⟨ S t a t W t h 1 7 ⟩

1. Deskriptive und explorative Statistik

Werner Gurker

Institut für Stochastik und Wirtschaftsmathematik

Technische Universität Wien

Deskriptive Statistik

Die deskriptive (oder beschreibende) Statistik beschäftigt sich mit der tabellarischen und grafischen Aufbereitung von Daten sowie mit ihrer zahlenmäßigen Beschreibung (Berechnung von Kenngrößen).

In der deskriptiven Statistik verwendet man keine statistischen (oder stochastischen) Modelle, sodass die auf Basis der Daten gewonnenen Erkenntnisse nicht durch Fehlerwahrscheinlichkeiten abgesichert werden können.

Dies lässt sich mit Hilfe der schließenden (oder inferentiellen) Statistik erreichen, sofern die unterstellten Modellannahmen – zumindest näherungsweise – zutreffen.

Explorative Statistik

Die explorative Datenanalyse (kurz EDA) wurde in den 1970er Jahren vom US-amer. Mathematiker, Statistiker und Informatiker J. W. Tukey begründet.

Die EDA versucht unbekannte Strukturen und Zusammenhänge in den Daten aufzudecken und Hypothesen über den datengenerierenden Prozess zu formulieren.

Wird u.a. auch beim Data Mining (Verarbeitung großer Datenmengen) verwendet.

Grundgesamtheit

Daten werden an statistischen Einheiten erhoben, durch Experimente oder durch Beobachtungsstudien.

Im ersten Fall spricht man auch von Versuchseinheiten, im zweiten Fall auch von Beobachtungseinheiten.

Die statistischen Einheiten, über die – deskriptiv und/oder explorativ – Aussagen getroffen werden sollen, bilden die Grundgesamtheit oder Population.

Bsp: Kleine oder mittlere IT-Unternehmen (von Interesse ist z.B. deren Wirtschaftskraft); Hörer_innen der VO "StatWth17" (von Interesse ist z.B. der hauptsächlich verwendete Web–Browser); etc.

Assoziation ist nicht Kausalität

Statistische Untersuchungen werden häufig zur Bestätigung (oder zur Widerlegung) von kausalen Zusammenhängen herangezogen.

Achtung: Im strengen Sinn erlauben nur – adäquat durchgeführte – Experimentalstudien Rückschlüsse auf kausale Zusammenhänge, nicht aber Beobachtungsstudien.

Auf Basis von Beobachtungsstudien bekommt man nur Hinweise auf assoziative Zusammenhänge.

Wegen ihrer stärkeren Aussage sind Experimentalstudien zu bevorzugen, aber – aus prinzipiellen, zeitlichen, ethischen, oder anderen Gründen – nicht immer durchführbar.

Stichprobe

Eine Untersuchung aller Elemente einer Grundgesamtheit (d. h. eine Gesamterhebung) ist aus verschiedenen Gründen (zeitlicher, finanzieller, prinzipieller, oder sonstiger Art) nicht immer durchführbar.

In solchen Fällen beschränkt man sich auf eine Stichprobe, d. h. auf eine repräsentative Teilauswahl aus der Grundgesamtheit.

Um ein getreues Abbild der Grundgesamtheit zu bekommen, sollte die Auswahl rein zufällig erfolgen.

Stichprobe (Forts.)

Besteht die Grundgesamtheit aus N (unterscheidbaren) Elementen und möchte man eine Stichprobe des Umfangs n ziehen, so gibt es dafür

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

verschiedene Möglichkeiten.

Werden die n Elemente so ausgewählt, dass jede der $\binom{N}{n}$ möglichen Stichproben die gleiche Auswahlwahrscheinlichkeit hat, spricht man von einer (einfachen) Zufallsstichprobe.

Bei Zufallsstichproben hat jedes Element der Grundgesamtheit die gleiche Chance, in die Stichprobe zu gelangen.

Ziehungen ohne/mit Zurücklegen

Bei der oben beschriebenen Form der Stichprobenziehung wird jedes Element der Grundgesamtheit höchstens einmal ausgewählt.

Das nennt man Ziehen ohne Zurücklegen.

Man kann aber auch eine bereits erhobene Einheit ein weiteres Mal berücksichtigen.

Das nennt man Ziehen mit Zurücklegen.

Besteht die Grundgesamtheit aus N (unterscheidbaren) Elementen, gibt es dafür N^n verschiedene Möglichkeiten.

Da eine Zufallsauswahl der beschriebenen Art nicht immer durchführbar (oder adäquat) ist, verwendet man in der Praxis noch eine Reihe von anderen Formen der Stichprobenentnahme.

Merkmal

Im nächsten Schritt werden an den Elementen der Stichprobe die interessierenden Größen erhoben, Merkmale oder Variablen genannt.

Die (möglichen) Werte, die ein Merkmal annehmen kann, nennt man die Merkmalsausprägungen.

Der Menge M der möglichen Merkmalsausprägungen nennt man den Merkmalraum.

Merkmal (Forts.)

Merkmalsausprägungen können von ganz unterschiedlicher Art sein:

- Das (physische) Geschlecht hat zwei Ausprägungen (die allein der Unterscheidung dienen).
- Die Mitarbeiterzahl eines Unternehmens ist eine Zählvariable mit (potenziell) unbeschränkt vielen Ausprägungen.
- Die Funktionsdauer einer Batterie (in Betriebsstunden) ist ein auf ein Intervall beschränktes metrisches Merkmal.
- etc.

Merkmale sind Abbildungen

Aus mathematischer Sicht ist ein Merkmal eine Abbildung (Funktion) $X : G \longrightarrow M$ von der Grundgesamtheit G in den Merkmalraum M.

Jeder statistischen Einheit aus G wird ein Element aus M zuordnet:

$$g \in G \longmapsto X(g) \in M$$

Das zugeordnete Element kann auch ein Vektor sein.

Misst man beispielsweise an Personen die Körpergröße (h) und das Körpergewicht (w), so gilt:

$$X(\text{Person}) = (h, w) \in (\mathbb{R}^+)^2$$

Messniveau

Merkmalsausprägungen werden meist durch Zahlen repräsentiert.

Daraus folgt nicht, dass alle Rechenoperationen (oder Vergleiche) mit diesen Zahlen durchgeführt werden dürfen bzw. sinnvoll sind.

Der Umfang der zulässigen Operationen (der zur Verfügung stehenden Methoden der statistischen Analyse) ist abhängig vom Messniveau des Merkmals.

Allgemein unterscheidet man zwischen qualitativen und quantitativen (oder metrischen) Merkmalen:

Geschlecht (qualitativ), Körpergröße (quantitativ/metrisch), ...

oder zwischen diskreten und stetigen Merkmalen:

Mitarbeiterzahl (diskret), Blutdruck (stetig), ...

Nominalskalen

Hierbei handelt es sich um eine reine Klassifikation.

Sonst bestehen keine weiteren Relationen zwischen den Elementen der Grundgesamtheit.

Zahlenmäßige Ausprägungen eines solchen Merkmals sind nur eine zweckmäßige Codierung.

Bsp: Geschlecht, Familienstand, Religionsbekenntnis, ...

Ordinalskalen

Kennzeichnend für Rangmerkmale ist eine lineare Ordnungsbeziehung.

Sonst sind keine weiteren Beziehungen vorhanden.

Zahlenmäßige Ausprägungen eines solchen Merkmals spiegeln diese Ordnung wider.

Bsp: Prüfungsnoten, Güteklassen von Obst, Windstärke (z. B. Beaufort–Skala von 0 bis 12), ...

Häufig wird ein an sich metrisch skaliertes Merkmal auf ein Rangmerkmal reduziert (ein Beispiel ist die vorhin erwähnte Beaufort–Skala).

Intervallskalen

Die Ausprägungen sind reelle Zahlen (oder Vektoren).

Der Nullpunkt – sofern überhaupt vorhanden – hat keine absolut festgelegte Bedeutung, dient nur der Definition der Skala.

Differenzen lassen sich sinnvoll interpretieren.

Aussagen wie "doppelt so warm", "halb so spät", ... haben keine sinnvolle Interpretation.

Bsp: Zeiteinteilung (0 bis 24 Uhr), Temperatur in Grad Celsius oder Grad Fahrenheit ($F = \frac{9}{5} C + 32$), ...

Verhältnisskalen

Hierbei handelt es sich um Intervallskalen mit ausgeprägtem und interpretierbarem Nullpunkt.

Aussagen wie "doppelt so hoch", "halb so schnell", ... sind sinnvoll.

Bsp: Körpergröße, Geschwindigkeit, Temperatur in Kelvin, ...

Als Folge der durch den Nullpunkt gegebenen – linksseitigen – Beschränkung der Messwerte, weisen verhältnisskalierte Merkmale häufig eine schiefe Verteilung auf.

Datenmatrix

Ausgangspunkt für eine tabellarische oder grafische Aufbereitung von Datensätzen sind zunächst die Rohdaten (oder Urdaten, Primärdaten).

Die erhobenen Ausprägungen werden in Form einer Datenmatrix (oder eines Datenframe) dargestellt.

Die Spalten einer Datenmatrix entsprechen den Variablen (Merkmalen).

Die Zeilen entsprechen den Untersuchungseinheiten.

Beispiel

Ausschnitt aus einem umfangreichen Datensatz (`body.txt`), bestehend aus einer Reihe von anthropometrischen Messwerten.

Wir betrachten nur 6 von insgesamt 25 Variablen:

- Biacromial diameter (cm)
- Waist girth (cm)
- Age (years)
- Weight (kg)
- Height (cm)
- Gender (1/0 = male/female)

Beispiel (Forts.)

```
dat <- read.table("body.txt", header=TRUE)
dat <- dat[,c(1,12,22,23,24,25)]
#  V1 = Biacromial diameter (Schulter) (cm)
# V12 = Waist girth (Taille) (cm)
# V22 = Age (years)
# V23 = Weight (kg)
# V24 = Height (cm)
# V25 = Gender (0 - female, 1 - male)
names(dat) <- c("Biacromial","Waist","Age","Weight",
                "Height","Gender")
head(dat, n=10)
tail(dat, n=10)
```

Beispiel (Forts.)

	Biacromial	Waist	Age	Weight	Height	Gender
1	42.9	71.5	21	65.6	174.0	1
2	43.7	79.0	23	71.8	175.3	1
3	40.1	83.2	28	80.7	193.5	1
4	44.3	77.8	23	72.6	186.5	1
5	42.5	80.0	22	78.8	187.2	1
	.	.	.	.		

$n\,(=507) \times p\,(=6)$ – Datenmatrix:

In der i-ten Zeile der Datenmatrix stehen die an der i-ten statistischen Einheit beobachteten Ausprägungen.

In der j-ten Spalte stehen die beobachteten Werte des j-ten Merkmals.

Univariate / Multivariate Daten

Hat die Datenmatrix nur eine Spalte – d. h. ist $p = 1$ – spricht man von univariaten Daten.

Für $p > 1$ spricht man von multivariaten Daten.

Die n beobachteten Ausprägungen eines univariaten Merkmals werden häufig in einem n–dimensionalen Datenvektor $\mathbf{x}$ zusammengefasst:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)' \in \mathbb{R}^n \quad \text{(Spaltenvektor)}$$

Diskrete univariate Merkmale

Die Darstellung von diskreten – d. h. in erster Linie von nominalen und ordinalen – Daten erfolgt durch Bestimmung von Häufigkeiten und einer geeigneten Visualisierung.

Bezüglich der grafischen Darstellung trifft man – speziell in den Medien – auf eine Fülle von Umsetzungen, die manchmal mit einer gewissen Skepsis zu betrachten sind.

Häufigkeiten

Ein diskretes Merkmal, das die Werte $x_1 < x_2 < \cdots$ annehmen kann, werde insgesamt n Mal beobachtet.

Die absolute Häufigkeit mit der die Ausprägung x_i beobachtet wird, werde mit n_i bezeichnet.

Es gilt: $\sum_i n_i = n$

Die relativen Häufigkeiten seien mit $f_i = n_i/n$ bezeichnet.

Es gilt: $\sum_i f_i = 1$

Beispiel

Hauptsächliche Verwendung von Webbrowsern:

	May14	May15	May16	May17
Chrome	59.2	64.9	71.4	75.8
Firefox	24.9	21.5	16.9	13.6
IE/Edge	8.9	7.1	5.7	4.6
Safari	3.8	3.8	3.6	3.4
Opera	1.8	1.6	1.2	1.1
Un.	1.4	1.1	1.2	1.5

(Quelle: `www.w3schools.com`)

Anordnung nach Häufigkeit der Verwendung May14.

Kreisdiagramm

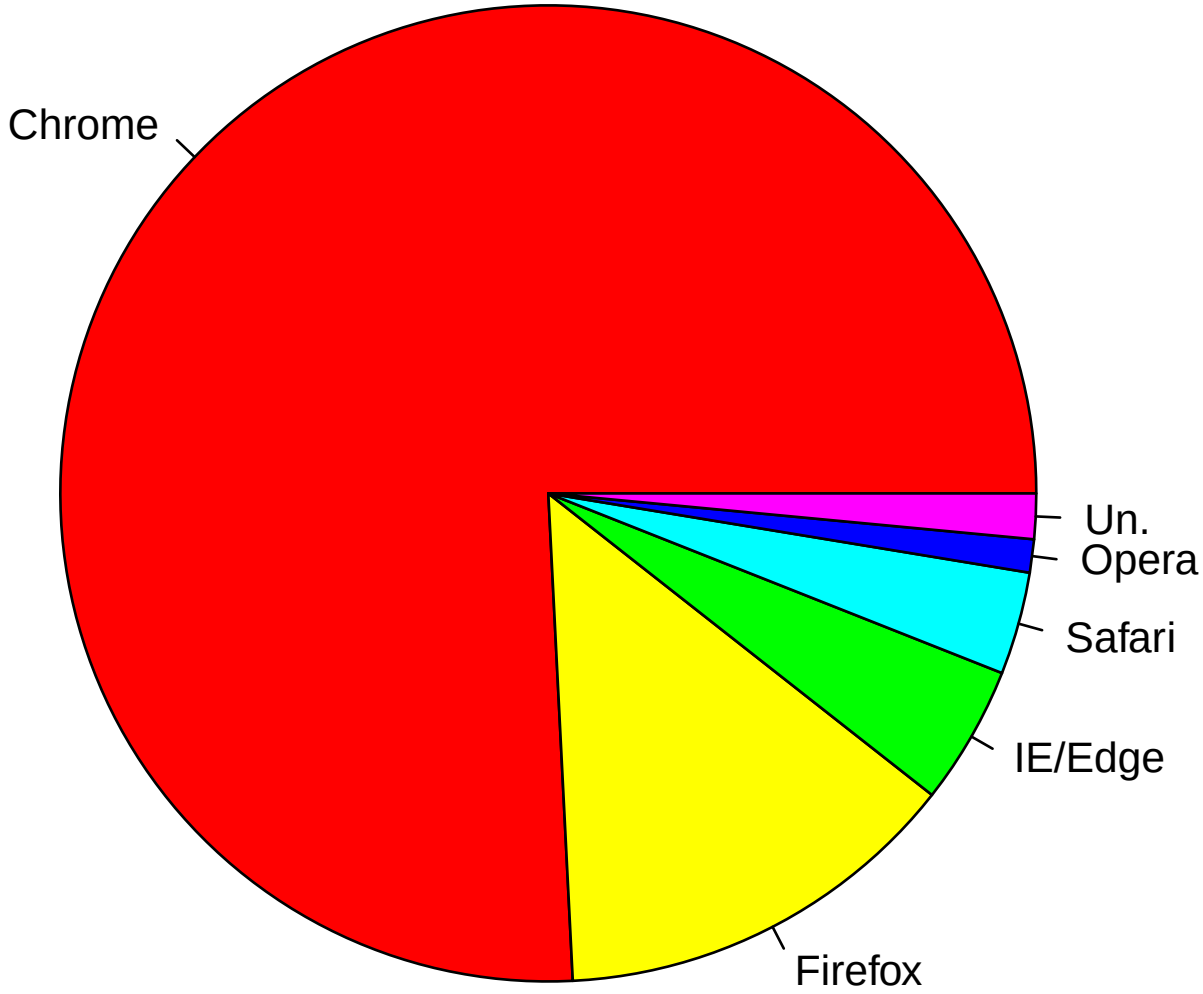
Bei einem Kreisdiagramm (auch Kuchen– oder Tortendiagramm) wird bei einem Kreis der Gesamtwinkel von 360° (bzw. 2π [rad]) entsprechend den absoluten oder relativen Häufigkeiten aufgeteilt.

Zur relativen Häufigkeit f_i gehört also der Winkel:

$$\varphi_i = f_i \cdot 360^\circ \quad \text{oder} \quad 2\pi f_i \text{ [rad]}$$

Die einzelnen Kreissegmente werden mit unterschiedlichen Farben (oä.) dargestellt.

Beispiel: Webbrowser May 17

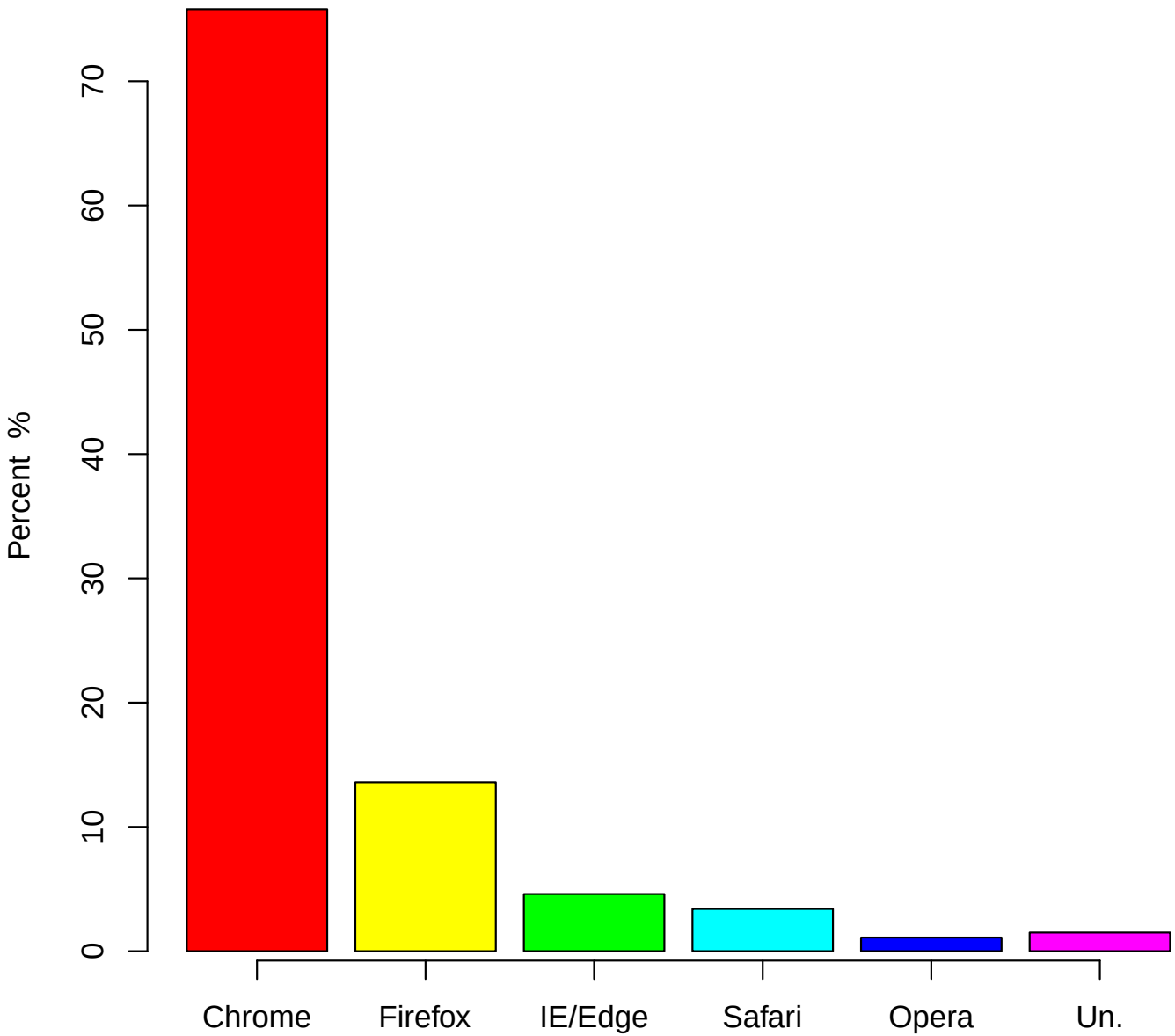


Balkendiagramm

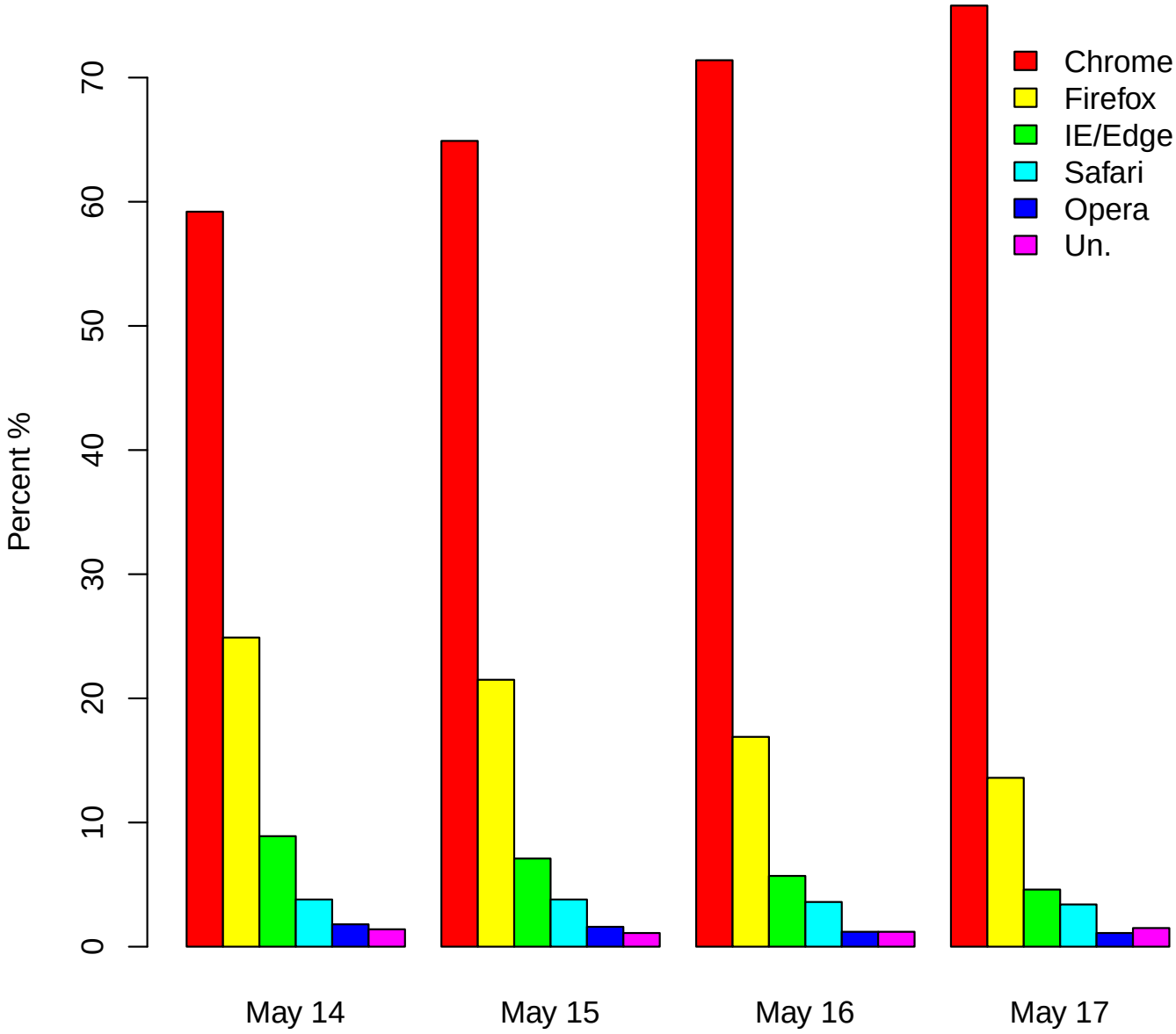
Das Balkendiagramm (auch Stabdiagramm, Barplot) ist eine grafische Darstellung der absoluten oder relativen Häufigkeiten mit senkrechten (manchmal auch waagrechten) Balken der Länge n_i (oder f_i) über den Merkmalswerten x_i.

