Statistik – Univariate Deskriptivstatistik

Häufigkeitsverteilungen

- Bei einer Erhebung wird an n Untersuchungseinheiten ein Merkmal X erfasst, d.h. jeder Einheit kann eine Ausprägung des Merkmals zugewiesen werden

- Die vorliegenden n Merkmalswerte x_i stellen die Beobachtungsreihe bzw. Urliste dar

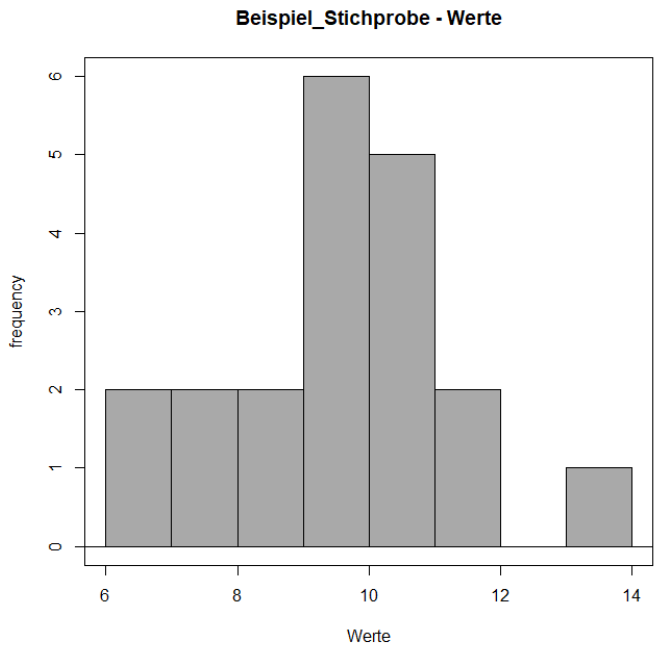
- Beispiel: 20 Messwerte Maßabweichung Welle (9,36 ; 10,83 ; 10,03 ; 10,54 ; 11 ; 9,09 ; 9,8 ; 9,5 ; 11,14 ; 8,82 ; 7,85 ; 9 ; 6,96 ; 9,28 ; 7,39 ; 10,77 ; 13,12 ; 11,35 ; 6,92 ; 9,85) (Beispiel_Stichprobe.xlsx)

- Wenig Aussagekraft

Häufigkeitsverteilungen

Wir benötigen andere Darstellungsformen, aus denen die wichtigen Informationen sofort zu erkennen sind und die Daten vergleichbar machen

Beispiel:
Mittelwert = 9,6300
Standardabweichung (SD) = 1,5781



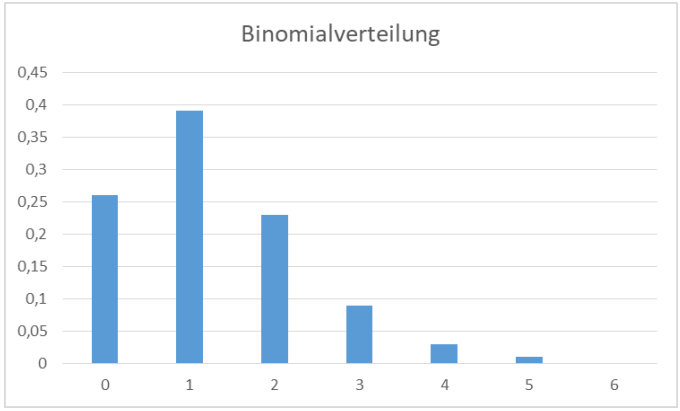
Deskriptivstatistik

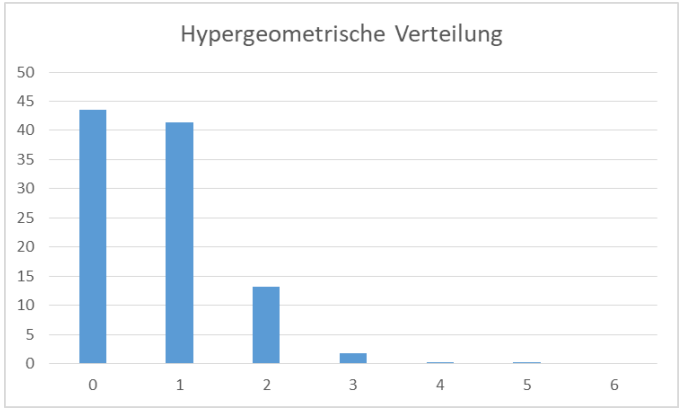
- Daten werden in geeigneter Weise beschrieben, aufbereitet und zusammengefasst
- Verdichtung von quantitativen Daten zu Tabellen, grafischen Darstellungen und Kennzahlen
- Bei Stichproben erfolgt kein Schluss auf die Grundgesamtheit
- Überprüfung der Daten auf mögliche Fehler, fehlende Werte, Auffälligkeiten
- Planung der weiteren Untersuchung

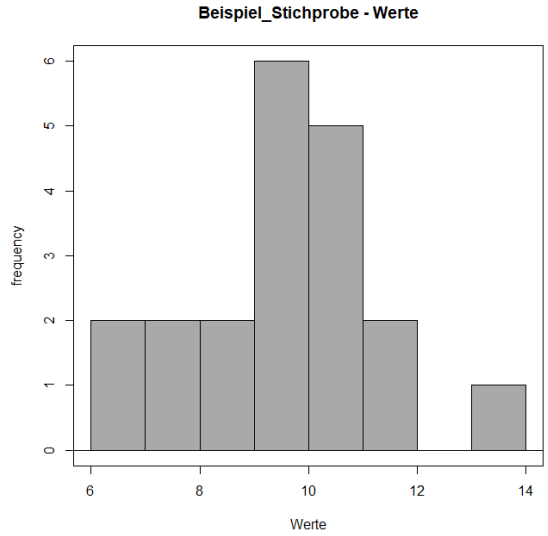
Univariat

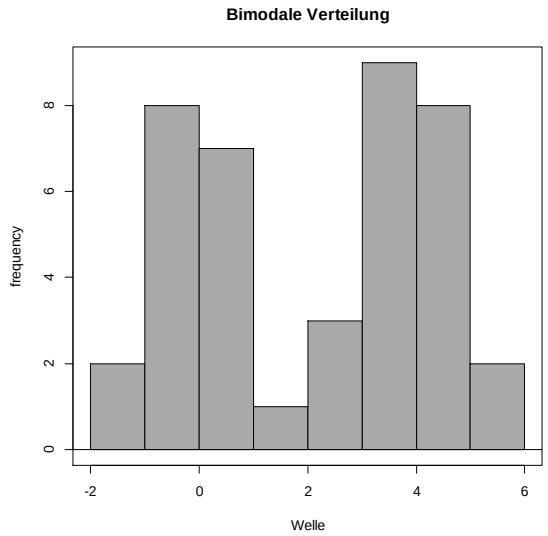
- Merkmal wird als eindimensional angesehen, d.h. es wird als unabhängig von anderen Größen untersucht
- Beispiel: Es liegen zwar Angaben zu Körpergewicht und Größe vor, beide Merkmale werden aber erst einmal unabhängig voneinander beschrieben
- Erst bei einer bivariaten oder multivariaten Beschreibung wird auf mögliche Zusammenhänge eingegangen

Häufigkeitsverteilungen









Häufigkeitsverteilungen

Aus den unterschiedlichen Formen, die eine Verteilung annehmen kann, wird ersichtlich, dass Mittelwert bzw. Standardabweichung nicht alle möglichen Verteilungsformen sinnvoll beschreiben.

