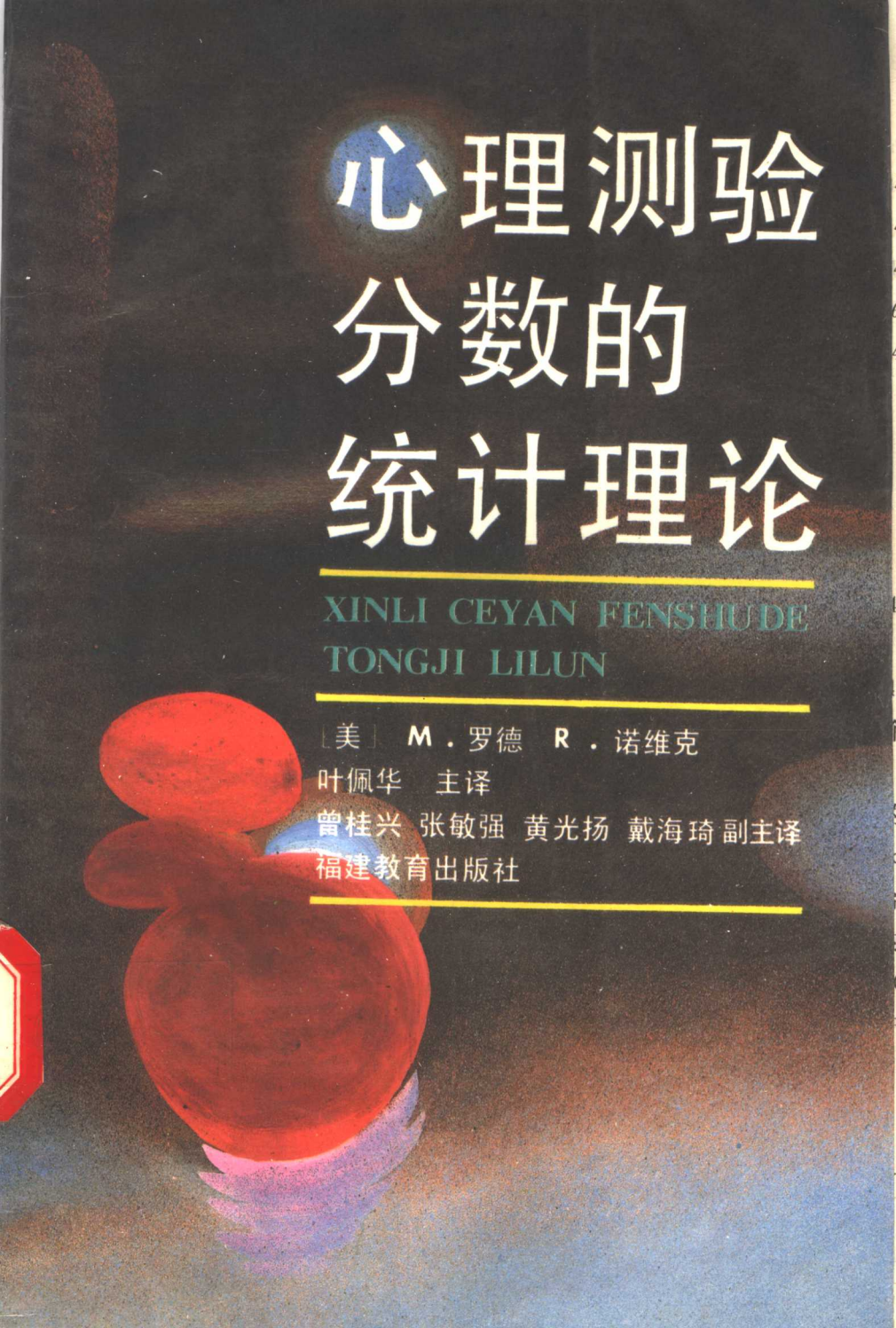
心理测验分数的统计理论

XINLI CEYAN FENSHUDE TONGJI LILUN

[美] M．罗德 R．诺维克
叶佩华 主译
曾桂兴 张敏强 黄光扬 戴海琦 副主译

福建教育出版社



心理测验分数的统计理论

[美]M. 罗德　R. 诺维克　著
叶佩华　　主　译
曾桂兴　张敏强
黄光扬　戴海琦　副主译

福建教育出版社
1991年·福州

(闽)新登字 02 号

STATISTICAL THEORIES OF MENTAL TEST SCORES

Frederic M. Lord and Melvin R. Novick

Addison—Wesley Publishing Company, Inc. 1968

心理测验分数的统计理论

M. 罗德 R. 诺维克 著

叶佩华 主译

曾桂兴 张敏强 黄光扬 戴海琦 副主译

福建教育出版社出版发行

(福州大梦山 7 号)

福州晚报印刷厂印刷

开本 850×1168 1/32 印张 20.5 字数 476 千

1992 年 2 月第 1 版 1992 年 2 月第 1 次印刷

印数:1—3,300

ISBN 7—5334—0748—2/G・448 定价:12.40元

译者的话

心理与教育测量在人才选拔及教育决策过程中，有着极其重要的作用。目前，心理与教育测量的方法在我国教育界得到了较为广泛的应用。如应用心理与教育测量的理论指导高考改革，进行高考标准化试验；应用心理与教育测量的方法指导教学改革及考试改革，等等。这些，在近年来都取得了一定的成效。

但是，作为心理与教育测量这一门学科，是基于一定的公理和原则建立起来的理论和方法体系。所以，了解和掌握心理与教育测量的理论和方法体系，对正确应用心理与教育测量的理论和方法，对心理与教育测量结果的解析和分析，都是非常重要的。

