# 概率论知识点整理 $\text{by gMr}$

> 叠甲：只是 **个人认为** 比较需要记忆的知识点，有些对我来说比较显然或者好记的就没有放进去，因此不是齐全的知识点。完整复习还请参照书或 $PPT$
>
> 要是哪里打错了欢迎指出，本人不对其中错误导致的分数损失负责（
>
> 以及温馨提醒，若追求高分刷题是很有必要的，许多易错点考点唯有自己做过才能知晓。

**你说得对，但是 emu 可爱捏**

# 第一章 随机事件

### 事件

$A-B=A-AB=A\overline{B}$

### 条件概率

$P(A\mid B)=\dfrac{P(AB)}{P(B)}$

$P(\overline{A}\mid B)=1-P(A\mid B)$

### 乘法公式推广

$P(ABC)=P(A)P(B\mid A)P(C\mid AB)$

### 全概率公式

$P(B)=\sum\limits_{i=1}^{n}P(A_i)P(B\mid A_i)$

### 贝叶斯公式

$P(A_i\mid B)=\dfrac{P(A_i)P(B\mid A_i)}{\sum\limits_{i=1}^{n}P(A_i)P(B\mid A_i)}$

# 第二章 随机变量及其分布

### 一维随机变量分布律（分布列）

$P\{X=x_i\}=p_i,(i=1,2,\ldots,)$

或

| $X$ | $x_1$ | $x_2$ | $\ldots$ | $x_n$ | $\ldots$ |
|---|---|---|---|---|---|
| $P$ | $p_1$ | $p_1$ | $\ldots$ | $p_n$ | $\ldots$ |

### 典型一维离散型随机变量及其分布

- **退化分布**

  $P\{X=c\}=1$
- **离散型均匀分布**

  $P\{X=x_k\}=\dfrac{1}{n},(k=1,2,\ldots,n)$
- **两点分布**

  $X\sim B(1,p)$

  $P\{X=k\}=p^k(1-p)^{1-k},(k=0,1)$
- **二项分布**

  $X\sim B(n,p)$

  $P\{X=k\}=C_n^kp^k(1-p)^{n-k},(n\in N,k=0,1,\ldots,n)$
- **泊松分布**

  $X\sim P(\lambda)$

  $P\{X=k\}=\dfrac{\lambda^k}{k!}e^{-\lambda},(k=0,1,\ldots)$

  其中 $\lambda$ 表示 $X$ 的平均值

  > 泊松分布是二项分布的极限分布，即
  >
  > $B(n,p)\xrightarrow{n\to\infty,p_n=\frac{\lambda}{n}\to 0}P(\lambda)$

- **几何分布**

  $X$：伯努利实验中 $A$ 首次发生时实验的次数

  $X \sim Ge(p)$

  $P\{X=k\}=(1-p)^{k-1}p$

- **超几何分布**

  $X$：表示 $N$ 件产品中有 $M$ 件次品 $(M \leq N)$，从中任取 $n$ 件 $(n \leq N)$，其中的次品数

  $X \sim H(N,n,M)$

  $P\{X=k\}=\frac{C_M^k C_{N-M}^{n-k}}{C_N^n},(k=0,1,\dots,\min\{M,n\})$

## 典型一维连续型随机变量及其分布

- **均匀分布**

  $X \sim U[a,b]$

  $$p(x)=\begin{cases}\frac{1}{b-a}, & a \leq x \leq b,\\ 0, & \text{其他}.\end{cases}$$

  分布函数：

  $$F(x)=\begin{cases}0, & x<a,\\ \frac{x-a}{b-a}, & a \leq x < b,\\ 1, & b \leq x.\end{cases}$$

- **正态分布**

  $X \sim N(\mu,\sigma^2)$

  $$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},(-\infty<x<\infty)$$

  > 这个公式要记住

  **特别地，标准正态分布**

  $X \sim N(0,1)$

  $$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},(-\infty<x<\infty)$$

  其分布函数为 $\Phi(x)$

  **性质：**

  $\Phi(-x)=1-\Phi(x)$，即密度函数关于 $y$ 轴对称

  **标准化：**

  $$X \sim N(\mu,\sigma^2) \implies \frac{X-\mu}{\sigma} \sim N(0,1)$$

- **指数分布**

  $X \sim Exp(\lambda)$

  $$p(x)=\begin{cases}\lambda e^{-\lambda x}, & x \geq 0,\\ 0, & x<0.\end{cases}$$

  $$F(x)=\begin{cases}1-e^{-\lambda x}, & x \geq 0,\\ 0, & x<0.\end{cases}$$

  > 泊松分布表示单位时间内随机事件发生的 **次数**，指数分布表示单位时间内随机事件发生的 **时间间隔**
  >
  > $N \sim p(\lambda), T \sim Exp(\lambda)$
  >
  > $E(N)=\lambda, E(T)=\frac{1}{\lambda}$

