Tutorium Datenauswertung

Darstellung metrischer Daten

Agenda

- Modalwert, Mittelwert, Median
- Spannweite, Interquartilsabstand, Varianz, Standardabweichung
- Schiefe, Exzess
- Darstellungen in R
  - Tabelle der Verteilungsparameter
  - Histogramm, Dichteplot
  - Boxplot

Lagemaße

Lagemaße

- Geben an, wo die Mitte der Verteilung liegt → zentrale Tendenz
- Sagen sehr wenig über die Form der Verteilung aus
- Gängige Lagemaße:
  - Mittelwert (arithmetisches Mittel)
  - Modalwert
  - Median

Modalwert

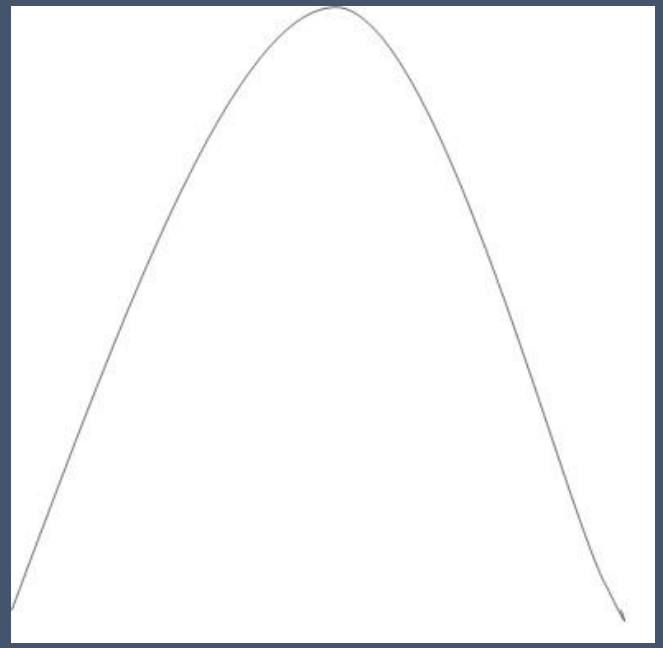
Der Modalwert (engl. „mode“)

- Ist die Ausprägung einer Variable, die am häufigsten vorkommt (nicht die höchste Ausprägung!)
  - Häufigkeitstabelle: höchster Wert in der Prozentspalte
  - Diagramme: höchster Wert auf der y-Achse
- Kann mehrmals in einer Verteilung auftreten
  - Unimodal: ein Modalwert
  - Bimodal: zwei Modalwerte
  - Multimodal: mehr als zwei Modalwerte

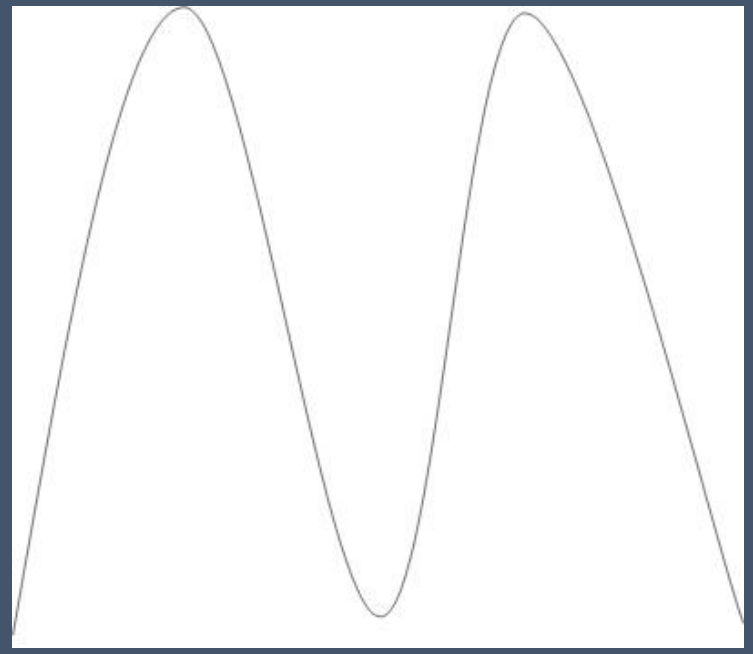
Modalwert – Beispiel

Fall	1	2	3	4	5
Ausprägung	15	0	10	20	15

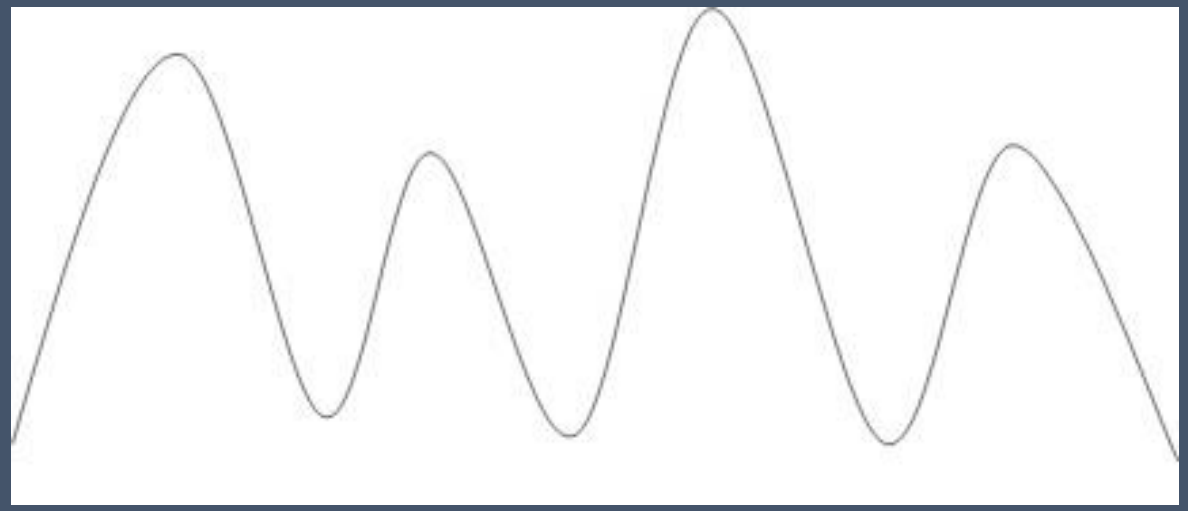
- Der Modalwert ist 15, weil es der Wert ist, der in der Verteilung am häufigsten auftritt
- Die Verteilung ist unimodal (sonst tritt keine Ausprägung 2 Mal auf)