Bei der vergleichenden Darstellung mehrerer Häufigkeitsverteilungen können die Balken auch übereinander gestapelt gezeichnet werden.

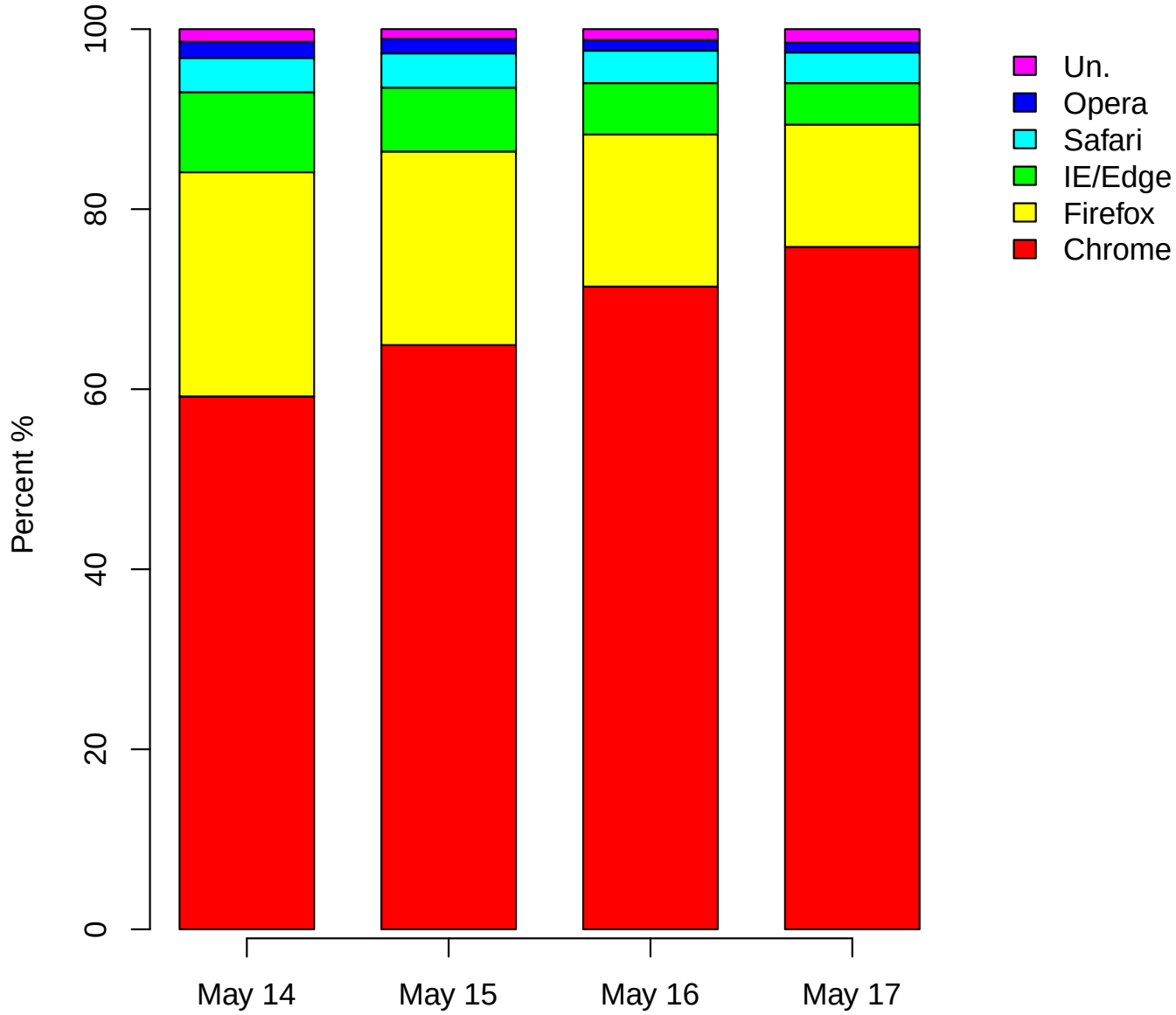
Beispiel: Webbrowser May 17



Beispiel: Webbrowser May 14 – 17



Beispiel: Webbrowser May 14 – 17



Stetige univariate Merkmale

Nun betrachten wir verschiedene grafische Darstellungsmöglichkeiten für Beobachtungen von stetigen Merkmalen.

Da das Messniveau nun höher ist (Intervall–, Verhältnisskalen) hat man mehr Möglichkeiten als bei qualitativen Merkmalen.

Ordnungsstatistiken

Ein natürlicher erster Schritt in der Aufbereitung von metrischen (oder ordinalen) Merkmalen ist ihre Sortierung nach der Größe.

Urliste: $x_1, x_2, \ldots, x_n$ (ungeordnet)

Rangfolge: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ (geordnet)

(i–te) Ordnungsstatistik: $x_{(i)}$ (i ist die Rangzahl)

Bindungen

Vielfach – z. B. als Folge einer nur beschränkten Messgenauigkeit – sind mehrere Beobachtungen identisch:

$$x_{(i-1)} < x_{(i)} = x_{(i+1)} = \cdots = x_{(i+c)} < x_{(i+c+1)}$$

In diesem Fall spricht man von einer Bindung vom Ausmaß $c + 1$.

Allen Werten von $x_{(i)}$ bis $x_{(i+c)}$ wird die mittlere Rangzahl $i + c/2$ zugeordnet.

Beispiel

10 Beobachtungen, 1 Bindung vom Ausmaß 2

Urliste									
0.15	−0.84	−0.83	0.15	−0.50	−1.62	−0.52	0.49	0.08	−0.66

Rangzahlen									
8.5	2	3	8.5	6	1	5	10	7	4

Rangtransformation

Wird jede Beobachtung durch ihre Rangzahl – unter Verwendung der Bindungsregel – ersetzt, spricht man von der Rangtransformation.

Dadurch verzichtet man auf einen Großteil der in den ursprünglichen Daten enthaltenen metrischen Information und verwendet für weitere Berechnungen nur mehr die relativen Positionen der Beobachtungen innerhalb des Datensatzes.

Nichtparametrische Statistik

Ordnungsstatistiken – und die Rangtransformation – spielen generell eine große Rolle in der Statistik, insbesondere aber in der sogenannten nichtparametrischen Statistik.

In der nichtparametrischen Statistik versucht man mit nur ganz wenigen Voraussetzungen hinsichtlich des zugrunde liegenden statistischen Modells auszukommen.

Vorteile: breitere Anwendbarkeit, größere Allgemeinheit

Nachteile: schwächere Aussagen, geringere Methodenvielfalt

Empirische Verteilungsfunktion

Eine Funktion von grundlegender Bedeutung in der Statistik ist die empirische Verteilungsfunktion, definiert für $x \in \mathbb{R}$ durch:

$$\widehat{F}_n(x) = \begin{cases} 0 & \text{für} \quad x < x_{(1)} \\ \dfrac{i}{n} & \text{für} \quad x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, 2, \ldots, n-1 \\ 1 & \text{für} \quad x_{(n)} \leq x \end{cases}$$

$\widehat{F}_n(x)$ ist eine Treppenfunktion mit Sprüngen an den Stellen $x_{(i)}$ der Höhe $1/n$ – oder der Höhe c/n, falls es bei $x_{(i)}$ eine Bindung vom Ausmaß c gibt.

Empirische Verteilungsfunktion (Forts.)

Äquivalente Definition:

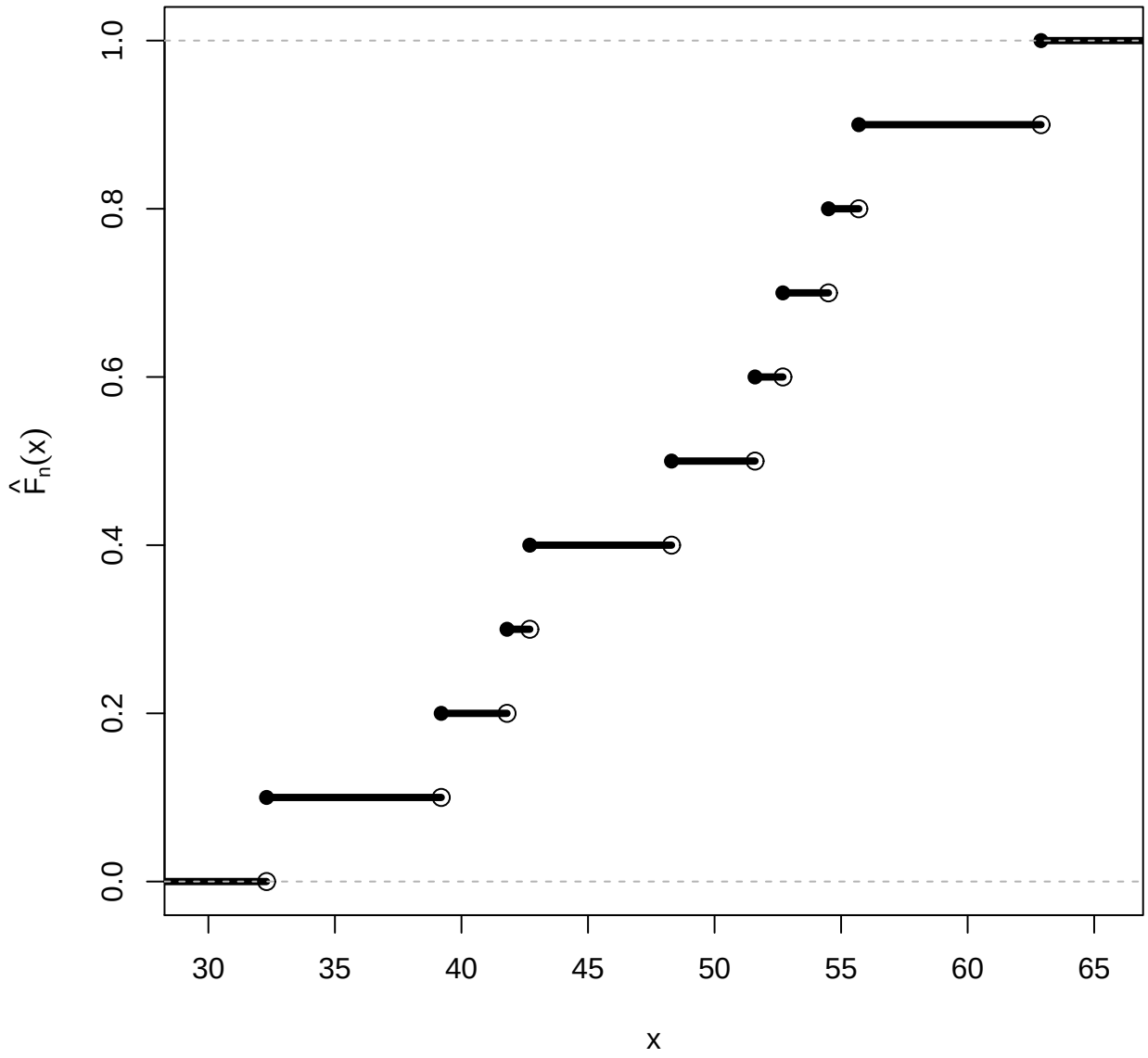
$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(x_i), \quad x \in \mathbb{R}$$

Dabei ist $I_A(x)$ die Indikatorfunktion der Menge A ($A \subseteq \mathbb{R}$):

$$I_A(x) = \begin{cases} 1 & \text{für } x \in A \\ 0 & \text{sonst} \end{cases}$$

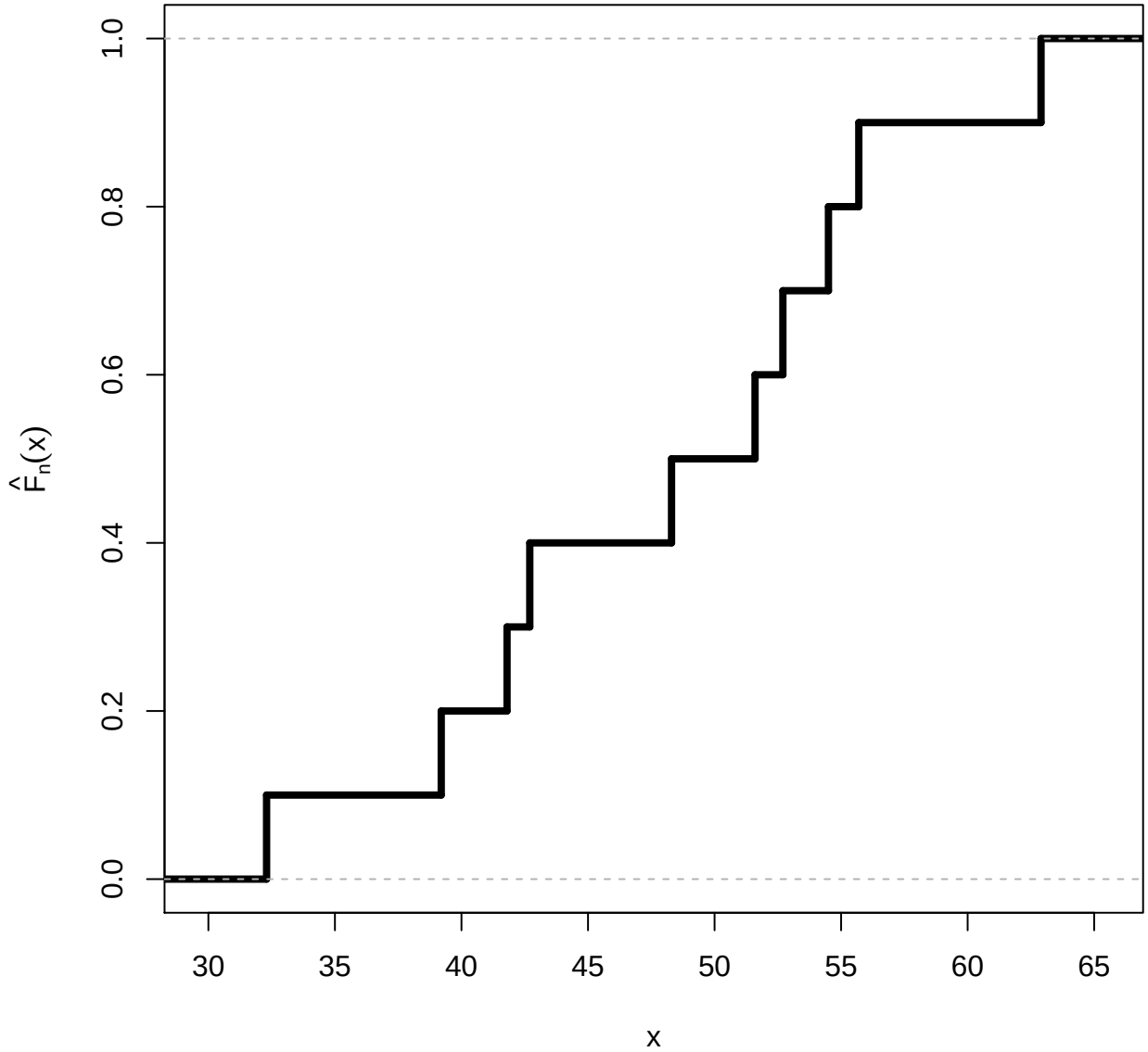
Beispiel

Bei Sprüngen gilt stets der obere Punkt.



Beispiel (Forts.)

Aus optischen Gründen zeichnet man die Stufen meist aus.



Klassierung

Bei größeren Stichprobenumfängen – ab etwa 30 – ist es sinnvoll, eine Klassenbildung vorzunehmen.

Hinsichtlich der Anzahl und Breite der Klassen – auch Bins genannt – gibt es verschiedene Regeln.

Es ist darauf zu achten, dass der gesamte Wertebereich überdeckt wird und jede Beobachtung eindeutig einer Klasse zugeordnet werden kann.

Meist nimmt man halboffene Klassen der Form $(a, b]$.

Falls die Verteilung nicht sehr schief ist, sind Klassierungen mit äquidistanten Klassenbreiten w zu bevorzugen.

Klassierungsregeln

(1) Bestimme den kleinsten $x_{(1)}$ und größten Wert $x_{(n)}$ der Stichprobe, sowie die Spannweite $R = x_{(n)} - x_{(1)}$.

(2) In der Praxis sind Beobachtungen gerundete (oder abgeschnittene) Zahlen. Ist der kleinste Wert z.B. 69.6, so steht er für einen Mess–wert zwischen 69.55 und 69.65. Als unteren Rand der ersten Klasse kann man daher 69.55 nehmen.

(3) Eine einfache Regel besagt:

$$\text{Klassenbreite:} \quad w = \begin{cases} \dfrac{R}{\sqrt{n}} & \text{falls} \quad 30 < n \leq 400 \\ \dfrac{R}{20} & \text{falls} \quad n > 400 \end{cases}$$

Sturges' Rule (R–Default)

Daten: $x_1, x_2, \ldots, x_n$

Nimm a Klassen gleicher Breite, wobei

$$2^{a-1} < n \leq 2^a \quad \Longrightarrow \quad a \approx \log_2(n) \quad \text{Klassen}$$

Weitere gebräuchliche Regeln: Scott, Freedman–Diaconis (FD), ...

Beispiel

Variable `Biacromial` (Datensatz: `body.txt`)

```
his <- hist(datm[,1], plot=FALSE)

# Absolute Häufigkeiten
(tab <- table(cut(datm[,1], breaks=his$breaks)))

(34,36] (36,38] (38,40] (40,42] (42,44] (44,46] (46,48]
      3      15      44      98      68      17       2

# Relative Häufigkeiten (%)
round(prop.table(tab)*100, 2)

(34,36] (36,38] (38,40] (40,42] (42,44] (44,46] (46,48]
   1.21    6.07   17.81   39.68   27.53    6.88    0.81
```

Histogramm

Eine grafische Darstellung einer Häufigkeitsverteilung, basierend auf einer vorherigen Klassierung der Daten.

Dabei sollte man sich an das folgende Prinzip halten:

Prinzip der Flächentreue: Zeichne über den Klassen Rechtecke der Höhe:

$$h_i = \frac{f_i}{w_i}, \quad i = 1, 2, \ldots, k$$

k ... Anzahl der Klassen

f_i ... Relative Häufigkeit der i–ten Klasse

w_i ... Breite der i–ten Klasse

Histogramm (Forts.)

Das so gezeichnete Histogramm nennt man ein flächentreues Histogramm (oder ein Dichtehistogramm), da unabhängig von der Klasseneinteilung die Summe der Rechtecksflächen genau Eins beträgt:

$$\text{Fläche des Histogramms } = \sum_{i=1}^{k} h_i w_i = \sum_{i=1}^{k} f_i = 1$$

Wann muss man ein flächentreues Histogramm zeichnen?

- Wenn man eine nicht äquidistante Klassierung verwendet.
- Wenn man mehrere Häufigkeitsverteilungen – mit unterschiedlichen Klassierungen – miteinander vergleichen möchte.

Histogramm (Forts.)

Klasseneinteilungen sind nicht eindeutig bestimmt, daher kann auch das Erscheinungsbild eines Histogramms, abhängig von der verwendeten Klasseneinteilung, u. U. beträchtlich variieren.

M. a. W., Histogramme sind nicht robust bezüglich der Klasseneinteilung.

In der Praxis variiert man die Anzahl der Klassen, die Klassenbreite und den Anfangspunkt der Klasseneinteilung und nimmt jenes Histogramm, das die wenigsten unechten Täler und Gipfel aufweist.

