A10. Statistik

10.1 Allgemeines – Was ist Statistik?

1. Daten sammeln: Durch Umfragen, Zählung, Messung, . . .
2. Daten präsentieren: Tabellen, Grafiken
3. Daten beschreiben/charakterisieren: durch Mittelwert, Maximum, Minimum. . .
4. Auf eine Grundgesamtheit schließen: Hypothese testen
5. Zusammenhänge finden: Lineare Regression, ...

1.-3.: Beschreibend
4.-5.: Schließend

Warum Datenanalyse?
Um einen Ist-Zustand zu beschreiben!
Um Entscheidungen (bei Unsicherheit) treffen zu können!
Um Hypothese zu testen!
Um von einer Stichprobe auf die Grundgesamtheit zu schließen!
Um Prognosen erstellen zu können!

10.2 Skalen

Skala: Einteilung und Reihung von Werten

Nominalskala (qualitativ)
Besitzt ein Merkmal für das eine Reihung der Ausprägungen nicht möglich bzw. sinnvoll ist.
Die Merkmalsausprägungen heißen **Kategorien**.

z.B: Geschlecht (m,w), Zustand (Ja, Nein, . . .), Fellfarbe (. . .), . . .

Ordinalskala (qualitativ)
Zwischen einzelnen Merkmalsausprägungen gibt es eine natürliche Rangordnung. Differenzen lassen sich aber nicht quantifizieren.

z.B: Ausbildung (Lehre, Fachschule,…), Prüfungsnoten (1,2,3,4,5), Güteklassen (A,B,C), . . .

Metrische Skala (quantitativ)
Merkmalsausprägungen lassen sich ordnen. Und es können Abstände numerisch angegeben werden.
Es gibt meist viele (verschiedene) Merkmalsausprägungen.
Z.B.: Körpergröße, Alter, Masse, . . .

Skalenniveaus bilden eine Hierarchie, i.e., ein metrisch skaliertes Merkmal ist auch ordinal- oder nominalskaliert, aber nicht umgekehrt.

10.3 Präsentation qualitativer Daten

10.3.1 Tabellarisch - Häufigkeitstabelle

1. Häufigkeitstabelle eines Merkmals gibt die Kategorien (Merkmalsausprägungen) und die Zahl der Elemente pro Kategorie an.
2. Erhält man, indem die Antworten den Kategorien zugeordnet werden (Strichliste).
3. Angegeben werden die absoluten oder relativen Häufigkeiten (Prozentsätze), oder beides.

Beispiel:
Beobachtung von 200 Tieren:

Fellfarbe	Anzahl	Anteil
hellbraun	130	65%
mittelbraun	20	10%
mit schwarz	50	25%
Gesamt	200	100%

Wobei: Fellfarbe = Merkmal,
hellbraun…= Kategorie (Ausprägung),
Anzahl aus Strichliste

10.3.2 Graphisch

a) Balkendiagramm (siehe Kapitel 9)

b) Tortendiagramm (siehe Kapitel 9)

10.4 Präsentation quantitativer Daten

Präsentation/Beschreibung der Verteilung, d.h. mit welcher Häufigkeit kommen bestimmte Daten vor.

10.4.1 Tabellarisch

a) Häufigkeitsverteilung

Vorgangsweise bei der Klasseneinteilung:
1. Bereich (kleinsten und größten Wert) bestimmen.
2. Geeignete Zahl an Klassen auswählen.
(Anzahl der Klassen: Wurzel aus Anzahl der Rohdatenwerte, Typischerweise zwischen 5 und 15)
3. Klassenbreite bestimmen. Die Klassen sind in der Regel gleich breit.
4. Klassengrenzen bestimmen.
5. Klassenmitte berechnen.
6. Beobachtungen abzählen und den Klassen zuordnen.

