第二章:概率统计回顾

梁龙跃

Email: lyliang@gzu.edu.cn

贵州大学 经济学院

2018 年 3 月

本章内容

1. 概率分布
2. 蒙特卡洛模拟
3. 统计推断

概率统计的比较

一种观点认为:概率论是数理统计的数学基础, 数理统计是概率论的重要应用. 另一种观点认为:概率和统计是两个独立又相关的数学领域.

我们用两个盒子来说明二者的联系和差异.

- 概率盒子: 盒子里装有 10 个白球, 10 个红球, 10 个黑球. 概率试图回答这样的问题:随机地从盒子中抽取一个球, 它是红色的可能性有多大?
- 统计盒子: 我们不知道盒子里装有哪些球. 从盒子中抽取一个样本, 我们可以根据样本信息猜测盒子里装着哪些球以及它们的比例.

概率关注的是在你知道所有可能的结果(即知道整个总体)时某件事的可能性. 而**统计关注的是**先抽样, 描述样本(描述统计), 再根据样本信息对总体做出推断(推断统计).

随机变量

概率的基石: $(\Omega, \mathcal{F}, \mathbb{P})$ 三元组, **概率空间**.

1. Ω 是样本空间, 就是全体可能结果(即全体基本事件)组成的集合, 它不必是有限集合.
2. $\mathcal{F}$ 称为**事件域**(也称为 σ-代数), $\mathcal{F}$ 满足 $\Omega \in \mathcal{F}$, $\mathcal{F}$ 对于补运算封闭, $\mathcal{F}$ 对于可数并运算封闭.
3. $\mathbb{P}$ 为**概率函数**. $\mathbb{P}$ 是在随机事件上"取概率"的运算, 也即 $\mathbb{P}$ 是 $\mathcal{F}$ 上定义的一个函数, 满足 非负性、规范性、可数可加性.

概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 完整地描述了样本空间 Ω 上的随机试验的统计性质, 因此一个概率空间就是描写一个随机试验的数学模型.

随机变量

为了统计方便, 将试验结果数量化, 定义随机变量. 如掷骰子的试验, 可以定义随机变量:

$$X = \begin{cases} 1 & \text{出现 } \omega_1 \\ 2 & \text{出现 } \omega_2 \\ 3 & \text{出现 } \omega_3 \\ 4 & \text{出现 } \omega_4 \\ 5 & \text{出现 } \omega_5 \\ 6 & \text{出现 } \omega_6 \end{cases}$$

随机变量是从样本空间到实数集的一个函数. 根据取值分为是离散型的和连续型的.

概率分布

随机变量 X 的取值及其相应的概率为概率分布.

1. 离散随机变量 X, 取值 $x_1, x_2, \cdots, x_n$, 则概率分布为:

$$f(x) = \begin{cases} \mathbb{P}(X = x_i) & i = 1, 2, \cdots, n \\ 0 & X \neq x_i, i = 1, 2, \cdots, n \end{cases}$$

累积概率分布函数定义为:

$$F(x) := \mathbb{P}(X \leq x) = \sum_{X \leq x} f(x).$$

2. 连续随机变量 X, 设其概率密度函数为 $f(x)$, 则累积概率分布函数为:

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

连续随机变量在某一点的概率为 0.

期望

1. 期望: 随机变量取值集中趋势的度量, 定义为:

$$\mathbb{E}(X) = \begin{cases} \sum xf(x) & \text{离散} \\ \int_X xf(x)dx & \text{连续} \end{cases}$$

随机变量 X 的函数 $g(X)$ 的期望:

$$\mathbb{E}[g(X)] = \begin{cases} \sum g(x)f(x) & \text{离散} \\ \int_X g(x)f(x)dx & \text{连续} \end{cases}$$

期望的线性性质:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b, \quad a, b \in \mathbb{R}.$$

方差

2. 方差: 随机变量取值离散程度的变量, 定义为:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X-\mathbb{E}(X)]^2 \\ &= \begin{cases}\sum[x-\mathbb{E}(X)]^2 f(x) & \text{离散} \\ \int_X [x-\mathbb{E}(X)]^2 f(x)dx & \text{连续}\end{cases} \\ &= \mathbb{E}(X^2)-[\mathbb{E}(X)]^2\end{aligned}$$

标准差: $\sigma_X = \sqrt{\mathrm{Var}(X)}$. 标准差的作用是统一量纲.
a 和 b 为常数时, 则

$$\mathrm{Var}(a)=0,\ \ \mathrm{Var}(a+bX)=b^2\mathrm{Var}(X).$$

协方差

3. 协方差: 随机变量 X 和 Y 的协方差定义为:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

4. 相关系数: 为了克服受到量纲的影响, 引入了相关系数. 随机变量 X 和 Y 的相关系数定义为:

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

- $-1 \le \rho \le 1$;
- 若 $\rho = 0$, 表示两个变量不是线性相关的;
- 若 $\rho < 0$, 表示两个变量是负相关的;
- 若 $\rho > 0$, 表示两个变量是正相关的.

分位数

对于给定的 $0 < \alpha < 1$, 随机变量 X 的 α 分位数 $Q(\alpha)$ 是满足如下条件的数值:

$$\mathbb{P}(X \leq Q(\alpha)) \leq \alpha, \quad \mathbb{P}(X \geq Q(\alpha)) \leq 1 - \alpha.$$

- 若已知随机变量的分布函数, 分位数可以通过求概率分布的逆运算给出 $F^{-1}(\alpha)$.
- 四分位数: 将已排序数据分为四个部分, 任何一组数据都可找到 3 个四分位数. 第一个四分位数 Q_1, 使得正好 25% 的数据比它小, 75% 的数据比它大. 第二个四分位数即中位数. 第三个四分位数, Q_3, 使得正好 75% 的数据比它小, 25% 的数据比它大.

其它数字特征

- r 阶中心矩:

$$\mu_r = \mathbb{E}[(X-\mu)^r],$$

其中$\mu = \mathbb{E}(X)$, 当 $\mu = 0$ 时为 r 阶原点矩, 记为 a_r.

- 偏度: 概率分布对称性的度量.

$$\nu = \frac{\mathbb{E}[X-\mu_X]^3}{\sigma_X^3},$$

$\nu > 0$, 右偏; $\nu < 0$, 左偏.

- 峰度: 概率分布密度函数高低(胖瘦)的度量.

$$\kappa = \frac{\mathbb{E}[X-\mu_X]^4}{\sigma_X^4},$$

$\kappa = 3$, 正态分布; $\kappa > (<)3$, 尖峰(低峰), 或者薄(厚)尾分布.

正态分布

定义 若随机变量 X 的概率密度函数为:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),\quad x\in\mathbb{R}.$$

则称 X 服从正态分布(也称 X 为正态随机变量), 记为 $X\sim N(\mu,\sigma^2)$. 若 $X\sim N(\mu,\sigma^2)$, 则变换

$$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$

$N(0,1)$ 为标准正态分布.

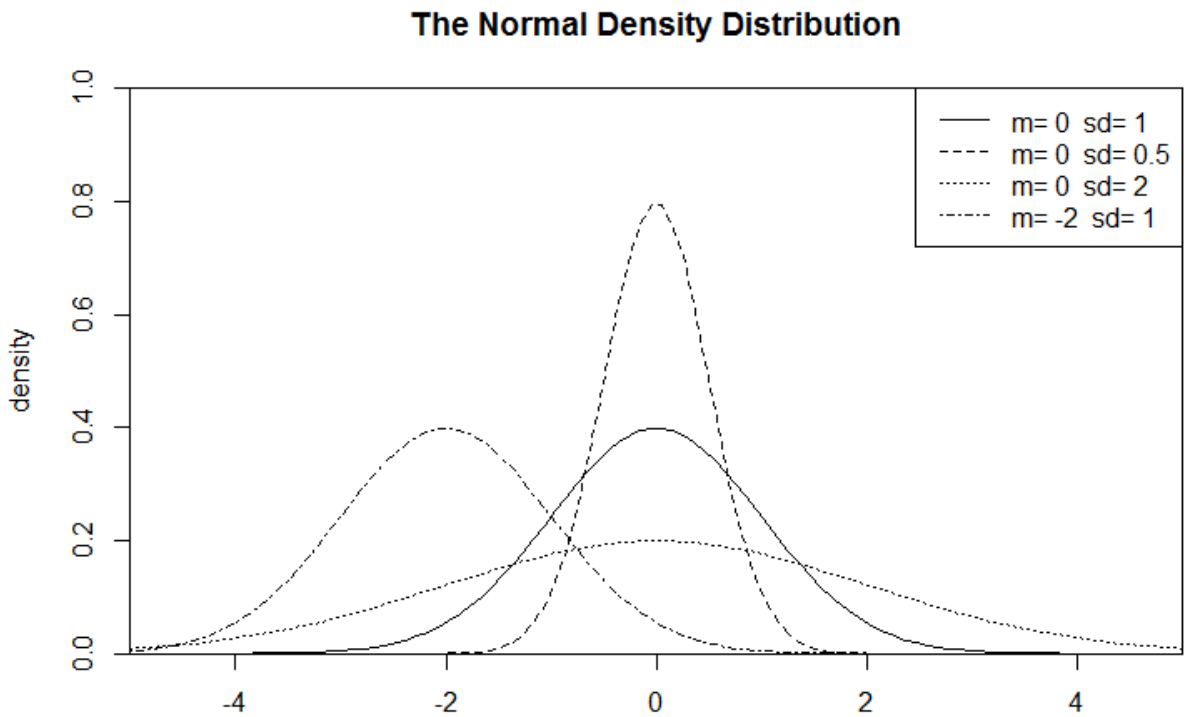
- 正态分布是单峰对称的, μ 称为位置参数, σ 称为形状参数.
- 3 倍标准差原理: 正态随机变量的值几乎全部落入区间 $(\mu-3\sigma,\mu+3\sigma)$, 概率为 99.7%.

