数据科学基础

Foundations of Data Science

4.4 常用抽样分布

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

抽样分布定理

【定理】设某总体的均值为μ，方差为σ^2。 X_1， $\cdots$ ，X_n为总体一样本，$\bar{X}$为样本均值，S^2为样本方差，则

(1) $\mathrm{E}(\bar{X}) = \mu$

(2) $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

(3) $\mathrm{E}(S^2) = \sigma^2$

常用抽样分布

三大抽样分布：

- χ^2-分布
- t-分布
- F-分布

Karl Pearson　　W. S. Gosset　　R. A. Fisher

χ^2分布的故事

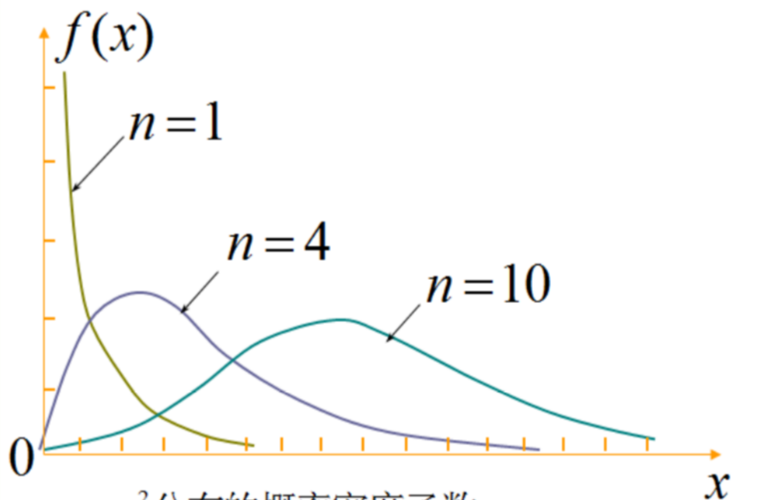
- 卡尔·皮尔逊(Karl Pearson)--χ^2分布。
- χ^2分布最早发现者是物理学家麦克斯韦，他发现分子速度在三个坐标轴上的分量是正态分布，而分子运动速度的平方符合自由度为3的χ^2分布。
- 卡尔.皮尔逊在分布曲线和数据的拟合优度检验中，采用 χ^2分布，这个工作被认为是假设检验的开山之作。

χ^2分布

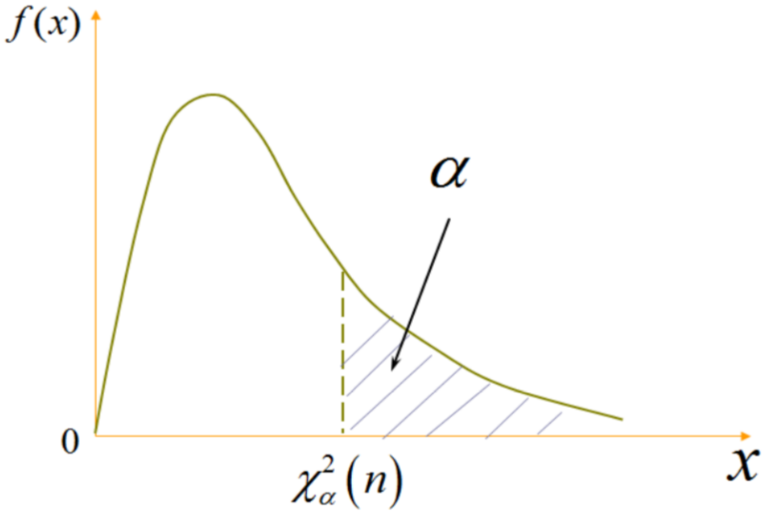
【卡方分布】设X_1，⋯，X_n相互独立，且$X_i \sim N(0，1)$，则称随机变量

