美国医师执照考试高效复习丛书（中英文对照）

High-Yield BIOSTATISTICS

17

生物统计学

[美]安东尼·N·格拉泽◆著
(Anthony N. Glaser)

（第 2 版）

中英对照 高效快捷 条理清晰 图文并茂

中信出版社
辽宁教育出版社

美国医师执照考试高效复习丛书(中英文对照)

[美] 安东尼·N·格拉泽 著

生物统计学

High – Yield Biostatistics

(第2版)

译 者 常 彤
审 校 姚 晨

中信出版社
辽宁教育出版社

图书在版编目（CIP）数据

生物统计学/（美）格拉泽著；常彤译. —北京：中信出版社，2004.2
（美国医师执照考试高效复习丛书）
书名原文：High - Yield Biostatistics
ISBN 7 - 5086 - 0119 - X

Ⅰ.生… Ⅱ.①格…②常…③姚… Ⅲ.生物统计 - 医师 - 资格考核 - 美国 - 自学参考资料 - 汉、英
Ⅳ.Q332

中国版本图书馆 CIP 数据核字(2003)第 003481 号

本书中药物的适应证、不良反应和剂量及用法有可能变化，读者在用药时应注意阅读厂商在包装盒上提供的信息。

生物统计学
SHENGWU TONGJIXUE

著　　者：[美] 安东尼·N·格拉泽
译　　者：常　彤
责任编辑：贾增福　靳纯桥
出版发行：中信出版社（北京朝阳区东外大街亮马河南路 14 号 塔园外交办公大楼 100600）
经 销 者：中信联合发行有限公司
承 印 者：北京牛山世兴印刷厂
开　　本：787mm × 1092mm 1/16　　**印　　张**：12　　**字　　数**：215 千字
版　　次：2004 年 2 月第 1 版　　**印　　次**：2004 年 2 月第 1 次印刷
京权图字：01 - 2003 - 8658
书　　号：ISBN 7 - 5086 - 0119 - X/R·34
定　　价：30.00 元

译者前言

《美国医师执照考试高效复习丛书》由 Lippincott Williams & Wilkins 公司出版，为参加美国医师行医执照考试（United States Medical Licensing Examination，USMLE）所用的培训教材，其主要读者对象是美国国内准备参加考试的医学生或毕业生和有志获取美国行医执照的外国医生或医学生。为了满足我国广大医学生和医生的需求，适应双语教学的需要，中信出版社和辽宁教育出版社委托首都医科大学组织学校及各附属医院相关学科的专家教授翻译了这套丛书。

丛书共 17 个分册，涵盖 USMLE 第一阶段（Step 1）基础医学和第二阶段（Step 2）临床医学的主要课程。丛书复习的高效性主要体现在：内容高度概括，重点突出，利于考生抓住重点，快速记忆；内容选择针对性强，用较少的时间便可掌握更多更重要的知识。各分册均由相关专业的专家教授编写，使丛书内容更具有权威性。

丛书的主要特点：(1)编排新颖、图文并茂：既有基础知识要点的分类介绍，又有以疾病为核心的综合复习，同时还有相关学科的横向比较和归纳；该丛书收集了大量丰富多彩的图片，使内容直观易懂；运用了大量表格对重要概念和问题进行比较、归纳和总结，便于快速理解和记忆。(2)理论联系实践，基础与临床结合：基础医学部分在讨论基础医学知识的同时，设有“与临床联系”等类似内容。临床医学部分在学习临床理论的同时，给出各种“病例分析”，使理论与实践紧密结合。这对医学教育的思维模式是一种创新。(3)丛书出版采取中英文合出的形式，即前面是中文，后面是英文，可供对照阅读。

丛书既可作为教学材料，又可供学生课后参考，适应于医学院校开展双语教学；也可作为我国执业医师资格考试复习的参考书，以及有志于获取美国行医执照的中国医学生和医生参考。

