Statistik

Johannes Bonnekoh

Inhaltsverzeichnis

1. Definitionen

Statistik ist ein Teilgebiet der Mathematik und beschäftigt sich mit der Untersuchung gegebener Daten. Im Gegensatz zur Stochastik geht es hier nicht um die Vorhersage von Wahrscheinlichkeiten mit denen ein bestimmtes Ereignis eintritt, sondern um die Auswertung von Häufigkeiten mit denen bestimmte Merkmale aufgetreten sind.

Mit Hilfe der so gewonnen Ergebnisse können Zusammenhänge zwischen unterschiedlichen Datensätzen untersucht und evtl. Vorhersagen über zu erwartende Ergebnisse gemacht werden.

Definitionen: (Grundbegriffe)

1. Die untersuchte Eigenschaft wird als **Merkmal** bezeichnet.
2. Die verschiedenen Formen in denen ein Merkmal vorkommen kann werden als **Merkmalsausprägungen** bezeichnet.
3. Die Gegenstände und Personen, die ein solches Merkmal besitzen sind die **Merkmalsträger**.
4. Alle Merkmalsträger zusammen bilden die **Grundgesamtheit**.
5. Die Anzahl der Elemente in der Grundgesamtheit ist der **Erhebungsumfang**.
6. Die erfassten Merkmale werden während der Erfassung in der **Urliste** notiert.
7. Eine Auswahl aus dieser Urliste wird als **Stichprobe** bezeichnet.
8. Die Anzahl der Merkmalsträger in einer Stichprobe wird **Stichprobenumfang** genannt.
9. Die einzelnen Merkmalsausprägungen einer Stichprobe nennt man **Stichprobenwerte**.

Beispiele:

1. An einer Straßenkreuzung wird die Farbe der vorbeifahrenden Autos erfasst.
   Merkmal: Autofarbe
   Merkmalsausprägungen: rot, blau, grün, weiß, etc.
   Merkmalsträger: die einzelnen Autos
   Grundgesamtheit: die Menge aller Autos die im Erhebungszeitraum an dieser Straßenkreuzung vorbeigefahren sind
   Erhebungsumfang: Anzahl der Autos die im Erhebungszeitraum an dieser Straßenkreuzung vorbeigefahren sind
2. An einer Straßenkreuzung wird der durch jedes vorbeifahrende Auto verursachte Lärm in dB (Dezibel) gemessen.
   Merkmal: verursachter Lärm
   Merkmalsausprägungen: 80, 79,5, 81,233, etc.
   Merkmalsträger: die einzelnen Autos
   Grundgesamtheit: die Menge aller Autos die im Erhebungszeitraum an dieser Straßenkreuzung vorbeigefahren sind
   Erhebungsumfang: Anzahl der Autos die im Erhebungszeitraum an dieser Straßenkreuzung vorbeigefahren sind
3. In einer Schulklasse mit 28 Schülerinnen und Schülern werden die Schuhgrößen der Schülerinnen und Schüler erfasst.
   Merkmal: Schuhgröße
   Merkmalsausprägungen: 34, 37, etc.

Merkmalsträger: die einzelnen Schülerinnen und Schüler
Grundgesamtheit: die Menge aller Schülerinnen und Schüler in der Klasse
Erhebungsumfang: 28

4. Nach der Benotung von 25 Klausuren werden die einzelnen Noten inkl. Tendenzen erfasst.
   Merkmal: Note
   Merkmalsausprägungen: 1+, 1, 1-, 2+, 2, 2-, 3+, 3, 3-, 4+, 4, 4-, 5+, 5, 5-, 6
   Merkmalsträger: die einzelnen Klausuren
   Grundgesamtheit: die Menge aller Klausuren der Klasse
   Erhebungsumfang: 25

Definitionen: (Skalen)
Wir unterscheiden drei verschiedene Skalenarten.

1. **Nominalskala**
   Mit einer Nominalskala werden qualitative Merkmale beschrieben. Sie gibt lediglich Namen und Bezeichnungen wieder.
2. **Ordinalskala**
   Auch die Ordinalskala beschreibt qualitative Merkmale. Sie wird für Reihenfolgen genutzt.
3. **Metrische Skala**
   Die metrische Skala beschreibt quantitative Merkmale, d.h. Merkmale, die direkt durch Zahlen abgelesen werden können. Wir unterscheiden **diskrete** und **stetige** Skalen. Bei diskreten Skalen können nur bestimmte Zahlen, z.B. natürliche Zahlen als Anzahlen, ohne die Zahlen dazwischen, z.B. 1,5, als Merkmalsausprägungen vorkommen. Bei stetigen Skalen können alle Zahlen, z.B. Körpergrößen, in einem gewissen Zahlenbereich als Merkmalsausprägungen vorkommen.

Anwendung auf die obigen Beispiele:

1. Nominalskala
2. Stetige metrische Skala
3. Ordinalskala
4. Ordinalskala

Aufgaben:
Geben Sie Merkmal, Merkmalsausprägungen, Merkmalsträger, Grundgesamtheit, Erhebungsumfang und die Skala an:

1. Bei einer Blutspendenaktion mit 214 Teilnehmern wird die Blutgruppe erfasst.
2. Das Jugendamt veranstaltet bei 500 Jugendlichen zwischen 12 und 16 eine Umfrage nach den am häufigsten geschauten Fernsehsendern.
3. An einer Straßenkreuzung wird zwischen 6:00 und 12:00 Uhr die Automarke erfasst.
4. Die Zuschauerzahlen aller Fernsehsendungen sonntags ab 20:15 Uhr werden erfasst.
5. An einer Straßenkreuzung wird gezählt, wie viele Motorräder, Autos, Quads und LKWs vorbeifahren.
6. Bei einem großen Fernsehsender werden für alle Sendungen im Nachmittagsprogramm die Zuschauerzahlen ermittelt.
7. In einer Schulklasse mit 28 Schülerinnen und Schülern wird die Körpergröße der Schülerinnen und Schüler gemessen.

