Zur Frage der Fehlergrenzen bei der Leistungsbewertung gemäss UfAB-Methodik

Dipl. Geophysiker
Heribert Genreith*

30. September 2003

1 Einleitung

Bei der Bewertung der Leistung gemäss UfAB-Methodik ist es von erheblicher Bedeutung in welcher Grössenordnung die Fehler der gemäss Bewertungstabelle anfallenden Leistungsbewertung L ist.

Grundsätzlich zu unterscheiden ist zwischen systematischen und statistischen Fehlern. Systematische Fehler gehen immer in dieselbe Richtung und haben daher eine geringere oder fehlende Tendenz zur Ausmittelung. Zudem sind sie, da meist unentdeckt, kaum kalkulierbar. Andererseits sollten bei korrekter Anwendung der UfAB i.V.m. der VOL eben keine solche systematischen Fehler vorhanden sein. Statistische Fehler dagegen sind unvermeidbar, jedoch besitzen sie die Tendenz zur Ausmittelung. Wir gehen daher davon aus, dass keine systematischen Fehler bei der Bewertung gemacht werden, so dass eine statistische Schwankung der Bepunktungen um den korrekten Wert stattfindet.

2 Statistische Fehlerfortpflanzung

Die Bewertung von Angeboten erfolge nach einem Bewertungsschema mit n bepunkteten Fragen. Jede Frage kann gleichermassen mit maximal l_{max} Punkten bewertet werden. Jedes Kriterium hat ein bestimmtes Gesamtgewicht von α_i an der Gesamtbewertung L und wird mit $l \leq l_{max}$ bewertet:

$$L = \sum_{i=1}^{n} \alpha_i l_i$$

*Beschaffungsamt des BMI, St. Augustiner Str. 86, 53225 Bonn; Ref. B2; Az.: B2.30-0002/02/UfABIII-BewF

Dann kann die Gauss'sche Fehlerfortpflanzung angesetzt werden:

$$\Delta L = \sqrt{\sum_{i=1}^{n} (\frac{\partial L}{\partial l_i} \Delta l_i)^2} = \sqrt{\sum_{i=1}^{n} (\alpha_i \Delta l_i)^2}$$

Der relative Fehler L_{rel} ist dann

$$L_{rel} = \frac{\Delta L}{L} = \frac{\sqrt{\sum_{i=1}^{n} (\alpha_i \Delta l_i)^2}}{\sum_{i=1}^{n} \alpha_i l_i}$$

Dieser Fehler wäre nun im Einzelfall zu berechnen.

2.1 Corollar: Skalenunabhängigkeit

Der relative Fehler der Leistungsbewertung $L_{rel} = \frac{\Delta L}{L} = \frac{\sqrt{\sum_{i=1}^{n} (\alpha_i \Delta l_i)^2}}{\sum_{i=1}^{n} \alpha_i l_i}$ ist skalenunabhängig.

Annahme: man führe andere Gewichte ein (z.B. mehrere Haupt- und Untergruppengewichte), so dass

$$L = \sum_{i=1}^{n} \prod_{j=1}^{k} \alpha_{ij} l_i$$

gilt. Dann ist der relative Fehler:

$$L_{rel} = \frac{\Delta L}{L} = \frac{\sqrt{\sum_{i=1}^{n} (\prod_{j=1}^{k} \alpha_{ij} \Delta l_i)^2}}{\sum_{i=1}^{n} \prod_{j=1}^{k} \alpha_{ij} l_i}$$

Mit der Identität

$$\alpha_i \equiv \prod_{j=1}^{k} \alpha_{ij} = const.$$

gelten dann alle hier aufgeführten Abschätzung weiterhin.

In der hier vorgelegten Analyse werden der Einfachheit halber immer auf den Bereich $\alpha_i \in [0, 1]$ normierte Gewichte verwendet, die Gültigkeit ist aber auf diese nicht beschränkt.

3 Näherungsformel des statistischen Fehlers

Für eine Abschätzung jedoch reicht die Annahme dass alle relevanten Grössen von der gleichen durchschnittlichen Grössenordnung seien:

$$\alpha := \overline{\alpha_i} \qquad \Delta l := \overline{\Delta l_i} \qquad l := \overline{l_i}$$

Dann lässt sich der relative Fehler schreiben als:

$$L_{rel} = \frac{\Delta L}{L} \approx \frac{\sqrt{n(\alpha\Delta l)^2}}{n\alpha l} = \frac{\Delta l}{\sqrt{n} l}$$

