Schätzungen für Varianz und Schiefe der Verteilung von Versicherungssummen

Schätzungen für Varianz und Schiefe der Verteilung von Versicherungssummen

Von Kurt Wettenschwiler, Zürich

In der mathematischen Statistik werden Stichproben von kontinuierlich verteilten Variablen oft in Gruppen zusammengefasst, indem die Grössen, die bei vorgegebener Intervalleinteilung in ein bestimmtes Intervall fallen, gewöhnlich dessen Mittelpunkt zugeordnet werden[1]). Bei der Berechnung von Momenten und andern charakteristischen Grössen der Stichprobe haben wir es dann allerdings mit einer diskreten Verteilung zu tun.

Vielfach sind wir aber nicht an den Momenten der diskreten Verteilung sondern an den Momenten der zugrunde liegenden kontinuierlichen Verteilung interessiert.

Studiert man den Zusammenhang zwischen den Momenten der beiden Verteilungen, so findet man die bekannten *Sheppardschen Korrekturen*[2]). So ist z.B. die Varianz der diskreten Verteilung gegenüber derjenigen der kontinuierlichen Verteilung durchschnittlich um $\frac{1}{12}h^2$ grösser, wenn h die einheitliche Intervallänge bedeutet. Man kann also aus den Momenten der einen Verteilung auf die entsprechenden der andern schliessen.

Im Versicherungswesen ist mit zunehmend kollektiven Betrachtungsmethoden (kollektive Risikotheorie, nicht proportionale Deckungen, Stop-Loss- und Katastrophen-Verträge) die Kenntnis der Verteilung der Versicherungssummen eines Portefeuilles (Nominal- oder Risikosummen) von eminent wichtiger Bedeutung.

---

1) Standardbeispiele: Körperlängen von Rekruten, meteorologische Statistiken über mittlere Temperaturen u.a.

2) Gewisse Voraussetzungen über die Verteilungsfunktion müssen noch erfüllt sein, vgl. etwa H.Cramer: Mathematical Methods of Statistics, p.361, Princeton University 1946.

Im allgemeinen wird eine Aufstellung folgender Art erhältlich sein:

	N_1 Risiken zwischen 0 und Fr. 25 000.– mit einer Versicherungssumme von	Fr. S_1
	N_2 Risiken zwischen Fr. 25 000 und Fr. 50 000.– mit einer Versicherungssumme von	Fr. S_2
	N_i Risiken zwischen x_{i-1} und x_i mit einer Versicherungssumme von	Fr. S_i
Total	N Risiken mit einer totalen Summe von	Fr. S

Gesucht ist eine möglichst genaue Abschätzung für Varianz und Schiefe (2. und 3. Moment) der Verteilung $F(x)$ der Versicherungssummen des gesamten Bestandes. Die Funktion $F(x)$ bedeutet dabei die Anzahl der Risiken, für welche die Versicherungssummen kleiner oder gleich x sind.

Gegenüber dem eingangs erwähnten Problem der mathematischen Statistik, bei dem lediglich die Anzahl der Elemente einer Stichprobe, die in ein bestimmtes Klassenintervall fallen, bekannt ist, haben wir in unserem Fall die zusätzliche Information über die durchschnittliche Versicherungssumme in jedem einzelnen Intervall. Diese zusätzliche Kenntnis wollen wir im folgenden auswerten.

Für ein Intervall i gilt

Mittelwert $$m_i = \frac{\int\limits_{(i)} x dF(x)}{\int\limits_{(i)} dF(x)} = \frac{S_i}{N_i},$$

Varianz $$s_i^2 = \frac{\int\limits_{(i)} (x-m_i)^2 dF(x)}{\int\limits_{(i)} dF(x)} = \frac{1}{N_i} \int\limits_{(i)} x^2 dF(x) - m_i^2,$$

und für den ganzen Versicherungsbestand finden wir darnach

Mittelwert

$$m = \frac{\int x dF(x)}{\int dF(x)} = \frac{1}{N} \sum \int_{(i)} x dF(x) = \frac{\sum S_i}{N} = \frac{S}{N} ,$$

