# Methoden der Analyse Qualitativer Daten

Mitschrift der Vorlesung
von Dr. Rainer Scheuchenpflug

im WS 07/08

Roland Pfister

Julius-Maximilians-Universität
Würzburg

# Inhaltsverzeichnis

# 0. Vorwort

Dozenten: Dr. Rainer Scheuchenpflug, Dr. Christian Maag, Dipl.-Psych. Ingo Totzke

Termin: Dienstag, 12:00 – 13:30, Külpe-Hörsaal

Klausur: 29.01.2008, 12:00 Uhr. Wird von Christian Maag gestellt.

Web: www.izvw.de bzw. www.psychologie.uni-wuerzburg.de/methoden

# 1. Konfigurationsfrequenzanalyse

## 1.1. Klassische KFA

Das Ziel wissenschaftlichen Handelns lässt sich in folgende Punkte untergliedern (Definition nach Zimbardo):

- Beschreiben
- Erklären
- Vorhersagen
- Beeinflussen.

Die Konfigurationsfrequenzanalyse (KFA) ist dabei klar dem ersten Punkt zugeordnet und dient der Beschreibung, Systematisierung und Klassifikation von Daten. Dabei sollen vor allem charakteristische Muster in den Daten erkannt werden.

Diese Muster werden hierbei als Syndrom bezeichnet. Ein Syndrom stellt somit eine Kombination von Symptomen dar.

### 1.1.1. Anwendungsgebiet

Die KFA ist ein Verfahren zur Analyse mehrdimensionaler Kontingenztafeln und kann somit als Erweiterung der $X^2$-Tests gesehen werden. Folglich sind auch die zugrunde liegenden Hypothesen aus den $X^2$-Tests abgeleitet:

$H_0$: Unabhängigkeit
$H_1$: Abhängigkeit

Mehrdimensional heißt in diesem Zusammenhang 3 oder mehr Variablen bzw. Merkmale. Die klassische KFA wird dabei bei mehrdimensionalen Kontingenztafeln verwendet, bei denen die Merkmale jeweils nur 2 Stufen haben: liegt vor vs. liegt nicht vor.

#### 1.1.1.1. Exkurs: r-x-c-$X^2$-Tests mit Excel

Erstellen einer Tabelle für die erwarteten Häufigkeiten ($f_e$) auf Basis der Tabelle der beobachteten Häufigkeiten ($f_b$):

- Formel zur Berechnung der ersten Zellhäufigkeit in die linke obere Zelle eintragen.
- Spalte fixieren → nach unten aufziehen
- Zeile fixieren → nach rechts aufziehen

Zellen- $X^2$-Werte:

- Formel eingeben: $\frac{(f_b - f_e)^2}{f_e}$
- Aufziehen

Teststatistiken:

- $X^2_{ges}$: SUMME
- $df_{ges}$: ANZAHL
- p: CHIVERT(PG;df)
- $X^2_{crit}$: CHIINV(W'keit;df)

Anmerkung: Für die Umkehrfunktion der $X^2$-Verteilung (CHIINV) wird als Wahrscheinlichkeit der abgeschnittene Bereich angegeben (α).

#### 1.1.1.2. Datenlage

- Nominalskalierte (dichotome) Merkmale
- Mehrere Merkmale
- Gegeben sind die Häufigkeiten der Merkmalskombinationen

## 1.1.2. Typen und Syndrome

Ausgangspunkt der Entwicklung der KFA war ein Befund von Leuner (1962) zur Wirkung von LSD, der als psychotoxisches Basissyndrom bezeichnet wird:

- Bewusstseinstrübung
- Denkstörung
- Affektivitätsbeeinflussung

Interessant an Leuners Basissyndrom ist die Tatsache, dass entweder alle drei Symptome auftreten oder eines davon, nie jedoch zwei. Es fanden sich also keine paarweisen (bivariate) Korrelationen zwischen den Symptomen, wohl aber trivariate Zusammenhänge. Hier stellt sich die Frage, ob solche höherrangigen Zusammenhänge überhaupt auftreten können, wenn die Daten je paarweise unabhängig sind.

Um dieses Problem zu klären, stellte Lienert (1970) ebenfalls LSD-Versuche an, auf deren Basis er die KFA entwickelte. 65 studentischen Probanden wurde hierfür LSD verabreicht und die Symptome Bewusstseinstrübung B, Denkstörung D und Affektstörung A wurden als vorhanden (+) vs. nicht vorhanden (-) klassifiziert.

Die $2^3 = 8$ Kombinationen werden nun nicht mehr in mehrdimensionalen Tabellen angegeben, sondern es werden alle möglichen Kombinationen untereinander geschrieben.

| B | D | A | f |
|---|---|---|---|
| + | + | + | 20 |
| + | + | - | 1 |
| + | - | + | 4 |
| + | - | - | 12 |
| - | + | + | 3 |
| - | + | - | 10 |
| - | - | + | 15 |
| - | - | - | 0 |

Anschließend werden die beobachteten Häufigkeiten f der einzelnen Kombinationen ausgezählt. Es zeigt sich, dass sowohl Leuners Basissyndrom als auch die Einzelhäufigkeiten relativ hoch sind, die Kombination von nur 2 Symptomen dagegen relativ selten.

Wichtig ist in diesem Zusammenhang das Problem der Basisraten: Kommt das Basissyndrom nur deswegen so häufig vor, weil auch B, D, A sehr häufig vorkommen? Und: Sind die Unterschiede auch als statistisch signifikant zu interpretieren?

#### 1.1.2.1. Erwartete Häufigkeiten e

Wenn kein Zusammenhang zwischen den Merkmalen besteht, ergibt sich jede Zellhäufigkeit aus dem Produkt der Randhäufigkeiten (Kombinationsregel für unabhängige Wahrscheinlichkeiten). Im dreidimensionalen Fall:

$$e_{ijk} = \frac{f_{i..} \cdot f_{.j.} \cdot f_{..k}}{N^2}$$ mit f = Randsummen.

Bzw. im allgemeinen Fall:

$$e_{i...t} = \frac{f_i \cdot ... \cdot f_t}{N^{t-1}}$$ mit t = Anzahl der Variablen bzw. Merkmale.

Auf Basis Abweichung der beobachteten Häufigkeiten f von den erwarteten Häufigkeiten e kann nun für jede Symptomkombination (Syndrom) ein X²-Wert berechnet werden:

$$\chi^2_{ijk} \frac{\left(f_{ijk} - e_{ijk}\right)^2}{e_{ijk}}$$

### 1.1.2.2. Test auf Typen (Syndrome)

Lienert (1970) schlägt nun vor, jede Zelle gegen $X^2_{crit} = X^2_1 = 3{,}841$ zu prüfen. Wird diese arbiträre Schranke überschritten, so nennt man die Konfiguration über- bzw. unterfrequentiert (= Syndrom). Meist werden dabei nur überfrequentierte Zellen (Typen) betrachtet, unterfrequentierte (Antitypen) werden ignoriert – obwohl diese genauso aussagekräftig sind.

Da keine Nullhypothese existiert sondern die Zellen gegen eine beliebig festgelegte Schranke getestet werden, kann man hier nicht von einem echten Signifikanztest sondern lediglich von einer Heuristik sprechen (→ Bauchweh).

Ein echter Signifikanztest wäre der Omnibus-Test auf $X^2_{ges}$. Die Freiheitsgrade berechnen sich dabei nach:

df = Anzahl der Zellen – Anzahl der zu schätzenden Rand-Häufigkeiten – 1 für die feste Stichprobengröße

Für t dichotome Merkmale lässt sich die Formel wie folgt angeben:

$df = 2^t - t - 1$

Für polychotome Merkmale gilt allgemein:

$$df = I \cdot J \cdot ... \cdot T - [(I-1) + (J-1) + ... \cdot + (T-1)] - 1,$$

wobei I, J,…T die Kategorienzahl der jeweiligen Variablen beschreiben.

Das statistische korrekte Vorgehen beginnt also mit einem Omnibus-Test („Gibt es überhaupt einen Unterschied?“) mit nachfolgenden Einzeltests (ggf. Alpha-Adjustierung).[1]

| B | D | A | f | e | X=(f-e)²/e |
|---|---|---|---|---|---|
| + | + | + | 20 | 12.5 | 4.50 |
| + | + | - | 1 | 6.8 | 4.95 |
| + | - | + | 4 | 11.4 | 4.80 |
| + | - | - | 12 | 6.2 | 5.43 |
| - | + | + | 3 | 9.5 | 4.45 |
| - | + | - | 10 | 5.2 | 4.43 |
| - | - | + | 15 | 8.6 | 4.76 |
| - | - | - | 0 | 4.7 | 4.70 |
| | | | | Summe | 38.02 |

df = 8 - 3 -1 = 4

### 1.1.2.3. Voraussetzungen

- Unabhängige Daten (jede VP ein Datum)
- Erwartete Zellhäufigkeiten > 5

[1] Vgl. ANOVA mit geplanten Kontraste bzw. post-hoc Tests.

### 1.1.3. Zusammenfassung

Die KFA dient der Suche nach Verteilungsinhomogenitäten im Merkmalsraum. Sie ist ein heuristisches Verfahren, bei dem Syndrome/Typen mangels inhaltlicher Basis aufgrund eines statistischen Kriteriums ($X^2$-Wert) identifiziert werden.

Dies war zu Beginn der Entwicklung der KFA ein klarer Vorteil, weil bis dahin Typen und Syndrome vor allem auf Basis der klinischen Intuition und Erfahrung definiert wurden.

Auch können mit der KFA höhergeordnete Zusammenhänge erfasst werden.

#### 1.1.3.1. Begriffe

- Typen = Höhere Zellbesetzung als erwartet.
- Antitypen = Geringere Zellbesetzung als erwartet.
- Cluster = Hohe Zellbesetzung.
- Anticluster = Geringe Zellbesetzung.

Wichtig ist hierbei, dass nicht jede hohe Zellbesetzung auch gleich eine überfrequentierte Zelle markiert: Problem der Basisraten.

#### 1.1.3.2. Probleme

Es gibt viele potentielle Merkmale, aber immer nur begrenzte Stichprobengrößen. Dies ist problematisch, weil für die $X^2$-Tests alle erwarteten Zellhäufigkeiten > 5 sein müssen.

Um eine geeignete Teilmenge von Merkmalen zu finden, wird die hierarchische KFA verwendet.

## 1.2. Hierarchische KFA

### 1.2.1. Vorgehen

#### 1.2.1.1. Stufe 1: Ein Merkmal weglassen

Bei t Merkmalen (Symptomen) werden verschiedene KFAn berechnet, bei denen je eines der t Merkmale weggelassen wird. Für jede dieser t KFAn wird ein Gesamt-$X^2$ berechnet. Die KFA mit dem höchsten Gesamt-$X^2$ wird gewählt.

#### 1.2.1.2. Stufen 2+: Zwei und mehr Merkmale weglassen

Von t Merkmalen (Symptomen) werden jeweils zwei Merkmale weggelassen und das Gesamt-$X^2$ jeder KFA berechnet. Dabei gibt es

$$\binom{t}{t-2} = \binom{t}{2} = \frac{t(t-1)}{2}$$ [t = Anzahl der Merkmale]

Möglichkeiten. Aus diesen KFAn wird wieder die KFA mit dem höchsten Gesamt-$X^2$ gewählt. Schließlich setzt man diese Prozedur mit 3, 4, …, t-2 Merkmalen fort.

Insgesamt berechnet man also $2^t - t - 1$ KFAn. Verwendet wird schlussendlich diejenige KFA, die am „signifikantesten“ wird, also deren Gesamt-$X^2$ den kleinsten p-Wert aufweist.

### 1.2.1.3. Anmerkungen zur Berechnung

Lässt man ein Merkmal weg, muss man die Zellhäufigkeiten (für M+ und M-) aufaddieren. Dabei addieren sich auch die erwarteten Häufigkeiten.

Die KFAn innerhalb einer Hierarchiestufe (Anzahl weggelassener Merkmale) sind mit denselben df versehen; hier können die $X^2$-Werte also direkt verglichen werden. Zwischen den Stufen muss eine Normaltransformation angewandt werden.

Bei mehr als 5 Merkmalen (df > 26) – Fisher-z-Transformation:

$$z = \sqrt{2\chi^2} - \sqrt{2df - 1}$$

Bei weniger als 26 df – Wilson-Hilferty-Transformation:

$$z = \frac{\sqrt[3]{\frac{\chi^2}{df}} - \left(1 - \frac{2}{9df}\right)}{\sqrt{\frac{2}{9df}}}$$

Bei 1 df – einfache z-Transformation:

$$z = \sqrt{\chi^2}$$

| # Merkmale | D | Sta | Sol | Val | z |
|---|---|---|---|---|---|
| 4 | x | x | x | x | 5.85 |
| 3 | x | x | x | | 3.52 |
| | x | x | | x | 5.33 |
| | x | | x | x | 5.90 |
| | | x | x | x | 2.34 |
| 2 | x | x | | | 1.2 |
| | x | | x | | 4.11 |
| | x | | | x | 5.26 |
| | | x | x | | 0.32 |
| | | x | | x | 3.15 |
| | | | x | x | 1.44 |

*Optimale Merkmalsmenge*

## 1.2.2. Kritik

Interpretiert werden nur Typen, also überfrequentierte Zellen, für die statistische Bewertung werden aber unterfrequentierte Zellen gleichermaßen berücksichtigt. D.h. man wählt zwar die „geeignete Menge“ an Symptomen anhand eines statistischen Kriteriums, vergisst aber dann, sich auch die Antitypen anzusehen.

Ein weiteres Problem besteht darin, dass man viele Tests ohne α-Adjustierung durchführt. Möglichkeiten, dieses Problem zu umgehen sind die (konservative) Bonferroni-Adjustierung oder das sequentielle Verfahren von Bonferroni-Holm. Für die Bonferroni-Adjustierung gilt k = Anzahl der Zeilen.

Noch besser: Kreuzvalidierung.

### 1.2.3. Kreuzvalidierung

Wie bei allen sog. heuristischen Verfahren müssen die Befunde der KFA durch eine Kreuzvalidierung abgesichert werden. Lienert schlägt hierzu vor, die Ausgangsstichprobe zu halbieren (Split-half) und über beide Hälften eine eigene KFA zu rechnen. Wenn beide KFAn die gleichen Typen ergeben, spricht das für eine gewisse Stabilität.

Formaler ist der Vergleich mehrerer Stichproben. Dabei sind zwei Fragestellungen möglich:

- Kreuzvalidierung: Ergeben sich bei mehreren Stichproben aus derselben Population dieselben Konfigurationen (Typen)?
- Diagnostik: Sind die Konfigurationen bei Stichproben aus verschiedenen Populationen hinreichend verschieden?

Der Vergleich mehrerer Stichproben erfolgt über den Konfigurationshomogenitätstest.

## 1.3. Konfigurationshomogenitätstest

### 1.3.1. Vorgehen

1. Man summiert die Zellhäufigkeiten jeder Konfiguration über die Stichproben zum Wert $s_{ijk}$; ijk ist der Index für die Konfiguration (hier: 3 Merkmale).
2. Man berechnet für jede Stichprobe die erwarteten Zellhäufigkeiten unter $H_0$ und die X²-Werte:

$$e_{1ijk} = s_{ijk} \cdot \frac{N_1}{N} \qquad \chi^2_{1ijk} = \frac{\left(f_{1ijk} - e_{1ijk}\right)^2}{e_{1ijk}}$$

3. Sinnvollerweise summiert man die X²-Werte noch für jede Konfiguration über die Stichproben (für später).
4. Dann summiert man die X²-Werte über alle Konfigurationen auf und beurteilt nach der X²-Verteilung mit

   df = (Anzahl_Stichproben - 1) * (Anzahl_Konfigurationen -1).

5. Ist das Gesamt-X² signifikant (die Verteilung also inhomogen), betrachtet man die X²-Werte für jede Konfiguration. X²-Werte, die eine arbiträre Schranke überschreiten, weisen auf Konfigurationen hin, die zwischen den Stichproben diskriminieren.
6. Für diagnostische Zwecke benutzt man nur die Konfigurationen, die zwischen den Stichproben diskriminieren. Ein Individuum wird der Population zugeordnet, in der seine Konfiguration häufiger vertreten ist (zu der sein Typ besser passt).

### 1.3.2. Beispiel

| Sta | Sol | Val | Patienten | Geheilte | s | $e_{Patient}$ | $e_{geheilt}$ | X² |
|---|---|---|---|---|---|---|---|---|
| + | + | + | 15 | 25 | 40 | 17.8 | 22.2 | 0.79 |
| + | + | - | 30 | 22 | 52 | 23.1 | 28.9 | 3.67 |
| + | - | + | 9 | 46 | 55 | 24.5 | 30.5 | 17.60 |
| + | - | - | 32 | 27 | 59 | 26.4 | 32.8 | 2.28 |
| - | + | + | 23 | 14 | 37 | 16.5 | 20.5 | 4.68 |
| - | + | - | 22 | 8 | 30 | 13.3 | 16.7 | 10.12 |
| - | - | + | 14 | 47 | 61 | 27.1 | 33.9 | 11.44 |
| - | - | - | 16 | 12 | 28 | 12.5 | 15.5 | 1.82 |
| | | Σ | 161 | 201 | | | | 52.41 |

Das Gesamt-$X^2$ ist signifikant (df = (2-1)*(8-1) = 7). Markiert sind alle Zeilen, deren Werte nun signifikant von der Schranke $X^2_1$ = 3,841 abweichen. Orange Zeilen – Merkmalsvektor (+-+) und (--+) – sind der Gruppe der Geheilten zuzuordnen, grüne – Merkmalsvektor (-++) und (-+-) der Gruppe der Patienten.

Für alle anderen Merkmalsvektoren wird auf die Diagnose verzichtet bzw. unbekannt ausgegeben (!).

Problem: Es kann vorkommen, dass bestimmte Konfigurationen in beiden Stichproben nicht auftreten (Zellbesetzung = 0). Damit kann kein Homogenitätstest berechnet werden, da der Erwartungswert dann ebenfalls 0 beträgt. Die Bestimmung diskriminierender Konfigurationen ist dennoch möglich, aber nicht mehr zufallskritisch abgesichert.

### 1.3.3. Beispiel II: Maxwell (1961) mit 3 Stpn

| Konfiguration | | | Stichproben | | | |
|---|---|---|---|---|---|---|
| D | U | S | f_Z | f_A | f_nD | $X^2_{Zeile}$ |
| + | + | + | 11 | 19 | 3 | 18.8 |
| + | + | - | 13 | 9 | 6 | 2.2 |
| + | - | + | 3 | 13 | 0 | 25.6 |
| + | - | - | 4 | 12 | 1 | 17.8 |
| - | + | + | 30 | 14 | 44 | 10.1 |
| - | + | - | 38 | 11 | 23 | 7.0 |
| - | - | + | 18 | 9 | 23 | 3.3 |
| - | - | - | 31 | 13 | 32 | 3.7 |
| | | $X^2_{Spalte}$ | 7.7 | 54.4 | 26.4 | |

Interpretation:

- Individuen mit den gelb markierten Konfigurationen gehören mit einer hohen Wahrscheinlichkeit zu der Gruppe der Angstneurotiker (A) und mit einer sehr geringen Wahrscheinlichkeit zu der Gruppe der neurotisch Depressiven (nD).
- Die blau markierte Konfiguration spricht für neurotische Depression, nicht jedoch gegen andere Diagnosen.
- Die Spalten-$X^2$ zeigen, für welche Diagnose die Symptome besonders gut geeignet sind: Angstneurose.
- Für die Diagnose „Zyklothymie“ spielen scheinbar andere Symptome eine Rolle.[2]

## 1.4. Polychotome Merkmale

### 1.4.1. Kontingenz-Strukturanalyse

Die Kontingenz-Strukturanalyse stellt eine Erweiterung der KFA auf Merkmale mit mehr als 2 Stufen dar. Sie entspricht also einem rxc-$X^2$-Test mit mehr als 2 Merkmalen.

Vorgehen: Gesamt-$X^2$ → Typen suchen ($X^2$_Zeile > 3,841) → Hierarchische KFA mit je 2 aus 3 Feldern.

[2] Anmerkung: Zyklothymie ist eine alte Bezeichnung für manisch-depressive Erkrankungen („Himmelhoch jauchzend, zu Tode betrübt.“).

Klassisches Beispiel – Suizidaltypenbestimmung:

- Merkmal 1: Geschlecht (binär/dichotom)
- Merkmal 2: Epoche (1944 vs. 1952; binär/dichotom)
- Merkmal 3: Tötungsmittel (Gas, Erhängen, Schlafmittel, Ertränken, Öffnen der Pulsader, Erschießen, Herabstürzen; multinär bzw. polychotom mit 7 Stufen)

Das Gesamt-$X^2$ der Untersuchung nach Lienert und Krauth (1973) beträgt 348.6 bei 19 df. Dabei ergaben sich folgende Typen:

| Geschlecht | Epoche | Tötungsart | f | e | $X^2$ |
|---|---|---|---|---|---|
| m | 52 | Puls | 22 | 11.43 | 9.73 |
| m | 44 | Erhängen | 76 | 39.22 | 34.48 |
| m | 44 | Erschießen | 35 | 12.32 | 41.75 |
| m | 44 | Herabstürzen | 9 | 3.77 | 7.25 |
| w | 52 | Schlafmittel | 97 | 39.0 | 86.12 |
| w | 44 | Gas | 61 | 44.25 | 6.34 |
| w | 44 | Ertränken | 54 | 25.9 | 30.5 |

Eine hierarchische KFA ergibt:

- Kein Zusammenhang zwischen Geschlecht und Epoche ($X^2$ = 0,004; df = 1).
- Zusammenhang zwischen Geschlecht und Tötungsart (Männer bevorzugen harte Mittel, Frauen eher weiche); $X^2$ = 91.2, df = 6.
- Zusammenhang zwischen Epoche und Tötungsmittel: In Friedenszeiten eher weiche Methoden, in Kriegszeiten harte; $X^2$ = 190.5; df = 6.

Es gibt also bivariate Zusammenhänge zwischen Geschlecht und Tötungsart, Epoche und Tötungsart, aber nicht zwischen Geschlecht und Epoche. Dies ist eine andere Struktur als bei den Leuner-Daten, bei denen es einen trivariaten, aber keine bivariaten Zusammenhänge gab.

Es stellt sich nun die Frage, ob der signifikante Gesamt-$X^2$-Test auf die bivariaten Zusammenhänge zurückgeht, oder ob noch ein trivariater Zusammenhang enthalten ist.[3] Diese Frage wird im Rahmen der Loglinearen Modelle (Kap. 2) genauer dargestellt.

Zur Bewertung der Dreifach-Kontingenz werden die $X^2$-Werte der Zweifach-Kontingenzen vom Gesamt-$X^2$ abgezogen:

$$X^2\_trivariat = 348.6 – 0.004 – 91.2 – 190.5 = 66.896.$$

Die Freiheitsgrade dieses Wertes ergeben sich aus der Differenz der df des Gesamt-$X^2$ und der Summer der df der Zweifach-$X^2$:

$$df\_trivariat = 19 – (1 + 6 + 6) = 6 \text{ df}$$

Da dieser Wert ebenfalls signifikant ist, muss neben den Zweifach-Kontingenzen auch von einer Tripel-Kontingenz gesprochen werden. Die Interpretation ist genauso komplex wie die Interpretation einer 3-fach IA bei einer ANOVA.

[3] Vgl. HE und IA bei ANOVA.

## 1.4.2. Interaktionsstrukturanalyse

Ähnlich wie bei loglinearen Modellen (Kap. 2) vs. kategorialer/logistischer Regression, kann man auch bei der „parameterfreien“ Betrachtung von Lienert zwischen zwei Einsatzzwecken unterscheiden:

- Zusammenhangsanalyse (≈ Korrelation)
- Regressionsanalyse mit abhängigen und unabhängigen Variablen, die eine Wirkrichtung (Kausalität) implizieren (≈ Regression)

Hier werden die beiden Analysearten als

- Kontingenzstrukturanalyse (Zusammenhang) und
- Interaktionsstrukturanalyse (Regression) bezeichnet.

### 1.4.2.1. Suizidalitätstypen

Im Beispiel wäre es sinnvoll, die Wahl des Tötungsmittels als abhängige Variable zu betrachten. Geschlecht und Epoche sind dann die UVn. Die Wahl des Tötungsmittels kann somit abhängen vom Geschlecht („Hauptfaktor 1“), der Epoche („Hauptfaktor 2“) oder der Kombination der Faktoren („Interaktion“).

### 1.4.2.2. Vorgehen

Ob ein Interaktionseffekt vorliegt, wird über den X²-Wert einer zweidimensionalen Tafel bestimmt. Die Variablen sind Tötungsmittel (7 Stufen) vs. Kombination Geschlecht x Zeit (4 Stufen).

Man schreibt die Daten also anders hin:

| | m&52 | m&44 | w&52 | w&44 | Summe |
|---|---|---|---|---|---|
| Gas | 52 | 16 | 47 | 61 | 176 |
| Erhängen | 31 | 76 | 14 | 35 | 156 |
| Schlafmittel | 44 | 7 | 97 | 9 | 157 |
| Ertränken | 20 | 19 | 10 | 54 | 103 |
| Puls | 22 | 15 | 5 | 4 | 46 |
| Erschießen | 3 | 35 | 0 | 11 | 49 |
| Stürzen | 2 | 9 | 2 | 2 | 15 |
| Summe | 174 | 177 | 175 | 176 | 702 |

Die Interaktion ist signifikant, wenn der X²-Wert der Tabelle X²_crit mit (7-1) * (4 -1) df überschreitet.

Die erwarteten Häufigkeiten berechnen sich nun nicht mehr als

$$e_{Gas \times Männer \times 1952} = \frac{\sum(Gas) \cdot \sum(Männer) \cdot \sum(1952)}{N^2}$$

sondern als

$$e_{Gas \times Männer \times 1952} = \frac{\sum(Gas) \cdot \sum(Männer \,\&\, 1952)}{N}$$

Hierdurch ergeben sich andere X²-Werte. Im Beispiel besteht nur deswegen eine geringe Abweichung, weil Geschlecht und Epoche praktisch unabhängig sind.

#### 1.4.2.3. Ergebnis

Aus den Suizidalitätsdaten ergibt sich bei 18 df eine signifikante Interaktion mit $X^2(GxE) = 348.1$.

Die „Hauptfaktoren“ wurden im Rahmen der hierarchischen KFA bereits berechnet – man könnte sie aber auch in diesem Rahmen berechnen, indem man über Geschlecht oder Epoche kollabiert. Diese Einflüsse werden durch eine signifikante „Interaktion“ modifiziert. Interpretation: ?.

### 1.4.3. Inferenzstatistik

Eine zufallskritische Bewertung verlangt natürlich auch hier nach einer Alpha-Adjustierung. Lienert schlägt in diesem Fall die Bonferroni-Adjustierung vor:

$$\alpha^* = \frac{\alpha}{k} \qquad \text{mit k = Anzahl der Tests}$$

Bei einer KSA (Zusammenhangsanalyse ohne Richtung) müsste man also 3 doppelte und 1 Tripelkontingenz prüfen: k = 4.

Bei einer ISA (gerichtete Fragestellung) sind nur 3 Tests durchzuführen (Faktor 1, Faktor 2, Interaktion): k = 3.

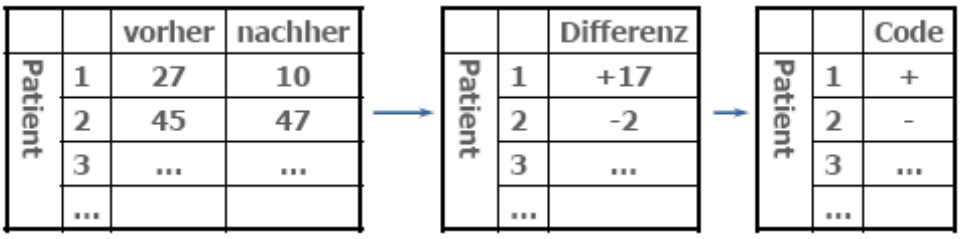
Will man auch die Typen inferenzstatistisch absichern, muss entsprechend mit der Anzahl der Zellen der Tafel (= Anzahl der Konfigurationen = 2 * 2 * 7) adjustiert werden. Damit wird das α* sehr klein: α* = 0.05/2*2*7 = 0.0018.

### 1.4.4. Zusammenfassung KSA und ISA

Sowohl KSA als auch ISA stellen eine Anwendung der KFA auf mehr als 2 Merkmale mit mehr als 2 Kategorien dar. Die KSA untersucht dabei Zusammenhänge und betrachtet alle Variablen daher als gleichwertig (≈ Korrelation), die ISA hingegen untersucht gerichtete Fragestellungen (AV vs. UVn; ≈ Regression).

## 1.5. Abhängige Messungen

Abhängige Messungen lassen sich nicht über die KFA auswerten, da diese unabhängige Daten zwingend voraussetzt. Analog zum t-Test wird in diesem Fall ein Trick verwendet: Statt zwei Werten derselben Person zu verschiedenen Zeitpunkten werden nur die Differenzen betrachtet. Dabei wird kodiert, ob sich jemand verbessert (+) oder verschlechtert hat (-).




Somit hat man wieder eine Fragestellung / Datensituation geschaffen, auf die die KFA anwendbar ist.[4]

[4] Im Beispiel auf den Folien kann ausnahmsweise gerichtet geprüft werden, da eine einseitige Hypothese vorliegt. Praktisch wird hierfür in der $X^2$-Tabelle bei α = .10 statt α = .05 nachgeschlagen, sodass bei 1 df ein $X^2\_crit = 2.71$ resultiert.

## 1.6. Zusammenfassung

Die KFA ist ein Verfahren zur Auswertung mehrdimensionaler Kontingenztafeln und somit eine Erweiterung der $X^2$-Verfahren. Aufgrund der Voraussetzung e > 5 für alle Zellen eignet sie sich vor allem für aktuarische Analysen (großes N).

