Messen und Skalen

Frank Fuhlbrück

Seminar: Noten, Studienpunkte und Automatisierung WiSe 2010/11

2010-11-29

Relative

Definition (Relativ)

Sei A eine Menge und $R_{i,1\leq i\leq k}$ Relationen auf A. Dann heißt $\mathbb{A} = \langle A, R_1, \ldots, R_k\rangle$ Relationalstruktur oder Relativ und A Grundmenge von $\mathbb{A}$.

empirisches und numerisches Relativ

Grundmenge enthält „Objekte der Anschauung“ $\Rightarrow$ empirisches Relativ
Grundmenge enthält Zahlen $\Rightarrow$ numerisches Relativ

Messen, Skalieren, Skala

Die folgende Definition sollte semantisch der in [Ath74] entsprechen (mathematisch Formulierung eher der in [Fri72]).

Definition (Skalieren,Skala)

Seien $\mathbb{A} = \langle A, R_1, \ldots, R_k \rangle$ und $\mathbb{B} = \langle B, S_1, \ldots, S_k \rangle$ Relationalstrukturen (R_i, S_i Relationen auf A bzw. B). *Skalieren* ist die Konstruktion einer homomorphen Funktion f von $\mathbb{A}$ nach $\mathbb{B}$. Das Tripel $\mathcal{S} = (\mathbb{A}, \mathbb{B}, f)$ wird dann als *Skala* bezeichnet.

Definition (Messen)

Abhängig vom Skalenniveau (s.u.) wird *Skalieren* als *Messen* bezeichnet.

fundamentale und abgeleitete Skalierung

fundamental und abgeleitet

Skala $\mathcal{S} = (\mathbb{A}, \mathbb{B}, f)$:
$\mathbb{A}$ empirisch: fundamentale Skalierung/Messung
$\mathbb{A}$ bereits numerisch: abgeleitete Messung

Beispiel (fundamental und abgeleitet)

Strecke und Zeit werden fundamental gemessen, Geschwindigkeit abgeleitet

Zufallsvariable

implizit definierte ZV

Skala $\mathcal{S} = (\mathbb{A}, \mathbb{B}, f)$ und $|A| = n$:
f kann auch als B-wertige ZV aufgefasst werden
Wahrscheinlichkeitsraum $(A, \wp(A), P)$ mit $P(A') = \frac{|A'|}{n}$ f.a. $A' \subseteq A$
aus zufälliger, gleichverteilter Ziehung aus A.
Notation: $X_{\mathcal{S}} := f$

Klassen von Skalen

Transformationen

Transformation t: isomorphe Abbildung von $\mathbb{B}$ nach $\mathbb{C}$ für zwei Skalen $(\mathbb{A}, \mathbb{B}, f)$ und $(\mathbb{A}, \mathbb{C}, g)$

Klassifikation von Skalen

B und $C \subseteq \mathbb{R}$: Klasse/Niveau von $(\mathbb{A}, \mathbb{B}, f)$ ist Menge der möglichen Transformationen.
Je kleiner die Menge der möglichen Transformationen, desto höher der Informationsgehalt.
$\exists$ 4 gängige Skalentypen (keinesfalls die einzigen).

Nominalskala

Transformationen

Jede bijektive Abbildung.

Relation

zweistellige Relation $\equiv$
$a \equiv b$: Objekte stimmen in Eigenschaft überein

mögliche Maßzahlen und Tests

Modalwert, einzelne Häufigkeiten, χ^2 - Unabhängigkeitstest

Beispiel

Zugehörigkeit zu Gruppe o.ä.

Ordinalskala

Transformationen

Jede streng monotone Abbildung.

Relation

zusätzlich zweistellige Relation $\prec$
$a \prec b$: Eigenschaft von a steht in Rangfolge vor der von b

mögliche Maßzahlen

Median, Extremwerte, Quantile, Spearman-Korrelationskoeffizient

Beispiel

Präferenzordnung (z.B. bei Wahlen)

Intervallskala

Transformationen

Jede affine Abbildung ohne Multiplikation mit 0.