2. Darstellung von Häufigkeiten

Definition: (absolute und relative Häufigkeiten)
Seien $n \in \mathbb{N}$ der Erhebungsumfang und $a_1, \ldots, a_m$ mit $1 \le m \le n$ und $m \in \mathbb{N}$ die möglichen Merkmalsausprägungen bei einer statistischen Untersuchung und $1 \le i \le m$.

1. Tritt die Merkmalsausprägung a_i bei der statistischen Untersuchung H_i-mal auf, so heißt H_i die **absolute Häufigkeit** von a_i und wir schreiben auch $H(a_i)$.
2. Die Größe
$$h(a_i) := \frac{H(a_i)}{n}$$
heißt **relative Häufigkeit** von a_i und wir können auch kurz h_i schreiben.
3. Eine Tabelle, in der jeder Merkmalsausprägung ihre absolute oder relative Häufigkeit zugeordnet wird, heißt **Häufigkeitsverteilung** oder **Häufigkeitstabelle**.

Satz:
Es gilt immer
$$0 \le h_i \le 1 \text{ und } \sum_{i=1}^{m} h_i = h_1 + h_2 + \cdots + h_m = 1$$
für alle $1 \le i \le m$.

Beispiel:
Für das Werfen 30-malige Werfen eines sechsseitigen Würfels erhalten wir folgende Urliste:
5, 4, 2, 2, 1, 5, 6, 3, 6, 4, 3, 1, 1, 5, 3, 3, 5, 1, 6, 3, 2, 3, 1, 4, 6, 2, 5, 2, 6, 3.
Wir erhalten also die folgende Häufigkeitsverteilung.

a_i	„1“	„2“	„3“	„4“	„5“	„6“
$H(a_i)$	5	5	7	3	5	5

Mit relativen Häufigkeiten würde sie dann wie folgt aussehen.

a_i	„1“	„2“	„3“	„4“	„5“	„6“
$h(a_i)$	$\frac{5}{30} = \frac{1}{6}$	$\frac{5}{30} = \frac{1}{6}$	$\frac{7}{30}$	$\frac{3}{30} = \frac{1}{10} = 0{,}1$	$\frac{5}{30} = \frac{1}{6}$	$\frac{5}{30} = \frac{1}{6}$

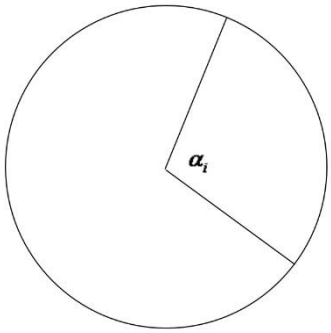
Im Falle einer unendlichen Dezimalzahl ist es am sinnvollsten den Bruch einfach stehen zu lassen. Falls mit der Zahl weitergerechnet werden muss ist dies die beste Lösung.

Definition: (Säulen- und Kreisdiagramm)
Häufigkeitsverteilungen lassen sich u.a. durch Balken- und Kreisdiagramme veranschaulichen.

1. Bei **Säulendiagrammen** steht ein Intervall auf der x – Achse für die Merkmalsausprägung und die y-Achse für die absolute oder relative Häufigkeit. Die Häufigkeiten werden nun durch senkrechte Säulen angezeigt.
2. Bei **Kreisdiagrammen** steht der Vollkreis (360°) für die relative Häufigkeit 1. Der zu einer bestimmten relativen Häufigkeit h_i gehörende Winkel α_i berechnet sich wie folgt:

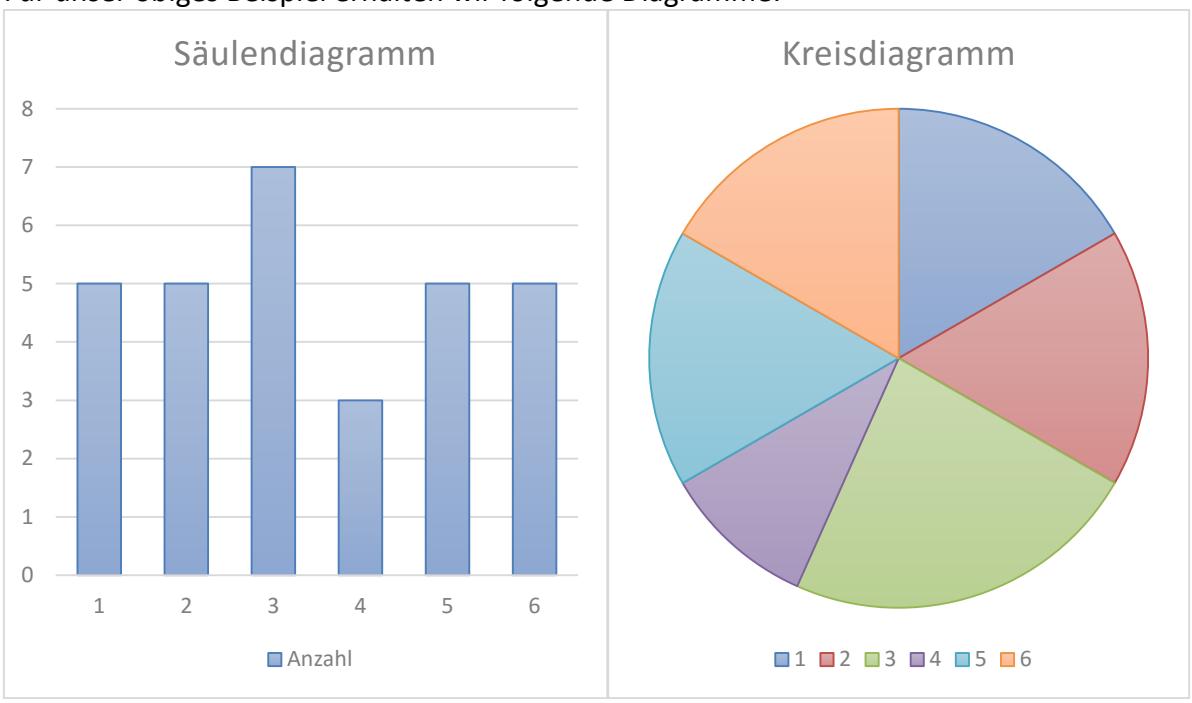
$$\alpha_i = 360° \cdot h_i = \frac{360° \cdot H_i}{n} \ ,$$

wobei n der Erhebungsumfang ist.



