Bootstrap módszerek és alkalmazásuk összefüggő adatsorokra

Varga László

Valószínűségelméleti és Statisztika Tanszék
Természettudományi Kar
Eötvös Loránd Tudományegyetem

TÁMOP Kutatószeminárium
2010. november 19.

Vázlat

Újramintavételezési módszerek

- Cél: adott $x_1, \ldots, x_n$ mintából minél több információt kisajtolni
- Jackknife ($\sim$50-es évek)
  - 1-törléses jackknife:
    i-edik jackknife minta: $\mathbf{x_i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ $i = 1, \ldots, n$
  - d-törléses jackknife:
    $\binom{n}{d}$ mintát csinálunk d mintaelem elhagyásával
- Bootstrap (Efron 1979.)
  $\mathbf{x_i^*} = \{x_1^*, \ldots, x_m^*\}$visszatevéses mintavétellel az eredeti mintából
  általában m=n

Újramintavételezési módszerek

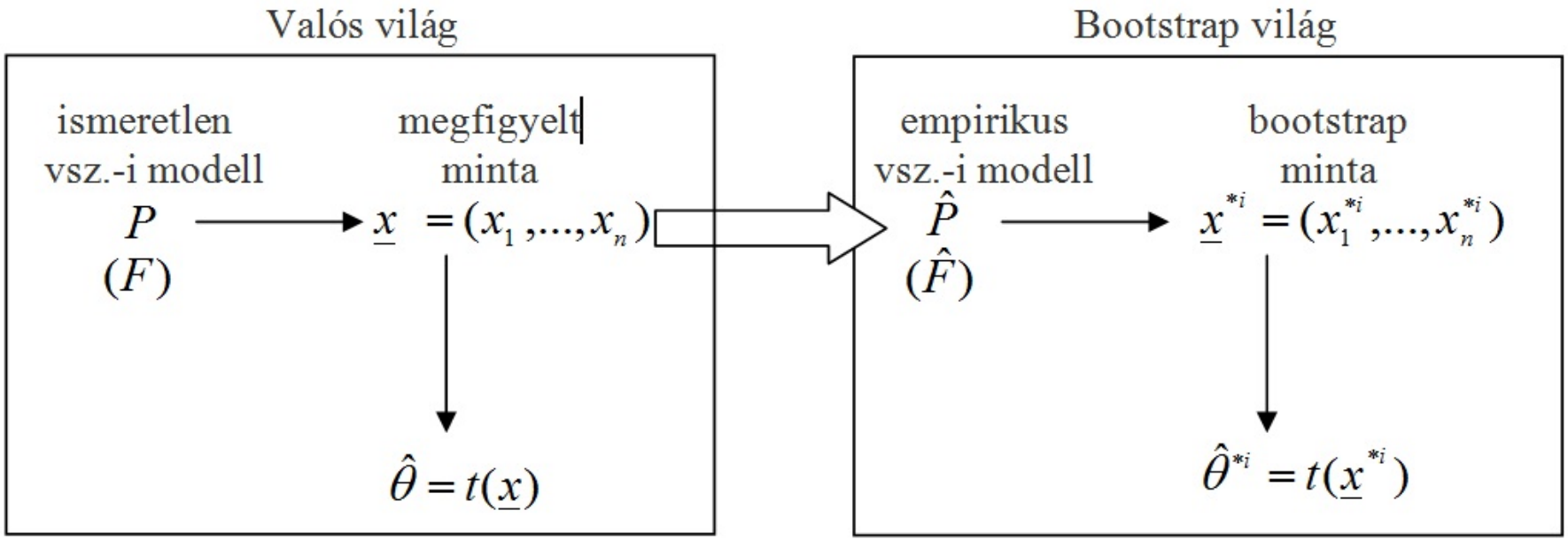
- Cél: adott $x_1, \ldots, x_n$ mintából minél több információt kisajtolni
- Jackknife ($\sim$50-es évek)
  - 1-törléses jackknife:
    i-edik jackknife minta: $\mathbf{x_i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ $i = 1, \ldots, n$
  - d-törléses jackknife:
    $\binom{n}{d}$ mintát csinálunk d mintaelem elhagyásával
- Bootstrap (Efron 1979.)
  $\mathbf{x_i^*} = \{x_1^*, \ldots, x_m^*\}$visszatevéses mintavétellel az eredeti mintából általában m=n

