

Statistik

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: <https://github.com/rsling/VL-Deutsche-Syntax>

- 1 Statistik, Inferenz und probabilistische Grammatik
 - Probative Wissenschaft
 - Elemente der Empirie
 - Validität
 - Ronald A. Fisher, Wahrscheinlichkeit, Ereignisraum, Teetassen
- 2 Deskriptive Statistik
 - Deskriptive Statistik
 - Motivation
 - Skalenniveau
 - Zentraltendenz
 - Dispersionsmaße
 - Bivariate Statistiken
 - Konfidenzintervalle
- 3 Nichtparametrische Verfahren
 - Testverfahren für Zähldaten
 - Vierfelder- χ^2 -Unterschiedstest
 - Fisher-Exakt-Test
 - Effektstärke: Cramér's v und ϕ
 - Chancenverhältnis
 - Binomialtest
- 4 z-Test und t-Test
 - Übersicht
 - Wiederholungen
 - Logik von statistischen Tests
 - t-Test
 - t-Test mit einer Stichprobe
 - t-Test mit zwei Stichproben
- 5 ANOVA
 - ANOVA
 - Überblick
 - Graphische Einführung
 - Einfaktorielle ANOVA
 - Zweifaktorielle ANOVA
- 6 Freiheitsgrade und Effektstärken
 - Freiheitsgrade
 - Mehr zu Zähltesttests
 - Effektstärke für χ^2 : Cramér's v und ϕ
 - Chancenverhältnis
 - Binomialtest
 - Effektstärken bei t-Test und ANOVA
 - Ein-Stichproben-t-Test
 - Zwei-Stichproben-t-Test
 - ANOVA
 - Voraussetzungen für t-Test und ANOVA
 - Nichtparametrische Alternativen zu t-Test und ANOVA
 - Mann-Whitney U-Test
 - Kruskal-Wallis H-Test
- 7 Power
- 8 Lineare Modelle
 - Lineare Modelle
 - Korrelation und Signifikanz
 - Lineare Regression
 - Multiple Regression
 - ANOVA und LMs
 - In R
- 9 Generalisierte Lineare Modelle
 - Generalisierte Lineare Modelle
 - LM und GLM
 - GLM Grundlagen
 - Maximum Likelihood
 - Nominale Unabhängige
 - Modellselektion
 - Modellevaluation
 - Alternativen und Lösungen
 - In R
- 10 Gemischte Modelle

Statistik, Inferenz und probabilistische Grammatik

- beobachtbare Phänomene
- Beobachtungen reproduzierbar
- messbar = beobachtbar (Sinneswahrnehmung an sich irrelevant)
- Realismus | wirkliche Phänomene und ihre Mechanismen
- keine postmoderne Realitäts- und Objektivitätsverweigerung
- kontrolliertes Experiment

- intrinsische Ungenauigkeiten der Messung (Wirkung plus Störeinflüsse)
 - potentiell inadäquate Messung des theoretischen Konstrukts
- Vermeidung von Fehlschluss auf unechte Ursachen
- relevante Ursachen
- insgesamt Stärkung der Validität

- Gegenstand: interne (mentale) Grammatik (I-Grammatik)
universeller und individueller Teil
 - I-Grammatik bei jedem Sprecher (leicht) verschieden
 - I-Grammatik erlaubt immer binäre Grammatikalitätsentscheidung
- Linguisten können eigene I-Grammatik untersuchen (Introspektion)!

Das Ergebnis ist die aktuelle Krise der Linguistik.

(Logischer) Positivismus

Formale Ableitung von Wissen (= Theorien) aus Beobachtbarem und irgendeiner Logik.
Induktion. Keine Metaphysik. Keine Kreativität erwünscht. (Carnap 1928, ...)

Aber suchen wir wirklich nur nach **Mustern**, z. B. in Korpusdaten?

- Was ist der **zugrundeliegende Mechanismus**?
- Wie kommen wir zu **erklärenden Theorien** von Mustern in Daten?
- **Datenaufbereitung** (z. B. im Korpus) kann dann nicht theoriegeleitet sein.
- **Die ART folgt auch nicht einfach so aus Daten!**

Rationalistischer Probativismus

Theorien werden aufgestellt von **Menschen, die die Welt beobachten**. Theorien werden getestet an Daten, aber nicht logisch aus Daten abgeleitet. (Popper 1962, Mayo 1996, ...)

Unter dieser Philosophie werden plötzlich Dinge wichtig ...

- Ist eine **Stichprobe repräsentativ** für das, was man zeigen will?
- Welche **Methode der statistischen Analyse** wird verwendet?
- Für eine Korpusstudie muss die Datenaufbereitung damit theoriegeleitet sein!
- Liefert die Studie **a serious Argument from Error**?

*There is evidence an error is absent to the extent that a **procedure with a very high capability of signalling the error**, if and only if it is present, nevertheless detects no error.* (Mayo 2018: 16)

Die konkreten Hypothesen, die in einem Experiment getestet werden, sind **nie** die Primärhypothesen der Theorie.

- **abgeleitete Partikularhypothesen** über konkrete Erwartungen im Experiment
- Einfluss zahlreicher **Auxiliarhypothesen**, z. B. über Messprozeduren
Duhem (1914), Quine (1951), Laudan (1990)
- „interessante“ Hypothesen
 - ▶ Formulierung relevanter **Kausationsbedingung** (wenn, dann)
 - ▶ **universelle Gültigkeit** | ein Sprecher vs. alle Sprecher
 - ▶ also z. B. **uninteressant** | *Welchen Kasus nimmt wegen?*

Kann die Hypothese weiter angenommen werden,
oder liefert das Experiment starke Evidenz gegen sie?

- Probleme bei Prüfung
 - ▶ falsch abgeleitete Partikularhypothese
 - ▶ falsche Sekundärhypothesen
 - ▶ Störeinflüsse, intrinsische Messungenauigkeit
 - ▶ mangelhafte **Operationalisierung**
 - ▶ zu wenige Daten (oder zu viele Daten?)

- von Interesse | **allgemeine Gesetzmäßigkeiten**
- also Untersuchungsgegenstand: **alle x** (Sprecher, Sätze, ...)
- untersuchbar | kleine Menge von x

Grundgesamtheit | alle x

datengenerierender Prozess (DGP) | Prozess, der **alle x** hervorbringt

Stichprobe | eine kleine Menge x, aus der auf Grundgesamtheit
bzw. DGP geschlossen werden soll

uniform zufällige Stichprobe

jedes Element der Grundgesamtheit hat die gleiche Chance beim Ziehen

stratifizierte Stichprobe

Stichprobe so zusammengesetzt, dass wichtige Eigenschaften proportional repräsentiert sind

- Problem bei Letzterem: haufenweise Auxiliärhypothesen

- **Operationalisierung** | präzise Formulierung der Messmethode für ein theoretisches Konstrukt
- Bsp. Konstrukt „Satzlänge“: Wortanzahl? Phonemanzahl? Phrasenanzahl?
- Bsp. Konstrukt „Satztopik“: Oha!?! (Cook & Bildhauer 2013)
- alle genannten Beispiele: **abhängig von Auxiliarahypothesen** bzw. anderen theoretischen Konstrukten (Wort, Phonem, Phrase, ...)

- uninteressanter Typ Fragestellung | „Wieviel Prozent X haben Eigenschaft A?“
- **Fehlen jeglicher Aussagen über kausale Zusammenhänge**
- Bsp. | Wie oft wird *wegen* mit Dat bzw. Gen verwendet?
- Besser | „Wie bedingt Eigenschaft B die Wahrscheinlichkeit von A bei X?“
- Bsp. | Per Hypothese nehmen denominales Präpositionen eher den Gen als den Dat.

konzeptuell:

	denominale P	andere P
Dat	x_1	x_2
Gen	x_3	x_4

Operationalisierte und gemessene Eigenschaften sind Variablen.

- im Experiment:
 - ▶ kontrolliere für Theorie irrelevante Variablen (Störvariablen)
 - ▶ variiere „Ursachen-Variablen“ (unabhängige Variablen)
 - ▶ beobachte „Wirkung-Variablen“ (abhängige Variablen)

- Problem in Astronomie, Korpuslinguistik usw. | keine Experimente möglich
- unabhängige Variablen nicht variierbar
- Daten liegen bereits vor bzw. fallen vom Himmel
- Auswahl von Datensätzen, so dass von den unabhängigen Variablen die zur Theorieprüfung nötigen Permutationen im Datensatz vorkommen
- dabei Zusatzproblem bei Korpuslinguistik: Korpus meist nicht das eigene, wenig Informationen über mögliche Verzerrungen
- Was ist die Grundgesamtheit bzw. der DGP?

Gefahren für statistische Schlussverfahren

- mathematische Vorbedingungen für das Testverfahren nicht
- zu viele Partikultests einer übergeordneten Hypothese aus denselben Daten
- zu kleine Stichprobe
- zu große Variation in der Grundgesamtheit

- bei Korpora | schlechte Zusammensetzung des Korpus

- Irrtum beim **Herstellen des Kausalzusammenhangs**
- Fiktives Bsp.:
 - ▶ Hypothese | Im DECOW2012 kommt öfter das Pronomen *son* vor als im DWDS Kernkorpus, weil es erst nach 2000 zum eigenständigen Pronomen wurde.
 - ▶ Die Hypothese wird bestätigt anhand von Stichproben aus den beiden Korpora.
 - ▶ **Die wirkliche Ursache sind aber Registerunterschiede.**

- Korrektheit des **theoretischen Konstrukts**
- eigentlich aus der Psychologie
- aber riesiges Problem in der Linguistik
- Echtes Bsp.
 - ▶ Beobachtung | Das Deutsche bewahrt genus-typische Pluralflexion am Substantiv.
 - ▶ Konstrukt | Nominalklammer/Klammerprinzip (NP-Kongruenzklammer Art – Subst) (Ronneberger-Sibold 2010)
 - ▶ Hypothese zu Beobachtung | Flexionserhalt stärkt Klammerprinzip
 - ▶ **Das Konstrukt ist hochgradig beliebig und unterdefiniert, damit nicht testbar.**
 - ▶ **Abhilfe: nur Konstrukte/Hypothesen, die starke Vorhersagen generieren**

- Generalisierbarkeit der Ergebnisse (über Raum, Zeit usw.)
- Problem | zu große Homogenität der Stichprobe
(was für statistische Validität wiederum gut ist)
- Bezug auf Korpora:
 - ▶ zu spezifische Stratifikation (DeReKo)
 - ▶ verzerrte Stichprobe (Webkorpora)

- Statistik als Teil der rationalen wissenschaftlichen Argumentation, der Interpretation von Experimenten
- daher: möglichst kein Mathematik-Jargon
- eingeschränkte Induktion als theoriegeleitete Dateninterpretation
- Kontrolle aller unabhängigen Variablen
- alle anderen (Stör-)Variablen konzeptuell zufallsgebunden

- Behauptung: Dame X kann am Geschmack erkennen, ob der Tee oder die Milch zuerst in die Tasse gegossen wurde.
- prä-fishersches Konzept: **alle Störvariablen kontrollieren** und gleich machen, sonst keine valide Inferenz möglich
- Fisher: Das ist prinzipiell unmöglich, umständlich, teuer, **unnötig!**
- wenn alle irrelevanten Stör-Faktoren zufällig, dann:
 - ▶ Variiere die relevante unabhängige Variable.
 - ▶ Vergleiche das Ergebnis mit zufällig erwartbaren Ergebnissen.
 - ▶ **Wie unwahrscheinlich ist das erzielte Ergebnis unter der Zufallsannahme?**

- acht Tassen (zwei Milch zuerst, zwei Tee zuerst)
- Mit wie vielen richtigen Treffern wären Sie zufrieden?
- Es muss die Wahrscheinlichkeit errechnet werden, eine, zwei, drei oder vier Tassen richtig zu raten.
- Typischerweise schätzen Menschen solche Kombinatorikprobleme intuitiv falsch ein.

$$P(\text{richtig per Zufall}) = \frac{\text{Anzahl richtiger Zuweisungen}}{\text{Anzahl aller potentiellen Zuweisungen}} \quad (1)$$

- Anzahl richtiger Zuweisungen: 1
- mögliche Zuweisungen: einfaches kombinatorisches Problem

mögliche Zuweisung von acht Tassen zu Milch/Tee zuerst

- erste MZ-Tasse: eine von 8
- zweite MZ-Tasse: eine von 7
- dritte MZ-Tasse: eine von 6
- vierte MZ-Tasse: eine von 5
- fünfte MZ-Tasse: STOPP (automatisch TZ)
- Möglichkeiten, 4 Tassen aus 8 auszuwählen:
 $8 \cdot 7 \cdot 6 \cdot 5 = 1680$

- bisher: jedes Set von 4 MZ-Tassen ist in verschiedenen Reihenfolgen in der Menge der Möglichkeiten
- Möglichkeiten, vier Tassen zu ordnen:
 $4 \cdot 3 \cdot 2 \cdot 1 = 24$

- Reihenfolge hier egal, also:

$$\text{Anzahl aller potentiellen Zuweisungen} = \frac{1680}{24} = 70 \quad (2)$$

Wenn Dame X also genau richtig liegt

- per Zufall genau richtig:
in einem von 70 Fällen, $p = 0.014$
- α -Niveau von 0.05 also erreicht

- eigentlich **Binomialkoeffizient**
- „Lotto-Kombinationen“: k aus n
ohne Zurücklegen und ohne Beachtung der Reihenfolge

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (3)$$

- die drei richtigen: $\binom{4}{3}$
- die eine falsche aus vier TZ: $\binom{4}{1}$
- Anzahl der Möglichkeiten drei richtige aus vier MZ
und dann eine falsche aus vier TZ zu ziehen: $\binom{4}{3} \cdot \binom{4}{1} = 4 \cdot 4 = 16$

$$P(\text{drei richtig per Zufall}) = \frac{16}{70} = 0.229 \quad (4)$$

- Bei $\alpha = 0.05$ reichen also drei richtige nicht im Ansatz!
- alle schlechteren Ergebnisse: folglich auch nicht ausreichend
- Und bei 30 von 40 richtigen (also insgesamt 80 Tassen)?

Deskriptive Statistik

- deskriptive Statistik als Datenaggregation
- Verteilungen in Stichproben und Grundgesamtheiten:
 - ▶ Zentralmaße
 - ▶ Streuung (Varianz)
- Beziehungen zwischen ko-variiierenden Messungen
- Genauigkeiten von Schätzungen quantifizieren (Konfidenzintervalle)

- Gravetter & Wallnau (2007)
Achtung! Vermittelt eine falsche Philosophie!
Nur für die Mathematik benutzen.
- Bortz & Schuster (2010)

- Mit unbewaffnetem Auge auf Datenmengen zu blicken, ist meistens sinnlos.
- In großen Zahlenkolonnen sehen Menschen nur schlecht Tendenzen und Zusammenhänge.
- Um dies zu erleichtern, gruppieren und visualisieren wir die Daten.

- Definition und (geschätzte) Größe der Grundgesamtheit.
(z. B. alle lebenden deutschen Erwachsenen)
- Stichprobengröße (N)
- Stichprobenmethode
 - ▶ Zufallsstichprobe (größere Stichprobe)
 - ▶ proportional stratifizierte Stichprobe (Quotenstichprobe)

Variablen sind folgendermaßen **skaliert**:

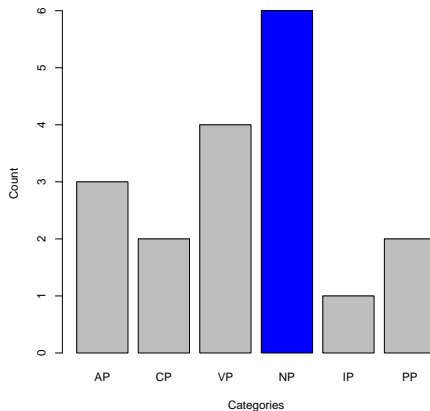
- **dichotom** (binär) = zwei Kategorien:
männlich, weiblich ; Präteritum, Perfekt
- **nominal** (kategorial) = disjunkte Kategorien ohne numerische Interpretation:
Parteizugehörigkeit ; NP, AP, VP
- **ordinal** = disjunkte Kategorien, nach Rang geordnet:
Schulnoten ; 5-point oder 7-point scales (Likert scales)
- **intervall~** = geordnete Werte mit definierten Abständen,
aber mit arbiträrem Nullpunkt: Celsius
- **verhältnis~** = wie intervall-,
aber der Nullpunkt ist ein echter Nullpunkt: Kelvin

- Wir messen die Größe von Menschen in cm auf einer Verhältnisskala.
 - ▶ 200cm sind das doppelte von 100cm.
 - ▶ Niemand kann kleiner sein als 0cm.
- Dieselbe Messung als **Abweichung vom Mittel** ergibt eine Intervallskala.
 - ▶ Wer 3 cm größer ist als der Durchschnitt ist doppelt soviel größer wie jemand, der 1.5 cm größer ist.
 - ▶ Die erste Person ist aber nicht doppelt so groß wie die zweite.
 - ▶ Außerdem kann man z.B. -3 cm vom Durchschnitt abweichen.

