STADA

STATISTIK:

- **VORHANDENE** Daten: (Info über Sachverhalte zb innerbetrieblich, außerbetrieblich, Absatzzahlen, Statistik Austria,…)

- **ERHOBENE** Daten: Wissenschaft, die sich mit Erhebung von Daten, Methoden und Techniken zur Datenanalyse und der Aufbereitung von numerischen Daten beschäftigt

- Methodensammlung: Grafiken, komprimierte Kennzahlen zur Entscheidungsfindung

DATENANALYSEN:

- Wissenschaften beschäftigen sich mit Menschen, Medien, Marken, Produkten, Dienstleistungen usw. Diese Analyseeinheiten werden bezüglich ausgewählter, für eine bestimmte Fragestellung relevanter Merkmale beschrieben.
    - **Merkmale** der Analyseeinheiten = **Variablen**
    - **Ausprägungen** der **Merkmale** je Analyseeinheit = **Werte**
    - **Menge aller Merkmalsmessungen** über alle Analyseeinheiten = **Daten**

(groß-) **X**	**Merkmal**
(klein-) **x**	**Ausprägung des Merkmals**
x_i	**einzelne Messwerte (x) des Merkmals bei unterschiedlichen Subjekten/Objekten**
x1, x2,…, xn	**Ausprägungen der n Elemente in einer Stichprobe**
x15	**Ausprägung des 15. Elements der Stichprobe**
x1, x2,…, xN	**Ausprägung der N Elemente in einer Grundgesamtheit**

Urliste = N oder n Messwerte x_i werden in der Reihenfolge ihres Auftretens notiert (ungeordnete Rohdaten)

STATISTISCHE DATENANALYSE – ARTEN:

- **Deskriptive Statistik**
    - Zum **Beschreiben**, **Ordnen** und **Darstellen** von **Merkmalsverteilungen**
    - Zur **Informationsverdichtung** bereits erhobener Daten
    - Zur **Analyse** von **Zusammenhängen** einzelner Merkmale

- **Schließende Statistik** (Inferenzenstatistik, induktive Statistik)
    - Ist der **Zusammenhang** zwischen einzelnen **Merkmalen signifikant**?
    - Kann von **Ergebnissen** der Stichprobe **auf** die **Grundgesamtheit rückgeschlossen** werden?
    - Erlauben die **Ergebnisse** Aussagen über Grundgesamtheit, sind sie „**generalisierbar**" – gehen also über Ergebnisse einzelner Fälle hinaus?

BEGRIFFE:

Univariat	Untersuchung **EINZELNER** Merkmale (Variablen)
Bivariat	Untersuchung Zusammenhänge **zweier** Merkmale (Variablen)
Multivariat	Untersuchung Zusammenhänge **mehrerer** Merkmale (Variablen)

STATISTISCHE DATENANALYSE – **VERFAHREN**:

- **Deskriptive** Analyseverfahren
    - Zählung Häufigkeiten, deskriptive Statistiken, explorative Datenanalyse
    - Kreuztabellen, Mittelwertsvergleiche, Korrelationen
- **Schließende** Analyseverfahren
    - Chi^2-Test
    - Mittelwertsvergleichtests (T-Test, Varianzanalyse, U-Test,…)
    - Korrelationsanalysen
- **Multivariate** Analyseverfahren
    - Faktorenanalyse, Clusteranalyse
    - Diskriminanzanalyse, Regressionsanalyse
    - Kausalmodelle

QUALITATIV ≠ QUANTITATIV:

- Quantitativ = Statistik

MESSEN:

- Zuordnen von Zahlen zu Merkmalen/Antworten von Befragten
- Verhältnis der Zahlen zueinander entspricht den Relationen unter den Untersuchungsobjekten
- Dadurch werden präzise & systematische Informationen in leicht ablesbarer Form verfügbar

SKALEN:

- Bereich in dem die Messergebnisse schwanken
- Nicht die Antworten einer Frage bilden die Skala
  → sondern die ZAHLEN, die den Ausprägungen zugeordnet wurden

MESSNIVEAUS:

- je nach Messniveau sind verschiedene Rechenoperationen bei Auswertung sinnvoll & zulässig
- und die in Zahlen erfassten Antworten unterschiedlich zu interpretieren

- diskrete Skalen: (nicht aufwärtskompatibel, nur Häufigkeiten, keine Mittelwerte)
  - ordinal (A < B < C)
  - nominal (A ≠ B ≠ C)

- metrische Skalen: (abwärtskompatibel, Mittelwerte & Häufigkeiten)
  - rational (A = x*B)
  - intervall (B – A = D – C)

QUASIMETRISCHE SKALEN:

- Sozialforschung beschäftigt sich oft mit Konstrukten, die nur mit Schulnoten oder anderen Ratingskalen abgebildet werden können.
- Für alle Befragten gelernt und gut verständlich
- Wenn völlige Zustimmung = 1 = trifft sehr zu, völlige Ablehnung = 5 = trifft gar nicht zu, dann gilt: 1 und 2 sind EINE Skalenposition auseinander, 4 und 5 ebenso

→ In SOWI Forschungspraxis werden ALLE eigentlich ordinalen Ratingskalen fast immer und überall als quasimetrisch behandelt und damit tauglich für Mittelwerts und weitere statistische Berechnungen

DISKRET (KATEGORIAL) – STETIG (METRISCH)

- Diskrete Werte (Zahlen ohne rechnerische Bedeutung):
  Merkmal kann nur definierte Werte annehmen.
  (zb bis 30 Jahre alt = 1 | 30 Jahre und älter = 2)
  - Nominal und ordinalskalen
  - Auswertung: Häufigkeit, Kreuztabellen

- Stetige Werte (Zahlen mit rechnerischer Bedeutung)
  Merkmal kann jeden beliebigen Wert zwischen einem Minimal und Maximalwert annehmen.
  (zb Alter in Jahren)
  - Intervall und Ratioskalen
  - Auswertung: Mittelwert, Streuungsmaße, Korrelationen

INDIKATOREN:

In empirischer Sozialforschung steht meist nicht direkt Erfahrbares (theoretische Konstrukte) im Mittelpunkt → nicht direkt messbar

→ Theorie wird mit Indikatoren empirisch überprüfbar gemacht
Ein Indikator ist ein empirisch beobachtbarer Sachverhalt, der es zulässt, etwas nicht direkt Erfahrbares zu messen

- Indikator repräsentiert das nicht Messbare
- Ein richtiger Indikator korreliert hoch mit dem nicht direkt beobachtbaren Merkmal und muss **valide** und **reliabel** sein

EINSTELLUNGSSKALA

Einstellungen oder Meinungen können meist
- Nicht explizit
- Nur implizit gemessen werden (zb über Zustimmung zu bestimmten Aussagen)

- Oft umfassen derartige Messinstrumente mehrere Indikatoren (Fragen bzw .Aussagen), die die gesuchteN EinstellungEN (=Konstrukte) jeweils widerspiegeln sollen (Einstellungsskala)
- Bei der Auswertung wird dann aus diesen Fragen bzw. Aussagen **je Konstrukt ein gemittelter Gesamtwert** berechnet

Einstellungsskalen müssen valide sein!

Bsp.:
- Wie sehr treffen die folgenden Eigenschaften auf Sie zu?
  Urteilen Sie bitte zwischen 1 (trifft sehr zu) und 5 trifft garnicht zu. Dazwischen können Sie abstufen.

  „Ich trachte danach, täglich mit dem Öffi zur Arbeit zu fahren"
  → Als EIN Indikator für Umweltbewusstsein

  ! tägliche Fahrt mit ÖFFI muss nicht unbedingt mit Umweltbewusstsein zusammenhängen, sondern kann andere Gründe haben (kein Auto, keinen Führerschein, wenig Parkplatz, zu teuer,…)
  → **Dieser Indikator wäre fehlerhaft (nicht valide)**
  → Misst etwas anderes oder gar nichts

OBJEKTIVITÄT, VALIDITÄT, RELIABILITÄT

Eine Erhebung ist objektiv
- Wenn sie nicht durch die durchführende Person verzerrt wird
- Also **frei von subjektiven Einflüssen** ist

Ein Erhebungsinstrument (Fragebogen, Skala) ist **reliabel**, wenn das zu erhebende Merkmal
- Bei wiederholter Erhebung
- Unter den **gleichen Bedingungen**
- In **geringem zeitlichem Abstand**

In **gleicher Weise** ausgeprägt ist (konsistent).

Ein Erhebungsinstrument (Fragebogen, Skala) ist **valide**,
- Wenn es das **Merkmal**, das gemessen werden soll, auch **tatsächlich misst**.

FORSCHUNGSFRAGEN UND HYPOTHESEN

- Forschungsfragen:
  - Drücken neutrales Erkenntnisinteresse in Frageform aus
  - Definieren genaue Inhalte und Formulierungen im Erhebungsinstrument (Fragebogen, Leitfaden, Codierschema, Beobachtungsprotokoll)

  → Empirie beantwortet Forschungsfragen

- Hypothesen
  - Stellen Annahmen/Behauptungen auf, beruhend auf Basiswissen
  - Sind vermutete Antworten auf Forschungsfragen
  - Sollten nicht bloße Aussagen sein
    - Empfohlen: Konditional (Wenn, dann) und Vergleichssätze (Je desto, umso desto)

  → Empirie prüft vorab formulierte Hypothesen (quanti.) oder aus den empirischen Erkenntnissen werden Hypothesen formuliert (quali.)

- Forschungsfragen und Hypothesen sind unverzichtbar
  - Fragebogen benötigt GENAU SO VIELE Fragen wie zur Abdeckung der Forschungsfragen/Hypothesen erforderlich sind
  - Nicht jede Forschungsfrage benötigt eine Hypothese, nicht jede Hypothese eine Forschungsfrage

- Hypothesen (inhaltlich, nicht statistisch)
    - = Vermutungen über Ergebnisse von Datenerhebungen
    - H. formuliert das, was zu untersuchen ist, als (wissenschaftlich) theoriegestützt überprüfbare Aussage.

    - H. kann gerichtet (präziser) oder ungerichtet sein
    - H. soll möglichst kurz und leicht fassbar sein
    - SOWI – H. ist eine Wahrscheinlichkeitsaussage
    - H. zielt auf neuartige Erkenntnisse ab
    - H. kann nie vollständig verifiziert, höchstens momentan gestützt werden. Niemals können alle denkbaren Möglichkeiten der Überprüfung ausgeschöpft werden.

- Inhaltliche Hypothesen:
    - Formulieren inhaltl. Zusammenhänge: Wenn Personen älter als 40 sind, essen sie ehe Schnitzel, als wenn sie jünger sind.