Wir benötigen verschiedene Kennwerte, die Aussagekraft haben, um Lage und Streuung unserer Daten zu beschreiben.

Dazu zählen:

Lage:	Mittelwert, Median, Modus und Quantile
Streuung:	Standardabweichung, Spannweite und Quantile

Lagemaße

Arithmetischer Mittelwert (i.Allg. *Mittelwert*)

- Lagemaß für metrisch skalierte Größen
- Durchschnittlicher Wert der Einzelwerte eines Datensatzes
- Reagiert empfindlich auf *Ausreißer*
- Bei Betrachtung einer Population wird der Mittelwert als μ angegeben, für Stichproben ist die Bezeichnung $\bar{x}$ üblich
- Sinnvoll vor allem für symmetrische, unimodale Verteilungen

Lagemaße

Arithmetischer Mittelwert

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

x_i: *Ausprägung eines Merkmals*
n: *Anzahl der Werte*

Lagemaße

Beispiel Arithmetischer Mittelwert

(Beispiel_Stichprobe.xlsx – Werte)

$$\overline{x} = \frac{1}{20} * (9,36 + \cdots + 9,85) = 9,6300$$

R-Berechnung

```
Werte
Min.    : 6.920
1st Qu.: 8.955
Median : 9.650
Mean   : 9.630
3rd Qu.:10.785
Max.    :13.120
```

Lagemaße

Getrimmtes arithmetisches Mittel

- Wirkung von Ausreißern auf das arithmetische Mittel kann durch Trimmung entschärft werden
- Kappung von sehr großen/kleinen Werten (Typischerweise auf 99% oder 95% der ursprünglichen Daten)
- Daten werden symmetrisch entfernt (oben / unten)
- Gefahr, dass auch Nicht-Ausreißer entfernt werden

Lagemaße

Geometrisches Mittel

- **Lagemaß für relative Änderung**
- **Nur für positive Zahlen definiert**

$$\overline{x}_g = \sqrt[n]{x_1 * x_2 * \cdots * x_n}$$

Beispiel: Jährliche Produktionssteigerung (2%; 2,5%; 1,7%; 2,3%)

$$\overline{x}_g = \sqrt[4]{1,02 * 1,025 * 1,017 * 1,023} = 1,021$$

Im Mittel beträgt die Steigerung 2,1%

Lagemaße

Median

- Mittlerer Wert einer geordneten Datenreihe
- Der Median teilt einen Datensatz nach Anzahl der Elemente
- Bei einem nach Größe geordneten Datensatz liegen unter bzw. über dem Median nicht mehr als 50% der Einzelwerte
- Robust gegen Ausreißer

Lagemaße

Median

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & , falls\ n\ ungerade \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right) & , falls\ n\ gerade \end{cases}$$

Lagemaße

Beispiel Median
(Beispiel_Stichprobe.xlsx – Werte)

- n ist gerade: Nach Ordnung der Werte liegt der Median zwischen den Werten $x_{10} = 9{,}50$ und $x_{11} = 9{,}80$

$$\tilde{x} = \frac{9,50 + 9,80}{2} = 9,65$$

Lagemaße

α-Quantil

- Definition eines bestimmten Anteils einer Datenmenge
- Wie viele Werte liegen unter oder über einer bestimmten Grenze
- Der Median und die beiden anderen Quartile (25%, 50%, 75%) sind spezielle Quantile

Lagemaße

α-Quantil

$$x_{\alpha} = \begin{cases} x_{(k)} & \text{falls } n*\alpha \text{ keine ganze Zahl ist (k ist dann die auf } n*\alpha \text{ folgende Ganzzahl)} \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{falls } n*\alpha \text{ eine ganze Zahl ist} \end{cases}$$

Beispiel: Gesucht ist das 25%-Quantil einer geordneten Datenreihe mit n = 30 Elementen

$n*\alpha = 7{,}5$ (keine ganze Zahl: aufrunden) => k = 8, $x_{25\%} = x_8$

Lagemaße

Beispiel Quantil

Werte
-0,48
1,23
1,49
-1,14
-0,71
-1,56
1,89
0,56
-1,1
-0,01
0,02
0,14
0,23
0,56
-0,01
-0,11
-0,03
-0,14
0,67
-0,51

Bestimmen Sie das 5%-Quantil aus Beispiel_Stichprobe.xlsx / Werte

Lagemaße

Beispiel Quantil 5% n =20; α = 5% = 0,05
(Beispiel_Stichprobe.xlsx – Werte)

lfd.Nr.	Werte
1	6,92
2	6,96
3	7,39
4	7,85
5	8,82

Auszug der geordneten Daten

1. n*α = 20* 0,05 = 1
2. n*α ist eine ganze Zahl
3. k = 1
4. Gesucht : $\frac{x_k+x_{k+1}}{2} = \frac{x_1+x_2}{2} = \frac{6{,}92+6{,}96}{2} = 6{,}94$

(R rechnet hier mit einer etwas anderen Formulierung, so dass Werte voneinander abweichen können)

Lagemaße

Modus (Modalwert)

- Der häufigste Wert in einer Datenreihe
- Vor allem für nominalskalierte bzw. diskrete Daten sinnvoll
- Interpretationsproblem bei mehreren gleichhäufigen Werten

Lagemaße

Beispiel Modus
(Beispiel_Stichprobe.xlsx – Modal)

Modal			
3	5	3	2
3	4	1	1
5	5	3	2
2	5	5	5
2	1	5	2

Bestimmen Sie den Modus der vorliegenden Datenreihe

Lagemaße

Beispiel Modus

$x_{mod} = 5$ (Häufigster Wert in der Tabelle)

Ergebnis aus R

```
counts:
Modal2
1 2 3 4 5
3 5 4 1 7

percentages:
Modal2
 1  2  3  4  5
15 25 20  5 35
```

Streuungsmaße

Spannweite

- Absoluter Abstand zwischen größtem (Maximum) und kleinstem (Minimum) Wert einer Datenreihe
- Extrem empfindlich hinsichtlich Ausreißern
- Auch hier ist eine Trimmung möglich (üblich sind 99% bzw. 95% der Werte)

