

Methoden zur Bestimmung der wichtigsten Merkmale einer statistischen Zahlenreihe

Von Charles Willigens, Bern

Die Methode, welche hier besprochen werden soll, ist ein besonderer Fall der Pearsonschen Darstellungsmethode. Da ihre Anwendung einfach ist und, sobald die Schlüsse vorliegen, keine besondere mathematische Bildung erfordert, muss man sich nur wundern, dass sie nicht mehr verwendet wird. Eine besondere theoretische Behandlung ist mir nicht bekannt. Erwähnt wird das Verfahren von Professor G. F. Lipps ¹⁾ und es wird von Professor Dr. Haemig in Zürich, mit gutem Erfolge auf wirtschaftliche Untersuchungen angewendet. Es scheint mir daher nicht überflüssig, die theoretische Begründung der Methode darzulegen, und sie auf einige Beispiele anzuwenden.

Die Frequenz und das arithmetische Mittel

Betrachten wir eine Reihe von N -Zahlen, $x_1, x_2 \dots x_i, \dots x_N$, welche wir uns nach steigenden Werten geordnet denken können. Die Zahl der Glieder mit dem Werte x_i sei gleich n_i . Das arithmetische Mittel dieser Zahlen ist die Summe von Ausdrücken von der Form $\frac{n_i}{N} x_i$. Der Bruch $\frac{n_i}{N} = f_i$ ist die Frequenz des Wertes x_i , und das arithmetische Mittel nimmt die Form $\sum f_i x_i = m$ an. Ebenso ist das arithmetische Mittel einer Potenz derselben Zahlen von der Form $\sum f_i x_i^r$, wo f_i denselben Zahlenwert bedeutet, wie im Falle der ersten Potenz.

Kennt man eine Funktion $\varphi(x)$ mit der Eigenschaft, dass der Ausdruck $N \varphi(x) dx$ die Zahl der Glieder einer Reihe angibt, deren Werte zwischen x und $x + dx$ liegen, so bezeichnet man $\varphi(x)$ als Frequenzfunktion dieser Reihe. Da die Summe der Werte n_i gleich N und also $\sum f_i = 1$, so muss auch das Integral

$$\int \varphi(x) dx = 1$$

sein, wobei die Integration über ein geeignetes Intervall sich erstreckt. Das arithmetische Mittel einer Potenz von x ist durch das Integral $\int x^r \varphi(x) dx$ gegeben, über das gleiche Intervall erstreckt.

Die Pearsonsche Frequenzfunktion. Wegen der Rolle, welche die Frequenzfunktion bei der Berechnung der arithmetischen Mittel der Potenzen von x spielt, ist es gegeben, diese Zahlen zur Bestimmung der Konstanten der

¹⁾ G. F. Lipps, Grundriss der Psychophysik (Sammlung Göschen, Nr. 98), Seite 78 f.

Frequenzfunktion zu benützen. Pearson betrachtet Funktionen, welche einer Differentialgleichung erster Ordnung der Form genügen:

$$(1) \quad \frac{y'}{y} = \frac{x - a}{b_0 + b_1 x + b_2 x^2}.$$

Für $x = a$ ist $y' = 0$, folglich entspricht diesem Werte ein Extremum (Maximum oder Minimum). Bringt man die Gleichung unter die Form:

$$y' (b_0 + b_1 x + b_2 x^2) = (x - a) y$$

und multipliziert man beide Glieder mit der Potenz x^s , so ergibt sich durch Integration

$$\int y' (b_0 x^s + b_1 x^{s+1} + b_2 x^{s+2}) dx = \int (x^{s+1} - a x^s) y dx.$$

Durch partielle Integration des ersten Gliedes erhält man

$$\left[y (b_0 x^s + b_1 x^{s+1} + b_2 x^{s+2}) \right]_{x_1}^{x_2} - \int_{x_1}^{x_2} y \left[s b_0 x^{s-1} + (s+1) b_1 x^s + (s+2) b_2 x^{s+1} \right] dx.$$

Wählt man nun die Integrationsgrenzen so, dass die erste Klammer verschwindet, so reduziert sich die Formel auf

$$(2) \quad - \int_{x_1}^{x_2} y \left[s b_0 x^{s-1} + (s+1) b_1 x^s + (s+2) b_2 x^{s+1} \right] dx = \int_{x_1}^{x_2} (x^{s+1} - a x^s) y dx.$$