- Statistische Hypothesen
    - Formulieren Operationalisierung und statistische Kennwerte:
        - **Nullhypothese**: Der Prozentanteil von Personen, die mehrmals wöchentlich Schnitzel essen, ist unter Personen, die älter als 40 sind, gleich (ähnlich) wie unter jüngeren Personen.
        - **Alternativhypothese**: Prozentanteil von Personen, die. Mehrmals wöchentlich Schnitzel essen, ist unter Personen, die älter als 40 sind höher, als unter jüngeren Personen.

      → in der Regel werden inhaltliche und statistische Hypothesen aufgestellt

VORGEHENSWEISE BEI STATISTISCHEN ANALYSEN

1. Forschungsfrage formulieren, Hypothesen aufstellen
2. Daten erheben
3. Daten analysieren

4. Unbedingt diese Fehler vermeiden:
    - FF/H. können nicht anhand derselben Daten geprüft werden, die für deren Formulierung verwendet wurden
    - FF/H. können nicht erst dann formuliert werden, wenn jemand Daten bereits erhoben hat
    - Fälle oder Variablen werden so lange ausgeschlossen, es wird so lange herumgerechnet bis Ergebnis vorliegt → höchstens zur Unterstützung des Nachdenkens VOR der Theoriebildung und tatsächlicher Erhebung zulässig!

ZUERST THEMENDETAILS (1), DANN ERST OPERATIONALISIERUNG (2)

(1) Forschungsfragen, Hypothesen, Themen (=Detailfragen)
Sammlung aller Befragungsinhalte
Exakte, ausführliche Problemformulierung
Zb: Welche Einstellung zu Schnitzelgerichten auf Speisekarten hat die österreichische Wohnbevölkerung?

(2) Konkretes Erhebungsinstrument (Operationalisierung):
Konkrete Fragen, Codebogen, Protokoll
Im konkreten Wortlaut ausformuliert
Zb: Frage in Fragebogen:
Wie sehr stimmen Sie der folgenden Aussage zu? Urteilen Sie bitte mit 0= stimme gar nicht zu, bis 7= stimme völlig zu
Dazwischen können Sie abstufen.

„Auf einer Speisekarte dürfen Schnitzelgerichte keinesfalls fehlen."

CODIERUNG
= **Zuordnung von Merkmalsausprägungen zu einer Variable** (Klassifikation) **in Form von Zahlen**
Erst diese Verschlüsselung der Antworten ermöglicht eine Messung der Variable

- Codierung VOR Datenerfassung
  Bei geschlossenen Fragen/fixierten Merkmalsausprägungen, wo die Codes von Beginn an vergeben und klar vorgegeben sind

- Codieren NACH Datenerfassung
  Bei offenen Fragen/nicht fixierten Merkmalsausprägungen, wo die Codes erst aus den Daten heraus konstruiert werden müssen.
  Hier zuerst Bildung von Dimensionen, dann Codierung, dann Antwortzuordnung

REGELN FÜR CODIERUNGEN:

- Codieren = jeder Antwort(kategorie) einen numerischen Wert zuzuordnen.
  - Jede Antwort muss sich einer Kategorie zuordnen lassen
  - Kategorien müssen einander ausschließen
  - Kategorien müssen eindimensional sein

- Umfang der Codes
  - Von Merkmal vorgegeben (zb. Bundesland)
  - Überschnittlichkeit! – zu viele versus wenige Codes (Gruppengröße)

  - Bei Mehrfachnennungen dichotome Codierung (1 oder 0)

- Jede Codierung muss klar dokumentiert werden
  - Sind zb Personen aus der Stadt Code 1 oder jene vom Land?

WEG DER DATENAUWERTUNG

Jede Datenanalyse durchläuft bestimmte Schritte

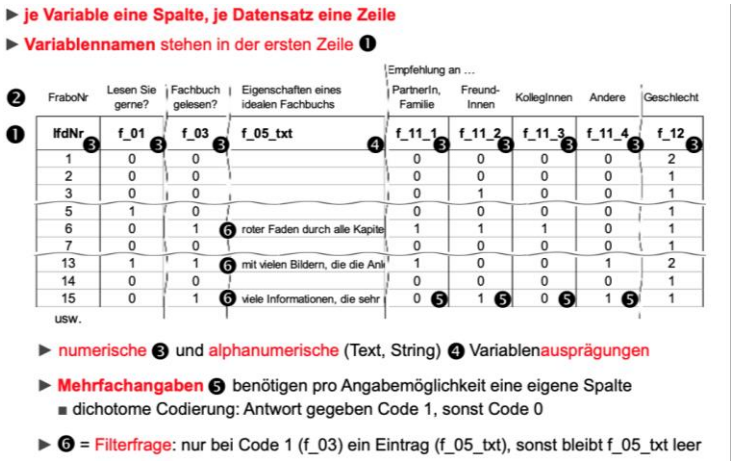
- Check Datenrücklauf: Datenstruktur? Repräsentativität? Gewichtung?
- Datenerfassung/Daten-Export (Erhebungssoftware) und -Import in zb. SPSS
- Konsistenzcheck (auf erlaubte bzw sinnvolle Codierungen, sowie falsche oder irreale Angaben)

- Technische Auswertung meist noch ohne Ergebnisinterpretation

- Offene Fragen werden anders ausgewertet als geschlossene Fragen

  → erst ganz am Ende: Ergebnisinterpretation: Aufbereiten und Interpretieren, Summary, Key Findings, Beantwortung der forschungsfragen usw.

AUFBAU EINES DATENFILES



MESSNIVEAUS UND AUSWERTUNG

VARIABLE 1	VARIABLE 2	VERFAHREN
Nominal/ordinal		Häufigkeitszählung
Metrisch/Skala		Mittelwert berechnen
Nominal/ordinal	Nominal/ordinal	Kreuztabelle
Metrisch/Skala	Neben metrisch/Skala Nach nominal/ordinal	Mittelwertsvergleich
Metrisch/Skala	Metrisch/Skala	Korrelation

HÄUFIGKEITSVERTEILUNG
Wie oft kommt jede Merkmalsausprägung in den Daten vor?

- Wenn ein Merkmal zu viele Ausprägungen besitzt, kann die Übersichtlichkeit gesteigert werden, indem man die Merkmale zu Klassen zusammenfasst

  Aber: je breiter die Klassen, desto weniger Information!
  Je weniger Klassen, desto weniger Information!

SYMBOLE

- **Absolute Häufigkeit** gibt an, wie oft eine Ausprägung i des Merkmals X in einer Grundgesamtheit mit Umfang N oder Stichprobe mit Umfang n auftritt
- **Relative Häufigkeit** gibt an, wie häufig eine Ausprägung i des Merkmals X im Verhältnis zu einer Grundgesamtheit mit dem Umfang N oder im Verhältnis zu einer Stichprobe mit Umfang n auftritt.

Für die Merkmalsausprägung xi bzw. die Klasse mit der Nummer i (einer Stichprobe) gilt:

- Absolute Häufigkeit ni
- Relative Häufigkeit (fi) ni/n
- Prozentuelle Häufigkeit ni/n 100[%]

ARTEN VON HÄUFIGKEITEN

1) Häufigkeit = Anzahl der Fälle je Merkmalsausprägung → Absolute Häufigkeiten
2) Prozent = Anzahl der Fälle je Ausprägung, relativiert zu ALLEN Fällen → relative Häufigkeiten
3) Gültige Prozente = Anzahl der Fälle je Ausprägung, relativiert an nur jenen Fällen, die eine Merkmalsausprägung haben (also ohne Fehlend)
4) Kumulierte Prozente = Prozentwerte summiert in steigender Reihenfolge der Merkmalsausprägungen, beginnend beim kleinsten Wert

ZENTRALMAßE:

Mittelwert und Median sind in der Praxis die beliebtesten Werte zur Beschreibung einer Verteilung

- Mittelwert (arithmetisches Mittel)
  - Hat Nachteile bei unsymmetrischen Verteilungen und rangvariablen
  - Achtung Nullwert # fehlender Wert!!!

- Median
  - Bietet Vorteile bei Ausreißern und schiefen Verteilungen
  - Möglich auch bei ungleichen Klassenintervallen
  - Und bei offenen Randklassen (zb einmal pro Woche, zweimal pro Woche, Öfter)

MITTELWERT UND MEDIAN

- Mittelwert: alle einzelnen Werte werden addiert und durch die Zahl der Messungen dividiert

- Median: teilt eine Verteilung so, dass sich 50% der in steigender Reihenfolge geordneten Messwerte unterhalb und 50% oberhalb befinden; bei gerader Fallzahl werden die beiden mittleren Werte addiert und dividiert

- Mittelwert aus Häufigkeitstabellen:

Note 1:	25 Personen
Note 2:	34 Personen
Note 3:	23 Personen
Note 4:	14 Personen
Note 5:	14 Personen
	n = 110

$$\bar{x} = \frac{1}{110} \cdot ((25 \cdot 1) + (34 \cdot 2) + (23 \cdot 3) + (14 \cdot 4) + (14 \cdot 5)) = 2{,}62$$

NORMALVERTEILUNG:

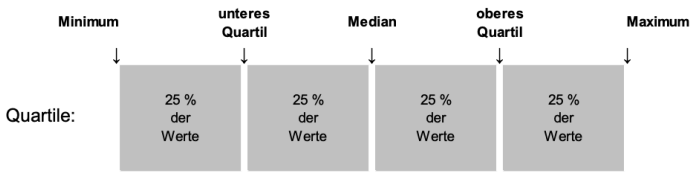
- Gesamtfläche unter Normalverteilungskurve ist immer 1 (100%)
- Normalverteilung ist symmetrisch
- Maximum der Kurve x = μ (Höchster Punkt der Kurve = Mittelwert)
- Wendepunkte der Kurve liegen bei μ-σ und μ+σ
- Je kleiner σ desto schlanker die Glockenkurve

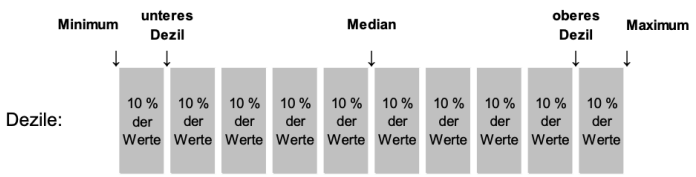
... und **schiefe Verteilung**

❶ Mittelwert
❷ Median
❸ Modalwert
► **Alter**
– die Bevölkerung wird älter

❸ Modalwert
❷ Median
❶ Mittelwert
► **Einkommen**
– wenige verdienen (sehr) viel

LAGEMAßE QUANTILE:





LAGEMAßE – (P-)QUANTILE:

QUANTIL = n * p

- Median: $x : p = 0{,}5$
- (4) Quartile: Q1 : p = 0,25 | Q2 : p = 0,5 | Q3 : p = 0,75
- (10) Dezile: D1 : p = 0,1 ……. D9 : p = 0,9

LAGEMAßE – ERMITTLUNG

In welchen Bereichen der Mess-Skala kommen die Beobachtungseinheiten zu liegen?

