دانشگاه علامه طباطبایی

کارگاه آموزشی
آمار ناپارامتری کاربردی

اولین سمینار تخصصی آمار ناپارامتری و کاربردهای آن

دکتر فرزاد اسکندری

دکتر احسان ارمز

حسین رضازاده

بهار ۱۳۹۵

۱ آمار ناپارامتری مقدماتی

۲ آزمون های نرمالیتی

۳ آزمون ها ی ناپارامتری

۴ برآورد تابع توزیع و چگالی

۵ رگرسیون ناپارامتری

❖آمار ناپارامتری:

روشهای آماری که بدون در نظر گرفتن توزیع آماری خاصی برای داده ها توسعه پیدا کرده اند.

❖ماهیت داده ها:

نوع آزمون	شرایط داده ها
پارامتری	✓ داده های نرمال ✓ داده های پیوسته غیرنرمال با حجم نمونه بیشتر از ۲۰ ✓ مرکزیت داده ها با میانگین بهتر توصیف می شود
ناپارامتری	✓ داده های رسته بندی شده ✓ داده های پیوسته غیر نرمال با حجم نمونه کمتر از ۲۰ ✓ مرکزیت داده ها با میانه بهتر توصیف می شود ✓ داده های پیوسته با نقاط تاثیرگذار غیرقابل حذف


آزمون های نرمالیتی
Kolmogorov-Smirnov
Anderson-Darling
Cramer-Von Mises
Lilliefors
Shapiro-Wilk
Shapiro-Francia
Jarque-Bera
D'Agostino's K-squared

آزمون کلموگروف اسمیرنف:

- فاصله بین توزیع تجربی نمونه را با توزیع احتمال نرمال مقایسه می کند.
- در حالت‌های بزرگ نمونه‌ای که تعداد گره ها کم باشد نتایج نسبتاً خوبی به دست می‌دهد.
- برای آزمودن سایر توزیع های پیوسته به جز نرمال نیز به کار می رود.
- در کل آزمون پرتوانی نیست به ویژه برای نمونه های کوچک.
- این آزمون به مرکز توزیع حساس تر از دم های توزیع است.

Package: stats

```
ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL)
```

Package: fBasics

```
ksnormTest(x, title = NULL, description = NULL)
```

آزمون اندرسون-دارلینگ:

- فاصله بین توزیع تجربی نمونه را با توزیع احتمال نرمال مقایسه می کند.
- نسبت به آزمون های کرامر- ون میزس و کلموگروف- اسمیرنف به مشاهدات موجود در دم ها وزن بیشتری می دهد.
- به خوبی آزمون شاپیرو ویلک نیست اما از بقیه آزمون ها بهتر عمل می کند.
- در عمل این آزمون می‌تواند برای نمونه‌هایی با اندازه بزرگتر از ۵ به کار رود

Package: nortest

```
ad.test(x)
```

Package: fBasics

```
adTest(x, title = NULL, description = NULL)
```

آزمون کرامر-ون میزس:

- فاصله بین توزیع تجربی نمونه را با توزیع احتمال نرمال مقایسه می کند.
- توان این آزمون بیشتر از آزمون کلموگروف اسمیرنف است.
- به مشاهدات موجود در دم ها نسبت به آزمون کلموگروف- اسمیرنف وزن بیشتر و نسبت به آزمون اندرسون دارلینگ وزن کمتری می دهد.

Package: nortest

cvm.test(x)

Package: fBasics

cvmTest(x, title = NULL, description = NULL)

آزمون لیلیفورس:

- توزیع تجربی نمونه را با توزیع احتمال تحت فرض صفر مقایسه می کند.
- بازنویسی شده آزمون کلموگروف اسمیرنف برای مواقعی است که پارامترهای توزیع نرمال مشخص نیست. در این موارد بهتر از آزمون کلموگروف اسمیرنف جواب می دهد.

Package: nortest

lillie.test(x)

Package: fBasics

lillieTest(x, title = NULL, description = NULL)

آزمون شاپیرو ویلک:

- این آزمون برای اندازه‌های نمونه بین ۳ تا ۲۰۰۰ بکار می‌رود.
- برای اندازه نمونه کم با گره های کم بهترین آزمون است و توان بالایی دارد.
- قادر است انحراف از نرمال بودن را با توجه به چولگی یا کشیدگی یا هر دو تشخیص دهد.
- مقادیر آماره آزمون بین صفر و یک قرار می‌گیرد و مقادیر کوچک منجر به رد فرض نرمالیتی داده‌ها می‌شود.

Package: stats

```
shapiro.test(x)
```

Package: fBasics

```
shapiroTest(x, title = NULL, description = NULL)
```

آزمون شاپیرو فرانسیا:

- آماره شاپیرو- فرانسیا تصحیح شده آماره شاپیرو- ویلک است.
- این آزمون برای اندازه‌های نمونه بین ۵ تا ۵۰۰۰ بکار می‌رود.
- برای اندازه نمونه کم با گره های کم بهترین آزمون است و توان بالایی دارد.

Package: nortest

```
sf.test(x)
```

Package: fBasics

```
sfTest(x, title = NULL, description = NULL)
```

آزمون جارکو-برا:

- آماره این آزمون به طور مجانبی دارای توزیع کای دو با ۲ درجه آزادی است.
- میزان تطابق چولگی و کشیدگی داده ها را با توزیع نرمال می سنجد.
- برای اندازه های نمونه کوچک خطای نوع اول این آزمون زیاد است.
- بهتر است برای نمونه های بالای ۲۰۰۰ از تقریب نرمال استفاده شود.
- توصیه می شود همراه با آزمون های دیگر نرمالیتی استفاده شود. چرا که مواردی وجود دارد که چولگی و کشیدگی مانند توزیع نرمال است اما توزیع نرمال نیست!!!

Package: fBasics

```
jarque.bera.test(x)
```

Package: tseries

```
jarque.bera.test(x)
```

آزمون مربع k دی آگوستین:

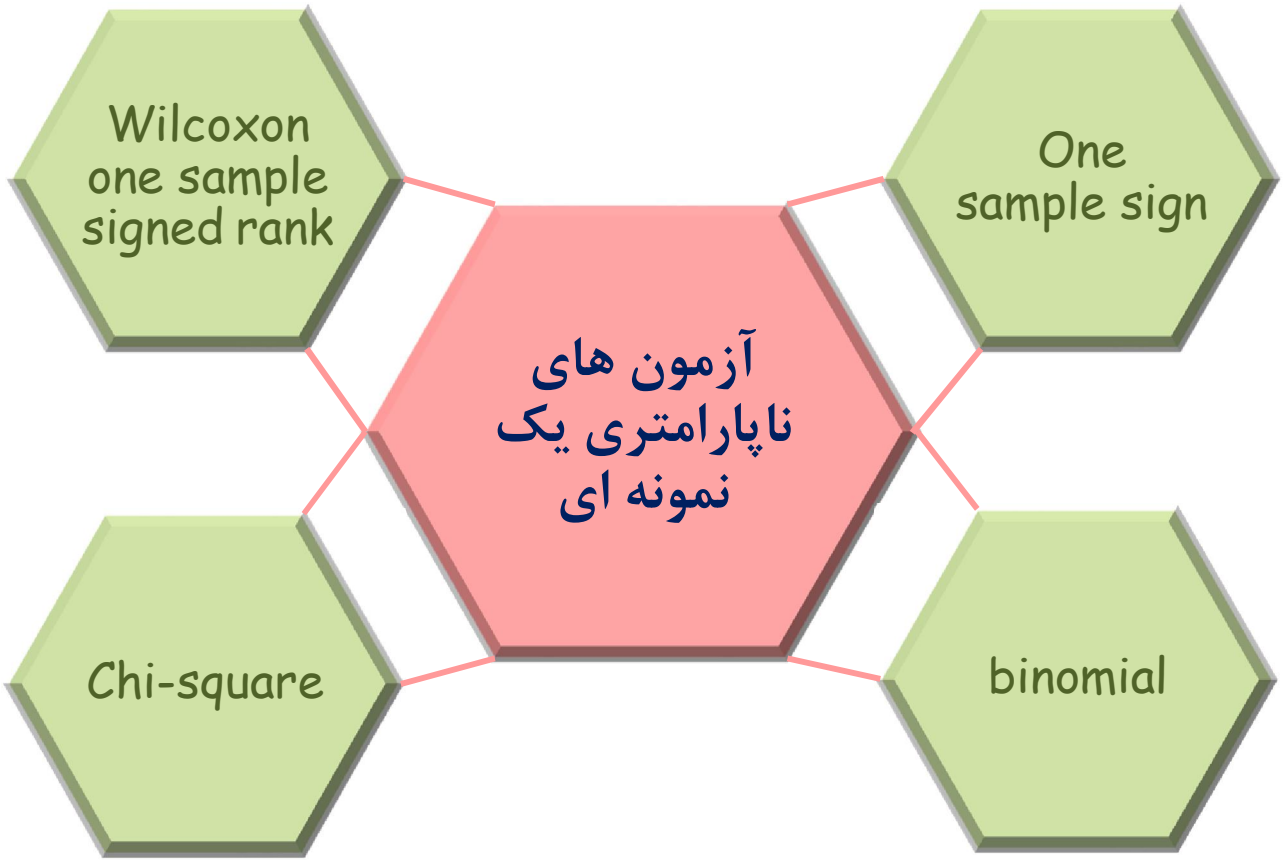
- این آزمون نیز همانند آزمون جارکو ـ برا به بررسی انحراف چولگی و کشیدگی داده‌ها از چولگی و کشیدگی توزیع نرمال می‌پردازد
- آماره این آزمون دارای توزیع کای دو با ۲ درجه آزادی است.
- توصیه می شود همراه با آزمون های دیگر نرمالیتی استفاده شود. چرا که مواردی وجود دارد که چولگی و کشیدگی مانند توزیع نرمال است اما توزیع نرمال نیست!!!

Package: fBasics

```
dagoTest(x, title = NULL, description = NULL)
```

Package: moments

```
agostino.test(x, alternative = c("two.sided", "less", "greater"))
```


Wilcoxon one sample signed rank
One sample sign
آزمون های ناپارامتری یک نمونه ای
Chi-square
binomial

آزمون علامت یک نمونه ای:

- برای آزمون برابری، کوچکی یا بزرگی میانه یک نمونه با یک عدد ثابت استفاده می شود.
- وقتی مشاهدات حول میانه نامتقارن بوده و در داده ها چولگی وجود داشته باشد، استفاده می شود.
- در تابع R از آرگومان y استفاده نمی کنیم.

Package: BSDA

```
SIGN.test(x, md = 0, alternative = "two.sided", conf.level = 0.95)
```

آزمون رتبه‌ای علامت‌دار ویلکاکسون یک نمونه ای:

- برای آزمون متقارن بودن توزیع مشاهدات حول میانگین به کار می رود.
- متغیرها حداقل رتبه بندی شده هستند
- برای متغیرهای پیوسته بهتر جواب می دهد.
- برای متغیرهای اسمی مناسب نیست.
- در تابع R از آرگومان y استفاده نمی کنیم.

