Data Science Einstieg

Praxisnahe Business Analyse mit Excel

Workshop im Rahmen des DIH SÜD

Data Science and Artificial Intelligence
Institute of Business Informatics and Data Science
FH Joanneum – University of Applied Sciences

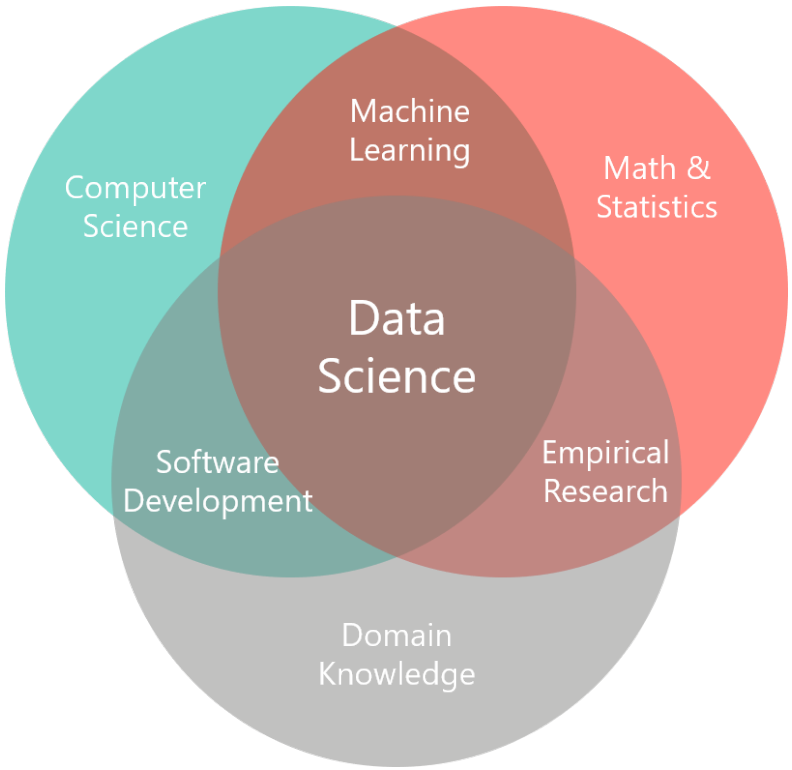
9. November 2022

DIHSÜD
FFG Forschung wirkt.
Das Land Steiermark
LAND KÄRNTEN
FH JOANNEUM University of Applied Sciences

Inhalt

Inhalt

Einleitung – Data Science



Quelle: www.finbridge.de

Bereiche/Grundaufgaben der Statistik

- Deskription (Beschreibung) – Deskriptive Statistik
  - Beschreibung und graphische Aufbereitung/Darstellung der Daten
  - Komprimierung der Daten z.B. in Tabellen (Extraktion der relevanten Information)
  - Verwendung von Lage- (arithmetisches Mittel, Quantile, Modalwert, ...) und Streuungs- (Standardabweichung, Varianz, Interquartilsabstand, ...) und Assoziationsmaßen (z.B. Korrelationskoeffizient)

- Exploration (Suchen nach Mustern und Strukturen) – Explorative Datenanalyse (EDA)
  - Auffinden von Objekten bzw. Objektgruppen, die bezüglich gewisser Merkmale eine mehr oder weniger homogene Einheit bilden, Identifikation besonderer Beobachtungen (Ausreißer)
  - Identifikation von Zusammenhängen in den Daten, Formulierung von Hypothesen
  - ... man geht auf Entdeckungsreise in der Welt der zuvor gemessenen oder erhobenen Daten ...

- Inferenz (schließende Statistik)
  - Schluß von der Stichprobe auf die Grundgesamtheit mit Mitteln der Wahrscheinlichkeitsrechnung/Wahrscheinlichkeitstheorie
  - Punktschätzung, Intervallschätzung und Hypothesentests
  - Wie muß eine Stichprobe gezogen werden, damit diese Rückschlüsse machbar sind?

Objekte und Merkmale

Im Rahmen jeder statistischen Untersuchung werden interessierende Größen, die sogenannten Merkmale an gewissen Objekten ermittelt.

Beispiel

- Ist man im Rahmen einer Absolventenstudie daran interessiert, mit welchem Anfangsgehalt Absolventen der Studienrichtungen *Informationsmanagement* und *Data Science and Artificial Intelligence* zu rechnen haben, so ist das erhobene Merkmal das (Monats-)bruttoeinkommen. Die Objekte sind in diesem Fall die Absolventen der beiden Studienrichtungen.
- Will man im Rahmen einer Wahlumfrage ermitteln, welche mit welchen Prozentsätzen die politischen Parteien bei einer (hypothetischen) Wahl am folgenden Sonntag in Österreich zu rechnen haben (*Sonntagsfrage*), befragt man die wahlberechtigte Bevölkerung (Objekte), das Merkmal ist die Parteienpräferenz.

Objekte und Merkmale

Den konkreten Wert eines Merkmals an einem Objekt nennt man Merkmalswert. Alle möglichen (nicht notwendigerweise numerischen) Werte, die ein Merkmal annehmen kann, nennt man den Wertebereich eines Merkmals oder die Merkmalsausprägungen, die zumeist in einer Menge M zusammengefasst werden.

Beispiel

- Will man die Geschlechterzusammensetzung im Vergleich zwischen technischen und sozialwissenschaftlichen Studienrichtungen untersuchen, so ist das erhobene Merkmal das Geschlecht der Studenten mit den möglichen Ausprägungen w(eiblich), m(ännlich) oder d(ivers). Somit ist die Menge $M = \{\text{w, m, d}\}$ der Wertebereich des Merkmals *Geschlecht*.
- Ist man an den Verkaufszahlen eines bestimmtem Produkts in den Filialen einer Kette interessiert, so sind die möglichen Ausprägungen $0, 1, 2, \ldots$. Bei geeignetem Design kann z.B. festgestellt werden, ob es (statistische signifikante) Unterschiede in den Verkaufszahlen zwischen den einzelnen Filialen gibt.
- Bestimmt man die Lebensdauer (Merkmal) eines technischen Produkts, so sind die möglichen Ausprägungen alle numerische Werte im Intervall $[0, \infty)$. Das Ergebnis einer solchen Untersuchung könnten die 5 Werte (in h)

$$x_1 = 125 \qquad x_2 = 67 \qquad x_3 = 423 \qquad x_4 = 40 \qquad x_5 = 212$$

sein.

Grundgesamtheit

Definition (Grundgesamtheit/Population)

Die Menge **aller** in Frage kommender Objekte, über die im Rahmen einer statistischen Untersuchung eine Aussage getroffen werden soll, nennt man Population oder Grundgesamtheit. Sie ist räumlich, zeitlich und sachlich genau abzugrenzen. Die Objekte der Grundgesamtheit werden in diesem Zusammenhang oft auch als statistische Einheiten oder Merkmalsträger bezeichnet.

Das Ziel einer statistischen Untersuchung ist es, Aussagen über die zugrundeliegende Grundgesamtheit zu treffen. Zumeist ist aber aus zeitlichen, finanziellen und/oder technischen Gründen eine Untersuchung **jeder** statistischen Einheit der Grundgesamtheit (dies wäre eine sogenannte Vollerhebung oder Totalerhebung bzw. Zensus) **nicht möglich** oder **sinnvoll**.

Beispiel

- Nicht sinnvoll ist eine Vollerhebung insbesondere dann, wenn das untersuchte Objekt bei der Untersuchung zerstört wird (Materialprüfung, technische Statistik).
- Beispiel für Vollerhebung: Volkszählung in Österreich (seit 1754); ab 1951 alle 10 Jahre bis 2001, dann durch eine Registerzählung ersetzt.
- Vollerhebung: mathematischer Eingangstest bei verschiedenen (technischen/naturwissenschaftlichen) Studienrichtungen.

Stichproben

Wird keine Vollerhebung angestrebt, so sind Stichprobenverfahren anzuwenden. Dabei wird nur ein (oft sehr kleiner) Teil der Grundgesamtheit untersucht. Stichproben sollten repräsentativ für die zugrundeliegende Population sein.

Definition (Einfache Zufallsstichprobe)

Unter einer einfachen Zufallsstichprobe versteht man eine Teilmenge der Grundgesamtheit (Stichprobe), wobei jede statistische Einheit die gleiche Chance hat, Teil der Stichprobe zu werden und die Ziehungen der Einheiten voneinander unabhängig zu erfolgen hat.

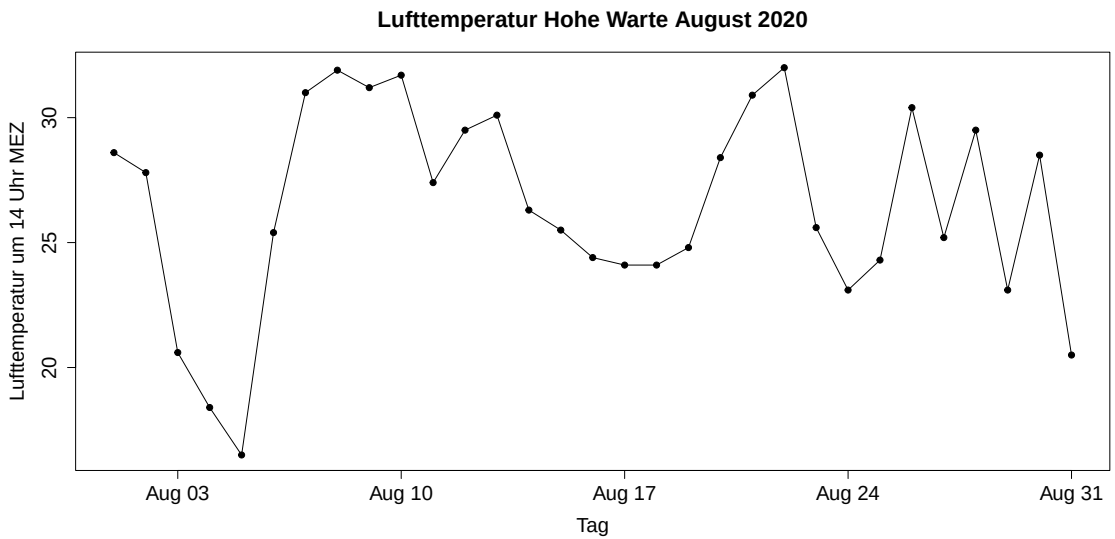
Beispiel

- Ist man am Ausschußanteil eines Produkts einer bestimmten Charge interessiert, werden alle statistischen Einheiten von 1 bis N durchnummeriert und eine Menge von n dieser Einheiten zufällig aus der Charge entnommen und geprüft (Zufallszahlengenerator, früher Tafeln von Zufallszahlen).
- **Negativbeispiel**: Amerikanische Präsidentschaftswahl 1936 Landon vs. Roosevelt; Magazin *Literary Digest* Umfrage prophezeit klaren Sieg von Landon ($57\,\%$) gegen Roosevelt ($43\,\%$) basierend auf einer Umfrage unter $n = 2.400.000$ (!) Personen (damalige Bevölkerung der USA ca. 125 Millionen). Das wahre Ergebnis der Wahl: $62\,\%$ Roosevelt, $38\,\%$ Landon – warum? Ein anderes Institut (George Gallup) konnte mit einer Stichprobengröße von $n = 50.000$ den Gewinner korrekt vorhersagen.