Varianz

$$s^2 = \frac{1}{N} \int (x-m)^2 dF(x) = \frac{1}{N} \sum \int_{(i)} [(x-m_i) + (m_i-m)]^2 \, dF(x)$$

$$= \frac{1}{N} \sum \left[\int_{(i)} (x-m_i)^2 dF(x) + 2(m_i-m) \int_{(i)} (x-m_i) \, dF(x) + (m_i-m)^2 \int_{(i)} dF(x) \right]$$

$$= \frac{1}{N} \sum [N_i s_i^2 + 0 + N_i (m_i-m)^2]$$

$$s^2 = \frac{1}{N} [\sum (m_i-m)^2 N_i + \sum s_i^2 N_i] .$$

Die erste Teilsumme dieser Hauptformel kann aus den vorliegenden Werten pro Intervall sofort berechnet werden. Für das Korrekturglied (zweite Teilsumme), welches den Fehler angibt, den man bei dieser Konzentration der Versicherungssummen in m_i begeht, muss eine Abschätzung vorgenommen werden.

Berechnung von s_i^2

Betrachten wir das Intervall i mit der Intervallbreite $x_i - x_{i-1} = \triangle_i$ und dem Mittelpunkt $\bar{x}_i = \frac{x_{i-1} + x_i}{2}$, welcher im folgenden der Ursprung eines neuen Koordinatensystems sei. Die Transformation $\xi = x - \bar{x}_i$ diene diesem Zweck. Es wird nun gefordert, dass die Versicherungssummen in diesem Intervall durch eine Dichtefunktion $f(\xi) = a\xi + b$ *linear verteilt* werden und zwar so dass *Anzahl* und *Durchschnittssumme* der Risiken mit der wirklichen Verteilung übereinstimmen:

$$\int_{-\Delta_{i/2}}^{\Delta_{i/2}} (a\xi + b)\, d\xi = N_i$$

$$\int_{-\Delta_{i/2}}^{\Delta_{i/2}} \xi(a\xi + b)\, d\xi = (m_i - \bar{x}_i)\, N_i = \mu_i N_i\,,$$

woraus sich a und b ergeben:

$$a = \frac{12\,\mu_i N_i}{\Delta_i^3} \qquad\qquad b = \frac{N_i}{\Delta_i}\,.$$

Damit können wir die Varianz berechnen:

$$s_i^2 = \frac{1}{N_i} \int_{-\Delta_{i/2}}^{\Delta_{i/2}} (\xi - \mu_i)^2 (a\xi + b)\, d\xi = \frac{1}{N_i} \frac{b\,\Delta_i^3}{12} - \mu_i^2$$

$$s_i^2 = \frac{\Delta_i^2}{12} - (m_i - \bar{x}_i)^2\,.$$

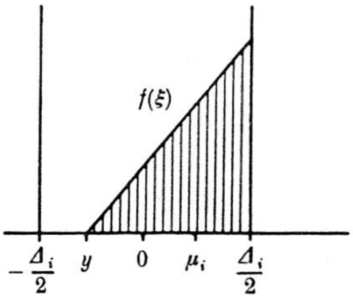
Obige Ableitungen sind allerdings nur dann sinnvoll, wenn die Dichtefunktion $f(\xi)$ nirgends negative Werte annimmt. Aus der Forderung, dass die Randwerte $f(-\Delta_{i/2}) \geqq 0$ und $f(\Delta_{i/2}) \geqq 0$, folgt sofort $|m_i - \bar{x}_i| \leqq \frac{1}{6}\Delta_i$.

Also gilt die Formel

$$s_i^2 = \frac{\Delta_i^2}{12} - (m_i - \bar{x}_i)^2 \quad \text{für} \quad |m_i - \bar{x}_i| \leqq \tfrac{1}{6}\Delta_i\,.$$

Berechnung von s_i^2, falls $|m_i - \bar{x}_i| > \frac{1}{6}\Delta_i$

i) $m_i - \bar{x}_i > \frac{1}{6}\Delta_i$.



s_i^2 wird berechnet für die Funktion

$$f(\xi) = \begin{cases} 0 & \text{für } -\frac{\Delta_i}{2} \leqq \xi \leqq y \\ a\xi + b & \text{für } \quad y < \xi \leqq \frac{\Delta_i}{2}, \end{cases}$$

wobei $ay + b = 0$ gilt,

also $\qquad y = -\frac{b}{a}$ bzw. $b = -ay$.

