Inference on a Distribution Function from Ranked Set Samples

Lutz Dümbgen (Univ. of Bern)

Ehsan Zamanzade (Univ. of Isfahan)

October 17, 2013

Swiss Statistics Meeting, Basel

I. The Setting

In some situations, ranking observations is much easier than obtaining their precise values ...

Consider $n \cdot k$ independent random variables

$$X_{ij}, \quad 1 \le i \le n,\ 1 \le j \le k,$$

with unknown continuous c.d.f. F.

Instead of complete i-th sample $(X_{ij})_{j=1}^{k}$ observe only

- one of its elements, X_i, and
- its rank $R_i \in \{1, 2, \ldots, k\}$ within $(X_{ij})_{j=1}^{k}$.

Example 1: Ranked set sampling (RSS)
(McIntyre 1952)

For $1 \leq i \leq n$ let

$$X_{i:1} \ < \ X_{i:2} \ < \ \cdots \ < \ X_{i:k}$$

be the order statistics of $(X_{ij})_{j=1}^{k}$.

- Pick a rank $R_i \in \{1, 2, \ldots, k\}$
- Obtain precise value $X_i := X_{i:R_i}$.

Special case: $k = n$ and $R_i = i$.

Example 2: Judgement post-stratification (JPS)
(MacEachern et al. 2004)

- $X_i := X_{i1}$,
- $R_i := \sum_{j=1}^{k} 1_{[x_{ij} \leq x_{i1}]}$.

General setting:

Independent observations (X_1, R_1), (X_2, R_2), $\ldots$, (X_n, R_n).

Conditional on $R_i = r$,

$$X_i \sim F_r(x) := B_r(F(x))$$

with

$$B_r(p) := \sum_{i=r}^{k} \binom{k}{i} p^i (1-p)^{k-i} = \int_0^p \beta_r(u)\, du$$

$$\beta_r(u) := k \binom{k-1}{r-1} u^{r-1} (1-u)^{k-r}$$

(c.d.f. and density of $\mathrm{Beta}(r, k+1-r)$).

Basic fact

$$\frac{1}{k}\sum_{r=1}^{k} \mathrm{Beta}(r, k+1-r) \;=\; \mathrm{Unif}[0,1],$$

whence

$$\frac{1}{k}\sum_{r=1}^{k} F_r \;=\; F.$$

II. Estimation paradigms

Paradigm 0: Naive empirical distribution function

$$\widehat{F}(x) \;:=\; \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \le x]}.$$

In JPS setting: Okay but inefficient.

In unbalanced RSS settings: Potentially severely biased.

Paradigm 1: Stratification (Stokes and Sager 1988)

Groupwise empirical c.d.f.s:

$$\widehat{F}_r(x) \;:=\; \frac{1}{N_r} \sum_{i\,:\,R_i=r} 1_{[X_i \le x]} \quad \text{with} \quad N_r := \#\{i : R_i = r\}.$$

Stratified estimator

$$\widehat{F}^{\mathrm{S}} \;:=\; \frac{1}{k} \sum_{r=1}^{k} \widehat{F}_r.$$

Potential problems with small/zero numbers N_r.

Paradigm 2: Likelihood (Kvam and Samaniego 1994)

$$\widehat{F}^{\mathrm{L}}(x) \;:=\; \underset{p\in[0,1]}{\arg\max}\, L(x,p)$$

with (conditional) log-likelihood function

$$\begin{aligned} L(x,p) \;&:=\; \sum_{i=1}^{n}\left[1_{[X_i\le x]}\log B_{R_i}(p) + 1_{[X_i>x]}\log(1-B_{R_i}(p))\right] \\ &=\; \sum_{r=1}^{k} N_r\left[\widehat{F}_r(x)\log B_r(p) + (1-\widehat{F}_r(x))\log(1-B_r(p))\right] \end{aligned}$$

Paradigm 3: Moments

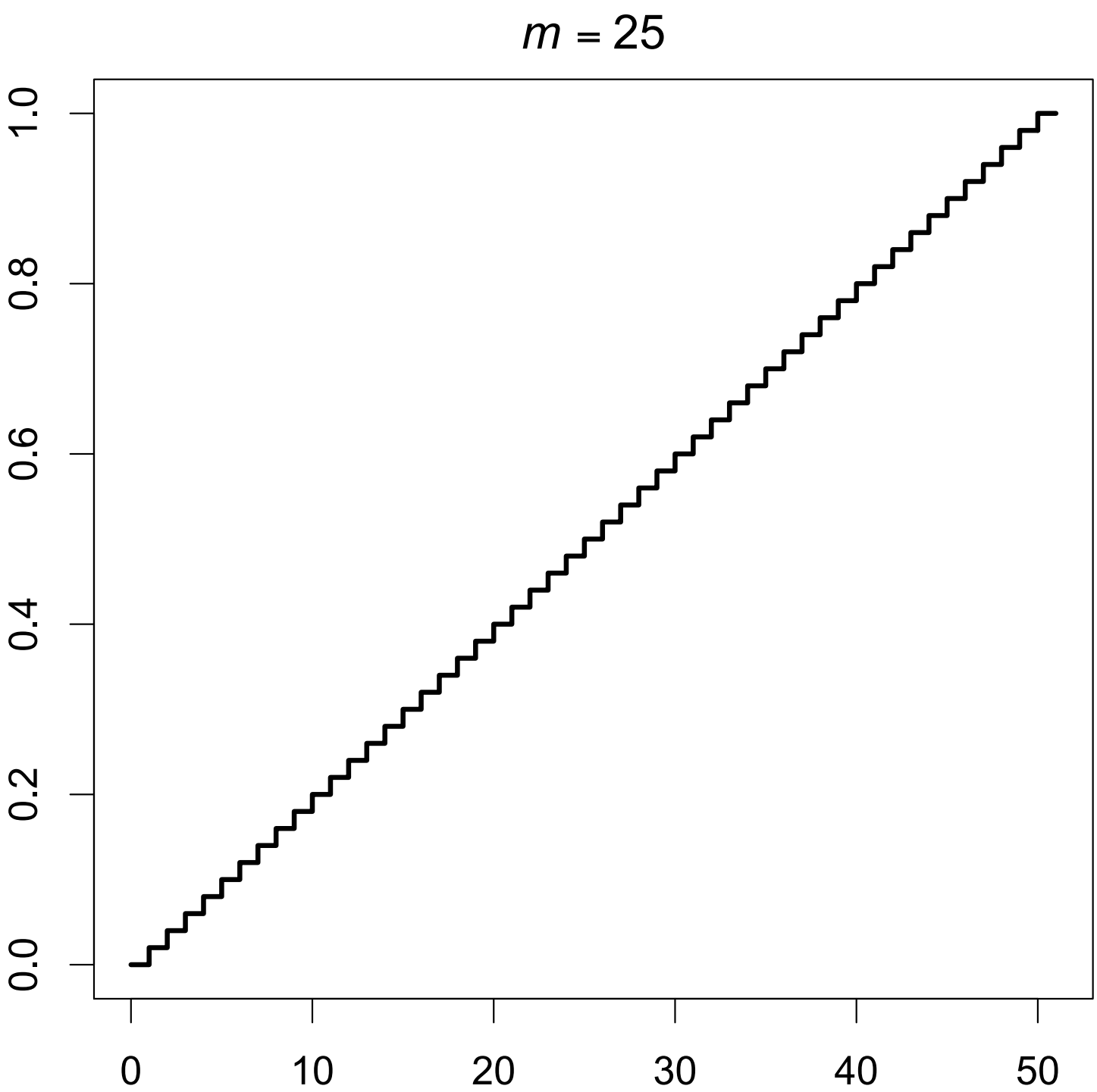
$$\mathbb{E}\big(\widehat{F}(x) \,\big|\, \mathbf{R}\big) \;=\; \sum_{r=1}^{k} \frac{N_r}{n} B_r(F(x)).$$

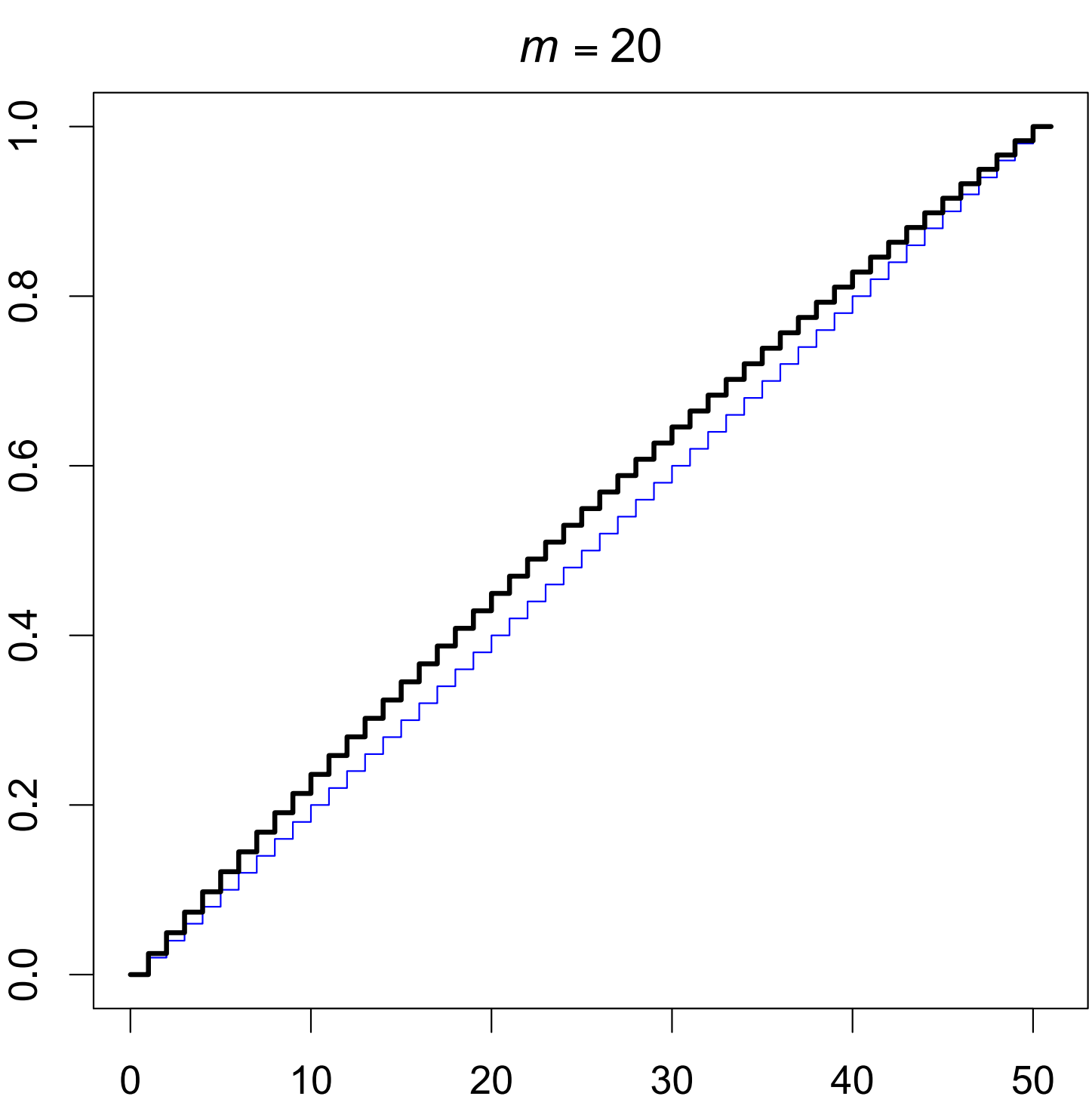
$$\widehat{F}(x) \;\stackrel{!}{=}\; \sum_{r=1}^{k} \frac{N_r}{n} B_r(\widehat{F}^{\mathrm{M}}(x)).$$

