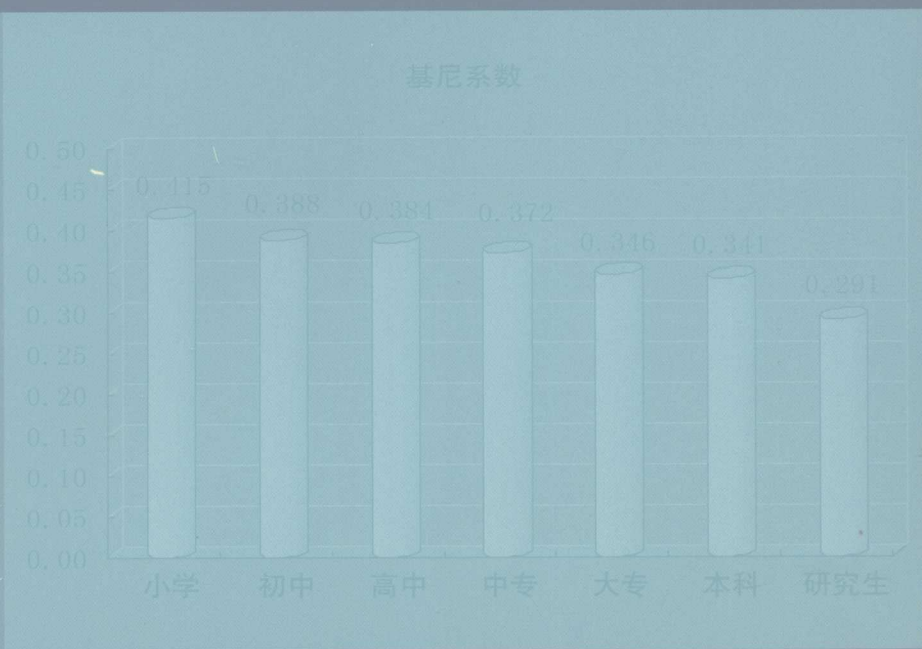
JIAOYU JILIANGXUE

教育计量学

岳昌君◎著



北京大学出版社
PEKING UNIVERSITY PRESS

教育计量学

岳昌君 著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

教育计量学/岳昌君著. —北京：北京大学出版社，2009.3
ISBN 978-7-301-14772-6

Ⅰ.教… Ⅱ.岳… Ⅲ.教育经济学—研究 Ⅳ.G40-054

中国版本图书馆 CIP 数据核字(2008)第 194244 号

书　　名：教育计量学
著作责任者：岳昌君　著
责 任 编 辑：刘　维
标 准 书 号：ISBN 978-7-301-14772-6/G·2559
出 版 发 行：北京大学出版社
地　　址：北京市海淀区成府路 205 号　100871
网　　站：http://www.jycb.org　http://www.pup.cn
电 子 信 箱：zyl@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62767346　出版部 62754962
印 刷 者：河北滦县鑫华书刊印刷厂
经 销 者：新华书店
730 毫米×980 毫米　16 开本　13.25 印张　220 千字
2009 年 3 月第 1 版　2009 年 3 月第 1 次印刷
定　　价：26.00 元

前　言

改革开放以来，我国经济、社会、教育、文化等各个方面都取得了巨大进步。同时，我国也出现了一些新的问题和矛盾亟待解决。在教育研究领域，学者们对我国教育发展中出现的各种问题进行了热烈的讨论。在研究中，人们越来越重视研究过程和方法的规范化，越来越注重对社会现实的深入调查和分析。在各种研究方法中，定量研究方法逐渐受到重视并得到广泛的应用。在学术期刊和研究生学位论文中，含有实证研究的文章和论文所占的比重越来越大。

多年来，北京大学教育学院一贯注重对研究生进行研究方法方面的训练。我从2003年春季开始，每年都为硕士和博士研究生开设教育统计与SPSS应用和高级定量研究方法课程，在指导研究生论文写作过程中和自己在课题研究过程中也经常使用定量研究方法。本书就是根据本人多年来的教学体会和学术研究成果编辑撰写而成的。之所以取名为《教育计量学》，是参考了其他学科中诸如《经济计量学》和《生物计量学》等课程名称。所谓教育计量学，是以一定的教育理论和实际统计资料为基础，运用数学、统计学方法与计算机技术，以建立教育计量模型为主要手段，定量分析研究具有随机性特征的教育变量关系。