Studiendesigns

Je nachdem, wie die Daten gewonnen werden, unterscheidet man

- Querschnittsstudie: zu einem fixen Zeitpunkt werden die Merkmale an den statistischen Einheiten (Vollerhebung oder – zumeist – Stichprobe) erhoben. Als Beispiel möge die PISA (*Programme for international student assessment*) Studie bzw. jede herkömmliche Wahlumfrage dienen.
- Longitudinalstudie: an einer fixen Stichprobe (ein sogenanntes Panel) werden die Merkmale zu unterschiedlichen Zeitpunkten erhoben. Beispiel: langfristige Wirkung eines Medikaments.
- Zeitreihe: es werden an einem Objekt in gewissen (nicht notwendigerweise gleichen) zeitlichen Abständen Messungen/Erhebungen durchgeführt. Beispiel: Messung der Lufttemperatur am Standort Hohe Warte durch die Zentralanstalt für Meteorologie und Geodynamik.



Stetige und Diskrete Merkmale

Man kann Merkmale hinsichtlich der Anzahl der Merksmalsausprägungen einteilen:

- Diskretes Merkmal: die Anzahl an möglichen Merkmalsausprägungen ist endlich oder abzählbar unendlich.
- Stetiges Merkmal: das Merkmal kann alle Werte in einem (reellen) Intervall annehmen.

Beispiel

- Das Geschlecht, die Anzahl an Autos (oder Kinder) im Haushalt, der Familienstand, die Anzahl an Spielrunden im Roulette, bis der Spieler bankrott ist, die Güteklasse einer bestimmten Obstsorte, ... sind diskrete Merkmale
- Die Körpergröße, die Zeitdauer bis zum Entritt eines Ereignisses (z.B. Ausfall eines elektronischen Bauteils), die innerhalb eines Jahres zurückgelegte Strecke mit dem eigenen KfZ, ... sind stetige Merkmale

Oft werden Merkmale, die nur diskret gemessen werden (können), aber in einer sehr feinen Abstufung vorliegen, als quasi-stetig bezeichnet, z.B. Einkommen einer Person, Mietpreis einer Wohnung, gemessene Zeiten bei einer Leichtathletik-Veranstaltung, ...

Gruppierte Merkmale

Eine weitere Zwischenform stellen gruppierte Daten dar. Dabei wird der Wertebereich eines stetigen Merkmals X in nicht-überlappende Gruppen oder Klassen eingeteilt, wie beispielsweise das Alter in Altersklassen

$$[0, 10) \quad [10, 20) \quad [20, 30) \quad \dots \quad [90, +\infty)$$

oder das Gehalt einer Person in Gehaltsklassen

$$[0, 500) \quad [500, 1000) \quad [1000, 2000) \quad \dots \quad [10000, +\infty)$$

- Man beachte die 1-seitig offenen und 1-seitig geschlossenen Intervalle.
- Werden die Daten in stetiger Form erhoben/gemessen, bedeutet eine Gruppierung einen Verlust an Information.
- Werden die Rohdaten bereits in gruppierter/klassierter/kategorisierter Form erhoben, erhöht dies in vielen Fällen die Antwortbereitschaft der befragten Person bzw. dient es dem Datenschutz.

Inhalt

Skalenniveaus von Daten

Das sogenannte Skalenniveau eines Merkmals charakterisiert den Informationsgehalt der gemessenen/beobachteten Werte bzw. der Ausprägungen eines Merkmals.

1. Nominalskala:
   - Die Ausprägungen eines nominalskalierten Merkmals weisen keine innere Ordnung der Ausprägungen auf.
   - **Beispiele**: Geschlecht (mit den Ausprägungen w, m und d), Nationalität (Ausprägungen z.B. *österreichisch*, *deutsch*, ...), Verwendungszweck einer Überweisung (Ausprägungen *beruflich* oder *privat*), gelesene Tageszeitungen (*Krone*, *Kurier*, *Presse*, *Standard*, ...)
   - Die Kodierung der Kategorien in Zahlen, wie beispielsweise m, w und d als $1, 2$ und 3 (oder auch 0, 1 und 2) dient **nur** zur Unterscheidung und effizienten Speicherung. Dementsprechend haben Ausdrücke wie $1 < 2$ oder $1 + 2$, $2 - 1$ oder $\frac{1}{3}$ keinen Sinn.
2. Ordinalskala (oder Rangskala):
   - Zwischen den Ausprägungen eines ordinalskalierten Merkmals herrscht eine innere/natürliche Ordnung, d.h. die Werte bzw. Kategorien lassen sich ordnen.
   - Differenzen (Abstände) der Kategorien können nicht interpretiert werden, arithmetische Operationen nicht ebenso nicht sinnvoll
   - **Beispiele**: Schulnoten (mit den Ausprägungen 1, 2, 3, 4, 5) mit der Ordnung $1 > 2 > 3 > 4 > 5$ (*größer* hier im Sinne von *besser*), Temperatur mit den Ausprägungen *sehr kalt*, *kalt*, *warm* und *sehr warm*

Skalenniveaus von Daten

3. Intervallskala:
   - Intervallskalierte Merkmale verfügen über keinen natürlichen Nullpunkt.
   - Differenzen zwischen Ausprägungen können interpretiert werden, Quotienten allerdings nicht.
   - **Beispiele**: Temperatur in Grad Fahrenheit oder Grad Celsius (°C): zwischen $T = 10\,°\text{C}$ und $T = 30\,°\text{C}$ besteht der gleiche Abstand/Unterschied, wie zwischen $T = 70\,°\text{C}$ und $T = 90\,°\text{C}$ (nämlich $\Delta T = 20\,°\text{C}$); der Temperaturwert $T = 30\,°\text{C}$ ist allerdings **nicht** dreimal so warm wie $T = 10\,°\text{C}$.
4. Verhältnisskala:
   - Der Nullpunkt verhältnisskalierter Merkmale ist inhaltlich interpretierbar und *natürlich*, alle arithmetischen Operationen sind sinnvoll
   - **Beispiele**: Geldbetrag, Wohnfläche, Temperatur in Kelvin (K), Zeitdauer in Sekunden s

Merkmale mit nominalem und ordinalem Skalenniveau werden auch als kategorische oder qualitative Merkmale oder Variablen bezeichnet. Intervall- oder verhältnisskalierte Merkmale/Variablen werden zusammenfassend auch als metrische Merkmale/Variablen bezeichnet.

Skalenniveaus von Daten

	sinnvoll interpretierbare Berechnungen			
Skalenart	auszählen	ordnen	Differenzen bilden	Quotienten bilden
Nominalskala	ja	nein	nein	nein
Ordinalskala	ja	ja	nein	nein
Intervallskala	ja	ja	ja	nein
Verhältnisskala	ja	ja	ja	ja

Der zuvor erwähnte Informationsverlust beim Übergang von der metrischen (verhältnisskalierten) Variable *Alter* zu Altersklassen bedeutet einen Rückschritt zu einer Ordinalskala.

Inhalt

Beispieldatensatz Mietspiegel

Daten Mietspiegel 2015 (nach Fahrmeier et. al); Variablenbedeutung: Nettomiete `nm`, Nettomiete pro m^2 `nmqm`, Wohnfläche `wfl`, Anzahl an Zimmer `rooms`, Baujahr `bj`, Bezirk (`bez`), gute und Bestlage `wohngut` und `wohnbest`, Warmwasserversorgung ja/nein in `ww0`, Zentralheizung ja/nein in `zh0`, gehobene Ausstattung des Bades und Küche in `badkach0`, `badextra` und `kueche`

Excel – `MietenA.csv`

Warum ist diese Darstellung gefährlich? Was wäre eine Alternative? Welche Variablentypen kommen vor?

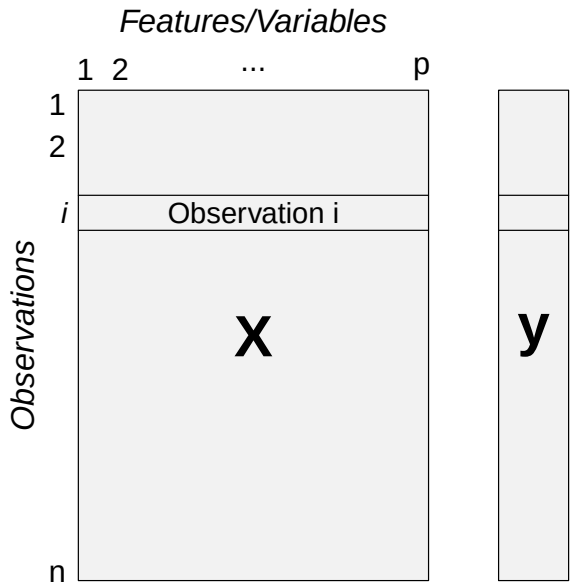
Excel – `MietenB.csv`

Datenstruktur in Beobachtungen (in Zeilen) und Variablen (in Spalten); Typischerweise n Anzahl an Beobachtungen, p Anzahl an Variablen, daher Format einer $n \times p$ Matrix; Erstellung von Pivottables.

Inhalt

Supervised Setting – Regression/Classification

Im sogenannten supervised learning haben wir eine uns interessierende Variable (oft genannt response y oder abhängige Variable), die aber schwer direkt zugänglich ist. Daher wollen wir sie anhand leicht und/oder billig meßbarer Größen $x_1, x_2, \ldots, x_p$ (sogenannte Prädiktoren) modellieren, d.h. vorhersagen. Dafür benötigen wir Daten – ein sogenanntes Trainingsset mit n Beobachtungen bekannter x- und y-Werte.



Ist y eine kategorielle Größe, so spricht man von Klassifikation, ist y eine numerische Größe, so handelt es sich um ein Regressionsproblem. Beispiel Mietdatensatz.

Unsupervised Learning

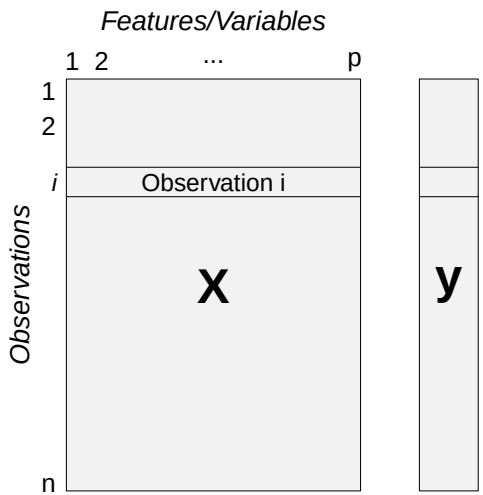
Oft hat man keine Größe y, die von besonderem Interesse und vorherzusagen ist, sondern lediglich eine Datenmatrix $\mathbf{X}$ und will interessante Strukturen in den Daten erkennen.

- Gibt es Gruppen ähnlicher Objekte (sogenannte Cluster) in den Daten, die sich von Objekten anderer Gruppen mehr oder weniger deutlich unterscheiden?
- Gibt es Ausreißer in den Daten, d.h. einzelne Objekte, die nicht in die Verteilung der anderen Objekte passen?
- Wie können wir solche (hochdimensionalen) Daten visualisieren?
- Angenommen wir können eine Menge an Gruppen untereinander ähnlicher Objekte identifizieren und haben nun eine neue/weitere Beobachtung – welcher Gruppe/welchem Cluster sollen wir die Beobachtung zuordnen?
- ...

