交通数据统计分析理论与方法

JIAOTONG SHUJU TONGJI FENXI——LILUN YU FANGFA

邵长桥 编著

人民交通出版社
China Communications Press

Jiaotong Shuju Tongji Fenxi Lilun yu Fangfa

交通数据统计分析理论与方法

邵长桥　编著

人民交通出版社

内 容 提 要

本书共分11章，主要内容包括：交通数据分析统计学基础、交通工程中常用统计分布、参数估计和假设检验方法、交通数据分析初步、相关性分析和线性回归分析、广义线性回归分析和非线性回归分析等。

本书可作为交通工程、交通规划、交通运输和交通管理专业研究生的教学用书，也可作为交通运输工程领域的教学、科研、管理人员的参考书。

图书在版编目(CIP)数据

交通数据统计分析理论与方法/邵长桥编著. --北京：人民交通出版社，2012.9

ISBN 978-7-114-09962-5

Ⅰ.①交… Ⅱ.①邵 Ⅲ.①交通工程—数据—统计分析 Ⅳ.①U491

中国版本图书馆CIP数据核字(2012)第169140号

书　　名：交通数据统计分析理论与方法
著 作 者：邵长桥
责任编辑：任雪莲
出版发行：人民交通出版社
地　　址：(100011)北京市朝阳区安定门外外馆斜街3号
网　　址：http://www.ccpress.com.cn
销售电话：(010)59757969、59757973
总 经 销：人民交通出版社发行部
经　　销：各地新华书店
印　　刷：北京市密东印刷有限公司
开　　本：787×1092　1/16
印　　张：10.5
字　　数：237千
版　　次：2012年9月　第1版
印　　次：2012年9月　第1次印刷
书　　号：ISBN 978-7-114-09962-5
定　　价：50.00元

前　言

笔者在教学过程中发现，尽管我国目前出版了很多关于数据统计分析的书籍，但这些书籍多偏重于理论或针对某些专业所编写，不适合交通工程专业的学生学习使用。为了将数理统计和交通工程中的统计分析问题相融合，笔者一直希望撰写出一本适合交通工程、道路和交通管理专业的学生使用的数据统计分析教材，同时也为从事汽车运输专业和数据分析人员提供一本实用的参考书。

本书围绕着交通分析中常见的数据分析问题，介绍了相关的数据统计分析理论和方法，并引入了一些实例来说明如何应用统计方法解决和分析交通问题。全书共11章。第1章介绍了数据分析的统计学基础；第2章介绍了交通工程中常用的统计分布模型，讨论了统计分布在交通工程中的应用；第3章和第4章分别介绍了参数估计和统计假设检验问题；第5章首先介绍了数据分析的图表方法和统计量方法，其后则对数据异常值处理及数据变换方法进行了介绍；第6～11章介绍了常用的统计分析模型，包括相关分析、一元线性回归模型、多元线性回归模型、广义线性回归模型、Logistic回归模型和非线性回归模型。

本书不仅介绍了各种统计分析方法的背景和实际意义，还列举了各种应用实例，将理论、方法与实践结合在一起，易于读者使用。这是本书的一大特点。

本书在整理的过程中参阅了大量国内外著作、学位论文和有关文章，有的文献可能由于疏忽遗漏未能在参考文献中列出，在此谨向本书所直接或间接引用的研究成果的作者表示深切的谢意。本书的编写，得益于许多人的帮助。感谢我的学生们，他们的学习热情和富有启发的提问促使我多次对讲义进行了修改并整理成此书。还要特别感谢北京工业大学的任福田先生和杨振海先生的鼓励，两位先生不仅是我的授业恩师，也对此书的出版给予了大力的支持。

