Methoden der Psychologie

Deskriptive Statistik

0. Einleitung 2
0.1. Definitionen 2
0.2. Grundbegriffe 2
1. Daten in der Psychologie 3
1.1. Merkmal und Merkmalsausprägung 3
1.2. Messen, Maßeinheiten und Messinstrumente 3
1.3. Tests in der Psychologie und Testgütekriterien 4
1.4. Datenniveaus und Skalenniveaus und ihr Informationswert 4
1.5 Grundgesamtheit und Stichprobe 5
1.6 Datenerfassung und Datenträger 6
2. Datenaufbereitung 6
2.1 Primäre und sekundäre Verteilungstafel 6
2.2 Häufigkeitsarten: absolut, relativ & kumulativ 6
2.3 Ausgleichung monovariater Häufigkeitsverteilungen 6
2.4 Monovariate, bivariate und multivariate Häufigkeitsverteilung 7
2.5 Normalverteilung (NV) nach Gauß 7
2.6. Weitere theoretische Verteilungen 8
3. Darstellung von Daten 8
3.1. Tabellarische Darstellung 8
3.2. Graphische Darstellung 9
3.3. Statistische Maßzahlen univariater Verteilungen 10
3.3.1. Maße für die zentrale Tendenz 10
3.3.2. Quantile 13
3.3.3. Streuungsmaße bzw. Dispersionsmaße (SQ-Werte) 13
3.4. Zusammenhangsmaße 15
3.4.1. Korrelation und Kontingenz (SP-Werte) 16
3.4.2. Spezielle Korrelationstechniken 19
3.4.3. Lineare Regression 21

0. Einleitung

0.1. Definitionen

Deskriptive Statistik

„Die deskriptive (beschreibende) Statistik hat zum Ziel, empirische Daten durch Tabellen und Grafiken übersichtlich darzustellen und zu ordnen, sowie durch geeignete grundlegende Kenngrößen zahlenmäßig zu beschreiben." (Quelle: http://de.wikipedia.org/wiki/Deskriptive_Statistik)

0.2. Grundbegriffe

Datenmatrix

Ganz allgemein werden allen gesammelten Informationen Zahlen zugeordnet.

Schreibweise: $\frac{X}{[Zeilen; Spalten]}$

Fehlende Werte (missing value)

Fehlende Werte (=**missing values**) werden häufig mit nicht leer gelassen, sondern mit einer bestimmten Zahl versehen (z.B.: weiblich = 1, männlich = 2, fehlend = 9)

a priori

Als **a-priori-Wahrscheinlichkeit** wird eine Vorwahrscheinlichkeit bezeichnet. Also eine Wahrscheinlichkeit, die absehbar ist (z.B. über geometrische Eigenschaften → Würfel).

Kriterium und Prädiktor

Die **unabhängige Variable** (Prädiktor-Variable) ist jene, welche variiert wird. Die **abhängige Variable** (Kriterium) ändert sich in einem bestimmten Verhältnis zur unabhängigen Variable. Die unabhängige Variable wird beobachtet.
In jeder Untersuchung gibt es **Störvariablen**, also Einflüsse, welche den Effekt der unabhängigen auf die abhängige Variable zusätzlich beeinflussen. Ziel ist es die Störvariablen soweit wie möglich zu minimieren.

Urliste

Die **Urliste** (siehe auch **primärer Datenträger** ⇨ 1.6.) entspricht der rohen Datenaufnahme. Jedem Merkmal ist einzeln die Merkmalsausprägung zugeordnet.

Verteilungstafel

In der **primären Verteilungstafel** sind für alle einzelnen Merkmalsausprägungen die Häufigkeiten aufgelistet. In der **sekundären Verteilungstafel** werden die Merkmalsausprägungen zusammengefasst zu Kategorien (von-bis-Bereiche). Siehe z.B. arithmetisches Mittel ⇨ 3.3.1.a.

Für die die sekundäre Verteilungstafel wird also eine **Spannweite** für die Kategoriegröße definiert.

Dispersion und zentrale Tendenz

Als Dispersionsmaße werden alle Berechnungen zusammengefasst, welche sich mit der Streuung befassen, während die zentrale Tendenz alle Berechnungen zusammenfasst, welche sich mit der Berechnung von Mittelwerten befassen.

1. Daten in der Psychologie

1.1. Merkmal und Merkmalsausprägung

Ein **Merkmal** ist eine Eigenschaft, die einer Person oder einer Sache zugwiesen ist und durch die sich die Peron oder die Sache von anderen abgrenzt und unterscheidet.

Merkmalsausprägungen sind konkrete Werte (Zahlen bzw. codierte Zahlen) von Merkmalen. Eine Merkmalsausprägung lässt sich grob in **qualitativ** und **quantitativ** unterscheiden. Zu den **qualitativen Merkmalsausprägungen** gehören Merkmale, welche **nominal** skaliert sind, zu den **quantitativen Merkmalsausprägungen** gehören **metrisch** und **ordinal** skalierte Merkmale (siehe auch 1.4.).

Als **Merkmalsart** wird die Unterteilung in **stetig** (grenzenloser „fließender" Verlauf) und **diskret** (Einteilung in Kategorien ohne „fließenden" Übergang).

Außerdem kann zwischen **manifesten Merkmalen** (direkt gemessen) und **latenten Merkmalen** (indirekt gemessen) unterschieden werden.

Stetig

Der Abstand der x-Werte bei **stetigen Wahrscheinlichkeitsverteilungen** (auch **kontinuierliche Wahrscheinlichkeitsverteilung**) ist **unendlich klein**. Außerdem kann einem einzelnen x-Wert **keine bestimmte Wahrscheinlichkeit** zugeordnet werden. Stetige Werte sind in der Praxis Messwerte (0,5m).

Eine Werteverteilung wird als **Quasistetig** bezeichnet, wenn die Werte zwar per Definition diskret sind, aber dem stetigen Prinzip sehr nahe sind (z.B. Größen 1,90m, 191m, 192m, etc.). Die Werte werden dann eher als stetig betrachtet.

Diskret

Bei **diskreten Wahrscheinlichkeitsverteilungen** befinden sich alle Werte auf der x-Achse **getrennt voneinander** und **jedem einzelnen x-Wert** ist eine bestimmte Wahrscheinlichkeit zugeordnet. Diskrete Werte werden in der Praxis nicht gemessen, da sie eine Art Kategorisierung sind (z.B. gut, mittel, schlecht).

1.2. Messen, Maßeinheiten und Messinstrumente

Als **Messen** wir das Zuordnen von **Zahlen** (**numerisches Relativ**) zu **Objekten** (**empirisches Relativ**) bezeichnet, während die Zahlen in einer bestimmten Relation die Objekte wiedergeben. Das numerische Relativ bildet die **Skala**. Das numerische Relativ repräsentiert das empirische Relativ in Form von Zahlen.

Als **Kategorisierung** wird die Zusammenfassung bestimmter **qualitativer Merkmalsausprägungen** in Gruppen, Klassen oder Kategorien bezeichnet.

Mit Hilfe **Messinstrument** erlangt man das numerische Relativ (**Zahlenwerte**). Wichtig sind hier **Objektivität** (Ergebnis muss bei Wiederholung gleich bleiben), **Reliabilität** (Messgenauigkeit) und **Validität** (Messung der wirklich beabsichtigen Merkmale).

Maßeinheiten müssen den zu messenden Größen **art- und dimensionsgleich** sein. Sie müssen entweder durch eine **Messvorschrift** festgelegt werden oder sie müssen sich an **anerkannten Naturgrößen** orientieren.

