励学·管理学系列

统计学

——原理及应用

吴兰德　编著

南京大学出版社

励学·管理学系

统计学

——原理及应用

吴兰德　编著

南京大学出版社

图书在版编目(CIP)数据

统计学:原理及应用 / 吴兰德编著. —南京:南京大学出版社,2015.5
(励学.管理学系列)
ISBN 978-7-305-14856-9

Ⅰ.①统… Ⅱ.①吴… Ⅲ.①统计学 Ⅳ.①C8

中国版本图书馆 CIP 数据核字(2015)第 048122 号

出版发行 南京大学出版社
社 址 南京市汉口路 22 号 邮 编 210093
出 版 人 金鑫荣

丛 书 名 励学·管理学系列
书 名 统计学——原理及应用
编 著 者 吴兰德
责任编辑 刘群烨 府剑萍 编辑热线 025-83592193

照 排 江苏南大印刷厂
印 刷 宜兴市盛世文化印刷有限公司
开 本 787×1092 1/16 印张 12.25 字数 320 千
版 次 2015 年 5 月第 1 版 2015 年 5 月第 1 次印刷
ISBN 978-7-305-14856-9
定 价 27.00 元

网 址:http://www.njupco.com
官方微博:http://weibo.com/njupco
官方微信号:njupress
销售咨询热线:(025)83594756

前　言

本教材是南京大学金陵学院“321”工程精品教材项目——《统计学——原理与应用》的成果，是应用型人才培养的教改成果之一。感谢南京大学金陵学院相关领导的大力支持和相关部门的资助。

作为应用性人才培养的本科统计学教材，本教材的基本改革思路是注重基本概念的理解和基本统计学方法的应用。

本教材的主要特色与创新在于基本舍弃了对统计学基本原理的数学式推断，代之于各种生活中的例子。很多基本概念使用了一些例子作为辅助，以增加教材的可读性。在习题的选择方面，本教材也尽量使用生活中真实的例子，而不是抽象空洞的假设一些人和事，并且参考了国外的一些优秀教科书，让学生觉得所学的统计学知识能在实践中得到应用。教材中主要的统计学术语和表达都列出了英文表达，对进一步理解概念有一定帮助。

全书一共分为两大部分十二个章节，其中第一章至第四章为描述统计的内容，第五章至第十二章为推断统计的内容。

感谢南京大学出版社府剑萍编辑和刘群烨编辑的辛勤劳动。

由于本教材作者水平有限，遗漏错误之处在所难免，恳请读者批评指正。

吴兰德

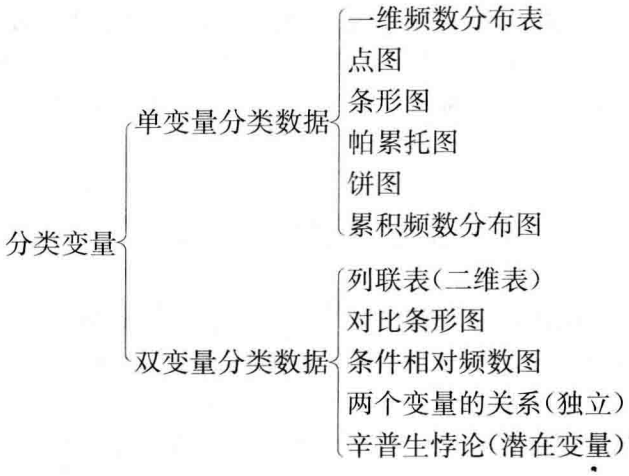
2015 年元月

目　录

第一章　分类变量的描述统计

描述统计主要分为分类变量(categorical variable)的描述统计和数值型变量(quantitative variable)的描述统计。本章先讲分类变量的描述统计。