需要说明的是，书中部分图片是引用其他作者的，因在英文部分均有交代，在中文部分未列出。

首都医科大学

Preface

Biostatistics is a course that most people dislike and try to scrape by on—often because it is taught by a statistician who assumes that everyone is adept at mathematics and can cope with blackboards full of formulae and equations. However, today's medical students will not, for the most part, become producers of research (and those who do will undoubtedly seek the advice of professional statisticians when designing their studies). On the other hand, all will become consumers of research and will need to understand the inferential statistical principles behind the reports that they read. Consequently, while this book does not pretend to be an in-depth statistics textbook (which few medical students have the need, the time, or the desire for), it aims to be more than a set of notes to be memorized for the purpose of examinations.

High-Yield Biostatistics explains concepts, provides examples, and also contains review exercises for the more difficult material in the first three chapters. Unlike other biostatistics books, it covers the complete range of biostatistics material that can be expected to appear on the USMLE Step 1, without going beyond that range.

The term "high yield" has become something of a buzzword among students. Its use here reflects that the core of information presented in *High-Yield Biostatistics* will be tested on the "Boards." The book contains little extraneous information, although the material does tread beyond the irreducible minimum in some instances.

For the more inquisitive reader, a small amount of additional material is presented in notes at the end of some chapters, but this information is neither necessary for a general understanding of the material nor to answer USMLE questions.

If you have any suggestions for changes or improvements to this book, or if you find a biostatistics question on Step 1 that this book does not equip you to answer, please drop me a line.

Anthony N. Glaser
tonyglaser@mindspring.com

目　录

Contents

1 统计描述

统计方法可以分成两大类:统计描述和统计推断。

• 统计描述　是对数据进行描述、组织及求和,此处仅指可用的实际数据。例如,一组病人的平均血压、外科手术的成功率等。

• 统计推断　对超出实际范围的数据做出推断。通常包含归纳推理(即在仅观察了一个样本之后就综合为一个总体)。例如,全美国人的平均血压、病人在做外科手术前对手术的期望成功率等。

Ⅰ. 总体、样本和观察单位

总体是研究者希望得出结论的全域。它不必由人群构成,而是一批测量数据。如果研究者希望对美国人的血压得出结论,总体由血压测量数据构成,而不是由美国人本身构成。

样本是总体的子集,是正在观察或研究的部分。因为研究者几乎不可能研究整个总体,因此,当只研究一个样本而想得出总体的结论时,几乎总是需要统计推断。

一次单一的观察——例如一个人的血压——是一个观察单位,用 X 表示。在一个总体中观察单位的总数用 N 表示,在一个样本中观察单位的个数用 n 表示。因此,总体是由从 X_1 到 X_N 的全部观察单位构成,样本由这些 N 个观察单位中的 n 个个体构成。

用于生物医学研究的大多数样本都是概率型样本——在该样本中,研究者能够指定该总体所包含的任一个观察单位的概率。例如,假如我们从 52 张扑克牌中随机地抽取 4 张作为一个样本,那么包含一张牌的概率是 4/52。概率型样本允许应用统计推断,而非概率型样本仅允许用统计描述。四种基本的概率型样本是:单纯随机样本、分层随机样本、整群样本和系统样本。

单纯随机样本

单纯随机样本是一种最简单的概率样本。在这种样本中,总体中的每个观察单位都有相同的机会被抽中,如同上面玩牌的例子。对随机样本的定义是根据它的抽样方法,而不是根据它的结果。如果摸到 4 张红心牌,这本身并不意味着样本不是随机的。

如果样本与抽取它的总体非常相似,则样本具有代表性。所有的随机样本倾向于可以代表总体,但它们并不能保证其代表性。非代表性样本能够引起严重问题(很清楚,4 张红心牌并不是一次摸到的牌的代表)。

