概率论与数理统计复习提纲

1 概率论基本概念与方法

- 样本空间;
  ◇ 随即试验所有可能结果的组成的集合
- 事件;
  ◇ 样本空间的子集
- 事件的发生;
  ◇ 这一子集中的样本点出现
- 事件的关系与运算;
  ◇ $\subset; =; \cup; \cap; -$
  ◇ 互不相容；对立
- 互不相容与相互独立, 互不相关与相互独立;
- 概率的性质;
  ◇ 定义中的三条基本性质$\Longrightarrow$概率的六条性质
- 古典概率计算公式;
  ◇ $P(A) = \frac{n_A}{n}$
- 条件概率$P(B|A)$的意义及计算;
  ◇ 事件A发生的条件下，事件B发生的概率
- 全概率公式与Bayes公式.
  ◇ 前者是将一复杂事件的概率转换成不同原因/情况下一系列简单事件的概率求和
  ◇ 后者是一事件已出现，考察引发该事件发生的各种原因的可能性大小
  ◇ **注意:两者的考察可能会在同一题目中出现**

2 一维随机变量及其分布

- 离散型随机变量与连续型随机变量的定义;
  ◇ 前者基于随机变量取值的数目
  ◇ 后者是：r.v 的分布函数定义为非负可积函数的变上限积分
- 分布函数的定义及性质;
  ◇ 定义：$F(x) = P(X \leq x)$
  ◇ 性质：单调非减；$F(-\infty) = 0, F(\infty) = 1$;右连续性

- 离散型随机变量分布率的性质;
  $\diamond$ $p_i > 0; \sum p_i = 1$
- 连续型随机变量密度函数的性质
  $\diamond$ $f(x) > 0; \int_{-\infty}^{\infty} f(x) = 1$
- 分布函数与密度函数的关系;
  $\diamond$ 离散型：$F(x)$是分段函数，p_i是跳跃点处的跳跃高度
  $\diamond$ 连续型：$f(x) = F'(x)$
- 概率计算:
  - $P(X \le a) = F(a)$, $P\{a < X \le b\} = F(b) - F(a)$, 特别$P\{X = c\} = F(c) - F(c^-)$;(适用于离散及连续r.v.)
  - $P\{X \in A\} = \sum\limits_{x_i \in A} p_i$;
  - $P\{X \in A\} = \int_A f(x)dx$.
- 6个常用分布的分布及其数字特征(了解它们的性质及相互关系);
- 求离散型随机变量的分布:
  1. X的取值范围;
  2. X取每个值的概率.
- 随机变量函数$Y = g(X)$的分布.
  $\diamond$ 离散型
  $\diamond$ 连续型：$F_Y(y) = \int_{\{x|g(x)\le y\}} f_X(x)dx$，或者由$P_{56}$定理1：$f_Y(y) = f_X(h(y))|h'(y)|$，其中$x = h(y)$是$y = g(x)$的反函数。

例1 设随机变量X的概率密度函数为:

$$f(x) = \begin{cases} Cx\mathrm{e}^{-x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

1. (5分) 求常数C;
2. (5分) 写出X的分布函数$F(x)$;
3. (5分) 计算$P\{1 \le X \le 2\}$;
4. (5分) 设Y与X独立同分布, 求$X + Y$的概率密度.

1答:

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)dx &= C\int_0^{+\infty} x\mathrm{e}^{-x}dx \\ &= -C\int_0^{+\infty} xd\mathrm{e}^{-x} \\ &= -C\, x\mathrm{e}^{-x}\Big|_0^{+\infty} + C\int_0^{+\infty} \mathrm{e}^{-x}dx \\ &= C \end{aligned}$$

得$C = 1$.