Unter den Integralzeichen treten Ausdrücke von der Formel $x^r y dx$ auf. Die entsprechenden Integrale geben also die arithmetischen Mittel der betreffenden Potenzen von x , deren Zahlenwert unmittelbar aus der gegebenen Zahlenreihe berechnet werden kann. Zur Vereinfachung der Berechnung nimmt man das arithmetische Mittel als Ausgangswert, d. h., wenn m das arithmetische Mittel bedeutet, dass jede Zahl x_i durch den Wert $x_i - m$ ersetzt wird. Das arithmetische Mittel dieser neuen Zahlenreihe ist gleich Null. Wir haben also

$$\int_{x_1}^{x_2} y dx = 1. \quad \int_{x_1}^{x_2} x y dx = 0. \quad \int_{x_1}^{x_2} x^r y dx = \mu_r.$$

μ_r bedeutet das arithmetische Mittel der r -ten Potenzen von x und ist somit eine bekannte Zahl. Gibt man im Vorhergehenden dem Exponenten s die Werte 0, 1, 2, 3, so erhält man die vier Gleichungen:

$$(3) \quad \begin{aligned} b_1 &= a \\ b_0 + 3 b_2 \mu_2 &= -\mu_2 \\ 3 b_1 \mu_2 + 4 b_2 \mu_3 &= -\mu_3 + a \mu_2 \\ 3 b_0 \mu_2 + 4 b_1 \mu_3 + 5 b_2 \mu_4 &= -\mu_4 + a \mu_3 \end{aligned}$$

aus welchen sich die Werte der Koeffizienten berechnen lassen.

Zur Vereinfachung der Schreibweise führt' Pearson die beiden Verhältniszahlen

$$\beta_1 = \frac{\mu_3^2}{\mu_2^2}, \quad \beta_2 = \frac{\mu_4}{\mu_2^2}.$$

Diese beiden Ausdrücke sind reine Zahlenwerte und werden durch einen Wechsel in der Masseinheit der Zahlen x nicht verändert. Für den allgemeinen Fall sei hier nur das Resultat gegeben

$$(4) \quad a = \frac{\mu_3}{\mu_2} \cdot \frac{-(\beta_2 + 3)}{10\beta_2 - 12\beta_1 - 18}.$$

Der Wert für β_2 ist immer grösser als 1. β_1 kann hingegen jeden positiven Wert annehmen.

Ist $\mu_3 = 0$, so ist die Verteilung der Reihe symmetrisch in bezug auf das arithmetische Mittel, und in diesem Falle erfahren die Formeln 3 eine bedeutende Vereinfachung. Wenn μ_3 nicht allzu gross ist, oder noch besser, wenn der Wert für a aus der Formel 4 genügend klein ist, so wird man mit Vorteil die Glieder mit μ_3 vernachlässigen und sich mit einer symmetrischen Frequenzkurve begnügen können.

Der Wert für a gibt die Abscisse des Extremums an. Ist also a von Null verschieden, so ist das arithmetische Mittel nicht der Wert grösster Frequenz in der Zahlenreihe. Setzen wir in den Formeln (3) $\mu_3 = 0$, so erhalten wir die Gleichungen

$$(5) \quad \begin{aligned} b_1 &= a = 0 \\ b_0 + 3 b_2 \mu_2 &= -\mu_2 \\ 3 b_0 \mu_2 + 5 b_2 \mu_4 &= -\mu_4. \end{aligned}$$

Die Auflösung nach b_0 und b_2 gibt

$$(6) \quad \begin{aligned} b_0 &= \frac{2 \mu_2 \mu_4}{9 \mu_2^2 - 5 \mu_4} = 2 \mu_2 \frac{\beta_2}{9 - 5 \beta_2} \\ b_2 &= \frac{\mu_4 - 3 \mu_2^2}{9 \mu_2^2 - 5 \mu_4} = \frac{\beta_2 - 3}{9 - 5 \beta_2}. \end{aligned}$$

Die Differentialgleichung nimmt die Form an:

$$\frac{y'}{y} = \frac{x}{b_0 + b_2 x^2} \text{ oder}$$

$$\frac{y'}{y} = \frac{1}{2 b_2} \frac{2 \frac{b_2}{b_0} x}{1 + \frac{b_2}{b_0} x^2}.$$

