当代国外语言学与应用语言学文库

Statistical Analyses for Language Assessment

语言测评中的统计分析

Lyle F. Bachman
Antony J. Kunnan

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
剑桥大学出版社
CAMBRIDGE UNIVERSITY PRESS

当代国外语言学与应用语言学文库

Statistical Analyses for Language Assessment

语言测评中的统计分析

Lyle F. Bachman
Antony J. Kunnan

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

剑桥大学出版社
CAMBRIDGE UNIVERSITY PRESS

北京 BEIJING

京权图字：01-2014-0574

This is a reprint edition of the following title published by Cambridge University Press:

Statistical Analyses for Language Assessment (ISBN:9780521003285)
© Cambridge University Press 2004

This reprint edition for the People's Republic of China (excluding Hong Kong SAR, Macao SAR and Taiwan Province) is published by arrangement with the Press Syndicate of the University of Cambridge, Cambridge, United Kingdom.

© Foreign Language Teaching and Research Press and Cambridge University Press 2014

This reprint edition is authorized for sale in the People's Republic of China (excluding Hong Kong SAR, Macao SAR and Taiwan Province) only. Unauthorized export of this reprint edition is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of both Cambridge University Press and Foreign Language Teaching and Research Press.

本书版权由剑桥大学出版社和外语教学与研究出版社共同所有。本书任何部分之文字及图片，如未获得两社书面同意，不得用任何方式抄袭、节录或翻印。
只限中华人民共和国境内销售，不包括香港特别行政区、澳门特别行政区及台湾省。不得出口。

图书在版编目 (CIP) 数据

语言测评中的统计分析 = Statistical analyses for language assessment：英文 /（美）巴克曼 (Bachman, L.F.)，（美）昆南 (Kunnan, A.J.) 著. — 北京：外语教学与研究出版社，2013.12

（当代国外语言学与应用语言学文库）
ISBN 978-7-5135-3963-0

Ⅰ. ①语… Ⅱ. ①巴… ②昆… Ⅲ. ①语言－测试－统计分析－研究－英文 Ⅳ. ① H09

中国版本图书馆 CIP 数据核字 (2014) 第 007495 号

出 版 人　蔡剑峰
责任编辑　李　云　钱垂君
封面设计　刘昱莲
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　北京京科印刷有限公司
开　　本　650×980　1/16
印　　张　35.25
版　　次　2014 年 10 月第 1 版　2014 年 10 月第 1 次印刷
书　　号　ISBN 978-7-5135-3963-0
定　　价　74.00 元（含 CD-ROM 光盘一张）

购书咨询：（010）88819929　电子邮箱：club@fltrp.com
外研书店：http://www.fltrpstore.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所　刘旭东律师
　　中咨律师事务所　殷　斌律师
物料号：239630001

藏書 图書馆

当代国外语言学与应用语言学文库

专家委员会

（按姓氏笔画排列）

丁言仁　马秋武　王宗炎　王才仁　王立弟　王克非
王初明　王逢鑫　王嘉龄　王继辉　文秋芳　文　旭
方　立　冯　蒸　冯志伟　史宝辉　宁春岩　田贵森
申　丹　石定栩　冉永平　刘润清　刘世生　刘丹青
朱永生　江　怡　吴一安　沈家煊　陆俭明　陈国华
严辰松　何兆熊　何安平　何自然　张绍杰　张柏然
张德禄　李　兵　李宇明　李延福　李行德　李福印
李筱菊　杜学增　汪榕培　纳　日　邵永真　陈治安
陈新仁　罗选民　杨永林　杨信彰　杨惠中　周流溪
周　燕　林连书　金利民　胡文仲　胡壮麟　姚小平
赵忠德　封宗信　祝畹瑾　姜望琪　桂诗春　顾曰国
徐烈炯　徐盛桓　徐大明　徐赳赳　涂纪亮　秦秀白
贾玉新　钱　军　顾　阳　高　远　高一虹　黄国文
惠　宇　程　工　程晓堂　董燕萍　蒋祖康　韩宝成
蓝　纯　熊学亮　潘永樑　戴炜栋　戴曼纯