  **特点：**无记忆性

  已经使用时间 $s$，还能使用至少时间 $t$ 的概率 $=$ 至少使用时间 $t$ 的概率

  $$P\{X>s+t \mid X>s\}=\frac{P\{X>s+t,X>s\}}{P\{X>s\}}=\frac{P\{X>s+t\}}{P\{X>s\}}$$

$\because P\{X>s\}=1-F(s)=e^{-\lambda s}$

$\therefore P\{X>s+t \mid X>s\}=\dfrac{e^{-\lambda(s+t)}}{e^{-\lambda s}}=e^{-\lambda t}=P\{X>t\}$

## $(X,Y)$ 的联合分布律（分布列）

**离散型：**

$P\{X=x_i,Y=y_i\}=p_{ij},(i,j=1,2,\ldots)$

或

| $X\backslash Y$ | $y_1$ | $y_2$ | $\cdots$ | $y_j$ | $\cdots$ |
|---|---|---|---|---|---|
| $x_1$ | $p_{11}$ | $p_{12}$ | $\cdots$ | $p_{1j}$ | $\cdots$ |
| $x_2$ | $p_{21}$ | $p_{22}$ | $\cdots$ | $p_{2j}$ | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | |
| $x_i$ | $p_{i1}$ | $p_{i2}$ | $\cdots$ | $p_{ij}$ | $\cdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | |

### 典型二维连续型随机变量及其分布

- **均匀分布**

  $p(x,y)=\begin{cases}\dfrac{1}{S}, & (x,y)\in D,\\ 0, & \text{其他}.\end{cases}$

- **二维正态分布**

  $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$

  $p(x,y)=\dfrac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-\frac{2\rho(x-\mu_1)(x-\mu_2)}{\sigma_1\sigma_2}+\frac{(x-\mu_2)^2}{\sigma_2^2}\right]},(-\infty<x<\infty,-\infty<y<\infty)$

  > 这么长个公式疑似不用背

### 独立

$p(x,y)=p_X(x)p_Y(y)$ 则称 $X,Y$ 相互独立。

### 边缘分布

- **密度函数**

  $p_X(x)=\int_{-\infty}^{\infty}p(x,y)\,dy$

  $p_Y(y)=\int_{-\infty}^{\infty}p(x,y)\,dx$

- **分布函数**

  $F_X(x)=\int_{-\infty}^{x}\left[\int_{-\infty}^{\infty}p(u,y)\,dy\right]du=F_{XY}(x,+\infty)$

  $F_Y(y)=\int_{-\infty}^{y}\left[\int_{-\infty}^{\infty}p(x,u)\,dx\right]du=F_{XY}(+\infty,y)$

> 二元正态分布的边缘分布是一元正态分布。
>
> 边缘分布均为正态分布的随机变量不一定是二维正态分布。

联合分布 **能推出** 边缘分布，边缘分布 **推不出** 联合分布。

### 条件分布

- **离散型**

  $p_{j|i}=P\{Y=y_j\mid X=x_i\}=\dfrac{P\{X=x_i,Y=y_j\}}{P\{X=x_i\}}=\dfrac{p_{ij}}{p_{i\cdot}}$

- **连续型**

  $p(x\mid y)=\dfrac{p(x,y)}{p_Y(y)}\implies P\{X\le x\mid Y=y\}=\int_{-\infty}^{x}p(u\mid y)\,du$

## 随机变量的函数的分布

- **一维随机变量函数** $Y = f(X)$
  - **离散型**

    $P\{Y = y_k\} = P\{f(X) = y_k\}$
  - **连续型**

    **方法1：**

    $F_Y(y) = P\{Y \le y\} = P\{f(X) \le y\} = \int_{f(x) \le y} p_X(x)\,dx$

    $p_Y(y) = F'_Y(y)$

    **方法2：**

    要求：$f(x)$ 可导且单调，即存在反函数。

    > 原函数不单调可以分成多段单调函数。

    $f(x)$ **单增：**

    $F_Y(y) = P\{Y \le y\} = P\{f(X) \le y\} = P\{X \le f^{-1}(y)\} = \int_{-\infty}^{f^{-1}(y)} p_X(x)\,dx$

    求导得：

    $p_Y(y) = p_X[f^{-1}(y)] \cdot [f^{-1}(y)]'$

    $f(x)$ **单减：**

    $F_Y(y) = P\{Y \le y\} = P\{f(X) \le y\} = P\{X \ge f^{-1}(y)\} = \int_{f^{-1}(y)}^{\infty} p_X(x)\,dx$

    求导得：

    $p_Y(y) = p_X[f^{-1}(y)] \cdot [-f^{-1}(y)]'$

    > 做题发现这两种都可以记成 $p_Y(y) = p_X[f^{-1}(y)] \cdot |[f^{-1}(y)]'|$ ，即导数取个绝对值
    >
    > 这边很常考

- **二维随机变量函数** $Z = f(X, Y)$
  - **离散型**

    $P\{Z = z_k\} = P\{f(X, Y) = z_k\}$

    > 若 $X_1 \sim P(\lambda_1), X_2 \sim P(\lambda_2)$ 则 $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$

