Statistik

Rainer Hauser

Dezember 2012

1 Einleitung

1.1 Population und Merkmale

Gegeben ist eine *Population* (oder Grundgesamtheit), und die Frage ist, welche *Elemente* dieser Population ein bestimmtes *Merkmal* besitzen. Das ist eine typische Aufgabe aus dem Gebiet der Statistik.

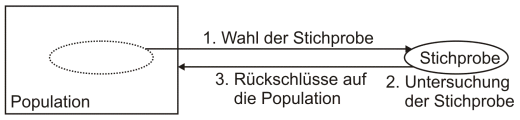
Beispiel:
Man möchte wissen, wie gross der prozentuale Anteil der Rotgrünfarbenblinden in der Bevölkerung der Schweiz ist. Die Population ist also die Schweizer Bevölkerung und das Merkmal die Farbenblindheit.

Wird nur ein Merkmal betrachtet, nennt man die Analyse *univariat*. Werden hingegen zwei Merkmale, hier dann meist *Variablen* genannt, gleichzeitig zusammen mit den Zusammenhängen zwischen ihnen untersucht, ist sie *bivariat* und bei mehr als zwei Merkmalen oder Variablen *multivariat*.

Beispiel:
Möchte man Körpergrösse und Körpergewicht der erwachsenen Schweizer Bevölkerung untersuchen, so macht es Sinn, die beiden Merkmale nicht unabhängig zu evaluieren, sondern auch zu analysieren, ob es zwischen den beiden Merkmalen einen Zusammenhang gibt.

1.2 Wahl einer Stichprobe

Häufig ist die Population zu umfangreich, um alle Elemente auf das Merkmal oder die Merkmale zu untersuchen. Man wählt deshalb eine geeignete *Stichprobe* aus und betrachtet nur die Elemente dieser Stichprobe. Die Schritte sind gemäss nebenstehender Abbildung:



1. Wahl der Stichprobe (sammelnde Statistik):
   Zufällige Stichproben sollen ein möglichst genaues Bild der gesamten Population geben und müssen deshalb sorgfältig ausgewählt werden, damit sie *repräsentativ* sind. (Geht es um die gesamte Bevölkerung der Schweiz, müssen zum Beispiel Frauen und Männer zu je 50% vertreten sein und das Verhältnis von Stadt und Land muss korrekt berücksichtigt werden.)
2. Untersuchung der Stichprobe (beschreibende Statistik):
   Das gesammelte Datenmaterial muss erst untersucht, bearbeitet und verständlich dargestellt werden, bevor man Schlüsse ziehen kann.
3. Rückschlüsse auf die Population (beurteilende Statistik):
   Rückschlüsse von der Stichprobe auf die gesamte Population sind immer mit einer Unsicherheit behaftet, auch wenn die Stichprobe noch so sorgfältig ausgewählt worden ist.

Im Folgenden wird im Wesentlichen die beschreibende Statistik behandelt.

Beispiel:
Bei den amerikanischen Präsidentschaftswahlen 1936 hat die Zeitschrift "Literary Digest" nach Auswertung von zwei Millionen Fragebögen einen Sieg von Landon vorausgesagt, während George Gallup nach Befragung von nur ein paar Tausend Leuten einen Sieg von Roosevelt prognostizierte. Gallup lag richtig,

weil er die Stichprobe sorgfältig auswählte, während "Literary Digest" die Fragebögen an Personen verschickte, die ein Auto oder Telefon besassen, was zu jener Zeit kurz nach dem Börsencrash 1929 nur für eine privilegierte Oberschicht zutraf. Es kommt also nicht in erster Linie auf die Grösse der Stichprobe an, sondern darauf, wie repräsentativ sie ist.

2 Sichtung der Daten

2.1 Quantitative und qualitative Daten

Das Feststellen eines Merkmals der Elemente einer Stichprobe führt zu *Daten*. Werden die Merkmale in Form von Wörtern, nicht aber in Form von Zahlen angegeben, so spricht man von *qualitativen* Daten. Im Gegensatz dazu liegen *quantitative* Daten in Form von Zahlen vor, wobei *diskrete* Daten als exakte Werte angegeben werden, die häufig von einem Zählvorgang stammen, während *kontinuierliche* Daten von einem Messvorgang stammen, sodass sie immer mit einem Messfehler behaftet sind. Qualitative Daten werden oft durch willkürliche Zuordnung von Zahlenwerten in quantitative, diskrete Daten umgewandelt.

