Wahrscheinlichkeit und Statistik für Business und Datenforschung

- Teil 1: Daten

Einführung

Wahrscheinlichkeit und Statistik

- **Statistik** ist die mathematische Wissenschaft hinter dem Problem "was kann ich über eine Population wissen, wenn es mir nicht möglich ist, jedes Mitglied zu erreichen?"

Wahrscheinlichkeit und Statistik

- Wenn wir die Größe jedes Bewohners Australiens messen könnten, dann könnten wir eine Aussage über die durchschnittliche Größe der Australier zum Zeitpunkt unserer Messung treffen.
- Hier kommt das Thema **Stichproben** ins Spiel.

Wahrscheinlichkeit und Statistik

- Wenn wir eine vernünftig große Stichprobe an Australiern nehmen und deren Größe messen, dann können wir statistische Rückschlüsse für die Gesamtbevölkerung Australiens daraus ziehen.
- Wahrscheinlichkeiten helfen uns dabei, zu wissen, wie sicher wir in unserer Annahme sein können.

Beispiel PARSHIP

- „Alle 11 Minuten verliebt sich ein Single über PARSHIP“

Frage: Sind das gute oder nicht so gute Neuigkeiten?

Beispiel PARSHIP

- PARSHIP hatte 5 Mio. Nutzer beim Start dieser Werbung

- Wenn sich alle **10 Minuten** zwei von denen verlieben und diese durch zwei neuen Singles ersetzt werden, dann sind die Chancen das sich ein zufälliger Single verliebt **2%** für jedes Jahr.
- Das heißt die Wahrscheinlichkeit **KEINEN** Partner über PARSHIP zu finden liegt bei **98%.**

Beispiel PARSHIP Erläuterung

- Wenn sich alle 10 Minuten zwei Singles verlieben, dann passiert das 6 mal pro Stunde, 144 mal am Tag oder **52.560 mal pro Jahr.**
- Die Wahrscheinlichkeit dass du der Nutzer bist der den Partner aus den 52.560 findet ist 2 zu 5.000.000 (=2/5.000.000 = 0.0000004)!
- Oder wenn man das **gesamte Jahr betrachtet (2/5.000.000)* 52.560 = 0.021 (=2,1 %)**
- **Antwort:** nicht so gut...

Daten

Was sind Daten?

- **Daten** = die gesammelten Beobachtungen, die wir über etwas haben.

- Daten können **kontinuierlich** sein:

„Wie verhält sich der Aktienkurs?„

- oder **kategorisch:**

"Welches Auto hat die beste Reparaturhistorie?"

Warum Daten wichtig sind

- Sie helfen uns **die Dinge zu verstehen, wie sie sind**:

"Welche Beziehungen bestehen zwischen zwei Ereignissen?"

„Müssen Menschen, die einen Apfel am Tag essen, weniger häufig zum Arzt als solche, die das nicht tun?"

Warum Daten wichtig sind

- Sie helfen uns, **zukünftiges Verhalten vorherzusagen**, um Geschäftsentscheidungen zu treffen:

"Basierend auf der Klick-Historie, welche Werbung bringt diesen Nutzer am ehesten auf unsere Seite?"

Datenvisualisierung

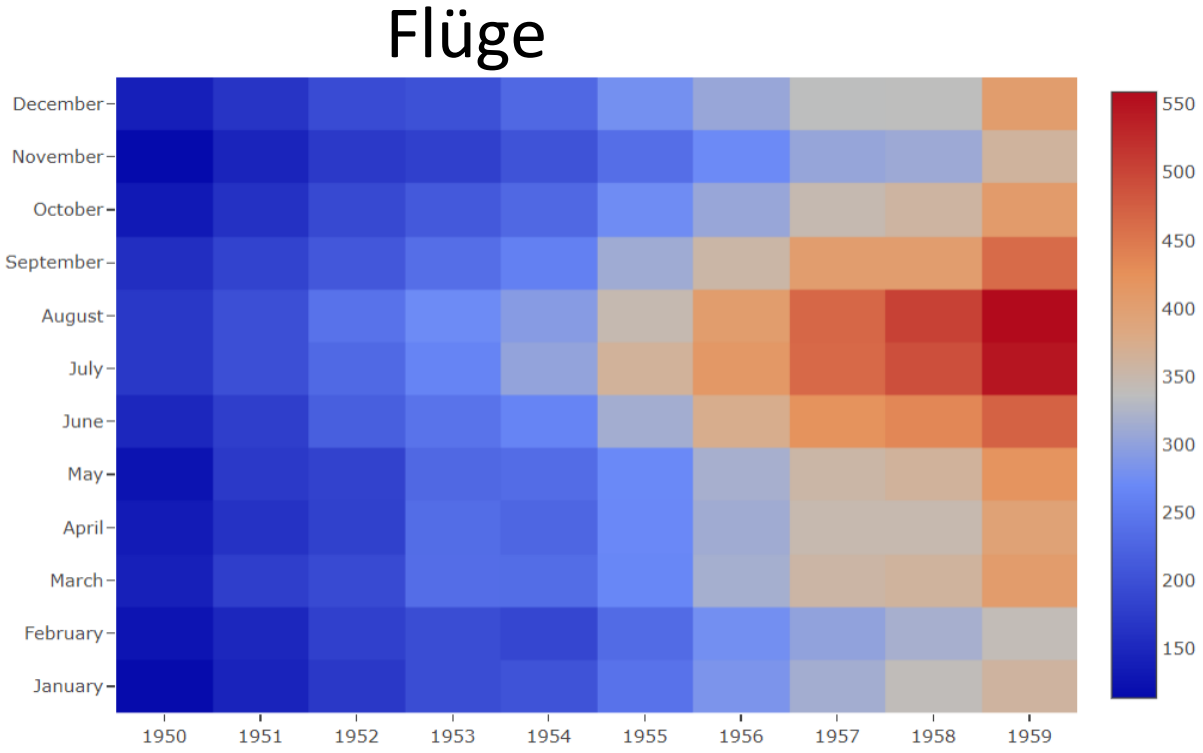
- Eine **Tabelle**:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**year**	**month**	**passengers**		**year**	**month**	**passengers**		**year**	**month**	**passengers**		**year**	**month**	**passengers**
2	1950	January	115		1952	July	230		1955	January	242		1957	July	465
3	1950	February	126		1952	August	242		1955	February	233		1957	August	467
4	1950	March	141		1952	September	209		1955	March	267		1957	September	404
5	1950	April	135		1952	October	191		1955	April	269		1957	October	347
6	1950	May	125		1952	November	172		1955	May	270		1957	November	305
7	1950	June	149		1952	December	194		1955	June	315		1957	December	336
8	1950	July	170		1953	January	196		1955	July	364		1958	January	340
9	1950	August	170		1953	February	196		1955	August	347		1958	February	318
10	1950	September	158		1953	March	236		1955	September	312		1958	March	362
11	1950	October	133		1953	April	235		1955	October	274		1958	April	348
12	1950	November	114		1953	May	229		1955	November	237		1958	May	363
13	1950	December	140		1953	June	243		1955	December	278		1958	June	435
14	1951	January	145		1953	July	264		1956	January	284		1958	July	491
15	1951	February	150		1953	August	272		1956	February	277		1958	August	505
16	1951	March	178		1953	September	237		1956	March	317		1958	September	404
17	1951	April	163		1953	October	211		1956	April	313		1958	October	359
18	1951	May	172		1953	November	180		1956	May	318		1958	November	310

Es kann hier nicht viel heraus gelesen werden

Datenvisualisierung

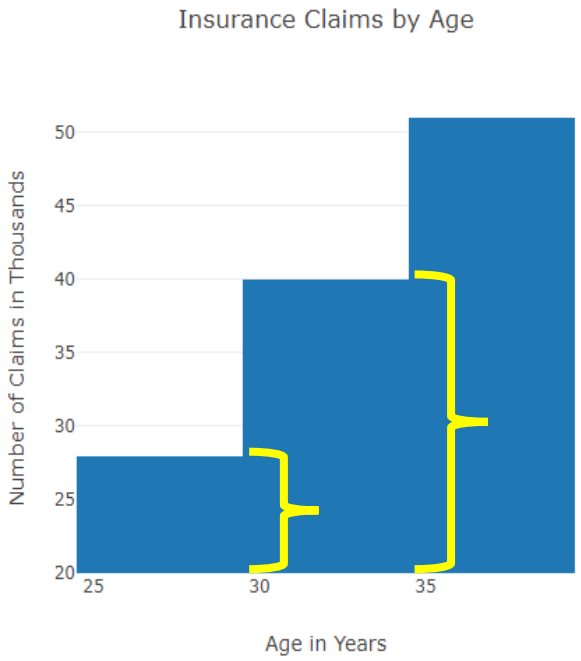
- mit einer **Grafik** vergleichen:

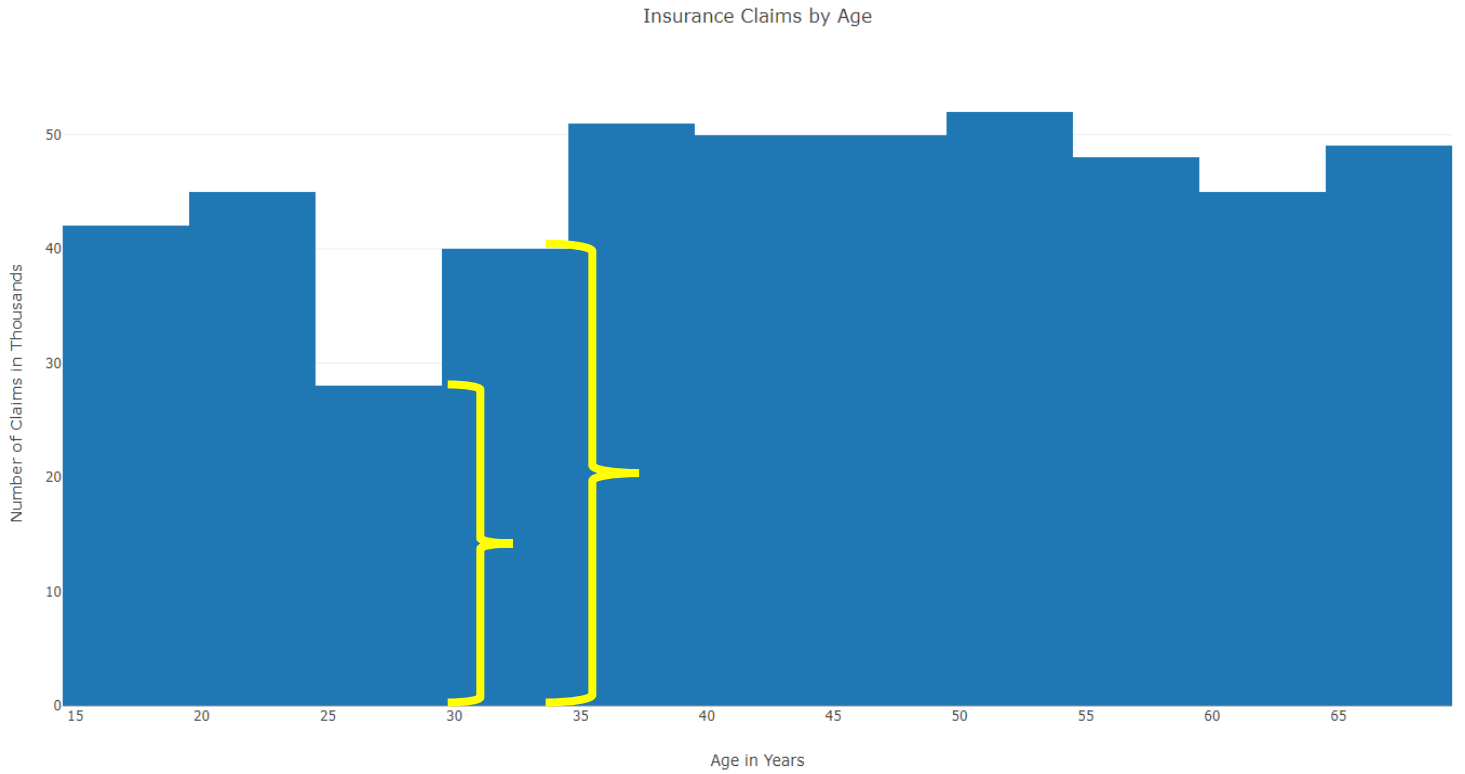


Die Grafik zeigt zwei unterschiedliche Trends auf: eine Zunahme der Fluggäste im Laufe der Jahre und eine größere Anzahl von Passagieren, die in den Sommermonaten fliegen.

Darstellungen kritisch analysieren

- Grafiken können irreführend sein:





Datenmessung

Ebenen der Datenmessungen

- **Nominal**
  - Vordefinierte Kategorien
  - Können nicht sortiert werden

Klassifikationen von Tieren (Säugetiere, Fische, Reptilien)

Augenfarbe (Blau, Grün, Braun)

Ebenen der Datenmessungen

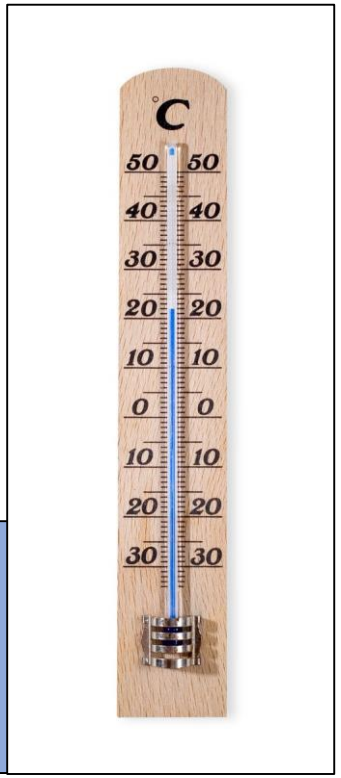
- **Ordinal**
  - Können sortiert werden
  - Sind nicht skalierbar

Meinungsumfragen

Ebenen der Datenmessungen

- **Intervall**
  - Skalierbar
  - Fehlender “Null”-Punkt
  - 20°C ist nicht “doppelt so heiß” wie 10°C

Temperatur



Ebenen der Datenmessungen

- **Verhältnis** (Ratio)

  - Die Werte haben einen echten “Null”-Punkt
  - Die Werte können negative sein

Alter, Gewicht, Gehalt, Entfernungen

Population vs. Stichprobe

- **Population** = jedes Mitglied einer Gruppe

- **Stichprobe** = eine Auswahl von Mitglieder, welche aufgrund von Zeit und Ressourcen zur Messung herangezogen werden

Mathematische Symbole & Syntax

Symbol/Expression	Gesprochen als	Beschreibung
x^2	x quadrat	x auf die zweite Potenz erhöht $x^2 = x \times x$
x_i	x-sub-i	Eine subskribierte Variable (der Index dient hierbei als Label)
$x!$	x faktoriell	$4! = 4 \times 3 \times 2 \times 1$
$\overline{x}$	x Strich	Symbol für den Stichprobenmittelwert
μ	"mü"	Symbol für den Mittelwert der Grundgesamtheit (griechischer Kleinbuchstabe mu)
Σ	sigma	Syntax für Summe (griechischer Großbuchstabe Sigma)

Exponenten

$$x^5 = x \times x \times x \times x \times x$$

1 2 3 4 5

Beispiel: $3^4 = 3 \times 3 \times 3 \times 3 = 81$

Exponenten - Spezialfälle

$$x^{-3} = \frac{1}{x \times x \times x}$$

Beispiel: $\mathbf{2^{-3} = \frac{1}{2 \times 2 \times 2} = \frac{1}{8} = 0.125}$

$$x^{\left(\frac{1}{n}\right)} = \sqrt[n]{x}$$

Beispiel: $\mathbf{8^{\left(\frac{1}{3}\right)} = \sqrt[3]{8} = 2}$

Faktoren

$$x! = x \times (x-1) \times (x-2) \times \cdots \times 1$$

Beispiel: $\mathbf{6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720}$

Beispiel: $\frac{5!}{3!} = \frac{5 \times 4 \times \cancel{3 \times 2 \times 1}}{\cancel{3 \times 2 \times 1}} = 5 \times 4 = 20$

Einfache Summen

$$\sum_{x=1}^{n} x = 1 + 2 + 3 + \cdots + n$$

Beispiel: $\sum_{x=1}^{4} x = 1 + 2 + 3 + 4 = 10$

Beispiel: $\sum_{x=1}^{4} x^2 = 1 + 4 + 9 + 16 = 30$

Reihensummen

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

Beispiel:

$$x = \{5, 3, 2, 8\}$$

$$n = \#\,Elemente\ in\ x = 4$$

$$\sum_{i=1}^{4} x_i = 5 + 3 + 2 + 8 = 18$$

Beispielformel

- Formel zur Berechnung des Mittelwerts einer Stichprobe:

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

- Bitte lies dies laut durch:

"x Strich (das Symbol für den Stichprobenmittelwert) ist gleich der Summe (dargestellt durch den griechischen Buchstaben Sigma) aller x-sub-i Werte in der Reihe, da i von 1 bis zur Zahl n Punkte in der Reihe geteilt wird durch n. "

Beispielformel

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

1. Beginne mit einer Reihe von Werten:

   {7 8 9 10}

2. Weise jedem Wert einen Platzhalter zu

   {7 8 9 10}

   1 2 3 4 **n=4**

3. Diese werden zu x_1 x_2 etc.

   $\boldsymbol{x_1 = 7 \quad x_2 = 8 \quad x_3 = 9 \quad x_4 = 10}$

Beispielformel

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

4. Setze nun diese in die Formel ein:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 \ldots + x_n}{n}$$

$$= \frac{7 + 8 + 9 + 10}{4} = \frac{34}{4} = 8,5$$

Zentrale Tendenz der Messgrößen

Datenmessgrößen

- "Was war die durchschnittliche Rendite?"

  Messgröße der zentralen Tendenz

- "Wie weit vom Durchschnitt sind einzelne Werte abgewichen?"

