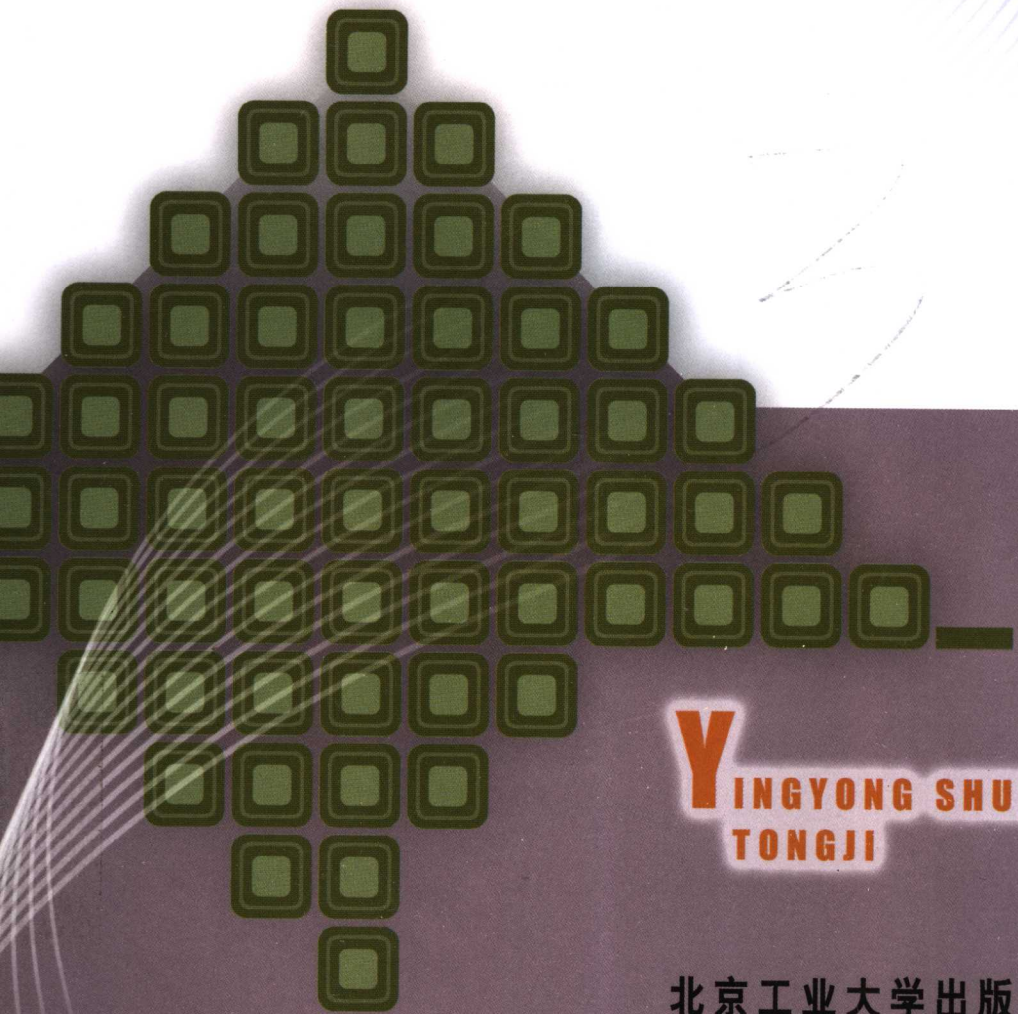
应用数理统计

杨振海 张忠占 主编



YINGYONG SHULI TONGJI

北京工业大学出版社

应用数理统计

杨振海　张忠占　主编

北京工业大学出版社

图书在版编目（CIP）数据

应用数理统计/杨振海，张忠占主编. —北京：北京工业大学出版社，2005.8
ISBN 7-5639-1564-8

Ⅰ.应... Ⅱ.①杨... ②张... Ⅲ.数理统计 Ⅳ.0212

中国版本图书馆 CIP 数据核字（2005）第 077257 号

应用数理统计

杨振海 张忠占 主编

*

北京工业大学出版社出版发行
各地新华书店总经销
徐水宏远印刷厂印刷

*

2005 年 9 月第 1 版 2005 年 9 月第 1 次印刷
787mm×960mm 16 开本 18.75 印张 315 千字
印数：1~3000 册
ISBN 7-5639-1564-8/G·794
定价：28.00 元

