


清华大学 大学数学实验

# 大学数学实验



## 实验11 数据的统计与分析

清华大学数学科学系

1 2 3

清华大学 大学数学实验

### 统计学(statistics): 一级学科

**统计学是收集、分析、表述和解释数据的科学**

统计学 vs 数学      统计学 vs 数据科学?

**统计学是一门方法论科学, 而不是一门实质性科学**

- 实质性科学研究某领域现象的本质关系和变化规律
- 统计学为研究这些关系和规律提供数量分析方法

**统计学是横跨社会科学和自然科学的多学科性的科学**

理论统计学、应用统计学 (复合/交叉/边缘学科);  
社会统计、工业统计、保险统计、生物统计、...

1 2 3

清华大学 大学数学实验

### 数据的统计与分析的两类方法

**第一类: 普通意义的统计 (普查) → “数数” (counting)**

对生产的全部1000件产品逐一检验, 发现18件次品  
对全区居民逐一调查, 得到月平均支出为828元

→ 次品率: 1.8%; 月平均支出为828元

**优点:** 结果完全确定, 可信  
**缺点:** 调查、收集的数据量可能很大, 经费投入大;  
有些产品不允许全部检验, 如灯泡、电器的寿命等

**描述统计 (Descriptive Statistics)**

1 2 3

清华大学 大学数学实验

### 第二类: 数理统计 (抽查) → “推理” (inferring)

全部产品中随机抽取100件, 发现2件次品  
随机调查了200位居民, 得到月平均支出为788元

→ 次品率: 2%; 月平均支出788元

**优点:** 调查、收集的数据量小, 经费投入小, 适合不允许全部检验的产品, 如灯泡、电器的寿命等  
**缺点:** 结果是随机的, 是否可信?


**任务:** 怎样用它来估计整体的状况 (全部产品的次品率, 全体居民的月平均支出)

**推断统计 (Inferential Statistics)**

1 2 3

清华大学 大学数学实验

### 本实验基本内容

- 一. 实例及其分析
- 二. 数据的整理和描述 
- 三. 随机变量的概率分布及数字特征
- 四. 用随机模拟计算数值积分
- 五. 实例的建模和求解

1 2 3

清华大学 大学数学实验

### 一. 实例及其分析

1 2 3

清华大学 大学数学实验

### 实例1：报童的利润

报童每天购进报纸零售，晚上将卖不掉的报纸退回；  
每份报纸购进价 $a$ ，零售价 $b$ ，退回价 $c$ ： $b \geq a \geq c$ ；  
为获得最大利润，该报童每天应购进多少份报纸？

**159天报纸需求量的情况**

需求量	100	120	140	160	180	200	220	240	260	280
天数	3	9	13	22	32	35	20	15	8	2

设 $a=0.8$ 元， $b=1$ 元， $c=0.75$ 元，为报童提供最佳决策

1 2 3 7

清华大学 大学数学实验

### 实例1：报童的利润（续）

**分析：**每天报纸需求量随机，报童每天利润也随机；  
以每天平均利润最大为目标，确定最佳决策。

**数学模型近似：**  
每天需求为 $r$ 的天数所占的百分比，记做 $f(r)$ ；  
如200(-219)份所占的百分比为35/159=22%

**决策变量：**报童每天购进报纸的份数 $n$

**平均利润：** $V(n)$

$$V(n) = \sum_{r=0}^{n-1} [(b-a)r - (a-c)(n-r)]f(r) + \sum_{r=n}^{\infty} [(b-a)n]f(r)$$

1 2 3 8

清华大学 大学数学实验

### 实例2：路灯更换策略

**路政部门：**路灯维护

**条件：**需要专用云梯车进行线路检测和更换灯泡；  
向相应管理部门提出电力使用和道路管制申请；  
向雇用的各类人员支付报酬等

**更换策略：**整批更换

**管理部门：**不亮灯泡，折合计时进行罚款。

**路政部门的问题：**多长时间进行一次灯泡的全部更换？

- 换早了，很多灯泡还没有坏；
- 换晚了，要承受太多的罚款。

1 2 3 9

清华大学 大学数学实验

## 二. 数据的整理和描述

- 数据的收集和样本的概念
- 数据的整理、频数表和直方图
- 统计量
- MATLAB命令

1 2 3 10

清华大学 大学数学实验

### 数据的收集：顾客感觉舒适时的柜台高度

- 银行随机选了50名顾客进行调查
- 测量每个顾客感觉舒适时的柜台高度(单位：厘米)

100	110	136	97	104	100	95	120	119	99
126	113	115	108	93	116	102	122	121	122
118	117	114	106	110	119	127	119	125	119
105	95	117	109	140	121	122	131	108	120
115	112	130	116	119	134	124	128	115	110

