Deskriptive Statistik

H. Wuschke

14. August 2019

Ziele der Sitzung

- verschiedene Typen von Daten und ihre Skalen beurteilen können
- Daten in passenden Diagrammen veranschaulichen können
- Kenngrößen von statistischen Daten beschreiben und interpretieren können
- Boxplots interpretieren und darstellen können

In der deskriptiven (beschreibenden) Statistik erhebt man Daten und wertet diese aus, um Erkenntnisse über Zusammenhänge zu generieren.

Grundbegriffe der Statistik

Daten sind meist in einer **Urliste** gegeben und werden dann in einer **Strichliste** geordnet. Daraus können dann die **absolute Häufigkeiten** (ausgezählte Zahlen) und die **relativen Häufigkeiten** (Anteil an der Gesamtheit) gebildet werden.

In der Klasse 6B hat Herr Wuschke die Lieblingstiere erfragt. Dabei ist folgendes Ergebnis herausgekommen:

Lieblingstiere der 6B

Lieblingstier	Strichliste	abs. Hfkt.	rel. Hfkt.
Kaninchen	|	1	$\frac{1}{24} \approx 0{,}0417$
Pferd	||||	4	$\frac{4}{24} \approx 0{,}1667$
Hund	$\cancel{\text{\|\|\|\|}}$|	6	$\frac{6}{24} = 0{,}25$
Einhorn	$\cancel{\text{\|\|\|\|}}$|||	8	$\frac{8}{24} \approx 0{,}3333$
Katze	|||	3	$\frac{3}{24} = 0{,}125$

Beispiel – Besitz von Mediengeräte

Bei der JIM-Studie[a] wurden Jugendliche von 12 Jahren bis 19 Jahren gefragt, welche Mediengeräte sie besitzen. Hier ist ein Beispiel von 5 Jugendlichen.

Urliste:
Smartphone; Laptop; Spielekonsole; Tablet; Smartphone; Fernsehgerät; Smartphone; Laptop; iPod; Smartphone; Spielekonsole; Tablet; Fernsehgerät; Tablet; Streamingbox; Smartphone; Laptop; Spielekonsole; iPod; Tablet

Gerät	Smartph.	Laptop	Konsole	Tablet	TV	SBox	iPod
Strichliste	𝍸	𝍷	𝍷	𝍸	𝍵	𝍶	𝍵
abs. Hfkt.	5	3	3	4	2	1	2

[a] Jährlich durchgeführte Studie „Jugend, Information, (Multi-)Media“

Skalenniveaus

Daten werden in drei unterschiedliche Skalen eingeteilt:

Nominalskala – Ordinalskala – metrische Skala

- **Nominalskala** – Daten können nur anhand unterschiedlicher Namen angegeben werden (z. B. Haarfarben, Geschlecht, Krankheiten, ...)
- **Ordinalskala** – Daten können geordnet werden unter bestimmten Gesichtspunkten (z. B. bevorzugte Produkte nach dem Motto: „*Ich würde eher ... kaufen, als ...*“)
- **metrische Skala** – Daten können geordnet werden und in sinnvollen Zahlen ausgedrückt werden (z. B. Daten aus den Naturwissenschaften, Daten aus dem Sport, ... nach dem Motto: „*Das kann ich messen.*“)

Bei metrischen Daten ist es häufig sinnvoll, diese zu klassieren. Dabei werden sie in selbstgewählten oder vorgegebenen Klassen zusammengefasst.

Beispiel – Gewicht von Säuglingen im Nordkurier 2018

Im Nordkurier erscheinen wöchentlich Informationen über Neugeborene. Das Geburtsgewicht der letzten Zeit war folgendermaßen: 3280g, 3165g, 3055g, 2850g, 3100g, 4190g, 3730g, 3540g, 3220g, 2774g, 2955g, 3280g, 2760g, 3570g, 3860g, 2915g, 3720g, 3500g, 3550g, 2760g, 4440g, 3720g, 3630g, 3450g

Gewicht x	abs. Hfkt.	Gewicht x	abs. Hfkt.
$2500g \leqq x < 3000g$	6	$3500g \leqq x < 4000g$	9
$3000g \leqq x < 3500g$	7	$4000g \leqq x < 4500$	2

Diagrammtypen

- Für die Darstellung von absoluten Häufigkeiten nutzt man ein **Säulendiagramm** oder **Balkendiagramm**. Wenn die x-Achse dabei geordnet ist (z. B. bei Jahreszahlen), kann man auch ein **Liniendiagramm** nutzen.
- Für die Darstellung von relativen Häufigkeiten nutzt man ein **Kreisdiagramm** oder **Streifendiagramm**.
  Diese Diagramme dürfen nur verwendet werden, wenn **keine Mehrfachnennung** möglich war.
- Klassierte Daten werden in einem **Histogramm** dargestellt. Dies ist eine besondere Form des Säulendiagramms.
- Wenn die statistischen Kenngrößen berechnet wurden, können die Daten auch mithilfe eines **Boxplots** dargestellt werden.

