**Kapitel Deskriptive Statistik**

1. Einführung
2. Nominalskala
3. Ordinalskala
4. Intervallskala I
5. Intervallskala II
6. Intervallskala III
7. Bivariate Zusammenhangsmaße I
8. Bivariate Zusammenhangsmaße II

**1. Einführung**

**Allgemein**

* Statistik Grundlage **empirischer Wissenschaft** (Empirie = Erfahrungswissen) → auf Messung und systematische Beobachtung beruhende Forschung
* empirische Forschung: Produktion **Daten** (meist in numerischer Form) → zur Bestätigung/Widerlegung Hypothese
* **Deskriptive Statistik** ist ein Verfahren zur Beschreibung von vielen Daten durch wenige Worte („**Kennwerte**") → Reduktion Datenmenge auf überschaubare Anzahl an Charaktersierungen

**Variablen**

* **Merkmale**: Isolierte Eigenschaft eines größeren Ganzen, z.B. Farbe, Intelligenz
* **Ausprägung**: Zustand des Merkmals, z.B. Farbe = grün, IQ =115
* ein Merkmal hat mind. 2 Ausprägungen, z.B. verbal (jung/alt), numerisch (0/1)
* **Merkmalsträger** (statistische Einheiten, Beobachtungseinheiten) sind alle Objekte mit Merkmalsausprägungen → in Humanwissenschaften meist Menschen oder Tiere
* **Beobachtungen**: Feststellung Merkmalsausprägung bei Merkmalsträgern; **im engeren Sinn**: echte Verhaltensbeobachtungen oder bildgebende Verfahren; **im weiteren Sinn**: Ergebnisse Leistungstest, Selbst-/Fremdauskünfte Fragebogen
* **Daten** sind sämtliche Beobachtung bei Informationssammlung
* **Statistik** sind Methoden zur Sammlung/Analyse der Daten
* zur Benutzbarkeit der Daten: **Variablen**; dabei werden Ausprägungsmerkmalen Zahlen zugeordnet → **Realisationen / Werte**
* **Variablen**: mit Großbuchstaben symbolisiert (X und Y); **Realisationen**: mit Kleinbuchstaben (x und y) → Bsp: Würfel: Augenzahl X, die Werte 2,5... x
* **Wertebereich**: Menge aller möglichen Realisationen
* Variablen müssen immer über mathematische Formulierung definiert werden
* **Extensionale Definition**: Aufzählung alle Realisationen (1)
* **Intensionale Definition**: gibt eine Vorschrift an, die die Variable eindeutig spezifiziert (aber nicht alle aufzählen) (2)

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Messung**: empirische Feststellung der Realisation einer Variablen
* Unterscheidung:

→ Beobachtung: Feststellung Merkmalsausprägung (verbale, bildliche Information)

→ Messung: Feststellung Werte einer Variable (numerische Information)

* Gemessene Zahlen heißen Messwerte oder Ergebnisse
* **Variablentypen**:
* **diskrete Variable**: besitzt endlich viele und feste Werte (Ganzzahlen);

→ **dichotome** Variable: mit genau zwei diskreten Werten (Frau/Mann);

→ **polytome** Variablen: mit mehr als zwei diskreten Werten

* stetige (kontinuierliche) Variable: kann (unendlich viele) beliebige Werte annehmen (reelle Zahlen), z. B. Körpergewicht
* Achtung: strenge Unterscheidung zwischen Typen von Merkmalen und Variablen, z.B. Alter als stetiges Merkmal, das diskret definiert wird (auch Intelligenz, Schulleistung...)
* stetige Variable eigentlich nicht vorhanden

**Skalen**

* Def.: Eine Skale ist die Festlegung von Einheiten, in denen ein gegebenes Merkmal gemessen wird
* Einheiten meist numerisch, kann aber auch andere Symbole haben (Farben)
* nur Variablen, die auf derselben Skala gemessen wurden, können verglichen werden; ansonsten Skalentransformation (Skalen ineinander überführen), z.B. von X zu Y(X) → Noten 1- 3 in Y gut, 4-6 in Y schlecht

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Skalenniveaus Typen:**
* **Nominalskala** (qualitativ)
* **Ordinalskala** (qualitativ/quantitativ)
* **Intervallskala** (quantitativ)
* **Verhältnisskala** (quantitativ)
* **Absolutskala** (quantitativ)

→ unterscheiden sich im Informationsgehalt; nimmt von Nominal- zur Absolutskala zu → **qualitativ**: zeigen Unterschiede auf d **quantitativ**: Zahlen sagen etwas über Größe von Unterschieden aus → Skalenniveau bestimmt die erlaubten mathematischen Operatoren (+,-, =), welche mathematischen Transformationen auf Variablen angewandt werden dürfen ohne Informationsverlust und welche statistische Verfahren auf Daten angewandt werden dürfen (Mittelwert...)

**2. Nominalskala**

* **Ziel**: Unterscheidung Kategorien durch Zuordnung von Zahlen für die Variablen; diese Zahlen sind vollständig beliebig und damit nicht interpretierbar
* die neuen Zahlen dürfen beliebig verändert werden (+, -, ...)
* Achtung: es dürfen nicht zwei Variablen ein neuer Wert zugeordnet werden (z.B. 2 und 3 dürfen nicht beide die 100 bekommen)
* **zulässige** **Operatoren**: ausschließlich Äquivalenzrelationen, d.h. **Gleich und Ungleich** → alle anderen Aussagen (x₁ größer x₂) ist unzulässig!
* **zulässige** **Transformationen**: eindeutige Abbildungen, sodass Unterscheidbarkeit erhalten bleibt

**Häufigkeiten**

* **Numerische** **Beschreibung**: Nominalskalierte Variablen sind praktisch immer diskret und endlich
* die empirische beobachtete Häufigkeit des Auftretens einer Ausprägung X=x wird als h (X=x) oder h(x) bezeichnet → **absolute Häufigkeit**
* f (X=x) bzw. f(x) ist definiert als Quotient aus absoluter Häufigkeit und der Anzahl n aller Beobachtungen (f(x) zwischen 0 und 1) → **relative Häufigkeit**
* **Achtung**: Relative Häufigkeiten sind nicht Wahrscheinlichkeiten (Prozent)

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

**Univariate Kreuztabellen**

* 1. Spalte: Wert der Variablen 2. Spalte: Häufigkeit (absolute und relative)
* Die Sammlung der Werte der absoluten und relativen Häufigkeit wird als **diskrete** **Häufigkeitsverteilung** bezeichnet
* Die tabellarische Darstellung erfolgt über **Häufigkeitstabellen**

**Bivariate Kreuztabellen**

* oft Betrachtung Häufigkeiten für das gemeinsame Auftreten zweier Merkmale
* Bsp.: Frauen/Männer unter- /normal- /übergewichtig → X: Geschlecht (x₁, x₂) Y: Gewichtstatus (y₁, y₂, y₃)
* Kombination Häufigkeiten → Verbundhäufigkeiten
* **absolute Verbundhäufigkeiten** werden als h(X=x, Y=y) bzw. h(x, y) bezeichnet
* **relative Verbundhäufigkeite**n als f (X=x, Y=y) bzw. f(x, y)
* Tabellarische Darstellung über **Kreuztabelle** (oder **Kontigenztabellen**)
* **Randhäufigkeiten**: Summe der einzelnen Variablen (Männer, Frauen, Übergewicht, Normalgewicht, Untergewicht); Addition jeder einzelnen Spalte und Zeile → Gesamtanzahl (Stichprobe) ist **n**

**Multivariate Kreuztabellen**

* auch gemeinsames Vorkommen von mehr als zwei Merkmalen
* Absolute Verbundhäufigkeiten werden hier als h(X=×, Y=y, ...) bzw. h(x, y,...) bezeichnet
* Relative Verbundhäufigkeiten mit f(...)
* Tabellarische Darstellung über **geschachtelte Kreuztabelle**

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Tipp: Bei mehr als 4 Variablen mehrere Tabellen
* Anzahl Zellen/Kombinationen: Multiplikation der Variablen (hier: Gewicht 3 \* Geschlecht 2 \* Freizeit 2 = 12)

**Kennwerte**

* **Def**.: statistisches Maß, das eine Menge von Beobachtungen über zumeist nur eine Zahl beschreibt
* **Ziel**: Datenreduktion
* charakterisieren nur bestimmte Eigenschaften der Menge von Beobachtungen → **Informationsverlust**
* **Achtung**: Mittelwert ergibt keinen Sinn, weil Daten nominalskaliert (1,2,3 oder 100,-5, 8 ergeben komplett anderen Mittelwert)
* ein Kennwert für nominalskalierte Daten ist der **Modalwert** (oder „**Modus**") → bezeichnet den unter den Beobachtungen am häufigsten vorkommenden Wert



* Wichtig: Der Modalwert ist nicht die Häufigkeit des Messwertes, sondern der Wert selbst
* Bei mehreren Maxima sinkt die Aussagekraft von dem Modalwert

**Diagramme**

* allgemein: Kategorien/Beschriftungen/Skala nicht vergessen
* **Kreisdiagramm**: stellt die absoluten oder relativen Häufigkeiten von Klassen als Kreissegmente eines Vollkreises dar; die Summe der Tortenstücke sollte wieder 360° ergeben; der Öffnungswinkel (alpha) eines Tortenstück wird so berechnet:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Säulen-/Balkendiagramm**: stellt die absoluten oder relativen Häufigkeiten von Werten einer Variablen als Balken (waagerecht) oder Säulen (senkrecht) dar; die verschiedene möglichen Werte werden auch als **Klassen** bezeichnet; Länge der Säulen/Balken: relative oder absolute Häufigkeit; die Breite variiert niemals!

**3. Ordinalskala**

* es können die Realisationen einer Variablen (natürlich) **geordnet** werden
* die Zuordnung der Zahlen zu den Ausprägungen des Merkmals spiegelt deren Ordnung wider
* Numerische Abstände zwischen den Realisationen können **nicht** interpretiert werden → Reihenfolge wichtig, Abstände können größer/kleiner sein
* Anwendung von Rechenoperationen auf die Werte ist an sich erlaubt, aber eigentlich nicht sinnvoll → nur > oder < interpretierbar
* **Zulässige** **Operationen**: Äquivalenzrelationen (Gleich und Ungleich) und qualitative Vergleichsrelationen (Größer oder Kleiner) → **Wichtig**: keine quantitativen Vergleiche (Abstände zwischen Variablen ist nicht unbedingt real)
* **Zulässige Transformationen**: alle streng monotonen Transformationen, sodass die Rangordnung der Werte erhalten bleibt (streng, weil keine zwei Werte zu einem neuen Wert transformiert werden dürfen; die Reihenfolge würde so aber trotzdem erhalten bleiben)
* **Kritische Betrachtung**: bei Ordinalskalen oder höhere Skalenniveaus können **Intransitivitäten** auftreten → eine angenommene Ordnung gilt nicht für bestimmte einzelne Paarungen (Bsp.: Nahrungskette Mensch isst Hund, Hund isst Ratte, Ratte isst Mensch (wenn tot) → ordinalskalierte Ordnung besteht nicht); **Lösungen**: Annahme eines niedrigeren Skalenniveaus oder Einführung neuer Skalenstufen (Mensch vorne ungleich Mensch hinten)

**Häufigkeiten**

* Ordinalskalierte Variablen sind sehr häufig diskret (abzählbar) und endlich
* es gelten die bereits eingeführten Notationen und Berechnungsvorschriften für Häufigkeiten
* Neben der **Häufigkeitsverteilung** kann auch noch die **empirische** **Verteilungsfunktion** bestimmt werden → gibt an, wie viele Beobachtungen kleiner oder gleich eine bestimmte Realisation sind; die Realisationen müssen hierfür **der Größe nach geordnet** werden
* **Empirische Verteilungsfunktion**: Aufsummieren dieser und alle vorherigen Wahrscheinlichkeiten (relative und absolute Häufigkeiten)

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Bsp.: Alkoholspiegel ist ordinalskaliert, weil durch Rundungen man nicht sagen kann, dass jemand mit 2% doppelt so hoch sind wie 1%
* Verteilungsfunktion: Frage, wer schlechter/so gut/besser war → Berechnung:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Für Ordinaldaten gelten gleiche Regeln zur Erstellung von Kreuztabellen

**Kennwerte**

* Maße der zentralen Tendenz: **Modalwert** und **Median**
* Andere Lagemaße: **Extrema (Minimum, Maximum)** und **Quantile, Quantilsrang**
* **Generell** **gilt**: Kennwerte für niedrigere Skalenniveaus sind für alle höheren anwendbar
* **Modalwert**: genau so definiert wie für nominalskalierte Daten; Realisation x der Variablen X mit dem häufigsten Vorkommen h(x); **aber**: sagt mehr über Häufigkeitsverteilung aus → markiert die Lage der häufigsten Realisation relativ zu den anderen Realisationen; man kann uni-, bi- und multimodale Verteilungen unterscheiden (Bimodal auch mit lokalen Maxima)
* **Median**: ist der Wert, der einen Datensatz mit n Messwerten in zwei gleich große Hälften teilt (50: 50); stimmt häufig mit keinem tatsächlichen Messwert überein; ist **äquivariant** (gleichbleibend) gegenüber gewissen Transformationen, insbesondere bei Addition und Multiplikation einer Konstanten; besitzt folgende Eigenschaften: 1. Mindestens 0.5\*n (50%) der Messwerte sind kleiner oder gleich dem Median 2. Mindestens 0.5\*n (50%) der Messwerte sind größer oder gleich **→ Bei gerader Anzahl den Mittelwert der zwei mittleren Werte nehmen; Notation:**

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Quantile**: Ergebnis in Personen
  + sind Zahlen, die einen Datensatz in bestimmtem Verhältnis teilt
  + **Median ist Spezialfall eines Quantils (2. Quartil)**
  + je nach Anzahl von Unterteilungen unterscheidet man **Perzentile** (100er Einteilung), **Dezentile** (10er Einteilung) und **Quartile** (4er Einteilung)
  + 25% Perzentil → unteres Quartil; 75% Perzentil → oberes Quartil;
  + besitzt folgende Eigenschaften:

1. Mindestens n\*p Messwerte sind kleiner oder gleich dem Quantil

2. Mindestens n\*(1-p) Messwerte sind größer oder gleich dem Quantil

* + **Notation:**

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* + **Achtung**: Bei Werten 2,3 und 2,4 nicht 2,35 sondern 2,325 → Berechnungsvorschriften für Quantile unterschiedlich definiert
* **Quantilsrang**: Ergebnis in Prozent
  + entgegengesetzte Sichtweise zu Quantil
  + Bsp.: Leistungstest mit Wert 105 Punkten; Wie viele Personen in der Stichprobe sind nun besser/schlechter?
  + bei der Rangbildung von n Messwerten einer Variablen X können maximal n Ränge vergeben werden
  + niedrigste Ausprägung Rangplatz 1, der höchste Rangplatz n **(kleinere Zahl = kleinerer Rang)**
  + wenn zwei gleiche Ausprägungen bekommen beide Werte gleichen Rangplatz (statt 6,7,8,9... dann 6,7,7,8...) → bei mehreren gleichen Werten („**Ties**" ) wird **mittlerer** **Rangplatz** nach der Regel vergeben:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* + Berechnung des Quantilsrangs:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

→ Abzug 0.5 Erklärung: es gibt nicht 0% Quantilsrang → Problem: Quantilsrang geht nicht von 0 bis 1, sondern liegt in schmalerem Bereich

* + **Korrekturformel**:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* + **einfachere** **Korrekturformel**:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* + **Quantil.Inkl = 0 bis 1!**