Vorteile:

- Wenige Annahmen
- Relativ einfache Struktur der Tests (aber: exakte Tests aufwendig)
- Leicht erklär- und durchführbar (versteht sogar ein Mediziner)

Nachteile:

- In Originalvariante sehr heuristisch
- Deutsche Spezifität; übliches Verfahren: loglineare Modelle[5]

## 1.7. Übungssitzung: KFA in Excel

- Nach der Berechnung der Randsummen und der erwarteten Häufigkeiten am besten mit Σ(Randsummen) = $\Sigma(e_{ijk})$ = n auf Fehler prüfen
- Typen: =WENN(Referenz > 3.84;"<==";"")
- $X^2$_ges: =CHIINV(W'keit über Quantil; df); in der Excel-Hilfe zu den einzelnen Verteilungsfunktionen steht, was jeweils unter „Wahrscheinlichkeit" zu verstehen ist
- p: =CHIVERT(Wert;df)
- Bonferroni-Alpha: =0.05/ANZAHL(Zeilen)
- Prinzipiell toll: Hierarchische KFA mit Verweisen auf einmal vorbereitete 3er-KFA

Anmerkungen zu den Übungsblättern:

- Aufgabe (1): Es war nach Typen (überfrequentiert) gefragt und nicht nach Antitypen! Hier: ---+- war ein Antityp. Nach der Bonferroni-Adjustierung finden sich immer noch 7 Typen.
- Aufgabe (3): Bei den Patienten ist nur +-+- ein Syndrom, +--- und -+-- fallen durch die Bonferroni-Adjustierung weg. Das war aber nicht zwingend verlangt.
- Generell: Bei der Wilson-Hilferty-Transformation sollte man seine Eingabe der Formel an bekannten Werten (z.B. der Präsentation) checken
- Kritisch ist auch die Angabe von Wahrscheinlichkeiten bei der CHIINV-Funktion (α oder 1-α?). Hier hilft es, sich bekannte Werte, z.B. CHIINV(0,05;1) = 3,841 ausgeben zu lassen.

[5] Daher sind in gängigen Statistikpaketen (SPSS, Statistica…) keine Anwendungen zur KFA enthalten. Nur für R existieren KFA-Module.

# 2. Loglineare Modelle

## 2.1. Datenlage

Auch loglineare Modelle sind für die Analyse mehrdimensionaler Kontingenztafeln entwickelt worden. Wie die KFA analysieren sie mehrere kategoriale Variablen, die sich nicht explizit als abhängige/unabhängige Variablen interpretieren lassen. Es liegen also Häufigkeiten für Gruppen/Kombinationen von Merkmalen vor, in denen nach Zusammenhängen gesucht wird.

Loglineare Modelle sind damit das parametrische Gegenstück zur KFA. Auch hier sind abhängige Designs schwer möglich.

Vorteile der LM:

- Anwendung der etablierten Theorie und Schätzverfahren der generalisierten linearen Modelle
- International übliche Methode
- Ist in gängigen Statistikpaketen integriert; die KFA findet man nur selten

Nachteile der LM:

- Formalismus nötig (etwas schwerer zu verstehen)
- Möglicherweise schwer zu interpretierende Zusammenhänge (analog zur KFA)

### 2.1.1. Unterschied zur logistischen Regression

1. Die Fragestellung der LM ist nicht gerichtet; es gibt normalerweise keine expliziten AVn und UVn.
2. Die (unabhängigen) Variablen sind nur kategorial. Die logistische Regression kann auch metrische Daten verwerten.

Die logistische Regression lässt sich aber aus einem loglinearen Modell der Daten herleiten. Die Verfahren sind also eng verwandt.

### 2.1.2. Begriffe und Definitionen

- Randhäufigkeiten: $f_{1.}...f_{i.}$, $f_{.1}...f_{.j}$
- Anzahl der Beobachtungen: $f_{..} = N$
- $H_0$: Unabhängigkeit, $H_1$: Abhängigkeit
- Erwartete Zellhäufigkeiten ergeben sich aus dem Produkt der Randhäufigkeiten geteilt durch N:

$$e_{ij} = \frac{e_{i.} \cdot e_{.j}}{e_{..}}$$

Da $e_{i.}$ meist nicht bekannt ist, wird stattdessen $f_{i.}$ genommen.

Der Einfachheit halber logarithmiert man auf beiden Seiten:

$$\ln e_{ij} = \ln e_{i.} + \ln e_{.j} - \ln e_{..}$$

Es handelt sich also um ein lineares Modell (es kommen nur Summen vor), das auf Logarithmen der Häufigkeiten beruht: loglineares Modell.

Die übliche Formulierung benutzt Parameter ähnlich der Varianzanalyse:

$$\ln e_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)}$$

Dabei entsprechen μ dem Grand Mean und die $\mu_{X(z)}$ den Haupteffekten der ANOVA.

Entsprechend gilt, dass sich die einzelnen Ausprägungen der Haupteffekte zu 0 summieren:

$$\sum_{i=1}^{I} \mu_{A(i)} = \sum_{j=1}^{J} \mu_{B(j)} = 0$$

Hier ist also noch kein Interaktionsterm enthalten (Unabhängigkeit), wobei geprüft werden kann, ob dieser benötigt wird, um die Daten zu erklären.

### 2.1.2.1. Saturiertes Modell

Ein Modell, das alle möglichen Abhängigkeiten zulässt (enthält), wird als saturiertes Modell bezeichnet. Im zweidimensionalen Fall:

$$\ln e_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(ij)}$$

Auch die Effekte der Interaktion müssen über alle Variablen selbstverständlich 0 ergeben. Unabhängigkeit bedeutet $\mu_{AB(ij)} = 0$ für alle Paare i, j.

Im zweidimensionalen Fall ist das LM jedoch nicht angebracht (hier genügen $X^2$-Tests). Im mehrdimensionalen Fall:

$$\ln e_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \underbrace{\mu_{AB(ij)} + \mu_{AC(ik)} + \mu_{BC(jk)}}_{\text{Bivariate Interaktionen}} + \underbrace{\mu_{ABC(ijk)}}_{\text{3-fach Zusammenhang}}$$

vgl. Leuner

Das saturierte Modell liefert jedoch keine Informationen über Zusammenhänge, sondern stellt lediglich eine Umparametrisierung der Daten dar (statt Tabelle eine Formel zur Berechnung der Zellbesetzungen).

### 2.1.2.2. Unabhängigkeitsgrade

Je nach fehlenden Parametern ergeben sich verschiedene „Unabhängigkeitsgrade“. Aus dem saturierten Modell werden also IA-Terme gestrichen:

$$\ln e_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \cancel{\mu_{AB(ij)}} + \cancel{\mu_{AC(ik)}} + \cancel{\mu_{BC(jk)}} + \cancel{\mu_{ABC(ijk)}}$$

$$\ln e_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AB(ij)} + \mu_{AC(ik)} + \mu_{BC(jk)} + \cancel{\mu_{ABC(ijk)}}$$

Zweifach-Zusammenhänge können dabei von der Ausprägung der dritten Variable abhängen, z.B. kein Zusammenhang zwischen Zulassung und Studiengang, wenn Geschlecht = m, aber starker Zusammenhang, wenn Geschlecht = w. Ist dies der Fall, liegt eine 3-fach-Interaktion vor.

Wird dagegen der letzte Term weggelassen, so wird von den Variablen eine gewisse Abhängigkeitsstruktur gefordert, d.h. die Art des Zusammenhangs zwischen A und C hängt nicht mehr von der Ausprägung von B ab.

Fehlt neben der 3-fach-IA auch noch ein 2-fach-Interaktionsterm, z.B. $\mu_{AB(ij)}$ gilt: A und B sind bedingt unabhängig (gegeben C). Fehlt zusätzlich noch ein weiterer 2-fach-IA-Term, z.B. $\mu_{BC(jk)}$, so gilt: A und C sind zusammen unabhängig von B.

Fehlen alle IA-Terme, so sind die Variablen unabhängig.

#### 2.1.2.3. Nicht umfassende Modelle

Prinzipiell sind Modelle denkbar, in denen nicht alle Effekte niedrigerer Ordnung enthalten sind. Modelle, die nur den Interaktionsterm enthalten, aber nicht die zugehörigen Haupteffekte oder Interaktionen niedrigerer Ordnung, werden wegen Interpretationsschwierigkeiten nicht betrachtet.

Wird eine IA signifikant, so müssen also auch die zugehörigen Haupteffekte aufgenommen werden. Modelle, die nach dieser Vorschrift gebildet werden, nennt man hierarchische (lineare) Modelle.

# 2.2. Berechnung

## 2.2.1. Saturiertes Modell

$$\ln e_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AB(ij)} + \mu_{AC(ik)} + \mu_{BC(jk)} + \mu_{ABC(ijk)}$$

Die Parameter des saturierten Modells lassen sich direkt berechnen:

$$\mu = \frac{1}{IJK}\sum_{i,j,k} \ln e_{ijk}$$

$$\mu_{A(i)} = \frac{1}{JK}\sum_{j,k} \ln e_{ijk} - \mu \qquad \mu_{AB(ij)} = \frac{1}{K}\sum_{k} \ln e_{ijk} - \mu_{A(i)} - \mu_{B(j)} - \mu$$

$$\mu_{B(j)} = \frac{1}{IK}\sum_{i,k} \ln e_{ijk} - \mu \qquad \mu_{AC(ik)} = \frac{1}{J}\sum_{j} \ln e_{ijk} - \mu_{A(i)} - \mu_{C(k)} - \mu$$

$$\mu_{C(k)} = \frac{1}{IJ}\sum_{i,J} \ln e_{ijk} - \mu \qquad \mu_{BC(jk)} = \frac{1}{I}\sum_{i} \ln e_{ijk} - \mu_{B(j)} - \mu_{C(k)} - \mu$$

$$\mu_{ABC(ijk)} = \ln e_{ijk} - \mu - \mu_{A(i)} - \mu_{B(j)} - \mu_{C(k)} - \mu_{AB(ij)} - \mu_{AC(ik)} - \mu_{BC(jk)}$$

## 2.2.2. Likelihood-Quotiententest

Mithilfe des eines Likelihood-Quotiententests[6] lässt sich nun zwischen verschiedenen loglinearen Modellen entscheiden:

$$LQ = \frac{L\left(Modell\_1 \mid Daten\right)}{L\left(Modell\_2 \mid Daten\right)}$$

Im Zähler steht das restriktivere Modell (das eine gewisse Anzahl von Unabhängigkeiten impliziert), im Nenner steht das allgemeinere Modell (mit mehr Parametern), im schlimmsten Fall das saturierte Modell.

L = Wahrscheinlichkeit, mit dem jeweiligen Modell genau die vorliegenden Daten zu erzeugen.[7] Wenn die L's hinreichend ähnlich sind (LQ ≈ 1), wird das restriktivere Modell 1 beibehalten.

Man streicht also so lange Parameter aus dem Modell, bis der LQ-Test eine schlechte Passung bescheinigt und nimmt dann das vorherige Modell.

[6] Eigentlich sind alle gängigen Tests Likelihood-Quotiententests, das alleine hilft also nicht viel. Man muss jetzt noch bestimmen, wie die Likelihoods berechnet werden.

[7] Streng genommen sind Likelihoods „Funktionen der Parameter, gegeben die Daten"; dagegen sind Wahrscheinlichkeitsfunktionen „Funktionen der Daten, gegeben die Parameter".

#### 2.2.2.1. Berechnung der Likelihoods

Auf Basis der Multinomialverteilung ergibt sich das Likelihood zu:

$$L = \frac{x!}{x_{11}! \cdot x_{12}! \cdot \ldots \cdot x_{IJ}!} \cdot p_{11}^{x_{11}} \cdot p_{12}^{x_{12}} \cdot \ldots \cdot p_{IJ}^{x_{IJ}}$$

$$\text{mit } p_{IJ}^{x_{IJ}} = W'keit_{Zelle}^{H'keit/Anzahl}$$

Die $x_{ij}$ sind Daten (Zellhäufigkeiten), die $p_{ij}$ ergeben sich aus dem Modell:

$$p_{IJ}^{x_{IJ}} = \frac{e_{ij}}{N},$$

wobei $e_{ij}$ aus dem Modell entnommen wird:

$$\ln e_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AB(ij)} + \mu_{AC(ik)} + \mu_{BC(jk)} + \mu_{ABC(ijk)}$$

Gesucht wird nun die Menge von Parametern $\mu^{***(???)}$, die L maximiert. Diese Rechenoperation kann nicht mehr mit der Hand durchgeführt werden.

#### 2.2.2.2. Prüfung des Likelihood-Quotienten

Die LQ folgen einer X²-Verteilung. Genauer: $(-2 - \ln LQ) \sim \chi^2$. Es gibt also ein X²-Kriterium für die Passung des Modells zu den Daten.

Wie bereits erwähnt werden nun so lange Parameter des Modells gestrichen, bis das Modell nicht mehr passt. Dabei sollte nach einer bestimmten Regel vorgegangen werden (hier für 3 Variable): Zunächst wird die 3-fach-Interaktion gestrichen, danach zusätzlich je eine 2-fach-IA. Schließlich wird geprüft, ob zusätzlich eine weitere IA entfernt werden kann oder schließlich sogar Haupteffekte.

#### 2.2.2.3. Test auf $H_0$

Im LQ-Test soll eigentlich die $H_0$ beibehalten werden ($H_0$ = Die Modelle erklären die Daten gleich gut, Abweichungen der Likelihoods sind nur Zufall).

Daher sollte eigentlich ein größerer Ablehnungsbereich gewählt werden, z.B. α = 0.2, um den β-Fehler möglichst klein zu halten. Für Klausur und Übungsaufgaben gilt – wie in der Praxis ☺ – α = 0.05.

### 2.2.3. Logit-Analyse

Loglineare Modelle analysieren Zusammenhänge zwischen Variablen, ohne bestimmte Variablen als UVn oder AVn zu spezifizieren. Wird eine Variable als abhängig betrachtet, bezeichnet man die loglineare Modellierung als Logit-Analyse (vgl. 4.2.1).[8]

Anstelle der Haupteffekte interessieren bei Logit-Analysen nur die Interaktionen der UVn mit der AV. Bsp.: Vorhersage „Vegetarier ja/nein“ aus den Variablen Geschlecht, Alter und geographische Herkunft.

Es interessieren die Zusammenhänge zwischen Veg und Geschlecht (Interaktion VxG), zwischen Veg und Alter, etc., und ob diese Zusammenhänge durch die anderen Variablen modifiziert werden: 3-fach Interaktion.

[8] Wenn die UVn metrisch und nicht nur kategorial sind, landet man bei der logistischen Regression.

# 2.3. Hinweise

## 2.3.1. Voraussetzungen

Die erwarteten Zellhäufigkeiten für alle Zweifach-Zusammenhänge sollten:

- größer als 1 sein und
- nicht mehr als 20% sollten kleiner 5 sein

Sind diese Voraussetzungen nicht erfüllt, so verliert das Modell Teststärke. Zusätzliche Voraussetzungen:

- Unabhängige Beobachtungen
- N > 5 x Anzahl der Zellen; sonst konvergieren die Modelle evtl. nicht
- Bei zu vielen Variablen nicht mehr interpretierbar
- Seltene Ereignisse vermeiden, da die nötigen erwarteten Häufigkeiten sonst nur bei extrem großen N realisierbar sind. Lösung: Produkt-multinomial erheben. Symptomträger suchen und danach auffüllen mit Nicht-Symptomträgern. Da man hierdurch Randhäufigkeiten festlegt verliert man Freiheitsgrade.

Häufig wird eine Konstante (0.5) zu allen Zellen addiert (auch von SPSS). Dies ist jedoch nicht theoretisch begründbar.

## 2.3.2. Generating Class

Als Kurzschreibweise für das gefundene hierarchische Modell im Text eines Artikels wird die generating class (modellbildende Klasse) angegeben. Dabei werden die höchstwertigen Interaktionen/Effekte, die noch im hierarchischen Modell enthalten sein müssen, angegeben.

Alle niedergeordneten Effekte, die in Interaktionen enthalten sind gelten also als automatisch gegeben. Nur solche die noch nicht durch die höchstgeordneten Interaktionen festgelegt sind müssen zusätzlich angegeben werden.

Die generating classes werden auch von SPSS ausgegeben.

Beispiele für generating classes (Z = Zulassung, G = Geschlecht, F = Fach):

- „Zulassung x Geschlecht“:

$$\ln e_{ijk} = \mu + \mu_{Z(i)} + \mu_{G(j)} + \cancel{\mu_{F(k)}} + \mu_{ZxG(ij)} + \cancel{\mu_{ZxF(ik)}} + \cancel{\mu_{GxF(jk)}} + \cancel{\mu_{ZxFxG(ijk)}}$$

- „Zulassung x Geschlecht; Fach“:

$$\ln e_{ijk} = \mu + \mu_{Z(i)} + \mu_{G(j)} + \mu_{F(k)} + \mu_{ZxG(ij)} + \cancel{\mu_{ZxF(ik)}} + \cancel{\mu_{GxF(jk)}} + \cancel{\mu_{ZxFxG(ijk)}}$$

- „Zulassung x Geschlecht x Fach“ (Saturiertes Modell):

$$\ln e_{ijk} = \mu + \mu_{Z(i)} + \mu_{G(j)} + \mu_{F(k)} + \mu_{ZxG(ij)} + \mu_{ZxF(ik)} + \mu_{GxF(jk)} + \mu_{ZxFxG(ijk)}$$

# 2.4. Anwendungsbeispiele

## 2.4.1. Desegregationsprogramme

Auf Basis der Idee, dass kooperativer, sozial sanktionierter Kontakt zwischen Mitgliedern verschiedener Ethnien (gleiche Hierarchiestufe!) Vorurteile reduzieren kann, wurden verschiedene Desegregationsprogramme entworfen. Diese waren grundsätzlich auf einer Makroebene angesiedelt, z.B. Mischung von Stadtvierteln oder Durchmischung von Schulen durch Busing-Programme.

Clack, Dixon und Tredoux (2005) untersuchten, inwiefern sich diese Programme auch auf das alltägliche Leben (Mikrolevel) auswirken, da aus der Literatur bekannt ist, dass derartige Versuche nur mäßige Resultate erzielen. Hierzu beobachteten sie eine Cafete auf einem Uni-Campus. Die gesamte Universität (in England) kann dabei als desegregiert gesehen werden: 83% Weiße, 17% andere Ethnien.

Beobachtet wurden verschiedene Bereiche der Mensa an 3 Tagen. Jeweils wurde kodiert, welche ethnische Zugehörigkeit und welchen Sitzplatz Personen hatten (Weiße, Asiaten, Schwarze und andere) sowie ob diese Personen eine soziale Einheit bilden oder isoliert sitzen. Zusätzlich wurde das Geschlecht kodiert.

→ Die beiden erwarteten Haupteffekte (Ethnie und Bereich der Cafete) sowie deren Interaktion wurden signifikant.

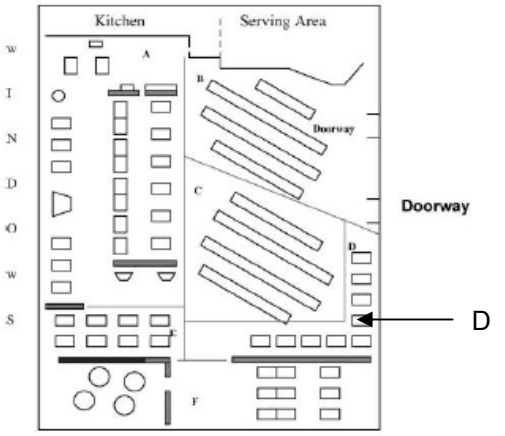


Figure 1. Section boundaries of seating arrangements; shaded areas indicate screens.

$HE_{Ethnie}$: In der Stichprobe finden sich mehr Weiße (Basisrate). $HE_{Bereich}$: Manche Bereiche sind beliebter als andere. Interaktion: In Bereich D finden sich viele Asiaten. Es findet sich also tatsächlich eine Segregation auf dem Mikrolevel.

Anschließend erfolgte eine gerichtete Analyse (logistische Regression), die zwei Moderatorvariablen aufdeckte:

- Wenn mehr als 40 Personen in der Cafete waren, dann war die Wahrscheinlichkeit, dass eine gemischte Tischgruppe auftritt, halb so groß wie bei kleineren Besetzungsdichten.
- Wenn eine Tischgruppe nur aus Frauen bestand, war die W'keit, dass eine gemischte Tischgruppe vorliegt, etwa halb so groß wie bei geschlechtsgemischten oder reinen Männergruppen.

Die Wahrscheinlichkeit für gemischte Tischgruppen wird also durch Geschlechtszusammensetzung und das Gedränge modifiziert. Dabei entspricht der Geschlechtsbefund bekannten Ergebnissen aus der Literatur. Der Gedrängebefund ist dagegen kontraintuitiv.

## 2.4.2. Mathefehler

Marshall (1983) untersuchte die Stabilität von Fehlern, die Jungen und Mädchen in Mathematiktests machen. Dabei bestand die Vermutung, dass die beiden Probandengruppen unterschiedliche Arten von Fehlern „bevorzugen".

Insgesamt wurden in 4 aufeinander folgenden Jahren Erhebungen durchgeführt, sodass sich ein 2 x 4 x 4 – Plan (Geschlecht x Erhebungsjahr x Fehler (3 Distraktoren und keine Antwort)) für jedes der 160 Items des kalifornischen Reihentests ergibt.

#### 2.4.2.1. Hypothesen

- Fehler x Geschlecht: Parameter muss im Modell verbleiben
- 3-fach IA sollte nicht signifikant werden (stabiler Effekt)
- Haupteffekte uninteressant, müssen aber aufgenommen werden (hierarchisches Modell)
- Braucht man noch weitere 2-fach Interaktionen?

#### 2.4.2.2. Ergebnis

Das Modell wird an allen 160 Items getestet (Bauchweh: α-Adjustierung!), wobei ein Modell am besten passt, das nur zweifach Effekte enthält: Fehler x Geschlecht und Fehler x Jahr.

Letzteres ist schwer interpretierbar. Die IA von Geschlecht und Fehlern entspricht jedoch der Erwartung und müsste durch nachgeschaltete Fehlerklassifikationen weiter untersucht werden. Dies wurde in der Studie jedoch nicht realisiert.

## 2.5. Beziehung loglinearer Modelle zu anderen Verfahren

### 2.5.1. Loglineare Modelle und KFA

Loglineare Modelle und die klassische KFA lassen sich nicht ineinander überführen – es handelt sich also um völlig verschiedene Verfahren. Die KFA kann dabei allerdings als Residualanalyse eines loglinearen Modells aufgefasst werden, bei dem Unabhängigkeit gilt (also nur Haupteffekte vorhanden sind).

Das loglineare Modell interessiert sich als für Zusammenhänge bzw. Klassen von Zusammenhängen zwischen Variablen. Die KFA hingegen interessiert sich für einzelne abweichende Zellen („Gibt's irgendwo einen Typ?“).

### 2.5.2. Loglineare Modelle und logistische Regression

Loglineare Modelle haben keine Richtung, werden aber häufig mit Richtung verwendet (vgl. Kontingenz-Strukturanalyse).

→ Korrelation

Bei der logistischen Regression interessiert die Vorhersage einer abhängigen kategorialen (dichotomen) Variablen durch mehrere unabhängige kategoriale oder (wie meistens) metrische Prädiktoren (vgl. Interaktionsstrukturanalyse).

→ Multiple Regression

Die beiden Verfahren sind jedoch ineinander überführbar.

# 2.6. Übung: Loglineare Modelle

Anmerkung: Die beiden Autorinnen Tabachnik und Fidel haben sich über einen Bauchtanzkurs an der Uni kennen gelernt und dabei beschlossen, ein Buch über multivariate Statistik zu schreiben ☺.

## 2.6.1. Hierarchische Modellsuche

SPSS bietet zwei verschiedene Verfahren mit leicht unterschiedlichen Funktionen:

- HILOGLIN (alt): Automatische Modellauswahl
- GENLOG (neu): Allgemeine (general) Modelltests

Wird ein Modell für die loglineare Analyse ausgewählt, so kann man in „Bereich definieren“ bestimmte Variablen ausschließen. Dies erlaubt die Analyse von Subtafeln.

## 2.6.2. Alte Verfahren

Anmerkung: SPSS streicht niemals alte Verfahren, da sich ja irgendjemand dran gewöhnt haben könnte. Daher findet sich im Output immer jede Menge Zeug, das man nicht im Geringsten brauchen kann.

Die Tabelle zu „Effekten der Ordnung k und höher“ beruht auf einem alten Verfahren und zeigt die Modellanpassung, wenn man bestimmte Parameter aus dem Modell entfernt.

Der obere Teil der Tabelle ist hierarchisch, der untere nicht hierarchisch (pfui!). Großes Problem: Wird der Test bspw. für 2-fach Interaktionen signifikant, so sagt er nur, dass bestimmte Zweifach-Interaktionen gebraucht werden. Welche das sind lässt sich über das Verfahren nicht ermitteln.

Diese Aussage ist jedoch in der Assoziationstabelle enthalten, die unter *Modellauswahl\Optionen* [dt.: Partielle Zusammenhänge] ausgewählt werden kann.

Wichtig: Partielle Assoziationstests sind keine Tests auf $H_0$ sondern klassische Signifikanztests. Man kann also die signifikanten Werte verwenden.

Folgt man der Tabelle resultiert für die Beispieldaten folgendes Modell:

$$\ln e_{ijk} = \mu + \mu_{Lesestoff} + \mu_{Beruf*Geschlecht}.$$

Es handelt sich also um ein nicht hierarchisches Modell, sodass keine Interpretation möglich ist. Um zu hierarchischen Modellen zu kommen empfiehlt sich die Funktion der automatischen Modellsuche.

## 2.6.3. Automatische Modellsuche (HILOGLIN)

Die automatische Modellsuche resultiert immer in hierarchischen Modellen. In der Tabelle *Zusammenfassung der Schritte* werden nun die verschiedenen Modelle und deren jeweilige Passung ausgegeben.

An den Beispieldaten zeigt sich für die 3-Fach-Interaktion (Beruf x Geschlecht x Lesestoff) ein $p(X^2) = .397 > .05$. Daher kann die 3-fach Interaktion entfernt werden.

Im nächsten Schritt wird jeweils eine der 2-fach Interaktionen entfernt. Diejenige, deren Fehlen die Passung am wenigsten beeinflusst wird entfernt, dann werden die verbleibenden 2-fach Interaktionen betrachtet, sodass am Ende nur noch die Interaktion Beruf x Geschlecht im Modell verbleibt.

Da alle Interaktionen, die Lesestoff enthalten, aus dem Modell entfernt wurden muss schließlich der Haupteffekt Lesestoff geprüft werden, der jedoch nicht entfernt werden darf.

→ generating class: „Beruf x Geschlecht; Lesestoff“

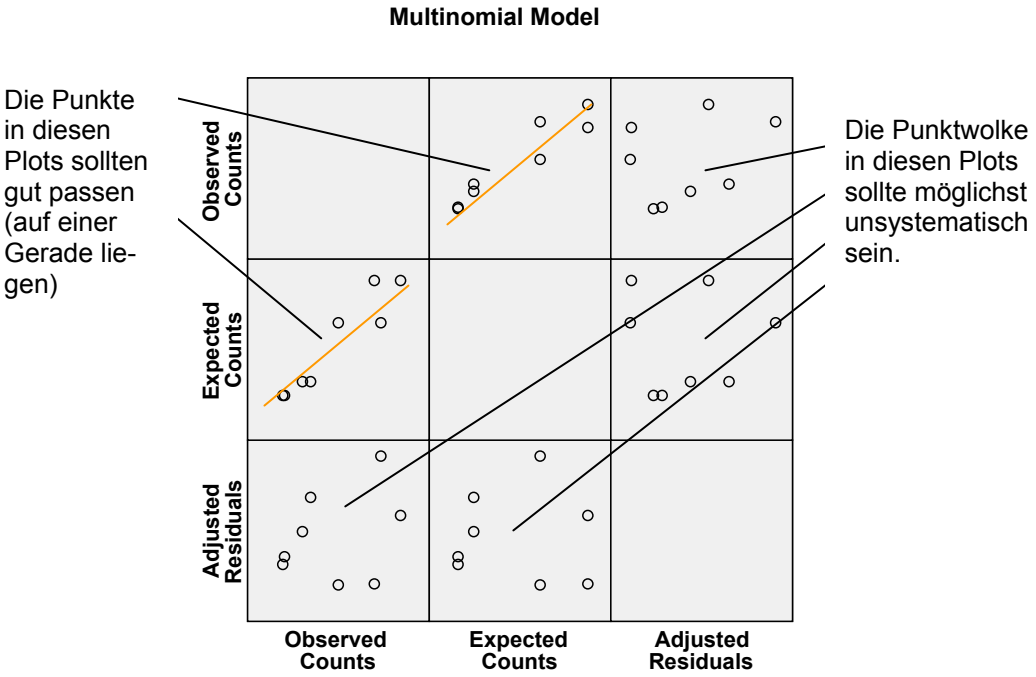
Beruf und Geschlecht sind also zusammen unabhängig von Lesestoff und die Lesestoffverteilung ist nicht homogen (mehr SciFi als Krimis). Letzteres ist aber meistens wenig interessant.

**ABER:** Eigentlich ist das alles nur rumstochern und gucken was rauskommt. Wesentlich typischer ist der Test eines hypothesenbasierten Modells.

