

Schiefte für geordnet-kategoriale Variablen

Ingo Klein

Überlingen am Ried, 3. Juni 2012

Lehrstuhl für Statistik und Ökonometrie



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Agenda

Fragestellung

Schiefeeigenschaft, -ordnung und Hypothesen

Schiefemaße

- Klassische Schiefemaße

- Lineare Schiefefunktionale nach Klein (2001)

- Drittes standardisiertes Moment

- Lineare Rangstatistiken

- Verallgemeinerter Likelihoodverhältnistest für geordnete Alternativen

Power

- Binomialverteilung

- Lehmann-Alternative

Anwendungen

- Links-Rechts-Selbsteinstufung in West- und Ostdeutschland

- Datensatz von Kemp & Conover (1973)

- Notenverteilung im WS 2011/12

- Empirische Studien aus Bhattacharya & Nandram (2011)

Fragestellung



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Geordnet-kategoriale Variablen

- Geordnet-kategoriale Variablen: k -Punkt-Ratingskalen.
- Kategorienzahl k (z.B. 5, 7, 10) klein im Vergleich zum Stichprobenumfang n (z.B. 1000, 2000).
- Kennzeichen geordnet-kategorialer Variablen:
 1. Ratingskalen sind Ordinalskalen: Keine sinnvolle Interpretation von Abständen der Beobachtungen.
 2. Bindungen (=Häufigkeitsverteilung der Kategorien).
 3. Wenn doch quantitativ, dann diskret.
- Mit wenigen Ausnahmen werden für geordnet-kategoriale Variablen in der Literatur nur Lokalisationsmaße (z.B. Median) diskutiert.

Schiefe vs. Lokalisation

- Behauptung: Für geordnet-kategoriale Variablen ist Schiefe informativer als Lokalisation.
- Lokalisationsmaße für geordnet-kategoriale Variablen können unter schwachen und plausiblen Invarianzforderungen nur von einer Beobachtung abhängen (Klein (1984), Klein (1994), Ovchinnikov (1996), Marichal (2002)).
- Ungeeignet: Mittelwerte.
Geeignet: Quantile (z.B. Median).
- Beispiel:

i	1	2	3	4	5
f_i	0.1	0.2	0.4	0.2	0.1
f'_i	0.3	0.19	0.19	0.16	0.16

Median beider Häufigkeitsverteilungen: 3,

$f = (f_1, \dots, f_5)$ ist symmetrisch um 3,

$f' = (f'_1, \dots, f'_5)$ ist rechtsschief.

Schiefe (allgemein) I

- Drei Aspekte der Schiefemessung:
 1. Schiefeigenschaft
 2. Schiefeordnung
 3. Schiefemaße und Schiefeteststatistiken
- Schiefeigenschaft: Wann ist eine Verteilung schief?
- Schiefeordnung: Wann ist eine Verteilung (links-) schiefer als eine andere?
- Schiefemaße: Abbildungen der Menge aller univariaten Verteilungen in die reellen Zahlen, so dass plausible Bedingungen (Axiome) eingehalten werden.
- Wichtigste plausible Forderung: Einhalten einer geeigneten Schiefeordnung.
D.h.: Wenn eine Häufigkeitsverteilung linksschiefer als eine andere ist, dann soll die Maßzahl kleinere Werte annehmen.

Schiefe (allgemein) II

- Forderung: Maßzahlwerte für
 1. Symmetrie: 0
 2. extreme Linksschiefe: -1
 3. extreme Rechtsschiefe: $+1$
- Forderung: Translations- und gelegentlich auch Skaleninvarianz.

Schiefe (Literatur für stetigen Fall) I

- Fechner (1897): Skalenparametersplit bei stetigen Verteilungen.
- Karl Pearson (1895): Gammaverteilung als Modell für schief-verteilte Daten und standardisierten Abstand von Mittelwert und Modus als Schiefemaß.
- Pareto (1897): Schiefe Verteilungen für ökonomische Fragestellungen (z.B. Einkommen).
- Charlier (1905), Edgeworth (1904). Drittes standardisiertes Moment als Schiefemaß.
- Bowley (1920): Quantilsbasiertes Schiefemaß.
- Yule (1912): Standardisierte Differenz zwischen Mittelwert und Median als Schiefemaß.

Schiefe (Literatur für stetigen Fall) II

- van Zwet (1964): Konvexe Ordnung zum Vergleich der Schiefe zweier Verteilungen.
- Oja (1981), MacGillivray (1986), Arnold & Groeneveld (1995): Weitere Schiefeordnungen und Axiomatisierungen von Schiefemaßen.
- Rayner et al. (1995), Tabor (2010), Doane & Seward (2011): Powervergleich diverser Test auf Schiefe.
- Tukey (1960), Azzalini (1986), Fernandez et al. (1995), Arrelano-Valle (2005): Transformation von symmetrisch in schief.
- Premaratne & Bera (2005): Score Test auf Asymmetrie.
- Kennzeichen: Stetige Verteilungen.

Schiefe (Nichtparametrische Statistik)

- Annahme: Stetige Verteilungsfunktion bedingt Fehlen von Bindungen.
- Lehmann (1953): Lehmann-Alternative $(F(x)^\lambda)$ ist eine Alternative von schiefen Verteilungen für $\lambda \neq 1$ und Wilcoxon signed rank-test lokal optimal für $pF + (1 - p)F^2$.
- Hájek & Sidak (1967): „signed rank test“ für die Nullhypothese der Symmetrie gegen Alternative der Lokalisationsverschiebung.
- Yanagimoto & Sibuya (1972): Alternativen wie „positively biasedness“ mit Schiefeigenschaft und Schiefeordnung ohne Bezug auf Quantilsfunktion.
- Cassart et al. (2007): Rank Tests auf „Edgeworth“ Asymmetrie.
- Ley & Pandavaine (2008): Rank Tests (inkl. Wilcoxon signed rank test) auf spezielle Ferreira & Steel Asymmetrie $(u(F(x)))$.
- Cassart et al. (2008): Neuer Rank Test auf (Fechner) Asymmetrie.

Schiefe (Diskrete Verteilungen)

- Hotelling & Solomon (1932): Grenzen für standardisierte Differenz von Mittelwert und Median auch für diskrete Verteilungen.
- Majindar (1962), Groeneveld (1991): Verfeinerung und Verallgemeinerung der Resultate von Hotelling & Solomon (1932).
- Groeneveld (1991): Schiefediskussion schließt als Beispiele Poisson-, Binomial-, geometrische und Zipf Verteilung ein.
- Rohatgi & Szekely (1989): Ungleichungen zwischen Schiefe und Kurtosis für alle Verteilungen.
- Pratt (1959), Chanda (1963), Conover (1972), Vorlickova (1972): Nichtparametrische Tests auf Symmetrie(?) bei (wenigen) Bindungen.
- von Hippel (2005): Vergleich Mittelwert, Median und Modus für die Entscheidung der Schieferichtung kann zu Fehlschlüssen führen.
- Beachte: Die Publikationen zum Testen betrachten als Alternative Lokalisationsshift und nicht Asymmetrie.
- Beachte: Schiefeigenschaft und Schiefeordnung werden für diskrete Verteilungen nicht betrachtet.

Schiefe (Geordnet-kategoriale Variablen I)

- Bowker (1948): Schiefe in Kontingenztabellen und χ^2 -Test.
- Bhapkar (1960), (1966): Testen linearer Restriktionen auf die Wahrscheinlichkeitsparameter einer Multinomialverteilung mittels Verallgemeinertem Likelihoodverhältnistest.
Kennzeichen: Omnibusalternative.
Resultate übertragbar auf stochastische Ordnung und Schiefe.
- Robertson (1978): Testen für und gegen Ordnungsrestriktionen in Multinomialverteilungen.
- Dykstra et al. (1995): Testen von Symmetrie gegen einseitige Alternativen (was Schiefeordnung impliziert).
- El Barmi & Dykstra (1998): Konkave ROC-Kurve von Sensitivität und Spezifität erzeugt konvexe Nebenbedingungen für ML-Schätzung.
- Klein (2001): Schiefemaße für geordnet-kategoriale Variablen (ohne Stochastik).
- Bhattacharya & Nandram (2011): Bayesianische Inferenz für multinomisch verteilte Grundgesamtheiten unter stochastischer Ordnung.

Schiefeeigenschaft, -ordnung und Hypothesen



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Klassische Schiefeordnungen I

- van Zwet (1964), Oja (1981), McGillivray (1986), Arnold & Groeneveld (1993)
- Konvexe Ordnung \preceq_1 : F linksschiefer als G , wenn $G^{-1}(F(x))$ ist konvex auf Träger I_F von F .
 D.h.: Sekanten liegen oberhalb von $G^{-1}(F(x))$.
 D.h.: $dG^{-1}(F(x))/dx$ ist monoton wachsend.
- „Star“-Ordnung mit Median als festem Referenzpunkt \preceq_2 : Seien m_F bzw. m_G Mediane von F bzw. G . F ist linksschiefer als G , wenn

$$\frac{G^{-1}(F(x)) - m_G}{x - m_F} \text{ monoton wachsend für } x \in I_F$$

D.h.: Sekantensteigung wächst mit x .

Klassische Schiefeordnungen II

- Doksum-Ordnung mit Median als Referenzpunkt \preceq_3 : Differenz zwischen $G^{-1}(F(x))$ und der Tangente $y = m_G - f(m_F)/g(m_G)(x - m_F)$ (im Punkt (m_F, m_G))

$$G^{-1}(F(x)) - \frac{m_F}{m_G}x + C$$

ist monoton abnehmend (zunehmend) für $x < m_F$ ($x > m_F$).

- \preceq_4 : Sekantensteigung ist für $x > m_F$ ($x < m_F$) größer (kleiner) als Tangentensteigung $f(m_F)/g(m_G)$.
- \preceq_5 : Sekantensteigung von $G^{-1}(F(x))$ zwischen p -Quantil von F und Median ist kleiner als Sekantensteigung zwischen $(1 - p)$ -Quantil und Median:

$$\frac{G^{-1}(F(x)) - m_g}{x - m_F} \leq \frac{G^{-1}(1 - F(x)) - m_G}{F^{-1}(1 - F(x)) - m_F} \quad x > m_F.$$

Klassische Schiefeordnungen III

- \preceq_6 : Seien X bzw. Y Zufallsvariablen zu F bzw. G . $(X - m_F)^+$ bzw. $(X - m_F)^-$ bezeichnen den positiven bzw. negativen Part von $X - m_F$ (analog für $Y - m_G$).
 X ist linksschiefer als Y , wenn $(X - m_F)^+$ stochastisch kleiner als $(Y - m_G)^+$ und $(X - m_F)^-$ stochastisch größer als $(Y - m_G)^-$.
- \preceq_7 : Analog für Lorenz-Ordnung

$$X \preceq_L Y : \iff \frac{\int_0^t F^{-1}(u) du}{\int_0^1 F^{-1}(u) du} \geq \frac{\int_0^t G^{-1}(u) du}{\int_0^1 G^{-1}(u) du} \quad t \in (0, 1).$$

- Starke oder gleichmäßige Ordnung \preceq_S : F ist linksschiefer als G , wenn

$$\begin{aligned} & F^{-1}(1 - u) - m_F - (m_F - F^{-1}(u)) \\ & \leq G^{-1}(1 - u) - m_G - (m_G - G^{-1}(u)) \end{aligned}$$

für alle $u \in (0, 0.5)$.

Klassische Schiefeordnungen IV

- Beziehungen:

$$\preceq_1 \Rightarrow \preceq_2 \Rightarrow \preceq_3 \Rightarrow \preceq_4 \Rightarrow \preceq_5$$

$$\preceq_6 \Rightarrow \preceq_s$$

$$\preceq_2 \Rightarrow \preceq_7 .$$

Formalismus geordnet kategorialer Variablen

- Ausprägungen: $u_1 < u_2 < \dots < u_k$
- Punktwahrscheinlichkeiten: $p_i = P(U = u_i), i = 1, 2, \dots, k$
- Kumulierte Punktwahrscheinlichkeiten: $P_i = \sum_{j=1}^i p_j, i = 1, 2, \dots, k,$
 $P_k = 1$
- Symmetrie:

$$p_i = p_{k+1-i}, \quad i = 1, 2, \dots, k$$

- Symmetrie (alternativ):

$$P_i + P_{k-i} = 1 \quad \text{bzw.} \quad Q_i = \frac{P_i + P_{k-i}}{2} = \frac{1}{2},$$

$$i = 1, 2, \dots, k - 1$$

Schiefeeigenschaft I

- Vorschläge von Klein (2001) und Dykstra et al. (1995)

1. $p = (p_1, \dots, p_k)$ ist linksschief (=negative Schiefe), wenn

$$Q_i \leq 1/2, \quad i = 1, 2, \dots, k - 1.$$

Extreme Linksschiefe: $p = (0, 0, \dots, 1)$.

Extreme Rechtsschiefe: $p = (1, 0, \dots, 0)$.