Beispiel: **Einteilung in Klassen**

Rohdaten: 24, 26, 15, 21, 27, 27, 30, 45, 32, 38

Klasse	Anzahl	Anteil	Prozent
15 – 25	3	0,3	30%
25 – 35	5	0,5	50%
35 – 45	2	0,2	20%
Summe	10	1,0	100%

Konvention: 15 bis unter 25, d.h., 15 gehört zur Klasse, 25 nicht.

b) Tabellarisch - Relative Summenhäufigkeitsverteilung (Kumuliertehäufigkeit)

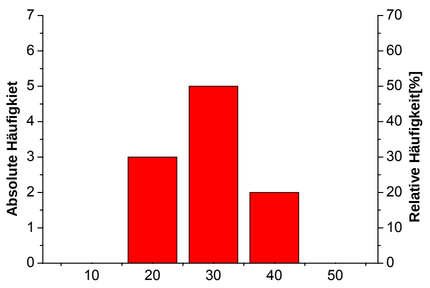
Gibt an, wie viele Beobachtungen – relativ oder prozentuell – einen Wert besitzen, der sich links der jeweiligen oberen Klassengrenze befindet.
(Links bedeutet: Kleiner als die obere Klassengrenze.)

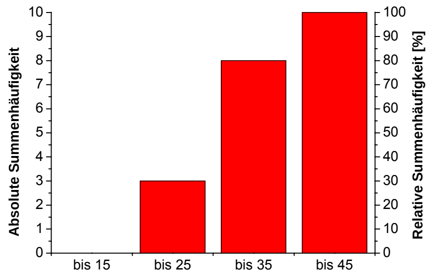
Beispiel: **Summenhäufigkeiten**

Rohdaten: 24, 26, 24, 21, 27, 27, 30, 41, 32, 38

Klassen	Anteil	Summe relativ	Summe prozentuell
15 – 25	0.3	0.3	30%
25 – 35	0.5	0.8	80% (30%+50%)
35 – 45	0.2	1.0	100% (30%+50%)+20%

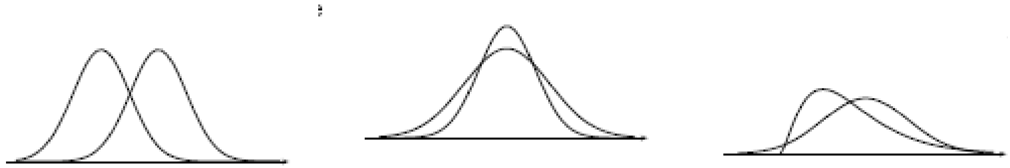
10.4.2 Graphisch - Histogramm





10.5 Numerische Beschreibung/Charakterisierung quantitativer Daten

= Numerische Charakterisierung einer Verteilung



Lage: Lagemaße (Geben die „Mitte“ der Verteilung an)
- arithmetisches Mittel
- Median
- Modus

Streuung: Streuungsmaße (Geben die „Breite“ und „Höhe“ einer Verteilung an)
- Spannweite
- Interquartilsabstand
- Varianz
- Standardabweichung

Form: Formmaße (Geben die „Form“ der Verteilung an)
- Schiefe

10.5.1 Lagemaße

10.5.1.1 (Arithmetisches) Mittel
- Maß für die Lage der Verteilung: Für metrisch skalierte Daten
- Am häufigsten verwendetes Maß
- Auch „Durchschnitt“
- „Durchschnittlicher Wert“ der Daten
- Empfindlich gegen „Ausreißer“ (Werte die deutlich anders sind als der Rest, z.B. Messfehler, Dezimalstelle,…)
- Achtung: Mittelwert hat die gleiche Einheit wie die Messwerte und muss natürlich angegeben werden!