2答: 对于$x \geq 0$,

$$
\begin{aligned}
F(x) &= \int_{-\infty}^{x} f(u)du \\
&= \int_{0}^{x} u\mathrm{e}^{-u} du \\
&= -\int_{0}^{x} u d\mathrm{e}^{-u} \\
&= -u\mathrm{e}^{-u}\big|_0^x + \int_0^x \mathrm{e}^{-u} du \\
&= 1-(x+1)\mathrm{e}^{-x}
\end{aligned}
$$

所以X的分布函数为:

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

3答:

$$
P\{1 \leq X \leq 2\} = F(2) - F(1) = 2\mathrm{e}^{-1} - 3\mathrm{e}^{-2} = 0.3298
$$

4答: 当$z < 0$时, $f_{X+Y}(z) = 0$; 对于$z \geq 0$,

$$
\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{+\infty} f(x)f(z-x)dx \\
&= \int_0^z x\mathrm{e}^{-x}(z-x)\mathrm{e}^{-(z-x)}dx \\
&= \mathrm{e}^{-z}\int_0^z x(z-x)dx \\
&= \frac{z^3}{6}\mathrm{e}^{-z}
\end{aligned}
$$

例2:设随机变量X的分布函数为:

$$
F(x) = A + B\arctan(x),\ -\infty < x < +\infty
$$

试计算

1. (5分) 常数A和B;
2. (5分) X的概率密度函数$f(x)$;
3. (5分) $P\{|X| \leq 1\}$;
4. (5分) $Y = X^2$的概率密度$f_Y(y)$.

1答: 根据分布函数的性质,

$$
F(-\infty) = A - \frac{\pi}{2}B, \quad F(+\infty) = A + \frac{\pi}{2}B = 1
$$

得$A = \frac{1}{2}$, $B = \frac{1}{\pi}$.

2答: X的概率密度函数为:

$$
f(x) = F'(x) = \frac{1}{\pi}\frac{1}{1+x^2},\ -\infty < x < +\infty
$$

3答:

$$
P\{|X| \leq 1\} = F(1) - F(-1) = \frac{2}{\pi}\arctan(1) = \frac{1}{2}
$$

4答: 根据分布函数的定义, Y的分布函数为:

$$F_Y(y) = P\{Y \le y\} = P\{X^2 \le y\}$$

(i) 如果$y \le 0$, 则$F_Y(y) = P\{X^2 \le y\} = 0$.

(ii) 如果$y > 0$, 则

$$\begin{aligned} F_Y(y) &= P\{X^2 \le y\} = P\{-\sqrt{y} \le X \le \sqrt{y}\} \\ &= \frac{2}{\pi}\arctan\sqrt{y} \end{aligned}$$

此时

$$f_Y(y) = F_Y'(y) = \frac{2}{\pi}\frac{1}{1+y}\cdot\frac{1}{2\sqrt{y}}$$

所以

$$f_Y(y) = \begin{cases} \dfrac{1}{\pi\sqrt{y}(1+y)}, & y > 0 \\ 0, & y \le 0 \end{cases}$$

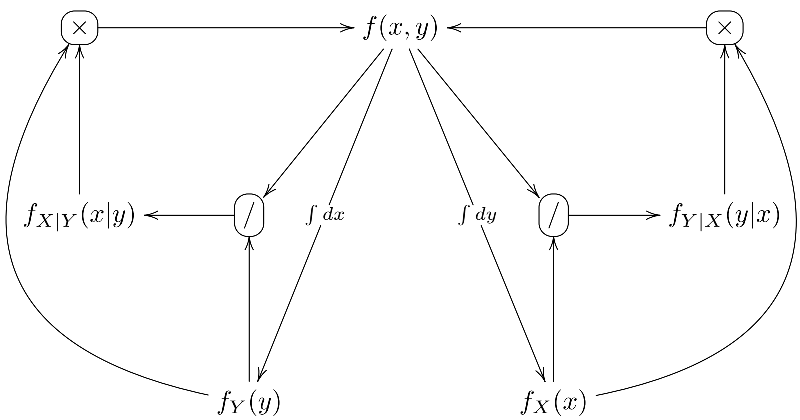
3 二维随机变量及其分布

- 二维联合分布函数的定义及性质;
  ◇ 定义：$F(x,y) = P(X \le x, Y \le y)$
  ◇ 性质：关于x 和y 均为单调非减；$F(-\infty,y) = F(x,-\infty) = F(-\infty,-\infty) = 0, F(\infty,\infty) = 1$; 关于$x$ 和y 均为右连续; $F(x_2,y_2) - F(x_2,y_1) - F(x_1,y_2) + F(x_1,y_1) \ge 0)$

