第1章

总论

国家教育部把统计学列为经济类、管理类大学本科学生的专业(核心)基础课。为什么要学习统计学？什么是统计？如何进行统计？在这一章中，我们将学习一些统计学的基本问题。

归纳推断出现象总体的数量特征。例如调查万分之一的城市居民户的收入水平，推断出城市全部居民户的收入水平；调查1‰的农田的收获量，推断出上万亩农田的收获量等等。

1.3 统计学中的基本概念

§1.3.1 总体、个体与样本

总体是在一定的研究目的下，所要研究事物的全体，它是由客观存在的、具有某种共同性质的众多个别事物构成的整体。①

构成总体的个别事物是个体或总体单位。个体是所要研究具体问题的承担者。在统计调查中，常常称总体为调查对象，称个体为调查单位。

样本是从总体中抽取的一部分个体的集合，构成样本的个体的数目称为样本容量。从总体中随机抽取一部分个体作为样本，目的是要根据样本提供的有关信息去推断总体的特征。

比如，了解某校学生的学习情况，学习情况具体体现在学生身上，所以全校所有的学生是总体，每一个学生是个体，从全校所有的学生中随机抽取400名学生就构成了一个样本，通过400名学生的学习情况如平均成绩、及格率等，可以推断全校学生的学习情况；若要研究某市的工业生产情况，工业生产情况具体体现在工业企业身上，该市每一个工业企业是个体，所有的工业企业是总体，从中抽取的若干个工业企业构成一个样本，通过样本工业企业的产值、利润、上缴税金、劳动生产率等，可以推算全市工业产值、利润、上缴税金、劳动生产率等；若要研究某市的工业生产设备情况，工业生产设备情况具体体现在设备上，所以每一台工业生产设备是个体，该市所有的工业生产设备是总体，从中抽取的部分工业生产设备是样本，通过这些设备的净值、生产能力等，可以推算全市所有工业生产设备的净值、生产能力等。

在这些例子中，“学习情况”、“工业生产情况”、“工业生产设备情况”是研究目的；某个学校的学籍、进行工业生产、用于工业生产的设备分别是这些学生、工业企业、工业生产设备的“共同性质”；若干名学生、若干个工业企业、很多很多的工业生产设备分别是“众多个别事物”。

总体具有以下特点：

总体具有同质性。这是指构成总体的总体单位在某一方面性质是相同的，只有性质

① 本章的“总体”指实物总体，在推断统计学中，称随机变量为总体（见第6章）。

第3章

统计指标

统计数据经过加工整理形成数列后，我们对它的分布规律已有了一个直观的了解。然而，要进一步挖掘统计数据，作更深入的分析，仅靠直观了解是远远不够的，还需要寻找一些能充分度量统计数据特征的统计指标，对现象进行分析研究。对统计数据的度量包括：总量水平度量、比较关系度量、集中趋势和离中趋势的度量、分布形态的度量等。

式是：

$$\text{算术平均数} = \frac{\text{总体标志总量}}{\text{总体单位总量}} \tag{3.10}$$

根据掌握的变量值情况，算术平均数的计算方法分为简单算术平均法和加权算术平均法。

(1)简单算术平均法

简单算术平均法是在统计数据未分组的情况下，将各个数据直接相加除以数据的个数计算平均数的方法。这样计算的平均数称为简单算术平均数。若以 $x_1, x_2, \cdots, x_n$ 表示变量值，$\bar{x}$ 表示平均数，则简单算术平均数的计算公式[①]为：

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n} \tag{3.11}$$

如某小组 8 个学生的英语考试成绩分别为 80、72、84、88、75、73、90、78。则该小组成绩的均值为：

$$\bar{x} = \frac{80 + 72 + 84 + 88 + 75 + 73 + 90 + 78}{8} = \frac{640}{8} = 80\ (\text{分})$$

(2)加权算术平均法

根据变量数列计算算术平均数，要用加权算术平均法：用次数对变量值加权求平均数的方法。用加权算术平均法计算的平均数称为加权算术平均数。所谓“加权”是指变量数列中，各个变量值出现的次数不一样，次数出现多的变量值对平均数的影响大一些，次数出现少的变量值对平均数的影响小一些，对各个变量值不能等同看待。计算平均数时，必须以变量值出现的次数与变量值相乘，以权衡其轻重，这就是“加权”。变量值出现的次数或比重称为“权数”。

若用 x_i 表示变量值，f_i 表示变量值 x_i 出现的次数，k 表示组数，则加权算术平均法的计算公式为：

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum xf}{\sum f} \tag{3.12}$$

在社会经济统计中经常用比重权数加权：

$$\bar{x} = x_1 \frac{f_1}{\sum f_1} + x_2 \frac{f_2}{\sum f_2} + \cdots + x_k \frac{f_k}{\sum f_k} = \sum x \frac{f}{\sum f} \tag{3.13}$$

