CJJC
财经教材

全国高等院校财经类专业规划教材

李晓林　编著

风险统计模型

FengXian TongJi MoXing

中国财政经济出版社

全国高等院校财经类专业规划教材

风险统计模型

李晓林 编著

中国财政经济出版社

图书在版编目（CIP）数据

风险统计模型/李晓林编著．—北京：中国财政经济出版社，2008.9
全国高等院校财经类专业规划教材
ISBN 978-7-5095-0925-8

Ⅰ．风… Ⅱ．李… Ⅲ．保险-风险分析-统计模型-高等学校-教材 Ⅳ．F840.32

中国版本图书馆 CIP 数据核字（2008）第 133710 号

中国财政经济出版社 出版
URL：http：//www.cfeph.cn
E-mail：cfeph @ cfeph.cn

社址：北京市海淀区阜成路甲 28 号 邮政编码：100142
发行处电话：88190406 财经书店电话：64033436
北京富生印刷厂印刷 各地新华书店经销
787×1092 毫米 16 开 16.75 印张 410 000 字
2008 年 9 月第 1 版 2008 年 9 月北京第 1 次印刷
印数：1—3060 定价：35.00 元
ISBN 978-7-5095-0925-8/C·0006
（图书出现印装问题，本社负责调换）
本社质量投诉电话：010-88190744

编写说明

风险统计模型，又称风险模型，是非寿险精算学的核心内容，它揭示了对不确定事件的财务后果提供数量化意见的方法。它以概率统计为基础，运用多种模型，刻画了保险索赔额等多种损失变量的规律及其影响。

《风险统计模型》一书，是高等学校同名课程教材，也是非寿险精算基础课程教材。本书是在1999年出版的《风险统计基础》等书的基础上修订而成的。原教材的雏形是中央财经大学1993年开始的中英精算教育合作项目的成果之一，是作者在多位英国精算专家的支持和帮助下完成的讲义。本书在中央财经大学的精算专业教学中使用多年，是研究生入学考试的指定参考书之一。本书于2002年被立项为北京市精品教材。

全书共分十一章，分别是数据的整理、随机变量与随机向量、概率母函数和矩母函数、大数定律和中心极限定理、统计推断、风险模型、破产分析理论、贝叶斯统计推断、置信度理论、无赔款优待、递推三角形；内容涵盖了非寿险精算中主要的风险统计模型。

本书的编写，得到了数十位国内外精算专家学者的帮助；原书的出版，曾得到苏黎世金融服务集团的大力支持；修订中，我的同事和学生提出了许多很好的建议；本书的出版过程中，中国财政经济出版社的编辑们提出了宝贵的修改意见；在此，一并表示衷心的感谢。

由于作者水平所限，书中一定存在缺点，请广大读者不吝赐教。

作　者

2008年8月8日于北京

目录

第一章 数据的整理

在实际中，一个完整的统计调查一般包括以下几个步骤：收集数据、数据的整理、统计推断和结果报告。其中，数据的整理是统计调查的基础。本章将从数据的描述开始，探讨数据的整理方法。

第一节 数据的描述

我们先来明确几个概念。批数据（batch data），是一组相关的观察数据，例如：当前世界各国的通货膨胀率、我国各省的年度预算、某校各班级的学生数。

样本数据（sample data），是一组从总体中抽出的，同时代表那个总体的数据，例如：从某保险公司人身意外伤害险保单中抽出的100份保单，其保额组成的样本；从某保险公司汽车险索赔案中抽出的300个案例组成的样本；某养老金计划中180个被保险人的年龄组成的样本。

批数据分析是以数据的重要特征为指标对数据进行分类整理。样本数据分析，除了是为按数据的重要特征进行分类整理之外，还有一个重要的目的就是要作出关于样本总体的推断。本书的主要内容将与样本数据和推断有关。

数据涉及到有关变量的值，我们把变量分为如下几种类型：

1. 数值性（numerical）数据：包括（1）离散型数据，产生于计数（如精算师的人数、索赔的件数）；（2）连续型数据，产生于测量（如比率、数额、年龄）。

2. 范畴性（categorical）数据：包括（1）属性数据，只有两个类型（如是与否、男与女、索赔与不索赔）；（2）名义性数据，有多种不规则的类型（如保单的类型、索赔的性质）；（3）序列性数据，有多种不同程度的类型（如调查表显示诸如大力支持，……，强烈反对）。

