第五章：数理统计的基本概念与抽样分布

张伟平

第五章：数理统计的基本概念与抽样分布

4.1 引言

4.1.1 数理统计学

本课程的前四章介绍了概率论的基本内容, 为数理统计学建立了重要的数学基础. 从本章起, 我们转入本课程的第二部分 —数理统计学. 下面我们首先说明什么是数理统计学.

统计学的任务是研究怎样有效地收集、整理和分析带有随机性影响的数据, 从而对所考虑的问题作出一定结论的方法和理论. 它是一门实用性很强的学科, 在人类活动的各个领域有着广泛的应用. 研究统计学方法的理论基础问题的那一部分构成 “数理统计学” 的内容. 一般地可以认为

数理统计是数学的一个分支, 它是研究如何有效地收集和有效地使用带有随机性影响的数据的一门学科.

下面通过例子对此加以说明.

1. 有效地收集数据

收集数据的方法有: **全面观察 (或普查)、抽样调查和安排试验**等方式.

↑Example

人口普查和抽样调查. 我国在 2000 年进行了第五次人口普查. 如果普查的数据是准确无误的, 无随机性可言, 不需用数理统计方法. 由于人口普查, 调查项目很多, 我国有 13 亿人口, 普查工作量极大, 而训练有素的工作人员缺乏. 因此虽是全面调查, 但数据并不可靠, 农村超计划生育瞒报、漏报人口的情况时有发生. 针对普查数据不可靠, 国家统计局在人口普查的同时还派出专业人员对全国人口进行抽样调查, 根据抽样调查的结果, 对人口普查的数字进行适当的修正. 抽样调查在普查不可靠时是一种补充办法.

↓Example

如何安排抽样调查, 这是有效收集数据的重要问题, 这构成数理统计学的一个重要分支 —《抽样调查方法》.

考察某地区 10000 农户的经济状况. 从中挑选 100 户做抽样调查. 若该地区分成平原和山区两部分, 平原地区较富, 占该地区农户的 70%, 山区的 30% 农户较穷. 我们的抽样方案规定在抽取的 100 户中, 从平原地区抽 70 户, 山区抽 30 户, 在各自范围内用随机化方法抽取.

↑Example

↓Example

在本例中有效收集数据是通过合理地设计抽样方案来实现的. 在通过试验收集数据的情形如何做到有效收集数据, 请看下例:

↑Example

某化工产品的得率与温度、压力和原料配方有关. 为提高得率, 通过试验寻找最佳生产条件. 试验因素和水平如下

因素 \ 水平	1	2	3	4
温度	800	1000	1200	1400
压力	10	20	30	40
配方	A	B	C	D

3 个因素, 每个因素 4 个水平共要做 $4^3 = 64$ 次试验. 做这么多试验人力、物力、财力都不可能. 因此, 如何通过尽可能少的试验获得尽可能多的信息? 比如采用正交表安排试验就是一种有效的方法.

↓Example

如何安排试验方案和分析试验结果, 这构成数理统计的另一分支

—《试验的设计和分析》. 在本例中有效收集数据是通过科学安排试验的方法来实现的.

在有效收集数据中一个重要问题是: 数据必须具有随机性.

2. 有效的使用数据

获取数据后, 需要用有效的方法, 去集中和提取数据中的有关信息, 以对所研究的问题作出一定的结论, 在统计上称为 "**推断**".

为了有效的使用数据进行统计推断, 需要对数据建立一个统计模型, 并给定某些准则去评判不同统计推断方法的优劣.

为估计一个物体的重量 a, 把它在天平上称 5 次获得数据 $x_1, x_2, \cdots, x_5$, 它们都受到随机性因素的影响 (天平的精度反映了影响的大小). 估计 a 的大小有下列三种不同方法: (1) 用 5 个数的算术平均值 $\bar{x} = \frac{1}{5}(x_1 + \cdots + x_5)$ 去估计 a; (2) 将 $x_1, x_2, \cdots, x_5$ 按大小排列为 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(5)}$, 取中间一个值 $x_{(3)}$ 去估计 a; (3) 用 $W = \frac{1}{2}(x_{(1)} + x_{(5)})$ 去估计 a. 你可能认为 $\bar{x}$ 优于 $x_{(3)}$, 而 $x_{(3)}$ 优于 W. 这是不是对的? 为什么是这样? 在什么条件下才对? 事实上, 对这些问题的研究正是数理统计学的任务.

↑Example

↓Example

要回答这些问题我们需要对数据建立一个统计模型和制定评判不同统计推断方法的准则. 本例中在适当的假定下, 可认为数据服从正态模型.

下面我们举一个例子说明采用合适的统计方法也是有效使用数据的一个重要方面.

