

# 概率论与数理统计

该手册为面向计算学部同学的《概率论与数理统计》的开箱手册。

小建议：这门课的内容最好能够在通过考试或考高分基础上能够进一步的理解，这有助于之后选择机器学习，人工智能的同学的学习。

**该手册仅仅作为开箱手册用于同学对于该门课有一个初步的认识和理解，细节部分以及更加深入部分请同学上课认真听讲或翻看书籍。考虑到学习知识的时效性以及书写手册的可能遗漏的错误，一切与教材，老师讲课内容不符合的地方请与老师交流，然后以老师为准。**

这门课包含两个部分，概率论部分，与数理统计部分。

## 一、 概率论部分

**知识点1：经典的概率模型,以及概率的性质，概率的相关公式**

1) 古典概率，几何概率，条件概率

古典概率  $P(A) = \frac{A \text{ 所含样本点数}}{S \text{ 所含样本点数}}$

几何概率  $P(A) = \frac{A \text{ 几何度量}}{S \text{ 几何度量}}$

条件概率  $P(A|B) = \frac{P(AB)}{P(B)}, P(B) > 0$

2) 概率的性质

如  $0 \leq P(A) \leq 1$  等性质，其余性质可自行查阅PPT或书籍

3) 乘法公式，全概率公式，贝叶斯公式

乘法公式就是描述了观察AB两个事件同时发生的状态，通过用A事件，和在A发生情况下B发生的概率来描述

全概率公式，也是描述性质的公式，

贝叶斯公式，这个公式非常重建议而且经典。最一般的写法为， $P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^n P(A_j)P(B|A_j)}$ ，把  $A_i$  代入，然后再把下面的  $P(A)$  展开即为常见的结果了。贝叶斯公式在之后的机器学习里非常重建议。

**知识点2：事件的独立性**

这个知识点侧重于理解事件，事件之间的关系，怎么描述事件是不是独立的。

建议掌握到，给出一个概率的公式观察上的结果，判断其中的事件是否是独立的。

另外建议注意到相互独立能够推出两两独立，而两两独立不能推出相互独立。

从公式上来看，区别在于是否能够满足  $P(ABC) = P(A)P(B)P(C)$

从理解角度来看就是，A和B之间没有影响，A和C之间没有影响，B和C之间没有影响，但是有可能AB（BC，AC同理）对于另外一个有影响

### 知识点3：随机变量，分布列，分布函数，概率，概率密度

这里首先建议理解什么是随机变量，然后理解什么是离散型随机变量，什么是连续型随机变量，在此基础上，掌握描述这些随机变量的方式。

对于离散型随机变量，建议会求分布列，分布函数和概率

建议掌握常用的离散型分布列，如两点分布，二项分布，泊松分布，几何分布等。

对于连续型随机变量，建议理解分布函数，概率密度，以及相互转化的方法

概率密度不同于其他的函数的本质区别在于，它满足了这样的约束  $f(x) \geq 0$  ,  $\int_{-\infty}^{+\infty} f(x)dx = 1$  , 计算题里面常用这样的约束来进行计算，也可以用这个性质来判断这样的函数是不是对应于一个概率分布

建议掌握的常用的分布函数，均匀分布，指数分布，正态分布，标准正态分布等

### 知识点4：掌握一维随机变量函数的分布，二维随机变量函数的分布

从一维随机变量函数本身出发，可以容易发现，可能的求解方式：

有分布函数->概率密度，这个适用于当分布函数比较好求解，但是概率密度不容易直接求解的时候有借助另外的已知或间接可知的变量的概率密度来求解。

具体细节见书或PPT

对于二维随机变量函数，建议掌握一些相关性质，详见书

比如：

$$P(x_1 \leq x \leq x_2, y_1 \leq y \leq y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \geq 0$$

联合分布函数能够确定边缘分布函数，但是只通过边缘分布函数是不足够确定联合分布函数的。在增加一些条件之后，才能通过边缘分布函数确定联合分布函数。

对于二维离散型随机变量，建议掌握定义，性质，分布列的书写，以及一些运算和性质，边缘分布列以及求解，条件分布列及其求解。

对于二维连续型随机变量，建议掌握定义，性质，边缘概率密度及其求解，条件分布函数，条件概率密度及其求解。

建议掌握二维均匀分布，二维正态分布的分布函数，概率密度函数。

想建议求解二维随机变量函数的分布，离散型：常见的题型为给出一个已知的概率分布，求解一个有相关关系的概率分布，这类题型，把每个取值的概率求解出来即可。连续型：利用分布函数求解概率密度是常见的做法。但是如果此时不方便

