

Excel

在统计分析中的应用

王文中 编著

本书特色：

- 既可以了解统计学，又可以马上用Excel来作数据分析。
- 以循序渐进的方式说明Excel的各个步骤，使您能够迅速将理论与实际相结合。
- 附有精心设计的范例，您可边学边做，效果更加突出。
- 每一章的后面附有作业和习题，可供您自我评价学习成果。
- 可作为统计学的教科书，也可作为数据分析的自学参考书。

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

Excel

在统计分析中的应用

王文中 编著

中国铁道出版社

2003 • 北京

(京)新登字 063 号

北京市版权局著作合同登记号: 01-2002-4089 号

版 权 声 明

本书中文繁体字版由台湾博硕文化股份有限公司出版。本文中文简体字版经台湾博硕文化股份有限公司授权由中国铁道出版社出版。任何单位或个人未经出版者书面允许,不得以任何手段复制或抄袭本书内容。

本书贴有博硕文化激光防伪标签,无标签者不得销售。版权所有,侵权必究。

图书在版编目(CIP)数据

Excel 在统计分析中的应用/王文中编著. —北京:中国铁道出版社, 2002. 11

ISBN 7-113-04999-0

I. E… II. 王… III. 电子表格系统, Excel-应用-统计分析 IV. C819

中国版本图书馆 CIP 数据核字(2002)第 088770 号

书 名: Excel 在统计分析中的应用

作 者: 王文中

出版发行: 中国铁道出版社(100054, 北京市宣武区右安门西街 8 号)

策划编辑: 严晓舟 郭毅鹏

责任编辑: 苏 茜 彭立辉

封面设计: 孙天昭

印 刷: 河北省遵化市胶印厂

开 本: 787×1092 1/16 印张: 27.75 字数: 657 千

版 本: 2003 年 1 月第 1 版 2003 年 1 月第 1 次印刷

印 数: 1~5000 册

书 号: ISBN 7-113-04999-0/TP. 813

定 价: 42.00 元

版权所有 侵权必究

凡购买铁道版的图书,如有缺页、倒页、脱页者,请与本社计算机图书批销部调换。

出版说明

学过统计学的人，大概都不会忘记背诵那些艰深符号和公式的痛苦经历，尤其在计算机统计方面更是如此。事实上，统计软件足以解决这个窘境，但一些专业软件过于昂贵，个人很难负担，且需要学习很久，因此难以满足使用者的迫切要求。Excel 则刚好克服了这些缺点，它非常普及且易学易用，可立即进行数据分析。

本书一方面讲述了统计学的基本概念和数据分析方法，另一方面利用 Excel 的数据分析功能，通过实例一步步说明了操作步骤及其分析结果。

另外本书还配有范例及解答，如果读者需要，请到 <http://www.tqbooks.net> 处下载。

本书由台湾博硕文化股份有限公司提供版权，经中国铁道出版社计算机图书中心审选，由方玲、张向阳、范永生、朱霞、肖志军、廖康良、陈贤淑和裘伟力等完成整稿及排版工作。

中国铁道出版社
2002 年 11 月

本书导读

Excel 的魅力

近年来，由于计算机技术的发展，很少有人再利用手算或是计算器来进行数据分析。市面上有关统计分析的计算机软件非常多，例如 SPSS、SAS、BMDP、STATISTICA、SYSTAT 等。只可惜这些计算机软件存在着以下两大缺点。第一：软件价格过于昂贵（高达数万元至数十万元），通常是学校等机构购买，SAS 软件更是只租不卖，对一般学生、民众、甚至公司而言，都负担不起，因此普及性很低。第二：这些软件艰深难懂，非短期所能学会，通常需要数月甚至数年才能熟练。

在没有适当统计软件的情形下，我们常常被迫拿起计算器，以最原始的分析方法来解决。这在 21 世纪的今天，显得有些滑稽。这种窘境，相信不少人都经历过。这正反映出我们迫切需要一套经济实用的软件，人人负担得起，普及性够高，而且容易操作，一学就会。

