第一讲　非参数估计分布函数及分位数

张伟平

统计与金融系

第一讲　简介

1.1 课程内容

■ 本课程主要介绍现代非参数统计的理论，包括

- 经验分布函数与统计泛函，再抽样技术，经验似然
- U 统计量
- 非参数密度估计，非参数回归
- 半参数模型，可加模型

■ 课程预修要求：
概率论，数理统计，回归分析，R 统计软件

■ 参考教材:

- 孙志华等，非参数与半参数统计，清华大学出版社，2016
- 薛留根，现代非参数统计，科学出版社，2015
- Härdle, W., Müller, M., Sperlich, St. and Werwatz, A.(2004) Nonparametric and Semiparametric Models. Springer Verlag, Heidelberg.
- Wasserman L. All of nonparametric statistics. Springer.

■ 成绩评定:

作业 (30%)、期末考试 (70%)

1.2 参数与非参数统计

■ 非参数统计是什么?

- Wolfowitz (1942):

  我们将这种情形（分布是完全的由其有限参数集的知识确定）称为参数情形，并将相反的情况（分布的函数形式是未知的）称为非参数情形。

- Randles, Hettmansperger and Casella (2004)

  非参数统计可以而且应该广义地定义为包括不使用基于参数分布族的模型之外的所有方法。

- Wasserman (2005)

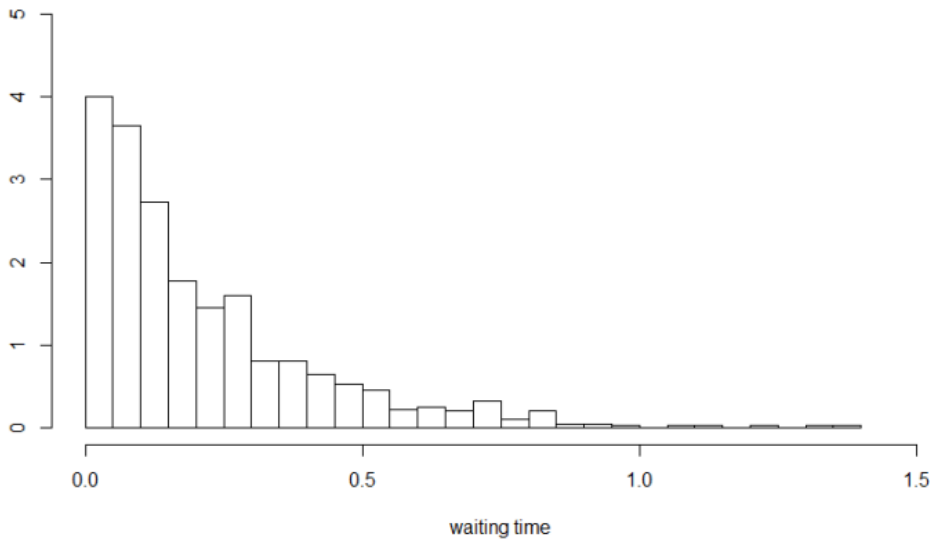
  非参数推断的基本思想是使用数据来推断一个未知的量，且假设尽可能的少。

参数统计

Definition

设数据 $X \sim P_\theta$，其中 $\theta \in \Theta \subset R^d$，$P_\theta$ 的分布形式已知，而 θ 未知，则称为参数模型。

