Contents

[《数论》 1](#_Toc45287153)

[基本概念 1](#_Toc45287154)

[算法 2](#_Toc45287155)

[定理 3](#_Toc45287156)

[<Statistics>统计学 5](#_Toc45287157)

[基本概念 6](#_Toc45287158)

[统计推断 7](#_Toc45287159)

[变量间关系 8](#_Toc45287160)

[多元分析 9](#_Toc45287161)

[如何生成概率分布？ 9](#_Toc45287162)

[R语言统计 10](#_Toc45287163)

# 《数论》

## 基本概念

* 倍数和约数(因数)，公倍数和公因数

最小公倍数(两个或多个整数的公倍数里最小的一个)

最大公因数(两个数的公因数中最大的一个)

a称为b的倍数，b称为a的约数

* 质数（素数）：只有1和它本身两个因数的自然数

100以内共25个质数：2 3 5 7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79 83 89 97...

* 合数：除了1与本身以外，还有其他约数

{质数} + {合数} + {1} = 正整数

* 孪生素数就是差为2的素数对，例如11和13
* 真约数，真因子：除了自身以外的约数
* 完全数，完美数，完备数: 真约数之和恰好等于自已

如：{6, 28, 496, 8128, …}, 截至2018年，相关研究仅找到51个完全数

6 = 1+2+3

28 = 1+2+4+7+14

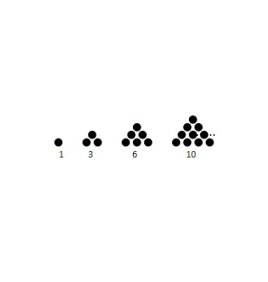
496 = 1+2+4+8+16+31+62+124+248

8128

* 亲和数(a,b)： a的真约数之和 = b， 且a = b的真约数之和

如： (220 284), (1184 1210), (2620 2924), …

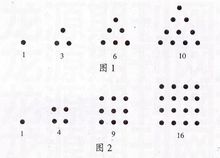
* 三角形数:一定数目的点或圆在等距离的排列下可以形成一个等边三角形



第n个三角形数的公式是n(n+1)/2 或者 [(2n+1)2 – 1]/8

如：{1, 3, 6, 10, 15, 21, 28, 36, 45, 55, 66, …}

* 平方数，正方形数: 平方根为整数的数



如：{1, 4, 9, 16, 25, …}

## 算法

* 最大公因数与最小公倍数求法？

短除法

最大公因数 = 2\*3 （除数乘积）

最小公倍数 = 2\*3 \* 2\*3 （除数与商乘积）

* 欧几里得算法

假定a>b, 令b = r0

a = r0 \* q1 + r1

r0 = r1 \* q2 + r2

直至整除为止，最后一个非零余数即为最大公因数

如

8251 = 6105\*1 + 2146

6105 = 2146\*2 + 1813

2146 = 1813\*1 + 333

1813 = 333\*5 + 148

333 = 148\*2 + 37

148 = 37\*4

则最大公因数为37

* 判断倍数？

2的倍数：　一个数的末尾是偶数（0，2，4，6，8）

3的倍数：　一个数的各位数之和是3的倍数

4的倍数：　一个数的末两位是4的倍数

5的倍数：　一个数的末尾是0或5

6的倍数：　一个数只要能同时被2和3整除

8的倍数：　一个数的末三位是8的倍数

9的倍数：　若一个整数的数字和能被9整除

10的倍数：　若一个整数的末位是0

12的倍数：　若一个整数能被3和4整除

* 判断是否质数？

若m不能被2~sqrt(m)间任一整数整除，则m必定是素数

如17不能被range(2, 5)之间的每一个整数整除

* 判断是否是三角形数？

求n = [sqrt(8x+1) – 1 ]/2

若n是整数，那么x就是第n个三角形数。

若n不是整数，那么x不是三角形数。

* 哥德巴赫猜想：是否每个大于2的偶数都可写成两个素数之和？
* 证明：任意两个奇数的平方差是8的倍数

设任意奇数2n+1, 2m+1, 其中m, n ∈ N

(2m+1)2 – (2n+1)2 = 4(m+n+1)(m-n)