Package: stats

```
wilcox.test(x, alternative = c("two.sided", "less", "greater"), mu = 0, exact =
NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95, ...)
```

آزمون دو جمله ای:

- یک آزمون تطابق توزیع برای داده‌های اسمی است.
- متغیر مورد بررسی فقط دو رده باید داشته باشد.
- مشاهدات از هم مستقل باشند.

Package: stats

```
binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)
```

آزمون کای دو:

- کار اصلی این آزمون بررسی معناداری تفاوت بین فراوانی های مشاهده شده و مورد انتظار است
- این آزمون تنها راه برای آزمون همگنی در مورد متغیرهای مقیاس اسمی با بیش از دو رسته است.
- این آزمون نسبت به حجم نمونه حساس است.
- کل رسته هایی که تعداد مشاهدات آنها کمتر از ۵ است، نباید بیش از ۲۰ درصد کل رسته ها باشد.
- مشاهدات از هم مستقل باشند.

Package: stats

```
chisq.test(x,  correct = TRUE, p = rep(1/length(x), length(x)), rescale.p =
FALSE, simulate.p.value = FALSE, B = 2000)
```


sign
McNemar
Wilcoxon rank signed
آزمون های ناپارامتری دو نمونه ای
Mood's Median
Kolmogorov Smirnov
U Mann-Whitney

آزمون مک نمار:

- برای داده های زوج شده استفاده می شود.
- متغیر مورد بررسی دو رسته ای با مقیاس اسمی و یا رتبه ای است.

Package: stats

```
mcnemar.test(x, y = NULL, correct = TRUE)
```

آزمون علامت:

- برای داده های زوج شده استفاده می شود.
- وقتی تفاضل زوج مشاهدات حول میانه اش نامتقارن بوده و توزیع آنها چولگی داشته باشد، از این آزمون استفاده می شود.
- .اگر تفاضل زوج مشاهدات صفر شود به آن گره میگویند و از مجموعه داده حذف می شود.

Package: BSDA

```
SIGN.test(x, y, alternative = "two.sided", conf.level = 0.95)
```

آزمون رتبه‌ای علامت‌دار ویلکاکسون:

- برای آزمون متقارن بودن توزیع تفاضل زوج مشاهدات حول میانگین به کار می رود.
- متغیرها حداقل رتبه بندی شده هستند.
- برای متغیرهای پیوسته بهتر جواب می دهد.
- برای متغیرهای اسمی مناسب نیست.

Package: stats

wilcox.test(x, y, alternative = c("two.sided", "less", "greater"), paired=TRUE, exact = NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95, ...)

آزمون کلموگروف اسمیرنف:

- فاصله بین توزیع تجربی دو نمونه مستقل از هم را مقایسه می کند.
- در حالت‌های بزرگ نمونه‌ای که تعداد گره ها کم باشد نتایج نسبتاً خوبی به دست می‌دهد.
- در کل آزمون پرتوانی نیست به ویژه برای نمونه های کوچک.

Package: stats

```
ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL)
```

Package: fBasics

```
ksnormTest(x, title = NULL, description = NULL)
```

آزمون U من ویتنی:

- برای مقایسه دو نمونه مستقل از هم استفاده می شود.
- متغیر مورد بررسی دارای مقیاس رتبه ای یا پیوسته است.
- تعداد گره ها کم باشد نتایج بهتر است.
- وقتی اندازه هر کدام از دو نمونه مستقل حداقل ۳۰ باشد توزیع مجانبی آماره U نرمال است.

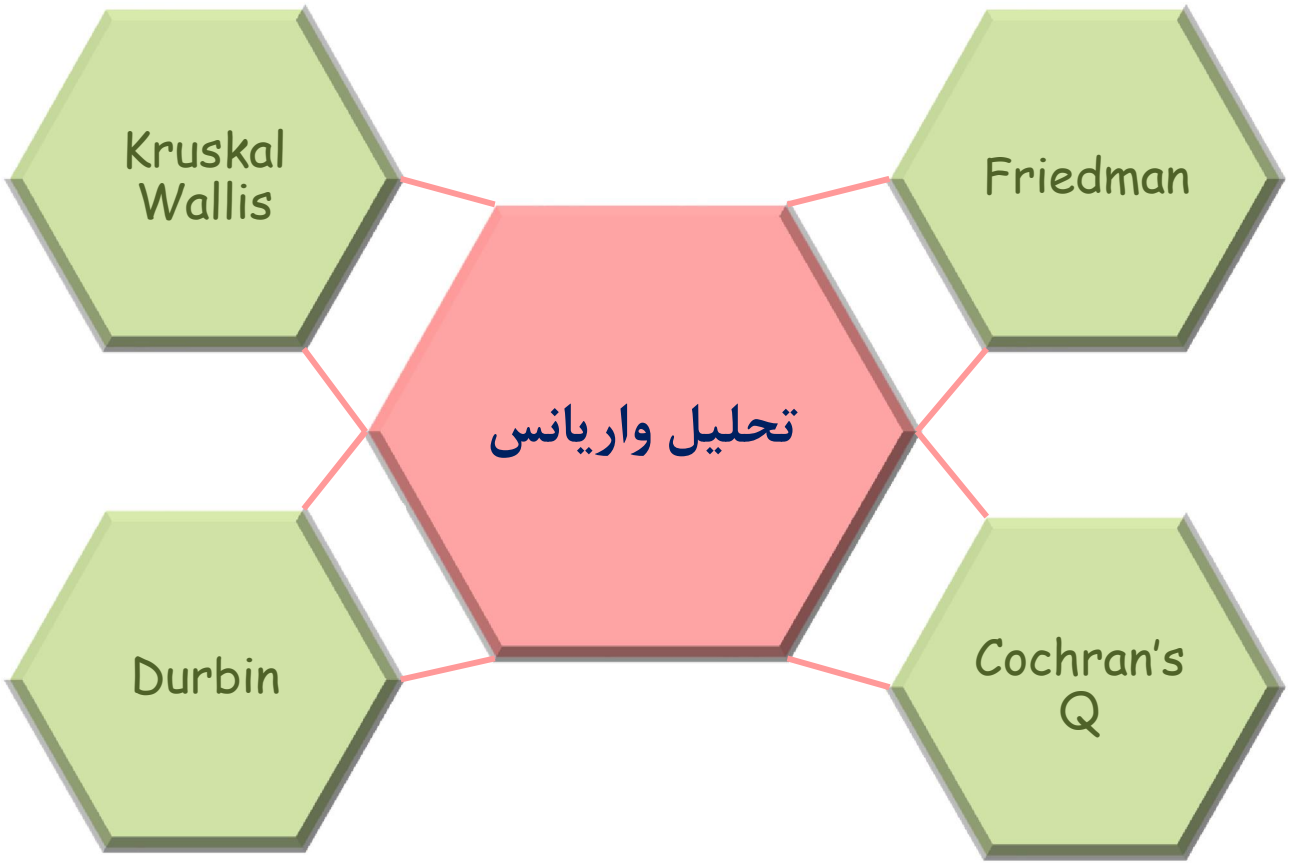
Package: stats

```
wilcox.test(x, y , alternative = c("two.sided", "less", "greater"), paired =
FALSE, exact = NULL, correct = TRUE, conf.int = FALSE, conf.level = 0.95, ...)
```

آزمون میانه مود:

- برای مقایسه دو جامعه مستقل به کار می رود.
- توانش از آزمون های دیگر کمتر است.
- مقیاس متغیرها حداقل رتبه ای است.
- تعداد گره ها زیاد باشد نتایج نامعتبر است.

```
median.test<-function(x,y){ z<-c(x,y) g <- rep(1:2, c(length(x),length(y)))
m<-median(z) fisher.test(z<m,g)$p.value }
```


Kruskal Wallis
Friedman
تحلیل واریانس
Durbin
Cochran's Q

آزمون کروسکال والیس:

- آزمونی برای انجام تحلیل واریانس یکطرفه است.
- توسعه آزمون U من ویتنی برای مقایسه بیش از دو نمونه مستقل است.
- مقیاس متغیر مورد بررسی حداقل رتبه ای است.
- توزیع مجانبی آماره آزمون، کای دو با درجه آزادی برابر تعداد نمونه های مورد مقایسه منهای یک است..

Package: stats

kruskal.test(x, g, ...)

آزمون فریدمن:

- آزمونی برای انجام تحلیل واریانس دو طرفه (طرح بلوک تصادفی کامل) یا تحلیل واریانس با اندازه های تکراری است.
- مقیاس متغیر مورد بررسی حداقل رتبه ای است.

Package: stats

friedman.test(y, groups, blocks, ...)

آزمون Q کوکران:

- برای تحلیل طرح های بلوکی تصادفی کامل با متغیر پاسخ دودویی مناسب است.
- آزمون کوکران به مقایسه بیش از دو گروه مستقل وقتی متغیر مورد بررسی دودویی است، می پردازد.
- تعمیمی از آزمون مک نمار است.
- آماره آزمون به صورت مجانبی دارای توزیع کای دو با درجه آزادی برابر تعداد گروه ها منهای یک است.

Package: RVAideMemoire

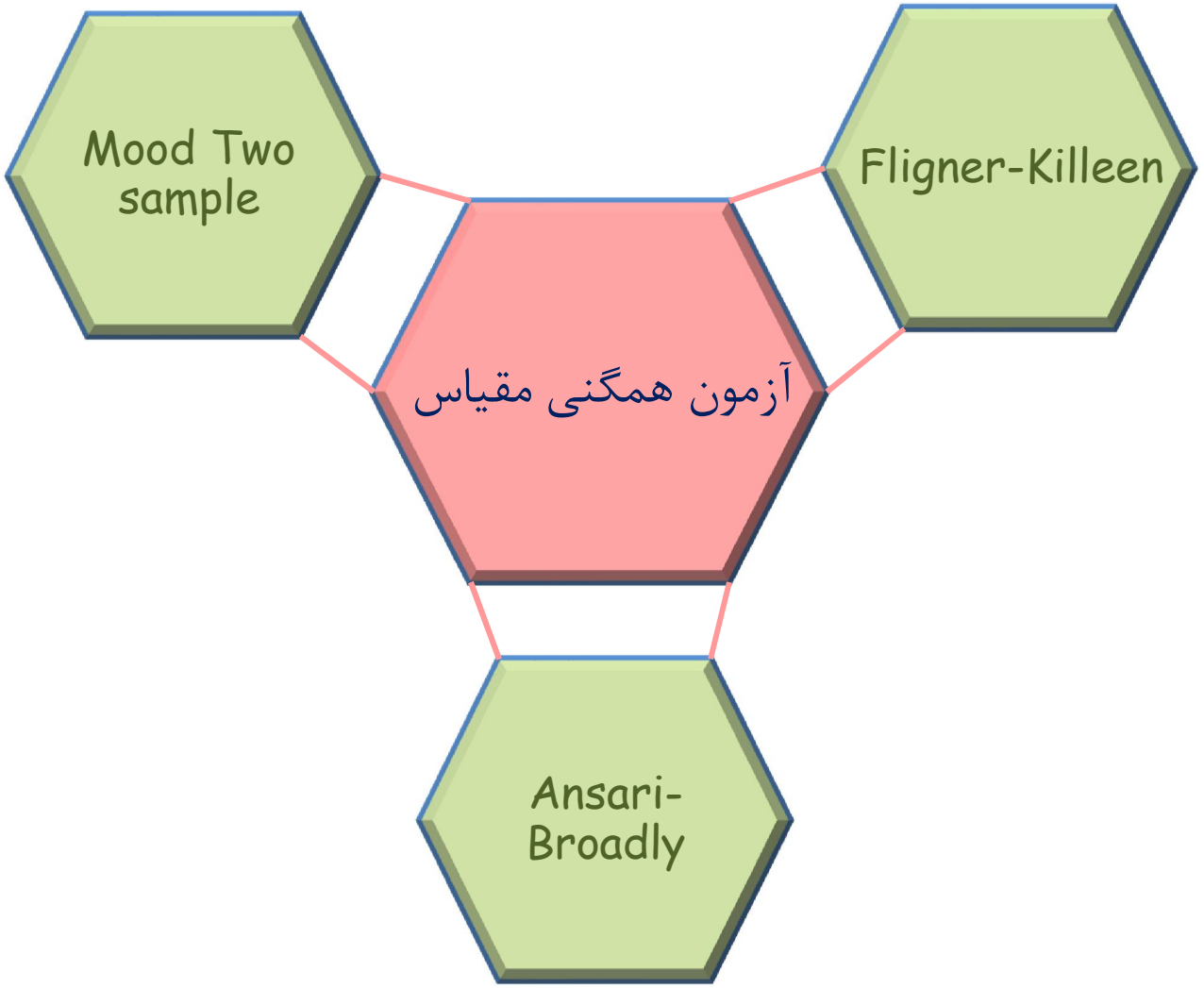
```
cochran.qtest(formula, data, alpha = 0.05, p.method = "fdr")
```