Relation

zusätzlich vierstellige Relation R_d
$(a, b, c, d) \in R_d$: Unterschied von a zu b entspricht dem von c zu d.

mögliche Maßzahlen

ar. Mittelwert, (emp.) Varianz, MAD, Interquartilsabstand, Pearson-Kor., lineare Modelle

Beispiel

Temperatur

Intervallskala

Transformationen

Jede affine Abbildung ohne Multiplikation mit 0.

Relation

zusätzlich vierstellige Relation R_d
$(a, b, c, d) \in R_d$: Unterschied von a zu b entspricht dem von c zu d.

mögliche Maßzahlen

ar. Mittelwert, (emp.) Varianz, MAD, Interquartilsabstand, Pearson-Kor., lineare Modelle

Beispiel

Jahreszahlen

Verhältnisskala

Transformationen

Jede lineare Abbildung außer Multiplikation mit 0.

Relation

zusätzlich vierstellige Relation R_r
$(a, b, c, d) \in R_r$: a verhält sich zu b, wie c zu d.

mögliche Maßzahlen

geom. Mittelwert

Beispiel

Temperatur

Maßzahlen

Definition (empirisches Quantil)

Sei $x = (x_1, \ldots, x_n)$ eine Stichprobe. Für $p \in [0, 1]$ ist ein p-Quantil x_p ein Wert, sodass $|\{x_i | x_i \leq x_p\}| \cdot \frac{1}{n} \geq p$ und $|\{x_i | x_i \geq x_p\}| \cdot \frac{1}{n} \geq 1 - p$. Das 0.5-Quantil $x_{0.5}$ heißt *Median*.

Definition (arithmetisches Mittel)

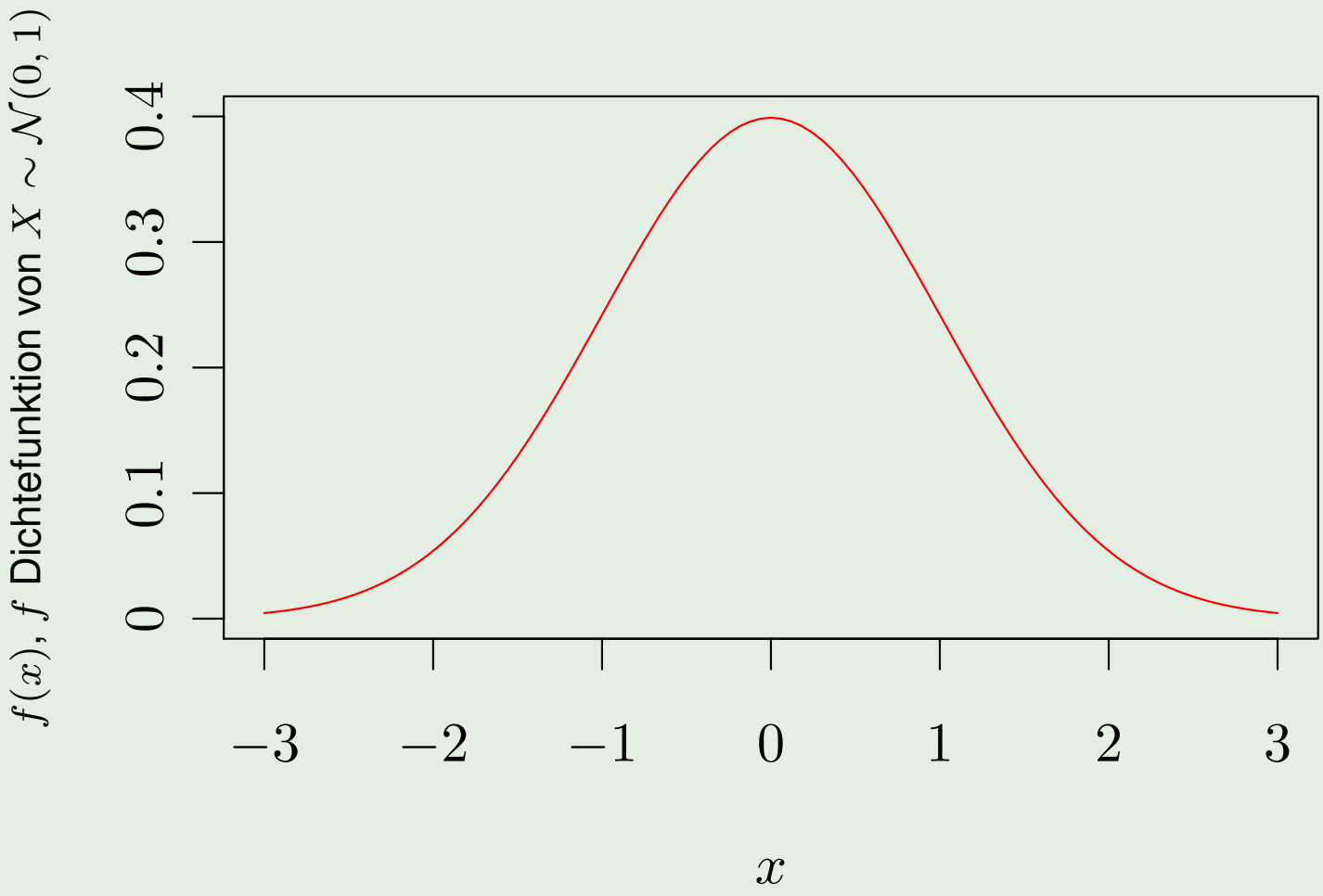
Sei $x = (x_1, \ldots, x_n)$ eine Stichprobe. $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ heißt *arithmetisches Mittel*.

