第四讲

统计基础

- 总体和样本
- 统计量
- 抽样分布
- 点估计
- 区间估计

总体和样本

- 数理统计的核心问题——由样本推断总体

数据资料的收集

↓

数据的整理、分析

↓

统计推断

总体和样本

- 总体的分布一般来说是未知的，统计学的主要任务正是要对总体的未知分布进行推断.

总体(理论分布)？ ← 样本值

总体(理论分布) → 样本 → 样本值

总体X的概率分布为$p(x) = P\{X = x\}$,

则样本的概率分布为

$$P(x_1,\ x_2, \cdots, x_n) = P\{X_1 = x_1, \cdots, X_n = x_n\} = \prod_{i=1}^{n} p(x_i).$$

第四讲

统计基础

- 总体和样本
- 统计量
- 抽样分布
- 点估计
- 区间估计

统计量

小象学院 ChinaHadoop.cn

设$X_1, X_2, \cdots, X_n$为来自总体X的一个样本，称此样本的任一不含总体分布未知参数的函数$g(X_1, X_2, \cdots, X_n)$为该样本的一个**统计量**.

1. **样本均值** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i;$

2. **未修正的样本方差**

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right).$$

3. **修正的样本方差(样本方差)**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right).$$

第四讲

统计基础

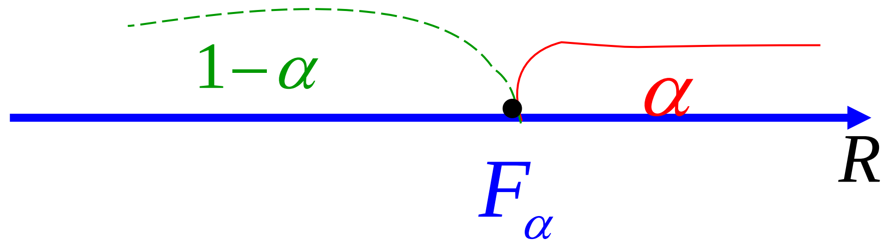
- 总体和样本
- 统计量
- 抽样分布
- 点估计
- 区间估计

抽样分布

小象学院 ChinaHadoop.cn

- **分位数**

$$P\{X > F_\alpha\} = \alpha,$$



即 $\quad 1 - F(F_\alpha) = \alpha \quad$ 或 $\quad F(F_\alpha) = 1 - \alpha$

则称 F_α 为随机变量 X 的 α 水平的上侧分位数.

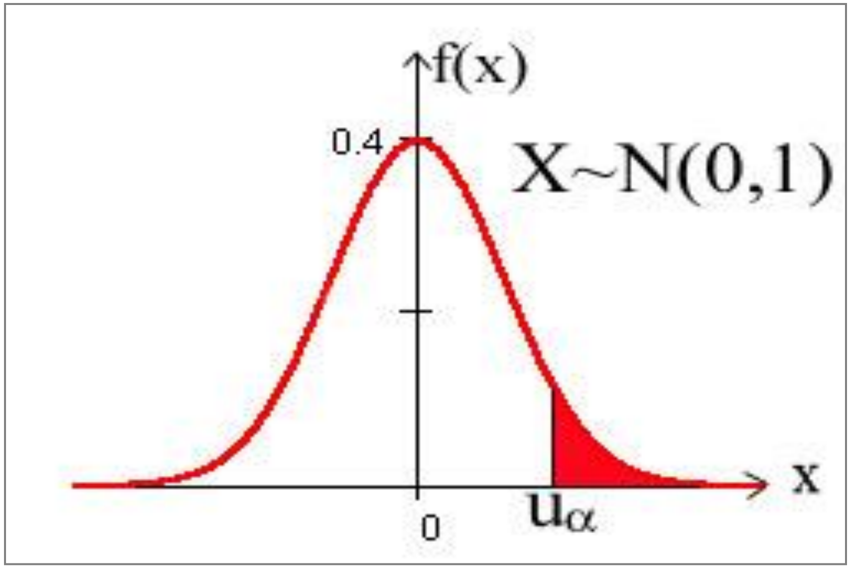
简称 α 上侧分位数.