آزمون دوربین:

- تعمیمی از آزمون فریدمن برای طرح های بلوکی تصادفی ناقص بالانس است.
- متغیر مورد بررسی حداقل مقیاس رتبه ای دارد.

Package: agricolae

durbin.test(judge, trt, evaluation, alpha = 0.05, group =TRUE, main = NULL, console=FALSE)


Mood Two sample
Fligner-Killeen
آزمون همگنی مقیاس
Ansari-Broadly

آزمون مود دو نمونه ای، آزمون انصاری-بردلی، آزمون فلیگنر-کیلین

- اختلاف در پارامتر مقیاس دو نمونه مستقل را می سنجند.

Package: stats

- mood.test(x, y, alternative = c("two.sided", "less", "greater"), ...)
- ansari.test(x, y, alternative = c("two.sided", "less", "greater"), exact = NULL, conf.int = FALSE, conf.level = 0.95, ...)
- fligner.test(x, g, ...)

Estimating the CDF

Let $X_1, \ldots, X_n \sim F$ *and let* $\widehat{F}_n$ *be the empirical* CDF $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$

1. *At any fixed value of* x,

$$\mathbb{E}\left(\widehat{F}_n(x)\right) = F(x) \quad \text{and} \quad \mathbb{V}\left(\widehat{F}_n(x)\right) = \frac{F(x)(1-F(x))}{n}.$$

Thus, $\mathsf{MSE} = \frac{F(x)(1-F(x))}{n} \to 0$ *and hence* $\widehat{F}_n(x) \xrightarrow{\mathrm{P}} F(x)$.

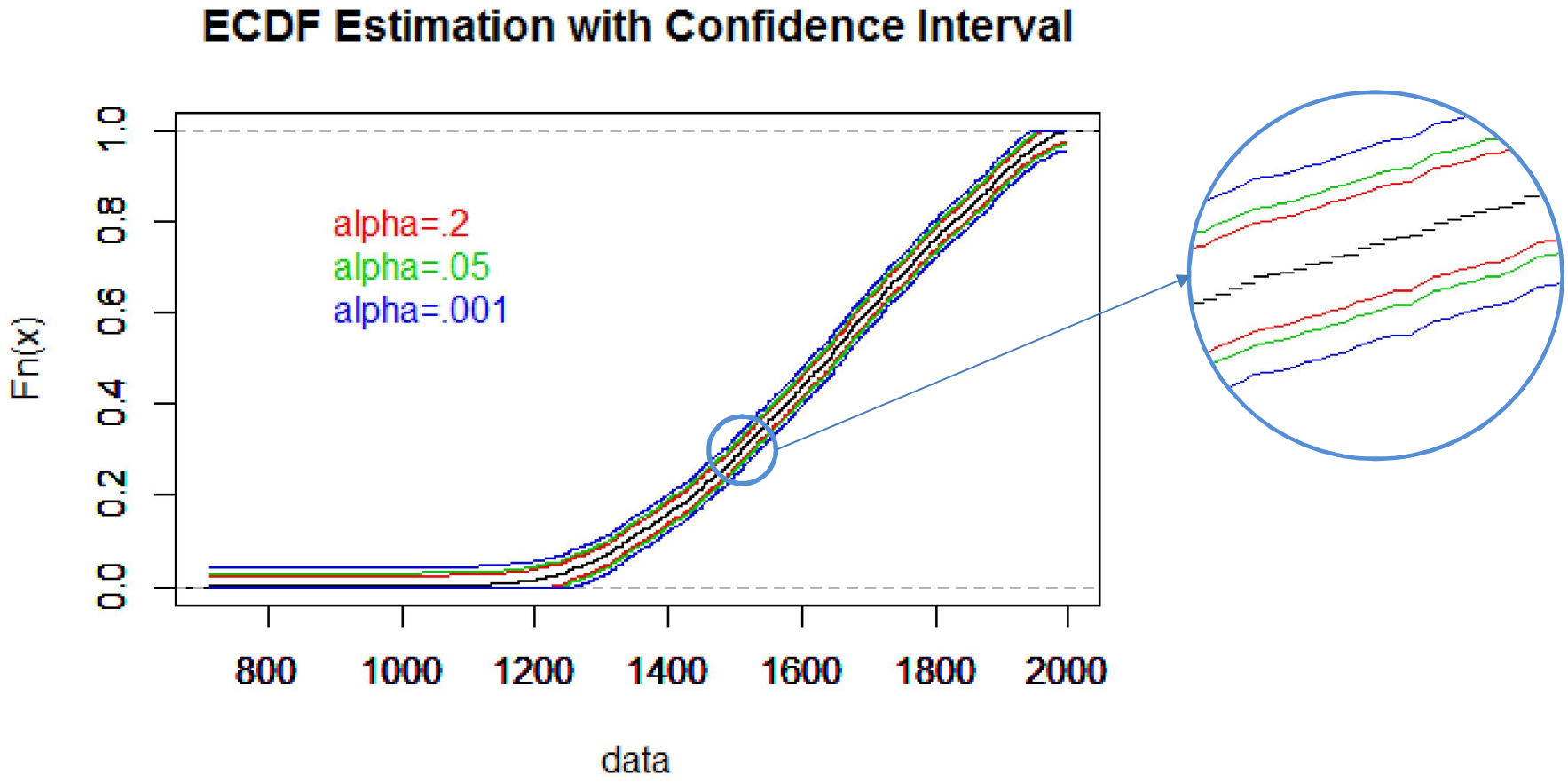
2. (Dvoretzky–Kiefer–Wolfowitz (DKW) inequality). *For any* $\epsilon > 0$,

$$\mathbb{P}\left(\sup_x |F(x) - \widehat{F}_n(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

$$L(x) = \max\{\widehat{F}_n(x) - \epsilon_n, 0\}$$
$$U(x) = \min\{\widehat{F}_n(x) + \epsilon_n, 1\}$$
$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}.$$

$$\mathbb{P}\left(L(x) \leq F(x) \leq U(x) \quad \text{for all } x\right) \geq 1 - \alpha.$$

Running : **ECDF.R**



Definition

A statistical functional $T(F)$ is any function of F. Examples are the mean $\mu = \int x\, dF(x)$, the variance $\sigma^2 = \int (x-\mu)^2 dF(x)$ and the median $m = F^{-1}(1/2)$.

The **plug-in estimator** *of* $\theta = T(F)$ *is defined by* $\widehat{\theta}_n = T(\widehat{F}_n)$.

The bootstrap is a method for estimating the variance and the distribution of a statistic $T_n = g(X_1, \ldots, X_n)$. We can also use the bootstrap to construct confidence intervals.

Bootstrap Variance Estimation

1. Draw $X_1^*, \ldots, X_n^*$ with replacement from $X_1, \ldots, X_n$
2. Compute $T_n^* = g(X_1^*, \ldots, X_n^*)$.
3. Repeat steps 1 and 2, B times to get $T_{n,1}^*, \ldots, T_{n,B}^*$.
4. Let

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*\right)^2$$

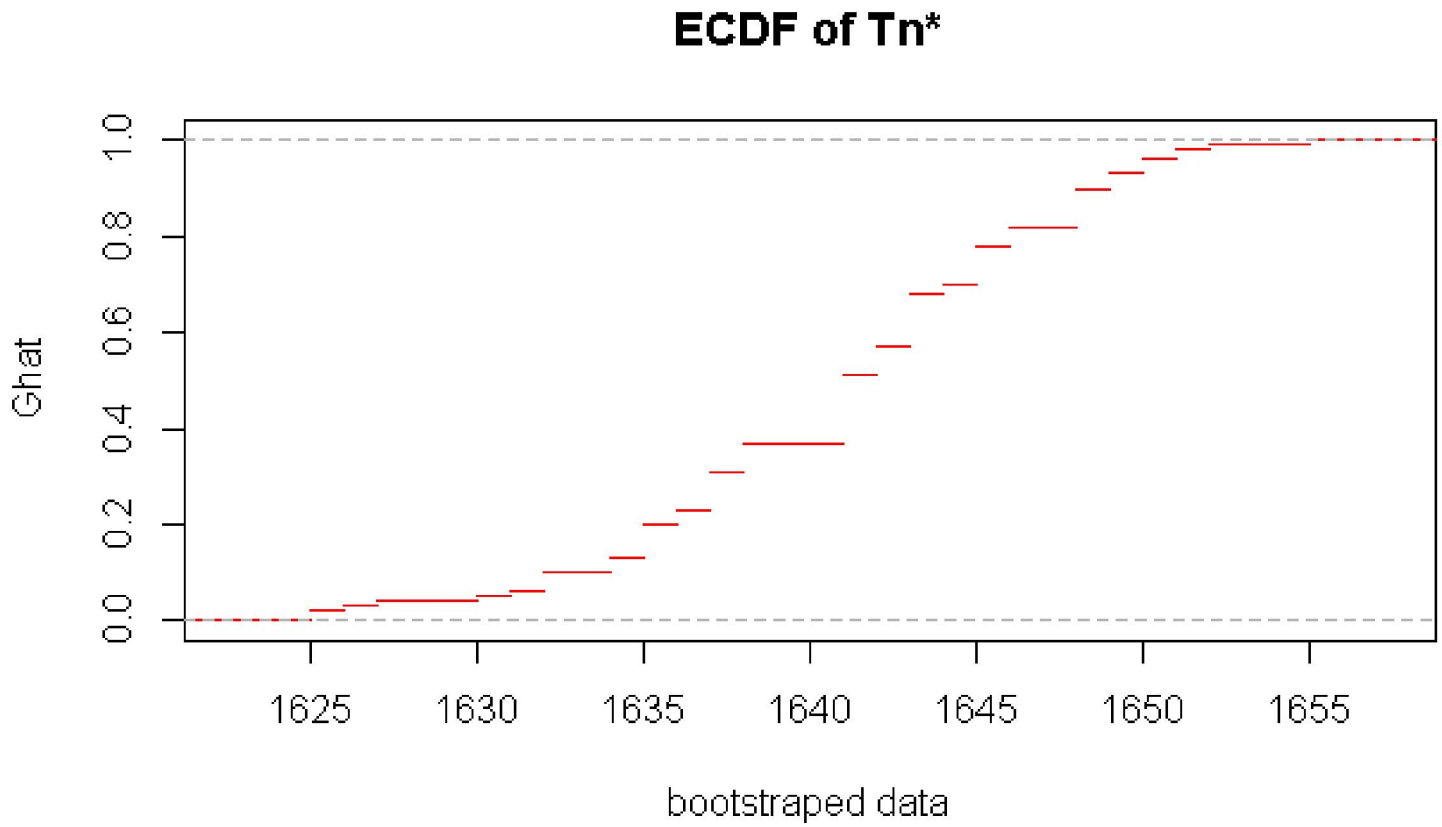
By the law of large numbers, $v_{\text{boot}} \xrightarrow{\text{a.s.}} \mathbb{V}_{\widehat{F}_n}(T_n)$ as $B \to \infty$

