中国石油大学(北京)克拉玛依校区
CHINA UNIVERSITY OF PETROLEUM - BEIJING AT KARAMAY

Multivariate Statistical Analysis

多元统计分析

2026 年 3 月 31 日

已学知识点 (Recap)

第 4 章 多元分布

4.1 分布函数与密度函数

- 累积分布函数 (cdf) 的定义为 $F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} < \boldsymbol{x})$.
- 如果概率密度函数 (pdf) 存在，则 $F(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$.
- 概率密度函数 (pdf) 满足 $\int_{-\infty}^{\infty} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$.
- 将随机向量 $\boldsymbol{X}$ 分块为 $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$，设 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 的联合 (即 $\boldsymbol{X}$ 的) 分布函数为 F，则 $\boldsymbol{X}_1$ 的边际分布函数为 $F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = \mathbb{P}(\boldsymbol{X}_1 < \boldsymbol{x}_1)$，$\boldsymbol{X}_1$ 的边际概率密度函数为 $f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = \int_{-\infty}^{\infty} f(\boldsymbol{x}_1, \boldsymbol{x}_2)\,\mathrm{d}\boldsymbol{x}_2$. 对联合概率密度函数求导亦可得到相同的边际概率密度函数.
- 给定 $\boldsymbol{X}_1 = \boldsymbol{x}_1$ 时 $\boldsymbol{X}_2$ 条件概率密度函数为 $f(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) = \dfrac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)}$.

已学知识点 (Recap)

第 4 章 多元分布

4.1 分布函数与密度函数

- 两个随机向量 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 独立，当且仅当 $f\left(\boldsymbol{x}_1,\boldsymbol{x}_2\right)=f_{X_1}\left(\boldsymbol{x}_1\right)\cdot f_{X_2}\left(\boldsymbol{x}_2\right)$，也等价于 $f\left(\boldsymbol{x}_2\middle|\boldsymbol{x}_1\right)=f_{X_2}\left(\boldsymbol{x}_2\right)$，$f\left(\boldsymbol{x}_1\middle|\boldsymbol{x}_2\right)=f_{X_1}\left(\boldsymbol{x}_1\right)$.
- 不同的联合概率密度函数可以有相同的边际概率密度函数.

已学知识点 (Recap)

第 4 章 多元分布

4.2 矩与特征函数

- 随机向量 $\boldsymbol{X}$ 的期望为 $\boldsymbol{\mu}=\int \boldsymbol{x} f(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}$，协方差矩阵为 $\boldsymbol{\Sigma}=\operatorname{Var}(\boldsymbol{X})=\mathbb{E}\left[\left(\boldsymbol{X}-\boldsymbol{\mu}\right)\left(\boldsymbol{X}-\boldsymbol{\mu}\right)^{\mathrm{T}}\right]$，将其记为 $\boldsymbol{X} \sim\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$.
- 求期望是线性运算，即 $\mathbb{E}\left(\alpha \boldsymbol{X}+\beta \boldsymbol{Y}\right)=\alpha \mathbb{E}(\boldsymbol{X})+\beta \mathbb{E}(\boldsymbol{Y})$. 如果 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 相互独立，则 $\mathbb{E}\left(\boldsymbol{X} \boldsymbol{Y}^{\mathrm{T}}\right)=\mathbb{E}(\boldsymbol{X}) \mathbb{E}\left(\boldsymbol{Y}^{\mathrm{T}}\right)$.
- 两个随机向量 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的协方差矩阵为

$$\boldsymbol{\Sigma}_{\boldsymbol{X} \boldsymbol{Y}}=\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y})=\mathbb{E}\left[(\boldsymbol{X}-\mathbb{E}(\boldsymbol{X}))(\boldsymbol{Y}-\mathbb{E}(\boldsymbol{Y}))^{\mathrm{T}}\right]=\mathbb{E}\left(\boldsymbol{X} \boldsymbol{Y}^{\mathrm{T}}\right)-\mathbb{E}(\boldsymbol{X}) \mathbb{E}\left(\boldsymbol{Y}^{\mathrm{T}}\right).$$

  如果 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立，则 $\operatorname{Cov}(\boldsymbol{X}, \boldsymbol{Y})=0$.
- 随机向量 $\boldsymbol{X}$ 的特征函数 (cf) 为 $\varphi_{\boldsymbol{X}}(\boldsymbol{t})=\mathbb{E}\left(\mathrm{e}^{\mathrm{i} \boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}}\right)$.
- p 维随机向量 $\boldsymbol{X}$ 的分布完全由形如 $\boldsymbol{t}^{\mathrm{T}} \boldsymbol{X}$ 的所有一维分布确定，其中 $\boldsymbol{t} \in \mathbb{R}^{p}$ (Cramer-Wold 定理).
- 条件期望 $\mathbb{E}\left(\boldsymbol{X}_{2} \middle| \boldsymbol{X}_{1}\right)$ 是用 $\boldsymbol{X}_{1}$ 的一个函数对 $\boldsymbol{X}_{2}$ 在均方误差 (MSE) 意义下的最佳近似.

已学知识点 (Recap)

第 4 章 多元分布

4.3 变换

- 如果 $\boldsymbol{X}$ 的密度函数为 $f_{\boldsymbol{X}}(\boldsymbol{x})$，设 $\boldsymbol{X}=u(\boldsymbol{Y})$，则变换后的随机向量 $\boldsymbol{Y}$ 的概率密度函数为 $f_{\boldsymbol{Y}}(\boldsymbol{y})=\operatorname{abs}\left(\left|\mathscr{J}\right|\right)\cdot f_{\boldsymbol{X}}\left[u(\boldsymbol{y})\right]$，其中 $\mathscr{J}$ 表示变换的 Jacobian 矩阵 $\mathscr{J}=\left(\dfrac{\partial u(y_i)}{\partial y_j}\right)$.
- 对于线性关系 $\boldsymbol{Y}=\mathscr{A}\boldsymbol{X}+\boldsymbol{b}$ 的情形，$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的概率密度函数之间的关系为

$$f_{\boldsymbol{Y}}(\boldsymbol{y})=\operatorname{abs}\left(\left|\mathscr{A}\right|^{-1}\right)\cdot f_{\boldsymbol{X}}\left\{\mathscr{A}^{-1}(\boldsymbol{y}-\boldsymbol{b})\right\}.$$

已学知识点 (Recap)

第 4 章 多元分布

4.4 多元正态分布

- p 维正态分布 $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的概率密度函数为

$$f(\boldsymbol{x}) = \left|2\pi\Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}} \Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\}.$$

  多元正态分布的密度函数的等高线是椭球面，其半轴的长度与 $\sqrt{\lambda_i}$ 成比例，其中 λ_i 表示 Σ 的特征值 $\left(i = 1, 2, \ldots, p\right)$.

- 设 $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right)$，经 Mahalanobis 变换有 $\boldsymbol{Y} = \Sigma^{-1/2}\left(\boldsymbol{X}-\boldsymbol{\mu}\right) \sim N_p\left(\boldsymbol{0}, \mathscr{I}_p\right)$. 反之，我们可以由 $\boldsymbol{Y} \sim N_p\left(\boldsymbol{0}, \mathscr{I}_p\right)$ 经过变换 $\boldsymbol{X} = \Sigma^{1/2}\boldsymbol{Y} + \boldsymbol{\mu}$ 得到 $\boldsymbol{X} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right)$.

多元正态分布

- 分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的几何特征

定理 4.8 多元正态分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的特征函数为 $\varphi_X(\boldsymbol{t}) = \exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right).$

- 特征函数和分布函数是相互唯一确定的.
  - 分布函数可以确定特征函数：$f_\xi(t) \triangleq E\left(\mathrm{e}^{it\xi}\right) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\,\mathrm{d}F_\xi(x)$
  - (唯一性定理) 分布函数由特征函数唯一确定.
  - 若特征函数 $f(t)$ 绝对可积，则相应分布函数 $F(x)$ 的导数存在并连续，并且 $F'(x) = p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx} f(t)\,\mathrm{d}t$.

$$\Longrightarrow f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^p}\int \exp\left(-\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{x}\right)\cdot\exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right)\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|2\pi\mathscr{I}_p\right|}\int \exp\left\{-\frac{1}{2}\left[\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t} + 2\mathbf{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|2\pi\Sigma^{-1/2}\Sigma^{1/2}\right|}\int \exp\left\{-\frac{1}{2}\left[\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t} + 2\mathbf{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|\sqrt{2\pi}\Sigma^{-1/2}\cdot\sqrt{2\pi}\Sigma^{1/2}\right|}\int \exp\left\{-\frac{1}{2}\left[\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t} + 2\mathbf{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}\mathrm{d}\boldsymbol{t}$$

多元正态分布

- 分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的几何特征

定理 4.8 多元正态分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的特征函数为 $\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right).$

- 特征函数和分布函数是相互唯一确定的.
  - 分布函数可以确定特征函数：$f_{\xi}(t) \triangleq E\left(\mathrm{e}^{it\xi}\right) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\,\mathrm{d}F_{\xi}(x)$
  - (唯一性定理) 分布函数由特征函数唯一确定.
  - 若特征函数 $f(t)$ 绝对可积，则相应分布函数 $F(x)$ 的导数存在并连续，并且 $F'(x) = p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx} f(t)\,\mathrm{d}t$.

$$\Longrightarrow \quad f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^p}\int \exp\left(-\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{x}\right)\cdot\exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right)\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|2\pi\Sigma^{-1}\right|^{1/2}\cdot\left|2\pi\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$\cdot\int\exp\left\{-\frac{1}{2}\left[\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t} + 2\mathbf{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|\sqrt{2\pi}\Sigma^{-1/2}\cdot\sqrt{2\pi}\Sigma^{1/2}\right|}\int\exp\left\{-\frac{1}{2}\left[\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t} + 2\mathbf{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}\mathrm{d}\boldsymbol{t}$$

多元正态分布

- 分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的几何特征

定理 4.8 多元正态分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的特征函数为 $\varphi_X(\boldsymbol{t}) = \exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right).$

- 特征函数和分布函数是相互唯一确定的.
  - 分布函数可以确定特征函数：$f_\xi(t) \triangleq E\left(\mathrm{e}^{it\xi}\right) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\,\mathrm{d}F_\xi(x)$
  - (唯一性定理) 分布函数由特征函数唯一确定.
  - 若特征函数 $f(t)$ 绝对可积，则相应分布函数 $F(x)$ 的导数存在并连续，并且 $F'(x) = p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx} f(t)\,\mathrm{d}t$.

$$\Longrightarrow\quad f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^p}\int \exp\left(-\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{x}\right)\cdot\exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right)\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|2\pi\Sigma^{-1}\right|^{1/2}\cdot\left|2\pi\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$\cdot\int\exp\left\{-\frac{1}{2}\left[\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t} + 2\mathbf{i}\boldsymbol{t}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|2\pi\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$\cdot\int\frac{1}{\left|2\pi\Sigma^{-1}\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{t}+\mathbf{i}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{\mathrm{T}}\Sigma\left(\boldsymbol{t}+\mathbf{i}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right]\right\}d\boldsymbol{t}$$

多元正态分布

- 分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的几何特征

定理 4.8 多元正态分布 $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ 的特征函数为 $\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right)$.

- 特征函数和分布函数是相互唯一确定的.
  - 分布函数可以确定特征函数：$f_{\xi}(t) \triangleq E\left(\mathrm{e}^{it\xi}\right) = \int_{-\infty}^{+\infty} \mathrm{e}^{itx}\,\mathrm{d}F_{\xi}(x)$
  - (唯一性定理) 分布函数由特征函数唯一确定.
  - 若特征函数 $f(t)$ 绝对可积，则相应分布函数 $F(x)$ 的导数存在并连续，并且 $F'(x) = p(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-itx} f(t)\,\mathrm{d}t$.

$$\Longrightarrow f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^p}\int \exp\left(-\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{x}\right)\cdot\exp\left(\mathbf{i}\boldsymbol{t}^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^{\mathrm{T}}\Sigma\boldsymbol{t}\right)\mathrm{d}\boldsymbol{t}$$

$$= \frac{1}{\left|2\pi\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$= \frac{1}{\left|2\pi\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]\right\}$$

$$\cdot\int \frac{1}{\left|2\pi\Sigma^{-1}\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{t}+\mathbf{i}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)^{\mathrm{T}}\Sigma\left(\boldsymbol{t}+\mathbf{i}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\right]\right\}\mathrm{d}\boldsymbol{t}$$

$= 1$

$N_p\left(-\mathbf{i}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}), \Sigma^{-1}\right)$ 的概率密度函数

多元正态分布

- 奇异正态分布
  - 与 k 维多元正态分布的联系： $\boldsymbol{Y} \sim N_k\left(\mathbf{0}, \Lambda_1\right)$ ， $\Lambda_1=\operatorname{diag}\left(\lambda_1, \ldots, \lambda_k\right)$

    存在满足 $\mathscr{B}^{\mathrm{T}} \mathscr{B}=\mathscr{I}_k$ 的正交矩阵 $\mathscr{B}_{p \times k}$，使得 $\boldsymbol{X}=\mathscr{B} \boldsymbol{Y}+\boldsymbol{\mu}$ 具有如下形式的奇异密度函数

$$f_X(\boldsymbol{x})=\frac{(2 \pi)^{-k / 2}}{\left(\lambda_1 \cdots \lambda_k\right)^{1 / 2}} \exp \left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-}(\boldsymbol{x}-\boldsymbol{\mu})\right\} d t$$

抽样分布与极限定理

- 总体：随机向量 $\boldsymbol{X}=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$.

数据结构：

$$\mathcal{X}=\begin{array}{c} \begin{matrix} X_1 & X_2 & \cdots & X_p \end{matrix} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \end{array}$$

- 随机样本：$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \overset{\text{i.i.d.}}{\sim} \boldsymbol{X}$.

- 统计推断：通过样本数据 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ 分析总体变量 $\boldsymbol{X}$ 的性质.

样本均值 $\bar{\boldsymbol{x}}$，样本协方差矩阵 $\mathcal{S}$ 等

均值向量 $\boldsymbol{\mu}$，协方差矩阵 Σ 等

抽样分布与极限定理

- 例: 总体 $\boldsymbol{X} \in \mathbb{R}^p$ 且 $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$. $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 是其独立同分布的一个样本.

样本观测值为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$$

样本均值:

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n \rightarrow \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$

$$\mathrm{E}(\bar{\boldsymbol{x}}) = \begin{pmatrix} \mathrm{E}(\bar{x}_1) \\ \mathrm{E}(\bar{x}_2) \\ \vdots \\ \mathrm{E}(\bar{x}_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

$$\mathrm{E}(\bar{\boldsymbol{x}}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\boldsymbol{x}_i) = \boldsymbol{\mu}$$

$\bar{\boldsymbol{x}}$ 是 $\boldsymbol{\mu}$ 的无偏估计量

$$\mathrm{Var}(\bar{\boldsymbol{x}}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(\boldsymbol{x}_i) = \frac{1}{n}\Sigma = \mathrm{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$$

抽样分布与极限定理

- 例: 总体 $\boldsymbol{X} \in \mathbb{R}^p$ 且 $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$. $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 是其独立同分布的一个样本.

样本观测值为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$$

样本协方差矩阵: $\mathcal{S} = \dfrac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}}$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

$$= \frac{1}{n}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X} - \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathcal{X}\right)$$

$$\mathrm{E}(\mathcal{S}) = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}}\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right) - \mathrm{E}\left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{Var}\left(\boldsymbol{x}_i\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right] - \left(\frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) - \left(\frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) = \frac{n-1}{n}\Sigma$$

$$\mathrm{Var}\left(\bar{\boldsymbol{x}}\right) = \frac{1}{n}\Sigma = \mathrm{E}\left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right) - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$$

$$\Longrightarrow \quad \mathrm{E}\left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right) = \frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$$

抽样分布与极限定理

- 例: 总体 $\boldsymbol{X} \in \mathbb{R}^p$ 且 $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$. $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 是其独立同分布的一个样本.

样本观测值为 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$.

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$$

样本协方差矩阵: $\mathcal{S} = \dfrac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}}$

$$= \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathcal{X} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$$

$$= \frac{1}{n}\left(\mathcal{X}^{\mathrm{T}}\mathcal{X} - \frac{1}{n}\mathcal{X}^{\mathrm{T}}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\mathcal{X}\right)$$

$$\mathrm{E}(\mathcal{S}) = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}}\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right) - \mathrm{E}\left(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{Var}\left(\boldsymbol{x}_i\right) + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right] - \left(\frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) - \left(\frac{1}{n}\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\right) = \frac{n-1}{n}\Sigma \neq \Sigma$$

$\mathcal{S}$ 是 Σ 的有偏估计量

$\Longrightarrow$ $\mathcal{S}_u = \dfrac{n}{n-1}\mathcal{S}$ 是 Σ 的无偏估计量

抽样分布与极限定理

定理 4.9 设 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 独立同分布，且 $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$. 则 $\bar{\boldsymbol{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right)$.