在我的教学经验里，常苦于无法让学生进行统计实习，因为如果要教学生利用上述的统计软件包来进行数据分析，恐怕得花上半个学期至一个学期的时间，根本不可行。如果只让学生用手或计算器来算，实在跟不上潮流，也不切实际，当然也无法进行较为复杂的统计分析。就算学生学会手算，过一阵子后也会忘得一干二净。一旦有数据要进行统计分析，就束手无策。这个困境直到最近几年我尝试利用 Excel 进行统计分析，才获得解决。在我的教学里，除了课堂上讲解统计学的理论外，课后（或实习课）还要求学生根据本书，进行数据分析与结果解释。学生的反应出奇的好，他们觉得通过 Excel，可以快速的分析数据，让他们觉得统计学是门实用的学问，而不是纸上谈兵，完全派不上用场。

Excel 价格便宜，而且非常普及，几乎所有的个人计算机都装有 Office 软件。除此之外，Excel 易学易用，非常简单。虽然它的统计分析功能不及上述这些专业软件，不过对绝大多数的用户而言，已经足够。更何况 Excel 除了统计分析外，还有更为强大的会计、绘图、数据库管理等功能，这是上述软件所欠缺的。可惜的是，几乎绝大多数的 Excel 用户，或是一般需要进行统计分析的人，并不知道原来可以利用 Excel 来进行统计分析，当然也就不知道如何使用了。因此，我们需要一本好书，除了介绍基本的统计学理论外，还要介绍如何利用 Excel 来进行数据分析，让我们能够轻松地完成数据分析的任务。

本书就在这样的理念下产生。一方面讲解统计学基本原理、区间估计、假设检验等重要议题。另一方面，配合 Excel 的画面，深入浅出地逐一说明利用 Excel 进行数据分析的步骤，以及解读报表的方法。

本书的特色与适用对象

这是本一举两得的书，既可了解统计学又可用 Excel 来进行数据分析。本书每一章的撰

写方式大多先介绍统计学的基本原理,然后再辅以 Excel 实习,以循序渐进的方式,说明 Excel 的各个步骤,让读者能够马上将理论与实际结合。每一章的后面附有作业和计算机习题,可供读者自我评价学习成果。务求通过边做边学,掌握数据分析的全貌。

本书可以当作大专院校统计学的教科书,也可供下列对象自修使用:

略有统计学概念,想学一种简单易用的软件来分析数据者。

略有 Excel 基础,想了解如何用它来进行数据分析者。

未具备统计学和 Excel 的基础,但想了解统计学和 Excel 数据分析者。

本书适用一整学年的授课。如果只是一学期,可将某些章略过,如第 14 章的多因素方差分析,第 16 章的复回归,第 17 章的一般线性模式,第 18 章的非参数统计等。除此之外,教师可依实际授课情形,再将某些较为复杂的单元略过。

激活 Excel

在计算机开机后,请检查是否有 Excel 软件。通常会在屏幕上发现 Excel 的符号,类似图 1 所示。然后在该符号上双击,即可激活 Excel。



图 1

如果没发现,则可单击屏幕左下角的“开始”按钮,若“程序”菜单中出现“Microsoft Excel”,就表示这部计算机装有 Excel。

一旦顺利激活 Excel,就可以发现如图 2 所示的画面。此即为 Excel 中的基本界面。可以在此工作底稿内输入两大类型的数据。一是常数,包括数值、日期、时间、货币、百分比、分数、科学记号或者是文字。另一是公式,公式是一连串的常数、储存格参照地址、名称、函数或是运算式。公式通常以“=”开始,如键入“=SUM(3,5)”,就会出现 8,其中 SUM 是汇总的函数。

本书受限于篇幅,无法详述 Excel 的基本操作,读者若对 Excel 完全陌生,必须参阅 Excel 使用手册或其他的参考书。最佳的学习方式就是直接上机,边操作边阅读使用手册或参考书,当然也可以多使用 Excel 的 help 功能。