纵观我国近年来出版的心理与教育测量的专著和译著，它们都较少涉及到心理与教育测量的理论和方法体系问题，故在心理与教育测量的理论和方法的应用中，常有误用或错用的情况。因此，为了使教育工作者能对心理与教育测量的理论和方法有系统的认识，并能了解和研究新近发展起来的各种测验理论，在叶佩华教授的主持下，我们集体翻译了美国当代著名的心理与教育测量学家及统计学家 M. 罗德和 R. 诺维克合著的《心理测验分数的统计理论》一书。该书系统地介绍了心理与教育测量的理论与方法体系，介绍了各种计算公式的来源及使用条件和范围，还介绍了近 20 年才发展起来的题目反应理论。该书在美国多次重版，堪称心理与教育测量方面的权威性著作。我们翻译此书的目的，就是希望

能对我国的心理与教育测量的理论研究和应用有所推动。

本书是集体翻译的。叶佩华教授主持了翻译工作并统校全书。曾桂兴同志负责翻译过程的具体组织工作。各章的译者如下：曾桂兴：前言、目录、导言、符号；景民、曾桂兴：第一章；景民：第二章；徐虹、曾桂兴：第三章；徐虹：第四章；王力发、黄光扬：第五章；王力发：第六章；黄光扬：第七、八章；龙文祥、黄光扬：第九章；龙文祥：第十章；李大红、戴海琦：第十一章；李大红：第十二章；亚新：第十三章；任建胜：第十四章；戴海琦：第十五、十六章；臧铁军：第十七、十八章；李域明、张敏强：第十九章；李域明：第二十章；任建胜、张敏强：第二十一章；张敏强：第二十二章；汪艳丽、张敏强：第二十三章；汪艳丽：第二十四章。在翻译过程中，为保证初译稿的质量，由曾桂兴、张敏强、黄光扬、戴海琦四位同志分工对全书的术语符号及有关章节内容再次进行修改与校译，随后由黄光扬同志对译著中的语言表达、术语与符号译写的规范要求以及可能存在的笔误作了统核。

