第一章　统计方法简介

第一节　统计的含义

统计作为一种社会实践活动是随着人类社会经济的发展，适应于国家治理和处置各种社会事务的需要而产生的，是伴随人类文明的进步一同发展起来的。现在已经发展成为一门重要的科学。

“统计”最基本的含义是对客观事物进行数量方面的计量和分析，是对所关心的对象获取量的信息的实践活动。统计本身可能无法直接得到解决问题的方法，但根据实际需要所做的信息的整理与收集却是解决问题、制订计划或预测事件发展规律的有效途径。一般情况下，通过统计方法所获取的信息可以为下一步工作计划的展开提供方法上的指导，有效的统计甚至可能直接成为解决实际问题的工具。作为一门科学，它主要通过客观对象数量上的特征或者表征客观对象特定属性间的数量的相关性来反映事件发展的规律，而这种规律本身可能是诸多非数量特征的客观原因所导致的必然结果。比如，股市的价格浮动可能是银行利息调整、国家经济政策宏观调控、企业经营状况以及有雄厚实力的投资者的人为操纵等原因的直接表现，但单纯任何一个单一的因素都不足以准确描述股市价格变动的规律。由于影响股市因素的多样性和不确定性，考察全部的影响因素显然是不现实的，因此基于影响股价的因素而建立的数学模型，无论多么复杂，都不可能对股市的走向

做出长期准确的预测。现在的经济学家更倾向于用统计的方法来揭示这一规律，原因在于事物数量上的相关特性必然使得其在演变过程中留下数字变化特征的痕迹。统计方法完全抛开各种具体的影响因素而纯粹从事件发展过程中所表现出的数量特征来捕捉其演变规律，简单说来就是从客观对象数量的变化关系中寻找量与量之间的函数关系。用数字的相关性所总结或归纳出来的演变规律其实是包含了影响事件发展的全部因素，从这个意义上讲，统计作为一种方法论或一门科学有着其他物理手段不可替代的作用，其广泛的应用领域和科学严密的数学手段必然使其具有极其深刻而十分丰富的内容。

人们的实践活动是复杂的、具体的，统计作为获取信息的手段，其方法也是多样的。而且在为实现同一目的所作的统计中，统计的内容也各不相同。比如，要获取一个单位的职工数量，最简单的办法就是直接找到该单位的人事部门或其主管部门询问。有时这种方法可能不太现实，我们则可以从对该单位的食堂规模、职工宿舍的建筑面积、停车场的大小以及单位时间内出入单位的平均人数等信息的统计中推断出一个大致的人数范围。这里需要强调的一点就是，如果不是对所考察的对象作直接的统计分析，统计的对象和内容必须要与所考察的对象有数量上的关联，而且这种关联具有事先的确定性。例如，统计了一个学校预订教材的数量就可以大体知道该校的学生人数，因为每个学生只能订一份教材是事先确知的，而且一般情况下所有的学生都必须拥有教材。

由此看来，统计是我们获取信息的最直接的手段，科学的统计方法和全面的统计资料有助于我们对未知事件作出合理的预测，从而有助于我们制定完善的计划或对解决实际问题提出有效的方法。春秋时期，秦国著名政治家商鞅在《商君书·去疆》中强调："强国知十三数，竟内仓、口之数，壮男、壮女之数，老、弱之数，官、士之数，以言说取食者之数，利民之数，马、牛、刍稿之数。欲强国，

不知国十三数，地虽利，民虽众，国愈弱至削。”意思是要实现国家的强盛必须先对国情国力有个基本的统计，并以此来制定正确的治国方针，否则即使地广人丰，国家也会因情况不明而导致决策失误，从而使国家渐趋衰落，甚至灭亡。《孙子兵法》中亦有言，“知彼知己，百战不殆”。其实这个“知”的过程必然是一个统计的过程，是为“战”服务的，卓越的军事家总是善于运用统计的方法以获取用于指导战争的翔实的资料，他们能够将各种统计方法和资料应用得出神入化。这些都是统计资料在预测和决策中所处地位的重要性的具体体现。人们可能比较关心的问题是，统计作为获取信息的工具是否只是解决问题的辅助手段？统计能否成为解决问题的直接的方法？对这个问题的正确回答可能会有助于我们加深对统计这一科学的真正理解。一般情况下，统计活动的结果作为一种信息或资料在人们处理问题的过程中确实是一种重要的辅助，但随着科技的发展，尤其是计算机技术的广泛运用，再加上统计方法本身的不断完善，统计成为一种直接解决实际问题的手段已是不争的事实。统计自产生之日起就赋有服务于人类实践需要的属性，只是由于时代不同和应用水平的局限，其作用的体现有所侧重而已。