内 容 提 要

本书主要介绍了数理统计学的基本知识，内容包括数理统计的基本概念、参数估计、假设检验、回归分析以及方差分析。在保持严谨叙述的同时，本书注重从直观上讲解数理统计的基本概念、基本结论，以便于读者尽快抓住这些内容的要旨。阅读本书时，读者需要具备基本的数学分析、线性代数和概率论的知识。为方便读者，本书在附录中列出了一些基本的概率论知识，以此作为不同背景的读者在阅读本书时的参考。

本书是为数学类专业的本科生编写的数理统计课程的教材，也可作为其他专业本科生或各类专业的研究生学习数理统计时的参考。

序　言

本书是作者多年来在数学类专业讲授数理统计课程的教学成果。北京工业大学数学类专业自20世纪70年代末开设数理统计课程以来，至今已有25年了。在此期间，数理统计课程组针对不同学生的特点、不同的学时和教学要求，注意总结经验，并积累了一些教学素材。虽然在当时曾经由北京工业大学出版社出版过一本数理统计的教材，但近年来，我国高等教育的形势发生了巨大的变化，高等教育逐步走向大众化。这对数理统计这样一门具有丰富应用特点背景的课程提出了新的要求。

随着高等教育向大众化的发展，学生的学习需求更加多样化。同时，由于素质教育的需要，一些高校专业课程的学时有所压缩。为此，我们本着打好基础、加强应用能力培养的精神，经过不断探索，在原来教材的基础上进行了重大的修订，逐渐形成了本书基本架构。在理论方面，本书既照顾到叙述的严谨，又尽量避免复杂的推演，力求把某些理论上比较难的证明放到选学内容；在应用方面，本书强调数理统计的概念和思想方法与现实问题的联系，重视从生活中通过一定的哲学抽象而产生统计思想的过程，并通过众多的例题和习题体现数理统计的应用。本书在每章后面均有一定数量的习题，并按各节的内容对应列出。个别稍有难度的习题前加了*号。

在读者具备了基本概率论相关知识的前提下，讲授本书大约需要72学时。其中基本的内容可以用60学时讲完，但不包括无偏检验、非参数检验、回归模型的选取与改进（章节标题前加了*号）等内容。本书出版前曾经作为讲义试用过3年。

在本书的写作过程中，得到了北京工业大学概率统计学科部多位老师的协助，并得到了北京工业大学出版基金的支持，在此一并致谢。由于作者水平所限，书中错讹之处在所难免，欢迎广大读者批评指正。

编　者

2005年7月

目　录

第 1 章　数理统计的基本概念

1.1　引言

数理统计是一门以概率论作为工具，研究如何分析带有随机性影响的数据的学科，包括如何有效地收集数据、如何用适当的方法分析数据，并通过数据分析的结果认识所研究对象的性质或规律。为了进一步理解数理统计学科的研究范畴，考察下面几个例子。

例 1.1.1　某工厂生产一批灯泡共 M 只，其中有次品 m 只（假设寿命小于 1 000 h 的灯泡为次品）。一般来说 m 是未知的，因此次品率 $p=m/M$ 也未知。现在某销售单位要从工厂进货，购买这批灯泡，因而有必要了解该批灯泡的次品率 p。要确切知道 p，必须对这 M 只灯泡逐一进行寿命试验。这是不现实的，不仅因为这样做工作量巨大，同时也因为试验是破坏性的，即经过试验的灯泡已经不能再使用了。因此，一般只能从 M 只灯泡中抽出一部分，比如选 100 只灯泡进行寿命试验，测量出它们的寿命，然后根据这 100 只灯泡的寿命数据，对整批灯泡的次品率 p 进行估计。

更进一步，显然用于试验的灯泡越多，越能把 p 估计得准确。那么，在试验费用一定或给定估计精度的条件下，应该抽取多少只灯泡进行试验也是需要决定的问题。

例 1.1.2　某公司为提高糖化饲料中所含还原糖量，研究温度、时间、pH 值和投曲量对还原糖量的影响。根据经验，确定这 4 个因素在试验中的取值如表 1.1 所示。