$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ 是相互独立的正态分布随机变量的一个线性组合.

第 5 章我们会证明它也服从正态分布.

抽样分布与极限定理

- 大数定律：概率论的核心定理，描述 $\frac{1}{n}\sum_{i=1}^{n} X_i$ 在样本容量趋于无穷时的收敛现象.

① Bernoulli 大数定律：$X_i \overset{\text{i.i.d.}}{\sim} b(1,p) \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} p \quad (n \to \infty).$

② Chebyshev 弱大数定律：X_i 两两不相关且 $\mathbb{E}\left(X_i\right)=\mu$，$\mathbb{D}\left(X_i\right) \le C$ (常数)，则

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu \quad (n \to \infty).$$

③ Khinchin 弱大数定律：X_i 独立同分布且 $\mathbb{E}\left(X_i\right)=\mu$，则 $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu \quad (n \to \infty).$

④ Kolmogorov 强大数定律：X_i 独立同分布且 $\mathbb{E}\left(X_i\right)=\mu$，则

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu \quad (n \to \infty).$$

抽样分布与极限定理

- 中心极限定理：概率论的核心定理，描述 $\sum_{i=1}^{n} X_i$ 的分布趋近于正态分布的规律.

① De Moivre-Laplace 定理：$X \sim b(n,p) \Longrightarrow \dfrac{X-np}{\sqrt{np(1-p)}} \xrightarrow{\mathscr{L}} N(0,1) \quad (n \to \infty).$

② 设 $X_1, X_2, \ldots, X_n, \ldots$ 独立同分布，且 $\mathbb{E}\left(X_i\right)=\mu, \mathbb{D}\left(X_i\right)=\sigma^2<\infty$，则

$$\frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)-\mu}{\frac{\sigma}{\sqrt{n}}}=\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathscr{L}} N(0,1), \qquad (n \to \infty).$$

- Slutsky 引理：设 $\{X_n\}$ 和 $\{Y_n\}$ 是两个随机变量序列，且当 $n \to \infty$ 时 $X_n \xrightarrow{\mathscr{L}} X$，$Y_n \xrightarrow{p} c$，其中 $-\infty < c < \infty$ 为常数，则

① $X_n \pm Y_n \xrightarrow{\mathscr{L}} X \pm c.$

② $X_n Y_n \xrightarrow{\mathscr{L}} cX.$

③ $\dfrac{X_n}{Y_n} \xrightarrow{\mathscr{L}} \dfrac{X}{c} \quad (c \neq 0).$

抽样分布与极限定理

定理 4.10 (中心极限定理) 设 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 独立同分布且 $\boldsymbol{X}_i \sim (\boldsymbol{\mu}, \Sigma)$. 则 $\sqrt{n}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)$ 渐近服从正态分布 $N_p(\mathbf{0}, \Sigma)$，即

依分布收敛

$$\sqrt{n}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \xrightarrow{\mathscr{L}} N_p(\mathbf{0}, \Sigma) \quad \text{as } n \to \infty .$$

- 例：$X_1, X_2, \ldots, X_n$ 独立同分布于 $p = \dfrac{1}{2}$ 的 Bernoulli 分布，则 $\mu = p = \dfrac{1}{2}$,

$\Sigma = p(1-p) = \dfrac{1}{4}$，于是

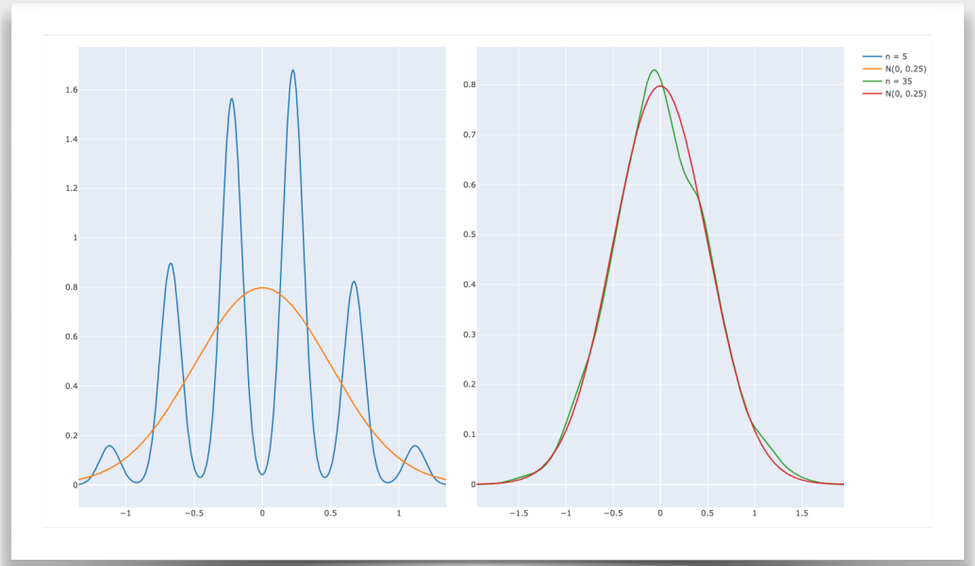
$$\sqrt{n}\left(\bar{x} - \frac{1}{2}\right) \xrightarrow{\mathscr{L}} N_1\left(0, \frac{1}{4}\right) \quad \text{as } n \to \infty$$

抽样分布与极限定理

```
x = array(0, dim = 1000)
for (i in 1:1000){
  y = rbinom(5, 1, prob = 1/2)
  x[i] = sqrt(5) * (mean(y) - 1/2)
}
fig8 = plot_ly()
d = density(x, kernel = 'gaussian')
fig8 = add_lines(fig8, x = d$x, y = d$y, name = "n = 5")
fig8 = add_lines(fig8, x = d$x, y = dnorm(d$x, 0, 0.5), name = "N(0, 0.25)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = '#ffff', zerolinewidth = 2, gridcolor = 'ffff'),
         yaxis = list(zerolinecolor = '#ffff', zerolinewidth = 2, gridcolor = 'ffff'))

x <- array(0, dim = 1000)
for (i in 1:1000){
  y = rbinom(35, 1, prob = 1/2)
  x[i] = sqrt(35) * (mean(y) - 1/2)
}
fig9 = plot_ly()
d = density(x, kernel = 'gaussian')
fig9 = add_lines(fig9, x = d$x, y = d$y, name = "n = 35")
fig9 = add_lines(fig9, x = d$x, y = dnorm(d$x, 0, 0.5), name = "N(0, 0.25)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = '#ffff', zerolinewidth = 2, gridcolor = 'ffff'),
         yaxis = list(zerolinecolor = '#ffff', zerolinewidth = 2, gridcolor = 'ffff'))

subplot(fig8, fig9)
```



抽样分布与极限定理

- 例：设 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 是一个二维随机样本，来自参数均为 $p = 0.5$ 的两个独立的 Bernoulli 分布构成的总体.

  - $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ 的联合分布为

X_1 \ X_2	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

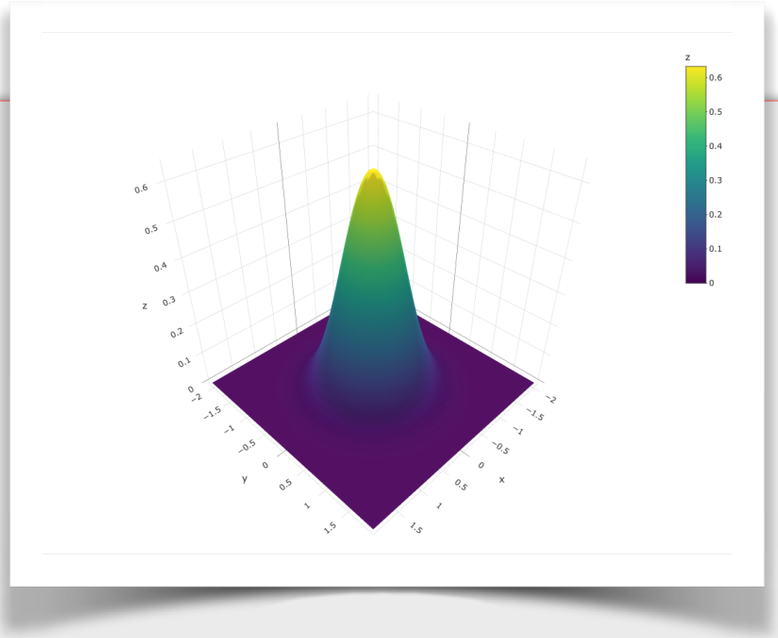
$$\Longrightarrow \quad \boldsymbol{\mu} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$$

$$\sqrt{n}\left(\bar{x} - \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}\right) \xrightarrow{\mathscr{L}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{pmatrix}\right), \qquad n \to \infty$$

抽样分布与极限定理

- 例：设 $X_1, X_2, \ldots, X_n$ 是一个二维随机样本，来自参数均为 $p = 0.5$ 的两个独立的 Bernoulli 分布构成的总体.

```
n = 50
mu1 = 0
mu2 = 0
s1 = 1/2
s2 = 1/2
rho = 0
x = seq(-4, 4, length = n) * s1
y = seq(-4, 4, length = n) * s2
f = function(x, y){
  (2 * pi * s1 * s2 * sqrt(1-rho^2))^-1 * exp(-0.5 * (1 - rho^2)^-1 *
      ((x-mu1)^2/s1^2 - 2 * rho * (x - mu1) * ( y - mu2) / (s1 * s2) + (y - mu2)^2/s2^2))
}
z = outer(x , y, f)
fig10 = plot_ly(x = ~x, y = ~y, z = ~z) %>%
  add_surface()
fig10
```



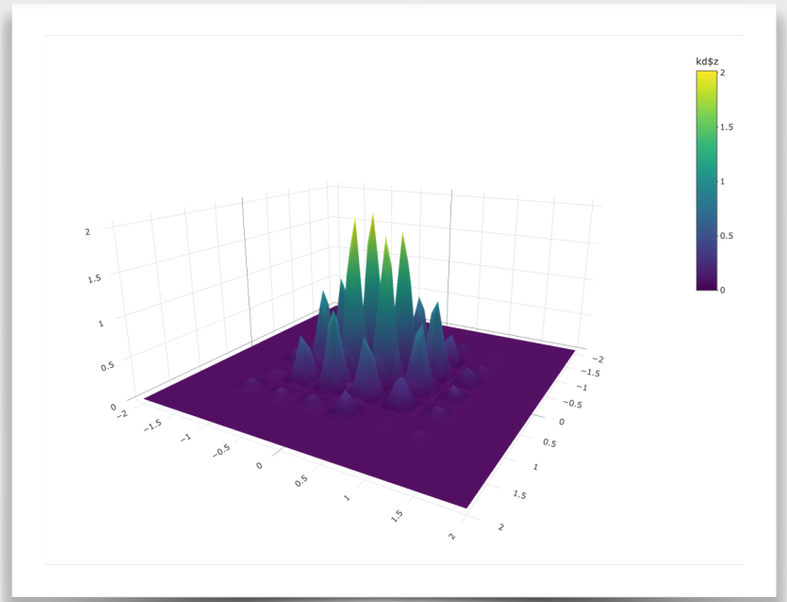
$N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\frac{1}{4} & 0\\ 0 & \frac{1}{4}\end{pmatrix}\right)$ 的联合密度

抽样分布与极限定理

- 例：设 $X_1, X_2, \ldots, X_n$ 是一个二维随机样本，来自参数均为 $p = 0.5$ 的两个独立的 Bernoulli 分布构成的总体.

样本容量 $n = 5$ 时的渐近分布

```
rm(list = ls(all = TRUE))
library(MASS)
x = array(0, dim = 1000)
y = array(0, dim = 1000)
for (i in 1:1000){
  a = rbinom(5, 1, prob = 1/2)
  x[i] = sqrt(5) * (mean(a) - 1/2)
  b = rbinom(5, 1, prob = 1/2)
  y[i] = sqrt(5) * (mean(b) - 1/2)
}
kd = kde2d(x, y, n=50, lims = c(-2, 2, -2, 2))
fig11 = plot_ly(x = ~kd$x, y = ~kd$y, z = ~kd$z) %>%
  add_surface() %>%
  layout(
    scene = list(
      xaxis = list(title = ""),
      yaxis = list(title = ""),
      zaxis = list(title = "")
    )
  )
fig11
```

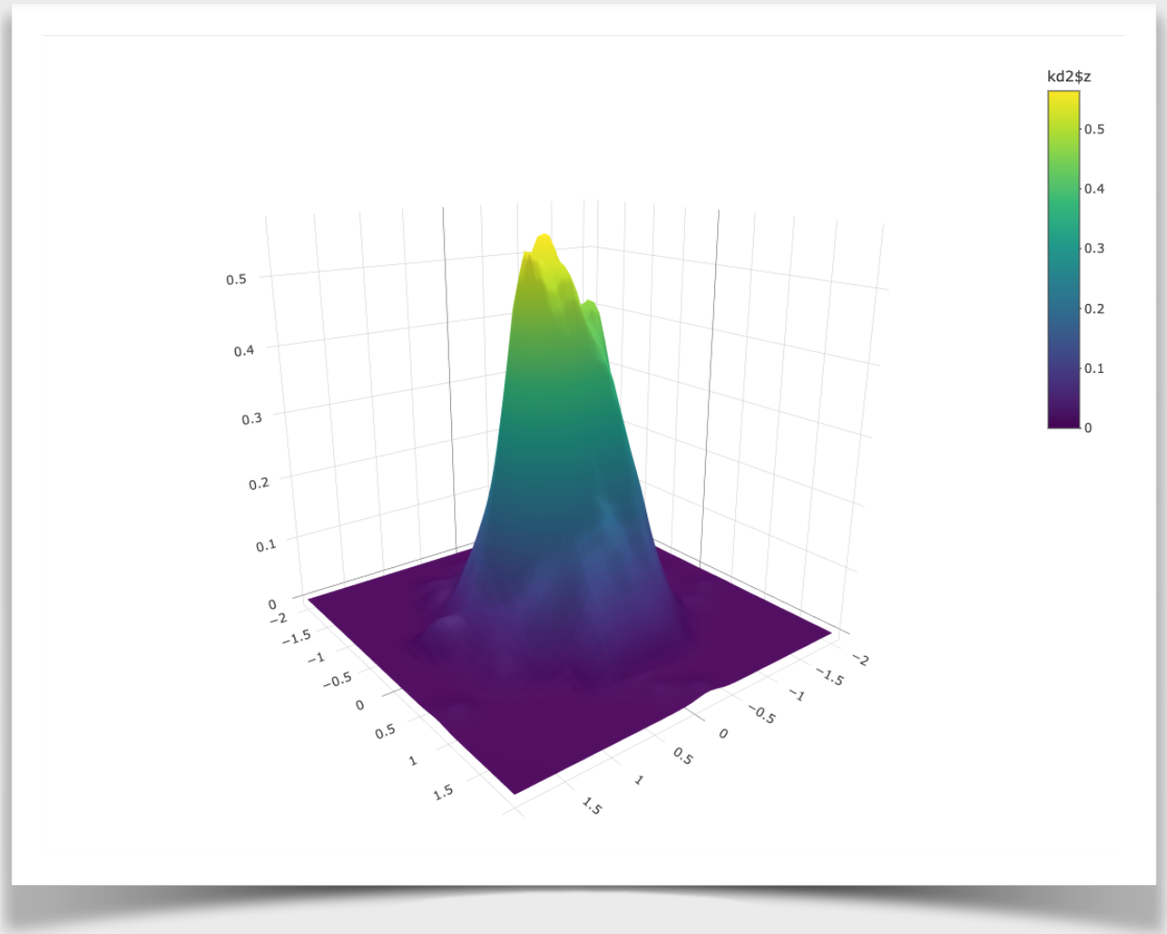


抽样分布与极限定理

- 例：设 $X_1, X_2, \ldots, X_n$ 是一个二维随机样本，来自参数均为 $p = 0.5$ 的两个独立的 Bernoulli 分布构成的总体.