2. Strengere Schiefeeigenschaften nach Dykstra et al. (1995):

$$p_i \leq p_{k+1-i}, \quad i = 1, 2, \dots, l = \lfloor k/2 \rfloor$$

3. Noch strengere Schiefeeigenschaft:

$$p_1 \leq p_2 \leq \dots \leq p_k$$

Schiefeeigenschaft II

- Sei $k = 6$.

$$Q_1 \leq 1/2 \iff (p_1 + (1 - p_6))/2 \leq 1/2 \iff p_1 \leq p_6,$$

$$Q_2 \leq 1/2 \iff (p_1 + p_2 + 1 - p_5 - p_6)/2 \leq 1/2$$

$$\iff p_1 + p_2 \leq p_5 + p_6,$$

$$Q_3 \leq 1/2 \iff (p_1 + p_2 + p_3 + 1 - p_4 + p_5 + p_6)/2 \leq 1/2$$

$$\iff p_1 + p_2 + p_3 \leq p_4 + p_5 + p_6.$$

- $Q_4 = Q_2, Q_5 = Q_1$.
- Beispiel: $p = (0, 0, 0.5, 0, 0, 0.5)$

$$p_1 = 0 \leq p_6 = 0.5, \quad p_1 + p_2 = 0 \leq p_5 + p_6 = 0.5,$$

$$p_1 + p_2 + p_3 = 0.5 \leq p_4 + p_5 + p_6 = 0.5.$$

Aber: $p_3 = 0.5 > p_4 = 0$.

Schiefeeigenschaft III

- Lehmann-Alternative: Sei $p = (p_1, \dots, p_k)$ symmetrisch und $P'_i = P_i^\lambda$, $\lambda > 1$, dann ist

$$Q'_i \leq 0.5, \quad i = 1, 2, \dots, k - 1.$$

D.h. $p' = (p'_1, \dots, p'_k)$ mit $p'_i = P_i^\lambda - P_{i-1}^\lambda$, $i = 1, 2, \dots, k$, wobei $P_0 = 0$, ist linksschief.

Schiefeordnung I

- $p' = (p'_1, \dots, p'_k)$ mit $P'_i = \sum_{j=1}^i p'_j$ und $Q'_i = (P'_i + P'_{k-i})/2$,
 $i = 1, 2, \dots, k - 1$
- Schiefeordnung nach Klein (2001):
 1. p ist linksschiefer als p' , wenn

$$Q_i \leq Q'_i, \quad i = 1, 2, \dots, k - 1.$$

2. Hinreichende Bedingung: Stochastische Ordnung

$$P_i \leq P'_i, \quad i = 1, 2, \dots, k - 1$$

oder

$$pp'_i \leq pp'_{k+1-i}, \quad i = 1, 2, \dots, l = [k/2]$$

mit

$$pp'_i = (p_i + p'_{k+1-i})/2, \quad i = 1, 2, \dots, l = [k/2]$$

Schiefeordnung II

- Sei $k = 6$.

$$\begin{aligned}
 Q_1 \leq Q'_1 &\iff (p_1 + 1 - p_6)/2 \leq (p'_1 + 1 - p'_6)/2 \\
 &\iff (p_1 + p'_6)/2 \leq (p'_1 + p_6)/2 \iff pp_1 \leq pp_6, \\
 Q_2 \leq Q'_2 &\iff (p_1 + p_2 + 1 - p_5 - p_6)/2 \leq (p'_1 + p'_2 + 1 - p'_5 - p'_6)/2 \\
 &\iff (p_1 + p'_6)/2 + (p_2 + p'_5)/2 \leq (p'_1 + p_6)/2 + (p'_2 + p_5)/2 \\
 &\iff pp'_1 + pp'_2 \leq pp'_5 + pp'_6, \\
 Q_3 \leq Q'_3 &\iff (p_1 + p'_6)/2 + (p_2 + p'_5)/2 + (p_3 + p'_4)/2 \\
 &\leq (p'_1 + p_6)/2 + (p'_2 + p_5)/2 + (p'_3 + p_4)/2 \\
 &\iff pp'_1 + pp'_2 + pp'_3 \leq pp'_4 + pp'_5 + pp'_6.
 \end{aligned}$$

Schiefeordnung III

- $pp'_i \leq pp'_{k+1-i}, i = 1, 2, \dots, l = \lfloor k/2 \rfloor$ ist hinreichend, aber nicht notwendig für „ $Q_i \leq Q'_i$ “:
 $p = (0, 0, 0.5, 0, 0, 0.5), p' = (0.5, 0, 0, 0.5, 0, 0)$ liefert
 $pp' = (0, 0, 0.5, 0, 0, 0.5)$ mit $pp'_3 > pp'_4$ und
 $Q = (0.25, 0.25, 0.5, 0.25, 0.25) < Q' = (0.75, 0.75, 0.5, 0.75, 0.75)$.
- Stochastische Ordnung ist hinreichend, aber nicht notwendig für „ $Q_i \leq Q'_i$ “:
 Sei
 $(P_1, P_2, P_3) = (0.1, 0.3, 0.7)$ und $(P'_1, P'_2, P'_3) = (0.15, 0.35, 0.64)$.
 mit
 $Q = (0.4, 0.3, 0.4) < Q' = (0.405, 0.35, 0.405)$
 und $P'_3 < P_3$.

Hypothesen I

- Nullhypothese: Symmetrie im Einstichprobenfall

$$H_0 : Q_i = 1/2, \quad i = 1, 2, \dots, l = [k/2].$$

- Nullhypothese: Identische Schiefe im Zweistichprobenfall

$$H'_0 : Q_i = Q'_i \quad i = 1, 2, \dots, l = [k/2]$$

oder äquivalent

$$H'_0 : pp'_i = pp'_{k+1-i}, \quad i = 1, 2, \dots, l = [k/2].$$

D.h. Nullhypothese ist Symmetrie des synthetischen Vektors von Wahrscheinlichkeiten $q = (q_1, \dots, q_k)$.

Hypothesen II

- Zweiseitige Alternativen: \bar{H}_0 bzw. \bar{H}'_0 .
- Einseitige Alternativen: $H_1 \setminus H_0$ bzw. $H'_1 \setminus H'_0$ mit

$$H_1 : Q_i \leq 1/2 \text{ for } i = 1, 2, \dots, l$$

$$H'_1 : Q_i \leq Q'_i \text{ for } i = 1, 2, \dots, l.$$

Hypothesen III

- Dykstra et al. (1995): $H_2 \setminus H_0$ bzw. $H'_2 \setminus H'_0$ mit

$$H_2 : p_i \leq p_{k+1-i} \text{ for } i = 1, 2, \dots, l$$

$$H'_2 : pp'_i \leq pp'_{k+1-i} \text{ for } i = 1, 2, \dots, l$$

Schiefemaße



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Klassische Schiefemaße

- Pearsons Schiefemaß:

$$\frac{E(X) - F^{-1}(1/2)}{\sqrt{\text{VAR}(X)}}$$

- Drittes standardisiertes Moment.
- Quantilsbasierte Schiefemaße:

$$\frac{F^{-1}(1 - u) + F^{-1}(u) - 2F^{-1}(1/2)}{F^{-1}(1 - u) - F^{-1}(u)} \quad \text{für } u \in (0, 1) \text{ fixiert}$$

- Problem: Sinnvolle Interpretation des Abstandes der geordneten Kategorien

Schiefefunktionale (Formaler Hintergrund) I

- $u = (u_1, \dots, u_k)$, $p = (p_1, \dots, p_k)$ mit $p_i = P(U = u_i)$, $i = 1, 2, \dots, k$,
 $u' = (u'_1, \dots, u'_k)$, $p' = (p'_1, \dots, p'_k)$ mit $p'_i = P(U' = u'_i)$, $i = 1, 2, \dots, k$,
 $\Delta_k = \{p = (p_1, \dots, p_k) \in [0, 1]^k \mid \sum_{i=1}^k p_i = 1\}$,
 $G = \{g^k \mid g \text{ streng monoton zunehmend auf } \mathbb{R}\}$ Transformationsgruppe.

Schiefefunktionale (Formaler Hintergrund) II

- Ergebnis von Klein (1994):

Sei $S : \mathbb{R}^k \times \Delta_k \rightarrow \mathbb{R}$. Wenn

$$S(u, p) = S(u', p') \iff S(g(u), p) = S(g(u'), p), \quad u \in \mathbb{R}^k, \quad g \in G$$

und für alle $c \in \mathbb{R}$

$$S(u + c, p) = S(u, p), \quad u \in \mathbb{R}^k,$$

dann gilt

$$S(u, p) = S(g(u), p), \quad u \in \mathbb{R}^k, \quad g \in G.$$

- Konsequenz: $P = (P_1, \dots, P_k)$ mit $P_i = \sum_{j=1}^i p_j$, $i = 1, 2, \dots, k - 1$ ist maximal-invariant, so dass

$$S(u, p) = S(p).$$

- Schiefemaße sollten translationsinvariant sein.
- D.h. klassische Schiefemaße fallen weg.

Schiefefunktionale (Formaler Hintergrund) III

- Vorgabe einer geeigneten Schiefeordnung \preceq .
- Definition: $S : \Delta_k \rightarrow \mathbf{R}$ ist ein Schiefemaß : \iff
 1. Seien $p_1 = (p_1, p_2, \dots, p_k) \in \Delta_k$ und $p' = (p_k, p_{k-1}, \dots, p_1)$. Dann gilt $S(p) = -S(p')$.
 2. Seien $p, p' \in \Delta_k$ mit $p \preceq p'$. Dann gilt: $S(p) \leq S(p')$.

Lineare Schiefefunktionale (Definition)

- Lineare Schiefefunktionale:

$$S^b(p) = \frac{1}{k-1} \sum_{i=1}^{k-1} b(Q_i)$$

mit erzeugender Funktion b .

- Eigenschaften von b :
 1. $b : [0, 1] \rightarrow \mathbb{R}$,
 2. b stetig auf $(0, 1)$,
 3. b antisymmetrisch, d.h. $b(p) = -b(1 - p)$ for $p \in (0, 1)$,
 4. b beschränkt und streng monoton zunehmend auf $[0, 1]$.
- Bessere Interpretierbarkeit für extreme Schiefe, wenn $b(0) = -1$ und $b(1) = 1$.
- Klein (2001): Schiefeigenschaft „ $Q_i \leq 1/2$ “ und Schiefeordnung „ $Q_i \leq Q'_i$ “ wird eingehalten.

Lineare Schiefefunktionale (Beispiele)

- Kiesl (2005): Positions-basiertes Streuungsmaß.
Übertragung auf Schiefe:

$$b_1(p) = 2(p - 1/2), \quad p \in [0, 1].$$

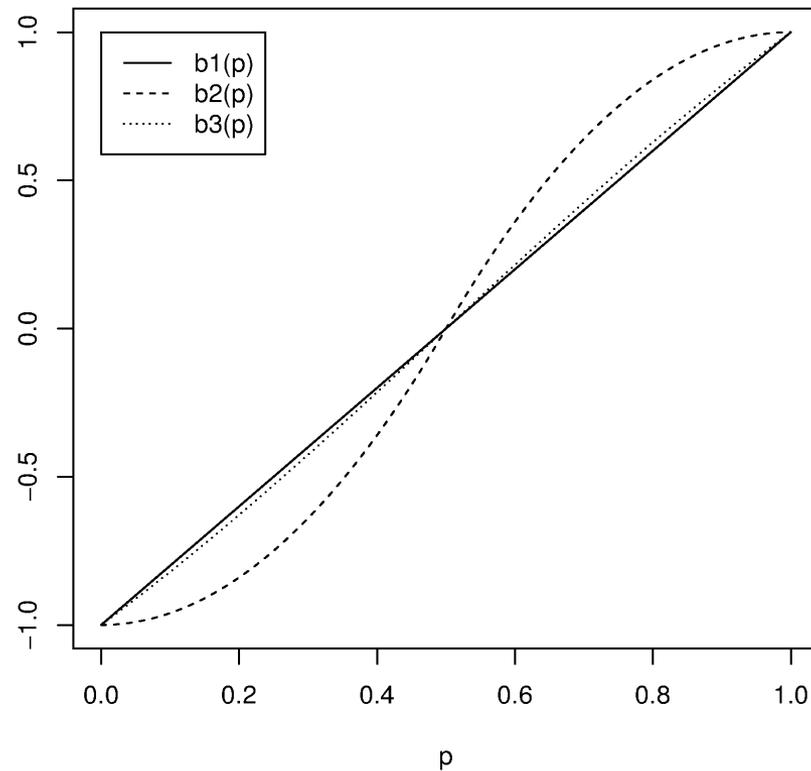
- Gini (1955):

$$b_2(p) = \begin{cases} 4(p^2 - 1/4) & \text{for } p \leq 1/2 \\ 4(p - 1/2)(3/2 - p) & \text{for } p > 1/2 \end{cases}$$

- Premaratne & Bera (2005), p. 174:

$$b_3(p) = -\frac{\tan^{-1}(p - 1/2)}{\tan^{-1}(1/2)}.$$

Lineare Schiefefunktionale (Beispiele)



Lineare Schiefefunktionale und Schiefeigenschaft

„ $p_i < p_{k+1-i}$ “

- Sei $k = 2l = 6$.

$$\begin{aligned}
 \sum_{i=1}^3 Q_i &= \frac{p_1 - p_6 + 1}{2} + \frac{p_1 + p_2 - p_5 - p_6 + 1}{2} \\
 &\quad + \frac{p_1 + p_2 + p_3 - p_4 - p_5 - p_6 + 1}{2} \\
 &= \frac{3}{2}(p_1 - p_6) + (p_2 - p_5) + \frac{1}{2}(p_3 - p_4) + \frac{3}{2}.
 \end{aligned}$$

Lineare Schiefefunktionale und Mittelwerte I

- Arithmetische Mittel können via Integration der Verteilungsfunktion bzw. der Survivor Funktion berechnet werden.
- Sei U eine Zufallsvariable mit Träger $\{i_1, i_1 + 1, \dots, i_1 + (k - 1)\}$, $i_1 \in \mathbb{G}$, $k \in \mathbb{N}$. Seien $p_i = P(U = i_1 + (i - 1))$, $i = 1, 2, \dots, k$. Dann ist $\mu = E(U) = \sum_{i=1}^k (i_1 + (i - 1))p_i$. Es gilt:

$$S^{b_1}(p) = \frac{2}{k-1} \left(i_1 + \frac{k-1}{2} - \mu \right).$$

- D.h.: Je größer der Mittelwert, desto linksschiefer.
- Beachte: Nur formale Beziehung, die gültig ist, wenn U quantitativ.

