大数据分析的数学原理学习笔记

本文为夏志宏老师所授课程数据分析的数学原理的学习笔记，为方便同学们学习、吸收老师所讲的知识点。

一、随机并非任意，随机系统的确定性

第一讲 随机系统的确定性

随机并非任意，随机系统的确定性的案例：

1. 投硬币的学问。

掷硬币200次，其结果为

(01001101011001100101001011001011010010000101010010011000001101010011010101010010111010001101010100000100110101100110010100101100101111001000010101001001100000110101001101010101001011101000110101010000)

1. 89个1， 111个 0；可信度低于1%！
2. 99 个1； 101个 0，好像没问题。但是 ，111出现过3次 ； 而1111一次都没有出现；
3. 只出现三次111的可信度低于千分之一，一次都没出现1111的可信度低于1%！

二、如何发现论文数据造假



一组实验数据，从某实验室得来（物理、化学、生物、社会、…），怎么发现数据有没有问题

1. 40个数字里没有一个最后数字为0！
2. 倒数第二位没有一个1！
3. 前面几位数字因其物理意义不会是随机的，而末尾几位数是很随机的！

当实验数据量很大的时候，打假会更精确。

三、随机数据的确定性

在原子、分子层次，物质有很大的不确定性

比如气体的每个分子的运动都是不确定的，但整体确有惊人的确定性！

按照遍历论，这种现象不仅可能而且一定会发生：

1. 房间的分子全部跑到左边，右边形成真空，
2. 但概率实在太小，需要等很长时间才会发生。
3. 实际情况下，可以说不可能发生。

100万次硬币，在1.3%（三个sigma）的置信度下，正反面的平均值误差不会超过千分之1.5。

四、随机问卷保护隐私

防疫部门需要统计某传染病发病情况，问卷调查对象不愿如实回答（隐私）。

可以给每个答卷者一个骰子，在回答问题前自己投骰子。如果骰子是1、2、3、4，如实回答；如果骰子是5、6，撒谎。

1. 收到答卷者并不知道每个人具体情况，因为答案可能是假的（保护隐私）。
2. 整体统计数据可以非常精确。

练习：

3万份问卷里有1万2千人回答有传染病，真实应该是多少人？你感觉误差会有多大？

30万份问卷里有12万人回答有传染病，真实应该是多少人？你感觉误差会有多大？

第二讲 随机变量与概率分布

1.1 随机变量

设试验E的样本空间为Ω，如果对于每一个样本点ω∈Ω，都有一个实数X (ω)与之对应，则称定义在Ω上的单值实值函数X (ω)为**随机变量**，简记为X . 通常用 X ,Y, Z 等表示随机变量.

**概念解析：**

**随机事件**：抛硬币得正、反两面,骰子的1、2、3、4、5、6,生男生女。

**随机函数**: X(x)依赖于随机事件的函数。函数值可以是在任何空间。比如骰子1、2撒谎;骰子3、4、5、6说真话；比如硬币正面吃川菜,反面吃粤菜: X(正面)=川菜，X(反面)=粤菜。

**期望值**:多次试验平均所得（平均值又称数学期望）。

1.1 概率分布

随机事件分两种：

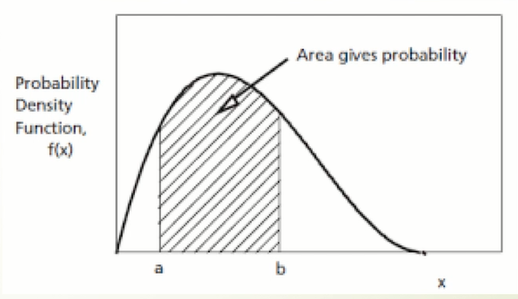
离散:n种可能，每种可能有个概率，第j种的概率是pj ≥ 0， 当然

1. 公平硬币正反面概率都是0.5
2. 公平骰子,每个数字概率都是1/6
3. 生男孩0.5，女孩0.5

连续: 如果随机事件取值为任意实数,我们用概率密度函数( PDF )来描述

如图，在[a, b]区间概率为一个积分E(x) =

这里的f(p)≥0为概率密度函数，当然， = 1



离散和连续的区别就在于和

**特征计算：**

**随机函数X(x)期望值为加权平均：**

* **E(x) =** （离散概率分布）
* **E(x) =** （连续概率分布）

**方差：**

* 期望值代表平均,但有些分布比较集中,有些分布广,这个特性可以用方差来衡量。
* 方差 **Var**(X) = **E**([ **X** - **E**(**X**)]²)
* 方差 = 取样与平均值之差的平方的期望值