这 4 个因素不同取值的搭配共有 $4\times4\times4\times2=128$ 种。因受实际情况限制，不能对每种搭配的生产条件都进行试验，要从全体 128 种搭配中挑出一部分作为试验性生产条件。这里，当然希望被选出的搭配条件具有代表性，最好还应便于分析试验结果。

表 1.1 4 个因素试验取值表

水 平	因 素			
	温度/℃	时间/h	pH 值	投曲量/(%)
1	15	12	7	5
2	20	24	6	10
3	25	36	5	—
4	30	48	4	—

例 1.1.3 在歌唱比赛中，有 10 名评委给歌手打分，并根据打分的结果确定歌手的水平。所得 10 个分数为 x_1，x_2，…，x_{10}。一般可以采用平均数 $\overline{x}=\left(\sum_{i=1}^{10}x_i\right)/10$ 作为这名歌手的水平的估计值，但也可采用去掉这 10 个分数中的最高分和最低分，然后求平均的方法，或采用去掉最小的 4 个分数和最大的 4 个分数，以剩下的 2 个分数的平均作为歌手得分的评分方法。这些方法是否具有合理性？在这些方法中，哪一种方法的效果更好些？这都是应研究解决的问题。

例 1.1.4 某公司希望了解潜在顾客对于某类产品的消费意向，为此进行调查。如何确定调查方案以及如何从调查所得到的数据理解顾客的消费偏好是调查成功的基础。

在以上这些例子中，解决问题的方法首先是采集数据。要了解 p，先得对 100 只灯泡进行寿命试验，得到 100 个寿命数据；要分析影响还原糖量的因素，必须先对几个因素的搭配做试验，获取在这些搭配的生产条件下的还原糖量；要确定歌手的水平，就要先给歌手打分；要了解顾客的消费意向，就要先进行调查，得到有关的数据。

为了达到解决问题的目的，采集数据时一般不能随意进行。比如，在例 1.1.1中，希望挑出来进行试验的灯泡要有代表性，因而随意拿一些灯泡进行试验的方法是不可取的。一个简单的原则是随机抽取，形象地说就是抓阄。另一方面，试验前应该依据试验的经费条件或精度要求确定出要抽取多少只灯泡。在例 1.1.2中，由于只挑选一部分生产条件做试验，希望在挑选出来的条件中各个因素的不同取值都能出现，并且具有可比较性。在例 1.1.4中，必须确定调查对象（如潜在顾客群是老年人群还是青年人群），调查时间和调查方式（如街头调查还是邮寄问卷），确保被调查到的人员在调查对象中具有代表性。

数据的采集方法和过程也影响到数据的分析方法，所以数据的采集应有利于数据的分析处理。采集数据的这些要求的重要性推动了人们在这个方面

的研究，形成了诸如试验设计（针对类似于例 1.1.2的问题）、抽样调查（针对类似于例 1.1.4的问题）等统计学的分支。

采集到的数据是带有随机色彩的。在例 1.1.1 的问题中，用来采集寿命数据的 100 只灯泡其选取是随机的，这 100 只灯泡的寿命数据与另 100 只灯泡的寿命数据未必相同。因此，得到的这 100 个寿命数据是带有随机性的。另外，在寿命试验的过程中，由于环境因素的变化，比如电压的高低波动等也可能会影响灯泡的寿命，从而使得到的数据受到随机性的影响。由此可知，数据中包含的随机性至少有两方面的原因：一是由于在全部对象中选取一部分进行观察时，因代表性等方面的需要，采取了随机的方式；另一方面，在数据采集中，由于某些未加控制的因素或周围环境因素的偶然变化而使得到的数据带有随机性。

数据的获取过程就是概率论中所讨论的随机试验。在概率论中，用概率模型来刻画各种随机试验，在那里主要用模型来描述随机试验的特性。然而，在许多问题中，需要利用随机试验中得到的数据进一步对于模型中未知部分做出推断，甚至于需要用所得到的数据建立模型。比如，在例 1.1.1 中，如果采用独立重复的有放回随机抽样，就可以用二项分布模型 $B(100, p)$来描述该试验中次品数的分布规律，但这里的问题在于，需要先依据这 100 只灯泡的试验结果把次品率 p 估计出来。