- 银行怎样依据它确定柜台高度呢？

1 2 3 11

清华大学 大学数学实验

### 样本：统计研究的主要对象

- 总体(population)：**研究对象的全体（母体）
  - 如：所有顾客感觉舒适的高度
- 个体(individual)：**总体中一个基本单位（总体单位）
  - 如：一位顾客的舒适高度
- 样本(sample)：**若干个体的集合（抽样，取样）
  - 如：50位顾客的舒适高度
- 样本容量(sample size)：**样本中个体数(样本数)
  - 如：50

1 2 3 12

清华大学 大学数学实验

## 总体和样本是随机变量，对吗？

总体的三层含义：

- **群体**：所有顾客
- **数据**：所有顾客——感觉舒适的高度（数据）
- **随机变量**：数据的特征，用随机变量描述

■ 顾客群体的舒适高度~随机变量 $X$ ，概率分布 $F(x)$

■ **（简单随机抽样，sampling）**：有放回，随机抽样

➔ **（简单随机）样本**： $\{x_i, i=1, \dots, n\}$

**两层含义**

- 一组独立的、同分布 (**i.i.d.**) 的随机变量 ( $\sim F(x)$ )
- 抽样后，是一组具体数据 (观测值, **observations**)

清华大学 大学数学实验

## 数据的整理

### 北京地区SARS患者的统计数据（截至2003年5月5日）

年龄	10岁以下	11-20岁	21-30岁	31-40岁	41-50岁	51岁以上	总数
人数	24	145	677	382	332	337	1897
比例	1.27%	7.64%	35.69%	20.14%	17.50%	17.77%	100%

**比较直观，比较清晰的结论**

21—50岁的中青年患者大约占总发病人数的 3/4，提醒民众中青年是易感人群。

清华大学 大学数学实验

## 给定样本（数据，观测值）：频数表

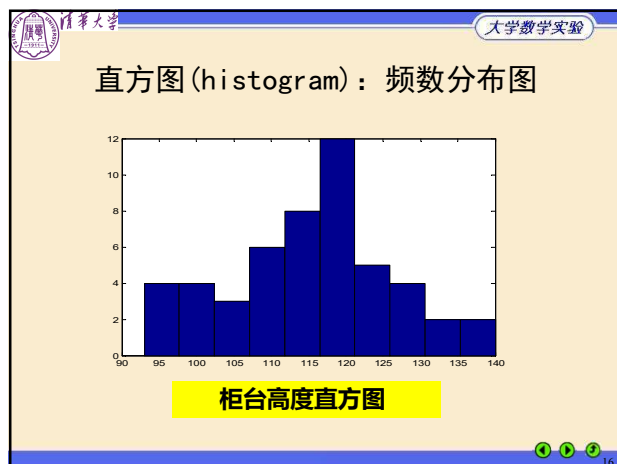
将数据的取值范围划分为若干个区间，统计这组数据在每个区间中出现的次数，称为**频数**，得到一个**频数表**。

**柜台高度频数表**      区间长=(140-93)/10=4.7

中点	95.35	100.05	104.75	109.45	114.15	118.85	123.55	128.25	132.95	137.65
频数	4	4	3	6	8	12	5	4	2	2

**作用：推测出总体的某些简单性质。**

如上表表明选择柜台高度在107.10至125.90的有**31**人，占总人数的62%，柜台高度设计在这个范围内，会得到大多数顾客的满意。



清华大学 大学数学实验

## 平均值

频数表和直方图给出某个范围的状况，无法直接给出具体值，如确定柜台具体高度

**平均值 (mean, 简称样本均值) 定义为**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{x} = 115.26$$

可作为设计柜台高度的参考值

清华大学 大学数学实验

## 例：两个班的一次考试成绩

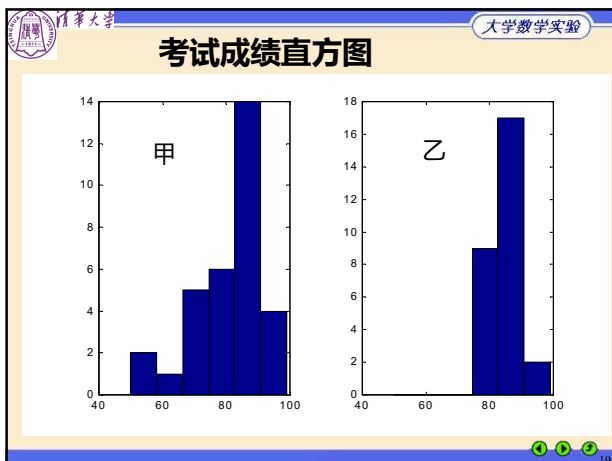
序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
甲班	92	88	85	92	95	79	84	87	88	65	93	73	88	87	94	80
乙班	84	83	82	85	82	81	82	90	84	78	75	83	78	85	84	79
序号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
甲班	69	86	88	78	79	68	88	87	55	93	79	85	90	53	99	81
乙班	85	73	90	77	81	82	82	80	86	83	77	78				