当代国外语言学与应用语言学文库 第三辑*

Psychology of Language *(Fifth Edition)* / D.W. Carroll
《语言心理学》（第五版）

A Course in Phonetics *(Fifth Edition)* / P. Ladefoged
《语音学教程》（第五版）

Linguistics: An Introduction to Language and Communication *(Fifth Edition)* / A. Akmajian, R. A. Demers, A. K. Farmer & R. M. Harnish
《语言学：语言与交际导论》（第五版）

The Minimalist Program / N. Chomsky
《乔姆斯基的最简方案》

Speaking: From Intention to Articulation / W. J. M. Levelt
《说话的认知心理过程》

Introducing Second Language Acquisition / M. Saville-Troike
《二语习得引论》

Minimalist Syntax: Exploring the Structure of English / A. Radford
《最简句法入门：探究英语的结构》

Analyzing Discourse: A Manual of Basic Concepts / R. A. Dooley & S. H. Levinsohn
《话语分析中的基本概念》

Curriculum Development in Language Teaching / J. C. Richards
《语言教学中的课程设计》

Fossilization in Adult Second Language Acquisition / ZhaoHong Han
《成人二语习得中的僵化现象》

A Student's Introduction to English Grammar / R. Huddleston & G. K. Pullum
《剑桥学生英语语法》

Introducing Phonology / D. Odden
《音系学导论》

Introducing Functional Grammar *(Second Edition)* / G. Thompson
《功能语法入门》（第二版）

* 本文库采用开放式结构，今后还将陆续出版其他有影响的语言学著作

An Introduction to Functional Grammar *(Third Edition)* / M. A. K. Halliday & C. Matthiessen
《功能语法导论》（第三版）

An Introduction to Cognitive Linguistics *(Second Edition)* / F. Ungerer & H.-J. Schmid
《认知语言学入门》（第二版）

Typology and Universals *(Second Edition)* / W. Croft
《语言类型学与普通语法特征》（第二版）

English Phonetics and Phonology: A Practical Course *(Third Edition)* / P. Roach
《英语语音学与音系学实用教程》（第三版）

Approaches and Methods in Language Teaching *(Second Edition)* / J. C. Richards & T. S. Rodgers
《语言教学的流派》（第二版）

Understanding Phonology *(Second Edition)* / C. Gussenhoven & H. Jacobs
《音系学通解》（第二版）

Metadiscourse / K. Hyland
《元话语》

The Language of Evaluation: Appraisal in English / J. R. Martin & P. R. R. White
《评估语言：英语评价系统》

Language in Literature: An Introduction to Stylistics / M. Toolan
《文学中的语言：文体学导论》

Pragmatics / Yan Huang
《语用学》

The Oxford Handbook of Computational Linguistics / R. Mitkov (ed)
《牛津计算语言学手册》

Intercultural Communication in Contexts / J. N. Martin & T. K. Nakayama
《社会、历史背景下的跨文化交际》

Handbook of Bilingualism: Psycholinguistic Approaches / J. F. Kroll & A. M. de Groot

《双语认知的心理语言学研究》

English: Meaning and Culture / A. Wierzbicka

《英语：意义和文化》

Foundations of Language: Brain, Meaning, Grammar, Evolution / R. Jackendoff

《语言的基础：大脑、意义、语法和演变》

Meaning in Interaction: An Introduction to Pragmatics / J. Thomas

《言谈互动中的意义：语用学引论》

Sociolinguistics: The Study of Speakers' Choices / F. Coulmas

《社会语言学：说话者如何作出选择》

Dialogic Inquiry: Toward a Sociocultural Practice and Theory of Education / G. Wells

《在对话中学习：社会文化理论下的课堂实践》

Handbook for Writing Research Papers, Reports, and Theses / C. Slade & R. Perrin

《如何写研究论文与学术报告》

Language and Society *(Second Edition)* / W. Downes

《语言与社会》（第二版）

Constructing a Language: A Usage-Based Theory of Language Acquisition / M. Tomasello