Lieblingstiere der 6B

Kaninchen: 1; Pferd: 4; Hund: 6; Einhorn: 8; Katze: 3

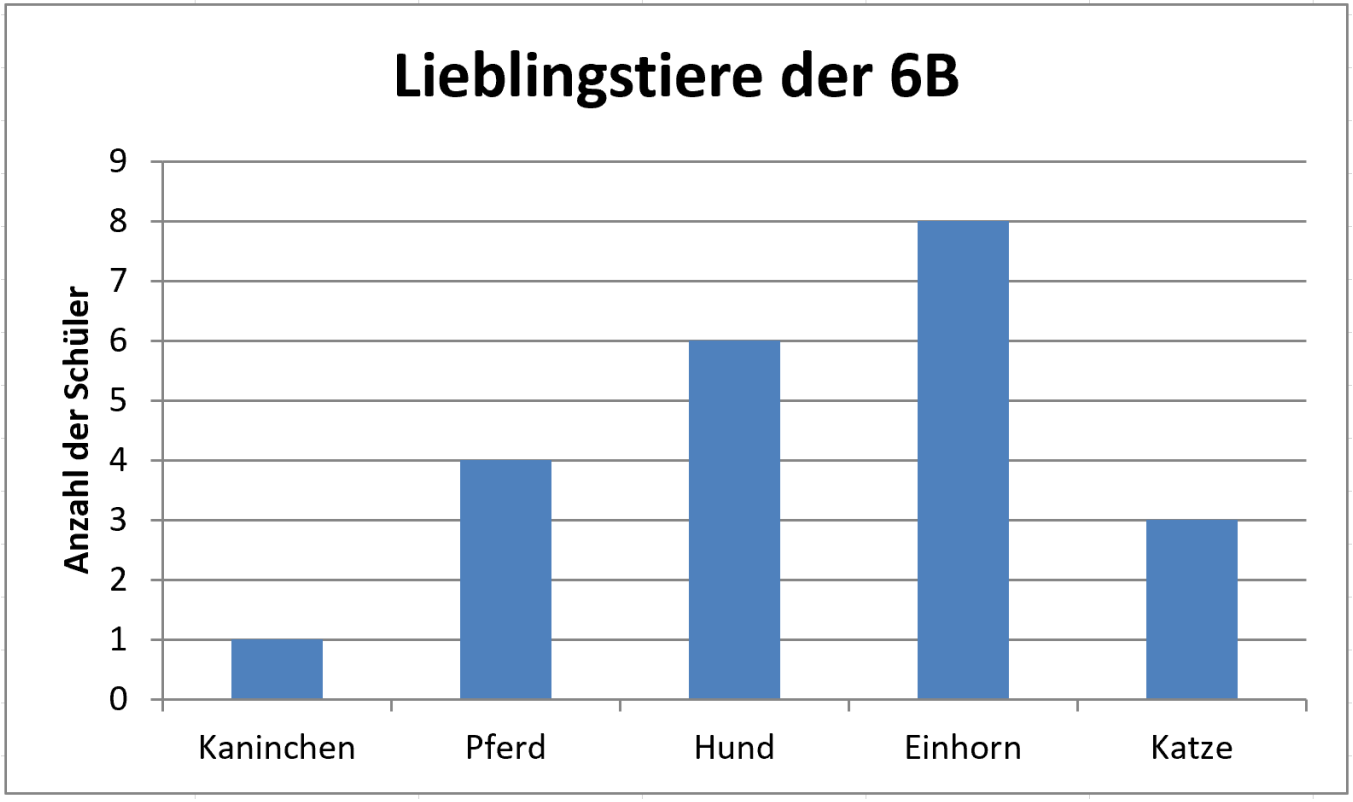

Abbildung: Säulendiagramm [Felix Röhricht 2018, Excel]