样本容量 $n = 85$ 时的渐近分布

```
x = array(0, dim = 1000)
y = array(0, dim = 1000)
for (i in 1:1000){
  a = rbinom(85, 1, prob = 1/2)
  x[i] = sqrt(85) * (mean(a) - 1/2)
  b = rbinom(85, 1, prob = 1/2)
  y[i] = sqrt(85) * (mean(b) - 1/2)
}
kd2 =kde2d(x, y, n=50, lims = c(-2, 2, -2, 2))
fig12 = plot_ly(x = ~kd2$x, y = ~kd2$y, z = ~kd2$z) %>%
  add_surface() %>%
  layout(
    scene = list(
      xaxis = list(title = ""),
      yaxis = list(title = ""),
      zaxis = list(title = "")
    )
  )
fig12
```



抽样分布与极限定理

- 渐近正态分布常用于构造未知参数的置信区间.
  - 置信度为 $1-\alpha$ 的置信区间，$\alpha \in (0\ ,\ 1)$：

$$P\left(\theta \in \left[\widehat{\theta_l}\ ,\ \widehat{\theta_u}\right]\right) = 1-\alpha$$

- 例：考虑独立同分布的随机变量 $X_1, X_2, \ldots, X_n$，其中 $X_i \sim \left(\mu, \sigma^2\right)$，$\sigma^2$ 已知.

$$\sqrt{n}\left(\bar{x}-\mu\right) \xrightarrow{\mathscr{L}} N\left(0, \sigma^2\right) \quad \Longrightarrow \quad \sqrt{n}\cdot\frac{\bar{x}-\mu}{\sigma} \xrightarrow{\mathscr{L}} N(0, 1)$$

$$\Longrightarrow \quad P\left(-u_{\frac{\alpha}{2}} \le \sqrt{n}\cdot\frac{\bar{x}-\mu}{\sigma} \le u_{\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha\ , \qquad n\to\infty$$

上侧 $\frac{\alpha}{2}$ 分位数

$$\Longrightarrow \quad P\left(\bar{x}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}} \le \mu \le \bar{x}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right) \longrightarrow 1-\alpha\ , \qquad n\to\infty$$

$$\Longrightarrow \quad \left[\bar{x}-\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\ ,\ \bar{x}+\frac{\sigma}{\sqrt{n}}u_{\frac{\alpha}{2}}\right] \qquad \mu \text{ 的置信度为 } 1-\alpha \text{ 的近似置信区间}$$

抽样分布与极限定理

推论 4.1 如果 $\widehat{\Sigma}$ 是 Σ 的一致估计，则中心极限定理成立，即

$$\sqrt{n}\,\widehat{\Sigma}^{-1/2}\left(\bar{\boldsymbol{x}}-\boldsymbol{\mu}\right)\xrightarrow{\mathscr{L}}N_p\left(\mathbf{0},\,\mathscr{I}\right)\quad\text{as } n\to\infty.$$

- 例：考虑独立同分布的随机变量 $X_1, X_2, \ldots, X_n$，其中 $X_i\sim\left(\mu,\,\sigma^2\right)$，$\sigma^2$ 未知.

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\quad\Longrightarrow\quad\frac{\sqrt{n}}{\hat{\sigma}}\left(\bar{x}-\mu\right)\xrightarrow{\mathscr{L}}N(0,\,1)\qquad as\ n\longrightarrow\infty$$

$$\Longrightarrow\quad P\left(-u_{\frac{\alpha}{2}}\le\frac{\sqrt{n}}{\hat{\sigma}}\left(\bar{x}-\mu\right)\le u_{\frac{\alpha}{2}}\right)\longrightarrow 1-\alpha\,,\qquad n\to\infty$$

$$\Longrightarrow\quad P\left(\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}u_{\frac{\alpha}{2}}\le\mu\ \le\bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}u_{\frac{\alpha}{2}}\right)\longrightarrow 1-\alpha\,,\qquad n\to\infty$$

$$\Longrightarrow\quad\left[\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}u_{\frac{\alpha}{2}}\,,\,\bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}u_{\frac{\alpha}{2}}\right]\qquad\mu\text{ 的置信度为 }1-\alpha\text{ 的近似置信区间}$$

抽样分布与极限定理

- 在实际应用中，n 取多大才能得到合理的近似呢?
  - 无明确答案：取决于具体问题 (X_i 分布的形状，X_i 的维数).
  - 如果 X_i 服从正态分布，则 $n = 1$ 开始的所有 $\bar{x}$ 均服从正态分布.
  - 多数情形下，一维问题中当 $n > 50$ 即可得到有效的近似.

抽样分布与极限定理

- 统计量的变换

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\mathscr{X}^{\mathrm{T}}\mathbf{1}_n$$

$$\mathscr{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}$$

  - 给定矩阵 $\mathscr{A} > 0$，定义 $f(\boldsymbol{\mu}) = \boldsymbol{\mu}^{\mathrm{T}}\mathscr{A}\boldsymbol{\mu}$，统计量 $f(\bar{\boldsymbol{x}})$ 服从什么分布?

定理 4.11 设 $\sqrt{n}(\boldsymbol{t}-\boldsymbol{\mu}) \xrightarrow{\mathscr{L}} N_p(\mathbf{0}, \Sigma)$，$f = \left(f_1, f_2, \ldots f_q\right)^{\mathrm{T}} : \mathbb{R}^p \to \mathbb{R}^q$ 为实值函数，且在 $\boldsymbol{\mu} \in \mathbb{R}^p$ 可微，则 $f(\boldsymbol{t})$ 渐近服从均值向量为 $f(\boldsymbol{\mu})$、协方差矩阵为 $\mathscr{D}^{\mathrm{T}}\Sigma\mathscr{D}$ 的正态分布，即

$$\sqrt{n}\left[f(\boldsymbol{t}) - f(\boldsymbol{\mu})\right] \xrightarrow{\mathscr{L}} N_q\left(\mathbf{0}, \mathscr{D}^{\mathrm{T}}\Sigma\mathscr{D}\right), \qquad n \longrightarrow \infty$$

其中

$$\mathscr{D} = \left(\frac{\partial f_j}{\partial t_i}\right)(\boldsymbol{t})\Bigg|_{\boldsymbol{t}=\boldsymbol{\mu}}$$

是所有偏导数的 $(p \times q)$ 矩阵.

抽样分布与极限定理

- 例：对二次型 $f(\boldsymbol{\mu}) = \boldsymbol{\mu}^{\mathrm{T}}\mathscr{A}\boldsymbol{\mu}$，其中 $\mathscr{A} > 0$.

$$f(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}}^{\mathrm{T}}\mathscr{A}\bar{\boldsymbol{x}} \quad \Longrightarrow \quad \frac{\partial f(\bar{\boldsymbol{x}})}{\partial \bar{\boldsymbol{x}}} = 2\mathscr{A}\bar{\boldsymbol{x}}$$

$$\Longrightarrow \quad \mathscr{D} = \left(\frac{\partial f(\bar{\boldsymbol{x}})}{\partial \bar{\boldsymbol{x}}}\right)\Bigg|_{\bar{\boldsymbol{x}}=\boldsymbol{\mu}} = 2\mathscr{A}\boldsymbol{\mu}$$

定理 4.11 $$\Longrightarrow \quad \sqrt{n}\left(\bar{\boldsymbol{x}}^{\mathrm{T}}\mathscr{A}\bar{\boldsymbol{x}} - \boldsymbol{\mu}^{\mathrm{T}}\mathscr{A}\boldsymbol{\mu}\right) \xrightarrow{\mathscr{L}} N_1\left(0,\ 4\boldsymbol{\mu}^{\mathrm{T}}\mathscr{A}^{\mathrm{T}}\Sigma\mathscr{A}\boldsymbol{\mu}\right)$$

抽样分布与极限定理

- 例：设 $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$；$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$，$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$，$p = 2$.

$$\text{CLT (定理 4.10)} \implies \sqrt{n}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \xrightarrow{\mathscr{L}} N\left(\mathbf{0}, \Sigma\right), \; n \to \infty$$

  - 如果我们想确定 $\begin{pmatrix} \bar{x}_1^2 - \bar{x}_2 \\ \bar{x}_1 + 3\bar{x}_2 \end{pmatrix}$ 的分布.

$$f = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} x_1^2 - x_2 \\ x_1 + 3x_2 \end{pmatrix}$$

$$\mathscr{D} = \left(\frac{\partial f_j}{\partial x_i}\right)(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} \end{pmatrix}\Bigg|_{\boldsymbol{x}=\boldsymbol{\mu}} = \begin{pmatrix} 2x_1 & 1 \\ -1 & 3 \end{pmatrix}\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = \begin{pmatrix} 0 & 1 \\ -1 & 3 \end{pmatrix}$$

$$\text{定理 4.11} \implies \sqrt{n}\begin{pmatrix} \bar{x}_1^2 - \bar{x}_2 \\ \bar{x}_1 + 3\bar{x}_2 \end{pmatrix} \xrightarrow{\mathscr{L}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\frac{7}{2} \\ -\frac{7}{2} & 13 \end{pmatrix}\right)$$

$\mathscr{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathscr{D}$

抽样分布与极限定理

- 例：设 $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$；$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$，$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$，$p = 2$.

$$\text{CLT (定理 4.10)} \implies \sqrt{n}\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \xrightarrow{\mathscr{L}} N\left(\mathbf{0}, \Sigma\right), \quad n \to \infty$$

  - 对更为复杂的一个函数 $f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} x_1^2 - x_2 \\ x_1 + 3x_2 \\ x_2^3 \end{pmatrix}$，$q = 3 > 2 = p$.

$$\mathscr{D} = \left(\frac{\partial f_j}{\partial x_i}\right)(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \\ \frac{\partial f_1}{\partial x_3} & \frac{\partial f_2}{\partial x_3} & \frac{\partial f_3}{\partial x_3} \end{pmatrix}\Bigg|_{\boldsymbol{x}=\boldsymbol{\mu}} = \begin{pmatrix} 2x_1 & 1 & 0 \\ -1 & 3 & 3x_2^2 \end{pmatrix}\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 3 & 0 \end{pmatrix}$$

$\mathscr{D}^{\mathrm{T}}\Sigma\mathscr{D}$

$$\text{定理 4.11} \implies \sqrt{n}\begin{pmatrix} x_1^2 - x_2 \\ x_1 + 3x_2 \\ x_2^3 \end{pmatrix} \xrightarrow{\mathscr{L}} N_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -\frac{7}{2} & 0 \\ -\frac{7}{2} & 13 & 0 \\ 0 & 0 & 0 \end{pmatrix}\right)$$

奇异正态分布！

厚尾分布 (heavy-tailed distributions)

- 位置参数为 l、尺度参数为 s 的 Cauchy 分布 $\mathrm{C}(l,\ s)$ 的概率密度函数为

$$f(x)=\frac{1}{\pi s}\cdot\frac{1}{1+\left(\frac{x-l}{s}\right)^2},\quad -\infty<x<\infty$$

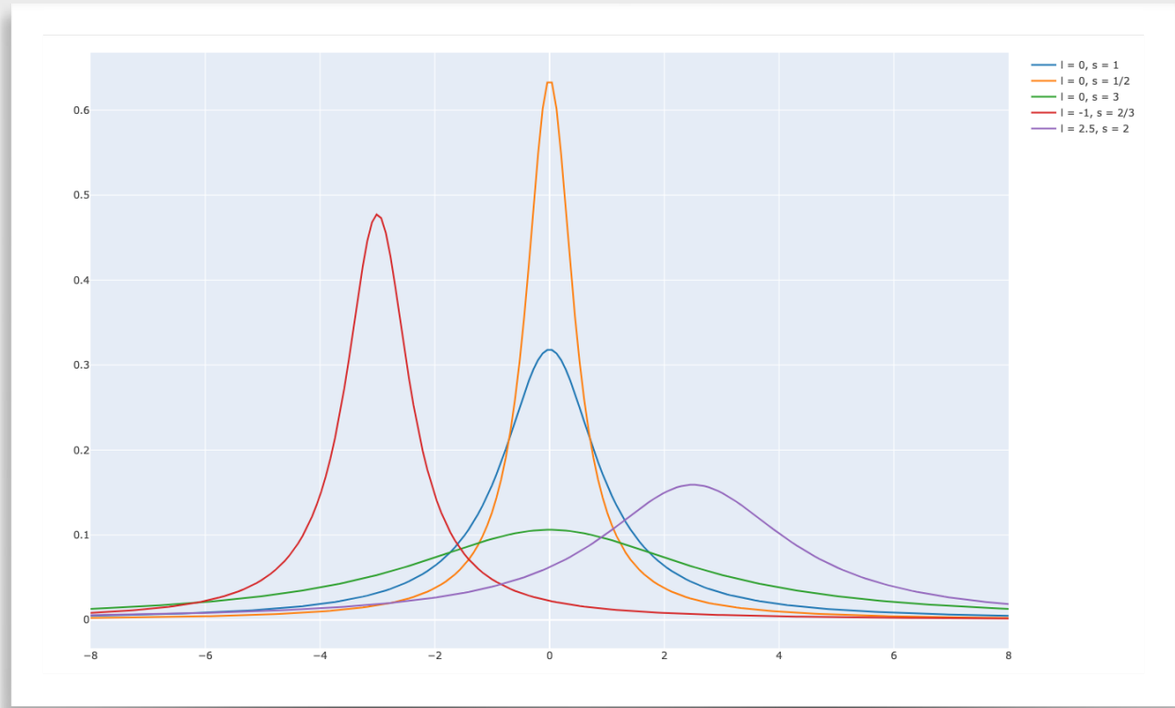
  - 若 $X\sim\mathrm{C}(l,\ s)$，则其均值 $\mathrm{E}(X)$，方差 $\mathrm{Var}(X)$ 以及更高阶矩均不存在.
  - 若 $X_i\overset{\text{i.i.d.}}{\sim}\mathrm{C}(l,\ s)$，则 $\sum_{i=1}^{n}X_i\sim\mathrm{C}(nl,\ ns)$.
  - 若 $X\sim\mathrm{C}(l,\ s)$，则 $\frac{1}{X}\sim\mathrm{C}\left(\frac{l}{l^2+s^2},\ \frac{s}{l^2+s^2}\right)$.
  - 若 $X_1,\ X_2\overset{\text{i.i.d.}}{\sim}N(0,\ 1)$，则 $\frac{X_1}{X_2}\sim\mathrm{C}\,(1\ ,\ 0)$.
  - 若 $X\sim U\left(-\frac{\pi}{2},\ \frac{\pi}{2}\right)$，则 $\tan X\sim\mathrm{C}\,(1\ ,\ 0)$.
  - 若 $X\sim\mathrm{C}\,(1\ ,\ 0)$，则 $X\sim t_1$.

厚尾分布 (heavy-tailed distributions)

- 位置参数为 l、尺度参数为 s 的 Cauchy 分布 $C(l, s)$ 的概率密度函数为

$$f(x) = \frac{1}{\pi s} \cdot \frac{1}{1 + \left(\frac{x - l}{s}\right)^2}, \quad -\infty < x < \infty$$

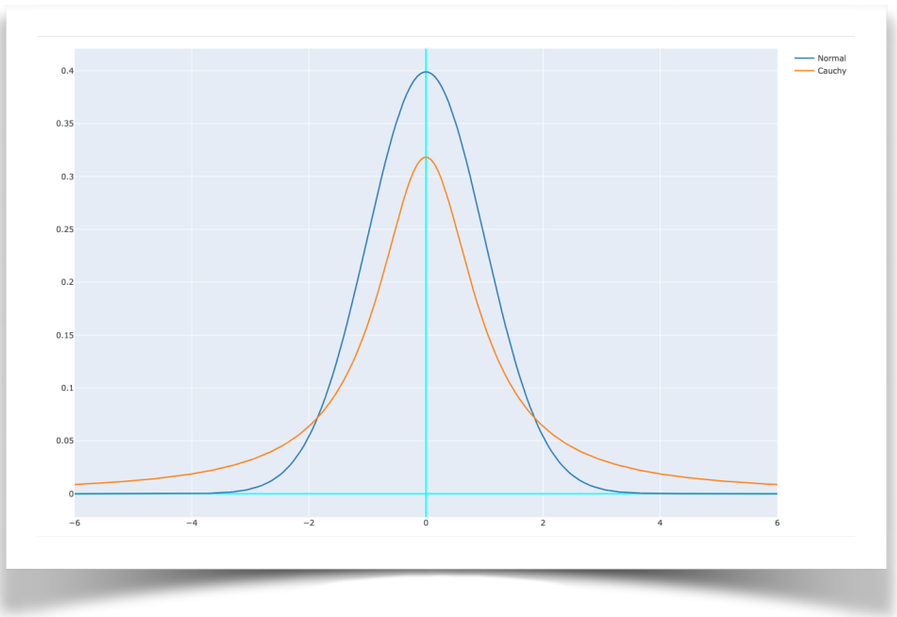
```
rm(list = ls(all = TRUE))
x = seq(-8, 8, length = 200)
y_1 = dcauchy(x, location = 0, scale = 1)
y_2 = dcauchy(x, location = 0, scale = 1/2)
y_3 = dcauchy(x, location = 0, scale = 3)
y_4 = dcauchy(x, location = -3, scale = 2/3)
y_5 = dcauchy(x, location = 2.5, scale = 2)
fig_13 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "l = 0, s = 1") %>%
  add_lines(x = ~x, y = ~y_2, name = "l = 0, s = 1/2") %>%
  add_lines(x = ~x, y = ~y_3, name = "l = 0, s = 3") %>%
  add_lines(x = ~x, y = ~y_4, name = "l = -1, s = 2/3") %>%
  add_lines(x = ~x, y = ~y_5, name = "l = 2.5, s = 2")
fig_13 = fig_13 %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = '#ffff', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = '#ffff', zerolinewidth = 2, gridcolor = 'ffff', title = ""))
fig_13
```



厚尾分布 (heavy-tailed distributions)

- 一个分布称为**重尾** (heavy-tailed) 是指其尾部的面积相比相同均值 μ 和方差 σ^2 的正态分布具有更高的概率密度.

```
x = seq(-6, 6, length = 200)
y_1 = dnorm(x, mean = 0, sd = 1)
y_2 = dcauchy(x, location = 0, scale = 1)
fig_14 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "Normal") %>%
  add_lines(x = ~x, y = ~y_2, name = "Cauchy")
fig_14 = fig_14 %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = ""))
fig_14
```



厚尾分布 (heavy-tailed distributions)

- 一个分布称为**厚尾** (heavy-tailed) 是指其尾部的面积相比相同均值 μ 和方差 σ^2 的正态分布具有更高的概率密度.
  - 峰度 (kurtosis)：$\gamma_4 = \dfrac{\mathrm{E}\left[(X-\mu)^4\right]}{\sigma^4}$.
  - 厚尾分布的峰度大于 3.

$$\begin{cases} \text{尖峰 (leptokurtic) 分布:} & \gamma_4 > 3 \\ \text{常峰 (mesokurtic) 分布:} & \gamma_4 = 3 \\ \text{低峰 (platykurtic) 分布:} & \gamma_4 < 3 \end{cases}$$

- 由于单变量厚尾分布是其多变量对应分布的基础，并且其密度特性已被证明即使在多元情形下也很有用，因此我们从引入一些单变量的厚尾分布开始. 然后我们再分析它们对应的多元分布及其尾部的行为.

厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)
  - 该分布由 Barndorff-Nielsen 教授引入，最早应用于模拟风吹砂的粒度分布.
  - 如今，它当前最重要的应用之一是用于股票价格的建模和市场风险的度量.
  - 该分布的名称来源于它的对数密度曲线构成了一条双曲线 (hyperbola)，而正态分布的对数密度曲线是一条抛物线 (parabola).

March 18, 1935 ~ June 26, 2022

厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)
  - 对 $x \in \mathbb{R}$，一维广义双曲 (GH) 分布的概率密度函数为

$$f_{\mathrm{GH}}\left(x;\ \lambda,\ \alpha,\ \beta,\ \delta,\ \mu\right) = \frac{\left(\dfrac{\sqrt{\alpha^2-\beta^2}}{\delta}\right)^{\lambda}}{\sqrt{2\pi}\ K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} \cdot \frac{K_{\lambda-1/2}\left[\alpha\sqrt{\delta^2+(x-\mu)^2}\right]}{\left(\dfrac{\sqrt{\delta^2+(x-\mu)^2}}{\alpha}\right)^{1/2-\lambda}} \cdot \mathrm{e}^{\beta(x-\mu)}$$

$$K_{\lambda}(x) = \frac{1}{2}\int_0^{\infty} y^{\lambda-1}\mathrm{e}^{-\frac{x}{2}\left(y+y^{-1}\right)}dy\ ,$$ 修正的指数为 λ 第三类贝塞尔 (Bessel) 函数

参数的取值范围：

$$\begin{cases} \mu \in \mathbb{R} & \\ \delta \geq 0\ , \quad \left|\beta\right| < \alpha\ , & \text{if } \lambda > 0 \\ \delta > 0\ , \quad \left|\beta\right| < \alpha\ , & \text{if } \lambda = 0 \\ \delta \geq 0\ , \quad \left|\beta\right| \leq \alpha\ , & \text{if } \lambda < 0 \end{cases}$$

厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)
  - 广义双曲分布的均值和方差如下:

$$\mathrm{E}(X) = \mu + \frac{\delta\,\beta}{\sqrt{\alpha^2-\beta^2}} \cdot \frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}$$

$$\mathrm{Var}(X) = \delta^2 \left\{ \frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{\delta\sqrt{\alpha^2-\beta^2}\;K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} + \frac{\beta^2}{\alpha^2-\beta^2}\left[\frac{K_{\lambda+2}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} - \left(\frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)} \right)^2 \right] \right\}$$

  - 其中的 μ 和 δ 分别在密度的位置参数和尺度参数中发挥着重要作用.

厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)
  - 对特定的 λ 值，可以得到广义双曲 (GH) 分布的不同类型.

    当 $\lambda = 1$ 时，得到双曲 (HYP) 分布

$$f_{\mathrm{HYP}}\left(x;\ \alpha, \beta, \delta, \mu\right) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1\left(\delta\sqrt{\alpha^2 - \beta^2}\right)} \mathrm{e}^{-\alpha\sqrt{\delta^2 + (x-\mu)^2} + \beta(x-\mu)}, \quad x, \mu \in \mathbb{R}, \delta \geq 0, \ \left|\beta\right| < \alpha$$

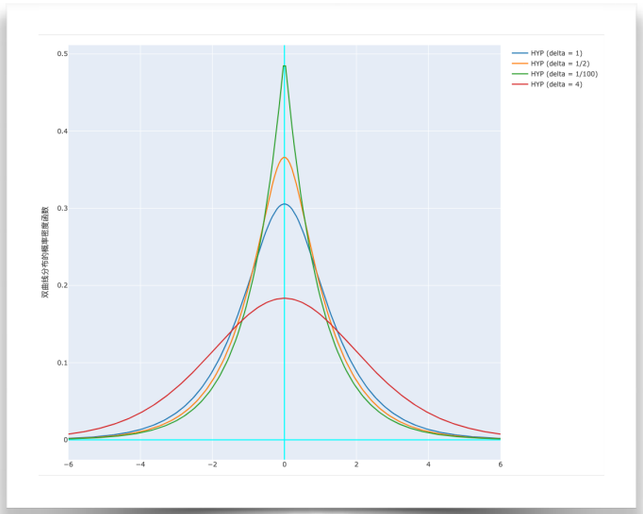
当 $\lambda = -\frac{1}{2}$ 时，得到正态逆高斯 (NIG) 分布

$$f_{\mathrm{NIG}}\left(x;\ \alpha, \beta, \delta, \mu\right) = \frac{\alpha\delta}{\pi} \frac{K_1\left(\alpha\sqrt{\delta^2 + (x-\mu)^2}\right)}{\sqrt{\delta^2 + (x-\mu)^2}} \mathrm{e}^{\delta\sqrt{\alpha^2 - \beta^2} + \beta(x-\mu)}$$

厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

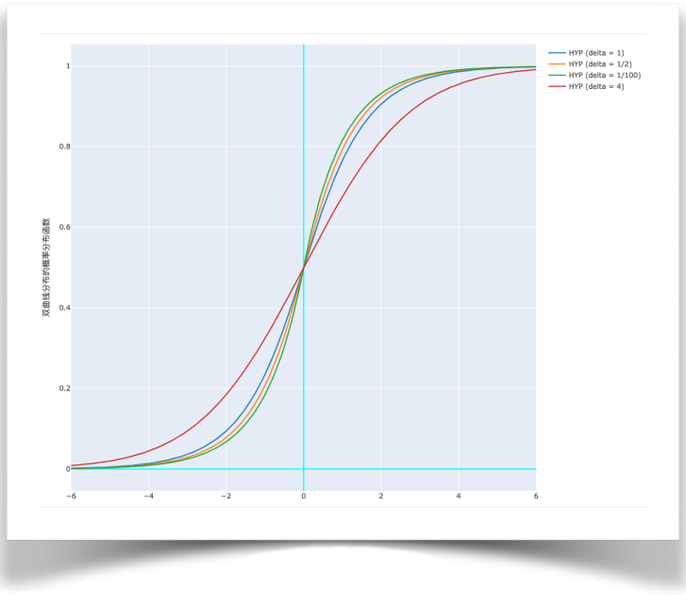
```
library(fBasics)
x = seq(-6, 6, length = 200)
y_1 = dhyp(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_2 = dhyp(x, alpha = 1, beta = 0, delta = 1/2, mu = 0)
y_3 = dhyp(x, alpha = 1, beta = 0, delta = 1/100, mu = 0)
y_4 = dhyp(x, alpha = 1, beta = 0, delta = 4, mu = 0)
fig_15 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "HYP (delta = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "HYP (delta = 1/2)") %>%
  add_lines(x = ~x, y = ~y_3, name = "HYP (delta = 1/100)") %>%
  add_lines(x = ~x, y = ~y_4, name = "HYP (delta = 4)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "双曲线分布的概率密度函数"))
fig_15 # 双曲线分布的概率密度曲线
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

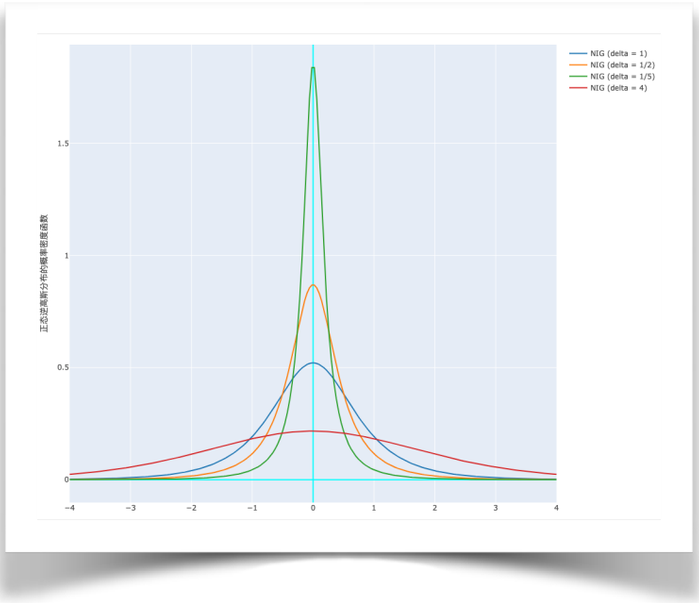
```
x = seq(-6, 6, length = 200)
y_1 = phyp(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_2 = phyp(x, alpha = 1, beta = 0, delta = 1/2, mu = 0)
y_3 = phyp(x, alpha = 1, beta = 0, delta = 1/100, mu = 0)
y_4 = phyp(x, alpha = 1, beta = 0, delta = 4, mu = 0)
fig_16 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "HYP (delta = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "HYP (delta = 1/2)") %>%
  add_lines(x = ~x, y = ~y_3, name = "HYP (delta = 1/100)") %>%
  add_lines(x = ~x, y = ~y_4, name = "HYP (delta = 4)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "双曲线分布的概率分布函数"))
fig_16 # 双曲线分布的分布函数曲线
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

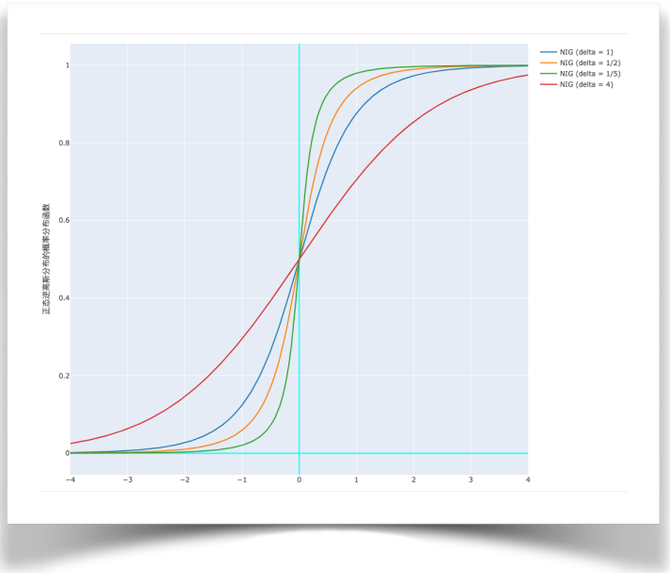
```
x = seq(-4, 4, length = 200)
y_1 = dnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_2 = dnig(x, alpha = 1, beta = 0, delta = 1/2, mu = 0)
y_3 = dnig(x, alpha = 1, beta = 0, delta = 1/5, mu = 0)
y_4 = dnig(x, alpha = 1, beta = 0, delta = 4, mu = 0)
fig_17 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "NIG (delta = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "NIG (delta = 1/2)") %>%
  add_lines(x = ~x, y = ~y_3, name = "NIG (delta = 1/5)") %>%
  add_lines(x = ~x, y = ~y_4, name = "NIG (delta = 4)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "正态逆高斯分布的概率密度函数"))
fig_17 # 正态逆高斯分布的概率密度曲线
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

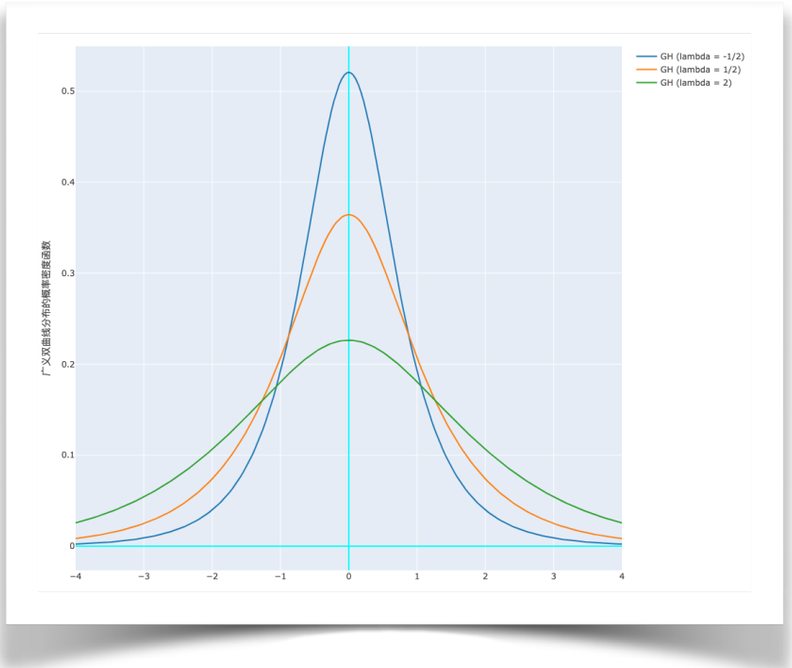
```
x = seq(-4, 4, length = 200)
y_1 = pnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_2 = pnig(x, alpha = 1, beta = 0, delta = 1/2, mu = 0)
y_3 = pnig(x, alpha = 1, beta = 0, delta = 1/5, mu = 0)
y_4 = pnig(x, alpha = 1, beta = 0, delta = 4, mu = 0)
fig_18 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "NIG (delta = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "NIG (delta = 1/2)") %>%
  add_lines(x = ~x, y = ~y_3, name = "NIG (delta = 1/5)") %>%
  add_lines(x = ~x, y = ~y_4, name = "NIG (delta = 4)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "正态逆高斯分布的概率分布函数"))
fig_18 # 正态逆高斯分布的分布函数曲线
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