**标准差：**

* **σ =**

案例

一、测试你对概率论的直觉

小王在香港机场赶飞机,她需要在24分钟内通过海关，海关有两个通道,一个通道前有10人在排队,每个人通关时间不同,但平均需要2分钟;第二个通道有两个服务窗口,前面共有20人在排队,同样每个人通关时间平均为2分钟(对任一服务窗口)

* 小王应该选择哪个通道?（20）
* 假如小王必须在18分钟内过关,小王应该选择哪个通道?（10）

二、 考试成绩对比

假如本课程有两次期中考试,第一-次和第二次考 试平均分均为80。小张第一次考80，第二次考70；小李第一次考70，第二次考80。

* 小李、小张成绩一样?（不能直接确定）
* 如果第一次标准差是10 ,第二次标准差是20，小李成绩好还是小张成绩好？（小张，标准差代表了全班同学的水平，由于第一次成绩的标准差小于第二次的标准差，标准差越小，成绩的离散性越小，小李第一次少考10分的排名比小张第二次少考10分的排名距离平均值的排名更靠后的）

三、再次测试对概率论的直觉

小李在公司表现优秀,老板决定给一个年终红包。老板手上拿着两个红包,告诉小李其中一个红包是另外一个红包的两倍。老板先随意给小李其中一个,小李打开红包一看,里面有一张一万元银行卡。老板问:想不想换成另外一个？

您说小李该换、还是不该换？

第三课 大数定理与正态分布

1.2 大数定理

概率论历史上第一个极限定理属于伯努利，后人称之为“大数定律”。概率论中讨论随机变量序列的算术平均值向随机变量各数学期望的算术平均值收敛的定律。

**实例描述**

1. 每次扔骰子得出1, 2, 3, 4, 5, 6中的一个数，平均值为3.5
2. 扔一次的误差为0.5到2.5不等,平均误差1.5
3. 扔两次的总数可能是2, 3, ...12.平均值还是3.5;误差是0到2.5不等,但平均误差低于1
4. n次的平均值还是3.5，误差最大还是2.5，但平均误差随着n增加而下降

**定理解析：**

假定X1, X2, … 是同一分布的随机变量，E(X1) = E(X2) = … = μ，平均值

收敛于期望值

方差为

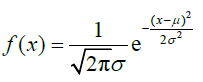
从另外一个角度看

（Pr是指概率）

一句话概括：随着样本容量n的增加，样本平均数将接近于总体平均数。

1.2 正态分布

正态分布（Normal distribution），也称“常态分布”，又名高斯分布（Gaussian distribution）。正态曲线呈钟型，两头低，中间高，左右对称因其曲线呈钟形，因此人们又经常称之为钟形曲线。



若随机变量X服从一个数学期望为μ、方差为σ^2的正态分布，记为N(μ，σ^2)。其概率密度函数为正态分布的期望值μ决定了其位置，其标准差σ决定了分布的幅度。当μ = 0,σ = 1时的正态分布是标准正态分布。

