



西安交通大学

XI'AN JIAOTONG UNIVERSITY

第五章 数理统计的基本概念

言涪 特聘研究员
网络空间安全学院

2025年4月

课程相关信息

课程资料地址：

<https://github.com/liyan2015/MATH200327>

课程讨论区：

<https://class.xjtu.edu.cn/course/93430/forum#/>

<https://class.xjtu.edu.cn/course/93436/forum#/>



课程资料



课程讨论区1



课程讨论区2

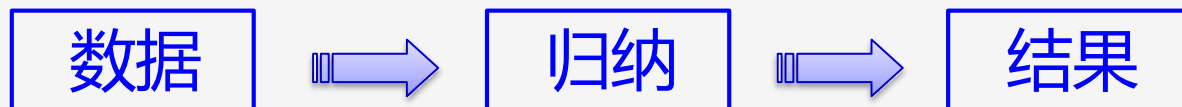
从宿舍到教室需要花多少时间?

相信大家心里对此都有一个大概的“数”。



问题 你是怎么得到这个“数”的?

这就是一个典型的统计思维过程



数理统计就是一个归纳推断过程

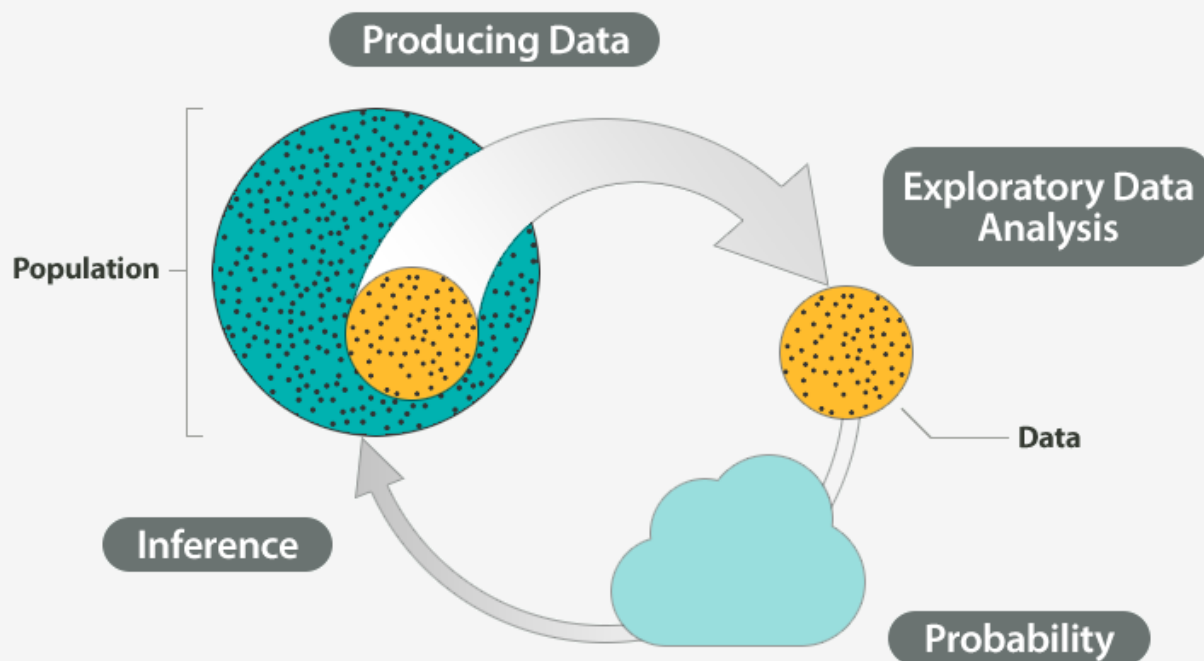
数理统计是以概率论为基础, 关于**实验数据**的**收集、整理、分析与推断**的一门科学与艺术

● **问题** 什么是实验数据?

科学试验, 或对某事物、现象进行观察获得的数据称为**试验数据**

● **特点** 数据受随机因素的影响
--可以通过某种概率分布来描述

统计推断：
研究如何
有效分析
已获得的
随机数据



从样本中得到数据 ➡ 对数据进行分析 ➡ 根据概率推断样本特性

2016

US NEWS



Clinton calls off Election Night fireworks

By Natalie Musumeci

Published Nov. 7, 2016 | Updated Nov. 7, 2016, 12:41 p.m. ET



皮尤研

One likely kinds of pe when certain rtunity outreach to all parts of the electorate. We know that some groups – including the less

● 对三组各100个零件进行有损抽样检查，结果如下：

A. 抽样10个零件，合格率0.5；

B. 抽样50个零件，合格率0.5；

C. 抽样90个零件，合格率0.5。

● 哪组未抽样零件的合格率在 $[0.4, 0.6]$ 区间的概率最高？

● **问题** 实验数据的处理过程?

数据 收集, 整理, 分析, 推断

《数理统计》围绕这
四个过程来进行研究

本讲主要介绍数据“收集”和“整理”
环节中的一些相关概念



西安交通大学
XI'AN JIAOTONG UNIVERSITY

5.1 总体与样本

5.1 总体与样本

- **总体** 研究对象的全体称为总体
- **个体** 总体中的每一个具体对象称为个体

例 分析某班级学生的英语考试成绩

总体 -- 该班级所有学生的英语考试成绩

个体 -- 每一个学生的英语考试成绩

5.1 总体与样本

例 分析某工厂生产的灯泡的使用寿命



总体 -- 该厂生产的所有灯泡的使用寿命

个体 -- 每一个灯泡的使用寿命



5.1 总体与样本

总体 研究对象的数量指标 X

$$X \sim F(x)$$

个体 总体 X 的可能取值

例 分析某工厂生产的灯泡的使用寿命

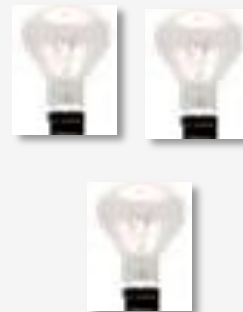
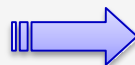
总体 -- 该厂生产的所有灯泡的使用寿命 X

$$X \sim N(\mu, \sigma^2)$$

个体 -- 每一个灯泡的使用寿命, 即 X 的一个可能取值

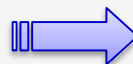
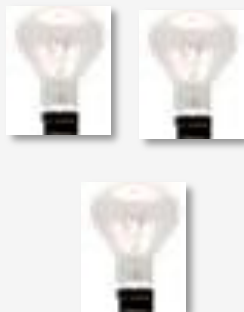
5.1 总体与样本

● **问题** 如果对总体完全了解的情况下，能否对个体进行预测？



5.1 总体与样本

● **问题** 如果知道部分个体的值, 能否预测总体?



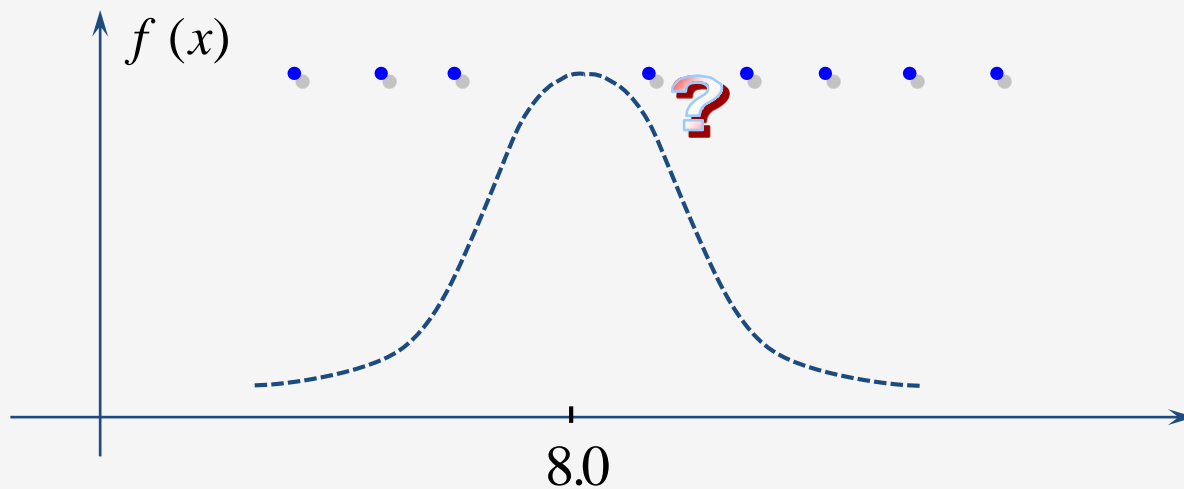
灯泡寿命 $X \sim N(\mu, \sigma^2)$

$\mu = ? \sigma^2 = ?$

5.1 总体与样本

例 设某台机床加工的零件的长度 $X \sim N(\mu, 0.1)$
实测了其中 8 个零件, 得到它们的测量值为

8.3, 7.7, 8.6, 8.0, 8.6, 7.7, 8.6, 8.0



定义1 从总体 X 中抽取的部分个体, 得到的数量指标 X_1, X_2, \dots, X_n , 若满足下条件:

(1) X_1, X_2, \dots, X_n 与 X 同分布;

(2) X_1, X_2, \dots, X_n 相互独立.

则称 X_1, X_2, \dots, X_n 是来自总体 X 的一个简单随机样本, 简称样本.

5.1 总体与样本

对样本 X_1, X_2, \dots, X_n 进行观测后, 得到的观测值: x_1, x_2, \dots, x_n 称为**样本观测值**.

注:

观测前: X_1, X_2, \dots, X_n 是随机变量;

观测后: x_1, x_2, \dots, x_n 是具体的数据.

样本的联合分布

设总体 $X \sim F(x)$, 则样本 X_1, X_2, \dots, X_n 的联合分布函数为:

$$F(x_1, \dots, x_n) = P\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \prod_{i=1}^n F(x_i)$$

若总体 X 的密度函数为 $f(x)$, 则样本 X_1, X_2, \dots, X_n 的联合密度函数为:

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

样本的联合分布

若总体 X 的分布律为：

$$P\{X = a_k\} = p_k, \quad k = 1, 2, \dots$$

则样本 X_1, X_2, \dots, X_n 的联合分布律为：

$$\begin{aligned} & P\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} \\ &= P\{X_1 = x_1\}P\{X_2 = x_2\} \cdots P\{X_n = x_n\} \\ &= \prod_{i=1}^n P\{X = x_i\} \end{aligned}$$

5.1 总体与样本

例 设 X_1, X_2, \dots, X_n 是来自总体 $N(\mu, \sigma^2)$ 的样本, 则样本的联合密度函数为:

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= \prod_{i=1}^n f(x_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2} \end{aligned}$$

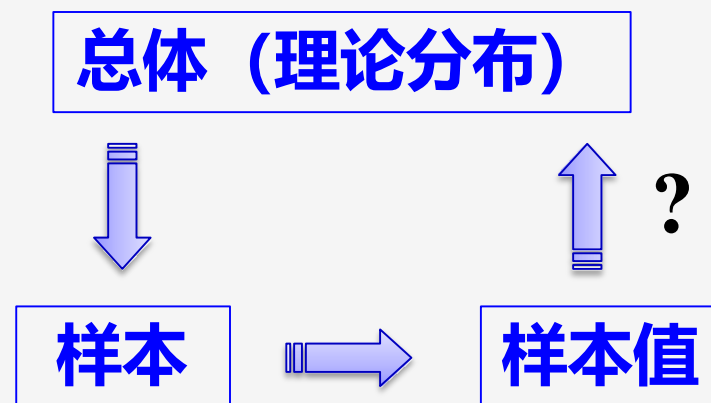
5.1 总体与样本

例 设 X_1, X_2, \dots, X_n 是来自总体 $B(1, p)$ ($0 < p < 1$) 的样本, 则样本的联合分布律为:

$$\begin{aligned} P\{X_1 = x_1, \dots, X_n = x_n\} &= \prod_{i=1}^n P\{X = x_i\} \\ &= \prod_{i=1}^n p^{x_i} p^{(1-x_i)} \\ &= p^{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n (1-x_i)} \end{aligned}$$

5.1 总体与样本

总体、样本、样本值的关系



对样本的一些认识

设 X_1, X_2, \dots, X_n 是来自总体 $X \sim F(x)$ 的样本

1. X_1, X_2, \dots, X_n 是一堆“杂乱无章”的数据；
2. X_1, X_2, \dots, X_n 包含总体的相关“信息”；
3. X_1, X_2, \dots, X_n 是对总体进行推断的依据；
4. 观测前, X_1, X_2, \dots, X_n 是 *i.i.d.* 随机变量,
观测后, x_1, x_2, \dots, x_n 是具体的数据.



西安交通大学
XI'AN JIAOTONG UNIVERSITY

5.2 常用的统计量

● 统计推断的基础：收集数据

从总体 $X \sim F(x)$ 中抽取样本：

$$X_1, X_2, \dots, X_n$$

“杂乱无章”的数据

包含了有用的“信息”

● 问题

如何提炼出有用的“信息”？

5.2 常用的统计量

例 设某班级英语考试后，全班同学的成绩分别为： X_1, X_2, \dots, X_n

● **问题** 你除了希望知道自己的成绩外，还关心哪个成绩？

● **问题** 如何评价该班级的英语整体学习情况？

$$\max\{X_1, X_2, \dots, X_n\} \quad \frac{1}{n} \sum_{i=1}^n X_i$$

--对样本进行“整理”后得到的数据

数据的整理：统计量

定义2 设 X_1, X_2, \dots, X_n 是来自总体 $X \sim F(x)$ 的样本, $g(x_1, x_2, \dots, x_n)$ 是 n 元实值连续函数, 若函数 $g(x_1, x_2, \dots, x_n)$ 不含未知参数, 则称随机变量 $g(X_1, X_2, \dots, X_n)$ 为统计量

5.2 常用的统计量

例 设 X_1, X_2, \dots, X_n 是来自总体 $N(\mu, \sigma^2)$ 的样本, 其中 μ, σ^2 均未知。则下列中哪些是统计量?

$$\frac{1}{n} \sum_{i=1}^n X_i, \quad \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2, \quad \min\{X_1, X_2, \dots, X_n\}$$

常用统计量

设 X_1, X_2, \dots, X_n 是来自总体 X 的样本, 则称

$$(1) \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{为样本均值}$$

$$(2) \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{为样本方差}$$

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{为样本标准差}$$

5.2 常用的统计量

$$(3) A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

为样本的 k 阶原点矩

$$(4) B_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

为样本的 k 阶中心矩

5.2 常用的统计量

(5) 将样本 X_1, X_2, \dots, X_n 按由小到大的顺序排成

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

则称统计量 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$

为顺序统计量

称 $X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$

为样本极小值

称 $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$

为样本极大值

称 $R_n = X_{(n)} - X_{(1)}$

为样本极差

样本均值与样本方差的数字特征

命题1 设 X_1, X_2, \dots, X_n 是来总体 X 的样本, 且总体的均值与方差存在, 记为

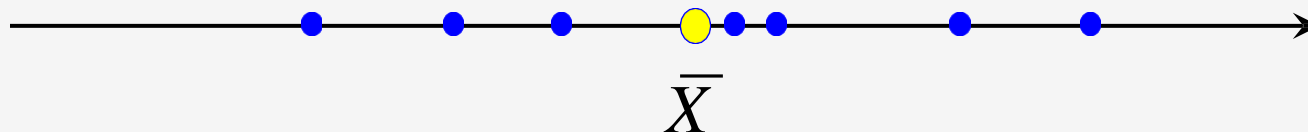
$$E(X) = \mu, \quad D(X) = \sigma^2$$

则有

- (1) $E(\bar{X}) = \mu, \quad D(\bar{X}) = \frac{1}{n} \sigma^2$
- (2) $E(S^2) = \sigma^2$

样本均值与样本方差的含义

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ — 是观测数据 X_1, X_2, \dots, X_n 的平均值
是观测数据 X_1, X_2, \dots, X_n 的 “中心”

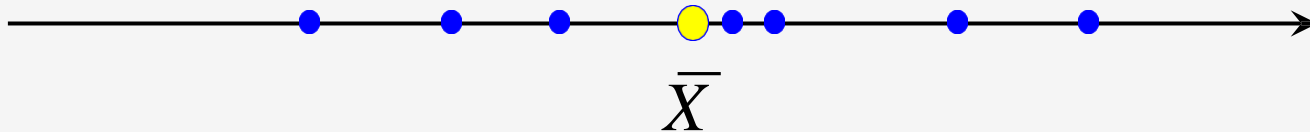


样本均值与样本方差的含义

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

—— 反映了观测数据 X_1, X_2, \dots, X_n 与观测数据
中心点的偏离程度

反映了观测数据 X_1, X_2, \dots, X_n 的离散程度



问题

下结果说明了什么？

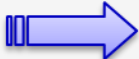
$$E(\bar{X}) = \mu, \quad D(\bar{X}) = \frac{\sigma^2}{n}, \quad E(S^2) = \sigma^2$$

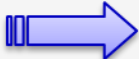


西安交通大学
XI'AN JIAOTONG UNIVERSITY

5.3 常见的抽样分布

5.3 常见的抽样分布

数据收集  样本、样本观测值
—包含了总体的有用信息

数据整理  统计量
—提炼数据中包含的信息

统计量 $g(X_1, X_2, \dots, X_n)$ 是随机变量

● 确定统计量的分布是概率统计的基本问题之一

5.3 常见的抽样分布

定义1 统计量 $g(X_1, X_2, \dots, X_n)$ 的分布称为**抽样分布**

本讲主要介绍与标准正态总体相关的抽样分布：

χ^2 - 分布 t - 分布 F - 分布

5.3 常见的抽样分布

一、 χ^2 - 分布

定义1 设 X_1, X_2, \dots, X_n 相互独立, 且都服从标准正态分布 $N(0, 1)$, 则称随机变量

$$X_1^2 + X_2^2 + \dots + X_n^2$$

服从自由度为 n 的 χ^2 分布, 记为

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

5.3 常见的抽样分布

χ^2 -分布的密度函数

$$p(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} e^{-\frac{x}{2}} x^{\frac{n}{2}-1}, & x > 0 \\ 0 & x \leq 0. \end{cases}$$

其中

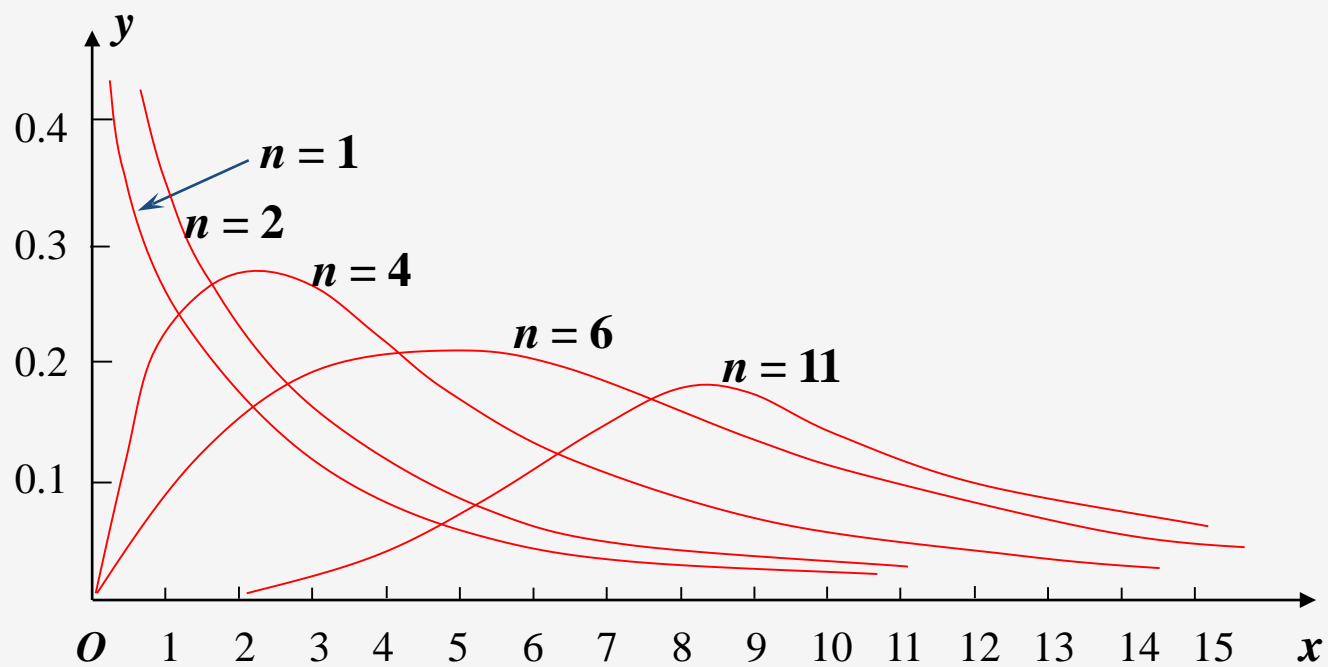
$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt, \quad x > 0$$

称为 Γ 函数, 具有性质

$$\Gamma(x+1) = x\Gamma(x), \quad \Gamma(1) = 1, \quad \Gamma(n+1) = n!$$

5.3 常见的抽样分布

χ^2 -分布密度函数的图像



5.3 常见的抽样分布

χ^2 -分布的性质

1. 可加性

若 $Y_1 \sim \chi^2(n)$, $Y_2 \sim \chi^2(m)$, 且 Y_1, Y_2 相互独立, 则有

$$Y_1 + Y_2 \sim \chi^2(n + m)$$

推广: 若 Y_1, Y_2, \dots, Y_k 相互独立, 且

$Y_i \sim \chi^2(n_i) \quad (i = 1, 2, \dots, k)$ 则有

$$\sum_{i=1}^k Y_i^2 \sim \chi^2\left(\sum_{i=1}^k n_i\right)$$

2. 数字特征

若 $Y \sim \chi^2(n)$, 则有 $E(Y) = n$, $D(Y) = 2n$,

证明 存在独立同分布的 X_1, X_2, \dots, X_n , 都服从标准正态分布 $N(0, 1)$, 使得

$$Y = X_1^2 + X_2^2 + \dots + X_n^2$$

$$\begin{aligned} E(Y) &= E(X_1^2 + X_2^2 + \dots + X_n^2) \\ &= E(X_1^2) + E(X_2^2) + \dots + E(X_n^2) = nE(X_1^2) = n \end{aligned}$$

$$D(X) = E(X^2) - (E(X))^2$$

5.3 常见的抽样分布

$$\begin{aligned} D(Y) &= D(X_1^2 + X_2^2 + \cdots + X_n^2) \\ &= D(X_1^2) + D(X_2^2) + \cdots + D(X_n^2) = nD(X_1^2) \end{aligned}$$

$$D(X_1^2) = E(X_1^4) - [E(X_1^2)]^2 = E(X_1^4) - 1$$

$$\begin{aligned} E(X_1^4) &= \int_{-\infty}^{+\infty} x^4 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 3 \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= 3E(X_1^2) = 3 \end{aligned}$$

$$D(Y) = nD(X_1^2) = n(3-1) = 2n$$

二、 t 分布

定义3 设 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 则称随机变量

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

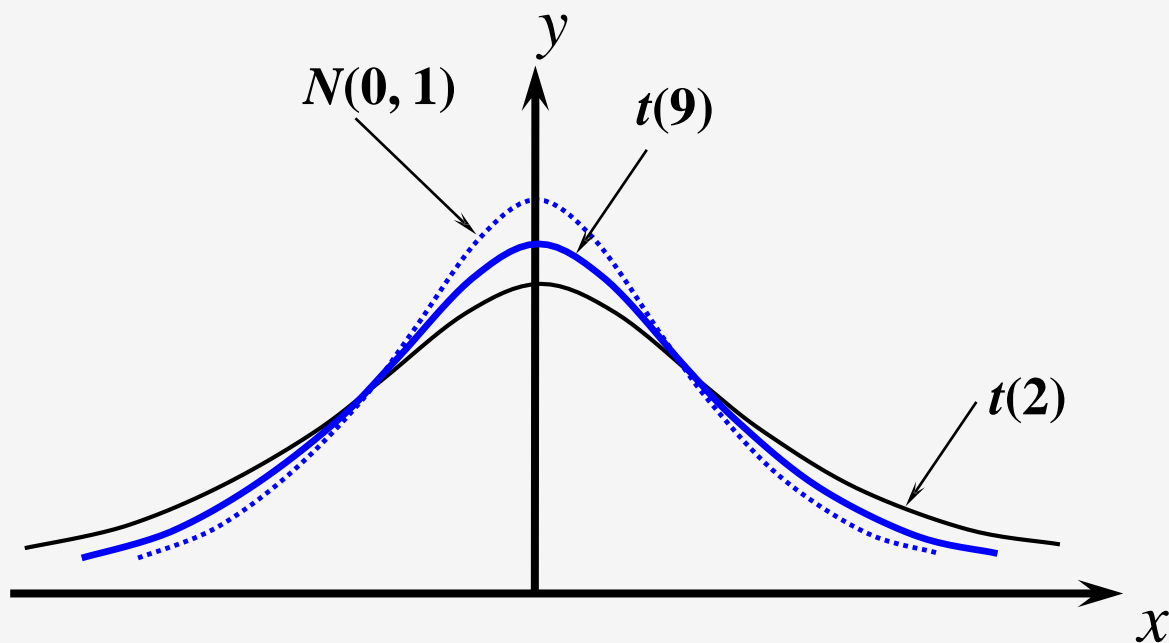
服从自由度为 n 的 t 分布, 记为

$$T \sim t(n)$$

5.3 常见的抽样分布

t 分布的密度函数及其图形

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



三、 F 分布

定义4 设 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 则称随机变量

$$F = \frac{X / m}{Y / n}$$

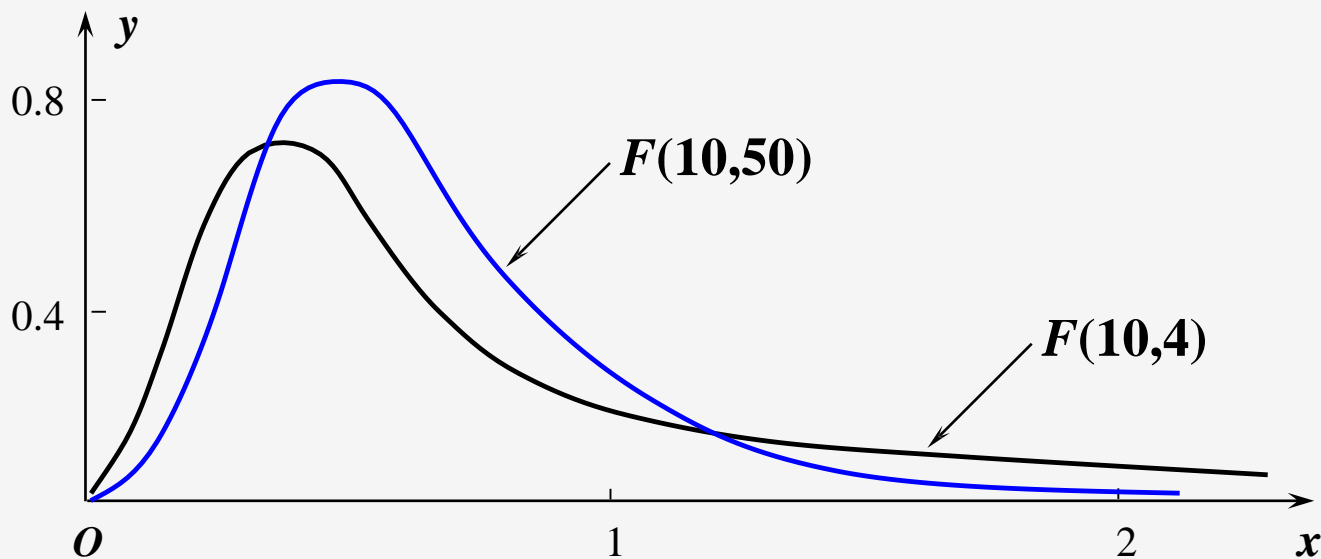
服从自由度为 (m, n) 的 **F 分布**, 记为

$$F \sim F(m, n)$$

5.3 常见的抽样分布

F 分布的密度函数及其图形

$$f(x) = \begin{cases} \frac{F(m,n)[(m+n)/2]}{F(m,n)(m/2)F(m,n)(n/2)} m^{\frac{n_1}{2}} n^{\frac{n_2}{2}} \frac{x^{\frac{n_1}{2}-1}}{(mx+n)^{\frac{n_1+n_2}{2}}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$



5.3 常见的抽样分布

F 分布的一个重要性质

若 $F \sim F(m, n)$, 则 $\frac{1}{F} \sim F(n, m)$

事实上, 因 $F \sim F(m, n)$, 所以由 F 分布定义知, 存在 $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 使得

$$F = \frac{X / m}{Y / n}$$

所以有 $\frac{1}{F} = \frac{Y / n}{X / m} \sim F(n, m)$

四、分位数

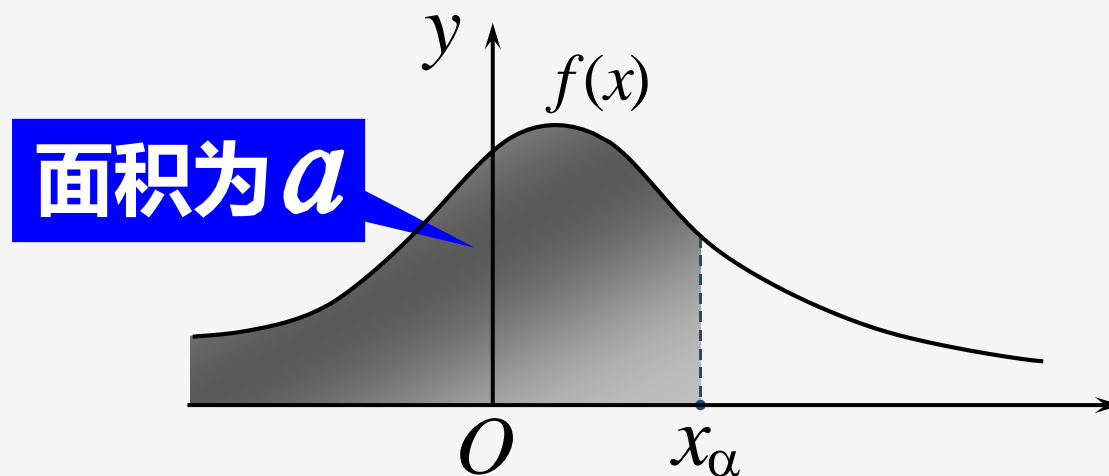
定义4 设连续型随机变量 $X \sim f(x)$, 对给定的 $0 < \alpha < 1$, 存在一个实数 x_α , 使得

$$P\{X \leq x_\alpha\} = \int_{-\infty}^{x_\alpha} f(x)dx = \alpha$$

则称 x_α 为密度函数 $f(x)$ 的 α 分位数

5.3 常见的抽样分布

x_α 为密度函数 $f(x)$ 的 α 分位数



5.3 常见的抽样分布

1. 标准正态分布 $N(0, 1)$ 的 α 分位数记为 u_α

$$\Phi(u_\alpha) = \int_{-\infty}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \alpha$$

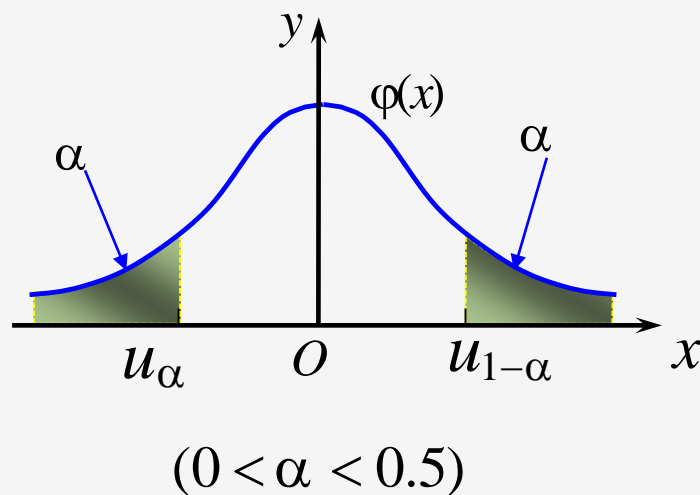
$$u_\alpha = -u_{1-\alpha}$$

查标准正态分布表

$$u_{0.975} = 1.96$$

$$u_{0.95} = 1.645$$

$$u_{0.05} = -1.645$$



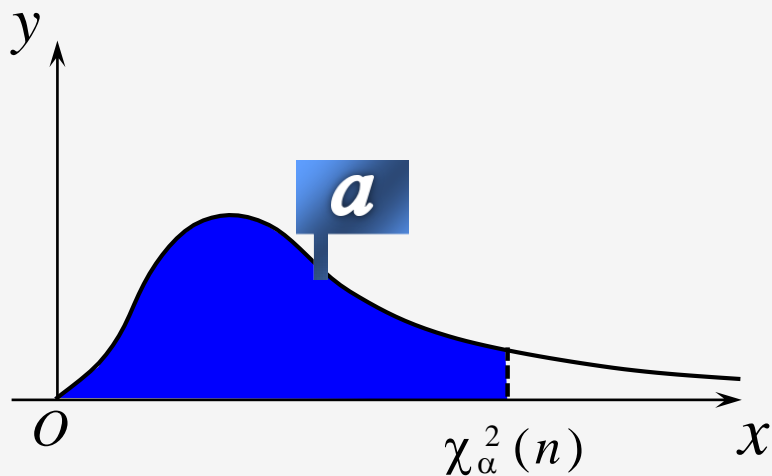
5.3 常见的抽样分布

2. $\chi^2(n)$ 分布的 α 分位数记为 $\chi_{\alpha}^2(n)$.

查 χ^2 分布表

$$\chi_{0.05}^2(10) = 3.940$$

$$\chi_{0.95}^2(10) = 18.307$$



5.3 常见的抽样分布

3. $t(n)$ 分布 α 分位数记为 $t_\alpha(n)$

$$t_\alpha(n) = -t_{1-\alpha}(n)$$

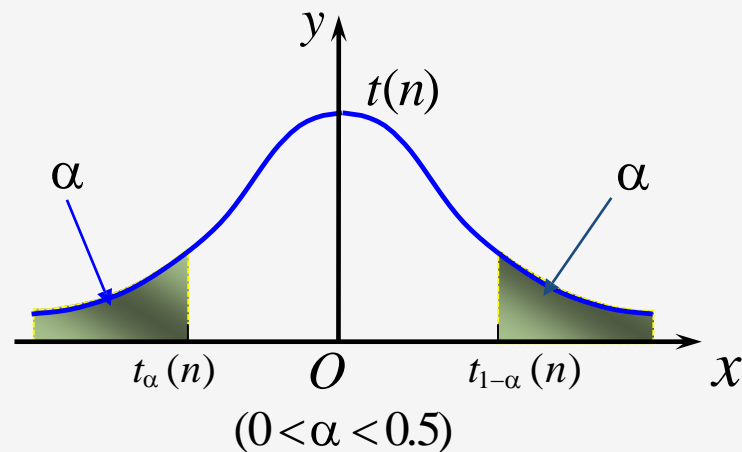
查 t 分布表

$$t_{0.025}(10) = -2.2281$$

$$t_{0.05}(18) = -1.7341$$

$$t_{0.95}(20) = -t_{0.05}(20) = 1.7247$$

当 $n > 45$ 时, $t_\alpha(n) \approx u_\alpha$



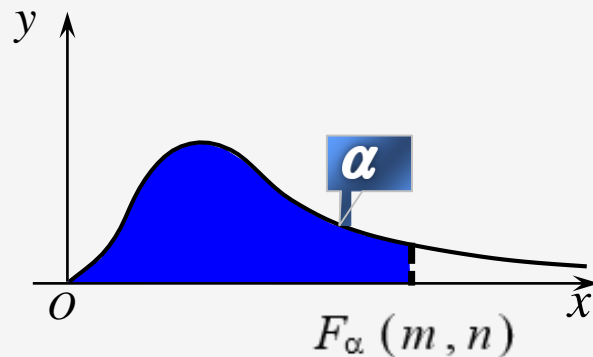
5.3 常见的抽样分布

4. $F(m, n)$ 分布的 α 分位数记为 $F_\alpha(m, n)$.

查 $F(m, n)$ 分布表

$$F_{0.95}(10, 20) = 2.35$$

$$F_{0.05}(15, 10) = \frac{1}{F_{0.95}(10, 15)} \\ = \frac{1}{2.54}$$



若 $F \sim F(m, n)$, 则

$$\frac{1}{F} \sim F(n, m) \Rightarrow F_\alpha(m, n) = \frac{1}{F_{1-\alpha}(n, m)}$$



西安交通大学
XI'AN JIAOTONG UNIVERSITY

谢谢大家！