【例 3-1】某企业某日工人日产量资料如表 3.1 所示。试计算工人日平均产量。

① 去掉公式中的下标意义也很明确，本章以下的公式均省略下标。

或 $\bar{x} = \sum x \frac{f}{\sum f} = 55 \times 0.0357 + 65 \times 0.2679 + 75 \times 0.3571 + 85 \times 0.2679 + 95 \times 0.0714$

$= 75.71$(分)

权数起权衡轻重的作用就体现在各组单位数占总体比重的大小上面。哪一组单位数所占比重大,其变量值对算术平均数的影响就大。比重权数更清楚地说明了权数的实质。

(3)调和平均法

计算算术平均数,有时只掌握了各组变量值(x)和各组变量值之和(xf)的资料,为了符合基本公式,应该首先经过除法运算把分母数据(f)求得,再计算平均数。这样计算平均数的方法称为"调和平均法",得到的平均数称为"调和平均数"。调和平均数的计算公式为:

$$\bar{x} = \frac{\sum xf}{\sum \frac{1}{x}xf} \tag{3.14}$$

【例3-3】根据表3.3资料栏数据计算工人平均日产量。

表3.3 某企业某日工人日产量

日产量(件)x	日总产量(件)xf	工人人数(人)$f=xf/f$
10	700	70
11	1100	100
12	4560	380
13	1950	150
14	1400	100
合 计	9170	800

资料栏　　计算栏

$$\bar{x} = \frac{700 + 1100 + 4560 + 195 + 1400}{\frac{700}{10} + \frac{1100}{11} + \frac{4560}{12} + \frac{1950}{13} + \frac{1400}{14}} = \frac{970}{800} = 12.1375(\text{件})$$

计算表明:用调和平均法计算的工人日均产量与用加权算术平均法的一致,只是计算形式不一样。在社会经济现象中,调和平均法往往是加权算术平均法的变形。

如果计算相对数的平均数,则应符合所要求的相对数本身的公式,将分子视为总体标志总量,分母视为总体单位总量,视掌握资料的情况采用算术平均法或调和平均法。

【例3-4】某工业公司18个工业企业产值计划完成程度分组资料如表3.4所示,试计算18个工业企业产值平均计划完成程度。

C	D
列1	
平均	75.23636364
标准误差	1.548533277
中位数	76
众数	76
标准差	11.48423015
方差	131.8875421
峰度	2.501493179
偏度	-0.871626358
区域	67
最小值	32
最大值	99
求和	4138
观测数	55
最大(1)	99
最小(1)	32
置信度(95.0%)	3.104622274

图 3.5　EXCEL 的描述统计结果

在这里 EXCEL 把数据作为样本进行处理。标准差和方差分别指样本标准差和方差，标准误差是指抽样平均误差①。

若要根据未分组数据计算总体的平均差、标准差和方差，可以利用 EXCEL 的函数功能来实现，它们的函数名分别为 AVEDEV，STDEVP 和 VARP。如图 3.4 的数据，可用“=STDEVP(A1:A55)”即可得到总体标准差为 11.379，用“=VARP (A1:A55)”即可得到总体方差为 129.490。

3.6　箱形图

§3.6.1　箱形图的绘制

箱形图(boxplot，也称为箱线图、盒式图、盒须图)，是用最小值、第一个四分位数、中位数、第三个四分位数与最大值五个统计量来描述数据分布的一种常用方法。通过盒式图可以粗略地看出数据是否具有有对称性，分布的分散程度等信息，特别可以用于对几个现象的比较。

绘制箱线图的步骤是：①计算上四分位数 Q_1，中位数 Me，下四分位数 Q_3；②在纵轴上设置适当的尺度，以 Q_1 为底、Q_3 为顶绘制一个矩形箱体(也可以在横轴上绘制箱体)；③在矩形箱体内用一条线代表 Me；④延长箱体上端至最大值、延长箱体下端至最小值作横线，但是，这两条横线与 Q_1 或 Q_3 的距离不大于箱体长度 1.5 倍。箱线图如图 3.6 所示。在 SPSS，SigmaPlot，R，SPlus，Origin 等软件中，有绘制箱线图的程序。