Das Integral dieses Ausdruckes lautet:

$$\text{Log } y = \frac{1}{2b_2} \text{Log} \left(1 + \frac{b_2}{b_0} x^2 \right) + \text{Log } y_0$$

$$(7) \quad y = y_0 \left(1 + \frac{b_2}{b_0} x^2 \right)^{\frac{1}{2b_2}}$$

y_0 wird durch die Bedingung bestimmt

$$\int_{x_1}^{x_2} y \, dx = y_0 \int_{x_1}^{x_2} \left(1 + \frac{b_2}{b_0} x^2 \right)^{\frac{1}{2b_2}} dx = 1.$$

Wir wollen nun untersuchen, wie sich die Koeffizienten b_0 und b_2 verändern, wenn μ_2 als fest angenommen wird und β_2 alle positiven Werte von 1 bis ∞ durchläuft.

Gestalt der Frequenzkurve. Lassen wir in den Formeln 6 die Veränderliche β_2 die angegebenen Werte durchlaufen und nehmen an, dass μ_2 einen konstanten Wert behält. Da wir die Abweichungen vom arithmetischen Mittel untersuchen, hat μ_2 den kleinsten möglichen Wert für die betrachtete Reihe.

Die zweite der Gleichungen 5 lässt sich schreiben

$$b_0 + \mu_2 (3b_2 + 1) = 0.$$

nimmt man nun b_0 und b_2 als Koordinaten, so liegen die entsprechenden Punkte auf einer Geraden, welche durch μ_2 vollständig bestimmt ist. Diese Gerade schneidet die Abscissenachse in $b_0 = -\mu_2$, die Ordinatenachse in $b_2 = -\frac{1}{3}$. Alle

Geraden schneiden sich also in einem festen Punkte auf der Ordinatenachse, wenn sich μ_2 verändert, und, da μ_2 stets positiv ist, schneiden sie immer den negativen Teil der Abscissenachse.

$$b_0 = 2\mu_2 \frac{\beta_2}{9 - 5\beta_2} \quad \frac{db_0}{d\beta_2} = 2\mu_2 \frac{9}{(9 - 5\beta_2)^2} > 0;$$

mit wachsendem β_2 wächst also b_0 beständig.

$$b_2 = \frac{\beta_2 - 3}{9 - 5\beta_2} \quad \frac{db_2}{d\beta_2} = \frac{-6}{(9 - 5\beta_2)^2} < 0;$$

b_2 ist eine beständig abnehmende Funktion.

$$\frac{b_2}{b_0} = \frac{1}{2\mu_2} \frac{\beta_2 - 3}{\beta_2} \quad \frac{d}{d\beta_2} \left(\frac{b_2}{b_0} \right) = \frac{3}{2\mu_2 \beta_2^2} > 0;$$

das Verhältnis $\frac{b_2}{b_0}$ wächst beständig.

$$\beta_2 = 1 \quad b_0 = \frac{\mu_2}{2} \quad b_2 = -\frac{1}{2};$$

b_0 wächst und b_2 nimmt ab.

$$b_0 > 0, \quad b_2 < 0$$

$$\beta_2 = \frac{9}{5} = 1,8 \quad b_0 = \pm \infty \quad b_2 = \mp \infty;$$

b_0 und b_2 werden unendlich gross und wechseln ihr Vorzeichen.

$$b_0 < 0 \quad b_2 > 0. \quad \beta_2 = 3 \quad b_0 = -\mu_2 \quad b_2 = 0$$

$$b_0 < 0 \quad b_2 < 0$$

$$\beta_2 = +\infty \quad b_0 = -\frac{2}{5}\mu_2 \quad b_2 = -\frac{1}{5}.$$

Daraus geht hervor, dass b_0 nie Werte zwischen $-\frac{2}{5}\mu_2$ und $\frac{1}{2}\mu_2$ annimmt, ebenso kann b_2 nie zwischen $-\frac{1}{5}$ und $-\frac{1}{2}$ liegen.

Untersuchen wir noch die Kurve für die Werte $\beta_2 = \frac{9}{5}$ und $\beta_2 = 3$.