$$\chi^2 = \sum_{i=1}^{n} {X_i}^2$$

所服从的分布为自由度为n的χ^2分布，记为$\chi^2(n)$。



χ^2分布的概率密度函数



χ^2分布的分位数

更多详情，请咨询课程助教。

χ^2分布性质

【定理】 $\chi^2 \sim \chi^2(n)$，则

$$E(\chi^2) = n, D(\chi^2) = 2n$$

【定理】 设$X_1 \sim \chi^2(n_1)$,$X_2 \sim \chi^2(n_2)$，且X_1, X_2相互独立，则

$$X_1 + X_2 \sim \chi^2(n_1 + n_2)$$

对于$0 < \alpha < 1$,其上分位点$\chi_\alpha^2(n)$可以通过χ^2分布表查询可得。

正态样本均值和方差的抽样分布

【定理】设$X_1, \cdots, X_n$是总体$\mathbb{N}\left(\mu, \sigma^2\right)$的样本,样本均值和样本方差分别为$\overline{X}$和$s^2$, 则有

- 均值抽样分布

$$\overline{X} \sim \mathbb{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- 方差抽样分布

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

卡方分布示例

【例】设总体$X \sim N\left(\mu, \sigma^2\right)$, 其中$\mu, \sigma$已知.$X_1, \cdots, X_n$是取自总体$X$的样本.

1. 求统计量$X' = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_2 - \mu)^2$的分布

2. 设$n = 5, a(X_1 - X_2)^2 + b(2X_3 - X_4 - X_5)^2 \sim \chi^2(k)$, 则$a, b, k$分别为多少

解: (1)令$Y_i = \frac{X_i - \mu}{\sigma}$, 显然$Y_i$独立且$Y_i \sim N(0,1)$, 所以

$$X' = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_2 - \mu)^2 = \sum_{i=1}^{n} Y_i^2 \sim \chi^2(n)$$

卡方分布示例

(2)$X_1 - X_2 \sim N\left(0,2\sigma^2\right)$,所以$\frac{(X_1-X_2)^2}{2\sigma^2} \sim \chi^2(1)$

$2X_3 - X_4 - X_5$\sim $N(0,6\sigma^2)$, 所以, $\frac{(2X_3-X_4-X_5)^2}{6\sigma^2} \sim \chi^2(1)$

$X_1 - X_2$与$2X_3 - X_4 - X_5$独立, 则有

$$\frac{\left((X_1 - X_2)^2\right)}{2\sigma^2} + \frac{(2X_3 - X_4 - X_5)^2}{6\sigma^2} \sim \chi^2(2)$$

所以$a = \frac{1}{2\sigma^2}, b = \frac{1}{6\sigma^2}, k = 2$

t分布的故事

- 戈塞特(W.S.Gosset)在酿酒厂工作期间考虑酿酒配方实验中的统计学问题，并追随卡尔·皮尔逊学习统计学，提出了t分布。
- 戈塞特笔名是学生氏(Student)，因此该分布也成为学生分布，或者简称t分布。
- 1908年，戈塞特提出了正态样本中样本均值和标准差的比值的分布，并给出了应用上极其重要的第一个分布表。戈塞特在t分布的工作是开创了小样本统计学的先河。