某农村有 100 户农户, 要调查此村农民是否脱贫. 脱贫的标准是每户年均收入超过 1 万元. 经调查此村 90 户农户年收入 5000 元, 10 户农户年收入 10 万元, 问此村农民是否脱贫?

↑Example

↓Example

(1) 用算术平均值计算该村农户年均收入如下:

$$\bar{x} = (90 \times 0.5 + 10 \times 10)/100 = 1.45(\text{万})$$

按此方法得出结论: 该村农民已脱贫. 但 90% 的农户年均收入只有 5000 元, 事实上并未脱贫.

(2) 用样本中位数计算该村农户年均收入: 即将 100 户的年收入记为 $x_1, x_2, \cdots, x_{100}$, 将其按大小排列为 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(100)}$. 样本中位数定义为排在最中间两户的平均值, 即

$$(x_{(50)} + x_{(51)})/2 = 0.5(\text{万})$$

按此方法得出结论: 该村农民尚未脱贫. 这与实际情况相符.

3. 数理统计方法的归纳性质

数理统计是数学的一个分支, 但是它的推理方法是不一样的. **统计方法的本质是归纳式的, 而数学则是演绎式的**. 统计方法的归纳性质, 源于它在作结论时, 是根据所观察到的大量的 “个别” 情况, “归纳” 起来所得. 而不是从一些假设、命题或已知事实出发按一定的逻辑推理得出来的 (这后者称为演绎推理) . 举一例子说明: 统计学家通过大量的观察资料发现, 吸烟与某种呼吸系统的疾病有关. 他得出这一结论的根据是: 从观察到的大量例子, 看到吸烟者中患此种疾病的比例远高于不吸烟者. 他不可能用逻辑推理的方法证明这一点. 试拿统计学与几何学进行比较就可以清楚地看出二者方法的差别所在. 在几何学中要证明 “等腰三角形两底角相等”, 只需从等腰这个前提出发, 运用几何公理, 一步步地推出这个结论 (这一方法属于演绎推理) . 而一个习惯于统计方法的人, 就可能想出这样的方法: 作很多大小形状不一的等腰三角形, 实际测量它的底角查看区别如何, 根据所得数据, 看看可否作出底角相等的结论, 这属于归纳推理的方法.

众所周知, **归纳推理是要冒风险的**. 事实上归纳推理的不确定性

的出现, 是一种逻辑的必然. 人们不可能做出十分肯定的结论, 因为归纳推理所依据的数据具有随机性. 然而, 不确定性的推理是可行的, 所以推理的不确定性程度是可以计算的. 统计学的作用之一就是提供归纳推理和计算不确定性程度的方法. 不确定性是用概率计算的. 以后会见到我们求参数的**区间估计**, 不但给出区间估计的表达式, 而且给出这一估计区间包含未知参数的可靠程度的大小.

4.2 数理统计的若干基本概念

4.2.1 总体和样本

通过下面的例子说明总体、个体和样本的概念.

↑Example

假定一批产品有 10000 件, 其中有正品也有废品, 为估计废品率, 我们往往从中抽取一部分, 如 100 件进行检查. 此时这批 10000 件产品称为**总体**, 其中的每件产品称为**个体**, 而从中抽取的 100 件产品称为**样本**. 样本中个体的数目称为**样本的大小**, 也称为**样本容量**. 而抽取样本的行为称为**抽样**.

↓Example

从本例我们可对总体和样本作如下直观的定义:
总体是与我们所研究的问题有关的所有个体组成, 而**样本**是总体

中抽取的一部分个体.

若总体中个体的数目为有限个, 则称为**有限总体**, 否则称为**无限总体**.

在统计研究中, 人们所关心的不是总体内个体的本身, 而是关心个体上的一项 (或几项) 数量指标, 如日光灯的寿命, 零件的尺寸. 在例4.2.1中若产品为正品用 0 表示, 若产品为废品用 1 表示, 我们关心的个体取值是 0 还是 1. 因此我又可获得总体的如下定义:

总体可以看成是由所有个体上的某种**数量指标**构成的集合, 因此它是**数**的集合.

由于每个个体在抽样时的出现是随机的, 所以相应的个体上的数量指标的出现也带有随机性. 从而可以把此种数量指标看成随机变量, 随机变量的分布就是该数量指标在总体中的分布. 以例4.2.1来说明, 假定 10000 只产品中废品数为 100 件, 其余的为正品, 废品率为 0.01. 我们定义随机变量 X 如下:

$$X = \begin{cases} 1 & \text{废品} \\ 0 & \text{正品}, \end{cases}$$

其概率分布为 0–1 分布, 且有 $P(X=1)=0.01$. 因此, 特定个体上的数量指标是随机变量 X 的观察值. 这样一来, 总体可以用一个随机变量 X 及其分布来描述, 获得如下定义:

一个统计问题所研究的对象的全体称为总体. 在数理统计学中总体可以用一个随机变量及其概率分布来描述.