求解分布函数，但是给出了比如  $Z = aX + Y$  这样的关系，而且给出了  $f(x), f(y)$ ，这样就可以考虑用这个关系来求解  $f(z)$ ， $f_Z(z) = \int_{-\infty}^{+\infty} f(x, z - ax) dx$ ，特别的，当  $a=1$ ， $X$  与  $Y$  独立时，有卷积公式，详见书。

$Z = f(X, Y)$ ，可以熟悉一些常见的条件，比如

$Z = XY$ ， $Z = Y/X$ ， $Z = \max(X, Y)$ ， $Z = \min(X, Y)$  的结论

## 知识点5:对于随机变量的特征的理解

这里主要想讲的是一些理解上的东西。

谈到一个随机变量，我们会先考虑到它取值不一定是固定的，秉承一个从粗略到细致的刻画思路，我们能够接下来怎么慢慢摸索这个随机变量的性质呢？我下面举一个例子来更好的讲述这个事情怎么引入的。

比如说一个公司A每个月都会有盈利，这个盈利额是一个随机变量，那投资人想建议为这公司下一、年的投资做一个规划，它就想建议知道，这个公司大概每年能够获利多少呢，那这个就是数学期望。但是光有数学期望，这个还不够，为什么呢？因为可能会出现，1-4每个月盈利2万，5-8每个月盈利5万，9-12月盈利8万，而我们此时有另外一个公司B，它1-12月每个月盈利都是5万，那么对于投资人而言，这件事就不一样了，很显然B和A一年的盈利虽然不一样，但是他们的盈利的稳定性是不一样的，那这里就引入了方差。我们进一步考虑，如果盈利这个事情和别的事情有了一些关系，我们想建议描述这个事情怎么办，比如和这个公司盈利有关的比如用工成本的浮动，比如租房成本的变化，市场销量也肯定有关系，盈利和这些因素都会有一些关系，但是我如果只用均值和方差是不够的，那么我就需考虑盈利和这些因素的相关程度。而如果说方差是用来衡量一个变量的采样中样本值的偏离均值的情况的话，那引入协方差就能够衡量一个样本的值的偏离，会对另外一个样本的值的偏离产生多大的影响。这样引入协方差和相关系数也就是比较自然的事情了。对于一些经典的二维分布，比如二维正态分布，是有很强的研究意义的，事实上在后面的机器学习学习中，还会再遇到它的，至于在哪遇到留作一个彩蛋。

而谈到原点矩和中心矩的提出，一定程度上是为了能够用这样的一个特征，考量了随机变量全部的取值情况，并且能够比较合理的描述随机变量的某种角度的分布情况。

直观上认为如果我试验的越多，那么随机变量将会更加趋近于其期望，切比雪夫不等式刻画了这一件事，而切比雪夫大数定律，则进一步的刻画了相互独立的随机变量序列的情况。辛钦大数定律刻画的是独立同分布的随机变量序列的情况，伯努利大数定律是刻画的n重伯努利试验中的情况。

## 二、数理统计部分

### 知识点6：数理统计的定义，研究对象，意义，常见概念

数理统计的研究对象的全体叫总体，总体又分为有限总体和无限总体。而总体中的每个成员叫做个体。

理解简单随机样本，样本容量，样本值，样本空间，统计量的概念或定义。

注意常用统计量，样本均值，样本方差，样本标准差，样本k阶原点矩，样本k阶中心原点矩，顺序统计量等

### 知识点7:数理统计的三大分布

三大分布为  $\chi^2$  分布，t分布，F分布，具体公式详见教材。

$\chi^2$  分布的性质用起来比较多，最好注意一下

上侧  $\alpha$  分位数或临界值，设  $\alpha(0 < \alpha < 1)$ ，称满足等式  $P(T \geq t_\alpha(n)) = \alpha$  的点  $t_\alpha(n)$  为t(n)的上侧  $\alpha$  分位数或临界值。直观理解的话，就是我们给出了一个概率分布，那么我想要得到一些重要的信息，比如说我们发生一个事情或者不发生一个事情的概率为  $\alpha$  的时候，我们的随机变量是什么样的一个情况。注意一些简单的变化，比如发生一个事情和不发生该事情的概率加起来为1，以及一些微积分，代数和概率公式上的变换，基本上这部分的习题就不难做了。