本书可以作为交通工程专业研究生教材，也可以作为汽车运输专业、土木工程专业的教学参考书。同时可供城市交通规划、交通运输、公共交通和交通管理部门的技术人员参考。

由于作者水平有限，书中难免存在不足之处，恳请读者批评指正。

编　者

2012年5月

目　　录

第 1 章　绪论 …… 1
§1.1　数据分析的统计学基础 …… 1
§1.2　顺序统计量和经验分布 …… 4
§1.3　总体分位数和样本分位数 …… 5
§1.4　抽样分布 …… 6
思考题 …… 8
第 2 章　常用统计分布及其应用 …… 9
§2.1　离散分布 …… 9
§2.2　连续型分布 …… 12
§2.3　统计分布在交通工程中的应用 …… 18
思考题 …… 23
第 3 章　参数估计方法 …… 25
§3.1　基本概念 …… 25
§3.2　矩法及矩估计 …… 27
§3.3　极大似然估计 …… 29
§3.4　贝叶斯估计和经验贝叶斯估计 …… 32
§3.5　区间估计 …… 35
思考题 …… 40
第 4 章　分析数据的统计检验 …… 41
§4.1　基本概念 …… 41
§4.2　t 检验 …… 43
§4.3　U 检验与基于大样本理论的检验 …… 46
§4.4　F 检验——两总体方差比较 …… 47
§4.5　分布的拟合优度检验 …… 48
§4.6　正态性检验 …… 51
思考题 …… 55
第 5 章　数据分析初步 …… 56
§5.1　单样本数据汇总分析 …… 56
§5.2　两样本数据汇总分析 …… 63
§5.3　异常值的处理 …… 65
§5.4　数据变换 …… 69
思考题 …… 70

第 6 章　相关性分析 …… 71
§6.1　引言 …… 71
§6.2　线性相关 …… 72
§6.3　秩相关分析 …… 74
§6.4　偏相关分析 …… 78
§6.5　复相关分析 …… 79
§6.6　典型相关分析 …… 79
思考题 …… 82
第 7 章　一元线性回归模型 …… 84
§7.1　一元线性回归模型与参数估计 …… 84
§7.2　模型参数分析 …… 87
§7.3　模型假设检验 …… 89
§7.4　模型预测精度的度量 …… 92
§7.5　预测置信区间 …… 93
§7.6　预测实例 …… 95
思考题 …… 97
第 8 章　多元线性回归分析 …… 99
§8.1　多元线性回归模型 …… 99
§8.2　模型参数估计及其性质 …… 100
§8.3　多元线性回归模型假设检验 …… 100
§8.4　预测置信区间 …… 103
§8.5　自变量的选择 …… 103
§8.6　共线性诊断 …… 107
§8.7　建模过程中注意的几个问题 …… 109
思考题 …… 113
第 9 章　广义线性回归分析 …… 114
§9.1　引言 …… 114
§9.2　广义线性模型 …… 114
§9.3　广义线性模型参数估计和检验 …… 117
§9.4　广义线性模型选择 …… 118
§9.5　广义线性模型在交通工程中的应用 …… 119
思考题 …… 120
第 10 章　Logistic 回归分析 …… 121
§10.1　Logistic 线性回归模型 …… 121
§10.2　Logistic 回归模型参数估计 …… 122
§10.3　模型假设检验 …… 125
§10.4　多项 Logistic 回归模型 …… 129
思考题 …… 131

第 11 章　非线性回归分析 …… 132
§11.1　引言 …… 132
§11.2　非线性回归模型 …… 132
§11.3　非线性回归模型参数估计 …… 135
§11.4　建模过程中常见问题 …… 136
§11.5　实例分析 …… 137
思考题 …… 139
附录 …… 140
附表 1　标准正态分布函数表 …… 140
附表 2　t 分布临界值(t_α)表 …… 142
附表 3　χ^2 分布分位数表 …… 144
附表 4　F 分布的分位数表 …… 146
附表 5　柯尔莫哥洛夫检验的临界值表 …… 148
附表 6　计算 W 的系数$\{a_{n+1-i}\}$(正态性检验) …… 150
附表 7　W 统计量分位数(正态性检验) …… 153
附表 8　爱泼斯-普利(Epps-Pully)检验:检验统计量 T_{EP}的分位数表 …… 155
参考文献 …… 156

第1章 绪 论

交通数据中往往含有一定的信息，这些信息可以是交通系统自身运行的特征和规律，也可以是人们的交通行为或经济行为在交通活动中的反映。交通科研人员日渐重视如何正确使用统计分析方法来挖掘数据中蕴涵的信息，并对分析结果给予正确解释的问题。

本章将主要介绍交通数据统计分析基础知识和一些常用的定义。

§1.1 数据分析的统计学基础

1.1.1 随机变量

关于随机变量在经典的概率论与数理统计中都有定义。其实质就是对于一次试验（或观测），由于受众多因素或试验本身误差的影响，其结果或观测值具有一定的随机性。如某个信号交叉口进口车道一个信号周期内到达的车辆数为 X，由于交通需求是变化的，在没有观测之前，不能确定 X 的具体取值；又如，经过路段上某点的车辆运行速度 v，由于不同的驾驶员驾驶习惯不同或其他车辆的干扰等原因，每辆车运行的速度在没有测量之前是不确定的。

