第 0 讲　概统基本概念一览

一　随机事件与概率

1. 随机试验

· 对随机现象进行观察、记录或试验，称为随机试验

2. 样本空间

· 称随机试验的所有可能结果构成的集合为样本空间，常用字母 S 来表示

3. 样本点

· 样本空间的每一个元素，即试验的每一个结果称为样本点

4. 随机事件

· 样本空间的任一子集称为随机事件，用大写字母 A 等表示；只含有一个样本点的事件称为基本事件

5. 发生/不发生

· 一次试验完成时，若试验所出现的结果属于某一事件，就称该事件发生，否则称该事件不发生

6. 必然事件与不可能事件

· 样本空间代表的事件必然发生，称为必然事件；空集称为不可能事件

7. 划分

· 设 S 为某一随机试验的样本空间， $B_1,B_2,\cdots,B_n$ 为该试验的一组事件，且满足

① $B_iB_j=\varnothing$， $i\neq j$

② $B_1\cup B_2\cup\cdots\cup B_n=S$

则称 $B_1,B_2,\cdots,B_n$ 为 S 的一个划分，或称 S 的一个**完备事件组**

8. 概率

· 设某随机试验对应的样本空间为 S，对 S 中的任一事件 A，定义一个实数 $P(A)$，若它满足：

① 非负性： $P(A)\geq 0$

② 规范性： $P(S)=1$

③ 可列可加性：对 S 中的可列个两两互不相容的事件 $A_1,A_2,\cdots,A_n,\cdots$，有

$$P(\bigcup_{i=1}^{+\infty}A_j)=\sum_{i=1}^{+\infty}P(A_i)$$

则称 $P(A)$ 为事件 A 发生的概率

9. 等可能概型（古典概型）

· 如果一个随机试验满足下面两个条件：

① 样本空间中样本点数有限（有限性）

② 出现每一个样本点的概率相等（等可能性）

就称这个试验为**等可能概型**，又称古典概型

10. **条件概率**

· 称在已知 B 事件发生的条件下事件 A 发生的概率为对应的条件概率

11. **独立**

· 设 A,B 为两随机事件，当 $P(A)P(B)=P(AB)$ 时，称 A,B 相互独立

二 随机变量

1. **随机变量**

· 设随机试验的样本空间为 S，若 X 为定义在样本空间上的单实值函数，$e\in S$，则称 X 为随机变量

2. **离散型随机变量**

· 若随机变量 X 的所有可能取值为有限个或可列个，则随机变量 X 为离散型随机变量

3. **概率分布律**

· 设 X 为离散型随机变量，若其可能取值为 $x_1,x_2,\cdots,x_k,\cdots$ 则称

$$P(X=x_k)=p_k$$

为 X 的概率分布律或**概率分布列**，简称 X 的分布律

4. **概率分布函数**

· 设 x 为任意实数，函数 $F(x)=P(X\leq x)$ 称为随机变量 X 的概率分布函数，简称分布函数

5. **连续型随机变量**

· 设随机变量 X 的分布函数为 $F(x)$，若存在一个非负实值函数 $f(x)$，$-\infty<x<+\infty$，使得对任意实数 x，有

$$F(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t$$

则称 X 为连续型随机变量

6. **概率密度函数**

· 称上个定义中的 $f(x)$ 为 X 的概率密度函数，简称密度函数

三 多维随机变量

1. **联合分布律**

· 称 $P(X=x_i,Y=y_j)=p_{ij}$ 为 (X,Y) 的联合概率分布律，简称联合分布律

2. **边际分布律**

· 称多维随机变量中单个随机变量的分布律为 (X,Y) 的边际分布律

3. 条件分布律

- 已知 (X,Y) 中一个随机变量取值确定的条件下，另一个随机变量的分布律称为 (X,Y) 的条件分布律

4. 联合分布函数

- 设二维随机变量 (X,Y)，对于任意的实数 x,y，称函数

$$F(x,y)=P(X\leq x,Y\leq y)$$

为随机变量 (X,Y) 的联合分布函数

5. 边际分布函数

- 二维随机变量中单个随机变量的分布函数为边际分布函数

6. 条件分布函数

- 已知 (X,Y) 中一个随机变量取值确定的条件下，另一个随机变量的分布函数为条件分布函数

7. 概率密度函数

- 设二维随机变量 (X,Y) 的联合分布函数 $F(x,y)$，若存在二元非负函数 $f(x,y)$，使对任意的实数 x,y 有

