概 率 论 与 数 理 统 计

Probability and Statistics

南京大学 张绍群 & 徐科

更新： November 17, 2025

Part I

统计 (Statistics)

Part IX – Ch09: 统计的基本概念

Ch09: 统计的基本概念

Basic Concepts of Statistics

November 17, 2025

提纲

- 总体 vs 个体
- 统计量
  - 样本均值、样本方差、样本矩
  - 次序统计量
- 常用的三个统计分布
  - Beta 分布、Dirichlet 分布、Gamma 分布
- 三大抽样分布
  - χ^2 分布、t 分布、F 分布

引言

之前的课程属于概率论的范畴. 随机变量及其概率分布全面地描述了随机现象的统计性质. 在概率论的许多问题中, 随机分布被假定为已知的, 而一切的计算及推理都基于已知的分布函数进行. 但在实际问题中, 情况往往并非如此.

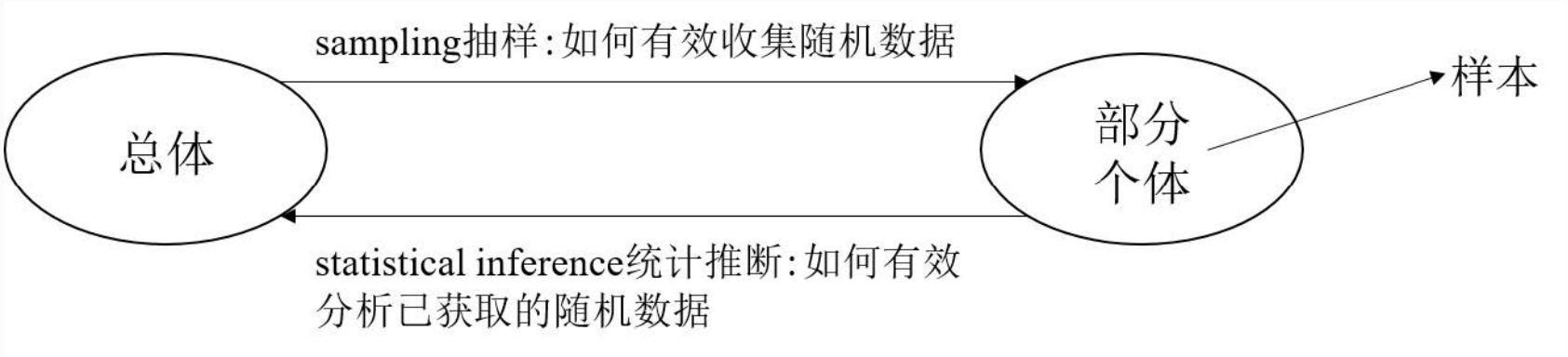
例如调查一批产品的不合格率, 某个省份的人均收入等诸如此类的问题, 是无法事先知道其分布函数的. 这类问题属于统计学的范畴. 一般认为, 统计学是一门研究如何有效地收集和分析所获数据的统计规律的学科. 统计学的研究内容包括: 抽样调查、参数估计、假设检验等.

基本概念

从总体中随机抽取一些个体, 表示为 $X_1, X_2, \ldots, X_n$, 称 $X_1, X_2, \ldots, X_n$ 为取自总体 X 的随机样本, 其样本容量为 n

- 总体: 研究对象的全体, 用随机变量 X 表示 (分布未知)
- 抽样: 抽取样本的过程
- 样本值: 观察样本得到的数值, 例如: $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ 为样本观察值或样本值
- 样本的二重性:
  - 就一次观察而言, 样本值是确定的数值
  - 不同的抽样下, 样本值会发生变化, 某次 sampling 可看作随机变量

基本概念 – 示图


sampling抽样：如何有效收集随机数据
总体
部分
个体
样本
statistical inference统计推断：如何有效
分析已获取的随机数据

简单随机样本

定义 0.1 (简单随机样本) 样本 $X_1, X_2, \ldots, X_n$ 是总体 X 的 简单随机样本, 简称 样本, 如果 $X_1, X_2, \ldots, X_n$ 满足

- 代表性: X_i 与 X 同分布
- 独立性: $X_1, X_2, \ldots, X_n$ 之间相互独立

Remarks: 课程后面所考虑的样本均为简单随机样本.

样本的分布

定义 0.2 (因为独立性) 总体 X 的联合分布函数为 $F(x), x_1, x_2, \ldots, x_n$ 为取自该总体的容量为 n 的样本, 则样本的联合分布函数为

$$F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F(x_i) \, .$$

若总体 X 的概率密度为 $f(x)$, 则样本 $x_1, x_2, \ldots, x_n$ 的联合概率密度为

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \, .$$

若总体 X 的分布列 $P(X = x_i)$, 则样本 $x_1, x_2, \ldots, x_n$ 的联合分布列为

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

统计量

样本来自于总体, 因此样本中含有总体各方面的信息. 为将这些分散在样本中的有关总体的信息集中起来以反映总体的各种特征, 需要对样本进行加工, 较有效的加工方法是构建样本的函数, 不同样本函数反映总体的不同特征.