Újramintavételezési módszerek

- Cél: adott $x_1, \ldots, x_n$ mintából minél több információt kisajtolni
- Jackknife ($\sim$50-es évek)
  - 1-törléses jackknife:
    i-edik jackknife minta: $\mathbf{x_i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ $i = 1, \ldots, n$
  - d-törléses jackknife:
    $\binom{n}{d}$ mintát csinálunk d mintaelem elhagyásával
- Bootstrap (Efron 1979.)
  $\mathbf{x_i^*} = \{x_1^*, \ldots, x_m^*\}$visszatevéses mintavétellel az eredeti mintából általában m=n

Bootstrap - bevezetés

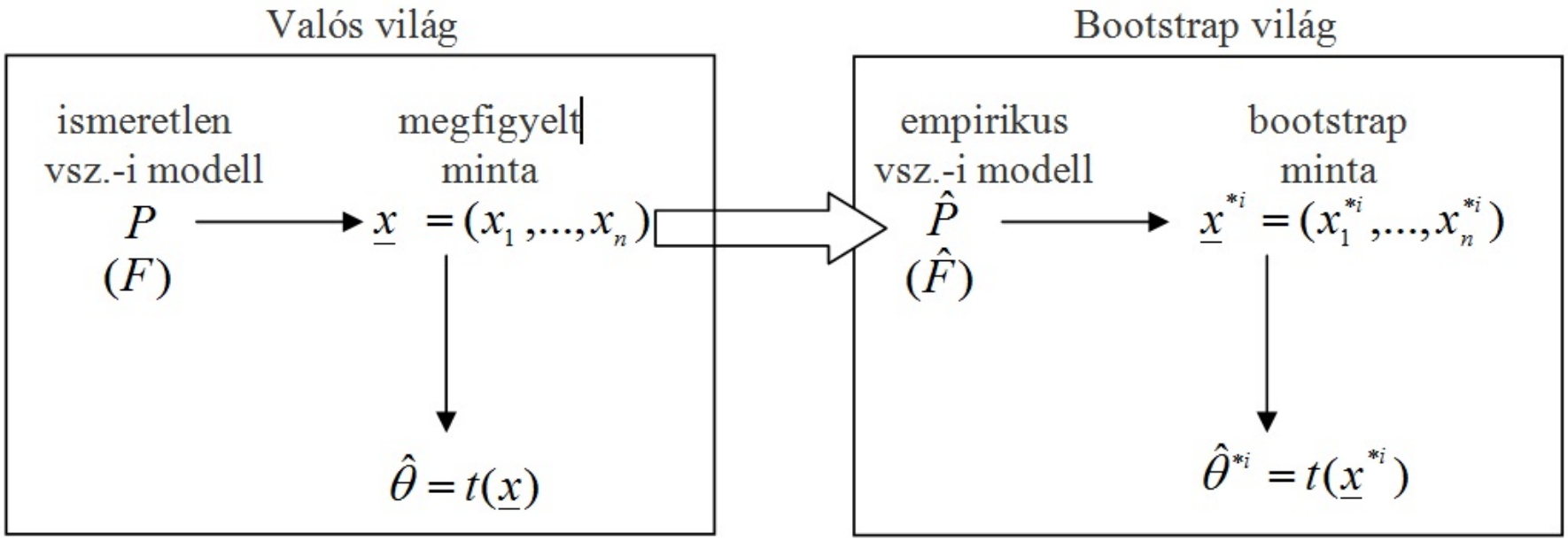
- A bootstrap módszerek lényege



- Nehézségek a gyakorlatban:
  1. $\underline{x} \Longrightarrow \hat{P}$ minden modellnél más és más
  2. $\hat{P} \longrightarrow \underline{x}^*$ a sok ismétlés megterheli a számítógépet

