在机器学习模型中，大量借用了概率和统计的理论。很多模型本身就是概率模型，比如朴素贝叶斯、最大熵模型、马尔科夫链、条件随机场等。

本篇文章基于《概率论与数理统计 第四版》，粗略但系统地总结了概率和统计的基本知识。借助这些知识，你可以更深刻地理解机器学习问题。

# 随机事件与概率

**第一节．基本概念**

1、随机试验

我们通常谈到概率，都是指某个随机事件（简称事件）的概率，即概率是针对随机事件而言的。而要定义随机事件，就要从随机试验（简称试验）说起：

随机试验是指这样的试验：

1. 可以在相同条件下重复进行；
2. 每次试验可能的结果不止一个，且事先明确试验的所有可能结果；
3. 进行一次试验之前不能明确哪个结果会出现。

可以看出，随机试验需要保证可以重复进行，也就是说试验必须具有可复现性，在无尽的时间长河中，能够一直重复地做下去；而“试验结果不止一个”和“进行一次试验前不能明确哪个结果会出现”保证了试验结果的随机性。

随机试验的概念也是理解后面要介绍到的“随机变量”概念和“大数定律”等的基石。

2、样本空间与样本点

样本空间和样本点的概念是针对随机试验而言的。随机试验的所有可能结果构成的集合称为样本空间；集合中每个元素（即每个可能的结果）称为样本点。

3、随机事件

有了以上概念，我们就可以定义随机事件：一个随机事件，就是一个随机试验的样本空间中的子集。

可以看出，我们是利用集合论的语言来定义的随机事件。所以在随机事件的基础上定义概率论，也必然与集合论有千丝万缕的联系。

下面来了解一些特殊的事件：

1. 基本事件：由一个样本点组成的单点集；
2. 必然事件：样本空间本身；
3. 不可能事件：

这些定义比较严格，但是非常抽象。然而它们对应的现实世界中的对象，跟我们直观上对“随机”、“事件”的印象是一致的。

举一个简单的例子：一个袋子里装有红、黄、蓝三种颜色的小球，我们每次从袋子中抽取一个小球，那么抽取的小球的颜色的可能性当然是红、黄、蓝其一。

在这个例子中，抽取小球就是一个随机试验，它可以重复进行，试验的结果就是抽取的小球的颜色，显然结果不止一个，且试验前不能确定。所有试验结果（红、黄、蓝）构成了这个试验的样本空间。“取出小球为红色（或黄色、蓝色）”这一事件，就是该试验中可能发生的事件。但是注意，“取出小球是红色或蓝色”、“取出小球不是黄色”的事实也是该试验可能发生的事件，它们也是该试验样本空间中的子集。

在这个例子中，取出的小球为红色（或黄色、蓝色）是基本事件；取出的小球为红黄蓝其一是必然事件；取出的小球是黑色为不可能事件。

4、随机事件的运算

设对某个试验，有两个事件A和B，则定义：

1. 和事件：
2. 积事件：
3. 差事件：

基于随机事件的运算，可以定义事件之间的一些关系：

1. 互不相容事件（互斥事件）：
2. 互逆事件（对立事件）：且

（对于事件A，其互逆事件记作：）

**第二节．**概率与频率

1、概率的定义

概率的概念是基于随机试验和随机事件的。我们常说某个事件发生的概率，我们把概率看成是针对事件的一个实数值（或者说针对集合的实数值）。

定义如下：给定试验E、其样本空间为S，对于E的每一个事件A，定义概率为满足以下条件的实数值：

（1）

（2）对必然事件S，有

（3）对两两互不相容事件，有

可以看出，概率的定义并不是一个精确的、具体的定义，它只给出了概率需要满足的必要条件。在实际中有很多具体的定义都满足这些必要条件，它们都属于概率，称为不同的“概型”，如：古典概型、几何概型等。

2、概率的性质

（1）非负性：

（2）规范性：对于必然事件S，

（3）可加性：对两两互不相容事件，有

上面三条其实就是概率的定义中的条件，由这三个条件，可以导出概率的其他性质：

（4）

（5）

（6）

可以看出，概率P的性质与集合的card的运算性质十分类似。是因为P和card都是基于集合定义的，二者具有相同的精神内核。

3、概率概念的延伸

（1）条件概率：

“事件A发生的条件下，事件B发生的条件概率”定义为：



注意，这个式子就是条件概率的定义式，而不是根据其定义推导出来的式子。它的定义并没有明确地指出它是一种概率（尽管可以证明，它确实满足概率的定义，因此是一种概率）。

概率是针对一个事件的发生的实数值，而条件概率也可以被这样看待，但它针对的并不是简单的事件，而是“条件事件”，即事件A发生的条件下，事件B的发生。

条件概率的实际意义是，若事件B的发生可能会受到A发生的影响，则“已知A发生时，B发生”的概率，与“单纯B发生”的概率会有不同。

“条件概率”的概念是贝叶斯思想的基础。

条件概率的衍生定理：

乘法定理：

全概率定理：

其中，是S的一个划分，且

贝叶斯定理：

其中，是S的一个划分，且

贝叶斯定理在实际中应用有很多，参数估计中的贝叶斯估计、机器学习中的朴素贝叶斯模型等，都依靠贝叶斯定理的思想。

（2）用概率定义事件的独立性：

若事件A和B满足：



则称A与B是互相独立的。

可以看出，两事件的独立性是用概率公式定义的。实际中，我们对事件独立的直观感受是，两个事件的独立意味着两个事件的发生互不影响。那么为什么这与上面的概率公式是对应的呢？这可以用条件概率解释。由上面的式子结合条件概率公式可以推出：



且

第一个式子表明，“B发生的条件下，A发生的概率”，跟“单纯考虑A发生的概率”是一样的，也就是说，B是否发生不会影响A的发生的概率，也即B不影响A，同理从第二个式子我们能够知道A是否发生不会影响B发生的概率。综上可知，A和B互不影响，也就是互相独立。

