# Lec-19. 数理统计介绍、随机样本、统计量

主讲教师：吴利苏 (**wulisu@sdust.edu.cn**)
主　　页：wulisu.cn

# 本章内容

1. 数理统计部分介绍

2. 统计量与常用统计量

3. 三个重要抽样分布

## 数理统计

- 数理统计是以概率论为基础，根据试验或观察数据，来研究随机现象，对研究对象的客观规律做出合理的估计和判断.

## 数理统计

- 数理统计是以概率论为基础，根据试验或观察数据，来研究随机现象，对研究对象的客观规律做出合理的估计和判断.
- 数理统计主要内容：

# 数理统计

- 数理统计是以概率论为基础，根据试验或观察数据，来研究随机现象，对研究对象的客观规律做出合理的估计和判断.
- 数理统计主要内容：
  - 数据收集（获取、预处理、数据清洗）；

# 数理统计

- 数理统计是以概率论为基础，根据试验或观察数据，来研究随机现象，对研究对象的客观规律做出合理的估计和判断.
- 数理统计主要内容：
  - 数据收集（获取、预处理、数据清洗）；
  - 数据处理（聚类分析、特征分析、降维分析、主成分分析、特征提取）；

# 数理统计

- 数理统计是以概率论为基础，根据试验或观察数据，来研究随机现象，对研究对象的客观规律做出合理的估计和判断.
- 数理统计主要内容：
  - 数据收集（获取、预处理、数据清洗）；
  - 数据处理（聚类分析、特征分析、降维分析、主成分分析、特征提取）；
  - 结果分析（统计推断）.

## 概率论和数理统计

- 在概率论中，已知随机变量的分布的前提下，研究它的性质、特点和规律性. 例如，求数字特征、求随机变量函数的分布等等.
- 在数理统计中, 随机变量的分布未知或者部分未知，通过数据对随机变量作出推断.

## 数理统计学习内容

**Chap-6** 总体、随机样本、统计量、常用的统计量和抽样分布

- $\chi^2$ 分布、$t$ 分布、$F$ 分布.

**Chap-7** 估计问题

- 点估计：分布函数已知，参数未知，估计未知参数.
  - ★ 矩估计、极大似然估计.
- 区间估计：对参数处在某个区间的可信程度的估计.

**Chap-8** 假设检验问题

- 分布函数未知，或分布函数只知道形式而参数未知，提出某些假设，进行检验推断.

## 总体和个体

- 总体　试验的全部可能的观察值;
- 个体　总体中的每个可能观察值;
- 总体的容量　总体中所包含的个体数;
- 有限总体　容量有限的总体;
- 无限总体　容量无限的总体，通常将容量非常大的有限总体也按无限总体处理.

## 例

- 研究 2000 名学生的年龄, 这些学生的年龄的全体构成一个总体, 每个学生的年龄就是个体.
- 考察一湖泊中某种鱼的含汞量, 所得总体是有限总体.
- 考察全国正在使用的某种型号灯泡的寿命所形成的总体. 由于可能观察值的个数很多, 可认为是无限总体.
- 一城市空气质量，PM2.5 值, 无限总体.

## 总体分布

- 实际中人们通常只关注总体的某个（或几个）指标.
- 总体的某个指标 $X$, 对于不同的个体来说有不同的取值, 这些取值构成一个分布, 因此 $X$ 可以看成一个随机变量.
- 有时候直接将 $X$ 称为总体. 假设 $X$ 的分布函数为 $F(x)$, 也称总体 $X$ 具有分布 $F(x)$.

## 如何推断总体分布的未知参数（或分布）?

在实际中, 总体的分布未知, 或总体的分布已知, 但某些参数未知. 要对总体进行推断, 研究所有个体是不可能的, 故须抽出部分个体进行研究.

- 样本　从总体中抽出的部分个体.
- 样本容量　样本中所含个体的个数..

- 简单随机样本 满足以下两个条件的随机样本 $(X_1, ..., X_n)$ 称为容量是 $n$ 的简单随机样本.
  - 代表性：每个 $X_i$ 与 X 同分布;
  - 独立性：$X_1, ..., X_n$ 是相互独立的随机变量.
- 样本值 $X_1, ..., X_n$ 的观察值 $x_1, ..., x_n$.

[注]: 后面提到的样本均指简单随机样本。

## 简单随机抽样

- 获得简单随机样本的抽样称为简单随机抽样.

如何进行简单随机抽样?

- 对于有限总体, 采用放回抽样.
- 对于无限总体, 一般采取不放回抽样.
- 但当总体容量很大的时候, 放回抽样有时候很不方便, 因此在实际中当总体容量比较大时, 通常将不放回抽样所得到的样本近似当作简单随机样本来处理.