抽样分布

➢ **分位数**

例如：$X \sim N(0,1)$, 记水平α的上侧分位数为u_α，

则 $1-\Phi_0(u_\alpha)=\alpha$



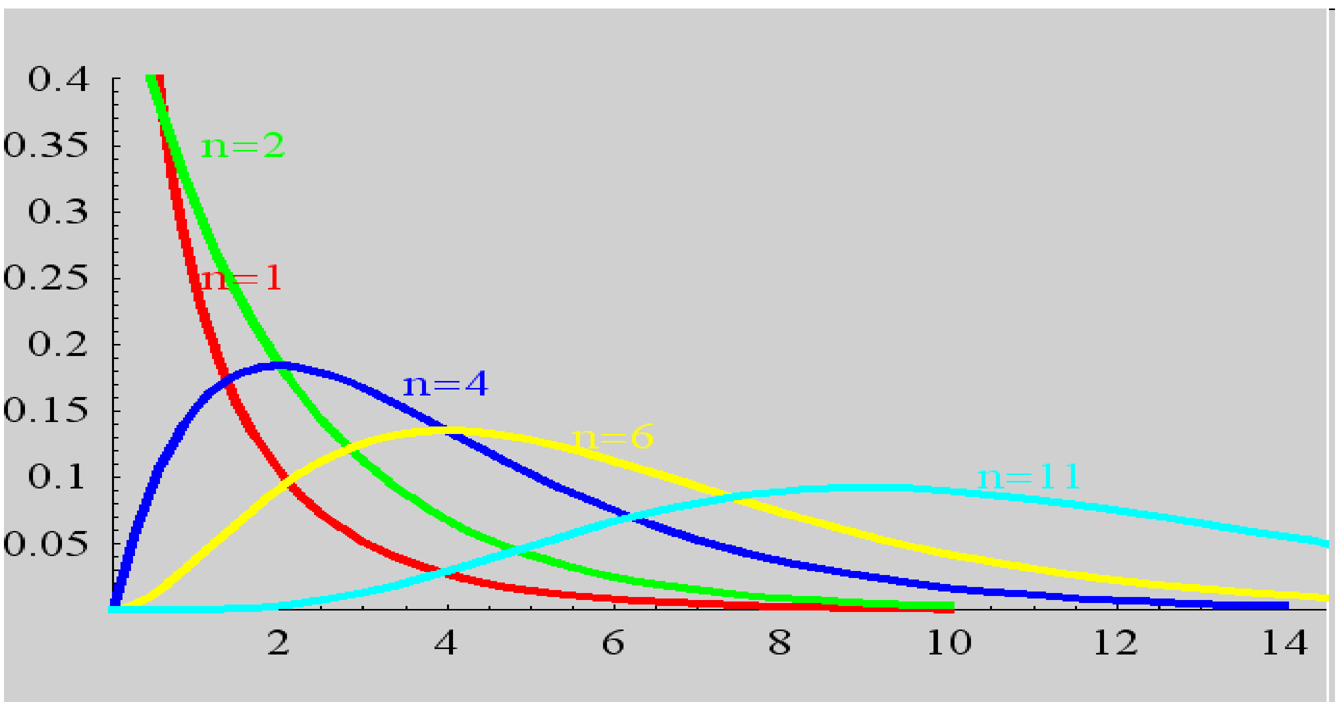
$u_{0.05}=1.64$

$u_{0.025}=1.96,$

1.χ^2分布的定义

设 $X_1, X_2, \cdots, X_n$ 是来自总体$X \sim N(0, 1)$的样本，

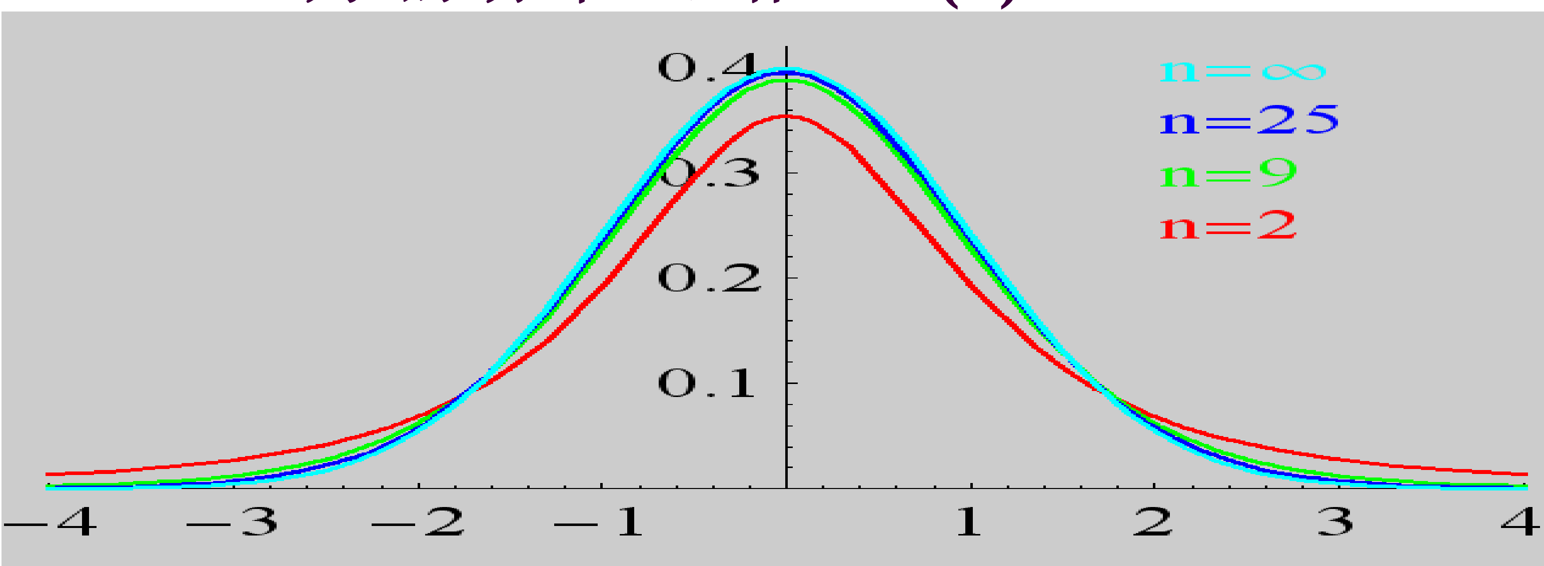
则$X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi^2(n)$.



抽样分布

➢ **t分布** 设随机变量X服从标准正态分布$N(0,1)$，Y服从$\chi^2(n)$，且X与Y相互独立，

记 $T=\dfrac{X}{\sqrt{Y/n}}$，则随机变量T服从自由度为n的t分布，记作$T\sim t(n)$.



抽样分布

小象学院 ChinaHadoop.cn

- ***t* 分布**

设总体 $X \sim N(\mu,\sigma^2)$，$X_1, X_1, \cdots, X_n$（$n \geq 2$）是来自X的一个样本，$\bar{X}$ 与 S^2 分别为样本均值与样本方差，则随机变量

$$\mathrm{t}=\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

服从自由度为$n-1$的t分布.

小象学院 ChinaHadoop.cn

- ***F*分布**

设 $X \sim \chi^2(m), Y \sim \chi^2(n)$, 且 X 与 Y 相互独立,

记
$$Z = \frac{X/m}{Y/n} = \frac{nX}{mY}$$

则 Z 的密度函数为 $f(x;m,n)$, 因此 $Z \sim F(m,n)$.

由定理5.3.4不难看出，若 $X \sim F(m,n)$，则 $X^{-1} \sim F(n,m)$.

抽样分布

- ***F*分布**