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)\,\mathrm{d}u\mathrm{d}v$$

则称 (X,Y) 为二维连续型随机变量，称 $f(x,y)$ 为 (X,Y) 的联合概率密度函数

8. 边际密度函数

- 二维随机变量中单个随机变量的密度函数为边际密度函数

9. 条件密度函数

- 某个随机变量取值确定的条件下，另一个随机变量的密度函数为条件密度函数

四　随机变量的数字特征

1. 数学期望

- 设离散型随机变量 X 的概率分布律为 $p(X_i)=p_i$， $i=1,2,\cdots$，若级数

$$\sum_{i=1}^{+\infty}x_ip_i$$

绝对收敛，则称该级数为 X 的数学期望，记为 $E(X)$

- 设连续型随机变量 X 的概率密度函数为 $f(x)$，若

$$\int_{-\infty}^{+\infty}|x|f(x)\mathrm{d}x$$

存在，则称 $\int_{-\infty}^{+\infty}xf(x)\mathrm{d}x$ 为 X 的数学期望，记为 $E(X)$

2. 方差

- 设随机变量 X 的数学期望 $E(X)$ 存在，若 $E[(X-E(X))^2]$ 存在，则称其为 X 的方差，记为 $\mathrm{Var}(X)$ 或 $D(X)$

3. **协方差**

· 对于数学期望都存在的随机变量 X 和 Y，当 $X-E(X)$ 和 $Y-E(Y)$ 的数学期望都存在时，称

$$\mathrm{Cov}(X,Y)=E[(X-E(X))(Y-E(Y))]$$

为 X 与 Y 的协方差

4. **相关系数**

· 对于随机变量 X 和 Y，当 $E(X^2)$ 与 $E(Y^2)$ 均在且 $\mathrm{Var}(X)\mathrm{Var}(Y)\neq 0$ 时，称

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

为 X 与 Y 的相关系数

5. **不相关**

· 若 X 与 Y 的相关系数 $\rho_{XY}=0$，则称 X 与 Y 不相关

五　大数定律

1. **依概率收敛**

· 设 $\{Y_n\}$ 为一随机变量序列，c 为一常数，若对任意的 $\varepsilon>0$，都有

$$\lim_{n\to+\infty}P(|Y_n-c|\geq\varepsilon)=0 \quad 或 \quad \lim_{n\to+\infty}P(|Y_n-c|<\varepsilon)=1$$

则称 $\{Y_n\}$ 依概率收敛于 c，记为 $Y_n\xrightarrow{P}c$，$n\to+\infty$

六　数理统计基础

1. **总体 个体**

· 研究对象的全体称为总体 X，总体中的每个成员称为个体

2. **样本**

· 数理统计的任务是从总体中抽取一部分个体进行分析，称为样本 X_i

3. **样本容量**

· 称样本的数量为样本容量

4. **样本值**

· 称样本 X_i 的观测值 x_i 为样本值

5. **随机样本**

· 如果从总体中抽取样本是随机的，称为随机样本

6. **简单随机样本**

· 设总体 X 是具有分布函数 F 的随机变量，$X_1,\cdots,X_n$ 是来自总体 X 的随机样本。若满足

①（独立性）$X_1,\cdots,X_n$ 是相互独立的随机变量

②（代表性）每个 X_i 都与总体 X 具有相同的分布函数

则称 $X_1,\cdots,X_n$ 为取自总体 X 的简单随机样本

7. **统计量**

· 设 $X_1,\cdots,X_n$ 是来自总体 X 的一个样本，$g(X_1,\cdots,X_n)$ 是样本 $X_1,\cdots,X_n$ 的函数，若 g 不含未知参数，则称 $g(X_1,\cdots,X_n)$ 是一个统计量

七　参数估计

1. **点估计**

· 构造一个适当的统计量 $\hat{\theta}$ 用来估计分布中的未知参数 θ 的值，称为点估计

2. **无偏估计**

· 设 $\theta\in\Theta$ 是总体 X 的待估参数，$X_1,\cdots,X_n$ 是来自总体 X 的样本。若估计量 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$ 的数学期望存在，且满足 $E(\hat{\theta})=\theta$，$\forall\theta\in\Theta$，则称 $\hat{\theta}$ 是 θ 的无偏估计