一个非代表性样本的经典例子是在 1936 年之前美国总统选举的民意测验。在样本为 200 多万人的基础上,Atfred Landon 想达到超过 Franklin Delano Roosevelt 的压倒性胜利,但

结果却相反。问题在哪里？样本来自电话记录和汽车所有者——在那经济不景气的年代，拥有这种产品的人不是整体选民的代表。

如果一个样本(或一个结果)一贯地在一个特定方向上犯错误，那么它是有偏的。例如，在由500个白人和500个黑人组成的总体中抽取一个10人样本，一种一贯地产生多于5个白人的抽样方法即是有偏性的。有偏性的样本是不具代表性的。可以证明，真正的随机化是无偏的。

分层随机样本

在一种分层随机样本中，首先把总体分成相对均匀的组(或层)，从这组(或层)中进行随机抽样。这种分层产生更大的代表性。例如，代替从500个白人和500个黑人的总体中抽取10人的一个样本，我们随机地从500个白人中抽取5个白人和从500个黑人中抽取5个黑人，这样就保证了一个10人样本的结果中种族的代表性。

整群样本

当进行单纯随机抽样或分层随机抽样太费钱费力时，可以采用整群样本。例如，在美国对100名医学院学生调查中，调查者可以选择一个随机组或"群"——例如10个美国医学院校为一个随机组——然后在那10个院校中对所有的学生进行调查。这种方法比试图在全部美国医学院学生中直接抽取100个随机样本要经济和实用得多。

系统样本

这些包含以系统的方法选择观察单位，例如每1/5的病人允许去一个医院，或者每1/3的婴儿在指定地区出生。这类抽样通常只提供单纯随机样本的等效样本，而实际上没有应用随机化。

在临床研究中，抽样问题是常见的。

例如，如果一个研究者在报纸上做广告以招收患有某种疾病的人——痤疮、糖尿病或抑郁症——响应者构成一种自选择的人群。在所有有这些疾病的人群构成的总体中，这些人或许并不具有代表性。

类似地，假如一个皮肤病学家报告了对痤疮的一种新的治疗方法，这方法已经应用于他的病人。但这样本对所有有痤疮的病人的总体来说也许并不具有代表性，因为或许只有具有更严重痤疮的人(或者具有良好保险的人)才从一个皮肤病学家那里寻找治疗方法。在任何情况下，他的实践或许局限于某一特定的地理、气候和人文地区。在这种情况下，虽然就病人而论，他的研究是有效的(这可叫做内部效度)，但是推广他的成果到有痤疮的其他人或许就不那么有效了(这又叫做缺乏外部效度)。

Ⅱ. 概率

一个事件的概率用 P 表示。概率通常用十进制的分数表达，不用百分数表达，并且它应当在0(零概率)和1(绝对肯定)之间。一个事件的概率不可能是负值。一个事件的概率也可表达为希望的结果与可能的结果之间的比值。

例如，假如一个硬币投掷无数次，出现正面的次数占投掷数的50%，那么，正面的概率 P 是0.50。假如一个10人的随机样本从100人的总体中抽样无数次，每个人包含在样本

中的次数是10%，因此，P(包含在任一样本中的概率)是0.10。

一个事件不发生的概率(用 q 表示)等于1减去事件发生的概率。在上述例子中，任何一个人不包含在任一样本中的概率 q 是$(1-P)=(1-0.10)=0.90$。

美国医师执照考试(USMLE)要求熟悉计算概率的三种主要方法：加法原则、乘法原则、二项分布。

加法原则

概率加法原则陈述为：若干事件之中任一个发生的概率等于各个事件概率之和，如果事件是互相排斥(即它们不可能同时发生)的话。

如果一次玩牌中摸一张红心的概率是0.25，则摸一张方块的概率也是0.25，按照加法原则，摸到一张红心或一张方块的概率是$0.25+0.25=0.50$。由于不存在又是红心又是方块的牌，这些事件满足事件互相排斥的要求。

