Eckehardt Spenhoff, Holzwickede

Paarweiser Vergleich

Ein universelles Werkzeug zur Definition von Utilitätsskalen

Der paarweise Vergleich ist ein vielfältig einsetzbares Werkzeug zur Prüfung sensorischer Unterschiede mit einer Beurteilung der Eignung von Beurteilern. Dieser Artikel zeigt auch die Anwendung der Methodik zur Priorisierung bei komplexen Entscheidungen. Die Einsatzmöglichkeiten werden nach statistischer (sensorische Prüfung) und nicht statistischer (Priorisierung) Methodik unterschieden.

Grundlagen

Utilitätsskalen sind nur bedingt mit Skalen für Messwerte vergleichbar, weil sie keine Absolutwerte enthält und Differenzen im allgemeinen nicht äquidistant sind. Demzufolge sind alle Rangskalen als Utilitätsskalen anzusehen. Utilitätsskalen definieren also Rangfolgen, welche zum Vergleich oder zur Setzung von Prioritäten geeignet sind.
Sind mehrere Objekte in eine Rangordnung zu bringen wird dies um so schwieriger je mehr Objekte vorhanden sind oder je komplexer die Beurteilung der Objekte ist. Einfacher ist der Vergleich, wenn immer nur zwei Objekte miteinander verglichen werden. Für N Objekte sind

$$M = \binom{N}{2}$$

paarweise Vergleiche möglich. Die binären Urteile der paarweisen Vergleiche werden so zusammengefasst, dass sich für alle Objekte eine Rangfolge ergibt.

Managementmethoden

Der paarweise Vergleich ist ein häufig gebrauchtes Instrument, so ist er Bestandteil in der QFD (Quality Function Deployment), in der FMEA (Failure Mode and Effect Analysis, deutsch Ausfalleffektanalyse) und den New Seven Tools. Im Vordergrund dieser Anwendungen steht immer die Festlegung von Prioritäten. Bewertet werden zum Beispiel, alternative Lösungen eines Problems, einer Kundenforderung usw. bezüglich der Merkmale wie der Wirksamkeit, der Realisierbarkeit usw. in administrativen und technischen Bereichen. Für die N alternativen Lösungen ergeben sich die M Paarvergleiche mit denen die Prioritäten oder Gewichtungen definiert werden. In den folgenden Abschnitten wird dieses Vorgehen erläutert.

Prioritätsdiagramme

Zur Lösung von Problemen ergeben sich oft eine Vielzahl von vorgeschlagenen Aktivitäten zur Erfüllung von Kundenforderungen. Es ist aber kaum möglich alle Maßnahmen gleichzeitig zu bearbeiten. Weitere Restriktionen ergeben sich, wenn die Zuständigkeiten woanders liegen und so die Umsetzung erschwert ist. Man wird deshalb Prioritäten setzen müssen. Das Prioritätsdiagramm hilft bei der Festlegung der weiteren Schwerpunkte. In das Prioritätsdiagramm werden die Aktivitäten in die Spalten- und Zeilenköpfe eingegeben (Tab.:1). Dann wird paarweise ein Vergleich der Aktivitäten durchgeführt.

Prioritätendiagramm						
Forderungen	Telefax frei	Sachbearbeiter verfügbar	geringste Formalitäten	direkter Weg zur Fertigung	kurze Einrichtzeiten	direkter Weg zum Versand
Telefax frei		1.00	5.00	5.00	10.00	5.00
Sachbearbeiter verfügbar	1.00		1.00	0.20	5.00	5.00
geringste Formalitäten	0.20	1.00		0.20	5.00	1.00
direkter Weg zur Fertigung	0.20	5.00	5.00		10.00	1.00
kurze Einrichtzeiten	0.10	0.20	0.20	0.10		0.10
direkter Weg zum Versand	0.20	0.20	1.00	1.00	10.00	
Summe	1.70	7.40	12.20	6.50	40.00	12.10
Summe (Rangfolge)	0.02	0.09	0.15	0.08	0.50	0.15

Tab.: 1 Beispiel eines Prioritätendiagramms

Zuerst muss ein Bewertungsmaßstab definiert werden. Dabei empfiehlt es sich mit einer Ratingskala zu arbeiten. Das Aussehen solcher Ratingskalen ist in folgender Tabelle dargestellt.

mögliche Bewertungsmaßstäbe		
numerische Bewertung	Wichtigkeit	Realisierbarkeit
0.1	viel unwichtiger	viel langsamer
0.2	unwichtiger	langsamer
1.0	gleichgewichtig	gleich schnell
5.0	wichtiger	schneller
10.0	viel wichtiger	viel schneller

Tab.: 2 Bewertungsmaßstäbe

Bei der eigentlichen Bewertung der Prioritäten muss beachtet werden, dass die Bewertung spaltenweise erfolgt, d.h., ist die Aktivität des Spaltenkopfes wichtiger (schneller) als die Zeilenaktivität wird eine 5 oder 10 vergeben. Umgekehrt wird eine 0.1 oder 0.2 vergeben. Sind beide Aktivitäten gleich wichtig notiert man eine 1. Diese Entscheidungen sind aber nur für die obere Dreiecksmatrix durchzuführen. Die untere Dreiecksmatrix wird als Kehrwert ergänzt. Ist z.B. in der oberen Dreiecksmatrix der Wert 5 eingetragen, ergibt sich spiegelbildlich in der unteren ein Wert von 0.2. Nun kann man die Bewertung von den Mitgliedern der Projektgruppe einzeln durchführen lassen und bildet dann eine akkumulierte Matrix, die zur endgültigen Priorisierung benutzt wird, oder man bildet eine gemeinsame Matrix und sucht den Konsens. Die Bewertung kann auch mit bis zu drei Kriterien durchgeführt und in einem Portfolio dargestellt werden.