Lieblingstiere der 6B

Kaninchen: 1; Pferd: 4; Hund: 6; Einhorn: 8; Katze: 3

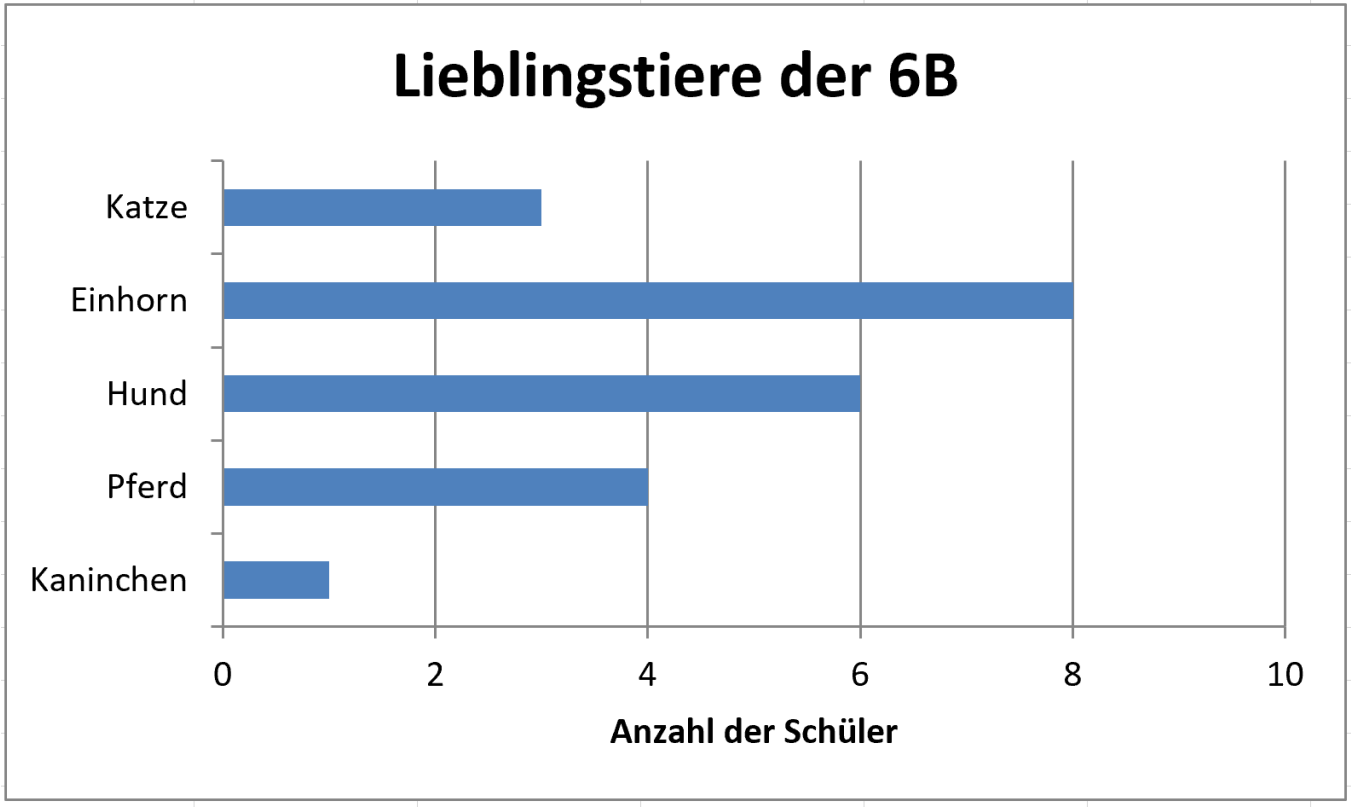

Abbildung: Balkendiagramm [Felix Röhricht 2018, Excel]

Lieblingstiere der 6B

Kaninchen: 1; Pferd: 4; Hund: 6; Einhorn: 8; Katze: 3

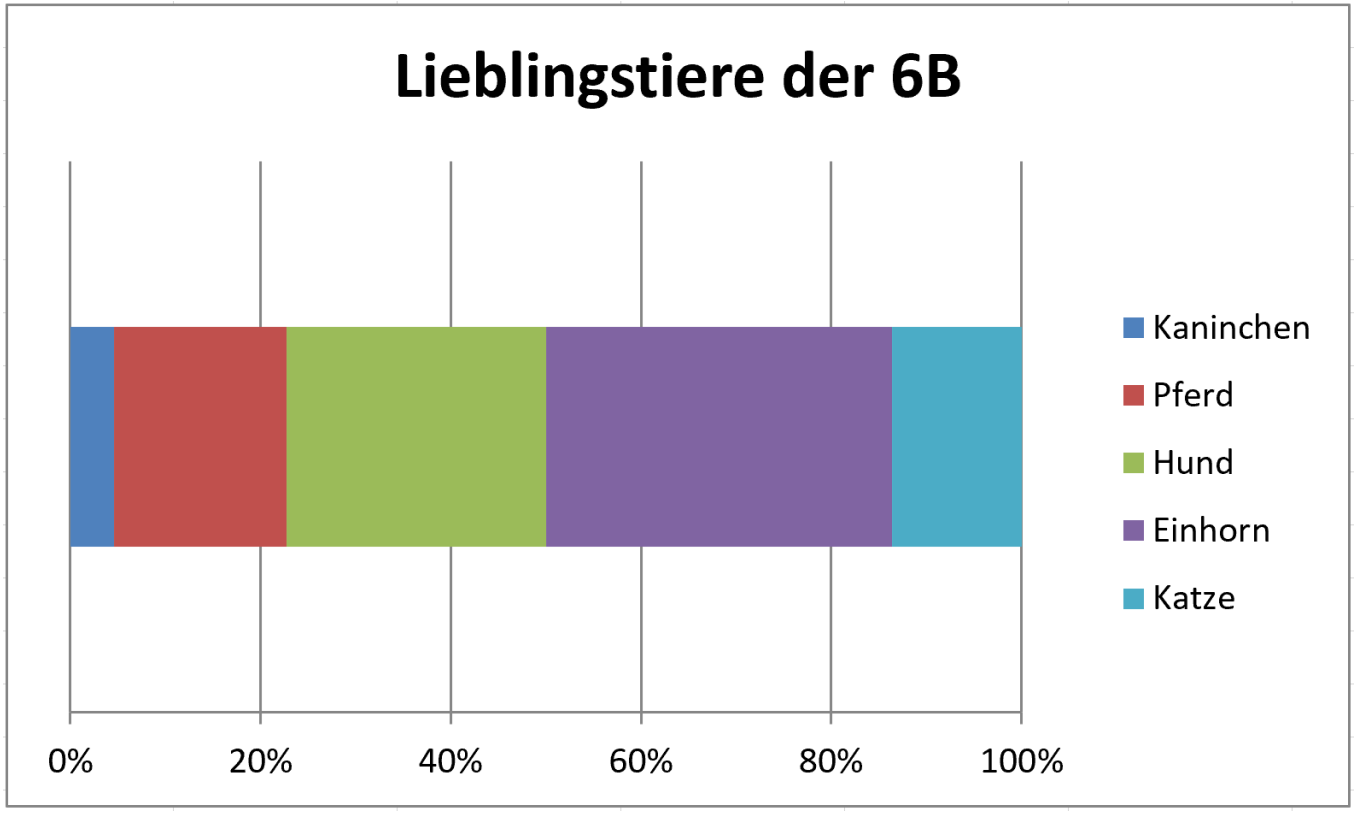

Abbildung: Streifendiagramm [Felix Röhricht 2018, Excel]