- Das SN bestimmt die **zulässigen mathematischen Operationen** (z.B. Rechenarten).
- Also kommen je nach SN nur bestimmte **deskriptive Statistiken** in Frage.
- Das gleiche gilt für die Zulässigkeit bestimmter **inferenzstatistischer Tests** je nach Skalenniveau.

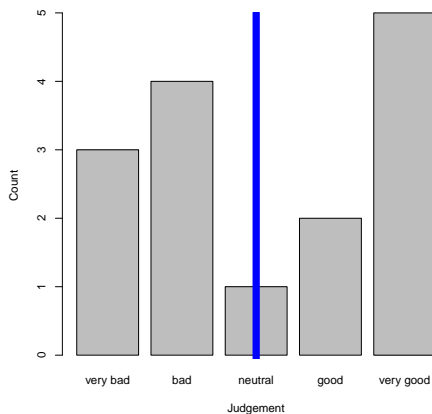
Zentraltendenz I

Der **Modus** ist der **häufigste Wert** in einer Grundgesamtheit oder Stichprobe.
Geht bei jedem Skalenniveau.



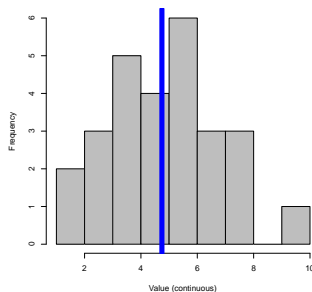
Zentraltendenz II

Der **Median** ist der Wert **über und unter dem gleichviele Werte liegen**. Ordinalskala oder höher.



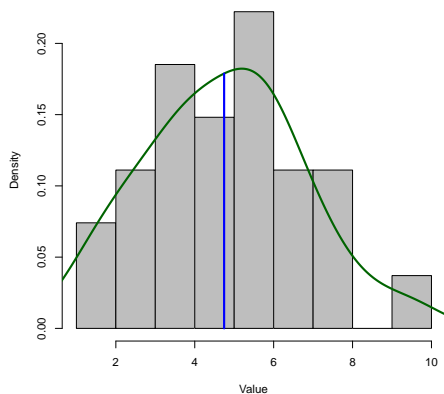
Das **arithmetische Mittel** \bar{x} ist die Summe aller Werte x dividiert durch Stichprobengröße n .
Intervallskala oder höher.

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



Zentraltendenz IV

Kontinuierliche Variablen und ihr arithmetisches Mittel lassen sich in **Dichteplots** gut visualisieren (per Software).

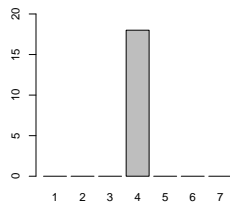
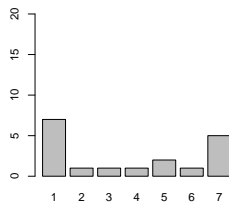
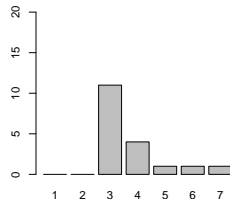
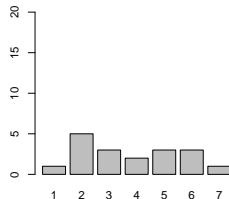


Warum sind Dispersionsmaße wichtig?

- 1 Das Wissen um die Zentraltendenz ist wichtig als grobe allgemeine Information über die Population.
- 2 Aber dieselbe Zentraltendenz kann das Ergebnis ganz verschiedener Werte sein.
- 3 Die Verteilung kann flach, chaotisch, glockenförmig usw. sein.

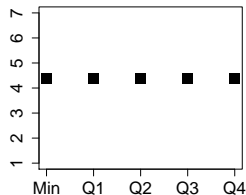
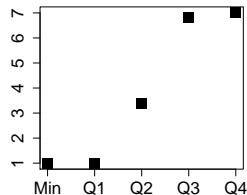
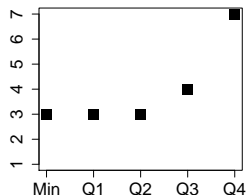
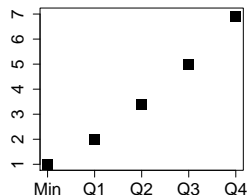
Verteilungsformen

Histogramme von vier Stichproben
mit $\bar{x} = 4.389$ und $n = 18$.



Quartile

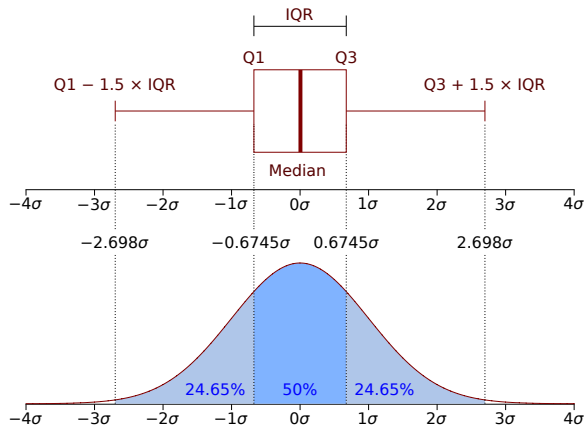
Quartile sind die Punkte, unterhalb derer 25%, 50%, 75% und 100% (Maximum) der Werte liegen. Dazu gibt es noch das Minimum (niedrigster Wert).



Quartile und Inter-Quartil-Bereich

$$IQR = Q_3 - Q_1$$

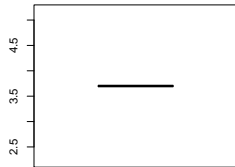
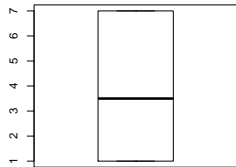
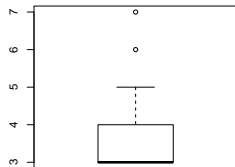
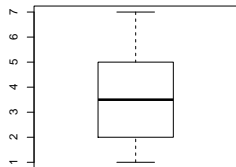
oder ganz einfach: die mittleren 50%



Attribution: Jhguch (<http://en.wikipedia.org/wiki/User:Jhguch>) at en.wikipedia

Boxplots als bessere Zusammenfassung

Boxplots zeigen Median (Linie in der Mitte), oberes und unteres Quartil (Boxen), 1,5-fachen Interquartilabstand zu diesen (gestrichelte Hebel) und Ausreißer (Punkte).



Die **Varianz** s^2 ist die quadrierte mittlere Abweichung vom Mittel:

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

Die **Standardabweichung** s ist die Quadratwurzel der Varianz:

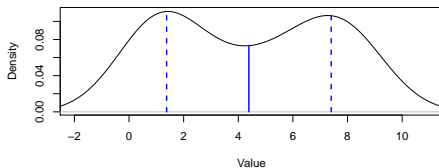
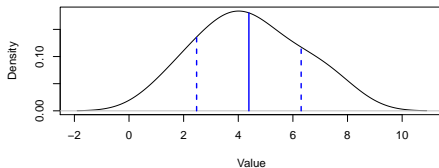
$$s(x) = \sqrt{s^2(x)}$$

Der Zählerterm der Varianz heißt auch **Summe der Quadrate**:

$$SQ(x) = \sum_{i=1}^n (x_i - \bar{x})^2$$

Unterschiedliche Stabw

Die erste Stichprobe hat $s = 1.91$,
die zweite $s = 3.01$ (beide $\bar{x} = 4.389$).



Um wie viele Standardabweichungen weicht jeder Datenpunkt vom Mittel ab?

Für jeden Punkt: $z(x_i) = \frac{x_i - \bar{x}}{s(x)}$

Bsp.: $x = [3.9, 4.3, 7.2, 8.5, 11.1, 12.1, 14.0, 20.7]$

$$\bar{x} = 10.225$$

$$s^2(x) = \frac{(3.9 - 10.225)^2 + \dots + (20.7 - 10.225)^2}{8 - 1} = \frac{215.495}{7} = 30.785$$

$$s(x) = \sqrt{30.785} = 5.548$$

$$z = \left[\frac{3.9 - 10.225}{5.548}, \dots, \frac{20.7 - 10.225}{5.548} \right] =$$
$$[-1.140, -1.068, -0.545, -0.311, 0.158, 0.338, 0.680, 1.888]$$

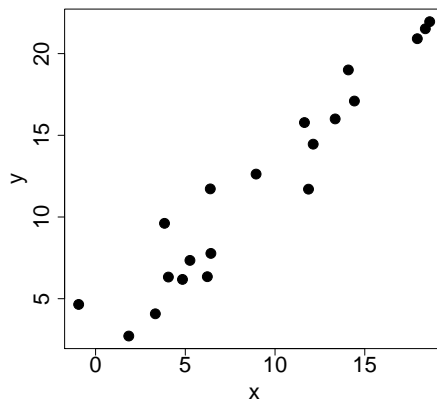
Zähldaten von zwei Variablen (egal wieviel Ausprägungen)
sind ideal als **Kreuztabelle** darstellbar.

	Variable 1: Wert 1	Variable 1: Wert2
Variable 2: Wert 1	Anzahl x_{11}	Anzahl x_{12}
Variable 2: Wert 2	Anzahl x_{21}	Anzahl x_{22}

Korrelationen

Korrelationskoeffizienten helfen, den Zusammenhang zwischen Variablen, die mindestens ordinalskaliert sind, numerisch zu erfassen.

Z. B. die hier geplotteten x und y:



Die Kovarianz kombiniert die Maße, zu denen die **zwei Messwerte** pro Datenpunkt vom **jeweiligen Mittel der Messwertreihen** abweichen.

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

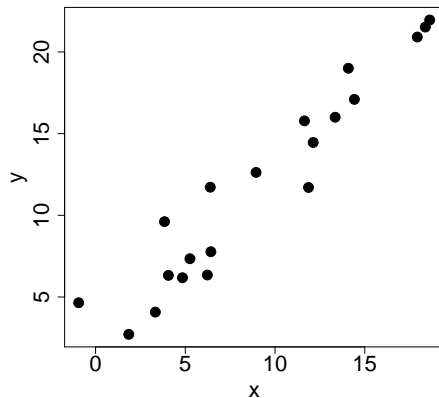
Sind $x_i - \bar{x}$ und $y_i - \bar{y}$ positiv oder negativ, ist der Beitrag ihres Produkts zur Kovarianz positiv, bei ungleichen Vorzeichen negativ.

Der Zählerterm heißt auch **Summe der Produkte**:

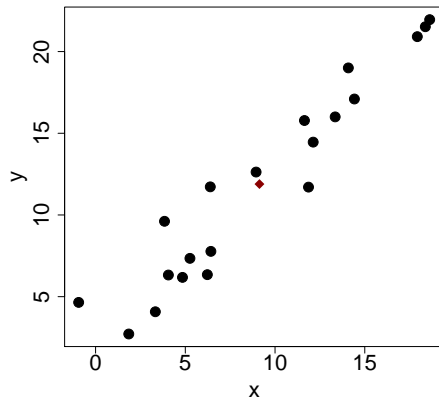
$$SP(x, y) = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Kovarianz: Illustration 1

Zwei Messvariablen (Vektoren): x und y



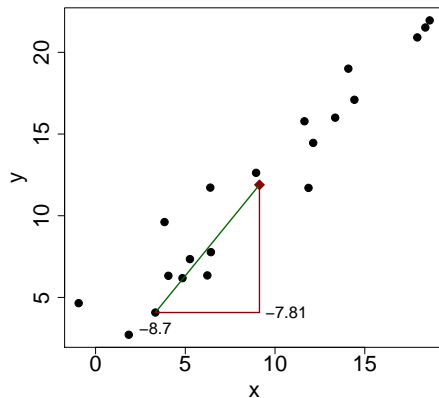
Koordinate von $\langle \bar{x}, \bar{y} \rangle$



Kovarianz: Illustration 3

Punktvarianzen: $x_3 - \bar{x} = -7.81$ und $y_3 - \bar{y} = -5.80$

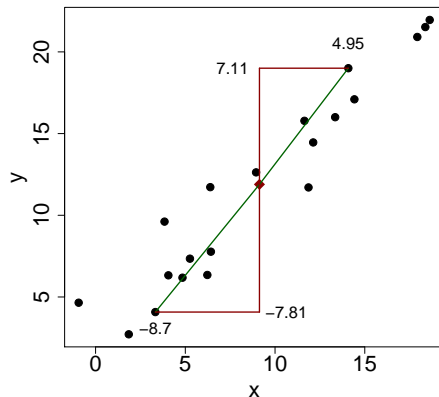
$$-7.81 \cdot -5.80 = 45.30$$



Kovarianz: Illustration 4

Punktvarianzen: $x_{17} - \bar{x} = 4.95$ und $y_{17} - \bar{y} = 7.11$

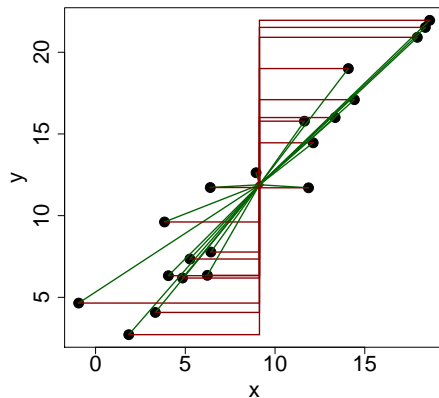
$$4.95 \cdot 7.11 = 35.19$$



Kovarianz: Illustration 5

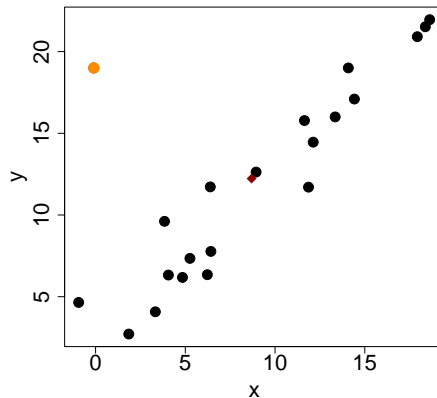
Puntvarianzen für alle $\langle x_i, y_i \rangle$

$$\text{cov}(x, y) = 34.52$$



Kovarianz: Illustration 6

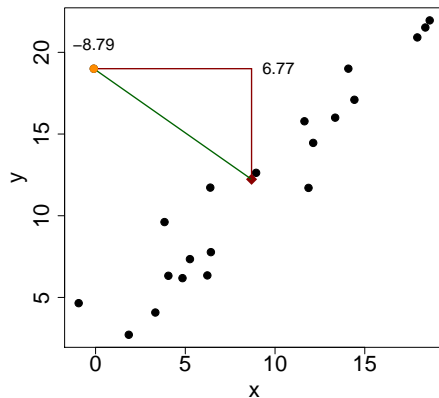
„Ausreißer“ bei – im Prinzip – positiver Kovarianz:
Negatives Produkt der Punktvarianzen



Kovarianz: Illustration 7

Punktvarianzen: $x_{21} - \bar{x} = 6.77$ und $y_{21} - \bar{y} = -8.79$

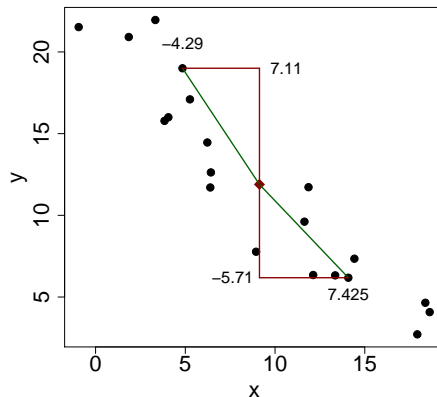
$$6.77 \cdot -8.79 = -59.51$$



Kovarianz: Negative Kovarianz

Wenn die Abhängigkeit zwischen den Werten tendentiell negativ ist, sind die Produkte der Punktvarianzen überwiegend negativ.

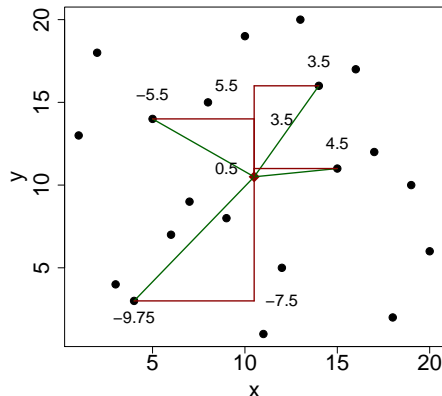
$$\text{cov}(x, y) = -33.77$$



Kovarianz: Null annähernd

Wenn es keine besondere Abhängigkeit gibt,
näht sich die Kovarianz 0:

$$\text{cov}(x, y) = -1.74$$



Während die Kovarianz **von der Größe der Werte** abhängt, macht der Korrelationskoeffizient Kovarianzen vergleichbar:

$$r(x, y) = \frac{\text{cov}(x, y)}{s(x) \cdot s(y)}$$

Dies ist die **Pearson-Korrelation**, später kommen noch andere Korrelationen.