1.3. Tests in der Psychologie und Testgütekriterien

Ein **Test** hat das Ziel mehrere abgrenzbare **Persönlichkeitsmerkmale** zu messen und letztlich eine quantitative Aussage über den **relativen Grad** der **individuellen Ausprägung** zu machen.

Die **Testgütekriterien** entsprechen der **Objektivität**, **Reliabilität** und **Validität** (siehe 1.2.).

1.4. Datenniveaus und Skalenniveaus und ihr Informationswert

Sobald bei der Abbildung des numerischen Relativs eine **Skala** angegeben wird, muss zunächst klar sein welcher **Skalentyp** gemeint ist bzw. welche Information abgebildet wird. Die Informationstypen werden als **Daten- oder Skalenniveau** bezeichnet.

Das **Skalenniveau** oder Messniveau oder Skalendignität ist in der Statistik und Empirie eine wichtige Eigenschaft von Merkmalen bzw. von Variablen. Je nach der Art eines Merkmals bzw. je nachdem, welche Vorschriften bei seiner Messung eingehalten werden können, lassen sich verschiedene Stufen der Skalierbarkeit unterscheiden. (Quelle: http://de.wikipedia.org/wiki/Skalenniveau)

In einer Skala werden bestimmten Eigenschaften Zahlen zugeordnet. Die **Skalenniveaus** sind nach den möglichen mathematischen Operationen unterschieden, die man an den Zahlen der Ausprägungen anwenden kann. Je höher die Skalenniveaus, desto weniger **Transformationen** (mathematische Umwandlungen) sind erlaubt, desto besser ist jedoch die Aussagekraft.

Nominalskala

Die **Nominalskala** macht eine Aussage über die **Gleichheit / Verschiedenheit** von Merkmalsausprägungen und erlaubt alle **eindeutigen** Transformationen.

- Die Nominalskala ist das niedrigste Skalenniveau. Sie enthält zwei Bedingungen:

Exklusivität: Unterschiedliche Merkmalsausprägungen werden unterschiedl. Zahlen zugeordnet.
Exhaustivität: Genau eine Zahl für jede Merkmalsausprägung.

Besonderheit: **Dichotome** nominalskalierte Merkmalsausprägungen
Als **dichotom** wird ein Merkmal bezeichnet, wenn es genau **zwei Ausprägungsmöglichkeiten** gibt (Bsp.: ♂/♀). **Künstlich Dichotom** wiederum bezeichnet eine Klassifizierung / Kategorisierung (**sekundäre Verteilungstafel** → siehe 2.1.) einer ursprünglich nicht nominalskalierten Merkmalsausprägung (z.B. Unterteilung der Größe in < 1,80 und > 1,80). Als **natürliche Dichotomie** wird eine Merkmalsausprägung bezeichnet, welche bereits nur in zwei Ausprägungsmöglichkeiten gemessen wurde.

Operationen: =/≠

Ordinalskala

Die **Ordinalskala** (auch **Rangskala**) macht zusätzlich zur Gleichheit / Verschiedenheit Aussagen über **Größer-Kleiner-Relationen** von Merkmalsausprägungen und erlaubt alle **monotonen** Transformationen.

- Die Ordinalskala enthält eine weitere Bedingung:

Die Zahlen repräsentieren Unterschiede einer bestimmten Größe in Bezug auf die

Merkmalsausprägung. Ausprägungen mit größerer Bedeutung bekommen entsprechend größere Werte zugewiesen.

Operationen: =/≠ ; </>

Intervallskala

Die **Intervallskala** (= **metrische** Skalierung) macht Aussagen über die **Größe der Unterschiede** zwischen den Merkmalsausprägungen und erlaubt nur **lineare** Transformationen.

- Die Intervallskala enthält eine weitere Bedingung:

Gleich große Abstände zwischen zugeordneten Zahlen repräsentieren gleich große Einheiten des Konstrukts.

Operationen: =/≠ ; </> ; +/−

Die **Äquidistanz** muss gleich bleiben. ($y = a * x + b$)

Verhältnisskala

Die **Verhältnisskala** macht Aussagen über das **Verhältnis** von Merkmalsausprägungen und erlaubt alle **Ähnlichkeitstransformationen**.

- Die Verhältnisskale enthält eine weitere Bedingung:

Der Anfangspunkt der Skala kennzeichnet einen definierten Nullpunkt.

Operationen: =/≠ ; </> ; +/− ; ×/÷

Ähnlichkeitstransformationen: ($y = a * x$)

1.5 Grundgesamtheit und Stichprobe

Die **Population** ist die **Grundgesamtheit** von Personen, während die **Stichprobe** bloß ein **Ausschnitt** ist. Von einer Stichprobe lasen sich Rückschlüsse auf die Population schließen. Für Stichprobe, Population und Schätzung bekommen die Kennwerte jeweils andere mathematische Symbole.

Mittelwert: $\bar{x}$ (Stichprobe), $\boldsymbol{\mu}$ (Population), $\hat{\boldsymbol{\mu}}$ (Schätzwert)
Streuung: $\boldsymbol{s}$ (Stichprobe), $\boldsymbol{\sigma}$ (Population), $\hat{\boldsymbol{\sigma}}$ (Schätzwert)

Zur **Stichprobenauswahl** gibt es verschiedene Techniken:

- **Zufallsauswahl**
  Zufällige Auswahl der Probanden
- **Geschichtete Auswahl**
  Jede Schicht der Population wird durch eine entsprechende Schicht in der Stichprobe im richtigen Verhältnis repräsentiert
- **Mehrstufige Auswahl**
  Mehrfaches Wiederholen der Stichprobe. Jedes Mal wird jedoch eine neue Stichprobe (mit neuen Probanden) aus der gleichen Grundgesamtheit genommen.
- **Clusterauswahl**
  Am ehesten zutreffende Gruppen werden aus der Population für die Stichprobe verwendet.

1.6 Datenerfassung und Datenträger

Die **Datenerfassung** dient der **Gewinnung und Fixierung** von Merkmalen. Die Daten werden in **alpha-numerischen Zeichen** gespeichert (Zahlen, Buchstaben, Sonderzeichen).

Bei der **Kodierung** eines Merkmals werden den **Ausprägungen** der entsprechenden **Zahlen**-Variablen **sinngemäße Begriffe** zugewiesen (z.B.: weiblich = 1, männlich = 2). Es entsteht also ein Zuordnung von Zahlen zu Merkmalsausprägungen.

Als **primärer Datenträger** wird der Datenträger bezeichnet, auf dem sich die **Rohdaten** befinden (z.B. Fragebogen). Auf dem **sekundären Datenträger** befinden sich bereits **abgeschriebene oder verwertete Daten** (z.B. Excle oder SPSS).

2. Datenaufbereitung

2.1 Primäre und sekundäre Verteilungstafel

Verteilungstafeln dienen der **Verdichtung** von Daten. Bestimmte Wertebereiche werden also in Verteilungstafeln zusammengefasst.

In einer **primären Verteilungstafel** werden die Werte noch nicht zusammengefasst, sie werden jedoch der Größe nach **sortiert**.

In der **sekundären Verteilungstafel** werden die Werte nun **verdichtet**. Bestimmte Wertebereiche werden hier zusammengefasst bzw. kategorisiert. Benachbarte Merkmalsausprägungen werden dadurch gebündelt. Dadurch können stetige Merkmale einen diskreten Charakter bekommen.

2.2 Häufigkeitsarten: absolut, relativ & kumulativ

Die einzelnen Werte sind nicht mehr nachvollziehbar (bzw. der Ursprung nicht mehr bestimmbar). Das Gesamtergebnis wird übersichtlich dargestellt.