  Messgröße der Dispersion

Messgrößen der zentralen Tendenz (**Mittelwert, Median, Modalwert**)

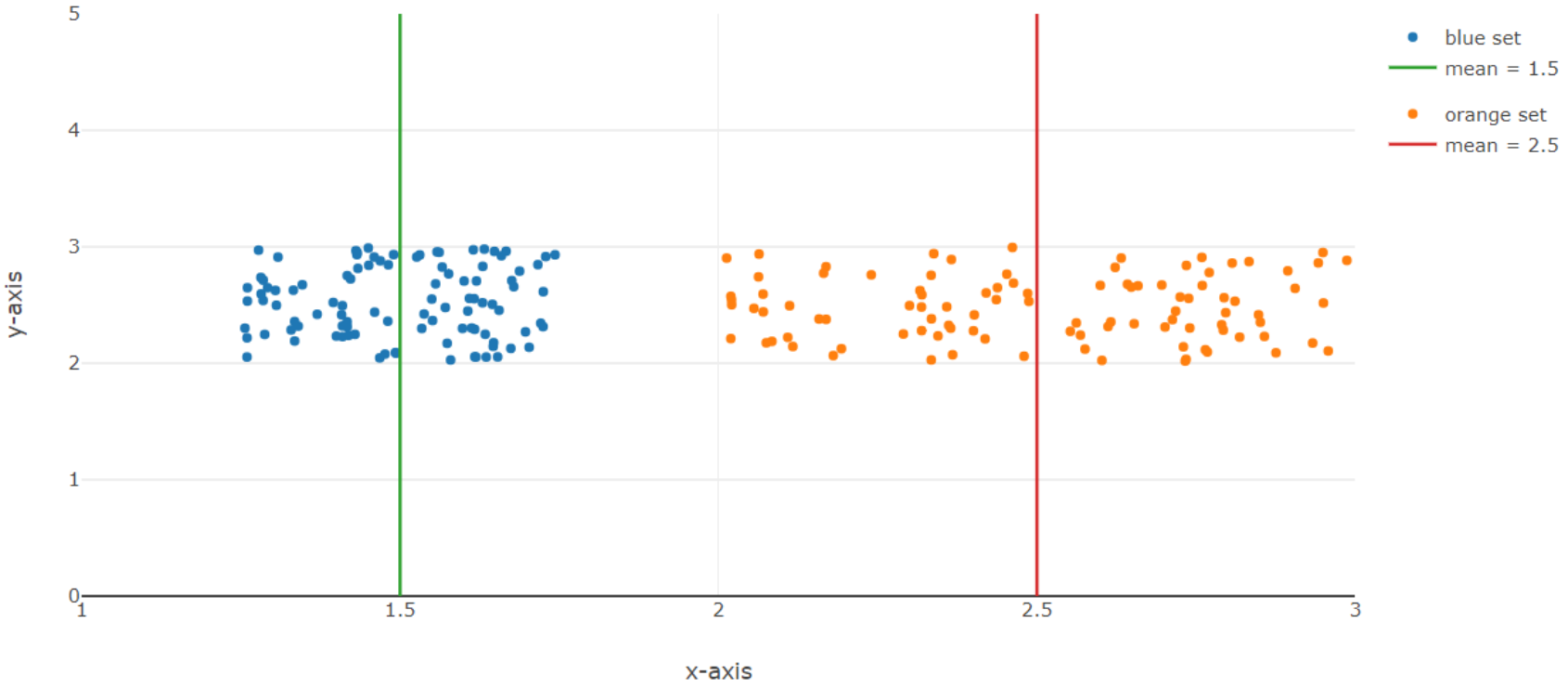
- Beschreibt die „Lokalisierung“ der Daten
- Die "Form" der Daten kann nicht beschrieben werden

Mittelwert (mean) = "berechneter Durchschnitt"

Median = "mittlerer Wert"

Modalwert (mode) = „am häufigsten auftretender Wert"

Mittelwert



- Zeigt „wo“ sich die Daten befinden, nicht jedoch deren „Verteilung“

Median – *ungerade Anzahl von Werten*

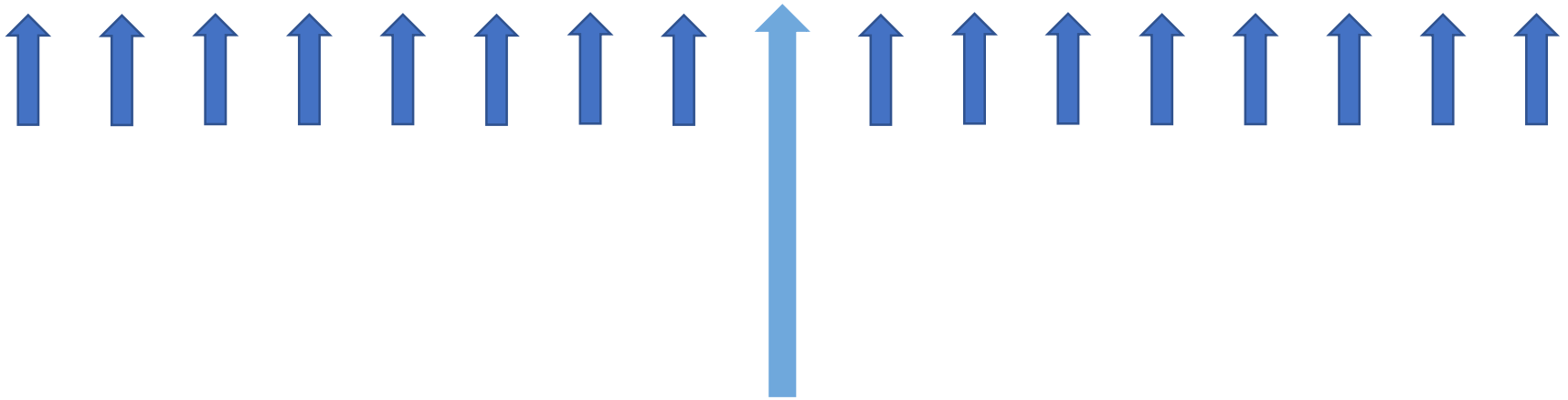
9 10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

33 9 10 19 10 44 11 16 19 21 23 13 28 30 34 36 15

- Zuerst sortieren wir die Datenserie

Median – *ungerade Anzahl von Werten*

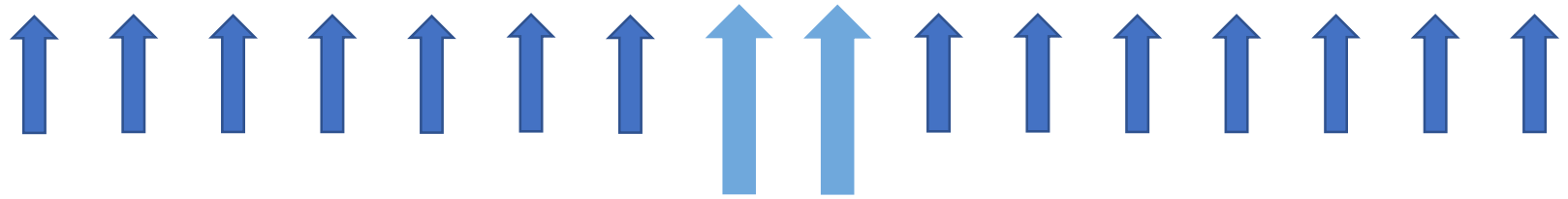
9 10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44



= 19

Median – *gerade Anzahl von Werten*

10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44



$$\frac{19 + 21}{2} = 20$$

Mittelwert vs. Median

- Der Mittelwert kann durch Ausreißer beeinflusst werden.
- Der Mittelwert von {2,3,2,3,2,12} ist 4
- Der Median beträgt 2,5
- Der Median liegt viel näher bei den meisten Werten in der Serie!

Modalwert

10 10 11 13 15 16 16 16 21 23 28 30 33 34 36 44

= 16

- Der Modalwert ist der am häufigsten vorkommende Wert
- Er kann zum Beispiel bei der Betrachtung von “Gewichten” hilfreich sein

Messgröße
Dispersion

Messgrößen der Dispersion/Streuungsmaß (**Streuung, Varianz, Standardabweichung**)

9 10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

- In diesem Beispiel ist der Mittelwert 22,25
- Wie beschreiben wir, wie das Beispiel “verteilt” ist?

Messgrößen der Dispersion/Streuungsmaß (**Streuung, Varianz, Standardabweichung**)

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

- In diesem Beispiel ist der Mittelwert 22,25
- Wie beschreiben wir, wie das Beispiel “verteilt” ist?

Range/Streuung

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

$$Streuung = max - min$$

$$= 39 - 9$$

$$= 30$$

Range/Streuung

- Streuung berücksichtigt nur zwei Werte (max / min)
- Kann durch "Ausreißer" beeinflusst werden

- Überlegung: 9 10 11 11 12 12 39
- Beschreibt dies wirklich die Daten?