Definition (empirische Standardabweichung)

Sei $x = (x_1, \ldots, x_n)$ eine Stichprobe. s_x mit $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2$ heißt *empirische Standardabweichung*.

Normalverteilung

Funktionsgraph



Normalverteilung

Definition (normalverteilt)

Stetige ZV X mit Dichte $f(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma^2}\right)}$ heißt *normalverteilt* $(X \sim \mathcal{N}(\mu, \sigma))$.

Satz (Zentraler Grenzwertsatz)

Seien $X_{i,1\leq i\leq n}$ unabhängige, identisch verteilte ZV mit Erwartungswert μ und Varianz σ^2, sowie $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Dann gilt für

$$Z := \lim_{n\to\infty} \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}: \qquad Z \sim \mathcal{N}(0,1).$$

Normalverteilung: Verteilung von Messfehlern

Fehlermodell beim Messen/Skalieren

X tatsächlicher Wert, Y idealer Wert , E Messfehler

$X = Y + E$

Normalverteilungsannahme für E

Ann.: Fehler ist Summe vieler einzelner Fehler (mit ähnlicher Vert.)

$\Rightarrow E$ ist annähernd normalverteilt.

Normalverteilung: Verteilung in Populationen

Beobachtungen

- viele Eigenschaften in Popul. sind annähernd normalverteilt: Körpergröße etc.
- Eigenschaften des Individuums beruhen auf vielen unabhängigen Faktoren:
  Genetik, (abiotische) Umwelt, Kultur . . .

Schluss

Postulat: (bestimmte) psychologische und soziologische Eigenschaften werden als normalverteilt angenommen.

Was wird gemessen/skaliert?

Theoretisches Konstrukt

- häufig vorwissenschaftliche Bedeutung
- grobe Klassifikation (Zustand/Prozess/Individuum/Gruppe)
- Zusammenhang mit anderen Konstrukten

Beispiel (Schulleistung [Fri72, S.25])

Schulleistung ist das Ergebnis [Zustand] des schulischen und außerschulischen Lernens [Prozess] eines Schülers.

Wie wird gemessen/skaliert?

Operationalisiertes Konstrukt

- Welche Beobachtung entspricht welcher Ausprägung des Konstrukts?

Repräsentationsproblem

Begründung des postulierten Homomorphismus

- Notfalls Anpassung der operationalisierten Definition (z.B. NV-Annahme).

Beispiel (Schulleistung)

Schulleistung ist die Anzahl der Punkte in meinem Test.
Schulleistung ist das Ergebnis eines Standardtests $\Rightarrow$ mein Test muss die Schüler in der selben Reihenfolge ordnen.

