Lecture 1 - Introduction and the Empirical CDF

记录者：陈小树, Xiaoshu Chen

1 导入：非参数统计

- **didtribution-free**：不对训练样本的分布做出任何假设，而是仅仅假设样本是独立同分布于一个未知的总体分布
- **non-parametric**：如果模型没有设定有限维参数，我们称之为非参数模型
    - **parametric model**：模型能完全地被确定的有限维参数描述,$X \sim P_\theta$，其中$\theta \in \Theta \subseteq R^d$。参数模型的优点是，操作方便、效率高、预测简单；缺点是，有些问题难以找到合适的参数模型，通常仅适用于interval-scaled数据，对outlier敏感，易错误假定(mis-specification)。
    - **semiparametric**：参数(θ, η)，其中θ为欧几里得参数，η为无限维参数

2 非参数模型分布函数与分位数的估计

2.1 ECDF经验分布函数

- $\hat{F}_n(x) = \frac{1}{n}\sum\limits_{i=1}^{n} I\{X_i \leq x\}$，其中$X_i$是独立同分布一未知分布的变量。
    - 经验分布函数依概率（in probability）或者以概率1（almost surely）收敛到分布函数。
    - Chebyshev's inequality：$P(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq \frac{F(x)(1-F(x))}{n\epsilon^2}$, rather loose
    - Hoeffding's inequality：$P(|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$
    - DKW inequality：$P(sup_{x \in R}|\hat{F}_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$

2.2 置信带与置信区间（以二项分布为例）

- **Exact（Clopper-Pearson）计算二项分布的置信带**:
    - an obeserved y，$Y \sim Bin(n, p_0)$，有

$$\{p : P_p(Y \geq y) > \alpha/2 \& P_p(Y \leq y) > \alpha/2\}$$

    - “Exact”是因为知道真实的分布是二项分布，但此置信带通常保守（conservative），且由于Y的离散性（discreteness）不能够精确计算区间（exact coverage）。

> Let $\hat{p}_n = Y/n$，我们观察到：**（接下来的三种置信带都是在此基础上采取不同解法）**

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

- **Asymptotic:（Wald）**

- 利用中心极限定理计算比例的置信带，根据Slutsky's theorem将分母的p换成$\hat{p}_n$
- $\hat{p}_n = Y/n$

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

- **Asymptotic，using a variance stabilizing transformation变异数稳定变换**
  - 不用Slutsky's theorem，根据δ-method，取$\phi(x) = 2arcsin\sqrt{x}$

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

- **Wilson Method**
  - 直接根据正态分布求解

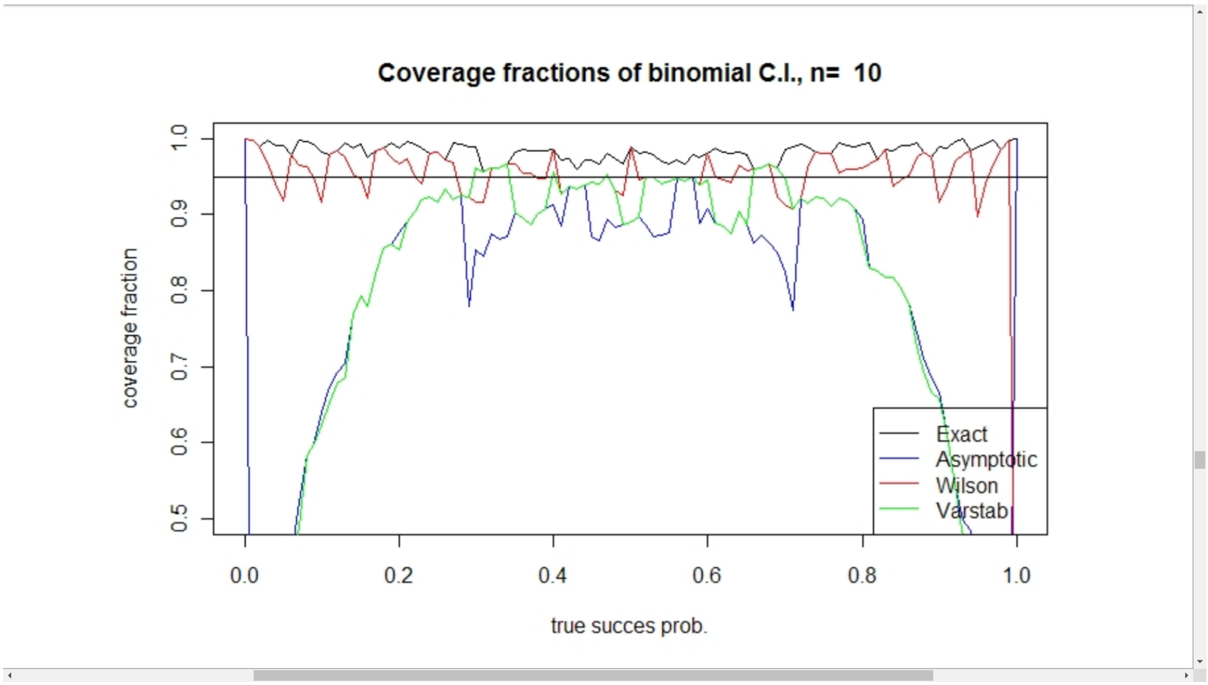
$$\frac{\hat{p}_n + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)+z^2_{\alpha/2}/(4n)}{n}}}{1+z^2_{\alpha/2}/2}$$