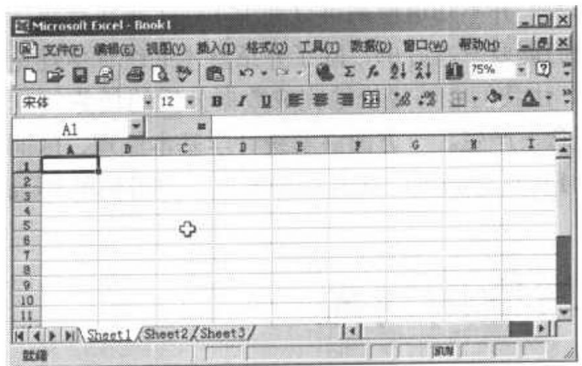


图 2

安装分析工具库

在默认的情况下，Excel 并没有安装“分析工具库”，因此无法进行统计分析，请查看计算机是否已经安装。请先激活 Excel，打开菜单栏中的“工具”菜单，如图 3 所示。检查里面是否有“数据分析”的功能。



图 3

在图 3 中的“工具”菜单内并没发现“数据分析”命令，这表示还没安装，可依照下列步骤来安装。(1) 首先选取“工具”中的“加载宏”(Add-ins)，即可出现如图 4 所示的对话框。(2) 然后选取“分析工具库”，再单击“确定”按钮即安装完毕。此时请再次查看“工具”菜单，即可发现“数据分析”命令，如图 5 所示。这表示安装完成，已经可以进行统计分析。以后若要进行统计分析，先选取“数据分析”命令，进入对话框即可。

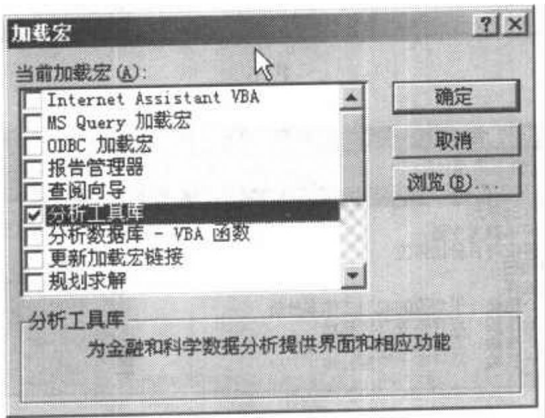


图 4



图 5

Excel 的数据分析功能

Excel 的“数据分析”中提供了 16 种主要的统计方法，当选取“工具”内的“数据分析”命令时，即可进入如图 6 和图 7 所示的对话框，可以发现这 16 种主要的统计方法。