- Das Verb *essen* kommt manchmal mit, manchmal ohne Akkusativ (direktes Objekt) vor.
- mit dO 39, ohne dO 61.
- Wenn wir in dieser Situation Stichproben mit $n=100$ ziehen, werden wir nicht immer genau diese Werte messen, sondern sie zwar häufig gut approximieren, manchmal aber auch stark abweichende Anteilswerte messen.
- In welchem Bereich liegen 95% aller Messwerte bei $n=100$?
- Diese Frage beantwortet das 95%-Konfidenzintervall.
- Es sagt uns, wie gut Stichproben einer bestimmten Größe bestimmte Anteilswerte approximieren.

- Annahme: Wahrer Anteilswert in der Grungesamtheit ist P .
- In Stichproben der Größe n misst man einen Stichprobenanteil p .
- Die meisten p liegen nah an P , sehr wenige weit weg davon.
- Wenn man beliebig viele p hat, verteilen sie sich so um P , dass eine Standardabweichung dem Standardfehler entspricht.
- Der Standardfehler ist der Erwartungswert für die Standardabweichung sehr vieler Messwerte (um den wahren Wert).
- Außerdem weiß man, dass die p normalverteilt um P sind.
Das folgt für groß genug Stichproben aus dem Zentralen Grenzwertsatz.
- Bei einer Normalverteilung weiß man, wieviel Prozent der Messwerte in einem Bereich $\pm q \cdot s$ (für beliebige q) vom Mittel liegen.

Wir brauchen also für Stichproben der Größe n den SF für den tatsächlichen Anteilswert P .

$$SF(P) = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$\text{Bsp. für } p = 0.39 \text{ und } n = 100: SF(p) = \sqrt{\frac{0.39 \cdot (1-0.39)}{100}} = 0.0488$$

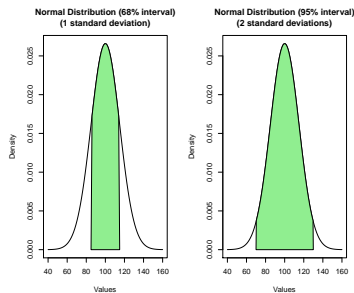
$$SF(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$\text{Bsp.: } SF(p) = \sqrt{\frac{0.39 \cdot (1-0.39)}{100}} = 0.0488$$

- Anders gesagt: Wenn man beliebig viele Stichproben der Größe $n = 100$ aus einer Grundgesamtheit zieht, in der der **wahre Anteilswert** $P = 0.39$ ist, ist eine Standardabweichung aller p (also der Standardfehler) $SF = 0.0488$.

Normalverteilung und z-Wert für Konfidenzniveau

- Um das KI für die gewünschte Konfidenzniveau zu ermitteln, müssen wir wissen, wie sich Werte um das geschätzte Mittel verteilen.
- Schätzverteilung dank Zentralem Grenzwertsatz: **Normalverteilung**
- Vorteil: Es ist genau bekannt, wieviel Werte je nach s in einem bestimmten Intervall liegen.



- Wir müssen nun wissen, wieviele Standardabweichungen bei der Normalverteilung 95% der Fläche definieren.
- Wenn es **symmetrische 95%** werden sollen, müssen **oben und unten je 2.5%** abgetrennt werden.
- Dazu gibt es Tabellen oder die **Quantil-Funktion der Normalverteilung `qnorm()`** in R.
- `qnorm(0.025, lower.tail=FALSE) \Rightarrow 1.959964`
- Also: **$z = 1.96$**

- Da der Standardfehler genau einer Standardabweichung entspricht, muss er nun mit dem z-Wert multipliziert werden.

$$KI = p \pm z \cdot SF(p)$$

$$\text{Bsp.: } KI = 0.39 \pm 1.96 \cdot 0.0488 = 0.39 \pm 0.096 = 0.29, 0.49$$

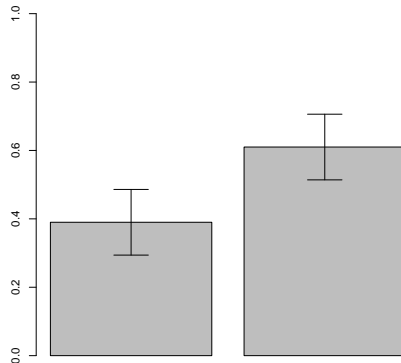
Das **Konfidenzintervall** ist in unserem Fall also

0.29 bis 0.49

- In 95% aller Stichproben mit $n = 100$ läge der Messwert beim wahren Anteil von 0.39 zwischen 0.29 und 0.49.
- Oft wird auf Basis einer Stichprobe mit der Größe n ein Anteilswert p geschätzt und dann für diesen das Konfidenzintervall ausgerechnet.
- Das kann man zwar machen, aber man lernt dadurch nichts über die GG!
- Ggf. kann uns das so errechnete KI einen Eindruck davon geben, wie genau Stichproben der Größe n bei einem Anteil wie dem gemessenen ungefähr sind.
- **Der gemessene Anteil p kann aber eine totale Fehlschätzung sein!**
- Die Philosophie bezieht sich auf das **wiederholte Berechnen** von KIs.

Verboten: Balkendiagramm mit Konfidenzintervall

Ein solches Diagramm signalisiert **fälschlicherweise**,
dass das Konfidenzintervall uns etwas über die GG sagt!



Nichtparametrische Verfahren

- Unterschiede in Zähldaten
- Signifikanz und Effektstärke
- Unterschiede bei Ja/Nein-Experimenten

- Gravetter & Wallnau 2007
- Bortz & Lienert 2008

Beobachtungen von zwei **kategorialen Variablen**.

Auxiliarwahl beim Perfekt: haben, sein

Herkunft des Belegs: nord, sued

Fall	Aux	Region
1	haben	nord
2	haben	nord
3	sein	nord
4	sein	sued
5	sein	sued
6	haben	nord
7	haben	sued
8	haben	sued

	Aux	
Region	haben	sein
nord	3	1
sued	2	2

Kreuztabelle mit Randsummen

Spaltensumme für Spalte i : $\sum_k x_{ik}$

Zeilensumme für Zeile j : $\sum_k x_{kj}$

	haben	sein	Zeilensummen
nord	3	1	4
sued	2	2	4
Spaltensummen	5	3	8

Beobachtete vs. erwartete Häufigkeiten

$n=100$

50 mal *haben*, 50 mal *sein* (= Spaltensummen)

50 mal Norden, 50 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der H_0
= kein Zusammenhang zwischen Hilfsverb und Region?

	haben	sein	Zeilensummen
nord	25	25	50
sued	25	25	50
Spaltensummen	50	50	100

Beobachtete vs. erwartete Häufigkeiten

$n=100$

50 mal *haben*, 50 mal *sein* (= Spaltensummen)

30 mal Norden, 70 mal Süden (= Zeilensummen)

- erwartete Häufigkeiten unter Annahme der H_0 ?

	haben	sein	Zeilensummen
nord	15	15	30
sued	35	35	70
Spaltensummen	50	50	100

Beobachtete vs. erwartete Häufigkeiten

n=100

30 mal Norden, 70 mal Süden

40 mal *haben*, 60 mal *sein*

	haben	sein	Zeilensummen
nord	12	18	30
sued	28	42	70
Spaltensummen	40	60	100

Allgemein: erwartete Häufigkeit für Zellen: $\frac{\text{Spaltensumme} \cdot \text{Zeilensumme}}{n}$

$$\text{bzw.: } EH(x_{ij}) = \frac{\sum_k x_{ik} \cdot \sum_k x_{kj}}{n}$$

Beobachtete vs. erwartete Häufigkeiten

beobachtete Häufigkeiten für eine DeReKo-Stichprobe (*geschwebt*):

	haben	sein	Zeilensummen
nord	27	33	60
sued	3	34	37
Spaltensummen	30	67	97

erwartete Häufigkeiten:

	haben	sein	Zeilensummen
nord	18.56	41.44	60
sued	11.44	25.56	37
Spaltensummen	30	67	97

- Beobachtete und erwartete Häufigkeit weichen ab.
- H_0 : kein Zusammenhang zwischen Region und Aux.
- Ab wann ist der Unterschied „signifikant“?
- Ein gemessener Unterschied ist **signifikant**, wenn er angesichts der Stichprobengröße groß genug ist, dass wir das im Experiment gefundene Ergebnis nur sehr selten (typischerweise in unter 5% der Fälle) erwarten würden, wenn er gar nicht bestünde.
- Diese 5% (als **Anteil** 0.05) sind das **Signifikanzniveau**.
- In Fishers Philosophie abgekürzt *sig*, nicht wie oft zu lesen „ α -Niveau“.

beobachtet:

	haben	sein
nord	27	33
sued	3	34

erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\chi^2 = \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}}$$

$$\text{bzw.: } \chi^2 = \sum_{ij} \frac{(x_{ij} - EH(x_{ij}))^2}{EH(x_{ij})}$$

Berechnung des χ^2 -Werts

$$\chi^2 = \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}}$$

beobachtet:

	haben	sein
nord	27	33
sued	3	34

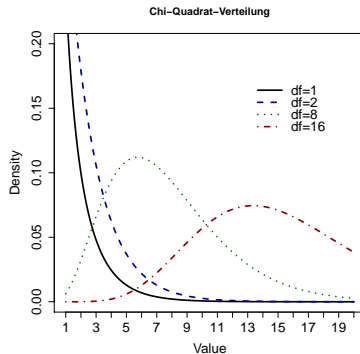
erwartet:

	haben	sein
nord	18.56	41.44
sued	11.44	25.56

$$\begin{aligned}\chi^2 &= \frac{(27-18.56)^2}{18.56} + \frac{(33-41.44)^2}{41.44} + \frac{(3-11.44)^2}{11.44} + \frac{(34-25.56)^2}{25.56} \\ \chi^2 &= 3.84 + 1.72 + 6.23 + 2.79 = 14.58\end{aligned}$$

Die χ^2 -Verteilung

Die χ^2 -Verteilung für Stichproben
aus Grundgesamtheiten ohne Zusammenhang:



Was sind „Freiheitsgrade“ oder *degrees of freedom (df)*?

- Das kommt später noch ausführlicher.
- Für n-Felder-Tests: $(\text{Zeilenzahl} - 1) \cdot (\text{Spaltenzahl} - 1)$
- Bei Vierfelder-Test also: $df = 1$

- Wahrscheinlichkeit eines bestimmten χ^2 -Werts unter Annahme der H_0 ?
VOR dem Experiment! Nach dem Experiment ist die Wahrscheinlichkeit des gemessenen p-Werts immer 1.
- In Fishers Philosophie Entscheidung nach Signifikanzniveau (*sig*):
Der χ^2 -Wert muss in den extremen *sig*-Anteilen liegen, um die H_0 zu *sig* zurückzuweisen.

In R ähnlich wie bei Normalverteilung:

`> qchisq(0.95, df=1) \Rightarrow 3.84`

- Also ist für $\chi^2 = 14.58$ auf jeden Fall $p < 0.05$ (weil $14.58 > 3.84$).

Mehr oder weniger signifikant?

- Oft liest man etwas von „ α -Niveau“ wie:
 - ▶ 5% („signifikant“)
 - ▶ 1%
 - ▶ 0.1% („hochsignifikant“)
- Diese Niveaus entsprechen einem falsch interpretierten *sig.*
- Die Idee von „mehr oder weniger signifikant“ ist **kompletter Schwachsinn**.
- Entweder ist das gesetzte Niveau akzeptabel, und dann bringt ein kleineres p aber auch nicht mehr.
- Oder es müsste eigtl. ein strengeres *sig*-Niveau gewählt werden, und dann ist $p < 0.05$ schlicht nicht ausreichend (s. Fishers **Sensitivität**).
- Die Entscheidung für ein bestimmtes *sig*-Niveau muss auf Basis konzeptueller/inhaltlicher Gründe gefällt werden.
- **EIN signifikantes Testergebnis alleine sagt nicht viel aus!!!**

Voraussetzungen für χ^2 -Tests

- 1 Die Beobachtungen sind voneinander unabhängig.
- 2 In jeder Zelle ist die erwartete Häufigkeit mindestens 5.
- 3 Keine Beschränkung auf vier Felder!

Mit einer Matrix `my.matrix`:

```
> chisq.test(my.matrix)
```

Eingabe einer einfachen Vierfeldermatrix:

```
> my.matrix <- matrix(c(27,33,3,34), 2, 2, byrow=TRUE)
```

Ausgeben der erwarteten Häufigkeiten:

```
> chisq.test(my.matrix)$expected
```

Der Fisher-Exakt-Test ist eine Alternative zum χ^2 -Test.

- exakter Test: direkte Berechnung der Wahrscheinlichkeit
- **keine** allgemein bessere Alternative zu χ^2
- robuster bei sehr kleinen Stichproben
- **aber nur für feststehende Randsummen geeignet!**
- ohne feste Randsummen: **Barnards Test**

Fisher-Exakt in R:

```
> fisher.test(my.matrix)
> fisher.test(my.vector.1, my.vector.2)
```

Der χ^2 -Wert sagt nichts über die **Stärke eines Zusammenhangs**!
Bei höheren absoluten Frequenzen wird auch der χ^2 -Wert größer.

	haben	sein
nord	27	33
sued	3	34

$$\chi^2 = 12,89$$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

	haben	sein
nord	54	66
sued	6	68

$$\chi^2 = 27,46$$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

Pearsons ϕ : Maß für die Stärke des Zusammenhangs in 2×2 -Tabellen

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

ϕ ist eine Zahl zwischen 0 und 1:

Je größer, desto stärker der Zusammenhang zwischen den Variablen.

$$\text{Beispiel: } \phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{12.89}{97}} = 0.3648$$

Cramér's v for $n \times n$ -Tables with $n > 2$ or $m > 2$

$$v = \sqrt{\frac{\frac{\chi^2}{n}}{\min(s-1, z-1)}}$$

mit: s die Spaltenzahl und z die Zeilenzahl

Beachte: für 2×2 -Tabellen: $s - 1 = 1$ und $z - 1 = 1$,

also $\min(s - 1, z - 1) = 1$

$$\text{daher: } v = \sqrt{\frac{\frac{\chi^2}{n}}{1}} = \sqrt{\frac{\chi^2}{n}} = \phi$$

Speichern des Test-Objekts:

```
> my.chi2.test <- chisq.test(my.matrix)
```

Speichern des χ^2 -Werts mit:

```
> my.chi2.value <- as.numeric(my.chi2.test$statistic)
```

Speichern von n :

```
> my.n <- sum(my.matrix)
```

Also Effektstärke (mit Ausgabe):

```
> my.phi <- sqrt( my.chi2.value / my.n ); my.phi
```

- Die **Chance (odds)** o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

$$o(E) = \frac{p(E)}{1-p(E)}$$

und damit

$$p(E) = \frac{o(E)}{1+o(E)}$$

- Ein Ereignis ist in Korpusstudien i. d. R. das Auftreten einer **Variablenausprägung**.
- Die Information in den Maßen Wahrscheinlichkeit und Chance ist dieselbe (s. Umrechenbarkeit ineinander).

Aux	Anzahl
haben	27
sein	33

$$p(\text{haben}) = \frac{27}{27+33} = \frac{27}{60} = 0.45 \text{ (Wahrscheinlichkeit)}$$

$$1 - p(\text{haben}) = p(\neg\text{haben}) = \frac{33}{27+33} = \frac{33}{60} = 0.55 \text{ (Gegenwahrscheinlichkeit)}$$

$$\text{Beachte: } p(\text{haben}) + p(\neg\text{haben}) = 1$$

$$o(\text{haben}) = \frac{\frac{27}{60}}{\frac{33}{60}} = \frac{27}{60} \cdot \frac{60}{33} = \frac{27}{33} = 0.82$$

$$\text{allgemein: } p(E) = \frac{\text{Anzahl}(E)}{\text{Anzahl}(E) + \text{Anzahl}(\neg E)} \text{ und } o(E) = \frac{\text{Anzahl}(E)}{\text{Anzahl}(\neg E)}$$

- Das Chancenverhältnis (odds ratio) gibt das Verhältnis an, wie sich die Chancen einer Variablenausprägung E unter Bedingung A – also $o(E|A)$ – und unter Bedingung B – also $o(E|B)$ – zueinander Verhalten:

$$r(E|A, E|B) = \frac{o(E|A)}{o(E|B)}$$

Beispiel zum Chancenverhältnis (1)

- Wir haben Texte aus Süddeutschland und Norddeutschland auf das Auftreten des Perfektauxiliars *haben* und *sein* bei bestimmten Verben untersucht.
- Die Kreuztabelle:

	nord	sued
haben	27	3
sein	33	34

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(haben|nord) = \frac{27}{33} = 0.82$
- $o(haben|sued) = \frac{3}{34} = 0.09$
- Verhältnis zwischen den Chancen: $or = \frac{0.82}{0.09} = 9.11$
- D. h. die Chance von *haben* ist 9.11 mal größer, wenn die Region *nord* ist.
- Ersatz für Effektstärke bei Fisher-Test

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*.“
 - ▶ „frei alternieren“ = beide Kasus haben den gleichen Anteil.
 - ▶ Grundgesamtheit per Null-Hypothese: 50% Genitive und 50% Dative
- Korpusstichprobe: $F(\text{Genitiv})=41$ und $F(\text{Dativ})=59$
- Stimmt das mit der Null überein bei $\text{sig} = 0.05$?