- 二维分布函数与密度函数的关系;
  ◇ 离散型：$F(x,y) = \sum_{x_i \le x, y_i \le y} p_{ij}$
  ◇ 连续型：$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv$

- 二维随机变量的概率计算:
  - $P\{(X,Y) \in A\} = \sum\limits_{(x_i,y_j)\in A} p_{ij}$;
  - $P\{(X,Y) \in A\} = \iint_A f(x,y)dxdy$.

- 联合分布, 边缘分布与条件分布的关系:



- 条件分布函数的定义$F_{Y|X}(y|x) = P\{Y \le y | X = x\}$;

- 条件密度计算(比如$f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}$, 用分布函数表示时就是$f_{Y|X}(y|x)=\frac{\partial^2 F(x,y)}{\partial x\partial y}/F'_X(x)$);

- 二维正态分布$N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$的5个参数的意义及性质
  ◇ 边缘分布是正态
  ◇ 线性变换仍是正态(如，$Z=aX+bY\sim N(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2+2ab\sigma_1\sigma_2\rho)$)
  ◇ 独立与不相关等价

- 如何验证两个随机变量的独立性;
  ◇ $F(x,y)=F_X(x)\cdot F_Y(y)$
  ◇ 离散型：$P(X=x_i,Y=y_i)=P(X=x_i)\cdot P(Y=y_i)$
  ◇ 连续型：$f(x,y)=f_X(x)\cdot f_Y(y)$, $f_{X|Y}(x|y)=f_X(x)$ 或者$f_{Y|X}(y|x)=f_Y(y)$)

- **连续型随机变量** X,Y的函数：$X+Y$, $\max(X,Y)$, $\min(X,Y)$分布的计算.
  ◇ $X+Y$的分布计算中，当X,Y 相互独立时用卷积公式
  ◇ 两个相互独立的且服从泊松分布的r.v.,其和仍然服从泊松分布，且参数为原参数之和.
  ◇ 两个相互独立的且服从正态分布的r.v.,其和仍然服从正态分布，且参数为原参数之和.

- **离散型随机变量** X,Y的函数：$\max(X,Y)$, $\min(X,Y)$分布的计算.
  例3：设X, Y 都服从$b(1,p)$，且独立，问：$\max(X,Y)$（或者$\min(X,Y)$）的分布?

  例4：已知二维连续型随机变量(X,Y)的联合概率密度为:

$$f(x,y)=\begin{cases} x+y & 0\le x,y\le 1\\ 0 & \text{其它}\end{cases}$$

  求:

  1. X的边缘概率密度函数$f_X(x)$;
  2. $X=x$时, Y的条件密度函数$f_{Y|X}(y|x)$;
  3. X与Y的协方差.

1答:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f(x,y)dy \\ &= \int_0^1 (x+y)dy \\ &= x+\frac{1}{2} \quad 0\le x\le 1 \end{aligned}$$

2答:

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x,y)}{f_X(x)} \\ &= \frac{x+y}{x+\frac{1}{2}} \\ &= \frac{2(x+y)}{2x+1} \quad 0\le y\le 1 \end{aligned}$$

3答:

$$\mathrm{E}(X)=\int_0^1 x\left(x+\frac{1}{2}\right)dx=\frac{7}{12}$$

由分布的对称性, $\mathrm{E}(Y)=\frac{7}{12}$.

$$\begin{aligned}\mathrm{E}(XY) &= \int_0^1\int_0^1 xy(x+y)dxdy \\ &= \int_0^1 y\left(\frac{1}{3}+\frac{y}{2}\right)dy \\ &= \frac{1}{3}\end{aligned}$$

于是

$$\mathrm{Cov}(X,Y)=\mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)=\frac{1}{3}-\left(\frac{7}{12}\right)^2=-\frac{1}{144}$$

例5：已知二维离散型随机变量(X,Y)的联合分布率为:

$$P\{X=i,Y=j\}=\frac{i+j}{36},\quad i,j=1,2,3$$

求:

1. (5分) X的边缘分布律;
2. (5分) $X=1$时, Y的条件分布律.