## 2.6.4. Modelltest (GENLOG)

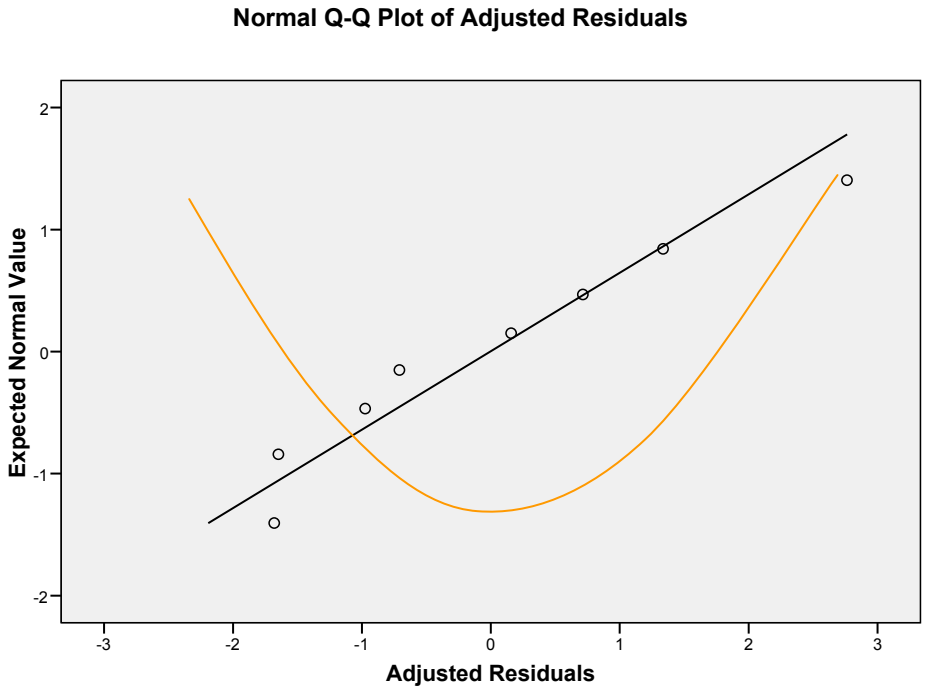
Der Test eines hypothesenbasierten Modells findet sich im Menüpunkt *Hierarchische Modelltests* (GENLOG). Hier lässt sich das Modell spezifizieren. Wichtig ist, dass nicht die generating class angegeben wird, sondern tatsächlich alle enthaltenen Effekte – also auch die Haupteffekte höhergeordneter Interaktionen.

GENLOG liefert dabei auch diagnostische Plots:




Bei deutlichen Ausreißern in einem der beiden Diagramme links oben bzw. im Q-Q-Diagramm sollte nochmals ein Blick in die Daten geworfen werden.

Auch das Q-Q-Diagramm sollte möglichst wenig um die Gerade streuen. Nimmt die Punktwolke andere Formen an (z.B. eine Parabel wie angedeutet), so ist dies ein Hinweis auf ein fehlspezifiziertes Modell. Evtl. wurde beispielsweise ein bedeutsames Symptom vergessen.




Praktisch: Unter der eigentlich wenig nützlichen Tabelle zu *Zellhäufigkeiten & Residuen* ist das verwendete Modell nochmals aufgeführt.

## 2.6.5. Reanalyse: LSD-Daten

Die LSD-Daten aus Leuners (1962) Versuch zur Konfigurationsfrequenzanalyse lassen sich auch mit loglinearen Modellen untersuchen. Zunächst muss hier die Datenbasis (Tabelle mit Häufigkeiten der Symptomkombinationen) in SPSS eingetragen werden. Hierfür stehen zwei Möglichkeiten zur Verfügung:

- Daten stupide als Liste eintragen:

| Fälle | B | D | A |
|---|---|---|---|
| 1 | + | + | + |
| 2 | + | + | + |
| 3 | + | + | + |
| … | | | |
| 43 | + | + | - |

- Daten als Tabelle inkl. „Anzahl“ einlesen lassen und unter *Daten\Fälle* gewichten die Anzahl wirklich als solche verwenden. Im SPSS-Hauptfenster wird jetzt rechts unten „Gewichtung an“ angezeigt. Außerdem erkennt man die aktive Gewichtung in Tabellen und Syntax.

Leuners Fragestellung lässt sich in die Sprache der loglinearen Modelle überführen, indem nur das saturierte Modell zutreffen sollte: die 3-fach Interaktion Bewusstseinstrübung * Denkstörung * Affektivitätsstörung wird gebraucht um die Daten zu erklären. [Anmerkung: Eigentlich sollte man doch ein nichthierarchisches Modell testen???].

Sowohl historische Verfahren als auch die automatische Modellsuche ergeben, dass die 3-fach Interaktion tatsächlich gebraucht wird.

# 3. Epidemiologie

## 3.1. Was ist Epidemiologie


*„Alle, die von dem Heilmittel trinken,*
*erholen sich nach kurzer Zeit.*
*Ausgenommen diejenigen, denen es nicht hilft,*
*die sterben allesamt.*
*Es ist daher offensichtlich,*
*dass es nur bei unheilbaren Fällen versagt.“*

Galen (129-199 n. Chr.)

Das Wort Epidemiologie leitet sich aus den griechischen Wörtern epi (über), demos (das Volk) und logos (die Lehre) ab, sie ist also die „Lehre über das Volk“. Im heutigen Sinne untersucht die Epidemiologie Auftreten und Verbreitung von Krankheiten in der Bevölkerung sowie beteiligte Faktoren in der Person und der Umwelt (Suche nach Risikofaktoren).

Die Epidemiologie ist vor allem eine beobachtende Wissenschaft an menschlichen Populationen, was spezifische Fehler mit sich bringt. Sie ist durch ein hohes Maß an Interdisziplinarität gekennzeichnet.

### 3.1.1. Ziele

Allgemein lassen sich folgende Ziele definieren:

- Erkennen von Krankheitsursachen (Ätiologie) und Risikofaktoren
- Beschreibung von Häufigkeit und Verteilung von Krankheiten
- Verlauf, Prognose, Diagnose von Krankheiten
- Bestimmung von Erkrankungsrisiken
- Evaluation von präventiven und therapeutischen Maßnahmen
- Entscheidungshilfe für die Gesundheitspolitik

Auf einem abstrakteren Niveau lassen sich zudem verschiedene Zielebenen unterscheiden: Deskriptive Feststellung eines Zusammenhangs und Hypothesengenerierung (**deskriptive Epidemiologie**), beobachtende Überprüfung der Hypothesen sowie Quantifizierung der Effekte (**analytisch-beobachtende Epidemiologie**) und schließlich das Ableiten von Interventionen sowie deren Evaluation **(analytisch-experimentelle Epidemiologie**).

Beispielhaft lassen sich diese Zielebenen, die auch den voranschreitenden Forschungsprozess charakterisieren, am Beispiel des Zusammenhangs zwischen Fluor und Karies verdeutlichen.

Zunächst wurde beobachtet, dass die Karieshäufigkeit scheinbar vom Fluoridgehalt im Trinkwasser abhängt, indem verschiedene Städte miteinander verglichen wurden (deskriptiv).

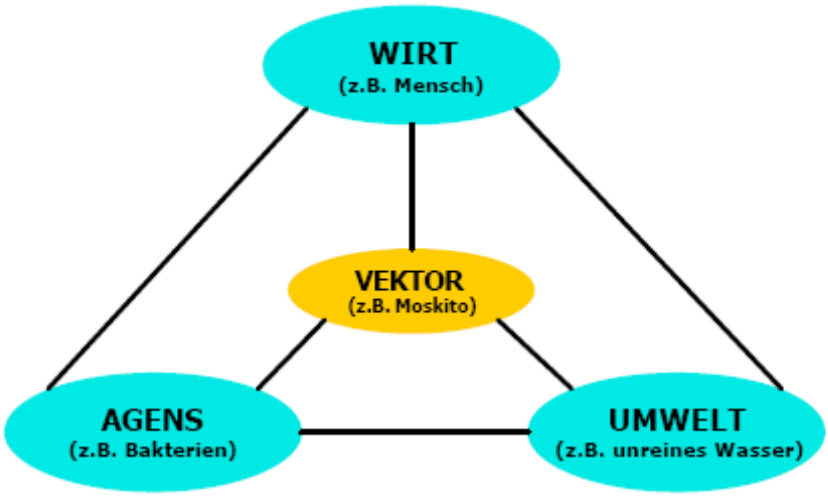
Als quasi-experimentelle Überprüfung der abgeleiteten Hypothese wurde das Trinkwasser zweier vergleichbarer Städte mit Fluorid versetzt vs. nicht. Nach 10 Jahren fanden sich in der Stadt mit fluoridiertem Trinkwasser deutlich weniger Kariesfälle. [Anmerkung zur Folie: DMF = diseased, missing filled; DMF-T (tooth/teeth) ist auf den ganzen Zahn bezogen. Für das Milchzahngebiss werden nur Kleinbuchstaben verwendet.]

Ein anderes Beispiel ist die Untersuchung der Steigerung der Lebenserwartung von 1900 – 1996, wobei sich zeigte, dass diese vor allem auf eine verminderte Kindersterblichkeit zurückgeht.

Weitere interessante Fragestellungen:

- Unterschiedliche Verbreitung von Cholera in den Stadtvierteln von London. Diese lag weder an unterschiedlich guter Luft noch unterschiedlicher Höhe, sondern an der Nähe zur Themse, die damals als Trinkwasserquelle diente.
- Zigaretten und Lungenkrebs
- KHK (Framingham)
- Pockenerkrankungen bei Kuhmägden wesentlich seltener als in der Normalbevölkerung. Häufig erfolgt jedoch eine Ansteckung bei den Kühen mit den wesentlich harmloseren Kuhpocken. Diese scheinen also gegen normale Pocken zu impfen (obwohl man zu der Zeit noch nicht wusste, was impfen ist).

### 3.1.2. Epidemiologische Triade




Unter der epidemiologischen Triade versteht man ein Rahmenmodell bzgl. der Ursachen sowie Kausalstrukturen von Krankheiten.

Vektoren sind nur für manche Krankheiten relevant und bezeichnen Zwischenstationen über die ein Krankheitserreger zum Menschen gelangen kann. Dies können beispielsweise Moskitos sein, die nicht an bestimmten Erregern erkranken die jedoch für den Menschen eine akute Gefährdung darstellen. Zur Prävention von Krankheiten bietet sich neben der Impfung also auch die Eliminierung der relevanten Vektoren an.

Die anderen 3 Komponenten der epidemiologischen Triade – Wirt, Agens und Umwelt – haben zudem spezifische risikomodifizierende Faktoren:

| Merkmale des Wirts | Art des Agens | Umwelt-Faktoren |
|---|---|---|
| Alter | Biologisch | Temperatur |
| Geschlecht | z.B. Bakterien, Viren | Feuchtigkeit |
| Ethnie/Rasse | Chemisch | Bevölkerungsdichte |
| Religion | z.B. Giftstoffe, Alkohol, Rauch | Wohnverhältnisse |
| Gewohnheiten | | Luftverschmutzung |

| | | |
|---|---|---|
| Genetisches Profil | Physikalisch | Lärm |
| Familienstand | z.B. Unfall, Strahlung | Strahlung |
| Familiärer Hintergrund | Ernährung | |
| Immunitätslage | z.B. Mangelernährung | |

### 3.1.3. Arten der Übertragung

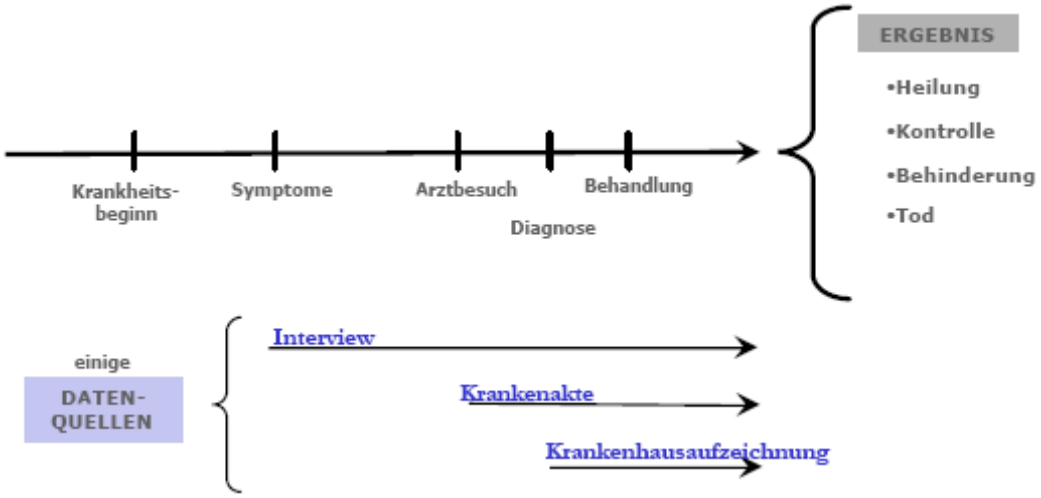
Schließlich muss noch zwischen verschiedenen Arten der Übertragung unterschieden werden:

- Direkt (Mensch-zu-Mensch-Kontakt)
- Indirekt (Gemeinsame Quelle, z.B. Themse, oder Vektor)

## 3.2. Epidemiologische Maßzahlen

### 3.2.1. Natürlicher Krankheitsverlauf

Zunächst soll kurz ein prototypischer Krankheitsverlauf geschildert werden, um zu zeigen, auf welcher Basis epidemiologische Maßzahlen berechnet werden.




Nach dem Krankheitsbeginn dauert es einige Zeit, bis sich die ersten Symptome zeigen. Ab diesem Zeitpunkt können Studien einsetzen. Protokolliert werden diese Symptome schließlich beim ersten Arztbesuch. Aus diesem folgt eine Diagnose sowie die daran anschließende Behandlung. Schließlich stellt sich das Ergebnis ein.

Als Datenquellen können also sowohl Interviewdaten (retrospektiv), Krankenakte des Arztes oder eventuelle Krankenhausaufzeichnungen fungieren. Auf Basis dieser Daten lassen sich nun die verschiedenen Maßzahlen der Epidemiologie berechnen.

### 3.2.2. Prävalenz

Die Prävalenz gibt den Anteil (in Prozent: Proportion) Erkrankter zu einem bestimmten Zeitpunkt (Punktprävalenz) oder eines gewissen Zeitraumes an (Periodenprävalenz). Sie besteht also aus den beiden Komponenten „Bestehende Erkrankung“ und „Bezugsbevölkerung“.

Die Prävalenz kann also als Wahrscheinlichkeit definiert werden, dass eine zufällig ausgewählte Person an einem Stichtag erkrankt ist.

Für diesen Stichtag gilt:

$$P = \frac{Anzahl\_Erkrankter}{Größe\_der\_Population} \leftarrow \text{[Bezugsgröße]}; \qquad W = [0;1]$$

Anmerkungen:

- Bei der Periodenprävalenz werden mehrmals Erkrankte als ein Fall gezählt. Typische Periodenprävalenzen beziehen sich auf 1, 3 oder 12 Monate. Ein weiteres wichtiges Konzept ist die Lebenszeitprävalenz.
- Bei beiden: Es werden sowohl Neuerkrankte als auch bereits Erkrankte gezählt.

## 3.2.3. Inzidenz

Die Inzidenz gibt das Risiko an, in einem bestimmten Zeitintervall zu erkranken. Zur Berechnung werden also die Anzahl der Neuerkrankungen und die Anzahl gesunder Fälle zu Anfang der Beobachtung in Beziehung gesetzt. Sie kann also ebenfalls als Proportion (Prozentzahl) ausgedrückt werden.

Die Inzidenz besteht also immer aus drei Komponenten: Neuerkrankungen, Bezugsbevölkerung und Beobachtungszeit.

Zur Quantifizierung werden zwei verschiedene Inzidenzmaße verwendet:

- Kumulative Inzidenz
- Inzidenzrate

### 3.2.3.1. Kumulative Inzidenz (-rate) CI

Es werden nur Personen berücksichtigt, die zu Beginn des Zeitraums nicht erkrankt sind, aber potentiell erkranken können (also z.B. keine Männer, wenn es um Gebärmutterhalskrebs geht).

$$CI = \frac{Anzahl\_Neuerkrankungen}{Größe\_der\_Population}$$

Die Bezugsbevölkerung bleibt während einer Studie unverändert (Anzahl der Gesunden zu Beginn der Untersuchung; „geschlossene Kohorte“).

Zu beachten ist hierbei, dass die Höhe der Inzidenz immer abhängig vom betrachteten Zeitraum ist. Die Mortalität (= $CI_{Tod}$) von Neugeborenen in 160 Jahren beträgt 1.

Der Wertebereich der Kumulativen Inzidenz ist $W = [0;1]$ (vgl. Prävalenz)

### 3.2.3.2. Inzidenzdichte ID

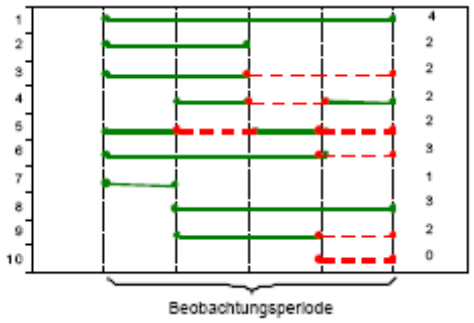
Die Inzidenzdichte bezieht sich nicht auf die Anfangsbevölkerung sondern auf die Personenzeit, also die Zeit die ein Proband in der Studie verbringt. Sie beginnt mit dem Eintritt in die Studie („offene / dynamische Kohorte“). Eine Person kann also bei mehrfacher Erkrankung auch mehrfach gezählt werden. Personen die von Anfang an krank waren werden nicht gezählt.

$$ID = \frac{Anzahl\_Neuerkrankungen}{Personenzeit}$$

Die Personenzeit läuft entweder bis zum Verlassen der Studie (Drop-Out) oder bis zum Auftreten der interessierenden Erkrankung. Nachdem die Person wieder genesen ist, startet die Personenzeit erneut.

Der Wertebereich der Inzidenzdichte liegt damit abweichend von den anderen Maßzahlen bei $W = [0;\infty]$. Natürlich gilt auch hier, dass jeder Proband prinzipiell erkranken können muss.

Beispiel:




Personenzeit:

4 + 2 + 2 + 2 + 2 + 3 +1 + 3 + 2 = 21

Anzahl der Neuerkrankungen: 7

$$ID = \frac{7}{21} = 0.33$$

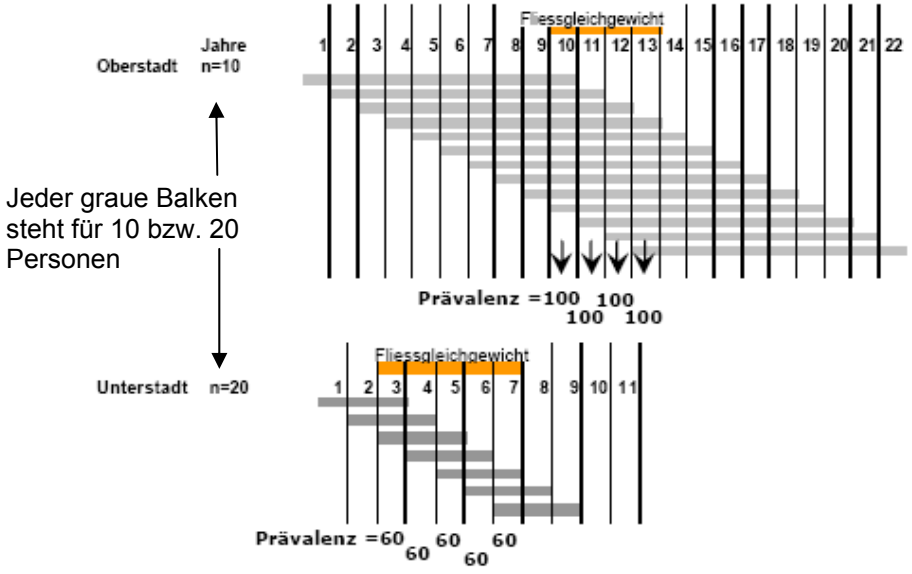
## 3.2.4. Zusammenhang von Prävalenz und Inzidenz

Bei einem Fließgleichgewicht – Zu- = Abwanderung – gilt zumindest für seltene Krankheiten:

$$Prävalenz = Inzidenzdichte \cdot Krankheitsdauer$$

Beispiel zu den Daten aus der Untersuchung zum Wasser der Themse:

| Untersuchte Bevölkerung | Punktprävalenz pro 1000 EW | Inzidenz (pro Jahr) | Dauer (Jahre) |
|---|---|---|---|
| Oberstadt | 100 | 10 | 10 |
| Unterstadt | 60 | 20 | 3 |




Das etwas seltsame Ergebnis – Oberstädter haben zu einer höheren Wahrscheinlichkeit Tuberkulose (TBC) – lässt sich einfach dadurch erklären, dass die Unterstädter früher sterben.

Krankheitsrate (Prävalenz) und Erkrankungsrisiko (Inzidenz) sind also zwei Paar Schuhe. Beide Maße sind dabei immer stark von Reliabilität und Validität der diagnostischen Tests abhängig (s.u.).

## 3.2.5. Weitere Maßzahlen

### 3.2.5.1. Mortalität (Sterblichkeit)

Die Mortalitätsrate gibt an, wie viele Todesfälle in der definierten Population in einem definierten Zeitraum auftreten. Es gibt auch altersspezifische bzw. krankheitsspezifische Mortalitätsraten.

$$Mortalitätsrate = \frac{Gesamtzahl\_der\_Todesfälle\_in\_definiertem\_Zeitraum}{Mittlere\_Populationszahl\_im\_definierten\_Zeitraum}$$

Beim Vergleich von Mortalitätsraten muss beachtet werden, dass die Bezugspopulationen vergleichbar sind (z.B. hinsichtlich der Altersstruktur). Mögliche Erklärungen für Veränderungen in der Mortalität können sowohl auf reale Effekte (z.B. Änderung der Inzidenz) oder auch Artefakte (veränderte Kodierung) hervorgerufen werden.

### 3.2.5.2. Letalitätsrate (Tödlichkeit)

Während die Mortalität ein Maß für die Anzahl der Todesfälle im generellen ist, ist die Letalitätsrate ein Maß für die Schwere einer Krankheit.

$$Letalitätsrate = \frac{Zahl\_der\_an\_Krankheit\_X\_Verstorbenen}{Zahl\_der\_an\_X\_Erkrankten}$$

Auch die Letalität muss sich immer auf einen bestimmten Zeitraum beziehen. Ihr Wertebereich beträgt wie der der Mortalität $W = [0;1]$.

# 3.3.Schätzung der Validität diagnostischer Tests

Die statistischen Maße der Sensitivität und Spezifität haben in der Epidemiologie als Maßzahlen für die Validität diagnostischer Tests eine besondere Bedeutung.

- **Sensitivität SE:** Wahrscheinlichkeit dafür, dass ein Erkrankter auch als krank klassifiziert wird.
- **Spezifität SP:** Wahrscheinlichkeit dafür, dass ein Gesunder auch als Gesund klassifiziert wird.

Beide Maße lassen sich auf Basis einer Vierfeldertafel berechnen:

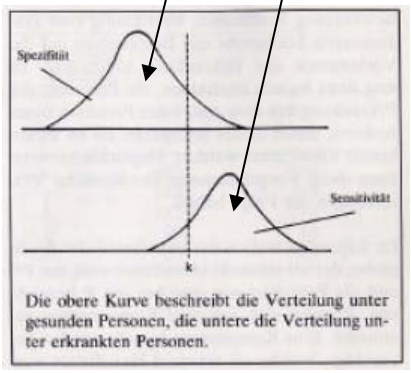
| | | Kriterium | | |
|---|---|---|---|---|
| | | Krank | Gesund | |
| Test | Krank | a | b | a + b |
| | Gesund | c | d | c + d |
| | | a + c | b + d | |

$$SE = \frac{a}{a+c}$$

$$SP = \frac{d}{b+d}$$

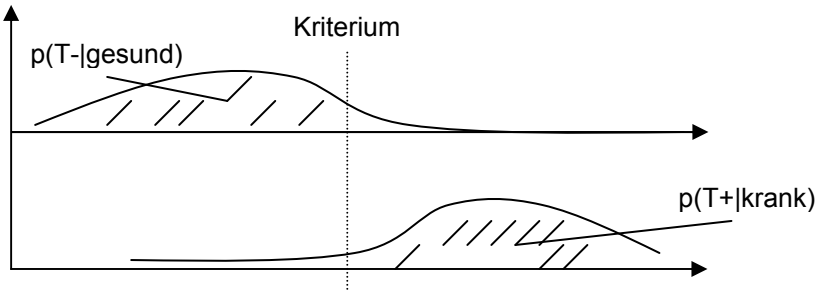
Die Werte verhalten sich im Regelfall gegenläufig, da die Verteilungen der Symptome unter Gesunden und Kranken immer überlappen. Mit dem Setzen eines Kriteriums ist also immer ein Trade-Off zwischen SE und SP verbunden.

Welcher der beiden Kennwerte bedeutsamer ist, muss im Einzelfall entschieden werden.



Die obere Kurve beschreibt die Verteilung unter gesunden Personen, die untere die Verteilung unter erkrankten Personen.

Eigene Grafik:




Auf Basis der obigen Vierfeldertafel lassen sich noch zwei weitere Kennwerte berechnen:

- **Positiver Vorhersagewert PW:** p(krank|T+). $PW = \frac{a}{a+b}$
- **Negativer Vorhersagewert NW:** p(gesund|T-). $NW = \frac{d}{c+d}$

Zu Beachten ist hierbei jedoch immer, dass PW und NW stak an Aussagefähigkeit verlieren, wenn die Prävalenz sehr niedrig ist. Auch bei hohen Werten für SE und SP.

Bei niedrigen Prävalenzen gilt:

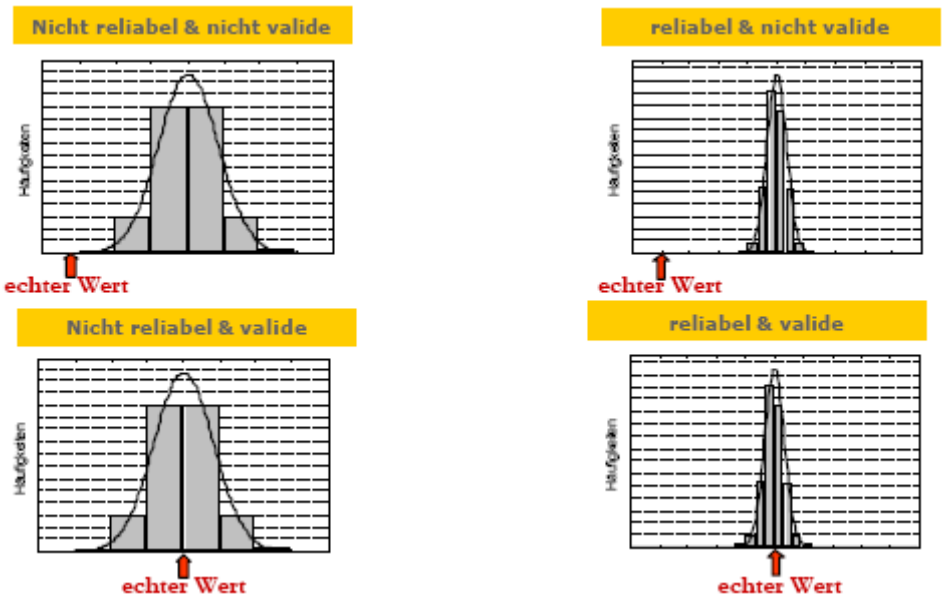
- Maximierung des NW: Sensitivität erhöhen
- Maximierung des PW: Spezifität erhöhen

Tests sollten daher immer an Bevölkerungsgruppen durchgeführt werden, bei denen die Prävalenz auch relativ hoch ist, also bspw. spezifische Risikofaktoren vorhanden sind.

Bei niedrigen Prävalenzen (Normalfall) bringt eine Steigerung der Spezifität mehr als eine Steigerung der Sensitivität.

# 3.4. Schätzung der Reliabilität diagnostischer Tests

→ Beobachterübereinstimmung.



# 3.5. Typen epidemiologischer Studien

Im Folgenden sollen die drei vorherrschenden Designs der Epidemiologie kurz umrissen werden: Kohortenstudien (Längsschnittstudie), Fall-Kontroll-Studie und Querschnittsstudie.

## 3.5.1. Kohortenstudie

In der Kohortenstudie werden zwei (nicht-zufallsgesteuerte) Stichproben verglichen, von denen eine exponiert ist und die andere nicht, z.B. Raucher und Nicht-Raucher.

Auch wenn es sich nicht um Zufallsstichproben handelt, sollten die Stichproben selbstverständlich in möglichst vielen Merkmalen vergleichbar sein.

Wichtig: Zu Beginn der Studie sind alle Pbn gesund. Erkrankte und Gesunde ergeben sich erst mit der Zeit. Die Kohortenstudie untersucht also die Inzidenz.

Vorteile:

- Inzidenz kann genau untersucht werden
- Entwicklungsgeschichte und Prozesse beschreibbar

Nachteile:

- Aufwendig und teuer
- Zeitintensiv (Induktionszeit bis zur Wirkung der Exposition)
- Drop-outs
- Nicht geeignet für seltene Erkrankungen

## 3.5.2. Fall-Kontroll-Studie

In der Fall-Kontroll-Studie (case-control-study) sind zum Zeitpunkt der Studie Fälle und Kontrollen gegeben. Dabei wird die Prävalenz der Exposition in den beiden Gruppen verglichen.

Vorteile:

- Billig und schnell
- Viele Expositionsfaktoren untersuchbar
- Auch für seltene Krankheiten geeignet
- Kein Drop-out

Nachteile:

- Schwierigkeit bei der Auswahl der Kontrollen (Wer ist vergleichbar?): Sollen bspw. Magenkrebs-Patienten in einem Spezialklinikum für gastrointestinale Erkrankungen untersucht werden, müsste eine adäquate Kontrollgruppe am besten aus demselben Krankenhaus kommen. Diese Personen sind aber ebenfalls von gastrointestinalen Erkrankungen betroffen, sodass sie mit hoher Wahrscheinlichkeit weniger Alkohol und Kaffee konsumieren als Normalbürger. Dieser Risikofaktor wird also durch eine solche Untersuchung systematisch unterschätzt.
- Probleme bei Erfassung der Exposition (Recall-Bias): Fälle haben vielleicht schon früh Symptome gezeigt und somit vermehrt auf die Exposition geachtet.
- Keine Angaben zu Inzidenzraten und absolutem Risiko möglich

### 3.5.3. Querschnittsstudie

In einer Querschnittsstudie (cross-sectional-study) wird die Prävalenz der Erkrankung in Populationen mit unterschiedlicher Exposition verglichen.