《如何建构语言：基于使用的语言习得理论》

Second Language Needs Analysis / M. H. Long

《第二语言需求分析》

Cognitive Linguistics and Language Teaching / R. Holme

《认知语言学和语言教学》

Intercultural Interaction: A Multidisciplinary Approach to Intercultural Communication / H. Spencer-Oatey & P. Franklin

《跨文化互动：跨文化交际的多学科研究》

Tasks in Second Language Learning / V. Samuda & M. Bygate
《第二语言学习中的任务》

English: One Tongue, Many Voices / J. Svartvik & G. Leech
《英语的变迁：一种语言，多种声音》

Language Testing and Validation: An Evidence-Based Approach / C. J. Weir
《语言测试与效度验证：基于证据的研究方法》

Second Language Learning and Language Teaching *(Fourth Edition)* / V. Cook
《第二语言学习与教学》（第四版）

The Oxford History of English / L. Mugglestone
《牛津英语语言史》

Toward a Cognitive Semantics (Volume I): Concept Structuring Systems / L. Talmy
《认知语义学（卷 I）：概念结构系统》

Toward a Cognitive Semantics (Volume II): Typology and Process in Concept Structuring / L. Talmy
《认知语义学（卷 II）：概念结构中的类型及过程》

Statistical Analyses for Language Assessment / L. F. Bachman & A. J. Kunnan
《语言测评中的统计分析》

Research Perspectives on English for Academic Purposes / J. Flowerdew & M. Peacock
《学术英语的多维研究视角》

Meaning in Language: An Introduction to Semantics and Pragmatics *(Third Edition)* / A. Cruse
《语言的意义：语义学与语用学导论》（第三版）

The Grammar of Words: An Introduction to Linguistic Morphology / G. Booij
《词的语法：形态学导论》

Genre Relations: Mapping Culture / J. R. Martin & D. Rose
《语类关系与文化映射》

Linguistics: An Introduction *(Second Edition)* / A. Radford, M. Atkinson, D. Britain, H. Clahsen & A. Spencer
《语言学教程》（第二版）

导读

⊙ 邵永真

一、《语言测评中的统计分析》概览

《语言测评中的统计分析》(*Statistical Analyses for Language Assessment*)是剑桥语言评估系列丛书之一，是继剑桥大学出版社出版的《专门用途英语测评》(*Assessing Languages for Specific Purposes*)、《词汇测评》(*Assessing Vocabulary*)、《阅读测评》(*Assessing Reading*)、《写作测评》(*Assessing Writing*)、《听力测评》(*Assessing Listening*)、《口语测评》(*Assessing Speaking*)以及《语法测评》(*Assessing Grammar*)等专著之后一部重要的理论综述参考书。

语言测试与语言教学是不可分离的。但有的语言教师却感到语言测试深奥难懂，遇到统计分析中的一些概念、符号、公式就会敬而远之，甚至把语言测试学看成一门与语言教学格格不入的学科。

的确，英语测试已经发展成为应用语言学的一个独立的学科分支。但是，它是与语言教学理论紧密联系的。按照剑桥语言评估系列丛书的观点，语言测试的核心还是语言本身。它测评的是语言的构想成分。如何在严格的理论基础上建立起一个语言构想成分测试和评估的框架，是应用语言学家所关注的问题。语言测试的首要手段就是测量，就是要进行量化，这就离不开统计分析。这就对从事语言测试的语言教师提出了更高的要求：除了要懂得语言教学本身之外，还要懂得统计，特别是要熟悉用于语言测试分析的统计理论和方法，并运用这些理论和方法来解决自己遇到的实际问题。简而言之，学习有关的统计理论和方法，学用结合，是掌握语言测试的统计分析最好的途径。

