第1章

数理统计基础

众所周知，数理统计学是伴随着概率论的发展而发展起来的。数理统计学研究怎样有效地收集、处理、分析、解释带有随机性的数据，以对所考查的问题做出推断或预测，直至为做出一定的决策和行动提供依据与建议。数理统计是应用数学中最重要、最活跃的学科之一。随着计算机技术的发展，数理统计的理论和应用也得到了长足的进展，在科学研究和国民经济的各个领域中发挥着重要作用。目前，数理统计已经涉及金融、经济、生物、工程技术、医学、地质等诸多领域。

本章主要介绍数理统计中的一些基本概念、常用的统计量和抽样分布，以及几个抽样分布定理，为后续内容的学习奠定基础。

1.1 数理统计的基本概念

用数理统计解决实际问题时，首先要确定研究对象和研究目的，其次是科学地收集数据并进行整理，然后对数据进行合理地分析，最后把数据分析的结果应用到实际问题中。掌握数理统计的思想和方法对确定研究对象、提出问题以及如何准确地解释数据分析的结论具有指导意义。

用数理统计方法解决一个实际问题时，一般可分为如下几个步骤：建立数学模型→收集数据→整理数据→进行统计推断→做出预测→决策。需要注意的是，这些环节互相交错，不可分割。各个步骤的内容如下。

(1) 选择和建立模型。模型是指关于所研究总体的某种假定，一般是确定总体分布的类型。另外，建立模型要依据概率的知识、所研究问题的专业知识、以往的经验以及从总体中抽取的样本(数据)。

(2) 收集数据。收集数据的方式一般有 3 种，包括全面观测、抽样观测和安排特定的实验。**全面观测**又称为普查，即对总体中每个个体都加以观测，测定所需要的指标。**抽样观测**又称为抽查，是指从总体中抽取一部分测定其相关的指标值，这方面的研究内容构成数理统计的一个分支学科，称为抽样调查。**安排特定的试验**以收集数据，这些特定的试验要有代表性，并使所得的数据便于分析。这里所包含的数学问题，构成数理统计的另一分支学科——试验设计的内容。

(3) 整理数据。整理数据的目的是把包含在数据中的有用信息提取出来。整理数据通常有两种形式：一种是绘制适当的图表(如散点图等)，以反映隐含在数据中的粗略的规律性或一般趋势；另一种是计算归纳若干数字特征，以总结出样本某些方面的性质，如样本均

值、样本方差等简单描述性统计量。

(4) 统计推断。统计推断是指根据总体模型及从总体中抽出的样本,做出有关总体分布的某种论断。数据的收集和整理是进行统计推断的必要准备。统计推断也是数理统计学的主要任务。

(5) 统计预测。统计预测的对象是随机变量在未来某个时刻所取的值,或设想在某种条件下对该变量进行观测时将取的值。例如,预测一种产品在未来 3 年内的市场销售额;预测某个 10 岁男孩 3 年后的身高、体重等。

(6) 统计决策。统计决策是指依据所做的统计推断或预测,并考虑到行动的后果而制订的一种行动方案。其目的是使损失尽可能小,反过来说,是使收益尽可能大。例如,一个商店要确定今年内某种商品的进货数量,商店的统计工作者根据抽样调查,预测本商店该产品今年销售量为 1000 件。假定每积压一件商品损失 20 元,而少销售一件商品损失 10 元,据此做出关于进货数量的决策。