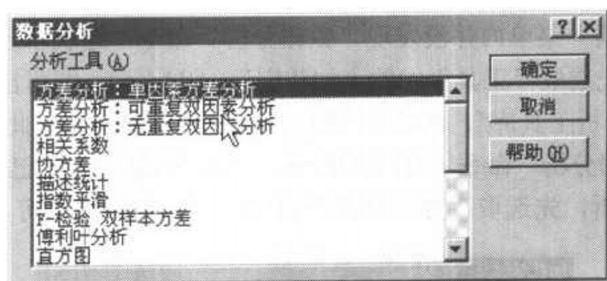


图 6

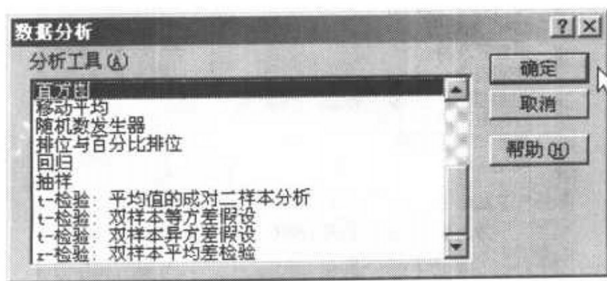


图 7

我们将在稍后的各章内陆续使用这些内设的分析方法。除此之外，Excel 还提供了相当多的函数，可以帮助进行统计分析。同时也可绘制图表，增加数据的可读性。

目 录

Chapter1 绪论.....	1
1-1 统计学的意义	2
1-2 变量的分类	2
1-3 四种测量尺度	3
1-4 作业	5
Chapter2 集中指标.....	7
2-1 平均数	8
2-2 其他集中指标	9
2-3 平均数、中位数、众数的比较.....	10
2-4 Excel 练习	13
2-5 作业	21
Chapter3 变异指标与分布形状.....	23
3-1 变异指标	24
3-2 各种变异指标的比较	26
3-3 切比雪夫不等式定理	27
3-4 分布形状	27
3-5 Excel 练习	29
3-6 作业	32
Chapter4 正态分布.....	35
4-1 正态分布的特性	36
4-2 标准正态分布	41
4-3 峰度与偏态	43
4-4 Excel 练习	45
4-5 作业	47
Chapter5 离散概率分布.....	49
5-1 概率理论	50
5-2 联合、边缘、条件概率.....	50
5-3 贝叶斯定理	53
5-4 离散概率分布	54
5-5 Excel 练习	61
5-6 作业	64

Chapter6 连续概率分布	65
6-1 连续变量	66
6-2 连续概率分布	67
6-3 随机变量的一些特性	80
6-4 Excel 练习	82
6-5 作业	85
Chapter7 抽样分布与估计式	87
7-1 抽样误差	88
7-2 抽样方法	89
7-3 抽样分布	90
7-4 估计式	96
7-5 Excel 练习	102
7-6 作业	105
Chapter8 区间估计	109
8-1 平均数的区间估计	110
8-2 两平均数之差的区间估计	116
8-3 方差的区间估计	123
8-4 比例的区间估计	125
8-5 Excel 练习	129
8-6 作业	138
Chapter9 假设检验	141
9-1 统计假设	142
9-2 一个总体平均数的检验	146
9-3 两总体平均数之差的假设检验	149
9-4 方差的假设检验	155
9-5 比例的假设检验	157
9-6 统计检验力	160
9-7 Excel 练习	165
9-8 作业	175
Chapter10 分类数据分析	179
10-1 适合度检验	180
10-2 独立性检验	181
10-3 同质性检验	183
10-4 改变的方向性检验	184
10-5 标准化残差	185
10-6 期望值的大小	186
10-7 相关度的测量	187

10-8	Excel 练习	189
10-9	作业	194
Chapter11	线性相关	197
11-1	协方差	198
11-2	皮耳森积差相关	200
11-3	相关系数的假设检验	202
11-4	相关系数的区间估计	205
11-5	相关系数的一些现象	207
11-6	其他类似的相关系数	208
11-7	Excel 练习	211
11-8	作业	217
Chapter12	单因素方差分析 I	219
12-1	结构模型	220
12-2	平方和的分解	222
12-3	平均数的估计与检验	227
12-4	多重比较	231
12-5	Excel 练习	237
12-6	作业	241
Chapter13	单因素方差分析 II	243
13-1	模型假设的检验	244
13-2	因变量尺度转换	251
13-3	实验效果值	252
13-4	统计检验力	254
13-5	Excel 练习	255
13-6	作业	257
Chapter14	多因素方差分析	259
14-1	结构模型	260
14-2	平方和的分解	263
14-3	因素效果的分析	268
14-4	无重复试验	274
14-5	多因素设计	277
14-6	Excel 练习	280
14-7	作业	285
Chapter15	简单线性回归	289
15-1	五种主要的回归线	290
15-2	结构模型与参数估计	291
15-3	预测效果的方差分析	296