Beispiele:

Für unser obiges Beispiel erhalten wir folgende Diagramme:



Aufgabe 1:

Wirf eine Münze 50-mal und erfasse auf der Urliste die Merkmalsausprägungen „Kopf“ und „Zahl“. Berechne die absoluten und die relativen Häufigkeiten und fertige ein Säulen- und ein Kreisdiagramm an!

Aufgabe 2:

Geh auf den Lehrerparkplatz und erfasse

a) Automarke und
b) Farbe (ohne Feinheiten)

der dort stehenden Autos. Berechne die relativen Häufigkeiten und stelle die Ergebnisse als Säulen- und als Kreisdiagramm dar!

Aufgabe 3:

Suchen Sie sich jeweils zu zweit einen Würfel beliebiger Seitenzahl aus!

a) Erstellen Sie eine Urliste für 50 Würfe!
b) Erstellen Sie eine Häufigkeitsverteilung mit relativen Häufigkeiten!
c) Erstellen Sie ein Kreisdiagramm!

3. Lagemaße von Häufigkeitsverteilungen

3.1 Der Modalwert

Definition: (Modalwert)

Die Merkmalsausprägung mit der höchsten absoluten Häufigkeit nennt man auch Modalwert und wird mit a_{Mod} bezeichnet.

Beispiel: An einer Ampel wird eine Stunde lang die Farbe jedes vorbeifahrenden Wagens aufgeschrieben. Es ergibt sich die folgende Häufigkeitsverteilung.

a_i	„rot“	„weiß“	„blau“	„schwarz“	„silber“	„grün“	„orange“
H_i	14	12	15	14	15	7	3

Wir erhalten also zwei Modalwerte:

$$x_{Mod,1} = "blau"$$
$$x_{Mod,2} = "silber".$$

3.2 Positionsunabhängige Lagemaße

Definition: (arithmetischer Mittelwert)

Seien $a_1, \ldots, a_m$ mit dem Erhebungsumfang n und $n, m \in \mathbb{N}$ die möglichen Merkmalsausprägungen einer metrischen Skala und $H_1, \ldots, H_m$ die zugehörigen absoluten Häufigkeiten. Dann ist der arithmetische Mittelwert definiert als

$$\boxed{\overline{x} := \frac{1}{n}\left(a_1 \cdot H_1 + \cdots + a_m \cdot H_m\right) = \frac{1}{n}\sum_{i=1}^{m} a_i \cdot H_i = \sum_{i=1}^{m} a_i \cdot h_i}.$$

Beispiel: Bei der Häufigkeitsverteilung aus Kapitel 2 erhalten wir also:

a_i	„1“	„2“	„3“	„4“	„5“	„6“
$H(a_i)$	5	5	7	3	5	5

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{m} a_i \cdot H_i$$
$$= \frac{1}{30}\left(1\cdot 5 + 2\cdot 5 + 3\cdot 7 + 4\cdot 3 + 5\cdot 5 + 6\cdot 5\right)$$
$$= \frac{1}{30}\cdot\left(5 + 10 + 21 + 12 + 25 + 30\right)$$
$$= \frac{1}{30}\cdot 103$$
$$= 3,43$$

Definition: (geometrischer Mittelwert)

Seien $a_1, \ldots, a_m$ mit dem Erhebungsumfang n und $n, m \in \mathbb{N}$ die möglichen Merkmalsausprägungen einer metrischen Skala und $H_1, \ldots, H_m$ die zugehörigen absoluten Häufigkeiten. Dann ist der geometrische Mittelwert definiert als

$$\boxed{\overline{x}_{geom} := \sqrt[n]{a_1^{H_1} \cdot \dots \cdot a_m^{H_m}} = \sqrt[n]{\prod_{i=1}^{m} a_i^{H_i}}}.$$

Beispiel: Bei der Häufigkeitsverteilung aus Kapitel 2 erhalten wir also:

a_i	„1“	„2“	„3“	„4“	„5“	„6“
$H(a_i)$	5	5	7	3	5	5

$$\begin{aligned}\overline{x}_{geom} &= \sqrt[n]{\prod_{i=1}^{m} a_i^{H_i}} \\ &= \sqrt[30]{1^5 \cdot 2^5 \cdot 3^7 \cdot 4^3 \cdot 5^5 \cdot 6^5} \\ &= \sqrt[30]{1 \cdot 32 \cdot 2187 \cdot 64 \cdot 3125 \cdot 7776} \\ &= \sqrt[30]{1,09 \cdot 10^{14}} \\ &= 2,94\end{aligned}$$

Was sich bei diesem Beispiel bereits abzeichnet gilt auch allgemein.

Satz:
Für jede Häufigkeitsverteilung gilt:

$$\overline{x}_{geom} \leq \overline{x}.$$

3.3 Das positionsabhängige Lagemaß

Definition: (Median)
Sei $x_1, \dots, x_n$ eine größenmäßig sortierte Urliste mit metrischen Merkmalsausprägungen. Der Median ist dann definiert als

$$\boxed{\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{für } n \text{ ungerade} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{für } n \text{ gerade} \end{cases}}.$$

Beispiel: Bei der Häufigkeitsverteilung aus Kapitel 2 erhalten wir also:

a_i	„1“	„2“	„3“	„4“	„5“	„6“
$H(a_i)$	5	5	7	3	5	5

$n = 30$ ist gerade. $\frac{30}{2} = 15$. Wir suchen also in der sortierten Liste die Zahl an der 15. Stelle.

$$1,1,1,1,1,2,2,2,2,2,3,3,3,3,\underset{x_{15}}{3}, \underset{x_{16}}{3}, 3,4,4,4,5,5,5,5,5,6,6,6,6,6$$

$$\Rightarrow \tilde{x} = \frac{x_{15} + x_{16}}{2} = \frac{3+3}{2} = 3.$$

Bemerkung:
Der Vorteil des Median ist, dass evtl. Ausreißer nach oben oder unten im Gegensatz zu den anderen Lagemaßen nicht berücksichtigt werden. Genau dasselbe ist aber auch der Nachteil falls es sich nicht um Ausreißer handelt.