**现象1：**甲班平均值：82.75分，乙班平均值：81.75分

**结论：**大致表明甲班的平均成绩稍高于乙班

**现象2：**甲班90分以上7人，但有2人不及格，分数分散

乙班全在73分到90分之间，分数相对集中



### 标准差

描述数据的分散程度（统计上称为**变异**，variation）

样本 $x=(x_1, x_2, \dots, x_n)$ 的**标准差**(Standard deviation)为：

$$s = \left[ \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2} \quad s_1 = \left[ \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2}$$

标准差即**标准偏差**，也称**均方差**；**偏差**，**离差**  
 其平方称为**方差**（variance）**(deviation)**

例：甲班成绩的标准差为10.98分，乙班为3.98分，表明甲班成绩的分散程度远大于乙班。

### 统计量

由样本加工出来的、反映样本数量特征的函数  
 （简单说，就是样本的函数；注意：不含未知参数）

常用统计量：**样本矩**（sample moment）

**样本k阶原点矩**  $\bar{\alpha}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$   $k=1$ : 均值

**样本k阶中心矩**  $\bar{\beta}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$   $k=2$ : 方差

三类常用统计量：**位置**，**变异程度**，**分布形状**

### 统计量

**表示位置的还有：**

**中位数(median)**：将数据由小到大排序后处于中间位置的那个数值。  
 $n$ 为奇数时，中位数唯一确定；  
 $n$ 为偶数时，定义为中间两数的平均值

**表示变异程度的还有：**

**极差(range)**： $x_1, x_2, \dots, x_n$ 的最大值与最小值之差

**表示分布形状的：**

**偏度(skewness)**：分布对称性  $g_1 = \frac{1}{ns_1^3} \sum_{i=1}^n (x_i - \bar{x})^3$

**峰度(kurtosis)**：分布形状  $g_2 = \frac{1}{ns_1^4} \sum_{i=1}^n (x_i - \bar{x})^4$

### 多选题 1分

关于统计量 (假设 $n>5$ )

设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的简单随机样本，其中 $\mu, \sigma^2$ 未知，则下面是统计量的是

☐ A  $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$

☐ B  $\text{Min} \{X_i | i=1, \dots, n\}$

☐ C  $(X_2 + \dots + X_{n-1})/2$

☐ D  $X_5$

☐ E 没有正确答案，以上都不是统计量

提交

### MATLAB数据描述的常用命令

命令	名称	输入	输出	注意事项
<code>[n,y]=hist(x,k)</code>	频数表	x: 原始数据行向量 k: 等分区间数	n: 频数行向量 y: 区间中点行向量	<code>[n,y]=hist(x)</code> 中k取缺省值10
<code>hist(x,k)</code>	直方图	同上	直方图	同上
<code>mean(x)</code>	均值	x: 原始数据行向量		
<code>median(x)</code>	中位数	同上	中位数	
<code>range(x)</code>	极差	同上	极差	
<code>std(x)</code>	标准差	同上	标准差 $s$	<code>std(x,1)</code> : $s_1$
<code>var(x)</code>	方差	同上	方差 $s^2$	<code>var(x,1)</code> : $s_1^2$
<code>skewness(x)</code>	偏度	同上	偏度 $g_1$	
<code>kurtosis(x)</code>	峰度	同上	峰度 $g_2$	

清华大学 大学数学实验

### 示例

求银行柜台高度的频数表、直方图及均值等统计量:

Demo1101.m  
Demo1102.m

输出图和下列结果:

```
N=4 4 3 6 8 12 5 4 2 2
Y= 95.3500 100.0500 104.7500 109.4500 114.1500 118.8500 123.5500
128.2500 132.9500 137.6500
x1 = 115.2600, x2 = 116.5000
x3 = 47, x4 = 10.9690
x5 = -0.0971, x6 = 2.6216
```

