

精算学基础

第二卷

风险统计 基础

李晓林 编著

INGSUANXUE JICHU

中国财政经济出版社

精算学基础

第二卷

風金充十基出

4.0 a3

F224.0

L35a3

222

精算学基础

第二卷

风险统计基础

李晓林 编著



A0909096

中国财政经济出版社

图书在版编目 (CIP) 数据

精算学基础 第2卷: 风险统计基础/李晓林编著. - 北京:
中国财政经济出版社, 1999.5

ISBN 7-5005-4216-x

I. 精… II. 李… III. ①经济数学-计算方法②风险-统计基础 IV. F224.0

中国版本图书馆 CIP 数据核字 (1999) 第 17432 号

中国财政经济出版社出版

URL: <http://www.cfeph.com>

E-mail: cfeph@ drc.gov.cn

(版权所有 翻印必究)

社址: 北京东城大佛寺东街 8 号 邮政编码: 100010

发行处电话: 64033095 财经书店电话: 64033436

北京财经印刷厂印刷 各地新华书店经销

850×1168 毫米 32 开 11.875 印张 279 000 字

1999 年 6 月第 1 版 1999 年 6 月北京第 1 次印刷

印数: 1—3 000 定价: 32.00 元

ISBN 7-5005-4216-x/F·3825

(图书出现印装问题, 本社负责调换)

第一章 数 据

在实际工作中，一个完整的统计调查一般包括以下几个步骤：
①收集数据；②数据的描述；③正式的统计推断；④结果报告。
本章将从数据的描述开始，探讨数据的整理方法。

§ 1 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；

中央财经大学各班级的学生数。

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。

涉及到批数据的分析是为了按数据的重要特征为指标来进行分类整理。涉及到样本数据的分析，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：

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：

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，有好几种不规则的类型（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 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	3	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说出现的次数称为频数，把

这种描述方式称为频数分布。本书中的频数用 m 来表示，即：

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，对于这些数据在其可能值上的分布，我们可以通过上述列表的方法得到清晰的认识。

条形图可能比表格表达的更为恰当，它能更好地表明数据的离散属性。如图 1-1：

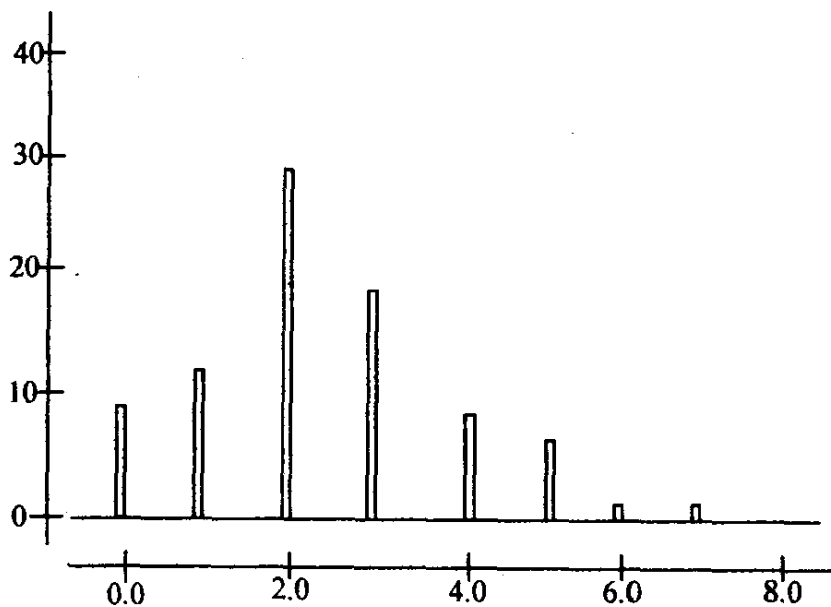


图 1-1

现在，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视觉印象，从而有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 100 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	236	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
291	299	265	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 523，一个合理的分组是 50—99，100—149，150—199，…，500—549，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情况有一个清晰的印象。