下面将通过例子来做说明，大部分例子涉及数值性数据。

一、频数分布和条形图

[例 1-1] 假设某保险公司的由 80 份家庭储蓄保单组成的样本，其家庭中 16 岁以下儿童的人数如下：

2	1	3	1	1	4	5	2	2	1	4	5	4	2	2	0
3	2	2	2	2	2	2	1	2	3	3	1	1	4	3	2
1	3	0	3	0	0	3	2	3	2	2	2	2	3	4	3
3	1	6	2	2	1	3	0	2	3	1	7	4	0	0	5
2	2	4	3	1	3	3	2	0	3	2	2	2	5	2	2

试用简明的方式描述这些数据，并用恰当的图表表示出来。

解：这是一个典型的离散型变量，其可能值为 0，1，2，3，…。很明显，通过计算出 0，1，2 等的个数可以很好地描述这些数据的特征。通常，我们把这些数字的个数或者出现的次数称为频数，把这种描述方式称为频数分布。在此用 m 来表示频数，可将频数列表如下：

16 岁以下的儿童人数，x	样本中的家庭数，m
0	8
1	12
2	28
3	19
4	7
5	4
6	1
7	1
8 或 8 以上	0

显然，频数列表可以清晰地表述这些数据在其可能值上的分布。

条形图常常能比表格更好地表明数据的离散属性，如图 1-1。

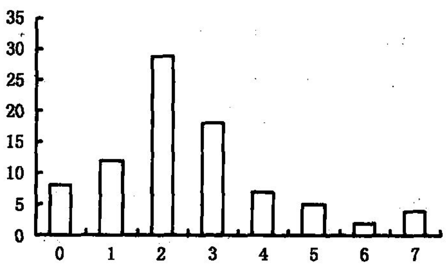

图 1-1

通过以上描述，对于这些数据是如何分布在其可能值上，我们已经有了一个清晰的印象，从而初步建立了该保险公司家庭储蓄保单中 16 岁以下儿童个数分布情况的概念。

二、群频数和直方图

群频数又称分组频数或组频数，我们通过实例来说明群频数和直方图的概念。实例中的数据是用最接近100元的方式分类的。现金额如用角或分的方式给出将是真正离散的，但在这里鉴于数额如此之大，我们可以认为它们是连续的。实际中，由于数据都只能以特定的精确度近似测量，而没有数据能以无限小的精度测量，所以所有数据都是离散的。

［例1－2］ 保险公司家庭财产保单中某100件由于渗漏引起的索赔组成的样本，索赔金额如下（单位：元）：

243	306	271	396	287	399	466	269	295	330
425	324	228	113	226	176	320	230	404	487
127	74	523	164	366	343	330	436	141	388
293	464	200	392	265	403	372	259	426	262
221	355	324	374	347	261	278	113	135	291
176	342	443	239	302	483	231	292	373	346
293	23	223	371	287	400	314	468	337	308
359	352	273	267	277	184	286	214	351	270
330	238	248	419	330	319	440	427	314	414
219	299	165	318	415	372	238	323	411	494

试用简明的方式描述这些数据，并用恰当的图表表示出来。

解：如果考虑频数分布，可能值太多了，因此我们将它们分组，并计算每组中的件数。数值从74到523，所以一个合理的分组是50－99，100－149，150－199，…，500－549等等。这将得到10组数据：

组　别	频　数
50－99	1
100－149	5
150－199	4
200－249	14
250－299	22
300－349	20
350－399	14
400－449	13
450－499	6
500－549	1