例如：某 IT 服务柜台 799 位客户的等待时间：



公司要求至少 99% 的概率保证等待服务时间不超过 1 分钟，问基于该数据是否可以认为此要求达到？

我们建立一个概率模型。假设样本 $X_1, \ldots, X_n$ 来自下面的分布

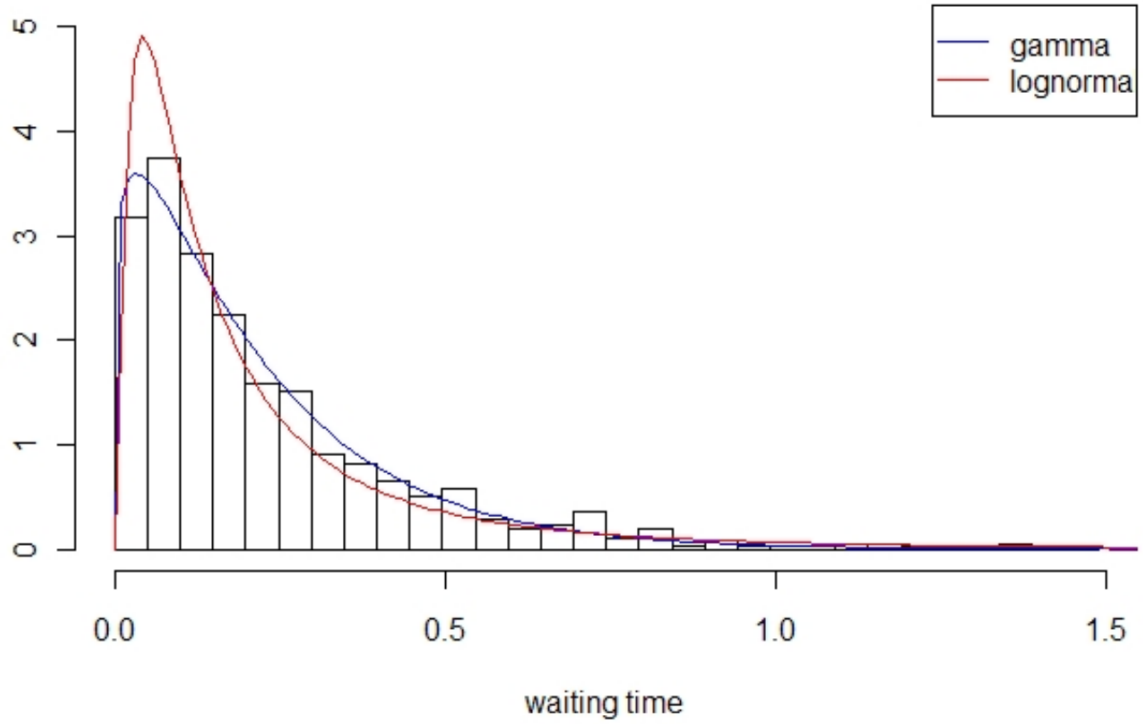
(A) Gamma 分布

$$p_{\theta,m}(x) = \frac{\theta e^{-\theta x}(\theta x)^{m-1}}{(m-1)!}, x > 0$$

(B) 或者对数正态模型

$$p_{\mu,\sigma}(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\Big(\frac{(logx-\mu)^2}{2\sigma^2}\Big)^2, x > 0$$

从而有



计算得到

$$P_A(W \leq 1) \approx 0.993 \qquad P_B(W \leq 1) \approx 0.968$$

参数统计的优缺点

优点：

- 方便：参数模型一般容易使用。
- 效率：如果参数模型是正确的，则参数方法比非参数方法更有效（但是非参数方法的效率损失经常很小）。
- 解释性：参数模型一般容易解释。

缺点：

- 有些时候不容易找到一个合适的参数模型。
- 错误假定（参数）模型的风险很高。

1.3 经验分布函数

设 $X_1, \ldots, X_n i.i.d \sim (unknown)F$，则

经验分布函数（ECDF）定义为

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le x\}.$$

Definition

其中

$$1\{X_i \le x\} = \begin{cases} 1, X_i \le x \\ 0, otherwise \end{cases}$$

性质

- 对固定的 x, 由于 $nF_n(x) \sim B(n, F(x))$, 因此

$$EF_n(x) = F(x), \quad Var(F_n(x)) = \frac{F(x)(1-F(x))}{n}.$$

- Glivenko-Cantelli Theorem (fundamental theorem of statistics): 对任意分布 F,

$$\sup_{x \in R} |F_n(x) - F(x)| \to 0, as.(n \to \infty)$$

- Dvoretzky-Kiefer-Wolfowith(DKW) Inequality: 对任意 $\epsilon > 0$, 任意 n,

$$P\Big(\sup_{x \in R} |F_n(x) - F(x)| > \epsilon \Big) \leq 2e^{-2n\epsilon^2}.$$