① 抽样平均误差及置信度含义，见第 7 章“参数估计”。

就会产生偏高(或偏低)的滞后偏差,即预测值的变化滞后于实际趋势值的变化。移动平均的项数 k 越大,滞后偏差就越大。

利用 EXCEL 计算移动平均序列的方法是:点击“数据”(2003 版工具点击“工具”)菜单下的“数据分析”,选择“移动平均”,在随即弹出的对话框中指定数据所在区域、间隔(即移动平均的项数)和输出区域的起点单元格后确定即显示输出结果。图 4.2 是利用 EXCEL 对表 4.4 中的用电量计算的 12 月移动平均的对话框和图表输出结果。

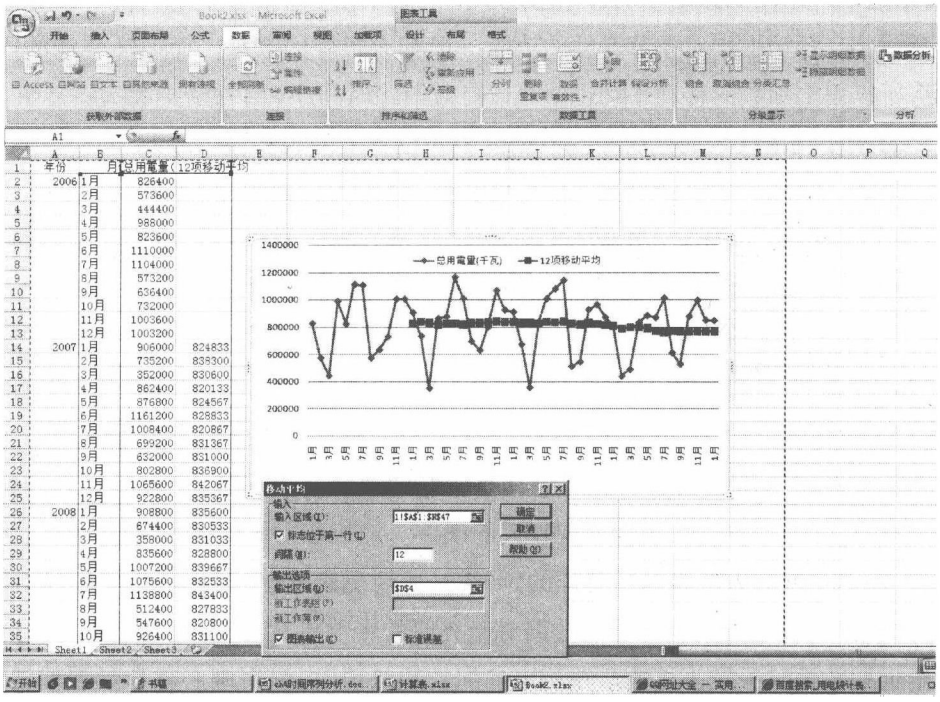

图 4.2 EXCEL 中移动平均的对话框

2. 指数平滑法

指数平滑法是在加权移动平均法基础上改进而来的一种广泛使用的统计分析方法。它通过计算一系列指数平滑值来消除不规则变动,以反映时间序列的长期趋势。[①] 指数平滑法既是对时间序列进行修匀的一种方法,也可以直接用于预测,还可以用于估计预测模型的参数。

用 E_t 表示第 t 期的指数平滑值,其计算公式为:

$$E_t = \alpha y_t + (1-\alpha) E_{t-1} \tag{4.25}$$

式(4.25)中 E_t 和 E_{t-1} 分别表示第 t 期和第 $t-1$ 期的指数平滑值,y_t 为第 t 期的观测

① 指数平滑有一次指数平滑、二次指数平滑和多次指数平滑之分。本教材只介绍一次指数平滑法。

$t+1$ 期的预测值等于上期预测值加用 α 调整后的上期预测误差。这个公式体现了一次指数平滑预测的基本思想：如果第 t 期的预测没有误差，则第 t 期预测值仍然是第 $t+1$ 期的预测值；如果有预测误差，则认为这种误差不外乎包括两部分：一部分是随机波动所引起的误差，预测时应尽可能予以剔除；另一部分是由于 t 期的现象与以前比较确实有了实质性变化而造成的误差，对此需及时跟踪反应，这就要求根据预测误差调整预测值。α 值实质上就体现了预测者对预测误差中实质性变化所占比重的一种估计。

指数平滑值的计算可利用 EXCEL 的“数据分析”宏。方法是：进入 EXCEL 工作簿的“数据分析”→“指数平滑”→根据对话框提示输入原始数据区域、阻尼系数（注：这里的阻尼系数指 $1-\alpha$）、输出区域等即可（如表 4.4(5)、(6)、(7)列）。但指数平滑值的位置是按 $t-1$ 期的平滑值作为 t 期趋势估计值处理的。