15-4	平均数的估计误差	299
15-5	新观测值的预测	300
15-6	模型假设的检验	301
15-7	Excel 练习	302
15-8	作业	307
Chapter16	复回归	309
16-1	一般线性模型	310
16-2	回归系数的估计与检验	311
16-3	预测效果的方差分析	314
16-4	平均数的估计误差	316
16-5	新观测值的预测	317
16-6	标准化回归系数	318
16-7	共线	319
16-8	多项式回归方程式	321
16-9	自变量的选择	322
16-10	Excel 练习	323
16-11	作业	328
Chapter17	一般线性模型	331
17-1	单因素方差分析	332
17-2	多因素方差分析	333
17-3	协方差分析	338
17-4	Excel 练习	343
Chapter18	非参数统计	351
18-1	符号检验	352
18-2	威尔科克森带符号等级检验	354
18-3	克瓦二氏单因素方差分析	354
18-4	弗氏相依样本等级检验	356
18-5	作业	357
Chapter19	统计图	359
19-1	图表向导	360
19-2	覆盖图	363
19-3	图的编辑	366
19-4	图表类型的意义与使用	369
Chapter20	数据透视	375
20-1	次数与百分比	376
20-2	计算次数与平均数	380

Chapter21 抽样调查的实例演练.....	383
21-1 研究设计	384
21-2 结果与解释	387
附表.....	399
作业解答	413

chapter 1

绪论

统计学已经成为大专院校的共同必修或选修科目，这是因为统计学和社会经济发展、人们的日常生活有着密不可分的关系。本章说明统计学的意义，它包含两种功能：描述所收集到的数据的特征，以及将此特征推论到未收集到的数据上。接着说明变量的种类，包括定量变量，如身高；定性变量，如性别。最后依照变量被测量的特性，说明四种尺度：定类尺度（如性别）、定序尺度（如硬度）、定距尺度（如温度）、定比尺度（如身高）。

本章重点

- 统计学的意义
- 变量的分类
- 四种测量尺度

1-1 统计学的意义

statistics（统计学）这个英文单词源自欧语系字根 standing, status, state, understand 等。基本上，这几个字根的意义，都是在描述事物的现象。这也就是统计学的基本意义：描述事物的现象。除了这类描述统计学（descriptive statistics）外，统计学还有另一功能：利用样本（sample）所收集的数据，推论总体（population）的状况，这就是推论统计学（inferential statistics）。

在描述统计学里，注重数据的整理、分析、展示与解释。并没有要将所得的结果推论（generalize）至别的资料或更大的范围上。换句话说，数据的出处就是总体，而不是样本。例如某研究者普查了台北市内所有的幼儿园，结果发现幼儿园教师具有大学学历的百分比占 80%，女性教师占 95%，平均年龄为 25.2 岁。由于本研究关心的是台北市的幼儿园，并不关心其他县市的幼儿园，因此台北市所有的幼儿园就是总体，这就是属于描述统计学的范围。

此外，在未进行推论统计之前，会先对数据加以整理，试图找出一些规则或是有问题的数据（如极端值），称为探索性数据分析（exploratory data analysis），这也算是描述统计学分析的一部分。

在推论统计学里，在理解了数据的意义后，意图通过整理和分析来推论总体的状况。例如某灯泡厂商关心其生产的灯泡寿命，该厂商年产灯泡 100 万个，随机抽取 50 个灯泡，测其寿命，结果发现平均寿命为 1235h，寿命最短为 985h，最长为 1392h。在这个实例里，我们并不只关心这 50 个灯泡的寿命，而是希望通过这 50 个灯泡，来推论所有该厂商生产的灯泡。我们无法将所有的灯泡拿来一一测试，因为一旦测试完毕，灯泡就报废了。

要将 50 个灯泡的结果，推论到所有灯泡上，并不是件简单的事。首先这 50 个灯泡应该能够充分代表所有的灯泡。所谓“充分”在此有两个意义。第一：这 50 个灯泡不可以是故意挑选出来的，而是随机挑选出来，因此具有代表性。第二：受测的样本数不可太小。如果只挑极少数灯泡（如 2 个）进行测试，万一这些灯泡特别好或特别差，就会导致推论失真。反之，数量不必太大（如 5000 个），因为数量反应出人力和物力，如果每个灯泡的寿命差异不大，只要测试一些灯泡即可。

