

Deskriptive Statistik

1.1) Wie entsteht eine empirische Verteilung? Wie gewinnt man die Tabelle? Wie stellt man die Verteilung grafisch dar? Was ist ein Histogramm, was ein Polygonzug?

Wie entsteht eine empirische Verteilung?

Eine **empirische Verteilung** entsteht durch die Aufbereitung von in empirischen Untersuchungen/Beobachtungen gewonnenen Daten. Grundlage dafür sind die Mess- oder Beobachtungsprotokolle der Untersuchung.

Wie gewinnt man die Tabelle?

Zur **tabellarischen Aufbereitung** des Materials wird zunächst eine Urliste angefertigt, in der die Versuchspersonen mit einem Laufindex gekennzeichnet werden und jeder dieser Laufvariablen in einer weiteren Spalte der entsprechende Messwert zugeordnet wird. Diese Urliste wird dann, evtl. nach Festlegung von Kategoriebreiten, in eine zusammenfassende Strichliste überführt. Dabei wird jeder Messwert oder jede Kategorie nur einmal aufgeführt und erhält eine Laufvariable für die Verteilung. Aufgrund der Strichliste kann nun die Häufigkeit der einzelnen Messwerte oder Kategorien bestimmt werden.

Die tabellarische Beschreibung der Verteilung kann, ausgehend von der Strichliste, durch eine Häufigkeitsverteilung, eine kumulierte Häufigkeitsverteilung (sukzessiv aufsummierte Häufigkeiten → wobei der letzte Wert dem Stichprobenumfang n entspricht – Kontrollfunktion) eine Prozentwertverteilung und/oder eine kumulierte Prozentwertverteilung erfolgen.

Wie stellt man die Verteilung grafisch dar?

Die **graphische Darstellung** einer empirischen Verteilung kann z. B. durch ein Histogramm, einen Polygonzug oder auch durch ein Kreisdiagramm erfolgen.

Was ist ein Histogramm, was ein Polygonzug?

- Für die Darstellung mittels **Polygonzug** werden die Kategoriemitten benötigt.

$$\text{Kategorienmitte} = \frac{\text{obere Kategoriengrenze} + \text{untere Kategoriengrenze}}{2}$$

Diese Werte werden in gleichen Abständen auf der Abszisse (x-Achse) abgetragen. Die Häufigkeiten werden auf der Ordinate (y-Achse) abgetragen. In den die Kategoriemitten kennzeichnenden Punkte werden Lote errichtet, deren Länge der jeweiligen Kategorienhäufigkeit entspricht. Die Endpunkte der Lote werden verbunden = Polygon. Bei dem Polygonzug handelt es sich nicht um eine Kurve, sondern um einen Streckenzug!!!

Der Polygonzug beginnt und endet auf der Abszisse, wobei die Fläche unter dem Polygonzug dem Stichprobenumfang n bzw. bei einer Prozentwertverteilungen 100% entspricht. Da die einem Polygon zugrunde liegende Variable stetig ist (stetige Variable: das Merkmal kann unendlich viele Messwerte annehmen, z.B. Alter, Gewicht, Körpergröße), dürften sich theoretisch keine Knicke im Linienverlauf ergeben (möglich bei sehr großen Stichproben). Eine Glättung des Kurvenverlaufes ist durch das Verfahren der gleitenden Durchschnitte (drei-, fünf- o. siebengliedrige Ausgleichung) möglich.

- Bei der Darstellung durch ein **Histogramm** werden auf der Abszisse die Kategoriengrenzen und auf der Ordinate, wie beim **Polygon** auch die Häufigkeiten (absolut oder prozentual) abgetragen. Im Histogramm erhält jeder X-Wert eine Säule von der Breite der Maßzahlklasse und der Höhe f . Die seitlichen Begrenzungslinien jeder Säule stellen dabei die Maßzahlklassengrenzen dar; der über der Maßzahl X in der Höhe f gezeichnete Punkt liegt also in der Mitte der Säule.

Wobei zu beachten ist, dass die Fläche und nicht die Höhe einer Säule im Histogramm die Häufigkeiten in dieser Kategorie/Messwert angibt (- bei gleich breiten Säulen ist die Höhe aber proportional!). Die Gesamtfläche des Histogramms entspricht ebenfalls dem Stichprobenumfang n oder 100%.

→ Achtung: Spalten zwischen den einzelnen „Säulen“ sind nicht möglich

Lt Bortz (S.31ff) ist ein Polygon nur zur Darstellung einer stetigen Variablen geeignet, während das Histogramm der graphischen Darstellung einer diskreten Variablen (diskrete Variable: Wertebereich

nicht beliebig unterteilbar, kann nur bestimmte Abstufungen einnehmen, z.B. Geschlecht, Parteizugehörigkeit, also qualitative Merkmale) vorbehalten bleiben sollte.
Ergänzend: Das Kreisdiagramm eignet sich besonders für die Veranschaulichung von Häufigkeiten einer Nominalskala.

1.2) Welcher Zusammenhang besteht zwischen der Häufigkeitsverteilung als Histogramm und der kumulierten Häufigkeitsverteilung als Polygonzug? Mit welcher geometrischen Konstruktion kann man aus der kumulierten Häufigkeitsverteilung Quantile entnehmen?

Welcher Zusammenhang besteht zwischen der Häufigkeitsverteilung als Histogramm und der kumulierten Häufigkeitsverteilung als Polygonzug?

Allgemein gesagt ist die **kumulierte Häufigkeitsverteilung als Polygonzug** das Integral der **Häufigkeitsverteilung als Histogramm**.

Satz: $f_{\text{kumk}}(x_0) = \int_{-\infty}^{x_0} f(x) dx.$

Mit welcher geometrischen Konstruktion kann man aus der kumulierten Häufigkeitsverteilung Quantile entnehmen?

- Bei der Häufigkeitsverteilung als Histogramm (Dichtefunktion) entsprechen stets die Flächen der Säulen den Häufigkeiten, d. h. wenn ein bestimmter Flächenanteil ermittelt werden soll, z.B. ein Quartil, müssten dafür die Flächen der Säulen bis zu dem Abszissenpunkt $\frac{n}{4}$ aufsummiert werden.
- Der Polygonzug für die kumulierte Verteilung dagegen stellt eine Verteilungsfunktion dar, bei der nicht die Fläche sondern die Funktion eines bestimmten Abszissenwertes den kumulierten Häufigkeiten bis zu diesem Wert entspricht. Dadurch, daß dieser Polygonzug der kumulierten Häufigkeiten das Integral der Häufigkeitsverteilung darstellt, ist es möglich den gesuchten Flächenanteil, der ja z. B. bei einem Quartil $\frac{n}{4}$ beträgt, geometrisch zu ermitteln. Dazu wird auf der Ordinate (bei einem Quartil) der Wert $\frac{n}{4}$ abgetragen und dann abgelesen welcher Abszissenwert zu diesem Punkt gehört. Der abgelesene „x“- Wert entspricht damit dem Punkt unter der Fläche, der $\frac{1}{4}$ der Fläche abschneidet. Andere Quantile können entsprechend ermittelt werden.

1.3) Was ist eine Maßzahlklasse? Was bedeutet Maßzahlklassenzusammenfassung, was Maßzahlenttransformation? Was ist eine lineare Maßzahlenttransformation? Wie hängen Maßzahlklassenzusammenfassung und Maßzahlenttransformation miteinander zusammen?

Was ist eine Maßzahlklasse?

Um die Verteilungseigenschaften empirischer Daten besser veranschaulichen zu können, werden die individuellen Messwerte in Kategorien bzw. Intervallen zusammengefasst.
Eine **Maßzahlklasse**, die *immer durch ihren mittleren Messwert bezeichnet* wird, enthält also je nach Klassenbreite eine bestimmte Anzahl von aufeinanderfolgenden Messwerten, deren Häufigkeiten aufsummiert die Häufigkeit in dieser Maßzahlklasse ergeben.

Was bedeutet Maßzahlklassenzusammenfassung, was Maßzahlenttransformation?

⇒ Die Maßzahlklassenzusammenfassung und ihre Transformation wird bei großen Stichproben vorgenommen um Kenngrößen, wie Mittelwert, Standardabweichung, etc., besser und einfacher berechnen zu können und um ein übersichtlicheres Histogramm zu erhalten.

- Wenn man mehrere alte Maßzahlklassen zu wenigen neuen zusammenfasst spricht man von **Maßzahlklassenzusammenfassung**. Dies bietet sich an, wenn mehr als ca. 50 Beobachtungswerte vorliegen hat. Man bildet entsprechend dem Umfang n einer Stichprobe etwa 7 bis 20 neu

Maßzahlklassen (\sqrt{n}) mit gleicher Klassenbreite b . Auch hier wird die Maßzahlklasse wieder nach ihrer Mitte bezeichnet. Angenommen, wir haben nach obigem Beispiel neue Maßzahlklassen mit der Klassenbreite $b = 3$ gebildet. Dann haben wir jetzt eine Maßzahlklasse von 13,5 – 16,5 mit der Mitte von 15. Die nächste Maßzahlklasse hätte eine Mitte von 18, 21, 24, usw..

