第四章 随机变量的数字特征

每个随机变量都有一个概率分布，这个分布完整地刻画了随机变量的统计规律性.然而在许多实际应用问题中，人们更关注这个概率分布的一些综合特征，这些综合特征是概率分布某方面信息的概括并且可用一个数值表示.这种由随机变量的概率分布确定的，能刻画随机变量某方面特征的数值统称为随机变量（或其概率分布）的数字特征或特征数.

例如，考虑某种元件的寿命，如果知道了其寿命的概率分布，那么就把握了元件寿命的所有概率信息.比如,我们可以计算出寿命在任一指定范围内的概率.根据寿命的概率分布，我们还可以确定用以反映元件寿命的平均水平的特征数：数学期望，以及用以刻画寿命的散布程度的特征数：方差.这些特征数虽不能对寿命状况提供完整刻画,但却往往是人们最为关注的信息.无论在理论上还是在实际应用中，这些特征数都有着极重要的意义.尤其是在实际应用中,概率分布虽很“完美”,但难以准确把握;而特征数则容易把握,并且特征数是以一个“醒目”的数值刻画随机变量的某种特征,这使得应用更为方便.

本章我们介绍几个常用的数字特征：数学期望、方差、协方差和相关系数.

4.1 随机变量的数学期望

4.1.1 数学期望的定义

一个随机变量的数学期望是指用以刻划该随机变量取值的平均水平的一个数值.那么如何找到这样一个数值呢?我们用概率的频率解释来考虑这个问题.

假设试验有若干个结果,结果出现的概率为,若出现,则得到分.设表示一次试验中的得分,则的分布律为

，,

将试验独立重复次,记为次试验中结果出现的频数,那么这次试验的平均得分为

,

由频率的稳定性知,当试验次数足够大时,频率会接近于概率, 从而

，

与频率与概率的关系一样，上式左端是有波动的，右端是一个由的概率分布确定的一个数值.当试验次数无限增大时,左端会稳定于这个数值.由此可见,用数值刻划取值的平均水平是合适的,这个数值就称为的数学期望.于是有下面定义.

定义 设离散型随机变量的分布律为

，,

如果级数绝对收敛，则称级数的和为的数学期望，记为，即

，

若级数不是绝对收敛的，则称的数学期望不存在.

由以上定义可看出，若只取有限个值，则它的数学期望总是存在的.如果取可列无穷个值，它的数学期望不一定存在，是否存在就看级数是否绝对收敛.这个要求的目的在于使得期望值唯一.因为若无穷级数只是条件收敛，则可通过改变这个级数各项的次序，使得改变后的级数不收敛或收敛到任意指定的值，这意味着这个级数的和存在与否，以及等于多少，与的取值的排列次序有关，而作为刻画取值的平均水平的特征数，具有客观意义，不应与的取值的排列次序有关.

由定义，的期望就是其所有可能取值的加权平均，每个可能值的权重就是取该值的概率，因此的数学期望又称为的平均值或均值.同时还可看出的数学期望只依赖于的概率分布，因此随机变量的数学期望又叫分布的数学期望.

对于连续型随机变量,以积分代替求和,从而得到连续型随机变量的数学期望的定义.

定义 设连续型随机变量的密度函数为，如果

,

则积分值称为的数学期望，简称为期望或均值,也记为.若不收敛，则称的数学期望不存在.

注:期望这一概念可类比于质量分布的质心这一物理概念.把概率分布看作一个单位的质量在轴上的分布.在离散场合,概率看作点处的质量,那么该质量分布的质心坐标为，即为期望值.在连续场合,概率密度函数相应于单位质量的分布密度,质量分布的质心坐标为，也是期望值.质心是质量分布的一个中心，同样地期望是概率分布的一个中心.

例4.1.1 设随机变量的分布律为

 ,

的数学期望是否存在?

解 由于，即无穷级数不是绝对收敛的,故的数学期望不存在.