25

清华大学 大学数学实验

### 三. 随机变量的概率分布及数字特征

- 频率与概率
- 概率密度与分布函数
- 期望和方差
- 常用的概率分布
- MATLAB命令

26

清华大学 大学数学实验

### 频率与概率

**频率:** 样本数据在一个确定区间  $(a, b]$  的频数  $k$  与样本容量  $n$  的比值

$$f(a < X \leq b) = \frac{k}{n}$$

保证抽取样本的随机性和独立性:  
样本容量无限增大时, 频率会趋向一个确定值;  
这个值称为随机变量  $X$  落入区间  $(a, b]$  的**概率** (Probability), 记作

$$P(a < X \leq b)$$

27

清华大学 大学数学实验

### 概率密度与分布函数

对于连续随机变量  $P(a < X \leq b) = \int_a^b p(x)dx$

**概率密度函数** (Probability density function, 简称**概率密度**(pdf)):

$$p(x) \geq 0 \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

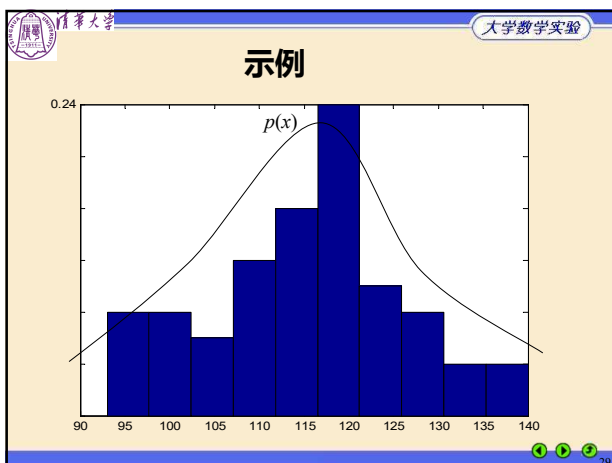
**累积(概率)分布函数** (Cumulative distribution function, 简称**分布函数**(cdf))

$$F(x) = P(X \leq x) = \int_{-\infty}^x p(x)dx$$

$$F(-\infty) = 0, F(\infty) = 1 \quad P\{a < X \leq b\} = F(b) - F(a)$$

$$p(x) = \frac{dF}{dx}$$

28



清华大学 大学数学实验

### 二维随机变量 $(X, Y)$

**联合分布密度函数**  $p(x, y) \geq 0 \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y)dx dy = 1$

**边际分布密度函数**  $p_X(x) = \int_{-\infty}^{\infty} p(x, y)dy, \quad p_Y(y) = \int_{-\infty}^{\infty} p(x, y)dx$

当  $p(x, y) = p_X(x)p_Y(y)$

称随机变量  $X$  和  $Y$  相互独立。

类似可定义多维连续随机变量的概率密度、概率计算、边际概率密度和多维随机变量的相互独立等。

30

期望和方差

随机变量 $X$ 的期望就是平均值的意思, 记作 $EX$ 或 $\mu$

$$EX = \int_{-\infty}^{\infty} xp(x)dx$$

方差

$$DX = \int_{-\infty}^{\infty} (x - EX)^2 p(x)dx$$

样本均值(随机变量)的均值与方差:

$$E\bar{X} = \frac{1}{n} \sum_{i=1}^n EX_i = EX \quad D\bar{X} = \frac{1}{n^2} \sum_{i=1}^n DX_i = \frac{DX}{n}$$

(均值)标准误(差) (SEM: standard error of mean):  
样本均值(统计量)的标准差:  $\sqrt{DX/n}$

多选题 1分

关于标准误 (SEM), 下列哪些说法正确?

- ☐ A 标准误就是总体的标准差
- ☐ B 标准误就是样本的标准差
- ☒ C 标准误就是样本均值的标准差
- ☒ D 标准误描述均值抽样分布的离散程度, 可用于衡量均值抽样误差的大小
- ☒ E 标准误越小, 样本均值对总体均值越有代表性
- ☐ F 以上都不对

提交

标准误 (SEM) vs 均方误差 (MSE)

MSE对象和目的: 与样本均值及偏差 (deviation) 不同

这里的误差(error), 是测量值与真实值之差  
或者观测值与预测值之差

均方根误差 (root mean squared error, RMSE, 形似标准差)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\text{observed}_i - \text{predicted}_i)^2}$$

均方误差(MSE): RMSE的平方 (形似方差; vs “均方差”)

平均绝对误差(MAE, MAD): 误差绝对值求和/N

平均绝对百分误差(MAPE): 相对误差绝对值求和/N

常用的概率分布

均匀分布(Uniform distribution):  $X \sim U(a, b)$

$$p(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b], \\ 0, & \text{其他} \end{cases} \quad EX = \frac{a+b}{2}, \quad DX = \frac{(b-a)^2}{12}$$