$X_1, X_1, \cdots, X_{n_1}$和$\mathrm{Y}_1, \mathrm{Y}_2, \cdots, \mathrm{Y}_{n_2}$是分别来自两个相互独立的正态总体$N(\mu_1, \sigma_1{}^2)$和$N(\mu_2, \sigma_2{}^2)$的两个随机样本（$n_1$，$n_2 \geq 2$），则

$$F = \frac{S_1{}^2 / \sigma_1{}^2}{S_2{}^2 / \sigma_2{}^2} \sim F(n_1 - 1，n_2 - 1)$$

特别，当$\sigma_1{}^2 = \sigma_2{}^2$时，统计量“两个样本方差之比”服从F分布，即

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

抽样分布

➤ **关系图**

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

$$(2)\ U = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1);$$

$$(3)\quad \frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1);$$

$\bar{X}$ 与 S^2 相互独立.

$$(4)\ T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

第四讲

统计基础

- 总体和样本
- 统计量
- 抽样分布
- 点估计
- 区间估计

点估计

- 使用来自总体X的样本值构造一个统计量来估计总体分布的某参数的真实值。这个统计量称为某参数的估计量
- 好的估计量：无偏性，有效性，相合性

- 统计推断
  - 统计估计
    - 参数估计
      - 点估计
      - 区间估计
    - 非参数估计
  - 统计假设检验
    - 参数
      - 单参数
      - 多参数
    - 非参数

点估计

用样本均值$\bar{X}$来估计总体的期望EX.