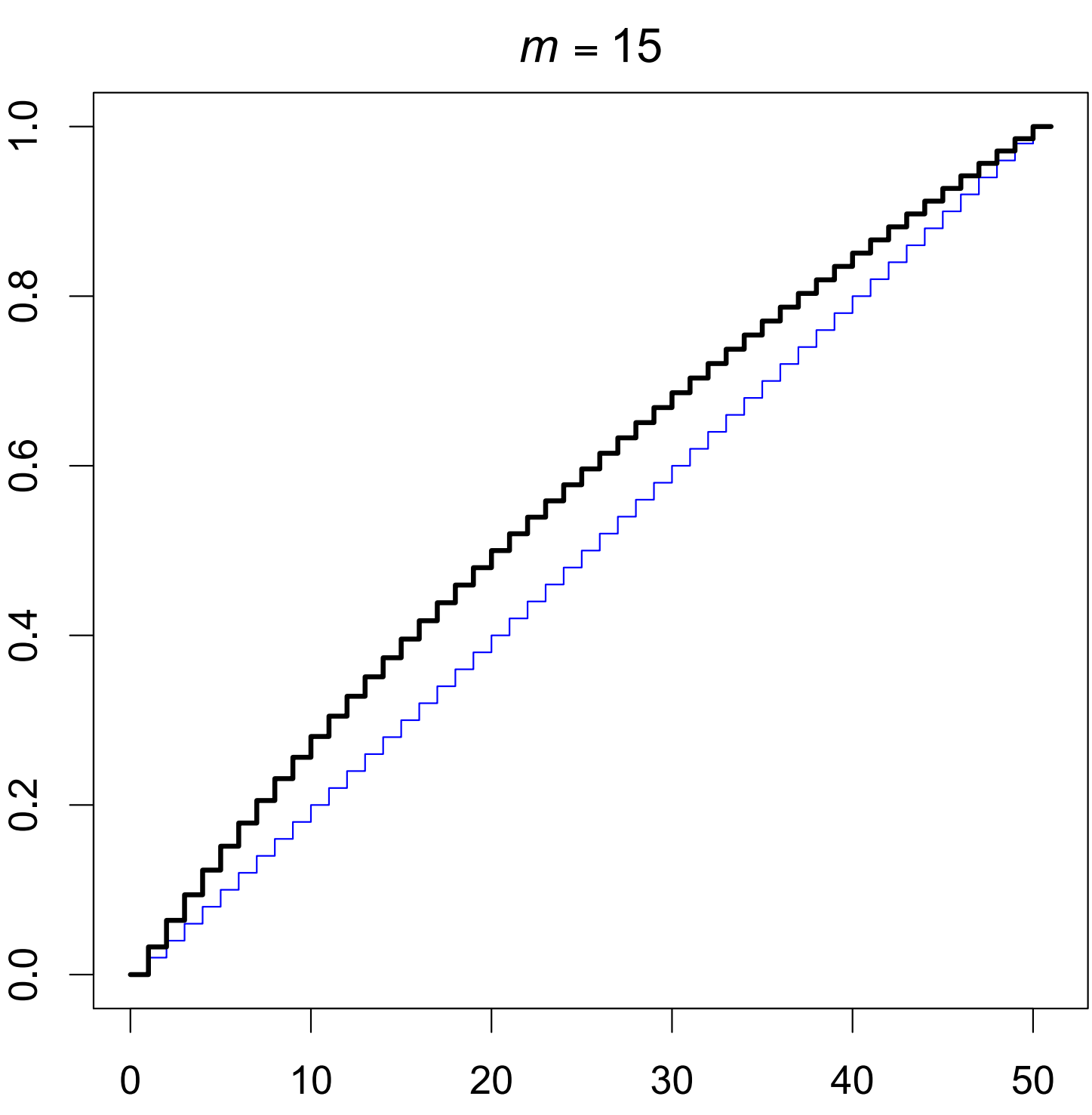
Paradigms 2-3 deal with small/zero numbers N_r properly.

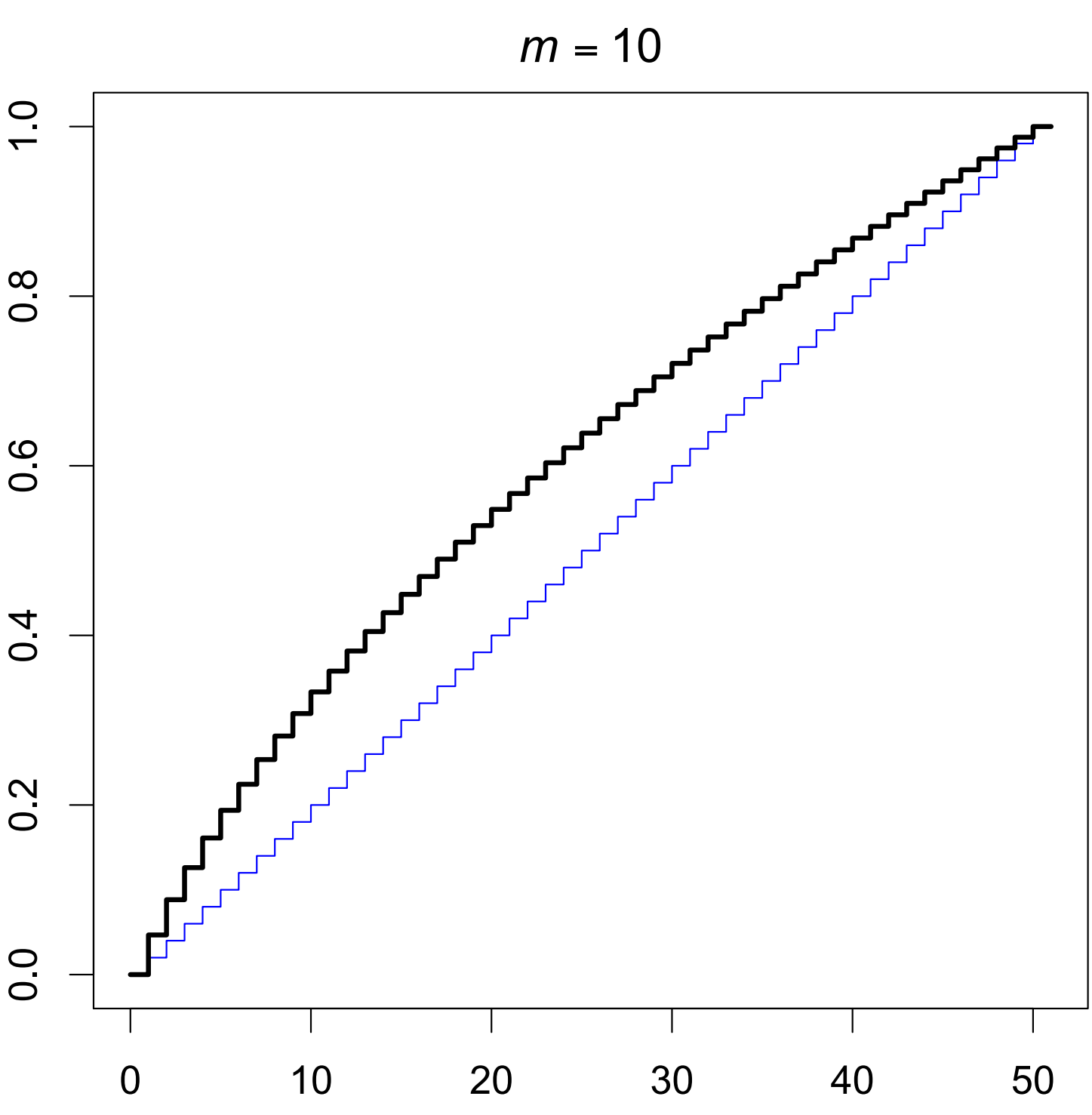
Estimator $\widehat{F}^{\mathrm{M}}$ when

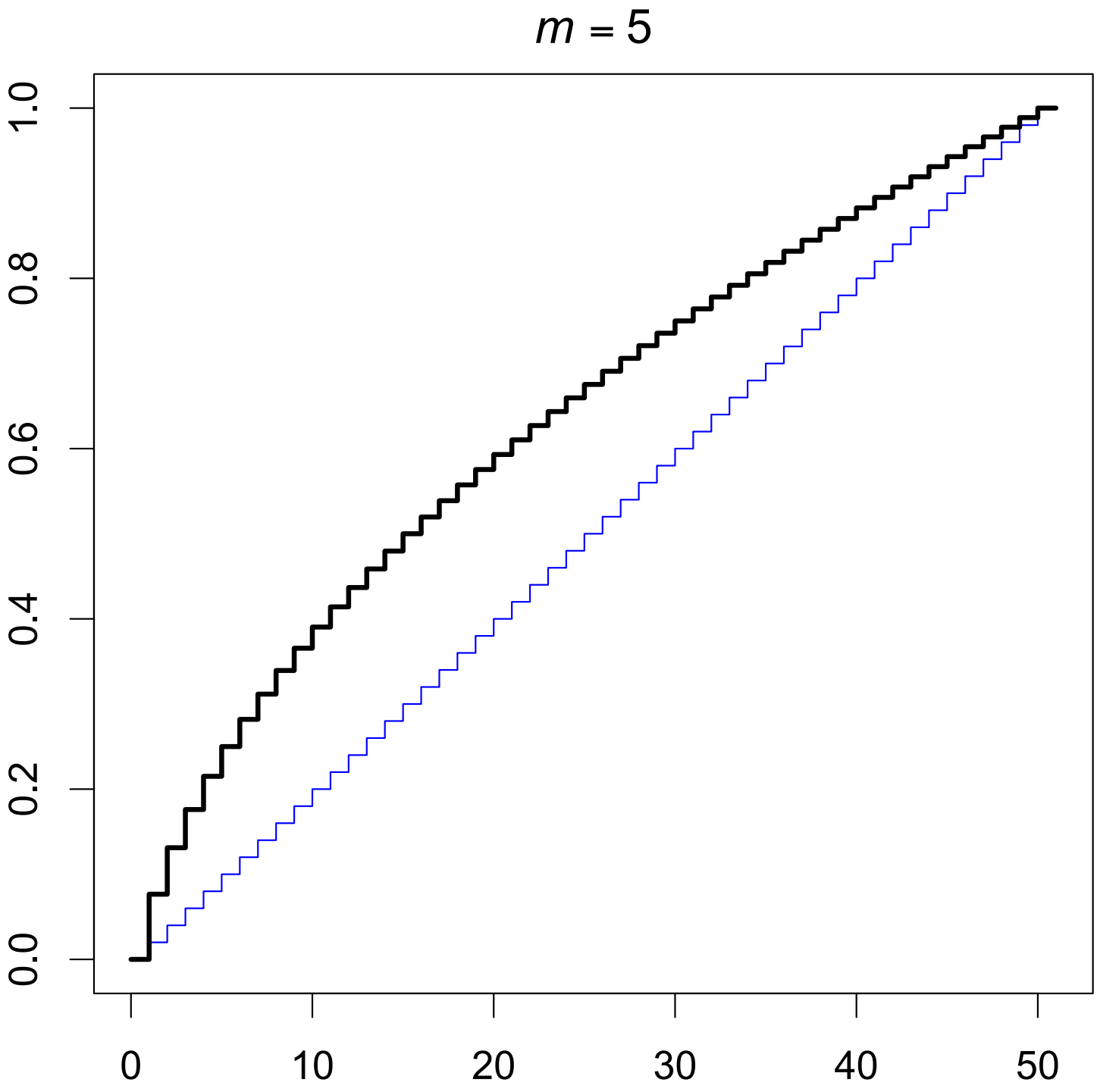
$$k = 2 \quad \text{and} \quad \mathbf{N} = (m, 50 - m)^{\top}$$