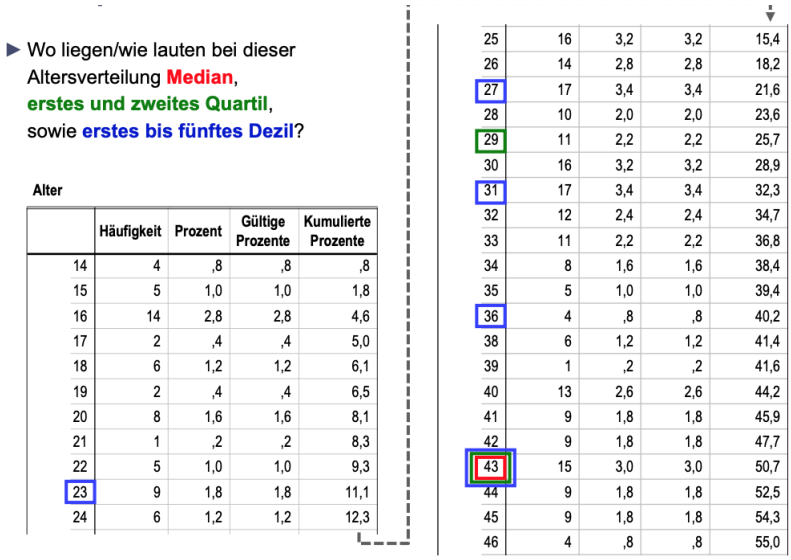
- Auch bei Ordinalskalen sinnvoll

→ Ermittlung des p-Quantils aus n Messwerten:

1. Erstellung einer Rangliste der n Messwerte
2. Bildung des Produkts aus n*p

→ wenn n*p = eine ganze Zahl:
Der Rangplatz dieser Zahl muss gesucht werden
Der Wert auf diesem Platz und der darauffolgende Wert werden gemittelt. Dieser Mittelwert bezeichnet den Wert des gesuchten p-Quantils

→ Wenn n*p = keine ganze Zahl
Die Zahl wird auf die nächstgrößere Zahl gerundet
Der Rangplatz der gerundeten Zahl muss gesucht werden
Der Wert auf diesem Platz bezeichnet das gesuchte p-Quantil.

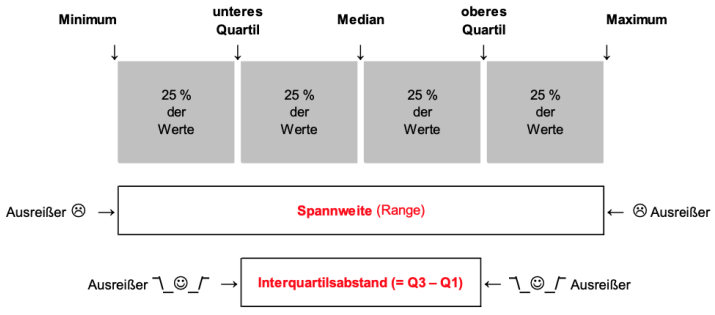


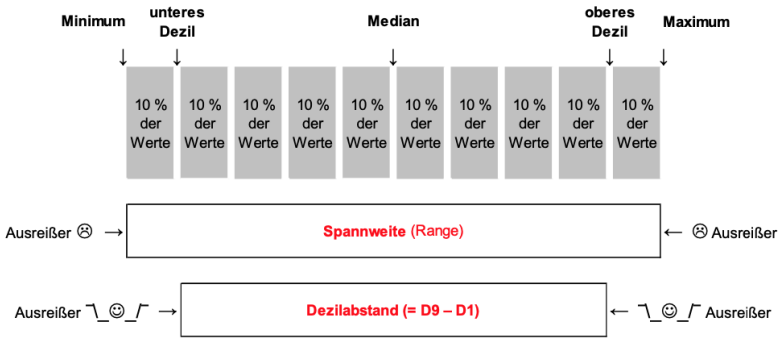
Alter

	Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
14	4	,8	,8	,8
15	5	1,0	1,0	1,8
16	14	2,8	2,8	4,6
17	2	,4	,4	5,0
18	6	1,2	1,2	6,1
19	2	,4	,4	6,5
20	8	1,6	1,6	8,1
21	1	,2	,2	8,3
22	5	1,0	1,0	9,3
23	9	1,8	1,8	11,1
24	6	1,2	1,2	12,3
25	16	3,2	3,2	15,4
26	14	2,8	2,8	18,2
27	17	3,4	3,4	21,6
28	10	2,0	2,0	23,6
29	11	2,2	2,2	25,7
30	16	3,2	3,2	28,9
31	17	3,4	3,4	32,3
32	12	2,4	2,4	34,7
33	11	2,2	2,2	36,8
34	8	1,6	1,6	38,4
35	5	1,0	1,0	39,4
36	4	,8	,8	40,2
38	6	1,2	1,2	41,4
39	1	,2	,2	41,6
40	13	2,6	2,6	44,2
41	9	1,8	1,8	45,9
42	9	1,8	1,8	47,7
43	15	3,0	3,0	50,7
44	9	1,8	1,8	52,5
45	9	1,8	1,8	54,3
46	4	,8	,8	55,0

(AUF PROZENTWERTE ACHTEN; MEDIAN AB 50%, ZWEITES QUARTIL AB 25%, ERSTES DEZIL AB 10%...)

SPANNWEITE, INTERQUARTILSABSTAND





BOXPLOTS

- Dienen zur Visualisierung und schnellen Beurteilung der Verteilung und Streuung von Variablen
- Auch Gruppenvergleiche bzw das Gegenüberstellen verschiedener Verteilungen sind damit sehr plakativ möglich

- Extremwerte (größer als 3x die Box vom dritten bzw. ersten Quartil entfernt
- Ausreißer größer/gleich 1,5x die Box vom dritten bzw. ersten Quartil entfernt
- Höchster Wert (=Maximum, ohne Ausreißer)
- Drittes Quartil (75%)
- Mittelwert (50%)
- Erstes Quartil (25%)
- Niedrigster Wert (Minimum, ohne Ausreißer)

STREUUNGSMAßE VARIANZ UND STANDARDABWEICHUNG

= geben die Stärke der Streuung der Merkmalsausprägungen um den Mittelwert (Variabilität) an.

Die gebräuchlichsten Streuungsmaße sind:

- **Varianz** Summe aller **quadrierten Abweichungen** der einzelnen Messwerte metrischer Variablen **vom Mittelwert**, dividiert durch die Anzahl aller Messwerte
  → durchschnittliche Abweichung der quadrierten Differenzen vom arithmetischen Mittel

- **Standardabweichung** = **Wurzel der Varianz** und anschaulicher interpretierbar
  → durchschnittliche Abweichung der Merkmalsausprägungen vom arithmetischen Mittel

MITTELWERT, VARIANZ UND STANDARDABWEICHUNG

- VARIANZ: = durchschnittliche quadrierte Abweichung aller Messwerte vom Mittelwert

  @ Grundgesamtheit: $\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - \mu)^2$ @ Stichprobe: $s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$
- STANDARDABWEICHUNG: = Wurzel der Varianz

  @ Grundgesamtheit: $\sigma = \sqrt{\sigma^2}$ @ Stichprobe: $s = \sqrt{s^2}$

→ WARUM dividiert durch n-1 bei Stichprobenvarianz?

- Bei Stichprobe kann nicht die tatsächliche Varianz aller Werte (Varianz der Grundgesamtheit) ermittelt werden, sondern nur eine Teilmenge daraus (Stichprobe). Diese Stichprobe hätte auch anders ausfallen können und eine andere Varianz ergeben können.
- Ein Mittelwert ist dermaßen definiert, dass die Summe der Abweichungen aller Messwerte von ihm Null ergibt. Theoretisch könnten damit alle Messwerte bis auf einen verändert werden um die Summe der Abweichung aller Messwerte auf Null zu behalten. Somit hat die Varianz einer Stichprobe n-1 Freiheitsgrade.

METHODENTRANSPARENZ, FORSCHUNGSETHIK UND DATENSCHUTZ

Jede seriöse Erhebung muss klar und transparent offenlegen,

- Worüber/über wen genau sie etwas aussagen will,
  - Grundgesamtheit, Zielgruppe

- Ob und wenn, WARUM sie Anspruch auf Repräsentativität und Validität erhebt,
    - Woran (an welchen Merkmalen/Kriterien) wird Repräsentativität festgemacht?
    - Wie und wann wurde von wem und warum für wen was genau gemessen?
- Wie die Ergebnisse zu interpretieren sind!
    - Sind statistische Unschärfen zu beachten? Wenn ja: Größe der Schwankungsbreiten!

Darüber hinaus muss empirische Sozialforschung
- Objektiv, freiwillig, anonym und vertraulich erfolgen
- Und ethische Normen berücksichtigen

Außerdem unterliegen personenbezogene Adressierungen und Datenanalysen
- Den Telekommunikations- und
- Datenschutzgesetzen

SPSS: DIE FENSTER

Daten Editor: - Anzeige des Inhalts der Datendatei
- neue Daten(dateien) erstellen, bestehende bearbeiten
- Dateien enden mit *.sav

Viewer: - Anzeige aller statistischen Ergebnisse, Tabellen & Diagramme
- bearbeitbar und speicherbar
- öffnet sich automatisch
- Dateien enden mit *.spv

Syntax Editor: - Auswertungen automatisiert und sekundenschnell ablaufen lassen
- Auswertungen über Befehlstext anstelle Mausklicks
- Syntax ist speicher- und bearbeitbar (text editor)
- Dateien enden mit *.sps

VARIABLEN UND/ODER WERTEBESCHRIFTUNGEN

- Variablenbeschriftungen benennen und charakterisieren Variablen

- Wertebeschriftungen der Variablenausprägungen benennen die Codierungen der Variable

FEHLENDE WERTE:

Was sind/woher kommen fehlende Werte?
- Vergessen zu antworten
- Weiß nicht
- Antwort verweigert
- Nicht zutreffend (Filterfrage)
- Keine Daten vorhanden

Fehlende Werte haben sehr große Relevanz bei
- Prozentrechnungen:
  Auf welcher Basis wird prozentuiert? - % aller Fälle oder nur % vorhandener Werte?

- Ermittlung von Mittelwerten:
  Werden Null-Werte mitgerechnet oder als fehlend betrachtet?

→ Mittelwerte und Prozentzahlen ändern sich, je nachdem, ob sie MIT oder OHNE Missings interpretiert werden!

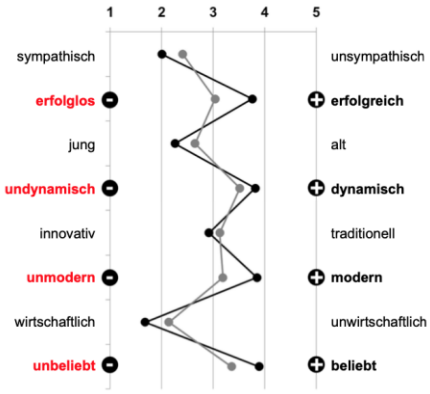
In SPSS werden fehlende Werte als Punkt (.) angegeben und mit Werteausprägung „systemdefiniert fehlend“ interpretiert.