Beispiel:
Die Anzahl Autoverkäufe in einer Garage in einem bestimmten Monat ist ein diskretes Merkmal, während die Körpergrösse der Rekruten eines Jahrgangs ein kontinuierliches Merkmal ist. Beschreibt man das Merkmal Haarfarbe mit Wörtern wie blond, braun und so weiter, so ist das ein qualitatives Merkmal, das man mit einer willkürlichen Zuordnung wie 1 für blond, 2 für braun und so weiter künstlich in ein diskretes, quantitatives Merkmal umwandeln kann.

2.2 Darstellung von Daten

Daten in Zahlenform sind meist schwierig zu erfassen. Sie lassen sich aber zur besseren Verständlichkeit auf vielfältige Weise *graphisch* darstellen. Tabellenkalkulationsprogramme bieten die gängigsten Darstellungsformen wie Pie Charts, Bar Charts oder Punkt- und Liniendiagramme an.

Wenn sie jedoch in Zahlenform vorliegen, werden Daten meist in *Häufigkeitstabellen* aufgelistet, in denen Merkmal und Anzahl Elemente mit diesem Merkmal in Tabellenform einander zugeordnet werden. Bei kontinuierlichen Datenmengen oder bei unübersichtlich grossen diskreten Datenmengen macht es meist Sinn, sie in *Klassen* einzuteilen.

Beispiel:
Weil die Körpergrösse ein kontinuierliches Merkmal ist, kann man die bei einem Jahrgang Rekruten gemessenen Werte in Klassen einteilen. So könnten alle Grössen bis und mit 160 cm in eine erste, die Grössen grösser als 160 cm bis und mit 175 cm in eine zweite, die Grössen ab 175 cm bis und mit 190 cm in eine dritte und alle Werte darüber in eine vierte Klasse eingeteilt werden. Die Häufigkeitstabelle

bis 160 cm	161 − 175 cm	176 − 190 cm	ab 191 cm
52	219	376	121

beschreibt mögliche Daten für $52 + 219 + 376 + 121 = 768$ Rekruten einer Einheit.

Die so bestimmten Grössen nennt man *absolute* Häufigkeiten. Diese Werte werden manchmal zu *kumulativen* Häufigkeiten zusammengezählt. Das ergibt

bis 160 cm	bis 175 cm	bis 190 cm	alle Grössen
52	271	647	768

für die Rekruten der obigen Einheit. In einer Tabelle, welche die Autoverkäufe einer Garage für den jeweiligen Monat angibt, sind kumulative Häufigkeiten üblich, weil man so immer die bisher im Jahr getätigten Verkäufe verfügbar hat.

Statt absoluten Häufigkeiten kann man auch *relative* Häufigkeiten aufzeigen. Wenn die i-te Klasse n_i Elemente umfasst und der Umfang der Stichprobe n beträgt, so ist $\frac{n_i}{n}$ die relative Häufigkeit dieser Klasse. Für die Rekruten im obigen Beispiel ergibt das

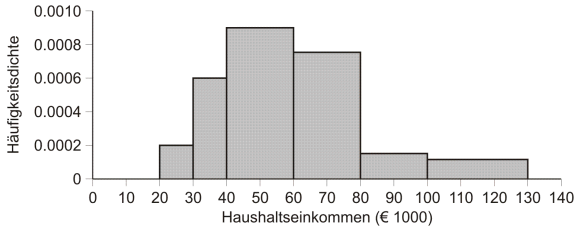
	bis 160 cm	161 − 175 cm	176 − 190 cm	ab 191 cm
absolut	52	219	376	121
relative	0.068	0.285	0.490	0.158

für die absoluten und relativen Häufigkeiten.