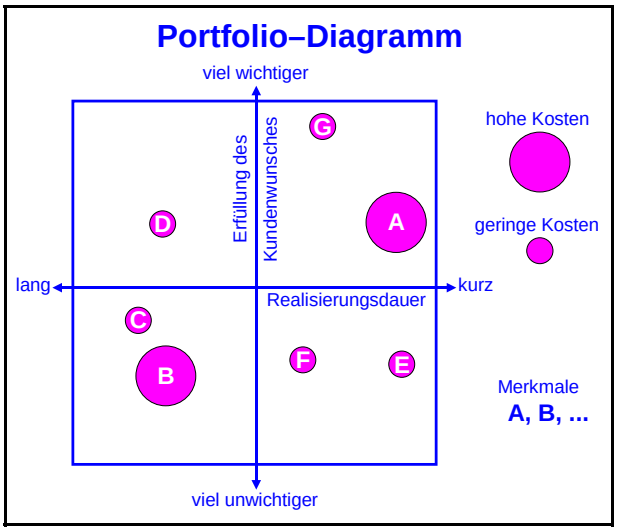

Abb.: 1 Beispiel für ein Portfolio

Das Portfolio dient der Gegenüberstellung von mehreren Betrachtungseinheiten hinsichtlich zweier oder dreier ausgewählter Dimensionen. Es hat zum Ziel, Strategien für das Positionieren eines Produktes oder einer Aktivität abzuleiten. Um ein Portfolio zu erstellen, legt man zunächst die Dimensionen fest, anhand derer die Gegenüberstellung durchzuführen ist. Diese sollen einen in Richtung und Stärke kontinuierlichen Verlauf haben. Dann positioniert man die betrachteten Objekte oder Aktivitäten mit Hilfe geeigneter Symbole. Entwicklungsmöglichkeiten und strategische Ziele lassen sich entsprechend darstellen. Dies unterstützt die Projektgruppe bei inhaltlichen Diskussionen.

Sensorische Prüfung

In der sensorischen Prüfung wurde die Methodik des paarweisen Vergleichs entwickelt und vervollkommnet. Der Urteiler muss bei jedem von M möglichen Paaren von Objekten entscheiden, welches Objekt dem jeweils anderen Objekt hinsichtlich eines vorgegebenen Merkmals überlegen bzw. vorzuziehen ist. Die Fragestellung ist immer quantitativ, kann dabei aber spezifisch (bestimmtes Merkmal) oder unspezifisch (Gesamteindruck) gestellt werden. Zum Beispiel

- **spezifisch:**
  Welches der Objekte ist hinsichtlich des zu prüfenden Merkmals intensiver, besser, reiner usw.?
- **unspezifisch:**
  Welches Objekt wird bevorzugt?

Beim Paarvergleichsurteil muss der Urteiler bei jedem von i möglichen Paaren von Objekten entscheiden, welches Objekt dem jeweils anderen Objekt hinsichtlich eines vorgegebenen Merkmals überlegen ist (Dominanzpaarvergleiche). Paarvergleichsurteile dieser Art erlauben nicht nur die Überprüfung der Urteilskonkordanz, sondern sind darüber hinaus geeignet, die Konsistenz, d.h. die Widerspruchsfreiheit der individuellen Urteile zu untersuchen.
Wie können wir entscheiden, ob ein Beurteiler wenigstens das Minimumerfordernis konsistenter Urteile erfüllt, unter der Voraussetzung natürlich, dass die Beurteilungsinstruktion auf einen eindimensionalen Aspekt eines unter Umständen komplexen Merkmals zugeschnitten wird?

Urteilskonsistenz

Angenommen, es liegen *N* Beurteilungsobjekte zur eindimensionalen Abschätzung vor. Ein auf Urteilskonsistenz zu untersuchender Beurteiler erhält dann die Instruktion, $\binom{N}{2}$ Vergleiche zwischen Paaren von Objekten (Paarvergleiche) durchzuführen und bei jedem Paarvergleich eine Präferenzschätzung mit den Abschätzungen 0 und 1 vorzunehmen. Wenn seine Präferenzen auf subjektiv realen Unterschieden zwischen den Objekten beruhen, gilt für $N = 3$ Objekte mit $A > B > C$, dass $A > B$, wenn A dem Objekt B vorgezogen wird, dass $B > C$, wenn B gegenüber C bevorzugt wird und dass $A > C$, wenn konsistenterweise A gegenüber C präferiert wird.