Regression, Klassifikation, Clustering

Supervised

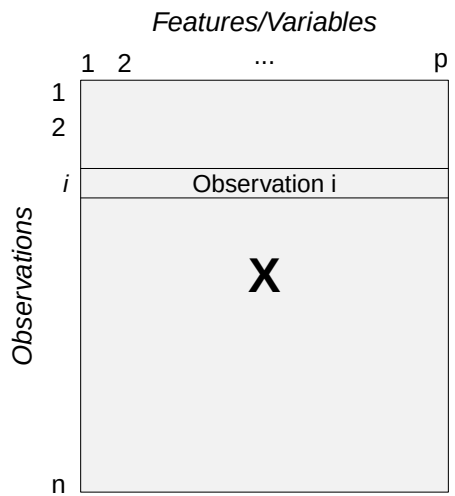
- Vorhersage einer response y mit Prädiktoren x_j
- Klassifikation: y ist qualitativ (kategoriell)
- Regression: y ist quantitativ (numerisch)



`MLR`, `PCR`, `PLS`, `Lasso`, Ridge Regression, Elastic Net, Trees, Random Forests, ... (regression) and `LDA`, `QDA`, `kNN`, `SVM`, ...

Unsupervised

- Entdeckung interessanter Strukturen
- kein y vorherzusagen
- oft Teil der `EDA` (exploratory data analysis)



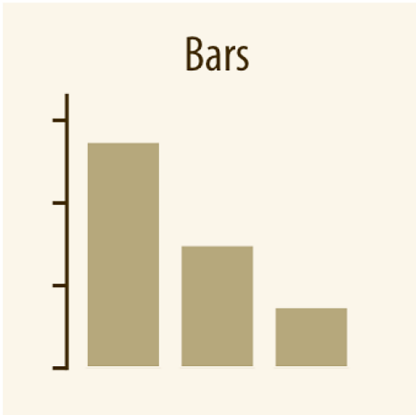
`PCA`, `MDS`, Factor Analysis, Kohonen maps, Hierarchical clustering, model based clustering, `kmeans`, ...

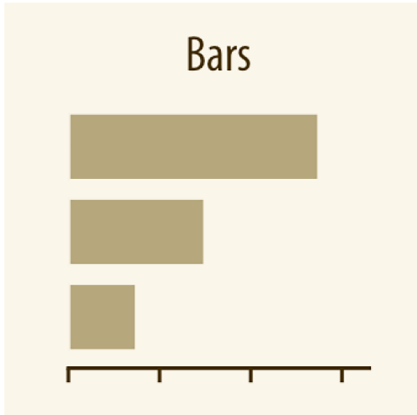
Inhalt

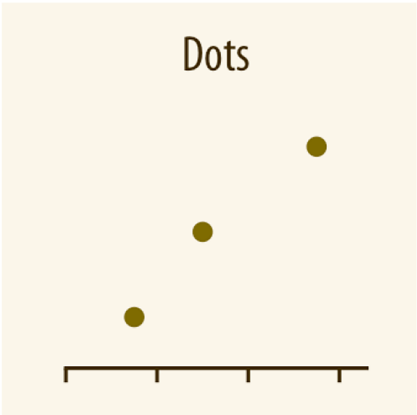
Inhalt

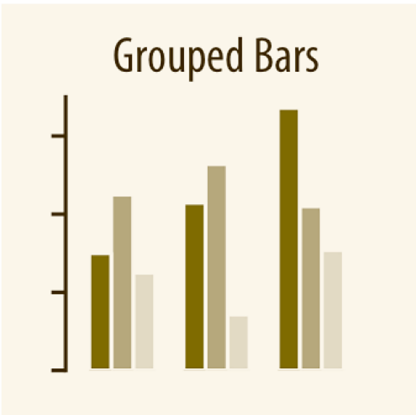
Übersicht

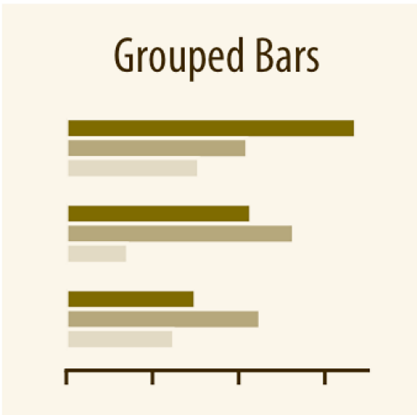
Grafiken entnommen von *Wilke: Fundamentals of Data Visualization. O'Reilly*, 2019

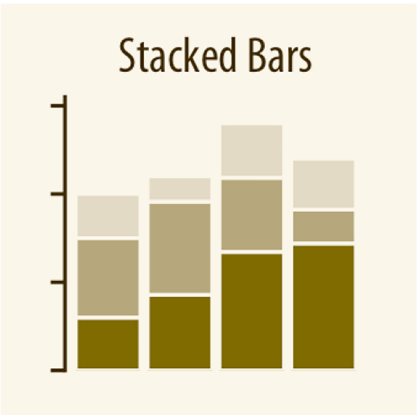




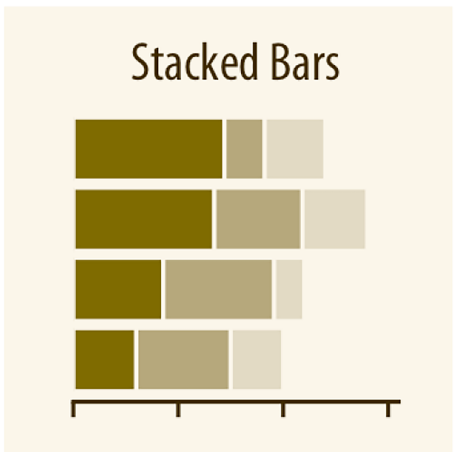








Übersicht



Wie sehen die Originaldaten im Falle von Barplots bzw. grouped/stacked barplots aus?

Darstellung kategorieller Daten

Im univariaten Fall (d.h. es wird nur eine Variable gleichzeitig betrachtet) ist man im Fall kategorieller Daten relativ beschränkt ...

- Balkendiagramm, Säulendiagramm (Unterschied?) sowie sogenannte Dotcharts
- Plotten der absoluten Werte (Anzahlen) oder der Anteile
- Torten- oder Kreisdiagramme
- 3D Diagramme?

Zu beachten:

- Barplots beginnen stets bei 0 und erstrecken sich bis zu dem Wert, der dargestellt werden soll.
- Oft hilfreich: grid lines zum besseren Ablesen der Zahlenwerte.
- Oft problematisch bei Säulendiagrammen (vertikale Darstellung): nicht genügend Platz für die Labels der Balken.
- Abhilfen: rotierte Darstellung (welche aber oft schlecht lesbar ist, ein Leser benötigt ca. um einen Zeitfaktor 2 länger) oder horizontale Darstellung (Balkendiagramm; einfaches Vertauschen der x- und y-Achsen). Von ersterer Lösung ist zumeist abzuraten.

Stacked oder Grouped Barplots

- Eine Variante des Bar- oder Column-Plots ist ein grouped oder ein stacked Barplot: quantitative Variable (oft Anzahl der Beobachtungen) in Abhängigkeit zweier kategorieller Variablen (bi- oder multivariater Fall). Hier werden die Balken nebeneinander oder übereinander (stacked) dargestellt.
- Variationen innerhalb der grouped/stacked barplots: welche kategorielle Variable ist die primäre Variable, welche die sekundäre, nach der gruppiert wird? Dies hängt im wesentlichen davon ab, welche Vergleiche man einfacher machen will, d.h. vor allem erwünscht sind.
- Oft schwer zu lesen ... – welche Vergleiche können einfach durchgeführt werden? Die möglichen bzw. tatsächlich gemachten Vergleiche werden auch durch dominante Farben beeinflusst.

Beispiel Mietspiegel

Pivottables mit 1 und 2 kategoriellen Variable $+$ Pivotchart (2D, 3D Plots, gruppiertes und stacked Balken- bzw. Säulendiagramm)

Darstellung kategorieller Daten

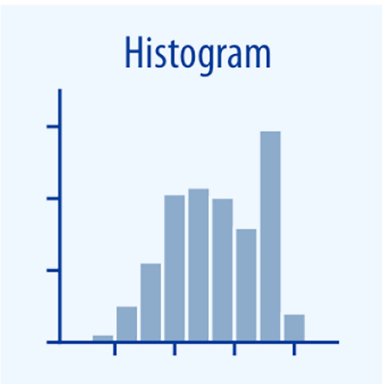
- Interessante Darstellungsart im Falle zweier kategorieller Variablen und einer quantitativen Größe: treemap (hierarchische Darstellungsart)
- Oft gestellte Frage: wie hängen (oder hängen überhaupt) zwei kategorielle Größen zusammen? Antwort siehe später im Bereich der *induktiven Statistik*.

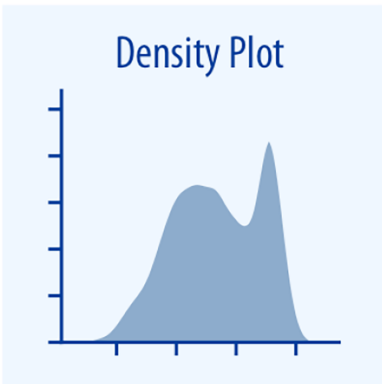
Beispiel Treemap anhand Automobilverkaufszahlen

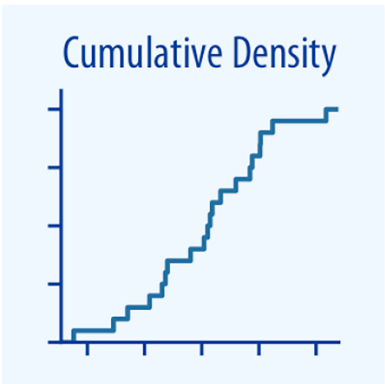
Table mit 2 hierarchischen kategoriellen Variablen und einer numerischen Größe