Die Zusammenfassung von Messwerten in Maßzahlklassen wird auch als Reduktionslage einer Häufigkeitsverteilung bezeichnet.

- Zuvor hatten wir Maßzahlklassen mit Schritten von 1, nun haben wir Maßzahlklassen mit Schritten von 3, um diese wieder in Schritten von 1 zu bekommen könnte man z.B. durch $b = 3$ teilen und hätte dann die Maßzahlklassen 5, 6, 7 und 8. Dies nennt man lineare Maßzahltransformation.

Was ist eine lineare Maßzahltransformation?

- Um eine weitere Vereinfachung der Daten zu erreichen, können die so erhaltenen Maßzahlen nach

der Formel $x_t = \frac{(x - K)}{a}$ linear transformiert werden, wobei a die Klassenbreite und K eine festgelegte

Konstante darstellt. Um eine neue einfachere Skala zu finden, kann z. B. die am häufigsten besetzte Maßzahlklasse, die transformierte Klasse $x_t = 0$ erhalten, die vorangehenden und nachfolgenden Klassen werden den entsprechend mit -1 oder 1 bezeichnet.

- Noch eine weitere Möglichkeit der linearen Maßzahltransformation wäre:

$$X_t = k X_i - b$$

Wie hängen Maßzahlklassenzusammenfassung und Maßzahltransformation miteinander zusammen?

Durch die lineare **Transformation** wird der Ursprung und die Einheit der Skala neu im numerischen Relativ neu festgelegt, die für die Auswertung wichtigen quantitativen Aussagen der Intervalldaten bleibt aber erhalten.

1.4) Was bedeutet der Ausdruck $\sum_{i=1}^n x_i$? Auf welche Anordnung von Zahlen kann er

bezogen werden? Was muß dabei berechnet werden? Was ergibt $\sum_{i=1}^n a$, was

$\sum_{i=1}^n (x_i - a)$, was $\sum_{i=1}^n a * x_i$, was $\sum_{i=1}^n x_i * y_i$?

Was bedeutet der Ausdruck $\sum_{i=1}^n x_i$?

Das griechische Sigma (Σ) stellt das Operationszeichen für eine Summe dar. Dabei steht $\sum_{i=1}^n x_i$ für x_1

+ x_2 + x_3 + ... + x_n , also die Summe aller x_i -Werte für $i = 1$ bis n . „ Der Laufindex i kann durch beliebige andere Buchstaben ersetzt werden. Unterhalb des Summenzeichens wird der Laufindex mit der unteren Grenze aller Werte ($i = 1$) gleichgesetzt, und oberhalb des Summenzeichens steht die obere Grenze (n).

Auf welche Anordnung von Zahlen kann er bezogen werden?

Bezogen wird der Ausdruck auf die Werte der Variablen x , wobei der Laufindex i untere Grenze aller Werte (also den 1. Wert) bezeichnet und oberhalb des Summenzeichens die obere Grenze (der letzte Wert) steht, der mit aufsummiert wird.

Was muß dabei berechnet werden?

$$\sum_{i=1}^n x_i = X_1 + X_2 + X_3 + X_4 + \dots + X_n$$

Was ergibt $\sum_{i=1}^n a$, was $\sum_{i=1}^n (x_i - a)$, was $\sum_{i=1}^n a * x_i$, was $\sum_{i=1}^n x_i * y_i$?

- Da bei dem Ausdruck $\sum_{i=1}^n a$ der Buchstabe a mit keinen Laufindex versehen ist, kann es sich nur um eine Konstante/freie Variable handeln, d. h. es kann keine Summe gebildet werden und damit ist

$$\sum_{i=1}^n a = n * a.$$

- $\sum_{i=1}^n (x_i - a)$ steht für $(x_1 - a) + (x_2 - a) + \dots + (x_{n-1} - a) + (x_n - a)$, es werden

also zuerst die Differenzen (oder Abweichungen) der einzelnen Werte x_i von einer Konstanten a ermittelt und dann diese Differenzen aufsummiert

→ Mögliche Umformung: $\sum_{i=1}^n (x_i - a) = \left(\sum_{i=1}^n x_i \right) - n * a$, also zuerst Summe aller x_i -Werte

bilden und dann die Konstante n -mal davon subtrahieren.

- $\sum_{i=1}^n a * x_i$ ist die Summe der Produkte $(a * x_1) + (a * x_2) + (a * x_3) + \dots + (a * x_n)$.

Da a eine Konstante darstellt ist es auch möglich zuerst die Summe aller x_i -Werte zu bilden und dann mit der Konstanten a zu multiplizieren,

d h. $\sum_{i=1}^n a * x_i = a * \sum_{i=1}^n x_i$.

- $\sum_{i=1}^n x_i * y_i = (x_1 * y_1) + (x_2 * y_2) + (x_3 * y_3) + \dots + (x_n * y_n)$, also die Summe der

Variablenprodukte $x * y$ mit den jeweils gleichen Indizes.

1.5 Was versteht man unter der Zentraltendenz einer Verteilung? Welche Maße gibt es dafür? Welche Eigenschaften haben sie? Welche Überlegungen führen zur Herleitung der einzelnen Maße der Zentraltendenz? An welche Anwendungsvoraussetzungen sind sie geknüpft?

Was versteht man unter der Zentraltendenz einer Verteilung?

Die **Zentraltendenz einer Verteilung** kennzeichnet die *Lage einer Verteilung auf dem Zahlenstrahl* (x- Achse).

Masse der Zentraltendenz geben an, durch welchen Wert eine Verteilung *am besten repräsentiert* werden kann. Allgemein ist die Zentraltendenz ein Maß für Pegel, Niveau oder Durchschnitt einer Verteilung und kann z. B. als mittlere oder durchschnittliche Leistung einer Stichprobe interpretiert werden.

Welche Maße gibt es dafür?

→ Maße der Zentraltendenz:

- 1.) **Modus/Modalwert**
- 2.) **Median** (= „typischer Fall“)
- 3.) **Arithmetisches Mittel/Mittelwert** (= Durchschnitt, z.B. bei Verbrauch,...)
- 4.) **Geometrisches Mittel** (*siehe Frage 1.6*)
- 5.) **Harmonisches Mittel** (*siehe Frage 1.6*)
- 6.) **Gewogenes arithmetisches Mittel** (*siehe Frage 1.6*)

Welche Eigenschaften haben sie?

zu 1.) Modus

Der Modalwert ist der *Mittelpunkt* der am **häufigsten besetzten Maßzahlklasse** = derjenige Wert, der am häufigsten vorkommt bzw. in der graph. Verteilung der Wert, bei dem die Verteilung ihr Maximum hat (Bortz).

→ Es gibt **bimodale** Verteilungen, bei denen 2 mit gleicher Häufigkeit besetzten Kategorien dazwischen mindestens eine weniger besetzte ist (= Zweigipfelig).

(Gibt es zwischen den Intervallen mit den höchsten Frequenzen keine weitere Kategorie, so ist die Verteilung **breitgipflig**).

⇒ Der Modus besitzt den geringsten Informationswert, da er **keine Voraussetzungen** an das Skalenniveau macht und auch die Verteilung der Maßzahlen nicht berücksichtigt.

zu 2.) Median

Der Median gibt den Punkt wieder, der die **Fläche**/die Maßzahlen der Verteilung **halbirt**.

→ der Wert, von dem alle übrigen Werte in der Weise abweichen, dass die **Summe der Absolutbeträge der Abweichungen** ein **Minimum** ergibt.

Anwendung: Bei **schiefen Verteilungen**, **offenen Maßzahlklassen** am Anfang oder Ende der Verteilung und **keinem Intervallskalenniveau**, sowie **extrem geringer Anzahl von Messwerten** ist er das günstigste Maß der Zentraltendenz.

⇒ Median wird berechnet durch *lineare Interpolation* (das ist die Formel, mit u und f_{kum} und f_{kumu} ...).

$$Md = u + \frac{\frac{n}{2} - f_{kumu}}{f_{krit}} \times Kb$$

zu 3.) Arithmetisches Mittel

- bezeichnet den *Durchschnitt einer Verteilung*.

- derjenige Wert, von dem die **Summe aller quadrierten Abweichungen aller X_i - Werte von diesem Wert ein Minimum** ergibt.

- Summe aller Maßzahlen dividiert durch Anzahl aller Maßzahlen:

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

Anwendung: *durchschnittlicher* Verbrauch, Kaufkraft,...

(zu 7): **Geometrisches Mittel** (siehe Frage 1.6)

(zu 8): **Harmonisches Mittel** (siehe Frage 1.6)

(zu 9): **Gewogenes arithmetisches Mittel** (siehe Frage 1.6)

An welche Anwendungsvoraussetzungen sind sie geknüpft?

Modus = kein Skalenniveau

Median = mindestens *Ordinalskalenniveau*

Arithmetisches Mittel = mindestens *Intervallskalenniveau*

1.6 Was versteht man unter dem geometrischen Mittel, was unter dem harmonischen Mittel und was unter dem gewogenen arithmetischen Mittel? Wann werden diese Maße der Zentraltendenz angewandt?