Lineare Schiefefunktionale und Mittelwerte II

$$\begin{aligned}\mu &= E(U) = i_1 p_1 + (i_1 + 1)p_2 + \dots + (i_1 + (k - 1))p_k \\ &= i_1 + (1 - p_1) + \dots + (1 - \sum_{i=1}^{k-1} p_i) = i_1 + (k - 1) - \sum_{i=1}^{k-1} P_i\end{aligned}$$

mit $P_i = \sum_{j=1}^i p_j$, $i = 1, 2, \dots, k - 1$.

D.h.:

$$\sum_{i=1}^{k-1} P_i = i_1 + (k - 1) - \mu.$$

S^{b_1} ist Funktion von $\sum_{i=1}^{k-1} P_i$ wegen

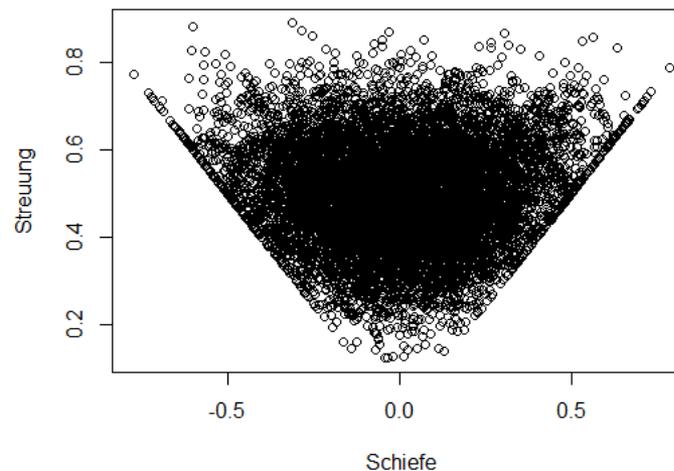
$$\begin{aligned}S^{b_1} &= \frac{2}{k-1} \sum_{i=1}^{k-1} \frac{P_i + P_{k-1}}{2} - 1 = \frac{2}{k-1} \sum_{i=1}^{k-1} Q_i - 1 \\ &= \frac{2}{k-1} (i_1 + (k-1) - \mu) - 1 = \frac{2}{k-1} \left(i_1 + \frac{k-1}{2} - \mu \right).\end{aligned}$$

Lineare Streuungs- und Schiefefunktionale

- Korrespondierendes lineares Streuungsfunktional für $b(p) = 2(p - 1/2)$:

$$S = \frac{2}{k-1} \sum_{i=1}^{k-1} a(P_i) \quad \text{mit } a(p) = |p - 1/2|, p \in [0, 1].$$

- Zusammenhang:



Lineare Schiefestatistik (Definition)

- Bezeichnungen:
 - Kategorienhäufigkeiten: $f = (n_1/n, \dots, n_k/n)$.
 - Kumulierte Kategorienhäufigkeiten: $F = (F_1, \dots, F_{k-1})$.
 - $H = (H_1, \dots, H_{k-1})$ mit $H_i = (F_i + F_{k-i})/2, i = 1, 2, \dots, k - 1$.
- Lineare Schiefestatistik:

$$S^b(f) = \frac{1}{k-1} \sum_{i=1}^{k-1} b(H_i)$$

- Exakte Verteilung: Via Totalenumeration bzw. Simulation durch Ziehen aus einer multinomischen (k -dimensionalen) Verteilung mit den Parametervektor $p \in \Delta_k$ und Stichprobenumfang n .

Lineare Schiefestatistik (Asymptotik) I

- $N = (N_1, \dots, N_k)$ multinomial verteilt
mit $p = (p_1, \dots, p_k)$,
 $n = \sum_{i=1}^k N_i = n$, $\sum_{i=1}^k p_i = 1$.

$$F_i = \sum_{j=1}^i N_j / n,$$

$$P_i = \sum_{j=1}^i p_j$$

$$H_i = (F_i + F_{k-1}) / 2,$$

$$Q_i = (P_i + P_{k-i}) / 2$$

Lineare Schiefestatistik (Asymptotik) II

- Wenn $b : [0, 1] \rightarrow [-1, 1]$ differenzierbar auf $(0, 1)$ und $b'(Q_i) \neq 0$ für $i = 1, 2, \dots, k - 1$, dann ist $S_k(N/n) = \sum_{i=1}^{k-1} b(H_i)$ asymptotisch normalverteilt mit Mittelwert $\sum_{i=1}^{k-1} b(P_i)$ und Varianz

$$\sigma_n^2 = \frac{1}{4n} \left(\sum_{i=1}^{k-1} P_i(1 - P_i)b'(Q_i)^2 + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k-1} P_i(1 - P_j)b'(Q_i)b'(Q_j) \right).$$

- Konsistenter Schätzer für asymptotische Varianz:

$$\hat{\sigma}_n^2 = \frac{1}{4n} \left(\sum_{i=1}^{k-1} F_i(1 - F_i)b'(H_i)^2 + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k-1} F_i(1 - F_j)b'(H_i)b'(H_j) \right).$$

Lineare Schiefestatistik (Asymptotik) III

Wegen des Grenzwertsatzes von deMoivre & Laplace ist $(F_1, F_2, \dots, F_{k-1})$ asymptotisch normal mit Mittelwert $(P_1, P_2, \dots, P_{k-1})$ und Kovarianzmatrix $1/n\Sigma$ mit Elementen

$$\sigma_{ij} = \text{Cov}(F_i, F_j) = \begin{cases} P_i(1 - P_i)/n & \text{für } i = j \\ P_i(1 - P_j)/n & \text{für } i < j \\ P_j(1 - P_i)/n & \text{für } i > j \end{cases}$$

für $i, j = 1, 2, \dots, k - 1$ (siehe z.B. Kiesl (2003), S. 98).

$S^b(N) = \frac{1}{k-1} \sum_{i=1}^{k-1} b(H_i)$ ist eine skalarwertige, differenzierbare Funktion von $(F_1, F_2, \dots, F_{k-1})$. Wenn das Differential von $S^b(F)$ an der Stelle $\mu = (P_1, \dots, P_{k-1})$ nicht verschwindet, ist $S^b(F)$ auch asymptotisch normal mit Mittelwert $\sum_{i=1}^{k-1} b(P_i)$ and Varianz

$$\frac{1}{n} \left(\frac{\partial S^b(F)}{\partial F_1}, \dots, \frac{\partial S^b(F)}{\partial F_{k-1}} \right) \Sigma \left(\frac{\partial S^b(F)}{\partial F_1}, \dots, \frac{\partial S^b(F)}{\partial F_{k-1}} \right)'$$

(siehe Serfling (1980), p. 124). Mit

$$\frac{\partial S^b(F)}{\partial F_i} = b'(H_i) \frac{\partial H_i}{\partial F_i} = b'(H_i) \frac{1}{2}$$

für $i = 1, 2, \dots, k - 1$ folgt die Varianz der asymptotischen Verteilung.

Lineare Schiefestatistik (Asymptotik) IV

- Spezialfall Positions-basierte Statistik:

$$\sigma_n^2 = \frac{1}{n} \left(\sum_{i=1}^{k-1} P_i(1 - P_i) + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k-1} P_i(1 - P_j) \right)$$

- Spezialfall Gini: Mit

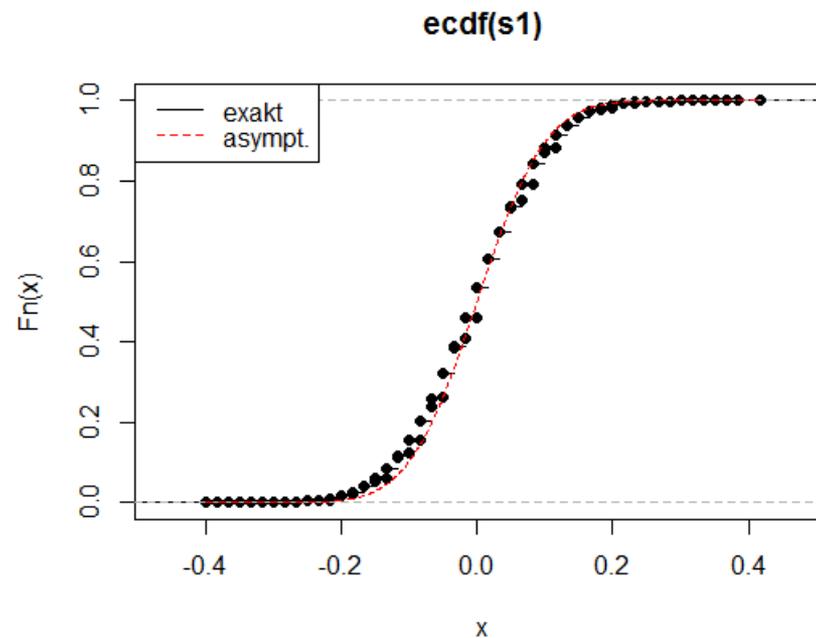
$$b'_2(p) = \begin{cases} 8p & \text{for } p \in (0, 1/2) \\ 8(1 - p) & \text{for } p \in (1/2, 1) \end{cases}$$

ist

$$\sigma_n^2 = \frac{1}{4n} \left(\sum_{i=1}^{k-1} P_i(1 - P_i)b'_2(Q_i)^2 + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k-1} P_i(1 - P_j)b'_2(Q_i)b'_2(Q_j) \right)$$

Lineare Schiefestatistik (Asymptotik) V

- Exakte versus asymptotische Verteilung für $k = 7$ und $n = 20$



Lineares Schiefefunktional im Zweistichprobenfall I

- Seien $p = (p_1, \dots, p_k)$, $p' = (p'_1, \dots, p'_k)$ mit $P_i = \sum_{j=1}^i p_j$, $P'_i = \sum_{j=1}^i p'_j$,
 $Q_i = (P_i + P_{k-i})/2$, $Q'_i = (P'_i + P'_{k-i})/2$, $i = 1, 2, \dots, k-1$.
- Dann

$$pp' = (p + (p'_k, \dots, p'_1))/2,$$

so dass

$$PP'_i = \sum_{j=1}^i pp'_j = \frac{1}{2} \left(\sum_{j=1}^i p_j + p'_{k+1-j} \right) = \frac{1}{2} (P_i + (1 - P'_{k-i}))$$

und

$$QQ'_i = \frac{1}{2} (PP'_i + PP'_{k-1-i}) = \frac{1}{4} (P_i + 1 - P'_{k-i} + P_{k-i} + 1 - P'_i) = \frac{1}{2} (Q_i - Q'_i + 1).$$

D.h.

$$QQ'_i \leq 0.5 \iff Q_i \leq Q'_i, \quad i = 1, 2, \dots, k-1.$$

Lineares Schiefefunktional im Zweistichprobenfall II

- Lineares Schiefefunktional im Zweistichprobenfall:

$$S(p, p') = S^b(pp') = \frac{1}{k-1} \sum_{i=1}^{k-1} b(Q_i Q'_i) = \frac{1}{k-1} b \left(\frac{1}{2} (Q_i - Q'_i + 1) \right)$$

mit $b(\cdot)$ wie oben.

- Linearer Spezialfall $b_1(p) = 2(p - 1/2)$:

$$\begin{aligned} S^{b_1}(pp') &= \frac{1}{k-1} \sum_{i=1}^{k-1} 2 \left(\frac{1}{2} (Q_i - Q'_i + 1) - 1/2 \right) \\ &= \frac{1}{k-1} \sum_{i=1}^{k-1} 2(Q_i - 1/2 - (Q'_i - 1/2)) = S^{b_1}(p) - S^{b_1}(p'). \end{aligned}$$

D.h. im linearen Fall ergibt sich Differenz der einzelnen Schiefefunktionale.

Lineare Schiefestatistik (Asymptotik im Zweistichprobenfall) I

- $N = (N_1, \dots, N_k)$ bzw. $N' = (N'_1, \dots, N'_k)$ multinomial verteilt mit $p = (p_1, \dots, p_k)$ bzw. $p' = (p'_1, \dots, p'_k)$,
 $n_1 = \sum_{i=1}^k N_i$, $\sum_{i=1}^k p_i = 1$ bzw. $n_2 = \sum_{i=1}^k N'_i$, $\sum_{i=1}^k p'_i = 1$.

$$f_i = F_i/n_1, f'_i = F'_i/n_2, F_i = \sum_{j=1}^i N_j/n_1, F'_i = \sum_{j=1}^i N'_j/n_2$$

$$P_i = \sum_{j=1}^i p_j, P'_i = \sum_{j=1}^i p'_j$$

$$H_i = (F_i + F_{k-1})/2, H'_i = (F'_i + F'_{k-1})/2$$

$$Q_i = (P_i + P_{k-i})/2, Q'_i = (P'_i + P'_{k-i})/2$$

Lineare Schiefestatistik (Asymptotik im Zweistichprobenfall) II

- Fall zweier unabhängiger Stichproben: N und N' bzw. F und F' sind stochastisch unabhängig.
D.h. (F, F') ist asymptotisch normalverteilt mit Mittelwertvektor (P, P') und Varianzkovarianzmatrix $\text{diag}(\Sigma, \Sigma')$ mit Elementen von Σ die

$$\sigma_{ij} = \text{Cov}(F_i, F_j) = \begin{cases} P_i(1 - P_i)/n_1 & \text{für } i = j \\ P_i(1 - P_j)/n_1 & \text{für } i < j \\ P_j(1 - P_i)/n_1 & \text{für } i > j \end{cases}$$

und Elementen von Σ'

$$\sigma'_{ij} = \text{Cov}(F'_i, F'_j) = \begin{cases} P'_i(1 - P'_i)/n_2 & \text{für } i = j \\ P'_i(1 - P'_j)/n_2 & \text{für } i < j \\ P'_j(1 - P'_i)/n_2 & \text{für } i > j \end{cases}$$

für $i, j = 1, 2, \dots, k - 1$.

