概率论与数理统计

（第三版）

盛 骤 编

上海交通大學出版社
SHANGHAI JIAO TONG UNIVERSITY PRESS

目　录

图书在版编目（**CIP**）数据

概率论与数理统计/盛骤编. --3版. --上海：上海交通大学出版社，2011

ISBN 978-7-313-02024-6

Ⅰ. ①概… Ⅱ. ①盛… Ⅲ. ①概率论—高等学校—教材 ②数理统计—高等学校—教材 Ⅳ. ①021

中国版本图书馆CIP数据核字（2011）第127966号

概率论与数理统计

（第三版）

盛 骤 编

出版社出版发行

（上海市番禺路951号 邮政编码200030）

电 话：64071208

出版人：韩建民

浙江云广印业有限公司印刷 全国新华书店经销

开 本：880mm×1230mm 1/32

印 张：9.25

字 数：260千字

1998年8月第1版　2011年8月第3版　2011年8月第15次印刷

印　　数：5030

ISBN 978-7-313-02024-6/O

定　　价：18.00元

告读者：如发现本书有质量问题请与印刷厂质量科联系

联系电话：0573-86572317

内容提要

本书是在2006年版的基础上增订而成的，新增内容主要有：bootstrap方法和在数理统计中应用Excel软件。全书共分9章：事件的概率、随机变量、随机变量的数字特征、正态分布、参数的点估计、假设检验与区间估计、回归分析与方差分析、bootstrap方法和在数理统计中应用Excel软件。各章配有适量的习题，并附有习题答案。

本书可作为高等院校工科各专业、理科（非数学专业）各专业概率论与数理统计课程的教材，也可供相关专业技术人员参考。

第三版前言

教材应该力求与时俱进，本版新增了以下内容：

（1）简单介绍了用bootstrap方法求参数的点估计和区间估计的具体做法。非参数bootstrap方法和参数bootstrap方法可用于当人们对总体知之甚少的情况，它们是近代统计中用于数据处理的重要的实用方法。

（2）新增了在数理统计中应用Excel软件一章，介绍了Excel软件及其在数理统计中的一些应用，举例介绍了应用VBA语言编写“宏”求解具体的数理统计问题。

（3）新增了点图、茎叶图、箱线图，新增了假设检验问题的p值法和列联表的独立性检验等内容。

新增内容与第二版内容相对独立，使用本教材时可视学时的多少作选择和安排。

诚恳地希望读者批评、指正。

盛骤

2011年5月

前言

本书是按照国家教育委员会高等学校工科数学课程教学指导委员会制订的《概率论与数理统计课程基本要求，Ⅱ类（概率少，统计多）》所规定的内容的广度和深度编写的，可作为高等学校工科本科各专业、理科（非数学专业）本科各专业概率论与数理统计课程的教材，也可供各类专业技术人员参考。

本书致力于讲清基本概念、基本理论和基本方法；在引入基本概念时，注意揭示其直观背景和实际意义；在叙述基本概念和基本方法时，特别注意阐明概率和统计的意义和思想；在选配例题和习题时，着力使学生理解基本理论和基本方法是怎样用于解决实际问题的，以培养学生运用概率统计的方法解决实际问题的能力。

本书致力于内容安排紧凑，讲述深入浅出，思路清晰，便于教师教学和学生学习。

考虑到学时的限制，对于《基本要求》中要求相对较低的那部分内容，本书力求叙述简明扼要，讲求实效，凡讲到的内容，一定讲清楚，使学生做到“了解”或“会”，不留有疑虑。对于这部分内容的习题，有意识地减轻了分量，以此来节约教学学时。本书对理论的论证作了适当的处理，做到详略适当，对于不证明的定理也说清楚条件、结论和意义。

本书在内容的表述和例题、习题的选配上注意能引起读者的兴趣。希望读者能感到这是一门不难学好的课程。

浙江大学应用数学系范大茵教授对本书书稿提出了宝贵的意见，在此表示衷心的感谢。诚恳地希望广大读者批评指正。

盛骤

1998年1月

1　事件的概率

在自然界和人们的活动中，存在着这样的一类现象：在一定条件下既可能出现这种结果，也可能出现另一种结果，出现哪一种结果具有不确定性，所出现的结果在事先是不能预知的. 这种现象称为随机现象 .

例如，投掷一枚硬币，它可能出现正面，也可能出现反面，在投掷之前不能预知.　又如用包装机包装水泥，规定每袋25kg，在生产线上任取一袋，其净重可能大于25kg，也可能小于25kg，在事先不能预知，若取10袋一一复称会得到不尽相同的值.　以上例子所说的现象都是随机现象.　　又如在考察一个地区的年降雨量时，在观察新生儿的性别或体重时，在考察晶体管的寿命时，都呈现出随机现象.

然而，若对一随机现象进行多次重复观察，人们可以发现其出现的结果呈现出规律性.　例如，多次投掷硬币则出现正面的次数约占一半；复称一批25kg装的袋装水泥，其净重是按照一定规律分布的，在25kg附近占绝大多数而远离25kg的占极少数.　这种在多次重复观察中，随机现象所显示的规律性，称为统计规律性 .

概率论与数理统计是研究随机现象所具有的统计规律性的数学学科.

1.1　随机事件

1.1.1　随机事件的概念

在概率论与数理统计中，习惯上将任一观察或测量的过程，视作是一个试验.　我们考虑这样的试验：它可以在相同的条件下重复进行，它的所有可能结果在试验之前是知道的，但对一次试验而言，它的结果是不能预知的，这种类型的试验称为随机试验，简称试验.

我们将试验的所有可能结果组成的集合称为试验的样本空间　，记为S. 样本空间的每个元素，即试验的每个结果，称为样本点.

下面是几个随机试验的例子.

（1）抛掷一枚硬币，观察出现正面（记为H）还是反面（记为T），则样本空间

$$S_1 = \{H，T\}.$$

（2）在一个班级中选一名学生，观察他的概率论与数理统计课程期终考试的得分（设以百分制记分），则样本空间

$$S_2 = \{0，1，2，\ldots，100\}.$$

（3）某种食品在制成后，经检验将食品分成Ⅰ，Ⅱ，Ⅲ，Ⅳ四个等级：Ⅰ级出售给食品商店；Ⅱ，Ⅲ级降价出售；Ⅳ级作为饲料出售. 取一份食品，观察它属于哪一级别，则样本空间

$$S_3 = \{Ⅰ，Ⅱ，Ⅲ，Ⅳ\}.$$

（4）在生产线的出口处测试闪光灯电池的电压，一只接一只地测试，直到发现一只次品为止. 若记正品为N，次品为D，则样本空间

$$S_4 = \{D，ND，NND，\ldots\}.$$

（5）在一批圆钢中取一条，测量它的抗拉强度f（以N/mm^2 计），则样本空间

$$S_5 = \{f|f>0\}.$$

在每次试验中，有一个结果（样本点）出现，也只有一个结果出现. 在研究试验的结果时，人们不但对试验的单一结果感兴趣，而且常常对试验的某些结果所组成的集合更感兴趣. 例如在上述试验2中，教师关心“考试成绩及格”，也就是关心S_2 的子集A={60，61，…，100}，我们称A是试验2的一个随机事件；在一次试验中若A的一个样本点出现，就说在这一次试验中A发生了. 又如在试验3中，我们关心“食品不作为饲料出售”，即关心S_3 的子集B={Ⅰ，Ⅱ，Ⅲ}，称B是试验3的一个随机事件；在一次试验中若B中有一个样本点出现，就说在这一次试验中B发生了.

一般，设试验E的样本空间为S，由S中的一些样本点组成的集合，称为试验E的随机事件 ，简称事件 . 随机事件是样本空间S的一个子集. 在一次试验中当且仅当这个子集中的一个样本点出现，就称这一事件发生 . 由于随机事件是由S中的一部分样本点组成的，因而在一次试验中，这一事件可能发生也可能不发生.

随机事件用大写的字母如A，B，C等来表示.

例如在试验2中事件A_1 ：“成绩为优良”，即A_1 ={81，82，…，100}；在试验3中事件A_3：“食品降价出售”，即A_3 ={Ⅱ，Ⅲ}.

特别，只含一个样本点的集合，称为基本事件 ，例如试验1有两个基本事件{H}，{T}；试验3有4个基本事件{Ⅰ}，{Ⅱ}，{Ⅲ}，{Ⅳ}；试验4有可列个基本事件{D}，{ND}，…；试验5含不可列个基本事件，如{f|f=1000}，{f|f=1500}等.

我们将样本空间S也作为一个随机事件，因为每次试验必然出现S中的某个样本点，因而在每次试验中S必然发生，称S为必然事件 . 我们也将不包含任何样本点的空集∅作为一个随机事件，它在每次试验中都不发生，称为不可能事件 .

1.1.2 随机事件间的关系和运算

包含 若事件A发生必导致事件B发生，则称事件B包含 事件A，记为A⊂B. 此时A中包含的样本点都含于B中.

相等 若对于两个事件A，B有关系A⊂B和B⊂A，则称A与B相等，记为A=B. 此时A和B包含的样本点相同.

和事件 若A，B是两事件，则“事件A和事件B至少有一个发生”这一事件称为A与B的和事件 ，记为A∪B. A∪B包含且只包含所有单属于A的样本点、单属于B的样本点以及同属于A与B两者的样本点.

积事件 若A，B是两事件，则“事件A和事件B同时发生”这一事件称为A与B的积事件 ，记为A∩B或AB. A∩B包含且只包含所有同属于A

和B两者的样本点.

差事件　若A，B是两事件，则“事件A发生而事件B不发生”这一事件称为A与B的差事件 ，记为A-B. A-B包含且只包含所有属于A且不属于B的样本点.

不相容　若事件A和事件B不能同时发生，则称事件A和事件B是不相容的 . 此时，A和B没有公共的样本点，即AB=∅. 若A，B中有一个发生，则另一个一定不发生. 基本事件是两两不相容的.

逆事件　若两事件A，B满足A∪B=S，AB=∅，则称B是A的逆事件或A是B的逆事件 . 记为或. 此时包含且只包含所有不属于A的样本点. 当且仅当A不发生时发生，.

我们可以用一种称为文氏图（Venn diagram）的图形来形象地示意事件的关系及运算. 在图1-1中，以长方形来示意样本空间S，事件A用一个圆表示，事件B用一个椭圆表示.

图1−1

图1-1a表示事件B包含事件A；图1-1b，c，d中有阴影线的区域分别表示A∪B，A∩B，B-A；图1-1e表示事件A和事件B互不相容；图1-1f表示事件B和事件互为逆事件.

类似地，对于有限个事件A_1，A_2，…，A_n，“事件A_1，A_2，…，A_n 至少有一个发生”，这一事件称为A_1 ，A_2 ，…，A_n 的和事件，记为，“事件A_1 ，A_2 ，…，A_n 同时发生”这一事件称为事件A_1 ，A_2 ，…，A_n 的积事件，记为. 也可以类似地定义可列个事件A_1 ，A_2 ，…，

A_n，…的和事件与积事件.

【例**1-1**】　　在一通道的出口处，观察相继驶过的三辆汽车向左（L）转弯还是向右（R）转弯. 在这一试验中事件

A={至少有一辆车向左转弯}

={LLL，RLL，LRL，LLR，LRR，RLR，RRL}，

B={第二辆车向左转弯}={LLL，RLL，LLR，RLR}，

C={第三辆车向右转弯}={LLR，LRR，RLR，RRR}，

D={三辆车均向右转弯}={RRR}.

（这里，例如LLR表示第一辆、第二辆车向左转弯而第三辆车向右转弯.）则有B⊂A，，BD=∅，A∪D=S，BC={LLR，RLR}，B-C={LLL，RLL}，B∪C={LLL，RLL，LLR，RLR，LRR，RRR}，，A（B∪C）={LLL，RLL，LLR，LRR，RLR}.

【例**1-2**】　　设A，B，C都是试验E的事件，试用A，B，C的运算表示下列事件：（1）G_1 ：A，B，C中至少有一个发生；（2）G_2 ：A，B，C同时发生；（3）G_3：A发生，而B，C都不发生.

解　G_1 =A∪B∪C，G_2 =ABC，.

事件是一个集合，以上事件间的关系和运算当然可以用集合论的术语来表达，而集合的运算规则也就是事件的运算规则. 例如“事件A和事件B的积事件A∩B”，“事件A与事件B不相容”，用集合论的术语来说就

是“集合A和集合B的交集A∩B”，“集合A和集合B的交集是空集”. 事件的运算满足以下的规则（这里A，B，C是试验E的事件）：

交换律　A∪B=B∪A，A∩B=B∩A;

结合律　　A∪（B∪C）=（A∪B）∪C，A∩（B∩C）=（A∩B）∩C;

分配律　　A∪（B∩C）=（A∪B）∩（A∪C），A∩（B∪C）=（A∩B）∪（A∩C）;

1.2　随机事件的概率

在一次试验中，一个事件（除不可能事件和必然事件外）可能发生也可能不发生.　我们观察试验的各个事件，一般来说，会发现有些事件在一次试验中发生的可能性较大，而另一些发生的可能性较小.　例如，在抛一颗骰子观察它的点数的试验中，事件“出现偶数点”比事件“出现2点”发生的可能性要大. 我们希望对每个事件都能指定一个数来表示事件在一次试验中发生的可能性的大小. 下面先从“频率”讲起.

1.2.1　随机事件的频率

在相同的条件下将试验重复进行n次，在n次试验中，事件A发生了f_A 次，f_A 称为事件A在这n次试验中发生的频数 ，而比值

$$R_n（A）=f_A /n$$

称为事件A在这n次试验中发生的频率 .

【例**1-3**】　将一枚硬币抛5次，10次，20次，…，5000次得到如表1-1的数据. 表中n表示试验的次数，f_H 表示在这n次试验中H（正面）发生的频数，R_n （H）表示这n次试验中H发生的频率. 我们还将表中的数据描在图1-2上.

表1−1

图1−2

从表1-1和图1-2看到，当n较小时，频率R_n （H）在0与1之间随机波动，其幅度较大；但随着n增加，波动的幅度逐渐减小，呈现出稳定性，稳定在0.5附近.

【例**1-4**】　表1-2给出了波兰从1927年到1932年间出生的婴儿总数，以及其中的男婴数. 试考察新生婴儿的性别.

表1−2

从表1-2可以看到出生男婴的频率稳定在0.517附近.

大量实验证实，随机事件A发生的频率R_n （A），当重复试验的次数n增大时，总呈现出稳定性，稳定在某一个数的附近. 这是随机现象固有的性质. “频率的稳定性”就是我们通常所说的统计规律性.

由于事件A发生的频率是它发生的次数与试验次数之比，其大小表示A发生的频繁程度，频率较大，事件A发生较频繁，这意味着A在一次试验中发生的可能性较大；反之亦然.　而频率又具有稳定性，当试验次

数n增大时，频率稳定在某一个数的附近. 这表明，对于一个事件A，存在着一个数，这个数可用来表示A发生的可能性的大小.

对于试验E的每一个事件，指定一个数，这个数称为事件的概率 ，它用来表示事件发生的可能性的大小. 事件A的概率记为P（A）.

概率是表示事件发生的可能性大小的数量指标. 那么对于E中的事件，怎样确定它的概率呢？这将留待下一小节去讨论. 这里，先讲一下频率的几条性质.

以R_n（A），R_n（B）分别表示事件A和事件B在n次重复试验中发生的频率，则有以下性质：

性质**1** 对于任意事件A，有$0 \le R_n$（A）≤ 1；

性质**2** 对于必然事件S，有R_n（S）=1；

性质**3** 若事件A，B不相容，即AB=∅，则有

$$R_n（A \cup B）=R_n（A）+R_n（B）.$$

性质1，2显然成立. 现在来证明性质3：

设在n次重复试验中，事件A，B分别发生f_A ，f_B 次，由AB=∅知$A \cup B$发生的次数$f_{A \cup B}$ 应等于$f_A + f_B$，故有

1.2.2 古典概率模型

下面考察最简单的一类随机试验，它们具有以下两个特点：

（1）试验的样本空间只包含有限个样本点；

（2）由于某种对称性，在每次试验中，各个基本事件发生的可能性相同.

例如，抛一颗骰子，观察其出现的点数就属于这一类试验. 具有这两个特点的试验称为古典概率模型 ，它在概率论发展初期曾是主要的研究对象.

对于古典概率模型，若设试验的样本空间

$$S=\{\omega_1，\omega_2，\ldots，\omega_n\}=\{\omega_1\}\cup\{\omega_2\}\cup\ldots\cup\{\omega_n\}，$$

事件A包含r个基本事件，则

$$A=\{\omega_{i_1}，\omega_{i_2}，\ldots，\omega_{i_r}\}=\{\omega_{i_1}\}\cup\{\omega_{i_2}\}\cup\ldots\cup\{\omega_{i_r}\}，$$

式中i_1 ，i_2 ，…，i_r 是1，2，…，n中某r个不同的数. 这里，基本事件$\{\omega_1\}$，$\{\omega_2\}$，…，$\{\omega_n\}$发生的可能性相同，而A包含r个基本事件$\{\omega_{i_1}\}$，$\{\omega_{i_2}\}$，…，$\{\omega_{i_r}\}$，当这r个基本事件之一发生时事件A发生，也称这r个基本事件有利于事件A. 我们自然想到用有利于事件A的基本事件数在全部基本事件数n中所占的比例r/n来表示事件A在一次试验中发生的可能性的大小.

我们称数r/n为事件A的概率，即

（1-1）式1-1称为概率的古典定义 .

【例**1-5**】　抛掷两颗骰子，观察它们出现的点数.

（1）写出试验的样本空间；

（2）设事件A为“第一颗骰子的点数为2”，事件B为“两颗骰子的点数之和为5”，求P（A），P（B）.

解　（1）样本空间

S={（1，1），（1，2），…，（1，6），（2，1）（2，2），…，

（2，6），…，（6，1），（6，2），…，（6，6）}.

式中（i，j）表示第一颗骰子为i点，第二颗骰子为j点（i，j=1，2，…，6）. 共有36个基本事件.

（2）A={（2，1），（2，2），…，（2，6）}，包含6个基本事件；

B={（1，4），（2，3），（3，2），（4，1）}，包含4个基本事件.

这是古典概率模型，故由式1-1可得

P（A）=6/36=1/6，P（B）=4/36=1/9.

【例**1-6**】　在100，101，…，999这900个3位数中，随机地取一个3位数，求不包含数字“1”的概率（这里“随机地”是指取到数100，…，999是等可能的）.

解　以A表示事件“取到的3位数不包含数字1”. 在100，101，…，

999中取一个3位数，每一种取法是一个基本事件，基本事件的总数n=900. 由题意可知各个基本事件发生的可能性相同，因此可以利用式1-1来计算概率. 由于取到的3位数要求不包含数字1，即知百位的数有8种取法，十位的数有9种取法，个位的数有9种取法，由组合法的乘法原理，取到不含数字1的3位数共有8×9×9=648种取法，即有利于事件A的基本事件数为648，故由式1-1即得

$$P（A）=648/900=18/25.$$

【例1-7】 设在100个产品中有4个次品，今从中随机地取12个，求其中恰有2个次品的概率.

解 在100个产品中取12个，所有可能的取法共有[1]种，每一种取法为一基本事件. 在4个次品中取2个次品，所有可能的取法共有种. 在96个正品中取10个正品，所有可能的取法共有种. 由组合法的乘法原理，在100个产品中取12个，其中恰有2个次品的取法共有种. 由式1-1可得所求概率

【例1-8】 袋中有a个白球，b个红球，k个人依次在袋中取一个球.

（1）作放回抽样（即前一人取一个球观察其颜色后，放回，后一人再去取球）；

（2）作不放回抽样（即前一人取一个球观察其颜色后，不放回，后一人再去取球）. 求第i（i=1，2，…，k）人取到白球（记为事件B）的概率（k≤a+b）.

解 （1）在放回抽样的情况，显然有

$$P（B）=a/（a+b）.$$

（2）在不放回抽样的情况，k个人各取到一个球，每种取法是一个基本事件，注意到各人取球的次序有先后，故考虑用排列法计算，共有个基本事件.　　当事件B发生时，由题意可知第i个人取的一个球应是白球，它可以是a个白球中的任一个，有a种取法，其余被取的k-1个球可以是其余a+b-1个球中的任意k-1个，共有种取法，于是事件B共包含个基本事件. 故由式1-1即得

值得注意的是P（B）与i无关，即k个人取球，尽管取球的次序不同，各人取到白球的概率是一样的，大家机会均等；另外放回抽样情况与不放回抽样情况P（B）也是一样的.

由概率的古典定义式1-1可知古典概率也具有与频率同样的三条性质：

性质**1**　对于任意事件A，有$0\le P（A）\le 1$；

性质**2**　对于必然事件S，有P（S）=1；

性质**3**　若事件A，B不相容，即$AB=\varnothing$，则有

$$P（A\cup B）=P（A）+P（B）.$$

事实上，由定义即可知性质1和性质2成立. 现在来证明性质3.

设试验的基本事件的总数为n，其中A所含的基本事件数为r_1 ，B所含的基本事件数为r_2 ，由$AB=\varnothing$知$A\cup B$共包含r_1+r_2 个基本事件，于是

$$P（A\cup B）=（r_1+r_2）/n=r_1/n+r_2/n=P（A）+P（B）.$$

1.2.3　概率的一般定义

古典概率模型不能用于试验有无限多个可能结果的情况，也不能用于虽然试验只包含有限个可能结果，但它们的发生不是等可能的情况. 例如1.1节中的试验2~试验5都不能使用古典概率模型.　近代概率论应用所涉及的试验多数不属于古典概率模型.

下面将引入概率的一般定义.　我们看到频率和古典概率都具有上述性质1~性质3. 频率的稳定性是我们能确定一个数（称为概率）用来表示事件发生可能性大小的依据，因而应该要求所确定的概率满足频率所满足的基本性质1~性质3. 另一方面，考虑到概率的一般定义应适用于古典概率模型，这样也要求概率满足性质1~性质3. 我们得到启发给出以下的定义：

定义　　设E是随机试验，S是它的样本空间，若集合（事件）的函数P（·）满足下列假设：

假设**1**　对于任意事件A，有0≤P（A）≤1；

假设**2**　对于必然事件S，有P（S）=1；

假设**3**　　若事件A_1 ，A_2 ，…，A_n ，…两两不相容，即A_i A_j =∅（i≠j；i，j=1，2，3，…），有

$$P(A_1 \cup A_2 \cup \ldots \cup A_n)=P(A_1)+P(A_2)+\ldots+P(A_n),$$

$$P(A_1 \cup A_2 \cup \ldots \cup A_n \cup \ldots)=P(A_1)+P(A_2)+\ldots+P(A_n)+\ldots;$$

则称P（·）为概率函数 ，简称概率 ，函数值P（A）称为事件 A的概率 .

在第3章的伯努利大数定理中，将会阐明：当试验次数n充分大时，事件A的频率R_n （A）在一定意义下接近于事件A的概率P（A）. 基于这一事实，我们能用P（A）来度量事件A在一次试验中发生的可能性的大小.

要指出的是，上述定义只给出了概率必须要满足的三条假设，并未给事件A的概率P（A）选定一个具体的数值. 只有在古典概率模型下，对于每个事件A给出了概率P（A）=r/n（即式1-1）. 一般，我们可以进行重复试验，得到事件A的频率，就以频率来作为P（A）的估计值.

由上述3条假设，可以推导出概率的下述性质：

性质**1** 设是事件A的对立事件，则有

性质**2** P（∅）=0；

性质**3** 设A，B是两个事件，且A⊂B，则有

$$P（B-A）=P（B）-P（A），$$

$$P（A）≤P（B）；$$

性质**4** 对于任意两个事件A，B，有

$$P（A∪B）=P（A）+P（B）-P（AB）.$$

证 （1）由于，由概率的假设2，3有

（2）由于S∪∅=S，S∅=∅，由假设3有

$$P（S）=P（S∪∅）$$

$$=P（S）+P（∅），$$

故得

$$P（∅）=0.$$

（3）因为A⊂B，有B=A∪（B-A）（见图1-1a）且A（B-A）=∅，由假设3得

$$P（B）=P（A）+P（B-A），$$

故有

$$P（B-A）=P（B）-P（A），$$

再由假设1可知

$$P（B）≥P（A）.$$

（4）如图1-3所示，将A∪B分成两两不相容的三个事件Ⅰ，Ⅱ，Ⅲ，即有

图1-3

$$A∪B=Ⅰ∪Ⅱ∪Ⅲ，$$

$$A=Ⅰ∪Ⅱ，$$

$$B=Ⅱ\cup Ⅲ,$$

于是

$$P（A\cup B）=P（Ⅰ）+P（Ⅱ）+P（Ⅲ），$$

$$P（A）=P（Ⅰ）+P（Ⅱ），$$

$$P（B）=P（Ⅱ）+P（Ⅲ），$$

从而

$$P（A\cup B）=P（A）+P（B）-P（Ⅱ），$$

但Ⅱ=AB，于是性质4得证.

【例1-9】 已知P（A）=1/2，P（B）=1/3，P（AB）=1/6. 求：.

解 （1）由性质1可知

（3）由性质1和性质4可得

【例1-10】 一个小班共有36名学生，其中有12人选修日语课，10人选修德语课，5人同时选修这两门课. 在这一小班中任选一人，求其至少选修这两门课程中的一门课的概率.

解 以A，B分别表示“选到的学生选修日语课”和“选到的学生选修德语课”，按题意需要求P（A∪B）. 由概率的性质4可得

1.3 条件概率与乘法公式

1.3.1　条件概率

我们将用一个例子引出条件概率这一重要的概念.

【例1-11】　袋中有4个分别编号为1，2，3，4的白球，另有3个分别编号为1，2，3的红球. 现在要在袋中随机地取一个球. 样本空间S={$1_{白}$，$2_{白}$，$3_{白}$，$4_{白}$，$1_{红}$，$2_{红}$，$3_{红}$}，其中“$1_{白}$”表示编号为1的白球. 以A记事件“取到的是白球”，以B记事件“取到的是1号球”，则有

$$A=\{1_{白}，2_{白}，3_{白}，4_{白}\},$$

$$B=\{1_{白}，1_{红}\}，AB=\{1_{白}\}.$$

我们来求已知A已经发生的条件下，事件B的概率. 本来试验的所有可能结果有7个，而有利于B发生的有2个，因而

$$P（B）=2/7.$$

而现在，我们知道了“A已经发生”这一信息，根据这一信息，不在A中的样本点，就不可能出现了，因而试验所有可能结果所组成的集合就是A.而A共有4个可能结果，其中只有结果“$1_{白}$”有利于B发生，由式1-1可知，当A已发生的条件下B发生的概率［记为P（B|A）］

$$P（B|A）=1/4.$$

在这个例子中可以看到两点：

在一般的情况下，应该怎样定义在事件A已发生的条件下事件B的

概率呢？下面先从事件的频率来分析.

设试验E的样本空间为S. A，B是E的事件，设在n次试验中事件A，AB各发生f_A，f_{AB}次，这时A已发生的条件下B发生的次数就是f_{AB}（因为A已发生，在此条件下B发生，那么B只能是伴随着A一起发生）. 因此，在A已发生的条件下B发生的频率［记为R_n（B|A）］

由此启发我们给出以下的定义：

定义　　设A，B是两个事件，且P（A）>0，则A已发生的条件下B发生的条件概率 记为P（B|A），定义为

容易验证，条件概率P（·|A）满足概率定义中的三条假设，即

（1）对于任意事件B，有0≤P（B|A）≤1；

（2）对于必然事件S，有P（S|A）=1；

（3）若事件B_1，B_2，…，B_n，…两两不相容，有

$P(B_1 \cup B_2 \cup \ldots \cup B_n \mid A)=P(B_1 \mid A)+P(B_2 \mid A)+\ldots+P(B_n \mid A)$，

$P(B_1 \cup B_2 \cup \ldots \cup B_n \cup \ldots \mid A)=P(B_1 \mid A)+P(B_2 \mid A)+$

$\ldots+P(B_n \mid A)+\ldots$

（证明作为习题）.

由于条件概率满足上述三条假设，故上一节中对概率所证明的四条

基本性质都适用于条件概率. 例如有

对于任意两个事件B，C，有

$$P（B\cup C|A）=P（B|A）+P（C|A）-P（BC|A）.$$

【例1-12】 掷两颗骰子，以A记事件“两颗骰子点数之和为7”，以B记事件“两颗骰子中有一颗为1点”，求P（B|A）.

解 样本空间S={（1，1），（1，2），…，（1，6），…，（6，1），（6，2），…，（6，6）}，共36个样本点；A={（1，6），（2，5），（3，4），（4，3），（5，2），（6，1）}，共6个样本点；AB={（1，6），（6，1）}. 由式1-2可得

我们也可直接按条件概率的含义来求P（B|A）. 因为已知事件A已发生，于是试验的所有可能结果所成的集合就是A. A中有6个样本点，其中只有2个样本点（1，6）和（6，1）属于B，故有

$$P（B|A）=2/6=1/3.$$

1.3.2 乘法公式

以P（A）乘式1-2两边，得到以下定理：

乘法定理 设P（A）>0，则有

$$P（AB）=P（B|A）P（A）.$$

（1-3）

式1-3称为乘法公式 . 它可以用来求积事件的概率.

乘法公式还能推广到多个事件的积事件的情况，例如设A，B，C是三个事件，且P（AB）>0，则有

P（ABC）=P（C|AB）P（AB）=P（C|AB）P（B|A）P（A）.

由假设P（AB）>0可得P（A）≥P（AB）>0，故上式右端有意义.

【例1-13】 一个次品与4个正品混在了一起，需要逐个进行检验将次品找出来. （1）求至少需要检验3次才能找出这个次品的概率；（2）求这个次品在第3次检验时被找出的概率.

解 以G_1 ，G_2 分别记事件第1次、第2次检验时得到正品，以D_3 记事件在第3次检验时得到次品，则有

1.4 事件的独立性

事件独立性的概念是一个十分重要的概念.

对于A，B两个事件，一般来说P（B|A）≠P（B），这就是说“A已发生”这一信息会影响B发生的概率. 然而，也有很多情况却有P（B|A）=P（B），就是“A已发生”这一信息对B发生的概率没有影响.

当P（B|A）=P（B）时，就有

P（AB）=P（B|A）P（A）=P（B）P（A），

于是我们有以下的定义：

定义　设A，B是两个事件，若

$$P（AB）=P（A）P（B），\tag{1-4}$$

则称事件A和事件B相互独立 .

定理　　设A，B是两个事件，且P（A）>0，若A，B相互独立，则P（B|A）=P（B）；反之亦然[2]

【例1-14】　加工某种零件，共需经过两道工序，设第一、第二道工序的次品率分别为0.1，0.2，且两道工序的工作是相互独立的，试求成品的次品率.

解　　以A表示事件“成品是次品”，以A_i （i=1，2）表示事件“第i道工序加工出来的是次品”. 因为当且仅当A_1 ，A_2 中至少有一个发生时A发生，即有$A=A_1 \cup A_2$ ，故有

$$P（A）=P（A_1 \cup A_2 ）=P（A_1 ）+P（A_2 ）-P（A_1 A_2 ）.$$

由独立性可得

$$P（A）=P（A_1 ）+P（A_2 ）-P（A_1 ）P（A_2 ）$$

$$=0.1+0.2-0.1\times0.2=0.28.$$

我们还可将独立性的概念推广到多于两个事件的情况.

定义　设A_1 ，A_2 ，…，A_n 是n（n>2）个事件，如果对于其中任

意2个、任意3个、……、任意n个事件的积事件的概率都等于各个事件概率之积，则称A_1，A_2，…，A_n 是相互独立的.

例如，对于3个事件A_1，A_2，A_3，若

$$P（A_1 A_2）=P（A_1）P（A_2），$$

$$P（A_2 A_3）=P（A_2）P（A_3），$$

$$P（A_3 A_1）=P（A_3）P（A_1），$$

$$P（A_1 A_2 A_3）=P（A_1）P（A_2）P（A_3），$$

则称A_1，A_2，A_3 是相互独立的.

【例1-15】 设有一系统由4个元件1，2，3，4组成，其连接方式如图1-4所示. 设各元件工作相互独立，各元件能正常工作的概率依次为p_1，p_2，p_3，p_4，求系统能正常工作的概率.

图1−4

解 以A_i（i=1，2，3，4）表示事件“元件i能正常工作”，以A表示事件“系统能正常工作”，则有

$$A=A_1（A_2 A_3 \cup A_4）=A_1 A_2 A_3 \cup A_1 A_4，$$

于是

$$P（A）=P（A_1 A_2 A_3）+P（A_1 A_4）-P（A_1 A_2 A_3 A_4），$$

由事件的独立性得

$$P（A）=P（A_1）P（A_2）P（A_3）+P（A_1）P（A_4）-$$

$$P（A_1）P（A_2）P（A_3）P（A_4）$$

$$=p_1 p_2 p_3 +p_1 p_4 -p_1 p_2 p_3 p_4 .$$

由上述定理可知，两事件A，B相互独立的直观意义是它们中一个已发生不影响另一个发生的概率. 在实际问题中，我们常从事件的实际含义去判断它们是否相互独立. 一般由实际情况分析，A，B之间的关联很微弱，那就认为它们是相互独立的. 例如以A，B分别表示事件甲、乙两地明天为晴天，若甲乙两地相距甚远就认为A，B相互独立，若两地紧挨着就不能认为A，B相互独立了.

1.5 全概率公式与贝叶斯公式

本节介绍两个与条件概率有关的公式，对某些较为复杂的情况可以用它们来计算事件的概率.

1.5.1 全概率公式

设S是试验E的样本空间，B_1，B_2，…，B_n 是E的n个两两不相容的事件，且有

$$B_1 \cup B_2 \cup \ldots \cup B_n =S,$$

也就是说将S划分成n个两两不相容的事件：B_1，B_2，…，B_n.

又若A是试验E的任一事件，则有

$$A=AS=A（B_1 \cup B_2 \cup \ldots \cup B_n）=AB_1 \cup AB_2 \cup \ldots \cup AB_n，$$

其中（AB_i ）（AB_j ）=∅（i≠j）. 这样就将A分成n个两两不相容的事件：AB_1，AB_2，…，AB_n. 设P（B_i ）>0（i=1，2，…，n），就有

$$P（A）=P（AB_1）+P（AB_2）+\ldots+P（AB_n）$$

$$=P（A|B_1）P（B_1）+P（A|B_2）P（B_2）+$$

$$\ldots+P（A|B_n）P（B_n）.$$

（1-5）

式1-5称为全概率公式.

全概率公式的意义在于：在较复杂的情况下，直接求P（A）不容易，可适当地选取一组两两不相容的事件B_1 ，B_2 ，…，B_n ，且有$B_1 \cup B_2 \cup \ldots \cup B_n$ =S，将求P（A）的问题归结为求P（AB_i ）（i=1，2，…，n）的问题，若能容易求出P（$A|B_i$ ）和P（B_i ）就能得到P（AB_i ），由此便得到P（A）.

【例1-16】 一运输系统，在不同运输量水平下其故障率不相同. 在低运输量水平、中等运输量水平、高运输量水平下其故障率分别为0，0.1，0.5. 已知运输系统处于低、中等、高运输量水平的概率分别为

0.6，0.3，0.1，求运输系统的故障率.

解　记A为事件系统出现故障，分别记B_1，B_2，B_3为事件路线处于低、中等、高运输量水平. 按题意有

$$B_1 \cup B_2 \cup B_3 = S，B_i B_j = \varnothing（i \neq j），$$

且有　$P（B_1）=0.6$，$P（B_2）=0.3$，$P（B_3）=0.1$，$P（A|B_1）=0$，$P（A|B_2）=0.1$，$P（A|B_3）=0.5$，由全概率公式可得

1.5.2　贝叶斯（Bayes）公式

在全概率公式的各个假定下，又设$P（A）>0$，则由条件概率的定义可有

式1-6称为贝叶斯公式，这是一个用来求条件概率的重要公式.

【例1-17】　在例1-16中，若已观察到发生故障，求此时路线处于中等运输量水平的概率.

解　按题意需求$P（B_2|A）$，由例1-16可得$P（A）=0.08$，由式1-6有

特别在式1-5，1-6中取n=2，并将B_1记为B，此时B_2就是，那么全概率公式和贝叶斯公式分别成为

这两个公式是常用的.

【例**1-18**】　一通信通道，使用信号“0”和“1”传输信息. 以A记事件收到信号“1”，以B记事件发出信号“1”.　　已知P（B）=0.4，P（A|B）=0.95，. 现在已知收到的是信号“1”，求发出的是信号“1”的概率.

解　需求的是P（B|A）. 由假设可知P（）=0.6，0.1，则由式1-8有

习题1

【答案链接】

1. 写出下列随机试验的样本空间：

（1）连续投掷一枚硬币直至正面出现为止，观察正反面出现的情况；

（2）连续投掷一枚硬币直至正面出现为止，记录投掷的次数；

（3）投掷一颗骰子直至6个结果中有一个结果出现2次为止，记录投掷的次数.

2.（1）设A，B是两个事件，A，B不相容.　　已知P（A）=0.35，P（B）=0.15，求P（A∪B），.

（2）设A，B是两个事件，已知P（A）=0.2，P（B）=0.3，P（A∪B）=0.4，求P（AB），.

3. 一袋中装有编号为1，2，3，4的4个球. 在袋中取球2次，每次随机地取一个，考虑两种情况：（1）放回抽样；（2）不放回抽样. 分别就这两种情况求两次均未取到1号球的概率.

4. 在11张卡片上分别写上engineering这11个字母，从中随机地抽出2张，求这2张卡片上分别写着e和g的概率.

5. 某地区的电话号码是7位数，前三位固定为452，后四位分别可在0，1，…，9的10个数中等可能地取一个. 求电话号码7个数字不相同的概率.

6. 10个人去钓鱼，共钓得3条鱼（设已钓得的鱼是被各人钓到的可能性相同）.

（1）求3条鱼是由同一个人钓得的概率；

（2）求3条鱼是由3个不同的人钓得的概率.

7. 20个产品中有5个是次品，从中随机地取4个.

（1）求恰有1个是次品的概率；

（2）求至少有1个是次品的概率.

8. 一盒子中有10个产品，其中有4个是次品，每次随机地取1个进行检验，直到这4个次品都找到为止. 求第4个次品在第6次检验时被发现的概率.

9. 一袋中有3个白球、4个红球和5个黑球，在袋中任取7个球，求其中有2个白球、3个红球和2个黑球的概率.

10. 在1~100的整数中，随机地取一个数，求取到的数能被2或3整除的概率.

11. 设P（A）=0.5，P（B）=0.3，P（AB）=0.1. 求P（A|B），

P（B|A），P（A|A∪B），P（A|AB）.

12. 疾病Ⅰ和Ⅱ在某地区流行，有10%的人患过疾病Ⅰ，有15%的人患过疾病Ⅱ，且有3%的人患过这两种病.

（1）在人群中任选一人，求他至少患过其中一种病的概率；

（2）在人群中任选一人，已知他至少患过其中一种病，求他患过这两种病的概率.

13. 一袋中有3个红球、7个白球，从袋中任意取出1个球，然后放进1个另一颜色的球（例如取出1个白球就放进1个红球）.

（1）如此连续取3个球，求第一次取出的是红球，而第二次、第三次取出的是白球的概率；

（2）如此连续取出2个球，已知取出的2个球具有相同的颜色，求它们都是白球的概率.

14.（1）设A，B是两个事件，已知A，B相互独立. 证明相互独立.

（2）设A，B，C是3个相互独立的事件. 验证A∪B与C相互独立；AB与C相互独立.

（3）设P（A）=0.5，P（B）=0.2，A，B相互独立. 求P（A∪B），.

15. 3人进行射击，击中目标的概率分别为1/6，1/4，1/3. 求：

（1）恰有一人击中；

（2）恰有两人击中；

（3）至少有一人击中的概率（设各人是否击中相互独立）.

16. 一枚二级火箭执行一项空间任务. 已知火箭起飞失败的概率为0.1；若起飞成功，一、二级火箭分离失败的概率为0.05；若一、二级火箭分离成功，第二级火箭失败的概率为0.03.

（1）求火箭完成整个任务的概率；

（2）求执行任务失败的概率.

17. 一系统由6个独立工作的元件1，2，3，4，5，6组成，各元件的连接方式如图所示. 设各元件能正常工作的概率依次为p_1，p_2，p_3，p_4，p_5，p_6，求系统能正常工作的概率.

18. 某间房门上锁的概率为1/2，这个门上的钥匙是架子上12把钥匙中的一把. 某人要进该房门，若门锁着，那么他在架子上任意取两把钥匙去开门，求他能打开门的概率.

19. 编号为1，2，3的三台仪器正在工作的概率分别为0.9，0.8和0.4，在其中任选一台：

（1）求选到的一台仪器正在工作的概率；

（2）已知选到的仪器正在工作，分别求这台仪器是编号为1，2，3的仪器的概率.

20. 某一城市有25%的汽车废气排放量超过规定，一废气排放量超标的汽车有0.99的概率不能通过城市检验站的检验，而一废气排放量未

超标的汽车也有0.17的概率不能通过检验. 求一辆未通过检验的汽车，它的废气排放量是超标的概率.