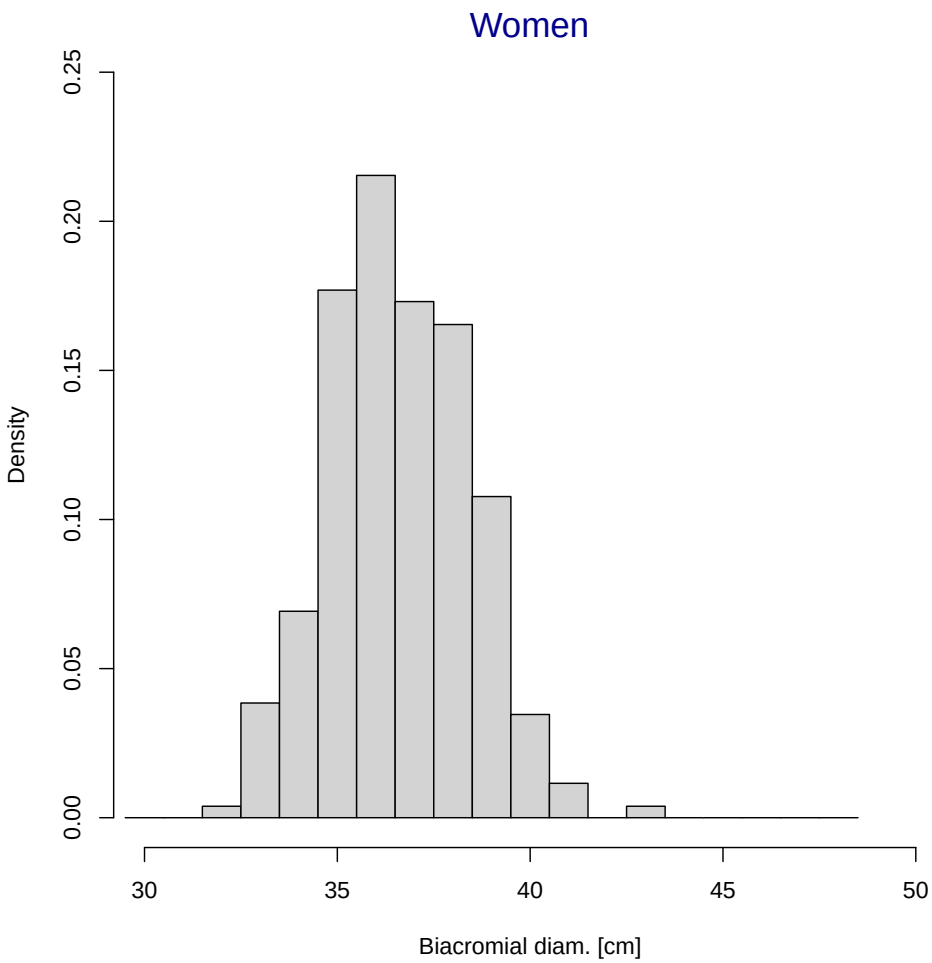
Das ist natürlich eine subjektive Entscheidung, bei der man aber auch sonstige Informationen über den Datensatz berücksichtigen sollte.

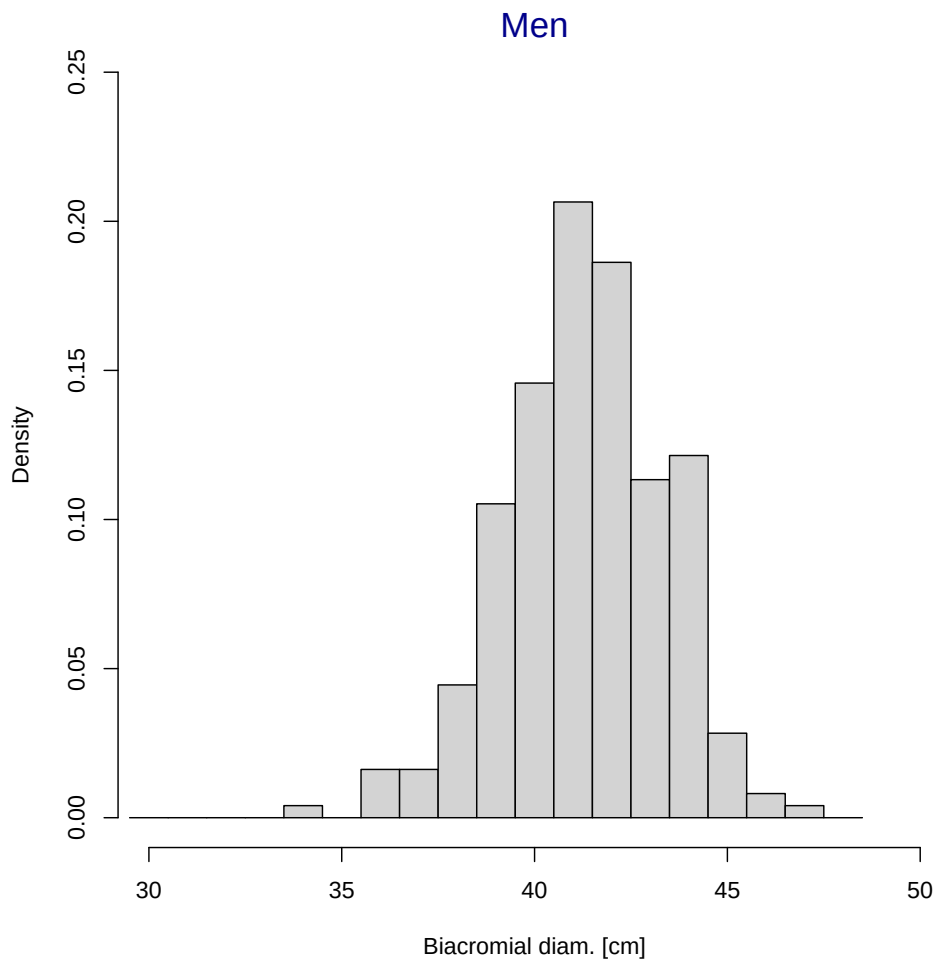
Beispiel

Variable: `Biacromial` (Datensatz: `body.txt`)

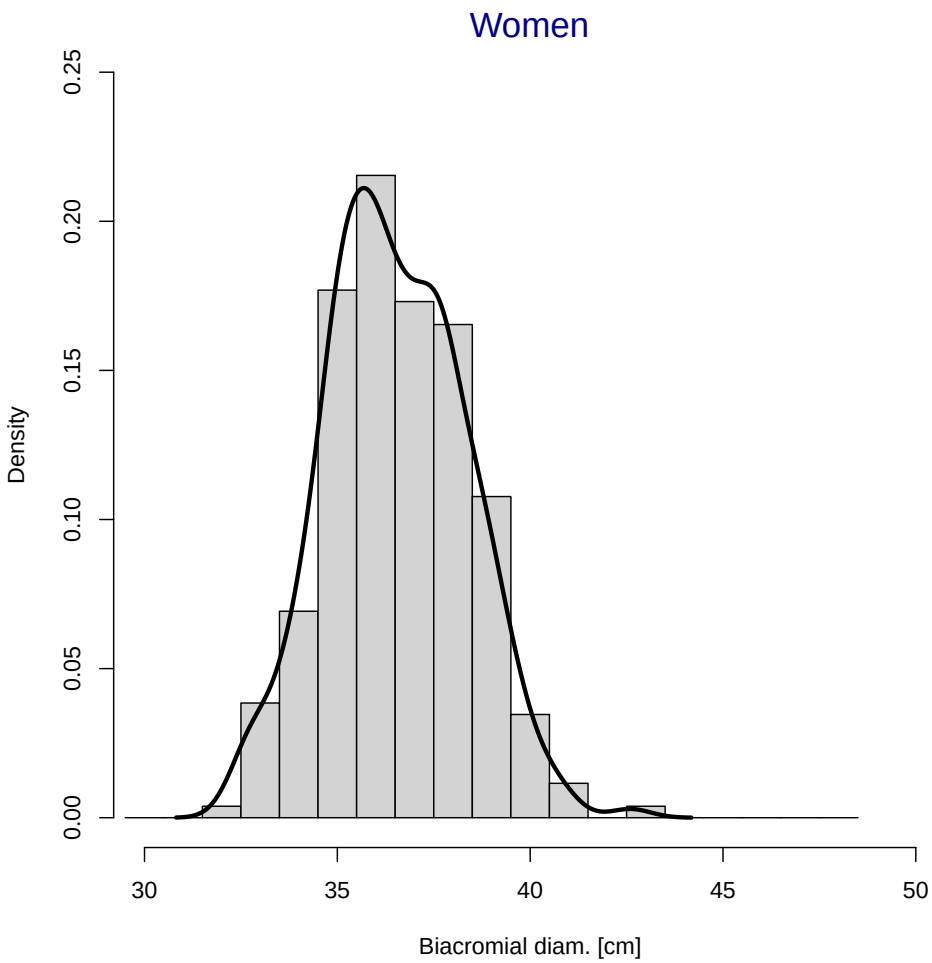
```
brk <- seq(29.5, 48.5, by=1)

old.par <- par(mfrow=c(1,2))

hist(datf[,1], breaks=brk, prob=T, col="lightgrey", right=F,
  xlab="Biacromial diam. [cm]", main="Women", xlim=c(30,50))

hist(datm[,1], breaks=brk, prob=T, col="lightgrey", right=F,
  xlab="Biacromial diam. [cm]", main="Men", xlim=c(30,50))

par(old.par)
```

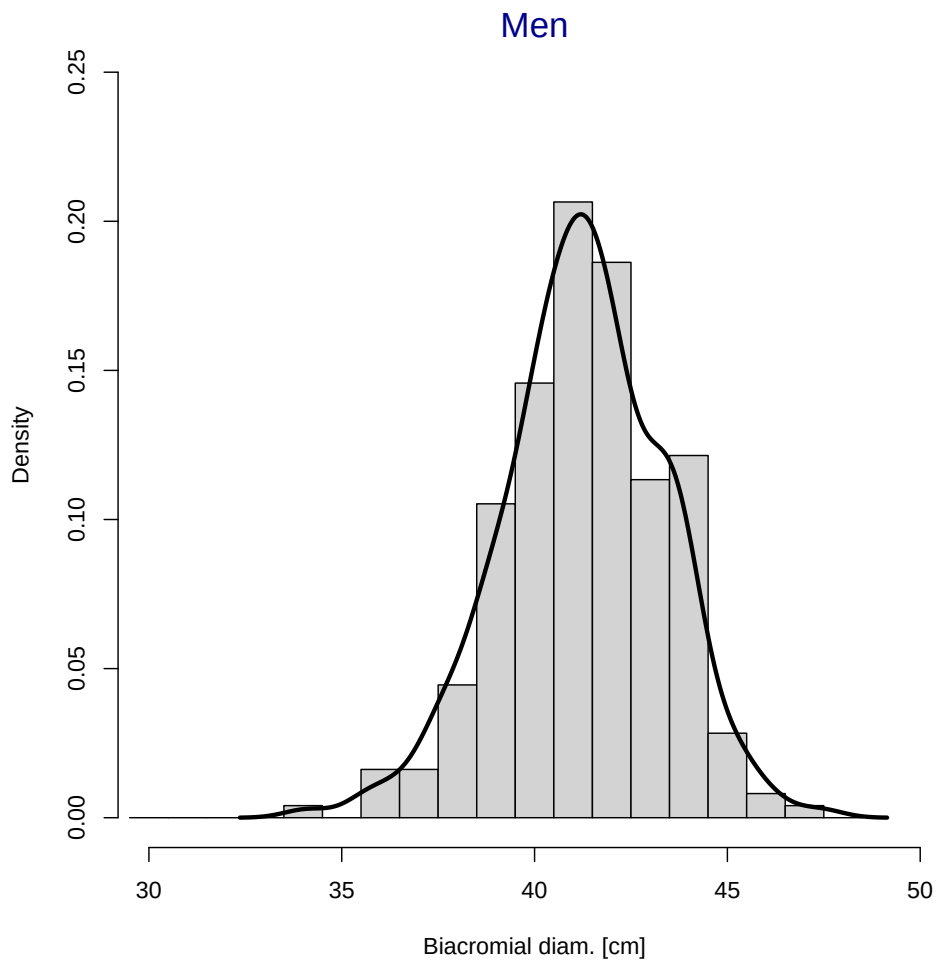
Beispiel (Forts.)





Beispiel (Forts.)





Quantile

Empirische (d. h. auf Daten basierende) Quantile sind nicht einheitlich definiert.

Allgemein gesagt ist ein p–Quantil – mit $0 \leq p \leq 1$ – ein Wert x_p, der den Datensatz (etwa) im Verhältnis $p : (1 - p)$ teilt.

$$\text{Daten:} \quad x_1, x_2, \ldots, x_n \quad \Longrightarrow \quad \frac{\#\{x_i \leq x_p\}}{n} \approx p$$

Die Definitionen können danach eingeteilt werden, ob für Quantile nur Datenwerte oder auch Werte dazwischen in Frage kommen.

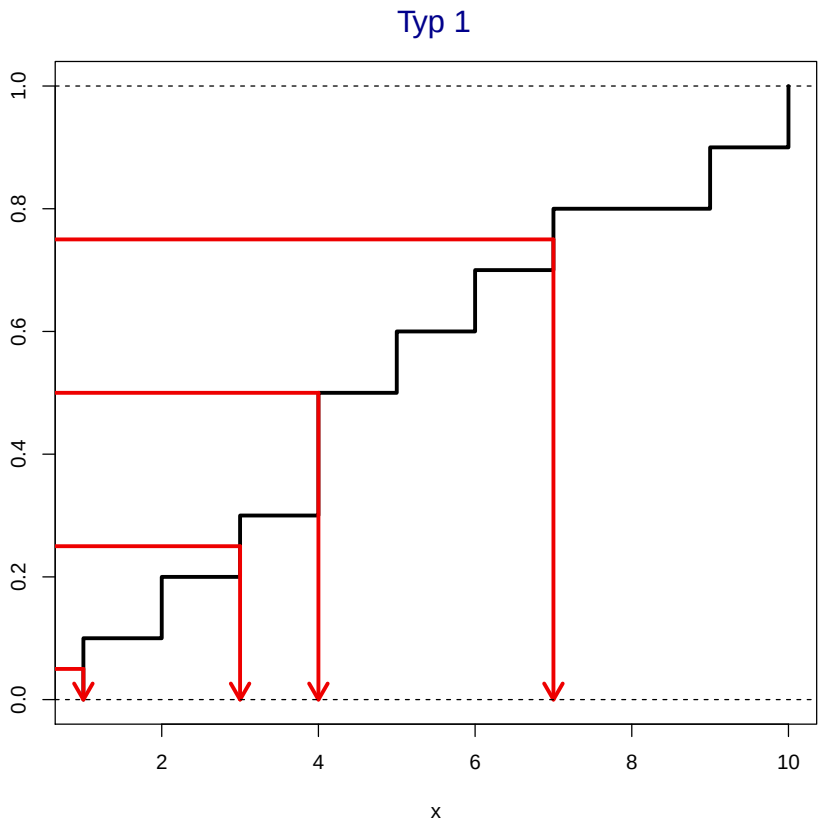
In R werden neun verschiedene Typen unterschieden:

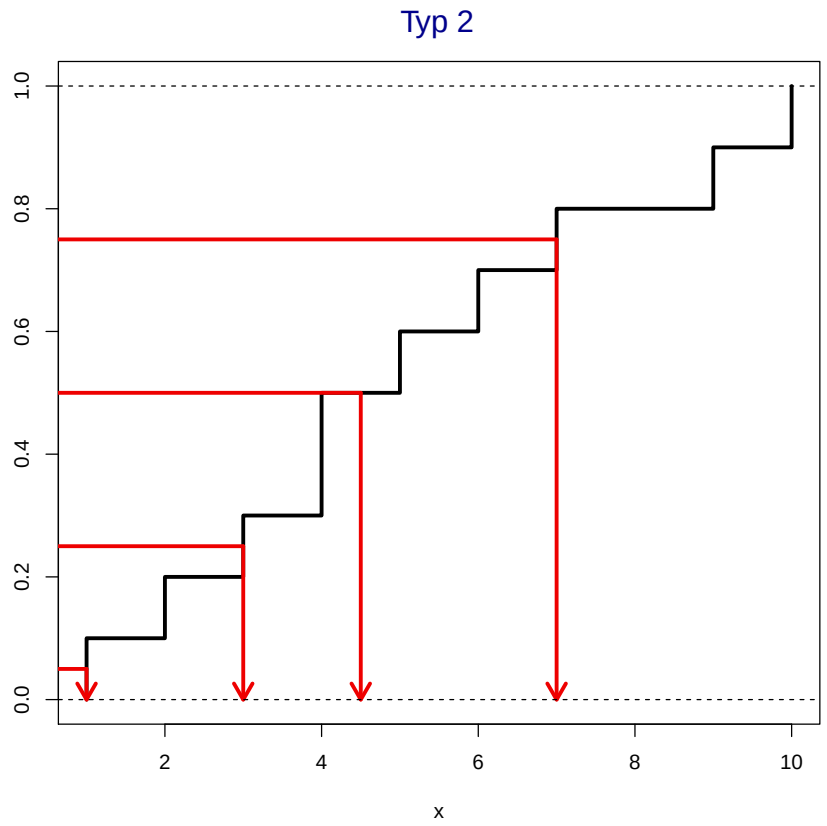
```
quantile(x, c(0.1, 0.5, 0.75), type=4)  # default: type=7
```

Quantile: Typ 1 und 2

Daten (geordnet): 1, 2, 3, 4, 4, 5, 6, 7, 9, 10

Quantile: 5%, 25%, 50%, 75%





Verallgemeinerte Inverse

Typ 1 bezieht sich auf die empirische Verteilungsfunktion $\widehat{F}_n$ und ist definiert durch:

$$x_p = \min\left\{x \in \mathbb{R} : \widehat{F}_n(x) \geq p\right\}$$

Das so definierte x_p entspricht stets einem Datenwert.

Diese Definition entspricht der verallgemeinerten Inversen von $\widehat{F}_n$.

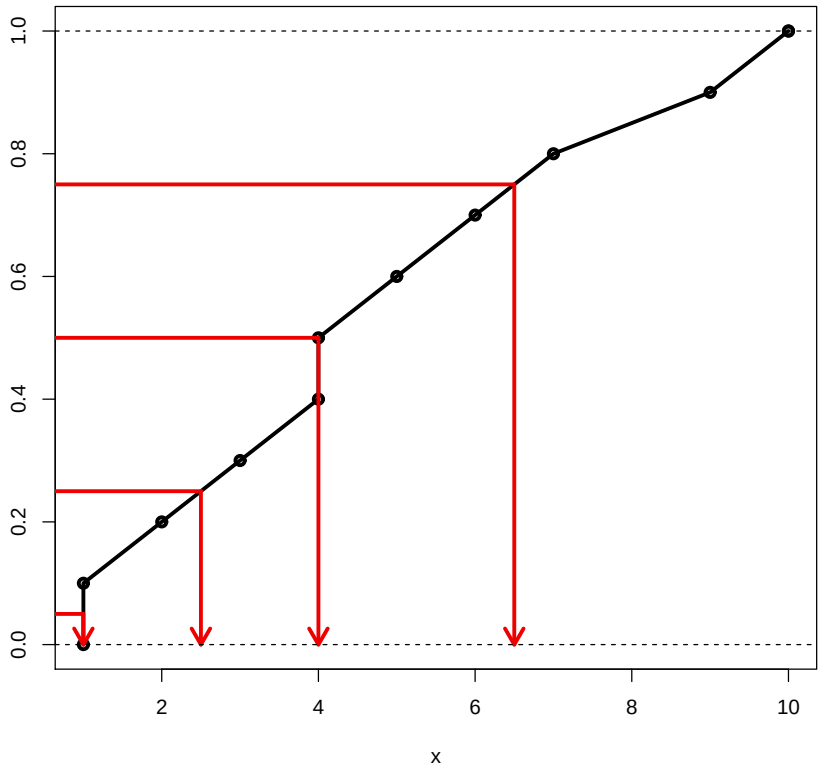
$$\text{Schreibweise:} \quad x_p = \widehat{F}_n^{-1}(p)$$

Quantile: Typ 4 und 7

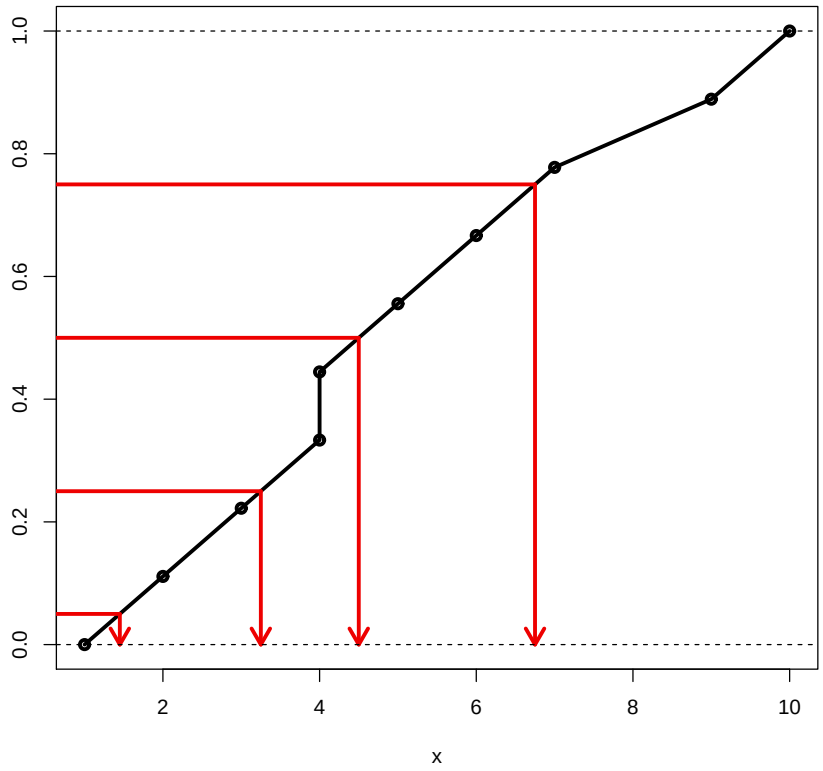
Daten (geordnet): 1, 2, 3, 4, 4, 5, 6, 7, 9, 10

Quantile: 5%, 25%, 50%, 75%

Typ 4



Typ 7



Spezielle Quantile: Median

Der Median ist das 50%–Quantil, teilt also den Datensatz in zwei gleich große Hälften.