指数分布(Exponential distribution):  $X \sim \text{Exp}(\lambda)$

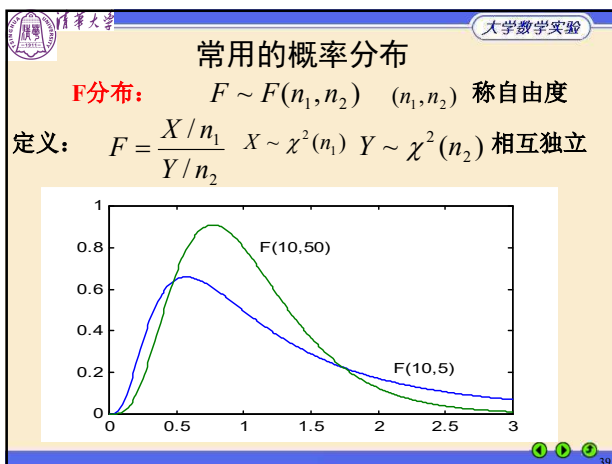
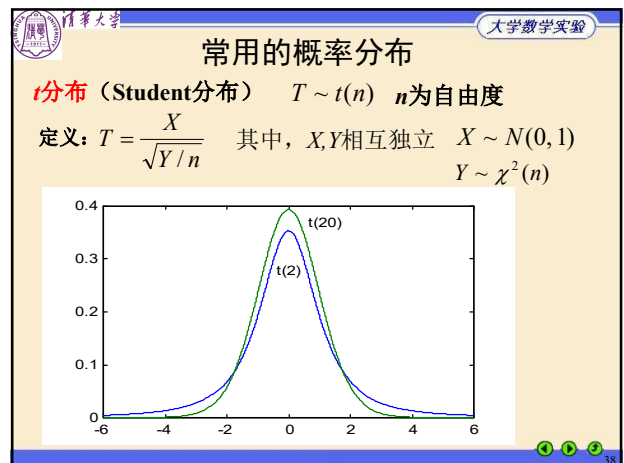
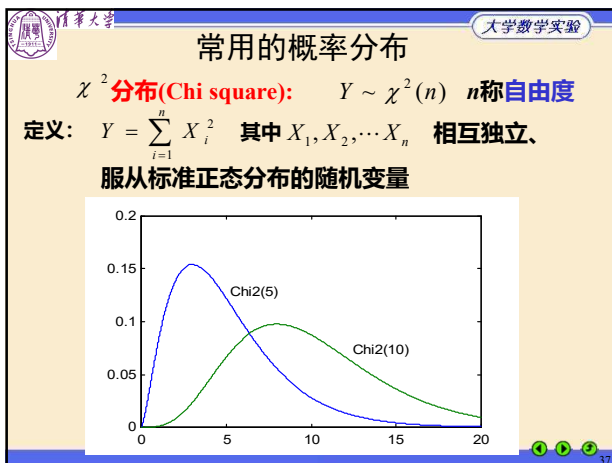
$$p(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, & x \geq 0 \\ 0, & \text{其他} \end{cases} \quad EX = \lambda, \quad DX = \lambda^2$$

相应的密度函数

常用的概率分布

正态分布(Norm distribution):  $X \sim N(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad EX = \mu, \quad DX = \sigma^2$$



多选题 1分

设总体  $Z \sim N(0, 1)$ ,  $Z_1, Z_2, \dots, Z_n$  为简单随机样本, 其中  $n > 3$ , 则下列表达正确的是

☒ A  $\sum_{i=1}^n Z_i^2 \sim \chi^2(n)$

☐ B  $\frac{1}{n} \sum_{i=1}^n Z_i \sim N(0, 1)$

☒ C  $\frac{\sqrt{n-1}Z_n}{\sqrt{\sum_{i=1}^{n-1} Z_i^2}} \sim t(n-1)$

☒ D  $\left[ \frac{\left(\frac{n-1}{2}\right) \sum_{i=1}^2 Z_i^2}{\sum_{i=1}^n Z_i^2} \right] \sim F(2, n-2)$

☐ E 以上都不正确

提交

常用的概率分布: 离散分布

**贝努利试验:**

- 一次试验只有两种结果 (成功和失败)
- 记成功的概率为  $p$ ,  $q=1-p$

**二项分布** (Binomial distribution)  $X \sim B(n, p)$

记  $n$  次独立试验中成功的次数是随机变量  $X$

$$P(X=k) = \binom{n}{k} p^k q^{n-k}, \quad k=0, 1, \dots, n$$

$$EX = np, \quad DX = npq$$

背景问题: 产品检验中的废品个数

常用的概率分布: 离散分布

**泊松分布** (Poisson distribution)  $X \sim \text{Poiss}(\lambda)$ ,

当二项分布的  $n \rightarrow \infty$ ,  $np \rightarrow \lambda$  (常数) 时

$$P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k=0, 1, 2, \dots$$

$$EX = \lambda, \quad DX = \lambda$$

背景问题:  
服务系统在一定时间内接到的呼唤数(到达率);  
到达间隔服从 iid 的 **指数分布** (参数: 均值  $1/\lambda$ )