- **Hoeffding's inequality**：$\hat{F}_n(x) - \sqrt{\frac{1}{2n}log\frac{2}{\alpha}} \leq F(x) \leq \hat{F}_n(x) + \sqrt{\frac{1}{2n}log\frac{2}{\alpha}}$

> 💡 **置信带覆盖能力比较**：1. Exact比较保守，通常以高于$1-\alpha$的概率包含真值；2.Wilson表现很好，但是在边界值0,1上表现欠佳；3.变异数稳定变换下的近似，n变大时，覆盖效果提升.

实验通过模拟进行:

1. 设置真值p，以及抽取的样本数n
2. 模拟收取一次，并计算95%置信区间
3. 记录置信区间是否包含真值
4. 实验结束后，计算每组模拟的覆盖率（coverage fraction）



2.2 次序统计量与分位数

- p分位数的定义$F^{-1}(y) := inf\{x : F(x) \geq y\}$
- ECDF的另一种表达方式$\hat{F}_n(x) = \frac{1}{n}\sum\limits_{i=1}^{n} I\{X_{(i)} \leq x\}$

综合上述两个定义，$\hat{q}_n(p) = \hat{F}_n^{-1}(p)$可以作为$q_p$的估计量，并且我们注意到，如果$p \in (\frac{i-1}{n}, \frac{1}{n}]$，有$\hat{F}_n^{-1}(p) = X_{(i)}$。

- 构建p分位数的置信带
  - 注意到$\{X_r \leq u\}$与$\{\sum\limits_{i=1}^{n} I(X_i \leq u) \geq r\}$的等价性，我们可以得到如下概率公式

$$P(X_{(r)} \leq u) = p(\{\sum_{i=1}^{n} I(X_i \leq u) \geq r\}) = \sum_{i=r}^{n} C_n^i F(u)^i (1 - F(u))^{n-i}$$

  - 取$u = q_p = F^{-1}(p)$，可以得到分位数的置信带

$$p(X_{(r)} < q_p \leq X_{(s)}) = \sum_{i=r}^{s-1} C_n^i p^i (1-p)^{n-i}$$

- 基于上述部分的结论，我们可以证明，二项分布是泊松分布的一个近似
  - 若$lim_{n\to\infty} n(1 - F(u_n)) = \tau \in (0, \infty)$，有如下结论

$$lim_{n\to\infty} PX_{(n-k)} \leq u_n = e^{-\tau} \sum_{j=0}^{k} \frac{\tau^j}{j!}$$

  - 若$F \sim exp(\lambda), u_n = (log(n) - x)/\lambda$

$$lim_{n\to\infty} \lambda PX_{(n-k)} - log(n) \leq -x = e^{-e^x} \sum_{j=1}^{k} \frac{e^{jx}}{j!}$$