3. 趋势方程拟合法

趋势方程拟合法，是通过拟合以时间 t 为解释变量、所考察指标为被解释变量的回归方程来测定现象的长期趋势。此回归方程也称为趋势方程。

客观现象的实际变化是复杂多样的，有的呈现直线趋势，有的呈现曲线趋势，有的呈现混合形式①。对这些趋势形态各异的各种序列，要准确判断其发展变化的趋势规律，选择出准确的函数形式确非易事，实际操作中有几种作法可供参考：

(1)定性分析。即根据经济理论、经济常识和现象的客观性质判断该现象在一般情况下遵循什么规律发展，从定性角度选择拟合的曲线。如人口增长、耐用消费品销售量等通常选择 S 曲线进行拟合。

(2)图形分析。绘制观测值散点图或时间序列折线图。这些图形常能很直观地表现出序列的趋势类型，配合定性分析，一般能对拟合曲线做出选择。这是最常用的有效方法。

(3)数据特征分析。如果时间序列的逐期增长量（一次差）大致相同，可配合线性方程；推而广之，时间序列的 k 次差大致相同，可配合 k 次曲线；当现象的逐期发展速度或增长速度大体相同时，可配合指数方程。

(4)分段拟合。对混合趋势形式的序列，可采取分段拟合的方法，分别考察各阶段的趋势变化。但若要对未来的趋势发展做出预测，只能根据最后一阶段的趋势方程进行外推预测。

线性趋势方程如下：

$$\hat{y}_t = a + bt \tag{4.28}$$

① 本教材只介绍线性趋势方程拟合法。

根据各月的季节指数可以绘制出季节指数图(见图4.6)。从表4.5和图4.6可以看出,某大学用电量的旺季是夏季和冬季,其中用电量最旺的季节是7月,比全年高出31.59%;用电量最淡的季节是3月,只为全年的50.83%。

由于同期平均法计算的季节指数实是相对于平均水平的变化程度,所以,欲对现象的未来发展做出预测,只需要以一个水平趋势值乘以相应的季节比率即可。公式为:

$$\hat{y}_{t+i} = \hat{a} \times \hat{S}_i \quad (i = 1,2,\cdots,L) \tag{4.34}$$

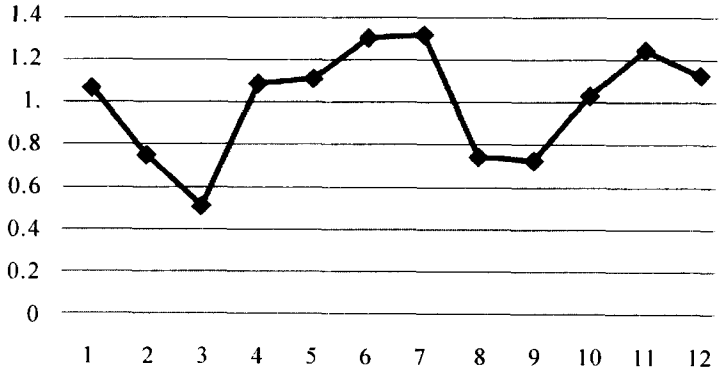

图4.6 某大学用电量季节指数图

若时间序列呈现明显的上升或下降趋势,同期平均法计算的季节指数就不够准确了。比如,当存在上升趋势时,即使完全没有季节变动,按同期平均法计算,年末季节指数也会大于年初季节指数。所以,按同期平均法计算,当现象呈现明显上升趋势时,总会高估年末季节指数,相应地低估年初季节指数;相反,若现象呈现明显的下降趋势,则会高估年初季节指数,相应地低估年末季节指数。为了避免这种局限性,测定季节变动时就应先剔除长期趋势。

2. 测定季节变动的趋势剔除法

该法适合于"趋势季节模式"$Y = T \cdot S \cdot I$。趋势剔除法的步骤是:首先测定出时间序列各期的趋势值,然后从原序列中消除趋势成分,最后再通过平均的方法消除不规则变动,从而测定出季节变动程度。

长期趋势的测定可用移动平均法,也可用趋势方程拟合法,还可以先采用移动平均法修匀时间序列,再采用趋势方程拟合法。但在计算季节指数的过程中,测定长期趋势最简便、最常用的方法是移动平均法。这是因为在长期趋势、季节变动和不规则变动三种因素共存时,若用趋势方程拟合法直接对原序列计算趋势值,会因为季节变动的影响而使趋势值不准确。而移动平均法可以同时消除不规则变动和季节变动的影响,只反映出长期趋势。

移动平均趋势剔除法计算季节指数的具体方法和步骤如下:

(1)计算移动平均值(M)。对原序列计算平均项数等于季节周期L的中心化移动平

其中，I 为个体指数：$I_p = p_1/p_0$；$I_q = q_1/q_0$

采用比重权数的平均法指数，不仅可以避免每次编制指数收集权数资料的困难，而且也便于前后不同时期的对比；但无法根据分子分母之差反映指数所引起的总额变化的绝对差额。

(2)加权调和平均法指数

加权调和平均法指数是采用调和平均形式对个体指数进行加权平均计算总指数的方法。为了使分子、分母有明确的经济意义，权数应为报告期总额。公式为：

$$\bar{I}_q = \frac{\sum q_1 p_1}{\sum \frac{1}{\frac{q_1}{q_0}} q_1 p_1} \tag{5.12}$$

$$\bar{I}_p = \frac{\sum q_1 p_1}{\sum \frac{1}{\frac{p_1}{p_0}} q_1 p_1} \tag{5.13}$$

根据表5.2资料，用加权调和平均法计算的销售量总指数和价格总指数如下：

销售量总指数及销售量变动对销售总额的影响绝对额：

$$\bar{I}_q = \frac{\sum q_1 p_1}{\sum \frac{1}{\frac{q_1}{q_0}} q_1 p_1} \times 100\% = \frac{2300 + 1080 + 162}{\frac{2300}{0.833} + \frac{1080}{1.20} + \frac{162}{1.316}} \times 100\% = \frac{3542}{3783.12} \times 100\%$$

$$= 93.63\%$$

$$\sum q_1 p_1 - \sum \frac{1}{\frac{q_1}{q_0}} q_1 p_1 = 3542 - 3783.12 = -241.12(\text{万元})$$

价格总指数及价格变动对销售总额的影响绝对额：

$$\bar{I}_p = \frac{\sum q_1 p_1}{\sum \frac{1}{\frac{p_1}{p_0}} q_1 p_1} \times 100\% = \frac{2300 + 1080 + 162}{\frac{2300}{1.15} + \frac{1080}{1.125} + \frac{162}{1.08}} \times 100\% = \frac{3542}{3110} \times 100\% = 113.89\%$$

$$\sum q_1 p_1 - \sum \frac{1}{\frac{p_1}{p_0}} q_1 p_1 = 3542 - 3110 = 432(\text{万元})$$

可见，其计算结果与前面用综合法帕氏指数计算的结果完全相同。从计算公式(5.12)和公式(5.13)不难看出，就同一资料，以报告期总额 q_1p_1 为权数的加权调和平均法指数与帕氏综合法指数是一致的。两者只是计算形式不同，经济意义和计算结果完全相同。

§5.4.3 综合评价的常用方法

要将反映客观现象总体不同内容、具有不同表现形式和不同计量单位的众多指标进行综合，使其得到一个反映总体状况的综合评价指标，这只能在不同指标同度量的基础上才可能实现。因此，凡能进行同度量的方法，都可以作为综合评价的方法。这里介绍几种实践中常用的方法。

1. 排队计分法

排队计分法的具体方法如下：将评价单位的各项评价指标依优劣秩序排队，并按如下公式计算各名次单位的具体得分：

$$y_i = 100 - \frac{k-1}{n-1} \times 100 \tag{5.20}$$

式中：y_i 为单项指标的得分；k 为排队名次$(1,2,\cdots,n)$；n 为参加评比的单位数。

对得分一般采用加权算术平均法，将各单位参评的各单项评价指标得分综合为总分。总分数的多少综合说明各单位整体状况的优劣高低，并可据以确定各单位在总体中所处的具体位置。

排队计分法具有以下一些优点：①不必人为寻找比较标准，被评价单位的单项评价值由该单位在总体中的相对位置来确定。②不必事先将指标作同向化处理，确定名次时已考虑了正指标和逆指标的不同。③各单项指标的评价值都有统一的变化范围，即介于(0,100)的区间内，因此，不会出现某一单项评价值过高从而对总评价值影响过大的情况，即评价结果不易受极端值的影响。④对数据的项数多少和分布状况没有严格要求。⑤不仅适用于一般的数值型评价指标，也适用于包含定性指标(只要可以区分出各单位的顺序或等级)的综合评价问题。

排队计分法也有其缺点：它是由指标值在全部评价单位中的位置即名次而不是其数值本身的大小来决定单项评价值。例如，某指标第一名和第二名之间的差异可能远远大于第二名与第三名之间的差异，但由于名次差异相同，体现在评价值上的差异也就相同。