Skalierung und Interpretation?

Eindeutigkeitsproblem

Welche numerischen Zusammenhänge dürfen interpretiert werden?

- Welches Skalenniveau liegt tatsächlich vor?

Testgütekriterien

Validität

Messe ich, was ich verspreche zu messen?

Reliabilität

Liefert mehrmaliges Testen ähnliche Ergebnisse?

Objektivität

Hängt der Test nicht vom Versuchsleiter ab?

(Normierbarkeit)

Lassen sich die Testergebnisse von Individuen in die einer Population einordnen?

Definition der Noten

Beispiel (Schulnoten Deutschland)

[. . .]Skala anzuwenden:
1. „sehr gut“(1) - wenn die Leistung den Anforderungen in besonderem Maße entspricht,
2. „gut“(2) - wenn die Leistung den Anforderungen voll entspricht,
3. „befriedigend“(3) - wenn die Leistung im Allgemeinen den Anforderungen entspricht,
4. „ausreichend“(4) - wenn die Leistung zwar Mängel aufweist, aber im Ganzen den An- forderungen noch entspricht,
5. „mangelhaft“(5) - wenn die Leistung den Anforderungen nicht entspricht, jedoch er- kennen lässt, dass die notwendigen Grundkenntnisse vorhanden sind und die Mängel in absehbarer Zeit behoben werden können,
6. „ungenügend“(6) - wenn die Leistung den Anforderungen nicht entspricht und selbst die Grundkenntnisse so lückenhaft sind, dass die Mängel in absehbarer Zeit nicht behoben werden können.

1.Fall Zensur als erste (fundamentale) Skalierung

Was wird skaliert?

- unterschiedliche theoretische Konstrukte (s. Beispiel)
- gemeinsam: Fähigkeit, Leistung o.ä.
- Zeit teilweise explizit in Definition [Ath74]

Wie wird skaliert?

- Gutachter beurteilt/en anhand von nominalen Kriterien (s. Beispiel) (und einigen sich ggf.)
- Ausrichtung an Aufgaben-/Anforderungsbereichen

1.Fall Zensur als erste (fundamentale) Skalierung

Beispiel (mündl. Abitur deutsch [Kul02, S. 32])

Eine Leistung kann mit „gut“ bewertet werden, wenn

- der Inhalt des vorgegebenen Materials präzise erfasst und eigenständig dargestellt wird
- das Thema bzw. Problem differenziert erläutert wird
- Struktur, Funktion und Intention des Materials und dessen Stilmittel zutreffend bezeichnet
- sowie Wirkungsmöglichkeiten überzeugend eingeschätzt werden
- differenzierte Kenntnisse und Einsichten nachgewiesen werden
- Zusammenhänge eigenständig erkannt werden
- ggf. ein Urteil oder eine Stellungnahme begründet dargelegt werden
- der Vortrag strukturiert erfolgt
- eine überzeugende sprachliche Darstellung in Vortrag und Gespräch geleistet wird.

1.Fall Zensur als erste (fundamentale) Skalierung

Beispiel (Theoretisches Konstrukt)

Die Diplomarbeit soll zeigen, dass der Prüfling in der Lage ist, innerhalb einer vorgegebenen Frist **ein Problem** aus seinem Fach **selbständig** nach **wissenschaftlichen Methoden** zu bearbeiten.
[Prüfungsordnung Informatik 2003 §19 Abs. 1, Hervorhebung hinzugefügt]
Eine operationale Definition wird nicht angegeben!

2.Fall Zensur als zweite, abgeleitete Skalierung

Was wird skaliert?

- unterschiedliche theoretische Konstrukte (s. Beispiel)
- gemeinsam: Fähigkeit, Leistung o.ä.
- Zeit teilweise explizit in Definition [Ath74]

Wie wird skaliert?