Varianz

- Varianz berücksichtigt **jeden** Datenpunkt
- Wir beginnen mit der Berechnung der Summe der Abstände im Quadrat von **jedem Punkt** zum Mittelwert
- Es gibt einen Unterschied zwischen der Stichprobenvarianz und der Varianz der Grundgesamtheit
- Faktor der Bessel-Korrektur ($\boldsymbol{n}$-$\mathbf{1}$)

Varianz

- Stichprobenvarianz:

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

- Populationsvarianz:

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Stichprobenvarianz

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

4 7 9 8 11

$$\bar{x} = \frac{4 + 7 + 9 + 8 + 11}{5} = \frac{39}{5} = 7.8$$

Stichproben-mittelwert

$$s^2 = \frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5-1}$$

$$= 6.7$$

Stichproben-varianz

Standardabweichung

- Quadratwurzel der Varianz
- Vorteil: gleiche Einheiten wie die Stichprobe (sample)
- Wichtig dabei ist:

"Werte, die innerhalb einer Standardabweichung um den Mittelwert liegen"

Stichproben Standardabweichung

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

4 7 9 8 11

$$\bar{x} = \frac{4 + 7 + 9 + 8 + 11}{5} = \frac{39}{5} = 7.8$$

Stichproben-mittelwert

$$s = \sqrt{\frac{(4 - 7.8)^2 + (7 - 7.8)^2 + (9 - 7.8)^2 + (8 - 7.8)^2 + (11 - 7.8)^2}{5 - 1}}$$

$$= \sqrt{6.7} = 2.59$$

Stichproben-standardabweichung

Standardabweichung Population

$$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

Population:

4 7 9 8 11

$$\mu = \frac{4+7+9+8+11}{5} = \frac{39}{5} = 7.8$$

Mittelwert Population

$$\sigma = \sqrt{\frac{(4-7.8)^2+(7-7.8)^2+(9-7.8)^2+(8-7.8)^2+(11-7.8)^2}{5}}$$

$$= \sqrt{5.36} = 2.32$$

Standardabweichung Population

Messgröße Quartil

Quartile und Interquartilabstände (IQR)

- Eine weitere Art Daten zu beschreiben ist über die **Quartile** und die **Interquartilsabstände (**interquartile range**/IQR)**

- Das hat den großen Vorteil, dass jeder Datenpunkt mit einbezogen wird und nicht aggregiert wird!

Quartile und Interquartilabstände (IQR)

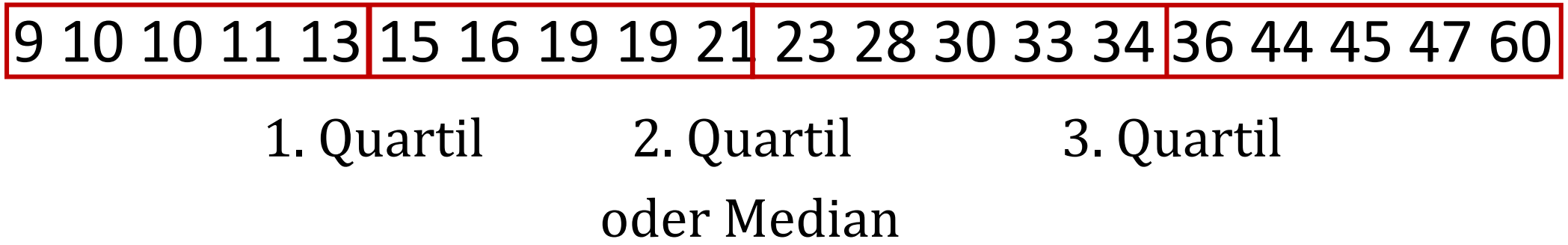
- Betrachten wir die folgende Datenserie mit 20 Werten

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60
1. Quartil	2. Quartil oder Median		3. Quartil

1. Teile die Serie auf
2. Teile jede Unterserie auf
3. Diese bilden die **Quartile**

Quartile und Interquartilabstände (IQR)

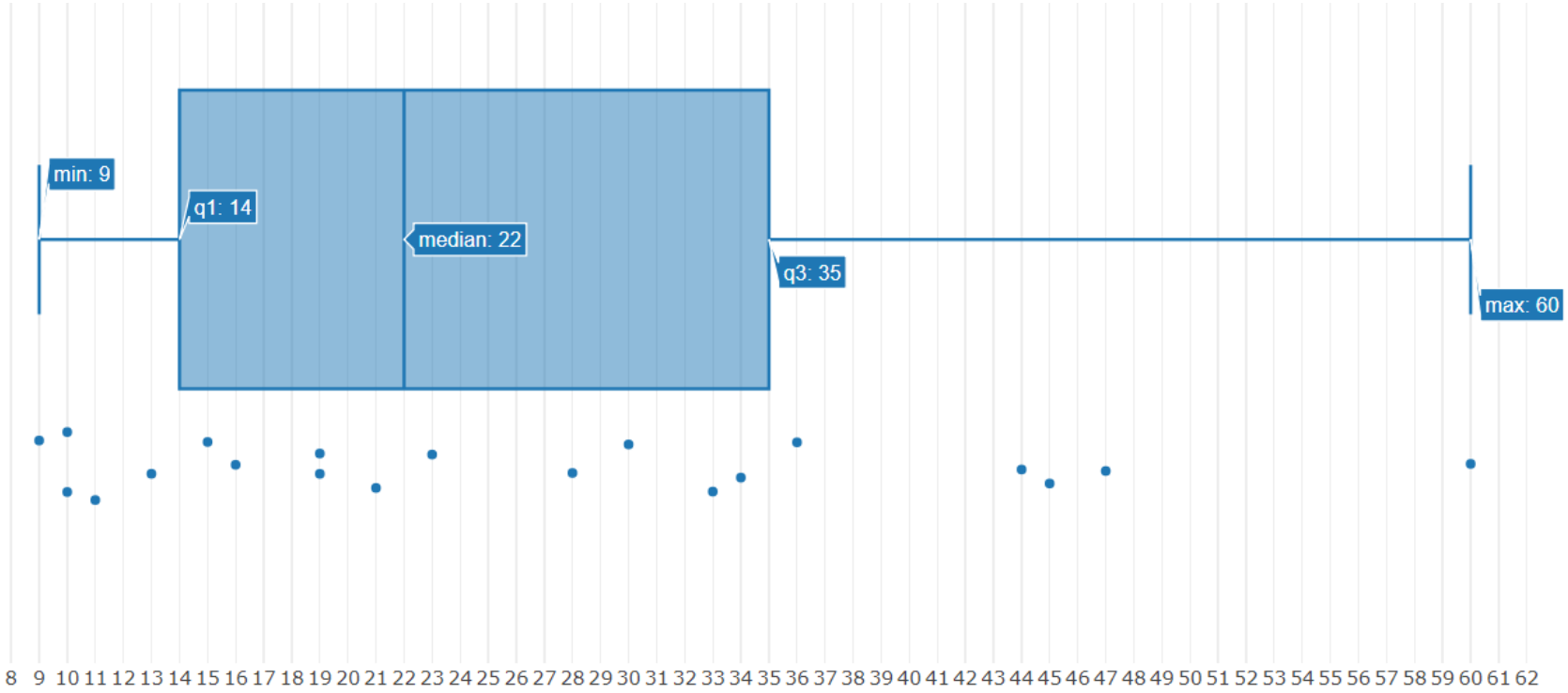
- Betrachten wir die folgende Datenserie mit 20 Werten