正态分布

在 R 中产生正态随机数的语句为

```
x=rnorm(100,3.5,1.2)    #产生 100 个服从 N(3.5,1.2) 的随机数据
```

pnorm(x,mean,sd),dnorm,qnorm,rnorm 分别为正态分布函数, 分布密度, 分位数函数, 随机数函数. 期望为 mean, 标准差为 sd.



Q-Q 图

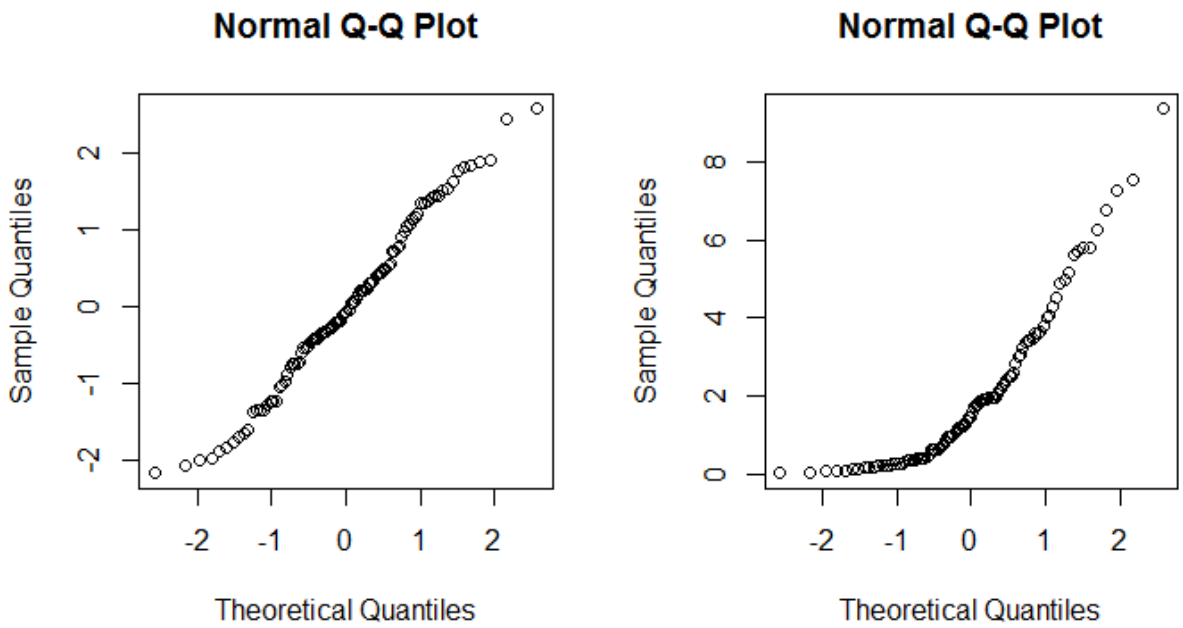
Q-Q 图, 这里的两个"Q" 都代表了 quantile, 即分位数. 比如 0.25 的分位数意味着有 25% 的数据小于它. 对于理论分布来说, 0.25 分位数意味着变量小于它的概率大约(对于正态分布就是刚好)等于 0.25.

对于正态分布, 如果 $\mathbb{P}(X \leq x) = p$, 那么 x 就是该正态分布的 p 分位数.

正态 Q-Q 图基于下列原理: 数据中一串数目的每个点都是该数据的某分位数, 把这些点(称为样本分位数点)和相应的理论上的分位数配对作出散点图, 如果该数据的确服从正态分布, 那么该图看上去应该像一条直线, 否则就不一定服从正态分布.

Q-Q 图

```
x=rnorm(100)
y=rchisq(100,2)
qqnorm(x);qqnorm(y)
par(mfrow=c(1,2))
```



随机向量及其分布函数

随机向量 $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}$, 其中每个分量 X_i 都是随机变量.

1. 联合分布函数:

$$F(x_1, x_2, \cdots, x_n) = \mathbb{P}(X_1 \leq x_1, X_1 \leq x_1, \cdots, X_n \leq x_n).$$

2. 联合概率密度函数: $\mathbf{X}$ 的联合概率密度函数 $f(x_1, x_2, \cdots, x_n)$ 为一个非负可积的函数, 且满足

$$F(x_1, x_2, \cdots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \cdots, t_n) dt_1 \cdots dt_n.$$

随机向量及其分布函数

3. 期望: $\mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), \mathbb{E}(X_2), \cdots, \mathbb{E}(X_n))'$.
4. 方差:

$$\begin{aligned} \mathrm{Var}(\mathbf{X}) &= \mathbb{E}\left(\mathbf{X} - \mathbb{E}(\mathbf{X})\right)\left(\mathbf{X} - \mathbb{E}(\mathbf{X})'\right) \\ &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} =: \Sigma, \end{aligned}$$

这里 $\sigma_{ij} = \mathbb{E}(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)), i, j = 1, \cdots, n.$

随机向量及其分布函数

5. 协方差: 两个随机向量 $\mathbf{X}$ 和 $\mathbf{Y}$, 其协方差矩阵为:

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) = \mathbb{E}\left(\mathbf{X}-\mathbb{E}(\mathbf{X})\right)\left(\mathbf{Y}-\mathbb{E}(\mathbf{Y})'\right).$$

如果 $\mathrm{Cov}(\mathbf{X},\mathbf{Y})=0$, 则称 $\mathbf{X}$ 和 $\mathbf{Y}$ 线性不相关.

6. 相关系数: 随机向量 $\mathbf{X}$ 的相关系数矩阵为:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix},$$

其中

$$\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}.$$

随机向量及其分布函数

7. 随机向量 $\mathbf{X}_{n\times 1}$, $\mathbf{a}_{n\times 1}$ 为常数向量, $\mathbf{A}$ 为常数矩阵, 有

$$\mathbb{E}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\mathbb{E}(\mathbf{X}), \quad \mathbb{E}(\mathbf{A}\mathbf{X}) = \mathbf{A}\mathbb{E}(\mathbf{X}).$$

8. 随机向量 $\mathbf{X}_{n\times 1}$, $\mathbf{Y}_{m\times 1}$, 常数矩阵 $\mathbf{A}_{p\times n}$ 和 $\mathbf{B}_{n\times m}$, 有

$$\begin{aligned} \mathrm{Cov}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) &= \mathbf{A}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}', \\ \mathrm{Var}(\mathbf{A}\mathbf{X}) &= \mathbf{A}\mathrm{Var}(\mathbf{X})\mathbf{A}'. \end{aligned}$$

9. 随机向量 $\mathbf{X}_{n\times 1}$, 记 $\Sigma = \mathrm{Var}(\mathbf{X})$, 则 Σ 对称, 正定.

随机变量的独立

X 与 Y 独立是指: Y 结果发生的概率与 X 结果发生的概率无关.

1. X 与 Y 独立:

$$\Longleftrightarrow F(x,y)=F(x)F(y)\Longleftrightarrow f(x,y)=f(x)f(y)$$

2. n 个随机变量 $X_1,X_2,\cdots,X_n$ 相互独立:

$$\begin{aligned}\Longleftrightarrow\quad & F(x_1,x_2,\cdots,x_n)=F(x_1)F(x_2)\cdots F(x_n)\\ \Longleftrightarrow\quad & f(x_1,x_2,\cdots,x_n)=f(x_1)f(x_2)\cdots f(x_n).\end{aligned}$$

3. 如果 $X_1,X_2,\cdots,X_n$ 相互独立, 则其中 k 个随机变量也相互独立.
4. 相互独立的随机变量(向量)的函数也相互独立.
5. 如果 X 与 Y 独立, 则 X 与 Y 不相关, 反之不一定成立. 对正态分布, 独立与不相关等价.