H_0 : Es gibt keine Abweichung
von den erwarteten gleich großen Anteilen.

$H_0: p(\textit{Dativ}) = 0.5$ (p für proportion)

Benötigte Größen:

- Stichproben der Größe n
- Proportion p (hier $p = 0.5$)
- Anzahl der beobachteten Ereignisse: X (hier $X(\text{Dativ}) = 59$)

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der Ho!) für die Normalverteilung:
 - ▶ Mittel: $\mu = p \cdot n$
 - ▶ Standardabweichung: $s = \sqrt{n \cdot p \cdot (1 - p)}$
 - ▶ Wir können für den gemessenen Wert den z-Wert ausrechnen.

$$z = \frac{x - \mu}{s} = \frac{x - p \cdot n}{\sqrt{n \cdot p \cdot (1 - p)}}$$

$$z = \frac{59 - (0.5 \cdot 100)}{\sqrt{100 \cdot 0.5 \cdot 0.5}} = \frac{59 - 50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

- Der gemessene Wert liegt 1.8 Standardabweichungen vom HO-Mittel entfernt.
- Wir kennen bereits die kritischen Werte für Normalverteilungen und $\text{sig} = 0.05$: $-1.96..1.96$
- Die H_0 kann also nicht zurückgewiesen werden bei $\text{sig} = 0.05$.
- Interpretation: Entweder ist die Variation nicht genau gleich verteilt oder ein seltenes Ereignis ist eingetreten.

```
> binom.test(59, 100, 0.5)
```

Exact binomial test

data: 59 and 100

number of successes = 59, number of trials = 100, p-value = 0.08863

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.4871442 0.6873800 sample estimates:

probability of success 0.59

z-Test und t-Test

- Wann sind Unterschiede zwischen Mittelwerten signifikant?
- Mittelwerte in Grundgesamtheiten und Stichproben

- Gravetter & Wallnau (2007)
- Bortz & Schuster (2010)
- oder eben gleich Fisher (1935)

- 1 **Nullhypothese** (H_0) festlegen: Der theoretisch angenommene Effekt existiert **nicht** (z. B.: Die Versuchsperson [VP] kann **nicht** erkennen, ob Tee oder Milch zuerst in der Tasse war).
- 2 **Stichprobengröße** und **Versuchsaufbau** festlegen (z. B. acht Tassen mit vier *Tee zuerst*-Tassen; VP kennt das Verhältnis)
- 3 **sig-Niveau** festlegen: Wie unwahrscheinlich darf das Ergebnis unter Annahme der H_0 sein, damit wir die H_0 zurückweisen.
- 4 Experiment durchführen, Ergebnis messen.
- 5 **p-Wert** berechnen: Wie wahrscheinlich **war** es, dieses Ergebnis oder ein extremeres Ergebnis zu erreichen, wenn die H_0 die Welt korrekt beschreibt.
- 6 Wenn $p \leq \text{sig}$, dann H_0 zurückweisen: Entweder der Effekt existiert (z. B. die VP kann die Reihenfolge des Einschenkens erkennen) **oder ein seltenes Ereignis ist eingetreten.**

- Voraussetzung: **echte Zufallsstichprobe**
- Ergebnis: **kein Beweis**
- keine Auskunft darüber, wie „wahrscheinlich“ der Effekt ist
- keine Auskunft darüber, wie stark wir von der Existenz des Effekts überzeugt sein sollten (= *inverse probability*)
- jede Ho-Zurückweisung: nur ein kleinteiliger Hinweis auf einen Effekt
- **substantielle** theoretische Hypothese oft und hart testen!

- **Sensitivity**: keine Auskunft über die **Stärke** des Effekts
 - ▶ große Stichprobe → hohe Sensitivität
 - ▶ kleine Stichprobe → niedrige Sensitivität
 - ▶ je sensibler desto leichter werden schwache Effekte signifikant
 - ▶ Abhilfe bei Neyman-Pearson: **Power** (Teststärke) vor dem Experiment
 - ▶ quasi-kompatibel zu Fisher: **Effektstärke** nach dem Experiment

Und beim Konfidenzintervall?

Am Beispiel des 95%-Konfidenzintervalls (KI)

- **Falsch:** Wir können zu 95% sicher sein, dass der wahre Wert im KI liegt.
- **Falsch:** Der wahre Wert liegt mit 95% Wahrscheinlichkeit im KI.
- Warum? Wenn der wahre Wert nicht im geschätzten KI liegt, ist die Wahrscheinlichkeit 1, dass er nicht im KI liegt.
- Fakten haben die Wahrscheinlichkeit 1.
- Richtig: Entweder liegt der wahre Wert im KI oder ein seltenes Ereignis ist eingetreten
- „selten“ heißt: nur in 5 von 100 Fällen (im Grenzwert)

- **exakter** Test:

- ▶ Die Wahrscheinlichkeitsverteilung ist bekannt und wird direkt zugrunde gelegt (= Berechnung der exakten Wahrscheinlichkeit).
- ▶ Fisher-Test, Binomialtest
- ▶ hohe Sensitivität
- ▶ geeignet für kleine Stichproben
- ▶ oft rechenintensiv

- **approximativer** oder **asymptotischer** Test:

- ▶ Die Wahrscheinlichkeitsverteilung ist nicht bekannt (oder kann mathematisch nicht effizient zugrundegelegt werden) und es wird ein Differenzwert berechnet, der asymptotisch eine bekannte Verteilung hat.
- ▶ χ^2 -Test, t-Test, ANOVA
- ▶ oft wird Normalverteilung approximiert
- ▶ wegen asymptotischer Natur weniger sensitiv (= größere Stichprobe)

- parametrischer Test:

- ▶ Messung eines Parameters/mehrerer Parameter der Grundgesamtheit
- ▶ (Parameter entsprechen in der Messung einer Variable)
- ▶ zum Beispiel Mittelwert oder Varianz
- ▶ Voraussetzung: **bekannte Wahrscheinlichkeitsverteilung der Variable**
- ▶ z. B. t-Test (mittel), ANOVA (Varianz)

- nichtparametrischer Test:

- ▶ keine direkte Messung eines zufallsverteilten Parameters
- ▶ zum Beispiel Ränge oder Zähldaten
- ▶ keine Verteilungsannahmen (auch: *verteilungsfreier Test*)
- ▶ z. B. χ^2 , Binomialtest, H-Test, U-Test

- Mittel μ über X in der Grundgesamtheit bekannt (z. B. mittlere Satzlänge im Korpus).
- Stichprobe (z. B. der Grundriss von PE) zeigt gemessenes Mittel \bar{x} .
- Ist die Abweichung signifikant?
- $H_0: \bar{x} = \mu$

Wäre die **Varianz der GG** als $s^2(X)$ bekannt:

- $SF(X)$ bei Stichprobengröße n ausrechnen, und...
- mit $z = \frac{\bar{x} - \mu}{SF(X)}$ einen Signifikanztest über Normalverteilung rechnen
- Problem aber leider: $SF(X) = \frac{s(X)}{\sqrt{n}}$
- und $s^2(X)$ meist nicht bekannt!

Aufgabe: Mit Ihrer Stichprobe aus NaB und $\mu = 6.8$ sowie $s^2(X) = 10.8$ z-Test rechnen. (Bzw. erstmal die nötigen Werte ausrechnen. Wir besprechen dann die Interpretation als Test.)

- Wir kennen μ oder haben eine Hypothese (z. B. $\mu = 0.5$).
- Wir haben eine Stichprobe x mit n und bekannten \bar{x} und $s^2(x)$.
- anders als bei z-Test: **Wir schätzen $SF(X) \approx SF(x)$!**

$$t = \frac{\bar{x} - \mu}{SF(\bar{x})}$$

Bitte rechnen für Satzlängen (in Wörtern):

$$\mu = 7.3$$

$$x = [6, 3, 12, 16, 8, 15, 9, 9, 2, 11]$$

1 $\bar{x} = 9.1$

2 $s^2(x) = 21.43$

3 $s(x) = 4.63$

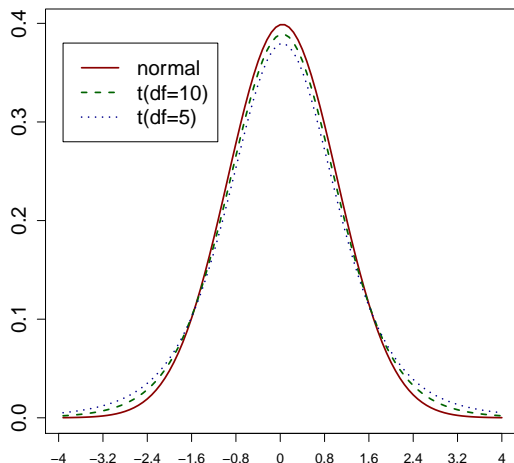
4 $SF(x) = \frac{4.63}{\sqrt{10}} = 1.464$

5 $t = \frac{9.1-7.3}{1.464} = 1.229$

Und was sagt uns $t = 1.229$?

t-Verteilung

Während die z-Werte normalverteilt sind,
flacht die Verteilung der t-Werte durch die Schätzung
je nach df verglichen mit der Normalverteilung ab.



- $df = n - 1$ (\bar{x} muss für $s^s(x)$ bekannt sein)
- Welche t-Werte machen $1 - \alpha$ der Werte aus?
- $> qt(c(0+0.05/2, 1-0.05/2), df=9)$
 $\Rightarrow 2.262157.. - 2.262157$
- Der errechnete t-Wert ist nicht signifikant.
- $H_0: \mu = \bar{x}$ nicht zurückgewiesen.

- Signifikanz \neq starker Effekt
- Effektstärke beim t-Test für Stichprobe x:

$$\text{Cohens } d = \frac{\bar{x} - \mu}{s(x)}$$

- Herleitung/Erklärung: Gravetter & Wallnau, Kap. 9

- ähnlich der Effektstärke:
Welcher Anteil der Varianz in den Daten
wird durch die Unabhängige erklärt?

$$\text{Cohens } r^2 = \frac{t^2}{t^2 + df}$$

- Herleitung/Erklärung: Gravetter & Wallnau, Kap. 9

- zwei Grundgesamtheiten (z. B. dt. Sätze im 19. und im 20. Jh.)
- dazu: zwei Stichproben (je eine) mit einem Mittelwert (z. B. Länge)
- Interesse: anhand der zwei Stichproben zeigen, dass sie (sehr wahrscheinlich) aus zwei Grundgesamtheiten kommen
- $H_0: \mu_1 - \mu_0 = 0$
- hier also: eine unabhängige Variable (Jahrhundert) und eine abhängige Variable (Satzlänge), gemessen als Mittel

Allgemein funktioniert der t-Test **immer** so:

$$t = \frac{\text{Stichprobenwert} - \text{Grundgesamtheitswert}}{\text{Standardfehler}}$$

Jetzt geht man per Hypothese von zwei GG und zwei Stichproben aus, also:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SF(x_1 - x_2)}$$

- Wir testen also auf die **Differenz der Unterschiede**.
- Per H_0 wird gesetzt: $\mu_1 - \mu_2 = 0$

Für gleichgroße Stichproben:

$$SF(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s^2(x_1)}{n_1} + \frac{s^2(x_2)}{n_2}}$$

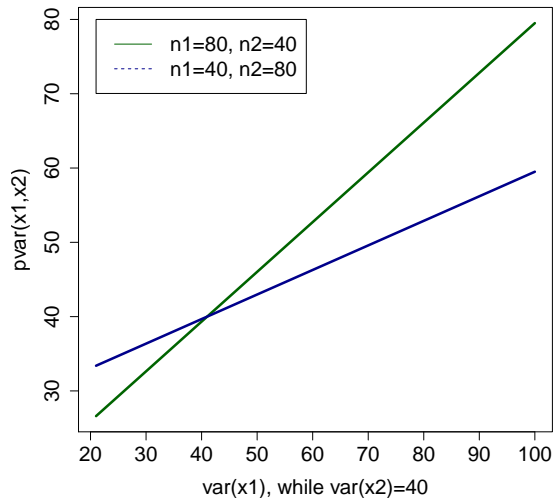
- Problem: Beitrag zum SF von beiden Stichproben gleich.
- Besser: **zusammengefasste Varianz**, und daraus dann SF.

$$s_p^2(x_1, x_2) = \frac{(\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2) + (\sum_{i=1}^n (x_{2,i} - \bar{x}_2)^2)}{(n_1 - 1) + (n_2 - 1)} = \frac{SQ(x_1) + SQ(x_2)}{(n_1 - 1) + (n_2 - 1)}$$

$$SF(x_1 - x_2) = \sqrt{\frac{s_p^2(x_1, x_2)}{n_1} + \frac{s_p^2(x_1, x_2)}{n_2}}$$

Mehr: Gravetter & Wallnau, Kap. 10

Illustration der zusammengefassten Varianz



t-Wert

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SF(x_1 - x_2)} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SF(x_1 - x_2)} = \frac{\bar{x}_1 - \bar{x}_2}{SF(x_1 - x_2)}$$

Freiheitsgrade

$$df = df(x_1) + df(x_2) = (n_1 - 1) + (n_2 - 1)$$

Effektstärke

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2}}$$

Erklärung der Varianz

$$r^2 = \frac{t^2}{t^2 + df}$$

Bitte „von Hand in R“ t-Test für folgende zwei Stichproben
bei $\alpha = 0.05$ rechnen:

$$\begin{aligned}x_1 &= [11, 11, 8, 8, 11, 9, 8, 11, 9, 8] \\x_2 &= [10, 14, 14, 13, 11, 14, 10, 14, 12, 10]\end{aligned}$$

Und überprüfen mit:
`> t.test(x1, x2)`

Die GGs müssen normalverteilt sein:

```
shapiro.test(x)
```

Wenn $p \leq 0.05$ wird die Nullhypothese des Shapiro-Wilk-Tests verworfen –
Ho: Die Werte stammen aus einer normalverteilten GG.

Die Varianzen müssen homogen sein:

```
var.test(x1, x2)
```

Auch hier: $p \leq 0.05$ weist die Ho zurück (sehr informell) –
Ho: Die Varianzen von x1 und x2 sind homogen.

Solche Tests sind umstritten, weil sie i. d. R. viel zu empfindlich reagieren.
Zuur u. a. (2009) empfehlen z. B. grafische Methoden (bei linearen Modellen).

Wenn Voraussetzungen nicht erfüllt sind:

- steigt das Risiko für Typ 1-Fehler
- nicht-parametrische Alternative nehmen
- Daten transformieren
- sich über Robustheit des Test ggü. verletzten Annahmen informieren (oft schwer zugängliche und kontroverse Spezialliteratur)

ANOVA

- Vergleiche von Mittelwerten zwischen mehr als zwei Gruppen
- Mittelwertvergleiche mit mehreren Unabhängigen
- Warum kann man über Varianzen Mittelwerte vergleichen?

- Gravetter & Wallnau (2007)
- Bortz & Schuster (2010)
- indirekt: Maxwell & Delaney (2004)

- Einschränkung beim t-test: immer nur 2 Gruppen
- t-Test bei mehr als 2 Gruppen: komplizierte paarweise Vergleiche
- stattdessen ANOVA: ANalysis Of VAriance
- Vergleich von Varianzen zwischen beliebigen Gruppen
- Schluss auf Mittelwerte nur indirekt über die Varianzen
- bei zwei Gruppen: Konvergenz von t-Test und ANOVA

- ANOVA vergleicht immer **mehrere Gruppen**
- Gruppen bei der einfaktoriellen ANOVA = den Ausprägungen **einer unabhängigen Variable** (z. B. Text-Register)
- diese Variablen heißen hier **Faktoren**.
- Einfluss der Faktoren auf **eine abhängige** (z. B. Satzlänge, Lesezeit)
- bei mehreren Faktoren (z. B. Text-Register und Jahrhundert): **mehrfaktorielle ANOVA**.