我们所处的时代是文明高度发达的时代，是信息化的时代，在这样的历史条件下，统计比以往更具有实际的意义。自然科学的飞速发展和技术的不断进步已经为统计赋予了新的历史使命。可以用马寅初先生的一段话来总结其重要性：“人类社会，日臻繁复，耳目有所未周，则不能无赖于统计焉。盖个人动作，在与社会有关，倘于社会事实，未尽了了，则闭门造车，难期合辙。自然界现象，变化万端，亦非一二人所能穷，则综合统计又为必要。是故学者不能离统计而研学，政治家不能离统计而施政，事业家不能离统计而执业也。”总之，统计作为一门科学，其应用范围已经渗透到自然科学和社会科学的各个领域。其作用和价值可以概括为以下几点：

（1）统计是人们揭示自然规律，认识世界的重要手段；

(2) 统计是制定计划、处理事务的有力保障;

(3) 统计是科学研究的有效工具。我们作为社会的一员,一切工作的进一步展开,都有赖于一定的统计知识,因此学好用好统计,也是我们的必然选择。

第二节　统计学中常用的基本概念和术语

鉴于编写本分册的初衷是介绍统计方法的常用知识和应用技巧,以实现在科技论文的编辑过程中及时准确地辨识一些常见的统计错误,而且由于统计方法在科技论文中的应用具有相对的独立性,以及限于篇幅,本分册不准备过多地讲述概率论方面的相关知识,尽管有些理论是统计学的基础。如果读者对于统计上涉及概率论的某些概念或理论感觉模糊或抽象,可以查阅概率论方面的有关书籍。以下只对统计中经常用到的一些概率中的基本概念和术语作简单的陈述。

一、统计中有关概率方面的基本概念

1. 随机试验

满足如下条件的试验称为随机试验:

(1) 试验可以在相同条件下重复进行;

(2) 每次试验的可能结果不止一个,但所有可能的结果都是事先已知的;

(3) 进行一次试验前不能确定会出现哪个结果,但试验结果必是已知结果中的一个。

2. 样本空间

随机试验的所有可能结果组成的集合称为样本空间,常用 S 来表示。S 中的元素称为样本点。S 中包含的样本点可以是有限个,也可能是无穷多个,或在区间及整个数轴上取值。

3. 随机事件

样本空间的子集称为随机事件，简称事件。它在每次试验中可能发生，也可能不发生，带有一定的随机性。常用大写字母 $A,B,C,\cdots$ 来表示。每次试验中，当且仅当这一子集合中的一个样本点出现，则称这一事件发生。比如，若 $e \in A$，每次试验中，如果样本点 e 出现，则称事件 A 发生；反之，如果事件 A 发生了，则这一子集合中必有某一个样本点 e 出现。以下是随机事件的特殊情形：

基本事件：随机试验 E 的每一个可能结果，称为基本事件，随机事件就是由若干基本事件组成的集合。

必然事件：随机试验中一定发生的事件，称为必然事件。样本空间 S 包含所有的样本点，它是 S 自身的集合，在每次试验中它必然发生，所以称为必然事件。

不可能事件：每次试验中不可能发生的事件，称为不可能事件，常记为 $\varnothing$。它不包含任何样本点，作为样本空间的子集，它在每次试验中都不发生，所以称为不可能事件。

4. 频率与概率

频率：在相同条件下进行 n 次试验，其中事件 A 发生的次数 n_A 称为事件 A 发生的频数，比值 $\frac{n_A}{n}$ 称为事件 A 发生的频率，记为 $f_n(A) = \frac{n_A}{n}$。当 n 较小时，用频率来表达事件发生的可能性的大小是不恰当的。但随着 n 的逐渐增大，频率 $f_n(A)$ 会趋于某个常数 p，这个常数 p 就是度量事件发生可能性大小的概率，记为常数 $P(A) = p$。这是概率的统计定义。

概率的公理化定义：记所有随机事件的集合为 F，即 $F = \{A \mid A \in S; A$ 为随机事件$\}$。对于每个事件 A，取一个实数与之对应，记为 $P(A)$，$P(A)$ 是定义在 F 上的集合函数。如果 $P(A)$ 满足下列条件：① 非负性：$P(A) \geqslant 0$；② 规范性：$P(S) = 1$；③ 可列可加性：设

$A_1, A_2, \cdots, A_n, \cdots$ 是两两互不相容的事件，则有 $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$。则称 $P(A)$ 为事件 A 发生的概率。

5. 随机变量

设随机试验 E 的样本空间为 S，如果对于每一个样本点 $e \in S$，变量 X 都有一个确定的实数值与之对应，则 X 是定义在 S 上的实值函数，即 $X = X(e)$，称这样的变量 X 为随机变量。显然随机变量是样本空间到实数集的单值映射，如果映射的范围为有限个或可列个，则称随机变量是离散型随机变量；若映射的范围为某个实数区间，则称随机变量是非离散型随机变量。