例4.1.2 按规定，某车站每天8:00～9:00,9:00～10:00都各有一辆客车到站，且到站时刻是随机的，两者到站时间相互独立，其规律为

|  |  |  |  |
| --- | --- | --- | --- |
| 到站时刻 | 8:10  9:10 | 8:30  9:30 | 8:50  9:50 |
| 概率 |  |  |  |

一旅客8:20到车站，求他候车时间的数学期望。

解 设旅客候车时间记为（单位：min）,则的分布律为

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 10 | 30 | 50 | 70 | 90 |
|  |  |  |  |  |  |

因此的数学期望为

(min).

例4.1.3 在一个人数很多的团体中普查某种疾病,需要抽血化验,可以用两种方法进行.(i)将每个人的血分别去化验,这需要化验次.(ii)按个人一组进行分组,把个人的血混在一起进行化验,如果这混合血液呈阴性反应,说明个人的血都呈阴性反应.这样, 个人的血液就只需验一次.若呈阳性,则再对这个人的血液逐一化验.这样, 个人的血液就共需验次.假设每个人血液化验呈阳性的概率为,且这些人的化验结果是相互独立的.试说明当较小时,选取适当的,按第二种方法可以减少化验次数,并说明取什么值时最适宜.

解 设个人为一组时, 个人共需化验的次数为,则的分布律为

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

这里.的数学期望为

,

由此可知,只需选取使得

,即,

便可使第二种方法减少平均化验次数.

要使最大程度地减少平均化验次数,需选取使得

,

小于1且取到最小值.

例如, ,则当时得到最好的分组方法.若,则按第二次方法平均化验次数为

,

这样平均来说,可以减少40%的工作量.

例4.1.4 随机变量的密度函数为



求.

解：

.

4.1.2 随机变量函数的期望

有了随机变量的概率分布，我们可以确定的数学期望。而计算的函数的期望也是经常遇到的问题.本身也是一个随机变量，有自身的概率分布，我们可以利用的概率分布计算.这种方法需要先导出的概率分布,但在许多场合这会比较麻烦,尤其是在涉及多个随机变量函数的场合.对于给定的函数,的数学期望完全取决于的概率分布，自然地我们会想到:能否直接利用的概率分布去计算的数学期望？下面定理回答了这个问题.

定理 设是随机变量的函数.

（1）设的分布律为，若绝对收敛，那么的数学期望为

.

（2）设的密度函数为，若，那么的数学期望为

.

这个定理的严格证明要涉及更多的概率论知识，在此省略了。

上述定理告诉我们,在求时,不必先计算出的概率分布,可直接利用的概率分布去计算的期望.这种方法对于求多个随机变量函数的期望尤其有价值.我们以两个随机变量函数的情形给出结论.

设二维随机向量的联合分布律为

,

若，则的数学期望为

.

设二维随机向量的概率密度为，若，则的数学期望为

.

特别地，若二维连续型随机向量的概率密度为,求分量的期望,或其函数的期望时，我们可以先求出的边缘密度,然后再计算或的期望.也可以直接利用的概率密度为:

, .

对于离散情形也类似,具体计算时就看哪种方法更方便.

例4.1.5 设随机变量的密度函数为

，

(1)问的数学期望是否存在?

(2)求.

解 (1)由于

所以的数学期望不存在.

(2) 



.

细心的同学可以发现，本例中随机变量既不是离散型随机变量，也不是连续型随机变量，但我们可以利用的概率密度求出的数学期望.

例4.1.6 设随机向量的概率密度为



求；.

解: ，

.

例4.1.7设随机变量独立同分布于指数分布，令，求.

解:的分布函数为



从而的分布函数为



故的密度函数为

,

所以

.

注意：一般而言，求的概率分布并不方便.但有些场合下，的概率分布可以方便地求出来，此时我们也可考虑先求出的概率分布，然后求.比如在为独立同分布的连续型随机变量时，他们的最大值，或最小值的概率密度常可容易地求出来.

4.1.3 数学期望的性质

由随机变量函数的期望的计算公式，可以得到数学期望的许多性质. 下面给出数学期望的几条重要性质.