21. 一货栈存有分别装有高、中、低质量灯管的盒子，盒子只数的比例为1:2:2. 这三种灯管不合用的概率依次为0，0.1和0.2. 随机地取一盒子，在其中任取2支灯管进行测试，结果2支都合用. 求选出的盒子所装的灯管的质量分别是高、中、低的概率（设灯管测试所得的结果相互独立）.

---

[1] 记号表示自n个元素中取k个元素的组合数，即.

[2] 在定理中A，B的位置互换，结论当然也成立. 即若P（B）>0，且A，B相互独立，则P（A|B）=P（A）；反之亦然.

2　随机变量

2.1　随机变量的概念

对于随机试验，常常有这种情况，人们不是关心试验的各个结果本身，而是对于与试验结果联系着的某个数感兴趣.

【例2-1】　在进行掷两颗骰子观察其点数的试验时，其样本空间S={（1，1），（1，2），…，（1，6），…，（6，1），（6，2），…，（6，6）}，以X记两骰子点数之和. 投掷者只关心X的取值，例如他关心的是X=5，而不在乎出现的是（1，4），（2，3），（3，2）还是（4，1）. 在这里，对于试验的每一个结果都有一个实数与之对应，我们将对应关系列成如下的表格.

表2-1

我们看到，对于试验的每一个结果ω=（i，j）∈S，X都有一个指定的实数i+j与之对应，因而X是定义在样本空间S上的一个实值函数，使用函数的记号可写成

$$X((i，j))=i+j，i，j=1，2，\ldots，6.$$

【例2-2】　一病毒专家对5个人进行某种疫苗接种试验，观察每个人的反应情况，其样本空间为S={（x_1，x_2，x_3，x_4，x_5），x_i =0，

1，i=1，2，3，4，5}，其中x_i 取1或0，x_i =1表示第i人反应为阳性，x_i =0表示第i人反应为阴性. 以X记具有阳性反应的人数. 病毒专家只关心X取什么值而对哪一个人反应为阳性不在乎. 例如他不关心出现的结果是（0，1，0，1，0）或是（1，1，0，0，0）或是（0，1，1，0，0），等等，而只关心当这些点出现时X=2. 在这里，对于试验的每一个结果，都有X的一个指定的实数与之对应. X是定义在样本空间S上的实值函数. 具体写出来就是

$$X（（x_1，x_2，…，x_5））=x_1+x_2+…+x_5 .$$

【例2-3】 一运动员进行打靶，观察弹着点的位置. 靶可认为是中心为（0，0）半径为r的圆盘，并设射击不会脱靶. 样本空间为S={（x，y）$|x^2+y^2 \le r^2$ }，人们并不关心弹着点（x，y）的具体位置而只关心弹着点离靶心（0，0）的距离X. X是定义在样本空间S上的实值函数. 具体写出来就是

定义 设随机试验的样本空间为S，对于试验的每一个结果ω∈S，X都有一个指定的实数X=X（ω）与之对应，则称X为随机变量 .

根据定义，随机变量是定义在样本空间上的实值单值函数.

在本书中，我们以大写的字母X，Y，Z，W等表示随机变量，而以小写字母x，y，z，w等表示实数.

有许多试验它的结果ω本身就是一个数，我们令X=X（ω）=ω，那么X是一个随机变量. 例如，用Y记某医院一天中挂号就诊的病人数；以Z记某邮局一天收到的信件数；以W记某种灯泡的寿命，那么，Y，Z，W都是随机变量.

随机变量是一个变量，它的取值随试验的结果而定，在试验之前，人们只知道它可能取值的范围，而不知道它取什么值. 又由于试验的各个结果在一次试验中发生有一定的概率，因而随机变量取各个值有一定的概率. 例如在例2-1中，X取值为5，记为{X=5}，对应于样本点的集合{（1，4），（2，3），（3，2），（4，1）}，这是一个随机事件，当且仅当这一事件发生时有{X=5}，于是我们就称概率P{（1，4），（2，3），（3，2），（4，1）}为{X=5}的概率，即

$$P\{X=5\}=P\{（1，4），（2，3），（3，2），（4，1）\}=4/36=1/9.$$

以后，还将事件{（1，4），（2，3），（3，2），（4，1）}说成是事件{X=5}. 类似地有

$$P\{X\geq 11\}=P\{（5，6），（6，5），（6，6）\}=3/36=1/12，$$

$$P\{X=2\}=P\{（1，1）\}=1/36.$$

引入了随机变量后，不论对什么样的随机现象都可以用随机变量来描述. 这使我们有可能用微积分的方法对各种有关的问题进行深入的讨论.

2.2 离散型随机变量与连续型随机变量

2.2.1 离散型随机变量

有些随机变量，它所有可能取的值是有限个或可列无限个，这种随机变量称为离散型随机变量 .

如例2-1，例2-2中的随机变量，它所有可能取的值是有限个，它们是离散型随机变量. 又如抛一枚硬币直到出现H为止，以X表示所需抛掷的次数，则X可能取的值为1，2，3，…，是可列无限个，X是离散型随机变量.

对于离散型随机变量X来说，我们只要将X所有可能取的值以及取各个值的概率说清楚，那么就将随机变量X完全描述清楚了.

定义　设离散型随机变量X所有可能取的值为x_1，x_2，…，x_n，…，取各个可能值的概率分别为p_1，p_2，…，p_n，…，且满足条件：

即

$$P\{X=x_k\}=p_k \quad (k=1，2，\ldots)，$$

（2-3）

则称式2-3为随机变量X的概率分布律，简称分布律 . X的分布律还可写成如下的表格形式：

上表直观地表示了随机变量X取各个值的概率的规律. X取各个值各占一些概率，这些概率合起来是1. 可以想象成：整个概率1以一定的规律分布在各个可能值上，这就是称以上表格为分布律的原因.

【例2-4】　写出例2-1中随机变量X的分布律.

解　由表2-1容易写出X的分布律如下：

我们还能用图形来表示X的分布律. 第一种表示法：将X的取值2，3，…，12描在Ox轴上，再在各点处作Ox轴的垂线使其长度等于X取该点值的概率，如图2-1所示. 这种图形称为棒图 . 第二种表示法：在Ox轴上方作小长方形，使小长方形的高度等于X取底边中点值的概率，而底边长取为1，每个小长方形的面积就等于X取底边中点值的概率，所有长方形的面积之和为1，如图2-2所示. 这种图形称为概率直方图 .

图2-1

图2-2

【例2-5】 在句子“IT IS TOO GOOD TO BE TRUE”中随机地取一个单词，以X表示取出的单词中包含的字母数，求X的分布律（认为取到各个单词是等可能的）.

解 先将试验的样本空间S以及X的取值情况列出如下：

X所有可能取的值为2，3，4，取这些值的概率分别为4/7，1/7，2/7，得X的分布律为

X	2	3	4
p_k	4/7	1/7	2/7

其棒图和概率直方图分别如图2-3和图2-4所示.

图2-3

图2-4

下面介绍两种重要的离散型随机变量.

1）二项分布

设试验E只有两个可能结果：A和，记P（A）=p（0<p<1），称E是一个伯努利试验 .

将试验E独立地重复n次，这一串n次试验称为n重伯努利试验 . 这里“重复”是指每次伯努利试验中P（A）=p保持不变，“独立”是指每次试验A发生与否与其他各次试验的结果无关，即若记A_i 为第i次试验A发生，则事件C_1 ，C_2 ，…，C_n （其中C_i 为A_i 或，i=1，2，…，n）相互独立.

例如，E是抛一枚硬币，观察出现正反面的情况，则E是一个伯努利试验，A表示抛得正面. 如将硬币抛n次，就是n重伯努利试验. 又如在袋中装有a个白球，b个黑球，在袋中随机地取一个球观察其颜色，放回，然后再取一个球（即作放回抽样），如此重复n次. 将取一个球观察其颜色作为一次试验，则n次试验是n重伯努利试验. 设A是事件“取到白球”，P（A）=a/（a+b）. 然而若作不放回抽样（n≤a+b），那么，虽然每次试验都有P（A）=a/（a+b），但各次试验不再是相互独立 [1] ，因而不再是n重伯努利试验了.

现作一n重伯努利试验. 设在试验中事件A的概率为p（0<p<1），X表示在n重伯努利试验中事件A发生的次数. X是一个随机变量，X所有可能取的值为0，1，2，…，n，现在来求P{X=k}（k=0，1，…，n）.

由于各次试验是相互独立的，因此事件A在指定的k次试验中发生而在其他n-k次试验中不发生的概率就是k个p的乘积再乘以n-k个1-p之积，即为p^k （1-p）$^{n-k}$. 又由于这种指定的方式有种，它们是两两不相容的，故在n次试验中A恰发生k次的概率

显然，且=［p+（1-p）］^{n} =1，因而P{X=k}（k=0，1，2，…，n）满足条件2-1，2-2. 这样，我们有以下的定义：

设随机变量X具有分布律为

其中0<p<1为常数，则称X服从以n，p为参数的二项分布 [2] ，记为X~B（n，p）.

特别当n=1时式2-4成为

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

（2-5）

或写成

X	0	1
p_k	1-p	p

此时称X服从以p为参数的伯努利分布 或称（0-1）分布 .

【例2-6】 在例2-2中设反应为阳性的概率p=0.45，且设各人的反应相互独立，以X记5人中反应为阳性的人数. （1）写出X的分布律；（2）求恰有3人反应为阳性的概率；（3）求至少有2人反应为阳性的概率.

解 将观察一个人反应是阳性或阴性看成是一次试验，这是伯努利试验. 按题意可知X~B（5，0.45），于是

（1）X的分布律为

【例2-7】　　有一大批产品已知其次品率为0.05，现在随机地抽10个，求其中至多有2个次品的概率.

解　　将抽取一个产品观察它是否为次品作为一次试验，连续抽10个，看成是连续进行了10次试验，由于是不放回抽样，因而各次试验不相互独立，不是10重伯努利试验. 但由于产品的总数很多，而只需抽出10个（即总数大大多于10），故可近似地作为放回抽样来处理，即可近似地看成是10重伯努利试验. 记X为抽出的10个中包含的次品数，即有X~B（10，0.05），于是所求概率

2）泊松（**Poisson**）分布

设随机变量X的分布律为

其中λ>0为常数，则称随机变量X服从以λ为参数的泊松分布　　，记为X~π（λ）.

易知（k=0，1，2，…），且$e^{-\lambda}$ e^{λ} =1，即P{X=k}（k=0，1，2，…）满足条件2-1和2-2.

【例2-8】　某种放射性物质，在一毫秒中所放射出经过计数器的α粒子数X，服从以4为参数的泊松分布.　　求在一给定的毫秒中X=6的概率，并求P{X≥3}.

解　X的分布律为

服从或近似服从泊松分布的随机变量在实际中是很多的. 例如，某医院一天中的急诊病人数，一天中某地区拨错号的电话呼唤次数，一本

书一页中的印刷错误数等都服从或近似服从泊松分布.

2.2.2 连续型随机变量

上面讨论的离散型随机变量，它所有可能取的值是有限个或者是可列无限个. 而另外有许多随机变量，例如灯泡的寿命、测量的误差、乘客在公共汽车站的候车时间等，它们可能取的值充满一个区间，是不可列的. 它们不是离散型随机变量. 若随机变量不是离散型的，由于它可能取的值不能一一列举出来，因此不再能用分布律来描述，就需要另想办法来对它们进行描述. 下面我们由一个例子引出用来描述“连续型”随机变量的“概率密度函数”.

【例2-9】 一射击运动员进行射击，设靶是中心在原点半径为r的圆盘，且射击不会脱靶. 以X记弹着点到靶心的距离，X可能取的值充满整个区间［0，r］，X显然不是离散型随机变量. 我们先取cm作为度量距离的单位，X取整数值，这样就将X的取值离散化，从而得到一个离散型随机变量. 据运动员以前射击的成绩，可以写出这一离散型随机变量的分布律，然后作出对应的概率直方图（如图2-5所示）. 接着取0.5cm作为度量距离的单位，又得到一个离散型随机变量，又可作图2-5出对应的概率直方图（如图2-6所示）. 这样继续缩小度量距离的单位，作出一系列的概率直方图，这些直方图顶部的台阶型曲线趋于一条光滑曲线C：y=f（x）（如图2-7所示），由曲线C的形成过程，可以看出位于Ox轴的上方，曲线C之下的整个面积等于1，而区间［a，b］上曲线C下的面积恰为X落在［a，b］的概率. 这就是说我们得到一个函数f（x），它具有性质f（x）≥0，，P{a≤X≤b}=，其中0≤a<b≤r. 我们对这样的函数f（x）感兴趣. 为说明问题方便起见，将f（x）的定义域扩充为

（-∞，∞），即当x∈（-∞，0）∪（r，∞）时定义f（x）=0.　由以上讨论知道，对于随机变量X，有一个函数f（x），使得X落在任意区间［a，b］的概率

图2-5

图2-6

图2-7

我们就利用f（x）来描述随机变量X.

我们给出以下的定义：

定义　　设X是随机变量，如果存在在整个实数轴上的可积函数f（x），满足

且对于任意两个实数a，b（a≤b），a也可为-∞，b也可为∞，有

则称X是连续型随机变量，而f（x）称为X的概率密度函数，简称概率密度.

按定义，X取实轴上的任一区间上的值的概率都能通过f（x）在该区间上的积分求得，从这一意义上来说，f（x）完整地描述了随机变量X的性质.

本书只讨论两类常见的随机变量——离散型随机变量和连续型随机变量.

图2-8画出了一条典型的概率密度函数的曲线，它的图形位于Ox轴的上方；位于Ox上方曲线下方的整个面积为1（见图2-8a）；概率

P{a≤X≤b}等于［a，b］上的曲边梯形的面积（见图2-8b）. 随机变量以一定的概率取Ox轴上任一区间的值，这表示整个概率1以一定的规律分布在各个区间上.

图2-8

在式2-9中令a=b，得

$$P\{X=a\}=0，$$

（2-10）

这表示连续型随机变量X取任一个特定值a的概率均为0. 因此，对于连续型随机变量来说，在计算它落在某一区间的概率时，可以不必考虑该区间是否包括端点. 例如有

$$P\{a\le X\le b\}=P\{a<X\le b\}=P\{a<X<b\}.$$

再者，若设f（x）在点x连续，则有

从式2-11可见，概率密度的定义与物理学中线密度的定义相似. 这就是“概率密度”这一名称的来源. 又由式2-11可知，若不计高阶无穷小，则有

$$P\{x\le X\le x+\Delta x\}\approx f（x）\Delta x，$$

这表示，X落在小区间［x，x+Δx］上的概率近似地等于f（x）Δx.

注意到在式2-9中，改变被积函数f（x）在有限个点的函数值不影响积分的值，亦即不影响概率P{a≤X≤b}. 因此，对于概率密度函数而言，改变它在有限个点的值，是被允许的.

【例**2-10**】　　从甲地到乙地的某一航班，飞机晚点或早到的时间X（以min计）是一个随机变量.　　X取负值表示早到，X取正值表示晚点，已知X的概率密度

其中k是某一常数，试确定k，并求事件（1）至少早到2min，（2）至少晚点1min，（3）早到1~3min的概率.

解　由式2-8可知

求出k=1/288. 由式2-9可得

下面介绍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具有概率密度函数

则称X在区间（a，b）上服从均匀分布，记为X~U（a，b）.

容易验证f（x）符合条件2-7和2-8. f（x）的图形如图2-9所示.

图2-9

易知X落在区间（a，b）内任意等长小区间上的概率相等［如图2-9所示，在（a，b）内两个等长小区间上，f（x）之下的小长方形面积相等］，这就是称为均匀分布的原因.

2）指数分布

设连续型随机变量X具有概率密度

其中β>0为常数，则称X服从以β为参数的指数分布，记为X~E（β）.

易知函数f（x）满足条件2-7和2-8. 图2-10中分别画出了β=1/3，1，2时f（x）的图形.　　服从指数分布的随机变量在工程技术中是经常遇到的，例如电子元件的寿命往往服从指数分布.

图2-10

【例2-11】　　某种轮胎在损坏之前所能行驶的路程X（以1000km计）是一个随机变量，已知它的概率密度

试求：（1）P{4≤X≤5}；（2）P｛X≤30｝.

这表明轮胎行驶4000~5000km之间损坏的概率仅为0.06，轮胎能行驶的路程不大于30000km的概率为0.95.

3）正态分布

设连续型随机变量X，具有概率密度

其中μ和σ（-∞<μ<∞，σ>0）为常数，则称X服从以μ，σ为参数的正态分布，又称高斯（Gauss）分布. 记为X~N（μ，σ^2）.

因此f（x）满足条件2-8.

正态分布是概率论与数理统计中最为重要的分布，我们将在第4章专门加以讨论.

2.3　分布函数

以上我们引入了随机变量的分布律以及概率密度函数，分别用来描

述离散型和连续型随机变量. 这一节我们将引入“分布函数”，它可以用来描述任一类随机变量，从而给我们研究随机变量带来了方便.

定义 设X是随机变量，x是任意实数，函数

$$F(x)=P\{X\le x\}\quad(-\infty<x<\infty)\tag{2-14}$$

称为X的分布函数 .

若将X看成是数轴上随机点的坐标，那么，分布函数F（x）在x处的函数值F（x）表示随机点X落在区间（-∞，x］的概率.

分布函数具有以下性质：

（1）F（x）是一个单调不减函数.

事实上，对于任意实数x_1，x_2（$x_1<x_2$），由于事件$\{X\le x_2\}$包含事件$\{X\le x_1\}$，因而$P\{X\le x_1\}\le P\{X\le x_2\}$，即$F(x_1)\le F(x_2)$.

（2）$0\le F(x)\le 1$，且

可以从几何上来说明这一性质，在图2-11中将区间的右端点x沿Ox轴向右无限移动（即$x\to\infty$），则事件“点X落在点x左边”趋于必然事件，从而其概率趋于1，即

图2-11

（3）F（x）是一个右连续函数，也就是说，对于任意实数x，有

证略.

可以证明，满足上述性质（1），（2），（3）的函数F（x），必定是某个随机变量的分布函数.

对于定义在整个实轴上的可积函数f（x），若它满足条件f（x）≥0，-∞<x<∞及，作，可知G（x）是某一随机变量X的分布函数，而f（x）是X的概率密度.

有了分布函数，对于任意实数a，b（a<b），由于

$$P\{a<X\leq b\}=P\{X\leq b\}-P\{X\leq a\},$$

即有

$$P\{a<X\leq b\}=F（b）-F（a），$$

（2-15）

这就是说，X落在任意区间（a，b］的概率，都能借助于分布函数来计算.

设X是离散型随机变量，其分布律为

$$P\{X=x_k\}=p_k \quad （k=1，2，…），$$

则由F（x）的上述几何解释，即分布函数在x处的函数值F（x）就是点X落在区间（-∞，x］的概率，可知

这里和式中∑是对于所有满足条件x_k ≤x的k求和. F（x）在点$x=x_k$

（k=1，2，…）处有跳跃点，其跳跃值p_k =P{X=x_k }．一般来说F（x）的图形是一条台阶形曲线.

【例**2-12**】　设随机变量X的分布律如下：

（1）求X的分布函数，并画出它的图形；

（2）求概率P{-3<X≤}，P{1≤X<4}.

解　（1）当x<-1时，F（x）=P{X≤x}=P（∅）=0；

当-1≤x<1时，F（x）=P{X≤x}=P{X=-1}=；

当1≤x<2时，

F（x）=P{X≤x}=P{X=-1}+P{X=1}=；

当x≥2时，F（x）=P{X≤x}=P（S）=1，即

F（x）的图形如图2-12所示，它是一条台阶形曲线，在x=-1，1，2处有跳跃点，跳跃值分别为1/2，1/3，1/6.

图2-12

（2）由式2-15可有

设X是连续型随机变量，其概率密度为f（x），由式2-14则有

如图2-13所示在点x的函数值，等于曲线f（x）之下，Ox轴的区间（-∞，x］之上的曲边梯形的面积．根据数学分析的知识知道，连续型随机变量的分布函数F（x）在（-∞，∞）是连续的，而且在f（x）的连续点

处F（x）的导数等于f（x），即

图2-13

【例**2-13**】　（1）设X服从均匀分布，其概率密度

求X的分布函数.

（2）设X服从指数分布，其概率密度

求X的分布函数.

解　（1）

即

其图形如图2-14所示.

图2-14

（2）当x≤0时，有

当x>0时，

故有

其图形如图2-15所示.

图2-15

【例**2-14**】　某种型号电子元件的寿命X（以h计）是一个随机变量，其分布函数

（1）求P{X>1500}，P{1500<X<2000}；（2）求X的概率密度f（x）.

解　（1）P{X>1500}=1-P{X≤1500}

=1-F（1500）=2/3，

P{1500<X<2000}=P{1500<X≤2000}

=F（2000）-F（1500）=1/6.

（2）在x≠1000处F（x）具有连续导数，故有

在x=1000处f（x）可取任意给定的值，例如取f（1000）=0，即有概率密度

以后，当我们提到一个随机变量X的“概率分布”时，指的是它的分布函数；或者，当X是离散型时，指的是它的分布律，当X是连续型时，指的是它的概率密度.

2.4　二维随机变量

以上讨论了单个随机变量的问题，在实际中许多试验需要同时观察或测量两个或多个随机变量. 例如，在研究一个家庭的生活水平时，需要同时观察家庭的月收入（X）和月支出（Y），这里X，Y都是随机变量. 在研究一条生产线的效益时，不但要记录一个生产周期中生产的合格品数（X），还要记录需返工的产品数（Y）和废品数（Z），这里X，Y，Z都是随机变量. 对水资源的污染作研究时，需同时测量水体中各种污染物的含量，这里各种污染物的含量都是随机变量. 在以上各种

情况中，我们需要同时去研究有关两个或多个随机变量的问题.

一般，设试验E的样本空间S={ω}，又设X=X（ω），Y=Y（ω）是定义在S上的随机变量，由它们组成的向量（X，Y）称为二维随机变量或二维随机向量 . 一般，设X_1 =X_1 （ω），X_2 =X_2 （ω），…，X_n =X_n（ω）都是定义在S上的随机变量，则向量（X_1 ，X_2 ，…，X_n ）称为n维随机变量 或n维随机向量 . 以前讨论的单个随机变量也称一维随机变量 . 以下我们着重讨论二维随机变量.

与一维情况类似，若二维随机变量（X，Y）所有可能取的值是有限个数偶或可列无限个数偶时，称为二维离散型随机变量 . 与一维情况类似，我们也引入二维分布律来描述二维离散型随机变量.

设二维随机变量（X，Y）所有可能取的数偶为（x_i ，y_j ）（i，j=1，2，…）；（X，Y）取（x_i ，y_j ）的概率为p_{ij} . 记P{（X=x_i ）∩（Y=y_j ）}=P{X=x_i ，Y=y_j }，称

$$P\{X=x_i，Y=y_j\}=p_{ij}\ （i，j=1，2，…）$$

（2-19）

为二维随机变量（X，Y）的分布律 ，也称为随机变量X和Y的联合分布律 . 我们也可用如下的表格来表示（X，Y）的分布律：

由概率的定义知p_{ij} 满足条件

【例2-15】 在句子“IT IS TOO GOOD TO BE TRUE”中随机地取一个单词，以X，Y分别记所取单词中含字母的个数和含字母O的个数，

试写出（X，Y）的分布律.

解　先将试验的样本空间S以及X，Y的取值情况列出如下：

X所有可能取的值为2，3，4；Y所有可能取的值为0，1，2.　容易得到（X，Y）取（i，j），i=2，3，4，j=0，1，2的概率. 例如：

$$P\{X=2，Y=0\}=3/7，P\{X=3，Y=1\}=0，$$

$$P\{X=4，Y=2\}=1/7.$$

于是得到（X，Y）的分布律如下：

与一维随机变量类似，如果存在一个定义在全平面上的可积函数f（x，y），满足

则称（X，Y）为二维连续型随机变量　，而f（x，y）称为二维随机变量（X，Y）的概率密度函数　或称为随机变量X和Y的联合概率密度函数 .

【例2-16】　设（X，Y）有概率密度

（1）试确定常数k；（2）求P{Y<X/2}.

解　（1）如图2-16所示，仅在区域D上有f（x，y）≥0，否则f（x，y）=0. 由式2-23可得

图2-16

（2）将（X，Y）看成是平面上随机点的坐标，事件{Y<X/2}={（X，Y）∈D_0 }，其中D_0 是直线y=x/2下方的无限区域. 因此由式2-24

可得

定义　设（X，Y）是二维随机变量，x，y是任意实数，二元函数

称为二维随机变量（X，Y）的分布函数　，或称为随机变量X和Y的联合分布函数 .

如果将随机变量（X，Y）看成是平面上随机点的坐标，那么，分布函数F（x，y）在点（x，y）处的函数值就是随机点（X，Y）落在如图2-17所示的，以点（x，y）为顶点而位于该点左下方的无穷矩形域内的概率.

图2-17

2.5　边缘分布

设（X，Y）是一个二维随机变量，其分量X，Y都是一维随机变量. 我们自然会问（X，Y）的分布与X，Y各自的分布存在着什么关系？这一节就来讨论这一问题.

设二维离散型随机变量（X，Y）具有分布律为

$$P\{X=x_i，Y=y_j\}=p_{ij}\ （i，j=1，2，…），$$

则X和Y分别为离散型随机变量，它们的分布律可求得如下.　　由于{（$Y=y_1$）∪（$Y=y_2$）∪…∪（$Y=y_j$）∪…}是必然事件，故有

$$\{X=x_i\}=\{X=x_i\}\{（Y=y_1）∪（Y=y_2）∪…∪（Y=y_j）∪…\}$$

$=\{X=x_i，Y=y_1\}\cup\{X=x_i，Y=y_2\}\cup\ldots\cup\{X=x_i，Y=y_j\}\cup\ldots$

上式右端各事件两两不相容，故得X的分布律为

同样可得Y的分布律为

式2-25和式2-26分别称为（X，Y）关于X，关于Y的边缘分布律 .

设二维连续型随机变量（X，Y）具有概率密度f（x，y），则X和Y分别为连续型随机变量，它们的概率密度可求得如下：

对于任意区间［a，b］由式2-24可得

与一维随机变量的概率密度的定义比较，即得X的概率密度

同样可得Y的概率密度

f_X（x），f_Y（y）分别称为（X，Y）关于X，关于Y的边缘概率密度 .

【例2-17】 某计算站一天中死机（因机器故障而中止运行）的次数X与操作员出错次数Y的联合分布律如下：

由式2-25可知，在上表中同一行各数相加就得到（X，Y）关于X的边缘分布律，即有

$$P\{X=0\}=0.40+0.15+0.02=0.57，$$

$$P\{X=1\}=0.30+0.05+0.01=0.36，$$

$$P\{X=2\}=0.04+0.03+0=0.07.$$

于是得（X，Y）关于X的边缘分布律为

X	0	1	2
p_k	0.57	0.36	0.07

同样在上表中同一列各数相加得到（X，Y）关于Y的边缘分布律为

Y	0	1	2
p_k	0.74	0.23	0.03

人们常将边缘分布律写在联合分布律表格的边缘上，如上表所示，这就是“边缘分布律”这一名称的来源.

【例2-18】 在例2-16中求二维随机变量（X，Y）关于X和关于Y的边缘概率密度f_X（x）和f_Y（y）.

解 （X，Y）的概率密度

如图2-18所示，当-1≤x≤1时，有

图2-18

当x<-1，x>1时，f_X（x）=0，故有

而

根据上面的讨论得知，已知二维随机变量（X，Y）的分布便能确定一维随机变量X，Y各自的分布. 然而，一般来说，已知X，Y各自的

分布却不能确定（X，Y）的分布. 下面举一个例子.

【例**2-19**】　设随机变量（X_1 ，Y_1 ）和（X_2 ，Y_2 ）的分布律以及相应的边缘分布律为

我们看到这两个二维随机变量的分布律是不相同的，然而却具有相同的边缘分布律.　这表明由边缘分布律一般是不能确定二维随机变量的分布的.

2.6　条件分布

以下我们由条件概率引出条件分布的概念.

设（X，Y）是二维离散型随机变量，其分布律为

$$P\{X=x_i，Y=y_j\}=p_{ij}\ （i，j=1，2，…），$$

（X，Y）关于X的边缘分布律为

设对于某一固定的i，$P\{X=x_i\}>0$，我们来研究在事件$\{X=x_i\}$已发生的条件下，事件$\{Y=y_j\}$（j=1，2，…）的条件概率：

易知上述条件概率具有分布律的性质：

于是，我们引入以下的定义：

定义　设（X，Y）是二维离散型随机变量，具有分布律为

$$P\{X=x_i，Y=y_j\}=p_{ij}\ （i，j=1，2，…），$$

对于固定的i，若$P\{X=x_i\}>0$，则称

为在$X=x_i$条件下随机变量Y的条件分布律 .

同样，对于固定的j，若$P\{Y=y_j\}>0$，则称

为在$Y=y_j$条件下随机变量X的条件分布律 .

【例2-20】 在例2-17中，求在一天中死机的次数X=1的条件下，操作员出错次数Y的条件分布律$P\{Y=k|X=1\}$（k=0，1，2）以及条件分布律$P\{X=k|Y=0\}$（k=0，1，2）.

解 题中X，Y的联合分布律和边缘分布律如下：

得在X=1的条件下Y的条件分布律：

或写成

Y=k	0	1	2
P{Y=k|X=1}	30/36	5/36	1/36

同样，得在Y=0的条件下X的条件分布律为

X=k	0	1	2
P{X=k|Y=0}	40/74	30/74	4/74

设（X，Y）是二维连续型随机变量，这时X，Y都是连续型随机变量，因而对于任意x，y有P{X=x}=0，P{Y=y}=0，这样就不能直接利用条件概率公式引入“条件概率密度”的定义了．设（X，Y）的概率密度为f（x，y），（X，Y）的边缘密度为f_X（x），f_Y（y），对于固定的x，任意a，b（a<b）以及ε>0，考虑条件概率：

$$P\{a\le Y\le b|x<X\le x+\varepsilon\},$$

设$P\{x<X\le x+\varepsilon\}>0$，即有

在一些条件下，当ε很小时，上式右端分子、分母分别近似于，εf_X（x）. 于是当ε很小时，有

此外，易见

与一维随机变量的概率密度的定义（式2-7，式2-8，式2-9）比较，我们给出以下的定义.

定义　　设（X，Y）是二维连续型随机变量，它的概率密度为f（x，y），（X，Y）关于X的边缘密度为f_X（x）. 若对于固定的x，f_X（x）>0，则称f（x，y）/f_X（x）为在X=x的条件下，Y的条件概率密度，记为

由式2-29可知，当ε很小时，有

上式说明了条件概率密度$f_{Y|X}$（y|x）的含义.

运用条件概率密度$f_{Y|X}$（y|x），我们可以定义在给定X=x的条件下，Y的条件概率：

特别，将$P\{Y\le y|X=x\}$记为$F_{Y|X}$（y|x），即

称为在X=x条件下，Y的条件分布函数 .

类似地，可以定义在Y=y的条件下，X的条件概率密度和条件分布函数：

【例2-21】 设G是平面上的有界区域，其面积为A，若二维随机变量（X，Y）具有概率密度

则称（X，Y）在G上服从均匀分布.

现自顶点为（0，0），（1，0），（0，1）的三角形G中随机地取一点，其坐标为（X，Y），试求条件概率密度$f_{X|Y}$（x|y）.

解 按题意（X，Y）在三角形G上服从均匀分布，于是（X，Y）具有概率密度

且有边缘密度

当0<y<1时，对于固定的y，可得在Y=y的条件下，X的条件概率密度

【例2-22】 在一周开始时在一储油罐中装入Y（吨）油，设Y在区间（0，1）上均匀分布. 设以X表示在一周内售出的油的重量. 已知当Y取值为y时X在区间（0，y）上均匀分布.

（1）求X和Y的联合概率密度；

（2）求X的概率密度；

（3）已知在一周开始时装入吨油的条件下，求这一周中售出多于吨油的条件概率.

解　（1）按题意，已知Y的概率密度

对于任意给定的值y（0<y<1），在Y=y条件下X的条件概率密度

由式2-33可得X和Y的联合概率密度

（2）X的概率密度

2.7　随机变量的独立性

随机变量的独立性是一个非常重要的概念.　在绝大多数情形下，概率论与数理统计是以独立随机变量为研究的主要对象.　我们从随机事件独立性的定义出发引出随机变量独立性的定义.

设A，B是两个事件，我们知道若P（AB）=P（A）P（B），则称A，B相互独立. 对于两个随机变量X，Y有以下的定义：

定义　　设X，Y是两个随机变量，若对于任意实数a，b（a<b），c，d（c<d），事件{a<X≤b}和{c<Y≤d}相互独立，即

$$P\{a<X\le b，c<Y\le d\}=P\{a<X\le b\}P\{c<Y\le d\}，$$

（2-35）

则称随机变量X，Y相互独立 .

可以证明：

离散型随机变量X，Y相互独立的充分必要条件是对于（X，Y）所有可能取的数偶（x_i，y_j），有

$$P\{X=x_i，Y=y_j\}=P\{X=x_i\}P\{Y=y_j\}；$$

（2-36）

连续型随机变量X，Y相互独立的充分必要条件是

$$f（x，y）=f_X（x）f_Y（y），$$

（2-37）

在平面上几乎处处 [3] 成立，其中f，f_X，f_Y 分别是（X，Y）的概率密度和边缘概率密度.

我们可以利用式2-36和式2-37来判断X，Y的独立性.

【例2-23】 问：例2-19中的X_1，Y_1 是否相互独立？又X_2，Y_2 是否相互独立？

解 对于X_1 和Y_1 来说有

$$P\{X_1=0，Y_1=0\}=4/9=P\{X_1=0\}P\{Y_1=0\}，$$

$$P\{X_1=0，Y_1=1\}=2/9=P\{X_1=0\}P\{Y_1=1\}，$$

$$P\{X_1=1，Y_1=0\}=2/9=P\{X_1=1\}P\{Y_1=0\}，$$

$$P\{X_1=1，Y_1=1\}=1/9=P\{X_1=1\}P\{Y_1=1\}，$$

故X_1，Y_1相互独立.

对于X_2和Y_2来说，由于

$$P\{X_2=0，Y_2=0\}=1/3，$$

而

$$P\{X_2=0\}=2/3，P\{Y_2=0\}=2/3，$$

$$P\{X_2=0，Y_2=0\}\neq P\{X_2=0\}P\{Y_2=0\}，$$

故X_2，Y_2不是相互独立的.

【例2-24】 设（X，Y）的概率密度

问：X，Y是否相互独立？

解 （X，Y）的边缘概率密度

对于所有x，y有f（x，y）=f_X（x）f_Y（y），故X，Y相互独立.

【例2-25】 例2-18中的随机变量X，Y不是相互独立的，这是因为f（x，y）与f_X（x）f_Y（y）不是几乎处处相等.

由定义可知X，Y相互独立就是对于任意两个事件{a<X≤b}，{c<Y≤d}相互独立. 从直观上看就是X和Y的取值互不影响. 在实际问题

中，我们常常根据问题的实际背景来判断两个随机变量的独立性.　　例如，若分别记甲、乙两个工厂生产的灯泡的寿命为X，Y，则可认为X，Y相互独立.

2.4节、2.5节及本节中与二维随机变量（X，Y）有关的各个定义都能推广到n（n>2）维随机变量的情况. 例如，n元函数

$$F（x_1，x_2，…，x_n）=P\{X_1 \le x_1，X_2 \le x_2，…，X_n \le x_n\}（-\infty<x_1，x_2，…，x_n<\infty）$$

称为n维随机变量（X_1，X_2，…，X_n）的分布函数 .

设f（x_1，x_2，…，x_n）是连续型随机变量（X_1，X_2，…，X_n）的概率密度，（X_1，X_2，…，X_n）关于X_1 的边缘概率密度

关于X_1，X_2 的边缘概率密度

设X_1，X_2，…，X_n 是连续型随机变量，X_1，X_2，…，X_n 相互独立的充分必要条件为

$$f（x_1，x_2，…，x_n）=f_{X_1}（x_1）f_{X_2}（x_2）…f_{X_n}（x_n），$$

在n维空间几乎处处成立，其中f，f_{X_i}（i=1，2，…，n）分别是（X_1，X_2，…，X_n）的概率密度和边缘概率密度.

设**X** =（X_1，X_2，…，X_m），**Y** =（Y_1，Y_2，…，Y_n）是两个多维随机变量，若对于任意G_1，G_2（G_1，G_2 分别为m维、n维空间的

立方体），事件{**X** $\in G_1$ }和事件{**Y** $\in G_2$ }相互独立，则称**X** ，**Y** 相互独立.[4]

还有以下的结论：

设（X_1 ，X_2 ，…，X_m ），（Y_1 ，Y_2 ，…，Y_n ）是相互独立的多维随机变量，则h（X_1 ，X_2 ，…，X_m ）和g（Y_1 ，Y_2 ，…，Y_n ）相互独立，其中h，g是连续函数.

2.8　随机变量函数的分布

在理论和实际应用中，经常有这种情况，一随机变量X的概率密度是已知的，而我们却需要知道X的某个函数的概率密度.　例如已知产品产量X的概率密度，而我们需要知道的是产值U的概率密度，这里U是随机变量X的函数. 本节先举例说明如何由已知的随机变量X的分布去确定它的函数Y=g（X）（g是已知的连续函数）的分布.

【例**2-26**】　设随机变量X具有分布律如下：

求$Y=X^2+1$的分布律.

解　Y所有可能取的值为1，2，10，由

$$P\{Y=1\}=P（X^2+1=1）=P\{X=0\}=3/10，$$

$$P\{Y=2\}=P\{X^2+1=2\}=P\{（X=1）\cup（X=-1）\}$$

$$=P\{X=1\}+P\{X=-1\}=3/10，$$

$$P\{Y=10\}=P\{X^2+1=10\}=P\{X=3\}=4/10,$$

即得Y的分布律为

Y	1	2	10
p_k	3/10	3/10	4/10

从这里可以看到，只要将Y所有可能取的值以及取这些值的概率找出来，就能写出Y的分布律.

【例2-27】 一食品工厂，一天的产量X（以吨计）具有概率密度

一天的产值是U=3X+1（以千元计），求U的概率密度f_U（u）.

解 分别记X，U的分布函数为F_X（x）和F_U（u）. 现在先来求F_U（u）.

$$F_U（u）=P\{U\le u\}=P\{3X+1\le u\}$$

$$=P\{X\le（u-1）/3\}=F_X（（u-1）/3），$$

将F_U（u）关于u求导数，由式2-18得U的概率密度

【例2-28】 设随机变量X具有概率密度f_X（x）（$-\infty<x<\infty$），求$Y=X^2$的概率密度f_Y（y）.

解 分别记X和Y的分布函数为F_X（x）和F_Y（y）. 先来求F_Y

（y）．由于$Y=X^2 \geq 0$，因此当y<0时，F_Y（y）=P{Y≤y}=0，而当y>0时，有

将F_Y（y）关于y求导，得Y的概率密度

【例2-29】 设某种电子元件的寿命X（以h计）服从瑞利分布，概率密度

由式2-38得$Y=X^2$ 的概率密度

Y是一个服从指数分布的随机变量.

例2-27和例2-28所说的是由已知随机变量X的概率密度去求Y=g（X）的概率密度. 其做法是先求出Y的分布函数F_Y（y）=P{Y≤y}，然后将F_Y（y）关于y求导，一般来说能得到Y的概率密度.

下面讨论两个独立随机变量之和的分布.

设X，Y是两个相互独立的随机变量，已知X，Y的概率密度分别为f_X（x），f_Y（y），现在来求U=X+Y的概率密度f_U（u）．先来求U的分布函数F_U（u）． 由X，Y的独立性知道（X，Y）的概率密度为f_X（x）f_Y（y），于是

这里的积分域是直线x+y=u左下方的半平面（如图2-19所示）．将这个二重积分化成累次积分：

图2-19

其中F_X（·）是X的分布函数. 对上式两边关于u求导（这里假设积分号内求导是允许的），就得到U的概率密度

由X，Y的对称性，f_U（u）又可写成

这就是两个相互独立随机变量X，Y之和U=X+Y的概率密度.

这两个公式称为卷积公式，常记为$f_X * f_Y$，即

【例2-30】 设一系统由两个独立工作的元件L_1，L_2组成，当L_1损坏时L_2立即开始工作，又设L_1，L_2的寿命X_1，X_2的概率密度均为f（t），

求系统寿命$U=X_1+X_2$的概率密度.

解 由式2-39可知U的概率密度

由f（t）的定义易知当且仅当

时上述积分的被积函数不等于零，于是（参见图2-20）得U的概率密度

图2-20

在本节的最后我们讨论M=max（X，Y）及N=min（X，Y）的分布.

设X，Y是两个相互独立的随机变量，它们的分布函数分别为F_X（x）和F_Y（y）. 现在来求M=max（X，Y）及N=min（X，Y）的分布函数.