Inhalt

Übersicht

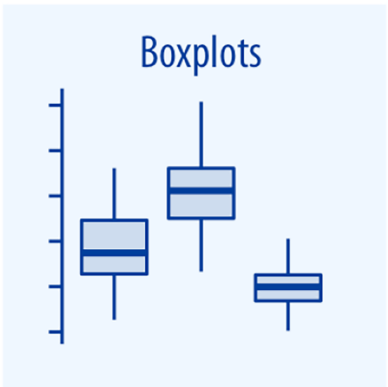

Histogram

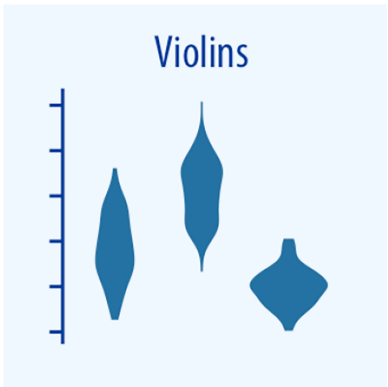

Density Plot

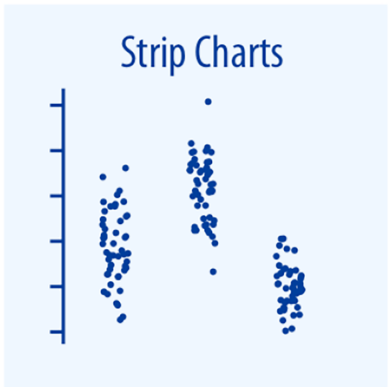

Cumulative Density

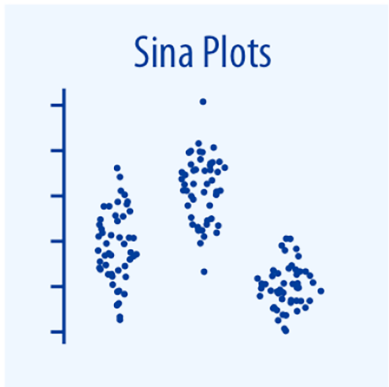

Quantile-Quantile Plot

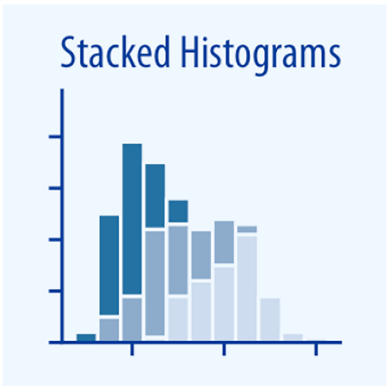
Übersicht

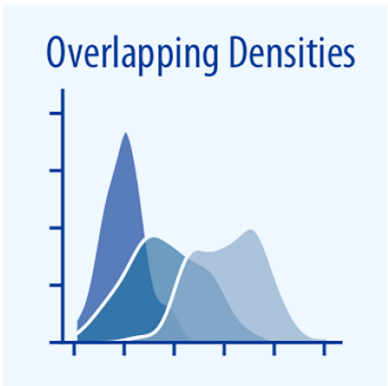

Boxplots

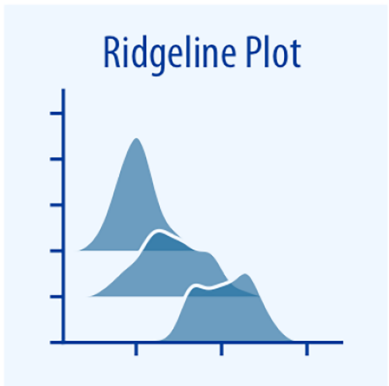

Violins


Strip Charts


Sina Plots


Stacked Histograms


Overlapping Densities


Ridgeline Plot

Histogramme und Density Plots

- Histogramme und Density Plots haben jeweils einen (wesentlichen) Freiheitsgrad (Klassenbreite im Fall von Histogrammen bzw. die Bandbreite oder die Art des kernels im Falle von density plots). Da real auftretenden Wahrscheinlichkeitsdichten oft glatte/stetige Kurven sind, werden density plots oft bevorzugt.
- Oftmals werden Histogramme zusammen mit der geschätzten Dichte dargestellt.
- Histogramme und Density Plots schätzen die zugrundeliegende Wahrscheinlichkeitsdichte, während die Cumulative Density (zu deutsch: empirische Verteilungsfunktion) die Verteilungsfunktion schätzt.
- Histogramme, Density Plots und empirische Verteilungen werden zumeist im 1-sample setting verwendet, können aber auch zum Vergleich mehrerer Verteilungen verwendet werden (Verwendung semitransparenter Farben).
- Müssen bei Histogrammen die Klassen gleiche Breite aufweisen?

Wahl der Klassenbreite in Histogrammen

Zur Wahl einer Klassenbreite in Histogrammen existieren zahlreiche Regeln. Im folgenden bezeichnet k die Anzahl der (gleichbreiten) Klassen, h die (einheitliche) Klassenbreite, IQR den (empirischen) Interquartilsabstand der Daten und $\hat{\sigma}$ einen Schätzwert für die Standardabweichung (zumeist die empirische Standardabweichung).

- Square Root Rule:

$$k = \sqrt{n}$$

- Sturges Rule:

$$k = \lceil 1 + \log_2 n \rceil$$

- Scott's Rule:

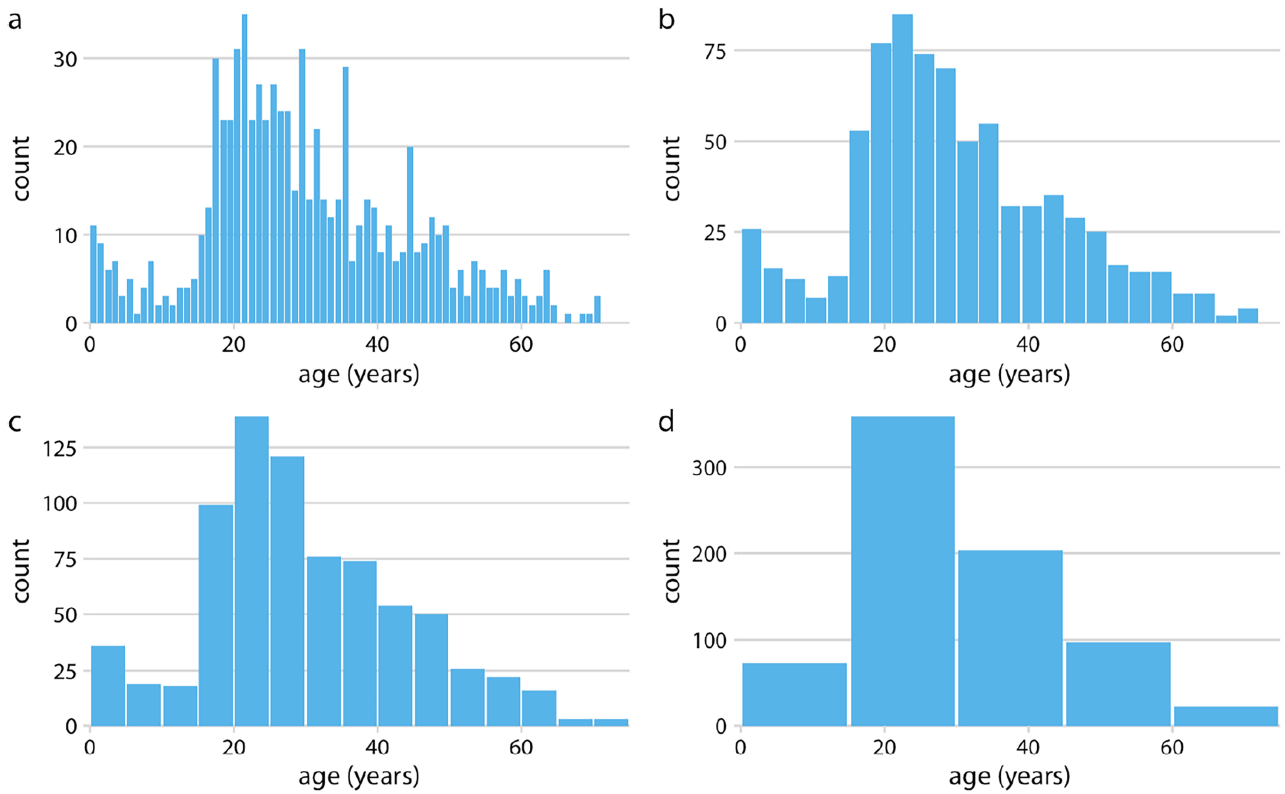
$$h = \frac{3.5 \cdot \hat{\sigma}}{n^{\frac{1}{3}}}$$

- Freedman-Diaconis (FD) Rule:

$$h = \frac{2 \cdot \mathrm{IQR}\,(x)}{n^{\frac{1}{3}}}$$

Wahl der Klassenbreite in Histogrammen

- Die Wahl einer richtigen Klassenbreite bzw. Anzahl an Klassen ist entscheidend – eine zu große Klassenbreite lässt Details verschwinden, eine zu kleine Klassenbreite erzeugt ein Histogramm mit zu vielen (und nicht interpretierbaren) Peaks (untenstehendes Beispiel: Altersverteilung der Titanic-Passagiere).
- vertikale Achse: absolute oder relative Häufigkeiten.

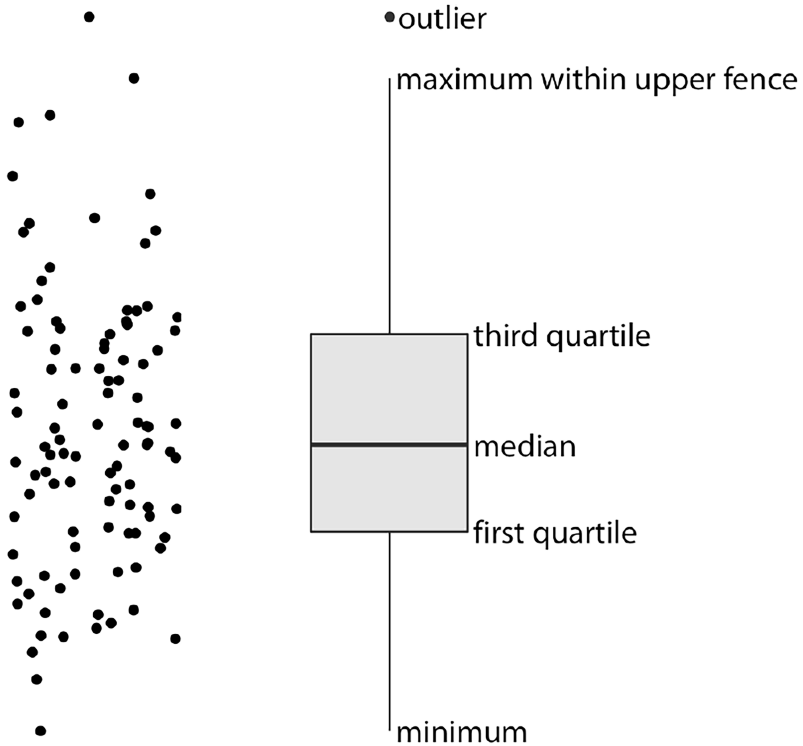


Histogramme in Excel

Excel – Histogramm einer numerischen Variable

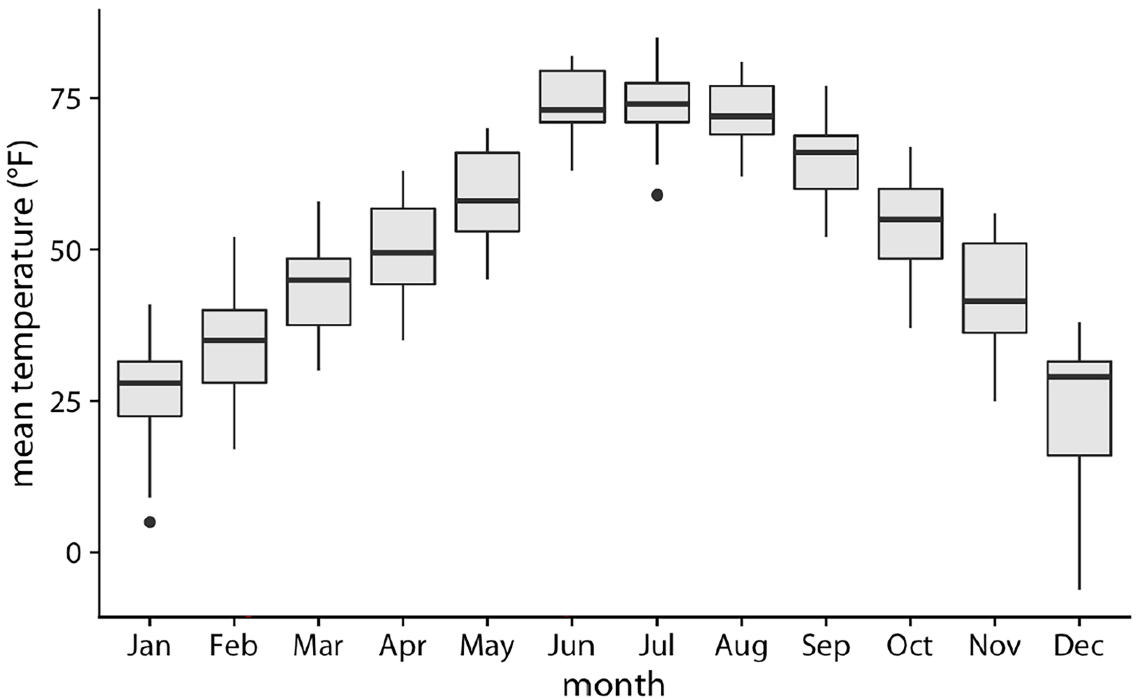
Boxplots

- Neben Histogrammen sind Boxplots die wohl am weitest verbreiteten Plots zur Visualisierung der Verteilung von Daten.
- Was ist in einem Boxplot dargestellt bzw. kann abgelesen/abgeleitet werden? Was kann nicht abgelesen werden bzw. welche Art von Verteilungen sind nicht als solche in Boxplots erkennbar?