CODIEREN:

- Umcodieren in dieselbe oder neue Variable:
  Ich…. Liebe Schnitzel 1 2 3 4 5 hasse Schnitzel
  Hasse Gemüse 1 2 3 4 5 liebe Gemüse
  → Richtungsbereinigung: 1→ 5 | 2→4 | 3→ 3 | 4→2 | 5→ 1

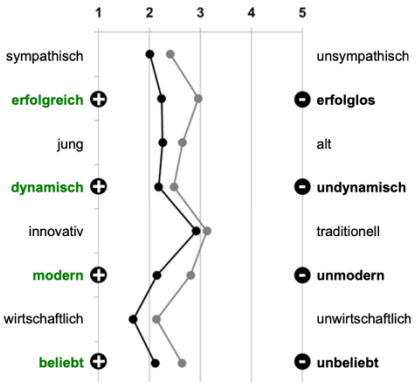
- Berechnen eines Index auf Einzelfallbasis:
  Ich liebe…. Tomaten 1 2 3 4 5
  Paprika 1 2 3 4 5
  Gurken 1 2 3 4 5
  → Gesamtliebe $_{\text{Gemüse}}$ = mean (Note$_{\text{Tomaten}}$, Note$_{\text{Paprika}}$, Note$_{\text{Gurken}}$)

SEMANTISCHES DIFFERENZIAL



Unterschiedliche Polung
→ eher schlecht da unübersichtlich

Idente Polung aller Items
→ Besser



ARTEN DER CODIERUNG:

① multiple Dichotomien

	f_05_txt	f_05_01	f_05_02	f_05_03	f_05_04	f_05_05	f_05_06	f_05_07	f_05_08	f_05_09	f_05_10	f_05_11	f_05_12	f_05_13
12		.	.	.	.	.	.	.	.	.	.	.	.	.
13		.	.	.	.	.	.	.	.	.	.	.	.	.
14		.	.	.	.	.	.	.	.	.	.	.	.	.
15	viele Informationen, die sehr realitätsnah sind;...	0	0	0	0	0	0	0	1	0	0	0	0	0
16	kompakt, einfache Formulierung, Bildmaterial...	1	1	1	1	1	0	0	0	0	0	0	0	0
17	Praxis mit Beispielen	0	0	1	0	0	0	0	0	0	0	0	0	0
18		.	.	.	.	.	.	.	.	.	.	.	.	.
19		.	.	.	.	.	.	.	.	.	.	.	.	.
20		.	.	.	.	.	.	.	.	.	.	.	.	.

② multiple Kategorien

	f_07_txt	f_07_1	f_07_2	f_07_3
81		.	.	.
82		.	.	.
83	kompakt, State of the Art	1 1	2 1	3 0
84	praxisbezogen, langweilig, zu lang	1	5	5
85		.	.	.
86	strukturiert, dick, teuer	4 1	5 3	6 5
87		.	.	.
88		.	.	.
89		.	.	.
90		.	.	.

MEHRFACHANTWORTEN SPSS:
→ Fasst einzelne Variablen, die zu einer Frage mit Mehrfachantworten gehören, in einem Mehrfachantwortset zusammen
→ Jedes Set kann dann über Analysieren – Mehrfachantworten Häufigkeiten/Kreuztabellen ausgewertet werden

GRUNDGESAMTHEIT

Menge aller gleichartigen Objekte (Untersuchungseinheiten, Merkmalsträger), auf die sich eine Erhebung bezieht. (Universum, Population, Kollektiv…)

→ Exakte Definition der Grundgesamtheit ist unumgänglich: Für wen besitzen die Untersuchungsergebnisse Gültigkeit?

→ Definition der Grundgesamtheit ist die Basis aller Forschungen!

Ohne Grundgesamtheit keine Forschung

SCHWANKUNGSBREITE

- Je kleiner die Stichprobe desto größer die Schwankungsbreite
- Je näher das Ergebnis bei 50% desto größer die Schwankungsbreite

Analysen, die auf Zufallsstichproben beruhen, liefern keine 100% sicheren Ergebnisse!

→ Stichprobenergebnisse weichen in der repräsentierten Grundgesamtheit vom Erhebungsergebnis ab.

Mathematische Wahrscheinlichkeitstheorien geben Auskunft,

- Mit welcher Fehlerspanne
- Wie wahrscheinlich

das Stichprobenergebnis auf die Grundgesamtheit übertragen werden kann.

Das Ausmaß der. Fehlerspanne hängt ab

- Von der Stichprobengröße
- Vom ermittelten Prozentwert

Ergebnis einer Erhebung für eine Grundgesamtheit = Verteilung in der Stichprobe plusminus Schwankungsbreite

ERGEBNISSE WIEDERHOLTER ZUFALLSSTICHPROBEN SIND NORMALVERTEILT

- Werden wiederholt Stichproben genügend großen Umfangs aus einer Grundgesamtheit gezogen, verteilen sich die Ergebnisse der Stichproben normal. Unabhängig davon, wie sich das Merkmal selbst in der Grundgesamtheit verteilt.

IRRTUMSWAHRSCHEINLICHKEIT UND KONFIDENZNIVEAU

- In der empirischen Sozialforschung wird es toleriert sich zu irren, als maximale Grenze 5% Fehlerrisiko (= Irrtumswahrscheinlichkeit = α)
- Das Konfidenzniveau ist die Gegenmenge dazu → 1 - α - also 95%

Z-WERT

- Die Größe des Konfidenzintervalls bestimmt die Größe der Fläche unter der Normalverteilungskurve und damit den z-Wert (=Standardabweichung) in der Verteilungstabelle.

SCHWANKUNGSBREITE BERECHNEN:

$$\sigma = \sqrt{\frac{p \bullet (100 - p)}{n}}$$

p = ermittelter Wert in %
n = Stichprobengröße
σ = Standardabweichung

Einfache Standardabweichung:

→ Das Stichprobenergebnis gilt aber nur mit 68,3% Wahrscheinlichkeit für die Grundgesamtheit

Doppelte Standardabweichung (2*σ):

→ Stichprobenergebnis gilt mit 95,5% Wahrscheinlichkeit auch für die Grundgesamtheit

Auf dieser Berechnung beruhen alle Schwankungsbreiten von Prozentergebnissen bei ZUFALLSstichproben. Für Quotenstichproben ist diese Berechnung eigentlich nicht zulässig!

- BSP:
  In einer Zufallsstichprobe mit 1000 Fällen besitzen 10% der Befragten ein Produkt einer bestimmten Marke.

Die Schwankungsbreite berechnet sich… $\sigma = \sqrt{\frac{10 \bullet (100 - 10)}{1.000}} = \sqrt{\frac{900}{1.000}} = \pm 0{,}95\%$

Aber diese Unschärfe gilt nur mit einer Wahrscheinlichkeit von 68,3%.
→ Deshalb: doppelte Standardabweichung **(2*σ):**
2* 0,95% = +- 1,9%

Ergebnis:
In der Grundgesamtheit wird das Stichprobenergebnis zwischen 10% plusminus 1,9% liegen, also zw 8,1% und 11,9%

KONFIDENZINTERVALL VON MITTELWERTEN

- Nur dort wo Fehlerspannen einander nicht überlappen, bestehen tatsächliche Unterschiede

STANDARDFEHLER UND KONFIDENZINTERVALL DES MITTELWERTS

- Standardfehler des Mittelwerts:
  = Standardabweichung der Mittelwerte von gleich großen Zufallsstichproben einer Population. Wird bei zunehmender Stichprobengröße kleiner.

  $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ Etwa 60% aller Stichprobenergebnisse liegen im Bereich von plusminus 1 Standardfehler um den wahren Wert in der Grundgesamtheit

- Konfidenzintervall des Mittelwerts bei 95,5%:
  $s_{\bar{x}} \bullet 2{,}0$ 95,5% aller Stichprobenergebnisse liegen im Bereich von plusminus 2 Standardfehlern um den wahren Wert.

- Konfidenzintervall des Mittelwerts bei 95%
  $s_{\bar{x}} \bullet 1{,}96$ Exakt 95% aller Stichprobenergebnisse liegen im Bereich von plusminus 1,96 Standardfehlern um den wahren Wert.

STICHPROBENGRÖßE (OHNE GRUNDGESAMTHEIT)

- Individuell nötige Stichprobengröße ist aus den Standardtabellen für Schwankungsbreiten ablesbar

ENDLICHKEITSFAKTOR

Gibt es einen Zusammenhang zwischen Größe der Grundgesamtheit und Größe der Stichprobe?

→ Bei kleinem N und demgegenüber großen n muss σ mit dem Endlichkeitsfaktor multipliziert werden.

$$\sigma = \sqrt{\frac{p \bullet (100 - p)}{n}} \bullet \sqrt{\frac{N - n}{N - 1}}$$

→ Diese Berechnung ist vernachlässigbar, wenn die Stichprobe weniger als rund 1% der Grundgesamtheit ausmacht. Der Endlichkeitsfaktor nähert sich dann immer mehr an 1 an.

SCHWANKUNGSBREITE → MINDESTSTICHPROBENGRÖßE

- Stichprobengröße ist abhängig von max. tolerierter Schwankungsbreite
- Stichprobengröße ≠ Repräsentativität

MINDEST-STICHPROBENGRÖßE

- Die Mindestgröße einer Stichprobe sollte 100 bis 200 Fälle umfassen!
  Optimal sind aber 400 Fälle

- Die kleinsten Teile einer Stichprobe, über die noch Aussagen getroffen werden, sollten 100 Fälle umfassen
- Nur im Notfall sollten Teilgruppen ab 30-50 Fällen interpretiert werden

Aufgrund dieser notwendigen Subsamplegröße müssen proportionale Stichproben oft sehr gro0ß oder disproportional angelegt werden.

Brutto ≠ Nettostichprobe
→ Stichprobe größer anlegen als nötig
Stichproben repräsentieren die Grundgesamtheit

HYPOTHESEN
= Wahrscheinlichkeitsaussagen

- Variablenbeziehungen sind
  - Nicht deterministisch (100% Zusammenhang: NaWi)
  - Sondern probabilistisch (Wahrscheinlichkeit: SoWi)

- Hypothesen lassen sich nicht vollkommen verifizieren
  - Man kann nur vorläufig durch eigene Daten bestätigen/stützen mit einer bestimmten Wahrscheinlichkeit
    → sonst müssten ja alle existierenden Fälle untersucht werden was praktisch unmöglich ist

Werte der Grundgesamtheit sind unbekannt und werden versucht, aus einer Stichprobe zu schätzen

- Nicht deterministisch („Alle älteren essen lieber Schnitzel als alle Jüngeren)
- Sondern probabilistisch (Mehr Ältere … (dargestellt durch den Gruppenmittelwert))

SIGNIFIKANZPRÜFUNG AUF GRUNDGESAMTHEIT SCHLIEßEN

- Signifikanzprüfungen drehen sich um folgende Fragen:
  - Besteht ein Ergebnisunterschied zwischen einzelnen Merkmalen?
  - Bewirkt ein Merkmal einen Ergebnisunterschied bei einem anderen?
  - Sind diese Unterschiede (=Zusammenhänge) signifikant?