6. 分布

一个随机现象的规律通常通过随机事件及其概率来描述。一个随机试验的所有结局事件与对应的概率的排列称为分布。对应于样本数量值的分布称其为频率分布；对应于总体数量值的分布称为概率分布。对于离散型随机变量，有以下几种重要的概率分布：

(1) 伯努力分布：设随机变量 X 的所有可能的取值为 0 和 1，它的分布为

$$P(X = k) = p^k(1-p)^{1-k},\ k = 0,1,\ 0 < p < 1$$

称 X 服从伯努力分布，又称 $(0,1)$ 分布，简记为 $X \sim B(1,p)$。

(2) 超几何分布：设随机变量 X 的概率分布为

$$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^M},\ \max\{0, n-N+M\} \leqslant k \leqslant \min\{n, M\}$$

其中 n, M, N 都是正整数，$M \leqslant N$，则称随机变量 X 服从超几何分布，简记为 $X \sim H(n,M,N)$，其中 n, M, N 是分布的参数。

(3) 二项分布：设随机变量 X 的概率分布为

$$P(X = k) = C_n^k p^k q^{n-k},\ k = 0,1,2,\cdots,n$$

其中 n 为正整数，$0 < p < 1, p + q = 1$，则称随机变量 X 服从参数

为(n,p)的二项分布,简记为$X \sim B(n,p)$。它满足$P(X=k) \geqslant 0$,$\sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1$。对于二项分布,当$k$增加时,概率$P(X=k)$随之增加,直至达到最大值,然后随着$k$的继续增加,概率单调减少。

(4) 泊松分布:设随机变量X的概率分布为

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0,1,2,\cdots$$

其中$\lambda > 0$,则称随机变量X服从参数为λ的泊松分布,简记为$X \sim P(\lambda)$,它满足$P(X=k) \geqslant 0, \sum_{k=0}^{n} P(X=k) = \sum_{k=0}^{n} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$。

(5) 几何分布:在n重伯努力试验中,每次试验事件A发生的概率都为p,直到事件A发生为止,所进行的试验次数X服从几何分布。X的概率分布为

$$P(X=k) = pq^{k-1}, k = 1,2,\cdots$$

对于非离散型随机变量,由于所有可能的结果不能一一列举,因此不能用分布率来表示。我们所关心的是这种随机变量落在某一范围内的概率,而不是它取某个值的概率。如等车时常关心的是多长时间等到车的概率,而不是某一时刻等到车的概率;考察产品的使用寿命,关心的是产品的寿命大于某个值的概率,而不是等于某值的概率等等。为此,引入随机变量分布函数的概念:设X是一个随机变量,对于任意实数x,称$F(x) = P(X \leqslant x)$为随机变量X的分布函数。

不论随机变量是离散型随机变量或非离散型随机变量,分布函数$F(x)$全面地描述了随机变量X的统计规律性。

7. 概率密度

如果对于随机变量X的分布函数$F(x)$,存在非负函数$f(x)$,使对于任意实数x,有$F(x) = \int_{-\infty}^{x} f(t)\,dt$,则称$X$是连续型随机变

量,其中函数$f(x)$称为X的概率密度函数,简称概率密度。概率密度有以下性质:①$f(x) \geqslant 0$;②$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$;③对于任意实数$x_1, x_2 (x_1 \neq x_2)$,有$p\{x_1 < X \leqslant x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\mathrm{d}x$;④若$f(x)$在点$x$处连续,则有$F'(x) = f(x)$。

对于连续型随机变量X,由定义知,改变概率密度$f(x)$在个别点的值并不影响分布函数$F(x)$的取值。事实上,它取任意一指定实数值a的概率均为0,即$P(X = a) = 0$。因为由分布函数知,$P\{x_0 < X \leqslant x_0 + \Delta x\} = \int_{x_0}^{x_0+\Delta x} f(x)\mathrm{d}x$,当$\Delta x \to 0$时,由拉格朗日中值定理,$\lim\limits_{\Delta x \to 0} P\{x_0 < X \leqslant x_0 + \Delta x\} = 0$。因此,在计算连续型随机变量落在某区间的概率时,可以不必考虑该区间是开区间还是闭区间。注意事件$\{X = a\}$并非不可能事件,但有$P\{X = a\} = 0$。

8. 三种重要的连续型随机变量

(1) 均匀分布:设连续型随机变量X具有概率密度

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \leqslant x \leqslant b \\ 0, & \text{其他} \end{cases}$$

则称X在区间$[a,b]$上服从均匀分布,记为$X \sim U(a,b)$。X的分布函数为

$$F(x) = \begin{cases} 0, & x \leqslant a \\ \dfrac{x-a}{b-a}, & a < x < b \\ 1, & x \geqslant b \end{cases}$$