因参加翻译的人员较多，本书的专业性较强，加上译者水平所限，故书中难免还会出现错漏，敬请读者行家予以指正。

译　者

1987年7月于广东省教育科学研究所（初稿）

1989年5月于华南师大考试研究中心（定稿）

前　言

自 Harold Gulliksen 的《心理测验理论》出版以来，已过去 15 年多了。这个时期，心理测验理论已取得了飞速的发展，因而有必要对心理测验理论方面进行新的综合处理。Gulliksen 教授看出这种必要已有好几年了。在一定程度上，编写本书是由于他建议和鼓励的结果。这几年来我们也看到测验理论日益依赖于有关的数理统计模型，因此显而易见的是，对测验理论的任何最新的综合处理，都必须以这些模型的陈述和推导为基础。

在寻求理论方面进行新的综合处理中，我们最初的尝试是：对有关模型提出非常仔细的数学陈述，将对许多已经研究过的课题提供另外的见解。为此，在编写本书的过程中，我们在这个方向试图作更仔细的数学陈述。然而，我们又试图避免无助于测验理论理解的专门性数学讨论。因此，我们可以保持本书的各部分处于更适度的数学水平上，而不是在某个看来合理的论点上。

我们工作的一个主要部分是重申和精炼许多人的著作，使之成为一种综合发展。然而，本书又要体现各原作者的理论倾向和兴趣。为此，在某种程度上说，分配给各课题的相对篇幅量正是由于这一原因。我们把那些围绕本书目标且已发表的研究，综合成为我们的论述，这样本书就包含了大量的测验理论文献。我们自己的某些迄今尚未发表的研究也包括在内。对已发表的有关文章且主要不属于我们探讨范围的研究，我们一般只提及而不讨论，本书没有

提及的某些基本论文和许多专门论文，其原因或者是这些著作不适合我们的探讨，或者这些主题超出我们任务的范围。因此，包括或排除任何特定论文都不应理解为对该论文价值的评价。

在计划编写本书期间，我们意识到有一项重要的研究（那时仍未发表）应包括在心理测验理论的综合处理之中。这项研究是 Allan Birnbaum 在潜在特质理论方面的著作（包括他的 Logistic 反应模型）。我们确实很幸运，因为 Birnbaum 教授已同意在本书中首先发表这种题材，他以自己的方式得心应手地发挥了他的贡献。我们曾与他密切合作，努力综合他的题材，使之成为本书发展的主流。本书余下各章由署名人编写，而对这些章的处理、对投稿人著作的选择和综合，以及本书的提纲、格式、观点等，则由我们单独负责。

本书提出的许多概念是在教育测验中心（ETS）心理测验研究组举办的数学心理学研究班上孕育的。这种研究班的历次成员有 Allan Birnbaum、Michael W. Brown、Karl Jöreskog、Walter Kristof、Michael Levine、Frederic M. Lord、Willam Meredith、Samuel Messick、Roderick McDonald、Melvin R. Novick、Fumiko Samejima、J. Philip Sutcliffe 和 Joseph L. Zinnes. 在我们多次讨论统计模型与心理学的理论问题、方法论问题之间的关系时，由于多次与 Norman Frederiksen、Nathan Kogan 和 Samuel Messick 交谈而受益。第三部分则由于 Julian Stanley 的精心校阅而受益。我们要感谢 Louis Guttman 教授为我们提供了他未发表的 1953 年的手稿《短评：信度与效度的概念和代数学》。Frederick Mosteller 和 Robert L. Thorndike 审阅了全部原稿，从他们的建议中我们得到极大的好处，谨向他们致以深切的感谢，而对那些可能遗留的错误，我们不把责任强加在他们身上。

在过去的两年里，本书初版的许多章节曾被下列各大学和测验理论研究班作为课程教材。

大　学	教　师
芝加哥大学	R. Darrell Bock、David Wiley、Benjamin Wright
哈佛大学	John B. Carroll
伦敦大学学院	Melvin R. Novick
北卡罗来纳州大学	Murray Aitkin
安大略教育研究院	Ross E. Traub
宾夕法尼亚州大学	Melvin R. Novick
普林斯顿大学	Frederic M. Lord、Melvin R. Novick
斯坦福大学	Lee J. Cronbach
田纳西州大学	Edward Cureton