1. Quartil = 14
2. Quartil = 22
3. Quartil = 35

Quartile grafisch darstellen

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60

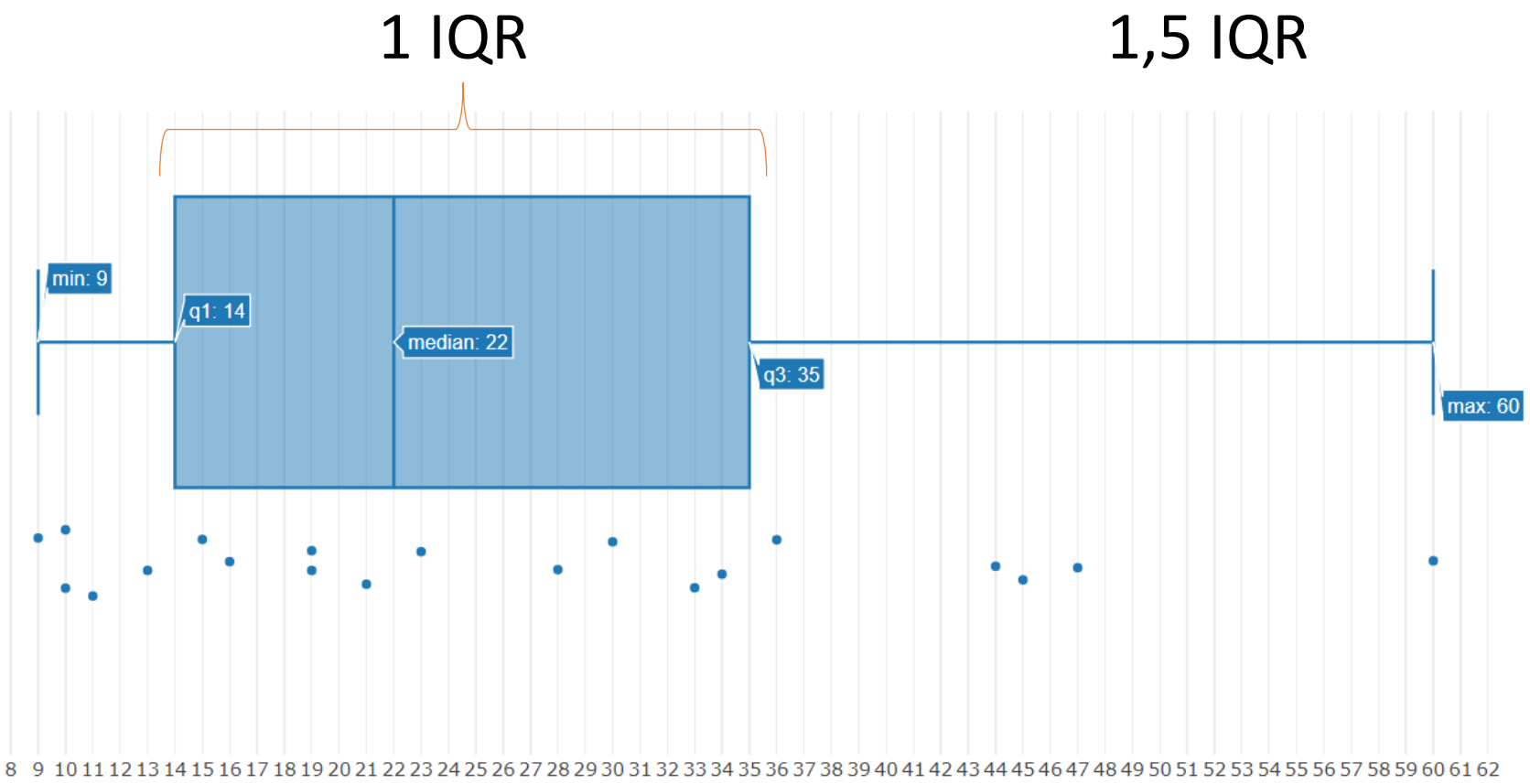


Quartile haben selten die selbe Größe!

Grenzen und Ausreißer

- Was gilt als "Ausreißer"?
- Eine gängige Praxis ist es, eine „Abgrenzung" zu setzen, 1,5 mal so groß wie der IQR
- Alles außerhalb dieser Begrenzung ist ein Ausreißer
- Dies wird durch die **Daten** bestimmt, nicht ein beliebiger Prozentsatz!

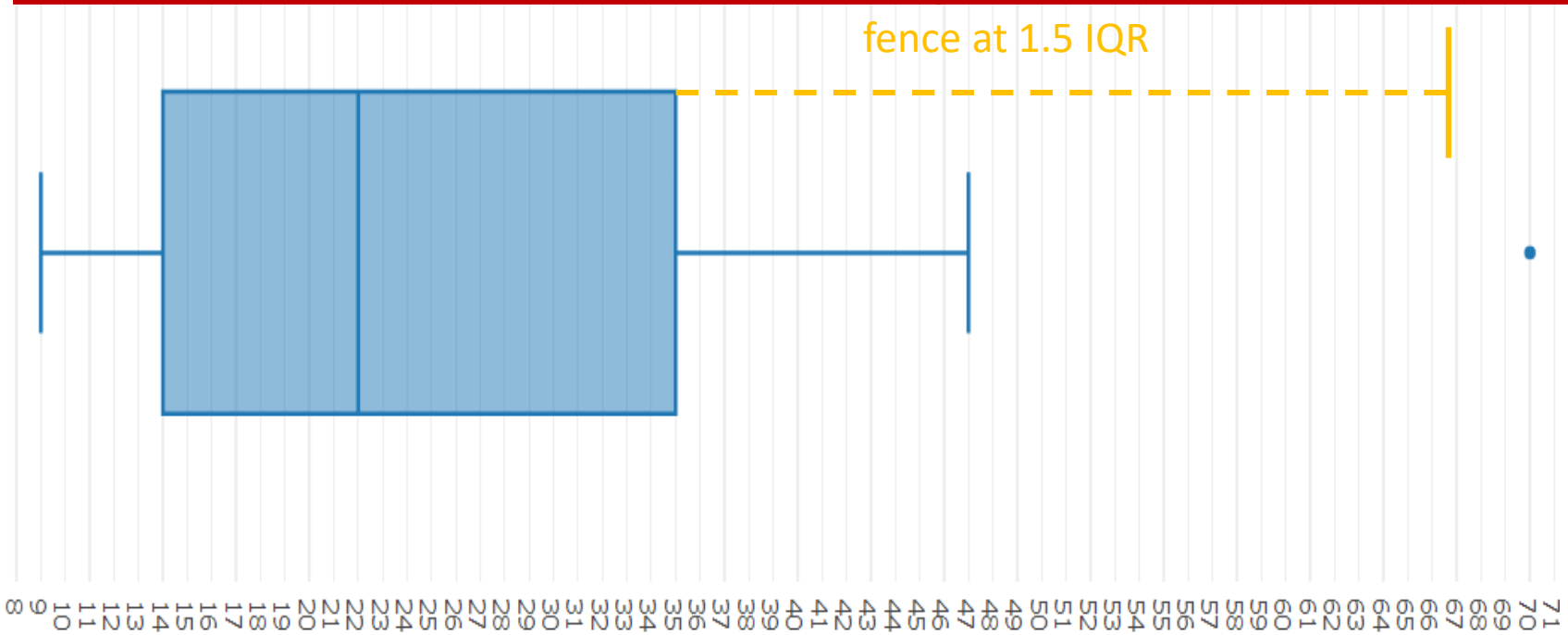
Grenzen und Ausreißer



In diesem Datenset ist 60 kein Ausreißer, 70 wäre es jedoch

Grenzen und Ausreißer

9 10 10 11 13 | 15 16 19 19 21 | 23 28 30 33 34 | 36 44 45 47 **70**



Hier ist 70 ein echter Ausreißer

Bei der Darstellung im Box-Whiskers-Diagramm (Kastengrafik) werden die Antennen (Whisker) an die äußersten Ränder der Begrenzung gezogen.

Bivariate Daten

Bivariate Daten

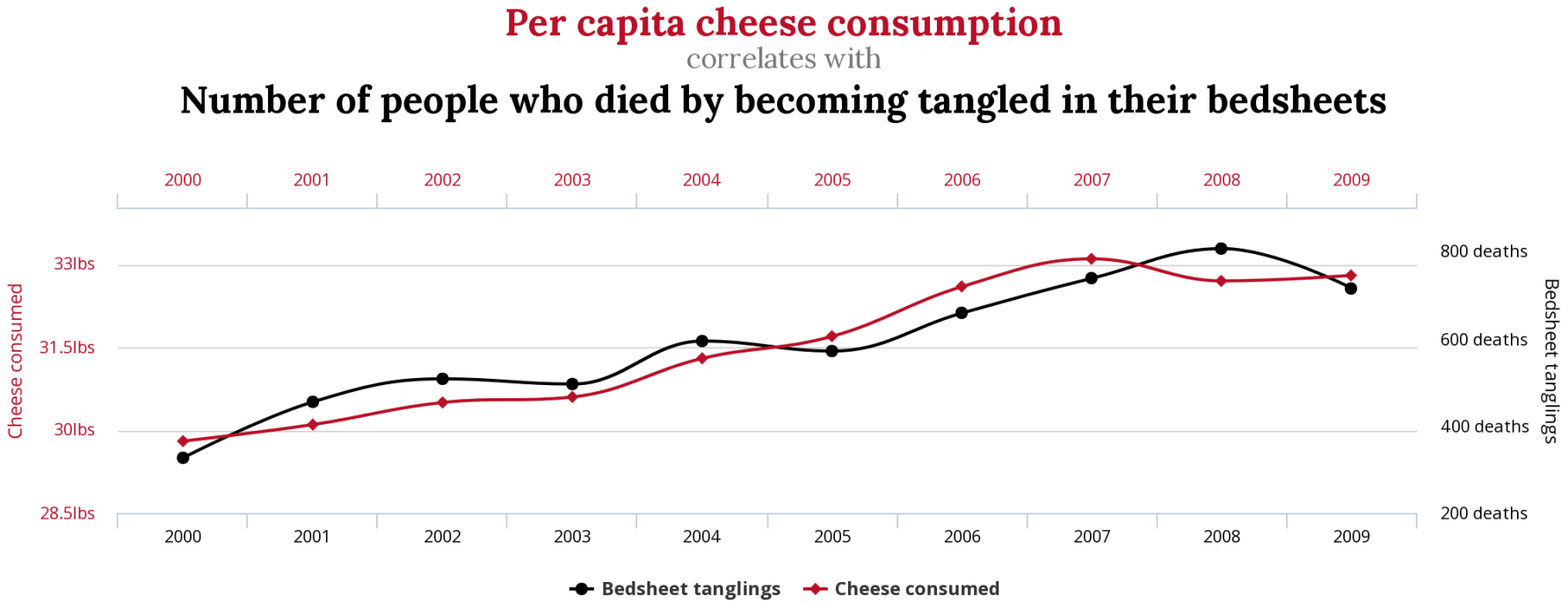
- Vergleicht zwei Variablen
- Die x-Achse wird auf die **unabhängige Variable** gesetzt
- Die y-Achse wird auf die **abhängige Variable** gesetzt, oder diejenige, die relativ zu x gemessen wird

Bivariate Daten

- Streudiagramme (scatter plots) können eine **Korrelation** zwischen zwei Variablen aufdecken
- Sie können ***keine*** **Kausalität** zeigen!