Idee bei ANOVA (z. B. drei Gruppen)

- $H_0: \bar{x}_1 = \bar{x}_2 = \bar{x}_3$
- aber: Es gibt keinen “Differenzwert” für drei Mittel (also sowas wie den t-Wert).
- daher Varianzvergleich
- F-Wert (Verteilung unter H_0 bekannt) als Test-Statistik

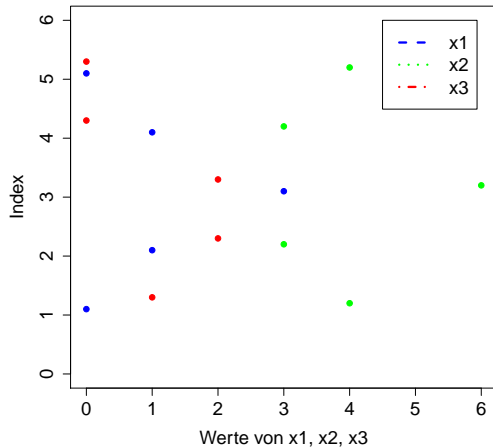
$$F = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz in den Stichproben}} = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz per Zufall}}$$

Drei Stichproben

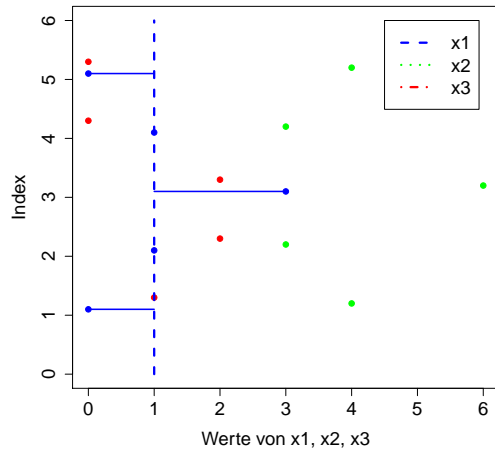
$$x_1 = [0, 1, 3, 1, 0]$$

$$x_2 = [4, 3, 6, 3, 4]$$

$$x_3 = [1, 2, 2, 0, 0]$$

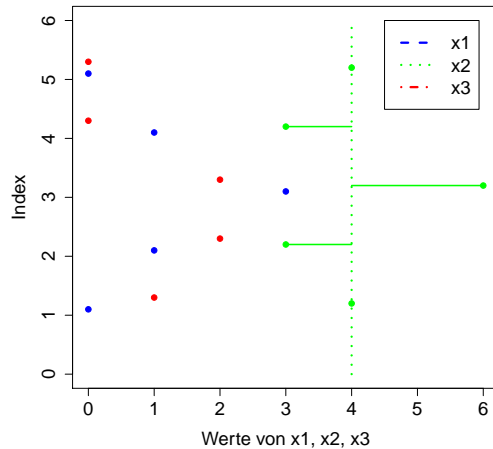


Komponenten der Varianz von x_1



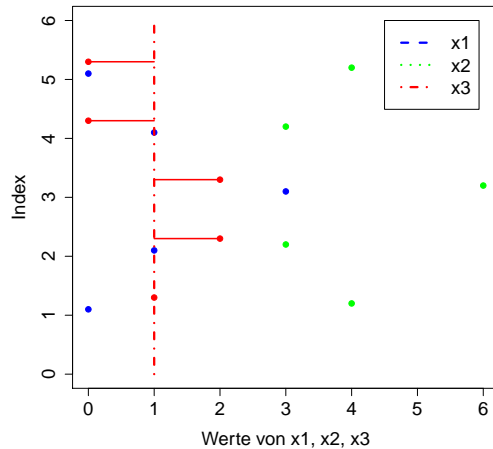
$$s^2(x_1) = 1.5$$

Komponenten der Varianz von x_2



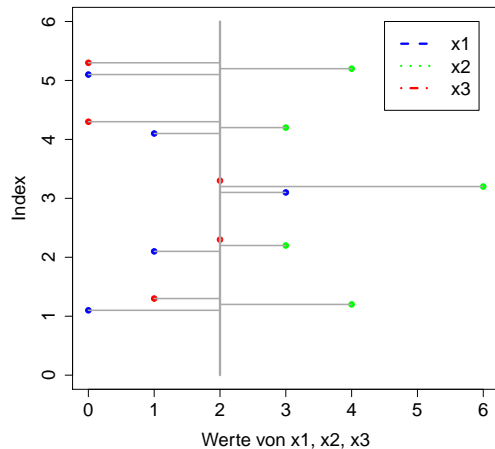
$$s^2(x_2) = 1.5$$

Komponenten der Varianz von x_3



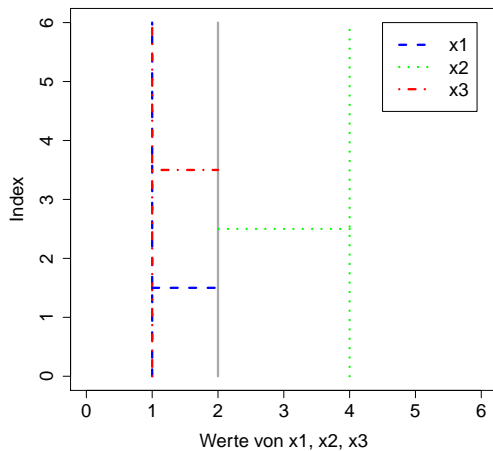
$$s^2(x_3) = 1$$

Varianz in der zusammengefassten Stichprobe X



$$s^2(X) = 3.29$$

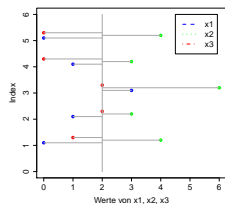
Varianz zwischen den drei Gruppen



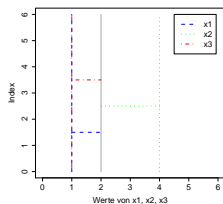
$$s^2([\bar{x}_1, \bar{x}_2, \bar{x}_3]) = 1.33$$

Achtung: Bei unterschiedlichen Stichprobengrößen

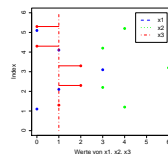
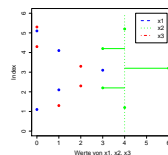
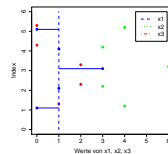
Es gilt bezüglich der Varianzen



=



+

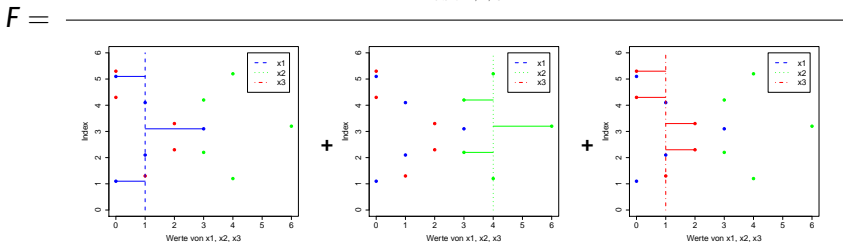
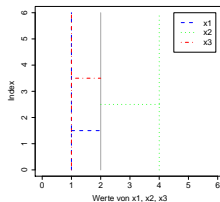


$$s^2(X) = s^2(\bar{x}_1, \bar{x}_2, \bar{x}_3) + s^2(x_1) + s^2(x_2) + s^2(x_3)$$

Wenn man den Abstand zwischen den Mitteln verschiebt

Graphische Verdeutlichung des F-Werts

$$F = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz in den Stichproben}} = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz per Zufall}}$$



Wenn man den Abstand zwischen den Mitteln verschiebt,
muss die Gesamtvarianz größer werden!

Wie funktioniert der F-Wert

- $F = \frac{\text{Varianz zwischen Stichprobenmitteln}}{\text{Varianz in den Stichproben}}$
- Warum?
- $F = \frac{\text{Unterschied durch Effekt} + \text{Unterschiede durch restliche Varianz}}{\text{Unterschied durch restliche Varianz}}$
- Unter Annahme der H_0 gibt es keinen Effekt, ...
- also $\text{Unterschied durch Effekt} = 0$
- dann: $F = \frac{0 + \text{Unterschiede durch restliche Varianz}}{\text{Unterschied durch restliche Varianz}} = 1$

- Anzahl der Gruppen x_i : k
- Größe der Gruppen: n_i
- Größe der Gesamtstichprobe X : N
- Summen der Gruppen: T_i
- Gesamtsumme: G
- Mittel (anders als G&W): \bar{x}_i, \bar{X}
- Summe der Quadrate (=Zähler der Varianz): $SQ(x_i), SQ(X)$

Zur Erinnerung: $s^2(x) = \frac{\sum(x-\bar{x})}{n-1} = \frac{SQ(x)}{df(x)}$

$$F = \frac{\text{Varianz zwischen den Gruppen}}{\text{Varianz in den Gruppen}} = \frac{s_{\text{zwischen}}^2}{s_{\text{in}}^2} = \frac{\frac{SQ_{\text{zwischen}}}{df_{\text{zwischen}}}}{\frac{SQ_{\text{in}}}{df_{\text{in}}}}$$

denn

$$s^2(x) = \frac{SQ(x)}{df(x)}$$

Am einfachsten unter Beachtung von:

$$SQ_{gesamt} = SQ_{zwischen} + SQ_{in}$$

$$\text{Es gilt: } SQ_{gesamt} = SQ(X) = \sum (X - \bar{X})^2$$

$$\text{Außerdem: } SQ_{in} = \sum SQ(x_i)$$

$$\text{Damit: } SQ_{zwischen} = SQ_{gesamt} - SQ_{in}$$

SQ_{zwischen} kann man auch direkt ausrechnen:

$$SQ_{\text{zwischen}} = \sum_i \left(\frac{T_i^2}{n_i} \right) - \frac{G^2}{N}$$

$$\mathbf{x}_1 = [0, 1, 3, 1, 0]$$

$$\mathbf{x}_2 = [4, 3, 6, 3, 4]$$

$$\mathbf{x}_3 = [1, 2, 2, 0, 0]$$

Bitte alle SQ ausrechnen, inkl. SQ_{zwischen} direkt.

Tipp: Sie brauchen als Vorwissen **nur**
den Stoff der ersten Statistik-Sitzung:

- arithmetisches Mittel
- SQ

Es gilt auch hier, ähnlich wie bei den SQ:

$$df_{gesamt} = df_{zwischen} + df_{in}$$

$$df_{gesamt} = N - 1$$

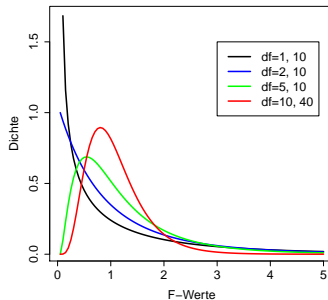
$$df_{zwischen} = k - 1$$

$$df_{in} = \sum_{i=1}^k (n_i - 1) = (N - 1) - (k - 1)$$

$$F = \frac{s_{\text{zwischen}}^2}{s_{\text{in}}^2} = \frac{\frac{SQ_{\text{zwischen}}}{df_{\text{zwischen}}}}{\frac{SQ_{\text{in}}}{df_{\text{in}}}}$$

Bitte ausrechnen für o. g. Beispiel.

F-Verteilung:



In R für $df_{\text{zwischen}} = 2$ und $df_{\text{in}} = 12$ bei $\text{sig}=0.05$:

`> qf(0.95, 2, 12) ⇒ 3.885294`

$$\eta^2 = \frac{SQ_{zwischen}}{SQ_{gesamt}}$$

(wieder ein r^2 -Maß)

- Problem: Welche Gruppen unterscheiden sich denn nun?
- Lösung: Post(-Hoc)-Tests, z. B. Scheffé-Test:
 - ▶ paarweise ANOVA
 - ▶ aber: k wird gesetzt wie bei ursprünglicher ANOVA
 - ▶ dadurch Vermeidung kumulierten Alpha-Fehlers (Vorteil ggü. paarweisen t-Tests)
 - ▶ weiterer Vorteil: paarweise Post-Tests nur erforderlich, wenn Omnibus-ANOVA bereits Signifikanz gezeigt hat
 - ▶ und: Generalisierbarkeit zu mehrfaktorieller ANOVA (geht mit t-Test nicht)

Bitte ausrechnen für die oben gerechnete ANOVA.

Oft vermutet man den Einfluss **mehrerer Unabhängiger** auf eine Abhängige.
Beispiel: Satzlängen

		Textsorte		
		Fiktion	Zeitung	Wissenschaft
Jahrhundert	19	X_{11}	X_{12}	X_{13}
	20	X_{21}	X_{22}	X_{23}

Hier also: $2 \cdot 3 = 6$ Gruppen

Ablauf der zweifaktoriellen ANOVA

- 1 erste ANOVA zwischen Zeilen
- 2 zweite ANOVA zwischen Spalten
- 3 dritte ANOVA für **Interaktionen** zwischen Zeilen und Spalten
- 4 Interaktion: Ungleichverteilung in Gruppen, die nicht durch die Spalten- und Zeileneffekte erklärt werden kann
- 5 Alle drei ANOVAs sind **unabhängig** voneinander!

- **Gesamtvarianz** = Varianz zwischen Gruppen + Varianz in den Gruppen
- **Varianz zwischen den Gruppen** =
Haupt-Faktoren-Varianz + **Interaktions-Varianz**
- **Haupt-Faktoren-Varianz** =
Varianz zwischen Faktor A-Gruppen +
Varianz zwischen Faktor B-Gruppen

Schritt 1(1): SQ/df zwischen den Gruppen

Jede Zelle der Tabelle ist eine Gruppe.

$$SQ_{\text{zwischen}} = \sum_i \left(\frac{T_i^2}{n_i} \right) - \frac{G^2}{N}$$
$$df_{\text{zwischen}} = k - 1 \text{ (k = Anzahl der Zellen/Gruppen)}$$

Beachte: Keine Änderung verglichen mit einfaktorieller ANOVA!

Schritt 1(2): SQ/df in den Gruppen

Jede Zelle der Tabelle ist eine Gruppe.

$$\begin{aligned}SQ_{in} &= \sum SQ(x_i) \\ df_{in} &= \sum df(x_i)\end{aligned}$$

Beachte: **Keine** Änderung verglichen mit einfaktorieller ANOVA!

Schritt 2(2): SQ/df für Gruppe A

Berechnung nach dem Schema für Zwischen-Gruppen-Varianz

		Textsorte			
		Fiktion	Zeitung	Wissenschaft	
Jahrhundert	19	x_{11}	x_{12}	x_{13}	A_1
	20	x_{21}	x_{22}	x_{23}	A_2

Auch hier keine wesentliche Änderung:

$$SQ_A = \sum_i \left(\frac{T_{A_i}^2}{n_{A_i}} \right) - \frac{G^2}{N}$$

$$df_A = k_A - 1 \quad (k_A = \text{Anzahl der Zeilen})$$

Schritt 2(2): SQ/df für Gruppe A

Berechnung nach dem Schema für Zwischen-Gruppen-Varianz

		Textsorte		
		Fiktion	Zeitung	Wissenschaft
Jahrhundert	19	X_{11}	X_{21}	X_{31}
	20	X_{12}	X_{22}	X_{32}
		B_1	B_2	B_3

Auch hier keine Änderung:

$$SQ_B = \sum_i \left(\frac{T_{B_i}^2}{n_{B_i}} \right) - \frac{G^2}{N}$$

$$df_B = k_B - 1 \quad (k_B = \text{hier Anzahl der Spalten})$$

Schritt 2(3): SQ/df für Interaktion $A \times B$

Die Varianz, die auf Kosten der Interaktion geht, ist
die Zwischen-Gruppen-Varianz ohne die Einzelfaktor-Varianz.

$$\begin{aligned} SQ_{A \times B} &= SQ_{\text{zwischen}} - SQ_A - SQ_B \\ df_{A \times B} &= df_{\text{zwischen}} - df_A - df_B \end{aligned}$$

Alle drei F-Werte ausrechnen

Die zweifaktorielle ANOVA
erfordert wie gesagt drei Einzel-ANOVAs.

$$F_A = \frac{\frac{SQ_A}{df_A}}{\frac{SQ_{zwischen}}{df_{zwischen}}} = \frac{s_A^2}{s_{zwischen}^2}$$

$$F_B = \frac{\frac{SQ_B}{df_B}}{\frac{SQ_{zwischen}}{df_{zwischen}}} = \frac{s_B^2}{s_{zwischen}^2}$$

$$F_{A \times B} = \frac{\frac{SQ_{A \times B}}{df_{A \times B}}}{\frac{SQ_{zwischen}}{df_{zwischen}}} = \frac{s_{A \times B}^2}{s_{zwischen}^2}$$

Entsprechend sind **drei** η^2 auszurechnen:

$$\eta_A^2 = \frac{SQ_A}{SQ_{gesamt} - SQ_B - SQ_{A \times B}}$$

$$\eta_B^2 = \frac{SQ_B}{SQ_{gesamt} - SQ_A - SQ_{A \times B}}$$

$$\eta_{A \times B}^2 = \frac{SQ_{A \times B}}{SQ_{gesamt} - SQ_A - SQ_B}$$

Wir fragen jeweils, welchen Anteil an der Varianz, die die anderen beiden Faktoren **nicht** erklären, der jeweilige dritte Faktor hat.