随机变量X在区间$[a,b]$上服从均匀分布,具有下述意义的等可能性,即它落在区间$[a,b]$中任意等长度的子区间内的可能性是相同的,或者说它落在区间$[a,b]$中子区间内的概率只依赖于子区间的长度而与子区间的位置无关。

(2) 指数分布:设连续型随机变量 X 具有概率密度

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & x > 0 \\ 0, & \text{其他} \end{cases}$$

其中 $\theta > 0$ 为常数。则称 X 服从参数为 θ 的指数分布,记为 $X \sim E\left(\frac{1}{\theta}\right)$,$X$ 的分布函数为

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{\theta}}, & x > 0 \\ 0, & \text{其他} \end{cases}$$

(3) 正态分布:设连续型随机变量 X 具有概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty$$

其中 $\mu, \sigma (\sigma > 0)$ 为常数,则称 X 服从参数为 μ, σ 的正态分布或高斯分布,记为 $X \sim N(\mu, \sigma^2)$。X 的分布函数为

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

特别地,当 $\mu = 0, \sigma = 1$ 时,称 X 服从标准正态分布。

正态分布是概率统计中最重要的分布,有广泛的实际应用背景,如测量中的误差,人体生理特征的数量指标(身高、体重等),产品的数量指标(直径、长度、体积等),学生的成绩等都服从或近似正态分布。可以说正态分布是自然界和社会现象中最常见的一种分布。一般情况下,一个变量如果受到大量微小的、独立的随机因素的影响,这个变量就是一个正态随机变量。另一方面,有些分布(二项分布、泊松分布)的极限是正态分布,有些分布(如 χ^2 分布,t 分布)又可以通过正态分布导出。因此,无论在实际中还是在理论上,正态分布都占有重要的地位。

9. 二维随机变量及其分布

有些随机试验的结果需要同时用两个或两个以上的随机变量

来描述，如炮弹着地点的位置需要由它的横坐标和纵坐标两个变量来确定，而横坐标和纵坐标是定义在同一个样本空间中的两个随机变量。设E是一个随机试验，它的样本空间是$C=\{e\}$，设$X=X(e)$和$Y=Y(e)$是定义在S上的随机变量，由它们构成的一个向量(X,Y)，叫做二维随机向量或二维随机变量。对于任意实数x,y，二元函数$F(X,Y)=P(X\leqslant x,Y\leqslant y)$称为二维随机变量$(X,Y)$的分布函数或随机变量$(X,Y)$的联合分布函数。

二维随机变量可以推广到多维的情形。

10. 随机变量的数学期望

(1) 离散型：设离散型随机变量X为$P\{X=x_k\}=p_k,k=1,2,\cdots$，若级数$\sum_{k=1}^{+\infty}x_kp_k$绝对收敛，则称级数$\sum_{k=1}^{+\infty}x_kp_k$的和为随机变量$X$的数学期望，记为$E(X)$，即$E(X)=\sum_{k=1}^{+\infty}x_kp_k$。

(2) 连续型：设连续型随机变量X的概率密度为$f(x)$，若积分$\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$绝对收敛，则称积分$\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$的值为随机变量$X$的数学期望，记为$E(X)$，即$E(X)=\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$。

数学期望简称期望，又称为均值。随机变量的分布函数能够完整地描述随机变量的统计特征，但在一些实际问题中，不需要去全面考察随机变量的变化情况，而只需要知道随机变量的某些特征，因而并不需要求出它的分布函数。数学期望虽然不能完整地描述随机变量，但能描述随机变量在某些方面的重要特性，它反映随机变量所取数值的集中位置，因而在理论上和实践中都具有重要意义。

11. 方差

设X是一个随机变量，若$E[X-E(X)]^2$存在，则称$E[X-E(X)]^2$为X的方差，记为$D(X)$或$\mathrm{Var}(X)$。对离散型随机变量，

有 $D(X) = \sum_{k=1}^{n} [x_k - E(X)]^2 P(X = x_k)$；对连续型随机变量，有 $D(X) = \int_{-\infty}^{+\infty} [x_k - E(X)]^2 f(x)\mathrm{d}x$，其中 $f(x)$ 是随机变量 X 的概率密度。

方差反映了随机变量 X 的取值与数学期望的偏离程度，其计算公式为：

$$D(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

12. *协方差及相关系数*

对于二维随机变量 (X,Y)，除讨论 X 与 Y 的数学期望和方差之外，还需要讨论描述 X 与 Y 之间相互关系的数字特征。当 X 与 Y 独立时，$D(X+Y) = D(X) + D(Y)$，但一般情况下，$D(X+Y) = D(X) + D(Y) + 2E[X - E(X)(Y - E(Y)]$。易见 $E[X - E(X)(Y - E(Y)]$ 是说明 X 与 Y 不独立的一个指标，称其为随机变量 X 与 Y 的协方差，记为 $\mathrm{Cov}(X,Y)$，即 $\mathrm{Cov}(X,Y) = E[X - E(X)(Y - E(Y)]$，或 $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$。