本书具有以下特色：

一、包括许多实例。本书不仅包括本人近年来发表的学术文章、北京大学教育学院的研究生学位论文还包括国内外期刊和《国际教育经济手册》上的一些内容。在这些具体的实例中，涉及个人教育收益率、教育生产函数、公共教育投资比例的国际比较、教育政策评价、入学机会和教育经费的公平性研究、课外补习现象研究、高等教育需求分析、高等教育学生学业成就的影响因素分析、高校毕业生就业满意度研究、劳动力行业分布的影响因素分析等。

二、内容涉及面广。(1) 本书全面收集了近年来在教育研究中用到的统计指标。例如，包括了基尼系数、塞尔系数、广义熵系数等用来反映教育不平等程度的统计指标。(2) 除了介绍一般线性回归模型外，还特别包括了随机前沿分析(SFA)、数据包络分析(DEA)、分位数回归、多分因变量模型、定序因变量模型、格兰杰因果检验方法、单位根检验、协整分析等内容。

三、兼顾软件应用。本书在介绍各种计量方法时，都采用具体数据在STATA统计软件中实现。读者可以重复书中的例题，达到将理论知识、计量方法、软件应用三者紧密结合的目的。对于从事教育研究的人来说，通过对本书的学习可以很快掌握定量研究的方法并将之应用到研究实践中。

四、提供练习数据。本书在介绍计量方法的同时，特别注重数据的采集，介绍了教育经济学研究中常用的数据来源。读者如果想要获得本书例题中使用的数据，可以与作者联系。

本书获得了北京大学出版社的资助，在此深表谢意。另外，作者还要特别感谢北京大学出版社刘维、李淑方对本书出版所提供的大力帮助。由于本人水平有限，撰写过程中错误和疏漏在所难免，敬请读者批评指正。

编者

2008年8月18日

目 录

第一章 统计学基础知识

在数字化的时代里，人们越来越喜欢用数字来描述世界。在教育研究中，实证研究方法越来越受到重视并得到广泛的应用。例如，分析一个国家的人力资源发展水平时，常用人均受教育年限这一指标，在研究教育财政经费配置的公平性时，常常用生均教育经费投入的极值比、变异系数、基尼系数等统计指标。为此，本章将介绍教育统计中常用的统计指标。当我们研究单个变量的特征时，主要从集中趋势、变异趋势、分布状况等三个方面来分析。

第一节 表示集中趋势的统计指标

在介绍统计指标之前，我们首先来区分数据的类型和度量尺度。按照类型，数据可以分为*数值数据*(numerical data 或 quantitative data)和*属性数据*(attribute data 或 qualitative data)。数值数据是指可以用数字表示数量或者测量结果的数据，如学生的身高、班级的学生数等。属性数据是指可以用非数字特征进行分类的数据，如学生的性别、教师的学历等。

按照数据度量尺度，数据可以分为定类、定序、定距、定比等4种尺度，各种度量尺度的特点和例子见表1.1。度量尺度的信息含量由弱到强排列为：定类、定序、定距、定比。一般来说，定类尺度和定序尺度用于属性数据，信息量较低；定距尺度和定比尺度用于数值数据，信息量较高。

表 1.1 数据的度量尺度

尺度	特点	例子
(1) 定类尺度(Nominal)	无等级次序排列	性别,行业,省份
(2) 定序尺度(Ordinal)	可作等级次序排列	教师职称,班级排名
(3) 定距尺度(Interval)	没有真正的零点	标准分,年份,温度
(4) 定比尺度(Ratio)	存在真正的零点,倍数有意义	身高,体重

1. 算数平均数

算数平均数(arithmetic mean)是度量数据集中趋势最常用的指标,其定义式如下:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{1.1.1}$$

例如,人均受教育年限是衡量人力资本的重要指标,用算数平均数计算的我国城市和农村人均受教育年限的结果如表 1.2 所示。

表 1.2 中国城乡居民平均受教育年限(单位:年)

年份	1982 年	1990 年	2000 年
城市	7.93	8.82	10.2
农村	5.01	6.04	7.33

资料来源:中国教育与人力资源问题报告课题组.从人口大国迈向人力资源强国[M].北京:高等教育出版社,2003:49.