Für $\beta_2 = \frac{9}{5}$ wird die Differentialgleichung $\frac{y'}{y} = 0$, also $y = \text{constant}$.

Die Kurve ist eine horizontale Gerade.

$$\beta_2 = 3 \quad \frac{y'}{y} = -\frac{x}{\mu_2} \quad y = y_0 e^{-\frac{x^2}{2\mu_2}};$$

man erhält die Gauss'sche Fehlerkurve.

In den verschiedenen Intervallen nimmt die Kurve folgende Formen an

$$1 < \beta_2 < \frac{9}{5} \quad \text{es sei z. B.} \quad \beta_2 = \frac{8}{5}$$

$$b_0 = \frac{16}{5}\mu_2 \quad b_2 = -\frac{7}{5}$$

$$\frac{b_2}{b_0} = -\frac{7}{16\mu_2} \quad y = \frac{y_0}{\left(1 - \frac{7}{16\mu_2}x^2\right)^{\frac{5}{14}}};$$

y wird für die Werte

$$x_1, x_2 = \pm \sqrt{\frac{7}{16\mu_2}} = \sqrt{-\frac{b_0}{b_2}}$$

unendlich. $x = 0$ gibt für y ein Minimum. Die Kurve hat eine Uförmige Gestalt; und das arithmetische Mittel ist der Wert geringster Frequenz. Die Werte der Reihe häufen sich in der Nähe der Grenzwerte x_1 und x_2 an, welche auch die Grenzwerte für die Integration sind. Mit wachsendem β_2 wird die Kurve immer flacher und artet für $\beta_2 = \frac{9}{5}$ in eine horizontale Gerade aus.

$$\frac{9}{5} < \beta_2 < 3 \quad \text{es sei z. B. } \beta_2 = 2$$

$$b_0 = -4\mu_2 \quad b_2 = +1 \quad \frac{b_2}{b_0} = -\frac{1}{4\mu_2}$$

$$y = y_0 \left(1 - \frac{x^2}{4\mu_2} \right)^{\frac{1}{2}}$$

Die Kurve ist glockenförmig und schneidet die x -Achse in den Punkten

$$x_1, x_2 = \pm \sqrt{4\mu_2} = \sqrt{-\frac{b_0}{b_2}}$$

In diesem besonderen Falle $\beta_2 = 2$ ist die Kurve der obere Teil der Ellipse

$$\frac{x^2}{4\mu_2} + \frac{y^2}{y_0^2} = 1.$$

Damit der Flächeninhalt dieser halben Ellipse gleich 1 ist, muss

$$y_0 = \frac{1}{\pi\sqrt{\mu_2}} \text{ sein.}$$

Wenn β_2 wächst, nehmen $\frac{b_2}{b_0}$ und $-\frac{b_0}{b_2}$ immer zu. Die Schnittpunkte x_1 und x_2 entfernen sich immer mehr vom Ursprunge und rücken für $\beta_2 = 3$ ins Unendliche.

$$\beta_2 > 3 \quad \text{es sei } \beta_2 = 4$$

$$b_0 = -\frac{8}{11}\mu_2 \quad b_2 = -\frac{1}{11} \quad \frac{b_2}{b_0} = \frac{1}{8\mu_2}$$

$$y = \frac{y_0}{\left(1 + \frac{x^2}{8\mu_2} \right)^{22}}$$

Die Kurve weist ein Maximum auf; da $\frac{b_2}{b_0} > 0$, kann der Nenner niemals gleich Null werden. Für $x = \pm \infty$ wird $y = 0$. Die Kurve ist glockenförmig und hat eine Gestalt ähnlich der Gauss'schen Fehlerkurve.

Untersuchung der Streuungsverhältnisse

Im Falle des Gauss'schen Verteilungsgesetzes versteht man unter Streuung die Zahl $\sqrt{\mu_2}$. Diese Zahl gibt die Abscissen der Wendepunkte der Darstellungskurve

$$y = \frac{1}{\sqrt{2\pi\mu_2}} e^{-\frac{x^2}{2\mu_2}}$$

und wir wollen auch im allgemeinen Falle unter Streuung der gegebenen Wertreihe den halben Abstand der Wendepunkte verstehen, oder auch die halbe Spannweite des nach unten konkaven Teiles der Kurve.