	1 Prüfer		A	B	C	D	E		2 Prüfer		A	B	C	D	E	
Konsistenzmaß:	1.000	A		1	1	1	1	4	1.000	A		1	1	1	1	4
Freiheitsgrad:	60.000	B	0		0	0	1	1	60.000	B	0		0	0	1	1
Chi ² prüf:	84.000	C	0	1		1	1	3	84.000	C	0	1		0	1	2
Chi ² tab:	79.082	D	0	1	0		1	2	79.082	D	0	1	1		1	3
		E	0	0	0	0		0		E	0	0	0	0		0

	3 Prüfer		A	B	C	D	E		4 Prüfer		A	B	C	D	E	
Konsistenzmaß:	1.000	A		1	0	1	1	3	0.600	A		1	1	1	0	3
Freiheitsgrad:	60.000	B	0		0	0	0	0	60.000	B	0		0	0	0	0
Chi ² prüf:	84.000	C	1	1		1	1	4	68.000	C	0	1		0	1	2
Chi ² tab:	79.082	D	0	1	0		1	2	79.082	D	0	1	1		1	3
		E	0	1	0	0		1		E	1	1	0	0		2

Tab.: 3 Matrizen von vier Prüfern

Zur Veranschaulichung lassen wir die 3 Objekte die Ecken eines gleichseitigen Dreiecks bilden, in dessen Seiten die Größer-Kleiner-Relationen als Pfeile eingelassen sind, wobei wir z.B. für $A \leftarrow B$ vereinbaren, dass A dem Objekt B vorgezogen wird $(A > B)$. Wenn alle 3 Präferenzen transitiv, d.h. die Vergleichsurteile konsistent sind, dann sind die Pfeile nicht kreisförmig (azirkulär) angeordnet, und es resultiert das Schema der Abb.: 2a. Wenn die Präferenzen hingegen intransitiv und damit die Vergleichsurteile inkonsistent sind, ordnen sich die Pfeile entsprechend einer gerichteten Kreislinie an (zirkulär), und es resultiert das Schema der Abb. 2b.

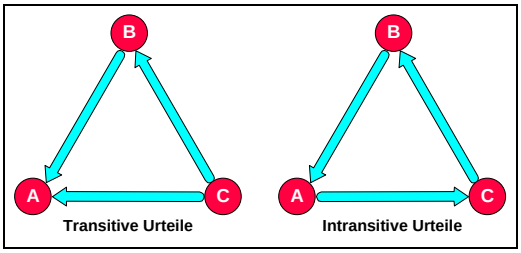

Abb.: 2 Paarvergleichsurteile für 3 Objekte

Betrachten wir statt dreier Objekte N Objekte, so lässt sich ein Maß für den Grad der Inkonsistenz der Paarvergleichsurteile aus der Zahl der zirkulären Triaden gewinnen, die man leicht erkennt, wenn man alle Ecken eines regelmäßigen N-Ecks durch gerichtete Pfeile verbindet. Je größer die Zahl der zirkulären Triaden, um so geringer ist die Konsistenz der Urteile. Um einen für einen bestimmten Beurteiler geltenden Konsistenzkoeffizienten zu definieren, müssen wir fragen, wie groß die Zahl der zirkulären Triaden bei N Objekten maximal werden kann. Betrachten wir dazu ein Beispiel mit $N = 5$ Vergleichsobjekten. Die Gesamtzahl aller Dreiecke (Triaden) beträgt $\binom{N}{3} = \binom{5}{3} = 10$. Unsere Annahme ist, dass 5 Triaden zirkulär bzw. Inkonsistent sind, und zwar die Triaden *ABD, ABE, ACD, BCE* und *CDE*. Wollten wir versuchen, die Zahl der zirkulären Triaden dadurch zu erhöhen, dass wir eine konsistente (azirkuläre) Triade (z.B. die Triade *ABC*) inkonsistent (zirkulär) machen, indem wir die Relation etwa zwischen A und B umkehren, so würden dadurch die inkonsistenten Triaden *ABD* und *ABE* konsistent werden. Die inkonsistenten Triaden wären also nicht, wie beabsichtigt, um eine vermehrt, sondern um eine vermindert.
Es lässt sich zeigen, dass sich die Zahl der inkonsistenten Triaden nicht vermehren, sondern nur vermindern lässt. Daraus ergibt sich, dass hier ein Maximum inkonsistenter Triaden vorliegt. Ein Beurteiler, der durch seine Paarvergleiche diese Maximalzahl inkonsistenter Triaden erreicht hat, urteilt also absolut inkonsistent; seinem Urteil entspräche ein Konsistenzkoeffizient K von Null, wenn man K von 0 bis 1 variieren lässt. Dass die Minimalzahl inkonsistenter Triaden gleich Null sein muss, ergibt sich intuitiv aus der Tatsache, dass eine subjektiv eindeutig differenzierende Rangordnung der 5 Objekte, z.B. $A < B < C < D < E$, nur konsistente Triaden zur Folge haben kann. Bezeichnen wir die beobachtete Zahl der inkonsistenten (zirkulären) Triaden eines Beurteilers mit d und die Maximalzahl inkonsistenter Triaden mit d_{max}, lässt sich ein individueller Konsistenzkoeffizient K dadurch definieren, dass man den Anteil der beobachteten an den höchstmöglichen Zirkulärtriaden von 1 subtrahiert:

$$K = 1 - \frac{d}{d_{\max}}$$

Bei fehlenden zirkulären Triaden ist $K = 1$, und bei maximaler Anzahl resultiert $K = 0$. Wie sich zeigen lässt, gelten für gerad- und ungeradzahlige N die folgenden Gleichungen für d_{max}.