  - **连续型**
    - $Z = X + Y$

      $p_Z(z) = \int_{-\infty}^{+\infty} p(x, z - x)\,dx$

      > $X, Y$ 独立时 $p(x, z - x) = p_X(x)\,p_Y(z - x)$

    - $Z = X - Y$

      $p_Z(z) = \int_{-\infty}^{+\infty} p(x, x - z)\,dx$

      > $X, Y$ 独立时 $p(x, x - z) = p_X(x)\,p_Y(x - z)$

    - $Z = \dfrac{X}{Y}$

      $p_Z(z) = \int_{-\infty}^{+\infty} |y|\,p(yz, y)\,dy$

      > $X, Y$ 独立时 $p(yz, y) = p_X(yz)\,p_Y(y)$

    - $M = \max\{X, Y\}, N = \min\{X, Y\}$

      $X, Y$ 独立时，

      $F_M(z) = F_X(z)F_Y(z)$

      $F_N(z) = 1 - [1 - F_X(z)][1 - F_Y(z)]$

      **更一般地，**

      $M = \max\{X_1, X_2, \ldots, X_n\}, N = \min\{X_1, X_2, \ldots, X_n\}$

      $X_1, X_2, \ldots, X_n$ 独立时，

      $F_M(z) = \prod_{i=1}^{n} F_{X_i}(z)$

$F_N(z) = 1 - \prod_{i=1}^{n}[1 - F_{X_i}(z)]$

> 这边也很常考，公式要记

# 第三章 随机变量的数字特征

### 常见分布的期望与方差

| 分布 | 期望 | 方差 |
|---|---|---|
| $X \sim B(1,p)$ 两点分布 | $p$ | $p(1-p)$ |
| $X \sim B(n,p)$ 二项分布 | $np$ | $np(1-p)$ |
| $X \sim P(\lambda)$ 泊松分布 | $\lambda$ | $\lambda$ |
| $X \sim Ge(p)$ 几何分布 | $\frac{1}{p}$ | $\frac{1-p}{p^2}$ |
| $X \sim U(a,b)$ 均匀分布 | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ |
| $X \sim N(\mu,\sigma^2)$ 正态分布 | $\mu$ | $\sigma^2$ |
| $X \sim Exp(\lambda)$ 指数分布 | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |

### 期望

- **期望计算**
  - $Y = f(X)$
    $E(Y) = E[f(X)] = \int_{-\infty}^{+\infty} f(x)p_X(x)\,dx$
  - $Z = f(X,Y)$
    $E(Z) = E[f(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)p(x,y)\,dxdy$
- **期望性质**
  1. $E(C) = C$
  2. $E(CX) = CE(X)$
  3. $E(X \pm Y) = E(X) \pm E(Y)$，没有限制，很神奇吧
  4. $E(XY) = E(X)E(Y)$，当 $X, Y$ 独立时

> 特别特别喜欢考期望计算，而且还都是贼够吧难算那种
>
> 小技巧：
>
> **伽玛函数**
>
> - $\Gamma(s) = \int_0^{+\infty} t^{s-1}e^{-t}dt$
>   $\Gamma(n+1) = n!$
> - 可以用于计算积分 $\int_0^{+\infty} x^n e^{-x}dx = \Gamma(n+1) = n!$

### 方差

- **方差计算**
  $D(X) = E[X - E(X)]^2 = E(X^2) - E^2(X)$
  $\implies E(X^2) = D(X) + E^2(X)$
- **方差性质**
  1. $D(C) = 0$
  2. $D(kX) = k^2D(X)$
  3. $D(X \pm Y) = D(X) + D(Y)$，当 $X, Y$ 独立时
  4. $P\{|X - \mu| < \varepsilon\} > 1 - \frac{\sigma^2}{\varepsilon^2}$

$$P\{|X-\mu|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2}$$

**（切比雪夫不等式）** $\leftarrow$ 竟然考过，，，感觉下面那个形式比较好记

5. $D(X)\leq E(X-C)^2$

### 协方差

- **协方差计算**

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

  $\implies E(XY)=E(X)E(Y)+Cov(X,Y)$

- **协方差性质**

  1. $Cov(X,Y)=Cov(Y,X)$
  2. $Cov(aX,bY)=abCov(X,Y)$
  3. $Cov(X_1\pm X_2,Y)=Cov(X_1,Y)\pm Cov(X_2,Y)$
  4. $X,Y$ 独立 $\implies X,Y$ 不相关 $\iff Cov(X,Y)=0$
  5. $Cov(X,X)=D(X)$
  6. $D(X\pm Y)=D(X)+D(Y)\pm 2Cov(X,Y)$

> $\text{X,Y独立}\implies\text{X,Y不相关},\quad\text{X,Y不相关}\not\Longrightarrow\text{X,Y独立}$
>
> 只有在 $X,Y$ 服从二维正态分布时，$Cov(X,Y)=0\iff X,Y$ 不相关 $\iff X,Y$ 独立
>
> $X,Y$ **分别服从正态分布时不能推出该结论**