随机游走假说

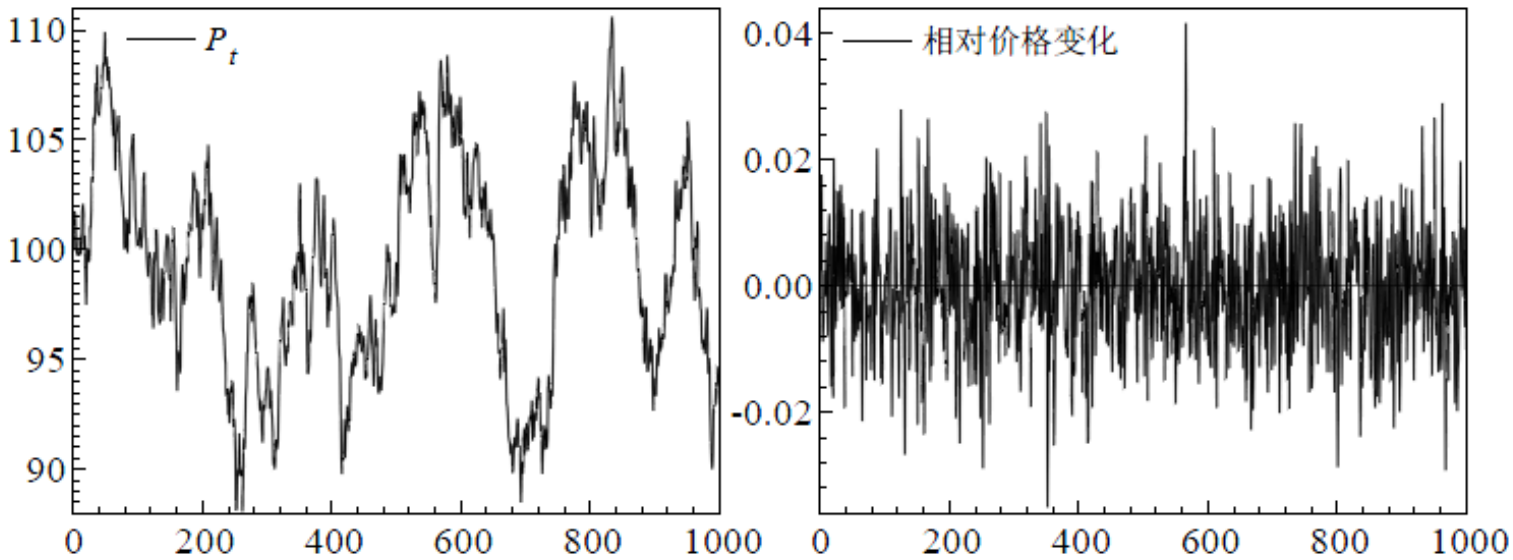
例(随机游走假说 Fama 1970): 若某股票价格 P_t 满足 $P_t = P_{t-1} + X_t$, 其中 $\{X_t\}$ 跨期独立, 则称 $\{P_t\}$ 服从随机游走. 注意到 $X_t = P_t - P_{t-1}$ 是从第 $t-1$ 期到第 t 期的股票价格变化. 通常考虑股票的收益率. 若 $\{P_t\}$ 满足: $\log P_t = \log P_{t-1} + X_t$, 其中 X_t 跨期独立, 则称其服从几何随机游走. 注意到

$$X_t = \log(P_t/P_{t-1}) = \log\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right) \approx \frac{P_t - P_{t-1}}{P_{t-1}}.$$

因此, X_t 可解释为股票价格的相对变化, 或股票的收益率.

随机游走假说

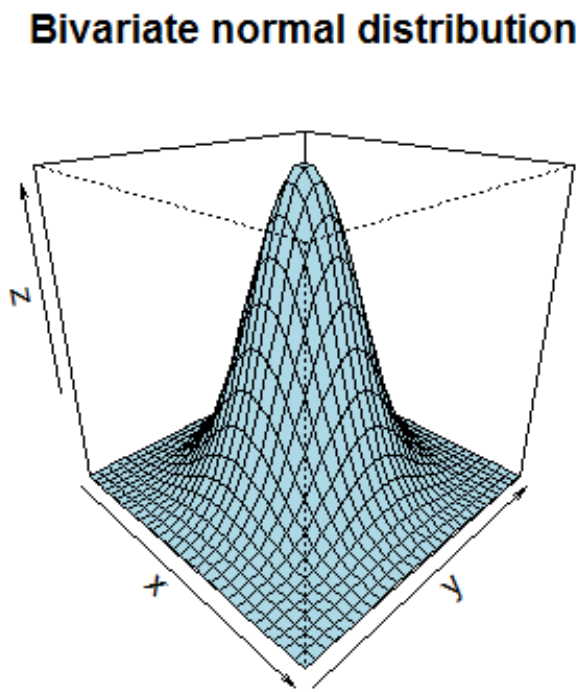
随机游走假设最重要的意义在于: 若 X_t 在不同期是序列独立的, 则未来股票市场的价格变化 X_t 就无法用股票价格的历史信息预测. 这种情况下, 我们称股票市场是**信息有效的**. 下图是计算机随机模拟生成的服从几何随机游走的价格观测值序列和相对价格的变化序列.



多元正态分布

正态随机向量: $\mathbf{X}_{n\times 1} \sim N(\boldsymbol{\mu}, \Sigma)$, 其联合概率密度函数为:

$$
\begin{aligned}
& f(x_1, x_2, \cdots, x_n) \\
= \quad & (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \right\}.
\end{aligned}
$$



边际分布

二元随机向量 (X,Y), 联合概率密度为 $f(x,y)$, 分布函数为 $F(x,y)$.

- 边际概率密度为:

$$f_x(x)=\begin{cases}\sum_y f(x,y) & Y \text{ 离散}\\ \int_y f(x,y)dy & Y \text{ 连续},\end{cases}$$

- 边际分布函数为:

$$F_x(x)=\begin{cases}F(x,+\infty)=\sum_{X\le x}\sum_y f(x,y)\\ F(x,+\infty)=\int_{-\infty}^{x}\int_{-\infty}^{+\infty} f(t,y)dtdy.\end{cases}$$

边际分布与联合分布

- 一般情况下, X 和 Y 的边缘分布(或者概率密度)无法完整地刻画 X 和 Y 的联合分布.

  $F(x,y) \Rightarrow F(x),\ F(y), \quad$ 但是, $\quad F(x),\ F(y) \nRightarrow F(x,y)$

- 在一种重要的特殊情形下, 即X **与** Y **相互独立时**, 可以用边缘分布确定联合分布.
- 如何解决? Copula 函数.

强大数定律

强大数定律: 若 $X_1, X_2, \cdots, X_n$ 独立同分布(iid)且 $\mathbb{E}(X_i) = \mu < \infty$, 则有几乎处处(a.s.)有:

$$\frac{\sum_{i=1}^{n} X_i}{n} \to \mathbb{E}(X_i) = \mu.$$

- 几乎处处: 以概率 1 成立.
- 若 $X_n \overset{a.s.}{\to} X$（几乎处处收敛）, 则有 $X_n \overset{P}{\to} X$（依概率收敛）.
- 强大数定律的一个直接应用: 蒙特卡洛(Monte Carlo).

蒙特卡洛

- 蒙特卡洛积分是一种基于随机抽样的统计方法.
- 假设 g 是一个可积函数, 我们要计算 $\int_a^b g(x)dx$.
- **回顾**: 若随机变量 X 的密度函数为 $f(x)$, 则随机变量 $Y = g(X)$ 的期望为:

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

- 从 X 的分布中产生一些随机数, 则利用强大数定律 $\mathbb{E}g(X)$ 的估计就是相应的样本平均值.

蒙特卡洛

计算 $\int_a^b g(x)dx$.

$$\int_a^b g(x)dx = (b-a)\int_a^b g(x)\cdot \underbrace{\frac{1}{b-a}}_{U(a,b)\text{的 pdf}} dx.$$

具体步骤

1. 从均匀分布 $U(a,b)$ 中产生 i.i.d 样本 $X_1, \cdots, X_n$;
2. 计算均值 $\overline{g_n(X)} = \frac{1}{n}\sum_{i=1}^n g(X_i)$;
3. $\int_a^b g(x)dx \approx (b-a)\overline{g_n(X)}$.