Bitte vollständige zweifaktorielle ANOVA
bei $\text{sig}=0.05$ und $\text{sig}=0.01$ rechnen:

	B1	B2	B3
A1	1, 3, 1, 4	4, 3, 3, 6	8, 6, 8, 10
A2	8, 6, 6, 8	1, 6, 8, 1	1, 4, 1, 4

Freiheitsgrade und Effektstärken

- Beispiel: Schätzung eines Parameters (z. B. Mittel) auf Basis von 1000 gemessenen Werten
- Wenn 999 Werte bekannt sind, steht abhängig vom Mittel der 1000ste Wert fest.
- Für jedes Mittel μ einer Stichprobe mit n Messungen sind also nur $n - 1$ frei wählbar.

(Unintuitive) Erweiterung(en)

- generell: $df = n - |E|$
wobei E die zu schätzenden Parameter sind. $|E|$ ist ihre Anzahl.
- Warum bei χ^2 dann $df = (\text{Zeilenzahl} - 1) \cdot (\text{Spaltenzahl} - 1)$?
- Bsp.: Tabelle mit 2×3 Feldern, also $df = (2 - 1)(3 - 1) = 1 \cdot 2 = 2...$
- Bei bekannten Randsummen sind aber tatsächlich nur 2 Felder frei wählbar!

	X1	X2	
Y1	\oplus		ZS1
Y2	\oplus		ZS2
Y3			ZS3
	SQ1	SQ2	

Der χ^2 -Wert sagt nichts über die **Stärke eines Zusammenhangs**!
Bei höheren absoluten Frequenzen wird auch der χ^2 -Wert größer.

	haben	sein
nord	27	33
sued	3	34

$$\chi^2 = 12,89$$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

	haben	sein
nord	54	66
sued	6	68

$$\chi^2 = 27,46$$

	haben	sein
nord	27.84%	34.02%
sued	3.09%	35.05%

Pearsons ϕ : Maß für die Stärke des Zusammenhangs in 2×2 -Tabellen

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

ϕ ist eine Zahl zwischen 0 und 1:

Je größer, desto stärker der Zusammenhang zwischen den Variablen.

$$\text{Beispiel: } \phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{12.89}{97}} = 0.3648$$

Cramér's v for $n \times n$ -Tables with $n > 2$ or $m > 2$

$$v = \sqrt{\frac{\frac{\chi^2}{n}}{\min(s-1, z-1)}}$$

mit: s die Spaltenzahl und z die Zeilenzahl

Beachte: für 2×2 -Tabellen: $s - 1 = 1$ und $z - 1 = 1$,

also $\min(s - 1, z - 1) = 1$

$$\text{daher: } v = \sqrt{\frac{\frac{\chi^2}{n}}{1}} = \sqrt{\frac{\chi^2}{n}} = \phi$$

Speichern des Test-Objekts:

```
> my.chi2.test <- chisq.test(my.matrix)
```

Speichern des χ^2 -Werts mit:

```
> my.chi2.value <- as.numeric(my.chi2.test$statistic)
```

Speichern von n :

```
> my.n <- sum(my.matrix)
```

Also Effektstärke (mit Ausgabe):

```
> my.phi <- sqrt( my.chi2.value / my.n ); my.phi
```

- Die **Chance (odds)** o setzt die Wahrscheinlichkeit p eines Ereignisses E in Relation zur Gegenwahrscheinlichkeit:

$$o(E) = \frac{p(E)}{1-p(E)}$$

und damit

$$p(E) = \frac{o(E)}{1+o(E)}$$

- Ein Ereignis ist in Korpusstudien i. d. R. das Auftreten einer **Variablenausprägung**.
- Die Information in den Maßen Wahrscheinlichkeit und Chance ist dieselbe (s. Umrechenbarkeit ineinander).

Aux	Anzahl
haben	27
sein	33

$$p(\text{haben}) = \frac{27}{27+33} = \frac{27}{60} = 0.45 \text{ (Wahrscheinlichkeit)}$$

$$1 - p(\text{haben}) = p(\neg\text{haben}) = \frac{33}{27+33} = \frac{33}{60} = 0.55 \text{ (Gegenwahrscheinlichkeit)}$$

$$\text{Beachte: } p(\text{haben}) + p(\neg\text{haben}) = 1$$

$$o(\text{haben}) = \frac{\frac{27}{60}}{\frac{33}{60}} = \frac{27}{60} \cdot \frac{60}{33} = \frac{27}{33} = 0.82$$

$$\text{allgemein: } p(E) = \frac{\text{Anzahl}(E)}{\text{Anzahl}(E) + \text{Anzahl}(\neg E)} \text{ und } o(E) = \frac{\text{Anzahl}(E)}{\text{Anzahl}(\neg E)}$$

- Das Chancenverhältnis (odds ratio) gibt das Verhältnis an, wie sich die Chancen einer Variablenausprägung E unter Bedingung A – also $o(E|A)$ – und unter Bedingung B – also $o(E|B)$ – zueinander Verhalten:

$$r(E|A, E|B) = \frac{o(E|A)}{o(E|B)}$$

Beispiel zum Chancenverhältnis (1)

- Wir haben Texte aus Süddeutschland und Norddeutschland auf das Auftreten des Perfektauxiliars *haben* und *sein* bei bestimmten Verben untersucht.
- Die Kreuztabelle:

	nord	sued
haben	27	3
sein	33	34

Beispiel zum Chancenverhältnis (2)

	nord	sued
haben	27	3
sein	33	34

- $o(\textit{haben}|\textit{nord}) = \frac{27}{33} = 0.82$
- $o(\textit{haben}|\textit{sued}) = \frac{3}{34} = 0.09$
- Verhältnis zwischen den Chancen: $or = \frac{0.82}{0.09} = 9.11$
- D. h. die Chance von *haben* ist 9.11 mal größer, wenn *Region nord* ist.
- Ersatz für Effektstärke bei Fisher-Test

- binäre Daten: Ereignis vs. Nicht-Ereignis bzw. Ja/Nein
- Vgl. Behauptung: „Gen/Dat alternieren frei bei *wegen*.“
 - ▶ „frei alternieren“ = beide Kasus haben die gleiche Chance.
 - ▶ Grundgesamtheit per Hypothese: 50% Genitive und 50% Dative
- Korpusstichprobe: $F(\text{Genitiv})=41$ und $F(\text{Dativ})=59$
- Passt das zur Hypothese bei $\text{sig}=0.05$?

- H_0 : Es gibt keine Abweichung von der erwarteten Wahrscheinlichkeit.
- $H_0: p(\textit{Dativ}) = 0.5$

Benötigte Größen:

- Stichproben der Größe n
- Ho-Wahrscheinlichkeit p (hier $p = 0.5$)
- Anzahl der beobachteten Ereignisse: X (hier $X(\text{Dativ}) = 59$)

- Wenn $p \cdot n > 10$ und $(1 - p) \cdot n > 10$
approximiert die Binomialverteilung die Normalverteilung.
- Es gilt dann (unter Annahme der Ho!) für die Normalverteilung:
 - ▶ Mittel: $\mu = p \cdot n$
 - ▶ Standardabweichung: $s = \sqrt{n \cdot p \cdot (1 - p)}$
 - ▶ Wir können für den gemessenen Wert den z-Wert ausrechnen.

$$z = \frac{X - \mu}{s} = \frac{X - p \cdot n}{\sqrt{n \cdot p \cdot (1 - p)}}$$

$$z = \frac{59 - (0.5 \cdot 100)}{\sqrt{100 \cdot 0.5 \cdot 0.5}} = \frac{59 - 50}{\sqrt{25}} = \frac{9}{5} = 1.8$$

- Der gemessene Wert liegt 1.8 Standardabweichungen vom Ho-Mittel entfernt.
- Wir kennen bereits die kritischen Werte für Normalverteilungen und $\text{sig}=0.05$: $-1.96..1.96$
- Die Ho kann also nicht zurückgewiesen werden bei $\text{sig}=0.05$.
- Interpretation: Wir haben keine Evidenz dafür, dass die Variation in der Grundgesamtheit von einer 50:50-Verteilung abweicht.
- Falsche Interpretation: Wir haben Evidenz dafür, dass die Verteilung in der Grundgesamtheit 50:50 ist.


```
> binom.test(59, 100, 0.5)
```

Exact binomial test

data: 59 and 100

number of successes = 59, number of trials = 100, p-value = 0.08863

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.4871442 0.6873800 sample estimates:

probability of success 0.59

- Signifikanz \neq starker Effekt
- Effektstärke beim t-Test für Stichprobe x:

$$\text{Cohens } d = \frac{\bar{x} - \mu}{s(x)}$$

- Herleitung/Erklärung: Gravetter & Wallnau, Kap. 9

- ähnlich der Effektstärke:
Welcher Anteil der Varianz in den Daten
wird durch die Unabhängige erklärt?

$$\text{Cohens } r^2 = \frac{t^2}{t^2 + df}$$

- Herleitung/Erklärung: Gravetter & Wallnau, Kap. 9

Effektstärke

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2}}$$

Erklärung der Varianz

$$r^2 = \frac{t^2}{t^2 + df}$$

$$\eta^2 = \frac{SQ_{zwischen}}{SQ_{gesamt}}$$

(wieder ein r^2 -Maß)

Entsprechend sind **drei** η^2 auszurechnen:

$$\eta_A^2 = \frac{SQ_A}{SQ_{gesamt} - SQ_B - SQ_{A \times B}}$$

$$\eta_B^2 = \frac{SQ_B}{SQ_{gesamt} - SQ_A - SQ_{A \times B}}$$

$$\eta_{A \times B}^2 = \frac{SQ_{A \times B}}{SQ_{gesamt} - SQ_A - SQ_B}$$

Wir fragen jeweils, welchen Anteil an der Varianz, die die anderen beiden Faktoren **nicht** erklären, der jeweilige dritte Faktor hat.

Bedingung für **alle** Tests:
Unabhängigkeit der Messungen

Wenn bei t-Test oder ANOVA also gepaarte Stichproben vorliegen
(Messung derselben Proband*innen unter Bedingung 1 und 2 usw.):
Besondere Versionen für geparte Stichproben nehmen!

Details hier nicht besprochen.

Die GGs müssen normalverteilt sein:

```
shapiro.test(x)
```

Wenn $p \leq 0.05$ wird die Nullhypothese des Shapiro-Wilk-Tests verworfen.
Ho: Die Werte stammen aus einer normalverteilten GG.

Die Varianzen müssen homogen sein:

```
var.test(x1, x2)
```

Auch hier: $p \leq 0.05$ weist die Ho zurück.
Ho: Die Varianzen von x1 und x2 sind homogen.

Solche Tests sind umstritten, weil sie angeblich zu empfindlich reagieren.
Zuur u. a. 2009 empfehlen z. B. grafische Methoden. Ich nicht.

Wenn Voraussetzungen nicht erfüllt sind:

- steigt das Risiko für Typ 1-Fehler
- nicht-parametrische Alternative nehmen
- Daten transformieren (Logarithmus für Normalverteilung)
- sich über Robustheit des Test ggü. verletzten Annahmen informieren (oft schwer zugängliche und kontroverse Spezialliteratur)

- Alternativen, wenn Bedingungen für t-Test und ANOVA nicht erfüllt sind (Normalverteilung, Varianzhomogenität)
- Prinzip: **Umrechnen von Werten in Ränge**
- nicht-parametrische Tests

- Bortz & Lienert 2008
- Gravetter & Wallnau 2007

- Mann-Whitney U-Test: Alternative zum t-Test mit zwei Stichproben
- Kruskal-Wallis H-Test: Alternative zur einfaktoriellen ANOVA

- Intervallskalierung der Abhängigen
- Normalität der Abhängigen
- Varianzhomogenität der Abhängigen in den Gruppen
- Unabhängigkeit der Messungen

Alle bis auf die letzte entfallen beim Mann-Whitney U-Test.

Gruppen/Stichproben (Messwerte):

$$x_1 = [9, 8, 12, 16]$$

$$x_2 = [4, 11, 7, 13]$$

Ränge in der **zusammengelegten** Stichprobe:

$$X = [4, 7, 8, 9, 11, 12, 13, 16]$$

$$R(x_1) = [4, 3, 6, 8]$$

$$R(x_2) = [1, 5, 2, 7]$$

Addiere für jeden Wert beider Gruppen die Anzahl der **niedrigeren Ränge (=höhere Rangzahl!)** in der anderen Gruppe:

$$U(x_1) = 2 + 2 + 1 + 0 = 5$$

$$U(x_2) = 4 + 2 + 4 + 1 = 11$$

$$U = \min(U_{x_1}, U_{x_2}) = U_{x_1} = 5$$

$$U(x_\alpha) = n_1 \cdot n_2 + \frac{n_\alpha(n_\alpha+1)}{2} - \sum R(x_\alpha)$$

- $\sum R(x_1) = 4 + 3 + 6 + 8 = 21$
- $\sum R(x_2) = 1 + 5 + 2 + 7 = 15$
- $n_1 \cdot n_2 = 4 \cdot 4 = 16$
- $n_1(n_1 + 1) = n_2(n_2 + 1) = 4 \cdot 5 = 20$
- $U(x_1) = 16 + 10 - 21 = 5$
- $U(x_2) = 16 + 10 - 15 = 11$
- $U = 5$

- Signifikanz für kleine Stichproben: [Tabelle](#)
- bei großen Stichproben: U ggf. normalverteilt, also [z-Test](#)
- in R:

```
> wilcox.test(x1,x2, paired = FALSE)
```

- Effektstärke: Punkt-biserielle Korrelation
- entspricht Pearson-Korrelation, aber Unabhängige ist dichotom
- In R: `cor(c(x1,x2), c(rep(0,4),rep(1,4)))`
- alternativ: „relativer Effekt“ (Bortz & Lienert, S. 142)

- Bei sehr vielen gleichen Rängen ist der Mann-Whitney U-Test unzuverlässig.
- Bei gleichen Rängen generell: korrigierte Version (s. Bortz & Lienert, S. 146).
- Er ist daher nur begrenzt geeignet für Dinge wie 5-Punkt-Skalen.
- generell am stärksten bei gleich großen und gleich stark streuenden Stichproben
- letzter Ausweg: **Mediantest** (Bortz & Lienert, S. 137)

Wie vom t-Test zur ANOVA...

$$x_1 = [9, 8, 12, 16]$$

$$x_2 = [4, 11, 7, 13]$$

$$x_3 = [13, 12, 5, 15]$$

Gleiches Vorgehen wie bei Mann-Whitney über

Rang in der zusammengelegten Stichprobe:

X	4	5	7	8	9	11	12	12	13	13	15	16
R(X)	1	2	3	4	5	6	7.5		9.5		11	12

$$R(x_1) = [5, 4, 7.5, 12]$$

$$R(x_2) = [1, 6, 3, 9.5]$$

$$R(x_3) = [9.5, 7.5, 2, 11]$$

$$H = \frac{12}{N(N+1)} \cdot \sum_i \frac{(\sum R(x_i))^2}{n_i} - 3(N+1)$$

Am Beispiel:

- Gruppen-Rang-Summen:

- ▶ $R(x_1) = [5, 4, 7.5, 12], \sum R(x_1) = 28.5$

- ▶ $R(x_2) = [1, 6, 3, 9.5], \sum R(x_2) = 19.5$

- ▶ $R(x_3) = [9.5, 7.5, 2, 11], \sum R(x_3) = 30$

- $H = \frac{12}{12 \cdot (12+1)} \cdot \left(\frac{28.5^2}{4} + \frac{19.5^2}{4} + \frac{30^2}{4} \right) - 3(12+1) =$

- $0.077 \cdot (203.06 + 95.06 + 225) - 39 = 1.28$

- Bei $n > 5$ ist H unter der H_0 χ^2 -verteilt.
- mit $df = k - 1$ (k ist die Anzahl der Gruppen)
- Effektstärke: t_j ...
- „relative Effekte“ sind rechenbar (Bortz & Lienert, S. 159)

```
> kruskal.test(c(x1,x2,x3) c(rep(0,4),rep(1,4),rep(2,4)))
```