  Signifikante Ergebnisse:
  - In einer (Zufalls!-) Stichprobe ermittelte Ergebnisunterschiede
  - Sind auf die Grundgesamtheit real übertragbar
  - Treten nicht nur in einer Stichprobe auf sondern kommen wahrscheinlich in fast allen Stichproben, die zufällig aus der Grundgesamtheit gezogen werden können, so oder ähnlich vor.

NULL UND ALTERNATIVHYPOTHESE

- Nullhypothese H0:
  1) Zwischen zwei Werten in einer Grundgesamtheit besteht real KEINE Differenz
  2) In einer ihr entstammenden Stichprobe gefundene Ergebnisse (Ergebnisunterschiede) sind bloß Zufall
  3) Zwischen unabhängiger und abhängiger Variable besteht kein systematischer Zusammenhang
     Unabhängige Variable hat keinen Einfluss auf die abhängige.

- Alternativhypothese H1:
  1) Zwischen zwei Werten in einer Grundgesamtheit besteht real EINE Differenz.
  2) In einer ihr entstammenden Stichprobe gefundene Ergebnisse (Ergebnisunterschiede) sind KEIN Zufall
  3) Zwischen unabhängiger und abhängiger Variable besteht EIN systematischer Zusammenhang
     Unabhängige Variable hat EINEN Einfluss auf die abhängige.

FORSCHUNGSFRAGEN, HYPOTHESEN, SIGNIFIKANZ IN DER PRAXIS

- INHALTLICHE Forschungsfrage:
  Welchen Zusammenhang gibt es zwischen Alter und Affinität zu Schnitzeln?

- INHALTLICHE Hypothese:
  - Ungerichtet: Wenn Personen jünger sind (bis 40), dann unterscheiden sie sich im durchschnittlichen Schnitzelkonsum von älteren
  - Gerichtet: Je älter Personen sind, desto öfter essen sie Schnitzel.

- STATISTISCHE Nullypothese:
  Alter und Schnitzelkonsumationsmenge weisen keinen Zusammenhang auf. Junge und ältere Personen haben idente (ähnliche) Schnitzelkonsumationsmittelwerte.

- STATISTISCHE Alternativhypothese H1:
  Ungerichtet: Alter und Schnitzelkonsumationsmenge weisen EINEN Zusammenhang auf. Junge und ältere Personen haben unterschiedliche Schnitzelkonsumationsmittelwerte.

- SIGNIFIKANZ: wenn Wahrscheinlichkeit für das Ergebnis <5%
  Die Nullhypothese wird verworfen, die Alternativhypothese angenommen:
  Es besteht ein signifikanter Unterschied zwischen jüngeren und älteren Personen in Bezug auf Affinität zu Schnitzeln.

WAS BEDEUTET SIGNIFIKANT:

- **Signifikanz**:
  In einem ZUFALLSsample ermittelte Ergebnisse sind **nicht zufällig**, sondern auf die dahinterstehende Grundgesamtheit WIRKLICH übertragbar.

  - Gefundene **Unterschiede** zwischen Teilgruppen (oder Werten) sind **groß genug**, um sie **in** (fast) **jeder Zufallsstichprobe** aus dieser Grundgesamtheit **zu vermuten**.
  - Sie treten **nicht nur in EINER** oder wenigen **Stichproben** zufällig auf:
    Sie kommen wahrscheinlich in (fast) ALLEN Stichproben (die immer wieder gezogen werden könnten) in dieser oder ähnlicher Form vor.

- **Signifikanztest** (in SOWI am Signifikanzniveau von 5%)
  Wie wahrscheinlich ist ein in der Stichprobe gefundenes (oder extremes) Ergebnis, wenn für die Grundgesamtheit Nullhypothese gilt?
  → Wahrscheinlichkeiten unter 5% bedeuten signifikant

P-WERT

P Wert drückt die Wahrscheinlichkeit aus, dass
- Unter der Annahme, in der Grundgesamtheit gilt die Nullhypothese,
- Das Ergebnis der Stichprobe
- Mindestens den analysierten Wert oder einen größeren Wert (mindestens den Wert der Teststatistik oder einen größeren Wert der Teststatistik) annimmt.

→ Der p-Wert ist eigentlich Irrtumswahrscheinlichkeit, mit der die Nullhypothese gerade noch abgelehnt werden kann.

SIGNIFIKANZPRÜFUNG, SIGNIFIKANZNIVEAU

- Signifikanzprüfung:
  - Wie groß ist die Wahrscheinlichkeit in % (=p-Wert),
    - Für ein Stichprobenergebnis
    - Wenn in der Grundgesamtheit die Nullhypothese (kein Unterschied, kein Zusammenhang) gilt?

- Signifikanzniveau:
  - Üblich in den Sozialwissenschaften ist ein Schwellwert von 5%
    - Forschende müssen diese Schwelle vorab selbst festlegen

SIGNIFIKANZ ODER NICHT

- **Signifikanz**: Wenn ein Ergebnis – bei Gültigkeit der Nullhypothese – eine **Wahrscheinlichkeit** (p-Wert) von **unter 5%** besitzt, gilt es als **signifikant**.

  Diese Verhältnisse in der Grundgesamtheit sind unbekannt. Die Erhebung versucht, sie vorherzusagen. Würde tatsächlich, real, Wertegleichheit bestehen (=H0), wäre ein Ergebnis wie das gefundene (extrem) unwahrscheinlich (unter 5%).
  Ein **signifikantes Ergebnis** ist also eines, das sich **mit** der **Nullhypothese nicht vereinbaren** lässt.
  → Die **Nullhypothese** wird **verworfen**, der **Alternativhypothese Vorzug** eingeräumt.

- **Keine Signifikanz:** Wenn ein **Ergebnis** – bei Gültigkeit der Nullhypothese – eine **Wahrscheinlichkeit** von **5% oder mehr** besitzt, gilt es als **NICHT signifikant**.

  Die Werte liegen dann nahe beieinander, ein **Ergebnis** wie das gefundene tritt – **unter H0 Bedingungen** – wahrscheinlich viel öfter auf (**5% oder öfter**).
  → **Die Nullhypothese wird angenommen**.

WIE STARK SIGNIFIKANT IST EIN ERGEBNIS?

- **Nullhypothese** wird **verworfen**, wenn der **Prozentwert** der Prüfung **KLEINER als** das **Signifikanzniveau** ist.
  „H0 wird verworfen, wenn p-Wert < α ist."

- Je **kleiner** der **p-Wert** ist, desto stärker sprechen die Daten **GEGEN die Nullhypothese** und **FÜR die Alternativhypothese**

FEHLER ERSTER α UND WEITER ART β

Wenn H0 zurückgewiesen wird, kann das fälschlich geschehen. Dabei können 2 Fehlertypen auftreten:

- Fehler erster Art (α-Fehler)
  H0 wird **verworfen**, **obwohl** sie **zutrifft**. Jemand **glaubt**, etwas **WÄRE signifikant**, obwohl **e**s das **nicht ist.**

- Fehler zweiter Art (β-Fehler)
  H0 wird nicht verworfen, obwohl sie nicht zutrifft.
  Jemand **glaubt**, etwas wäre **NICHT signifikant**, obwohl **es** das **ist**.

SIGNIFIKANZ – STICHPROBENGRÖßE – UNTERSCHIEDSSTÄRKE

Signifikanz hängt auch mit Samplegröße, Unterschiedsstärke und allgemeinen Rahmenbedingungen zusammen.

- Bei großen Stichproben:
  Auch bereits **kleine Unterschiede zwischen** zu vergleichenden **Gruppen** (=schwache Zusammenhänge) können **signifikant** sein. Solche **Unterschiede** sind dann **inhaltlich kaum relevant**.

- Bei kleinen Stichproben:
  Unterschiede müssen ziemlich groß sein, um signifikant zu werden.

- Auch Stärke von Ergebnisunterschieden und das Rundherum (gesellschaftlicher Wandel, Setting, Fragestellung, Erhebungsinstrument, Stichprobengröße, Auswertung…) haben oft Einfluss.

→ Fazit: Signifikanz ist nicht Alles
- Schwellenwert von 5% sollte nicht zum Erfolgsschalter werden
- Besser: die Nullhypothese wird vorläufig angenommen, danach wird noch umfassend inhaltlich interpretiert.
- Signifikanz bedeutet nicht Kausalität!

EINSEITIG UND ZWEISEITIG SIGNIFIKANTE UNTERSCHIEDE

EINSEITIGE (GERICHTETE) UND ZWEISEITIGE TESTPROBLEME

- BSP:
  Befragt werden Personen, die bis 40 Jahre alt sind, und solche, die älter sind. Beide Gruppen werden befragt, ob sie gerne Schnitzel essen.

  Zweiseitiges Testproblem:
  H0: Der Anteil der Personen, die **gerne Schnitzel** essen, ist unter **Jüngeren (bis** 40) und **Älteren gleich groß.**
  H1: Der **Anteil** ist ein **anderer**.

  Linksseitiges Testproblem:
  H0: Der **Anteil** unter **Jüngeren (bis 40**) ist **gleich** oder **größer** als bei Älteren.
  H1: Der **Anteil** unter **jüngeren** ist **kleiner**.

  Rechtsseitiges Testproblem:
  H0: Der **Anteil** unter **Jüngeren (bis 40**) ist **gleich** oder **kleiner** als bei Älteren.
  H1: Der **Anteil** ist unter **Jüngeren größer**.

SIGNIFIKANZ EIN UND ZWEISEITIG PRÜFEN

α = Irrtumswahrscheinlichkeit = Signifikanzniveau = 5% (= Risiko für Fehlentscheidung)
= Konfidenzniveau = 95%)

p-Wert = Wahrscheinlichkeit für dieses Ergebnis, wenn H0 gilt

SPSS berechnet zweiseitigen p-Wert

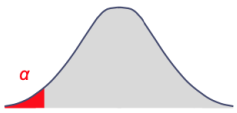
Ein zweiseitiger p-Wert muss für einseitige Tests halbiert werden.

zweiseitiger Hypothesentest

signifikant, wenn: $p\text{-}Wert < \alpha$

[Eigentlich, wenn: $2 \cdot \frac{p\text{-}Wert}{2} < 2 \cdot \frac{\alpha}{2}$]

einseitiger Hypothesentest



signifikant, wenn: $\frac{p\text{-}Wert}{2} <$

signifikant, wenn: $\frac{p\text{-}Wert}{2} < \alpha$

► **Entscheidung für einseitig – ein Muss im VORFELD!**

PRO: ist genauer und „schneller" signifikant

CONTRA: „gefährlich", wenn das tatsächliche Ergebnis eine zur H_1 „gegenteilige" Differenz aufweist, muss H_0 trotz Unterschieds akzeptiert werden!