Definition:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

TABELLENKALKULATIONSPROGRAMM:

=MITTELWERT(Wert_1; Wert_2; …Wert_n)
Bzw.
=MITTELWERT(Zelle_1:Zelle_n)

Oder
=SUMME(Zelle_1:Zelle_n) / ANZAHL (Zelle_1:Zelle_n)

Beispiel
Urinabgabe: Tier1:10.3ml, Tier2: 4.9ml, Tier3: 8.9ml, Tier4: 11.7ml, Tier5: 6.3ml, Tier6: 7.7ml

1. Anzahl: n=6
2. Summe: (10.3+4.9+8.9+11.7+6.3+7.7)
3. Ergebnis: (10.3+4.9+8.9+11.7+6.3+7.7) ml / 6 = 8.30 ml

10.5.1.2 Gewichtetes Mittel

Notwendig zur Berechnung des arithmetischen Mittels aus einer Häufigkeitstabelle.

k verschiedene Werte $x_1, \ldots, x_k$ mit den Häufigkeiten $h_1, \ldots, h_k$.

Definition

$$\boxed{\bar{x} = \sum_{i=1}^{k} x_i \frac{h_i}{n} = \frac{1}{n} \sum_{i=1}^{k} x_i h_i} \qquad \text{wobei } n = \sum_{i=1}^{k} h_i$$

Beispiel
Es wurde die Wurfgröße einer Tierart in einem bestimmten Zeitraum erhoben:
Wie groß ist die mittlere Wurfgröße?

Wurfgröße x_i	0	1	4	3	4
Häuigkeit h_i	12	14	5	2	0

Lösung:
Summe der Häufigkeiten: n = 12+14+5+2 = 33 (Anzahl der „Muttertiere“)

$$\bar{x} = \frac{1}{33}\,(0 \cdot 12 + 1 \cdot 14 + 2 \cdot 5 + 3 \cdot 2) = \frac{30}{33} = 0.909$$

10.5.1.3 Median

Maß für die Lage der Verteilung
für ordinalskalierte Daten

Definition:

Mittlerer Wert in der geordneten Liste

Ungerade Anzahl *n*: Mittlerer Wert
Gerade Anzahl *n*: Durchschnitt der beiden mittleren Werten

Position des Medians= (*n*+1)/2

Robust (unempfindlich) gegenüber Ausreißern.

Beispiel: Ungerade Anzahl von Daten:
Rohdaten: 24.1 22.6 21.5 23.7 22.6
Sortiert : 21.5 22.6 22.6 23.7 24.1
Position : 1 2 **3** 4 5
Position des Median= (n+1)/2 = (5+1)/2 = 3
Median = 22.6

Beispiel: Gerade Anzahl von Daten:
Rohdaten: 10.3 4.9 8.9 11.7 6.3 7.7
Sortiert: 4.9 6.3 7.7 8.9 10.3 11.7
Position: 1 2 **3 4** 5 6
Position des Median = (6+1) / 2 = 3.5
Median=(7.7+8.9) /2 = 8.3

TABELLENKALKULATIONSPROGRAMM:
=MEDIAN(Wert_1; Wert_2; …; Wert_n)
bzw
=MEDIAN(Zelle_1:Zelle_n)

10.5.1.4 Modus

Maß für die Lage der Verteilung
Kann auch für nominalskalierte Daten verwendet werden

Definition:
Häufigster Wert

Kein, ein oder mehrere Modi sind möglich
Kann auch bei qualitativen und quantitativen Daten verwendet werden
Robust (unempfindlich) gegenüber Ausreißern

Beispiele:

Kein Modus:
Rohdaten: 10.3 4.9 8.9 11.7 6.3 7.7

Ein Modus:
Rohdaten 6.3 **4.9** 8.9 6.3 **4.9** **4.9**

Mehrere Modi:
Rohdaten: 21 **28** **28** 41 **43** **43**

10.5.1.5 Zusammenfassung

Begriff	Formel	Beschreibung	Skalenniveau
Mittel (arithm.)	$\bar{x} = \frac{\sum x_i}{n}$	„Durchschnitt“	metrisch
Median	Position $= \frac{n+1}{2}$	mittlerer Wert bei geordneten Daten	ordinalsk.
Modus		häufigster Wert	nominalsk.