### 相关系数

- **相关系数定义**

  $\rho_{XY}=\dfrac{Cov(X,Y)}{\sqrt{D(X)}\cdot\sqrt{D(Y)}}$

  $|\rho_{X,Y}|$ 越大表示越相关，$\rho_{X,Y}$ 正负性代表了正相关还是负相关

  $\rho=0\iff Cov(X,Y)=0\iff\text{X与Y不相关}$

- **相关系数性质**

  1. $|\rho_{XY}|\leq 1$
  2. $|\rho_{XY}|=1\iff P\{Y=aX+b\}=1$

> 二维正态分布密度函数中 $\rho$ 代表了 $X$ 与 $Y$ 的相关系数

# 第四章 极限定理

> 这边几个定理倒是考

### 四种收敛

- **依分布收敛**

  随机变量序列 $\{Y_n\}$ 和随机变量 $Y$

  $\lim\limits_{n\to\infty}F_n(x)=F(x)$

  即 $\lim\limits_{n\to\infty}P\{Y_n\leq x\}=P\{Y\leq x\}$

  则称随机变量序列 $\{Y_n\}$ **依分布收敛** 于随机变量 $Y$ 。是较弱的一种收敛性

- **依概率收敛**

  随机变量序列 $\{Y_n\}$ 和随机变量 $Y$ ，若对于任意实数 $\varepsilon\geq 0$ ，

  有 $\lim\limits_{n\to\infty}P\{|Y_n-Y|\geq\varepsilon\}=0$

  则称随机变量序列 $\{Y_n\}$ **依概率收敛** 于随机变量 $Y$ 。简记为 $Y_n\xrightarrow{p}Y$

- **$r$ 阶收敛**

  $\lim\limits_{n\to\infty}E|Y_n-Y|^r=0$

- **以概率 $1$ 收敛**

$$P\{\lim_{n\to\infty} Y_n = Y\} = 1$$

> 听说不咋考，考了别怪我（

### 大数定律

对于随机变量序列 $X_1, X_2, ..., X_n$ ，存在一个常数序列 $a_n$

且 $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} a_n$

则称随机变量序列 $\{Y_n\}$ 服从大数定律。

- **切比雪夫大数定律**

  随机变量序列 $X_1, X_2, ..., X_n$ **两两不相关，且方差存在公共上界** $C$

  则 $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n} E(X_i)$

- **伯努利大数定律**

  随机变量序列 $X_1, X_2, ..., X_n$ **独立同分布**， $X_i \sim B(1,p)$ , $\mu_n$ 为事件发生次数

  则 $\frac{\mu_n}{n} \xrightarrow{p} p$

  > 因此可以用频率估计概率

- **辛钦大数定律**

  随机变量序列 $X_1, X_2, ..., X_n$ **独立同分布**，且 $E(X_i) = \mu$ 存在

  则 $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E(X_i)$

### 中心极限定理

- **棣莫佛-拉普拉斯中心极限定理**

  若 $Y_n \sim B(n,p)$ ，则 $n \to \infty$ 时

  $$Y_n \sim AN(np, np(1-p))$$

  $$\frac{Y_n - np}{\sqrt{np(1-p)}} \sim AN(0,1)$$

  > 正态分布是二项分布的极限分布

- **林德伯格-列维中心极限定理**

  随机变量序列 $X_1, X_2, ..., X_n$ **独立同分布**，且 $E(X_i) = \mu, D(X_i) = \sigma^2$ ，则 $n \to \infty$ 时

  $$\sum_{i=1}^{n} X_i \sim AN(n\mu, n\sigma^2)$$

  $$Y_n^* = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \sim AN(0,1)$$

  $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim AN(\mu, \frac{\sigma^2}{n})$$

- **李雅普诺夫定理**

  随机变量序列 $X_1, X_2, ..., X_n$ **独立不同分布。**且 $E(X_i) = \mu_i, D(X_i) = \sigma_i^2$ ，则 $n \to \infty$ 时

  $$\sum_{i=1}^{n} X_i \sim AN(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$$

  $$Y_n^* = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \sim AN(0,1)$$

  > 条件懒得写了应该不考吧。

# 第五章 统计量及抽样分布

## 统计量

- **定义**

  设 $(X_1,X_2,\ldots,X_n)$ 是来自总体 $X$ 的一个样本，$f(X_1,X_2,\ldots,X_n)$ 是 $X_1,X_2,\ldots,X_n$ 的函数，若 $f$ 中不含任何 **关于总体 $X$ 的未知参数**，则称 $f(X_1,X_2,\ldots,X_n)$ 是一个 **统计量** 。