Absolute Häufigkeit: $\boldsymbol{f_k}$

Relative Häufigkeit: $\boldsymbol{f_{rel_k} = \frac{f_k}{n}}$

Prozentuale Häufigkeit: $\%_k = \boldsymbol{f_{rel_k} * 100\%}$

Gültige Prozente: Alle gültigen Werte sind 100%!

Kumulierte Prozente: Aufaddierung der Prozentwerte

Anmerkung: f = frequency

Beispiel

Der Wert „6“ kommt insgesamt „3“-mal vor: $f_{k=6} = 3$

Die relative Häufigkeit ist daher (n=10): $f_{rel_{k=6}} = \frac{3}{10} = 0{,}3$

Die prozentuale Häufigkeit ergibt sich daraus: $\%_k = 0{,}3 * 100\% = 30\%$

2.3 Ausgleichung monovariater Häufigkeitsverteilungen

Zum Ausgleich einer **monovariaten Häufigkeitsverteilung** kann das Verfahren des **gleitenden Durchschnitts** verwendet werden. Die Ausschläge der Verteilung gehen bei diesem Verfahren verloren. Die Verteilung wird geglättet.

Zur Neuberechnung der Werte wird ein beliebiger Wert genommen und mit dem vorigen und dem folgenden addiert. Anschließend werden die drei addierten Werte durch drei geteilt. Man erhält den Durchschnittswert der drei Werte. f_i ist dabei der aktuell zu betrachtende Wert.

$$f_{i'} = \frac{f_{i-1} + f_i + f_{i+1}}{3}$$

2.4 Monovariate, bivariate und multivariate Häufigkeitsverteilung

Eine **monovariate** (**univariate**) Häufigkeitsverteilung bildet ein Merkmal mit den zugehörigen Häufigkeiten ab. Eine **bivariate** Häufigkeitsverteilung bildet die Häufigkeiten von zwei Merkmalen ab, während die **multivariate** Häufigkeitsverteilung drei oder mehr Merkmalshäufigkeiten abbildet.

2.5 Normalverteilung (NV) nach Gauß

Eine Normalverteilung ist **eingipflig** und **unimodal** mit einem **glockenförmigen** Verlauf mit einer **asymptotischen** Annährung an die x-Achse und einer **Symmetrie**. Die Werte für Median, Modus und arithmetischem Mittel sind gleich.

Die **Bedingung** für die Normalverteilung ist eine große Anzahl von **zufälligen Einflüssen**, die auf die Werte einwirken. Eine optimale Normalverteilung ergibt sich nur für den Fall des absoluten Zufalls.

Eigenschaften

Das arithmetische Mittel entspricht dem **Hochpunkt**, während die Streuung im Abstand von Mittelwert und **Wendepunkt** zu erkennen ist.

Die **Fläche der Normalverteilungsfunktion** ergibt die Wahrscheinlichkeit in dem entsprechenden Bereich. Im Abstand von ± 1s (einer Standardabweichung) liegen 68,26% der Werte. Im Bereich von ± 2s (zweifache Standardabweichung) liegen 95,44% der Werte.

Berechnung

Die **Normalverteilung** ergibt sich aus dem **Mittelwert** (arithmetisches Mittel) und der **Varianz**.

$$\boldsymbol{N(\mu;\sigma^2)} = f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Standardnormalverteilung

Die **Standardnormalverteilung** ergibt sich aus $\mathrm{V(x)}$ (**Variationskoeffizient**). Dabei ist der Mittelwert 0 und die Streuung liegt bei 1 (Wendepunkte bei -1 und +1). Mit Hilfe der Standardnormalverteilung können verschiedene Stichproben miteinander verglichen werden. Um die Werte der Normalverteilung in die Standardnormalverteilung zu transformieren ist eine z-Transformation nötig:

$$\boldsymbol{N(0;1)} z_i = \frac{x_i - \mu}{\sigma}$$

Außerdem kann eine Normalverteilung zunächst nur **zentriert** werden (Mittelwert bei null).

$$z_i = x_i - \mu$$

Im Gegenzug kann die Normalverteilung auch zunächst **normiert** werden (Streuungsverhältnis).

$$z_i = \frac{x_i}{\sigma}$$

Konfidenzintervall

Das **Konfidenzintervall** gibt eine Breite an Werten an. Innerhalb dieser Spanne liegen dann ein bestimmter Prozentsatz (95%, 99%, …) von Werten. Damit kann eine Aussage getroffen werden mit welcher Wahrscheinlichkeit bestimmte Werte auch tatsächlich vorkommen.

Um das Konfidenzintervall zu bestimmen muss die Normalverteilung der Stichprobenverteilung jedoch zunächst durch die z-Transformation in die **Standardnormalverteilung** umgeformt werden.

Graphisch dargestellt wird das Konfidenzintervall in Form eines **Fehlerbalkens**.

2.6. Weitere theoretische Verteilungen

Gleichverteilung

Bei einer **Gleichverteilung** ist die Wahrscheinlichkeit jeder Merkmalsausprägung **konstant gleich groß** (z.B. Würfel, Münze).
Im **diskreten Fall** hat die Gleichverteilung an jeder Merkmalsausprägung einen exakt gleich hohen Wert (Im Diagramm überall **gleich hohe Balken**).

$$G(x) = \frac{1}{n}$$

Im **stetigen Fall** ergibt sich in einem Intervall (a,b) ebenfalls ein gleich hoher Wert (Im Diagramm eine dauerhaft **gleich hohe Linie**).

$$G(a;b) = \frac{1}{b-a}$$

Binomialverteilung

Die **Binomialverteilung** ist eine **diskrete Verteilung**. Ähnlich der Normalverteilung ist sie **unimodal** und **eingipflig**. Eine **Symmetrie** herrscht nur wenn $p = 0{,}5$.Die Binomialverteilung beschreibt die Ergebnisse **mehrerer zufälliger unabhängiger Versuche**, bei dem ein Merkmal nur **zwei Ausprägungen** hat (z.B. Münze).

$$\boldsymbol{B(n;p)} = \binom{n}{k} \cdot p^k \cdot q^{n-k}$$

3. Darstellung von Daten

3.1. Tabellarische Darstellung

(a) Häufigkeitstabellen

Aus der **Urliste** (alle Rohwerte) kann zunächst eine **Häufigkeitstabelle** erstellt werden. In ihr wird jedes **Merkmal** in einer und die entsprechenden **Häufigkeiten** in einer anderen **Spalte** dargestellt. Jede Merkmalsausprägung mit der entsprechenden Häufigkeit ergibt eine **Zeile**.

In weiteren Spalten können auch **relative** oder **kumulierte** Werte stehen.

(c) Kreuztabelle

In einer **Kreuztabelle** werden mehrere Merkmale in einem **Zusammenhang** dargestellt. Die Merkmale befinden sich in der **ersten Zeile** und der **ersten Spalte**. In die entsprechenden **Zell-Kombinationen** werden dann die **Häufigkeiten** eingetragen.

(d) Linearisieren höherdimensionaler Strukturen

Unter der **Linearisierung** versteht man die Abbildung einer **mehrdimensionalen Struktur** in einer eindimensionalen Form. Im Normalfall wird dies durch **Verschachtelung** erreicht.

3.2. Graphische Darstellung

Graphische Darstellungen beinhalten immer die Gefahr der der **Verzerrung**! Daher sind einige Regeln (Richtwerte) zu beachten.

- Unter der **Schiefe** versteht man eine **Asymmetrie der Verteilung**. Ist der Gipfel weiter rechts von der Mitte, so wird die Verteilung als **linksschief** oder **rechtssteil** bezeichnet. Im umgekehrten Fall (weiter links von der Mitte) als **rechtsschief** oder **linkssteil**.