10.5.1.6 Beispiele

Beispiel 1:
Beschreiben Sie die Lage der Verteilung (Mittel, Median, Modus)
Stimmen die Lagemaße gut überein?
Rohdaten: 17g, 16g, 21g, 18g, 13g, 16g, 12g und 11g

Mittel = (17g+16g+21g+18g+13g+16g+12g+11g) / 8 =15.5 g

Median:

Sortiert: 11g 12g 13g 16g 16g 17g 18g 21g
Position: 1 2 3 **4** **5** 6 7 8 (Anzahl gerade)
Position des Medians = (8+1) / 2 = 4.5
Median = (16g+16g) / 2 =16g

Modus:

Sortiert: 11g 12g 13g **16g** **16g** 17g 18g 21g
Modus = 16g

Vergleich: Mittel=15.5g, Median=16g, Modus=16g => Gute Übereinstimmung

Beispiel 2:
Beschreiben Sie die Lage der Verteilung (Mittel, Median, Modus)
Stimmen die Lagemaße gut überein?
Rohdaten: 17g, 33g, 25g, 21g, 13g, 11g, 12g und 11g

Mittel: Mittel = (17g+33g+25g+21g+13g+11g+12g+11g) / 8 =17.9 g
Median:

Sortiert: 11g 11g 12g 13g 17g 21g 25g 33g
Position: 1 2 3 **4** **5** 6 7 8 (Anzahl gerade)
Median = (13g+17g) / 2 =15g

Modus:

Sortiert: **11g 11g** 12g 13g 17g 19g 25g 33g
Modus = 11g

Vergleich: Mittel=17.9g, Median=15g, Modus=11g => Keine gute Übereinstimmung

10.5.2 Streuungsmaße

10.5.2.1 Spannweite
Maß für die Streuung metrisch skalierter Merkmale
Definition:
Spannweite ist die Differenz zwischen größter (x_{max}) und kleinster (x_{min})Beobachtung

Spannweite = $x_{max} - x_{min}$

Nachteil: Sehr empfindlich gegenüber Ausreißern

TABELLENKALKULATIONSPROGRAMM:

Maximalwert: =MAX(Wert_1; Wert_2; …; Wert_n) bzw. =MAX(Zelle_1:Zelle_n)

Minimalwert: =MIN(Wert_1; Wert_2; …; Wert_n) bzw. =MIN(Zelle_1:Zelle_n)

10.5.2.2 Varianz

Maß für die Streuung für metrisch skalierte Merkmale
Berücksichtigt die Verteilung
Zeigt die Abweichung von Mittelwert x bzw. vom Erwartungswert μ

Definition der Varianz:

Grundgesamtheit

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$$

Stichprobe

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

Die Varianz berechnet sich leichter durch folgende Formeln (Verschiebungssatz):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N}x_i^2}{N} - \mu^2$$

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

10.5.2.3 Standardabweichung

Maß für die Streuung für metrisch skalierte Merkmale
Berücksichtigt die Verteilung
Die Standardabweichung ist die positive Quadratwurzel der Varianz.
Die Standardabweichung hat die gleiche Dimension wie Mittelwert.

Definition der Standardabweichung:

Grundgesamtheit

$$\boxed{\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}}$$

Stichprobe

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

Beispiel
Rohdaten: 10.3 4.9 8.9 11.7 6.3 7.7
Varianz in der Stichprobe:

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \qquad \text{wobei} \qquad \bar{x} = \frac{\sum_{i=1}^{n}x_i}{n} = 8.300 \\ &= \frac{(10.3-8.3)^2+(4.9-8.3)^2+\ldots+(7.7-8.3)^2}{6-1} \\ &= 6.368 \end{aligned}$$