Boxplots

- Boxplots sind gut geeignet, eine (begrenzte) Anzahl an Datenreihen miteinander zu vergleichen.
- Man beachte, dass es auch hier **nicht** notwendig/sinnvoll ist, jeden Boxplot in einer anderen Farbe darzustellen (außer man hat dafür triftige Gründe).



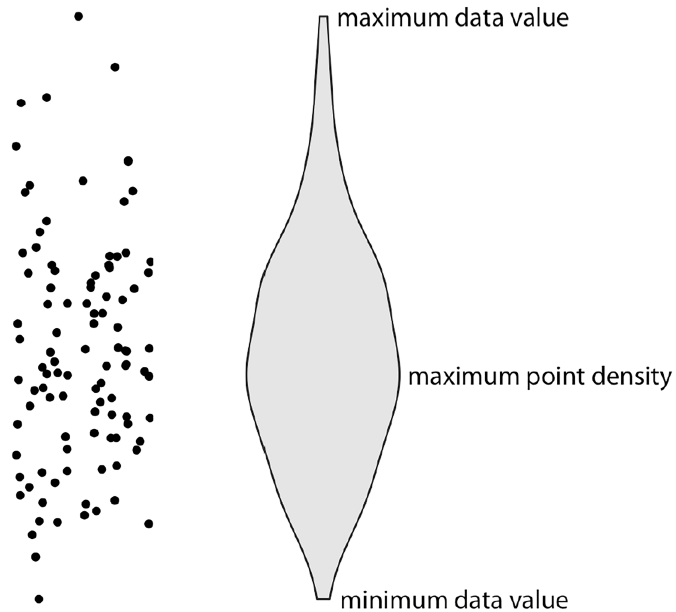
Histogramme in Excel

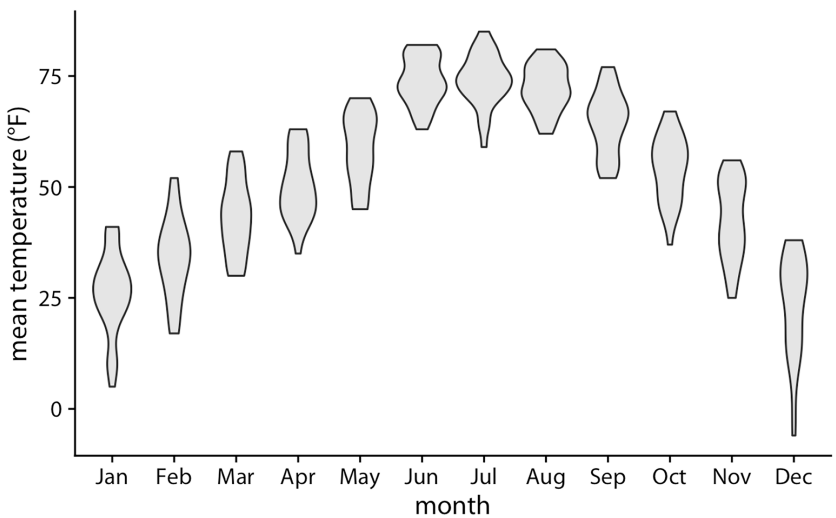
Excel – Boxplots einer numerischen Variable

Eventuelle Aufgliederung nach einer kategoriellen Variable (Gruppierungsvariable) zum Vergleich zweier (oder mehrerer) Verteilungen

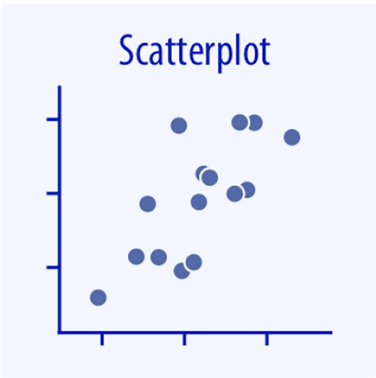
Violinplots

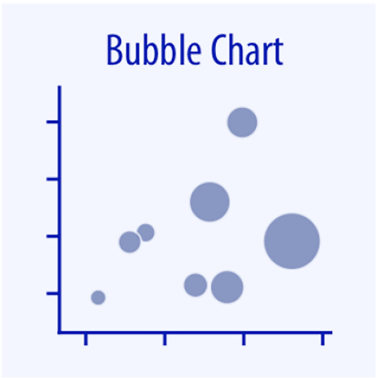
- Eine Weiterentwicklung von Boxplots stellen Violinplots dar – es sind um $90°$ rotierte Dichteschätzungen der Daten.
- Im Unterschied zu Boxplots können so auch bi- oder multimodale Verteilungen erkannt werden.