数理统计的主要内容之一就是通过对数据进行收集和整理后,进行合理分析,进而做出统计推断。首先我们需要讨论如何收集和整理数据,即**随机抽样**。

数据是进行统计推断的依据。区分数据类型是十分重要的,因为不同类型的数据要用不同的方法来处理和分析。数据的类型很多,根据所采用的计量尺度,一般可分为**定性数据**和**定量数据**。**定性数据**又包括**分类数据**和**顺序数据**。**分类数据**表示对事物进行分类的结果,数据代表类别,用文字表述,由分类尺度计量形成。例如,按照天气状况可以把天气分为晴天、多云、降雨等;按照植物类别可以把植物分为花、草、树等。为了便于数据处理,可以用数字表示各个类别,但这些数字只是代码,没有大小关系,也不能用于计算。**顺序数据**也是对事物进行分类的结果,但这些类别是有顺序的,由顺序尺度计量形成。例如,雨可以分为小雨、中雨、大雨、暴雨等;人的受教育程度可以分为小学、中学、大学等。也可以用数字来表示其分类与顺序,但这些数字只起顺序作用,类与类之间的差别不能做运算。**定量数据**又称为数值型数据或数量数据,是使用自然数或度量衡单位对事物进行测量的结果,其结果表示为具体数值。例如,人的年龄、身高、体重等;产品的重量、规格等。数据也可以按照收集方法的不同,分为**观测数据**和**实验数据**。**观测数据**是指通过调查或观测而收集到的数据。例如,有关社会经济现象的统计数据几乎都是观测数据。**实验数据**是指在实验中控制实验对象而收集到的数据。自然科学领域的大多数数据都是实验数据。数据还可按照被描述对象与时间的关系,分为**截(横断)面数据**和**时间序列数据**。**截(横断)面数据**是指在同一时间点上收集到的数据。例如,同一时间全国各地不同的温度、湿度、气压等。**时间序列数据**是指在不同的时间点上收集到的数据。例如,2023 年 7 月 1 日到 31 日南昌市每天的温度、湿度、气压等。当数据同时具有时间序列和截面两个维度时,可称为时间序列—截面数据,简称为**面板数据**,通常以表格形式来表达。

一般可从以下三方面来描述和了解数据的分布特征:一是数据分布的**集中趋势**,反映各数据向中心位置靠拢或聚集的程度;二是数据分布的**离散程度**,反映各数据远离中心数据的程度;三是数据分布的**偏度与峰度**,反映数据分布的形状。

描述数据**集中趋势**的常见方法有**平均数**、**众数**、**中位数**、**分位数**。

定义 1.1.1 设 $x_1,x_2,\cdots,x_n$ 是一组观测数据,则称 $\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$ 为该组数据的**平均**

数，也称为**均值**。

平均数是衡量集中位置最主要的度量值，主要用于定量数据，不适合用于分类数据和顺序数据。

有时得到的数据可能是分组数据，例如调查工资收入时，收入可能位于某个区间范围，如 4000～5000 元。当数据是分组数据时，此时平均数的近似公式为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{m}x_i\nu_i$，其中 m 为分组的组数，x_i 是第 i 组数据的组中值，ν_i 是第 i 组数据的频数，即有 $\sum_{i=1}^{m}\nu_i=n$。

例 1.1.1 为了了解某公司员工的收入情况，对不同岗位的 20 名员工进行了调查，得到其工资收入数据，如表 1-1 所示。

表 1-1 工资收入数据 单位：元

7500	1250	2630	3760	4520	8460	2010	5500	6700	5150
3970	6500	4280	5600	3080	3500	4800	4000	4500	8050

则可计算得这 20 名员工的平均工资为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=4788$ 元。

若得到的工资数据为分组数据，如表 1-2 所示。

表 1-2 员工工资分组数据 单位：元

序　　号	分组区间	组中值	频数(人数)
1	(1000,3000]	2000	3
2	(3000,5000]	4000	9
3	(5000,7000]	6000	5
4	(7000,9000]	8000	3
总计	—	—	20

则平均工资的近似值为 $\bar{x}=4800$ 元。

定义 1.1.2 在观测数据 $x_1,x_2,\cdots,x_n$ 中，出现频数最多或频率最高的数值称为该组数据的**众数**。


Excel 软件实现

众数主要描述分类数据的集中位置，也可用于定性数据或定量数据集中位置的度量。一般情况下，只有在数据量较大时众数才有意义。

众数不受极端数据的影响，具有较强的稳定性。有时，频数最多的数可能不止一个，这时就存在多个众数。如果在数据中有两个众数，则称为**双众数数据**。如果有 3 个或 3 个以上的众数，则称为**多众数数据**。在多众数的情况下，众数对于描述定性数据的平均大小就没有多大意义了。

设 $x_1,x_2,\cdots,x_n$ 是一组观测数据，将该组数据按从小到大的顺序排序，记为 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，则称 $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 为有序数据。显然，$x_{(1)}$ 为该组数据中的最小值，而 $x_{(n)}$ 为最大值。

定义 1.1.3 称有序数据 $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 中间位置的数为数据 $x_1,x_2,\cdots,x_n$ 的**中位数**，记为 $m_{0.5}$，即

$$m_{0.5}=\begin{cases}x_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数}\\ \dfrac{1}{2}\left[x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right], & n\text{ 为偶数}\end{cases}\tag{1.1.1}$$