Vorteile:

- Kurzfristig und billig
- Für seltene Krankheiten geeignet
- Exposition zuverlässig erfassbar
- Vergleich verschiedener Populationen möglich

Nachteile:

- Problematisch bei Krankheiten mit kurzer Dauer
- Nur korrelative Interpretation möglich
- Keine Angaben zur Inzidenz möglich

## 3.6. Vergleichende Maßzahlen – Bestimmung von Risiken

Die Bedeutung von protektiven und Risikofaktoren für eine Krankheit lässt sich über zwei verschiedene Maße quantifizieren:

- Chance einer Krankheit $Chance = \frac{Anzahl\_Kranke}{Anzahl\_Gesunde}$
- Risiko einer Krankheit $Risiko = \frac{Anzahl\_Kranke}{Anzahl\_Kranke + Anzahl\_Geunde}$

Die Chance einer Krankheit wird als reelle Zahl definiert, das Risiko wird als Prozentzahl angegeben. Während das relative Risiko nur in Kohortenstudien berechenbar ist, muss dieses Maß in Fall-Kontroll-Studien über die Chance geschätzt werden.

| | EXPOSITION | | |
|---|---|---|---|
| ERKRANKUNG | exponiert | nicht | |
| krank | a | b | a+b |
| gesund | c | d | c+d |
| | a+c | b+d | a+b+c+d |

### 3.6.1. Risikomaß: Kohortenstudie (RR)

$$Risiko = \frac{Anzahl\_Kranke}{Anzahl\_Kranke + Anzahl\_Geunde}$$

Da bei Kohortenstudien zunächst alle Pbn gesund sind, werden in den Zähler des Bruchs die Anzahl die Neuerkrankungen eingetragen. Das Risiko entspricht also der Inzidenz, die durch die Kohortenstudie gefunden wird.

Zentral ist das Konzept des relativen Risikos, das Inzidenz bei Exposition mit Inzidenz ohne Exposition vergleicht:

$$RR = \frac{Risiko \mid Exposition}{Risiko \mid Keine\_Expoistion} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}}$$

## 3.6.2. Risikomaß: Fall-Kontroll-Studie (OR)

Im Gegensatz zur Kohortenstudie geht eine Fall-Kontrollstudie von dem Fakt der Erkrankung aus.

In einer Fall-Kontroll-Studie wird das Verhältnis der Fälle (a+b) zu den Kontrolle (c+d) willkürlich festgelegt. Daher darf kein relatives Risiko berechnet werden, sodass das Odds Ratio (OR) als Chance einer Krankheit berechnet wird:

$$OR = \frac{Chance \mid Exposition}{Chance \mid Keine\_Exposition} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a \cdot d}{b \cdot c}$$

Das Odds Ratio gibt also das Verhältnis von Kranken und Gesunden bei Exposition bzw. ohne Exposition aus – die Chance einer Krankheit. Ein Odds Ratio von 3 bedeutet also, dass die Chance an einer Krankheit zu erkranken bei Exposition 3x so hoch ist wie ohne Exposition.

Anmerkung: Das Konzept des OR hängt eng mit der logistischen Regression (s. 4.) zusammen. Das Ergebnis dieses Verfahrens ist ein OR und muss auch dementsprechend als Chance und nicht als Risiko interpretiert werden.

## 3.6.3. RR und OR

Das Odds Ratio ist unabhängig vom Verhältnis von Fällen und Kontrollen. Je kleiner die Prävalenz einer Krankheit i (a+c ≈ c und b+d ≈ d), desto genauer schätzt das OR jedoch das relative Risiko:

$$RR = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} \approx \frac{\frac{a}{c}}{\frac{b}{d}} = OR$$

Für beide Maße gilt: $W = [0;\infty[$ . Für RR/OR

- < 1: protektiver Faktor
- > 1: risikoerhöhender Faktor.

Interpretieren lassen sich die beiden Maße etwa folgendermaßen:

- RR = 10: „Durch die Exposition wird das Erkrankungsrisiko um das 10-fache erhöht."
- OR = 10: „Durch die Exposition wird die Chance zu erkranken um das 10-fache erhöht." Oder „Durch die Exposition wird das Verhältnis von Kranken zu Gesunden um das 10-fache erhöht."

Vorteile:

- RR: Direkte Beschreibung der Risikoerhöhung durch die Exposition. Das Risiko ist als relative Häufigkeit (%) gut interpretierbar.
- OR: Liefert auch bei Fall-Kontroll-Studien eine zuverlässige Schätzung des durch die Exposition bedingten Krankheitsrisikos. V.a. bei seltenen Krankheiten wird das RR gut geschätzt. Das OR findet zudem in der logistischen Regression Anwendung.

Nachteile:

- RR: Bei Fall-Kontroll-Studien nicht anwendbar
- OR: Nur eine *Schätzung* des Relativen Risikos

### 3.6.4. Confounding

Ist ein nicht erfasstes Merkmal sowohl mit der Krankheitshäufigkeit als auch mit der Expositionshäufigkeit korreliert (Confounding), kommt es zu gravierenden Fehleinschätzungen des Krankheitsrisikos, z.B. im Sinne des Simpsonschen Paradoxons.

Es müssen also alle relevanten Einflussfaktoren kontrolliert und als Prädiktoren ins Modell aufgenommen werden.

### 3.6.5. Attributables Risiko

Das Attributable Risiko beschreibt den Prozentanteil der Inzidenz, der einem bestimmten Faktor zugeschrieben werden kann. Das Attributable Risiko kann sowohl für die Gruppe der Exponierten als auch für die Gesamtpopulation berechnet werden.

Attributables Risiko in der Gruppe der Exponierten:

$$\frac{Inzidenz(Expositionsgruppe) - Inzidenz(Nicht - Expositionsgruppe)}{Inzidenz(Expositionsgruppe)}$$

Dieser Wert beschreibt die maximal erreichbare Senkung des Erkrankungsrisikos, wenn die Exposition vollkommen verhindert wird (Präventionspotential).

Attributables Risiko in der Gesamtbevölkerung:

$$\frac{Inzidenz(Gesamtpopulation) - Inzidenz(Nicht - Expositionsgruppe)}{Inzidenz(Gesamtpopulation)}$$

Dieser Wert beschreibt den Anteil der Inzidenz einer Erkrankung in der Gesamtbevölkerung der der Exposition zugeschrieben werden kann.

# 4. Logistische Regression

Technische Anmerkung: Die zweite angegebene Literatur,

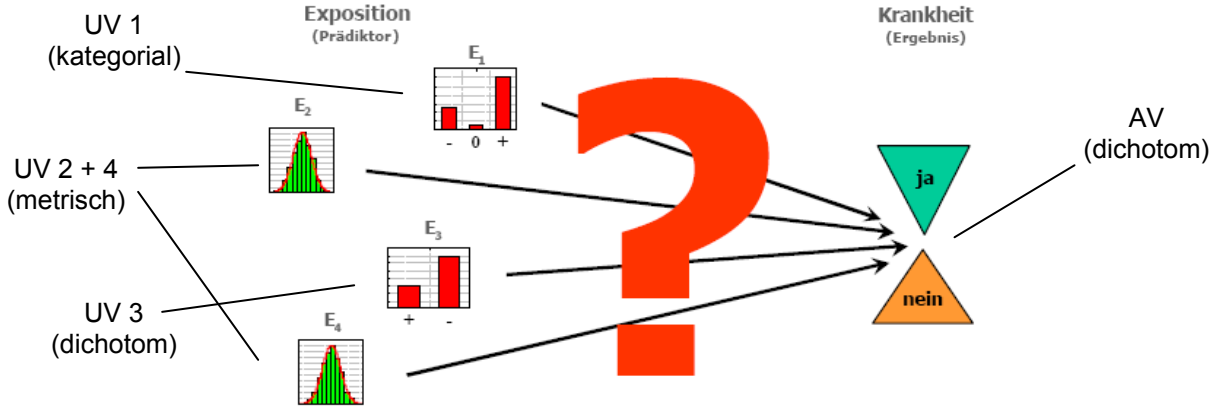
> Diaz-Bone, R. & Künemund, H. (2003*). Einführung in die binäre logistische Regression* (Heft Nr. 56). Berlin: Mitteilungen aus dem Schwerpunktbereich Methodenlehre,

findet sich auch im Internet auf den Seiten des Berliner Instituts (heruntergeladen).

## 4.1. Grundgedanke

Die logistische Regression sollte besser als binäre logistische Regression bezeichnet werden, da sie eingesetzt wird, um eine dichotome AV vorherzusagen. Ein häufiges Anwendungsgebiet ist dabei die Epidemiologie (AV: krank vs. gesund).

Die UVn (Prädiktoren) können dabei nominalskaliert, ordinal oder metrisch sein. Es können eine oder mehrere UVn verwendet werden.




Das Ziel der logistischen Regression ist also die statistische Beurteilung des Zusammenhangs zwischen einer nominalskalierten, dichotomen AV und mindestens einer UV. Daraus ergeben sich folgende Fragestellungen:

- Wovon hängt das Eintreten eines Ereignisses ab, bzw. wodurch wird es beeinflusst?
- Wie stark beeinflussen welche Prädiktoren die AV in welche Richtung (Gewichtungsfaktoren)?
- Vorhersage für den Einzelfall auf Basis der Kenntnis der Ausprägung der UVn.

Ausgangspunkt ist hierbei die (multiple) lineare Regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i$$

mit $x_{ij}$ = Ausprägung der UVn j bei Pb i, $\beta_0$ = Absolutglied, $\beta_i$ = Gewichtungsfaktor (Koeffizient), $u_i$ = Residuum bei Pb i und $y_i$ = Vorhergesagte Ausprägung der AV bei Pb i.

Bei der vorliegenden Datenlage ist die klassische Regressionsanalyse jedoch nicht durchführbar, da diese immer mit einem metrischen Kriterium arbeitet. Hier soll jedoch ein dichotomes Kriterium vorhergesagt werde; die Voraussetzungen der klassischen Regression sind also verletzt.

Die Verletzung der Modellspezifikation ist v.a. problematisch, da die Residuen weder normalverteilt (W = [0;1]), noch varianzhomogen sind.

# 4.2. Logistische Funktion und Koeffizienten

## 4.2.1. Herleitung

Um eine Regression mit einem dichotomen Kriterium durchführen zu können müssen also einige Zusatzschritte eingeführt werden. Ziel dieser Schritte ist die Entwicklung einer Gleichung, die auf der rechten Seite den bekannten Ausdruck $\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + u_i$ enthält sowie auf der linken Seite einen Ausdruck mit dem Wertebereich $]-\infty;+\infty[$, der eine Dichotomisierung erlaubt.

**Schritt 1:** Statt der Gruppenzugehörigkeit (0/1) wird die beobachtbare Wahrscheinlichkeit der Zugehörigkeit $p(y_i=1)$ betrachtet. → Stetiges Merkmal

**Schritt 2:** Nicht $p(y_i=1)$ wird betrachtet, sondern das Chancenverhältnis (Odds) mit Odds = $\frac{p(y_i=1)}{1-p(y_i=1)}$. → W = [0;∞[

**Schritt 3:** Logarithmierung; *log*(Odd) wird auch als *Logit* $p(y_i=1)$ bezeichnet.

$$Logit\_p(y_i=1) = \ln\left(\frac{p(y_i=1)}{1-p(y_i=1)}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} \left(+u_i\right)$$

**AV** **UVn**

Hierdurch erhält man den gewünschten Wertebereich. Beachte: $\ln\left(\frac{0,5}{1-0,5}\right) = 0$

Wenn $x_k$ um eine Einheit zunimmt, steigt Logit p(y=1) um $\beta_k$ an.

## 4.2.2. β-Koeffizienten

Das Ergebnis der logistischen Regression sind also die Logitkoeffizienten β (Regressionskoeffizienten), die mit dem ML-Verfahren geschätzt werden. Das ML-Verfahren ist eine Funktion der Parameter, gegeben die Daten: Die Auftretenswahrscheinlichkeit der gefundenen Daten soll maximiert werden.

Wenn keine Exposition erfolgt (alle $x_i = 0$), dann ist Logit $p(y_i=1) = \beta_0$. $\beta_0$ wird auch als **background log odd** bezeichnet. $\beta_0$ kann nur in Kohortenstudien interpretiert werden, in Fall-Kontroll- und Querschnittsstudien müsste das wahre Verhältnis von Fällen und Kontrollen berücksichtigt werden. Auch bei bestimmten Prädiktoren ist die Interpretation des background log odd erschwert: W'keit für „Krank" wenn Alter = 0?

Die erhaltenen β-Gewichte geben an, welche Veränderung des Logits durch eine Veränderung der Prädiktoren hervorgerufen wird. Zur inhaltlichen Interpretation lassen sich die Koeffizienten in Odds-Ratios umwandeln.

## 4.2.3. Logistische Funktion

Die Logit-Gleichung lässt sich über einige Zwischenschritte in die logistische Funktion umformen:

$$p(y_i=1) = \frac{1}{1+e^{-z}} \quad \text{mit } z = (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}) \quad \text{oder}$$

$$f(z) = \frac{1}{1+e^{-z}} \quad \rightarrow \textbf{Logistische Funktion}$$

Wichtig: z ist also ein „Index“, der alle Einflüsse der unabhängigen Variablen kombiniert.

Anmerkung zur Herleitung auf Folie 17: Die etwas seltsame Umformung

$$\frac{p(y_i=1)}{1-p(y_i=1)} = e^z = \frac{1}{e^{-z}} = \frac{\frac{1}{1+e^{-z}}}{\frac{e^{-z}}{1+e^{-z}}} = \frac{\frac{1}{1+e^{-z}}}{1-\frac{1}{1+e^{-z}}} \qquad \frac{\frac{a}{1+a}}{\frac{1+a}{1+a}-\frac{1}{1+a}} = \frac{\frac{1+a-1}{1+a}}{1-\frac{1}{1+a}}$$

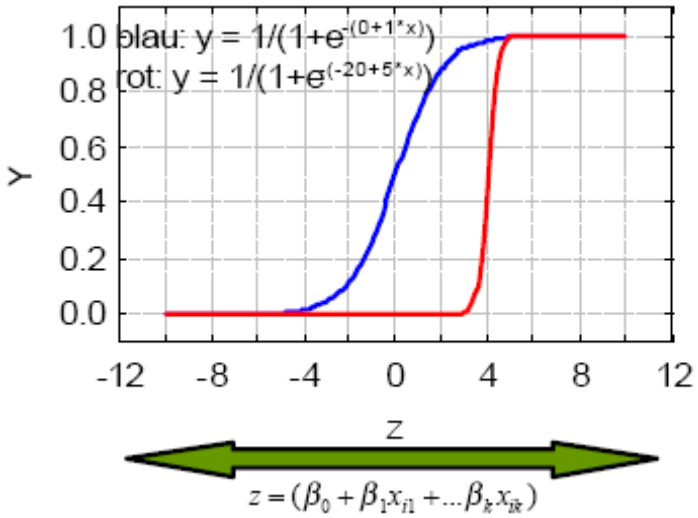
resultiert in einer Beziehung, die auf beiden Seiten des Gleichheitszeichens dieselbe Struktur aufweist:

$$\frac{x}{1-x} = \ldots = \frac{y}{1-y}.$$

Daraus lässt sich schlussfolgern: x = y bzw. $p(y_i=1) = 1/(1+e^{-z})$. Die der Wahrscheinlichkeit, ein Symptom zu zeigen kann damit berechnet werden. Über die logistische Funktion kann also tatsächlich berechnet werden, mit welcher Wahrscheinlichkeit das Kriterium bei einem einzelnen Individuum auftritt, wenn die Ausprägung der Risikofaktoren bekannt ist.

Diese Berechnung ist jedoch nur möglich, wenn $\beta_0$ interpretierbar ist – es muss also durch eine Kohortenstudie geschätzt worden sein.

Insgesamt lässt sich zur logistischen Funktion folgendes festhalten:




Die logistische Funktion beschreibt einen nichtlinearen Zusammenhang zwischen UVn und dichotomer AV. Das background log odd $\beta_0$ bestimmt die Lage in der Horizontalen, während die restlichen Gewichte $\beta_k$ die Steilheit bestimmen.

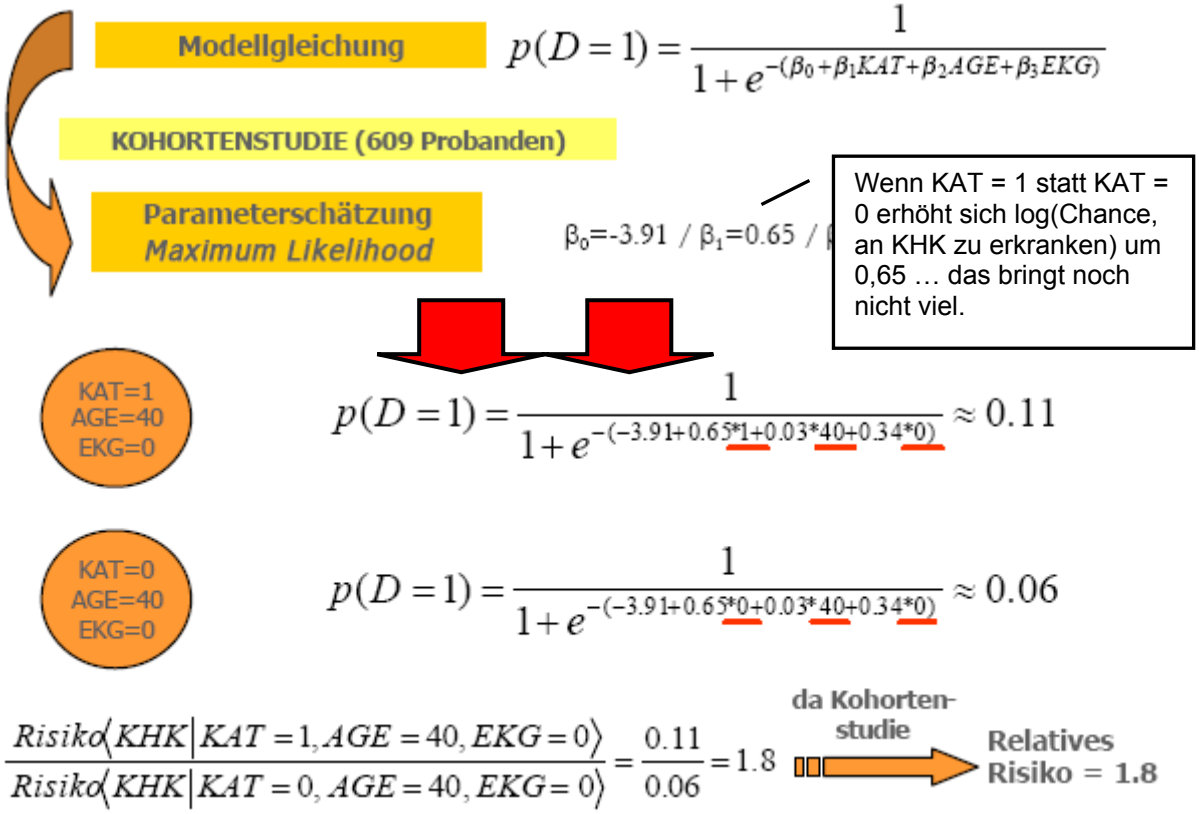
Sie zeigt einen s-förmigen Verlauf, was gut zum theoretischen Hintergrund der epidemiologischen Anwendung passt. Man kann eine Menge Risikofaktoren haben, ohne dass etwas passiert. Ist man jedoch fettleibig, bewegt sich wenig und fängt dann auch noch das Rauchen an zeigt sich plötzlich ein starker Anstieg der Erkrankungswahrscheinlichkeit für KHK („Schwellengedanke“). Andererseits kann das Risiko auch nicht ins Unendliche gesteigert werden („Sättigungs-Effekt“).

# 4.3. Anwendung

Die logistische Regression modelliert den Wahrscheinlichkeitsübergang einer kategorialen (hier: dichotomen) Variablen in Abhängigkeit von der Ausprägung der unabhängigen Variablen unter Annahme der logistischen Verteilung der Residuen.

Das Ergebnis ist zunächst die Schätzung der Gewichtungsfaktoren $\beta_k$. Über die logistische Funktion lässt sich jedoch auf Basis einer Kohortenstudie sowie unter Kenntnis der Ausprägung der betrachteten Risikofaktoren die Wahrscheinlichkeit (Risiko) berechnen, dass ein Kriterium bei einem Individuum auftritt:

- D=Koronare Herzkrankheit KHK (0=nein / 1=ja)
- KAT=Katecholaminlevel (0=niedrig / 1= hoch)
- AGE=Alter (kontinuierlich)
- EKG=EKG-Befund (0=unauffällig / 1=auffällig)




Aus dem Verhältnis der Risiken zweier Individuen ist also eine direkte Schätzung des Risikoverhältnisses, also des Relativen Risikos möglich. Diese direkte Einschätzung des Relativen Risikos ist allerdings nur möglich, wenn eine Kohortenstudie vorliegt und alle unabhängigen Variablen spezifiziert sind.

In der Regel wird in der Logistischen Regression jedoch das Odds Ratio verwendet, da seine Schätzung des Relativen Risikos für alle Studientypen und ohne Zusatzannahmen gilt.

## 4.3.1. Odds Ratio (Effektkoeffizient)

Das Odds Ratio gibt an, um wie viel höher die Chance zu erkranken bei Exponierten vs. nicht exponierten Personen ist: $OR_{X1,X0}=e^{\left(\sum_{i=1}^{k}\beta_i(X_{1i}-X_{0i})\right)}$.

Um das OR zu berechnen nimmt man die $\beta_i$ also einfach in den Exponenten der Eulerschen Zahl. Dabei kürzen sich natürlich sämtliche $\beta_i$ mit den gleichen Ausprägungen der jeweiligen UVn, sodass sich bspw. ergibt:

$e^{\beta_1}$ = OR für auffälligen KAT.

Interpretation: Die Erhöhung der UV i um eine Einheit bewirkt eine Veränderung der Chance um den Faktor $e^{\beta_i}$.

Der Wertebereich eines Odds Ratios liegt dabei im Intervall $[0;\infty[$, wobei ein OR < 1 für einen protektiven Faktor steht, OR > 1 für einen Risikofaktor. OR = 1 bedeutet, dass die Ausprägungen des Faktors völlig irrelevant sind.

Wichtig: Das OR ist eine Chance und darf somit niemals als Wahrscheinlichkeit interpretiert werden. Bei Kohortenstudien sind jedoch nicht nur die $\beta_i$ interpretierbar, sondern zusätzlich das $\beta_0$. Die Gewichte lassen sich somit in die logistische Funktion einsetzten, wodurch die Wahrscheinlichkeit berechnet werden kann (s. 4.2.3).

## 4.3.2. Anmerkungen

### 4.3.2.1. Kategoriale Prädiktoren

Kategoriale Variablen müssen in dichotome Variablen zerlegt werden, wobei eine Referenzkategorie bestimmt wird. Das Odds Ratio gibt das Chancenverhältnis zu der vorher bestimmten Referenzgruppe an.

### 4.3.2.2. Metrische Prädiktoren

Das OR für einen metrischen Prädiktor stellt die Veränderung der Chance bei Veränderung des Prädiktors um eine Einheit dar.

Bsp.: OR(Alter) = 1,02. Mit jedem Lebensjahr vervielfacht sich die Chance, das Kriterium „Wahlteilnahme“ zu zeigen um den Faktor 1,02. Für 10 Jahre ergibt sich eine Veränderung der Chance um den Faktor $1{,}02^{10}$ = 1,22, also einer Steigerung von 22%.

Für 10 Jahre jüngere Probanden ergibt sich folglich eine Veränderung der Chance um den Faktor $1{,}02^{-10}$.

### 4.3.2.3. Wechselwirkungen

Im Standardfall geht das Modell von unabhängigen Prädiktoren aus. In die Modellgleichung können jedoch auch Wechselwirkungen mit einbezogen werden:

$$Logit\_p(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

HWen WW

### 4.3.2.4. Prädiktorblöcke und Modellsuche

Es besteht die Möglichkeit der Verwendung mehrerer Prädiktorblöcke. Die sukzessive Modellerweiterung um Blöcke von Regressoren kann mit einem F-Test darauf geprüft werden, ob der zuletzt einbezogene Block eine signifikante Verbesserung der Vorhersage bewirkt hat.

Es gibt eine automatische Modellsuche in SPSS (vorwärts und rückwärts). Die Modellsuche sollte aber i.d.R. vom Untersucher vorgenommen werden.

### 4.3.3. Voraussetzungen

Für den binären Fall sollten mindestens 50, besser jedoch 100 Fälle in die Analyse aufgenommen werden. Bei ungleichen Zellverteilungen (krank / gesund) sollte die kleinste Zellbesetzung mindestens 25 betragen. Auch sollten umso mehr Personen untersucht werden, je mehr UVn verwendet werden.

Jede Prädiktorvariable sollte zudem einen eigenen Anteil zur Vorhersage leisten (das ist auf F. 31 mit unkorreliert gemeint). Wirkliche Unkorreliertheit ist dabei nicht nötig – in der Praxis ist alles bis r = .9 noch OK. R = .6-.7 ist völlig normal.

Auch müssen alle relevanten Regressoren im Modell enthalten sein, da sich sonst das Problem des Confounding zeigen kann (s. 3.6.4).

### 4.3.4. Vergleich zur Diskriminanzanalyse

Die logistische Regression ist sehr robust und hat weniger statistische Voraussetzungen als die Diskriminanzanalyse. Diese verlangt Intervallniveau und Normalverteilung aller Prädiktoren sowie eine homogene Varianz-Kovarianz-Matrix in allen Gruppen.

## 4.4. Beurteilung des Modells

Das verwendete bzw. erhaltene Modell kann anhand verschiedener Verfahren auf die Güte der Schätzung geprüft werden.

- **$X^2$-Test (Omnibus-Test):** Sollte für das Modell signifikant sein, da er die Verbesserung der Vorhersage im Vergleich zum Nullmodell ohne Prädiktoren (nur $\beta_0$) prüft.
- **Pseudo-$R^2$-Maße:** Ab $R^2 > .2$ liegt eine akzeptable Schätzung vor, ab $R^2 > .4$ kann die Schätzung als gut bezeichnet werden. Das **McFadden-$R^2$** kann Werte zwischen 0 und 1 annehmen, das **Cox and Snell $R^2$** kann den Maximalwert von 1 nicht erreichen. Interessant ist das **Nagelkerke-$R^2$**, da es wie der Determinationskoeffizient der linearen Regression interpretiert werden kann.
- **t-Test / Wald-Test:** Statistische Signifikanz der Effektkoeffizienten. Wenn dieser Test signifikant wird, dürfen die Effektkoeffizienten als Odds Ratios interpretiert werden.
- **Hosmer und Lemeshow:** Dieser $X^2$-Test gibt die Abweichung der vorhergesagten von den tatsächlichen Werten der AV aus (die Gruppen werden dabei anhand der vorhergesagten W'keiten gebildet). Dieser Wert sollte also nicht signifikant werden.
- **Klassifikationsmatrix:** Prozentuales Zutreffen der Klassifikation durch den Vergleich von vorhergesagten und tatsächlichen Gruppenzugehörigkeiten.

Die Relevanz des Modells ist über eine Kreuzvalidierung prüfbar. Zunächst wird das Modell an einer Analysestichprobe entwickelt und dann auf eine Validierungsstichprobe angewendet.

## 4.5. Anwendungsbeispiele

Anmerkung: Die Anwendungsbeispiele 1 und 2 wurden mit Statistica gerechnet und in der Vorlesung ausgelassen.

Anwendungsbeispiel 3 ist aus Diaz-Bone, R. & Künemund, H. (2003; vgl. 4.) entnommen und beschäftigen sich mit der Frage, ob politisches Desinteresse in höheren Altersgruppen stärker ausgeprägt ist.

Hinweis: Hier zeigt sich, dass es immer wichtig ist, die Kodierung der Variablen zu beachten. AV = 1 steht hier für Desinteresse. Dies ist v. a. für die Interpretation von Bedeutung.

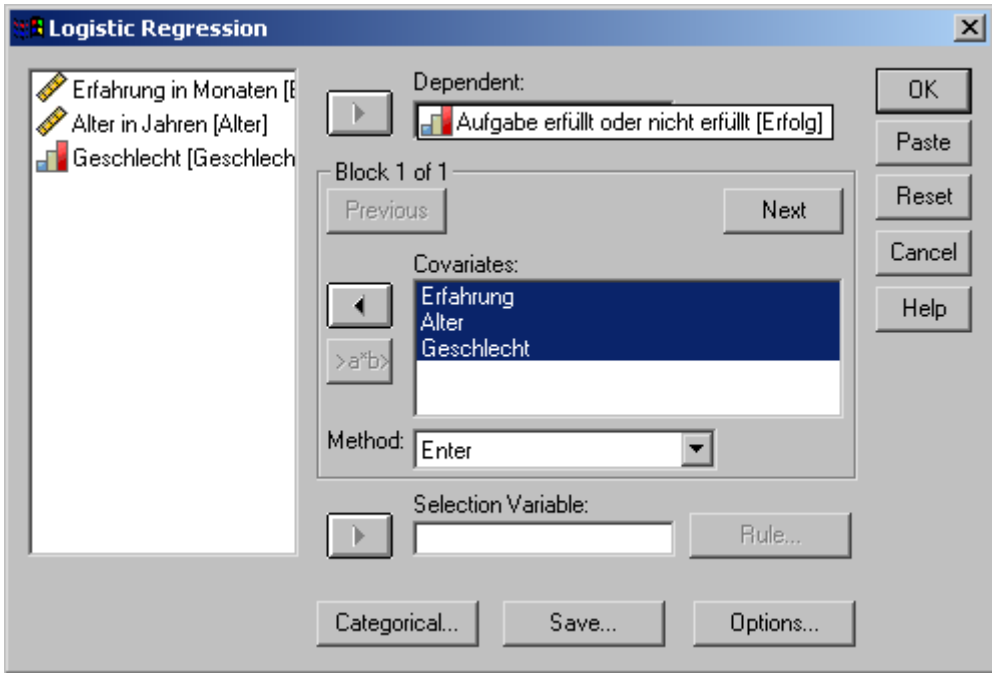
Zu Folie 46: Das Endergebnis der logistischen Regression steht selbstverständlich im letzten Schritt. Schritt 1: Alterseffekte wenn keine andere Variablen berücksichtigt werden. Hier hat Alter einen deutlichen Effekt. Dieser verschwindet jedoch im 4. Schritt, wenn zusätzlich noch Geschlecht als UV hinzugekommen ist.

