Alibaba Group 阿里巴巴集团 | 技术丛书

Broadview www.broadview.com.cn

大数据丛书 | 十二五国家重点图书出版规划项目

算法优化

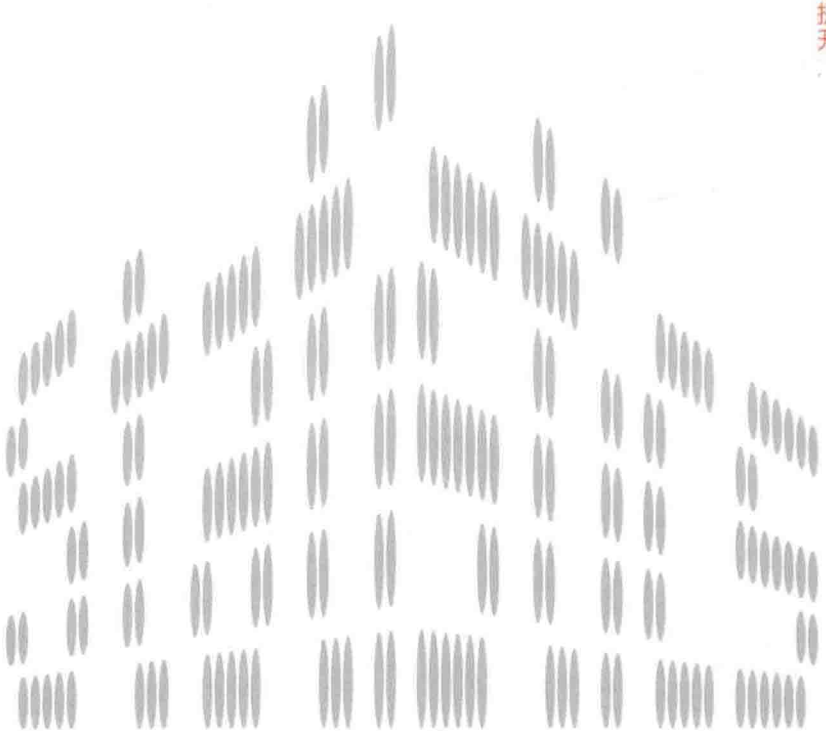
数据挖掘

数据分析

价值提升

重构大数据统计

杨旭 著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Alibaba Group 阿里巴巴集团 | 技术丛书

大数据丛书 | 十二五国家重点图书出版规划项目

重构
大数据统计

杨旭 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

大数据的统计计算是进行数据探索和分析挖掘的基础，在实际应用中会遇到两个问题：一个是需要使用多少资源；另一个是计算时间，它关系到数据探索分析的效率和效果。人们都希望花更少的钱，并且希望计算时间更短，但对于某个确定的计算过程，它们是成反比的。本书作者就是从统计计算的算法入手，重构其计算过程，从而同时降低资源使用量和计算时间。本书提出了一套完整的关于大数据统计的计算理论，包括常用的各种统计量和统计方法。基于本书内容开发的数据分析工具已经在阿里巴巴集团内部的多个部门使用，并取得显著效果。另外，本书还提供大量的示例程序代码帮助读者进一步了解算法细节，便于将书中的方法运用于实际计算。

本书适合对大数据分析感兴趣的读者阅读，本书前面章节比较容易理解，包含了常用统计量的计算；后面的各章节需要读者具备一些基础知识。建议读者根据自己的兴趣和工作需要，选择相应的内容进行参考。

图书在版编目（CIP）数据

重构大数据统计 / 杨旭著. —北京：电子工业出版社，2014.8
（大数据丛书. 阿里巴巴集团技术丛书）
ISBN 978-7-121-22500-0

Ⅰ. ①重… Ⅱ. ①杨… Ⅲ. ①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2014)第 030832 号

策划编辑：刘　皎
责任编辑：李利健
印　　刷：北京中新伟业印刷有限公司
装　　订：河北省三河市路通装订厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×980　1/16　印张：25.25　字数：502 千字
版　　次：2014 年 8 月第 1 版
印　　次：2014 年 8 月第 1 次印刷
定　　价：79.00 元