## 样本的分布函数和概率密度函数

若 $X_1, ..., X_n$ 是总体 $X$ 的样本, $X$ 的分布函数 $F(x)$. 则 $X_1, ..., X_n$ 的联合分布函数

$$F^*(X_1, ..., X_n) = \prod F(x_i).$$

又若 $X$ 具有概率密度 $f$, 则 $X_1, ..., X_n$ 的联合概率密度为

$$f^*(x_1, ..., x_n) = \prod f(x_i).$$

**例**

设一批灯泡的寿命 $X$( 小时) 服从参数为 $\theta$ 的指数分布，$\theta$ 未知. 从该批灯泡中采用简单随机抽样抽取容量为 10 的样本 $X_1, ..., X_{10}$. 对样本实施观测, 得到样本值为

| | | | | |
|---|---|---|---|---|
| 6394 | 1105 | 4717 | 1399 | 7952 |
| 17424 | 3275 | 21639 | 2360 | 2896 |

写出样本的概率密度.

解：总体 $X \sim Exp(\theta)$, $X_1, ..., X_{10}$ 为来自总体 $X$ 的一个样本, 则 $(X_1, ..., X_{10})$ 的联合概率密度为

$$f^*(x_1, ..., x_{10}) = \prod_{i=1}^{10} f(x_i)$$

$$= \begin{cases} \frac{1}{\theta^{10}} e^{-\frac{1}{\theta}\sum_{i=1}^{10} x_i} & x_1 > 0, .., x_n > 0; \\ 0 & \text{其他}. \end{cases}$$

□

- 如何由已知样本值来估计未知参数 $\theta$?

为了估计指数分布的参数 $\theta$, 进行抽样观测，得到样本 $X_1, ..., X_{10}$ 和样本值

| | | | | |
|---|---|---|---|---|
| 6394 | 1105 | 4717 | 1399 | 7952 |
| 17424 | 3275 | 21639 | 2360 | 2896 |

样本中包含了许多信息。
对于推断总体的参数或分布而言，有些是有用的、重要的信息，有些则并不重要。
上例的样本至少提供了两种信息：

1) 10 个灯泡的平均寿命；–有用且重要的信息
2) 灯泡寿命的序号 (如 6394 是第 1 个).–不重要信息

## 构造统计量

从样本中提取有用的信息来研究总体的分布及各种特征数.–构造统计量.

- 统计量 设 $X_1, ..., X_n$ 是来自总体 $X$ 的一个样本, $g(X_1, ..., X_n)$ 是 $X_1, ..., X_n$ 的函数, 若 $g$ 中不含未知数, 则称 $g(X_1, ..., X_n)$ 是一个统计量.

注: $X_1, ..., X_n$ 是随机变量, 而统计量 $g(X_1, ..., X_n)$ 是随机变量的一个函数. 设 $x_1, ..., x_n$ 是相应于样本 $X_1, ..., X_n$ 的一个样本值, 则称 $g(x_1, ..., x_n)$ 是 $g(X_1, ..., X_n)$ 的观察值.

**例**

设 $X_1, X_2, X_3$ 是来自总体 $N(\mu, \sigma^2)$ 的一个总体, 其中 $\mu$ 已知, $\sigma^2$ 未知, 判断下列各式哪些是统计量.

$$T_1 = X_1, \qquad T_2 = X_1 + X_2 e^{X_3}$$

$$T_3 = \frac{1}{3}(X_1 + X_2 + X_3), \quad T_4 = \max\{X_1, X_2, X_3\}$$

$$T_5 = X_1 + X_2 - 2\mu, \qquad T_6 = \frac{1}{\sigma^2}(X_1^2 + X_2^2 + X_3^2)$$

## 常用的统计量

设 $X_1, ..., X_n$ 是来自总体的一个样本,.

- 样本均值 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
- 样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right);$$

样本标准差

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

## 常用的统计量

- 样本 $k$ 阶原点矩 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ k = 1, 2, ...;$
- 样本 $k$ 阶中心矩 $B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k.$

注: $B_2$ 与 $S^2$ 不一样. 样本方差 $S^2$ 中, 除 $n$ 会低估方差, 为保证无偏性, 修正除 $n-1$.

(Chap7-3)

## 当总体数字特征未知时

- 用样本均值 $\overline{X}$ 估计总体均值 $\mu = E(X)$;
- 用样本方差 $S^2$ 估计总体方差 $\sigma^2 = E(X-\mu)^2$;
- 用样本原点矩 $A_k$ 估计总体原点矩 $\mu_k = E(X^k)$;
- 用样本中心矩 $B_k$ 估计总体中心矩 $\nu_k = E(X-\mu)^k$.