Leitet man die Gleichung

$$y'(b_0 + b_2 x^2) = x y$$

nach x ab, so erhält man

$$y''(b_0 + b_2 x^2) + y' 2 b_2 x = x y' + y;$$

wenn man nach y'' auflöst und y' in Funktion von x und y ersetzt, so erhält man

$$y'' = y \frac{x^2(1 - b_2) + b_0}{(b_0 + b_2 x^2)^2};$$

die Abscissen der Wendepunkte haben also die Werte

$$x = \pm \sqrt{\frac{b_0}{b_2 - 1}} = \pm \sqrt{\frac{\mu_2 \beta_2}{3(\beta_2 - 2)}}$$

und für $\beta_2 = 3$ findet man wieder den Wert $\sqrt{\mu_2}$. Damit die Kurve Wendepunkte aufweist, muss der Ausdruck unter dem Wurzelzeichen positiv sein und dies erfordert $\beta_2 > 2$. Mit wachsenden β_2 nimmt die Streuung beständig ab, der mittlere Teil der Kurve zieht sich also immer mehr zusammen.

Wenn wir den Fall des Gauss'schen Gesetzes als normale Verteilung bezeichnen, so haben wir für

$\beta_2 < 3$ eine übernormale Streuung,

$\beta_2 = 3$ eine normale Streuung,

$\beta_2 > 3$ eine unternormale Streuung.

Anwendung des Verfahrens. Es sei eine Reihe von Zahlen, x_1, x_2, \dots, x_N gegeben, die wir uns steigend geordnet denken können. Man geht folgendermassen vor:

1. Man bildet das arithmetische Mittel m aus den Zahlen x .
2. Man berechnet die Zahlen $x - m$ und bildet die arithmetischen Mittel μ_2, μ_3, μ_4 der zweiten, dritten und vierten Potenz dieser Zahlen.
3. Man berechnet die beiden Zahlen

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{und} \quad \beta_2 = \frac{\mu_4}{\mu_2^2},$$

sowie den Wert

$$a = -\frac{\mu_3}{\mu_2} \frac{\beta_2 + 3}{10\beta_2 - 12\beta_1 - 18},$$

ist das Mass der Asymmetrie, d. h. der Wert $\frac{a}{\sqrt{\mu_2}}$, klein, so ist man zu folgenden

Schlüssen berechtigt:

$1 < \beta_2 < 1,8$. Die Frequenzkurve ist Uförmig und nähert sich einer horizontalen Geraden, wenn β_2 von der oberen Grenze wenig verschieden ist. Das arithmetische Mittel ist in der Darstellung der Wert geringster Frequenz.

$1,8 < \beta_2 < 3$. Die Frequenzkurve weist ein Maximum auf und schneidet die x -Achse im Endlichen. Mit wachsenden β_2 nimmt die Streuung ab, für $\beta_2 = 3$ hat man die Normalverteilung von Gauss.

$3 < \beta_2$. Die Kurve hat ein Maximum und nähert sich asymptotisch der x -Achse. Die Streuung ist unternormal. m ist der Wert extremer Frequenz.

Beispiel 1. Zahlen der Todesfälle im ersten Lebensjahre auf 1000 lebendgeborene Knaben der Kalenderjahre 1901—1920.

| x | $x - m$ | x | $x - m$ |
|-----|---------|-----|---------|
| 87 | — 30,4 | 118 | + 0,6 |
| 88 | — 29,4 | 122 | + 4,6 |
| 90 | — 27,4 | 125 | + 7,6 |
| 93 | — 24,4 | 127 | + 9,6 |
| 96 | — 21,4 | 128 | + 10,6 |
| 99 | — 18,4 | 137 | + 19,6 |
| 99 | — 18,4 | 141 | + 23,6 |
| 104 | — 13,4 | 145 | + 27,6 |
| 104 | — 13,4 | 145 | + 27,6 |
| | | 148 | + 30,6 |
| | | 152 | + 34,6 |

$$m = \frac{2348}{20} = 117,4.$$

$$\mu_2 = \frac{9510,80}{20} = 475,54. \quad \sqrt{\mu_2} = 21,806.$$

$$\mu_3 = + \frac{19.733,654}{20} = + 986,6827.$$

$$\mu_4 = \frac{6.975.945,1040}{20} = 348.797,2552.$$

$$\beta_1 = 0,009.053. \quad \underline{\beta_2 = 1,5424.}$$

$$a = - 3,659. \quad \frac{a}{\sqrt{\mu_2}} = - 0,1678. \quad m + a = 113,74.$$

da $\beta_2 < 1,8$ hat die Kurve ein Minimum und in diesem Falle ist das arithmetische Mittel kein gutes Merkmal der Wertereihe.