Quantil.Exkl = nicht ganz 0 bis 1

* **Extrema**:
  + sind die Grenzen des beobachteten Bereichs von Messwerten (unterster und oberste Wert)
  + können als 0. und 4. Quartil bezeichnet werden
  + **Achtung**: Extrema sind nicht notwendigerweise identisch mit Grenzen des Wertebereichs (sondern nur kleinster/ größter erhobene Wert)

**Grafische Darstellung**

* für **empirische Häufigkeitsverteilung**: dieselben Darstellungen wie bei Nominaldaten, **Achtung**: die Abfolge der x-Achse ist geordnet!
* für **empirische Verteilungsfunktion**: Aufsummieren relativer Häufigkeiten

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

→ Balkendiagramm an letzter Stelle voll ausgefüllt

**4. Intervallskala I**

* es wird eine **Einheit** (festgelegter Abstand zwischen zwei Werten) definiert
* es existiert **kein natürlicher Nullpunkt**
* **Differenzen von Werten** können verglichen werden, nicht aber die Werte selbst
* Bsp.: Temperaturskala → Aussage: es sind 10°C mehr, nicht doppelt so viel (Kelvin Skala würde dem widersprechen)
* wird am häufigsten bei empirischen Untersuchungen angenommen
* Intervallskalierte Variablen können diskret **oder** stetig sein
* **Kritische Betrachtung**: die bekanntesten und am meisten verbreiteten statistischen Verfahren setzen eine Intervallskala voraus; der Umgang mit niedrigen Skalenniveaus ist mathematisch oftmals weitaus komplexer; die ungeprüfte Annahme der Intervallskala ist oft problematisch
* **Zulässige Transformationen**: Äquivalenzrelationen (Gleich und Ungleich); qualitative Vergleichsrelationen (Größer oder Kleiner); quantitative Vergleichsrelationen, die sich auf **Differenzen** beziehen (nicht Wert A doppelt so groß wie Wert B); alle **linearen** Transformationen (Grundrechenarten) zulässig, sodass Verhältnisse zwischen Differenzen erhalten bleiben
* **Problem Häufigkeit**: Intervallskalierte Variablen können u.U. beliebige Ausprägungen besitzen, die sich nicht mehr sinnvoll in einer Tabelle darstellen lassen → zu viele Realisationen; Bsp: Körpergrößen; **Lösung**: es muss eine **Aggregation** vieler Realisationen in wenige Kategorien („**Klassen**") stattfinden

**Klassenbildung**

* Transformation Variable X in eine neue Variable C:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Wie viele Klassen sollen gebildet werden? Wie werden die Grenzen dieser Klassen festgelegt?

→ es kommt anscheinend darauf an, ob die Daten **diskret oder stetig** sind

* zur Bestimmung der Anzahl k der Klassen gibt es viele Formeln; **Faustregel**:

5-50 n → 5-8 k

50-100 n → 6-10 k

100-250 n → 7-12 k

>250 n → 8-25 k

* **einfache Formel**: [ ] = Aufrundung:



→ statt der Beobachtungen n wird zuweilen auch die Anzahl der Realisationen verwendet

→ man kann sich auch die Werte und Extrema anschauen

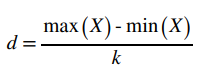
* die k Messwertklassen dürfen sich nicht überschneiden, sie sind also **wechselseitig** **ausschließend**
* die unteren und oberen Klassengrenze UG und OG der Variable j gehören zur Klasse c, die OG von j-1 aber nicht

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

→ 1.Eckige Klammer: Wert gehört zur Klasse 2.Runde Klammer vorne: Wert gehört zur vorigen Klasse → Grund der Unterscheidung: bei stetigen Daten (4,9999) → Problem bei 1. ; es muss 2. angewendet werden

* alle Massen haben im Normalfall **dieselbe Breite**
* Anzahl der Klassen: möglichst wenige leere Klassen; keine in den Daten enthaltene Informationen "herausaggregieren"
* **Klassenbreite** **d**:



→ X ist die ursprüngliche intervallskalierte Variable; es können die beobachteten Messwerte oder die Realisationen von X verwendet werden

* bei der Berechnung der Klassenbreite muss auf **Ausreißer** geachtet werden
* bei diskreten Daten werden die Klassengrenzen **nicht-überlappend** angegeben
* bei **kontinuierlichen/stetigen Daten** werden die Klassengrenzen überlappend angegeben (die obere Grenze gehört zur Klasse, die untere aber nicht (Ausnahme 1. Klasse))
* Die Klassenbreite allgemein ist:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Achtung**: für die unterste Klasse (j=1) nimmt man UG1

**Histogramm**

* stellt die Häufigkeiten vieler Realisationen in einem Säulendiagram mit **weniger Klassen als Realisationen** dar
* Klassen müssen nicht notwendig gleich breit sein
* es gelten dieselben Faustregeln wie bei der Kreuztabelle
* es können absolute oder relative Häufigkeiten sein (absolutes oder relatives Histogramm)
* bei gleichen Klassenbreiten zeigt zumeist die **Höhe einer Säule** die Häufigkeit der Elemente in der Klasse
* beim Histogramm: Säulen **ohne** Abstand!
* warum gleich breit? → Flächenbewertung würde Ergebnis verfälschen
* **Regel**: Wählt man ungleiche Klassenbreiten, muss das Histogramm normiert werden (wegen der Flächenbeurteilung der menschlichen Wahrnehmung)
* Wenn nicht die Höhe, sondern die Fläche Aj einer Säule die Häufigkeit repräsentieren soll, gilt für eine Klasse xj:

Ein Bild, das Text, Person enthält.

Automatisch generierte Beschreibung

* Dies gilt auch für die Darstellung mit absoluten Häufigkeiten h(xj)
* Gründe gegen Histogrammnormierung: Ausreißerproblem (Lösung: weitere Klasse für Ausreißer)

**Grafische Darstellung**

* je nach Anzahl der (lokalen) Maxima unterscheidet man **uni-, bi- und multimodale** **Verteilungen**
* **Histogramme**: Achtung, die Wahl der Klassen kann für die Aussage entscheidend sein (z.B. zu wenige Klassen → Informationen gehen verloren)
* **Schiefe**:
* **Symmetrische Verteilungen**: Häufigkeiten für die Ausprägungen einer Zufallsvariablen verlaufen (annähernd) gleichartig um den Mittelpunkt
* **Linkssteile / rechtsschiefe Verteilungen**: Häufigkeiten laufen rechts des Mittelpunktes flacher aus
* **Rechtssteile / linksschiefe Verteilungen**: Häufigkeiten laufen links des Mittelpunktes flacher aus

→ Schiefe erst jetzt, weil bei Ordinalskala Abstand ×-Werte keine Aussage

**5. Intervallskala II**

**Kennwerte**

* Maße der zentralen Tendenz
* Modalwert
* Median
* Mittelwert
* Andere Lagemaße
* Extrema
* Quantile, Quantilsrang

→ Modus, Median, Quantile und Quantilsrang identisch definiert wie bei Nominal- bzw. Ordinalskala → dieselben Regeln der Berechnung und Interpretation → **Äquivarianz** gegenüber linearen Transformationen bleibt erhalten

 Omega: Kennwert

* **Mittelwert**:

Ein Bild, das Text, Antenne enthält.

Automatisch generierte Beschreibung

* „× quer" ist durch extreme Werte beeinflussbar (ausreißerempfindlich)
* Ist der Schwerpunkt der Beobachtungen, d.h.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Der Mittelwert stimmt häufig mit keiner beobachteten Realisation überein
* Der Mittelwert ist äquivariant gegenüber bestimmten (z.B. linearen) Transformationen
* Insbesondere

1. Addition einer Konstante b zu alle n Beobachtungen



2. Multiplikation aller n Beobachtungen mit einer Konstanten a



* **Lageregeln**:
* für die Maße der zentralen Tendenz bei unimodalen Verteilungen
* Bei symmetrischen Verteilungen:



* Bei linkssteilen Verteilungen:



* Bei rechtssteilen Verteilungen:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Streuungsmaße (Dispersionsmaße)**
* Spannweite
* neben Lage der Verteilung noch eine wichtige Information
* Die Spannweite d ist die Differenz zwischen dem kleinsten und größten Wert aller n Messwerte
* Sie ist definiert als:



* Die Spannweite ist ausreißerempfindlich
* Die Spannweite ist eher uninformativ, da sie nur zwei von n Messwerten berücksichtigt (keine Information über Verteilung zwischen Minimum und Maximum)
* Interquartilsabstand
* Der Interquartilsabstand d ist die Differenz zwischen dem 1. und 3. Quartil
* Er ist definiert als:

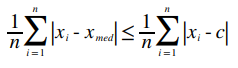


* Manchmal wird er auch ein halber Interquartilsabstand berechnet als d/2
* Ausreißerunempfindlich, weil es die inneren 50% abbildet, unabhängig davon, ob es außen Ausreißer gibt
* noch nicht bei Ordinalskala erlaubt, weil Differenz dort nicht erlaubt (da Differenz keine Aussage hätte)
* Mittlere Differenz/ Mittlere Abweichung
* Als mittlere Abweichung (MD) von n Beobachtungen in einem Datensatz wird die Summe aller Abweichungsbeträge zum Median bezeichnet

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* immer positive Zahl; je weiter ×-Wert, desto größer wird MD
* „durchschnittliche Abweichung vom Median, die auf jeden einzelnen Wert zurückgeht"
* sagt etwas über die Streuung der Daten aus (größere Streuung, größere mittlere Abweichung)
* Für jeden anderen Wert als für den Median ist der mittlere Abweichungsbetrag größer, d. h.:



→ wenn verschiedene Werte für c eingesetzt werden, ergibt sich eine Parabel mit dem Minimum beim Median

→ der Median minimiert die Summe (Herleitung siehe Unterlagen)

* (Abweichungs-)Quadratsumme
* Probleme mittlere Abweichung: Beträge "blöd" zum Rechnen; beinhaltet nur Median (Ordinaldaten), nicht Mittelwert (Intervallskala)
* Die Abweichungsquadratsumme (oder auch: Fehlerquadratsumme oder einfach Quadratsumme) ist die Summe der quadrierten Abweichungen aller n Beobachtungen vom Mittelwert

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Erfasst die Streuung um den Mittelwert
* Nur falls keine Streuung besteht, ist QS = 0, d.h. alle beobachteten Werten sind gleich; sonst: QS>0
* Je größer die Streuung, desto größer die QS
* Problem: Die Fehlerquadratsumme wird um so größer, je mehr Beobachtungen vorliegen
* Stichprobengröße muss berücksichtigt werden (kann nicht einfach verglichen werden)
* Für jeden anderen Wert als den Mittelwert ist die Summe der Abweichungsquadrate höher

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* Der Mittelwert minimiert also die quadrierten Abweichungen alle Beobachtungen
* Varianz
* Die Varianz ist das mittlere Abweichungsquadrat aller n Beobachtungen vom Mittelwert

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* Erfasst die mittlere quadrierte Streuung um den Mittelwert
* Nur falls keine Streuung besteht, ist s^2=0, d. h. alle beobachteten Werte sind gleich; sonst s^2> 0
* Je größer die Streuung, desto größer ist die Varianz
* Ist anfällig gegenüber Ausreißern
* Die Formel lässt sich leicht umformen:

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* Die Varianz ist also die Differenz des Mittelwerts der quadrierten Daten und dem quadrierten Mittelwert der Daten
* Dies wird auch als **Momentenschreibweise der Varianz** bezeichnet
* Standardabweichung
* besser als Varianz, weil Varianz nicht äquivariant (wenn x\*10, wird Varianz \*100)
* Problem: Die Varianz ist nicht äquivariant zu erlaubten Skalentransformationen



* Durch Wurzelziehen erhält man die Standardabweichung (SD, standard deviation)

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* Die Standardabweichung ist äquivariant zu den erlaubten Skalentransformationen → Verhalten von Varianz und Standardabweichung bei Transformationen der n Beobachtungen

1. Die Addition einer Konstanten a zu allen Werten x verändert Varianz und Standardabweichung nicht

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

1. Die Multiplikation aller Werte x mit einer Konstanten a führt zu einer Erhöhung der Varianz um a² und der Standardabweichung um a

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

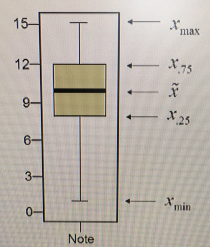
**Das Skalenniveau dichotomer Variablen**

* Ausprägungen
* Möglichkeiten der Skalenniveaus: → Nominalskala: (Un)gleichheit (Ungleichheit Datenwerte nicht Ungleichheit Ausprägungen (Haselnuss = Noccioli)) → Ordinalskala: Ordnung/Transitivität (A>B>C!) → Intervallskala: Abstände (Achtung: Schulnoten nicht wirklich Intervallskala, weil Abstand 1 zu 2 und 4 zu 5 anders)
* Intransitivität braucht immer mindestens 3 Werte; es gibt nur einen Abstand (zwischen den zwei Werten) → **dichotome Variablen** **als intervallskaliert** (natürliche Ordnung nicht so wichtig nehmen)
* Besonders gut Ausprägungen mit 0 und 1 → Mittelwert der gemessenen Daten ist identisch mit der relativen Häufigkeit (siehe Herleitung Unterlagen)

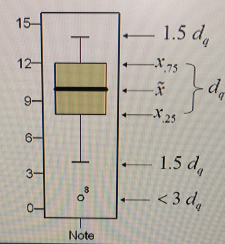
→ bei Interpretation darauf achten, wie die beiden möglichen Ausprägungen kodiert wurden (0 und 1 **oder** 1 und 0)

**Grafische Darstellung**

* **Box-Whisker-Plot**
* Mithilfe der **Fünf-Punkte-Zusammenfassung** (Minimum, 25%, Median, 75%, Maximum) können Daten grafisch am **Boxplot** veranschaulicht werden



* Diese Variante ist problematisch, weil Ausreißer die Länge der Whisker erheblich vergrößern können
* Eine zweite, häufiger verwendete Variante des Boxplots verwendet den 1.5fachen Interquartilsabstand d für die Länge der Whisker
* Box und Whisker enden am letzten Datenpunkt innerhalb ihrer Reichweite (auch wenn der Whisker länger gehen würde)
* Datenpunkte außerhalb der Whisker werden explizit eingetragen (o) (oft Nr. der Person dazu)
* Ausreißer > 3d werden mit Sternchen (\*) markiert (oft Nr. der Person dazu)



* **Das Fehlerbalkendiagramm**
* Das Fehlerbalkendiagramm (Error Bar) veranschaulicht Mittelwerte und die Streuung von Daten für mindestens eine Stichprobe
* Für die Länge der Fehlerbalken existieren verschiedene Konventionen (+/- 1\*SD, +/- 1,96\*SD, +/- 2,58\*SD) → verschiedene Konventionen für die Länge der Whisker

→ **Mittelwert: Kennwert der Intervallskala**

→ **Fehlerbalkendiagramm**: **Darstellung der Intervallskala**

**6. Intervallskala III**

**Mittelwert und Varianz aus kategorisierten Daten**

* notwendig, wenn nur kategorisierte Daten/Histogramme
* wenn man einen einzelnen Wert aus einer Kategorie haben möchte, nimmt man den Mittelwert der Kategorie (20-30 → 25) → das gilt für alle Werte in diese Kategorie (z.B. 7 Werte darin → alle 25)

→ ergibt neue geschätzte Datenreihe mit nur Mittelwerte der Kategorien

→ neue SD/MW/QS/Varianz

* Mittelwert mit den Klassen C:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Varianz mit den Klassen C:

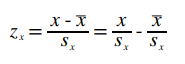
Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Standardabweichung: Wurzel der Varianz ziehen
* Quadratsumme: Varianz mit n multiplizieren
* Liegen intervallskalierte Daten bereits in kategorisierte Form vor, so können daraus Mittelwert und Varianz näherungsweise bestimmt werden
* Es sei x = (OG + UG) /2; gilt sowohl bei überlappenden als auch nicht überlappenden Grenzen
* bei nicht überlappend mit Ganzzahlen: sehr nah an Werten, die auf Originaldaten basieren

**z-Standardisierung**

* Ziel: Angabe der relativen Lage von Werten in einer Verteilung (Vergleichbarkeit mit anderen Verteilungen) 1. Quantile: wie bereits gesehen 2. Angabe einer normierten Differenz eines Messwertes zum Mittelwert
* Berechnungsvorschrift: Jede Differenz durch die Standardabweichung aller Messwerte geteilt. Die erhaltenen Werte werden als z-Werte bezeichnet



* Der Mittelwert von z-Werten ist immer 0
* Die Standardabweichung von z-Werten ist immer 1
* **Skalentransformation:**
* Mithilfe der z-Transformation können Messdaten mit einem beliebigen Mittelwert und Standardabweichung in Daten transformiert werden, die einen definierten Mittelwert und Standardabweichung aufweisen
* Schritt 1: z-Standardisierung jedes Datenpunktes
* Schritt 2: Transformation jedes Datenpunktes in die neue Skala



**7. Bivariate Zusammenhangsmaße I**

* Bisher wurden Kennwerte für den **univariaten** **Fall** betrachtet, d.h. für Daten **einer** Variablen
* Mit geschachtelten Kontigenztabellen wurde eine kompakte Darstellungsmöglichkeit für den **multivarianten** **Fall** beschrieben, d.h. für Daten **mehrerer** Variablen
* In der Statistik sind weitere Verfahren gebräuchlich, die speziell den Zusammenhang **zweier Variablen (bivariate Fall)** beschreiben

**Grundlagen**

* Intervallskala trägt Informationen über die Ordnung von Ausprägungen und hat feste Einheit dazwischen
* Die Werte sind nicht direkt vergleichbar, nur Unterschiede zwischen Werten
* Intervallskalierte Daten leicht grafisch oder numerisch behandelbar

**Scatterplot (Grafische Beschreibung) -> erst ab Intervallskalenniveau; linearer Zusammenhang**

* jeder Wert hat zwei Variablen x und y
* Punkte Darstellung: **Punktewolke**

**Kovarianz**

* Gewünschte Eigenschaften eines Zusammenhangskoeffizenten → sollte Stärke des Zusammenhangs numerisch ausdrücken; sollte die Richtung des Zusammenhangs ausdrücken; sollte invariant unter zulässigen Transformationen sein (z.B. m in cm); sollte einfach interpretierbar sein
* Für n Beobachtungen aus zwei Variablen ×₁, ..., ×n und y1, ..., yn ist die Kovarianz definiert als:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* positiver Zusammenhang: positive Differenz bei x mit positiver Differenz bei y **oder** negative Differenz bei × mit negativer Differenz bei y **→ immer positive Zahl bei Multiplikation**
* negativer Zusammenhang: positive Differenz bei x mit negativer Differenz bei y oder negative Differenz bei × mit positiver Differenz bei y → **immer negative Zahl bei Multiplikation**
* Die Kovarianz ist Null, wenn **kein Zusammenhang** zwischen den Realisationen der Variablen besteht
* Die Kovarianz ist positiv, wenn ein **gleichsinniger Zusammenhang** besteht
* Die Kovarianz ist negativ, wenn ein **gegensinniger Zusammenhang** besteht
* Problem: Die Kovarianz **erfüllt nicht die Forderungen der Invarianz** gegenüber erlaubten Transformationen
* Addition einer Konstante zu × und y (Form der Wolke ändert sich nicht, nur Lage):



* **Aber**: Multiplikation von x und y mit einer Konstanten (Kovarianz verändert sich äquivalent):



→ Die Äquivarianz macht die Kovarianz numerisch schwer zu interpretieren

**Korrelation (linearer Zusammengang, „standardisiert“)**

* Für n Beobachtungen aus einem Zufallsexperiment x1, ..., ×n und y1, ..., yn ist der Korrelationskoeffizient definiert als:

Ein Bild, das Text, Whiteboard enthält.

Automatisch generierte Beschreibung

* keine z-Standardisierung notwendig, wenn wir Standardabweichungen der ursprünglichen Daten kennen
* Für die **Richtungsinformation** (Vorzeichen) gelten dieselben Regeln wie bei der Kovarianz
* Bei der Korrelation ist zudem die **Stärke** (der Betrag) des Zusammenhangs interpretier- und vergleichbar
* Der so definierte Korrelationskoeffizient r wird auch als **Produkt-Moment-Korrelation** oder **Korrelationskoeffizient nach Pearson** bezeichnet
* Für Daten unterhalb Intervallskalenniveau gibt es andere Berechnungsformeln für die Korrelation (weil es auf Differenzen beruht!)
* Die Korrelation ist Null, wenn kein Zusammenhang der Ausprägungen der Zufallsvariablen besteht
* Die Korrelation liegt immer zwischen -1 und 1 (1 nur bei "perfekter" Geraden, ansonsten weniger)
* Addition und Multiplikation verändern Korrelationskoeffizienten nicht
* Negative Werte zeigen einen gegensinnigen Zusammenhang an; positive Werte zeigen einen gleichsinnigen Zusammenhang an
* Die Korrelation ist anfällig gegenüber Ausreißern

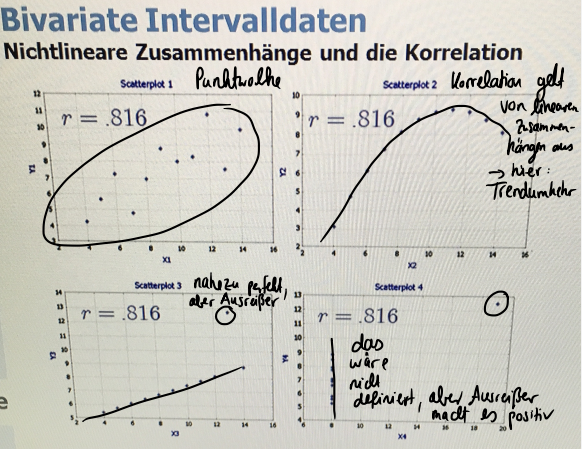
**Vergleich Kovarianz und Korrelation**

* bei beiden: Reihenfolge der Betrachtung der Daten egal

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

**Nichtlineare Zusammenhänge und die Korrelation**



→ Vorsichtig, wenn nur Korrelationskoeffizient; immer noch in Scatterplot schauen

* **Korrelation: Numerische Beschreibung – Faustregeln**
* Für die Bewertung der absoluten Höhe der Produkt-Moment-Korrelation existieren Faustregeln nach Cohen (1988):

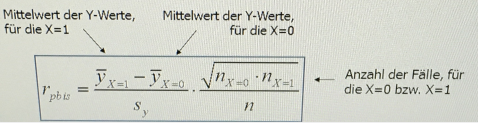
**r < ± 0.10 → keine Korrelation**

**r < ± 0.30 → kleine Korrelation**

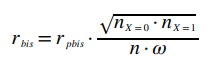
**r < ± 0.50 → mittlere Korrelation**

**r ≥ ± 0.50 → hohe Korrelation**

* In Experimentellen Forschung hohe Korrelation mindestens erwartet für aussagekräftiges Ergebnis
* Vorsichtig bei zu kleinen Stichprobengrößen (Zufallskorrelationen
* **Punktbiseriale Korrelation**
* Gegeben seien zwei Variablen X und Y. X sei dichotom nominalskaliert (mit zwei Ausprägungen 0 und 1), Y intervallskaliert; Datentabelle nicht mit x und y, sondern 0 und 1 als zwei Gruppen, wo jeweils die y-Werte eingetragen werden
* Hier kann, wie auch bei zwei intervallskalierten Variablen die Produkt-Moment-Korrelation berechnet werden
* Die Formel lässt sich aber auch zur Formel für die punktbiseriale Korrelation vereinfachen



* **Biseriale Korrelation**
* Häufig werden in empirischen Untersuchungen eigentlich (mindestens) intervallskalierte Merkmale **künstlich auf dichotome Variablen** **reduziert**
* Bsp.: Alter (unter 25, über 25), Einkommen (niedrig, hoch), Depression (ja, nein)
* Hier führt die konkrete **Setzung des impliziten Kriteriums**, welches die intervallskalierte Variable in zwei Gruppen teilt, zu beliebigen Ergebnissen, obwohl der "wahre" Zusammenhang unverändert ist → Dichotome Variable kann nur "Zwischenvariable" sein (Zusammenhang unverändert)
* Korrektur der Punktbiserialen Korrelation
* Die Korrektur dieser kriteriumsabhängigen Veränderung des Zusammenhangs leistet die **biseriale Korrelation**:

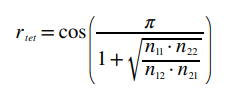


* Dabei ist Omega die Ordinate der Standardnormalverteilung für den z-Wert an der Stelle der Dichotomisierung (p)
* rpbis und rbis Korrelation haben dieselben Eigenschaften wie der Produkt-Moment- Korrelationskoeffizient
* rpbis ist zumeist vorzuziehen, da hier keine Normalverteilungsannahme gemacht werden muss
* **Tetrachorische Korrelation**
* doppelte Biseriale Korrelation
* Sind beide Variablen künstlich dichotomisiert und eigentlich normalverteilt, so kann der Zusammenhang durch die tetrachonische Korrelation ausgedrückt werden
* Ausgegangen wird zunächst von einer 2×2 Kontigenztabelle

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

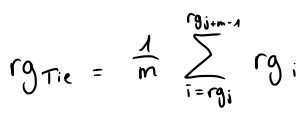
* Daraus berechnet sich die tetrachorische Korrelation als:



* rtet überschätzt die wahre Korrelation, wenn die Randverteilungen stark asymmetrisch sind oder eine der Verbundhäufigheiten nxy < 5 ist

**Bivariate Ordinaldaten**

* **Grundlagen**
* Bei der Ordinalskala ist der numerische Abstand zwischen zwei Realisationen einer Variablen nicht interpretierbar
* Die Ordinalskala trägt lediglich Information über die **Ordnung** der Realisationen
* Damit sind mathematische Transformationen direkt auf den Werten einer ordinalskalierte Variablen nicht sinnvoll, also auch nicht die Produkt-Moment-Korrelation
* **Ansatz**: Die Ordnung selbst muss genutzt werden, um Kennwerte zu berechnen
* **Rangbildung**
* Bei der Rangbildung von n Messungen x1, ...,×n einer Variablen X können maximal n Ränge vergeben werden
* Per Konvention erhält der kleinste Messwert von X den Rangplatz 1, der höchste Messwert den Rangplatz n (**kleinere Zahl = kleinerer Rang**)
* Bei mehreren gleichen Werten („**Ties**") von X wird der mittlere Rangplatz vergeben nach der Regel: Es gäbe m gleiche Werte von X. Wären sie unterschiedlich und direkt aufeinander folgend, erhielten sie die Rangplätze rgj, ..., rgj+m-1
* Der **mittlere Rang** ist dann:



* **Rangkorrelation - Spearman's r**
* Nach der Rangbildung ordinalskalierter Daten für zwei Variablen X und Y kann die Produkt-Moment-Korrelation der Ränge rg(X) und rg(Y) berechnet werden
* neue ranggebildete Variable als intervallskaliert
* Diese wird **Spearman's r** oder **Rangkorrelation** genannt und berechnet als:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

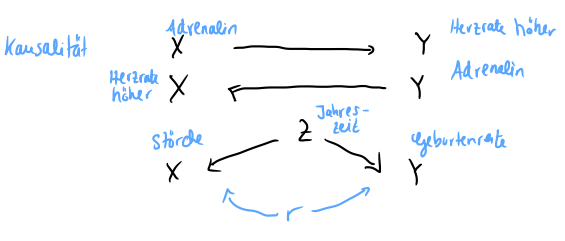
Oben: Kovarianz; Unten Standardabweichung (jeweils Wurzel)

* Vereinfacht/Umgeschrieben:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Wertebereich von -1 bis +1
* Vorzeichen gibt die Richtung des Zusammenhangs an
* Ist robust bezüglich Ausreißern → Ausreißer bekommen "nur" höchsten Rang
* Ist invariant bei streng monotonen Transformationen
* Liegen wenige Ties vor, gibt es vereinfachte näherungsweise Berechnungsformeln, die aber kaum mehr Anwendung finden (PC schneller)
* **Rangkorrelation - Weitere Kennwerte**
* Neben Spearman's r existieren weitere Kennwerte für den Zusammenhang zweier ordinalskalierter Merkmale
* Die bekanntesten sind der **Konkordanzkoeffizient γ (gamma)** nach Goodman-Kruskal und die daraus abgeleitete Weiterentwicklung **Kendall's τ (tau)** für zwei ordinalskalierte Variablen (wird kaum noch angewendet)
* Die Interpretation diese Koeffizienten verläuft analog zu r und rs
* **Interpretation von Korrelationen**
* Eine vorhandene (hohe) Korrelation zwischen zwei Variablen X und Y darf nicht ohne weiteres als Kausalität zwischen den Variablen interpretiert werden (z.B. Störche und Geburtenrate)
* Eine vorhandene Korrelation zeigt zunächst nur eine Assoziation an. Diese kann viele Ursachen haben, z.B.:



→ **Korrelation ist nicht Kausalität, sondern Assoziation**

* Frage: Wann darf in der empirischen Forschung von einer Korrelation auf Kausalität geschlossen werden?