Klasseneinteilungen mit absoluten (oder relativen) Häufigkeiten werden häufig in Form von *Histogrammen* dargestellt. Sind die Klassenbreiten alle gleich gross, so sind Histogramme einfach, andernfalls müssen die verschiedenen Klassenbreiten berücksichtigt werden. In beiden Fällen die *Häufigkeitsdichte* durch $\frac{f}{w}$ definiert, wobei f die Häufigkeit und w die Klassenbreite ist.

Beispiel:
Bei einer Befragung haben 48 Personen Angaben über das Einkommen ihres Haushaltes gemacht.

Einkommen (in €)	Haushalte
20 000 − 30 000	2
30 000 − 40 000	6
40 000 − 60 000	18
60 000 − 80 000	15
80 000 − 100 000	3
100 000 − 130 000	4

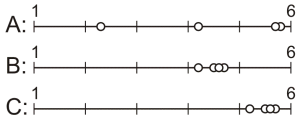


3 Lage- und Streuparameter

3.1 Unterschied zwischen Lage und Streuung

Im Folgenden ist n der Umfang der Stichprobe, und x_i sind die festgestellten Werte für das betrachtete Merkmal. Falls die Daten in k Klassen eingeteilt worden sind, werden die Klassengrössen mit n_i angegeben und die x_i sind die für die Klassen festgelegten Werte. Um die Daten besser zu verstehen, haben sich verschiedene Kenngrössen als nützlich erwiesen. Die Lageparameter drücken aus, wie eine ganze Reihe von Messwerten als Gesamtheit liegt, während die Steuparameter beschreiben, wie kompakt oder verstreut die Daten zu einander liegen.

Beispiel:
Wenn beispielsweise die drei Schüler A, B und C in vier Mathematikprüfungen die Noten



Schüler A	2.3	5.8	4.2	5.7
Schüler B	4.7	4.2	4.5	4.6
Schüler C	5.7	5.2	5.5	5.6

bekommen haben, so lassen sich diese Werte graphisch wie in der Abbildung oben rechts gezeigt darstellen. Der Notendurchschnitt als Lageparameter ist bei den beiden mittelmässigen Schülern A und B 4.5, während er beim guten Schüler C bei 5.5 liegt. Die beiden Schüler B und C zeigen sehr konstante Leistungen, während Schüler A grosse Leistungsunterschiede zeigt, sodass seine Noten breit streuen.

3.2 Lageparameter

Das, was umgangssprachlich als Durchschnitt bezeichnet wird, kann auf verschiedene Arten mathematisch festgelegt werden. Der *Mittelwert* $\overline{x}$ ist der gebräuchlichste Wert. Er wird durch

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad \overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i \tag{1}$$

festgelegt. Manchmal eignet sich aber der *Median*, der dem mittleren Wert entspricht, wenn man die Daten der Grösse nach sortiert, und für den somit gilt, dass die Hälfte der Daten links davon und die

andere Hälfte rechts davon liegt, für gegebene Merkmale besser. In anderen Fällen eignet sich auch der *Modus* (oder Modalwert), der dem häufigsten Wert entspricht. (Es ist zu beachten, dass für qualitative Daten, denen künstlich Zahlen zugeordnet worden sind, diese Kenngrössen ausser dem Modus nicht viel aussagen, weil ihre Werte von der willkürlichen Wahl der Zahlen abhängen.)

Beispiel:
Gegeben sind die Werte 5, 4, 5, 1, 3, 5, 6, 2, 1, 4 als Messwerte eines Merkmals. Der Mittelwert ist 3.6, weil es zehn Werte sind, die zusammengezählt 36 ergeben. Der Modus ist 5, weil dieser Wert dreimal vorkommt, während die übrigen Werte nur einmal oder zweimal vorkommen. Sortiert man die Zahlen der Grösse nach, bekommt man 1, 1, 2, 3, 4, 4, 5, 5, 5, 6 und sieht, dass der mittlere Wert als Median 4 ist.

3.3 Streuparameter

Der einfachste Parameter zur Beschreibung der Streuung ist die *Variationsbreite* (auch Spannweite genannt), die die Differenz zwischen grösstem und kleinstem Wert angibt. Wichtiger sind aber die *Varianz* σ^2 und die *Standardabweichung* σ als Wurzel daraus, die durch

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad\qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2}$$

definiert sind, wobei $\overline{x}$ der Mittelwert gemäss (1) ist.