这就是有相同间距的群频数分布，我们称之为组容相同，由此可以对数据分布在各组上的情况有一个清晰的印象。

直方图能更好地表明现金额的几乎连续属性，如图1－2。

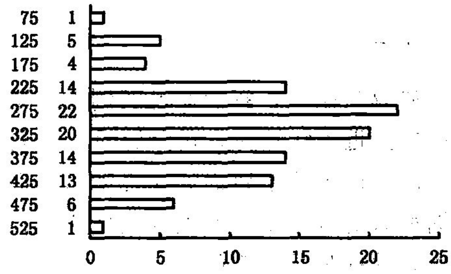

图 1－2

如此可以对这些数据的分布情况有清晰地了解，进而掌握公司此类业务的索赔额分布。

在例 1－2 中我们把数据分成了 10 组，这是舍弃细枝末节和获得清晰认识之间的妥协。如果分 5 组，会失去太多的细节，而分 20 组将不会有一个如此清晰的概括。

上述的群频数分布和直方图的组容相等。在某些情况下，在两端分一个或两个更宽的组可能会更方便。在这些情况下必须注明长方形的面积而不是高度更与频数成比例，原因在于直觉上的比较用的是面积。例如两个边长之比为 1:2的正方形，它们的面积之比为 1:4，眼睛很自然地判断它们之比为 1:4。

三、枝叶图

常用的另一种直方图为枝叶图。它的视觉印象与直方图相似，但是没有丢失组中数据的变化细节。这里是例 1－2 中数据的枝叶图（图 1－3）。

0	7
1	11344
1	6888
2	0122333344444
2	5666777777788999999999
3	00011122222233334444
3	55556677777799
4	000001122333444
4	677899
5	2

图 1－3

0	7
1	113446888
2	01223333444445666777777788999999999
3	0001112222223333444455556677777799
4	000001122333444677899
5	2

图 1－4

图中左边的枝单位是 100，右边的叶单位是 10。由此单位数据能被表达出来，虽然是以最接近 10 的数额近似表述出来。

构造一个枝叶图可以用与直方图相同的组容。图 1－3 的枝叶图我们称为半枝图，因为 100、200、300 等数据的两个半部分是分别描述的。图 1－4 的枝叶图称为全枝图。相比之下，半枝图的描述更清晰一些。

四、点图

对于更小的数据组常常用点图（dotplot）或线图（lineplot）来描述。其做法是把数据沿着一条有刻度的直线用点或叉号表示出来。

五、其他描述方式

1. 相对频数，即发生的频数或群频数与观察的次数的比值。它给出了数额为某值或属于某组的数据所占的比例。

例如，例 1-1 中的儿童人数的相对频数分布为：

16 岁以下的儿童人数	相对频数
0	0.1
1	0.15
2	0.35
3	0.2375
4	0.0875
5	0.05
6	0.0125
7	0.0125

在第二章你将看到，这是一个等价于随机变量概率的数据，它被看作是从数据中得到的经验概率。

2. 累积频数，即把数据小于和等于某值或某数据组的频数累加后的和（包括取其自身和比其更小值的频数或群频数）。

对分组的数据来说，我们很自然把累积频数与各组的上界联系起来。例如，例 1-2 中的索赔额分布，其累积频数为：

各组上限	累积频数
100	1
150	6
200	10
250	24
300	46
350	66
400	80
450	93
500	99
550	100

累积频数也能用图形表示出来。

我们也可以用相对形式来研究累积频数，在第二章将会看到数据的相对累积频数分布相似于随机变量的分布函数。

还有其他描述数据的方式，如饼分图和象形图等等，本书不一一叙述了。

但是必须指出的是一种图示中的危险倾向：一种特定的图能引起误导。我们可能在传媒上看过这样的图片，如瓶的图形被用来对不同时期某种酒的销量进行比较，视觉上的比较是被察觉出来的体积引起的，但是这类图片往往错误地用高度来代表频数。这与前面讲的正方形面积与长度的问题是类似的。

六、精确度

在描述数据时，我们需要考虑它的精确度，但并非在所有情况下越精确越好。例如，在例 1－2 的索赔额中，用角和分来记录数据是没有意义的。当然，如果不考虑其他因素，知道索赔额是 796.33 元，比知道是 796 元更好。