Definition

由于总体的特征由其分布来刻画, 因此统计学上常把总体和总体分布视为同义语. 由于这个缘故, 常用随机变量的符号或分布的符号来表示总体. 比如研究某批日光灯寿命时, 人们关心的数量指标是寿命 X, 那么此总体就可以用随机变量 X 来表示, 或用其分布函数 F 来表示. 若 F 有密度, 记为 f, 则此总体也可用密度函数 f 来表示. 有时也根据总体分布的类型来称呼总体的名称, 如正态总体、二项分布总体、0–1 分布总体. 若总体分布函数记为 F, 当有一个从该总体

中抽取的相互独立同分布 (i.i.d.) 的大小为 n 的样本 $X_1, \cdots, X_n$, 则常记为

$$X_1, \cdots, X_n \text{ i.i.d. } \sim F \tag{4.1}$$

若 F 有密度 f, 可记为

$$X_1, \cdots, X_n \text{ i.i.d. } \sim f \tag{4.2}$$

若所考虑的总体用随机变量 X 表示其分布函数为 F, 则样本 $X_1, \cdots, X_n$ 可视为随机变量 X 的观察值, 亦可记为

$$X_1, \cdots, X_n \text{ i.i.d. } \sim X \tag{4.3}$$

(4.1)、(4.1) 和 (4.3) 表示相同的意思.

当个体上的数量指标不止一项时, 我们用随机向量来表示总体. 例如研究某地区小学生的发育状况时, 人们关心的是其身高 X 和体重 Y 这两个数量指标, 此时总体就可以用二维随机向量 (X, Y) 或其联合分布 $F(x, y)$ 表示.

4.2.2 样本的两重性和简单随机样本

1. 样本空间

我们知道样本是由总体中抽取的一部分个体组成. 设 $X = (X_1, \cdots, X_n)$ 是从总体中抽取的一个样本, 其样本空间如下:

Definition

样本 $X = (X_1, \cdots, X_n)$ 可能取值的全体成为**样本空间** , 记为 $\mathscr{X}$.

例如在前面称重例中, 样本空间为 $\mathscr{X} = \{(x_1, \cdots, x_5) : \ 0 < x_i < \infty, \ i = 1, 2, \cdots 5\}$, 也可以写成 $\mathscr{X} = \{(x_1, \cdots, x_5) : \ -\infty < x_i < \infty, \ i = 1, 2, \cdots 5\}$. 虽然物重不可以取负数, 但这无关紧要, 因为在考虑样本分布时, 可令样本取负值的概率为 0. 再看下例:

打靶试验, 每次打三发, 考察中靶的环数. 如样本 $X = (5, 1, 9)$ 表示三次打靶分别中 5 环、1 环和 9 环. 此时样本空间为

$$\mathscr{X} = \{(x_1, x_2, x_3) : \ x_i = 0, 1, 2, \cdots 10, \ i = 1, 2, 3\}$$

这个样本空间中样本点数是有限的, 上例称重问题中样本空间中的样本点数是无限的.

↑Example

↓Example

1、样本的两重性

当我们从总体中作具体抽样时, 每次抽样的结果都是些具体的数, 如例 5.2.3 的打靶问题中, 3 维样本 $X = (X_1, X_2, X_3)$, 其中 $0 \leq X_i \leq 10$为整数, $i = 1, 2, 3$, 它是数字向量. 但若是在相同条件下, 再打三发, 由于种种不可控制的随机因素的影响, 中靶的环数不可能和上一次完全一样, 具有随机性. 如果无穷次打下去, 每次打三发, 出现的结果可视为随机向量 (X_1, X_2, X_3) 的观察值.

样本的两重性是说, **样本既可看成具体的数, 又可以看成随机变量 (或随机向量)**. 在完成抽样后, 它是具体的数; 在实施抽样前, 它被看成随机变量. 因为在实施具体抽样之前无法预料抽样的结果, 只能预料它可能取值的范围, 故可把它看成一个随机变量, 因此才有概率分布可言。为区别起见, **今后用大写的英文字母表示随机变量或随机向量, 用小写字母表示具体的观察值**.

对理论工作者, 更重视样本是随机变量这一点, 而对应用工作者虽则将样本看成具体的数字, 但仍不可忽视样本是随机变量 (或随机向量) 这一背景. 否则, 样本就是一堆杂乱无章毫无规律可言的数字, 无法进行任何统计处理. 样本既然是随机变量 (或随机向量), 就有分布而言, 这样才存在统计推断问题.