由于当 $X = Y$ 时 $\mathrm{Cov}(X,Y) = D(X)$ 表示随机变量 X 的偏差程度（即方差），故 $\mathrm{Cov}(X,Y)$ 反映了 X 与 Y 的偏差的关联程度，另外由于当 $a > 0, b > 0$ 时 $\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$ 以及 $\sqrt{D(aX)} = a\sqrt{D(X)}$，$\sqrt{D(bY)} = b\sqrt{D(Y)}$，因此，一旦 X 扩大 a 倍，Y 扩大 b 倍，则 X 和 Y 的协方差就会扩大 ab 倍，由此可见，协方差的大小是依赖于 X 和 Y 的线性变换，为了克服随机变量的线性效应，定义 $\rho(X,Y) = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)D(Y)}}$，称为随机变量 X 与 Y 的相关系数。它是一个无量纲的数，反映了 X 与 Y 的相关程度。由定义易见，$\rho(aX + c, bY + d) = \pm\rho(X,Y)$。当 a,b 同号时为正，异号时为负。这说明当 X 和 Y 同时线性增加时，相关系数不变，而当 X 和 Y 一个线性增加，而另一个线性减少时，相关系数大小不变，符号相反，

因而$\rho(X,Y)$反映了消去线性效应后的相关关系。

二、统计中的一些常用概念和术语

1. 总体与总体单位

根据研究目的所确定的研究对象的全体组成的集合称为总体,组成总体的每个元素称为个体,或总体单位。在统计工作中,确定总体是为了确定调查研究的对象和范围,确定总体单位则是确定调查登记项目的承担者。总体根据总体单位的数量可分为有限总体和无限总体。

2. 同质与变异

同一总体中的每一个体都具有相同性质类别的特征称为同质,同一总体中的各个个体间的差异称为变异。

3. 标志和指标

标志是指构成事物的各个单位所具有的属性和特征。原始数据资料必须按某一标志分类、整理(如按时间、地域、数量等标志分类)才能使反映客观对象的统计数据一目了然,条理清晰。按其是否具有数字特征,标志可分为品质标志和数量标志。标志以品质属性表达的,称为品质标志,如某一统计人群的性别分为男、女,文化程度可分为小学、初中、高中、大学等。说明总体单位数量特征的标志,称为数量标志,如工厂的职工人数、资金总额、产值、利润以及工人的工资、工龄等。标志按总体单位的属性倾向还可分为不变标志和可变标志。所有总体单位共同具有的特征,叫不变标志,它是构成总体的必要条件和确定总体范围的标准。在各总体单位之间必然存在差异的标志,叫可变标志。可变标志中既有品质标志,也有数量标志,其中可变的数量标志称为变量。

指标是说明总体数量特征的概念。将标志的具体表现(变异和变量)进行登记、汇总,最后形成说明事物综合特征的各种数字资料,称为统计指标。例如在工业调查中,所有工业企业构成总体,

工业企业总数、工人总数、工资总额、平均工资、固定资产总值、总利润等，都从不同方面反映总体的数量特征，它们都是统计指标。

在统计实践中，任何一个统计指标都只能反映总体某一方面的数量特征。要全面系统地对总体进行了解与把握，必须同时使用多个指标。这种以共同的研究目的为纽带而相互联系的一系列统计指标，称为指标体系。具体统计过程中，根据统计的目的，必须预先设计精简合理的统计指标体系。

4. 随机化

能使总体中每一观察单位均能以相同概率进入样本，或分配到实验组与对照组的过程，称为随机化。其特点是“机会均等”地选取或分配研究对象的“公平性”原则。现实生活中的许多“公众法则”，如“抽签”、“抓阄”、“猜拳”以及一些福利彩票活动等之所以被认为是合理的，原因在于它的“公平性”，即对每个人的机会均等性。游戏者即使失败了，也会自认“公道”。这种公认的合理性在统计学中称为“等概率”，该规则在数学上的名称就是“随机化”。

5. 统计量

统计量是指不含任何未知参数的样本$(X_1, X_2, \cdots, X_n)$的函数$T = T(X_1, X_2, \cdots, X_n)$，它是根据样本数据计算出来的一个量。比如样本均值$(\bar{x})$、样本标准差$(s)$、样本比例$(p)$等统计指标，都称为统计量，它是一个随机变量。

引进统计量的目的是为了将杂乱无章的样本值整理成便于对所研究问题进行统计推断、分析的形式，将样本中所含的研究问题的有关信息集中起来，从而更有效地提示出问题的实质，进而得到研究对象的统计规律。当然，选择的统计量应较好地集中样本所含的关于所研究问题的信息。由于统计量的使用目的是对所研究的问题进行推断分析，因此，当用统计量对未知参数进行估计时，如果统计量本身仍含有参数，那么，就无法根据所测得的样本值求得