算数平均数如同平衡点,容易受极端数值和数据分布情况的影响。因此,在计算算术平均数之前,一般要对异常值进行处理,或者对平均数做必要的解释。例如,在统计大学毕业生的平均收入水平时,如果个别人的收入特别高,就会使得平均数变大,而多数人的收入都不会达到平均数的水平。

2. 加权算数平均数

加权算数平均数(weighted arithmetic mean)是将各个数据乘以反映其重要性的权数(weight)再求平均的方法。其定义式如下:

$$\overline{X}_W = \frac{w_1 X_1 + w_2 X_2 + \cdots + w_n X_n}{w_1 + w_2 + \cdots + w_n} \tag{1.1.2}$$

在教育研究中,很多时候会用到加权算数平均数。例如,对某门课程的成绩进行计算时,常常对平时成绩、期中成绩和期末成绩赋予不同的权重,然后计算加权的平均成绩。有些时候,当我们对"平均数"进行平均时,容易忽略加权的问题,在计算平均值的时候要格外小心。例如,教育部每年在统计全国本科毕业生的就业率时,先由各省(区、市)上报各省(区、市)的就业率,然后进行平均得到全国本科毕业生的就业率。因为各省(区、市)的本科毕业生规模不一样,因此,不能直接用算

数平均数计算，而应该用各省(区、市)的本科毕业生数进行加权。下面看一个具体的例子。

例 1.1 已知 2004 年京、津、沪的小学生均教育经费支出分别为 6 411 元、3 621元和 9 039 元，求京、津、沪三个直辖市的小学生均教育经费支出是多少？

解：因为三个城市的小学生人数不同，不能用简单算数平均数，而应该以各市的小学生数为权重，进行加权平均。根据 2005 年《中国统计年鉴》的数据可知，2004 年京、津、沪的小学在校生数分别为 516 042 人、554 844 人和 542 898 人，因此三个直辖市的小学生均教育经费支出为

$$\overline{X}_{\mathrm{w}}=\frac{516\,042\times 6\,411+554\,844\times 3\,621+542\,898\times 9\,039}{516\,042+554\,844+542\,898}\approx 6\,336(\text{元})$$

3. 几何平均数

几何平均数(geometric mean)是 n 个数乘积的 n 次方根。一般来说几何平均数会小于或等于算数平均数。几何平均数的定义式如下：

$$g=\sqrt[n]{X_1\times X_2\times\cdots\times X_n} \tag{1.1.3}$$

几何平均数常用于计算增长率、百分比、比率、指数等指标的平均数。上述公式要求每个 X 都大于 0，但是有时这个条件很难满足。例如，当计算平均增长率时，每个 X_i 都表示增长率，而增长率常常会出现 0 值或者负值。为了避免零增长率或负增长率，在应用中几何平均数的计算一般改为如下公式：

$$g=\sqrt[(n-1)]{\frac{X_2}{X_1}\times\frac{X_3}{X_2}\times\cdots\times\frac{X_n}{X_{n-1}}}-1=\sqrt[(n-1)]{\frac{X_n}{X_1}}-1 \tag{1.1.4}$$

此处的 X_i 不是增长率，而是绝对水平值，一般不会出现 0 或者负值。而且这一公式还有个优点，即它只用到首尾两个数，中间的数据可以忽略。在计算一国 GDP 的增长率时通常用的就是实际 GDP 的几何平均数。

例 1.2 计算我国普通高校招生规模的增长率。根据 2007 年《中国统计年鉴》公布的数据，1978 年我国普通高校招生规模为 40.2 万人，2006 年为 546.1 万人，则 1978 年到 2006 年的增长率

$$g=\sqrt[28]{\frac{546.1}{40.2}}-1\approx 0.097\,7=9.77\%$$

上述公式适合于计算器计算，如果用 EXCEL 统计软件计算的话，可以利用先取对数再求指数的技巧，将开方的问题转化为乘除问题。具体的表达式如下：

$$g=\exp(\ln(X_n/X_1)/(n-1))-1 \tag{1.1.5}$$

上述例子的计算过程可以表示为：$g=\exp(\ln(546.1/40.2)/28)-1=0.097\,7=9.77\%$。

4. 中位数

中位数(median)表示经排序后的一列数据中位于中间位置(50%)的数。也就

是说，在这组数据中，有一半的数据大于中位数，另外一半的数据小于中位数。中位数的位置为$(n+1)/2$。如果n是奇数，中位数即为排序序列居中位置的观测值；如果n是偶数，则中位数为排序序列两个居中位置的观测值的均值。中位数的优点是不受数据极端值的影响。