此外，我们也收到用过本书稿的教师和学生的评论，这些评论对我们编写最新的书稿大有裨益。

在检查手稿的错误方面，Dorothy Thayer 夫人起了重要的作用，她在编拟或检查大多数习题以及编制图表方面承担了主要的责任。Charles Lewis 先生提供了许多深刻的评论，这些评论迫使我们在许多地方加强了论证。Carl Frederiksen、Jon Kettenring、Philip Piserchia 和 Larry G. Richards 先生等，以暑期研究助教的身份帮助我们找出初稿的错误及模糊之处。为本书打印书稿是件很乏味的工作，它由 Beatrice Stricklin 夫人和 Kathleen Rohe 小姐熟练地完成了，Mary Evelyn Kunyon 和 Eleanor Hibbs 夫人也临时帮助了我们；Fay F. Richardson 夫人协助我们作长条校样。Ann King 夫人核检到最后一页的校样，并在整个编写计划期间提供了无法估计的编辑帮助。Michael Friendly 先生仔细地检查了许多章的印刷错误。Sara B. Matlack 夫人安排和检查了完成本计划所需的大量资助。

我们要感谢 Addison-Wesley 出版公司的所有人，他们在完成本书计划中，为我们提供了许多细致的技术帮助；令我们深谢的

是，他们对我们的许多请求均表示同情并且尽力做到。

我们要感谢海军研究局准许转述他们的某些研究。自 1952 年起，人事培训部门在一定程度上支持了这种测验理论研究。编写本书的资金部分地来自 logistic 和数理统计部门的资助。

教育测验中心(ETS)对本书计划提供了很大的支持，我们非常感谢他们为促使本书的成功所做出的努力。我们要特别感谢心理研究处的执行副总裁 William W. Turnbull 和处长 Norman Frederiksen，是他们为基础研究创立和维持了一个繁荣的环境。

普林斯顿　新泽西　　Frederic M. Lord

1967 **年** 11 **月**　　Melvin R. Novick

乘第二次印刷的机会，我们对本教材作一些小的改动。这些改动包括：印刷错误及其他错误，重新界定某些关键的句子以减少曲解的可能性，并增加一些重要而有用的参考资料。

普林斯顿　新泽西　　Frederic M. Lord

衣阿华市　衣阿华州　　Melvin R. Novick

导　言

本书的主要目的是在心理测验数据的解释方面，提高读者的技能、经验和直觉(intuition)，并在编制与使用心理测验上既作心理理论的工具，又作选择、评价和指导实际问题的工具。为了实现这一目的，我们试图使读者接触某些心理测验分数的富有意义的统计理论。

虽然本书是根据测验分数理论和模型来编写的，但所研究的每一模型的实际应用及限制都列为实质重点。而且我们尽量以非专门性的方式提出这些讨论。由于本书选编了许多测验理论模型和公式，故它可作参考手册之用。另外，就某些高级专家而言，除迄今可得到的研究外，本书旨在为进一步的理论研究提供了更严密的基础。

本书的目标之一是，提出一组所选统计模型的假定陈述和结论推导。著者认为这些统计模型在测验编制与运用的实践中是有用的入门书。除极少例外，对书中出现的每个主要结果我们都给出完整的证明。许多情况下，这些证明比原来的证明更简单、更完整和更明确。当省略证明或部分证明时，我们一般都提供一本相应的参考文献。仅当一般的证明方法已被证实时，我们才留下一些证明给读者作练习。当一般证明无助于另外的见解而实质上只是更复杂的数学问题时，我们就只证明一般定理中的特殊情形。