乘法原则

概率乘法原则陈述为：2个或更多统计独立事件全发生的概率等于它们单独概率的乘积。

例如一个癌症病的生存概率是0.25，一个精神分裂症病人的生存概率是0.01，那么一个同时是癌症病人又同时是精神分裂症病人的生存概率是$0.25\times0.01=0.0025$，只要这两种病是独立的——换句话说，患了一种病，但这既不增加也不减少患其它病的危险。

二项分布

互相排斥、独立的事件的一种特定的组合事件发生的概率能够应用二项分布来确定。二项分布是这样一种分布，即只存在两种可能性。例如，是/不是，男/女，以及健康/生病等。假如一次实验严格地有两种可能的结果(其中之一一般称之为"成功")，二项分布给出在一系列独立尝试后获得成功的精确次数的概率。

二项分布典型的医学应用是遗传学咨询。一种遗传性疾病，如泰-萨(Tay-Sachs)病的遗传性，遵循二项分布：有两种可能情况(疾病遗传或不遗传)是相互排斥的(一个人不可能既有又没有这种疾病)，并且可能性是独立的(例如，在一个家庭中，一个小孩已经遗传上这种疾病，但是并不影响家庭中别的小孩遗传此病的机会)。

因此，医生能够应用二项分布通知一对夫妇：谁是遗传性疾病的携带者，某些特定的结合事件(例如，如果他们有2个小孩，这2个小孩都不得遗传疾病的概率)发生的可能性有多大。对于USMLE，并不需要学习或应用二项分布的公式。

Ⅲ. 数据的类型

适当的统计技术的选择依赖于问题中数据的类型。数据有四种度量尺度：定名测量(Normial)、定序测量(Ordinal)、定距测量(Interval)或定比测量(Ratio)。帮助记忆的符号："NOIR"能够用来记住这些度量尺度的顺序。数据还具有离散或连续的特征。

定名测量数据

定名测量数据被定性地分成类或组，如男/女、黑/白、大城市/小城市/农村及红/绿等。没有隐含顺序或比值。仅分成两组的叫做二项分布数据。

定序测量数据

定序测量数据能够以一种名义顺序排列(例如,学生可以按他们的班分成一班、二班、三班),但是不知道间隔的大小——例如,一班学生减去二班学生的数量与二班学生减去三班学生的数量是否相等,没有结论。

定距测量数据

定距测量数据像定序测量数据一样能够按名义次序排列。另外,它们在记录(这是通常的测量值)之间有等距的意义。例如,在摄氏温度表中,在100℃和90℃之间的差别与50℃和40℃之间的差别相同。然而,因为定距测量数据没有绝对的零,数值的比值是没有意义的,例如,100℃并不是50℃的热度的两倍,因为0℃并不代表一点热度也没有。

定比测量数据

定比测量数据有与定距测量数据同样的性质,然而因为它有绝对的零,比值意义的确存在。大多数生物变量形成一种定比测量数据:以克或磅表示重量,以秒或天表示时间,以"mmHg"表示血压,以及每分钟的脉搏率都是定比测量数据。温度的惟一定比测量数据是Kelvin尺度,其中零度表示绝热,正如零脉搏表示心脏跳动的绝对缺乏。因此,以下说法是正确的:120次/min是60次/min的2倍,300°K是150°K的2倍。

离散型变量

离散型变量仅能取自一定的数值,在这些数值之间没有值。例如,在医院人口普查中,病人数可以是178或179,但在178和179之间没有值。一个诊所在一天中用的注射器的数目,仅能以单位1的数量来增加或减少。

连续型变量

连续型变量可以取一定极限范围的任何数值。大多数生物医学变量是连续的(例如,病人的年龄、身高、体重和血压)。然而,在测量或报告连续型变量的过程中将它们简化成离散型变量。例如,血压可以报告成最接近整数的mmHg数,体重可以报告成最接近的磅数,年龄可以报告成最接近的年数。

