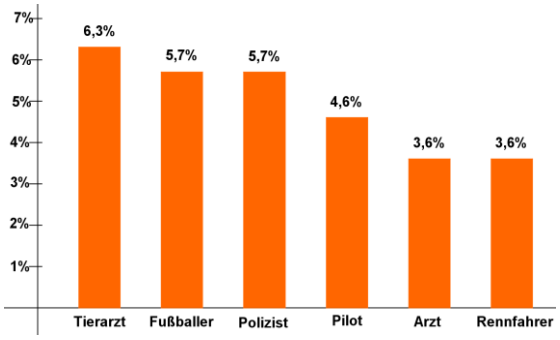
DATENANALYSE III

Das Histogramm

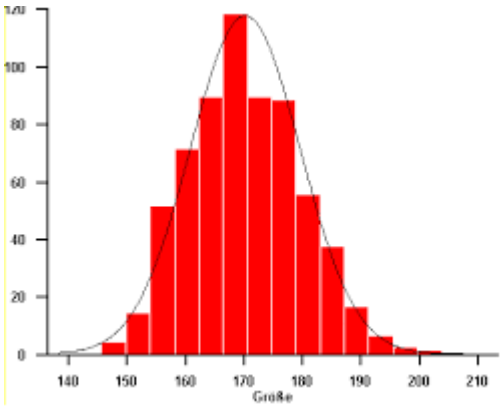
Merkmale (Ausprägungen) werden hier in Klassen eingeteilt. Dargestellt werden dann die Häufigkeiten in den einzelnen Klassen. Für jede Klasse wird ein Rechteck gezeichnet, dessen Höhe der Anzahl der Beobachtungen entspricht.

Die Rechtecke eine Histogramms werden vielfach durch die Seitenlinien getrennt, also ohne Zwischenraum wie beim Säulendiagramm.

Säulendiagramm mit Zwischenräumen (diskrete Daten)



Histogramm ohne Zwischenräume (stetige Daten)



Erstellen eines Histogramms

- Die Anzahl k von Klassen festlegen, idealerweise zwischen 5 und 20. Als Grössenordnung für die Klassenanzahl kann die Formel $k \approx \sqrt{n}$ berücksichtigt werden, wobei n für den Stichprobenumfang steht.
- Die Klassen müssen durch die Klassengrenzen exakt definiert sein. Die Klassengrenzen sollten möglichst «einfach lesbar» sein. Normalerweise haben alle Klassen eines Histogramms die gleiche Breite.
- Die absoluten Häufigkeiten (= die Anzahl Beobachtungen in jeder Klasse) bestimmen.
- Die absolute Häufigkeit jeder Klasse wird als Rechteck dargestellt, dessen Höhe proportional zur Häufigkeit ist.
- Die Grafik wird dem Inhalt und Kontext entsprechend beschriftet und bezeichnet.

Im Folgenden werden wir ein Histogramm für zwanzig gemessene Körpergrössen erstellen:

Beispiel zum Histogramm: Körpergrösse (→ Arbeitsblatt austeilen)

Histogramme sehen ähnlich aus wie Säulen- oder Balkendiagramme, werden aber für stetige statt diskrete Daten verwendet. Um ein Histogramm zu zeichnen, muss man seine Daten zuerst klassieren, d.h. Klassen bilden und sie ihnen zuordnen. Am einfachsten sind Histogramme zu zeichnen, wenn diese Klassen gleich breit sind.

Misst man zum Beispiel die Körpergröße von 20 Personen, könnte man diese Klassen in 10cm-Abständen bilden, also von 150-159cm, von 160-169cm, und so weiter.

Veranschaulichung des Erstellens eines Histogramms für zwanzig Körpergrößen:

172 164 160 162 173 180 158 185 171 181 162 184 177 175 177 174 158 151 192 177

Zuerst müssen die Klassen festgelegt werden, in welche die Daten zugeordnet werden.
Wir wählen fünf gleich breite Klassen mit entsprechenden Intervallen:

Klasse	***1***	***2***	***3***	***4***	***5***
Intervall	[150, 160)	[160, 170)	[170, 180)	[180, 190)	[190, 200)

Die eckigen und runden Klammern beschreiben die jeweiligen Grenzen des Intervalls. In der zweiten Spalte ist z.B. die 160 enthalten, da davor eine eckige Klammer steht, aber die 170 ist nicht enthalten, da dort eine runde Klammer ist. Wenn also jemand genau 170cm groß ist, fällt er in die dritte Klasse. Falls jemand 169.8cm groß ist, fällt er in die zweite Klasse.