在衡量劳动力市场上从业人员的平均收入水平时，一般不仅要统计算数平均数，通常也要统计中位数。因为收入的算数平均数一般都大于中位数，即多数人的收入在算数平均数之下。如果只知道算数平均数，我们并不能确定有多大比例的人其收入低于算数平均数。中位数对于求职者来说是一个非常重要的参考指标，把收入期望定在中位数（而不是算数平均数）的水平上，求职者既能保证收入高于大多数人，而且相对而言更容易实现求职目标。

5. 众数

众数（mode）是指出现次数最多的那个数的数值，可从排序数组中观察得到。众数不受极端值的影响，一组数据可能没有众数或有多个众数，并且对数值数据和类型数据均适用。

6. 值域中点

值域中点（midrange）也是用于度量数据的集中趋势，是最小和最大观测值的平均值。它对数据的极端值非常敏感，常用于金融分析和气象预报，在教育研究中用得不多。

为了说明算数平均数、中位数、众数和值域中点之间的区别和联系，表1.3给出这4个统计指标的比较。如果数据分布是对称的，则算数平均数、中位数、众数和值域中点大致相等。如果数据分布是非对称的，那么最好既报告算数平均数也报告中位数。算数平均数是相对可靠的，当从样本数据估计总体特征的时候，算数平均数和其他平均数相比有更高的一致性，也就是说从不同样本得到的算数平均数差别不大。

表1.3 算数平均数、中位数、众数和值域中点的比较

平均	算数平均数	中位数	众数	值域中点
定义	$\overline{X}=\frac{1}{n}\sum X_i$	排序序列居中位置的观测值	频数最大的数值	最大值与最小值的平均
是否常用	最常见的“平均”	常用	有时用	少用
存在性	总是存在	总是存在	可能不存在；也可能不唯一	总是存在
考虑每个值	是	不是	不是	不是
受极端值影响	是	不是	不是	是
优缺点	容易计算；有良好的统计性质	有极端值存在时是很好的选择	对定类数据最适合	对极端值敏感

资料来源：Triola, Mario F. Elementary Statistics [M]. Addison Wesley Longman Inc. 2003. pp. 63.

7. 四分位数

四分位数(quartile)是用于度量数据的集中趋势以及非集中趋势的指标，把排序数据等分为4个区间，第 i 个四分位数的位置

$$\text{position of } Q_i = \frac{i \times (n+1)}{4} \tag{1.1.6}$$

Q_1 表示第一四分位数(first quartile)，是处于 $(n+1)/4$ 位置上的观测值。25%的观测值比第一四分位数小。

Q_2 表示第二四分位数(second quartile)，就是中位数。处于 $2(n+1)/4=(n+1)/2$ 的位置上。50%的观测值比中位数小。

Q_3 表示第三四分位数(third quartile)，是处于 $3(n+1)/4$ 位置上的观测值。75%的观测值比第三四分位数小。

也就是说，第一、二、三四分位数是位于第25%、50%、75%位置上的数值。

8. 中轴数

中轴数(midhinge)是用于度量数据的集中趋势的指标，是第一和第三四分位数的平均值。因为只与两个数有关，特别是排除了最小的25%和最大的25%的数据，因此不受数据极端值的影响。

第二节 表示变异程度的统计指标

1. 方差与标准差

方差(variance)是指每个数据与算数平均数之差的平方的算术平均数，简称为“离差平方和的平均”。它表示全部观测值相对于均值的平均变异程度，度量的是数据的离散程度。方差的定义式为