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

Streuungsmaße

Beispiel Spannweite
(Beispiel_Stichprobe.xlsx – Werte)

```
Werte
 Min.    : 6.920
 1st Qu.: 8.955
 Median : 9.650
 Mean   : 9.630
 3rd Qu.:10.785
 Max.   :13.120
```

$$x_{min} = 6,920$$
$$x_{max} = 13,120$$
$$R = x_{max} - x_{min} = 6,200$$

Streuungsmaße

Interquartilsabstand (IQR; Inter Quartile Range)

- Abstand zwischen dem oberen (75%) und dem unteren (25%) Quartil
- Sehr robust
- Wird nicht durch Ausreißer beeinflusst
- Bestandteil des Box Plot

Streuungsmaße

Beispiel Interquartilsabstand
(Beispiel_Stichprobe.xlsx – Werte)

Gesucht: $x_{75\%}$ - $x_{25\%}$

```
Werte
 Min.    : 6.920
 1st Qu.: 8.955
 Median : 9.650
 Mean    : 9.630
 3rd Qu.:10.785
 Max.    :13.120
```

$$IQR = x_{75\%} - x_{25\%} = 10{,}785 - 8{,}955 = 1{,}830$$

(Eine Handrechnung würde hier zu etwas anderen Werten führen)

Streuungsmaße

Varianz / Standardabweichung

- Varianz: Mittlere quadratische Abweichung der gemessenen Werte vom arithmetischen Mittelwert

- Standardabweichung: Quadratwurzel der Varianz

Streuungsmaße

Varianz:

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

Standardabweichung:

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

Hier handelt es sich um die Berechnung für eine Stichproben, für die Grundgesamtheit wird der Divisor $\mathbf{n-1}$ durch **N** ersetzt

Streuungsmaße

Beispiel Varianz / Standardabweichung
(Beispiel_Stichprobe.xlsx – Werte)

$\bar{x} = 9{,}63$

Varianz

$$V = \frac{1}{20-1}\{(9{,}36-9{,}63)^2 + \cdots + (9{,}85-9{,}63)^2\} = 2{,}4903$$

Standardabweichung

$$s = \sqrt{2{,}4903} = 1{,}5781$$

Streuungsmaße

Beispiel Varianz / Standardabweichung

Aus R:

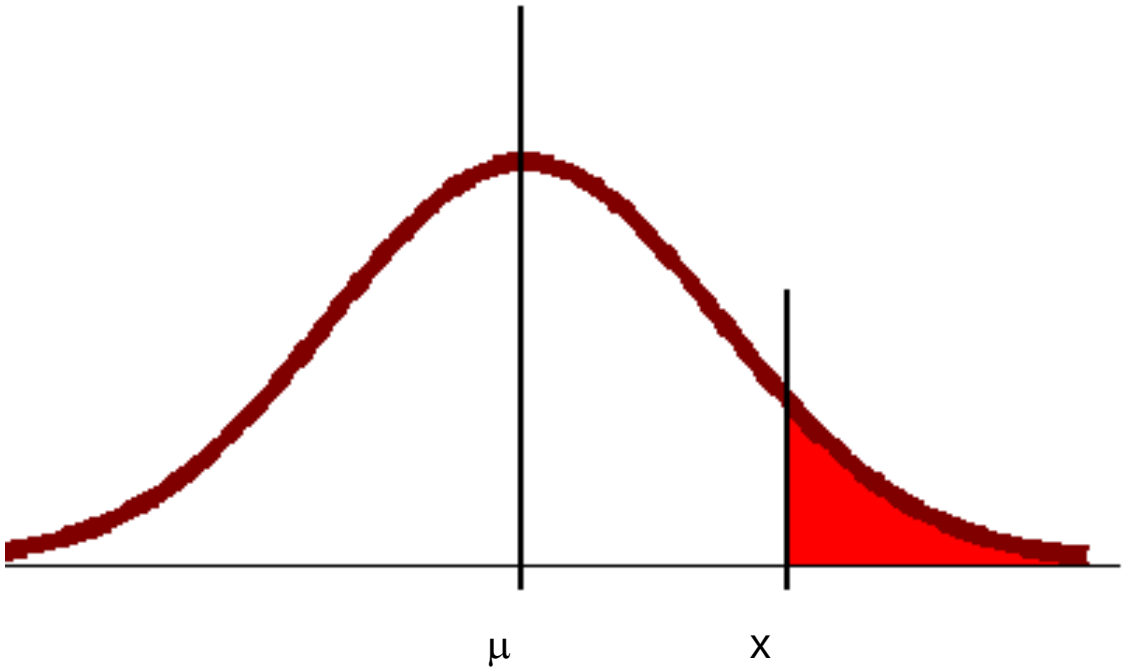
```
mean       sd  se(mean)  IQR        cv   skewness   kurtosis   0%   25%  50%
9.63 1.57808 0.3528694 1.83 0.1638712 0.01241656 0.08303501 6.92 8.955 9.65
   75%  100%  n
10.785 13.12 20
```

Standardwerte

- Auch bekannt als z-Wert
- Maß für den Abstand einer Beobachtung vom Mittelwert, ausgedrückt in Standardabweichungen für normalverteilte Daten
- Es werden Mittelwert und Standardabweichung benötigt
- Der z-Wert transformiert die aktuellen Daten in den Bereich einer standardisierten Normalverteilung
- Mit Hilfe von Tabellenwerken oder Software kann ein Bereich unter der Normalverteilung bestimmt werden