我们试图按某些归类(grouping)的次序提出所选的测验理论

模型，这些归类是由构成模型的假定性质所决定的。这种次序导致了许多测验分数理论的解释。我们指出这些理论的相互关系，并更进一步把它们结合成一种潜在特质的总理论。这种结合构成了本书的基础。

本书从事的一项重要任务是，对具有语义(semantic)(如现实世界)意义的概念提供明确的句法(syntactic)(如数学的)定义，其中最主要的是真分数的概念。这个概念以前通常在句法上定义，而在语义上解释为：某人在无限长的测验中获得的观测分数。在本书中真分数则按句法定义为所期望的观测分数。于是用大数定律说明了，以前的测验理论定义是一种有效的语义解释。著者感到，对基本测验理论概念如真分数和误差，当按严格的句法定义和语义定义来解释与分辨时，许多令人厌烦的测验理论的争论就消失了。

由于我们的主要兴趣是测验理论的科学应用而不是技术应用，所以本书使用的统计方法主要属于有时称为信息推断(informative inference)范围之列。总估计和置信区间的某些经典方法作为信息推断方法而被广泛使用。Bayesian 学派懂得，这些估计常常是对导自 Bayesian 分析相似估计的良好近似值。

我们并没有对测验理论问题提出决策理论方法的全面评述。尽管决策理论模型目前在概念上是极其有用的，而且我们希望在测验课题的应用上最终是有价值的，但是我们并不认为：现有的模型通常足够灵活或逼真到可以立即正式应用。我们也认为：那些被流行决策理论公式所鼓励的学科专家，不明智地使用过度正式的模型和放弃决策责任，都应予阻止，尤其在科学关系方面应予阻止。我们赞同 Cronbach 和 Gleser(1965)的看法，“对发展和应用测验而言，流行的决策理论作为一种观点比作为一种正式数学方法源更为重要。”有兴趣的读者可参阅 Cronbach 和 Gleser 著的《心理测验与人事决策》以及 Herbert Soloman(1961)编的《题目分析与预

测之研究》。另一方面，我们看到 Birnbaum 的第五部分著作，正是试图填补在推断公式与决策理论公式之间的空白。此外，我们还在多处表明，从决策理论观点来看，某些标准的推断方法会更好理解。

在教育测验中妨碍决策理论的正式应用的一个重要问题是，缺乏一个可作分析依据的简单可测标准。例如，如果我们为医科学校选拔学生，我们要求选拔那些经过正规训练后可以成为“最好医生”的学生。但遗憾的是，正如我们没有“构成一个好教师”的明确测度(measure)一样，我们也没有“构成一个好医生”的明确测度。我们常用的标准是以可用性作为主要效能(virtue)。这个标准问题贯穿教育测验的所有预测问题，此可参阅 Kelly(1964)的书。

应用决策理论的另一个问题是，难于对所选标准的各种水平给出价值测度(value measures)。即使我们假定一个医科学校毕业生的未来价值(future value)，是与他在全国医科考生考试委员会的分数单调相关，我们也难以证明这种关系是线性的、或二次的、或递减边际效用的(decreasing marginal utility)。而统计决策理论的制定却要求我们相当明确地规定这种关系的性质。

至少还有另一个严重的问题是，限制某些已经发展的统计方法的应用性。除少数很专门的测验外，测验均可用于不同条件的广大范围之中。没有一个测验发行者能够在每一可能应用的条件下来验证其每个测验。因此，对测验理论的纯预测方法并没有得到完全满意的证明。

测验发行者和心理学家都承认，正式决策理论和预测方法都有其固有的困难。因此，他们对那些构成最早心理测验基础的测验发展方法，都持保留、扩展和改进的态度。这种方法是以包括学术倾向和能力等概念的理论体系为基础的；这种方法被认为会影响某人在心理测验中的成绩及其在各种兴趣效标上的成绩。在现代

