UNIVERSITÄT PADERBORN

Prof. Dr. Christoph Kulgemeyer

AG Didaktik der Physik

Fakultät für Naturwissenschaften

Einführung in quantitative Methoden der Datenauswertung

Ausgangssituation

- Sie haben in einer größeren Stichprobe einen Test zum Verständnis der newtonschen Mechanik eingesetzt (Force Concept Inventory, FCI).
- Sie wollen überprüfen, ob der Test
  - ein zusammenhängendes Fähigkeitskonstrukt erhebt („Verständnis der Mechanik“)
  - in der Lage ist, innerhalb dieses Konstrukts Teilfähigkeiten zu messen (z.B. „Verständnis des 3. Axioms „actio = reactio“)
  - von Frauen und Männern unterschiedlich gut gelöst wird
- Psychometrisch spricht man hier von „Reliabilitätsanalysen“
- Gezeigt wird das Vorgehen nach klassischer Statistik

Schritt 1: Die Daten tabellarisch erfassen

- Datentabelle anlegen
  - geordnete Variablennamen und prägnante Labels
  - Typ und Messniveau der Variablen festlegen (metrisch, ordinal, nominal)
  - Codes für fehlende Werte festlegen (missing values)
- Daten eingeben
  - oder aus einer Excel-Tabelle importieren
  - !! Alle Datensätze in einer Tabelle gemeinsam erfassen
  - !! Zunächst die Rohdaten erfassen: Was wurde angekreuzt — statt „die Antwort ist richtig“!
- Werte für berechnete Variablen berechnen
  `Transformieren -> Variable berechnen`
  - Ankreuzung --> richtig / falsch
  - Summe der richtigen Antworten in einer Skala (Test)

Schritt 2: Einen ersten Blick auf die Daten werfen

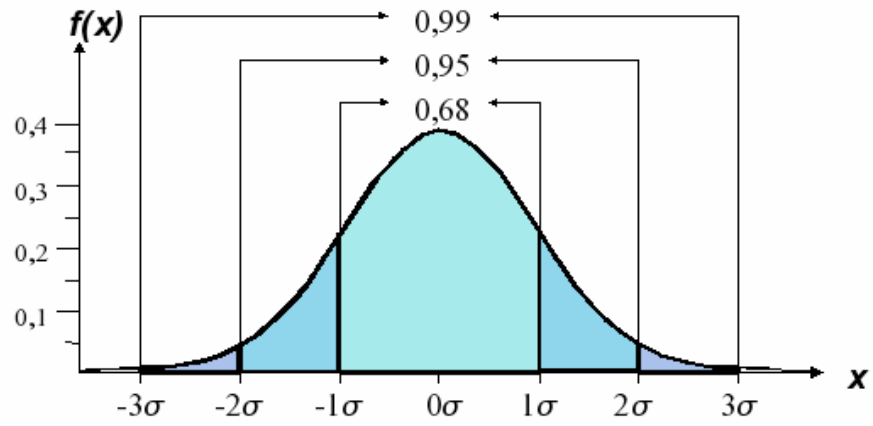
- Datentabelle sichten
  - viele fehlende Werte?
  - auffällige Werte (außerhalb des erwartbaren Wertebereichs)?
- Diagramme ausgeben
  - Balkendiagramme für Häufigkeitsverteilungen
  - x-y-Diagramme
- Tabellen ausgeben
  - Mittelwert / Median / Modus
  - Spannweiten, Varianzen
  - Kreuztabellen

Schritt 2: Einen ersten Blick auf die Daten werfen

- Verteilungen anschauen
  - Kann man bei intervallskalierten Daten von Normalverteilung ausgehen?
    - Voraussetzung z.B. für t-Test, Standardabweichung etc.
    - `Analysieren -> nichtparametrische Tests -> K-S bei einer Stichprobe`
  - oder sind sie nicht-normalverteilt, z.B. links-/ rechtschief / zweihöckrig?
    - dann sind nicht-parametrische Analysen zu wählen
  - Sind die Werte einer nominalen Messgröße so wie erwartet verteilt?
    - z.B. Mädchen und Jungen in der Stichprobe gleichverteilt?