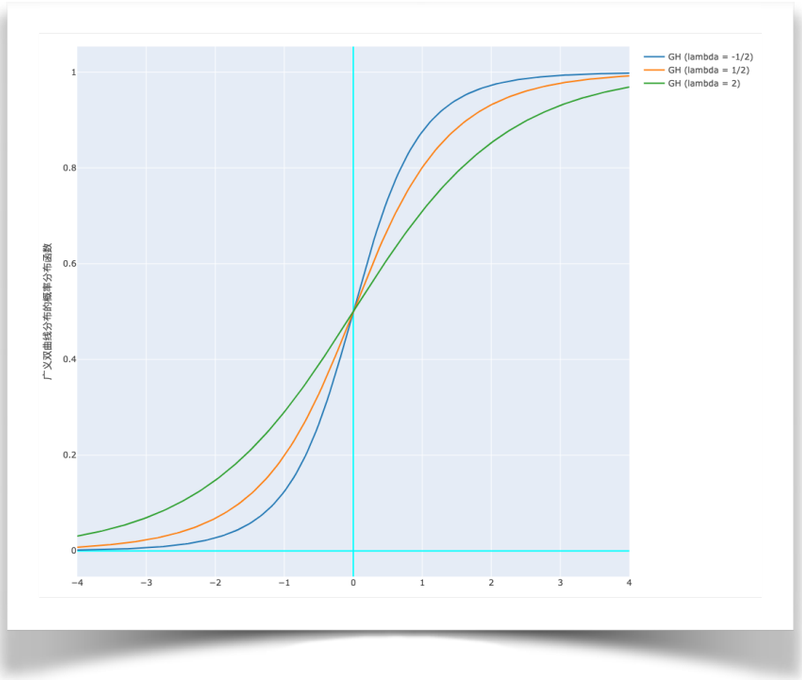
```
# 广义双曲线分布的概率密度函数
x = seq(-4, 4, length = 200)
y_1 = dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = -1/2)
y_2 = dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 1/2)
y_3 = dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 2)
fig_19 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "GH (lambda = -1/2)") %>%
  add_lines(x = ~x, y = ~y_2, name = "GH (lambda = 1/2)") %>%
  add_lines(x = ~x, y = ~y_3, name = "GH (lambda = 2)") %>%
   layout(plot_bgcolor='#e5ecf6',
       xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
       yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "广义双曲线分布的概率密度函数"))
fig_19
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

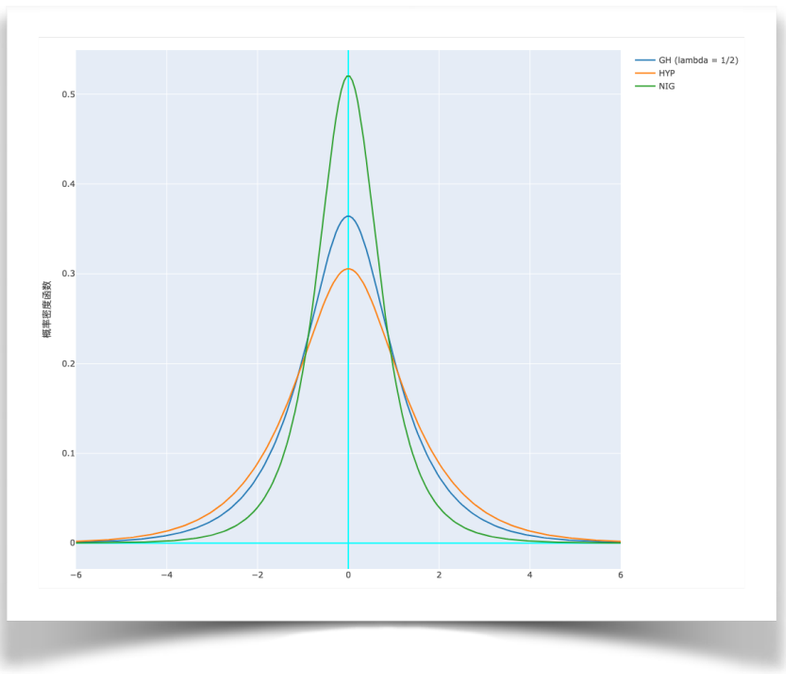
```
# 广义双曲线分布的概率分布函数
x = seq(-4, 4, length = 200)
y_1 = pgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = -1/2)
y_2 = pgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 1/2)
y_3 = pgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 2)
fig_20 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "GH (lambda = -1/2)") %>%
  add_lines(x = ~x, y = ~y_2, name = "GH (lambda = 1/2)") %>%
  add_lines(x = ~x, y = ~y_3, name = "GH (lambda = 2)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "广义双曲线分布的概率分布函数"))
fig_20
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布 (Generalized Hyperbolic Distribution)

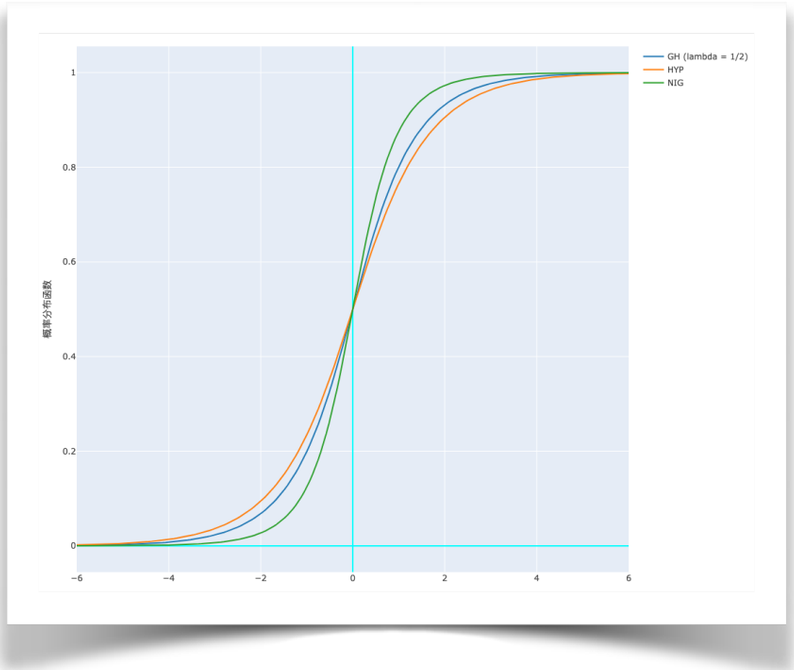
```
# 不同概率密度函数的对比
x = seq(-6, 6, length = 200)
y_1 = dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 0.5)
y_2 = dhyp(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_3 = dnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
fig_21 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "GH (lambda = 1/2)") %>%
  add_lines(x = ~x, y = ~y_2, name = "HYP") %>%
  add_lines(x = ~x, y = ~y_3, name = "NIG") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "概率密度函数"))
fig_21
```



厚尾分布 (heavy-tailed distributions)

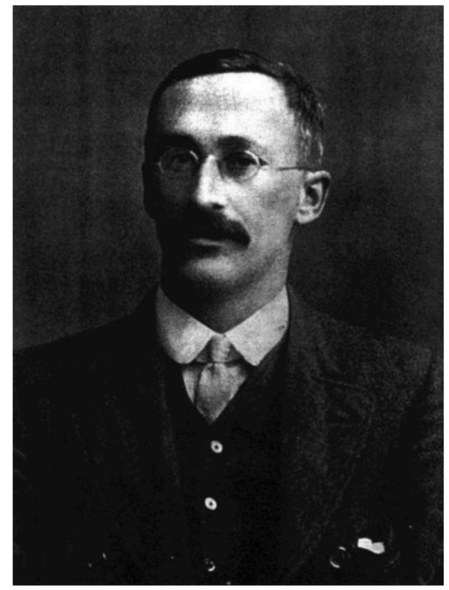
- 广义双曲分布 (Generalized Hyperbolic Distribution)

```
# 不同概率分布函数的对比
x = seq(-6, 6, length = 200)
y_1 = pgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 0.5)
y_2 = phyp(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_3 = pnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
fig_22 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "GH (lambda = 1/2)") %>%
  add_lines(x = ~x, y = ~y_2, name = "HYP") %>%
  add_lines(x = ~x, y = ~y_3, name = "NIG") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "概率分布函数"))
fig_22
```



厚尾分布 (heavy-tailed distributions)

- 学生 t 分布
  - t 分布是由 Gosset (1908) 最先提出的，应雇主要求他只得用笔名 Student 发表.
  - 设 $X \sim N(0,\ 1)$，$Y \sim \chi_n^2$，且 $X,\ Y$ 相互独立，则称 $t = \dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ 服从自由度为 n 的学生 t 分布.



  - 学生 t 分布的概率密度函数为

$$f_t(x;\ n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < \infty$$

自由度

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\mathrm{e}^{-x}dx$$

厚尾分布 (heavy-tailed distributions)

- 学生 t 分布
  - 自由度为 $n\,(n>4)$ 的学生 t 分布的均值、方差、斜度、峰度分别为

$$\text{均值 } \mu = 0$$

$$\text{方差 } \sigma^2 = \frac{n}{n-2}$$

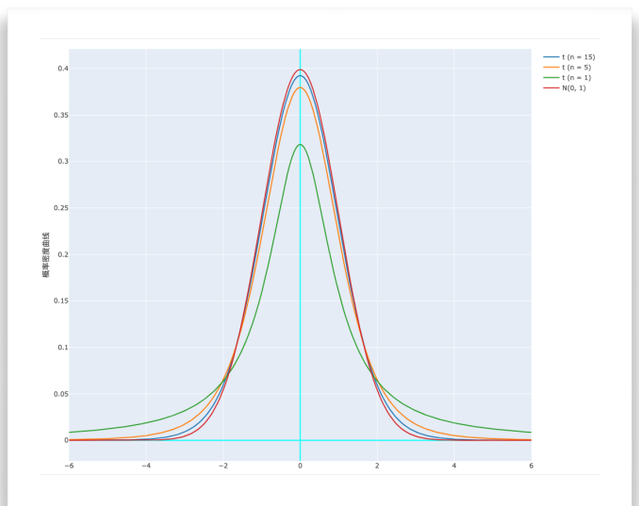
$$\text{斜度} = 0$$

$$\text{峰度} = 3 + \frac{6}{n-4}$$

厚尾分布 (heavy-tailed distributions)

- 学生 t 分布

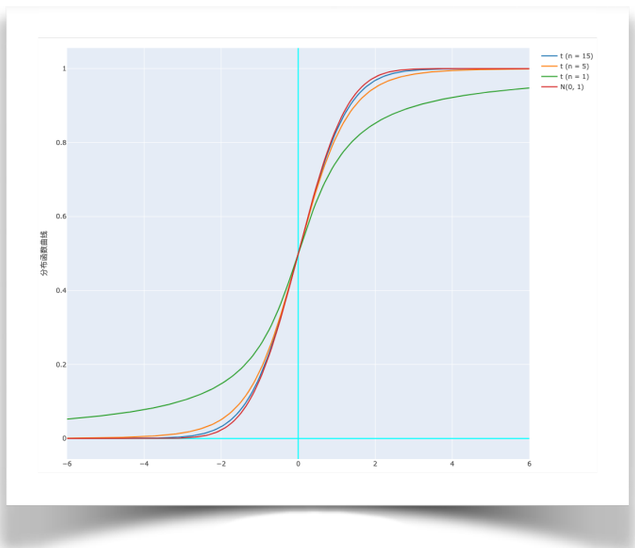
```
# t 分布的密度函数曲线
x = seq(-6, 6, length = 200)
y_1 = dt(x, 15)
y_2 = dt(x, 5)
y_3 = dt(x, 1)
y_4 = dnorm(x, 0, 1)
fig_23 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "t (n = 15)") %>%
  add_lines(x = ~x, y = ~y_2, name = "t (n = 5)") %>%
  add_lines(x = ~x, y = ~y_3, name = "t (n = 1)") %>%
  add_lines(x = ~x, y = ~y_4, name = "N(0, 1)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "概率密度曲线"))
fig_23
```



厚尾分布 (heavy-tailed distributions)

- 学生 t 分布

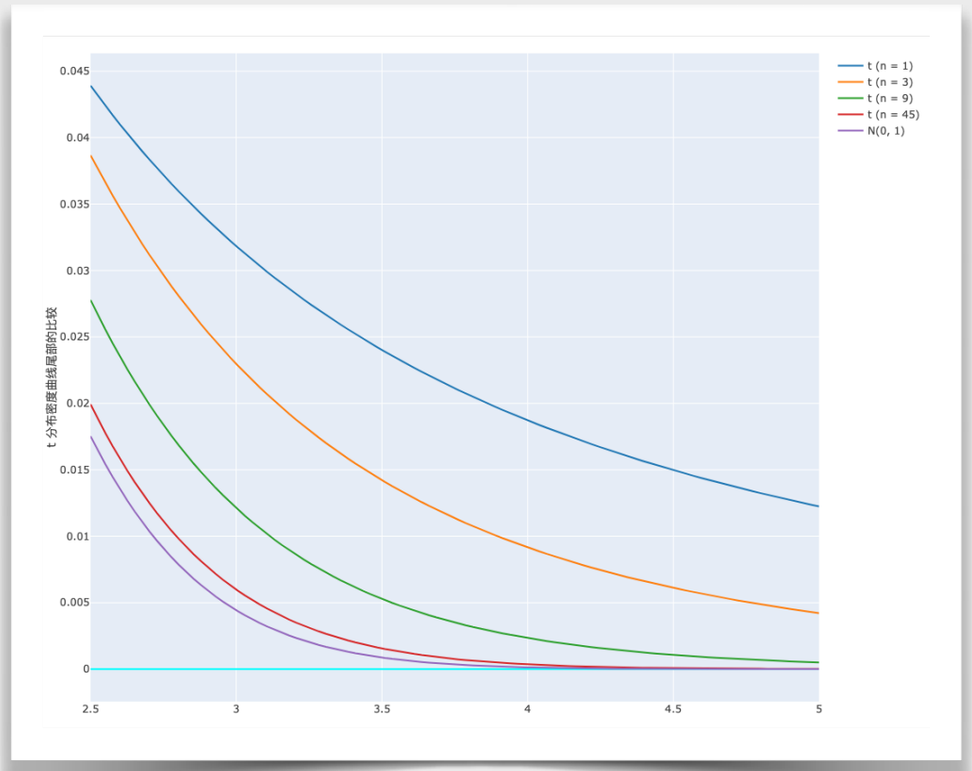
```
# t 分布的分布函数曲线
x = seq(-6, 6, length = 200)
y_1 = pt(x, 15)
y_2 = pt(x, 5)
y_3 = pt(x, 1)
y_4 = pnorm(x, 0, 1)
fig_24 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "t (n = 15)") %>%
  add_lines(x = ~x, y = ~y_2, name = "t (n = 5)") %>%
  add_lines(x = ~x, y = ~y_3, name = "t (n = 1)") %>%
  add_lines(x = ~x, y = ~y_4, name = "N(0, 1)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "分布函数曲线"))
fig_24
```



厚尾分布 (heavy-tailed distributions)

- 学生 t 分布
  - 学生 t 分布随 n 的增大趋于正态分布 $\lim_{n\to\infty} f_t(x;\ n) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{2}}.$
  - 在 t 分布的尾部，x 与 $|x|^{-(n+1)}$ 成正比.

```
# t 分布尾部的比较
x = seq(2.5, 5, length = 100)
y_1 = dt(x, 1)
y_2 = dt(x, 3)
y_3 = dt(x, 9)
y_4 = dt(x, 45)
y_5 = dnorm(x, 0, 1)
fig_25 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "t (n = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "t (n = 3)") %>%
  add_lines(x = ~x, y = ~y_3, name = "t (n = 9)") %>%
  add_lines(x = ~x, y = ~y_4, name = "t (n = 45)") %>%
  add_lines(x = ~x, y = ~y_5, name = "N(0, 1)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "t 分布密度曲线尾部的比较"))
fig_25
```



厚尾分布 (heavy-tailed distributions)

- Laplace 分布
  - 均值为零的一元 Laplace 分布由 Laplace (1774) 引入.

双指数分布

  - Laplace 分布可定义为相互独立且服从同一指数分布的两个随机变量之差的分布.

  - 均值为 μ，尺度参数为 θ 的 Laplace 分布的概率密度函数为

$$f_{\text{Laplace}}(x;\ \mu,\ \theta) = \frac{1}{2\theta}\mathrm{e}^{-\frac{|x-\mu|}{\theta}}\ ,\quad -\infty < x < \infty$$

  - 其分布函数为

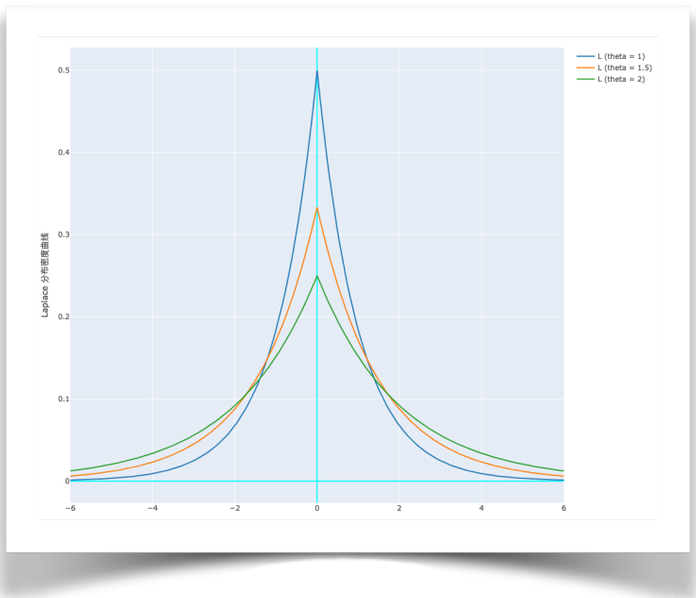
$$\text{sign}(x-\mu) = \begin{cases} 1, & x > \mu \\ 0, & x = \mu \\ -1, & x < \mu \end{cases}$$

$$F_{\text{Laplace}}(x;\ \mu,\ \theta) = \frac{1}{2}\left[1 + \text{sign}(x-\mu)\left(1 - \mathrm{e}^{-\frac{|x-\mu|}{\theta}}\right)\right]$$

厚尾分布 (heavy-tailed distributions)