交通工程中常用的随机变量主要有两种类型：离散型随机变量和连续型随机变量。离散型随机变量只能取有限或可数个值，如某条段路上一定时间内发生的交通事故数或交通事故死亡人数只能是整数值，它们都是离散型的随机变量。连续型随机变量取值则布满某个区间，如车头时距和车辆运行速度等，这类变量的特点是其取值可为某个区间上的任何值。

随机性产生的原因主要有两点：交通现象本身的随机性和观测误差的存在。例如，在对驾驶人的反应时间进行测量的试验中，不同的驾驶人由于生理、心理条件的不同，测量的结果是不同的。这就是自身的随机性。在相同的道路或交通条件下，应用相同的调查仪器对某个交通参数进行重复观测，每次测量的结果可能是不同的，部分原因就是存在观测误差。

1.1.2 总体和样本

在数理统计中，一般把研究对象的全体称为总体（或母体）。把总体中每个成员称为总体中的个体。例如，在研究驾驶人的行为特性时，则所有的驾驶人就是要研究的总体，而每一个驾驶人就是一个个体。在实际研究中，虽然研究的问题是针对总体的特性，但我们往往并不关心每个个体的所有的特殊属性，而只是关心某个（些）指标 X 及这些指标在总体上的分布特征。如研究驾驶人的反应特性，测量的指标只是驾驶人的反应时间，而对于驾驶人的收入情况、是否结婚都不关心。对于每个个体，指标 X 是确定的，但对总体中不同的个体而言，X 是不同的。因此，对总体而言，指标 X 是随机变量（或随机向量）。如果 X 的分布函数为 $F(x)$，则称 $F(x)$ 为总体分布。由于总体特性可以用其分布来刻画，因此，通常把总体分布与总体视为同义词。

用观测或试验的方法从总体中获取的总体中一部分个体称为样本。一般用大写字母表示样本，用小写字母表示样本的取值（观测值），如用$(X_1, X_2, \cdots, X_n)$表示来自同一个总体的一组样本，其观测值记为$(x_1, x_2, \cdots, x_n)$，n称为样本量。对于具体的一次观测（或试验），$(x_1, x_2, \cdots, x_n)$是一个确定的值；但对于另一个观测（或试验），其可能取另一个值。因此，在一般教科书中，往往把$(X_1, X_2, \cdots, X_n)$看作随机向量。在本书中，在不引起混淆的情况下，两者不加区别。

由上述可知，样本是随机变量。样本的概率分布就称为样本分布。样本的分布取决于总体的性质和样本获取方法。当获取的一个样本$(X_1, X_2, \cdots, X_n)$，其每个分量X_i的分布都与总体X有相同的分布，并且$X_1, X_2, \cdots, X_n$是相互独立的（样本中个体的选取是完全独立的），则称$(X_1, X_2, \cdots, X_n)$为总体X的（一个）简单随机样本，简记为$X_1, X_2, \cdots, X_n$ $iid.\ X_i \sim X$。如果总体X的分布为$F(x)$［密度为$f(x)$］，也记为$X_1, X_2, \cdots, X_n \overset{iid}{\sim} F(x)$［或$X_1, X_2, \cdots, X_n \overset{iid}{\sim} f(x)$］。

一般情况下，本书所提到的样本都为简单随机样本。即如无特殊说明，则对给定一个样本假设，它们是相互独立同分布的，并简单表述为：$X_1, X_2, \cdots, X_n$为来自总体$X \sim F(x)$的一个样本。

在交通实践中，由于随机变量的测量结果和记录方式不同，其测量值通常有四种表现形式（数据类型）：计量数据、计数数据、名义数据和有序数据。

计量数据：随机变量的测量值可以是某个区间内任意一个实数，例如，车辆运行速度和驾驶人的反应时间。

计数数据：就是只能在整数范围内取值。如一定统计时间段内通过某道路（车道）断面的车辆数；某路段上在一定时间内发生的事故数等。

名义数据：测量结果不是数，而是事物的属性，如人的性别、婚姻状况、路面的干燥状况等。在进行数据分析时，为了分析（或统计）的便利，常用数来表示属性的分类，例如用“0”和“1”分别表示男和女；在选择行为中，“1”和“0”分别表示接受和拒绝。名义数据之间没有内在的次序，也没有数值距离，只起一个名义的作用，不表示大小关系，更不能进行运算。

有序数据：在实际测量中，有时需要用到表征事物的次序关系的定性变量，通常的做法是采用数值来表示排序信息。如评价道路设施的服务水平，用1、2、3、4等表示服务水平的质量。与次序对应的数值只是反映某一特定属性上的排序，数值之间距离并不相等，这些数只起一个顺序作用。这类数据称为有序数据。