- $S^b(f, f') = \frac{1}{k-1} \sum_{i=1}^{k-1} b(1/2(H_i - H'_i + 1))$ ist eine skalarwertige, differenzierbare Funktion von $(F_1, F_2, \dots, F_{k-1}, F'_1, F'_2, \dots, F'_{k-1})$.
Wenn das Differential von $S^b(N/n_1, N'/n_2)$ an der Stelle $\mu = (P_1, \dots, P_{k-1}, P'_1, \dots, P'_{k-1})$ nicht verschwindet, ist $S^b(f, f')$ auch asymptotisch normal mit Mittelwert $1/(k-1) \sum_{i=1}^{k-1} b(P_i)$ and Varianz

$$\frac{1}{(k_1)^2} \left(\left(\frac{\partial S^b(f, f')}{\partial F_1}, \dots, \frac{\partial S^b(f, f')}{\partial F_{k-1}} \right) \Sigma_1 \left(\frac{\partial S^b(f, f')}{\partial F_1}, \dots, \frac{\partial S^b(f, f')}{\partial F_{k-1}} \right)' + \left(\frac{\partial S^b(f, f')}{\partial F'_1}, \dots, \frac{\partial S^b(f, f')}{\partial F'_{k-1}} \right) \Sigma_2 \left(\frac{\partial S^b(f, f')}{\partial F'_1}, \dots, \frac{\partial S^b(f, f')}{\partial F'_{k-1}} \right)' \right)$$

Mit

$$\frac{\partial S^b(f, f')}{\partial F_i} = \frac{\partial S^b(f, f')}{\partial F'_i} = b'(1/2(H_i - H'_i + 1)) \frac{1}{2}$$

für $i = 1, 2, \dots, k - 1$ folgt die Varianz der asymptotischen Verteilung.

Lineare Schiefestatistik (Asymptotik im Zweistichprobenfall) III

- Linearer Spezialfall $b = 1(p)2(p - 1/2)$.

$$S^{b_1}(f) - S^{b_2}(f') \stackrel{asy}{\sim} N(S^{b_1}(p) - S^{b_1}(p'), \sigma_{n_1}^2 + \sigma_{n_2}^2)$$

mit

$$\sigma_{n_i}^2 = \frac{1}{4n_{i'}} \left(\sum_{i=1}^{k-1} P_{i'i} (1 - P_{i'i}) b'(Q_{i'i})^2 + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k-1} P_{i'i} (1 - P_{i'j}) b'(Q_{i'i}) b'(Q_{i'j}) \right).$$

- σ^2 kann konsistent geschätzt werden mittels $F_{i',j} = \sum_{j=1}^{i'} f_{i',j}$ und $H_{i',j} = (F_{i',j} + F_{i',k-i})/2$, $i = 1, 2, \dots, k - 1$, $i' = 1, 2$.

Drittes standardisiertes Moment (Definition)

- Diskrete quantitative Zufallsvariablen sind auch geordnet-kategoriale Zufallsvariablen.
- Sei U eine diskrete Zufallsvariable mit Träger $u_1 < u_2 < \dots < u_k$ und $p_i = P(U = u_i)$, $i = 1, 2, \dots, k$.
- Drittes standardisiertes Moment:

$$SK(u, p) = \frac{\sum_{i=1}^k (u_i - \mu)^3 p_i}{(\sigma^2)^{3/2}}$$

mit

$$\mu = \sum_{i=1}^k u_i p_i, \quad \sigma^2 = \sum_{i=1}^k (u_i - \bar{u})^2 p_i$$

- $\sqrt{b_1}$ -Test (lt. Karl Pearson) ist Test auf Schiefe mittels $SK(u, f)$, wenn $f = (f_1, \dots, f_k)$ die relativen Stichprobenhäufigkeiten sind.
- Exakte Verteilung: Totalenumeration und Simulation.

Drittes standardisiertes Moment (Definition)

- Klein (2001) gibt Beispiel, dass SK die Schiefeordnung nicht einhält.
- Weitere Überlegung: $p = (0, 0, 0, 0, 0.5, 0.5)$ ergibt $SK = 0$, da extreme Nullen in die Berchnung nicht einfließen, obwohl

$$Q = (0.25, 0, 0, 0, 0.25) \text{ mit } Q_i < 0.5, i = 1, 2, 3, 4, 5,$$

was Linksschiefe anzeigt.

Drittes standardisiertes Moment (Asymptotik) I

- $SK(u, f)$ ist asymptotisch normalverteilt mit Mittelwert $SK(u, p)$ und Varianz $1/n \nabla' \Sigma \nabla$, wobei ∇ der Gradientenvektor mit Elementen

$$\frac{\partial SK(u, f)}{\partial p_i} = -3 \frac{u_i}{\sqrt{S^2(u, p)}} + \frac{(u_i - M_1(u, p))^3}{S^2(u, p)^{3/2}} - \frac{3}{2} SK(u, p) \frac{u_i^2 - 2M_1(u, p)u_i}{S^2(u, p)},$$

$i = 1, 2, \dots, k$ und Σ die Kovarianzmatrix mit Elementen

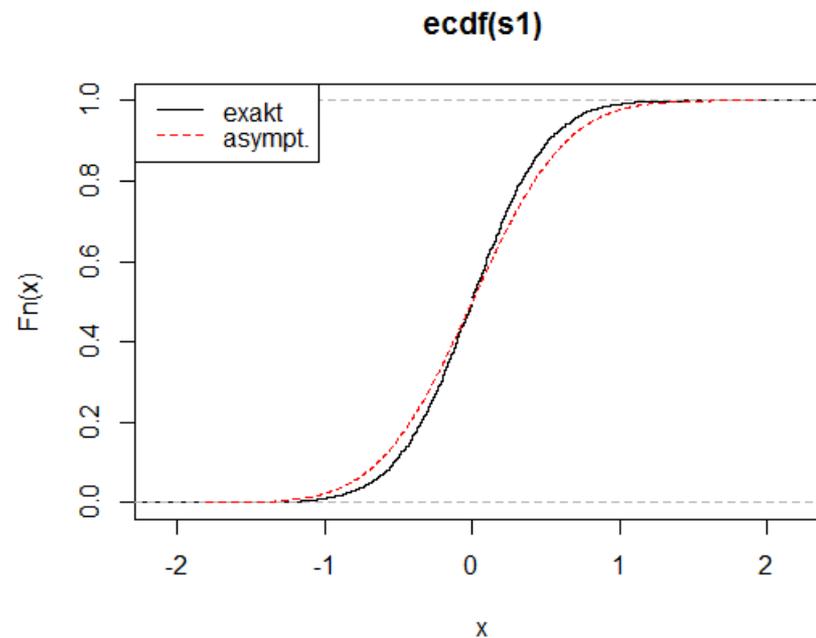
$$\sigma_{ij} = Cov(f_i, f_j) = \begin{cases} p_i(1 - p_i)/n & \text{für } i = j \\ -p_i p_j / n & \text{für } i \neq j \end{cases}$$

sind.

- ∇ und σ_{ij} können konsistent durch das Stichprobenanaloga geschätzt werden.

Drittes standardisiertes Moment (Asymptotik) II

- Exakte versus asymptotische Verteilung für $k = 7$ und $n = 20$



Drittes standardisiertes Stichprobenmoment (Asymptotik im Zweistichprobenfall)

- $SK(u, f_1) - SK(u, f_2)$ ist asymptotisch normalverteilt mit Mittelwert $SK(u, p_1) - SK(u, p_2)$ und Kovarianzmatrix $\nabla'_1 \Sigma_1 \nabla_1 + \nabla'_2 \Sigma_2 \nabla_2$.
- Elemente von $\nabla'_{i'}$:

$$\frac{\partial SK(u, p_{i'})}{\partial p_{i'j}} = -3 \frac{u_j}{\sqrt{S^2(u, p_{i'})}} + \frac{(u_j - M_1(u, p_{i'}))^3}{S^2(u, p_{i'})^{3/2}} - \frac{3}{2} SK(u, p_{i'}) \frac{u_j^2 - 2M_1(u, p_{i'})u_j}{S^2(u, p_{i'})},$$

$i = 1, 2, \dots, k$ and $i' = 1, 2$.

- Elemente von $\Sigma_{i'}$:

$$\sigma_{ij} = \text{Cov}(f_{i'i}, f_{i'j}) = \begin{cases} p_{i'i}(1 - p_{i'i})/n & \text{für } i = j \\ p_{i'i}(1 - p_{i'j})/n & \text{für } i \neq j \end{cases}$$

für $i, j = 1, 2, \dots, k$ und $i' = 1, 2$.

Symmetrie in der nichtparametrischen Statistik

- Signed rank tests: Test für Lokalisationsalternativen.
- Hajék & Šidák (1967): Überschrift von Section III.5.1 ist "Tests of Symmetry".
- Aber: Diskutiert werden Tests für Lokalisationsshifts in Familien symmetrischer Verteilungen.
- Yanagimoto & Sibuya (1972), p. 423:

"The purpose of this paper is to make clear the notion of a 'positively biased' one-dimensional random variable as an alternative to 'symmetric about zero'. This notion is useful to make more precise statements on the test of symmetry than discussed in previous publications"

Signed rank tests

- Seien X_1, \dots, X_n eine einfache Zufallsstichprobe aus Verteilung mit Verteilungsfunktion $F(\cdot)$, $|X|^{(1)}, \dots, |X|^{(n)}$ Ordnungsstatistiken der Absolutbeträge von X_i , $i = 1, 2, \dots, n$ und $Z_i = 1$ ($= -1$) wenn $|X|^{(i)}$ zu einem positiven (negativen) X_i gehört für $i = 1, 2, \dots, n$.

Dann definiert

$$S^a = \sum_{i=1}^n a(i)Z_i$$

eine signed rank test-Statistik mit scores $a(i)$, $i = 1, 2, \dots, n$.

- Yanagimoto & Sibuya (1972): S^a kann für einen unverzerrten Test der Nullhypothese (Symmetrie) gegen die Alternative $H_1 \setminus H_0$ benutzt werden, wenn die Scores nicht abnehmend sind.
- Spezielle Scores:
 - Median scores ($a(i) = 1$),
 - Wilcoxon scores ($a(i) = i$),
 - Expected normal scores ($E(\Phi^{-1}(U^{(i)}))$),
 - Normal scores ($a(i) = \Phi^{-1}(i/(n+1))$).

Bindungen in der nichtparametrischen Statistik

- Annahme von Yanagimoto & Sibuya (1972): F ist stetig.
- Große Stichproben (Uumfang n) für geordnet-kategoriale Variablen mit k Kategorien, so dass $k \ll n$, führen zu zahlreichen Bindungen.
- Conover (1973), S. 1105:

"Because of the difficulty in working with ranks in the presence of tied observations, most of the theoretical research involving rank test begins with the assumption that all distribution functions are continuous. In applied work, however, ties do occur, and experimental workers are sometimes hesitant to apply rank tests in the presence of many ties, for fear that the rank tests are no longer valid."

Signed rank test-Statistik und Midranks (gerader Fall) I

- Seien $u_1 < \dots < u_k$ und $k = 2l$, $l \in \mathbb{N}$, $p_i = P(U = u_i)$, $i = 1, 2, \dots, k$.
Symmetrie:

$$p_{k/2+1-i} = p_{k/2+i}, \quad i = 1, 2, \dots, l.$$

- Setze $pp_i = p_{k/2+1-i} + p_{k/2+i}$, $i = 1, 2, \dots, l$.
- Berechne kumulierte Wahrscheinlichkeiten: $PP_i = \sum_{j=1}^i pp_j$,
 $i = 1, 2, \dots, l$.
- "Mid-ranks": $\overline{PP}_1 = 1/2 PP_1$ und

$$\overline{PP}_i = \frac{1}{2} (PP_{i-1} + PP_i), \quad i = 2, \dots, l.$$

- "Linear rank functional":

$$S^a(p) = \sum_{i=1}^l a(\overline{PP}_i) p_{k/2+i}.$$

Signed rank test-Statistik und Midranks (gerader Fall) II

- $l = 3$
- $p = (0.1, 0.2, 0.2, 0.3, 0.1, 0.1)$
- $pp = (0.2 + 0.3, 0.2 + 0.1, 0.1 + 0.1) = (0.5, 0.3, 0.2)$
- $PP = (0.5, 0.8, 1)$
- $\overline{PP} = (0.5/2, (0.5 + 0.8)/2, (0.8 + 1)/2) = (0.25, 0.65, 0.9)$
- $a(p) = p: S^a(p) = 0.25 \cdot 0.3 + 0.65 \cdot 0.1 + 0.9 \cdot 0.1 = 0.23$

Signed rank test-Statistik und Midranks (gerader Fall) III

- Symmetrie für $l = 3$: $p = (p_6, p_5, p_4, p_4, p_5, p_6)$.
- $pp = (2p_4, 2p_5, 2p_6)$.
- $PP = (2p_4, 2(p_4 + p_5), 1)$.
- $\overline{PP} = (p_4, 2p_4 + p_5, 2(p_4 + p_5) + p_6)$.
- $a(p)$ allgemein;

$$S^a(p) = a(p_4)p_4 + a(2p_4 + p_5)p_5 + a(2(p_4 + p_5) + p_6)p_6.$$
- $a(p) = p$:

$$S^a(p) = p_4^2 + (2p_4 + p_5)p_5 + (2(p_4 + p_5) + p_6)p_6 = (p_4 + p_5 + p_6)^2 = 0.25.$$
- $a(p) = 1$:

$$S^a(p) = p_4 + p_5 + p_6 = 0.5.$$
- Ergebnis gilt allgemein für beliebige k .