Die Schiefe kann auch (als Zahl) berechnet werden (Abweichung Median / Mittelwert). Eine **linkschiefe Verteilung** hat einen **negativen** Wert, während eine **rechtschiefe Verteilung** einen **positiven** Wert besitzt. Eine Symmetrie würde bei einem Wert gleich Null auftreten.

- Ein weiteres optisches Merkmal kann auch die **Ausprägung des Gipfels** sein. Ist dieser sehr breit wird die Verteilung als **breitgipflig**, ist sie eher dünn als **schmalgipflig** bezeichnet.

- Die **Kurtosis** beschreib dir Form des Gipfels. Ist dieser eher spitz zulaufend wird die Verteilung als **steilgipflig**, ist sie eher rund geformt als **breitgipflig** bezeichnet.

- Eine univariate Verteilung kann **eingipflig**, **zweigipflig** oder **mehrgipfig** sein. Die Gipligkeit ergibt sich über die **Anzahl der Hochpunkte**.

- Eine univariate Verteilung kann außerdem **unimodal**, **bimodal** oder **multimodal** sein. Die Modalität ergibt sich über die **Anzahl der Modalwerte**.

(a) Univariate graphische Darstellung

Eine **univariate Verteilung** lässt sich als **Histogramm** (**Balken- / Stabdiagramm**) darstellen. Auf der **x-Achse** (**Ordinate**) befinden sich die **Merkmalsausprägungen**, auf der **y-Achse** (**Abszisse**) die **Häufigkeiten**.

In einem **Histogramm** lassen sich neben **absoluten** und **relativen** auch **kumulierte** Werte (siehe b) eintragen.

(b) Kumulative graphische Darstellung

Eine **kumulative Verteilung** lässt sich als **Stufendiagramm** bzw. einer **Ogive** darstellen. In einer **diskreten Verteilung** (bzw. einer stetigen Verteilung in einer **sekundären Verteilungstafel**) entsteht ein **Stufendiagramm** (kategorisierte Kumulierung). Bei **stetigen Verteilungen** entsteht eine **Ogive** mit unendlich vielen Werten.

(c) Bivariate graphische Darstellung

Eine **bivariate Verteilung** lässt sich als **Punkt- bzw. Streudiagramm** darstellen. Auf den beiden Achsen (**Ordinate**, **Abszisse**) werden jeweils die beiden **Merkmalsausprägungen** eingetragen. Dadurch ergibt sich für jede kombinierte Merkmalsausprägung ein **Punkt** im Diagramm.

Um die **Häufigkeiten** der Punkte darzustellen gibt es verschiedene Möglichkeiten.

- Eine denkbare Möglichkeit wäre ein **dreidimensionales Diagramm**, in dem die x- und z-Achse die beiden Merkmale und die **y-Achse die Häufigkeiten** darstellt. Hier würde es sich um ein **dreidimensionales Stab- oder Balkendiagramm** handeln.
- Eine andere (einfache) Möglichkeit ist, die Punkte in **verschiedener Stärke** darzustellen (z.B. größer / kleiner).

(d) Prozessuale graphische Darstellung

Mögliche Darstellungen sind hier das **Mittelwertdiagramm mit Fehlerbalken** und der **Box- und Whisker-Plot** (bzw. Boxplot). Es gibt noch viele weitere.

3.3. Statistische Maßzahlen univariater Verteilungen

Zu den statistischen Kennwerten gehören die **zentrale Tendenz** (Modalwert, Medianwert & Arithmetisches Mittel) und **Dispersionsmaße** (Variationsbreite, Varianz & Standardabweichung / Streuung).

Die wichtigste Komponente für die statistischen Maßzahlen sind die **SQ-Werte**. Sie geben die **Summe der Abweichungsquadrate** vom Mittelwert an.

$$SQ = \sum_{i=1}^{n} (x_i - \bar{x})$$

3.3.1. Maße für die zentrale Tendenz

	stetig	**diskret**
metrisch	Arithmetisches Mittel Median (+ Quantile) Modalwert Geometrisches Mittel Harmonisches Mittel	Arithmetisches Mittel Median (+ Quantile) Modalwert Geometrisches Mittel Harmonisches Mittel
ordinal	Median (+ Quantile) Modalwert	Median (+ Quantile) Modalwert
nominal	-	-

(a) Arithmetisches Mittel

Das **arithmetische Mittel** oder der **Mittelwert** gibt den Durchschnitt der Verteilung an. Jeder Wert, der sich in der Verteilung ändert bewirkt eine unmittelbare Änderung des Mittelwertes (sensitiv). Um das arithmetische Mittel berechnen zu können, muss mindestens eine **Intervallskala** eine **Symmetrie** und eine **Normalverteilung** vorliegen.

Das arithmetische Mittel ist der Wert, bei dem die **Summe der quadrierten Abweichungen** aller Werte von diesem Mittelwert **minimal** wird.

$$\sum_{x=i}^{n}(x_i - \bar{x})^2 = Min$$

Die Summe der Abweichungen des arithmetischen Mittels ergibt daher immer null.

$$\sum_{x=i}^{n}(x_i - \bar{x}) = 0$$

Berechnung

- Berechnung des arithmetischen Mittels aus einer Gruppe ohne Kategorisierung der Werte (primäre Verteilungstafel).

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

- Berechnung des arithmetischen Mittels aus einer Gruppe mit Kategorisierung der Werte (sekundäre Verteilungstafel).

$$\bar{x}_{sek} = \frac{1}{n} \cdot \sum_{i=1}^{m} (f_i * x_i)$$

m : Anzahl der Kategorien
f_i : Häufigkeit in der Kategorie / Gruppe k
x_i : Kategorienmittel der Kategorie k

- Berechnung des arithmetischen Mittels mit mehreren gleich großen Gruppen

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{p} \bar{x}_i}{p}$$

p : Anzahl der der Mittelwerte bzw. Gruppen
$\bar{\bar{x}}$: Mittelwert der Mittelwerte
$\bar{x}_i$: Mittelwert der Gruppe i

- Berechnung des arithmetischen Mittels mit mehreren unterschiedlich großen Gruppen

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{p}(n_i * \bar{x}_i)}{\sum_{i=1}^{p} n_i}$$

n_i : Jeweilige Gruppengröße

(b) Media

Der **Median** oder auch **Zentralwert** halbiert die Verteilung. Über und unter dem Median liegen exakt gleich viele Messwerte. Damit ist die **Summe der Abweichungsbeträge** minimal (im Gegensatz zum arithmetischen Mittel bzw. Durchschnitt). Um den Median berechnen zu können, muss mindestens ein **ordinales Skalenniveau** vorliegen.

$$\sum_{x=i}^{n}|x_i - \bar{x}|^2 = Min$$

Beispiel:
Wenn die Werte 5,8,10,13,17 sind, dann wäre der Median bei **10**. Die Stabilität gegenüber der Abweichung wird deutlich, wenn man die Werte verändert: 1,2,10,16,22.
Der Median bleibt stabil bei **10**. Hat die Verteilung eine gerade Anzahl an Werten, dann wird die Mitte aus den beiden mittleren Werten gebildet. Bei 1,3,5,7,9,11 wäre der Median also:

$$Z = \frac{x_3 + x_4}{2} = \frac{7+5}{2} = 6$$

In einer **sekundären Verteilungstafel** wird die Berechnung etwas komplizierter. Zunächst wird bei den Häufigkeiten die Mitte gesucht. Anschließend wird die **untere Grenze** der entsprechenden Gruppe benötigt. Angenommen bei 1-3, 4-6, 7-9, 10-12 wäre der Median irgendwo in der 7-9 Gruppe. Dann wäre der untere Grenzwert **9,5**. Außerdem wir die **Klassenbreite** (in diesem Fall 3), der **Häufigkeit der Klasse** in der der Median liegt und die **kumulierte Häufigkeit der vorigen Klasse** gebraucht.

$$Z_{sek} = x_{ug} + b \cdot \frac{n/_2 - c_{fn}}{f_z}$$

x_{ug} : Untere Grenze der Klasse
b : Klassenbreite
f_z : Häufigkeit in der gefundenen Klasse
c_{fn} : Kumulierte Häufigkeit der vorigen Klasse

(c) Modalwert

Der **Modalwert** oder auch **Modus** ist der Wert, der in einer diskreten Verteilung am häufigsten vorkommt. In einer graphischen Darstellung zeigt sich der Modalwert als Maximum.