【知识结构图】



【学习目标】

(1) 用频数分布表描述数据的分布
(2) 用点图、条形图等图形展示数据
(3) 用列联表分析两个分类变量之间的关系

【重难点】

(1) 掌握用条件相对频数表示的对比条形图
(2) 两个分类变量是否独立

1.1　变量的类型

变量(variable)是统计学中一个重要的概念,它是指一个可以取两个或更多个可能值的特征、特质或属性。例如,“性别”,该变量取两个值——男和女;人的“寿命”;“身高”等。而变量的具体取值称为变量值,例如“身高为 170 cm”。

变量主要有如下两种类型。

1.1.1　分类变量(Categorical variable)

分类变量是指只能归于某一类别的非数字的数据,也称定性变量(qualitative variable),“性别”就是分类变量。分类变量的取值一般用文字表示。在计算机软件中,为了

便于处理，可以给它们赋值，如用“0”表示“男”、用“1”表示“女”。

其实，分类变量也可以分为两种类型：一种是不可排序的，像上述的“性别”；一种是可以排序的，如“产品等级”，它的取值是“一等品”“二等品”“三等品”，这些变量的取值虽然也是分类型的数据，但是可以排序。

[**注意**] 分类变量和定性变量是一样的。

1.1.2 数值型变量(Quantitative variable)

这类变量就是我们一般默认的数据类型，它的取值是数字。如“人的寿命”、“身高”、“打靶的环数”等。数值型变量也可以分成两类：一类叫做离散型变量(discrete variable)，它的取值可以一一列举，如“扔骰子的点数”，其结果是1、2、3、4、5、6中的一个；另一类叫做连续型变量(continuous variable)，这种变量研究的比较多，其取值不能一一列举，如“人的寿命”、“身高”等。

1.2 频数分布表

为了做一个正确的决策，一个很重要的事情就是知道变量取值的分布情况。例如，为了确定学校的饮料自动售卖机中哪些饮料多放些，我们就需要确定每种饮料的销售量。这可以通过画频数分布表(frequency distribution table)来展示。

如果是单变量数据，我们画的是一维表；如果是双变量数据，我们画的是二维表。

下面是几个相关概念：

(1) 频数(frequency)。落在某一特定类别中的数据的个数。

(2) 相对频数(relative frequency)。落在某一特定类别中的数据的个数除以样本数据总数。

(3) 百分比频数(percentage frequency)。相对频数乘以100%。

1.2.1 单变量分类数据的频数分布

例1-1 一家市场调查公司为研究不同品牌饮料的市场占有率，随机抽取一家超市进行调查。调查员在某天对50名顾客购买饮料的品牌进行记录，如果一个顾客购买某一品牌的饮料，就将这一饮料的品牌名字记录一次。其原始数据如表1-1所示。

	A	B	C	D	E
1	旭日升冰茶	可口可乐	旭日升冰茶	汇源果汁	露露
2	露露	旭日升冰茶	可口可乐	露露	可口可乐
3	旭日升冰茶	可口可乐	可口可乐	百事可乐	旭日升冰茶
4	可口可乐	百事可乐	旭日升冰茶	可口可乐	百事可乐
5	百事可乐	露露	露露	百事可乐	露露
6	可口可乐	旭日升冰茶	旭日升冰茶	汇源果汁	汇源果汁
7	汇源果汁	旭日升冰茶	可口可乐	可口可乐	可口可乐
8	可口可乐	百事可乐	露露	汇源果汁	百事可乐
9	露露	可口可乐	百事可乐	可口可乐	露露
10	可口可乐	旭日升冰茶	百事可乐	汇源果汁	旭日升冰茶

表1-1 不同品牌饮料市场占有率调查的原始数据

频数分布表如表 1－2 所示。

表 1－2　频数分布表

品牌名称	频数	频率	百分比频数
百事可乐	9	0.18	18%
汇源果汁	6	0.12	12%
可口可乐	15	0.3	30%
露露	9	0.18	18%
旭日升冰茶	11	0.22	22%

对于可排序的分类数据，我们还可以作累积频数表(cumulative frequency table)。

累积方法有两种：一种是从类别顺序的开始一方向最后一方累积频数，称为“向上累积”；另一种是从类别顺序的最后一方向开始一方累计频数，称为“向下累积”。

累积频数(cumulative frequency)。各类别频数的逐级累加。

累积频率(cumulative percentage)。各类别频率(百分比)的逐级累加。

例 1－2　在一项城市物价问题的研究中，研究人员在甲城市抽样调查 300 户，其中的一个问题是：“您对目前的物价水平是否满意?”