Ⅳ. 频数分布

未组织好的数据集是难以消化和理解的。考虑一个200人的血浆胆固醇样本:200个值的列表是其原始值的排列。第一种简单地组织数据的方式是按顺序从高至低列出所有可能值,记录每个值发生的频数(f)。这就形成了频数分布。假如最高的血胆固醇水平是260mg/dl*,最低值是161mg/dl,频数分布如表1-1所示。

成组频数分布

表1-1是一种不常用的数据表达。通过创建成组频数分布,这些数据能够变得更加易于管理,如表1-2所示。单个数值被成组化(分成7~20组通常是适当的)。每组数值围绕一个相等的组距。在这个例子中,分成了10组组距为10的数。(161~170,171~180等)。

---

* 1mg/dl = 0.026mmol/L,下同

表 1-1

值(mg/dl)	频数	值(mg/dl)	频数	值(mg/dl)	频数	值(mg/dl)	频数	值(mg/dl)	频数
260	1	240	2	220	4	200	3	180	0
259	0	239	1	219	2	199	0	179	2
258	1	238	2	218	1	198	1	178	1
257	0	237	0	217	3	197	3	177	0
256	0	236	3	216	4	196	2	176	0
255	0	235	1	215	5	195	0	175	0
254	1	234	2	214	3	194	3	174	1
253	0	233	2	213	4	193	1	173	0
252	1	232	4	212	6	192	0	172	0
251	1	231	2	211	5	191	2	171	1
250	0	230	3	210	8	190	2	170	1
249	2	229	1	209	9	189	1	169	1
248	1	228	0	208	1	188	2	168	0
247	1	227	2	207	9	187	1	167	0
246	0	226	3	206	8	186	0	165	0
245	1	225	3	205	6	186	2	164	1
244	2	224	2	204	8	184	1	163	0
243	3	223	1	203	4	183	1	162	0
242	2	222	2	202	5	182	1	161	0
241	1	221	1	201	4	181	1	160	1

相对频数分布

如表 1-2 所示,成组频数分布能够转换成相对频数分布,它表示落在每个组距之内所有观察单位的百分比。任一给定组距内观察单位的相对频数通过用 f 去除 n(样本数,这里为 200)而获得。将结果乘以 100,就转换成百分数了。这样,这种分布显示 211 和 220 之间血胆固醇水平是 19%。

累计频数分布

表 1-2 也给出了累计频数分布。它可以用百分比表示,表明处在每一组距及其以下的观察单位所占的百分比。虽然有一组可以叫做 211 ~ 220 组,但这组实际上包含分值的极差是从 210.5 ~ 220.5(含 220.5)——那么,图 1-1 表示出该组精确的上限和下限。

处于 161 ~ 170 组的相对频数为 2%,处于 171 ~ 180 组的相对频数为 2.5%,因此在 180.5 以下(含 180.5)的总的累计频数为 4.5%,正如表 1-2 中累计频数列中所表示的。进一步看,处于 181 ~ 190 组的相对频数为 6%,因此处在 190.5 以下(含 190.5)的总累计频数为(2 + 2.5 + 6)% = 10.5%。根据这一抽样,约有 10%的人的血胆固醇低于 190mg/dl,约有 90%的人高于此值。最高组(251 ~ 260)的累计频数应当是 100%,表明处于 260.5 以下(包

含260.5)的累计频数为100%。

表 1-2

区间	频数(f)	相对频数(f%)	累积频数(f%)
251 ~ 260	5	2.5	100.0
241 ~ 250	13	6.5	97.5
231 ~ 240	19	9.5	91.0
221 ~ 230	18	9.0	81.5
211 ~ 220	38	19.0	72.5
201 ~ 210	72	36.0	53.5
191 ~ 200	14	7.0	17.5
181 ~ 190	12	6.0	10.5
171 ~ 180	5	2.5	4.5
161 ~ 170	4	2.0	2.0