Bezeichnung: $\widetilde{x}$, $x_{0.5}$, ...

Der Median ist einheitlich wie folgt definiert:

geordnete Daten: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

$$\widetilde{x} = \begin{cases} x_{(k+1)} & \text{falls} \quad n = 2k+1 \quad (\text{d. h. } n \text{ ist ungerade}) \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{falls} \quad n = 2k \quad (\text{d. h. } n \text{ ist gerade}) \end{cases}$$

Spezielle Quantile: Quartile

Die Quartile (= Viertel) teilen den Datensatz in (etwa) vier gleich große Stücke:

$$Q_1 = x_{1/4} \quad (= 1.\ \text{Quartil})$$

$$Q_2 = x_{1/2} \quad (= 2.\ \text{Quartil} = \text{Median})$$

$$Q_3 = x_{3/4} \quad (= 3.\ \text{Quartil})$$

Zwischen dem 1. und 3. Quartil liegen die mittleren 50% der Daten.

Weitere spezielle Quantile: Oktile (= Achtel), Dezile (= Zehntel), ...

Hinges

hinge engl. = Türangel, Drehachse, Gelenk, ...

unterer (lower) Hinge = Median der unteren Hälfte der (geordneten) Daten

oberer (upper) Hinge = Median der oberen Hälfte der (geordneten) Daten

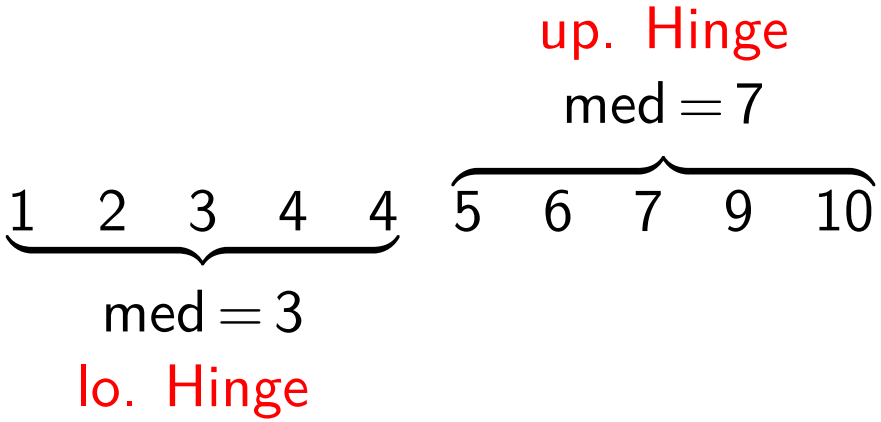
Bei ungerader Anzahl von Daten zählt der Median zu beiden Hälften.

Die Hinges entsprechen dem 1./3. Quartil, sind aber einfacher und schneller zu bestimmen.

Beispiel 1 [52]

Daten (geordnet): 1, 2, 3, 4, 4, 5, 6, 7, 9, 10

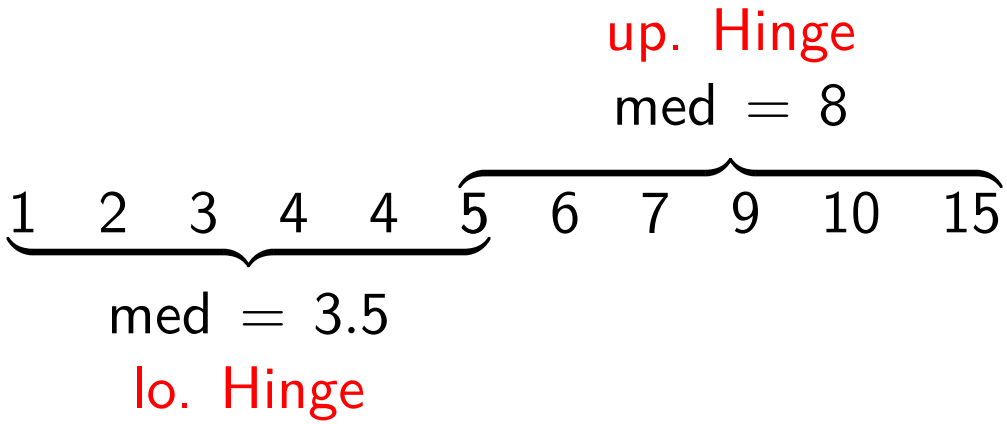
Median: $n = 10$ gerade $\Longrightarrow$ $\tilde{x} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{4+5}{2} = 4.5$



Beispiel 2

Daten (geordnet): 1, 2, 3, 4, 4, 5, 6, 7, 9, 10, 15

Median: $n = 11$ ungerade $\Longrightarrow$ $\widetilde{x} = x_{(6)} = 5$



QQ–Plot

Der Q(uantilen)Q(uantilen)–Plot ist eine Art Streudiagramm zum grafischen Vergleich zweier Datensätze.

Ist die Größe der beiden Datensätze identisch, so zeichnet man die beiden geordneten Stichproben gegeneinander:

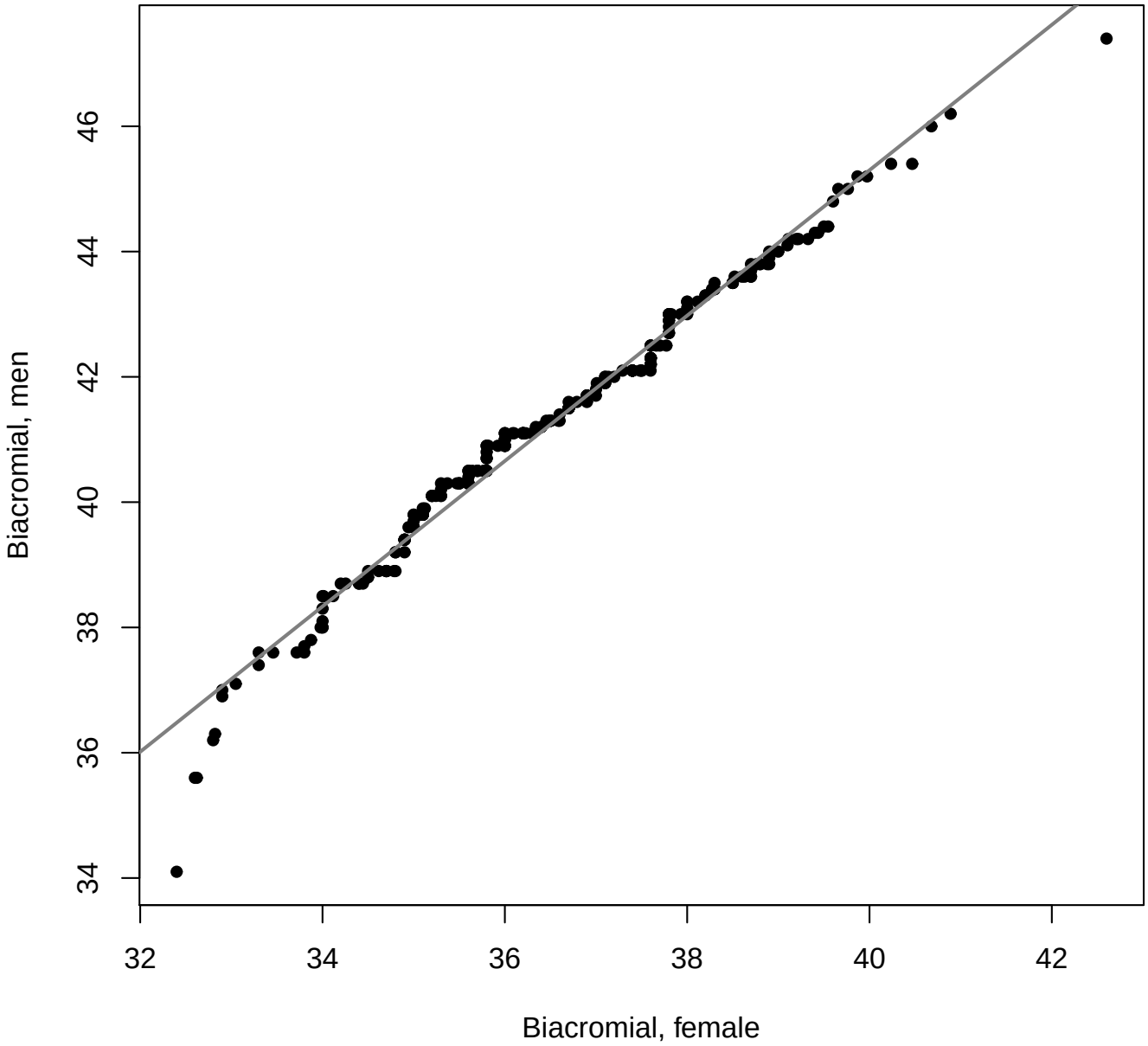
$$(x_{(i)}, y_{(i)}), \quad i = 1, 2, \ldots, n$$

Sind die Stichprobengrößen unterschiedlich, muss man die Datensätze einander angleichen.

Dabei geht man so vor, dass der größere Datensatz reduziert wird.

Man behält Minimum und Maximum und wählt gleichmäßig aufgeteilte (empirische) Quantile dazwischen.

Beispiel


Biacromial, men
34
36
38
40
42
44
46
32
34
36
38
40
42
Biacromial, female

Boxplot

Der Boxplot ist eine grafische Darstellung eines Datensatzes auf Basis der Quartile oder der Hinges.

Auf diese Weise können auch mehrere Datensätze schnell miteinander verglichen werden.

(1) Zeichne zunächst die Box, d. h. zeichne ein Rechteck vom 1. zum 3. Quartil, oder vom unteren zum oberen Hinge ($\longrightarrow$ R).

Die Box umfasst also die mittleren 50% der Daten.

(2) Der Median wird durch eine dicke Linie hervorgehoben.

Boxplot (Forts.)

(3) Bestimme die lower und upper Fences (= Einzäunungen):

$$\text{LF} = Q_1 - \underbrace{1.5\,(Q_3 - Q_1)}_{=:\,h}, \qquad \text{UF} = Q_3 + h$$

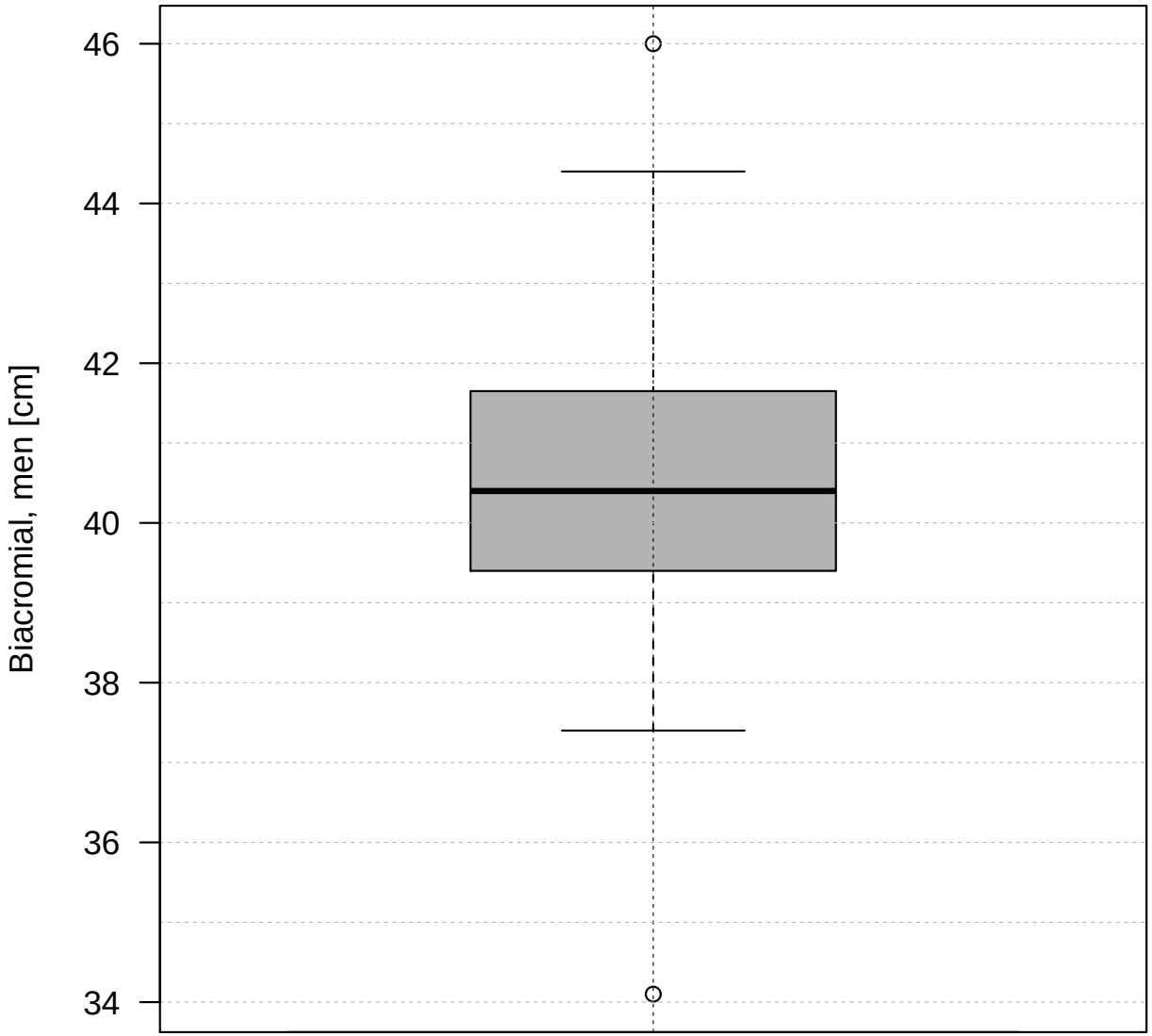
(4) Nun zeichnet man die Whiskers (= Barthaare), d. h. Linien (mit Querstrichen), die sich vom Rand der Box bis zu den äußersten Datenpunkten, die noch innerhalb der Fences liegen, erstrecken.

(5) Punkte die außerhalb der Fences liegen, werden extra gezeichnet. Sie gelten als (potenzielle) Ausreißer, d. h. als Punkte, die sich vom Gros der Daten absetzen.

Beispiel 1: Biacromial Men (Zufallsauswahl)

$$LF = 36.025$$

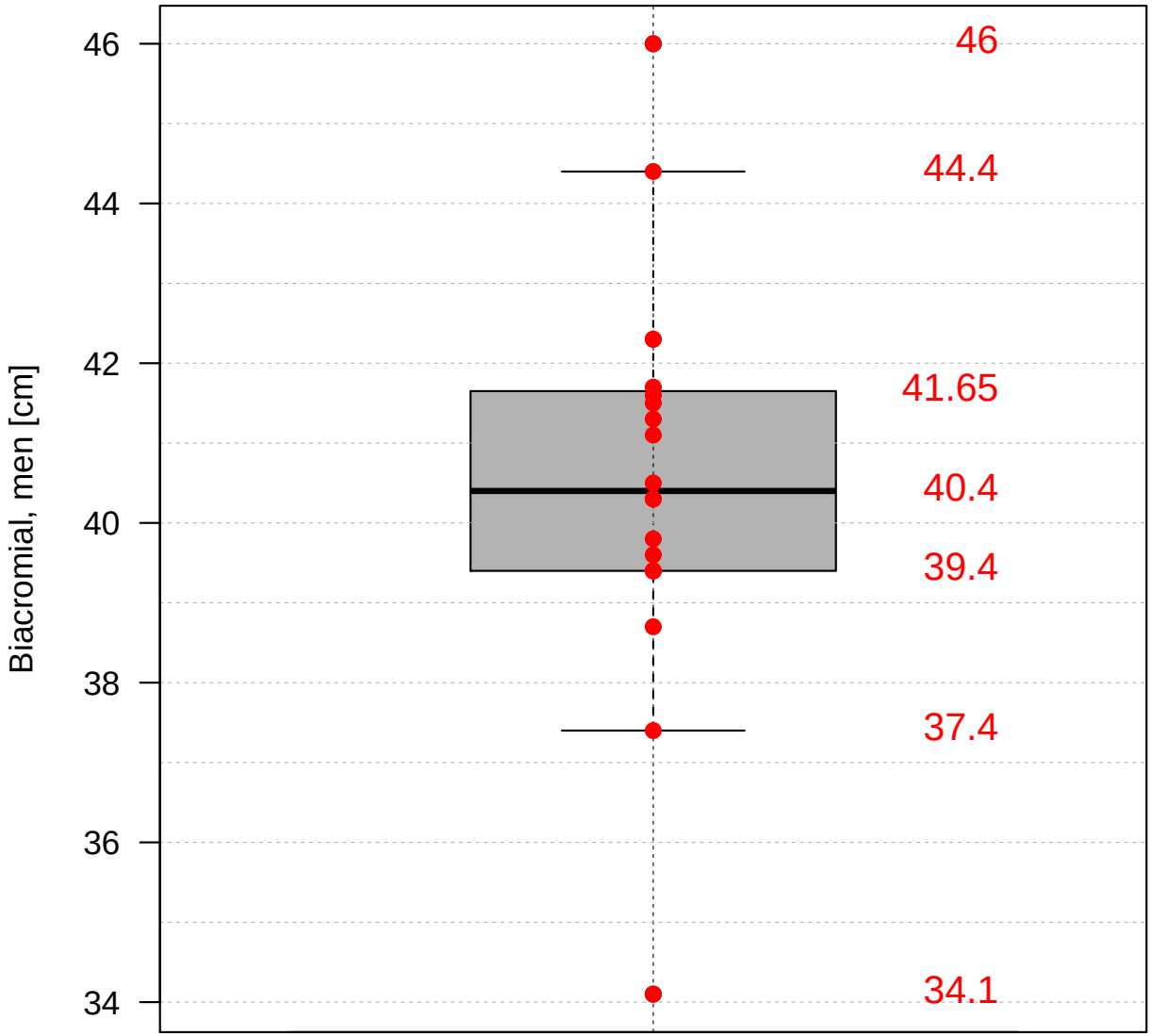
$$UF = 45.025$$



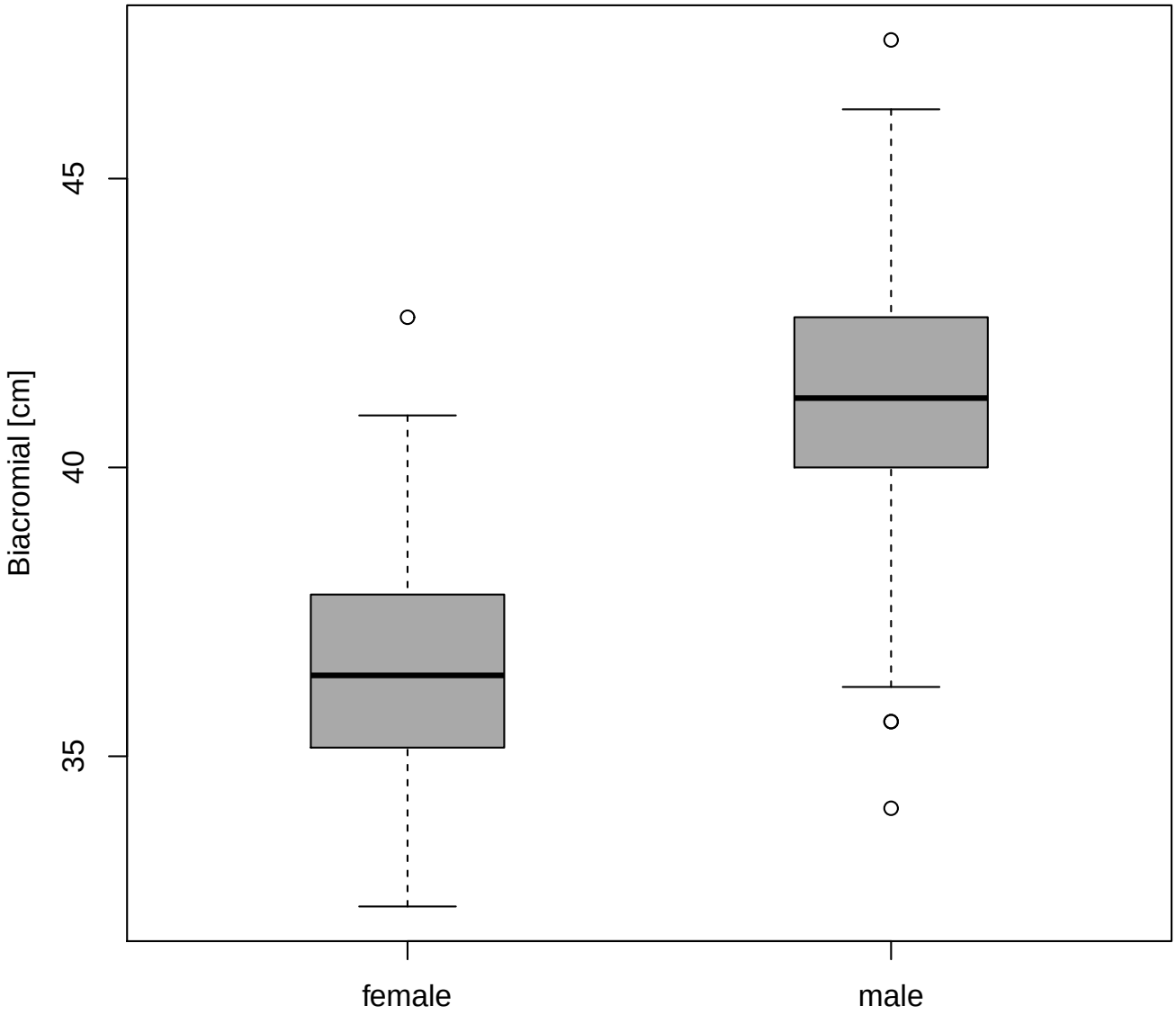
Beispiel 1: Biacromial Men (Zufallsauswahl)

$$LF = 36.025$$

$$UF = 45.025$$



Beispiel 2: Biacromial Men/Women



Kennzahlen

Neben der grafischen Aufbereitung von Daten ist die Berechnung von Stichprobenparametern (oder empirischen Parametern) eine unbedingt notwendige Ergänzung.