Lieblingstiere der 6B

Kaninchen: 1; Pferd: 4; Hund: 6; Einhorn: 8; Katze: 3

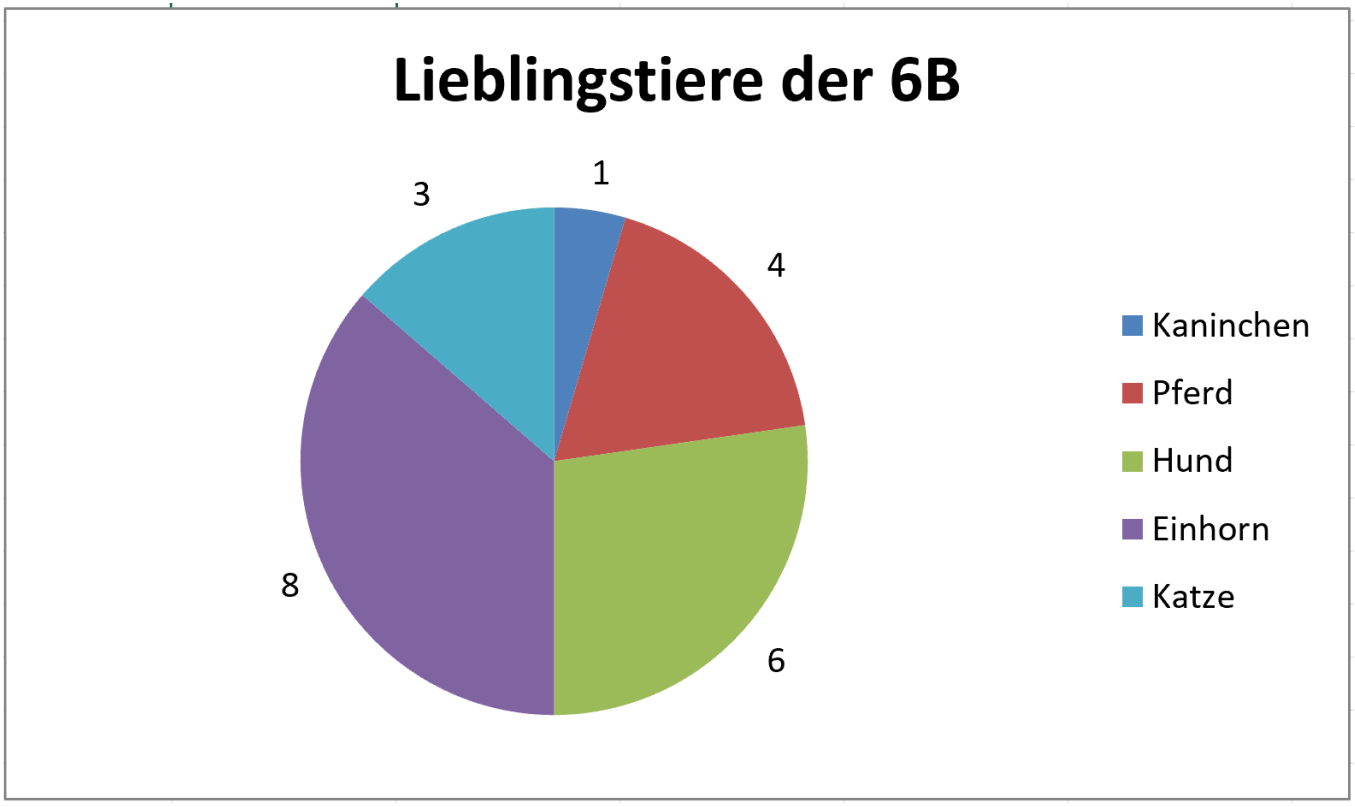

Abbildung: Kreisdiagramm [Felix Röhricht 2018, Excel]