- Punktbewertung (Bewertungseinheiten) ähnlich wie Zensurenbewertung im Fall 1
- Skalierung Punkte $\rightarrow$ Zensur mittels Tabelle oder Prozentwerten und/oder
- Skalierung sodass (Normal-)Verteilung in Gruppe erfüllt ist
- Verteilung der Punkte an AFB orientiert
- Punkte aus verschiedenen Teilgebieten (z.B. Analysis und Algebra) werden addiert

2.Fall Zensur als zweite, abgeleitete Skalierung

Beispiel (Schriftliche Abiturprüfungen Berlin [Sen10, S.13])

- 1+ ab 95%, dann in 5%-Schritten bis
- 4 ab 45%, dann in 9%-Schritten bis
- 5- ab 9%

  $\Rightarrow$ zwei getrennte (affine) Transformationen von Intervallen (z.B. $[0.95, 1]$ auf 15(Punkte))

Operationen

gewichtetes arithmetisches Mittel

Noten werden mit Faktoren gewichtet und (ggf. in mehreren Schritten) addiert, dann wird durch Gesamtgewicht geteilt.

mögliche stochastische Erklärung/Deutung

- Für Zensurenskala $\mathcal{Z}$ wird $\mathbf{E}(X_{\mathcal{Z}})$ geschätzt mit (mod.) $\overline{x}$.
- $\overline{x}$ ist ML-Schätzer für $\mathcal{N}$ und erwartungstreu.
- Gewicht entspricht der Zuverlässigkeit der Messung oder der Auftretenswahrscheinlichkeit der Situation.
- Bei bekannter Verteilung (inkl. Varianz): Schätzung für Erwartungswert zeigt Leistungsspektrum

auftretende Skalen

Bewertungseinheitenskala

wird als Intervallskala oder sogar Verhältnisskala („ausreichend“ bei der Hälfte der maximalen Leistung) begriffen

Zensurenskala

wird als Intervallskala oder Ordinalskala begriffen (s. Operationen)

Verteilungsannahmen

Verteilung der Zensuren in Klasse/Kurs . . .

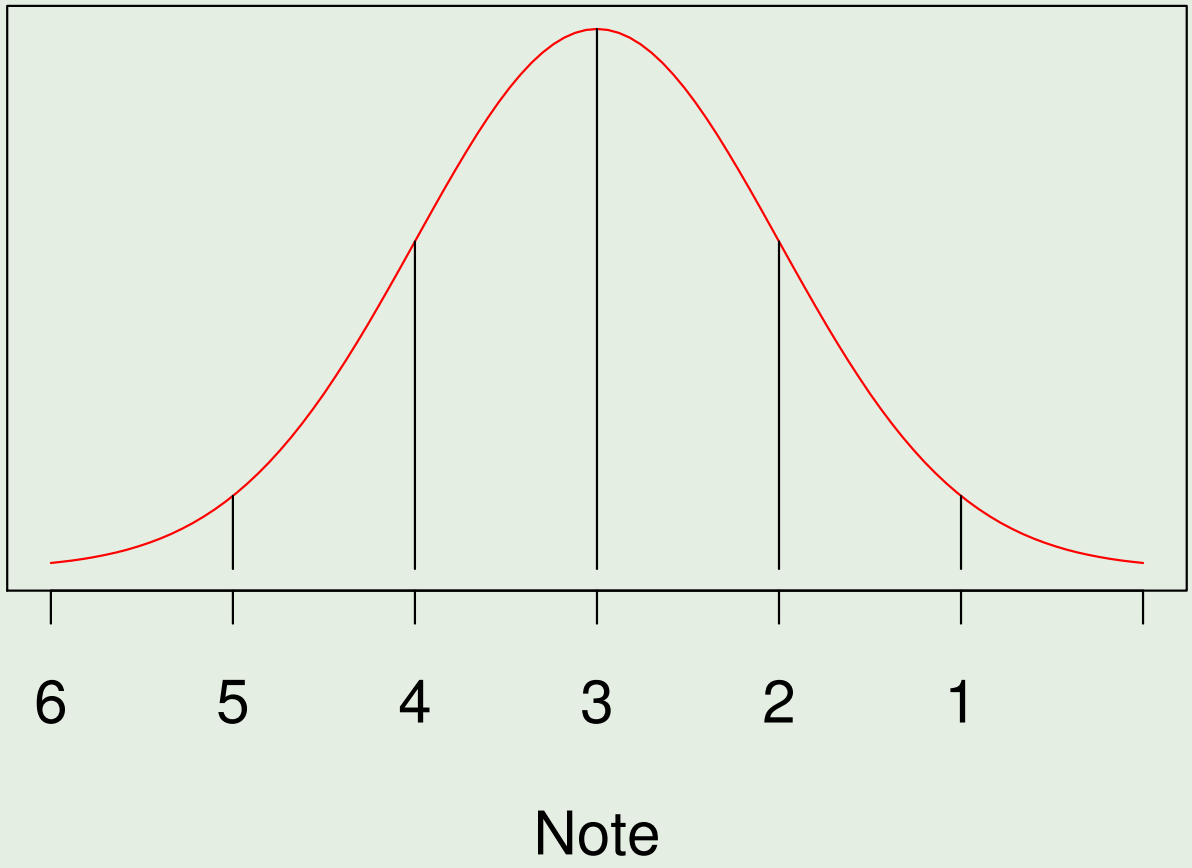
„Glockenkurve“ ist zumindest erwünschtes Ergebnis.
Annahme Normalverteilung der Leistung in Population (ähnlich Intelligenz).