为了基于这种带有随机性影响的数据进行合理地推断或预测，需要用概率论的的方法对数据的获取过程进行尽可能精确的描述，建立模型，并在此基础上研究适当的统计推断方法，就所关心的问题作出尽可能精确、可靠的结论。不同的推断方法往往有不同的特点，比如在例 1.1.3 中，由于打分具有一定的随机性，$x_1, x_2, \cdots, x_{10}$是一组带有随机性的数据，因此基于它们对歌手水平的推断也有随机性。一般说来，用同样一种推断方法，这 10 次打分的结果给出的估计值与另外 10 次打分结果给出的估计值会有一些不同。同时，不同的推断方法从数据中提取信息的方式不同：10 个分数平均的方法充分地利用了各个评委的意见，而去掉最高分和最低分的方法易于排除某个评委的特殊偏好，因此这些方法具有各自不同的性质。如何从试验的特点出发，找到适合的推断方法，是应该解决的重要问题。

以上这些问题，即如何合理有效地收集数据、如何用数据做出合理的推断，以及研究各种推断方法的性质都是数理统计所面临的任务。数理统计在收集数据或分析数据的工作中发挥着重要的作用，因此有着广泛的应用。数理统计在不同领域的应用，或者形成了具有特色的统计方法以至统计学分支

(如生物统计学)，或者形成了新的学科（如计量经济学和数量遗传学等)。因而数理统计学内容丰富，发展迅速。本书作为一本基础教材，主要论述数理统计的基本概念和一些基础知识，而不涉及深入的专门内容。

1.2 基本概念

1.2.1 总体和样本

通常，把所研究问题涉及到对象的全体称为**总体**，把总体中的每个元素称为**个体**。例 1.1.1 中，考察 100 000 只灯泡的质量，这 100 000 只灯泡的全体就是总体，而其中的每一只灯泡都是个体。又比如，要研究一群人的身高与体重，那么，这群人就是总体，这群人中的每一个人都是个体。

然而，在这些统计问题中，真正关心的并不是这些个体本身，而只是关心个体的某些指标及这些指标在总体中的分布。在身高与体重的例子中，这群人中具体某一个人是男是女，是年长还是年幼等都不是研究人员所关心的，这里仅仅关心他的身高 H 和体重 W，以及身高和体重这 2 个指标在这群人中的分布情况。常用 X（标量）或 $\boldsymbol{X}$（向量）来记这些指标，此例中 $\boldsymbol{X}=(H, W)^{\mathrm{T}}$（这里及以后 T 表示转置)。而例 1.1.1 中的指标 X 就是灯泡的质量（次品或合格品)。由于在数理统计中只关心指标 X，通常把所有个体指标 X 取值的全体与总体等同看待。若用 $X=0$ 表示合格品，$X=1$ 表示次品，那么例 1.1.1 中的总体就是由 $100\,000-m$ 个 0 和 m 个 1 组成的“集合”，这里 m 为这批灯泡中次品的个数。这个总体的表示有些臃肿。事实上，根据研究的目的，只要知道 X 取 0，1 两个值之一，以及取 1 的比例 p 就够了，而这恰是两点分布 $B(1, p)$所表示的，即从所有个体中随机抽取一个个体，其指标 X 的概率分布。由此，用 $B(1,p)$ 来表示产品品质指标 X 的总体。同样，在一群人身高与体重的问题中，总体也指这群人中指标 $\boldsymbol{X}=(H,W)^{\mathrm{T}}$ 取值的分布。这个分布表明了，身高 1.70 米以下、体重 60 公斤以下的人的比例有多大，身高 1.75 米以下、体重 61 公斤以下的人的比例有多大等所有这样的信息，以及身高和体重两个变量之间的关系。

总体的抽象含义是指**所感兴趣的指标及其取值的分布**，这个分布又称为**总体分布**。总体分布与对指标进行一次随机观测所对应的随机变量的概率分

布相同，这里所说的随机观测的随机是指在进行观测时所有个体被观测到的可能性都一样。由于这个原因，本书对这二者不加区别。

总体（总体分布）是对客观研究对象的指标取值情况的数学描述。如 1.1 节中所说，数理统计的重要任务之一是通过对总体中个体的观测结果来推断总体的特性，达到了解总体的目的。在本书中，主要讨论一维总体。