HÖHERE TESTSTÄRKE BEI EINSEITIGEN TESTS

Der p-Wert muss einseitig zwar nur halb so groß sein, die Wahrscheinlichkeit für den ß-Fehler ist aber bei identer Datenlage kleiner.
Teststärke (power) = Wahrscheinlichkeit, ß-Fehler zu vermeiden:
Zweiseitig = stärker irrtumsbehaftet, die HA fälschlich abzulehnen → größerer ß-Fehler → weniger starke Power

Z-WERT

- **Größe** des **Konfidenzintervalls** bestimmt die **Größe** der **Fläche** innerhalb der **Normalverteilung** und damit den **z-Wert (=Standardabweichung)** in der Verteilungstabelle.

$$p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \bullet (100 - p)}{n}}$$

TESTSTATISTIKEN

Werden benötigt

- Für **Hypothesentests** in der **schließenden Statistik** (interferenzenstatistik, induktive statistik)
- Teststatistiken **beschreiben** die **Verteilung** einer **Variablen**, die **für** einen **Signifikanztest konstruiert** wurde
- **Per Zufall** wurden bereits **sehr viele Ausprägungen** dieser **Testvariablen erzeugt**:
    - Jeder dieser zufälligen Ausprägungen tritt mit einer bestimmten Wahrscheinlichkeit auf

  Eine Teststatistik einer Testvariable ist eine Übersichtstabelle über die Wahrscheinlichkeit des Eintretens ihrer Ausprägungen

FREIHEITSGRADE

= voneinander unabhängige Beobachtungen/Anzahl frei wählbarer Elemente in einer Berechnung
BSP:

- Aus 3 Zahlen wird ein Mittelwert berechnet
  Soll er immer gleich bleiben, wären zwei der drei Zahlen frei veränderbar. Die dritte Zahl würde sich aus den beiden veränderten ergeben (sind von diesen also vorgegeben) um den identen Mittelwert zu erhalten.

  2 | 3 | 4 → Mittelwert = 3

  Wenn der Mittelwert 3 bleiben soll, können max 2 Zahlen verändert werden, zb 7 | 1 → Die dritte Zahl MUSS dann 1 sein, um wieder auf einen Mittelwert von 3 zu kommen.

FREIHEITSGRADE UND SIGNIFIKANZPRÜFUNG

Freiheitsgrade dienen zur **Identifizierung** der **kritischen Wertgrenzen** in Teststatistiken.
Bei **Signifikanztests** wie zb **Chi2** oder T-Test werden

- Neben dem angestrebten **Konfidenzintervall**
- Auch die **Freiheitsgrade**

Des auf Signifikanz zu prüfenden Ergebnisses benötigt

► Wie viele Werte des vorliegenden Ergebnisses (%, Ø) einer Zufallsstichprobe könnten durch den Zufall (in Form theoretisch möglicher anderer Stichprobenelemente) maximal variiert werden, ohne dass sich das vorliegende Ergebnis (= das aus der Stichprobe geschätzte Ergebnis der Grundgesamtheit) verändert?
Welche rechnerische „Flexibilität“ ist maximal möglich?

FREIHEITSGRADE INTERPRETATION

In der Praxis werden Freiheitsgrade dazu benötigt, die Variabilität (und damit die Güte) eines Ergebnisses zu beurteilen.

→ Bei kategorialen Kreuztabellen sind viele df schlecht:
Hier referenzieren die df auf die Anzahl der Zellen in der Kreuztabelle

- Je mehr Zellen (df) eine Kreuztabelle enthält, desto schwieriger wird die Interpretation des Variablenzusammenhangs. Welche Ausprägungen hängen wie zusammen?

→ Bei metrischen Mittelwertsvergleichen (t-Test) sind viele df gut:
Hier referenzieren die df auf die Anzahl der Messungen, aus denen sich die zu vergleichenden Mittelwerte zusammensetzen.

- Je mehr df einem Ergebnis zugrunde liegen, desto hochwertiger ist es zu betrachten.

STATISTISCHE PARAMETER FÜR GRUNDGESAMTHEIT UND STICHPROBE

Symbol	Grundgesamtheit (Parameter)	Stichprobe (Schätzung der Parameter)
arithmetisches Mittel	μ	$\bar{x}$
Median	$\tilde{\mu}$	$\tilde{X}$
Standardabweichung	σ	s
Varianz	σ^2	s^2
Fallzahl	N	n
Proportion	π	p

KONTINGENZTAFEL (KREUZTABELLE)

Wie viele Beobachtungseinheiten weisen für jede mögliche Kombination beider Merkmale die jeweilige Kombination auf?

→ gebräuchlichste Form der Zusammenhangsdarstellung zwischen zwei nominal- oder ordinal skalierten Variablen

- Ein Merkmal wird den Zeilen i, das andere den Spalten j zugeordnet
- Eine Kontingenztafel mit k Zeilen und m Spalten = eine „k*m-Kontingenztafel“

	Fastfood-Konsum			
Stress	**wöchentlich**	**monatlich**	**seltener**	**Summe**
nein	5	5	14	**24**
ja	47	17	9	**73**
Summe	**52**	**22**	**23**	**97**

- Gesamt-Tabellensumme untere rechte Ecke
- Summe der Zeilensummen muss Gesamtsumme ergeben
- Summe der Spaltensummen muss Gesamtsumme ergeben
- Auch Zeilen und/oder Spaltenprozent sind ausgewiesen

KONTINGENZTAFEL BEOBACHTETE UND ERWARTETE WERTE

E_{ij} = theoretisch zu erwartende Werte, wenn die Merkmale unabhängig sind

$$E_{ij} = \frac{\text{Zeilensumme}}{n} \cdot \text{Spaltensumme}$$

→ Der Zusammenhang ist umso größer, je mehr der erwartete Wert von dem beobachteten Wert abweicht.

ABHÄNGIGE UNABHÄNGIGE VARIABLE

Die **unabhängige Variable (UV)** ist die **Ursache** die **Abhängige (AV)** die **Wirkung**.

- Anhand der **UV** wird eine **Grundgesamtheit** oder **Stichprobe** in **Gruppen** aufgeteilt.
- Innerhalb der **AV** wird die **Verteilung** dieser **Gruppen untersucht**.
- Hängt der Schnitzelkonsum vom Alter ab oder hängt das Alter vom Schnitzelkonsum ab?

UV und AV können wechselseitig sein:
Zb: Besuchshäufigkeit unabhängig, Kaufmenge abhängig:
→ Kaufen OFT-Besucher einer Firma mehr als Wenig-Käufer?

KONTINGENZTAFEL – QUADRATISCHE KONTINGENZ (CHI^2)

$$\chi^2 = \sum_{\text{alle Felder}} \frac{(\text{beobachtete Häufigkeit} - \text{erwartete Häufigkeit})^2}{\text{erwartete Häufigkeit}}$$

Da B_{ij}-E_{ij} auch negativ sein kann, wird die Differenz quadriert (=normiert).

→ Chi^2 allein ist schwer bewertbar, da es von 0 bis unendlich große Werte annehmen kann.
→ Es bildet aber die Basis für die Berechnung von Zusammenhangsmaßen (zb Cramers V) oder Hypothesenprüfungen.

ZUSAMMENHANGSMAß CRAMERS V

- Cramers V ist ein Maß für die Stärke des Zusammenhangs zwischen nominalen Variablen
- Es ist unerheblich, wie viele Zeilen und Spalten die Kontingenztabelle hat
- Cramers V nimmt Werte zwischen 0 und 1 ein:
    - 0 – 0,2 = schwacher Zusammenhang
    - 0,2 – 0,6 = mittlerer Zusammenhang
    - Ab 0,6 starker Zusammenhang

$$V = \sqrt{\frac{\chi^2}{n \cdot (m-1)}}$$

m = Anzahl der Zeilen oder Anzahl der Spalten, je nachdem, welcher Wert der beiden der kleinere ist (im Beispiel mit 2 Zeilen und 3 Spalten = i = 2)

X^2 = Chi^2 Test Ergebnis
n = Gesamtanzahl (In Kreuztabelle ganz unten rechts)

CHI^2 VERTEILUNG

Für die Chi2 Verteilung gibt es eine SEHR große Tabelle für jede mögliche Anzahl an Freiheitsgraden. Am häufigsten verwendet wird das 95% Quantil, da dieses die kritische Schranke für einen Chi2 Test mit Signifikanzniveau 0,05 darstellt.

KREUZTABELLEN – NULL UND ALTERNATIVHYPOTHESE

Bestehen Gruppenunterschiede?
Haben Personen, die gerne lesen, im letzten Jahr zu einem anderen Anteil ein Fachbuch gelesen als Leute, die NICHT gerne lesen?

- Nullhypothese:
  Es gibt KEINEN wirklichen Unterschied, das sind Zufallsschwankungen.

➔ Menschen die gerne lesen, und Menschen, die NICHT gerne lesen, haben zu IDENTEN (ÄHNLICHEN) Anteilen ein Fachbuch gelesen.

- Alternativhypothese:
  Es gibt EINEN wirklichen Unterschied, das sind KEINE Zufallsergebnisse.

➔ Menschen, die gerne lesen, haben zu einem ANDEREN ANTEIL ein Fachbuch gelesen als Menschen, die NICHT gerne lesen.

- Signifikanzprüfung:

➔ Wie groß ist die Wahrscheinlichkeit des vorhandenen Stichprobenergebnisses, wenn in der Grundgesamtheit die Nullhypothese gilt?

KREUZTABELLE: STARKER VARIABLENZUSAMMENHANG

Zur **Analyse** von **Zusammenhängen** zwischen **nominalen** bzw. **ordinalen Variablen** mit **überschaubarer Anzahl** an **Ausprägungen**

Ohne % sind Zusammenhänge kaum beurteilbar.

KREUZTABELLEN: CHI2 TEST

Statistische Prüfung auf die **Signifikanz nominaler** oder **ordinaler** Variablenzusammenhänge.

➔ Das **Ergebnis** ist ein Prozentwert (**p-Wert**), der die **Wahrscheinlichkeit** des Ergebnisses unter der Annahme „**Nullhypothese gilt**“ ausdrückt.