序　一

在大数据时代，数理统计是研究和挖掘数据价值的不可或缺的工具。尽管数理统计学科中有大量的理论研究成果，但如何将这些经典成果应用到大数据挖掘问题中，则还是近几年的事情。本书立足于将数理统计基础知识应用到大数据计算上，通过理论介绍和算法程序相结合的描述方法，把复杂的计算问题转变为相对简单、高效的计算过程。

本书提出了一套适合于分布式计算的统计计算方法，通过对一些基础统计量的计算，可获得有关数据集更全面的统计信息，进而可以进行高级的统计分析，例如，区间估计、线性回归、主成分分析等。可以这样说，本书从统计计算的角度，梳理出一套对大数据分析有实用价值的统计理论，并形成参考代码。

对于非数学背景，甚至非数理统计专业背景的读者来说，本书中所涉及的理论基础可能会比较陌生，特别是后半部分章节中提到的一些定理和推演过程。以我自己阅读为例，本书前半部分阅读起来比较愉悦和轻松，后半部分有些章节着实“不明觉厉”了。对于大数据分析的工程师或者分析师而言，本书的内容极其有参考价值，可以让你在先验统计知识的基础上，获得更多的统计信息，甚至捕捉到数据集内在的规律。

我也相信，大多数读者并不具备数理统计专业的基础，但这不会成为阅读本书和使用高级统计量的障碍。相反，阅读本书会开阔大数据分析的思路，学习统计分析的理论基础，并快速找到大数据统计分析的正确方法。建议每一位正在从事大数据分析的工程师都读一读这本书。

本书作者杨旭曾经获得了南开大学数理统计专业的博士学位，并先后在微软亚洲研究院和阿里巴巴从事高性能计算和大数据计算方面的工作。两年前，他曾经向我提到，他计划将工作中积累的算法经验写成一本书，到了 2014 年 1 月，他告诉我书稿已经完成，询问我能否作序或

写推荐。我粗略看了一下目录，认为对于大数据计算非常有价值，也正好有兴趣学习这方面的知识，所以答应春节假期阅读完书稿后再做决定。

是以作序推荐。

潘爱民

阿里云 OS 首席架构师

序　二

相信大多数从事数值计算的技术人员都读过一系列叫作 *Numerical Recipe in C/Fortran/Java* 的书。眼前这本《重构大数据统计》或许可以叫作 *Numerical Statistics Recipes in Java*，和那个系列一样，对于重新快速学习和使用数值统计算法来实现数据分析应用的读者来说，这本书是一本不错的入门手册。

数理统计和多变量统计分析的数值算法存在已久，并不是新的 Rocket Science。无论冠以数据分析还是大数据分析之名，本书所涵盖的统计量和统计方法的知识都是数据分析人员必须具备的基本功。大数据分析是一门应用科学，根本原因在于互联网和计算技术，硬件技术的发展使得海量数据的产生和实时持续处理成为可能，这样由海量数据驱动的数据挖掘、机器学习等基于传统的统计数据分析方法真正成为基于理论建模、实验和数值模拟之外新的范式。认识到这一点就会对大数据持平常心，学习的核心就在于如何掌握数理统计和多变量分析在大数据应用框架下的实现和性能优化。本书简明扼要地介绍了常用的数理统计算法，重点介绍了数值算法的实现。

囿于篇幅，本书未能介绍每一种统计量和统计方法在实际的大数据分析场景中的可能应用，我们期待作者后续能完善这方面的研究，将本书的下一版内容提升到应用指南的层面。

何万青博士

英特尔技术计算集团技术计算架构师

前　言

大数据的统计计算是进行数据探索和分析挖掘的基础。在实际应用中，随着数据规模的快速增长，数据会分布式存储在多台计算机上，即使最简单的求和操作都需要多台计算机协同完成，并且需要分钟级别的计算时间，这样我们将面对如下两个问题：

- 需要使用多少资源，即所要付出的成本。
- 计算需要多少时间，它关系到数据探索分析的效率和效果。

人们都希望花更少的钱，并且希望计算时间更短，但对于某个确定的计算过程，它们是成反比的。这就是我在大数据统计实践中经常遇到的情况，当研究了一个又一个计算性能问题之后，我惊奇地发现，这些不同的算法间是有共性的，这就吸引我去深入研究，最终形成了一套完整的理论，包括常用的各种统计量和统计方法。基于本书内容开发的数据分析工具已经在阿里巴巴集团内部的多个部门使用，并取得了显著效果。

首先举一个例子，使大家有一个直观的印象：对于 10TB 的数据，大约有 1.25 万亿个数据，以求和计算为例，使用很多人熟悉的分布式 SQL 进行计算：

```
SELECT SUM(COL1) AS COL1_SUM, COUNT(COL2) AS COL2_CNT, … FROM DATA_TABLE;
```

共运行了 4 分 44.062 秒，在该 SQL 语句中，每列只算了一个统计量。

然后使用本书的计算方法，计算更多的统计量，包括：总个数、总和、均值、方差、标准差、标准误、变异系数、立方和、四次方和、二阶原点矩、三阶原点矩、四阶原点矩、二阶中心矩、三阶中心矩、四阶中心矩、偏度、峰度；最大值、最小值、极差、最大的 100 个值、最小的 100 个值；数据分布直方图、经验分布函数、近似百分位值。如果不同数值的个数小于 10000 个，会将其频数信息计算出来，并有精确百分位值、中值、众数；协方差矩阵、相关系数矩阵。

得到所有的这些统计量使用的计算节点数目与用 SQL 语句获得的基本相同，花费的计算时间为：4 分 53.673 秒。计算这么多内容才多花约 10 秒，说明本书介绍的算法够高效吧！但这还只是一个开头。

接下来，做一个更有挑战的实验，除了上面这些统计量，我们再加入一些高级的统计计算，区间估计、参数检验、非参数检验、线性回归、共线性分析、方差分析、主成分分析，完成这些需要多久呢？答案是 4 分 53.766 秒。多么神奇的事情！对这些大数据进行高级统计计算只多用了不到 0.1 秒。这种计算效率的提高够显著了吧，我们无须再为资源和时间发愁了。

本书通过文字描述、数学表达式和程序代码，将整个统计计算过程清晰地展现在读者眼前。全书揭示了各种统计概念和方法，以及它们内在的关联，并根据其特点，对各自的计算公式进行恒等变换，找到更适合大数据的计算方式。书中提供的示例程序代码可以帮助读者进一步了解算法细节，便于将书中的方法运用于实际计算。

本书适合对大数据分析感兴趣的读者阅读，本书前面的章节比较容易理解，包含了常用统计量的计算；后面的各章节需要读者具备一些基础知识，建议读者根据自己的兴趣和工作需要，选择相应的内容进行参考。

在本书编写过程中，感谢初敏、陈一宁、张东晖的支持和帮助，感谢蔡宁、高志涵在算法方面的讨论和交流，感谢邓钟强、蔡宁、高志涵、蒋耘、罗毅、谭望达、代斌、周俊、王少萌、姜晓燕、王乐珩、曹传宇等同事，一同将此理论应用于实际。

因作者水平有限，书中难免有不妥或疏漏之处，敬请广大专家和读者批评、指正！我的电子邮件地址为：yangxu@alibaba-inc.com。

杨旭

2014 年 7 月

目　录

第 1 章　基本概念 ····· 1

1.1　数据类型 ····· 1

1.2　总体和样本 ····· 2

1.3　参数和统计量 ····· 2

1.4　分布式计算 ····· 3

第 2 章　单变量基本统计量 ····· 5

2.1　数量统计量 ····· 5

2.1.1　样本方差为何除以 *n*-1 ····· 7

2.1.2　数据分布与标准差的关系 ····· 10

2.1.3　新的计算公式 ····· 11

2.1.4　代码实现 ····· 16

2.2　频数统计量 ····· 18

2.3　次序统计量 ····· 23

2.3.1　通过排序方法计算次序统计量 ····· 25

2.3.2　不需排序就可计算的次序统计量 ····· 29

2.3.3　基于频数信息计算次序统计量 ····· 31

2.3.4　中位数、众数和均值的关系 ····· 34

第 3 章　单变量数据的分布 ····· 36

3.1　直方图 ····· 36

3.1.1　直方图的计算 ····· 39

3.1.2　算法实现 ····· 42

3.1.3　已知数据频数的情况下求直方图 ····· 49

3.1.4 日期类型直方图……49
3.2 经验分布……57
3.3 近似分位数和近似百分位数……61
3.4 PP、QQ 概率图……65
3.5 单变量的基本统计信息……69
第 4 章 多变量的数据特征……77
4.1 协方差……77
4.2 相关系数……79
4.3 协方差和相关系数的计算实现……80
4.4 数据表的基本统计结果……84
第 5 章 数据探索……88
5.1 扩展直方图……88
5.1.1 计算方法……90
5.1.2 代码实现……91
5.2 交叉表……110
第 6 章 极限定理……116
6.1 大数定理……116
6.2 中心极限定理……117
第 7 章 常用的分布函数介绍……123
7.1 基本定义……123
7.2 标准正态分布（Z 分布或 U 分布）……124
7.3 卡方分布（χ^2分布）……129
7.4 学生 T 分布……133
7.5 F 分布……139
第 8 章 常用分布函数计算……145
8.1 函数定义……145
8.2 函数性质及相互间的关系……147
8.3 分布函数关系图……164
8.4 分布函数的计算……166
8.4.1 计算$\Gamma(x)$……166

8.4.2 计算CDF_Γ ····170
8.4.3 计算CDF_B ····173
8.4.4 计算IDF_Γ和CDF_B ····176
8.4.5 其他函数的计算 ····178
8.5 生成常用分布的随机数 ····180
第 9 章 参数估计 ····187
9.1 点估计与区间估计 ····187
9.2 单个总体的参数估计 ····190
9.2.1 不同情况的参数估计表达式 ····190
9.2.2 单个总体参数估计的实现 ····191
9.3 两个总体的参数估计 ····196
9.3.1 不同情况的参数估计表达式 ····196
9.3.2 两个总体参数估计的实现 ····199
第 10 章 假设检验 ····207
10.1 基本概念 ····207
10.2 参数检验 ····209
10.3 单个总体参数的检验 ····212
10.3.1 各种情况下的检验方法 ····212
10.3.2 单个总体参数检验方法的实现 ····214
10.3.3 不同检验方法的选择 ····223
10.4 两个总体参数的检验 ····227
10.4.1 各种情况下的检验方法 ····227
10.4.2 两个总体参数检验方法的实现 ····231
10.4.3 不同检验方法的选择 ····237
第 11 章 非参数检验 ····244
11.1 Pearson 拟合优度χ^2检验 ····245
11.2 两个变量的列联表检验 ····248
11.3 K-S 检验 ····250
11.3.1 单样本 K-S 检验 ····251
11.3.2 双样本 K-S 检验 ····256
11.4 符号检验 ····258
11.5 秩统计量和秩检验方法 ····260
11.5.1 Wilcoxon 秩和检验 ····260

11.5.2 Wilcoxon 符号秩和检验 ……266
11.5.3 Kruskal-Wallis 检验 ……268
11.5.4 Friedman 检验 ……273

第 12 章 方差分析 ……277

12.1 单因素方差分析 ……278
12.1.1 计算流程 ……278
12.1.2 代码实现 ……280
12.1.3 方差分析与 T 检验的关系 ……283
12.1.4 方差分析中的多重比较方法 ……285
12.2 双因素方差分析 ……289
12.2.1 无交互作用的双因素方差分析 ……289
12.2.2 有交互作用的双因素方差分析 ……295

第 13 章 多元线性回归 ……302

13.1 数学模型 ……302
13.2 显著性检验 ……308
13.3 计算步骤 ……309
13.4 代码实现 ……313
13.5 多重共线性 ……320
13.5.1 度量指标 ……320
13.5.2 代码实现 ……323
13.5.3 应用示例 ……328
13.6 逐步回归 ……330

第 14 章 主成分分析 ……340

14.1 计算步骤 ……342
14.2 代码实现 ……345
14.3 应用举例 ……350

第 15 章 判别分析 ……359

15.1 距离判别 ……359
15.1.1 Mahalanobis 距离 ……360
15.1.2 模型训练和预测 ……361
15.2 Fisher 判别 ……364
15.3 Bayes 判别 ……369

15.3.1 朴素 Bayes 判别……369
15.3.2 模型训练和预测……370
15.4 判别算法的综合模型……377
15.5 应用举例……378
第 16 章 模型评估曲线……383
16.1 相关概念……383
16.2 定义……384
16.2.1 ROC 曲线……384
16.2.2 上升图和反馈率—精确率线……386
16.3 计算实现……386
参考文献……391

第1章 基本概念

本章将简要介绍一些最基本的概念：数据类型、变量、总体、样本、参数和统计量，并通过具体的例子和说明介绍本书常用的分布式计算思想，为读者深入阅读本书作铺垫。

1.1 数据类型

我们接触到的数据有很多，例如：用户姓名、性别、交易金额、商品单价、用户评分、交易时间等。按照所采用的计量尺度不同，可以分为三类：名义数据、有序数据和数值型数据。

1．名义数据

名义数据（Nominal Data）是指对事物分类的结果不区分顺序，但有分类尺度计量形成的数据。

各个名义数据间无大小、高低和等级之分，唯一可行的是对发生的频数进行计算。例如，用户姓名和性别都为名义数据。名义数据可以用数字表示，例如，1 表示男，0 表示女。显然，这里的 1 并不意味着比 0 大。

2．有序数据

有序数据（Ordinal Data）是指对事物分类的结果有顺序、有分类尺度计量形成的数据。

该类型数据可以进行排序操作，也可以对发生的频数进行计算。例如：用户评分（好、中、差），受教育水平（小学、初中、高中、大学及以上）。有序数据也可用数值表示，例如：对评分用 3 表示好，2 表示中，1 表示差；对受教育水平用 1 表示小学，2 表示初中，3 表示高中，4

表示大学及以上，其中的 4 意味着比 2 受教育水平更高。其数值计算结果也没有意义，例如：1+1+1= 3 不能说明 3 个差评等于一个好评；2+2=4 不能说明受了两次初中教育相当于大学毕业。

3. 数量数据

数量数据（Quantitative Data）是按自然单位、度量衡单位、价值单位对事物进行测量的结果，该结果表现为具体的数值，取值为实数，可以进行所有的计算（求和、平均值等 ），包括前两种数据类型的排序和计算发生的频数。例如：购买商品的个数、交易金额等。

上述三种数据类型的关系如图 1-1 所示。

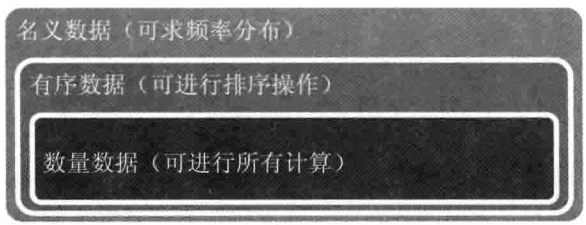


图 1-1

1.2 总体和样本

说明事物某种特征的概念，称为变量（因素或者元）。例如：灯泡的寿命、购物时间、物品单价、物品个数等。

个体是由一个或多个变量（多元或多个因素）构成的。例如：某个灯泡的寿命是 1200 小时；一条网购记录为“用户名称：张三；购物时间：2013-9-1；物品单价：99.99；物品个数：10”。

包含所研究的全部个体的集合，称为总体。

对于所要研究的总体，通过观测或试验而得到的个体集合$X_1, X_2, \cdots, X_n$，称为样本。这里的X_i称为第i个样本，n称为样本大小或样本容量（Sample Size）。

通常，我们用表格来记录个体的集合，表格的行数对应个体的数量，每一列对应一个变量。

1.3 参数和统计量

用来描述总体特征的概括性数字度量，称为参数（Parameter）。

例如：某工厂生产的一批灯泡，把它们看作一个总体，灯泡的平均使用寿命就是一个重要的参数。但需要测试整批灯泡的寿命，才可以得到这个参数，得到参数的同时，这批灯泡也就都费掉了。我们能否只拿一小部分的灯泡来测试，从而估计出这个参数呢？这就需要下面的概念：样本统计量。

用来描述样本特征的概括性数字度量（简单地说，就是由样本计算出来的量），称为统计量（Statistic）。

例如：在上面的例子中，从这一批灯泡中抽样出 20 个，测试并计算其平均使用寿命为 1200 小时，则整批灯泡的平均使用寿命应该在 1200 小时“附近”。如何精确描述和确定这个“附近”值？这需要统计推断的一个重要内容是：参数估计。相关内容将在后面详细介绍。

1.4　分布式计算

分布式计算（Distributed Computing）是将大的计算任务（需要巨大的计算能力或需处理巨大的数据量）分解成许多小的部分，并分配给许多计算机进行处理，最后将所有小部分的计算结果综合起来得到最终的结果。

我们所说的大数据（如几百亿条的交易记录）一般就存储在由很多机器构成的分布式存储系统里，而这些机器同时也是我们进行数据分析的计算资源。如何通过分布式计算进行高效的统计计算，就是本书的重点。

用数学计算的例子来解释分布式计算，例如，要计算下面的式子：

$$a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7 + a_8 + a_9 + a_{10} + a_{11} + a_{12} + a_{13} + a_{14} + a_{15}$$

利用加法结合律，我们有

$$(a_1 + a_2 + a_3 + a_4 + a_5) + (a_6 + a_7 + a_8 + a_9 + a_{10}) + (a_{11} + a_{12} + a_{13} + a_{14} + a_{15})$$

对于新的表达式，每个括号内的内容可以分别计算，最后将 3 个中间结果再进行加法计算，得到最终结果。通过括号将原本的计算问题分成 3 个小部分，每个小部分都可以独立进行计算，它是分布式计算能够处理大问题、提高计算速度的关键。而括号间的加号代表汇总过程，它保证我们可以得到原问题的正确结果。

上面的例子说明了数据的求和可以通过分布式方法进行计算。特别地，当每个a_i都取 1 的时候，就相当于求数据的总个数，每个小部分独立地计算出各自包含的数据的个数，再汇总到一起，就是数据的总个数。

分布式计算可以求得数据的总个数和数据的和。那么对于数据的平方和呢？下式：

$$a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2 + a_6^2 + a_7^2 + a_8^2 + a_9^2 + a_{10}^2 + a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{14}^2 + a_{15}^2$$

也使用加法结合律，有

$$(a_1^2 + a_2^2 + a_3^2 + a_4^2 + a_5^2) + (a_6^2 + a_7^2 + a_8^2 + a_9^2 + a_{10}^2) + (a_{11}^2 + a_{12}^2 + a_{13}^2 + a_{14}^2 + a_{15}^2)$$

对于新的表达式，每个括号内的数据都是总体数据的一个小部分，可以独立进行计算；3个括号内的计算完成后，就可进行汇总，即进行括号间的加法运算，得到全部数据的平方和结果。类似的，我们可以使用分布式计算得到数据的立方和、数据的四次方和。

我们再看一下另外两个非常熟悉的统计量：最大值和最小值。它们该如何通过分布式计算得到？仍以数据a_i为例进行说明，数据a_i的最大值表示为：

$$\max(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, a_{14}, a_{15})$$

这个表达式意味着计算时，从左边的两个数据a_1、a_2开始，选出最大的，然后和下一个数a_3进行比较，留下最大的数，再比较下一个，……，直到比较完最后一个数。由于整体数据的最大值一定包含该数据的任意个小部分的最大值，且所有部分的最大值中最大的就是整体数据的最大值。有如下等价的表达式：

$$\max\{\max(a_1, a_2, a_3, a_4, a_5), \max(a_6, a_7, a_8, a_9, a_{10}), \max(a_{11}, a_{12}, a_{13}, a_{14}, a_{15})\}$$

根据这个表达式，每个小部分分别求出最大值，然后所有小部分的最大值中最大的就是我们要求的结果。

同样也适用于计算最小值

$$\min(a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, a_{14}, a_{15})$$
$$= \min\{\min(a_1, a_2, a_3, a_4, a_5), \min(a_6, a_7, a_8, a_9, a_{10}), \min(a_{11}, a_{12}, a_{13}, a_{14}, a_{15})\}$$

综上所述，使用分布式计算可以很容易求得数据的总个数、和、平方和、立方和、四次方和、最大值和最小值。对于其他统计量（如方差、峰度等），该如何计算呢？接下来的章节就会回答这个问题。