蒙特卡洛

例 1 计算 $\int_0^1 e^{-x}dx$

```
x=runif(10000)  #产生10000个服从U(0,1)的随机数
theta.hat1=mean(exp(-x))  #计算均值
theta.hat1
```

```
[1] 0.6319105
```

```
1-exp(-1)  #直接计算
```

```
[1] 0.6321206
```

蒙特卡洛

例 2 计算 $\int_2^4 e^{-x}dx$

```
x=runif(10000,min=2,max=4)  #产生10000个服从U(2,4)的随机数
theta.hat2=(4-2)*mean(exp(-x))  #计算均值
theta.hat2
```

```
[1] 0.1167818
```

```
exp(-2)-exp(-4)  #直接计算
```

```
[1] 0.1170196
```

蒙特卡洛

例 3 计算 $\int_0^\infty x^{-x}dx$

$$\int_0^\infty x^{-x}dx = \int_0^\infty \underbrace{x^{-x}e^x}_{=:g(x)} \underbrace{e^{-x}}_{Exp(1)的pdf} dx$$

```
x=rexp(10000,1)  #产生10000个服从Exp(1)的随机数
theta.hat3=mean(exp(x)/(x^x))  #计算均值
theta.hat3
```

```
[1] 1.993042
```

蒙特卡洛

例 4 计算 $\int_0^1 \int_0^1 \sin(x+y)dxdy$

```
u=runif(10000)
v=runif(10000)
theta.hat4=mean(sin(u+v))  #计算均值
theta.hat4
```

```
[1] 0.7723024
```

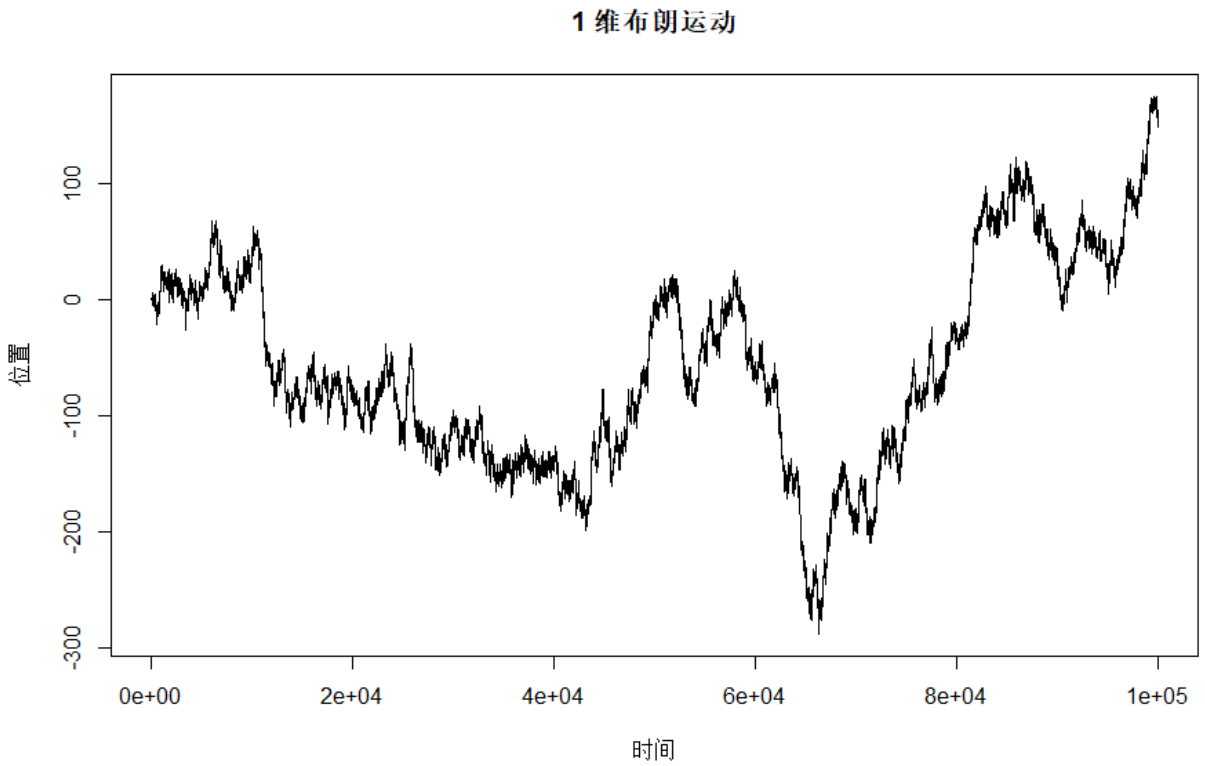
例 5 计算 $\int_3^7 \int_1^5 \sin(x+y)dxdy$

```
u1=runif(10000,3,7)
v1=runif(10000,1,5)
theta.hat5=mean(sin(u1+v1))*(7-3)*(5-1)  #计算均值
theta.hat5
```

```
[1] 3.401881
```

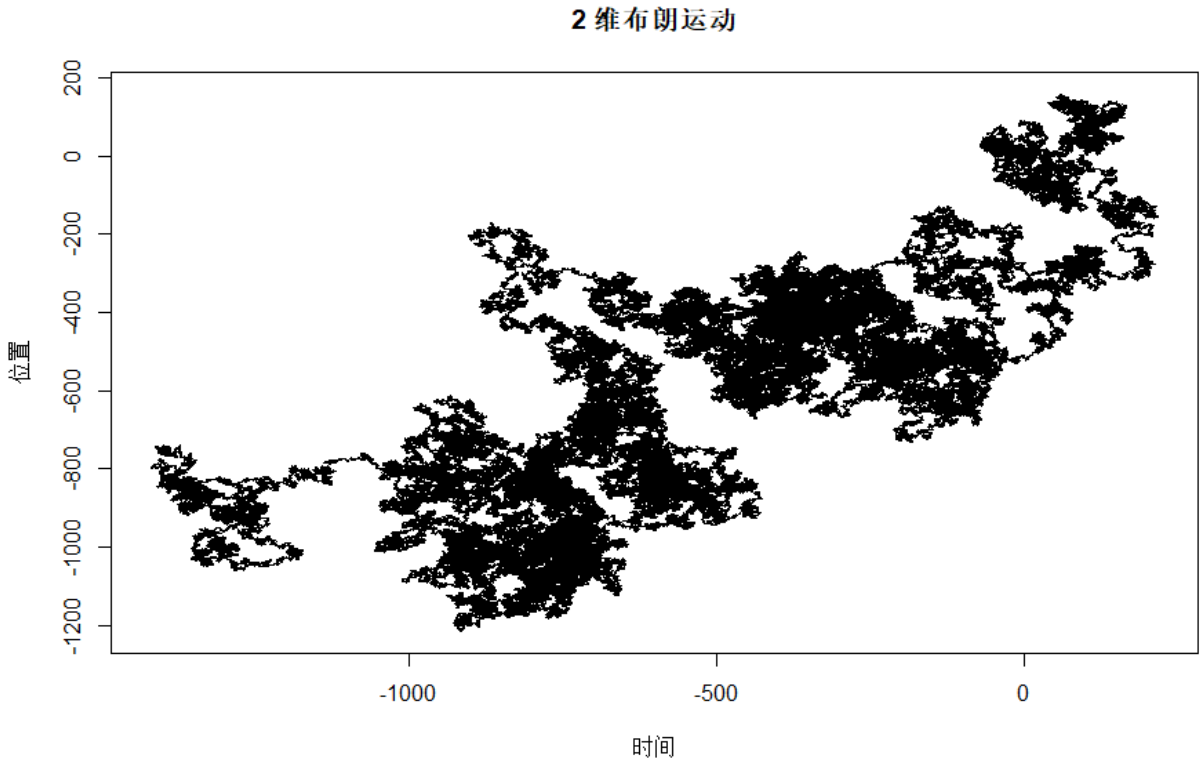
一维布朗运动模拟

```
set.seed(100)
Noise=rnorm(100000)
BM=cumsum(Noise)
plot(BM, type="l", main="1 维布朗运动", xlab="时间", ylab="位置")
```