当m, n都是奇数或都是偶数时，m-n是偶数，被2整除

当m, n一奇一偶时，m+n+1是偶数，被2整除

## 定理

* a\*b = (a,b) \* [a, b] 整数a,b, 最大公因子(a,b), 最小公倍数[a,b]

正整数

* 大于1的正整数N可以**唯一分解**成有限个质数的乘积

N = p1a1 \* p2a2 \* ...\*pnan

奇数和偶数

* 任意两个奇数的平方差是8的倍数

素数

* 大于3的素数只分布在6n-1和6n+1两数列中
* 所有大于10的质数中，个位数只有1,3,7,9

完全数

* 所有的完全数都是三角形数
* 完全数的所有约数的倒数和等于2

1/1+1/2+1/3+1/6 = 2

1/1+1/2+1/4+1/7+1/14 = 2

…

* 完全数(>6)可以表示成连续奇数的立方和

28 = 13 + 33

496 = 13 + 33 + 53 + 73

8128 = 13 + 33 + 53 + 73 + ...153

* 完全数可以表示2的连续正整数次幂和

6 = 21 + 22

28 = 22 + 23  + 24

496 = 24 + 25 + 26 + 27 + 28

8128 = 26 + 27 + 28 + … + 212

* 完全数都是以6或8结尾,如果以8结尾，那么就肯定是以28结尾

三角形数

* 第n个三角形数是开始的n个自然数的和。

1 = 1

3 = 1 + 2

6 = 1 + 2 + 3

10 = 1 + 2 + 3 + 4

15 = 1 + 2 + 3 + 4 + 5

…

* 开始的n个立方数的和是第n个三角形数的平方

12 = 13

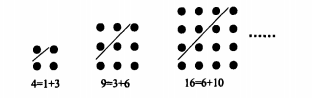
32 = 13 + 23

62 = 13 + 23 + 33

102 = 13 + 23 + 33 + 43

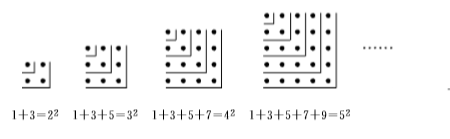
…

* 所有三角形数的倒数之和是2
* 任何三角形数乘以8再加1是一个平方数
* 两个相继的三角形数之和是平方数

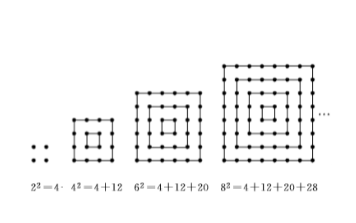


平方数

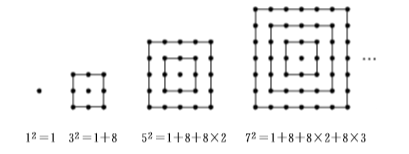
* 平方数都可以表示为从“1”开始的连续奇数之和的形式



* 任何一个非零偶数的平方数都可以表示为首项为4，公差为8的一串数之和



* 任何一个奇数的平方都可以表示为从1开始，然后依次是8的连续倍数的几个数之和



* 每4个连续的自然数相乘加 1，必定会等于一个平方数，

即a(a+ 1)(a+ 2)(a+ 3) + 1 = (a+ 3a+ 1)。

* 四平方和定理说明所有正整数均可表示为最多四个平方数的和
* 平方数必定是3的倍数或者3的倍数+1。
* 平方数必定是4的倍数或者4的倍数+1。

立方数

前n个整数的立方数累加等于累加的平方

13 + 23 + 33 + … + n3 = (1 + 2 + 3 + … + n)2

# 代数

## 因式分解

长除法

## 求解一元多次方程

1. 求一元二次方程？

可用配方法推导出根公式

1. 求一元三次方程？

推导如下：

令

令,

(1)

令

(2)