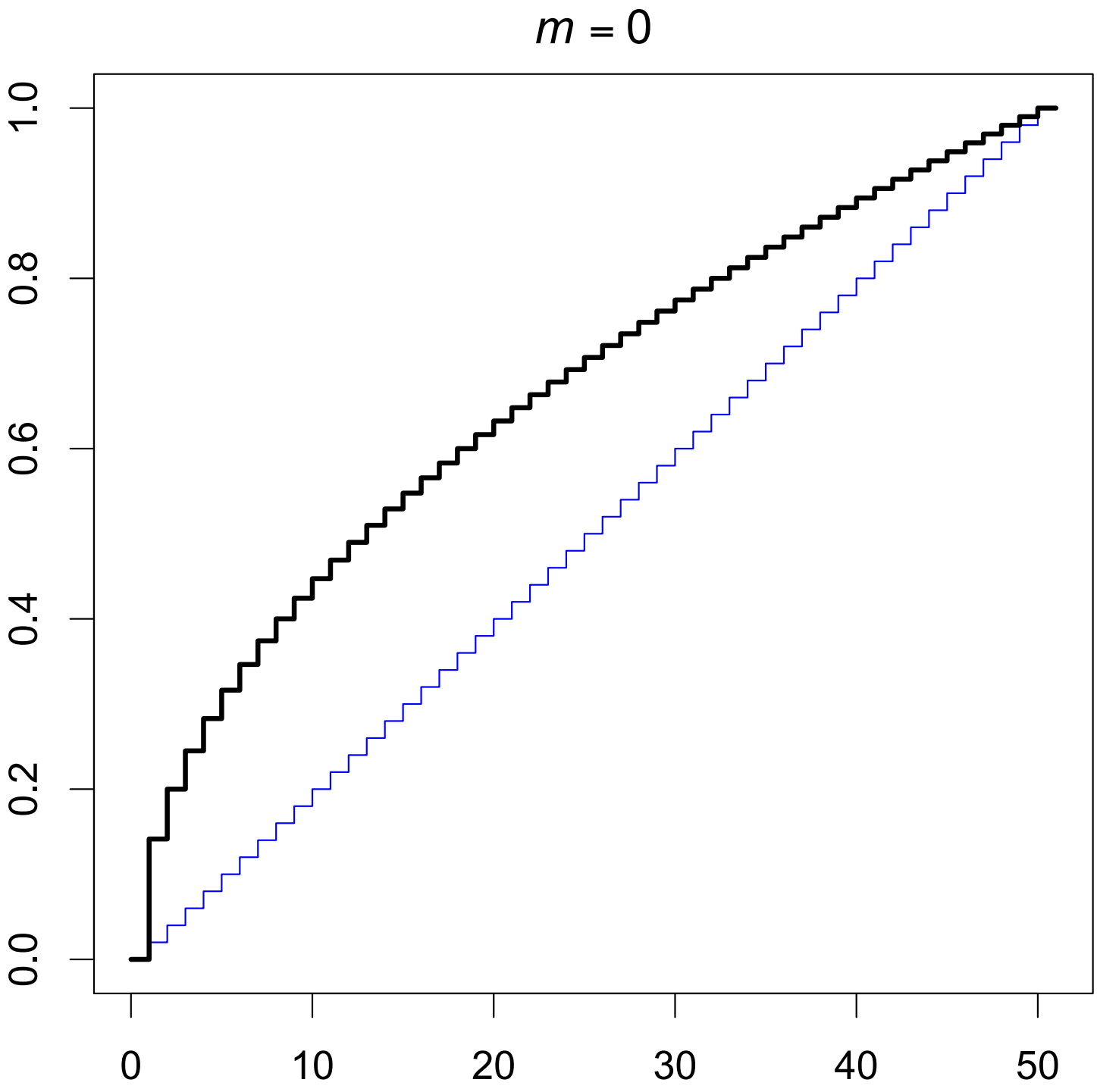

$m = 25$
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


$m = 20$
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


m = 15
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


m = 10
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


$m = 5$
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


$m = 0$
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

III. Asymptotics

Refine and generalize results by
Stokes and Sager (1988) and Huang (1997)

Condition on **R** and assume that

$$\frac{N_r}{n} \;\to\; \pi_r \quad \text{for } 1 \le r \le k$$

where

$$\begin{cases} \pi_1, \pi_2, \ldots, \pi_k > 0 & \text{for } \widehat{F}^{\mathrm{S}}, \\ \pi_1, \pi_k > 0 & \text{for } \widehat{F}^{\mathrm{M}}, \widehat{F}^{\mathrm{L}}. \end{cases}$$

Without loss of generality let

$$F \mathrel{\widehat{=}} \mathrm{Unif}[0, 1].$$

Empirical processes:

$$\mathbb{V}_r(t) \;:=\; \sqrt{N_r}\big(\widehat{F}_r(B_r^{-1}(t)) - t\big), \quad t \in [0,1],$$

(Standard empirical process of sample of size N_r from $\mathrm{Unif}[0,1]$.)

Linear expansion: For $\mathrm{Z} = \mathrm{S}, \mathrm{L}, \mathrm{M}$,

$$\sqrt{n}\,(\widehat{F}^{\mathrm{Z}} - F) \;\approx\; \sum_{r=1}^{k} \gamma_r^{\mathrm{Z}}\; \mathbb{V}_r \circ B_r$$

with certain weight functions γ_1^{Z}, γ_2^{Z}, $\ldots$, γ_k^{Z}.

Weights for $\widehat{F}^{\mathrm{S}}$

$$\begin{aligned}
\sqrt{n}(\widehat{F}^{\mathrm{S}} - F) &= \sum_{r=1}^{k} \frac{\sqrt{n}\,(\widehat{F}_r - B_r)}{k} \\
&= \sum_{r=1}^{k} \frac{\sqrt{n}\,\mathbb{V}_r \circ B_r}{k\sqrt{N_r}} \\
&\approx \sum_{r=1}^{k} \gamma_r^{\mathrm{S}}\ \mathbb{V}_r \circ B_r
\end{aligned}$$

with

$$\gamma_r^{\mathrm{S}} := \frac{1}{k\sqrt{\pi_r}}.$$

Weights for $\widehat{F}^{\mathrm{M}}$

$$
\begin{aligned}
0 \stackrel{!}{=} & \; n\widehat{F}(t) - \sum_{r=1}^{k} N_r B_r(p) \\
= & \sum_{r=1}^{k} N_r \big(\widehat{F}_r(t) - B_r(t) + B_r(t) - B_r(p)\big) \\
\approx & \sum_{r=1}^{k} N_r \big(\widehat{F}_r(t) - B_r(t) - \beta_r(t)(p-t)\big) \\
\approx & \sqrt{n} \sum_{r=1}^{k} \sqrt{\pi_r}\, \mathbb{V}_r(B_r(t)) - n(p-t) \sum_{r=1}^{k} \pi_r \beta_r(t).
\end{aligned}
$$

$$0 \overset{!}{\approx} \sum_{r=1}^{k} \sqrt{\pi_r}\, \mathbb{V}_r(B_r(t)) - \sqrt{n}\,(p-t) \sum_{r=1}^{k} \pi_r \beta_r(t).$$

Hence we expect that

$$\sqrt{n}\,(\widehat{F}^{\mathrm{M}} - F) \approx \sum_{r=1}^{k} \gamma_r^{\mathrm{M}}\, \mathbb{V}_r \circ B_r$$

with

$$\gamma_r^{\mathrm{M}} := \frac{\sqrt{\pi_r}}{\sum_{s=1}^{k} \pi_s \beta_s}.$$

Weights for $\widehat{F}^{\mathrm{L}}$

With

$$w_r \;:=\; \frac{\beta_r}{B_r(1-B_r)}$$

we write

$$\begin{aligned}
0 \;\stackrel{!}{=}\; \frac{\partial}{\partial p} L(t,p) \;&=\; \sum_{r=1}^{k} N_r w_r(p)\big[\widehat{F}_r(t) - B_r(p)\big] \\
&\approx\; \sum_{r=1}^{k} N_r w_r(t)\big(\widehat{F}_r(t) - B_r(t) - \beta_r(t)(p-t)\big) \\
&\approx\; \sqrt{n}\sum_{r=1}^{k} \sqrt{\pi_r} w_r(t)\,\mathbb{V}_r(B_r(t)) - n(p-t)\sum_{r=1}^{k} \pi_r w_r(t)\beta_r(t).
\end{aligned}$$

$$0 \overset{!}{\approx} \sum_{r=1}^{k} \sqrt{\pi_r} w_r(t) \, \mathbb{V}_r(B_r(t)) - \sqrt{n}\,(p-t) \sum_{r=1}^{k} \pi_r w_r(t) \beta_r(t).$$

Hence we expect that

$$\sqrt{n}\,(\widehat{F}^{\mathrm{L}} - F) \;\approx\; \sum_{r=1}^{k} \gamma_r^{\mathrm{L}} \; \mathbb{V}_r \circ B_r$$

with

$$\gamma_r^{\mathrm{L}} \;:=\; \frac{\sqrt{\pi_r} w_r}{\sum_{s=1}^{k} \pi_s w_s \beta_s}.$$

Theorem 1 (Linear expansions). For $\mathrm{Z} = \mathrm{S}, \mathrm{L}, \mathrm{M}$ let

$$\mathbb{V}^{\mathrm{Z}} := \sum_{r=1}^{k} \gamma_r^{\mathrm{Z}} \, \mathbb{V}_r \circ B_r.$$

For any $\delta \in [0, 1/2)$,

$$\sup_{t \in (0,1)} \frac{\left| \sqrt{n} \left(\widehat{F}^{\mathrm{Z}}(t) - t \right) - \mathbb{V}^{\mathrm{Z}}(t) \right|}{t^{\delta}(1-t)^{\delta}} \to_p 0 \quad \text{as } n \to \infty.$$

$$\sup_{t \in (0,c) \cup (1-c,1)} \frac{\sqrt{n} \left| \mathbb{V}^{\mathrm{Z}}(t) \right|}{t^{\delta}(1-t)^{\delta}} \to_p 0 \quad \text{as } n \to \infty, c \downarrow 0.$$

All three estimators asymptotically equivalent in the tail regions!