Was versteht man unter dem geometrischen Mittel, was unter dem harmonischen Mittel und was unter dem gewogenen arithmetischen Mittel?

zu 4.) **Geometrisches Mittel**

Das **Geometrisches Mittel** ist die n-te Wurzel aus dem Produkt aller Einzelmesswerte

$$GM = \sqrt[n]{x_1 \times x_2 \times x_3 \times \dots \times x_n}$$

Voraussetzung: alle Werte müssen positiv sein

° zu 5.) **Harmonisches Mittel**

Das **Harmonisches Mittel** ist die Anzahl der Einzelmesswerte dividiert durch die Summe der Kehrwerte der Einzelmesswerte.

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

⇒ Damit ist das HM der Reziprokwert zum arithm. Mittel.

zu 6.) **Gewogenes arithmetisches Mittel**

⇒ Dies ist der *Gesamtmittelwert aus mehreren Einzelmittelwerten*.

Aus den Mittelwerten eines Merkmals in mehreren Kollektiven wird die Gesamtsumme aller Messwerte berechnet und durch die Summe aller Kollektivgrößen geteilt.

$$GAM = \frac{\sum_{j=1}^e n_j \times \bar{x}_j}{\sum_{j=1}^k n_j}$$

Wann werden diese Maße der Zentraltendenz angewandt?

Geometrisches Mittel

Anwendung: Es wird bei verhältnisskalierten Daten oder Merkmalen angewandt Bei subjektiven Empfindungsstärken, Wachstums-/Inflationsraten, Schwundquoten und (Lern-) Zuwachs wird im Normalfall dieses Maß berechnet.

Harmonisches Mittel

Anwendung: zur Berechnung von Durchschnittswerten bei konstanten und expliziten Maßeinheiten, z. B. Durchschnittsverbrauch von Treibstoff (in Liter/ km), Durchschnittsgeschwindigkeit (in km/h)...

***Merke:** bei konstanter Nennervariable (Fahrzeit, Litermenge ...) ergibt sich der *durchschnittliche Index über das arithm. Mittel der Einzelindizes*.

Gewogenes arithmetisches Mittel

Anwendung: Dieses Maß macht eine **Aussage über die Gruppe** (und nicht z.B. über den einzelnen Schüler), er wird auch verwendet, wenn verschiedenen Mittelwerten eine **unterschiedliche Gewichtung** zukommen soll (z.B. unterschiedliche Gewichtung von Haupt- und Nebenfächern beim Durchschnitt).

1.7 Was versteht man unter dem gleitenden arithmetischen Mittel? Wozu wird es angewandt?

Was versteht man unter dem gleitenden arithmetischen Mittel?

- damit ist das **Verfahren der gleitenden Durchschnitte** oder auch **Ausgleichung** genannt, gemeint
- glättet den Kurvenverlauf eines Polygonzuges
die einem Polygonzug zugrundeliegende Variable ist stetig, damit dürften sich theoretisch keine Knicke im Linienvverlauf ergeben
 1. Möglichkeit der Annäherung an einen „geglätteten“ Verlauf:
 - großer Stichprobenumfang bei sehr engen Kategorien
 2. Möglichkeit der Annäherung an einen „geglätteten“ Verlauf :
 - Verfahren der gleitenden Durchschnitte

Annahme des Verfahren der gleitenden Durchschnitte:

- Häufigkeiten in benachbarten Kategorien auf einer stetigen Variable verändern sich kontinuierlich und nicht sprunghaft
- unter dieser Annahme kann die Häufigkeit einer Kategorie durch die Häufigkeiten der benachbarten Kategorien im Interpolationsverfahren bestimmt werden
- zufällig bedingte Irregularitäten und Sprünge können damit ausgeglichen werden
statt der Häufigkeiten einer Kategorie k wird der Durchschnitt der Häufigkeiten der Kategorien k-1, K und k + 1 eingesetzt

a) 3- gliedrige Ausgleichung

$$f(\text{quer})_k = \frac{f_{(k-1)} + f_k + f_{(k+1)}}{3}$$

b) 5- gliedrige Ausgleichung

$$f(\text{quer})_k = \frac{f_{k-2} + f_{k-1} + f_k + f_{k+1} + f_{k+2}}{5}$$

Wozu wird es angewandt?

Den Polygonzug einer graphischen Darstellung zu glätten, da die zugrundeliegende Variable stetig ist, dürfte sich theoretisch keine Knicke im Linienvverlauf ergeben.

1.8 Was versteht man unter der Dispersion einer Verteilung? Welche Maße gibt es dafür? Welche Eigenschaften haben sie? Welche Überlegungen führen zur Herleitung der einzelnen Maße der Dispersion? An welche Anwendungsvoraussetzungen sind sie geknüpft?

Was versteht man unter der Dispersion einer Verteilung?

Dispersionsmaße kennzeichnen die Breite oder Ausdehnung einer Verteilung. (= Maße der Streuung)

Die Dispersion ist ein Maß für die Abweichungen der einzelnen Maßzahlen voneinander. Sie ist damit ein Maß für die Homogenität (Gleichheit, innere Übereinstimmung) einer Stichprobe, bzw. wie ähnlich die Mitglieder hinsichtlich des gemessenen Merkmals sind

→ Sie informieren also über die Unterschiedlichkeit der Werte und charakterisieren die Variabilität eines Merkmals.

→ Bei empirischen Untersuchungen, helfen sie bei der Beantwortung der Frage, wie die bezüglich eines Merkmals angetroffene Unterschiedlichkeit von Personen oder anderen Untersuchungseinheiten zu erklären ist.

Dispersionsmaße können zudem die Zuverlässigkeit einer Messung/Prognose oder die Größe eines Messfehlers bestimmen helfen.

Die Maße der Dispersion werden formal analog zu den Maßen der Zentraltendenz definiert. Sie entsprechen diesen hinsichtlich Informationsgehalt und Anwendungsvoraussetzungen:

Jedoch: ähneln sich 2 Verteilungen hinsichtlich ihrer zentralen Tendenz, können sie dennoch auf Grund unterschiedlicher Streuungen (Dispersion) der einzelnen Werte stark voneinander abweichen

Welche Maße gibt es dafür?

- 1) Variationsbreite („absoluter Streubereich“) (R, „range“)
- 2) Quartilabstand (Q) (Interquartilbereich)
- 3) AD- Streuung
- 4) Varianz (s^2 , σ^2)
- 5) Standardabweichung (s , σ)

Welche Eigenschaften haben sie?

Zu 1) *Variationsbreite* (R, „range“): „entspricht“ dem Modus und ist die Differenz zwischen größtem und kleinstem Wert einer Verteilung. Ist das einfachste Streuungsmaß, das nur die beiden extremsten Werte berücksichtigt → Streubreite aller Werte

$$R = x_{\max} - x_{\min}$$

Er gibt also an, in welchem Bereich sich die Messwerte befinden.

Da der absolute Streubereich beim Auftreten von großen Extremwertem kein adäquates Abbild der vorliegenden Daten liefert, ist er nicht besonders zur Bestimmung der Dispersion einer Verteilung geeignet. (Er verliert besonders an Bedeutung, wenn Ausreißer dabei sind.)

Zu 2) Quartilabstand (Q) (Interquartilbereich)

a) *Interquartilbereichsbestimmung*

Sie geben also denjenigen Ausschnitt der Messskala wieder, in dem sich ein bestimmter Prozentsatz aller Werte befindet. E sind Maßzahlen mit der gleichen Einheit wie die ursprünglichen Werte.

$$Q_1 = u + \frac{\frac{1 \times n}{4} - f_{kumu}}{f_{krit}} \times Kb$$

$$Q_2 = u + \frac{\frac{2 \times n}{4} - f_{kumu}}{f_{krit}} \times Kb$$

$$Q_3 = u + \frac{\frac{3 \times n}{4} - f_{kumu}}{f_{krit}} \times Kb$$

b) *Interquartilbereich*

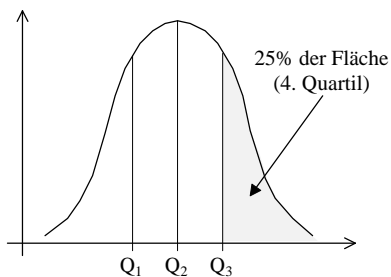
Breite des Intervalls, in der die Mitte, die Hälfte aller Fälle liegt

$$IQ = Q_3 - Q_1$$

c) *mittlerer Quartilabstand (Q)*: Er ist quasi das Pendant zum Median und ist definiert als:

$$Q = \frac{Q_3 - Q_1}{2} \quad Q_3 - Q_1 \text{ wird als Interquartilbereich bezeichnet.}$$

Q_1 und Q_3 sind die Punkte auf der Skala, die die unteren bzw. oberen 25% (1. bzw. 4. Quartil) der Maßzahlen der Verteilung abschneiden. Q_2 entspricht dem Median und halbiert die Verteilung.