$$d_{\max} = \frac{N^3 - N}{24} \quad (N \textit{ ungeradzahlig})$$

$$d_{\max} = \frac{N^3 - 4 \cdot N}{24} \quad (N \textit{ geradzahlig})$$

Für $N = 5$ erhalten wir also den bereits ausgezählten Wert von $d_{max} = 5$; wäre $d = 4$ gewesen, hätte sich $K = 1 - 4/5 = 0,2$ ergeben, was einer nur geringen Urteilskonsistenz entspricht. Setzen wir die Bestimmungsgleichungen für d_{max} in die Gleichung für den Koeffizienten ein, resultiert

$$K = 1 - \frac{24 \cdot d}{N^3 - N} \quad (N \text{ ungeradzahlig})$$

$$K = 1 - \frac{24 \cdot d}{N^3 - 4 \cdot N} \quad (N \text{ geradzahlig})$$

Bei einer größeren Zahl von Beurteilungsobjekten - etwa mehr als 6 - wird es sehr aufwendig, ein graphisches Verfahren zur Abzählung der Zirkulärtriaden heranzuziehen. Man geht hier zweckmäßigerweise so vor, dass man die Ergebnisse der $\binom{N}{2}$ Paarvergleiche in einer quadratischen, nicht symmetrischen Matrix zur Darstellung bringt. Wie dies geschieht, wird in Tab.: 4 verdeutlicht.

	A	B	C	D	E	F	
A		1	1	1	1	1	5
B	0		0	1	0	1	2
C	0	1		0	0	0	1
D	0	0	1		1	1	3
E	0	1	1	0		1	3
F	0	0	1	0	0		1

Tab.: 4 Beispiel einer Bewertung

Die Tab.: 4 zeigt die Resultate von $\binom{6}{2} = 15$ Paarvergleichsurteilen für $N = 6$ Objekte. Es wurde eine 1 signiert, wenn ein die Zeilen kennzeichnendes Objekt (kurz: Zeilenobjekt) einem die Spalten kennzeichnendes Objekt (kurz: Spaltenobjekt) vorgezogen wurde. Dieser Regel folgend, ist Tab.: 4 z.B. zu entnehmen, dass Objekt *A* über alle anderen Objekte dominiert, dass Objekt *E* dem Objekt *B* vorgezogen wird etc.. Bildet man nun die Zeilensummen der Präferenzen und bezeichnet diese mit $S_i (i = 1, ..., N)$, lässt sich zeigen, dass die Zahl der inkonsistenten Triaden algebraisch gegeben ist durch

$$d = \frac{N(N-1)(2N-1)}{12} - \frac{1}{2} \sum_{i=1}^{N} S_i^2$$

Diesen - stets ganzzahligen - Wert setzen wir in die Gleichung für den Koeffizienten ein und erhalten so auf einfachste Weise den gesuchten Konsistenzkoeffizienten $K = 0.625$.
Ein Test zur Beantwortung der Frage, ob ein Konsistenzkoeffizient *K* überzufällig hoch oder überzufällig niedrig ist, muss prüfen, ob die Zahl der Zirkulärtriaden d größer (oder geringer) ist als die Anzahl der Zirkulärtriaden, die man nach Zufall - z.B. durch Paarvergleich nach Münzwurf - erwarten kann.
Bei *N* Vergleichsobjekten sind $\binom{N}{2}$ Paarvergleiche möglich. Jeder dieser Vergleiche kann mit einer Wahrscheinlichkeit von 1/2 positiv oder negativ ausfallen, wenn die Nullhypothese einer Zufallsentscheidung gilt. Versieht man die $\binom{N}{2}$ Verbindungen zwischen den Eckpunkten eines *N*-Ecks, mit Pfeilzeichen, so ergeben sich

$$M = 2^{\binom{N}{2}}$$

Pfeilzeichenpermutationen. Für all diese Permutationen bleibt die Gesamtzahl der Triaden gleich, nämlich $\binom{N}{3}$, nicht jedoch die Zahl d der Zirkulärtriaden, die von den jeweiligen Pfeilrichtungen abhängt. Kombinatorisch lässt sich nun ermitteln, wieviele der *M* Pfeilrichtungen $d = d_{max}$ Zirkulärtriaden liefern. Die Summe aller Permutationen ist