Verteilung der Zensuren einer Person

Normalverteilungsannahme für Messfehler?
Zeitpunkt der Bewertung wird meist kaum berücksichtigt

Verteilungsannahmen

Beispielbewertung nach NV



Prozess

Theoretisches Konstrukt, Was?

existiert z.T., aber variiert

operationalisiertes Konstrukt, Wie?

existiert kaum (nicht z.B. bei [Ath74], PO 2003)

Begründung des Homomorphismus, Wie?

ohne operationalisiertes Konstrukt nicht möglich

Operationen und Deutung, Wie?

normierte Deutung (bundesweite Zulassungsgrundlage) ohne Normierung,
Intervallskaladeutung von Ordinalskala (s.u.)

Testgütekriterien

Validität

Konstruktvalidität ohne operationale Definition kaum möglich,
Vorhersagevalidität Abitur für Studiumsleistung gut ($r = 0,35$ laut [KB10])
für Arbeitsleistung nicht unbedingt

Reliabilität

Wiederholung schwierig

Objektivität

variiert (mündliche Mitarbeitsnoten über mündl. Prüfungen bis zentrale Prüfung)
(aber $\leq$ z.B. Scholastic Aptitude Test)

(Normierbarkeit)

insbesondere nicht wie bei z.B. Intelligenztests

Verteilungsannahmen Individuum

Normalverteilung bei Individuum (Messfehler)

- führt Funktion der Rückmeldung zur Verbesserung ad absurdum
- Problem bei Extremen: kein Ausgleich möglich
  (obwohl nach NV bel. gutes Ergebnis möglich sein müsste)

Beispiel

Noten: 4,3,2,1,1,1,1 $\Rightarrow$ Gesamtnote: 2
Noten: 6,5,4,3,4,3,1,2,1,2,1,1,1,1,1,1,1,1,1,1,1 $\Rightarrow$ Gesamtnote: 2
(bei $p_{\chi^2} < 0.05$)
Noten: 2,3,4,1,1 $\Rightarrow$ Gesamtnote: 2
Noten: 1,2,3,1+,1+ $\Rightarrow$ Gesamtnote(ohne Beachtung Tendenz): 2
(tritt bei Tendenz entsprechend für 15+ Punkte auf)

Verteilungsannahmen Individuum

Alternativen

- zeitliche Entwicklung einrechnen (z.B. lineares Modell mit Zeit als UV)?
- auf Verteilungen testen und ggf. weitere Maße (z.B. Streuung) mit angeben?
- anderen (robusteren) Schätzer für Erwartungswert (z.B. getrimmte Mittel, Median)?

Verteilungsannahmen Gruppe

Normalverteilung in Gruppe

- sign. Unterschiede in Vergleichsarbeiten u. Tests $\Rightarrow$ nicht jede Gruppe repräsentativ
- Unterschiede in Jahrgang u.ä. wird nicht sichtbar

Alternativen

- keine Annahme über Verteilung in Gruppe nutzen?
- reale Verteilung vorab testen und an diese anpassen?