未知参数的估计值。利用统计量估计未知参数就会失去意义。这是要求它不含参数的原因之一。

由于统计学是研究随机现象规律性的科学，因此，具有较多的数学学科特征，上面所简单陈述的基本概念、公式和术语，只是统计学中所用到的极少部分，可以说只是冰山一角，读者要想系统地掌握统计学的原理和方法然后加以应用，必须立足于实际，有针对性地强化对抽象的数学概念、符号和公式的理解与学习，必须从应用的角度理解统计方法在解决实际问题中所发挥的作用。

第三节　统计工作的步骤和过程

从程序上来看，统计工作是一项复杂的系统化的实践活动。可分为统计设计、资料收集、资料整理和统计分析四个步骤。其中统计设计是最关键的一步，它从总体上决定了统计工作的效率。

一、统计设计

统计设计是与统计目的相联系的对整个统计工作的整体安排和计划。其定义是，用统计学原理对研究的全过程所作出的周密合理的统筹规划，如确定研究对象，拟定研究因素及其分配，如何执行随机、对照与重复的统计学原则，如何观察与度量效应，如何对数据进行收集、整理与分析，如何控制系统误差，怎样以尽可能小的资源耗费（最少的人力、物力、财力和时间等）来获取准确、可靠的信息资料，使效益最大化。它还直接影响到统计结论的真实性和可靠性。科学的设计原则有利于树立统计结论的权威，而不合理的统计设计会造成统计结论的模糊和不确定，进而导致人们对统计结论产生置疑。如20世纪80年代初，有研究报道孕期补充维生素可以减少生育神经管缺陷婴儿的风险，即先服用维生素后怀孕的妇女比怀孕后才开始服用维生素的妇女和拒绝参加试验的怀孕妇

女所生的婴儿神经管缺陷的发生率要低得多。但由于参加服用维生素试验和拒绝参加试验的孕妇间存在某些生理特征上的系统差别，致使在解释试验结果时发生困难。为补救先前研究的不足，后续的研究采用了随机化分配受试者的方法，分叶酸补充组和安慰剂组。但在观察结果时，又因样本过少而无法做出肯定的结论。直到1991年，研究者报道了一个大样本的随机化试验，才获得了肯定的科学结论。在安慰剂组的602名怀孕妇女中有21名分娩出的新生儿有神经管缺陷，在叶酸补充组的592名怀孕妇女中出现新生儿有神经管缺陷的只有6例，而其他维生素（不含叶酸）的补充对新生儿有神经管缺陷的发生无明显影响。其间的显著性差异说明叶酸对预防新生儿有神经管缺陷有明显效果[1]。可见在研究之前预先制定一套科学的统计方案对研究的结果有重要的影响。作为统计工作的关键环节，统计设计应该按以下原则展开：一是根据研究目的，做出具体的统计任务和所要获取有用信息的技术路线与方法，科学、合理、周密地安排整个统计工作的全过程；二是尽可能以较为经济的人力、物力、财力和时间进行试验，最大限度地减少误差，从而获得可靠的结果。具体工作中，统计设计的内容应主要考虑以下方面：

1. 确立明确的统计指标体系，营造畅通、准确的统计调查渠道

科学研究或试验的顺利开展，很大程度上取决于对于研究本身的统计指标体系的确立。它是统计分析的先决条件，是获得预期结果的重要保证。如果没有明确的统计指标，必然会造成实验数据统计上的遗漏和混乱，从而使整个研究陷于被动。而建立畅通、准确的统计调查渠道无疑是获取有用信息的必要手段，在此必须要做好如下几个方面的工作：

（1）检查一切所用的实验仪器和设备，确保实验工具合乎质量上的要求；

（2）建立与被调查对象（人群）之间的良好关系，以保证被调查者（或实验对象）对试验具有积极的心态倾向；

（3）排除各种人为因素的干扰，对统计数据所表现出的差异

不要盲目倾向于权威；

(4) 杜绝仅凭经验和想象而随意编造试验数据的严重错误。

其中建立与被调查人群的良好关系，在以人为研究对象的试验中，具有特别重要的作用。因为被调查者的主观态度决定了他对试验的配合程度，不积极的配合可能直接导致试验结果的完全背离。

2. 选择具有随机性的试验对象，并根据试验要求确定样本容量

试验对象的随机性可以使试验结果不会因为主观因素而产生人为的偏差，而试验对象的数量指标则是减少由其内在因素的差异所导致的实验误差的必要保证。

3. 制定统计资料收集的科学方法

对统计资料的收集要根据研究对象或调查单位的特点以及研究内容的特殊要求来确定科学的方法。无论是调查研究还是实验研究，也不论研究者同时确定多少统计指标，都必须严格按照周密、条理、严谨的统计程序来观测、记录试验数据，有步骤、分层次地组织数据结构，合理、明确地确定变量的类型，为后续的分析研究作好材料上的充分准备。

