



Wie aus einzelnen Bewertungen ein Gesamtergebnis wird

Faires Gesamtergebnis – das Alpha-getrimmte Mittel

Jeder Text erhält über die Bewertungszeit hinweg eine ganze Reihe von Bewertungen. Dabei gehen erfahrungsgemäß die einzelnen Federwerte stark auseinander, es ist keine Seltenheit, dass Texte in ihrem Bewertungsrepertoire neben allen anderen möglichen Werten sowohl die Höchstwertung 9 als auch die Tiefstwertung 1 aufweisen. Wie kommt man zu einer einzigen fairen Gesamtzahl angesichts dieser Variabilität?

Administrator Boro hat für die Ermittlung einer einzigen charakteristischen Zahl aus allen Bewertungen für einen Text das *alpha-getrimmte Mittel* gewählt. Das wird euch jetzt wahrscheinlich auch nicht so viel weiterhelfen. Also schauen wir uns das doch mal an einem ganz einfachen Beispiel an.

Wir nehmen an, wir haben als Bewertungen für unseren imaginären Herzblut-getränkten Text zweimal 6 Federn, einmal 7 und einmal 8 erhalten. Normalerweise würde man nun den *Mittelwert* berechnen, weil der unter den meisten Umständen die zuverlässigste Charakterisierung ist. Er wird ganz einfach wie folgt ermittelt:

Summe_der_Bewertungen geteilt durch *Anzahl_der_Bewertungen*,
also in unserem Beispiel: $27 / 4 = 6,75$

Nun betritt ein missgelaunter Zeitgenosse die Bühne (oder auch nur jemand mit radikal abweichendem Geschmack) und gibt nur eine Feder. Wegen der geringen Anzahl von Bewertungen (alles unter 10 gilt in der Statistik als extrem klein) hat diese eine Bewertung einen überproportionalen Einfluss auf das Endergebnis, nämlich, $28 / 5 = 5,6$. Die endgültige Zahl liegt unter allen Einzelbewertungen mit Ausnahme der unseres missgelaunten Zeitgenossen.

Nun, ruft jemand mit Statistikkenntnissen, das ist doch wohlbekannt, dass der Mittelwert nicht gut ist, wenn n (Anzahl der Beobachtungen = Bewertungen) klein ist. Da, so der Statistiker, nimmt man den *Median*. Was ist das nun schon wieder?

Der Median ist der Einzelwert, den man bekommt, wenn man alle Bewertungen der Größe nach sortiert, also 1, 6, 6, 7, 8

in unserem Beispiel, und den nimmt, der sich in der Mitte befindet, also die 6. Wunderbar. Wir schauen noch in der statistischen Fachliteratur nach und in der Tat, es wurde gezeigt, dass der Median die beste Kennziffer ist bei kleinem n . Alle sind glücklich und tanzen um ein Freudenfeuer.

Fußnote: Wenn man eine gerade Anzahl von Bewertungen hat, nimmt man die beiden Werte in der Mitte und mittelt. Im Beispiel, bevor wir die Bewertung mit nur einer Feder bekommen haben, (6, 6, 7, 8) ist es 6 und 7 und daraus wird 6,5 als Endergebnis.

Dann schauen sich die Wettbewerbsorganisatoren die Ergebnisse an und finden, dass sie acht Sieger mit 7,0 haben und neunzehn Texte dahinter mit Bewertung 6,0 oder so ähnlich. Bei allen bisherigen Wettbewerben hätte es auf jeden Fall eine ganze Schar Gewinnertexte mit einer einheitlichen Endbewertung von 7,0 oder 6,0 gegeben. Und auf den Plätzen darunter hätten sich noch mehr Texte auf der gleichen Ebene (5,0 4,0 3,0 ...) gedrängt. Denn es gibt ja kaum noch Zwischenstufen. Was tun?

Eine der Alternativen zum Median ist das *alpha-getrimmte Mittel*. Die Idee ist folgende: Wir gehen zurück



Wie aus einzelnen Bewertungen ein Gesamtergebnis wird

zum Mittelwert, aber wir schließen die Extreme aus. Und zwar unabhängig davon, wie extrem sie sind, und von beiden Seiten.

Wir ordnen die Bewertungen wieder der Größe nach an

1, 6, 6, 7, 8

und nehmen die äußeren Werte weg. Wie viele genau, das verrät uns das *alpha*, das wir *vorher* festgelegt haben. Sagen wir hier im Beispiel 20%, dann fällt genau ein Wert weg und wir kriegen

6, 6, 7

Mittelwert davon ist $19 / 3 = 6,3$

Wie man sieht, hat nun weder der *Ausreißer*, die Bewertung mit einer Feder, einen unzulässig großen Einfluss, noch erzielen Massen von Texten das exakt gleiche Endergebnis. Also, wenn das nicht fair und praktisch ist

...

Zusätzliche Informationen:

(a) Das alpha-getrimmte Mittel ist für den Wettbewerb auf $\alpha = 15\%$ eingestellt.

(b) Jede Bewertung hat einen Einfluss auf das Endergebnis. Auch die Extrembewertungen, nur haben die keinen überproportional großen.

(c) Es ist nicht sinnvoll, eine tiefere oder höhere Bewertung abzugeben, damit »die Bewertung nicht abgeschnitten« würde. Kehren wir noch einmal zu dem obigen Zahlenbeispiel zurück mit den Bewertungen:

1, 6, 6, 7, 8

Wenn ein weiterer Bewerter nun 9 Federn vergibt, haben wir

1, 6, 6, 7, 8, 9

und nach Anwendung der Trimmung verbleiben

6, 6, 7, 8

was zu einem finalen Wert von $27 / 4 = 6,75$ führt, also deutlich höher als das vorige 6,3.

Wenn der Benutzer stattdessen 7 Federn vergeben würde, erhielten wir

1, 6, 6, 7, 7, 8

und nach dem Abschneiden

6, 6, 7, 7

was den endgültigen Wert von $26 / 4 = 6,5$ ergibt. Wie erwartet und sinnvoll, hat die tiefere Bewertung zu einer relativen Erniedrigung des Endwertes geführt, obwohl sie selbst nicht wie die vorige 9 »weggefallen« ist.

Danke an alle Beteiligten für diese Erklärung!

Diskutieren Sie [hier](#) online mit!