d) *Interdezilbereich I_D* (Breite der Verteilung innerhalb der sich die mittleren 80% befinden)

$$D_1 = u + \frac{\frac{1 \times n}{10} - f_{kumu}}{f_{krit}} \times Kb$$

$$D_9 = u + \frac{\frac{9 \times n}{10} - f_{kumu}}{f_{krit}} \times Kb$$

$$D = D_9 - D_1$$

→ I_Q / I_D = Maß für die Schmalgipfligkeit

zu 3) *AD- Streuung (= average-deviation) /Mittlere Abweichung/Mittlere Variation*

Gibt den Durchschnitt der in Absolutbeträgen gemessenen Abweichungen aller Messwerte von \bar{x} (AM) an. (Auf die großen und kleinen Maßzahlen wird ganz verzichtet, es gehen nur die mittleren 50% der Fälle in das Maß ein.)

$$AD = \frac{\sum_{i=1}^n (|x_i - M|)}{n}$$

nicht so empfindlich gegen Ausreißer

zu 4) *Varianz (s^2 , σ^2)*: Die Varianz ist definiert als die Summe der quadrierten Abweichungen der Maßzahlen von ihrem arithmetischen Mittel. (D.h. jede Abweichung wird also mit sich selbst gewichtet)

$$s^2 = \frac{\sum_{i=1}^n (x_i - M)^2}{n}$$

Sie ist also die durchschnittliche quadrierte Abweichung der Einzelmaßzahlen von ihrem arithmetischen Mittel.

- Varianz ist neben der Standardabweichung das gebräuchlichste Maß zur Kennzeichnung der Dispersion einer Verteilung
 - Die Varianz wird von jedem Messwert der Gruppe beeinflusst
 - sämtliche Werte werden einzeln berücksichtigt
 - größere Abweichungen werden durch die Quadrierung stärker berücksichtigt, als kleinere Abweichungen
- ist empfindlich gegen Ausreißer

Nachteil: Wir erhalten hier nicht Maßzahlen mit der gleichen Einheit wie die ursprünglichen Werte, sondern das Quadrat der ursprünglichen Einheit der Messwerte liegen hier zu Grunde. Da dieses Maß schwer interpretierbar ist, wird die Quadrierung gewissermaßen wieder rückgängig gemacht, in dem man die Wurzel aus der Varianz berechnet (= Standardabweichung!!!)

Zu 5) *Standardabweichung* oder *Streuung* (s , σ): Sie ist die positive Quadratwurzel aus der Varianz. (= Maß für die Stärke der Variabilität der Rohwerte). Sie gibt in etwa an, wie weit die einzelnen Werte im Durchschnitt vom Mittelwert abweichen.

$$s = \sqrt{s^2}$$

Standardabweichung ist größer als AD- Streuung: $s > AD$

- durch die Quadrierung werden größere Abweichungen stärker berücksichtigt

- AD- Streuung gewichtet alle Abweichungen gleich

Differenz zwischen AD- Streuung und Standardabweichung nimmt deshalb mit steigender Dispersion zu

Welche Überlegungen führen zur Herleitung der einzelnen Maße der Dispersion?

Zu 1) absoluter Streubereich oder Variationsbreite (R , „range“)

→ „entspricht“ dem Modus und ist die Differenz zwischen größtem und kleinstem Wert einer Verteilung:

$$R = x_{\max} - x_{\min}$$

Zu 2) Quartilabstand (Q) (Interquartilbereich)

a) *Interquartilbereichsbestimmung*

→ Hier liegt die Flächenberechnung zu Grunde

$$Fläche = f_{kumu} + \frac{f_{krit} \times (x - u)}{Kb}$$

b) *Interquartilbereich (IQ)*

→ Fläche 3 – Fläche 1

c) *mittlerer Quartilabstand (Q)*

→ Um den mittleren Quartilabstand zu berechnen,

$$Q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{(Q_3 - Q_1)}{2}$$

3) AD- Streuung

→ nix mehr gefunden

4) Varianz (s^2 , σ^2)

→ Dieses Streumaß ist eine Art Mittelwert 2. Ordnung, d.h. Mittelwert der Mittelwertsabweichung

5) Standardabweichung (s , σ)

→ Da das Maß der Varianz schwer interpretierbar ist, wird die Quadrierung gewissermaßen wieder rückgängig gemacht, in dem man die Wurzel aus der Varianz berechnet (= Standardabweichung!!!)

An welche Anwendungsvoraussetzungen sind sie geknüpft?

• Streumaße für mindestens ordinalskalierte Messwerte

1) absoluter Streubereich oder Variationsbreite (R, „range“)

2) Quartilabstand (Q) (Interquartilbereich)

a) *Interquartilbereichsbestimmung*

b) *Interquartilbereich (IQ)*

c) *mittlerer Quartilabstand (Q)*

→ ist stets dann zu berechnen, wenn als Maß der Zentraltendenz der Median bestimmt worden ist, also vor allem bei schiefen Verteilungen und bei offenen Maßzahlklassen

• Streumaße für mindestens intervallskalierte Messwerte

3) AD- Streuung (?)

4) Varianz (s^2 , σ^2)

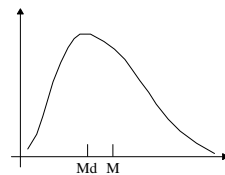
5) Standardabweichung (s , σ)

1.9 Was versteht man unter Schiefe, was unter Exzess einer Verteilung? Welche Maße gibt es dafür? Aufgrund welcher Überlegungen werden sie hergeleitet?

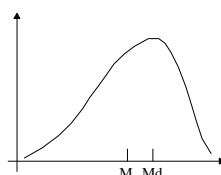
Was versteht man unter Schiefe, was unter Exzess einer Verteilung?

• Die **Schiefe** ist ein Maß für die Abweichung einer Verteilung von der Symmetrie. Unter schiefen Verteilungen versteht man Verteilungen, deren graphische Darstellung nicht achsensymmetrisch ist. Bei schiefen Verteilungen fallen arithmetisches Mittel und Median nicht zusammen. Liegt das arithmetische Mittel rechts vom Median, so spricht man von einer linkssteilen Verteilung, anderenfalls von einer rechtssteilen Verteilung

rechtsschief (linksgipflig)



linksschief (rechtsgipflig)



• Der **Exzess** bezeichnet die Schmal- bzw. Breitgipfligkeit einer Verteilung

Welche Maße gibt es dafür?

A) Das Maß für die **Schiefe**

A1) ist das sogenannte **3. Potenzmoment** (α_3) einer Verteilung:

$$\alpha_3 = \frac{\sum_{i=1}^n z_i^3}{n}$$

In diese Formel gehen z-standardisierte Maßzahlen ein!

$$\text{Formel zur Berechnung von z-Werten: } z_i = \frac{x_i - M}{s}$$

Rechtssteil = negativer α -Wert

Symmetrisch = 0

Linkssteil = positiver α - Wert

A2) oder bei **Quantilen** berechnet sich die Schiefe: $Sch = \frac{AM - Md}{S}$

Ist $Sch > 0$ - linkssteil, $Mo < Md < AM$

Ist $Sch < 0$ - rechtssteil, $AM < Md < Mo$

$Sch = 0$ - symmetr., $AM = Md = Mo$

B) Das Maß für den **Exzess** ist

B1) das **4. Potenzmoment** (α_4) einer Verteilung:

$$\alpha_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n} \times \frac{1}{s^4} - 3$$

In diese Formel gehen z-standardisierte Maßzahlen

ein!

α_4 bei einer Normalverteilung = 0, kleiner Werte bedeuten Breitgipfligkeit, größere Werte bedeuten Schmalgipfligkeit. Der Exzess sollte nur bei einer unimodalen Verteilung berechnet werden.

B2) über die **Quantile** betrachtet.

$$Ex = \frac{Q3 - Q1}{2(D9 - D1)}$$

Der Exzess einer Normalverteilung = 0,263, je größer umso breitgipfliger und umgekehrt.

Aufgrund welcher Überlegungen werden sie hergeleitet?

Überlegungen zur Herleitung:

Schiefe: Ausgangspunkt sind die z-Werte. Bei der 3. Potenz werden z-Werte mit größerem Betrag stärker gewichtet als kleinere, d.h. vom \bar{x} weiter entfernte x_i erhalten stärkeres Gewicht. Außerdem bleiben die Vorzeichen bei der 3. Potenz erhalten, d.h. die größeren (negativen) z-Werte liegen bei rechtssteilen Verteilung links $\Rightarrow \alpha_3$ wird negativ, bei linkssteilen Verteilungen liegen größere (positive) z-Werte rechts $\Rightarrow \alpha_3$ wird positiv.

Eine grobe Abschätzung für die Größe der Schiefe (Sch) einer Verteilung nannte bereits PEARSON

(1895): $Sch = \frac{\bar{x} - Mo}{s}$ (BORTZ Gl.1.28).

$$\text{GLASER: } Sch = \frac{\bar{x} - Md}{s} \quad (10).$$

Exzess: Ausgangspunkt sind die z-Werte. Bei einer Normalverteilung erwartet man den Wert von $\alpha_4 = 0$.