4、频率

在相同条件下进行n次试验，在这n次试验中，事件A发生的次数称为事件A发生的频数，比值称为事件A发生的频率，并记为。

可以看出，概率与频率的概念都是基于“试验”、“事件”定义的，实际上，二者都是针对“事件”的一个实数。但概率是一个不具体的定义，频率则是一个具体的定义；频率的定义需要给定一个n次独立重复试验，而概率则不需要，概率只是针对一次试验而言的。

后面在“随机变量”概念的加持下，根据大数定律，我们可以严格证明，大量重复试验中，一个事件的频率可以逼近概率。这个结论是统计推断（参数估计、假设检验）的理论基础。

**第三节．**实际推断原理

实际上，概率的严格定义是很困难的，公理化的概率论是基于测度论建立的。在概率与统计的课本中，给出了基于集合论的一个不那么抽象，但是也挺抽象的定义。尽管如此，现实中我们都对概率有很直观的认识，概率表示事件发生的可能性。我们常常依靠对概率的评估做出决策，比如，我们常常使用“实际推断原理”：

概率很小的事件在一次试验中实际上几乎是不发生的。

# 第二章 随机变量及其分布

第一节．一维随机变量

1、随机变量的定义

设随机试验的样本空间为，是定义在样本空间S上的实值单值函数，则称为随机变量。

从定义中并不能明显地看出随机变量有什么用，或者说为什么要这么定义，你得细品。实际上，随机变量就是用有序的数值代表样本空间中的元素，也就是把样本空间给数值化了。这样，就可以把概率统计的问题转化成函数性质或者函数计算的问题。其本质思想就是数学建模。

2、离散型随机变量和连续型随机变量

离散型随机变量是指取值为有限个或可数无限个；连续型随机变量，简单的理解就是其取值为实数轴上连续的区间，但这样说并不严格，其需要用分布函数对其进行严格的定义。显然，对于连续型随机变量，其取任意值的概率都为0，这是它很重要的一个性质。

3、随机变量的分布规律

分布规律是指随机变量取每一个实值的概率（也就是试验中每个事情发生的相应的概率），对于离散随机变量和连续随机变量，它们的分布规律的形式有很大不同。分布规律用数学的语言刻画了随机试验的统计规律。

（1）离散型随机变量的分布律：

离散型随机变量用“分布律”来表征其分布规律，分布律就是对随机变量每一个取值，给出其对应的概率。



（2）连续型随机变量的分布函数和概率密度函数：

首先来介绍一下分布函数，分布函数是针对任意随机变量的分布规律的表征，既可以用于离散型，也可以用于连续型，其定义为：

对随机变量X，，定义为X的分布函数。

根据分布函数，我们可以计算X的取值落在任意区间上的概率：



显然，分布函数是一个不减的函数。

尽管分布函数普适于所有随机变量，但实际中，离散型随机变量多用分布律，而连续型随机变量多用概率密度函数。概率密度函数是在分布函数的基础上定义的，下面我们来介绍它。

对于随机变量X的分布函数F(X)，如果存在非负可积函数f(x)，使对任意实数x，有：



则称f(x)是X的概率密度函数。

可以看出，概率密度函数是分布函数的导数。那么如何理解概率密度函数呢？我们知道，连续型随机变量取任意值的概率都为0，因此我们一般不会计算连续型随机变量取单个值的概率。我们通常考虑X取值落在某一区间内的概率，由概率密度与分布函数的关系可知，X取值落在某区间中的概率，就等于概率密度函数在该区间上曲线下的面积。从局部来看，概率密度函数在某个x上的取值越大，说明X取值在x附近的概率越大。

4、常见的分布函数

离散型：0-1分布、二项分布、泊松分布；

连续型：均匀分布、指数分布、正态分布。

5、随机变量的函数及其分布

类似于一般的变量的函数，随机变量也有函数的概念，那么其含义是什么呢？

对于随机变量X，其函数Y=g(X)也是一个随机变量，它的意义是，当X取值为x时，Y取值为g(x)。

已知X的概率密度，如何求Y的概率密度？

设Y=g(X)，g是单调函数，其反函数为h，

则有：



证明：g是单调函数，则g(x)>0，或g(x)<0.

当g(x)>0时：



由X和Y的关系，可知，则上面等式右端第一项为1。

根据概率密度与分布函数的关系，上面等式右端第二项为，即。

显然，上面等式右端第三项为。

当g(x)<0时，证明方法类似。

由此得证。

第二节．**多维**随机变量

我们为什么要研究多维随机变量呢？首先我们回顾一下随机变量的概念，前面我们已经解释了，随机变量是样本空间到R的映射，它的意义在于将样本空间的元素对应到实数值。用实数值来代表样本空间的元素，有助于用函数来处理相关的概率问题。

在实际中，多个随机变量之间可能具有实际意义上的关联，我们需要把它们综合到一起进行考虑。很自然地，类比一般的变量与向量的概念，我们可以将多个随机变量构成一个随机向量（或者说多维随机变量）进行研究。这就是多维随机变量的意义。

下面我们以二维随机变量为例进行介绍。

1、二维随机变量的定义

设样本空间，和是定义在S上的两个随机变量，则由它俩构成二维随机变量(X,Y)。

二维随机变量(X,Y)的性质不仅与X和Y有关，还依赖于这两个随机变量间的相互关系。

2、二维随机变量的分布

（1）联合分布函数：类似于一维随机变量，二维随机变量也有分布函数。二维随机变量的分布函数分为联合分布函数和边缘分布函数。

联合分布函数的定义如下：



定义式中的表示“X取值小于x，且Y取值小于y”事件。

需要注意，理解“联合分布”的要点在于，**X和Y要同时取值**。只有同时取值，才能看出X和Y之间的互相影响（也就是相关性）。对后面的随机变量的函数的分布的理解也是以此为基石的。