这些非常直观的想法，有什么理论依据吗？

这部分内容我们会在 Chap7 中介绍。

**性质**

若总体 $X$ 的 $k$ 阶矩 $E(X^k) = \mu_k$ 存在, 则当 $n \to \infty$ 时, 有 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k$.

## 性质

若总体 $X$ 的 $k$ 阶矩 $E(X^k) = \mu_k$ 存在, 则当 $n \to \infty$ 时, 有 $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k$.

证明: 由于 $X_1, ..., X_n$ 独立且同 $X$ 同分布,

$$E(X_1^k) = E(X_2^k) = ... = E(X_n^k) = \mu_k,$$

由辛钦大数定律知, $A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mu_k$. □

进一步由依概率收敛的性质知

$$g(A_1, ..., A_k) \xrightarrow{P} g(\mu_1, ..., \mu_k),$$

其中 $g$ 是连续函数. Chap7 矩估计的理论依据.

## 经验分布函数

### 定义

设 $x_1, ..., x_n$ 是来自分布函数 $F(x)$ 的总体 $X$ 的样本观察值. $X$ 的经验分布函数 $F_n(x)$ 定义为

$$F_n(x) = \frac{\sharp(x_i \leq x)}{n}, \quad -\infty < x < +\infty$$

其中 $\sharp(x_i \leq x)$ 表示 $x_1, ..., x_n$ 中小于或等于 $x$ 的个数.

注: 由定义, 当给定样本观察值 $x_1, ..., x_n$ 时, $F_n(x)$ 是 $X$ 的函数, 具有分布函数的三个条件:

1. $F_n(x)$ 是 $x$ 的不减函数.
2. $0 \leq F_n(x) \leq 1$, $F(-\infty) = 0$, $F(+\infty) = 1$.
3. $F_n(x)$ 是一个右连续函数.

故 $F_n(x)$ 是一个分布函数. 当 $x_1, ..., x_n$ 各不同时, $F_n(x)$ 是以等概率 $\frac{1}{n}$ 取 $x_1, .., x_n$ 的离散型随机变量的分布函数.

一般地, 设 $x_1, ..., x_n$ 是总体 $X$ 的容量为 $n$ 的样本观察值, 先将 $x_1, .., x_n$ 按自小到大的次序排序, 重新编号为 $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$, 则

$$F_n(x) = \begin{cases} 0 & x < x_{(1)}; \\ \frac{k}{n} & x_{(k)} \leq x < x_{(k+1)} \quad k = 1, 2, ..., n-1;. \\ 1 & x \geq x_{(n)}. \end{cases}$$

**例**

设 $X$ 有样本观察值 $-1, 1, 2$, 则

$$F_3(x) = \begin{cases} 0 & x < -1; \\ \frac{1}{3} & -1 \le x < 1; \\ \frac{2}{3} & 1 \le x < 2; \\ 1 & x \ge 2. \end{cases}$$

当给定 $x$ 时, $F_n(x)$ 是样本 $X_1, ..., X_n$ 的函数, 故它是一个统计量.

**定理 (格里汶科定理)**

设 $X_1, ..., X_n$ 是来自以 $F(x)$ 为分布函数的总体 $X$ 的样本, $F_n(x)$ 是经验分布函数, 则有

$$P\left\{\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\right\} = 1.$$

上面定理表明 $F_n(x)$ 在整个实数轴上以概率 1 均匀收敛于 $F(x)$. 所以, 当 $n$ 很大时, $F_n(x)$ 可以很好地近似总体分布函数 $F(x)$. 这是以样本推断总体的依据.

## 抽样分布

- 统计量的分布被称为抽样分布.
- 当总体 $X$ 服从一般分布 (如指数分布、均匀分布等), 要得出统计量的分布是很困难的.
- 当总体 $X$ 服从正态分布时，统计量 $\overline{X}, S^2$ 是可以计算的，那么服从什么分布呢?
- 下面将介绍数理统计中三个重要的抽样分布—$\chi^2$ 分布, $t$ 分布, $F$ 分布.

## 1. $\chi^2$ 分布

**定义**

设 $X_1, ..., X_n$ 是来自总体 $N(0,1)$ 的样本, 则称统计量

$$\chi^2 = X_1^2 + ... + X_n^2$$

为服从自由度为 $n$ 的$\chi^2$ 分布, 记为 $\chi^2 \sim \chi^2(n)$.

自由度指右端包含独立变量的个数.

## $\chi^2(n)$ 的概率密度和图像

$\chi^2(n)$ 的概率密度为

$$f(y)=\begin{cases}\frac{1}{2^{n/2}\Gamma(n/2)}y^{\frac{n}{2}-1}e^{-\frac{y}{2}} & y>0;\\ 0 & \text{其他.}\end{cases}.$$

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

$f(y)$ 的图像.