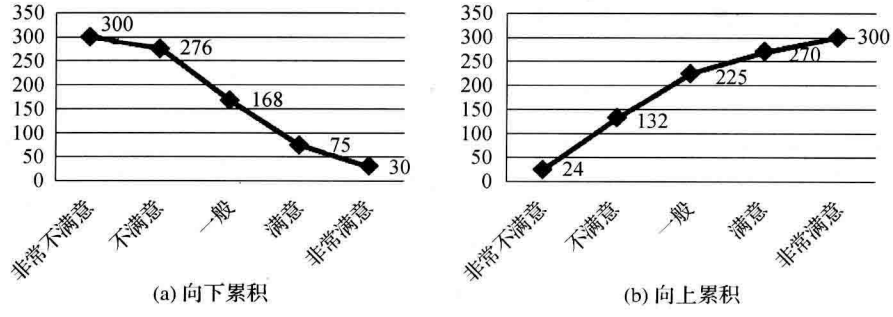
答案：1. 非常不满意；2. 不满意；3. 一般；4. 满意；5. 非常满意。

其频数分布表如表 1－3 所示。

表 1－3　甲城市居民对物价水平评价的频数分布

回答类别	频数	百分比频数%	向上累积		向下累积	
			累积频数	累积百分比%	累积频数	累积百分比%
非常不满意	24	8	24	8	300	100
不满意	108	36	132	44	276	92
一般	93	31	225	75	168	56
满意	45	15	270	90	75	25
非常满意	30	10	300	100	30	10
总计	300	100	—	—	—	—

累积频数分布图如图 1－1 所示。



(a) 向下累积　　(b) 向上累积

图 1－1　累积频数分布图

1.2.2 单变量分类数据的图形展示

统计数据想获得更生动的展示，可以使用统计图。利用统计图表现统计数据，能更加鲜明、一目了然、形象具体地显示现象之间的相互关系。分类数据的展示主要有点图(dotplot)、条形图(bar chart)和饼图(pie chart)，点图和条形图本质上是相同的。