Rechnen Sie bitte mal die U- und H-Tests von diese Folien
und vergleichen Sie die p-Werte mit denen von t-Test und ANOVA
über die gleichen Daten:

$$x_1 = [9, 8, 12, 16]$$

$$x_2 = [4, 11, 7, 13]$$

$$x_3 = [13, 12, 5, 15]$$

Power

Lineare Modelle

- Gravetter & Wallnau 2007
- Zuur u. a. 2009
- Maxwell & Delaney 2004

- Wiederholung der Pearson-Korrelation (r , r^2)
- Signifikanztests mit Korrelationen
- Unterschied von Pearsons r zu Spearman's Rang-Korrelation
- Unterschiede zwischen Korrelation und Regression
- Berechnung linearer Regressionsmodelle
- Signifikanztests für Modell und Koeffizienten

$$r(x_1, x_2) = \frac{\text{cov}(x_1, x_2)}{s(x_1) \cdot s(x_2)}$$

In Gravetter & Wallnau, Kap. 16 lautet die Formel:

$$r = \frac{SP}{\sqrt{SQ_x \cdot SQ_y}}$$

Die Formeln sind äquivalent, weil (mit x, y statt x_1, x_2):

$$\begin{aligned} r(x, y) &= \frac{\text{cov}(x, y)}{s(x) \cdot s(y)} = \frac{\frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1} \cdot \frac{\sum (y_i - \bar{y})^2}{n-1}}} = \frac{\frac{SP(x, y)}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x}) \cdot \sum (y_i - \bar{y})}{n-1}}} = \\ &= \frac{\frac{SP(x, y)}{n-1}}{\frac{\sqrt{\sum (x_i - \bar{x}) \cdot \sum (y_i - \bar{y})}}{n-1}} = \frac{SP(x, y)}{n-1} \cdot \frac{n-1}{\sqrt{SQ(x) \cdot SQ(y)}} = \frac{SP(x, y)}{\sqrt{SQ(x) \cdot SQ(y)}} \end{aligned}$$

- Maß der Varianzerklärung durch r : r^2 (vgl. t-Test)
- **Signifikanztest** möglich: Schluss auf Korrelation in der Grundgesamtheit
- $df_r = n - 2$
- Unter der H_0 (keine Korrelation) t-verteilt:
$$t = r \sqrt{\frac{n-2}{1-r^2}}$$
- ...oder Tabellen (z. B. G&W, B.6)

- Intervallskalierung
- lineare Abhängigkeit
- bei kleinen n : Normalverteilung für x und y

- wenn nicht: Spearmans Rang-Korrelation

- mathematisch nicht andere als eine Pearson-Korrleation
- vorher: Umrechnung der rohen x,y-Werte in Ränge
- bei gleichen Werten: alle gleichen Werte bekommen Rang-Mittel

Werte in Ränge umrechnen

Ein Beispiel zur Umwandlung in Ränge:

Index:	1	2	3	4	5
Messwerte x:	4	7	3	1	3
Messwerte y:	9	12	11	2	8

Statt der Messwerte arbeitet man mit den Rängen der Messwerte an den jeweiligen Indexen.

Index:	1	2	3	4	5
Ränge der Messwerte x:	4	5	2.5	1	2.5
Ränge der Messwerte y:	3	5	4	1	2

Wenn $Rang(x_i)$ der Rang für x_i in x ist:

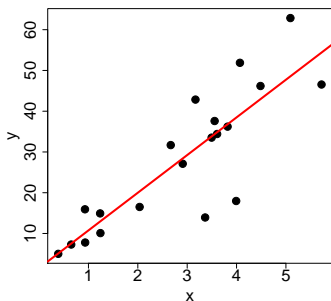
Spearman's Rang-Korrelation:

$$r_S = 1 - \frac{6 \sum_{i=1}^n (Rang(x_i) - Rang(y_i))^2}{n(n^2 - 1)}$$

Unterschiede zwischen Korrelation und Regression

- Korrelation: Stärke des Zusammenhangs
- Regression: genaue Funktion zur Modellierung des Zusammenhangs
- Korrelation: Diagnostik/Test
- Regression: Vorhersage (und Test)

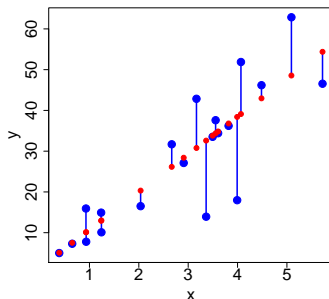
Spezifikation der Funktion für die Regressionsgerade



- Schnittpunkt mit der y-Achse (**Intercept**): a
- Steigung (**Slope**): b (b heißt auch **Koeffizient**)
- **Regressionsgleichung (=Modell)**: $\hat{y} = b \cdot x + a$
- Für jeden beobachteten Wert: $y_i = b \cdot x_i + a + e_i$ (e_i als Fehlerterm)

Idee der kleinsten Quadrate

Die vom Modell vorhergesagten Werte (rot, auf der Regressionsgerade) sollen insgesamt einen so geringen Abstand wie möglich zu den Beobachtungen (blau) haben.



Die Summe der **quadratierten** negativen und positiven Differenzen (blau) soll **minimiert** werden (=kleinste Quadrate): Minimierung von $\sum e^2$

Berechnung der Regressionsgleichung

- Slope/Steigung: $b = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$
- Intercept: $a = \bar{y} - b \cdot \bar{x}$
- Der Beweis, dass dies die Gerade mit den kleinsten Quadraten schätzt, erfordert bereits erheblichen mathematischen Aufwand, den wir uns sparen.
- Determinationskoeffizient: $r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$

- Wie stark variiert der Fehler für Stichproben einer Größe?
- $SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$
- Je kleiner $SF_{residual}$, desto besser das Modell.
- Beachte: n wird größer (größere Stichprobe): $SF_{residual}$ wird kleiner.
- Und: Fehler e werden kleiner: $SF_{residual}$ wird kleiner.

- Wie bei ANOVA: $F = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} = \frac{S_{\text{regression}}^2}{S_{\text{residual}}^2}$
- zufällige Varianz: $S_{\text{residual}}^2 = \frac{(1-r^2) \cdot \text{SQ}(y)}{1}$
- erklärte Varianz: $S_{\text{regression}}^2 = \frac{r^2 \cdot \text{SQ}(y)}{n-2}$
- Freiheitsgrade sind immer $df_1 = 1$ und $df_2 = n - 1$.
- Beachte: r^2 ist in $[0..1]$ und teilt die Varianz von y auf.

- Für b und a kann je ein Standardfehler angegeben werden.

- $SF(b) = \frac{\sqrt{\frac{\sum e^2}{n-1}}}{\sqrt{SQ(x)}}$

- Unter der H_0 : $b = 0$ ist dann t-verteilt:

$$t = \frac{b}{SF(b)}$$

- Design bei einfachem LM:
 - ▶ eine intervallskalierte Abhängige
 - ▶ eine Unabhängige
- wie bei mehrfaktorieller ANOVA:
 - ▶ oft interessiert **mehrfaktorielle Abhängigkeit**

Mehrere Koeffizienten im **allgemeinen linearen Modell**:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 \dots b_n \cdot x_n + a$$

Konzeptuell bleibt die Berechnung aller Werte und Tests gleich,
die Mathematik wird ungleich komplizierter.

Man schreibt R^2 statt r^2 .

Die Residuen müssen normalverteilt sein.
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität
- denn: Für jedes x_i müssen die e normalverteilt sein.
- erfordert mehrere Messungen pro x_i oder Intervallbildung
- größere Stichproben, kleinere Probleme
- visuelle Diagnose: Q-Q-Plots (hier nicht behandelt)

Jedes y_i darf nur von x_i abhängen,
niemals zusätzlich von x_j mit $i \neq j$.

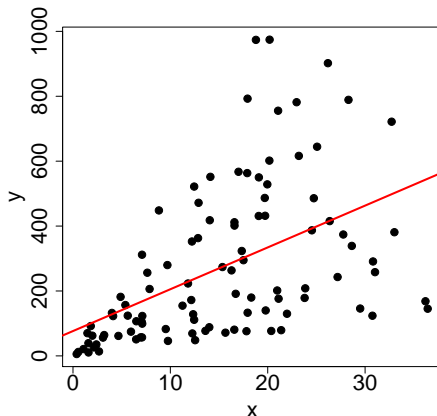
- mathematisch: nicht-lineare Abhängigkeit
- konzeptuell: Zeitserien
- konzeptuell: Sequenzen in Texten
- Lösung: andere Modellspezifikation

Die Residuen müssen homoskedastisch verteilt sein.

- Bedeutung: Die Varianz der e muss über alle x homogen sein.
- vgl. die Forderung der „Varianzhomogenität“ bei t-Test und ANOVA

Darstellung heteroskedastischer Residuen

Hier wird die Varianz der Residuen mit steigendem x immer größer.
Ein lineares Modell versagt hier wegen verletzter Verteilungsannahmen.



- mehr Daten ziehen, Daten transformieren
- generalisierte lineare Modelle (GLM)
legen andere Verteilungsannahmen zugrunde
- (generalisiert) additive Modelle (GAM)
schätzen Smoothingfunktionen für Koeffizienten

ANOVA als Modell mit kategorialen Regressoren

n Gruppen der ANOVA können als n dichotome Variablen dargestellt werden:

		ANOVA-Gruppen		
		A_1	A_2	A_3
Regressor	$x_1 =$	1	0	0
	$x_2 =$	0	1	0
	$x_3 =$	0	0	1

Normale Modellspezifikation:

$$\hat{y} = b_1x_1 + b_2x_2 + \cdots + b_nx_n + a$$

Da jeweils nur eins der $x_i = 1$ und alle anderen immer 0 werden, wird einfach der Wert des entsprechenden β_i (plus a) vorhergesagt.

Die Funktion `cor()` hat ein Argument `method`, das als "spearman" angegeben werden kann.

```
> cor(x, y, method = "spearman")
```


- Modellformeln: $y \sim x$
„y abhängig von x“
- Mehrere Unabhängige: $y \sim x_1 + x_2$
- Mehrere Unabhängige mit Interaktion: $y \sim x_1 * x_2$
- Mehrere Unabhängige nur Interaktion: $y \sim x_1 : x_2$

- Lineares Modell schätzen und speichern:

```
> m <- lm(y~x)
```
- Ausgabe Evaluation:

```
> summary(m)
```

Interpretieren Sie diese Ausgabe anhand der Folien:

```
Call:
lm(formula = y ~ x)

Residuals:
Min      1Q  Median      3Q      Max
-20.4298 -2.4920 -0.2625  3.8038 14.2922

Coefficients:
(Intercept)      1.513      4.321      0.350      0.73
x              9.242      1.333      6.933 1.77e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.008 on 18 degrees of freedom
Multiple R-squared:  0.7275, Adjusted R-squared:  0.7124
F-statistic: 48.06 on 1 and 18 DF,  p-value: 1.768e-06
```

Generalisierte Lineare Modelle

- Generalisierte Lineare Modelle mit Logit-Link = Logistische Regression
- Regression zur Modellierung dichotomer Abhängiger
- Modellselektion für GLMs
- Modellevaluation für GLMs
- Problemlösungen (Ausblick):
Zufallseffekte (GLMMs), Kreuzvalidierung, Bootstrapping, GAMs

- Backhaus u. a. 2011
- Zuur u. a. 2009
- Fahrmeir u. a. 2009

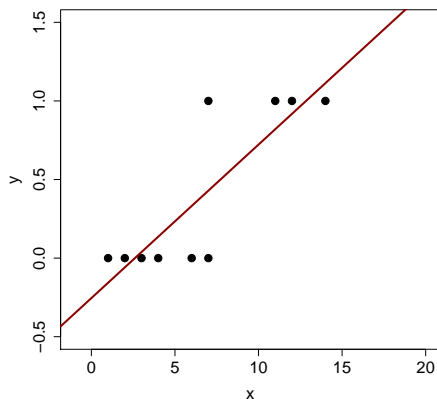
Alternation von Genitiv und Kasusidentität
in der Maßangabe im Deutschen:

- *Wir trinken eine Flasche guten Wein.* (Agree=1)
- *Wir trinken eine Flasche guten Weines.* (Agree=0)
- Welche Faktoren beeinflussen die Wahl von Agree=1 oder Agree=0?
- Unabhängige hier:
 - ▶ Kasus der Maßangabe (Nom, Akk, Dat)
 - ▶ Definitheit der NP (0, 1)
 - ▶ Maß ist als Zahl geschrieben (0, 1)
- Das Beispiel kommt dann in der R-Session tatsächlich dran.

- LM sagt **kontinuierliche Werte** voraus
- unplausibel für dichotome Abhängige
- auch als Eintrittswahrscheinlichkeit unplausibel (außerhalb $[0,1]$)
- **Normalitätsannahmen nicht erfüllt**

Illustration der Probleme

Datenpunkte einer dichotomen Abhängigen y
zu einer intervallskalierten Unabhängigen x
und lineares Modell $y \sim x$



- Vorhersage der Eintrittswahrscheinlichkeiten
- lineare Kombination der Regressoren wie beim LM
- Linearkombination ergibt die Logits (z):

$$z = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0$$

Die Logits werden transformiert in Eintrittswahrscheinlichkeiten mittels der **logistischen Funktion** (e ist die Euler-Konstante):

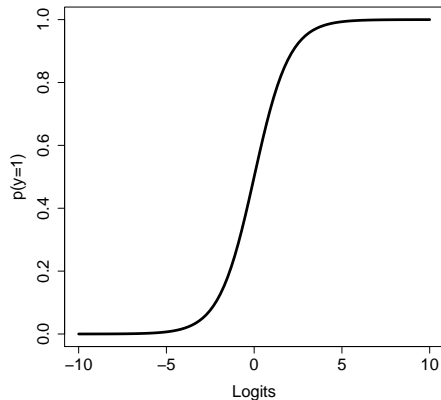
$$\hat{p}(y = 1) = \frac{1}{1+e^{-z}}$$

Bei der **binären Vorhersage** dann:

$$\hat{y} = \begin{cases} 0 & \text{wenn } \hat{p}(y = 1) \leq 0.5 \\ 1 & \text{wenn } \hat{p}(y = 1) > 0.5 \end{cases}$$

Darstellung des Effekts der Logit-Transformation

Die transformierten Logits als $\hat{p}(y = 1)$:



- Interpretation der Koeffizienten nur **indirekt** möglich
- β_i positiv \Rightarrow positiver Einfluss auf $\hat{p}(y = 1)$
- β_i negativ \Rightarrow negativer Einfluss auf $\hat{p}(y = 1)$
- Stärke des Einflusses: **nicht linear**
- linearer Einfluss nur auf die Logits, nicht auf $\hat{p}(y = 1)$

- Chance (Odds): $o(y=1) = \frac{p(y=1)}{1-p(y=1)}$
- Die Chancen des Modells verteilen sich (zum Glück) einfach:

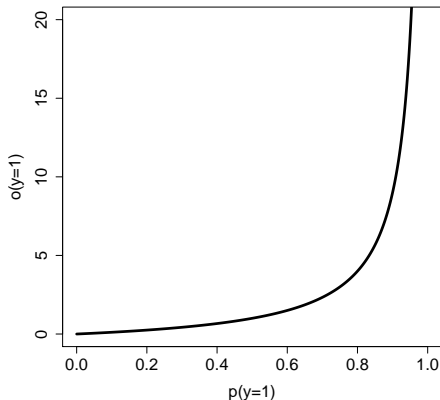
$$o(y=1) = \frac{p(y=1)}{1-p(y=1)} = e^z$$

Beachte: $\ln(e^z) = z = \text{Logits}$

- Die Chance liegt offensichtlich in $[0, \infty]$.
- Mit steigender Wahrscheinlichkeit gehen die Odds gegen ∞ .
- Bei einem Logit von 3 ist die Chance für $y = 1$ doppelt so hoch wie bei einem Logit von 1.5 usw.

Beziehung zwischen Wahrscheinlichkeit und Odds

In der Interpretation stellen die Odds die Linearität her, die den Wahrscheinlichkeiten bei der log. Regression fehlen.



Für die Interpretation der einzelnen Koeffizienten β_i
im Sinne eines Chancenverhältnisses:

$$or(y = 1|x_i) = e^{\beta_i}$$

In Worten: Steigt x_i (intervallskaliert!) um eine Einheit,
dann steigt die Chance für $y = 1$ um e^{β_i} .

Ein Chancenverhältnis von 1 entspricht einem Koeffizienten 0,
also einem ohne jeglichen Effekt.

Beziehungen zwischen den Maßen
sowie ihre Wertebereiche.

Einzel-Koeffizient		Gesamtmodell		
Koeffizient	Chancenverhältnis	Logit	Chance	$\hat{p}(y = 1)$
$\beta > 0$	$e^{\beta} > 1$	steigt um βx	steigt um $e^{\beta x}$	steigt
$\beta < 0$	$e^{\beta} < 1$	sinkt um βx	sinkt um $e^{\beta x}$	sinkt
$[-\infty, +\infty]$	$[0, +\infty]$	$[-\infty, +\infty]$	$[0, +\infty]$	$[0, 1]$

- Es gibt keine direkte Lösung für die Koeffizientenberechnung.
- Das Schätzverfahren funktioniert iterativ.
- Es kommt der sog. Maximum-Likelihood-Schätzer zum Einsatz.