Für das schraffierte Dreieck interpretieren wir geometrisch:

Fläche $\qquad N_i = \left(\frac{\Delta_i}{2} - y\right)^2 \frac{a}{2} = (\Delta_i - 2y)^2 \frac{a}{8}$

Schwerpunkt $\qquad \mu_i = \frac{\Delta_i}{2} - \frac{\frac{\Delta_i}{2} - y}{3} = \frac{\Delta_i + y}{3}$.

Die Konstanten haben demnach folgende Werte:

$$y = 3\mu_i - \Delta_i \qquad a = \frac{8N_i}{(\Delta_i - 2y)^2}, \qquad b = -ay.$$

Damit bestimmen wir die Varianz

$$s_i^2 = \frac{1}{N_i} \int\limits_y^{\Delta_{i/2}} (\xi - \mu_i)^2 \, (a\xi + b) \, d\xi$$

$$= \frac{1}{N_i} \int\limits_y^{\Delta_{i/2}} (a\xi^3 + b\xi^2) \, d\xi - \mu_i^2 = \frac{1}{N_i} \left(\frac{a\Delta_i^4}{64} + \frac{b\Delta_i^3}{24} - \frac{ay^4}{4} + \frac{ay^4}{3} \right) - \mu_i^2$$

$$= \frac{1}{N_i} \left[\frac{8N_i}{192} \left(\frac{3\Delta_i^4 - 8y\Delta_i^3 + 16y^4}{(\Delta_i - 2y)^2} \right) \right] - \mu_i^2$$

$$= \frac{1}{24} (4y^2 + 4y\Delta_i + 3\Delta_i^2) - \left(\frac{\Delta_i + y}{3} \right)^2$$

$$s_i^2 = \frac{1}{8} (2\mu_i - \Delta_i)^2, \quad \text{für} \quad \mu_i = m_i - \bar{x}_i > \tfrac{1}{6} \Delta .$$

ii) $\bar{x}_i - m_i > \frac{1}{6} \Delta_i$.

Dieser Fall kann analog dem Fall i) behandelt werden, wobei lediglich zu beachten ist dass μ_i negativ wird: $-\dfrac{\Delta_i}{2} \leqq \mu_i \leqq -\dfrac{\Delta_i}{6}$.

Man erhält

$$s_i^2 = \tfrac{1}{8} (2\mu_i + \Delta_i)^2$$

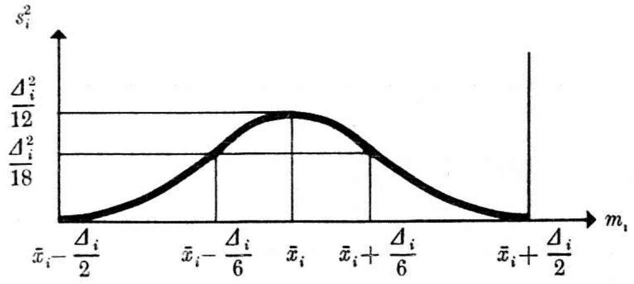
oder im ursprünglichen Koordinatensystem

$$s_i^2 = \tfrac{1}{8} (2m_i - 2\bar{x}_i + \Delta_i)^2 = \tfrac{1}{8} (2\,|\,m_i - \bar{x}_i\,| - \Delta_i)^2 .$$

Der letzte Ausdruck kann in dieser Form auch für Fall i) verwendet werden und wir haben zusammenfassend die

Resultate

$$s_i^2 = \begin{cases} \dfrac{\Delta_i^2}{12} - (m_i - \bar{x}_i)^2\,, & \text{falls} \quad |m_i - \bar{x}_i| \leqq \frac{1}{6}\Delta_i \\[2ex] \frac{1}{8}(2\,|m_i - \bar{x}_i| - \Delta_i)^2. & \text{falls} \quad |m_i - \bar{x}_i| > \frac{1}{6}\Delta_i\,. \end{cases}$$