Kleinere α_4 -Werte kennzeichnen eine breitgipflige und größere α_4 -Werte eine schmalgipflige Verteilung.

Der Exzess einer Verteilung sollte nur bei unimodalen Verteilungen berechnet werden.

Der Exzess (Ex) kann auch über Perzentilwerte nach folgender Gleichung geschätzt werden:

$$\text{BORTZ (1.29): } Ex = \frac{P_{75} - P_{25}}{2 * (P_{90} - P_{10})}$$

Der Exzess einer Normalverteilung beträgt $Ex = 0,263$. Je größer der Exzess einer Verteilung, um so breitgipfliger ist ihr Verlauf.

$$\text{GLASER: } Ex = \frac{Q_3 - Q_1}{2 * (D_9 - D_1)}$$

1.10 Was versteht man unter dem Moment einer Maßzahl bezüglich des Koordinatenanfangs, bezüglich des arithmetischen Mittelwertes und eines beliebigen Punktes a? Was versteht man unter einem linearen, quadratischen, kubischen, quartischen Moment? Was versteht man unter dem Moment einer Verteilung?

Was versteht man unter dem Moment einer Maßzahl bezüglich des Koordinatenanfangs, bezüglich des arithmetischen Mittelwertes und eines beliebigen Punktes a?

Moment einer Maßzahl = x_i bezüglich a \rightarrow das b^{te} Moment $\sum (x_i - \bar{x})^b$

Das Moment ist die Abweichung der Messwerte vom Bezugswert:

bzgl. des Koordinatenanfangs: Messwert $x_i - 0$

bzgl. des AM: Abweichung vom AM ($x_i - \bar{x}$)

bzgl. eines Punktes a: ($x_i - a$)

\Rightarrow Moment Maßzahl – beliebiger Punkt A: Punkt a entspricht dem Lot des Waagebalkens

Der Punkt A ist das Lot der Waagebalken der Verteilung (Vorstellung daß Verteilung auf Waagebalken drückt.) Waage ist genau dann im Gleichgewicht wenn $a = AM =$ Schwerpunkt einer Verteilung.

\rightarrow Wenn alle Fälle auf x-Achse wie Hebelarme am Schwerpunkt drehen + sich Gleichgewicht einstellt. \Rightarrow AM hat Schwerpunkteigenschaft, wenn $a = AM$

Was versteht man unter einem linearen, quadratischen, kubischen, quartischen Moment?

Da die Abweichungen potenziert werden können, spricht man vom ersten, zweiten bis n-ten Moment bzw. vom linearen,Moment:

lineares Moment: $\sum (x_i - a)$ \rightarrow Lage

quadratisches Moment: $\sum (x_i - a)^2$ \rightarrow Dispersion/Varianz: $a = \bar{x}$

kubisches Moment: $\sum (x_i - a)^3$ \rightarrow Schiefe: $a = \bar{x}$

quartisches Moment: $\sum (x_i - a)^4$ \rightarrow Exzess: $a = \bar{x}$

Was versteht man unter dem Moment einer Verteilung?

Das lineare Moment einer Verteilung bezüglich der Summe a ist genau dann 0 wenn der Wert dem arithmetischen Mittelwert entspricht

$$\sum (x_i - a) = 0 \quad \text{Gleichgewicht } a = \bar{x}$$

1.11. Was versteht man unter einer standardisierten Verteilung? Zu welchem Zweck werden Verteilungen standardisiert?

Was versteht man unter einer standardisierten Verteilung?

Eine standardisierte Verteilung ist eine Verteilung deren Werte z-transformiert wurden und die somit den Mittelwert 0 und die Standardabweichung 1 hat.

Zu welchem Zweck werden Verteilungen standardisiert?

Eine Verteilung wird standardisiert um verschiedene Gruppen oder Populationen vergleichbar zu machen. Dies wird durch eine Relativierung der individuellen Leistungen an denen der Gruppe erreicht.

a) Die einfachste Art ist dabei den Prozentrang zu bilden: es wird für jede Person ermittelt wie viel Prozent aller Mitglieder der Population einen größeren bzw. kleineren Wert erhalten. Der Prozentrang wird dann anhand kumulierter Prozentwertverteilungen bestimmt.

b) Eine andere Möglichkeit ist der Vergleich der Abweichungen der individuellen Leistung von den Durchschnittsleistungen der Gruppe.

Um die Abweichungen zweier Leistungen vom Mittelwert besser vergleichbar machen zu können müssen sie zuvor an der Unterschiedlichkeit aller Werte in der jeweiligen Gruppe relativiert werden. Dabei werden die Abweichungen durch die Standardabweichung der jeweiligen Gruppe dividiert und man erhält somit einen z- Wert :

$$z_i = \frac{x_i - \bar{x}}{s}$$

1.12. Was versteht man unter einer bivariaten Verteilung ? Wie sieht die zugehörige Urliste aus? Wie gewinnt man die bivariate Verteilung als Tabelle aus der Urliste ? Wie wird sie graphisch dargestellt? Was versteht man dabei unter einer univariaten Randverteilung?

Was versteht man unter einer bivariaten Verteilung?

Eine bivariate Verteilung ist eine Häufigkeitsverteilung bei der n voneinander unabhängige Beobachtungen zwei Merkmalsalternativen (= 2 Variablen) zugeordnet werden. In der Urliste entsprechen die beiden Zeilen den verschiedenen Merkmalen, die Spalten den Messwerten der Vpn.

Kriterien:

- 1) bivariat normalverteilte Grundgesamtheit
 - Normalverteilung der x-Werte
 - Normalverteilung der y-Werte

Wie sieht die zugehörige Urliste aus?

X 3 3 2 1 1 6 4 3 6 5 4 1

Y 5 4 3 5 2 2 1 6 1 2 1 5

→ Es werden die Merkmalsausprägungen x_i des Merkmals X und die Merkmalsausprägungen y_i des Merkmals Y an denselben statistischen Einheiten erhoben

Wie gewinnt man die bivariate Verteilung als Tabelle aus der Urliste?

Um eine bivariate Tabelle aus der Urliste zu erhalten werden die beiden Variablen als Koordinatensystem aufgespannt, wobei auf der Abszisse (I) die Werte von y, auf der Ordinate (—) die Werte von x abgetragen werden. (Die Beobachtungen $(x_i; y_i)$ ordnet man so, daß die Paare zuerst nach den Ausprägungen des Merkmals X geordnet werden. Bei gleicher Ausprägung x_j ordnet man nach den Ausprägungen des Merkmals Y (lexikographische Ordnung))
Dabei wird jeweils ein kontinuierlicher Zahlenstrahl vom kleinsten bis zum größten Wert verwendet. In diesem Koordinatensystem können nun die jeweiligen x den dazugehörigen y- Werten zugeordnet werden (man erhält eine Strichliste)

$f_v \cdot y^2$	$f_v \cdot y$	y	$f_v \cdot y^2$	$f_v \cdot y$	f_v	y							
							6			/			
							5	//		/			
							4			/			
							3		/				
							2				/	/	
							1	/			//	/	
								1	2	3	4	5	6

Wie wird sie graphisch dargestellt?

Graphisch wird eine bivariate Verteilung durch eine Punktwolke dargestellt, wobei mehrfach besetzte Zellen durch eine häufigere Umrandung (z.B. Sunflowers) gekennzeichnet werden. Dabei sind allerdings auch andere Kennzeichnungen möglich.

Was versteht man dabei unter einer univariaten Randverteilung?

Eine univariate Randverteilung entspricht den einzelnen Verteilungen der beiden Merkmale unabhängig voneinander. Die entsprechenden Kennwerte können jeweils am Rand der bivariaten Verteilung (s.o.) errechnet werden und auch graphisch dargestellt werden. Dabei wird die Verteilungsfunktion eines Merkmals direkt an die jeweilige Tabelle angelegt.

1.13) Was versteht man bei einer bivariaten Verteilung unter linearer Regression? Welche beiden Parameter hat die lineare Regressionsgleichung? Wie gewinnt man sie aus der bivariaten Urliste oder Verteilung? Welche Gleichungen lassen sich nach dem Kriterium der kleinsten Abweichungsquadrate herleiten? Warum liefert das Verfahren im allgemeinen für jede bivariate Verteilung zwei verschiedene Geraden? In welchem Punkt schneiden sich diese Geraden? Wann stehen sie senkrecht aufeinander, wann fallen sie zusammen?

Was versteht man bei einer bivariaten Verteilung unter linearer Regression?

Ein linearer Zusammenhang zwischen X und Y liegt dann vor, wenn die gemessenen Punkte im Streudiagramm von X und Y auf einer Gerade liegen

Bivariate Verteilung untersucht den Zusammenhang zwischen zwei Variablen.

⇒ Regression ist das Voraussagen oder Schätzen des Wertes einer Variablen aus der Kenntnis des Wertes einer anderen Variablen. Die Regressionsgleichung ermöglicht die Vorhersage des Wertes der abhängigen Variablen (Prädiktorvariable) aufgrund eines beliebigen Wertes der unabhängigen Variablen (Kriteriumsvariable). Dazu muß Höhe und Richtung (positiv oder negativ) ihrer Korrelation bekannt sein. Geschätzt wird mit Hilfe sogenannter Regressionsgeraden.

