

概率统计分析与量测技术笔记

adamanteye

分布	分布律	期望	方差
Exp(θ)	$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{where } x > 0 \\ 0 & \text{elsewhere} \end{cases}$	θ	θ^2

表 1 常见概率分布表

1. 集合以及事件

定义 1 **σ -field**: σ -field, or σ -algebra, is a collection of subsets of a set S that is closed under countable unions, countable intersections, and complements.

定理 2: Let $\mathcal{O}$ be one of the 8 set operations, let $\{\mathcal{C}_t, t \in T\}$ be an indexed family of subsets such that for each t , $\mathcal{C}_t$ is closed under $\mathcal{O}$. Then

$$\mathcal{C} = \bigcap_{t \in T} \mathcal{C}_t \text{ is closed under } \mathcal{O}$$

推论 2.1: The intersection of a σ fields is a σ field.

定义 3 **minimal σ -field**: Let $\mathcal{C}$ be a collection of subsets of Ω . The σ -field generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is a σ -field satisfying the following conditions:

- $\mathcal{C} \subset \sigma(\mathcal{C})$
- If $\mathcal{B}'$ is some other σ -field containing $\mathcal{C}$, then $\sigma(\mathcal{C}) \subset \mathcal{B}'$

定理 4: Given a class $\mathcal{C}$ of subsets of Ω , there exists a unique smallest σ -field containing $\mathcal{C}$.

定义 5 **Borel Sets**: Suppose $\Omega = \mathbb{R}$ and let

$$\mathcal{C} = \{(a, b] : -\infty \leq a \leq b \leq +\infty\}$$

Then $\mathcal{B}(\mathbb{R}) := \sigma(\mathcal{C})$ the Borel subsets of $\mathbb{R}$.

There are many equivalent ways to define the Borel sets.

2. 随机事件与概率

定义 6 **事件域**: 设 S 为样本空间, $\mathcal{F}$ 为 S 的某些子集组成的集合类. 如果 $\mathcal{F}$ 满足下列条件, 称 $\mathcal{F}$ 为 S 的一个事件域.

- $S \in \mathcal{F}$
- $A \in \mathcal{F} \rightarrow \bar{A} \in \mathcal{F}$
- $A_n \in \mathcal{F}, n = 1, 2, \dots \rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$

定义 7 **概率**: 1933 年柯尔莫哥洛夫(Kolmogorov) 基于集合论给出.

设 S 为样本空间, $\mathcal{F}$ 是由 S 的某些子集组成的一个事件域. 如果对任意事件 $A \in \mathcal{F}$, 定义在 $\mathcal{F}$ 上的一个实值函数 $P(A)$ 满足

非负性公理 $A \in \mathcal{F} \rightarrow P(A) \geq 0$

正则性公理(规范性公理) $P(S) = 1$

可列可加性公理 若 $A_1, A_2, \dots, A_n, \dots$ 互斥, 则

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

那么称 P 为概率. $(S, \mathcal{F}, P)$ 为概率空间.

2.1. 几何概型

古典概型的局限是样本空间离散, 基本事件数有限.

当随机试验的样本空间是某连续区域 S , 并且任意一点落在度量(长度, 面积, 体积)相同的子区域是等可能的, 则事件 A 的概率可定义为

$$P(A) = \frac{m(A)}{m(S)}$$

几何概型基于现代的“测度”的概念,

2.2. 贝叶斯概率

贝叶斯概率的样本空间中的样本点为一系列**假设(hypotheses)**

3. 连续型随机变量

定义 8 **连续型随机变量**: 设 X 是随机变量, 若存在一个非负可积函数 $f(x)$, 使得

$$F(x) = \int_{-\infty}^x f(t) dt, -\infty < x < +\infty$$

则称 X 是**连续性随机变量**, 函数 $F(x)$ 是它的**分布函数(distribution function)**, 函数 $f(x)$ 是它的**概率密度函数**, 简称概率密度或密度函数.

分布函数的有用之处在于, 把连续型随机变量与离散型随机变量统一了起来.

定义 9 **指数分布**: 设 X 是一个连续型随机变量, 若它的概率密度函数为

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

则称 X 服从参数为 λ 的指数分布, 记为 $X \sim \text{Exp}(\lambda)$.

常作为各种“寿命”分布的近似

- 不稳定粒子的寿命
- 无线电元件的寿命

定理 10 **指数分布无记忆**: 若 $X \sim \text{Exp}(\lambda)$, 则 $P(X > s + t | X > s) = P(X > t)$. 指数分布是“永远年轻”的分布.

证明:

$$\begin{aligned} P(X > s + t | X > s) &= \frac{P(X > s + t)}{P(X > s)} \\ &= \frac{1 - P(X \leq s + t)}{1 - P(X \leq s)} \\ &= \frac{1 - F(s + t)}{1 - F(s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} = P(X > t) \end{aligned}$$

指数分布作为“寿命”分布的近似, 并不是例如人的寿命的实际分布.

几何分布作为也是无记忆的, 可以认为是离散型随机变量中的无记忆分布.

离散型随机变量是右连续的.

定义 11 **柯西分布**: 又称 Breit-Wigner 分布, 概率密度函数为

$$f(x; x_0, \gamma) = \frac{1}{\pi} \frac{\gamma}{(x - x_0)^2 + \gamma^2}$$

$x_0 = 0, \gamma = 1$ 的特例被称为标准柯西分布.

定理 12 **随机变量概率密度的复合**: 设随机变量 X 具有概率密度 $f_X(x)$, $-\infty < x < \infty$. 设函数 $g(x)$ 处处可导且恒有 $g'(x) > 0$ 或 $g'(x) < 0$. 则随机变量 $Y = g(X)$ 是连续型随机变量, 其概率密度为

$$f_Y(y) = \begin{cases} f_X(h(y)) |h'(y)| & \text{if } \alpha < y < \beta \\ 0 & \text{elsewhere} \end{cases}$$

其中 $h(y)$ 是 $g(x)$ 的反函数

$$\alpha = \min(g(-\infty), g(\infty)), \beta = \max(g(-\infty), g(\infty))$$

定理 13 **正态分布的可加性**: 对于 n 个独立正态随机变量之和 $Z = X_1 + X_2 + \dots + X_n$, 有

$$Z \sim N(\mu_1 + \mu_2 + \dots + \mu_n, \sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2)$$

定义 14 **Γ 分布**: 对于 $X \sim \Gamma(\alpha, \theta)$

$$f_X(x) = \begin{cases} \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\theta}} & \text{where } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

定义 15 **Γ 函数**:

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

定理 16 **Γ 分布的可加性**: 设 X, Y 互相独立, 分别服从参数为 $\alpha, \theta; \beta, \theta$ 的 Γ 分布, 则 $X + Y$ 服从参数为 $\alpha + \beta, \theta$ 的 Γ 分布.

证明: 注意使用概率密度的归一性. ■

定义 17 **Beta 函数**:

$$B(\alpha, \beta) =: \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt, \alpha, \beta > 0$$
$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

4. 二维随机变量

注意 X, Y 来自同一个样本空间, 意味着 X, Y 可以不独立.

5. 期望

定理 18 **柯西-施瓦茨不等式**:

$$E(X^2)E(Y^2) \geq E(XY)^2$$

6. 大数定律

定理 19 **Chebyshev 不等式**: 设随机变量 X 有数学期望 $E(X) = \mu$ 和方差 $\text{Var}(X) = \sigma^2$, 则:

$$\forall \varepsilon > 0, P(|x - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

定义 20 **Khinchin**: 辛钦大数定律是弱大数定律, 设 $X_1, X_2, \dots$ 是相互独立, 服从同一分布的随机变量序列, 且具有数学期望 $E(X_k) = \mu, k = 1, 2, \dots$, 则 $\forall \varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{1}{n} \sum_{k=1}^n X_k - \mu\right| < \varepsilon\right) = 1$$

即 $\bar{X} \xrightarrow{P} \mu$

推论 20.1 **Bernoulli**: 伯努利大数定律是辛钦大数定律的重要推论, 设 f_A 是 n 次独立重复试验中事件 A 发生的次数, p 是事件 A 在每次试验中发生的概率, 则 $\forall \varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{f_A}{n} - p\right| < \varepsilon\right) = 1$$

或

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{f_A}{n} - p\right| \geq \varepsilon\right) = 0$$

7. 极限定理

极限定理是概率论的核心内容之一.

中心极限定理 什么条件下 $\sum_{i=1}^n X_i$ 的分布收敛

于正态分布

- 独立不同分布
- 李雅普诺夫
- 独立同分布
- 林德伯格-列维
- 棣莫弗-拉普拉斯

定义 21 **依概率收敛**: 设 $X_1, X_2, \dots, X_n, \dots$ 是一个随机变量序列, a 是一个常数, 若对于任意的 $\varepsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P(|X_n - a| < \varepsilon) = 1$$

则称随机变量序列 $X_1, X_2, \dots, X_n, \dots$ 依概率收敛于常数 a , 记作 $X_n \xrightarrow{P} a$

定理 22 **Lindberg Levi**: 假设随机变量序列 $X_1, X_2, \dots$ 独立同分布, 且数学期望和方差存在

$$E(X_k) = \mu, \text{Var}(X_k) = \sigma^2 > 0$$

则随机变量之和 $X = \sum_{k=1}^n X_k$ 的标准化变量

$$Y_n = \frac{X - n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_n(x)$ 对于任意实数 x 满足

$$\lim_{n \rightarrow \infty} F_n(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x)$$

n 足够大时, X 近似服从 $N(n\mu, n\sigma^2)$

证明: 对于

$$Y_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n}\sigma} = \sum_{i=1}^n \frac{X_i - \mu}{\sqrt{n}\sigma}$$

$$\Rightarrow \varphi_{Y_n}(t) = \left(\varphi_{X_i} - \mu \left(\frac{t}{\sqrt{n}\sigma} \right) \right)^n$$

推论 22.1 **De Moivre-Laplace**: 这是 Lindberg-Levi 中心极限定理的二项分布特例.