Diskussion

Die Funktion $s_i^2(m_i)$ erreicht ihren höchsten Wert $\dfrac{\Delta_i^2}{12}$, wenn $m_i = \bar{x}_i$, und nimmt von diesem Maximum gegen die Intervallgrenzen hin bis auf Null ab. Sie ist im ganzen Intervall, insbesondere auch an den Nahtstellen $|m_i - \bar{x}_i| = \frac{1}{6}\Delta_i$, stetig differenzierbar. Ist ein Klassenintervall nicht besetzt, so liefert es richtigerweise an die Varianz des gesamten Bestandes keinen Beitrag, was aus der betreffenden Formel (auf Seite 163) ersichtlich ist. Ist eine Klasse besetzt und ist die mittlere Versicherungssumme (*a*) in der Nähe oder ganz im Intervallmittelpunkt oder (*b*) in der Nähe der Intervallgrenzen, so wird die totale Versicherungssumme dieser Klasse im ersten Fall über das ganze Intervall bzw. im letzteren Fall nur über einen Teil des Intervalls linear verteilt. Fällt der Mittelwert m_i im Extremfall mit einer Intervallgrenze zusammen, so können die Versicherungssummen nicht über das Intervall verteilt werden; es ist für den Betrag, den dieses Intervall zur Varianz des gesamten Bestandes beiträgt, auch gar kein Korrekturglied s_i^2 nötig, da mit Sicherheit der exakte Wert benützt wird. Mithin scheint diese Funktion zur Abschätzung der Varianz von Versicherungssummen brauchbar. Sie ist speziell geeignet für gut besetzte Klassen, wobei man nicht an eine konstante Intervallänge gebunden ist. Wenn eine Klasse nur mit einem Element besetzt ist, braucht man für das betreffende Intervall kein Korrekturglied zu verwenden.

Berechnung der Schiefe

Nachdem wir die Varianz s^2 kennen, müssen wir zur Berechnung der Schiefe $\gamma = \frac{h^{(3)}}{s^3}$ noch das 3. Hauptmoment $h^{(3)}$ der Verteilung bestimmen:

$$\begin{aligned}
h^{(3)} &= \frac{1}{N} \int (x-m)^3 dF(x) \\
&= \frac{1}{N} \sum \int_{(i)} [(x-m_i) + (m_i-m)]^3 dF(x) \\
&= \frac{1}{N} \sum \Bigg[\int_{(i)} (x-m_i)^3 dF(x) + 3(m_i-m) \int_{(i)} (x-m_i)^2 dF(x) \\
&\qquad + 3(m_i-m)^2 \int_{(i)} (x-m_i)\, dF(x) - (m_i-m)^3 \int_{(i)} dF \Bigg] \\
&= \frac{1}{N} \left(\sum N_i h_i^{(3)} + 3 \sum (m_i-m) N_i s_i^2 + 0 + \sum (m_i-m)^3 N_i \right).
\end{aligned}$$

Die Terme $h_i^{(3)}$ können sowohl positive wie negative Werte annehmen. Bei der Summation über den gesamten Bestand werden sie sich mehr oder weniger aufheben. Ausserdem ist ihr Gewicht im Vergleich zu den andern Termen im allgemeinen gering, so dass wir $\sum N_i h_i^{(3)} \approx 0$ vernachlässigen können.

Damit erhalten wir als *Schätzung für das 3. Hauptmoment*

$$h^{(3)} \approx \frac{1}{N} \left[\sum (m_i-m)^3 N_i + 3 \sum (m_i-m)\, s_i^2 N_i \right],$$

wobei für die s_i^2 die früher abgeleiteten Schätzungen einzusetzen sind.

Hat man erst die Hauptmomente der Gesamtverteilung bestimmt, so kann man sofort die entsprechenden Momente bezüglich des Ursprungs finden nach den Beziehungen:

$$p_1 = \frac{1}{N} \int x dF(x) = m$$

$$p_2 = \frac{1}{N} \int x^2 dF(x) = s^2 + m^2$$

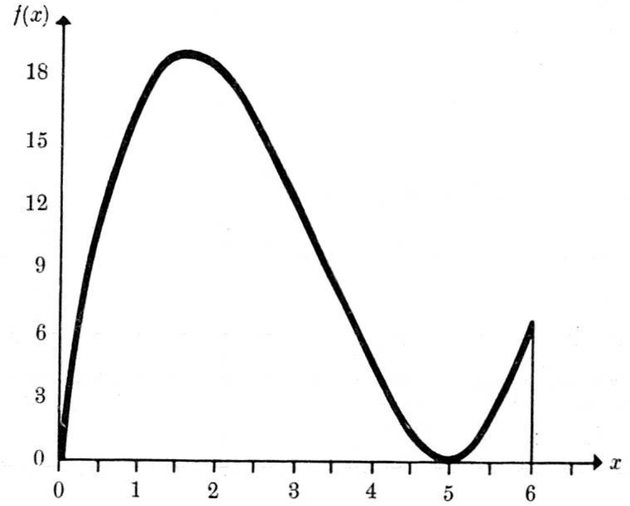
$$p_3 = \frac{1}{N} \int x^3 dF(x) = h^{(3)} + 3ms^2 + m^3 .$$

Beispiel

Spezielle Verteilung

$f(x) = x(x-5)^2$

6 Gruppen
Intervallbreite 1



i	N_i	S_i	m_i	$(m_i-m)^2$	$\|m_i-\bar{x}_i\|$	s_i	$(m_i-m)^3$	$(m_i-m)s_i^2$
1	9,4167	6,03333	0,64071	2,2279	0,14071	0,06353	—3,3254	—0,09483
2	17,9167	27,0334	1,50855	0,3900	0,00855	0,08325	—0,2437	—0,05201
3	15,4167	38,0333	2,46702	0,1113	0,03298	0,08224	0,0371	0,02745
4	7,9167	27,0333	3,41472	1,6420	0,08528	0,07606	2,1040	0,09746
5	1,4167	6,0334	4,25907	4,5188	0,24093	0,03356	9,6058	0,07134
6	1,9167	11,0333	5,75640	13,1266	0,25640	0,02967	47,5586	0,10750
1-6	54,0000	115,2	2,13333					

$$\frac{1}{N}\sum N_i(m_i-m)^2 = 1{,}3748 \qquad \frac{1}{N}\sum N_i(m_i-m)^3 = 1{,}5984$$

$$\frac{1}{N}\sum N_i s_i^2 = 0{,}0753 \qquad \frac{3}{N}\sum N_i(m_i-m)s_i^2 = -0{,}0165$$

$$s^2 = 1{,}4501 \qquad h^{(3)} = 1{,}5819$$

Als Schätzung für die Varianz erhalten wir $s^2 = 1\,4501$ gegenüber dem genauen Wert von 1 4489. Eine Schätzung ohne Berücksichtigung der durchschnittlichen Versicherungssummen pro Intervall ergäbe mit den Sheppardschen Korrekturen den Wert 1,4039.

Für die Schiefe erhalten wir $= \dfrac{h^{(3)}}{s^3} = 0{,}906$ gegenüber dem genauen Wert 0,911.

Zusammenfassung

Es werden möglichst genaue Abschätzungen für Varianz und Schiefe der Verteilung von Versicherungssummen gesucht, wenn die Anzahl Risiken und die Versicherungssumme pro Klassenintervall bekannt sind. Unter Verwendung der Information über die durchschnittliche Versicherungssumme in jedem einzelnen Intervall werden einfache Formeln mit Korrekturgliedern für den Fehler, den man bei der Konzentration der Versicherungssumme in den Schwerpunkten der Intervalle begeht, hergeleitet.

Summary

It is attempted to find the most exact estimates possible for variance and skewness of the distribution of sums assured when the number of risks and sum assured per class interval are known. By using the information on the average sum assured in each individual interval, simple formulae are derived with correction terms for the error produced by concentrating the sums assured at the mean values of the intervals.

Résumé

On peut arriver à une estimation aussi exacte que possible dans le calcul de la variance et de l'asymétrie de la distribution des sommes assurées lorsque le nombre de risques et la somme assurée pour chaque intervalle sont connus. En utilisant des sommes assurées moyennes on peut développer une formule avec termes correctifs pour réduire les erreurs produites en concentrant les sommes assurées au centre de gravité de chaque intervalle.

Riassunto

Si cerca di trovare una buona stima per la varianza e la dissimetria della distribuzione delle somme assicurate quando si conoscono il numero dei rischi e la somma assicurata per ogni classe d'intervallo. Utilizzando l'informazione sulle somme assicurate medie in ogni singolo intervallo, si possono dedurre semplici formule con termini correttivi per gli errori dovuti alla concentrazione delle somme assicurate ai valori medi degli intervalli.