Werte der kumulativen Verteilungsfunktion (*CDF*) $F(x)$ an der rechten Intervallgrenze	35/750 ≈ 0.047	217/750 ≈ 0.29	534/750 ≈ 0.712	618/750 ≈ 0.824	1	
Klassenmitten	150	350	650	900	1500	

Stichprobenmittelwert:

$$\bar{x} = \frac{1}{750}[150 \cdot 35 + 350 \cdot 182 + 650 \cdot 317 + 900 \cdot 84 + 1500 \cdot 132] = 10972/15$$
$$\approx 731.5 \text{ [CHF]}$$

Stichprobenvarianz:

$$\tilde{s}^2 = \frac{1}{750}[150^2 \cdot 35 + 350^2 \cdot 182 + 650^2 \cdot 317 + 900^2 \cdot 84 + 1500^2 \cdot 132] - \bar{x}^2$$
$$\approx 161030\ [\text{CHF}^2]$$

Standardabweichung:

$$\tilde{s} = \sqrt{\tilde{s}^2} \approx 401.29\ [\text{CHF}]$$

Korrigierte Stichprobenvarianz:

$$s^2 = \frac{750}{749}\tilde{s}^2 \approx 161245\ [\text{Franken}^2]$$

Median:

Gesucht ist der Wert $R_{0.5} = F^{-1}(0.5)$ für die klassierten Daten. Dazu nehmen wir die Klasse 500-800 und lösen die Gleichung:

$$\frac{F(800) - F(500)}{800 - 500} = \frac{0.5 - F(500)}{R_{0.5} - 500}$$
$$\Leftrightarrow R_{0.5} = 500 + \frac{300 \cdot (0.5 - F(500))}{F(800) - F(500)} = 500 + \frac{300 \cdot (0.5 - 217/750)}{317/750} \approx 649.53$$

```
#Berechnung von Mittelwert und Varianz bei klassierten Daten
import numpy as np

K_m=np.array([150, 350, 650, 900, 1500]) #Mitten der Klassen
h=np.array([35,  182,  317, 84, 132])  #Absolute Häufigkeiten der Klassen
n=np.sum(h) #Stichprobengrösse
Mean=1/n*np.sum(K_m*h) #Mittelwert
Varianz=1/n*np.sum((K_m-Mean)**2*h) #Varianz
Varianz=1/n*np.sum((K_m**2)*h)-Mean**2 #Varianz alternativ mit weniger Rechenaufwand

#Kennwerte
StAbw=np.sqrt(Varianz) #Standardabweichung
kVarianz=n/(n-1)*Varianz #korrigierte Varianz
kVarianz=1/(n-1)*(np.sum((K_m**2)*h)-n*Mean**2) #korrigierte Varianz alternativ mit weniger
Rechenaufwand
kStAbw=np.sqrt(kVarianz) #korrigierte Standardabweichung
```

```
#Mittelwert und Varianz alternativ berechnet mit Statistikbefehlen
Stichprobe=[]
for k in range (0,K_m.size):
    Stichprobe=np.append([Stichprobe], np.ones(h[k])*[K_m[k]]) #Rohdaten werden generiert

#Numpy-Funktionen auf Rohdaten
Mean_s=np.mean(Stichprobe) #Mittelwert
Varianz_s=np.var(Stichprobe) #Varianz
StAbw_s=np.std(Stichprobe) #Standardabweichung
kVarianz_s= np.var(Stichprobe,ddof=1) #korrigierte Varianz
kStAbw_s=np.std(Stichprobe, ddof=1) #korrigierte Standardabweichung
```

1. Quartil:

$$\frac{F(500)-F(200)}{500-200}=\frac{0.25-F(200)}{R_{0.25}-200}$$

$$\Leftrightarrow R_{0.25}=200+\frac{300\cdot(0.25-F(200))}{F(500)-F(200)}=200+\frac{300\cdot(0.25-35/750)}{182/750}\approx 451.37$$

3. Quartil:

$$\frac{F(1000)-F(800)}{1000-800}=\frac{0.75-F(800)}{R_{0.75}-800}$$

$$\Leftrightarrow R_{0.75}=800+\frac{200\cdot(0.75-F(800))}{F(1000)-F(800)}=800+\frac{200\cdot(0.75-534/750)}{84/750}\approx 867.86$$

Interquartilsabstand:

$$Q_3-Q_1=R_{0.75}-R_{0.25}\approx 416.48$$

```
#Median, 1. Quantil und 3. Quantil
Median=np.median(Stichprobe)
Q1=np.quantile(Stichprobe,0.25)
Q3=np.quantile(Stichprobe,0.75)
IntQuAbs=Q3-Q1

#Modalwert klassierte Daten
max_num=max(h)
Mod=K_m[(h==max_num)]
```