1答:

$$p_{i\cdot}=\frac{1}{36}\sum_{j=1}^{3}(i+j)=\frac{1}{6}+\frac{i}{12},\quad i=1,2,3$$

或

X	1	2	3
$p_{i\cdot}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{5}{12}$

2答:

$$P\{Y=j|X=1\}=\frac{\frac{1+j}{36}}{\frac{1}{4}}=\frac{1+j}{9},\quad j=1,2,3$$

或

$Y\|X=1$	1	2	3
$p_{ij}/p_{1\cdot}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{4}{9}$

4 随机变量的数字特征

- $\mathrm{E}(g(X))$, $\mathrm{E}(g(X,Y))$的计算;

r.v.类型	$\mathrm{E}(g(X))$	$\mathrm{E}(g(X,Y))$
离散型	$\sum_i g(x_i)p_i$	$\sum_i\sum_j g(x_i,y_j)p_{ij}$
连续型	$\int g(x)f(x)dx$	$\int\int g(x,y)f(x,y)dxdy$

- 期望, 方差, 协方差以及相关系数的计算;
  ◇ 期望的计算由上述公式
  ◇ $D(X) = E[(X - EX)^2] = E(X^2) - (EX)^2$
  ◇ $cov(X,Y) = E[(X - EX)(Y - EY)]$
  ◇ $\rho_{XY} = \frac{cov(X,Y)}{\sqrt{D(X)D(Y)}}$

- 期望, 方差, 协方差以及相关系数的性质;
  ◇ 期望性质P_{97}(4条)
  ◇ 方差性质P_{103}(3条)
  ◇ 协方差性质P_{108}(6条)
  ◇ 相关系数性质P_{110}(3条)

- 正态线性函数分布的计算归结为期望, 方差的计算.

5 大数定律与中心极限定理

- Chebyshev不等式;
  ◇ $P(|X - EX| \geq \varepsilon) \leq \frac{DX}{\varepsilon^2}$ 或者$P(|X - EX| < \varepsilon) \geq 1 - \frac{DX}{\varepsilon^2}$

- 大数定律(条件和结论);
  ◇ 样本$X_1, \ldots, X_n$相互独立，且$EX_i = \mu$, $DX_i = \sigma^2$, 对任给的$\varepsilon > 0$, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1$
  ◇ Bernoulli大数定律：样本$X_1, \ldots, X_n$相互独立，服从两点分布，且$EX_i = p$, 对任给的$\varepsilon > 0$, $\lim_{n\to\infty} P(|\bar{X}_n - p| < \varepsilon) = 1$

- 中心极限定理
  ◇ 近似计算: $\{X_i\}_1^n$独立同分布, 当n很大时, $\sum\limits_{i=1}^{n} X_i \overset{\text{近似地}}{\sim} N(n\mathrm{E}(X_1), n\mathrm{D}(X_1))$，
  即$\frac{\sum\limits_{i=1}^{n} X_i - n\mathrm{E}(X_1)}{\sqrt{n\mathrm{D}(X_1)}} \overset{\text{近似地}}{\sim} N(0,1)$
  ◇ Demoivre-Laplace 定理：$\{X_i\}_1^n \overset{i.i.d}{\sim} b(1,p)$, 则$\frac{\sum\limits_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \overset{\text{近似地}}{\sim} N(0,1)$
  例6: 设随机变量序列$X_1, X_2, \cdots$相互独立, 且期望均为1, 方差均为2. 利用Chebyshev 不等式估计$P\left\{80 \leq \sum\limits_{i=1}^{100} X_i \leq 120\right\} \geq \underline{\frac{1}{2}}$; 为使$P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n} X_i - 1\right| \leq 0.1\right\} \geq 0.9$, 利用中心极限定理, 估计$n$至少需要达到$\underline{542}$.