Aufgabe:
Berechne in den Häufigkeitsverteilungen aus Kapitel 2 alle Lagemaße.

3.4 Wann nutzen wir welchen Mittelwert?

Der Vorteil des Median liegt darin, dass eventuelle Ausreißer an den Rändern nicht berücksichtigt werden. Zeichnet es sich also bei der Betrachtung der Häufigkeitsverteilung ab, dass solche Ausreißer existieren, nutzen wir den Median, ansonsten nutzen wir den arithmetischen Mittelwert.
Der Modalwert wird meistens bei Merkmalsausprägungen nicht metrischer Skalen angewandt.

4. Streuungsmaße

Streuungsmaße geben an, wie stark die in der Urliste erfassten Merkmalsausprägungen um den Mittelwert streuen. Sie sind ein Maß für die mittlere Abweichung der einzelnen Merkmalsausprägungen vom gewählten Lagemaß.
Zur Vereinfachung werden wir hier nicht alle Streuungsmaße behandeln. Als Lagemaß wählen wir nun, ebenfalls zur Vereinfachung, den arithmetischen Mittelwert.

4.1 Die Hilfssummen

Für die beiden hier verwendeten Streuungsmaße ist es sinnvoll einen Zwischenschritt zur Berechnung durchzuführen. Hierzu führen wir die Hilfssummen

$$\boxed{s_{aa} := H_1 \cdot (a_1 - \bar{a})^2 + H_2 \cdot (a_2 - \bar{a})^2 + \cdots + H_m \cdot (a_m - \bar{a})^2} \ ,$$

bzw. in der bereits bekannten Kurzschreibweise

$$s_{aa} = \sum_{i=1}^{m} H_i \cdot (a_i - \bar{a})^2 \ ,$$

ein. Das m steht hierbei für die Anzahl der vorkommenden Merkmalsausprägungen. Im Index steht jeweils die für die verschiedenen Merkmalsausprägungen verwendete Variable. Wurden die Merkmalsausprägungen z.B. mit x_i bezeichnet, so wird die Hilfssumme mit s_{xx} bezeichnet.

4.2 Die Varianz

Die Varianz ist ein eigenständiges Streuungsmaß, wird aber meistens nur als Zwischenergebnis zur Berechnung des optimalen Streuungsmaßes, der Standardabweichung verwandt.

Definition: (Varianz)
Seien $a_1, \ldots, a_m$ die Merkmalsausprägungen auf einer metrischen Skala, $m \in \mathbb{N}$ und $n \in \mathbb{N}$ der Erhebungsumfang. Die Varianz V_a ist dann definiert als

$$\boxed{V_a := \frac{s_{aa}}{n}.}$$

Bemerkung:
Genauso wie bei den Hilfssummen steht im Index immer die Variable mit der die Merkmalsausprägungen bezeichnet wurden.

4.3 Die Standardabweichung

Definition: (Varianz)
Seien $a_1, \ldots, a_m$ die Merkmalsausprägungen auf einer metrischen Skala, $m \in \mathbb{N}$ und $n \in \mathbb{N}$ der Erhebungsumfang. Die Standardabweichung s_a oder s ist dann definiert als

$$\boxed{s_a := \sqrt{V_a}.}$$

Bemerkungen:
Genauso wie bei den Hilfssummen und der Varianz steht im Index immer die Variable mit der die Merkmalsausprägungen bezeichnet wurden. Sollte es hier nur eine Variable geben, kann diese auch weggelassen werden.

Satz:
Seien $a_1,\ldots,a_m$, $m \in \mathbb{N}$ und $n \in \mathbb{N}$ der Erhebungsumfang die Merkmalsausprägungen auf einer metrischen Skala. Dann gilt:

Ca. zwei Drittel der Merkmalsausprägungen befinden sich im Intervall $\left[\overline{a} - s_a; \overline{a} + s_a\right]$.

5. Statistischer Zusammenhang

5.1 Einführungsaufgabe

Sie arbeiten als BerufspraktikantIn in der Kindertagesstätte St. Christopherus der Stadt Euskirchen. Die Einrichtung nimmt seit sieben Jahren an dem Projektversuch „Spritze gegen Pusteln“ zu einer neuartigen Impfung gegen Windpocken für unter Dreijährige teil. Jedes Jahr wird genau erfasst, wie viele Kinder geimpft wurden und wie viele der geimpften Kinder trotzdem an Windpocken erkranken. Einmal geimpfte Kinder werden nicht noch einmal geimpft. Alle in den umliegenden Kindertagesstätten erfassten Ergebnisse werden von Ihnen gesammelt. Dieses Jahr, im achten Jahr, sollen die bisherigen Ergebnisse ausgewertet werden.

Es besteht die Gefahr, dass die Schutzimpfung bei den Geimpften zu einem Ausbruch der Windpocken führen könnte. Aufgrund Ihrer umfassenden mathematischen Kenntnisse werden Sie mit der statistischen Auswertung der Untersuchungsreihe betraut und sollen zum Schluss erklären, ob durch die Impfung tatsächlich bei den Geimpften die Masern ausbrechen können.

x: Anzahl der geimpften Kinder
y: Anzahl der im späteren Verlauf erkrankten Kinder

	2005	2006	2007	2008	2009	2010	2011	2012
x	50	35	40	23	60	40	30	55
y	23	19	21	10	32	20	17	25

Frage: Gibt es einen statistischen Zusammenhang zwischen der Impfung und der Erkrankung?

Bisher haben wir immer nur ein Merkmal untersucht. Nun haben wir zwei Merkmale, in diesem Beispiel die Anzahl der geimpften Kinder und die Anzahl der im späteren Verlauf erkrankten Kinder.

Einen solchen statistischen Zusammenhang zwischen zwei Merkmalen bezeichnet man als **Korrelation**. Prinzipiell kann der Zusammenhang ein beliebiger funktionaler sein. Wir beschränken uns hier jedoch auf einen **linearen Zusammenhang**, d.h. die Frage, **ob die gegebenen Daten annähernd auf einer Geraden liegen**.