Da Ausreißer in der Praxis eher die Regel sind, spielt die Robustheit bei der Berechnung der Parameter eine wichtige Rolle.

Bei klassierten Daten werden die Kennzahlen so berechnet, als ob alle Daten einer Klasse in deren Mittelpunkt liegen.

Um einen Informationsverlust zu vermeiden, sollten Kennzahlen nach Möglichkeit auf Basis der unklassierten Daten (Rohdaten, Urdaten) berechnet werden.

Die Kennzahlen lassen sich in solche zur Beschreibung der Lage, des Streuverhaltens und der Form einer (empirisch gegebenen) Verteilung unterteilen.

Mittelwert

Der wichtigste Lageparameter ist der (empirische) Mittelwert (oder Stichprobenmittelwert), bezeichnet mit $\overline{x}_n$ (oder $\overline{x}$; sprich: „x quer“).

Sind $x_1, x_2, \ldots, x_n$ die Daten, so ist $\overline{x}_n$ ihr arithmetisches Mittel:

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

Der Mittelwert hat eine Minimumseigenschaft:

$$\sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \leq \sum_{i=1}^{n} (x_i - c)^2 \quad \text{für alle} \quad c \in \mathbb{R}$$

Berechnung aus Teilmittelwerten

Gegeben seien m Teilmittelwerte (basierend auf unterschiedlichen Anzahlen n_j, $j = 1, \ldots, m$, von Daten):

$$\overline{x}_{n_1}, \overline{x}_{n_2}, \ldots, \overline{x}_{n_m}$$

Gesamtmittelwert $\overline{\overline{x}}$ („x quer quer“) aller Daten:

$$\overline{\overline{x}} = \frac{1}{n} \sum_{i=1}^{m} n_j \overline{x}_{n_j} \quad \text{mit} \quad n = \sum_{j=1}^{m} n_j$$

Dabei handelt es sich um ein gewichtetes Mittel (mit den Gewichten n_j/n) der Teilmittelwerte.

Gewichteter Mittelwert

$$\overline{x}_g = \sum_{i=1}^{n} g_i x_i \quad \text{mit} \quad g_i \geq 0, \ \sum_{i=1}^{n} g_i = 1$$

Mittelwert aus klassierten Daten:

Klassen: $K_j, \ j = 1, \ldots, k$

Absolute (Relative) Klassenhäufigkeiten: $H_j \ (f_j), \ j = 1, \ldots, k$

Klassenmittelpunkte: $x_j^*, \ j = 1, \ldots, k$

Der Gesamtmittelwert ist das gewichtete Mittel der Klassenmitten:

$$\overline{x}_g = \frac{1}{\sum_{j=1}^{k} H_j} \sum_{j=1}^{k} H_j x_j^* = \sum_{j=1}^{k} f_j x_j^*$$

Geometrisches Mittel

Daten (nichtnegativ): $x_1, x_2, \ldots, x_n$

$$\overline{x}_n^{(g)} = \sqrt[n]{x_1 x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

Gewichtetes geometrisches Mittel:

$$\overline{x}_g^{(g)} = \prod_{i=1}^{n} x_i^{g_i} \quad \text{mit} \quad g_i \geq 0, \sum_{i=1}^{n} g_i = 1$$

Für $g_i = 1/n$ ergibt sich das gewöhnliche geometrische Mittel.

Anwendung bei relativen Änderungen: Lohnerhöhung, Bevölkerungs–wachstum, Preisindex, ...

Harmonisches Mittel

Daten (positiv): $x_1, x_2, \ldots, x_n$

$$\overline{x}_n^{(h)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Gewichtetes harmonisches Mittel:

$$\overline{x}_g^{(h)} = \frac{1}{\sum_{i=1}^{n} \frac{g_i}{x_i}} \quad \text{mit} \quad g_i \geq 0, \sum_{i=1}^{n} g_i = 1$$

Für $g_i = 1/n$ ergibt sich das gewöhnliche harmonische Mittel.

Typische Anwendung: Durchschnittsgeschwindigkeit

Beziehung zwischen den verschiedenen Mittelwerten

Hat man nur positive Beobachtungswerte $x_1, x_2, \ldots, x_n$, so gilt:

$$\overline{x}_n^{(h)} \leq \overline{x}_n^{(g)} \leq \overline{x}_n$$

Gleichheit besteht nur für $x_1 = x_2 = \cdots = x_n$.

Gilt auch für die verschiedenen gewichteten Mittelwerte.

Bsp: 1, 2, 3, 4, 5, 6

$$\overline{x} = \frac{1}{6}\sum_{i=1}^{6} i = 3.5, \quad \overline{x}^{(g)} = \left(\prod_{i=1}^{6} i\right)^{1/6} \doteq 2.994, \quad \overline{x}^{(h)} = \frac{6}{\sum_{i=1}^{6} \frac{1}{i}} \doteq 2.449$$

Median

Der Median wurde bereits als 50% Quantil (oder 2. Quartil) eines Datensatzes eingeführt.

Äquivalente Definition:

$$\widetilde{x} = \begin{cases} x_{((n+1)/2)} & n \text{ ungerade} \\ \frac{1}{2}\left[x_{(n/2)} + x_{((n+2)/2)}\right] & n \text{ gerade} \end{cases}$$

Der Median hat eine Minimumseigenschaft:

$$\sum_{i=1}^{n} |x_i - \widetilde{x}| \leq \sum_{i=1}^{n} |x_i - c| \quad \text{für alle} \quad c \in \mathbb{R}$$

Bruchpunkt

Die Robustheit eines Schätzers in Bezug auf Ausreißer lässt sich u. a. durch seinen Bruchpunkt bemessen.

Man versteht darunter den kleinsten Anteil (in %) der Datenwerte, den man ersetzen müsste, um den Schätzwert beliebig zu verändern.

Daten: $x_1, x_2, \ldots, x_n$

Ist n groß, beträgt der Bruchpunkt des

$$\text{Mittelwerts} \quad \bar{x} \quad \approx \quad 0\%$$

$$\text{Medians} \quad \tilde{x} \quad \approx \quad 50\%$$

Insofern ist der Median das robusteste Lagemaß.

Varianz

Neben Kennzahlen für die Lage benötigt man auch Kennzahlen zur Beschreibung des Streuverhaltens einer empirisch gegebenen Verteilung.

Die wichtigste Kennzahl dieser Art ist die (empirische) Varianz (oder Stichprobenvarianz) s_n^2 (kurz s^2):

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \qquad \text{(Bruchpkt. } \approx \text{ 0\%)}$$

Interpretation: s_n^2 ist mittlere quadratische Abweichung der Daten von ihrem Mittelwert $\overline{x}_n$.

Die Bedeutung des Faktors $1/(n-1)$ wird später erklärt ...

Varianz (Forts.)

Die Stichprobenstreuung (oder Standardabweichung) ist die (positive) Wurzel aus der Varianz:

$$s_n = \sqrt{s_n^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2} \qquad \text{(Bruchpkt. } \approx \text{ 0\%)}$$

Ist $\{x_1, x_2, \ldots, x_n\}$ die Gesamtpopulation, definiert man:

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

Achtung: Spricht man von der „Varianz“ oder der „Streuung“ eines Datensatzes ist stets s_n^2 bzw. s_n gemeint!

Verschiebungssatz

Die Varianz s_n^2 lässt sich auch wie folgt berechnen:

$$s_n^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n(\overline{x}_n)^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

Diese Form ist für praktische Berechnungen vorzuziehen.

Beweis: UE–Aufgabe!

Mean Absolute Deviation (MAD)

Verwendet man den Median $\widetilde{x}$ als Lageparameter kann man die folgenden Abstände bilden:

$$|x_1 - \widetilde{x}|, |x_2 - \widetilde{x}|, \ldots, |x_n - \widetilde{x}|$$

Den Mittelwert dieser Abstände nennt man den MAD:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \widetilde{x}| \qquad (\text{Bruchpkt.} \approx 0\%)$$

Wegen der mangelnden Robustheit des MAD nimmt man häufig wiederum den Median der Abstände:

$$\text{med}\{|x_1 - \widetilde{x}|, |x_2 - \widetilde{x}|, \ldots, |x_n - \widetilde{x}|\} \qquad (\text{Bruchpkt.} \approx 50\%)$$

Weitere Streuungsparameter

Spannweite (Range): „Spreizung“ der Daten

$$R = \text{Max} - \text{Min} = x_{(n)} - x_{(1)} \qquad (\text{Bruchpkt.} \approx 0\%)$$

Interquartilabstand (IQA, IQR): „Spreizung“ der mittleren 50% der Daten

$$\text{IQR} = 3.\text{Quartil} - 1.\text{Quartil} = Q_3 - Q_1 \qquad (\text{Bruchpkt.} \approx 25\%)$$

Hingeabstand (HA): „Spreizung“ der mittleren 50% der Daten

$$\text{HA} = \text{up. Hinge} - \text{lo. Hinge} \qquad (\text{Bruchpkt.} \approx 25\%)$$

5-Zahlen–Zusammenfassung

Übersichtliche Darstellung einiger Kennzahlen der Lage und Streuung:

$$x_{(1)}\,(\text{Min}),\ Q_1\,(1.\ \text{Quartil}),\ \widetilde{x}\,(\text{Med}),\ Q_3\,(3.\text{Quartil}),\ x_{(n)}\,(\text{Max})$$

oder

$$x_{(1)}\,(\text{Min}),\ \text{lo. Hinge},\ \widetilde{x}\,(\text{Med}),\ \text{up. Hinge},\ x_{(n)}\,(\text{Max})$$

Manchmal wird zusätzlich zum Median auch der Mittelwert $\overline{x}$ angezeigt.

Der Boxplot [62] ist i. W. eine grafische Darstellung der 5-Zahlen–Zusammenfassung.

Beispiel

Datensatz: `body.txt`

Variable: `Biacromial`

```
attach(dat)

by(Biacromial, Gender, summary)

Gender: 0
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.40   35.18   36.40   36.50   37.80   42.60
------------------------------------------------
Gender: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.10   40.00   41.20   41.24   42.60   47.40
```

Beispiel (Forts.)

```
by(Biacromial, Gender, fivenum)

Gender: 0
[1] 32.40 35.15 36.40 37.80 42.60
------------------------------------------------------------
Gender: 1
[1] 34.1 40.0 41.2 42.6 47.4

detach(dat)
```

Einheiten der Kenngrößen

Zu beachten ist, dass die meisten Kenngrößen – Ausnahmen sind der Variationskoeffizient (vgl. UE–Aufgabe 1.14) und Maßzahlen der Schiefe und Kurtosis (vgl. Skriptum S. 41ff) – auch Einheiten haben.

Der Mittelwert, die Quantile, die Hinges, der MAD, etc. haben die Dimension [D] der Beobachtungen.

Die Einheit der Varianz ist allerdings $[D^2]$.

Das macht die direkte Interpretation dieser Größe schwierig.

Andererseits hat aber die Streuung wiederum die Dimension [D] und lässt sich einfacher interpretieren.

Modalwert

Ein Bezugspunkt bei der Beurteilung der Verteilungsform ist der Modalwert (oder Modus).

Allgemein versteht man darunter eine Merkmalsausprägung mit höchster „Dichte“.

Diskrete Merkmale: Ausprägung mit der höchsten Beobachtungshäufigkeit

Stetige Merkmale: Mittelpunkt der Klasse mit der höchsten beobachteten (relativen) Häufigkeit ($\longrightarrow$ Modalklasse).

In vielen Fällen ist der Modalwert mehr oder weniger deutlich ausgeprägt.

Manchmal gibt es aber zwei oder mehr deutlich erkennbare – i. A. nicht gleich hohe – „Gipfel“ ($\longrightarrow$ bimodale, multimodale Verteilung).

Bsp: Verteilungsmischung (vgl. Skriptum S. 40)

Momente

Daten: $x_1, x_2, \ldots, x_n$

(Empirische) Momente: Momente um den Nullpunkt

$$r\text{–tes Moment:} \quad m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r \quad \text{für} \quad r = 1, 2, \ldots$$

Zentrale (empirische) Momente: Momente um $\overline{x}$

$$r\text{–tes zentrales Moment:} \quad m_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^r \quad \text{für} \quad r = 1, 2, \ldots$$

Bsp: Der Mittelwert $\overline{x}$ $(= m'_1)$ ist das 1. Moment, die Varianz s_n^2 ist – bis auf den Faktor $1/(n-1)$ – das 2. zentrale Moment der Daten.

Schiefe

Um die Schiefe einer (empirisch gegebenen) Verteilung zu charakterisieren, kann man sich der (zentralen) Momente der Ordnung 2 und 3 bedienen.

Klassische Definition:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\sqrt{n}}{(n-1)\sqrt{n-1}} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^3$$

$$g_1 > 0 \quad \Longrightarrow \quad \text{linkssteil / rechtsschief}$$

$$g_1 \approx 0 \quad \Longrightarrow \quad \text{symmetrisch}$$

$$g_1 < 0 \quad \Longrightarrow \quad \text{rechtssteil / linksschief}$$

Kurtosis (Wölbung)

Um die Kurtosis (Wölbung) einer (empirisch gegebenen) Verteilung zu charakterisieren, kann man sich der (zentralen) Momente der Ordnung 2 und 4 bedienen.

Klassische Definition:

$$g_2 = \frac{m_4}{m_2^2} = \frac{n}{(n-1)^2} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s} \right)^4$$

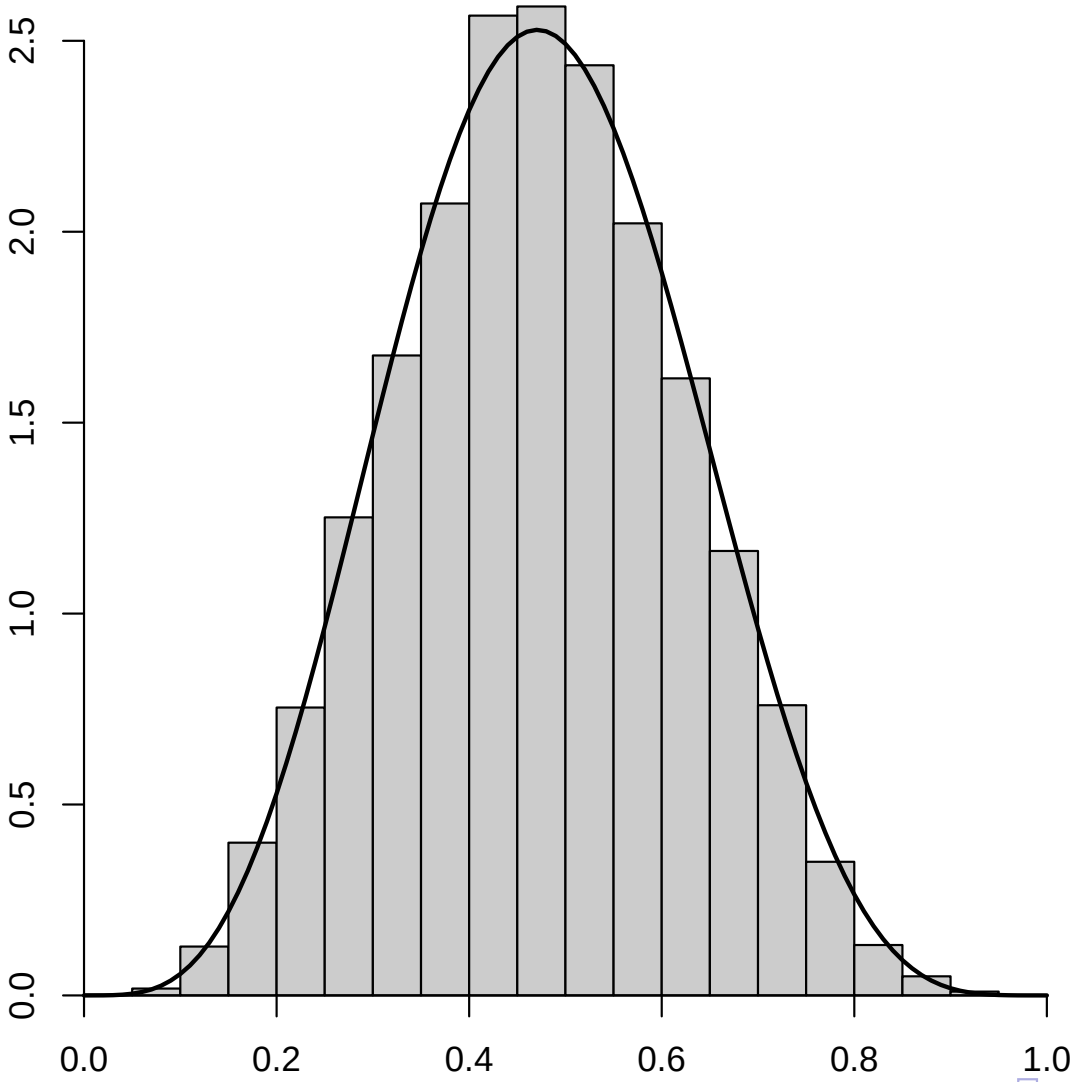
$g_2 < 3 \quad \Longrightarrow \quad$ platykurtisch / flach gewölbt