综合(1),(2)式得：

令

一元二次方程

故 的方程根：

其中

塔尔塔利亚标准一元三次方程 解：

解： while and while

## 复数

*其中i为虚数单位,*

指数形式： 其中

所有实数能用一条数轴表示，复数用平面上的点表示

复数可以视为复平面上的点(a, b), 也可以看作是向量 (a, b)， 横轴为实轴，竖轴为虚轴

向量z的幅度为r, 与实轴正方向的夹角为

幂公式：

棣莫弗定理：

开方公式：

欧拉公式：

据麦克劳林展开式

将代入指数展开式，即证

# <Statistics>统计学

统计描述　＋　统计推断

Statistical inference

Descriptive statistics is solely concerned with properties of the observed data, descriptive statistics are typically used as a preliminary step before more formal inferences are drawn

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. The majority of the problems in statistical inference: How the translation from subject-matter problem to statistical model

Some common forms of statistical proposition

* point estimate,
* interval estimate,
* credible interval,
* rejection of a hypothesis,
* clustering or classification of data points into groups.

Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data

three levels of modeling assumptions:

* Fully parametric

The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters

* Non-parametric:
* Semi-parametric:

Importance of valid models/assumptions

## 基本概念

统计学定义为收集数据，分析数据，由数据得出结论而组成的概念，原则和方法

统计可被定义为在随机性中寻找规律性

经验变量：特征，字段，属性

理论变量：经验变量→变换，如z, t, gamma, F变量

抽样误差：如果研究被再做一遍，结果未必会和上次一模一样。抽样误差告诉我们，样本离总体的实际值可能有多远

抽样误差的大小依赖于得到样本的方式和样本中包含的观测的个数。

在公布任何一次抽样调查的结果时都应说明抽样误差的大小，不管是比例，均值还是其它形式

标准得分：(观察值-均值)/标准差

不同的变量一般有不同的均值和标准差，统计上，均值和标准差不同时，一个变量的值不能与另一变量的值相比较。将变量值转换化标准得分，从而一个变量的任何值都可以和任何其它变量的值相比较，因为我们知道任何一个得分与均值的相对距离

mean-2\*std mean-std mean mean+std mean-2\*std

-2 -1 0 1 2

任何变量的标准得分的值大部分落在[-2, 2]，2/3的标准得分值落在[-1, 1]。如果一个观察值的标准得分>2 or <-2，那么这个观察值就不寻常的大或小。标准得分常称为t值

## 统计推断

从样本数据得出与总体参数值有关的结论　＝　估计＋假设检验

尽管样本中的信息不完全，而且来自样本的结果一般者不等于总体真值，为了弥补样本结果的不准确，研究者们计算抽样误差－－这个数能使95%的抽样结果都位于由总体真值加减样本误差所得到的区间内

点估计：是一个用来估计总体参数的数，一般用样本统计量作为总体参数的点估计

最大似然法，Bayes估计法

注意：特定数据训练得到的估计方法的准确性不能通过相同的数据进行测试，（举个例子，考试题目如果与课堂习题完全相同，对知识的理解效果就得不到检验，因为仅靠死记硬背就能得到最佳答案。所谓的实力，是解决之前未曾遇到过的问题的能力）这涉及到模型选择的方法：一般采用交叉检验法

置信区间：包含总体参数真值的区间　＝　点估计＋抽样误差

无法回答某个特定的置信区间是否包含总体真值，我们知道的仅仅是－－在多次抽样中有95%的样本得到的区间包含真值，95%称为置信水平

置信区间是一个随机区间，因为置信区间会因样本的不同而不同，而且不是所有的区间都包含参数真值。

一次抽样无法回答，因为每次抽样的样本集不同，导致点估计和区间估计均不同，但无数次抽样，得到的无数个点估计和区间估计，

无数个点估计的平均为总体真值；

无数个区间估计中，至少有95%的区间估计中包含真值，则95%称为置信水平

换句话，95%置信水平是指多次抽样中有95%的置信区间包含未知的总体参数值，至于在一次抽样得到的置信区间是包含总体参数的众多区间中的一员呢，还是属于个别不含参数值的区间就永远不得而知了。