2. 加权指数法

这是利用加权算术平均法指数的形式，对单项评价指标指数加权平均，求得总指数。这里单项评价指标指数不一定以基准期水平作比较的标准，也可以用总体平均水平或规划值或其他某个公认值作为比较标准。它的一般做法是：分别计算各项评价指标值与对比标准值的指数，实现其单项评价指标的无量纲化；用各指标经济重要程度确定的权数作加权平均计算，得综合评价总指数。综合评价总指数公式为：

$$\bar{k} = \frac{\sum \left(\frac{x_i}{X_i}\right) w_i}{\sum w_i} \tag{5.21}$$

第7章

参数估计

在一些实际问题中，研究对象的总体分布类型可以从理论或实际经验得到，但总体参数常常是未知的，需要利用样本提供的信息对未知参数做出估计，其分布函数才能完全确定。由于随机变量的数字特征与它的概率分布中的参数有一定的关系，因而对数字特征的估计也称参数估计。参数估计有两种基本形式——点估计和区间估计。

知时常用的一种点估计法。

介绍极大似然估计的思想之前，先看一个例子。设有甲、乙两个形状相同的盒子，已知两盒各有6个球，在甲盒中装入了4个白色的球和2个黑色的球。乙盒中装入了2个白色的球和4个黑色的球。现任取一只盒子，再从其中有放回地任抽2个球，结果均为白球。我们自然会想到，这两个球来自甲盒的可能性更大。即随机试验的结果更可能来自概率最大的事件，也就是人们常说的概率最大原则。

极大似然估计的基本思想可简单叙述如下：假设总体 X 的分布函数已知，但参数 θ 未知，它可以有很多取值，我们要在参数 θ 的一切可能取值中选出一个使样本观察值出现概率最大的 $\hat{\theta}$ 作为参数 θ 的估计，并称 $\hat{\theta}$ 为 θ 的极大似然估计。在实际计算中，这种概率最大化的是通过求解参数，使得联合概率函数最大化来体现的。极大似然估计的数学表述为：

记总体的概率函数为 $f(x,\theta_1,\theta_2,\cdots,\theta_k)$，$\theta_1,\theta_2,\cdots,\theta_k$ 为未知参数。当 X 为离散型随机变量时，$f(x,\theta_1,\theta_2,\cdots,\theta_k) = P(X = x,\theta_1,\theta_2,\cdots,\theta_k)$；当 X 为连续型随机变量时，$f(x,\theta_1,\theta_2,\cdots,\theta_k)$ 为 $X = x$ 处的密度函数。$X_1,X_2\cdots,X_n$ 为来自总体的随机样本，$x_1,x_2\cdots,x_n$ 为样本观察值，当总体概率函数 $f(x,\theta)$ 的形式已知时，我们把样本的联合密度函数式(7.3)称为似然函数。

$$L = L(x_1,\cdots,x_n;\theta_1,\theta_2,\cdots,\theta_k) = \prod_{i=1}^{n} f(x_i;\theta) \tag{7.3}$$

似然函数常简记为 $L(\theta_1,\theta_2,\cdots,\theta_k)$，它是未知参数 θ 的函数。对于离散型随机变量，似然函数是样本出现的概率；对于连续型随机变量，似然函数是样本的联合密度函数。似然函数的大小取决于未知参数 $\theta_1,\theta_2,\cdots,\theta_k$，若有 $\hat{\theta}_i(x_1,\cdots,x_n)$（其中，$i=1,2,\cdots,k$），使得：

$$L(x_1,\cdots,x_n;_1,_2,\cdots,_k) = \max_{(\theta_1,\theta_2,\cdots,\theta_k)} L(x_1,\cdots,x_n;\theta_1,\theta_2,\cdots,\theta_k) \tag{7.4}$$

其中，称 $\hat{\theta}_i(X_1,\cdots,X_n)$ 为参数 θ_i 的极大似然估计量，$\hat{\theta}_i(x_1,\cdots,x_n)$ 为参数 θ_i 的极大似然估计值。

当似然函数关于参数 θ_i 可导时，求似然函数的极大值可以通过求导来获得。为便于求解最大值，通常将似然函数取对数转化为对数似然函数求解 $\hat{\theta}$[①]。实际用时，可以按照如下步骤进行：

(1)由总体分布推导出样本联合密度函数；

(2)根据联合密度(或概率)函数，写出在样本 $(x_1,\cdots,x_n)$ 处的似然函数 $L(x_1,\cdots,x_n;\theta)$ 或对数似然函数 $\ell nL(x_1,\cdots,x_n;\theta)$；