[**注意**] 所有分布图的横轴是变量的取值，纵轴是该取值的频数或频率。

1. 点图

点图使用点表示频数，某类别的频数是多少就画多少点，同一类别的点画在一列上。

例 1-3 下图是将例 1-1 中饮料种类的频数分布表画成点图，如图 1-2 所示。

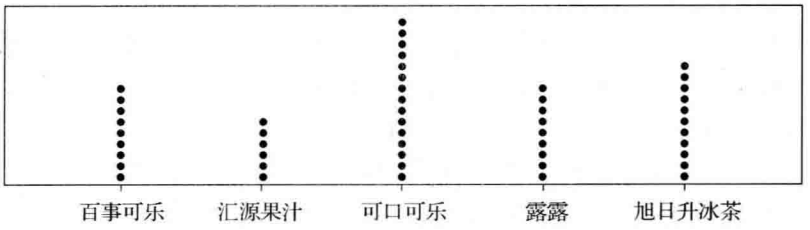


图 1-2 饮料种类分布的点图

每一个类别的频数用该类别点的个数表示，频数越大，点的数量越多，点列也就越高。所以和条形图类似。

2. 条形图

条形图是用宽度相同的柱子的高度或长短来表示各类别数据的图形

例 1-4 下面是将例 1-1 中饮料种类的频数分布表画成条形图，如图 1-3 所示。

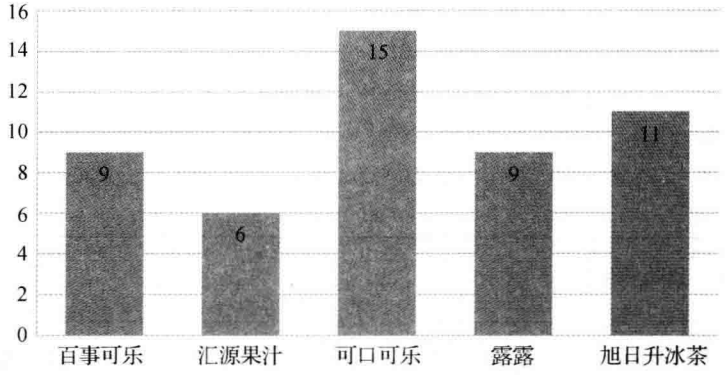


图 1-3 饮料种类分布的条形图

当条形图的纵轴是相对频数时，我们可以作相对频数条形图(relative frequency bar chart)，图 1-4 是上述饮料种类的相对频数条形图。

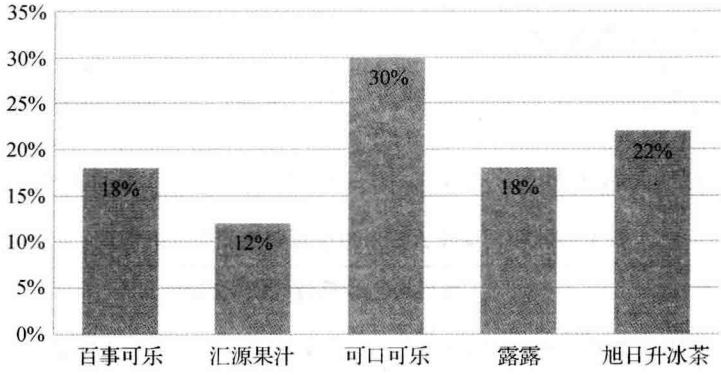


图 1－4　百分比表示的条形图

3. 帕累托图

帕累托图就是把普通条形图中每个变量的取值出现的频数从左到右、从大到小排列。图 1－5 是上述饮料种类的帕累托图。

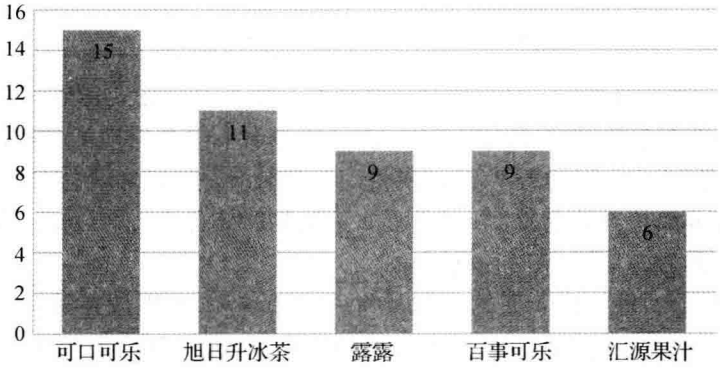
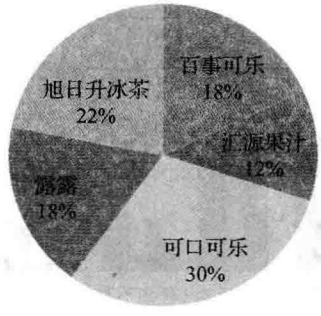