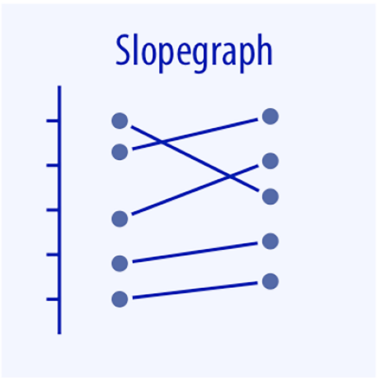




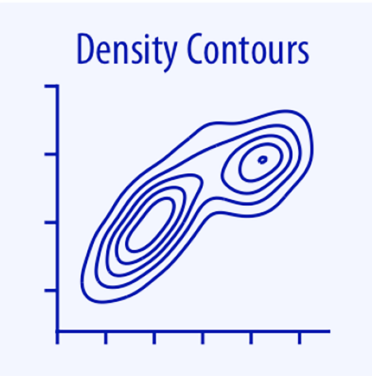
Übersicht

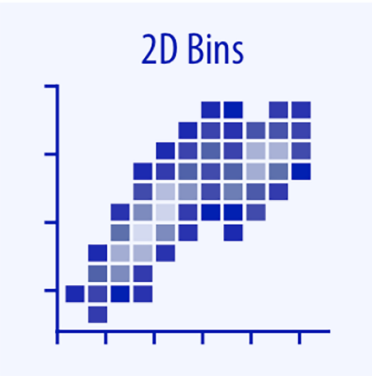


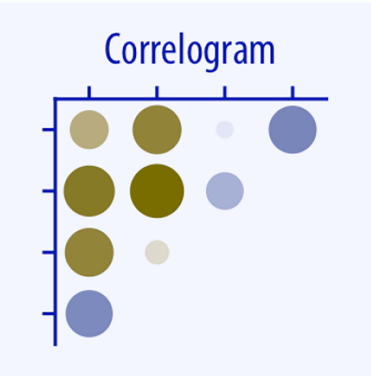




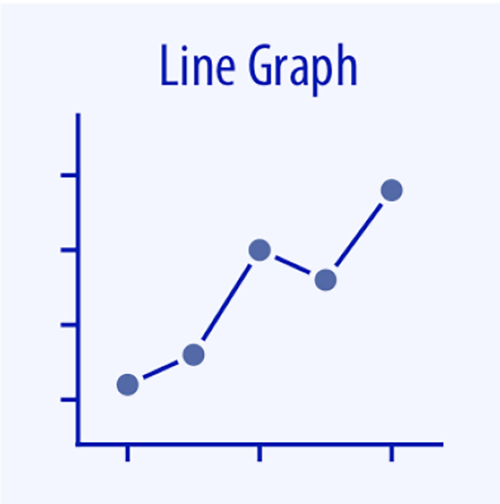
Übersicht

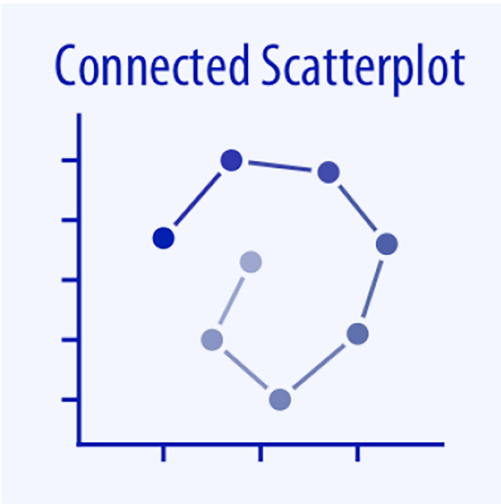


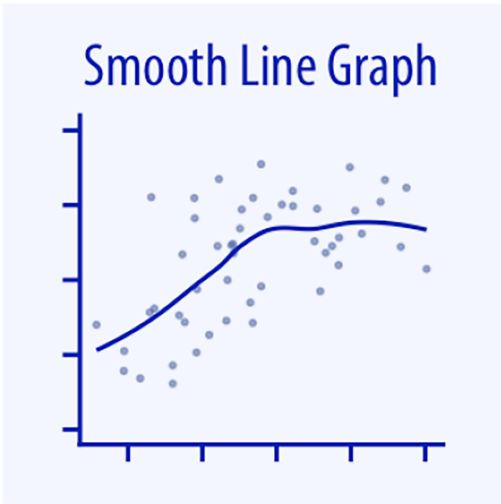




Übersicht





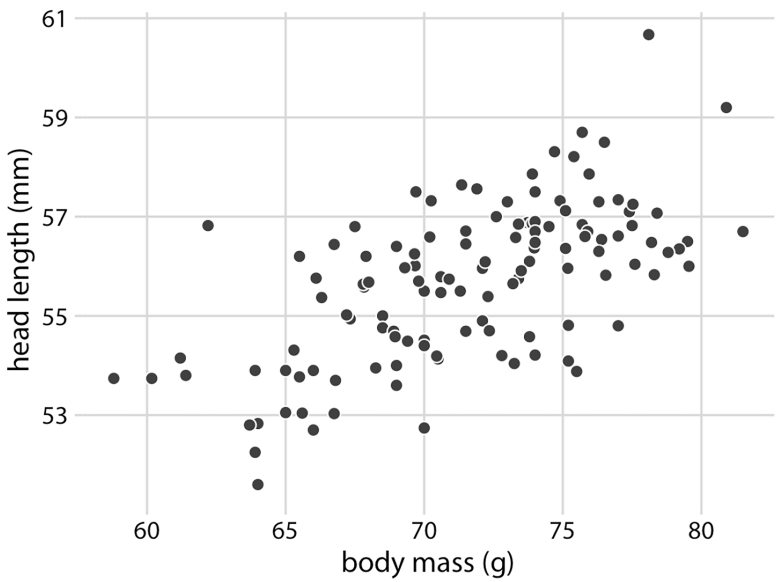


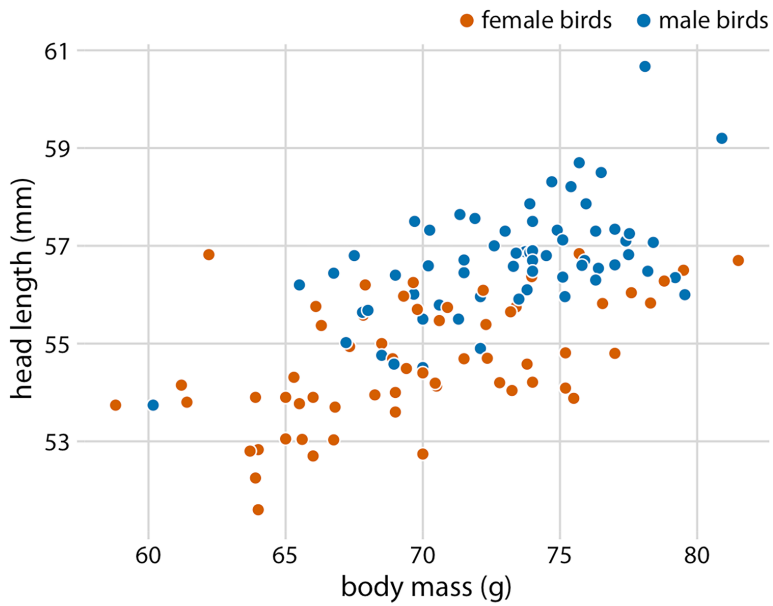
Scatter- und Bubbleplots

- Ein Scatterplot ist der wohl am häufigsten verwendete Diagrammtyp zur Darstellung zweier quantitativer Größen.
- Soll eine weitere quantitative Variable dargestellt werden, ist prinzipiell ein dreidimensionaler Scatterplot vorstellbar, allerdings nur im interaktiven Umfeld (Drehungen der Punktwolke durchführbar). Zielführender ist meist ein sogenannter Bubbleplot, indem eine dritte Variable über die Größe des plotting characters kodiert wird.
- Verschiedene Gruppen (Subpopulationen) können beispielsweise über verschiedenfärbige plotting Symbole untergebracht werden.

Scatter- und Bubbleplot

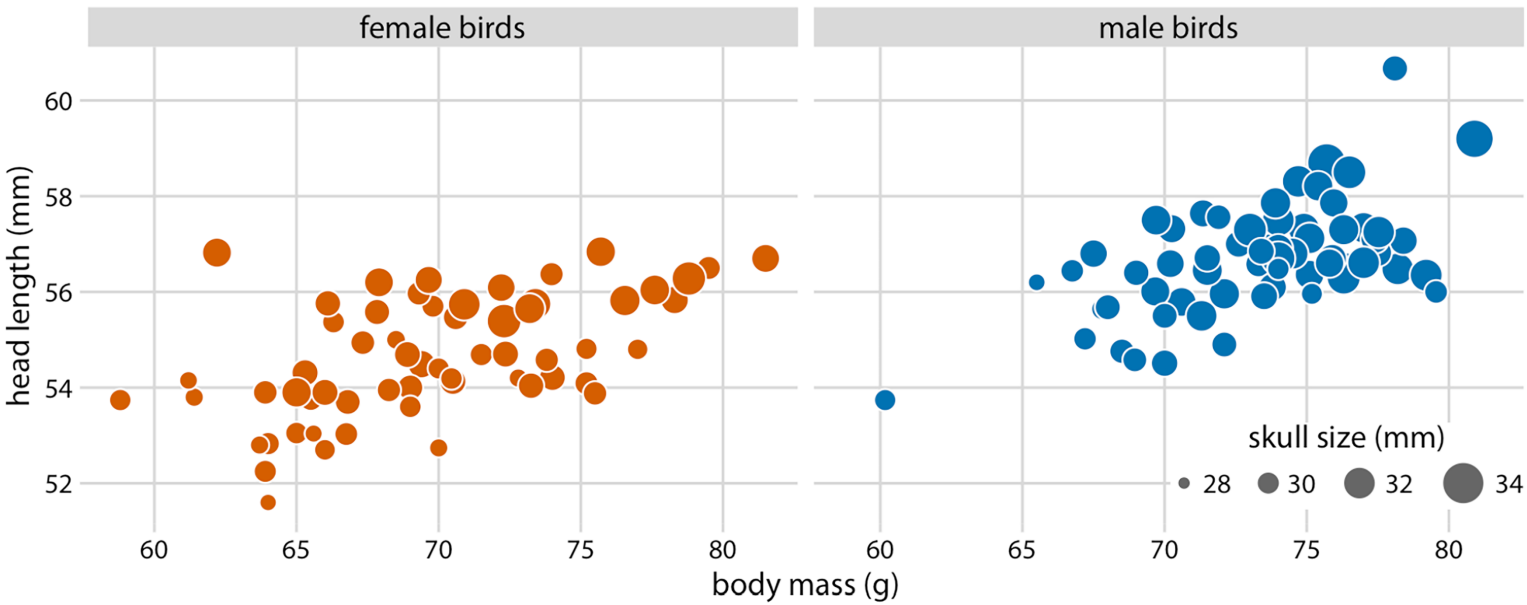
Blue jay (Blauhäher) Datensatz: $n = 123$ Beobachtungen der Kopflänge, der Schädelgröße und des Körpergewichts für männliche und weibliche Vertreter dieser Vogelart.





Scatter- und Bubbleplot

Man beachte die separate Beschriftung der Schichten (*male* und *female* birds), die einfache Beschriftung sowohl der horizontalen als auch der vertikalen Achse sowie die **nicht** doppelt ausgeführte Legende und schließlich die gleiche Skalierung der Achsen in beiden Teilplots.



Grundsätzlich ist auch **ein** plot für beide Untergruppen (*male* und *female* birds) denkbar, die einzelnen Punkte wären allerdings schlecht wahrzunehmen und würden deutlich überlappen. Die Anzahl an Größenklassen sollte relativ klein gehalten werden (Unterscheidbarkeit der Größen der plotting character), d.h. maximal ca. 4.

Histogramme in Excel

Excel – Scatterplots und Bubbleplots

Interpretation der Regressionsgeraden (Trendlinie) sowie des Bestimmtheitsmasses R^2.

Inhalt

Beschreibung von Stichproben

- Neben Grafiken (z.B. Histogrammen) zur Beschreibung von Verteilungen sind oft auch numerische Beschreibungen notwendig, sogenannte statistische Kennzahlen oder Kenngrößen. Sie sollen die Daten verdichten/komprimieren.
- Wir unterscheiden Kennzahlen zur Beschreibung
  - der Lage (Lokation) – wo liegt das **Zentrum**/der Schwerpunkt der Verteilung/der Daten? arithmetisches Mittel und Median
  - der Variabilität – wie stark **schwanken** die Beobachtungen rund um das Zentrum? Varianz, Standardabweichung und Variationskoeffizient
  - der Form der Verteilung – ist eine Verteilung schief oder symmetrisch?

Inhalt

Arithmetisches Mittel $\overline{x}$

Für metrische Beobachtungen $x_1, x_2, \ldots, x_n$ ist der arithmetische Mittelwert (das arithmetische Mittel, Stichprobenmittel) definiert als

$$\overline{x} = \overline{x}_n := \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Ein Nachteil des arithmetischen Mittels ist seine Empfindlichkeit gegenüber extremen Werten. Ein möglicher Ausweg ist das getrimmte Mittel.

In `Excel` werden beim α-getrimmten Mittel von beiden Seiten der sortierten Daten $\lfloor \frac{\alpha}{2} \cdot n \rfloor$ Datenpunkte weggelassen und von den restlichen Daten der herkömmliche Mittelwert gebildet.

Median x_{med}

- Der Median ist ein weiteres Beispiel für ein robustes oder resistentes Lagemaß.
- Man geht aus von der (aufsteigend) geordneten Stichprobe

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Man beachte, dass im allgemeinen **nicht** gilt, dass $x_i = x_{(i)}$. Außerdem müssen die Daten mindestens ordinales Skalenniveau haben.

Definition (Median)

Der Median x_{med} ist ein Lagemaß, das die Mitte der Daten $x_1, x_2, \ldots, x_n$ beschreibt. Für eine ungerade Anzahl n an Beobachtungen, ist der Median der mittlere Datenwert in der geordneten Stichprobe. Im Falle eines geraden n ist der Median als das arithmetische Mittel der beiden mittleren Werte definiert:

$$x_{\text{med}} = \begin{cases} x_{((n+1)/2)} & n \text{ ungerade} \\ \frac{1}{2} \cdot \left(x_{(n/2)} + x_{(n/2+1)} \right) & n \text{ gerade} \end{cases}$$

Dabei ist $x_{(i)}$ der i-te Wert in der aufsteigend geordneten Stichprobe.

Excel – Funktionen MITTELWERT, GESTUTZTMITTEL und MEDIAN

Inhalt

Streuung

- Eine Information über die Lage einer Verteilung ohne begleitende Information zur Streuung ist oft wertlos, zumindest aber unvollständig.
- Die Streuung enthält die Information, wie weit die Beobachtungen x_i (im Mittel) von der Lokation der Verteilung entfernt sind.

Varianz und Standardabweichung

- Die empirische Standardabweichung s bzw. die empirische Varianz s^2 (= Quadrat der Standardabweichung) messen die Streuung der Daten x_i $(i = 1, \ldots, n)$ um den Mittelwert $\overline{x}$.
- Beide Streuungsmaße sind nur für metrische Daten (intervall- oder verhältnisskaliert) geeignet.

Die empirische Varianz bzw. Standardabweichung sind wie folgt definiert:

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

Sind die Beobachtungswerte x_i dimensionsbehaftet, so trägt die Varianz diese Einheit^2, während die Standardabweichung die gleiche Einheit wie die ursprünglichen Beobachtungen aufweist. Daher ist letztere in den allermeisten Fällen einfacher zu interpretieren.

Variationskoeffizient v

Ein Vergleich zweier Standardabweichungen ist problematisch – da s dimensionsbehaftet ist, ändert sich der Wert von s z.B. mit einem Wechsel der Einheiten. Es scheint daher sinnvoll, s auf den Mittelwert $\overline{x}$ zu beziehen, um eine dimensionslose Größe zu erhalten.

Definition (Variationskoeffizient)

Für gegebene Daten $x_1, x_2, \ldots, x_n \geq 0$ mit arithmetischem Mittel $\overline{x} \neq 0$ und Standardabweichung s ist der Variationskoeffizient v definiert als

$$v = \frac{s}{\overline{x}}$$

`Excel` – Funktionen `VAR.p`, `STABW.s`

Mittel/Median der Absolutdistanzen

Ein weiteres Streuungsmaß für metrische Merkmale ist der Median der Absolutabweichungen. Für gegebene Daten $x_1, x_2, \ldots, x_n$ ermitteln sich die Absolutabweichungen der x_i vom Median x_{med} als:

$$|x_1 - x_{\text{med}}|, \quad |x_2 - x_{\text{med}}|, \quad \ldots, \quad |x_n - x_{\text{med}}|$$

Der Median dieser n Absolutabweichungen wird oft als `MAD` oder `MEDMED` bezeichnet.

Definition

Die mittlere absolute Abweichung `MAD` für gegebene Daten $x_1, x_2, \ldots, x_n$ ist gegeben durch

$$\text{MAD} = \text{MedMed} = \text{median}(|x_i - x_{\text{med}}|)$$

Einfache Interpretation: Die Hälfte der Werte x_i ist weniger als `MedMed` vom Median entfernt, die andere Hälfte weiter als `MedMed` weg. Der `MedMed` ist (im Gegensatz zu s, s^2 oder v) ein robustes Streuungsmaß.