① 对数化处理在优化求解指数分布族函数参数时尤其有效，能将密度函数的指数部分线性化，提高运算效率。

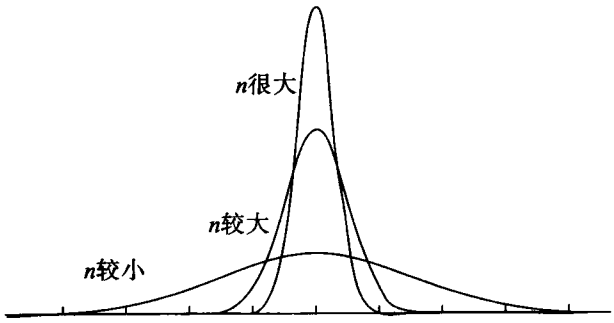

图 7.3　估计量的一致性示意图

【例 7-6】已知 $X \sim N(\mu,\sigma^2)$，$X_1,\cdots,X_n$ 是来自该总体的随机样本，$\frac{1}{n}\sum_{i=1}^{n} X_i$ 是总体的未知参数 μ 的极大似然估计，求证，$\frac{1}{n}\sum_{i=1}^{n} X_i$ 是 μ 的一致估计量。

证明：对于任意小的正数 ε，由切比雪夫不等式有，

$$p\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i-\mu\right|\geqslant\varepsilon\right)\leqslant\frac{D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2}$$

$$\frac{D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\varepsilon^2}=\frac{\sigma^2}{n\varepsilon^2}$$，当 $n\to\infty$ 时，$\frac{\sigma^2}{n\varepsilon^2}\to 0$

因此，$\frac{1}{n}\sum_{i=1}^{n} X_i$ 是 μ 的一致估计量。

进一步由切比雪夫不等式可证，对于 θ 的点估计 $\hat{\theta}_n$，如果 $E(\hat{\theta}_n)=\theta$ 或者 $\lim\limits_{n\to\infty}E(\hat{\theta})=\theta$，并且 $\lim\limits_{n\to\infty}D(\hat{\theta})=0$，则 $\hat{\theta}$ 是 θ 的一致估计量。样本平均数和方差分别是总体平均数及方差的一致估计量。

7.2　区间估计

点估计能够给出明确的估计值大小，但是无法说明估计的精确程度。与点估计不同，区间估计是在一定概率把握程度下，构造待估参数可能的区间范围作为未知参数的估计。

§7.2.1　区间估计的概念

1. 区间估计

设 θ 是总体的一个未知参数，其参数空间为 Θ，$X_1,\cdots,X_n$ 为来自总体的一个样本，对

9.1 相关与回归分析的概述

§9.1.1 相关关系的概念

客观现象之间的相互依存关系是多种多样的，按其数量上是否确定可分为函数关系和相关关系两大类型。

函数关系是指变量之间确定性的数量依存关系，即一个变量 x（或几个变量）取一定数值时，另一个变量 y 总有确定的值与之相对应。例如圆的半径与面积的关系，个人所得与应纳个人所得税的关系，存款额、存款期限和利率与存款利息的关系等等。只就两个变量而言，函数关系可写为 $y=f(x)$，通常将 x 称为自变量，将 y 称为因变量。一般情况下，函数关系可在平面坐标图上表示为一条直线或一条曲线。如图 9.1 是某市出租车乘车里程与乘车费的函数关系：$y=10+1.8x$。

相关关系是指变量之间不确定性的数量依存关系，即指当一个变量 x（或几个变量）取一定数值时，与之相关的某一个变量 y 的值依某种规律在一定范围内变化，不存在一一对应的数量关系。例如，通常企业的广告投入越多，产品销售额就会越多，但是具有相同广告支出的企业，其产品销售额未必都相同，因为企业销售额不仅受广告的影响，同时还受许多其他因素的影响，这些影响因素存在不确定性，甚至有些是无法观察的，所以，企业广告投入与产品销售额的关系就属于相关关系。

如果两个变量呈相关关系，将它们的若干实际观测值标示在平面坐标图上，各观测值点总是在一条直线或一条曲线而上下波动的，如图 9.2 表示的是广告费与销售额的相关关系。

