Datenvisualisierung und Statistik

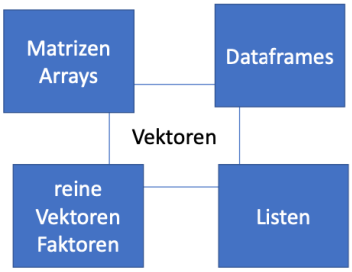
„Um Berechnungen in R zu verstehen sind zwei Sätze hilfreich: Alles, was existiert, ist ein Objekt. Alles, was passiert, ist ein Funktionsaufruf.“

Objekt = Datenstrukturen
Zwei Merkmale 1. Dimension: 1, 2 oder mehrere // 2. Homogenität: eine Art von Daten

Funktion = Befehl

Objekttypen

1. Vektoren
   - Geordnete Folgen von Werten
   - Eindimensional
   - 1 oder mehreren Elementen eines Datentyps
   - *"apple", "pear", "orange"*



Wichtigste Datentypen:
- Numeric: reelle Zahlen (3.14)
- Integer: ganze Zahlen (1, 2, 3)
- Character: Text (apples, pears)
- Logical: logische Ausdrücke (True, False)

2. Faktoren
   - Nützlich um Informationen von Gruppenzugehörigkeiten zu hinterlegen
   - Standardmässig alphabetisch geordnet
   - Typische Anwendung ist die Datenvisualisierung
   - Unterschiedliche Elemente werden als Stufen (levels) bezeichnet

```
days <- factor(c("Monday","Tuesday","Thursday","Monday"))
levels(days)
```

3. Listen
   - Nicht homogen!
   - Elemente mit unterschiedlicher Länge und Form

```
mylist <- list(1, c(1,2), c(TRUE, FALSE), c("Monday","Tuesday","Thursday"))
str(mylist)
List of 4
 $ : num 1
 $ : num [1:2] 1 2
 $ : logi [1:2] TRUE FALSE
 $ : chr [1:3] "Monday" "Tuesday" "Thursday"
```

4. Matrizen und Arrays
   - Matrizen: numerische Vekoren mit 2 Dimensionen
   - Arrays: n-Dimensionen

```
mymatrix <- matrix(c(1:6), nrow = 2, ncol = 3, byrow = TRUE)
mymatrix
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6

class(mymatrix)
 [1] "matrix" "array"
```

5. Dataframes und Tibbles

- Dataframes: Intuition wie bei "klassischer" Tabelle
- Spalten und Zeilen
- Homogen und Elemente mit gleicher Länge (nicht wie Listen!)
- Spalten haben einen Namen
- Elemente von Dataframes sind Vektoren

Dimension	Homogen	Heterogen
1	Reiner Vektor, Faktor	Liste
2	Matrix	Dataframe (Tibble)
Beliebig	Array	

Grundlagen Messen

Viele Bedeutungen von "Variable": Statistische Variable = Merkmal einer Erhebungseinheit mit verschiedenen Ausprägungen

Merkmal: Farbe *Mögliche Ausprägung: Schwarz, Grau, Weiss, ...*

1 Merkmal = 1 Spalte
1 Einheit = 1 Zeile

Einheit	Farbe	Muster	Geschlecht	Gewicht	...
Katze 1	Schwarz	Uni	Weiblich	3.2	
Katze 2	Weiss	Uni	Weiblich	2.5	
Katze 3	Weiss	gestreift	Weiblich	3.1	
Katze 4	Grau	gefleckt	Männlich	3.9	

Skalen

Nominalskala (kategoriale variable)

Merkmale mit "qualitativen" Ausprägungen, Einteilungen und Gruppenzugehörigkeiten
Beispiele: Augenfarbe, Geschlecht, Klasse, Kanton, Religion

Ordinalskala

Merkmale mit "qualitativen" Ausprägungen, die in einer "natürlichen" Rangordnung stehen. Abstände zwischen Ausprägungen sind nicht definiert!
Beispiele: Prüfungsnoten, Energielabel, Klassierung Turnier, Bildungsniveau