图 1－5　帕累托图

4. 饼图

饼图也称圆形图，是用圆形及圆内扇形的角度来表示数值大小的图形，主要用于表示样本或总体中各组成部分所占的比例，用于研究结构性问题。

例 1－5　图 1－6 是将例 1－1 中饮料种类的频数分布表画成饼图。



图－6　饮料的频数分布的饼图

1.2.3 双变量分类数据的频数分布

双变量分类数据的频数分布常常表现为一张二维表(two-way table),我们把它叫做列联表(contingency table)。

例 1-6 不同城市的女性对新款夏装接受态度的调查数据表如表 1-4 所示。

表 1-4 不同城市女性对新款夏装接受态度的调查

城市	女性对新款夏装的接受态度						行总和
	非常喜欢	有些喜欢	既不反对也不喜欢	有些不喜欢	完全不喜欢	不知道	
南京	52	58	25	12	3	1	151
上海	35	48	40	21	9	2	155
苏州	96	28	13	7	10	0	154
杭州	21	41	50	23	18	3	156
北京	31	48	45	19	10	3	156
列总和	235	223	173	82	50	9	772

列联表(contingency table)是由两个或两个以上变量进行交叉分类得到的频数分布表。列联表中间的各个变量不同水平的交汇处,就是这种水平组合出现的频数或计数(count)。比如上表中的"南京"这一行的数字 52,表明有 52 名南京女性顾客非常喜欢新夏装的款式。构成列联表的变量都是分类变量。一个 r 行 c 列的列联表称为 $r\times c$ 列联表,一般的把 2×2 的二维列联表又称为交叉表(cross table)。

例 1-6 还展示了每一行人数的总和及每一列的总和,分别放在最后一列和最后一行中。实际上,最后一行就是对新款夏装态度的频数分布,最后一列反映的是变量"城市"的频数分布。在统计上就把列联表的这两部分数据称为对应变量的边缘分布(marginal distribution)。

1.2.4 双变量分类数据的图形展示——对比条形图(Double bar chart)

根据例 1-6,画出对比条形图如图 1-7 所示。

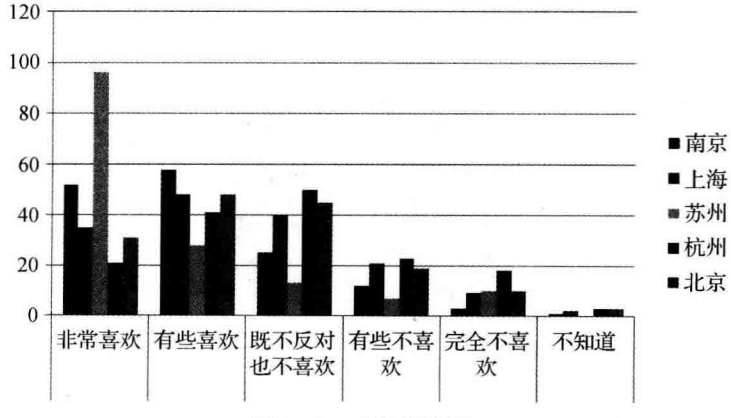


图 1-7 对比条形图 1

从这个对比条形图中我们可以知道不同态度的消费者在各个城市中的对比，也可以画另一个方向的对比条形图，看相同城市的不同态度消费者数量的对比。如图 1－8 所示。

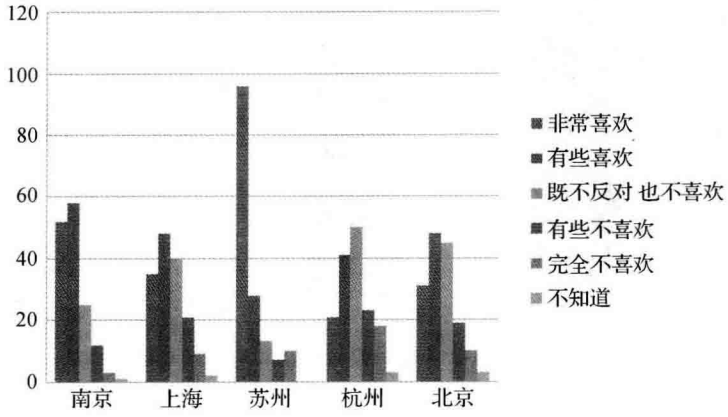


图 1－8　对比条形图 2

进一步地，我们看“非常喜欢”这一列，可以看到苏州喜欢人数为 96 人，杭州喜欢人数仅 21 人。我们能否认为相对于苏州的顾客，杭州顾客不太喜欢新夏装的款式呢？对这类问题的回答可以采用百分数。