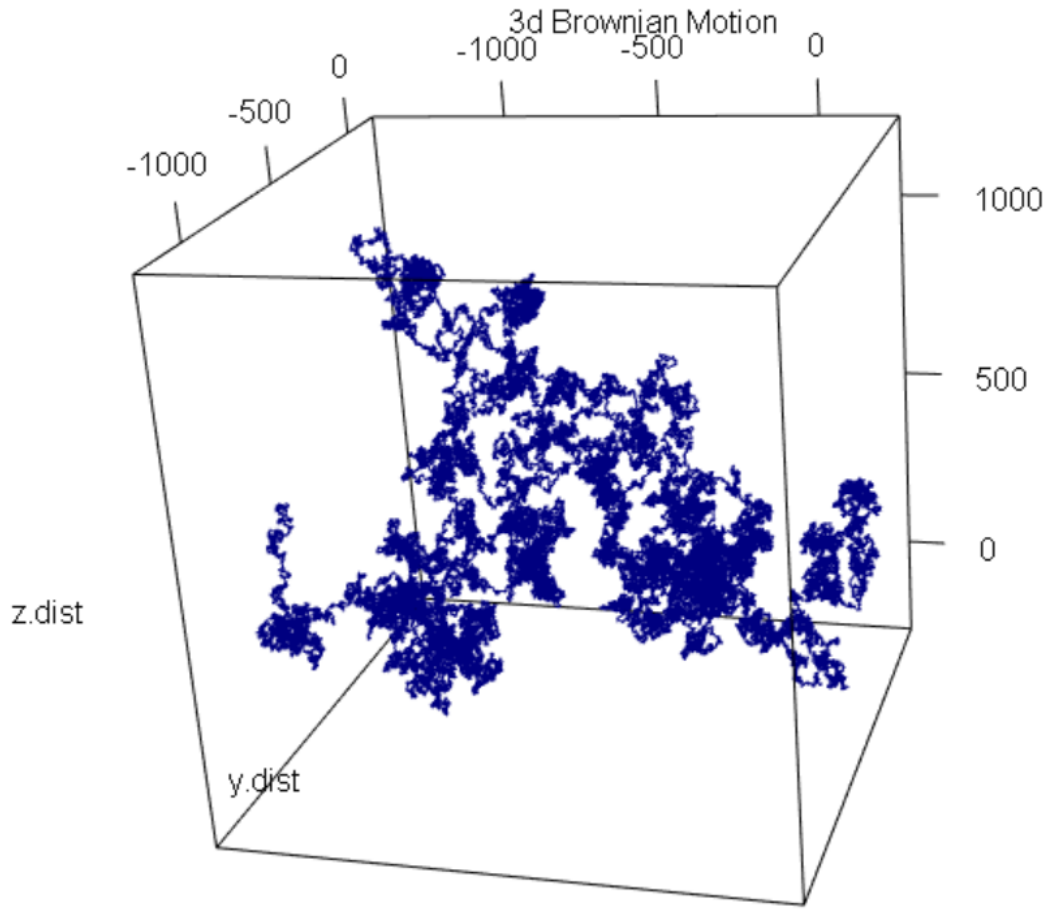
二维布朗运动模拟

```
set.seed(100)
x.noise <- rnorm(1000000);y.noise <- rnorm(1000000)
x.dist <- cumsum(x.noise);y.dist <- cumsum(y.noise)
plot(x.dist, y.dist, type="l", main="2 维布朗运动", xlab="时间", ylab="位置")
```



三维布朗运动模拟

```
set.seed(100)
z.noise <- rnorm(1000000);z.dist <- cumsum(z.noise)
library(rgl)
plot3d(x.dist, y.dist, z.dist, type="l",col='navyblue',main="3 维布朗运动")
```



大数定律与中心极限定理

大数定律和**中心极限定理**是概率论极限理论的主要分支, 也是概率论和数理统计的重要基础. 下面给予形式上的介绍.

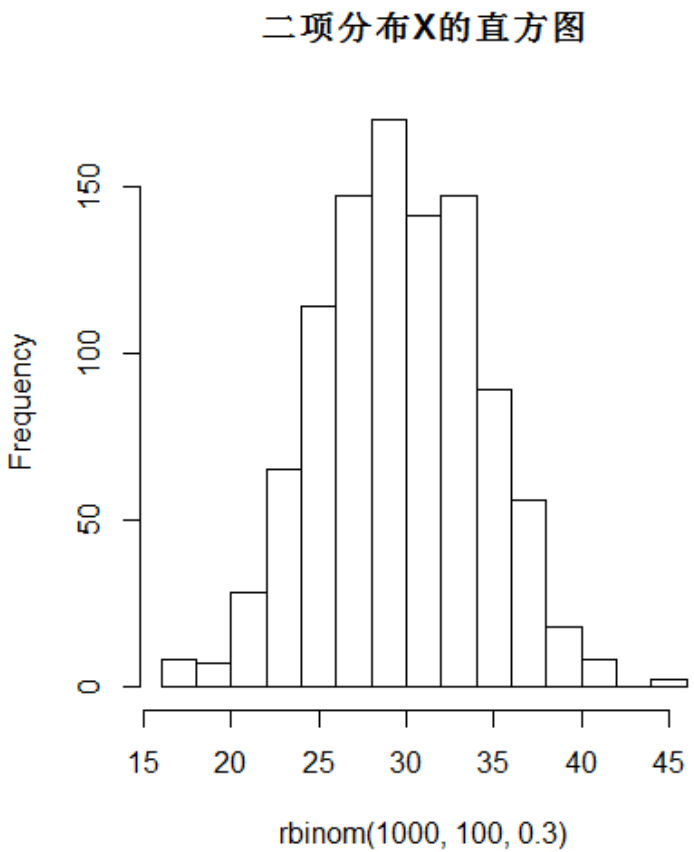
若 $X_1, X_2, \cdots, X_n$ 独立同分布(iid)于某一期望为 μ, 方差为 σ^2 的分布. 利用独立性以及期望的线性性和方差的性质, 我们有:

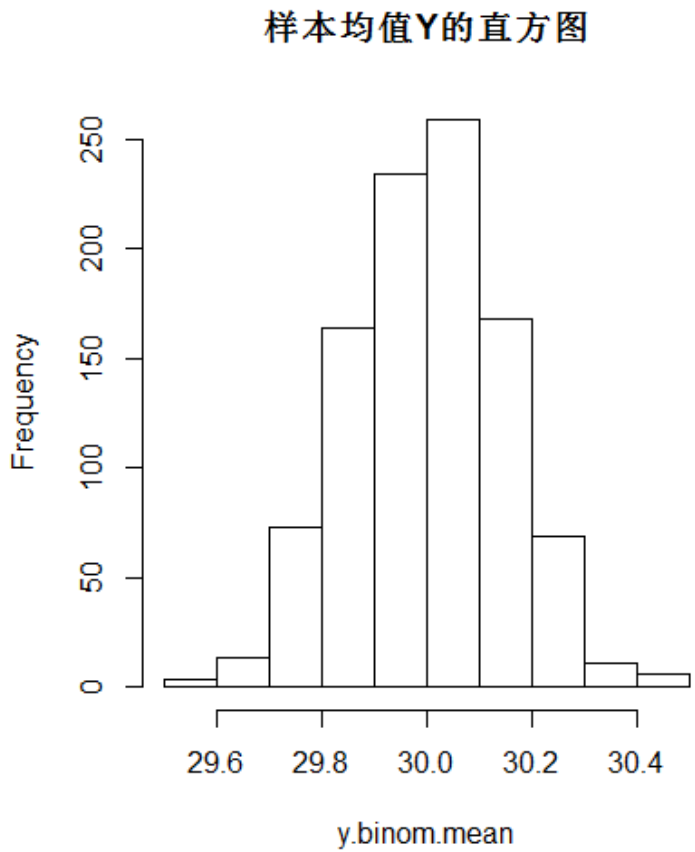
$$\mathbb{E}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = \mu,$$
$$\mathbf{D}\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{\sum_{i=1}^n \mathbf{D}(X_i)}{n^2} = \frac{\sigma^2}{n}.$$

- **注** 这两个结果在后续中经常用到, 应熟记.
- 大数定律: $\frac{\sum_{i=1}^n X_i}{n}$ 收敛到 μ.
- 中心极限定理: $\frac{\sum_{i=1}^n X_i}{n}$ 的分布收敛到 $N(\mu, \frac{\sigma^2}{n})$.

中心极限定理

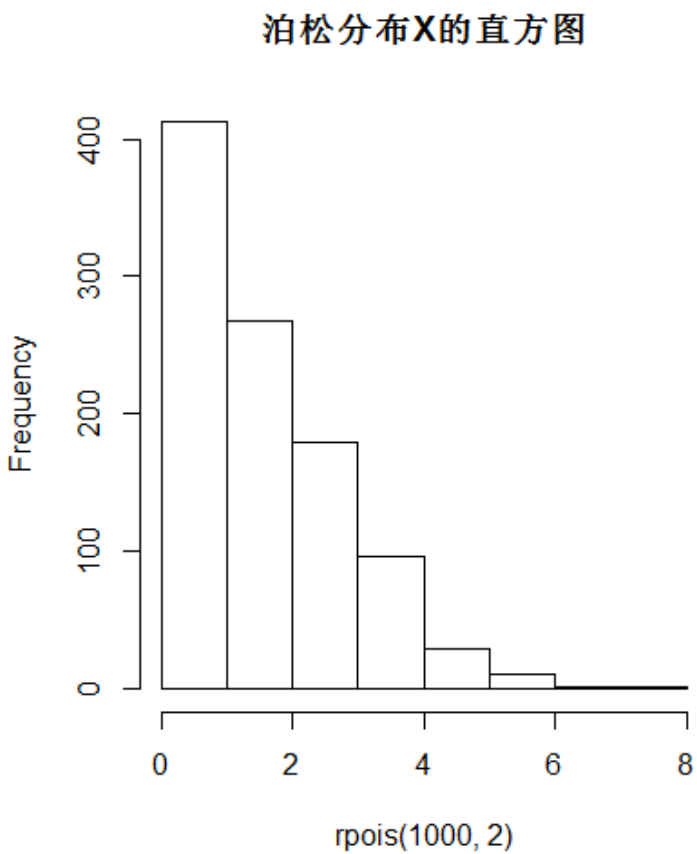
例(二项分布) 假设 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} binom(100, 0.3)$, 其样本均值 $Y = \frac{\sum_{i=1}^{n} X_i}{n} \overset{approx}{\sim} N(30, 0.021)$. 下图左边为 1000 组随机产生的 1000 个服从 $binom(100, 0.3)$ 分布样本点的直方图, 右边为其样本均值的直方图.