从联合分布的定义中，我们容易计算出X取值落在区间，且Y取值落在的概率：



（2）联合分布律：用来表征离散型二维随机变量的分布规律，其形式为：



其中，表示“X取值为，且Y取值为”事件。

（3）联合概率密度函数：我们知道一维的连续型随机变量的概率密度函数是其分布函数的导数。对于二维连续型随机变量，相应地也有概率密度函数，那么其概率密度函数如何定义呢？其与分布函数又有何关系呢？

类比一维的情况，二维连续型随机变量的概率密度函数定义为满足如下条件的非负可积函数：



可见，二维连续型随机变量的概率密度函数是其对应的分布函数的二阶混合偏导数：



对二维连续型随机变量的概率密度函数的意义的理解，与一维的情况类似。曲面下的体积表示两个随机变量的取值落在该区块内的概率。

（4）边缘分布：对于二维随机变量X,Y的分布函数F(X,Y)，其边缘分布是指X和Y的单独的分布函数，分别记为：和。

边缘分布可以由联合分布得到：



边缘分布律可以由联合分布律得到：



边缘概率密度也可以由联合概率密度得到：



可以看到，由联合分布可以推出边缘分布，但由边缘分布推不出联合分布。如何由联合分布求取一个变量的边缘分布呢？就是联合分布对另一个变量积分。

从两个随机变量的边缘分布中，只能知道它们各自的分布情况，而不能知道它们之间的相关性，只有从联合分布中才能知道它们之间的相关性。比如，你知道了X的分布和Y的分布，那么你确实完全地掌握了X和Y各自的分布规律，但是你以为这就掌握了全部的信息了吗？No，你仍然不确定X和Y的关系，此时X和Y的关系有无数种可能，它们可能互相独立，也可能有很强的相关性。但如果你知道了二者的联合分布，那么X和Y的关系就完全地呈现在你的眼前，这时你才能说你掌握了关于X和Y的所有信息。

3、条件分布

由随机事件的条件概率的概念，我们很容易引出条件分布的概念。

对于离散型随机变量，就是条件分布律，对于连续型随机变量，就是条件概率密度。

条件分布律：



其中，表示“在‘Y取值为’事件已经发生的条件下，‘X取值为’事件的发生”。

条件概率密度：



条件分布和联合分布的概念在机器学习的概率模型中非常重要。实际上，所谓的概率模型本质上就是求一个分布，其中生成模型是求一个联合分布，判别模型是求一个条件分布。

注意，从形式上看，条件分布和都是关于x和y的二元函数（x、y分别是X和Y的取值变量）。

4、随机变量独立性

由随机事件的独立性概念，我们很容易引出随机变量的独立性的概念。我们将用随机变量的分布来定义随机变量的独立性。

二维随机变量(X,Y)的联合分布函数和边缘分布函数分别是F(X,Y)、和，则若对所有x,y，有：



则称随机变量X和Y相互独立。

可以看出，两个随机变量互相独立意味着任意的“X取某个值”事件与任意的“Y取某个值”事件互相独立。

对于离散型随机变量X,Y，它们互相独立的条件等价于：

对任意的和，有：



对于连续的随机变量X,Y，它们互相独立的条件等价于：



在平面上几乎处处成立。

5、两个随机变量的函数的分布

两个随机变量的函数h(X,Y)也是一个随机变量，它的意义是，当X取值为x同时Y取值为y时，h(X,Y)的取值为h(x,y)。

h(X,Y)的概率密度函数可以由(X,Y)的联合概率密度求出，以Z=X+Y为例，其概率密度函数为：



其意义可以这样理解：比如Z取值为的概率密度，就等于X,Y的联合概率密度中，所有X的取值x与Y的取值y满足x+y=的f(x,y)进行积分。

注意，随机变量(X,Y)的函数h(X,Y)的分布，甚至其取值，与X和Y的相关性有很大关系。

（理解这个结论的一个关键点是，随机变量的函数h(X,Y)的取值空间并不是对于所有的X的可能的取值x和所有的Y的可能的取值y，得到的所有h(x,y)构成的集合。而是需要X和Y同时取值。）

比如，给定X和Y的分布，分别为和，当X和Y互相独立时，和不互相独立时，二者的函数的分布是不同的。

举个简单的例子，设X与Y同分布，它们取值空间都为{1,2,3}且取每个值的概率都是1/3。如果X与Y独立，也就是说X和Y的取值与对方无关，那么随机变量Z=X+Y的取值空间为{1,2,3,4,5,6}；但若两个随机变量不独立，比如，设想一个极端的情况：Y=X，则Z=X+Y=X+X的取值空间为{1,4,6}，显然，两种情况下取值空间都不一样，分布也肯定不一样了。

（使用“同时取值”的认识，很容易理解这个例子的意思）

为了研究随变量的分布，我们定义了很多分布函数。无论是一维随机变量的分布函数还是二维随机变量的联合分布函数和边缘分布函数；无论是针对离散型随机变量的分布律还是针对连续型随机变量的概率密度函数，它们都是普通函数。虽然随机变量并不是普通的变量，但是表征它们分布规律的函数，都是普通函数。只不过它们都具有概率的意义，我们用它们来研究随机变量的统计规律。

# 第三章 随机变量的数字特征

前面我们定义了随机变量，研究了随机变量的分布、多维随机变量的分布和随机变量的函数的分布，我们知道随机变量的分布完全地表征了其统计规律，多维随机变量中也包含了其中每一维随机变量的统计规律和各维随机变量之间关系的全部信息。但实际中我们有时候很难获取准确的分布，这时我们希望通过简单的指标大概地了解随机变量的统计规律和不同随机变量之间的关系，这指标就是随机变量的“数字特征”。