Unimodal



bimodal



multimodal

Mittelwert

Der Mittelwert (engl „mean“)

- Gibt den durchschnittlichen Wert einer Variable an
- Kann schwer interpretierbar sein (Beispiel: 1,34 Kinder)
- Formel:

formal	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$
Nicht-formal	$\text{Mittelwert} = \frac{Ausprägung\ erster\ Fall\ + \cdots +\ Ausprägung\ letzter\ Fall}{Anzahl\ der\ Fälle\ insgesamt}$

Mittelwert - Beispiel

Fall	1	2	3	4	5
Ausprägung	15	0	10	20	15

Formel: $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

Berechnung: $\bar{x} = \frac{15+0+10+20+15}{5} = 12$

Median

Der Median (engl. „median“)

- Gibt den Wert an, der, wenn man die Werte nach Größe ordnet, in der Mitte liegt → ohne Berechnung ablesbar
- Formel für eine geordnete Stichprobe (Wf = Wert für):

	Ungerade Anzahl von Werten	Gerade Anzahl von Werten
formal	$\tilde{x} = x\frac{n+1}{2}$	$\tilde{x} = \frac{1}{2}\left(x\frac{n}{2} + x\frac{n}{2} + 1)\right)$
Nicht-formal	$Median = Wf\frac{Anzahl\ Werte + 1}{2}$	$Median = \frac{1}{2}\left(Wf\frac{Anzahl\ Werte}{2} + Wf\left(\frac{Anzahl\ Werte}{2} + 1\right)\right)$

Median - Beispiel

Unsortiert:

Fall	1	2	3	4	5
Ausprägung	15	0	10	20	15

Der Größe nach sortiert:

Fall (x)	1	2	3	4	5
Ausprägung	0	10	15	15	20

Formel: $\tilde{x} = x\frac{n+1}{2}$

Berechnung: $\tilde{x} = x\frac{5+1}{2} = x3 = 15$

Lagemaße im Vergleich

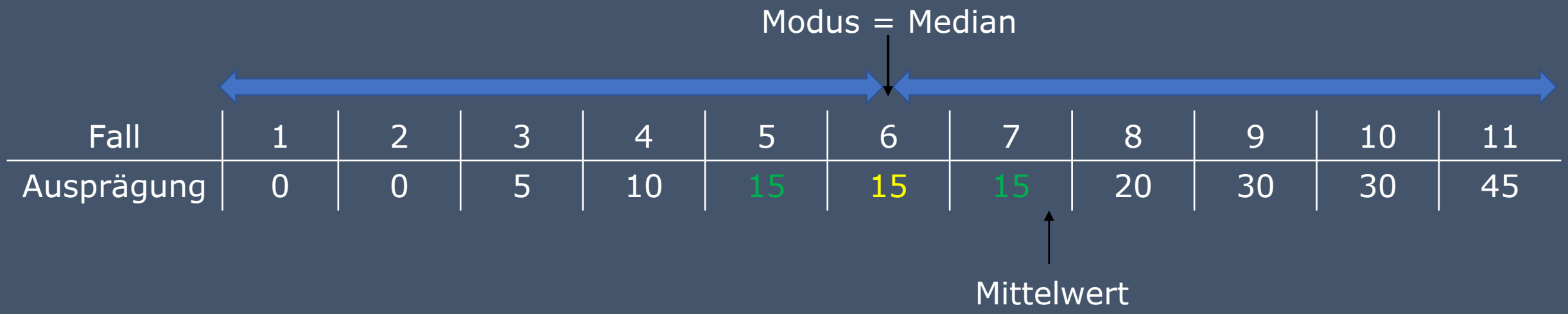
	Modalwert	**Median**	**Mittelwert**
Skalenniveau	ab Nominalskala	ab Ordinalskala	ab Intervallskala
Vorteile	Einfachstes Lagemaß	Gut geeignet für schiefverteilte Variabeln	gut geeignet für normalverteile Variablen
Nachteile	Nur sinnvoll für unimodale Verteilungen	Ungenau bei kleinen Datensätzen	stark durch Ausreißer beeinflusst

Lageparameter – Aufgabe

Aufgabe: Berechnet Modus, Median und Mittelwert für folgende Verteilung:

Fall	1	2	3	4	5	6	7	8	9	10	11
Ausprägung	15	0	10	20	15	5	30	15	45	30	0

Lageparameter – Lösung



Fall	1	2	3	4	5	6	7	8	9	10	11
Ausprägung	0	0	5	10	15	15	15	20	30	30	45

- Modus: 15
- Median: 15
- Mittelwert: $\bar{x} = \frac{15+0+10+20+15+5+30+15+45+30+0}{11} \approx 16{,}8$

Streuungsmaße

Streuungsmaße

- Geben an, wie nah die Werte einer Verteilung aneinander liegen
  - Homogen: Werte liegen nah aneinander
  - Heterogen: Werte liegen weit voneinander weg
- Sollten immer mit einem Lageparameter interpretiert werden
- Gängige Streuungsmaße:
  - Spannweite
  - Interquartilsabstand
  - Varianz, Standardabweichung

Spannweite

Die Spannweite (engl. „range“)

- Gibt den Abstand zwischen dem größten und dem kleinsten Wert einer Variablen an
- Ist extrem ausreißerempfindlich
- Formel:

formal	$R = x_{max} - x_{min}$
Nicht-formal	$Spannweite = größter\ Wert - kleinster\ Wert$

Spannweite - Beispiel

Unsortiert:

Fall	1	2	3	4	5
Ausprägung	15	0	10	20	15

Der Größe nach sortiert:

Fall (x)	1	2	3	4	5
Ausprägung	0	10	15	15	20

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

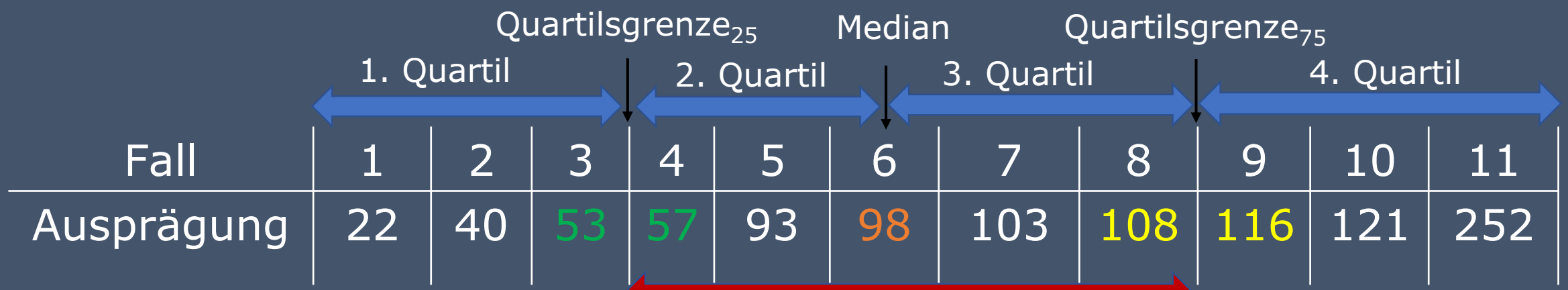
Berechnung: $R = 20 - 0 = 20$

Interquartilsabstand

Der Interquartilsabstand (engl. „interquartile range“)

- gibt die Spannweite zwischen der 25% und der 75% Quartilsgrenze der Verteilung an
- Quartile geben den Wertebereich eines Viertels der nach Größe sortierten Verteilung an
- Quartilsgrenzen
  - sind die äußeren Werte eines Quartils
  - 75% Quartilsgrenze: Mitte vom Median zum größten Wert
  - 25% Quartilsgrenze: Median vom kleinsten Wert zur Mitte
- Formel: $IQA = x_{0,75} - x_{0,25}$

Interquartilsabstand – Beispiel



Fall	1	2	3	4	5	6	7	8	9	10	11
Ausprägung	22	40	53	57	93	98	103	108	116	121	252

Interquartilsabstand

Formel: $IQA = x_{0,75} - x_{0,25}$

$$x_{0,25} = \frac{53 + 57}{2} = 55 \qquad x_{0,75} = \frac{108 + 116}{2} = 112$$

$$\rightarrow IQA = 112 - 55 = 57$$

Varianz

Die Varianz (engl. „variance“)

- Gibt die durchschnittliche Abweichung eines Werts vom Mittelwert zum Quadrat an
- Grund für‘s Quadrieren:
  - Abweichungen nach oben (plus) und nach unten (minus) heben sich auf (0)
  - Idee: $x * x = x^2$ und $(-x) * (-x) = x^2$
- Ist schwer interpretierbar

Standardabweichung

Die Standardabweichung (engl. „standard deviation“)

- Gibt die durchschnittliche Abweichung eines Werts vom Mittelwert an
- Ist das am häufigsten genutzte Streuungsmaß
- Gibt Auskunft darüber, wie gut der Mittelwert die Stichprobe repräsentiert:
  - Geringe Standardabweichung: gute Repräsentation
  - Hohe Standardabweichung: schlechte Repräsentation
- Ist die Wurzel der Varianz → gut interpretierbar!

Varianz und Standardabweichung - Formeln

	Varianz	Standardabweichung
formal	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
Nicht-formal	$Varianz = \frac{(Wert_1 - Mittelwert)^2 + \cdots + (Wert_n - Mittelwert)^2}{Anzahl\ der\ Werte - 1}$	$Standardabweichung = \sqrt{Varianz}$

Varianz und Standardabweichung - Beispiel

Fall	1	2	3	4	5
Ausprägung	15	0	10	20	15

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
15 – 12 = 3	9
0 – 12 = -12	144
10 – 12 = -2	4
20 – 12 = 8	64
15 – 12 = 3	9
	9 + 144 + 4 + 64 + 9 = 230

$$\bar{x} = \frac{15 + 0 + 10 + 20 + 15}{5} = 12$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = \frac{230}{5 - 1} = 57{,}5$$

$$s = \sqrt{s^2} = \sqrt{57{,}5} \approx 7{,}6$$

Hier ist die Summe 0!

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

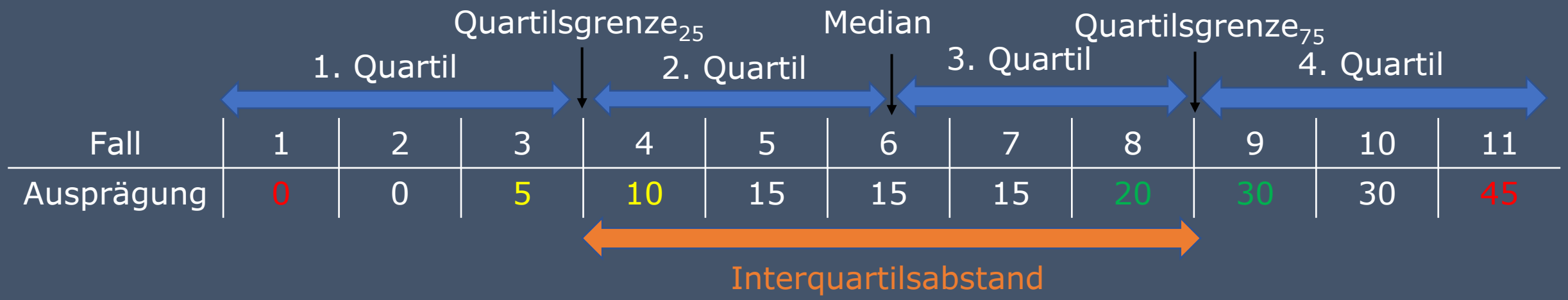
Streuungsmaße – Aufgabe

Aufgabe: Berechnet Spannweite und Interquartilsabstand für die gesamte Verteilung sowie Varianz und Standardabweichung für Fälle 7 bis 11.