大学数学实验

### MATLAB命令

分布	均匀分布	指数分布	正态分布	$\chi^2$ 分布	t分布	F分布	二项分布	泊松分布
字符	unif	exp	norm	chi2	t	f	bin	poiss

功能	概率密度	分布函数	逆概率分布	均值与方差	随机数生成
字符	pdf	cdf	inv	stat	rnd

`y=normpdf(1.5,1,2)` 正态分布( $\mu=1, \sigma=2$ ) 在 $x=1.5$ 处的概率密度 (标准正态分布的 $\mu, \sigma$ 可省略)  
`y=normcdf([-1 0 1.5],0,2)`  $N(0,2^2)$  在 $x=-1, 0, 1.5$ 处分布函数值  
`[m,v]=fstat(3,5)` 计算 $F(3,5)$ 的期望和方差  
`x=ttinv(0.3,10)` 计算 $t(10)$ 的0.3-分位数 (点)

大学数学实验

### 二维随机变量

**联合分布密度函数**  $p(x,y) \geq 0 \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) dx dy = 1$

**边缘分布密度函数**  $p_X(x) = \int_{-\infty}^{\infty} p(x,y) dy, \quad p_Y(y) = \int_{-\infty}^{\infty} p(x,y) dx$

**协方差**  $Cov(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-EX)(y-EY)p(x,y) dx dy$

**相关系数**  $r_{XY} = Cov(X,Y) / [\sqrt{DX} \sqrt{DY}]$

当 $r=0$ 时 $X, Y$ 不相关,  $r=1$ 时正(线性)相关,  $r=-1$ 时负(线性)相关。当 $X, Y$ 相互独立时 $E(XY)=(EX)(EY)$ , 可知 $X, Y$ 不相关。

**二维正态分布**

$$p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2r\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

$EX = \mu_1, EY = \mu_2, DX = \sigma_1^2, DY = \sigma_2^2, r_{XY} = r$

大学数学实验

### 二维随机变量: MATLAB命令

**二维随机数生成/二维密度函数(例)**

`mu = [1 -1],`  
`Sigma = [.9 .4; .4 .3],`  
`X = mvnrnd(mu,Sigma,10),`  
`p = mvnpdf(X,mu,Sigma)`

**二维数据处理**

$$s_{xy}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

**cov(x,y)** 计算协方差(矩阵)

$$\begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}$$

**corrcoef(x,y)** 计算相关系数(矩阵)

$$\begin{bmatrix} 1 & r_{xy} \\ r_{xy} & 1 \end{bmatrix}$$

多选题 1分

对于给定的正数  $\alpha (0 < \alpha < 1)$ , 设  $Z_\alpha, t_\alpha(n), \chi_\alpha^2(n), F_\alpha(n_1, n_2)$  分别是  $N(0, 1)$  分布、 $t(n)$  分布、 $\chi^2(n)$  分布、 $F(n_1, n_2)$  分布上的  $\alpha$  分位点, 则下列表达正确的是

☒ A  $Z_{1-\alpha} = -Z_\alpha$ 
☒ B  $t_{1-\alpha}(n) = -t_\alpha(n)$ 
☐ C  $\chi_{1-\alpha}^2(n) = -\chi_\alpha^2(n)$ 
☒ D  $F_{1-\alpha}(n_1, n_2) = \frac{1}{F_\alpha(n_2, n_1)}$ 
☐ E 以上都不正确

提交

大学数学实验

### 四. 用随机模拟计算数值积分

- 4.1 定积分的计算
- 4.2 重积分的计算
- 4.3 MATLAB实现

大学数学实验

### 1) 随机投点法

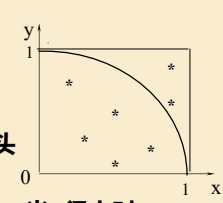
**方法的直观解释——随机投石**

目的: 计算1/4单位圆的面积

向单位正方形里随机投 $n$ 块小石头

若有 $k$ 块小石头落在1/4单位圆内, 当 $n$ 很大时

1/4单位圆的面积  $\frac{\pi}{4} \approx \frac{k}{n}$  (计算 $\pi$ 的一种方法)





大学数学实验

### 大数定律 (雅各布·贝努利定理, 1713)

设  $k$  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, 则对任意的正数  $\varepsilon$ , 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right) = 1$$

投点坐标  $(x_i, y_i)$ ,  $x_i, y_i$  是相互独立、  $(0,1)$  内均匀分布的随机变量 ((0,1)随机数)