Standardwerte

$$z = \frac{(x - \mu)}{\sigma} bzw. \frac{(x - \bar{x})}{s}$$



Standardwerte

Zusammenhang von z-Werten und Anteil der Normalverteilung

z-Werte	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,49601	0,49202	0,48803	0,48405	0,48006	0,47608	0,47210	0,46812	0,46414
0,1	0,46017	0,45620	0,45224	0,44828	0,44433	0,44038	0,43644	0,43251	0,42858	0,42465
0,2	0,42074	0,41683	0,41294	0,40905	0,40517	0,40129	0,39743	0,39358	0,38974	0,38591
0,3	0,38209	0,37828	0,37448	0,37070	0,36693	0,36317	0,35942	0,35569	0,35197	0,34827
0,4	0,34458	0,34090	0,33724	0,33360	0,32997	0,32636	0,32276	0,31918	0,31561	0,31207
0,5	0,30854	0,30503	0,30153	0,29806	0,29460	0,29116	0,28774	0,28434	0,28096	0,27760
0,6	0,27425	0,27093	0,26763	0,26435	0,26109	0,25785	0,25463	0,25143	0,24825	0,24510
0,7	0,24196	0,23885	0,23576	0,23270	0,22965	0,22663	0,22363	0,22065	0,21770	0,21476
0,8	0,21186	0,20897	0,20611	0,20327	0,20045	0,19766	0,19489	0,19215	0,18943	0,18673
0,9	0,18406	0,18141	0,17879	0,17619	0,17361	0,17106	0,16853	0,16602	0,16354	0,16109
1,0	0,15866	0,15625	0,15386	0,15151	0,14917	0,14686	0,14457	0,14231	0,14007	0,13786
1,1	0,13567	0,13350	0,13136	0,12924	0,12714	0,12507	0,12302	0,12100	0,11900	0,11702
1,2	0,11507	0,11314	0,11123	0,10935	0,10749	0,10565	0,10383	0,10204	0,10027	0,09853
1,3	0,09680	0,09510	0,09342	0,09176	0,09012	0,08851	0,08692	0,08534	0,08379	0,08226
1,4	0,08076	0,07927	0,07780	0,07636	0,07493	0,07353	0,07215	0,07078	0,06944	0,06811
1,5	0,06681	0,06552	0,06426	0,06301	0,06178	0,06057	0,05938	0,05821	0,05705	0,05592
1,6	0,05480	0,05370	0,05262	0,05155	0,05050	0,04947	0,04846	0,04746	0,04648	0,04551
1,7	0,04457	0,04363	0,04272	0,04182	0,04093	0,04006	0,03920	0,03836	0,03754	0,03673
1,8	0,03593	0,03515	0,03438	0,03362	0,03288	0,03216	0,03144	0,03074	0,03005	0,02938
1,9	0,02872	0,02807	0,02743	0,02680	0,02619	0,02559	0,02500	0,02442	0,02385	0,02330
2,0	0,02275	0,02222	0,02169	0,02118	0,02068	0,02018	0,01970	0,01923	0,01876	0,01831

Standardwerte

Beispiel Standardwerte

Ihnen liegt Mittelwert (50) und Standardabweichung (3) einer Grundgesamtheit vor

Bestimmen Sie den Anteil, der über x = 54 liegt.

$$z = \frac{54 - 50}{3} = 1{,}33$$

Aus der Tabelle für z=1,33: p = 0,09176

In Ihrer Grundgesamtheit liegen 9,176% der Werte oberhalb von x = 54

Einführung Grafik

Graphische Werkzeuge helfen dabei:

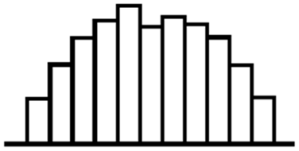
- Überblick über vorhandene Daten zu gewinnen
- Mögliche Beziehungen zwischen Variablen aufzuzeigen
- Risiken zu identifizieren, dass bestimmte Anforderungen nicht erfüllen werden
- Einblick zu bekommen, welcher Input (x) einen Einfluss auf das Ergebnis (y) hat und welcher nicht

Histogramme

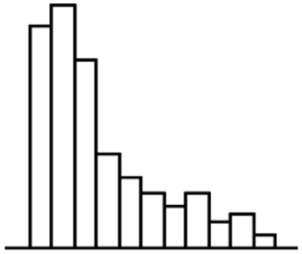
- Häufigkeitsverteilung nach Größe geordneter intervallskalierter Merkmale
- Klassen müssen nicht zwingend die gleiche Breite besitzen
- Überblick über die Streuung/Variation, die in einem Prozess auftritt
- Flächeninhalt der Klassen sind proportional zur Häufigkeit
- Visualisierung von Daten in erster Linie für stetige Merkmale mit einer großen Anzahl an Ausprägungen
- Oft wird eine Normalverteilung überlagert

Histogramme

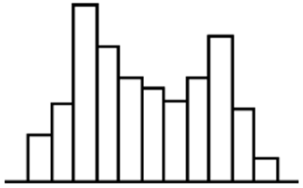
Typische Auffälligkeiten



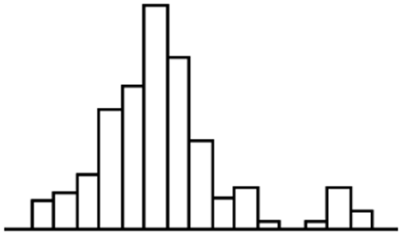
Alle Klassen haben ähnliche Datendichte – häufig der Fall bei zwei überlagerten Verteilungen



Deutlicher Shift – tritt häufig auf, wenn eine technische Grenze sehr nahe liegt oder es eine 100% Prüfung gibt



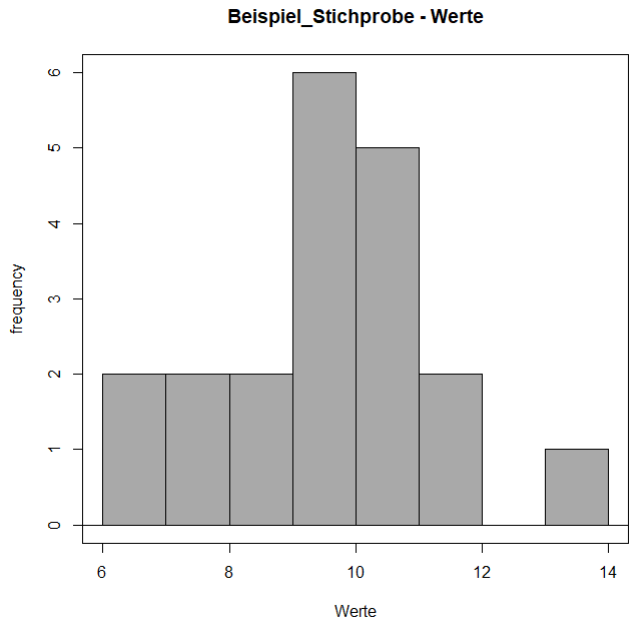
Zwei überlagerte Verteilungen mit unterschiedlichen Mittelwerten



Zwei Verteilungen, eine hat aber sehr wenig Daten – es könnte sich um ein kurzfristiges Prozessproblem handeln

Histogramme

Beispiel Histogramme
(Beispiel_Stichprobe.xlsx – Werte)



Augenmerk auf:

- Form (Welche Verteilung erkenne ich?)
- Schiefe (Verschiebung nach links oder rechts)
- Peaks (uni-modal, bi-/multi-modal)
- Long Tail (Daten-Cluster im Zentrum, ausgedehnte Seiten)

Balken- und Kreisdiagramme

- Besondere Eignung für die Darstellung diskreter Merkmale
- Stetige Merkmale sollten vorab in Klassen eingeteilt werden
- Darstellung von absoluten und relativen Häufigkeiten möglich
- Nur für überschaubare Datenmengen sinnvoll
- Bei relativen Häufigkeiten auf die Angabe der Basis achten (*2 von 3 Zahnärzten empfehlen…*)