1. Die betrachteten Variablen müssen **kovariieren** → die Korrelation muss ungleich Null sein → Probleme: Standards (wann ist eine Korrelation ungleich Null) sind normativ; je kleiner n, desto größere Korrelationen können per Zufall auftreten
2. Die Ursache muss der Wirkung **zeitlich vorausgehen** (z.B. Pretest- Treatment-Posttest, 2-mal Y und 1-Mal X (Treatment) erheben)
3. Andere plausible Erklärungen für die Kovariation müssen ausgeschlossen werden können (Schmerz beim Spritzen Adrenalin)
4. Die Kovariation muss raum-zeitlich indifferent sein → Generalisierung auf eine Population zu jeder Zeit

**8. Bivariate Zusammenhangsmaße I**

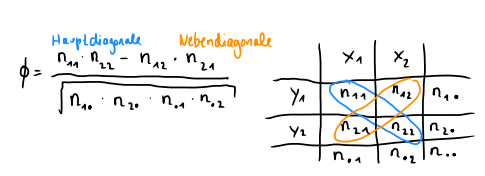
**Bivariate Nominaldaten**

* **Recap: Kontigenztabellen**
* Kontigenztabellen empirischer Verbundhäufigkeiten
* schreibt man statt h(xi, yj) kurz nij, so lautet die vereinfachte Notation:

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

* Analoge Notation für relative Häufigkeiten (mit fij statt nij)
* **Unabhängigkeit**: Ausprägung Variablen X sagt nichts über Ausprägung Variablen Y aus (und andersherum); Randhäufigkeit (und deren Wahrscheinlichkeit) als gleichmäßige Aufteilung
* **Abhängigkeit**: Wissen über Ausprägung Variablen X sagt eine ganze Menge über Ausprägung Variablen Y aus (und andersherum); Randhäufigkeiten ignorieren und dafür auf Verbundhäufigkeiten schauen
* **Zusammenhangsmaße für 2×2 Kontigenztabellen**
* viele empirische Fragestellungen über Zusammenhänge von Variablen beziehen sich auf 2 Merkmale mit je 2 Ausprägungen
* Bsp.: Auftreten eines bestimmten Genmarkers bei Frauen/Männern (x(0,1) und y(0,1))
* natürliche dichotome Variablen (tetrachorische Korrelation nur für künstlich hergestellte dichotome Variablen)
* aus Kontigenztabelle 2×2 zur Tabelle mit 2 Ausprägungen (Schuldspruch, Geständnis), in die die Wertepaare eingetragen werden
* Weil hier eine Intervallskala erzwungen wird (genau ein Abstand zwischen Skalenwerten = konstanter Abstand zwischen Skalenwerten), kann immer die **Produkt-Moment-Korrelation r** als Zusammenhangsmaß berechnet werden
* von Kontigenztabelle zur neuen Tabelle sehr mühselig; Idee: Der 2×2 Fall bei Nominaldaten kann immer auf den ja/nein bzw. 0/1 Fall zurückgeführt werden **→ Umformung r zum Phi-Koeffizient**
* der **Phi-Koeffizient** beschreibt die Stärke des Zusammenhangs zweier dichotomer Variablen (in der Praxis auch für nicht natürliche Variablen); kann mit relativen und absoluten Häufigkeiten berechnet werden



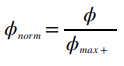
* Phi liegt zwischen -1 und 1 (abhängig von Anordnung der Realisationen (0 → 1 oder 1 → 0)
* Problem: Bei schiefen Randverteilungen kann der Phi-Koeffizient selbst bei maximalem Zusammenhang zwischen den Variablen die Grenze +/- 1 **nicht erreichen**
* Bei schiefen Randverteilungen sollte Phi daher an der maximal möglichen Korrelation normiert werden

→ Die **maximale positive / negative Korrelation** berechnet sich als:

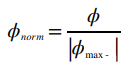
Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Ist also der tatsächliche Phi-Koeffizient **positiv**, kann/soll er an der maximal möglichen positiven Korrelation **normiert** werden über (Phi oben kann r sein):



* Ist er **negativ**, berechnet sich die Normierung ganz analog aus:

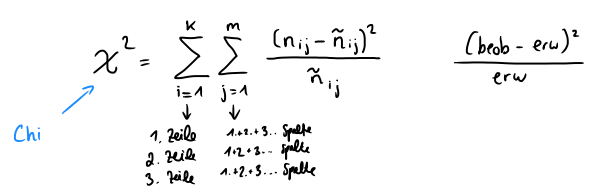


* **Phi** bei sehr schiefen Randverteilungen niedriger → eignet sich nicht so
* **Phi (Norm)** bei schiefen Randverteilungen größer → korrigiert
* Phi und Phi(Norm) sollen sehr ähnlich sein, ansonsten sehr schiefe Randverteilungen (z.B. Schuhe-Geschlecht; 30 Männer und 300 Frauen → schiefe Verteilung)
* **Zusammenhangsmaße für k ✗ m Kontigenztabelle**
* Ansatz: Man vergleicht die beobachtete Kontigenztabelle (absolute Häufigkeiten) mit einer fiktiven Kontigenztabelle, die entstanden wäre, hätte **kein Zusammenhang** (Unabhängigkeit) bestanden
* Abweichungen der beobachteten von den erwarteten Häufigkeiten sind dann als Abweichungen von der Unabhängigkeit aufzufassen
* Zur Konstruktion der Indifferenztabelle (erwartete Unabhängigkeit) rechnet man für absolute Häufigkeiten aus n Beobachtungen:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Aus den beobachteten und unter der Annahme keines Zusammenhangs (**Indifferenz**) erwarteten Häufigkeiten berechnet sich nun:



* Chi2 ist Null bei perfekte Unabhängigkeit, ansonsten größer Null
* Chi2 kann beliebig große Werte annehmen, abhängig von der Anzahl der Ausprägungen und der Beobachtungen (wenn Tabelle \*10, wird Chi2 auch \*10 größer) → nicht sinnvoll, weil nicht relativ zu Werten!
* **Nur berechnen, wenn alle erwarteten Häufigkeiten >5 sind!!**
* Um aus dem nicht normierten Chi2-Koeffizienten ein als Korrelationskoeffizient interpretierbares Maß zu berechnen, wird folgende Formel verwendet:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Cramérs V ist wie Chi2 Null bei perfekter Unabhängigkeit, ansonsten größer Null
* Cramérs V schwankt zwischen 0 und 1; ohne Richtungsinformation; desto größer Zahl, desto abhängiger
* Faustregeln nach Cohen anwendbar

**Kapitel Regressionsrechnung**

1. Grundlagen der Linearen Regression
2. Kennwerte der Regressionsrechnung I
3. Voraussetzungen der Regressionsrechnung
4. Polynomiale Regression
5. Partialkorrelation

**1. Grundlagen der Linearen Regression**

**Multiple Regression: Grundlagen**

* Betrachtung **mehrerer unabhängigen Variablen**, die eine **abhängige Variable** beeinflussen (oder vorhersagen sollen); Zusammenhang mehr als 2 Variablen
* **Beispiele**: Abhängigkeit Lebenszufriedenheit von sozialem, ökonomischem und Gesundheitsstatus
* Solche Fragestellungen werden auch als **multifaktoriell** bezeichnet
* **Drei Hauptfragestellungen der Regressionsrechnung:**

1. Gibt es eine statistische Beziehung zwischen mehreren Variablen, die die Vorhersage der AV aus der UV erlaubt?
2. Kann eine möglichst einfache mathematische Regel formuliert werden, die diesen Zusammenhang beschreibt?
3. Wie gut ist diese Regel im Hinblick auf die Vorhersage? (exakte Vorhersage)

* **Ziel**: Die unabhängigen Variablen sollen mithilfe einer mathematischen Regel so miteinander verrechnet werden, dass eine möglichst gute Vorhersage der abhängigen Variablen erreicht wird; Regel, wie man abhängige Variable aus den unabhängigen berechnen kann, um somit eine Aussage für die ganze Population zu entwickeln (Alter + Attraktivität → „Nett-Skala“)

**Grundgleichung**

* Die vorherzusagende Variable (AV, y-Wert) wird als **Kriterium** oder **Response** bezeichnet, die vorhersagenden Variablen (UVn, x-Werte) als **Prädiktoren** oder **erklärende** **Variablen**.
* Die Vorhersagegleichung der multiplen Regression mit k Prädiktoren wird geschrieben als:

 (y=a\*x+b)

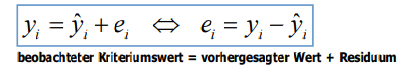
* y^ ist das vorhergesagte Kriterium, nur y ist der tatsächlich gemessene Wert; jedes x ist eine andere UVn
* Bei z-standardisierten Daten verwendet man das Symbol β für die k **Regressionsparameter (bzw. „-gewichte“) (oben b)** → hier fällt b0 weg!



* Gründe für die Annahme einer linearen Gleichung:
* Lineare Zusammenhänge sind **einfach zu verstehen**
* Lineare Zusammenhänge sind mathematisch und statistisch **einfach zu behandeln**
* Lineare Gleichungen haben sich in der Praxis vielfach als **gute Approximationen** für komplexe Beziehungen erwiesen
* **Achtung**: Auch wenn die Beziehung zwischen zwei Variablen linear „aussieht“, muss es sich nicht zwangsläufig um einen linearen Zusammenhang handeln (z.B. zu wenige Stichproben)

**Methode der kleinsten Quadrate (KQ-Kriterium)**

* Für eine Versuchsperson i aus allen n gelte:



* **Residuum ei**: Abweichung zwischen dem tatsächlichen beobachteten y-Wert und der im Rahmen der resrechnung vorhergesagten y-Wert
* Zur Minimierung des Vorhersagefehlers (jedes y-Wert) wird oft das **Kleinste-Quadrate Kriterium** verwendet (KQ; oder Ordinary Least Squares, OLS)
* Parameter der multiplen Regressionsgleichung werden so gewählt, dass das **Quadrat der Abweichungen** von gemessenem und geschätztem Wert **minimiert** wird
* Dann soll für alle n Datenwerte erreicht werden, dass die Summe der quadrierten Vorhersagefehler minimal wird:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Mithilfe der Allgemeinen Gleichung der einfachen linearen Regression lässt sich für die **Streuung des Vorhersagefehlers QSe** also schreiben:



bzw. in der standardisierten Form als Funkion wäre es f(b / β), nicht von x!



* Die Minimierung der Regressionsparameter erfolgt über partielle Differenzierung (Ableiten, um Minimum zu finden) nach jedem einzelnen der b- bzw. β-Gewichte

**Normalgleichungen der multiplen Regression**

* Die partielle Differenzierung der nichtstandardisierten Gleichung mit k Prädiktoren führt immer auf ein System von k+1 Normalgleichungen, das wie folgt aufgebaut ist

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* In der standardisierten Form ergibt sich ein System von k Normalgleichungen:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Hier kein b0, also kein k+1

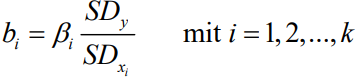
* Die Normalgleichungssysteme können nun durch Substitution oder Diagonalisierung gelöst werden

**Matrixalgebraische Berechnung**

* **Lösung über Lineare Algebra**
* Beide Verfahren sind eher mühselig. In der Praxis wird deshalb ein drittes, einfacheres Verfahren verwendet, das auf vektoralgebraischen Methoden basiert
* Mit wenigen Rechenschritten gelangt man hier zu den β-Gewichten
* Benötigt werden dazu lediglich:

1. Alle **Prädiktorinterkorrelationen**: die Korrelationen zwischen den Prädiktoren x1…k
2. Alle **Kriteriumskorrelationen**: die Korrelationen zwischen den Prädiktoren x1…k und dem Kriterium y

* **Umrechnung zwischen b und β**
* Zwischen β-Parametern für die z-standardisierten Daten und b-Parametern für die unstandardisierten Daten existiert ein einfacher Zusammenhang, der beliebige Hin- und Rücktransformationen zulässt:



* Die Konstante b0 wird dann berechnet als



da die Regressionsgerade immer durch den Punkt der Mittelwerte von Prädiktoren und Kriterium geht

* Bei Standardisierung ist b0 = 0, weil bei z-standardisierten Daten der Mittelwert immer 0 ist
* **Spezialfall: Nur 1 Prädiktor (Einfache Lineare Regression)**
* Bei nur einem Prädiktor vereinfacht sich die Berechnung der Regressionsgewichte erheblich



1. Es gilt dann:

 **r =Korrelation**

1. Steigung:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

1. y-Achsenabschnitt:

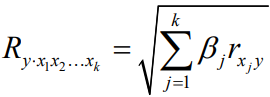


* **Es gilt: R2 = rxy2**
* **Interpretation der b und β**
* Die Größe eines b-Gewichtes gibt an, um wie viele Einheiten sich der Wert des unstandardisierten Kriteriums verändert, wenn der Betrag des unstandardisierten Prädiktors um 1 steigt
* Die Größe des β-Gewichtes gibt dasselbe für die standardisierten Variablen an (Prädiktor mit höherem β wichtiger/bedeutsamer)
* Bsp.: Das b-Gewicht beantwortet die Frage: „Ich möchte einen der Prädiktoren um 1 erhöhen. Welchen sollte ich wählen, damit das Kriterium maximal steigt?“
* Bsp.: Das β-Gewicht beantwortet die Frage: „Mit welchem Prädiktor erhöhe ich das Kriterium am effizientesten?“
* Das b-Gewicht liefert also eine **absolute**, das β-Gewicht eine **relative** **Information**
* **Vorsicht bei der Interpretation der Regressionsgleichung**
* Bei der **Korrelationsrechnung** bedeutet ein Zusammenhang niemals Kausalität, **lediglich Assoziation**
* Bei der **Regressionsrechnung** gilt zunächst dasselbe
* Die Kausalitätsvermutung wird (wenn überhaupt) schon bei der Aufstellung der Regressionsgleichung getroffen, nicht erst bei der Interpretation der Ergebnisse
* Um tatsächlich Kausalität festzustellen, müssen weitere Randbedingungen vorliegen (i.e. zeitliche Antezedenz von Ursache vor Wirkung, Generalisierbarkeit etc.)

**2. Kennwerte der Regressionsrechnung**

**Der multiple Korrelationskoeffizient R**

* **Def**.: Der **multiple Korrelationskoeffizient R** repräsentiert die Korrelation zwischen dem Kriterium y und **allen** Prädiktoren x1…xk bzw. Korrelation zwischen dem beobachteten Kriteriumswerten und den vorhergesagten Kriteriumswerten.
* Dabei berücksichtigt R etwaige Interkorrelationen zwischen den Prädiktoren (und entfernt sie)
* Der multiple Korrelationskoeffizient R ist definiert als:

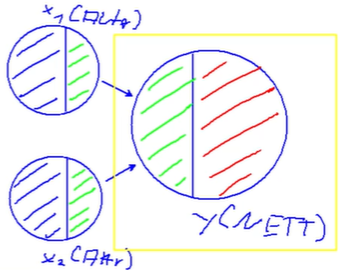


* Er ist mathematisch äquivalent zur Korrelation zwischen den gemessenen y-Werten und den vorhergesagten ydach-Werten, also:



**Der multiple Determinationskoeffizient R²**

* Definition: Der **multiple Determinationskoeffizient R²** repräsentiert die **Varianzaufklärung (Aufspaltung Abweichungen)**, die **alle** Prädiktoren x1…xk am Kriterium y leisten



* **Erklärung**: rot nicht erklärbar; grün ist Zusammenhang; blau ist egal; Streuung des Kriteriums wichtig (Varianz wird benutzt, weil sie nicht von der Stichprobengröße abhängig ist)
* **Gesamtvarianz = Aufgeklärte Varianz + Nicht aufgeklärte Varianz**
* Gesamtvarianz: Varianz aus y-Werten; Aufgeklärte Varianz: Varianz aus ydach-Werten; Nicht aufgeklärte Varianz: Varianz aus e-Werten
* Der multiple Determinationskoeffizient R² ist definiert als (wird öfter angewendet):

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Rechnerisch:

Ein Bild, das Text, orange enthält.

Automatisch generierte Beschreibung

**R und R²**

* R und R² sind tatsächlich direkt ineinander transformierbar

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* Für die Bewertung des R können wieder die Daumenregeln nach Cohen (1988) verwendet werden:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Dies führt aber auf ein Problem bei der Bewertung des R², denn die Quadratur der Daumenregeln liefert

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* In der Praxis bedeuten 25% aufgeklärte Varianz, dass 75% der Streuung in der AV nicht durch die Regressionsgleichung, d.h. die Prädiktoren erklärt wird
* Daher haalternative **Daumenregeln für die Bewertung des R²** vorgeschlagen (beruhend auf seiner Definition der Effektstärke)

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung**NICHT NEGATIV!!**

* Diese Regeln sind recht streng, insbesondere in der Feldforschung, wo 20-30% Varianzaufklärung bereits als gutes Ergebnis gewertet werden

**Abhängigkeit (Beziehung Prädiktoren)**

* Sind die Prädiktoren **unabhängig**, so sind die ß-Gewichte gleich den Kriteriumskorrelationen und die aufgeklärte Varianz ist die Summe der Quadrate der ß-Gewichte
* Erklärung: Bei perfekt unabhängigen Prädiktoren ist die Prädiktorinterkorrelationsmatrix Rxx gleich der Identitätsmatrix I.