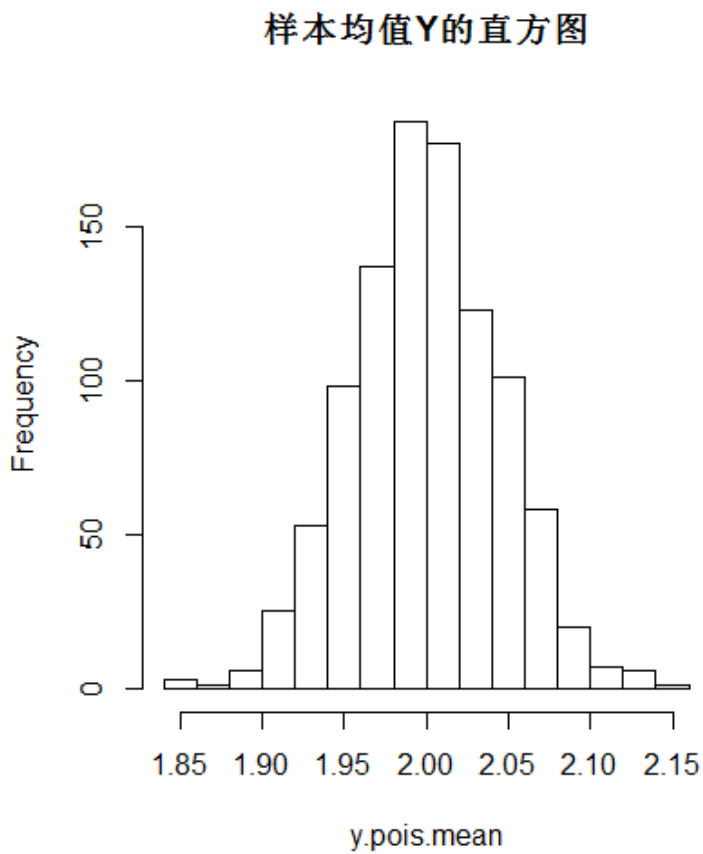




中心极限定理

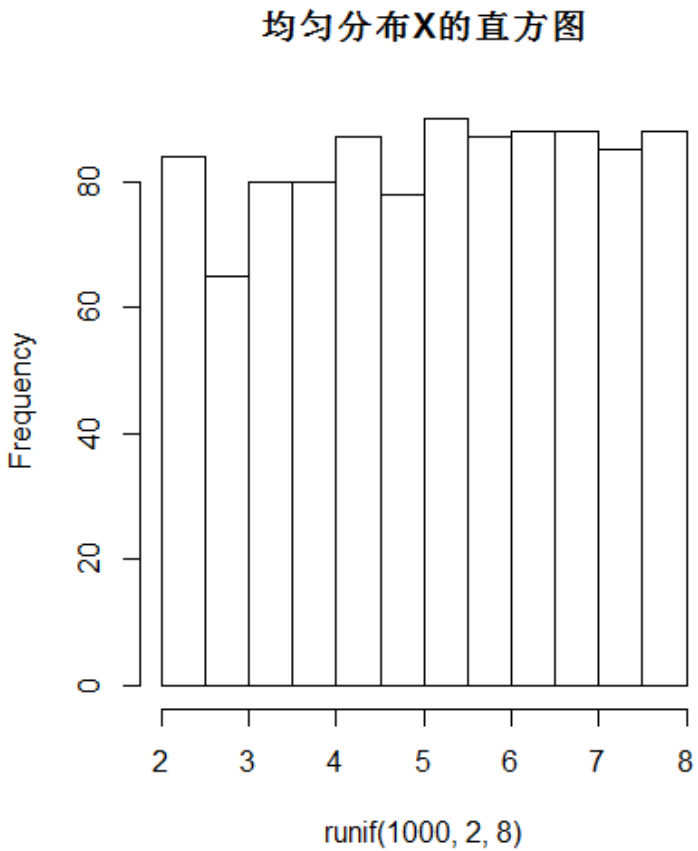
例(Poisson 分布) 假设 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} \pi(2)$, 其样本均值 $Y = \frac{\sum_{i=1}^{n} X_i}{n} \overset{approx}{\sim} N(2, 0.002)$. 下图左边为 1000 组随机产生的 1000 个服从 $\pi(2)$ 分布样本点的直方图, 右边为其样本均值的直方图.

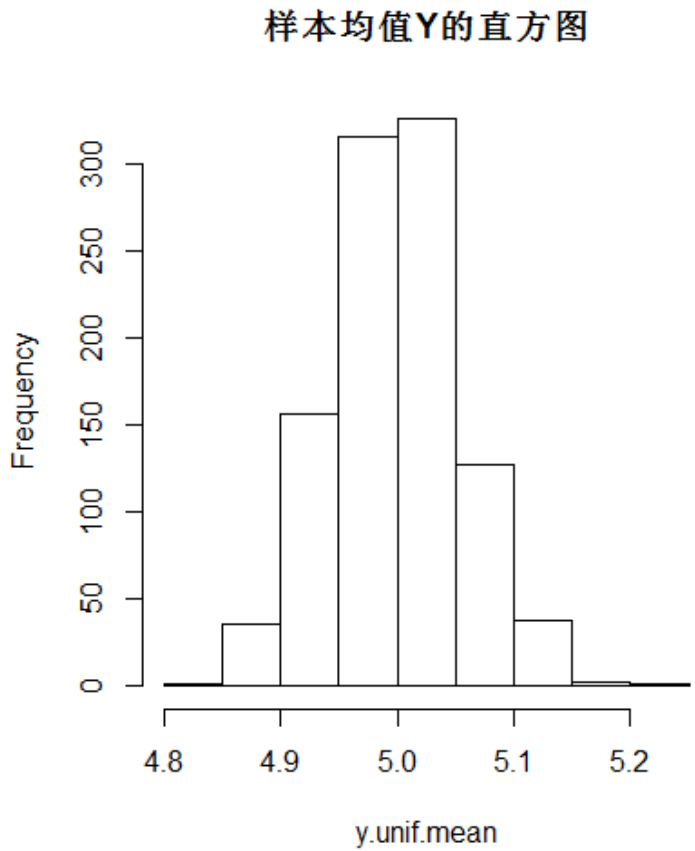




中心极限定理

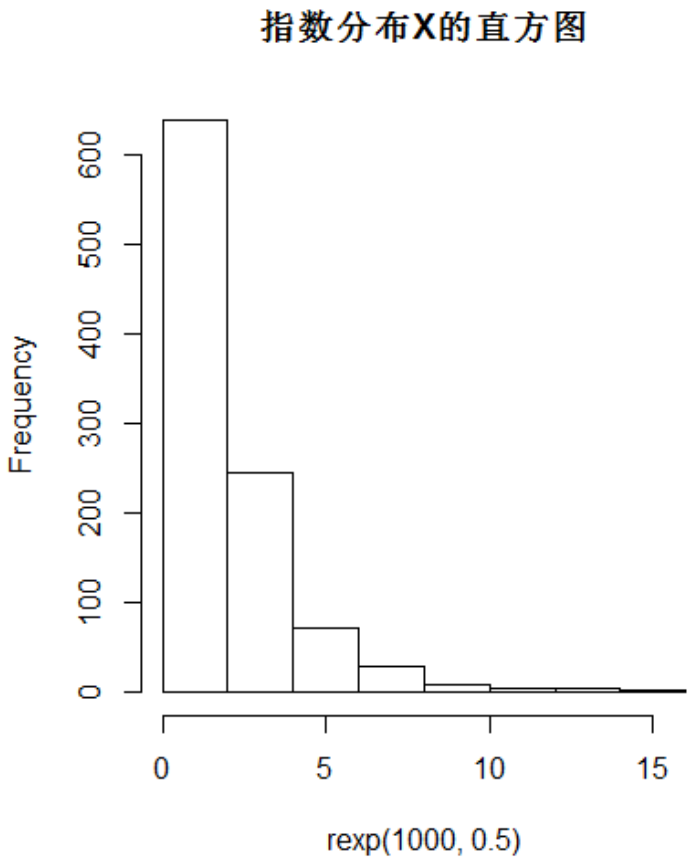
例(均匀分布) 假设 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} U(2,8)$, 其样本均值 $Y = \frac{\sum_{i=1}^{n} X_i}{n} \overset{approx}{\sim} N(5, 0.003)$. 下图左边为 1000 组随机产生的 1000 个服从 $U(2,8)$ 分布样本点的直方图, 右边为其样本均值的直方图.