Bootstrap - bevezetés

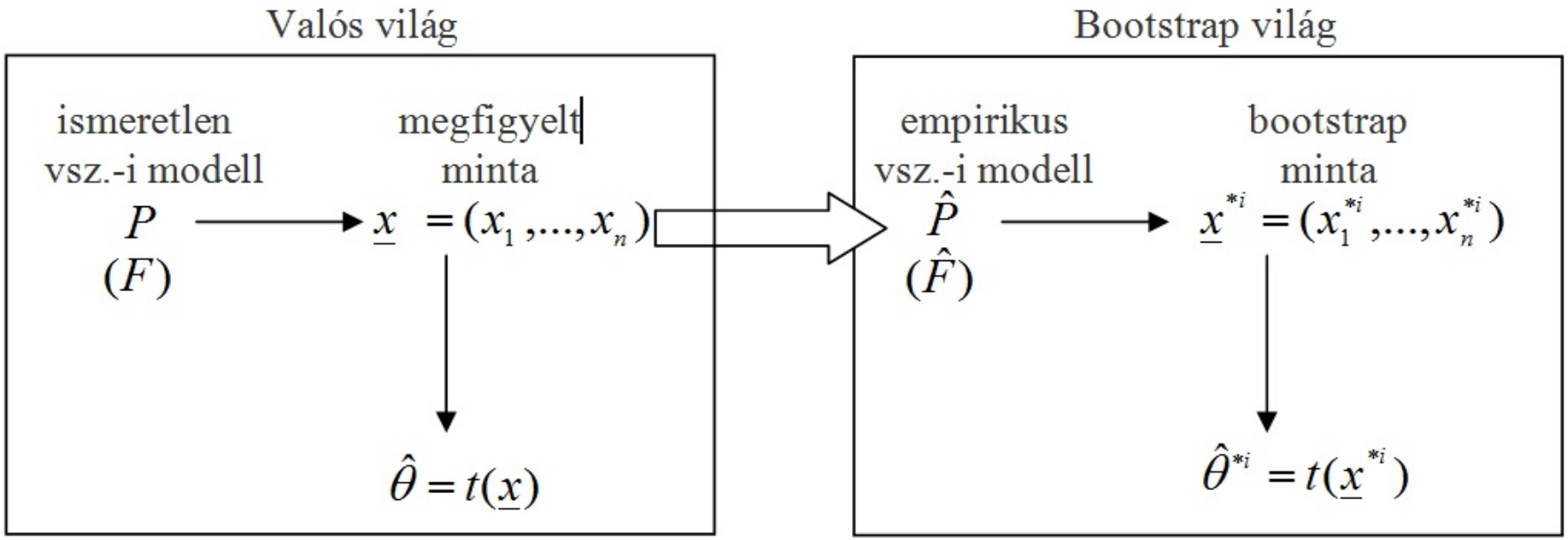
- A bootstrap módszerek lényege



- Nehézségek a gyakorlatban:
  1. $\underline{x} \Longrightarrow \hat{P}$ minden modellnél más és más
  2. $\hat{P} \longrightarrow \underline{x}^*$ a sok ismétlés megterheli a számítógépet

Bootstrap - bevezetés

- A bootstrap módszerek lényege



- Nehézségek a gyakorlatban:
  1. $\underline{x} \Longrightarrow \hat{P}$ minden modellnél más és más
  2. $\hat{P} \longrightarrow \underline{x}^*$ a sok ismétlés megterheli a számítógépet