CHI2 TEST BESONDERHEITEN

- Chi2 Test ist nur sinnvoll interpretierbar, wenn
    - Die zu vergleichenden Stichproben UNABHÄNGIG voneinander sind (andere Personen, keine Messwiederholung)
    - Alle erwarteten Häufigkeiten > 0 sind

- Nicht mehr als 20% der erwarteten Häufigkeiten <5 sind (bei 2x2 Tabellen sollte keine erwartete Zellhäufigkeit <5 sein)
- Wenn erwartete Häufigkeiten = 0 oder mehr als 20% <5:
  - Bei 2x2 Tabellen: Fishers Exact Test für 2x2 Tabellen
  - Bei größeren Tabellen:
    - Ausprägungen zusammenfassen oder
    - Mehr Daten erheben (größere Stichprobe)

➔ Sonst ist CHI2 NICHT INTERPRETIERBAR

Der Chi2 sagt nichts über die Stärke des Effekts aus. Dazu dient Cramers V oder Kontingenzkoeffizienten C.

ABHÄNGIGE UND UNABHÄNGIGE STICHPROBEN(TEILE)

➔ Zwei Stichproben sind dann voneinander **abhängig** (gepaart, verbunden)
- Wenn **jedem Wert** der Stichprobe eindeutig
- Genau **ein Wert** der **anderen Stichprobe zugeordnet** werden kann.

Abhängig sind:
- **Wiederholungsmessungen** bei **denselben Person**en
- **Vergleich** mehrerer **Variablen DERSELBEN Datensätze**

➔ Zwei Stichproben sind voneinander **unabhängig**
- Wenn **einem Wert** der **anderen Stichprobe**
- **KEIN WERT** der anderen **zugeteilt** werden kann

Unabhängig sind:
- **Stichprobenteile unterschiedlicher Grundgesamtheiten**
- **Vergleich einzelner Variablen** jeweils **ANDERER Datensätze**

VERFAHREN FÜR MITTELWERTSVERGLEICHE – ZWEI ARTEN

Mittelwertsdifferenz zwischen zwei oder mehr Gruppen signifikant und damit real (auch in der Grundgesamtheit vorhanden) oder bloß mit zufälligen Schwankungen zu erklären?

→ Zur Signifikanzprüfung gibt es zwei Arten von Verfahren:

- Parametervrfahren
  - Sind besser als parameterfreie Verfahren (rechnen höherwertiger)
  - Basieren auf den Parametern Mittelwert und Varianz
  - Sind nur unter gewissen Bedingungen anwendbar

- Parameterfreie Verfahren
  - Rechnen auf einem niedrigeren Niveau als Parameterverfahren
  - Sind verteilungsunabhängig
  - Sind IMMER anwendbar

➔ Voraussetzungen:
- Für Parameterverfahren:
  1) Mindestens (Quasi)Intervallskalierung (Ratingskalierung)
  2) Normalverteilung der Werte (je Gruppe, die verglichen werden soll, JE GRUPPE eigener Test
  3) Homogenität der Varianzen (Prüfung mittels Levene Test, wird bei Parameterverfahren mitgeliefert)

  → Wenn 1. Oder 2. Voraussetzung nicht zutrifft: parameterfreies Verfahren

- Für parameterfreies Verfahren
  1) Mindestens Ordinalskalierung
  2) Keine weiteren Voraussetzungen

  → sind IMMER anwendbar, auch wenn die Voraussetzungen für Parameterverfahren erfüllt wären!

Parameterverfahren

[1]	**Stichproben** (bzw. Variablen)	**ABhängig** = gepaart = verbunden	**zwei**	**T-Test für ABhängige Stichproben**	[1]
[2]			**drei** (und mehr)	**Varianzanalyse mit Messwiederholung**	[2]
[3]	**Stichproben** (bzw. Variablen	**UNabhängig**	**zwei**	**T-Test für UNabhängige Stichproben**	[3]
[4]			**drei** (und mehr)	**Varianzanalyse (ANOVA)**	[4]

parameterFREIE Verfahren

[5]	***Stichproben*** *(bzw. Variablen)*	***ABhängig*** *= gepaart = verbunden*	***zwei***	***Wilcoxon-Test***	*[5]*
[6]			***drei*** *(und mehr)*	***Friedman-Test***	*[6]*
[7]	***Stichproben*** *(bzw. Variablen*	***UNabhängig***	***zwei***	***U-Test***	*[7]*
[8]			***drei*** *(und mehr)*	***Kruskal-Wallis-Test***	*[8]*

F TEST

➔ Zur Überprüfung der Varianzgleichheit von zwei unabhängigen Stichprobenteilen
➔ Kann unter anderem dazu benutzt werden, um die Annahme der Varianzhomogenität bei einem T-Test zu überprüfen.

- **Nullhypothese**: Die Varianzen sind gleich (in beiden Grundgesamtheiten, aus denen Stichproben stammen)
- **Alternativhypothese**: Varianzen sind verschieden

- Berechnung der Prüfgröße: $F = \frac{s_1^2}{s_2^2}$ $df_1 = n_1 - 1$ $df_2 = n_2 - 1$

➔ Zum Vergleich der Prüfgröße mit den kritischen Werten aus F-Tabellen sollte der Ergebniswert der Formel einen Wert größer als 1 ergeben – Stichprobe 1 (im **Zähler**) muss deshalb das **größere s^2** aufweisen.
→ Bei den Varianzen der zwei Stichproben gehört die größere Varianz in den Zähler (oben, =s12) und die kleinere in den Nenner (unten, =s2^2).

	monatliche Fastfood-Häufigkeit										Ø	s^2
Personen ohne Stress	2	2	3	4	1	2	1	3	0		2,00	1,50
Personen mit Stress	4	4	5	6	3	8	9	6	7	8	6,00	4,00 ▶ **größerer Wert!**

$F_{empirisch} = \frac{s_1^2}{s_2^2}$ $df_1 = n_1 - 1$ $df_2 = n_2 - 1$

$F_{empirisch} = \frac{4{,}00}{1{,}50} = 2{,}66$

$df_1 = 10 - 1 = 9$ ▶ **größerer Wert!**
$df_2 = 9 - 1 = 8$
$F_{kritisch} = 3{,}39$

Wenn: $F_{empirisch} < F_{kritisch}$
▶ H_0 bleibt aufrecht
▶ **Die Varianzen SIND homogen!**

Wenn aber: $F_{empirisch} \geq F_{kritisch}$ ▶ ~~H_0~~

F-Verteilung für (1-a)=0,95

df_2 (Nenner)	df_1 (Zähler) 1	2	3	4	5	6	7	8	9	10
1	161,45	199,5	216	225	230	234	237	239	240,54	241,88
2	18,51	19	19	19	19	19	19	19	19,38	19,4
3	10,13	9,55	9,3	9,1	9	8,9	8,9	8,9	8,81	8,79
4	7,71	6,94	6,6	6,4	6,3	6,2	6,1	6	6	5,96
5	6,61	5,79	5,4	5,2	5,1	5	4,9	4,8	4,77	4,74
6	5,99	5,14	4,8	4,5	4,4	4,3	4,2	4,2	4,1	4,06
7	5,59	4,74	4,4	4,1	4	3,9	3,8	3,7	3,68	3,64
8	5,32	4,46	4,1	3,8	3,7	3,6	3,5	3,4	3,39	3,35
9	5,12	4,26	3,9	3,6	3,5	3,4	3,3	3,2	3,18	3,14
10	4,96	4,1	3,7	3,5	3,3	3,2	3,1	3,1	3,02	2,98
11	4,84	3,98	3,6	3,4	3,2	3,1	3	3	2,9	2,85

Quelle: http://eswf.uni-koeln.de/glossar/fvert3.htm [01.09.2018]

→ F-empirisch = erst Varianzen ausrechnen, die größere Varianz durch die kleinere dividieren.
→ F-kritisch = größere n in den Zähler, kleinere n in den Nenner und in F-Verteilungstabelle Wert nachsehen.

➔ Ist der empirische Wert kleiner als der kritische Wert bleibt H0 aufrecht und die VARIANZEN SIND HOMOGEN
➔ Ist der empirische Wert größer/gleich als der kritische verfällt H0.

T-TEST

Zur Überprüfung der Mittelwertsunterschiede von zwei unabhängigen oder abhängigen Stichproben

Voraussetzungen:

- Zumindest QUASI Intervallskalenniveau
- Varianzhomogenität → F-Test (nur bei unabhängigen Vergleichen)
- Normalverteilung
  - Subjektive Annahme Grenzwertsatz, keine Ausreißer in den Differenzwerten bzw. unabhängigen Werten

- Nullhypothese
  Es gibt keine signifikanten Mittelwertsunterschiede (in den Grundgesamthieten, aus denen gezogen wurde)

- Alternativhypothese:
  Es gibt einen signifikanten Mittelwertsunterschied

T-Test für unabhängige Stichproben

Berechnung der Prüfgröße (t-Wert kann auch negativ sein, dann wird sein Betrag verwendet)

$$|t_{empirisch}| = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2}\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad df = n_1 + n_2 - 2$$

	monatliche Fastfood-Häufigkeit										Ø	s^2
Personen ohne Stress	2	2	3	4	1	2	1	3	0		2,00	1,50
Personen mit Stress	4	4	5	6	3	8	9	6	7	8	6,00	4,00

$$|t_{empirisch}| = \frac{2-6}{\sqrt{\frac{(9-1)\cdot 1{,}50+(10-1)\cdot 4{,}00}{9+10-2}\cdot\left(\frac{1}{9}+\frac{1}{10}\right)}} = \frac{-4}{\sqrt{\frac{12+36}{17}\cdot(0{,}11+0{,}1)}}$$

$$= \frac{-4}{\sqrt{2{,}82\cdot 0{,}21}} = \frac{-4}{0{,}77} = |-5{,}19| = 5{,}19$$

$$df = 9 + 10 - 2 = 17$$

$t_{kritisch[einseitig]} = 1{,}74$ ▶ $t_{empirisch} \geq t_{kritisch}$

$t_{kritisch[zweiseitig]} = 2{,}11$ ▶ $t_{empirisch} \geq t_{kritisch}$

▶ H_0 wird verworfen.

▶ Die Mittelwerte unterscheiden sich signifikant!

einseitig α = 5% zweiseitig α = 5%

df	kritische T-Werte (einseitig)		
	0,95	0,975	0,99
1	6,314	12,706	31,821
2	2,92	4,303	6,965
3	2,353	3,182	4,541
4	2,132	2,776	3,747
5	2,015	2,571	3,365
6	1,943	2,447	3,143
7	1,895	2,365	2,998
8	1,86	2,306	2,896
9	1,833	2,262	2,821
10	1,812	2,228	2,764
11	1,796	2,201	2,718
12	1,782	2,179	2,681
13	1,771	2,16	2,65
14	1,761	2,145	2,624
15	1,753	2,131	2,602
16	1,746	2,12	2,583
17	1,74	2,11	2,567
18	1,734	2,101	2,552
19	1,729	2,093	2,539
20	1,725	2,086	2,528

Test bei unabhängigen Stichproben

■ Wenn der Levene-Test ein p > 0,05 ergibt, besteht Varianz-Gleichheit.