Beispiel:
Betrachtet man nochmals die drei Schüler A, B und C aus dem obigen Beispiel mit den Noten

Schüler A	2.3	5.8	4.2	5.7
Schüler B	4.7	4.2	4.5	4.6
Schüler C	5.7	5.2	5.5	5.6

und dem Notendurchschnitt 4.5 für A und B beziehungsweise 5.5 für C, so ist $\sigma \approx 1.64$ für A und $\sigma \approx 0.22$ für die beiden anderen.

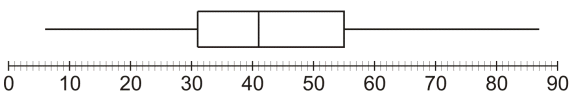
3.4 Kastengraphik

Weil es bei Daten immer auch so genannte Ausreisser gibt, die das Bild verzerren, sind Methoden erfunden worden, die diese eliminieren. Möchte man beispielsweise die Einkommen in der Schweiz untersuchen, so gibt es auf beiden Seiten der Skala problematische Werte. Auf der einen Seiten arbeiten Leute für einen Hungerlohn, um sich ihr Studium zu finanzieren, und auf der anderen Seiten gestehen sich gewisse Manager unanständig hohe Saläre zu. Weil sehr hohe Einkommen stark ins Gewicht fallen, eignet sich der Median besser, um einen sinnvollen Durchschnitt zu berechnen, als der Mittelwert.

Das *Quantil* Q_p ist der Punkt in den nach Grösse angeordneten Daten, unter dem der Anteil p aller Fälle liegt, und ist deshalb eine Verallgemeinerung des Medians. Das Quantil ist das generelle Konzept, das die Daten als Median in zwei, als Terzil in drei, als Dezil in zehn und als Perzentil in hundert Teile teilt. Besonders häufig gebraucht ist das *Quartil*, mit dem die Daten in vier Viertel geteilt werden. Mit $Q_{0.25}$ wird das erste oder untere und mit $Q_{0.75}$ das dritte oder obere Quartil bezeichnet, während der Median $Q_{0.5}$ auch zweites Quartil heisst.

Das Quartil ist die Grundlage der so genannten *Kastengraphik* (auch Box-Plot oder Box-and-Whisker-Plot genannt). Darin wird der kleinste und der grösste Wert zusammen mit dem ersten Quartil Q_1, dem Median Q_2 und dem dritten Quartil Q_3 gezeigt. Der Bereich von Q_1 bis Q_3 heisst *Interquartilbereich*. Manchmal wird statt Minimum und Maximum auch das unterste und oberste Dezil benutzt.

Beispiel:
Die nebenstehende Abbildung zeigt Daten, bei denen das Minimum 6 und das Maximum 87 ist, und bei denen $Q_1 = 31$, $Q_2 = 41$ und $Q_3 = 55$ ist.



4 Bivariate Analyse und Korrelation

4.1 Zusammenhang zwischen zwei Merkmalen

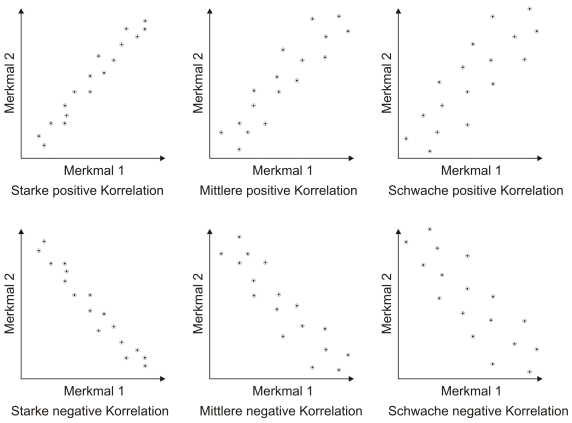
Zwischen zwei Merkmalen kann es einen Zusammenhang geben, der nicht einfach als Funktion angegeben werden kann. So sind grössere Menschen tendenziell zwar schwerer, aber aus der Körpergrösse kann man das Körpergewicht nicht exakt bestimmen. So einen Zusammenhang nennt man *Korrelation*. Viele wissenschaftliche Fragen basieren auf der Suche nach einer Korrelation zwischen zwei Merkmalen, die oft einfach Variablen genannt werden, wobei die eine als unabhängig und die andere als abhängig betrachtet wird. Je stärker die Korrelation ist, umso genauere Vorhersagen kann man von der unabhängigen Variablen auf die abhängige Variable machen.