Diese Abschätzung ist unabhängig von der Natur der Gewichte α. Für die durchschnittliche Bewertung eines Kriteriums lässt sich zudem $l \approx l_{max}/2$ ansetzen, so dass sich

$$L_{rel} = \frac{\Delta L}{L} \approx \frac{2\Delta l}{\sqrt{n} l_{max}}$$

schreiben lässt. Setzt man nun folgende Werte

$$l_{max} = 5 \qquad \Delta l = \pm 1 \qquad n = 25$$

ein, so ergibt sich als relativer Fehler

$$L_{rel} \approx \frac{2\Delta l}{\sqrt{n} l_{max}} = \pm \frac{2}{\sqrt{25} \cdot 5} = \pm 0.08 = \pm 8\%$$

Dieser relative Fehler ist umso kleiner, je höher die Anzahl der bewerteten Kriterien n ist und je feiner die Auflösung der Kriterienbepunktung l_{max} ist und umso höher, je grösser die durchschnittliche Unsicherheit bei der Bewertung der Kriterien Δl ist. In der Praxis hängen jedoch l_{max} und Δl eng zusammen, denn je feiner die Auflösung der Kriterien ist, desto grösser ist i.a. auch die Unsicherheit der Bewerter bei der Vergabe der Punkte.

3.1 Corollar: Minimaler statistischer Fehler

Bei obiger Abschätzung handelt es sich um einen minimalen statistischen Fehler: Bei n Kriterien ist dass durchschnittliche Gewicht $\alpha = 1/n$. Nun sei aber o.E.d.A. ein Kriterium mit einem höheren Gewicht $\alpha_0 = p > \alpha = \frac{1-p}{n-1}$ versehen. Dann ist die Leistungsbewertung

$$L = \alpha_0 l_0 + \sum_{i=1}^{n-1} \alpha_i l_i$$

und der relative Fehler unter Einbeziehung obiger Näherungen demnach

$$L_{rel} = \frac{\sqrt{\alpha_0^2 \Delta l_0^2 + (n-1)\alpha^2 \Delta l^2}}{\alpha_0^2 l_0^2 + (n-1)\alpha^2 l^2} = \frac{\Delta l}{l} \cdot \frac{\sqrt{1 + (n-1)(\frac{\alpha}{\alpha_0})^2}}{1 + (n-1)\frac{\alpha}{\alpha_0}}$$

Es gilt zudem

$$0 < \frac{\alpha}{\alpha_0} < 1$$

und damit weiterhin

$$\frac{\sqrt{1 + (n-1)(\frac{\alpha}{\alpha_0})^2}}{1 + (n-1)\frac{\alpha}{\alpha_0}} > \frac{\sqrt{1 + (n-1) \cdot 0}}{1 + (n-1)\frac{\alpha}{\alpha_0}} > \frac{1}{\sqrt{1 + (n-1) \cdot 1}} = \frac{1}{\sqrt{n}}$$

Anderseits möge $\alpha_0 = p < \alpha = \frac{1-p}{n-1}$ gegeben sein. Dann gilt

$$\frac{\alpha}{\alpha_0} > 1$$

und damit weiterhin

$$\frac{\sqrt{1+(n-1)(\frac{\alpha}{\alpha_0})^2}}{1+(n-1)\frac{\alpha}{\alpha_0}} > \frac{\sqrt{1+(n-1)\cdot 1}}{1+(n-1)\frac{\alpha}{\alpha_0}} > \frac{\sqrt{n}}{\sqrt{1+(n-1)\cdot 0}} = \frac{\sqrt{n}}{1} \geq \frac{1}{\sqrt{n}}$$

Somit handelt es sich bei der hier gezeigten Abschätzung um den minimalen statistischen Fehler. *Der tatsächliche Fehler liegt daher i.d.R. etwas höher.*

4 Der Fehler bei mehreren unabhängigen Bewertungen

Zur weiteren Minimierung des statistischen Fehlers können Angebote durch mehrere *unabhängige Begutachter* bewertet werden. Der Leistungswert L wird dann durch das arithmetische Mittel der k unabhängigen Bewertungen

$$\overline{L} = \frac{1}{k}\sum_{j=1}^{k} L_j = \sum_{j=1}^{k}\sum_{i=1}^{n}\frac{\alpha_i}{k} l_{ij}$$

gebildet. Der relative Fehler ist dann

$$L_{rel} = \frac{\Delta\overline{L}}{\overline{L}} = \frac{\sqrt{\sum_{j=1}^{k}\sum_{i=1}^{n}(\frac{\alpha_i}{k}\Delta l_{ij})^2}}{\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{\alpha_i}{k} l_{ij}}$$