Bivariate Daten

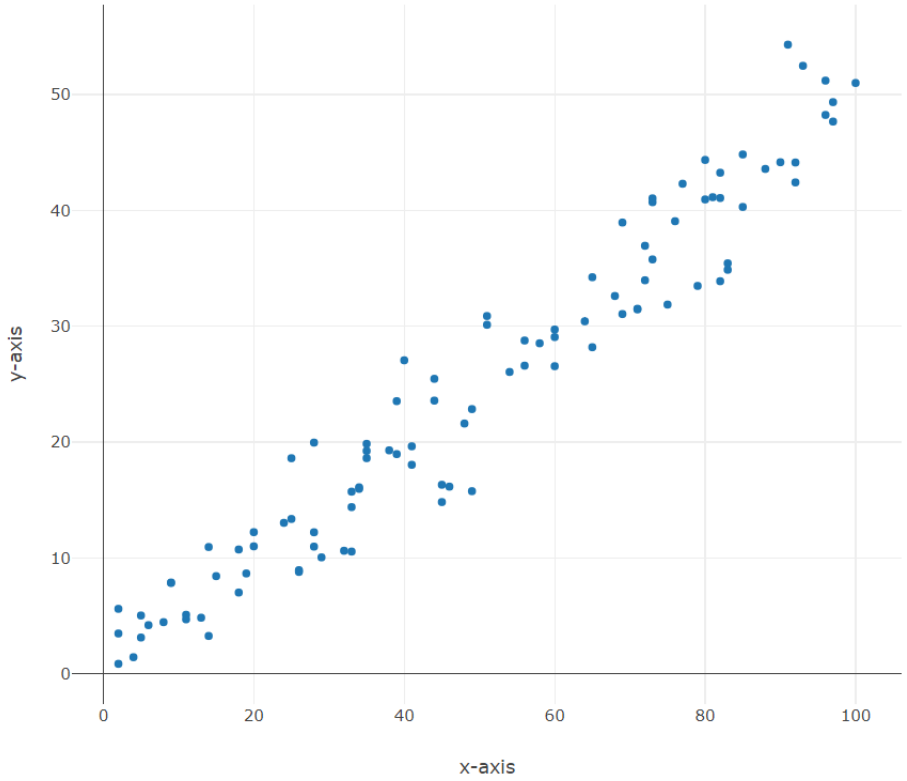
- Die K**orrelation** zweier Variablen
- Sie zeigt keine **Kausalität** auf!



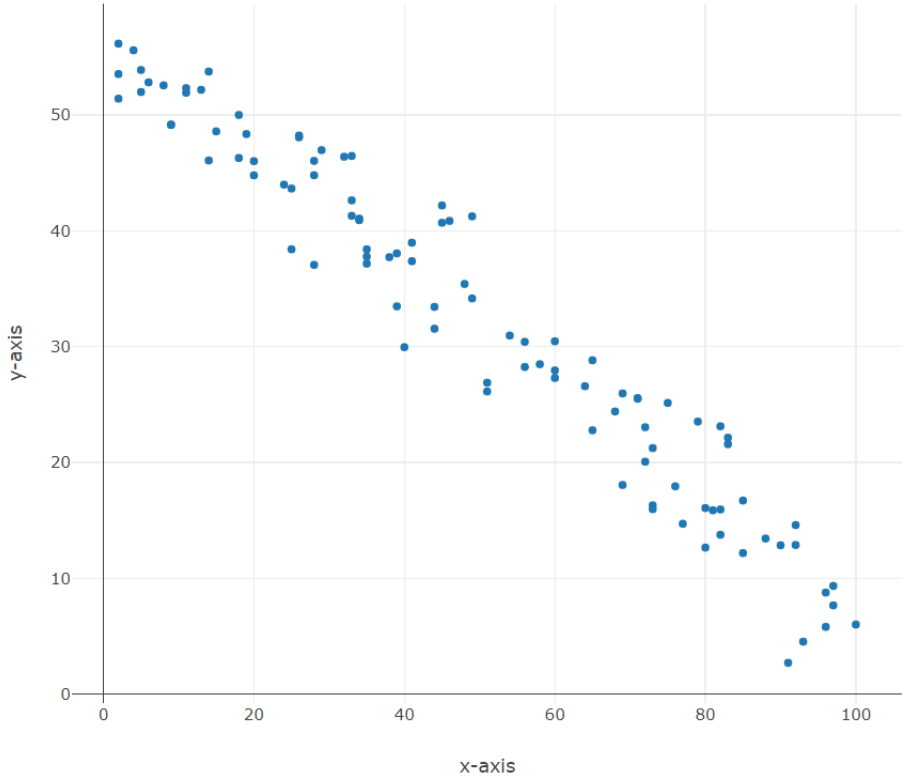
Bivariate Daten

- Um die **Kausalität** zu bestimmen, sind weitere statistische Analysen erforderlich!
- Zum Beispiel: "Verringert die zunehmende Zahl von Polizeibeamten die Kriminalität?"
- Wir werden die Korrelation betrachten und weitere Analysen durchführen, um die Kausalität zu verstehen.

Bivariate Daten



Positive Korrelation



Negative oder inverse Korrelation

Kovarianz

- Eine übliche Methode, zwei Variablen zu vergleichen, ist, ihre Varianzen zu vergleichen - wie weit sind die typischen Werte vom Mittelwert jedes Elements entfernt?
- Die erste Herausforderung besteht darin, den Maßstab zu finden. Beispielsweise ist ein Vergleich der Höhe in cm mit dem Gewicht in kg nicht sinnvoll, es sei denn, wir entwickeln einen **Standard-Score**, um die Daten zu **normalisieren**.

Kovarianz

- Der Einfachheit halber betrachten wir die

  Populationskovarianz:

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Kovarianz Übung

- Vergleiche die beiden Tabellen

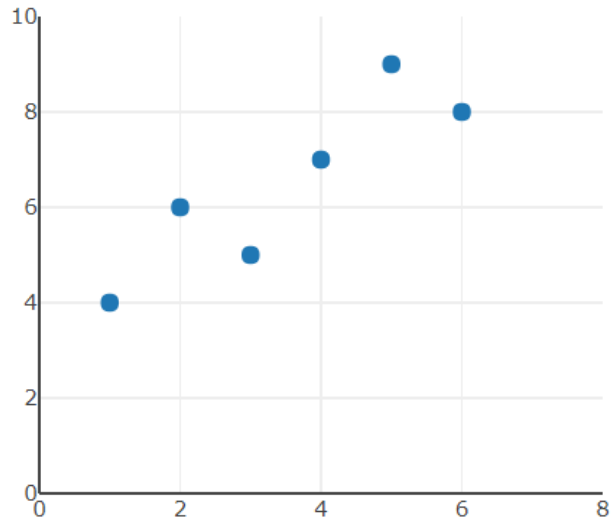
x	y
1	4
2	6
3	5
4	7
5	9
6	8

x	y
1	5
2	9
3	7
4	4
5	8
6	6

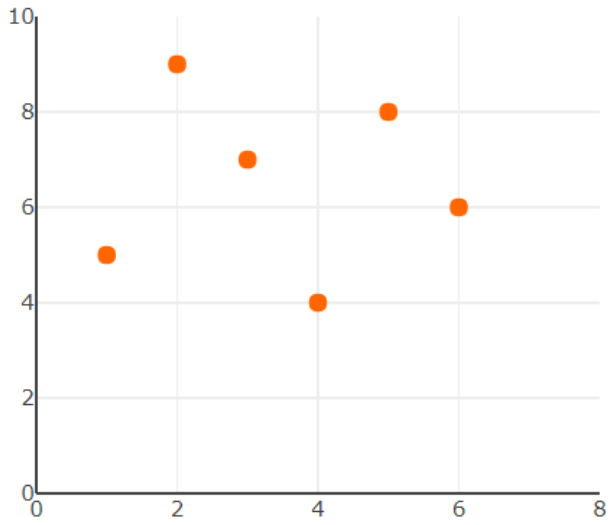
Kovarianz Übung

- Stelle diese nun grafisch dar

x	y
1	4
2	6
3	5
4	7
5	9
6	8



x	y
1	5
2	9
3	7
4	4
5	8
6	6



Kovarianz Übung

$\bar{x} = 3.5, \bar{y} = 6.5$

- Berechne den Mittelwert

x	y
1	4
2	6
3	5
4	7
5	9
6	8

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{y} = \frac{4+6+5+7+9+8}{6} = 6.5$$

x	y
1	5
2	9
3	7
4	4
5	8
6	6

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{y} = \frac{5+9+7+4+8+6}{6} = 6.5$$

Kovarianz Übung

$\bar{x} = 3.5, \bar{y} = 6.5$

- Kalkuliere nun $(x - \bar{x})$ und $(y - \bar{y})$

x	y	$(x - \bar{x})$	$(y - \bar{y})$
1	4	-2.5	-2.5
2	6	-1.5	-0.5
3	5	-0.5	-1.5
4	7	0.5	0.5
5	9	1.5	2.5
6	8	2.5	1.5

x	y	$(x - \bar{x})$	$(y - \bar{y})$
1	5	-2.5	-1.5
2	9	-1.5	2.5
3	7	-0.5	0.5
4	4	0.5	-2.5
5	8	1.5	1.5
6	6	2.5	-0.5