Jetzt zählen sie, wie viele Personen in jede Klasse fallen. Es gibt z.B. drei Personen in der Klasse von 150 (einschließlich) bis 160 (ausschließlich). Die Höhe (wird auch Dichte genannt) der einzelnen Balken, berechnen wir wie folgt:

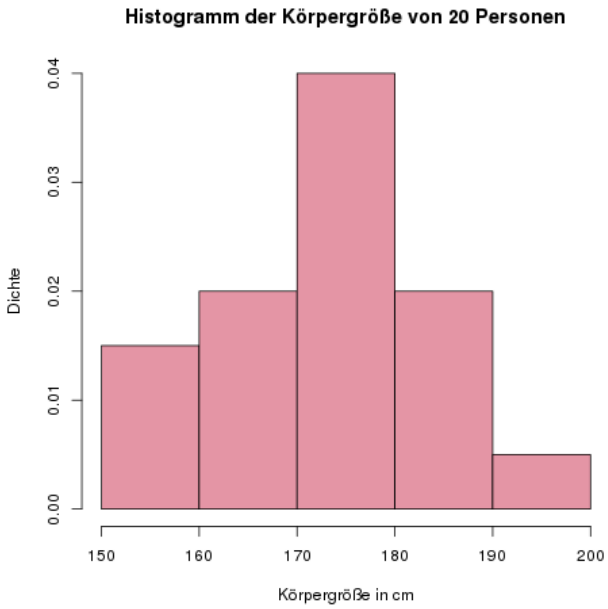
$$hi = \frac{a}{N*b}$$

Hier ist *hi* die Höhe des i-ten Histogrammbalkens, *a* ist die Anzahl der Personen in dieser Klasse i, *N* ist die Gesamtzahl an Personen (bei uns N=20), und *b* ist die Breite der i-ten Klasse (bei uns sind alle Klassen gleich breit, also *b=10* für alle Klassen). In der ersten Klasse ist die Höhe zum Beispiel *h1 = 320 * 10 = 0.015 (entspricht der relativen Häufigkeit der Beobachtung).*

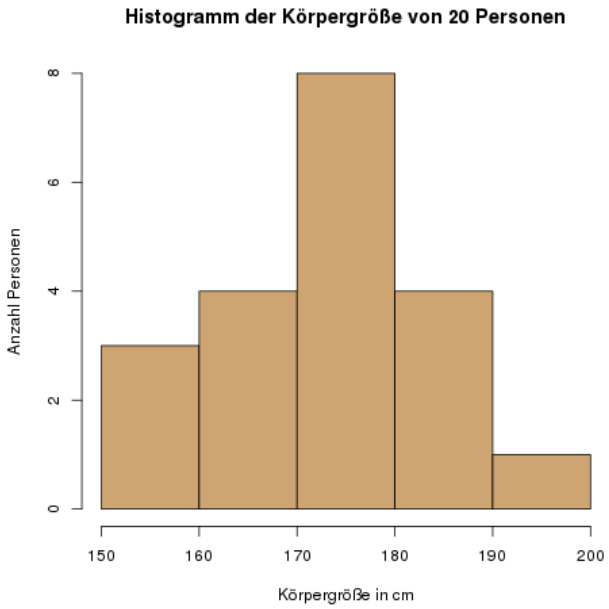
→ Am einfachsten erstellt man all diese Daten in einer Tabelle:

Klasse i	1	2	3	4	5
Intervall	[150, 160)	[160, 170)	[170, 180)	[180, 190)	[190, 200)
Anzahl Personen in dieser Klasse *ni*	3	4	8	4	1
Histogrammhöhe *hi* (Dichte: relative Häufigkeit)	0.015	0.02	0.04	0.02	0.005

Damit kann man das Histogramm zeichen:



Die Daten für das linke Histogramm haben wir gerade berechnet.