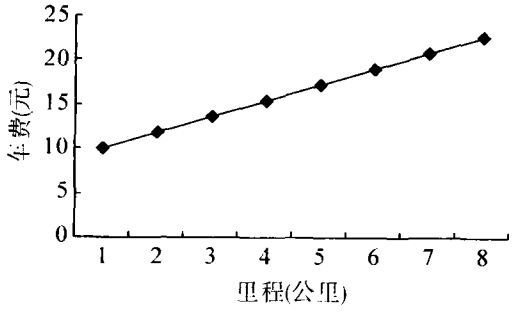

图 9.1 乘车里程与乘车费的函数关系

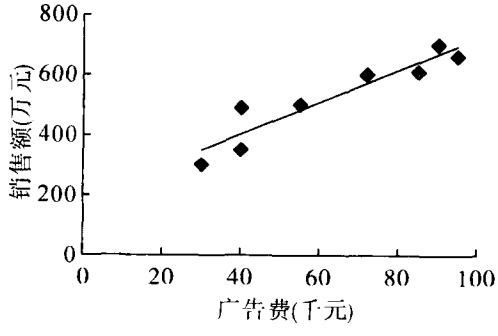

图 9.2 广告费与销售额的相关关系

研究相关关系时，必须区别真相关与假相关。真相关是指现象之间确实存在某种客观的内在联系，不是主观臆造的或者是数据上偶然的巧合。假相关或称伪相关是指原本

$$\begin{cases} b = \dfrac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \\ a = \dfrac{\sum y}{n} - b \times \dfrac{\sum x}{n} = \bar{y} - b\bar{x} \end{cases} \tag{9.12}$$

可证明，当基本假定满足的情况下，a 和 b 分别是 α 和 β 的无偏估计，即：

$E(a) = \alpha$，　$E(b) = \beta$

a 和 b 的方差分别为：

$$D(a) = \sigma^2 \frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}, \quad D(b) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

而且，在回归参数的所有线性无偏估计中，最小二乘估计 a 和 b 的方差最小，即最小二乘估计是最佳线性无偏估计量，这一结论称为高斯—马尔可夫定理。

【例 9－3】根据例 9－1 的数据，试建立企业广告费与销售额之间的回归方程。

解：广告费支出显然是影响销售额的一个重要因素，应该以广告费为自变量 x，以销售额为因变量 y。根据表 9.1 中的计算结果，由式(9.12)可得：

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{24\,041.25}{4527.875} = 5.310$$

或：$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{8 \times 290\,850 - 507 \times 4210}{8 \times 36\,659 - 507^2} = \frac{192\,330}{36\,223} = 5.310$$

$$a = \frac{\sum y}{n} - b \times \frac{\sum x}{n} = \frac{4210}{8} - 5.310 \times \frac{507}{8} = 189.754$$

所求的回归方程为：$= 189.754 + 5.310x$

上述回归方程表明，如果没有广告投入($x=0$ 时)，销售额只有 189.754(十万元)；广告费每增加 1 万元，企业的销售额将平均增加 5.31(十万元)。

3. 运用 EXCEL 估计回归方程

在“工具”菜单的“数据分析”中选择“回归”，点“确定”后出现一个回归对话框(见图 9.5)，在“Y 值输入区域”一栏输入因变量观测数据的起止单元格(在本例中为 C1：C9)，在“X 值输入区域”中一栏输入自变量观测数据的起止单元格(在本例中为 B1：B9)，点击“标志”(因为这里输入区域的第一行是变量名，如果输入区域只有观测值，就不选此项)，在“输出区域”一栏指定显示输出结果的单元格起点(本例为 E1)。最后点“确定”，即可得到回归估计结果(见图 9.6)。

5. 判定系数 r^2 表示因变量 y 总的变异中样本回归线能够解释的部分所占比重，其值介于0和1之间，数值越大，表示回归方程的拟合效果越好。回归估计标准误差 S_e 反映因变量的实际观测值与回归估计值之间的平均误差程度，其值越小，表示回归方程的代表性越好。

6. 对一元线性回归方程的显著性检验就是检验总体回归系数是否等于零，此时 t 检验和 F 检验是等价的。

7. 回归方程可用于对因变量进行估计或预测，可作点预测，也可作区间预测。

8. 相关分析和回归分析中的计算与散点图绘制都可以运用 EXCEL 去实现。

思考题与练习题

9-1 举例说明相关关系与函数关系、因果关系的联系与区别。

9-2 相关分析与回归分析有何联系与区别？

9-3 简单线性相关系数与 Spearman 等级相关系数有何异同？

9-4 相关系数、判定系数、回归估计标准误差之间有何联系？

9-5 简述估计回归方程参数的最小平方法的基本原理。

9-6 解释因变量的离差平方和、回归平方和及残差平方和的含义，分析这三者有何意义？

9-7 回归预测应该注意一些什么问题？

9-8 某生产线上的管理人员认为，工人加工产品的速度可能影响到加工产品的质量。于是一天随机抽取了6名工人进行观测，他们的加工速度和优质品率如下表所示：

工人序号	加工速度(件/分钟)	优质品率(%)
1	25	70
2	40	60
3	55	63
4	30	78
5	60	60
6	20	85

(1)试说明工人的加工速度与其产品的优质品率有什么样的相关关系？试对总体相关性进行显著性检验。

(2)根据样本观测值建立一个回归方程，并说明该回归方程的拟合效果，并且检验回归方程的显著性。

9-9 我国西部12个地区2008年农村居民家庭人均纯收入与人均消费支出的数据