Theorem 2 (Tail behaviour). Define

$$\mathbb{V}^{(\ell)} := \frac{\mathbb{V}_1 \circ B_1}{k\sqrt{N_1/n}} \quad \text{and} \quad \mathbb{V}^{(r)} := \frac{\mathbb{V}_k \circ B_k}{k\sqrt{N_k/n}}.$$

For $\mathrm{Z} = \mathrm{S}, \mathrm{L}, \mathrm{M}$ and any fixed $\kappa \in [1/2, 1)$,

$$\sup_{t\in(0,c)} \frac{\left|\sqrt{n}\,(\widehat{F}^{\mathrm{Z}}(t) - t) - \mathbb{V}^{(\ell)}(t)\right|}{t^{\kappa}} \to_p 0$$

$$\sup_{t\in(1-c,1)} \frac{\left|\sqrt{n}\,(\widehat{F}^{\mathrm{Z}}(t) - t) - \mathbb{V}^{(r)}(t)\right|}{(1-t)^{\kappa}} \to_p 0$$

as $n \to \infty$ and $c \downarrow 0$.

Theorem 3 (Limiting distributions). For $\mathrm{Z} = \mathrm{S}, \mathrm{L}, \mathrm{M}$,

$$\left(\sqrt{n}(\widehat{F}^{\mathrm{Z}}(t) - t)\right)_{t\in[0,1]} \;\to_{\mathcal{L}}\; \mathbb{G}^{\mathrm{Z}} \;=\; \sum_{r=1}^{k} \gamma_r^{\mathrm{Z}}\; \mathbb{G}_r \circ B_r$$

with independent Brownian bridges $\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_k$.

The covariance function K^{Z} of $\mathbb{G}^{\mathrm{Z}}$ equals

$$K^{\mathrm{Z}}(s,t) \;=\; \sum_{r=1}^{k} \gamma_r^{\mathrm{Z}}(s)\gamma_r^{\mathrm{Z}}(t)\; K\big(B_r(s), B_r(t)\big)$$

with

$$K(s,t) \;:=\; \min\{s,t\} - st.$$

In the balanced case, i.e. $\pi_1 = \cdots = \pi_k = 1/k$,

$$\mathbb{G}^{\mathrm{S}} \;\equiv\; \mathbb{G}^{\mathrm{M}}.$$

Theorem 4 (Asymptotic relative efficiencies). For any $t \in (0,1)$,

$$K^{\mathrm{L}}(t,t) \;\le\; K^{\mathrm{S}}(t,t) \quad \text{(equality for at most one } t\text{)}$$
$$K^{\mathrm{L}}(t,t) \;\le\; K^{\mathrm{M}}(t,t) \quad \text{(equality iff } k = 2 \text{ and } t = 1/2\text{)}.$$

On the other hand,

$$\sup_{\pi} \frac{K^{\mathrm{S}}(t,t)}{K^{\mathrm{L}}(t,t)} \;=\; \infty$$
$$\sup_{\pi} \frac{K^{\mathrm{M}}(t,t)}{K^{\mathrm{L}}(t,t)} \;=\; \frac{\rho(t) + \rho(t)^{-1} + 2}{4} \;<\; \infty$$

with

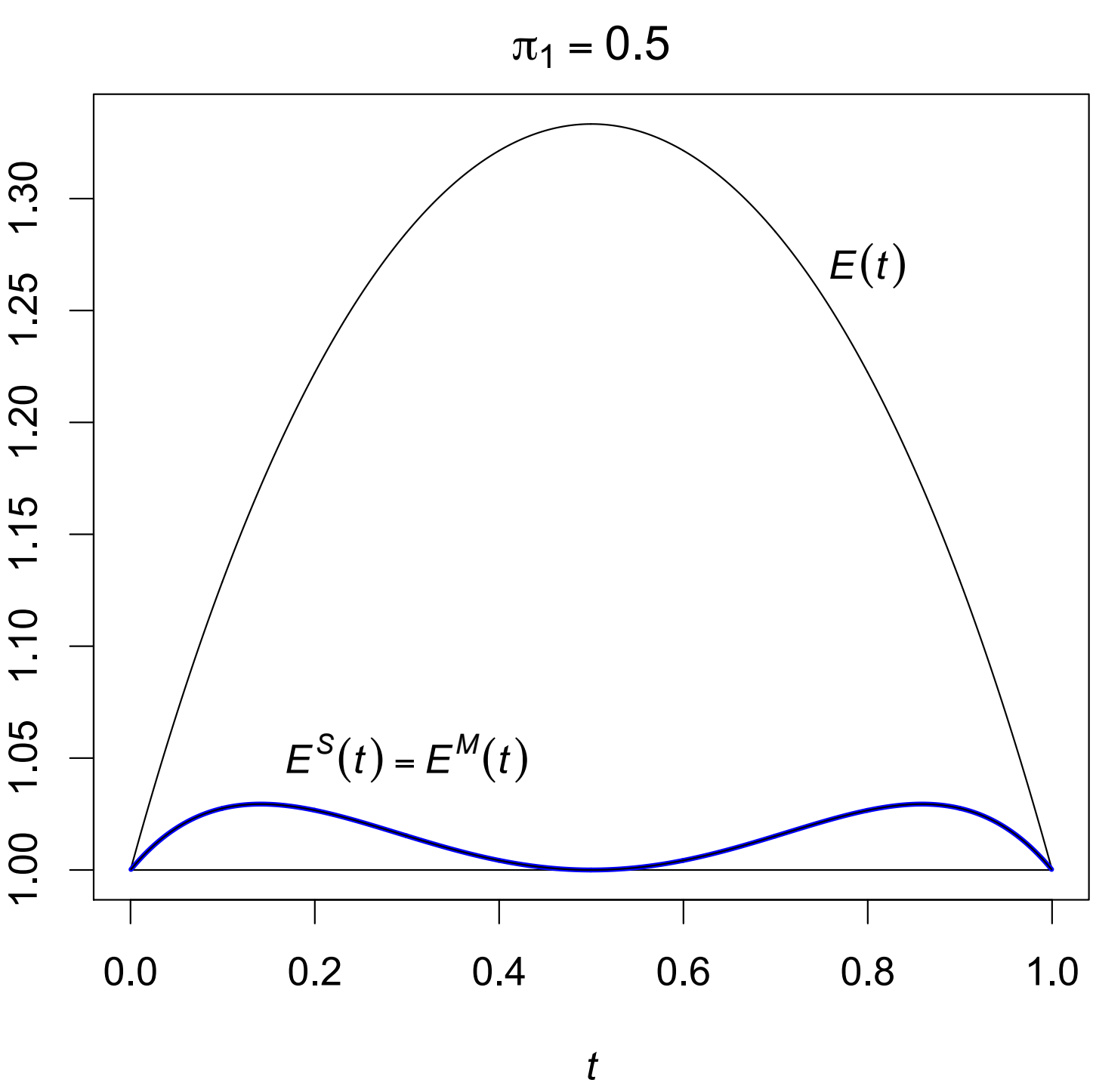
$$\rho(t) \;:=\; \max_r w_r(t) \big/ \min_r w_r(t).$$