Es trifft wohl die Alternativerklärung zu: Männer interessieren sich mehr für Politik als Frauen, aber Frauen werden älter. Daher zeigt sich ein niedrigeres Interesse für Politik (= höheres Desinteresse) in höheren Altersstufen (Confounding).

# 4.6. Übung

Die folgenden Angaben über das Vorgehen beziehen sich nicht mehr auf die Beispiele aus der Vorlesung sondern auf den Übungsdatensatz. Die binäre logistische Regression findet sich unter *Analysieren\Regression\Binär Logistisch.*

In SPSS werden die Variablen im Menü der binären logistischen Regression zunächst in die entsprechenden Zellen eingetragen:




## 4.6.1. Methode

Unter Methode kann das Vorgehen der logistischen Regression eingestellt werden. Einschluss (Enter) bedeutet, dass sofort alle Variablen verwendet werden.

Schrittweise vorwärts geht zunächst vom Nullmodell ohne Prädiktoren aus und nimmt die UVn schrittweise in das Modell mit auf. Rückwärts bedeutet, dass aus dem vollständigen Modell nach und nach irrelevante Prädiktoren entfernt werden.

## 4.6.2. Kategoriale Daten

Das Alter ist im vorliegenden Datensatz in 9 Kategorien unterteilt, sodass eine Umkodierung in 8 dichotome Variablen nötig ist aus der schließlich auch 8 ORs resultieren. Nachdem die Prädiktoren in SPSS in das Kovariatenfenster gezogen wurden muss also eingestellt werden, welche Prädiktoren metrisch und welche kategorial sind.

Wenn unter *Kategorial* bestimmt wurde, welche Referenzkategorie verwendet werden soll, muss zusätzlich auf *Ändern* geklickt werden, da bei *OK* die Änderungen nicht automatisch übernommen werden. Standardmäßig sind in diesem Fenster zusätzlich Indikatorkontraste eingestellt. Die verschiedenen Kontrastarten sind in der Hilfe beschrieben.

## 4.6.3. Speichern

Die vorhergesagte Wahrscheinlichkeit lässt sich für jeden Fall speichern. Es empfiehlt sich zudem, unter Einfluss *Cook* zu aktivieren sowie standardisierte Residuen ($\mu = 0$, $\sigma = 1$) zu speichern.

Bei nicht standardisierten Residuen wird einfach die Abweichung von vorhergesagter und tatsächlicher Wahrscheinlichkeit ausgegeben. Wurde bspw. eine Wahrscheinlichkeit von $p = 0.8$ vorhergesagt und das Ergebnis tritt tatsächlich auf ($p = 1$), so wird ein Residuum von .2 ausgegeben.

Wichtig: Extrem abweichende Fälle (gesunder 4-Zentner-Mann, der raucht, säuft und sich nie bewegt) können einen starken Einfluss auf die gesamte Parameterschätzung haben. Dieser Einfluss wird z.B. durch *Cook* ausgegeben.

## 4.6.4. Output

Block 0: Anfangsblock (Beginning Block) beschreibt das Nullmodell. Hier ist vor allem die korrekte Vorhersage in % relevant, da diese durch die verwendeten Prädiktoren (Block 1) stark verbessert werden sollte.

[Lauter Nuller in Parameterkodierung: Referenzkategorie].

Die Bedeutung der ausgegebenen Teststatistiken lässt sich unter Beurteilung des Modell (s. 4.4) nachlesen.

Das Hauptergebnis ist die Tabelle *Variablen in der Gleichung* (*Variables in the Equation*). Diese Tabelle wird im Endeffekt auch interpretiert, wobei es immer wichtig ist, auf das Skalenniveau der Prädiktoren zu achten.

Für kategoriale Prädiktoren gibt die Spalte $e^{\beta}$ die Veränderung der Chance an, wenn statt der Referenzkategorie (in Klammern) die Kategorie der jeweiligen Zeile vorliegt.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1(a) | Erfahrung | ,172 | ,053 | 10,711 | 1 | ,001 | 1,187 |
| | Alter | -,002 | ,032 | ,006 | 1 | ,939 | ,998 |
| | Geschlecht(1) | 2,234 | ,823 | 7,367 | 1 | ,007 | 9,337 |
| | Constant | -4,091 | 1,800 | 5,164 | 1 | ,023 | ,017 |

a Variable(s) entered on step 1: Erfahrung, Alter, Geschlecht.

**Wichtig: Exp(B) = OR.**

Beispielhafte Interpretation für Geschlecht: Die Chance auf Programmiererfolg bei Frauen ist um das 9.337-fache höher als die Chance auf Programmiererfolg bei Männern (Referenzkategorie der Variable Geschlecht).

Beispielhafte Interpretation für Erfahrung (in Monaten gemessen): Mit jedem Monat Erfahrung steigt die Chance auf Erfolg um den Faktor 1.187 oder: Mit jedem Monat Erfahrung wird das Verhältnis von Erfolg zu Misserfolg um den Faktor 1.187 erhöht.

# 5. Metaanalyse

## 5.1. Übersicht

Traditionell wurden vor der Erfindung der Metaanalyse Review-Verfahren eingesetzt, um zumindest eine qualitative Integration größerer Datenmengen zu gewährleisten. Hierbei wurde auf rein narrativer Basis versucht, die Essentials eines Forschungsparadigmas herauszuarbeiten.

Die Metaanalyse hingegen dient der quantitativen Integration von Daten. Beide Verfahren sollen hier kurz umrissen werden.

### 5.1.1. Review

Wie bereits erwähnt stellt das Review einen längeren, ausführlichen und systematischen Übersichtsartikel dar („narrativ-qualitatives Review“). Ziel ist die Reduktion singulärer Ergebnisse aus Primärstudien auf wesentliche Gesamtaussagen zum aktuellen Forschungsstand.

Darin sind neben der reinen Zusammenfassung auch eine Gesamtbewertung sowie Aussagen zur Generalisierbarkeit der Befunde (auf Populationen, Situationen und Zeiträume) und somit eine Spezifizierung des empirischen Anwendungsbereichs enthalten.

Ein gutes Review sollte also das Studium der Originalliteratur erübrigen.

Es lassen sich jedoch einige methodische, inhaltliche und statistische Schwächen des Verfahrens anführen[9]. Die kursiv geschriebenen Punkte werden in der Metaanalyse zumindest größtenteils vermieden.

#### 5.1.1.1. Review: Schwächen

**Methodische Schwächen:**

- *Lediglich Einbezug einer Teilmenge relevanter Studien*
- Fehlen objektiver Maßstäbe bei Darstellung und Bewertung von Primärstudien sowie bei der Interpretation der Befundlage
- Bedeutsamkeit nicht-signifikanter Befunde für den Gesamteffekt wird unterschätzt
- *Die Methode der Zusammenfassung einzelner Primärstudien wird selten berichtet*
- *Lediglich eine grobe Klassifikation der Ergebnisse (z.B. einfaches Auszählen)*
- Ergebnisdarstellung oft unübersichtlich und fehlgeleitet

**Statistische Schwächen:**

- *Stichprobenfehler oder Fehler 1. Art werden nicht berücksichtigt*
- Umfangreichen Stichproben wird zu viel Gewicht beigemessen
- Statistische Absicherung zusammenfassender Aussagen fehlt
- Aussagen über Effektstärken und Interaktionseffekte sind nicht möglich

[9] „Ein qualitatives Verfahren kann am Krüger-Lehrstuhl nicht gelobt werden“. Die blauen Schwächen auf den Folien werden durch die Metaanalyse behoben.

**Inhaltliche Schwächen:**

- *Keine kritische Bewertung früherer Reviews*
- Immenser Informationsverlust
- Narratives Review nicht adäquat für Aufklärung von Widersprüchen

### 5.1.1.2. Review: Bewertung

Es lassen sich massive methodische, statistische und inhaltliche Schwächen ausmachen; „es sieht schlecht aus um dieses Verfahren". Allerdings ist auch immer zu beachten, dass die großen Kritiker („Ein Review ist mehr eine Kunst als eine Wissenschaft") zugleich Gurus der Metaanalyse sind.

Ein direkter Vergleich von Review und Metaanalyse wurde von Cooper und Rosenthal (1980) vorgenommen. Die Pbn sollten dabei 7 Studien per Review bzw. per Metaanalyse aggregieren, in denen objektiv ein Zusammenhang zwischen Geschlecht und Ausdauer bei der Aufgabenbewertung vorhanden war [was heißt hier objektiv?].

Die Metaanalyse kam hier weitaus häufiger zum korrekten Ergebnis, allerdings ist sie auch nicht die Lösung aller Probleme (nur 68% korrekt). Zudem zeigte sich kein Unterschied zwischen Dozenten und Diplomanden.

Anmerkung: Sehr problematisch sind nicht signifikante Studien. Schachter und Singer (1962) wurde z.B. laut Ingo nicht signifikant und trotzdem gibt's eine 2-Faktoren-Theorie. Prof. Janke (Vorgänger von Prof. Pauli) hat bspw. lange versucht die Studie zu replizieren, jedoch ohne Erfolg.

## 5.1.2. Metaanalyse

Unter einer Metaanalyse versteht man eine Analyse von Analysen mit dem Ziel einer zusammenfassenden Interpretation der Befunde.

Das Vorgehen bei einer Metaanalyse lässt sich dabei in verschiedene Schritte aufteilen. Die einzelnen Schritte sind hier als Rückkopplungsschleifen zu verstehen: Man kann z.B. die Fragestellung auch in einem der darauf folgenden Schritte abändern und die nachfolgenden Schritte erneut durchlaufen.

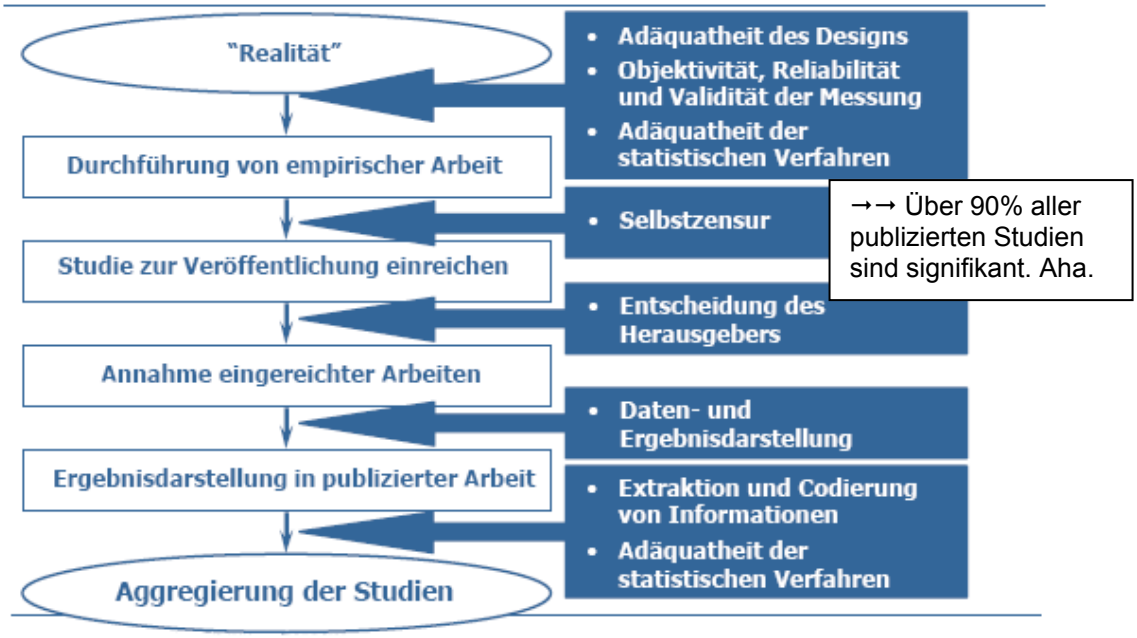
1.) Formulierung einer empirisch bereits geprüften Fragestellung. Dies ist also ein großer Unterschied bspw. zur Diplomarbeit. Dort wird eine Nische gesucht, die noch nicht beforscht wurde, wobei Literatur aus angrenzenden Bereichen herangezogen wird. Hier muss die Fragestellung natürlich bereits empirisch geprüft worden sein.
2.) Systematische Erfassung und Erhebung der empirischen Primärbefunde
3.) Kodierung und Bewertung der Studien hinsichtlich inhaltlicher und methodischer Merkmale
4.) Aggregation der quantitativ-summativen Befunde
5.) Interpretation der Ergebnisse
6.) Dokumentation und Präsentation

Das wichtigste Kennzeichnungsmerkmal der Metaanalyse ist die quantitative Herangehensweise. Diese umfasst:

- Effektstärkenorientierung bzw. Kombination von W'keitswerten
- Betonung methodischer Aspekte (Stichprobenfehler, Messfehler und andere Artefakte)
- Berücksichtigung von Moderatorvariablen

### 5.1.2.1. Anmerkung: Primärstudien

Es ist äußerst wichtig zu beachten, dass die verfügbaren Primärstudien erst nach einem Prozess veröffentlicht werden, der die Metaanalyse verfälschen kann.




### 5.1.2.2. Metaanalyse: Vorteile und Güte

Die methodischen Vorteile der Metaanalyse sind:

- Hohe statistische Validität durch hohe Power
- Hohe interne Validität durch gegenseitiges Ausgleichen methodischer Artefakte (z.B. unterschiedlicher Operationalisierung)
- Hohe Konstruktvalidität durch Realisierung unterschiedlicher Operationalisierungen
- Hohe externe Validität (heterogene Replikationen, unterschiedliche Personengruppen, Settings und Zeitpunkte)

Die Validität ist also generell höher als bei Primärstudien.

Merkregel: Insgesamt ist eine Metaanalyse umso erfolgsversprechender, a) je größer die Anzahl der enthaltenen Primärstudien ist, b) je mehr differenzierte und quantitative Ergebnisparameter (nicht nur Befragung!) verwendet wurden und c) je homogener das für die Fragestellung zentrale Konstrukt ist.

Der erste und dritte Punkt sind hierbei auf die Person bezogen, die die Metaanalyse durchführt, der zweite und dritte Punkt auf die Untersucher der Primärstudien.

Beispiel zu c): Eine Metaanalyse sollte also nicht generell für den Vergleich von Psychoanalyse und VT verwendet werden. Besser ist der Vergleich PA ↔ VT für Kurzzeittherapien bei Panikstörungen des Typs XY.

### 5.1.2.3. Metaanalyse: Probleme

Es werden 4 zentrale Probleme vorgestellt: Methodische Qualität der Primärstudien, Auswahl der Primärstudien, Dokumentation der Primärstudien sowie das Uniformitätsproblem.

### a) Methodische Qualität der Primärstudien

Rosenthal (1978) hat ein einer Studie die Autoren von 27 Primärstudien angeschrieben und sich die Datensätze schicken lassen. Eine Reanalyse ergab, dass im schlimmsten Fall jeder zweite Wert falsch berechnet war – und Fehler in 64% zugunsten der Forschungshypothese gemacht wurden.

Dabei zeigt sich auch ein starker Einfluss der Forscherloyalität (Elliot, 1996). Während Anhänger der Gesprächspsychotherapie eine Effektstärke von .63 ermittelten kamen VTler auf Basis der gleichen Daten zu einer ES von -.31 und neutrale Personen zu einer ES von -.02.

Das Problem der methodischen Qualität von Primärstudien lässt sich über den Ausschluss von Studien kontrollieren. Jedoch ist auch diese Option problematisch, weil verschiedene Kriterien herangezogen werden können:

- Inhaltlich: Primärstudien einschlägig? Mögliche Indikatoren sind Zitationsindizes. Diese sind aber konfundiert mit Publikationsdatum, Forschername, Zitierzirkelzugehörigkeit und Bereich (Verkehrspsychologen zitieren nicht viel). Dieses Kriterium ist also kritisch zu betrachten, wird jedoch trotzdem meistens herangezogen.
- Inhaltlich: Externe Validität erreicht mindestens den Grad der internen Validität. Dieses Kriterium lässt sich jedoch nicht überprüfen. Ingo: „Keine Idee. Ich habe keine Idee. Für diese Veranstaltung habe wich wirklich lange gesucht und habe nichts gefunden."
- Methodisch: Methodisches Vorgehen schließt Verzerrungen aus. Allerdings sind Störvariablen immer vorhanden, sodass Verzerrungen niemals ausgeschlossen werden. Kontrollierbar ist dieses Kriterium über eine Schätzung der Reliabilität und Objektivität. Hierfür existieren Daumenregeln nach Cook und Campbell (1979), nach denen man auch eine Checkliste entwerfen könnte.

Es gibt also drei klare Kriterien, die sich jedoch nur unzulänglich umsetzten lassen.

### b) Auswahl der Primärstudien

Die Aggregation von Einzelfalldaten sowie Einzelfall- + Gruppendaten ist grundsätzlich problematisch.

Jedoch ist auch bei reinen Gruppendaten die Repräsentativität durch den Publication Bias eingeschränkt. Darunter versteht man einen „seltsamen Zusammenhang zwischen Signifikanz einer Studie und der Wahrscheinlichkeit, veröffentlicht zu werden". Es lässt sich ein OR = 6.15 für die Veröffentlichung signifikanter Befunde schätzen.

Zusätzlich ist die Effektstärke veröffentlichter Befunde ca. um 1/3 höher als die Effektstärke nicht veröffentlichter Befunde. Wird diese Eigenheit nicht beachtet resultiert also eine Überschätzung der Mittleren Effekte.

### c) Dokumentation der Primärstudien

Häufig fehlen Angaben zu Studienmerkmalen (Stichprobe, Treatmentdurchführung, Ergebnisse). Derartige Studien sollten ausgeschlossen werden, was jedoch zu einem massiven Informationsverlust führt. Es bietet sich also an, die Autoren nach den Daten zu fragen bzw. fehlende Daten zu schätzen. Oder: Subgruppenanalyse auf unterschiedlichen Datenniveaus.

[Subgruppenanalyse: Optimale Informationen → Analyse mit gutem Verfahren, schlechte Daten → anderes Analysetool].

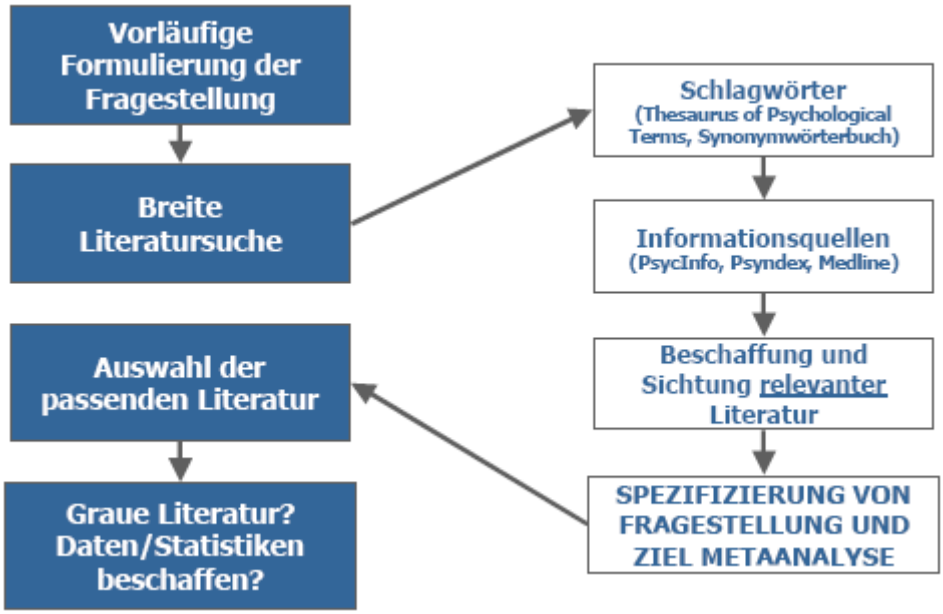
**d) Uniformitätsproblem (Äpfel-Birnen-Problem)**

Das Uniformitätsproblem umfasst Probleme bei der Explikation der Fragestellung sowie der Definition der in der Fragestellung enthaltenen Konstrukte und Variablen (UV, AV; s. Bortz & Döring, S. 675).

Bsp.: Klinik im Allgäu vs. Klinik in Potsdam (?).

# 5.2. Planung und Vorbereitung einer Metanalyse

Die Planung einer Metaanalyse lässt sich in vier Schritte unterteilen:




Diese Schritte sollen im Einzelnen näher betrachtet werden.

## 5.2.1. Vorläufige Formulierung der Fragestellung

Die Fragestellung wird immer erst vorläufig formuliert. Auf Basis der nachfolgenden Erhebung der Primärbefunde kann diese noch verändert werden.

## 5.2.2. Breite Literatursuche

Hinsichtlich der Schlagwörter (Erhebungsumfang) gibt es ein Trade-Off zwischen Vollständigkeitsquote und Genauigkeitsquote (Recall-Precision-Kurve). Es gilt jedoch beides möglichst hoch zu halten.

$$\text{Recall} = \frac{\text{Zahl der gefundenen relevanten Dokumente}}{\text{[Unbekannte] Gesamtzahl relevanter Dokumente}}$$

$$\text{Precision} = \frac{\text{Zahl der gefundenen relevanten Dokumente}}{\text{Gesamtzahl der gefundenen Dokumente}}$$

Thesaurus erlaubt hierbei eine Suche nach verwandten Begriffen und ist daher für die Generierung von Schlagwörtern äußerst hilfreich.

Die Auswahl der Quelle (PsycInfo, Psyndex, Medline…) sollte sich nach dem Gegenstand der Fragestellung richten. Sollen Studien zu einem Bereich der

Architekturpsychologie analysiert werden, ist eine Architekturdatenbank evtl. Medline vorzuziehen.

Ist die Beschaffung und Sichtung relevanter Literatur abgeschlossen, so können Fragestellung und Ziel der Metaanalyse spezifiziert werden.

## 5.2.3. Auswahl der passenden Literatur

Nach Auswahl und Sichtung der Literatur können die Studien mit Hinblick auf die Fragestellung ausgewählt werden. Hier sollten einige Merkmale berücksichtigt werden.

### 5.2.3.1. Validitätsbereiche von Primärstudien

Cook & Campbell (1979) nennen 4 Validitätsbereiche von Primärstudien und geben entsprechende Hinweise für deren Beurteilung (Ausschluss):

**1.) Validität des statistischen Schlusses**

| | |
|---|---|
| Mangelnde Teststärke | Ungenügende Reliabilität der Messinstrumente |
| Voraussetzungsverletzungen | Inflationierung des α-Fehlers |
| Mangelnde Reliabilität der Behandlungsimplementierung | |

**2.) Interne Validität**

| | |
|---|---|
| Geschichtliche Einflüsse | Mehrdeutigkeit der kausalen Richtung |
| Reifungsprozesse | Behandlungsdiffusion |
| Testeffekte | Kompensatorische Behandlungsapplikation |
| Messeffekte | Kompensatorische Rivalität der Kontrollgruppen |
| Regression zu Mitte | Demoralisierung in den KGs |
| Selektionseffekte | Behandlungsimitation |
| Drop-Out | |
| Interaktion zwischen Selektion den übrigen Faktoren | |

**3.) Konstruktvalidität**

| | |
|---|---|
| Ungenügende a priori Explikation der Konstrukte | HypothesenkonformesVerhalten |
| Eindimensionale Operationalisierung | Soziale Erwünschtheit |
| Eindimensionale Methoden | Rosenthal-Effekt |
| Verwechslung der Konstruktebenen | Interaktionen verschiedener Behandlungen (BxB, Bxt) |
| | Eingeschränkte Generalisierung über die Konstrukte |

**4.) Externe Validität**

| | |
|---|---|
| IA: Selektion x Behandlung | IA: Geschichte x Behandlung |
| IA: Setting x Behandlung | |

Neben den angeführten 4 Validitätsbereichen (statistisch, intern, Konstrukt- und extern) lassen sich noch weitere Validitätsbereiche angeben, u. a. **Authentizität** (authenticity; Interne Konsistenz: passt alles zusammen?), **Glaubwürdigkeit** (credibility), **Repräsentativität** (representativeness) und **Bedeutung bzw. Sinngehalt** (meaning).

Für Journals gilt hinsichtlich der Authentizität: Ein Journal das Werbung abdruckt kann sich nicht selbst tragen und ist damit nicht besonders beliebt.

Ebenfalls nachdenkenswert: Was ist mit Drittmittelforschung? Die Forscher werden ja schließlich für die Studien von externen Geldgebern finanziert („ja, natürlich sind diese Ergebnisse valide“).

Eine Kontrolle der Validität ist möglich über:

- **Bewertung von Primärstudien**, z.B. über Kriterienkataloge validitätsmindernder Faktoren (z.B. Cook & Campbell, 1979) oder definierte objektive Studienmerkmale wie N oder Teststärke der Verfahren
- **Gewichtung oder Ausschluss von Primärstudien** nach den unter 5.1.2.3 a) genannten Kriterien zur Bewertung der methodischen Qualität von Primärstudien.

#### 5.2.3.2. Reliabilität von Quellen

Hier gilt es u. a. zu beachten, dass Diplomarbeiten und Dissertationen eine geringere mittlere Effektstärke aufweisen. Dies liegt an deren geringeren Qualität hinsichtlich der Versuchsdurchführung trotz höherer Qualität der Versuchsplanung.

### 5.2.4. Exkurs: Graue Literatur

Neben der in Datenbanken zu findenden Literatur gibt es auch die Möglichkeit, graue Literatur heranzuziehen. Beispiele hierfür sind:

- Unveröffentlichte Berichte/Paper
- Dissertationen und Diplomarbeiten
- Artikel in obskuren Zeitschriften oder Online-Journals
- Abstractbände von Tagungen und Konferenzen
- Politische Dokumente
- Abgelehnte oder nicht eingereichte Manuskripte
- Nicht-englischsprachige Artikel ☺

Typische Merkmale grauer Literatur sind daher kleine Stichproben, Pilotstudien, schwer zugängliche Stichproben und hoch-innovative Behandlungen.

Hier sollte man berücksichtigen, dass ein signifikanter Effekt bei einer kleinen Stichprobe evtl. mehr aussagt als ein signifikanter Effekt bei einer sehr großen Stichprobe. Auch ist graue Literatur innovativer, weil sie schneller ist. Während es 2-3 Jahre dauert, bis ein Paper in einem Journal veröffentlicht wird, ist graue Literatur sehr schnell verfügbar. Daher: Innovative Forschung ist grau!

Anmerkung: Wikipedia. Wikipedia zählt ebenfalls zur grauen Literatur. Ein Vergleich von 50 Texten aus Wikipedia und Brockhaus ergab für Wikipedia eine Gesamtnote von 1,7 und eine Gesamtnote von 2,7 für den Brockhaus. Wiki ist die Lösung ☺.

In der geschilderten Studie wurden Themen aus der Allgemeinbildung verwendet, da sich auch nur diese für einen Vergleich mit dem Brockhaus eignen. Bei Wikipedia sollte allerdings beachtet werden, dass bei vielen Themen nur wenige Autoren Beiträge leisten oder leisten können. Hier gilt es kritisch zu sein. Insgesamt gilt also: Die Vernachlässigung grauer Literatur führ zur Ausgrenzung innovativer Forschung.

## 5.2.5. Exkurs: Publikationsverzerrung (Publication bias)

Da hauptsächlich publizierte Primärstudien verwendet werden, greift man auf zu homogene Stichproben zurück, die unrepräsentativ für den Forschungsbereich sein können. Man unterscheidet zwei Arten der Publikationsverzerrung:

- **Report Bias:** Im Manuskript werden nicht signifikante Ergebnisse nicht dargestellt. „Von meinen 10 AVn werden nur zwei signifikant…huch, ich hab ja auch nur die 2 erhoben…“.
- **Retrieval Bias (File Drawer Problem):** Nicht-Einreichen von Manuskripten.

Wenn nur veröffentlichte Studien berücksichtigt werden resultiert also eine systematische Überschätzung der integrierten Effekte.

### 5.2.5.1. Fail-Safe N (Widerlegungssichere Untersuchungszahl)

Das Fail-Safe N (Rosenthal, 1979) beschreibt die Anzahl der Studien mit Nulleffekten, die erforderlich sind, um die Signifikanz des Gesamtergebnisses unter ein festzulegendes Signifikanzniveau zu senken.

$$FSN = \frac{\left[\left(\sum z_i\right)^2 - k \cdot z_\alpha^2\right]}{z_\alpha^2}$$

mit $z_i$ = z-Werte der einzelnen Studien
$k$ = Anzahl der Studien
$z_\alpha$ = z-Wert des α-Niveaus

Daumenregel: Ist FSN > 5k + 10, dann ist nicht davon auszugehen, dass eine Signifikanz des Gesamtergebnisses nur durch Selektion signifikanter Ergebnisse zustande kommt (FSN > 5k + 10 → Wahrer Effekt).