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.2.1}$$

标准差(standard deviation)是指方差的平方根。其计算公式为

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{1.2.2}$$

方差与标准差的值越大，表示数据的分散程度越大；相反，方差与标准差的值越小，表示数据的分散程度越小，数据向算术平均数的集中程度就越高。

上述方差和标准差的计算公式是针对普通的一组数的，如果是用样本数据估计总体的方差和标准差，为了得到无偏估计量，则公式中分母应该由 n 变为 $n-1$。

2. 全距

全距(range)是指最大值减去最小值，不考虑数据是如何分布的，只取决于数据的极端值。全距与原数据的单位相同，全距也被称为极差。全距的定义式如下：

$$R = X_{\max} - X_{\min} \tag{1.2.3}$$

例如，在比较我国各省(区、市)人均受教育年限的差异时，全距可以用来表示落差的大小。

3. 极值比

极值比(range ratio)是指最大值与最小值之比，也不考虑数据是如何分布的，只取决于数据的极端值。极值比没有单位。极值比的定义式为

$$RR = \frac{X_{\max}}{X_{\min}} \tag{1.2.4}$$

在一般情况下，全距和极值比所反映的内容是一致的，都是看两个极端值之间的差别。在有些情况下，当我们要进行比较研究时，计量单位可能出现不一致或者不可比较的情况，这时用极值比更合适。例如，比较 1978 年与 2008 年各省(区、市)之间小学教师工资的差别时，如果用全距就不合适，因为不同年份的物价水平是不同的。

4. 变异系数

变异系数(coefficient of variation)是标准差相对于均值的比率，度量的是相对离散程度，可以用来比较两组或多组计量单位不同的数据的变异程度。其公式为

$$CV = \frac{S}{\overline{X}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}{\overline{X}} \tag{1.2.5}$$

5. 基尼系数

基尼系数(Gini coefficient)是 20 世纪初意大利经济学家基尼根据洛伦兹曲线设计的判断收入分配平等程度的指标。基尼系数取值范围在[0,1]之间，收入分配越是趋向平等，基尼系数也越小；反之，收入分配越是趋向不平等，那么基尼系数也越大。

基尼系数表示的是相对差距，取值多大才算是差距显著？对这一问题并没有统计意义上的明确标准。当用基尼系数衡量居民收入的差距时，通常将取值范围划分为六个区间：低于 0.2 表示收入绝对平均；0.2～0.3 表示比较平均；0.3～0.4 表示相对合理；0.4～0.5 表示收入差距较大；0.5～0.6 表示收入差距很大；0.6 以上表示收入差距悬殊。通常，0.4 被看做基尼系数的警戒线。

基尼系数的计算方法有很多种：几何方法、基尼的平均差方法(或相对平均差方法)、斜方差方法、矩阵方法、回归拟合方法等。每种方法都有其自身的优点和特

殊的用途。[①] 下面介绍的是几何方法。

基尼系数的图形表示如图 1.1 所示。横坐标表示按照个人收入从低到高排列后的累计人数比率；纵坐标表示的是相应的累计收入比率。将两者构成的散点连接起来形成的曲线，就形成了洛伦兹曲线。直线 OD 被称为完全平等线，收入分布越平等，则洛伦兹曲线就越逼近完全平等线。

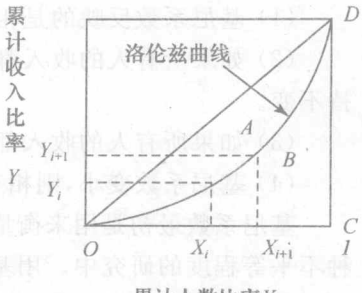


图 1.1 洛伦兹曲线

用 A 表示直线 OD 与曲线 OD 所围成区域的面积，用 B 表示曲线 OD 与直线 OC、CD 所围成区域的面积，则面积 A 占面积 A 和 B 的比值称之为基尼系数。

显然，基尼系数与洛伦兹曲线的位置有密切的关系。特别地：(1) 当洛伦兹曲线变为对角线 OD 时，基尼系数为 0，表示收入绝对的平等。累计人数每增加一个百分点，累计收入也增加一个百分点。(2) 当洛伦兹曲线变为折线 OCD 时，基尼系数为 1，表示收入绝对的不平等。这时全部的收入都被一个人所占有。

基尼系数的计算公式：

$$G = 1 - \sum_{i=0}^{n-1} (X_{i+1} - X_i)(Y_i + Y_{i+1}) \tag{1.2.6}$$

其中，X_i 为累计人数比率；Y_i 为累计收入比率；n 表示不同收入人群组的个数，一般情况下它不等于总的观测人数。上述公式可以改写为

$$G = \sum_{i=1}^{n-1} X_i Y_{i+1} - \sum_{i=1}^{n-1} X_{i+1} Y_i \tag{1.2.7}$$

当每个收入组都只含有一个人时，n 就等于总的观测人数，假定 X_i 表示第 i 个人的收入，则基尼系数可以改写为

$$G = 1 - \sum_{i=1}^{n} \frac{1}{n} \left[2 \sum_{j=1}^{i} \frac{X_j}{\sum_{k=1}^{n} X_k} - \frac{X_i}{\sum_{k=1}^{n} X_k} \right] \tag{1.2.8}$$

利用此公式计算基尼系数的步骤如下：

(1) 对数据 X_i 进行排序。

(2) 计算每个收入占全体收入的比重。

(3) 计算累积比重。

(4) 计算“2 倍”的梯形面积。

(5) 计算基尼系数。

① 徐宽. 基尼系数的研究文献在过去八十年是如何拓展的[J]. 经济学季刊，2003，2(4)：757—778.