* Damit gilt für den multiplen Korrelationskoeffizienten R und R² ist einfach die Summe der quadrierten Kriteriumskorrelationen:

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* In der Praxis nicht vorhanden
* Sind die Prädiktoren **abhängig** (interkorreliert), so sind 3 Fälle zu unterscheiden:

1. Der Prädiktor klärt zumindest Teile der Varianz am Kriterium auf, die andere Prädiktoren nicht aufklären; zusätzliche Informationen: er ist **nützlich**.
2. Der Prädiktor enthält (nur) Information, die auch andere Prädiktoren enthalten: er ist **redundant**.
3. Der Prädiktor unterdrückt irrelevante Varianz in anderen Prädiktoren: er ist ein **Suppressor**.

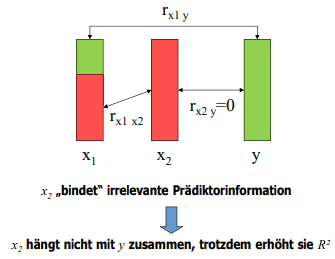
* **Nützlichkeit:**
* = Der Beitrag, den eine Variable zur Varianzaufklärung des Kriteriums leistet, der von den anderen Variablen nicht geleistet wird; Differenz mit und ohne Prädiktor, der uns interessiert
* Nützlichkeit immer relativ zu anderen Prädiktoren
* Die Nützlichkeit (utility) einer Variablen xj berechnet sich als:



* Uj ist also der Betrag, um den R² wächst, wenn die Variable xj in die multiple Regressionsgleichung aufgenommen wird.
* **Redundanz:**
* = die vielen Variablen messen Aspekte gemeinsam, sodass man prinzipiell weniger Prädiktoren benötigt → **unerwünschter Aspekt**
* Die Variable xj ist redundant zur Vorhersage von Variable y, wenn gilt



* Prädiktoren enthalten empirisch nahezu immer gemeinsame Varianzanteile und sind somit „teilweise redundant“. Echte Redundanz liegt erst gemäß obiger Definition vor.
* **Multikollinearität**: Die Kovarianz eines Prädiktors mit dem Kriterium ist in den anderen Prädiktoren (fast) vollständig enthalten → extremer Fall von Redundanz, der **unbedingt zu vermeiden** ist.
* **Suppression:**



* Definition: Eine Variable xj ist ein Suppressor, wenn gilt:



* Die Zunahme der erklärten Varianz durch Aufnahme der Variable ist also größer als die einzelne Varianzaufklärung.
* **Anwendung von Nützlichkeit, Redundanz, Suppression:**
* Nützlichkeit, Redundanz und Suppression werden verwendet, um durch die **Selektion geeigneter Prädiktoren** eine möglichst **sparsame Regressionsgleichung** zu generieren
* Die Suppression hilft oft dabei, theoretisch unerwartete Prädiktoren zu identifizieren und **beizubehalten**
* Nützlichkeit und Redundanz haben eher das Ziel, ungeeignete Prädiktoren **auszuschließen** (Schritt für Schritt, weil sich alles ändert, wenn wir einen schlechten Prädiktor rausnehmen)
* Dazu gibt es statistische Verfahren, die die Frage beantworten, wann eine Nützlichkeit **hoch genug** ist
* Es gibt weitere Verfahren zur Prädiktorselektion, z. B. den statistischen **Test der β-Gewichte**

**3. Voraussetzungen der Regressionsrechnung**

**Voraussetzung**

* **Mathematische Voraussetzung:**
* praktisch **immer zu rechnen**, da nur in Ausnahmefällen die Invertierung der Prädiktorinterkorrelationsmatrix fehlschlägt
* **Statistische Voraussetzungen:**
* Eine Reihe von Voraussetzungen, damit Kennwerte und inferenzstatistische Verfahren (z.B. der statistische Test der β–Gewichte) anwendbar sind und die Regressionsgleichung empirische Aussagekraft besitzt (sagen R und R2 überhaupt etwas über die Güte aus)
* **1. Skalenniveaus:**
* Die **Prädiktoren** können entweder intervallskaliert oder dichotom sein
* Das **Kriterium** **muss** intervallskaliert sein und die Skala soll unbeschränkt sein (keine untere und obere Schranke; Minimum/Maximum → **Ungebundenheit**) → es würden sonst Werte vorhergesagt werden können, die es gar nicht geben kann (Kriterium y = Schulnote 1-6)
* Manchmal wird doch Kriterium angenommen (z.B. IQ), welches ein Minimum hat → weit genug weg von Grenzen, dass wir sie mit Prädiktorenkonstellation nicht erreichen dürften (IQ=0)
* Für andere Skalenniveaus des Kriteriums existieren verschiedene Regressionsvarianten (sehr speziell):

- **Logistische** **Regression** für dichotome Kriteriumsvariablen (gesund vs. krank)

- **Multinomiale** **Regression** für nominalskalierte Kriterien (mehrere mögliche Ausprägungen, die aber nicht geordnet werden können)

- **Ordinale Regression** für ordinalskalierte Kriterien

* **2. Eigenschaften der Prädiktoren:**
* Keine zu hohen Interkorrelationen zwischen den Prädiktoren, i.e. **Vermeidung von** **Multikollinearität**
* Es sollen alle wesentlichen Einflussvariablen (durch Prädiktoren) des Kriteriums erfasst werden, d.h. **hinreichend hohes R²**
* Zusammenhang zwischen den Prädiktoren und dem des Kriteriums soll dem Modell der Regressionsgleichung entsprechen (linear, polynomisch etc.) → Zusammenhang zwischen jedem einzelnen Prädiktor und jeweils dem Kriterium sollte auch linear sein, wenn wir die (einfache/multiple) lineare Regression anwenden wollen
* **hinreichend hohe Stichprobengröße**, Daumenregeln empfehlen hier zwischen 10 und 25 Personen pro Prädiktor
* **3. Eigenschaften der Fehler bzw. Residuen:**
* **Erinnerung**: Der Vorhersagefehler in der Regression wird auch als **Residuum** bezeichnet für **jeden** Merkmalsträgers i aus allen n:



* Die Residuen dürfen **nicht untereinander korreliert** sein, d.h. die Höhe des Vorhersagefehlers für Merkmalsträger 1 darf nicht den Fehler für Merkmalsträger 2 beeinflussen (z.B. Waage die durch zu schweres Gewicht danach (bei nächster Person) mehr anzeigt)
* Die Residuen sollen **normalverteilt** sein
* Für die Residuen („Fehler“) soll der **erwartete Mittelwert 0** sein (aufgrund Normalverteilung gleichen sich negative und positive Residuen aus; kann aber auch nicht so sein!)
* Die Residuen sollen dem Gebot der **Homoskedastizität** genügen, d.h. ihre Varianz soll unabhängig vom Kriteriumswert sein
* **Prüfung der Voraussetzungen:**
* Für die meisten der Fehlereigenschaften gibt es **statistische Tests** zur Voraussetzungsprüfung (meist bezogen auf Residuen):

- Variance Inflation Factor (VIF) für Multikollinearität (ob zu hohe Prädiktorinterkorrelationen)

- Durbin-Watson Test für Unkorreliertheit der Residuen

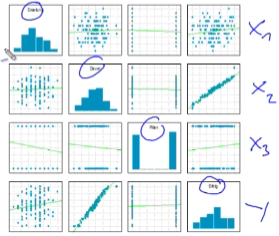
- Levene-Test für Homoskedastizität

- Kolmogoroff-Smirnov Test für Normalverteilung

* Zur ersten optischen Prüfung eignen sich spezielle Darstellungsvarianten, z.B. der **Fit-Plot des Kriteriums** sowie der **Residualplot**

**Darstellungen**

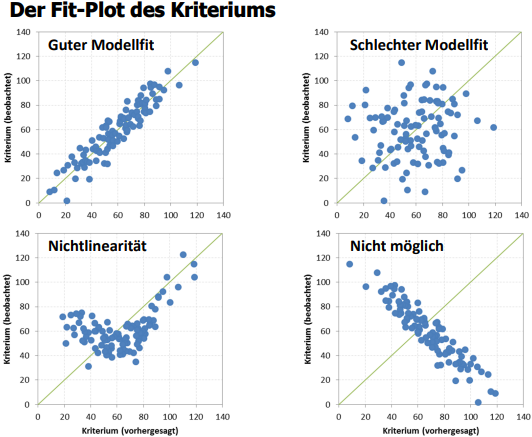
* **Die Matrixdarstellung der Scatterplots:**
* Der **Matrixplot** ist eine platzsparende Möglichkeit, die paarweisen Zusammenhänge zwischen allen an der Regression beteiligten Variablen (Prädiktoren + Kriterium) darzustellen
* Es handelt sich einfach um eine geordnete Darstellung aller möglichen Scatterplots für die Daten
* Der Matrixplot erlaubt eine schnelle Beurteilung von **Form und Stärke** der Zusammenhänge innerhalb der Prädiktoren sowie zwischen den Prädiktoren und dem Kriterium



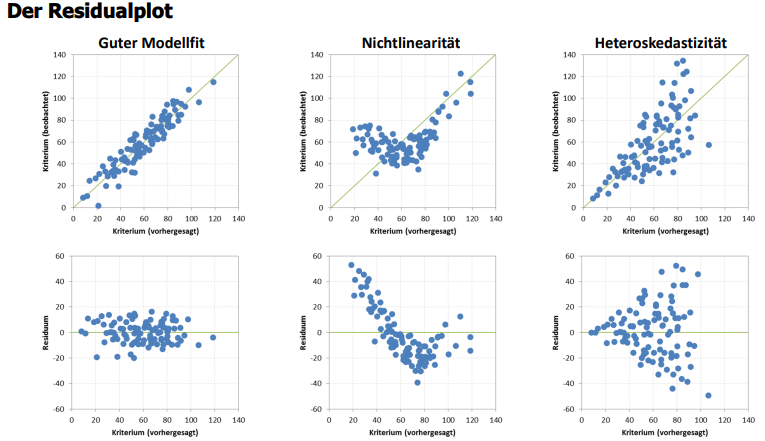


* x1: Erwartung, x2: Dauer, x3: Pillen, y: Erfolg Therapie
* Hauptdiagonale: Histogramme (Matrixplot kann auch ohne davon dargestellt werden)
* Nebendreiecke: Paarweise Darstellung von Zusammenhang aller 4 Variablen; jede Zeile und Spalte jeweils für x1, x2, … (Scatterplot gedreht) → waagerecht: immer x1 auf y-Achse, senkrecht: immer auf x-Achse
* Grüne Linie: einfache lineare Regression der zwei Variablen (muss nicht dabei sein)
* Es kann sein, dass eine Hälfte (also ein Nebendreieck + Histogramme) fehlt und nur die 6 Scatterplots übrigbleiben
* Soll vor der Berechnung der Multiplen Regression gemacht werden!
* **Der Fit-Plot des Kriteriums:**
* Um Zusammenhänge in Koordinatensystem darzustellen, brauchen wir eine neue Darstellung, weil man sonst in 3D, 4D … gehen müsste, um alle Prädiktoren darzustellen
* Der **Fit-Plot** des Kriteriums ist ein optisches Verfahren zur Prüfung der Übereinstimmung zwischen vorhergesagten und beobachteten Kriteriumswerten; dafür muss die Multiple Regressionsgleichung und für jeden Merkmalsträger seinen vorhergesagten Kriteriumswert berechnet worden sein
* Er stellt die vorhergesagten Kriteriumswerte (x-Achse) und die beobachteten (y-Achse) gegenüber
* Er ist eine **direkte** **Darstellung** des Multiplen Korrelations- bzw. Determinationskoeffizienten
* Bei perfekter Passung der Regressionsgleichung gilt (die Wertpaare für alle i Merkmalsträger liegen auf der Winkelhalbierenden):





* **Der Residualplot:**
* Der **Residualplot** ist ein optisches Verfahren zur Prüfung der Voraussetzungen an unsere Residuen; Verbesserung Fit-Plot
* Er stellt die **vorhergesagten** Kriteriumswerte (x-Achse) und die **Regressionsresiduen (nicht beobachtete Kriteriumswerte!)** (y-Achse) gegenüber
* An ihm kann man **Homoskedastizität** (verletzte Homoskedastizität heißt **Heteroskedastizität** →**nicht gut**) und **Modellpassung** (links und Mitte!) optisch gut überprüfen



* Bei jedem oberen Fit-Plot: senkrechte Abstände jedes Punktes zur Winkelhalbierenden schauen → Differenz beobachteten und vorhergesagten Kriteriumswert ist Regressionsresiduum
* Oft werden Residuen (und auch die Kriteriumswerte) der **standardisierten Daten** dargestellt, was die Normalverteilung der Residuen (Voraussetzung der Regressionsrechnung) optisch gut einschätzbar macht → beim links unten: zwar guter Modellfit, aber keine Normalverteilung (nichts im sehr negativen oder sehr positiven Bereich)

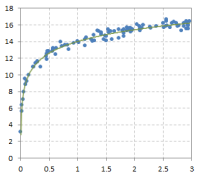
**4. Polynomiale Regression**

**Nichtlineare Regression**

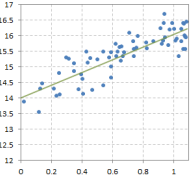
* **Grundlagen:**
* Bei einer Reihe psychologischer Fragestellungen ergeben sich nichtlineare Zusammenhänge zwischen UV & AV
* Bsp.: Reaktionszeit mit wachsender Ermüdung, Blutalkohol (Prädiktor) mit motorischer Präzision (Kriterium)
* Solche nichtlinearen Zusammenhänge lassen sich in **zwei Klassen** einteilen:

1. Zusammenhänge, die sich durch eine einfache (nichtlineare) Transformationen in lineare Zusammenhänge überführen lassen (siehe Beispiel)
2. Zusammenhänge, für die eine nichtlineare Regressionsgleichung gelöst werden muss

* **Beispiel: Logistische Regression:**
* Gemessene Daten verlaufen logarithmisch und variieren zwischen 0 und 1



* Umformung der x-Werte durch Logarithmieren bewirkt eine **Linearisierung** der Datenlage (hier stärker gestaucht, desto größer sie sind):



* Mithilfe dieser neuen x-Werte kann eine lineare Regression bestimmt werden, um die Parameter b0 und b1 zu errechnen; man darf nicht mehr über normale x-Werte reden, sondern nur noch logarithmierte x-Werte; b0 und b1 nicht mehr einfach so zu interpretieren
* **Linearisierbare und polynomiale Formen:**
* **Fall 1**: Linearisierende Transformation hilft nicht weiter (nicht weiter behandelt), z.B.



* **Fall 2**: Nicht (einfach) linearisierbar (Beschreibung durch Polynom):



**Polynomiale Regression**

* **Grundlagen und Durchführung:**
* Häufig können Merkmalszusammenhänge durch Polynome **2. oder 3. Ordnung** gut beschrieben werden, d.h.

 **oder** 

* Dies ist formal eine **lineare multiple Regression**, allerdings nicht mit mehreren Prädiktoren, sondern mit einem Prädiktor sowie Transformationen seiner selbst
* Eine solche **polynomiale Regression** wird berechnet, indem einfach die transformierten Prädiktorterme x², x³ usw. bestimmt werden
* Dann wird auf diesen eine übliche **lineare multiple Regression** durchgeführt
* Höher als der 3. Grad wird in der Praxis selten gegangen, weil sonst die Daten schwieriger interpretierbar werden
* Es können alle von **Kennwerten** **und** **Gütemaße** der multiplen Regression bestimmt werden (R und R2 wie bei multipler linearer Regression und gleiche Interpretation) → beschreiben nur Daten, erklären sie nicht
* Zur Reduzierung von Multikollinearität (Zusammenhänge Prädiktoren) kann der Prädiktor vor der Transformation auch in seinem Mittelwert zentriert werden (von jedem Prädiktor x den Mittelwert abziehen; dann ist sein Mittelwert 0 → Interkorrelation zwischen x und x2 wird kleiner)
* Die polynomiale Regression ist auch über die **KQ-Methode** (inkl. Normalgleichungen) herzuleiten. Dies führt auf dasselbe Ergebnis wie der hier verfolgte Ansatz

**→** **Die polynomiale Regression ist eine multiple lineare Regression (nur Schreibweise x2 bzw. x2 zeigt die Unterschiedlichkeit)**

**→ nur für betrachteten Bereich gültige Vorhersage!**

**5. Partialkorrelation**

**Partialkorrelation**

* Deutungsmöglichkeiten der bivariaten Korrelation:

1. Direkte Kausalität

Ein Bild, das Uhr enthält.