因为$\bar{X}$是EX的既是无偏估计量，又是相合估计量，而且在EX的一切无偏估计量中$\bar{X}$的方差最小，即最有效。

设$(X_1, X_2, \cdots, X_n)$为来自总体X的样本，且其方差存在，则样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

是方差DX的无偏估计量。

点估计

例： 人的身高$X \sim N(\mu,\sigma^2)$,其中μ,σ^2未知。现抽样得样本（$X_1, X_2 \cdots X_{10}$），样本值为

$168, 170, 172, 183, 200, 175, 174, 180, 165, 178.$

试估计μ,σ^2的值。

参数的最大似然估计

小象学院 ChinaHadoop.cn

应寻找使试验结果(即样本值）出现的可能性最大的那个θ作为θ真值的估计值。

似然函数

$$L(\theta_1,\theta_2,\cdots\theta_m)=P\{X_1=x_1,X_2=x_2\cdots X_n=x_n\}$$

$$=\prod_{i=1}^{n}p(x_i;\theta_1,\theta_2,\cdots\theta_m)$$

即 $L(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_m)=\max\limits_{(\theta_1,\theta_2,\cdots,\theta_m)\in\Theta}L(\theta_1,\theta_2,\cdots,\theta_m)$

则称 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_m$分别为$\theta_1,\theta_2,\cdots,\theta_m$的

最大似然估计值 (MLE).

参数的最大似然估计

小象学院 ChinaHadoop.cn

（3）求最大似然估计(MLE)的一般步骤

1° 写出似然函数 $L(\theta_1, \theta_2, \cdots, \theta_m)$

似然方程组---- $\begin{cases} \dfrac{\partial L}{\partial \theta_1} = 0 \\ \dfrac{\partial L}{\partial \theta_2} = 0 \\ \quad \vdots \\ \dfrac{\partial L}{\partial \theta_m} = 0 \end{cases}$

2° 取对数 $\ln L(\theta_1, \theta_2, \cdots, \theta_m)$ ------对数似然函数

参数的最大似然估计

小象学院 ChinaHadoop.cn

解 $X \sim \varphi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$,

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\ln L(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2,$$

令 $\begin{cases} \frac{\partial}{\partial\mu}\ln L(\mu,\sigma^2) = 0 \\ \frac{\partial}{\partial\sigma^2}\ln L(\mu,\sigma^2) = 0 \end{cases}$ $\begin{cases} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \end{cases}$

解得 $\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \end{cases}$ $\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2. \end{cases}$

第四讲
统计基础

- 总体和样本
- 统计量
- 抽样分布
- 点估计
- 区间估计

区间估计

所谓参数θ的点估计，是指用一个估计量$\hat{\theta}(X_1, X_2, \cdots X_n)$的值去估计$\theta$的真值。

但估计效果的好坏并没有指出，即估计的精确性与可靠性未给出，这种估计是没有多大意义的，这时需要引入区间估计。

设θ是总体X的分布的未知参数$\theta \in \Theta$. $X_1, X_2, \cdots, X_n$ 为来自X的样本。对给定的α（$0<\alpha<1$),若存在两个统计量

$\underline{\theta} = \underline{\theta}(X_1, X_2, \cdots, X_n)$和$\overline{\theta} = \overline{\theta}(X_1, X_2, \cdots, X_n)$，使得

$$P\{\underline{\theta} < \theta < \overline{\theta}\} = 1 - \alpha$$

则随机区间$(\underline{\theta}, \overline{\theta})$称为参数$\theta$的$1-\alpha$置信区间;

$1-\alpha$称为置信水平（置信度）；

$\underline{\theta}$与$\overline{\theta}$称为θ的置信下限与置信上限.

正态分布的μ的区间估计

总体$X \sim N(\mu, \sigma_0^2)$，$\sigma_0^2$已知，$\mu$未知.

$X_1, X_2, \cdots, X_n$是来自X的一个样本，$x_1, x_2, \cdots, x_n$为样本值，$\alpha = 0.05$求μ的$1-\alpha$置信区间.

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma_0^2}{n})$$

寻找未知参数的一个良好估计.

令 $U = \dfrac{\bar{X} - \hat{\mu}}{\sigma_0 / \sqrt{n}} \sim N(0, 1)$

寻找一个待估参数和估计量的函数，其分布确定.