$$M = 2^{\binom{n}{2}} = 2^{\binom{5}{2}} = 2^{10} = 1024$$

Wir hatten genau $d = 5$ Zirkulärtriaden ermittelt, und wir können nun fragen, wie groß die Punktwahrscheinlichkeit *p* ist, genau 5 Zirkulärtriaden bei Geltung von H_0 zu finden. Die Ant-

wort lautet $p = 24/1024 = 0.023$. Die Überschreitungswahrscheinlichkeit P, 5 oder mehr Zirkulärtriaden zu finden, beträgt natürlich ebenfalls $P = 0.023$, da mehr als 5 Zirkulärtriaden, wie festgestellt, nicht möglich sind. Legt man $\alpha = 0.05$ zugrunde, wären die $d = 5$ zirkulären Triaden bzw. $K = 0$ bei $\binom{5}{2} = 10$ Paarvergleichsurteilen als signifikant inkonsistentes Urteilsverhalten zu interpretieren.
Wie man sieht, ist die H_0-Verteilung der Prüfgröße d so kumuliert worden, dass die Überschreitungswahrscheinlichkeiten auf Urteilsinkonsistenz statt auf Urteilskonsistenz gerichtet sind. Will man auf Urteilskonsistenz prüfen, was die Regel ist, muss man $P^* = 1-P(d+1)$ bilden und dieses P^* mit dem vorgegebenen α-Risiko vergleichen. Für $d = 1$, $N = 5$ und damit $K = 0.8$ ergibt sich $P^* = 1-0.766 = (120+120)/1024 = 0.234$. Diese Konsistenz wäre nicht signifikant. Ausgehend vom Erwartungswert für d

$$E(d) = \frac{1}{4} \cdot \binom{N}{3}$$

weisen Ergebnisse mit $d > E(d)$ auf tendenziell inkonsistente und Ergebnisse mit $d < E(d)$ auf konsistente Urteile hin. Für größere Stichproben von Beurteilungsobjekten ($N > 8$) ist die folgende Funktion der Prüfgröße d asymptotisch χ^2-verteilt:

$$\chi^2 = \frac{8}{N-4} \cdot \left[\frac{1}{4} \cdot \binom{N}{3} - d + \frac{1}{2} \right] + Fg$$

wobei

$$Fg = \frac{N(N-1)(N-2)}{(N-4)^2}$$

zugrunde zu legen sind. Man erkennt, dass der χ^2-Wert mit wachsendem d sinkt, d.h. kleine χ^2-Werte ($\chi^2 < \chi^2_{0.05}$) sprechen für inkonsistente und große χ^2-Werte ($\chi^2 > \chi^2_{0.95}$) für konsistente Urteile. Der asymptotische Test ist gegenüber dem exakten Test eher progressiv. Signifikanztests zur Überprüfung der Konsistenz von Paarvergleichsurteilen setzen voraus, dass die Paarvergleichsurteile voneinander unabhängig sind. Man sollte deshalb zumindest dafür Sorge tragen, dass die Abfolge der Paarvergleiche zufällig ist.

Urteilskonkordanz

Mit der Konsistenzanalyse überprüfen wir, inwieweit ein einzelner Urteiler widerspruchsfreie Paarvergleichsurteile abgegeben hat. Werden die Paarvergleiche nun von m Beurteilern durchgeführt, stellt sich die Frage, wie gut diese Paarvergleichsurteile übereinstimmen. Diese Frage zu beantworten ist Aufgabe der Konkordanzanalyse.
Bereits an dieser Stelle können wir vorwegnehmen, dass eine hohe Konkordanz der Urteile durchaus möglich ist, obwohl die einzelnen Urteiler jeder für sich inkonsistent geurteilt haben. Dies wäre beispielsweise der Fall, wenn die Urteiler einheitlich bei bestimmten Paarvergleichen das Urteilskriterium wechseln. Selbstverständlich kann natürlich auch mangelnde Konsistenz zu einer mangelnden Konkordanz führen - ein Ergebnis, mit dem man eher bei urteilerspezifischen Urteilsfehlern rechnen würde. Auf der anderen Seite bedeutet eine durchgehend bei allen Beurteilern festgestellte perfekte Konsistenz keineswegs gleichzeitig eine hohe Urteilskonkordanz. Eine perfekte Konsistenz erhalten wir, wenn $d = 0$ ist, wenn also keine zirkulären Triaden auftreten. Wie jedoch zeigt, gibt es bei gegebenem N jeweils mehrere Urteilskonstellationen mit $d = 0$, d.h. verschiedene Urteiler können trotz hoher individueller Konsistenz diskordant urteilen. Mit diesem Ergebnis wäre zu rechnen, wenn jeder Urteiler seine Paarvergleichsurteile konsistent nach einem anderen Kriterium abgeben würde und die verschiedenen benutzten Kriterien wechselseitig schwach oder gar nicht korreliert sind.
Zusammengenommen ist also davon auszugehen, dass Konsistenz und Konkordanz bei Paarvergleichsurteilen zumindest theoretisch zwei von einander unabhängige Konzepte darstellen. Wir wollen das Prinzip der Konkordanzbestimmung und -überprüfung bei Paarvergleichsprüfungen

anhand der Daten in Tab.: 5 verdeutlichen. Dabei mag es sich um die Paarvergleichsurteile von vier Graphologen handeln, die bei jeweils 2 von $N = 5$ Probanden zu entscheiden haben, welcher der beiden Probanden aufgrund seiner Handschrift vermutlich der intelligentere (A) sei.