我们主要看列联表的两种百分数：一是用列联表中的数据除以最后一列的数据得到行百分数(row percent)。例如，在南京顾客中，“非常喜欢”的比例(行百分数)为 52/151＝34.44%；二是用列联表中的数据除以最后一行的数据得到列百分数(column percent)，例如，在“非常喜欢”的顾客中，南京顾客的占比(列百分数)为 52/235＝22.13%。

下面主要对比南京和上海这两个城市，具体如表 1－5 所示。

表 1－5　南京和上海两城市女性对新款夏装的态度调查

城市	态度						合计
	非常喜欢	有些喜欢	既不反对也不喜欢	有些不喜欢	完全不喜欢	不知道	
南京	34.44%	38.41%	16.56%	7.95%	1.99%	0.66%	100.00%
上海	22.58%	30.97%	25.81%	13.55%	5.81%	1.29%	100.00%

当我们把一个变量某一取值的相对频数限制在另一变量某一取值的条件下后，则称之为条件相对频数(conditional relative frequency)。如在所有南京的女性顾客中，“非常喜欢”的占 34.44%，这里的条件就是南京，然后再关心南京范围内喜欢的比例有多少。这时，可以用条件相对频数表示的对比条形图(double bar charts)，如图 1－9 所示。

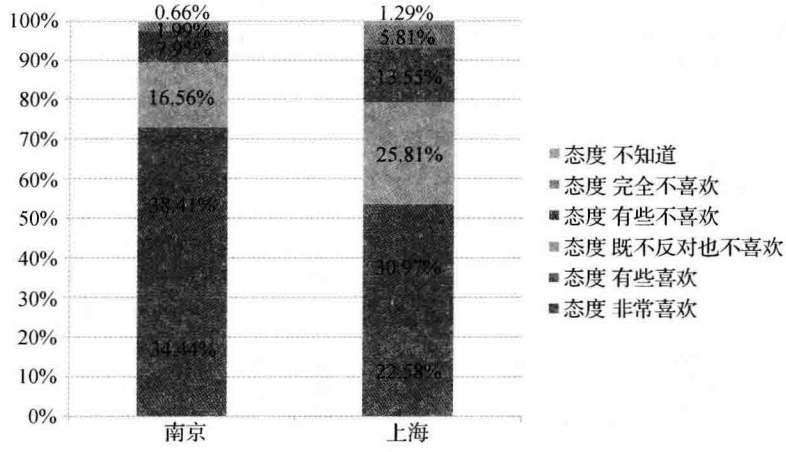


图 1-9　以条件相对频数表示的对比条形图

1.2.5　列联表中两个变量间的关系

列联表中两个变量间的关系分为独立(independent)和不独立(not independent)(即相关)两种。

如果一个变量的分布对于另一个变量的所有取值保持不变,则可以说这两个变量是独立的,表明这些变量之间没有关联。

例 1-7　世纪海难泰坦尼克号沉没事件中乘客性别和生还与否的列联表如表 1-6 所示。

表 1-6　性别与生还的列联表

性别	生还与否	
	否	是
男性	1 364	367
女性	126	344

如果假设这两个分类变量(性别和是否生还)没有关系、相互独立,那么二维表中的每一格对应的期望数量(expected count)是多少呢?