2、简单随机样本

抽样是指从总体中按一定方式抽取样本的行为. 抽样的目的是通过取得的样本对总体分布中的某些未知因素做出推断, 为了使抽取的样本能很好的反映总体的信息, 必须考虑抽样方法. 最常用的一种抽样方法叫作 “简单随机抽样”, 它要求满足下列两条:

(1) 代表性. 总体中的每一个体都有同等机会被抽入样本, 这意味着样本中每个个体与所考察的总体具有相同分布. 因此, 任一样本中的个体都具有代表性.

(2) 独立性. 样本中每一个体取什么值并不影响其它个体取什么值. 这意味着, 样本中各个体 $X_1, X_2, \cdots, X_n$ 是相互独立的随机变量.

由简单随机抽样获得的样本 $(X_1, \cdots, X_n)$ 称为简单随机样本. 用数学语言将这一定义叙述如下:

Definition

设有一总体 F, $X_1, \cdots, X_n$ 为从 F 中抽取的容量为 n 的样本, 若

(i) $X_1, \cdots, X_n$ 相互独立,

(ii) $X_1, \cdots, X_n$ 相同分布, 即同有分布 F,

则称 $(X_1, \cdots, X_n)$ 为简单随机样本, 有时简称简单样本或随机样本.

设总体为 F, $(X_1, \cdots, X_n)$ 为从此总体中抽取的简单样本, 则 $X_1, \cdots, X_n$ 的联合分布为:

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

若 F 有密度 f, 则其联合密度为

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

显然, 有放回抽样获得的样本是简单样本. 当总体中个体数较大或所抽样本在总体中所占比例较小时, 无放回抽样获得的样本可以近似认为是简单样本.

4.2.3 统计模型

样本既然是随机变量, 就有一定的概率分布, 这个概率分布就叫作**样本分布**. 样本分布是样本所受随机性影响的最完整的描述.

要决定样本分布, 就要根据观察值的具体指标的性质 (这往往涉及有关的专业知识), 以及对抽样方式和对试验进行的方式的了解, 此外常常还必须加一些人为的假定. 下面看一些例子:

↑Example

一大批产品共有 N 个, 其中废品 M 个, N 已知, 而 M 未知. 现在从中抽出 n 个加以检验, 用以估计 M 或废品率 $p = M/N$.

(1) 有放回抽样, 即每次抽样后记下结果, 然后将其放回去, 再抽第二个, 直到抽完 n 个为止. 求样本分布.

(2) 不放回抽样, 即一次抽一个, 依次抽取, 直到抽完 n 个为止. 求样本分布.

↓Example

解: (1) 在有放回抽样情形, 每次抽样时, N 个产品中的每一个皆以 $1/N$ 的概率被抽出, 此时 $P(X_i = 1) = M/N$, $P(X_i = 0) =$

$(N-M)/N$, 故有

$$P(X_1 = x_1, \cdots, X_n = x_n) = \left(\frac{M}{N}\right)^a \left(\frac{N-M}{N}\right)^{n-a}, \tag{4.4}$$

当 $x_1, \cdots x_n$都为 0或 1, 且 $\sum_{i=1}^{n} x_i = a$ 时为上述结果 (其余情形为 0).

(2) 若采取不放回抽样, 则计算要复杂的多, 读者可作为练习, 现将结果给出如下: 记 $\sum_{i=1}^{n} x_i = a$, 利用概率乘法公式易求

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= \frac{M}{N} \cdot \frac{M-1}{N-1} \cdots \frac{M-a+1}{N-a+1} \cdot \frac{N-M}{N-a} \cdots \frac{N-M-n+a+1}{N-n+1}, \end{aligned} \tag{4.5}$$

当 $x_1, \cdots x_n$都为 $0, 1$, 且 $\sum_{i=1}^{n} x_i = a$ 时为上述结果 (其余情形为 0) .

上述计算之所以复杂, 是因为在不放回情形, 样本 $X_1, \cdots, X_n$ 不是相互独立的, 样本分布是利用乘法公式, 通过条件概率计算出来的.

而在有放回的情形, 样本 $X_1, \cdots, X_n$ 是独立同分布的, 因此要简单得多.

当 n/N 很小时, (4.5) 和 (4.4) 差别很小. 因而当 n/N 很小时可把上例中的无放回抽样近似当作有放回抽样来处理.

所谓一个问题的**统计模型**, 就是指研究该问题时所抽样本的**样本分布**, 也常称为**概率模型**或**数学模型**.

由于模型只取决于样本的分布, 故常把分布的名称作为模型的名称. 如下列例4.2.3中样本分布为正态，可称其为正态模型。因此把模型和样本紧密联系起来是必要的. 统计分析的依据是样本, 从统计上说, 只有规定了样本的分布, 问题才算真正明确了.

下例告诉我们是怎样由一个具体问题建立统计模型的.