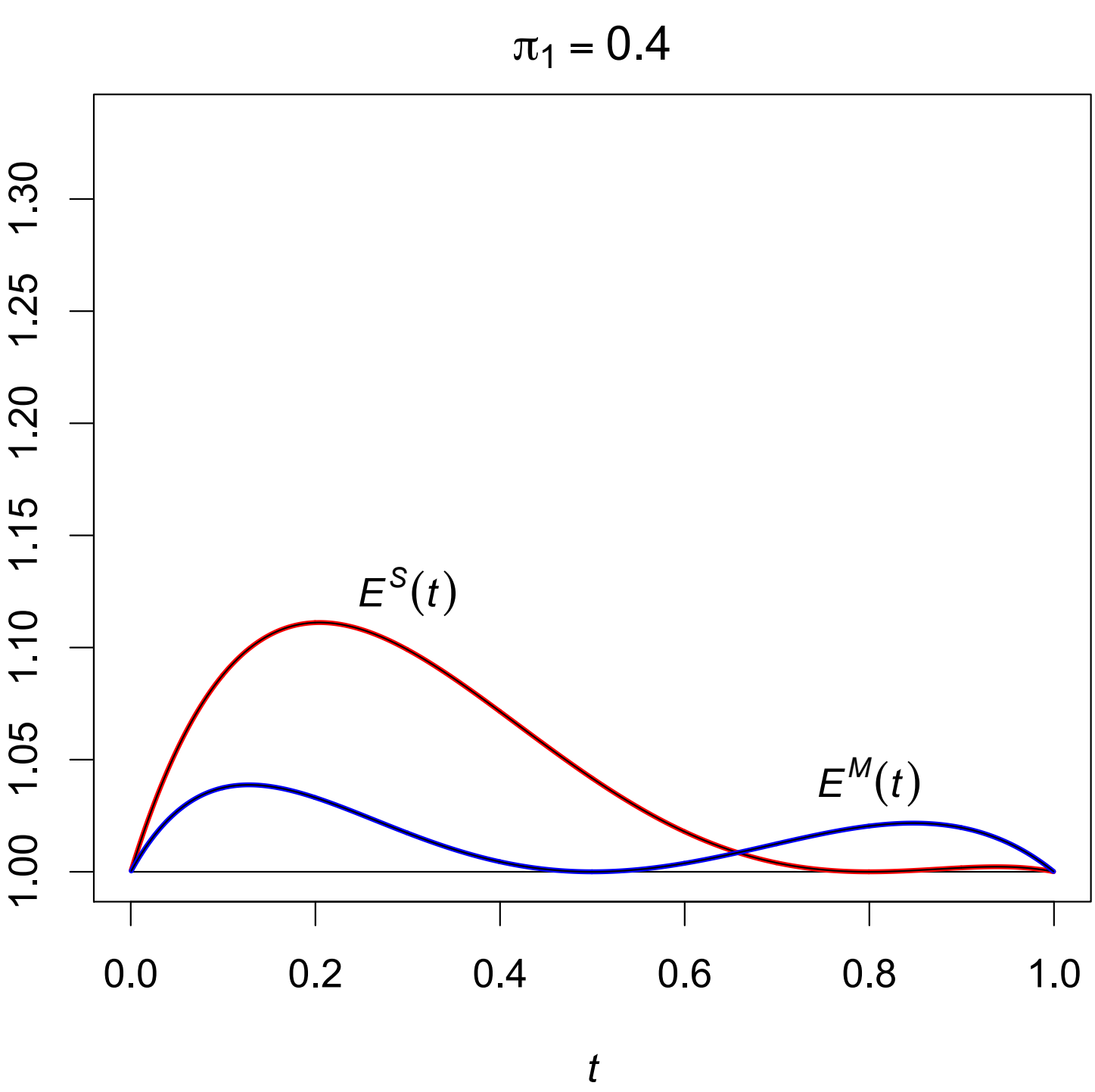
The case $k = 2$

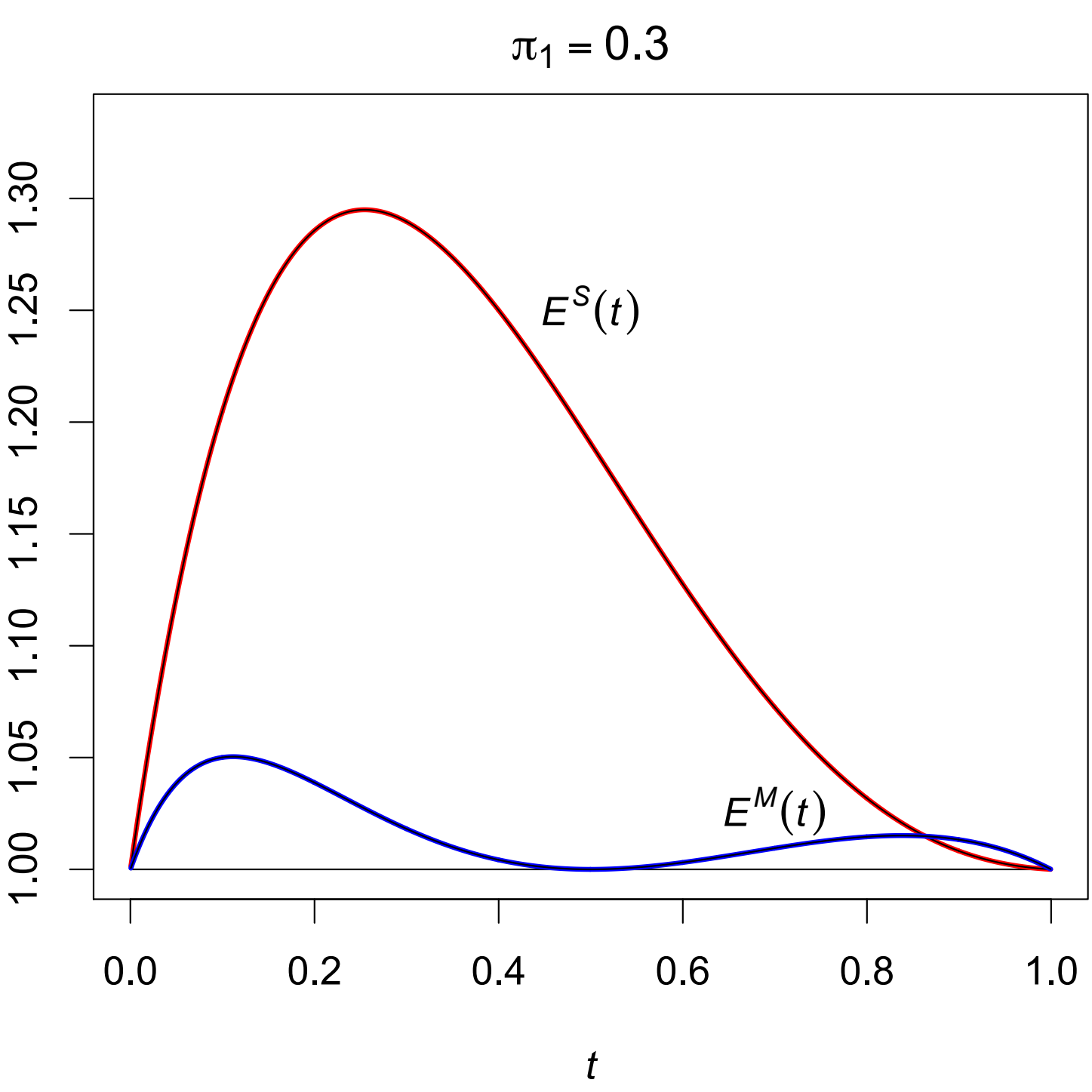
$$\frac{K^{\mathrm{M}}(t,t)}{K^{\mathrm{L}}(t,t)} \leq \frac{1}{1-(2t-1)^2/9} \leq 1.125.$$

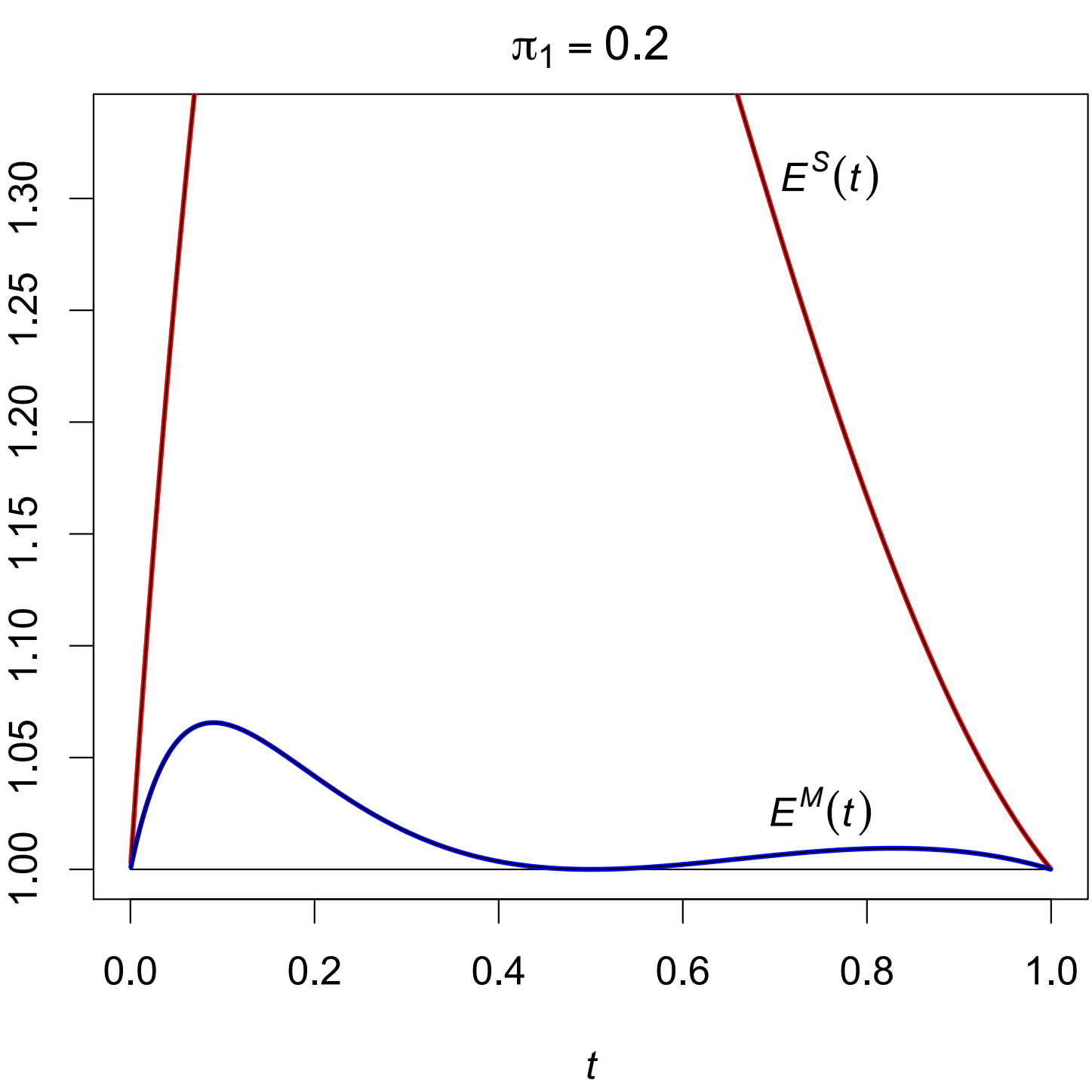
Plots of asymptotic relative (in)efficiencies

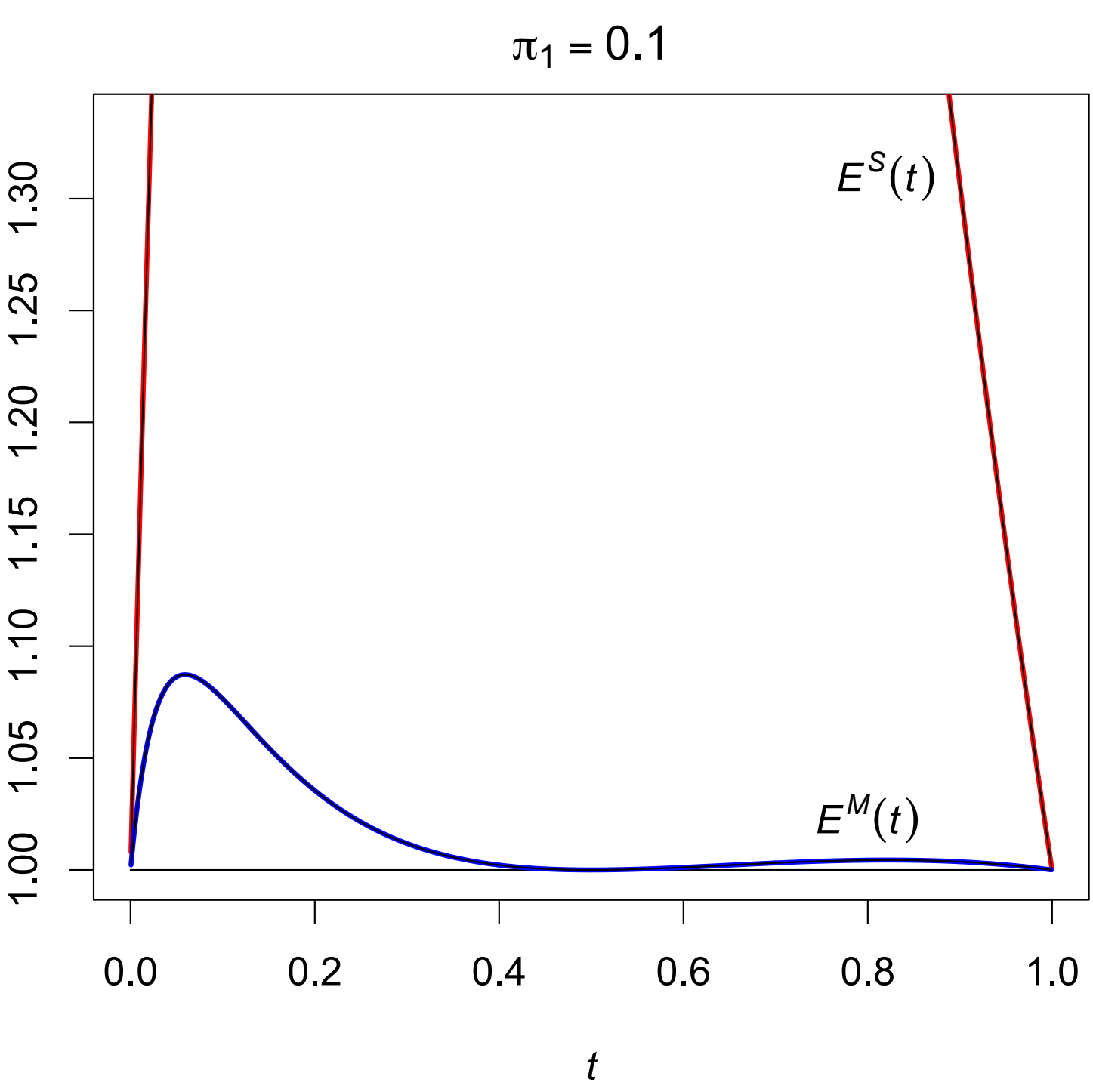
$$(0,1) \ni t \mapsto \begin{cases} E^{\mathrm{S}}(t) & := & K^{\mathrm{S}}(t,t)/K^{\mathrm{L}}(t,t) \\ E^{\mathrm{M}}(t) & := & K^{\mathrm{M}}(t,t)/K^{\mathrm{L}}(t,t) \\ E(t) & := & K(t,t)/K^{\mathrm{L}}(t,t) \end{cases}$$