置信区间特点：

大的样本产生较短的置信区间

低的置信水平产生较短的区间

按照惯例，抽样误差默认是基于95%置信水平的，所有好的调查都应该提供抽样误差，这样读者自己就可以构造置信区间了

多个总体，抽取各自的样本，若样本均值有差异，能否推出总体均值有差异？若有差异，则称为统计显著

因为样本均值的差异，有可能仅仅归于随机性

零假设：假设的内容总是没有差异，或没有改变，或变量间没有关系等等

零假设所提问题的答案是以样本数据为基础的，值得强调的是零假设总是一个与总体参数有关的问题，关于样本统计量如样本均值或方差的零假设是没有意义的，因为样本统计量是已知的

零假设：两人总体均值无差异

假设检验机制：

将样本均值之差1.3，标准差及样本数　→ 　标准得分(t变量)　→ 查t分布表，得出样本均值之差>1.3的概率，即p-value, → 若p-value < 0.05

两种解释：

* 零假设是正确的，观测到的数据恰好是不常发生的那一类；
* 数据倒是常见的那一类，只是零假设是错的

若零假设为真，那么样本均值差异1.3应很常见，而实际计算出基本样本均值差异1.3的概率小于0.05, 说明不常见，那么选择第二种解释，即认为导致这个小概率的假设－－两总体均值无差异是错的，即应认为两人总体均值差异不为0.

总结：首先要对我们研究的事物作出某种假设，然后收集数据，并在假设的基础上计算得到该数据的概率，如果这个概率非常小，如小于0.05，则一开始的假设就是错误的

假设检验只考虑到一个可能值，置信区间给了我们一个参数值的可能范围。即使当我们拒绝了零假设，并得到均值之差不为零的结论之后，紧接着的一个问题就是该差异有多大，这个问题是由置信区间来负责回答

## 变量间关系

许多变量可能有统计关系但没有因果关系

比如冰淇淋销量与车祸中儿童受伤数，只是统计关系，背后的第三个变量，温度则是与这个变量有因果关系

用一个变量的值预测另一个变量的值时，它们之间不必非得有因果关系

自变量：先发生的变量

因变量：受影响的变量

两个分类变量的相关分析　－－卡方分析

列联表：描述两个分类变量分布的频率表　contingency table

若自左上角至右下角的对角线上有更多的观测，这种频率结构表明在这些数据中这两个变量是相关的

若两列的百分比分布不同，则两变量存在着关系

若相关系数<0.3，认为关系比较弱; 若[0.3 0.7],则关系适中；若>0.7,则关系强

相关系数计算见公式

两个数值型变量的回归分析和相关分析

离群点对相关系数影响大

列差变量：除自变量外其他所有变量对因变量的效应 = sum(yi - a-bxi)2 观测值－估计值

总的平方和：度量所有的变量对因变量的效应　＝ sum(yi – u)2 观测值－均值

回归平方和：度量自变量对因变量的效应 ＝sum( a+bxi - u)2 估计值－均值

r = sqrt(回归平方和/总的平方和)

相关系数的平方　＝　因变量取值变化的效应可归于自变量的比例

分类自变量和数值因变量的方差分析

若分类自变量只有两种取值（如西海岸和东海岸），那么可以定义虚拟变量，西海岸映射为0, 东海岸映射为1，从而做回归分析

若分类自变量有多种取值，采用方差分析anova

方差分析：用来对比因变量在不同组中的平均值的统计方法

常用盒子图来对比各组中因变量的不同

假定自变量区域，因变量暴力率

方差分析：先根据自变量（地区）分组，再求出每一组的因变量的平均值，我们感兴趣的是因变量的均值在各组之间是否有差异，需要借助方差

自变量平方和：　sum(组均值－总均值)2 度量地区变量的效应

残差平方和：sum(观测－组均值)2

总平方和：sum(观测－总均值)2 = 自变量平方和+残差平方和

r = sqrt(自变量平方和/总平方和)

地区变量与暴力犯罪率变化的R2有关联，而残差变量与剩下的1-R2有关联

类别型变量~类别型变量

独立检验: 卡方检验 （p值：独立概率,若p值小，则拒绝二者独立）

Library(vcd)

Mytable <- xtabs(~Treatment+Improved, data=Arthritis)

Chisq.test(mytable)

相关性度量:相关性强弱系数Phi

Assocstats(mytable)

数值型变量~二元类别型变量 （即样本分成两个组）

用于检验两个总体的均值是否相等？（若p值小，拒绝两组相同）

两组样本独立的t检验，且从正态总体中抽样

t.test(数值变量~二元变量,data=..)