对于任意实数z，由于

$$\{M\le z\}=\{\max（X，Y）\le z\}=\{（X\le z）\cap（Y\le z）\}，$$

由X和Y的独立性得到M=max（X，Y）的分布函数

$$F_{max}（z）=P\{M\le z\}=P\{（X\le z）\cap（Y\le z）\}$$

$$=P\{X\le z\}P\{Y\le z\}，$$

即有

$$F_{max}（z）=F_X（z）F_Y（z）.$$

（2-41）

类似地，可得N=min（X，Y）的分布函数

$$F_{min}（z）=P\{N\le z\}=P\{\min（X，Y）\le z\}$$

$$=1-P\{\min（X，Y）>z\}$$

$$=1-P\{X>z，Y>z\}=1-P\{X>z\}P\{Y>z\}$$

$$=1-［1-P\{X\le z\}］［1-P\{Y\le z\}］，$$

即有

$$F_{min}（z）=1-［1-F_X（z）］［1-F_Y（z）］.$$

（2-42）

以上结果可推广到n个相互独立的随机变量的情况. 设X_1 ，X_2 ，…，X_n 是n个相互独立的随机变量，它们的分布函数分别为F_{X_1}（x_1），F_{X_2}（x_2），…，F_{X_n}（x_n），则M=max（X_1，X_2，…，X_n）的分布函数

$$F_{max}（z）=F_{X_1}（z）F_{X_2}（z）…F_{X_n}（z），$$

（2-43）

N=min（X_1，X_2，…，X_n）的分布函数

$$F_{min}（z）=1-［1-F_{X_1}（z）］［1-F_{X_2}（z）］…［1-F_{X_n}（z）］.$$

（2-44）

特别，当X_1，X_2，…，X_n 相互独立且具有相同的分布函数F（x）时，有

$$F_{max}（z）=［F（z）］^n，$$

（2-45）

$$F_{min}（z）=1-［1-F（z）］^n.$$

（2-46）

【例2-31】 设系统L由两个独立工作的电子元件L_1 ，L_2 连接而

成.　　连接的方式分别为（1）串联，（2）并联（如图2-21和图2-22所示）. 设L_1，L_2 的寿命分别为X，Y，已知它们的概率密度分别为

其中$\lambda_1 >0$，$\lambda_2 >0$. 试分别就以上两种连接方式求系统L的寿命Z的概率密度.

图2-21

图2-22

解　（1）串联的情况

由于当L_1，L_2 中有一个损坏时，系统L就停止工作，因而系统L的寿命

$$Z=\min（X，Y）.$$

由式2-47和式2-48可知X，Y的分布函数分别为

于是得Z=min（X，Y）的分布函数

Z=min（X，Y）的概率密度

（2）并联的情况

由于当L_1，L_2 都损坏时系统L才停止工作，所以L的寿命

$$Z=\max（X，Y），$$

于是得Z=max（X，Y）的分布函数

Z=max（X，Y）的概率密度

习题2

【答案链接】

1. 投掷一枚硬币直至正面出现为止，以Y表示投掷的次数. 求Y的分布律及Y是偶数的概率.

2. 已知4个元件中有2个是次品，检验员每次检验1个元件，当2个次品都找到时即停止检验，以X表示检验的次数. 求X的分布律.

3. 设X的分布律为$P\{X=k\}=p^k$ （1-p），其中k=0，1，2，…. 求：Y=max（X，3）的分布律.

4. 据以往的资料知，某大学某一专业的毕业生中有10%成为硕士研究生. 今年有20个学生毕业，求：（1）恰有3人；（2）至少有1人；（3）不少于1人且不多于3人成为硕士研究生的概率；（4）已知至少有1人成为硕士研究生，求有3人成为硕士研究生的条件概率（设各毕业生是否成为硕士研究生是相互独立的）.

5. 一部件包含5个元件，各元件能正常工作的概率为0.9，各元件独立工作. （1）若仅当5个元件均正常工作时，部件正常工作；（2）若仅当至少有4个元件正常工作时，部件正常工作. 求这两种情况下部件能正常工作的概率.

6. 一机场的接客轿车有4个座位，旅客需预订座位. 经验表明有20%的旅客预订了座位而不来乘车，设各旅客是否来乘车相互独立.（1）若

已有6名旅客来订座，求届时各人都有座位的概率（即求至多有4名旅客来乘车的概率）；（2）若已有6名旅客订座，求届时恰有1名旅客没有座位的概率.

7. 设某地区一年内刮龙卷风的次数X~π（5）. 求P{X≤4}，P{X>3}，P{2<X≤5}.

8. 学生完成一次某种课程的测验需要的时间X（以h计）是随机变量，它具有概率密度

（1）确定常数c；（2）求半个小时以内完成的概率；（3）求完成时间在30~40min的概率.

9. 某种易腐烂的食品的货架寿命X（以h计）的概率密度

求P{X≥200}，P{X≤100}，P{X=300}.

10. 某种元件的寿命（以h计）服从以100为参数的指数分布. 一系统由3个独立工作的元件组成，若其中至少有2个元件失效则系统失效. 求系统的寿命大于200h的概率.

11.（1）设随机变量X的概率密度

求分布函数F（x），并作出F（x）的图形.

（2）设随机变量X的分布律为

X	0	1	2
p_k	1/3	1/2	1/6

求分布函数F（x），作出F（x）的图形，并利用F（x）求P{0≤X<3}.

12. 某一地点的风速（以km/h计）X是一随机变量，具有分布函数

（1）求P{X≤30}，P{10<X≤80}，P{70≤X≤120}，P{X≥110}；

（2）求概率密度函数.

13. 一条长为200m悬挂着的绳子受力拉断，其断点均匀分布在整条绳子上，以X记绳子的一个端点至断点的距离. 试写出X的概率密度，并求P{X≤100}，P{X>50}，P{40≤X≤80}.

14. 设随机变量X在区间（0，1）上服从均匀分布，引入随机变量

证明：Y~B（1，p）.

在一般的计算机中，都储有产生在区间（0，1）上均匀分布的随机变量X的观察值（称为伪随机数）的程序. 由本题的结论我们能利用这种程序来产生服从B（1，p）分布的随机变量Y的观察值. 其做法：由上述程序产生随机变量X的观察值，若它小于p就得到Y的观察值为1；否则得到Y的观察值为0. 在区间（0，1）上均匀分布的随机变量在随机模拟（蒙特-卡洛（Monte-Carlo）方法）中有广泛的应用.

15. 设随机变量X在1，2，3三个数中等可能地取一个值，随机变量Y在1~X中等可能地取一个整数值，求X和Y的联合分布律.

16. 将A，B两枚硬币各投掷一次，以X表示A币得到的正面数，以Y表示A，B两枚硬币得到的正面总数. 求X和Y的联合分布律.

17.（1）设随机变量（X，Y）的概率密度

试确定常数c，并求P{Y>2X}.

（2）设随机变量（X，Y）的概率密度

试求$P\{Y>1/X^2\}$.

18. 设随机变量（X，Y）的概率密度

求边缘概率密度f_X（x），f_Y（y）.

（3）在15题中求边缘分布律.

19. 在集合{1，2，3}中取数两次，每次任取一个数，作不放回抽样；以X表示第一次取到的数，以Y表示第二次取到的数.

（1）求X，Y的联合分布律；

（2）求在Y=3的条件下X的条件分布律；

（3）求在X=1的条件下Y的条件分布律.

20. 设随机变量（X，Y）的概率密度

（1）对于每一个x∈（0，1），求条件概率密度$f_{Y|X}$（y|x），并写出；

22. 设随机变量（X，Y）的概率密度

求条件概率密度$f_{Y|X}$（y|x）和$f_{X|Y}$（x|y），并求条件概率$P\{0<Y<2|X=1\}$.

23. 设随机变量（X，Y）的概率密度

问：X，Y是否相互独立.

（3）问：第15题中的随机变量X和Y是否相互独立？

24. 设随机变量X在区间（0，2）均匀分布，随机变量Y的概率密度

且X，Y相互独立.

（1）求X和Y的联合概率密度；（2）求.

25. 设有两个独立工作的元件，它们的寿命T_1，T_2服从指数分布，其概率密度分别为

求：概率$P\{T_1 > T_2\}$，$P\{T_1 > 2T_2\}$.

26. 设二维随机变量（X，Y）的概率密度

称（X，Y）服从参数为μ_1，μ_2，σ_1，σ_2，ρ的二维正态分布，记为

（1）证明（X，Y）的边缘概率密度为

（2）证明当$\rho=0$时X，Y相互独立.

提示：

即能求得f_X（x）.

27. 设随机变量X的分布律为

求$Y=2X^2-1$的分布律.

28. 设随机变量X在区间（-1，1）服从均匀分布，求Y=（X+1）/2的概率密度.

29. 设随机变量X的概率密度

求Y=1/X的概率密度.

30. 设随机变量X~N（0，1），即知X的概率密度，求U=|X|的概率密度.

31. 设随机变量X，Y相互独立，且都服从以1为参数的指数分布，求Z=X+Y的概率密度.

32. 设随机变量X在区间（-1，1）均匀分布，随机变量Y具有概率密度（$-\infty<y<\infty$），且X，Y相互独立. 求Z=X+Y的概率密度.

33. 设随机变量X在区间（0，1）服从均匀分布，随机变量Y具有概率密度

且X，Y相互独立.

（1）求M=max（X，Y）的分布函数和概率密度；

（2）求概率.

34. 一种物品在两个同类的商店出售，价格分别为X，Y，它们都具有概率密度

某人选择一价格低的商店购置了该物品，求他购入物品的价格的概率密

度（设X，Y相互独立）.

35. 一工厂的工人完成某项任务所需的时间（以h计）X是一个随机变量，它具有概率密度

其中θ>0，是一常数（θ表示完成任务的最小时间）. 自这一工厂中随机地选出n个工人，分别以X_1，X_2，…，X_n 记这n个工人完成任务所需的时间，设X_1，X_2，…，X_n 相互独立. 求Z=min（X_1，X_2，…，X_n）的概率密度.

36. 设随机变量（X，Y）的分布律为

（1）求V=max（X，Y）的分布律；

（2）求U=min（X，Y）的分布律；

（3）求W=X+Y的分布律；

（4）求P{X=2|Y=2}，P{Y=3|X=1}.

---

[1] 对于不放回抽样，以A_1，A_2 分别记第一次、第二次取到白球，则有$P(A_2|A_1)=$，而$P(A_2)=a/(a+b)$，$P(A_2|A_1)\neq P(A_2)$，故第一次、第二次试验不相互独立.

[2] 这里是二项式$(p+q)^n$（q=1−p）展开式的第k+1项，故得名.

[3] 在平面上几乎处处成立是指允许在平面上一面积为零的集合上不成立.

[4] 两个多维随机变量相互独立的概念以及下面的结论在数理统计中有用.

3　随机变量的数字特征

分布函数、概率密度函数和分布律都能完整地描述随机变量，但人们还常感兴趣于某些能描述随机变量某一个方面特征的常数. 例如，一个排球队上场比赛运动员的身高是一个随机变量，人们常关心上场运动员的平均身高；一个城市一户家庭拥有的自行车数是一个随机变量，在考察城市的交通情况时，人们关心户均拥有自行车的辆数；一个乡的人年收入是一个随机变量，考察这个乡农民的生活水平时，人们不但关心人均年收入，还关心个人年收入与人均年收入的偏离程度. 这种由随机变量分布所确定的，能刻画随机变量某一方面的特征的常数称为数字特征，它们在理论和应用上都很重要. 本章着重介绍随机变量的两个最重要的数字特征：数学期望和方差.

3.1　数学期望

我们由一个例题引入数学期望的概念.

【例3-1】　一书店购入一批（共N本）次年的挂历. 在当年11月底前售出可盈利10元/本，当年12月份以折扣价售出盈利6元/本，次年1月份以进货价售出盈利0元/本，次年2月份作为废纸售出亏本9.7元/本. 售出一本挂历盈利X（元）是一个随机变量，据往年经验知X的分布律如下：

问：预期平均一本挂历能盈利多少？

解　　如果书店分别在当年11月底前、12月份、次年1月份、次年2月份售出n_1，n_2，n_3，n_4本，$n_1+n_2+n_3+n_4=N$，那么平均一本盈利为

然而这个数事先并不知道，要等到次年2月份结算时才能知道. 注意到这里n_k/N是事件$\{X=x_k\}$发生的频率. 在3.4节中将会讲到当N充分大时n_k/N在某种意义下接近于事件$\{X=x_k\}$的概率p_k，于是平均一本挂历盈利为

这就是说，书店购入一大批挂历，可以预期平均一本挂历盈利7.415元左右.

式3-1表明随机变量X的观察值的算术平均，当N充分大时接近于数. 我们称数为随机变量X的数学期望，记为E（X）［在本例中E（X）=7.415］. 这就是说，X的数学期望是在大量次数试验下，X在各次试验中的观察值的算术平均的近似值. 数学期望刻画了X的平均大小.

定义　　设离散型随机变量X具有分布律$P\{X=x_k\}=p_k$（k=1，2，…），且级数绝对收敛；设连续型随机变量X具有概率密度f（x），且积分绝对收敛. 随机变量X的数学期望记为E（X），定义为

数学期望简称为期望，或称为均值.

【例**3-2**】　　以X记某城市一户家庭拥有自行车的辆数，由调查得知X具有分布律如下：

X的数学期望

E（X）=0×0.08+1×0.15+2×0.45+3×0.27+4×0.05

=2.06（辆），

这意味着考虑大量家庭时，例如1000户，那么平均一户拥有自行车约2.06辆，1000户家庭总共拥有自行车约2060辆.

【例3-3】 某种电子元件的寿命X（以年计）具有概率密度

求元件寿命的数学期望.

这意味着，测量大量这种电子元件，可求得它们的平均寿命约为7/3年.

【例3-4】 设X服从泊松分布，其分布律为

泊松分布的参数λ就是X的数学期望，因而只要知道泊松分布变量的数学期望，就能完全确定它的分布了.

例如，若一本书中一页的印刷错误的个数X~π（0.1），即有E（X）=0.1，这表示平均10页约有一个印刷错误，如这本书共400页，则约有40个印刷错误.

有许多实际问题，我们需要求随机变量函数的数学期望. 例如，某商店一天的营业额X是一个随机变量，而一天的利润Y是营业额X的函数，Y也是随机变量. 如已知X的概率密度，而我们需要求的却是Y的数学期望E（Y）. 一种方法是先求出Y的概率密度，然后按定义（式3-2）求出Y的数学期望E（Y），这样做往往会很麻烦. 下面的定理给出了简便的方法.

定理 设Y是随机变量X的函数：Y=g（X）（g是连续函数）. 其中离散型随机变量X具有分布律$P\{X=x_k\}=p_k$ （k=1，2，…），且级数绝对收敛，连续型随机变量X具有概率密度函数f（x），且积分绝对收敛，则有

证明略. 这一定理的意义在于为了计算X的某一函数g（X）的数学期望E［g（X）］，并不需要求出g（X）的概率密度或分布律，只需利用X的概率密度或分布律就可以了，这样做的好处是明显的.

【例3-5】 某商店一天的营业额X（以万元计）是一个随机变量，具有概率密度

一天的盈利Y是X的函数：

求E（Y）.

解 由式3-3可有

【例3-6】 某医院一天消耗某种药剂的数量X（以100L计）是一个随机变量，其概率密度

当0<X≤1时，每100L药剂价格为800元；当1<X≤2时，每100L药剂价格为500元. 求医院一天花在这种药剂的费用的数学期望.

解 以g（X）表示该医院一天花在这种药剂上的费用，则

按题意需要求E［g（X）］. 由式3-3可有

上述定理还能推广到两个随机变量的函数的情况. 设W是随机变量X，Y的函数：

$$W=g（X，Y）$$

（g是连续函数），其中二维随机变量（X，Y）是离散型随机变量，其分布律为$P\{X=x_i，Y=y_j\}=p_{ij}$（i，j=1，2，…），则有

又若（X，Y）是连续型随机变量，其概率密度为f（x，y），则有

这里设式3-4和式3-5右边的级数或积分绝对收敛.

【例**3-7**】　设随机变量（X，Y）具有概率密度

求数学期望E（XY）.

图3−1

解　由式3-5可得

【例**3-8**】　一次考试由两次测验组成. 以X，Y分别表示某班级一名学生第一次、第二次测验的得分，（X，Y）的分布律如下：

按校方规定，学生这次考试的得分W=max（X，Y）.　试求E（W）=E{max（X，Y）}.

这就是说这个班这次考试的平均成绩约为9.6分.

数学期望具有以下的性质（设这里遇到的随机变量其数学期望都存在）.

（1）设c是常数，则有E（c）=c.

（2）设c是常数，X是随机变量，则有

$$E（cX）=cE（X）.$$

（3）设X，Y是两个任意的随机变量，则有

$$E（X+Y）=E（X）+E（Y）.$$

这一性质可以推广到有限个随机变量之和的情况.

（4）设X，Y是两个相互独立的随机变量，则有

$$E（XY）=E（X）E（Y）.$$

这一性质可以推广到有限个相互独立的随机变量之积的情况.

证　（1）c是这样的随机变量，它只可能取值c，因而它取c的概率为1，于是E（c）=c·1=c.

以下只就X为连续型随机变量的情况来证明.　对于离散型的情况，其证明类似. 设X的概率密度为f（x），由式3-3可有

（3）设（X，Y）的概率密度为f（x，y），由式3-5可有

（4）又若X，Y相互独立，此时f（x，y）=f_X（x）f_Y（y），f_X（x），f_Y（y）是（X，Y）的边缘密度. 由式3-5和式3-2可有

3.2　方差

现在介绍另一个重要的数字特征——方差.　数学期望即均值给出了随机变量的平均大小，然而我们还常常关心随机变量的取值在均值周围的散布程度.　如前面曾提到的，在考察一个地区农民的贫富情况时，我们不但关心农民的人均年收入，还关心各个农民的个人年收入与人均年收入的偏离程度.　例如有甲、乙两个乡的人均年收入都是6000元，而两个乡农民的个人年收入的总的情况却不一样，甲乡各人的年收入大多集中在6000元附近，而乙乡农民的个人年收入与6000元的偏离程度较大，

即贫富差别较大. 如果分别作出甲、乙两个乡个人年收入X，Y的概率密度曲线[1]，可以看出两者的差别（图3-2中曲线①，②分别是X，Y的概率密度）． 方差就是用来刻画随机变量的取值和均值的偏离程度（即散布的程度）的数字特征. 下面引入方差的定义.

图3-2

定义　　设X是随机变量，若E{［X-E（X）］^{2} }存在，称它为X的方差，记为D（X）或var（X），即

$$D（X）=var（X）=E\{［X-E（X）］^2\}.$$

（3-6）

方差的算术平方根称为X的均方差 或标准差 . 与X具有相同的量纲.

若X是离散型随机变量，其分布律为P{X=x_k　　}=p_k　　（k=1，2，…），或X为连续型随机变量，其概率密度为f（x），按方差的定义则有

将式3-6展开，由数学期望的性质得到

$$D（X）=E\{［X-E（X）］^2\}$$

$$=E\{X^2-2XE（X）+［E（X）］^2\}$$

$$=E（X^2）-2E（X）E（X）+［E（X）］^2,$$

即

$$D（X）=E（X^2）-［E（X）］^2.$$

（3-8）

我们常利用式3-8来求D（X）.

【例3-9】 一商店出售的干电池有5种包装规格，分别为1个，2个，4个，6个，12个. 以X表示一个顾客购买干电池的个数，据以往经验，X具有分布律如下：

求E（X），D（X），.

解 E（X）=5.64（个），

$D（X）=E（X^2）-［E（X）］^2$

$=（1^2 \times 0.08+2^2 \times 0.27+4^2 \times 0.10+6^2 \times 0.33+12^2 \times 0.22）-5.64^2$

$=14.51（个^2）$，

这就是说一个顾客平均购买5.64个干电池，而3.81表示X的取值与均值的偏离程度.

【例3-10】 设随机变量X~π（λ），求D（X）.

解 X的分布律为

在例3-4中已求得E（X）=λ，现在来求$E（X^2）$.

于是

$$D（X）=E（X^2）-［E（X）］^2=\lambda^2+\lambda-\lambda^2=\lambda.$$

由此知泊松分布随机变量的数学期望与方差相等，都等于λ.

【例3-11】 设随机变量X在（a，b）上具有均匀分布，求E（X），D（X）.

解　X的密度函数

【例3-12】 设随机变量X服从参数为β的指数分布，求E（X），D（X）.

解　X的密度函数

方差具有以下的性质（设遇到的随机变量其方差都存在）：

（1）设c是常数，则有

$$D（c）=0.$$

（2）设X是随机变量，c是常数，则有

$$D（cX）=c^2 D（X），$$

$$D（X+c）=D（X）.$$

（3）设X，Y是两个任意的随机变量，则有

$$D（X+Y）=D（X）+D（Y）+2［E（XY）-E（X）E（Y）］.$$

（3-9）

特别，若X，Y相互独立，则有

$$D（X+Y）=D（X）+D（Y）.$$

（3-10）

这一性质可推广到有限多个相互独立的随机变量之和的情况.

（4）D（X）=0的充要条件是X以概率1取常数c，即

$$P\{X=c\}=1,$$

显然这里c=E（X）.

证　（1）$D（c）=E（c^2）-［E（c）］^2=c^2-c^2=0.$

（2）$D（cX）=E（c^2X^2）-［E（cX）］^2$

$=c^2［E（X^2）-（E（X））^2］$

$=c^2D（X）.$

$D（X+c）=E\{（X+c）^2\}-\{E（X+c）\}^2$

$=E（X^2）+2cE（X）+c^2-\{［E（X）］^2+2cE（X）+c^2\}$

$=E（X^2）-［E（X）］^2=D（X）.$

（3）$D（X+Y）=E\{（X+Y）^2\}-［E（X+Y）］^2$

$=E\{X^2+2XY+Y^2\}-［E（X）+E（Y）］^2$

$=[E(X^2)-(E(X))^2]+[E(Y^2)-(E(Y))^2]+2\{E(XY)-E(X)E(Y)\}$

$=D(X)+D(Y)+2[E(XY)-E(X)E(Y)]$.

特别，若X，Y相互独立，则由数学期望的性质可知$E(XY)=E(X)E(Y)$，此时式3-9成为

$$D(X+Y)=D(X)+D(Y).$$

证毕.

（4）证略.

【例3-13】 设$X\sim B(n,p)$，求$E(X)$，$D(X)$.

解 由二项分布的定义可知随机变量X是n重伯努利试验中事件A发生的次数，且在每次试验中A发生的概率为p. 引入随机变量

易知

$$X=X_1+X_2+\ldots+X_n.$$

（3-11）

由于X_k 只依赖于第k次试验，而各次试验相互独立，于是X_1 ，X_2 ，…，X_n 相互独立，又知X_k 服从（0-1）分布，其分布律为

式3-11表明以n，p为参数的二项分布变量，可分解为n个相互独立且都服从以p为参数的（0-1）分布的随机变量之和.

由式3-12可得　$E(X_k)=p$，

故由式3-11有

$$E(X)=E(X_1)+E(X_2)+\cdots+E(X_n)=np.$$

又由X_1，X_2，…，X_n的独立性可得

$$D(X)=D(X_1)+D(X_2)+\cdots+D(X_n)=np(1-p)，$$

即

$$E(X)=np，D(X)=np(1-p).$$

【例3-14】　设随机变量X具有数学期望$E(X)=\mu$，方差$D(X)=\sigma^2\neq 0$. 记

即$Y=(X-\mu)/\sigma$具有数学期望为0，方差为1. Y称为X的标准化变量 .

3.3　协方差与相关系数

本节讨论表征二维随机变量（X，Y）中X与Y的相互关系的数字特征——协方差与相关系数.

定义　若$E\{[X-E(X)][Y-E(Y)]\}$存在，称它为随机变量X和Y的协方差，记为

$$cov(X，Y)=E\{[X-E(X)][Y-E(Y)]\}，$$

（3-14）

又称

为X和Y的相关系数 .

由定义看出cov（X，Y）=cov（Y，X），cov（X，X）=D（X），又由于

$$E\{[X-E(X)][Y-E(Y)]\}$$

$$=E\{XY-XE(Y)-YE(X)+E(X)\cdot E(Y)\}$$

$$=E(XY)-E(X)E(Y),$$

因而协方差又可写成

$$\mathrm{cov}(X,Y)=E(XY)-E(X)E(Y).$$

（3-16）

这样，式3-9就成为

$$D(X+Y)=D(X)+D(Y)+2\mathrm{cov}(X,Y).$$

（3-17）

【例3-15】 设随机变量（X，Y）具有概率密度

求：cov（X，Y），ρ_{XY}，D（X+Y）.

图3-3

$$D（X）=E（X^2）-［E（X）］^2=1/18，$$

$$D（Y）=D（X）=1/18，$$

于是

相关系数具有以下两条重要性质.

定理　设ρ_{XY}是随机变量X，Y的相关系数，则有

（1）$|\rho_{XY}|\leq 1$；

（2）$|\rho_{XY}|=1$的充要条件是X和Y以概率1存在线性关系，即

$$P\{Y=a+bX\}=1（a，b是常数，b\neq 0）.$$

于是$\rho_{XY}\geq -1$. 类似地有

从而$\rho_{XY}\leq 1$.

（2）必要性　设$\rho_{XY}=1$，则

由方差的性质4可知

上式可改写成

$$P\{Y=a+bX\}=1（a，b为常数，且b\neq 0），$$

这就是说，存在常数a，b，且b≠0，使P{Y=a+bX}=1. 对于$\rho_{XY}=-1$可得同样的结果.

充分性　可由相关系数的定义直接证明.

定义　若$\rho_{XY}=0$，则称X，Y不相关 .

由ρ_{XY}的定义及数学期望的性质可得以下的定理：

定理　若随机变量X，Y相互独立，则$\rho_{XY}=0$，即X，Y不相关.

然而，需要指出：两个不相关的随机变量，却不一定是相互独立的，现举例如下.

【例3-16】　设二维随机变量（X，Y）具有分布律如下：

即有　E（X）=-1/3，E（Y）=0，E（XY）=0，

于是　cov（X，Y）=E（XY）-E（X）E（Y）=0，

即$\rho_{XY}=0$，亦即X，Y不相关.　而P{X=-1，Y=-1}=1/6≠P{X=-1}P{Y=-1}=2/9，可知X，Y不相互独立.

由以上的讨论知道，若$\rho_{XY}\neq0$，则X，Y一定不相互独立，也就是X和Y存在着某种关系. 特别当$\rho_{XY}=1$或$\rho_{XY}=-1$时，X，Y几乎总存在着线性关系[2]. 若$\rho_{XY}=0$，则X，Y可能相互独立，也可能不相互独立，但肯定不会存在着线性关系.

3.4　随机变量的另几个数字特征

本节介绍随机变量的另几个数字特征.

3.4.1 矩和中心矩

定义　设X是随机变量，若

$$E(X^k)，k=1，2，\cdots$$

存在，称它为X的k阶原点矩 或k阶矩 .

若

$$E\{[X-E(X)]^k\}，k=2，3，\cdots$$

存在，称它为X的k阶中心矩 .

3.4.2 分位数

定义　设连续型随机变量X的分布函数为F（x），概率密度函数为f（x），

1°对于任意正数α（0<α<1），称满足条件

的数为此分布的α分位数 或下 α分位数 .

2°对于任意正数α（0<α<1），称满足条件

的数x_α 为此分布的上 α分位数.

特别地，当α=0.5时，

$x_{0.5}$ 称为此分布的中位数.

下α分位数将概率密度曲线下的面积分为两部分，左侧的面积恰为α（如图3-4a）. 上α分位数x_{α} 也将概率密度曲线下的面积分为两部分，右侧的面积恰为α（如图3-4b）.

图3-4

下α分位数与上α分位数有以下的关系

类似地，可定义离散型随机变量X的分位数.

定义　对于任意正数α（0<α<1），称满足条件

的数为此分布的α分位数 或下 α分位数.

分位数又称分位点.

【例**3-17**】　设随机变量X服从指数分布，其分布函数为

求中位数$x_{0.5}$ 及0.25上分位数$x_{0.25}$.

解　由F（$x_{0.5}$ ）=1-$e^{-x_{0.5}}$ =0.5，得$e^{-x_{0.5}}$ =0.5，故中位数

$$x_{0.5} =-\ln 0.5=0.69.$$

由F（$x_{0.25}$ ）=1-$e^{-x_{0.25}}$ =0.25，得$e^{-x_{0.25}}$ =0.75. 故

$$x_{0.25} =-\ln 0.75=0.29.$$

3.4.3　变异系数

定义　设X是随机变量，若

存在，称它为X的变异系数，记为$(CV)_X$，即

我们常用随机变量X的标准差作为衡量X取值分散程度的尺度，变异系数是单位均值上的标准差，用它作为衡量X取值分散程度的尺度更为合理.

3.5　大数定理

在第1章中曾讲过，人们在实践中认识到，随着试验次数的增加，事件发生的频率逐渐稳定于某个常数附近，频率的稳定性是概率定义的客观基础. 本节我们将对频率的稳定性作出理论的说明.

先介绍一个有用的不等式.

定理**1**［契比雪夫（**Chebyshev**）不等式］　设随机变量X具有数学期望$E(X)=\mu$，方差$D(X)=\sigma^2$，则对于任意$k>0$，有

$$P\{|X-\mu|\geq k\sigma\}\leq 1/k^2,$$

（3-18）

或

$$P\{|X-\mu|<k\sigma\}\geq 1-1/k^2 .$$

（3-18'）

式3-18和式3-18'称为契比雪夫不等式 .

证　我们只就X是连续型的情况来证明（离散型情况的证明与此类似）. 设X的概率密度为f（x），则有

以（kσ）^{2} 除上式两边，式3-18得证.

我们用σ作为度量X与μ的偏差：|X-μ|的尺度，在式3-18'中分别令k=3，4得到：对于任意随机变量都有

$$P\{|X-\mu|<3\sigma\}\geq 1-1/9=8/9,$$

$$P\{|X-\mu|<4\sigma\}\geq 1-1/16=15/16.$$

契比雪夫不等式给出了在随机变量X的分布未知，而只知道μ和σ的情况下，对事件{|X-μ|<kσ}概率的估计. 这个估计是比较保守的 [3] . 当然，如果随机变量的分布已知时，就可以得到这一概率的精确值，此时就没有必要求助于契比雪夫不等式来作估计了.

定理**2**（辛钦大数定理）　设X_1，X_2，…，X_n，…是相互独立 [4] 服从同一分布的随机变量序列，且具有数学期望E（X_k）=μ（k=1，2，…），作前n个变量的算术平均，则对任意ε>0，有

证　我们只在随机变量的方差D（X_k）=σ^2 （k=1，2，…）存在这一附加条件下证明上述结果. 因为

由独立性得

由契比雪夫不等式3-18知

在上式中令$n\to\infty$，即得

设Y_1，Y_2，…，Y_n，…是一个随机变量序列，a是一个常数. 若对于任意正数ε，有

则称序列Y_1，Y_2，…，Y_n，…依概率收敛于a，记为

这样，定理2又可叙述为：

定理**2'**（辛钦大数定理）　设X_1，X_2，…，X_n，…是相互独立，服从同一分布的随机变量序列，且具有数学期望$E(X_k)=\mu$（k=1，2，…），则序列依概率收敛于μ，即.

作为定理2的特殊情况，得到如下的定理：

定理**3**（伯努利大数定理）　设f_A是n重伯努利试验中事件A发生的次数，设P（A）=p，则对于任意ε>0，有

这是因为$f_A \sim B(n, p)$，由例3-13知f_A可写成$f_A=X_1+X_2+\dots+X_n$，其中X_1，X_2，…，X_n相互独立且都具有以p为参数的（0-1）分布，它们的数学期望为p，由式3-19即得式3-20'.

人们在长期实践中认识到“一个概率很小的事件在一次试验中，实际上几乎是不发生的”，这在概率论中称为小概率原理 . 式3-20表明，对于任意ε>0，只要重复试验的次数n充分大，事件$\{|f_A/n-p|\geq\varepsilon\}$是一个小概

率事件. 这表明事件{|f_A /n-p|≥ε}实际上几乎是不发生的. 亦即事件“频率f_A /n与概率p的偏差小于ε”即事件{|f_A /n-p|<ε}，实际上几乎是必定要发生的. 这就是在大量重复试验中频率f_A /n接近于概率p的真正含义，也就是我们所说的频率稳定性的真正含义.

而定理2表明，对于任意ε>0，只要n充分大，事件实际上几乎是必定要发生的. 这就是我们通常所说的具有稳定性的真正含义.

现将上面讲过的几种重要分布汇总如下（见表3-1），以备查用.

表3-1

① 正态分布随机变量的数学期望和方差将在下一章给出，现在将结果先写在这里.

习题3

【答案链接】

1. 据以往资料，某人进行射击，击中目标的概率为0.7，现在射击3次，以X记击中目标的次数（设各次射击击中与否相互独立）. 求X的分布律和E（X）.

2. 据以往资料，某人打一次电话的持续时间（以min计）X的概率密度

求E（X）.

3. 某工程队完成某种工程的天数X是一随机变量，具有分布律如下：

所得利润（以万元计）

$$Y=1000（12-X）.$$

（1）求E（X）；（2）求E（Y）.

4. 设X~B（4，p），求.

5. 设随机变量X具有概率密度

求Y=2X-1的数学期望E（Y）.

6. 由某种机器切割而成的圆盘的直径（以cm计）是一个随机变量，其概率密度

求圆盘面积的数学期望.

7. 两个生物种属在一个地区争夺某种有限的资源，以X记种属I所占有资源的比例，X在（0，1）上服从均匀分布. 两种种属中的优胜者（即占有较多资源者），其占有的资源的比例

$$h（X）=max（X，1-X）.$$

求E［h（X）］.

8. 已知随机变量（X，Y）的分布律为

求$E（XY^2）$.

9. 设随机变量（X，Y）具有概率密度

求E（X），E（Y），E（XY）.

10. 某人每天上班相继要乘两条线路的公共汽车，乘各辆车的候车时间（以min计）都服从（0，5）上的均匀分布，求他一天上班花在候车上的平均时间.

11. 一快餐店，以Y_1 记顾客到达餐厅直至离开服务窗口的时间（以min计），以Y_2 记一顾客排队等待的时间（以min计）. 设Y_1 ，Y_2 的概率密度分别为

求窗口服务时间$Y_1 - Y_2$ 的数学期望.

12. 设随机变量X在（-1/2，1/2）上服从均匀分布，Y=sinπX. 求：E（Y），D（Y）.

13. 设随机变量X服从瑞利分布. 其概率密度

求：E（X），D（X）.

14.（1）在第3题中求D（X），D（Y）；

（2）设随机变量X的概率密度

15. 证明在c=E（x）时，E［（X-c）^{2}］达到最小.

16. 证明协方差具有如下性质：

（1）cov（aX，bY）=abcov（X，Y），a，b是常数；

（2）cov（X_1+X_2，Y）=cov（X_1，Y）+cov（X_2，Y）.

17. 设随机变量（X_1，X_2）具有概率密度

求：cov（X_1，X_2），$\rho_{X_1X_2}$，D（X_1-X_2）.

18. 设随机变量（X，Y）具有分布律如下：

求：cov（X，Y），ρ_{XY}.

19. 一盒装4个编号分别为1，2，3，4的球，在盒中取球2次，每次取1个，作不放回抽样. 以X，Y分别记取到的2个球中的小号码和大号码.

（1）求X，Y的联合分布律；（2）求E（X），D（X），E（Y），D（Y），E（XY），cov（X，Y），ρ_{XY}.

20. 设随机变量（X，Y）的分布律为

验证X，Y不相关，但X，Y不相互独立.

---

[1] 只要经过调查，就可根据得到的数据分别作出X，Y的频率直方图（见5.2节），以此作为概率密度曲线的近似.

[2] 参见3.5节中的小概率原理.

[3] 例如，若X~U（0，8），则有E（X）=4，D（X）=16/3，在式3-18'中取，得P{|X-4|<4}≥1-1/3=2/3，但准确结果为P{|X-4|<4}=1.

[4] 是指对于任意n>1，X_1，X_2，…，X_n 相互独立.

4　正态分布

正态分布在2.2节中已提到过，在这一章里我们将进行较多的讨论.

在概率论与数理统计的研究和实际应用中，正态分布的重要性居各种分布的首位.　这是因为：实际中遇到的随机变量有许多是服从或近似服从正态分布的（例如，气象学中的温度、湿度、降雨量，有机体的长度、重量、智能测试的评分，实验中的测量误差，经济学中的众多量度等都服从或近似服从正态分布）；正态分布是许多重要分布的极限分布；许多非正态分布变量是正态分布变量的函数；正态分布的概率密度和分布函数具有各种优良性；等等.

4.1　正态分布

我们已经知道正态随机变量X的概率密度

而其分布函数

其中，$-\infty<\mu<\infty$，$\sigma>0$.　我们将服从以μ，σ为参数的正态分布的随机变量X记为$X\sim N(\mu,\sigma^2)$.

特别，当μ=0，σ=1时，得到特别重要的情况：X~N（0，1），称X服从标准正态分布　.　它的密度函数和分布函数通常分别记为φ（x）和Φ（x），即

它们的图形分别如图4-1和图4-2所示.

图4-1　　图4-2

由于φ（x）是偶函数，φ（x）的曲线关于纵轴对称，曲线以横轴为水平渐近线，又由曲线的对称性可知（见图4-1），对于a>0，有

$$P\{X\leq -a\}=P\{X>a\}=1-P\{X\leq a\},$$

即

$$\Phi(-a)=1-\Phi(a).$$

（4-5）

这一关系式给计算Φ（x）的值带来了方便.

在书末附有Φ（x）函数表可供查用，表中只列出当a≥0时的函数值Φ（a）. 当a<0时可利用式4-5及Φ（x）函数表得到Φ（a）的值. 例如对于a=-0.51，在函数表上查到Φ（0.51）=0.6950. 于是有

$$\Phi(-0.51)=1-\Phi(0.51)$$

$$=1-0.6950$$

$$=0.3050.$$

设X~N（0，1），对于给定的正数α（0<α<1），满足条件

的点z_α 就是标准正态分布的上 α分位点 . 这表示如图4-3所示的右侧阴影部分的面积等于α. 由关系式

$$\Phi(z_\alpha)=1-P\{X>z_\alpha\}$$

$$=1-\alpha,$$

（4-6）

图4-3

对于给定的α的值，从Φ（x）的函数表可查到z_α 的值. 下面列出了几个常用的z_α 的值.

另外由φ（x）图形的对称性知道$z_{1-\alpha}=-z_\alpha$.

现在来讨论一般的正态分布，即考察X~N（μ，σ^2 ）. 事实上，借助于一个线性变换就能将一般的正态分布化成标准正态分布. 我们有以下的引理.

引理　若X~N（μ，σ^2 ），则Z=（X-μ）/σ~N（0，1）.

证　Z=（X-μ）/σ的分布函数

令（t-μ）/σ=u，得到

由此知Z=（X-μ）/σ~N（0，1）.

由上述引理，若X~N（μ，σ^2 ），则它的分布函数F（x）可写成

这样我们就能利用Φ（x）的函数表来求F（x）的函数值了.　　例如若X~N（2.5，16），则（X-2.5）/4~N（0，1），X的分布函数可写成

【例**4-1**】　　在车床上加工金属圆杆，已知圆杆直径（以cm计）$X\sim N(12.4,\sigma^2)$，规定直径在12.0~12.8cm为合格品，要求产品合格的概率至少为0.95. 试确定σ至多为多少?

解　按题意需确定σ，使得满足条件

即

$$\Phi(0.4/\sigma)-[1-\Phi(0.4/\sigma)]\geq 0.95,$$

即

$$\Phi(0.4/\sigma)\geq 0.975=\Phi(1.96).$$

由Φ（x）的单调性知必须有

$$0.4/\sigma\geq 1.96,\ \sigma\leq 0.204,$$

这就是说σ至多为0.204cm.

设$X\sim N(\mu,\sigma^2)$，现在来求X的数学期望和方差. 先来求标准正态变量$Z=(X-\mu)/\sigma$的数学期望和方差. Z的概率密度

因为$X=\mu+\sigma Z$，即得

$$E(X)=E(\mu+\sigma Z)=\mu,$$

$$D(X)=D(\mu+\sigma Z)=D(\sigma Z)=\sigma^2 D(Z)=\sigma^2,$$

即

$$E（X）=\mu，D（X）=\sigma^2 .$$

这就是说正态分布N（μ，σ^2 ）的概率密度

中的两个参数μ和σ分别就是该分布的数学期望和标准差，因而正态分布完全由它的数学期望μ和标准差σ所确定. 图4-4a画出了f（x）的图形，它具有以下的性质. 曲线呈钟形，关于x=μ对称，这表明对于任意h>0，有

$$P\{\mu-h<X<\mu\}=P\{\mu<X<\mu+h\}，$$

图4−4

当x=μ时，取到最大值

x离μ越远，f（x）的值越小，因而对于同样长度的区间，当区间离μ越远时，X落在该区间上的概率越小.

若固定σ，改变μ的值，则图形沿着Ox轴平移而不改变形状（见图4-4a）.对于相同的μ不同的σ，图形的对称轴相同，而图形的形状不相同，σ大的图形较为平坦，σ小的图形较为陡峭，因而X落在X=μ附近的概率较大（见图4-4b）. μ称为位置参数，σ称为形状参数.

由Φ（x）的函数表还能得到（见图4-5）

图4−5

$P\{\mu-\sigma<X\leq\mu+\sigma\}=\Phi（1）-\Phi（-1）=68.26\%，$

$P\{\mu-2\sigma<X\leq\mu+2\sigma\}=95.44\%，$

$P\{\mu-3\sigma<X\leq\mu+3\sigma\}=99.74\%.$

由第三个数据知正态变量的值落在（μ-3σ，μ+3σ）上几乎是肯定的，这就是人们所说的“3σ”法则.