Bootstrap approximate CDF

The bootstrap can be used to approximate the CDF of a statistic T_n. Let $G_n(t) = \mathbb{P}(T_n \leq t)$ be the CDF of T_n. The bootstrap approximation to G_n is

$$\widehat{G}_n^*(t) = \frac{1}{B} \sum_{b=1}^{B} I\left(T_{n,b}^* \leq t\right)$$

Running : **BOOT.R**



Bootstrap Confidence Intervals

Normal Interval. The simplest is the Normal interval

$$T_n \pm z_{\alpha/2}\ \widehat{\text{se}}_{\text{boot}}$$

Pivotal Intervals. Let $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$ and define the **pivot** $R_n = \widehat{\theta}_n - \theta$.

The $1-\alpha$ **bootstrap pivotal confidence** interval is

$$C_n = \left(2\widehat{\theta}_n - \widehat{\theta}^*_{((1-\alpha/2)B)},\ 2\widehat{\theta}_n - \widehat{\theta}^*_{((\alpha/2)B)}\right)$$

θ^*_β denote the β sample quantile of $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$

Bootstrap Confidence Intervals

Studentized Pivotal Interval. There is a different version of the pivotal interval which has some advantages

The $1-\alpha$ **bootstrap studentized pivotal interval** is

$$\left(T_n - z^*_{1-\alpha/2}\,\widehat{\mathsf{se}}_{\text{boot}},\ T_n - z^*_{\alpha/2}\,\widehat{\mathsf{se}}_{\text{boot}}\right)$$

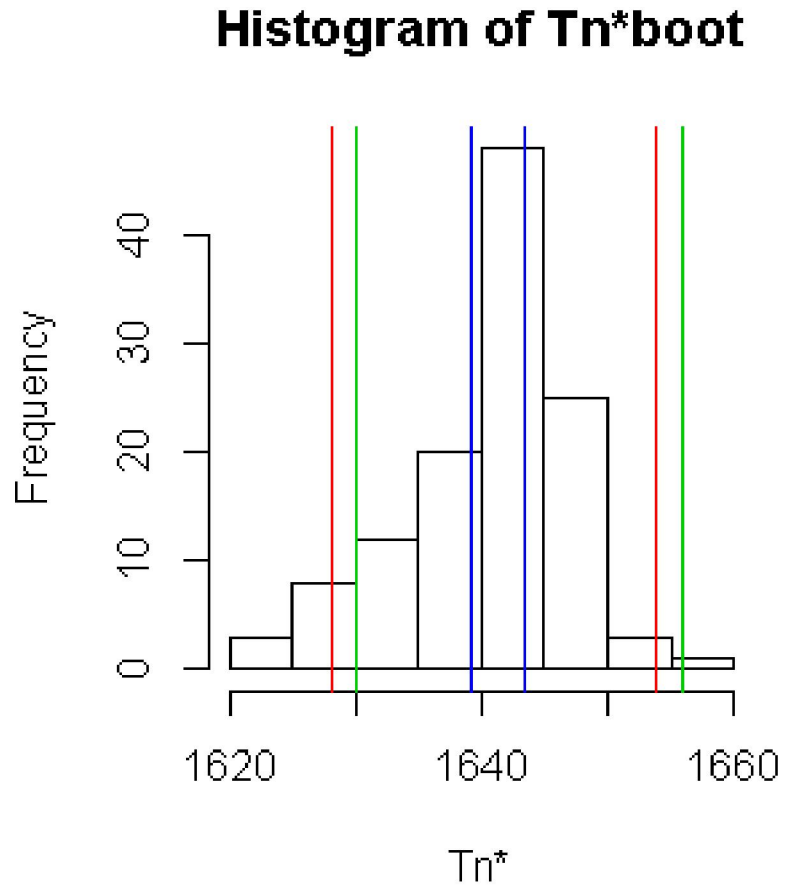
where z^*_β is the β quantile of $Z^*_{n,1},\ldots,Z^*_{n,B}$ and

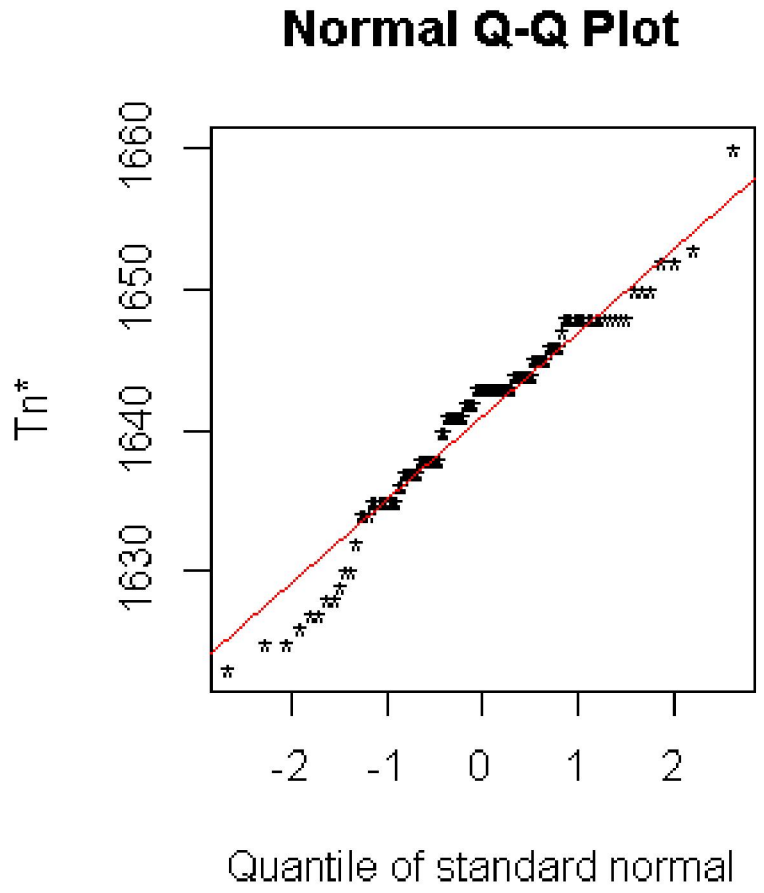
$$Z^*_{n,b} = \frac{T^*_{n,b} - T_n}{\widehat{\mathsf{se}}^*_b}.$$

where $\widehat{\mathsf{se}}^*_b$ is an estimate of the standard error of $T^*_{n,b}$ **not** T_n

to compute $\widehat{\mathsf{se}}^*_b$ for each bootstrap sample you need a second bootstrap within each bootstrap

Running : **BOOT.R**





بوت استرپ ناپارامتری همواره به خوبی کار می کند؟

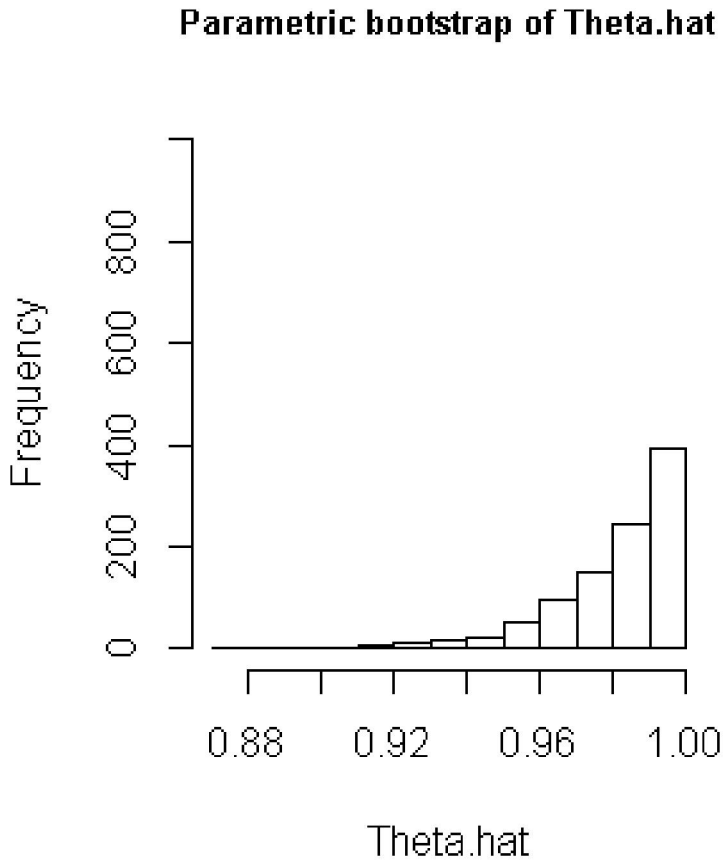
مثال نقض: برای مثال فرض کنیم $X_1, \dots, X_n \sim U(0,\theta)$ باشد و برآورد زیر برای پارامتر مد نظر باشد: $\hat{\theta} = \max(X_1, \dots, X_n)$