Fall	1	2	3	4	5	6	7	8	9	10	11
Ausprägung	15	0	10	20	15	5	30	15	45	30	0

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad R = x_{max} - x_{min} \qquad IQA = x_{0,75} - x_{0,25}$$

Streuungsmaße – Lösung I



$$R = 45 - 0 = 45$$

$$\text{IQA} = \frac{20 + 30}{2} - \frac{5 + 10}{2} = 17{,}5$$

Streuungsmaße – Lösung II

Fall	1	2	3	4	5
Ausprägung	30	15	45	30	0

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
30 – 24 = 6	36
15 – 24 = -9	81
45 – 24 = 21	441
30 – 24 = 6	36
0 – 24 = -24	576
	36 + 81 + 441 + 36 + 576 = 1170

$$\bar{x} = \frac{30 + 15 + 45 + 30 + 0}{5} = 24$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1170}{5-1} = 292{,}5$$

$$s = \sqrt{s^2} = \sqrt{292{,}5} \approx 17{,}1$$

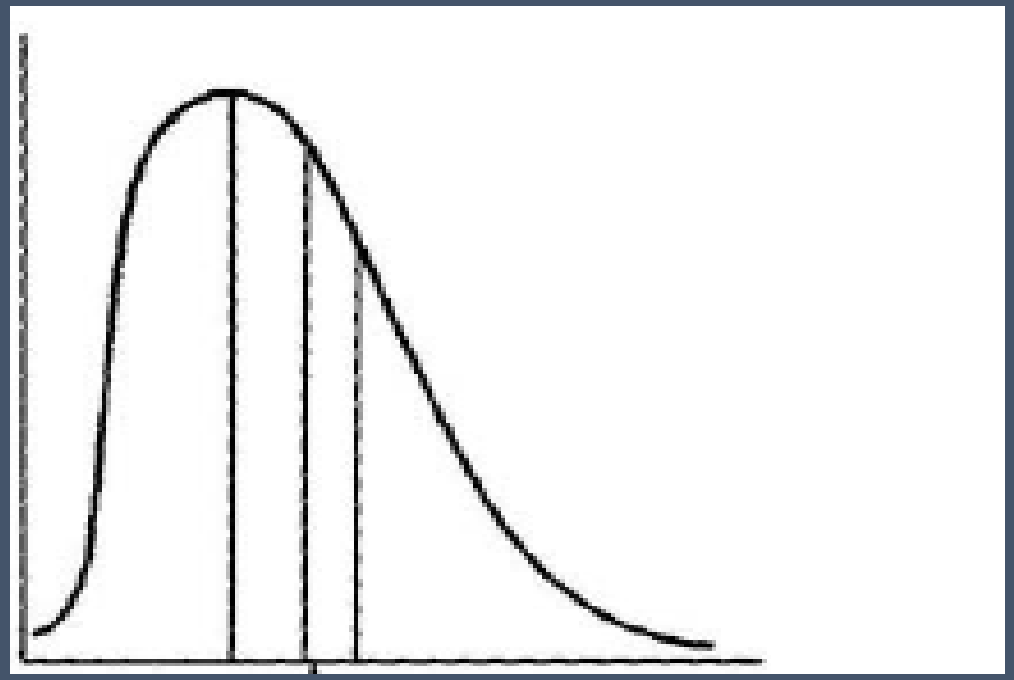
Hier ist die Summe 0!

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

Schiefe und Exzess

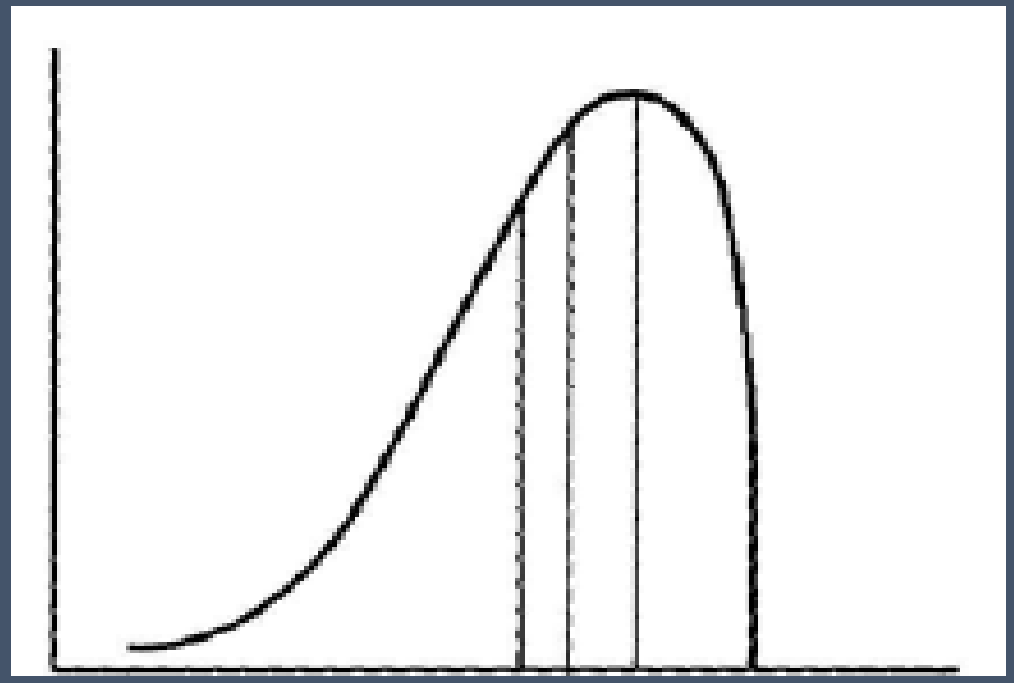
- Schiefe (engl. „skew“):
  - Gibt an, wie symmetrisch die Verteilung ist
  - Interpretation der Werte:
    - 0 → perfekt symmetrisch
    - Positiv → rechtsschief
    - Negativ → linksschief
- Exzess (engl. „kurtosis“)
  - Gibt an, wie flach/steil die Verteilung ist
  - Interpretation der Werte:
    - 0 → so flach wie die Normalverteilung (siehe nächste Sitzung)
    - Positiv → steiler als die Normalverteilung (gestaucht)
    - Negativ → flacher als die Normalverteilung (gestreckt)

Schiefe



Rechtsschief:

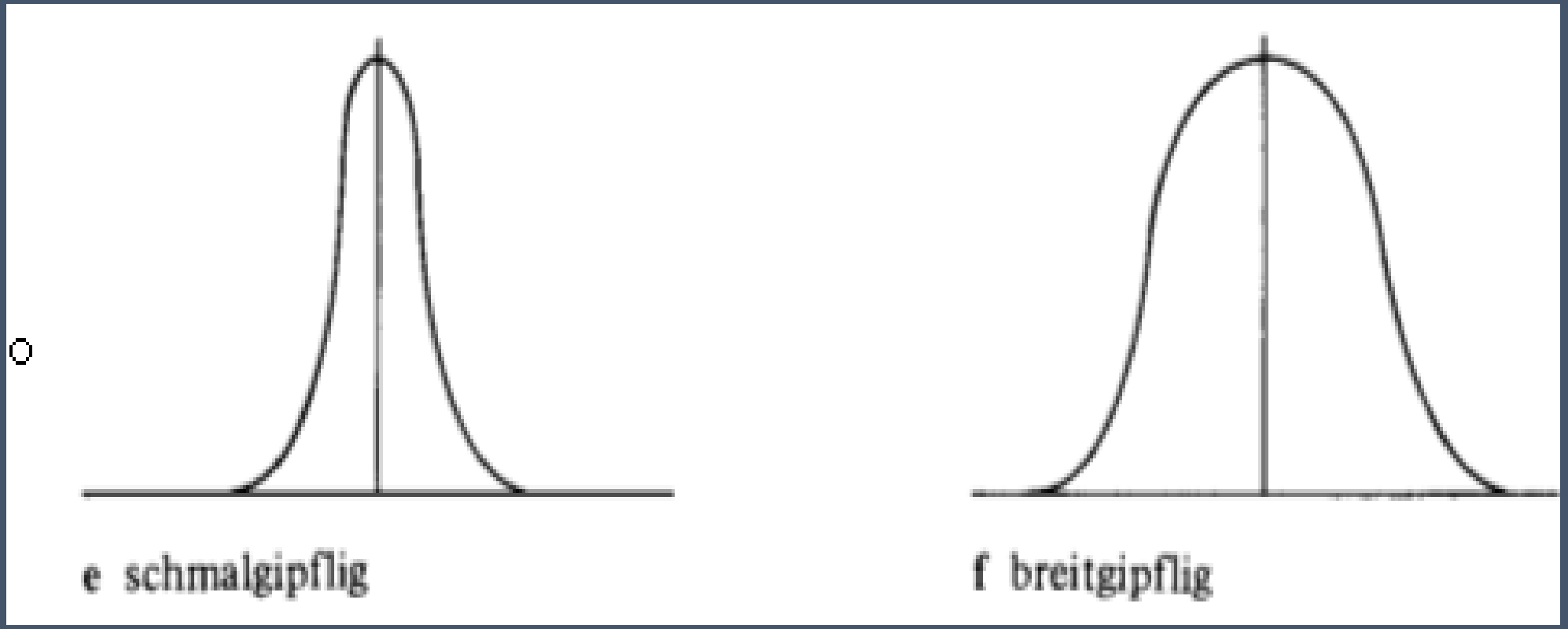
- Positiver Schiefekoeffizient
- Modus < Median < Mittelwert



Linksschief:

- Negativer Schiefekoeffizient
- Mittelwert < Median < Modus

Exzess



gestaucht: positiver Exzess

gestreckt: negativer Exzess

Beschreibung metrischer Daten in R

Lagemaß	**Funktion in R**
Modalwert	names(which.max(table(variable)))
Mittelwert	mean(variable)
Median	median(variable)

Streuungsmaß	**Funktion in R**
Spannweite	range(variable)
Interquartilsabstand	iqr(variable) [nur mit mosaic!]
Varianz	var(variable)
Standardabweichung	sd(variable)

→ Bei fehlenden Werten: Argument „na.rm=T“ hinzufügen!

Tabelle der Verteilungsparameter in R

```
#Laden der Daten und Pakete
library(knitr)
library(mosaic)
load("daten_x.RData")
datensatz <- daten_x

#Gesamttabelle

kable(round(rbind("name variable_1" =
favstats(datensatz$variable_1), "name variable_2" =
favstats(datensatz$variable_2)), 2))
```

Tabelle der Verteilungsparameter in R - Beispiel

```
#Laden der Daten und Pakete
library(knitr)
library(mosaic)
load("Beispieldatensatz.RData")
datensatz <- beispiel

#Gesamttabelle

kable(round(rbind("TV-Konsum" = favstats(datensatz$Minuten.TV.Konsum),

"Internetkonsum" = favstats(datensatz$Minuten.Internetkonsum)), 2))
```

	min	Q1	median	Q3	max	mean	sd	n	missing
TV-Konsum	0	30	90	150	240	95.45	80.29	11	0
Internetkonsum	60	150	180	270	360	196.36	97.08	11	0

Histogramme in R

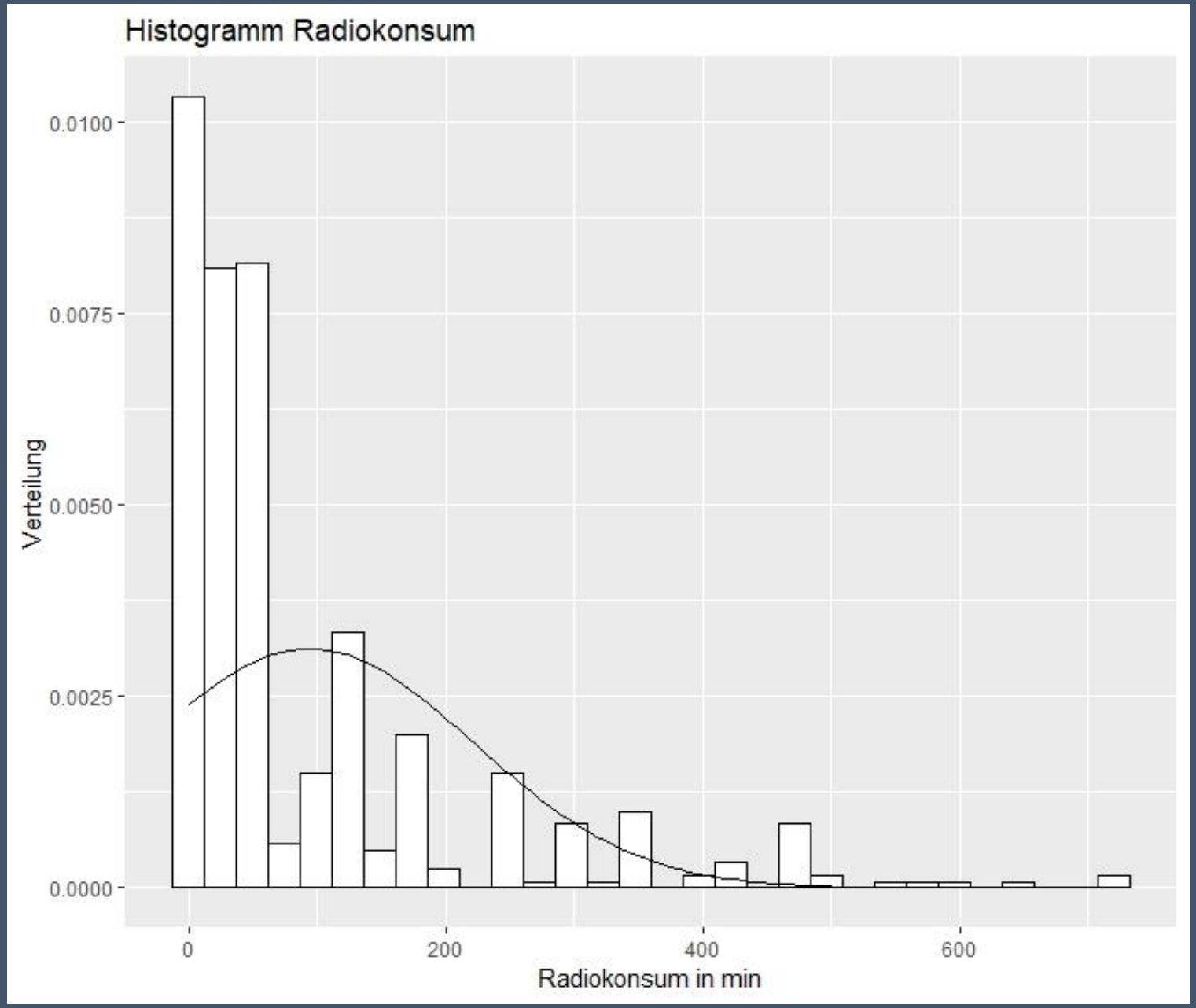
```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_x.RData")
datensatz <- daten_x

#Erstellen eines Histogramms mit eingezeichneter Normalverteilungskurve
ggplot(datensatz, aes(x = variable)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(title = "überschrift", x = "beschriftung x-Achse", y = "beschriftung y-Achse") +
  stat_function(fun = dnorm, args = list(mean = mean(datensatz$variable, na.rm = T), sd =
sd(datensatz$variable, na.rm = T)), colour = "black", size = 0.5)
```