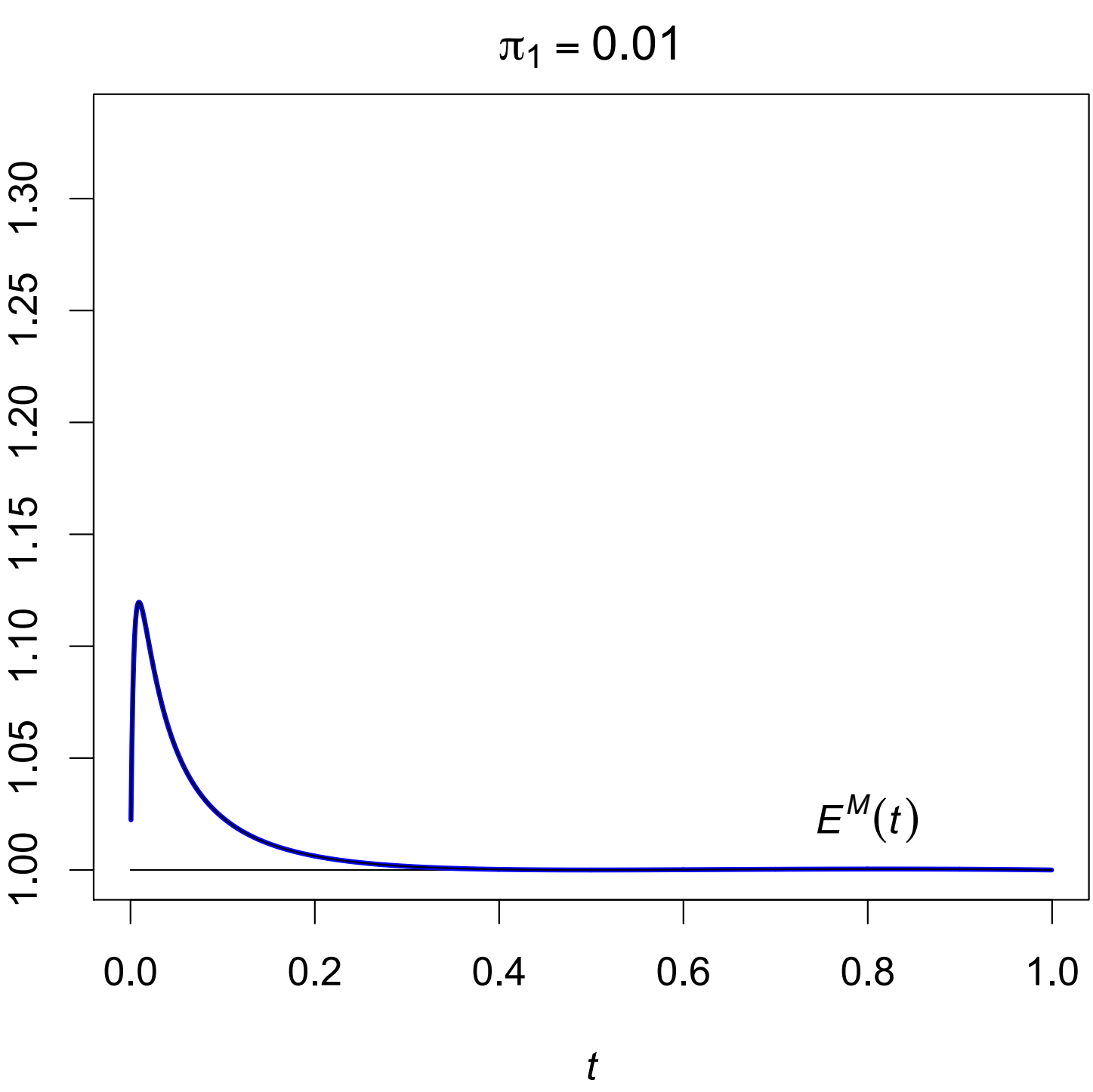

$\pi_1 = 0.5$
$E(t)$
$E^S(t) = E^M(t)$
1.00
1.05
1.10
1.15
1.20
1.25
1.30
0.0
0.2
0.4
0.6
0.8
1.0
t

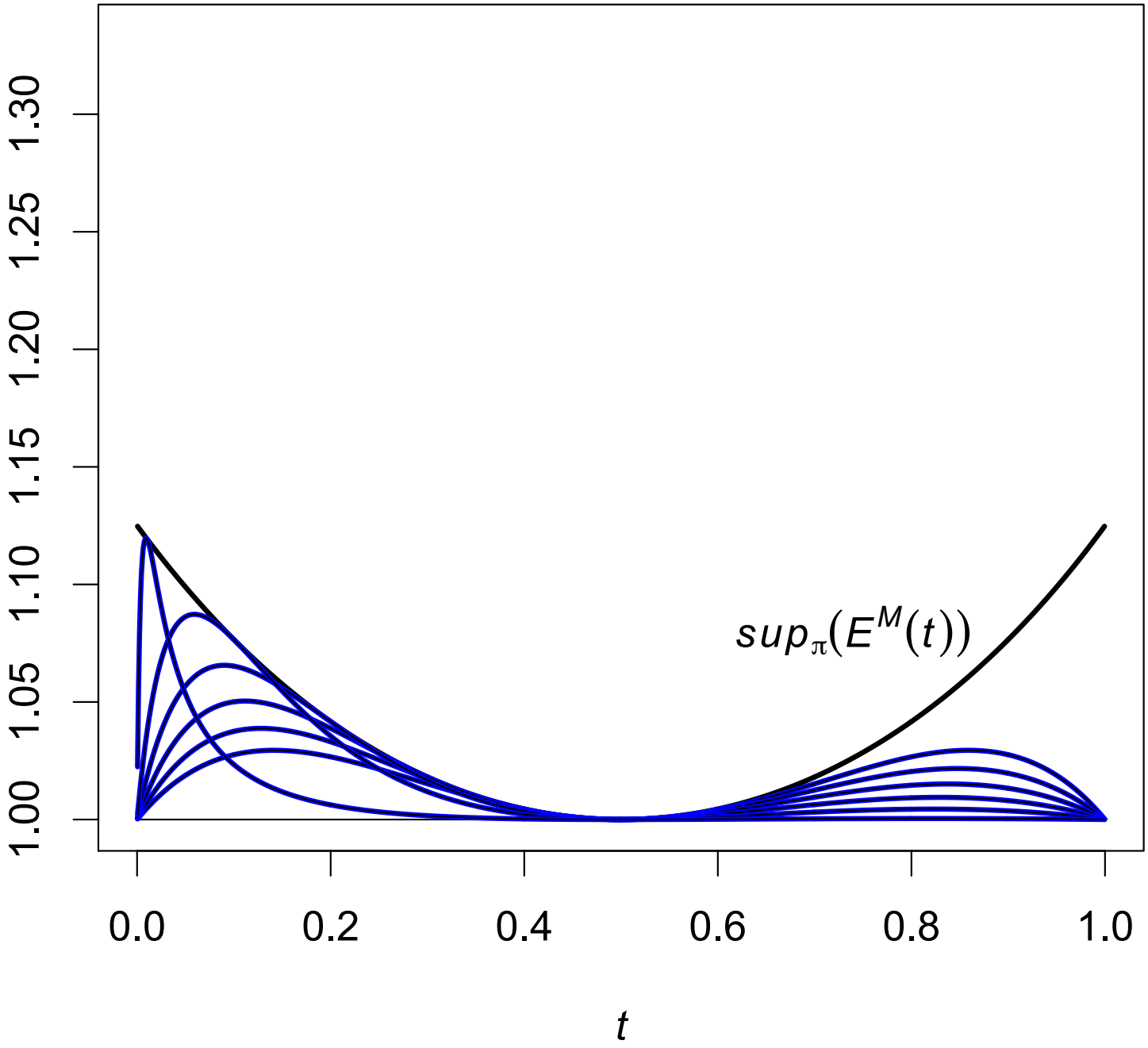

$\pi_1 = 0.4$
1.00
1.05
1.10
1.15
1.20
1.25
1.30
$E^S(t)$
$E^M(t)$
0.0
0.2
0.4
0.6
0.8
1.0
t


$\pi_1 = 0.3$
$E^S(t)$
$E^M(t)$
1.00
1.05
1.10
1.15
1.20
1.25
1.30
0.0
0.2
0.4
0.6
0.8
1.0
t


$\pi_1 = 0.2$
$E^S(t)$
$E^M(t)$
1.00
1.05
1.10
1.15
1.20
1.25
1.30
0.0
0.2
0.4
0.6
0.8
1.0
t


$\pi_1 = 0.1$
$E^S(t)$
$E^M(t)$
1.00
1.05
1.10
1.15
1.20
1.25
1.30
0.0
0.2
0.4
0.6
0.8
1.0
t


$\pi_1 = 0.01$
1.00
1.05
1.10
1.15
1.20
1.25
1.30
0.0
0.2
0.4
0.6
0.8
1.0
t
$E^M(t)$


1.00
1.05
1.10
1.15
1.20
1.25
1.30
0.0
0.2
0.4
0.6
0.8
1.0
t
$sup_{\pi}(E^{M}(t))$

IV. Exact Inference for Finite Samples

Pointwise confidence intervals

Adaptation of a method by Terpstra and Miller (2006) . . .