$g_2 \approx 3 \quad \Longrightarrow \quad$ mesokurtisch / mittel gewölbt

$g_2 > 3 \quad \Longrightarrow \quad$ leptokurtisch / steilgipfelig

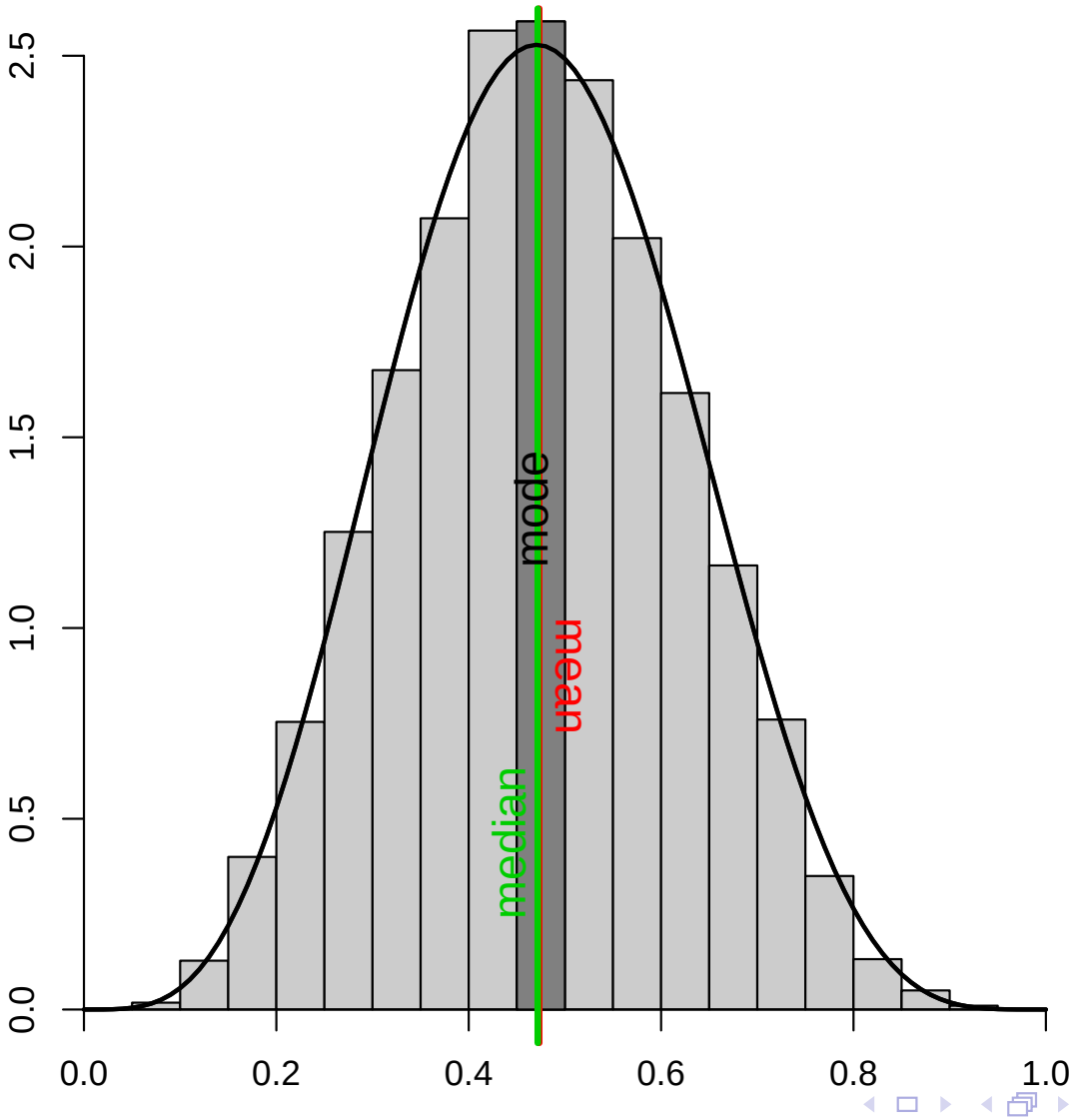
Beispiel 1

$g_1 = 0.04$ $g_2 = 2.58$

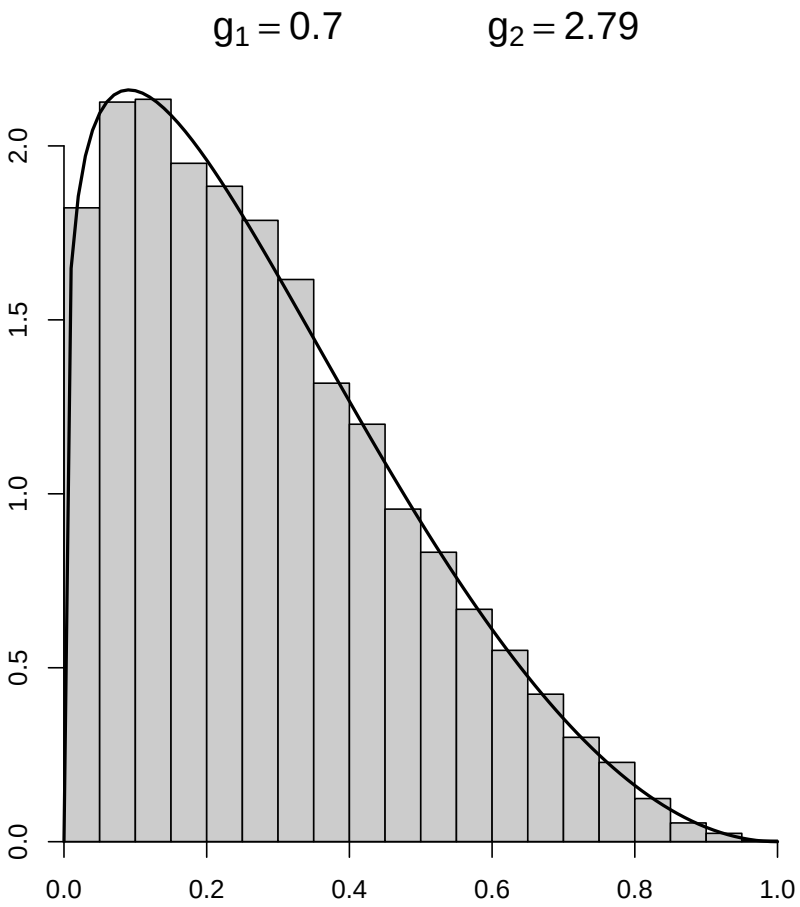


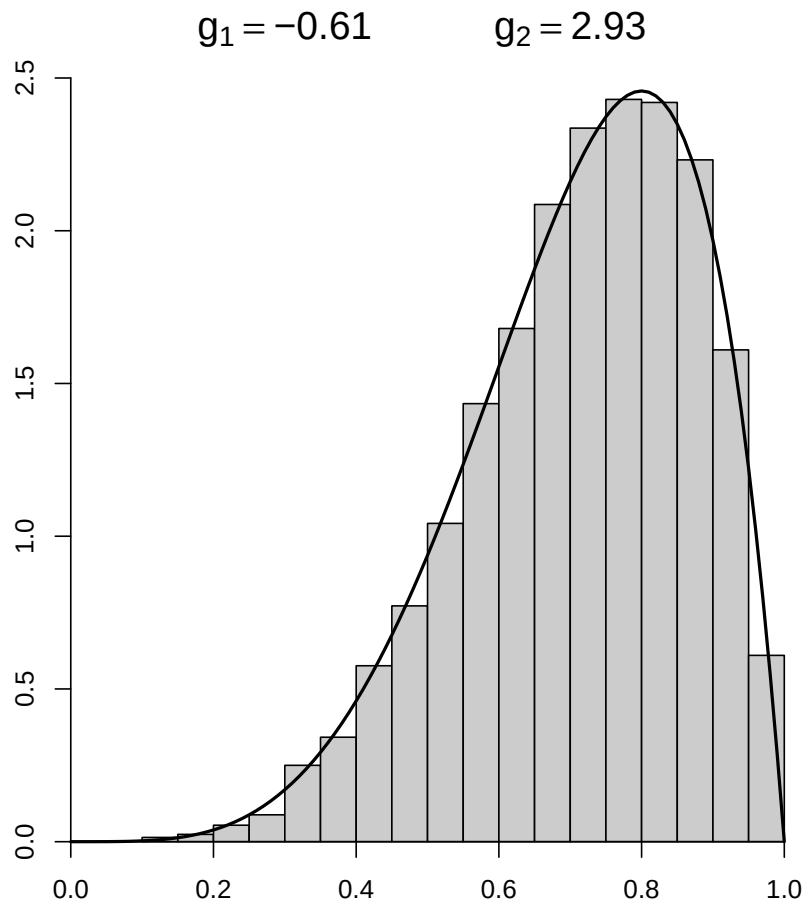
Beispiel 1

$g_1 = 0.04$ $g_2 = 2.58$

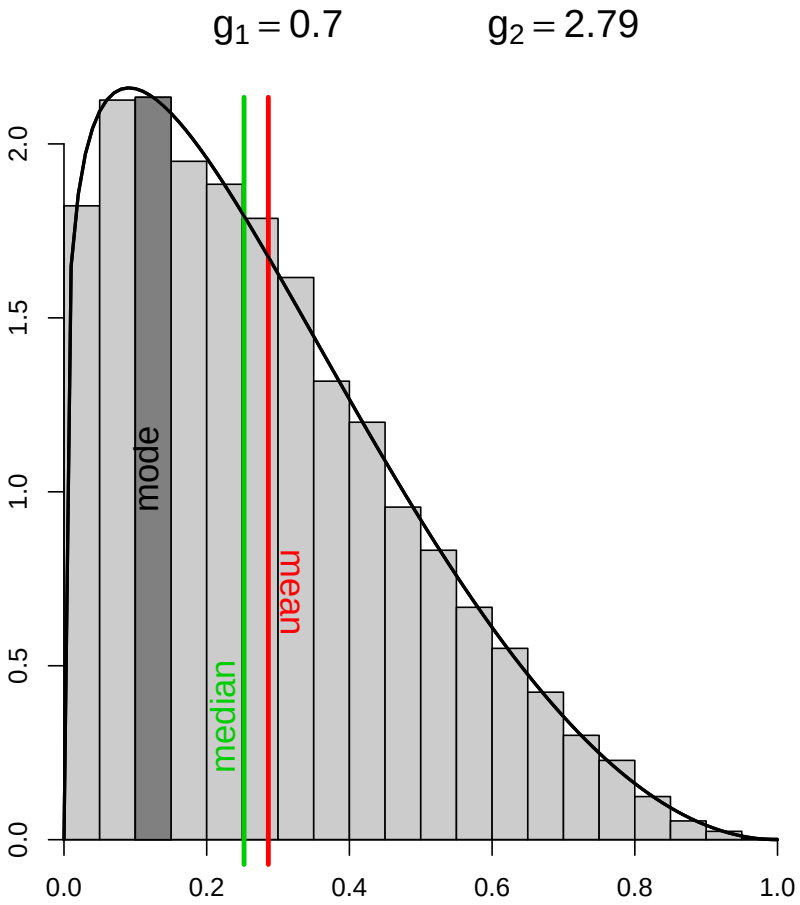


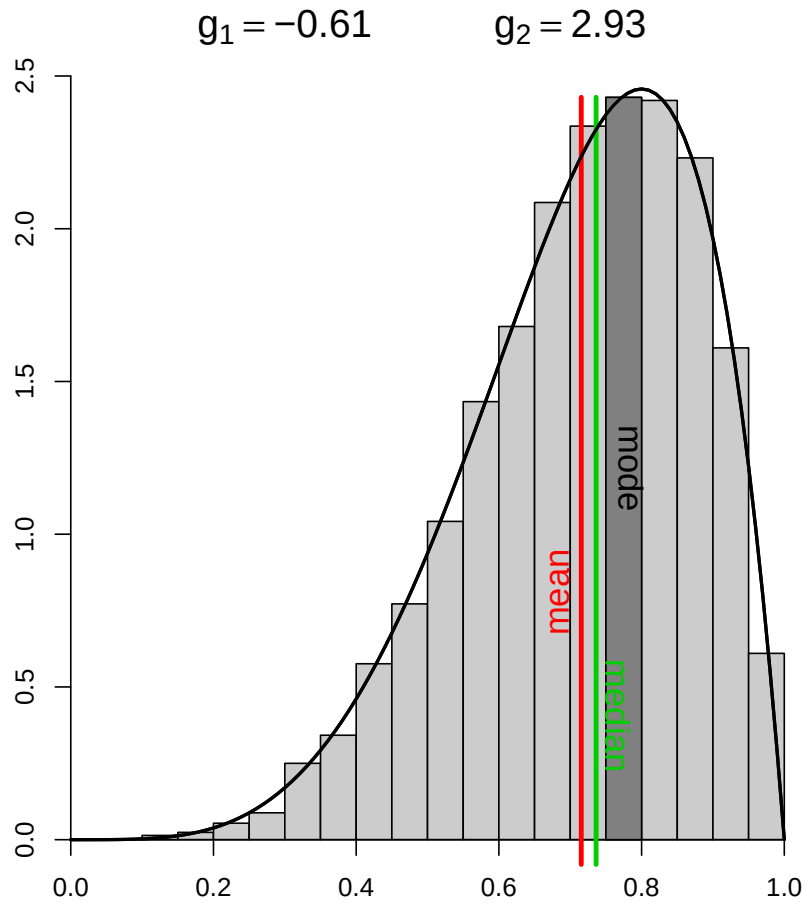
Beispiel 2


$g_1 = 0.7$
$g_2 = 2.79$
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0


$g_1 = -0.61$
$g_2 = 2.93$
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0

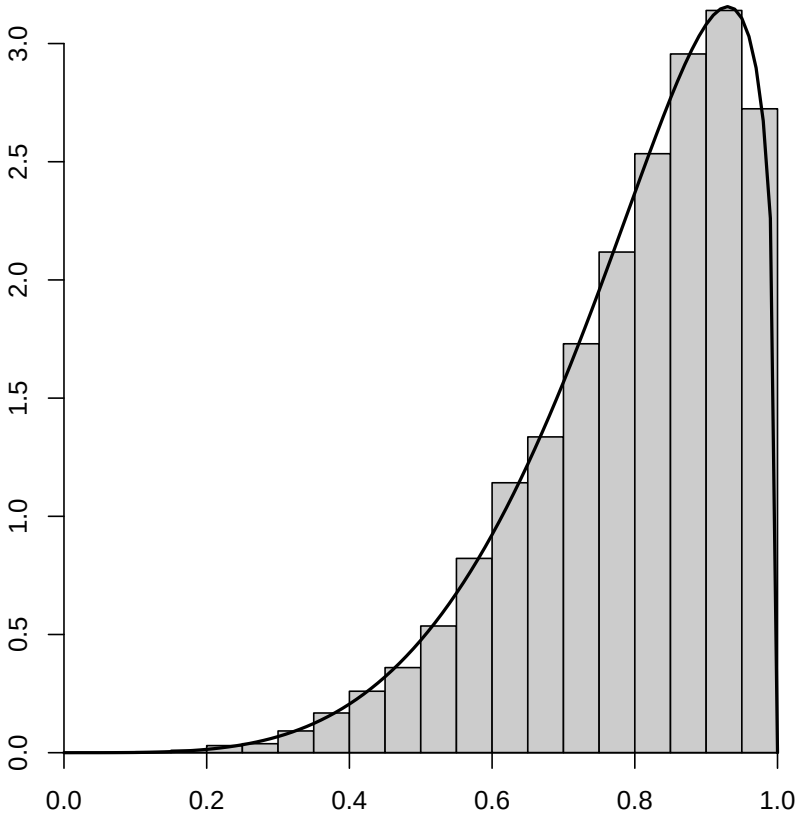
Beispiel 2


$g_1 = 0.7$
$g_2 = 2.79$
mode
median
mean
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0

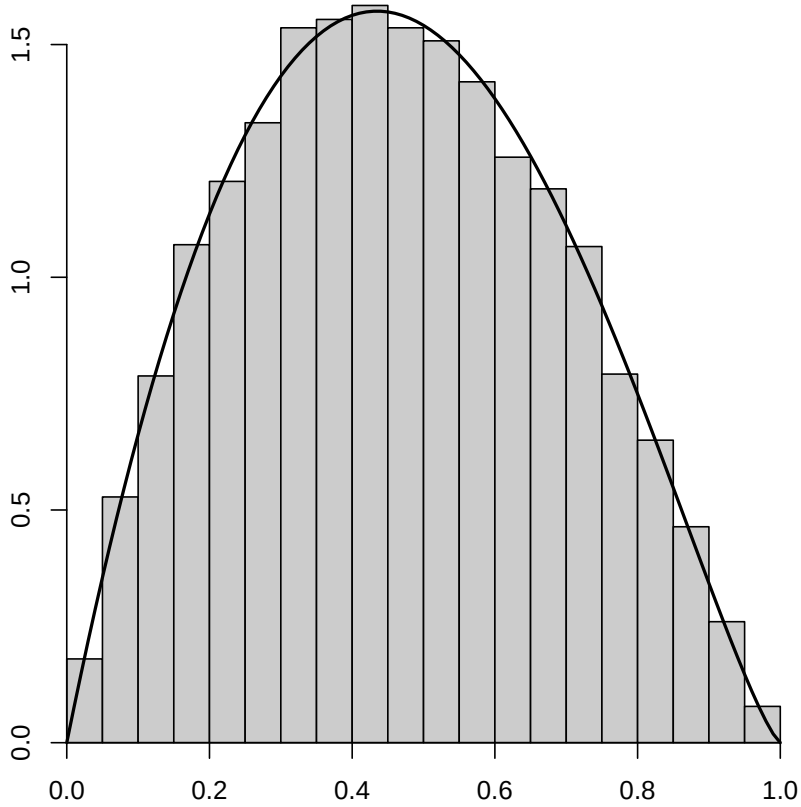

$g_1 = -0.61$
$g_2 = 2.93$
mean
median
mode
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0

Beispiel 3

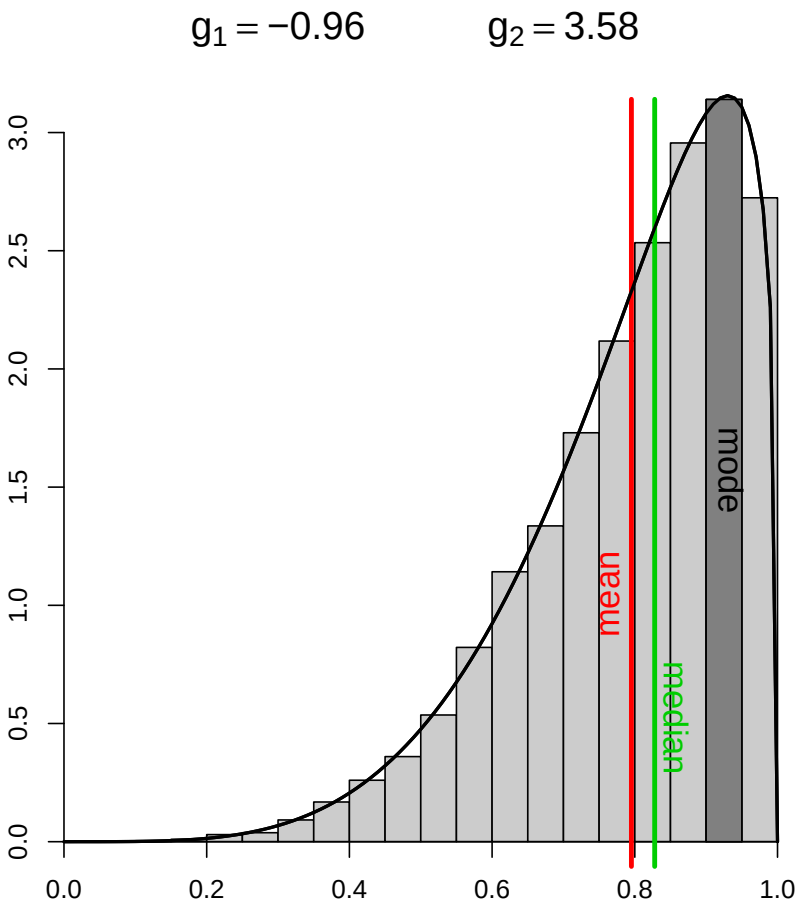
$g_1 = -0.96$ $g_2 = 3.58$

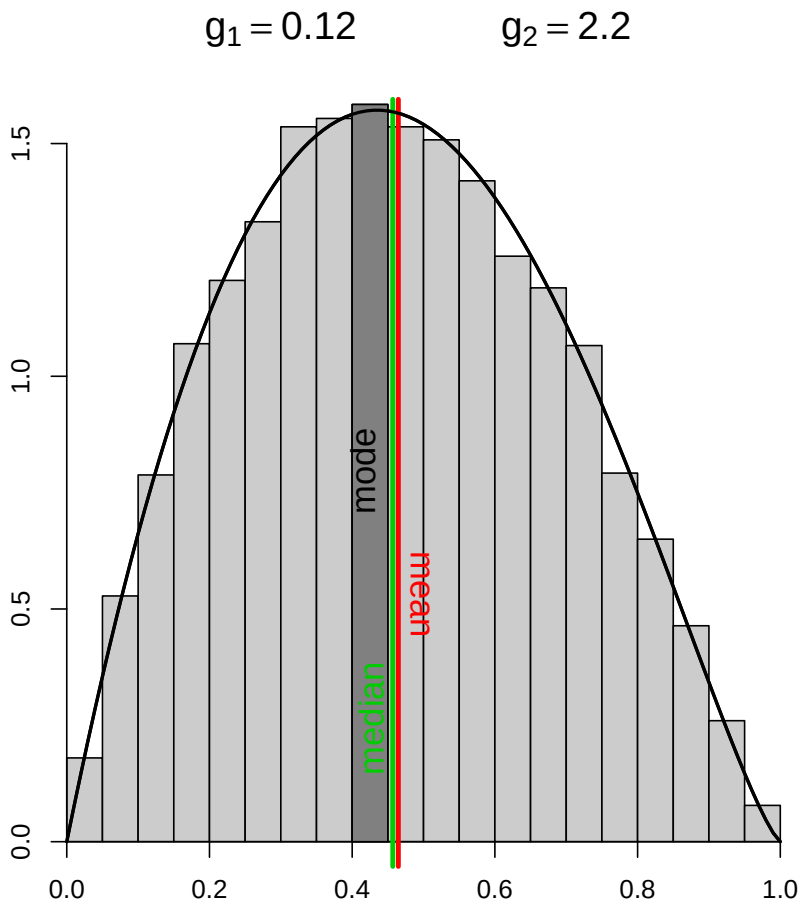


$g_1 = 0.12$ $g_2 = 2.2$



Beispiel 3


$g_1 = -0.96$
$g_2 = 3.58$
mean
median
mode
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0


$g_1 = 0.12$
$g_2 = 2.2$
mode
median
mean
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0

Multivariate Daten

Werden an den Einheiten Beobachtungen (Messungen) für mehrere Merkmale vorgenommen, spricht man von multivariaten Daten.

Neben der Analyse der einzelnen Merkmale stehen insbesondere die Beziehungen zwischen den Merkmalen im Mittelpunkt des Interesses.

Dazu kann man sich der vielfältigen Methoden der multivariaten Statistik bedienen.

Im Folgenden nur einige einfache grafische Methoden und ein kurzer Einblick in die Korrelations– und Regressionsrechnung für die Analyse von quantitativen (d. h., metrischen) mehrdimensionalen (speziell 2–dimensionalen) Merkmalen.

Datenframes

n Beobachtungen zu je p Variablen (oder Merkmalen):

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}), \quad i = 1, 2, \ldots, n$$

Die Beobachtungsvektoren $\mathbf{x}_i$ werden in einer Datenmatrix – oder einem Datenframe – zusammengefasst:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \qquad (n \times p) - \text{Matrix}$$

Die Zeilen von $\mathbf{X}$ entsprechen den Beobachtungen.

Die Spalten von $\mathbf{X}$ entsprechen den Merkmalen.

Beispiel: body.txt

	Biacromial	Waist	Age	Weight	Height	Gender
1	42.9	71.5	21	65.6	174.0	1
2	43.7	79.0	23	71.8	175.3	1
3	40.1	83.2	28	80.7	193.5	1
4	44.3	77.8	23	72.6	186.5	1
5	42.5	80.0	22	78.8	187.2	1
.....						
505	34.7	57.9	33	48.6	160.7	0
506	38.5	72.2	33	66.4	174.0	0
507	35.6	80.4	38	67.3	163.8	0

Mit Ausnahme des nominellen Merkmals `Gender` sind alle Variablen metrisch skalierte Merkmale auf einer Verhältnisskala.