为估计一物件的重量 a, 用一架天平将它重复称 n 次, 结果记为 $X_1, \cdots, X_n$, 求样本 $X_1, \cdots, X_n$ 的联合分布.

↑Example

↓Example

解: 要定出 $X_1, \cdots, X_n$ 的分布, 就没有前面例子那种简单的算法, 需作一些假定: (1) 假定各次称重是独立进行的, 即某次称重结果不受其它次称重结果的影响. 这样 $X_1, \cdots, X_n$ 就可以认为是相互独立的随机变量. (2) 假定各次称重是在 "相同条件" 下进行的, 可理解为每次用同一天平, 每次称重由同一人操作, 且周围环境 (如温度、湿度等) 都相同. 在这个假定下, 可认为 $X_1, \cdots, X_n$ 是同分布的. 在上述两个假定下, $X_1, \cdots, X_n$ 是 n 个独立同分布的随机变量, 即为**简单随机样本**.

为确定 $X_1, \cdots, X_n$ 的联合分布, 在以上假定之下求出 X_1 的分布即可. 在此考虑称重误差的特性: 这种误差一般由大量的、彼此独立起作用的随机误差迭加而成, 而每一个起的作用都很小. 由概率论中的中心极限定理可知这种误差近似服从正态分布. 再假定天平没有系统误差, 则可进一步假定此误差为均值为 0 的正态分布. 可以把 X_1 (它可视为物重 a 加上称量误差之和) 的概率分布为 $N(a, \sigma^2)$. 因

此简单随机样本 $X_1, \cdots, X_n$ 的联合分布为

$$f(x_1, \cdots, x_n) = (\sqrt{2\pi}\sigma)^{-n} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\} \qquad (4.6)$$

本例中求样本分布, 引入两种假定: (i) 导出样本 $X_1, \cdots, X_n$ i.i.d. 的假定, (ii) 正态假定, 这一点依据问题的性质、概率论的极限理论和以往经验.

在有了研究统计模型后, 很多性质不一样的问题, 可以归入到同一模型下. 例如涉及到测量误差的问题, 只要例4.2.3中叙述的假定误差服从正态分布的理由成立, 则都可以用正态模型 (4.6). 只要把这个模型中的统计问题研究清楚了, 就可以解决许多不同专业部门中的这样一类问题.

另一方面, 同一模型下可以提出很多不同的统计问题. 如例4.2.3的 $N(a, \sigma^2)$ 模型中, 有了样本 $X_1, \cdots, X_n$, 并规定分布 (4.6) 后就有了一个统计模型. 在这个模型下可提出一些统计问题, 如在例4.2.3中, 我们的问题是估计物重 a. 为了考察天平的精度我们可以

提出估计 σ^2 的问题, 当然我们还可以对 a 和 σ^2 提出**假设检验**和**区间估计**问题等等.

4.2.4 统计推断

从总体中抽取一定大小的样本去推断总体的概率分布的方法称为**统计推断**.

数理统计是着手于样本, 着眼于总体, 其任务是用样本去推断总体. 当样本分布完全已知时是不存在任何统计推断问题.

当样本的分布形式已知, 但含有未知参数时, 有关其参数的推断, 称为**参数统计推断**.

在另一些问题中, 情况就要复杂一些. 这类问题中样本分布的形式完全未知, 有关其分布的统计推断问题称为**非参数统计推断**问题.

参数统计推断有种种不同的形式: 主要有**参数估计**和**假设检验**问题. 如例4.2.3中样本分布 (亦即总体分布) $N(a, \sigma^2)$ 中, 当 a 和 σ^2 未知时, 从总体中抽取大小为 n 的样本 $X_1, \cdots, X_n$, 对 a 和 σ^2 的取值

作出估计, 或对断言 “$a \leq 1$” 作出接受或拒绝这一假设的结论.

非参数问题中, 统计推断的主要任务是通过样本对总体分布的形式作出推断.

由于样本的随机性, 统计推断的结论不可能 100% 的正确, 但我们可以给出衡量推断正确程度的指标. 如在例4.2.3中, 若用 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 估计 a, 可以算出 $\bar{X}$ 与 a 的偏差大于 c 的概率, 即 $P(|\bar{X} - a| > c)$, 作为用 $\bar{X}$ 推断 a 的正确性的合理指标.

统计推断包括下列三方面内容: (1) 提出种种的统计推断的方法. (2) 计算有关统计推断方法性能的数量指标, 如前述例子中用 $\bar{X}$ 估计 $N(a, \sigma^2)$ 中的 a, 用 $P(|\bar{X} - a| > c)$ 表示推断性能的数量指标. (3) 在一定的条件和优良性准则下寻找最优的统计推断方法, 或证明某种统计推断方法是最优的.