optional (zeichnet Normalverteilungskurve ein)

Histogramme in R – Beispiel

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_2019.RData")
datensatz <- daten_2019

#Erstellen eines Histogramms mit eingezeichneter Normalverteilungskurve
ggplot(datensatz, aes(x = radio_minuten)) +
  geom_histogram(aes(y = ..density..), colour = "black", fill = "white") +
  labs(title = "Histogramm Radiokonsum", x = "Radiokonsum in min", y = "Verteilung") +
  stat_function(fun = dnorm, args = list(mean = mean(datensatz$radio_minuten, na.rm = T), sd =
sd(datensatz$radio_minuten, na.rm = T)), colour = "black", size = 0.5)
```

optional (zeichnet Normalverteilungskurve ein)

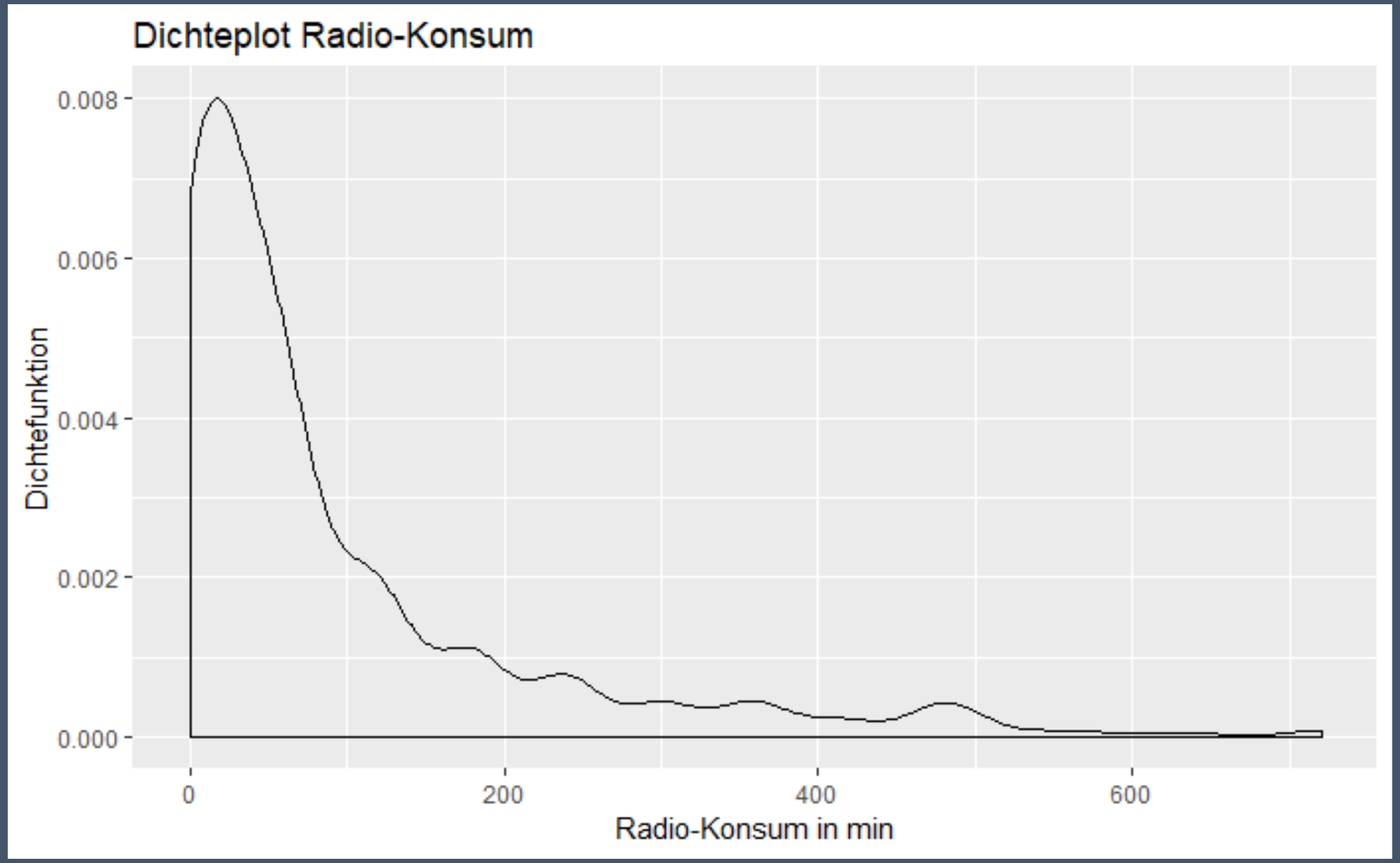

Histogramm Radiokonsum
0.0100
0.0075
0.0050
0.0025
0.0000
Verteilung
0
200
400
600
Radiokonsum in min

Dichteplots in R

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_x.RData")
datensatz <- daten_x

#Erstellen des Dichteplots
ggplot(datensatz, aes(x = variable)) +
  geom_density() +
  labs(title = "überschrift", x = "beschriftung x-Achse", y =
"beschriftung y-Achse")
```