频数分布的图形表示

频数分布常常用图形来表示,最常用的是直方图。图1-1是表1-2的成组频数分布的直方图,横坐标(*X* 轴或水平轴),表示各组的数值;纵坐标(*Y* 轴或垂直轴)表示频数。

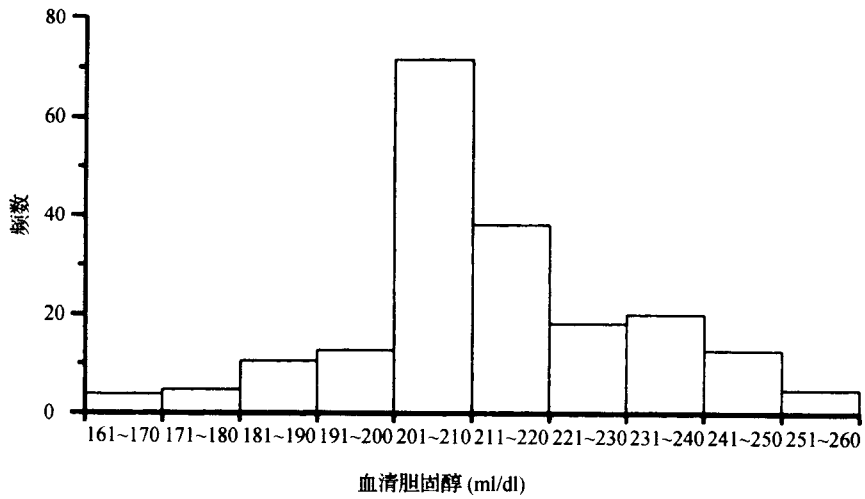

图 1-1

为了显示定名测量数据,典型的是采用直条图。例如,100例男性的一组血胆固醇数值,其平均值为212mg/dl;100例女性的一组血胆固醇数值,其平均值为185mg/dl。这两组数值的意义能够用一种直条图来表示,如图1-2所示。