6 统计量与抽样分布

- 统计量的概念
  ◇ 样本的函数，且不含有任何参数

- χ^2-分布, t-分布, F-分布的定义
  ◇ $\{X_i\}_1^n \overset{i.i.d}{\sim} N(0,1)$, $X = \sum\limits_{i=1}^{n} X_i^2 \sim \chi^2(n)$
  ◇ $X \sim N(0,1), Y \sim \chi^2(n)$,且$X, Y$独立，则$T = \frac{X}{\sqrt{Y/n}} \sim t(n)$
  ◇ $X \sim \chi^2(m), Y \sim \chi^2(n)$,且$X, Y$独立，则$T = \frac{X/m}{Y/n} \sim F(m,n)$

- χ^2-分布, t-分布, F-分布的性质, α分位点查表;
  ⋄ χ^2-分布($X \sim \chi^2(n)$): 图像；$E(X)=n$,$D(X)=2n$；可加性；分位点($P(X \geq \chi^2_\alpha(n))=\alpha$)
  ⋄ t-分布($X \sim t(n)$): 图像；分位点($P(X \geq t_\alpha(n))=\alpha$)；$t_{1-\alpha}(n)=-t_\alpha(n)$
  ⋄ F-分布($X \sim F(m,n)$): 图像；$\frac{1}{X}=F(n,m)$；可加性；分位点($P(X \geq F_\alpha(m,n))=\alpha$)；$F_{1-\alpha}(m,n)=\frac{1}{F_\alpha(n,m)}$

- 正态总体八种抽样分布:

$\dfrac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$	σ已知时, 对μ的推断
$\dfrac{\bar{X}-\mu}{S}\sqrt{n} \sim t(n-1)$	σ未知时, 对μ的推断
$\sum \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$	μ 已知时, 对σ的推断
$\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$	μ未知时, 对σ的推断
$\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1}{m}+\frac{\sigma_2}{n}}} \sim N(0,1)$	σ_1,σ_2 已知时, 对$\mu_1-\mu_2$的推断
$\dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2)$, 其中$S_w=\sqrt{\frac{(m-1)S_X^2+(n-1)S_Y^2}{m+n-2}}$	$\sigma_1=\sigma_2$且未知时, 对$\mu_1-\mu_2$的推断
$\dfrac{\frac{1}{m}\sum(X_i-\mu_1)^2/\sigma_1^2}{\frac{1}{n}\sum(Y_i-\mu_2)^2/\sigma_2^2} \sim F(m,n)$	μ_1和μ_2已知时, 对$\dfrac{\sigma_1}{\sigma_2}$的推断
$\dfrac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F(m-1,n-1)$	μ_1和μ_2未知时, 对$\dfrac{\sigma_1}{\sigma_2}$的推断

例7：已知X,Y,Z独立同分布于$N(0,\sigma^2)$, 那么$\frac{X+Y}{\sqrt{2}|Z|} \sim \underline{t(1)}$; 为使$AX^2$服从$\chi^2$分布, 则$A=\underline{\frac{1}{\sigma^2}}$

7 参数点估计

- 无偏估计; 无偏估计的有效性; 相合估计的概念;
  ⋄ $E(\hat{\theta})=\theta$
  ⋄ $E(\hat{\theta}_1)=\theta$, $E(\hat{\theta}_2)=\theta$, 若$D(\hat{\theta}_1)<D(\hat{\theta}_2)$
  ⋄ $\hat{\theta} \xrightarrow{P} \theta$，即任意$\varepsilon>0$, $\lim_{n\to\infty} P(|\hat{\theta}_n-\theta|<\varepsilon)=1$ 或者$\lim_{n\to\infty} P(|\hat{\theta}_n-\theta|\geq\varepsilon)=0$

- 矩估计方法与最大似然估计法.