Signed rank test-Statistik und Midranks (gerader Fall) IV

- Sei $k = 2l = 6$: $p = (p_1, p_2, p_3, p_4, p_5, p_6)$.

$$S^a(p) = a((p_3 + p_4)/2)p_4 + a((2(p_3 + p_4) + p_2 + p_5)/2)p_5 \\ + a((2(p_3 + p_4 + p_2 + p_5) + p_1 + p_6)/2)p_6.$$

- Schiefeigenschaft: Wenn a monoton nicht abnehmend auf $(0, 1)$ und $p_i \leq p_{k+1-i}$, $i = 1, 2, \dots, l = [k/2]$, dann

$$S^a(p) \geq a(p_4)p_4 + a(2p_4 + p_5)p_5 + a(2(p_4 + p_5) + p_6)p_6$$

D.h.:

1. $a(p) = p$: $S^a(p) \geq 0.25$,
 2. $a(p) = 1$: $S^a(p) \geq 0.5$.
- Schiefeigenschaft: Dass „ $Q_i \leq 1/2$ “, $i = 1, 2, 3$ eingehalten wird, ist noch zu überprüfen.

Signed rank test-Statistik und Midranks (ungerader Fall)

- $k = 2l + 1, \in \mathbb{N}$
- Symmetriepunkt: $u_{(k+1)/2}$
- Problem: Bindungen $p_{(k+1)/2}$
- Lösung: Eliminationsstrategie

D.h. Elimination aller Beobachtungen, die mit $u_{(k+1)/2}$ übereinstimmen und Division der verbleibenden $k - 1$ Wahrscheinlichkeiten durch $1 - p_{(k+1)/2}$ (siehe z.B. Büning & Trenkler (1994), p. 90).

Pratt (1959) zeigt Mängel dieser Strategie auf.

- Setze $pp_i = (p_{(k+1)/2-i} + p_{(k+1)/2+i}) / (1 - p_{(k+1)/2}), i = 1, 2, \dots, l$.
- "Linear rank functional":

$$S^a(p) = \sum_{i=1}^l a(\overline{PP}_i) p_{(k+1)/2+i}.$$

Rangbasierte Schiefefunktionale

- Extreme Schiefe: Es existieren i mit $p_i \in \{0, 1\}$.
- D.h. Scoreserzeugende Funktion sollte in 0 und 1 definiert und beschränkt sein.
- Wilcoxon Schiefefunktional (ungerader Fall):

$$W = -4 \left(\sum_{i=1}^l \overline{PP}_i p_{(k+1)/2+i} - 0.25 \right)$$

- Vorzeichen Schiefefunktional (ungerader Fall):

$$S = -2 \left(\sum_{i=1}^l a(\overline{PP}_i) p_{(k+1)/2+i} - 0.5 \right)$$

- Problem: Nachweis, dass Schiefeordnung eingehalten wird, steht noch aus.

Rangbasierte Schiefestatistiken

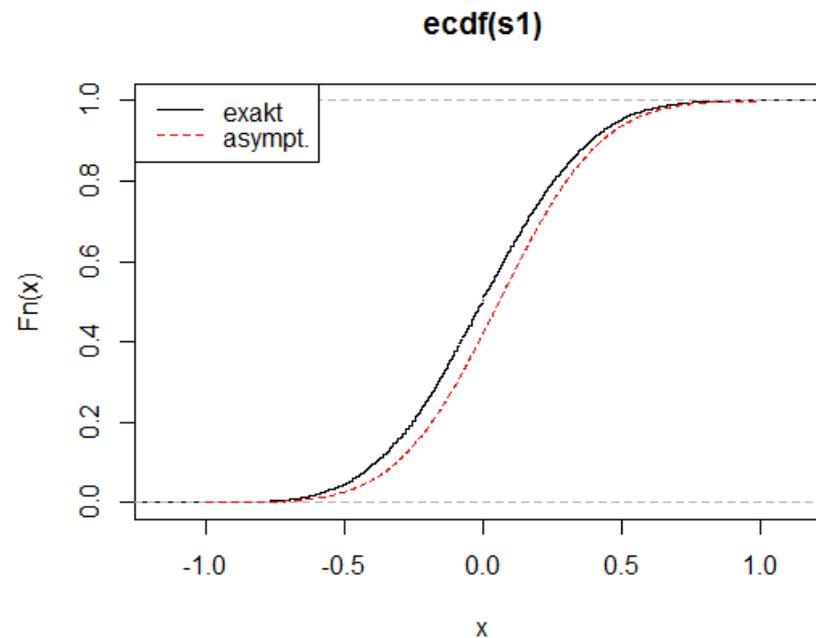
- Schätzen von p durch den Vektor der beobachteten relativen Häufigkeiten.
- Exakte Verteilung: Totalenumeration oder Simulation bzw. Binomialverteilung für Vorzeichenteststatistik.
- Asymptotische Verteilung für Wilcoxon Schiefestatistik: Normalverteilung mit Mittelwert $-(n+1)/n - 0.25$ und Varianz $16n(n+1)(2n+1)/(24n^4)$. Korrektur der Varianz bei Bindungen: $16/(48n^4) \sum_{j=1}^l ((nb_j)^3 - nb_j)$ mit

$$b_j = p_j + p_{k+1-j}, \quad j = 1, 2, \dots, l.$$

- Alternativ: Via Asymptotik von Funktionen multinomialer Zufallsvariablen.

Asymptotik Wilcoxon

- Exakte versus asymptotische Verteilung für $k = 7$ und $n = 20$



Rangbasierte Schiefestatistiken (Zweistichprobenfall)

- Wilcoxon-Mann-Whitney-Statistik und Median-Teststatistik
- Wilcoxon-Mann-Whitney-Statistik ist asymptotisch normalverteilt mit der Varianz

$$\frac{n_1 n_2}{12} \left(n + 1 + \frac{1}{n(n-1)} \sum_{j=1}^l (b_j^3 - b_j) \right),$$

wenn Bindungen der Länge $b_j, j = 1, 2, \dots, r$ auftreten und Mid-Ranks verwendet werden.

$n_i, i = 1, 2$ sind die beiden Stichprobenumfänge und $n = n_1 + n_2$ (siehe z.B. Büning & Trenkler (1994), S. 134 oder Lehmann (1975), S. 20).

- Median-Teststatistik ist hypergeometrisch verteilt.

Distanzbasierte Schiefestatistiken (Unspezifizierte Alternative)

- Sei Let $p = (p_1, \dots, p_k)$ mit $\sum_{i=1}^k p_i = 1$.
Symmetrie als Nullhypothese:

$$H_0 : p_i = p_{k+1-i}, \quad i = 1, 2, \dots, k.$$

Alternative: Beliebig schief.

- Testverfahren: u.a. verallgemeinerter Likelihoodverhältnistest oder χ^2 -Test.
- Identische Asymmetrie im Zweistichprobenfall:

$$H'_0 : Q_{1i} = Q_{2i}, \quad i = 1, 2, \dots, k - 1$$

- Alternative: Asymmetrie unterschiedlich.
- Testverfahren: Verallgemeinerter Likelihoodverhältnistest, Wald-Test (siehe Bhapkar (1961) und Bhapkar (1966)).

Distanzbasierte Schiefestatistiken (Gerichtete Alternative) I

- Alternativhypothese: Rechtsschief bzw. linkschief.
- Dykstra et al. (1995): Verallgemeinerter Likelihoodverhältnistest auf Symmetrie diskreter Verteilungen.

Diskrete Verteilung mit Träger $\{1, 2, \dots, 2l + 1\}$.

Alternativhypothesen:

$$H_1 : P_i \leq 1 - P_{k-i}, \quad i = 1, 2, \dots, l$$

oder

$$H_2 : p_i \leq p_{k+1-i}, \quad i = 1, 2, \dots, k$$

Distanzbasierte Schiefestatistiken (Gerichtete Alternative) II

- Statistik für Test von H_0 gegen $H_1 \setminus H_0$:

$$T_1 = 2 \sum_{i=1}^k n_i (\ln \hat{p}_i^{(1)} - \ln \hat{p}_i^{(0)})$$

$\hat{p}_i^{(0)} = (n_i + n_{k+1-i}) / (2n)$: ML-Schätzer von p_i unter H_0

$\hat{p}_i^{(1)}$: ML-Schätzer von p_i unter der Restriktion von H_1

- Berechnung von $\hat{p}_i^{(1)}$ mittels isotoner Regression:

Setze $\hat{p}_- = (\hat{p}_{k+1-i}, \hat{p}_{k-i}, \dots, \hat{p}_1)$. Dann

$$\hat{p}^{(1)} = \hat{p} E_{\hat{p}} \left(\frac{\hat{p} + \hat{p}_-}{2\hat{p}} \mid I \right),$$

wobei $E_w(x|I)$ LS-Projektion mit Gewichten w des Vektors x auf Kegel aller nichtabnehmenden Vektoren.

Distanzbasierte Schiefestatistiken (Gerichtete Alternative) III

- Numerik: Pool adjacent violators algorithm (PAVA) (implementiert in R, siehe de Leeuw et al. (2010)).

Distanzbasierte Schiefestatistiken (Gerichtete Alternative) IV

- Exakte Verteilung: Totalenumeration oder Simulation.
- Asymptotische Verteilung unter H_0 (Dykstra et al. (1995)):
Wenn p H_0 erfüllt und $p_i > 0, i = 1, 2, \dots, k = 2l + 1$, dann gilt

$$\lim_{n \rightarrow \infty} P(T_1 \geq t) = \sum_{i=0}^l p(i, k, p_r) P(\chi_{l-i}^2 \geq t),$$

wobei $p(0, k, p_r)$ Wahrscheinlichkeit, dass $E_{p_r}(V_r | J) = 0$ und $p(i, k, p_r)$ für $i = 1, 2, \dots, l$ Wahrscheinlichkeit, dass $E_{p_r}(V_r | J)$ i verschiedene Werte hat.

- Least favorable asymptotische Verteilung:

$$\sup_p \lim_{n \rightarrow \infty} P(T_1 \geq t) = \frac{1}{2} P(\chi_{l-1}^2 \geq t) + \frac{1}{2} P(\chi_l^2 \geq t)$$

Distanzbasierte Schiefestatistiken (Gerichtete Alternative) V

- Alternative $H_2 \setminus H_0$.
- Dykstra et al. (1995), S. 726: ML-Schätzer von p unter H_2 ist

$$\hat{p}_i^{(2)} = \begin{cases} \frac{n_i + n_{k+1-i}}{n} \left(\max \left\{ \frac{1}{2}, \frac{n_i}{n_i + n_{k+1-i}} \right\} \right) & \text{for } i = l + 1, \dots, k \\ \frac{n_i + n_{k+1-i}}{n} \left(\min \left\{ \frac{1}{2}, \frac{n_i}{n_i + n_{k+1-i}} \right\} \right) & \text{for } i = 1, \dots, l \\ \frac{n_i}{n} & \text{for } i = l + 1 \end{cases}$$

- Teststatistik:

$$T_2 = 2 \sum_{i=1}^k n_i (\ln \hat{p}_i^{(2)} - \ln \hat{p}_i^{(0)})$$

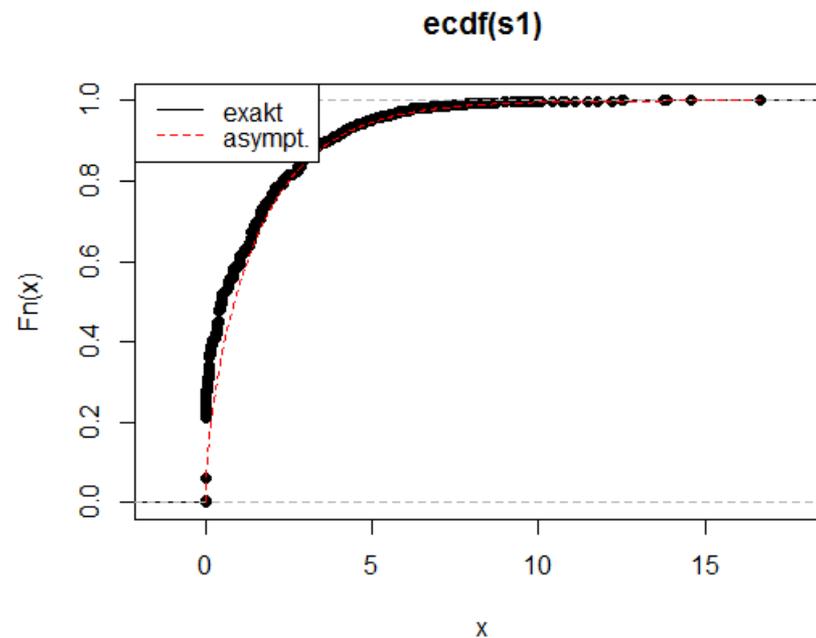
- Asymptotische Verteilung unter H_0 (Dykstra et al. (1995), S. 726):

$$\lim_{n \rightarrow \infty} P(T_2 \geq t) = \sum_{i=0}^l \binom{l}{i} \left(\frac{1}{2}\right)^k P(\chi_i^2 \geq t)$$

für alle $t \in \mathbb{R}$. Setze $\chi_0^2 = 0$.