中位数主要用于描述有序数据的集中位置，也可用于度量定量数据的集中位置，但不适用于分类数据。

中位数将数据分为两部分，一部分数据比中位数大，另一部分数据比中位数小，但每部分包含的数据个数相同。对于对称分布的数据，平均数与中位数比较接近；对于非对称分布的数据，平均数与中位数则有一定的差异。与众数类似，中位数也不受异常值的影响，具有稳定性，是数据分析中重要的数字特征。

定义 1.1.4 对有序数据 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，给定常数 $p(0<p<1)$，称

$$m_p=\begin{cases}x_{([np]+1)}, & np\text{ 不是整数}\\ \dfrac{1}{2}(x_{(np)}+x_{(np+1)}), & np\text{ 是整数}\end{cases}\tag{1.1.2}$$

为数据 $x_1,x_2,\cdots,x_n$ 的 p **分位点**或 p **分位数**，其中$[np]$表示 np 的整数部分。

分位数是一种度量定量数据位置的方法。易见中位数是特殊的分位数，即 0.5 分位数。称 0.25 分位数为下四分位数，记为 m_L；0.75 分位数为上四分位数，记为 m_U。下四分位数、中位数、上四分位数把全部数据四等分，每一份各占数据总量的 25%。

例 1.1.2 为了解高等数学课程某学期的期末考试成绩，在全校随机抽取了 30 名学生，将其考试成绩由小到大排序，如表 1-3 所示。

表 1-3 学生考试成绩 单位：分

35	42	48	51	55	59	60	64	66	68
70	71	71	71	71	74	76	76	78	78
79	80	82	85	88	88	88	90	91	94

试求该组数据的平均数、众数、中位数、0.3 分位数和 0.8 分位数。

解 平均数为 $\bar{x}=\dfrac{1}{30}\sum\limits_{i=1}^{30}x_i=71.633$ 分；众数为 71 分，即 30 名学生的考试成绩中，考试成绩为 71 分的学生最多；中位数为 $\dfrac{1}{2}(x_{(15)}+x_{(16)})=72.5$ 分，说明有一半的学生成绩低于 72.5 分，一半的学生成绩高于 72.5 分；0.3 分位数为 $m_{0.3}=\dfrac{1}{2}(x_{(9)}+x_{(10)})=67$ 分，说明有 30%的学生成绩低于 67 分；0.8 分位数为 $m_{0.8}=\dfrac{1}{2}(x_{(24)}+x_{(25)})=86.5$ 分，说明有 80%的学生成绩低于 86.5 分。

Excel 软件实现

集中趋势的度量值是一组数据水平的一个概括和代表，但它不能反映对数据组的代表程度。对数据组的代表程度取决于该组数据的离散程度。离散程度越大，集中趋势的测量值对数据组的代表性就越差；离散程度越小，其代表性就越好。描述定量数据离散程度的方法通常有**极差**、**方差**、**标准差**、**变异系数**等。

定义 1.1.5 设一组数据为 $x_1,x_2,\cdots,x_n$，按由小到大的顺序排列后的有序数据为 $x_{(1)},x_{(2)},\cdots,x_{(n)}$，称 $R=x_{(n)}-x_{(1)}$ 为该组数据的**极差**。

极差又称**全距**或**范围误差**，反映数据的变异范围和离散幅度，极差越大，反映数据的取值范围越大，其离散程度越大。反之，极差越小，反映数据的取值范围越小，数据越集中。但极差只利用了一组数据两端的信息，未能反映出中间数据的分散情况，因而不能准确地描述出整体数据的分散程度。

定义 1.1.6 对一组观测数据 $x_1,x_2,\cdots,x_n$，称 $s^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2$ 为该组数据的**方差**，称 $s=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$ 为**标准差**。

方差(标准差)是实际应用和理论研究中使用比较广泛的离散程度描述方法，它反映的是每个数据与其平均数相比平均偏离的程度。因此，方差能较好地反映数据的整体离散程度。当数据比较分散时，方差就比较大；当数据分布比较集中时，方差就比较小。标准差与数据的计量单位相同，其实际意义要比方差更好。因此，在实际应用中，更多地使用标准差。

定义 1.1.7 设 $x_1,x_2,\cdots,x_n$ 是一组观测数据，称 $\mathrm{CV}=\dfrac{s}{\bar{x}}\times 100\%$ 为该数据组的**变异系数**，其中，s 为标准差，$\bar{x}$ 为平均值。