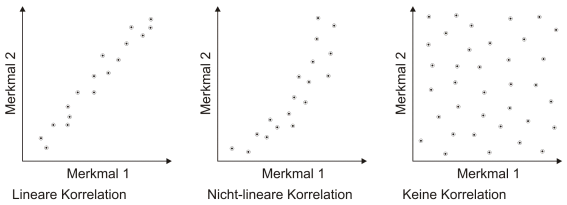
Dass zwei Merkmale korrelieren, heisst nicht, dass ein kausaler Zusammenhang zwischen ihnen besteht. Generell kann die Statistik keine kausalen Zusammenhänge aufdecken. Sie kann einen Zusammenhang aufdecken, aber was das für ein Zusammenhang ist, muss mit anderen Mitteln eruiert werden.

4.2 Streudiagramme

Zur Darstellung des Zusammenhangs zwischen zwei Merkmalen werden *Streudiagramme* (auch Scatterdiagramme oder Scatterplots genannt) benutzt, wie sie in der nebenstehenden Abbildung gezeigt werden. Nimmt das zweite Merkmal mit dem ersten zu, so spricht man von positiver Korrelation, und nimmt das zweite Merkmal mit steigendem ersten Merkmal ab, so nennt man die Korrelation negativ.

Je stärker die Korrelation ist, umso mehr entspricht sie einer Funktion, bei der man dem Merkmal 1 exakt einen Wert für das Merkmal 2 zuordnen kann. Je schwächer die Korrelation ist, desto unsicherer und ungenauer ist diese Zuordnung.

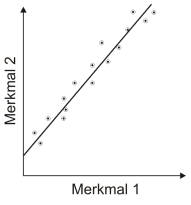




Die Abbildung links zeigt das Diagramm einer linearen Korrelation im Vergleich mit demjenigen einer nicht-linearen Korrelation und einem Streudiagramm, das überhaupt keine Korrelation aufweist. Im Folgenden kann die Korrelation positiv oder negativ sein, soll aber immer sinnvoll durch eine Gerade approximiert werden können.

4.3 Lineare Regression

Sind x und y die beiden Merkmale, für welche eine Korrelation besteht, so kann man nach einer Funktion $y = f(x)$ suchen, welche diesen Zusammenhang möglichst exakt darstellt. Diese Suche nach einer geeigneten Funktion f nennt man das *Regressionsproblem*.



Bei der *linearen Regression* wie in der nebenstehenden Abbildung soll f eine lineare Funktion mit $y - \overline{y} = a(x - \overline{x})$ beziehungsweise $y = ax + (\overline{y} - a\overline{x})$ sein. Der Punkt mit den Koordinaten $(\overline{x}, \overline{y})$ liegt also auf dieser Geraden. Um eine ungefähre Lösung zu finden, kann man entweder einen typischen Datenpunkt als zweiten Punkt auf der Geraden wählen, oder man legt die Gerade so, dass etwa gleich viele Datenpunkte auf beiden Seiten der Geraden liegen. Will man die Gerade genauer legen, bestimmt man sie mit der Methode der kleinsten Quadraten so, dass die Summe der quadrierten Abstände der Datenpunkte von der Geraden minimal wird. (Jedes Tabellenkalkulationsprogramm erlaubt es, lineare oder andere Funktionen zu finden, welche die Datenpunkte unter dem Stichwort *Trendlinie* optimal approximieren.)