Distanzbasierte Schiefestatistiken (Asymptotik)

- Exakte versus asymptotische Verteilung für $k = 7$ und $n = 20$



Distanzbasierte Schiefestatistiken (Gerichtete Alternative im Zweistichprobenfall) I

- Zweistichprobenfall fehlt bei Dykstra et al. (1995).
- Nullhypothese identischer Schiefe:

$$H'_0 : Q_i = Q'_i, \quad i = 1, 2, \dots, l = [k/2].$$

- Darstellung als Einstichprobenproblem für synthetische Wahrscheinlichkeiten:

$$H'_0 : q_i = q_{k+1-i}, \quad i = 1, 2, \dots, l$$

mit $q_i = (p_i + p'_{k+1-i})/2, i = 1, 2, \dots, k$.

- Alternativhypothese: „Linksschiefer als“, falls

$$H'_1 : Q_i \leq Q'_i, \quad i = 1, 2, \dots, l$$

oder äquivalent

$$H'_1 : \sum_{j=1}^i q_j \leq \sum_{j=1}^i q_{k+1-j} \quad i = 1, 2, \dots, l.$$

Distanzbasierte Schiefestatistiken (Gerichtete Alternative im Zweistichprobenfall) II

- Restriktivere Alternative;

$$H'_2 : q_i \leq q_{k+1-i}, \quad i = 1, 2, \dots, l.$$

Distanzbasierte Schiefestatistiken (Gerichtete Alternative im Zweistichprobenfall) III

- ML-Schätzer:

- Ohne Restriktion: $\hat{q}_i = (n_i/n + n'_{k+1-i}/n')/2$.

- Unter H'_0 :

$$\hat{q}_i^{(0)} = \frac{n_i/n + n'_{k+1-i}/n' + n'_i/n' + n_{k+1-i}/n}{4}$$

- Unter $H'_1 \setminus H'_0$:

$$\hat{q}_i^{(1)} = \hat{q} E \left(\frac{\hat{q} + \hat{q}_-}{2\hat{q}} \parallel I \right).$$

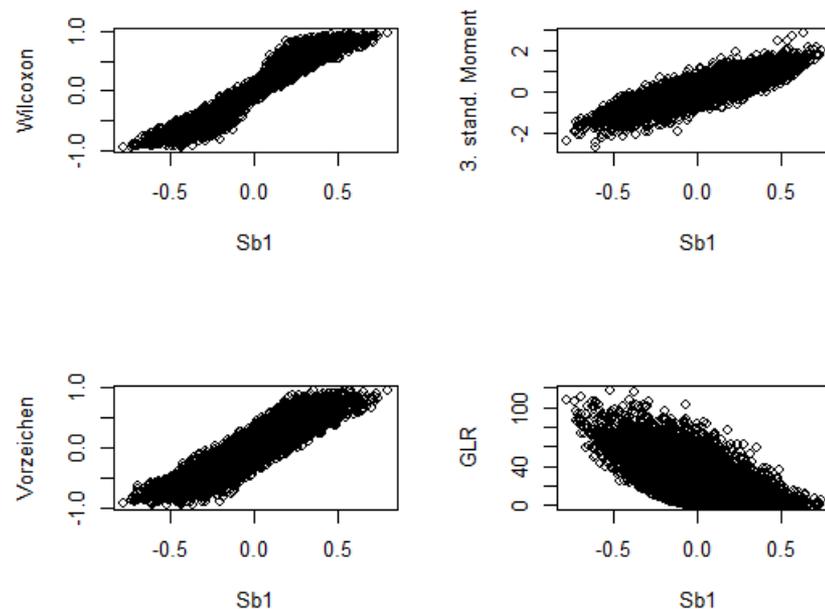
- Teststatistik:

$$T_1 = 2 \sum_{i=1}^k \frac{n_i + n'_{k+1-i}}{2} \left(\ln \hat{q}_i^{(1)} - \ln \hat{q}_i^{(0)} \right).$$

- Asymptotik analog zum Einstichprobenfall.

Vergleich der Werte von Schiefefunktionalen (Einstichprobenfall)

- $k = 7$ und 10000 Züge aus einer Dirichlet-Verteilung mit $\alpha_i = 1$, $i = 1, 2, \dots, 7$.



Power



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Power im Einstichprobenfall (Binomialverteilung) I

- Steuerung der Alternative über Parameter: θ als Schiefeparameter der Binomialverteilung:

$$p(x; k - 1, \theta) = \binom{k - 1}{u} \theta^u (1 - \theta)^{(k-1)-u}, \quad x = 0, 1, \dots, k - 1.$$

$$\mu_k = E(X) = (k - 1)\theta \text{ und}$$

$$S^{b_1}(p(.; \theta)) = 1 - \frac{2}{k - 1}(k - 1)\theta = 1 - 2\theta.$$

- Binomialverteilung ist symmetrisch für $\theta = 1/2$.
Für $\theta > 1/2$ ($\theta < 1/2$) ergibt sich linkschiefe (rechtsschiefe) Verteilung.

Power im Einstichprobenfall (Binomialverteilung) II

- Für einfache Zufallsstichprobe X_1, \dots, X_n ist ML-Schätzer für θ durch $1/(n(k-1)) \sum_{i=1}^n X_i$ gegeben.
UMP-Test mit Nullhypothese $\theta = 0.5$ gegen einseitige Hypothesen kann auf $\sum_{i=1}^n X_i$ oder äquivalent auf

$$S^{b_1} = 1 - \frac{2}{n(k-1)} \sum_{i=1}^n X_i$$

basieren.

- Alternative: Abgeschnittene diskrete Verteilungen in der linearen Exponentialfamilie (z.B. Poisson-, negative Binomialverteilung).

Power im Einstichprobenfall (Binomialverteilung, exakt)

θ_1	$\sqrt{b_1}$	S^{b_1}	S^{b_2}	Sign	Wilcoxon	GLR
$n = 20$						
0.5	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
0.52	0.0580	0.1222	0.1244	0.1078	0.1157	0.0873
0.55	0.0688	0.3021	0.3004	0.2513	0.2806	0.1987
0.6	0.0911	0.7204	0.7264	0.6062	0.6722	0.5338
0.7	0.1575	0.9991	0.9985	0.9992	0.9969	0.9868
0.8	0.2786	1.0000	1.0000	1.0000	1.0000	1.0000
0.9	0.6035	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 100$						
0.5	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
0.52	0.0668	0.2674	0.2625	0.2404	0.2602	0.1978
0.55	0.1146	0.7987	0.7931	0.7850	0.7859	0.6572
0.6	0.2162	0.9997	0.9996	0.9997	0.9997	0.9963
0.7	0.5451	1.0000	1.0000	1.0000	1.0000	1.0000
0.8	0.9097	1.0000	1.0000	1.0000	1.0000	1.0000
0.9	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 1000$						
0.5	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
0.52	0.1309	0.9282	0.9258	0.8903	0.9205	0.8455
0.55	0.3934	1.0000	1.0000	1.0000	1.0000	1.0000
0.6	0.8763	1.0000	1.0000	1.0000	1.0000	1.0000
0.7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Power im Einstichprobenfall (Binomialverteilung, asympt.)

θ_1	$\sqrt{b_1}$	S^{b_1}	S^{b_2}	Sign	Wilcoxon	GLR
$n = 20$						
0.5	0.0499	0.0388	0.0274	0.0515	0.0210	0.0304
0.52	0.0580	0.0990	0.0712	0.1113	0.0541	0.0594
0.55	0.0686	0.2614	0.2044	0.2568	0.1582	0.1451
0.6	0.0908	0.6801	0.6143	0.6170	0.4939	0.4482
0.7	0.1567	0.9981	0.9965	0.9992	0.9828	0.9751
0.8	0.2784	1.0000	1.0000	1.0000	1.0000	1.0000
0.9	0.6035	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 100$						
0.5	0.0520	0.0462	0.0387	0.0483	0.0315	0.0364
0.52	0.0684	0.2491	0.2229	0.2333	0.2023	0.1548
0.55	0.1181	0.7809	0.7538	0.7787	0.7247	0.5963
0.6	0.2199	0.9995	0.9995	0.9996	0.9993	0.9950
0.7	0.5504	1.0000	1.0000	1.0000	1.0000	1.0000
0.8	0.9114	1.0000	1.0000	1.0000	1.0000	1.0000
0.9	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 1000$						
0.5	0.0514	0.0496	0.0496	0.0514	0.0408	0.0326
0.52	0.1322	0.9248	0.9257	0.8917	0.9065	0.7926
0.55	0.3959	1.0000	1.0000	1.0000	1.0000	1.0000
0.6	0.8770	1.0000	1.0000	1.0000	1.0000	1.0000
0.7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Power im Einstichprobenfall (Binomialverteilung)

```
function (n,k,alpha,p0)
{
  set.seed(12345)
  n1=10000
  sb=numeric(n1)
  p=dbinom(seq(0,k-1,1),k-1,1/2)
  nk=n
  l=k/2
  for (i in 1:n1)
  {
    per=rmultinom(1,n,p)/n
    sb[i]=wilcox_ties(per)
  }
}
```

Power im Zweistichprobenfall (Binomialverteilung) I

Seien p_1 and p_2 Parameter zweier Binomialverteilungen und $\delta = p_1 - p_2$.

δ	$\sqrt{b_1}$	S^{b_1}	S^{b_2}	Wilcoxon	GLR
$m = 20, n = 20$					
0	0.0503	0.04908	0.0506	0.0488	0.0497
0.02	0.0710	0.09154	0.0834	0.1012	0.0794
0.05	0.1001	0.21695	0.1650	0.2535	0.1565
0.1	0.1387	0.54040	0.3837	0.5790	0.3527
0.2	0.1986	0.96914	0.8775	0.9589	0.7923
0.3	0.2568	0.99980	0.9978	0.9999	0.9858
0.4	0.3193	1.00000	1.0000	1.0000	1.0000
$m = 100, n = 100$					
0	0.0501	0.0497	0.0456	0.0503	0.0507
0.02	0.1131	0.1878	0.1384	0.2349	0.1492
0.05	0.1718	0.6721	0.4161	0.6864	0.4374
0.1	0.2792	0.9898	0.9052	0.9842	0.8877
0.2	0.4710	1.0000	1.0000	1.0000	0.9999
0.3	0.6510	1.0000	1.0000	1.0000	1.0000
0.4	0.8248	1.0000	1.0000	1.0000	1.0000
$m = 1000, n = 1000$					
0	0.0525	0.05100	0.0490	0.0511	0.0509
0.02	0.2249	0.85133	0.5518	0.8398	0.6014
0.05	0.4125	1.00000	0.9764	1.0000	0.9995
0.1	0.7418	1.00000	1.0000	1.0000	1.0000
0.2	0.9926	1.00000	1.0000	1.0000	1.0000
0.3	0.9992	1.00000	1.0000	1.0000	1.0000
0.4	0.9990	1.00000	1.0000	1.0000	1.0000

Power im Zweistichprobenfall (Binomialverteilung) II

```
function (n1,n2,k,alpha,delta)
{
  u=seq(0,k-1,1)
  set.seed(12345)
  nwh=10000
  s1=numeric(nwh)
  s2=numeric(nwh)
  s=numeric(nwh)
  x=numeric(10000)
  i=0
  while (i < nwh)
  {
    p0=runif(1)
    if ((p0+delta)<1)
    {
      i=i+1
      p1=dbinom(seq(0,k-1,1),k-1,p0)
      per1=rmultinom(1,n1,p1)/n1
      p1=dbinom(seq(0,k-1,1),k-1,p0)
      per2=rmultinom(1,n2,p1)/n2
      s1[i]=wilcox.two_ties(n1,n2,per1,per2)
      p2=dbinom(seq(0,k-1,1),k-1,p0+delta)
      per2=rmultinom(1,n2,p2)/n2
      s2[i]=wilcox.two_ties(n1,n2,per1,per2)
    }
  }
  t=s1[1:i]
  t=sort(t)
  Fnp=ecdf(t)
  krit1=t[alpha*i]
  print(krit1)
  t=s2[1:i]
  beta=sum(t<krit1)/i
  return(beta)
}
```

Power im Einstichprobenfall (Lehmann-Alternative) I

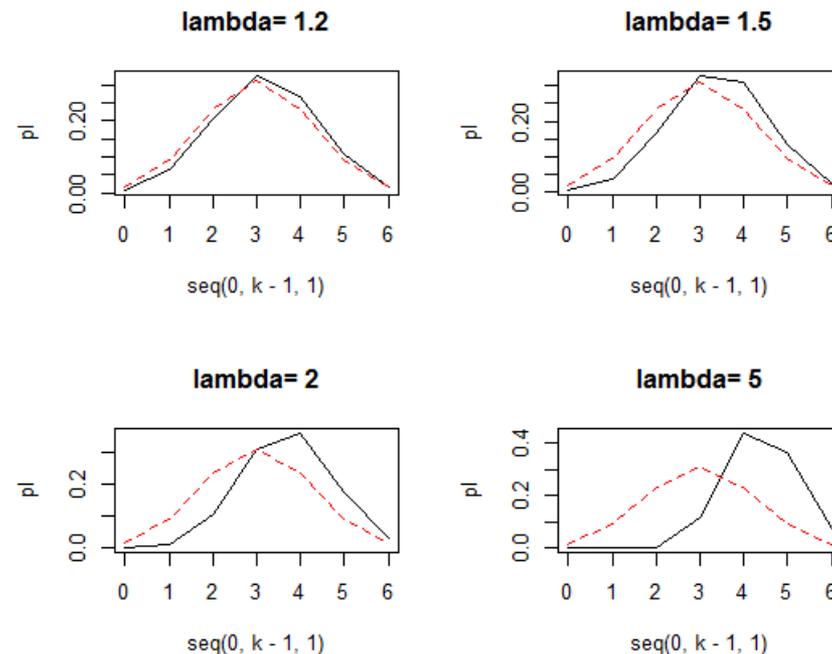
- Steuerung der Alternative über Parameter λ der Lehmann-Alternative.
- Sei $p = (p_1, \dots, p_k) \in \Delta_k$, dann liefert

$$p^s = (p + (p_k, \dots, p_1))/2$$

einen symmetrischen Wahrscheinlichkeitsvektor mit kumulativen Wahrscheinlichkeiten $P_i^s = \sum_{j=1}^i p_j^s$.