## $\chi^2$ 分布和 $\Gamma$ 分布

$\chi^2(n)$ 的概率密度为

$$f(y)=\begin{cases}\frac{1}{2^{n/2}\Gamma(n/2)}y^{\frac{n}{2}-1}e^{-\frac{y}{2}} & y>0;\\ 0 & \text{其他.}\end{cases}.$$

$\Gamma(\alpha,\theta)(\alpha>0,\theta>0)$ 的概率密度为

$$f(x)=\begin{cases}\frac{1}{\theta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\theta}} & x>0;\\ 0 & \text{其他.}\end{cases}$$

## 性质

- $\chi^2(1) \sim \Gamma(\frac{1}{2}, 2)$;
- $\chi^2(n) \sim \Gamma(\frac{n}{2}, 2)$.

证明：$X_i \sim N(0,1)$，所以 (Page-52, 80)

$$X_i^2 \sim \chi^2(1) \sim \Gamma(\frac{1}{2}, 2),$$

$X_i^2$ 相互独立, 由 $\Gamma$ 分布的可加性,

$$\chi^2 = \sum_{i=1}^{n} X_i^2 = \Gamma(\frac{n}{2}, 2).$$

## $\chi^2$ 分布的性质

**性质 (可加性)**

设 $\chi_1^2 \sim \chi^2(n_1)$, $\chi_2^2 \sim \chi^2(n_2)$, 且 $\chi_1^2, \chi_2^2$ 相互独立, 则

$$\chi_1^2 + \chi_2^2 \sim \chi^2(n_1 + n_2).$$

- 一般地, 设 $\chi_i^2 \sim \chi^2(n_i)$, 且 $\chi_i^2 (i = 1, ..., m)$ 相互独立, 则

$$\sum_{i=1}^{m} \chi_i^2 \sim \chi^2(n_1 + ... + n_m).$$

## $\chi^2$ 分布的性质

**性质 (期望和方差)**

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

证: $X_i \sim N(0,1)$, $E(X_i^2) = D(X_i) = 1$,

$$D(X_i^2) = E(X_i^4) - [E(X_i^2)]^2 = 3 - 1 = 2.$$

则 $E(\chi^2) = E(\sum_{i=1}^{n} X_2^2) = \sum E(X_i^2) = n.$

$D(\chi^2) = D(\sum_{i=1}^{n} X_i^2) = \sum D(X_i^2) = 2n.$ □

## 上/下 $\alpha$ 分位数（Page-120）

任意给定随机变量 $X$, 分布函数 $F(x)$, 概率密度函数 $f(x)$,

- 下 $\alpha$ 分位数

$$P\{X \leq \chi_{\underline{\alpha}}\} = F(\chi_{\underline{\alpha}}) = \int_{-\infty}^{\underline{\alpha}} f(x)\,dx = \alpha.$$

- 上 $\alpha$ 分位数

$$P\{X > \chi_{\alpha}\} = 1 - F(\chi_{\alpha}) = \int_{\chi_{\alpha}}^{+\infty} f(x)\,dx = \alpha.$$

## $\chi^2$ 分布的性质

**性质 (上分位数)**

给定 $0 < \alpha < 1$, 满足条件

$$P\{\chi^2 > \chi^2_\alpha(n)\} = \int_{\chi^2_\alpha(n)}^{\infty} f(y)dy = \alpha$$

的 $\chi^2_\alpha(n)$ 称为 $\chi^2$ 分布的上 $\alpha$ 分位数.

## 求 $\chi^2$ 分布的上分位数

- 查附录 5(Page-400, $n = 40$ 为止)
  $\alpha = 0.05$, $n = 20$, $\chi^2_{0.05}(20) = 31.410$.
  $\alpha = 0.1$, $n = 25$, $\chi^2_{0.1}(25) = 34.382$.
- 当 $n$ 充分大时, 费希尔证明

$$\chi^2_\alpha \approx \frac{1}{2}(z_\alpha + \sqrt{2n-1})^2.$$

  其中 $z_\alpha$ 是标准正态分布上的上 $\alpha$ 分位数.
- 当 $n > 40$ 时, 可用上式求,
  $\chi^2_{0.05}(50) \approx \frac{1}{2}(1.645 + \sqrt{99})^2 = 67.221$

**例**

设总体 $X \sim N\left(\mu, \sigma^2\right), \mu, \sigma^2$ 已知. $(X_1, \cdots, X_n)$ 是取自总体 $X$ 的样本. 求统计量

$$\chi^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

的分布.

证明：令 $Y_i = \frac{X_i - \mu}{\sigma}$, 则 $Y_i \sim N(0, 1)$.

因此 $\chi^2 = \sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum Y_i \sim \chi^2(n)$. □