Lieblingstiere der 6B

Kaninchen: 1; Pferd: 4; Hund: 6; Einhorn: 8; Katze: 3

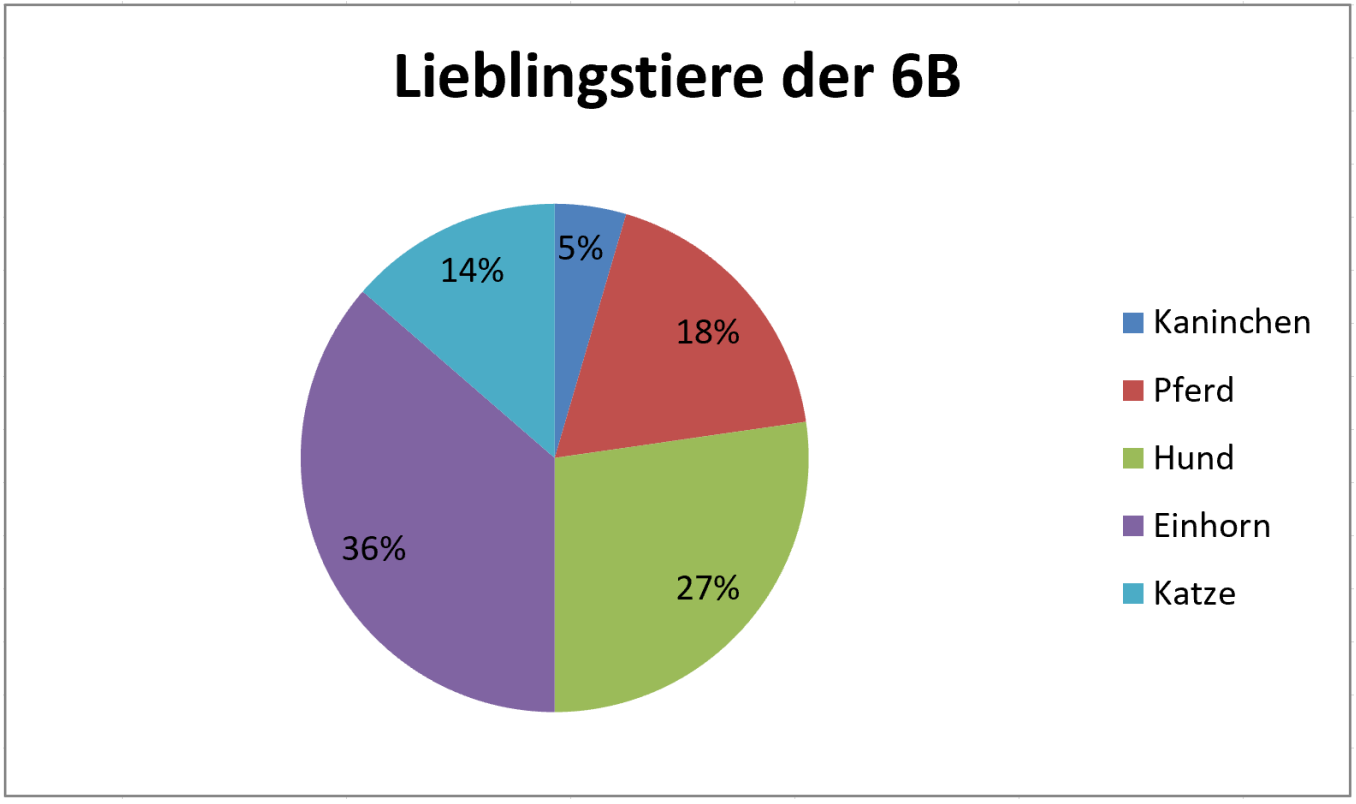

Abbildung: Kreisdiagramm [Felix Röhricht 2018, Excel]

Gewichtsklassen von Säuglingen aus dem Nordkurier

Gewicht x	abs. Hfkt.
$2500g \leqq x < 3000g$	6
$3000g \leqq x < 3500g$	7
$3500g \leqq x < 4000g$	9
$4000g \leqq x < 4500$	2

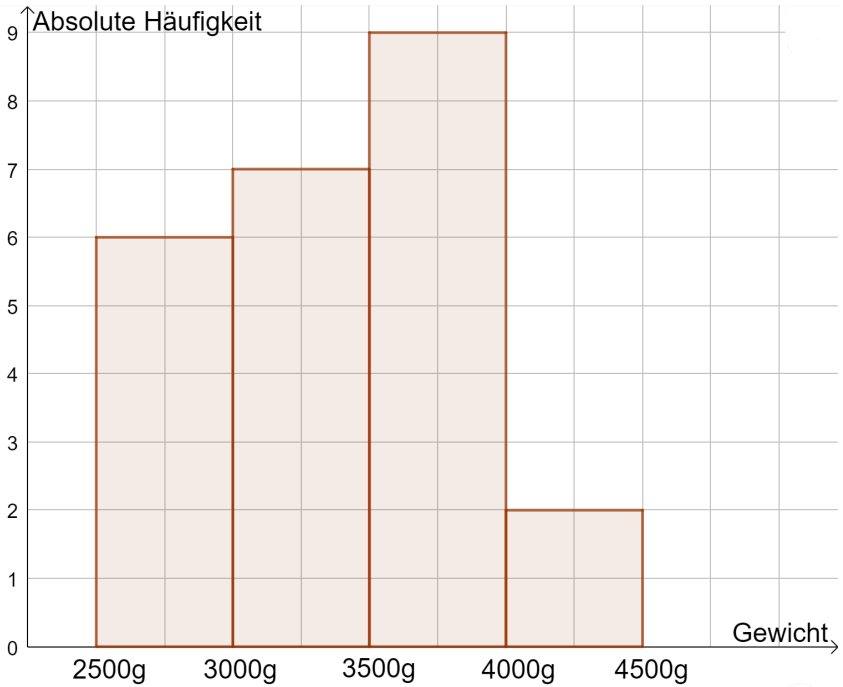

Abbildung: Histogramm [HW 2018, GeoGebra]

Je besser die Skala ist, umso mehr statistische Kenngrößen können erhoben werden. Bei einer Nominalskala kann man den **Modalwert** angeben.

Modalwert

Der **Modalwert** x_{mod} ist die Größe, welche am häufigsten vorkommt. Es kann mehre Modalwerte geben.

Netflix-Konsum

Hannah erzählt über ihren Netflix-Konsum am Wochenende: „Ich habe alle 8 neuen Folgen *Haus des Geldes* geschaut, dann noch 3 Folgen von *Dark* und zwischendurch 5 Folgen *Modern Family*."