计量数据和计数数据称为定量数据，这些数据具有一定的物理意义，可以进行各种运算。**名义数据和有序数据称为定性数据**（或属性数据），这些数据只是一种代码，不表示实际数量上的含义，不能进行数学运算。正是由于定量数据和定性数据存在着这种差异，决定了定量数据和定性数据统计分析方法的不同。因此，在进行数据分析时，需要注意测量数据类型，并选择合适的分析方法。

1.1.3 参数和分布族

在数理统计中，称出现在分布中的常数为参数；如果不知道其值时称为未知参数。例如，正态分布密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1-1}$$

式中，μ、σ^2 为常数，就是参数。由于式(1-1)中，参数 μ、σ^2 取不同值时，式(1-1)都对应着一个分布。因此，式(1-1)实际上是表示一族分布。

不失一般性，假设总体为 $X \sim F(x;\theta)$，$\theta \in \Theta$，对固定的 θ(可以是向量)，$f(x,\theta)$ 都是分布密度函数，其中，Θ 为由 θ 可能取值的全体组成，称为参数空间。称 $\{F(x;\theta):\theta \in \Theta\}$ 为总体分布族。同样，对一个样本 $X_1, X_2, \cdots, X_n$，可以定义样本分布族：$\{\prod_{i=1}^{n} F(x_i;\theta):\theta \in \Theta\}$。

分布族规定了样本所来自的母体，反映了对研究对象的了解程度。因此，分布族(参数空间)确定了所要研究问题的范围，为统计分析指明了方向。

1.1.4 统计量

为了研究某个总体 X，一般方法是按照一定的试验设计获取其一个样本 $X_1, X_2, \cdots, X_n$。根据前面的叙述，$X_1, X_2, \cdots, X_n$ 具有一定的随机性。所以，一般不能由样本直接获得统计推断，必须对其进行有效的处理和加工，使样本中分散的信息集中起来，用样本的某个函数表示。这种函数就是数理统计学中常用的统计量。

设 $X_1, X_2, \cdots, X_n$ 为总体 X 的一个样本，$T(X_1, X_2, \cdots, X_n)$ 为一连续函数。如果 $T(X_1, X_2, \cdots, X_n)$ 中不含有任何未知参数，则称 $T(X_1, X_2, \cdots, X_n)$ 为一个统计量。

因此，统计量完全由样本确定，其不依赖于任何未知的量。例如，设 $X \sim N(\mu, \sigma^2)$，$X_1, X_2, \cdots, X_n$ 为其一个简单随机样本，则 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 都是统计量。其中 $\overline{X}$ 称为样本均值，S^2 称为样本方差。而 $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$ 则不是统计量，原因是其含有未知参数 μ。

统计量是一个随机变量，应用统计量可以对研究问题作出统计意义上的推断(如某个信号周期内到达左转车辆数小于给定数值的概率是多少?)。应用统计量来对所研究问题作出一定论断的过程称为统计推断。这部分内容将在后续章节中介绍。

1.1.5 秩统计量

设 X_j 为 $X_1, X_2, \cdots, X_n$ 中第 R_j 个最小值，则称 X_j 的秩为 R_j。由于 $X_1, X_2, \cdots, X_n$ 是一个随机样本，因此，秩 R_j 也是随机变量。记 $R = (R_1, R_2, \cdots, R_n)$，则 R 是样本 $X_1, X_2, \cdots, X_n$ 的一个统计量。由秩得到的统计量统称为秩统计量。

1.1.6 矩统计量

(1)样本矩

一类重要的统计量就是样本矩，分为原点样本矩和样本中心矩。

$$u_k = \frac{1}{n}\sum_{i=1}^{n}(X_i)^k \tag{1-2}$$

称为 k 阶样本原点矩。特别地，$k=1$ 时，则是样本均值。

$$\eta_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \tag{1-3}$$

称为 k 阶样本中心矩。应当注意的是，当 $k=2$ 时，二阶样本中心矩与样本方差只差一个系数，特别当 n 较大时，则差异不显著。

(2)偏度、峰度

样本偏度和峰度也是两个常用的统计量,其分别反映了总体分布的对称性与尖峰程度。其在交通安全中有着一定的应用,有的研究人员把地点速度分布偏斜程度作为交通事故潜在程度的表征。

样本偏度定义为:

$$\beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{\frac{3}{2}}} \tag{1-4}$$

样本峰度定义为:

$$\beta_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^{\frac{4}{2}}} \tag{1-5}$$