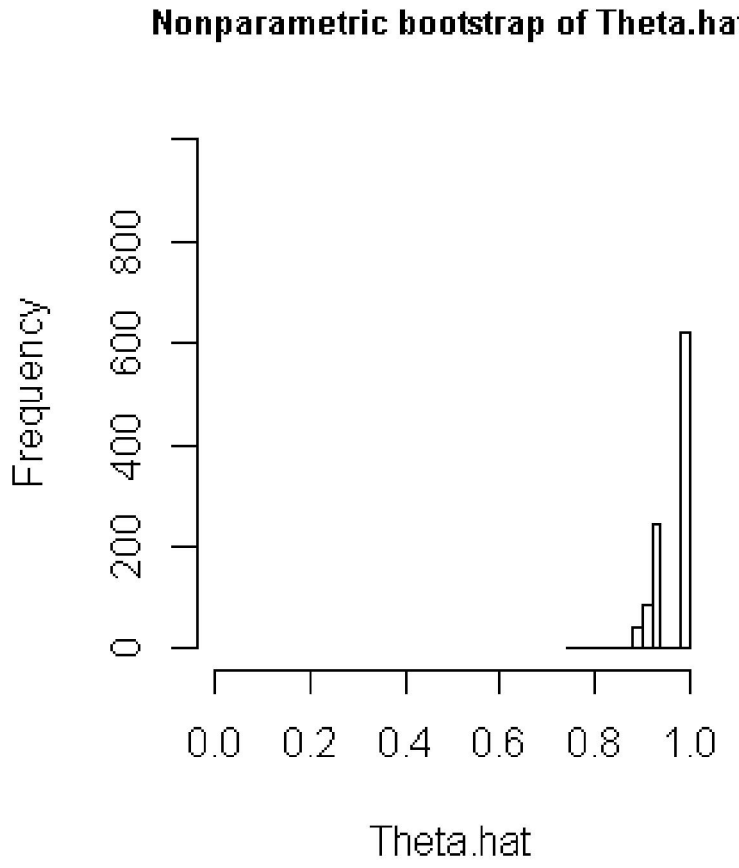
می دانیم $\hat{\theta} \sim \text{beta}(n, 1)$

توزیع واقعی $\hat{\theta}$ را با استفاده از هیستوگرام های بوت استرپ ناپارامتری و پارامتری با هم مقایسه می کنیم. برای انجام این کار:

- یک نمونه B=1000 تایی از توزیع بتا با n=50 تولید کرده و هیستوگرام آنرا رسم می کنیم.
- یک نمونه 50 تایی از توزیع یکنواخت در بازه (۱و۰) اختیار می کنیم و B=1000 بار نمونه 50 تایی با جایگذاری از این نمونه اولیه استخراج کرده و ماکسیمم آنها را در برداری ذخیره کرده و هیستوگرام آنرا رسم می کنیم.

Running : **uniform(0,theta).R**





هموارسازی

❖ به منظور برآورد یک منحنی باید داده ها را هموار کرد. در این حوزه دو مسئله کلی مطرح است:

• برآورد چگالی: اگر $X_1, \ldots, X_n \sim f$ به دنبال برآورد چگالی f هستیم.

• رگرسیون: اگر جفت مشاهدات $(X_1, Y_1), \ldots, (X_n, Y_n)$ را داشته باشیم که

$$Y_i = r(X_i) + \varepsilon_i \quad \text{و} \quad E(\varepsilon_i) = 0$$

به دنبال برآورد تابع رگرسیون r هستیم.

روشهای برآورد تابع چگالی

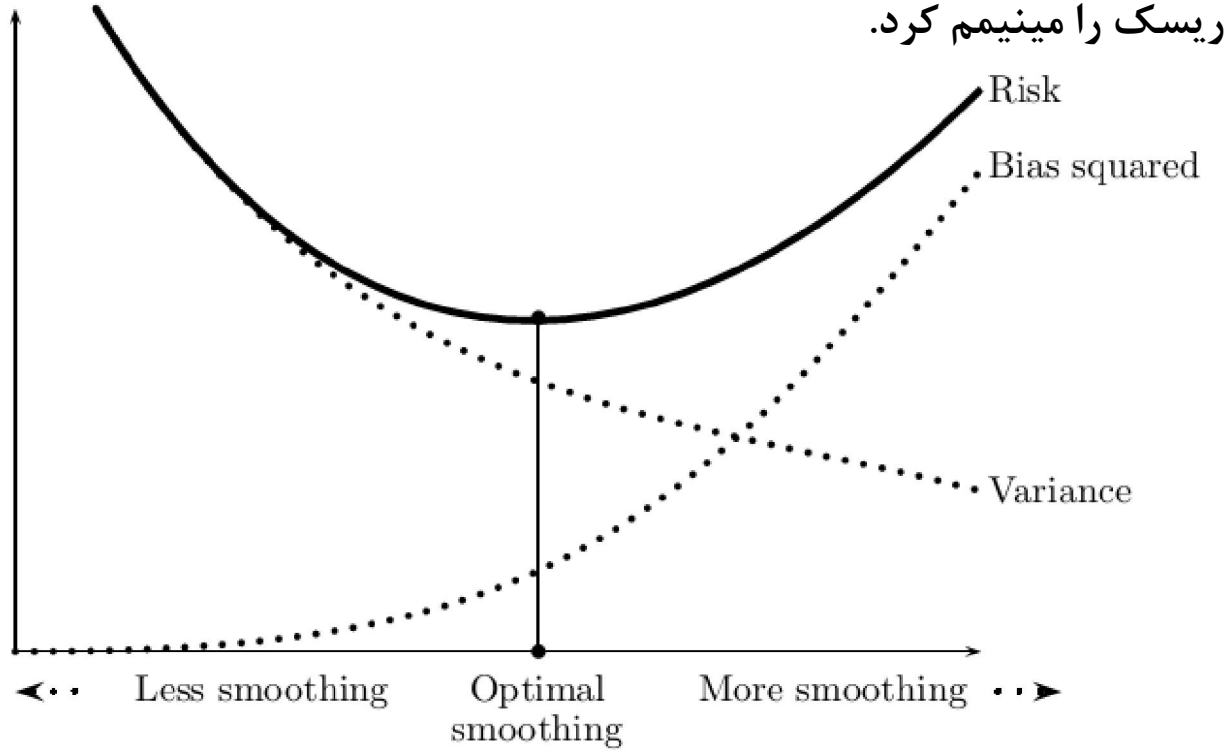
❖ در حالت کلی روشهای برآورد تابع چگالی عبارتند از:

- روش هیستوگرام: (ساده ترین روش است)
- روش تابع هسته

❖ ملاک انتخاب پارامتر هموار سازی در هر دو روش استفاده از اعتبار سنجی متقابل (**Cross Validation**) است که با این انتخاب تابع ریسک مینیمم می شود.

مصالحه اریبی واریانس (Bias-Variance Tradeoff)

- چالش اصلی در هموار سازی تعیین میزان هموارسازی است. وقتی داده ها بیش هموار شوند اریبی بالا و واریانس پایین است. و وقتی داده ها کم هموار شوند برعکس. با تعدیل اریبی و واریانس می توان ریسک را مینیمم کرد.



توابع هموارساز هسته (کرنل)

❖ تابع هموار ساز هسته (کرنل) به تابعی گفته می شود که شرایط زیر را داشته باشد:

$$K(x) \geq 0 \qquad \int K(x)\,dx = 1, \qquad \int xK(x)dx = 0 \quad \text{and} \quad \sigma_K^2 \equiv \int x^2 K(x)dx > 0.$$

❖ معروفترین توابع هسته مورد استفاده عبارتند از:

$$\text{Cosine}: \; K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)\mathbf{1}_{\{|u|\leq 1\}}$$

$$\text{Gaussian}: \; K(u) = \frac{\mathbf{1}}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}$$

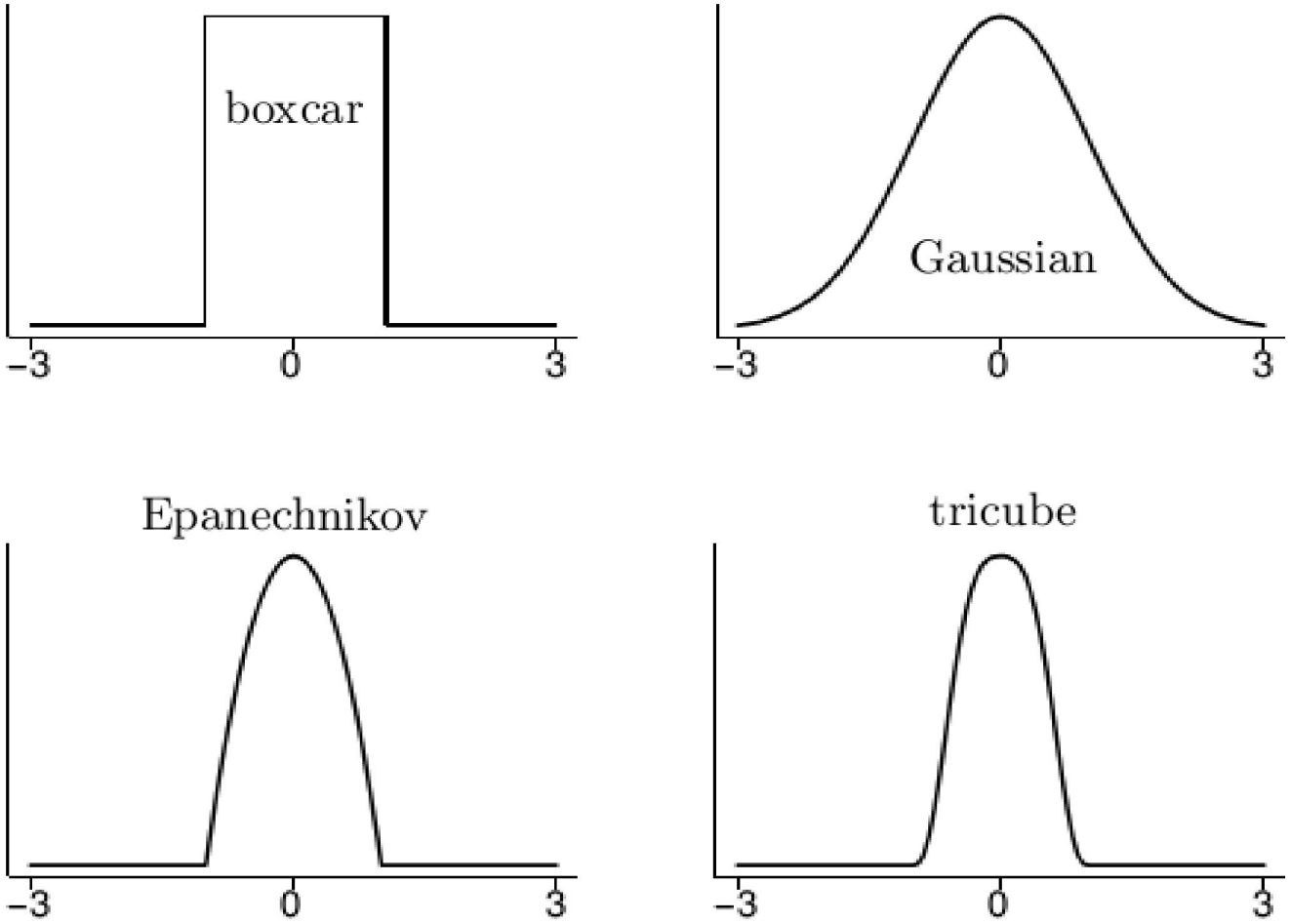
$$\text{Epanechnikov}: \; K(u) = \frac{\mathbf{3}}{\mathbf{4}}(1-u^2)\,\mathbf{1}_{\{|u|\leq 1\}}$$

$$\text{Triangular}: \; K(u) = (1-|u|)\,\mathbf{1}_{\{|u|\leq 1\}}$$

$$\text{Tricube}: \; K(u) = \frac{\mathbf{70}}{\mathbf{81}}(1-|u|^3)^3\,\mathbf{1}_{\{|u|\leq 1\}}$$

$$\text{Rectangular}: \; K(z) = \begin{cases} 1 & \text{for } -\frac{1}{2} < z < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