4. 建立减小统计误差的有效的数学模型

无论何种类型的试验，都必须首先考虑减少误差的有效方法和途径，这是保证试验取得预期结果的必要前提。一个好的统计模型可以使试验数据的收集更快速、更准确，从而可以极大地提高统计工作的效率。这也是统计设计重点考虑的内容。

5. 合理规划试验步骤，确定科学的试验方法

试验究竟按怎样的步骤进行，试验内容的先后次序的确定，试验方法的选择与制定都将影响到试验的效率和试验结果的可靠性。在统计设计中必须充分考虑到试验方法对结果的影响，根据试验的目的和要求制定最科学的实验方法。

6. 设计完善的数据分析模式

在试验中获取的数据往往是大量的、复杂的，如何整理、分析

这些数据，选择什么样的工具（软件）来分析，如何建立试验数据与预测结果之间的联系等等，都是研究者所要面对的具体问题。统计设计中必须在宏观上充分考虑到这些因素并预先确定出大体的处置方案，具体实践中才能做到有条不紊地对各种复杂数据进行正确的处理，否则，可能会面对大量冗杂的数据而不知所措。

统计设计对于各种不同的试验类型还会有不同的特点，总的说来，研究者必须综合考虑上述各种因素，立足于全局，从宏观上把握整个研究的全过程，避免和减少主观因素的影响，克服和杜绝各个环节中的任意性和盲目性，使之成为研究计划的全面顺利展开的参照和基础。

二、资料收集

资料收集是统计工作的关键环节，统计工作的核心就是根据研究的目的来收集或记录统计数据。有了这些基础资料，我们才能对总体的数量特征和数量关系作进一步的描述和推断。在调查性的试验研究中，由于行业间的利益关系和人们心理倾向的差异，往往使获取令人满意的资料的工作变得十分困难和艰巨。数据的收集与整理历来是统计学家十分关注和致力研究的领域，它在整个统计研究中占有重要的地位[2]。统计研究中对资料的收集根据研究的需要可分为原始资料的收集和次级资料的收集。所谓原始资料是指试验中通过直接观察或实地调查而获得的，未经任何加工的第一手资料。次级资料是指已经加工过的，往往是公开发表过的资料（如从《统计年鉴》、会计报表及各种期刊或杂志上获得的资料）。由于次级资料由原始资料经加工、整理而得到，因此对原始资料收集必须做到准确、完整，它是保证统计数据信息质量的关键。对原始资料的收集，有以下几种组织形式：

1. 普查

这是针对调查性试验研究所采取的一种数据收集形式。主要用于一些重要项目的调查，如人口普查、耕地面积普查、环境质量

普查等。普查涉及面广、调查对象多、周期长，需要花费较多的人力、物力和财力，但可获得较准确、全面的信息。

2. 典型调查

典型调查是根据调查的目的和要求，在对调查对象进行全面分析的基础上，有针对性地选择部分有代表性的单位所作的调查。它主要用于对某些尚处于萌芽状态的新生事物或具有某种倾向性的社会问题的研究。但这种方式所取得的资料往往存在较大误差。

3. 重点调查

这种调查方式是通过在总体中选择具有举足轻重地位的单位作调查而得名。这些单位尽管数量上可能不占多数，但它们所具有的调查的标志值在总体标志总量中占有绝对优势。通过对这些单位的调查，一般就可以掌握总体的基本情况。如要了解全国钢铁企业的生产情况，只需调查宝钢、鞍钢、太钢等大型钢铁企业，就能达到调查的目的。因此，当调查研究目的只要求了解调查对象的基本情况，而在总体中确有部分单位能较集中地反映所要研究的问题时，进行重点调查是比较适宜的[2]。

4. 随机抽样调查

随机抽样调查是指按随机性原则从总体中抽取部分单位进行调查，借以推断、认识总体的一种统计方法。它是在对总体的各研究指标没有任何先期了解的情况下进行的，是各种非全面调查方法中最科学的一种方法，正逐步成为我国统计调查方法体系中的主要方法，也是现代推断统计的核心。因为无论是对总体的参数估计或假设检验，都是以测定样本得到的样本指标为依据的。与典型调查和重点调查两种非全面调查相比较，随机抽样调查有以下几个显著特点：