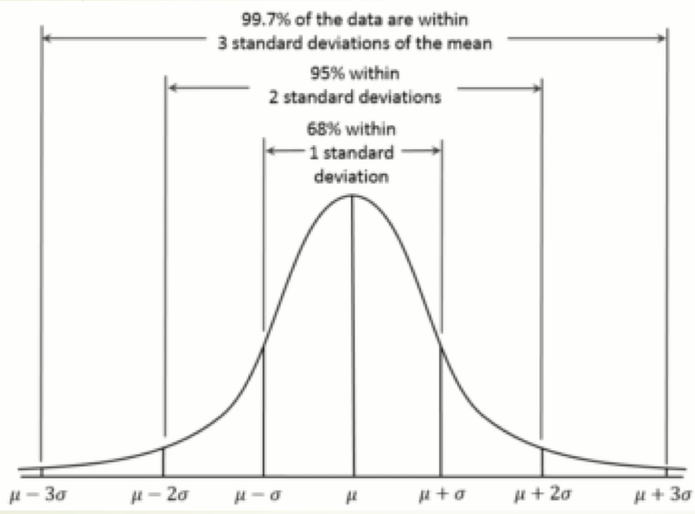
**正态分布3σ原则为**：

数值分布在（μ-σ,μ+σ)中的概率为0.6826

数值分布在（μ-2σ,μ+2σ)中的概率为0.9544

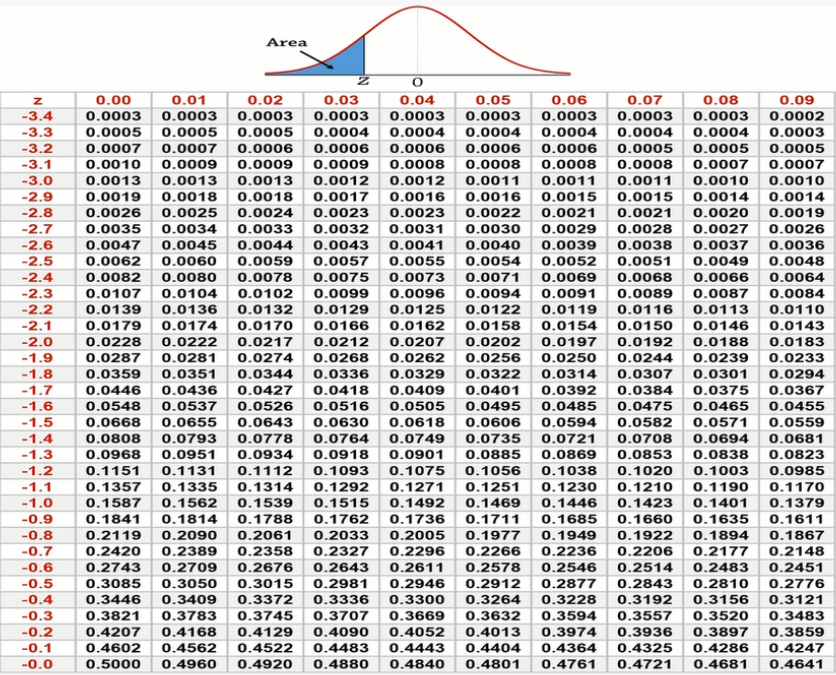
数值分布在（μ-3σ,μ+3σ)中的概率为0.9974

可以认为，Y 的取值几乎全部集中在（μ-3σ,μ+3σ)区间内，超出这个范围的可能性仅占不到0.3%。



**标准正态分布表（Z值表）**

标准正态分布表 (n(0,1)，Φ（x），通过查找实数x的位置，从而得到p(z<=x))。表的纵向代表x的整数部分和小数点后第一位，横向代表x的小数点后第二位，然后就找到了x的位置。（可由概率反推x）



第四讲 中心极限定理的应用

*本讲主要讲解中心极限定理以及如何应用它解决问题*

**中心极限定理**

中心极限定理 是指概率论中讨论随机变量序列部分和分布渐近于正态分布的一类定理。这组定理是数理统计学和误差分析的理论基础，指出了大量随机变量近似服从正态分布的条件。它是概率论中最重要的一类定理，有广泛的实际应用背景。在自然界与生产中，一些现象受到许多相互独立的随机因素的影响，如果每个因素所产生的影响都很微小时，总的影响可以看作是服从正态分布的。中心极限定理就是从数学上证明了这一现象。最早的中心极限定理是讨论重点，伯努利试验中，事件A出现的次数渐近于正态分布的问题。

*伯努利试验（Bernoulli experiment）是在同样的条件下重复地、相互独立地进行的一种随机试验，其特点是该随机试验只有两种可能结果：发生或者不发生。我们假设该项试验独立重复地进行了n次，那么就称这一系列重复独立的随机试验为n重伯努利试验，或称为伯努利概型。单个伯努利试验是没有多大意义的，然而，当我们反复进行伯努利试验，去观察这些试验有多少是成功的，多少是失败的，事情就变得有意义了，这些累计记录包含了很多潜在的非常有用的信息。*

- 一个随机变量不管是连续分布还是离散分布，在取了平均值以后他的分布都符合**正态分布**

- 标准化以后则符合标准正态分布

- **条件是数据量足够大**

**用中心极限定理估计**

扔10000次硬币平均值的误差

1.正面出现4900次，可信度有多大?