变异系数又称**标准差率**，主要用于比较不同数据的离散程度，变异系数大说明数据的离散程度大，变异系数小说明数据的离散程度小。

例 1.1.3 接例 1.1.2，计算数据的极差、方差、标准差和变异系数。

解

$$\text{极差 } R=x_{(30)}-x_{(1)}=59$$

$$\text{方差 } s^2=\frac{1}{29}\sum_{i=1}^{30}(x_i-71.633)^2=221.8954$$

$$\text{标准差 } s=\sqrt{\frac{1}{29}\sum_{i=1}^{30}(x_i-71.633)^2}=14.8962$$

$$\text{变异系数 } \mathrm{CV}=\frac{14.8962}{71.633}\times 100\%=20.8\%$$

集中趋势和离散程度是数据分布的两个重要特征，如果要全面了解数据分布的特点，有时还需要知道数据分布的形状是否对称、偏斜的程度以及分布的扁平程度等，这就要考查数据组的偏度和峰度。

偏度由统计学家皮尔逊于 1895 年首次提出，是度量数据分布对称性的指标。

定义 1.1.8 设一组数据为 $x_1,x_2,\cdots,x_n$，称

$$\mathrm{sk}=\frac{\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^3}{\left[\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{3}{2}}} \tag{1.1.3}$$

为数据组的**偏度**。

峰度由皮尔逊于 1905 年首次提出，是度量数据平峰或尖峰程度的指标。

定义 1.1.9 设一组数据为 $x_1,x_2,\cdots,x_n$，称

$$\mathrm{ku}=\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}-3 \tag{1.1.4}$$

为数据组的**峰度**。

峰度通常是与标准正态分布比较而言的。如果一组数据服从标准正态分布，则峰度等于零；若数据分布比正态分布更平，则峰度小于零；若数据分布比正态分布更尖，则峰度大于零。峰度越大，说明该数据组中的极端值越多。

类似于平均值、中位数和众数，极差、方差、标准差、变异系数、偏度和峰度均可通过相应的函数在 Excel 中实现，此处不再详细介绍。

习题 1.1

(1) 在对大一新生进行体检时，测得某学院 30 名男生的身高数据如表 1-4 所示。

表 1-4　身高数据　　单位：cm

171	169	180	175	165	177	172	169	170	170
165	167	169	168	175	177	180	170	178	177
167	166	175	173	175	168	163	174	169	182

试计算上述数据的众数、中位数、0.8 分位数、变异系数。

(2) 某高校有相距 23.5km 的新、老两个校区，为了分析教师从老校区到新校区乘坐校车所花的时间，现从某一学期中随机抽取 14 趟校车进行检测，得到运行时间(单位：min)为：67.0，71.0，55.0，62.5，80.2，58.7，64.5，77.0，89.5，59.7，70.5，61.2，66.5，65.8。试计算这组数据的均值、方差、标准差、峰度、偏度、下四分位数、中位数。

(3) 工业工程师通常会定期进行“工作量”分析，以确定生产一个单位产品所需要的时间。表 1-5 记录了某个假期放假前后各 30 天工人执行某项任务每天所需的总工时数。

表 1-5　总工时数　　单位：h

放假前	128	119	95	97	124	128	142	98	108	120
	113	109	124	132	97	138	133	136	120	112
	146	128	103	135	114	109	100	111	131	113
放假后	116	118	138	85	105	117	108	119	95	105
	108	119	95	97	124	128	112	98	86	120
	115	115	98	101	111	89	113	115	108	122

试计算放假前和放假后数据的极差、方差、标准差、偏度以及峰度。

1.2　总体与样本、统计量

1.2.1　总体与样本

在数理统计中，把研究对象的全体称为**总体**，构成总体的每个元素称为**个体**。例如，要研

究某大学学生的身高情况，那么该大学的全体学生是该问题的总体，每个学生是一个个体；要研究鄱阳湖中鱼类的重量，那么鄱阳湖中鱼类的全体就是该问题的总体，每条鱼是一个个体。

每个个体有很多特征，例如学生有身高、年龄、体重、学习成绩等；电子元件有规格、耗电量、寿命等。然而我们只关心某些数量的指标值，例如学生的身高、电子元件的寿命等。如果不考虑实际背景，总体就是一些数的集合，这些数有一定的统计规律。若考查的数量指标用 X 表示，则 X 是一个随机变量，X 的可能取值就是总体中的数，研究总体的分布规律实际上就是研究随机变量 X 的分布规律。因此，总体就是一个随机变量。