4.2　正态随机变量的线性组合

正态随机变量具有如下的重要性质：两个或多个相互独立的正态随机变量的线性组合仍是正态变量. 我们从最简单的情况讲起.

设X，Y相互独立，且X~N（0，1），Y~N（0，1），则

$$X+Y\sim N(0,1+1),$$

这是因为X，Y的概率密度均为f（x），

由式2-40得U=X+Y的概率密度

令t=x-u/2，得

故U=X+Y~N（0，2），即U~N（0，1+1）.

一般，设X，Y相互独立，且X~N，Y~N，则由式2-40经计算知U=X+Y~N.

更一般地有以下的定理：

定理　设X_1，X_2，…，X_n 相互独立，且X_i ~N（i=1，2，…，n），则对于任意不全为零的常数c_1，c_2，…，c_n，有

证明略. 在定理中特别取$c_1=c_2=\ldots=c_n=1/n$，就能得到下列结果.

推论　设X_1，X_2，…，X_n 相互独立且具有同一分布N（μ，σ^2），是X_1，X_2，…，X_n 的算术平均，则

定理和推论都是十分重要的结果.

【例4-2】　设垫圈内直径（以mm计）X~N（11，0.25），螺栓直径（以mm计）Y~N（10，1），X，Y相互独立. 随机取一个螺栓和一个垫圈，求螺栓能装入垫圈的概率.

解　按题意需要求概率P{X>Y}=P{X-Y>0}. 由上述定理知道X-Y服从正态分布，又

$$E（X-Y）=E（X）-E（Y）=1，$$

$$D（X-Y）=D（X）+（-1）^2 D（Y）=1.25，$$

于是

$$X-Y\sim N（1，1.25），$$

故螺栓能装入垫圈的概率约为0.81.

【例4-3】　设瓷砖的长度（以cm计）服从N（25，0.4）.（1）随机地取20块瓷砖，将它们排成一行，求总长度超过505的概率；（2）分别随机取40块，各自排成一行，求两行的长度之差小于10的概率（设各瓷砖的长度相互独立）.

解　（1）记20块瓷砖的长度分别为X_1，X_2，…，X_{20}，由上述定理可知服从正态分布，且有

（2）分别记第一行瓷砖的长度为X_1，X_2，…，X_{40}，第二行瓷砖的长度为Y_1，Y_2，…，Y_{40}，于是

所求概率

4.3　中心极限定理

上一节我们看到独立的正态随机变量的和仍是正态随机变量，下面将介绍概率论中的一个重要结果：在相当一般的条件下，充分多个独立的非正态随机变量的和近似地服从正态分布.　这一事实大大增加了正态分布的重要性.

设随机变量X_1，X_2，…，X_n，…相互独立，且具有相同的分布，具有数学期望$E(X_i)=\mu$，方差$D(X_i)=\sigma^2\neq 0$（i=1，2，…），那么

且有以下的定理：

定理**1**（独立同分布的中心极限定理）　设随机变量X_1，X_2，…，X_n，…相互独立，服从相同的分布，具有数学期望$E(X_i)=\mu$，方差$D(X_i)=\sigma^2\neq 0$（i=1，2，…），则随机变量之和的标准化变量

的分布函数对于任意x，满足

证明略. 这就是说，均值为μ，方差为σ^2 的独立同分布的随机变量X_1，X_2，…，X_n，当n充分大时，有

在一般情况下，我们很难求出n个随机变量之和的分布的确切形式. 式4-9告诉我们，当n充分大时，可以通过Φ（x）给出其近似，这样就可以利用正态分布对作理论研究或作实际计算了，其好处是不言而喻的. 由于

于是式4-9可以写成：当n充分大时，有

这就是说，均值为μ，方差为σ^2 的独立同分布的随机变量X_1 ，X_2 ，…，X_n 的算术平均，当n充分大时近似地服从均值为μ，方差为σ^2 /n的正态分布. 这一结果是数理统计中大样本理论的基础.

将上述定理1应用于具有（0-1）分布的随机变量，即设X_1 ，X_2 ，…，X_n ，…相互独立，且都服从参数为p的（0-1）分布：

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

此时E（X_i ）=p，D（X_i ）=p（1-p）. 又记

由式3-11知η_n ~B（n，p），此时式4-8可写成

于是得到下述定理：

定理**2**（德莫佛-拉普拉斯**De Moivre-Laplace**定理） 设随机变量η_n （n=1，2，…），服从参数为n，p（0<p<1）的二项分布，则对于任意x，满足

这个定理表明，当n充分大时，二项分布随机变量η_n 的标准化随机变量近似服从标准正态分布，即

我们可以利用式4-12来近似计算二项分布的概率.

【例4-4】 某种短波无线电接收机中有一个关键性的组件，它的寿命（以h计）服从均值为500的指数分布. 现在有一个组件在工作，另有19个备用，当一个组件损坏时备用件立即换上. （1）求20个组件至少能使用1年（8760h）的概率；（2）问至少需多少个组件才能保证接收机至少能工作1年的概率不小于0.9.

解 （1）记第i（i=1，2，…，20）个组件的寿命为X_i ，则X_1 ，X_2 ，…，X_{20} 相互独立，且都具有概率密度

于是由表3-1知E（X_i ）=500，D（X_i ）=500^2 （i=1，2，…，20）. 按题意需求. 由定理1可知，随机变量

近似服从N（0，1）分布，于是

（2）设需要n个组件，按题意需要确定n，使得

式4-13左端即为

于是式4-13成为

现在取n=24，即至少需24个组件才能保证达到题中的要求.

【例4-5】 一工厂生产的某种产品其次品率为0.005.产品按每100只包装成为一箱，一箱中如含有的次品数超过3只就不能通过验收. 今有10000箱产品，求多于25箱不能通过验收的概率（设各只产品是否为次品相互独立，各箱是否不能通过验收相互独立）.

解　　以X记一箱产品中所含次品数，则X~B（100，0.005），于是一箱产品不能通过验收的概率

以Y表示10000箱产品中不能通过验收的箱数，则Y~B（10000，0.0017），于是由定理2得10000箱中多于25箱不能通过验收的概率近似地为

4.4　χ^2 分布、t分布与F分布

本节介绍在数理统计中占有极其重要地位的三个分布：χ^2　分布，t分布与F分布，有时称它们为“统计学中的三大分布”，这三个分布都是正态随机变量的函数的分布.

4.4.1　χ^2 分布

定义　设随机变量Z_1，Z_2，…，Z_n 相互独立，且都服从标准正态分布N（0，1），则称随机变量

服从自由度为n（自由度常记为df，df=n）的χ^2　　分布，记为χ^2　　~χ^2（n）. 此处自由度是指式4-14右边包含的独立变量的个数.

可以证明χ^2（n）分布具有概率密度

图4-6画出了$f_{\chi^2(n)}$（x）的图形.

图4-6

容易得到χ^2（n）分布的数学期望和方差分别为[1]

$$E(\chi^2)=n,\ D(\chi^2)=2n. \tag{4-16}$$

χ^2（n）分布还具有以下的“可加性”：

设$\chi^2 \sim \chi^2$（n），对于给定的正数α（0<α<1），满足条件

的点就是χ^2（n）分布的上α分位点（见图4-7）.

图4-7

对于不同α，n的上α分位点的值已制备成表格可以查用（见附表3）[2]，例如有.

4.4.2 t分布

定义　设Z~N（0，1），Y~χ^2（n）且Z，Y相互独立，则称随机变量

服从自由度为n（df=n）的t分布，记为T~t（n）.

可以证明t（n）分布具有概率密度

还可证明

图4-8中画出了f_T（x）的图形. 曲线关于纵轴对称，从图上可以看到t分布的概率密度曲线的尾部要比标准正态分布的要长一些.

图4-8

设T~t（n），对于给定的正数α（0<α<1），满足条件

的点t_α（n）就是t（n）分布的上α分位点（见图4-9）. t_α（n）的值可以查附表2；在n>45时，可利用t_α（n）≈z_α 得到.

图4-9

由t分布上α分位点的定义以及f_T（x）图形的对称性知

$$t_{1-\alpha}（n）=-t_\alpha（n）.$$

4.4.3 F分布

定义　设V_1 ~χ^2（n_1），V_2 ~χ^2（n_2），且V_1，V_2 相互独立，则称随机变量

服从自由度为n_1，n_2（df=（n_1，n_2））的F分布，记为F~F（n_1，n_2）.

由定义知道，若F~F（n_1，n_2），则

$$1/F~F（n_2，n_1）.$$

（4-24）

图4-10中画出了对应于不同自由度的F分布的概率密度的两条曲线[3].

图4-10

设F~F（n_1，n_2），对于给定的正数α（0<α<1），满足条件

的点F_α（n_1，n_2）就是F（n_1，n_2）分布的上α分位点（见图4-11），此处f_F（x）是F（n_1，n_2）分布的概率密度函数.

图4-11

F分布的上α分位点可自附表4中查到，它还具有以下的性质：

式4-26可用来求表上没有列出的一些F_α（n_1，n_2）的值，例如：

$$F_{0.95}（10，20）=1/F_{0.05}（20，10）=1/2.77=0.36.$$

习题4

【答案链接】

1. 设X~N（μ，σ^2），验证Y=a+bX［a，b（b≠0）为常数］也服从正态分布.

2. 一大学生运动员投掷铅球的距离（以m计）X~N（17，4）.

（1）求P{X≤18.5}，P{X≥15}，P{16<X≤18}；

（2）若P{X>d}=0.95，求d.

3. 设某地区成年男子的体重X（以kg计）服从正态分布N（μ，σ^2），已知P{X≤70}=1/2，P{X≤60}=1/4.（1）求μ，σ；（2）若在这一地区随机地选出5名成年男子，问其中至少有2人体重大于60kg的概率是多少？

4. 盒装人造黄油的重量（以kg计）X~N（0.5，0.003^2 ）.（1）求重量至少为0.495的概率；（2）要使P（0.5-d<X<0.5+d）≥0.95，问d至少是多少？

5. 设随机变量X，Y相互独立，且X服从数学期望为150，方差为9的正态分布；Y服从数学期望为100，方差为16的正态分布.

（1）求X+Y，X-Y，（X+Y）/2的分布；

（2）求P{X+Y<242.6}，P{|（X+Y）/2-125|>5}.

6. 已知随机变量X_1，X_2，X_3，X_4，Y_1，Y_2，Y_3，Y_4 相互独立，且有X_i ~N（4，1）（i=1，2，3，4），Y_i ~N（3，4）（i=1，2，3，4）.

（1）求$X_1+2X_2+3X_3+4X_4$ 的分布；

（2）求P{$X_1<Y_1$ }；

7. 一大批药片每颗重量（以g计）X~N（0.65，0.02^2 ），求30颗药

片的平均重量落在区间（0.64，0.66）之外的概率（设各药片的重量相互独立）.

8. 一个城市的某一出售汽车的商店一年365天都营业，商店一天售出的汽车辆数服从以参数λ=2的泊松分布且各天售出的汽车辆数相互独立，以Y记一年365天中售出的汽车的总数，求：P{Y≥700}，P{Y≤800}.

9. 一零售商店的计算机，为一个顾客结账所花的时间是一个随机变量，均值为1.5min，方差为1min^2 . 各顾客使用计算机的时间相互独立，服从同一分布. 求100个顾客使用计算机的总时间小于2h的概率.

10. n件货物的重量分别为X_1 ，X_2 ，…，X_n （以kg计），将它们装在一辆车上运输，各货物的重量相互独立，服从同一分布，数学期望为0.5，标准差为0.1，要求车上货物的总重量超过2500的概率不大于0.05. 问n至多是多少？

11. 据以往的资料知道，人们患了某种严重疾病，治疗后的存活率为0.4，求100个病人中至少有50人存活的概率.

12. 一工厂生产某种零件的次品率为0.1，用2个这种零件组装成一成品出售. 若2个零件全部完好，可得利润1元；若2个零件中一个为次品、另一个完好，得利润0元；若2个全是次品，损失10元（包括罚款和成本）.

（1）将工厂出售1个成品所得的利润记为X，写出X的分布律；

（2）求工厂卖出1000个成品获得的利润至少为620元的概率（设出售各个成品所得利润相互独立且服从同一分布）.

13. 以X记某地区新生儿的体重（以g计），已知E（X）=3320，D（X）=660^2 ，取n=225个新生儿，以记n个新生儿体重的算术平均. 求概率P{3233.76≤≤3406.24}（设各新生儿的体重相互独立）.

14. 设X_1 ，X_2 ，…，X_n 相互独立，且都服从N（μ，0.2^2 ），要求P{|-μ|>0.01}≤0.001，问n至少是多少？

15. 设X_1 ，X_2 ，…，相互独立，且都在区间（0，1）上服从均匀分布. 验证：对于任意x，有

在上式中取n=12，得

人们常常利用这个式子，再利用伪随机数（参见习题2第14题）产生标准正态分布变量Z的观察值. 又若有Y=μ+σZ，则Y~N（μ，σ^2 ），由此可以产生正态分布随机变量Y的观察值.

---

[1]

[2]

[3] F分布的概率密度函数的表达式比较冗长，这里不写了.

5　参数的点估计

自本章开始将展示数理统计的内容.　数理统计与概率论一样也是研究随机现象的一门学科，它是一门应用性很强的学科.　它以概率论的理论为基础，研究如何有效地、合理地收集数据，建立有效的数学方法，对获得的数据进行分析、处理、研究，从而推断随机现象的客观规律性.

统计推断是数理统计的核心部分，它包括两个基本问题：统计估计和统计检验. 我们将分两章讲述统计推断的一些基本内容.

5.1　总体与样本

在数理统计中，人们将所研究的对象的全体称为总体　，总体中的每一个元素称为个体 . 例如某工厂生产的电阻器是一个总体，每个电阻器是一个个体. 一个湖泊中某种2岁以上的鱼的全体是一个总体，每一条这种鱼是一个个体.　在实际中，我们所要研究的并不是总体中个体的各方面的性质，而只研究个体的某一项数量指标 [1] . 例如，在上述电阻器这一总体中，我们只研究电阻器的电阻值这一数量指标.　又如在湖泊中某种2岁以上的鱼这一总体中，只研究鱼的含汞量这一数量指标. 在总体中，个体的某项数量指标（例如电阻器的电阻值）取各个区间上的值具有一定的百分比.　我们自总体中随机地取一个个体观察它的数量指标，以X记这一数量指标的值，则X是一个随机变量. 我们对总体的研究就是

对相应的随机变量X的分布的研究. X的分布函数和数字特征就称为总体的分布函数和数字特征. 今后将不区分总体和相应的随机变量，笼统称为总体X[2].

在实际中，总体的分布一般是未知的，或部分未知，如只知道它具有某种形式而其中包含未知参数. 在数理统计中，人们都是通过从总体中抽取一部分个体，对它们的所要研究的数量指标进行观察，根据观察获得的数据来对总体分布作出推断的. 我们从总体X中独立地、随机地取n个个体，逐个观察其数量指标，将n次观察结果按观察的次序排列为X_1，X_2，…，X_n. 各次观察结果X_i（i=1，2，…，n）都是随机变量，称X_1，X_2，…，X_n 是一个来自总体X的样本，n称为这个样本的容量. 当n次观察一经完成，就得到一组实数x_1，x_2，…，x_n，它们依次是X_1，X_2，…，X_n 的观察值，称为样本值. 也称为X的n个独立的观察值. 我们用示意图5-1来说明总体、样本、样本值的关系.

图5-1

统计推断的内容是数理统计的核心部分. 统计推断就是利用来自样本的信息推断总体，得到有关总体分布的种种结论. 样本是进行统计推断的依据，对于抽取样本的方式当然要有一定的要求. 最常用的样本需满足以下两个条件：

条件**1** X_1，X_2，…，X_n 相互独立；

条件**2** X_1，X_2，…，X_n 都与总体X具有相同的分布.

满足这两个条件的样本X_1，X_2，…，X_n 称为来自总体X的一个简单随

机样本 .

上述条件2是容易理解的，这表述了通常所说的样本具有代表性；条件1则是为了理论上的探讨和研究的需要.

对于个体总数为有限的总体，用放回抽样就能得到简单随机样本. 但放回抽样用起来不方便，当总体的个体数N比要得到的样本的容量n大得多时，用不放回抽样抽得的样本也可当作简单随机样本使用. 对于个体总数为无限的总体，因抽取一个个体不影响它的分布，故总是用不放回抽样. 本书提到的样本都是指简单随机样本. 此外，可将样本看成随机向量，记成（X_1，X_2，…，X_n）.

5.2 样本数据的图形显示

为了研究总体分布的性质，人们通过试验得到许多观察值，即得到一个数据集. 一般来说，这些数据是杂乱无章的. 为了利用它们进行统计分析，可将这些数据加以整理，还常借助表格或图形对它们加以描述. 用图形来显示数据，也就是将大量数据概括地画在图中，使人们一看就能理解，有一个直观的印象. 作图能帮助研究者从数据集中提取信息并将信息传送给别人. 本节将依次介绍“点图”、“茎叶图”、“直方图”和“箱线图”.

5.2.1 点图

当数据较少时，用点图来显示数据集是合适的. 点图能显示数据在

数据集中的位置. 例如，有19双不同品牌的鞋子，它们的价格（元）分别为

90 70 70 70 75 70 65 68 60 74

70 95 75 70 68 65 40 65 70

画一条水平的数轴，轴上的刻度表明价格. 然后在轴的上方适当位置画上小圆点表示每双鞋的价格. 当数据重复时，所对应的小圆点可画在同一垂线上. 例如，在刻度70之上画了7个小圆点. 上述数据集的点图如图5-2所示.

图5−2

对于两个及两个以上数据集的比较，点图是很有用的.

5.2.2 茎叶图

我们以例题说明“茎叶图”的作法.

【例5-1】 一门诊中心在20天中各天完成心电图检测的人数为

25 31 20 32 13 14 43 2 67 23

36 32 33 32 44 32 62 44 61 45

试作数据集的茎叶图.

解 将数据集中的每个数据分成两部分，领头的数字称为茎 ，后

面的数字称为叶 . 例如，将数31分成两部分，“3”为茎，“1”为叶. 画一条长的竖线，将数据集中各数据的茎自小到大沿竖线左侧自上到下排成一列，如图5-3所示.　　然后将每一个数据的叶在竖线右侧找到合适的位置（相应的茎所在的行），逐一写上 [3] . 这样作成的图形称为数据集的茎叶图.

图5-3

【例5-2】　　下面给出了20个学生一次概率论课程考试的成绩

70　77　86　68　64　95　74　72　88　74

72　86　60　94　92　79　91　75　76　78

试画出数据集的茎叶图.

解　　作出茎叶图如图5-4所示.

图5-4

在图（a）中，一个茎占一行.　由于茎太少，图纵向太短，且其中茎7的叶太多，看起来效果不好. 在图（b）中，一个茎占两行（第一行对应于叶0~4，第二行对应于叶5~9），图（b）比图（a）好.

【例5-3】　　测得两种不同的植物在生长20天后的高度（以cm计）如表5-1所示.

表5-1

为了便于比较两组数据，我们将两组数据作成如图5-5所示的“背对背茎叶图”. 此图的“茎”居中，为两组数据的拾位数部分. 两组数据的个位数

部分，作为“叶”分别写在茎的左右两侧（左侧为数据集A的叶，右侧为数据集B的叶）. 从数据集A、数据集B的茎叶图看到两数据集数据的分布图形类似，数据集B较为分散.

图5-5

茎叶图在数据归类和图形绘制上都很方便，且不需要任何计算. 茎叶图的最大优点是原数据集中关于数据分布的信息仍保留在图中，从茎叶图可以重建全部数据（但重建时数据的顺序一般不同于原始的顺序）. 茎叶图的缺点是一般不能用于容量大的数据集，且对于含有多位数字的数据也难以应用.

5.2.3 直方图

我们将通过一个例子对连续型随机变量X引入“频率直方图”. 它使我们对X的分布有一个粗略的了解，使我们能够评价所假设的模型是否合适，能够估计随机变量落在各个区间上的概率，等等.

【例5-4】 为研究患某种疾病的21~59岁男子的血压（收缩压）这一总体X，抽查了100个男子，测得的数据如表5-2所示（单位：mmHg）.

表5-2

这些数据杂乱无章，先要将它们进行整理. 在表上看到最大和最小观察值分别为174和90. 考虑到表中数据是将实测数据经四舍五入后得到的，取区间I=［89.5，174.5］使得所有实测数据都落在区间I上. 将区间I等分成若干个小区间，小区间的个数与数据的个数n有关，实用上小区

间的个数取为左右为佳. 各小区间的端点坐标常取比表中数据的精度高一位，以免数据落在区间的端点上. 本题取小区间个数为9，于是小区间的长度为（174.5-89.5）/9=9.44. 这一长度用起来不方便，为此将区间I的上限向外延伸至179.5，这样每个小区间的长度调整为

$$\Delta=（179.5-89.5）/9=10.$$

Δ叫做组距 ，小区间的端点叫做组限 . 算出组限，数出落在每个小区间内的数据的个数f_i （即频数；i=1，2，…，9），算出数据落在各个小区间的频率f_i /n（n=100；i=1，2，…，9），所得结果如表5-3所示.

表5-3

如图5-6所示，在每个小区间上作出以对应的频率除以Δ为高，即以（f_i /n）/Δ为高，以小区间为底的小长方形，小长方形的面积就是［（f_i /n）/Δ］×Δ=f_i /n. 这样的图形称为频率直方图 .

图5-6

由于数据来自连续型总体X，由大数定理可知频率接近于概率，因此每个小区间上的小长方形的面积近似于以该小区间为底，以X的概率密度曲线为曲边的小曲边梯形的面积. 因而所得直方图顶部的台阶形曲线近似于X的概率密度曲线（见图5-6）.

由图上看到直方图顶部的台阶形曲线两头低，中间高，有一个峰，且关于中心线比较对称，看来像是接近于某一正态变量的概率密度曲线（将在下一章进一步讨论）. 由直方图可以大致看出总体的分布属于什么类型. 例如在本例中，如果你假设的模型认为X的分布具有指数分布，那么这样的假设显然是不合适的. 从直方图上可以估计出X落在某

一区间上的概率，例如，从图上看到有49%的人血压在（109.5，129.5）之内，仅有12%的人血压高于149.5，血压在124附近的概率为最大，等等.

5.2.4 箱线图

先介绍样本中位数和样本四分位数.

设x_1，x_2，…，x_n是来自总体X的样本值，将它们按自小到大的次序排列为

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} .$$

样本中位数　若n是奇数，则在$x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$中排在最中间的那个数称为样本中位数. 若n是偶数，则在$x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$中排在最中间的两个数的算术平均值称为样本中位数.

例如，对于样本6　9　7　5　4　4　10，将这些数自小到大排序为

4　4　5　6　7　9　10

n=7为奇数，样本中位数为6.

对于样本15　18　21　35　14　3，将这些数自小到大排序为

3　14　15　18　21　35

n=6为偶数，样本中位数是（15+18）/2=16.5.

样本四分位数　如上所述，样本中位数将数据集$x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$分成两个个数相等的部分. 下面介绍的四分位数则几乎将数据集分成四个个数相等的部分. 一个数据集有3个四分位数，将它们自左至右依次记为Q_1，Q_2，Q_3，依次称为第一、第二、第三样本四分位数（Q_2即为样本中位数）. 图5-7依次画出了均匀的分布、对称的分布、右斜的分布、左斜的分布这四种情况的四分位数.

图5-7

有多种计算样本四分位数的方法，各种方法给出的结果都近似相等. 我们采用以下的方法. 先将样本按自小到大的次序排列成

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

然后来找Q_1. 计算0.25（n+1），若这个值是整数，则排在这一整数位置上的数就是Q_1；如果不是，则Q_1就是位于数0.25（n+1）相邻两边的数值的平均值.对于Q_2只要改用0.5（n+1），对于Q_3只要改用0.75（n+1）即可.

例如，对于样本（样本已经过排序）

4　6　6　8　10　11　15　17　22　25　25　27　29（n=13，n+1=14）

Q_1：0.25×14=3.5，不是整数，它位于第3个数与第4个数之间，故有. Q_2：0.5×14=7是整数，故有Q_2 =15. Q_3：0.75×14=10.5不是整数，它位于第10个数与第11个数之间，故有Q_3 =.

将数据集中的最小值$x_{(1)}$（记为Min）、最大值$x_{(n)}$（记为Max）以及Q_1，Q_2，Q_3这5个数按自小到大的次序排列成

$$\text{Min}\quad Q_1\quad Q_2\quad Q_3\quad \text{Max}$$

用来描述数据集的一些特征（例如，Q_2表示数据集的中心，Max-Q_3表示数据集的第4个四分之一部分的离散程度），称为数据集的五数概括 .

基于5个数Min，Q_1，Q_2，Q_3，Max，我们来构造“箱线图”，作法如下.

（1）画一水平数轴，在轴上标上Min，Q_1，Q_2，Q_3，Max. 在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于Q_1、Q_3的上方. 在Q_2点的上方画一条垂直线段. 线段位于箱子内部.

（2）自箱子左侧引一条水平线直至最小值Min，在同一水平高度自箱子右侧引一条水平线直至最大值Max. 这样就将箱线图作好了，如图5-8. 箱线图也可以沿垂直数轴来作.

图5-8

由箱线图可以形象地看出数据集的以下重要性质.

① 中心位置：中位数所在的位置就是数据集的中心.

② 散布程度：全部数据都落在［Min，Max］之内，在区间［Min，Q_1］，［Q_1，Q_2］，［Q_2，Q_3］，［Q_3，Max］内的数据个数约各占1/4，区间较短时，表示落在该区间的点较集中，反之较为

分散.

③ 倾斜：若Min离Q_2 的距离较Max离Q_2 的距离大，则表示数据分布向左倾斜，反之表示数据向右倾斜，且能看出分布尾部的长短.

【例5-5】 在一个班级中随机地取9名学生，他们的数理统计课程期终考试得分为（数据已经过排序）

65 68 72 75 82 85 87 91 95

试画出这一数据集的箱线图.

解 算得下列数据：Min=65，Q_1 =70，Q_2 =82，Q_3 =89，Max=95. 画出箱线图如图5-9所示. 从图上看到，考试得分的分布向左倾斜.

图5-9

将两个或两个以上数据集的箱线图画在同一数轴上，就能比较各个数据集的性质.

【例5-6】 营养学家感兴趣于比较真乳酪与乳酪代用品两者的钠含量，下面分别给出两者的一个样本（两组数据单位相同）如下：

真乳酪（数据集Ⅰ）	310	420	45	40	220	240	180	90
代用品（数据集Ⅱ）	270	180	250	290	130	260	340	310

试在同一数轴上分别画出它们的箱线图，并比较它们的性质.

解　将数据自小到大排序，得到

数据集Ⅰ	40	45	90	180	220	240	310	420
数据集Ⅱ	130	180	250	260	270	290	310	340

易知，对于数据集Ⅰ有：Min=40，Q_1 =67.5，Q_2 =200，Q_3 =275，Max=420. 对于数据集Ⅱ有：Min=130，Q_1 =215，Q_2 =265，Q_3 =300，Max=340. 作出箱线图如图5-10所示.

图5-10

比较这两个图形可以看到，代用品数据的中位数比真乳酪的中位数要大，这两个图形都向左倾斜. 还可以看到真乳酪的钠含量数据的分布较为分散.

在数据集中某一个观察值不寻常地大于或小于该数集中的其他数据，称为疑似异常值 . 疑似异常值的存在，会对随后的计算结果产生不适当的影响. 检查疑似异常值并加以适当的处理是十分重要的. 箱线图只要稍加修改，就能用来检测数据集是否存在疑似异常值.

第一四分位数Q_1 与第三四分位数Q_3 之间的距离Q_3 -，称为四分位数间距 . 若数据小于Q_1 -1.5×IQR或大于Q_3 +1.5×IQR，就认为它是疑似异常值. 我们将上述箱线图的作法作如下的改变：

（1'）同（1）.

（2'）计算IQR=Q_3 -Q_1 ，若一个数据小于Q_1 -1.5×IQR或大于Q_3 +1.5×IQR，则认为它是一个疑似异常值. 画出疑似异常值，并以*表示.

（3'）自箱子左侧引一水平线段直至数据集中除去疑似异常值后的最小值，又自箱子右侧引一水平线直至数据集中除去疑似异常值后的最大值.按（1'）、（2'）、（3'）作出的图形称为修正箱线图 .

【例5-7】　下面给出一医院随机选取的21名病人的住院天数，

4　4　12　18　9　6　12　3　6　15　7　3　55　1　10　13　5　7
1　23　9

试画出这一数据集的箱线图，并识别数据集的疑似异常值.

解　将数据自小到大排列成为

1　1　3　3　4　4　5　6　6　7　7　9　9　10　12　12　13　15
18　23　55（n=21）

现在有Min=1，Max=55，Q_2 =7. 有22×0.25=5.5，故有Q_1 =4. 又22×0.75=16.5，故有，IQR=Q_3 -Q_1 =12.5-4=8.5. Q_3 +1.5×IQR=12.5+1.5×8.5=25.25.

观察值55>25.25，故55是疑似异常值且只有这一个疑似异常值；去掉55以后，得到的最大值为23. 画出修正箱线图如图5-11所示.

图5-11

数据集中，疑似异常值的产生源于①数据的测量、记录或输入计算机时的错误；②数据来自不同的总体；③数据是正确的，但它只体现小概率事件. 当检测出疑似异常值时，人们需对疑似异常值出现的原因加以分析. 如果是由于测量或记录的错误，或某些其他明显的原因造成的，将这些疑似异常值加以更正或从数据集中丢弃就可以了. 然而当出

现的原因无法解释时要作出丢弃或保留这些值的决策无疑是困难的. 此时在对数据集作分析时尽量选用一种方法，使得疑似异常值对结论的影响较小. 例如我们采用中位数来描述数据集的中心，而不使用数据集的平均值，因为后者受疑似异常值的影响较大.

5.3 统计量

样本是进行统计推断的依据，然而在处理具体的理论和应用问题时，往往不是直接使用样本本身，而是针对不同的问题构造样本X_1，X_2，…，X_n 的适当函数，利用这些函数来进行统计推断的.

假设X_1，X_2，…，X_n 是来自总体X的样本，若样本的函数g（X_1，X_2，…，X_n）[4] 不包含未知参数，则称它是一个统计量；若x_1，x_2，…，x_n 是相应于样本X_1，X_2，…，X_n 的样本值，则称g（x_1，x_2，…，x_n）是统计量g（X_1，X_2，…，X_n）的观察值.

值得特别注意的是：统计量是n个独立同分布的随机变量X_1，X_2，…，X_n 的函数，因而统计量是一个随机变量.

下面是四个最常用的统计量，以及它们的一些重要性质.

设X_1，X_2，…，X_n 是来自总体X的样本，则

称为总体X的样本均值 .

称为总体X的样本方差 .

称为总体X的样本标准差 或样本均方差 .

称为总体X的k阶样本矩 .

若x_1 ， x_2 ， …， x_n 是相应于样本X_1 ， X_2 ， …， X_n 的样本值，上述四个统计量的观察值

仍分别称为样本均值、样本方差、样本标准差（样本均方差）和k阶样本矩.

设总体X（不论服从什么分布）的均值E（X）=μ，方差D（X）=σ^2 ，易知

即

$$E（S^2）=\sigma^2 .$$

（5-2）

进而，设X~N（μ，σ^2 ），由4.2节定理的推理可知

还有以下两个重要定理.

定理**1**　设X_1 ， X_2 ， …， X_n 是来自正态总体N（μ，σ^2 ）的样本，， S^2 分别是总体X的样本均值和样本方差，则有

证明略.

定理**2**　设X_1 ， X_2 ， …， X_n 是来自正态总体N（μ，σ^2 ）的样

本，，S分别是总体X的样本均值和样本标准差，则有

证　由式5-4及定理1可知

且两者相互独立. 按t分布的定义即有

化简上式左端即得式5-5.

本节介绍的统计量以及两个定理，在随后的内容中是常要用到的.

在这里，我们还要介绍一个与总体分布函数F（x）相应的统计量——经验分布函数.

定义　设x_1，x_2，…，x_n是来自分布函数为F（x）的总体X的样本观察值. X的经验分布函数记为F_n（x），定义为样本观察值x_1，x_2，…，x_n中小于或等于指定值x所占有的比率，即

其中（#$x_i \le x$）表示x_1，x_2，…，x_n中小于或等于x的个数.

经验分布函数又称样本分布函数 .

按定义，当给定样本观察值x_1，x_2，…，x_n时，F_n（x）是自变量x的函数，它具有分布函数的三条性质：①F_n（x）是x的不减函数；②$0 \le F_n$（x）≤1，且F（-∞）=0，F（∞）=1；③F_n（x）是一个右连续函数，由此知F_n（x）是一个分布函数. 当x_1，x_2，…，x_n各不相同时，F_n（x）是以等概率1/n取x_1，x_2，…，x_n的离散型随机变量的分布函数[5].

一般地，设x_1 ，x_2 ，…，x_n 是总体X的一个容量为n的样本观察值，先将x_1，x_2，…，x_n 按自小到大的次序排列，并重新编号为

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$$

则经验分布函数F_n（x）可写成

例如，设总体X有样本观察值2，-1，1，经排序为$x_{(1)}$ =-1，$x_{(2)}$ =1，$x_{(3)}$ =2，得经验分布函数为（如图5-12）

图5-12

另一方面，当给定x时，F_n （x）是样本X_1 ，X_2 ，…，X_n 的函数，因而它是一个统计量. 格里汶科（Glivenko）在1933年给出了以下的定理：

定理**3**（格里汶科定理） 设X_1，X_2，…，X_n 是来自以F（x）为分布函数的总体X的样本，F_n （x）是经验分布函数，则有

（证明略）.

定理的含义是F_n （x）在整个实轴上以概率1均匀收敛于F（x）. 于是当样本容量n充分大时，F_n （x）能够良好地逼近总体分布函数F（x）. 这是在统计学中以样本推断总体的依据.

5.4 参数的点估计

参数估计和假设检验是统计推断的两个基本问题. 参数估计分为点估计和区间估计. 本章只讨论参数的点估计，参数的区间估计问题以及假设检验问题将在下一章讨论.

有很多情况，人们对于所研究的总体已经有了某些信息，例如常常有理由假设总体的分布函数具有已知的形式，但其中包含一个或多个未知参数. 例如，已知总体具有正态分布N（μ，σ^2 ），但参数μ，σ^2 未知. 人们无法知道未知参数的真值，只能利用样本去估计未知参数. 为了了解总体的分布，我们需要对参数进行估计.

例如，已知某种电子元件的寿命X服从参数为β的指数分布，而参数β未知. 今抽查了6个元件测得以下的数据（单位：年）：1.9，2.7，4.8，3.1，3.4，2.4，我们需要利用这些数据来估计β. 由于β=E（X），由大数定理知道当n很大时样本均值接近于E（X）. 我们自然想到以的观察值来估计E（X）. 由已知数据得到

一般，设总体的分布函数F（x；θ）的形式已知，θ是待估计的未知参数 [6] ，X_1 ，X_2 ，…，X_n 是来自X的样本，选择一个合适的统计量h（X_1，X_2，…，X_n ）称为θ的估计量 . 每当有了一个样本值x_1 ，x_2 ，…，x_n ，将样本值代入统计量h，得到这一统计量的一个观察值h（x_1 ，x_2 ，…，x_n ），以此作为θ的估计，h（x_1 ，x_2 ，…，x_n ）称为θ的估计值 . 由于估计值是一个数值，画在直线上是一个点，因而称它为θ的点估计 . 在不致混淆的情况下估计量和估计值统称为估计 ，都记为.

要注意的是，估计量是一个随机变量，而估计值是一个数值，对于不同的样本值，估计值一般是不相同的.

下面介绍求估计量的两种常用的方法：矩估计法和最大似然估计法，并介绍评定估计量好坏的标准.

5.4.1 矩估计法

在上述例题中，以样本均值作为总体均值E（X）的估计量，也就是以一阶样本矩作为一阶总体矩的估计量，从而得到未知参数β的估计量，这种做法实际上是矩估计法.

一般，若总体X的分布函数的形式已知，其中只含一个未知参数θ. 设X_1 ，X_2 ，…，X_n 是来自X的样本，设X为连续型随机变量，其概率密度为f（x；θ）；或X为离散型随机变量，其分布律为$P\{X=x_i\}=p(x_i;\theta)$（i=1，2，…），那么一阶总体矩

它是θ的函数：$\mu_1=\mu_1(\theta)$，可用一阶样本矩作为μ_1 的估计量. 令

$$\mu_1=\mu_1(\theta)=A_1,$$

从中解出θ. 以记上述方程的解，就以=h（X_1 ，X_2 ，…，X_n ）作为θ的估计量，称为矩估计量 . 若以样本值x_1 ，x_2 ，…，x_n 代入h就得到θ的矩估计值 ：=h（x_1 ，x_2 ，…，x_n ）.

若总体的分布函数的形式已知，其中含两个未知参数θ_1 ，θ_2 ，那么当X为连续型，其概率密度为f（x；θ_1 ，θ_2 ）时，一阶、二阶总体矩为

它们是θ_1，θ_2 的函数：$\mu_1=\mu_1$（θ_1，θ_2），$\mu_2=\mu_2$（θ_1，θ_2），可用一阶样本矩，二阶样本矩分别作为μ_1，μ_2 的估计量. 令

在上述方程组中解出θ_1，θ_2，分别记为；分别以X_2，…，X_n），（X_1，X_2，…，X_n）作为θ_1，θ_2 的估计量，并称之为θ_1，θ_2 的矩估计量. 对于离散型总体X，可同样讨论[7].

【例5-8】 设总体X的概率密度

其中$\theta>-1$为待估参数. 设X_1，X_2，…，X_n 是来自X的一个样本，试求θ的矩估计量.

解 总体X的一阶矩

从中解出θ，即得θ的矩估计量

若有一个样本值x_1，x_2，…，x_n，则得θ的估计值

【例5-9】 设总体X的概率密度

其中μ，β（$\beta>0$）为待估参数；X_1，X_2，…，X_n 是来自X的样本. 试求μ，β的矩估计量.

解 总体X的一阶、二阶矩：

【例5-10】 设总体X的均值μ，方差$\sigma^2>0$均未知，X_1，X_2，…，X_n 是来自X的样本. 试求μ，σ^2 的矩估计量.

解　总体X的一阶、二阶矩：

解得μ，σ^2 的矩估计量分别为

所得结果表明，在μ，σ^2 均未知时，μ，σ^2 的矩估计量不因不同总体而异，即不论X服从什么分布，X的均值μ和方差σ^2　的矩估计量都分别是和.

5.4.2　最大似然估计法

先举一个简单的例子说明最大似然估计法的思路.

【例5-11】　临近新年时，人们常常收到朋友寄来的贺卡. 有一天某人收到4封信，不小心丢失了一封（设丢失哪一封是等可能的），将留下的3封拆开，其中有2封装有贺卡. 现在来估计4封信中装有贺卡的信件数k，显然k=2或3.　记A为事件“留下的3封中有2封装有贺卡”.当k=2或k=3时，A的概率分别为

现在观察到事件A已经发生，自然认为A发生的概率是比较大的. 由于当k=3时比当k=2时A的概率要大，可取3作为k的估计，亦即取参数k的估计值，使得事实上已经发生的事件A的概率为最大.

最大似然估计法的基本想法：若事件A的概率依赖未知参数θ，如果观察到A已经发生，那么就在θ的可能取值范围Θ内取θ的估计值使事件A的概率为最大.

设总体X为离散型，其分布律为$P\{X=x\}=p(x;\theta)$（$\theta\in\Theta$），θ为

待估参数，设X_1 ，X_2 ，…，X_n 是来自总体X的样本，则X_1 ，X_2 ，…，X_n 的联合分布律为

设现在已经得到样本X_1，X_2，…，X_n 的一个样本值x_1，x_2，…，x_n . 我们知道样本X_1，X_2，…，X_n 取到样本值x_1，x_2，…，x_n 的概率为

（注意：这里x_1 ，x_2 ，…，x_n 是已知样本值，是常数）. 它是θ的函数，对于不同的θ，这一概率是不相同的. 记

L（θ）称为似然函数 . 现在事件A={X_1 =x_1 ，X_2 =x_2 ，…，X_n =x_n }已经发生，基于上述最大似然估计法的基本想法，可选取θ的估计值，使得P（A）的概率L（θ）取到最大. 于是由下式确定：

设总体X为连续型，其概率密度为f（x；θ）（θ∈Θ），其中θ为待估参数，设X_1 ，X_2 ，…，X_n 是来自总体X的一个样本，则X_1 ，X_2 ，…，X_n 的联合密度

设现在已经得到样本X_1 ，X_2 ，…，X_n 的一个样本值x_1 ，x_2 ，…，x_n ，可知样本（X_1 ，X_2 ，…，X_n ）落在点（x_1 ，x_2 ，…，x_n ）邻域里的概率近似地为

（这里x_1 ，x_2 ，…，x_n ，是常数）. 它是θ的函数，对于不同的θ，这一概率是不相同的. 现在事件A：“（X_1 ，X_2 ，…，X_n ）落在点（x_1 ，x_2 ，…，x_n ）的邻域里”已经发生，基于上述最大似然估计法的基本想法，可选取θ的估计值使得A的概率（式5-7）取到最大. 考虑到为常数，于是由下式确定：

综上所述，求θ的最大似然估计的问题就归结为求似然函数L（θ）的最大值点的问题. 经常遇到的情况是，L（θ）关于θ的导数存在，于是θ的估计，可由方程

解得[8]. 更方便的是由方程

解得（方程式5-9与方程式5-10有相同的解）.方程5-9和方程5-10分别称为似然方程 和对数似然方程 . 以上所说的估计参数的方法称为最大似然估计法 .