《语言测评中的统计分析》的作者是国际著名语言测试专家 Lyle F. Bachman。他在一系列的论著中，对语言测试的基本理论作了系统的阐述。尤其是他与 Adrian Palmer 合著的《语言测试实践》(*Language Testing in Practice*)一书提出语言测试的效用性 (usefulness) 这个极其重要的概念。衡量语言测试是否具有较好的效

用性，就要综合考虑测试的信度、构想效度、真实性、交互性、影响以及可实践性这6个质量因素。而这6个因素的验证，离不开受试者在语言测试中的发挥（以考试分数的形式表现出来）。所以，语言测试的设计者必须采用合适的方法来解释考试的结果，说明这些考试分数是否真实反映了受试者的语言运用能力。对于语言教师来说，这一点也十分必要。因此，不仅是语言测试的设计和开发者，第一线的语言教师也有必要读一读这本专著，掌握书中所介绍的统计分析手段和工具，从而验证所设计或使用的语言测试是否有效，是否存在什么不足。与一般的统计学不同，本书侧重实用，没有深奥的理论和数学公式推导。作者在介绍有关的统计方法时总是先介绍用该方法能提供什么信息；如何使用这些信息；然后进一步说明在什么条件下可以应用该统计方法；以及在这种条件下，还可以采用的其他方法；接下来再对该统计方法的要领进行逐步介绍。

值得一提的是，本书在附录中特别设计了一百余页的研习专页，内容主要是有关正文中概念的练习及基于一些小型数据组的运算练习，还收录了实际语言测评中的真实数据，并分步骤详尽地演示如何使用SPSS进行运算。

所附CD共包含两部分的内容：第一部分是研习专页中练习的答案；第二部分是研习专页练习中部分数据组样本的电子版以及对每组数据的全面描述。每组数据举了几种不同的格式，以使读者能在各种不同的应用软件中使用。这些格式包括：在Windows 系统中可用SPSS格式（.sav），用制表符分隔的文本文件格式（.dat）或纯文本文件格式（.txt）以及EXCEL文件格式（.xls）。

全书由3部分组成：第1部分介绍语言测评的基本概念及统计方法；第2部分介绍分析和改进测试的统计方法；第3部分介绍测试应用中的统计方法。

二、《语言测评中的统计分析》

第1部分

第1部分由第1章至第3章组成，介绍了语言测评的基本概念及统计方法。

第1章　基本概念和术语

本章首先介绍了测试的效用（test usefulness）或有用性。它包括了6个方面的特性：信度、构想效度、真实性、交互性、影响和可实践性。接着，作者对几个有关“测试”的常用术语进行了解释，简要地说明测评（assessment）、测量（measurement）、测验（test）和评估（evaluation）之间的区别和联系。与之有关的是测评之后的决策：相对性决策和绝对性决策。例如，大学录取新生就是一个

相对性的决策，是从考生群中择优录取的；而一些证书考试，例如驾驶证、律师证书、医师证书考试，均要求达到设定的标准，是绝对性的决策。

通过语言测试，我们得到了考试成绩和丰富的数据。但如何用这些数据正确地解释我们所要测量的但又不易直接观察到的语言能力呢？我们在测评过程中必须遵循3个步骤：(1) 对语言能力的组成进行定义；(2) 根据语言能力的组成设计出具体的测量步骤和方法；(3) 把观察到的结果进行量化，其量化方式可以是数值、等级或比率。

第2章　考试成绩的统计分析和描述

经过语言测试得到的数据要进行统计分析。描述性统计可以为受试的某一个群体或某一个样本的测验成绩进行总结和描述；而推理性统计对更大的群体或受试总体的语言能力进行合理的推理和判断。在解释统计分析时一般有两种框架：常模参照性和尺度参照性。

在描述考试成绩的量化特征时，首先就是观察某一系列成绩的分布。这可以通过图示，也可以采用描述统计方法。用图示方法可以看出考试分数是否对称，分布曲线是平坦的还是呈尖峰形，是正偏态的还是负偏态的，是否呈现双众数形，等等。用描述统计方法可以准确地说明考试成绩的集中量和离散量。平均数、中数和众数是3个常用到的集中量。常用的离散量有全距、半四分距（或称四分差）和标准差。通过集中量和离散量等描述统计方法，可以进行成绩的报告和解释，估计考试的信度和效度。