	1 Prüfer		A	B	C	D	E		2 Prüfer		A	B	C	D	E	
Konsistenzmaß:	1.000	A		1	1	1	1	4	1.000	A		1	1	1	1	4
Freiheitsgrad:	60.000	B	0		0	0	1	1	60.000	B	0		0	0	1	1
Chi ² prüf:	84.000	C	0	1		1	1	3	84.000	C	0	1		0	1	2
Chi ² tab:	79.082	D	0	1	0		1	2	79.082	D	0	1	1		1	3
		E	0	0	0	0		0		E	0	0	0	0		0

	3 Prüfer		A	B	C	D	E		4 Prüfer		A	B	C	D	E	
Konsistenzmaß:	1.000	A		1	0	1	1	3	0.600	A		1	1	1	0	3
Freiheitsgrad:	60.000	B	0		0	0	0	0	60.000	B	0		0	0	0	0
Chi ² prüf:	84.000	C	1	1		1	1	4	68.000	C	0	1		0	1	2
Chi ² tab:	79.082	D	0	1	0		1	2	79.082	D	0	1	1		1	3
		E	0	1	0	0		1		E	1	1	0	0		2

Tab.: 5 Graphologische Beurteilung

Wie man sieht, urteilen die Graphologen 1, 2 und 3 voll konsistent ($K = 1$), der Graphologe 4 hingegen nur partiell konsistent ($K = 0.6$). Drei von vier sind sich also ihrer Sache sicher. Jeder von ihnen urteilt konsistent nach jenem Aspekt des Schriftbildes, das er für intelligenzrelevant hält. Ob dies für alle drei bzw. vier Beurteiler derselbe Aspekt ist, wollen wir mit Hilfe der folgenden Überlegungen klären. Um die Übereinstimmung der vier Konsistenztabellen in Tab.: 6 zu beurteilen, legen wir sie gewissermaßen übereinander und summieren pro Zelle die Eins: Bei $N = 4$ Urteilern müssen sich damit diagonalsymmetrische Felder zu 4 ergänzen.

	Gesamt		A	B	C	D	E	
Akkordanz	0.533	A		4	3	4	3	14
Freiheitsgrad:	30.000	B	0		0	0	2	2
Chi ² prüf:	62.000	C	1	4		2	4	11
Chi ² tab:	43.773	D	0	4	2		4	10
		E	1	2	0	0		3

Tab.: 6 Berechnung zur Konkordanz.

Diesen Häufigkeiten (f_{ij}) ist zu entnehmen, wieviele Urteilerpaare übereinstimmende Paarvergleichsurteile abgegeben haben. Es wurde beispielsweise von vier Urteilern das Urteil $A > B$ abgegeben, d.h. $\binom{4}{2} = 6$ Urteilerpaare waren sich in der vergleichenden Einschätzung der Intelligenz von A und B einig. Allgemein ergeben sich pro Zelle $\binom{f_{ij}}{2}$ Urteilerpaare, die für das Beispiel in Tab.: 7 zusammengestellt sind.

	A	B	C	D	E	
A		6	3	6	3	18
B	0		0	0	1	1
C	0	6		1	6	13
D	0	6	1		6	13
E	0	1	0	0		1
Übereinstimmungen						46

Tab.: 7 Übereinstimmende Urteilerpaare

Als Anzahl der übereinstimmenden Urteilerpaare J ermitteln wir für das Beispiel $J = 46$. Ausgehend von den Häufigkeiten f_{ij} in Tab.: 6 errechnet man diesen Wert allgemein nach der Gleichung:

$$J = \sum_{i=1}^{N}\sum_{j=1}^{N}\binom{f_{ij}}{2} = \frac{1}{2}\cdot\sum_{i=1}^{N}\sum_{j=1}^{N} f_{ij}^2 - \binom{N}{2}\cdot\frac{m}{2} \quad (i \neq j)$$

Um J auf N und m zu relativieren, definiert man das folgende Übereinstimmungsmaß, den wir als Akkordanzmaß A bezeichnen wollen:

$$A = \frac{J - \frac{1}{2} \cdot \binom{N}{2} \cdot \binom{m}{2}}{\frac{1}{2} \cdot \binom{N}{2} \cdot \binom{m}{2}}$$