Vázlat

Az i.i.d. bootstrap

- Legyenek $X_1, X_2, \ldots$ i.i.d. valószínűségi változók, F (ismeretlen) közös eloszlással
  - $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ véletlen minta
  - $T_n = t_n(\mathcal{X}_n; F)$ minket érdeklő val.változó, az eloszlása: G_n
- Cél: G_n eloszlásának becslése
- Bootstrap módszer:
  - Adott $\mathcal{X}$-re, visszatevéssel m elemű mintát veszünk: $\mathcal{X}_m^* = \{X_1^*, \ldots, X_n^*\}$ (általában m $\approx$ n)
  - az X_i^*-ok közös eloszlása: $F_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$
  - $T_{m,n}^* = t_m(\mathcal{X}_m^*; F_n)$
  - Ismétlések $\Rightarrow \hat{G}_{m,n}$

Az i.i.d. bootstrap

- Legyenek $X_1, X_2, \ldots$ i.i.d. valószínűségi változók, F (ismeretlen) közös eloszlással
  - $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ véletlen minta
  - $T_n = t_n(\mathcal{X}_n; F)$ minket érdeklő val.változó, az eloszlása: G_n
- Cél: G_n eloszlásának becslése
- Bootstrap módszer:
  - Adott $\mathcal{X}$-re, visszatevéssel m elemű mintát veszünk: $\mathcal{X}_m^* = \{X_1^*, \ldots, X_n^*\}$ (általában m $\approx$ n)
  - az X_i^*-ok közös eloszlása: $F_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$
  - $T_{m,n}^* = t_m(\mathcal{X}_m^*; F_n)$
  - Ismétlések $\Rightarrow \hat{G}_{m,n}$

Az i.i.d. bootstrap

- Legyenek $X_1, X_2, \ldots$ i.i.d. valószínűségi változók, F (ismeretlen) közös eloszlással
  - $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ véletlen minta
  - $T_n = t_n(\mathcal{X}_n; F)$ minket érdeklő val.változó, az eloszlása: G_n
- Cél: G_n eloszlásának becslése
- Bootstrap módszer:
  - Adott $\mathcal{X}$-re, visszatevéssel m elemű mintát veszünk: $\mathcal{X}_m^* = \{X_1^*, \ldots, X_n^*\}$ (általában m $\approx$ n)
  - az X_i^*-ok közös eloszlása: $F_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$
  - $T_{m,n}^* = t_m(\mathcal{X}_m^*; F_n)$
  - Ismétlések $\Rightarrow \hat{G}_{m,n}$

Az i.i.d. bootstrap előnyei és korlátai

- Az i.i.d. bootstrap előnyei a jackknife-hoz képest:
  1. Intervallumbecslésre és statisztika eloszlásának becslésére is használható
  2. "Bonyolultabb" statisztikák becslésére is alkalmas
- Az i.i.d. bootstrap korlátai:
  1. számításigényes
  2. bizonyos esetekben a becslés nem lesz konzisztens

Példa (Singh, 1981)

Def: $\{X_n\}_{n\geq 1}$ *m-függő* valamely $m \geq 0$ számra, ha $\{X_1, \ldots, X_k\}$ és $\{X_{k+m+1}, \ldots\}$ függetlenek minden $k \geq 0$-ra.

Jel. $\sigma_m^2 = Var(X_1) + 2\sum_{i=1}^{m-1} Cov(X_1, X_{1+i})$

Legyen a becsülendő statisztika: $T_n = \sqrt{n}(\overline{X}_n - \mu)$

Ekkor a bootstrap alteregója: $T_{n,n}^* = \sqrt{n}(\overline{X}_n^* - \overline{X}_n)$

Tétel: Legyen $\{X_n\}_{n\geq 1}$ stacionárius m-függő v.v. sorozat, $EX_1 = \mu$, $\sigma^2 = Var(X_1) \in (0, \infty)$, $\sum_{n=1}^{m} Cov(X_1, X_{1+i}) \neq 0$ és $\sigma_m^2 \neq 0$ Ekkor

$$\lim_{n\to\infty} \sup_x |P_*(T_{n,n}^* \leq x) - P(T_n \leq x)| \neq 0 \text{ m.m. } \forall x \neq 0 - ra$$

Az i.i.d. bootstrap előnyei és korlátai

- Az i.i.d. bootstrap előnyei a jackknife-hoz képest:
  1. Intervallumbecslésre és statisztika eloszlásának becslésére is használható
  2. "Bonyolultabb" statisztikák becslésére is alkalmas
- Az i.i.d. bootstrap korlátai:
  1. számításigényes
  2. bizonyos esetekben a becslés nem lesz konzisztens

Példa (Singh, 1981)

Def: $\{X_n\}_{n\geq 1}$ *m-függő* valamely $m \geq 0$ számra, ha $\{X_1, \ldots, X_k\}$ és $\{X_{k+m+1}, \ldots\}$ függetlenek minden $k \geq 0$-ra.

Jel. $\sigma_m^2 = Var(X_1) + 2\sum_{i=1}^{m-1} Cov(X_1, X_{1+i})$

Legyen a becsülendő statisztika: $T_n = \sqrt{n}(\overline{X}_n - \mu)$

Ekkor a bootstrap alteregója: $T_{n,n}^* = \sqrt{n}(\overline{X}_n^* - \overline{X}_n)$

Tétel: Legyen $\{X_n\}_{n\geq 1}$ stacionárius m-függő v.v. sorozat, $EX_1 = \mu$, $\sigma^2 = Var(X_1) \in (0, \infty)$, $\sum_{n=1}^{m} Cov(X_1, X_{1+i}) \neq 0$ és $\sigma_m^2 \neq 0$ Ekkor

$$\lim_{n\to\infty} \sup_x |P_*(T_{n,n}^* \leq x) - P(T_n \leq x)| \neq 0 \text{ m.m. } \forall x \neq 0 - ra$$

Az i.i.d. bootstrap előnyei és korlátai

- Az i.i.d. bootstrap előnyei a jackknife-hoz képest:
  1. Intervallumbecslésre és statisztika eloszlásának becslésére is használható
  2. "Bonyolultabb" statisztikák becslésére is alkalmas
- Az i.i.d. bootstrap korlátai:
  1. számításigényes
  2. bizonyos esetekben a becslés nem lesz konzisztens

Példa (Singh, 1981)

Def: $\{X_n\}_{n\geq 1}$ *m-függő* valamely $m \geq 0$ számra, ha $\{X_1, \ldots, X_k\}$ és $\{X_{k+m+1}, \ldots\}$ függetlenek minden $k \geq 0$-ra.

Jel. $\sigma_m^2 = Var(X_1) + 2\sum_{i=1}^{m-1} Cov(X_1, X_{1+i})$

Legyen a becsülendő statisztika: $T_n = \sqrt{n}(\overline{X}_n - \mu)$

Ekkor a bootstrap alteregója: $T_{n,n}^* = \sqrt{n}(\overline{X}_n^* - \overline{X}_n)$

Tétel: Legyen $\{X_n\}_{n\geq 1}$ stacionárius m-függő v.v. sorozat, $EX_1 = \mu$, $\sigma^2 = Var(X_1) \in (0, \infty)$, $\sum_{n=1}^{m} Cov(X_1, X_{1+i}) \neq 0$ és $\sigma_m^2 \neq 0$ Ekkor

$$\lim_{n\to\infty} \sup_x |P_*(T_{n,n}^* \leq x) - P(T_n \leq x)| \neq 0 \text{ m.m. } \forall x \neq 0 - ra$$

Az i.i.d. bootstrap előnyei és korlátai

- Az i.i.d. bootstrap előnyei a jackknife-hoz képest:
  1. Intervallumbecslésre és statisztika eloszlásának becslésére is használható
  2. "Bonyolultabb" statisztikák becslésére is alkalmas
- Az i.i.d. bootstrap korlátai:
  1. számításigényes
  2. bizonyos esetekben a becslés nem lesz konzisztens

Példa (Singh, 1981)

Def: $\{X_n\}_{n\geq 1}$ *m-függő* valamely $m \geq 0$ számra, ha $\{X_1, \ldots, X_k\}$ és $\{X_{k+m+1}, \ldots\}$ függetlenek minden $k \geq 0$-ra.