شکل توابع هموارساز هسته (کرنل)



برآورد تابع چگالی به روش هیستوگرام

a) تکیه گاه f را به بازه (۱و۰) منتقل می کنیم.

b) عدد صحیح m را در نظر گرفته و بازه (۱و۰) را به ئ زیر بازه با پهنای مساوی تقسیم می کنیم:

$$B_1=[0,1/m)\ ,\ B_2=[1/m,2/m)\ ,\ \cdots\ ,\ B_m=[(m-1)/m,1]\ ,$$

c) پهنای باند h را به صورت $h=1/m$ تعریف می کنیم. Y_j را تعداد مشاهدات در بازه B_j در نظر می گیریم و $\hat{p}_j = \frac{Y_j}{n}$

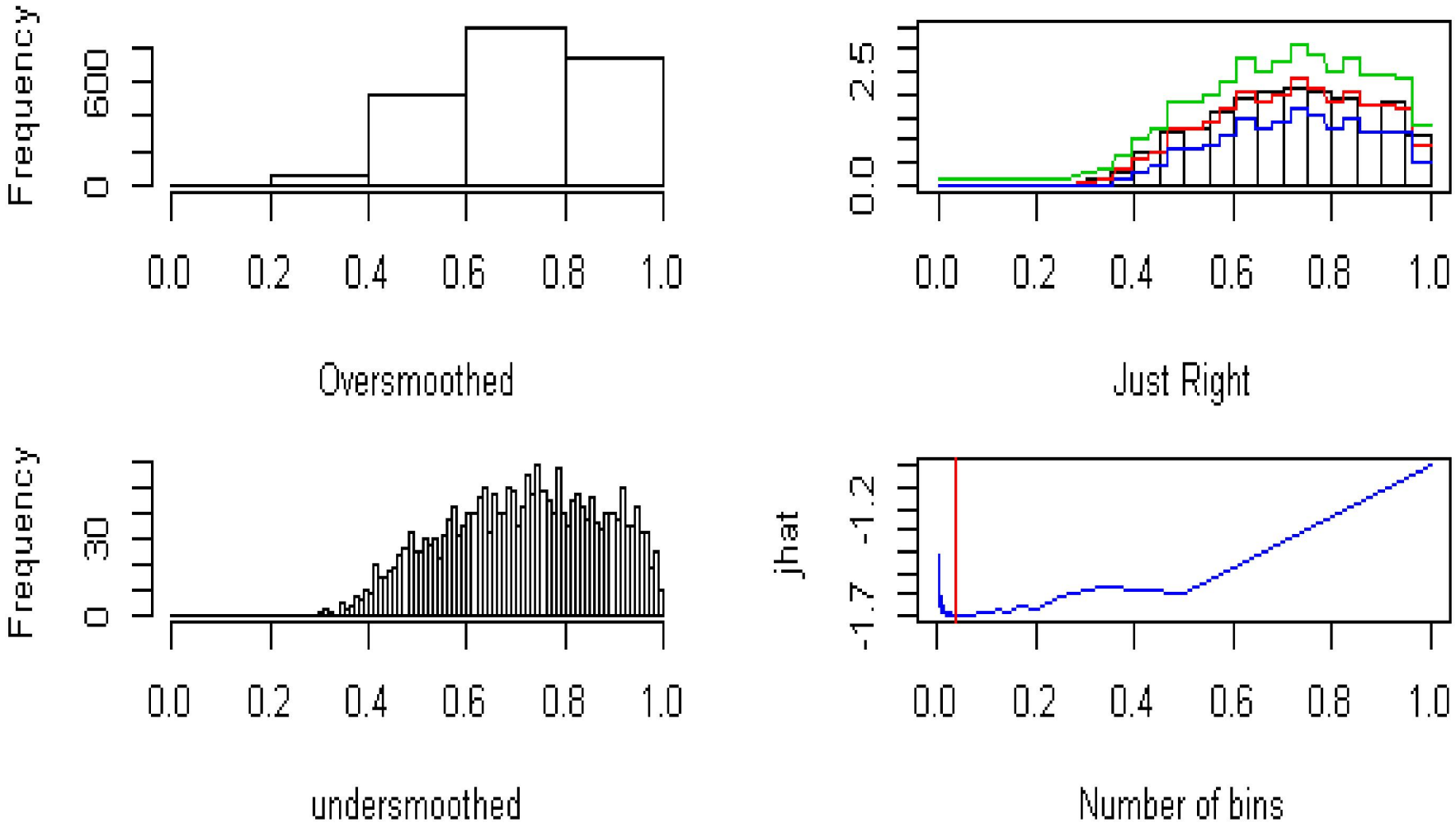
d) برآوردگر هیستوگرام را به صورت زیر به دست می آوریم:

$$\hat{f}_n(x) = \sum_{j=1}^{m} \frac{\hat{p}_j}{h} I(x \in B_j)$$

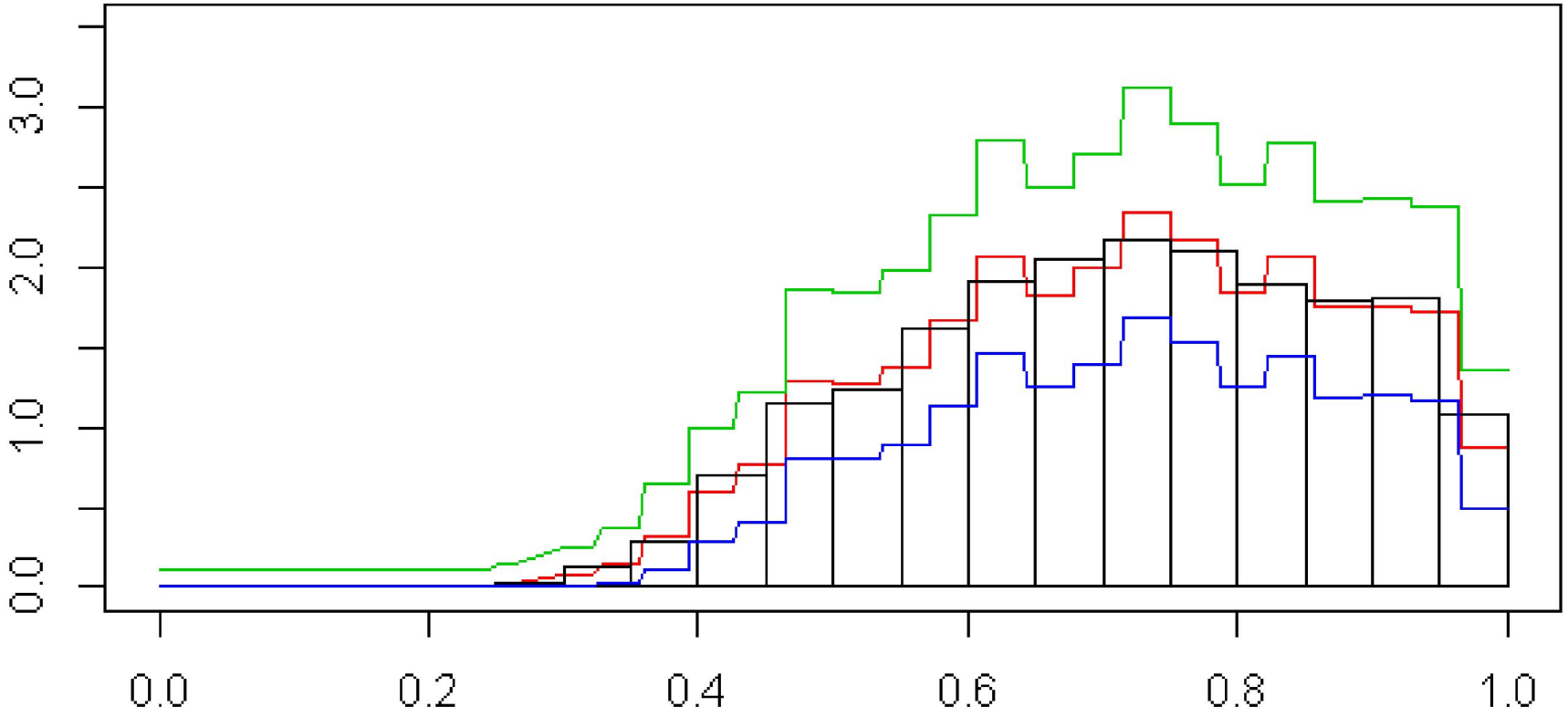
e) پهنای باند h بهینه با مینیم کردن برآوردگر اعتبارسنجی متقابل زیر به دست می آید:

$$\hat{J}(h) = \frac{2}{h(n-1)} - \frac{n+1}{h(n-1)} \sum_{j=1}^{m} \hat{p}_j{}^2$$

Running: **Jhathist.R**



برآورد تابع چگالی به روش هیستوگرام



برآورد تابع چگالی به روش تابع هسته

- روش هیستوگرام هموارکننده خیلی خوبی نیست. در حقیقت روش هسته هم هموار کننده بهتری است و هم همگرایی سریعتری دارد.
- با در نظر گرفتن پهنای باند مثبت h و تابع هسته K، برآوردگر چگالی هسته به صورت زیر تعریف می شود:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)$$

- انتخاب پهنای باند مناسب حیاتی تر از انتخاب نوع تابع هسته است.

a) (قاعده ارجاع نرمال) اگر تابع f خیلی هموار و نزدیک به توزیع نرمال باشد می توان از برآورد زیر برای h استفاده کرد:

$$h_n = \frac{1.06\,\hat{\sigma}}{n^{1/5}} \qquad \hat{\sigma} = \min\left\{s,\ \frac{Q}{1.34}\right\}$$