Standardabweichung in der Stichprobe:

$$s = \sqrt{s^2} = \sqrt{6.368} = 2.523$$

Beispiel
Rohdaten: 10.3 4.9 8.9 11.7 6.3 7.7
Varianz in der Stichprobe:

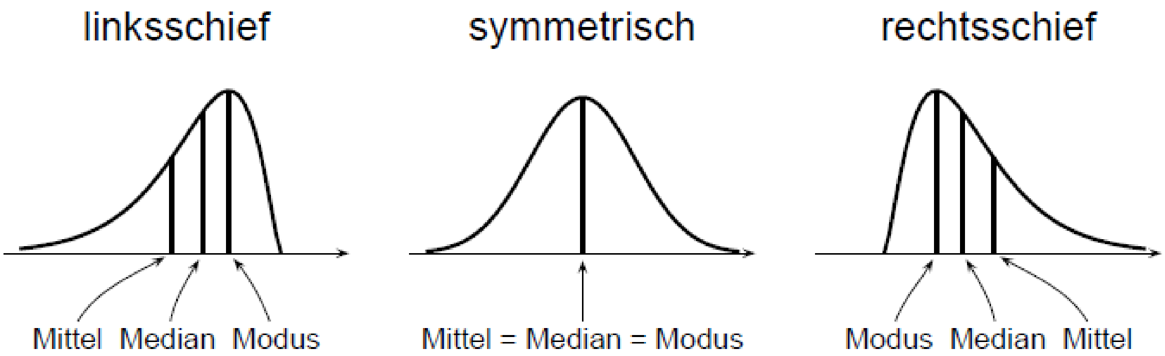
$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n}x_i^2}{n-1} - \frac{n}{n-1}\bar{x}^2 \qquad \text{wobei} \qquad \bar{x} = \frac{\sum_{i=1}^{n}x_i}{n} = 8.300 \\ &= \frac{10.3^2+4.9^2+8.9^2+11.7^2+6.3^2+7.7^2}{6-1} - \frac{6}{6-1}8.3^2 \\ &= 6.368 \end{aligned}$$

10.5.2.4 Zusammenfassung

Begriff	Formel	Beschreibung
Spannweite	$x_{\max} - x_{\min}$	maximale Streuung
Standardabweichung (Stichprobe)	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$	Streuung um das Stichprobenmittel
Standardabweichung (Grundgesamtheit)	$\sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$	. . . um das Mittel in der Grundgesamtheit
Varianz (Stichprobe)	$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$	Quadrierte Streuung um das SP-Mittel
Varianz (Grundgesamtheit)	$\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	. . . um das Mittel in der Grundgesamtheit

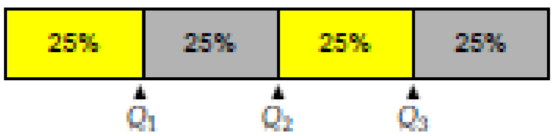
10.5.3 Form der Verteilung

Form: beschreibt die Verteilung der Daten
Maß für die Form ist die **Schiefe** (= Abweichung von der Symmetrie)



10.5.3.1 Quartile

Beschreiben die Verteilung der Daten x_j, $j=1, \ldots, n$.
Die **geordneten** Daten werden in **vier** gleiche Teile zerlegt.
Das i-te Quartil ist der maximale x-Wert des i-ten Teiles:



$$\text{Position des } i\text{-ten Quartils} = i\,\frac{n+1}{4}, \qquad i = 1, 2, 3$$

1/4 aller Daten sind kleiner oder gleich Q_1.
1/2 aller Daten sind kleiner oder gleich Q_2 (=Median).
3/4 aller Daten sind kleiner oder gleich Q_3.