Es ergeben sich also zwei Interpretationsmöglichkeiten für FSN = 9,85 bei k = 5 Studien:

- Es werden mindestens 10 nicht signifikante Studien benötigt, um als Gesamtergebnis „nicht signifikant“ zu produzieren
- 9,85 $\searrow 5 \cdot 5 + 10 = 35 \rightarrow$ Man kann nicht davon ausgehen, dass wirklich ein signifikanter Effekt vorliegt. Man könnte ihn noch ns kriegen, wenn man noch 10 nicht signifikante Studien findet.

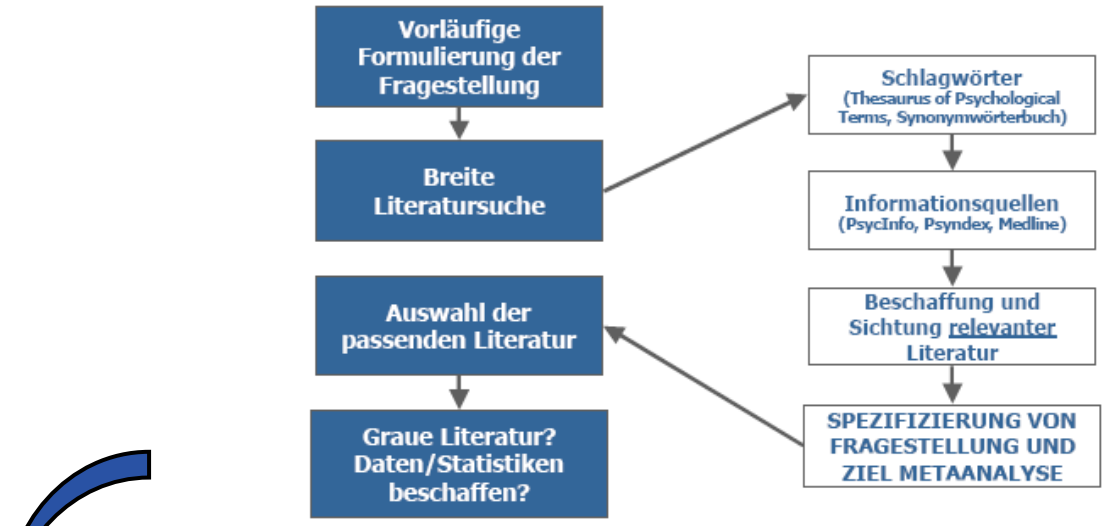
### 5.2.5.2. FSN: Funnel-Plots

Die Repräsentativität der Stichprobenauswahl lässt sich auch über sog. Funnel-Plots schätzen. Dafür wird der Schätzer für die Effektgröße auf die Abszisse angetragen und die Stichprobengröße bzw. der Standardfehler auf die Ordinate.

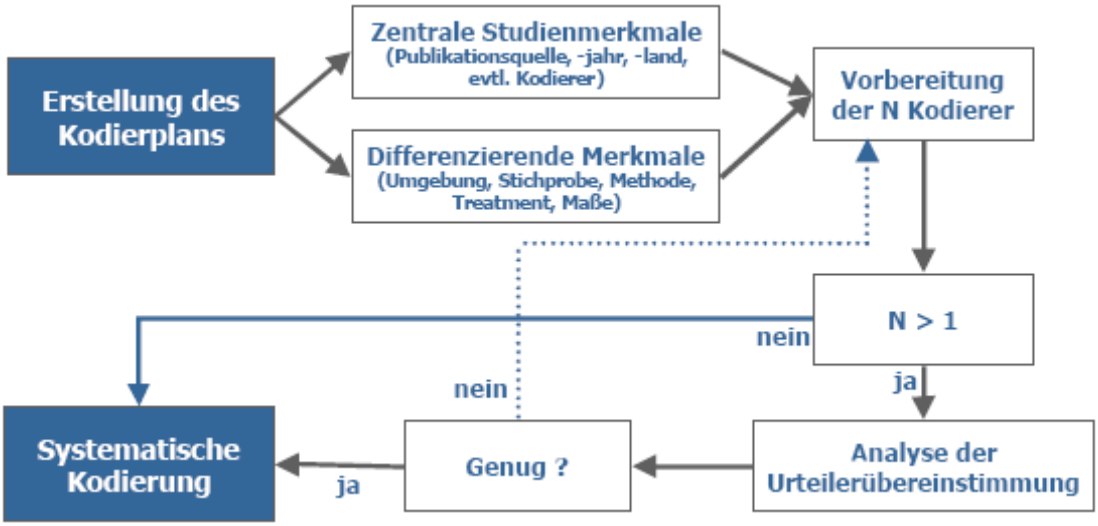
Annahme: Die Streuung der Effektgrößenschätzungen um die wahre Effektgröße wird bei zunehmendem N kleiner. Die graphische Darstellung sollte also eine Form eines gedrehten symmetrischen Trichters annahmen. Falls der Funnel-Plot asymmetrisch ist liegt vermutlich ein Publication Bias vor: Nicht-signifikante bzw. nicht-hypothesenkonforme Studien wurden nicht veröffentlicht.

# 5.3. Kodierung

Standpunkt: Literatur (inkl. Grauer Literatur) liegt vor und die Fragestellung ist spezifiziert (vgl. 5.2 bzw. Folie 25 zur Planung und Vorbereitung einer Metaanalyse):




Nun kann die Kodierung der Studien erfolgen, die mit der Erstellung eines Kodierplans beginnt. Anschließend werden die Kodierer trainiert und es erfolgt eine systematische Kodierung.




## 5.3.1. Kodierplan

Es werden die zu kodierenden Merkmale festgelegt. Hierbei wird zwischen **zentralen Studienmerkmalen** wie Publikationsquelle, -jahr und -land sowie ggf. Kodierer und **differenzierenden Merkmalen** wie Umgebung, Stichprobe, Methode, Treatment und Maße unterschieden.

Hierbei gibt es verschiedene Möglichkeiten zur Auswahl der zu kodierenden Variablen:

- Substanzielle Merkmale der Primärstudie: Inhaltlich relevante Merkmale von Untersuchungsgruppen, Behandlung, Rahmenbedingungen sowie Messinstrumente.
- Verzerrende Merkmale der Primärstudie: Qualitätsmerkmale des Versuchsdesigns.

- Extrinsische Merkmale der Primärstudie: Generelle Eigenschaften der Primärstudien (Autor, Jahr, Land, Institution) sowie editoriale Merkmale (Darstellungs- und Berichtsmerkmale, Titel der Quelle).

Die Zahl der Kategorien sollte in Abhängigkeit von der Zahl der Effekte, die in die Datenbank eingehen gewählt werden. Es sollte jeweils gewährleistet werden, dass alle Kategorien ein Mindestmaß an Besetzung zeigen, weswegen bei kleinen Datenbeständen nicht mehr als 5 Kategorien verwendet werden sollten.

Zudem ist die Auswahl der Kategorienzahl abhängig von den Zielen des Metaanalytikers. Bei einer hierarchischen Zerlegung der Wirkzusammenhänge verringert sich bspw. die Kategorienzahl (Hunter & Schmidt, 1990).

### 5.3.2. Vorbereitung der N Kodierer

Idealerweise werden mindestens zwei unabhängige Beurteiler eingesetzt, sodass Güte und Eindeutigkeit der Klassifikationen, Bewertungen und Datenextraktionen geprüft werden können.

Dabei sollte die Beurteilung blind gegenüber Methoden- und Theorieteil sein. Die beiden Publikationsteile sollten also auf verschiedene Kodierer aufgeteilt werden.

Für eine möglichst gute Kodierung sind die Entwicklung eines Kodierhandbuchs sowie spezifische Kodierblätter und ein Kodierertraining unerlässlich.

Das Training der Beurteiler lässt sich bspw. über das Kappa-Maß evaluieren. Das Ziel des Trainings ist dabei immer die Vereinheitlichung der Zuordnung einer Studie zu den Kategorien für die differenzierenden Merkmale.

Problematisch ist hierbei, dass die Urteilerübereinstimmungen von der Art der Informationen abhängen (Rosenthal, 1991):

- $r_{max}$ = 1.0 für „Median Alter“, „Zeitschriftenname“
- $r_{max}$ = .57 für „Anzahl der Subgruppen“
- $r_{max}$ = .44 für „Fragebogen als Hauptverfahren“

### 5.3.3. Systematische Kodierung

Ist eine genügend hohe Beobachterübereinstimmung erreicht, kann mit der systematischen Kodierung der Primärstudien begonnen werden. Das Ergebnis dieser Kodierung ist nun der Ausgangspunkt für die Anwendung der verschiedenen statistischen Techniken der Metaanalyse.

Während der Durchführung können systematische Kodierverzerrungen z.B. durch Anonymisierung der Studien eingeschränkt werden.

## 5.4. Statistische Analyse I: Deskriptive Statistik

Zunächst werden deskriptive Ansätze (Effektstärke, Korrelationen) vorgestellt. Anschließend folgen zwei inferenzstatistische Verfahren (Vote-Counting, Tests).

### 5.4.1. Effektstärke

Das zentrale Maß der Metaanalyse ist die Effektstärke ES, als standardisiertes Maß für die Differenz zwischen den Mittelwerten von Experimentalgruppen. Um eine ES zu berechnen muss folgendes bekannt sein:

- Versuchsdesign (Kontrollgruppe, Prä-post-Messung)
- Stichprobengröße in EG und KG
- Mittelwerte von EG und KG
- Standardabweichungen von EG und KG (bzw. F- oder t-Werte, um die Standardabweichung zu schätzen).

Die Grundformel der Effektstärke lässt sich also folgendermaßen angeben:

$$ES = \frac{M_a - M_b}{s}$$

Die Schätzung der Effektstärke folgt dabei einem standardisierten Vorgehen:

1.) Berechnung der Effektstärke pro Studie
2.) Berechnung der gesamten Effektstärke als Mittelwert dieser Effektstärken unter Berücksichtigung der jeweiligen Stichprobenumfänge (ES x N; gewogenes arithmetisches Mittel).
3.) Interpretation der Effektstärken in Einheiten der gemittelten Standardabweichungen.

Als Faustregel lässt sich zumindest für die Berechnung gemäß der klassischen Referenz (s.u.) folgender Interpretationsrahmen festlegen:

| | |
|---|---|
| ES < .2 | Kein Effekt |
| .2 < ES < .5 | Kleiner Effekt |
| .5 < ES < .8 | Mittlerer Effekt |
| ES > .8 | Großer Effekt |

### 5.4.1.1. Formeln

Die gebräuchlichste Form der ES-Berechnung ist die Methode der klassischen Referenz nach Smith, Glass und Miller (1980). Voraussetzung der Methode ist das Vorliegen eines Kontrollgruppendesigns und zwei Messzeitpunkte:

$$ES = \frac{M_{treat\_t2} - M_{control\_t2}}{s_{control\_t2}}$$

Liegt kein Kontrollgruppendesign vor (z.B. in Therapievergleichsstudien), so können die unpassenden Studien ausgeschlossen werden. Eine geschicktere Alternative ist die Entwicklung von alternativen ES-Formeln, die auch bei anderen Designs funktionieren.

Grundprinzip ist hierfür die Verwendung einer Mittelwertsdifferenz zweier Zeitpunkte im Zähler sowie eine Verknüpfung der Standardabweichungen im Nenner. Beispiel sind die Formeln von Grawe (1992) und McGaw und Glass (1980).

Die Klassische Referenz führt insgesamt zu geringeren mittleren ES als ihre Varianten. Auch führt die Verwendung von Post-Varianzen zu einer Erhöhung der mittleren Effektstärke (da die Varianz nach der Behandlung geringer sein sollte als davor und sie im Nenner steht).

Weiterhin gilt, dass im Bereich von 0 < ES < 1 alle ES-Formeln zu sehr ähnlichen Ergebnissen kommen. Für ES < 0 und ES > 1 sind große Unterschiede zu berücksichtigen.

**Wichtig:** Aus den bisherigen Erklärungen wird deutlich, dass „Effektstärke“ nicht gleich „Effektstärke“ ist. Die Auswahl der Formel hat einen Einfluss auf die Ergebnisse der Metaanalyse und daher sollte auch beim Lesen einer solchen immer auf die verwendete Methode geachtet werden.

Zu beachten ist hierbei, dass die Einteilung der ES in klein, mittel und groß nur für die Methode der klassischen Referenz, nicht jedoch für Varianten gilt. Trotzdem werden diese Werte auch hierfür gerne verwendet (pfusch).

Hohes Missbrauchspotential bietet auch der Einfluss der Varianzen auf die Effektstärke. Wird beispielsweise erwartet, dass die Behandlung zu einer Homogenisierung der Stichproben führt, so resultiert bei Verwendung der Post-Varianz natürlich eine höhere ES.

Führt man selbst eine Metaanalyse durch, so sollte eine ES-Variante für alle zu vergleichenden Studien festgelegt werden. Subgruppenanalysen für Studien, bei denen bspw. die klassische Referenz nicht verwendet werden kann sind zwar möglich, sollten aber eher gemieden werden.

### 5.4.1.2. Moderatorvariablen

Ein weiteres wichtiges Konzept im Zusammenhang mit Effektstärken ist das Moderatorkonzept[10]. Die bisher vorgestellte Berechnung geht dabei von **homogen strukturierten Studieneffekten** aus – es wird also *ein* wahrer Wert angenommen, den alle Studien mehr oder weniger treffen.

Dies muss jedoch nicht so sein, da evtl. mehrere wahre Werte vorliegen – die Effektstärken also durch Moderatorvariablen bedingt sind. Dabei deutet eine **heterogene ES-Verteilung** auf das Vorliegen von Stichprobenfehlern hin. Die ES-Verteilung ist über einen $\chi^2$-Test prüfbar (**Q** als $\chi^2$-verteilte Prüfgröße; $H_0$: Homogenität der ESn; es gibt einen wahren Wert. $H_1$: Heterogenität; Stichprobenfehler).

Wird die Nullhypothese eines solchen Tests verworfen, müssen Moderatorvariablen mit einbezogen werden. Sind keine solchen bekannt muss die Metaanalyse abgebrochen werden. Natürlich sollten schon vor Berechnung solcher Kennwerte Überlegungen zu möglichen Moderatorvariablen aus der Literatur abgeleitet werden.

Auch ist es möglich, Subgruppen mit homogenen Effektstärkeverteilungen zu bilden. Dieses Konzept ist erweiterbar auf weitere Unter-Untergruppen (Hierarchische Zerlegung nach Hunter & Schmidt, 1990).

Ein weiteres Indiz für nicht homogen strukturierte Studieneffekte ist, wenn ein größerer Varianzanteil nicht durch **Zufallsfehler** erklärt werden kann.

Hier gibt es die sagenumwobene 75%-Regel: „Wenn mindestens 75% der Varianz auf Stichprobenfehler zurückzuführen sind, liegt *Homogenität* vor.“

$$\frac{\text{Fehlervarianz}}{\text{Gesamtvarianz}} \geq 75\%$$

Das Problem der 75%-Regel ist, dass sie nicht statistisch begründet ist und sich auch keine Originalmanuskripte finden, in denen diese genauer beschrieben ist. Daher: „Verlasst die 75%-Regel; nehmt Q!“.

[10] Vorlesungsbeispiel: „Ingo und das schöne Wetter“: Starker Effekt am Strand, kein Einfluss an Fasching (Alkohol und Party), sogar negativer Effekt auf der langweiligen Konferenz/Messe.

[Anmerkung zu F 50: Natürlich ist eine heterogene Verteilung der ES anzunehmen, wenn der Homogenitätstest signifikant wird (und nicht „nicht-signifikant“).]

## 5.4.2. Korrelationen

Bisher wurden immer nur Mittelwerte betrachtet (in Form der ES). Die bisherigen Methoden sind bei anderen Datenniveaus also nicht anwendbar.

Eine Alternative sind (Produkt-Moment-)Korrelationen als Ersatz für die ES. Korrelationen können im Gegensatz zu Effektstärken bei jeder Fragestellung verwendet werden, also bspw. auch bei Rangdaten.

Für die Interpretation wird der Determinationskoeffizient $R^2$ als Anteil der aufgeklärten Varianz herangezogen. Auch hierfür gibt es Richtlinien für die Größe eines bestimmten Effekts (Umrechnungsformel aus ES; s. Bortz & Döring).

| | | |
|---|---|---|
| $R^2 < .01$ | $r < .1$ | Kein Effekt |
| $.01 < R^2 < .09$ | $.1 < r < .3$ | Kleiner Effekt |
| $.09 < R^2 < .25$ | $.3 < r < .5$ | Mittlerer Effekt |
| $R^2 > .25$ | $r > .5$ | Großer Effekt |

Diese Einteilung ist jedoch kritisch zu sehen. Ein kleiner Effekt mit $R^2 = .04$ leistet nur 4% Varianzaufklärung. Das ist fast nichts.

Auch werden Korrelationen als Effektmaße mit größeren Unterschieden in den Stichprobengrößen sehr ungenau.

### 5.4.2.1. Einsatzmöglichkeiten

Neben dem Einsatz als einfache Effektmaße sind Korrelationen auch nützlich, weil man alle Teststatistiken in PM-Korrelationen umrechnen kann: Standardisiertes Effektmaß Δ (= r). Somit können auch Studien ausgewertet werden, die nur einen F- oder t-Wert berichten.

Eine weitere Möglichkeit ist die Überführung einer ES in den Determinationskoeffizienten $R^2$ als Kennwert für die Varianzaufklärung. Hierfür rechnet man die standardisierte Mittelwertsdifferenz in eine entsprechende Teststatistik um (Prinzip: Teststatistik = Effektstärke * Studiengröße). Diese Teststatistik lässt sich in einem zweiten Schritt dann in eine PM-Korrelation umwandeln.

Durch die Umwandlung verschiedenster Kennwerte in Korrelationskoeffizienten lassen sich also auch verschiedenste Studien miteinander vergleichen und so in der Metaanalyse verarbeiten.

### 5.4.2.2. Binomial Effect Size Display

Die Effektstärke sagt zwar etwas über den Unterschied zwischen EG und KG aus, man kann jedoch keine Aussagen über die Effektivität des Treatments im Sinne einer Erfolgsquote treffen.

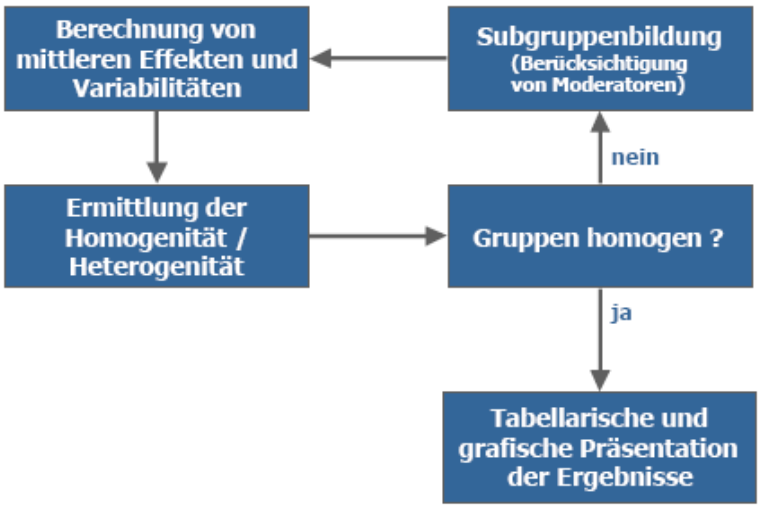
Zur leichteren Interpretation von metaanalytischen Kennwerten bei dichotomen Merkmalen wurde daher das BESD als Maß für die Erfolgsrate eingeführt.

$$BESD = .5 \pm r/2$$

mit EG: $.5 + r/2$ und KG: $.5 - r/2$

Das BESD lässt sich nun tatsächlich als Erfolgsrate interpretieren. Mit r = .7 bei einer Metaanalyse zur Wirksamkeit einer therapeutischen Maßnahme ergibt sich also eine Erfolgsrate von .5 + .7/2 = 85% und eine Erfolgsrate von .5 - .7/2 = 15% für die Kontrollbedingung.

### 5.4.3. Zusammenfassung: Deskriptive Statistik




### 5.4.4. Anmerkung: DGPs-Richtlinien

Seit Herbst 2007 muss nach den DGPs-Richtlinien zur Manuskriptgestaltung immer die Effektstärke oder ein vergleichbares Maß des Effekts berichtet werden (ES / d, $\eta^2$, r).

Anfang der 90er Jahre wurde postuliert, dass Primärstudien in Zukunft nur noch als Bausteine für Metaanalysen fungieren (Schmidt, 1992). Diese Position findet hier ihren Ausdruck.

Anmerkung: Ebenso muss seit dieser Auflage der Richtlinien auch immer das exakte Signifikanzniveau der Untersuchung berichtet werden, also p = .034 statt p < .05.

## 5.5. Statistische Analyse II: Inferenzstatistik

### 5.5.1. Vote-Counting

Bisher wurden ausschließlich Verfahren zur Beschreibung der Effekte in Primärstudien geschildert. Im Folgenden soll daher noch die inferenzstatistische Prüfung in Metaanalysen betrachtet werden. Hierfür stehen zwei Ansätze zur Verfügung: Vote-Counting und Summierung von Teststatistiken.

Unter Vote-Counting versteht man das einfache Auszählen von Prüfergebnissen („signifikant“ vs. „nicht-signifikant“). Ausgangspunkt dieser Methode ist die häufig mangelhafte Dokumentation der Studien hinsichtlich relevanter Kennwerte. Angaben zur Signifikanz finden sich als einzige mit Sicherheit in jedem Bericht.

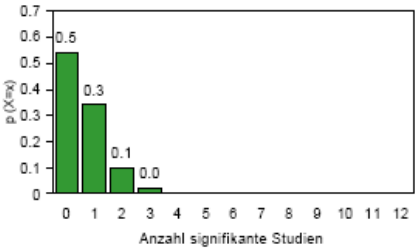
Eigentlich handelt es sich also um ein sehr grobes Verfahren und erlaubt auch kaum Rückschlüsse auf Existenz und Größe des wahren Effekts. Aber: Es gibt eine Weiterentwicklung aus Würzburg (fabrikneu), die eine derartige Auswer-

tung auf Basis der Binomialverteilung erlaubt. Dieses Verfahren soll am Beispiel „4 von 12 Studien signifikant“ illustriert werden.

Vorüberlegungen: Es gibt die zwei Extremfälle, dass alle (12/12) oder keine Studien (0/12) signifikant werden. Im ersten Fall liegt mit großer Sicherheit ein wahrer Effekt vor, im zweiten Fall mit großer Sicherheit nicht.

Wenn es in Wahrheit ($H_0$) keinen Effekt gäbe, sollte eine Studie zu $p = \alpha = 0.05$ signifikant werden. Es ist davon auszugehen, dass die einzelnen Studien unabhängig voneinander sind. Damit liegt ein Bernoulli-Prozess vor (2 Ausgänge, Stationaritätsannahme $p_1 = p_2 = \ldots = p_i$ erfüllt, unabhängige Durchgänge), der es erlaubt zu berechnen, wie viele signifikante bei n =12 Studien zu erwarten wären.

Es ist somit extrem unwahrscheinlich, dass 4 oder mehr Studien signifikant werden, wenn in Wirklichkeit kein Effekt vorliegt.

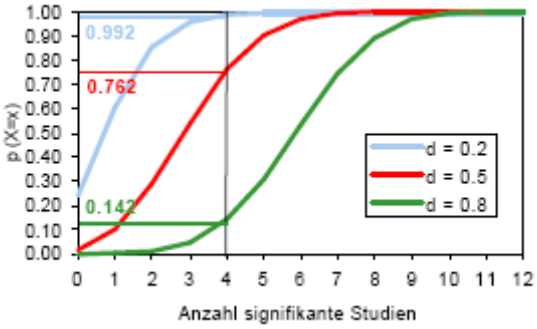



Für diesen Binomialtest ist also p deutlich kleiner als 0.05, sodass die Nullhypothese („In Wahrheit liegt kein Effekt vor.“) verworfen werden muss.

Die Existenz des Effekts ist also gesichert, sodass nun nach seiner Größe gefragt werden kann.

Während für das Zeigen des Effekts die normale Testlogik verwendet wurde wird diese für die Prüfung der Größe des Effekts umgedreht. Hier werden verschiedene Verteilungen für die Nullhypothesen „Es liegt ein kleiner / mittlerer / großer Effekt vor.“ berechnet. Es wird also wieder über die Binomialverteilung mit p = ES ausgewertet.

Ist das empirische Ergebnis unter den Annahmen des Modells hinreichend wahrscheinlich, so wird die Nullhypothese beibehalten. Der Effekt ist danach mindestens so groß wie die geprüfte ES.




Als Daumenregel wird dasjenige Ergebnis interpretiert, das eine Wahrscheinlichkeit > 95% aufweist.

Zur Abschließenden Bewertung des Vote-Countings bleibt festzuhalten, dass nur Informationen hinsichtlich der Signifikanz von Studien vorliegen müssen. Ein Problem ist die fehlende Differenzierung zwischen Effekten verschiedener Richtung (Vorzeichen). Hier ist es möglich, die Effektstärken nach positiv und negativ einzuteilen, wenn das Vorzeichen der Effektdifferenz bekannt ist. In so gebildeten Subgruppen lassen sich gesonderte Analysen durchführen.

Das Vote-Counting erlaubt dabei immer eine Aussage sowohl über Existenz als auch ungefähre Größe des untersuchten Effekts, was einen großen Vorteil des Verfahrens ausmacht.

### 5.5.2. Teststatistiken

Liegen in den Primäruntersuchungen auch Teststatistiken (t-, F- und/oder p-Werte) vor, so lassen sich auch diese inferenzstatistisch prüfen. Hierbei gibt es verschiedene Methoden:

- Adding of Logs
- Addieren von t-Werten
- Addieren von z-Werten

Alle diese Verfahren können dabei jedoch immer nur prüfen, ob ein Effekt vorhanden ist. Die Größe des Effekts wird nicht betrachtet.

Liegen deskriptive Statistiken (m, s) vor, so können natürlich auch Effektstärken berechnet und geprüft werden.

### 5.5.3. Zusammenfassung: Inferenzstatistik

Sobald die Signifikanz einer Studie berichtet wird, kann eine inferenzstatistische Auswertung im Rahmen der Metaanalyse erfolgen. Hier ist die Methode des Vote-Countings einschlägig. Diese ist ebenfalls einsetzbar, wenn nur die Signifikanzklasse berichtet wird.


Bei besseren Daten können auch Teststatistiken oder Effektstärken analysiert werden. Das ? bei Effektstärken soll darauf aufmerksam machen, dass hier immer bestimmte Voraussetzungen erfüllt sein müssen.

## 5.6. Bewertung der Metaanalyse

In den Anfängen der Metaanalyse wurde argumentiert, dass es sich um ein systematisches, quantitatives Vorgehen handelt, welches hochgradig objektiv und reliabel ist. Methodische Probleme wurden v.a. auf zugrunde liegende Primärstudien zurückgeführt.

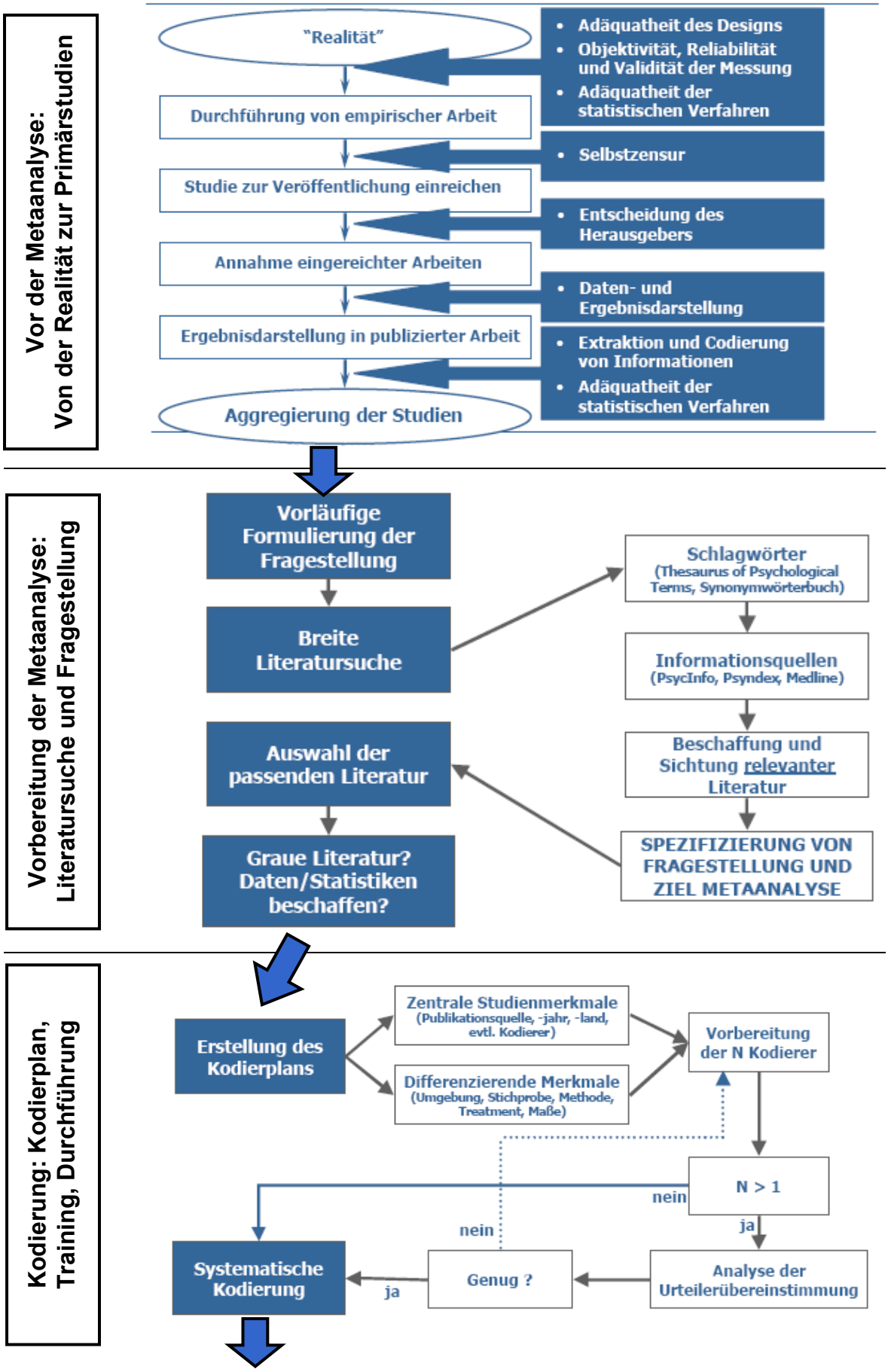
Heute ist jedoch deutlich, dass sowohl Datensammlung als auch Integration subjektiven Entscheidungen unterliegen, sodass auch Metaanalysen einer kritischen Evaluation zu unterziehen sind. Erforderlich sind metaanalytische Replikationsstudien und Re-Analysen (Meta-Metaanalysen). Auch zeigt sich ein Trend zur Erstellung eines Reviews über mehrere Meta-Analysen ☺.