定义 1.2.1 一个随机变量 X 或其相应的分布函数 $F(x)$ 称为一个**总体**。

如果要对每一个研究对象观测两个或多个数量指标，那么用多维随机向量表示总体，这是多元统计分析研究的对象。

根据总体中所含元素的个数是有限的还是无限的，可以分为**有限总体**和**无限总体**。本书主要讨论无限总体。

对于无限总体，要研究总体的分布规律或某些特征，但又无法一一测得数据（例如，要研究日光灯管的质量，不可能对每个日光灯管进行测定），因此，需要从总体中随机抽取一定数量的个体进行观测，这一过程称为**抽样**。对总体 X 进行 n 次独立重复观测，将 n 次观测结果依次记为 $X_1, X_2, \cdots, X_n$。由于某一次抽样与另外一次抽样所得的观测结果一般取不同的数值，因此重复抽样中每一个 X_i 应该看作一个随机变量。同时，为了使抽取出的 n 个个体能够较准确地反映总体的情况，要求每次抽样必须满足下面两个性质。

(1) 代表性：抽取出的每一个个体 X_i 都能反映总体 X 的特性，即要求每一个个体 X_i 与总体 X 有相同的分布。

(2) 独立性：上一次抽样不影响下一次抽样的结果，即要求 $X_1, X_2, \cdots, X_n$ 相互独立。

上述抽样方法称为**简单随机抽样**，利用简单随机抽样得到的样本，称为**简单随机样本**，简称**样本**。

因为总体中的每一个个体是随机试验的一个观察值，在随机试验中，它取的每一个值或取值区间都有对应的概率，所以总体对应一个随机变量，从而总体是有分布的。例如，元件寿命的分布往往是指数分布。

在后面章节中我们不区分总体与其对应的随机变量，一般都用随机变量表示总体，有时也用其分布作为总体。

下面是简单随机样本的定义。

定义 1.2.2 设总体 X 的分布函数是 $F(x)$（或概率密度函数为 $f(x)$），若随机变量 $X_1, X_2, \cdots, X_n$ 相互独立且每个 $X_i (i=1,2,\cdots,n)$ 与总体 X 有相同的分布，则称随机变量 $X_1, X_2, \cdots, X_n$ 为来自总体 X 的容量为 n 的**简单随机样本**，简称**样本**，它们的观察值 $x_1, x_2, \cdots, x_n$ 称为**样本值**，n 称为**样本容量**，样本 $X_1, X_2, \cdots, X_n$ 也记作 $(X_1, X_2, \cdots, X_n)$。

注意：本书如不作特别说明，所提到的样本均指简单随机样本。

例如，在大一新生中随机抽取 10 个学生，得到他们的身高 $x_1, x_2, \cdots, x_{10}$，就称这一组数据是一个样本，样本容量为 10。

每个容量为 n 的样本都可看作 n 维空间中的一个点，样本所有可能的取值构成了 n 维空间的一个子集，称为**样本空间**。样本作为随机变量具有概率分布，称为**样本分布**，样本分布可由总体分布完全确定。

设总体 X 的分布函数为 $F(x)$，由于 $X_1,X_2,\cdots,X_n$ 相互独立且与总体 X 有相同的分布，于是得到样本 $X_1,X_2,\cdots,X_n$ 的联合分布函数为

$$F(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}F(x_i) \tag{1.2.1}$$

设总体 X 是连续型随机变量，且概率密度函数为 $f(x)$，则样本 $X_1,X_2,\cdots,X_n$ 的联合概率密度函数为

$$f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i) \tag{1.2.2}$$

设总体 X 是离散型随机变量，其分布律为 $p_i=P\{X=x_i\}$，$i=1,2,\cdots$，则样本 X_1，$X_2,\cdots,X_n$ 的联合分布律为

$$p_{i_1i_2\cdots i_n}=P\{X_1=x_{i_1},X_2=x_{i_2},\cdots,X_n=x_{i_n}\}=\prod_{j=1}^{n}P\{X=x_{i_j}\}=\prod_{j=1}^{n}p_{i_j} \tag{1.2.3}$$