- **常用统计量**

| 统计量 | 公式 |
| --- | --- |
| 样本均值 $\bar{X}$ | $\bar{X}=\frac{1}{n}\sum\limits_{i=1}^{n}X_i$ |
| 样本方差 $S_n^2$ | $S_n^2=\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n}(\sum\limits_{i=1}^{n}X_i^2-n\bar{X}^2)$ |
| 样本标准差 $S_n$ | $S_n=\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}$ |
| 修正样本方差 $S_n^{*2}$ | $S_n^{*2}=\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2=\frac{n}{n-1}S_n^2$ |
| 修正样本标准差 $S_n^*$ | $S_n^*=\sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i-\bar{X})^2}$ |
| 样本 $k$ 阶原点矩 $A_k$ | $A_k=\frac{1}{n}\sum\limits_{i=1}^{n}X_i^k$ |
| 样本 $k$ 阶中心距 $B_k$ | $B_k=\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\bar{X})^k$ |

- **统计量常用性质**
  1. $E(\bar{X})=E(X)=\mu$
  2. $D(\bar{X})=\frac{1}{n}D(X)=\frac{1}{n}\sigma^2$
  3. $E(S_n^2)=\frac{n-1}{n}D(X)=\frac{n-1}{n}\sigma^2$
  4. $E(S_n^{*2})=D(X)=\sigma^2$
  5. 若 $E(A_k)=\alpha_k$，则 $A_k\xrightarrow{p}\alpha_k(n\to\infty)$
  6. $X\sim N(\mu,\sigma^2),D(S_n^2)=?$

     $\frac{nS_n^2}{\sigma^2}\sim\chi^2(n-1)$

     $D(\frac{nS_n^2}{\sigma^2})=2(n-1)\implies D(S_n^2)=\frac{2(n-1)\sigma^4}{n^2}$

- **次序统计量**

  设 $(X_1,X_2,\ldots,X_n)$ 是来自总体 $X$ 的一个样本，$(x_1,x_2,\ldots,x_n)$ 是其一个观测值，将观测值按 **由小到大** 的次序重新排列为

  $$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$$

  当 $(X_1,X_2,\ldots,X_n)$ 取值为 $(x_1,x_2,\ldots,x_n)$ 时，定义 $X_{(k)}$ 取值为 $x_{(k)}(k=1,2,\ldots,n)$ ，由此得到

  $$(X_{(1)},X_{(2)},\ldots,X_{(n)})$$

  称为 $(X_1,X_2,\ldots,X_n)$ 的 **次序统计量** 。

  **特别地，**

  $X_{(1)}=\min\limits_{1\le i\le n}X_i$ 称为 **最小次序统计量**

  $X_{(n)}=\max\limits_{1\le i\le n}X_i$ 称为 **最大次序统计量**

  最小次序统计量 $X_{(1)}$ 的**分布密度**为:

$p_{X_{(1)}}(x) = n[1-F(x)]^{n-1}f(x)$

最大次序统计量 $X_{(n)}$ 的**分布密度**为：

$p_{X_{(n)}}(x) = n[F(x)]^{n-1}f(x)$

> 爱考分布密度

- **经验分布函数**

  设 $(X_1, X_2, \ldots, X_n)$ 是总体 $X$ 的一个样本，$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ 为样本的次序统计量，$(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ 为其观测值，设 $x$ 为任一实数，称函数

  $$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \le x < x_{(k+1)} \\ 1, & x_{(n)} \le x \end{cases}$$

  为总体 $X$ 的 **经验分布函数** 。

## 统计分布

- **卡方分布** $\chi^2$
  - **定义**

    设 $X_1, X_2, \ldots, X_n$ 均是来自总体 $N(0,1)$ 的样本，则称 $\chi_n^2 = \sum\limits_{i=1}^{n} X_i^2$ 服从 **自由度** 为 $n$ 的 $\chi^2$ 分布，记作 $\chi_n^2 \sim \chi^2(n)$
  - **性质**
    1. $Y_1 \sim \chi^2(n_1), Y_2 \sim \chi^2(n_2)$，且 $Y_1, Y_2$ 独立

       则 $Y_1 + Y_2 \sim \chi^2(n_1+n_2)$
    2. $E(\chi_n^2) = n, D(\chi_n^2) = 2n \leftarrow$ 这个要用
    3. $n \to \infty$ 时，$\chi_n^2 \sim AN(n, 2n)$
- $t$ **分布**
  - **定义**

    设 $X \sim N(0,1), Y \sim \chi^2(n)$，且 $X, Y$ 独立，则称 $T = \dfrac{X}{\sqrt{Y/n}}$ 服从 **自由度** 为 $n$ 的 $t$ 分布，记为 $T \sim t(n)$
  - **性质**
    1. $t$ 分布概率密度函数 **关于 $y$ 轴对称**
    2. $E(T) = 0, D(T) = \dfrac{n}{n-2}$
    3. $n \to \infty$ 时，$T \sim AN(0,1)$
- $F$ **分布**
  - **定义**