In der Gleichung ist

$$E(J) = \frac{1}{2} \cdot \binom{N}{2} \cdot \binom{m}{2}$$

der Erwartungswert für die Zahl der Urteilerpaare, die unter der Nullhypothese rein zufällig identisch urteilen. Für unser Beispiel mit $N = 5$ Handschriften und $m = 4$ Graphologen beträgt für $J = 46$ die Akkordanz $A = 0.53$. Wir sind bislang von eindeutigen Präferenzen im Paarvergleichsurteil ausgegangen, d.h. Gleichurteile waren ausgeschlossen. Lässt man Gleichurteile zu, können diese mit Gleichheitszeichen symbolisiert werden und mit je 1/2 Punkt in die Berechnung von J gemäß Gleichung eingehen. Um zu einem exakten Test für den Akkondanzkoeffizienten A zu gelangen, müssen wir folgende Überlegung anstellen: Ein Beurteiler hat (bei unzulässigen Gleichurteilen) zwei Möglichkeiten, eine von zwei zum Paarvergleich gebotene Handschriften zu bevorzugen ($A > B$ oder $A < B$). Für jede dieser zwei Möglichkeiten hat ein anderer Beurteiler wiederum zwei Möglichkeiten der Präferenz usw. Wenn wir m von einander unabhängige Beurteiler heranziehen, gibt es $2m$ Möglichkeiten der Präferenz für ein einzelnes Schriftenpaar. Sofern wir aber, wie in unserem Beispiel, nicht mit zwei Schriften, sondern mit $N = 5$ Schriften operieren, aus denen wir $\binom{N}{2}$ Schriftenpaare bilden müssen, ergeben sich insgesamt

$$T = (2^m)^{\binom{N}{2}} = 2^{mN(N-1)/2}$$

Präferenzmöglichkeiten von der Art, wie sie als Matrix in Tab.: 5 dargestellt wurden. Aus jeder dieser T Tabellen resultiert ein J-Wert, den wir als Prüfgröße definieren. Die Verteilung dieser J-Werte ergibt die tabellierte Prüfverteilung von J. Für die $N = 5$ Handschriften und $m = 4$ Graphologen erhalten wir zu einem $J = 46$ ein $P = 0.00041$ ab. Damit ist der Akkordanzkoeffizient von $A = 0.53$ selbst bei einer Signifikanzforderung von 0.1 % gesichert. Wenn N oder m größer ist, prüft man asymptotisch über χ^2-Verteilung unter Benutzung von J:

$$\chi^2 = \frac{4}{m-2} \cdot \left(J - \frac{1}{2} \cdot \binom{N}{2} \cdot \binom{m}{2} \cdot \frac{m-3}{m-2} \right)$$

mit

$$Fg = \binom{N}{2} \cdot \frac{m(m-1)}{(m-2)^2}$$

Für weniger als $Fg = 30$ empfiehlt es sich, mit Kontinuitätskorrektur zu testen und J durch $J' = J-1$ zu ersetzen. Auf unser Handschriftenbeispiel angewendet, ergibt sich für $N = 5$ und $m = 4$ mit $J' = 45$. Dieser χ^2-Wert ist auf der 0.1%-Stufe signifikant; er entspricht dem Ergebnis des exakten Akkordanztests.

Bilden einer Empfindungsskala

Das Ziel ist die Bildung einer Empfindungsskala, welche die Ergebnisse der Dominanzmatrix (Tab.: 6) auf einen Wertebereich von -1 bis 1 darstellt und die Abstände untereinander den empfundenen Abständen entsprechen. Die Voraussetzung dafür ist, dass der Abstand zweier Beurteilungseinheiten auf der Skala proportional dem Beurteilungsüberschuss ist, den eine Beurteilungseinheit gegenüber einer anderen hat. Diese Voraussetzung ist nur erfüllt, wenn sich die Beurteiler in der Bevorzugung einer Beurteilungseinheit gegenüber einer anderen nicht alle einig sind.

Dominanzmatrix					
	A	B	C	D	E
A		4	3	4	3
B	0		0	0	2
C	1	4		2	4
D	0	4	2		4
E	1	2	0	0	

Distanzmatrix						
	A	B	C	D	E	Skalenwerte
A		1.0	0.5	1.0	0.5	0.750
B	-1.0		-1.0	-1.0	0.0	-0.750
C	-0.5	1.0		0.0	1.0	0.375
D	-1.0	1.0	0.0		1.0	0.250
E	-0.5	0.0	-1.0	-1.0		-0.625

Tab.: 8 Erstellung der Distanzmatrix

Um die Distanzmatrix zu erstellen werden die Differenzen AB-BA, AC-CA, AD-DA, AE-EA, BC-CB usw. gebildet und durch die Anzahl der Beurteiler dividiert. Dieses muss nur für die obere Dreiecksmatrix durchgeführt werden, weil die untere Dreiecksmatrix gleiche Werte mit allerdings anderem Vorzeichen hat. Die Addition der Werte von unterer und oberer Dreiecksmatrix müssen für die Distanzmatrix Null ergeben. Aus der Distanzmatrix können nun die Skalenwerte als Zeilenmittelwerte berechnet werden. Die Summe aller Mittelwerte ergibt wieder Null.