### 知识点8：单个正态总体统计量的分布

1.样本均值的分布，及其推论

2.样本方差的分布，及其推论

3.样本均值和方差，

$$\sqrt{(n)} \frac{(\bar{X} - \mu)}{S} \sim t(n-1) \text{ 以及相关的一些结论}$$

### 知识点9：参数估计

这部分内容在之后的学习工作中如果用到了的话，就会发现它非常有用。

对于一个统计总体，总体的分布函数为  $F(x, \theta)$ ，其中  $\theta$  为未知参数，这件事情，在后面的机器学习也会再次遇到的。比如当这个未知参数变得更加复杂，我们想要估计的更加精准一些，就需要一些更

加复杂的方法了，当然这门课里面不会涉及得那么深。

参数估计包括点估计，区间估计，点估计里面又包括矩估计和最大似然估计。

从结果上来看，点估计的结果是统计量  $\hat{\theta}$ ，区间估计的结果是  $P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$ ，也就是点估计是去对未知参数估计出来了一个结果，而区间估计是去估计这个未知参数的置信区间，这也符合这两个方法的名字。

矩估计是用样本矩估计总体矩，用样本矩函数估计总体矩函数，总体的前m阶原点矩可以用  $\theta_1, \dots, \theta_m$  表示，把这个方程式调整一下，转化到  $\theta = h(\alpha_1, \dots, \alpha_m)$  的形式，这个时候容易发现如果我们的总体矩的结果是没有问题的话，那么这个方程组的转化是没有带来偏差的，而我们如果还想要得到一个估计的结果，结合我们能够获得样本的总体矩，那么我们就可以用样本的总体矩替代这里的总体矩，从而获得一个对各个参数估计的结果。

最大似然估计的思想非常直接。考虑这样的一个直观想法，我们已经获得了一堆观察值，那么我们拿不同的参数代入进入未知参数的话，就会发现这些观察值出现的概率也不一样。在这样的情况下，我们还想要一个比较科学的可信的结果，那我们就可以期待能够让这些观察值出现的概率越高的参数越接近我们想要估计的未知参数。为了衡量这个结果，我们就引入了似然函数，和最大似然估计。所以似然，顾名思义，就是“像这回事”，“好像是对的”。

继续学习机器学习的同学将会在以后的学习中遇到一个叫做最大后验估计（MAP）的估计方法，也同样非常有意思。

求最大似然估计的步骤为写出似然函数，一般需要对似然函数取对数，如果似然函数十分简单而不需要取对数当然也可以不取，建立似然方程，解似然方程，从而求解最大似然估计。另外注意如果这里的似然函数不可以求微分的话，就需要用定义(求max),来做了。

当然我们获得了估计量之后，我们还需要对其进行评价，这里引入的评价标准为1.无偏性2.有效性3.相合性

简单理解的话，无偏性是期待我们通过这些样本来求到的参数的期望和未知参数是一致的，

即  $E(\hat{\theta}(X_1, X_2, \dots, X_n)) = \theta$ ，这样称  $\hat{\theta}$  为  $\theta$  的无偏估计量，比如样本均值和方差是总体均值和方差的无偏估计。

有效性，简单理解的话就是对于同样的一个总体参数的两个无偏估计，具有更小的标准差（ $\theta_1 \leq \theta_2$ ，且至少对于某个  $\theta_0$ ，取得小于号）的参数更有效。

相合性反应了我们这样的一种期待，随着样本容量越来越大，我们估计参数的结果越来越接近被估计的总体参数。

对于区间估计，我们期待能够得到这样的结果，区间越小，能够相信的程度越高，这样的结果越好。反映出来就是

置信空间，和置信度。单个正态总体参数的置信空间详见书。

## 知识点10：假设检验

这里涉及的是“小概率原理”，也是符合直觉的。一个事件的发生概率很小（小概率事件），那么它在一次试验中是几乎不可能发生的。那么我们对于想要检验的东西就可以构造这样的一个假设检验的方法了。先假设  $H_0$  是正确的，再构造一个小概率事件A，如果我们在一次抽样检验中，事件A发生了，则拒绝  $H_0$  ,否则接受  $H_0$  。

进一步的我们会发现我们做假设检验的时候，可能会犯错，考虑这样两种错误，第一种是原命题是真的，但是我们拒绝了原假设。第二种，原假设是错的，但是我们却接受了原假设。需要注意到这两类错误是相互制约的，直观理解就是随着我们试验的增加，我们犯某类错误的概率降低的同时，我们可能比原来更容易犯另外一类错误。

单个正态总体参数的总体检验（u检验，t检验， $\chi^2$  检验）详情见书。

到了这里就基本把这门课的主要内容和主干讲完了，该讲义仅作为开箱手册，用作帮助同学初步了解课程内容目的，考虑到学习知识的时效性以及书写手册的可能遗漏的错误，一切与教材，老师讲课内容不符合的地方请与老师交流，然后以老师为准。