Bewertungseinheiten und Zensuren

Auswirkung verschiedener Transformationen

- Tests mit Punkten 18/100 und 60/100 $\Rightarrow$ Schätzung $\mathbf{E}(X_{\mathcal{B}}) = 36$
- $\Rightarrow \mathbf{E}(X_{\mathcal{Z}}) = 4$ Punkte?
- aber Noten 2 Punkte und 8 Punkte $\Rightarrow \mathbf{E}(X_{\mathcal{Z}}) = 5$ Punkte?

Rundungsfehler

- Tests mit Punkten 94/100, 94/100 und 97/100 ergeben als Gesamtnote: 14 Punkte

Transformation

Die Notenermittlung ist keine erlaubte Transformation für Intervallskalen.

Median

Median und andere Quantile

- bei symm. Verteilung robuster Schätzer für $\mathbf{E}(X_{\not z})$
- ggf. zusätzliche andere Quantile $(x_{0.1}, x_{0.9})$ als „Streuungsmaß“
- auch für Ordinalskala definiert

Beispiel

Noten: 4,3,2,1,1,1,1 $\Rightarrow$ Gesamtnote: 1 (1,4)
Noten: 6,5,4,3,4,3,1,2,1,2,1,1,1,1,1,1,1,1,1,1,1 $\Rightarrow$ Gesamtnote: 1(1,4)
Noten: 2,3,4,1,1 $\Rightarrow$ Gesamtnote: 2 (1,4)
Noten: 1,2,3,1+,1+ $\Rightarrow$ Gesamtnote(ohne Beachtung Tendenz): 1(1,3)
Noten: 1,4,1,4,1,4 Gesamtnote: ?(1,4)
Noten: 1,4,1,4,1,4,1 Gesamtnote: 1(1,4)

Untersuchung von Athanasiadis [Ath74]

Vorgehen

- Bildung von Referenzgruppen nach Prüfern (schriftlich, mündlich)
- Zuordnung von Valenzen zu Prüfungsnoten nach NV in Referenzgruppen
- Untersuchung Unterschied Valenzen zu Noten an exemplarischen Prüfungsergebnissen

Resultate

- Existenz „milder“ und „strenger“ Prüfer
- Differenzen mittlerer Valenzen und Noten von 1 und mehr möglich (und realistisch)
- auch Umkehrung der ordinalen Relation
- Auswirkung auf Zulassung zur Promotion zu erwarten

Quellen und Verweise I

[Ath74] Th. Athanasiadis, *Zum System der Leistungsbewertung mit Zensuren*: *Ein Beitrag zur Quantifizierung von Leistungsvergleichen*, Statistische Studien, vol. 7, Franz Steiner Verlag, Wiesbaden, 1974.

[Fri72] R. Fricke, *Über Meßmodelle in der Schulleistungsdiagnostik*, Studien zur Lehrforschung, vol. 2, Pädagogischer Verlag Schwann, Düsseldorf, 1972.

[Wec76] H. Weck, *Leistungsermittlung und Leistungsbewertung im Unterricht*, Volk und Wissen Verlag, Berlin, 1976.

[Czi96] U. Czienskowski, *Wissenschaftliche Experimente*: *Planung Auswertung Interpretation*, Psychologische VerlagsUnion, Weinheim, 1996.

[Kai99] G. Kaiser, *Unterichtswirklichkeit in England und Deutschland*: *Vergleichende Untersuchungen am Beispiel des Mathematikunterrichts*, Deutscher Studien Verlag, Weinheim, 1999.

[Kul02] Kultusministerkonderenz, *Einheitliche Prüfungsanforderungen in der Abiturprüfung Deutsch* (2002).

[Sen10] Senatsverwaltung für Bildung, Wissenschaft und Forschung, *Ausführungsvorschriften über schulische Prüfungen* (2010).

Quellen und Verweise II

[KB10] O. Köller and J. Baumert, *Das Abitur - immer noch ein gültiger Indikator für die Studierfähigkeit?* (2010).