4.4 Korrelationskoeffizienten

Gesucht ist eine Grösse, die angibt, wie stark die Korrelation zwischen zwei Merkmalen ist. Die Grösse

$$C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{3}$$

heisst *Kovarianz*, und es gilt $C_{xx} = \sigma^2$ wegen (2). Normiert man die Kovarianz, indem man durch das Produkt der Standardabweichung der x_i und der y_i dividiert, bekommt man

$$r_{xy} = \frac{C_{xy}}{\sqrt{C_{xx}} \cdot \sqrt{C_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{4}$$

als *Korrelationskoeffizienten*. Für diese Grössen gilt $0 \leq |r_{xy}| \leq 1$.

Die Korrelationskoeffizienten sind ein Mass für die Stärke einer Korrelation. Für $0 \leq |r_{xy}| \leq 0.25$ ist die Korrelation sehr schwach und für $0.25 < |r_{xy}| \leq 0.5$ ist sie schwach. Von mittlerer Korrelation spricht man für $0.5 < |r_{xy}| \leq 0.75$, und für $0.75 < |r_{xy}| \leq 1$ ist die Korrelation stark. Positive Korrelationen haben positive Korrelationskoeffizienten und negative Korrelationen haben negative Korrelationskoeffizienten.

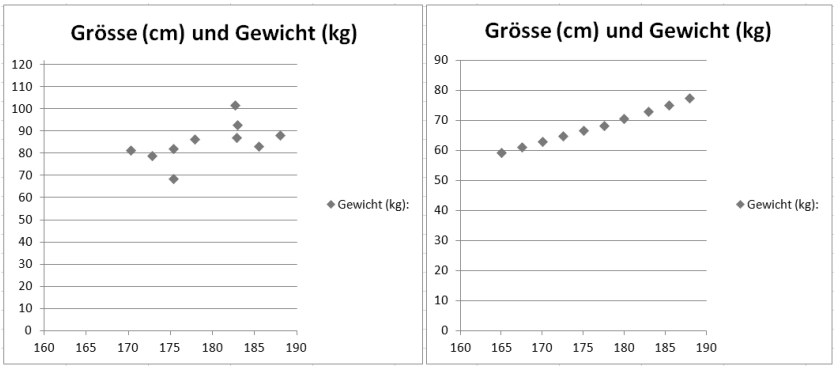
Beispiel:
Körpergrösse und Körpergewicht beim Menschen sind wie oben erwähnt nicht vollständig unabhängig, aber sind auch nicht so stark korreliert, dass man aus der Grösse das Gewicht exakt berechnen könnte. Eine Stichprobe bei zehn jungen Männern hat die Daten

Grösse (cm)	188.0	177.8	182.6	182.9	170.2	185.4	175.3	175.3	182.8	172.7
Gewicht (kg)	88.45	86.64	102.06	92.99	81.65	83.46	82.55	68.95	87.54	79.38

ergeben. Der Zusammenhang zwischen durchschnittlicher Grösse und durchschnittlichem Gewicht hingegen ist eine eindeutige Zuordnung wie in

Grösse (cm)	165.0	167.5	170.0	172.5	175.0	177.5	180.0	183.0	185.5	188.0
Gewicht (kg)	59.30	61.20	63.00	64.80	66.60	68.40	70.70	72.90	75.20	77.50

gezeigt. Die folgende Abbildung zeigt die beiden Zusammenhänge graphisch



mit der Stichprobe links und den Durchschnittswerten rechts. Mit (3) und (4) gilt für die Stichprobe links $C_{xx} \approx 34.40$, $C_{yy} \approx 75.93$ und $C_{xy} \approx 25.37$, woraus $r_{xy} \approx 0.5$ folgt, und für die Durchschnittwerte rechts $C_{xx} \approx 60.27$, $C_{yy} \approx 36.96$ und $C_{xy} \approx 42.44$, woraus $r_{xy} \approx 0.9$ folgt. Die Korrelation ist bei der Stichprobe also eher schwach, während sie bei den Durchschnittswerten erwartungsgemäss sehr stark ist. (Die Korrelationskoeffizienten mit einem Tabellenkalkulationsprogramm berechnet sind etwas höher. Der Wert für die Stichprobe ist 0.56, und für die Durchschnittswerte bekommt man genau 1.00.)