Balken- und Kreisdiagramme

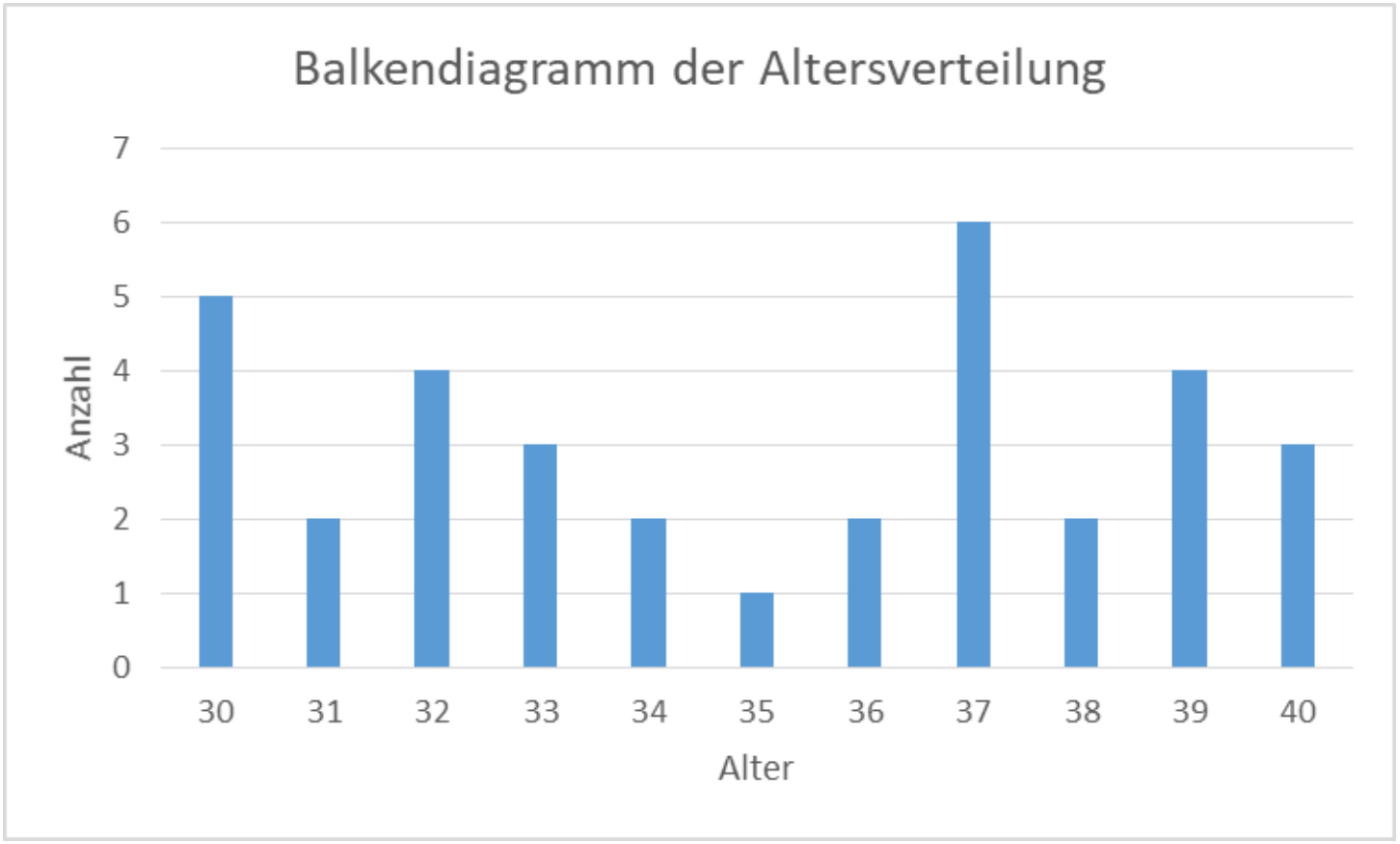
Beispiel Balken- und Kreisdiagramme

Alter	Anzahl	[%]
30	5	14,7
31	2	5,9
32	4	11,8
33	3	8,8
34	2	5,9
35	1	2,9
36	2	5,9
37	6	17,6
38	2	5,9
39	4	11,8
40	3	8,8

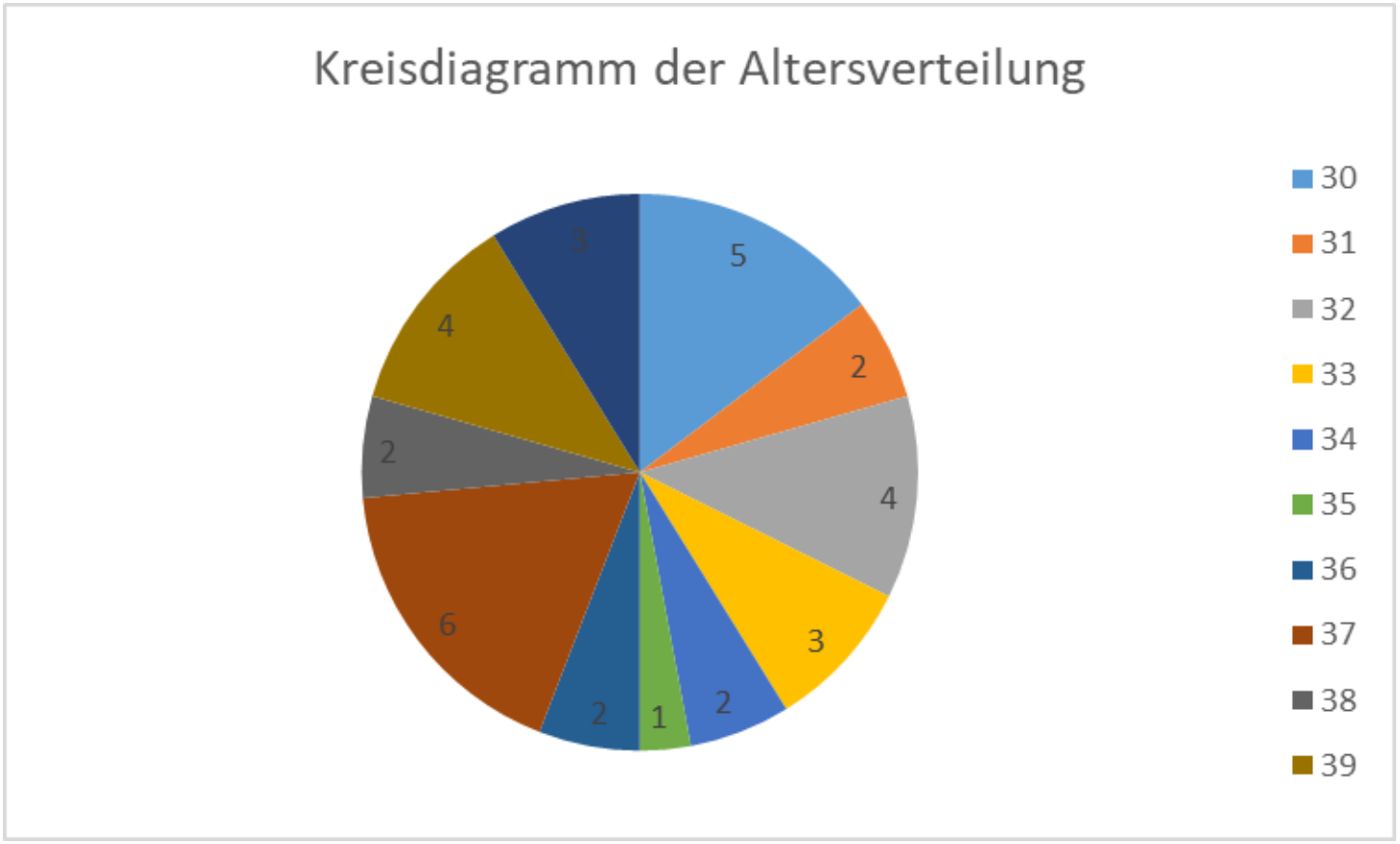
Stellen Sie folgende Altersverteilung als Balken- und Kreisdiagramm dar