通过观察或试验的方法，获得总体中一部分个体的指标 X 的取值，称之为**样本**。如检验 n 只灯泡是否为次品可得 n 个数据 x_1，x_2，…，x_n，它们的取值为 0 或 1。(x_1，x_2，…，x_n) 就是灯泡质量指标总体的一个样本。对于具体某一次试验而言，抽取 n 个个体进行观测，得到的 (x_1，x_2，…，x_n) 是一组确定的数，但在另一次试验中可能得到另一组数。在实际问题中，利用观测到的数据 (x_1，x_2，…，x_n) 对于总体的某些未知特性进行推断，得到的是这些特性的具体的值，比如 100 个灯泡中若有 3 个次品，那么可以把这批灯泡的次品率估计为 3%。但在数理统计学研究中，讨论的问题侧重于一个推断方法的总体表现。换句话说，首先关注上面的 3% 是怎么从数据中计算出来的，其次必须考虑这个方法的优良性。为此，需要考虑对于所有可能的观测或试验结果，这个算法得到结果的性质怎么样。这就涉及到观测结果的分布。因此，在研究统计方法时，把观测或试验的结果看成随机向量 $(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，这就是**样本的二重性**。今后，称一组具体的试验值 $(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为样本 $(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的**观察值**，而在提到样本时总是指作为随机向量的 $(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$。$n$ 称为**样本大小**或**样本容量**。

样本作为随机向量有其概率分布，称其为**样本分布**。显然，样本分布取决于总体的性质和样本的获取方法。一种常见的样本是简单随机样本，它有两个特点：(1) 同分布，即样本的每个分量 X_i 都与总体 X 有相同的分布；(2) 独立性，即样本中的 n 个个体的选取是完全独立的，或者说 X_1，X_2，…，X_n 是相互独立的随机变量。满足 (1) 和 (2) 的样本称为**简单随机样本**，简记为 X_1，X_2，…，X_n i.i.d. $X_1 \sim X$。若总体 X 的分布函数为 $F(x)$ (或概率密度函数为 $f(x)$)，也记做 $X_1 \sim F(x)$(或 $X_1 \sim f(x)$)。

在实施抽样时，要得到简单随机样本，需要保证以下两点：(1) 总体中每个个体都有同等可能性被选中作为样本中的个体；(2) 每一次抽取不改变下一次抽取的抽取环境，也不影响下一次抽取的抽取方式。由此，对于有限总体而言，需要进行有放回独立重复抽样。在实际问题中，当总体很大时，少量个体的抽取对于抽取环境的改变往往可以忽略，可用无放回独立重复抽样近似代替。

本书如不作特殊声明，所提到的样本都为简单随机样本。简单随机样本的样本分布可由总体分布完全确定。例如，若 $X_1,X_2,\cdots,X_n$ i.i.d. $X_1 \sim f(x)$，则样本 $(X_1,\cdots,X_n)^{\mathrm{T}}$ 的联合分布由 $\prod_{i=1}^{n} f(x_i)$ 给出，这里 $f(x)$ 为概率密度函数（当 X 为连续型随机变量时）或概率函数（当 X 为离散型随机变量时）。

例 1.2.1 要估计一物体的质量 a，用天平将物体重复测量 n 次，结果记为 X_1，X_2，$\cdots$，X_n。假定各次测量是相互独立的，并且结果具有相同分布 $N(a,\sigma^2)$，试确定样本 $(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的分布。

解： 总体的概率分布为 $N(a,\sigma^2)$，概率密度为

$$f_1(x;a,\sigma^2)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$$

其中 a 为物体的质量，σ^2 反映天平的精度。故 $(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 的概率密度为

$$\begin{aligned}\prod_{i=1}^{n} f_1(x_i;a;\sigma^2)&=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(x_i-a)^2}{2\sigma^2}\right)\\&=\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^{n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right)\end{aligned}$$

例 1.2.2 设某电子元件的寿命 X 服从指数分布 $\varepsilon(\lambda)$，其概率密度为

$$f_2(x,\lambda)=\begin{cases}\frac{1}{\lambda}\mathrm{e}^{-\frac{x}{\lambda}} & x>0\\ 0 & \text{其他}\end{cases}$$