Kovarianz Übung

$\bar{x} = 3.5, \bar{y} = 6.5$

- Kalkuliere weiter $(x - \bar{x})(y - \bar{y})$

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
1	4	-2.5	-2.5	6.25
2	6	-1.5	-0.5	0.75
3	5	-0.5	-1.5	0.75
4	7	0.5	0.5	0.25
5	9	1.5	2.5	3.75
6	8	2.5	1.5	3.75

x	y	$(x - \bar{x})$	$(x - \bar{x})(y - \bar{y})$
1	5	-2.5	3.75
2	9	-1.5	-3.75
3	7	-0.5	-0.25
4	4	0.5	-1.25
5	8	1.5	2.25
6	6	2.5	-1.25

Kovarianz Übung

$\bar{x} = 3.5, \bar{y} = 6.5$

- Berechne nun die Summen

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
1	4	-2.5	-2.5	6.25
2	6	-1.5	-0.5	0.75
3	5	-0.5	-1.5	0.75
4	7	0.5	0.5	0.25
5	9	1.5	2.5	3.75
6	8	2.5	1.5	3.75
			Σ	**15.5**

x	y	$(x - \bar{x})$	$(x - \bar{x})(y - \bar{y})$
1	5	-2.5	3.75
2	9	-1.5	-3.75
3	7	-0.5	-0.25
4	4	0.5	-1.25
5	8	1.5	2.25
6	6	2.5	-1.25
		Σ	**-0.5**

Kovarianz Übung

$\bar{x} = 3.5, \bar{y} = 6.5$

- Berechne jetzt die Kovarianz

x	y
1	4
2	6
3	5
4	7
5	9
6	8

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{15.5}{6} = \mathbf{2.583}$$

Σ	15.5

x	y
1	5
2	9
3	7
4	4
5	8
6	6

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

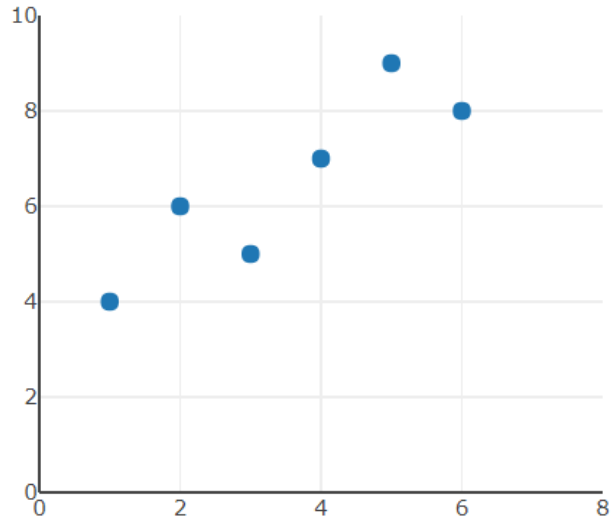
$$= \frac{-0.5}{6} = \mathbf{-0.083}$$

Σ	-0.5

Kovarianz Übung

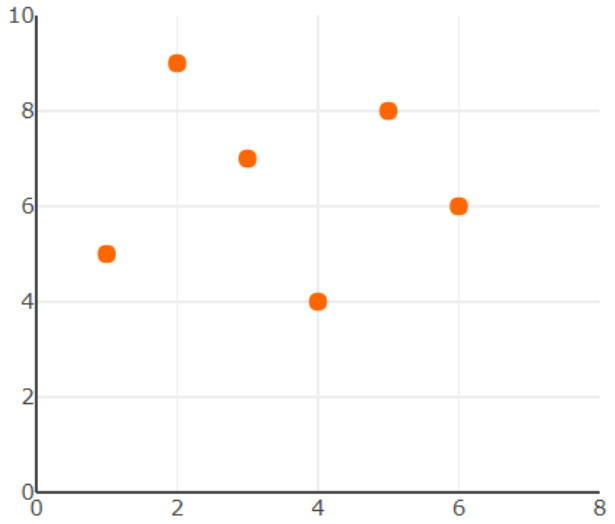
- Vergleich der Kovarianzen

x	y
1	4
2	6
3	5
4	7
5	9
6	8



cov(x,y) = 2.583

x	y
1	5
2	9
3	7
4	4
5	8
6	6



cov(x,y) = -0.083

Pearson-Korrelationskoeffizient

Pearson-Korrelationskoeffizient

- Um Werte aus zwei verschiedenen Verteilungen zu normalisieren, verwenden wir:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

ρ = griech. Buchstabe“rho” σ = Standardabweichung

cov = Kovarianz $\bar{x}$ = Mittelwert von X

Pearson-Korrelationskoeffizient

- Werte liegen zwischen +1 und -1, wobei

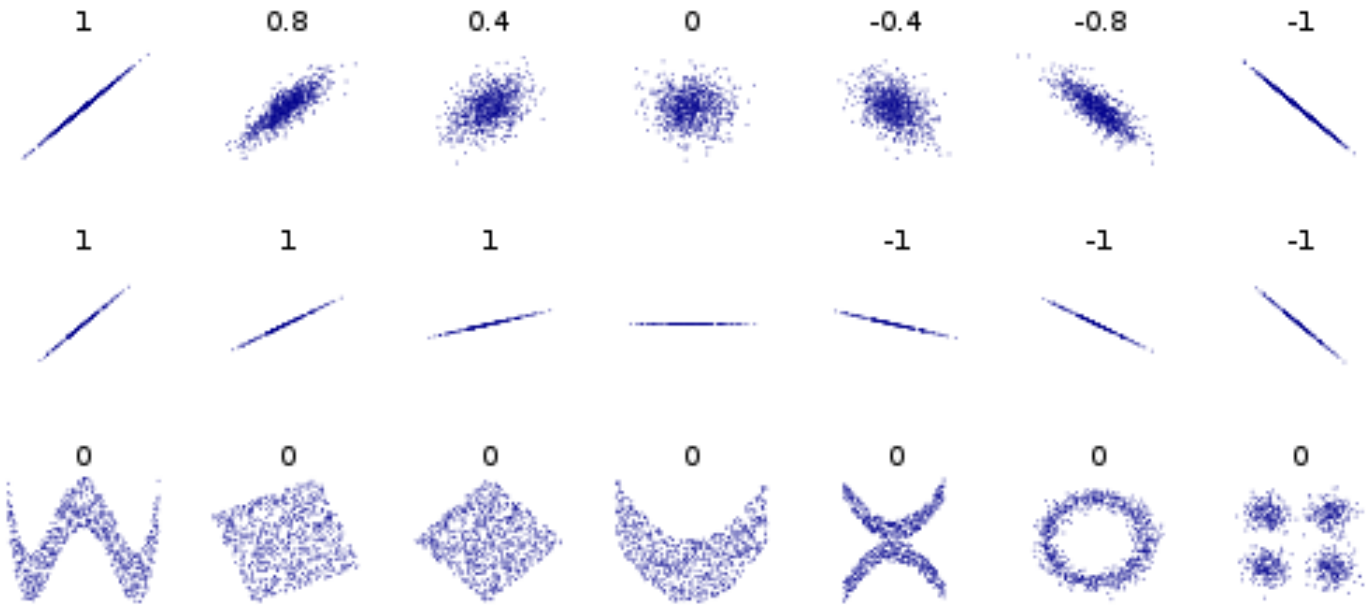
1 = insgesamt positive lineare Korrelation

0 = keine lineare Korrelation

-1 = insgesamt negative lineare Korrelation

Pearson-Korrelationskoeffizient

- Mehrere Reihen von (x, y) Punkten mit dem Korrelationskoeffizienten für jede Reihe:



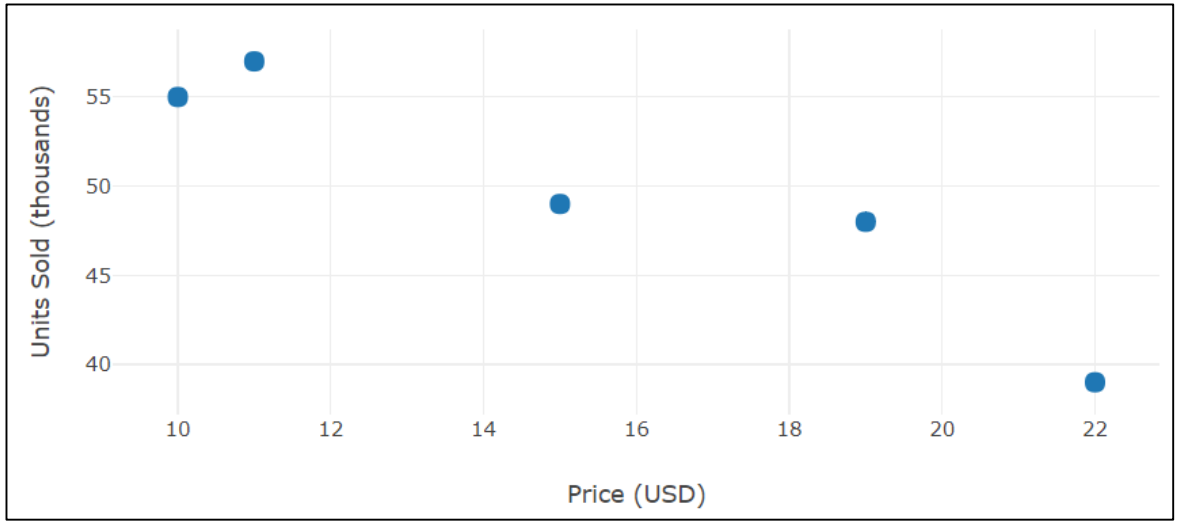
Übung Korrelation

- Ein Unternehmen entscheidet sich, den Verkauf eines neuen Produkts in fünf verschiedenen Märkten zu testen, um den besten Preis zu kalkulieren.
- Sie legen in jedem Markt einen anderen Preis fest und notieren das Verkaufsvolumen über einen 30-Tage-Zeitraum.