读者可能对“4 位小数”或“3 位有效数字”的概念已经非常熟悉。在风险统计中，我们更需要“重要数字”的概念。

例如，一个数集可能牵涉到计算的比率问题，该数集由以下数据组成：

1.0581	1.0366	1.0120	1.0404	1.0321
1.0156	1.0632	1.0026	1.0589	1.0333

由于所有数据都以 1.0 开头，出于比较的目的，主要考察第三和第四两位数字。实际上，把它们表示成超过 1 的百分数更好比较。于是，我们往往考察下列数据：

5.8　3.7　1.2　4.0　3.2　1.6　6.3　0.2　5.9　3.3

在另一个有如下数据的例子里：

33 232 元	7 677 元	65 652 元	86 675 元	98 329 元
40 020 元	65 526 元	4 484 元	85 113 元	52 886 元

用 1 000 元为单位进行比较会更清晰，即考察数据：

33　8　66　87　98　40　66　4　85　53

在报告中表达数据时，应避免过于精确，数字越简单，读者越容易理解和享用从数据中得到的信息。在实际中，一般用 2 或 3 个重要数字足够了。

第二节　数据分布位置的度量

通常，人们对数据分布情况关心的主要是数据分布的位置、密集或分散程度、偏斜情况等等。本节将探讨数据分布位置情况的各种度量。例如，看到例 1－2 的直方图，我们能发现索赔额集中在 300 元附近。第二节、第三节将探讨数据分布的密集或分散程度、偏斜情况。

一、均值

描述数据分布的最普通的度量是均值（mean）。严格来说应叫作算术平均值，因为还有其他“均值”，如调和均值和几何均值。但通常，我们还是把它简单地称为均值。

对一个有 n 个数据的数列：

$$x_1, x_2\cdots, x_n$$

或

$$x_i, i = 1, 2, \cdots, n$$

其均值是：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

读作“x - bar（巴）”。

实际上均值与大多数人常常称用的平均数是相同的概念。

对一个有可能值$x_1, x_2, \cdots, x_k$的频数分布，其对应频数为$m_1, m_2, \cdots, m_k$，其中 $\sum m_i = n$，其均值为：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i x_i$$

对一个群频数分布来说，其均值计算如上，不过这里的数值是各组的中点(中间值)①。

[例 1-3] 计算例 1-1 中的均值。

解：

$$\bar{x} = \frac{1}{80}\sum m_i x_i = \frac{186}{80} = 2.325$$

由于三位小数太多了，我们取 $\bar{x} = 2.3$。

[例 1-4] 以全部的数据和分组数据两种形式计算例 1-2 的均值。

解：用全部的数据计算：

$$\bar{x} = \frac{1}{100}\sum x_i = \frac{31\ 353}{100} = 313.53$$

我们取 $\bar{x} = 313.5$。

用分组数据计算：我们取各组中间值 75，125，…

$$\bar{x} = \frac{1}{100}\sum m_i x_i = \frac{31\ 750}{100} = 317.5$$

严格来说，组中值应是 74.5，124.5…，得出均值为 312.0，但注意到分组的误差，简单的选择是足够的。

由于分组的误差为 313.53 - 312.0 = 1.53，其相对误差为 0.5%。这个误差是由于分成群频数分布时丢失的细节造成的。

注意：上例中涉及了“过于精确”的概念。考虑到由于分组的误差或丢失的细节，用更高的精确度没有什么意义。

一般地，我们取比原始数据多一位小数作为均值。如果是一个较大的数列，我们可以考虑再多取两位小数，但不要取得更多。

---

① 在一般的统计著作中把它定义为样本均值，而且把它当作总体均值的估计值。如果这些数据是样本数据，则这是很明显的。如果这些数据是批数据，那么将没有相应的总体均值。然而，由于大部分是样本数据，本书一般地将其称为样本均值。

二、中位数

另一个有用的度量是中位数。把 n 个数据按大小排列，中位数是把这个数列分成两半的那个数，一半小于它，一半大于它；如果 n 是奇数，中位数就是中间的那个；如果 n 是偶数，中位数则是中间两数的中点（或平均值），可以表示成是第（n+1）/2 个数①。

中位数与均值是两个不同的概念。例如，有 5 个观测值：

1.1　　1.5　　1.6　　1.8　　2.2

均值为 1.64，中位数是 1.6，相当接近。然而，对于另外 5 个观测值：

1.1　　1.5　　1.6　　1.8　　202.2

均值为 41.64，中位数仍是 1.6，差距非常大。显然 202.2 对均值影响很大，而对中位数没有影响，这是中位数对考察某些数列的潜在优势之一，即它能在一定程度上抵消极端数据的影响。