    设 $X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$，且 $X, Y$ 独立，则称 $F = \dfrac{X/n_1}{Y/n_2}$ 服从 **自由度** 为 $(n_1, n_2)$ 的 $F$ 分布，记为 $F \sim F(n_1, n_2)$
  - **性质**
    1. $E(F) = \dfrac{n_2}{n_2-2}, D(F) = \dfrac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)} \leftarrow$ 疑似不咋考
    2. $F \sim F(n_1, n_2) \implies \dfrac{1}{F} \sim F(n_2, n_1)$
    3. $F$ 分布极限分布也是正态分布
    4. $T \sim t(n) \implies T^2 \sim F(1, n)$

## 概率分布的分位数

- **上 $\alpha$ 分位点**
  - **定义**

    对于给定总体 $X$ 和给定的 $\alpha(0 < \alpha < 1)$，若存在 $x_\alpha$ 使得 $P\{X > x_\alpha\} = \alpha$ 则称 $x_\alpha$ 为 $X$ 的分布的 **上 $\alpha$ 分位点**
  - **正态分布** $N$

    上 $\alpha$ 分位点称为 $u_\alpha$

显然 $\Phi(u_\alpha)=1-\alpha$

**对称性：** $u_\alpha=-u_{1-\alpha}$

- **$t$ 分布**

  上 $\alpha$ 分位点称为 $t_\alpha(n)$

  **对称性：** $t_\alpha(n)=-t_{1-\alpha}(n)$

  $n>45$ 时，$t_\alpha(n)\approx u_\alpha$
- **$\chi^2$ 分布**

  上 $\alpha$ 分位点称为 $\chi^2_\alpha(n)$
- **$F$ 分布**

  上 $\alpha$ 分位点称为 $F_\alpha(n_1,n_2)$

  $F_\alpha(n_1,n_2)=\dfrac{1}{F_{1-\alpha}(n_2,n_1)}$

### 统计分布

- **样本来自一个正态总体 $X\sim N(\mu,\sigma^2)$**
  - $\bar{X}\sim N(\mu,\dfrac{\sigma^2}{n})$

    $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$
  - $\dfrac{nS_n^2}{\sigma^2}=\dfrac{(n-1)S_n^{*2}}{\sigma^2}\sim\chi^2(n-1)$
  - $\dfrac{\bar{X}-\mu}{S_n^*/\sqrt{n}}=\dfrac{\bar{X}-\mu}{S_n/\sqrt{n-1}}\sim t(n-1)$
- **样本来自两个正态总体 $X\sim N(\mu_1,\sigma_1^2),Y\sim N(\mu_2,\sigma_2^2)$，且相互独立**
  - $\bar{X}-\bar{Y}\sim N(\mu_1-\mu_2,\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2})$

    $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim N(0,1)$
  - $\sigma_1^2=\sigma_2^2=\sigma^2$ **但未知时：**

    $\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$

    其中 $S_w=\sqrt{\dfrac{(n_1-1)S_1^{*2}+(n_2-1)S_2^{*2}}{n_1+n_2-2}}=\sqrt{\dfrac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}}$
  - $\dfrac{S_1^{*2}/\sigma_1^2}{S_2^{*2}/\sigma_2^2}\sim F(n_1-1,n_2-1)$

> $nS_n^2=(n-1)S_n^{*2}=\sum\limits_{i=1}^n(X_i-\bar{X})^2$ 很常用

# 第六章 参数估计

### 参数的点估计

- **矩估计法**
  - **基本思想**

    使用样本矩匹配总体矩的方式进行参数估计。
  - **理论基础**

    样本矩 $A_k$ 和偏差矩 $B_k$ 的极限分别趋向于总体矩和总体偏差矩：

    $A_k=\dfrac{1}{n}\sum\limits_{i=1}^n X_i^k\xrightarrow{p}E(X^k)$

    $B_k=\dfrac{1}{n}\sum\limits_{i=1}^n(X_i-\bar{X})^k\xrightarrow{p}E[X-E(X)]^k$

**特别地，**

$A_1 = \bar{X} \xrightarrow{p} E(X)$

$B_2 = S_n^2 \xrightarrow{p} D(X)$

- **步骤**
  1. 根据样本计算 **样本矩** $A_k, B_k$ 。
  2. 假设 **总体矩** 等于 **样本矩** ，解出参数。

- **最大似然估计**
  - **基本思想**
    使得样本的联合分布最有利于当前观测值出现，即取得概率最大。
  - **步骤**
    1. 构造 **似然函数**：$L(\theta_1, \theta_2, \ldots, \theta_n) = \prod_{i=1}^{n} p(x_i)$ ，其中 $\theta$ 为分布函数的未知参数。
    2. 解 **似然方程**：$\dfrac{\partial \ln L}{\partial \theta_i} = 0$ ，得到参数最大似然估计值。

- **常见结论**
  对 **几何分布、二项分布、泊松分布、指数分布**，$E(x)$ 最大似然估计量为 **样本均值** $\bar{X}$ 。
  对 **正态分布**，期望 $\mu$ 、方差 $\sigma^2$ 的最大似然估计量为 **样本均值** $\bar{X}$ 、**样本方差** $S_n^2$ 。
  对 **均匀分布**， 下限 $\theta_1$ 的最大似然估计量为 **最小次序统计量** $X_{(1)}$ ，上限 $\theta_2$ 的最大似然估计量为 **最大次序统计量** $X_{(n)}$ 。