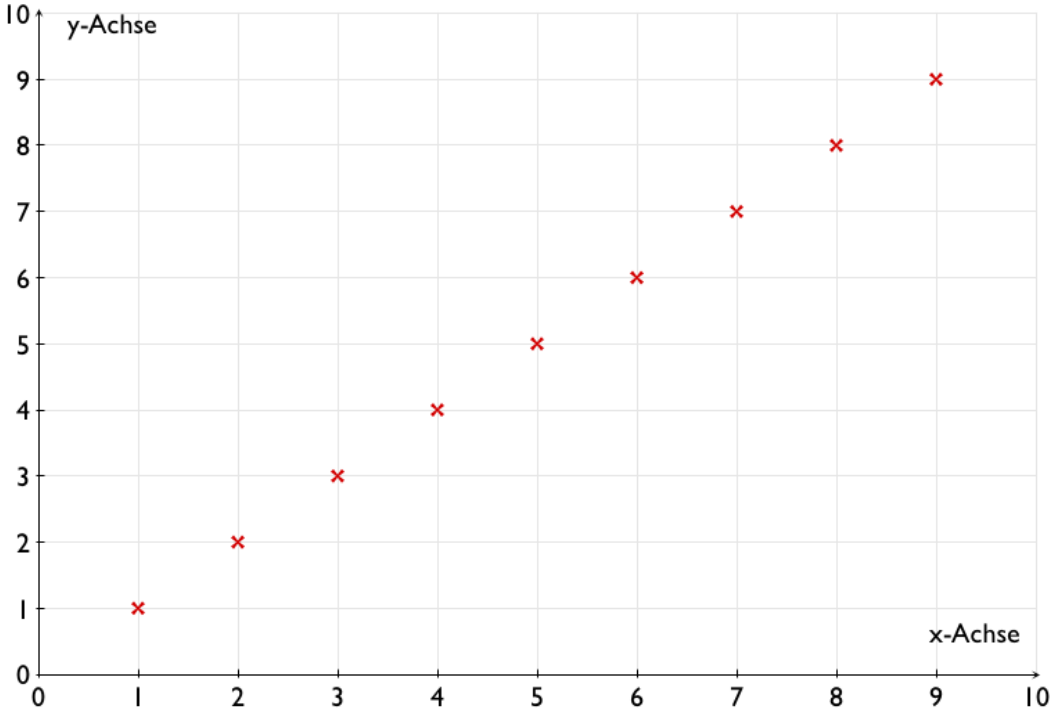
VORSICHT:
Da hier der Zusammenhang der MermalsausprägungsPAARE untersucht wird, dürfen die einzelnen WertePAARE N I C H T getrennt werden. Die Darstellung der Tabelle in Form einer Häufigkeitsverteilung ist also N I C H T sinnvoll!

Es ist sofort einsehbar, dass unsere bisherigen Untersuchungsmethoden für diese Fragestellung nicht ausreichend sind.

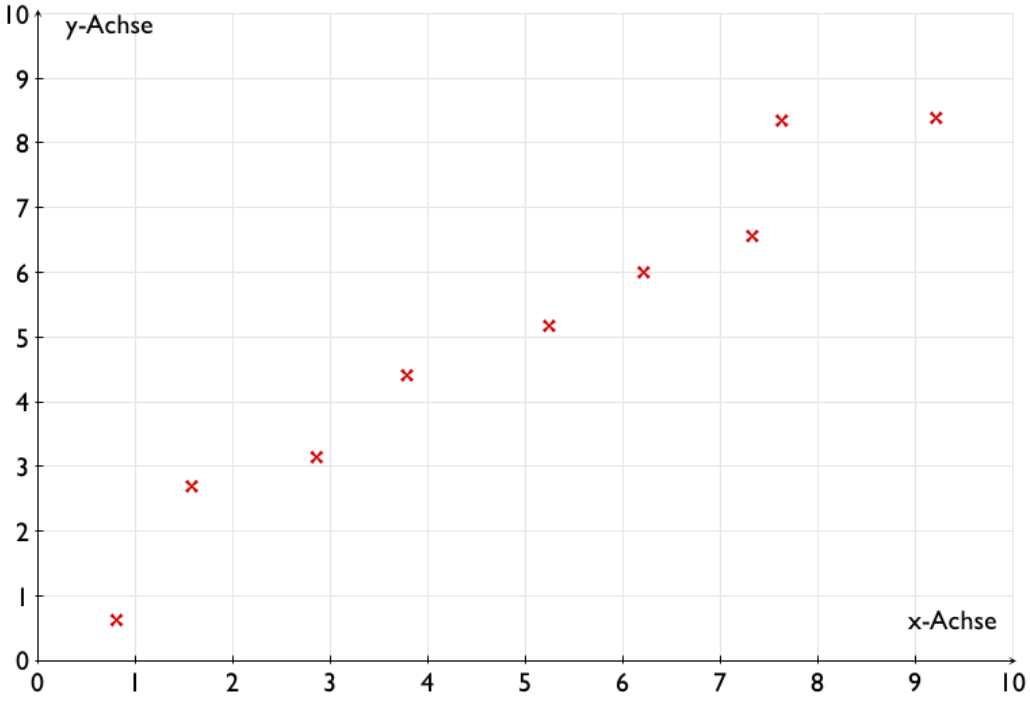
5.2 Graphische Untersuchung des statistischen Zusammenhangs

Das einfachste Verfahren ist die graphische Untersuchung des statistischen Zusammenhanges. Hierzu tragen wir die gegebenen Daten als Punkte in ein Koordinatensystem ein. Das entstehende Muster bezeichnet man auch als **Punktwolke**. Anschließend schauen wir uns die Punktwolke genauer an und ordnen sie einer der folgenden Kategorien zu:

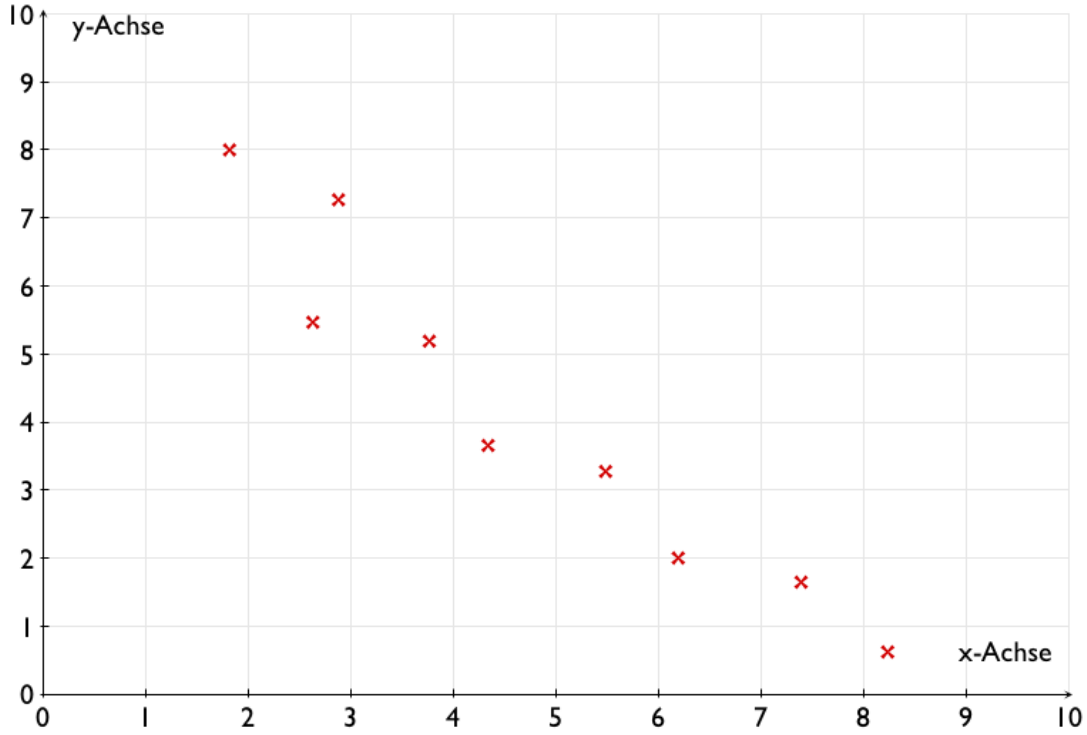
1.) Die Werte sind **total korreliert**, d.h. alle Werte liegen exakt auf einer Geraden.



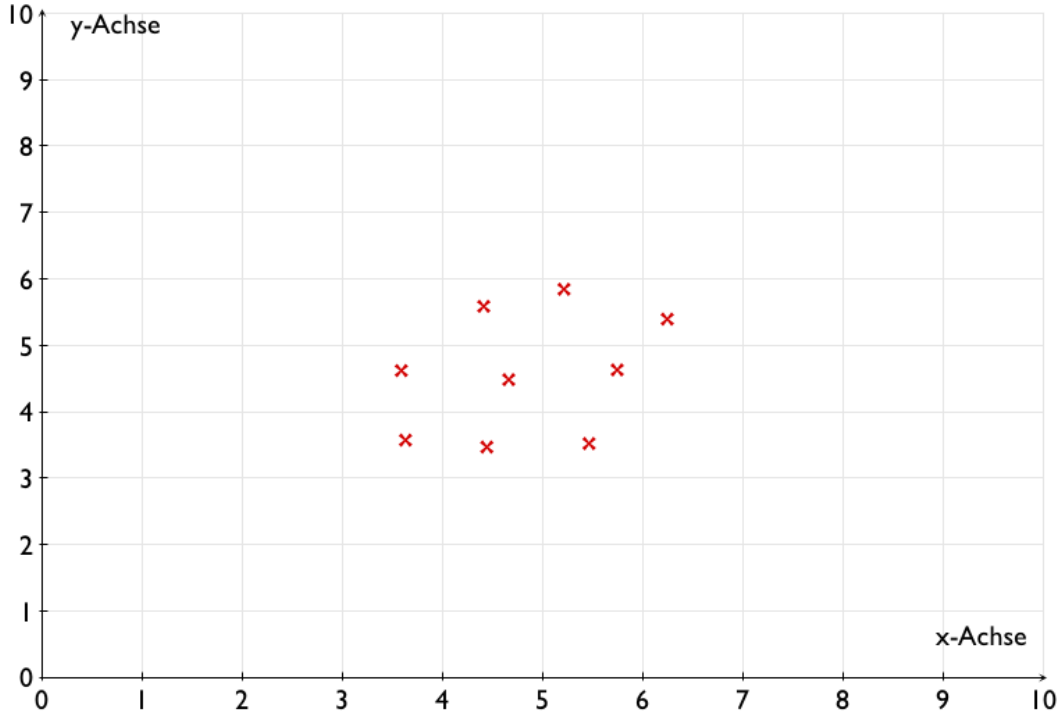
2.) Die Werte sind **stark korreliert**, d.h. alle Punkte sind so verteilt, dass sie um eine Gerade herum streuen.