使用 DKW 不等式，我们可以建立 F 的一个置信带。记 $\alpha \in (0,1)$，

$$L_n(x) = \max\{F_n(x) - \sqrt{\frac{1}{2n} log \frac{2}{\alpha}}, 0\}$$

$$U_n(x) = \min\{F_n(x) + \sqrt{\frac{1}{2n} log \frac{2}{\alpha}}, 1\}$$

从而有

DKW confidence band:

对任意 CDF F 和所有的 n，

$$P(L_n(x) \leq F(x) \leq U_n(x), \ for \ all \ x) > 1 - \alpha.$$

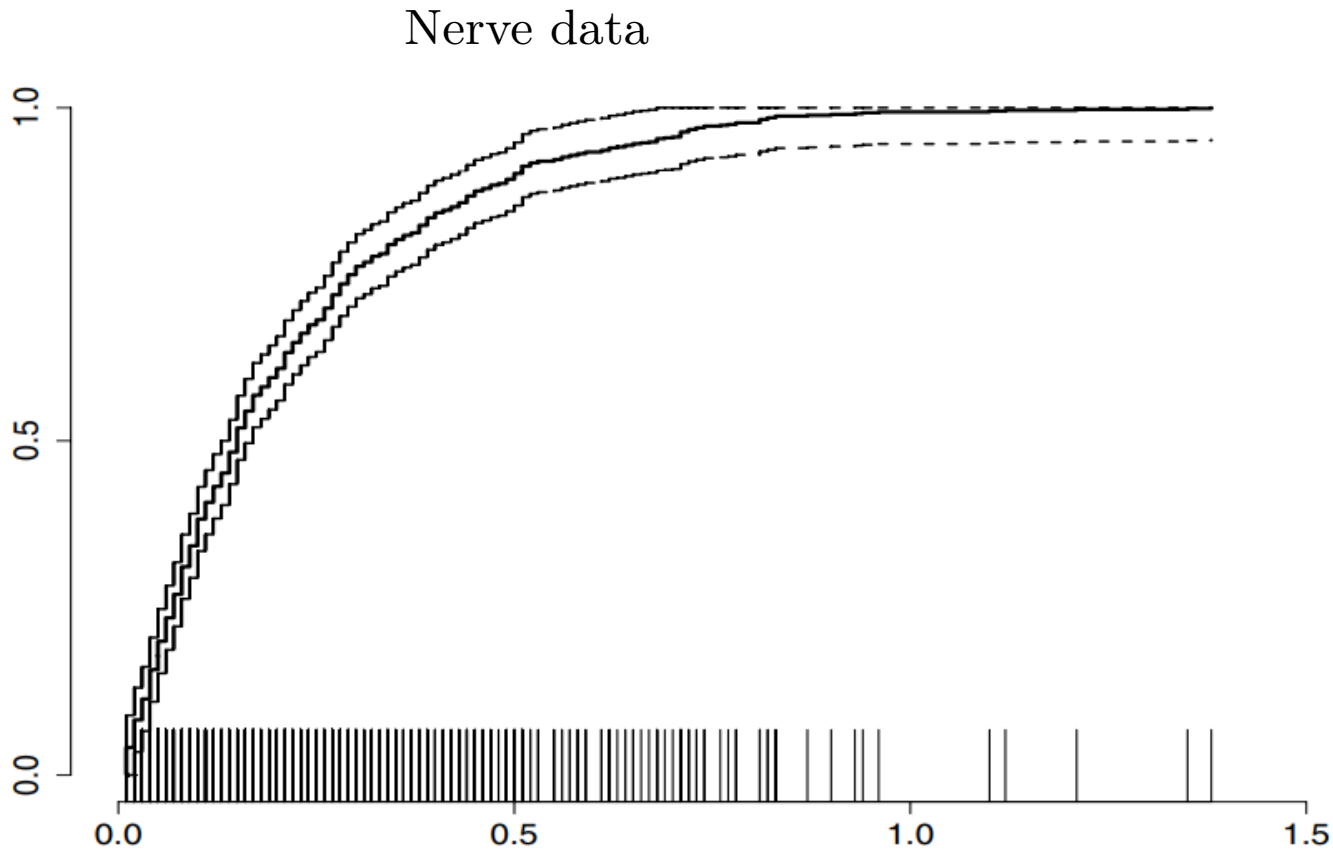


FIGURE 2.1. Nerve data. Each vertical line represents one data point. The solid line is the empirical distribution function. The lines above and below the middle line are a 95 percent confidence band.