`Excel` – händische Berechnung

Inhalt

Kovarianz und Korrelation

- Wir interessieren wir uns nun für den Zusammenhang zweier numerischer Zufallsgrößen X und Y bzw. zweier Datenreihen, z.B. im Mietpreisbeispiel die Größen Nettomiete und Wohnfläche.
- Haben wir n Wertepaare

$$(x_1, y_1)\,,\ (x_2, y_2)\,, \ldots,\ (x_n, y_n)$$

gegeben, so definieren wir zunächst die Stichprobenkovarianz s_{xy} durch

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n \cdot \overline{x} \cdot \overline{y}\right)$$

- **Problem**: Größe hängt von der Varianz von X und Y ab (nicht normiert).

Kovarianz und Korrelation

- Dividiert man die Kovarianz durch die beiden empirischen Standardabweichungen s_x und s_y erhält man als dimensionslose Größe den (Pearson–) Stichprobenkorrelationskoeffizienten r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- Die Werte des Korrelationskoeffizienten r_{xy} liegen im Bereich

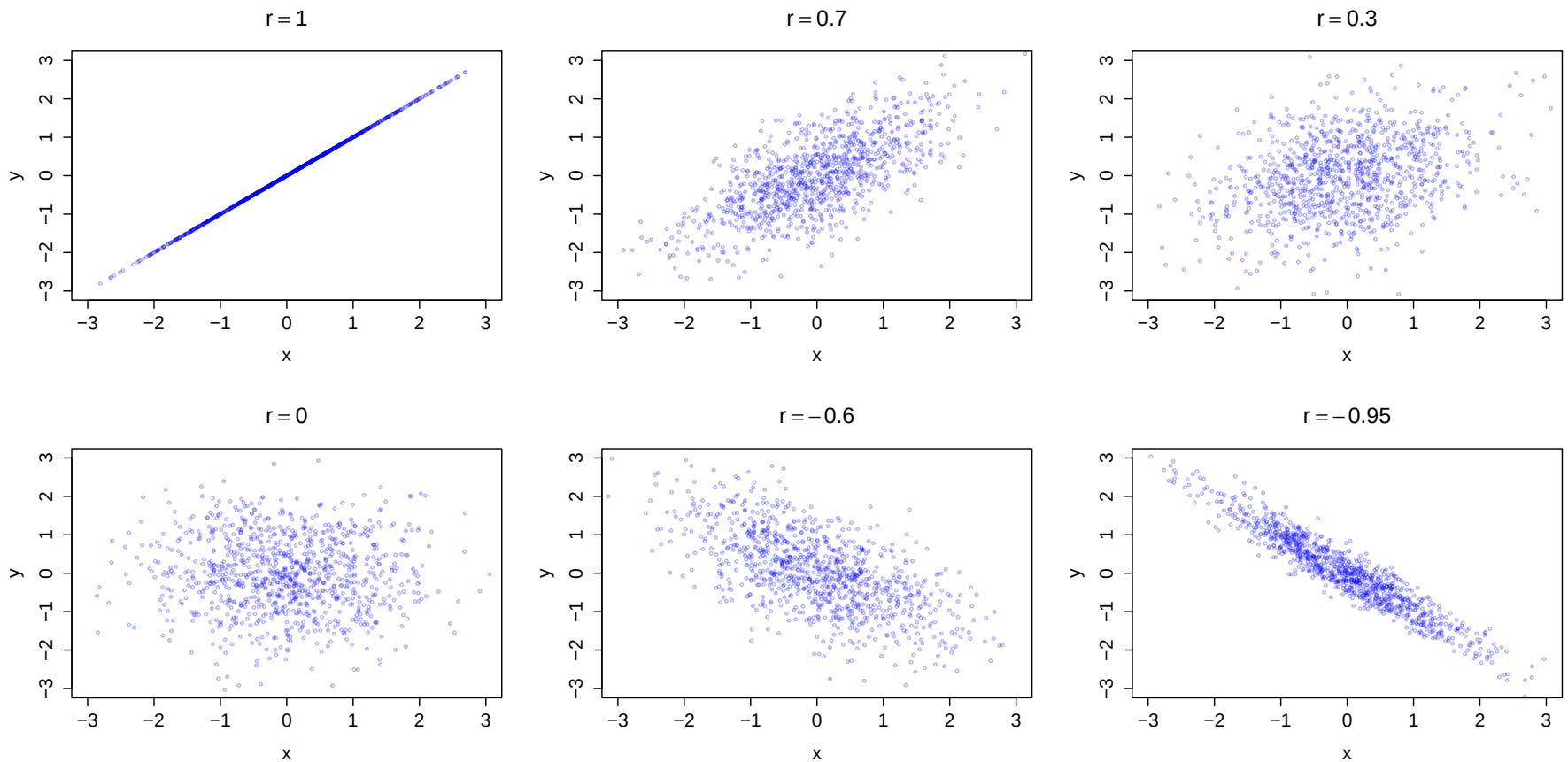
$$-1 \leq r_{xy} \leq +1$$

  Dabei bedeutet $r_{xy} = 0$ keinen linearen Zusammenhang und $|r_{xy}| = 1$ ... einen vollkommenen linearen Zusammenhang.

- Das Vorzeichen von r_{xy} gibt die Richtung des Zusammenhangs an, d.h. ob mit steigenden Werten von X auch die Werte von Y (tendenziell) ansteigen (positive Korrelation) oder fallen (negative Korrelation).

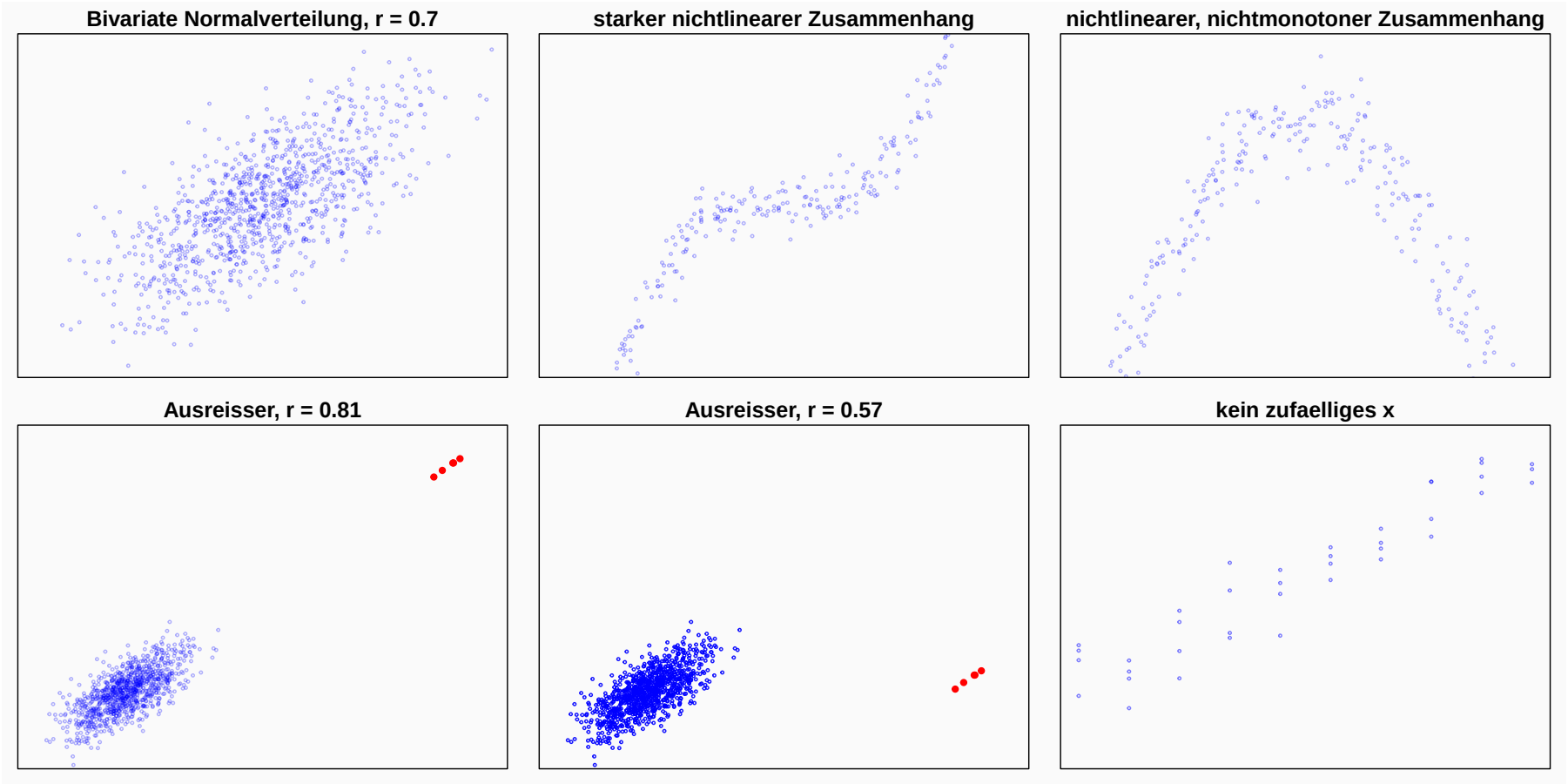
Pearson-Korrelation

Die Pearson Korrelation ist ein Maß für die Stärke des linearen Zusammenhangs.



Pearson-Korrelation

Nur in der ersten Situation ist die Berechnung der (Pearson) Korrelation uneingeschränkt zu empfehlen.



Pearson-Korrelation

Excel – `Daten - Datenanalyse - Korrelation/Kovarianz`

- Eventuell ist die Installation des `Datenanalyse` plugins notwendig (Datei – Optionen – Add-ins – auf `Los` und dort `Analyse`-Funktionen anhaken).
- Die Kovarianz und Korrelation kann auch für mehr als 2 Variablen berechnet werden – das Ergebnis ist die Kovarianz- bzw. Korrelationsmatrix.
- Haben die Daten lediglich ordinales Skalenniveau oder ist der Zusammenhang zwischen X und Y lediglich monoton und nicht linear oder befinden sich Ausreißer in den Daten, so ist der Pearson'sche Korrelationskoeffizient nicht geeignet.
- Der Spearman'sche Rangkorrelationskoeffizient und Kendall's τ sind auch für solche Daten geeignet und basieren auf den Rängen anstatt auf den Originaldaten.