如果性别是否生还之间没有关系,那么

$$P(\text{男性},\text{生还})=P(\text{男性})\times P(\text{生还})=\frac{1\,731}{2\,201}\times\frac{711}{2\,201}$$

这个关系对所有的单元格都成立。男性并且生还的期望数量应该是

$$2\,201\times P(\text{男性},\text{生还})=\frac{1\,731\times 711}{2\,201}=559.17$$

所以,

$$\text{期望数量}=\frac{(\text{行总和})\times(\text{列总和})}{\text{测量的总数}}$$

而男性并且生还的实际观测值为 367 人,说明性别和是否生还之间是有关系的。我们将在后面的章节中对这个问题作更进一步的讨论。

1.2.6　辛普森悖论

辛普森悖论是指当人们尝试探究两种变量是否具有相关性时，比如新生录取率与性别、报酬与性别等，会对之进行分组研究。这种研究中，在某些前提下会产生那些在分组比较中都占优势的一方，而在总评中则是失势的一方。

为什么分组比较中都占优势的一方，在总评中反而是失势的一方呢？这是因为存在潜在变量(lurking variable)，而分组就是按潜在变量的取值来分的。

例 1-8　打 100 场网球比赛以总胜率评价水平高低，选手 F 专找高手挑战 30 场胜 1 场，另外 70 场找平手挑战胜 35 场，胜率为 36%；选手 P 专找高手挑战 70 场胜 7 场，另外 30 场找平手挑战胜 20 场，胜率为 27%，比 36%低很多。所以表面看起来选手 F 的水平比选手 P 要高很多，但仔细观察挑战对象，选手 P 明显较有实力。具体分析如下：

用表 1-7 展示总胜率。

表 1-7　总胜率

比赛结果	选手 F	选手 P
胜	36	27
败	64	73

但按照对手的水平来分组分析，其结果如表 1-8、表 1-9 所示。

表 1-8　对手是高手

比赛结果	选手 F	选手 P
胜	1	7
败	29	63

表 1-9　对手是平手

比赛结果	选手 F	选手 P
胜	35	20
败	35	10

在对手是高手时，选手 F 的胜率为 3.3%，而选手 P 的胜率为 10%；在对手是平手时，选手 F 的胜率为 50%，选手 P 的胜率为 67%。所以无论在哪种情况下，选手 F 都比选手 P 差。这和总评是完全相反的。

总评的表隐藏了潜在变量的影响。这里的潜在变量就是对手的实力，取值为“高手”和“平手”。

辛普森悖论告诉我们：“量”(quantity)和“质”(quality)是不等价的。但实际生活中量要比质更容易测量，所以人们总是习惯用量来评定好坏。例如，用医生治疗病人的总存活率来衡量医生医术水平的高低，但这里存在潜在变量，那就是病人的病情。

【课后练习】

1. 下面的列联表(表 1-10)是根据一个城市的居民受教育水平(以获得了大学以上文凭和没有获得大学文凭分类)和就业状况(以全职和非全职分类)做出的。

表 1-10 列联表

受教育水平	全职工作	非全职工作	总计
获得大学以上文凭	56	45	101
没有获得大学文凭	28	30	58
总计	84	75	159

假定受教育水平和就业状况之间没有关系，那么下列选项是获得了大专以上文凭并且全职工作的期望值？（ ）

A. $\frac{101\times56}{159}$　　B. $\frac{101\times84}{159}$　　C. $\frac{56\times84}{101}$　　D. $\frac{58\times56}{101}$

2. 每一个成人都有他最喜爱的颜色，表 1-11 展示了一次实验中每个人喜欢颜色的情况按年龄分组的试验结果。

表 1-11 试验结果

年龄	红	黄	蓝	其他	总计
40 岁以下	18	44	18	33	113
40～60 岁	22	31	26	27	106
60 岁以上	27	20	31	13	91
总计	67	95	75	73	310

如果对于颜色的偏好和属于哪个年龄组相互独立，下列选项是年龄在 40～60 岁并且喜爱黄色人数的期望值的是（ ）

A. $\frac{95\times106}{310}$　　B. $\frac{67\times113}{310}$　　C. $\frac{95\times31}{106}$　　D. $\frac{31\times106}{310}$