至于样本数要多大才算足够，要看受测物体的特性以及推论的严谨程度而定。如果受测的物体的特性相当一致（如灯泡寿命），则样本数不用太大。反之，如果受测的物体是人的某种特质，如智力，而每个人的智力差异相当大，此时受测人数就要多些。如果推论的品质必须非常严谨，误差要非常小，此时样本数就要很大。反之，如果要求没那么精确，样本数可以小一些。

1-2 变量的分类

在收集资料的过程里，我们收集各式各样的变量（variable），如性别、宗教、班级人数、身高和智力等。基本上，变量可分为两大类：定性变量（qualitative variable）和定量变量（quantitative variable）。性别和宗教是定性变量，因为性别和宗教均无大小之分。班级人数、

身高、智力则是定量变量。

定量变量又可分为两种：离散变量（discrete variable）和连续变量（continuous variable）。班级人数是离散变量，因为它必须是整数。身高和智力则是连续变量，因为任何两个值之间，都可能存在第三个值（至少理论上是如此），因此是连续变量。定性变量必是离散变量。由于定性变量可分为数个类别，因此又称分类变量（categorical variable）。定量变量未必是连续变量，例如班级人数，就是定量变量，但并不连续。这些变量分类的关系如图 1-1 所示。

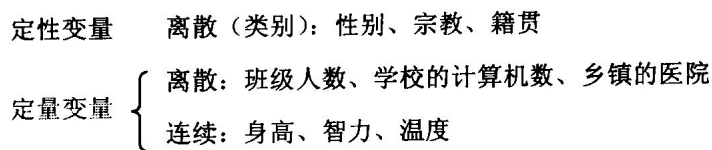


图 1-1

除了上述的分类外，在实验研究（experimental research）里，还可以将变量分为自变量（independent variable）和因变量（dependent variable）。例如调节温度，看生产力的变化。此时温度就是自变量，生产力就是因变量。调整教学方法，看学业成绩的表现。此时教学方法就是自变量，学业成绩就是因变量。在观察研究（observational research）里，关心的是两个变量共同改变的情形，而不是改变某个变量，看另一个变量的变化，因此通常无所谓自变量和因变量之分。例如关心身高和体重的关系，智商和学业成绩的关系，性别和党派的关系等。

1-3 四种测量尺度

除了上述关于变量的分类外，心理学家 S. S. Stevens（1946）提出了四种所谓的测量尺度（measurement scales）：定类尺度、定序尺度、定距尺度和定比尺度，依次说明如下。

性别、宗教等变量属于定类尺度（nominal scale），即使在计算机的文件登录上，男性为 1，女性为 2，但这并不表示 2 大于 1，或 1 大于 2。1 和 2 只是代号，没有顺序的差别。

定序尺度（ordinal scale）则有顺序的意义，例如甲生的学业成绩 90 是班上第一名，乙生的学业成绩 89 是第二名，丙生的学业成绩 88 是第三名。此时第一名在学业成绩上，就比第二名来得好，第二名又比第三名好。因此有着顺序上的意义。不过无法区分第一名和第二名的差距（90 - 89）是否和第二名与第三名（89 - 88）的差距一样。换句话说，虽有顺序意义，但却无法精确描述其间的差距。

例如在问卷调查里，常会要求受试者以所谓的五点尺度来作答，例如：非常不同意、不同意、差不多、同意、非常同意。研究者常分别以 1, 2, 3, 4, 5 代表这五点，但这其实不是定距尺度，应该只能算是定序尺度，因为我们没有证据显示 1 到 2 的差距会等于 2 到 3 的差距。