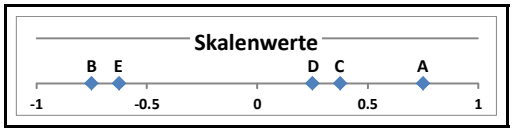

Abb.: 3 Empfindungsskala

Signifikanz der Skalenwerte

Will man prüfen welcher Skalenwert sich von einem anderen Skalenwert signifikant unterscheidet, dann können wir einen *t*-Test verwenden. Dazu muss als erstes die Varianz der Skalenwerte bestimmt werden. Man bildet die Matrix der reproduzierten Distanzen unter Verwendung der Skalenwerte.

reproduzierte Distanzmatrix						
	A	B	C	D	E	Skalenwerte
A		1.500	0.375	0.500	1.375	0.750
B	-1.500		-1.125	-1.000	-0.125	-0.750
C	-0.375	1.125		0.125	1.000	0.375
D	-0.500	1.000	-0.125		0.875	0.250
E	-1.375	0.125	-1.000	-0.875		-0.625

Tab.: 9 Erstellung der reproduzierten Distanzmatrix

Die obere Dreiecksmatrix wird berechnet, in dem folgender Algorithmus angewendet werden muss.

$$\bar{x}_A - \bar{x}_B = AB, \bar{x}_A - \bar{x}_C = AC, \bar{x}_A - \bar{x}_D = AD, \bar{x}_A - \bar{x}_E = AE, \bar{x}_B - \bar{x}_C = BC, \text{ usw.}$$

Danach wird die untere Dreiecksmatrix mit den selben Werten allerdings mit anderem Vorzeichen ergänzt. Nun können die Residuen berechnet werden, indem von den Werten der reproduzierten Distanzmatrix die Werte der Distanzmatrix abgezogen werden. Es ergibt sich die Residuengrafik und die quadrierte Residuengrafik.

Residuenmatrix					
	A	B	C	D	E
A		0.500	-0.125	-0.500	0.875
B	-0.500		-0.125	0.000	-0.125
C	0.125	0.125		0.125	0.000
D	0.500	0.000	-0.125		-0.125
E	-0.875	0.125	0.000	0.125	

quadrierte Residuenmatrix					
	A	B	C	D	E
A		0.250000	0.015625	0.250000	0.765625
B	0.250000		0.015625	0.000000	0.015625
C	0.015625	0.015625		0.015625	0.000000
D	0.250000	0.000000	0.015625		0.015625
E	0.765625	0.015625	0.000000	0.015625	

Tab.: 10 Erstellung der Residuengrafiken

Als nächstes wird die Quadratsumme (*QS = 2.6875*) berechnet. Die Gesamtvarianz ergibt sich mit

$$s^2 = \frac{QS}{(k-1)(k-2)}$$ k = Anzahl der Beurteilungseinheiten

Der Freiheitsgrad (*f = 6*) wird berechnet aus *f = [(k-1)(k-2)]/2*. Die Quadratsumme wurde durch die doppelte Anzahl von Freiheitsgraden geteilt, weil die Anzahl der Messwerte verdoppelt wurde, um die Matrix zu füllen. Die Varianz (s^2_{mD}) der mittleren Differenz zweier Skalenwerte hängt mit der Gesamtvarianz zusammen.

$$s_{mD}^2 = \frac{2 \cdot s}{k}$$

Nun können für alle Differenzen (*10 = [k(k-1)]/2*) *t*-Werte berechnet werden. Dies geschieht in dem man die Differenzen durch die Standardabweichung der mittleren Differenzen dividiert. Aus den *t*-Werten berechnet man die *p*-Werte.

$$t_{(AB)} = \frac{|\bar{x}_A - \bar{x}_B|}{s_{mD}}$$

$$p_{(AB)} = t_{t_{(AB)},f}^{-1}$$

Die *p*-Werte müssen noch angepasst werden, weil bei multiplen Mittelwertvergleichen die vorgegebene Schranke ($\alpha = 0.05$) nicht eingehalten wird. Für den Fall, dass die m Tests nicht unabhängig voneinander sind, haben Tukey, Ciminera und Heyse eine Variante der Basiskorrektur von Sidak vorgeschlagen.

$$\hat{p}_i = 1 - (1 - p_i)^{\sqrt{m}}$$

Nach Korrektur ergibt sich für das Beispiel folgende Tabelle.

Vergleich	t	p-Wert	adj. p-Wert
AB	5.011614	0.001205	**0.003805**
AC	1.252904	0.169682	0.444571
AD	1.670538	0.100541	0.284720
AE	4.593980	0.001953	**0.006164**
BC	3.758711	0.005535	**0.017399**
BD	3.341076	0.009669	**0.030257**
BE	0.417635	0.346209	0.739165
CD	0.417635	0.346209	0.739165
CE	3.341076	0.009669	**0.030257**
DE	2.923442	0.017249	0.053536

Tab.: 11 Multiple Mittelwertvergleiche

Literatur

Retzlaff, G., Rust, G., Waibel, J.:
Statistische Versuchsplanung, Verlag Chemie, 2. Auflage 1978

Bortz, J., Lienert, G. A., Boehnke, K.:
Verteilungsfreie Methoden in der Biostatistik, Springer Verlag, 4. Auflage 1990

Lüpsen, H.:
Multiple Mittelwertvergleiche, Universität zu Köln, Internet 2014