## 多元分析

在多元分析中，残差变量对因变量的影响减少了，这是因为我们将所有自变量的作用同时从残差变量中分离出来，而非逐一分离

统计方法的选择总是决定于所考虑的变量的性质

自变量 因变量 方法

分类型 分类型 偏phi法

数值，哑元 数值型 多元回归

回归模型无截距项，有m个特征，设置

偏phi法：

控制其余变量，计算单变量和因变量的phi

然后取各组中的phi的均值

回归，分类器

若自变量是数值型和分类型，对分类型变量引入哑元

若某分类变量level=5, 则引入5个哑元变量，每个哑元变量取值0/1，每个二进制特征对应分类数据的一个等级

categorical\_feature = pd.Series(['sunny', 'cloudy', 'snowy', 'rainy', 'foggy'])

dummies = pd.get\_dummies(categorical\_feature)

数值型变量~多元类别型变量(即样本分成多个组)

方差分析：名字令人误解，更适合的名字是均值分析：先根据自变量分组，再求出每一组因变量的均值，目的是讨论各组的均值是否不同。但这里判断均值之间是否有差异是借助于方差.

Fit <- aov(数值型变量~类别型变量)

## 如何生成概率分布？

伪随机序列：由随机的种子确定的长序列，只有种子是随机的

真随机：长序列的每个元素都是随机的

代码库一般提供[0 1]范围内的均匀分布函数，对生成的随机序列进行变换，使其遵从特定的分布

0-1分布： y = 正面朝上　if x < 0.5

均匀分布：y = (b-a)\*x + a

正态分布：y = 0.5\*exp(-x\*x/2) and z = std\*y+mean

多元正态分布：N(mean\_vec, cov\_mat)

随机生成变量(x1, x2, ...xn), 从而求f(x1, x2, ...xn)

采用scipy.stats工具包

from scipy import stats

# 二项分布binomial distribution

n次事件中，发生概率为0.5, 恰好发生x次的概率

#计算概率值p(x=650), p(x=651), ...p(x=1200), n=1200, p=0.5

stats.binom.pmf(range(650, 1201), 1200, 0.5)

#生成随机变量x, 服从二项分布n=1200, p=0.5

data = stats.binom.rvs(n=1200, p=0.5, size=1000)

plt.hist(data, bins=24, normed=True) #可视化随机变量pmf

# 泊松分布 Poisson distribution

某路口事件发生的比率是ramda(每天发生次数), 求一段固定时间间隔内(这里是1天）发生x次事件的概率

stats.poisson.pmf(range(0, 10), 2)

#生成随机变量x, 服从泊松分布ramda=2

data = stats.poisson.rvs(mu=2, size=1000)

# 正态分布

stats.norm.pdf(np.arange(-5, 5, 0.1), loc=0, scale=1)

data = stats.norm.rvs(loc=0, scale=1, size=1000)

## 可视化

可视化：盒子图boxplot(数值型~类别型)

两组样本不独立的t检验，组间差异呈正态分布

可视化：带有置信区间的组均值图形

Library(gplots)

Plotmeans(数值型变量~类型别变量)

boxplot

# 《概率统计》

## 基本概念

PDF (probability density function) : 连续型随机变量X的概率密度函数 = CDF的导数

描述随机变量的输出值，在某个确定的取值点附近的可能性的函数

CDF (cumulative distribution function): 分布函数 = PDF的积分

描述随机变量X的概率分布: F(x) = P(X<=x)