4.3 统计量

4.3.1 统计量的定义

数理统计的任务是通过样本去推断总体. 而样本自身是一些杂乱无章的数字, 要对这些数字进行加工整理, 计算出一些有用的量. 可以这样理解: 这种由样本算出来的量, 把样本中与所要解决的问题有关的信息集中起来了. 我们把这种量称为统计量, 其定义如下:

由样本算出的量是**统计量**，或曰,**统计量**是样本的函数.

Definition

对这一定义我们作如下几点说明:

(1) 统计量只与样本有关, 不能与未知参数有关. 例如 $X \sim$

$N(a,\sigma^2)$, $X_1,\cdots,X_n$ 是从总体 X 中抽取的 i.i.d. 样本, 则 $\sum\limits_{i=1}^{n} X_i$ 和 $\sum\limits_{i=1}^{n} X_i^2$ 都是统计量, 当 a 和 σ^2 皆为未知参数时, $\sum\limits_{i=1}^{n}(X_i-a)$ 和 $\sum\limits_{i=1}^{n} X_i^2/\sigma^2$ 都不是统计量.

(2) 由于样本具有两重性, 即样本既可以看成具体的数, 又可以看成随机变量; 统计量是样本的函数, 因此统计量也具有两重性. 正因为统计量可视为随机变量 (或随机向量), 因此才有概率分布可言, 这是我们利用统计量进行统计推断的依据.

(3) 在什么问题中选用什么统计量, 要看问题的性质. 一般说来, 所提出的统计量应是最好的集中了样本中与所讨论问题有关的信息, 这不是容易做到的.

4.3.2 若干常用的统计量

1. 样本均值: 设 $X_1, \cdots, X_n$ 是从某总体 X 中抽取的样本, 则称

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

为**样本均值**. 它分别反映了总体均值的信息.

2. 样本方差: 设 $X_1, \cdots, X_n$ 是从某总体 X 中抽取的样本, 则称

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

为**样本方差**, 它分别反映总体方差的信息. 而 S 称为样本标准差, 它反映了总体标准差的信息.

3. 样本矩: 设 $X_1, \cdots, X_n$ 为从总体 F 中抽取的样本, 则称

$$a_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \cdots$$

为**样本 k 阶原点矩**, 特别 $k=1$ 时, $a_1=\bar{X}$ 即**样本均值**. 称

$$m_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k,\quad k=2,3,\cdots$$

为**样本 k 阶中心矩**.

4. 次序统计量及其有关统计量: 设 $X_1,\cdots,X_n$ 为从总体 F 中抽取的样本, 把其按大小排列为 $X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$, 则称 $(X_{(1)},X_{(2)},\cdots,X_{(n)})$ 为**次序统计量**, $(X_{(1)},\cdots,X_{(n)})$ 的任一部分也称为**次序统计量**.

利用次序统计量可以定义下列统计量:

(1) **样本中位数**:

$$m_{\frac{1}{2}}=\begin{cases}X_{(\frac{n+1}{2})} & \text{当 n 为奇数}\\ \frac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}] & \text{当 n 为偶数}\end{cases}\tag{4.7}$$

样本中位数反映总体中位数的信息. 当总体分布关于某点对称时, 对称中心既是总体中位数又是总体均值, 故此时 $m_{1/2}$ 也反映总体均值的信息.

(2) **极值**: $X_{(1)}$ 和 $X_{(n)}$ 称为样本的极小值和极大值. 极值统计量在关于灾害问题和材料试验的统计分析中是常用的统计量.

5. 经验分布函数: 假设总体分布 F 有矩, 由于不知道 F, 也就不知道矩, 现从该总体中抽出样本 $X_1, \cdots, X_n$, 我们可以使用如下函数估计 F:

$$F_n(x) = \{X_1, \cdots, X_n \text{中} \leq x \text{的个数}\}/n$$

它称为样本 $X_1, \cdots, X_n$ 的经验分布函数.

4.3.3 正态总体样本均值和样本方差的分布

为方便讨论正态总体样本均值和样本方差的分布, 我们先给出正态随机变量的线性函数的分布.

1. 正态变量线性函数的分布

设随机变量 $X_1, \cdots, X_n$ $i.i.d.$ $\sim N(a, \sigma^2)$, $c_1, c_2, \cdots, c_n$ 为常

数, 则有

$$T=\sum_{k=1}^{n}c_kX_k \ \sim \ N\left(a\sum_{k=1}^{n}c_k,\sigma^2\sum_{k=1}^{n}c_k^2\right)$$

特别, 当 $c_1=\cdots=c_n=1/n$, 即 $T=\frac{1}{n}\sum\limits_{i=1}^{n}X_i=\bar{X}$ 时, 有

$$\bar{X}\sim N(a,\sigma^2/n).$$

2. 正态变量样本均值和样本方差的分布

下述定理给出了正态变量样本均值和样本方差的分布和它们的独立性.