Allgemein gilt:

- ⇒ Eine lineare Regressionsanalyse ist nur dann sinnvoll, wenn der Prädiktor X und das Kriterium Y korreliert sind,
- ⇒ Bei der einfachen linearen Regression soll die Kriteriumsvariable nur durch eine Prädiktorvariable X vorhergesagt werden
- ⇒ Die Konstanten a und b heißen *Regressionskonstanten*. a wird manchmal auch als *Regressionskoeffizient* bezeichnet.

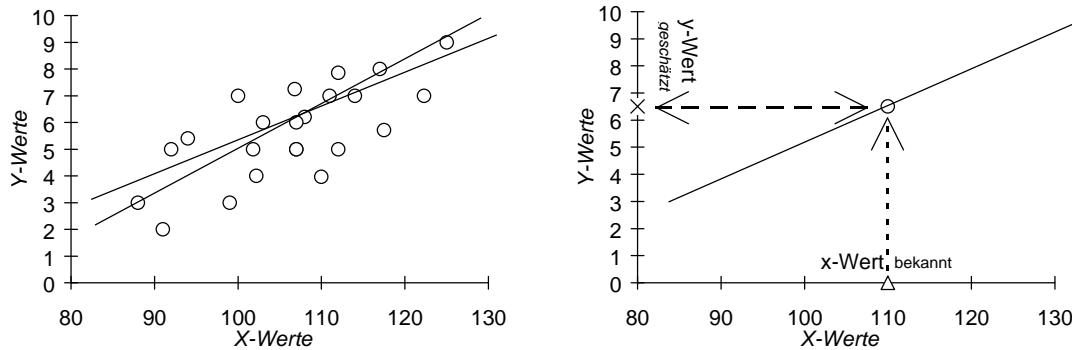


Abb. 3: Als Punkte dargestellte Datenpaare zweier korrelierender Variablen. Die Regressionsgleichungen lassen sich als Geraden darstellen. Auf der rechten Abbildung ist der Prozess des Schätzens grafisch dargestellt (Y' wird durch X geschätzt).

Welche beiden Parameter hat die lineare Regressionsgleichung?

Die Gerade wird durch 2 Bestimmungen (Steigung = b und Höhe = a) eindeutig festgelegt.

$$y^{\wedge} = b_{yx} \cdot x + a_{yx} \quad \text{und} \quad x^{\wedge} = b_{xy} \cdot y + a_{xy} \quad (y^{\wedge} \text{ und } x^{\wedge} \text{ bedeutet } y \text{ bzw. } x \text{ geschätzt})$$

Bei einem linearen Zusammenhang ($y^{\wedge} = b_{yx} \cdot x + a_{yx}$) gibt der Wert a die Steigung der Geraden an, auf der die Punkte liegen, und der Wert b gibt den Wert von Y an der Stelle X = 0 an.

→ Der vordere Buchstabe entspricht dem Kriterium (= Kriteriumsvariable; die noch nicht vorhanden ist, und die vorhergesagt werden soll).

→ Der zweite Buchstabe von b und a entspricht dem Prädiktor (= Prädiktorvariable; die vorhanden ist, und die zur Vorhersage eingesetzt wird)

⇒ Grundsätzlich reichen zwei Bestimmungsstücke einer Geraden, um die Koordinaten aller Punkte aus rechnen zu können.

Wie gewinnt man sie aus der bivariaten Urliste oder Verteilung?

Die Beziehung zwischen 2 Variablen kann aufgrund von Beobachtungen ermittelt werden, z. B. bei 2 VP die x- und die y-Leistungen registrieren. Man erfasst einen Wert für Merkmal y und einen Wert für Merkmal x, trägt nun die einander entsprechenden Koordinaten in ein Koordinatensystem ein. Der entstehenden Punktwolke entsprechen die Werte einzelner Personen. Im Idealfall liegen die Punkte auf einer Geraden, der Regressionsgerade.

→ Wir erhalten entsprechend mehr Messpunkte, wenn mehr VP gemessen werden, die dann aber auch nicht mehr alle auf einer Geraden liegen (Grund: z.B. Motivationsschwankungen oder Ermüdungseffekte) => Punkteschwarm. → Gesucht werden die Regressionskoeffizienten a und b, die zu einer Geraden führen, die Punkteschwarm am besten repräsentiert. Das Ergebnis einer empirischen Untersuchung ist also ein Punkteschwarm, die Regressionsgleichung ist die, die diesen am Besten wiedergibt. Diesen Regressionsgeraden entsprechen zwei Gleichungen, die für die Schätzungen eingesetzt werden können:

$$y^{\wedge} = b_{yx} \cdot x + a_{yx}$$

$$x^{\wedge} = b_{xy} \cdot y + a_{xy}$$

→ Wobei man zuerst **b** errechnen muss:

$$b_{yx} = \frac{\sum f \times x_i \times y_i - \frac{1}{n} \times (\sum f \times x_i) \times (\sum f \times y_i)}{\sum f \times x_i^2 - \frac{1}{n} \times (\sum f \times x_i)^2}$$

$$b_{xy} = \frac{\sum f \times x_i \times y_i - \frac{1}{n} \times (\sum f \times x_i) \times (\sum f \times y_i)}{\sum f \times y_i^2 - \frac{1}{n} \times (\sum f \times y_i)^2}$$

→ Und dann mit den Mittelwerten **a** bestimmen

$$a_{yx} = \bar{y} - b_{yx} \times \bar{x}$$

$$a_{xy} = \bar{x} - b_{xy} \times \bar{y}$$

Welche Gleichungen lassen sich nach dem Kriterium der kleinsten Abweichungsquadrate herleiten?

Allgemeines:

Gesucht wird also, diejenige Gerade, für die die Summe der quadrierten Abweichung der vorhergesagten \hat{y} - Werte von den beobachteten y-Werten minimal wird (= Kriterium der kleinsten Quadrate)

$\sum (y_i - \hat{y}_i)^2 = \min$ (← Das Kriterium bezieht sich auf die Abweichung der Punkte von der Geraden in y-Richtung. Dadurch ist gewährleistet, dass die Regressionsgleichung ihre Aufgabe, y-Werte möglichst präzise vorherzusagen, optimal erfüllt.)

⇒ Die Regressionsgerade ist diejenige Gerade, die die Summe der quadrierten Vorhersagefehler minimiert.

→ In einer Kurvendiskussion muß nun bestimmt werden, für welche Werte a und b diese Funktion ein Minimum annimmt bzw. die Summe der quadrierten Abweichungen minimal ist. Dies ist der Fall, wenn die erste Ableitung der Funktion gleich 0 ist und die Zweite Ableitung positiv ist.

Herleitbar sind:

- a

- b

⇒ Die Regressionsgerade für $y^{\wedge} = b_{yx} x_i + a_{yx}$ und die Gerade $x^{\wedge} = b_{xy} y_i + a_{xy}$. Nach dem Kriterium der kleinsten Quadrate können die Regressionskoeffizienten, die wie erwähnt wichtige Bestimmungsstücke der Geraden sind, ermittelt werden durch die Differentialrechnung.

Warum liefert das Verfahren im allgemeinen für jede bivariate Verteilung zwei verschiedene Geraden?

Es wird die Beziehung zwischen 2 Variablen aufgrund von Beobachtungen ermittelt

⇒ Mit den Regressionsgleichungen werden beide Vorhersage Richtungen beachtet.

Beispiel:

In einer Therapie soll geprüft werden, wie sich bei einem Klienten die Beeinträchtigung durch psychische Problem im Verlaufe der ersten 10 Sitzungen verändert. X sei die Anzahl der Sitzungen und Y das Ausmaß der Beeinträchtigungen.

a) Vorhersage der Beeinträchtigung Y auf Grund der Sitzungsanzahl X

$$y^{\wedge} = b_{yx} \cdot x + a_{yx}$$

b) Vorhersage der Sitzungsanzahl X auf Grund der Beeinträchtigung Y

$$x^{\wedge} = b_{xy} \cdot y + a_{xy}$$

In welchem Punkt schneiden sich diese Geraden?

Die Koordinaten des Schnittpunktes entsprechen den Mittelwerten \bar{x} und \bar{y} .

Wann stehen sie senkrecht aufeinander, wann fallen sie zusammen?

Bei normalverteilten Merkmalen folgt die Umhüllung eines Punkteschwarms einer Ellipse, die mit wachsender Kovarianz enger wird. Nähert sich die Verteilung der Punkte einem Kreis, so besteht keine Kovarianz. Kann der Punkteschwarm durch eine Gerade mit positiver (Gerade von links unten nach rechts oben) oder negativer Steigung (Gerade von links oben nach rechts unten) repräsentiert werden, spricht man von positiver oder negativer Kovarianz

1) Eine Kovarianz von 0 wirkt sich folgendermaßen auf Regressionsgeraden aus:

$$y^{\wedge} = \bar{y} \quad \text{und} \quad x^{\wedge} = \bar{x}$$

Diese Geraden stehen **senkrecht** aufeinander, sie stehen parallel zu x-Achse, bzw. die andere Gerade parallel zur y-Achse.