Automatisch generierte Beschreibung

1. Zufall
2. Drittvariable(n) → Wenn X eliminiert werden würde, wäre die Korrelation zwischen Y1 und Y2 r=0.

Ein Bild, das Text, Antenne enthält.

Automatisch generierte Beschreibung

1. Direkte + indirekte Kausalität → Es gibt eine Korrelation zwischen Y1 und Y2auch ohne X!

Ein Bild, das Antenne enthält.

Automatisch generierte Beschreibung

* **Definition**: Eine Partialkorrelation ist die Korrelation zweier Variablen, die vom Effekt einer dritten Variablen bereinigt wurden
* **Einsatzzweck**: Prüfung einer Kausalvermutung G

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* „Scheinkorrelation“ → Wenn die Korrelation zweier Variablen nach Herausrechnen von X r=0 ist
* Wenn-Diagramm → Kreisdarstellungen der Variablen; Kreisfläche beschreibt die Streuung
* **Berechnung und Prüfung:**

1. Sage y1 aus x voraus und berechne Residuen ey1
2. Sage y2 aus x voraus und berechne Residuen ey2
3. Berechne die Korrelation rey1ey2 → Schreibe: ry1y2 · x (Punkt heißt ohne)

Ein Bild, das Text, ClipArt enthält.

Automatisch generierte Beschreibung

* Ist Partialkorrelation **nahe Null**, so beruht die Korrelation ry1y2 tatsächlich vor allem auf der Einwirkung von x. (Prüfung mit Korrelationstest)
* **Vereinfachte Berechnung:**
* Für die **Varianz der Vorhersagefehler** gilt: 
* Die **Korrelation der Fehler** lässt sich schreiben als: Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung
* Man kann nun zeigen, dass gilt: 
* Damit errechnet sich die **Partialkorrelation** als:

Ein Bild, das Antenne, Uhr enthält.

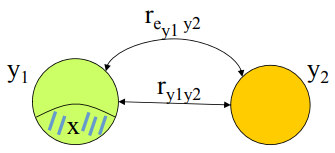
Automatisch generierte Beschreibung

→ Die Formel gibt dasselbe Ergebnis aus, wie bei der linearen Regression rauskam

**Semipartialkorrelation**

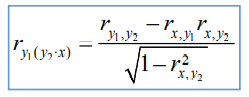
* Definition: Eine Semipartialkorrelation ist die Korrelation zweier Variablen, von denen eine vom Effekt **einer** anderen Variablen bereinigt wurden
* **Einsatzzweck**: Prüfung der zusätzlichen Information eines Prädiktors bei der Erklärung des Kriteriums
* Die Semipartialkorrelation ist eng verbunden mit der **Nützlichkeit**.

Es gilt nämlich Ux1 = r²y(x1 · x2)



* **Berechnung:**

1. Sage y2 aus x voraus und berechne Residuen ey2
2. Berechne die Korrelation ry1ey2 → Schreibe: ry1(y2 · x)
3. Oder verwende die vereinfachte Formel:



**(Semi-)Partialkorrelation höherer Ordnung**

* Soll der Zusammenhang zwischen zwei Variablen um **mehrere andere Variablen** bereinigt werden, spricht man von (Semi-)Partialkorrelationen höherer Ordnung
* Die Berechnung verläuft analog zu den (Semi-)Partialkorrelationen bei nur einer auszupartialisierenden Variable

Ein Bild, das Text, Uhr, ClipArt enthält.

Automatisch generierte Beschreibung

* **Berechnung über multiple Regression:**

1. Sage y1 aus den x1…xk voraus und berechne Residuen ey1
2. Sage y2 aus den x1…xk und berechne Residuen ey2
3. Berechne die Korrelation rey1ey2 → ry1y2 · x1…xk (Partialkorrelation)

oder

Berechne die Korrelation ry1 ey2 → ry1(y2 · x1…xk) (Semipartialkorrelation)

**Kapitel Varianzanalyse**

1. Grundlagen der Varianzanalyse für unabhängige Daten
2. Durchführung der Varianzanalyse für unabhängige Daten
3. Einzelvergleiche bei unabhängigen Daten
4. Die Varianzanalyse für Messwiederholungsdaten
5. Weitere Typen von Varianzanalysen

**1. Grundlagen der Varianzanalyse für unabhängige Daten**

**Ein-/Mehrfaktorielle ANOVA (ANalysis Of VAriance)**

* **Ziel**: Analyse des Einflusses unabhängiger Variablen (UVn) auf eine abhängige Variable (AV)
* **Beispiele**: Wie verändert sich die Schulleistung von Kindern über mehrere Nachhilfesitzungen hinweg?
* Generell versucht die ANOVA immer, Unterschiede **zwischen Gruppen** in der Höhe der AV aufzuzeigen
* Während bei der Regression der Fokus auf den **Zusammenhängen** zwischen UVn und AV liegt, soll die ANOVA gemessene **Unterschiede** in der AV erklären
* Die AV muss dabei **stetig** sein (intervallskaliert); die UVn sind i.d.R **nominal- oder ordinalskaliert**
* Die UVn werden im Rahmen der ANOVA auch als **Treatments** oder **Faktoren** bezeichnet, die Ausprägungen eines Treatments als **Treatmentstufen** oder **Faktorstufen**
* Nach der Anzahl der Treatments unterscheidet man die einfaktorielle, zweifaktorielle oder allgemein mehrfaktorielle ANOVA
* Beispiel:
* Experiment mit einem 3stufigen Treatment A
* Man habe an 3 Personengruppen (definiert durch die Treatmentstufen A1 bis A3) die Werte einer AV erhoben

Ein Bild, das Tisch enthält.

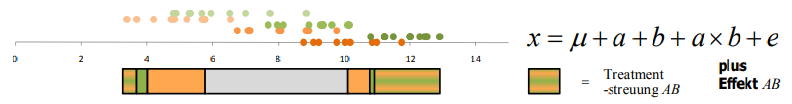
Automatisch generierte Beschreibung

* In der Praxis können die Gruppengrößen verschieden sein
* Nun nehme man das Geschlecht als „Treatment“ B mit zwei Stufen hinzu
* Man hat nun 6 Personengruppen (oder „Zellen“) Ein Bild, das Tisch enthält.

  Automatisch generierte Beschreibung
* Die abhängige Variable sei X genannt
* Mit der korrekten Indizierung lässt sich die Tabelle dann schreiben als: Ein Bild, das Tisch enthält.

  Automatisch generierte Beschreibung
* Benennung der Laufindizes: **Personenzahl in der Zelle; Zeile; Spalte**
* **Lineares Modell:**
* Die ANOVA geht davon aus, dass am Zustandekommen jedes Messwertes xi,j,k  (i = 1…nkj, j = 1…p, k = 1…q mit p=3, q=2) mehrere Komponenten beteiligt sind:

1. **Populationswert**, von allen Personen geteilt (Dieser ist für alle Personen konstant)
2. **Effekt der Treatmentstufe** j des Treatments A (Dieser ist für jede Person in Gruppe Aj konstant)
3. **Effekt der** **Treatmentstufe** k des Treatments B (Dieser ist für jede Person in Gruppe Bk konstant)
4. **Interaktionseffekt** jeder spezifischen Kombination von Aj und Bk (Dieser ist für jede Person in Zelle Aj Bk konstant)
5. **Zufallsfehler**, der Unterschiede zwischen Personen ausmacht, die nicht auf A oder B zurückgehen (Dieser ist individuell für jede Person)

* Wie die Regression geht die ANOVA von einem einfachen **linearen Modell** aus
* Sie nimmt an, dass der Messwert einer Person auf einer beliebigen AV additiv aus systematischen Komponenten und einer Fehlerkomponente besteht
* Wir betrachten hier nur den Fall **gleicher Zellhäufigkeiten**
* Die ANOVA zeigt die **Zerlegung von Datenstreuungen** in verschieden Effekte der Faktoren auf
* **Der Varianzanalyse liegt also formal ein sehr einfaches lineares Modell zugrunde:** Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung
* **Merke**: Der Messwert einer beliebigen aller N Personen setzt sich zusammen aus einem **Populationswert**, den **Treatmenteffekten** & einem **individuellen Messfehler**
* Der beobachtete Messwert jeder Person ist also in **drei Komponenten zerlegt**Ein Bild, das Tisch enthält.

  Automatisch generierte Beschreibung



* Die zentrale **Forschungsfrage**, die von der ANOVA beantwortet werden soll, lautet:Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung
* **In Worten**: Gibt es mindestens eine Treatmentstufe von A bzw. B, die auf die gemessene Variable der Versuchspersonen anders wirkt als die übrigen Treatmentstufen von A bzw. B?
* Die zentrale **Forschungsfrage**, die von der ANOVA beantwortet werden soll, lautet: Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung
* **In Worten**: Gibt es mindestens eine Interaktion von A und B, die auf die gemessene Variable der Versuchspersonen anders wirkt als die übrigen Kombinationen von A und B?
* **Streuungsvergleich:**
* Hat das Treatment **keinen Effekt**, so sind Unterschiede zwischen Mittelwerten der Treatmentstufen rein zufälligEin Bild, das Tisch enthält.

  Automatisch generierte Beschreibung
* Die Streuung der Stufenmittelwerte kann dann nur aus der Fehlerstreuung entstehen. Es muss gelten:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Ansatz**: Wenn man also feststellt, dass die Streuung der Stufenmittelwerte **deutlich größer** ist als die Fehlerstreuung, kann die Fehlerstreuung allein nicht mehr für die Streuung der Stufenmittelwerte verantwortlich sein
* Das Treatment hat dann einen **Effekt**, es gibt also **systematische Unterschiede** zwischen den Stufenmittelwerten
* **Die ANOVA leistet also das Auffinden von Unterschieden zwischen Gruppen-Mittelwerten, indem sie Streuungen von Fehlern und Treaments miteinander vergleicht.**
* **Problem**: Diese Logik von Fehler- und Treatmentstreuung ist nur dann hilfreich, wenn sie nicht nur für alle Treatmentstufen gemeinsam, sondern individuell für jedes Treatment bzw. die Interaktion gilt.
* **Denn**: Ziel der ANOVA ist das Treffen separater Aussagen über die Wirkung der einzelnen Treatments.
* **Ansatz**: Es ist zu zeigen, dass die Streuung der Stufenmittelwerte des Faktors A nicht auch von einem möglichen Effekt des Faktors B oder der Interaktion AB abhängt
* Diesen Beweis liefert die **Quadratsummenzerlegung** in der Varianzanalyse
* Exakt wie in der Multiplen Regression wird von der **Streuung aller Daten** um den Mittelwert ausgegangen
* Der **Mittelwert aller Daten (Grand Mean)** wird in der ANOVA oft als **G** oder **G quer** bezeichnet:

Ein Bild, das Text, Uhr, orange enthält.

Automatisch generierte Beschreibung

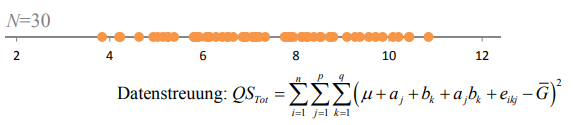
* Deren Streuung, ausgedrückt als Quadratsumme ist: Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung
* Man könnte statt der Quadratsumme auch die Varianz betrachten, hätte aber lediglich den zusätzlichen Faktor 1/N bzw. 1/n⋅p⋅q in voriger Gleichung (wir schauen uns aber nur die Quadratsumme an)
* Mit dem **linearen Modell der ANOVA** wird daraus:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

indem einfach der Messwert jeder Person xikj durch die Modellgleichung ersetzt wird

* Diese Quadratsumme repräsentiert die **gesamte Streuung in den Daten** um den gemeinsamen Mittelwert aller Daten
* Die Aufteilung der Streuung in Fehler- und Treatmentstreuung läuft nun darauf hinauszuverstehen, wie sich die Gesamtstreuung der Daten zerlegen lässt in: 
* **Problem**: aj und bk und ajbk und eijk und μ sind unbekannt
* Man kann aber aus den Daten andere, inhaltlich ähnliche Kennzahlen berechnenEin Bild, das Tisch enthält.

  Automatisch generierte Beschreibung
* So muss für den Mittelwert aller Personen in einer Faktorstufe des Faktors A bzw. B gelten:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Und natürlich auch für die Zellmittelwerte:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Für den Gesamtmittelwert aller Personen gilt zudem

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Diese Gleichungen kann man nach den bekannten Termen **umformen** und in die Gleichung der Quadratsumme **einsetzen**
* Nach weiteren Umformungen erhält man eine Gleichung, die zeigt, dass sich die gesamte Quadratsumme tatsächlich aufteilen lässt in **unabhängige** **Quadratsummen** für die Treatments und den Fehler
* **QS-Zerlegung:**
* Es zeigt sich nämlich, dass gilt: Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Gäbe es keine Interaktion, so sollte man annehmen, dass die beiden Faktoren additiv zusammenwirken, dann sollte gelten:



* Der Grand Mean muss einmal subtrahiert werden, da er in beiden Stufenmittelwerten enthalten ist
* Die Differenz dieser beiden Terme ist dann die **Interaktionswirkung** (d.h. der nicht-additive Teil des gemeinsamen Effektes von A und B)

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

**2. Durchführung der Varianzanalyse für unabhängige Daten**

**Ein-/Mehrfaktorielle ANOVA (ANalysis Of VAriance)**

* **QS-Zerlegung:**
* Zurückkehrend zur Ausgangsidee sollte man nun vermuten, dass **kein Effekt eines Treatments** dazu führt, dass die Streuung zwischen seinen Stufenmittelwerten ähnlich hoch ist wie die Fehlerstreuung, also:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Aber**: Es kann gezeigt werden, dass die Quadratsummen zuvor transformiert werden müssen, damit der Vergleich stimmt
* Diese transformierten Quadratsummen werden als **Populationsvarianzen** bezeichnet
* Die Berechnung der Populationsvarianzen erfordert statt des – bei der Varianzberechnung üblichen – Terms 1/n andere Nenner
* Statt des n werden die so genannten **Freiheitsgrade** (*df*, degrees of freedom) eingesetzt:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung **n: Anzahl in einzelnen Zellen**

**p: Anzahl Stufen 1. Faktor A q: Anzahl Stufen 2. Faktor B**

* Damit werden die Populationsvarianzen berechnet als

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung**s2: normale Varianz**

**σ2: Populationsvarianz**

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Prüfgröße & Ergebnistabebellen:**
* Mit den Populationsvarianzen stimmt die zuvor angenommene Beziehung!
* Ohne Effekt eines Treatments kann die Streuung der Stufenmittelwerte nur aus der Fehlerstreuung entstehen:

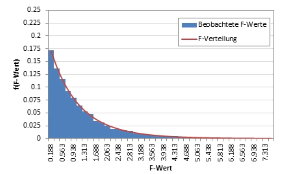
Ein Bild, das Text enthält.

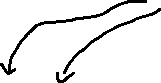
Automatisch generierte BeschreibungNiemals exakt 1 wegen Zufall!!