直方图的表达更恰当，它能更好地表明现金额的几乎所有连

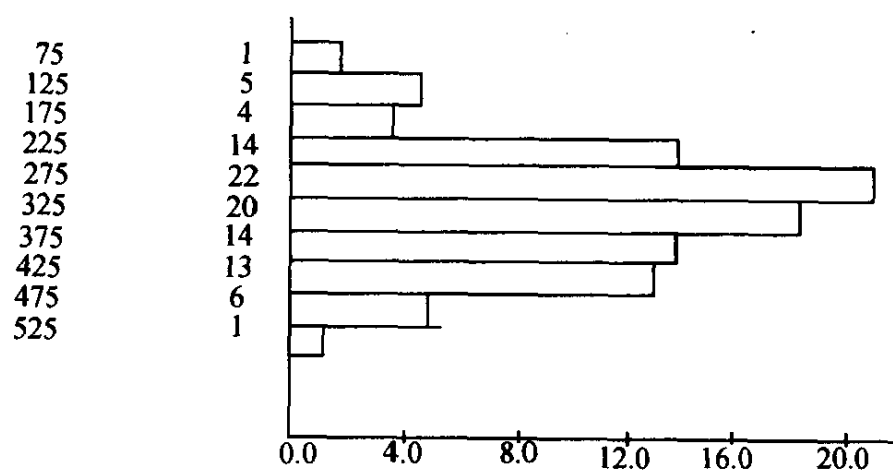


图 1-2

续属性。如图 1-2。

我们已经对这些数据的分布情况有了清晰的印象，进而了解了公司此类业务的索赔额分布的概念。

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

当今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）。

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。上面的枝叶图中的枝我们称为半枝图，因为 100、200、300 等数据的两个半部分是分别描述的。如下的全枝

0		7
1		11344
1		6888
2		0122333344444
2		566677777778899999999
3		00011122222233334444
3		55556677777799
4		000001122333444
4		677899
5		2

图 1-3

图（图 1-4）没有给出半枝图这样清晰的描述。

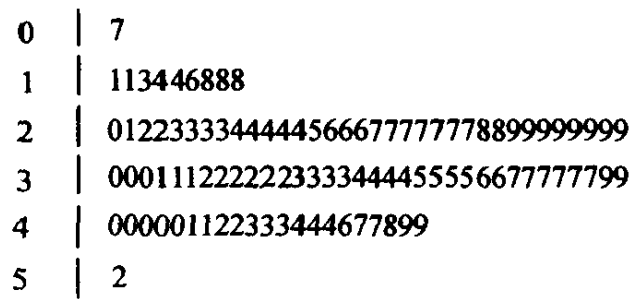


图 1-4

四、点 图

对于更小的数据组我们用点图（dotplot）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，它将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，本书不一一叙说了。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，我们不得不始终考虑它的精确度。例如，在例 1.2 的索赔额中，用角和分来记录数据是没有价值的。如果不

考虑其他因素，知道索赔额是 796.33 元，总比知道是 796 元好。

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。

§ 2 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。§ 3、§ 4 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 $m_i = n$, 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 $\bar{x} = 2.3$ 。

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。如果这些是样本数据, 则这是很明显的。如果这些是批数据, 那么将没有相应的总体均值。然而, 由于大部分是样本数据, 本书将一般地将其称为样本均值。

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二位小数，但一定不要多于此。

二、中位数

另一个有用的度量是中位数 (*median*)。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

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

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，这里中位数对考察某些数列的潜在优势之一，即它对极端数据的影响能在一定程度上抵消。

中位数有时也认为对数据分布高度偏斜的数列能够更好地描述（见 §4）。

另一个中位数有用的地方是，端点数据不能确切地知道，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) 参看枝叶图，得到第 50 和 51 位数分别为 310 和 320，所以中位数为 315。^①

① 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。

(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 (*mode*)。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，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。

§ 3 数据分布密集与分散程度的度量

本节将探讨数据是如何在均值附近分布的。

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 , $i = 1, 2, \dots, n$ ，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

考虑这些偏差的均值：

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

它等于 0，因为正的和负的相互抵消了，这很容易证明。

例 1.7

证明任何数列到均值的偏差之和为 0，即

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

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

但是
$$\bar{x} = \frac{1}{n} \sum x_i$$

因此
$$\sum (x_i - \bar{x}) = 0$$

显然，数列到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，也是没有意义的。

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

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

称为“偏差平方均值的根”。

由于统计推断的技术原因（见第八章），用 $(n-1)$ 来代替 n ，得到

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