2) Bei maximaler Kovarianz ($s_x \cdot s_y$) **fallen** die beiden Regressionsgeraden **zusammen**, je kleiner der Winkel, desto größer die Kovarianz. Die beiden Regressionsgeraden sind nur im Falle eines perfekt linearen Zusammenhangs identisch, d. h. wenn Bivariate nur aus Punkten besteht, die ohnehin schon auf einer Geraden liegen.

Erklärung Kovarianz:

In einer bivariaten Verteilung ist das mittlere Produkt der Abweichungspaare vom jeweiligen Mittelwert die Kovarianz. Sie ist ein Maß für die gemeinsame Variation zweier Variablen, hängt aber außerdem von den Streuungen der beiden Variablen ab, deshalb als statistische Kennziffer ungeeignet.

1.14) Was versteht man unter der Regression zur Mitte?

⇒ Auch Regression zum Mittelwert, Regressionsresiduen oder Regressionseffekt genannt.

Damit ist folgendes Phänomen gemeint:

→ wird bei einer Gruppe von Individuen die gleiche Variable 2 mal hintereinander gemessen, ist die Abweichung der individuellen Messwerte vom Gruppenmittel bei der Zweitmessung geringer als bei der Erstmessung. (Trautner Band 1)

oder:

Unter der Regression zur Mitte versteht man, dass geschätzte bzw. vorhergesagte Standardwerte einer abhängigen Variablen näher dem Stichprobenmittelwert liegen als die der unabhängigen Variablen.

Bsp. aus Bortz S. 192: Rechtschreibfähigkeit eines Schülers hängt nicht nur von dessen allgemeiner Intelligenz ab, sondern auch von weiteren Merkmalen wie Sprachverständnis, Merkfähigkeit.... Eine genaue Untersuchung der Residuen kann deshalb aufschlussreich dafür sein, durch welche Merkmale die geprüfte Variable noch determiniert ist.

[Der Ausdruck geht auf Francis Galton zurück, der die Beziehung der Körpergrößen von Vätern und Söhnen untersuchte. Er fand dass die Söhne von großen Vätern (entspräche der a.V.) im Durchschnitt weniger von der durchschnittlichen Größe aller männlichen Personen abweichen als die Väter selbst (entspräche der u.V.).]

⇒ Anders, weil verständlicher ausgedrückt: die Tendenz zur Mitte tritt immer dann auf, wenn man Personen mit extrem hoher und extrem niedriger Merkmalsausprägung ein zweites Mal untersucht. Die Werte nahe am Mittelwert einer Verteilung sind nämlich wahrscheinlicher als Extremwerte.

1.15) Was versteht man in einer bivariaten Verteilung unter Produktmomentkorrelation? Welche Zahlenwerte kann sie annehmen? Was sagen diese über den Zusammenhang der beiden Variablen aus? Was bedeutet das Vorzeichen? Was versteht man unter Determinationskoeffizient, was unter Standardschätzfehler?

Was versteht man in einer bivariaten Verteilung unter Produktmomentkorrelation?

Die **Produktmomentkorrelation**, oder der Korrelationskoeffizient $r = \frac{\text{cov}(x, y)}{s(x) \times s(y)}$

(r = Kovarianz durch Produkt beider Standardabweichungen (= geometrisches Mittel beider Varianzen)) gibt den **Zusammenhang** zwischen den zwei Merkmalen x und y an.

r ist also ein Maß zur Kennzeichnung von Zusammenhängen, wobei man davon ausgeht, dass zwischen x und y ein „wahrer“, irgendwie gearteter Zusammenhang unabhängig von deren Quantifizierung existiert (anders als bei der Kovarianz).

⇒ *Der Korrelationskoeffizient beschreibt die Enge und die Richtung des linearen Zusammenhangs zweier Merkmale durch eine Zahl r .*

Voraussetzungen zur Berechnung von r :

- beide Häufigkeitsverteilungen müssen symmetrisch und unimodal sein, d.h. annähernd normalverteilt.
- der Zusammenhang muss linear sein (d.h. die Punktwolke muss durch eine Ellipse begrenzt sein).

Welche Zahlenwerte kann sie annehmen?

⇒ r kann **Zahlenwerte** zwischen -1 und $+1$ annehmen.

Was sagen diese über den Zusammenhang der beiden Variablen aus?

Bei $r = +1$ spricht man von einem perfekt positivem Zusammenhang,

bei $r = -1$ spricht man von einem perfekt negativem Zusammenhang.

Bei $r = 0$ besteht kein linearer Zusammenhang zwischen den zwei Merkmalen.

Hoher Zusammenhang: =/größer 0,7

Mittlerer Zusammenhang: 0,7 - 0,3

Niedriger Zusammenhang: 0 - 0,3

Was bedeutet das Vorzeichen?

Das **Vorzeichen**, negativ oder positiv, gibt die Steigung der Regressionsgeraden an.

- Ein **positives Vorzeichen** bedeutet, dass bei größer werdendem x auch der y-Wert zunimmt,
 - bei einem **negativen Vorzeichen** wird mit größer werdendem x der y-Wert kleiner und andersherum.
- ⇒ Allgemein kann gesagt werden: je höher zwei Merkmale miteinander korrelieren, desto besser kann von der Ausprägung des einen Merkmals auf die des anderen geschlossen werden.

Dies ist u.a. von Vorteil, wenn ein Merkmal schwer zugänglich oder schwer zu messen ist und seine Ausprägung leichter mit Hilfe der anderen Variablen erschlossen werden kann.

Was versteht man unter Determinationskoeffizient, was unter Standardschätzfehler?

• Der **Determinationskoeffizient** r^2 ist definiert als das Quadrat des Korrelationskoeffizienten. Mit dem Zahlenwert $r^2 * 100$ wird der gemeinsame Varianzanteil zweier Variablen ausgedrückt. Man sagt auch, der Determinationskoeffizient gibt an, in welchem prozentualen Ausmaß die Varianz der einen Variablen durch die Varianz der anderen Variablen erklärt wird.

Dieses Maß spiegelt wider wie viel die zwei Variablen „gemeinsam“ haben.

Diese Interpretation des Determinationskoeffizienten geht dabei von der (nicht immer korrekten) Annahme aus, daß die Ausprägung der Werte zweier Variablen von einer Reihe anderer Einflussfaktoren verursacht werden. Die Korrelation zwischen zwei Variablen ist theoretisch immer dann hoch, wenn sie über möglichst viele gemeinsame Ursachen verfügen. Man kann sich dies folgendermaßen vorstellen: Eine Kriteriumsvariable X_1 wird von den (fiktiven) Einflussfaktoren A, B und C mitbestimmt. Eine Prädiktorvariable Y_1 hat als Ursachen die Faktoren A, B und D. Insgesamt sind also vier Einflussfaktoren bei der Ausprägung der Variablenwerte beteiligt. Davon beeinflussen zwei (nämlich A und B) *beide* Variablen und führen zur Ausprägung einer nach außen hin sichtbaren Korrelation. Der Statistiker spricht hier von „gemeinsamer“ Varianz. Betrachten wir nun ein anderes Beispiel. X_2 soll von A und B beeinflusst werden und Y_2 von E und F. Diese beiden Variablen enthalten keine gemeinsamen Einflussfaktoren, daher auch keine gemeinsamen Varianzanteile. Ihre Korrelation und demzufolge ihr Determinationskoeffizient ist gleich Null. Beachtet bitte, daß dieses Modell die tatsächlichen statistischen Sachverhalte nur sehr grob wiedergibt. Ich halte es trotzdem für vertretbar, weil es (hoffentlich) ein bisschen mehr Anschaulichkeit schafft.

Achtung: Korrelationskoeffizienten dürfen *nie* als Prozentwerte interpretiert werden! Und auch nie als Kausalbeziehungen

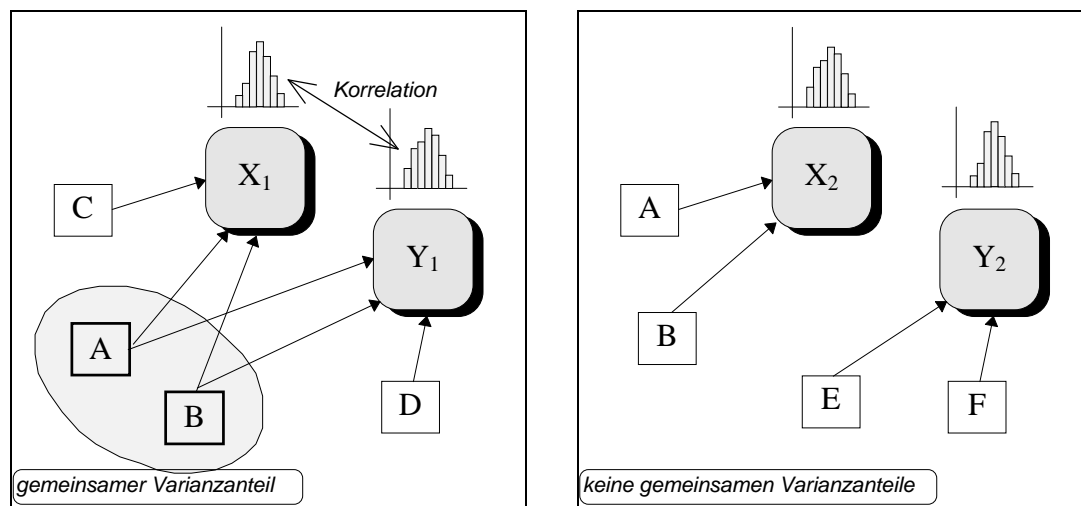


Abb. 2: Graphische Darstellung der „gemeinsamen Varianz“. Das dargestellte Modell ist nur eine sehr grobe Vereinfachung der tatsächlichen statistischen Sachverhalte.