Scatterplots

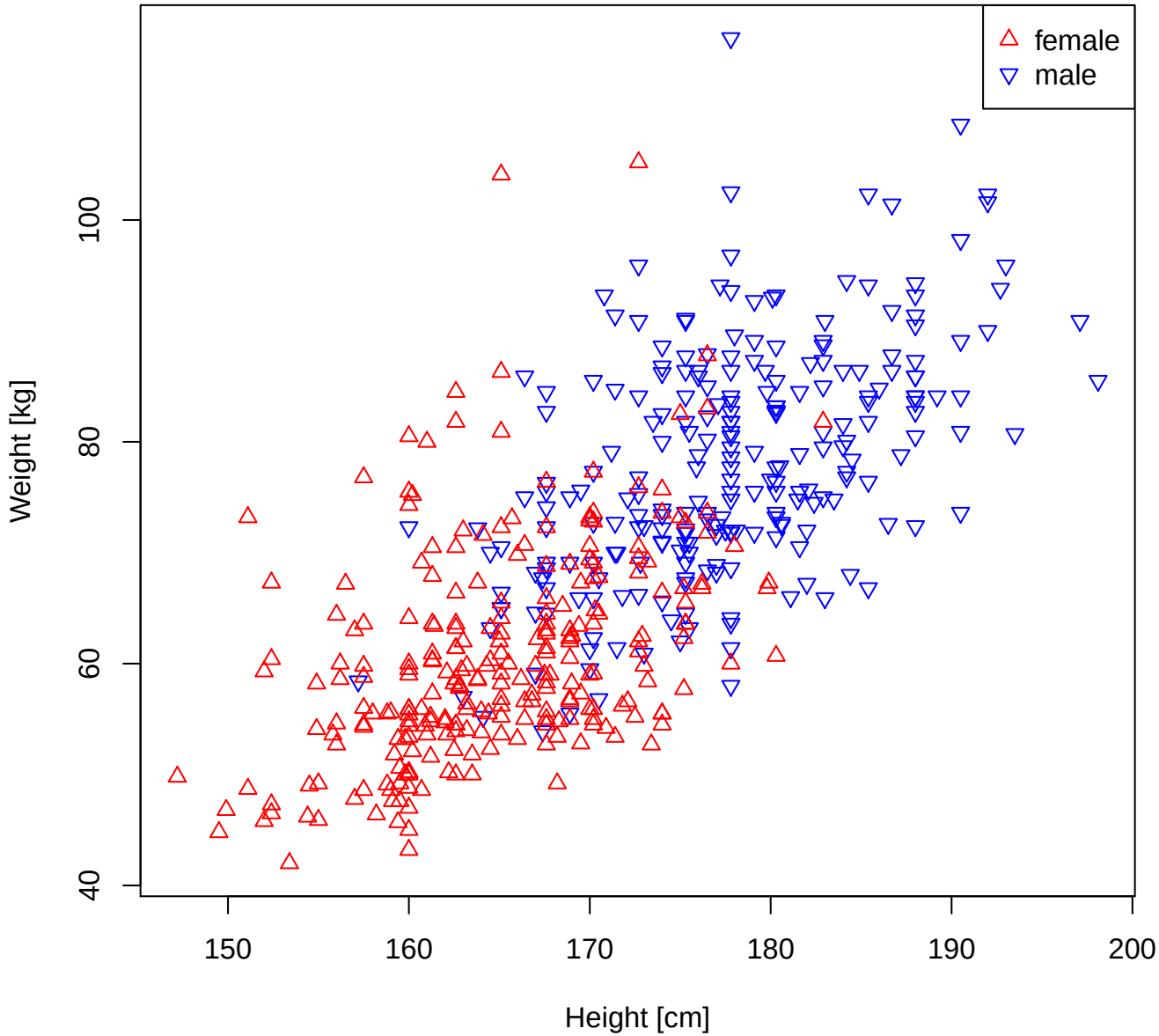
Bei zwei (metrischen) Merkmalen kann man die Beobachtungspaare (x_{1i}, x_{i2}), $i = 1, 2, \ldots, n$, als Punkte in einem Koordinatensystem interpretieren und in Form eines Scatterplots darstellen.

Durch „Überladen“ der Punkte eines Scatterplots (Farbe, Größe/Art der Punkte, u. Ä.) können weitere – meist nominelle – Merkmale repräsentiert werden.

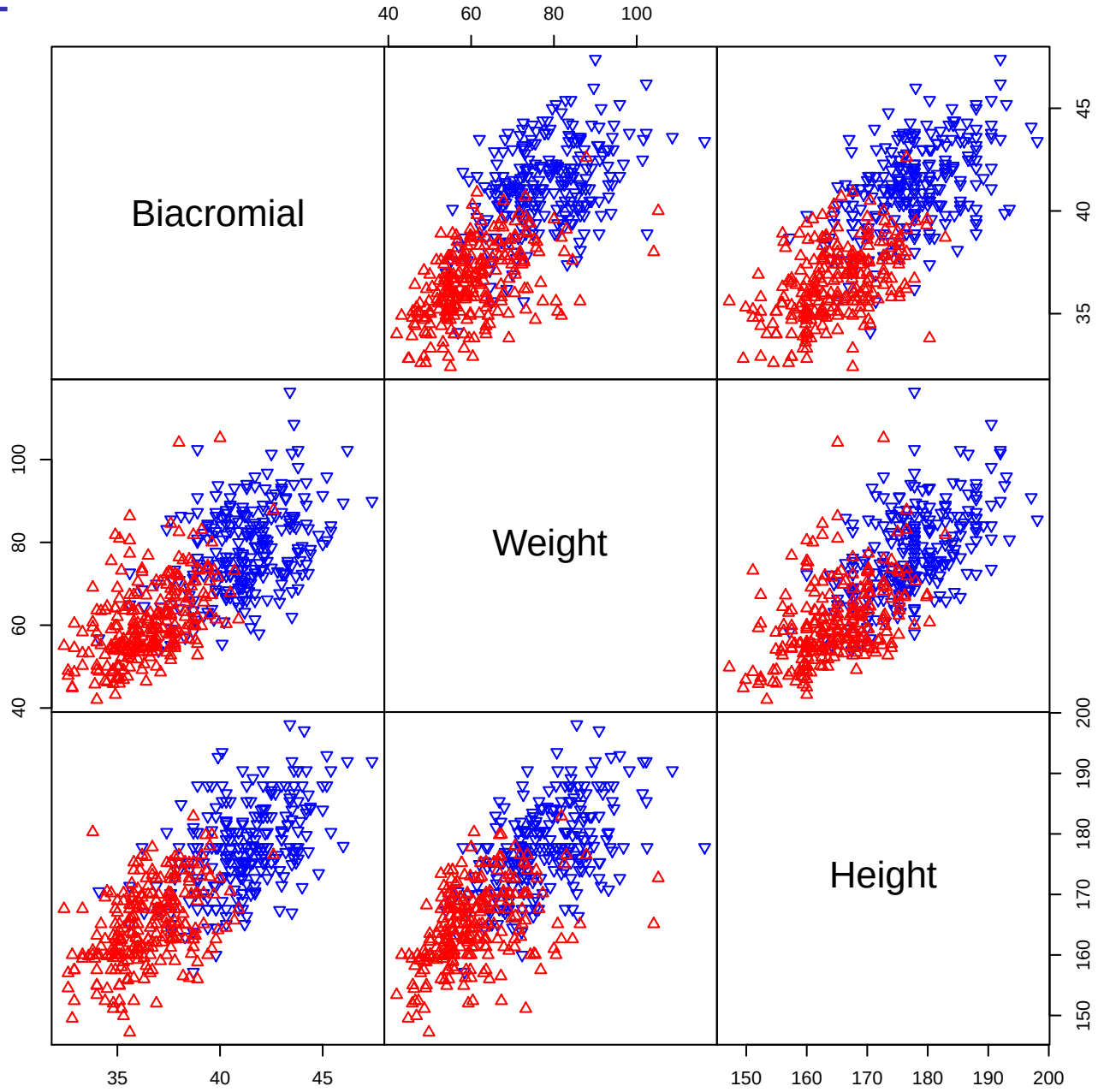
Bei mehr als zwei (metrischer) Merkmalen kann man die Merkmale paarweise gegeneinander zeichnen und in Form einer Scatterplotmatrix darstellen.

Scatterplots lassen sich durch zusätzliche grafische (oder numerische) Elemente (Histogramme, Boxplots, Trendkurven, etc.) ergänzen.

Beispiel 1


female
male
Weight [kg]
40
60
80
100
150
160
170
180
190
200
Height [cm]

Beispiel 2


Biacromial
Weight
Height
40
60
80
100
35
40
45
150
160
170
180
190
200

Linearer Zusammenhang

Ein Scatterplot gibt nicht nur eine grafische Veranschaulichung eines bivariaten Datensatzes, sondern zeigt auch die Art und Stärke eines eventuell vorhandenen Zusammenhangs zwischen den Merkmalen.

Beispielsweise [99] ist der Zusammenhang von `Weight` und `Height` bei beiden Geschlechtern – wenn auch nicht allzu stark ausgeprägt – grob linearer Natur mit einer positiven Tendenz.

Neben qualitativen Feststellungen der obigen Art möchte man aber auch Art und Stärke des linearen Zusammenhangs zwischen zwei Merkmalen zahlenmäßig beschreiben.

Das führt zum Konzept des Korrelationskoeffizienten.

Korrelationskoeffizient

Beobachtungspaare: $(x_i, y_i),\ i = 1, 2, \ldots, n$

(Empirische) Mittelwerte und Varianzen: $\overline{x},\ s_x^2,\ \overline{y},\ s_y^2$

Standardisierte Abweichungen (vom Mittelwert):

$$\frac{x_i - \overline{x}}{s_x}, \quad \frac{y_i - \overline{y}}{s_y}, \quad i = 1, 2, \ldots, n$$

Der „Durchschnitt“ aus den Produkten dieser Abweichungen ist der (empirische) Korrelationskoeffizient:

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

Korrelationskoeffizient (Forts.)

(Empirische) Kovarianz: $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$

Damit folgt: $r_{xy} = \frac{s_{xy}}{s_x \, s_y} = \frac{\text{Kovarianz } (x, y)}{\text{Streuung } (x) \, \text{Streuung } (y)}$

Eigenschaften:

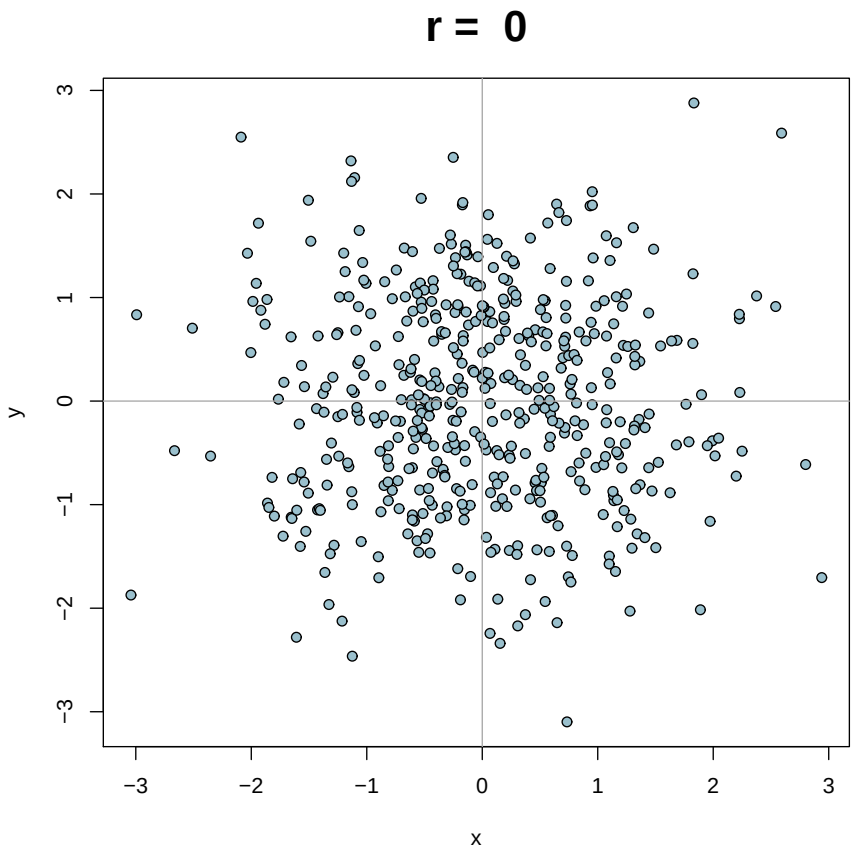
(1) $s_{xx} = s_x^2, \quad s_{yy} = s_y^2$ (Varianz ist spezielle Kovarianz)

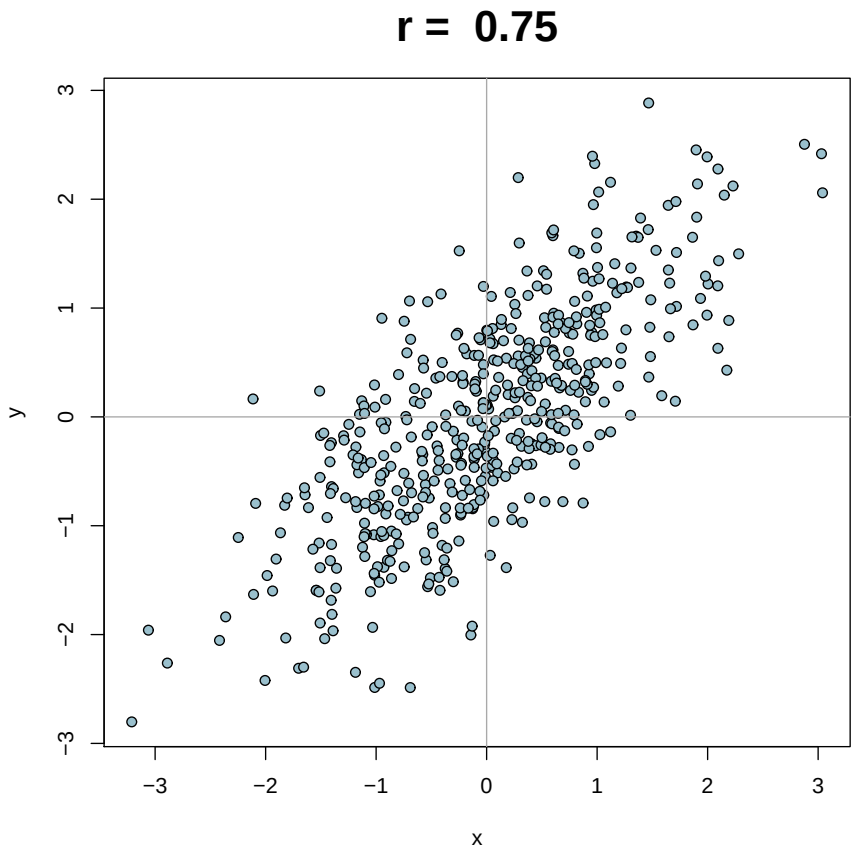
(2) $s_{xy} = s_{yx}, \quad r_{xy} = r_{yx}$ (symmetrisch in x und y)

(3) $-\infty < s_{xy} < \infty$ (nicht normiert, dimensionsbehaftet)

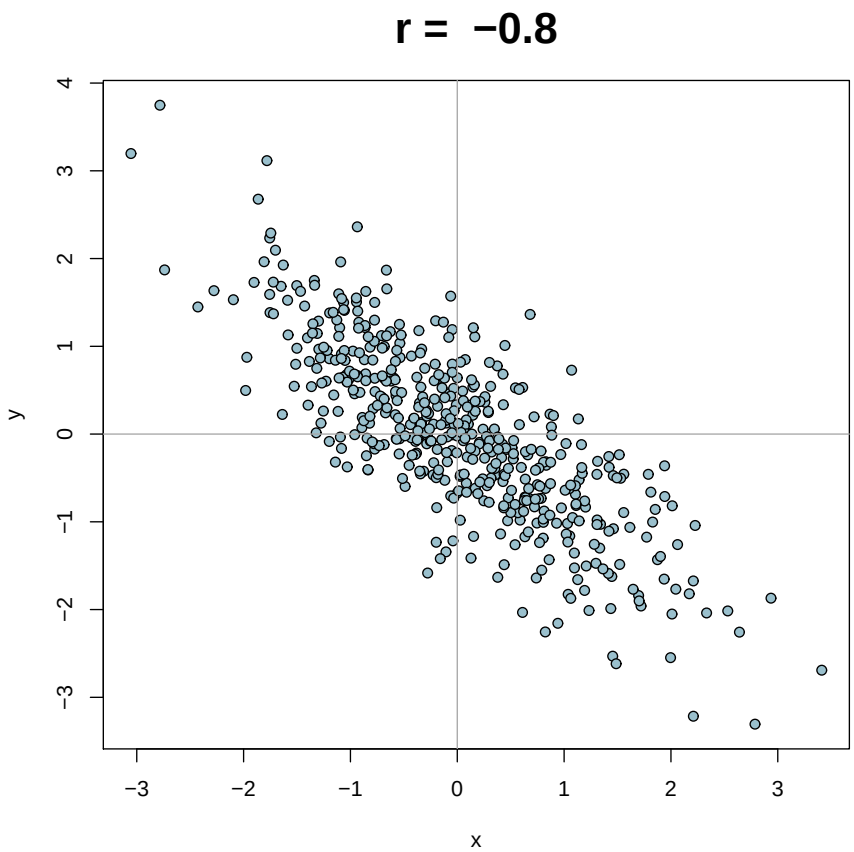
(4) $-1 \leq r_{xy} \leq 1$ (normiert, dimensionslos)

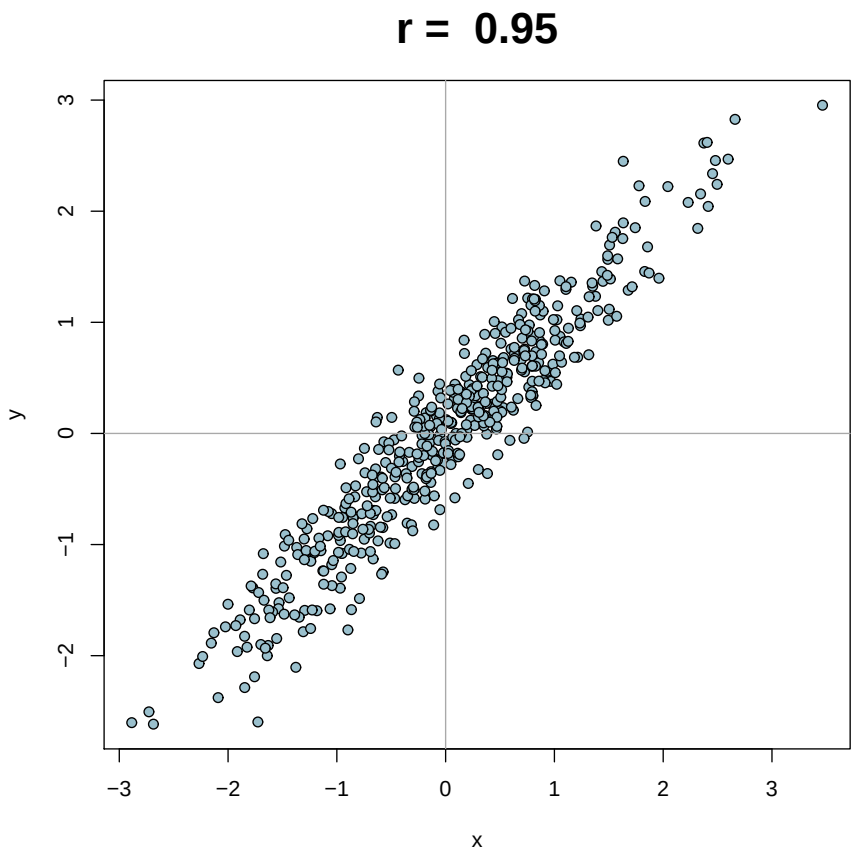
Beispiel





Beispiel (Forts.)





Wie ist r zu interpretieren?

- Das Vorzeichen von r zeigt die Richtung des Zusammenhangs.

  Ein positiver Wert signalisiert eine gleichsinnige Assoziation.

  Ist die Ausprägung eines Merkmals größer als der Durchschnitt, ist die Ausprägung des anderen Merkmals tendenziell auch größer als der Durchschnitt.

  Eine analoge Interpretation gilt für negative Werte.

- Der Absolutbetrag von r sagt etwas über die Stärke der Assoziation.

  Für $r = +1$ liegen die Punkte (x_i, y_i), $i = 1, 2, \ldots, n$, exakt auf einer Geraden $y_i = a + b\,x_i$ mit $b > 0$.

  Für $r = -1$ liegen die Punkte (x_i, y_i), $i = 1, 2, \ldots, n$, exakt auf einer Geraden $y_i = a + b\,x_i$ mit $b < 0$.

  Für r in der Nähe von 0 gibt es nur einen schwachen (linearen) Zusammenhang.

Wie ist r zu interpretieren? (Forts.)

- Der Korrelationskoeffizient ist nur ein Maß für die lineare Assoziation zwischen zwei Merkmalen.

  Andere, kompliziertere Formen der Assoziation werden von ihm nicht (ausreichend) erfasst.

- Achtung: Korrelation ist nicht gleichbedeutend mit Kausalität.

  Der Umstand, dass zwei Merkmale (stark) korrelieren bedeutet nicht notwendigerweise, dass auch eine Ursache-Wirkungsbeziehung zwischen ihnen besteht.

  Möglicherweise gibt es weitere Variablen („Confounder“), die beide Merkmale beeinflussen.

- Bei der Interpretation von r ist auch zu beachten, dass die Größe von r vom Beobachtungsbereich der beiden Merkmale abhängt.

Prinzip der kleinsten Quadrate

An n Punkte (x_i, y_i), $i = 1, 2 \ldots, n$, soll eine Kurve $y = h(x; \alpha, \beta)$, die von zwei Parametern α und β abhängt, angepasst werden.