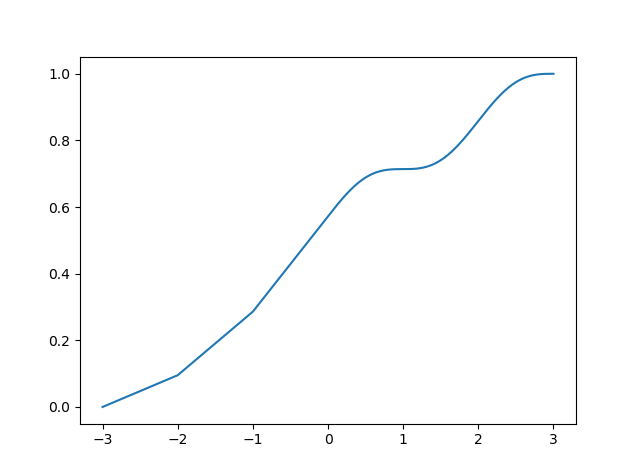
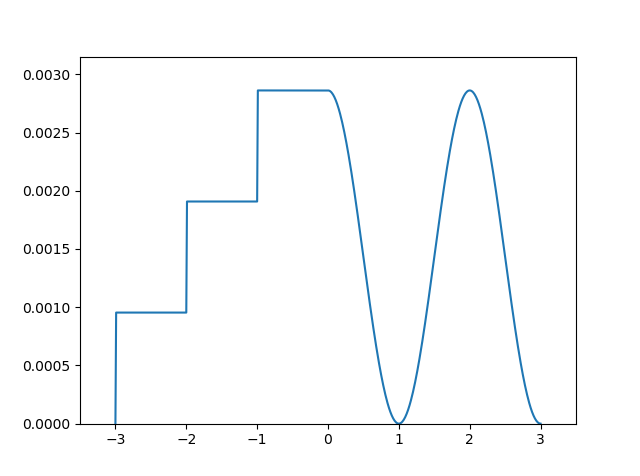
X落在区间(a, b)之内的概率为：P(a < X <= b) = F(b) – F(a)

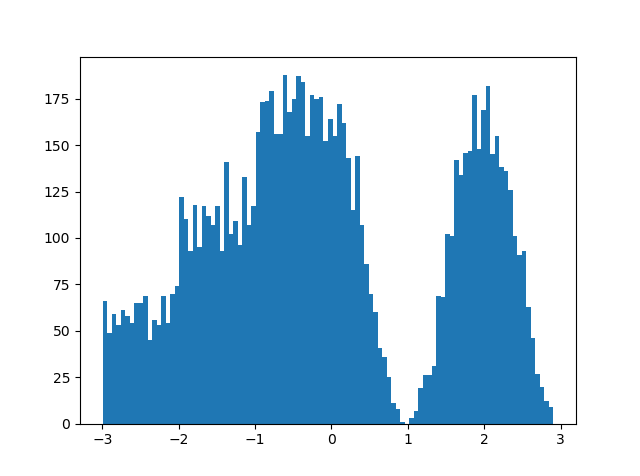
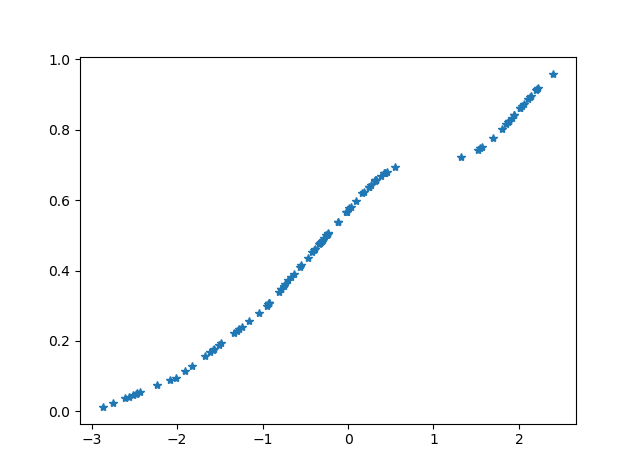
例：

## 生成模型：产生指定分布数据的模型 （均匀分布 -> 指定分布）

生成指定分布PDF （见左上图） -> CDF {x, y=[0, 1]} （见右上图）

生成均匀分布，作为CDF y轴，进行插值(见左下图)，得{x=指定分布, y=均匀分布} （见右下图）





Python 代码：

x\_samples = np.arange(-3, 3.01, 0.01)

pdf = define\_pdf(x\_samples)

cdf = calc\_cdf(pdf)

generate = partial(np.interp, xp=cdf, fp=x\_samples)

u\_rv = np.random.random(10000) #均匀分布[0, 1]

x = generate(u\_rv) #插值成指定分布