Hat die Verteilung mindestens ein ordinales Skalenniveau können mehrere gleich große Häufigkeitsgipfel **getrennt** voneinander auftreten. Die Bezeichnung ist wie folgt:
Eingipflig: **Unimodal**
Zweigipflig: **Bimodal**
Mehrgipflig: **Multimodal**

(d) Geometrisches Mittel

Das **geometrische Mittel** eignet sich für Größen, bei denen das Produkt anstelle der Summe interpretierbar ist. Es eignet sich zur Berechnung einer Steigung (bzw. **Zuwachsrate**). Beim geometrischen Mittel müssen alle Werte jedoch positiv sein. Das geometrische Mittel lässt sich nur bei **verhältnisskalierten Merkmalen** anwenden.

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

(e) Harmonisches Mittel

Wenn bei **Indexzahlen / Verhältniszahlen** (km/h, €/l, Einw./km², etc.) der Zähler konstant ist und die Werte zu mitteln sind, dann muss das harmonische Mittel verwendet werden. Ist der Nenner konstant wird das arithmetische Mittel verwendet. Das **harmonische Mittel** eignet sich somit für Größen, die in einem Bezug zu Einheiten stehen.

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

3.3.2. Quantile

Quantile sind **Unterteilungen einer Verteilung** in verschiedene Bereiche. Jeder Bereich ist der Häufigkeit nach **gleich groß**. Der **Median** bildet bereits zwei Quantile ab. Auf der einen Seite des Median befinden sich **50% der Werte**, auf der anderen Seite ebenfalls 50%. Eine weitere wichtige Unterteilung sind die **Quartile**, welche die Verteilung in **4 Bereiche** einteilen (zu je **25% der Werte**). Auch weitere Einteilungen in 10 Bereiche (**Perzentile**) oder 100 Bereiche (**Prozentile**) sind realisierbar.

Die Berechnung der **Quartile** erfolgt nach der Formel der **Medianberechnung**. Vor allem q_1 (Trennlinie zwischen dem 1. und 2. Quartil) und q_3 (Trennlinie zwischen dem 3. und 4. Quartil) sind von Bedeutung, da ihr Abstand, der **Innerquartilabstand** eine größere Bedeutung hat (siehe 3.3.3. f). Die Trennlinie q_2 ergibt sich aus der Berechnung des Medians und ist somit gleich Z.

$$q_1 = x_{ug} + b \cdot \frac{\frac{n}{4} - c_{fn}}{f_z}$$

$$q_3 = x_{ug} + b \cdot \frac{\frac{3n}{4} - c_{fn}}{f_z}$$

3.3.3. Streuungsmaße bzw. Dispersionsmaße (SQ-Werte)

	stetig	**diskret**
metrisch	Variationsweite Mittlere Abweichung Streuung Varianz Variationskoeffizient Quartilabstand	Variationsweite Mittlere Abweichung Streuung Varianz Variationskoeffizient Quartilabstand
ordinal	Variationsbreite Quartilabstand	Variationsbreite Quartilabstand
nominal	-	-

Die **Dispersionsmaße** bzw. **Dispersionsstatistiken** beschreiben, wie stark die einzelnen Werte in einer Verteilung vom Mittelwert abweichen. Die **Streuung** gibt an, wie stark die Werte vom Mittelwert abweichen. Ist die Streuung groß, dann weichen die Werte sehr stark vom Mittelwert ab.

(a) Variationsweite

Die **Variationsweite** (bzw. der **Range** oder die **Spannweite**) gibt an, wie groß der Bereich der Werte ist. Dazu wird von dem letzten vorkommenden Wert und von dem ersten vorkommendem Wert die Differenz gebildet.

$$w = x_{max} - x_{min}$$

(b) Mittlere Abweichung

Die mittlere Abweichung ergibt sich über die **Summe der Abweichungsbeträge** und gibt grundsätzlich an wie weit die Elemente der Verteilung vom Mittelwert abweichen.

$$\bar{d} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \bar{x}|$$

(c) Varianz

Die **Varianz** ist sehr wichtig, aber unanschaulich. Um die Varianz zu bilden wird die **Summe der quadrierten Abweichungen** durch die **Anzahl der Messwerte** minus eins. Je größer die Varianz, desto stärker ist die Abweichung der Werte vom Mittelwert (Streuung). Durch das Quadrieren fällt die Varianz deutlich höher aus, wenn sich die Streuung nur minimal erhöht.

- Zur Berechnung der Varianz in einer **primären Verteilungstafel** wird folgende Formel verwendet.

$$s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{SQ_x}{n-1} = \frac{SQ_x}{FG}$$

- Für eine **sekundäre Verteilungstafel** ist die Varianz wie folgt definiert.

$$s_x^2 = \frac{1}{n-1} \cdot \sum_{k=1}^{l} [(x_k - \bar{x})^2 \cdot f_k]$$

k : Aktuelle Klasse
l : Anzahl der Klassen
x_k : Median des entsprechenden Klassenbereichs
f_k : Absolute Häufigkeit der entsprechenden Klasse

- Neben der **Stichprobenvarianz**, kann auch die Formel zur Berechnung der **Populationsvarianz** ($N = \infty$) aufgestellt werden.

$$\sigma_x^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$$

(d) Standardabweichung

Die **Standardabweichung** oder **Streuung** wird aus der **Wurzel** der Varianz gebildet. Sie gibt den Abstand des Mittelwertes zum Wendepunkt einer Normalverteilung an. Die Standardabweichung gibt somit die Breite der Normalverteilung an.

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{SQ_x}{FG}}$$

(e) Variationskoeffizient

Um Statistiken mit verschiedenen Mittelwerten und unterschiedlichen Streuungen vergleichbar zu machen, gibt der **Variationskoeffizient** eine vergleichbare Größe an (z.B. aus verschiedenen Tests, die jedoch dieselbe Problematik untersuchen).

$$V = \frac{s}{\bar{x}}$$

Um bestimmte x-Werte zu vergleichen kann die Formel leicht verändert werden. Man spricht beim Ergebnis der Werte auch von **z-standardisierten Werten**.

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

Bei der z-Transformation entsteht aus einer Normalverteilung eine Standardnormalverteilung (bzw. auch **z-Verteilung**).