此外,偏度和峰度反映了数据的分布形态,例如正态分布的偏度为0,峰度为3。因此,样本峰度和偏度也常用来描述数据分布形态。

统计量的分布称为抽样分布。统计推断的结论是根据统计量的分布得到的。因此,确定统计量的分布是应用数理统计学进行推断的一个基本问题。统计量的分布在§1.3中介绍。

§1.2 顺序统计量和经验分布

1.2.1 顺序统计量

设 $X_1, X_2, \cdots, X_n$ 为总体 X 的一个样本,把 $X_1, X_2, \cdots, X_n$ 按从小到大的顺序排列为

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)} \tag{1-6}$$

则称 $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 为 $X_1, X_2, \cdots, X_n$ 的顺序统计量;通常称 $X_{(i)}$ 为"第 i 个顺序统计量"。特别地,由于 $X_{(1)}$ 和 $X_{(n)}$ 分别为极小值和极大值,又称为"极值"。

设 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 为 $X_1, X_2, \cdots, X_n$ 的顺序统计量,令

$$R = X_{(n)} - X_{(1)} \tag{1-7}$$

则称 R 为样本 $X_1, X_2, \cdots, X_n$ 的极差。极差是度量样本散布程度的一个统计量。

设 $X_1, X_2, \cdots, X_n \sim F(x)$,$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ 为其顺序统计量,则对 $1 \leqslant r \leqslant n$ 有

$$P(X_{(r)} \leqslant x) = r\binom{n}{r}\int_0^{F(x)} t^{r-1}(1-t)^{n-r}\mathrm{d}t \tag{1-8}$$

如果 $F(x)$ 有密度函数 $f(x)$,则

$$f_r(x) = r\binom{n}{r} f(x)[F(x)]^{r-1}[1-f(x)]^{n-r} \tag{1-9}$$

特别地,$r=1$ 或 $r=n$ 时,可分别得到 $X_{(1)}$ 和 $X_{(n)}$ 的分布函数

$$F_{(1)}(x) = 1 - [1 - F(x)]^n \tag{1-10}$$

$$F_{(n)}(x) = [F(x)]^n \tag{1-11}$$

1.2.2 经验分布

设 $X_1, X_2, \cdots, X_n \sim F(x)$，则称

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x) = \#\{X: X_i \leqslant x, i = 1, 2, \cdots, n\}/n \tag{1-12}$$

为 $X_1, X_2, \cdots, X_n$ 的经验分布函数。经验分布函数又可以写为：

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ k/n, & X_{(k)} \leqslant x < X_{(k+1)} \\ 1, & x \geqslant X_{(n)} \end{cases} \tag{1-13}$$

经验分布 $F_n(x)$ 可以看成是总体分布 $F(x)$ 的一个估计，并且 $F_n(x)$ 关于 x 收敛到 $F(x)$。根据经验分布可以绘出经验分布函数曲线，如图 1-1 所示。

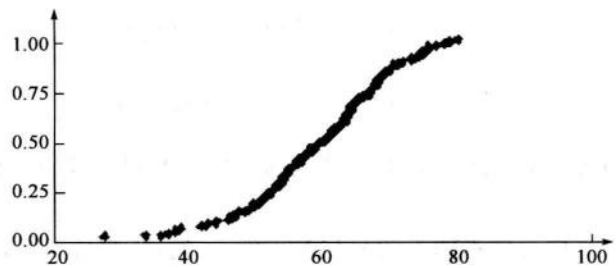

图 1-1　经验分布函数曲线示意图

§1.3　总体分位数和样本分位数

1.3.1 总体分位数

设 $F(x)$ 为一个一维分布，$0 < \alpha < 1$。如果 w_α 满足

$$F(w_\alpha - 0) \leqslant \alpha \leqslant F(w_\alpha) \tag{1-14}$$

则称 w_α 为 $F(x)$ 的 α 分位数（点）（或 $100\alpha\%$ 分位数），其意义是小于 w_α 的概率为 α。其中，$F(x_\alpha - 0)$ 为 F 在 w_α 点的左极限。

在统计分析中还常用到两个概念：下分位数和上分位数。由式(1-14)定义的分位数又称为 $F(x)$ 的下 α 分位数。同样可以定义上分位数。如果 u_α 满足：

$$F(u_\alpha - 0) \leqslant 1 - \alpha \leqslant F(u_\alpha) \tag{1-15}$$

则称 u_α 为 $F(x)$ 的上 α 分位数（即 $F(x)$ 的 $100(1-\alpha)\%$ 分位数）。其意义是大于 u_α 的概率为 α（如图 1-2 所示）。因此，上分位数与下分位数之间具有换算关系：$F(x)$ 的 α 上分位数就是 $F(x)$ 的 $(1-\alpha)$ 下分位数（图 1-2 为标准正态分布 $N(0,1)$ 的 α 上分位数与下分位数示意图）。