- Laplace 分布

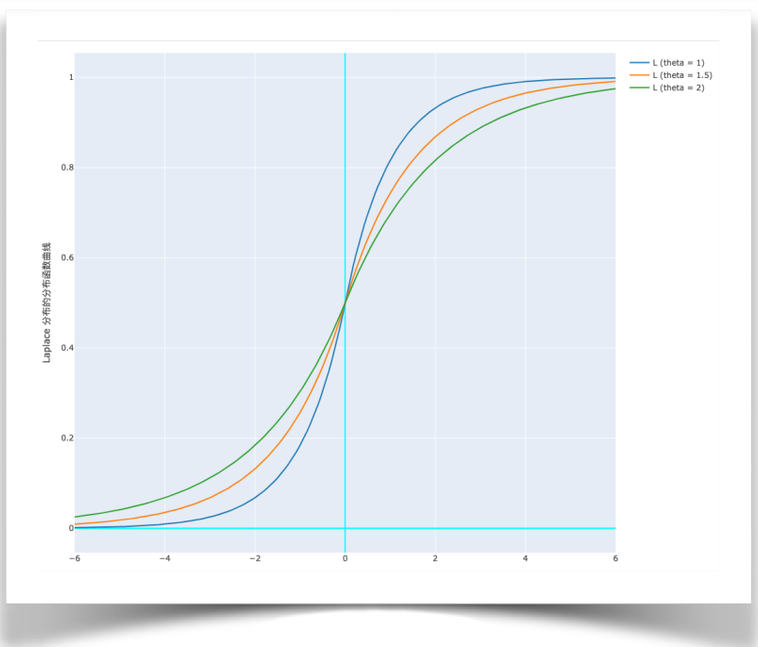
```
# Laplace 分布的密度函数曲线
library(LaplacesDemon)
x = seq(-6, 6, length = 201)
y_1 = dlaplace(x, location = 0, scale = 1)
y_2 = dlaplace(x, location = 0, scale = 1.5)
y_3 = dlaplace(x, location = 0, scale = 2)
fig_26 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "L (theta = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "L (theta = 1.5)") %>%
  add_lines(x = ~x, y = ~y_3, name = "L (theta = 2)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "Laplace 分布密度曲线"))
fig_26
```



厚尾分布 (heavy-tailed distributions)

- Laplace 分布

```
# Laplace 分布的分布函数曲线
x = seq(-6, 6, length = 201)
y_1 = plaplace(x, location = 0, scale = 1)
y_2 = plaplace(x, location = 0, scale = 1.5)
y_3 = plaplace(x, location = 0, scale = 2)
fig_27 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "L (theta = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "L (theta = 1.5)") %>%
  add_lines(x = ~x, y = ~y_3, name = "L (theta = 2)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "Laplace 分布的分布函数曲线"))
fig_27
```



厚尾分布 (heavy-tailed distributions)

- Laplace 分布
  - Laplace 分布的均值、方差、斜度、峰度为

$$\text{均值 } \mu = \mu$$
$$\text{方差 } \sigma^2 = 2\theta^2$$
$$\text{斜度} = 0$$
$$\text{峰度} = 6$$

  - 标准 Laplace 分布：$\mu = 0,\ \theta = 1$

$$f(x) = \frac{\mathrm{e}^{-|x|}}{2}$$

$$F(x) = \begin{cases} \frac{\mathrm{e}^{x}}{2}, & x < 0 \\ 1 - \frac{\mathrm{e}^{-x}}{2}, & x \geq 0 \end{cases}$$

厚尾分布 (heavy-tailed distributions)

- Cauchy 分布
  - 位置参数 m，尺度参数 s 的 Cauchy 分布的概率密度函数和分布函数分别为

$$f_{\text{Cauchy}}(x;\ m,\ s) = \frac{1}{s\pi} \cdot \frac{1}{1+\left(\frac{x-m}{s}\right)^2}, \quad -\infty < x < \infty$$

$$F_{\text{Cauchy}}(x;\ m,\ s) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-m}{s}\right)$$

  - 标准 Cauchy 分布：$m = 0,\ s = 1$

$$f_{\text{Cauchy}}(x) = \frac{1}{\pi\left(1+x^2\right)}, \quad -\infty < x < \infty$$

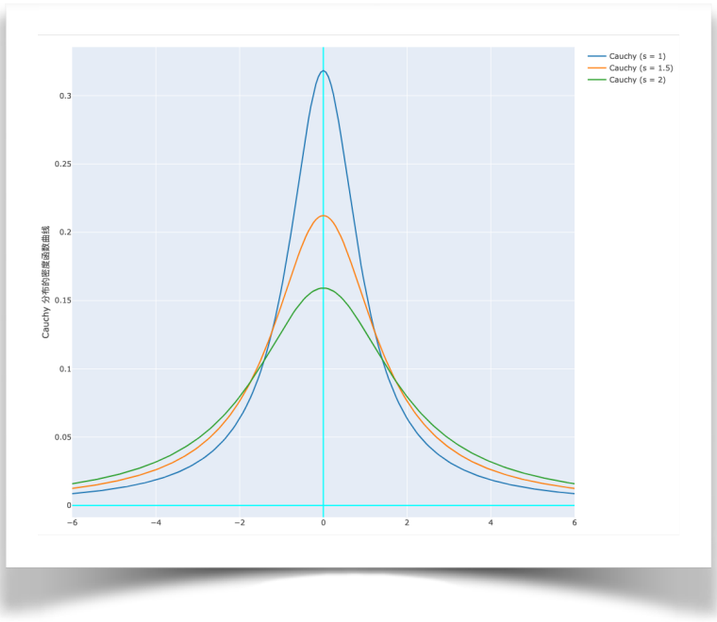
$$F_{\text{Cauchy}}(x) = \frac{1}{2} + \frac{\arctan x}{\pi}$$

  - Cauchy 分布的均值、方差、斜度、峰度不存在.
  - Cauchy 分布的众数 (mode) 和中位数存在，都等于位置参数 m.

厚尾分布 (heavy-tailed distributions)

- Cauchy 分布

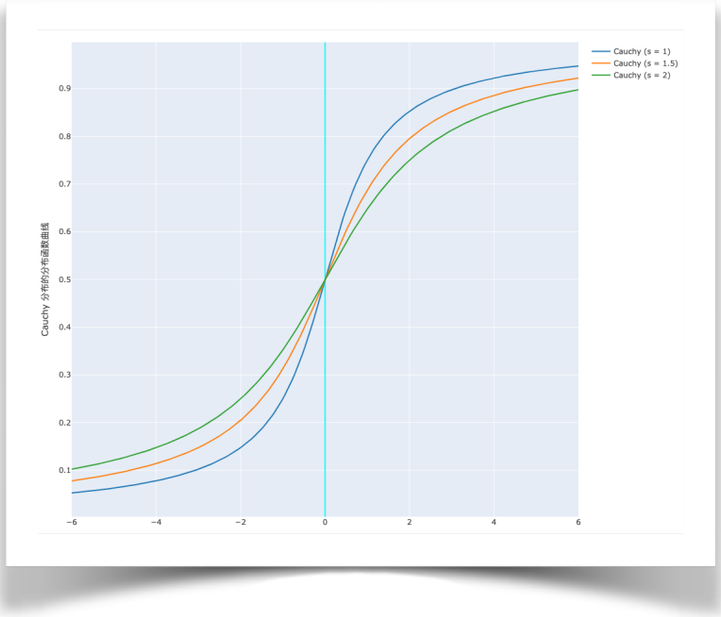
```
# Cauchy 分布的密度函数曲线
x = seq(-6, 6, length = 201)
y_1 = dcauchy(x, location = 0, scale = 1)
y_2 = dcauchy(x, location = 0, scale = 1.5)
y_3 = dcauchy(x, location = 0, scale = 2)
fig_28 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "Cauchy (s = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "Cauchy (s = 1.5)") %>%
  add_lines(x = ~x, y = ~y_3, name = "Cauchy (s = 2)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "Cauchy 分布的密度函数曲线"))
fig_28
```



厚尾分布 (heavy-tailed distributions)

- Cauchy 分布

```
# Cauchy 分布的分布函数曲线
x = seq(-6, 6, length = 201)
y_1 = pcauchy(x, location = 0, scale = 1)
y_2 = pcauchy(x, location = 0, scale = 1.5)
y_3 = pcauchy(x, location = 0, scale = 2)
fig_29 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "Cauchy (s = 1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "Cauchy (s = 1.5)") %>%
  add_lines(x = ~x, y = ~y_3, name = "Cauchy (s = 2)") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "Cauchy 分布的分布函数曲线"))
fig_29
```



厚尾分布 (heavy-tailed distributions)

- 混合模型
  - 由 L 个分布构成的混合分布 (mixture distribution) 的概率密度函数可表示为

$$f(x) = \sum_{l=1}^{L} w_l \cdot p_l(x)$$

权重

密度

约束条件：

$$0 \le w_l \le 1$$

$$\sum_{l=1}^{L} w_l = 1$$

$$\int p_l(x)dx = 1$$

厚尾分布 (heavy-tailed distributions)

- 混合模型
  - 混合分布的均值、方差、斜度、峰度为

$$\text{均值 } \mu = \sum_{l=1}^{L} w_l \mu_l$$

$$\text{方差 } \sigma^2 = \sum_{l=1}^{L} w_l \left[\sigma_l^2 + (\mu_l - \mu)^2 \right]$$

$$\text{斜度} = \sum_{l=1}^{L} w_l \left[\left(\frac{\sigma_l}{\sigma} \right)^3 SK_l + \frac{3\sigma_l^2 (\mu_l - \mu)}{\sigma^3} + \left(\frac{\mu_l - \mu}{\sigma} \right)^3 \right]$$

$$\text{峰度} = \sum_{l=1}^{L} w_l \left[\left(\frac{\sigma_l}{\sigma} \right)^4 K_l + \frac{6 (\mu_l - \mu)^2 \sigma_l^2}{\sigma^4} + \frac{4 (\mu_l - \mu) \sigma_l^3}{\sigma^4} SK_l + \left(\frac{\mu_l - \mu}{\sigma} \right)^4 \right]$$

其中 μ_l、σ_l、SK_l 以及 K_l 分别是分布 $p_l(x)$ 的均值、方差、斜度和峰度.

厚尾分布 (heavy-tailed distributions)

- 混合模型
  - 只要混合分布中密度函数的数量足够多，并且模型的参数选择正确的话，混合模型就能以任意的精度近似任何一个连续型的密度.
  - 统计分析、机器学习和数据挖掘的各方面中，混合模型几乎无处不在.
  - 对于含有连续型变量的数据集而言，最常见的方法均涉及正态分布的混合模型.

厚尾分布 (heavy-tailed distributions)

- 混合模型
  - 正态 (高斯) 混合分布的概率密度函数为

Gaussian mixture

$$f_{\mathrm{GM}}(x)=\sum_{l=1}^{L}\frac{w_l}{\sqrt{2\pi}\ \sigma_l}\ \mathrm{e}^{-\frac{(x-\mu_l)^2}{2\sigma_l^2}}$$

  - 由均值为 0 的正态分布构成的正态混合分布，其密度函数为

$$f_{\mathrm{GM}}(x)=\sum_{l=1}^{L}\frac{w_l}{\sqrt{2\pi}\ \sigma_l}\ \mathrm{e}^{-\frac{x^2}{2\sigma_l^2}}$$

其均值、方差、斜度以及峰度则为

$$\text{均值}\ \mu=0$$

$$\text{方差}\ \sigma^2=\sum_{l=1}^{L}w_l\sigma_l^2$$

$$\text{斜度}=0$$

$$\text{峰度}=\sum_{l=1}^{L}w_l\left(\frac{\sigma_l}{\sigma}\right)^4\cdot 3$$

厚尾分布 (heavy-tailed distributions)

- 例：考虑由 80% 的 $N(0,\ 1)$ 与 20% 的 $N(0,\ 9)$ 混合而成的正态混合分布.
  - $N(0,\ 1)$ 与 $N(0,\ 9)$ 的概率密度函数为

$$f_{N(0,\ 1)}(x)=\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{2}},\quad f_{N(0,\ 9)}(x)=\frac{1}{3\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{18}}$$

  - 则正态混合分布的概率密度函数为

$$f_{\mathrm{GM}}(x)=0.8\cdot f_{N(0,\ 1)}(x)+0.2\cdot f_{N(0,\ 9)}(x)=\frac{1}{5\sqrt{2\pi}}\left(4\mathrm{e}^{-\frac{x^2}{2}}+\frac{1}{3}\mathrm{e}^{-\frac{x^2}{18}}\right)$$

  - 正态混合分布不是正态分布：

$$\text{均值 }\mu=0$$

$$\text{方差 }\sigma^2=0.8\times 1+0.2\times 9=2.6$$

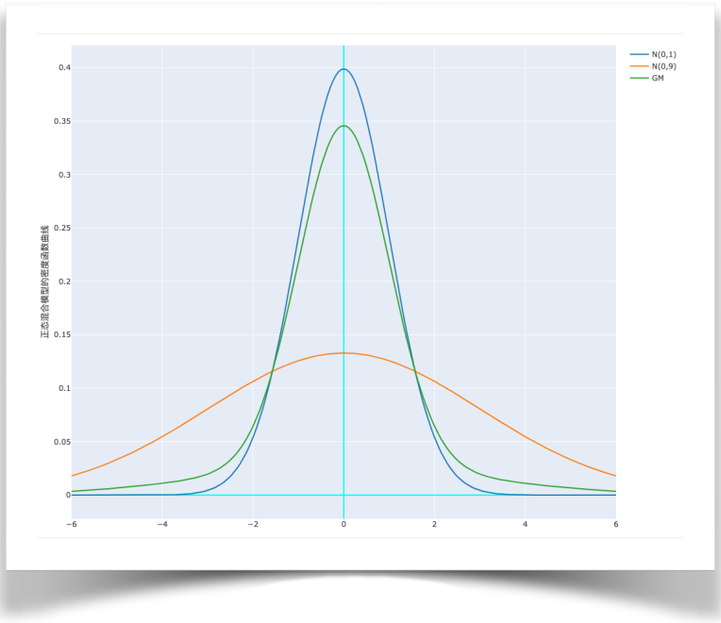
$$\text{斜度}=0$$

$$\text{峰度}=0.8\times\left(\frac{1}{\sqrt{2.6}}\right)^4\times 3+0.2\times\left(\frac{\sqrt{9}}{\sqrt{2.6}}\right)^4\times 3=7.54>3$$

厚尾分布 (heavy-tailed distributions)

- 例：考虑由 80% 的 $N(0, 1)$ 与 20% 的 $N(0, 9)$ 混合而成的正态混合分布.

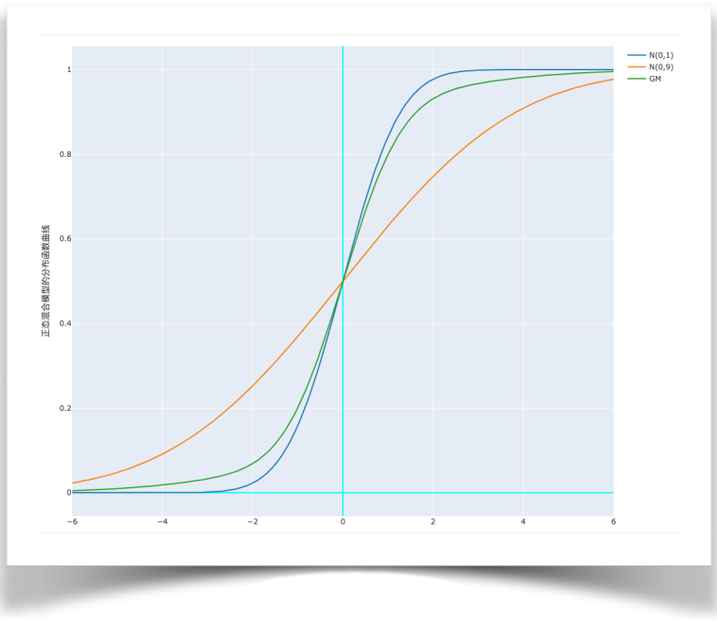
```
# 80% 的 N(0, 1) 与 20% 的 N(0, 9) 构成的正态混合分布的密度函数曲线
x = seq(-6, 6, length = 201)
y_1 = dnorm(x, mean = 0, sd = 1)
y_2 = dnorm(x, mean = 0, sd = 3)
y_3 = 0.8 * y_1 + 0.2 * y_2
fig_30 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "N(0,1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "N(0,9)") %>%
  add_lines(x = ~x, y = ~y_3, name = "GM") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "正态混合模型的密度函数曲线"))
fig_30
```



厚尾分布 (heavy-tailed distributions)

- 例：考虑由 80% 的 $N(0, 1)$ 与 20% 的 $N(0, 9)$ 混合而成的正态混合分布.

```
# 80% 的 N(0, 1) 与 20% 的 N(0, 9) 构成的正态混合分布的分布函数曲线
x = seq(-6, 6, length = 201)
y_1 = pnorm(x, mean = 0, sd = 1)
y_2 = pnorm(x, mean = 0, sd = 3)
y_3 = 0.8 * y_1 + 0.2 * y_2
fig_31 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "N(0,1)") %>%
  add_lines(x = ~x, y = ~y_2, name = "N(0,9)") %>%
  add_lines(x = ~x, y = ~y_3, name = "GM") %>%
  layout(plot_bgcolor='#e5ecf6',
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "正态混合模型的分布函数曲线"))
fig_31
```



厚尾分布 (heavy-tailed distributions)

t 分布，Laplace 分布与 Cauchy 分布的数字特征

	t	Laplace	Cauchy
Mean	0	μ	Not defined
Variance	$\frac{n}{n-2}$	$2\theta^2$	Not defined
Skewness	0	0	Not defined
Kurtosis	$3+\frac{6}{n-4}$	6	Not defined