点  $(x_i, y_i)$  落在  $1/4$  单位圆内概率

即满足  $y_i \leq \sqrt{1-x_i^2}$   $p = \frac{\pi}{4} \approx \frac{k}{n}$

**一般地** 随机变量  $(X, Y)$  在单位正方形内均匀分布

$$p(x, y) = 1, \quad 0 \leq x, y \leq 1$$

大学数学实验

### 随机投点法 (续)

$P((X, Y) \in \Omega) = \iint_{\Omega} p(x, y) dx dy$   $\Omega: 0 \leq x \leq 1, 0 \leq y \leq f(x) \leq 1$

$$= \int_0^1 dx \int_0^{f(x)} dy = \int_0^1 f(x) dx$$

产生  $n$  组  $(0, 1)$  随机数  $(x_i, y_i)$ , 其中  $k$  组满足  $y_i \leq f(x_i)$   $\Rightarrow P((X, Y) \in \Omega) \approx k/n$

**随机投点法**  $\int_0^1 f(x) dx \approx k/n, \quad 0 \leq f(x) \leq 1$

大学数学实验

### 2) 均值估计法

**大数定律 (辛钦定理, 1929)** 设随机变量  $Y_1, Y_2, \dots, Y_n$  相互独立, 服从同一个分布, 且具有数学期望  $EY_i = \mu (i=1, 2, \dots, n)$ , 则对任意的正数  $\varepsilon$  有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{i=1}^n Y_i - \mu\right| < \varepsilon\right) = 1$$

随机变量  $X$  的概率密度为  $p(x)$ ,  $a \leq x \leq b$

$Y=f(X)$  的期望为  $E(f(X)) = \int_a^b f(x) p(x) dx$

产生  $(0,1)$  随机数  $x_i (i=1, 2, \dots, n)$ ,  $n$  很大  $\Rightarrow \int_0^1 f(x) dx \approx \frac{1}{n} \sum_{i=1}^n f(x_i)$

大学数学实验

### 均值估计法 (续)

**均值估计法的优点**

- 没有  $0 \leq f(x) \leq 1$  限制;
- 不要产生  $y_i$ , 不用比较  $y_i \leq f(x_i)$

**用随机模拟方法计算任意区间上的积分**

$x = a + (b-a)u$

$$\int_a^b f(x) dx = (b-a) \int_0^1 f(a + (b-a)u) du$$

**均值估计法**  $\Rightarrow \approx \frac{b-a}{n} \sum_{i=1}^n f(a + (b-a)u_i)$

其中  $u_i$  为  $(0,1)$  随机数

大学数学实验

### 随机模拟法计算重积分

产生相互独立  $(0,1)$  随机数  $x_i, y_i, i=1, \dots, n$ ; 落在  $\Omega$  内  $m$  个点记作  $(x_k, y_k), k=1, \dots, m$

$$\iint_{\Omega} f(x, y) dx dy \approx \frac{1}{m} \sum_{k=1}^m f(x_k, y_k)$$

$\Omega: 0 \leq x \leq 1, 0 \leq g_1(x) \leq y \leq g_2(x) \leq 1$

- 可用于任意的  $f, \Omega$ , 且可推广至高维
- 结果的精度和收敛速度与维数无关
- 计算量大, 精度低, 结果具有随机性

大学数学实验

### 一般区间重积分的计算

$\iint_{\Omega} f(x, y) dx dy, \quad \Omega: a \leq x \leq b, \quad c \leq g_1(x) \leq y \leq g_2(x) \leq d$

$x_i, y_i (i=1, \dots, n)$  分别为  $[a, b]$  和  $[c, d]$  区间上的均匀分布随机数, 判断每个点是否落在  $\Omega$  域内, 将落在  $\Omega$  域内的  $m$  个点记作  $(x_k, y_k), k=1, \dots, m$

则

$$\iint_{\Omega} f(x, y) dx dy \approx \frac{(b-a)(d-c)}{n} \sum_{k=1}^m f(x_k, y_k)$$

多选题 1分

关于伯努利大数定律 (B)、辛钦大数定律 (K), 下列表达正确的是

☐ A B和K等价

☒ B B是K的特例

☐ C K是B的特例

☒ D 历史上, B先于K被发现和证明

☐ E 以上都不正确

提交

MATLAB实现

随机数的产生: `unifrnd(a,b,m,n)`

产生  $m$  行  $n$  列  $[a,b]$  区间上的均匀分布随机数。当  $a=0$ ,  $b=1$  时, 可用 `rand(m,n)`

随机投点法计算  $\pi$

```
n=10000;
x=rand(2,n);
k=0;
for i=1:n
    if x(1,i)^2+x(2,i)^2<=1
        k=k+1;
    end
end
p=4*k/n
```

Cal\_pi.m

例:炮弹命中概率

目标:  $a=1.2, b=0.8$  (椭圆)