第3章　不同系列的考试成绩之间的关系

相关系数是社会科学，特别是语言测试中许多领域均要涉及的概念，是测量信度和效度研究的重要环节。最常用的相关系数计算方法是皮尔逊积距相关和斯皮尔曼等级相关。这两种相关系数的涵义是相同的，但应用前提有所不同。对于两个线性的正态分布的连续变量，可以采用皮尔逊积距相关来计算它们的相关系数。而应用斯皮尔曼等级相关时只需知道两个线性变量的等级顺序。书中用同一组变量比较了这两种不同的计算方法，并通过众数、中位数、平均数、偏态值、峰值、测量误差等参数分析产生差异的原因。强调采用某种计算法必须注意其前提条件，否则就会影响计算的准确性。

相关系数可以用来估计考试题目的区分度，进而推算试卷的信度和构想效度，因而是十分重要的统计概念。

除了皮尔逊和斯皮尔曼这两种相关系数计算法之外，第3章还简单地介绍了

此为试读，需要完整PDF请访问：www.ertongbook.com

一些比较复杂的相关系数估算方法，例如多重线性回归、通径分析、因子分析和结构方程模型等。

第2部分

第2部分由第4章至第6章组成，介绍了分析和改进测试的统计方法。

我们设计专门的考试时，除了严格控制考试内容之外，还希望考试成绩能达到预期的结果。例如，在常模参照性考试中，成绩分布最好拉得开一些；在尺度参照性考试中，成绩分布就要集中一些，标准差小一些。另外，我们还要求考试的信度尽量大。因此，严格地说，一个重要的试卷设计过程，应该包含预测这一步骤，然后通过对具体的单个题目的统计特征的观察，筛选出合适的题目来组卷。

第4章　考试的分析

第4章着重讨论了项目分析和题目筛选的方法。这一章介绍了经典的项目分析（包括难度和区分度）计算方法。项目分析可以帮助考生了解本人对具体考题的应答发挥情况；可以帮助教师和教材编写者改进教学和教材；还可以帮助考试命题人改进试题质量，例如控制试题的难易度、区分度、信度，发现命题的缺陷。

本书与其他的语言测试学教科书有几点不同：(1) 一般的语言测试教科书只介绍常模参照性考试的项目分析，而本书还讨论尺度参照性考试的经典项目分析的计算方法。例如，在常模参照性考试中，题项的区分度可以通过对高分组和低分组中答对此题的人数之差异进行计算而求得。但在尺度参照性考试中，题项的区分度是要把考生分为达到预期尺度标准和没有达到标准的两群人，然后观察这两群考生中答对此题的人数之差异，以得出该项题目的区分度。书中还提出通过教学前后考生对同一试题应答的变化来计算尺度参照性考试试题区分度的方法。(2) 一些语言测试教科书仅仅讨论0－1赋分的题目的题项分析，但本书还介绍了连续性赋分的题目的题项分析的计算方法。与此相关的是如何利用点双列相关和双列相关分析来确定题项的区分度。毋庸置疑，连续性赋分的题目，应该用双列相关分析来确定其区分度。对于0－1赋分的题目，书中提出，要作具体分析。性别区分（男0－女1）或是否某一社团成员的区分（成员0－非成员1）是典型的0－1系列。某些题目虽然是0－1赋分的，但不同于性别区分或是否某一社团成员的区分。这些题目所测试的潜在能力是连续变量。出于这种考虑，采用双列分析是恰当的。(3) 有了题项分析的数据，筛选题目就有据可依了。例如，所选用的题目的区分度应在0.3以上，平均难度值可以放在0.5，适当地选用一些难度值在0.3至0.7（甚至于0.2至0.8）