  > 这几个结论貌似也不咋考，题目好像都是给你一个莫名其妙的密度函数让你算矩估计和最大似然估计，不过知道没坏处（
  >
  > ↑反转了，有几个结论还是考的。

### 估计量的评价标准

- **无偏性**：$E(\hat{\theta}) = \theta$
  样本均值 $\bar{X}$ 是总体期望 $\mu$ 的 **无偏估计** 。
  样本方差 $S_n^2$ 是总体方差 $\sigma^2$ 的 **渐进无偏估计** 。
  修正样本方差 $S_n^{*2}$ 是总体方差 $\sigma^2$ 的 **无偏估计** 。
- **有效性**：$D(\hat{\theta}_1) < D(\hat{\theta}_2)$
  最小方差无偏估计量 $\hat{\theta}$:
  满足 $E(\hat{\theta}) = \theta$ 且 $D(\hat{\theta}) = \dfrac{1}{nI(\theta)}$ ，
  其中 $I(\theta) = E\left[\dfrac{\partial \ln p}{\partial \theta}\right]^2 (= -E\left[\dfrac{\partial^2 \ln p}{\partial \theta^2}\right])$ 称为费希尔 (Fisher) 信息量，其中 $p$ 为总体密度函数。

  > 一般来说，样本均值 $\bar{X}$ 都是总体期望 $\mu$ 的 **最小方差** **无偏估计**。

- **相合性**：$\hat{\theta} \xrightarrow{p} \theta \implies \lim\limits_{n \to \infty} E(\hat{\theta}_n) = \theta, \lim\limits_{n \to \infty} D(\hat{\theta}_n) = 0$

  矩估计量一般是相合估计量。

  相合性只有在样本容量很大时才展现优越性，所以一般较少考虑。

  是估计量应当满足的基本要求。

> 这边这一坨好像也只需要知道无偏性和相合性，题目就喜欢让你算这俩，就是分别判断是否 $E(\hat{\theta}) = \theta$ 和 $D(\hat{\theta}) \xrightarrow{n \to +\infty} 0$
>
> 反转了有几个结论好像还是要背一下，有效性确实不怎么考

### 参数的区间估计

- **基本思想**
  尝试给出一个参数可能的取值区间。
- **定义**
  存在两个统计量 $\hat{\theta}_1, \hat{\theta}_2$ ，使得
  $P\{\hat{\theta}_1(X_1, X_2, \ldots, X_n) < \theta < \hat{\theta}_2(X_1, X_2, \ldots, X_n)\} = 1 - \alpha$
  则称区间 $[\hat{\theta}_1, \hat{\theta}_2]$ 为参数 $\theta$ 的 **置信度** 为 $1 - \alpha$ 的 **置信区间** 。

- **单个正态总体均值的区间估计**
  - **正态总体 $X$ 的方差 $\sigma^2$ 已知，求 $\mu$ 置信区间**

    $$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \implies P\{|\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}| \le u_{\alpha/2}\} = 1-\alpha$$

    故置信度为 $1-\alpha$ 的置信区间为 $[\bar{X}-u_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}+u_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$
  - **正态总体 $X$ 的方差 $\sigma^2$ 已知，求 $\mu$ 置信区间**

    $$\frac{\bar{X}-\mu}{S_n^*/\sqrt{n}} \sim t(n-1) \implies P\{|\frac{\bar{X}-\mu}{S_n^*/\sqrt{n}}| \le t_{\alpha/2}(n-1)\} = 1-\alpha$$

    故置信度为 1-$\alpha$ 的置信区间为 $[\bar{X}-t_{\alpha/2}(n-1)\frac{S_n^*}{\sqrt{n}}, \bar{X}+t_{\alpha/2}(n-1)\frac{S_n^*}{\sqrt{n}}]$
- **单个正态总体方差的区间估计**
  - **正态总体 $X$ 的期望 $\mu$ 未知，求 $\sigma^2$ 置信区间**

    $$\frac{(n-1)S_n^{*2}}{\sigma^2} \sim \chi^2(n-1) \implies P\{\chi^2_{1-\alpha/2}(n-1) \le \frac{(n-1)S_n^{*2}}{\sigma^2} \le \chi^2_{\alpha/2}(n-1)\} = 1-\alpha$$

    故置信度为 $1-\alpha$ 的置信区间为 $\left[\frac{(n-1)S_n^{*2}}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S_n^{*2}}{\chi^2_{1-\alpha/2}(n-1)}\right]$
- **两个正态总体均值差的区间估计**

  $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ 且独立
  - **方差 $\sigma_1^2, \sigma_2^2$ 已知，期望 $\mu_1, \mu_2$ 未知，求 $\mu_1-\mu_2$ 置信区间**

    $$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1) \implies P\{|\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}| \le u_{\alpha/2}\} = 1-\alpha$$