Exkurs: Normalverteilung

- gilt häufig für Merkmale, die in einer Population vorliegen
- Kennzeichen
  - Glockenform, symmetrisch
  - Mittelwert, Median und Modus fallen zusammen
- für die z-Verteilung oder Standard-Normalverteilung gilt:
  - Mittelwert $\mu = 0$
  - Standardabweichung $\sigma = 1$ (Sigma)
  - d.h.: Fläche unter der Glockenkurve = 1



- Normalverteilung ist Voraussetzung für die Berechnung vieler Kennwerte und Testverfahren, z.B.
  - Standardabweichung
  - t-Test (Normalverteilung beim t-Test als Voraussetzung ist allerdings relativierbar)

Schritt 3: Qualität der Skalen einschätzen

Oft werden Personenmerkmale (Messgrößen) nicht anhand einzelner Items, sondern von Skalen erhoben, d.h. anhand einer Gruppe von Items, die das gleiche Konstrukt erheben sollen, z.B. Fachwissen über Optik, Intelligenz oder Lesefähigkeit.

- Innere Konsistenz: Wie stark hängt die Beantwortung der Items dieser Gruppe zusammen?
  - Reliabilitätsanalysen: Cronbachs alpha berechnen
    `Analysieren -> Skalierung -> Reliabilität`
  - durch Ausschluss einzelner Items die Reliabilität erhöhen

Exkurs: Cronbachs alpha

- Cronbachs alpha ist das gängige Maß für die Abschätzung der inneren Konsistenz einer Gruppe von Items („Skala“), die ein bestimmtes Merkmal erheben sollen, z.B. Interesse an Physik
- In die Berechnung gehen ein (Abb. nach http://de.wikipedia.org/wiki/Korrelationskoeffizient (14.6.11)
  - die Anzahl der Items
  - die mittlere Interkorrelation zwischen dem Ankreuzverhalten bei den Einzelitems
  - bzw. die Varianzen der Einzelitems bezogen auf die Varianz des Skalengesamtwerts

$$\alpha_{st} = \frac{N \cdot \bar{r}}{1 + (N-1) \cdot \bar{r}}$$

$$\alpha = \frac{N}{N-1}\left(\frac{\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right)$$

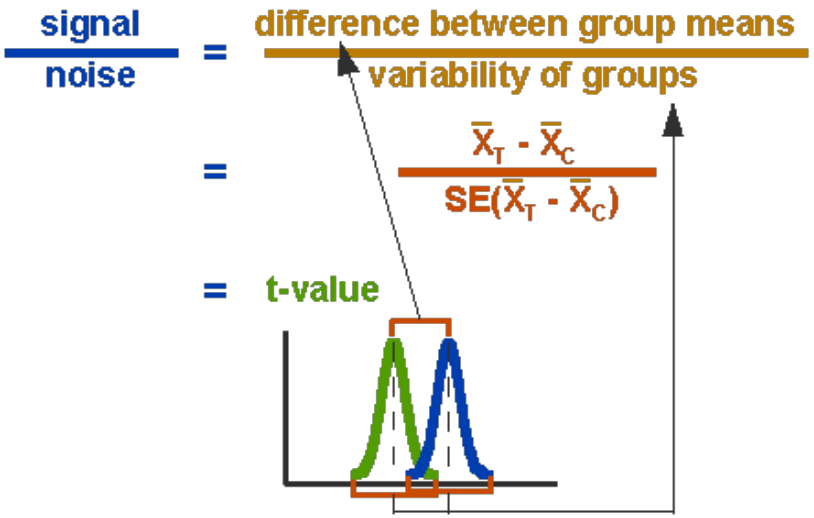
  - systematisch verbessert sich — bei gleicher mittlerer Interkorrelation – alpha mit steigender Itemzahl (--> excel-Berechn.)
- Durch Ausschluss bestimmter Items lässt sich der alpha-Wert steigern
  - dabei muss man fachdidaktisch allerdings darauf achten, dass nicht gerade die „interessanten“ Items herausfallen
- alpha ist niedrig, wenn die Skala in zwei wenig verbundene Subskalen zerfällt (--> Faktorenanalyse vornehmen)