定理 1. 设 $X_1,X_2,\cdots,X_n$ *i.i.d.* $\sim N(a,\sigma^2)$, $\bar{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$ 和 $S^2=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2$ 分别为样本均值和样本方差, 则有

(1) $\bar{X}\sim N(a,\frac{1}{n}\sigma^2)$;

(2) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$;

(3) $\bar{X}$ 和 S^2 独立.

证: (1) 由推论 5.5.2 立得.

(2) 设

$$\mathbf{A} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

为一正交阵 (这一正交阵的存在性由 Schmidt 正交化方法保证), 作正交变换 $Y = AX$, 故有

$$Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \sqrt{n}\bar{X},$$

由正交变换保持向量长度不变可知

$$Y_1^2 + \cdots + Y_n^2 = X_1^2 + \cdots + X_n^2.$$

所以

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2. \tag{4.8}$$

由定理**??** 可知 $Y_i \sim N(\mu_i, \sigma^2),\ i = 2, \cdots, n.$ 再由 A 的行向量正交性可知

$$\mu_i = a\sum_{k=1}^{n} a_{ik} = \sqrt{n}a \cdot \sum_{k=1}^{n} \frac{1}{\sqrt{n}} \cdot a_{ik} = 0. \tag{4.9}$$

以及

$$\begin{aligned} Cov(Y_i, Y_j) &= E[(Y_i - EY_i)(Y_j - EY_j)] = E\left[\sum_{k=1}^{n} a_{ik}(X_k - a) \cdot \sum_{l=1}^{n} a_{jl}(X_l - a)\right] \\ &= \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}a_{jl}E[(X_k - a)(X_l - a)] = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}a_{jl}\delta_{kl}\sigma^2 \\ &= \sigma^2 \sum_{k=1}^{n} a_{ik}a_{jk} = \begin{cases} \sigma^2 & \text{当 } i = j, \\ 0 & \text{当 } i \neq j. \end{cases} \end{aligned}$$

此处 $\delta_{kl} = 1$, 当 $k = l$; 否则为 0. 因此 $Y_2, \cdots, Y_n$ i.i.d. $\sim N(0, \sigma^2)$. 故 $Y_i/\sigma \sim N(0, 1)$, $i = 2, \cdots, n$, 因此由 (4.8) 得

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=2}^{n} (Y_i/\sigma)^2 \sim \chi^2_{n-1}.$$

(3) 由上述 (2) 的证明中可知 $Y_1, Y_2, \cdots, Y_n$ 相互独立, S^2 只和 $Y_2, \cdots, Y_n$ 有关, $\bar{X}$ 只和 Y_1 有关, 因此 $\bar{X}$ 和 S^2 独立定理的证明超出我们的要求, 只要求记住这一结论.

4.3.4 几个重要推论

下面几个推论在正态总体区间估计和假设检验问题中有着重要应用.

推论 1. 设 $X_1, X_2, \cdots, X_n$ 相互独立相同分布 *(i.i.d.)* $\sim N(a, \sigma^2)$, 则

$$T = \frac{\sqrt{n}(\bar{X} - a)}{S} \sim t_{n-1}.$$

证: 由注 5.4.3 可知 $\bar{X} \sim N(a, \sigma^2/n)$, 将其标准化得 $\sqrt{n}(\bar{X} - a)/\sigma \sim N(0, 1)$. 又 $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, 即 $S^2/\sigma^2 \sim \chi^2_{n-1}/(n-1)$, 且 $\bar{X}$ 和 S^2 独立, 按定义有

$$T = \frac{\sqrt{n}(\bar{X} - a)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{\sqrt{n}(\bar{X} - a)}{S} \sim t_{n-1}.$$

推论 2. 设 $X_1, X_2, \cdots, X_m$ *i.i.d.* $\sim N(a_1, \sigma_1^2)$, $Y_1, Y_2, \cdots, Y_n$ *i.i.d.* $\sim N(a_2, \sigma_2^2)$, 且假定 $\sigma_1^2 = \sigma_2^2 = \sigma^2$, 样本 $X_1, X_2, \cdots, X_m$ 与 $Y_1, Y_2, \cdots, Y_n$ 独立, 则

$$T = \frac{(\bar{X} - \bar{Y}) - (a_1 - a_2)}{S_w} \cdot \sqrt{\frac{mn}{n+m}} \sim t_{n+m-2},$$

此处 $(n+m-2)S_w^2 = (m-1)S_1^2 + (n-1)S_2^2$, 其中

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2, \qquad S_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2.$$