数学期望具有如下性质（我们假定下面涉及到的期望存在）.

1.设是常数，则.

2. 设是常数，则.

3. .

这一性质可推广至任意有限个随机变量的线性组合的情形,

.

（4）若相互独立，则

.

这一性质可推广至任意有限个相互独立的随机变量之积的情形.

下面就是连续型随机向量情形给出性质（3）和（4）的证明.

证明 假设是连续型随机向量，其概率密度为，





.

若相互独立，则，从而







.

例4.1.8 设独立同分布，且，令,求(1)的期望;(2) 的期望.

解：(1)..

（2）.

例4.1.9 将一骰子掷次,求次的点数之和的数学期望.

本例中,如果去求的分布律将会很困难,即使求出了其分布律,再求期望时,其运算量也是非常大的.如能将表示为一些简单随机变量之和,就可以极大地简化计算.

解 令表示第次的点数,,则

,

且



由期望的性质可得

.

补充内容

思考题

1.设为取非负整数的随机变量，且期望存在,证明

(1)，或,

(2),或.

2. 设为取非负实数的连续型随机变量，且期望存在,证明

(1)，

(2) .

注意,可以验证对于取非负整数值的离散型随机变量,思考题2的结论也成立.更进一步,思考题2的结论对任意非负随机变量都成立.这里以非负连续型随机变量的情形来证明这两个结果只是为了我们可以利用微积分的知识.同学们不妨利用思考题2的结论,对指数分布的期望算一遍.

补充

1.数学期望的一般记号.本节中我们只给出了离散型和连续型随机变量的数学期望的定义,对于一般的随机变量的数学期望的定义,由于涉及较多的数学知识,这里不予讨论,下面给出期望的一个形式记号:

,

其中为的分布函数,上式右端的积分这里不解释了,只是希望同学们以后能认识这个记号.

2(票券收集问题) 一个盒子中装有标上至的张票券,以有放回方式一张一张地取.如果想收集到张不同的票券,所需的抽取次数记为,求的期望.

解:记依次表示对一张新票券的等待时间(次数),即表示在得到第张新票券后到得到第张新票券所需的抽取次数(.而,那么



由于服从参数为的几何分布,所以

,,

故





下面看两种特殊情况:(假设为偶数), ,并假定足够大.

时,

,

时,

.

可见收集到一半票券平均需要抽取次数大约是总票券数的70%,而收集全部票券,则平均需要总票券数的倍的抽取次数,可见越往后越难以收集.

3.（快速排序算法）设有一组互不相同的数.将它们排成上升的序列.一种快速排序算法如下:随机地从中选一个数,设为,然后将其余的数都与作比较,将小于的数归入的左边一个集合,将大于的数归入的右边一个集合,然后再对左、右两个集合的数重复刚才的处理过程(如果集合是单点集就不用处理).直到把所有数排成上升序列为止.

和通常的排序算法的分析一样，用算法中比较的次数来衡量运行时间.随机排序算法中的比较次数是随机变量,自然地要分析比较次数的期望.记表示为实现排序所需的比较次数,是排序算法效率的一个度量,下面计算.

先将最小的数命名为，第二小的数命名为，…，最大的数命名为，对于，令



则.

由于，

故

且～.

注:随机排序算法尽管在运行中有随机性但总能给出正确的结果,这样的算法称为

Las Vegas算法.在第一章中,由蒲丰投针问题而引出的随机模拟方法则属于Monte Carlo算法. Monte Carlo算法特点是可以保证有很高概率输出正确结果，出现错误结果的概率非零.但可以以多花费运行时间为代价,做到失败概率尽可能小.

4.在内连续地随机取数,若第2次取出的数小于第1次取出的数就停止取数，否则再第3个数;若第3次取出的数小于第2次取出的数就停止取数, 否则再第4个数.按如此方式下去直至停止,表示取数的次数,求.

利用思考题1的结论来求解.

设分别表示第1次, 第2次, 取到的数.

，

，

……,

,

……,

注意到

,

,

,

,

……,

所以