Exkurs: Nullhypothese H_0

- **Eine Hypothese ist eine zu prüfende Annahme. Häufig formuliert man als Nullhypothese die Annahme, dass hinsichtlich einer Messgröße Gleichheit zwischen zwei Probandengruppen besteht oder dass ein Zusammenhang zwischen zwei Messgrößen nicht gegeben ist (obwohl man eigentlich vermutet [oder hofft], dass er besteht).**
  - **Beispiele für Nullhypothesen:**
    - **In meiner Stichprobe sind gleich viele Mädchen und Jungen**
    - **Mädchen und Jungen unterscheiden sich nicht in der Intelligenz**
    - **Es besteht kein Zusammenhang zwischen Lesefähigkeit und Intelligenz**
- Überprüft wird dann, wie hoch die Wahrscheinlichkeit ist, dass man einen Fehler macht, wenn man die H_0 verwirft (Irrtumswahrscheinlichkeit; „alpha-Fehler“).
  - Dafür setzt man sich einen Grenzwert, i.d.R. p = 1% oder 5%
  - Liegt die Irrtumswahrscheinlichkeit unter diesem Grenzwert, verwirft man die NH
  - Damit ist aber die Alternativhypothese nicht automatisch bestätigt!

t-Test Analysieren -> Mittelwerte vergleichen -> T-Test

- Der t-Test untersucht, ob ein Unterschied in den Mittelwerten einer Messgröße zwischen zwei Population als systematisch oder zufällig angenommen werden kann/muss.



http://www.socialresearchmethods.net/kb/Assets/images/stat_t3.gif

- Voraussetzung:
  - intervallskalierte Daten
  - Normalverteilung
  - die Varianzen des Parameters sind in den beiden Populationen gleich
  - ABER: der t-Test ist recht robust gegenüber Nicht-Normalverteilung oder ungleichen Varianzen, wenn die Stichproben ungefähr gleich groß sind und nicht zu klein sind ($n_1 = n_2 > 30$) (vgl. Rasch et al. 2006, 59)
- Varianten:
  - t-Test für unabhängige Stichproben (z.B. Schüler in zwei Kursen)
  - t-Test für verbundene Stichproben (z.B. prä-post-Messung in einem Kurs)

t-Test - SPSS-Output

Test bei unabhängigen Stichproben

		Levene-Test der Varianzgleichheit		T-Test für die Mittelwertgleichheit						
									95% Konfidenzintervall der Differenz	
		F	Signifikanz	T	df	Sig. (2-seitig)	Mittlere Differenz	Standardfehler der Differenz	Untere	Obere
Klausurpunkte	Varianzen sind gleich	5,039	,028	1,124	64	,265	7,61	6,772	-5,918	21,141
	Varianzen sind nicht gleich			1,085	49,037	,283	7,61	7,015	-6,486	21,708

Prüfgröße t

*Mit dem **Levene-Test** wird automatisch die Voraussetzung der Varianzhomogenität überprüft. Geprüft wird die H0, die Varianzen seien gleich. Wird der Test signifikant ($p < 0.05$) spricht dies für die H1, dass die Varianzen nicht gleich sind. Ist dies der Fall, werden die Werte in der Zeile „Varianzen sind nicht gleich“ verwendet. Ansonsten nimmt man die erste Zeile und ignoriert die zweite.*

Anzahl der Freiheitsgrade (für den Fall ungleicher Varianzen ist diese Anzahl korrigiert!