例8：设简单随机样本$X_1,X_2,\cdots,X_n$为取自密度函数为

$$f(x;\theta)=\begin{cases} \dfrac{2x}{\theta^2} & x\in[0,\theta] \\ 0 & \text{其它} \end{cases}$$

的总体. 求

1. θ的矩估计量$\hat{\theta}_M$;
2. θ的最大似然估计量$\hat{\theta}_L$.

1答:

$$
\begin{aligned}
\mathrm{E}(X) &= \int_0^\theta \frac{2x^2}{\theta^2} = \frac{2}{3}\theta \\
\theta &= \frac{3}{2}\mathrm{E}(X) \\
\hat{\theta}_M &= \frac{3}{2}\overline{X}
\end{aligned}
$$

2答: 似然函数为:

$$
\begin{aligned}
L(\theta) &= \begin{cases} \dfrac{2^n}{\theta^{2n}}\prod\limits_{i=1}^n x_i & x_1,\cdots,x_n \le \theta \\ 0 & \text{其它} \end{cases} \\
&= \begin{cases} \dfrac{2^n}{\theta^{2n}}\prod\limits_{i=1}^n x_i & \theta \ge x_{(n)} \\ 0 & \theta < x_{(n)} \end{cases}
\end{aligned}
$$

其中$x_{(n)} = \max(x_1,\cdots,x_n)$. 当$\theta \ge x_{(n)}$时, $L(\theta)$是θ的单调递减函数, 而当$\theta < x_{(n)}$时, $L(\theta) = 0$, 于是$L(\theta)$在$x_{(n)}$处达到最大值, 即得$\hat{\theta}_L = x_{(n)}$, θ的最大似然估计量为$\hat{\theta}_L = X_{(n)}$

例9: 设总体$\mathbb{X}$的期望为μ, $X_1,\cdots,X_n$是取自该总体的一组样本. 如果$\sum\limits_{i=1}^n a_iX_i$是$\mu$的无偏估计, 则常数$a_1,a_2,\cdots,a_n$必须满足条件:$\underline{\ a_1+a_2+\cdots+a_n=1\ }$; 如果$\hat{\mu}_n = \hat{\mu}(X_1,\cdots,X_n)$是$\mu$的相合估计, 则对于$\forall\varepsilon > 0$, 有$\lim\limits_{n\to\infty} P\{|\hat{\mu}_n - \mu| < \varepsilon\} = \underline{1}$.

8 正态总体参数的区间估计与假设检验

- μ, σ^2的置信度为$1-\alpha$区间估计:
  1. 从上表中选择合适的$T(X_1,\cdots,X_n;\theta)$;
  2. 根据$T(X_1,\cdots,X_n;\theta)$的分布, 确定分位点$c_{1-\frac{\alpha}{2}}$, $c_{\frac{\alpha}{2}}$, 使得$P\{c_{1-\frac{\alpha}{2}} \le T(X_1,\cdots,X_n;\theta) \le c_{\frac{\alpha}{2}}\} = 1-\alpha$. (注: 当$T(X_1,\cdots,X_n;\theta)$的密度关于$Y$-轴对称时, $c_{1-\frac{\alpha}{2}} = -c_{\frac{\alpha}{2}}$);
  3. 解不等式: $c_{1-\frac{\alpha}{2}} \le T(X_1,\cdots,X_n;\theta) \le c_{\frac{\alpha}{2}}$, 即得$\theta$的置信度为$1-\alpha$区间估计.
- 显著性水平为α的μ, σ^2的假设检验:
  1. 构造H_0,H_1,(注意:“=”一定含在H_0中)
  2. 从上表中选择合适统计量, 其中的θ取H_0中等号后的数值: $T(X_1,\cdots,X_n;\theta_0) \sim$ 完全已知的分布;
  3. 根据H_0和H_1的形式,判断$T(X_1,\cdots,X_n;\theta_0)$何时对$H_0$有利,何时不利, 比如越大对$H_1$越有利, 则拒绝域的形式为$W = \{T|T(X_1,\cdots,X_n;\theta) > c\}$;
  4. 根据$T(X_1,\cdots,X_n;\theta_0)$的分布, 确定$c$使得$P\{T \in W|\theta = \theta_0\} = \alpha$.
  5. 将样本观测值带入T 得到T_0,若$T_0 \in W$,则$RejH_0$,否则,$AccH_0$.
- 显著性水平α, 两类错误的概念