- Betrachte $P_i = (P_i^s)^\lambda$, $i = 1, 2, \dots, k$ als Alternative (Linksschiefe) zur Nullhypothese der Symmetrie.
- Generierung von p : Ziehe beliebigen Wahrscheinlichkeitsvektor p aus Dirichlet-Verteilung mit Hyperparameter $\alpha = (1, 1, \dots, k)$.

Power im Einstichprobenfall (Lehmann-Alternative) II



Power im Einstichprobenfall ($k = 7$, Lehmann-Alternative, exakt)

λ	$\sqrt{b_1}$	S^{b_1}	Wilcoxon	GLR
$n = 20$				
1.00	0.0491	0.0515	0.0492	0.0485
1.05	0.0621	0.0724	0.0705	0.0632
1.10	0.0766	0.0969	0.0976	0.0789
1.20	0.1119	0.1595	0.1669	0.1142
1.50	0.2511	0.4278	0.4674	0.3075
2.00	0.4706	0.8213	0.8711	0.7135
3.00	0.6457	0.9914	0.9988	0.9891
5.00	0.7025	1.0000	1.0000	1.0000
$n = 100$				
1.00	0.0499	0.0509	0.0501	0.0478
1.05	0.0861	0.1039	0.1064	0.0814
1.10	0.1374	0.1883	0.2000	0.1347
1.20	0.2803	0.4324	0.4653	0.3144
1.50	0.6836	0.9555	0.9733	0.9173
2.00	0.8305	0.9999	1.0000	1.0000
3.00	0.8740	1.0000	1.0000	1.0000
5.00	0.9201	1.0000	1.0000	1.0000
$n = 1000$				
1.00	0.0501	0.0500	0.0495	0.0496
1.05	0.2337	0.3419	0.3627	0.2417
1.10	0.5365	0.7877	0.8154	0.6716
1.20	0.8496	0.9986	0.9995	0.9957
1.50	0.9187	1.0000	1.0000	1.0000
2.00	0.9213	1.0000	1.0000	1.0000
3.00	0.9313	1.0000	1.0000	1.0000
5.00	0.9589	1.0000	1.0000	1.0000

Power im Zweistichprobenfall (Lehmann-Alternative)

Seien $\lambda_1, \lambda_2 > 0$.

λ_1	λ_2	$\sqrt{b_1}$	S^{b_1}	Wilcoxon	GLR
$n = 20, m = 20$					
1.00	1.00	0.0499	0.0529	0.0473	0.0509
1.50	1.00	0.1051	0.3004	0.2605	0.1407
2.00	1.00	0.1262	0.6119	0.5269	0.2026
3.00	1.00	0.1504	0.8888	0.8122	0.2468
1.50	1.50	0.0492	0.0520	0.0502	0.3075
2.00	1.50	0.0628	0.2040	0.1777	0.0961
3.00	1.50	0.0872	0.5509	0.4858	0.1375
2.00	2.00	0.0491	0.0552	0.0521	1.0000
3.00	2.00	0.0687	0.2618	0.2295	0.0903
3.00	3.00	0.0506	0.0525	0.0509	0.0521
$n = 100, m = 100$					
1.00	1.00	0.0494	0.0507	0.0488	0.0485
1.50	1.00	0.2164	0.7265	0.7236	0.2741
2.00	1.00	0.2265	0.9594	0.9513	0.3585
3.00	1.00	0.2426	0.9957	0.9923	0.3659
1.50	1.50	0.0524	0.0514	0.0527	0.0487
2.00	1.50	0.0640	0.3550	0.3616	0.1308
3.00	1.50	0.1080	0.7980	0.7676	0.1910
2.00	2.00	0.0497	0.0499	0.0518	0.0512
3.00	2.00	0.0721	0.3675	0.3550	0.1131
3.00	3.00	0.0503	0.0508	0.0510	0.0498

Anwendungen



FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

FACHBEREICH WIRTSCHAFTS-
WISSENSCHAFTEN

Anwendungsfall: Links-Rechts-Selbsteinstufung in West- und Ostdeutschland I

- Quelle: ALLBUS 2010 (im Netz nach Anmeldung bei GESIS verfügbar)
- 10-er Ratingskala mit "1=extrem links" und "10=extrem rechts".
- $n_w = 1810$ Antworten im Westen
 $n_o = 845$ Antworten im Osten.

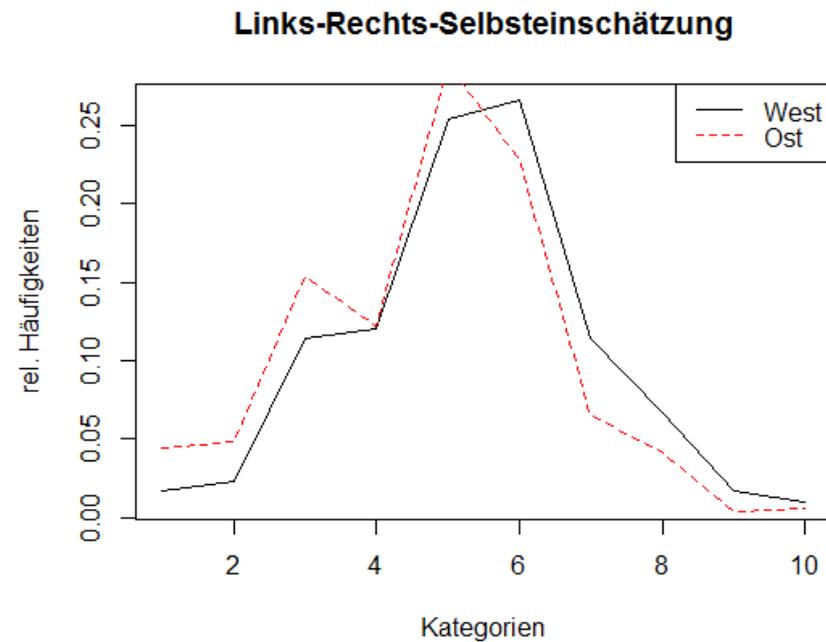
Anwendungsfall: Links-Rechts-Selbsteinstufung in West- und Ostdeutschland II

- Ergebnis der Befragung:

West					
i	1	2	3	4	5
f^w	0.01713	0.02320	0.11381	0.11989	0.25359
F^w	0.01713	0.04033	0.15414	0.27403	0.52762
Q^w	0.50387	0.506905	0.53066	0.53370	0.52762
i	6	7	8	9	10
f^w	0.26575	0.11381	0.06630	0.01713	0.00939
F^w	0.79337	0.90718	0.97348	0.99061	1.00000
Q^w	0.53370	0.53066	0.506905	0.50387	

Ost					
i	1	2	3	4	5
f^o	0.04379	0.04852	0.15384	0.12189	0.28876
F^o	0.04379	0.09231	0.24615	0.36804	0.65680
Q^o	0.518935	0.54142	0.59763	0.62603	0.65680
i	6	7	8	9	10
f^o	0.22722	0.06509	0.04142	0.00355	0.00592
F^o	0.88402	0.94911	0.99053	0.99408	1.00000
Q^o	0.62603	0.59763	0.54142	0.518935	

Anwendungsfall: Links-Rechts-Selbsteinstufung in West- und Ostdeutschland III



Anwendungsfall: Links-Rechts-Selbsteinstufung in West- und Ostdeutschland IV

- Beachte: $Q_i^W > 0.5$ und $Q_i^O > 0.5$, $i = 1, 2, \dots, 9$.
D.h. Rechtsschiefe Häufigkeitsverteilungen.
- Beachte: $Q_i^W < Q_i^O$ for $i = 1, 2, \dots, 9$.
D.h. Häufigkeitsverteilung im Osten linksschiefer als im Westen.
- Beachte: Kategorie 5 ist Median beider Häufigkeitsverteilungen.

Anwendungsfall: Links-Rechts-Selbsteinstufung in West- und Ostdeutschland V

Testergebnisse im Einstichprobenfall getrennt für West- und Ostdeutschland:

Test	sample value	critical value (exact)	p value (exact)	critical value (asympt.)	p value (asympt.)
West ($n = 1810$)					
$\sqrt{b_1}$	-0.0511	0.1134	0.7681	0.0907	0.8230
S^{b_1}	0.0395	0.0129	0.0000	0.01421	0.0000
S^{b_2}	0.0768	0.0256	0.0003	0.0244	0.0000
Wilcoxon	0.1084	0.0440	0.0000	0.0443	0.0000
Sign	0.0552	0.0375	0.0093	0.0387	0.0094
GLR	29.566	4.5651	0.0000	5.1250	0.0000
Ost ($n = 845$)					
$\sqrt{b_1}$	-0.1368	0.1134	0.9748	0.1286	0.9599
S^{b_1}	0.1611	0.0188	0.0000	0.0232	0.0000
S^{b_2}	0.2868	0.0365	0.0000	0.0345	0.0000
Wilcoxon	0.4380	0.0629	0.0000	0.0653	0.0000
Sign	0.3136	0.0556	0.0000	0.0566	0.0000
GLR	145.8970	4.4706	0.0000	5.1360	0.0000

Anwendungsfall: Links-Rechts-Selbsteinstufung in West- und Ostdeutschland VI

Testergebnisse im Zweistichprobenfall:

Test	sample value	critical value (exact)	p value (exact)	critical value (asympt.)	p value (asympt.)
$\sqrt{b_1}$	-0.0857	-0.1822	0.1993	-0.1569	0.1846
S^{b_1}	-0.1215	-0.0189	0.0000	-0.0205	0.0000
S^{b_2}	-0.1215	-0.0238	0.0000	-0.0215	0.0000
Wilcoxon	87623.5000	1094872.5000	0.0000	1092504.8857	0.0000
Median	412.0000	403.0000	0.2061	402.6015	0.1944
GLR	288.7651	9.6892	0.0000	9.5564	0.0000

Anwendungsfall: Datensatz von Kemp & Conover (1973) I

- Einsatz von ORTHENE und AKTON zur Bekämpfung von Larvenbefall auf 10 Parzellen mit „Bentgrass“.
- Zähldaten:

Plot	ORTHENE	AKTON
1	1	0
2	0	0
3	0	1
4	2	0
5	1	0

Anwendungsfall: Datensatz von Kemp & Conover (1973) II

- Häufigkeitsverteilungen:

		ORTHENE			AKTON		
i	x_i	f_i	F_i	Q_i	f_i	F_i	Q_i
1	0	0.4	0.4	0.6	0.8	0.8	0.9
2	1	0.4	0.8	0.6	0.2	1.0	0.9
3	2	0.2	1.0		0.0	1.0	

- Beachte: Beide Häufigkeitsverteilungen sind rechtsschief.
- Beachte: Häufigkeitsverteilung von ORTHENE linksschiefer.