式中，$i_j=1,2,\cdots$；$j=1,2,\cdots,n$。

1.2.2 统计量的概念

在统计推断问题中，样本只是原始数据，无法直接得到总体的规律。那么，往往需要对样本进行数学上的“加工”，这样才能有效地利用其中的信息，即构造样本的各种函数。例如，要了解某学期所有学习“概率论与数理统计”课程的学生的期末考试情况，学校随机地抽取 50 位学生，得到一个容量为 50 的样本 $X_1,X_2,\cdots,X_{50}$，计算其样本均值 $\overline{X}=\frac{1}{50}\sum_{i=1}^{50}X_i$，这样可以从样本均值得到所有学生的近似平均成绩；还可以通过计算样本方差 $S^2=\frac{1}{49}\sum_{i=1}^{50}(X_i-\overline{X})^2$ 判断学生的期末成绩差距是否较大。

由此可见，把分散在样本中我们关注的信息提炼出来，针对不同的研究目的构造不同的样本函数，是统计推断的基础；同时，为了使提炼出的信息是已知的，这个函数不能含有未知参数，这样的函数在统计学中称为统计量。

定义 1.2.3 设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，$T(X_1,X_2,\cdots,X_n)$是样本 $X_1,X_2,\cdots,X_n$ 的函数，若函数 $T(X_1,X_2,\cdots,X_n)$中不含任何未知参数，则称 $T(X_1,X_2,\cdots,X_n)$是一个**统计量**。

如果 $x_1,x_2,\cdots,x_n$ 为样本 $X_1,X_2,\cdots,X_n$ 的观测值，则称函数值 $T(x_1,x_2,\cdots,x_n)$为统计量 $T(X_1,X_2,\cdots,X_n)$的一个观测值。

例 1.2.1 设总体 $X\sim N(\mu,\sigma^2)$，$\mu\in\mathbf{R}$ 未知，$\sigma>0$ 已知，$X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，则 $\frac{1}{n}\sum_{i=1}^{n}X_i$，$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$，$\frac{1}{\sigma^2}\sum_{i=1}^{n}X_i$ 都是统计量，但 $\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2$ 和 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$ 含有未知参数 μ，因此都不是统计量。

下面介绍几种常见的统计量。

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的容量为 n 的一个样本。

(1) **样本均值** $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，它反映了总体均值 $E(X)$的信息。

(2) **样本方差** $S^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2$,它反映了样本中各分量相对于其平均值的离散程度。

(3) **样本标准差** $S=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\overline{X})^2}$,也称样本均方差。

(4) **样本 k 阶原点矩** $A_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i^k$,$k=1,2,\cdots$,它反映了总体 k 阶原点矩 $E(X^k)$ 的信息。

(5) **样本的 k 阶中心矩** $B_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}(X_i-\overline{X})^k$,$k=1,2,\cdots$,它反映了总体 k 阶中心矩 $E[(X-E(X))^k]$ 的信息。

设 $x_1,x_2,\cdots,x_n$ 为样本 $X_1,X_2,\cdots,X_n$ 的观测值,则 $\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$,$s^2=\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2$,$s=\sqrt{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}$,$a_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^k(k=1,2,\cdots)$,$b_k=\dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^k$ $(k=1,2,\cdots)$分别称为样本均值、样本方差、样本标准差、样本 k 阶原点矩、样本 k 阶中心矩的观测值。

例 1.2.2 为了了解某种橡胶的性能,随机抽取容量为 20 的样本,测量每个样本的硬度,得到样本值如下:

65,70,70,68,69,66,67,68,72,75,76,65,62,66,67,68,75,72,77,60

分别计算样本均值、样本方差、样本标准差。

解 利用公式,可得样本均值为 $\bar{x}=\dfrac{1}{20}\sum\limits_{i=1}^{20}x_i=68.9$,样本方差为 $s^2=\dfrac{1}{19}\sum\limits_{i=1}^{20}(x_i-\bar{x})^2=20.83$,样本标准差为 $s=\sqrt{\dfrac{1}{19}\sum\limits_{i=1}^{20}(x_i-\bar{x})^2}=4.564$。


Excel 软件实现

习题 1.2

(1) 某制药厂生产了一种感冒药,其不合格品率 p 未知,每 n 件产品包装为一盒。为了检查药物的质量,现随机抽取 m 盒,检测其中不合格的件数。问:在该次检查中,总体、样本的分布各是什么?