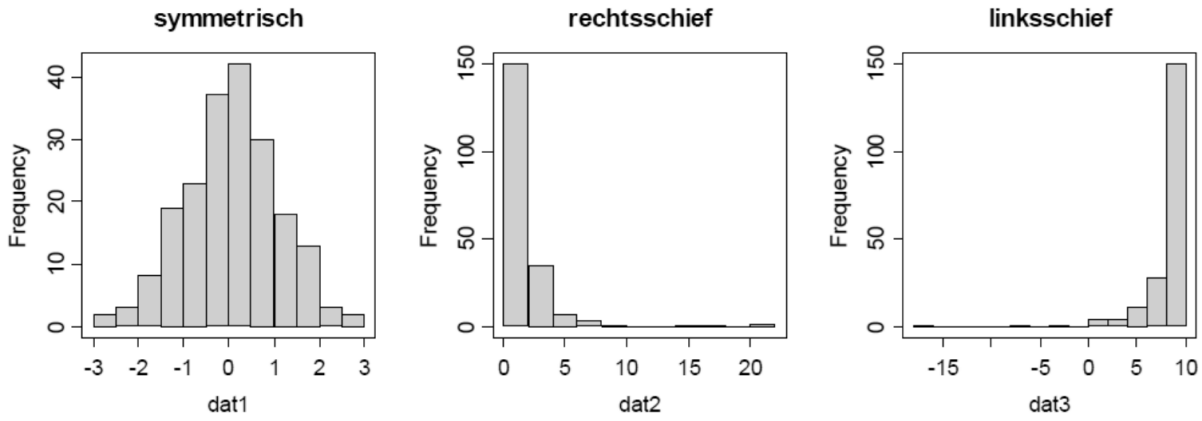
Für den Modus der klassierten Daten wählen wir die Klassenmitte der Klasse mit der grössten relativen Häufigkeit.
Die relativen Häufigkeiten f_i sind maximal für die Klasse 500-800:

$$x_{mod}=650$$

1.4.5 Form der Verteilung

Die Verteilung der Daten kann viele Formen besitzen. Insbesondere unterscheidet man zwischen symmetrischen und schiefen Verteilungen sowie unimodalen und multimodalen Verteilungen.
Eine Verteilung heisst *symmetrisch*, wenn es eine Symmetrieachse gibt, sodass die rechte und die linke Hälfte der Verteilung annähernd zueinander spiegelbildlich sind. Exakte Symmetrie ist bei empirischen Verteilungen selten gegeben.

Deutlich unsymmetrische Verteilungen heissen *schief*. Eine Verteilung ist *linkssteil* (oder *rechtsschief*), wenn der überwiegende Anteil von Daten linksseitig konzentriert ist. Dann fällt die Verteilung nach links deutlich steiler und nach rechts langsamer ab.
Entsprechend heisst eine Verteilung *rechtssteil* (oder *linksschief*), wenn die Verteilung nach rechts steiler und nach links flacher abfällt.



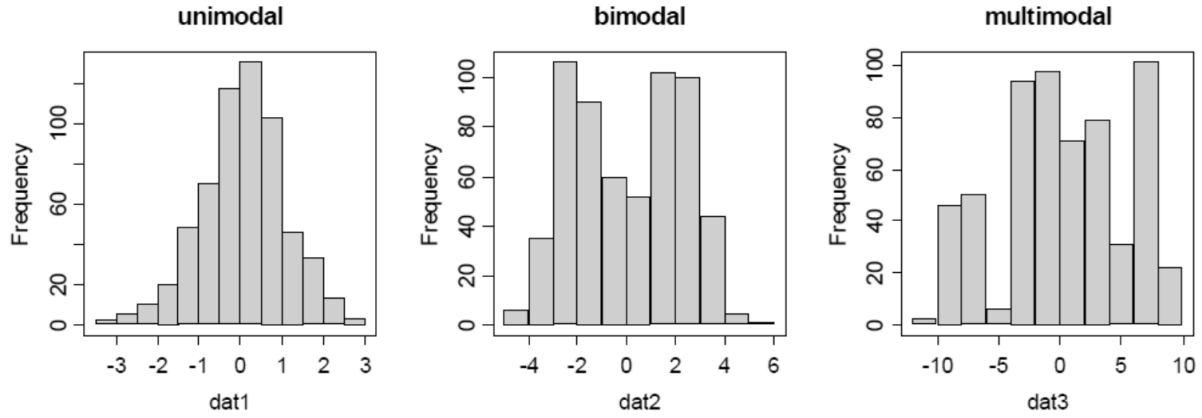
Als Faustregeln kann man für die meisten Verteilungen dabei festhalten:

rechtsschief: $x_{mod} < x_{med} < \overline{x}$

symmetrisch: $x_{mod} = x_{med} = \overline{x}$

linksschief: $x_{mod} > x_{med} > \overline{x}$

Existieren zwei oder mehrere (lokale) Maxima, heisst die Verteilungsfunktion *bimodal* oder *multimodal*; existiert nur ein Maximum, heisst die Verteilungsfunktion *unimodal*.