之间的题目。但是，本书特别指出，设计尺度参照性考试时，应该使成绩分布呈负偏态。成绩的众数应该接近合格分数线。

本章还分析了经典测试理论统计分析的局限性，并简要地介绍了项目反应理论。

经典的项目统计分析严重依赖于被试样本。项目难度以通过率表示，因此被试样本能力高时项目通过率就高，就被看作容易，反之就被看作难度大；区分度通常以项目与总分的相关或高低能力组的通过率之差表示，两组能力差别大时，区分度就高，反之则低；对被试能力的估计依赖于测试题目的难度。被试能力与测试题目的难度是相关的，参加不同难度的测试会得到不同的能力估计值，不同测试结果难以进行比较。另外，诸如口语考试之类的考试中，考生的成绩还受到考官的主观因素的影响。为了克服这些局限性，教育测量专家提出了项目反应理论（IRT）、多层面Rasch模型分析和概化理论。

项目反应理论具有以下优点：

(1) 难度和区分度的估计值与被试能力无关。

(2) 对被试能力的估计不依赖于特定的测试题目。

(3) 测试信息函数的概念代替了信度理论，用测试对能力估计所提供的信息量的多少来表示测量的精度，能给出不同能力被试者的测量精度。

多层面Rasch模型可以分析考生和评分人员之间、考生和试题之间以及评分人员和试题之间的相互作用，了解阅卷老师、考生、试题和评分标准等层面的表现。从阅卷老师层面反馈的信息可以帮助了解阅卷老师阅卷的态度以及对评分标准的掌握情况，从而判断是否过于严厉或者过于宽松；是否始终一致地准确掌握评分标准，不会因为考题不同或考生不同而发生偏离。从考生和试题层面反馈的信息则可以发现有没有非拟合考生或非拟合试题。

第5章　常模参照性考试的信度

第5章讨论常模参照性考试的信度的估算方法。

测试理论在估算信度时提出了“真分数理论”。真分数指的是没有测量误差时的理论上的观察值。但实际的观察值总是带有种种误差，因此，实际的观察值等于真分数加上误差。这些误差主要源于：(1) 测验方法和过程（包括评分者），属系统性因素；(2) 随机因素。

经典测试理论（CTT）估算信度，主要是针对随机因素的。其信度估算方法主要有以下几种：

(1) 内部一致性信度估算：具体有斯皮尔曼—布朗信度公式、哥德曼折半信度计算法、克龙巴赫α系数估算法、库德—理查森信度公式20和公式21等。采用内部一致性信度估算方法的前提是两组分数相互独立，而且彼此平行（注意：哥德曼折半信度计算法不要求平行这个条件）。如果不是相互独立的，那么估算信度往往会过高；如果两组分数不是平行题得出的，那么估算信度往往会过低（当然，哥德曼折半信度计算法不受影响）。

(2) 稳定信度（再测信度）：这是一种长期的信度。也就是指标在不同时间测试时，得到的结果是否相同。通常我们是用测试与再测试方法来检验试卷的稳定信度，也就是用同样的试卷在同一群体中重新施测，如果每次都得到同样的结果，此指标即有稳定信度。在实施时，把握两次测试的间隔时间很关键。相隔时间过长，则考生能力会发生不同程度的变化；相隔时间太短，则考生各自会对不同的题目产生不同的熟悉程度。

(3) 同等信度（平行测验信度）：请同一批受试者做两份相似的问卷，来估计平行测验信度；题目通常可互换，这表明题目同构型很高，亦能显示题目内在一致性的信度。但如果两次测验时间相隔两周以上，相关系数可能降低，这种信度反映出施测场合不同而造成的差异。因此，我们还可以采用两组能力相同的受试者交叉做两份平行的试卷，来平衡时间带来的误差。

(4) 评分员一致性信度：同一位评分员是否始终能保持同一个评分标准，例如作文评分中，会不会时而严格时而宽松，我们需要估算评分员的内信度（intra-rater reliability）。不同评分员对同一位考生的应答评分也可能不一致，例如在口语考试中，往往是几位考官同时给一位考生打分。这时，我们需要估算评分员间信度（inter-rater reliability）。