炮弹:  $\sigma_x=0.6, \sigma_y=0.4$  (独立)

$P = \iint_{\Omega} p(x,y) dx dy, \quad \Omega: \frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1$

$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right]$

积分域和被积函数的对称性

蒙特卡罗方法:

$x$  取  $(0,a)$  随机数,  $y$  取  $(0,b)$  随机数

Cal\_pao.m

介绍: 切比雪夫大数定律 (1866)

设随机变量序列  $\{X_i\}$  两两相对独立, 且期望存在  $E(X_i) = \mu_i$ , 方差存在且有共同有限上界  $D(X_i) = \sigma_i^2 < M$ , 则对

$$\forall \varepsilon > 0, \lim_{n \rightarrow +\infty} P\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu_i\right| < \varepsilon\right\} = 1$$

贝努利大数定律是其特例 但辛钦大数定律不是

“切”没有要求同分布, “辛”要求同分布, 但要求方差存在共同上限 但没有要求方差存在

5. 实例的建模和求解

- 报童的利润
- 路灯更换策略

报童的利润

假设:

- 1) 每份报纸的购进价  $a$ , 零售价  $b$ , 退价  $c$
- 2) 需求为连续随机变量  $x$ , 大致服从正态分布
- 3) 将历史的统计表看作需求量的频率, 由此可以计算需求量的均值和标准差

报童每天的平均利润  $V(n)$

$$V(n) = \int_0^n \{(b-a)x - (a-c)(n-x)\} p(x) dx + \int_n^\infty (b-a)np(x) dx$$

其中  $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$   $\mu$  和  $\sigma$  由 3) 的假设计算得到

大学数学实验

$$V'(n) = (b-a)np(n) - \int_0^n (a-c)p(x)dx - (b-a)np(n) + \int_n^\infty (b-a)p(x)dx$$

$$= -\int_0^n (a-c)p(x)dx + \int_n^\infty (b-a)p(x)dx = 0$$


$$\frac{\int_0^n p(x)dx}{\int_n^\infty p(x)dx} = \frac{b-a}{a-c} \quad \text{简化为} \quad \int_{-\infty}^n p(x)dx = \frac{b-a}{b-c}$$

**定性分析** 在  $b \geq a \geq c$  的条件下讨论  $a$ 、 $b$ 、 $c$  的变化对最佳决策  $n$  的影响。

1) 当  $b > a = c$  时; 2) 当  $b = a > c$  时; 3) 当  $b > a > c$  时

**定量求解**

$a=0.8$  元,  $b=1$  元,  $c=0.75$  元

 **报童应购进 232 份报纸**

Newsboy.m

大学数学实验

### 路灯的更换问题

**假设**

- 1) 每个灯泡更换价格:  $a$   
— 灯泡的成本和安装时分摊到每个灯泡的费用
- 2) 不亮灯泡单位时间罚款:  $b$
- 3) 假定灯泡寿命服从  $N(\mu, \sigma^2)$
- 4) 更换周期:  $T$
- 5) 灯泡总数:  $K$

**模型:**

$$F(T) = \frac{Ka + Kb \int_{-\infty}^T (T-x)p(x)dx}{T}$$

大学数学实验

### 路灯的更换问题

计算:  $\frac{dF}{dT} = 0$  可得:  $\int_{-\infty}^T xp(x)dx = \frac{a}{b}$

**定性结果分析**

**结论1:**  
 $a/b$  越大, 更换价格与惩罚费用之比越大, 更换周期  $T$  应越长

**结论2:** 若以灯泡的平均寿命为更换周期, 惩罚费用为:

$$b = \frac{a}{\int_{-\infty}^{\mu} xp(x)dx}$$

大学数学实验

### 路灯的更换问题

考虑灯泡寿命服从  $N(\mu, \sigma^2) \rightarrow b = \frac{a}{\frac{\mu}{2} - \frac{\sigma}{\sqrt{2\pi}}}$

**具体示例**

某品牌灯泡服从平均寿命为4000小时, 标准差为100小时的正态分布, 每个灯泡的安装价格为80元, 管理部门对每个不亮的灯泡制定的惩罚费用为0.02元/小时, 求最佳更换周期。

**计算结果** 最佳更换周期为4459 (小时)

**不同惩罚费用对更换周期的影响**

b	0.05	0.1	0.5	1	10
T	3977	3918	3828	3797	3715

**演示 lamp.m**

大学数学实验

### 布置实验

**目的**

- 1) 掌握数理统计的基本概念;
- 2) 掌握用随机的方法(蒙特卡罗法)计算积分;
- 3) 对实际问题建立概率模型和进行计算.

**内容** 见网络学堂