随机变量的数字特征包括：期望、方差、标准差、协方差、相关系数、矩。下面我们一一介绍它们。

第一节．期望

对于离散型随机变量X：

对于连续型随机变量X：

期望反映了随机变量取值的集中趋势，又称均值，通常记为。

实际中，给定随机变量X的一组观测值，我们就可以利用这组观测值来估算X的期望，具体的做法是，求这组观测值的平均值。其严格的依据需要到《大数定律》部分才讲到，这里我们针对离散型随机变量的情况，给出一个简单直观的解释：对于观测值，设它在所有观测值中出现的频数为（即它重复出现次），则所有观测值的平均值等于：



也就等于：



假设观测值足够多，使得观测值的频率能够近似等于其概率，那么近似等于，所以平均值就等于，也就是X的期望。

第二节．方差



对于离散型随机变量：



对于连续型随机变量：



方差反映随机变量X的所有取值相对期望的偏离的一个平均程度。或者说，方差表征了一个随机变量取值是比较集中还是比较分散。若方差为0，则，也就是X的几乎所有的取值都为其均值。方差通常记为。

实际中我们通常用一个随机变量的期望和方差对其进行变换，比如：



称为X的零均值化。



称为X的标准化。

第三节．协方差与相关系数

协方差与相关系数是表征两个随机变量之间关系的数字特征。

1、协方差



协方差相当于对X和Y分别进行零均值化，得到和,然后对二者乘积（也是一个随机变量，假设为Z）求期望。

我们前面讲过，两随机变量的函数的分布与这两个随机变量的相关性有很大关系。也就是说，Z的分布与和的相关性有很大关系，所以使用Z的期望可以简单地评估和的相关性，也就是X和Y的相关性。协方差可正可负，其值越大表明X和Y越相关，其中正值表示正相关，负值表示负相关；值越小则说明X与Y越不相关。

实际中，给定随机变量X和Y的一组联合观测值，我们可以根据这些观测值估算协方差：

观测值集合为：



估算协方差的公式为：



有人可能想问，这样计算是不是有问题，因为根据协方差的定义公式，应该将X零均值化后的每个取值跟Y零均值化后的每个取值相乘（而不只是对应值相乘），然后求期望。这样想是不对的，这个观点有两个错误：第一个错误是误认为Z的取值总是的每个取值跟的每个取值相乘。上面已经强调过，在协方差定义中，Z的取值与和的相关性有关，如果和是互相独立的，那么Z的取值确实是的每个取值跟的每个取值相乘，但若和具有一定相关性，那么Z的取值就不能这样计算了；第二个错误是误认为“观测值的对应相乘”与“的每个取值跟的每个取值相乘”是不一致的。这其实是把“随机变量的取值”和“随机变量的观测值”混为一谈，二者其实是不同的。“取值”是指随机变量所有可能取到的值，观测值是指在多次随机试验中，该随机变量的观测结果，显然，观测值的范围就随机变量的取值范围。当试验次数较多时，观测值会取到所有的随机变量的取值，当观测次数较少时，观测值可能取不到所有的随机变量的取值。

实际上，当与互相独立时，观测值对应相乘能够近似地等价于每个取值与每个取值相乘（假设观测值足够多），因为在大量重复试验中，总能取到所有可能的结果；而当与具有一定相关性时，观测值对应相乘也能近似等价于与相乘的取值，究其本质原因，是因为观测值是X与Y的“联合观测值”，即每一对观测值都是在一次试验中同时观测到X和Y的取值，也就对应于Z的取值空间的一个样本点。“同时观测X和Y的取值”就考虑到了X和Y的相关性。比方说，假如X和Y具有相关性，导致X取时，Y不能取，那么在观测值集合中就绝对不会出现；或者X和Y的相关性导致X取时，Y取的概率很小，那么在观测值的集合中也会表现出来，即会非常少。

当然，利用观测值计算协方差的公式还有另一种理解方式，就是将观测值集合看成是多个独立同分布的随机变量在一次试验中产生的结果的集合。这样可以利用大数定律的思想来推导该公式。

2、协方差矩阵

设有n维随机变量：，它们之间两两求协方差，就构成协方差矩阵：



3、相关系数

相关系数相当于用随机变量X和Y的方差对它们之间的协方差归一化：



相关系数取值范围在[-1,1]，它被用来评估X和Y的（线性）相关性。其绝对值越大，表明X和Y的（线性）相关性越强，其中为负值时，是负相关，为正值时，是正相关；其绝对值越小，表明X和Y的（线性）相关性越弱。当时，我们说X和Y不相关。

注意，这种评估是针对X和Y的线性相关性的。这可以用下面的结论来解释：

的充要条件是，存在常数a,b，使。其中，指的是在一次随机试验中观测(X,Y)（注意，是对X和Y同时观测）的取值(x,y)满足y=ax+b的概率。

需要注意的是两个随机变量的“相关性”和“独立性”的关系。两个随机变量互相独立能推出二者不相关，但不相关不能推出互相独立。不过对于二维正态分布来说，不相关与相互独立是等价的。

第四节．矩

矩也是随机变量的一种数字特征，它一般用于参数估计等统计推断问题中。下面给出一些矩的定义：

（1）k阶原点矩：（期望是一种特殊的原点矩）

（2）k阶中心矩：（方差是一种特殊的中心矩）

（3）k+l阶混合矩：

（4）k+l阶混合中心矩：（协方差是一种特殊的混合中心矩）

我们看到，随机变量的数字特征只反映了随机变量的部分信息，而随机变量的分布函数（或分布律、概率密度函数）才反映了其全部的统计规律信息。比如，两个随机变量的相关性（是一种“弱关系”），可以只通过随机变量的协方差或相关系数来确定；但它们之间的独立性（是一种“强关系”），必须通过联合分布才能确定。（相关性和独立性的区别还在于，相关性有程度的概念，相关程度的大小由协方差或相关系数的值给定；独立性没有程度的概念，独立就是独立，不独立就是不独立）。