Beispiel
Rohdaten: 10.3, 4.9, 8.9, 11.7, 6.3, 7.7
Sortiert: 4.9, 6.3, 7.7, 8.9, 10.3, 11.7
Position: 1 $\boxed{2}$ $\boxed{3 \quad 4}$ $\boxed{5}$ 6

$$Q_1\text{-Position} = \frac{1 \cdot (n+1)}{4} = \frac{1 \cdot (6+1)}{4} = 1.75 \simeq 2$$

$$Q_1 = 6.3$$

$$Q_2\text{-Position} = \frac{2 \cdot (n+1)}{4} = \frac{2 \cdot (6+1)}{4} = 3.5$$

$$Q_2 = \frac{7.7 + 8.9}{2} = 8.3$$

$$Q_3\text{-Position} = \frac{3 \cdot (n+1)}{4} = \frac{3 \cdot (6+1)}{4} = 5.25 \simeq 5$$

$$Q_3 = 10.3$$

Tabellenkalkulation

1.Quartil:	=QUARTILE(Wertebereich;1)		
2.Quartil:	=QUARTILE(Wertebereich;2)	oder:	=MEDIAN(Wertebereich)
3.Quartil:	=QUARTILE(Wertebereich;3)		

10.5.3.2 Interquartilsabstand

Streuungsmaß für metrisch skalierte Daten
Definition:
Differenz zwischen 3. und 1. Quartil

$$\boxed{\text{Interquartilsabstand} = Q_3 - Q_1}$$

Zeigt den Bereich der mittleren 50% der Daten
Robust (unempfindlich) gegenüber Ausreißern

Beispiel 1

Rohdaten:	10.3	4.9	8.9	11.7	6.3	7.7
Sortiert:	4.9	6.3	7.7	8.9	10.3	11.7
Position:	1	$\boxed{2}$	3	4	$\boxed{5}$	6
		Q_1			Q_3	

$Q_1 = 6.3$
$Q_3 = 10.3$
Interquartilsabstand: Q_3-Q_1=10.3 - 6.3 = 4.0

Beispiel 2
Sie haben in den letzten 8 Tagen folgenden Futtermittelbedarf festgestellt: Analysieren Sie die Verteilung mit Hilfe der Quartile Q_1 und Q_3 sowie mit dem Interquartilsabstand.

Rohdaten:	17kg	16kg	21kg	18kg	13kg	16kg	12kg	11kg
Sortiert:	11kg	12kg	13kg	16kg	16kg	17kg	18kg	21kg
Position:	1	$\boxed{2}$	3	4	5	6	$\boxed{7}$	8

Q_1,Q_3:

$$(\text{Position} = \frac{1\cdot(n+1)}{4} = \frac{1\cdot(8+1)}{4} = 2.25 \cong 2)$$

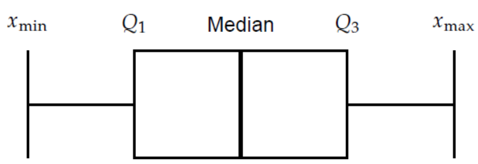
$$(\text{Position} = \frac{3\cdot(n+1)}{4} = \frac{3\cdot(8+1)}{4} = 6.75 \cong 7)$$

Interquartilsabstand: Q_3-Q_1=18kg-12kg=6kg

Tabellenkalkulation
=QUARTILE(Wertebereich;3) - QUARTILE(Wertebereich;1)

10.5.3.3 Box-Plot

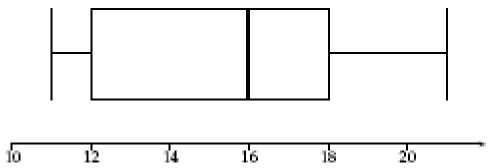
= Graphische Darstellung der Daten
verwendet 5 Kenngrößen:
x_{min}, Q_1, Median, Q_3, x_{max}