Abstände zwischen y_i (= beobachteter Wert zu x_i) und $h(x_i; \alpha, \beta)$ (= Wert der Kurve an der Stelle x_i):

$$d_i = y_i - h(x_i; \alpha, \beta), \quad i = 1, 2, \ldots, n$$

Nach dem Prinzip der kleinsten Quadrate werden die Parameter α und β so bestimmt, dass die Summe der Abstandsquadrate minimal wird:

$$S(\alpha, \beta) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left[y_i - h(x_i; \alpha, \beta) \right]^2 \longrightarrow \text{Min!}$$

Spezialfall: Ausgleichsgerade

Im Falle einer Geraden $y = h(x; \alpha, \beta) = \alpha + \beta x$, besagt die Methode der kleinsten Quadrate: Bestimme α und β so, dass

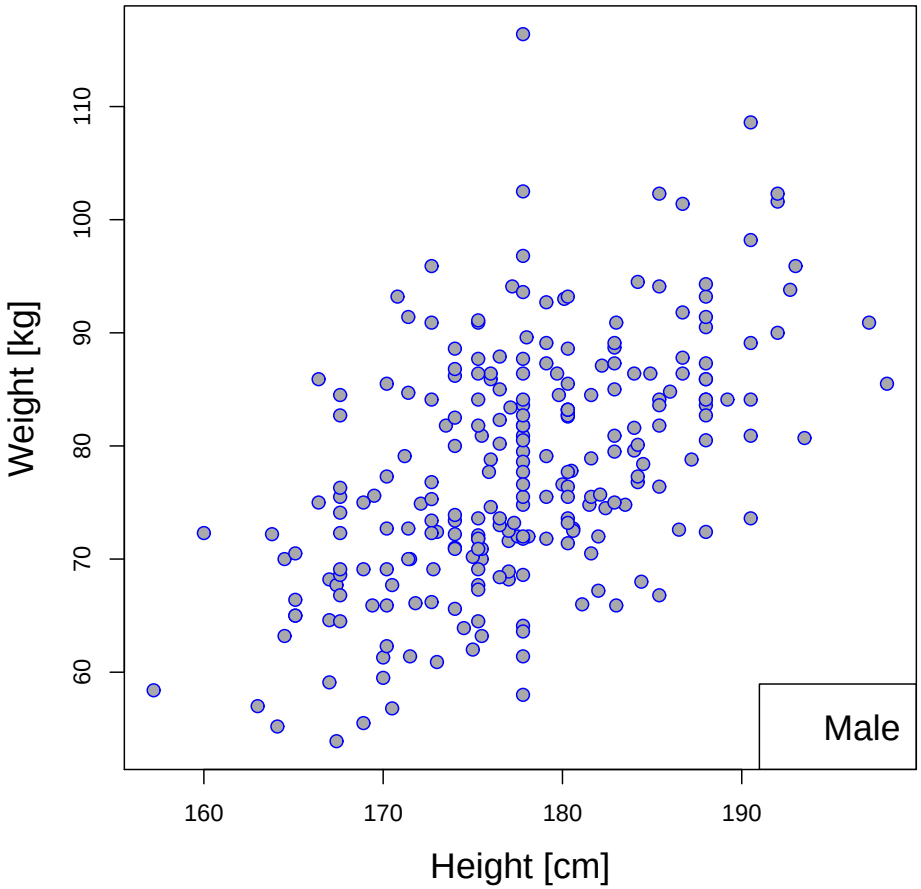
$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \longrightarrow \text{Min!}$$

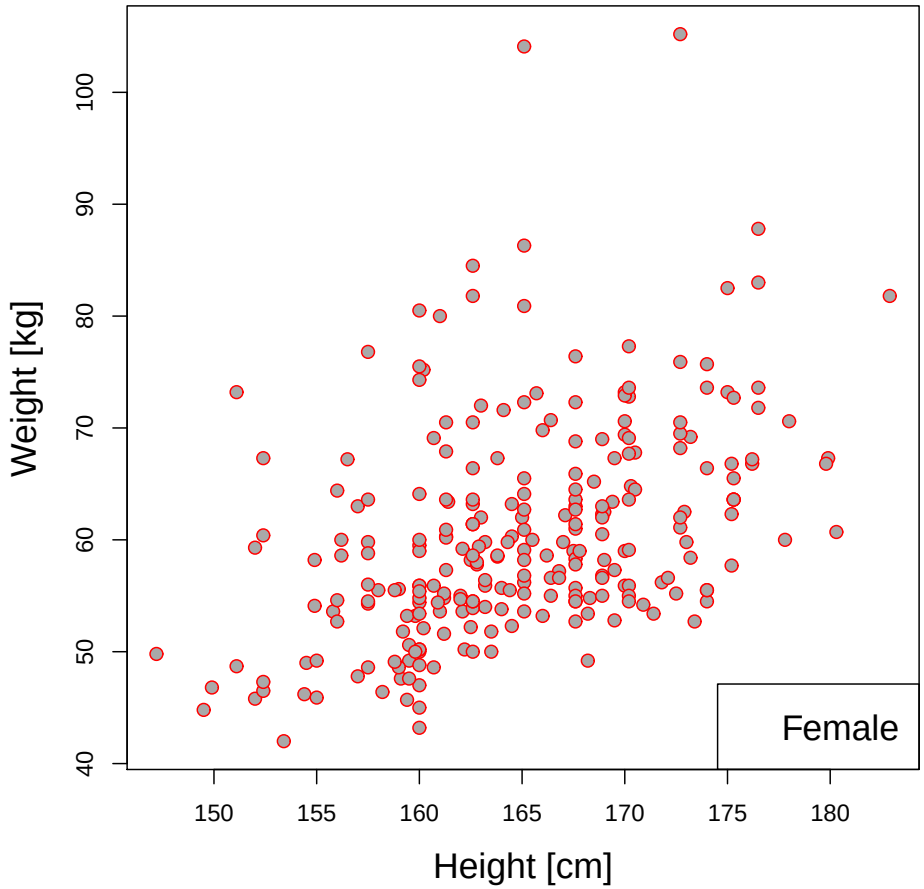
Dieses Minimierungsproblem führt auf ein lineares Gleichungssystem (vgl. Skriptum S. 57), dessen Lösung wie folgt lautet:

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}, \qquad \widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = r_{xy}\,\frac{s_y}{s_x}$$

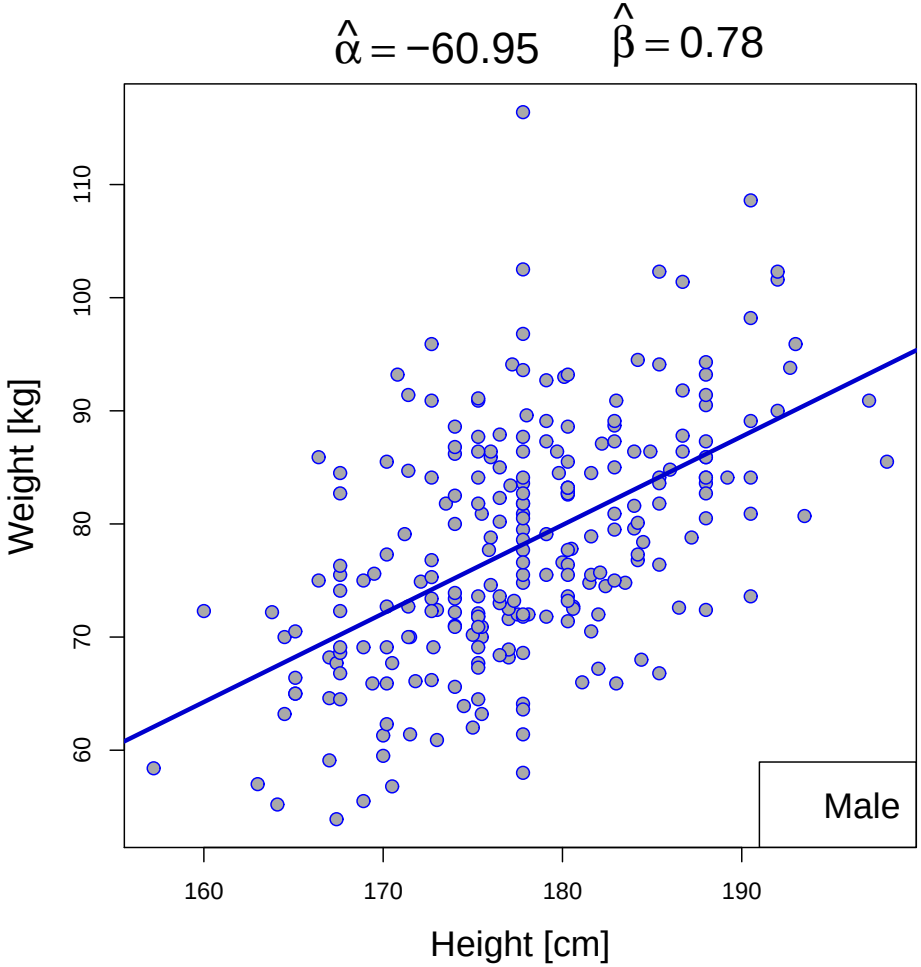
Die angepasste Gerade verläuft durch den Punkt $(\overline{x}, \overline{y})$.

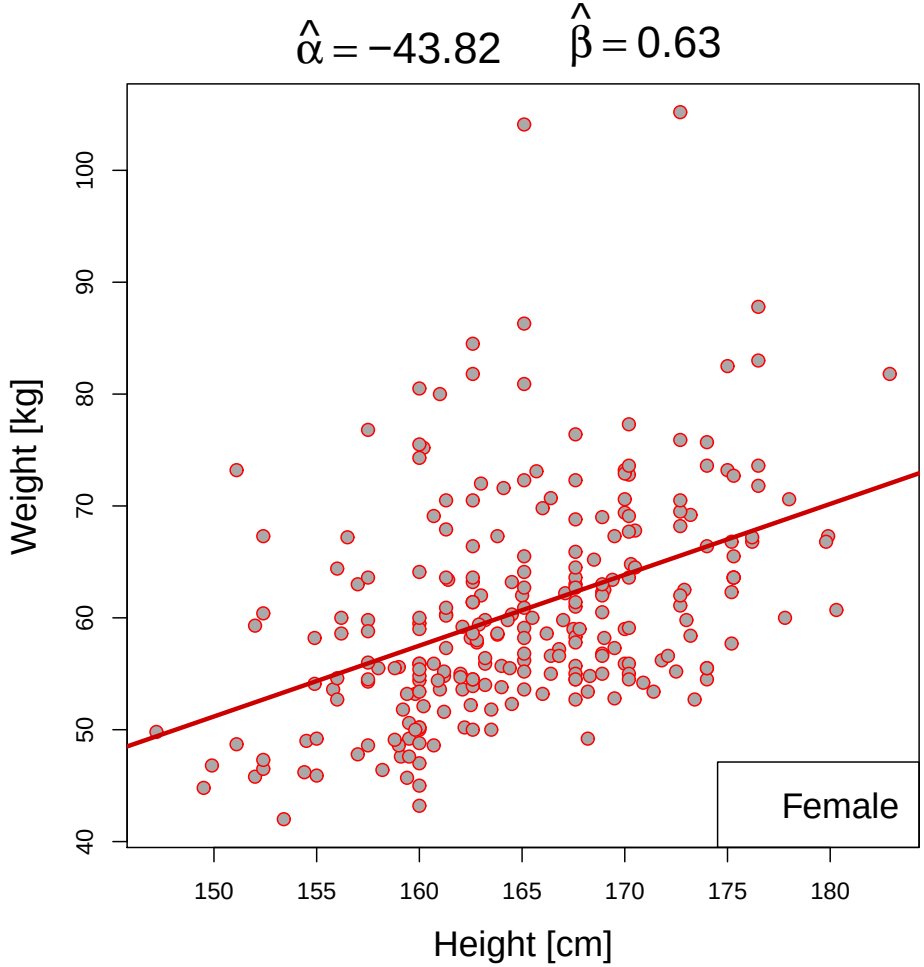
Beispiel: `body.txt`



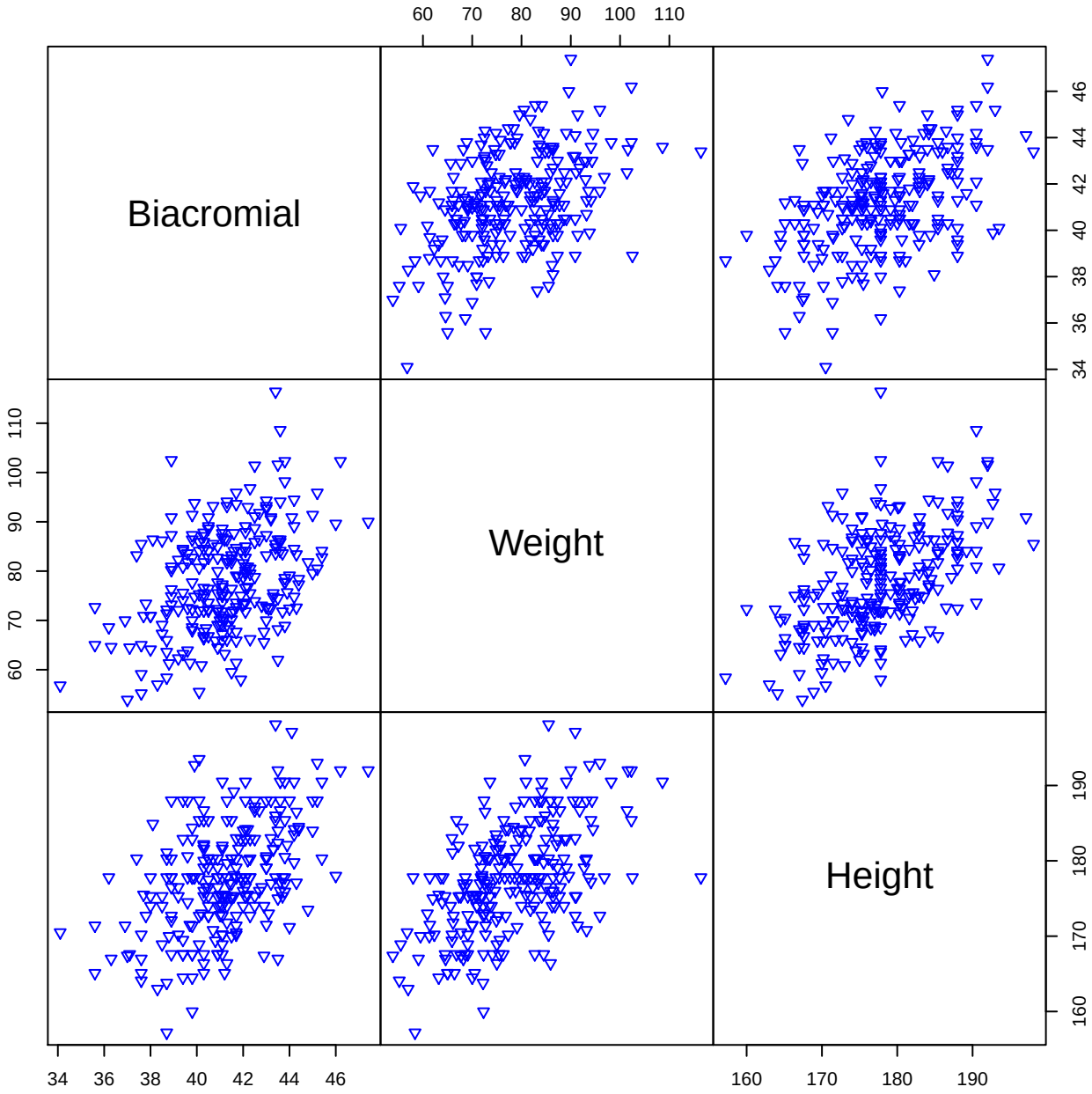


Beispiel: `body.txt`

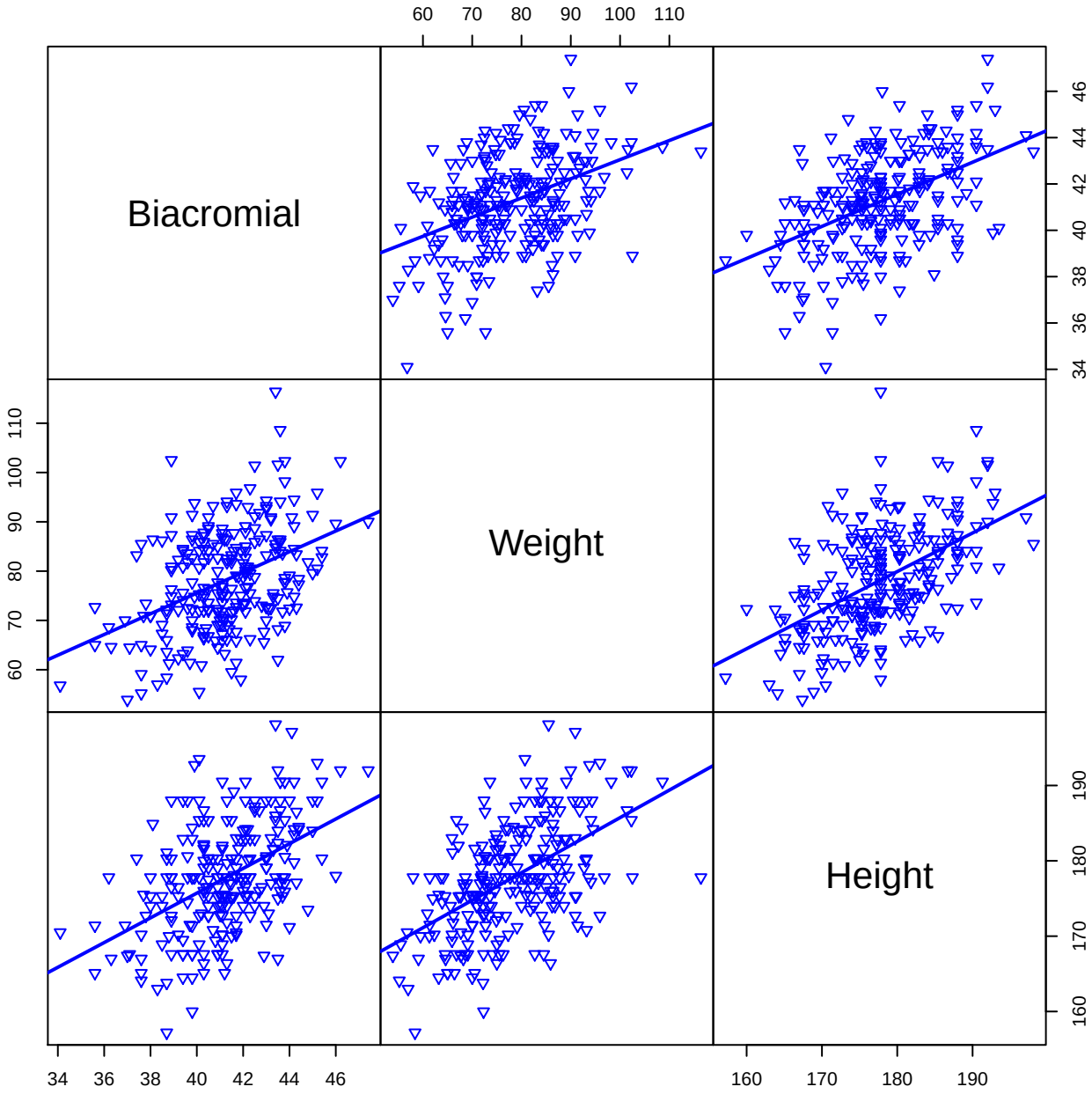




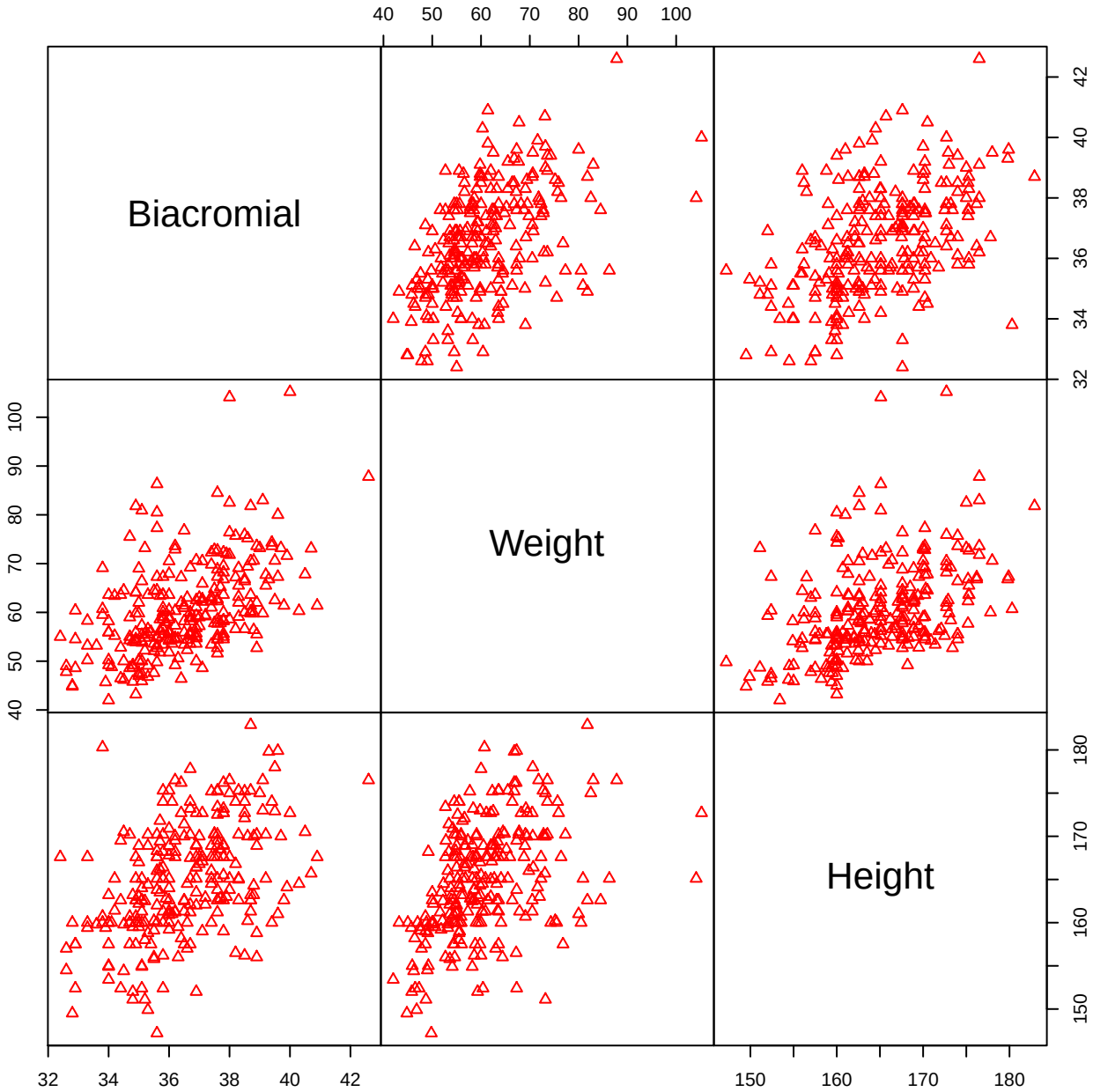
Kombiniert mit Scatterplotmatrix: `Gender = 1`



Kombiniert mit Scatterplotmatrix: `Gender = 1`



Kombiniert mit Scatterplotmatrix: `Gender = 0`



Kombiniert mit Scatterplotmatrix: `Gender = 0`