Anwendungsfall: Datensatz von Kemp & Conover (1973) III

Testergebnisse im Einstichprobenfall:

Test	sample value	critical value (exact)	p value (exact)	critical value (asympt.)	p value (asympt.)
FORTHENE					
$\sqrt{b_1}$	0.3436	1.5000	0.2607	0.5756	0.1631
S^{b_1}	0.2000	0.6000	0.1733	0.3121	0.1459
S^{b_2}	0.3600	0.8400	0.1759	0.8154	0.2339
Wilcoxon	0.3333	1.0000	0.2579	1.6933	0.5000
Sign	0.3333	1.0000	0.2579	0.9497	0.2819
GLR	0.3398	4.1587	0.4240	5.1361	0.7018
AKTON					
$\sqrt{b_1}$	1.5000	1.5000	0.0783	1.8112	0.0865
S^{b_1}	0.8000	0.6000	0.0009	0.4681	0.0025
S^{b_2}	0.9600	0.8400	0.0010	0.3058	0.0000
Wilcoxon	1.0000	1.0000	0.0000	1.3667	0.1346
Sign	1.0000	1.0000	0.0000	0.8224	0.0228
GLR	5.4750	4.1599	0.0261	5.1361	0.0420

Anwendungsfall: Datensatz von Kemp & Conover (1973) IV

Testergebnisse im Zweistichprobenfall:

Test	sample value	critical value (exact)	p value (exact)	critical value (asympt.)	p value (asympt.)
$\sqrt{b_1}$	-1.1943	0.0000	-1.9021	0.1509	
S^{b_1}	-0.5990	-0.6000	0.0844	-0.9477	0.1077
S^{b_2}	-0.5990	-0.8400	0.1105	-0.7619	0.0978
Wilcoxon	22.0075	20.5000	0.1005	20.4277	0.1007
GLR	1.3227	1.9274	0.0852	2.6977	0.1251

Notenverteilung Klausuren mit 10 ECTS und 7.5 ECTS im WS 2011/12 I

- Zusammenfassung der Ergebnisse zu Noten 1, 2, 3, 4, 5
- $n_{10} = 567$ Klausuren 10 ECTS
- $n_{7.5} = 357$ Klausuren 7.5 ECTS

Notenverteilung Klausuren mit 10 ECTS und 7.5 ECTS im WS 2011/12 II

Daten:

i	1	2	3	4	5
f_{10}	0.09171	0.20282	0.26984	0.23104	0.20459
F_{10}	0.09171	0.29453	0.56437	0.79541	1.00000
Q_{10}	0.44356	0.42945	0.42945	0.44356	
i	1	2	3	4	5
$f_{7.5}$	0.04762	0.11484	0.19328	0.29132	0.35294
$F_{7.5}$	0.04762	0.16246	0.35574	0.64706	1.00000
$Q_{7.5}$	0.34734	0.25910	0.25910	0.34734	

Notenverteilung Klausuren mit 10 ECTS und 7.5 ECTS im WS 2011/12 III

Testergebnisse im Einstichprobenfall:

Test	sample value	critical value (exact)	p value (exact)	critical value (asympt.)	p value (asympt.)
10 ECTS					
$\sqrt{b_1}$	0.1391	0.1098	0.0203	0.0896	0.00535
S^{b_1}	0.1270	0.3527	0.0000	0.0275	0.0000
Wilcoxon	0.2694	0.09193	0.0000	0.09444	0.0000
GLR	26.0492	4.6237	0.0000	5.1244	0.0000
7.5 ECTS					
$\sqrt{b_1}$	0.7015	0.1383	0.0000	0.1386	0.0000
S^{b_1}	0.3936	0.0434	0.0000	0.0429	0.0000
Wilcoxon	0.6792	0.1185	0.0000	0.1139	0.0000
GLR	122.2449	4.6665	0.0000	5.1226	0.0000

Notenverteilung Klausuren mit 10 ECTS und 7.5 ECTS im WS 2011/12 IV

Testergebnisse im Zweistichprobenfall:

Test	sample value	critical value (exact)	p value (exact)	critical value (asympt.)	p value (asympt.)
$\sqrt{b_1}$	-0.5624	-3.0892	0.0019	-0.1585	0.0000
S^{b_1}	-0.2267	-0.04601	0.0000	-0.0478	0.0000
Wilcoxon	237488	256519	0.0000	255916	0.0000
GLR	37.02748	3.9953	0.0000	5.1333	0.0000

Empirische Studien aus Bhattacharya & Nandram (2011) I

- Nichtraucherkampagne (aus: Gelb & Pickett (1983)):
Frage, ob man mit einer Antiraucheranzeige übereinstimmt.

	strongly agree	agree	neutral	disagree	strongly disagree
Raucher	8	14	35	21	19
Nichtraucher	31	42	78	61	69

- Arthritis Datensatz (aus Koch (1985))
Behandlungsgruppe: 73
Kontrollgruppe: 75
Frage, wie Behandlung wirkt.

	exzellent agree	gut	moderat	befriedigend	schwach disagree
Behandlung	7	23	20	10	13
Kontrolle	3	15	17	16	24

Empirische Studien aus Bhattacharya & Nandram (2011) II

- Geschmacksbewertung von 4 Joghurts *A*, *B*, *C*, *D* (aus: Newell (1982))

	strongly like	like	neutral	dislike	strongly dislike
A	11	22	9	1	0
B	5	9	14	10	5
C	0	2	15	12	14
D	0	3	9	14	17

Anhang: Maximalinvarianz I

- **Ausgangspunkt:** Gruppe G von Transformationen bezgl. der Operation der Hintereinanderausführung, mit denen die Beobachtungen (x_1, \dots, x_n) transformiert werden

$$g(x_1), \dots, g(x_n) \text{ für } g \in G.$$

- Eine **statistische Maßzahl** ist eine Abbildung

$$T = t(x_1, \dots, x_n).$$

- T heißt **G -invariant**, wenn

$$t(x_1, \dots, x_n) = t(g(x_1), \dots, g(x_n))$$

für alle $g \in G$.

- u heißt **G -maximalinvariant**, wenn u G -invariant ist und es für alle x und y mit $u(x) = u(y)$ ein $g \in G$ gibt, so daß

$$y = g(x).$$

Anhang: Maximalinvarianz II

- Beispiele (Klein (1994), Kapitel 6):

1. Gruppe der affin-linearen Transformationen (n **identische** Intervallskalen)

$$u(x_1, \dots, x_n) = \left(\frac{x_1 - \bar{x}}{s_x}, \dots, \frac{x_n - \bar{x}}{s_x} \right)$$

Konstruktion von g : Wenn

$$\left(\frac{y_1 - \bar{y}}{s_y}, \dots, \frac{y_n - \bar{y}}{s_y} \right) = \left(\frac{x_1 - \bar{x}}{s_x}, \dots, \frac{x_n - \bar{x}}{s_x} \right)$$

dann

$$y_i = \bar{y} - \bar{x} \frac{s_y}{s_x} + \frac{s_y}{s_x} x_i$$

für $i = 1, 2, \dots, n$.

2. Gruppe der linearen Transformationen (n **identische** Verhältnisskalen):

$$u(x_1, \dots, x_n) = \left(\frac{x_1}{G_x}, \dots, \frac{x_n}{G_x} \right)$$

mit $G_x = (\prod_{i=1}^n x_i)^{1/n}$.

Anhang: Maximalinvarianz III

- **Satz: (Lehmann, (1953), Büning, (1973))**
Existiert eine G -Maximalinvariante, dann ist jede G -invariante Funktion ausschließlich Funktion dieser Maximalinvarianten.

Anhang: Maximalinvarianz für stetige Verteilungsfamilien I

- Sei \mathcal{F} die Menge aller stetigen Verteilungsfunktionen, dann ist

$$T : \mathcal{F} \rightarrow \mathcal{F}$$

eine statistische Maßzahl.

- Sei G wiederum eine Transformationsgruppe, dann ist T G -invariant, wenn

$$T(F \circ g^{-1}) = T(F)$$

für alle $g \in G$.

- **Wahrscheinlichkeitsintegraltransformation:**

Sei X die zu F_X gehörende stetige Verteilungsfunktion. Betrachte

$$u(F_X) = F_X(X) = U,$$

dann bildet u jede stetige Verteilung auf die $[0, 1]$ -rechteckverteilte Zufallsvariable U ab.

Anhang: Maximalinvarianz für stetige Verteilungsfamilien II

- Wenn g eine **streng monoton zunehmende Transformation** ist, folgt für stetiges F auch

$$u(F_{g(X)}) = F_{g(X)}(g(X)) = U = F_X(X) = u(F),$$

so dass u g -invariant ist.

- Seien X und X' Zufallsvariablen mit Verteilungsfunktionen F_X und $F_{X'}$.
Wenn

$$u(F_X) = F_X(X) = F_{X'}(X') = u(F_{X'})$$

ist, dann existiert mit

$$X' = F_{X'}^{-1}(F_X(X))$$

eine streng monoton zunehmende Transformation $g = F_{X'}^{-1} \circ F_X$, so daß $X' = g(X)$ bzw. $F_{X'}(x) = F_X(g^{-1}(x))$.

D.h. u ist **maximalinvariant** bezüglich der Gruppe der streng monoton zunehmenden Transformationen.

Anhang: Maximalinvarianten streng monoton zunehmender Transformationen

- Ranganteile, wenn keine Bindungen vorliegen.
- Kumulierte relative Häufigkeiten, wenn Bindungen vorliegen.
- $U = F_X(X)$, wenn eine theoretische Verteilung vorliegt.

Anhang: Translationsinvarianz und Maximalinvarianz I

- Stevens (1946), Suppes (1959), Adams et al. (1965), Pfanzagl (1971): Statistische Maßzahlen sollen zumindest (formal) **"sinnvolle" Vergleiche** ermöglichen.
- Eine statistische Maßzahl $T = t(x_1, \dots, x_n)$ heißt **G-vergleichsinvariant** wenn für alle $g \in G$ gilt:

$$t(x_1, \dots, x_n) = t(y_1, \dots, y_n) \iff \begin{aligned} &t(g(x_1), \dots, g(x_n)) \\ &= t(g(y_1), \dots, g(y_n)). \end{aligned}$$

- **Satz (Klein, (1994), S. 178f.):** Sei G die Gruppe der auf \mathbb{R} streng monoton zunehmenden Transformationen, T G -vergleichsinvariant und T G -invariant bez. beliebiger **Translationen**, dann ist T G -invariant und damit Funktion von Maximalinvarianten.
- Translationsinvariante statistische Maße sind u.a. **Streuungsmaße**, **Schiefe- und Wölbungsmaße** und **Abhängigkeitsmaße**.

Anhang: Translationsinvarianz und Maximalinvarianz II

- **Ties:** Streuungsmaß nach Vogel & Dobbener:

$$- \sum F_i \ln F_i - \sum (1 - F_i) \ln(1 - F_i)$$

- **Keine Ties:** $F_n(x_i) = i/n$ zu setzen und dieses Maß nur noch von n abhängig wie auch z.B.

$$1/n \sum_{i=1}^n (r_i - \bar{r})^2 = (n^2 - 1)/12$$

- **Theoretische Verteilung:** Z.B.

$$\text{Var}[F_X(X)] = 1/12$$

Anhang: Isotone Regression I

- Sei $\mathcal{D} = \{d_1, \dots, d_K\}$. d_i können z.B. Dosisniveaus einer Medikamentengabe bezeichnen, wobei $d_i < d_{i+1}$ für $i = 1, 2, \dots, K - 1$ gilt.
- Sei g eine gegebene Funktion auf \mathcal{D} . Eine Funktion g^* ist eine isotone Regression von g mit Gewichtsvektor w , wenn g^* isoton ist und

$$\sum_{d \in \mathcal{D}} |g(d) - f(d)|^2 w(d)$$

in der Menge aller auf \mathcal{D} isotonen Funktion minimiert.

Anhang: Isotone Regression II

- Sei \mathcal{D} einfach geordnet. PAVA-Algorithmus:
 1. Wenn $g(d)$ nicht-abnehmend geordnet, dann $g^*(d) = g(d)$.
 2. Ansonsten gibt es ein d , so dass $g(d) > g(d + 1)$. Ersetze d und $d + 1$ durch ihr gewichtetes arithmetisches Mittel:

$$\bar{g} = \frac{w(d)g(d) + w(d + 1)g(d + 1)}{w(d) + w(d + 1)}.$$

3. Dann bilden d und $d + 1$ eine sog. Niveaumenge („level set“ oder „solution block“). Wenn g nicht abnehmend geordnet auf den verbliebenen $K - 1$ Werten ist, setze $g^*(d) = g^*(d + 1) = \bar{g}$ und $g^*(d) = g(d)$ sonst.
4. Ansonsten wiederhole Vorgehen mit den neuen Werten.

Anhang: Isotone Regression II

- Beispiel: $\mathcal{D} = \{1, 2, 3, 4, 5\}$ mit $g(1) = 1, g(2) = 0, g(3) = 2, g(4) = 1, g(5) = 3, w_i = 1, i = 1, 2, \dots, 5$.
 1. $\bar{g} = (g(1) + g(2))/2 = 0.5, g^*(1) = g^*(2) = 0.5.$
 2. $\bar{g} = (g(3) + g(4))/2 = 1.5, g^*(3) = g^*(4) = 1.5.$
 3. Lösung: $g^*(1) = g^*(2) = 0.5, g^*(3) = g^*(4) = 1.5, g^*(5) = 3.$

Anhang: Isotone Regression (Verteilung unter Nebenbedingungen) I

- Seien X_1, \dots, X_n stochastisch unabhängig, identisch normalverteilt mit Mittelwert μ und Varianz 1.
- Testproblem: $H_0 : \mu = 0$ gegen $H_1 : \mu > 0$.
- ML-Schätzer unter H_0 : $\hat{\mu}_0 = 0$.
- ML-Schätzer von μ unter der Nebenbedingung $\mu > 0$:

$$\hat{\mu}_r = \begin{cases} \bar{X}_n & \text{falls } \bar{X}_n > 0 \\ 0 & \text{sonst} \end{cases}$$

Anhang: Isotone Regression (Verteilung unter Nebenbedingungen) II

- Verallgemeinertes Likelihoodverhältnis

$$\Lambda = \begin{cases} \exp(-1/2n\bar{X}_n^2) & \text{falls } \bar{X}_n > 0 \\ 1 & \text{sonst} \end{cases}$$

bzw.

$$-2 \ln \Lambda = \begin{cases} n\bar{X}_n^2 & \text{falls } \bar{X}_n > 0 \\ 0 & \text{sonst} \end{cases}$$

- Verteilung von $T = -2 \ln \Lambda$ unter H_0 ist Mischung aus diskreter und stetiger Verteilung:

$$\begin{aligned} P(T \leq t | H_0) &= P(T = 0) + P(0 < T \leq t | \bar{X}_n > 0)P(\bar{X}_n > 0) \\ &= 1/2 + 1/2F_{\chi^2(1)}(t). \end{aligned}$$

Anhang: Zipf-Verteilung

$$f(x; s, N) = \frac{1/x^s}{\sum_{y=1}^N 1/y^s}, \quad x = 1, 2, \dots, N, \quad N \in \mathbb{N}, \quad s > 0$$