厚尾分布 (heavy-tailed distributions)

广义双曲 (GH) 分布的数字特征

GH	
均值	$\mu + \frac{\delta\,\beta}{\sqrt{\alpha^2+\beta^2}}\,\frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}$
方差	$\delta^2\left\{\frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}{\delta\sqrt{\alpha^2+\beta^2}\,K_{\lambda}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}+\frac{\beta^2}{\alpha^2+\beta^2}\left[\frac{K_{\lambda+2}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}-\left(\frac{K_{\lambda+1}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}{K_{\lambda}\left(\delta\sqrt{\alpha^2+\beta^2}\right)}\right)^2\right]\right\}$
混合分布	
均值	$\sum_{l=1}^{L} w_l\,\mu_l$
方差	$\sum_{l=1}^{L} w_l\left[\sigma_l^2+\left(\mu_l-\mu\right)^2\right]$
斜度	$\sum_{l=1}^{L} w_l\left[\left(\frac{\sigma_l}{\sigma}\right)^3 SK_l+\frac{3\sigma_l^2\left(\mu_l-\mu\right)}{\sigma^3}+\left(\frac{\mu_l-\mu}{\sigma}\right)^3\right]$
峰度	$\sum_{l=1}^{L} w_l\left[\left(\frac{\sigma_l}{\sigma}\right)^4 K_l+\frac{6\left(\mu_l-\mu\right)^2\sigma_l^2}{\sigma^4}+\frac{4\left(\mu_l-\mu\right)\sigma_l^3}{\sigma^4}SK_l+\left(\frac{\mu_l-\mu}{\sigma}\right)^4\right]$

厚尾分布 (heavy-tailed distributions)

- 多元广义双曲分布
  - 多元广义双曲分布 $\left(\mathrm{GH}_d\right)$ 的概率密度函数为

$$f_{\mathrm{GH}_d}\left(\boldsymbol{x};\ \lambda,\ \alpha,\ \boldsymbol{\beta},\ \delta,\ \Delta,\ \boldsymbol{\mu}\right)=a_d\ \frac{K_{\lambda-\frac{d}{2}}\left[\alpha\sqrt{\delta^2+\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}}\Delta^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}\right]}{\left[\frac{1}{\alpha}\sqrt{\delta^2+\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^{\mathrm{T}}\Delta^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)}\right]^{\frac{d}{2}-\lambda}}\ \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})}$$

$$a_d=a_d\left(\lambda,\ \alpha,\ \boldsymbol{\beta},\ \delta,\ \Delta\right)=\frac{\left(\frac{1}{\delta}\sqrt{\alpha^2-\boldsymbol{\beta}^{\mathrm{T}}\Delta\boldsymbol{\beta}}\right)^{\lambda}}{(2\pi)^{\frac{d}{2}}K_{\lambda}\left(\delta\sqrt{\alpha^2-\boldsymbol{\beta}^{\mathrm{T}}\Delta\boldsymbol{\beta}}\right)}$$

其中

$$\lambda\in\mathbb{R}\ ,\qquad \boldsymbol{\beta}\ ,\boldsymbol{\mu}\in\mathbb{R}^d\ ,\qquad \delta>0\ ,\qquad \alpha>\boldsymbol{\beta}^{\mathrm{T}}\Delta\boldsymbol{\beta}$$

$$\Delta\in\mathbb{R}^{d\times d}\quad \text{正定矩阵}\ ,\qquad \left|\Delta\right|=1$$

厚尾分布 (heavy-tailed distributions)

- 多元广义双曲分布
  - 多元广义双曲分布 $\left(\mathrm{GH}_d\right)$ 的特征函数为

$$\Phi(\boldsymbol{t}) = \left(\frac{\alpha^2 - \boldsymbol{\beta}^{\mathrm{T}} \Delta \boldsymbol{\beta}}{\alpha^2 - \boldsymbol{\beta}^{\mathrm{T}} \Delta \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{t}^{\mathrm{T}} \Delta \boldsymbol{t} - \mathrm{i} \boldsymbol{\beta}^{\mathrm{T}} \Delta \boldsymbol{t}} \right)^{\frac{\lambda}{2}} \cdot \frac{K_\lambda\left(\delta \sqrt{\alpha^2 - \boldsymbol{\beta}^{\mathrm{T}} \Delta \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{t}^{\mathrm{T}} \Delta \boldsymbol{t} - \mathrm{i} \boldsymbol{\beta}^{\mathrm{T}} \Delta \boldsymbol{t}} \right)}{K_\lambda\left(\delta \sqrt{\alpha^2 - \boldsymbol{\beta}^{\mathrm{T}} \Delta \boldsymbol{\beta}} \right)}$$

  - 多元双曲 (HYP) 分布：$\lambda = \frac{d+1}{2}$.
  - 多元正态逆高斯 (NIG) 分布：$\lambda = -\frac{1}{2}$.

厚尾分布 (heavy-tailed distributions)

- 多元广义双曲分布

定理 4.12 假设 $\boldsymbol{X}$ 服从 d 维广义双曲分布 GH_d，$\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right)$ 是 $\boldsymbol{X}$ 的一个分割，r 与 k 分别是 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 的维数，用 $\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ 与 $\left(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\right)$ 分别表示 $\boldsymbol{\beta}$ 与 $\boldsymbol{\mu}$ 的相应分割，记

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$$

是 Δ 的一个分块矩阵，其中 Δ_{11} 为 $r \times r$ 矩阵. 则有：

1. $\boldsymbol{X}_1$ 服从 r 维的广义双曲分布 $\mathrm{GH}_r\left(\lambda^*, \alpha^*, \boldsymbol{\beta}^*, \delta^*, \boldsymbol{\mu}^*, \Delta^*\right)$，其中

$$\lambda^* = \lambda$$

$$\alpha^* = \left|\Delta_{11}\right|^{-\frac{1}{2r}} \left[\alpha^2 - \boldsymbol{\beta}_2 \left(\Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right) \boldsymbol{\beta}_2^{\mathrm{T}}\right]^{\frac{1}{2}}$$

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 \Delta_{21} \Delta_{11}^{-1}$$

$$\delta^* = \delta \left|\Delta_{11}\right|^{\frac{1}{2r}}$$

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_1$$

$$\Delta^* = \left|\Delta\right|^{-\frac{1}{r}} \Delta_{11}$$

厚尾分布 (heavy-tailed distributions)

- 多元广义双曲分布

定理 4.12 假设 $\boldsymbol{X}$ 服从 d 维广义双曲分布 GH_d，$\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right)$ 是 $\boldsymbol{X}$ 的一个分割，r 与 k 分别是 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 的维数，用 $\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ 与 $\left(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\right)$ 分别表示 $\boldsymbol{\beta}$ 与 $\boldsymbol{\mu}$ 的相应分割，记

$$\Delta=\begin{pmatrix}\Delta_{11} & \Delta_{12}\\ \Delta_{21} & \Delta_{22}\end{pmatrix}$$

是 Δ 的一个分块矩阵，其中 Δ_{11} 为 $r\times r$ 矩阵. 则有：

2. 给定 $\boldsymbol{X}_1=\boldsymbol{x}_1$ 时 $\boldsymbol{X}_2$ 的条件分布为 k 维广义双曲分布 $\mathrm{GH}_k\left(\tilde{\lambda}, \tilde{\alpha}, \tilde{\boldsymbol{\beta}}, \tilde{\delta}, \tilde{\boldsymbol{\mu}}, \tilde{\Delta}\right)$，其中

$$\tilde{\lambda}=\lambda-\frac{r}{2}$$

$$\tilde{\alpha}=\alpha\left|\Delta_{11}\right|^{\frac{1}{2k}}$$

$$\tilde{\boldsymbol{\beta}}=\boldsymbol{\beta}_2$$

$$\tilde{\delta}=\left|\Delta_{11}\right|^{-\frac{1}{2k}}\left[\delta^2+\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right)\Delta_{11}^{-1}\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right)^{\mathrm{T}}\right]^{\frac{1}{2}}$$

$$\tilde{\boldsymbol{\mu}}=\boldsymbol{\mu}_2+\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right)\Delta_{11}^{-1}\Delta_{12}$$

$$\tilde{\Delta}=\left|\Delta_{11}\right|^{\frac{1}{k}}\left(\Delta_{22}-\Delta_{21}\Delta_{11}^{-1}\Delta_{12}\right)$$

厚尾分布 (heavy-tailed distributions)

- 多元广义双曲分布

定理 4.12 假设 $\boldsymbol{X}$ 服从 d 维广义双曲分布 GH_d，$\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right)$ 是 $\boldsymbol{X}$ 的一个分割，r 与 k 分别是 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 的维数，用 $\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ 与 $\left(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\right)$ 分别表示 $\boldsymbol{\beta}$ 与 $\boldsymbol{\mu}$ 的相应分割，记

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}$$

是 Δ 的一个分块矩阵，其中 Δ_{11} 为 $r \times r$ 矩阵. 则有：

3. 设 $\boldsymbol{Y} = \boldsymbol{X}A + B$ 是 $\boldsymbol{X}$ 的正则仿射变换，记 $\|A\|$ 为 A 的行列式的绝对值. 则 $\boldsymbol{Y}$ 服从 d 维广义双曲分布 $\mathrm{GH}_d\left(\lambda^+, \alpha^+, \boldsymbol{\beta}^+, \delta^+, \boldsymbol{\mu}^+, \Delta^+\right)$，其中

A 非奇异

线性变换 + 平移

$$\begin{aligned}
\lambda^+ &= \lambda \\
\alpha^+ &= \alpha \ \|A\|^{-\frac{1}{d}} \\
\boldsymbol{\beta}^+ &= \boldsymbol{\beta}\left(A^{-1}\right)^{\mathrm{T}} \\
\delta^+ &= \|A\|^{\frac{1}{d}} \\
\boldsymbol{\mu}^+ &= \boldsymbol{\mu}A + B \\
\Delta^+ &= \|A\|^{-\frac{2}{d}} A^{\mathrm{T}} \Delta A
\end{aligned}$$

厚尾分布 (heavy-tailed distributions)

- 多元 t 分布
  - 设 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，$Y \sim \chi^2_n$，且相互独立，则称

$$\boldsymbol{t} = \frac{\boldsymbol{X}}{\sqrt{Y/n}} + \boldsymbol{\mu}$$

服从自由度为 n、非中心参数为 $\boldsymbol{\mu}$ 的非中心 t 分布，$\boldsymbol{t}$ 的概率密度函数为

$$f_{\boldsymbol{t}}(\boldsymbol{t};\ n, \boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{\Gamma\left(\frac{n+p}{2}\right)}{\Gamma\left(\frac{n}{2}\right) n^{\frac{p}{2}}\ \pi^{\frac{p}{2}}\ \left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}\ \left[1 + \frac{1}{n}(\boldsymbol{t}-\boldsymbol{\mu})^{\mathrm{T}}\ \boldsymbol{\Sigma}^{-1}\ (\boldsymbol{t}-\boldsymbol{\mu})\right]^{\frac{n+p}{2}}}$$

厚尾分布 (heavy-tailed distributions)

- 多元 Laplace 分布
  - 设 g 与 G 分别表示 d 维正态分布 $N_d\ (\mathbf{0},\ \Sigma)$ 的概率密度函数和分布函数，则多元 Laplace 分布的概率密度函数和分布函数为

$$f_{M\text{Laplace}_d}(\boldsymbol{x};\ \boldsymbol{m},\ \Sigma) = \int_0^{\infty} g\left(z^{-\frac{1}{2}}\boldsymbol{x} - z^{\frac{1}{2}}\boldsymbol{m}\right) z^{-\frac{d}{2}}\mathrm{e}^{-z}dz$$

$$F_{M\text{Laplace}_d}(\boldsymbol{x};\ \boldsymbol{m},\ \Sigma) = \int_0^{\infty} G\left(z^{-\frac{1}{2}}\boldsymbol{x} - z^{\frac{1}{2}}\boldsymbol{m}\right) \mathrm{e}^{-z}dz$$

第三类修正的 Bessel 函数

  - 概率密度函数也可以表示为

$$f_{M\text{Laplace}_d}(\boldsymbol{x};\ m,\ \Sigma) = \frac{2\mathrm{e}^{\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{m}}}{(2\pi)^{\frac{d}{2}}\left|\Sigma\right|^{\frac{1}{2}}}\left(\frac{\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}}{2+\boldsymbol{m}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{m}}\right)^{\frac{\lambda}{2}} K_{\lambda}\left(\sqrt{\left(2+\boldsymbol{m}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{m}\right)\left(\boldsymbol{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{x}\right)}\right)$$

$$K_{\lambda} = \frac{1}{2}\left(\frac{x}{2}\right)^{\lambda}\int_0^{\infty} t^{-\lambda-1}\mathrm{e}^{-t-\frac{x^2}{4t}}dt\,,\ x>0$$

$$\lambda = \frac{2-d}{2}$$

  - 均值与方差

$$\mathrm{E}(\boldsymbol{X}) = \boldsymbol{m}$$

$$\mathrm{Var}(\boldsymbol{X}) = \Sigma + \boldsymbol{m}\boldsymbol{m}^{\mathrm{T}}$$

厚尾分布 (heavy-tailed distributions)

- 多元混合模型
  - 多元混合模型由多元分布构成.
  - 例如，多元正态混合分布的概率密度函数可以表示以为

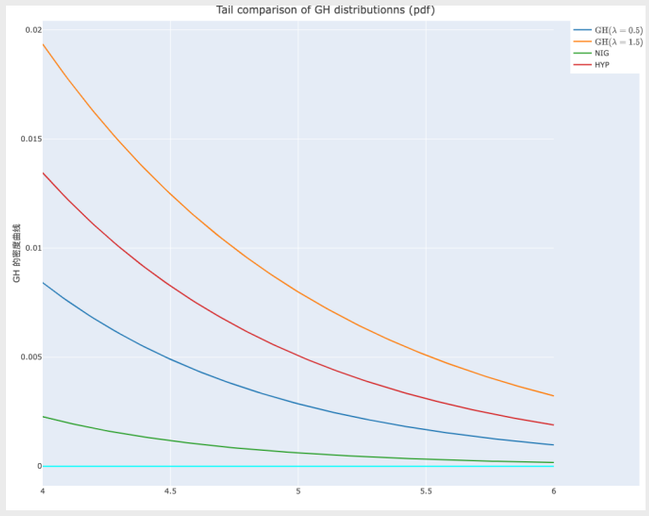
$$f(\boldsymbol{x}) = \sum_{l=1}^{L} \frac{w_l}{\left|2\pi\Sigma_l\right|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_l)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_l)}$$

厚尾分布 (heavy-tailed distributions)

- 广义双曲分布
  - 广义双曲分布以指数速度衰减

$$f_{\mathrm{GH}}(x;\ \lambda,\ \alpha,\ \beta,\ \delta,\ \mu=0) \sim x^{\lambda-1}\mathrm{e}^{-(\alpha-\beta)x}, \qquad x \to \infty$$