除了直条图中的每个矩形在空间上是互相分离的以外,直条图等于直方图。直条图可表示不同的数据(如男人和女人),而不是连续的数据。

对定距测量或定比测量数据,频数分布可以用频数多边形来描绘,其中每一区间分级

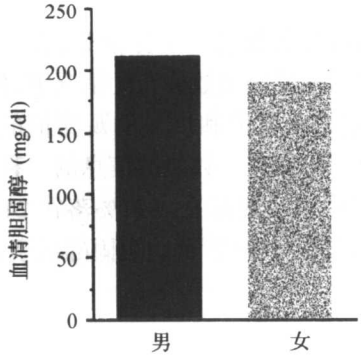

图 1-2

的中点用直线连接，如图 **1-3** 所示。

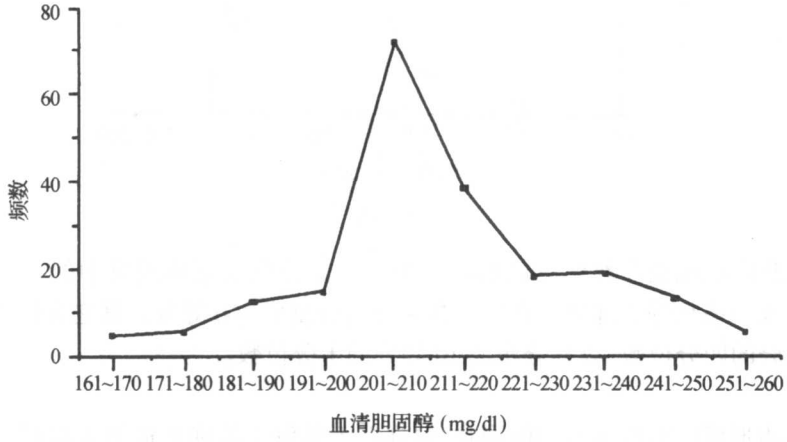

图 1-3

累计频数分布也能够用一个多边图形来表示，如图 1-4A 所示。累计频数多边形典型地形如 S 曲线，叫做尖顶曲线，图 1-4A 就是它的近似形状。

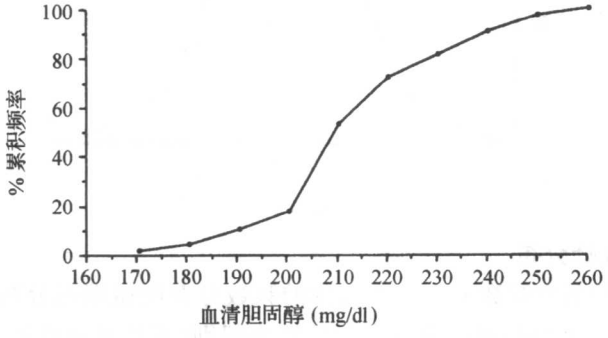

图 1-4A

百分位数和其它分位数

累计频数多边形和累计频数分布都说明了分位数(百分数)的概念,它表示落在任一具体数值下的观测值的百分比。在成组频数分布的情况下,例如表 1-2 的一组累计频数分布,百分位数表示落在任一给定组距以内和以下的观测值的百分比。百分位数提供了在一种分布中给出单一数值相对于所有其它数值的信息的一种途径。

例如,在表 1-2 累计频数列中表明观测值的 91%落在 240.5mg/dl 以下,因此它代表第 91 个百分位数(能够写成 C_{91}),如图 1-4B 所示。胆固醇低于 240mg/dl 的人占 91%,高于此值的人占 9%。

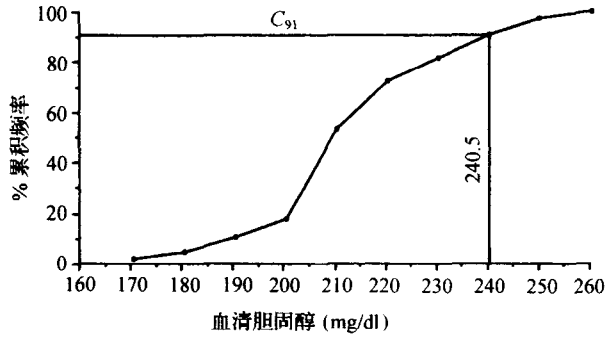

图 1-4B

百分位数广泛地用在教育测试的报告中。它是分位数家族成员中的一个成员,它们将分布分割成一些相等的部分。百分位数将分布分割成 100 等分。其它分位数有:将数据等分成 4 等分的四分位数,等分分布成 10 等分的十分位数。

正态分布

频数多边形可以有许多不同的形状,但许多天然发生的现象近似地按照一种对称的、类似钟形状的正态分布或高斯分布,如图 1-5 所示。

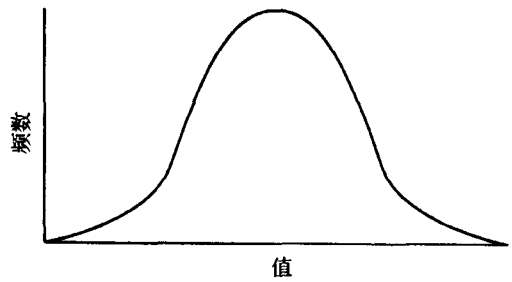

图 1-5

偏态、类 J 形状和双峰分布

图 1-6 表示某些其它频数分布。不对称的频数分布叫做偏态分布。正(或右)偏态分布和负(或左)偏态分布能够通过曲线尾部位置来识别(不是通过峰值—— 一种公共误差的位置来识别)。正偏态分布有相对较大数目的低值和相对较小数目的高值,而负偏态分