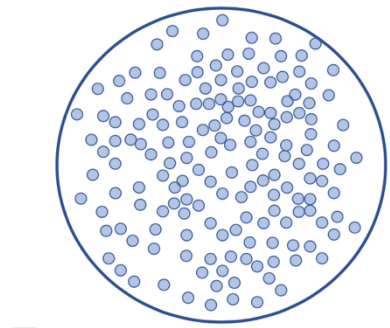
Metrische Skala (numerische Variable)

Merkmale mit "quantitativen" Ausprägungen wobei die Abstände zwischen zwei Werten mit einer Masseinheit definiert sind
Beispiele: Gewicht, Grösse, Alter, IQ, Distanz

Lagemasse

Mittelwert

- "average" oder "mean"
- Der Mittelwert wird als die Summe von der beobachteten Werte geteilt durch die Anzahl der Beobachtungen berechnet.



$\bar{\mu}$ Populationsmittelwert

$\bar{x}$ Stichprobenmittelwert

Median

- Median teilt die Daten in zwei Hälften
- Resistent gegenüber extremen Werten
- - nicht auf alle Beobachtungen basiert
- - beeinflusst durch Schwankungen bei der Probenahme
- Median und das zweite Quartil einer Verteilung sind identisch
- Perfekt symmetrischen Verteilung haben Median und Mittelwert gleichen Wert

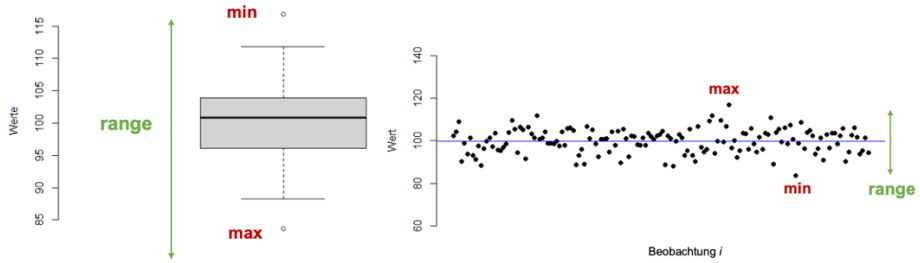
Streuungsparameter

= Dispersionsparameter

- Beschreiben die Variabilität der Ausprägung eines Merkmals in einem Datensatz

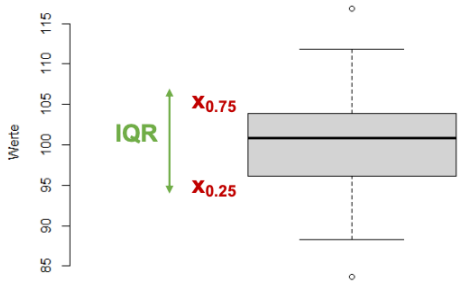
- Also wie dicht die Werte einer Häufigkeitsverteilung um den Mittelwert streuen

Minimum = kleinster Wert; maximum = grösster Wert; Range = Breite des beobachteten Wertebereiches



Interquartilsabstand = IQR

- Das Intervall, welches die mittleren 50% der gemessenen Werte enthält



Outlier/Ausreisser

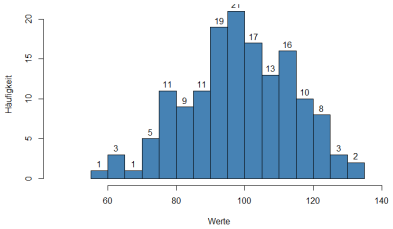
- "Extreme" Beobachtungen (weit entfernt vom grossen Rest der Beobachtung)
- Hinweise auf Messfehler
- "Problem" da Kennzahlen beeinflussen

Varianz
= durchschnittliche Abweichung aller Werte von ihrem Mittelwert im Quadrat.
Interpretation: immer positiv, grösserer Wert → mehr Varianz → mehr Streuung