Der Modalwert ist in diesem Beispiel $x_{mod} =$ *Haus des Geldes*

Schlafzeiten am Tag

Frank ist in letzter Zeit häufig erschöpft. Deshalb hat er seine tägliche Stundenanzahl Schlaf in zwei Wochen notiert:

Mo	Di	Mi	Do	Fr	Sa	So
6h	4,5 h	7 h	6,5 h	5 h	9 h	10 h
7 h	6,5 h	6,5 h	5 h	5,5 h	7 h	10 h

Die Modalwerte sind $x_{mod1} = 6,5$ h und $x_{mod2} = 7$ h.

Da die Daten bei der Ordinalskala geordnet werden können, ist es möglich einen Zentralwert zu bestimmen, den **Median**.

Median

Der **Median** (Zentralwert) x_{med} ist die Größe, welche in der geordneten Datenreihe in der Mitte steht.
Ist die Anzahl der Daten gerade, kann man den Median nicht immer bestimmen. Für metrische Größen (Zahlen) gilt bei einer Datenreihe von n Daten:

$$x_{med} = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$$

Beispiel Median

In einem Freizeitpark können bei der Auswertung verschiedene Zufriedenheitsstufen angegeben werden:

15 Personen haben folgendermaßen bewertet:

Ordnen der Ergebnisse führt zu:

$x_{mod} =$ $\qquad x_{med} =$

Schlafzeiten von Frank – Beispiel oben

Ordnet man die Schlafzeiten von Frank, erhält man:

4,5 h; 5 h; 5 h; 5,5 h; 6 h; 6,5 h; 6,5 h;
6,5 h; 7 h; 7 h; 7 h; 9 h; 10 h; 10 h

$$x_{med} = \frac{6,5h + 6,5h}{2} = 6,5h$$

Maximum, Minimum, Spannweite

Bei einer Ordinalskala kann man das **Maximum** und das **Minimum** angeben. Dies ist eindeutig, wenn Zahlenwerte vorliegen, ansonsten ist es abhängig von der Betrachtung.

Der Abstand zwischen Maximum und Minimum ist die **Spannweite**, es gilt also: Spannweite = Maximum – Minimum

Schlafzeiten von Frank

Mo	Di	Mi	Do	Fr	Sa	So
6h	4,5 h	7 h	6,5 h	5 h	9 h	10 h
7 h	6,5 h	6,5 h	5 h	5,5 h	7 h	10 h

Maximum = 10 h; Minimum = 4,5 h; Spannweite = 5,5 h

Quartile für die Boxplots

unteres/oberes Quartil

Bildet man wiederum den Median zwischen dem kleinsten Wert und dem Median[a], so erhält man das **untere Quartil** x_{Q1}.
Bildet man den Median zwischen dem größten Wert und dem Median, so erhält man das **obere Quartil** x_{Q3}.
Während der Median die Hälfte der Stichprobe markiert, zeigt x_{Q1} ein Viertel der Stichprobe und x_{Q3} drei Viertel der Stichprobe an.

[a] Bei einer geraden Anzahl an Daten, nutzt man die Daten, aus denen der Median gebildet wurde.

Schlafzeiten von Frank – Beispiel oben

4,5 h; 5 h; 5 h; 5,5 h; 6 h; 6,5 h; 6,5 h; 6,5 h; 7 h; 7 h; 7 h; 9 h; 10 h; 10 h

$x_{med} = 6,5h$

x_{Q1} wird zwischen 4,5 h und 6,5 h gebildet: $x_{Q1} = 5,5h$

x_{Q3} wird zwischen 6,5 h und 10 h gebildet: $x_{Q3} = 7h$

geordnete Geburtsgewichte – Beispiel oben

2760g; 2760g; 2774g; 2850g; 2915g; 2955g; 3055g; 3100g;
3165g; 3220g; 3280g; 3280g; 3450g; 3500g; 3540g; 3550g;
3570g; 3630g; 3720g; 3720g; 3730g; 3860g; 4190g; 4440g

$$x_{med} = \frac{3280g + 3450g}{2} = 3365g$$

Das untere Quartil wird zwischen 2760g und 3280g gebildet und das obere zwischen 3450g und 4440g.

$$x_{Q1} = \frac{2955g + 3055g}{2} = 3005g \qquad x_{Q3} = \frac{3630g + 3720g}{2} = 3675g$$

arithmetisches Mittel

arithmetisches Mittel

Das **arithmetische Mittel** (Durchschnitt) berechnet sich bei n metrischen Daten:

$$\overline{x} = \frac{x_1 + x_2 + \ldots x_n}{n}$$

Arithmetisches Mittel in den Beispielen

Bei den Geburtsgewichten ist

$$\overline{x} = \frac{2 \cdot 2760g + 2774g + \cdots + 4440g}{24} = \frac{81814g}{24} \approx 3375,58g$$

Bei Franks Schlafzeiten ist

$$\overline{x} = \frac{4,5h + 2 \cdot 5h + \cdots + 2 \cdot 10h}{14} = \frac{95,5h}{14} \approx 6,82h$$

arithmetisches Mittel bei klassierten Daten

Das arithmetische Mittel von klassierten Daten bestimmt man ähnlich, nur dass man die Klassenmitten zur Berechnung nutzt.