- Es gibt beliebig viele Modelle = Belegungen für die β -Koeffizienten
- Das **wahrscheinlichste Modell angesichts der Beobachtungen** ist zu finden.
- In den Beobachtungsdaten für jeden Fall k : $y_k = 1$ oder $y_k = 0$
- Für jeden Beobachtungswert y_k betrachtet man:

$$p_k = \left(\frac{1}{1+e^{-z_k}} \right)^{y_k} \cdot \left(1 - \frac{1}{1+e^{-z_k}} \right)^{1-y_k}$$

$$p_k = \left(\frac{1}{1+e^{-z_k}}\right)^{y_k} \cdot \left(1 - \frac{1}{1+e^{-z_k}}\right)^{1-y_k}$$

- z_k ist der Modell-Logit für die zu y_k empirische gemessenen x .
- In den () steht links die vom Model geschätzte Wahrscheinlichkeit $\hat{p}(y_k)$ und rechts jeweils die Gegenwahrscheinlichkeit dazu $1 - \hat{p}(y_k)$.
- Wenn der Modellwert nahe an 0 (z. B. 0.1) und $y_k = 0$ ist:
 $p_k = (0.1)^0 \cdot (0.9)^1 = 1 \cdot 0.9 = 0.9$ („gute“ Approximation)
- Wenn der Modellwert bei gleichen empirischen Daten umgekehrt ist:
 $p_k = (0.9)^0 \cdot (0.1)^1 = 1 \cdot 0.1 = 0.1$ („schlechte“ Approximation)
- Die p_k messen also die Güte der vom Modell vorhergesagten Wahrscheinlichkeit für jeden beobachteten Datenpunkt.

- Bei unabhängigen Ereignissen $E_{1..n}$ gilt:
$$P(E_1 + E_2 + \cdots + E_n) = \prod_i P(E_i)$$
- Die Wahrscheinlichkeit eines Modells (seine „Likelihood“) angesichts aller empirischen Werte y_k ist also:

$$L = \prod_k p_k$$

- Der Maximum Likelihood-Schätzer maximiert L für die Belegungen der β -Koeffizienten (= konkurrierende Modelle).

Wie bei der LM-Variante der ANOVA müssen kategoriale Unabhängige mit mehr als zwei Ausprägungen als dichotome Dummy-Variablen kodiert werden.

Beispiel für dreiwertige Variable A und Dummy-Regressoren $x_{1..3}$

	A = 1	A = 2	A = 3
$x_1 =$	1	0	0
$x_2 =$	0	1	0
$x_3 =$	0	0	1

Achtung! De facto gibt es für einen kategorialen Regressor mit k Ausprägungen nur $k - 1$ Dummies (s. Abschnitt zum Intercept).

Beispiel für eine als $x_{1..3}$ dummy-kodierte Unabhängige A und eine intervallskalierte Unabhängige x_4 :

$$\hat{p}(y = 1) = \frac{1}{1 + e^{-z}}$$

$$\text{mit } z = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_0$$

Dabei treten die Werte auf:

- $x_{1..3}$: 0 oder 1
- Wenn $x_1 = 1$, dann $x_2 = 0$ und $x_3 = 0$ usw.

(Wh.:) Für die Interpretation der einzelnen Koeffizienten β_i
im Sinne eines Chancenverhältnisses:

$$or(y = 1|x_i) = e^{\beta_i}$$

In Worten für nominale Regressoren bzw. ihr dichotomen Dummies:

Wenn $x_i = 1$ (x_i ist dichotom skaliert!),
dann ist die Chance $o(y = 1)$ um e^{β_i} höher als bei $x_i = 0$.
Andere Fälle gibt es wegen der dichotomen Skalierung nicht.

- „Intercept“ (β_0) in GLMs \neq Schnittpunkt mit y-Achse
- **intervallskalierte Regressoren:**
 - ▶ einfachstes binomiales GLM: $\hat{p}(y = 1) = \beta_1 x_1 + \beta_0$
 - ▶ Wenn $x_1 = 0$, wird β_0 vorhergesagt.
- bei **Dummy-Variablen** wird eine zur Referenz-Kategorie:
 - ▶ GLM mit drei Dummies: $\hat{p}(y = 1) = \beta_{Akk} \cdot x_{Akk} + \beta_{Dat} \cdot x_{Dat} + \beta_{Nom}$
 - ▶ „Alle Regressoren werden 0“ heißt hier, es liegt Nom vor.
 - ▶ Die Dummies modellieren den **Unterschied zwischen Referenz (Nom) und den anderen Fällen**.
 - ▶ Die Referenzkategorie sollte die häufigste sein, besonders bei Interaktionen.

- nichts wesentlich anderes als in LM
- vereinte Effekte, die über die Einzeleffekte hinausgehen
- bei Interpretationsschwierigkeiten ggf. nachlesen

- Signifikanz wird für das Modell und Koeffizienten bestimmt.
- Allerdings: Signifikanz heißt nicht automatisch Modellgüte.
- Je „weniger signifikant“ ein Regressor, desto wahrscheinlicher kann er ohne Güteverlust entfernt werden.
- Modellselektion: Auswahl des **einfachsten Modells** mit der **größten Modellgüte**.
- Achtung bei dichotomen Dummy-Regressoren:
Immer **alle** Dummies im Modell lassen oder herausnehmen, die zu einer kategorialen Unabhängigen gehören!

- 1 Weglassen des Regressors mit der geringsten Signifikanz
- 2 Vergleich des vollen und des reduzierten Modells
- 3 bei nicht-signifikantem Unterschied: Regressor weglassen
- 4 von vorne beginnen...

Log-Likelihood-Ratio für Likelihood des vollen (L_f) und reduzierten (L_r) Modells:

$$LR = (-2 \cdot \ln(L_r)) - (-2 \cdot \ln(L_f))$$

Test: Unter der H_0 $L_r = L_f$ ist die LR χ^2 -verteilt
mit $df = df_f - df_r$ (df jeweils: Zahl der Regressoren)

Ist die LR größer als der kritische Wert: Regressor im Modell lassen!

Regressoren-Selektion auf Basis des **Akaike Information Criterion**:

- Ablauf wie bei LR-Test
- Maß für Modellvergleich ist das AIC
- Informationstheoretisches Maß:
Distanz des Modells zur (geschätzten) absoluten Realität
- Je kleiner das AIC, desto besser das Modell.
- Achtung: Nur zum Vergleich **eingebetteter Modelle** verwenden, also bei gleichem Datensatz, und wenn das reduzierte Modell eine Teilmenge der Regressoren des vollen enthält.

- Signifikanzbestimmung für einzelne Regressoren
- wie bei LM: **Standardfehler** für jeden Regressor
- darauf basierend: **z-Wert** für jeden Regressor...
- und **z-Test** auf Basis der Normalverteilung

- Log-Likelihood-Ratio-Test für Gesamtheit aller Regressoren
- volles Modell (ggf. nach Eliminierung von Koeffizienten)
- **Nullmodell**, das nur einen konstanten Term zur Vorhersage nutzt
- ähnlich den Modellvergleichen im Kapitel „ANOVA als LM“

- auch Vergleich des vollen Modells und Nullmodells
- Interpretation wie gewohnt: Varianzerklärung

$$\text{Cox \& Snell: } R_C^2 = 1 - \left(\frac{L_0}{L_f}\right)^{\frac{2}{n}}$$

Problem: Geht nicht bis 1!

$$\text{Nagelkerke: } R_N^2 = \frac{R_C^2}{R_{max}^2}$$

$$\text{mit } R_{max}^2 = 1 - (L_0)^{\frac{2}{n}}$$

- gutes GLM \Rightarrow gute Vorhersagen
- einfache Vorhersagegüte: Anteil der richtigen Vorhersagen
- instruktiv: Vergleich mit „Baseline“
(= Anteil der richtigen Vorhersagen bei Vorhersage der modalen Kategorie)
- Problem wie bei Fehlerreduktion:
auch bei starkem Effekt nicht unbedingt Umkehrung der modalen Kategorie

- zugrundegelegte Verteilung: **Binomialverteilung**
- Überdispersion: Varianz ist größer als für Binomialverteilung angenommen
- mögliche Gründe:
 - ▶ unbeobachtete Heterogenität (fehlende erklärende Variablen)
 - ▶ Gruppenbildung (= Beobachtungen nicht unabhängig)

Schätzung des **Dispersionsparameters**:

$$\hat{\phi} = \sum \left(\frac{R_P}{df_R} \right)^2$$

wobei: R_P ist das **Pearson-Residual** (hier nicht behandelt) und

df_R die **Residual-Freiheitsgrade** $n - p$, p die Anzahl der Modellparameter

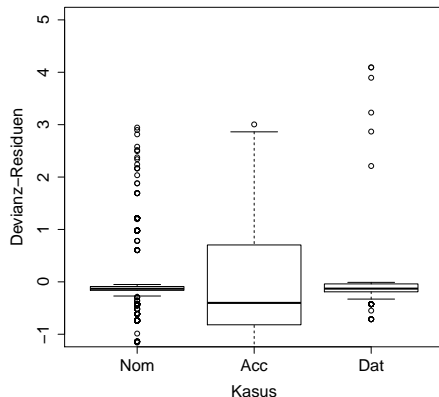
- Problem: $\hat{\phi}$ deutlich über 1
- Lösung: Schätzung der Parameter bleibt (im Ergebnis) gleich
- aber für die Evaluation der Koeffizienten:
 - ▶ Signifikanzschätzung mit größeren Standardfehlern
 - ▶ t-Verteilung statt Normalverteilung (z-Werte)
- Ein „Quasi-Likelihood-Modell“ folgt im Wesentlichen dieser Strategie.

- (Multi-)kollinearität: Abhängigkeit zwischen Regressoren
- Probleme: β -Fehler, Überanpassung, ungenaue Koeffizientenschätzung
- Test: Varianzinflations-Faktoren (nicht im Detail behandelt)
- Lösungen z. B.: mehr Daten, Regressoren weglassen
- Test des Modells auf Robustheit trotz Kollinearität (z. B. Kreuzvalidierung)

Varianzhomogenität

Die Residuen werden im GLM zwar anders berechnet, sind aber trotzdem ein Maß für die Varianz.

Die Varianz sollte nicht mit den Regressorausprägungen variieren!



- bei Problemen: Test auf **Robustheit des Modells**
- Idee bei k -facher Kreuzvalidierung:
 - 1 teile Daten in k Teile
 - 2 Modellanpassung auf $k - 1$ von k Teilen
 - 3 Prüfung der Vorhersage auf verbleibendem Teil
 - 4 Modell ist Robust, wenn die Parameter in der Kreuzvalidierung nicht wesentlich anders geschätzt werden als im Ursprungsmodell
- wenn $k = n$: **Leave-One-Out-Kreuzvalidierung**
- verwandtes Verfahren: **Bootstrapping** (mit Zurücklegen)

Einige typische Anwendungsfälle für nicht-binomiale GLMs:

- Zähldaten: **Poisson**
- Zähldaten mit Überdispersion: **negativ-binomial**
- bestimmte Intervalldaten in $[0, \infty]$: **Gamma**
- viele Nullen: **zero-inflated** Varianten

Das Vademecum, vor allem für R-Benutzer:
Zuur u. a. 2009

- typisches gemischtes Modell: mit Zufallseffekten
- Idee: Varianzunterschiede oder Dispersion durch Gruppen
- mögliche Gruppen in linguistischen Experimenten:
 - ▶ Werte von einem Probanden bei Befragung, Rating-Studie
 - ▶ Werte zu einem Lexem bei Korpusstudie
 - ▶ Werte aus einer Textsorte bei Korpusstudie
- ideal: Gruppeneffekte durch zusätzliche normale Regressoren auflösen
- sonst (vereinfacht): Schätzung eines Intercepts pro Gruppe
- Typisch für Zufallseffekte: In der GG sind vermutlich viel mehr Ausprägungen vorhanden, als gemessen (wie z. B. Sprecher oder Lexeme) wurden.

GAMs oder „nichtparametrische Regression“

$$\hat{y} = f_1(x_1) + f_2(x_2) + \dots + f_n(x_n) + \beta_0$$

- f_n : besondere Art von Funktion, die geschätzt wird
- Wenn die Funktionen ungefähr linear sind, ist ein GLM genauso gut.
- Interpretation von GAMs: viel schwieriger als GLMs
- letzter Ausweg bei schlechtem GLM

1 Modell-Anpassung:

```
> m <- glm(y ~ x1+x2*y3, data=mydata, family="binomial")  
> summary(m)
```

2 Chancenverhältnisse für Koeffizienten:

```
> exp(coef(m))
```

3 95%-Konfidenzintervalle für Chancenverhältnisse:

```
> exp(confint(m))
```

4 Log-Likelihood extrahieren:

```
> logLik(m)
```

5 Nagelkerke R^2 :

```
> library(fmsb); NagelkerkeR2(m)
```

6 LR-Test:

```
> m0 <- glm(y ~ 1, data=mydata, family="binomial")  
> lr <- (-2*logLik(m0)) - (-2*logLik(m))  
> pchisq(lr, m$rank-m0$rank)
```

- 7 Modellselektion (wenn nicht von Hand):
`> drop1(m)`
- 8 Varianzinflationsfaktoren:
`> library(car); vif(m)`
- 9 Dispersion $\hat{\phi}$ schätzen:
`> sum(resid(m, type="pear")^2 / df.residual(m))`
- 10 Vorhersagegüte:
`> pred <- ifelse(predict(m) <= 0.5, 0, 1)`
`> tab <- table(pred, mydata$response)`
`> sum(diag(tab))/sum(tab)`
- 11 Fehlerrate in Kreuzvalidierung (hier $k = 10$):
`library(boot); cv.glm(mydata, m, K=10)$delta`

Gemischte Modelle

- Backhaus, Klaus, Bernd Erichson, Wulff Plinke & Rolf Weiber. 2011. *Multivariate Analysemethoden*. 13. Aufl. Berlin etc.: Springer.
- Bortz, Jürgen & Gustav Lienert. 2008. *Kurzgefasste Statistik für die klinische Forschung*. Heidelberg: Springer.
- Bortz, Jürgen & Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.
- Carnap, Rudolf. 1928. *Der logische Aufbau der Welt*. Berlin: Weltkreis Verlag.
- Cook, Philippa & Felix Bildhauer. 2013. Identifying “aboutness topics”: two annotation experiments. *Dialogue and Discourse* 4(2), 118–141.
- Duhem, Pierre. 1914. *La Théorie Physique: Son Objet et sa Structure*. Marcel Riviera & Cie.
- Fahrmeir, Ludwig, Thomas Kneib & Stefan Lang. 2009. *Regression – Modelle, Methoden und Anwendungen*. 2. Aufl. Heidelberg etc.: Springer.
- Fisher, Ronald A. 1935. *The Design of Experiments*. London: Macmillan.
- Gravetter, Frederick J. & Larry B. Wallnau. 2007. *Statistics for the Behavioral Sciences*. 7. Aufl. Belmont: Thomson.
- Laudan, Larry. 1990. Demystifying Underdetermination. In C. Wade Savage (Hrsg.), *Scientific Theories*, 267–297. Minneapolis: University of Minnesota Press.
- Maxwell, Scott E. & Harold D. Delaney. 2004. *Designing experiments and analyzing data: a model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.

- Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
- Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.
- Popper, Karl Raimund. 1962. *Conjections and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books.
- Quine, Willard Van Orman. 1951. From a Logical Point of View. In 2. Aufl. Cambridge: Harvard University Press. Kap. Two Dogmas of Empiricism, 20–46.
- Ronneberger-Sibold, Elke. 2010. Der Numerus – das Genus – die Klammer : die Entstehung der deutschen Nominalklammer im innergermanischen Vergleich. In Antje Dammel, Sebastian Kürschner & Damaris Nübling (Hrsg.), *Kontrastive Germanistische Linguistik. Teilband 2*, Bd. 206/209 (Germanistische Linguistik), 719–748. Hildesheim: Olms.
- Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.

Kontakt

Prof. Dr. Roland Schäfer
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena
Fürstengraben 30
07743 Jena

<https://rolandschaefer.net>
roland.schaefer@uni-jena.de

Creative Commons BY-SA-3.0-DE

Dieses Werk ist unter einer Creative Commons Lizenz vom Typ *Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Deutschland* zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie

<http://creativecommons.org/licenses/by-sa/3.0/de/> oder wenden Sie sich brieflich an Creative Commons, Postfach 1866, Mountain View, California, 94042, USA.