正态分布的μ的区间估计

$$P\left\{\left|\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}}\right|<\frac{\lambda_0}{\sigma_0/\sqrt{n}}\right\}=P\{|U|<\lambda\}=2\Phi_0(\lambda)-1$$

令 $1-\alpha=0.95$ 则 $\lambda=u_{\alpha/2}=1.96$

$$\Phi_0(\lambda)=1-\frac{\alpha}{2}=0.975=\Phi_0(1.96)$$

μ的0.95置信区间为 $(\bar{X}-1.96\frac{\sigma_0}{\sqrt{n}},\ \bar{X}+1.96\frac{\sigma_0}{\sqrt{n}})$

区间估计

1、σ^2已知，均值 μ 的置信区间

简记为$\left(\bar{X} \pm u_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$.

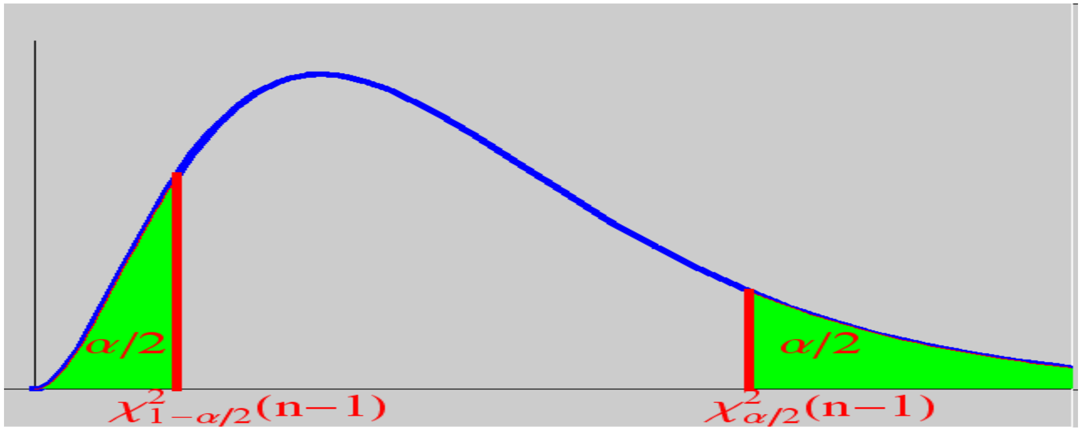
2. σ^2为未知，μ的置信区间

简记为$\left(\bar{X} \pm t_{\alpha/2}(n-1)\dfrac{S}{\sqrt{n}}\right)$.

区间估计

3 正态总体方差 σ^2 的置信区间

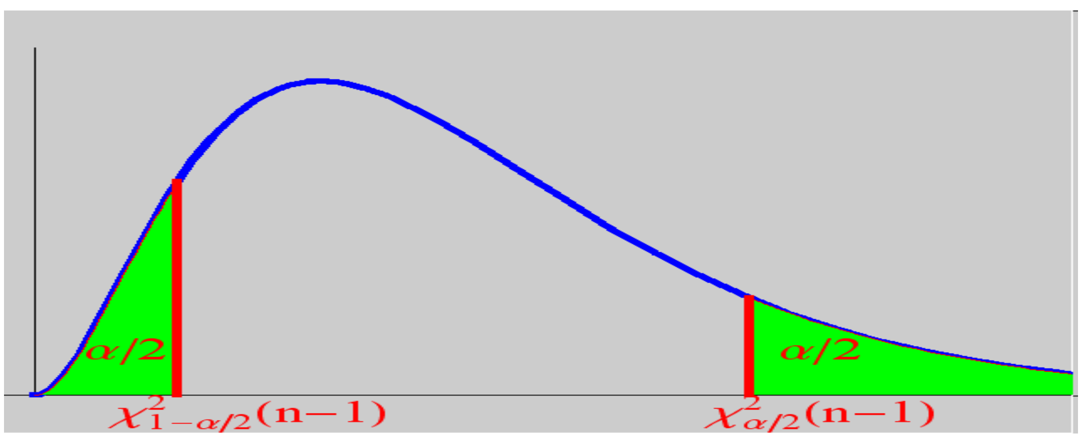
$$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)} \right)$$



4.μ 未知，方差的置信区间

方差σ^2的置信水平为$1-\alpha$的置信区间为

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right).$$



联系我们：

- 新浪微博：ChinaHadoop
- 微信公众号：ChinaHadoop
- 网站：http://chinahadoop.cn