# 第四章 大数定律和中心极限定理

大数定律和中心极限定理是基于概率理论的定律，也是统计推断方法的理论基础。所以它们可以说是连接概率理论和统计推断问题桥梁。

下面介绍两个大数定律和一个中心极限定理：

第一节．辛钦大数定律

设是互相独立且服从同一分布的随机变量序列，且数学期望为，作前n个变量的算术平均，则对任意，有：



辛钦大数定律说明，当n足够大时，独立同分布的n个随机变量的算术平均（也是一个随机变量）的取值与期望的差距可以达到任意小（按照概率意义）。

第二节．伯努利大数定律

设是n次独立重复试验中事件A发生的次数，p是事件A在每次试验中发生的概率，则对任意正数，有：



伯努利大数定律是说，如果独立重复试验的次数n足够大，那么事件A发生的频率与其概率就能够任意地接近（按照概率的意义）。

该定律的证明的一个关键点是，是一个随机变量，它是n个独立同分布(参数为p的0-1分布)的随机变量之和。

第三节．中心极限定理

是互相独立且服从同一分布的随机变量序列，其期望和方差分别是：

，则随机变量之和的标准化变量



的分布函数满足：



中心极限定理表明，大量的独立同分布的随机变量之和（也是一个随机变量）的分布逼近于正态分布。（更近一步，李雅普诺夫中心极限定理表明，无论这些随机变量服从什么分布（它们可以不是同分布），只要满足一定的条件（这个条件的实际意义可能就是：“每个影响因素的影响都是微小的”），那么它们的和的分布也逼近于正态分布）。

中心极限定理的实际意义是，若一个随机变量的取值由大量的影响因素结合而成（也就是它相当于大量的随机变量之和），且每一个影响因素的影响都是微小的，那么这个随机变量就服从正态分布。实际中很多现象都是正态分布的。

大数定律和中心极限定理应用到实际的统计推断问题，需要借助于抽样（这是下一章要介绍的内容）。抽样提供的大量样本对应于多个独立同分布的随机变量，这就实现了实际问题的抽象建模。

# 第五章 抽样

前面几章介绍了概率论的知识，包括随机事件、概率、随机变量及其分布、随机变量的数字特征等几个主要知识点。后面我们将介绍统计推断方法。概率论的知识都是为统计推断方法提供服务的，它们为统计推断方法提供了理论基础和思想。我们的终极目标就是利用统计推断方法，解决实际中的统计问题。

应用大数定律的思想解决统计推断问题的基本思想就是将大量样本（或者说观测值）都看成是随机变量，用这些随机变量的某种组合（或者说某种函数）的特征来逼近总体分布的相应特征。

在统计推断问题中，我们把实际数据抽象成概率论中的理论模型，再用概率论的方法予以解决。那么实际数据与理论模型的对应是怎样的？常见的有两种对应方式：

一就是上一章结尾说到的，认为采集的每个样本（或者说观测值）相当于一个随机变量在一次随机试验中的观测的结果，那么大量的样本对应于大量的独立同分布的随机变量的观测值（而不是对应于同一随机变量在多次随机试验中的不同观测值）；

二就是前面括号里所说的那种情况，比如在机器学习问题中，数据的每一维特征都看成是一个随机变量，而每个样本在该维特征的取值看成是一次试验中该随机变量的观测值。

统计推断的问题中，大多数都是使用第一种对应方式。这就是本章要着重讲述的。

为了对实际数据进行建模，首先要建立几个概念：总体、个体和样本。

第一节．基本概念

1、总体、个体

我们考察实际中某个对象的一项指标（如学生的身高），通过相应的随机试验（如逮住一个学生，测量他的身高）获取该指标的观察值，那么试验的全部可能的观察值就称为总体；

每个可能的观察值称为个体；

总体中包含的个体的个数称为总体的容量，容量有限的总体称为有限总体，容量无限的总体称为无限总体。

可以看出，“随机试验”、“随机变量”、“总体”这几个概念包含的内容是一样的，它们之间可以对等地理解。

2、样本

设X是具有分布函数F的随机变量，则与X分布相同且相互独立的随机变量称为从分布函数F得到的容量为n的样本。它们的观察值

称为样本值。

可以看到，总体和样本都有容量的概念。注意这里的总体的观察值和样本的观察值都可能重复，所以这里“总体”的概念并不相当于概率论中的“样本空间”的概念，“个体”的概念也不等价于概率论中的“样本点”的概念。只能说，总体中的每个个体的取值范围是相应的随机试验的样本空间。

统计推断的问题，简单地说就是用样本的观察值去推断总体的特征。

可以看出，数理统计的基本概念与概率论的基本概念是一脉相承的，但需要注意这些统计的基本概念与概率论中的概念的异同：这里涉及到的随机试验、随机变量的概念与概率论中是相同的；不过要特别注意这里的“样本”、“样本值”与概率论中的“样本空间”、“样本点”的区别。这些概念虽然都带有“样本”的字样，含义却大相径庭。

另外，我们一般将“样本”和“总体”对应地理解，认为前者是局部，后者是整体。这样理解并不确切，实际上，“样本值”和“总体”才是局部与整体的关系。而样本是随机变量的序列。

基于样本的概念，我们可以提出“统计量”的概念，它是解决统计推断问题的基本工具。

第二节．统计量

1、统计量的定义

设样本的一组样本值为，则称样本的函数为统计量。显然统计量也是一个随机变量，而是它的一个观察值。

常见的统计量有：

样本均值：

样本方差：

其他样本矩：…

2、统计量的经验分布函数

经验分布函数是一个统计量函数（是一个关于x的函数，对于每一个固定的x，它是一个统计量），它是用来逼近总体分布函数F(x)的：

设是总体F的一个样本，用表示中不大于x的随机变量的个数，然后定义经验分布函数：



可以看出，当n足够大时，可以无限逼近F(x)。这下看到了统计量的强大之处了吧？使用样本观测值，连分布函数都能逼近！

3、统计量的抽样分布

统计量的抽样分布是一个重要的概念，在使用统计量进行统计推断时，常常要知道它的分布。在后面的参数估计问题中，你可以体会到统计量分布的重要性。

常见的统计量的抽样分布：

分布：n个正态总体（均值为0，方差为1）的样本之和服从自由度为n的分布



t分布：设，且二者互相独立，则称随机变量t服从自由度为n的t分布



F分布：设，则称随机变量F服从自由度为的F分布



再次强调，统计量是解决统计推断问题的基本工具。求解统计推断问题的基本步骤就是先对给定的样本构造合适的统计量，然后利用统计量的分布的特点，进行统计推断，具体的过程可以在后面的参数估计和假设检验中体会。