Bemerkung:
Bei Histogrammen ist bei allen Aussagen zu Schiefe und Modalität die Anzahl Beobachtungen und Klassen entscheidend, mit welcher das Histogramm erzeugt wurde. Je nach gewählter Klassierung müssen wir davon ausgehen, dass die Höhe der Histogramm-Säulen um den Betrag Wert $\pm 2\sqrt{\text{Wert}}$ schwankt. Erst wenn sich bei zwei Gipfeln im Histogramm auch diese sogenannten Variabilitäts-Intervalle nicht mehr überschneiden, können wir generell davon ausgehen, dass diese Bestand haben. So würden wir z.B. im multimodalen Histogramm der absoluten Häufigkeiten oben nur 3 Gipfel zählen. Der kleine Nebengipfel für die Werte in [2,4[hebt sich bei einer Variabilität von ca. $\pm 2\sqrt{80}$ zu wenig stark vom daneben liegenden Hauptgipfel der Säule [2,0[ab.

1.5 Lernziele für dieses Kapitel

- ☐ Ich kann die Begriffe *Stichprobe*, *Stichprobengrösse*, *Grundgesamtheit*, *Merkmal*, *Merkmalstyp* an Beispielen erklären.
- ☐ Ich kenne die Einteilung von Merkmalstypen in *kategoriell*, *metrisch*, *nominal*, *ordinal*, *diskret*, *stetig* und kann diese unterscheiden und mit Beispielen erklären.
- ☐ Ich kann *absolute* und *relative Häufigkeiten* bei Stichproben berechnen.
- ☐ Ich kann den Unterschied *klassierter* und *nichtklassierter Daten* beschreiben.
- ☐ Ich kenne die *Faustregeln* für die Klassierung von Daten.
- ☐ Ich kann die *Dichtefunktion* (*PDF* bzw. *PMF*) für klassierte und nichtklassierte Stichproben berechnen und kenne ihre Eigenschaften.
- ☐ Ich kann die *Dichtefunktion* (*PMF*) für nichtklassierte Stichproben als *Säulen-* und *Stab-Diagramm* visualisieren.
- ☐ Ich kann die *Dichtefunktion* (*PDF*) für klassierte Stichproben als *Histogramm* visualisieren.
- ☐ Ich kann die *kumulative Verteilungsfunktion* (*CDF*) für klassierte und nichtklassierte Stichproben berechnen und kenne ihre Eigenschaften.
- ☐ Ich kann die *kumulative Verteilungsfunktion* (*CDF*) für nichtklassierte Stichproben als monoton wachsende, rechtsseitig stetige Treppenfunktion visualisieren.
- ☐ Ich kann die *kumulative Verteilungsfunktion* (*CDF*) für klassierte Stichproben als monoton wachsenden Polygonzug visualisieren.
- ☐ Ich kann die Begriffe *Lagemasse* und *Streuungsmasse* einer Stichprobe erklären.
- ☐ Ich kann die *p-Quantile* bei nichtklassierten und klassierten Stichproben berechnen.
- ☐ Ich kenne die Begriffe *Median*, *Quartil*, *Perzentile* und *Interquartilsabstand* und kann diese erklären.
- ☐ Ich kann die Definition der *Oberen* und die *Unteren Antenne* beim *Boxplot* erklären.
- ☐ Ich kann alle Daten für den *Boxplot* einer Stichprobe bestimmen und den Boxplot zeichnen.
- ☐ Ich kann die Definitionen des *Stichprobenmittelwerts*, der *Stichprobenvarianz* und der *Stichprobenstandardabweichung* für nichtklassierte und klassierte Stichproben aufschreiben.
- ☐ Ich kann den *Stichprobenmittelwert*, die *Stichprobenvarianz* und die *Stichprobenstandardabweichung* für nichtklassierte und klassierte Stichproben berechnen.
- ☐ Ich kann die Definition der *korrigierten Stichprobenvarianz* und der *korrigierten Stichprobenstandardabweichung* für nichtklassierte und klassierte Stichproben aufschreiben.

- ☐ Ich kann die *korrigierte Stichprobenvarianz* und die *korrigierte Stichprobenstandard-abweichung* für nichtklassierte und klassierte Stichproben berechnen.
- ☐ Ich kenne die Definition des *Modus* und kann diesen bestimmen.