*p-Wert (nicht signifikant, da $p >$ **0,05***

entnommen aus:
http://www.metheval.uni-jena.de/projekte/spss-interaktiv/spss-interaktiv.php

Exkurs: Effektstärke

- Statistisch signifikante Unterschiede können fachdidaktisch bedeutsam oder auch vernachlässigbar sein.
- Wenn sich ein Mittelwertunterschied als systematisch erwiesen hat, z.B. durch einen t-Test, kann man abschätzen, wie groß der Effekt ist.
- Beispiel: Cohens d bei gleichen Stichprobengrößen
  $d = (\mu_1 - \mu_2) / \sigma_{ges}$
  - $\mu_{1,2}$: Mittelwerte der beiden Stichproben
  - σ_{ges}: Standardabweichung in der Gesamtstichprobe
    geschätzt aus den mittleren Varianzen $\sigma_{ges} = (\sigma_1^2 + \sigma_2^2)^{1/2}$
  - 0,2: kleiner Effekt / 0,5: mittlerer Effekt / 0,8: starker Effekt
- Keine zu hohen Erwartungen an Effektstärken hegen
  bezogen auf unterrichtliche Innovationen sind Effekte einzelner Maßnahmen meist eher klein bis mittel.
- Weitere Größen für Effektstärken
  mit online-Berechnung (sehr empfehlenswert):
  - http://www.phil.uni-sb.de/~jakobs/seminar/vpl/bedeutung/effektstaerketool.htm

Mann-Whitney-U-Test

- Der U-Test untersucht, ob ein Unterschied in den Rangreihen bez. einer Messgröße zwischen zwei Population als systematisch oder zufällig angenommen werden kann/muss.
  `Analysieren -> Nichparamterische Tests -> Unabhängige Stichproben`
- Voraussetzung:
  - ordinalskalierte Daten
  - unabhängige Stichproben
  - implizite Annahme: die Ränge sind äquidistant
- auch für intervallskalierte Daten anwendbar
  - z.B. wenn keine Normalverteilung vorliegt
  - oder nur wenige Probanden einbezogen waren
- für abhängige Stichproben: Wilcoxon-Test verwenden)
  `Analysieren -> Nichparamterische Tests -> Verbundene Stichproben`

chi²-Test

- Wird für nominalskalierte Daten verwendet
  - Wie wahrscheinlich ist das Vorkommen der gemessenen Häufigkeiten der Werte einer Messgröße gegenüber einer Vergleichsverteilung?
  - Ausgangsannahmen
    - i.d.R. Gleichverteilung
    - ggf. auch selbst gesetzte Verteilung (auf Basis bereits vorliegender Forschungsergebnisse)
- eindimensionaler chi²-Test
  - Beispiel: Sind in der Stichprobe gleich viele Mädchen und Jungen?
  - Nullhypothese: Gleichverteilungsannahme
  - `Analysieren -> nichtparametrische Tests -> chi²`
- zweidimensionaler chi²-Test
  - Beispiel: Sind Jungen und Mächen gleich auf VG und KG verteilt?
  - Nullhypothese: Gleichverteilung
  - `Analysieren -> Deskriptive Statistik -> Kreuztabellen`

chi²-Test SPSS-Output

Geschlecht

	Beobachtetes N	Erwartete Anzahl	Residuum
Weiblich	44	53,5	-9,5
Männlich	63	53,5	9,5
Gesamt	107		

Differenzen ($f_b - f_e$)
Je größer die Differenzen, umso größer ist der χ^2-Wert!

in der Stichprobe beobachte Häufigkeiten (f_b)

Unter Annahme der Gleichverteilung erwartete Häufigkeiten (f_e)

Statistik für Test

	Geschlecht
Chi-Quadrat[a]	3,374
df	1
Asymptotische Signifikanz	,066

a. Bei 0 Zellen (,0%) werden weniger als 5 Häufigkeiten erwartet. Die kleinste erwartete Zellenhäufigkeit ist 53,5.