其中 $\lambda>0$。今从一批产品中独立地抽取 n 件进行寿命试验可得寿命数据 $X_1,X_2,\cdots,X_n$，求样本 $(X_1,\cdots,X_n)^{\mathrm{T}}$ 的概率分布。

解： 依题意有 $X_1,X_2,\cdots,X_n$ i.i.d. $X_1 \sim f_2(x,\lambda)$，故所求的概率密度为

$$\prod_{i=1}^{n} f_2(x_i;\lambda)=\begin{cases}\frac{1}{\lambda^n}\exp\left(-\frac{1}{\lambda}\sum_{i=1}^{n}x_i\right) & x_1,x_2,\cdots,x_n>0\\ 0 & \text{其他}\end{cases}$$

1.2.2 参数空间和分布族

例 1.2.1 中总体分布中的 a 与 σ^2 是确定分布的常数，即这些常数的不同

值对应于不同的总体分布，只要知道了它们的值，就可以把相应的总体分布确定下来。例 1.2.2 中总体分布为指数分布 $\varepsilon(\lambda)$，λ 是确定总体分布的常数。在数理统计中，称出现在总体分布中的常数为**参数**。因此 a，σ^2，λ 都是参数。a 和 a^2 出现在同一分布中，可记为 $(a,\sigma^2)^{\mathrm{T}}$，称为参数向量。若参数的值未知，称其为未知参数。例 1.2.1 中 σ^2 是否为未知参数，要看研究人员对天平精度的了解程度。若对天平精度足够了解可以给出 σ^2 的值，则 σ^2 就是已知参数；若对天平精度不够了解，无法给出 σ^2 的值，甚至于抽样的目的就是要估计这个精度，那么 σ^2 就是未知参数。参数所有可能的取值构成的集合称为**参数空间**。例 1.2.1 中，由于 a，σ^2 都是参数，参数空间为 $\Theta=\{(a,\sigma^2)^{\mathrm{T}}:a>0,\sigma^2>0\}$。例 1.2.2 中的参数空间为 $\Theta=\{\lambda:\lambda>0\}$。

由于不同的参数值一般对应于不同的总体分布，因此，参数空间中所有可能的参数取值对应于一族总体分布，称该分布族为**总体分布族**。同样，不同的参数值也对应不同的样本分布，这些分布构成**样本分布族**。例 1.2.1 中，若 a，σ^2 都为未知参数，则总体分布族为 $\{f_1(x;a,\sigma^2):a>0,\sigma^2>0\}$，样本分布族为 $\{\prod_{i=1}^{n} f_1(x_i;a,\sigma^2):a>0,\sigma^2>0\}$。例 1.2.2 的总体分布族为 $\{f_2(x,\lambda):\lambda>0\}$，样本分布族为 $\{\prod_{i=1}^{n} f_2(x_i;\lambda):\lambda>0\}$。

一般地，设总体概率密度函数（或概率函数）为 $f(x,\theta)$，其中含有未知参数 θ（可能是由多个参数构成的向量），参数空间为 Θ，则总体分布族为 $\{f(x,\theta):\theta\in\Theta\}$，样本分布族为 $\{\prod_{i=1}^{n} f(x_i,\theta):\theta\in\Theta\}$。

样本分布族及其参数空间给出了所考虑的总的范围，反映了对所研究问题的了解程度及对抽样方式的规定。今后，称样本分布族为统计模型。在例 1.2.1中，统计模型给出总体的分布形式为正态分布，只有参数 a，σ^2 未知。因此，只要对 a 和 σ^2 作出推断，就等于对总体作出了推断。例 1.2.2 中，模型规定了总体为指数分布，只有参数 λ 未知，所以对总体的统计推断归结为对 λ 的推断。类似这种总体分布形式已知，只需要对一些未知参数作出推断的问题称为**参数统计问题**。另有一类问题，不仅总体的分布参数未知，而且对总体分布函数的形式也不是完全了解（可能只知道是离散型或连续型，对称分布或偏态分布等），甚至于统计推断的目的就在于确定总体的分布形式，这类问题称为**非参数统计问题**。