Dichteplots in R

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_2019.RData")
datensatz <- daten_2019

#Erstellen des Dichteplots
ggplot(datensatz, aes(x = radio_minuten)) +
  geom_density() +
  labs(title = "Dichteplot Radiokonsum", x = "Radiokonsum in Minuten", y = "Dichte")
```


Dichteplot Radio-Konsum
0.008
0.006
0.004
0.002
0.000
Dichtefunktion
0
200
400
600
Radio-Konsum in min

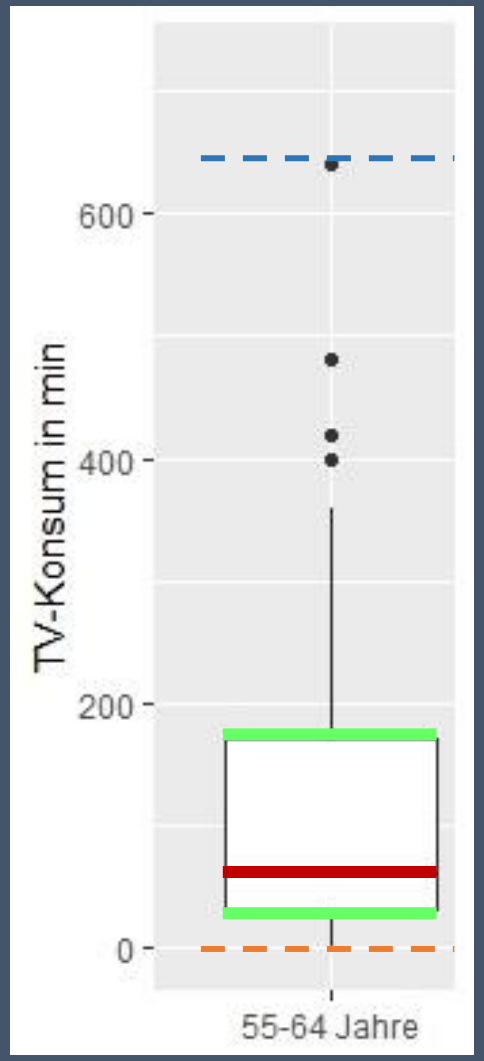
Boxplots – Erklärung

Parameter für den Radiokonsum von 55-64-Jährigen:

Min.	Q25	Median	Mean	Q75	Max.	NA
0	30	60	127,9	170	640	3

Die Striche, die im rechten Winkel zur Mitte der Box nach oben und nach unten gehen, nennen wir Whisker. Sie entsprechen in R 1,5 mal dem Interquartilsabstand (also der Länge der Box).

Die schwarzen Punkte über den Whiskern nennen wir Ausreißer. Sie stellen Werte dar, die über 1,5 mal dem Interquartilsabstand liegen. Je mehr Ausreißer, desto heterogener die Verteilung.

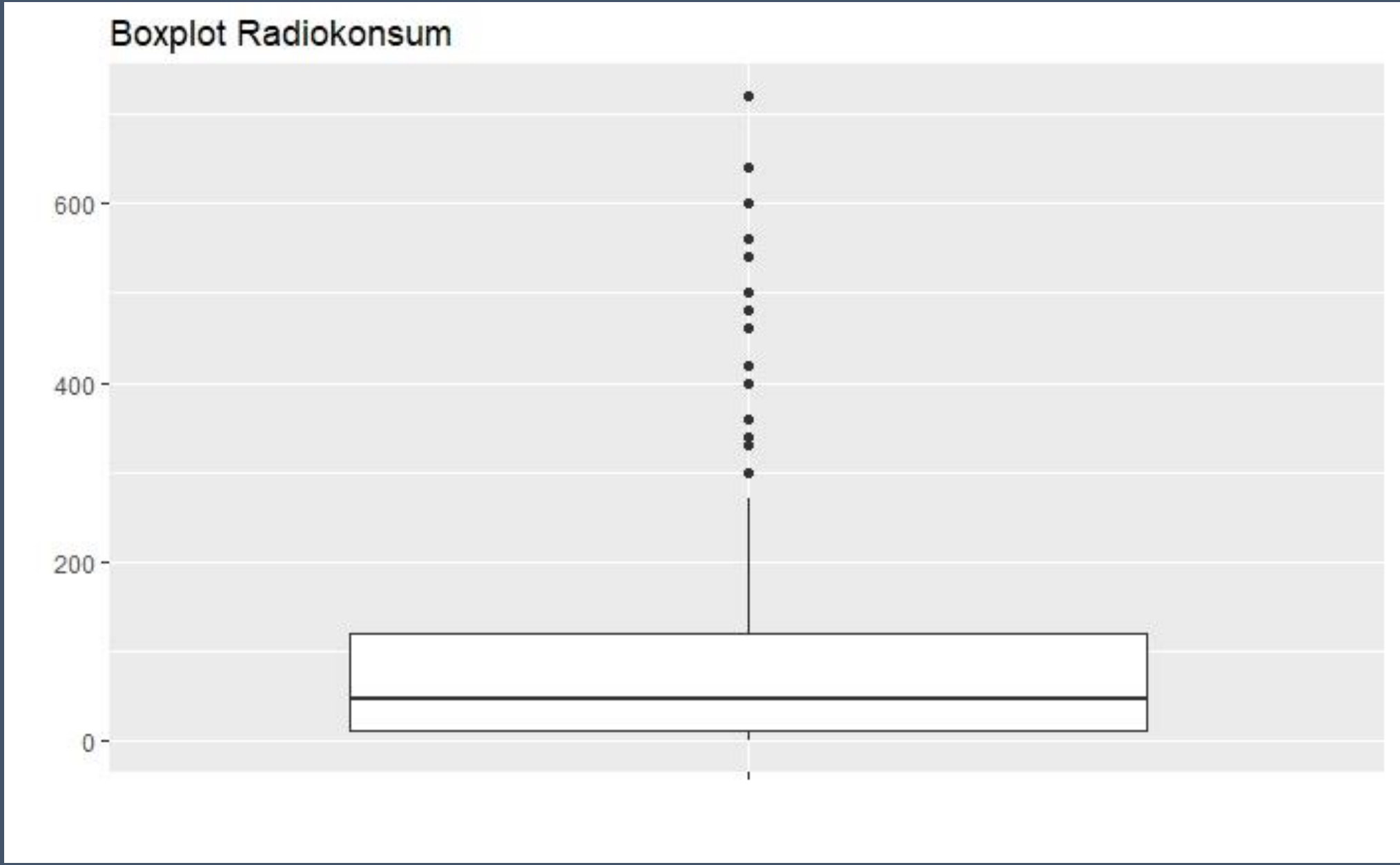


Ungruppierte Boxplots in R

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_x.RData")
datensatz <- daten_x

#Erstellen des Boxplots
ggplot(datensatz, aes(x = "", y = variable)) +
geom_boxplot() + labs(title = "überschrift", x="", y="")
```