基尼系数有以下性质：

(1) 基尼系数反映的是相对差异，不是绝对差异。

(2) 如果所有人的收入都有相同比例的增加或者减少，则基尼系数的数值保持不变。

(3) 如果所有人的收入都有相同数量的增加，则基尼系数的数值将变小。

(4) 基尼系数变小，则相对差异一定变小，但是绝对差异未必变小。

基尼系数最初是用来衡量收入不平等程度的，现在已经被广泛应用到衡量各种不平等程度的研究中。用基尼系数衡量同一经济指标或者教育指标时，可以做纵向比较和横向比较。但是，对于不同的指标，基尼系数不可做简单的比较。例如，用基尼系数衡量受教育年限的差距程度，一般来说基尼系数取值不会很大，不能简单地与收入的基尼系数做比较。

在使用基尼系数研究教育问题时，应该说明使用的是什么教育变量以及分析的基本单位是什么。托马斯等(Thomas *et al*，2001)使用 85 个国家和地区在 1960—1990 年间 15 岁以上人口的受教育年限数据，计算出教育成就的基尼系数。他们发现，绝大多数国家在这 30 年间教育不平等程度都呈现显著的减弱趋势；教育不平等程度与平均受教育年限呈负相关关系，意味着教育发展水平高的国家其教育平等程度也高。①

岳昌君(2008)采用 2003 年各省(区、市)的教育发展数据分别计算入学率和生均教育经费支出的基尼系数。从入学率的基尼系数来看，我国省份之间的教育不均衡程度随着教育层次的提高而扩大。相比之下，义务教育阶段的教育入学机会最为均衡，高中阶段的入学机会不均衡程度显著大于小学和初中，而大学阶段的入学机会不均衡程度最为严重。从生均教育经费支出的基尼系数来看，我国各省(区、市)之间的教育不均衡程度随着教育层次的降低而扩大。相比之下，大学的生均教育经费支出最为均衡，基尼系数为 0.14；高中的不均衡程度显著增加，基尼系数为 0.22；而初中和小学的不均衡程度最为严重，基尼系数分别达到 0.25 和0.26。图 1.2 显示出我国各级教育的入学机会与教育经费投入之间存在“剪刀差”关系：一方面，虽然义务教育的入学机会均衡程度很高，但是教育经费投入差异很大。这说明近年来我国的“普九”只做到了普及“数量”(入学率都非常高)，但是没有做到普及“质量”(薄弱学校的教育投入严重不足)；另一方面，各省份之间的大学教育经费投入虽然差别不大，但是入学机会却很不均衡。

① Vinod Thomas, Yan Wang, Xibo Fan. Measuring education inequality: Gini coefficients of education[R]. Policy Research Working Paper, 2001, no. WPS2525.

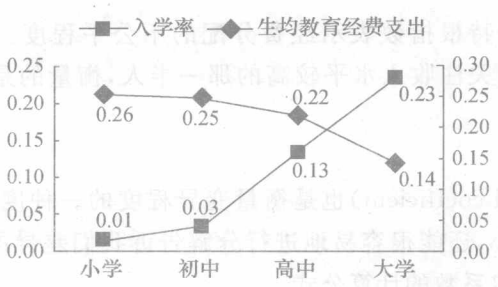


图 1.2 以入学率和生均教育经费支出为变量的基尼系数

6. 麦克伦指数

麦克伦指数(McLoone index)指数据落在中位数以及中位数以下的总和,与落在中位数以下者,如果都达到中位数时可以得到的总和,二者相除所得的比例。麦克伦指数取值范围在 0 和 1 之间,麦克伦指数越大表示分配越公平,越小表示分配越不公平,这是与其他分散指标最大的不同。其公式为

$$\text{McLoone index} = \frac{\sum_{i=1}^{m} X_i P_i}{X_{MP} \sum_{i=1}^{m} P_i} \tag{1.2.9}$$