(2) 设有两组样本,分别为 $X_1,X_2,\cdots,X_n$ 和 $Y_1,Y_2,\cdots,Y_n$,且具有如下关系:

$$Y_i=b(X_i-a),\quad i=1,2,\cdots,n$$

其中,a,b 均为常数。试求样本均值 $\overline{X}$ 和 $\overline{Y}$ 之间的关系,以及样本方差 S_X^2 和 S_Y^2 之间的关系。

(3) 设 X_1,X_2,X_3 是来自总体 $X\sim\Gamma(1,0.02)$的样本,求:

① 样本 X_1,X_2,X_3 的联合概率密度函数;

② $P\{\min\{X_1,X_2,X_3\}\leqslant 0.6\}$和 $P\{\max\{X_1,X_2,X_3\}\leqslant 10\}$。

(4) PM2.5 是重要的衡量空气污染程度的指标。三个检测员分别在华北、华东、华南

检测了不同数量空气质量检测子站的 PM2.5 数据。抽检的数量 n_i、样本均值 $\bar{x}_i$ 与样本标准差 s_i 如表 1-6 所示。

表 1-6　PM2.5 数据表

地　　区	样本容量 n	样本均值 $\bar{x}_i$	样本标准差 s_i
华北	25	65.78	9.489
华东	32	62.97	12.363
华南	27	58.52	11.552

试求合并后的样本均值和样本方差。

提示：设 $x_1,x_2,\cdots,x_n$ 和 $y_1,y_2,\cdots,y_m$ 是从总体 X 中先后两次取出的样本，其对应的样本均值和样本方差分别为：$\bar{x}$ 和 $\bar{y}$，s_x^2 和 s_y^2。将两次抽取的样本进行合并，得到样本容量为 $n+m$ 的样本，其样本均值和样本方差记为 $\bar{z}$ 和 s_z^2，则有

$$\bar{z}=\frac{n\bar{x}+m\bar{y}}{n+m},\quad s_z^2=\frac{(n-1)s_x^2+(m-1)s_y^2}{m+n-1}+\frac{nm(\bar{x}-\bar{y})^2}{(m+n)(m+n-1)}$$

1.3　抽样分布

1.3.1　抽样分布的定义

统计量的分布称为**抽样分布**。在数理统计中，抽样分布是进行统计推断的重要依据。根据概率论的知识，如果总体 $X\sim N(\mu,\sigma^2)$，$X_1,X_2,\cdots,X_n$ 是来自总体的一个容量为 n 的样本，则 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$；如果总体 $X\sim b(1,p)$，$X_1,X_2,\cdots,X_n$ 是来自总体的一个容量为 n 的样本，则 $T=\sum_{i=1}^{n}X_i\sim b(n,p)$。这都可以看作抽样分布。

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个容量为 n 的样本，对于 $1,2,\cdots,n$ 中的任意一个值 k，按如下方式定义样本 $X_1,X_2,\cdots,X_n$ 的一个函数 $X_{(k)}$：对于 $X_1,X_2,\cdots,X_n$ 的每个观察值 $x_1,x_2,\cdots,x_n$，把 $x_1,x_2,\cdots,x_n$ 按照从小到大的顺序排成一列：$x_{(1)},x_{(2)},\cdots,x_{(n)}$。规定 $X_{(k)}$ 相应的观察值为 $x_{(k)}$。记作

$$X_{(k)}=(X_1,X_2,\cdots,X_n\text{ 中第 }k\text{ 个小的值}),\quad k=1,2,\cdots,n$$

特别地，$X_{(1)}=\min\{X_1,X_2,\cdots,X_n\}$，$X_{(n)}=\max\{X_1,X_2,\cdots,X_n\}$。称 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 为**顺序统计量**，也称各个 $X_{(k)}$ 为顺序统计量。当顺序统计量 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 的值 $x_{(1)},x_{(2)},\cdots,x_{(n)}$ 给定时，对于任意实数 x，定义函数

$$F_n(x)=\begin{cases}0, & x<x_{(1)}\\ \dfrac{k}{n}, & x_{(k)}\leqslant x<x_{(k+1)},k=1,2,\cdots,n-1\\ 1, & x\geqslant x_{(n)}\end{cases}$$

称 $F_n(x)$ 为总体 X 的**经验分布函数**。

定理 1.3.1（格列汶科定理）　设总体 X 的分布函数为 $F(x)$，经验分布函数为 $F_n(x)$，