中位数有时也能够更好地描述数据分布高度偏斜的数列——见第四节。

另一个中位数有用的地方是，不能确切地知道端点数据，只是简单地知道它大于或小于某些数据。

[例 1-5] 计算例 1-1 的中位数。

解：查看频数分布，你将看到第 40 位和第 41 位都是 2，所以中位数是 2。

注意：累积频数对确定中位数是有用的。中位数就是累积频数 50% 所对应的那个数，而这能从累积频数分布图中看出来。

[例 1-6] 利用下列方式计算例 1-2 的中位数：（a）按枝叶图计算；（b）按群频数分布计算，采用简单的插值法。

解：（a）参见枝叶图，得到第 50 和 51 位数分别为 310 和 320，所以中位数为 315②。

（b）中位数位于 300-349 这一组中，因为在这一组之前有 46 个数据，这一组中有 20 个数据，我们可以合理地假定这 20 个数据均匀地分布在 300-349 之间。因此用简单插值法来得到第 50.5 个数，得中位数为：

$$300+\left(\frac{50.5-46}{20}\right)\times 50=311.3$$

注意：对这个数列来说，均值和中位数很接近。这是由于数据分布相当匀称的缘故。

三、众数

第三个度量是众数。它被定义为：出现频数最多或最典型的数。它在实际中的应用是有限的，但偶尔也会用到，如保险公司对最典型的投保人感兴趣。

例如，在例 1-1 中，由频数分布，我们会看到出现频数最多的数值是 2，

---

① 没有一个标准的符号表示中位数，一些书用 M，另一些书甚至用 $\tilde{x}$。

② 实际上，从全部数据看，第 50 和 51 位数分别为 314 和 318，中位数为 316。

因此众数为 2（注意：中位数也为 2，均值为 2.3）。而在例 1－2 中，我们发现，250－299 这一组的频数为 22，接着是 300－349 的组为 20，于是，我们可以大概地判断出，其众数约为 300。

第三节　数据分布密集与分散程度的度量

本节将探讨数据在均值附近分布的密集与分散程度。

一、标准差

最常用的度量是标准差，它是反映数据离均值远近程度的度量。

考虑一个均值为 $\bar{x}$ 的数列 $x_i, i=1,2,\cdots,n$，对于数值 x_i，$(x_i-\bar{x})$ 是 x_i 到均值的距离，即 x_i 与均值的偏差。现在考虑这些偏差的均值：

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})$$

很容易证明它等于 0，因为正的和负的相互抵消了。

[例 1－7]　证明任何数列与均值的偏差之和为 0，即：

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})=0$$

证明：

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})=\sum x_i-\sum\bar{x}=\sum x_i-n\bar{x}$$

而

$$\bar{x}=\frac{1}{n}\sum x_i$$

因此

$$\sum(x_i-\bar{x})=0$$

显然，数列与均值的偏差之和无法度量数据的密集或分散程度。为此，我们取绝对值：

$$\frac{1}{n}\sum|x_i-\bar{x}|$$

上式叫作“均值绝对偏差”，这是一个明显的对密集或分散程度的度量，然而它在计量上很困难，因此是没有意义的。

所以，我们不取绝对值，而用平方，取偏差平方的均值，然后开方，这又回到了度量的初始维数。

于是采用：

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

即“偏差平方均值的根”，称为样本标准差或样本均方差。

由于统计推断的技术原因（见第五章），用（n－1）来代替 n，得到：

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

称作样本修正标准差，或称样本修正均方差，记为 s，而

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

称作样本修正标准差，或称样本修正均方差。

在不至于发生混淆的情况下，有时将样本修正方差也称为样本方差，样本修正标准差也称为样本标准差。

上述公式用手工计算很不方便，这牵涉到采用合适的精确度计算均值，再用所有的 x_i 减去它，平方后求和等。为此，我们采用一个更适合手工计算的公式：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2-n\bar{x})^2$$

或者：

$$\frac{1}{n-1}\sum_{i=1}^{n}\left[x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$$