Beispiel

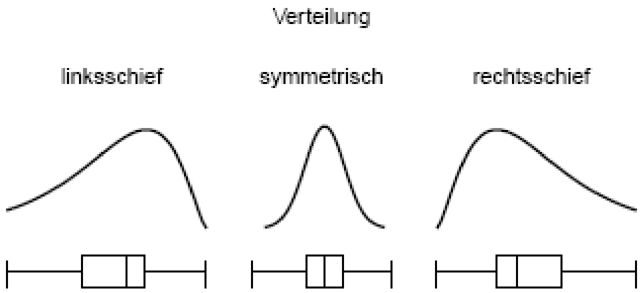
Rohdaten:	17kg	16kg	21kg	18kg	13kg	16kg	12kg	11kg
Sortiert:	11kg	12kg	13kg	16kg	16kg	17kg	18kg	21kg
Position:	$\boxed{1}$	$\boxed{2}$	3	$\boxed{4}$	$\boxed{5}$	6	$\boxed{7}$	$\boxed{8}$

n=8,
X_{min}= 11kg
Q_1= 12kg (Pos. : 1*9/4=2.25=2)
Median= 14kg
Q_3= 18kg (Pos.: 3*9/4=6.75=7
X_{max}= 21kg



10.5.3.4 Bestimmung der Form

Linkschief: Median liegt näher bei Q3 als bei Q1
Symmetrisch: Median liegt zwischen Q3 und Q1
Rechtsschief: Median liegt näher bei Q1 als bei Q3



10.5.3.5 Quantile

Beschreiben die Verteilung der Daten x_j, $j=1, \ldots, n$.
Die **geordneten** Daten können in **beliebig viele** gleiche Teile zerlegt.
Das *i*-te Quantil ist der maximale *x*-Wert des *i*-ten Teiles

Besondere Quantile: Percentil (100er), Dezil (10er-Teilung)

Tabellenkalkulation

=QUANTIL(Wertebereich;Alpha)
Alpha = [0,1]
z.B.:Alpha 0.25-> 1. Quartil

10.5.3.6 Interquantilsabstände

Interquantilsabstände sind analog zum Interquartilabstand definiert:

Interdezil: $Q_{90\%}$-$Q_{10\%}$ … Wertebereich in dem 80% aller Werte liegen (α2=0.9, α1=0.1)

Interpercentil: $Q_{99\%}$-$Q_{1\%}$ … Wertebereich in dem 98% aller Werte liegen (α2 =0.99, α1=0.01)

Im Gegensatz zum Quartileinteilung benötigt die Quantileinteilung mehr Werte und zwar mindestens so viele wie Unterteilungsintervalle (Dezil 10, Percentil 100)

10.5.3.7 Anwendung von Quantil und Quantilsabstände

Was ist „normal“? was ist außergewöhnlich?

Häufig gefragt:
Intervall in dem 75% aller Werte liegen? => Abstand=0.75
Intervall in dem 95% aller Werte liegen? => Abstand=0.95
Intervall in dem 99% aller Werte liegen? => Abstand=0.99

Frage:
$\alpha 1$=?, $\alpha 2$=?

Es gilt: $\alpha 2 = 1.0 - \alpha 1$
Abstand = $\alpha 2 - \alpha 1$

$\alpha\ 1 = (1.0 - \text{Abstand})/2$
$\alpha 2 = 1.0 - \alpha 1$
$\alpha 2 = (\alpha 1 + \text{Abstand})$

Intervall in dem 75% aller Werte liegen?
- ⇨ Abstand=0.75,
- ⇨ $\alpha 1 = (1.0-0.75)/2.0 = 0.125$,
- ⇨ $\alpha 2 = 0.125+0.75= 0.875$

Intervall in dem 95% aller Werte liegen?
- ⇨ Abstand=0.95,
- ⇨ $\alpha 1 = (1.0-0.95)/2.0 = 0.025$,
- ⇨ $\alpha 2 = 0.025+0.95= 0.975$

Intervall in dem 99% aller Werte liegen?
- ⇨ Abstand=0.99,
- ⇨ $\alpha 1 = (1.0-0.99)/2.0 = 0.005$,
- ⇨ $\alpha 2 = 0.005+0.99= 0.995$