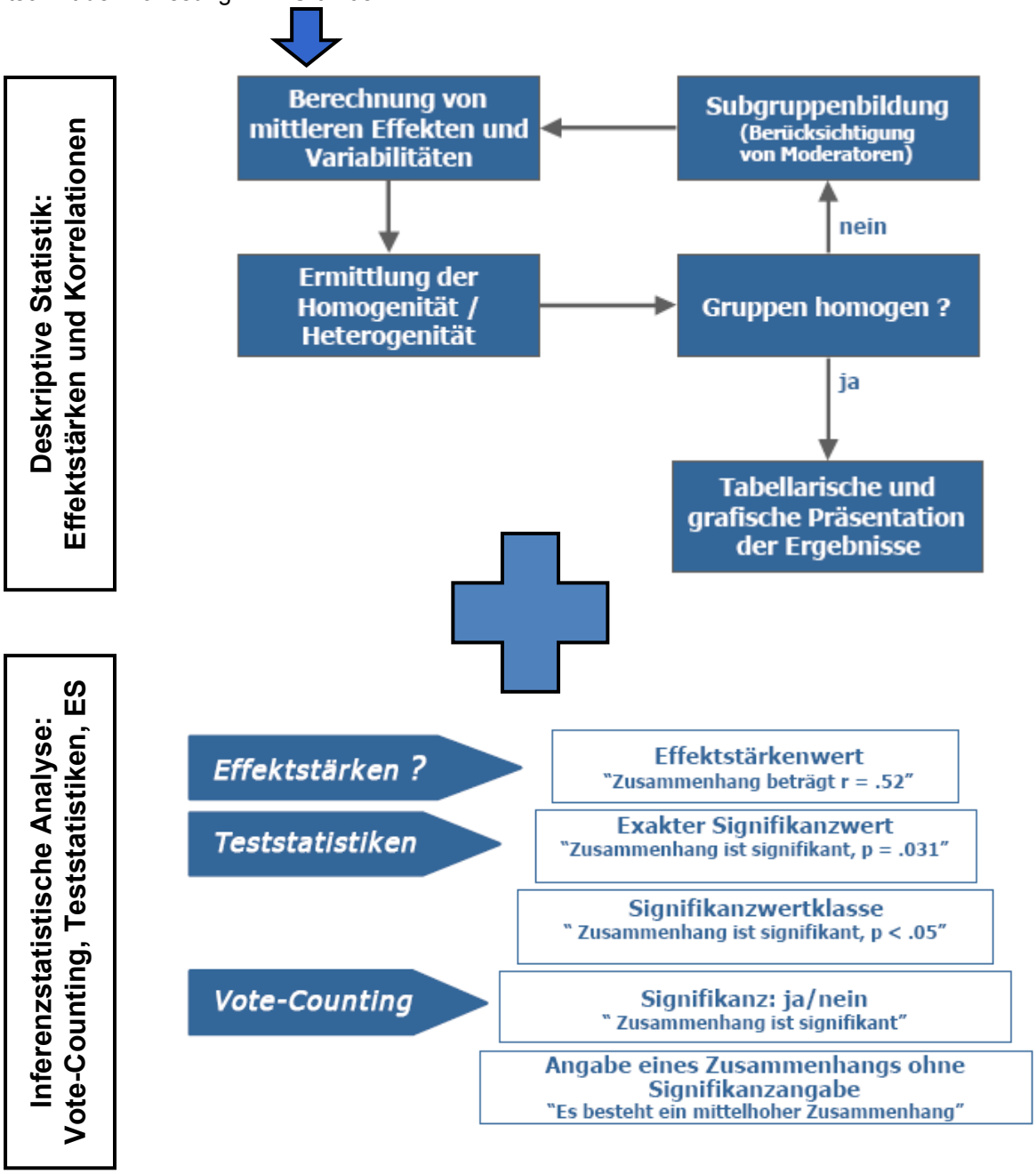
## 5.7. Checkliste: Ist eine Studie meta-analytische auswertbar?

Nach Czienskowski (2002):

1. Stellen Sie die theoretischen und empirischen Bezüge Ihrer Studie dar.
2. Achten Sie auf eine vollständige Aufstellung der Unabhängigen Variablen (UVn) und Abhängigen Variablen (AVn).
3. Stellen sie alle Hypothesen, insbesondere über Einzelvergleiche, dar.
4. Berichten Sie das Treatment detailliert, v.a. alle variierten Treatmentbedingungen.
5. Berichten Sie die Randbedingungen der durchgeführten Studie (z.B. Land, Population, Alter, Anzahl der untersuchten Einheiten).
7. Geben Sie Tabellen kumulierter Daten (M, SD) an.
8. Geben Sie möglichst alle statistischen Kennwerte und exakte Signifikanzwerte ($p=.034$ statt $p<.05$) vollständig wider.

# 5.8. Zusammenfassung




Deskriptive Statistik:
Effektstärken und Korrelationen
Berechnung von mittleren Effekten und Variabilitäten
Subgruppenbildung
(Berücksichtigung von Moderatoren)
nein
Ermittlung der Homogenität / Heterogenität
Gruppen homogen ?
ja
Tabellarische und grafische Präsentation der Ergebnisse
Inferenzstatistische Analyse:
Vote-Counting, Teststatistiken, ES
Effektstärken ?
Effektstärkenwert
"Zusammenhang beträgt r = .52"
Teststatistiken
Exakter Signifikanzwert
"Zusammenhang ist signifikant, p = .031"
Signifikanzwertklasse
" Zusammenhang ist signifikant, p < .05"
Vote-Counting
Signifikanz: ja/nein
" Zusammenhang ist signifikant"
Angabe eines Zusammenhangs ohne Signifikanzangabe
"Es besteht ein mittelhoher Zusammenhang"

# 6. Qualitatives Vorgehen

Ausgangspunkt der Abgrenzung von quantitativer und qualitativer Forschung sind folgende Überlegungen.

Quantitative Forschung beschäftigt sich u. a. mit der Frage der Operationalisierung bzw. Quantifizierung der zu erhebenden Merkmale, wofür verschiedene Versuchspläne und Datenquellen zur Verfügung stehen. Die erhaltenen Messwerte werden statistisch analysiert und verarbeitet.

Qualitative Forschung bezieht sich vor allem auf die Interpretation von verbalem Material[11] (Beobachtungsprotokolle, Interviewtexte, Briefe, Zeitungsartikel). Auch findet sie in der Verarbeitung nicht-numerischer Daten (Fotos, Zeichnungen, Filme, …) Anwendung. Meist werden qualitative Ansätze für Explorationsstudien oder in Kombination mit quantitativen Ansätzen verwendet.

## 6.1. Übersicht

### 6.1.1. Grundprinzipien qualitativen Vorgehens

Es ist schwer zu definieren, was einen qualitativen Ansatz ausmacht, da die qualitativen Ansätze eine äußerst heterogene Gruppe darstellen. Es gibt also nicht DAS Design qualitativer Art (so wie das Experiment der quantitativen Ansätze). Allerdings lassen sich typische Kennzeichen finden.

- Qualitative Forschung bedeutet **intensiven und/oder verlängerten Kontakt im Feld** bzw. in einer Lebenssituation (alltagsnah)
- **Zielsetzung: Holistischer Überblick** über Situationen und relevante Prozesse. Nicht nur UV → AV soll betrachtet werden, sondern auch Rückkopplungen: ┌→UV → AV ─┐
- **Datengewinnung: Der Forscher als zentrales Messinstrument.** Es liegen relativ wenige Standardverfahren vor bzw. werden so gut wie nicht eingesetzt. Grundgedanke dieser Strategie ist die Überwindung hinderlicher Vorannahmen über „tiefe Aufmerksamkeit und empathisches Verstehen“.
- **Datenanalyse:** Zumeist sprachgebundene Datenverarbeitung. Informationselemente werden isoliert bzw. hervorgehoben. Es wird davon ausgegangen, dass immer verschiedene Interpretationen möglich sind, manche jedoch aufgrund theoretischer Annahmen bedeutsamer sind als andere.
- **Fremdheitspostulat:** Die Sicht auf eine objektiv identische Situation kann zwischen zwei Personen starke Differenzen aufweisen (z.B. zwischen Untersucher und Untersuchtem).
- **Prinzip der Offenheit:** Es sollten vorab (ex ante) keine Hypothesen gebildet werden. Auch sollten die Methoden möglichst wenig vorstrukturiert sein.
- **Prinzip der Kommunikation:** Kommunikation und Reflexion der Forschers nicht als Störvariable sondern als explizierter Bestandteil der Erkenntnis.

[11] „Transkription? Huch, da geht’s in Richtung esoterisch.“

### 6.1.1.1. Lineare und zirkuläre Forschungsstrategien

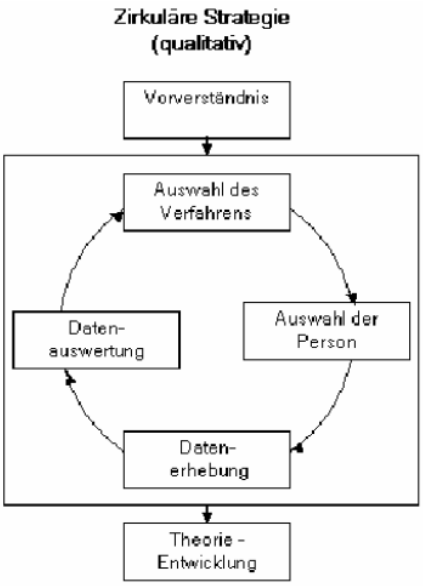
Die lineare Strategie ist mit deduktivem bzw. quantitativem Vorgehen verbunden. Zielsetzung ist die Vergleichbarkeit von Daten verschiedener Datenquellen. Dabei werden die einzelnen Schritte nacheinander abgearbeitet[12] – von der Formulierung einer Hypothese bis hin zum statistischen Test dieser Hypothese. Während der Untersuchungsdurchführung dürfen frühere Phasen nicht modifiziert werden.

Qualitative bzw. heuristische Ansätze hingegen **verwenden eine zirkuläre Forschungsstrategie** (Witt, 2001). Zielsetzung ist die holistische Erfassung des Untersuchungsgegenstandes. Dabei wird die Schrittfolge mehrmals durchlaufen, wobei zu Beginn der Forschung nur ein ungefähres Verständnis über den Forschungsgegenstand vorliegt.

Die endgültige Fragestellung, Umfang der Untersuchung, Größe der Stichprobe oder Bandbreite der verwendeten Verfahren ergeben sich erst im Laufe der Untersuchung. Dies führt zu einer generellen Unvergleichbarkeit der Daten.

Kennzeichen zirkulärer Forschung sind somit (Flick, 1995):

1.) **Theoretisches Sampling.** Die Fall- und Fallgruppenauswahl geschieht im Prozess der Forschung und nicht davor.
2.) **Theoretisches Kodieren.** Kategorien zur Zusammenfassung der Daten stehen nicht von Beginn an fest, sondern werden erst im Laufe der Untersuchung gebildet.
3.) **Schreiben der Theorie.** Eine Studie wird nicht in einem standardisierten Format (Abstract → Interpretation) dargestellt, sondern sie soll in Form einer Geschichte erzählt werden.




### 6.1.1.2. Problem: Strategienverschnitt

Häufig werden qualitative Verfahren in einer linearen Strategie verwendet. Im Untersuchungsdesign wird also der Einsatz eines qualitativen Verfahrens eingeplant. Meist ist dieser Einsatz nur einmalig, die Kreisbewegung wird also nicht mehrfach durchgeführt. Dabei werden nach Witt (2001) charakteristische Fehler gemacht.

- Erst Datenerhebung, dann Auswertung. Folge: Schwächen der Verfahren, Fehler der Versuchsleiter, Hinweise auf neue Aspekte werden nicht berücksichtigt.
- Stichproben zu homogen.
- Festhalten am bekannten Befund bzw. am Vorgehen.
- Auswertung nach festen Regeln oder vorher festgelegten Kategorien. Folge: Informationen können übersehen werden.
- Auszählen der Häufigkeiten bestimmter Inhaltskategorien. Aber: Evtl. ist gerade die Kombination von Kategorien wichtig.
- Verwendung nur eines einzigen Erhebungsverfahrens.

[12] Ähh…Sarris: „Forschung als simultaner und nicht sukzessiver Prozess.“!?

#### 6.1.1.3. Einsatz

„Warum werden die nicht häufiger eingesetzt?“. In Deutschland herrscht eine starre Schuldgebundenheit bzgl. quantitativer Vorgehensweisen: „Weiche von mir Satan – du verunreinigst meine Daten!“.

In Frankreich finden diese Verfahren auch in der psychologische Forschung wesentlich häufiger Anwendung. In Deutschland erfreuen sie sich vor allem in der Marktforschung großer Beliebtheit (Alleinstellungsmerkmal, wenn man sich genauer eingearbeitet hat).

Vielleicht ist die seltene Anwendung aber auch ganz gut…

### 6.1.2. Gütekriterien qualitativen Vorgehens

Die drei klassischen Gütekriterien Validität, Reliabilität und Objektivität werden von Vertretern der qualitativen Ansätze vehement abgelehnt. Folgende Argumentation wird gerne verwendet:

- Kriterienbezogene Validität: Ein neues Messinstrument will besser sein als das vorhergehende. Es will genau seinen Gegenstand treffen und nicht ähnliche Gegenstände.
- Interne Konsistenz: Wenn eine hohe interne Konsistenz vorliegt, so steckt jede Menge redundanter Information in einem Fragebogen. Den redundanten Teil hätte man sich auch sparen können.
- Retest-Reliabilität: Wirkung des Faktors Zeit sowie die Reaktivität des Untersuchten wird bei Retest-Korrelationen ignoriert.
- Durchführungs-, Auswertungs- und Interpretationsobjektivität: Entscheidend ist die Subjektivität des Forschenden als zentrales Messinstrument der qualitativen Ansätze.

…sic!

Mayring (1990), einer der deutschen Gurus der qualitativen Ansätze, nennt 6 andere Gütekriterien qualitativer Forschung:

- **Verfahrensdokumentation.** Forderung nach einer exakten Dokumentation der verwendeten Methoden.
- **Argumentative Interpretationsabsicherung.** Die eigene Interpretation soll insbesondere gegen Alternativerklärungen argumentativ abgesichert werden.
- **Regelgeleitetheit.** Forderung nach der Orientierung an Verhaltensregeln, z.B. Teilung des Materials in sinnvolle Einheiten.
- **Nähe zum Gegenstand.** Forderung nach der Nähe zum Alltag, was u.a. durch verstärkte Feldforschung realisiert werden kann.
- **Kommunikative Validierung.** Die Ergebnisse sollten mit den Probanden kritisch diskutiert und danach evtl. angepasst werden [gar keine so schlechte Idee…].
- **Triangulation.** Forderung nach der Verbindung bzw. dem Vergleich mehrerer Analysegänge (Datenquellen, Untersucher, Theorien und Methoden). Dadurch sollen Stärken und Schwächen der einzelnen Aspekte bestmöglich kombiniert werden.

### 6.1.2.1. Exkurs: Triangulation

Das Gütekriterium der Triangulation soll hier nochmals näher betrachtet werden. „Zusammengefasst beinhaltet methodologische Triangulation einen komplexen Prozess des Gegeneinander-Ausspielens jeder Methode gegen die andere, um die Validität von Feldforschung zu maximieren.“ (Denzin, 1978).

Ihren Ursprung hat die Triangulation in der Landvermessung. Dort bezieht sich der Begriff auf die exakte Positionsbestimmung eines Punktes von mindestens zwei unterschiedlichen anderen Punkten aus. Übertragen auf die empirische Sozialforschung sollte jeder Gegenstand mindestens durch zwei verschiedene Forschungsmethoden erforscht werden.

Dabei können 4 Subtypen der Triangulation unterschieden werden.

- **Datentriangulation.** Kombination und Nutzung mehrerer Datenquellen, die zu unterschiedlichen Zeiten und Orten an verschiedenen Personen erhoben werden.
- **Untersuchertriangulation.** Einsatz verschiedener Beobachter oder Interviewer, um Verzerrungen durch die Person des Untersuchers auszugleichen.
- **Theorien-Triangulation.** Anwendung unterschiedlicher Theorien, Perspektiven oder Hypothesen auf denselben Forschungsgegenstand.
- **Methoden-Triangulation. „Within-Method“:** Verwendung verschiedener Auswertungsverfahren innerhalb einer Methode bzw. eines Messinstruments. **„Between-Method“:** Einsatz verschiedener Methoden der Datengewinnung.

### 6.1.2.2. Konsequenzen der Gütekriterien

Die unterschiedlichen Gütekriterien von quantitativer und qualitativer Forschung führen auf unterschiedlichen Stufen des Forschungsprozesses zu unterschiedlichen Herangehensweisen.

In der mittleren Spalte der folgenden Darstellung sind die quantitativen Schritte dargestellt. Daneben wird deren mögliche Erweiterung durch qualitative Ansätze angedeutet.

Eine weitere Konsequenz der unterschiedlichen Gütekriterien sind unterschiedliche Richtlinien für die Veröffentlichung in den verschiedenen Schulen.

Sowohl für quantitative als auch für qualitative Forschung gilt:

- Expliziter wissenschaftlicher Kontext und Zielsetzung
- Angemessene Methodenauswahl
- Respekt gegenüber dem Probanden
- Spezifizierung der Methoden
- Angemessene Diskussion
- Klarheit der Präsentation
- Mehrwert für die Forschercommunity

Ausschließlich für qualitative Forschung gilt zusätzlich:

- Berücksichtigung eigener Perspektive (Autorensicht)
- Umfassende Beschreibung der Stichprobe
- Einbezug von Beispielen
- „Credibility Checks“, z.B. Triangulation
- Kohärenz der Darstellung
- Unterscheidung zwischen genereller vs. spezifischer Zielsetzung
- Auslösung positiver Resonanz beim Leser (aha…)

### 6.1.3. Abgrenzung qualitativ/quantitativ – sinnvoll?

Es wurde bereits kurz erwähnt, dass qualitative und quantitative Ansätze häufig kombiniert werden. Es stellt sich dabei die Frage, inwiefern die Abgrenzung der beiden Ansätze im Sinne von völlig gegensätzlichen Vorgehensweisen sinnvoll ist.

[Anmerkung: „Diese Folien sind uralt und existieren schon so lange wie unser Institut. Ich zeig sie trotzdem gerne, weil sie wirklich gut sind.“]

- **Grundorientierung.** Quantitativen Ansätzen wird eine nomothetische Vorgehensweise zugeschrieben, also die Beschreibung von Erleben und Verhalten durch Naturgesetze. Qualitative Forschung ist dagegen idiographisch orientiert und konzentriert sich auf die Interpretation von einmaligen Ereignissen und Sachverhalten.
- **Wissenschaftsverständnis.** Quantitative Forschung trennt Alltagstheorien und Wissenschaft und nimmt eine objektiv erfassbare Realität an. Qualitative Forschung betont die Ähnlichkeit von Alltagstheorien und Wissenschaft sowie das Subjekt als Konstrukteur seiner Wirklichkeit.
- **Forschungszweck.** Quantitativ: Theorieprüfung, qualitativ: Theorieentwicklung.
- **Forschungslogik.** Quantitativ: Deduktives Vorgehen; aus allgemeinen Theorien werden Hypothesen abgeleitet, die falsifiziert werden können. Qualitativ: Induktives Vorgehen; vom Besonderen zum Allgemeinen. Aus Beobachtungen werden Theorien generiert.
- **Untersuchungsdesign.** Quantitativ: Große Zufallsstichproben, Kontrollgruppen, Experimente, Statistik, Messung auf möglichst hohem Skalenniveau. Qualitativ: Einzelfallstudien, Feld- und Handlungsforschung, Inhaltliche statt statistischer Repräsentativität, quasi-statistische Messungen.

Es ist jedoch nicht sinnvoll, die Strategien als dichotom aufzufassen. Vielmehr sollten sie als Pole einer gemeinsamen Dimension aufgefasst werden. Dabei ist zu beachten, dass Forschungsziele und Forschungsmethoden zusammenhängen – bei Pilotstudien sollte z.B. eher qualitativ vorgegangen werden – und auch der Untersuchungsgegenstand bestimmt das Methodeninventar.

Es ist also nicht verboten, in einem qualitativen Ansatz zu messen (quantifizieren) und ein quantitativer Ansatz kann auch qualitative Techniken enthalten.

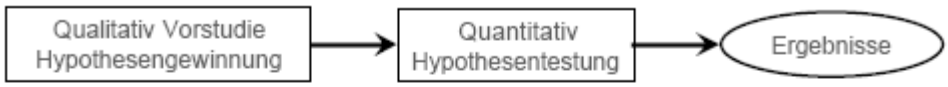
#### 6.1.3.1. Integration qual. & quant. Forschung

Mayring (2001; Qualitativer Guru) schlägt daher eine Integration der beiden Ansätze auf 4 verschiedenen Ebenen vor: Technische, Daten-, Personen- und Designebene.

- **Technische Ebene.** Auch in qualitativen Ansätzen ist bspw. eine **computergestützte Analyse** möglich. Dabei wird jedoch nicht wie in SPSS die komplette Analyse vom PC übernommen. Vielmehr können verschiedene qualitative Techniken, z.B. Transkribieren von Interviews, von Computern schneller und besser bewältigt werden als von Menschen (bspw. 120 Seiten Begriffslänge zählen).
- **Datenebene.** Die **qualitative Inhaltsanalyse** ermöglicht eine **induktive Kategorienbildung** – Beobachtungskategorien werden als am Untersuchungsmaterial gebildet. Mit einem Kodierleitfaden ist schließlich eine **deduktive Anwendung** der so gewonnenen Kategorien möglich.
- **Personenebene.** Typisierung: Bei **Einzelfallstudien** sollte der **Kontext** genau beschrieben werden, um eine Vergleichbarkeit über die jeweiligen Rahmeninformationen zu schaffen. Somit lassen sich ebenfalls allgemeine Gesetze ableiten (induktive Fallverallgemeinerung).
  Bei Interviews sollte bspw. nicht nur das Gesagte, sondern auch das „Wie“ erfasst werden, um möglichst genaue Informationen zu erhalten.
  Eine weitere Lösung der geringeren Verallgemeinerbarkeit von Einzelfallstudien ist das **theoretische Sampling**: Es werden so lange neue Probanden erhoben, bis der Untersucher mit seinem Erkenntnisstand zufrieden ist (theoretisches Sampling). Dagegen steht das statistische Sampling (Stp-Größe von vorneherein festgelegt).
- **Designebene.** Auf Designebene lassen sich verschiedene Kombinationen der beiden Ansätze finden. Häufig wird das **Vorstudienmodell** verwendet, sehr attraktiv ist jedoch auch das **Triangulationsmodell**, in dem qualitative und quantitative Ansätze gleichberechtigt nebeneinander stehen.
  Daneben existieren weitere Modelle, wie das **Vertiefungsmodell** oder das **Verallgemeinerungsmodell**. Die einzelnen Kombinationen werden auf der folgenden Seite beschrieben.
  In gewisser Weise stellen auch verbale Nachbefragungen nach einem Experiment einen Hauch von qualitativem Vorgehen dar.

Verschiedene Kombinationsmöglichkeiten auf Designebene:

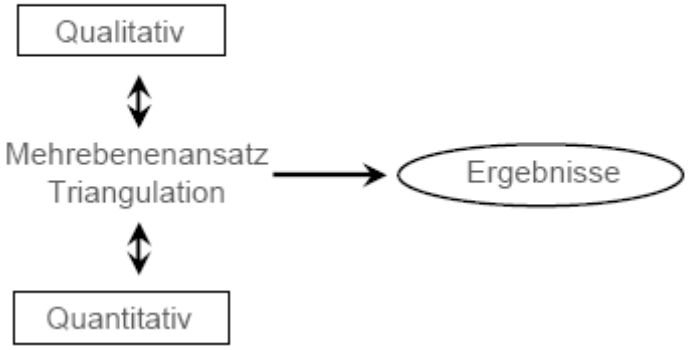
**Vorstudienmodell:**




**Verallgemeinerungsmodell:**


**Vertiefungsmodell:**


**Triangulationsmodell:**




## 6.1.3.2. Wann quantitativ, wann qualitativ?

Es lassen sich Empfehlungen aufstellen, wann welcher Ansatz angebracht ist.

| Quantitatives Vorgehen | Qualitatives Vorgehen |
|---|---|
| • Fragestellung spezifizierbar / klar formulierbar | • Untersuchung komplexer Fragestellungen in natürlicher Umgebung<br>• Fehlende Klarheit der Fragestellung bzw. zentraler Annahmen |
| • Theoretische/empirische Erfahrungen liegen vor (z.B. Literatur)<br>• Einschlägige quantitative Ansätze liegen vor<br>• (Standard-)Verfahren zur Datengewinnung verwendbar | • Fehlen einschlägiger Literatur bzw. theoretische und/oder empirische Ansätze werden Komplexität der Fragestellung nicht gerecht |
| • Kenntnisse und Erfahrungen im Umgang mit quantitativen Analysen (z.B. Deskriptive Statistik, Inferenzstatistik) | • Kenntnisse und Erfahrungen im Umgang mit qualitativer Forschung (auch: Anforderung an Betreuer!) |
| | • Verfügbarkeit von Geld und Zeit<br>• Umsetzung der „qualitativen Methodologie" und „Philosophie"<br>• Bestandteil ist Analyse von Sprache |

#### 6.1.3.3. Ausblick: Qualitatives Vorgehen

Gobo (2005) nennt einige Punkte, in denen sich die qualitative Forschung vermutlich in den nächsten Jahren stark weiterentwickeln wird. Darunter finden sich u. a. der verstärkte Einsatz von Computern bei der Datenauswertung (z.B. bei der Inhalts- oder Textanalyse von verbalem Material), eine gewisse Formalisierung der Methoden und die Entwicklung von standardisierten Datenanalysemethoden (Annäherung an das quantitative Vorgehen).

Prominente und vielversprechende Anwendungsfelder der qualitativen Techniken sind kulturvergleichende Fragestellungen und die Anwendungsforschung.

Bsp. „Kulturvergleichende Studien“: Hinsichtlich technischer Systeme brauchen Amis eigentlich nur einen Hilfebutton, der sie mit der Hotline verbindet, damit sie ja nicht selbst denken müssen. Deutsche wollen Luxus und einen Teuer-Look, dagegen ist typisch für Japaner: „Gebt mir mehr Menüebenen; je komplizierter desto besser“. Das ist wohl immer so ☺.

Bsp. „Anwendungsforschung“: Psychoanalytisches Institut für Verkehrspsychologie in Köln mit seinen Fahrertypen…sowas gibt’s wohl auch…

## 6.2. Qualitative Techniken

Im Folgenden sollen bestimmte Teilaspekte der qualitativen Methoden genauer betrachtet werden. Neben grundlegenden Kennzeichen zirkulärer Forschung werden verschiedene Versuchsdesigns, Erhebungstechniken, Dokumentationsverfahren und Auswertungstechniken betrachtet.

### 6.2.1. Kennzeichen zirkulärer Forschung

Die drei wichtigsten Kennzeichen zirkulärer Forschung sind:

- Theoretisches Sampling
- Theoretisches Kodieren
- Schreiben der Theorie

#### 6.2.1.1. Theoretisches Sampling

Die Fall- und Fallgruppenauswahl im Bereich der qualitativen Forschung erfolgt nach der Methode des theoretischen Samplings (vs. statistisches Sampling bei quantitativen Ansätzen).

Theoretisches Sampling kennzeichnet sich als Prozess der Datensammlung, bei dem Daten gleichzeitig gesammelt, kodiert und analysiert werden. Der Forscher entscheidet zu jedem Zeitpunkt, welche Daten als nächstes gesammelt werden sollen und wo sie zu suchen sind.

Dabei bildet sich im Laufe der Untersuchung Schritt für Schritt eine Theorie aus.

Kontrastierung von theoretischem und statistischem Sampling:

| Theoretisches Sampling | Statistisches Sampling |
|---|---|
| - Umfang der Grundgesamtheit ist vorab unbekannt. | - Umfang der Grundgesamtheit ist vorab bekannt. |

| | |
|---|---|
| - Merkmalsverteilung der Grungesamtheit ist vorab nicht bekannt.<br>- Mehrmaliges Ziehen von Stichprobenelementen nach neu festzulegenden Kriterien möglich.<br>- Stichprobengröße vorab nicht definiert.<br>- Sampling beendet, wenn theoretische Sättigung erreicht ist (d. h. sich nichts mehr Neues ergibt). | - Merkmalsverteilung in der Grundgesamtheit ist abschätzbar.<br>- Einmalige Ziehung einer Stichprobe nach vorab festgelegtem Plan.<br>- Stichprobengröße vorab definiert.<br>- Sampling beendet, wenn die gesamte Stichprobe untersucht wurde. |

### 6.2.1.2. Theoretisches Kodieren

Das theoretische Kodieren bezeichnet den Prozess in dem Daten verarbeitet und konzeptualisiert werden. Es ist der zentrale Prozess, durch den die Theorie aus Daten entwickelt wird. Wichtig ist hierbei, dass die Kategorien der Datenklassifikation am Datenmaterial entwickelt werden.