* Daraus konstruiert man die Prüfgröße:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung**F: Häufigkeitsverteilung (F-Verteilung)**





Diese zufällige Werte machen den Unterschied zwischen Fehler- und Treatmentstreuung



* Sie ist **F-verteilt** mit *df*Treat= p–1 Zählerfreiheitsgraden und *df*Fehler= p·(n-1) Nennerfreiheitsgraden
* Aus der F-Verteilung kann die **Wahrscheinlichkeit p(F)** für das Auftreten dieser Prüfgröße ermittelt werden
* **Wahrscheinlichkeitsverteilung und Bewertung der Prüfgröße F:**
* **Grundfrage**: Ist die Streuung der Mittelwerte zwischen Treatmentstufen hoch genug, damit statistisch behauptet werden kann, dass sie nicht mehr auf zufälligen Unterschieden aufgrund der Stichprobenziehung beruhen kann?
* Dazu muss bekannt sein, in welchem Bereich die Streuung der Mittelwerte zwischen Treatmentstufen **typischerweise** läge, wenn in Wahrheit keine Treatmentwirkung besteht.
* Die Statistik beantwortet dies „typischerweise“ mithilfe einer **Prüfgröße**.
* Aus Treatment- und Fehlerstreuung wird eine solche **Prüfgröße F** errechnet, deren zufällige Schwankungen bekannt sind, weil die der **F-Verteilung** folgen.
* Ein zu unwahrscheinlicher F-Wert belegt Unterschiede zwischen Treatmentstufen.
* Man berechnet also zunächst die **Prüfgröße F**
* Die F-Verteilung gibt nun an, welche **Wahrscheinlichkeit p(F)** das Auftreten der Prüfgröße hat unter der Annahme, dass Treatmentvarianz und Fehlervarianz gleich sind.
* Ist diese Wahrscheinlichkeit **zu klein**, so muss es Einflüsse auf die Treatmentstreuung geben, die nicht auf die Fehlerstreuung zurückzuführen sind.
* Diese Einflüsse können nur durch das Treatment hervorgerufen werden, es hat also eine Wirkung.
* Das bedeutet gleichzeitig, dass die Mittelwerte der Treatmentstufen systematisch unterschiedlich sind (dieser Unterschied erzeugt gerade die Treatmentstreuung).
* Ist die berechnete Wahrscheinlichkeit *p(F)* also zu klein, gibt es Unterschiede in den Stufenmittelwerten.
* **Problem**: Wie klein ist „zu unwahrscheinlich“?
* Es gibt hier Konventionen, die so genannten **Signifikanzniveaus:** Ein Bild, das Text enthält.

  Automatisch generierte Beschreibung
* **Zusammenfassung:**

1. Experimentelle Beobachtung (Messung Daten)
2. Datenzusammenfassung in Treatmentstufen/ ANOVA Zellen
3. Berechnung Quadratsummen
4. Berechnung Treatmentpopulationsvarianz mithilfe der Freiheitsgrade *df*



1. Vergleich VarianzenEin Bild, das Text enthält.

   Automatisch generierte BeschreibungEin Bild, das Text enthält.

   Automatisch generierte Beschreibung

Vergleich F-Wert

1. Man erstellt folgende Ergebnistabelle:

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

1. Zusätzlich ist zu empfehlen, die **Varianzaufklärung** (η2 = eta2) der Treatments zu bestimmen (= Anteil an der Gesamtstreuung, für den das Treatment verantwortlich ist):

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

1. Ebenso ist die nicht aufgeklärte Varianz zu bestimmen als der Anteil der verbleibenden Fehlerstreuung an der Gesamtstreuung, also:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

**Annahmen**

* **Voraussetzungen:**
* Die Beobachtungen müssen **unabhängig** sein, d.h. sie müssen von verschiedenen Merkmalsträgern stammen
* Die **Fehler** in jeder Treatmentstufe sollten **normalverteilt** sein mit einem erwarteten Mittelwert von 0

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Die Fehlervarianzen in jeder Zelle der ANOVA sollen erwartet (nicht numerisch) gleich sein (**Homoskedastizität**) (Streuung soll gleich sein; bezieht sich auf Zelllen)



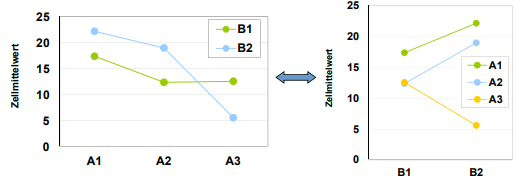
* Treatmenteffekte und Fehler müssen **additiv** sein, d.h. die Fehler dürfen nicht mit den Erwartungswerten der Treatmentstufen korrelieren (bezieht sich auf Faktorstufen)
* **Voraussetzungsprüfung:**
* Die **Unabhängigkeit der Beobachtungen** wird i.d.R. „begründet angenommen“ (educated guess) (nur annehmen, niemals beweisen)
* Zur Prüfung der **Normalverteilungsannahme** der Fehler (i.e. Zellresiduen) wird der Kolmogoroff-Smirnov Test verwendet
* Zur Prüfung der **Homogenität der Fehlervarianzen** wird zumeist der Bartlett Test, seltener der Levene Test bzw. der F-Test verwendet
* Die **Unabhängigkeit der Treatmenteffekte und Messfehler** kann über einen Korrelationstest von Zellmittelwerten und Varianzen geprüft werden (selten praktiziert)
* **Stichprobengröße:**
* Für alle Zellhäufigkeiten sollte nij… ≥ 10 gelten (n = pro Zelle)
* Zellhäufigkeiten nij… sollten (annähernd) gleich sein
* Sind die Zellhäufigkeiten ungleich…

1. müssen verschiedene Größen in der ANOVA-Berechnung **approximiert** bzw. **gewichtet** werden, z.B. der Grand Mean (kann nicht mehr einfach aufsummiert werden, sondern muss mit Faktoren gewichtet werden)
2. ist in der einfaktoriellen ANOVA die Annahme der Varianzhomogenität der Fehler schneller verletzt
3. funktioniert die Quadratsummenzerlegung in der mehrfaktoriellen ANOVA nicht mehr, so dass die Faktoren nicht mehr trennbar sind (nicht mehr zuordenbar)

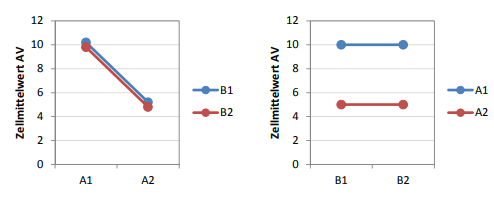
* **Zu große Stichproben** sind gleichermaßen problematisch: wenn n größer dann…
* Treatmentvarianzen größer
* Prüfgrößen F steigen
* Wahrscheinlichkeiten p sinken
* Signifikanzen steigen
* Die ANOVA hat dann zu viel statistische **Power**, d.h. sie kommt selbst bei klein(st)en Unterschieden zu einer Signifikanzaussage
* Dies ist auf die Größe der Fehlerfreiheitsgrade (*df*Error) zurückzuführen, in die der Stichprobenumfang direkt eingeht
* Die praktische Bedeutsamkeit von Effekten bei großen Stichproben sollte über die **Varianzaufklärung η²** bewertet werden!

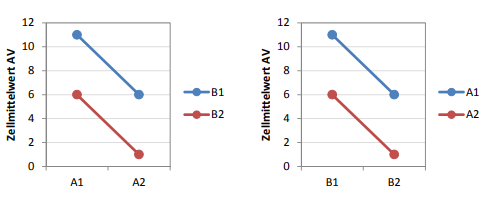
**Plots (Zweifaktorielle ANOVA)**

* **Interaktionsplots:**
* Zur Veranschaulichung der Effekte bei der mehrfaktoriellen ANOVA werden Interaktionsplots verwendet (mithilfe ANOVA Mittelwertstabelle)
* Sie enthalten die Informationen der Zellmittelwerte für zwei Faktoren
* Für solche Plots existieren immer zwei mögliche Darstellungen

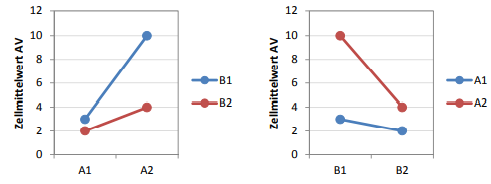


Plot (es könnten auch noch die Mittelwerte als Linie dargestellt werden) Gegenplot

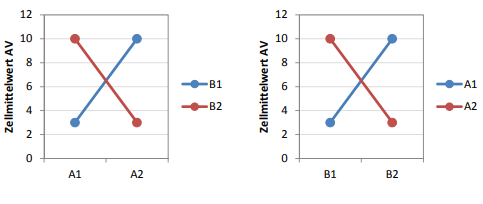
* **typische Effekte:**
* Effekt auf Faktor A, kein Effekt auf Faktor B, keine Interaktion zwischen A und B
* Effekt auf Faktor A, Effekt auf Faktor B, keine Interaktion zwischen A und B



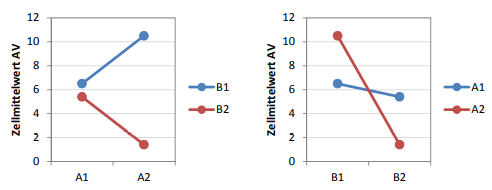
* Effekt auf Faktor A, Effekt auf Faktor B, Interaktion (Linien nicht parallel!) zwischen A und B ► **ordinal**, da sich die Geraden auch im Gegenplot nicht schneiden



* Kein Effekt auf Faktor A, kein Effekt auf Faktor B, Interaktion zwischen A und B ► **disordinal**, da sich die Geraden auch im Gegenplot schneiden



* Kein Effekt auf Faktor A, großer Effekt auf Faktor B, Interaktion ► **hybrid**, da sich die Geraden im Gegenplot schneiden



* TIPP: Wenn nur 1 Diagramm, dann das mit weniger Linien erst probieren!

**3. Einzelvergleiche bei unabhängigen Daten**

**Einzelvergleiche**

* **Problem**: Ein signifikantes Ergebnis in der ANOVA zeigt nicht an, zwischen welchen Treatmentstufen der Effekt besteht; einzelne Stufenmittelwerte miteinander vergleichen
* Für die **Prüfung der Mittelwerte** einzelner Faktorstufen gibt es zwei unterschiedliche Verfahrensweise:

1. **A-Priori Tests** zur Prüfung von Hypothesen, die bereits **vor** der Untersuchung formuliert worden sind (z.B. Faktor A wichtiger als Faktor B)
2. **A-Posteriori Tests** (Post-hoc Tests) zur Prüfung von Hypothesen, die nach Ansehen der Daten gebildet wurden (z.B. Interaktion)

* Bei den **Mittelwertevergleichen** im Rahmen der ANOVA können einzelne und Kombinationen von Mittelwerten verglichen werden
* Es können Fragen beantwortet werden wie:

1. Sind die Mittelwerte zweier Faktorstufen unterschiedlich? (z.B. Fahren Männer besser als Frauen?)
2. Ist eine Faktorstufe unterschiedlich zum Mittelwert aller vorhergehenden Faktorstufen? (z.B. Ist der Mittelwert bei 0% und 0,5% schlechter als 1%; Faktorstufen zusammennehmen)
3. Sind die letzten beiden Faktorstufen unterschiedlich zu den ersten beiden? (z.B. Depressivität der ersten Sitzungen mit den letzten Sitzungen)

* Sind die Mittelwerte zweier Faktorstufen unterschiedlich? (1)

**D = Differenz** z.B. Aj = 0 %, Ak = 0,5%

* Ist eine Faktorstufe unterschiedlich zum Mittelwert aller vorhergehenden Faktorstufen? (2)

Ein Bild, das Text enthält.

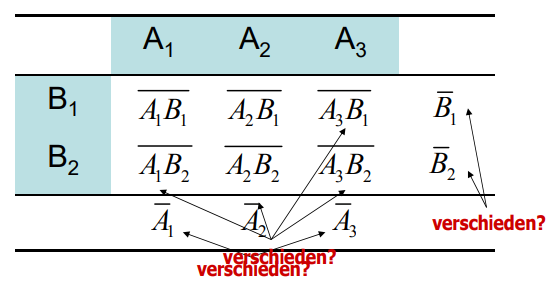
Automatisch generierte Beschreibungz.B. Ap = 1%

* Sind die letzten beiden Faktorstufen unterschiedlich zu den ersten beiden? (3) Ein Bild, das Text, Uhr enthält.

  Automatisch generierte Beschreibung
* Meist nur 1. Fall von Relevanz!!
* **Frage**: Wann ist eine so berechnete Mittelwertdifferenz *D* verschieden genug von Null, damit der Mittelwerteunterschied als statistisch signifikant bewertet wird?
* **Lösung**: Aus dem *D* muss eine **Prüfgröße** mit bekannter Häufigkeitsverteilung konstruiert werden, um die Auftretenswahrscheinlichkeit des gemessenen D zu ermitteln, wenn in Wahrheit kein Unterschied besteht.
* Zur Berechnung **einer F-verteilten Prüfgröße** wird die (theoretische) **Varianz der Mittelwerte** benötigt
* Zusätzlich wird die **Fehlervarianz** benötigt
* Die Berechnung dieser Varianzen ist dabei von der **Art des Mittelwertevergleichs** abhängig.

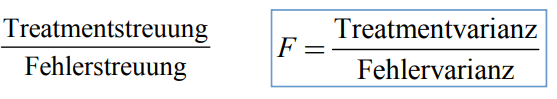
**A-Priori Kontraste**

* Es können sowohl **Stufen**- als auch **Zellmittelwerte** miteinander verglichen werden



B1 und B2 nicht sinnvoll (Unterschied ist nur durch die zwei Mittelwerte); A1 und A3 verschieden? Zellmittelwerte A1B2 und A3B2; auch A2B2 und A3B1 geht

* Ein solcher Vergleich ist **nur** dann erlaubt, wenn der zugehörige Faktor **signifikant** geworden ist
* Zur statistischen Prüfung von **Differenzen in der Mittelwertetabelle** der ANOVA kann eine ganz ähnliche Logik angewandt werden wie bei der ANOVA
* Sind Mittelwerte unterschiedlich, haben sie eine Streuung
* Diese Streuung entsteht a) aus dem **Messfehler** und b) einer eventuell vorhandenen **systematischen Wirkung** der beteiligten Faktorstufen
* Man kann dann wieder eine **Prüfgröße** definieren gemäß



* Zur statistischen Prüfung der Mittelwertsdifferenz lässt sich eine F-verteilte Prüfgröße konstruieren über

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Aus der F-Verteilung lässt sich nun die **Wahrscheinlichkeit *p(F)*** ermitteln, mit der der beobachtete Wert der Prüfgröße auftritt – unter der Annahme, dass es in Wahrheit keinen Unterschied zwischen den Stufenmittelwerten gibt (D=0)
* Ist *p(F)* zu klein, liegt ein statistisch signifikanter Unterschied zwischen den verglichenen Mittelwerten vor, die Treatmentstufen haben verschieden große Effekte
* Die Fehlervarianz zur Bewertung einer Mittelwertestreuung D² in der ANOVA ist prinzipiell:

Ein Bild, das Text, Whiteboard enthält.

Automatisch generierte Beschreibung mit N = Anzahl Personen, die in das D eingehen, wobei sich für spezifische Vergleiche andere Werte ergeben

* Es wird also die in der ANOVA bereits berechnete **Fehlervarianz** zugrunde gelegt und so angepasst, dass sie als Vergleichsgröße für das D² geeignet ist

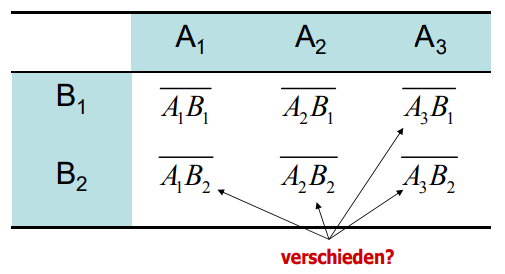
Ein Bild, das Text enthält.