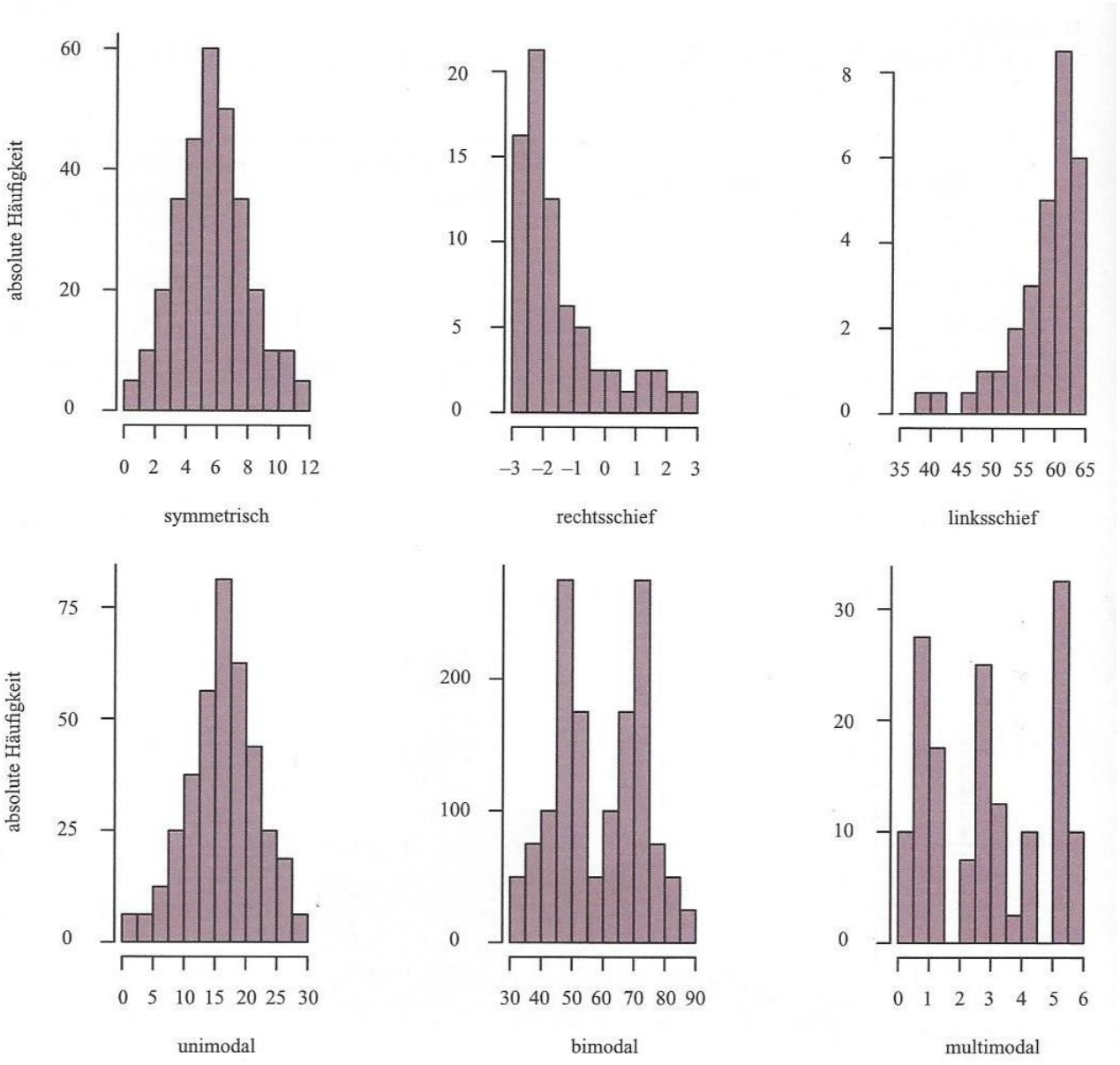


Das rechte unterscheidet sich nur darin, dass auf der y-Achse absolute Häufigkeiten verwendet wurden—es wurden also statt den Höhen hi die Anzahl an Personen, ni gezeichnet.

Man sieht im linken Histogramm also direkt, dass in der mittleren Klasse 8 Personen liegen.

Der Anteil an Beobachtungen in jeder Klasse entspricht nun der Fläche dieser Balken. In der ersten Klasse ist ein Anteil von $10 * 0.015 = 0.15$, also 15% der Daten, was bei 20 Personen genau 3 Personen entspricht.