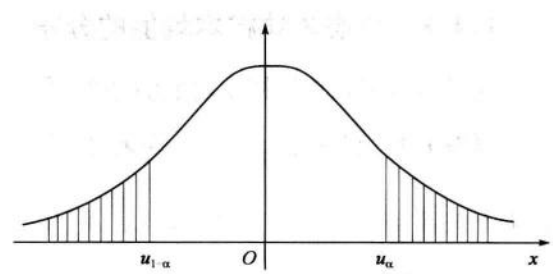

图 1-2　标准正态分布分位数示意图

在数理统计学中，常常讨论 α 分别取值为

0.25、0.50、0.75 的情况，因为这些数值对应的统计量在实际应用中是非常重要的。在交通工程学中则常常分析 α 取值 0.15、0.50、0.85 的情况，如在车辆限速管理中常常用速度的 15% 和 85% 分位数作为车辆限速的依据，车速中位数也是常用的一个指标。

由于总体的分布往往是不知道的，实践中常用样本分位数估计总体分位数。

1.3.2 样本分位数

设样本 $X_1, X_2, \cdots, X_n$ 的顺序统计量为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$。对 $0 < \alpha < 1$，令

$$m_{n,\alpha} = X_{[n\alpha]} + (n+1)\left(\alpha - \frac{[n\alpha]}{n+1}\right)(X_{([n\alpha]+1)} - X_{([n\alpha])}) \tag{1-16}$$

称 $m_{n,\alpha}$ 是 $X_1, X_2, \cdots, X_n$ 的“样本 p 分位数”，$[a]$ 表示不超过 a 的最大整数。

特别地，当 $\alpha = 0.5$ 时，$m_{n,\alpha}$ 称为“样本中位数”，常用 m_2 表示。并且

$$m_2 = \begin{cases} [x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]/2 & ,n \text{ 为偶数} \\ x_{(\frac{n+1}{2})} & ,n \text{ 为奇数} \end{cases} \tag{1-17}$$

在应用中，常用样本分位数 m_2 来估计总体的均值。并且由于 m_2 不受异常值的影响，是个稳健统计量。

当 $\alpha = 0.25$ 时，$m_{n,\alpha}$ 称为样本的下四分位数，用 m_1 表示；当 $\alpha = 0.75$ 时，$m_{n,\alpha}$ 称为样本的上四分位数，常用 m_3 表示。

设总体 X 的 α 分位数为 ξ_α，总体的密度函数 $f(x)$ 在 ξ_α 处连续，且 $f(\xi_\alpha) \neq 0$，则当样本量 n 足够大时，有

$$\sqrt{n}(m_{n,\alpha} - \xi_\alpha) \xrightarrow{L} N(0, p(1-p)/f^2(\xi_\alpha)) \tag{1-18}$$

即 $m_{n,\alpha}$ 渐近分布为正态分布 $N_r(\xi_\alpha, \alpha(1-\alpha)/f^2(\xi_\alpha))$。该结论只有在分布函数 $f(x)$ 已知的情况下才可以应用。在实际应用中，$f(x)$ 的分布往往是不知道的，这种情况下可以借助于 Bootstrap 方法，确定渐近分布。

§1.4 抽样分布

统计量的分布称为抽样分布。有些统计量的确切分布，如正态总体样本均值的分布是知道的；而有些统计量的确切分布是不知道的，只能给出渐近分布。

1.4.1 正态总体样本均值的分布

为了给出正态总体样本均值的分布，先看一个定理：

定理 1.1 设 $X_1, X_2, \cdots, X_n$ 为来自总体 $N(\mu, \sigma^2)$ 的一个样本，则样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 和样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$，满足以下结论：

(1) $\overline{X}$ 服从正态分布 $N(\mu, \sigma^2/n)$；

(2) $(n-1)S^2/\sigma^2$ 服从自由度为 $n-1$ 的 χ^2 分布；

(3) $\bar{X}$ 与 S^2 相互独立，并且 $t=\dfrac{\sqrt{n}(\bar{X}-\mu)}{S}$ 服从自由度为 $n-1$ 的 t 分布。