在信度估算的基础上，我们可以推算一次考试的标准误。标准误的大小与样本标准差成正比。同时，我们常用“置信区间估计”的方法，来估计观察成绩可能的偏移范围和偏移概率。即：X±1.00 SEM 表示成绩的68%置信区间；X±1.96 SEM 表示成绩的95%置信区间；X±2.58 SEM 表示成绩的99%置信区间。

由于经典测试理论的估算信度方法存在着一些缺陷，心理测试学家提出了概化理论和项目反应理论。

经典测试理论的估算信度主要的缺陷是：(1) 测试的误差来源很多，而且交互起作用，经典测试理论一次只能针对一种测量误差；(2) 只能笼统分析随机性误差，没有把系统性误差从随机性误差中区分出来；(3) 使用经典测试理论的标准误估计成绩对所有的考生都是一样的，不能按能力的高低来区别对待。另外，它估算

信度的基础假设是平行试题，而实际上，两份试题很难做到真正的平行。

针对经典测试理论的缺陷，出现了概化理论和项目反应理论。

概化理论为检查语言测量提供了灵活的、可行的理论框架。概化理论在测量多个误差来源维度上对经典测量理论进行了扩展，为常模参照性考试的信度分析提供了新的框架。该模型以因子设计和方差分析为基础，把某一个分数看成一个假设总体的各种可能的分数的一个样本。它的理论构架包括概化研究，即G研究（generalizability study）和决断研究，即D研究（decision study）两个阶段。

概化理论的优势是：(1) 测量的多种误差来源可以放在同一个分析中分别估计；(2) 可以指导决策者选择最优测量方案；(3) 提供可靠性系数，用于不同的决策任务；(4) 排除了严格平行测验的假设。例如，在考虑诸多的变异因素时，概化理论考虑到受试者、试题、评分员、试题和受试者的相互关系，受试者和评分员的相互关系，评分员和试题的相互关系，受试者和评分员以及试题的相互关系。本书没有提供具体运算的步骤，但提供了参考的网站。

在项目反应理论的框架中，一个考生答对一个考试题目的期望行为是题目难度水平和考生本人的能力水平两者的函数。在信度估算方面，项目反应理论提出了信息函数的概念，一个考试题目的信息函数指的是一个考试题目所提供的估计一个人的能力水平的信息，一次测试的信息函数指的是所有考试题目的信息函数之和，而测量误差是根据信息函数来估算的。因此，每一个考试题目都可以有自己的测量误差，每一个能力水平也可以有自己的测量误差。

第6章　尺度参照性考试的信度

尺度参照性考试的目的在于检测考生对指定内容的掌握情况，分数线的划分是按既定的标准进行的。这种考试是否可信并不依赖于考生成绩的分布和相互差异。因而，沿用常模参照性考试的信度估算方法不可能为其可靠性提供信息。

为区别起见，在讨论尺度参照性考试的信度时，书中用可靠度（dependability）、使用一致性（agreement）等术语。

尺度参照性考试的可靠度可以用phi系数的公式来计算。同时，还可以用标准误差和置信区间检验其可靠度。

这一章还介绍了临界缺损一致性（threshold loss agreement）分析法、平方误差缺损一致性（squared-error loss agreement）分析法和领域分数可靠度（domain score dependability）分析法。

第3部分

第3部分由第7章至第10章组成，介绍了使用测试的统计方法。

第7章　假设和推断

第7章首先介绍了统计显著性的检验。在分析考试成绩时，我们经常需要比较两组不同学生成绩的差异，以便了解他们能力是否相同。有时我们需要对同一批学生安排两次不同的考试，看看他们自身是否在不同的方面存在能力差异。为此，我们往往需要比较两组成绩的平均值的差异，或估算两组成绩的相关值。这时，我们需要对这些差异或相关进行统计推断，来确定它们是确确实实如此，还是偶然因素使然。这个统计推理过程包括提出假设、收集样本数据以及用数据验证假设等几个步骤。