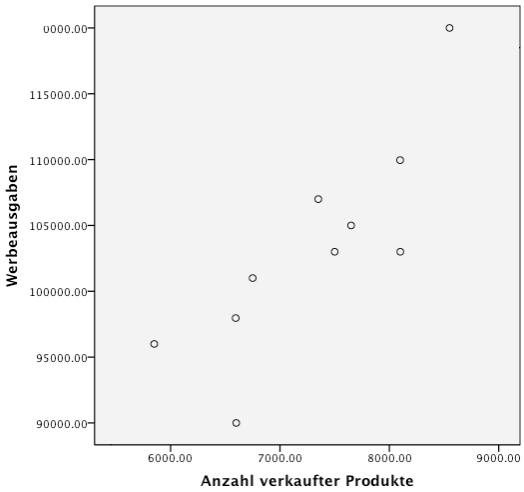
Mithilfe von Histogrammen können Verteilungen charakterisiert werden. Es wird einerseits zwischen **symmetrischen**, **rechtsschiefen** und **linksschiefen** andererseits zwischen **unimodalen**, **bimodalen** und **multimodalen** Verteilungen unterschieden.
Rechtsschiefe Verteilungen haben einen Ausläufer nach rechts, linksschiefe Verteilungen haben den Ausläufer nach links.
Mit den Begriffen uni-, bi- und multimodal wird die Anzahl Gipfel/Häufungspunkte im Histogramm beschrieben. Das Histogramm einer unimodalen Verteilung hat nur einen Gipfel. Bei einer bimodalen Verteilung sind es zwei Gipfel und das Histogramm einer multimodalen Verteilung hat mehrere Gipfel.

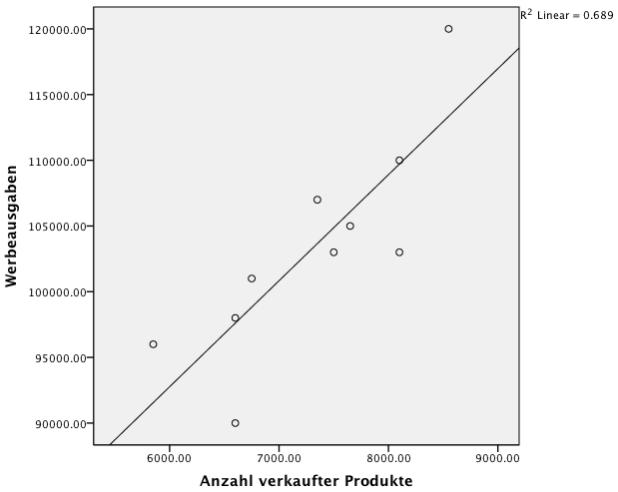


Das Streudiagramm

Es können zwei quantitative Merkmale dargestellt werden. Ein Punkt im Streudiagramm entspricht einer Beobachtung von diesen zwei Merkmalen.

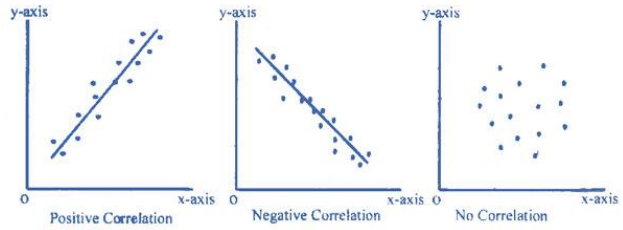
Es können lineare Zusammenhänge zwischen diesen Merkmalen dargestellt werden. Diesem Zusammenhang sagt man Korrelation. Die Häufung der Punkte wird durch eine Gerade (lineare Funktion) ersetzt, welche die Gesamtheit der Punkte möglichst gut wiedergeben soll.





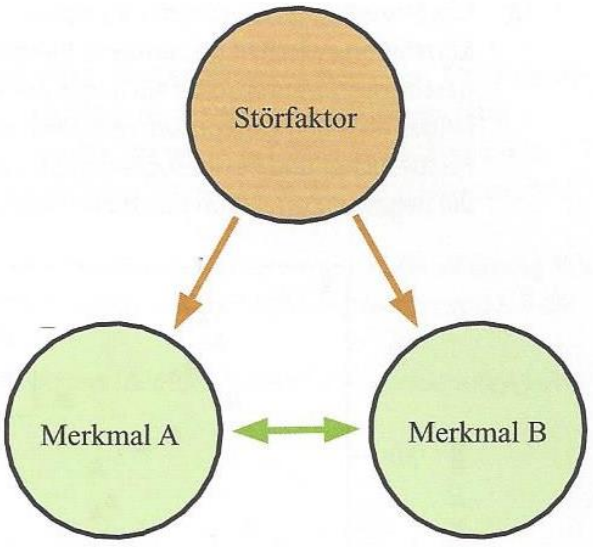
<u>Positive Korrelation</u>
Grosse Ausprägung des Merkmals A (x-Achse) sowie grosse Ausprägung des Merkmals B (y-A)