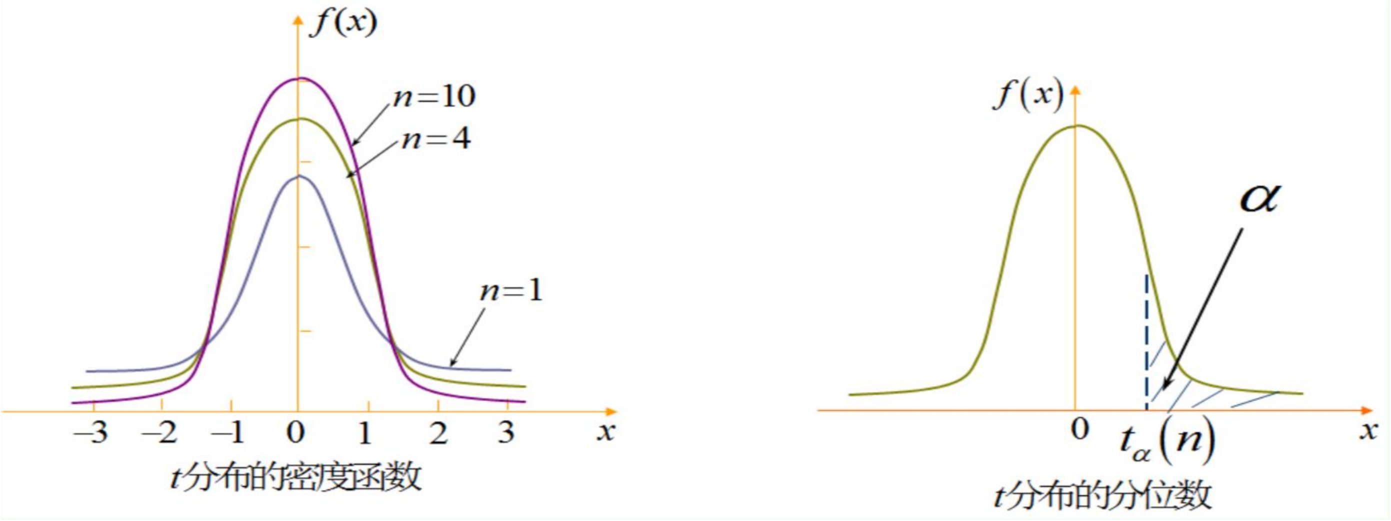
t分布

【t分布】设$X \sim N(0;1), Y \sim \chi^2(n)$, 且$X, Y$相互独立,则称随机变量

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

所服从的分布为自由度为n的t分布, 记为$t(n)$.

t分布性质



对于给定的$0 < \alpha < 1$,其上分位点$t_\alpha(n)$可以通过t分布表查询可得。由t分布的对称性可知$t_{1-\alpha}(n) = -t_\alpha(n)$。$n$足够大时, t分布近似于标准正态分布。

$$\lim_{n\to\infty} f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

更多详情，请咨询课程助教。

正态样本均值和方差的抽样分布

【定理】 设$X_1,\cdots,X_n$是总体$N(\mu,\sigma^2)$的样本,$\overline{X}$,S^2分别是样本均值和样本方差, 则

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

由$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$,$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$可知

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(\mathrm{n}-1)S^2}{\sigma^2(n-1)}} \sim t(n-1)$$

F分布的故事

- F分布就是为了纪念费希尔(R.A.Fisher)而用他的名字首字母命名的。F分布在方差分析等领域有重要用途。
- 费希尔对现代统计学具有卓越贡献。创立的极大似然估计，被尊为统计学参数估计的最重要的方法。