Jel. $\sigma_m^2 = Var(X_1) + 2\sum_{i=1}^{m-1} Cov(X_1, X_{1+i})$

Legyen a becsülendő statisztika: $T_n = \sqrt{n}(\overline{X}_n - \mu)$

Ekkor a bootstrap alteregója: $T_{n,n}^* = \sqrt{n}(\overline{X}_n^* - \overline{X}_n)$

Tétel: Legyen $\{X_n\}_{n\geq 1}$ stacionárius m-függő v.v. sorozat, $EX_1 = \mu$, $\sigma^2 = Var(X_1) \in (0, \infty)$, $\sum_{n=1}^{m} Cov(X_1, X_{1+i}) \neq 0$ és $\sigma_m^2 \neq 0$ Ekkor

$$\lim_{n\to\infty} \sup_x |P_*(T_{n,n}^* \leq x) - P(T_n \leq x)| \neq 0 \text{ m.m. } \forall x \neq 0 - ra$$

Vázlat

Circular blokk bootstrap (CBB)

1. $Y_t = X_{t_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
2. Legyenek $i_1, i_2, \ldots$ minta az {1,...,n} halmazon egyenletes eloszlásból
3. Adott b blokkméretre készítsünk $n'=mb$ ($n' \approx$ n) pszeudo-megfigyelést:

$$Y^*_{(k-1)b+j} = Y_{i_m+j-1} \quad \text{ahol j=1,...,}b\text{;} \quad \text{k=1,...,}m$$

4. A minket érdeklő statisztika kiszámítása a pszeudo-megfigyelésekből:

$$\overline{Y}^*_{n'} = (n')^{-1}(Y^*_1 + \ldots + Y^*_n)$$

Circular blokk bootstrap (CBB)

1. $Y_t = X_{t_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
2. Legyenek $i_1, i_2, \ldots$ minta az {1,...,n} halmazon egyenletes eloszlásból
3. Adott b blokkméretre készítsünk $n'=mb$ ($n' \approx$ n) pszeudo-megfigyelést:

$$Y^*_{(k-1)b+j} = Y_{i_m+j-1} \quad \text{ahol j=1,...,}b\text{;} \quad \text{k=1,...,}m$$

4. A minket érdeklő statisztika kiszámítása a pszeudo-megfigyelésekből:

$$\overline{Y}^*_{n'} = (n')^{-1}(Y^*_1 + \ldots + Y^*_n)$$

Circular blokk bootstrap (CBB)

1. $Y_t = X_{t_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
2. Legyenek $i_1, i_2, \ldots$ minta az {1,...,n} halmazon egyenletes eloszlásból
3. Adott b blokkméretre készítsünk $n'=mb$ ($n' \approx$ n) pszeudo-megfigyelést:

$$Y^*_{(k-1)b+j} = Y_{i_m+j-1} \quad \text{ahol j=1,...,}b\text{;} \quad \text{k=1,...,}m$$

4. A minket érdeklő statisztika kiszámítása a pszeudo-megfigyelésekből:

$$\overline{Y}^*_{n'} = (n')^{-1}(Y^*_1 + \ldots + Y^*_n)$$

Circular blokk bootstrap (CBB)