Übung Korrelation

- Das sind die Ergebnisse
- Stelle diese grafisch dar

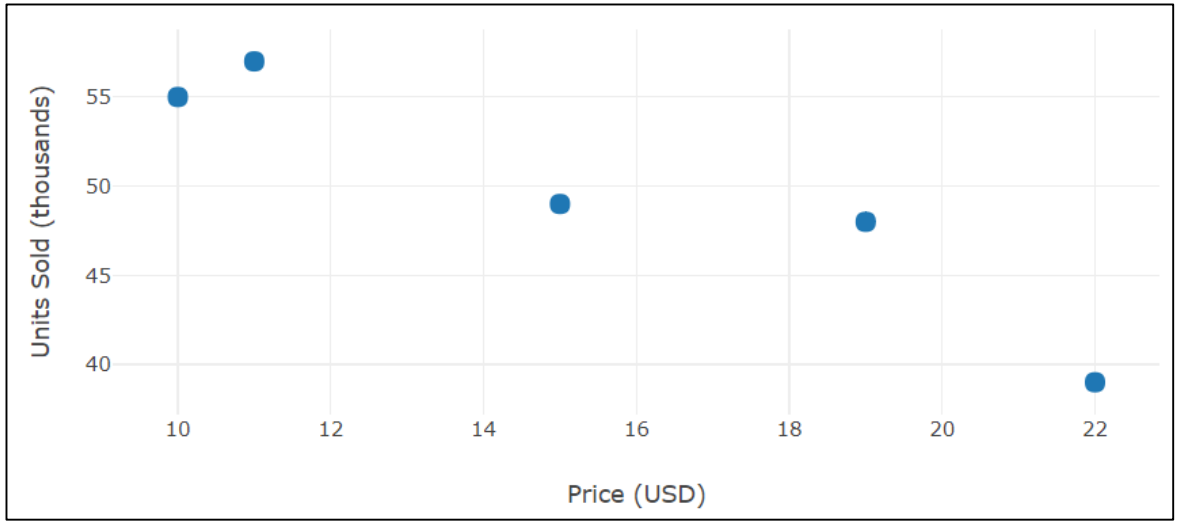
Preis (USD)	Verkaufte Einheiten (tausende)
10	55
11	57
15	49
19	48
22	39



Übung Korrelation

- Es scheint eine starke Korrelation zu geben, aber wie stark?

Preis (USD)	Verkaufte Einheiten (tausende)
10	55
11	57
15	49
19	48
22	39



Übung Korrelation

1. Erinnere dich an die vereinfachte Korrelationsformel:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

Preis (USD)	Verkaufte Einheiten (tausende)
10	55
11	57
15	49
19	48
22	39

2. Finde das Mittel von x und y:

$$\bar{x} = \frac{10 + 11 + 15 + 19 + 22}{5} = 15.4$$

$$\bar{y} = \frac{55 + 57 + 49 + 48 + 39}{5} = 49.6$$

Übung Korrelation

$$\rho_{X,Y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2}\sqrt{\Sigma(y - \bar{y})^2}}$$

$$\bar{x} = 15.4 \quad \bar{y} = 49.6$$

3. Kalkuliere nun $(x - \bar{x})$ und $(y - \bar{y})$:

Preis (USD)	Verkaufte Einheiten (tausende)	$(x - \bar{x})$	$(y - \bar{y})$
10	55	-5.4	5.4
11	57	-4.4	7.4
15	49	-0.4	-0.6
19	48	3.6	-1.6
22	39	6.6	-10.6

Übung Korrelation

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$\bar{x} = 15.4 \quad \bar{y} = 49.6$$

4. Kalkuliere nun $(x-\bar{x})$ $(y-\bar{y})$:

Preis (USD)	Verkaufte Einheiten (tausende)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
10	55	-5.4	5.4	-29.16
11	57	-4.4	7.4	-32.56
15	49	-0.4	-0.6	0.24
19	48	3.6	-1.6	-5.76
22	39	6.6	-10.6	-69.96

Übung Korrelation

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$\bar{x} = 15.4 \quad \bar{y} = 49.6$$

5. Kalkuliere nun $(x-\bar{x})^2$ und $(y-\bar{y})^2$:

Preis (USD)	Verkaufte Einheiten (tausende)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36

Übung Korrelation

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$\bar{x} = 15.4 \quad \bar{y} = 49.6$$

6. Bilde nun die Summen

Preis (USD)	Verkaufte Einheiten (tausende)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36
			Σ	**-137.2**	**105.2**	**199.2**

Übung Korrelation

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$\bar{x} = 15.4 \quad \bar{y} = 49.6$$

7. Setze diese in die Originalformel ein:

Preis (USD)	Verkaufte Einheiten (tausende)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36
			Σ	**-137.2**	**105.2**	**199.2**

Übung Korrelation

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$$\bar{x} = 15.4 \quad \bar{y} = 49.6$$

7. Setze diese in die Originalformel ein:

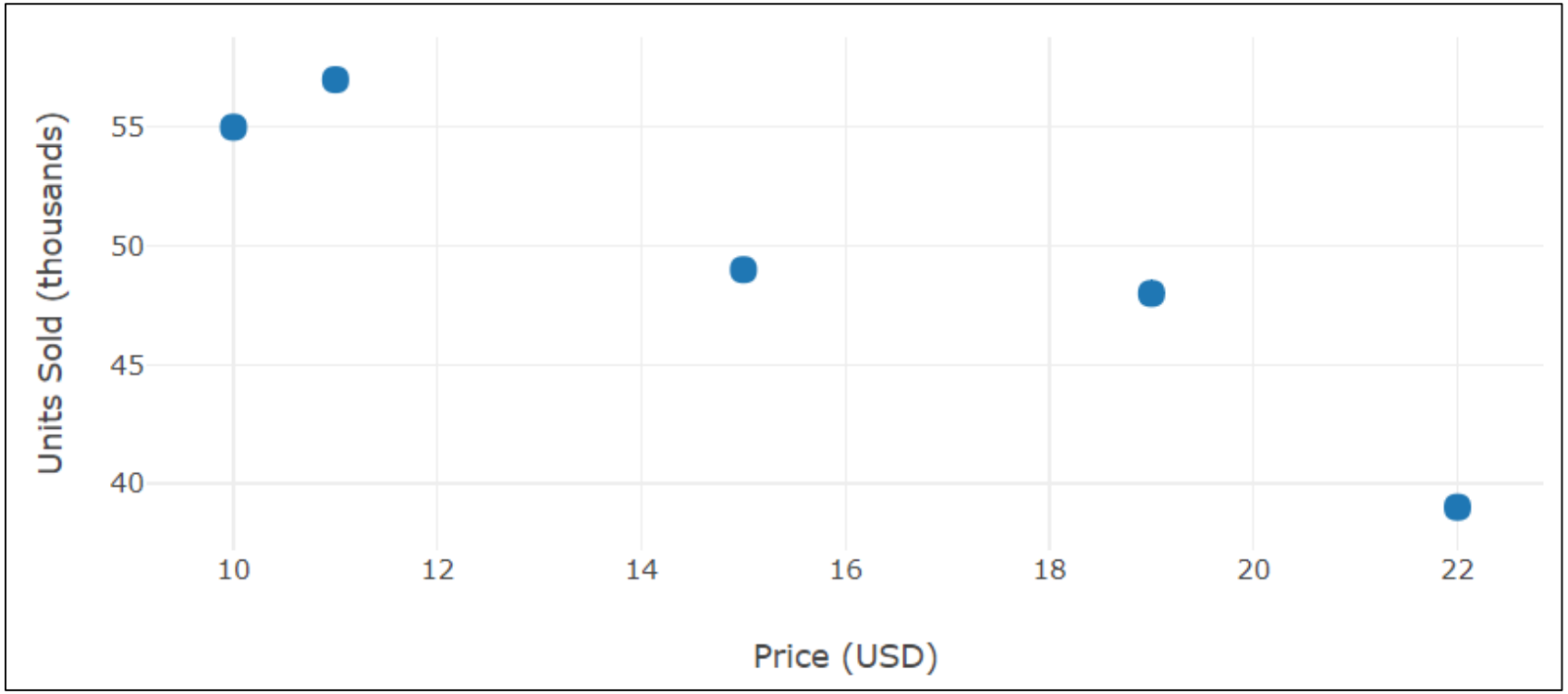
$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{-137.2}{\sqrt{105.2}\sqrt{199.2}}$$

$$= \frac{-137.2}{10.26 \times 14.11} = \frac{-137.2}{144.8} = \mathbf{-0.948}$$

Σ	-137.2	105.2	199.2

Übung Korrelation

- $\rho_{X,Y} = -0.948$ **Zeigt eine sehr hohe negative Korrelation!**



Als nächstes: Wahrscheinlichkeiten