因此，由定理 1.1 可推出样本均值 $\bar{X}$ 服从均值为 μ，方差为 σ^2/n 的正态分布，即

$$\bar{X} \sim N(\mu,\sigma^2/n) \tag{1-19}$$

对于非正态总体，样本均值则没有上述性质，其精确分布就很难求出，这时可以借助于渐近分布。

1.4.2 一般总体样本均值的渐近分布

在统计分析中，多数情况下是难以知道统计量（如均值）的确切分布的。这就需要借助统计量的极限分布。而求极限分布的一个重要工具就是中心极限定理。用数学的语言可以描述为：

设 $X_1,X_2,\cdots,X_n$ 为相互独立，来自同一总体的一个样本，并且期望和方差存在：$EX_k=u$，$\mathrm{Var}X_k=\sigma^2(k=1,2,\cdots)$，则

$$U_n=\frac{\sum_{k=1}^{n}X_k-nu}{\sqrt{n}\sigma}=\frac{\sqrt{n}(\bar{X}-u)}{\sigma} \tag{1-20}$$

当样本量 $n\to\infty$ 时，渐近服从正态分布 $N(0,1)$。

中心极限定理说明了一个问题，即在样本量很大的情况下，样本均值 $\bar{X}$ 近似服从正态分布 $\bar{X}\sim N\left(u,\dfrac{\sigma^2}{n}\right)$。特别地，当样本量 n 很大时，总体分布均值可用样本均值 $\bar{X}$ 来估计，σ^2 可以用样本方差 S^2 来估计。并且

$$U_n=\frac{\sqrt{n}(\bar{X}-u)}{S}\xrightarrow{L}N(0,1) \tag{1-21}$$

中心极限定理为统计推断和假设检验提供了理论基础。

1.4.3 样本方差的分布

假设 $X_1,X_2,\cdots,X_n$ 为相互独立，来自同一总体的一个样本，并且期望和方差存在：$EX_k=u$，$\mathrm{Var}X_k=\sigma^2(k=1,2,\cdots)$，则对于样本方差

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{1-22}$$

可以计算得到

$$E(S^2)=\sigma^2 \tag{1-23}$$

$$\mathrm{Var}(S^2)=\left\{[X-E(X)]^4-\frac{n-3}{n-1}(\sigma^2)^2\right\} \tag{1-24}$$

由式(1-23)可以发现，样本方差 S^2 的期望值为 σ^2。因此，样本方差 S^2 常用作总体方差 σ^2 的估计值。同样，样本标准差

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2} \tag{1-25}$$

作为总体标准误差 σ 的估计值。

由式(1-24)可知，当样本量 n 增大时，样本方差 S^2 的方差会减小。因此，S^2 作为总体方差

σ^2 的估计值是合适的。当总体服从正态分布时，$E[X-E(X)]^4=3\sigma^2$，S^2 标准差为

$$\sqrt{\mathrm{Var}(S^2)}=\sqrt{\frac{2}{n-1}}\sigma^2 \tag{1-26}$$

所以，当用样本方差估计总体标准方差时，可用式(1-26)近似确定样本量，即求满足 $\frac{\sqrt{\mathrm{Var}(S^2)}}{\sigma^2}=\sqrt{\frac{2}{n-1}}$ 小于给定的相对误差值时的 n 值即可。

1.4.4 正态总体样本方差的分布

假设 $X_1,X_2,\cdots,X_n$ 为来自总体 $N(\mu,\sigma^2)$ 的一个样本，则由定理1.1可知，$(n-1)S^2/\sigma^2$ 服从自由度为 $n-1$ 的 χ^2 分布，即

$$\frac{n-1}{\sigma^2}S^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/\sigma^2\sim\chi^2(n-1) \tag{1-27}$$

记 $\chi^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/\sigma^2$（称为 χ^2 随机变量），由于 χ^2 的分布是已知的，可容易由式(1-27)导出 S^2 的分布，这里没有给出其分布，留作练习。

思 考 题

1. 试推导正态总体样本方差的分布。
2. 试用式(1-27)推导出正态总体样本方差的方差。

第2章　常用统计分布及其应用

概率统计方法是最早应用于交通研究的数学方法之一。其在交通控制、驾驶人行为分析、通行能力研究和交通规划等研究方向都得到了较广泛的应用。随着交通研究的深入开展和对交通现象认识的加深，越来越多的概率统计方法被加以应用。本章主要介绍交通工程中常用统计分布以及这些分布的性质。

§2.1　离散分布

离散型分布常用于描述一定时间间隔内事件的发生次数。如某段时间内到达停车场的车辆数，某路段一年内发生的交通事故数等。交通工程中常用的离散型分布主要有三种：泊松分布、二项分布和负二项分布。