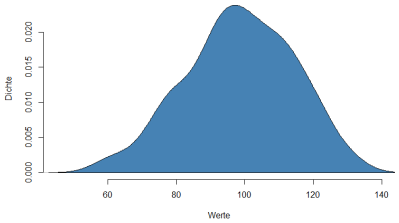
Standardabweichung
sd = Wurzel der Varianz
Interpretation: immer positiv; grösserer Wert → grössere sd → mehr Streuung

Verteilungen

Häufigkeitsverteilung
= eine Funktion, die beschreibt wie häufig welche Ausprägung eines Merkmals in der Stichprobe vorkommen (d.h. wie die Beobachtungen verteilt sind).
Wird durch die Kombination von Lage- und Streuungsmassen beschrieben

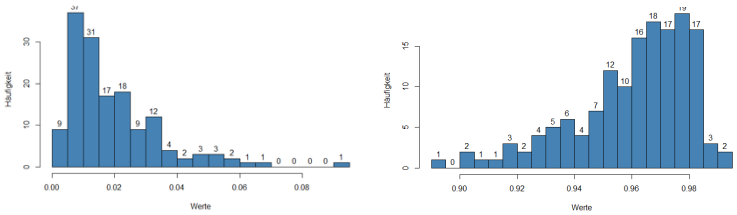


Wahrscheinlichkeitsverteilung
Abstrahierung der Häufigkeiten in Wahrscheinlichkeiten (Fläche)
Ziel: Rückschlüsse auf Population

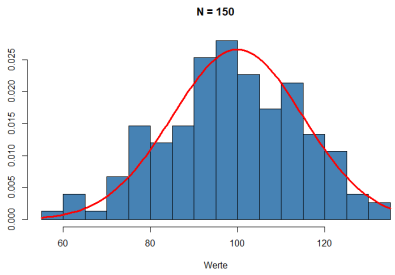


Symmetrie
Bei symmetrischer Verteilung sind Mittelwert, Modus und Median ungefähr identisch

Asymmetrie → Schief
Rechtsschief: Modus < Median < Mittelwert (mehr kleine Werte; weniger relativ grosse)
Linksschief: Modus > Median > Mittelwert (wenig relativ kleine Werte, mehr grosse Werte)

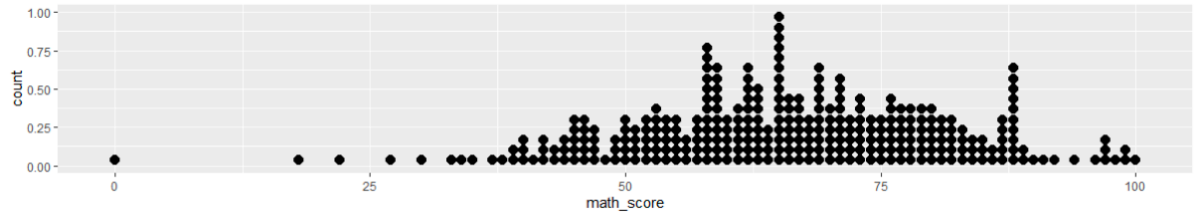


Normalverteilung
"Glockenkurve"
= Annäherung für viele Prozesse und Vorgänge in Natur, Gesellschaft, Wirtschaft, Technik
= immer symmetrisch