If $F(x) = p \in [0, 1]$, then conditional on $\mathbf{N} = (N_r)_{r=1}^k$,

$$n\widehat{F}(x) \ \sim \ \sum_{r=1}^{k} Y_{r,p}$$

with

$$Y_{1,p}, Y_{2,p}, \ldots, Y_{k,p} \text{ independent,}$$
$$Y_{r,p} \ \sim \ \mathrm{Bin}(N_r, B_r(p)).$$

Let

$$G_{\mathbf{N},p}(y) \;:=\; \mathbb{P}\Big(\sum_{r=1}^{k} Y_{r,p} \le y\Big).$$

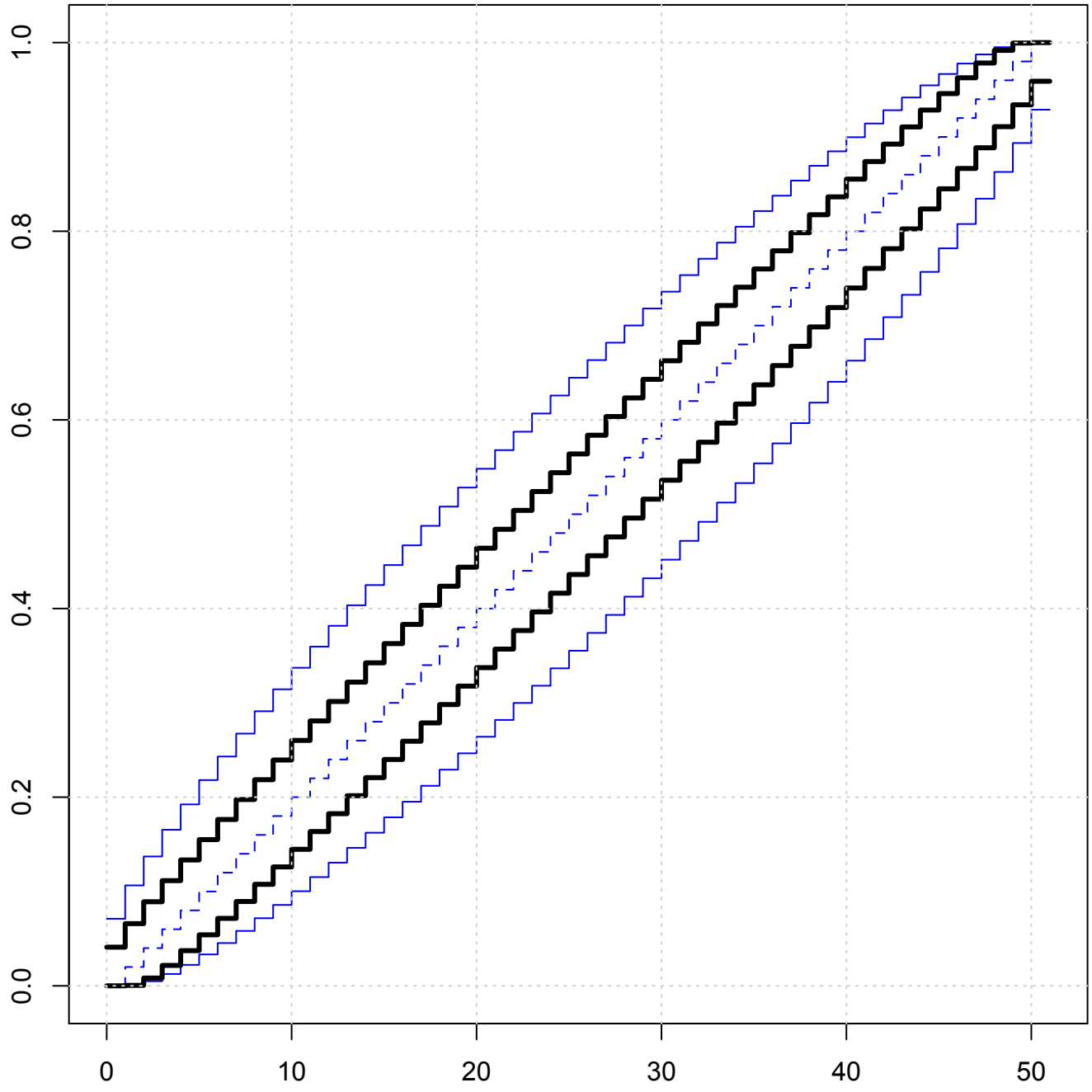
Exact p-values:

$$\begin{aligned}
&\text{Null hypothesis } \text{“}F(x) \ge p\text{”} \;:\quad G_{\mathbf{N},p}(n\widehat{F}(x)),\\
&\text{Null hypothesis } \text{“}F(x) \le p\text{”} \;:\quad 1 - G_{\mathbf{N},p}(n\widehat{F}(x) - 1).
\end{aligned}$$

Resulting one-sided confidence bounds:

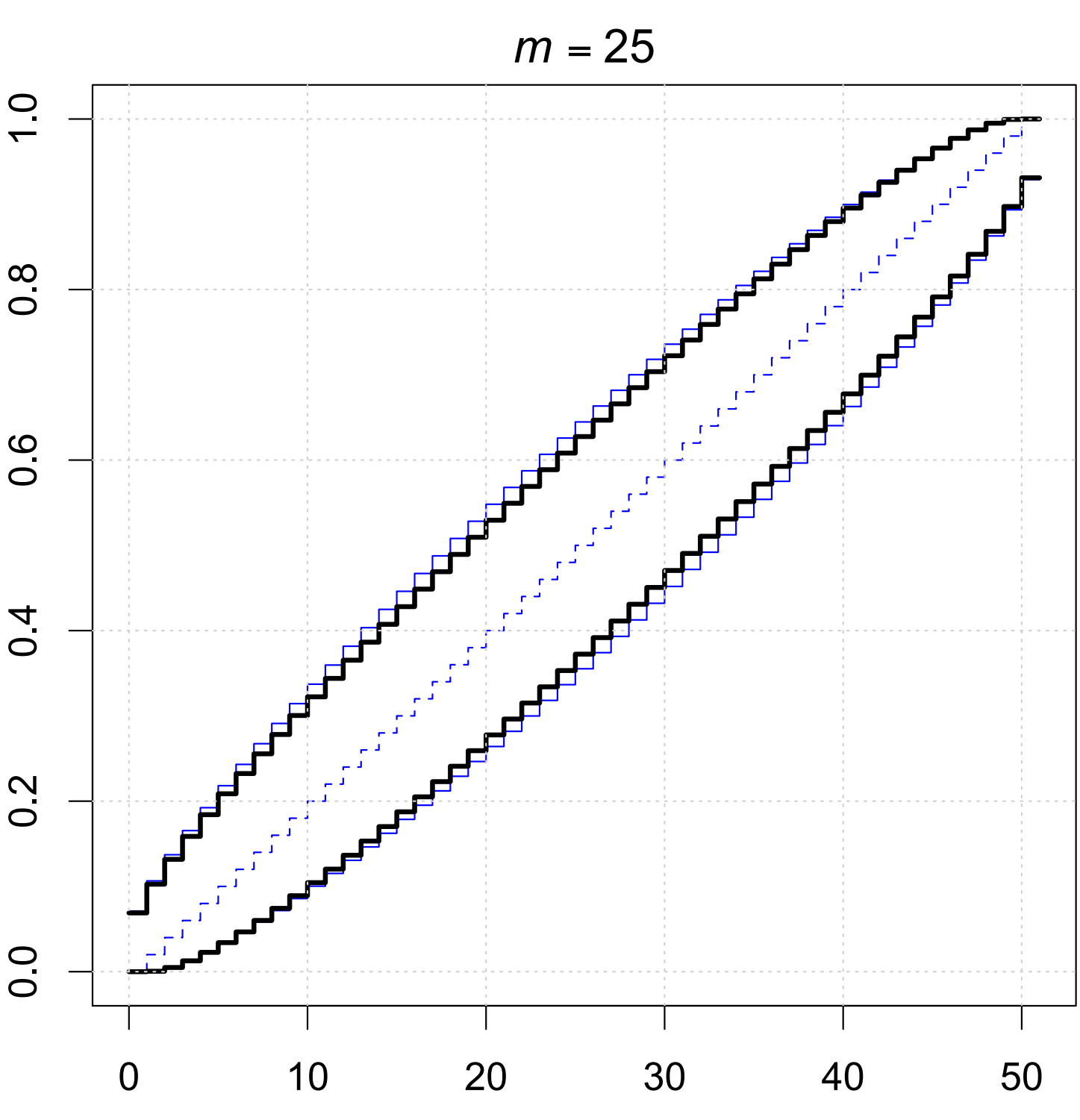
$$\begin{aligned}
\big\{p \in [0,1] : G_{\mathbf{N},p}(n\widehat{F}(x)) \ge \alpha\big\} &= \big[0, b_\alpha(\mathbf{N}, n\widehat{F}(x))\big],\\
\big\{p \in [0,1] : G_{\mathbf{N},p}(n\widehat{F}(x) - 1) \le 1 - \alpha\big\} &= \big[a_\alpha(\mathbf{N}, n\widehat{F}(x)), 1\big].
\end{aligned}$$

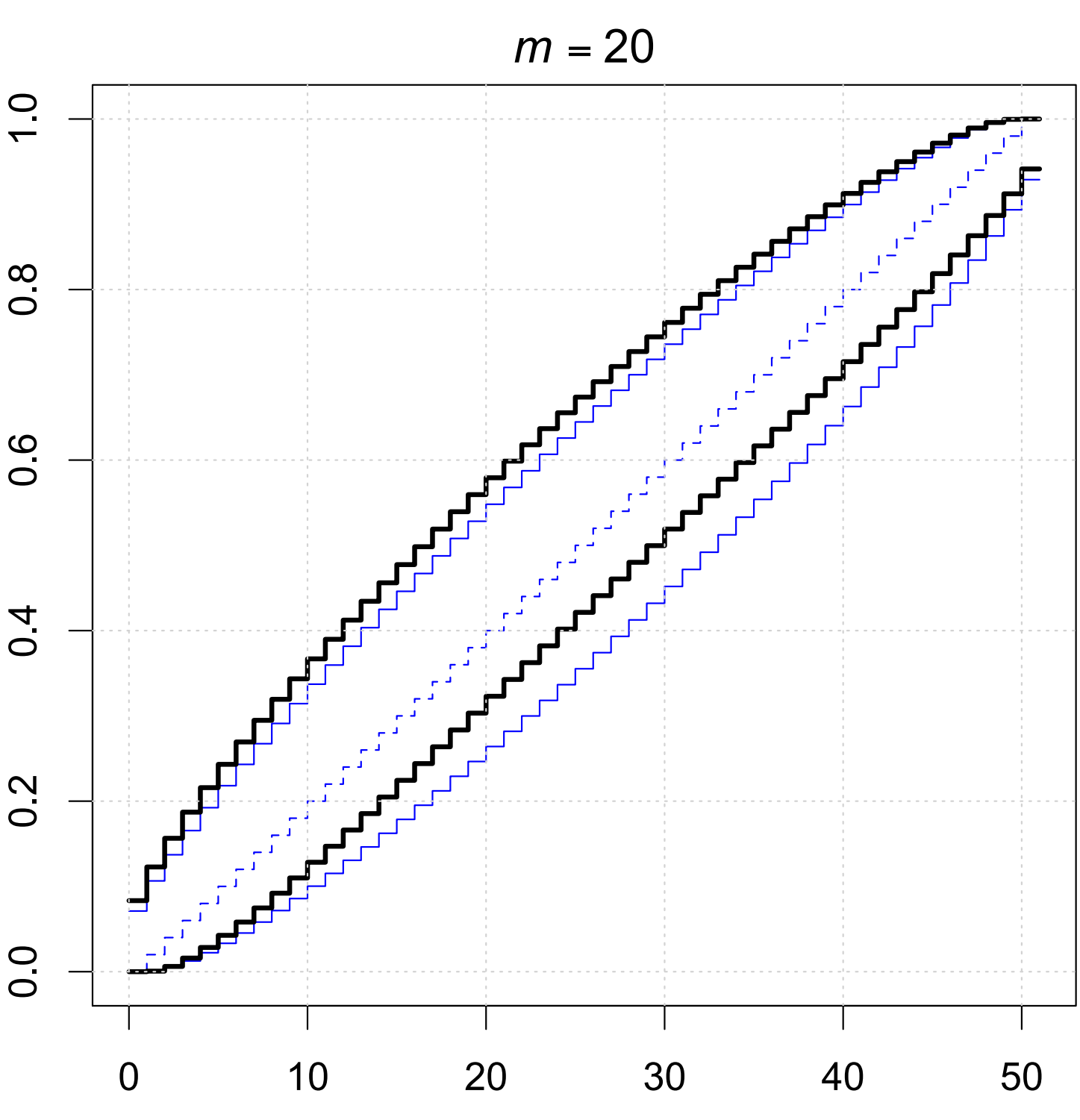
95%-confidence intervals in RSS setting, $n = k = 50$