定义 0.3 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的一个样本, 若 $T = g(X_1, X_2, \ldots, X_n)$ 是一个连续、且不含任意参数的函数. 称 T 是一个统计量.

统计量具有以下性质:

- 因为 $X_1, X_2, \ldots, X_n$ 是随机变量, 所以 $g(X_1, X_2, \ldots, X_n)$ 是随机变量;
- $g(x_1, x_2, \ldots, x_n)$ 为 $g(X_1, X_2, \ldots, X_n)$ 的一次观察值.

Remarks: g 是定义在随机变量 $(X_1, X_2, \ldots, X_n)$ 上的.

统计量 – 样本均值及样本方差

定义 0.4 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 其算术平均值称为样本均值, 一般用 $\bar{X}$ 表示, 即

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

样本关于它本身的样本均值 $\bar{X}$ 的平均偏差平方和称为 样本方差, 即

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

其算术根 $S_n = \sqrt{S_n^2}$ 称为 样本标准差. 当 n 不大时, 常用

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

作为样本方差 (也称 无偏方差).

无偏方差解释

直觉原因：样本均值靠得太近 $\rightarrow$ 方差被系统性压缩

真实方差围绕总体均值 μ，偏差为 $(X_i-\mu)^2$；但样本方差使用的是围绕样本均值的偏差 $(X_i-\bar{X})^2$。由于 $\bar{X}$ 是根据样本计算出来的，它会被样本“拉过去”，使偏差自然变小，导致方差被低估。严格可证样本方差 S_n^2 的期望是 $\mathbb{E}[S_n^2]=\frac{n-1}{n}\sigma^2$，因此低估比例恰好为 $\frac{n-1}{n}$。

自由度解释：偏差只能自由变化 $n-1$ 个

偏差满足 $\sum_{i=1}^{n}(X_i-\bar{X})=0$，意味着只要前 $n-1$ 个偏差确定，最后一个偏差就被强制决定。有效自由度因此只有 $n-1$ 个；但若仍用 n 去平均，就会把“$n-1$ 个自由信息”分摊到 n，从而压缩方差。修正方法就是除以 $n-1$，即无偏方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$。

样本均值及样本方差的性质

- 设总体 X 的期望 $\mathbb{E}[X] = \mu$ 以及方差 $\mathbb{VAR}(X) = \sigma^2$, 则有

$$\mathbb{E}[\bar{x}] = \mu, \quad \mathbb{VAR}(\bar{x}) = \sigma^2/n, \quad \bar{x} \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n)$$

- 样本方差 S_n^2 与总体方差 σ^2 之间存在偏差, 即

$$\mathbb{E}[S_n^2] = \frac{n-1}{n}\sigma^2$$

- 无偏方差 S^2 与总体方差 σ^2 相等

$$\mathbb{E}[S^2] = \sigma^2$$

Remarks: 在实际中, S^2 比 S_n^2 更常用, 因此以后讲样本方差通常是指 S^2. (可证)

样本均值及样本方差的性质 – Proof

$$
\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i^2-2X_i\bar{X}+\bar{X}^2)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}-2X_i\bar{X}\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bar{X}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] - \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}(X_jX_i)\right] + \mathbb{E}\left[\bar{X}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2\right] - \frac{2}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_iX_i)\right] - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}(X_jX_i) \\
&\quad + \frac{1}{n^2}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2)\right] + \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}(X_jX_i)\right] \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

样本均值及样本方差的性质 – Proof

$$
\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\big((X_i - \mu) - (\bar{X} - \mu)\big)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\big((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\big)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right] \; - \; \mathbb{E}\big[(\bar{X} - \mu)^2\big] \\
&= \sigma^2 \; - \; \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}
$$

统计量：例 0.1

例 0.1 在一批产品中随机检查了 10 箱, 发现每箱中的不合格品数为

$$4, 5, 6, 0, 3, 1, 4, 2, 1, 4$$

试计算其样本均值、样本方差和样本标准差.

解答：例 0.1

题目: 在一批产品中随机检查了 10 箱, 发现每箱中的不合格品数为

$$4, 5, 6, 0, 3, 1, 4, 2, 1, 4$$

试计算其样本均值、样本方差和样本标准差.

解答:

- 根据样本均值、样本方差和样本标准差的定义直接计算, 即

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{4 + 5 + \cdots + 4}{10} = 3.$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{9}[(4-3)^2 + (5-3)^2 + \cdots + (4-3)^2] = 3.78.$$

$$s = \sqrt{s^2} = 1.94.$$

统计量：例 0.2

例 0.2 设总体 $X \sim \mathcal{N}(20, 3)$, 从总体中抽取两独立样本, 容量分别为 10 和 15. 求这两个样本均值之差的绝对值大于 0.3 的概率.

解答：例 0.2

题目: 设总体 $X \sim \mathcal{N}(20, 3)$, 从总体中抽取两独立样本, 容量分别为 10 和 15. 求这两个样本均值之差的绝对值大于 0.3 的概率.

解答:

- 设 $x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}$ 和 $x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}$ 分别为来自总体 $X \sim \mathcal{N}(20, 3)$ 的两个独立样本. 根据正态分布的性质有

$$\bar{x^{(1)}} = \frac{1}{10}\sum_{i=1}^{10} x_i^{(1)} \sim \mathcal{N}(20, 3/10), \quad \bar{x^{(2)}} = \frac{1}{15}\sum_{i=1}^{15} x_i^{(2)} \sim \mathcal{N}(20, 1/5)$$

- 进一步根据正态分布的性质有 $\bar{x^{(1)}} - \bar{x^{(2)}} \sim \mathcal{N}(0, 1/2)$, 于是可得

$$P(|\bar{x^{(1)}} - \bar{x^{(2)}}| > 0.3) = 2 - 2\Phi(0.3/\sqrt{1/2}).$$

统计量 – 样本矩

样本均值和样本方差的更一般推广是样本矩, 这是一类常见的统计量.

定义 0.5 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, k 为正整数, 则统计量

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

称为 样本 k 阶原点矩. 特别的, 样本一阶原点矩就是样本均值. 统计量

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

称为 样本 k 阶中心矩. 特别的, 样本二阶中心矩就是样本方差.

次序统计量

除了样本矩以外, 另外一类常见的统计量是次序统计量. 它在实际应用及理论中都有广泛的应用.

定义 0.6 设 $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, $X_{(i)}$ 称为该样本的第 i 个次序统计量, 它的取值是将样本观测值从小到大排序后得到的第 i 个观测值. 其中

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

称为该样本的 最小次序统计量.

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

称为该样本的最大次序统计量. $R_n = X_{(n)} - X_{(1)}$ 称为样本极差.

单个次序统计量的分布

定理 0.1 设总体 X 的密度函数为 $f(x)$, 分布函数为 $F(x)$, $X_1, X_2, \ldots, X_n$ 是来自总体 X 的样本, 则第 k 个次序统计量 $X_{(k)}$ 的分布函数和密度函数分别为

$$F_k(x) = \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1 - F(x)]^{n-r} ,$$

$$f_k(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x) .$$

Remarks:

- 次序统计量 $X_{(k)}$ 表示 $X_1, X_2, \ldots, X_n$ 中有 k 个变量小于等于 $X_{(k)}$
- f_k 理解为 $X_{(k)}$ 在 x 附近的小区间 $(x, x + dx)$ 内的事件
- 令 $k = 1$ 和 $k = n$, 分别得到最小次序统计量和最大次序统计量的分布函数和密度函数

证明: 单个次序统计量的分布（应用莱布尼茨定理）

证明 根据题意有第 k 次序统计量 $X_{(k)}$ 的分布函数为

$$\begin{aligned} F_k(x) &= \Pr[X_{(k)} \leqslant x] = \Pr[X_1, X_2, \cdots, X_n \text{ 中至少有 } k \text{ 个随机变量 } \leqslant x] \\ &= \sum_{r=k}^{n} \Pr[X_1, X_2, \cdots, X_n \text{ 中恰有 } r \text{ 个随机变量 } \leqslant x,\ n-r \text{ 个随机变量 } > x] \\ &= \sum_{r=k}^{n} \binom{n}{r} [F(x)]^r [1-F(x)]^{n-r}. \end{aligned}$$

利用恒等式

$$\sum_{r=k}^{n} \binom{n}{r} p^r (1-p)^{n-r} = \frac{n!}{(k-1)!(n-k)!} \int_0^p t^{k-1}(1-t)^{n-k} dt \quad (r \in [n],\ p \in [0,1])$$

由此可知

$$F_k(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1}(1-t)^{n-k} dt,$$

根据积分函数求导完成证明.

证明: 单个次序统计量的分布（应用莱布尼茨定理）

令

$$G(p)=\sum_{r=k}^{n}\binom{n}{r}p^r(1-p)^{n-r},\qquad 0\le p\le 1.$$

显然 $G(0)=0$。对 $G(p)$ 求导，有

$$\begin{aligned}
G'(p)&=\sum_{r=k}^{n}\binom{n}{r}\left[rp^{r-1}(1-p)^{n-r}-(n-r)p^r(1-p)^{n-r-1}\right]\\
&=\sum_{r=k}^{n}n\binom{n-1}{r-1}p^{r-1}(1-p)^{n-r}-\sum_{r=k}^{n}n\binom{n-1}{r}p^r(1-p)^{n-r-1}\\
&=n\sum_{j=k-1}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j}-n\sum_{j=k}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j}\\
&=n\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}=\frac{n!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k}.
\end{aligned}$$

因此

$$G(p)=G(0)+\int_0^p G'(t)\,dt=\frac{n!}{(k-1)!(n-k)!}\int_0^p t^{k-1}(1-t)^{n-k}\,dt,$$

即

$$\sum_{r=k}^{n}\binom{n}{r}p^r(1-p)^{n-r}=\frac{n!}{(k-1)!(n-k)!}\int_0^p t^{k-1}(1-t)^{n-k}\,dt.$$

小结

通过前面的学习, 我们知道样本统计量也是一个随机变量. 以样本均数为例, 假设我们现在想了解南京大学男性学生的身高情况, 按照同样的方法重复 50 次抽样, 每次抽 100 人, 每组样本都可以计算一个样本均数, 假设分别为: 1.76, 1.72, 1.69, 1.77, . . . , 1.75, 样本均数会随着抽样的不同而随机变动.

进一步, 我们可以将样本统计量作为随机变量研究其概率分布 (又称“抽样分布”), 从而得到其分布的性质或计算特定情况下的概率.

本课程主要研究的抽样分布通常是样本均值的分布、样本方差的分布、样本标准差的分布.

Beta (贝塔) 分布

定义 0.7 (Beta 函数) 对任意给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 定义 Beta 函数为

$$\text{Beta}(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}\,\mathrm{d}x,$$

简记为 $B(\alpha_1, \alpha_2)$, 被称为第一类欧拉积分函数.

定义 0.8 给定 $\alpha_1 > 0$ 和 $\alpha_2 > 0$, 若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{B(\alpha_1,\alpha_2)} x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}, & x \in (0,1) \\ 0, & \text{其它} \end{cases}$$

称 X 服从参数为 α_1 和 α_2 的 Beta 分布, 记为 $X \sim B(\alpha_1, \alpha_2)$.

Beta (贝塔) 分布的数字特征

定理 0.2 若随机变量 $X \sim B(\alpha_1, \alpha_2)$, 则有

$$\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \quad \mathbb{VAR}(X) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}.$$

由期望定义得到

$$\mathbb{E}[X] = \int_0^1 x^{\alpha_1}(1-x)^{\alpha_2 - 1}\, dx = B(\alpha_1 + 1, \alpha_2),$$

故

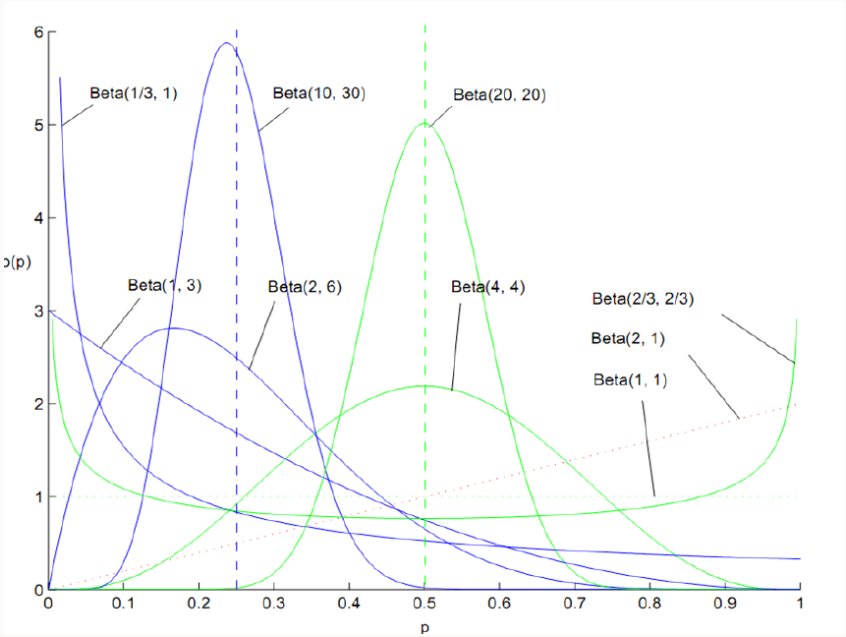
$$\mathbb{E}[X] = \frac{B(\alpha_1 + 1, \alpha_2)}{B(\alpha_1, \alpha_2)}.$$

利用 $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ 以及 $\Gamma(a+1) = a\Gamma(a)$, 可得

$$\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}.$$

Beta (贝塔) 分布的性质

Beta 分布的概率密度我们把它画成图, 会发现它是个百变星君, 它可以是凹的、凸的、单调上升的、单调下降的; 可以是曲线也可以是直线, 而均匀分布也是特殊的分布. 由于分布能够拟合如此之多的形状, 因此它在统计数据拟合和贝叶斯分析中被广泛使用.



Dirichlet (狄利克雷) 分布

定义 0.9 给定 $\alpha_1, \ldots, \alpha_k \in (0, +\infty)$, 若多元随机向量 $X = (X_1, \ldots, X_k)$ 的概率密度函数为

$$f(x_1, x_2, \ldots, x_k) = \begin{cases} \frac{x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_k^{\alpha_k - 1}}{B(\alpha_1, \alpha_2, \ldots, \alpha_k)}, & \sum_{i=1}^{k} x_i = 1 \text{且} x_i > 0 (i \in [k]) \\ 0, & \text{其它} \end{cases}$$

称 X 服从参数为 $\alpha_1, \ldots, \alpha_k$ 的 Dirichlet 分布, 记为 $X \sim \text{Dir}(\alpha_1, \ldots, \alpha_k)$.

Dirichlet 分布是 Beta 分布的一种高维推广. 当 $k = 2$ 时, Dirichlet 分布退化为 Beta 分布.

Dirichlet (狄利克雷) 分布的数字特征

定理 0.3 若随机变量 $X \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, 设 $\tilde{\alpha} = \alpha_1 + \alpha_2 + \cdots + \alpha_k$ 和 $\tilde{\alpha}_i = \alpha_i / \tilde{\alpha}$ 则有

$$\mathbb{E}[X_i] = \tilde{\alpha}_i, \quad \mathbb{VAR}(X_i, X_j) = \begin{cases} \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\tilde{\alpha}+1}, & i = j \\ -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha}+1}, & i \neq j \end{cases}$$

Remarks:

- $k = 2$ 时，$\mathbb{E}[X_i]$ 和 $\mathbb{VAR}(X_i, X_j)$ 就是 Beta 对应计算值
- 因为 Dirichlet 分布描述的是一个总和必须等于 1 的概率向量。所以如果某个 X_i 增大，那其他分量必须减少一点做补偿。因此不同分量之间天然是负相关的

Gamma (伽马) 分布

定义 0.10 (Γ 函数) 对任意给定 $\alpha > 0$, 定义 Γ 函数为

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

被称为第二类欧拉积分函数.

定义 0.11 若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

其中 $\alpha > 0$ 且 $\lambda > 0$, 称 X 服从参数为 α 和 λ 的 Γ 分布, 记为 $X \sim \Gamma(\alpha, \lambda)$.

Γ (伽马) 分布的数字特征

定理 0.4 若随机变量 $X \sim \Gamma(\alpha, \lambda)$, 则有

$$\mathbb{E}[X] = \alpha/\lambda\,, \quad \mathbb{VAR}(X) = \alpha/\lambda^2\,.$$

定理 0.5 (Gamma 分布的可加性) 若随机变量 $X \sim \Gamma(\alpha_1, \lambda)$ 和 $Y \sim \Gamma(\alpha_2, \lambda)$, 且 X 和 Y 相互独立, 则有 $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

另外, 对比指数分布和伽马分布的密度函数形式, 易知 $\Gamma(1, \lambda) = e(\lambda)$.

Remarks:

- 定理0.5说明，如果 X 表示等到第 α_1 次事件的时间，Y 表示等到第 α_2 次事件的时间，那么 $X + Y$ 就是等到第 $\alpha_1 + \alpha_2$ 次事件的时间。—这就是 Poisson 过程的本质

数字特征的分布: 图像、数字特征、例子

- Beta 分布: 数字特征

$$\mathbb{E}[X] = \frac{\alpha_1}{\alpha_1 + \alpha_2}, \quad \mathbb{VAR}(X) = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}$$

- Dirichlet 分布: 数字特征

$$\mathbb{E}[X_i] = \tilde{\alpha}_i, \quad \mathbb{VAR}(X_i, X_j) = \begin{cases} \frac{\tilde{\alpha}_i(1-\tilde{\alpha}_i)}{\tilde{\alpha}+1}, & i = j \\ -\frac{\tilde{\alpha}_i \tilde{\alpha}_j}{\tilde{\alpha}+1}, & i \neq j \end{cases}$$

- Gamma 分布: 数字特征

$$\mathbb{E}[X] = \alpha/\lambda, \quad \mathbb{VAR}(X) = \alpha/\lambda^2$$

Beta 分布：例 0.3

例 0.3 设总体分布 $U(0,1)$, $x_1, x_2, \ldots, x_n$ 为样本, 试求第 k 个次序统计量 $x_{(k)}$ 的密度函数.

解答：例 0.3

题目: 设总体分布 $U(0,1), x_1, x_2, \ldots, x_n$ 为样本, 试求第 k 个次序统计量 $x_{(k)}$ 的密度函数.

解答:

- 由题易知总体 X 的分布函数 $F(x) = x, x \in (0,1)$ 和密度函数 $f(x) = 1, x \in (0,1)$, 进一步根据定理 0.1, 可得

$$\begin{aligned} p_k(x) &= \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} \\ &= \frac{1}{B(k, n-k+1)} x^{k-1}(1-x)^{n-k}, \qquad 0 < x < 1. \end{aligned}$$

- 即贝塔分布 $B(k, n-k+1)$.

解答：例 0.3

本页求证

$$\int_0^1 C_{n-1}^{k-1} x^{k-1}(1-x)^{n-k}dx = \frac{1}{n} \quad \text{for any} \quad k\,.$$

Proof: 分布积分

$$\begin{aligned}\int_0^1 C_{n-1}^{k-1} x^{k-1}(1-x)^{n-k}dx &= C_{n-1}^{k-1}\left[\frac{x^k(1-x)^{n-k}}{k}\ |_0^1 - \int_0^1 -\frac{x^k}{k}(n-k)(1-x)^{n-k-1}\,\mathrm{d}x\right] \\ &= \int_0^1 C_{n-1}^{k} x^{k}(1-x)^{n-k-1}\,\mathrm{d}x \\ &= \int_0^1 x^{n-1}\,\mathrm{d}x \\ &= \frac{1}{n}\end{aligned}$$

因此，有

$$B(k, n-k+1) = \int_0^1 x^{k-1}(1-x)^{n-k}dx = \frac{1}{n}\frac{1}{C_{n-1}^{k-1}}\,.$$

Gamma 分布：例 0.4

例 0.4 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $X^2 \sim \Gamma(1/2, 1/2)$.

解答：例 0.4

题目: 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $X^2 \sim \Gamma(1/2, 1/2)$.

解答:

- 由题易知总体 $Y = X^2$ 的分布函数为: 当 $y \leq 0$ 时, 有 $F_Y(y) = 0$; 当 $y > 0$ 时, 有

$$F_Y(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x,$$

  由此可得概率密度函数为 $f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-\frac{y}{2}}$, 从而得到 $X^2 \sim \Gamma(1/2, 1/2)$.
- 利用高斯积分

$$\Gamma\left(\tfrac{1}{2}\right) = \int_0^\infty x^{-1/2} e^{-x}\, dx = \int_0^\infty \frac{1}{t} e^{-t^2} (2t\, dt) = 2\int_0^\infty e^{-t^2} dt,$$

$$\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi} \quad \Longrightarrow \quad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}.$$

三大 (运算) 抽样分布

中心极限定理揭示了样本均值的分布, 即不管总体是什么分布, 任意一个总体的样本平均值都会围绕在总体的平均值周围, 并且呈正态分布. 在数理统计中, 用于描述抽样分布的分布函数, 除了正态分布外, 最重要的三个分布分别是:

- χ^2 (卡方) 分布
- t 分布
- F 分布

这三个以标准正态分布而构造的统计量在统计推断中有广泛的应用, 不仅是因为这三个统计量有明确的构造背景, 而且其抽样分布的密度函数都有显性表达式, 它们被称为统计中的 “三大抽样分布”.

χ^2 (卡方) 分布及其密度函数

定义 0.12 若 $X_1, X_2, \ldots, X_n$ 是来自总体 $X \sim \mathcal{N}(0,1)$ 的一个样本, 称 $Y = X_1^2 + X_2^2 + \cdots + X_n^2$ 为服从自由度为 n 的 χ^2 分布, 记 $Y \sim \chi^2(n)$.

根据 $X_1^2 \sim \Gamma(1/2, 1/2)$ 和 Γ 函数的可加性可得 $Y \sim \Gamma(n/2, 1/2)$. 于是有随机变量 Y 的概率密度函数为

$$f_Y(y) = \begin{cases} \frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, & y > 0 \\ 0, & y \leq 0 \end{cases}$$

χ^2 分布的性质

- 若随机变量 $X \sim \chi^2(n)$, 则 $\mathbb{E}[X] = n$ 和 $\mathbb{VAR}(X) = 2n$;
- 若随机变量 $X \sim \chi^2(m)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则 $X + Y \sim \chi^2(m + n)$;
- 推广命题：若随机变量 $X \sim \mathcal{N}(0, 1)$, 则

$$\mathbb{E}[X^k] = \begin{cases} (k-1)!!, & k \text{ 为偶数} \\ 0, & k \text{ 为奇数} \end{cases}$$

其中,

$$\begin{cases} (2k)!! = 2k \cdot (2k-2) \cdots 2 \,, \\ (2k+1)!! = (2k+1) \cdot (2k-1) \cdots 1 \,, \end{cases}$$

χ^2 分布：例 0.5

例 0.5 若 X_1, X_2, X_3, X_4 是来自总体 $X \sim \mathcal{N}(0, 4)$ 的样本, 以及

$$Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$$

求 a, b 取何值时, Y 服从 χ^2 分布, 并求其自由度.

解答：例 0.5

题目: 若 X_1, X_2, X_3, X_4 是来自总体 $X \sim \mathcal{N}(0, 4)$ 的样本, 以及

$$Y = a(X_1 - 2X_2)^2 + b(3X_3 - 4X_4)^2$$

求 a, b 取何值时, Y 服从 χ^2 分布, 并求其自由度.

解答:

- 根据正态分布的性质有 $X_1 - 2X_2 \sim \mathcal{N}(0, 20)$ 和 $3X_3 - 4X_4 \sim \mathcal{N}(0, 100)$, 因此

$$\frac{X_1 - 2X_2}{2\sqrt{5}} \sim \mathcal{N}(0, 1), \qquad \frac{3X_3 - 4X_4}{10} \sim \mathcal{N}(0, 1),$$

  所以当 $a = 1/20, b = 1/100$ 时有 $Y \sim \chi^2(2)$ 成立.

χ^2 分布：例 0.6

例 0.6 相互独立的随机变量 $X_1, X_2, \ldots, X_n$ 满足 $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, 求 $\sum_{i=1}^{n}(X_i - \mu_i)^2/\sigma_i^2$ 的分布.

解答：例 0.6

题目: 相互独立的随机变量 $X_1, X_2, \ldots, X_n$ 满足 $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, 求 $\sum_{i=1}^n (X_i - \mu_i)^2/\sigma_i^2$ 的分布.

解答:

- 令 $Y_i = (X_i - \mu_i)/\sigma_i$, 由题意知 $Y_1, Y_2, \ldots, Y_n$ 是独立同分布的随机变量, 其共同分布为 $\mathcal{N}(0, 1)$, 于是由定义 0.12可知

$$Y = \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n (X_i - \mu_i)^2/\sigma_i^2 \sim \chi^2(n)$$

  即 $\sum_{i=1}^n (X_i - \mu_i)^2/\sigma_i^2$ 服从自由度为 n 的 χ^2 分布.

分布可加性小结

- 若随机变量 $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 和 $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ 相互独立, 那么

$$X \pm Y \sim \mathcal{N}(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2)$$

- 若随机变量 $X \sim B(n_1, p)$ 和 $Y \sim B(n_2, p)$ 相互独立, 那么

$$X + Y \sim B(n_1 + n_2, p)$$

- 若随机变量 $X \sim P(\lambda_1)$ 和 $Y \sim P(\lambda_2)$ 相互独立, 那么

$$X + Y \sim P(\lambda_1 + \lambda_2)$$

- 若随机变量 $X \sim \Gamma(\alpha_1, \lambda)$ 和 $Y \sim \Gamma(\alpha_2, \lambda)$ 相互独立, 那么

$$X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$$

- 若随机变量 $X \sim \chi^2(m)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则 $X+Y \sim \chi^2(m+n)$.

t 分布 (student distribution) 及其密度函数

定义 0.13 随机变量 $X \sim \mathcal{N}(0,1)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则随机变量

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

服从自由度为 n 的 t 分布, 记 $T \sim t(n)$.

随机变量 $T \sim t(n)$ 的概率密度为（具有对称性）

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

当 $n \to \infty$ 时, 随机变量 $T \sim t(n)$ 的概率密度为

$$f(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

因此当 n 足够大时,$f(x)$ 可被近似为 $\mathcal{N}(0,1)$ 的密度函数.

t 分布：例 0.7

例 0.7 设 $X_1, X_2, \ldots, X_9$ 和 $Y_1, Y_2, \ldots, Y_9$ 是分别来自于总体 $\mathcal{N}(0, 9)$ 的两个独立样本, 求 $(X_1 + X_2 + \cdots + X_9)/\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}$ 的分布.

解答：例 0.7

题目: 设 $X_1, X_2, \ldots, X_9$ 和 $Y_1, Y_2, \ldots, Y_9$ 是分别来自于总体 $\mathcal{N}(0, 9)$ 的两个独立样本, 求 $(X_1 + X_2 + \cdots + X_9)/\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}$ 的分布.

解答:

- 由题意知 $X_1 + X_2 + \cdots + X_9 \sim \mathcal{N}(0, 81)$, 标准化后可得 $X = \frac{X_1 + X_2 + \cdots + X_9}{9} \sim \mathcal{N}(0, 1)$; 同时, 根据定义 0.12可知, $Y = \left(\frac{Y_1}{3}\right)^2 + \left(\frac{Y_2}{3}\right)^2 + \cdots + \left(\frac{Y_9}{3}\right)^2 \sim \chi^2(9)$, 显然 X 与 Y 独立, 根据定义 0.13可知

$$\frac{\frac{X_1 + X_2 + \cdots + X_9}{9}}{\sqrt{\left(\left(\frac{Y_1}{3}\right)^2 + \left(\frac{Y_2}{3}\right)^2 + \cdots + \left(\frac{Y_9}{3}\right)^2\right)/9}} = \frac{(X_1 + X_2 + \cdots + X_9)}{\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}} \sim t(9).$$

  即 $(X_1 + X_2 + \cdots + X_9)/\sqrt{Y_1^2 + Y_2^2 + \cdots + Y_9^2}$ 服从自由度为 9 的 t 分布.

F 分布及其密度函数

定义 0.14 随机变量 $X \sim \chi^2(m)$ 和 $Y \sim \chi^2(n)$ 相互独立, 则随机变量

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

服从自由度为 (m, n) 的 F 分布, 记 $Z \sim F(m, n)$.

随机变量 $Z \sim F(m, n)$ 的概率密度为

$$f(z) = \begin{cases} \dfrac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}} z^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(1+\frac{mx}{n}\right)^{\frac{m+n}{2}}}, & z > 0 \\ 0, & z \leq 0 \end{cases}$$

若随机变量 $Z \sim F(m, n)$, 则 $\frac{1}{Z} \sim F(n, m)$.

F 分布：例 0.8

例 0.8 设 $X_1, X_2, \ldots, X_{2n}$ 是来自于总体 $\mathcal{N}(0, \sigma^2)$ 的样本, 求

$$(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/(X_2^2 + X_4^2 + \cdots + X_{2n}^2)$$

的分布.

解答：例 0.8

题目: 设 $X_1, X_2, \ldots, X_{2n}$ 是来自于总体 $\mathcal{N}(0, \sigma^2)$ 的样本, 求

$$(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/(X_2^2 + X_4^2 + \cdots + X_{2n}^2)$$

的分布.

解答:

- 由题意知 $\frac{X_i}{\sigma} \sim \mathcal{N}(0,1)$, 根据定义0.12可知, $A = \left(\frac{X_1}{\sigma}\right)^2 + \left(\frac{X_3}{\sigma}\right)^2 + \cdots + \left(\frac{X_{2n-1}}{\sigma}\right)^2 \sim \chi^2(n)$, $B = \left(\frac{X_2}{\sigma}\right)^2 + \left(\frac{X_4}{\sigma}\right)^2 + \cdots + \left(\frac{X_{2n}}{\sigma}\right)^2 \sim \chi^2(n)$, 显然 A 与 B 独立, 根据0.14可知

$$\frac{A/n}{B/n} = \frac{[(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/n^2]/n}{[(X_2^2 + X_4^2 + \cdots + X_{2n}^2)/n^2]/n} \sim F(n,n).$$

  即 $(X_1^2 + X_3^2 + \cdots + X_{2n-1}^2)/(X_2^2 + X_4^2 + \cdots + X_{2n}^2)$ 服从自由度为 (n,n) 的 F 分布.

三大抽样分布小结

若 $X_1, X_2, \ldots, X_n$ 和 $Y_1, Y_2, \ldots, Y_n$ 是来自标准正态分布的两个相互独立的样本, 因此三个统计量的构造及抽样分布如下表所示.

统计量的构造	抽样分布密度函数	期望	方差
$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$	$f_Y(y) = \frac{(\frac{1}{n})^{\frac{n}{2}}}{\Gamma(n/2)} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, (y > 0)$	n	$2n$
$F = \frac{(Y_1^2+Y_2^2+\cdots+Y_m^2)/m}{(X_1^2+X_2^2+\cdots+X_n^2)/n}$	$f_Y(y) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}} y^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(1+\frac{mx}{n}\right)^{\frac{m+n}{2}}}, (y > 0)$	$\frac{n}{n-2}, (n > 2)$	$\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, (n > 4)$
$t = \frac{Y_1}{\sqrt{(X_1^2+X_2^2+\cdots+X_n^2)/n}}$	$f_Y(y) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})\sqrt{n\pi}} \left(1 + \frac{y^2}{n}\right)^{-\frac{n+1}{2}}$	$0, (n > 1)$	$\frac{n}{n-2}, (n > 2)$