		Levene-Test der Varianzgleichheit		t-Test für die Mittelwertgleichheit						
		F	Sig.	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehlerdifferenz	95% Konfidenzintervall der Differenz	
									Unterer Wert	Oberer Wert
f_06_2 optisch ansprechend	Varianzen sind gleich	13,618	❹ ,000	-4,012	131	,000	-,789	,197	-1,178	-,400
	Varianzen sind nicht gleich			-4,261	118,186	❺ ,000	-,789	,185	-1,155	-,422

▶ Die Mittelwerte der beiden Teilstichproben unterscheiden sich höchst signifikant voneinander.

bbildung aus How to do

STREUDIAGRAMME, KORRELATIONEN

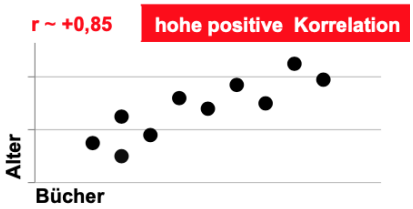
Experiment Münzen:
Euromünzen (je 2mm hoch) werden untershiedlich hoch gestapelt danach alle Stapel der Größe nach geordnet

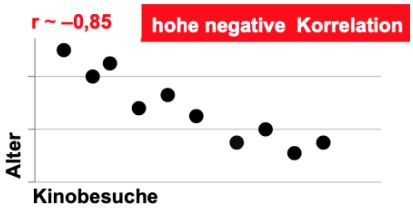
- ➔ Anzahl der Münzen pro Stapel werden abgezählt und Höhe der Stapel in mm bestimmt
- ➔ Da eine Münze 2mm hoch, ist die Höhe in mm immer doppelt so groß wie die Anzahl von Münzen
- ➔ Es existiert exakter linearer Zusammenhang zw. Anzahl der Münzen und Stapelhöhe:
  - Höhe in mm = 2*Anzahl Münzen

ZUSAMMENHANG 2 VARIABLEN X UND Y:

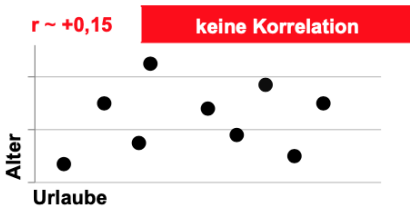
In der Praxis ist ein Variablenzusammenhang von zwei Variablen so gut wie nie linear:
→ Fast immer gibt es Abweichungen der beiden Messwerte von der linearen Gerade

Die Korrelation misst die Stärke der Abweichungen
→ Sie kann stark (nahe 1, wenig Abweichung)
oder schwach (nahe 0, mehr Abweichung),
POSITIV oder NEGATIV sein.





KORRELATIONEN:



- Liefern ein Maß für
  - Den Grad des Zusammenhangs zwischen
  - Zwei (zumindest ordinalen, besser metrischen) Merkmalen (Variablen)

- Für jeden Datensatz existieren dazu zwei Messwerte
- Der Grad des Zusammenhangs ist der Korrelationskoeffizient

Voraussetzungen:

- Ab Ordinalskala → **Spearman Korrelation**
- Ab Intervallskala UND ohne Ausreißer UND wenn bei Signifikanzprüfung Normalverteilung beider Variablen → **Pearson Korrelation**

- Zusammenhänge nominal skalierter Variablen werden beschrieben durch den Kontingenzkoeffizienten und **CRAMERS V**

STICHPROBEN KOVARIANZ:

→ Wie variieren bei bivariater Betrachtung die Ausprägungen zweier Merkmale zusammen?

- Bei positiver Korrelation liegen die meisten Punkte im 3. Und 1. Quadranten
- Bei negativer Korrelation liegen die meisten Punkte im 2. Und 4. Quadranten

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

STICHPROBENKOVARIANZ – GRUNDSÄTZE

- **Gleichen** sich die **positiven** und **negativen Multiplikationsergebnisse** aus, wird die **Stichprobenkovarianz Null**.
- Liegen **alle Beobachtungspunkte** auf den **beiden Achsen**, wird die **Stichprobenkovarianz Null**.
- Je **größer** die **Streuung**, desto **größer** die **Kovarianz** und desto **größer** der **Zusammenhang zwischen** den metrischen **Variablen**.

→ Je **stärker** die **Wertpaare überwiegen**, wo **große x Werte** mit **großen y Werten** einhergehen, desto **größer** die **Kovarianz**
→ Je **stärker Wertpaare überwiegen**, wo **große x Werte** mit **kleinen y Werten** einhergehen, desto **kleiner** die **Kovarianz**

STICHPROBENKOVARIANZ RECHNUNGSBEISPIEL

Ø 50	78	65	60	51	42	34	20	Alter
Ø 60	100	90	65	65	45	30	25	Arzneimittel-ausgaben p. M.

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

$$s_{xy} = \frac{1}{7-1}[(20-50)\cdot(25-60)+(34-50)\cdot(30-60)+(42-50)\cdot(45-60)+\ldots+(78-50)\cdot(100-60)]$$

KOVARIANZ UND KORRELATION

Die Kovarianz gibt wegen des Vorzeichens einen Hinweis auf das gemeinsame Wachstumsverhalten (Miteinander variieren) der beiden Merkmale,

- Sie erlaubt zwar die Aussage „je größer, desto stärker“ aber nicht über die genaue Stärke des Zusammenhangs.

➔ Aus diesem Grund wird eine Normierung so durchgeführt, dass das resultierende Maß – der Korrelationskoeffizient – immer zwischen +1 und -1 liegt.
  - Bei einem Korrelationskoeffizient von 1 (kaum real), liegt ein vollständig linearer Zusammenhang zwischen den beiden Variablen
  - Hohe positive Korrelation:
    - Hohe Werte der einen Variable weisen tendenziell auch
    - Hohe Werte der anderen Variable auf
  - Hohe negative Korrelation
    - Hohe Werte der einen Variable weisen tendenziell
    - Niedrige Werte der anderen Variable auf

→ Berechnungsformel der Korrelation nach Pearson:
(= Produkt Moment Korrelationskoeffizient r)

STREUDIAGRAMM:

→ Zur Visualisierung von Beziehungen metrischer Variablen zueinander

KORRELATION – STÄRKEN:

$$r = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \frac{s_{xy}}{s_x \cdot s_y}$$

positiver Korrelationskoeffizient	Korrelation	negativer Korrelationskoeffizient
bis +0,2	sehr gering	bis -0,2
> +0,2 bis +0,5	gering	< -0,2 bis -0,5
> +0,5 bis +0,7	mittel	< -0,5 bis -0,7
> +0,7 bis +0,9	hoch	< -0,7 bis -0,9
> +0,9	sehr hoch	< -0,9

NORMALVERTEILUNG

P >= 0,05 bedeutet NORMALVERTEILUNG!

Kolmogorov-Smirnov[a]			Shapiro-Wilk		
Statistik	df	Signifikanz	Statistik	df	Signifikanz
,108	505	❽ ,000	,964	505	❾ ,000
,250	505	,000	,798	505	,000

p ≥ 0,05 bedeutet Normalverteilung!

► KEINE der Variablen liegt normal-verteilt vor.

KORRELATION ARTEN

- Pearson Korrelation: Auf Basis von Ausprägungen berechnet
- Spearman Korrelation: Bezug auf Ränge der Ausprägungen

EXTREMWERTE UND DER SCHEIN KÖNNEN KORRELATIONEN FÄLSCHEN

- Der Begriff Zusammenhang bedeutet NICHT, dass zwischen zwei Merkmalen neben dem rechnerischen auch ein KAUSALER Zusammenhang besteht

- Scheinkorrelation durch Inhomogenität:
    - Komplexe Studienrichtungen benötigen längere Zeit…
    - Längere Zeit fürs Studium führt später zu höherem Einkommen…

➔ Studienkomplexität korreliert also mit Einkommen:
Je komplexer das Studium, desto höher das Einkommen.

- Scheinkorrelation zwischen Zeitreihen:
    - Hitzetage in Europa werden mehr…
    - Personen ohne eigenes Auto in Städten werden mehr.

➔ Je weniger Autos, desto wärmer das Klima.

REGRESSIONSANALYSE – ARTEN:

Regressionsanalyse ermittelt Zusammenhang zwischen zwei Variablen und schätzt aufgrund Stichprobe den wahren Zusammenhang in der Grundgesamtheit.

- Grundsätzliche Fragen sind die Art der Variablenbeziehung
    - **Lineare** Regression
    - **Nichtlineare** Regression

- Form der Regression
    - **Einfach**regression (Wie stark wirken Schneefallstunden auf Streuzalsverbrauch?)
    - **Mehrfach**regression (Welche Gefühlsdimensionen wirken wie stark auf die Kaufentscheidung?)

REGRESSIONSANALYSE – CHARAKTERISIERUNG:

Regressionsanalyse unterstellt eine **eindeutige Richtung eines Zusammenhangs** (Ursache, Wirkung) zwischen zwei Variablen
→ Je desto – Beziehung

Beschreibung und **Vorhersage der Wirkung**
- Einer **unabhängigen** Variable **Y**
- Auf **eine** oder **mehrere abhängige Variablen x1**, x2, x3,…,xk

→ Während die Korrelation
- Die STÄRKE des Zusammenhangs zwischen zwei Variablen ermittelt, dient die Regressionsanalyse dazu,
    - Die Art des Zusammenhangs aufzudecken
    - Den Wert der abhängigen Variable aus den Werten der unabhängigen Variable VORHERZUSAGEN

        - Um wieviel verändert sich die abhängige Variable, wenn die unabhängige Variable um einen bestimmten Betrag zu oder abnimmt?
        - Zb: Wie wirkt der Preis auf die Absatzmenge eines Produkts?

KORRELATION UND REGRESSION – BASIS

In der Praxis ist ein Variablenzusammenhang so gut wie nie linear

- Die **Korrelation** misst die **gemeinsame Streuung** von **X** und **Y** im **Verhältnis zu** den **Streuungen** jeder **einzelnen** dieser beiden **Variablen**.
- **Regression sucht** die **Geraden** deren **Prognosen** möglichst **nahe an** den **tatsächlichen Werten** liegen – die also die Quadratsumme der Residuen minimieren.

PRÜFUNGSVORBEREITUNG LETZTE EINHEIT:

SPSS: welche möglichkeiten gibt es umzucodieren
Wo codiert man
Fehlende Werte was ist dabei zu beachten
Wann codiert man
Standardabweichung mittelwert händisch auszurechnen mit einfachen zahlen
Chi2 test sagt aus ob zusammenhang signifikant zwischen nominalen variablen
Schema
Kein normalverteilungstest oder levine tet
Spss screenshot – keine befehle
Freiheitsgrade verstehen
Z wert aus tabelle ablesen (alpha = irrtumswahrscheinlichkeit