(f) Quartilabstand

Der **Quartilabstand** bzw. genauer der **Interquartilabstand** gibt den Abstand vom **ersten** zum **dritten** Quartil an und damit die inneren 50% einer Verteilung. In einem Box-Whisker-Plot werden z.B. die inneren 50% als Balken und die äußeren 50% als Linien dargestellt. Quartilabstände sind bereits ab **ordinal** skalierten Merkmalen möglich.

$$q = \frac{q_3 - q_1}{2}$$

3.4. Zusammenhangsmaße

	metrisch	**ordinal**	**nominal**	**dichotom**
metrisch	Maßkorrelation	Rangkorrelation	Kontingenz (K)	(Punkt-)Biseriale Korrelation
ordinal	Rangkorrelation	Rangkorrelation	Kontingenz (K)	Biseriale Rangkorrelation
nominal	Kontingenz (K)	Kontingenz (K)	Kontingenz (K)	Kontingenz (K)
dichotom	(Punkt-)Biseriale Korrelation	Biseriale Rangkorrelation	Kontingenz (K)	-

Kovarianz

Die **Kovarianz** gibt die Abweichung von zwei Werten im Produkt an. Die Abweichung des x-Wertes wird mit der Abweichung des y-Wertes für jeden x/y-Wert einzeln multipliziert. Die Kovarianz ist **nicht Standardisiert**!

„Die Kovarianz zweier Variablen ist das **durchschnittliche Abweichungsprodukt** aller Messwertepaare **von ihrem jeweiligen Mittelwert**." (QM1, 2010, S. 122)

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

Die im Zähler stehende Aufsummierung der Produkte wird als **Kreuzproduktsumme** bezeichnet.

Eigenschaften

- **Positive Kovarianz**

Erhöhen sich die jeweiligen Faktoren relativ stabil positiv oder negativ, dann wird das Ergebnis relativ stark positiv ausfallen (+ & + = + / - & - = +).

- **Negative Kovarianz**

Erhöhen sich die jeweiligen Faktoren entgegengesetzt, dann wird das Ergebnis relativ stark negativ ausfallen (+ & - = - / - & + = -).

- **Kovarianz von Null**

Erhöhen sich die Faktoren sehr unterschiedlich, dann lösen sich die Werte bei der Summierung auf. In diesem Fall wird das Ergebnis Null (bzw. nahe Null) und gibt es keinen Zusammenhang.

Maximale Kovarianz

Die Kovarianz von zwei Variablen kann in einem konkreten Fall nur ein bestimmtes Maximum annehmen. Dieses Maximum ergibt sich aus der x-Streuung multipliziert mit der y-Streuung.

$$|cov(max)| = \hat{\sigma}_x \cdot \hat{\sigma}_y$$

Überleitung zur Korrelation

Die Korrelation ist grundsätzlich die **Standardisierung** der Kovarianz. Elementar ist der **Korrelationskoeffizient** r.

Die Korrelation trifft keine Aussage über **kausale Zusammenhänge**. Ob x y beeinflusst, y x beeinflusst oder beide durch eine dritte Variable z beeinflusst werden kann mit der Korrelation nicht ermittelt werden.

3.4.1. Korrelation und Kontingenz (SP-Werte)

Die Zusammenhangsmaße basieren auf den SP-Werten (Kreuzprodukt).

$$SP_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Konvention für die **Effektstärke** des Korrelationskoeffizienten (nach Cohen 1988):

$r = 0{,}10$: Kleiner Effekt
$r = 0{,}30$: Mittlerer Effekt
$r = 0{,}50$: Großer Effekt

(a) Fechners Korrelationsindex

Fechners Korrelationsindex (r_f) ist eine sehr einfache Form der Korrelation. Hier werden die Anzahl der positiven und der negativen Kovarianz-Werte in ein Verhältnis gebracht.

$$r_f = \frac{n_k - n_d}{n_k + n_d}$$

n_k: Anzahl positive Abweichungsprodukte

n_d: Anzahl negative Abweichungsprodukte

Ein Abweichungsprodukt ergibt sich aus dem Kreuzprodukt: $(x_i - \bar{x}) \cdot (y_i - \bar{y})$

(b) Maßkorrelation nach Pearson

Die **Maßkorrelation** (bzw. **Produkt-Moment-Korrelation**) nach **Pearson** ist maßstabsunabhängig. Die konkrete errechnete Kovarianz wird einfach durch die **maximale Kovarianz** geteilt. Die Voraussetzung ist, dass beide Variablen **metrisch** und **Normalverteilt** sind. Außerdem gilt die

Maßkorrelation nur für **lineare Zusammenhänge** (nicht quadratische, kubisch, etc.). Für ordinal skalierte Merkmale eignet sich die Rangkorrelation (siehe d).

$$r_{xy} = \frac{cov_{emp}}{cov_{max}} = \frac{cov(x,y)}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

Der **Korrelationskoeffizient** (r_{xy}) liegt jetzt immer im Bereich zwischen -1 und +1.
-1 entspricht einem perfekten **negativen Zusammenhang**, +1 entspricht einem perfekten **positiven Zusammenhang** und 0 entspricht absolut **keinem Zusammenhang** (**Nullkorrelation**).

Wichtig: Der Korrelationskoeffizient ist nicht äquidistant (+1 ist nicht doppelt so viel wie +0,5).

(c) Bestimmtheitsmaß B

Das **Bestimmtheitsmaß** bzw. der **Determinationskoeffizient** ist ein **Gütemaß für die Korrelation**. Das Gütemaß gibt die **Abweichung (Varianz)** der Korrelationswerte (jedes y-Wertes auf der Verteilung der Korrelationswerte) an. Das Maß gibt an wie stark die Werte **vorherbestimmt** (determiniert) sind.

$$B = r_{xy}^2$$

*„Der Determinationskoeffizient gibt an, wie viel **Prozent** Varianz der **einen Variablen** durch die **andere aufgeklärt** wird.“* (QM1, 2010, S. 133) Damit wird also angegeben in welcher Qualität eine Vorhersage bzw. eine Modellbildung möglich ist.

(d) Rangkorrelation R nach Spearman

Die **Rangkorrelation nach Spearman** eignet sich für den Zusammenhang zwischen zwei **ordinalskalierten** Merkmalen. Allerdings müssen die Ränge trotzdem **äquidistante** sein, es reicht also nicht aus die Werte in ihrem Rohzustand nach ihrer Größe aufzulisten. Auch der Zusammenhang zwischen einem ordinal- und einem intervallskalierten Merkmal ist möglich, wenn das intervallskalierte Merkmal herabgestuft werden kann.

Zur Berechnung müssen die Häufigkeiten jedoch zunächst in Rangdaten geändert werden (wegen Äquidistanz). Also wird 12; 7; 45; 2 (Häufigkeiten) zu 3; 2; 4; 1 (Rangplätze). Diese Neuzuordnung muss für beide Merkmale erfolgen. Anschließend kann die Differenz d_i aus den Rängen gebildet werden. Kommen Ränge doppelt vor, so wird der mittlere Rang für alle verwendet. Also 1; 1; 2; 2; 3; 4; 5; 5;6 (Häufigkeiten) wird zu 1,5; 3,5; 5; 6; 7,5; 9 (Rangplätze).

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{N \cdot (N^2 - 1)}$$

d_i : Die Differenz der Rangplätze einer Untersuchungseinheit i (Konkret: $d = |x - y|$).
N : Anzahl der Untersuchungseinheiten

(e) Rangkorrelation T nach Kendall

Die **Rangkorrelation nach Kendall** hat vor allem den Vorteil auch bei **kleineren Stichproben** (n < 20) noch gute Ergebnisse zu liefern. Diese Korrelation bietet außerdem den Vorteil, dass **keine Normalverteilung** gegeben sein muss und eine **ordinale** Skalierung (im Gegensatz zu Pearson) wirklich ausreichend ist. Bei Kendall müssen jedoch alle Rangplätze getrennt vorkommen. Eine Doppelbelegung (wie bei Spearman) ist nicht möglich.