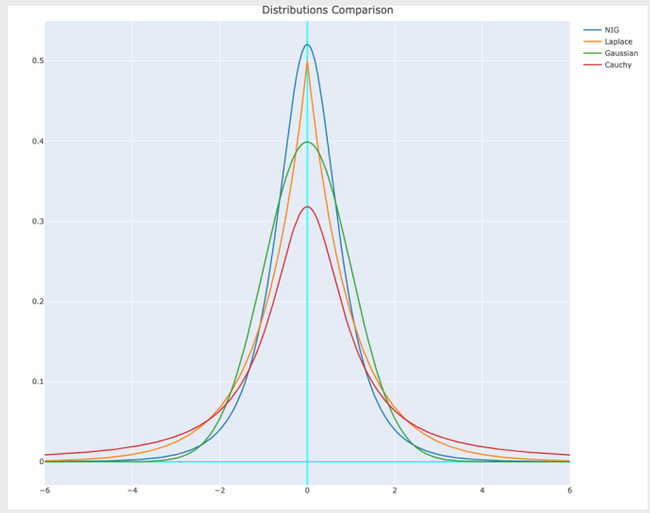
```
# 广义双曲分布以指数速度衰减
x = seq(4, 6, length = 101)
y_1 = dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 0.5)
y_2 = dgh(x, alpha = 1, beta = 0, delta = 1, mu = 0, lambda = 1.5)
y_3 = dnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_4 = dhyp(x, alpha = 1, beta = 0, delta = 1, mu = 0)
fig_32 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = TeX("\\text{GH} (\\lambda = 0.5)")) %>%
  add_lines(x = ~x, y = ~y_2, name = TeX("\\text{GH} (\\lambda = 1.5)")) %>%
  add_lines(x = ~x, y = ~y_3, name = "NIG") %>%
  add_lines(x = ~x, y = ~y_4, name = "HYP") %>%
  layout(plot_bgcolor='#e5ecf6', title = "Tail comparison of GH distributionns (pdf)",
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = "GH 的密度曲线")) %>%
  config(mathjax = "cdn")
fig_32
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布
  - 正态逆高斯 (NIG)、Laplace、Cauchy、标准正态分布 (Gaussian) 的比较.

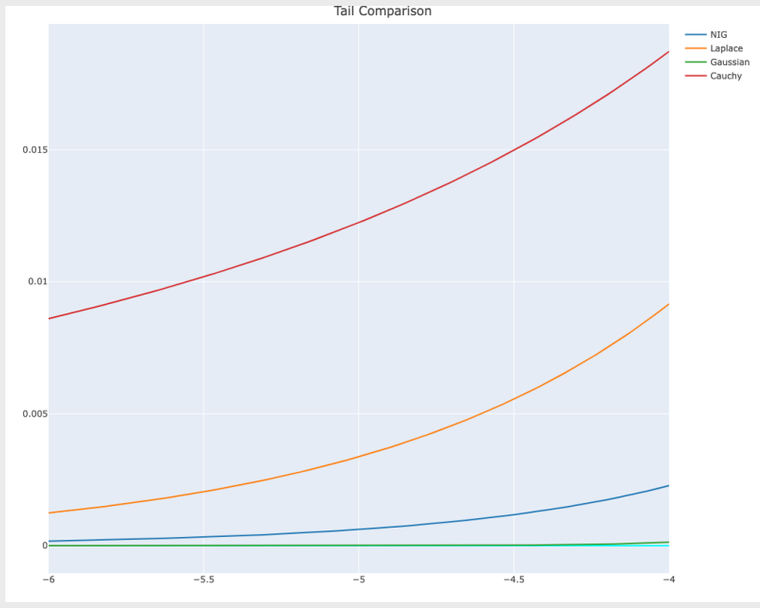
```
# 正态逆高斯(NIG)、Laplace、Cauchy、标准正态(Gaussian) 分布的密度曲线的比较
x = seq(-6, 6, length = 201)
y_1 = dnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_2 = dlaplace(x, location = 0, scale = 1)
y_3 = dnorm(x, 0, 1)
y_4 = dcauchy(x, location = 0, scale = 1)
fig_33 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "NIG") %>%
  add_lines(x = ~x, y = ~y_2, name = "Laplace") %>%
  add_lines(x = ~x, y = ~y_3, name = "Gaussian") %>%
  add_lines(x = ~x, y = ~y_4, name = "Cauchy") %>%
  layout(plot_bgcolor='#e5ecf6', title = "Distributions Comparison",
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = ""))
fig_33
```



厚尾分布 (heavy-tailed distributions)

- 广义双曲分布
  - 正态逆高斯 (NIG)、Laplace、Cauchy、标准正态分布 (Gaussian) 的比较.

```
# 正态逆高斯(NIG)、Laplace、Cauchy、标准正态(Gaussian) 分布尾部的比较
x = seq(-6, -4, length = 101)
y_1 = dnig(x, alpha = 1, beta = 0, delta = 1, mu = 0)
y_2 = dlaplace(x, location = 0, scale = 1)
y_3 = dnorm(x, 0, 1)
y_4 = dcauchy(x, location = 0, scale = 1)
fig_34 = plot_ly() %>%
  add_lines(x = ~x, y = ~y_1, name = "NIG") %>%
  add_lines(x = ~x, y = ~y_2, name = "Laplace") %>%
  add_lines(x = ~x, y = ~y_3, name = "Gaussian") %>%
  add_lines(x = ~x, y = ~y_4, name = "Cauchy") %>%
  layout(plot_bgcolor='#e5ecf6', title = "Tail Comparison",
         xaxis = list(zerolinecolor = 'cyan', zerolinewidth = 2, gridcolor = 'ffff', title = ""),
         yaxis = list(zerolinecolor = "cyan", zerolinewidth = 2, gridcolor = 'ffff', title = ""))
fig_34
```



自助法 (Bootstrap)

- 为通过中心极限定理 (CLT) 计算得到足够近似的临界值，我们需要样本容量很大.

一维数据一般要求 $n > 50$

- 样本容量较小时该如何构造置信区间? 自助法 (Bootstrap)
- 自助算法 (bootstrap algorithm) 会两次使用数据:
  - 估计我们关注的参数.
  - 从一个估计的分布进行模拟，以近似我们关注的统计量的渐近分布.

自助法 (Bootstrap)

- 一维数据的情形：

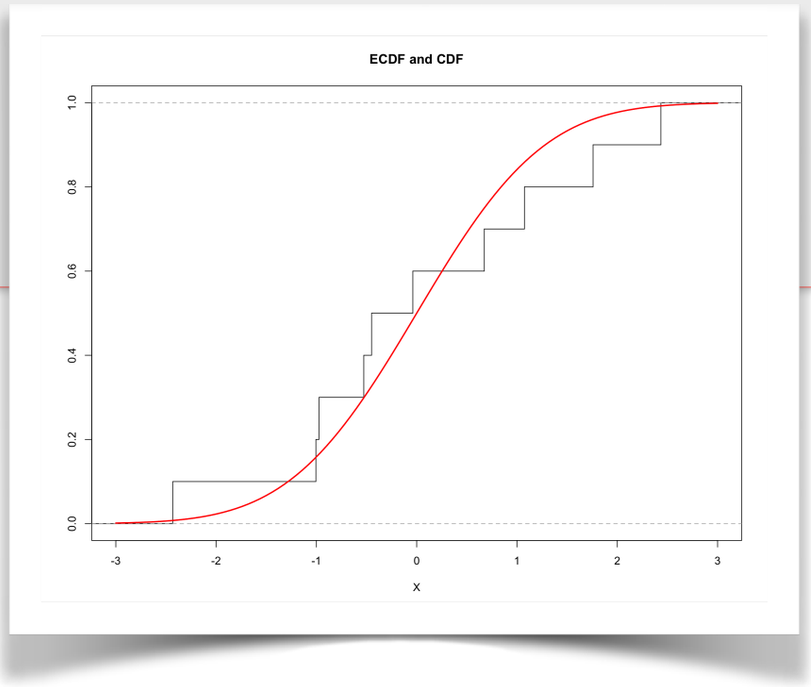
随机样本：$X_1, X_2, \ldots, X_n$　　观测值：$x_1, x_2, \ldots, x_n$

经验分布函数 (ecdf)：$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\left(X_i \le x\right)$

这是一个阶梯函数，在两个相邻的观测值之间的取值是一个常数.

- 例：从标准正态分布 $N(0, 1)$ 随机抽取 $n = 10$ 个数据点.

```
n = 10
x = rnorm(n, 0, 1)
plot(ecdf(x), xlab = 'X', ylab ='', xlim = c(-3, 3), main = 'ECDF and CDF',
     verticals = TRUE, do.points = FALSE)
curve(pnorm(x, 0, 1), -3, 3, add = TRUE, col = 'red', lwd = 2)
```



自助法 (Bootstrap)

- 一维数据的情形：

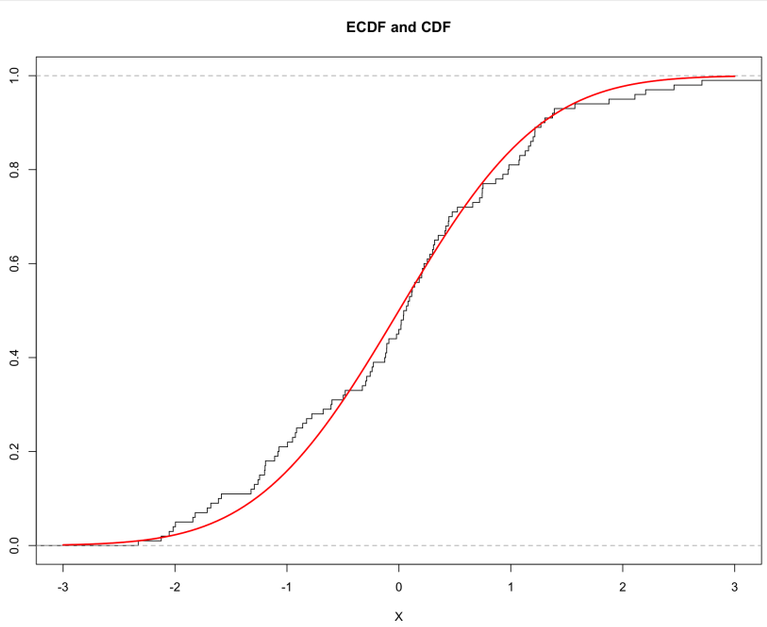
随机样本：$X_1, X_2, \ldots, X_n$　　观测值：$x_1, x_2, \ldots, x_n$

经验分布函数 (edf)：$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\left(X_i \le x\right)$

这是一个阶梯函数，在两个相邻的观测值之间的取值是一个常数.

- 例：从标准正态分布 $N(0, 1)$ 随机抽取 $n = 100$ 个数据点.

```
n = 100
x = rnorm(n, 0, 1)
plot(ecdf(x), xlab = 'X', ylab ='', xlim = c(-3, 3), main = 'ECDF and CDF',
     verticals = TRUE, do.points = FALSE)
curve(pnorm(x, 0, 1), -3, 3, add = TRUE, col = 'red', lwd = 2)
```



自助法 (Bootstrap)

- 一维数据的情形：

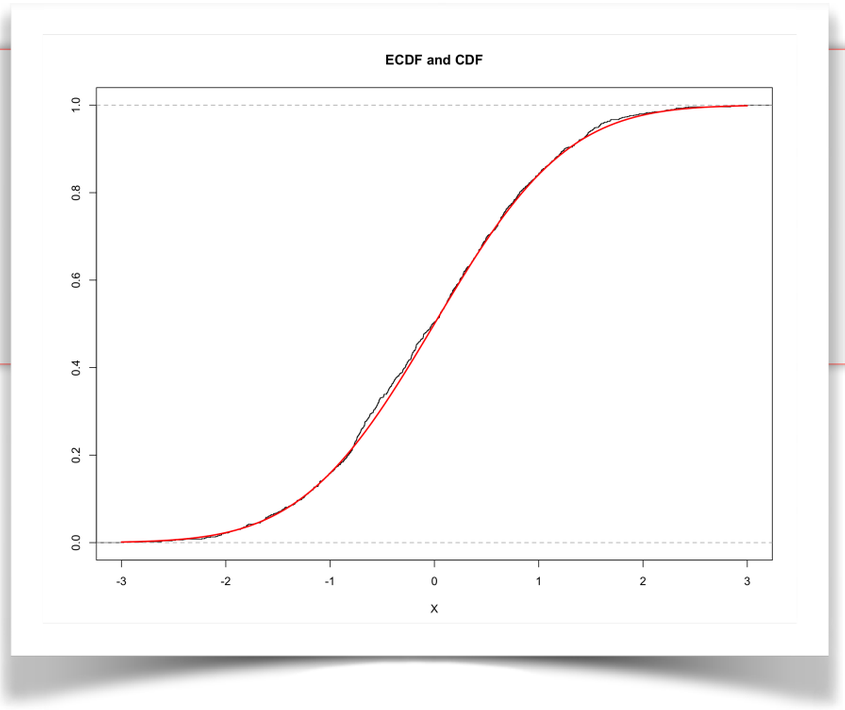
随机样本：$X_1, X_2, \ldots, X_n$　　观测值：$x_1, x_2, \ldots, x_n$

经验分布函数 (edf)：$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\left(X_i \le x\right)$

这是一个阶梯函数，在两个相邻的观测值之间的取值是一个常数.

- 例：从标准正态分布 $N(0, 1)$ 随机抽取 $n = 1000$ 个数据点.

```
n = 1000
x <- rnorm(n, 0, 1)
plot(ecdf(x), xlab = 'X', ylab ='', xlim = c(-3, 3), main = 'ECDF and CDF',
     verticals = TRUE, do.points = FALSE)
curve(pnorm(x, 0, 1), -3, 3, add = TRUE, col = 'red', lwd = 2)
```



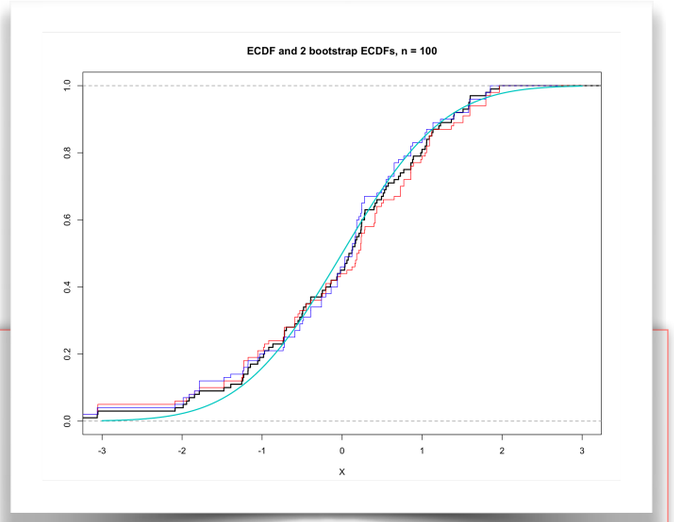
自助法 (Bootstrap)

- 自助样本 (Bootstrap sample): 从初始样本 $X_1, X_2, \ldots, X_n$ 中可重复的抽取的一个样本 $X_1^*, X_2^*, \ldots, X_{n^*}^*$, 一般我们取 $n^* = n$.
  - X_i^* 等可能取 $X_1, X_2, \ldots, X_n$ $(i = 1, 2, \ldots, n)$ 之一.

$$\mathrm{E}_{F_n}\left(X_i^*\right) = \sum_{i=1}^{n}\left(x_i \cdot \frac{1}{n}\right) = \bar{x}$$

  - 对方差也有同样的结果，即

$$\mathrm{Var}_{F_n}\left(X_i^*\right) = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$




```
n = 100
x = rnorm(n, 0, 1)
y1 = sample(x, n, replace = TRUE)
y2 = sample(x, n, replace = TRUE)
plot(ecdf(x), xlab = 'X', ylab ='', xlim = c(-3, 3), main = 'ECDF and 2 bootstrap ECDFs, n = 100',
    verticals = TRUE, do.points = FALSE, lwd = 2, col = 'black')
plot(ecdf(y1), verticals = TRUE, do.points = FALSE, col = 'red', add = TRUE)
plot(ecdf(y2), verticals = TRUE, do.points = FALSE, col = 'blue', add = TRUE)
curve(pnorm(x, 0, 1), -3, 3, add = TRUE, col = 'cyan3', lwd = 2)
```

自助法 (Bootstrap)

推论 4.2 如果 $X_1^*, X_2^*, \ldots, X_n^*$ 是取自 $X_1, X_2, \ldots, X_n$ 的一个自助样本，则

$$\sqrt{n}\left(\frac{\bar{x}^* - \bar{x}}{\hat{\sigma}^*}\right)$$

的分布渐进服从 $N(0, 1)$，其中

$$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} X_i^* \ , \quad (\hat{\sigma}^*)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^* - \bar{x}^*\right)^2$$

自助法 (Bootstrap)

- 如何利用自助法计算 μ 的置信区间?

$$\sqrt{n}\left(\frac{\bar{x}-\mu}{\hat{\sigma}}\right) \longrightarrow N(0,\,1)\,, \quad n \longrightarrow \infty \quad \Longrightarrow \quad \left[\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}\cdot u_{\frac{\alpha}{2}}\,,\ \bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}\cdot u_{\frac{\alpha}{2}}\right]$$

  - 对于较小的 n, $\sqrt{n}\left(\frac{\bar{x}-\mu}{\hat{\sigma}}\right)$ 可能与 $N(0,\,1)$ 的差异较大，因此上侧分位数 $u_{\frac{\alpha}{2}}$ 也可能很差.

  - 自助的思想使我们能够利用很多自助样本，通过计算

$$\sqrt{n}\left(\frac{\bar{x}^{*}-\bar{x}}{\hat{\sigma}^{*}}\right)$$

来“模拟”这一分布，然后据此来估计经验上侧分位数 $u^{*}_{\frac{\alpha}{2}}$.

自助法 (Bootstrap)

- 如何利用自助法计算 μ 的置信区间?

$$\sqrt{n}\left(\frac{\bar{x}-\mu}{\hat{\sigma}}\right) \longrightarrow N(0,\,1)\,, \quad n \longrightarrow \infty \quad \Longrightarrow \quad \left[\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}\cdot u_{\frac{\alpha}{2}}\,,\,\bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}\cdot u_{\frac{\alpha}{2}}\right]$$

  - 于是，自助法改进后的置信区间为

$$C^*_{1-\alpha}=\left[\bar{x}-\frac{\hat{\sigma}}{\sqrt{n}}\cdot u^*_{\frac{\alpha}{2}}\,,\,\bar{x}+\frac{\hat{\sigma}}{\sqrt{n}}\cdot u^*_{\frac{\alpha}{2}}\right]$$

  - 由推论 4.2 我们有

$$P\left(\mu\in C^*_{1-\alpha}\right)\longrightarrow 1-\alpha\,,\quad \text{as}\quad n\to\infty$$

  但是收敛的速度得以改善.