Automatisch generierte Beschreibung



Zu (4): p(F) Wahrscheinlichkeit, ob Mittelwerte gleich sind; wenn p(F) klein, dann ist es sehr unwahrscheinlich, dass sie eigentlich gleich sind

* **Kontraste I:**
* Oft sollen in einer ANOVA beliebige **Zellmittelwerte** paarweise miteinander verglichen werden
* Hat a-priori eine Hypothesenbildung stattgefunden, kann auch dies über **Kontraste** erreicht werden



* Für a-priori Vergleiche zweier beliebiger **Zellmittelwerte** in einer ANOVA mit n Personen je Zelle gilt für den Kontrast

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Fehlervarianz aus ANOVA (nicht nochmal quadrieren!) n= Werte in einer Zelle

* Für die Prüfgröße F gilt:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung *df* = Fehler-Freiheitsgrade in der ANOVA

* **Kontraste II:**
* Oft sollen auch **Stufenmittelwerte** statt der Zellmittelwerte verglichen werden
* Hier gehen in die Mittelwertsdifferenz D mehr Personen als nur die n Personen einer Zelle ein

Ein Bild, das Tisch enthält.

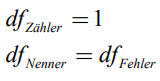
Automatisch generierte Beschreibung

* Für a-priori Vergleiche zweier **beliebiger Stufenmittelwerte** eines Faktors in einer ANOVA mit n Personen je Zelle gilt für den Kontrast

Ein Bild, das Text enthält.

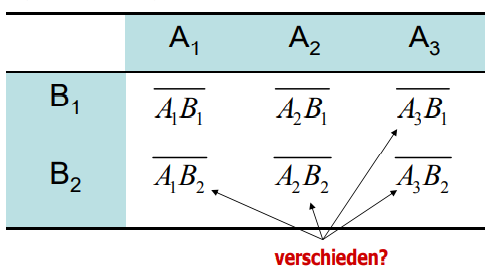
Automatisch generierte Beschreibung

* Für die Prüfgröße F gilt wieder:



**A-Posteriori Tests**

* Gerade bei umfangreicheren Versuchsdesigns und unklaren Hypothesen werden während der Auswertung Effekte entdeckt, für die zuvor **keine Hypothesen** bestanden
* In solchen Fällen ist es trotzdem sinnvoll zu prüfen, ob sich Signifikanzen ergeben, um gezielt Fragestellungen für weitere Untersuchungen zu entwickeln
* **Achtung**: Eine im Nachhinein aufgestellte Hypothese mit einem a-posteriori Test zu prüfen und zu belegen, hat faktisch **keine Aussagekraft** (ist jedoch in der empirischen Forschung durchaus verbreitet)
* **Scheffé Test I:**
* Mit dem Scheffé Test können **beliebige Zellen** miteinander verglichen werden
* Dieser Test ist ein **a-posteriori Test** und kommt eher spät (i.e. bei größeren Unterschieden) zu einer Signifikanzaussage



* Für a-posteriori Vergleiche zweier **beliebiger Zellmittelwerte** einer ANOVA mit *m* Zellen und *n* Personen je Zelle gilt nach Scheffé

Ein Bild, das Text enthält.

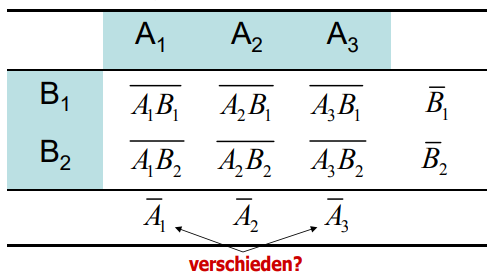
Automatisch generierte Beschreibung

* Für die Prüfgröße Fcorr gilt:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* **Scheffé Test II:**
* Oft sollen in einer ANOVA a-posteriori auch Mittelwerte von **Treatmentstufen** ohne Beachtung des anderen Faktors geprüft werden
* Es gehen hier wieder mehr als nur die n Personen einer Zelle in die Prüfgröße nach Scheffé ein



* Für a-posteriori Vergleiche zweier beliebiger Stufenmittelwerte eines Faktors mit *p* Stufen und *n* Personen je Zelle gilt nach Scheffé

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung hier k = 2 (B1 und B2)

* Die Prüfgröße Fcorr hat

Ein Bild, das Text, Uhr enthält.

Automatisch generierte Beschreibung

* **Weitere A-Posteriori Tests:**
* Der Scheffé Test ist der konservativste unter den üblichen a-posteriori Tests, d.h. er kommt erst bei größeren Mittelwertsunterschiede zu einer signifikanten Entscheidung
* Er ist robust gegenüber Verletzungen der Voraussetzungen der ANOVA (z.B zu kleine Stichproben)
* Andere, progressivere Tests sind der Least Significant Difference Test (LSD) nach Fisher, der Honest Significant Difference Test (HSD) nach Tukey, der Duncan-Test oder der Newman-Keuls Test

**4. Die Varianzanalyse für Messwiederholungsdaten**

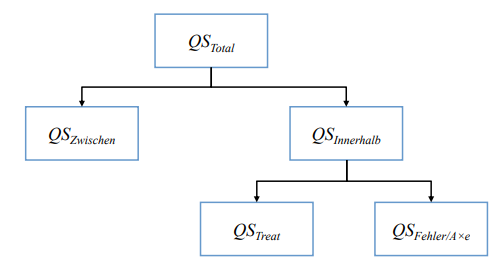
**Einfaktorielle rmANOVA**

* Bisher nur Varianzanalyse für unabhängige Daten (Person kommt nur einmal in ANOVA vor, nicht mehrmals)
* Jetzt Varianzanalyse für abhängige Daten (repeated measures) am Beispiel der **Einfaktoriellen rmANOVA (repeated measures)**
* Erweiterung auf Zwei- oder Mehrfaktorielle rmANOVA ist ohne weiteres möglich
* **Ziel**: Analyse des Einflusses einer **unabhängigen Variablen** (UVn) auf eine **abhängige Variable** (AV) (bei Mehrfaktoriellen rmANOVA mehr unabhängige Variablen dazu)
* Die AV muss dabei **stetig** sein (mind. intervallskaliert) die UVn sind i.d.R **nominal- oder ordinalskaliert**
* Anders als bei der einfaktoriellen ANOVA werden in der rmANOVA **dieselben Merkmalsträger** in allen Treatmentstufen beobachtet („Messwiederholung“)
* **Vorteil**: Die Fehlervarianz wird i.d.R. kleiner (also F eher größer und somit signifikant), da sich die zufällige Unterschiedlichkeit zwischen den Versuchspersonen auf alle Stufenmittelwerte gleichartig ist → **schärfere Testung**
* Es braucht auch nicht auf Varianzhomogenität zwischen den Versuchspersonen geachtet zu werden
* **Nachteil**: Die rmANOVA ist durch **anspruchsvollere Voraussetzungen** gekennzeichnet als die ANOVA
* Zudem wirkt sich der Ausfall eines Merkmalsträgers (Drop-Out) auf alle Stufen aus (Problem Längsschnitt)
* **Wenn anwendbar, ist die rmANOVA trotzdem zumeist das wünschenswertere Design**
* **Annahmen:**
* Die Beobachtungen müssen **abhängig** sein
* Die Fehler innerhalb einer Person sollen normalverteilt sein mit einem Erwartungswert von 0 (z.B. Messgerät bei einer Person verstellt)
* Die Varianzen der Differenzen zwischen den Treatmentstufen müssen gleich sein → **Sphärizitätsannahme**

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Die Sphärizitätsannahme ist eine wichtige Voraussetzung
* Verletzungen führen zu **progressiveren** Entscheidungen bei der der F-Prüfgröße, d.h. es stellen sich schneller Signifikanzen ein (unerwünscht)
* Die Sphärizitätsannahme wird mit dem **Mauchly Test** auf Sphärizität geprüft (wenn er signifikant wird, ist die Sphärizität nicht gegeben)
* Ist die Sphärizität verletzt, so ist eine **Korrektur der Freiheitsgrade** vorzunehmen (aber: Malte empfiehlt, die Freiheitsgradkorrektur **immer** durchzuführen).
* Korrekturmethoden sind die **Greenhouse-Geisser** Korrektur sowie die **Huynh-Feldt** Korrektur
* **QS-Zerlegung:**





QSZwischen = Streuung zwischen Personen (uninteressant) Interaktion

QSInnerhalb = Streuung innerhalb von Personen Person/Faktorstufe

QSTreat = Wirkung der Faktorstufen (interessant)

QSFehler = Streuung innerhalb einer Person (Prüfstreuung)

* Der einfaktoriellen rmANOVA liegt ein einfaches lineares Modell zugrunde:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

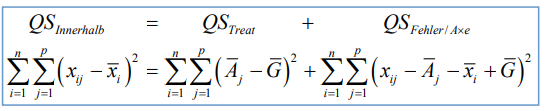


* Daraus lässt sich nun die Quadratsummenzerlegung in der Varianzanalyse ableiten:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Die QSInnerhalb zerfällt in:



* **Prüfgröße & Ergebnistabelle:**
* Aus den Quadratsummen lassen sich wieder spezielle Varianzen bestimmen, die **Populationsvarianzen** 
* Dazu werden wieder die **Freiheitsgrade** (*df*, degrees of freedom) benötigt
* In der rmANOVA gilt:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Damit werden die Populationsvarianzen berechnet als

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Hat das Treatment **keinen Effekt**, so sind Unterschiede zwischen Mittelwerten der Treatmentstufen rein zufällig
* Die Streuung der Stufenmittelwerte kann dann nur aus der Residualstreuung (Fehler) entstehen. Es muss gelten:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Daraus konstruiert man die **Prüfgröße**

Ein Bild, das Text enthält.

Automatisch generierte BeschreibungFehlervarianz innerhalb einer Person, nicht zwischen Personen!

* Sie ist **F-verteilt** mit *dfTreat= p–1* Zählerfreiheitsgraden und *dfFehler= (n-1) · (p-1)* Nennerfreiheitsgraden
* Aus der F-Verteilung kann die **Wahrscheinlichkeit p(F)** für das Auftreten dieser Prüfgröße ermittelt werden

Ein Bild, das Text enthält.

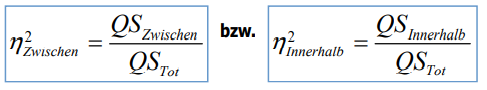
Automatisch generierte Beschreibung

* Man erstellt folgende Ergebnistabelle

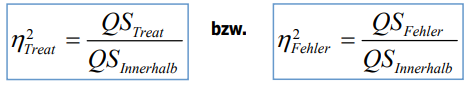
Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

* Zwischen + Innerhalb = Total; A + Fehler = Innerhalb
* Dabei wird die zweite Zeile (QSInnerhalb) manchmal auch weggelassen, v.a. von Statistikprogrammen
* Die Fehlervarianz wird auch als Residualvarianz bezeichnet, um den „Innerhalb“-Fehler in der rmANOVA vom „Zwischen“-Fehler in der ANOVA abzugrenzen
* Zusätzlich ist zu empfehlen, die Varianzaufklärung η² (eta²) für das Model zu bestimmen
* Die Varianzaufklärung für QSZwischen und QSInnerhalb verläuft analog zur ANOVA für unabhängige Daten



* Für den Messwiederholungsfaktor wird jedoch die QSInnerhalb als Referenz gewählt



**Mittelwertevergleiche**

* Wo liegt die Wirkung des Faktors, der einen Effekt hat?
* **Prinzip**: Die Mittelwertevergleiche in der rmANOVA folgen demselben Prinzip wie in der unabhängigen ANOVA
* **Frage**: Wann ist ein Mittelwertsunterschied zwischen Messzeitpunkten verschieden genug von Null, damit das berechnete D als statistisch signifikant bewertet wird?
* **Berechnung**: Es gelten praktisch dieselben Berechnungsformeln wie in der einfaktoriellen ANOVA; aber als Fehlervarianz wird die Residualvarianz genommen
* Wie auch dort existieren zwei verschiedene Testvarianten (gleiche Berechtigung):

1. **A-Priori Tests** zur Prüfung von Hypothesen, die vor der Untersuchung formuliert wurden
2. **A-Posteriori Tests** (Post-hoc Tests) zur Prüfung von Hypothesen, die nach Ansehen der Daten gebildet wurden

**5. Weitere Typen von Varianzanalysen**

**Spezialfall: Einfaktorielle ANOVA**

* Die ANOVA kann problemlos auch mit einer anderen Faktorenzahl als 2 berechnet werden
* Wird nur ein Faktor betrachtet, spricht man von einer **einfaktoriellen ANOVA**
* Es gibt bei der einfaktoriellen ANOVA nur einen Haupteffekt und keine Interaktion
* Die Ergebnistabelle ist dann stark vereinfacht

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

* Fehlerquadratsumme ist höher, weil der Effekt des Faktors B trotzdem noch vorhanden ist (Fehler = B + AxB + Fehler); genauso bei Varianzaufklärung

**Allgemeiner Fall: Mehrfaktorielle ANOVA**

* Die ANOVA kann problemlos auch mit einer anderen Faktorenzahl als 2 berechnet werden
* Die ANOVA Ergebnistabelle z. B. bei 3 Faktoren lautet

Ein Bild, das Tisch enthält.

Automatisch generierte BeschreibungA: Alkoholmenge; B: Geschlecht;

C: Jährlich gefahrene Kilometer

**Weitere wichtige Typen von Varianzanalysen**

* Es können in abhängigen und unabhängigen ANOVA Designs beliebig viele Faktoren mit beliebig vielen Stufen miteinander kombiniert werden
* Auch die Kombination von abhängigen und unabhängigen Faktoren ist möglich (**mixed ANOVA**)
* **Beispiel**: Therapiestudie mit einem Messwiederholungsfaktor, einem Geschlechtsfaktor und einem Faktor „Therapieart“
* **Messwiederholungsdesigns mit Gruppenfaktoren**
* Die Anzahl von Zellen in der ANOVA wird bei multiplen Faktoren mit vielen Stufen sehr schnell sehr groß (**kombinatorische Explosion**)
* Wenn der Experimentator nur an den Haupteffekten interessiert ist, kann der Versuchsplan drastisch reduziert werden, indem nur bestimmte Kombinationen von Stufen der Faktoren realisiert werden
* Interaktionen werden nicht mehr dargestellt

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

* **Lateinische Quadrate**
* In einigen Untersuchungsdesigns wird ein zweiter Faktor nur für bestimmte Stufen eines anderen Faktors realisiert
* **Beispiel**: Therapiestudie mit Placebogruppe, Medikamentengruppe und VT-Gruppe. Die Medikamentengruppe zerfällt in drei Untergruppen (kleine, mittlere, starke Dosis)
* **Hierarchische ANOVA Designs**
* Bei vielen Fragestellungen ist es sinnvoll, verschiedene Treatments an denselben Merkmalsträgern zu testen.
* **Beispiel**: Strahlungsstärken in der Krebstherapie, Reizstärken bei Tests auf Schmerzempfindlichkeit, Effektivität verschiedener Düngemittel
* **Problem**: Die einer Treatmentstufe ausgesetzte Stelle ist „kontaminiert“ und für weitere Stufen unbrauchbar
* **Lösung**: Man teilt den Merkmalsträger auf in verschiedene, als eigenschaftsgleich definierte Stücke („Plots“) und appliziert hier die verschiedenen Stufen des Treatments
* **Split Plot Designs**