Zur **Berechnung** müssen die Werte des **ersten** Merkmals zunächst nach Größe sortiert werden. Dadurch verschieben sich die Werte des **zweiten** Merkmals automatisch auch. Nun werden die Werte des **zweiten** Merkmals von Rangposition 1 des **ersten** Merkmals aus betrachtet durchgegangen. Der erste Werte des **zweiten** Merkmals wird also nun betrachtet und es wird aufsummiert wie oft es vorkommt, dass bei den folgenden Werten ein Kleinerer Vorkommt (Inversion). Anschließend wird der nächste Wert des **zweiten** Merkmals betrachtet. Auch hier wird geschaut wie oft es vorkommt, dass es einen **kleineren Wert** als den aktuell Betrachteten unter den Folgenden gibt. Das wird bis zum Ende fortgeführt und anschließend werden alle **Inversionen** addiert zu I. Auf dieser Basis kann τ (tau) berechnet werden.

$$\tau = 1 - \frac{4 \cdot I}{n \cdot (n^2 - 1)}$$

I : Summe der Inversionen

(f) Kontingenzkoeffizienten: Phi, K

Die **Kontingenzkoeffizienten** ergeben sich maßgeblich aus dem χ^2 (**Chi-Quadrat**).

Das **Chi-Quadrat** gibt die Stärke des Zusammenhangs von mehreren Merkmalsausprägungen an. Das Chi-Quadrat verlangt nur eine **nominale Skalierung**. Bei der Berechnung werden die **beobachteten** mit den **erwarteten** Häufigkeiten in einen Bezug gebracht. Die erwarteten Häufigkeiten ergeben sich aus der Kreuztabelle.

In der **Kreuztabelle** werden die Felder mit **a bis d** gekennzeichnet. a ist das linke obere Feld. Im Uhrzeigersinn werden die Felder bis d benannt.

		Merkmal 1		
		Ausprägung 1	Ausprägung 2	
Merkmal 2	Ausprägung 1	4 (**5,2**)	6 (**4,8**)	10
	Ausprägung 2	10 (**8,8**)	7 (**8,2**)	17
		14	13	n = 27

$$E_a = \frac{(a+d) \cdot (a+b)}{n} = \frac{14 \cdot 10}{27} \approx 5{,}2$$

Nun wird der erwartete mit dem beobachteten Wert verglichen:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{b_i} - f_{e_i})}{f_{e_i}}$$

f_{b_i} : Die beobachteten Werte (Im Bsp. 4).
f_{e_i} : Die erwarteten Werte (Im Bsp. 5,2)
k : Die Anzahl der Zellen

Je größer also der Unterschied zwischen den erwarteten und den beobachteten Werten ist, **desto größer wird Chi-Quadrat**.

Die **Freiheitsgrade** (u.a. zum Ablesen des kritischen Wertes aus der Tabelle für den Chi-Quadrat-Test) ergeben sich aus der Anzahl der Zellen: $df = k - 1$

- **Der Kontingenzkoeffizient Cramers V (K)**

Cramers V gibt den Zusammenhang zwischen den beobachteten und erwarteten Werten, also den Chi-Quadrat-Wert in **standardisierter Form** wieder. Für q wird das **Minimum der Spalten- bzw. Zeilenanzahl** eingesetzt. Wenn also eine Kreuztabelle aus 5 Spalten und 4 Zeilen besteht, dann liegt das Minimum bei 4. Der Stichprobenumfang würde **3 Mal so stark** gewichtet werden.

Der Kontingenzkoeffizient K kann nur für Kreuztabellen mit **mindestens 2 Spalten und 2 Zeilen** angewendet werden (Also mindestens zwei Ausprägungen pro Merkmal). Für den Mindestfall von 2 Zeilen und 2 Spalten entspricht Cramers V dem Kontingenzkoeffizienten Phi.

$$K^2 = \frac{\chi^2}{n(q-1)}$$

q : Minimum der Zeilen/Spalten

Bei einem Wert von 0 gib es **keinen** Zusammenhang, ab 0,6 ist der Zusammenhang **relativ Stark** und bei 1 **perfekt**. Die Werte gelten auch für Phi!

- **Der Kontingenzkoeffizient Phi (φ)**

Für den **Minimalfall** von 2 Spalten und 2 Zeilen (**4-Felder-Tafel**) ergibt q den Wert 2 und im Nenner ergibt sich für die Klammer Null. **Phi** beschreibt also die standardisierte Form für den **vereinfachten Spezialfall** der Vierfeldertafel.

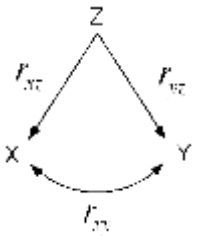
$$\varphi^2 = \frac{\chi^2}{N}$$

3.4.2. Spezielle Korrelationstechniken

(a) Partielle Korrelation

Bei einer Korrelation von zwei Variablen muss deshalb nicht sofort ein **kausaler Zusammenhang** bestehen. Um herauszufinden, wie der Zusammenhang von zwei Variablen unter dem Aspekt einer Dritten ist, kann man die **Partialkorrelation** verwenden. Hier ergibt sich ein **neuer Korrelationswert** unter Einbeziehung einer **dritten Variablen**.

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



(b) Multiple Korrelation

Unter der **multiplen Korrelation** versteht man das Einwirken **zweier Variablen auf eine**. Gibt es also einen Zusammenhang zwischen den Variablen **x und y** und einen Zusammenhang zwischen **x und z**, so kann mit der multiplen Korrelation der Zusammenhang von **y und z (gemeinsam) auf x** berechnet werden.

$$r_{x\cdot yz} = \sqrt{\frac{r_{xy}^2 + r_{xy}^2 - 2 \cdot r_{xy} . r_{yz} \cdot r_{xz}}{1 - r_{yz}^2}}$$

(c) Kanonische Korrelation

Die **kanonische Korrelation** dient allgemein der Aufklärung **wechselseitiger Beziehungen** (Abhängigkeiten) von zwei Variablengruppen. Die kanonische Korrelation wird daher eher in der **multivariaten Statistik** angewendet (→ 2. Semester).

(d) Biseriale Korrelation

Die **Biseriale Korrelation** eignet sich für den Zusammenhang zwischen **intervallskalierten** Merkmalen und **künstlich dichotomen** (vorher ebenfalls intervallskalierten) Merkmalen. Sie setzt normalverteilte Werte voraus.

$$r_{xy} = \frac{\bar{x}_p - \bar{x}_q}{s_x} \cdot \frac{p \cdot q}{y}$$

$\bar{x}_p$: Arithmetisches Mittel der Werte der 1. nominal dichotomen Klasse
$\bar{x}_q$: Arithmetisches Mittel der Werte der 2. nominal dichotomen Klasse
p : Prozentualer Anteil der Werte in der 1. nominal dichotomen Klasse
q : Prozentualer Anteil der Werte in der 2. nominal dichotomen Klasse
y : $\phi(z)$ bzw. $\phi(u)$ entspricht dem y-Wert in der Standardnormalverteilung (z-transformierten p-Wert einsetzen)

(e) Punktbiseriale Korrelation

Die **Biseriale Korrelation** eignet sich für den Zusammenhang zwischen **intervallskalierten** Merkmalen und **natürlich dichotomen** Merkmalen. Für eine punktbiseriale Korrelation muss keine Normalverteilung vorliegen.

$$r_{xy} = \frac{\bar{x}_p - \bar{x}_q}{s_x} \cdot \sqrt{p \cdot q}$$

$\bar{x}_p$: Arithmetisches Mittel der Werte der 1. nominal dichotomen Klasse
$\bar{x}_q$: Arithmetisches Mittel der Werte der 2. nominal dichotomen Klasse
p : Prozentualer Anteil der Werte in der 1. nominal dichotomen Klasse
q : Prozentualer Anteil der Werte in der 2. nominal dichotomen Klasse