2.平均值应该是μ =0.5,现在为0.49，0.49-0.5 =-0.01

3.假定正态分布,标准差为σ=0.5/100 = 0.005

4.离平均值两个σ，查表得0.028, 可信度在5%之内。

**假如三个σ，可信度只有千分之3之内**

1.n多大才可以用正态分布?

2.一般收敛速度大概是 ~ n的-1/2次方

**数据造假**

*如何发现数据造假行为*

1.随机数据有规律,而造假往往破坏了内在规律

2.数据越大，规律越强

3.大量数据分组以后，每组都有同样的规律

4.大数据分析常用方法是预留部分数据作为验证数据,对训练数据的结果进行验证

5.实验室的数据一般可以假定为正态分布，精度(有效数字)决定了标准差。在有效数字之外基本可以认为是随机的

6.在数字量不大的情况下，可以考虑二进制(奇数、偶数...) ，更有统计意义

**习题**

1. 一篇学术论文有500个数据 ,每个数据精确到6位，而有效数字为3位。倒数第二位数字偶数有230 ,奇数有270。你觉得可信吗?估计一下可信度有多少。

2. 传染病估计（第一讲留下的问题）

- 3万份问卷里有1万2千人回答有传染病,真实比例估计是多少?算一下误差会有多大?

- 30万份问卷里有12万人回答有传染病,真实比例应该是多少?算一下误差会有多大?

3. 圆周率π的随机性

- 在网上找出圆周率π精确到100万位的数值。用四位数表示你出生的月日，你的生日在π出现过多少次?

- 如果π中每个数字的出现是随机的,你的生日应该出现多少次?

- 随机取一个三位数,在r里出现多少次?

**结论**

- 随机数据有规律，数据越大规律越强，结果越可靠!

- 大数据分析就是要找出数据规律。

部分参考答案

第一讲

**一、投硬币问题**

源代码见附件

在统计学中，一个概率样本的置信区间（Confidence interval）是对这个样本的某个总体参数的区间估计。置信区间展现的是这个参数的真实值有一定概率落在测量结果的周围的程度。置信区间给出的是被测量参数的测量值的可信程度，即前面所要求的“一定概率”。这个概率被称为置信水平（即置信度）。

按中心极限定理，扔很多次后，单个H（1）、T（0），每个占50%；00、01、10、11，各占25%；HHH （111）占0.125。

以HHH为例，HHH的次数服从正态分布均值为0.125。标准差std ≈ 0.033 ，则x = （3/197 – 0.125）/ 0.033 = -3.3264，置信度 C = 1-2×Φ（x）= 0.9987 。

**习题**

**患病人数**：由已知得，说真话的概率为2/3，说假话的概率为1/3。设患病人数占总人数的比例为p，则健康人的比例为（1-p）。统计结果中，30000人中有12000人为患病人，比例为40%。建立等式：40% = 2/3 × p + 1/3 × （1-p）。易得p = 0.2 。则真实患病人数大致为30000 × 0.2 = 6000人；不用概率论的方法：我们要求的是，已知 （其中Xi为患病人），易得 ，设患病者概率为p概率，，易得p = 0.2 。

**误差**：已知， ， ，则3σ的置信区间端点计算为(标准正态)，则

第二问同理。

第二讲

**一、排哪个队**

A窗口两通道20人人均2分钟，B窗口单通道10人人均2分钟。假设这样的排队排N次，A窗口排队时间记为*A*分布，B窗口排队时间记为*B*分布,由大数定理，*A*分布与*B*分布收敛于正态分布，平均值相等（都是20分钟），但*A*分布比*B*分布的方差更小。