其中,X_{MP}表示中位数;当 n 为偶数,$m=\frac{n}{2}$;当 n 为奇数,$m=\frac{n+1}{2}$;P_i是权重。例如,假定 X_i表示各县小学的生均教育经费支出,P_i表示各县小学生的人数,则可用上述麦克伦指数表示经费分配的不公平程度。基尼系数考虑的是样本中的所有人,测量的所有人之间的不平等程度。而麦克伦指数只关注收入水平较低的那一半人,衡量的是"穷人"之间的不平等程度。

7. 沃斯特根指数

沃斯特根指数(Verstegen index)指数据落在中位数以及中位数以上的总和,与落在中位数以上者,如果都达到中位数时可以得到的总和,二者相除所得的比例。沃斯特根指数取值范围在 1 以上,沃斯特根指数越小表示分配越公平,越大表示分配越不公平。其公式为

$$\text{Verstegen index} = \frac{\sum_{i=m}^{n} X_i P_i}{X_{MP} \sum_{i=m}^{n} P_i} \tag{1.2.10}$$

其中,X_{MP}表示中位数;当 n 为偶数,$m=\frac{n}{2}+1$;当 n 为奇数,$m=\frac{n+1}{2}$;P_i是权重。例如,假定 X_i表示各县小学的生均教育经费支出,P_i表示各县小学生的人

数，则可用上述沃斯特根指数表示经费分配的不公平程度。与麦克伦指数相对应，沃斯特根指数只关注收入水平较高的那一半人，衡量的是“富人”之间的不平等程度。

8. 塞尔系数

塞尔系数(Theil coefficient)也是衡量变异程度的一种度量指标，它不仅可以告诉我们差异的大小，还能很容易地进行分解告诉我们差异的原因。下面以地区差异为例来说明塞尔系数的计算公式。

以省为单位的差异：

$$T_P = \sum_i \sum_j \left(\frac{Y_{ij}}{Y}\right) \ln \left(\frac{Y_{ij}/Y}{N_{ij}/N}\right) \tag{1.2.11}$$

以地区内的省份为单位的差异：

$$T_{Pi} = \sum_j \left(\frac{Y_{ij}}{Y_i}\right) \ln \left(\frac{Y_{ij}/Y_i}{N_{ij}/N_i}\right) \tag{1.2.12}$$

以地区为单位的差异：

$$T_{BR} = \sum_i \left(\frac{Y_i}{Y}\right) \ln \left(\frac{Y_i/Y}{N_i/N}\right) \tag{1.2.13}$$

易证，下列的分解公式成立：

$$T_P = \sum_i \sum_j \left(\frac{Y_{ij}}{Y}\right) \ln \left(\frac{Y_{ij}/Y}{N_{ij}/N}\right) = \sum_i \left(\frac{Y_i}{Y}\right) T_{Pi} + T_{BR} = T_{WR} + T_{BR} \tag{1.2.14}$$

塞尔系数的优点是，它可以比较容易地将差异分解为组内差异和组间差异。对于同样的分组可以比较不同年份的差异变化，以及引起差异变化的原因所占比例的变化。比如，如果有县级小学的生均教育经费的数据，我们可以以县为单位计算全国小学的生均教育经费的塞尔系数，并且分解为省内差异和省际差异，从而比较省内差异和省际差异对全国小学的生均教育经费差异的贡献大小。

从塞尔系数的公式来看，在塞尔系数给定的情况下，组间差异和组内差异的大小与分组有关。一般来说，在分组方式相同的情况下，分组的个数越多则组间差异越大。特别地，只分一组，则全部差异都归于组内差异；如果每个观测都是一组，即每组只包括一个观测值，则全部差异都归于组间差异。

如果分组方式不同，则分组的个数与组间差异的关系是不确定的。例如，假定样本数据包括中国大陆 31 个省(区、市)分城乡的平均年收入，共有 62 个观测值。我们可以计算收入的塞尔系数，并试图分析引起差异的原因。假定城乡差异是我国收入差异的主要原因，当我们按照城乡分组，即使只有两组，那么组间差异相对也会很大；当按照省区市分组，即使组数很多，每组只有两个观测值，那么组间差异相对也会不大。