【例5-12】 设总体X~π（λ），λ>0为待估参数；x_1，x_2，…，x_n是来自X的一个样本值. 试求λ的最大似然估计值和估计量.

解 总体X的分布律为

$$P\{X=x\}=p（x；λ）=λ^x e^{-λ}/x!（x=0，1，2，…），$$

似然函数

由于E（X）=λ，这就是说，以样本均值作为总体均值E（X）的最大似然估计量，与λ的矩估计量相同.

【例5-13】 在例5-8中求参数θ的最大似然估计. 设x_1，x_2，…，x_n是一个样本值.

解 似然函数

它与θ的矩估计量不一样.

【例5-14】 设总体X在区间［0，θ］上服从均匀分布，θ>0未知.

试由样本值x_1，x_2，…，x_n求θ的最大似然估计值.

解　X的概率密度

似然函数

记$x_{(n)}$ =max（x_1，x_2，…，x_n），可知$x_{(n)}$ >0. 而上述似然函数相当于

即知当θ<$x_{(n)}$ 时，L（θ）=0，而当θ≥$x_{(n)}$ 时，L（θ）随θ的增加而减少. 故L（θ）在$x_{(n)}$ 取到最大值（见图5-13），得θ的最大似然估计值

图5-13

本题的最大似然估计值在L（θ）的间断点处取到，不能利用对L（θ）求导的方法得到.

最大似然估计法也适用于总体分布函数中含有多个未知参数θ_1，θ_2，…，θ_k（k>1）的情况，此时似然函数L是这些参数的函数. 分别令

解上述方程组，一般来说就能得到θ_1，θ_2，…，θ_k 的最大似然估计.

【例5-15】　设X~N（μ，σ^2），μ，σ^2 >0未知，x_1，x_2，…，x_n是一个样本值. 求μ，σ^2 的最大似然估计.

解　似然函数

它们与相应的矩估计量是一样的.

今设有某种零件的长度（以cm计）X~N（μ，σ^2 ），随机地抽取8个零件，分别测得其长度如下：

37.0　37.4　38.0　37.3　38.1　37.1　37.6　37.9

则可得零件长度的均值及方差的最大似然估计分别为

5.4.3　评定估计量好坏的标准

对于总体分布中的一个未知参数可以提出不同的估计量，例如对于例5-8中的未知参数θ得到的矩估计量与最大似然估计量就不相同.　因此自然会提出比较好坏的问题，这就需要给出评定好坏的标准.

估计量是一个随机变量，对于不同的样本值，一般给出参数的不同估计值，因而在考虑估计量的好坏时，应从某种整体性能去衡量，而不能看它在个别样本之下的表现如何.　下面介绍三个常用的评定估计量好坏的标准.

设X_1 ，X_2 ，…，X_n 是来自总体X的一个样本，θ∈Θ是总体X分布中的待估参数，Θ是θ的可能取值范围.

1）无偏性

无偏估计量并不意味着对于每个样本值x_1 ，x_2 ，…，x_n 给出的θ的估计值就是θ的真值，只是说对于某些样本值得到的估计值相对于真值θ来说偏大，有些则偏小，反复使用多次，就“平均”来说偏差为0，因此无偏性可解释为不存在系统误差.

例如，设总体X的均值为μ，方差σ^2 >0均未知，则有（见式5-1和式5-2）

若估计量都是参数θ的无偏估计量，显然其取值更集中在θ附近的估计量要好一些. 因而，如果对于任意θ∈Θ，有

亦即

这就是说两个无偏估计量以方差小的为较好，这就引出了下述有效性的概念.

2）有效性

例如设X_1，X_2，…，X_n 是来自总体X的样本，则X_1，都是总体均值μ的无偏估计，而D（X_1 ）=D（X），D（）=D（X）/n，当n>1时，D（）<D（X_1 ），即当n>1时，较X_1 为有效.

无偏性和有效性，都是在样本容量n固定的前提下提出的. 当n增大时，一般来说样本中包含的信息会增多，我们要求当n增加时估计量能充分地接近于待估计参数的真值，这就引出了相合性的要求.

3）相合性

设（X_1 ，X_2 ，…，X_n ）是参数θ的估计量. 若对于任意θ∈Θ都满足：对于任意ε>0，有

相合估计量也称一致估计量.

由3.4节的定理2，知道样本均值是总体均值μ的相合估计量，还可

以证明样本方差S^2也是总体方差σ^2的相合估计量.

习题5

【答案链接】

1. 随机选取一个健康的人，对他的血红细胞数（以10^6个/微升计）进行了测试，共测试了15天，得到以下的数据：

54 52 50 52 55 53 54 52 51 53 53 49 54 52 52

试作出这一数据集的点图.

2. 为了改善驾驶员的工作条件，测试了他们的心率（以每分钟的搏动次数计）作研究，得数据如下：

74 52 67 63 77 57 80 77 53 76 54 73 54 60 77 6360 68 64 66 71 66 55 71 84 63 73 59 68 64 82

试作这一数据集的茎叶图：（1）以5，6，7，8为茎；（2）以5，5，6，6，7，7，8为茎.

3. 瓶装苏打水的规格为1升. 为了保证产品的容量，取了30瓶样品，测得苏打水的容量（以毫升计）如下：

1025 977 1018 975 977 990 959 957 1031 964 986 914 1010 988 1028989 1001 984 974 1017 1060 1030 991 999 997 996 1014 946 995 987

试以91，92，…，106为茎作数据集的茎叶图.

4. 下面给出某24个国家1995年的女子失业率和男子失业率数据：

女子（%）：　8.0　3.7　8.6　5.0　7.0　3.3　8.6　3.2　8.8　6.8
9.2　5.9　7.2　4.6　5.6　5.3　7.7　8.0　8.7　0.5　6.5　3.4　3.0　9.4

男子（%）：　8.8　1.9　5.6　4.6　1.5　2.2　5.6　3.1　5.9　6.6
9.8　8.7　6.0　5.2　5.6　4.4　9.6　6.6　6.0　0.3　4.6　3.1　4.1　7.7

试以0，1，2，…，9为茎作两数据集的背对背茎叶图.

5. 试在同一数轴上分别画出例5-3中数据集A和数据集B的箱线图.

6. 下面给出某发电厂内4个不同地点噪声（以分贝计）的数据：

地点1（Ⅰ）：30　12　35　65　24　59　68　57　100　61　32　45
92　56　44

地点2（Ⅱ）：64　99　87　59　23　16　94　78　57　32　52　78
59　55　55

地点3（Ⅲ）：100　59　78　97　84　64　53　59　89　88　94　66
57　62　64

地点4（Ⅳ）：25　15　30　20　61　56　34　22　24　21　32　52
14　10　33

试在同一数轴上分别作数据集Ⅰ，Ⅱ，Ⅲ，Ⅳ的箱线图.　如有疑似异常值，则画出疑似异常值和修正箱线图.

7. 下面给出某种氯苯那敏药片的氯苯那敏有效含量（以mg计）的数据：

4.02 3.86 3.96 3.97 4.00 3.82 3.98 3.99 4.02 3.93

试作箱线图. 如有疑似异常值，则画出疑似异常值和修正箱线图.

8. 下面给出的数据集Ⅰ是8个品牌的切片乳酪的含钠量（以mg计），数据集Ⅱ则是另外5个品牌的切片乳酪的含钠量（以mg计）.

数据集Ⅰ：340 300 520 340 320 290 260 330

数据集Ⅱ：300 300 320 290 180

试在同一数轴上分别画出数据集Ⅰ、数据集Ⅱ的箱线图. 如有疑似异常值，则画出疑似异常值和修正箱线图.

9. 下面给出了40个商店在某两个月内售出的某种新汽车的辆数. 试画出这些数据的频率直方图［将区间（59.5，94.5）分成7等份］.

10. 某妇产科医院在91天中诞生的婴儿数如下：

画出这些数据的频率直方图［将区间（-0.5，34.5）分成7个等份］.

11. 设总体X的概率密度为

X_1，X_2，X_3 是来自总体X的样本.（1）求X_1，X_2，X_3 的联合概率密度；（2）求E（X_1 X_2 X_3 ）；（3）求.

12. （1）设X_1，X_2，…，X_n 和Y_1，Y_2，…，Y_n 都是来自总体

N（0，4）的样本，且两样本独立，问n至少取多少才能使得.

（2）设X_1，X_2，…，X_9 和Y_1，Y_2，…，Y_{16} 是分别来自总体N{0，4}，N（1，9）的两个样本，且两样本独立. 求P{Y_1 <0}，P{X_1 +Y_1 >1}，并问服从什么分布？

13. 设总体X~χ^2（5），X_1，X_2 是来自总体X的样本.（1）问X_1 +X_2 服从什么分布？（2）设Y=X_1 +X_2，求E（Y），D（Y）.

14.（1）设总体X具有概率密度

X_1，X_2，…，X_{10} 是来自总体X的一个样本. 试写出样本均值和样本方差S^2，并求的均值和方差.

（2）在某工厂生产的轴承中随机地取10个，测得其重量（以kg计）为

2.36　2.42　2.38　2.34　2.40　2.42　2.39　2.43　2.39　2.37

求：样本均值、样本方差和样本标准差.

（3）对第9题中的样本求样本均值和样本方差.

（4）对第10题中的样本求样本均值和样本方差.

15. 求下列各题中未知参数的矩估计量. 设X_1，X_2，…，X_n 是来自总体X的一个样本，X的概率密度：

（3）设总体X~B（m，p）；m，p（0<p<1）均未知（若求得的不

是整数，则取与最接近的整数作为m的估计）.

16. 设x_1，x_2，…，x_n 是来自总体X的样本值，求未知参数的最大似然估计值.

（1）X的概率密度如上题（2）；

（2）X~B（m，p）（0<p<1，p为未知参数）；

（3）X服从瑞利分布，其概率密度

θ>0为未知参数；

（4）设X的分布律为

X	1	2	3
p_k	θ^2	2θ（1-θ）	（1-θ）^{2}

0<θ<1，θ为未知参数，已知取得一个样本值（x_1，x_2，x_3）=（1，2，1），求未知参数θ的最大似然估计值.

17. 设有来自总体N（μ，σ^2）的样本如下：

12.7　6.6　5.6　14.3　11.4　4.3　7.2　10.8

13.8　11.2　10.0　12.8　7.1　14.0　6.1

求μ，σ^2 的最大似然估计值.

18. 设元件的寿命X（以h计）服从以β为参数的指数分布. 今随机地

取5个元件，测得寿命为30.4，7.8，1.4，13.1，67.3，求β的最大似然估计值.

19. 某种小型计算机一星期中的故障次数Y~π（λ），设Y_1，Y_2，…，Y_n 是来自总体Y的样本.（1）验证是λ的无偏估计；（2）设一星期中故障修理费用$Z=3Y+Y^2$，求E（Z）；（3）验证：是E（Z）的无偏估计.

20. 用电压表测量电路的电压，电压表的读数在区间（θ，θ+1）服从均匀分布，设X_1，X_2，…，X_n 是一个读数样本.（1）验证若以作为θ的估计，它不是无偏的；（2）找出一个θ的无偏估计量.

---

[1] 在多元统计学中，研究总体中个体的多项数量指标.

[2] 举一个例子：设总体由标有号码的10个球组成，其中有2个标有号码1，有3个标有号码2，有5个标有号码3，我们研究球的号码这一数量指标. 这一数量指标取各个值有一定的百分比，它取1，2，3的百分比分别为20%，30%，50%. 我们在总体中随机地取一个球，以X表示它的号码，则X是一个随机变量，其分布律是

X	1	2	3
p_k	20%	30%	50%

X的分布律就称为总体的分布律.

[3] 图中，在茎5的右边是空格，这表示数据集中没有领头的数字

为“5”的数，但茎5仍需保留.

[4] 一般设g为连续函数.

[5] 当x_1，x_2，…，x_n中有相同的数时，例如$x_1=x_2=x_3$而x_4，x_5，…，x_n各不相同，则$F_n(x)$是以概率3/n取x_1，而以等概率1/n取x_4，x_5，…，x_n的离散型随机变量的分布函数.

[6] 当未知参数多于一个时，可同样讨论.

[7] 矩估计法也可以应用于含两个以上未知参数的情况，其做法与仅含两个未知参数时类似.

[8] 当然，仅当这个方程的解满足最大值的充分条件时，才是我们所要求的. 对于具体问题是容易看清楚的.

6　假设检验与区间估计

本章继续讲述统计推断的内容，介绍假设检验与区间估计.

6.1　假设检验

人们常常对所研究总体的某些感兴趣的未知特性提出某种陈述. 例如，提出：“某地区男子的平均高度为1.68m”，“一个班级概率论课程的考试分数服从正态分布”，“某一工厂生产的灯泡的平均寿命为1000h”等等. 这种关于总体分布函数形式的陈述或关于总体参数的陈述称为统计假设 . 为了判断一个统计假设是否成立，就需要进行随机试验，收集数据，对数据进行分析，然后作出接受或拒绝这一假设的决策. 假设检验就是一个决策过程.

下面用例题来说明假设检验的基本思想和做法.

【例6-1】　假设要从某工厂购买一批电阻值为200Ω的电阻器，已知该厂生产的电阻器的电阻值$X\sim N(\mu,\sigma^2)$. 据以往的经验，$\sigma=8\Omega$，一般不会改变. 某日该工厂送来一大批电阻器，我们要求这批电阻器电阻值总体的均值为200Ω. 为此提出两个相互对立的统计假设：

$$H_0：\mu=\mu_0=200，H_1：\mu\neq\mu_0，$$

（6-1）

并随机地取16个电阻器，测量它们的电阻值.得到下面的数据：

206.4　204.1　199.4　200.9　181.1　191.2　179.8　202.2

194.2　206.3　207.9　200.8　202.4　201.7　203.9　204.6

基于这些数据，我们将作出是接受H_0 还是拒绝H_0 的决策. 如果接受H_0，就认为电阻值的均值μ=200Ω，从而就购买这批电阻器；否则拒绝购买.

由于要检验的是关于总体均值μ的假设，自然想到借助于μ的无偏估计. 将标准化得到

当H_0 为真时$\mu=\mu_0$，此时

已知σ=8，n=16，因而Z中不含未知参数，它是一个统计量，且有

注意到当H_0 为真时，是μ_0 的无偏估计，的观察值应落在μ_0 的附近，即Z的观察值应落在0的附近，而落在Z的分布的两侧尾部是很少遇到的. 另一方面，如果H_1 为真，譬如说μ的真值为μ_1 （$\neq\mu_0$ ），此时的观察值应落在μ_1 的附近，如此|Z|的观察值就偏离0，有偏大的倾向.

如上所说，当H_0 为真时，Z的观察值落在Z的分布的两侧尾部是很少遇到的. 现在指定一个小概率α，当H_0 为真时，使得事件

的概率为α，即

（这里$P_{H_0}\{\ \}$表示当H_0 为真时，即当$\mu=\mu_0$ 时事件{ }的概率）. 由于当H_0

为真时

得到$k=z_{\alpha/2}$，即有

这就是说，当H_0 为真时，事件$A=\{|Z|\geq z_{\alpha/2}\}$是一个小概率事件. 基于“概率很小的事件在一次试验中，实际上几乎是不发生的.　”这一小概率原理，如果Z的观察值使得$|Z|\geq z_{\alpha/2}$，这表明，在一次试验中小概率事件$A=\{|Z|\geq z_{\alpha/2}\}$竟然发生了，我们自然怀疑$H_0$ 的真实性，从而作出拒绝H_0 的决策. 如果Z的观察值未能使$|Z|\geq z_{\alpha/2}$，这表明未能使小概率事件$A=\{|Z|\geq z_{\alpha/2}\}$发生，我们没有找到理由拒绝$H_0$，那就作出接受$H_0$ 的决策.

在本例中，若取$\alpha=0.05$，则$z_{\alpha/2}=z_{0.025}=1.96$. 已知$\sigma=8$，$n=16$，又由样本值计算得到，此时Z的观察值z使

于是可作出接受H_0 的决策，认为电阻值总体的均值$\mu=200$，从而购买这批电阻器.

以上的做法是，取统计量，给定一个小的数α，$0<\alpha<1$，确定k，使得

由当H_0 为真时，，得到$k=z_{\alpha/2}$，则当Z的观察值z使

时，拒绝H_0 （即接受H_1），认为H_0 为不真，而当时，接受H_0 （即拒绝H_1），认为H_0 为真.

像这种判断假设H_0 是真或不真的法则，称为检验法则 .

我们称H_0为原假设，H_1为对立假设 或备择假设 . 在作检验时所使用的统计量称为检验统计量 . 如上述统计量（式6-3）是用来检验假设检验问题（式6-1）的，它是检验统计量.　当检验统计量取某一区域的值时，拒绝原假设H_0，这个区域称为拒绝域 或称临界域 . 临界域的端点称为临界点 或临界值 . 例如在上例中拒绝域为$|z|\geq z_{\alpha/2}$，而$z=-z_{\alpha/2}$，$z=z_{\alpha/2}$为临界值. 在上述做法中先给定一个小的数α（$0<\alpha<1$），接着由检验统计量的分布确定拒绝域（式6-6），这里α称为检验的显著性水平 . 当H_0被拒绝时，也称μ与μ_0有显著差异.

在确定了拒绝域后，取一样本，当Z的观察值z落在拒绝域内就拒绝H_0，否则就接受H_0. 然而，由于样本的随机性，当H_0为真时，Z的观察值也会落入拒绝域，此时致使我们作出拒绝H_0的错误决策. 这种错误称为第Ⅰ类错误 . 由式6-5，当H_0为真时，Z的观察值z落入拒绝域的概率仅为α，因此，我们犯第Ⅰ类错误的概率就是检验的显著性水平α，即

$$P\{\text{犯第Ⅰ类错误}\}=P\{\text{当}H_0\text{为真拒绝}H_0\}=\alpha.$$

（6-7）

另一方面，当H_0为不真时，Z的观察值也会未落入拒绝域，此时致使我们作出接受H_0的错误决策. 这种错误称为第Ⅱ类错误，犯第Ⅱ类错误的概率常记为β，即

$$P\{\text{犯第Ⅱ类错误}\}=P\{\text{当}H_0\text{为不真接受}H_0\}=\beta.$$

（6-8）

我们希望α，β都很小，然而进一步讨论可知，当样本容量n取定时，若减小α则必增大β；反之亦然. 只有增加样本容量n，α和β才能同时减小. 以上所建立的检验法则，显著性水平α是事先选定的，因而我们可以控制犯第Ⅰ类错误的概率. 这种只考虑控制犯第Ⅰ类错误的概率而不考虑犯第Ⅱ类错误的概率的检验法则称为显著性检验 . 本书只介绍这种检验. 通常显著性水平α取0.1，0.05，0.01，0.005，0.001等[1] .

在检验问题6-1中，在数轴上备择假设H_1 中μ的值位于H_0 中μ的值μ_0 的左右两边，称式6-1为双边检验问题 . 在实际问题中，有时我们关心总体均值是否会增大. 例如，采用甜菜的新品种，我们关心新品种能否增加甜菜的糖分含量. 若能判断新品种生产的甜菜糖分含量这一总体的均值比老品种大，则采用新品种. 在这种情况下，若分别以μ，μ_0 记在新、老品种之下的糖分含量的均值，则我们需要检验假设：

$$H_0：\ \mu=\mu_0，\ H_1：\ \mu>\mu_0 .$$

（6-9）

此处设新品种的糖分含量不会比老的低，因此我们已摒除了$\mu<\mu_0$ 的可能性，故检验问题（式6-9）中的H_0 和H_1 是两个对立的假设. 因在数轴上，H_1 中μ的取值位于H_0 中μ的值μ_0 的右方，称式6-9为右边检验 问题. 有时则需检验下述假设：

$$H_0：\ \mu=\mu_0，\ H_1：\ \mu<\mu_0$$

（6-10）

（此处设$\mu \succ \mu_0$ ）. 式6-10称为左边检验问题 . 右边检验、左边检验统称为单边检验 .

今对于方差已知为σ^2 的总体N（μ，σ^2 ），在给定的显著性水平α下，讨论单边检验问题的拒绝域.

先考虑右边检验问题，使用统计量

在检验问题6-9中，当H_0 为真时，Z的观察值应落在0的附近；另一方面，若H_1 为真，譬如说μ的真值为μ_1 （$>\mu_0$ ），此时的观察值应落在μ_1 的附近，如此，Z的观察值就有偏大的倾向. 因而拒绝域的形式为

取显著性水平为α，因为当H_0 为真时Z~N（0，1），所以由

得$k=z_\alpha$ ，故检验问题6-9的拒绝域为

类似地，对于检验问题6-10，其拒绝域的形式为

由

得$c=-z_\alpha$ ，故检验问题6-10的拒绝域为

【例6-2】 在生产线上装配某种产品，在正常情况下，一件产品所需的装配时间（以min计）X~N（10，1.12^2 ）. 某日管理人员怀疑平均装配时间μ超过10，因而提出要检验假设（取显著性水平α=0.1）：

$$H_0：\mu=10，H_1：\mu>10.$$

（6-13）

据以往经验知σ=1.12一般不会改变. 现在随机地观察了25件产品的装配时间，得到样本均值. 由式6-11可知检验问题（式6-13）的拒绝域为

Z的观察值落在拒绝域内，故拒绝H_0 ，认为平均装配时间显著地大于10，于是车间管理人员需要暂时停止生产而去调整机器.

综上所述，可得当总体分布形式已知时，参数的假设检验的步骤如下：

（1）根据实际问题的要求提出原假设H_0 和备择假设H_1 .

（2）给出显著性水平α，选择合适的检验统计量，给出拒绝域的形式，然后按P{当H_0 为真拒绝H_0 }=α确定拒绝域.

（3）根据样本值计算检验统计量的观察值.

（4）作出决策，即当检验统计量的观察值落在拒绝域内，则拒绝原假设H_0 ；否则接受原假设H_0 .

6.2 正态总体均值的假设检验

6.2.1 单个正态总体N（μ，σ^2 ）均值μ的检验

设总体X~N（μ，σ^2 ），μ未知，σ^2 已知或未知，X_1 ，X_2 ，…，X_n

是来自总体X的样本，我们需要检验关于均值μ的假设：

（1）H_0：$\mu=\mu_0$，H_1：$\mu>\mu_0$；（6-14）

（2）H_0：$\mu=\mu_0$，H_1：$\mu<\mu_0$；（6-15）

（3）H_0：$\mu=\mu_0$，H_1：$\mu\neq\mu_0$.（6-16）

先设方差 σ^2 已知 .

在上节中，我们利用检验统计量

已经得到这些假设检验的拒绝域（见表6-1）：

表6-1

图6-1

以上检验所使用的检验统计量，当H_0 为真时服从正态分布，称为Z检验 .

再设方差 σ^2 未知 .

当σ^2 未知时，包含未知参数σ，不能用来作为检验统计量. 注意到样本方差S^2 是σ^2 的无偏估计，用S代替中的σ，得到

以T作为检验统计量. 当H_0 为真时，T~t（n-1），与上节类似可得到以下的结果（见表6-2）：

表6-2

图6–2

以上检验使用的检验统计量当H_0为真时服从t分布，称为t检验 .

【例**6-3**】　在某一地点对重力加速度（cm/s^2）进行了18次测试，得到样本均值为976，样本标准差为6.　设样本来自正态总体N（μ，σ^2），μ，σ^2 均为未知. 问：所得结果是否表示重力加速度总体均值（总体均值就是重力加速度常数g）显著小于980（取α=0.05）.

解　我们提出以下的假设：

$$H_0：\mu=\mu_0=980，H_1：\mu<\mu_0 .$$

今α=0.05，n=18，$t_{0.05}$（17）=1.7396，这是左边检验，由表6-2可知其拒绝域

T的观察值落在拒绝域内，故拒绝H_0，认为总体均值显著小于980，亦即认为重力加速度常数显著小于980.

t检验在实际中用得较多，因为总体方差往往是未知的.

6.2.2　两正态总体均值的比较

在实际中，人们常需要对两个总体的某个或某些参数进行比较. 例如：为了决定在甲工厂或乙工厂购买一批灯泡时，需要比较这两个工厂生产的灯泡寿命的均值，以选择购买寿命均值大的灯泡；在考虑两种不同的投资方法时，需要比较两种方法的资金的平均利润而选择具有高平

均利润的；一个环境保护工程师感兴趣于比较被污染的环境和自然环境下微生物的平均增长率.

下面讨论用来比较两正态总体的均值的假设检验.

设总体，μ_X ，μ_Y 未知，已知或未知. X_1 ，X_2 ，…，X_{n_1} 和Y_1 ，Y_2 ，…，Y_{n_2} 是分别来自总体X，Y的样本，两样本独立 [2] . 样本均值分别为，样本方差分别为.

我们需要检验假设：

1° H_0：$\mu_X=\mu_Y$，H_1：$\mu_X>\mu_Y$，

2° H_0：$\mu_X=\mu_Y$，H_1：$\mu_X<\mu_Y$，

3° H_0：$\mu_X=\mu_Y$，H_1：$\mu_X\neq\mu_Y$.

先设均为已知

现在来检验假设（显著性水平为α）

$$H_0：\mu_X=\mu_Y，H_1：\mu_X\neq\mu_Y .$$

（6-18）

当H_0 为真时，统计量

取式6-20左端作为检验统计量，可知检验问题式6-18的拒绝域的形式为

故得检验问题式6-18的拒绝域为

上述检验问题1°，2°的拒绝域在表6-3中列出.

表6-3

【例6-4】 为比较两种白酒A和B中的灰尘含量，测得以下的数据：

设样本依次来自正态总体，σ_X =0.027g/l，σ_Y =0.040g/l，两样本独立，试取显著性水平α=0.05检验假设

$$H_0：\mu_X=\mu_Y，H_1：\mu_X\neq\mu_Y.$$

解 n_1 =10，n_2 =14，拒绝域为

故接受H_0，认为两种白酒灰尘含量的平均值相等.

再设方差未知

现在来检验假设（显著性水平为α）

$$H_0：\mu_X=\mu_Y，H_1：\mu_X\neq\mu_Y$$

（6-21）

可以证明随机变量（式6-22）近似服从t分布，其自由度v按下式计算（使用时将自由度取为［v］）.

取

作为假设检验统计量，由式6-22知当H_0为真时，这一统计量近似服从自由度为v的t分布，从而得检验问题（式6-21）的拒绝域为

另外两个单边检验问题的拒绝域在表6-4中列出.

表6-4

上述检验法称为两样本t检验法[3].

【例**6-5**】　用两种药物A和B治疗偏头痛. 医生希望知道使用药物B解除头痛所需的时间（以min计）是否短于使用药物A. 下面是测得药物A、药物B解除头痛所需的时间：

设样本1、样本2分别来自总体X，Y，X~，Y~，μ_1 ，μ_2 ，均未知，两样本独立，试取显著性水平α=0.05检验假设

$$H_0：\mu_1=\mu_2，H_1：\mu_1>\mu_2$$

解　对于样本1，有n_1 =17，，s_1 =6.71，对于样本2，有n_2 =15，，s_2 =2.56.

故接受H_0，认为药物B解除头痛所需时间不比药物A所需时间短.

图6-3分别画出了样本1、样本2的箱线图.

图6-3

6.2.3　基于成对数据的假设检验

有时为了比较两种产品，或两种仪器、两种方法等的差异，我们在相同的条件下作对比试验，得到一批成对的观察值，然后分析观察数据作出推断，这种方法常称为逐对比较法 .

【例6-6】 为了增加母鸡的产蛋量，农场管理人员增加了鸡舍的光照时间. 随机选取10只母鸡，分别记录了每只母鸡在增加光照时间之前和之后同样长的时间段内的产蛋个数，得到以下的10对数据：

问能否认为增加了光照时间，母鸡产蛋量有显著的增加（取$\alpha=0.05$）？

解 本题中的数据是成对的，对同一只母鸡测得一对数据. 对应于不同母鸡产蛋量的差异是由各种因素引起的，例如由母鸡的种类、年龄、体重、遗传因素等因素所引起的. 由于各只母鸡产蛋情况有明显的差别，我们不能将上表中的x行（或y行）看成是一个样本值，因此就不能用表6-4中的t检验法作假设检验. 而数据表中同一对的两个数据的差异，则可看成是仅由两种不同光照时间所引起的. 这样，局限于各对中两个数据来比较，就能排除其他因素的影响，从而有可能研究单独由光照时间影响所引起的产蛋量的差异.

一般的，设有n对相互独立的观察结果（X_1 ，Y_1 ），（X_2 ，Y_2 ），…，（X_n ，Y_n ），记$D_1=X_1-Y_1$ ，$D_2=X_2-Y_2$ ，…，$D_n=X_n-Y_n$ ，则D_1 ，D_2 ，…，D_n 相互独立. 又由于D_1 ，D_2 ，…，D_n 是由同一因素引起的，可以认为它们服从同一分布. 今假设，i=1，2，…，n，这就是说D_1 ，D_2 ，…，D_n 构成正态总体的一个样本，其中μ_D ，未知. 我们需要基于这一样本检验假设

1° H_0 ： $\mu_D=0$，H_1 ： $\mu_D>0$.

2° H_0：$\mu_D=0$，H_1：$\mu_D<0$.

3° H_0：$\mu_D=0$，H_1：$\mu_D\neq0$.

分别记D_1，D_2，…，D_n 的样本均值和样本方差的观察值为，由表6-2中关于单个正态总体的均值的t检验，知检验问题1°，2°，3°的拒绝域如表6-5所列.

表6-5

现在来讨论本例的检验问题，先将母鸡在光照时间增加前后产蛋量之差列于数据表的d=x-y行，按题意需检验假设

$$H_0：\mu_D=0，H_1：\mu_D<0.$$

现在，n=10，t_α（9）=$t_{0.05}$（9）=1.8331，即知拒绝域为

接受H_0，认为产蛋量没有增加.

6.3 正态总体方差的假设检验

6.3.1 单个正态总体N（μ，σ^2 ）方差σ^2 的检验

设总体X~N（μ，σ^2 ），μ，σ^2 均为未知，X_1，X_2，…，X_n 是来

自总体X的样本，现在来检验关于方差σ^2 的假设（显著性水平为α）：

考虑到S^2 是σ^2 的无偏估计，又当H_0 为真时

分布χ^2 （n-1）不依赖于任何未知参数，可取作为检验统计量.

当H_0 为真时，S^2 是的无偏估计，S^2 的观察值s^2 应落在的附近，也就是χ^2 的观察值应落在n-1的附近；另一方面，若H_1 为真，χ^2 的观察值倾向于偏离n-1. 因而拒绝域的形式为

$$（\chi^2 < k_1 ）\cup（\chi^2 > k_2 ）（k_1 ，k_2 \text{ 为常数}）.$$

对于给定的显著性水平α，确定k_1 ，k_2 ，使得

图6-4

对于右边假设检验：

取作为检验统计量. 当H_0 为真时，χ^2 的观察值应落在n-1附近；又若H_1 为真时，χ^2 的观察值倾向于大于n-1. 因而拒绝域具有形式

$$\chi^2 \geq k（k\text{为常数}）.$$

对于给定的显著性水平α，确定k，使得

图6-5

类似地，对于左边假设检验（显著性水平为α）：

其拒绝域（见图6-6）为

图6-6

于是得到以下结果（见表6-6）：

表6-6

以上检验称为χ^2 检验 .

【例6-7】 在生产线上随机地取10个电阻器，测得电阻值（以Ω计）如下：

114.2　91.9　107.5　89.1　87.2

87.6　95.8　98.4　94.6　85.4

设电阻器的电阻值总体服从N（μ，σ^2 ），μ，σ^2 均未知. 问：在显著性水平α=0.1下，方差σ^2 与60是否有显著差异.

解　按题意需检验假设：

$$H_0：\sigma^2=60，H_1：\sigma^2\neq 60.$$

这里n-1=9，α/2=0.05，，. 经计算s^2 =87.682，由表6-6可知H_0 的拒绝域为

χ^2 的观察值未落在H_0 的拒绝域内，故接受H_0 ，认为在α=0.1下，方差σ^2 和60无显著差异.

6.3.2　两正态总体方差的比较

设总体，μ_X，μ_Y，均未知，X_1，X_2，…，X_{n_1}和Y_1，Y_2，…，Y_{n_2}是分别来自总体X，Y的样本，两样本独立，样本方差分别为.

我们需要检验假设：

先给出一个有用的统计量的分布. 因

且假设两样本独立，故上述两统计量独立. 由F分布的定义可知

化简上式左端，得

现在来检验假设（显著性水平为α）：

采用统计量

作为检验统计量. 由式6-26可知，当H_0为真时，

确定k，使

得$k=F_{1-\alpha}$（n_1 -1，n_2 -1），拒绝域为（见图6-7）

图6-7

我们将另外两个假设检验问题的拒绝域一并列在表6-7中.

表6-7

以上检验称为F检验 .

【例**6-8**】　用两种饲料分别喂养6只、5只兔子. 过一定时间后测量兔子增加的重量（kg），得到下面的数据：

解　n_1 =6，n_2 =5，α=0.05，$F_{0.025}$（5，4）=9.36，$F_{0.975}$（5，4）=1/$F_{0.025}$（4，5）=1/7.39，经计算得到，，由表6-7得拒绝域为

今，而$F_{0.975}$（5，4）<F<$F_{0.025}$（5，4），F的值未落在拒绝域，故接受H_0，认为两总体方差相同.

6.4　分布拟合检验

以上讨论的问题都是在总体分布的形式已知，而其中含有未知参数的情况下，对未知参数进行假设检验的.　在实际中还常会遇到总体X的分布形式未知的情况.　此时，我们希望检验假设：X服从某一形式的分布. 例如，我们希望去检验假设：X服从正态分布. 下面介绍用来检验关于总体分布的χ^2 拟合检验法 .

设总体X的分布未知，x_1 ，x_2 ，…，x_n 是来自X的样本值. 我们要依据样本值来检验假设：

H_0：总体X的分布函数为F（x）[4]，

H_1：总体X的分布函数不是F（x）.

（这里备择假设H_1 可以不必写出.）

先设H_0 所假设的X的分布函数F（x）不含未知参数. 我们将在H_0 下X可能取值的全体Ω分成k个互不相交的集合A_1 ，A_2 ，…，A_k ，并用A_i 表示事件{X的值落在子集A_i 内}（i=1，2，…，k），以n_i 记样本观察值x_1 ，x_2 ，…，x_n 中落在A_i 的个数，这表示事件A_i 在n次独立试验中发生n_i 次. 于是在n次独立试验中事件A_i 发生的频率为n_i /n. 另一方面，当H_0 为真时，可以根据H_0 中所假设的X的分布函数计算事件A_i 发生的概率，得到p_i =P（A_i ）. 频率n_i /n与概率p_i 会有差异，但当H_0 为真且当试验次数n甚大时，一般来说差异不会太大. 基于这种想法，我们采用形如

的统计量来度量样本与H_0 中所假设的分布的吻合程度，其中c_i （i=1，2，…，k）为给定的常数. 皮尔逊证明，如果选取c_i =n/p_i （i=1，2，…，k），则由式6-27定义的统计量具有下述定理中所述的简单性质. 于是我们就采用

作为检验统计量. 注意到，检验统计量式6-28又可写成

当H_0 中所假设的分布函数F（x）包含未知参数时，需先利用样本求出参数的最大似然估计（在H_0 下），以估计值作为参数值，然后根据H_0 中所假设的分布函数，求出p_i 的估计值. 在式6-28中以代替p_i ，取

作为检验统计量.

定理 若n≥50，则当H_0 为真时，统计量6-28近似地服从χ^2 （k-1）分布；而统计量6-29近似地服从χ^2 （k-r-1）分布，其中r是被估计的未知参数的个数（证略）.

据上讨论，当H_0 为真时，统计量6-28或6-29中的χ^2 不应太大，如χ^2过分大就拒绝H_0，因而拒绝域的形式为

$$\chi^2 \geq G（G为常数）.$$

对于给定的显著性水平α，确定G，使

$$P\{当H_0 为真拒绝H_0 \}=P_{H_0}\{\chi^2 \geq G\}=\alpha.$$

由上述定理可得[5]，于是得拒绝域

即当样本观察值使式6-28或式6-29的χ^2 的值有

则拒绝H_0，反之，则接受H_0 . 这就是χ^2 拟合检验法 .

在使用这一检验法时，n至少等于50（定理条件），也不能太小，以为佳. 如不满足这一要求，则需适当合并A_i 以满足之（见例6-9）.

【例6-9】 以X记某纺织厂一个工作日机器的故障数，下面记录了60个工作日的故障数. 试检验故障数总体X服从泊松分布，取显著性水平α=0.05.

解 按题意需检验假设（α=0.05）：

$$H_0：X\sim\pi（\lambda）.$$

本题中n=60可使用χ^2 拟合检验法. 在假设H_0 中，参数λ未知，这表示要检验的假设是“X的分布属于分布族π（λ）”. 需先求出λ的最大似然估计

（在H_0 下）. 由例5-12可知

在假设H_0　下，即在X服从泊松分布的假设下，X所有可能取的值为Ω={0，1，2，…}. 将Ω分成四部分：

$$A_1=\{X=0\}，A_2=\{X=1\}，A_3=\{X=2\}，A_4=\{X\geq 3\}，$$

所需计算见表6-8（n=60）.

表6-8

表6-8的第4列中，故将A_3 ，A_4 合并成一个事件，合并以后k=3. 现在χ^2 =62.939-60=2.939，r=1，而，故接受H_0 ，认为故障数总体X服从泊松分布，即X～π（λ）.

【例6-10】　　取显著性水平α=0.1，检验例5-4中患某种疾病的21~59岁男子血压（收缩压）总体X服从正态分布.

解　按题意需在α=0.1下检验假设：

$$H_0：X\sim N(\mu，\sigma^2)．$$

本题n=100可以使用χ^2 拟合检验法. 先需估计未知参数μ，σ^2 . 在H_0 下μ，σ^2 的最大似然估计（见例5-15）

在H_0　下Ω为全体实数（-∞，∞），将区间（-∞，∞）分为9个子区间A_1 ，A_2　，…，A_9　，如表6-9中所列，其中A_1　=（-∞，99.5］，A_9 =（169.5，∞），其他的与表5-2中所列的相同. 按公式

现将计算结果列于表6-9中，可见均小于5，因而将A_7 ，A_8 ，A_9 合并，合并后k=7.

表6-9　n=100

χ^2 =105.3678-100=5.3678，而=7.779>5.3678，故接受H_0 ，认为患某种疾病的21~59岁的男子血压（收缩压）总体X服从正态分布.

6.5　列联表的独立性检验

在实际问题中，当抽取了一个容量为n的样本时，常对样本中的每一个元素按其属性分类. 例如，选取了n=180个人，可按各人吸烟的程度分为不吸烟、中度吸烟、严重吸烟三类，又可按各人是否患有高血压病分为患高血压病、不患高血压病两类，这样可将180人按其所属的类别列成如下的一张2×3的二维表格.

这一表格称为2×3列联表 .

一般，设在一总体中抽取容量为n的样本，样本中各元素可按两种指标（或属性）X与Y来分类.　这里将这两种指标可能取值的范围分成r及c部分，以n_{ij}　表示样本中X及Y指标分别属于第i部分及第j部分的元素的个数（i=1，2，…，r；j=1，2，…，c），又令

用如下的表格表示样本中各元素的分类：

表6-10

这一表格称为r×c列联表.

我们感兴趣的是要检验假设

H_0：总体中指标X和Y是相互独立的，

H_1：X和Y不是相互独立的即有关联的.

（H_1 可不必写出）

以p_{ij} 表示总体中任意抽取的一个元素，它的X指标属于第i部分，且Y指标属于第j部分的概率. 又以$p_{i\cdot}$，$p_{\cdot j}$ 分别表示边缘分布概率，即，这样上述检验问题中的H_0 可写成

$$H_0：p_{ij}=p_{i\cdot}\,p_{\cdot j}，i=1，2，\ldots，r，j=1，2，\ldots，c. \tag{6-32}$$

由上节式6-28'知，如果$p_{i\cdot}$，$p_{\cdot j}$ 为已知，则在H_0 成立时，统计量

当n很大时近似地服从自由度为rc-1的χ^2 分布. 我们就能取式6-34作为检验统计量来检验假设6-32. 但在多数情况$p_{i\cdot}$，$p_{\cdot j}$ 是未知的，这就需用样本分别求出在H_0 下$p_{i\cdot}$，$p_{\cdot j}$ 的最大似然估计，，用它们代替式6-34中的$p_{i\cdot}$，$p_{\cdot j}$. 在这里包含r+c个未知参数$p_{i\cdot}$ 和$p_{\cdot j}$（i=1，2，…，r，j=1，2，…，c），但这些参数需满足条件式6-33，所以实际上只包含r+c-2个未知参数. 先写出似然函数

经计算可得$p_{i\cdot}$ 及$p_{\cdot j}$ 的最大似然估计分别为

将它们代入式6-34，得到统计量

当n很大时，在H_0 下，这一统计量近似服从χ^2 分布，其自由度为

$$rc-（r+c-2）-1=（r-1）（c-1）.$$

我们就用这一统计量作为检验统计量，于是得到对于给定的显著性水平α（0<α<1），假设H_0 的拒绝域为

【例6-11】 试在本节开始提到的例子中，取显著性水平α=0.05检验假设

H_0：吸烟的程度与患高血压是相互独立的.