<u>Negative Korrelation</u>
Grosse Ausprägung des Merkmals A bei kleiner Ausprägung des Merkmals B (oder umgekehrt).



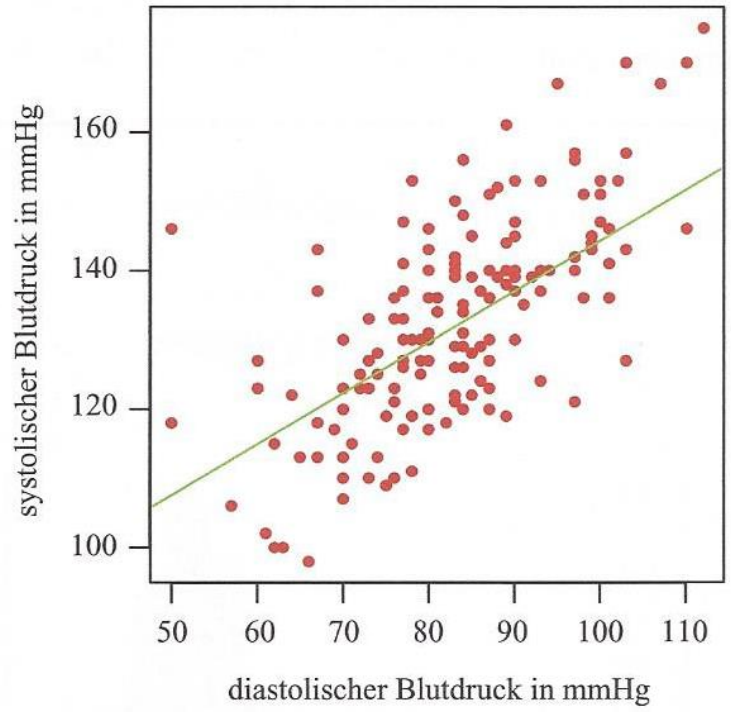
Wir beschränken uns jedoch auf die rein graphische Darstellung der Beobachtungen, also ohne Korrelationsfunktion.

Wird zwischen zwei quantitativen Merkmalen eine Korrelation festgestellt, bedeutet das nicht, dass eine Änderung in der einen Variablen eine Änderung in der anderen Variablen verursacht. Aus einer Korrelation darf also nicht auf einen **kausalen Zusammenhang** geschlossen werden. Ausser in einem geplanten Experiment kann nie ausgeschlossen werden, dass die Korrelation zwischen den Merkmalen A und B durch ein drittes Merkmal, einen sogenannten **Störfaktor**, verursacht wird, der sowohl Merkmal A als auch Merkmal B beeinflusst.



2 Beispiele

Im Beispiel 21.6, Übergewicht und Blutdruck, ist ein Streudiagramm mit den Variablen BMI und systolischer Blutdruck abgebildet. In diesem Diagramm lässt sich keine Korrelation erkennen. Im folgenden Streudiagramm mit den Variablen diastolischer und systolischer Blutdruck ist eine positive Korrelation erkennbar. Mit grüner Farbe ist diejenige Gerade eingezeichnet, um welche sich die Punkte nach bestimmten mathematischen Kriterien am besten streuen.



Das Streudiagramm, welches im Beispiel 21.12, Bierfest, abgebildet ist, lässt eine negative Korrelation zwischen getrunkener Biermenge und Haarlänge vermuten. Es darf nun nicht geschlossen werden, dass ein hoher Bierkonsum die Ursache für kurzes Haar ist oder gar dass langes Haar den Bierkonsum reduziert. Beim unten stehenden Streudiagramm wurden für Festbesucher und Festbesucherinnen unterschiedliche Symbole gewählt. Offensichtlich wird die negative Korrelation durch den Störfaktor Geschlecht verursacht.