24分钟情况：由于*A*分布方差更小，可能出现大于20分钟的可能性更小，固选A窗口。

18分钟情况：由于*B*分布方差更大，更可能出现小于20分钟的情况，故选B窗口。

**三、该不该换红包**

不换红包的收益为10000，换红包后的收益，有1/2概率为5000、1/2概率为20000，则换红包以后，有1/2概率亏损5000，也有1/2概率盈利10000，则1/2 \* 10000 – 1/2 \* 5000 = 2500 > 0，换句话说，换红包是偏向盈利方向的。从经济学中的理性人假设可得应该换。

**附件一**：

丢硬币源码

#!/usr/bin/python  
# -\*- coding: utf-8 -\*-  
# @Author: Jaylin  
  
import numpy as np  
from scipy.stats import norm  
  
  
def simple\_gen(num=10000, count=200):  
 *"""  
 样本生成  
 0 为 T， 1 为 H  
 返回'HHH'与'HHHH'的概率* ***@param*** *num:样本数，默认为10000* ***@param*** *count:掷硬币次数，默认为200  
 """* simples = []  
 for k in range(num):  
 simple\_cell = [0, 0] # 初始化样本单元，[三连出现的次数， 四连出现的次数]  
 toss = np.random.choice([0, 1], count) # 投掷200次  
 # print(toss) # Debug code  
 for i in range(len(toss)):  
 if i < len(toss) - 3:  
 sum\_4H = toss[i] + toss[i + 1] + toss[i + 2] + toss[i + 3]  
 if sum\_4H == 4:  
 simple\_cell[1] = simple\_cell[1] + 1  
 if i < len(toss) - 2:  
 sum\_3H = toss[i] + toss[i + 1] + toss[i + 2]  
 if sum\_3H == 3:  
 simple\_cell[0] = simple\_cell[0] + 1  
 # print(simple\_cell) # Debug code  
 simples.append(simple\_cell)  
 \_simples = [[], []]  
 for i in simples:  
 \_simples[0].append(i[0] / 200)  
 \_simples[1].append(i[1] / 200)  
 # \_simples[0].append(i[0])  
 # \_simples[1].append(i[1])  
 # simples为[[a, b], [a, b], [a, b]]  
 # \_simples为[[a, a, a], [b, b, b]]  
 # 取你喜欢的一个return  
 return \_simples  
  
  
def calculate(simple):  
 *"""  
 计算样本的平均值，方差，标准差，标准误差* ***@param*** *simple: 待求解的数字数组  
 """* mean = np.mean(simple)  
 var = np.var(simple)  
 std = np.std(simple)  
 std\_err = np.sqrt(std / len(simple))  
  
 return [mean, var, std, std\_err]  
  
  
def DoC\_3(\_mean, \_std):  
 *"""  
 计算3个'HHH'的置信值* ***@param*** *\_mean: 样本空间标均值* ***@param*** *\_std: 样本空间标准差  
 """* print((3 / 200 - \_mean) / \_std)  
 return 1 - 2 \* norm.cdf((3 / 200 - \_mean) / \_std)  
  
  
def DoC\_0(\_mean, \_std):  
 *"""  
 计算0个'HHHH'的置信值* ***@param*** *\_mean: 样本空间标均值* ***@param*** *\_std: 样本空间标准差  
 """* print((- \_mean) / \_std)  
 return 1 - 2\*norm.cdf((1 / 200 - \_mean) / \_std)  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 simples = simple\_gen(10000, 200) # 样本空间  
 [H\_3, H\_4] = [calculate(simples[0]), calculate(simples[1])] # 得到计算结果  
 # print(simples)  
 # print(H\_3, H\_4)  
 E1 = DoC\_3(H\_3[0], H\_3[2])  
 E2 = DoC\_0(H\_4[0], H\_4[2])  
 # print(H\_3[0] / H\_3[2])  
 # print(H\_4[0] / H\_4[2])  
 print(3/200)  
 print("出现三连H的平均值：", H\_3[0])  
 print("出现三连H的标准差：", H\_3[2])  
 print("出现四连H的平均值：", H\_4[0])  
 print("出现四连H的标准差：", H\_4[2])  
 print("出现三次三连H的置信值是：", E1)  
 print("出现一次四连H的置信值是：", E2)