Durchschnittliches Gewicht der Säuglinge

Gewicht x	Klassenmitte	abs. Hfkt.
$2500g \leqq x < 3000g$	2750g	6
$3000g \leqq x < 3500g$	3250g	7
$3500g \leqq x < 4000g$	3750g	9
$4000g \leqq x < 4500g$	4250g	2

$$\begin{aligned}\overline{x} &= \frac{6 \cdot 2750g + 7 \cdot 3250g + 9 \cdot 3750g + 2 \cdot 4250g}{24} \\ &= \frac{81500g}{24} \approx 3395,83g\end{aligned}$$

Zusammenhang Median und arithmetisches Mittel

statistische Ausreißer

Das arithmetische Mittel wird von extremen Daten stark beeinflusst. Der Median gibt einen realistischen zentralen Wert der Datenmenge an. Wenn das arithmetische Mittel und der Median stark voneinander abweichen, dann ist in der Stichprobe ein **Ausreißer** enthalten.

Gehalt in der Firma „Knausrig“

Chef: „*Bei mir verdient jeder im Schnitt 1.500€, es geht meinen Angestellten sehr gut!*“

Personal	Gehalt	Personal	Gehalt
Praktikant	200€	Reinigungskraft	450€
Werksstudentin	800€	Aushilfe	450€
Chef	6.000€	Mitarbeiter	1.100€

empirische Varianz

Die **empirische Varianz** σ^2 ist die mittlere quadratische Abweichung vom arithmetischen Mittel.

$$\sigma^2 = \frac{(\overline{x} - x_1)^2 + (\overline{x} - x_2)^2 + \cdots + (\overline{x} - x_n)^2}{n}$$

empirische Standardabweichung

Die **empirische Standardabweichung** σ macht die Quadrierung in der Varianz wieder rückgängig und gibt daher an, wie stark die Daten vom arithmetischen Mittel abweichen. σ heißt auch Streumaß, weil es beschreibt, wie stark die Stichprobe streut.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(\overline{x} - x_1)^2 + (\overline{x} - x_2)^2 + \cdots + (\overline{x} - x_n)^2}{n}}$$

Geburten von Zwillingen in Mecklenburg-Vorpommern

Jahr	♂♂	♂♀	♀♀
2000	76	51	63
2001	49	61	67
2002	64	60	52
2003	65	61	71
2004	70	71	72
2005	57	48	58
2006	64	66	58
2007	67	59	57
2008	64	76	55
2009	63	69	80
2010	71	75	67
2011	74	74	69
2012	63	70	57
2013	56	76	67
2014	74	81	64
2015	60	74	52

Statistische Kenngrößen

♂♂

Max $= 76$; Min $= 49$; Spannweite $= 27$

$x_{Q1} = 61,5$; $x_{med} = 64$; $x_{Q3} = 70,5$;

$x_{mod} = 64$; $\overline{x} = 64,81$; $\sigma = 7,00$

♂♀

Max $= 81$; Min $= 48$; Spannweite $= 33$

$x_{Q1} = 60,5$; $x_{med} = 69,5$; $x_{Q3} = 74,5$;

$x_{mod1} = 61$; $x_{mod2} = 74$;

$\overline{x} = 67,0$; $\sigma = 9,19$

♀♀

Max $= 80$; Min $= 52$; Spannweite $= 28$

$x_{Q1} = 57$; $x_{med} = 63,5$; $x_{Q3} = 68$;

$x_{mod} = 67$; $\overline{x} = 63,06$; $\sigma = 7,69$

Boxplots

Ein **Boxplot** stellt die Verteilung der erfassten Daten dar. Er besteht aus Minimum, unterem Quartil, Median, (arithmetischem Mittel,) oberem Quartil und Maximum. Ausreißer liegen außerhalb des Boxplots.