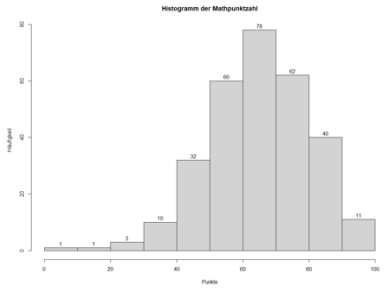


Visualisierungsformen

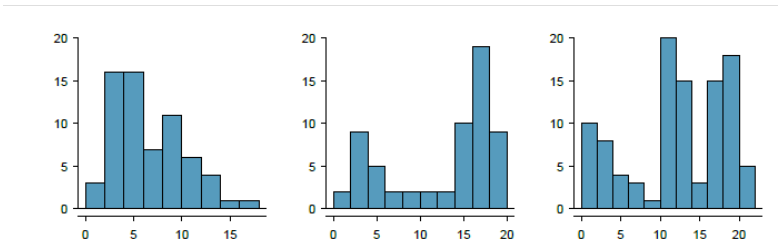
	Vorteil	**Nachteil**
Dotplot	1) sehr viele Daten sichtbar aber wenig Zusammenfassung	1) Mittelwert und Median nicht direkt ersichtlich 2) Y-Achse eher schwierig zu interpretieren.
Histogram	1) Leichte Orientierung (Achsen selbsterklärend; gute Übersicht der Häufigkeitsverteilung) 2) Beobachtungen gruppiert in Intervalle ("bins")	1) Lagemasse nicht direkt ersichtlich 2) Einzelne Beobachtung jetzt aggregiert
Boxplot	1) Viel Zusammenfassung 2) Viele Lage- und Streuungsparameter ersichtlich 3) Sehr geeignet zum Vergleich von Kategorien	1) Bedingt statistisches Wissen und Interpretation
Scatterplot	1) Alle Datenpunkte sichtbar (Ausreisser und extreme Werte gut erkennbar) 2) sehr intuitiv 3) Gängigste Visualisierung für bivariate Darstellung (Zusammenhänge lassen sich gut erkennen)	1) wenig Zusammenfassung (wird meistens zusätzlich gemacht)



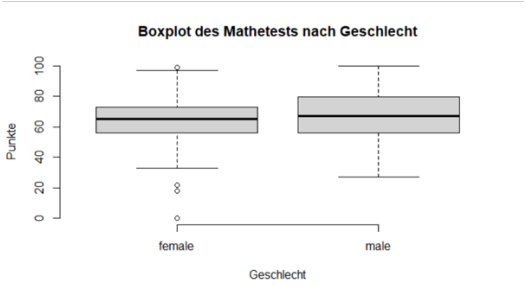
Dotplot



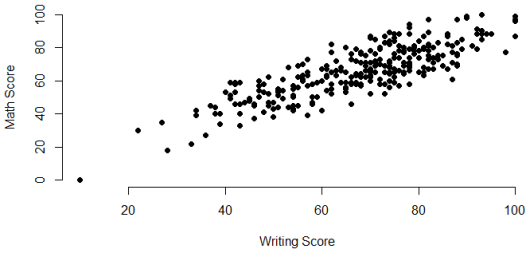
Histogram



unimodal bimodal multimodal



Boxplot



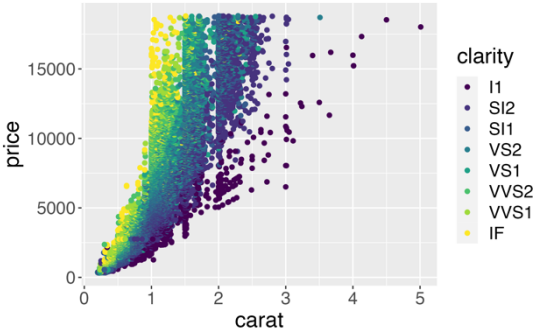
Scatterplot

7 verschiedene Grammatikalische Ebenen von ggplot2

1) die Daten die wir darstellen wollen
2) abbildende Aspekte ("aesthetics"): Achsen, Farben, usw.
3) Geometrie ("Art von Bild") (Histogram, Boxplot, point, text, bar, line)
4) Thema
5) Statistics
6) Coordinates
7) facete Layers

Overplotting = Überplottung

- Grosser Datensatz
- Ausgerichtete Wrte auf einer einzigen Achse
- Daten mit geringer Genauigkeit
- Ganzzahlige Daten
-



1. Lösung = Transparenz der Abbildenden Form (Shape) anpassen
2. Lösung = leere Kreise
3. Lösung = Jitter: Ausgerichtete Werte auf einer einzigen Achse