Mit obiger Näherung ergibt dies

$$L_{rel} \approx \frac{\sqrt{kn\frac{\alpha^2}{k^2}\Delta l^2}}{kn\frac{\alpha}{k} l} = \frac{\Delta l}{\sqrt{nk}l} \approx \frac{2\Delta l}{\sqrt{nk}l_{max}}$$

sodass sich der Fehler weiter um den Faktor $\frac{1}{\sqrt{k}}$ verringert. Bei drei unabhängigen Gutachtern reduziert sich im weiter oben aufgeführten Beispiel der Fehler von 8% auf nur noch

$$l_{max} = 5 \qquad \Delta l = \pm 1 \qquad n = 25 \qquad k = 3$$

$$L_{rel} \approx \frac{2\Delta l}{\sqrt{nk}l_{max}} = \pm\frac{2}{\sqrt{75}\cdot 5} = \pm 0.046 = \pm 4.6\%$$

5 Faustformel zur Kalkulation des Fehlers

Die hier angeführte Analyse der statistischen Fehler bei der Leistungsbewertung lässt sich in eine kurze Faustformel verwandeln. Die Vergabe von Punkten für ein individuelles Leistungskriterium l_i hat eine relative Unsicherheit, die durch

$$u = \frac{2\Delta l}{l_{max}}$$

ausgedrückt werden kann. Sie geht als wesentlicher Anteil in den relativen statistischen Fehler

$$L_{rel} = \frac{u}{\sqrt{nk}}$$

der Leistungsbewertung ein.

Wie gross ist die **relative Unsicherheit** bei der Vergabe von Punkten für ein Kriterium l_i nun im allgemeinen ? Wenn für die Beantwortung eines Kriteriums z.B. 0 bis 3 Punkte zur Verfügung stehen, so sind dies 4 Bewertungsstufen. Nimmt man an, dass ein Gutachter sich um maximal eine Bewertungstufe verschätzen kann, so ist eine relative Unsicherheit von $u = 2 \cdot 1/4 = 0.5 = 50\%$ anzusetzen; bei einer Feinauflösung der Kriterien von 1 bis 10 Punkten sind es 10 Bewertungsstufen und damit $u = 2 \cdot 1/10 = 20\%$. Die hohe relative Unsicherheit von immerhin 50% bei 4 Wertungsstufen mögen auf den ersten Blick heftig erscheinen, sie gibt aber lediglich der Tatsache Ausdruck, dass der Bewerter sich nach oben *oder* unten um nur eine Stufe verschätzen kann. Dabei wird zudem angenommen, dass die Bewertungsstufen ausreichend genau definiert sind, so dass der Gutachter eine bestimmte Bepunktung mit einiger Sicherheit erteilen kann. Sind diese Stufen dagegen nur vage definiert, so nützt auch eine besonders feine Aufteilung der Bewertungsstufen nicht viel. Als Faustformel für die relative Unsicherheit kann man daher ansetzen:

> Die durchschnittliche **relative Unsicherheit** bei der Vergabe von Punkten für ein Kriterium (in %) ist gleich 200 geteilt durch die Anzahl der Bewertungsstufen:
>
> $$u = \frac{200}{Anzahl\,der\,Bewertungsstufen}[\%]$$

Als Faustformel für den Fehler der Leistungsbewertung wird daher angesetzt:

Relativer Fehler der gesamten Leistungsbewertung in % = ±(Relative Unsicherheit in %) geteilt durch (Wurzel aus [Anzahl der bepunkteten Kriterien mal Anzahl der unabhängigen Bewertungen]):

$$L_{rel} = \pm \frac{u[\%]}{\sqrt{Anz.Kriterien \times Anz.unabh.Bewertungen}}$$

6 Ein kleines praktisches Beispiel

Es sei ein kleines Projekt gegeben mit nur $n = 4$ bewerteten Fragen. Es seien 11 Bewertungsstufen vorgegeben, so dass jede Frage 0 bis 10 Punkte bekommen kann. Die Gewichte seien 30%, 10%, 20% und 40%, dann gilt:

$$L = 0.3l_1 + 0.1l_2 + 0.2l_3 + 0.4l_4$$

Ein Angebot erhalte die Bepunktungen $l_i = 3, 5, 8, 6$ mit einem statistischen Fehler von $\Delta l = 1$ Punkt. Der *tatsächliche* relative statistische Fehler für dieses Angebot ist

$$L_{rel} = \frac{\Delta L}{L} = \frac{\sqrt{\sum_{i=1}^{4}(\alpha_i \Delta l_i)^2}}{\sum_{i=1}^{4} \alpha_i l_i}$$

$$= \pm \frac{\sqrt{0.3^2 + 0.1^2 + 0.2^2 + 0.4^2}}{0.3 \cdot 3 + 0.1 \cdot 5 + 0.2 \cdot 8 + 0.4 \cdot 6} = \pm \frac{0.548}{5.4} = \pm 0.101 = \pm 10.1\%$$