又如我们想研究一般人的社会支持（social support）程度，请其列出“有急难时可以寻求协助的人”，然后计算人数，人数越多，表示他的社会支持强度越高。此时，甲的人数为 7，乙的人数 6，丙的人数为 5。大致可以说甲的社会支持强度大于乙，乙大于丙。但无法确知，到底他们之间社会支持强度的差异有多大。尽管表面上数字为 7, 6, 5，但只有 $7 > 6 > 5$ 的意

义而已，并不表示 $7 - 6 = 6 - 5$ 。

在自然科学里，“硬度”的测量只能达到定序尺度而已。若要比较物品的硬度，只能两两碰撞，若某个物体的表面受损，就表示其硬度较小。利用这种方法，我们可以将所有物品的硬度加以排序，但无法说明两物品间的硬度差异是多少。

定距尺度 (interval scale) 不仅有顺序意义，还有差距的意义。例如 10°C 和 11°C 之间的差距，等于 11°C 和 12°C 之间的差距。简言之，1 摄氏度的差距在任何度数上都是相等距离，因此称为定距尺度。此时， $12 - 11 = 11 - 10$ 。这个 1 摄氏度的差距表现在水银汞柱的刻度上。

最高级的尺度是定比尺度 (ratio scale)，它除了具有定距尺度的特性外，还具有所谓“自然的零点”。例如身高和体重，有所谓零点，但是温度并没有。因此身高和体重是定比尺度，温度只是定距尺度。身高 200cm 者是 100cm 者的两倍，这个倍数并不会随着采用公制还是英制而改变。温度可就没有这个特性，例如说 20°C 是 10°C 的两倍，但如果换做华氏，则变为 68 度和 50 度，也就不是两倍了。

Stevens 的分类方法，至今仍广受大家采用，因为它的确能够非常简洁地表达出变量的差异。不过这个分类法有非常严重的缺点 (Michell, 1994)。对此感兴趣的读者请参见补充说明 1。

定序尺度只是一个迈向定距尺度的过渡期。以温度的测量而言，最早，我们只能凭着手的感觉去比较两物体的温度。但随着知识的进步，我们了解到热胀冷缩的原理，因此通过水银汞柱的刻度变化，可以精确地测量温度，甚至通过 $PV = nRT$ 的公式，无须测量，就可以直接计算出温度 T 。

现今对于硬度的测量，就像早期对于温度的知识，当时我们不晓得热胀冷缩的原理，也不知道 $PV = nRT$ 的公式。同样的，现在我们仍不知道何种因素会影响硬度，也不知道硬度会有何具体的外在表现，因此硬度的测量只能停留在两两比较上。不过一旦知识向前发展了，对于硬度的测量就会有突破，而达到等距的尺度。

定距尺度和定比尺度的区分，并无太大意义。又或者可以说，所有的定距尺度其实都是定比尺度，因为要产生等距必须要有个参照点，关键在于这个参照点（或称为零点）。温度没有自然的零点，身高则有。但事实上，所谓自然的零点，一点都不自然，因为零根本看不到，摸不到。零不是自然数！只要大伙对于参照点有共识，那么定距尺度就是定比尺度。例如从高雄开始算，到台北的距离就是到台中的两倍。如果没有共识（从高雄开始算），就不知道到台北的距离是到台中的几倍。

若我们同意以 0°C 为参照点，那么 20°C 就是 10°C 的两倍。 0°C 、 10°C 、 20°C 换算为华氏的话分别为 32 度、50 度、68 度。此时变为：“若以华氏 32 度为参照点，那么华氏 68 度就是华氏 50 度的两倍。”因此温度是定比尺度。

补充说明 1: Stevens 四种尺度存在的问题

根据 Michell 的说法，Stevens 的分类方法至少存在以下五种问题。

第一，测量的尺度并没有种类之分。因为测量的尺度来自选定一个单位，然后所有的物体和这个单位来作比较，再观察其数值的比例。定类尺度不是测量的尺度，因为根本没有测量发生。例如性别不是测量得到的，也没测量误差。但我们仍然可以将男生登录为 1，女生