Balken- und Kreisdiagramme

Beispiel Balken- und Kreisdiagramme



Balken- und Kreisdiagramme

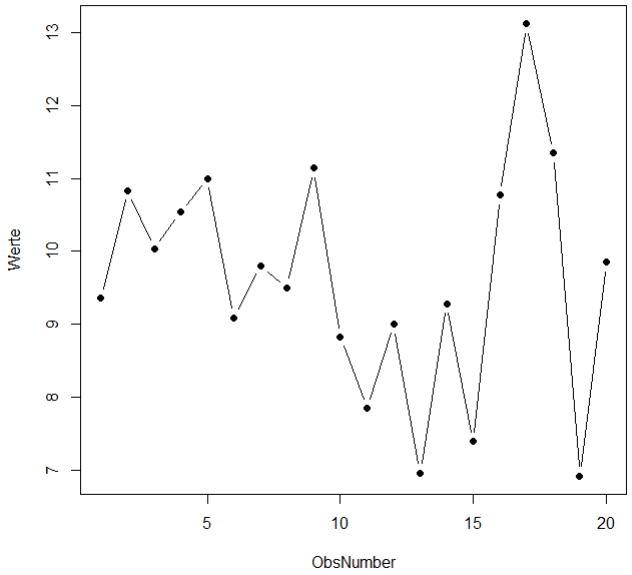


Liniendiagramme

- In der univariaten Darstellung wählt man die vorliegende Reihenfolge der Daten
- Vielfach fließt damit schon ein Zeitmuster ein (bivariat)
- Form und Auftreten von Auffälligkeiten können bei der Identifikation von Problemen der Datensammlung helfen

Liniendiagramm

Beispiel Liniendiagramm
(Beispiel_Stichprobe.xlsx – Werte)



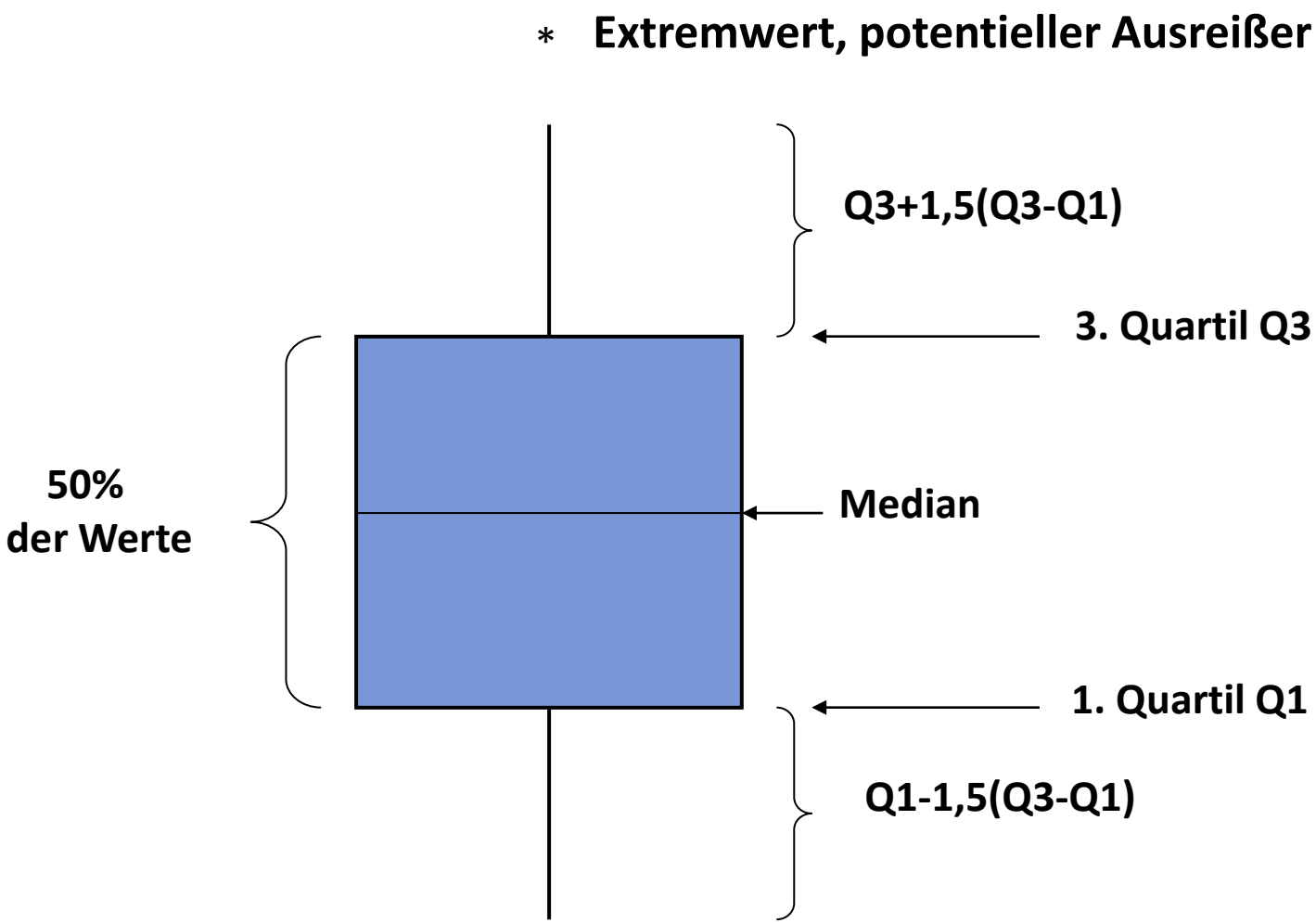
Augenmerk auf:

- Lage (Liegt das Zentrum der Daten wie erwartet?)
- Trends
- Mögliche Ausreißer
- Lücken
- Muster

Boxplots

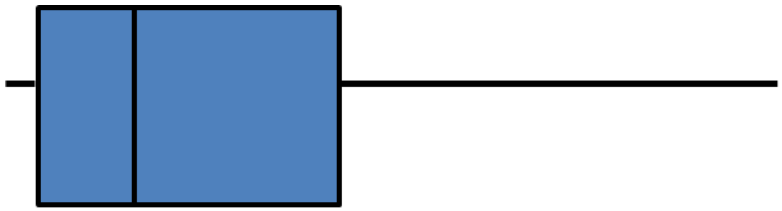
- Box-Whisker-Diagramm, Kastengrafik
- Vereinfachte graphische Darstellungen der Häufigkeitsverteilung von stetigen Daten
- Schneller Überblick über Lage und Verteilung
- Vergleich von Datensätzen
- Vorhandensein von möglichen Ausreißern
- Nicht für bi- oder multimodale Verteilungen geeignet

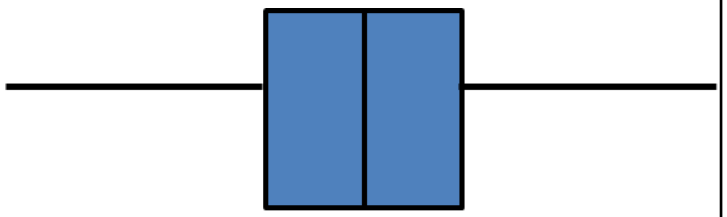
Boxplots

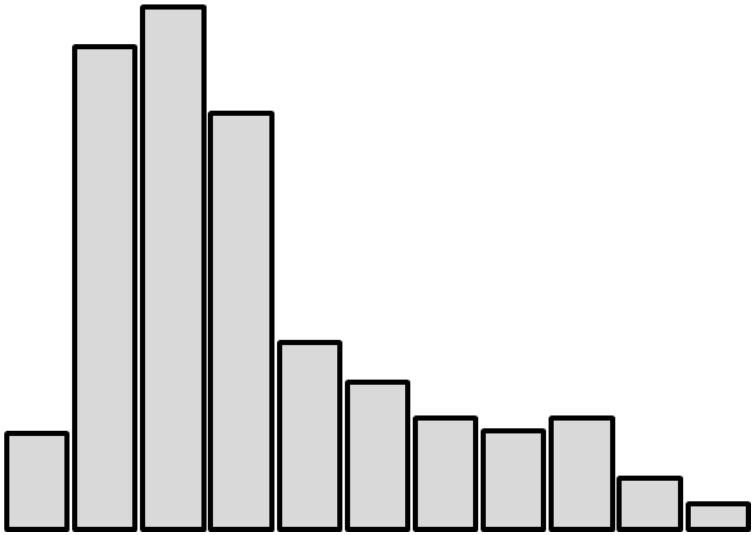

* Extremwert, potentieller Ausreißer
Q3+1,5(Q3-Q1)
3. Quartil Q3
50%
der Werte
Median
1. Quartil Q1
Q1-1,5(Q3-Q1)