3. **有效性**

· 设 $\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 与 $\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n)$ 都是参数 θ 的无偏估计，若 $\forall\theta\in\Theta$，$\mathrm{Var}(\hat{\theta}_1)\le\mathrm{Var}(\hat{\theta}_2)$，且至少存在一个 θ 使得不等号成立，则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效

4. **均方误差**

· 设 $\hat{\theta}=\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是总体参数 θ 的估计量，称 $E[(\hat{\theta}-\theta)^2]$ 为估计量 $\hat{\theta}$ 的均方误差，称为 $\mathrm{Mse}(\hat{\theta})$

5. **相合估计**

· 设 $\hat{\theta}_n=\hat{\theta}(X_1,X_2,\cdots,X_n)$ 是总体参数 θ 的估计量，若对任意 $\varepsilon>0$，有

$$\lim_{n\to+\infty}P(|\hat{\theta}_n-\theta|<\varepsilon)=1$$

即 $\hat{\theta}_n$ 依概率收敛于 θ，则称 $\hat{\theta}_n$ 是 $\hat{\theta}$ 的相合估计，记为 $\hat{\theta}_n\xrightarrow{P}\theta$，$n\to+\infty$

6. **双侧置信区间**

· 设总体为 X，θ 为待估参数，$X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本，统计量 $\hat{\theta}_L=\hat{\theta}_L(X_1,X_2,\cdots,X_n)$ 和 $\hat{\theta}_U=\hat{\theta}_U(X_1,X_2,\cdots,X_n)$ 满足 $\hat{\theta}_L<\hat{\theta}_U$，且对给定的 $\alpha\in(0,1)$ 和任意的 θ，有

$$P(\hat{\theta}_L<\theta<\hat{\theta}_U)\ge1-\alpha$$

则称随机区间 $\hat{\theta}_L<\hat{\theta}_U$ 是参数 θ 的置信水平为 $1-\alpha$ 的置信区间，$\hat{\theta}_L,\hat{\theta}_U$ 分别称为是 θ 的置信水平 $1-\alpha$ 的双侧置信下限和双侧置信上限

7. **单侧置信区间**

· 对给定的 $\alpha\in(0,1)$，如果统计量满足

$$P(\hat{\theta}_L < \theta) \geq 1-\alpha$$

$$P(\theta < \hat{\theta}_U) \geq 1-\alpha$$

则称 $\hat{\theta}_L, \hat{\theta}_U$ 分别是 θ 的置信水平 $1-\alpha$ 的单侧置信下限和单侧置信上限

9. 枢轴量

- 设总体 X 的密度函数 $f(x;\theta)$，其中 θ 为待估参数，并设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的样本，如果样本和参数 θ 的函数 $G(X_1, X_2, \cdots, X_n, \theta)$ 的分布完全已知，且形式上不依赖其它未知参数，称 $G(X_1, X_2, \cdots, X_n, \theta)$ 为枢轴量

八 假设检验

1. 统计假设

- 对总体的分布或其中的参数提出的假设称为统计假设，简称为假设，一般用 H 表示

2. 原假设 备择假设

- 根据样本资料想得到推翻的假设称为原假设 H_0，与其完全相反的假设称为备择假设 H_1

3. 检验形式

- 单侧检验：H_0： $\theta \geq \theta_0$ $\quad H_1$： $\theta < \theta_0$ （左侧检验）

  H_0： $\theta \leq \theta_0$ $\quad H_1$： $\theta > \theta_0$ （右侧检验）
- 双侧检验：H_0： $\theta = \theta_0$ $\quad H_1$： $\theta \neq \theta_0$

4. 检验统计量

- 取值与原假设是否成立有密切关系的统计量称为检验统计量

5. 拒绝域

- 如果检验统计量的值落入了拒绝域，则应拒绝原假设

6. 两类错误

- 如果拒绝了一个真实的原假设，则称为第 I 类错误
- 如果接受了一个错误的原假设，则称为第 II 类错误

7. 显著性水平

- 假设检验的要求是犯 I 类错误的概率不超过显著性水平 α 时，尽可能减少犯第 II 类错误的概率

8. P_- 值

- 当原假设 H_0 为真时，检验统计量取比观察到的结果更为极端的数值的概率称为 P_-值