Der Prozess des theoretischen Kodierens lässt sich wiederum in drei Stufen untergliedern: Offenes, axiales und selektives Kodieren.

Zunächst werden redundante Informationen aus den Daten entfernt (**offenes Kodieren**). Das Ziel hierbei ist die begriffliche Fassung von Daten und Phänomenen. Anschließend werden die durch das offene Kodieren gebildeten Kategorien verfeinert und differenziert (**axiales Kodieren**). Zudem werden die reduzierten Informationen miteinander in Beziehung gesetzt, sodass ein Modell entwickelt werden kann.

Im dritten und letzten Schritt (**selektives Kodieren**) wird eine Kernkategorie (zentrales Phänomen) herausgearbeitet, welche die Informationen bestmöglich zusammenfasst und zur Gruppierung der anderen Kategorien verwendet werden kann.

Ein Beispiel aus der IZVW-Forschung: Im Auftrag von BMW wurden 64 Autofahrer in ausführlichen Interviews befragt, was ihnen am Autofahren Spaß macht.

Die vielen erhaltenen Informationen wurden zunächst per offenem Kodieren zu gemeinsamen Begriffen reduziert. Dann wurden die einzelnen Informationen zu Gruppen zusammengefasst (axiales Kodieren). Dabei zeigten sich etwa Gruppen wie Streckenmerkmale (wie Landstraße…; weitaus am häufigsten genannt), Verkehrsbedingungen, Sichtbedingungen, Fahrzeugausstattung und Anforderungen. Diese Kategorien waren wenig spektakulär, sodass eine alternative Kodierung vorgenommen wurde, die in den drei Kategorien Sicherheit, Freiheit und Dynamik resultierte, wobei Dynamik deutlich am häufigsten genannt wurde.

Dies zeigt, dass gänzlich andere Ergebnisse resultieren können, wenn man beim axialen Kodieren das Kriterium wechselt.

Im letzten Schritt (selektives Kodieren) wurde dann ein Konzept für BMW entwickelt.

Insgesamt fragt sich jedoch, inwiefern dass theoretische Kodieren tatsächlich eine handfeste Methode ist, oder ob es eher als Kunstlehre betrachtet werden sollte. Potentiell gibt es unendlich viele Kodierungs- und Vergleichsmöglichkeiten und keine klar definierten Auswahl- und Abbruchkriterien. Ähnliche Kritik gilt für das theoretische Sampling.

#### 6.2.1.3. Schreiben der Geschichte

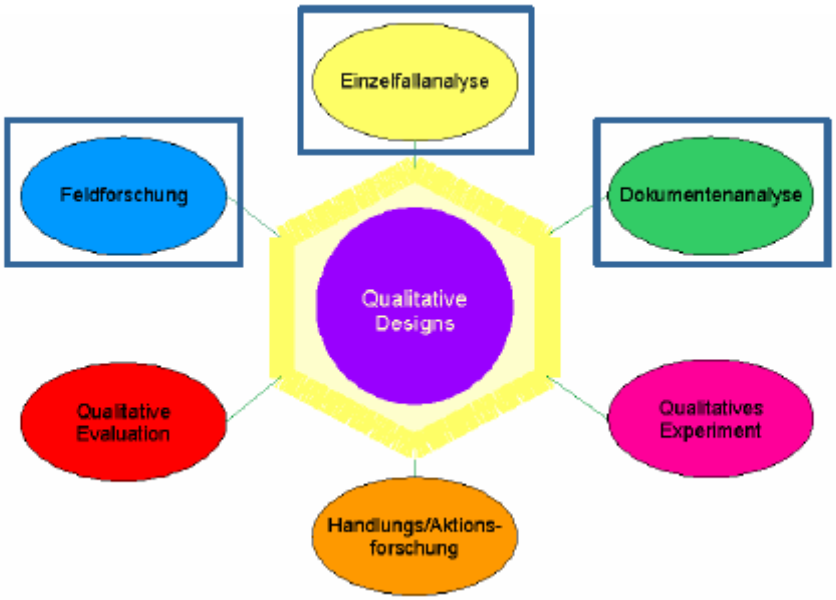
Die Darstellung qualitativer Theorien verlangt eine klare analytische Geschichte sowie eine klare Spezifizierung von Beziehungen zwischen den Kategorien.

Dies bedeutet die Skizzierung eines logischen Entwurfs einer Theorie sowie die zwingende Verdeutlichung anhand visueller Darstellungen.

Bisher waren die Kriterien also sehr schwammig. Etwas härter wird bei verschiedenen qualitativen Versuchsdesigns.

### 6.2.2. Versuchsdesigns

Mayring (1990; super Lehrbuch ☺) nennt 6 qualitative Versuchsdesigns:




Hier werden Dokumentenanalyse, Einzelfallanalyse und Feldforschung behandelt. Für qualitative Evaluation, Handlungs-/Aktionsforschung und qualitative Experimente sei auf Mayring (1990) verwiesen.

#### 6.2.2.1. Dokumentenanalyse

Als Dokumente zählen zwar vor allem verbale Informationsquellen, jedoch können auch andere Objekte wie z.B. Filme, Werkzeuge oder Kunstgegenstände, in eine Dokumentenanalyse einbezogen werden. Wichtig ist, dass nur nicht-individuelle Dokumente berücksichtigt werden. Tagebücher und ähnliches fallen unter das Design der Einzelfallanalyse.

Dokumente werden hier als Objektivationen (Vergegenständlichungen) der Psyche des Urhebers angesehen. Vorteile der Dokumentenanalyse ist die Vielfältigkeit potenzieller Analyseobjekte sowie die Nicht-Reaktivität des Ansatzes.

Anwendungsgebiete sind vor allem Forschungsthemen, bei denen kein direkter Zugang durch befragen, messen oder beobachten möglich ist.

Ein wichtiges zusätzliches Konzept der Dokumentenanalyse ist der Erkenntniswert, also potentielle Informationsquellen neben den enthaltenen verbalen Informationen:

1.) Art des Dokuments: Urkunden und Akten gelten als zuverlässiger als Zeitungsberichte, die politisch motiviert/zensiert sein können.
2.) Äußere Merkmale wie das Material und der Zustand eines Objekts sind ebenfalls aufschlussreich. Ein Buch mit kostbarem Einband war z.B. vermutlich bedeutsam.
3.) Innere Merkmale: Hierunter fällt v. a. der Inhalt verbaler Dokumente, jedoch auch die Aussagekraft anderer Quellen.
4.) Intendiertheit des Dokuments: Bei Büchern sollte bspw. der Herausgeber beachtet werden; ein Dokument der APA ist vermutlich vertrauenswürdiger als ein Medizinjournal eines Pharmakonzerns. Weiterhin sollten Dokumente, die explizit für die Nachwelt geschaffen wurden kritisch betrachtet werden.
5.) Nähe des Dokuments zum Gegenstand: Relevanz.
6.) Herkunft des Dokuments: Wo wurde ein Dokument gefunden, wo kommt es her und wie wurde es überliefert?

#### 6.2.2.2. Einzelfallanalyse

Ziel ist die umfassende Untersuchung eines Falles in seiner Gesamtheit. Auch hier werden vor allem sprachliche Informationen verwendet.

Hauptproblem ist die subjektive Verzerrung (v. a. bei biographischen Fallanalysen). Abhilfen sind die Prüfung der internen Konsistenz oder der Vergleich mit anderen Quellen (bei biographischen Untersuchungen etwa mit Tagebüchern von Personen, die am selben Ort aufgewachsen sind).

#### 6.2.2.3. Deskriptive Feldforschung

Grundgedanke ist die Untersuchung von Verhalten in möglichst natürlichen Kontexten. Dabei sollen vor allem Verzerrungen durch den Eingriff der Untersuchungsmethoden vermieden werden.

Ein Beispiel ist die IZVW-Untersuchung zu Drogenfahrten in Bayern von 1994–1997. Hierfür haben sich Hiwis in Nürnberg, München und Würzburg vor Diskotheken gestellt (hier: Labby, Airport) und herauskommende Personen um die Teilnahme an der Untersuchung gebeten.

Eine andere Alternative ist, dass Polizisten bei Routinekontrollen auch um die Teilnahme an einem kurzen Interview zu Forschungszwecken bitten.

Selbstverständlich ist Feldforschung nur dann anwendbar, wenn das Feld ohne große Störungen zugänglich ist. Der Forscher muss dabei seine Untersuchungen durchführen können, ohne die ablaufenden Prozesse zu beeinflussen. Hierfür ist eine gesonderte Schulung nötig. Auch muss das Vorhaben ethisch gerechtfertigt sein.

Größtes Problem: Störvariablen. Die Untersuchungen liefern also „im günstigsten Fall mitteleindeutige Ergebnisse“.

## 6.2.3. Erhebungstechniken

Qualitative Erhebungstechniken beziehen sich in erster Linie auf verbale Daten (wer hätt's gedacht) lassen sich grob in drei große Gruppen einteilen: **Leitfadeninterviews, Gruppenverfahren** und **Erzählung als Zugang**.


Die einzelnen Verfahren sollen hier nicht in epischer Breite behandelt werden. Die Grobgliederung ist jedoch relevant. Details (siehe Tabelle auf F 54) finden sich im Lehrbuch von Flick (1995; Guru 2 neben Mayring).

### 6.2.3.1. Leitfadeninterviews

Leitfadeninterviews sind eine Unterform der offenen, halbstandardisierten Interviews. Der Befragte kommt frei zu Wort, wird aber auf die vorgegebene Frage- bzw. Problemstellung explizit hingewiesen.

Anwendbar sind Leitfadeninterviews also vor allem bei Themen, für die bereits Vorwissen existiert. Ein großer Vorteil ist die Vergleichbarkeit der erhaltenen Daten.

### 6.2.3.2. Gruppenverfahren

Viele subjektive Bedeutungsstrukturen zeigen sich vor allem in sozialen Kontexten. Gruppenverfahren eignen sich dabei vor allem zur Erhebung kollektiver Einstellungen, Ideologien und Vorurteilen.

### 6.2.3.3. Erzählung als Zugang

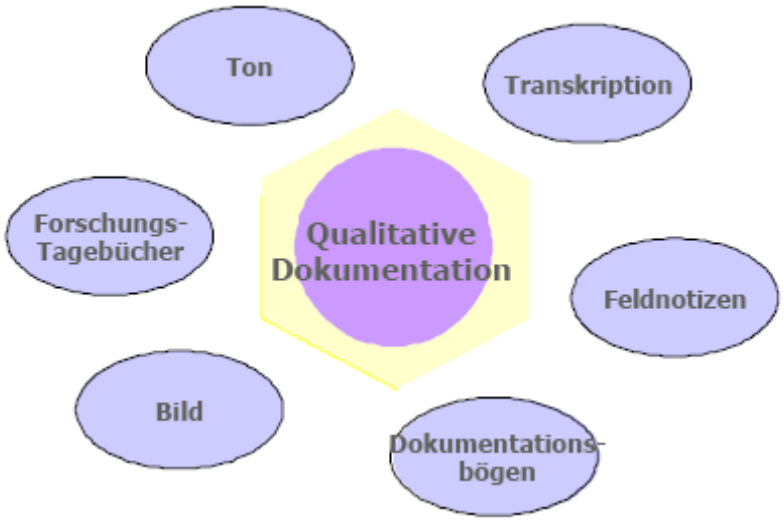
Die Erzählung als Zugang (narrative und episodische Interviews) ist eine weniger standardisierte Befragungsvariante. Über freies Erzählen von Geschichten sollen subjektiven Bedeutungsstrukturen ermittelt werde, die durch systematische Abfragen nicht erfasst werden können.

Bei einer Erzählung als Zugang erfolgt vom Interviewer lediglich ein Erzählanstoß. Nach der Haupterzählung erfolgt eine Nachfrage- und Bilanzierungsphase.

Anwendungsgebiete sind Themen mit starkem Handlungsbezug und explorative Fragestellungen. Wichtig ist hierbei immer Handlungsbezug: Nicht „Kohl vs. Merkel" (→Leitfaden) sondern **eigene Erlebnisse** (episodisch).

## 6.2.4. Dokumentationsverfahren

Aus den 6 prominentesten Dokumentationsverfahren der qualitativen Forschung soll nur die Transkription näher behandelt werden.




Unter **wörtlicher Transkription** versteht man die Herstellung einer vollständigen Textfassung von verbalem Material. Hierbei sollte nicht nur der semantische Inhalt erfasst werden, sondern auch die Dialektfärbung, Spracheigenheiten etc. Hierfür existiert ein internationales phonetisches Alphabet, wobei alternativ auch eine literarische Umschrift verwendet werden kann.

Über das Wortprotokoll hinaus können weitere Informationen über die **kommentierte Transkription** erhalten werden. Dabei werden Pausen, Betonungen etc. durch Sonderzeichen symbolisiert und evtl. zusätzliche Kommentare eingefügt.

Das Schema von Kallmeyer und Schütze (1976) konnte Ingo nur in einem Lehrbuch finden. Wesentlich häufiger ist das Jefferson Transkriptions-System von Howitt und Cramer (2005). Hier werden beispielsweise Unterbrechungen bzw. gleichzeitiges Reden über eckige Klammer dargestellt.

Vorlesungsbeispiel: Verhör (Suspect und Untersucher; Suspect hat wohl einer Elfjährigen irgendetwas Sexuelles gezeigt): Ohne kommentierte Transkription kann man sich kein besonders gutes Bild machen, mit Kommentaren wirkt das Ganze wesentlich lebendiger.

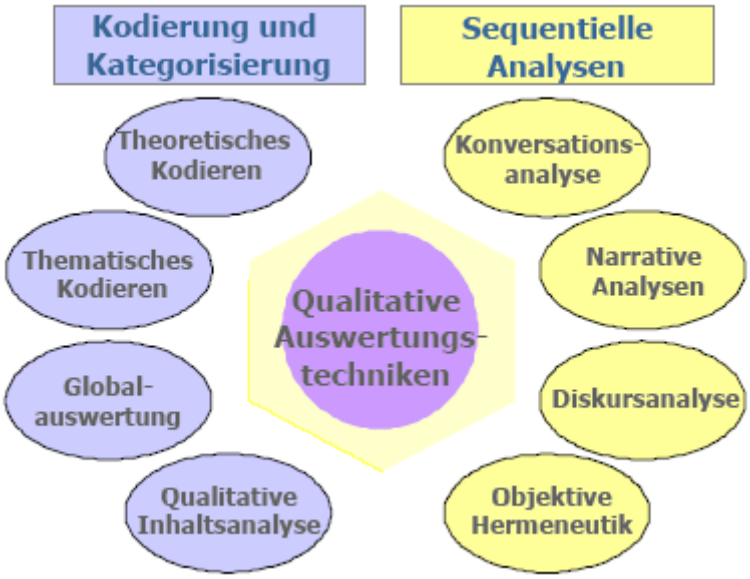
Insgesamt ist die Erstellung einer kommentierten Transkription also extrem aufwendig, dafür sind aber auch jede Menge Informationen enthalten.

## 6.2.5. Auswertungstechniken

Nach allgemeinen Kennzeichen, Versuchsdesigns, Erhebungstechniken und Dokumentationsverfahren sollen nun auch qualitative Auswertungstechniken abschließend betrachtet werden.

Auch hier gibt es eine Vielzahl verschiedener Methoden, wobei sich diese grob **in Kodierung und Kategorisierung** auf der einen und **Sequentielle Analysen** auf der anderen Seite untergliedern lassen.

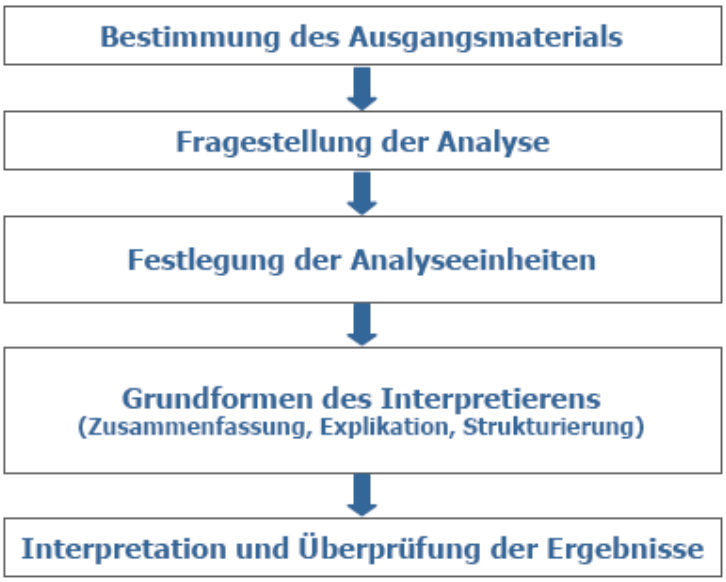
Hier wird als Beispiel für Kodierung und Kategorisierung die **qualitative Inhaltsanalyse** schematisch behandelt. Sequentielle Analysen werden vollständig ausgeklammert.




Die qualitative Inhaltsanalyse (auch: Aussagenanalyse, Contentanalyse) stellt eine methodisch kontrollierte Analyse von sprachlichem Material oder Texten dar. Dabei wird das Material zergliedert und schrittweise bearbeitet. Analyseaspekte werden in Kategoriensystemen vorher theoriegeleitet festgelegt.

Dabei findet der Kontext von Textbestandteilen explizit Berücksichtigung. Ebenfalls in Abgrenzung zur quantitativen Inhaltsanalyse ist das Ziel der Aufdeckung latenter Sinnstrukturen, die nicht explizit im Text enthalten sind. Dies geschieht vor allem anhand markanter Einzelfälle.

Die qualitative Inhaltsanalyse folgt dabei immer einem festen Ablaufschema:




Dieser Ablauf soll an einer Studie von Ulich et al. (1985) zur Kognitiven Kontrolle in Krisensituationen an arbeitslosen Lehrern dargestellt werden. Hierfür wurden Interviews mit N = 75 arbeitslosen Lehrern durchgeführt, die in 20.000 Seiten transkribierten Materials endeten.

Nun folgte eine qualitative Inhaltsanalyse, die sich in die o. g. Schritte - Bestimmung des Ausgangsmaterials, Fragestellung der Analyse, Festlegung der Analyseeinheiten, Grundformen des Interpretierens und die endgültige Interpretation unterteilt.

### 6.2.5.1. Bestimmung des Ausgangsmaterials

Das Material waren die per Tonband aufgenommenen Interviews, die nach vorher festgelegten Regeln transkribiert wurden. Hierbei gilt es auch festzulegen, wann das Material erhoben wurde und welche Interviewsituation im Hintergrund steht.

### 6.2.5.2. Fragestellung der Analyse

Durch die Analyse sollten Aussagen über den emotionalen, kognitiven und konativen (handlungsbezogenen) Hintergrund der Kommunikatoren abgeleitet werden. Differenzierter sollten Erfahrungen über den „Praxisschock“ sowie dessen Beeinflussung des Selbstvertrauens erhoben werden.

### 6.2.5.3. Festlegung der Analyseeinheiten

Unterschieden wird zwischen Kodiereinheit, Kontexteinheit und Auswertungseinheit. Unter **Kodiereinheit** versteht man den kleinsten Materialbestandteil, der ausgewertet werden soll: Was ist der minimale Textteil, der unter eine Kategorie fallen soll?

**Kontexteinheit:** Was ist der größte Textbestandteil, der unter eine Kategorie fallen kann?

**Auswertungseinheit:** Welche Textteile werden jeweils nacheinander ausgewertet?

### 6.2.5.4. Grundformen des Interpretierens

Bisher wurde das Material nur sortiert – jetzt wird es anspruchsvoller. Ziel der Interpretation ist eine Reduktion des Materials auf wesentliche Inhalte in Form einer **Zusammenfassung**. Danach erfolgt die **Explikation (Kontextanalyse)**, bei der zusätzliches Material zur Deutung herangezogen wird, um das Textverständnis zu steigern sowie abschließend die **Strukturierung** des Materials.

#### Grundform I: Zusammenfassung

Der Ausgangstext wird auf eine überschaubare Kurzversion reduziert, die nur noch die wichtigsten Inhalte umfasst. Dabei soll also das Allgemeinheitsniveau vereinheitlicht und schrittweise erhöht werden. Dies geschieht in den drei Arbeitsgängen **Paraphrasierung**, **Generalisierung** und **Reduktion**.

Unter Paraphrasierung versteht man eine sprachliche Zusammenfassung des Materials, wobei eine einheitliche Sprachebene erzielt werden soll. Wenig relevante Textbestandteile werden dabei gestrichen. Die erhaltenen Bedeutungseinheiten werden nun in der Generalisierung auf das angestrebte Abstraktionsniveau gebracht und in der Reduktion schließlich gebündelt und integriert, wodurch eine Aussage konstruiert wird.

Dies ist ein weiteres Beispiel für eine Induktive Kategorienbildung: Die Kategorien und Aussagen werden hier am Material abgeleitet. [Widerspruch zu F 60: Das Kategoriensystem wird vorher festgelegt?].

### Grundform II: Explikation

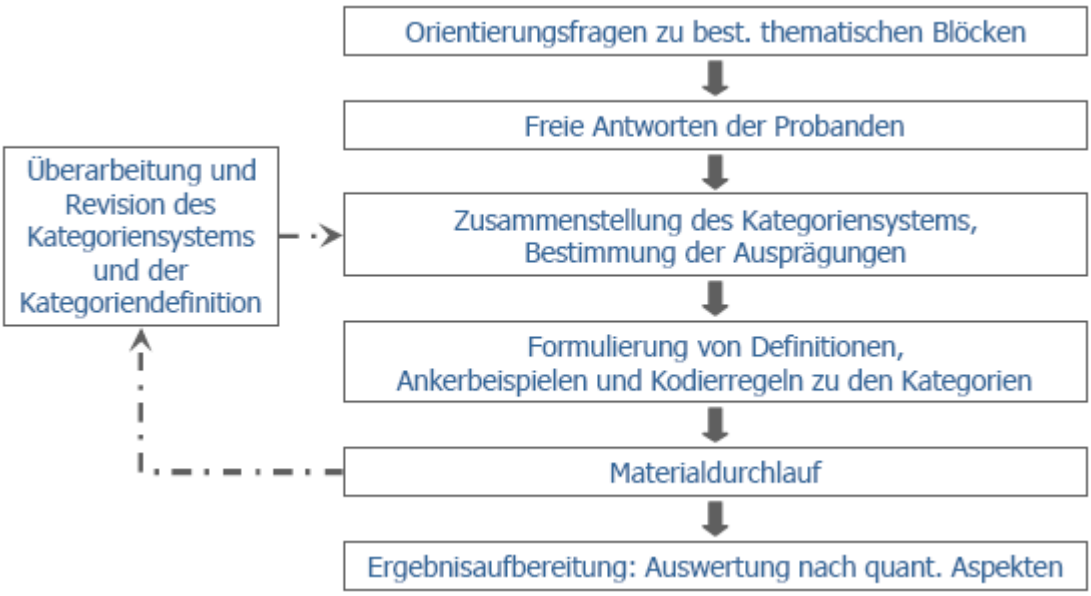
Im zweiten Schritt wird bei fraglichen Textteilen zusätzliches Material hinzugezogen. Man unterscheidet hier **enge Kontextanalyse** (Heranziehen von anderen Stellen im selben Text) und **weite Kontextanalyse** (über den Text hinausgehende Informationen).

### Grundform III: Skalierung/Strukturierung

Ziel dieses Schrittes ist das Herausfiltern einer bestimmten Struktur des Materials – also die Einschätzung aufgrund bestimmter Kriterien. Hierfür ist ein Kodierleitfaden mit genauen Regelungen zwingend erforderlich.

Man unterscheidet **formale** Strukturierung (z.B. anhand von Besonderheiten im Satzbau, **inhaltliche** Strukturierung, **typisierende** Strukturierung (markante Bedeutungsgegenstände) und **skalierende** Strukturierung.

Skalierende Strukturierung schätzt das Material auf einer Skala (i. d. R. ordinal) ein, um Inhalte auf einer interessierenden Variablen zu verankern. Auch hier kann ein standardisierter Ablaufplan mit Rückkopplungsschleifen aufgestellt werden.




## 6.3. Schlussbemerkung

Alle Ausführungen zu qualitativen Modellen sind sehr abstrakt und schwammig. Laut Ingo wird's erst dann spannend, wenn man es wirklich mal anwendet.

Wenn man das wirklich mal vorhaben sollte, so sind die Lehrbücher von Flick (1995) oder Mayring (1990) ein guter Leitfaden.

# 7. Anmerkungen

## 7.1. Allgemeines

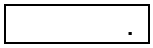
- Narratives Review: Meta-Analyse ohne Statistik („lesen und Bauch fragen“).
- Statistica gibt’s auf der Rechenzentrums-Homepage zum Download

## 7.2. Excel

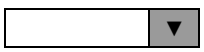
- Darstellung: Für Zahlen gibt es verschiedene Formate (z.B. normal vs. Datum). Das Programm rechnet also evtl. mit anderen Zahlen, als man in der Darstellung sieht.
- Absolute vs. relative Bezüge. Standardmäßig sind relative Bezüge eingestellt; wird die Formel verschoben, erfolgt eine ebensolche Verschiebung der Bezugsfelder. Absolute Bezüge können mit Dollarzeichen gesetzt werden (z.B. *$C$16*). Das Dollarzeichen vor dem Buchstaben setzt einen absoluten Zeilenbezug, das Dollarzeichen nach dem Buchstaben setzt einen absoluten Spaltenbezug.

## 7.3. SPSS

- Jede Vpn eine Zeile (Fall bzw. Case)
- Missing Values werden durch einen Punkt im unteren rechten Bereich der Zelle dargestellt. Diese Zellen werden bei der Berechnung ausgeschlossen.




- Variablen umkodieren (z.B. String → Numerisch) unter *Transformieren*
- Manche SPSS-Module können auch mit Strings rechnen, die meisten funktionieren aber ausschließlich mit Zahlen.
- Wertelabels: Angeben, was für die Zahlencodes dargestellt werden soll (z.B. „männlich“ und „weiblich“ statt „0“ und „1“). Unter *Ansicht* kann eingestellt werden, ob Wertelabel oder Zahl dargestellt werden soll. Wie in Excel zeigt sich also, dass das was dargestellt wird und das mit dem das Programm rechnet nicht das gleiche sein muss.
- Steht für das Programm eine Zahl in einem Kästchen, so findet sich eine Labelanzeige im Auswahlmenü:




- ANOVAS finden sich unter *Analysieren\Allgemeines Lineares Modell*
- Spezialität: Bei t-Tests werden Gruppenwerte vergeben. Man kann so z.B. Bedingung 1 mit Bedingung 5 vergleichen.
- SPSS-Algorithmen lassen sich mit visual basic programmieren.
- Hilfe: Command Syntax Reference, wenn’s wirklich ernst wird („Was rechnet das Programm eigentlich?)
- Analyseschritte im Output protokollieren zu lassen ist ziemlich mächtig. Lässt sich unter *Extras* einstellen.
- Tabellen im Output lassen sich per Doppelklick umformatieren

# 8. Klausuranmerkungen

- Neu sind die generating classes bei den loglinearen Modellen
- Bei logistischer Regression: $\beta_0$ ist nur bei Kohortenstudien interpretierbar. Damit lässt sich die Wahrscheinlichkeit für das Auftreten eines Kriteriums auch nur auf Basis von Kohortenstudien bzw. bei bekanntem $\beta_0$ schätzen. (1000x gesagt).
- Metanalyse F46: Die einzelnen ES-Formeln sind nicht relevant. Viel wichtiger ist der Grundgedanke ES = Mittelwertsdifferenz / SD
- Metaanalyse F55: Auch diese Formeln braucht man nicht zu können.
- Qualitative Analysen: Vor allem die erste Sitzung (bis qualitative Techniken) ist relevant, hier vor allem die Unterscheidung qual./quant. Aus der zweiten Sitzung (qualitative Techniken) nur die Grundgedanken.

Fragen:

- KFA – Konfigurationshomogenitätstest F84 – hier müsste der Merkmalsvektor +-+ ebenfalls signifikant sein.

| | | | | | | |
|---|---|---|---|---|---|---|
| - | + | - | 38 | 11 | 23 | 7.03119938 |

- Inzidenzdichte, Beispielfolie: Person 10 wird ausgeschlossen weil sie von Anfang an krank war. Wird aber trotzdem mitgerechnet. Was jetzt?
- Metaanalyse: Was genau ist das Uniformitätsproblem?
- Lord: KFA-Module für R?
- Qualitative Forschung: F60 vs. F70. Bei der qualitativen Inhaltsanalyse wird das Kategoriensystem einmal induktiv gebildet (F70) und einmal deduktiv vorgegeben (F60). Hm.

- Warum werden bei der HKFA $2^t$-t-1 KFAn berechnet?
- Epidemiologie: Ist Letalität eine krankheitsspezifische Mortalität?
- Metaanalyse: Was genau ist das Uniformitätsproblem? Heißt das einfach nur, dass in den Primärstudien jeweils gleiche UVn/AVn usw. verwendet wurden, wenn man sie vergleichen will, bzw. dass die AVn zumindest für dasselbe Konstrukt stehen?
- QualAnal: F.29 – Regelgeleitetheit; spricht das nicht eher für lineare Forschungsstrategien?
- QualAnal: Der Lord hat in der ersten Vorlesung zur MWH-ANOVA mal kurz eine „qualitative Inhaltsanalyse, die sie ja alle auch spontan anwenden würden“ erwähnt… das sollte man sich wohl nochmal anschauen.