Ungruppierte Boxplots in R - Beispiel

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_2019.RData")
datensatz <- daten_2019

#Erstellen des Boxplots
ggplot(datensatz, aes(x = "", y = radio_minuten)) +
geom_boxplot() + labs(title = "Boxplot Radiokonsum", x="",
y="")
```

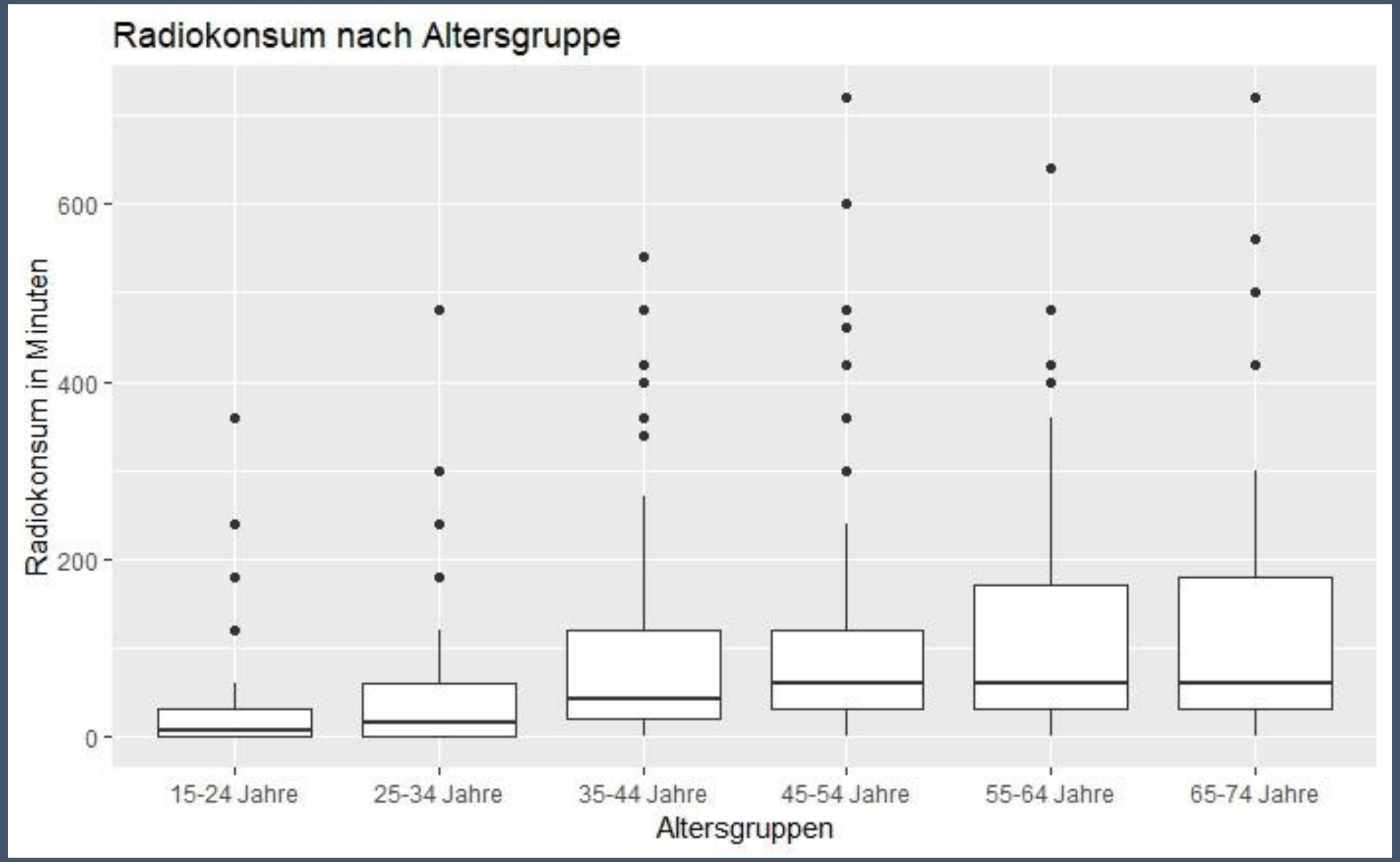

Boxplot Radiokonsum
600
400
200
0

Gruppierte Boxplots in R

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_x.RData")
datensatz <- subset(daten_x, Bedingung)

#Erstellen des Boxplots
ggplot(datensatz, aes(x = factor(UV), y = AV)) +
  geom_boxplot() +
  labs(title = "überschrift", x = "beschriftung x-Achse", y = "beschriftung
y-Achse")
```

Gruppierte Boxplots in R – Beispiel

```
#Laden der Daten und Pakete
library(ggplot2)
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr>2000)

#Erstellen des Boxplots

ggplot(datensatz, aes(x = factor(altersgruppe), y = radio_minuten)) +
geom_boxplot() + labs(title = "Radiokonsum nach Altersgruppe", x =
"Altersgruppen", y = "Radiokonsum in Minuten")
```


Radiokonsum nach Altersgruppe
Radiokonsum in Minuten
600
400
200
0
15-24 Jahre
25-34 Jahre
35-44 Jahre
45-54 Jahre
55-64 Jahre
65-74 Jahre
Altersgruppen

FRAGEN?

Selbststudium

- Forschungsbericht (EFB und ZFB):
  - Tabelle der Verteilungsparameter, Histogramm, Dichteplot, Boxplot für daten_2019$www_minuten erstellen
  - Genereller Text zur Verteilung der Variable daten_2019$www_minuten (siehe Gehraus Folien)
- Wiederholen:
  - Modalwert, Mittelwert, Median
  - Spannweite, IQA, Varianz, Standardabweichung
  - Hypothesen (!)

Bis nächste Woche!