证明如下：

证明一：

记 $\tau_n = n(1 - F(u_n))$

$$\sum_{i=n-k}^{n} C_n^i F(u_n)^i (1 - F(u_n))^{n-i} = \sum_{i=n-k}^{n} \frac{n!}{i!(n-i)!} F(u_n)^i (1 - F(u_n))^{n-i}$$

$$= \sum_{i=n-k}^{n} \frac{n(n-1)(n-2)......(i+!)}{(n-i)!} (\frac{\tau_n}{n})^{n-i} (1 - \frac{\tau_n}{n})^i$$

$$= \sum_{i=n-k}^{n} \frac{{\tau_n}^{n-i}}{(n-i)!} (1 - \frac{1}{n})(1 - \frac{2}{n})......(1 - \frac{i-1}{n})(1 - \frac{\tau_n}{n})^i$$

对于固定的 i，有 $\lim_{n\to\infty} \tau_n^{n-i} = \tau^{n-i}$，$\lim_{n\to\infty}(1 - \frac{\tau_n}{n})^i = e^{-\tau}$

上式 $= \sum\limits_{i=n-k}^{n} \frac{\tau^{n-i}}{(n-i)!} e^{-\tau} = e^{-\tau} \sum\limits_{j=0}^{k} \frac{\tau_j}{j!}$

证明二：

$\lim_{n\to\infty} n(1 - F(u_n)) = \lim_{n\to\infty} n(1 - \int_0^{\frac{log(n)-x}{\lambda}} \lambda \cdot e^{-\lambda x}) dx = e^x$

Code for Simulation

```
library(Hmisc)  # contains function binconf

# function to compute CI based on variance stabilizing transformation
binconf.varstab<-function (x,size,alpha)
{
    len<-length(x)
    ci.matrix<-matrix(0,len,2)
    for (r in 1:len)
    {
        frac.obs<-x[r]/size
        h1<-asin(sqrt(frac.obs))
        h2<-0.5*qnorm(alpha/2,lower.tail=F)/sqrt(size)
        ci.matrix[r,]<-c((sin(h1-h2))^2,(sin(h1+h2))^2)
    }
    return(ci.matrix)
}

N<-1000                                      # number of times we compute a CI
size<-20                                    # number of bernoulli trials
pvec<-seq(0,1,by=0.01)                       # vector of true binomial probabilities
alf<-0.05                                    # compute 2-sided (1-alf)*100% CI
results<-matrix(0,length(pvec),4)

for (j in 1:length(pvec))
{
    p<-pvec[j]
    x<-rbinom(N,size,prob=p)

    res.exact<-binconf(x,size,method="exact",alpha=alf,include.x=TRUE,include.n=TRUE)
    res.asymp<-binconf(x,size,method="asymptotic",alpha=alf,include.x=TRUE,include.n=TRUE)
    res.wilson<-binconf(x,size,method="wilson",alpha=alf,include.x=TRUE,include.n=TRUE)
    res.varstab<-binconf.varstab(x,size,alpha=alf)

    # compute coverage fractions
    exact<-sum( (p>= res.exact[,4]) & (p<= res.exact[,5]) )/N
    asymp<-sum( (p>= res.asymp[,4]) & (p<= res.asymp[,5]) )/N
    wil<-sum( (p>= res.wilson[,4]) & (p<= res.wilson[,5]) )/N
    varstab<-sum( (p>= res.varstab[,1]) & (p<= res.varstab[,2]) )/N
    results[j,]<-c(exact,asymp,wil,varstab)
}

# make plot
plot(pvec,results[,1],type="l",col="black",xlab='true success prob.',ylab='coverage fraction',ylim=c(0.5,1))
lines(pvec,results[,2],col="blue")
lines(pvec,results[,3],col="red")
lines(pvec,results[,4],col="green")
title(paste('Coverage fractions of binomial C.I., n= ',as.character(size)))
legend("bottomright", c("Exact", "Asymptotic", "Wilson","Varstab"), col = c("black","blue","red","green"),merge=TRUE,lty=rep(1,
abline(h=1-alf)
```