Der *überschlägige* Wert für alle Angebote nach Faustformel ist mit

$$u = 200/11 = 18.2\%$$

$$L_{rel} \approx \pm \frac{u}{\sqrt{nk}} = \pm \frac{18.2\%}{\sqrt{4 \cdot 1}} = \pm 9.1\%$$

nur geringfügig kleiner, obwohl grosse Schwankungen in den Gewichten vorliegen. Diese Faustformel ist daher als robust anzusehen. Werden in diesem Beispiel zudem drei unabhängige Gutachter eingesetzt, so verringert sich der relative Fehler weiter auf

$$L_{rel} \approx \pm \frac{u}{\sqrt{nk}} = \pm \frac{18.2}{\sqrt{4 \cdot 3}} = \pm 5.3\%.$$

Wie man sieht lassen sich in aufwendigen Verfahren die Fehlerquoten durch eine grössere Menge von Kriterien n, mehrere unabhängige Bewerter k sowie durch eine feinere Auflösung der Punkteskala l_{max} minimieren. Mit einem durchaus realistischen Szenario mit

$$n = 30 \qquad k = 2 \qquad l_{max} = 10 \qquad \Delta l = \pm 1$$

erreicht der statistische Fehler

$$L_{rel} \approx \frac{2\Delta l}{\sqrt{nk}\, l_{max}} = \frac{2}{\sqrt{60} \cdot 10} = \pm 2.6\%$$

nur noch eine geringe Grösse.

7 Gegenüberstellung der Angebote unter Beachtung der Fehlerbreiten

Sei der ermittelte relative Fehler 5%. Dann sind Angebote deren Leistungsbewertungen um nicht mehr als $2 \cdot 5\% = 10\%$ differieren als potenziell gleichwertig zu betrachten. Ein Beispiel:

$$L_1 = 8000 \pm 5\% \Rightarrow L_1 \in [7600, 8400]$$

$$L_2 = 7400 \pm 5\% \Rightarrow L_2 \in [7030, 7770]$$

$$L_3 = 7000 \pm 5\% \Rightarrow L_2 \in [6650, 7350]$$

Die Angebote L_1 und L_2 sind als gleichwertig zu betrachten, da der höchste Wert für L_2 innerhalb des Schwankungsbereiches von L_1 liegt. Dagegen ist das Angebot L_3 eindeutig schlechter als L_1, jedoch nicht grundsätzlich schlechter als L_2 anzusehen. Befindet sich das Angebot 1 in der engeren Auswahl, so ist auch das Angebot 2 zu erwägen. Scheidet Angebot 1 aus einem anderen Grunde aus, so ist nicht nur das Angebot 2, sondern auch das Angebot 3 in Erwägung zuziehen. Denkbar ist zudem noch, dass das Angebot 1 einzig durch weitere unabhängige Gutachter untersucht wurde und damit ein neuer sicherer Wert entstanden ist:

$$L_1 = 8050 \pm 3\% \Rightarrow L_1 \in [7809, 8292]$$

Damit wäre Angebot 1 als alleiniger Spitzenreiter anzusehen, da Angebot 2 auch bei weiterer Begutachtung und damit schmäleren Fehlerschranken nicht über 7770 Punkte hinauskommen dürfte.

7.1 Fortpflanzung des Fehlers auf Bewertungsformeln Z

Schlussendlich ist noch die Frage zu klären, wie sich der statistische Fehler von L auf eine eventuell angewendete Bewertungsformel $Z(L, K)$ fortpflanzt. Der relative statistische Fehler einer gegebenen Bewertungszahl ist:

$$Z_{rel} = \frac{\Delta Z}{Z(L,K)} = \frac{\sqrt{(\frac{\partial Z}{\partial L}\Delta L)^2 + (\frac{\partial Z}{\partial K}\Delta K)^2}}{Z(L,K)}$$

Falls die Kosten mit den Preisen identifiziert werden, können wir $\Delta K = 0$ annehmen. Dann gilt:

$$Z_{rel} = \frac{\partial Z}{\partial L}\frac{\Delta L}{Z}$$

Im Falle des standardmässigen Leistungs- zu Preisverhältnisses $Z = L/K$ oder auch des verallgemeinerten Verhältnisses $Z = L/K^n$ gilt dann

$$Z_{rel} = \frac{\partial Z}{\partial L}\frac{\Delta L}{Z} = \frac{1}{K^n}\frac{\Delta L}{\frac{L}{K^n}} = \frac{\Delta L}{L} = L_{rel}$$

und damit *ändert sich nichts am statistischen Fehler*. Anders jedoch bei additiven Formeln des Typs $Z(L, K) = aL \pm bK$:

$$Z_{rel} = \frac{\partial Z}{\partial L}\frac{\Delta L}{Z} = a\frac{\Delta L}{aL \pm bK} = \frac{\Delta L}{L \pm \frac{b}{a}K} \neq L_{rel}$$

Bei letzterem müsste der relative Fehler, da kostenabhängig, für jedes Angebot korrigiert werden.