设 $Y_n \sim b(n, p)$, $0 < p < 1$, $n = 1, 2, \dots$, 则 $\forall x \in \mathbb{R}$, 有

$$\lim_{n \rightarrow \infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt = \Phi(x)$$

如何理解中心极限定理?

卷积操作是高斯分布的不动点.

定理 23 **Lyapunov**: 李雅普诺夫定理说, 设 $X_1, X_2, \dots$ 是独立随机变量序列, 且数学期望和方差存在

$$E(X_k) = \mu_k, \text{Var}(X_k) = \sigma_k^2 > 0, 1 \leq k \leq n$$

记 $B_n^2 = \sum_{k=1}^n \sigma_k^2$. 如果存在 $\delta > 0$, 使得 Lyapunov 条件

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E(|X_k - \mu_k|) = 0$$

成立, 则随机变量之和 $X = \sum_{k=1}^n X_k$ 的标准化变量

$$Y_n = \frac{\sum_{i=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n}$$

的分布函数 $F_n(x)$ 对于任意 x 满足

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x)$$

例 24: 对于柯西分布, 由于其方差不存在, 所以中心极限定理不成立.

定理 25 **马尔科夫中心极限定理**: 对于相互不独立的随机变量, 即设 $X_1, X_2, \dots$ 是随机变量序列, 且满足无后效性, 即 $P(X_j | X_{j-1}, X_{j-2}, \dots) = P(X_j | X_{j-1})$ 且满足可逆性和可达性条件, 使它成为一个马尔科夫链, 那么

$$\mu = E(X_1)$$

$$\sigma^2 = \text{Var}(X_1) + 2 \sum_{k=1}^{\infty} \text{Cov}(X_1, X_{1+k}) < +\infty$$

8. 蒙特卡罗方法

定义 26 **第一类舍选法**: 希望采样分布 $X = x \in [a, b]$, $X \sim f(x)$ 且 $L = \max\{f(x) | a \leq x \leq b\}$

1. 对 $[a, b] \times [0, 1]$ 区域内均匀分布的二维随机变量 (U, V) 抽样

2. 保留曲线 $v = \frac{f(u)}{L}$ 下方的点, 取其横坐标

$$\left\{ u_i | v_i \leq \frac{f(u_i)}{L}, i = 1, 2, \dots, n \right\} = \{x_1, x_2, \dots, x_n\}$$

定理 27 **舍选法效率的期望**:

$$E = \frac{1}{(b-a)L}$$

证明: 舍选法可以采样到希望的分布,是因为

$$\begin{aligned} P(X \leq x) &= P\left(U \leq x | V \leq \frac{f(U)}{L}\right) \\ &= \frac{P\left(U \leq x, V \leq \frac{f(U)}{L}\right)}{P\left(V \leq \frac{f(U)}{L}\right)} \\ &= \frac{\int_a^x du \int_0^{\frac{f(u)}{L}} dv g(u, v)}{\int_a^b du \int_0^{\frac{f(u)}{L}} dv g(u, v)} \\ &= \int_a^x du f(u) \end{aligned}$$

其中 $g(u, v) = \frac{1}{b-a}$. 此外,效率的期望为

$$\begin{aligned} E &= P\left(V \leq \frac{f(U)}{L}\right) \\ &= \int_a^b du \int_0^{\frac{f(u)}{L}} dv g(u, v) \\ &= \frac{1}{(b-a)L} \end{aligned}$$

9. 分布族

定义 28 **指数分布族**: 一族的分布,它们的概率密度函数或分布律记为 $f(x|\vec{\theta})$,如果有如下的形式,则称为 s 维指数分布族

$$f(x|\vec{\theta}) = \exp\left(\sum_{i=1}^s \eta_i(\vec{\theta}) T_i(x) - B(\vec{\theta})\right) h(x)$$

定义 29 **指数分布族标准形式**: 把 $\eta_i(\vec{\theta})$ 当作自变量反解 θ , 那么形式有进一步化简.

$$f(x|\vec{\eta}) = \exp\left(\sum_{i=1}^s \eta_i T_i(x) - A(\vec{\eta})\right) h(x)$$

其中 $\eta_i = \eta_i(\vec{\theta})$, $A(\vec{\eta}(\vec{\theta})) = B(\vec{\theta})$