    故置信度为 $1-\alpha$ 的置信区间为 $\left[(\bar{X}-\bar{Y})-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, (\bar{X}-\bar{Y})+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$
  - **$\sigma_1^2=\sigma_2^2=\sigma^2$ 但未知，求 $\mu_1-\mu_2$ 置信区间**

    $$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \implies P\{|\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}| \le t_{\alpha/2}(n_1+n_2-2)\} = 1-\alpha$$

    故置信度为 $1-\alpha$ 的置信区间为

    $$\left[(\bar{X}-\bar{Y})-t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, (\bar{X}-\bar{Y})+t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$
- **两个正态总体方差比的区间估计**

  $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ 且独立
  - **$\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ 都未知**

    $$\frac{S_1^{*2}/S_2^{*2}}{\sigma_1^2/\sigma_2^2} \sim F(n_1-1, n_2-1) \implies P\{F_{1-\alpha/2}(n_1-1,n_2-1) \le \frac{S_1^{*2}/S_2^{*2}}{\sigma_1^2/\sigma_2^2} \le F_{\alpha/2}(n_1-1,n_2-1)\} = 1-\alpha$$

    故置信度为 $1-\alpha$ 的置信区间为

    $$\left[\frac{S_1^{*2}/S_2^{*2}}{F_{\alpha/2}(n_1-1,n_2-1)}, \frac{S_1^{*2}/S_2^{*2}}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right] = \left[\frac{S_1^{*2}/S_2^{*2}}{F_{\alpha/2}(n_1-1,n_2-1)}, F_{\alpha/2}(n_2-1,n_1-1)S_1^{*2}/S_2^{*2}\right]$$

> 真是一堆又臭又长的公式，，，还是背一下上面的然后推吧

# 第七章 假设检验

> 跟第六章参数的区间估计其实差不多，都是狂背公式就行了。因此省略一点。具体过程还是看题目吧。

### 两类错误

- **第 $I$ 类错误（弃真错误）**

  原假设为 **真** ，观察值落入 **拒绝域。**
- **第 $II$ 类错误（取伪错误）**

  原假设为 **假** ，观察值落入 **接受域。**

> 当样本容量 $n$ 一定时, 若减少犯第 $I$ 类错误的概率, 则犯第 $II$ 类错误的概率往往 **增大** 。
>
> 若要使犯两类错误的概率都减小, 除非增加样本容量。

### 一般步骤

1. 提出假设 $H_0/H_1$
2. 选择 **检验统计量** ，确定概率分布
3. 给定显著性水平 $\alpha$ 确定 **拒绝域** $W$
4. 根据样本观察值计算统计量的 **观察值**
5. 根据统计量值是否落入拒绝域内，做出拒绝或接受 $H_0$ 的 **推断**

### 单个总体参数的检验

- **$\sigma^2$ 已知，关于 $\mu=\mu_0$ 的检验（$U$ 检验法）**

  取 $U=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\sim N(0,1)$
- **$\sigma^2=\sigma_0^2$ 未知，关于 $\mu=\mu_0$ 的检验（$T$ 检验法）**

  取 $T=\dfrac{\bar{X}-\mu_0}{S_n^*/\sqrt{n}}\sim t(n-1)$
- **$\mu$ 未知，关于 $\sigma^2=\sigma_0^2$ 的检验（ $\chi^2$ 检验法 ）**

  取 $\chi^2=\dfrac{(n-1)S_n^{*2}}{\sigma_0^2}\sim \chi^2(n-1)$

### 两个总体参数的检验

- **$\sigma_1^2,\sigma_2^2$ 已知，关于 $\mu_1=\mu_2$ 的检验（$U$ 检验法）**

  取 $U=\dfrac{(\bar{X}-\bar{Y})}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}\sim N(0,1)$
- **$\sigma_1^2=\sigma_2^2=\sigma^2$ 但未知，关于 $\mu_1=\mu_2$ 的检验（$T$ 检验法）**

  取 $T=\dfrac{(\bar{X}-\bar{Y})}{S_w\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$
- **$\mu_1,\mu_2$ 未知，关于 $\sigma_1^2=\sigma_2^2$ 的检验（ $F$ 检验法 ）**

  取 $F=\dfrac{S_1^{*2}}{S_2^{*2}}\sim F(n_1-1,n_2-1)$

> 有一种题长这样:
>
> $\sigma_1^2,\sigma_2^2,\mu_1,\mu_2$ **都未知，关于 $\mu_1=\mu_2$ 的检验**
>
> 然而用 $T$ 检验法的前提是 $\sigma_1^2=\sigma_2^2$ ，因此我们需要先用 $F$ 检验法检验 $\sigma_1^2=\sigma_2^2$ ，然后再用 $T$ 检验法检验 $\mu_1=\mu_2$
>
> 二臂完了，，，，，

**完结撒花！！！辛苦了，祝看到这里的你考试成功！！！**