解 为使计算快捷且不易算错，引入记号

下列表格是将原有的数据再行列出，并将算好的E_{ij} 值写在相应位置的括号内.

得统计量

拒绝域为

现在χ^2 =13.89>5.992，故拒绝H_0 ，认为吸烟程度与患高血压不是相互独立的，而是有关联的.

【例6-12】 销售商调查了某种产品的品牌A、品牌B、品牌C在三个地区的销售量（件）如下表所示，试依据表上的信息检验假设

H_0：各品牌的销售量与销售地区是相互独立的（取α=0.05）.

解　算出，i=1，2，3，j=1，2，3，将它们分别写在上述表格相应位置的括弧内，得到

拒绝域为

现在χ^2 =4.14<9.488，故接受H_0 ，认为产品的各品牌的销售量与销售地区是相互独立的.

6.6　假设检验问题的p值法

以上讲的假设检验法称为临界值法. 这一方法只给出“接受原假设H_0 ”或“拒绝原假设H_0 ”的结论，并不能给出很多信息. 下面介绍被称为p值法的检验方法. 我们从例题讲起.

在本章例6-2中，取显著性水平α=0.1，给出临界值c=1.282，而由样本得到的检验统计量的观察值为

故拒绝H_0　.　数2.0089远大于1.282，若取α＝0.05，则给出临界值c=1.645，由于z_0 =2.0089>1.645，因此也能在显著性水平α=0.05下拒绝H_0 . 实际上，由P{Z≥z_0 =2.0089}=1-Φ（2.0089）=1-0.9777=0.0223知道，若取显著性水平α=0.0223，就有P{Z≥2.0089}=α，因此也拒绝H_0 . 但是，若取显著性水平α<0.0223，就有P{Z≥2.0089}>α，则接受H_0 . 据此，概率

是原假设H_0可被拒绝的最小显著性水平. 我们有以下的定义.

定义　假设检验问题的p值（probability value），是由检验统计量的样本观察值得出的原假设可被拒绝的最小显著性水平.

例如，对于表6-2中所列的3个关于正态总体N（μ，σ^2　）均值的检验，检验统计量为，若由样本求得的统计量的观察值为t_0，则对于假设检验问题：

（1）H_0：$\mu=\mu_0$，H_1：$\mu>\mu_0$

p值=$P_{H_0}\{t\geq t_0\}=t_0$右侧尾部面积（图6-8）.

图6-8

（2）H_0：$\mu=\mu_0$，H_1：$\mu<\mu_0$

p值=$P_{H_0}\{t\leq t_0\}=t_0$左侧尾部面积（图6-9）.

图6-9

（3）H_0：$\mu=\mu_0$，H_1：$\mu\neq\mu_0$

p值=$P_{H_0}\{|t|\geq|t_0|\}=2P_{H_0}\{t\geq|t_0|\}=2\times$（$|t_0|$右侧尾部面积）（图6-10）.

图6-10

又如，对于表6-7中所列的3个关于两正态总体与方差比较的F检验，检验统计量为，若由样本求得的统计量的观察值为F_0，则对于假设

检验问题：

p值=$P_{H_0}\{F\geq F_0\}=F_0$ 右侧尾部面积（图6-11）.

图6-11

p值=$P_{H_0}\{F\leq F_0\}=F_0$ 左侧尾部面积（图6-12）.

图6-12

若F_0 落在F分布的右侧尾部（图6-13a），则有

p值=$2P_{H_0}\{F\geq F_0\}=2\times\min\left[P_{H_0}\{F\geq F_0\},\ P_{H_0}\{F\leq F_0\}\right]$.

若F_0 落在F分布的左侧尾部（图6-13b），则有

p值=$2P_{H_0}\{F\leq F_0\}=2\times\min\left[P_{H_0}\{F\geq F_0\},\ P_{H_0}\{F\leq F_0\}\right]$.

图6-13

综上两种情况，都有

$$\text{p值}=2\times\min\left[P\{F\geq F_0\},\ P\{F\leq F_0\}\right]$$

$$=2\times\min\left[P\{F\geq F_0\},\ 1-P\{F\geq F_0\}\right].$$

（6-36）

对于一个假设检验问题，我们求出p值，那么对于任意的显著性水平（$0<\alpha<1$），按p值的定义知道

若$p\le\alpha$，则拒绝H_0；若$p>\alpha$，则接受H_0.

这种利用p值来确定是否拒绝原假设H_0 的方法称为假设检验问题的**p**值法.

【例**6-13**】 试用p值法检验本章6.2节例6-3的检验问题：

$$H_0：\mu=\mu_0=980，H_1：\mu<980.$$

解 本题是单边检验. 用t检验法检验统计量的观察值为

用计算机算得

$$p值=P_{H_0}\{t\le-2.828\}=P_{H_0}\{t\ge2.828\}=0.0058，$$

故若取显著性水平$\alpha\ge0.0058$，则拒绝H_0；若取$\alpha<0.0058$，则接受H_0. 例如，取$\alpha=0.005$，则接受H_0；若取$\alpha=0.05$，则拒绝H_0.

【例**6-14**】 试用p值法检验本章例6-8中的假设检验问题：

解 本题是双边检验. 用F检验法，检验统计量的观察值为

$$F_0=0.00408/0.01173=0.347826;$$

用计算机算得

$$P\{F\ge F_0\}=0.861372，得$$

$$p值=2［1-P\{F\ge F_0\}］=2［1-0.861372］=0.277256.$$

故若取$\alpha \geq 0.277256$，则拒绝H_0 ；若取$\alpha < 0.277256$，则接受H_0 ，例如取$\alpha = 0.05$，则接受H_0 .

当检验完成时，检验者同时报告所使用的检验统计量及p值，让决策人自己根据客观实际判断，来确定显著性水平α，将α与检验者所报告的p值比较，若$p \leq \alpha$，就作出决策，拒绝H_0 ；否则就接受H_0 .

p值是当H_0 为真时由检验统计量的观察值得出的尾部的概率. p值越小，表示在H_0 为真时出现这一观察值的可能性越小，因而反对H_0 的依据越充分（例如，若$p = 0.0001$，p是如此地小，以至在H_0 为真时几乎不可能出现这一观察值，这说明反对H_0 的理由很充分）. p值是衡量反对H_0 的依据的强度的尺度.

在杂志、案例或技术报告中，当叙述假设检验的结果时，常不论及显著性水平或临界值，代之以引用假设检验问题的p值，让读者基于p值来评价反对H_0 的依据的强度，从而作出决策，是否拒绝假设H_0 . 在作出决策时，以下的原则可以使用（但不是硬性的）.

若$p \leq 0.01$，拒绝H_0 ，反对H_0 的依据是很强的.

若$0.01 < p \leq 0.05$，拒绝H_0 ，反对H_0 的依据是强的.

若$0.05 < p \leq 0.10$，一般不拒绝H_0 . 若要拒绝H_0 ，应考虑犯第Ⅰ类错误的后果.

若$p > 0.10$，不拒绝H_0 ，没有依据拒绝H_0 .

【例**6-15**】　卫生部门规定，在一水体中，每cm^3 平均有70个某种细菌是最大可接受的水准. 如果越过，食用该水体中的蛤类会致病. 今测得一样本，含细菌的平均数为

69　74　75　70　72　73　71　73　68

设样本来自正态总体N（μ，σ^2 ），μ，σ^2 未知，试检验假设

$$H_0：\mu=\mu_0=70，H_1：\mu>70.$$

解　使用t检验法，检验统计量的观察值为

由计算机算得

$$p值=P_{H_0}\{t\geq 2.1320\}=0.0327.$$

由于$p<0.05$，由上一段所说，就拒绝H_0 ，认为$\mu>70$.

本书在解题时，若题中未写明当p小于何值时拒绝H_0　　，就按“当$p\leq 0.05$时拒绝H_0 ”来作决策.

6.7　参数的区间估计

上一章讨论的未知参数的点估计，只给出未知参数的一个估计值，不管所得到的点估计量具有多么好的性质，例如它具有无偏性、一致性，但是并未给出估计值与参数真值的接近程度，亦即没有给出估计的精度. 例如，我们用作为总体均值μ的估计，如果有了一个样本，得到的

观察值，但不能说μ的真值在区间（3.1-0.5，3.1+0.5）或（3.1-1，3.1+1）之内. 我们希望能够得到一个区间，这个区间包含参数真值的可信程度是知道的. 这就是区间估计所要讨论的问题.

【例6-16】 某种疾病患者的存活时间X（自确诊到死亡的时间，以月计）是一个随机变量，已知X~N（μ，σ^2 ），其中σ^2 =9，而μ未知. 设X_1 ，X_2 ，…，X_n 是来自总体X的样本，为了估计μ，我们考虑使用统计量. 将标准化得到

上式右端的分布N（0，1）不依赖于任何未知参数，按标准正态分布的上α分位点的定义，有（见图6-14）

图6-14

查表得$z_{0.05/2}$ =1.96，故上式可改写成

其中，是两个统计量. 这样，我们就得到了一个两端点都是随机变量的随机区间：

由式6-37可知这个随机区间包含μ的真值的概率为0.95. 式6-38表示的区间称为μ的置信水平为0.95的置信区间.

现在抽查16个这种疾病的患者，得到以下的数据（存活时间以月计）：

8.0 13.6 13.2 13.6 12.5 14.2 14.9 14.5

13.4 8.6 11.5 16.0 14.2 19.0 17.9 17.0

由这些数据得到. 又n=16，σ^2 =9，代入区间6-38的两端点，得到一

个数字区间：

（12.41，15.35）.

这个数字区间仍称为μ的置信水平为0.95的置信区间.

现在来阐明置信区间的含义. 在总体中抽得一个样本容量为n的样本值，将它代入区间6-38的两端点就得到一个数字区间. 若在总体中抽样多次（设每次样本容量均为n），就得到许多数字区间，它们的长度都是（是固定的），中心是（是随机的），故这些数字区间的位置各不相同（见图6-15）. 这些数字区间中，有一些包含参数μ的真值，而另一些则不包含. 由式6-37及伯努利大数定理可知在这么多的数字区间中，包含μ的真值的约占95%，而不包含μ的真值的仅约占5%. 例如抽样1000次，则得到1000个数字区间，其中包含μ的真值的约为950个，而不包含μ的真值的仅约50个.

图6-15

在实际中，一般我们只有一个样本值，因而只能得到一个数字区间. 例如上述的（12.41，15.35），它是否包含真值μ，我们不得而知（也不可能知道），但基于上述对区间6-38的解释，我们说“区间（12.41，15.35）包含μ的真值”这一陈述的可信程度为95%.

定义 设总体X的分布含有一个未知参数θ，θ∈Θ（Θ是θ可能取值的范围），X_1 ，X_2 ，…，X_n 是来自X的样本，对于给定值α（0<α<1），若两个统计量和，对于任意θ∈Θ满足

则称随机区间为θ的置信水平为1-α的置信区间 ，或置信水平为1-α的区间估计 . 分别称为置信区间的置信下限 和置信上限 ，1-α称为置信水平

.

对于样本X_1，X_2，…，X_n 的一个样本值x_1，x_2，…，x_n，相应地，置信区间有一个数字区间：，这一数字区间也称为θ的置信水平为1-α的置信区间，或置信水平为1-α的区间估计.

在例6-16中，以1-α代替0.95就得到μ的置信水平为1-α的置信区间：

评价一个置信区间的好坏有两个要素，一是其精度，这可以用区间的长度L来刻画，长度愈大，精度愈低；二是置信水平1-α，即“区间包含μ”这一陈述的可信程度. 由式6-41可看到：在n固定时，当置信水平1-α变大时，α变小，从而$z_{\alpha/2}$　变大，此时置信区间的长度L变大.　这就是说，置信区间的置信水平愈高，则精度愈低；反之精度愈高（即L愈小），则置信水平愈低.

一般，我们取置信水平1-α为0.90，0.95，0.99.

另外，由式6-41还可看到：在α固定时，区间长度L随n的增大而减小，从而，我们可以确定样本容量n，使置信区间的长度具有预先给定的长度. 若希望长度小，n就要大.

参考例6-16可得到一个寻找置信区间的一般方法，现叙述如下：

（1）设θ是待估参数，找一个统计量U，一般U是θ的点估计量（如在例6-16中），构造一个U和θ的函数h（U，θ），它的分布与θ以及其他未知参数无关. 函数h（U，θ）称为枢轴量 ［如例6-16中的h（U，μ）=］. 同时h（U，θ）的分布为已知.

（2）对于给定的置信水平1-α，选择常数a，b，使其满足

$$P\{a<h（U，\theta）<b\}=1-\alpha$$

（6-43）

（如在例6-16中，$a=-z_{\alpha/2}$ ，$b=z_{\alpha/2}$ ），若不等式$a<h（U，\theta）<b$可改写成，这样就可以得到与式6-43等价的概率等式：

则是θ的一个置信水平为1-α的置信区间.

本节我们只讨论正态总体参数的置信区间.

1）正态总体**N（μ，σ^2 ）**均值**μ**的置信区间

设X_1 ，X_2 ，…，X_n 为总体N（μ，σ^2 ）的样本，，S^2 分别是样本均值和样本方差，又给定置信水平为1-α.

（1）σ^2 已知时，已在例6-16中得到μ的置信水平为1-α的置信区间：

（2）σ^2 未知时，取枢轴量

得μ的置信水平为1-α的置信区间：

2）两正态总体均值差的置信区间均未知）

设X_1 ，X_2 ，…，X_{n_1} 和Y_1 ，Y_2 ，…，Y_{n_2} 是分别来自总体X~N（μ_X ，的样本，两样本独立. 设分别是总体X，Y的样本均值；分别是总体X，Y的样本方差. 给定置信水平为1-α.

取枢轴量

其中

（见式6-22，式6-23）得$\mu_X - \mu_Y$ 的一个置信水平为1-α的近似置信区间为

3）正态总体**N**（**μ**，$\boldsymbol{\sigma^2}$ ）方差$\boldsymbol{\sigma^2}$ 的置信区间（**μ**未知）

设X_1 ，X_2 ，…，X_n 为总体N（μ，σ^2 ）的样本，S^2 是样本方差. 给定置信水平为1-α.

取枢轴量

4）两正态总体方差比的置信区间（$\boldsymbol{\mu_X}$ ，$\boldsymbol{\mu_Y}$ 未知）

设X_1 ，X_2 ，…，X_{n_1} 和Y_1 ，Y_2 ，…，Y_{n_2} 分别是来自总体X~N（μ_X ，的样本，两样本独立，分别是总体X，Y的样本方差. 给定置信水平为1-α.

取枢轴量

我们将所得结果汇集于表6-11.

表6-11 （置信水平为1-α）

注：表中所列的各个枢轴量与相应的假设检验问题中的检验统计量（当H_0 为真时）有相同的分布.

【例**6-17**】 在某一计算机终端上调试程序，其响应时间X（以s

计）具有正态分布N（μ，σ^2），μ，σ^2 未知. 今测得X的样本值如下：

1.48　1.26　1.52　1.56　1.48　1.46　1.30　1.53

1.28　1.43　1.43　1.55　1.57　1.51　1.53　1.74

1.68　1.37　1.47　1.61　1.44　1.43　1.64　1.51

1.51　1.60　1.65　1.60　1.64　1.51

试求：（1）μ的置信水平为0.95的置信区间；（2）求σ^2　的置信水平为0.95的置信区间.

解　　（1）现在n=30，置信水平1-α=0.95，α=0.05，查表得$t_{0.025}$（29）=2.0452，经计算，s=0.1143.　由表6-11中第2栏得所求μ的置信区间为

即

（1.4670，1.5524）.

（2）现在n=30，α/2=0.025，1-α/2=0.975，查表得，s^2 =0.0131，由表6-11第3栏，得所求σ^2 的置信区间为

即

（0.0083，0.0237）.

【例6-18】　　以X_1 ，X_2 分别记棕色鹈鹕和蛎鹬这两种鸟在横向风下的飞行速度（以英里/小时计），随机地取9只棕色鹈鹕和12只蛎鹬，

测得以下数据：

设X_1 ，X_2 分别服从正态分布均未知，两样本独立，求μ_1 -μ_2 的置信水平为0.90的近似置信区间.

解　由表6-11的第4栏知

所求的近似置信区间为

【例6-19】　为比较两个煤矿所产煤的质量，测得以下的发热量（以4.1868×10^6 J/吨计）：

煤矿A　8500，8330，8480，7960，8030

煤矿B　7710，7890，7920，8270，7860

设样本依次来自总体均未知，且两样本独立.　试求方差比的置信水平为0.90的置信区间.

解　经计算得

现在1-α=0.90，α=0.10，n_1 =n_2 =5，查表得，$F_{0.05}$ （4，4）=6.39，$F_{0.95}$（4，4）=1/6.39，由表6-11第5栏得所求的置信区间为

习题6

【答案链接】

1. 自一批钢管抽取10根，测得其内径（mm）如下：

100.36　100.31　99.99　100.11　100.64

100.85　99.42　99.91　99.35　100.51

设这批钢管内直径服从正态分布N（μ，σ^2　　），试在显著性水平α=0.05下检验假设H_0：μ=100，H_1：μ≠100. 若（1）σ=0.5；（2）σ为未知.

2.　某一化工生产过程的日产量（吨）近似服从正态分布N（μ，σ^2），μ，σ^2 均未知，在正常的情况下日平均产量μ=810. 今测得上周日产量如下：

785　805　790　793　802

试在显著性水平α=0.05下检验假设H_0：μ=810，H_1：μ<810.

3.　为考察一鱼塘中某种鱼的含汞量（mg/kg），随机地取10条鱼，测得各条鱼的含汞量如下：

0.8　1.6　0.9　0.8　1.2　0.4　0.7　1.0　1.2　1.1

设鱼的含汞量服从正态分布N（μ，σ^2 ），μ，σ^2 均未知. 试检验假设H_0：μ=1.2，H_1：μ≠1.2（取显著性水平α=0.10）.

4. 研究某种合金的铜含量，取6个试块测得铜含量（%）如下：

8.031　9.994　9.920　7.745　11.652　14.640

设铜含量服从正态分布N（μ，σ^2 ），μ，σ^2 均未知. 问：在显著性

水平α=0.01下能否认为铜含量的均值显著地大于9.5？

5. 设样本1来自总体N（μ_1，9），样本容量n_1 =18，样本均值. 样本2来自总体N（μ_2 ，12），样本容量n_2 =16，样本均值. 试取显著性水平α=0.05检验假设

$$H_0：\ \mu_1 =\mu_2，\ H_1：\ \mu_1 >\mu_2 .$$

6. 一工厂的两车间A和B生产同一型号的器件，为比较两车间的日产量（件），分别测得一周内各天的产量如下：

设数据依次来自正态总体均未知，两样本相互独立，试检验假设H_0：$\mu_1 =\mu_2$，H_1：$\mu_1 \neq\mu_2$（显著性水平α=0.05）.

7. 用两种不同的方法自原油中提炼汽油，下面给出使用不同方法的得油率：

设这两批数据分别来自总体均未知，两样本独立. 试检验假设（取显著性水平α=0.05）：

$$H_0：\ \mu_X =\mu_Y，\ H_1：\ \mu_X \neq\mu_Y .$$

8. 用一种叫“混乱指标”的尺度去衡量工程师的英语文章的可理解性，对混乱指标的打分越低表示可理解性越高. 分别随机选取13篇刊载在工程杂志上的论文，以及10篇未出版的学术报告，对它们的打分列于下表：

设数据Ⅰ，Ⅱ分别来自正态总体均未知，两样本独立. 试检验假设（取显著性水平α=0.05）：

$$H_0：\ \mu_1 =\mu_2，\ H_1：\ \mu_1 <\mu_2 .$$

9. 以X，Y分别表示有过滤嘴的香烟和无过滤嘴的香烟的煤焦油含量（以毫克/支计），分别自总体X和Y抽得两样本如下：

设，，均未知，两样本相互独立，试取α=0.05，检验假设

$$H_0：\ \mu_1 =\mu_2，\ H_1：\ \mu_1 <\mu_2，$$

并在一数轴上分别画出两组数据的箱线图.

10. 两种汽油，汽油Ⅰ掺入了添加剂，汽油Ⅱ未掺入，考察同一汽车使用汽油Ⅰ，Ⅱ所能行驶的里程（以英里/加仑计），测得以下的数据：

设各对数据之差$D_i =X_i -Y_i$（i=1，2，…，10）是来自总体的样本，但参数未知，试以显著性水平α=0.05检验假设

$$H_0：\ \mu_D =0，\ H_1：\ \mu_D >0.$$

11. 对9个病人分别用口腔表和肛门表测量体温（以℃计），结果如下：

设各对数据之差$D_i =X_i -Y_i$（i=1，2，…，9）是来自总体的样本，但参数未知，试以显著性水平α=0.05检验假设

$$H_0：\ \mu_D =0，\ H_1：\ \mu_D <0.$$

12. 为考察患青光眼病是否会引起眼睛角膜厚度改变，抽取了8名有

一只眼睛患青光眼，另一只眼睛不患青光眼的病人，测得下列角膜厚度数据（以微米计）：

设各对数据之差$D_i = X_i - Y_i$（i=1，2，…，8）是来自总体的样本，但参数未知，试以显著性水平α=0.05检验假设

$$H_0：\mu_D = 0，H_1：\mu_D \neq 0.$$

13. 为了考察一含有某种矿物质的食品对于人体胆固醇指标（以毫克/分升计）的影响，随机地选择了6个人，他们食用一定分量的这种食品，历时6周，比较各人在食用之前和之后的胆固醇指标，得到以下的数据.

设各对数据之差$D_i = X_i - Y_i$（i=1，2，…，6）是来自总体的样本，但参数未知，试以显著性水平α=0.05检验假设

$$H_0：\mu_D = 0，H_1：\mu_D \neq 0.$$

14. 某生产厂家声称他们生产的一批部件的直径的方差不大于$0.0002cm^2$. 今从中随机抽取一个容量为10的样本，得到样本方差$s^2 = 0.0003cm^2$.

设这批部件直径服从正态分布N（μ，σ^2 ），μ，σ^2 均未知. 试检验假设$H_0：\sigma^2 = 0.0002$，$H_1：\sigma^2 > 0.0002$（显著性水平α=0.05）.

15. 在一批温度计中，随机地取16个，对它们进行测试，得到16个读数，其样本标准差s=0.3℃. 设温度计的读数服从正态分布N（μ，σ^2

），μ，σ^2 均未知，试检验假设（取显著性水平α=0.05）

$$H_0：\sigma^2=0.5^2，H_1：\sigma^2<0.5^2.$$

16. 在第4题中取显著性水平α=0.05检验假设

$$H_0：\sigma^2=5.5，H_1：\sigma^2\neq5.5.$$

17. 下面分别给出某地区不同年龄组男子的血压（收缩压）的数据（以mmHg计）. A组为21~44岁，B组为45~59岁. 设两样本依次来自总体，两样本独立，均未知. 试检验假设

18. 两批零件分别由机器A和机器B所生产，在这两批零件中分别取样，测得机器A生产的零件长度的样本方差，样本容量为12；机器B生产的零件长度的样本方差为，样本容量为10. 设两样本独立，且分别来自正态总体均未知. 试在显著性水平α=0.1下检验.

19. 随机取A，B两个班学生的数理统计课程期终考试的得分，得到：

设A，B两班学生的考试得分分别服从正态分布均未知，又设两样本独立. 试检验假设

20. 医师希望知道，吸烟者与不吸烟者的心率（次/分）的方差是否有差别，测得以下数据：

两样本分别来自总体均未知，两样本独立，试取显著性水平α=0.05检验假设

21. 下面给出了某个城市在某一时间段内不同颜色小汽车发生交通事故的次数：

试检验假设H_0 ：各种颜色小汽车发生交通事故是等可能的（取显著性水平α=0.05）.

22. 对于习题5第5题的数据检验假设H_0 ：数据来自正态分布N（μ，σ^2 ），取显著性水平α=0.05.

提示：先将整个实数轴分为（-∞，64.5］，（64.5，69.5］，（69.5，74.5］，…，（84.5，89.5］，（89.5，∞）.（注：利用χ^2 拟合检验进行检验时，要求样本容量n≥50；本题系作练习之用，n只有40.）

23. 测得300只电子元件的寿命（以小时计）如下

寿命	只数
0<t≤100	121
100<t≤200	78
200<t≤300	43
t>300	58

试取显著性水平d=0.05检验假设：H_0 ：寿命服从指数分布，其密度为

24. 在一副扑克牌中任抽3张，记录3张牌中含黑桃的张数后放回，然后再任意抽3张，再记录其中黑桃张数，如此重复64次，得到以下数据：

试取显著性水平α=0.01，检验假设.

25. 下面记录了某个城市1979年各天报火警的次数：

试检验假设H_0　　：一天报火警次数服从泊松分布（取显著性水平α=0.01）.

26.　有一项研究课题，研究人的左、右足大小的相异性，得到以下的数据（人数）（L，R分别代表左、右足大小）：

试检验假设H_0　　：人的左右足大小的相异性与人的性别是相互独立的（取显著性水平为0.05）.

27.　　为考察维生素C是否有防冷的作用，随机地选择了279名滑雪者，一部分人服用维生素C，另一部分人服用安慰剂，得到以下的试验结果（人数）：

试检验假设H_0：服用维生素C与防冷相互独立（取显著性水平α=0.05）.

28. 为了改进某种部件的结构设计，对于这种部件的3种不同的结构形式，分别统计了它们的4种损坏模式，得到下列数据（部件数）：

试取显著性水平α=0.05检验假设H_0：结构形式与损坏模式是相互独立的.

29. 某种钢瓶的内压力强度近似地服从正态分布N（μ，σ^2），其中σ=0.2（标准大气压）. 随机地取25个瓶子，测得内压强的样本均值（标准大气压），求μ的置信水平为0.95的置信区间.

30. 进入某一商场的顾客用于购物的时间X（以min计）服从正态分布N（μ，σ^2），μ，σ^2 未知，随机地选取64位顾客，他们购物时间的样本均值为64，样本标准差为16，试求均值μ的置信水平为0.90的置信区间.

31. 在一机器上加工某种零件，随机地取15个，测得其直径（mm）如下：

8.24　8.23　8.20　8.21　8.20　8.28　8.23　8.26

8.24　8.25　8.19　8.25　8.26　8.23　8.24

设该零件直径服从正态分布，参数未知.　求直径均值的置信水平为0.95的置信区间.

32.　为比较两种品牌（品牌1，品牌2）电话蓄电池的寿命，对每个品牌随机地各取容量为15的样本，依次测得各个样本的样本均值和样本标准差如下：

设两个样本分别来自正态总体均未知，试求$\mu_1 - \mu_2$ 的置信水平为0.95的近似置信区间.

33. 考察两种品牌滤水器（A和B）的性能，比较它们能减少水中杂质的分量（以ppm计），经测试得到以下数据：

设两样本依次来自总体均未知，两样本独立，求$\mu_1 - \mu_2$ 的置信水平为0.95的近似置信区间.

34. 在第2题中，求σ的置信水平为0.90的置信区间.

35. 在第3题中，求σ^2 的置信水平为0.90的置信区间.

36. 在18题中，求方差比的置信水平为0.95的置信区间.

---

[1] 在实际问题中，当假设H_0 被拒绝会造成严重后果时，应将α取小一些.

[2] 是指随机向量（X_1 ，X_2 ，…，X_{n_1} ）和（Y_1 ，Y_2 ，…，Y_{n_2} ）相互独立.

[3] 在以上的讨论中，若设两正态总体的方差相等，即设但未知，又设两样本独立，可得检验问题H_0 ：$\mu_X = \mu_Y$ ，H_1 ：$\mu_X \neq \mu_Y$ 的拒绝域为

其中. 这一检验法称为合并方差检验法. 这种检验法对于违反条件很是敏感，而核实条件是困难的（有一种作法，是使用6.3节中的F检验法作的预检验. 由于F检验对于两总体的正态性的违背非常敏感，因而这种做法不可取）. 容易产生以下的情况，实际上而误认为，以致使用了合并方差t检验法，从而导致错误的结果. 因此不推荐使用合并方差t检验法.

[4] 当X为离散型时常以分布律代替F（x），当X为连续型时以概率密度代替F（x）.

[5] 当H_0 中所假设的分布函数不含未知参数时，r=0.

7　回归分析与方差分析

本章我们将利用前两章学过的参数估计和假设检验的知识来研究数理统计中具有广泛实际应用的两个内容——回归分析和方差分析.

7.1　一元线性回归

在实际问题中，往往有这样的情况：有两个或多个变量，它们存在着内在的关系，其中有一个因变量Y在试验中是无法控制的，是一个随机变量；因变量Y依赖于一个自变量x，或多个自变量x_1 ，x_2 ，…，x_t ，这些自变量在试验中是可以测量的（测量误差可忽略不计），是可以由试验者控制的，它们可以被认为是普通变量（不是随机变量）. 回归分析是一种统计方法，可用来研究因变量Y与自变量之间的关系. Y与自变量的关系可以由称为回归方程的方程来表征. 利用回归方程能够解决人们感兴趣的依据自变量的取值预测因变量的取值的问题. 例如，我们可以由服药的剂量（x）预测血压的改变量（Y）；由一个家庭的收入（x）预测这个家庭花在娱乐方面的费用（Y）；可以由建筑物的高度（x）预测作用在建筑物上的风负荷（Y）；可以由新产品的定价（x_1 ）和广告费用（x_2 ）预测新产品的销售量（Y），等等.

下面来研究最简单的情况，即一元线性回归.

设随机变量Y（因变量）与普通变量x（自变量）之间存在着内在

的关系，由Y的随机性可知，对于x的各个确定的值，Y的取值随试验的结果而定，Y的取值有一定的概率分布（如图7-1中曲线C_1，C_2分别为x_1，x_2处Y的概率密度）. 例如，为考察儿童的身高（Y）与其年龄（x）的关系，对一群同一年龄的儿童逐一地去测量他们的身高，他们的身高各不相同，从而形成一定的概率分布. 用F（y|x）表示当x取确定值时所对应的Y的分布函数. 如果我们掌握了F（y|x）随着x的取值而变化的规律，那么就能完全掌握Y与x之间的关系了，但是这样做往往比较复杂. 作为一种近似，可转而去考察分布F（y|x）的数学期望. Y的数学期望［即F（y|x）的数学期望］随x的取值而定，它是x的函数. 将这一函数记为$\mu_{Y|x}$或μ（x），称为Y关于x的回归函数 .

图7-1

若η是一个随机变量，则E［（η-c）^{2} ］作为c的函数，在c=E（η）时E［（η-c）^{2} ］达到最小（见习题3第15题）. 这表明在一切x的函数中以回归函数μ（x）作为Y的近似，其均方误差E［（Y-μ（x））^{2} ］为最小. 因此，作为一种近似，为了研究Y与x的关系转而去研究μ（x）与x的关系是合适的. 在实际问题中，回归函数μ（x）一般是未知的，回归分析的任务是在于根据试验数据去估计回归函数，讨论有关的点估计、区间估计、假设检验以及对Y作预测等问题.

我们对于x取定一组不全相等的值x_1，x_2，…，x_n，设Y_1，Y_2，…，Y_n分别是在x_1，x_2，…，x_n处对Y的独立观察的结果. 这里Y_1，Y_2，…，Y_n是相互独立的，但一般它们未必同分布，为方便计，我们也称（x_1，Y_1），（x_2，Y_2），…，（x_n，Y_n）是一个样本.

我们要利用样本来估计回归函数，先要推测回归函数μ（x）的形式. 有时，根据所讨论的问题的专业知识能确定μ（x）的形式，更多的情况下，则要根据实测数据去推测μ（x）的形式. 其做法是将样本观察值（x_i ，y_i ）（i=1，2，…，n）在直角坐标系中描出相应的点，作出散点图 （见图7-2）. 从散点图上可大致看出μ（x）的形式. 如在图7-2a中这些点大致落在一条直线的附近，看起来μ（x）取线性函数α+bx的形式是适当的. 而在图7-2b中，μ（x）取x的线性形式不合适，取二次函数 $a+bx+cx^2$ 可能恰当一些.

图7−2

这一节我们只考虑回归函数μ（x）是x的线性函数的情况. 现在假设Y的回归函数具有形式

$$\mu(x)=\beta_0+\beta_1 x（\beta_0，\beta_1 为常数），$$

进一步还假设对于x（在某个区间）的每一个值，有

$$Y\sim N(\beta_0+\beta_1 x，\sigma^2），$$

其中参数β_0，β_1，σ^2 （$\sigma>0$）均为未知，且它们都不依赖x. 我们记$\varepsilon=Y-(\beta_0+\beta_1 x)$，即有$\varepsilon\sim N(0，\sigma^2)$，亦即假设

$$Y=\beta_0+\beta_1 x+\varepsilon，\varepsilon\sim N(0，\sigma^2），$$

（7-1）

其中参数β_0，β_1，σ^2 （$\sigma>0$）都不依赖x.

这里，β_1 称为回归系数 .

式7-1表明，试验结果Y可以看成是由两部分叠加而成，一部分是由x的线性函数$\beta_0 +\beta_1 x$所引起的，另一部分$\varepsilon \sim N(0, \sigma^2)$是随机误差，是人们不可控制的.

设有一样本(x_1, Y_1)，(x_2, Y_2)，…，(x_n, Y_n)，由式7-1有

$$Y_i =\beta_0 +\beta_1 x_i +\varepsilon_i \quad (i=1, 2, \ldots, n), \tag{7-2}$$

其中$\varepsilon_i \sim N(0, \sigma^2)$（i=1，2，…，n），且由$Y_1$，$Y_2$，…，$Y_n$ 的相互独立性知ε_1，ε_2，…，ε_n 相互独立.

式7-1和式7-2称为一元线性回归模型 . 为了以后讨论问题方便计，我们将式7-2改写成如下的形式[1]：

其中

现在来估计参数β'_0，β_1 . 设已经有了样本值(x_1, y_1)，(x_2, y_2)，…，(x_n, y_n)，将它们画在坐标平面上得到一个散点图（见图7-3）. 假设我们分别用b_0，b_1 去估计β'_0 和β_1，那么，对于给定的x，自然取作为回归函数的估计. 我们希望直线

尽可能“拟合”n个点(x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) . 在x_i

处作观察值y_i与直线（式7-4）的纵坐标的差：

它表示点（x_i ，y_i ）与直线（式7-4）的垂直偏差. 作各偏差的平方和（见图7-3）：

图7-3

我们以Q的大小衡量直线（式7-4）与点（x_1 ，y_1 ），（x_2 ，y_2 ），…，（x_n，y_n ）的拟合程度，Q越小表示拟合得越好. 设当时Q取到最小值，这表示直线

与n个点（x_1，y_1 ），（x_2，y_2 ），…，（x_n，y_n ）拟合得最好. 我们就取，分别作为β'_0，β_1 的估计，而以作为回归函数的估计.

这里的估计是使平方和取到最小值而得到的，这种估计参数的方法叫做最小二乘估计法 . 分别称为β'_0 ，β_1 的最小二乘估计 . 最小二乘估计法是应用数学中的一个重要方法.

由二元函数求极值的方法，知道可由解方程组

得到，化简后得到

式7-7称为正规方程组 . 由于x_1，x_2，…，x_n 不全相同，故0，正规方程组有解. 求解正规方程组，就得到β'_0，β_1 的估计分别为 [2]

有了参数β'_0，β_1 的估计，就可得到回归函数μ（x）=β'_0 +的估计：

式7-9称为Y关于x的一元经验线性回归方程 ，简称一元线性回归方程

，其图形称为回归直线 .

仍将回归函数写成μ（x）=β_0 +β_1 x，由式7-3得β_0 的估计

而一元线性回归方程（式7-9）又可写成

为计算和书写的方便，引入下述记号：

这样，β'_0 ，β_1 的估计式7-8可写成

【例7-1】 为研究一游泳池池水经化学处理后，水中氯气的残留量Y（μg/mL）与经历的时间x（自处理结束时算起，以h计）的关系，测得以下数据：

试画出散点图，并求出Y关于x的回归直线方程.

解 画出散点图如图7-4所示，从中大致看出回归函数μ（x）具有线性函数的形式. 为了求得回归直线方程，将所需计算列入表7-1.

图7-4

表7-1 （n=6）

由此得

=42/6=7，=7.8/6=1.3，

S_{xy} =48.6-（42×7.8）/6=-6，

S_{xx} =364-42^2 /6=70，

S_{yy} =10.68-7.8^2 /6=0.54 [3] ，

故有，=-6/70=-0.085714，于是得回归直线方程为

7.2　一元线性回归的统计分析

现在将逐个讨论有关一元线性回归模型中几个主要的统计量的统计分析问题（本节中涉及的有关条件和记号与上一节相同）.

7.2.1　最小二乘估计的一些性质

由式7-8可知的最小二乘估计量分别为 [4]

在式7-14中记

则式7-14可写成

这表示都是Y_1 ，Y_2 ，…，Y_n 的线性组合. 由Y_i 的正态性假设以及Y_1 ，Y_2 ，…，Y_n 相互独立性知道，均为正态随机变量.

由式7-15和式7-2'知道

又由式7-15和式7-2'知道

7.2.2　随机误差ε的方差σ^2 的估计

由式7-1可知

$$E\{[Y-(\beta_0+\beta_1 x)]^2\}=E(\varepsilon^2)=D(\varepsilon)+[E(\varepsilon)]^2=\sigma^2,$$

这表示σ^2 愈小，以回归函数$\mu(x)=\beta_0+\beta_1 x$作为Y的近似导致的均方误差就愈小，这样利用回归函数$\mu(x)$去研究随机变量Y与x的关系就愈有效. 然而σ^2 是未知的，因而需要利用样本来估计σ^2 .

称为残差 ，而平方和

称为残差平方和 .

直接利用式7-18来计算残差平方和是比较麻烦的，为便于计算，将式7-18改写如下：

与SSE相对应的统计量（仍记为SSE）

定理　在模型7-2'中残差平方和具有性质：

（1）$SSE/\sigma^2 \sim \chi^2(n-2)$，

（2）SSE，相互独立.

定理证明略. 由这一定理可知

$$E(SSE/\sigma^2)=n-2，或E(SSE/(n-2))=\sigma^2,$$

这就是说残差平方和SSE除以n-2（记为），即

是σ^2 的无偏估计.

7.2.3　线性假设的显著性检验

如前所述取Y关于x的回归函数μ（x）具有形式$\beta_0 + \beta_1 x$，这往往是一种假定，它是否符合实际，首先要根据有关专业知识和实践来判断，也可根据样本通过假设检验的方法来判断. 我们提出原假设H_0：$\beta_1 = 0$，若检验结果是接受H_0，就表明回归函数μ（x）与x不存在线性关系，此时就不宜使用线性回归模型了.　　为此，就需要利用样本来检验假设问题：

$$H_0：\beta_1 = 0，H_1：\beta_1 \neq 0.$$

下面将使用t检验法来进行检验. 由式7-17及上述定理，有

此处α为显著性水平.

当H_0 被拒绝时，就认为回归效果是显著的. 若接受H_0：$\beta_1 = 0$，则认为回归效果不显著，此时不宜使用线性回归模型，需另行研究.

7.2.4　回归函数μ（x）的置信区间

设x_0 是自变量x的某一指定值.　我们是以经验线性回归方程在$x=x_0$处的值作为回归函数值的点估计的（见式7-9）. 考虑对应的估计量

由于，因而这一点估计是无偏的. 又可知[5]

由此得到μ（x_0 ）的一个置信水平为1-α的置信区间：

这就是在$x=x_0$ 处回归函数值μ（x_0 ）的一个置信区间. μ（x_0 ）的置信区间的长度是x_0 的函数，它随的增加而增加，当时为最短.

7.2.5　新观察值的预测

回归模型的一个重要应用是，在x的指定值x_0　处，对于对应的Y的新观察值（或未来的观察值）Y_0 进行点预测或区间预测.

若我们对$x=x_0$ 时Y的观察值Y_0 感兴趣，然而我们在$x=x_0$ 并未进行观察或者暂时无法观察，则可以按以下的方法对它进行预测.

按式7-2'的假设新观察的结果Y_0 应满足

因Y_0　是将要做的一次独立试验的结果，因此它与已经得到的试验结果Y_1，Y_2，…，Y_n 相互独立.

我们就取x_0 处的回归函数值μ（x_0 ）的点估计

作为新观察值的点估计. 称为Y_0 的点预测值 . 下面介绍区间预测.

可以证明

从而对于给定的置信水平1-α，有

称为新观察值Y_0 的置信水平为1-α的预测区间 [6] .

可以看到预测区间的长度是x_0 的函数，它随的增加而增加，当时为最短. 比较式7-25和式7-29知道对于同一置信水平1-α，Y_0 的预测区间要比μ（x_0 ）的置信区间宽.

【例7-2】 设在例7-1中变量Y，x符合回归模型（式7-2'）.（1）求σ^2 的无偏估计；（2）检验回归效果是否显著（取α=0.05）；（3）如果回归效果是显著的，求回归函数μ（x）在x=5处的置信水平为0.95的置信区间；（4）求在$x=x_0$ 处Y的新观察值Y_0 的置信水平为1-α的预测区间.

解 （1）由式7-19和式7-20可知σ^2 的无偏估计

（2）由式7-22可知

故拒绝H_0：$\beta_1=0$，认为回归效果是显著的.

（3）由式7-23可得在x=5处μ（x）的点估计

这表示在x=5处平均含氯量的点估计为1.47μg/mL. 另一方面它也是在x=5处含氯量的观察值的一个预测值，这表示我们可以预测在x=5处的含氯量为1.47μg/mL.

由式7-25知回归函数μ（x）在x=5处的值的置信水平为0.95的置信区间为

（4）由式7-29可得$x=x_0$ 处Y的新观察值Y_0 的置信水平为0.95的预

测区间为

取x_0 为不同的值，得到在各点处对应的Y新观察值的预测区间（置信水平为0.95）如下：

分别将这些区间的下端点和上端点连接起来，得到两条曲线L_1 和L_2 ，回归直线位于由L_1 ，L_2 所围成的带域的中心线上（见图7-5）.

图7-5

7.3　可转化为一元线性回归的模型举例

以上我们所讨论的只限于Y关于x的回归μ（x）是x的线性函数的情况，实际上还经常会遇到μ（x）是x的非线性函数. 但在某些情况可以通过适当的变量变换将问题转化为一元线性回归来处理.　下面举出几种常见的可转化为一元线性回归的模型.

（1）$Y=\beta_0+\beta_1 \ln x+\varepsilon$，$\varepsilon\sim N(0,\sigma^2)$.

（7-30）

其中β_0 ，β_1 ，σ^2 （$\sigma>0$）是与x无关的未知参数. 此时，可以取新自变量$x'=\ln x$，将它转化为一元线性回归模型：

$$Y=\beta_0+\beta_1 x'+\varepsilon，\varepsilon\sim N(0,\sigma^2).$$

（7-31）

若在原模型下，对（x，Y）有样本（x_1 ，y_1 ），（x_2 ，y_2 ），…，（x_n，y_n ），就相当于在新模型下有样本（x'_1，y_1 ），（x'_2，y_2 ），…，（x'_n，y_n ），其中$x'_i=\ln x_i$，于是就能利用上节中的方法来估计β_0，β_1 或对β_1 作假设检验，或对Y进行预测. 在得到Y关于x'的回归方程后，再将原自变量x代回，就得到Y关于x的回归方程，它的图形是一条曲线，也称为曲线回归方程.

（2）$Y=\alpha x^{\beta}\varepsilon$，$\ln\varepsilon\sim N(0,\sigma^2)$.

（7-32）

其中α，β，σ^2 （α>0，σ>0）是与x无关的未知参数. 将$Y=\alpha x^{\beta}\varepsilon$两边取对数：

$$\ln Y=\ln\alpha+\beta\ln x+\ln\varepsilon,$$

令lnY=Y'，$\ln\alpha=\beta_0$ ，$\beta=\beta_1$ ，lnx=x'，lnε=ε'，即可将模型7-32转化为一元线性回归模型：

$$Y'=\beta_0+\beta_1 x'+\varepsilon',\ \varepsilon'\sim N(0,\sigma^2).$$

（7-33）

（3）$Y=\alpha e^{\beta x}\varepsilon$，$\ln\varepsilon\sim N(0,\sigma^2)$.

（7-34）

其中α，β，σ^2 （α>0，σ>0）是与x无关的未知参数. 将$Y=\alpha e^{\beta x}\varepsilon$两边取对数：

$$\ln Y=\ln\alpha+\beta x+\ln\varepsilon.$$

令lnY=Y'，lnα=β_0 ，β=β_1 ，x=x'，lnε=ε'，模型7-34可转化为一元线性回归模型：

$$Y'=\beta_0+\beta_1 x'+\varepsilon',\ \varepsilon'\sim N(0,\sigma^2).$$

（7-35）

【例7-3】 为了解废气中的含氨（NH_3 ）量，经试验得到以下数据：

已知电压读数x与含氨量Y符合模型$Y=\alpha e^{\beta x}$ ε，lnε~N（0，σ^2 ），其中α，β，σ^2 （α>0，σ>0）与x无关，试建立Y关于x的曲线回归方程.

解 本题符合模型7-34. 经变量变换后转化为模型7-35：

$$Y'=\beta_0+\beta_1 x'+\varepsilon',\ \varepsilon'\sim N(0,\sigma^2).$$

其中Y'=lnY，β_0 =lnα，β_1 =β，x'=x，ε'=lnε，数据经变换后得到

这就是我们所求的曲线回归方程. 有了这一回归方程，当测得电极的电压后就能预测废气中的含氨量. 例如，若测得x=-40mV就可以预测含氨量为

7.4　单因素试验方差分析

实际应用中常会遇到比较多个（两个以上）同方差正态总体均值的问题.　方差分析法就是解决这类问题的一种有效方法，它在实际中有广泛的应用.

我们将在试验中要考察的指标称为试验指标　，那些影响试验指标的可以控制的条件称为试验的因素．为了考察一个因素对试验指标的影响，一般将因素控制在几个不同的状态上，每一个状态称为因素的一个水平．若一项试验中只有一个因素在改变，而其他因素保持不变，称为单因素试验；多于一个因素在改变，称为多因素试验.本节只讨论单因素试验.

【例7-4】　为了比较编号为Ⅰ，Ⅱ，Ⅲ，Ⅳ的4种不同食品对增加人体重量的影响，将这4种食品分别喂养5只、5只、4只、6只老鼠，经过一定时间，测得其增加的体重见表7-2.

表7-2

这里试验指标是老鼠增加的体重.　在试验中，假设除食品品种这一因素在改变以外，其他条件保持不变，因而是单因素试验.　每一种品种是一个水平，共有4个水平.　我们要考察4种不同食品对增加老鼠体重的影响是否有显著差异，如果有显著差异就表明食品品种这一因素对增加老鼠体重的影响是显著的.

在这个例题中，我们在因素的每一水平下进行独立试验，其结果是随机变量.　表中每一横行的数据来自一个总体，表中的数据可看成是来自4个不同的总体，假设其都服从正态分布，且方差相同.这样，上述要

考察的问题可归结为检验这4个总体的均值是否相等.

【例7-5】　为了比较各个工作日进入某一商场的顾客人数，测得各工作日下午4~5时进入商场的顾客人数如表7-3所示.

表7-3

这一试验的试验指标是顾客人数. 在这里只有工作日这一因素在改变，是单因素试验. 一个工作日是一个水平，共有5个水平. 我们要考察各个工作日顾客的人数是否有显著差异. 如果有显著差异就表明工作日这一因素对顾客的人数的影响是显著的.

一般可设因素A有r个水平A_1，A_2，…，A_r，在水平A_i下进行了n_i（$n_i \geq 2$，i=1，2，…，r）次试验，所得结果如表7-4所示.

表7-4

我们假设水平A_i（i=1，2，…，r）下的样本X_{i1}，X_{i2}，…，X_{in_i}来自总体$N(\mu_i, \sigma^2)$，μ_i，σ^2未知，且设来自不同水平A_i下的样本之间相互独立.

因$X_{ij}-\mu_i \sim N(0, \sigma^2)$，故$X_{ij}-\mu_i$可看成是随机误差. 记$X_{ij}-\mu_i=\varepsilon_{ij}$，则$X_{ij}$可写成

式7-36称为单因素试验方差分析的数学模型 . 特别要注意的是，这里假设所涉及的r个正态总体的方差是相同的.

方差分析的首要任务是对于模型7-36，检验假设：

$$H_0：\ \mu_1 =\mu_2 =\ldots=\mu_r ，$$

$$H_1：\ \mu_1 ，\ \mu_2 ，\ \ldots，\ \mu_r \text{ 不全相等.}$$

（7-37）

下面来建立检验统计量.

记$n_1 +n_2 +\ldots+n_r =n$，引入全部数据的总平均

以及总离差平方和

注意到数据X_{ij} （i=1，2，…，r，j=1，2，…，n_i ）总是参差不齐的，其离散程度可用S_T 来描述. 若S_T 比较大，表示数据X_{ij} 的波动程度比较大；反之，数据的波动程度就比较小.而数据的波动是由试验的随机误差以及因素各水平的效应的差异所引起的. 若后者较前者大得多，则有理由认为因素A的各个水平对应的试验结果有显著差异，从而拒绝H_0 . 为此，我们设法将S_T 分解成两部分：一部分是纯粹由随机误差引起的；另一部分则是由因素A的各水平效应以及随机误差引起的.

注意到上式右端第三项为零，即

从而平方和S_T 就分解成为两个平方和之和：

$$S_T =S_E +S_A ，$$

（7-40）

其中

S_E 中各个加项是各组组内数据与组均值的离差平方和，这完全是由随机误差所引起的，S_E 称为误差平方和 （或组内平方和 ）；而S_A 是各水平A_i 的总体的样本均值与数据总平均的离差平方和，这是由各水平效应的差异以及随机误差引起的，S_A 称为因素A的效应平方和 （或组间平方和 ）. 式7-40就是我们所需要的平方和S_T 的分解式.

下面考察S_E ，S_A 的统计特性.

又因X_{ij} （i=1，2，…，r；j=1，2，…，n_i ）相互独立，故式7-43右端各平方和相互独立，由χ^2 分布的可加性知

$$S_E/\sigma^2 \sim \chi^2 \text{（n-r）}.$$

（7-44）

由此还得知E（S_E/σ^2 ）=n-r，从而有

$$E[S_E/\text{（n-r）}]=\sigma^2 .$$

（7-45）

对于S_A 可以证明有以下的结果：

（2）S_A 与S_E 相互独立，且当H_0 为真时$S_A/\sigma^2 \sim \chi^2$ （r-1）.

进而由式7-44及上述第2个结果，按F分布的定义知道，当H_0 为真时，有

亦即当H_0 为真时

现在我们可以建立用来检验假设7-37的检验法了. 取

为检验问题7-37的检验统计量. 当H_0 为真时，$\mu_1 =\mu_2 =\ldots=\mu_r =\mu$，，由式7-46知此时$E［S_A /（r-1）］=\sigma^2$ ，而当H_0 为不真即当H_1 为真时，由式7-46有$E［S_A /（r-1）］=$. 另一方面，由式7-45知不管H_0 是否为真，$E［S_E /（n-r）］=\sigma^2$. 这就是说当H_0 为真时，F的分子、分母的数学期望均为σ^2 ，而当H_1 为真时，分子的数学期望大于分母的数学期望. 又知分子、分母独立，这表明当H_1 为真时，分式有偏大的趋势，故知检验问题7-37的拒绝域具有形式

取检验的显著性水平为α，确定k的值，使

$$P\{当H_0 为真拒绝H_0 \}=P_{H_0}\{F>k\}=\alpha，$$

由式7-47得$k=F_\alpha$ （r-1，n-r）. 于是方差相等的多个正态总体均值检验问题7-37的拒绝域为

这种检验法称为方差分析法 . 由上面的讨论知道，这一检验法是基于将S_T （它表示数据总的离散程度）分解为S_E （是由随机误差引起的）和S_A （是由因素A各水平效应及随机误差引起的）两部分，然后比较S_A ，S_E 的大小而得到的. 这就是这里说的“方差分析”的含义.

上述分析结果常列成如下的表格（见表7-5），称为方差分析表 .

表7-5　单因素方差分析表

表中，S_T 的自由度是S_A ，S_E 的自由度之和. 是S_A 除以它自己的自由度，是S_E 除以它自己的自由度. 分别称为S_A ，S_E 的均方 （即平均平方和）. 表中各平方和按以下的式子计算较为方便. 记

【例7-6】　设在例7-4中食品品种Ⅰ，Ⅱ，Ⅲ，Ⅳ所对应的总体分别为N（μ_i ，σ^2 ）（i=1，2，3，4），μ_i ，σ^2 均为未知，且设自各个总体取得的样本之间相互独立. 试取显著性水平α=0.01，检验假设：

$$H_0：\ \mu_1 =\mu_2 =\mu_3 =\mu_4 ，$$

$$H_1：\ \mu_1 ，\ \mu_2 ，\ \mu_3 ，\ \mu_4 \text{ 不全相等.}$$

解　图7-6画出了各样本数据的箱线图. 由此可看出各样本数据的大致情况. 由表7-2，r=4，$n_1 =n_2 =5$，$n_3 =4$，$n_4 =6$，n=20. 算得

图7-6

$T_{1.} =533$，$T_{2.} =507$，$T_{3.} =458$，

$T_{4.} =741$，$T_{..} =2259$，

$S_T =2256.95$，$S_A =1372.05$，

$S_E =S_T -S_A =884.9.$

得方差分析表如下：

表7-6　例7.6的方差分析表

因为$F_{0.01}$（3，16）=5.29<8.27，故拒绝H_0，认为品种不同的食品对于增加老鼠的体重有显著差异.

在实际中，往往不是事先给出显著性水平α，而是将F比的值与F分布的上分位点$F_{0.1}$（r-1，n-r），$F_{0.01}$（r-1，n-r），$F_{0.001}$（r-1，n-r）比较，看看在什么水平上差异是显著的. 一般，在F比大于$F_{0.01}$（r-1，n-r）时就已认为差异是高度显著的. 因此，本题中差异是高度显著的.

7.5　双因素试验方差分析

上一节讨论的是单因素试验的方差分析法.　这一节讨论双因素试验的方差分析法. 先看以下的例题.

【例7-7】　用3个不同的温度A_1，A_2，A_3 和4种催化剂B_1，B_2，B_3，B_4 对某种产品进行转化率试验，每个温度与每种催化剂的组合各做两次试验，得到表7-7的结果.

表7-7

这里，试验指标是转化率，在试验中有两个因素——温度（因素A）与催化剂（因素B）在变化，它们分别有3个、4个水平，这是双因素试验. 我们的目的在于考察在两种因素的各种水平下，转化率是否有显著的差异，即考察温度、催化剂对转化率是否有显著的影响.

一般，设试验的指标受到因素A和因素B的作用. 因素A有r个水平A_1

，A_2，…，A_r；因素B有s个水平B_1，B_2，…，B_s. 对于两种因素各个水平的每一种组合（A_i，B_j）（i=1，2，…，r，j=1，2，…，s）在相同的条件下，都进行了l（l≥2）次重复试验. 这种试验称为等重复试验. 今得到以下的试验结果（见表7-8）.

表7-8

设X_{ijk} ~N（μ_{ij}，σ^2）（i=1，2，…，r；j=1，2，…，s；k=1，2，…，l），各X_{ijk} 独立，μ_{ij}，σ^2 >0均为未知参数. 记ε_{ijk} =X_{ijk} -μ_{ij}，就有ε_{ijk} ~N（0，σ^2），则X_{ijk} 可写成

（7-50）式7-50就是我们要研究的模型. 为了研究方便起见，引入以下的记号，将式7-50加以改写. 记

此时

$$\mu_{ij} =\mu+a_i +b_j + [（\mu_{ij} -\mu）-（a_i +b_j ）]$$

（i=1，2，…，r；j=1，2，…，s）.

记

$$（ab）_{ij} =（\mu_{ij} -\mu）-（a_i +b_j ）$$

$$=\mu_{ij} -\mu_i \cdot-\mu_{\cdot j} +\mu，$$

即有

$$\mu_{ij} = \mu + a_i + b_j + (ab)_{ij}$$

$$(i=1，2，\ldots，r；\ j=1，2，\ldots，s)，$$

（7-51）

易知

这样，式7-50就可写成

其中μ，a_i ，b_j ，$(ab)_{ij}$ ，$\sigma^2 >0$均为未知参数. 式7-52就是我们所要讨论的双因素试验方差分析的数学模型 .

在这里，μ称为总平均 . $a_i = \mu_{i\cdot} - \mu$反映了因素A的第i水平A_i 对试验结果的影响，称为因素A的第i水平A_i 的主效应 . 同样称$b_j = \mu_{\cdot j} - \mu$为因素B的第j水平$B_j$ 的主效应 . 在式7-51的右端除了a_i ，b_j 外，剩余的部分$(ab)_{ij}$ 是由于A处于i水平、B处于j水平所引起的. 因为除了A，B这两个因素外，没有其他因素，而它们各自的处于i水平、j水平的作用已在a_i ，b_j 中反映了，因而$(ab)_{ij}$ 就反映了A的i水平与B的j水平联合起来的作用（交互作用），$(ab)_{ij}$ 称为水平A_i 与水平B_j 的交互作用效应 [7] .

考察因素A，因素B以及交互作用对试验结果的影响是否显著，就需要分别检验假设：

我们分两种情况来讨论上述检验问题的处理方法.

1）交互作用效应为零的情况

在实际问题中，如果已知不存在交互作用，或已知交互作用对试验的结果影响很小，可以不考虑交互作用效应时，即$(ab)_{ij}=0$（i=1，2，…，r；j=1，2，…，s）时，试验不必重复，只要取l=1，即可对检验问题7-53和7-54进行检验. 现在取l=1，设对于两个因素各水平的每个组合（A_i ，B_j ）只做一次试验，所得结果（即在表7-8中令l=1，且记$X_{ij1}=X_{ij}$）如表7-9所示：

表7-9

此时模型7-52可写成（即在式7-52中记$X_{ij1}=X_{ij}$，$\varepsilon_{ij1}=\varepsilon_{ij}$）：

① 这种不考虑交互作用的模型，也称为可加效应模型，在这一模型中，两个因素作用在试验指标上的效应就等于各个因素效应之和.

与单因素情况类似，检验问题7-53和7-54的检验方法也是建立在平方和的分解上的. 引入记号

引入总平方和

则S_T 可分解成

即得总平方和S_T 的分解式：

$$S_T=S_E+S_A+S_B .$$

（7-57）

其中

S_E 称为误差平方和 ；S_A ，S_B 分别称为因素A，因素B的效应平方和 ；且可证

当H_{0A}：$a_1 = a_2 = \ldots = a_r = 0$为真时，可证

取显著性水平为α，得假设H_{0A} 的拒绝域为

类似地，在显著性水平α'下，假设H_{0B} 的拒绝域

于是得方差分析表如下：

表7-10　双因素方差分析表（可加效应模型）

表中的平方和可按下述式子来计算：

【例7-8】　为考察3个打字员的打字速度是否有显著差异，并考察4种不同品牌的打字机的性能是否有显著差异，测得各打字员在不同品牌的打字机上的打字速度（以个/分计）如表7-11所示.

表7-11

设本题符合模型7-52中的条件，且已知可以不考虑交互作用效应，试取显著性水平α=0.05检验假设问题7-53和7-54.

解　$T_{i.}$，$T_{.j}$ 的计算已载于上表. 现在r=3，s=4，由式7-58可知

得方差分析表如下：

表7-12　例7-8的方差分析表

由于$F_{0.05}$（2，6）=5.14<F_A，但$F_{0.05}$（3，6）=4.76>F_B，故拒绝H_{0A}而接受H_{0B}，即认为打字员的打字速度有显著差异，而各种品牌打字机的性能无显著差异.

2）不能知道交互作用的效应是否为零的情况

为要检验交互作用的效果是否显著，对于两个因素的每一种组合（A_i，B_j）至少要做2次试验. 这是因为在模型7-52中若l=1，则$(ab)_{ij}+\varepsilon_{ij1}$总以结合在一起的形式出现，这样就不能将交互作用效应与误差分离开来. 现在针对模型7-52在l≥2的情况来讨论检验问题7-53，7-54和7-55的处理方法. 引入记号

引入总离差平方和

则与式7-57类似，S_T可分解成

$$S_T = S_E + S_A + S_B + S_{A\times B}.$$

其中

以上四式中：S_E称为误差平方和；S_A，S_B分别称为因素A，因素B的效应平方和；$S_{A\times B}$称为因素A与因素B的交互效应平方和. 从而可得假设H_{0A}的拒绝域（显著性水平为α）为

假设H_{0B}的拒绝域（显著性水平为α'）为

假设H_{0AB}的拒绝域（显著性水平为α″）为

可写出方差分析表如下：

表7-13　双因素方差分析表

表中的各个平方和可按以下各式来计算.

【例7-9】　在例7-7中，假设符合模型7-52所需的条件，试在显著性水平0.01下检验催化剂（因素B），温度（因素A）以及这两者的交互作用对于产品的转化率是否有显著影响.

解　按题意需检验假设H_{0A}，H_{0B}，H_{0AB}（即式7-53，式7-54和式7-55）. $T_{...}$，$T_{i..}$，$T_{.j.}$ 的计算如表7-14所示，表中括弧内的数是$T_{ij.}$.

表7-14

现在r=3，s=4，l=2，由式7-59可知

$$S_T = 85^2 + 89^2 + \ldots + 89^2 - 1794^2/24 = 2828.5,$$

$$S_A = (604^2 + 595^2 + 595^2)/8 - 1794^2/24 = 6.75,$$

$$S_B = (453^2 + 513^2 + 363^2 + 465^2)/6 - 1794^2/24 = 1960.5,$$

$$S_{A\times B} = (174^2 + 142^2 + \ldots + 174^2)/2 - 1794^2/24 - S_A - S_B$$

$$= 797.25,$$

$$S_E = S_T - S_A - S_B - S_A \times B = 64,$$

得方差分析表如下：

表7-15　例7-9的方差分析表

由于$F_{0.01}$（2，12）=6.93>F_A，$F_{0.01}$（3，12）=5.95<F_B，$F_{0.01}$（6，12）=4.82<$F_{A\times B}$，所以接受H_{0A}，而拒绝H_{0B}，$H_{0A\times B}$，即认为温度对产品的转化率影响不显著，而催化剂以及催化剂与温度的交互作用对产品的转化率的影响显著.

习题7

【答案链接】

以下约定各道习题均符合涉及的回归分析或方差分析模型所要求的条件.

1. 为了考察机器速度x（m/min）与机器温度Y（℃）的关系，测得以下数据：

（1）画出散点图；（2）求Y关于x的回归直线方程.

2. 食盐（以g计）在不同温度（以℃计）下溶解在100mL水中的质量的数据如下：

（1）画出散点图；（2）求Y关于x的回归直线方程；（3）求在x=25处和在x=55处Y的点预测值.

3. 在机械加工镁合金时，为了考察钻高速钢的布氏硬度x对于切割速率Y（m/min）的效应，测得以下数据：

设$Y=\beta_0+\beta_1 x+\varepsilon$，$\varepsilon\sim N(0,\sigma^2)$，且$\beta_0$，$\beta_1$，$\sigma^2$与x无关.（1）画出散点图；（2）求线性回归方程；（3）求σ^2 的无偏估计；（4）检验回归效果是否显著（取显著性水平$\alpha=0.05$）；（5）求在点x=65处Y的预测值.

4. 在某个生产过程中，于一个大气压下，温度自0℃变到10℃，需考察得率Y与温度x的关系. 今测得以下的数据：

设$Y=\beta_0+\beta_1 x+\varepsilon$，$\varepsilon\sim N(0,\sigma^2)$，且$\beta_0$，$\beta_1$，$\sigma^2$与x无关.（1）求线性回归方程；（2）求$\sigma^2$ 的无偏估计；（3）检验回归效果是否显著（取显著性水平$\alpha=0.05$）；（4）求在点x=4处回归函数$\mu(x)$的置信水平为0.95的置信区间；（5）求在点x=4处Y的置信水平为0.95的预测区间.

5. 已知某种晶体的导热率Y与层厚x有关系：

其中β_0，β_1，σ^2（$\sigma>0$）是与x无关的未知参数. 现在测得样本如下：

试估计参数β_0，β_1，并求在点x=500处Y的预测值.

6. 设在例7-5中各工作日顾客人数的总体分别为$N(\mu_i,\sigma^2)$（i=1，2，3，4，5），μ_i，σ^2 均为未知，且各样本之间相互独立. 试检验假设H_0：$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$，$H_1$：$\mu_1$，$\mu_2$，$\mu_3$，$\mu_4$，$\mu_5$ 不全相等（显著性水平$\alpha=0.05$）.

7. 为了了解一种新研制的镇静剂的药效，取6只豚鼠进行试验，对这些豚鼠注射不同剂量的镇静剂，测得它们的睡眠时间（以min计）如

下：

设涉及的三个总体均为正态总体，且方差相同，样本之间相互独立. 问：镇静剂的三种不同剂量对豚鼠睡眠的时间是否有显著差异（显著性水平α=0.05）.并在同一坐标轴上画出各个样本的箱线图.

8. 为了考察不同型号的玻璃以及不同型号的黄磷对于产生电视显像管某一水平的辉度所需的电流（以μA计）大小的影响，使用了2种型号的玻璃和3种型号的黄磷. 对于玻璃和黄磷的不同型号的各种组合都进行了3次试验，得到以下结果：

设诸水平组合下所需电流量的总体服从正态分布，且方差相等，样本之间相互独立. 试在水平α=0.05下分别检验玻璃型号、黄磷型号以及两者的交互作用对所需电流量是否有显著影响.

---

[1] 在式7-2'中，β_1 后面的因子对i=1，2，…，n求和为0，将式7-2'称为是由式7-2经中心化而得的模型.

[2] 容易验证在点处满足Q取到最小值的充分条件.

[3] S_{yy} 将在下一节用到.

[4] 为方便计，我们将β'_0 的估计值和估计量都记为，又将β_1 的估计值和估计量都记为. 在用到它们时，视上下文是能够区分它们是估计值还是估计量的.

[5]

[6] 预测区间的意义与置信区间的意义相似，只是后者对未知参数而言，而前者对随机变量而言.

[7] 在实际问题中，两个因素之间往往会有交互作用. 例如在乒乓球比赛时，男、女单打冠军配对组成的混合双打选手，不一定能得到最好的成绩，而排名较低的某个男选手和某个女选手因搭配得当、发挥出色而夺冠. 这是因为该男、女选手的技术有交互作用，从而提高了成绩.

8 bootstrap方法

本章先讲述如何模拟服从指定分布的随机变量. 模拟一个随机变量，意指利用计算机去产生这个随机变量的样本观察值. 接着讲述一种数据处理方法——bootstrap方法.

8.1 模拟各种分布的随机变量

8.1.1 随机数和伪随机数

在概率统计的应用中，常需要模拟各种分布的随机变量. 某一分布随机变量的样本值，就称为这一分布的随机数. 例如指数分布随机变量的样本值就称为指数分布随机数. 特别地，区间（0，1）上均匀分布的随机变量的样本值称为均匀分布随机数 ，简称随机数 .我们先来考虑如何产生均匀分布随机数，其他分布随机数，一般可以由均匀分布随机数通过变换得到.

产生均匀分布随机数的方法很多，目前使用最广泛的方法是在计算机上利用数学递推公式来产生. 这种按确定性算法得到的数字序列，不可能是真正来自区间（0，1）上均匀分布的独立同分布样本值序列，我们称它为伪随机数 .

在大多数计算机中都装有产生伪随机数序列的算法程序，我们都是

假设由这些程序产生的伪随机数序列能通过独立性和均匀分布检验，可作为随机数序列来使用，需要时用特定的命令加以调用即可.

8.1.2 模拟离散型随机变量

设离散型随机变量X具有分布律

现在来产生X的随机数.

先产生伪随机数u，令

所以X具有给定的分布律.

产生随机变量X的样本值也叫做对随机变量X进行抽样. 上述模拟离散型随机变量的方法的算法为：

产生伪随机数u.

若$u<p_1$，令$X=x_1$，停止.

若$u<p_1+p_2$，令$X=x_2$，停止.

若$u<p_1+p_2+p_3$，令$X=x_3$，停止.

……

按照这一算法，不论使用何种算法语言编写程序，都能产生所需要的随机数.

【例**8-1**】　设随机变量X具有分布律如下：

试产生X的样本值.

解　取算法为：产生伪随机数u，

若u<0.3，令X=2，停止.

若u<0.5，令X=3，停止.

若u<0.6，令X=7，停止.

若u<0.8，令X=10，停止.

否则，令X=11.

【例**8-2**】　设随机变量X具有分布律

试产生X的样本值（X称为取值为1，2，…，n的离散型均匀分布随机变量）.

解　在式8-1中，令x_i =i，i=1，2，…，n；p_1 =p_2 =…=p_n =，就得到式8-3. 再由式8-2中写出的最下面的一个式子得：“若，则令X=i.”也就是，“若i-1≤nu<i，则令X=i=［nu］+1，i=1，2，…，n.”因此，若u是伪随机数，则

$$X=［nu］+1，$$

就是分布8-3的样本值.

【例**8-3**】　设随机变量X服从参数为λ的泊松分布，试产生X的样

本值.

解　X的分布律为

利用这一递推公式，并注意到$p_0 = e^{-\lambda}$，就容易得到产生泊松随机变量的样本值的算法如下（记F=F（i）=P{X≤i}）：

1°产生伪随机数u.

2°i=0，$p=e^{-\lambda}$，F=p.

3°若u<F，令X=i，停止.

4°$p=\lambda p/(i+1)$，F=F+p，i=i+1.

5°转向3°.

（上述算法中，例如，i=i+1意指i的值增加1）.

【例8-4】　设随机变量X服从参数为n，p的二项分布，试产生X的样本值.

解　X的分布律为

且有

$$p_0 = (1-p)^n,$$

可得产生X的样本值的算法如下（记F=F（i）=P{X≤i}）：

1°产生伪随机数u.

2°c=p/（1-p），i=0，g=（1-p）^{n}，F=g.

3°若u<F，令X=i，停止.

4°g=［c（n-i）/（i+1）］g，F=F+g，i=i+1.

5°转向3°.

下面介绍另一种方法.

设U_1，U_2，…，U_n 相互独立，且它们都在区间（0，1）上服从均匀分布. 令

则有$P\{X_i=1\}=P\{U_i<p\}=p$，$P\{X_i=0\}=1-p$，故X_i ~B（1，p）. 又因U_1，U_2，…，U_n 相互独立，故有. 据此，只要产生n个伪随机数u_1，u_2，…，u_n，若统计其中使得$u_i<p$（i=1，2，…，n）的u_i 个数为k，则得X的样本值为k.

8.1.3 模拟连续型随机变量

先证明一个定理.

定理　设随机变量U~U（0，1），F（x）是某一随机变量的分布函数，且F（x）为严格单调增加且连续的函数，则随机变量F^{-1}（U）具有分布函数F（x），其中F^{-1}（x）是F（x）的反函数.

证　由于F（x）严格单调增加且连续，因此其反函数F^{-1}（x）存

在（即有F［F^{-1} （x）］=x），且严格单调增加连续，即得随机变量F^{-1}（U）的分布函数为

$$P\{F^{-1}（U）\le x\}=P\{F［F^{-1}（U）］\le F（x）\}$$

$$=P\{U\le F（x）\}=F（x）.$$

由定理，若要产生以F（x）（F（x）严格单调增加且连续）为分布函数的随机变量X，只需产生U~U（0，1），令$X=F^{-1}$（U）就行了. 又若要产生X的样本值x，只需产生U的样本值u，令$x=F^{-1}$（u）即得. 这一产生X的样本值的方法，称为逆变换法 . 这种方法在随机变量具有严格单调增加且连续的分布函数F（x），而且F^{-1}（x）能够用显式表示时都能使用.

说明：在上述定理中对F（x）在（-∞，∞）上的严格单调连续的要求可放宽为F（x）在某一区间（有限或无限）上取值从0到1，且在此区间上严格单调增加且连续即可.

【例8-5】 设随机变量X具有指数分布，其分布函数为

试产生随机变量X.

解 设U~U（0，1），令$U=1-e^{-X/\theta}$，解得

$$X=-\theta\ln（1-U）.$$

因为当U~U（0，1）时，也有1-U~U（0，1），从而

$$X=-\theta\ln U$$

就是所要产生的指数分布的随机变量. 若有伪随机数u，就有X的随机数-θlnu.

【例**8-6**】　设随机变量X具有韦布尔（Weibull）分布，其分布函数为

试产生随机变量X.

解　设U~U（0，1），令$U=1-e^{-(X/\eta)^{\beta}}$，解得

$$X=\eta[-\ln(1-U)]^{1/\beta}.$$

因为1-U~U（0，1），故

$$X=\eta[-\ln U]^{1/\beta}$$

就是所要产生的韦布尔分布随机变量.

【例**8-7**】　正态随机变量的产生.

标准正态变量的分布函数Φ（x）的反函数不存在显式，故不能用逆变换法产生标准正态变量. 下面介绍一种近似方法.

设U_i ~U（0，1），i=1，2，…，n，且它们相互独立，由于E（U_i）=1/2，D（U_i）=1/12，由中心极限定理，当n较大时近似地有

取n=12，近似地有

这就是说，只需产生12个伪随机数u_1，u_2，…，u_{12}，将它们加起来，再减去6，就能近似地得到标准正态变量的样本值了.　这样做是很方便

的.

又若X~N（μ，σ^2），Z~N（0，1），利用关系式

$$X=\mu+\sigma Z$$

就能得到一般的正态随机变量X的样本值.

模拟随机变量的方法有很多，我们在这里就不一一介绍了.

【例8-8】 （1）试在均值为200的指数分布总体中抽取一容量为4的样本.

（2）一盒子装有5只电灯泡，它们的寿命（以h计）分别为2000，1800，1900，2050，2000，试以放回抽样的方式抽取一容量为5的样本.

解 （1）在随机数表中取4个伪随机数（一般我们都是在计算机上取伪随机数）

0.15544 0.97742 0.97081 0.42451，

由例8-5得到容量为4的样本为

-200ln0.15544 -200ln0.97742 -200ln0.97081 -200ln0.42451，

即

372.29909 4.56777 5.92490 171.36394.

（2）将5只灯泡依次编号为1，2，3，4，5，按题意需在具有分布律为

的总体中抽取一容量为5的样本. 为此，在随机数表中取到5个伪随机数

0.28739　　0.19210　　0.42723　　0.86530　　0.32768.

由例8-2，得分布8-4的5个随机数为

x_1 =［5×0.28739］+1=2，x_2 =［5×0.19210］+1=1，

x_3 =［5×0.42723］+1=3，x_4 =［5×0.86530］+1=5，

x_5 =［5×0.32768］+1=2.

即所求的样本为（x_1，x_2，x_3，x_4，x_5）=（2，1，3，5，2）或即

（1800，2000，1900，2000，1800）.

8.1.4　利用模拟样本进行统计推断

人们可以利用计算机在很短的时间里产生服从各种分布的样本观察值，将这些数据当成是由实际试验产生的数据来处理.　这样的数据叫模拟数据 或模拟样本 . 我们可以利用模拟样本来进行统计推断. 这种利用模拟样本进行统计推断的方法叫随机模拟法 或模拟法 . 特别当统计模型所涉及的统计量很难用解析方法来处理时，可以利用随机模拟法来解决.

下面举两个例子.

【例8-9】　设一系统由相互独立的元件A和元件B串联而成，元件

A、元件B的寿命分别为Y_1，Y_2（以h计），Y_1 服从韦布尔分布，其分布函数为

Y_2 服从指数分布，其分布函数为

试用模拟的方法估计：（1）系统的寿命大于8000h的概率，（2）系统的平均寿命、系统寿命的标准差.

解　（1）系统的寿命为

$$Y=\min（Y_1，Y_2）.$$

产生伪随机数对（U，V），随之分别按例8-6，例8-5产生随机数Y_1 $=10000（-\ln U）^{1/1.5}$ 和$Y_2=8500（-\ln V）$，得到随机数对（Y_1，Y_2），共产生10000对. 这样就得到$Y=\min（Y_1，Y_2）$的容量为10000的模拟样本如下表所示.

将$Y=\min（Y_1，Y_2）$的值按自小到大排序，得到

故得系统寿命大于8000h的概率近似地为

（2）由第3章3.5节的辛钦大数定理，样本均值依概率收敛于总体均值，而且还能证明，样本标准差依概率收敛于总体标准差，即当样本容量n很大时，样本均值以大的概率接近于总体均值，样本标准差以大的概率接近于总体标准差.　于是，我们以样本均值作为总体均值μ的估计，以样本标准差作为总体标准差的估计.

本例中，系统寿命的样本均值、样本标准差分别为

故得系统平均寿命近似地为8429.2752h，系统寿命的标准差近似地为5904.995h.

【例8-10】 长方形金属板的长度X和宽度Y分别服从分布X~U（2.9，3.1），Y~U（1.9，2.1），X，Y均以m计且两者相互独立. 试估计金属板的面积A=XY的数学期望.

解 X，Y的分布函数分别为

现在分别来产生X，Y的随机数. 设U~U（0，1），令U=得

$$X=2.9+0.2U. \tag{8-5}$$

设V~U（0，1），令得

$$Y=1.9+0.2V. \tag{8-6}$$

按式8-5产生随机变量X的10000个随机数，按式8-6产生随机变量Y的10000个随机数，这样就得到A=XY的样本容量为10000的样本值. 列表如下：

以A的样本均值作为金属板面积A的数学期望的估计，得到面积A的数学期望近似为6.000762m^2 .

此例说明，用模拟方法得到的A的数学期望，与A的数学期望的真值E（A）=E（XY）=E（X）·E（Y）=3×2=6m^2 非常非常接近.

8.2　非参数bootstrap方法

本节和下一节介绍非参数bootstrap方法和参数bootstrap方法. 它们是近代统计中的一种用于数据处理的重要方法. 这种方法的实现需要在计算机上作大量的计算，随着这一方法的发展以及计算机功能的增强，它们已成为一种流行的方法.

8.2.1　估计量的标准误差的bootstrap估计

来估计，其中（参见本章例8-9（2））. 然而F常常是未知的，这样就无法产生模拟样本，不能得到式8-7的结果，需要用另外的方法.

现在假设总体的分布函数F未知，但已经有一个容量为n的来自F的数据样本x_1，x_2，…，x_n，考虑到此时对应于x_1，x_2，…，x_n 的经验分布函数F_n　　是已知的，由格里汶科定理（见第5章5.3节），当n很大时，F_n 近似于F，我们就用F_n 代替F，在F_n 中抽取样本. 在F_n 中抽样就是在原始样本x_1 ，x_2 ，…，x_n 中一次随机地抽一个个体，放回，再在x_1 ，x_2 ，…，x_n 中抽一个个体，直至得到容量为n的样本，也就是对具有分布律为

的离散型均匀分布随机变量，以放回抽样的方式抽取容量为n的样本. 将得到的样本记为，称为**bootstrap**样本 .

对bootstrap样本按上一段中计算估计那样求出θ的估计，估计称为θ的**bootstrap**估计 ．相继地、独立地抽得B个bootstrap样本，以这些样本分别求出θ的相应的bootstrap估计如下：

综上所述得到求的bootstrap估计的步骤是：

1°自原始样本x=（x_1 ，x_2 ，…，x_n ）按放回抽样的方法，抽得容量为n的样本（称为bootstrap样本）.

2°相继地、独立地求出B个（B≥1000）容量为n的bootstrap样本，，i=1，2，…，B. 对于第i个bootstrap样本，计算，i=1，2，…，B（称为θ的第i个bootstrap估计）.

3°计算

【例8-11】 生物学家随机地选取了20只某种雄性绿蜘蛛（这种蜘蛛不织网而以追赶或跳跃去捕食），测量它们前腿的长度，得到下面的数据（以mm计）：

15.10 13.55 15.75 20.00 15.45 13.60 16.45 14.05 16.95 19.05

16.40 17.05 15.25 16.65 16.25 17.75 15.40 16.80 17.55 19.05

设前腿长度总体是具有分布函数F的连续型随机变量，F未知，总体的中位数θ是未知参数. 以样本中位数作为总体中位数θ的估计，求估计量的标准误差的bootstrap估计.

解 将原始样本自小到大排序，得到样本中位数为16.425，以

16.425作为总体中位数θ的估计，即.

相继地、独立地抽取10000个bootstrap样本如下：

对于以上每个bootstrap样本求得样本中位数依次为：

于是以原始样本的中位数作为总体中位数θ的估计，由式8-8知其标准误差的bootstrap估计为

8.2.2 用分位数法求未知参数的bootstrap置信区间

设X=（X_1，X_2，…，X_n）是来自总体F（F未知）的样本，θ是总体的待估参数，x=（x_1，x_2，…，x_n）是一个已知的样本值. 是θ的估计量，是θ的相应于原始样本（x_1，x_2，…，x_n）的bootstrap估计量.

就定义为参数θ的置信水平为1-α的bootstrap置信区间.

下面我们用模拟的方法来求bootstrap置信区间.

相继地、独立地从样本（x_1，x_2，…，x_n）中抽出B个容量为n的bootstrap样本，对于每个样本求出θ的bootstrap估计，…，，将它们自小到大排序，得

作为θ的置信水平为1-α的bootstrap置信区间.

上述求置信区间的方法称为分位数法 .

【例**8-12**】　下面给出某个班级20个学生数理统计课程期终考试的得分：

88　67　64　76　86　85　82　39　75　34

90　63　89　90　84　81　96　100　70　96

（1）以样本中位数M=M（X）作为总体中位数的估计，按分位数法求中位数的置信水平为0.95的bootstrap置信区间.

（2）以样本均值作为总体均值μ的估计，按分位数法求总体均值μ的置信水平为0.95的bootstrap置信区间.

（3）以样本标准差S作为总体标准差σ的估计，按分位数法求总体标准差σ的置信水平为0.95的bootstrap置信区间.

解　将原始样本自小到大排序，得

34　39　63　64　67　70　75　76　81　82

84　85　86　88　89　90　90　96　96　100

知原始样本的中位数为，均值=77.75，标准差=17.61.

相继地、独立地抽取10000个bootstrap样本，得到

样本1：89　84　85　70　75　90　34　90　90　89

34　81　96　90　76　100　96　39　96　76

样本中位数，样本均值，样本标准差20.29；

样本2：84　90　39　85　82　70　86　86　70　70

90　90　85　100　96　67　88　100　67　84

（1）将上述算出的bootstrap样本中位数（i=1，2，…，10000）自小到大排序，得

现在，α/2=0.025，1-α/2=0.975，B=10000，k_1 =［10000×0.025］=250，k_2 =［10000×0.975］=9750，由式8-10得总体中位数的置信水平为0.95的bootstrap置信区间为

（2）将上述算出的bootstrap样本均值（i=1，2，…，10000）自小到大排序，得

现在，α/2=0.025，1-α/2=0.975，B=10000，k_1 =［10000×0.025］=250，k_2 =［10000×0.975］=9750，由式8-10得总体均值μ的置信水平为0.95的bootstrap置信区间为

（3）将上述算出的bootstrap样本标准差（i=1，2，…，10000）自小到大排序，得

现在，k_1 =250，k_2 =9750，由式8-10得总体标准差σ的置信水平为0.95的bootstrap置信区间为

8.3　参数bootstrap方法

假设所研究的总体的分布函数F（x；β）的形式已知，但其中包含

未知参数β（β可以是向量）． 现在已知有一个来自F（x；β）的样本X_1，X_2，…，X_n．利用这一样本求出β（在F（x；β）下）的最大似然估计. 在中以代替β得到，接着在中产生容量为n的样本

① 意指样本来自以为分布函数的总体.

这种样本可以产生很多个，例如产生B个（B≥1000），就可以利用这些样本对总体进行统计推断，其做法与非参数bootstrap方法一样． 这种方法称为参数**bootstrap**法 .

【例8-13】 某种类型的热泵的寿命（以年计）服从指数分布，其分布函数为

今测得一样本：

2.0 1.3 6.0 1.9 5.1 0.4 1.0 5.3 15.7 0.7 4.8

0.9 12.2 5.3 0.6

（1）求总体X的中位数的置信水平为0.95的bootstrap置信区间.

（2）求总体X的均值μ的置信水平为0.95的bootstrap置信区间.

解 将原始样本自小到大排序，得

0.4 0.6 0.7 0.9 1.0 1.3 1.9 2.0 4.8 5.1 5.3

5.3 6.0 12.2 15.7

知原始样本的中位数为2.0，样本均值为4.21.

相继地、独立地以为分布函数产生10000个容量为15的bootstrap样本如下：

得总体X的均值μ的置信水平为0.95的bootstrap置信区间为

【例8-14】 设总体X的分布律为

X	1	2	3
p_k	θ	θ	1-2θ

其中参数θ>0，但未知，今测得一容量为16的样本：

X	1	2	3
出现的次数	x_1 =7	x_2 =6	x_3 =3

（1）求θ的最大似然估计.

（2）求θ的置信水平为0.95的bootstrap置信区间.

（3）以作为θ的估计，求标准误差的bootstrap估计.

解 （1）似然函数为

$$L=\theta^{x_1}\theta^{x_2}(1-2\theta)^{x_3}，\ x_1+x_2+x_3=16，$$

$$\ln L=x_1\ln\theta+x_2\ln\theta+x_3\ln(1-2\theta)，$$

以x_1 =7，x_2 =6代入上式，得θ的最大似然估计为.

（2）以代替上述分布律中的θ，得到总体X的近似分布律为

按这一分布律，相继地、独立地产生10000个容量为16的bootstrap样本如下：

作为θ的置信水平为0.95的bootstrap置信区间.

（3）以作为θ的估计，由式8-8，得标准误差的bootstrap估计为

bootstrap方法是由Bradley　Efron在20世纪70年代后期建立的，近30年来有很大的发展.　这一方法可以用于当人们对总体知之甚少的情况，它是近代统计中的一种用于数据处理的重要且实用的方法.

9　在数理统计中应用Excel软件

9.1　概述

微软（Microsoft）公司推出的办公软件包Office得到了广泛的应用，Excel是Office的重要成员之一. Excel是一个功能多、技术先进、使用方便的表格式数据综合管理和分析系统，它采用电子表格方式进行数据处理，工作直观方便；它提供了丰富的函数，可以进行数据处理、统计分析和决策辅助；还具有较好的制图功能.

本书中应用Excel处理数理统计问题，在Windows XP操作系统下使用Excel 2003. Excel 2007有一些较大变动，需要做相应的改变.

启动Excel后就会打开Excel的用户界面窗口，如图9-1所示，该窗口自上而下有标题栏、菜单栏、常用工具栏、格式工具栏、编辑栏、工作表区、工作表标签、水平滚动条和状态栏.

图9-1　Excel的用户界面窗口

工作表区由单元格组成，每个单元格由列标和行号标识. 工作表区的最上面一行为列标，用A，…，Z，AA，…，AZ，BA，…，BZ，…，IA，…，IV表示，最多可使用256列. 工作表区左边一列为行号，用1，2，…，65536表示，最多可使用65536行 [1] . 单元格“A1”表示单元格位于A列第1行. 单元格区域则规定为矩形，例如，“A1：F5”表示一矩形

区域，A1和F5为其主对角线两端的单元格，每张工作表有一个标签与之对应，例如，“Sheet 1”. 工作表隶属于工作簿，一个工作簿最多可由255个不同的工作表组成.

检查Excel的“工具”菜单，看是否已安装了分析工具. 如果在“工具”的菜单中没有“数据分析”项，则需调用“加载宏”来安装“分析工具库”.

“工具”菜单中有了“数据分析”命令项，单击它，就出现“数据分析”对话框，其中有19个模块，它们分别属于5类：

（1）基础分析：①随机数发生器，②抽样，③描述统计，④直方图，⑤排位与百分比排位.

（2）检验分析：①t检验，平均值的成对两样本分析，②t检验，双样本等方差假设，③t检验，双样本异方差假设，④Z检验，双样本平均差检验，⑤F检验，双样本方差.

（3）相关、回归：①相关系数，②协方差，③回归.

（4）方差分析：①方差分析，单因素方差分析，②方差分析，可重复双因素分析，③方差分析，无重复双因素分析.

（5）其他分析工具：①移动平均，②指数平滑，③傅里叶分析.

在本书中，只讲述Excel在几个问题上的应用.

9.2 假设检验

9.2.1　假设检验问题p值的求法

本书中遇到的假设检验问题，其p值的求法，可归结为调用函数NORMSDIST（·），TDIST（·），CHIDIST（·）或FDIST（·），下面举例来说明.

【例9-1】　用计算机求下列各题中关于z检验的p值（其中z_0 是检验统计量的观察值）.

（1）z_0 =2.03，右边检验；

（2）z_0 =-2.42，双边检验.

解　（1）打开一个Excel工作表，单击“插入”，对于弹出的菜单单击“函数”（或者，单击编辑栏中的“f_x　”），对接着弹出的菜单单击“统计”，再对弹出的菜单单击“NORMSDIST”，然后单击“确定”，即弹出对话框.

对显示的对话框，键入x=2.03，即得NORMSDIST（2.03）=Φ（2.03）=0.9788，本题为右边检验，p值=P{Z≥2.03}=1-Φ（2.03）=1-0.9788=0.0212.

选一个工作表中的单元格，键入“=NORMSDIST（2.03）”，也能得到0.9788.

（2）类似于（1），可得NORMSDIST（2.42）=Φ（2.42）=0.9922，本题为双边检验，p值=2P{Z≤-2.42}=2P{Z≥2.42}=2×［1-Φ（2.42）］=2×0.0078=0.0156.

【例**9-2**】　用计算机求下列各题中关于t检验的p值（其中t_0 是检验统计量的观察值，n是样本容量）.

（1）t_0 =2.321，n=15，右边检验；

（2）t_0 =1.945，n=28，双边检验；

（3）t_0 =-1.267，n=8，左边检验.

解　（1）做法参见上例，改为调用函数TDIST（·）. 对于函数TDIST（·），键入x=2.321，df（自由度）=14，Tails（尾数）=1，即得TDIST（2.321，14，1）=P{t≥2.321}=0.017945，p值=P{t≥t_0 }=P{t≥2.321}=0.017945.

（2）对于函数TDIST（·），键入x＝1.945，df（自由度）＝27，Tails（尾数）＝1，即得TDIST（1.945，27，1）=P{t≥1.945}=0.031129，本题为双边检验，p值＝2P{t≥1.945}＝2×0.031129＝0.062258. 也可以键入Tails＝2，即得p值＝TDIST（1.945，27，2）＝0.062258.

（3）对于函数TDIST（·），键入x＝1.267 [2]，df（自由度）＝7，Tails（尾数）＝1，即得TDIST（1.267，7，1）=P{t≥1.267}=0.122839，本题为左边检验，p值=P{t≤-1.267}=P{t≥1.267}=0.122839.

【例**9-3**】　用计算机求下列各题中关于χ^2 检验的p值（其中是检验统计量的观察值，n是样本容量）.

解　（1）做法参见例9-1，改为调用函数CHIDIST（·）.对于函数CHIDIST（·），键入x=29.321，df（自由度）=15，即得CHIDIST（29.321，15）=P{χ^2 ≥29.321}=0.014619，本题为右边检验，p值=P{χ^2 ≥29.321}=0.014619.

（2）对于函数CHIDIST（·），键入x=10.215，df（自由度）=24，即得CHIDIST（10.215，24）=P{χ^2 ≥10.215}=0.993602，本题为左边检验，p值=P{χ^2 ≤10.215}=1-P{χ^2 ≥10.215}=1-0.993602=0.006398.

（3）对于函数CHIDIST（·），键入x=24.672，df（自由度）=10，即得CHIDIST（24.672，10）=P{χ^2 ≥24.672}=0.006003，本题为双边检验，p值=2×min［P{χ^2 ≥24.672}，1-P{χ^2 ≥24.672}］[3] =2×P{χ^2 ≥24.672}=2×0.006003=0.012006.

（4）对于函数CHIDIST（·），键入x=13.974，df（自由度）=27，即得CHIDIST（13.974，27）=P{χ^2 ≥13.974}=0.981507，本题为双边检验，p值=2×min［P{χ^2 ≥13.974}，1-P{χ^2 ≥13.974}］=2×［1-P{χ^2 ≥13.974}］=2×（1-0.981507）=0.036986.

【例9-4】　用计算机求下列各题中关于F检验的p值（其中F_0 是检验统计量的观察值，df=（df_1 ，df_2 ）是自由度）.

（1）F_0 =2.97，df_1 =9，df_2 =14，右边检验；

（2）F_0 =3.32，df_1 =6，df_2 =12，双边检验；

（3）F_0 =0.413，df_1 =9，df_2 =9，双边检验.

解　　（1）做法参见例9-1，改为调用函数FDIST（·）.对于函数FDIST（·），键入x=2.97，df=（9，14），即得FDIST（2.97，9，14）=P{F≥2.97}=0.033334，本题为右边检验，p值=P（F≥F_0 ）=0.033334.

（2）对于函数FDIST（·），键入x=3.32，df=（6，12），即得FDIST（3.32，6，12）=P{F≥3.32}=0.036481，本题为双边检验，p值=2×min［P{F≥3.32}，1-P{F≥3.32}］[4]
=2×P{F≥3.32}=2×0.036481=0.072962.

（3）对于函数FDIST（·），键入x=0.413，df=（9，9），即得FDIST（0.413，9，9）=P{F≥0.413}=0.898071，本题为双边检验，p值=2×min［P{F≥0.413}，1-P{F≥0.413}］=2×（1-0.898071）=0.203856.

9.2.2　两个正态总体均值差的检验（t检验）

设X_1 ，X_2 ，…，X_{n_1} 是来自正态总体的样本，Y_1 ，Y_2 ，…，Y_{n_2} 是来自正态总体的样本，两样本独立，均未知，现用Excel来求解假设检验问题

H_0：$\mu_1=\mu_2$，H_1：$\mu_1>\mu_2$；

H_0：$\mu_1=\mu_2$，H_1：$\mu_1<\mu_2$；

H_0：$\mu_1=\mu_2$，H_1：$\mu_1\neq\mu_2$.

举例来说明.

【例**9-5**】　下面给出地区A的9名推销员和地区B的6名推销员在某一时间段内推销的产品数：

设两样本分别来自总体均未知，两样本相互独立，试取显著性水平α=0.01检验假设

$$H_0：\mu_A=\mu_B，H_1：\mu_A\neq\mu_B.$$

解　用Excel求解如下.

1° 打开Excel工作表，将数据分别输入A1：A10和B1：B7.

2° 依次单击“工具”、“数据分析”、“t检验：双样本异方差假设” [5] 和“确定”，弹出对话框.

3° 在对话框中，键入变量1的范围A1：A10，键入变量2的范围B1：B7，在假定均值差空格中键入“0”，单击“标志”，将Excel预设的“α=0.05”改成“α=0.01”，单击“确定”，跳出一个新的工作表如下（也可以选择使计算结果产生在本工作表中指定位置）.

t检验：双样本异方差假设

② 在Excel算法中，以公式算出v，将v四舍五入得df，而在本书第6章6.2节中是取df=［v］，两者有时会相差1.

4° 结果分析　可以用两种方法来判别检验的结果.

（1）临界值法　这是双边检验，α=0.01，t　　stat（统计量的观察

值）=0.988232，小于临界值3.249836，故接受H_0 .

（2）p值法　p值为0.348861，远大于0.05，故接受H_0 .

下面再举两个用p值法作假设检验的例子.

【例9-6】　据以往的调查得知某地区一周岁婴儿的平均身高为29吋（1吋=2.54cm），今在该地区一托儿所中随机地抽取了30个一周岁的婴儿，测得以下的身高数据：

25　32　35　25　30　26.5　26　25.5　29.5　32　30　28.5　30　32
28　31.5　29　29.5　30　34　29　32　27　28　33　28　27　32
29　29.5

设婴儿的身高服从正态分布N（μ，σ^2 ），μ，σ^2 未知，试用p值法检验假设

$$H_0：\mu=29，H_1：\mu\neq 29.$$

解　本题是单样本的双边检验问题.　由数据可得：样本容量为n=30，样本均值，样本标准差s=2.61082.而且，μ=29，从而可算出检验统计量.　以0.944053和df（自由度）=29，Tails（尾数）=2调用TDIST，得TDIST（0.944053，29，2）=0.352942.　这就是本题的p值，它远大于0.05，故接受H_0 .

【例9-7】　试用p值法检验第6章6.3节例6-7中的假设检验问题

解　本题是双边检验问题. 用χ^2 检验法，检验统计量χ^2 =的观察值为

由计算机得CHIDIST（13.152，9）=0.155848，

它远大于0.05，故接受H_0 .

9.3　一元线性回归

我们以例题说明用Excel求解一元线性回归问题的做法.

【例9-8】　有人做过一项试验，检测被测试者在英文文献中搜索到一特定单词所需的时间（以秒计），搜索内容是在文献中找出含字母k的由4个字符组成的单词，得到以下的数据：

设题目符合线性回归模型$Y=\beta_0+\beta_1 x+\varepsilon$，$\varepsilon\sim N(0,1)$所要求的条件.

（1）给出数据的散点图.

（2）作Y关于x的一元线性回归方程.

（3）检验回归效果是否显著，即检验假设H_0 ：$\beta_1=0$，H_1 ：$\beta_1\neq 0$（取显著性水平$\alpha=0.05$）. 如果回归效果显著，求β_1 的置信水平为0.95的置信区间.

解　打开Excel工作表，将数据键入A1：A15和B1：B15.

（1）依次单击“插入”、“图表”、“XY散点图”、“下一步”，弹出对话框，在“数据区域”键入“A1：B15”，在“系列产生在”认定“列”，单击“下一步”，弹出“图表选项”对话框，在“图表标题”键入“时间-行序”，

在“X轴”键入“x，行序”，在“Y轴”键入“y，时间”，单击“完成”，即显示散点图，如图9-2所示.

图9-2

（2）依次单击“工具”、“数据分析”、“回归”和“确定”，弹出对话框，在“Y值输入区域”键入“B1：B15”，在“X值输入区域”键入“A1：A15”，单击“标志”，认定置信水平为95%，选定“输出选项”，单击“确定”，即得计算结果表格如下：

表中系数一栏中显示“y”：0.249125，“x”：0.615092，它们分别是β_0，β_1 的估计，即，于是得y关于x的回归方程为

（3）上表中P-Value一栏中显示“x”：5.58E-0.9，这是关于β_1 的双边检验

$$H_0：\beta_1=0，H_1：\beta_1\neq 0$$

的p值，由于5.58E-09<0.05，故拒绝H_0，认为回归效果是显著的.

上表中，下限95%一栏中显示“x”：0.522872，上限95%一栏中显示“x”：0.707311，这表示β_1 的置信水平为0.95的置信区间为

（0.522872，0.707311）.

9.4 方差分析

9.4.1　单因素方差分析

【例**9-9**】　为比较某一地区种植的4种谷物（大麦、小麦、玉米和燕麦）的维生素B_1　的含量A，B，C，D（以μg/g计），今依次在总体A，B，C，D中取一个样本：

A：5.2　4.5　6.0　6.1　6.7　5.8

B：6.5　8.0　6.1　7.5　5.9　5.6

C：5.8　4.7　6.4　4.9　6.0　5.2

D：8.3　6.1　7.9　7.0　5.5　7.2

设总体A，B，C，D依次服从正态分布N（μ_i　，σ^2　）（i=1，2，3，4），并设各样本相互独立. 试画出各个样本数据的箱线图，并取显著性水平α=0.05，检验各种谷物的维生素B_1 含量的均值是否有显著的差异.

解　画出各样本数据的箱线图如图9-3所示，从图上可以看出4种谷物的维生素B_1 含量的大致情况.

图9-3

接着，用Excel来求解，步骤如下.

1° 建立给定问题的原假设和备择假设

H_0：$\mu_1=\mu_2=\mu_3=\mu_4$，H_1：μ_1，μ_2，μ_3，μ_4 不全相等.

2° 打开Excel工作表，将数据输入A1：D7.

3° 依次单击“工具”、“数据分析”、“方差分析：单因素方差分析”和“确定”，跳出对话框.

4° 在对话框中键入变量的输入范围“A1：D7”，单击“标志位于第一行”，确认α=0.05 [6] ，单击“确定”，显示结果. 有两张表，前一张是各种谷物的均值、方差等的汇总，后一张是本题的方差分析表：

方差分析

5° 结果分析 临界值法，F=3.956544大于临界值F_{crit} =$F_{0.05}$ （3，20）=3.098391，故拒绝H_0 ，认为各品种的维生素B_1 含量有显著差异. p值法，p值=0.022934，小于0.05，故拒绝H_0 ，认为差异是显著的.

9.4.2 双因素无重复试验的方差分析

【例**9-10**】 用计算机求解第7章例7-8的双因素无重复试验方差分析问题. 按题意需在显著性水平α=0.05下检验假设

H_{0A} ，H_{1A} ；H_{0B} ，H_{1B} （见式7-53，式7-54）.

解 1° 打开Excel工作表，将数据输入A1：E4：

2° 依次单击“工具”、“数据分析”、“方差分析：无重复双因素分析”和“确定”，跳出对话框.

3° 在对话框中键入变量的输入范围“A1：E4”，单击“标志”，确认α=0.05，单击“确定”，显示结果. 有两张表，后一张是本题的方差分析表：

方差分析

4° 结果分析 临界值法，在“行”中，由于F=10.69846大于$F_{crit}=F_{0.05}$（2，6）=5.143253，故拒绝H_{0A} ，认为打字员的打字速度有显著差异. 又在“列”中，由于F=1.48小于$F_{crit}=F_{0.05}$（3，6）=4.757063，故接受H_{0B}，认为各种品牌的打字机性能无显著差异.

p值法，在“行”中，p值=0.010504，小于0.05，故拒绝H_{0A} . 又在“列”中，p值=0.311799，大于0.05，故接受H_{0B} .

9.4.3 双因素等重复试验的方差分析

【例**9-11**】 用计算机求解第7章例7-9的双因素等重复试验的方差分析问题. 按题意需在显著性水平α=0.01下检验假设

H_{0A}，H_{1A}；H_{0B}，H_{1B}；H_{0AB}，H_{1AB}（见式7-53~式7-55）.

解 1° 打开Excel工作表，将数据输入A1：E7.

2° 依次单击“工具”、“数据分析”、“方差分析：可重复双因素方差分析”和“确定”，跳出对话框.

3° 在对话框中键入变量的输入范围“A1：E7”，再键入“每一样本的行数”为“2”，键入α=0.01，单击“确定”，即显示本题的方差分析表如下：

方差分析

4° 结果分析

	临界值法	**p**值法
对于因素A	$F=0.632813<F_{crit}=F_{0.01}$（2，12）=6.926608，接受$H_{0A}$	p值=0.547925>0.05，接受H_{0A}
对于因素B	$F=122.5313>F_{crit}=F_{0.01}$（3，12）=5.952545，拒绝$H_{0B}$	p值=2.89E-09<0.05，拒绝H_{0B}
对于交互作用A×B	$F=24.91406>F_{crit}=F_{0.01}$（6，12）=4.820574，拒绝$H_{0AB}$	p值=4.13E-06<0.05，拒绝H_{0AB}

9.5 bootstrap方法、宏、VBA语言

我们举例说明，在Excel环境中bootstrap方法在计算机上的实现.

Excel虽然有多方面的功能，但也不可能直接处理各种各样的具体问题，很多情况下需要读者自编称为“宏（Macro）”的程序以解决问题.

“宏”是包括一连串指令的一个小程序，编写宏要使用VBA（Visual Basic for Application）语言，VBA是Office软件包中的标准语言.

本书没有篇幅介绍VBA，请读者参阅有关书籍.

【例9-12】 从装载铁矿石的船上，随机地取22块铁矿石样品，测得以下各块样品中含铁的百分比数据：

62.66 63.22 62.87 63.22 63.22 63.08 63.01 62.87

62.1 61.68 63.43 62.45 63.22 62.1 63.57 62.87

61.75 62.87 63.15 62.94 63.08 62.38

（1）以样本中位数M=M（X）作为总体中位数的估计，求估计量M的标准误差的bootstrap估计.

（2）以样本中位数M=M（X）作为总体中位数的估计，按分位数法求中位数的置信水平为0.95的bootstrap置信区间.

（3）以样本均值作为总体均值μ的估计，按分位数法求总体均值μ的置信水平为0.95的bootstrap置信区间.

（4）以样本标准差s=s（X）作为总体标准差σ的估计，按分位数法求总体标准差σ的置信水平为0.95的bootstrap置信区间.

解 相继地、独立地抽取10000个bootstrap样本（样本1，样本2，…，样本10000），作法见下述的“宏”.

对于以上每一个bootstrap样本，依次求得其样本中位数为

并将它们按自小到大的次序排列，得到

对于以上每一个bootstrap样本，依次求得其样本均值为

并将它们按自小到大的次序排列，得到

对于以上每一个bootstrap样本，依次求得其样本标准差为

并将它们按自小到大的次序排列，得到

（1）以原始样本的中位数M=62.905作为总体中位数的估计，由第8章式8-8知其标准误差的bootstrap估计为

（2）取区间

作为总体中位数的置信水平为0.95的bootstrap置信区间.

（3）取区间

作为总体均值μ的置信水平为0.95的bootstrap置信区间.

（4）取区间

作为总体标准差σ的置信水平为0.95的bootstrap置信区间.

以上结果是从“宏”得到的. 下面给出求解这一问题的宏.

```
Sub Macro1()
Dim i As Integer, j As Integer, k As Integer, Sum As Double, Squ
    As Double
    For i = 1 To 10000
```

```
10   For j = 1 To 22
        k = Int(Rnd * 22)+ 1
         Cells(i, j + 9)= Cells(k, 7)
20   Next j
30   Cells(i, 1)= WorksheetFunction. Median(Cells(i, 10), Cells(i
40   Cells(i, 2)= Cells(i, 1)
50   Cells(i, 3)= WorksheetFunction. Average(Cells(i, 10), Cells(
60   Cells(i, 4)= Cells(i, 3)
  Next i
70  Sum = 0#
    For i = 1 To 10000
     Sum = Sum + Cells(i, 1)
    Next i
    Sum = Sum / 10000
    Cells(32, 7)= Sum
    Cells(34, 7)= Sum - 62.905
    Squ = 0#
    For i = 1 To 10000
        Squ = Squ +(Sum - Cells(i, 1))^2
    Next i
    Cells(36, 7)= Squ
80  Cells(38, 7)= Sqr(Squ / 9999)
90  For i =1 To 10000
        Squ = 0#
        For j = 1 To 22
            Squ = Squ +(Cells(i, 4)- Cells(i, j + 9))^2
        Next j
```

```
        Cells(i, 5)= Sqr(Squ / 21)
        Cells(i, 6)= Cells(i, 5)
100 Next i
End Sub
```

对这个宏，说明如下：

（1）对于一个Excel工作表，设计为：使用工作表的第1至第10000行，第1至第31列（A，B，…，Z，AA，…，AE列），用单元格G1:G22存放原始样本，用单元格J1:AE10000存放bootstrap样本，用单元格A1:A10000存放bootstrap样本中位数，用单元格C1:C10000存放bootstrap样本均值. 用单元格E1:E10000存放bootstrap样本标准差.

（2）在宏中出现的变量名和数组要用Dim语句来声明（Dim是dimension的缩写）：Integer，整数，从-32768到32767；Long，大整数，从-2147483648到2147483647；Single，单精度浮点数，负数从-3.402823E38到-1.401298E-45，正数从1.401298E-45到3.402823E38；Double，双精度浮点数，负数从-1.79769313486232E308到-4.94065645841247E-324，正数从4.94065645841247E-324到1.79769313486232E308. 程序中，数字后的#号表示浮点数.

（3）使用循环语句，以i计工作表中的行（1至10000），以j计工作表中的列（10至31），借助随机数Rnd的取值来产生bootstrap样本，存放在单元格Cells（i，j+9）中（这个单元格位于工作表中的第i行，第j+9列）. 以10标注的程序行到以20标注的程序行，完成这部分工作.

在Excel环境中（0，1）上的均匀分布随机数函数为RAND，但在VBA环境中，（0，1）上的均匀分布随机数函数为Rnd，在宏中需使用

Rnd.

k为整数1，2，3，…，22中之一.　若k=1，则以62.66给Cells（i，j+9）赋值；若k=2，则以63.22给Cells（i，j+9）赋值，…；若k=22，则以62.38给Cells（i，j+9）赋值.

（4）执行了上述，得到bootstrap样本：

（5）以30标注的程序行，使用Median函数求bootstrap样本的中位数. Median函数在VBA中没有，在Excel中才有，在宏中使用（即在VBA环境中使用）要在其前面加上“Worksheet　Function.”（凡是在VBA中没有而在Excel中才有的函数，在VBA环境中使用，都要如此处理）. 求得的中位数放入Cells（i，1）.

（6）以40标注的程序行，是给Cells（i，1）作一备份. 这是因为，诸bootstrap样本的中位数以后要按数值大小排序以求置信区间，如果不作一备份，中位数与bootstrap样本将不能保持逐行对应.

（7）以50标注的程序行，使用Average函数求bootstrap样本的均值，放入Cells（i，3）［与（5）类似］.

（8）以60标注的程序行，是给Cells（i，3）作一备份［与（6）类似］.

（9）以70标注至以80标注的程序行，用于求. 由第8章式8-8知，以原始样本的中位数M=62.905作为总体中位数的估计，其标准误差的bootstrap估计为

这就是题（1）的答案.

（10）以90标注至以100标注的程序行，产生各bootstrap样本的标准差，并作一备份.

上述程序运行结束、运算结果呈现以后，依次将光标指向A列列标和排序的工具，单击鼠标左键，即出现警示对话框，选择其“以当前选定区域排序”而不是“扩展选定区域”，单击“排序”，即可将A列数据自上而下按从小到大排序，从而得的置信水平为0.95的bootstrap置信区间

（A250，A9750）=（62.66，63.15）

这就是题（2）的答案.

改为指向C列列标，进行类似的操作，得μ的置信水平为0.95的bootstrap置信区间

（C250，C9750）=（62.59955，63.01）

这就是题（3）的答案.

改为指向E列列标，进行类似的操作，得σ的置信水平为0.95的bootstrap置信区间

（E250，E9750）=（0.344348，0.634688）

这就是题（4）的答案.

下面再举几个例子.

【例9-13】 给出例8-10的程序如下. 程序中产生随机数的语句的依据见第8章式8-5、式8-6.

```
Sub Macro7()
    Dim i As Integer, Sum As Double
    Sum = 0#
    For i = 1 To 10000
        Cells(i, 1)= i
        Cells(i, 2)= 2.9 + 0.2 * Rnd
        Cells(i, 3)= 1.9 + 0.2 * Rnd
        Cells(i, 4)= Cells(i, 2)* Cells(i, 3)
        Sum = Sum + Cells(i, 4)
        Cells(i, 6)= Sum
    Next i
    Cells(6, 7)= Sum / 10000
End Sub
```

【例9-14】 试给出一个宏，以产生例8-13中分布函数为

的指数随机变量的样本. 共需产生10000个相互独立的样本容量为15的样本.

解 给出的宏如下. 程序中产生随机数的语句的依据见例8-5.

```
Sub Macro4()
Dim i As Integer, j As Integer, Sum As Double, Squ As Double
    For i = 1 To 10000
      For j = 1 To 15
        Cells(i, j + 8)= -4.21 * WorksheetFunction. Ln(Rnd)
      Next j
    Next i
```

```
End Sub
```

【例**9-15**】　设总体X~N（98.7，4.67^2 ），试产生X的1000个相互独立的容量为10的样本.

解　解这一问题的“宏”如下[7]. 程序中产生随机数的语句的依据见例8-7.

```
Sub Macro7()
    Dim i As Integer, j As Integer, k As Integer, Sig As Single
    For i = 1 To 1000
        For j = 1 To 10
            Sig = 0#
            For k = 1 To 12
                Sig = Sig + Rnd
            Next k
            Cells(i, j + 2)= 98.7 + (Sig - 6) * 4.67
        Next j
    Next i
End Sub
```

由这个宏得到如下的样本：

本章内容可参考相关的Excel书籍[8][9][10].

习题**9**

【答案链接】

用计算机求下列各题的p值.

1. Z检验　（1）z_0 =1.24，右边检验，

（2）z_0 =-1.84，双边检验.

2. t检验　（1）t_0 =3.025，df=23，右边检验，

（2）t_0 =-1.145，df=4，左边检验.

4. F检验　（1）F_0 =2.28，df=（12，20），右边检验，

（2）F_0 =0.75，df=（30，12），双边检验.

5. 用Excel求解. 在两个不同地区的农场，测得他们生产的河虾的重量（以g计）如下：

农场A　15.5　12.7　12.1　14.4　16.1　15.0　16.2

农场B　11.9　13.3　15.8　11.6　10.4　13.6　13.8　12.4　13.6　13.0

设样本依次来自正态总体均未知，两样本独立.

试用（1）临界值法（取显著性水平α=0.05），（2）p值法检验假设

$$H_0：\mu_1=\mu_2，H_1：\mu_1>\mu_2.$$

6. 用Excel求解. 考察白鼠和棕鼠在迷宫中找到出口逃出迷宫需要花费的时间（以min计）. 以下数据分别是6只试验白鼠和6只试验棕鼠花费的时间：

设样本依次来自正态总体均未知，两样本独立. 试用p值法检验假设

$$H_0：\ \mu_1 = \mu_2，\ H_1：\ \mu_1 \neq \mu_2 .$$

7. 设有来自总体N（μ，σ^2 ）的样本：

10.05　10.00　10.02　9.97　10.07　10.03

9.98　10.10　9.95　9.99　10.00　10.08

试用p值法检验假设

$$H_0：\ \sigma^2 = 0.09^2，\ H_1：\ \sigma^2 < 0.09^2 .$$

8. 用Excel求解. 对于习题7的第1题和第2题，

（1）画出散点图；（2）求一元线性回归方程；（3）检验回归效果是否显著（取显著性水平α=0.05）；（4）若回归效果显著，求系数 β_1 的置信水平为0.95的置信区间.

9. 用Excel求解习题7第6题，并在同一数轴上画出各样本的箱线图.

10. 用Excel求解习题7第7题.

11. 下面给出美国某城市1970~1979年时间段中的年降雨量（以mm计）：

30.5　29.4　23.8　21.1　19.1　35.1　16.5　34.9　30.3　31.0

（1）以样本均值作为总体均值μ的估计，按分位数法求μ的置信水平为0.95的bootstrap置信区间（取B=1000）.

（2）以样本中位数M=M（X）作为总体中位数的估计，按分位数法求的置信水平为0.95的bootstrap置信区间（取B=1000）.

---

[1] Excel 2007增大了数据容量，列数增为16384，行数增为1048576.

[2] TDIST（x，df，Tails）不接受x为负值.

[3] 参见第6章式6-36.

[4] 参见第6章式6-36.

[5] Excel称这一检验为“t检验：双样本异方差检验”，事实上这是两样本方差未知的双样本t检验，在这里不需要异方差这一条件.

[6] Excel已设α=0.05，如需取α≠0.05，则需修正.

[7] Excel的“数据分析”中有“随机数发生器”. 调用之，可以给出均匀分布、正态分布、伯努利分布、二项分布、泊松分布、离散分布、模式分布计7种类型的随机数. 但是，只能将给出的随机数放入给定的单元格区域，然后在使用时调用，比较麻烦.

[8] 〔日〕Project-A & Dekiru系列编辑部. 办公宝典——Excel/2003/2002/2000 VBA大全［M］. 彭彬等译. 北京：人民邮电出版社，2007.

[9] 宇传华. Excel统计分析与电脑实验［M］. 北京：电子工业出版社，2009.

[10] John Walkenbach. Excel 2007宝典［M］. 杨艳等译. 北京：人民邮电出版社，2008.

附表1　标准正态分布表

附表2　t分布表

附表3　χ^2 分布表

附表4　F分布表

F分布表　α=（0.05）

F分布表　α=（0.025）

F分布表　α=（0.01）

F分布表　α=（0.005）

习题答案

习题1

1. （1）{H，TH，TTH，TTTH，…}；（2）{1，2，3，…}；（3）{2，3，…，7}.

3. （1）9/16；（2）1/2.

4. 6/55.

5. 0.084.

6. （1）1/100；（2）18/25.

7. （1）455/969；（2）232/323.

8. 1/21.

9. 5/33.

10. 67/100.

11. P（A|B）=1/3，P（B|A）=1/5，P（A|A∪B）=5/7，P（A|AB）=1.

12. （1）11/50； （2）3/22.

13. （1）21/125； （2）7/8.

14. （3）0.60，0.90.

15. （1）31/72； （2）5/36； （3）7/12.

16. （1）0.829； （2）0.171.

17. $[p_3 + p_1 p_2 - p_1 p_2 p_3]$ $[1-(1-p_4)(1-p_5)(1-p_6)]$.

18. 7/12.

19. （1）7/10； （2）3/7，8/21，4/21.

20. 0.66.

21. 10/39，16.2/39，12.8/39.

习题2

1. $P\{Y=k\}=1/2^k$ （k=1，2，3，…）； 1/3.

2.

X	2	3	4
p_k	1/6	2/6	3/6

3. $P\{Y=3\}=1-p^4$，$P\{Y=k\}=(1-p)p^k$（k=4，5，6，…）.

4.（1）0.190；（2）0.878；（3）0.745；（4）0.216.

5.（1）0.590；（2）0.919.

6.（1）0.345；（2）0.393.

7. $P\{X\le 4\}=0.440$，$P\{X>3\}=0.735$，$P\{2<X\le 5\}=0.491$.

8.（1）c=3/2；（2）3/16；（3）79/432.

9. $P\{X\ge 200\}=1/9$，$P\{X\le 100\}=3/4$，$P\{X=300\}=0$.

10. 0.05.

12.（1）$P\{X\le 30\}=15/22$，$P\{10<X\le 80\}=25/42$，$P\{70\le X\le 120\}=3/38$，$P\{X\ge 110\}=0$.

13. $P\{X>50\}=0.75$；$P\{40\le X\le 80\}=0.2$.

17.（1）c=6，1/2；（2）e^{-1}.

$P\{0<Y<2|X=1\}=e-e^{-1}$.

23.（1）不独立；（2）独立；（3）不独立.

习题3

1.

2. 8/3.

3. E（X）=11.6，E（Y）=400.

4. 4p（1-p）（1-2p）.

5. 3/2.

6. 3π/5.

7. 3/4.

8. 124/21.

9. E（X）=1，E（Y）=2，E（XY）=3.

10. 5.

11. 1.

12. E（Y）=0，D（Y）=1/2.

14. （1）1.44；1440000；（2）0.984.

17. cov（X_1，X_2）=0.019，$\rho_{X_1 X_2}$=0.397，D（X_1 -X_2）=0.059.

18. cov（X，Y）=-0.32，ρ_{XY} =-0.749.

（2）E（X）=5/3，D（X）=5/9，E（Y）=10/3，D（Y）=5/9，

E（XY）=35/6，cov（X，Y）=5/18，ρ_{XY} =1/2.

习题4

2.（1）0.7734，0.8413，0.3830；（2）13.71.

3.（1）μ=70，σ=14.81；（2）0.984.

4.（1）0.952；（2）0.006.

5. （1）X+Y~N（250，25），X-Y~N（50，25），（X+Y）/2~N（125，25/4）；（2）0.0694，0.0456.

6.（1）N（40，30），（2）0.3264，（3）N（1，5/4）.

7. 0.00616.

8.（1）0.8665；（2）0.9952.

9. 0.0013.

10. 4976.

11. 0.0207.

12.（1）

X	-10	0	1

p_k	0.01	0.18	0.81

（2）0.9936.

13. 0.95.

14. n至少为4356.

习题5

11.

12. （1）n=11；（2）0.3707，0.5，N（-1，145/144）.

13. （1）$X_1+X_2 \sim \chi^2$（10）；（2）E（Y）=10，D（Y）=20.

习题6

1. （1）接受H_0，（2）接受H_0.

2. 拒绝H_0.

3. 拒绝H_0. 认为鱼的含汞量的均值不等于1.2.

4. 接受H_0. 不能认为均值显著大于9.5.

5. 拒绝H_0 .

6. 接受H_0 .

7. 接受H_0 . 认为两总体均值相等.

8. 拒绝H_0 . 认为杂志的论文比未出版的报告可理解性高.

9. 拒绝H_0 .

10. 拒绝H_0 .

11. 拒绝H_0 .

12. 接受H_0 .

13. 接受H_0 .

14. 接受H_0 .

15. 拒绝H_0 .

16. 拒绝H_0 . 认为方差为5.5.

17. 拒绝H_0 .

18. 接受H_0 .

19. 接受H_0 .

20. 拒绝H_0 .

21. 拒绝H_0 . 认为各种颜色小汽车发生事故不是等可能的.

22. 接受H_0 . 认为数据来自正态分布N（μ，σ^2 ）.

23. 接受H_0 ，认为寿命服从密度为.

24. 接受H_0 .

25. 接受H_0 . 认为一天报警次数服从泊松分布.

26. 不独立.

27. 不独立，认为维生素C有防冷作用.

28. 不独立. 认为结构形式与损坏模式是有关联的.

29. （1.9±0.0784）.

30. （64±3.29）.

31. （8.234±0.014）.

32. （-4.9075，5.5075）.

33. （0.3695，2.6305）.

34.（5.413，19.774）.

35.（0.241，0.543）.

36.（0.285，4.011）.

习题7

1. =-445.9181+0.6649x

4. ，回归效果显著；在x=4处μ（x）的置信水平为0.95的置信区间为（9.2545±0.4138）；在x=4处置信水平为0.95的Y的预测区间为（9.2545±1.2316）.

6. 接受H_0，认为各工作日的顾客人数无显著差异.

7. 拒绝H_0 ，认为三种不同剂量的镇静剂使豚鼠的睡眠时间有显著差异.

8. $F_{A\times B}$ =0.32，F_A =192.08，F_B =8.96，拒绝H_{0A} 和H_{0B}，接受$H_{0A\times B}$，即认为玻璃型号、黄磷型号对所需电流影响显著，而两者的交互作用对所需电流影响不显著.

习题9

1.（1）0.107488，（2）0.065768.

2.（1）0.003013，（2）0.158031.

3.（1）0.088333，（2）0.006826.

4.（1）0.049786，（2）0.503654.

5. 自计算机跳出以下结果.（是单边检验）

t-检验：双样本异方差假设

（1）临界值法. α=0.05，t Stat（统计量的观测值）=2.123304，大于单尾临界1.782288，故拒绝H_0 .

（2）p值法. p值为0.027605，小于0.05，故拒绝H_0 .

6. p值=0.90683，接受H_0 .

7. p值=0.0091，拒绝H_0 .

8. 见习题7第1、2题答案.

9. 接受H_0 .

10. 拒绝H_0 .