1. $Y_t = X_{t_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
2. Legyenek $i_1, i_2, \ldots$ minta az {1,...,n} halmazon egyenletes eloszlásból
3. Adott b blokkméretre készítsünk $n'=mb$ ($n' \approx$ n) pszeudo-megfigyelést:

$$Y^*_{(k-1)b+j} = Y_{i_m+j-1} \quad \text{ahol j=1,...,}b\text{;} \quad \text{k=1,...,}m$$

4. A minket érdeklő statisztika kiszámítása a pszeudo-megfigyelésekből:

$$\overline{Y}^*_{n'} = (n')^{-1}(Y^*_1 + \ldots + Y^*_n)$$

Általánosított blokk bootstrap (GBB)

- $Y_i = X_{i_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
- $I_1, I_2, \ldots$ az adott $A \subseteq \{1, \ldots, n\}$ halmazon független, egyenletes eloszlású v.v. sorozat $\longrightarrow$ blokkok kezdőindexei
- $J_1, J_2, \ldots$ i.i.d. v.v. sorozat $\longrightarrow$ blokkok hosszai
- I_i és J_i nem feltétlenül függetlenek!
- Blokkok kiválasztása $B(I_1, J_1) = \{Y_{I_1}, Y_{I_1+1}, \ldots, Y_{I_1+J_1-1}\}$

Gyakran használt speciális esetek:

1. CBB (circular block bootstrap)
   - $A = \{1, \ldots, n\}$
   - P(J_1=b)=1 $\rightarrow$ fix blokkméret
   - I_i és J_i függetlenek
2. SBB(stationary block bootstrap)
   - $A = \{1, \ldots, n\}$
   - $J_1 \sim \text{Geo}(p)$ (p-t egy "elvárt" blokkmérethez határozzuk meg)
   - I_i és J_i függetlenek

Általánosított blokk bootstrap (GBB)

- $Y_i = X_{i_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
- $I_1, I_2, \ldots$ az adott $A \subseteq \{1, \ldots, n\}$ halmazon független, egyenletes eloszlású v.v. sorozat $\longrightarrow$ blokkok kezdőindexei
- $J_1, J_2, \ldots$ i.i.d. v.v. sorozat $\longrightarrow$ blokkok hosszai
- I_i és J_i nem feltétlenül függetlenek!
- Blokkok kiválasztása $B(I_1, J_1) = \{Y_{I_1}, Y_{I_1+1}, \ldots, Y_{I_1+J_1-1}\}$

Gyakran használt speciális esetek:

1. CBB (circular block bootstrap)
   - $A = \{1, \ldots, n\}$
   - P(J_1=b)=1 $\rightarrow$ fix blokkméret
   - I_i és J_i függetlenek
2. SBB(stationary block bootstrap)
   - $A = \{1, \ldots, n\}$
   - $J_1 \sim \text{Geo}(p)$ (p-t egy "elvárt" blokkmérethez határozzuk meg)
   - I_i és J_i függetlenek

Általánosított blokk bootstrap (GBB)

- $Y_i = X_{i_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
- $I_1, I_2, \ldots$ az adott $A \subseteq \{1, \ldots, n\}$ halmazon független, egyenletes eloszlású v.v. sorozat $\longrightarrow$ blokkok kezdőindexei
- $J_1, J_2, \ldots$ i.i.d. v.v. sorozat $\longrightarrow$ blokkok hosszai
- I_i és J_i nem feltétlenül függetlenek!
- Blokkok kiválasztása $B(I_1, J_1) = \{Y_{I_1}, Y_{I_1+1}, \ldots, Y_{I_1+J_1-1}\}$

Gyakran használt speciális esetek:

1. CBB (circular block bootstrap)
   - $A = \{1, \ldots, n\}$
   - P(J_1=b)=1 $\rightarrow$ fix blokkméret
   - I_i és J_i függetlenek
2. SBB(stationary block bootstrap)
   - $A = \{1, \ldots, n\}$
   - $J_1 \sim \text{Geo}(p)$ (p-t egy "elvárt" blokkmérethez határozzuk meg)
   - I_i és J_i függetlenek