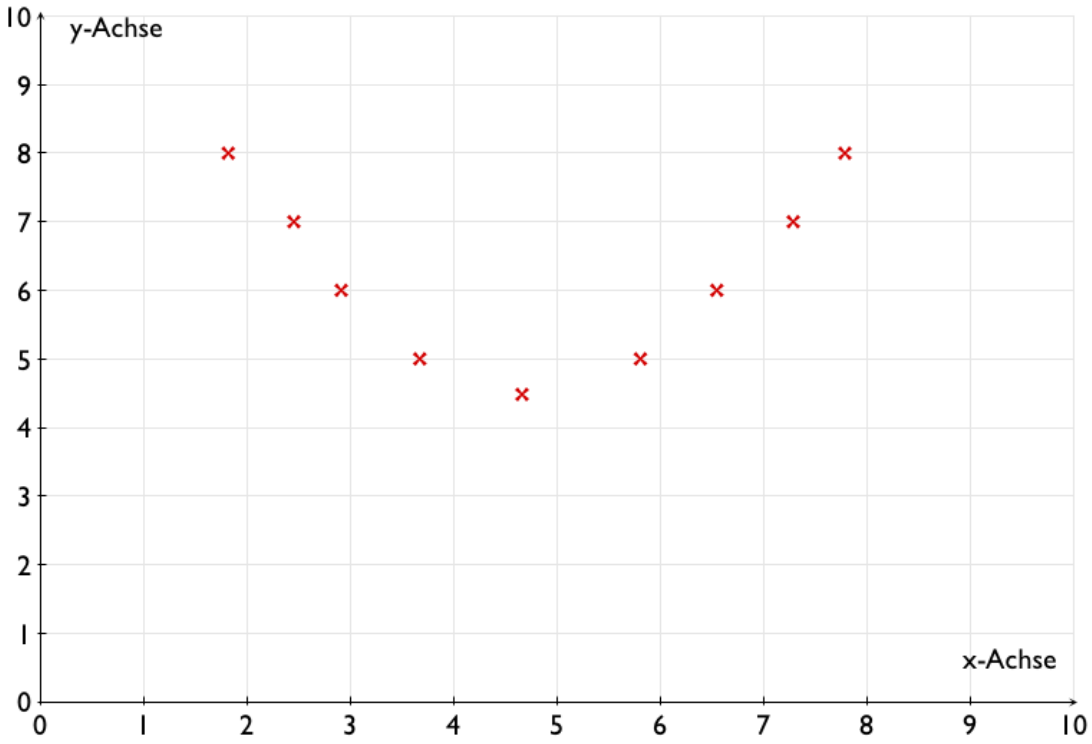
bzw.



3.) Die Werte sind **unbestimmbar** oder **schwach korreliert**, d.h. die Punkte bilden kaum eine Streuung um eine Gerade. Bei der später folgenden rechnerischen Untersuchung entstehen hier zwei verschiedene Fälle, unbestimmbar und schwach korreliert, die sich jedoch von der Punktwolke her kaum unterscheiden lassen.



4.) Die Werte sind **nicht linear korreliert**, d.h. es ist ein Zusammenhang erkennbar der sich nicht durch eine Gerade darstellen lässt. Dieser Fall hat für die rechnerische Untersuchung keine Bedeutung.



Aufgabe:
Untersuchen Sie den graphisch den statistischen Zusammenhang der beiden Merkmale aus der Einführungsaufgabe 5.1!

5.3 Rechnerische Untersuchung des statistischen Zusammenhangs

Die rechnerische Untersuchung erfolgt mit Hilfe des Korrelationskoeffizienten.

Definition: (Korrelationskoeffizient)

Seien (x_i, y_i) für $i = 1, \ldots, n$ die zusammenhängenden Merkmalsausprägungen zweier Merkmale und $n \in \mathbb{N}$ der Erhebungsumfang. Die Zahl

$$\boxed{r := \frac{s_{xy}}{\sqrt{s_{xx} \cdot s_{yy}}}}$$

heißt **Korrelationskoeffizient**, mit den Hilfssummen

$$s_{xy} = (x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + (x_2 - \bar{x}) \cdot (y_2 - \bar{y}) + \cdots + (x_n - \bar{x}) \cdot (y_n - \bar{y})$$

$$s_{xx} = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

$$s_{yy} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2,$$

wobei $\bar{x}$ und $\bar{y}$ hier jeweils den arithmetischen Mittelwert bezeichnen.

Satz: Es gilt immer:

$$-1 \leq r \leq 1.$$

Aus dem Wert des Korrelationskoeffizienten lässt sich nun die Korrelation wie folgt ablesen:

1.) **Total korreliert**: $r=-1$ und $r=1$
2.) **Stark korreliert**: $-1<r<-0,8$ und $0,8<r<1$
3.) **Unbestimmbar korreliert**: $-0,8\le r<-0,5$ und $0,5<r\le 0,8$
4.) **Schwach Korreliert**: $-0,5\le r\le 0,5$

Nur in den ersten beiden Fällen lässt sich eine Gerade durch die Punktwolke ziehen. Eine nicht lineare Korrelation fiele unter den dritten oder vierten Fall.

Beispiel: Schauen wir uns die Rechnung mit den folgenden Wertepaaren mal genauer an:

x_i	1	2	3	4	5
y_i	5	15	10	20	25

Die Berechnung der arithmetischen Mittelwerte ergibt:

$$\overline{x}=\frac{1+2+3+4+5}{5}=3$$

$$\overline{y}=\frac{5+15+10+20+25}{5}=15$$

Es hat sich als sehr sinnvoll erwiesen die obige Tabelle nun um die vier Zeilen $x_i-\overline{x}$, $\left(x_i-\overline{x}\right)^2$, $y_i-\overline{y}$ und $\left(y_i-\overline{y}\right)^2$ zu ergänzen.

x_i	1	2	3	4	5
$x_i-\overline{x}$	-2	-1	0	1	2
$\left(x_i-\overline{x}\right)^2$	4	1	0	1	4
y_i	5	15	10	20	25
$y_i-\overline{y}$	-10	0	-5	5	10
$\left(y_i-\overline{y}\right)^2$	100	0	25	25	100

Nun können wir ganz leicht die benötigten Hilfssummen berechnen.

$$\begin{aligned}s_{xy}&=\left(x_1-\overline{x}\right)\cdot\left(y_1-\overline{y}\right)+\left(x_2-\overline{x}\right)\cdot\left(y_2-\overline{y}\right)+\cdots+\left(x_n-\overline{x}\right)\cdot\left(y_n-\overline{y}\right)\\&=(-2)\cdot(-10)+(-1)\cdot 0+0\cdot(-5)+1\cdot 5+2\cdot 10=45\end{aligned}$$

$$\begin{aligned}s_{xx}&=\left(x_1-\overline{x}\right)^2+\left(x_2-\overline{x}\right)^2+\cdots+\left(x_n-\overline{x}\right)^2\\&=4+1+0+1+4=10\end{aligned}$$

$$\begin{aligned}s_{yy}&=\left(y_1-\overline{y}\right)^2+\left(y_2-\overline{y}\right)^2+\cdots+\left(y_n-\overline{y}\right)^2\\&=100+0+25+25+100=250\end{aligned}$$

Wir erhalten also den Korrelationskoeffizienten

$$r=\frac{s_{xy}}{\sqrt{s_{xx}\cdot s_{yy}}}=\frac{45}{\sqrt{10\cdot 250}}=0,9.$$

Die Wertepaare sind also stark korreliert.