  ◇ $P(T \in W|H_0) = \alpha$—第一类错误

  ◇ $P(T \in \bar{W}|H_1) = \beta$—第二类错误

例10. 设某种职业的年收入$\mathbb{X} \sim N(\mu, \sigma^2)$, 现从该种职业人群中随机抽取一组容量为36的样本, 算得样本均值$\bar{x} = 6.5$(万元), 样本标准差$s = 1.0$(万元).

1. 求总体标准差σ的置信度为0.95的区间估计;
2. 在显著性水平$\alpha = 0.05$下, 能否认为该种职业的平均年收入超过了6万元.

1答: 总体标准差σ的置信度为0.95的区间估计为:

$$\begin{aligned}
& \left[s\sqrt{\frac{n-1}{\chi^2_{\alpha/2}(n-1)}}, s\sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}(n-1)}}\right] \\
= & \left[\sqrt{\frac{35}{53.2033}}, \sqrt{\frac{35}{20.5694}}\right] \\
= & [0.8111, 1.3044]
\end{aligned}$$

2答: 构造假设:

$$H_0: \mu \le 6 \quad H_1: \mu > 6$$

的拒绝域为:

$$\begin{aligned}
W &= \left\{\boldsymbol{x} | \bar{x} > 6 + \frac{s}{\sqrt{n}} t_\alpha(n-1)\right\} \\
&= \left\{\boldsymbol{x} | \bar{x} > 6 + \frac{1}{\sqrt{36}} 1.6896\right\} \\
&= \{\boldsymbol{x} | \bar{x} > 6.2816\}
\end{aligned}$$

由于$\bar{x} = 6.5 \in W$, 因此拒绝H_0, 即可以认为该种职业的平均年收入超过了6万元.

9 分布拟合检验

- 一般步骤：
- H_0:$X \sim F(x)$
  1. 将总体X的取值范围分成k个互不相交的区间$A_1, \ldots, A_k$;
  2. 计算样本观察值落入每个小区间$(A_i, A_{i+1}]$的频数$f_i, i = 1, \ldots, k-1$.
  3. 计算H_0为真时，总体X落入区间$(A_i, A_{i+1}]$ 的概率$p_i = P(A_i < X \le A_{i+1}), i = 1, ldots, k-1$.
     (注意，若$F(x)$ 中含有r 个未知数，则需要先用MLE 估计未知数，然后再计算p_i.)
  4. 构造检验统计量$K = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i}$, 由Pearson定理知道，$K \sim \chi^2(k-1)$.
     (注意，若$F(x)$ 中含有r 个未知数，则需要先用MLE 估计这r个未知数，此时，$K \sim \chi^2(k-1-r)$.)
  5. 拒绝域$W = K > \chi^2_\alpha(k-1-r)$. 若$K$ 的观察值$K_0 \in W$, 则$RejH_0$, 否则，$AccH_0$.
- 掌握具体检验方法(例子).
  例11 掷一颗骰子60次，结果如下：

点数	1	2	3	4	5	6
次数	7	8	12	11	9	13

试在显著性水平为下检验这颗骰子是否均匀。

解：H_0：$F(x)$是离散分布 $P(X=k)=\frac{1}{6}$，$k=1,\cdots,6$。

组频率：

k	1	2	3	4	5	6
f_k	0.117	0.133	0.2	0.184	0.15	0.216

$\chi^2=\sum_{i=1}^{k}\frac{n}{p_i}(n_i/n-p_i)^2=2.780$，

拒绝域 $W=\{\chi^2=\sum_{i=1}^{k}\frac{n}{p_i}(n_i/n-p_i)^2\geq\chi_\alpha^2(k-1)\}=\{\chi^2\geq\chi_{0.05}^2(5)=11.07\}$

所以接受。