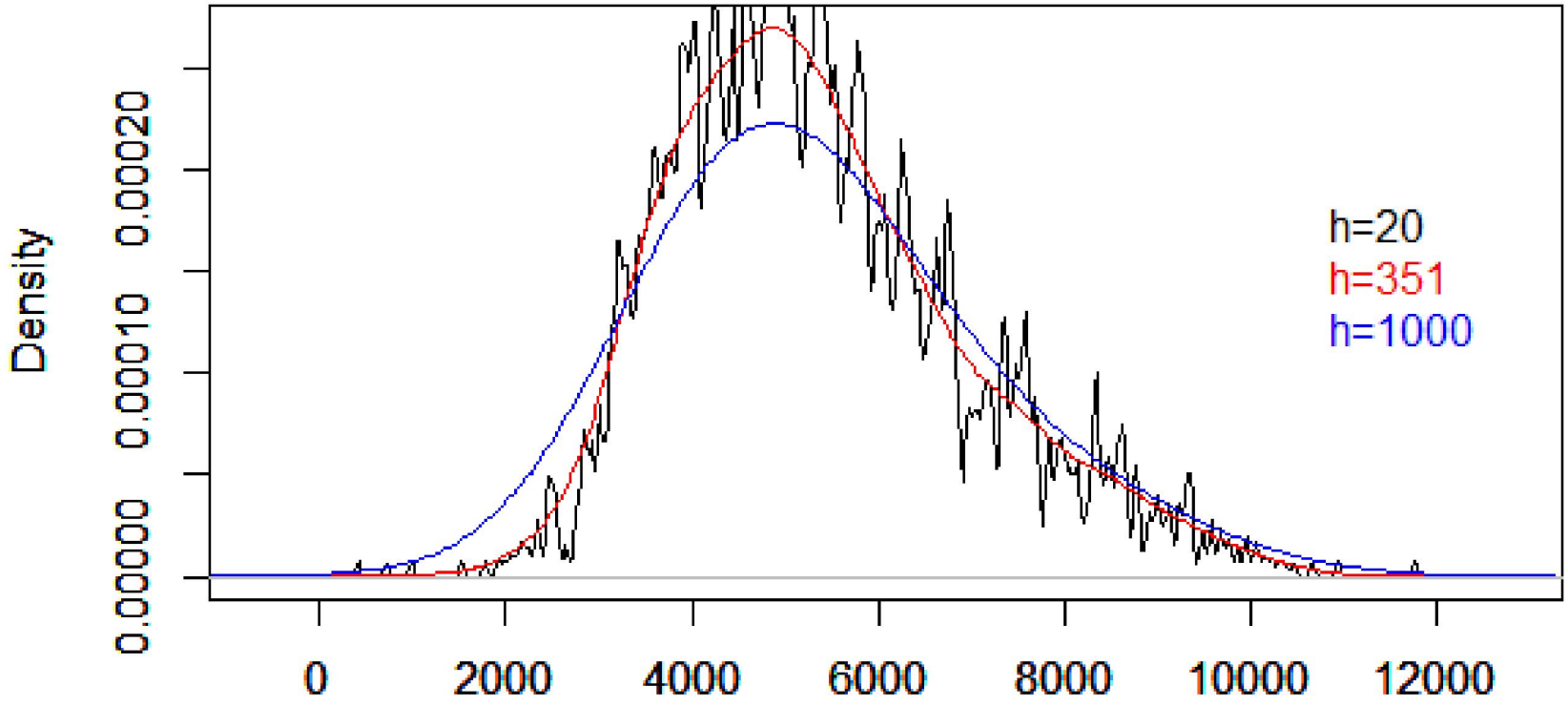
b) در حالت کلی پهنای باند h بهینه با مینیم کردن برآوردگر اعتبارسنجی متقابل زیر به دست می آید:

$$\hat{J}(h) = \frac{1}{hn^2}\sum_i\sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh}K(0) + O\left(\frac{1}{n^2}\right)$$

$$K^*(x) = K^{(2)}(x) - 2K(x) \ \ and \ K^{(2)}(z) = \int K(z-y)K(y)dy$$

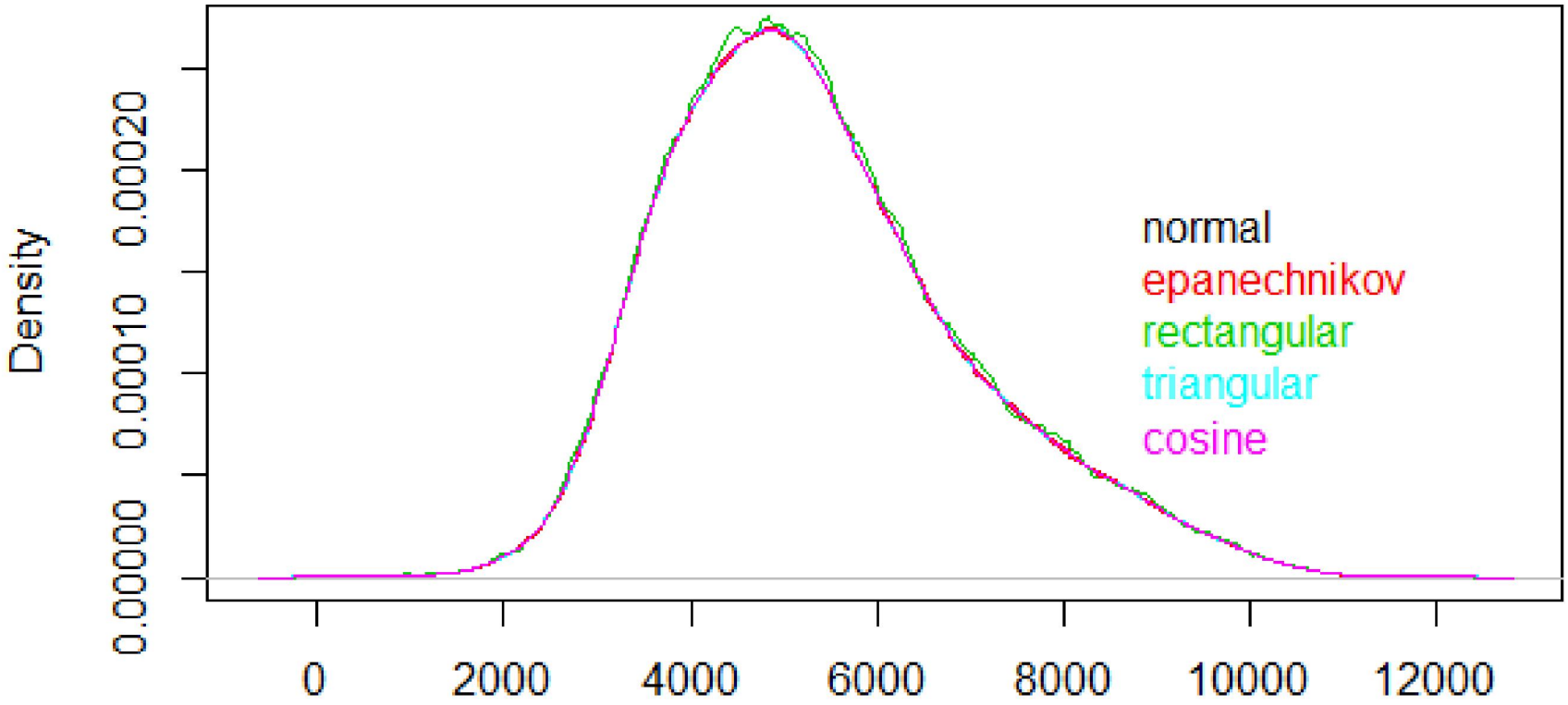
برآورد تابع چگالی به روش تابع هسته با استفاده از قاعده ارجاع نرمال

Running: **NRR.R**



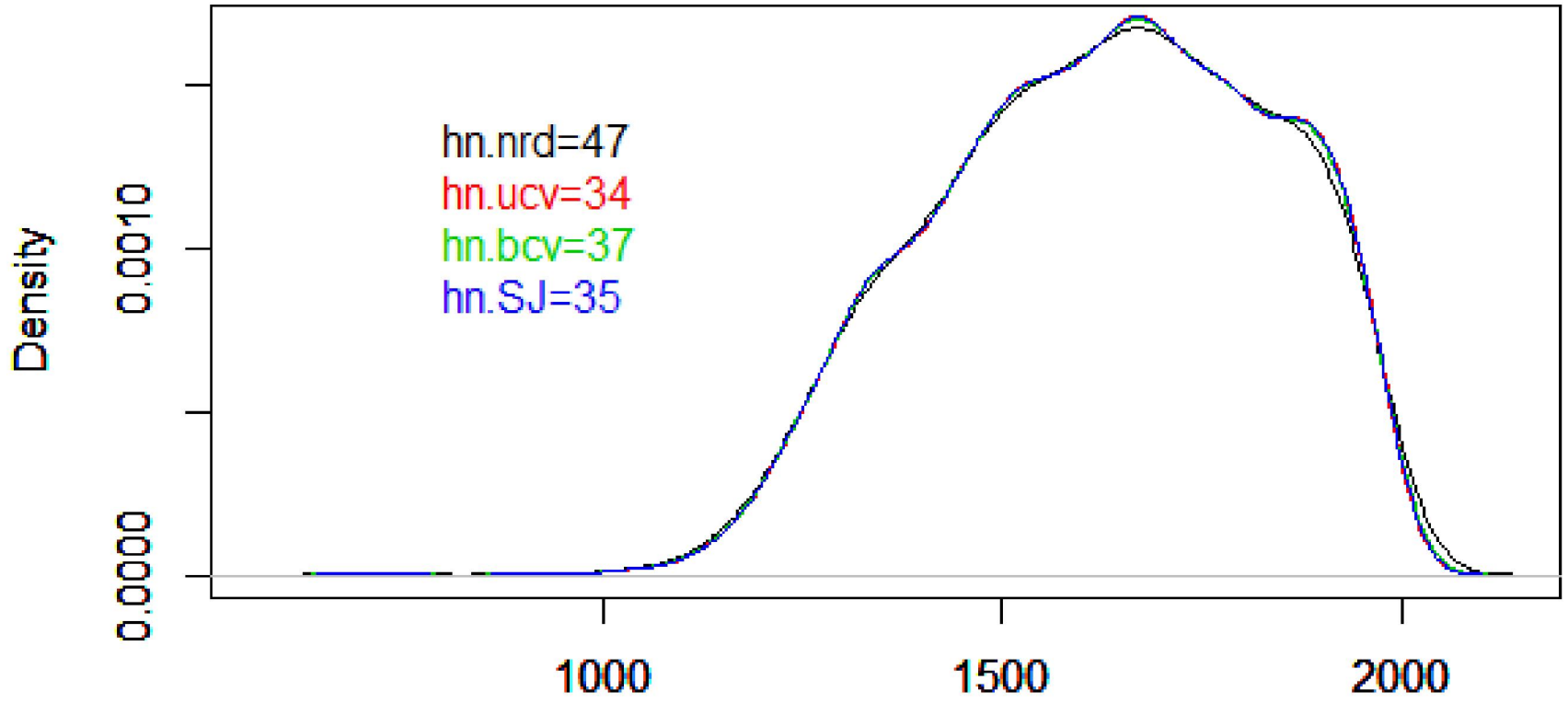
برآورد تابع چگالی به روش تابع هسته با استفاده از قاعده ارجاع نرمال

Running: **NRR.R**



برآورد تابع چگالی به روش تابع هسته با استفاده از قاعده ارجاع نرمال

Running: **denest.R**



رگرسیون ناپارامتری

- فرض می کنیم پهنای باند $h>0$ و برآوردگر هسته نادارایا-واتسون را به صورت زیر تعریف می کنیم:

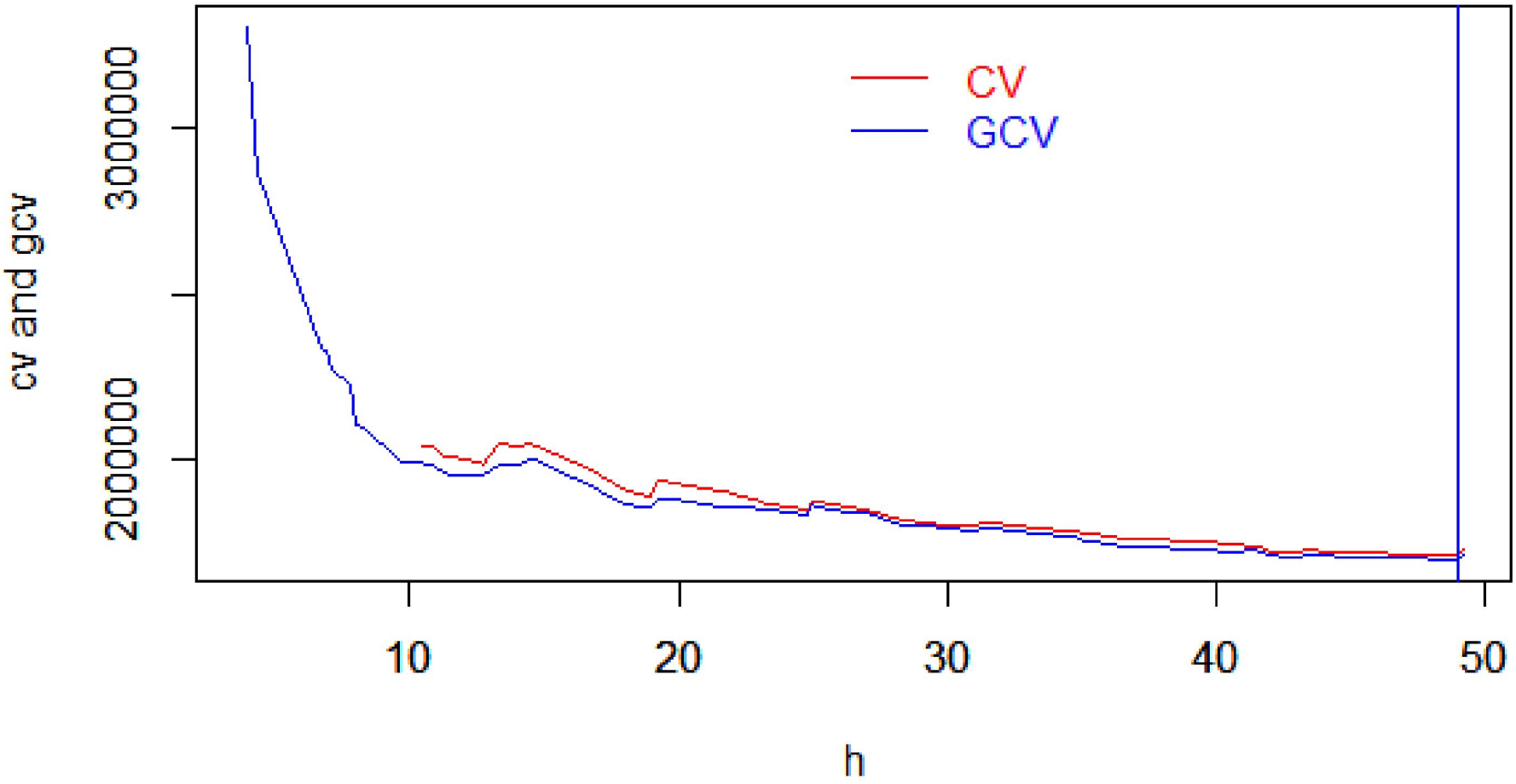
$$\widehat{r}_n(x) = \sum_{i=1}^{n} \ell_i(x) Y_i$$

- که $\ell_i(x)$ وزن هایی هستند که با توجه به تابع هسته مورد نظر به صورت زیر تعیین می شوند:

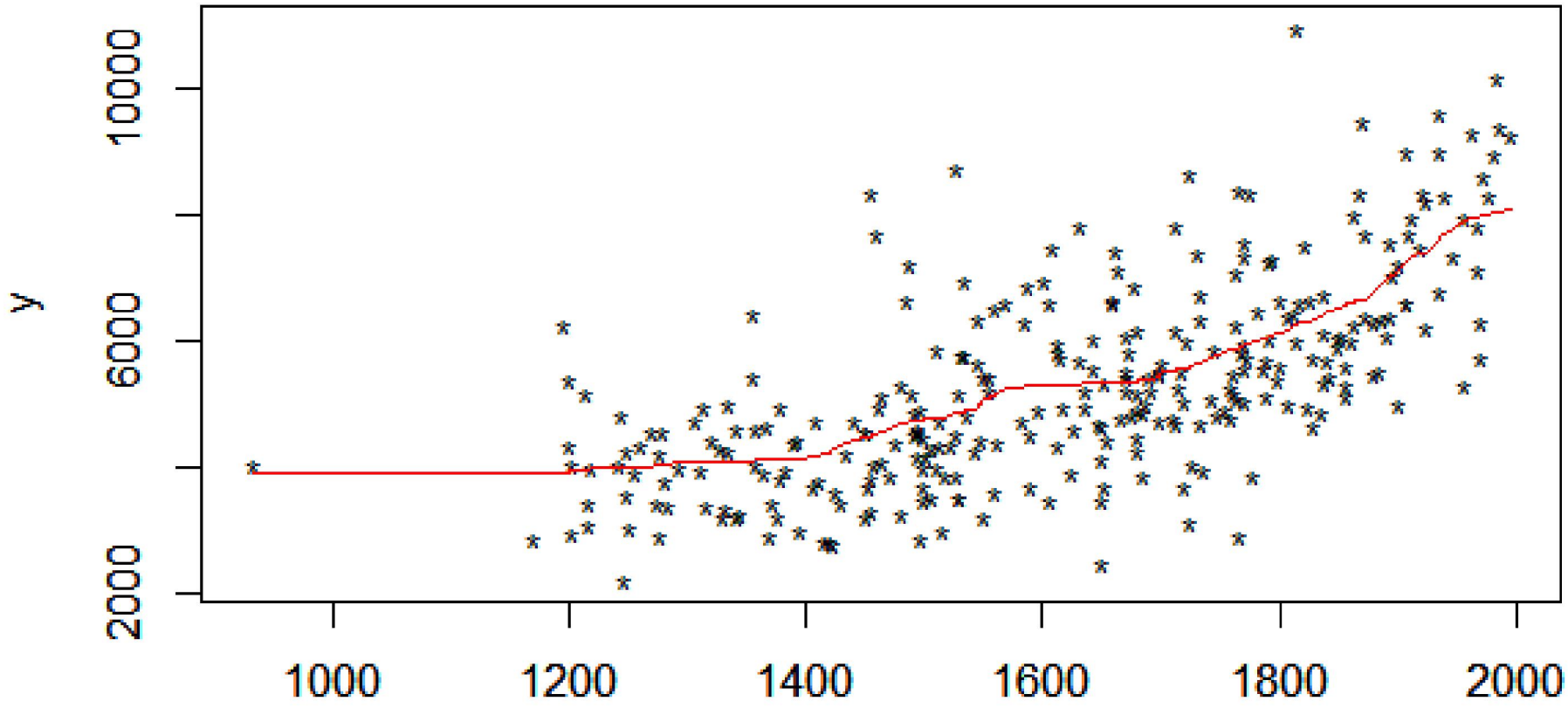
$$\ell_i(x) = \frac{K\left(\frac{x-x_1}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_1}{h}\right)}$$

- در اینجا نیز پهنای باند h بهینه با مینیم کردن برآوردگر اعتبارسنجی متقابل یا برآوردگر اعتبارسنجی متقابل تعمیم یافته به صورت زیر به دست می آید:

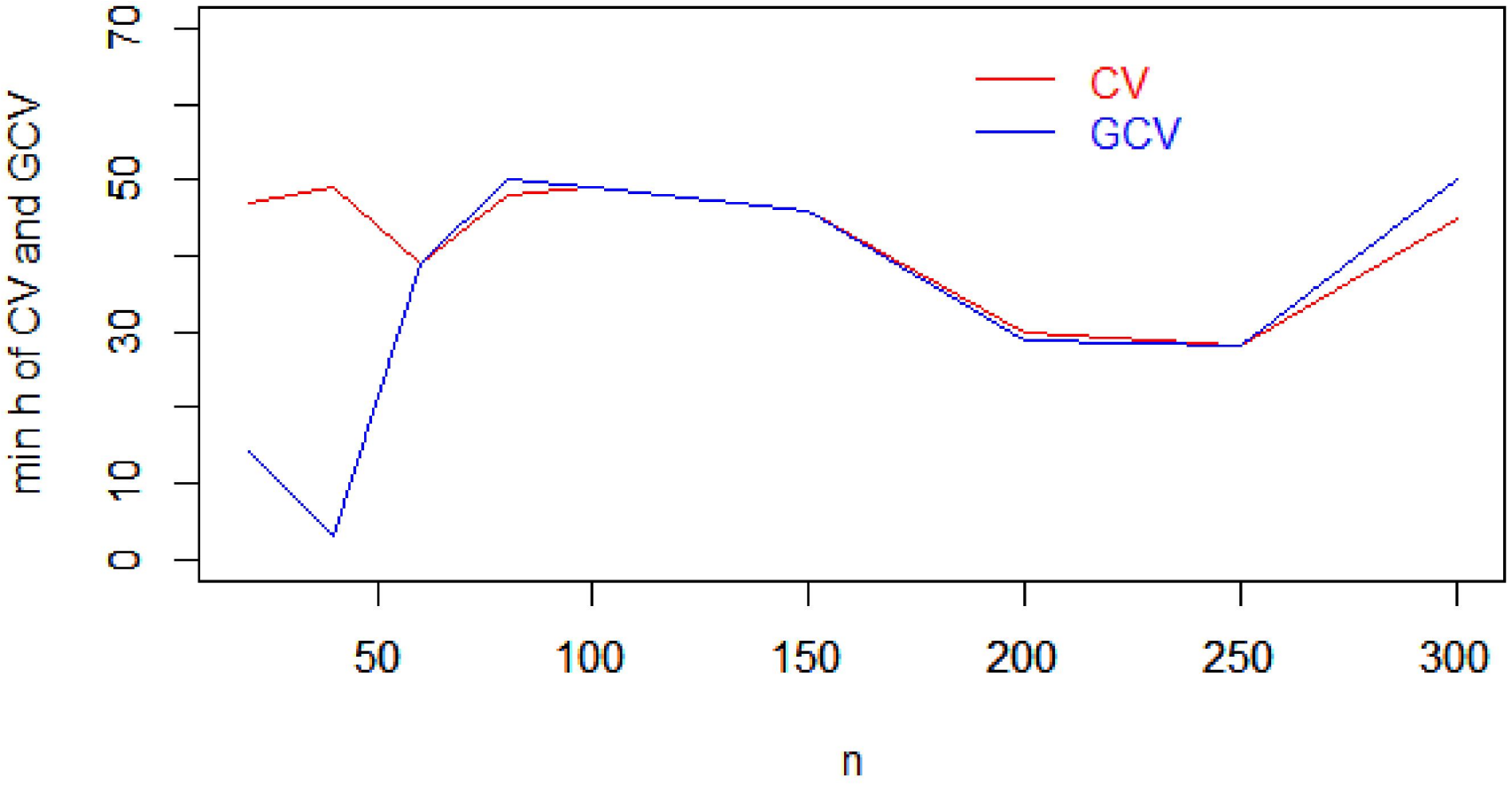
Running: **NWKE**



Running: **NWKE**



Running: **NWKE**



با تشکر از توجه شما