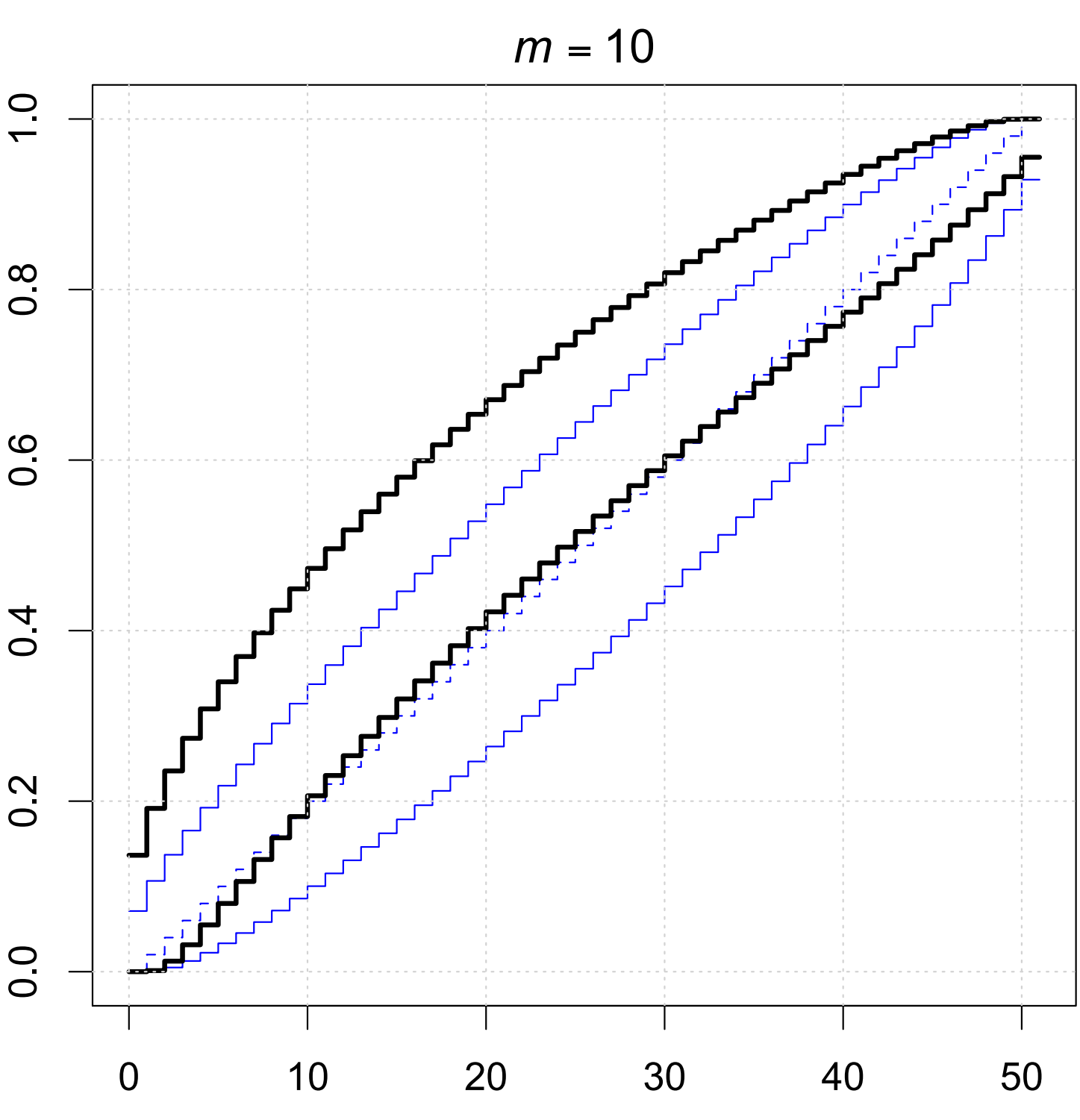


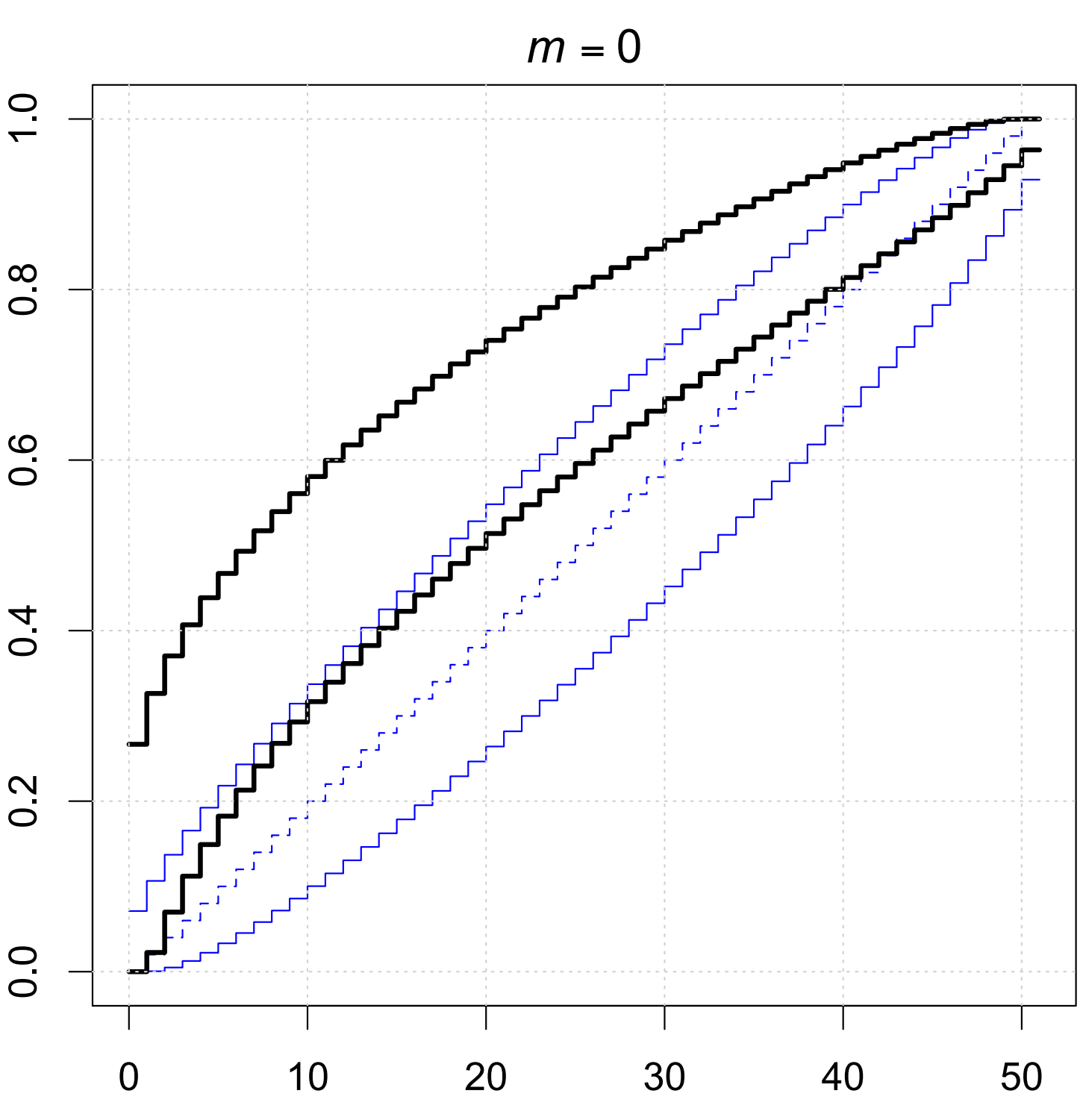
95%-confidence intervals when

$$k = 2 \quad \text{and} \quad \mathbf{N} = (m, 50 - m)^\top$$


$m = 25$
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


m = 20
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50


$m = 10$
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50


$m = 0$
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50

Confidence bands

With confidence $1-\alpha$,

$$F(x) \;\in\; \left[\widehat{F}^{\mathrm{Z}}(x) \pm \kappa^{\mathrm{Z}}(\mathbf{N},\alpha)\right] \quad \text{for all } x \in \mathbb{R}$$

where

$$\kappa^{\mathrm{Z}}(\mathbf{N},\alpha) \;:=\; (1-\alpha)\text{-quantile of } \mathcal{L}\left(\sup_{t\in[0,1]} \left|\widehat{B}^{\mathrm{Z}}(t) - t\right| \;\middle|\; \mathbf{N} \right)$$

$$\widehat{B}^{\mathrm{Z}} \;:\quad \text{distributed as } \widehat{F}^{\mathrm{Z}} \text{ in case of } F \mathrel{\widehat{=}} \mathrm{Unif}[0,1]$$

Estimation of $\kappa^{\mathrm{Z}}(\mathbf{N},\alpha)$ via Monte Carlo simulations,
particularly easy for $\mathrm{Z} = \mathrm{M}$ ...

$\kappa^{\mathrm{M}}(m) := \kappa^{\mathrm{M}}((m, 50 - m), 5\%)$: