Part III-B: Probability Theory and Mathematical Statistics

Lecture by 李漫漫
Note by THF

2024 年 11 月 5 日

目录

Lecture 13

10.31

Notation. 偏度 r_1：三阶标准化随机变量的矩，用于描述对称性
峰度 r_2：四阶标准化随机变量的矩，一般使正态分布的峰度 $r_2=0$，描述分布的陡峭程度

表 1: 常见分布的数字特征

分布	EX	DX	r_1	r_2
$B(1,p)$	p	$p(1-p)$	$\frac{1-2p}{\sqrt{p(1-p)}}$	$\frac{1}{p(1-p)-6}$
$B(n,p)$	np	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$\frac{1-6p(1-p)}{np(1-p)}$
$P(\lambda)$	λ	λ	$\frac{1}{\sqrt{\lambda}}$	$\frac{1}{\lambda}$
$G(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{2-p}{\sqrt{1-p}}$	$6+\frac{p^2}{1-p}$
$U[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0	$\frac{9}{5}-3$
$\Gamma(1,\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	2	6
$N(\mu,\sigma^2)$	μ	σ^2	0	0

1 数理统计基本概念

随机变量引入：使样本空间映射到实数轴上
分布函数：任意随机变量的概率
大数定律和中心极限定理：由概率论过渡到数理统计

$$\begin{cases} \text{描述统计学：过去的实验数据/相关分析图} \\ \text{推断统计学：根据现有的实验数据决策} \begin{cases} \text{参数估计：第七章} \\ \text{假设检验：第八章} \\ \text{回归分析：第九章} \end{cases} \end{cases}.$$

Definition. 总体：全部研究对象，可以用分布描述（随机变量组）

Definition. 个体：组成总体的成员，符合总体分布（每一个个体都是一个随机变量）

Example. 从总体中抽取 n 个样本

对数据记录：$x_1, x_2, \ldots, x_n$ 称为 n 维随机变量 $X_1, X_2, \ldots, X_n$ 对应的观测值，$X_1, X_2, \ldots, X_n$ 为来自总体 X 的一个样本

Notation. 简单样本：$X_1, X_2, \ldots, X_n$ *i.i.d*，且与总体分布相符

特点：
- 独立性
- 代表性

Definition. 样本空间：$\boldsymbol{\Omega} = \{(x_1, x_2, \ldots, x_n) \,|x_i \in \mathbb{R}, \mathrm{i} = 1, 2, \ldots, n\}$

Notation. 样本联合分布和总体分布的关系 (*i.i.d*)：

$$\begin{aligned} F(x_1, x_2, \ldots, x_n) &= P\{X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n\} \\ &= \prod_{i=1}^{n} P\{X_i \le x_i\} \\ &= \prod_{i=1}^{n} F(x_i). \end{aligned}$$

扩展：X 为连续型，密度函数的关系：

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f_{X_i}(x_i) \\ &= \prod_{i=1}^{n} f(x_i) \quad x_i \in \mathbb{R}, i = 1, 2, \ldots, n. \end{aligned}$$

1.1 经验分布函数

经验分布函数：$F_n(x)$

将样本观测值 $x_1, x_2, \ldots, x_n$ 按大小分类为 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$

$$\begin{aligned} F_n(x) &= f_n\{X \le x\} \\ &= \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x \in [x_{(k)}, x_{(k+1)}) \\ 1, & x \ge x_{(n)} \end{cases} \\ &\approx F(X). \end{aligned}$$

Corollary. 格利文科定理：

$$P\left\{\limsup_{n\to\infty, x\in\mathbb{R}} |F(x) - F_n(x) = 0|\right\} = 1.$$

根据格利文科定理：可以使用经验分布函数来估计理论分布函数

Lecture 14

11.05

1.2 密度函数

Notation. 密度函数和分布函数的关系：

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x \\ f_X(x) &= \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}. \end{aligned}$$

对于直方图：将中点光滑连接 = 密度函数

或：核密度

直方图

Notation. 直方图的面积代表频率：

$$\text{高度}h_i = \frac{\text{面积}f_i}{\text{区间长度}\Delta x_i}.$$

直方图的高度代表密度，直方图的横坐标的取值范围为观测值的取值范围，直方图分块的区间来源一般为经验公式：$m \approx 1.87(n-1)^{0.4}$，其中 m 为区间分组数量

计算直方图频率：

$$\text{频率}f_i = \frac{\text{落入区间的个数}y_i}{\text{总个数}y}.$$

1.3 统计量

统计量 (statistic)，统计学 (statistics)

Definition. 统计量：关于样本的函数，不含任何未知参数

完整定义：

Example. X_1, X_2 来自正态总体 $N\left(\mu, \sigma^2\right)$ 的样本（这两个任意抽出一个都属于一个样本），其中 μ, σ 均未知，以下表达式：

- $\frac{1}{4}\left(X_1 + X_2\right) - \mu$
- $\frac{X_1}{\sigma}$

均不是统计量（使用了未知的数），以下表达式都是统计量

- $3X_1$
- $X_1 - 8$
- $X_1^2 + X_2^2$

提出统计量的目的：通过样本估计或检测未知量，因此统计量不能含未知量

常见统计量：

- 样本均值（算术平均数）：$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 样本方差：$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$
- 样本标准差：$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$
- 样本阶原点矩
- 样本阶中心距

Notation. 样本均值：若 $X_1, X_2, \ldots X_n$ *i.i.d*：根据辛钦大数定律：$\bar{X} \xrightarrow[n\to+\infty]{P} E\bar{X} = EX$

Notation. 样本阶中心矩：

$$B_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k \xrightarrow[n\to\infty]{P} DX.$$

或：$S^2 = B_2 \times \frac{n}{n-1} \Rightarrow E\left(S^2\right) = DX$

证明.

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2 - 2X_i\bar{X}+\bar{X}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}Xi^2\right) - \sum_{i=1}^{n}\bar{X}^2\\
ES^2 &= E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\bar{X}^2\right)\right]\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}EX_i^2 - nE\bar{X}^2\right)\\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(DX_i+(EX_i)^2\right) - n\left(D\bar{X}+\left(E\bar{X}\right)^2\right)\right]\\
&= \frac{1}{n}\left[\sum_{i=1}^{n}\left(DX+(EX)^2\right) - n\left(\frac{DX}{n}+(EX)^2\right)\right]\\
&= \frac{1}{n-1}\left[nDX+n(EX)^2 - DX - n(EX)^2\right] = \frac{1}{n-1}(n-1)DX = DX.
\end{aligned}$$

即: $EB_2 = E\left(\frac{n-1}{n}S^2\right) = \frac{n-1}{n}DX$ □

用样本均值估计总体均值:

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

顺序统计量

令 $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ 为最小顺序统计量，最大同理
要求第几小的顺序统计量: R 成为样本极差，$\widetilde{X}$ 称为样本中位数

1.4 样本均值的分布

Theorem. $X_1, X_2, \ldots, X_n$ 来自 $N\left(\mu, \sigma^2\right)$ ，则

$$\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1).$$

定义 $\bar{X}$ 为 $X_1, X_2, \ldots, X_n$ 的线性函数，$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，计算期望和方差，将 $\bar{X}$ 标准化

Theorem. 标准化后的线性函数 $\frac{\bar{X}-\mu}{\sigma}\sqrt{n}$:

$$\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \xrightarrow[n\to\infty]{L} N(0,1).$$

Example. 总体: $X \sim N(20, 9)$ ，求样本容量 n 多大时使样本均值与总体均值的绝对值之差 ≤ 0.3 的概率 $> 95\%$

1.4.1 三大抽样分布

- 卡方分布: $\chi^2(n)$

Notation. 卡方分布实际上为 $\alpha=\frac{1}{2},\lambda=\frac{n}{2}$ 的 Gamma 分布

当 $n=2$ 时为参数为 $\frac{1}{2}$ 的指数分布

一般称 $f(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\lambda x}$ 为伽马分布族

Definition. 设 $X_1,X_2,\ldots,X_n\sim N(0,1)$ *i.i.d* ，令 $\chi^2=\sum_{i=1}^n X_i^2$ ，称 χ^2 为自由度为 n 的卡方分布

Notation. 卡方分布具有可加性:

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

从 $n=3$ 开始，卡方分布出现最大值，且 n 越大卡方分布的方差越大

卡方分布的性质:

- $E\left(\chi^2\right)=n,D\left(\chi^2\right)=2n$
- 可加性
- 分位点:

对性质 1:

证明.

$$\begin{aligned}E\left(\sum_{i=1}^n X_i^2\right)&=\sum_{i=1}^n E(X_i^2)=nEX^2\\&=n\left(DX+(EX)^2\right)\\D\left(\sum_{i=1}^n X_i^2\right)&=nDX^2=n\left(E\left(X^2\right)^2-\left(EX^2\right)^2\right)\\&=n\left(EX^4-1\right).\end{aligned}$$

□