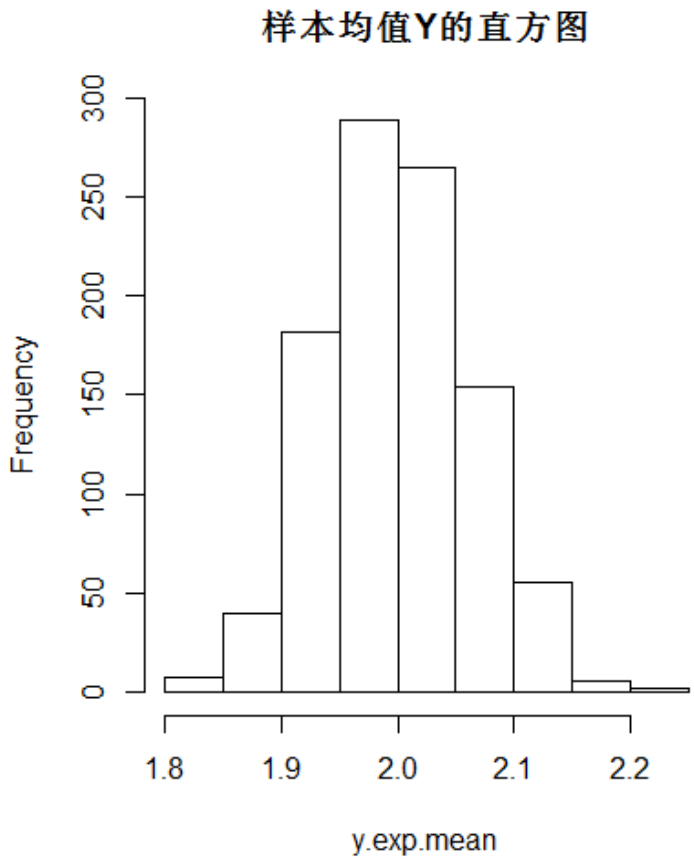




中心极限定理

例(指数分布) 假设 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} exp(0.5)$, 其样本均值 $Y = \frac{\sum_{i=1}^{n} X_i}{n} \overset{approx}{\sim} N(2, 0.004)$. 下图左边为 1000 组随机产生的 1000 个服从 $exp(0.5)$ 分布样本点的直方图, 右边为其样本均值的直方图.





总体和样本

定义

总体: 被研究对象的全体, 是一维或多维的随机变量.
样本: 从总体中抽取的部分单位的集合, 应该满足代表性和独立性原则.

设随机变量 $X \sim F(x), X_1, X_2, \cdots, X_n$ 为来自总体 X 的一个容量为 n 的子样.

1. 代表性是指子样中每个 X_i 都与总体 X 具有**相同的分布**.
2. 独立性是指 $X_1, X_2, \cdots, X_n$ 为**相互独立**的随机变量.

这样的子样称为简单随机子样.
统计量: 不含未知参数的样本的函数.

常用统计量

设随机变量 $X \sim F(x), X_1, X_2, \cdots, X_n$ 为来自总体 X 的一个容量为 n 的子样.

1. 样本均值: 描述中心化趋势的统计量.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

2. 样本方差: 描述观测值离散程度的统计量.

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

样本标准差: $S_X = \sqrt{S_X^2}$.

常用统计量

3. 样本协方差: 设随机变量 $Y \sim F(y), Y_1, Y_2, \cdots, Y_n$ 位来自总体 Y 的一个容量为 n 的子样.

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

样本相关系数: $r_{XY} = \frac{S_{XY}}{S_X S_Y}$.

4. 样本偏度

$$\hat{\nu} = \frac{1}{n-1}\sum_{i=1}\frac{(X_i - \bar{X})^3}{S_X^3}.$$

5. 样本峰度

$$\hat{\kappa} = \frac{1}{n-1}\sum_{i=1}\frac{(X_i - \bar{X})^4}{S_X^4}.$$

常用统计量

6. 样本 k 阶(原点)矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \cdots$$

7. 样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^k, \quad k = 2, 3 \cdots$$

- 若总体 X 的 k 阶矩 $\mathbb{E}(X^k) =: \mu_k$ 存在, 则当 $n \to \infty$ 时, $A_k \xrightarrow{P} \mu_k$, $k = 1, 2, \cdots$. 这是因为 $X_1, \cdots, X_n$ 独立同分布与 X, 所以 $X_1^k, \cdots, X_n^k$ 独立同分布于 X^k, 从而

$$\mathbb{E}(X_1^k) = \cdots = \mathbb{E}(X_n^k) = \mu_k$$

利用辛钦大数定理, 我们得到矩估计方法的理论基础:

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k, \quad k = 1, 2, \cdots$$

正态总体的抽样分布

设随机变量 $X \sim N(\mu_X, \sigma_X^2)$, $X_1, X_2, \cdots, X_n$ 为来自总体 X 的一个容量为 n 的子样.

1. 样本均值的分布:

$$\bar{X} \sim N\left(\mu_X, \frac{1}{n}\sigma_X^2\right), \quad \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0,1).$$

2. 样本方差的分布:

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1).$$

重要性质: $\bar{X}$ 与 S_X^2 相互独立.

正态总体的抽样分布

3. t 分布:

$$\frac{\bar{X}-\mu_X}{S_X/\sqrt{n}} \sim t(n-1).$$

4. F 分布: 设随机变量 $Y \sim N(\mu_Y, \sigma_Y^2)$, $Y_1, Y_2, \cdots, Y_m$ 位来自总体 Y 的容量为 m 的子样, 则

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(n-1, m-1).$$

点估计

设任意随机变量 $X \sim F(x;\theta); \theta \in \Theta$. $F(x;\theta)$ 形式已知, 但是参数 θ 未知. θ 可以是一维的参数, 也可以是 K 维的:
$\theta = (\theta_1, \theta_2, \cdots, \theta_K)'$. Θ 为参数空间.
若 $X_1, X_2, \cdots, X_n \sim X$, 点估计就是利用样本提供的信息确定参数 θ. 设参数是一维的, 若统计量:

$$\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$$

作为未知参数 θ 的估计, $\hat{\theta}$ 为参数的一个估计量. 记 $x_1, x_2, \cdots, x_n$ 位样本的观测值, 则 $\hat{\theta}(x_1, x_2, \cdots, x_n)$ 为参数 θ 的一个估计值.
一般情况下, 不区分估计量和估计值.

常见的点估计方法有:矩估计、极大似然估计、最小二乘法.

点估计量的性质

1. 无偏性: 设 $\hat{\theta}$ 是 θ 的估计, 若 $\mathbb{E}(\hat{\theta}) = \theta$, 则称 $\hat{\theta}$ 是 θ 的无偏估计.
   如果 $\mathbb{E}(\hat{\theta}) \neq \theta$, 则称 $\hat{\theta}$ 是有偏的, 定义其偏差为:

$$bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

2. 有效性: 对给定的样本容量 n, 如果无偏估计量 $\hat{\theta}$ 的方差小于任何其它无偏估计量的方差, 则称 $\hat{\theta}$ 是一个有效的无偏估计量.
3. 均方误差:

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 = \left[bias(\hat{\theta})\right]^2 + \mathrm{Var}(\hat{\theta}).$$

例子

设 $X_1, X_2, \cdots, X_n$ 为来自总体 X, 总体均值和方差分别为 μ, σ^2, $\bar{X}$, S_n 分别样本均值和样本方差, 则有:

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu; \\
\mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}; \\
\mathbb{E}(S_n^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i-\mu)-(\bar{X}-\mu)\right]^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i-\mu)^2 - \frac{n}{n-1}\mathbb{E}(\bar{X}-\mu)^2 = \sigma^2.
\end{aligned}$$

即样本均值 $\bar{X}$ 和样本方差 S_n^2 分别是总体均值和总体方差的无偏估计.