定理 30: 指数分布族的期望和方差为

$$E(X) = A'(\eta)$$

$$\text{Var}(X) = A''(\eta)$$

例 31 **二项分布属于指数分布族**:

$$\begin{aligned} P(k|p, n) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= h(k) \exp(T(k)\eta(p) - B(p)) \end{aligned}$$

其中

$$h(k) = \binom{n}{k}, T(k) = k, \eta(p) = \ln\left(\frac{p}{1-p}\right), B(p) = -n \ln(1-p)$$

例 32 **正态分布属于指数分布族**:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

10. 统计学概论

定义 33 **分位数**: 对于 α , 称 x_α, x'_α 为下侧 α -分位数和上侧 α -分位数, 若

$$P(X \leq x_\alpha) = F(x_\alpha) = \alpha, P(X \geq x'_\alpha) = \alpha$$

■

定义 34 **中位数**: $x_{0.5}$ 称为中位数.

定义 35 **箱线图**: 箱线图是一种用作显示一组数据分散情况的统计图表, 显示了一组数据的最大值, 最小值, 中位数, 上四分位数和下四分位数.

这五个分位数分别称为 $\text{Min}, Q_1, M, Q_3, \text{Max}$.

常用于比较不同组别的数据分布情况.

R 语言中分位函数(quantile function)可以用于计算对于给定概率的分位数.

11. 统计量

定义 36 **简单随机样本**: 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的一个样本, 若 $(X_1, X_2, \dots, X_n)$ 相互独立, 且都服从同一分布, 则称 $(X_1, X_2, \dots, X_n)$ 为来自总体 X 的一个简单随机样本.

定义 37 **样本均值**: 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

定义 38 **样本方差**: 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本

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

定义 39 **样本标准差**: 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本

$$S = \sqrt{S^2}$$

定义 40 **样本 k 阶原点矩**: 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本

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

定义 41 **样本 k 阶中心矩**: 设 $(X_1, X_2, \dots, X_n)$ 是来自总体 X 的样本

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

定理 42 **偏差平方和最小**: 数据观察值与样本均值的偏差平方和最小,即在形如 $\sum_{i=1}^n (X_i - c)^2$ 的函数中, $\sum_{i=1}^n (X_i - \bar{X})^2$

定理 43 **样本均值方差和矩的联系**:

- $E(\bar{X}) = \mu$
- $\text{Var}(\bar{X}) = \sigma^2/n$
- $E(S^2) = \sigma^2$

证明: 由于 $(X_1, X_2, \dots, X_n)$ 独立同分布于 X ,从而有

$$\begin{aligned} \text{Var}(\bar{X}) &= \sum_{i=1}^n \text{Var}\left(\frac{1}{n} X_i\right) \\ &= \sum_{i=1}^n \frac{1}{n^2} \text{Var}(X) \\ &= \frac{\sigma^2}{n} \end{aligned}$$

$$\begin{aligned} E(B_2) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) - E(\bar{X}^2) \\ &= \frac{1}{n} (n(\text{Var}(X) + E^2(X))) - (\text{Var}(\bar{X}) + E^2(\bar{X})) \\ &= \text{Var}(X) + \frac{E^2(X)}{n} - \frac{1}{n} \text{Var}(X) - \frac{1}{n} E^2(X) \\ &= \frac{n-1}{n} \text{Var}(X) \end{aligned}$$

由于 $E(B_2) = \frac{n-1}{n} E(S^2)$,从而得证. ■

再论中心极限定理

设 $X_1, X_2, \dots, X_n$ 是来自某个总体的样本, $\bar{X}$ 是样本均值.

- 若总体分布为 $N(\mu, \sigma^2)$,则 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- 若总体分布未知或不是正态分布,但 $E(X) = \mu$, $\text{Var}(X) = \sigma^2$ 存在,则 n 较大时, $\bar{X}$ 的渐进分布为 $N(\mu, \frac{\sigma^2}{n})$,参考[定理 23](#).

如何理解 χ^2 分布在自由度越大的时候越接近正态分布?

具有可加性的分布,因为中心极限定理,当自由度越大时,分布的形状越接近正态分布.

定义 44 **χ^2 分布**: 自由度为 n 的卡方分布是 n 个独立标准正态分布的平方和的分布.