:: & %>%

:: → doppelter Doppelpunkt (Funktion count vom Packet dplyr)
>%> → pipe. Leitet den Wert oder das Ergenis eines Ausdrucks auf den nächsten Funktionsaufruf / Ausdruck weiter

geom_col & geom_bar

geom_col: zeichnet den tatsächlichen Wert auf, den es im Datensatz findet

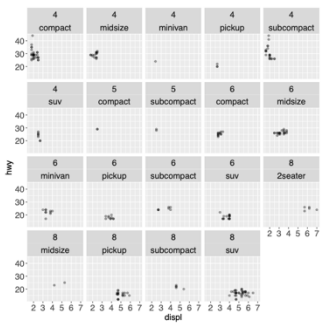
geom_bar: zählt die Anzahl der Fälle in jeder Kategorie der auf der x-Achse abgebildeten Variable.

Facettierung eines Diagrams

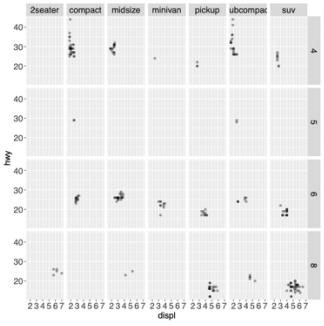
Ein Diagramm in Subgruppen aufteilen

Facet_wrap() erzeugt ein langes Band von Panels (erzeugt durch eine beliebe Anzahl von Variablen) und entiwckelt es in 2d

Facet_grid() legt die Diagramme in einem 2D-Gitter an, wie es durch die Formel definiert ist



facet_wrap()

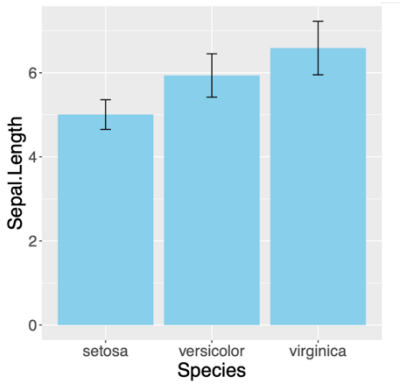


facet_grid()

facet_grid()
verteilt die Werte von a auf die Spalten . ~ a
verteilt die Werte von b auf die Zeilen b ~ .
verteilt a über die Spalten und b über die Zeilen b ~ a

X-Achse	Y-Achse	Diagrammtype
Kontinuierliche Variable		Histogramm, Dichtediagramm,
Kontinuierliche Variable	Kontinuierliche Variable	Scatterplot, Schachbrett-Diagramm, Liniendiagramm
Nominale Variable	Kontinuierliche Variable	Balkendiagramm
Nominale Variable	Nominale Variable	Mosaicplot, Fliesendiagramm
Nominale Variable	Kontinuierliche Variable	Punktdiagramm für Zusammenfassungen
Nominale Variable	Kontinuierliche Variable	Boxplot

Dynamite Blot



Vorteile
- sie sind den meisten Menschen vertraut
- sie verankern die Daten bei Null
- sie bieten mehr Fläche für die Darstellung von Farben

Nachteile
- niedrige "Data-to-Ink"-Werte
- roh Daten sind versteckt
- sie gehen von symmetrischen Konfidenzintervallen aus

Langes und breites Datenformat

Breit:

- Jede Untersuchungseinheit = 1 Zeile (d.h. keine Wiederholungen der Einheit)
- "viele" Spalten, d.h. verschiedene Messungen nebeinander verteilt
- Praktisch zum Rechnen
- Unpraktisch für Visualisierung

Lang:

- Jede Messung = 1 Zeile (d.h. keine widerholte Messungen nebeneinander aber wiederholte Zeilen für Einheiten)
- "Wenige" Spalten, d.h. verschiedenen Messungen gruppiert verteilt
- Praktisch für Visualisierung aber unpraktisch für direkte Berechnung