• Der Standardschätzfehler

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$$

kennzeichnet die Streuung der y-Werte um die Regressionsgerade und ist damit ein Gütemaßstab für die Genauigkeit der Regressionsvorhersagen, die man durch die Regressionsgleichung erhält. Die Genauigkeit einer Regressionsvorhersage wächst mit kleiner werdendem Standardschätzfehler. Je größer die Korrelation zwischen x und y, umso geringer die Streuung.

1.16) Welche Korrelation kann man bei zwei Variablen auf Ordinalniveau berechnen? In welchem Zusammenhang steht sie mit der Produktmomentkorrelation?

Welche Korrelation kann man bei zwei Variablen auf Ordinalniveau berechnen?

→ Auf Ordinalniveau kann man eine **Rangfolge** der gemessenen Objekte erstellen.

Hier wird eine Stichprobe untersucht und anschließend werden die Subjekte geordnet nach Rängen.

⇒ Ordinalskalen kommen bei der Verwendung der Paarvergleichsmethode oder der Methode der Rangreihenbildung vor. Die Ordinalskala macht keine Aussage über die Größe der Abstände zwischen den Rängen.

⇒ bei zwei Merkmalen auf Ordinalniveau z.B. sozialer Rang in einer Gruppe und IQ kann man die **Rangkorrelation nach Spearman** berechnen:

$$r_s = 1 - \frac{6 \times \sum d_i^2}{n \times (n^2 - 1)}$$

Man muss zuerst eine Rangreihe bilden, wobei $d = \text{Rangplatz 1} - \text{Rangplatz 2}$ ist. Davon den Betrag nehmen $|d|$.

Bsp: Vier Schüler werden von 2 Lehrern in Bezug auf ihre Intelligenz in Rangreihen gebracht.

Frage: Stimmen die Lehrer in ihrem Urteil über die Schüler überein?

i Lehrer 1 Lehrer 2 $|d|$ d^2

1	2	3	1	1
2	1	2	1	1
3	3	1	2	4
4	4	4	0	0

⇒ $\sum d^2 = 6$, in Formel eingesetzt: $r = 0,4$, die Lehrer haben eine mittlere gleichsinnige Übereinstimmung.

In welchem Zusammenhang steht sie mit der Produktmomentkorrelation?

Die Produktmomentkorrelation r kann man aus Intervall- (z.B. Ergebnisse von b Tests) und Verhältnisdaten (z.B. Messungen wie Temperatur, Geschwindigkeit und Körpergröße) berechnen.

$$r = \frac{\sum xy - 1/n (\sum x) (\sum y)}{\sqrt{(\sum x^2 - 1/n (\sum x)^2) (\sum y^2 - 1/n (\sum y)^2)}}$$

⇒ Der Zusammenhang zwischen r_s und r besteht darin, dass r_s aus der Produktmomentkorrelation hergeleitet wurde.

⇒ Sie ist mit der Produkt-Moment-Korrelation identisch, wenn beide Merkmale jeweils die Werte 1 bis n annehmen, was bei Rangreihen der Fall ist. Eine Rangkorrelation könnte somit berechnet werden, indem in die Produkt-Moment-Korrelationsformel statt der intervallskalierten Messwerte die Rangdaten eingesetzt werden.

1.17 Welche Korrelation kann man an einer Vierfeldertafel berechnen? In welchem Zusammenhang steht sie mit der Produktmomentkorrelation?

Welche Korrelation kann man an einer Vierfeldertafel berechnen?

a) für echt dichotome Merkmale

Die Vierfelderkorrelation (im Bortz auch "Phi-Koeffizient ϕ ")

→ Mit Hilfe des Φ -Koeffizienten kann man die Vierfelderkorrelation r_ϕ berechnen:

$$r_{\Phi} = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+b)(c+d)}}$$

In welchem Zusammenhang steht sie mit der Produktmomentkorrelation?

Diese Korrelation ist direkt aus der Produktmomentkorrelation

$$r = \frac{\sum_{i=1}^n f \times x_i \times y_i - \frac{1}{n} \left(\sum_{i=1}^n f \times x_i \right) \times \left(\sum_{i=1}^n f \times y_i \right)}{\sqrt{\left[\sum_{i=1}^n f \times (x_i)^2 - \frac{1}{n} \times \left(\sum_{i=1}^n f \times (x_i) \right)^2 \right] \times \left[\sum_{i=1}^n f \times (y_i)^2 - \frac{1}{n} \times \left(\sum_{i=1}^n f \times (y_i) \right)^2 \right]}}$$

ableitbar, da beide Variablen dichotom sind und somit nur die Ausprägung 0 und 1 besitzen, d.h. die Produktmomentkorrelation über die beiden Messwertreihen von x und y entspricht exakt dem Phi-Koeffizienten.

b) für unecht dichotome

Als Spezialfall für künstlich bzw. normalverteilte dichotomierte Variablen berechnet man die tetrachorische Korrelation (z.B. bei einem Test hat man ab der Hälfte der Punktzahl bestanden, alles darunter bedeutet durchgefallen):

$$r_{tet} = \cos \frac{180^\circ}{1 + \sqrt{\frac{b \cdot c}{a \cdot d}}}$$

1.18 Was versteht man bei einer Menge von drei oder mehr Variablen unter der Partialkorrelation zwischen zwei Variablen? Was versteht man unter dem Auspartialisieren einer oder mehrerer Variablen?

Was versteht man bei einer Menge von drei oder mehr Variablen unter der Partialkorrelation zwischen zwei Variablen?

Eine Partialkorrelation ist ein Verfahren, mit dem sich überprüfen lässt, ob die Beziehung zwischen 2 Merkmalen auf einer "Scheinkorrelation" beruht, also einer Korrelation, die nur durch die Wirksamkeit einer dritten oder weiterer Variablen zustande gekommen ist.

⇒ Der Grundgedanke dieses Verfahrens ist folgender: Wenn die Korrelation zwischen 2 Variablen x und y von einer dritten Variable z beeinflusst wird, kann dies nur in der Weise geschehen, daß die Variable z mit Variable x und zusätzlich mit Variable y korreliert. Suchen wir eine Korrelation zwischen x und y, die von der Variablen z nicht beeinflusst wird, müssen wir die Variablen x und y vom Einfluß der dritten Variablen z befreien.

Was versteht man unter dem Auspartialisieren einer oder mehrerer Variablen?

⇒ Die Partialkorrelation gibt an, wie stark die Korrelation zwischen x und y wäre, wenn sie von dem vermuteten erzeugenden Effekt der Störvariablen bereinigt oder befreit wird, wenn es diese Einflüsse also nicht gäbe.

⇒ Also werden die Störvariable/n aus der Korrelation von x und y herausgerechnet, das heißt dann herauspartialisiert.

Dies geschieht mit Hilfe der Regressionsrechnung.

➔ Wir bestimmen zunächst eine Regressionsgleichung, mit der geschätzte x-Werte (= x mit Dach) auf Grund der Variablen z vorhergesagt werden können. Die Varianz dieser vorhergesagten Werte wird ausschließlich durch die Variable z bestimmt. Subtrahieren wir die vorhergesagten x-Werte von den tatsächlichen x-Werten, resultieren Residualwerte bzw. Regressionsresiduen, deren Varianz von der

Variablen z unbeeinflusst ist. Diesen Vorgang der regressionsanalytischen Bereinigung bezeichnet man als "**Herauspartialisieren**" einer Variablen z aus einer Variablen x.

→ Genauso verfahren wir mit der Variablen y, aus der ebenfalls regressionsanalytisch der Einfluß der Variablen z herauspartialisiert wird.

→ Korrelieren wir die bezüglich der Variablen z "bereinigten" Variablen x und y, ergibt sich eine Partialkorrelation zwischen den Variablen x und y, die von der 3. Variablen unbeeinflusst ist.

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \times r_{yz}}{\sqrt{1 - r_{xz}^2} \times \sqrt{1 - r_{yz}^2}}$$

→ Eine Partialkorrelation stellt eine bivariate Korrelation zwischen Regressionsresiduen dar.

Partialkorrelationen können auch berechnet werden, wenn aus dem Zusammenhang zweier Variablen nicht nur eine, sondern mehrere Variablen herauspartialisiert werden sollen = Partialkorrelationen höherer Ordnung