Inhalt

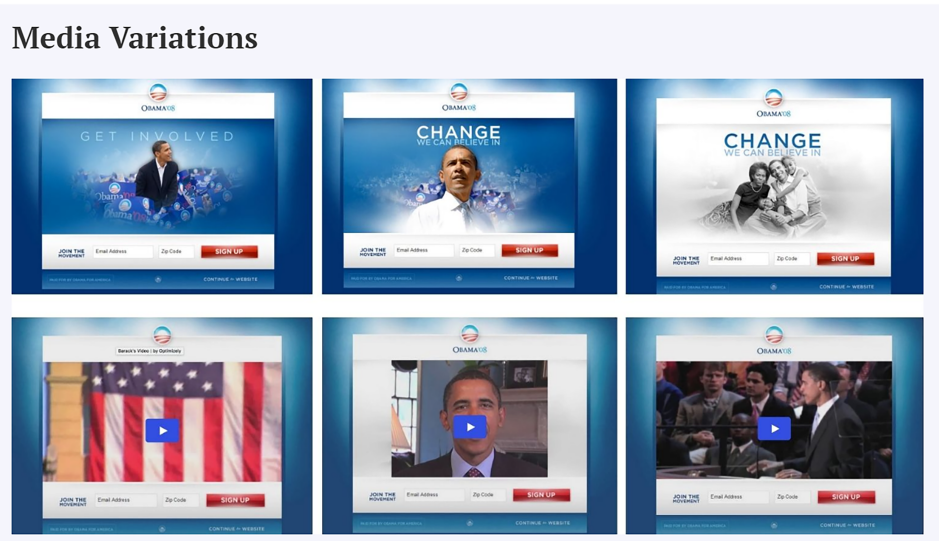
Inhalt

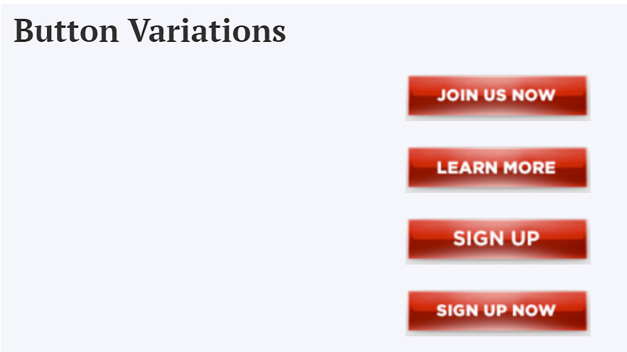
A/B Testing

- Ziel des A/B Testing: Testen einer Produktidee direkt am Kunden – was funktioniert, was nicht?
- Zumeist Testen eines neuen Produktdesigns (oder Newsletterdesigns) oder einer Veränderung an einer Software.
- Idee des A/B Testing: Erstellen eines Prototyps des Produkts (z.B. 2 Versionen einer Webpage) und sammeln des **user feedback**.
- Man erhofft sich **Antworten** auf Fragen wie *Welche Änderungen sind am Produkt noch notwendig?* oder *Soll das Produkt überhaupt noch weiterentwickelt werden?* oder *Sind die Features des Produkts verständlich und einfach zu bedienen?* oder *Enhält das Produkt Fehler?*

A/B Testing – Anwendungsgebiete

Webseiten – Design (welcher Hintergrund soll gewählt werden, welche Farbe hat ein bestimmter Button – rot oder grün, wie sollen einzelne Blöcke auf der Seite organisiert sein); Beispiel Auftritt von Politikern direkt vor Wahlen (Obama) oder bei Spendenaufrufen in Wahlkampagne etc.





Wahl 2008, untersucht wurde der Einfluß des Designs des **Call-to-action Buttons** (4 Varianten) und die gezeigten Medien (6 Varianten – 3 Bilder, 3 Videos).

A/B Testing – Anwendungsgebiete

- **Ergebnis**: Metrik war hier die **Sign-Up Rate**, d.h. die Anzahl an Anmeldungen im Verhältnis zur Anzahl der Seitenaufrufe.
- Gesamt: > 300.000 Seitenaufrufe, daher pro Kombination eines Mediums (Bild/Video) und eines Call-to-action Buttons immer noch > 10.000 Besucher.
- Die Originalseite hatte eine *sign-up rate* von $8.3\,\%$, die beste Variante $11.6\,\%$, entsprechend einer Verbesserung um ca. $40\,\%$.
- Im Rahmen der Kampagne wurden so mehr als 2.8 Millionen zusätzliche Emailadressen eingesammelt und ein zusätzlicher Spendenumsatz von 60 Millionen Dollar erzielt.
- Wie wird es gemacht? Beim A/B Testing wird der **traffic** in 2 gleiche Teile **gesplittet** – $50\,\%$ der Besucher der Webseite bekommen Variante A präsentiert, $50\,\%$ Variante B und die Metrik wird gemessen.
- Daten entnommen von https://www.mailmunch.com/blog/ab-testing-got-obama-60-million.

A/B Testing – Beispiel

- Beispiel: Bei einer sample size von $n = 200$ zeigen wir $n_A = 100$ Besuchern der Webseite die Variante A (control), $n_B = 100$ Besuchern der Webseite Variante B (treatment, neues Feature).
- Wir erhalten conversion rates von (Angaben als Anteile, A ist herkömmliche Variante, B neue Variante einer Webpage)

$$p_A = 0.08 \qquad \text{und} \qquad p_B = 0.12$$

- Welche Schlüsse ziehen wir? Sind wir damit schon fertig?

A/B Testing – Beispiel

- Wir sind interessiert am Verhalten der Population, haben aber nur die Daten einer (kleinen) Stichprobe zur Verfügung. Stichproben sind per definitionen ein Zufallsprodukt, d.h. würden wir die Untersuchung wiederholen, würden wir (höchstwahrscheinlich) andere Werte erhalten.
- Wir müssen uns fragen: Sind die in der Stichprobe festgestellten Unterschiede ($8\,\%$ für Variante A, $12\,\%$ für Variante B) noch mit der Zufälligkeit der Stichprobenziehung zu erklären oder ist der Unterschied dafür schon zu groß und wir können darauf vertrauen, dass ein solcher Unterschied auch in der Population vorhanden ist (signifikanter Unterschied).

A/B Testing – Beispiel

- In der Statistik geschieht dies mittels sogenannter Hypothesentests. Wir formulieren dazu eine sogenannte Null- und eine Alternativhypothese:

$$\mathcal{H}_0 : p_A = p_B \qquad \text{versus} \qquad \mathcal{H}_1 : p_A \neq p_B$$

  In der Nullhypothese steht zumeist die *harmlose Situation*, die besagt, dass die beiden conversion rates gleich sind, die Alternativhypothese behauptet grundsätzlich das Gegenteil.

- Man berechnet nun mittels der erhobenen Daten eine sogenannte Testgröße und lässt sich (von einer geeigneten Software) einen zugehörigen p-Wert ausgeben.

- Der p-Wert gibt an, wie wahrscheinlich es ist, einen solchen Wert für die Teststatistik zu erhalten wie jener, den man erhalten hat, oder noch extremer, wenn die Nullhypothese $\mathcal{H}_0$ timmt. Ist diese Wahrscheinlichkeit (also der p-Wert) gering (in der Regel nimmt man als Grenze 0.05), so schenkt man der Nullhypothese keinen Glauben und vertraut ab sofort der Alternativhypothese. Ist der p-Wert vergleichsweise groß, d.h. ≥ 0.05, so vertraut man weiterhin der Nullhypothese.

A/B Testing – Beispiel

- Wir haben einen 2-Stichproben Test für Anteile vor uns. In diesem Fall lautet die Testgröße

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot (\frac{1}{n_A} + \frac{1}{n_B})}}$$

  wobei $\hat{p}_A$ und $\hat{p}_B$ die Stichprobenanteile (nicht zu verwechseln mit den wahren Anteilen in der Population p_A und p_B) sind, n_A und n_B die Stichprobenumfänge in den beiden Gruppen und $\hat{p} = \frac{n_A \cdot \hat{p}_A + n_B \cdot \hat{p}_B}{n_A + n_B}$ (Erfolgsquote in den beiden Stichproben kombiniert).

- Wir erhalten zunächst für

$$\hat{p} = \frac{8 + 12}{200} = 0.1$$

  und für die Teststatistik

$$Z = \frac{0.08 - 0.12}{\sqrt{0.1 \cdot 0.9 \cdot (\frac{1}{100} + \frac{1}{100})}} = -0.94$$

- Der p-Wert lautet 0.34 (siehe nächste Folie) und ist damit > 0.05. Es ist also nicht so unwahrscheinlich, bei Gültigkeit der Nullhypothese solche Unterschiede zu beobachten, wir vertrauen also weiterhin darauf, dass die conversion rates in den beiden Populationen (Personen, die Varianten A bzw. B gezeigt bekommen), gleich sind. Das Ergebnis ist nicht signifikant.

- `Excel` – Funktion `STANDNORMVERT`

Inhalt

A/B Testing – Konfidenzintervalle

- Eng verbunden mit statistischen Hypothesentests sind sogenannte Konfidenzintervalle.
- Stellen wir uns vor, wir haben die Population aller potentiellen User/Kunden unserer Webseite, an der wir interessiert sind. Unglücklicherweise haben wir wieder nur Zugang zu einer Stichprobe.
- Von Interesse wäre die *conversion rate* der Population/Grundgesamtheit, r, die allerdings unzugänglich ist. Daher nehmen wir eine Stichprobe des Umfangs n und können damit die conversion rate in der Stichprobe berechnen:

$$\hat{r} = \text{CR (Stichprobe)} = \frac{\text{Conversions}}{\text{Anzahl Besucher}}$$

- Es liegt nahe, als Schätzwert für die conversion rate der Population die conversion rate der Stichprobe zu verwenden. Das statistische Gesetz der großen Zahlen sagt uns, dass das sinnvoll ist und dass dieser Schätzwert umso näher beim wahren Wert r liegt, je größer der Stichprobenumfang n ist.
- Wie groß ist die Wahrscheinlichkeit, dass dieser Schätzwert $\hat{r}$ den wahren Wert der Population, r, genau trifft?

A/B Testing – Konfidenzintervalle

- Es liegt somit nahe, nicht einen einzelnen Wert (sogenannter Punktschätzwert), sondern ein Intervall anzugeben, in dem sich der gesuchte Populationsparameter r mit hoher Sicherheit (z.B. $95\,\%$) befindet. Dies ist ein sogenanntes Konfidenzintervall.
- Es lautet:

$$\hat{r} - z_{1-\alpha} \cdot \sqrt{\frac{\hat{r}(1-\hat{r})}{n}} \leq p \leq \hat{r} + z_{1-\alpha} \cdot \sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$$

  Dabei ist $\hat{r}$ der Stichprobenwert der conversion rate, $z_{1-\alpha}$ ein Quantil der Standardnormalverteilung ($z = 1.96$ für eine $95\,\%$ige Sicherheit und $z = 2.58$ für eine $99\,\%$ige Sicherheit), der die Sicherheit widerspiegelt, die wir mit unserem Intervall verbinden und n der Stichprobenumfang.
- Man erkennt, dass das Intervall umso kleiner wird (die Schätzung also umso **genauer**), je höher der Stichprobenumfang n ist.

A/B Testing – Konfidenzintervalle

Beispiel

Wir nehmen eine Stichprobe von $n = 500$ Besuchern unserer Website und bemerken, dass sich 40 für den Newsletter anmelden. Wir haben einen Punktschätzwert für die conversion rate von

$$\hat{r} = \frac{40}{500} = 0.08$$

Ein $95\,\%$iger Konfidenzintervall ergibt sich zu

$$0.08 - 1.96 \cdot \sqrt{\frac{0.08 \cdot 0.92}{500}} \leq r \leq 0.08 + 1.96 \cdot \sqrt{\frac{0.08 \cdot 0.92}{500}}$$

und nach Ausrechnen $[0.056, 0.104]$. Man kann also mit einer Sicherheit von $95\,\%$annehmen, dass die wahre conversion rate in der Population in diesem Intervall liegt. Möchte man hingegen $99\,\%$ige Sicherheit, so lautet das Intervall $[0.049, 0.111]$, es ist also breiter.


DIHSÜD


FFG
Forschung wirkt.


Das Land
Steiermark


LAND KÄRNTEN


FH JOANNEUM
University of Applied Sciences