Aufgabe 1:

Untersuchen Sie den rechnerisch den statistischen Zusammenhang der beiden Merkmale aus der Einführungsaufgabe 5.1!

Aufgabe 2:
Untersuchen Sie den statistischen Zusammenhang der folgenden Tabellen graphisch und rechnerisch!

a)

x_i	4	5	4,5	5,25	4	5,5	6	6	5,5	4,5
y_i	3	3,5	2,5	2,75	3,5	4,5	3	4	4,5	3,5

b)

x_i	1	3	6	2	4	7	8	5	10	9
y_i	1,5	2,5	4	2	3	4,5	5	3,5	6	5,5

c)

x_i	10	1	9	2	8	3	7	5	6	4
y_i	6	0,5	4,5	2	6	3,5	4,5	3,5	4,5	2,5

5.4 Lineare Regression

Sind die Werte total oder stark korreliert, so lässt sich eine gerade mit einer möglichst geringen Abweichung von den einzelnen Wertepaaren durch die Punktwolke legen.

Satz: Seien die Wertepaare $(x_i | y_i)$ entweder stark oder total korreliert. Dann hat die Gerade

$$\boxed{y = \frac{s_{xy}}{s_{xx}}(x - \overline{x}) + \overline{y}}$$

mit den arithmetischen Mittelwerten $\overline{x}$ und $\overline{y}$ sowie den Hilfssummen

$$s_{xy} = (x_1 - \overline{x}) \cdot (y_1 - \overline{y}) + (x_2 - \overline{x}) \cdot (y_2 - \overline{y}) + \cdots + (x_n - \overline{x}) \cdot (y_n - \overline{y}) \text{ und}$$

$$s_{xx} = (x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2$$

die kleinste durchschnittliche Abweichung von den gegebenen Wertepaaren.

Bemerkung:
Dieser Satz ermöglicht uns drei verschiedene Dinge:

1. Das Aufstellen der Geradengleichung,
2. die Vorhersage eines y – Wertes bei einem gegebenen x – Wert und
3. die Vorhersage eines x – Wertes bei einem gegebenen y – Wert.

Beispiel: Schauen wir uns diese drei Rechnungen einmal mit der Beispielaufgabe aus 5.3 an.

1. Aufstellen der Geradengleichung
   Aus 5.3 wissen wir bereits:
   $\overline{x} = 3,$
   $\overline{y} = 15,$
   $s_{xy} = 45$ und
   $s_{xx} = 10.$
   Einsetzen in die Formel liefert nun die gesuchte Geradengleichung

$$y = \frac{s_{xy}}{s_{xx}}(x - \overline{x}) + \overline{y} = \frac{45}{10}(x - 3) + 15 = 4,5x - 13,5 + 15 = 4,5x + 1,5\,.$$

2. Berechnung von y
   Hierbei muss immer ein Wert für x vorgegeben werden. Als Beispiel wähle ich nun $x = 100$.
   Aus 5.3 wissen wir bereits:
   $\overline{x} = 3,$
   $\overline{y} = 15,$
   $s_{xy} = 45$ und
   $s_{xx} = 10.$
   Da die Formel bereits die richtige Form besitzt, müssen wir diese Werte nur noch einsetzen.
   $$y = \frac{s_{xy}}{s_{xx}}(x - \overline{x}) + \overline{y} = \frac{45}{10}(100 - 3) + 15 = 4,5 \cdot 97 + 15 = 451,5$$
   Der gesuchte y – Wert ist also 451,5.
3. Berechnung von x
   Hierbei muss immer ein Wert für y vorgegeben werden. Als Beispiel wähle ich nun $y = 50$.
   Aus 5.3 wissen wir bereits:
   $\overline{x} = 3,$
   $\overline{y} = 15,$
   $s_{xy} = 45$ und
   $s_{xx} = 10.$
   Im nächsten Schritt müssen wir nun die o.a. Formel nach x auflösen.

$$y = \frac{s_{xy}}{s_{xx}}(x - \overline{x}) + \overline{y} \qquad \left| -\overline{y} \right.$$

$$\Leftrightarrow \quad y - \overline{y} = \frac{s_{xy}}{s_{xx}}(x - \overline{x}) \qquad \left| : \frac{s_{xy}}{s_{xx}} \right.$$

$$\Leftrightarrow \quad \frac{s_{xx}}{s_{xy}}(y - \overline{y}) = x - \overline{x} \qquad \left| + \overline{x} \right.$$

$$\Leftrightarrow \frac{s_{xx}}{s_{xy}}(y - \overline{y}) + \overline{x} = x$$

   Nun können wir die beiden Seiten des Gleichheitszeichens vertauschen und die o.a. Werte einsetzen.
   $$x = \frac{s_{xx}}{s_{xy}}(y - \overline{y}) + \overline{x} = \frac{10}{45}(50 - 15) + 3 = \frac{10}{45} \cdot 35 + 3 = 10,78$$
   Der gesuchte x – Wert ist also 10,78.

<u>Aufgabe 1:</u>

a) Stellen Sie die Gleichung der Regressionsgerade für das Problem aus der Einführungsaufgabe 5.1 auf!
b) In einer weiteren Stadt werden insgesamt 125 Kinder geimpft. Berechnen Sie, wie viele voraussichtlich im weiteren Verlauf erkranken werden!
c) In einer anderen Stadt sind nach der Impfung 100 Kinder erkrankt. Berechnen Sie, wie viele Kinder geimpft wurden!

Aufgabe 2:

a) Berechnen Sie die Regressionsgeraden für die total oder stark korrelierten Wertepaare aus Aufgabe 2 in 5.3!
b) Berechnen Sie den zu $y = 15$ gehörenden x – Wert!
c) Berechnen Sie den zu $x = 15$ gehörenden y – Wert!