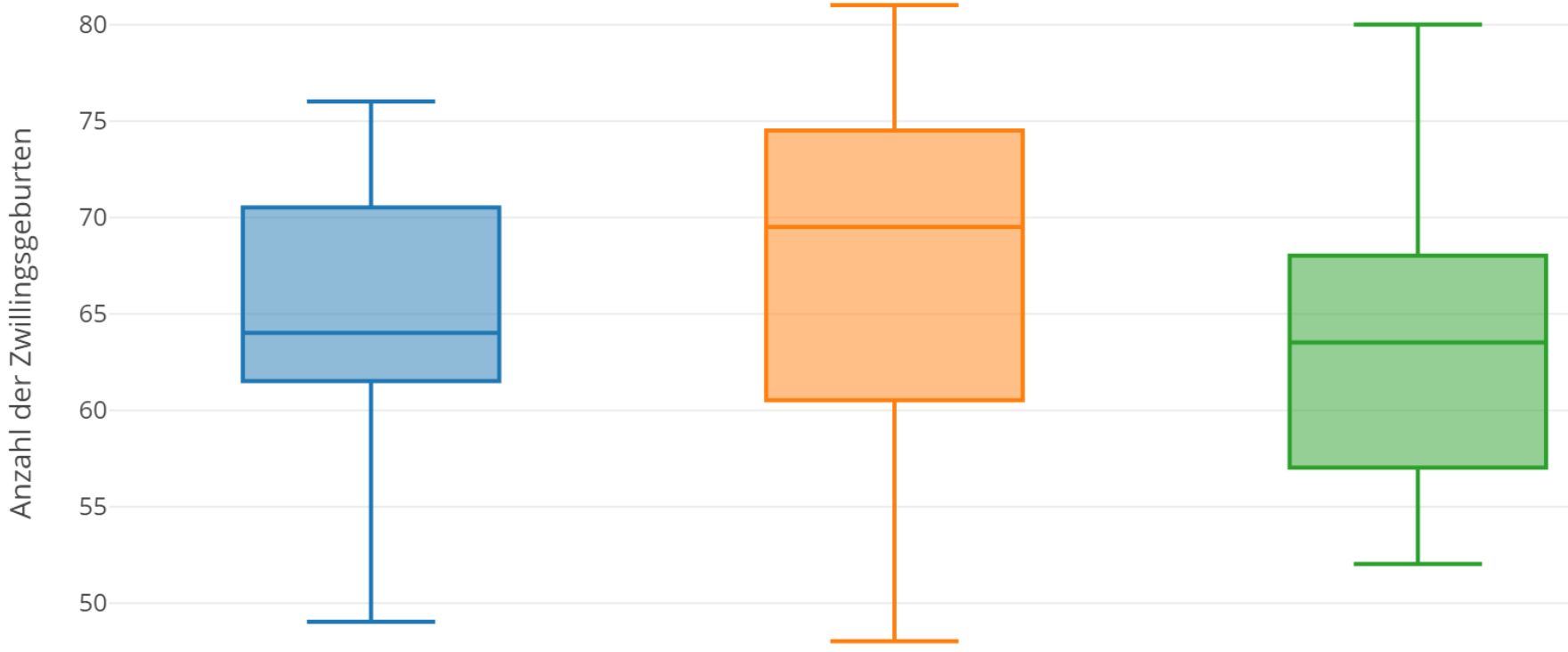

Abbildung: Boxplot [HW 2018, `https://plot.ly/create/box-plot/`]

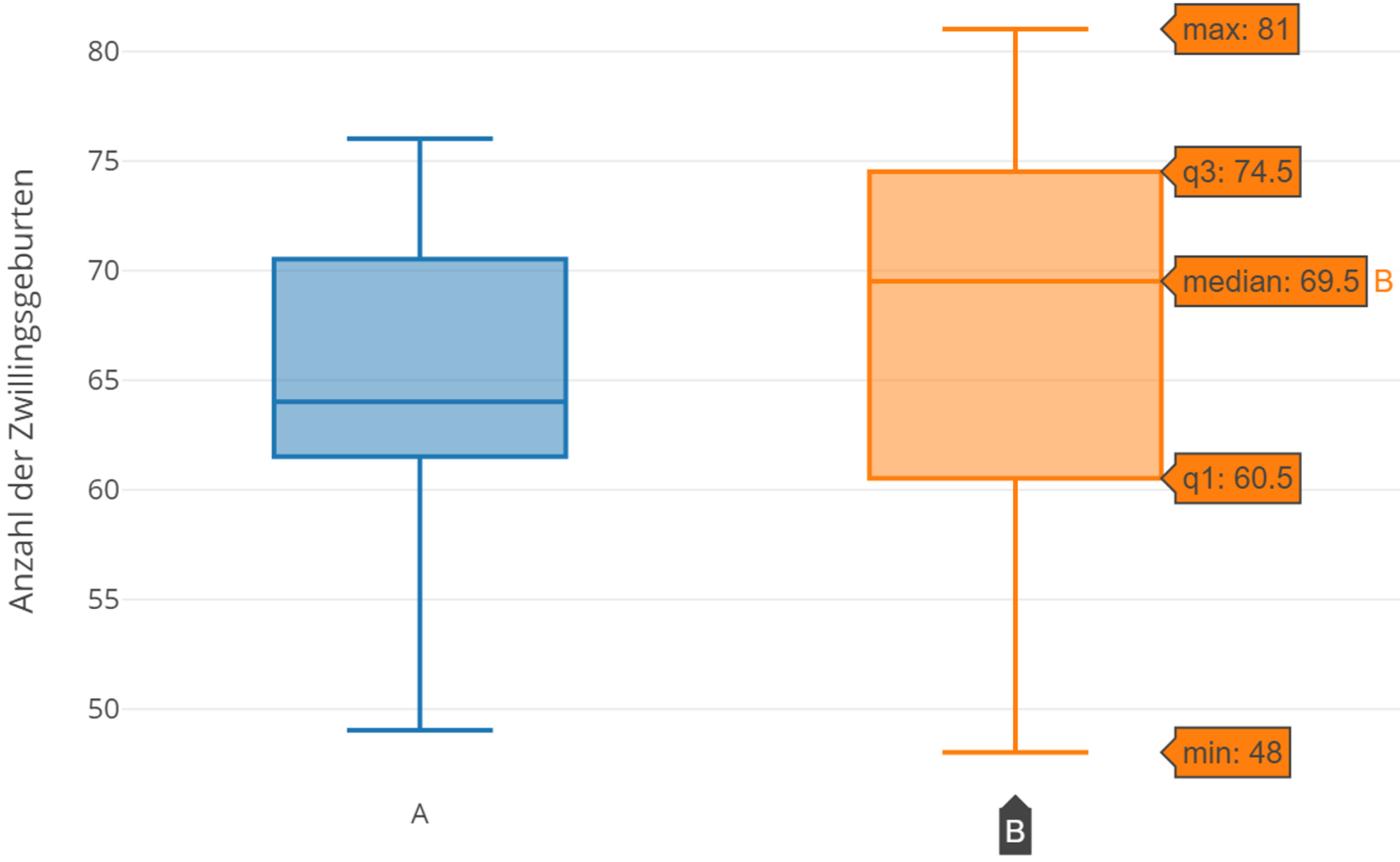

Abbildung: Boxplot [HW 2018, `https://plot.ly/create/box-plot/`]