证: 由注 5.4.3 可知 $\bar{X} \sim N(a, \sigma^2/m), \bar{Y} \sim N(a_2, \sigma^2/n)$, 故有 $\bar{X} - \bar{Y} \sim N\big(a_1 - a_2, (\frac{1}{m} + \frac{1}{n})\sigma^2\big) = N\big(a_1 - a_2, \frac{n+m}{mn}\sigma^2\big)$. 将其标准化得

$$\frac{\bar{X} - \bar{Y} - (a_1 - a_2)}{\sigma}\sqrt{\frac{mn}{m+n}} \sim N(0,1). \tag{4.10}$$

又 $(m-1)S_1^2/\sigma^2 \sim \chi_{m-1}^2$, $(n-1)S_2^2/\sigma^2 \sim \chi_{n-1}^2$, 再利用 χ^2 分布的性质可知

$$\frac{(m-1)S_1^2 + (n-1)S_2^2}{\sigma^2} \sim \chi_{n+m-2}^2. \tag{4.11}$$

再由 (4.10) 和 (4.11) 中 $(\bar{X}, \bar{Y})$ 与 (S_1^2, S_2^2) 相互独立, 由定义可知

$$\begin{aligned} T &= \frac{(\bar{X}-\bar{Y})-(a_1-a_2)}{\sigma}\sqrt{\frac{mn}{n+m}}\Big/\sqrt{\frac{(m-1)S_1^2+(n-1)S_2^2}{\sigma^2(n+m-2)}} \\ &= \frac{(\bar{X}-\bar{Y})-(a_1-a_2)}{S_w}\sqrt{\frac{nm}{n+m}} \sim t_{n+m-2}. \end{aligned}$$

推论 3. 设 $X_1, X_2, \cdots, X_m$ *i.i.d.* $\sim N(a_1, \sigma_1^2)$, $Y_1, Y_2, \cdots, Y_n$ *i.i.d.* $\sim N(a_2, \sigma_2^2)$, 且合样本 $X_1, X_2, \cdots, X_m$ 和 $Y_1, Y_2, \cdots, Y_n$ 相互独立, 则

$$F = \frac{S_1^2}{S_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1},$$

此处 S_1^2 和 S_2^2 定义如推论2所述.

证: 由注 5.4.3 可知 $(m-1)S_X^2/\sigma_1^2 \sim \chi_{m-1}^2$, $(n-1)S_Y^2/\sigma_2^2 \sim \chi_{n-1}^2$,

且二者独立, 由 F 分布的定义可知

$$F = \frac{\frac{(m-1)S_X^2}{\sigma_1^2}/(m-1)}{\frac{(n-1)S_Y^2}{\sigma_2^2}/(n-1)} = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1}.$$

证毕.

下列这一推论给出了服从指数分布随机变量的线性函数的分布与 χ^2 分布的关系. 这在指数分布总体的区间估计和假设检验问题中有重要应用.

推论 4. 设 $X_1, X_2, \cdots, X_n$ *i.i.d.* 服从指数分布: $f(x,\lambda) = \lambda e^{-\lambda x} I_{[x>0]}$, 则有

$$2\lambda n\bar{X} = 2\lambda \sum_{i=1}^{n} X_i \sim \chi_{2n}^2.$$

证: 首先证明 $2\lambda X_1 \sim \chi^2_2$. 因为

$$F(y) = P(2\lambda X_1 < y) = P\Big(X_1 < \frac{y}{2\lambda}\Big) = \int_0^{\frac{y}{2\lambda}} \lambda e^{-\lambda x} dx,$$

所以

$$f(y) = F^{'}(y) = \begin{cases} \frac{1}{2} e^{-\frac{y}{2}} & \text{当 } y > 0 \\ 0 & \text{当} y \leq 0. \end{cases}$$

因此 $f(y)$ 即为自由度为 2 的 χ^2 密度, 即 $2\lambda X_1 \sim \chi^2_2$.

再利用 χ^2 分布的性质 (3), $2\lambda X_i \sim \chi^2_2,\ i = 1, 2, \cdots, n$; 又它们相互独立, 故有 $2\lambda \sum_{i=1}^{n} X_i \sim \chi^2_{2n}$.

4.4 总结

获得有效数据后, 统计推断问题可以按照如下的步骤进行:

1. 确定用于统计推断的合适统计量;
2. 寻求统计量的精确分布; 在统计量的精确分布难以求出的情形, 可考虑利用中心极限定理或其它极限定理找出统计量的极限分布.
3. 基于该统计量的精确分布或极限分布, 求出统计推断问题的精确解或近似解.
4. 根据统计推断结果对问题作出解释.

其中第二步是最重要, 但也是最困难的一步. 正态总体下样本均值和样本方差的分布, 在寻求与正态变量有关的统计量精确分布时, 起着十分重要作用. 尤其在后面两章中求区间估计和假设检验问题时可以看得十分清楚.