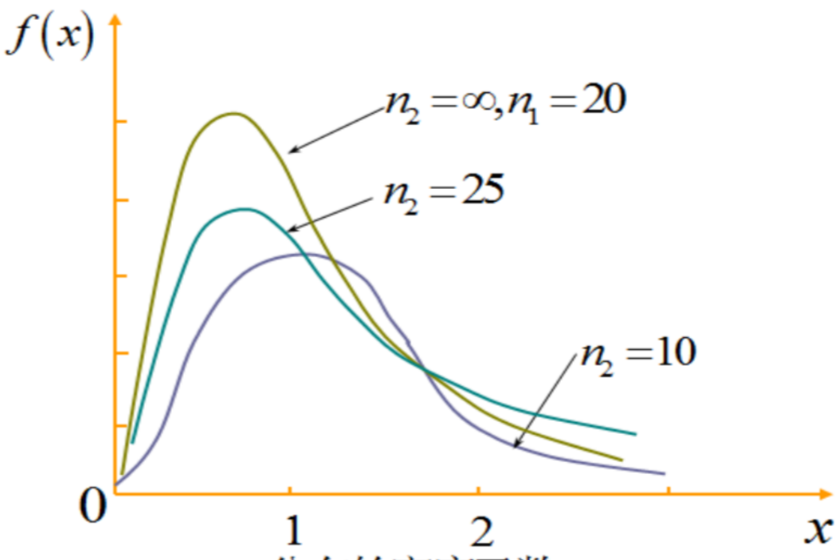
F分布

【F分布】设$X \sim \chi^2(n_1), Y \sim \chi^2(n_2)$, 且X, Y相互独立,则称随机变量

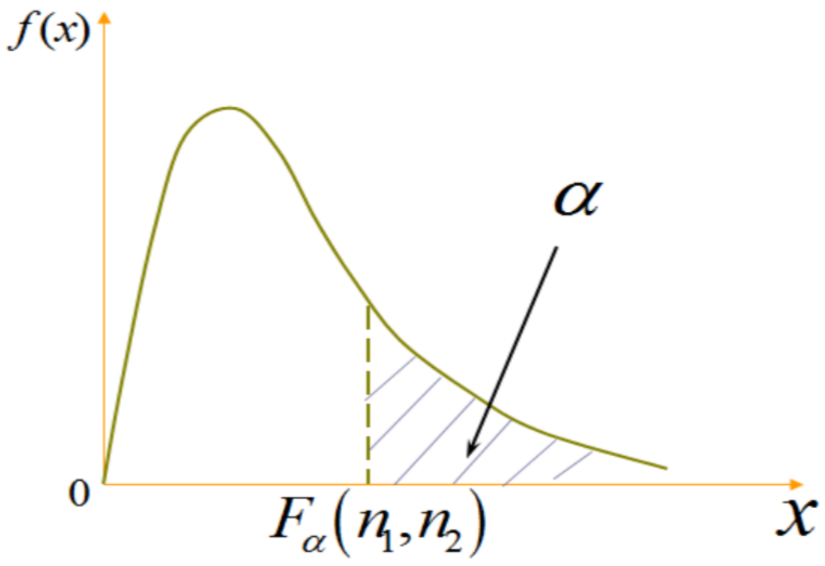
$$F = \frac{\frac{X}{n_1}}{\frac{Y}{n_2}} = \frac{n_2}{n_1}\frac{X}{Y}$$

所服从的分布为自由度为n_1, n_2)的F分布，记为$F(n_1, n_2)$,其中n_1为第一自由度, n_2为第二自由度。

F分布性质



F分布的密度函数



F分布的分位数

对于给定的$0<\alpha<1$,其上分位点$F_\alpha(n_1,n_2)$可以通过F分布表查询可得.

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

正态样本方差抽样分布

【定理】设 $X_1,\cdots,X_{n_1}$ 与 $Y_1,\cdots,Y_{n_2}$ 是分别来自具有相同方差的两正态总体 $N(\mu_1,{\sigma_1}^2)$，$N(\mu_2,{\sigma_2}^2)$ 的两个独立样本，s_1^2,s_2^2 分别为两个样本的方差，则

$$\frac{s_1^2}{s_2^2} \sim F(n_1-1,n_2-1)$$

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1}{n_1}+\frac{\sigma_2}{n_2}}} \sim N(0,1)$$

正态样本方差抽样分布

【定理】 设 $X_1,\cdots,X_{n_1}$ 与 $Y_1,\cdots,Y_{n_2}$ 是分别来自两正态总体 $N(\mu_1,\sigma_1^2)$，$N(\mu_2,\sigma_2^2)$的两个独立样本,$\overline{X},\overline{Y}$分别为两样本均值,$s_1^2,s_2^2$分别为两样本方差, 则

$$\frac{(s_1^2)/(s_2^2)}{(\sigma_1^2)/(\sigma_2^2)} \sim F(n_1-1,n_2-1)$$

更多详情，请咨询课程助教。