Alle drei Werte sollten in einem statistischen Bericht auftauchen!
☞ *Interpretation: Die Verteilung von Männern und Frauen in dieser Vorlesung weicht nicht signifikant von der erwarteten Gleichverteilung ab ($\chi^2 = 3.37$; $df = 1$; $p > .05$).*

entnommen aus:
http://www.metheval.uni-jena.de/projekte/spss-interaktiv/spss-interaktiv.php

Schritt 5: Zusammenhänge untersuchen

Hängt die Ausprägung zweiter Merkmale innerhalb einer Probandengruppe miteinander zusammen (z.B. Vorwissen und Leistungszuwachs)?

- Nullhypothesen festlegen
  - i.d.R.: „Die beiden Merkmale hängen nicht zusammen“
- Signifikanzniveau festlegen (maximale akzeptierte Irrtumswahrscheinlichkeit für das Ablehnen der NH)
- Tests durchführen, dabei Datenqualität beachten!
  - metrische Daten: z.B. Pearson Koeffizient *r*
    - Hängen Lesefähigkeit und Intelligenz zusammen?
  - ordinale Daten: z.B. Spearman, Kendall‘s tau-b
    - Hängen Lebensalter und Gewicht zusammen?
  - nominale Daten: z.B. Kontingenz-Koeffizient
    - Hängen Blutgruppe und Geschlecht zusammen?

Pearson Korrelation *r* (Produkt-Moment-Korrelation)

`Analysieren -> Korrelation -> Bivariat`

- Voraussetzungen
  - intervallskalierte Daten (oder intervallskaliert und dichotom; z.B. Intelligenz und Geschlecht)
  - Normalverteilung (bei intervallskalierten Daten)
  - halbwegs plausible Annahme der Möglichkeit eines linearen Zusammenhangs (bei Annahmen über andere Zusammenhänge, müssen die Messwerte erst linearisiert werden)
- Wertebereich von r: -1 ... +1
  - -1 (direkter gegenläufiger Zusammenhang) ...
  - 0 (kein Zusammenhang) ...
  - +1 (direkter gleichlaufender Zusammenhang)
- *r* kann direkt als Effektstärkemaß verwendet werden, Konvention:
  - r ≥ 0,1 kleiner Effekt / r ≥ 0,3 mittlerer Effekt / r ≥ 0,5 großer Effekt
  - Achtung: *r* ist nicht intervallskaliert
  - r^2 entspricht der aufgeklärten Varianz in Messgröße A durch B

Pearson Korrelation *r* SPSS Output

- H_0: Intelligenz (gemessen mit KFT) und Leseverständnis hängen nicht zusammen
- Ergebnis
  - die Korrelation beträgt ,419 (mittlerer Effekt)
  - sie ist hochsignifikant (Fehler bei Verwerfen von **H_0** ist 0,000)
    - d.h. die Irrtumswahrscheinlich, wenn man die Nullhypothese ablehnt, ist kleiner als 1%

Korrelationen

		KFT_ges KFT_Gesamt	LeseVerst Leseverständnis
KFT_ges KFT_Gesamt	Korrelation nach Pearson	1	,419**
	Signifikanz (2-seitig)		,000
	N	198	198
LeseVerst Leseverständnis	Korrelation nach Pearson	,419**	1
	Signifikanz (2-seitig)	,000	
	N	198	198

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Rangkorrelation

`Analysieren -> Korrelation -> Bivariat`

- Voraussetzungen
  - ordinal skalierte Daten
  - getestet wird, wie sehr die Rangreihen der Probanden bei zwei Messgrößen miteinander in Verbindung stehen
- Wertebereich - 1 ... 0 ... +1
- Koeffizienten
  - Spearman‘s rho
    - Grundlage: nach Rängen geordnete Messdaten
    - Annahme: die Ränge in einer Rangfolge sind äquidistant
  - Kendall‘s tau-b
    - Grundlage: nur die Rangfolge selbst (keine Zusatzannahme)