(1) 调查不带有任何主观倾向，完全以随机性原则抽取样本，总体中每个单位被抽中的机会均等。而典型调查和重点调查则受人的主观因素的影响过多。

(2) 随机抽样调查以样本指标(统计量) 为依据推断总体参

数或对总体的某种特征作假设，目的就是对总体数量特征作出估计或判断。由于它是以概率论的有关分布律为依据的估计，故可以计算其判断的可靠性和精确度。

（3）随机抽样调查的误差可以事先计算并加以控制。抽样误差是抽样调查所固有的，它是一个随机变量，其分布具有一定的规律性，可以依据这种规律和具体的抽样条件来计算抽样误差的大小，并根据研究精度的要求可以调整抽样的数量或比率来控制误差的范围。这在一定程度上提高了抽样调查的应用价值。

统计资料的收集是一个系统的工程，根据调查对象、调查单位的特点和调查内容的不同要求，具体实施时通常使用直接观察、采访、汇报、问卷调查、网上搜集等方法。

三、统计资料整理

由统计调查和观察记录获取到的大量原始资料往往是分散的、杂乱的。为了便于分析，必须按照一定的规则加以整理，使之条理化、系统化。统计整理是整个统计工作和研究过程的中间环节，它是统计调查的继续，又是统计分析的基础。统计调查所收集到的资料，只有通过科学的组织与整理，才能使之由个别的数量特征上升为反映总体的数量特征。统计资料的整理包括资料的审核、分组、汇总计算和统计制表、制图等。

1. 统计资料的审核

对统计资料的审核是统计整理的第一步，主要包括：

（1）审查资料的完整性；

（2）审查资料的准确性和可靠度；

（3）审查资料的一致性（主要是审查统计资料的获取是否具有相似的条件）。

2. 统计资料的分组

统计分组是数据整理的核心，是根据统计研究的目的要求和研究对象的特点，按某种重要标志把总体数据分成若干部分的科

学分类。从哲学观看，任何事物都是相互联系的，不同事物会在一定的环境中以一定的条件而相互关联，这种联系是构成各种统计总体的前提。但不同的事物之间，在产生的原因、存在的条件、表现的形式、运动的规律等方面，又具有各自的特点，甚至是千差万别。正是这些区别，使统计分组有了客观的依据。可以说，没有科学的统计分组，就没有科学的统计。

3. 统计资料的汇总

在统计分组的基础上，将统计资料归并到各组中去，并计算各组和总体的合计数（包括总体总量和标志总量）的工作过程称为统计资料汇总[3]。统计汇总也是一项复杂的工作，必须采用先进科学的汇总技术，才能节约成本、提高效率。统计汇总有逐级汇总和集中汇总两种组织形式。由于计算机技术的不断发展，对大量数据的汇总，可以借助于计算机来完成。

4. 统计制表、制图

为了使整理后的统计数据能更加直观地反映总体的特征，常常将经过加工整理过的统计数据集中有序地以表格或图形的形式表现出来。实现这一目的过程和手段就是统计制表和统计制图。层次清晰、结构合理的统计表不仅可以简明地反映研究对象的总体特征，避免冗长烦琐的文字叙述，而且可以深入地揭示研究对象的各指标或属性间的联系，甚至可以由此发现研究对象的发展规律。如果统计表在表达统计数据的关系方面尚有不足，那么可以借助于统计图的直观、形象和生动的特性来加以弥补。一般来说，统计图能更形象、准确地表达统计数据所反映的规律。根据统计数据的性质和特点，以及统计研究的类型，统计数据常以直方图、曲线图、饼图、柱形图等形式来表达。详见第二章第二节。

四、统计分析

统计分析是整个统计工作的最终环节，它决定了统计工作的

全部价值，是研究者进行决策分析的基础和前提。建立在对统计资料进行加工整理基础上的统计分析，是研究者进行探索、推理、找寻、描述客观规律的积极的活动。统计分析的结果直接决定了研究者对研究对象内在规律的判断。可以说统计分析的成功，是决策者制定计划、设定目标以及作出正确决策的有力保障。很难想象，错误的统计分析结果用于指导工作，制定计划会带来什么样的严重后果。从这一角度上讲，每一个统计工作者都应充分认识到统计分析的重要性和价值，决不可仅凭主观和想象来推断或臆造统计规律。

这里需要说明的一点是，尽管统计分析是基于对统计资料的加工整理之上的统计工作的最后阶段，但绝不可将统计分析与统计整理在主观上隔裂开来。事实上，很多情况下统计分析与统计整理是完全融合在一起的。因为统计分析的目的是为了得出规律性的认识，而这种分析本身又必然要求统计工作者对统计数据进行各种形式的整理，常常是统计整理必须是在统计分析的基础上进行的整理，是在统计分析“指导”下进行的整理。不含任何“分析”成分的整理一定是盲目的整理，甚至是无用的整理。由此看来，统计分析仍然是一个复杂的、系统化的过程，由于它以发现统计规律为目的，必然包含带有许多数学运算技巧的数据加工整理过程，因此要求统计工作者必须有一定的数学基础，我们将在后续的章节中对统计过程的一些重要环节逐一展开介绍。