(f) Biseriale Rangkorrelation

Die **Biseriale Rangkorrelation** eignet sich für den Zusammenhang zwischen **ordinalskalierten** Merkmalen und allgemeinen **dichotomen** Merkmalen. **X** bildet dabei die ordinale Variable ab, welche als **Rangplätze** angegeben werden muss.

$$r_{xy} = \frac{2}{n}(\bar{x}_p - \bar{x}_q)$$

$\bar{x}_p$: Arithmetisches Mittel der Rangplätze der 1. nominal dichotomen Klasse
$\bar{x}_q$: Arithmetisches Mittel der Rangplätze der 2. nominal dichotomen Klasse

3.4.3. Lineare Regression

(a) Korrelation und Kausalität

Falls eine Korrelation besteht, dann ist die Kausalität nicht eindeutig zuzuordnen. Es gibt mehrere denkbare Möglichkeiten für einen kausalen Zusammenhang.

x → y x beeinflusst y

x ← y y beeinflusst x

x ←→ y x und y werden von z beeinflusst

x ← z → y x und y beeinflussen sich gegenseitig

x → w → y y wird über w von x aus beeinflusst (Wirkungspfad)

(b) Das Modell der linearen Regression

Die **Regression** ist eine **Vorhersage** eines Merkmals y aus einem Merkmal x. Das zugrundeliegende Merkmal wird **Prädiktor** (unabhängige Variable, in diesem Fall x) und das vorhergesagte Merkmal **Kriterium** (abhängige Variable, in diesem Fall y) genannt. Geschrieben wird dieser Fall als eine **Regression von y auf x**. Diese Unterscheidung der Variablen passiert auf Grund der Unterteilung in **abhängige** (in diesem Fall x) und **unabhängige** (in diesem Fall y) Variable.

Die Regression ist deshalb **einfach**, weil es nur einen Prädiktor und ein Kriterium gibt. Gäbe es mehrere hieße sie **multiple Regression**. Außerdem wird in diesem Fall die Regression deshalb als **linear** bezeichnet, weil man einen linearen Zusammenhang der Variablen annimmt (also nicht exponentiell o.ä.).

- **Die Regressionsfunktion**

Die Funktion, welche der Regressionsgerade zugrunde liegt ist die **lineare Standardfunktion**. β_0 und β_1 bilden die **Regressionskoeffizienten**.

$$y = \beta_0 + \beta_1 \cdot x$$

β_0 : Regressionskonstante
β_1 : Anstieg

Auch eine umgekehrte Vorhersage (x auf y) ist möglich. In diesem Fall drehen sich bei der Berechnung x und y um. Zeichnet man beide Regressionsgeraden in ein Diagramm, so weicht die Zweite von der Ersten leicht ab (außer bei einer Korrelation von „1“).

$$x = \beta_0 + \beta_1 \cdot y$$

- **Eigenschaften der Regressionsgeraden**

Hat die Verteilung eine Kovarianz / Korrelation von „0“, dann ergibt sich für die Vorhersage der y-Werte eine Funktion, welche parallel zur x-Achse verläuft (Steigung = 0) und für die Vorhersage der x-

Werte eine Funktion, welche eine unendliche Steigung und sich damit nicht mit der y-Achse schneidet. Beide Geraden der zwei Sonderfälle stehen senkrecht aufeinander.

(c) Linearitätsproblem

Die **lineare Regression** ist eigentlich nur dort anzuwenden, wo auch ein **linearer Zusammenhang** zu erwarten ist. Ist der Zusammenhang z.B. ähnlich einer quadratischen Funktion, so kann es sein, dass man bei der linearen Regression keinen oder einen schwachen Anstieg in eine Richtung bekommt, obwohl der Zusammenhang in anderer Form eigentlich sehr ausgeprägt ist.

(d) Die Methode der kleinsten Quadrate

Grundsätzlich wird die Gerade so bestimmt, dass die Abstände aller Punkte **in der Summe minimal** sind ($y_i - \hat{y}_i$). Um Vorzeichenprobleme zu verhindern und stärkere Abweichungen auch deutlich stärker zu gewichten werden die **Abstände quadriert**.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = min!$$

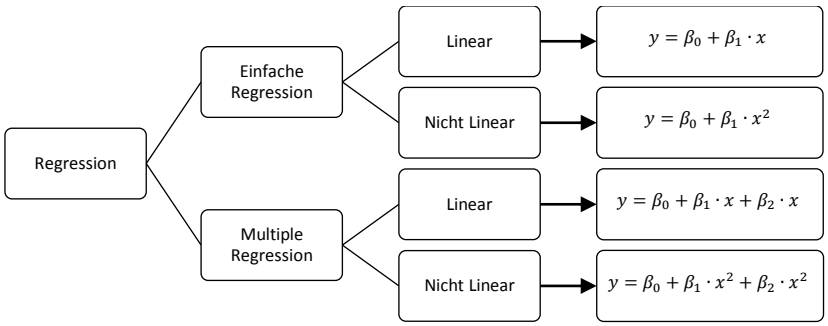
Aus dieser Grundannahme lassen sich die Werte für β_0 und β_1 aus der Regressionsfunktion bestimmen.

$$b_{yx} = \frac{SP_{xy}}{SQ_x} = \frac{cov(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

(e) Regressionsarten

Die Regressionen lassen sich nach mehreren Kriterien einteilen. Wichtig sind vor allem ob zwei oder mehr Variablen beteiligt sind und ob die Funktion linear und nicht linear ist.



(f) Vorhersagefehler und Güteabschätzung

„*Bei einem nicht perfekten Zusammenhang sind die Vorhersagen einer Regression fehlerhaft*." (QM1, 2010, S. 157)

Grundsätzlich ist eine Regression **kritisch** zu beurteilen, da sie unabhängig von der Streuung identisch bleiben kann. Eine fast lineare Punktwolke kann mit einer zufällig verteilten Punktwolke eine identische Regressionsgerade haben. Die **Residualvarianz** gibt Aufschluss über den **Fehlergrad** der Regression. Ist die Residualvarianz sehr hoch, so ist die Regression eher schlecht.

Es gibt drei Arten von Abweichung resultierend aus den drei verschiedenen Variablenformen. Einmal der **Wert** an sich (y_i), dann der **Mittelwert** ($\bar{y}_i$) und schließlich der **vorhergesagte Wert** aus der Regressionsgerade ($\hat{y}_i$). Daraus ergeben sich:

Varianz: $SQ_{ges} = \hat{\sigma}_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ mit $y_i - \bar{y}$

Regressionsvarianz: $SQ_{reg} = \hat{\sigma}_{\hat{y}}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n-1}$ mit $\hat{y}_i - \bar{y}$

Residualvarianz: $SQ_{res} = \hat{\sigma}_{[y/x]}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}$ mit $y_i - \hat{y}_i$

Der **Residualfehler** $y_i - \hat{y}_i$ wird auch als $\boldsymbol{\varepsilon_i^2}$ bezeichnet.

Die drei Varianzen stehen in einer additiven Verbindung:

$\hat{\sigma}_y^2 = \hat{\sigma}_{\hat{y}}^2 + \hat{\sigma}_{[y/x]}^2$ **bzw.** $SQ_{ges} = SQ_{reg} + SQ_{res}$

Bildlich:

SQ_Y Gesamt Total	=	**SQ** Regression Modell	+	**SQ** Residual Rest Fehler