2.1.1　泊松(Poisson)分布

泊松分布的分布函数：

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

式中：$P(X=x)$——在计数时间 T 内，事件 X 发生 x 次的概率；

λ——单位时间内平均发生的事件次数；

T——计数时间，如一个信号周期；

e——自然对数的底数，取值为2.718280。

若记 $m=\lambda T$，则 m 为时间 T 内平均发生的事件次数，式(2-1)可写为：

$$P(X=x)=\frac{(m)^x e^{-m}}{x!},x=0,1,2,\cdots \tag{2-2}$$

由式(2-2)可求得 X 的期望 $E(X)$ 和方差 $\mathrm{Var}(X)$：

$$E(X)=\sum_{x=0}^{\infty}x\frac{m^x e^{-m}}{x!}=m\sum_{x=1}^{\infty}\frac{m^{x-1}e^{-m}}{(x-1)!}=m \tag{2-3}$$

$$\mathrm{Var}(X)=\sum_{x=1}^{\infty}(x-m)^2\frac{m^x e^{-m}}{x!}=m \tag{2-4}$$

在实际应用中，期望 $m=E(X)$ 和方差 $\mathrm{Var}(X)$ 可分别由其样本均值 $\overline{m}$ 和样本方差 S^2 进行估计：

$$\overline{m}=\frac{\sum_{i=1}^{k}x_i f_i}{\sum_{i=1}^{k}f_i}=\frac{\sum_{i=1}^{k}x_i f_i}{n} \tag{2-5}$$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{N}(x_i-m)^2=\frac{1}{n-1}\sum_{j=1}^{k}(x_j-m)^2 f_j \tag{2-6}$$

式中：n——调查样本数；

f_j——观测时间内，事件 X 发生 x_j 次的频率；

k——观测数据分组数。

由式(2-3)和式(2-4)可以发现，泊松分布的期望 $E(X)$ 和方差 $\mathrm{Var}(X)$ 是相等的，这是泊松分布的一个重要特点。由概率论知识可知，泊松分布的样本均值 $\overline{m}$ 和样本方差 S^2 分别为总体均值和总体方差的无偏估计。因此，当 $\frac{S^2}{m}$ 显著不等于 1 时，则意味着用泊松分布拟合观测数据不合适。常用此作为能否应用泊松分布拟合观测数据分布的初始判据。

在实际应用时，常用以下递推式进行计算：

当 $x=0$ 时，

$$P(X=0)=\mathrm{e}^{-m} \tag{2-7}$$

当 $x\geqslant 1$ 时，

$$P(X=x)=\frac{m}{x}P(X=x-1) \tag{2-8}$$

在交通工程中，泊松分布最早用于描述一定时间内到达车辆数的分布规律的。当交通量不大且没有交通信号干扰时，基本上可用泊松分布拟合观测数据；当交通拥挤时，车辆之间的干扰较大，则应考虑用其他分布。此外，泊松分布还常用于描述一定时间内交通事故发生次数，故在交通安全中也有着广泛的应用。

例 2-1 假设一个商场停车场停车需求服从泊松分布。停车场每小时平均停车数为 10 辆，求 1 小时内到达车辆数小于等于 10 辆的概率；1 小时内到达车辆数大于 10 辆的概率；1 小时内到达车辆数大于 5 但不超过 10 的概率。

解：由式(2-2)可求得时间 T 内到达车辆数小于等于 x 辆的概率为：

$$P(X\leqslant x)=\sum_{i=0}^{x}\frac{m^i\mathrm{e}^{-m}}{i!}$$

对本题而言，$T=1$ 小时，$x=10$，$m=10$，所以

$$P(X\leqslant 10)=\sum_{i=0}^{x}\frac{m^i\mathrm{e}^{-m}}{i!}=\sum_{i=0}^{10}\frac{10^i\mathrm{e}^{-10}}{i!}=0.583$$

同样，1 小时内到达车辆数大于 10 的概率为：

$$P(X>10)=1-\sum_{i=0}^{10}\frac{10^i\mathrm{e}^{-10}}{i!}=0.417$$

1 小时内到达车辆数大于 5 但不超过 10 的概率为：

$$P(5<X\leqslant 10)=\sum_{i=6}^{10}\frac{m^i\mathrm{e}^{-m}}{i!}=0.516$$

2.1.2 二项分布

在交通工程中，描述计数事件发生次数的另一个常用分布是二项分布。其分布函数为：

$$P(X=x)=C_n^x p^x(1-p)^{n-x}\quad ,x=0,1,2,\cdots \tag{2-9}$$

式中：$C_n^x=\frac{n!}{x!\ (n-x)!}$；

p、n——二项分布参数，$0<p<1$，n 为正整数。