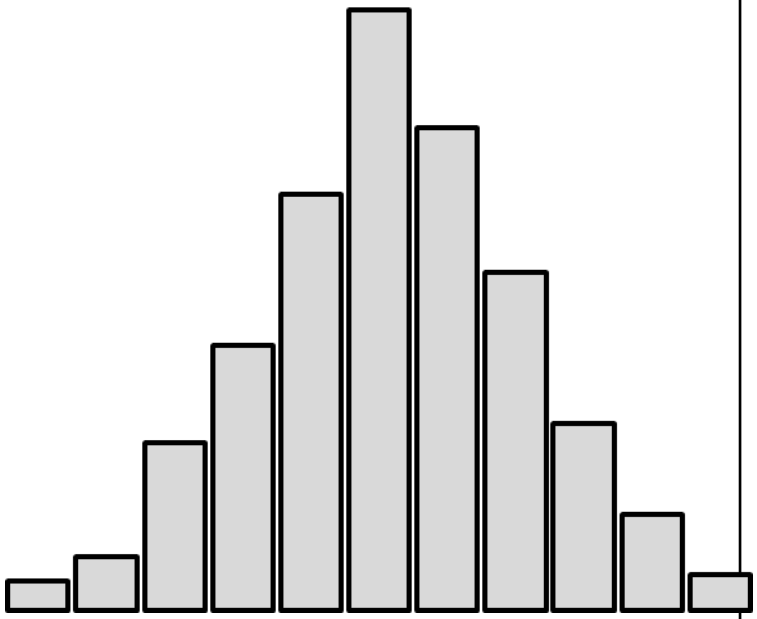
Boxplots

Ein Boxplot ist so etwas wie die Draufsicht auf unsere Daten



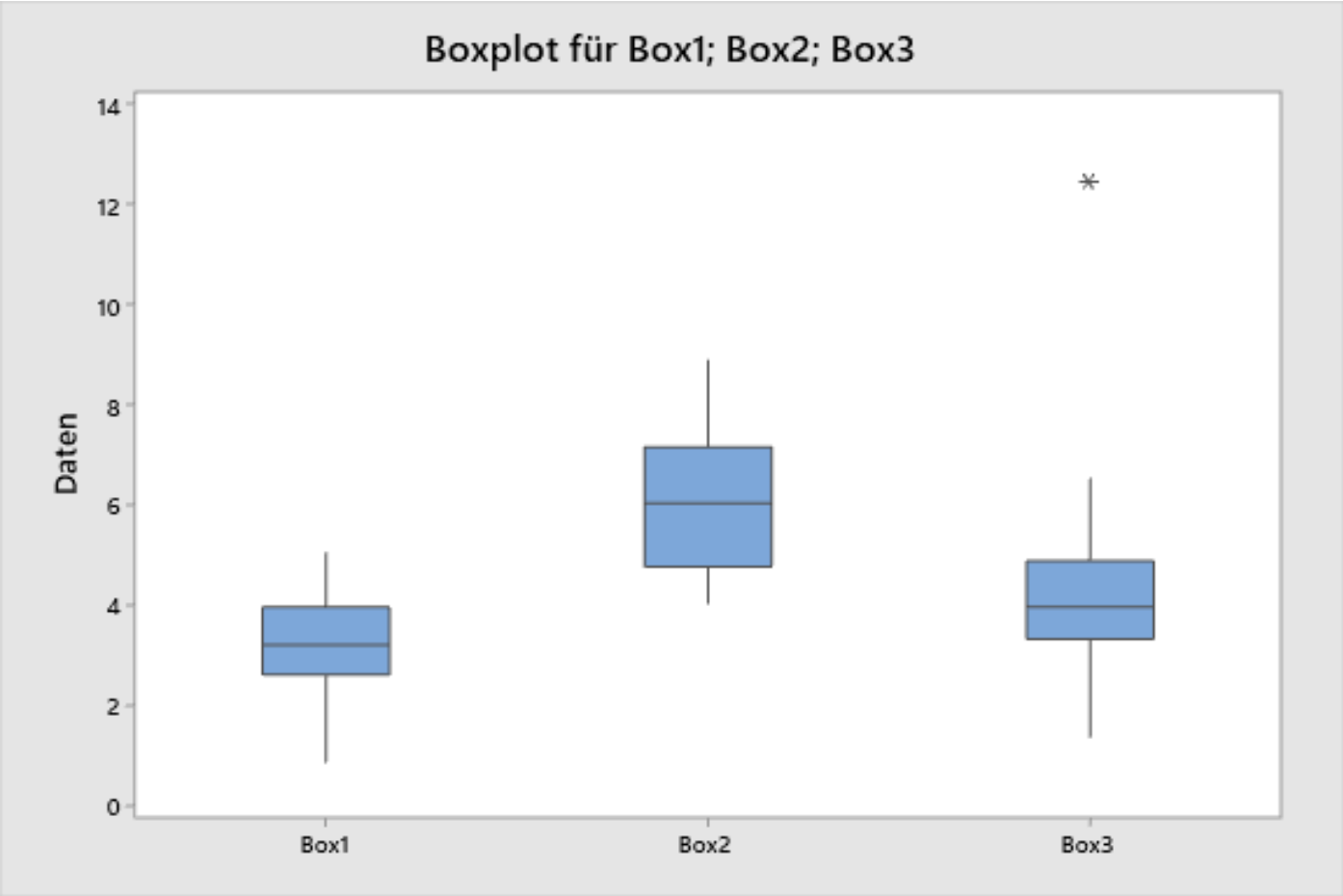






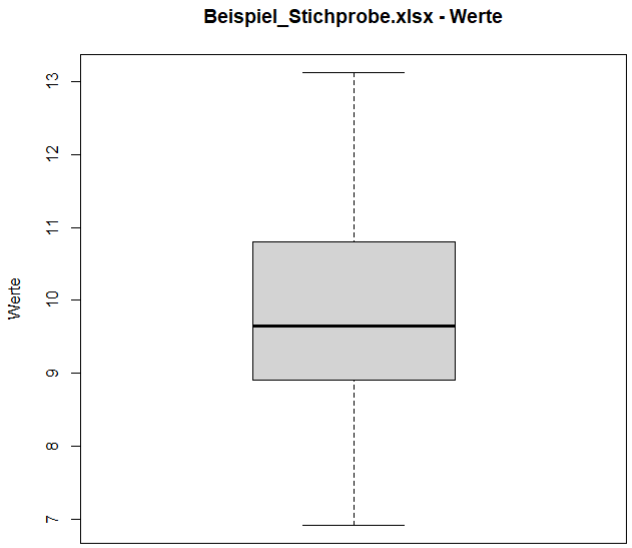
Boxplots

Seine wahre Stärke zeigt sich im Vergleich von Datensätzen



Boxplot

Beispiel Boxplot
(Beispiel_Stichprobe.xlsx – Werte)



Augenmerk auf:

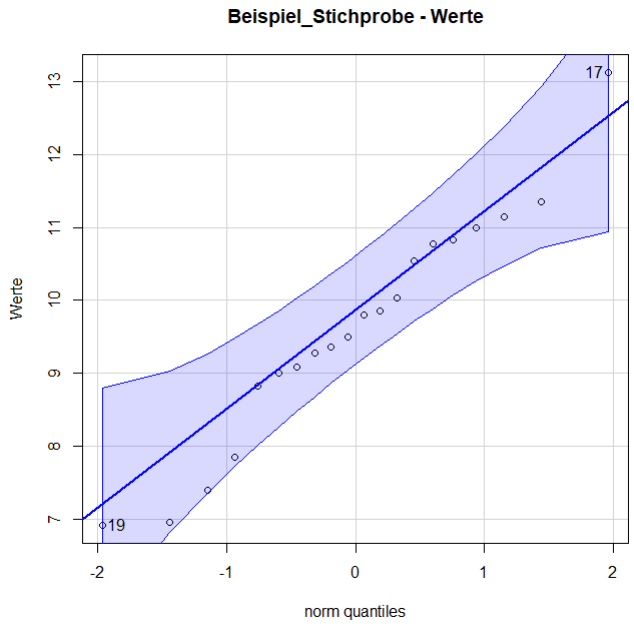
- Lage (Bei mehreren Boxplots)
- Symmetrie (Box bzw. Whisker)
- Mögliche Ausreißer

QQ – Diagramm

- Quantil-Quantil-Diagramm
- Darstellung der vorliegenden Daten im Vergleich zu einem theoretischen Verlauf
- Die Daten werden dazu in auf einer Quantil-Skala dargestellt, die für den theoretischen Verlauf eine Gerade darstellt
- Visuelle Prüfung der Normalverteilungsannahme (Vergleich mit anderen wählbaren Verteilungsformen)

QQ – Diagramm

Beispiel QQ-Diagramm
(Beispiel_Stichprobe.xlsx – Werte)



Augenmerk auf:

- Lage der Datenpunkte zur Geraden
- Datenpunkte innerhalb/außerhalb des Konfidenzintervalls
- Ungewöhnliche Formationen (S-Form)
- Brüche, Knicke usw.

Dot Plot

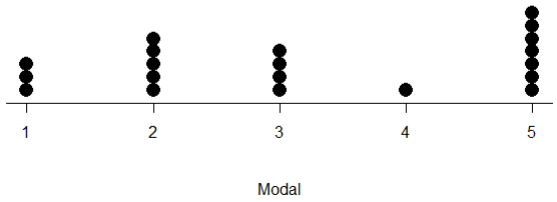
- Punktegrafik
- Darstellung der einzelnen Datenpunkte
- Geeignet zur Darstellung von ordinalen bzw. diskreten Verteilungen
- Hier: Das Gegenstück zum Histogramm
- Für kontinuierliche Daten weniger geeignet
- Nicht geeignet für sehr viele Daten

Dot Plot

Beispiel Dot Plot – ordinale/diskrete Daten
(Beispiel_Stichprobe.xlsx – Ordinal)

Augenmerk auf:

- Siehe Histogramm



Dot Plot

Beispiel Dot Plot – kontinuierliche Daten
(Beispiel_Stichprobe.xlsx – Werte)

Augenmerk auf:
- Verdichtung
- Lücken