Beispiel 2. Zahlen der Eheschliessungen auf 1000 Einwohner in den Jahren 1881—1910. Diese Zahlen wurden wegen ihres regelmässigen Verlaufs gewählt. Zur Platzersparnis verzichten wir auf die Wiedergabe, der kleinste Wert beträgt 6,8, der grösste 7,8. Die charakteristischen Werte sind

$$m = 7,3. \quad \mu_2 = 0,097. \quad \sqrt{\mu_2} = 0,3114. \quad \mu_3 = + 0,0027.$$

$$\mu_4 = 0,0175. \quad \beta_1 = 0,0823. \quad \underline{\beta_2 = 1,85.}$$

$$a = - 0,2765. \quad \frac{a}{\sqrt{\mu_2}} = - 0,8878. \quad m + a = 7,02.$$

Die Asymmetrie ist sehr stark, da die grösste Abweichung vom arithmetischen Mittel in der Zahlenreihe $\pm 0,5$ beträgt. Sieht man aber davon ab, so genügt der Wert von β_2 um zu zeigen, dass selbst die symmetrische Kurve, welche eine sehr starke Streuung aufweist, kein günstiges Ergebnis für das arithmetische Mittel liefert.

Beispiel 3. Zahlen der männlichen Geburten auf 10.000 Lebend- und Totgeburten in den Jahren von 1871—1930.

$m = 5.137$, grösste Abweichung vom Mittel ± 42 . Die anderen Werte sind folgende:

$$\mu_2 = 351,38. \quad \sqrt{\mu_2} = 18,745.$$

$$\mu_3 = + 929,52. \quad \mu_4 = 330.286,18.$$

$$\beta_1 = 0,019.915. \quad \beta_2 = 2,7374.$$

$$a = - 1,6614. \quad \frac{a}{\sqrt{\mu_2}} = - 0,08863. \quad m + a = 5135,34.$$

Die Asymmetrie ist schwach und die Streuung ist nur wenig übernormal. Das arithmetische Mittel kann als ein Wert betrachtet werden, von welchem die Zahlen der gegebenen Reihe durch Zufall abweichen.

Die drei angeführten Beispiele entsprechen ganz verschiedenen typischen Fällen. Beim ersten haben wir eine geringe Asymmetrie, aber die Werte der Reihe scharen sich nicht um den Mittelwert, welcher keinen Ersatz für die Zahlenreihe bilden kann. Beim zweiten Beispiel haben wir eine sehr starke Asymmetrie, welche von der mittleren quadratischen Abweichung wenig verschieden ist. Das arithmetische Mittel findet Verwendung in der Meinung, es sei der wahrscheinlichste Wert für die Zahlen der gegebenen Reihe, was aber nicht notwendig der Fall ist; der Umstand, dass für das arithmetische Mittel die Summe der Quadrate der Abweichungen ein Minimum ist, bleibt hierfür ohne Einfluss. Ist nun diese Bedingung nicht erfüllt, so sinkt natürlich der Wert des arithmetischen Mittels als Darstellung einer Reihe von Werten.

Das angeführte Verfahren erlaubt es z. B. die Änderungen von Beobachtungen von Jahr zu Jahr zu verfolgen. Man kann unter anderem auch feststellen, ob mit der Zeit die Zahlen sich immer dichter um einen bestimmten Betrag scharen, ob eine Stabilität der Verhältnisse eintritt, ob die Abweichungen vom Mittel als zufällige angesehen werden dürfen oder nicht. Diese Methode kann also wertvolle Aufschlüsse geben und sollte darum grössere Beachtung finden, es geht z. B. aus ihr klar hervor, dass die Sätze für die Kindersterblichkeit im ersten Lebensjahre keineswegs um einen festen Betrag schwanken und eine Stabilität dieser Sterblichkeitsverhältnisse noch nicht eingetreten ist.