心理学理论中，我们要用多种学术倾向和能力来思维。发展测验技术这与用最小可能误差来测量学术倾向和能力的测验发展有关。潜在特质理论是要识别那些与人类行为有关的特性(如学术倾向、能力、兴趣及个性等)，并且通过测量这些特质来解释和预测人类行为。虽然本书的着重点是关于潜在特质的推断及其测量理论，但我们将评述那些已被证明与测验应用直接有关的预测理论。

由于本书主要涉及测验理论，因而它对专门的人事选择问题的应用范围是不完全的。Horst(1955)所写的"编制分辨预测成套测验的一种方法"专题文章和 Thorndike(1949)写的《人事选择》，可作为我们在这个方面简化讨论的补充。有关教育测量理论与实践问题的广泛讨论，可阅读美国教育教材委员会编写的《教育测量》(第二版由 Thorndike 编辑)。美国心理协会的刊物《教育与心理测验的标准和手册》对任何从事心理测验的人都是一本有价值的指导书。

在一定的意义上，本书是一本教材或训练书。的确，这里所介绍的内容比纯理论文章更富于评论性。在技术和概念上，本书前几章的要求比后几章稍低。另外，在读者面临抽象概念之前，一般都给读者掌握具体材料的机会。许多理论上的细微点留作练习，在练习里可以找到许多数值例子。尽管在测验理论方面本书基本上不表现为一本教科书，但它有可能成为各种测验理论课程的部分内容。在心理测验理论方面，对许多课程(如因素分析、潜在结构理论等)，它也可以作为一种补充教材。

缺乏正规数学训练或统计训练的读者，也可以在本书里找到许多有用的材料。因为在对每种材料作正式描述的同时，我们也提供一种非技术的(即使很简要的)解释。除我们指出的几节可以省略而不失连续性之外，第 1 至 6 章和 12 至 15 章只不过要求熟悉一点期望代数学。这几章与第 7 章一起，适用于一学期(半年)测验

理论课的基础课程。具备模型Ⅱ方差分析的预备知识的读者，可阅读第8至11章及第7章的较易部分，尽管这些章实质上是独立的。我们设想，研究其余各章的读者，要具有适度的微积分能力和熟悉数理统计的语言、基本结论及基本技能。除了偶然有些被标明属高层次并可省略而不失连续性的小节之外，每当引进一种高等数理统计概念(如最小充分性)时，我们都给出了完整的推导过程。

尽管我们强调，任何地方都不要求完成一系列数理统计专业课程，但到一定程度后，数理统计知识越强的读者，可望在本书中得到越多的东西，也越容易阅读本书的后半部分。熟悉 Freeman (1963)、Mood 和 Glaybill(1963)或 Hoel(1954)等人著作的那种数理统计水平的人，几乎都能看懂全书。有几章的参考书目引自 Lindgren(1962)、Kendall 和 Stuart(1958,1961)。虽然我们假定读者熟悉这些教材或相应的读物，但我们并不假定读者具有这种教材的专业能力。实际上，如果我们为专业统计理论研究人员写书，我们就不会给出如此详细的推导了。

这并不是说，任何具有这种数理统计基础知识的人都能容易读懂全书的所有内容。尽管我们并没有使用高等数学或高等统计方法，但书中有标号的许多章节，主要是针对专家们的，这些章节由于所涉及材料的复杂性而显得困难得多。

心理学方面的先决条件几乎没有限定。一门心理学体系课、一门科学方法论课、一门应用实验设计课和一门心理测验应用课，这些课程都是有用的但肯定并非都是必需的。对应用统计学的一般熟悉程度若能达到可阅读 Hays(1963)和 Winer(1962)著作水平时就很不错了。那些阅读本书但没听过相应课程的读者，倘若能具备一些有关测验实际问题的感性认识或常识，这也是有好处的。如果读者缺乏这样的见识，就要查阅 Cronbach(1960)的资料性著作《心理测验纲要》以及 Jackson 与 Messich(1967)编写的74篇论文集

《人性评价中的若干问题》。

符　号

本书所用的符号，绝大多数都符合“统计协会会长会议”1965年6月在《美国统计学家》上所推荐的符号，但在某些细小方面作过修改. 本书符号体系的要点如下：

1. 大写字母表示随机变量，但某些多元变量的问题例外.