假设检验原理

- **原理**: 小概率事件原理. 小概率事件在一次试验中几乎是不可能发生的, 一旦它发生, 我们就有理由拒绝原假设.
- **方法**: 带有概率性质的反证法.
  在假定原假设 H_0 成立的情况下(与原假设对立的假设为备择假设 H_1), 构造一个小概率事件 A, $\mathbb{P}(A) \leq \alpha$(α 很小, 称之为检验的显著性水平, 通常取 $\alpha = 0.01, 0.05$), 经过一次试验或者观测(即获得的样本信息), 如果事件 A 发生, 则拒绝原假设 H_0; 如果事件 A 没有发生, 则接受原假设 H_0.
- **注**: 假设检验所采用的概率性质反证法, 是基于总体中一部分样本信息的, 不能完全代表总体, 而是在一定概率(置信度)下来推断总体特征. 通常把 $1-\alpha$ 称为置信度(或置信水平), 即对推断结果的把握程度, 可靠性.

假设检验的步骤:以 Z 检验为例

设随机变量 $X \sim N(\mu, \sigma^2)$, σ^2 已知, $X_1, X_2, \cdots, X_n \sim X$. 现检验 μ 是否等于 μ_0.

1. 建立统计假设(通常将要证明的陈述作为备择假设):

$$H_0: \ \mu = \mu_0; \quad H_1: \ \mu \neq \mu_0 \text{ (双边检验)}.$$

2. 构造检验统计量:
   由于 $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, 从而 H_0 成立的条件下, 有

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1). \quad \text{中心极限定理}$$

假设检验的步骤以 Z 检验为例

3. 对给定的显著性水平 α, 构造小概率事件:

$$\mathbb{P}_{H_0}\left(\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > Z_{\frac{\alpha}{2}}\right) = \alpha,$$

则小概率事件为:

$$\left\{\left|\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right| > Z_{\frac{\alpha}{2}}\right\}.$$

4. 做出统计推断: 将统计量 $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ 代入样本观测值为 Z_0, 判断小概率事件是否发生. 如果 $|Z_0| \geq Z_{\frac{\alpha}{2}}$, 则拒绝 H_0; 否则, 接受 H_0 (这种推断方法称为:临界值法). **现在常用 p-值来做决策.**

注 上述检验也称为 Z 检验.

p-值

p-值(也称为边际显著性水平):

$$p\text{-值} = \begin{cases} \mathbb{P}_{H_0}\left(\text{统计量 } Z \geq \text{ 样本值 } Z_0\right) & \text{单侧检验} \\ \mathbb{P}_{H_0}\left(\text{统计量 } |Z| \geq \text{ 样本值 } |Z_0|\right) & \text{双侧检验} \end{cases}.$$

- 可以把 p-值解释成"**拒绝原假设的最小显著性水平**".
- 可以用 p-值进行假设检验的判断: 如果 p-值$\leq \alpha$, 则拒绝 H_0; 否则不能拒绝 H_0.

注 当不能拒绝原假设时, **严格讲**, 我们不说"接受原假设", 因为没有证据可以说明原假设是真的.

例子: t-检验

问题: 从一个总体中得到一个样本, 根据给定的这个样本, 判断总体均值是否为一个特殊值 μ.
方法: 利用t.test(x,mu=m) 函数, 输出包括一个 p-值, 若 $p < 0.05$, 则拒绝原假设; 若 $p > 0.05$, 则接受原假设.

```
x<-rnorm(50,mu=100,sd=15)
t.test(x,mu=90)    #检验总体均值是否可能为~90.
```

```
One Sample t-test

data:  x
t = 3.0062, df = 49, p-value = 0.004164
alternative hypothesis: true mean is not equal to 90
95 percent confidence interval:
  92.26098 101.37903
sample estimates:
mean of x
 96.82001
```

p-值很小(< 0.05), 所以(根据样本数据)90 不可能是总体均值.

例子: t-检验

```
x<-rnorm(50,mu=100,sd=15)
t.test(x,mu=100)    #检验总体均值是否可能为~100.
```

```
One Sample t-test

data:  x
t = 0.6946, df = 49, p-value = 0.4906
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  98.01543 104.08099
sample estimates:
mean of x
 101.0482
```

p-值很大(> 0.05), 所以(根据样本数据)100 很可能是总体均值.

假设检验的两类错误

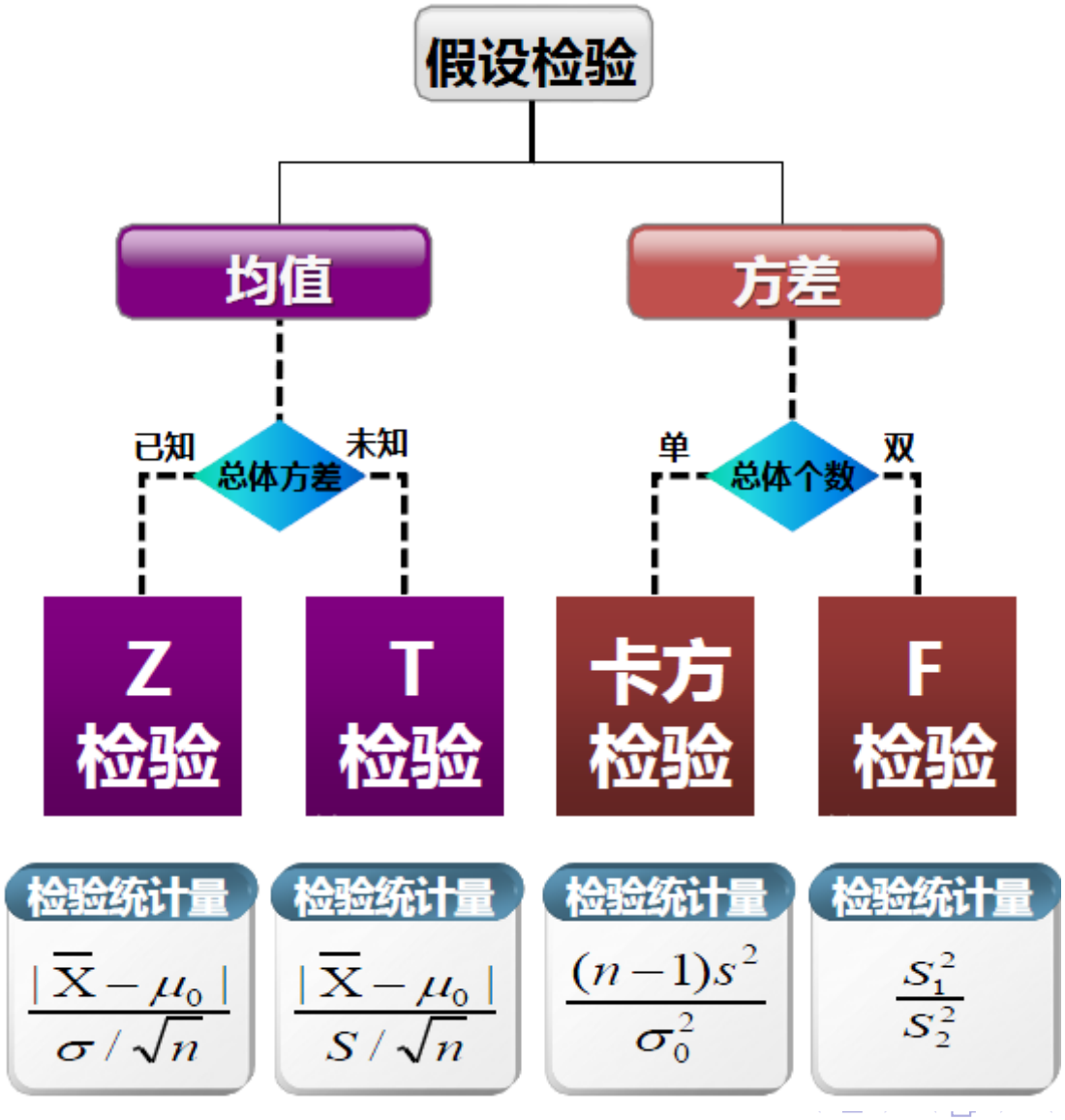
- 第一类错误: 原假设为真, 拒绝了原假设. 犯第一类错误的概率(拒真概率)为检验的显著性水平 α.

$$\alpha = \mathbb{P}(\text{拒绝 } H_0 | H_0 \text{ 为真})$$

- 第二类错误: 原假设为假, 接受了原假设. 犯第二类错误的概率(取伪概率)记为 β. $\beta \neq 1 - \alpha$.

$$\beta = \mathbb{P}(\text{接受 } H_0 | H_0 \text{ 为假})$$

常见的假设检验


假设检验
均值
方差
已知
总体方差
未知
单
总体个数
双
Z 检验
T 检验
卡方检验
F 检验
检验统计量
$\frac{|\overline{X}-\mu_0|}{\sigma/\sqrt{n}}$
检验统计量
$\frac{|\overline{X}-\mu_0|}{S/\sqrt{n}}$
检验统计量
$\frac{(n-1)s^2}{\sigma_0^2}$
检验统计量
$\frac{S_1^2}{S_2^2}$