# 第六章 参数估计

基于抽样的基本概念，尤其是统计量的基本概念，我们就可以导出统计推断问题的求解过程。

统计推断包含两类问题：参数估计和假设检验。本章我们介绍第一类问题：参数估计。

参数估计问题就是假设总体X的分布函数形式已知，但包含未知参数，借助X的一个样本的观测值来估计总体的位置参数的问题。

参数估计问题又包括两个子问题：点估计和区间估计。点估计是估计未知参数本身，而区间估计是对未知参数给出给定置信水平下的置信区间。

下面我们分别研究这两个子问题：

第一节．点估计

1、矩估计

点估计的最基本的思路是构造一个统计量，由样本的观察值可以得出统计量的观察值，用它作为未知参数

的近似值。

比如：用样本均值近似总体均值



一般地，可以用样本矩来近似总体矩。

我们把统计量称为估计量，把其观察值称为估计值。

可以看出，估计量就是一个用于点估计的统计量，本质上它是一个随机变量。

当未知参数是期望、方差或者其他矩时，可以直接用上面的思路进行点估计，但是当未知参数不是矩时，就不能直接使用上述思路，而是借助样本矩，间接地进行估计：

（1）首先推出总体矩与未知参数之间的关系：

以连续型随机变量X为例，其概率密度函数为，其l阶矩（总体矩）的表达式为



在本问题中，总体X的概率密度函数形式已知，是未知参数，因此上式就是l阶矩关于未知参数的函数。进一步，由于我们有k个参数，我们需要获得前k阶矩与未知参数的关系，构成方程组



（2）然后用样本矩代替总体矩



（3）解方程组，得到参数的值。

可以看出，我们根据总体矩与分布中的未知参数的关系，通过用样本矩近似总体矩，间接地计算出了未知参数，不可谓不巧妙也。这种未知参数的点估计方法称为矩估计。还有一种点估计方法称为极大似然估计，其思路与矩估计迥异。

2、极大似然估计

以离散型随机总体X为例，对于给定的n维样本，它们的联合分布律是。由于样本中的随机变量互相独立，所以联合分布律等于。对于样本对应的观察值（为了与联合分布律中的区分，这里的观察值采用上标的记法），可以计算出其出现的概率为，由于该观察值已经出现，所以我们认为最优的估计值是使观察值出现的概率最大的值。

因为总体分布的形式已知，观察值也已知，所以样本观察值出现的概率是关于未知参数的函数，称为似然函数。则问题就转化为求使似然函数极大的。该问题为优化问题。

对于实际中不同的问题，总体X的分布函数具有不同的形式，所以优化问题的目标函数也不同。

极大似然估计是不考虑条件概率这一概念的，即不考虑一个随机变量的取某个值的概率受别的随机变量的影响，而极大后验概率则考虑了条件概率的概念。

3、点估计的估计量选取标准

对于同一参数，可能有不同的估计量对它进行估计，看起来似乎都是合理的。但究竟哪个更好呢？这就需要我们建立一些估计量的评价标准：

（1）相合估计：这是估计量的基本要求。

相合估计是指对任意都满足的的估计。

（2）无偏估计：

无偏估计是指满足的的估计。

（3）估计量的有效性：

有效性是一种相对指标，若对参数的两个估计量和，有，则说比有效。

可以看出，点估计是用样本的估计量的估计值去近似未知参数的真值。显然，样本容量越大则估计越准确，但是只通过点估计，我们无法知道估计的准确度与样本容量大小的关系，或者说，我们无法衡量在有限样本容量下，估计的“准确度”。我们只知道当样本容量无限大时，估计值是否能无限接近未知参数，以及样本容量无限大时，估计值的一些性能（无偏性，有效性）。对有限样本容量下估计的准确度的衡量，就是区间估计要干的事。

第二节．区间估计

1、置信水平与置信区间

首先定义置信水平和置信区间的概念：

假设总体X的分布函数，对于给定值，如果样本的两个统计量和满足，则称随机区间为的置信水平是的置信区间。

注意，和都是随机变量，区间是一个随机区间。概率的意义是，我们把看成是一个确定的值，则这个概率就是“对随机区间进行观测的结果（是一个确定区间）包含”这件事发生的概率，也就是“落在随机区间的观测结果（是一个确定区间）中”这件事发生的概率。

随机区间的具体取值结果有很多，因此对于确定的随机区间，其满足的取值可能有多个。

那么我们如何构造随机区间呢？

一般的步骤如下：

1. 基于样本，构造“枢轴量”（也是个统计量），使它的分布不依赖于。
2. 求两个常数a，b，使（由于W不依赖于未知参数，故a，b很容易确定）
3. 求出使的和，则就是所求置信区间。

2、区间估计的具体步骤（举例说明）

设总体，其中已知，为未知，设是来自X的样本，求的置信水平为的置信区间。

解：（1）构造枢轴量

（2）根据正态分布的特点，可知

（3）则有，

则随机区间就是的一个置信水平为的置信区间。

然后我们根据的观察值计算出一个确定的区间，其意义是：“的真值落在这个确定区间内”这件事发生的概率是，或者说，“这个确定区间包含”这一陈述的可信程度为。

对于不同的区间估计问题，需要构造不同的枢轴量。

常见的正态总体的区间估计问题有：单个正态总体期望的估计（分方差已知、方差未知两种情况）；单个正态总体方差的估计；两个正态总体的期望之差的估计；两个正态总体的方差之比的估计。

上面的例子就是方差已知，估计期望的情况。

单侧置信区间：满足（或）的区间（或）称为的置信水平为的置信区间。

从我们对正态总体的期望的区间估计的例子中可以看出，置信水平为的置信区间为，显然，样本容量n越大，置信区间的长度就越小，当n为无穷大时，无论置信水平是多少，置信区间长度为0。实际上，对任何参数的区间估计问题，都有此结论。也就是说，对未知参数进行区间估计的“精确度”是受样本容量影响，样本容量越大，估计的越精确（显然，置信区间越小说明越精确），这也符合我们的直观理解。对于无限样本容量，就根本不用进行区间估计，因为依据大数定律，能够完全准确地得到未知参数的值。

但是样本容量有限且给定的情况下，给定置信水平的置信区间的大小，也跟总体分布的形式有关。从例子中容易看出，总体的方差越大，置信区间越大；总体方差越小，置信区间越小。

# 第七章 假设检验

第一节．概述

假设检验问题是统计推断的一类，与参数估计问题类似，也是利用样本观察值推断总体特性。

假设检验中的“假设”是指关于总体的未知参数或者总体的分布形式的论断。假设检验的终极目标就是利用样本观察值提供的依据，决定做出的假设是被接受还是拒绝。

假设检验的一般形式为：给定两个假设、，分别称为零假设（或原假设）和备择假设，这两个假设是对立的，即接受则拒绝，拒绝则接受。

然后通过一些手段决定是接受还是拒绝。

需要注意的是，在实际的假设检验问题中，这两个假设并不是“对称”的，也就是说，实际中我们对接受假设和拒绝假设通常会有一些偏向。比如对于疾病检测问题，我们的假设是某人有病，那么我们更偏向于接受这个假设，也就是说，尽量不要拒绝这个假设，因为如果一旦我们判断他没病，而不对其进行治疗，但是实际上他有病，那么对他而言结果是很严重的，所以针对这种情况，我们必须从统计数据中找到拒绝假设的非常充分的依据，或者说证明该假设成立的概率非常之小，才能拒绝该假设，否则必须接受该假设；再比如次品检测问题，我们假设一个公司的产品的次品率大于质检标准要求的最大阈值，那么我们更偏向于拒绝这个假设，也就是说，尽量不要接受这个假设，因为如果我们判断次品率超标，但实际上次品率并未超标，那么对该公司的声誉会造成很大负面影响。所以我们必须有充分的理由证明次品率确实超标，或者说我们证明次品率未超标的概率非常之小，才能接受这个假设，否则必须拒绝这个假设。

在实际中，我们就是依据上述思路进行假设检验的。上述介绍的两个典型例子中包括了判断失误了的两种情况：一是假设实际为真，而我们拒绝了该假设，这称为I类错误；二是假设实际上不真，而我们接受了该假设，这称为II类错误。显然，第一个例子中，我们相当于关心的是控制I类错误，使犯I类错误的概率尽量小（这种假设检验称为显著性检验）；第二个例子中，我们相当于关心的是控制II类错误，使犯II类错误的概率尽量小。

当然，我们可以看出，控制的I类错误，就相当于控制的II类错误（因为与是对立的）。所以我们只需研究控制I类错误的假设检验的方法（即显著性检验）即可。

第二节．举例说明

下面以关于正态总体的期望的假设检验问题为例，来介绍显著性检验的一般思路：

正态总体，其中已知，为未知，假设，其中是给定的常数。

根据显著性检验的思路，我们需要控制I类错误。即当为真时，这个概率非常小时，才拒绝，否则接受。为此，我们构造统计量，它的分布是确定的。显然，统计量的观察值的绝对值越大，则和越不可能相等，因此应保证小于一个较小的正数k，也就是：。此时，。由于的分布确定（均值为0，方差为1的正态分布），给定k的情况下，这个概率可以计算出来，假设为。根据正态分布的特性，可以得到k与的关系：。显然，当为真时，k越大，越小。由于我们需要控制I类错误，所以认为非常小时，也就是犯错误的概率非常小时，才拒绝。因此我们给取一个很小的数，一般为0.05，然后计算出k的值，然后根据样本的观察值计算统计量的观察值，若该值比k大，说明非常小概率的事件发生了，这充分说明假设不成立，这时才能拒绝，否则必须接受。

可以看出，假设检验的思路类似于 “反证法”，即假设我们的假设成立，那么可以构造一个小概率事件（），然后用观察值检验，如果使小概率事件为真，就认为假设不成立，因为现实中我们认为小概率事件不会发生。

我们也可以用另一种理解方式去看待假设检验的过程：当然，首先总是要构造一个统计量，该统计量的分布是确定的。然后给定以后，可以在统计量的分布函数上画出一个“拒绝域”，若统计量的观察值落在这个拒绝域，则拒绝假设，否则接受假设。

第三节．双边检验和单边检验

双边检验和单边检验都是关于总体中的未知参数的检验。

上面的例子中关于的检验问题，是双边检验，其拒绝域是双边的；像、这样的假设检验，称为单边检验，其拒绝域是单边的。

第四节．常见的假设检验问题

1、单个正态总体的均值检验

，当已知时，即上面例子介绍的情况，是Z检验；当未知时，是t检验。

2、两个正态总体均值差检验

，t检验

3、单个正态总体方差检验

，检验

4、两个正态总体方差检验

，F检验

5、分布拟合检验

，检验

6、分布族拟合检验

，检验

第五节．P值检验

P值检验与上面所述的临界值检验相对。

临界值检验是给定显著性水平，计算p值并将p值跟显著性水平比较以决定是否接受假设；p值检验是通过统计量先计算出p值，然后考虑显著性水平设置为多少会拒绝假设，设置为多少会接受假设。

p值的意义是假设假设为真，构造的统计量服从确定的正态分布，然后计算出在该分布下统计量的观测值产生的概率，它是原假设可以被拒绝的最小的显著性水平。也就是说，显著性水平比它再小就必须接受假设了。

第六节．假设检验与区间估计的关系

假设检验问题跟区间估计问题是对应的，假设检验问题可以转换成区间估计问题解决，区间估计问题也可以转换成假设检验问题解决。

比如一个双边假设检验问题：假设分布参数θ等于一个常数θ0（显著性水平为α）。该问题可以先求θ的置信水平为1-α的置信区间，若区间包含θ0则接受假设，否则拒绝假设。

反过来，一个区间估计问题：求分布参数θ的置信水平为1-α的置信区间，可以先求出上面假设检验问题的接受域，则接受域就是所求。

# 附录1 三种估计方法的比较

最大似然估计、最大后验估计和贝叶斯估计。

这三种方法都是基于给定的样本观察值的集合，估计总体的分布的具体形式。通常总体分布含有未知参数（设为），当分布的参数确定时，其具体形式就确定了，然后我们就可以用总体分布来计算任何给定的样本x的生成概率了。

1、最大似然估计

在第六章 《参数估计》中，已经介绍了极大似然估计方法，它属于点估计。这里为了本附录的内容的完整性，对其重述如下。

最大估计认为，总体分布的未知参数是一个未知的、但是固定的常量。我们无法知道真实的值，只能利用样本观察值集合T来求取某种准则下“最优”的参数作为真实的近似值。

这个准则就是“最大似然”。我们知道，样本观察值集合T是由样本联合分布生成的，由于T已经被观察到，也就是已经生成出来了，因此有理由认为，最优的相比其他而言，能够使联合分布生成T的概率最大：



又因为样本之间相互独立，所以上式可以写成：



对于具体的问题，我们会假设出总体分布的形式，带入上式，形成一个关于的最优化问题，求解之，得最优参数。

最大似然估计可以应用到机器学习模型中，就是条件分布的估计。设样本观察值集合为，模型即为条件分布，则应用最大似然法估计最优的目标为：



一般地，我们对最大似然目标函数取对数，称为最大对数似然。这能简化计算，且不影响最优的取值：



求出以后，就得到了总体分布的具体形式，则对任意给定的样本x，预测其生成概率为

2、最大后验估计

最大后验估计也属于点估计，它的目标也是求真实的的一个最佳近似值。但最大后验估计并不把总体分布的未知参数看作一个固定的常量，而是将其看作一个随机变量，我们要从这个随机变量的所有取值中求出一个最好的取值，作为真实的近似。既然是随机变量，那么最自然的想法就是，取哪个值的概率最大，哪个值就是最优的估计。则优化问题为：



根据贝叶斯定理，可将目标函数转化，得：



其中是的先验分布，在实际中常常是根据先验知识给出的，而不是基于样本集得到的。

目标函数中，分母相对于来说是一个常数，它不影响最优的取值，故可将其去掉：



再由样本的独立性，并取对数似然，得：



可以看出，最大后验估计相对于最大似然估计，就是多了一个的先验分布。

注意，在具体的分布估计的问题中，既可以使用最大似然估计，也可以使用最大后验估计。最大后验估计目标函数中的，与最大似然估计目标函数中的

，实际上拥有相同的具体形式，这里写法不同只是因为两种估计方法看待同一个问题的观点不同。

与最大似然估计一样，求出以后，就得到了总体分布的具体形式，则对任意给定的样本x，预测其生成概率为。

3、贝叶斯估计

贝叶斯估计与最大后验估计的相同点是，它们都把总体分布的未知参数看作一个随机变量，并且都考虑了的先验分布。

但贝叶斯估计与最大后验和最大似然估计都不同的是：它不属于点估计，并不是通过求取未知参数的最优近似值得到最优的总体分布，然后用总体分布预测给定样本x的生成概率。而是用基于数据集T的条件概率来计算给定样本x的出现的概率。

根据条件概率的计算公式，有：



注意，贝叶斯估计其实也给出了总体分布包含未知参数的形式，但不用它来直接预测给定样本x的概率（也不求最优的近似值），而是用条件概率来预测。

式中，是给定值，由分布形式确定的x的生成概率，则基于贝叶斯定理，由下式给出：



注意，我们知道，贝叶斯估计的计算式是根据条件概率公式得到的，但它还有另一种理解方式，即：对所有可能的，以作为权重，对加权平均的结果作为x产生的概率的预测x出现的概率。

# 附录2 对贝叶斯定理的一些理解

我们常说“贝叶斯定理是一个框架”，实际中有许多具体问题，如果能够从概率的角度上进行建模，就能够基于贝叶斯定理构建其具体的目标函数，进而求解。比如朴素贝叶斯算法、最大后验估计算法等。

贝叶斯定理的思想其实就是“反演”的思想。它其实就是反演的思想在概率模型中的体现。

“反演”是一种非常普遍的建模思想。实际中有许多问题，正着求解很简单，直接就能求解，但反过来求解就比较困难，不能直接求解。这时我们常利用正演来求解相应的反演问题。这里举几个正演-反演的例子：加法与减法是一对正演和反演过程；乘法和除法是一对正演和反演过程；给定函数f的形式和自变量x而求解y，这是正演过程，给定函数f的形式和y，求解x，这是相应的反演过程；求解机器学习模型最终都会转化成优化问题，优化问题实际上也属于反演。

优化问题的思路是：迭代地更新反演求解变量，每一次更新后，执行一次正演，看正演结果跟给定的输出之间的误差是否缩小到预期的范围内。

为什么说贝叶斯定理的思想是反演的思想呢？我们来看它的表达式：P(B|A)P(A)=P(A|B)P(B)，P(B|A)相当于是反演问题，难求；P(A|B)相当于是正演问题，好求。贝叶斯定理的思想就相当于通过求解P(A|B)问题来得到P(B|A)的结果，这不是反演是什么？

（这些理解还不是很成熟，需要继续思考，进行完善和改进）