2. 小写字母表示样本观测值或非随机变量.

3. 小写字母表示概率密度函数，如 $f(x)$；相应的大写字母表示累积概率密度函数，如 $F(x)$.

4. 除少数例外，希腊字母始终用以表示总体参数.

5. 在参数符号上面置一个脱字符(即 ^)，表示某总体参数及其值的样本估计量(估计值).

6. $\mathscr{E}(X)$ 或 $\mathscr{E}X$ 或 μ_X 表示随机变量 X 的平均数，必要时使用下标或括号，例如

$$\mu_1=\mu(X_1)=\mathscr{E}(X_1).$$

7. $\mathscr{E}(X^k)$ 或 $\mathscr{E}X^k$ 表示随机变量 X 的第 k 阶矩，也能用适当的括号或下标.

8. $\bar{x}$ 表示一数集 $x_1,\cdots,x_n$ 的算术平均数，即 $\bar{x}=\sum x_i/n$.

9. “点下标”表示样本平均数，如

$$x_{\cdot}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x},\qquad x_{g\cdot}=\frac{1}{N}\sum_{a=1}^{N}x_{ga},$$

$$x_{\cdot a}=\frac{1}{n}\sum_{g=1}^{n}x_{ga},\qquad x_{\cdot\cdot}=\frac{1}{nN}\sum_{g=1}^{n}\sum_{a=1}^{N}x_{ga}.$$

加号(+)下标,如 x_{g+},通常表示相应的和数.

10. $\sigma^2(X)$,$\sigma_X{}^2$ 或 $\mathrm{Var}(X)$表示随机变量 X 的方差.

11. $\sigma(X_1,X_2)$、σ_{12}或 $\mathrm{Cov}(X_1,X_2)$表示随机变量 X_1、X_2 的协方差. 若随机变量用 X、Y 表示,则用符号 σ_{XY}表示 X、Y 的协方差. 用 ρ 代 σ,则表示类似变量的相关系数.

12. 大写黑体字母,如 $\boldsymbol{A}$,表示矩阵,$\|a_{ij}\|$ 表示矩阵 $\boldsymbol{A}$ 中第 i、j 个元素 a_{ij}.

13. 小写黑体字母,如 $\boldsymbol{a}$,表示列向量;$\{a_i\}$表示列向量中第 i 个元素 a_i;用小写黑体字母加"撇上标"表示行向量,如 $\boldsymbol{a}'$.

14. Σ 表示随机变量 $X_1,\cdots,X_n$ 的方差矩阵和协方差矩阵,即

$$\Sigma=\|\sigma_{ij}\|.$$

15. $\rho_{0\cdot 123\cdots n}$表示某随机变量 X_0 与另外 n 个随机变量 $X_1,\cdots,X_n$ 之间的复相关系数.

16. 字母 β 及相应下标,表示线性回归函数的系数,而 α 表示该函数的常数. 于是,X_0 对 k 个独立变量 $x_1,\cdots,x_k$ 的线性回归可写成

$$\alpha+\beta_1x_1+\cdots+\beta_kx_k;$$

更确切地,可把第 i 个权数写成

$$\beta_{0i\cdot 12\cdots(i-1)(i+1)\cdots n}.$$

17. $\rho_{01\cdot 23\cdots n}$表示在考虑随机变量 $X_2,\cdots,X_n$ 联合分布情况下,随机变量 X_0 与 X_1 之间的偏相关系数.

18. 在 n 个数 $x_1,\cdots,x_n$ 的集合中,符号 s^2 定义为

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

19. 在 n 个数对$[(x_{11},x_{21}),\cdots,(x_{1n},x_{2n})]$的集合中,符号 s_{12}定