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2 \sim \chi^2(n), X_i \sim N(0, 1)$$

$\chi^2(n)$ 分布的概率密度为

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

注意[定义 14](#), $\chi^2(1) \sim \Gamma(\frac{1}{2}, 2)$

定理 45 χ^2 分布是可加的:

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

定理 46 χ^2 分布的数学期望和方差: 设 $X \sim \chi^2(n)$

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

定义 47 **F分布**: 设随机变量 $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$, 且 X, Y 相互独立. 称

$$F = \left(\frac{X}{n} \right) / \left(\frac{Y}{m} \right)$$

为服从第一自由度为 n , 第二自由度为 m 的 F 分布.

$$f_F(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2}) \Gamma(\frac{n}{2})}{\Gamma(\frac{m}{2}) \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \left(1 + \frac{n}{m}x\right)^{-\frac{m+n}{2}} & \text{elsewhere} \\ 0 & \text{where } x \leq 0 \end{cases}$$

定理 48 **F分布的特性**:

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

定义 49 **T分布**: 设随机变量 $X \sim N(0, 1)$, $Y \sim \chi^2(n)$, 且 X, Y 相互独立. 称

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

为服从自由度为 n 的 T 分布.

$$f_T(t) = t f_{T^2}(t^2), t \in (-\infty, +\infty)$$

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

证明:

$$T^2 = \left(\frac{X^2}{1} \right) / \left(\frac{Y}{n} \right) \Rightarrow T^2 \sim F(1, n)$$

利用 $T = \sqrt{T^2}$ 推得 T 分布的概率密度函数. ■

定理 50 **T分布的性质**:

- $n \rightarrow \infty$ 时为标准正态分布
- $f_n(t)$ 是偶函数

```
library(ggplot2)
x <- seq(-5, 5, length.out=100)
t_df <- data.frame(x=c(), pd=c(), n=c())
# t 分布在 n 越大时越接近正态分布
for (n in c(1, 2, 9, 25, 3600)) {
  t_df <- rbind(t_df, data.frame(x=x, pd=dt(x,
    df=n), n=n))
}
t_df$n <- as.factor(t_df$n)
plot <- ggplot(t_df, aes(x=x, y=pd, color=n)) +
  geom_line()
plot + labs(y="概率密度", color="自由度")
ggsave("t.svg")
```

使用 R 语言绘制 $F(10, 2)$ 与 $T(10)$, 可以参考

[geom_function](#) 了解更多如何绘制连续函数

```
library(ggplot2)
base <- ggplot() + xlim(-5, 5)
base +
  geom_function(
    aes(colour = "F(10,2)"),
    fun = df,
    args=list(df1 = 10, df2 = 2)
  ) +
  geom_function(
    aes(colour = "T(10)"),
    fun = dt,
    args = list(df = 10)
  )
ggsave("f_t.svg")
```

定理 51 **正态总体的性质**: 设 $X_1, X_2, \dots, X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的样本, $\bar{X}$ 和 S^2 分别是样本均值和样本方差.

- $E(\bar{X}) = \mu, D(\bar{X}) = \frac{\sigma^2}{n}$
- $E(S^2) = \sigma^2$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
- $\bar{X}$ 与 S^2 相互独立
- $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

```
# install.packages("ggplot2")
```

R R

定理 52 **两个正态总体的样本均值和样本方差:**

设总体 $X \sim N(\mu, \sigma^2)$, 样本为 $(X_1, X_2, \dots, X_n)$, 又

设总体 $X' \sim N(\mu', \sigma'^2)$, 样本为 $(X'_1, X'_2, \dots, X'_n)$,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \bar{X}' = \frac{1}{n'} \sum_{i=1}^{n'} X'_i$$

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

则

$$\frac{S^2/S'^2}{\sigma^2/\sigma'^2} \sim F(n-1, n'-1)$$

12. 统计推断

12.1. 前言

参数估计与非参数估计

参数估计 当推断的对象是有限个, 例如高斯总体的期望, 方差

非参数估计 当推断的对象是无限个, 例如未知分布总体的期望, 方差, 分布

参数估计类型

点估计 估计未知参数的值

区间估计 估计未知参数的取值范围, 并使此范围包含未知参数真值的概率为给定的值

例 53: $X \sim N(\mu, \sigma^2)$, 若 μ, σ 未知, 通过构建统计量, 给出它们的估计值(点估计)或取值范围(区间估计)就是参数估计的内容.

12.2. 参数估计方法

定义 54 **点估计**: 从总体的一个样本估计未知参数的值称为点估计.

设总体 X 的分布函数的形式已知, θ 是待估参数, $(X_1, X_2, \dots, X_n)$ 为总体的一个样本.

点估计构造一个恰当的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$, 用它的观察值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 作为待估参数 θ 的近似.

常见的两种构造估计量的方法有 [定义 55](#) 以及 [定义 58](#).

定义 55 **矩估计**: 用样本 k 阶矩作为总体 k 阶矩的估计量, 建立含待估参数的方程, 从而解出待估参数.

设随机变量 $X \sim f(x; \theta_1, \theta_2, \dots, \theta_k)$, 其中

$\theta_1, \theta_2, \dots, \theta_k$ 为待估参数. 假设样本总体的前 k 阶矩存在

$$E(X^r) = \mu_r(\theta_1, \theta_2, \dots, \theta_k), 1 \leq r \leq k$$

假设 $(X_1, X_2, \dots, X_n)$ 为来自 X 的一个样本, r 阶样本矩 $A_r = \frac{1}{n} \sum_{i=1}^n X_i^r$.

A_r 及其函数依概率收敛于相应的总体矩. 因此可以

- 用样本矩作为相应的总体矩的估计量
- 用样本矩的函数作为相应的总体矩函数的估计量

总体的前 k 阶矩构成联立方程组, 含 k 个未知参数. 一般情况下可以解出这 k 个参数 $\theta_1, \theta_2, \dots, \theta_k$.

$$\begin{cases} \mu_1 = \mu_1(\theta_1, \theta_2, \dots, \theta_k) \\ \mu_2 = \mu_2(\theta_1, \theta_2, \dots, \theta_k) \\ \vdots \\ \mu_k = \mu_k(\theta_1, \theta_2, \dots, \theta_k) \end{cases} \Rightarrow \begin{cases} \theta_1 = \theta_1(\mu_1, \mu_2, \dots, \mu_k) \\ \theta_2 = \theta_2(\mu_1, \mu_2, \dots, \mu_k) \\ \vdots \\ \theta_k = \theta_k(\mu_1, \mu_2, \dots, \mu_k) \end{cases}$$

用样本矩 A_r 代替总体矩 μ_r , 阶得到待估参数的估计量, 称为 **矩估计量**.

$$\hat{\theta}_i = \theta_i(A_1, A_2, \dots, A_k), i = 1, 2, \dots, k$$

矩估计量的观测值称为 **矩估计值**.

例 56 **最大似然法的引入**: 设总体 X 服从 0-1 分布, 且 $P(X = 1) = p$, 用最大似然法求 p 的估计值 设 $x_1, x_2, \dots, x_n$ 为总体的样本的估计值, 则得到该样本值的概率为

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

对于不同的 p , 有 $L(p)$ 不同, 取 p 使这个事件发生的概率最大

$$\hat{p} = \arg \max L(p) = \arg \max \log L(p)$$

由于

$$\begin{aligned} \frac{d}{dp} \log L &= \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} \\ &\Rightarrow \hat{p} = \bar{x} \end{aligned}$$

定义 57 **似然函数**:

- 设 X 为离散型随机变量, 分布律为 $P(X = x) = p(x, \theta)$, 则似然函数定义为

$$L(\vec{x}, \theta) = \prod_{i=1}^n p(x_i, \theta)$$

- 设 X 是连续型随机变量, 取 $f(X, \theta)$ 为 X 的密度函数, 则似然函数定义为

$$L(\vec{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

定义 58 **最大似然估计**:

$$\hat{\theta}(\vec{x}, \theta) = \arg \max L(\vec{x}; \vec{\theta})$$

称为最大似然估计估计值, 称统计量

$\hat{\theta}(X_1, X_2, \dots, X_n)$ 为参数 $\vec{\theta}$ 的最大参数估计量.

用 R 语言做最大似然估计

```
file <- read.csv("samples.csv")
sample <- file$sample
exp_func <- function(theta, x) {
  n <- length(x)
```

```
logL <- n * log(theta) - theta * sum(x)
return(-logL)
}
theta <- optimize(exp_func, c(0, 1), x = sample)
print(1 / theta$minimum)
```

例 59 **均匀分布的矩估计和最大似然估计**: 设 $X \sim U(a, b)$, 有 $X_1, X_2, \dots, X_n$ 是 X 的一个样本. 求 a, b 的矩估计和最大似然估计.

解答: 使用矩估计有

$$\begin{aligned} \mu_1 &= E(X) = \frac{a+b}{2} \\ \mu_2 &= E(X^2) = \text{Var}(X) + E^2(X) \\ &= \frac{(b-a)^2}{12} + \frac{(a+b)^2}{4} \end{aligned}$$

解得

$$\begin{aligned} \vec{a} &= A_1 - \sqrt{3(A_2 - A_1^2)} = A_1 - \sqrt{3B_2} \\ \vec{b} &= A_1 + \sqrt{3(A_2 - A_1^2)} = A_1 + \sqrt{3B_2} \end{aligned}$$

使用最大似然估计有

$$L(a, b) = \begin{cases} \frac{1}{(b-a)^n} & \text{where } a < x_i < b \\ 0 & \text{elsewhere} \end{cases}$$

得到 $a \leq x_1, x_2, \dots, x_n, b \geq x_1, x_2, \dots, x_n$

当 $a = \min x_i, b = \max x_i$ 时, $L(a, b)$ 最大, 所以

$$\hat{a} = \min X_i, \hat{b} = \max X_i$$

这两种估计的结果不同. ■

定理 60 **最大似然估计的不变性**: 如果 $\hat{\theta}$ 是 θ 的最大似然估计.

那么 $g(\hat{\theta})$ 是 $g(\theta)$ 的最大似然估计.

12.3. 估计的评价

定义 61 **无偏性**: 若估计量 $\hat{\theta}$ 的数学期望存在, 且 $E(\hat{\theta}) = \theta$, 则称 $\hat{\theta}$ 是 θ 的无偏估计量.

无偏估计的实际意义是无系统误差.

例 62: 设 $X_1, X_2, \dots, X_n$ 是来自总体 $X \sim b(n, p)$ 的一个样本, 求 p^2 的无偏估计量

解答: 求出 $\hat{p}^2$ 后注意验证无偏性

定义 63 **有效性:** 若两个估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$ 都是 θ 的无偏估计量, 且对于任意 $\theta \in \Theta$, 有

$$\text{Var}(\hat{\theta}_1) \leq \text{Var}(\hat{\theta}_2)$$

则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

定义 64 **一致最小方差无偏估计:** 如果一个估计量比任何其它估计量都有效, 则称之为一致最小方差无偏估计 (uniformly minimum variance unbiased estimator, UMVUE)

定理 65 **最大似然估计量理论:**

- 如果参数存在有效无偏估计量, 那么它一定是最大似然估计量
- 一般情况下, 最大似然估计量是一致的
- 最大似然估计量渐进服从正态分布

定义 66 **相合性之一:** 设 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 是 θ 的估计量, 若

$$\hat{\theta}(X_1, X_2, \dots, X_n) \xrightarrow{P} \theta$$

那么称 $\hat{\theta}$ 是 θ 的相合估计量.

定义 67 **相合性之二:** 设 $\hat{\theta}_n(X_1, \dots, X_n)$ 是 θ 的估计量. 若

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta, \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$$

则 $\hat{\theta}_n$ 是 θ 的相合估计量.

定义 68 **均方误差:**

$$\text{MSE}(\hat{\theta}_n) = E(\hat{\theta}_n - \theta)^2 = \text{Var}(\hat{\theta}_n) + E^2(\hat{\theta}_n - \theta)$$

要求 MSE 最小就等价于 [定义 67](#).

13. 假设检验

问题 提出假设 H_0 , 如何判断 H_0 是否成立?

要求 使得 H_0 为真时拒绝 H_0 的概率不超过 α .

根据 α 可以定出合适的 **拒绝域** (其边界称为 **临界点**), 常见的正态分布的双边检验的临界点为 $k = z_{\alpha/2}$.

定义 69 **备择假设:** 在原假设被拒绝后可选择的假设称为备择假设.

定义 70 **一类错误:** 当 H_0 为真时拒绝 H_0 的错误称为一类错误.

定义 71 **二类错误:** 当 H_0 为假时接受 H_0 的错误称为二类错误.

定义 72 **显著性检验:** 只控制一类错误的概率, 不考虑二类错误, 称为显著性检验.

13.1. Z 检验

Z 检验是 σ^2 已知, 关于 μ 的检验.

定义 73 **Z 检验统计量:**

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

13.2. t 检验

t 检验是 σ^2 未知, 关于 μ 的检验.

定义 74 **t 检验:** 注意到 S^2 是 σ^2 的无偏估计

$$t = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

根据 [定理 51](#), 有 $t \sim t(n-1)$.

$$P_{\mu_0} \left(\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \geq k \right) = \alpha$$

从而对双边检验, 有

$$|t| \geq k = t_{\alpha/2}(n-1)$$

如果是单边检验, 注意更换分位数.

用 R 语言做单边检验的例子如下

```
sample_data <- c(16, 25, 21, 20, 23, 21, 19,
15, 13, 23, 17, 20, 29, 18, 22, 16, 22)
```



```
print(t.test(sample_data, mu = 21, alternative =
"less", conf.level = 0.95))
```

`t.test` 会给出 95% 置信区间的值

95 percent confidence interval:

-Inf 21.68713

Plain Text

说明接受原假设.

当然,也可以使用 `qt` 函数计算临界值并手动进行比较.

```
qt(0.05, 17 - 1)
```

R

13.3. 正态分布均值的比较

比较两样本均值是否相同,可以使用 t 检验.前提假设为两样本方差相等(尽管未知).

13.4. 正态分布方差的比较

定义 75 χ^2 检验: 根据定理 51, 有

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

如果取

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

对于双边检验,习惯上取

$$P_{\sigma_0^2} \left(\frac{(n-1)S^2}{\sigma_0^2} \leq k_1 \right) = \frac{\alpha}{2}$$

$$P_{\sigma_0^2} \left(\frac{(n-1)S^2}{\sigma_0^2} \geq k_2 \right) = \frac{\alpha}{2}$$

对于单边检验,例如 $H_0: \sigma^2 \leq \sigma_0^2$, 取

$$P_{\sigma^2 \leq \sigma_0^2} \left(\frac{(n-1)S^2}{\sigma^2} \geq \frac{(n-1)k}{\sigma_0^2} \right) = \alpha$$

可得拒绝域为

$$s^2 \geq k = \frac{\sigma_0^2}{n-1} \chi_a^2(n-1)$$

R 语言当中的 `chisq.test` 是做独立性检验的,为了检验 $H_0: \sigma^2 \leq \sigma_0^2$, 可以手动计算

```
a <- c(90, 105, 101, 95, 100, 100, 101, 105,
93, 97)
```

R

```
n <- length(a)
```

```
alpha <- 0.01
```

```
sigma0 <- 14
```

```
print(qchisq(1 - alpha, n - 1) * sigma0^2 / (n -
1))
```

```
print(var(a))
```

定义 76 **F 检验**: 对于两个不同总体 $N(\mu_1, \sigma_1^2)$ 以及 $N(\mu_2, \sigma_2^2)$, 考虑它们的方差的大小关系. 即 $H_0: \sigma_1^2 \leq \sigma_2^2$.

拒绝域的形式为

$$\frac{s_1^2}{s_2^2} \geq k = F_\alpha(n_1 - 1, n_2 - 1)$$

如果两总体方差相等,则称它们有**方差齐性**.

13.5. 方差分析

方差分析又称 F 检验(纪念 Fisher), Analysis of Variance, 简称 AoV/ANOVA.

定义 77 **偏差平方和**:

$$S_T = \sum_j^s \sum_i^{n_j} (X_{ij} - \bar{X})^2 = \sum_j^s \sum_i^{n_j} X_{ij}^2 - \frac{T_{..}^2}{n}$$

$$S_A = \sum_j^s n_j \bar{X}_{.j}^2 - n \bar{X}^2 = \sum_j^s \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{n}$$

$$S_E = S_T - S_A$$

$$T_{.j} = \sum_i^{n_j} X_{ij}, T_{..} = \sum_j^s T_{.j}$$

称 S_E 为**误差平方和**, S_A 为**效应平方和**.

定理 78 **S_A 与 S_E 的关系**:

- S_A 与 S_E 相互独立
- 因此平方和与自由度具有可加性

问题 如何判断 S_A 是否显著(相比 S_E)?

构造统计量

$$F = \frac{S_A/\nu_A}{S_E/\nu_E}$$

	自由度	平方和	均方	F比
	Df	Sum Sq	Mean Sq	F value
效应	$s - 1$	S_A	$\overline{S_A} = \frac{S_A}{s - 1}$	$\frac{\overline{S_A}}{\overline{S_E}}$
误差	$n - s$	S_E	$\overline{S_E} = \frac{S_E}{n - s}$	
总和	$n - 1$	S_T		

表 2 单因素试验的方差分析表

拒绝域 对于 $H_0: \mu_1 = \mu_2 = \dots = \mu_s$, 在显著性水平 α 下

$$F = \frac{\overline{S_A}}{\overline{S_E}} \geq F_{\alpha}(s - 1, n - s)$$

用 R 语言做单因素方差分析的例子为

```
value <- c(c(8, 6, 4, 2), c(6, 6, 4, 4),
c(8, 10, 10, 10, 12), c(4, 4, 2))
group <- c(rep("A", 4), rep("B", 4), rep("C", 5),
rep("D", 3))
data <- data.frame(group = group, value = value)
summary(aov(value ~ group, data))
print(qf(0.05, 3, 12))
```

13.6. 回归分析

定义 79 **回归函数**: 对于随机变量 Y 和普通变量 x , 如果 $E(Y)$ 存在并且值随着 x 的取值而定, 则 $E(Y)$ 是 x 的函数, 并且记为 $\mu(x)$

定义 80 **一元线性回归模型**: 假设对于 x 的每一个值, 有

$$Y = a + bx + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

未知参数 a, b 及 σ^2 都不依赖于 x .

定义 81 **经验回归方程**:

$$\hat{y} = \hat{a} + \hat{b}x$$

$$\hat{b} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

定义 82 **log1P**:

$$\log 1P(N) =: \log(1 + N)$$

定义 83 **泊松回归**: 观测量 y_i 相对于预测量 $E(Y_i) = e^{a+bx_i}$ 为泊松分布的回归模型, 称泊松回归.

泊松回归假设因变量 Y 是泊松分布, 并假设它期望值 $E(Y)$ 的对数 $\log E(Y)$ 可由一组未知参数进行线性表达.

定义 84 **泊松分布对数似然距离**: 预测值 $E(\hat{Y})$

定义 85 **变权迭代最小二乘法**:

13.7. 广义线性回归

定义 86 **指数离散分布族**: 指数离散分布族称为 **Exponential Dispersion Family**, 是指数分布族上再配上一项 $\varphi > 0$