我们只需要计算出 x 的和及其平方和就可以了。

$$\begin{aligned}(n-1)s^2&=\sum_{i=1}^{n}(x_i-\bar{x})^2\\&=\sum_{i=1}^{n}(x_i^2-2x_i\bar{x}-\bar{x}^2)\\&=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2\\&=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2(\text{由于}\sum_{i=1}^{n}x_i=n\bar{x})\\&=\sum_{i=1}^{n}x_i^2-n\bar{x}^2\end{aligned}$$

所以，上式成立。

在例 1－1 中，由频数分布可得：

$$\sum m_ix_i=186,\ \sum m_ix_i^2=592$$

$$s^2=\frac{1}{79}\left(592-\frac{1}{80}\times186^2\right)=2.02$$

$s=1.4$（取我们计算均值时所用的相同的精度）

而在例 1－2 中，可用全部的详细数据或分组后的群频数分布分别计算标准差。

采用全部数据计算可得：

$$\sum x_j=31\,353,\ \sum x_j^2=10\,687\,041$$

$$s^2=\frac{1}{99}\left(10\,687\,041-\frac{1}{100}\times31\,353^2\right)=8\,655.91$$

$s=93.0$（取与均值一样的精度）

对分组数据，像以前一样用 75，125 等组中值：

$$\sum m_i x_i = 31\ 250,\ \sum m_i x_i^2 = 10\ 637\ 500$$

$$s^2 = \frac{1}{99}\left(10\ 637\ 500 - \frac{1}{100} \times 31\ 250^2\right) = 8\ 866.82$$

$$s = 93.8$$

由分组导致的误差为：93.8 - 93.0 = 0.8，相对误差为 0.99。

作为分别描述数据分布的位置及密集（或分散）程度的指标，均值和标准差常常联合使用。下面的例子就是这样。

［例 1 - 8］ 对于数集 $x_i, i = 1,2,3,\cdots,n$，如果 $y_i = \alpha + \beta x_i$，说明 y 的均值和标准差与 x 之间的关系。

解：

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(\alpha + \beta x_i) = \frac{1}{n}\left(n\alpha + \beta\sum x_i\right) = \alpha + \beta\bar{x}$$

$$y_i - \bar{y} = (\alpha + \beta x_i) - (\alpha + \beta)\bar{x} = \beta(x_i - \bar{x})$$

因此，显然有

$$s_y^2 = \beta^2 s^2,\ s_y = \beta s_x$$

注意：α 只影响数据分布的位置而不影响密集或分散程度，这是一个在其他场合有用的重要观点。

我们有了用 $\sum x_i$ 和 $\sum x_i^2$ 表示 $\bar{x}$ 和 s^2 的公式，反过来，也可以用 $\bar{x}$ 和 s^2 来表示 $\sum x_i$ 和 $\sum x_i^2$。

显然，$\bar{x} = \frac{1}{n}\sum x_i \Rightarrow \sum x_i = n\bar{x}$

$$(n-1)s^2 = \sum x_i^2 - n\bar{x}^2 \Rightarrow \sum x_i^2 = (n-1)s^2 + n\bar{x}^2$$

［例 1 - 9］ 在例 1 - 2 中，$\bar{x} = 313.5$ 元，s = 93.0 元，假设另外一例 50 个索赔案的索赔额均值、标准差分别为 $\bar{x} = 327.4$ 元，s = 105.1 元。计算上述 150 个索赔案合并在一起的均值和标准差。

解：第一例 100 个索赔案：

$$\sum x = 100 \times 313.5 = 31\ 350,$$

$$\sum x^2 = 99 \times (93.0)^2 + 100 \times (313.5)^2 = 10\ 684\ 476$$

注意：相应的精确值为 31 353 和 10 687 041。

第二例 50 个索赔案：

$$\sum x = 50 \times 327.4 = 16\ 370,$$

$$\bar{x}^2 = 44 \times (105.1)^2 + 50 \times (327.4)^2 = 5\ 900\ 792.5$$

上述两例合并：

$$\sum x = 31\ 350 + 16\ 370 = 47\ 720,$$

$$\bar{x} = \frac{47\ 720}{150} = 318.1$$

$$\bar{x}^2 = 10\ 684\ 476 + 5\ 900\ 792.5 = 16\ 585\ 268.5$$