在给定点 x 处 $F(x)$ 的置信区间: 寻找 l_n, u_n，使得

$$P(l_n \leq F(x) \leq u_n) \geq 1 - \alpha$$

由于 $nF_n(x) \sim B(n, F(x))$, 此问题可以等价为

$$Observe\ Y \sim B(n,p),\ find\ a\ CI\ for\ p$$

- 精确方法 (Clopper-Pearson)
- 渐近方法 (Wald)
- Wilson method
- 渐近方法（使用方差稳定化变换,VST)

1. Clopper-Pearson 置信区间

- $P_p(Y=k)=\binom{n}{k}p^k(1-p)^{n-k}, k=0,1,\ldots,n.$
- 对观测到的 y，Clopper-Pearson 置信区间定义为

$$[p_L, p_U]$$

其中 p_L 为下式的解

$$\frac{\alpha}{2}=P_{p_L}(Y\geq y)=\sum_{i=y}^{n}\binom{n}{i}p_L^i(1-p_L)^{n-i}.$$

以及 p_U 为下式的解

$$\frac{\alpha}{2}=P_{p_U}(Y\leq y)=\sum_{i=0}^{y}\binom{n}{i}p_U^i(1-p_U)^{n-i}.$$

此置信区间一般过于保守，其置信水平总是 $\geq 1-\alpha$。

2. Wald 置信区间

- 记 $\hat{p}_n = Y/n$, 则

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \to N(0,1)$$

- LLN: $\hat{p}_n \to p, a.s.$
- Slutsky 定理:

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \to N(0,1)$$

- 渐近 $1-\alpha$ 置信区间

$$\left[\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

3. Wilson 置信区间

- 利用 $\sqrt{n}\frac{\hat{p}_n-p}{\sqrt{p(1-p)}} \to N(0,1)$，对充分大的 n 有

$$P\Big(-z_{\alpha/2} \le \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \le z_{\alpha/2}\Big) \approx 1-\alpha$$

- 重新表示上述概率中的不等式，得到置信区间的端点为

$$\frac{\hat{p}_n + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)+z^2_{\alpha/2}/4n}{n}}}{1+z^2_{\alpha/2}/n}$$

4. 使用方差稳定化变换 (VST) 构造置信区间

- 设 ψ 为一个 $R \to R$ 且在 p 处导数非零的变换，根据 Delta 方法知

$$\sqrt{n}(\psi(\hat{p}_n) - \psi(p)) \to N(0, p(1-p)(\psi'(p))^2)$$

- 取 ψ 满足 $\psi'(p) = 1/\sqrt{p(1-p)}$，即 $\psi(p) = 2arcsin\sqrt{p}$
- 从而

$$Z_n(p) = 2\sqrt{n}(arcsin\sqrt{\hat{p}_n} - arcsin\sqrt{p}) \to N(0,1)$$

- 利用 $P(-z_{\alpha/2} \le Z_n(p) \le z_{\alpha/2}) \approx 1-\alpha$ 得到置信区间

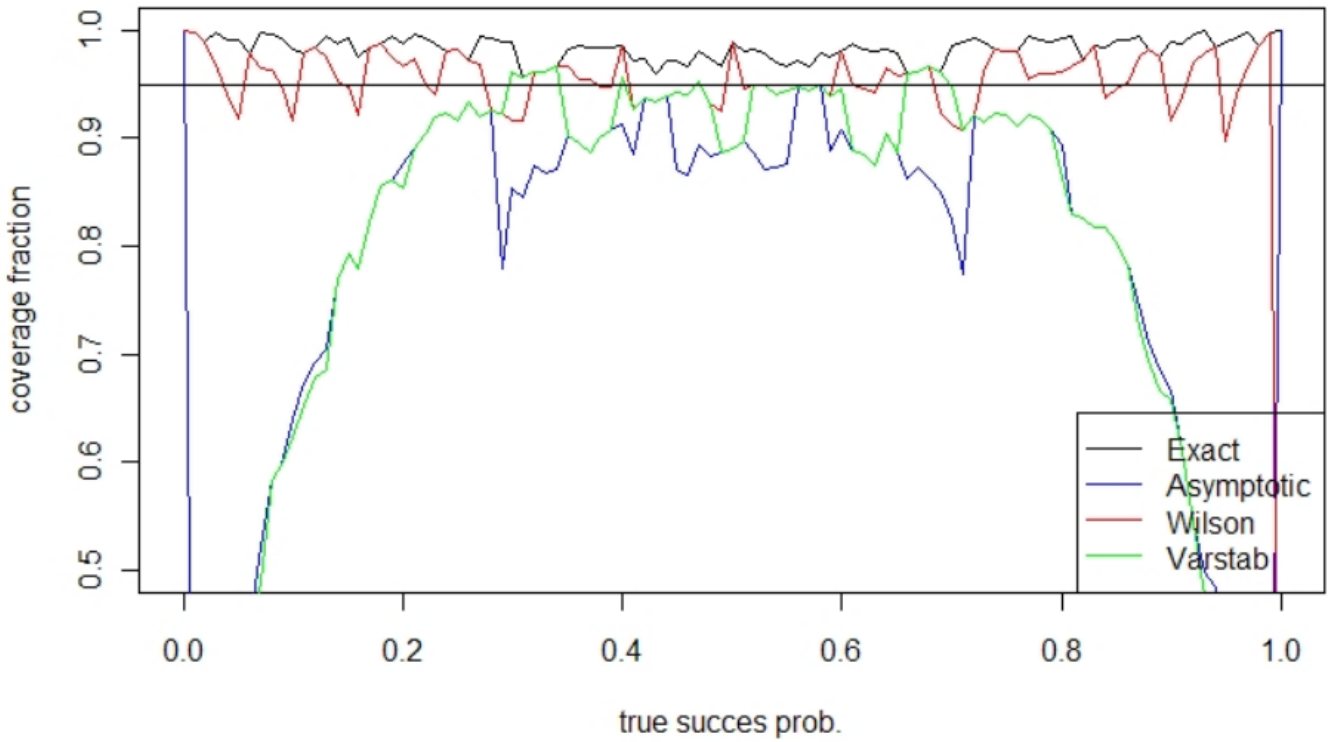
$$\left[sin^2\left(arcsin\sqrt{\hat{p}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right), sin^2\left(arcsin\sqrt{\hat{p}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right]$$

比较不同的置信区间

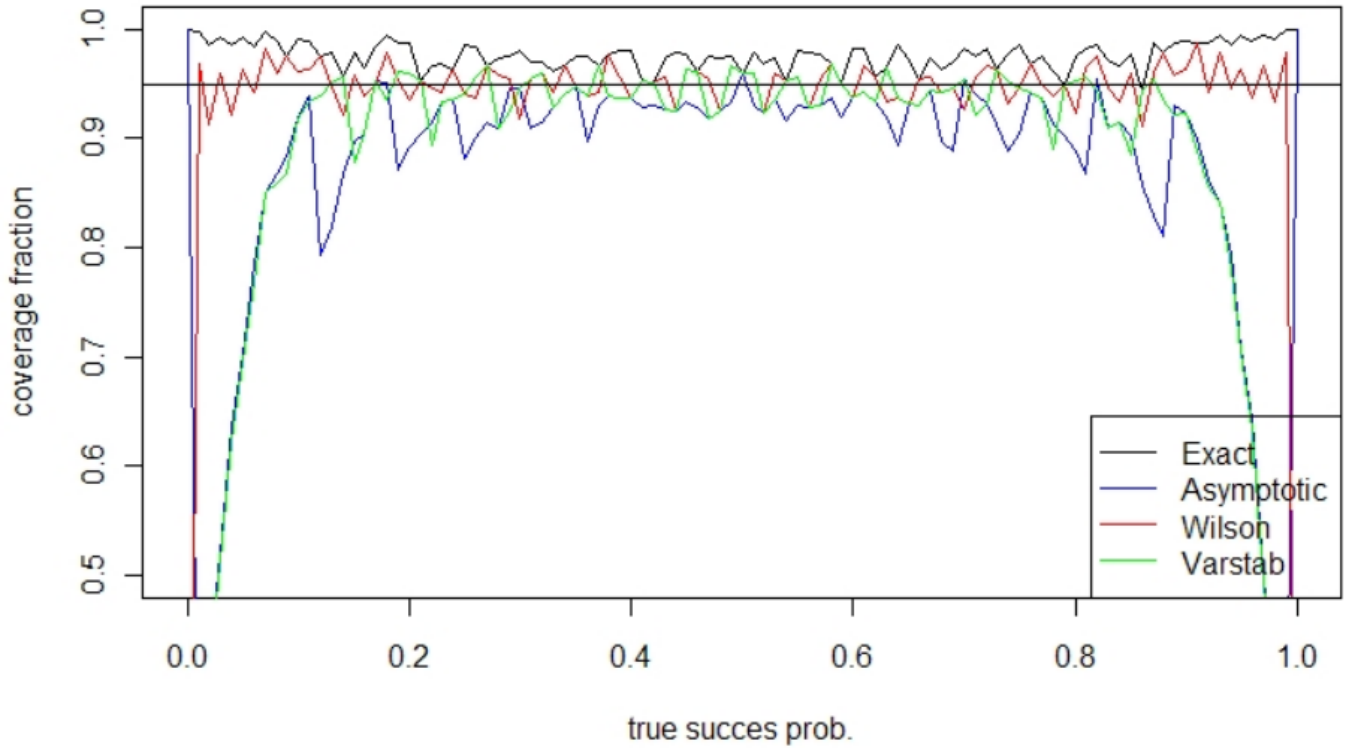
- 理论上比较覆盖概率不容易
- 利用 Monte Carlo 模拟来比较不同置信区间方法的实际覆盖率

- 选择 n 和 $p \in (0,1)$
- 产生 1000 个二项分布 $B(n,p)$ 随机数
- 对每个随机数，计算四种% 95 置信区间值
- 计算每种置信区间包含 p 值的比例

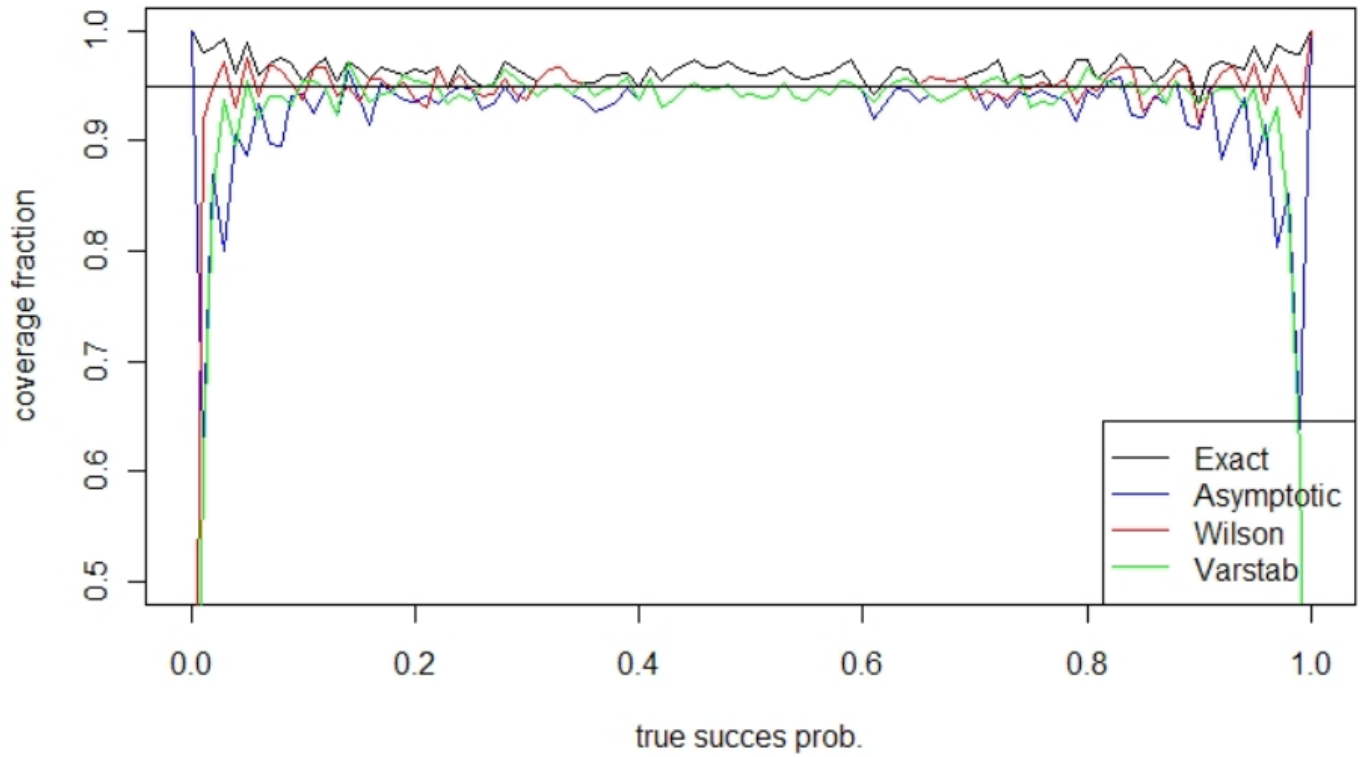
Coverage fractions of binomial C.I., n=10



Coverage fractions of binomial C.I., n=25



Coverage fractions of binomial C.I., n=100



1.4 估计分位数

- 分布 F 的 p 分位数定义为

$$\xi_p = F^{-1}(p) = \inf\{x|F(x) \geq p\}$$

- $F_n^{-1} \rightsquigarrow F^{-1}$ iff $F_n \rightsquigarrow F$

设随机变量 $X_1, \ldots, X_n$ $i.i.d$ $\sim F$，则记其次序统计量为

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

- 如果使用经验分布函数反函数 F_n^{-1} 来估计 F^{-1}，则 ξ_p 的估计量为次序统计量 $X_{(i)}, p \in (\frac{i-1}{n}, \frac{i}{n}]$.
- 构造 ξ_p 的 $1-\alpha$ 置信区间，主要想法是寻找 $r < s$，使得

$$P(X_{(r)} < \xi_p < X_{(s)}) \geq 1 - \alpha$$

- 首先注意到

$$P(X_{(r)} < \xi_p \leq X_{(s)}) = 1 - [P(X_{(r)} \geq \xi_p) + P(X_{(s)} < \xi_p)]$$

- 因此找 $r < s$，使得

$$P(X_{(r)} \geq \xi_p) \leq \alpha/2 \qquad P(X_{(s)} < \xi_p) \leq \alpha/2$$

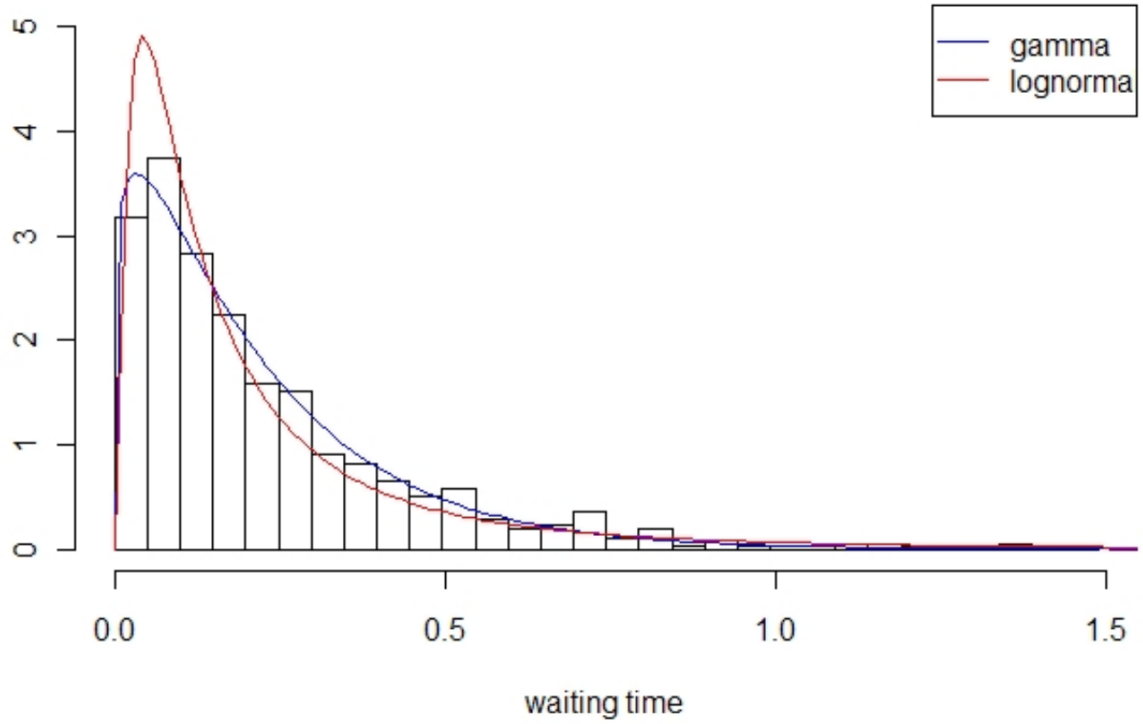
- $\{X_{(s)} < \xi_p\}$ 等价于至少有 s 个观测小于 ξ_p
- 记 $N = \sum_{i=1}^{n} 1\{X_i < \xi_p\}$，假设 F 是连续的，则 $P(X_1 < \xi_p) = P(X_1 \leq \xi_p) = p$. 则 $N \sim B(n, p)$ 且

$$P(X_{(s)} \leq \xi_p) = P(N \geq s) = \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

- 类似地，$P(X_{(r)} \geq \xi_p) = P(N < r)$
- 记 $X_{(0)} = -\infty, X_{(n+1)} = \infty$，则 ξ_p 的 $1-\alpha$ 置信区间为

$$(X_{(r)}, X_{(s)}]$$

其中 $r = \max\{k \in \{0, 1, \ldots, n\} : P(N < k) \leq \alpha/2\}$ 以及 $s = \min\{k \in \{1, \ldots, n+1\} : P(N \geq k) \leq \alpha/2\}$



计算得到

$$P_A(W \leq 1) \approx 0.993 \qquad P_B(W \leq 1) \approx 0.968$$

$\hat{\xi}_{0.99} \approx 0.90$, $\xi_{0.99}$ 的 95% 置信区间为 (0.81,1.21]

大样本性质

■ 设随机变量 $X_1, \ldots, X_n$ $i.i.d$, 其 ECDF 为 F_n, 总体分布函数 F 具有密度函数 f, 且 f 在 ξ_p 处连续满足 $f(\xi_p) > 0$。记 $\hat{\xi}_{n,p} = \inf\{x : F_n(x) \geq p\}$ 为样本 p 分位数, 则有

$$\sqrt{n}(\hat{\xi}_{n,p} - \xi_p) \rightsquigarrow N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$$

■ 对 $0 < p < 1$, ξ_p 是满足 $F(x) \geq p, F(x-0) \leq p$ 的总体 F 的 p 分位数. 如果 ξ_p 是唯一的, 则当 $n \to \infty$ 时候, $\hat{\xi}_{n,p} \to \xi_p$ a.s.

由此若记 $\hat{f}_n$ 为 f 的估计, 则可以得到 ξ_p 的渐近置信区间

$$\hat{\xi}_{n,p} \pm z_{1-\alpha/2}\sqrt{p(1-p)}/(\sqrt{n}\hat{f}_n(\hat{\xi}_{n,p}))$$