通过抽取样本来收集数据时，结果是随机的。关键的一点是：从总体中抽取的全部样本所构成的抽样分布平均值应该等于总体的平均值；随着样本容量的增加，抽样平均值的分布趋近于正态分布。

假设是统计研究中广泛使用的方法。根据已知的理论，我们先提出理论假设，然后根据已知的事实或参数提出操作上的假设，最后提出一项预想的希望证实的假设。这种假设就是统计术语中的研究假设，用H1表示。由于H1的真实性不能直接检验，需要建立与之对立的假设，先假设要比较的数据并无实质性差异，称为虚无假设或零假设，用H0表示。H1和H0两者对立，两者择一，所以H1有时又称为备择假设。假设检验，就是判断虚无假设H0是否正确，决定接受还是拒绝虚无假设。如果有充分理由拒绝虚无假设，则接受备择假设。反之，若虚无假设成立，则备择假设就不成立。

第8章　统计分析中显著性的检验

估算抽样统计中的平均值的标准偏差是估算总体平均值的基础。因为许多场合下，总体平均值是未知的，而样本的平均值与总体平均值有一定的离差。但在理论上可以认为，从总体中抽取的全部样本所构成的抽样分布的平均数应该等于总体的平均数。人们就考虑用实际上所有可能抽到的样本的实际抽样平均值的标准差代表抽样平均误差。这个抽样平均误差是度量样本平均值在总体平均值周围分散程度的一个指标。它的计算公式是：

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

注意公式中s为样本平均值的标准差。

从这一公式可以看出，影响抽样平均值标准误差的主要因素有两个，即总体标准差的无偏估计量和样本量的大小。也就是说，样本量越大，总体标准差的无偏估计量越小，样本平均值的标准差也越小。

当样本量比较大的时候，可以利用这个标准差建立总体平均值的置信区间。当样本量比较小（不足30）的时候，对总体参数的估计就要利用t分布。进行t检验时需要考虑一些应用条件。例如，有一些教科书把正态分布作为t检验的条件之一。事实上，双侧检验时并没有必要强调正态分布。另外，在两个总体参数进行比较时，是否要求方差齐性。作者介绍了方差齐性检验方法，即先进行F率的检验，来确定两个总体方差的差异是否显著，然后决定是否继续进行t检验。

在实际研究中，往往需要对两组以上的平均数进行比较。如果仍然用每对平均数差异检验，就会降低效率，增加出错概率。此时应该用方差分析（ANOVA）。进行方差分析时同样也要考虑应用条件，例如是否强调总体参数的正态分布。作者指出，除非样本容量太小，而总体参数偏态性太明显，一般情况下仍可以进行方差分析。其次，与t检验类似，需要检验各受试组内部的方差彼此有无明显的差异。着手进行方差分析之前可以通过F值来检验多个方差的齐性。另外，作者强调了方差分析中观察的独立性。违反了观察的独立性会影响分析的准确性，易犯I类错误。

这一章对方差分析的具体过程作了举例说明，还介绍了相关系数的显著性检验和非参数检验等基本概念。

第9章　效度研究

效度即有效性，指的是测试在多大程度上测出了它预期测量的东西。使用工具进行测量时，我们关心的不仅是测量工具是否准确，而且要看测量工具用来测量某种属性是否有效。效度实际上指的是我们通过测试所获得的证据在多大程度上支持我们根据分数所做出的判断。

在考试中，我们要利用考试成绩作为依据，进行推理分析，做出合理的决策。考试有不同的目的，这些目的是否能达到要看考试效度的高低。考试的效度高，则考试能实现原定的测量目的；如果效度低，不仅无法达成目标，甚至会提供不正确的数据而导致错误的决定。

作者在介绍效度的概念时提出：(1) 效度是考试结果的解释和使用是否适当有效，而不是考试本身或成绩是否适当有效；(2) 效度是程度上的差别，而不是完全有效或根本无效的问题；(3) 效度分析是有针对性的，应依据使用该考试的预定

此为试读，需要完整PDF请访问：www.ertongbook.com