Általánosított blokk bootstrap (GBB)

- $Y_i = X_{i_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
- $I_1, I_2, \ldots$ az adott $A \subseteq \{1,\ldots,n\}$ halmazon független, egyenletes eloszlású v.v. sorozat $\longrightarrow$ blokkok kezdőindexei
- $J_1, J_2, \ldots$ i.i.d. v.v. sorozat $\longrightarrow$ blokkok hosszai
- I_i és J_i nem feltétlenül függetlenek!
- Blokkok kiválasztása $B(I_1, J_1) = \{Y_{I_1}, Y_{I_1+1}, \ldots, Y_{I_1+J_1-1}\}$

Gyakran használt speciális esetek:

1. CBB (circular block bootstrap)
   - $A = \{1, \ldots, n\}$
   - P(J_1=b)=1 $\rightarrow$ fix blokkméret
   - I_i és J_i függetlenek
2. SBB(stationary block bootstrap)
   - $A = \{1, \ldots, n\}$
   - $J_1 \sim$ Geo(p) (p-t egy "elvárt" blokkmérethez határozzuk meg)
   - I_i és J_i függetlenek

Általánosított blokk bootstrap (GBB)

- $Y_i = X_{i_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
- $I_1, I_2, \ldots$ az adott $A \subseteq \{1, \ldots, n\}$ halmazon független, egyenletes eloszlású v.v. sorozat $\longrightarrow$ blokkok kezdőindexei
- $J_1, J_2, \ldots$ i.i.d. v.v. sorozat $\longrightarrow$ blokkok hosszai
- I_i és J_i nem feltétlenül függetlenek!
- Blokkok kiválasztása $B(I_1, J_1) = \{Y_{I_1}, Y_{I_1+1}, \ldots, Y_{I_1+J_1-1}\}$

Gyakran használt speciális esetek:

1. CBB (circular block bootstrap)
   - $A = \{1, \ldots, n\}$
   - P(J_1=b)=1 $\rightarrow$ fix blokkméret
   - I_i és J_i függetlenek
2. SBB(stationary block bootstrap)
   - $A = \{1, \ldots, n\}$
   - $J_1 \sim \text{Geo}(p)$ (p-t egy "elvárt" blokkmérethez határozzuk meg)
   - I_i és J_i függetlenek

Általánosított blokk bootstrap (GBB)

- $Y_i = X_{i_{mod(n)}}$ azaz periodikusan kiterjesztjük a mintát
- $I_1, I_2, \ldots$ az adott $A \subseteq \{1, \ldots, n\}$ halmazon független, egyenletes eloszlású v.v. sorozat $\longrightarrow$ blokkok kezdőindexei
- $J_1, J_2, \ldots$ i.i.d. v.v. sorozat $\longrightarrow$ blokkok hosszai
- I_i és J_i nem feltétlenül függetlenek!
- Blokkok kiválasztása $B(I_1, J_1) = \{Y_{I_1}, Y_{I_1+1}, \ldots, Y_{I_1+J_1-1}\}$

Gyakran használt speciális esetek:

1. CBB (circular block bootstrap)
   - $A = \{1, \ldots, n\}$
   - P(J_1=b)=1 $\rightarrow$ fix blokkméret
   - I_i és J_i függetlenek
2. SBB(stationary block bootstrap)
   - $A = \{1, \ldots, n\}$
   - $J_1 \sim$ Geo(p) (p-t egy "elvárt" blokkmérethez határozzuk meg)
   - I_i és J_i függetlenek

Vázlat

Blokkméret kiválasztása

Több tényezőtől is függ

- az adatgeneráló folyamat - van-e összefüggőség:
  - térbeli
  - időbeli
- a bennünket érdeklő statisztika, például
  - átlag
  - medián
  - lag(k) korreláció
- a statisztikából mit szeretnénk számolni, ha például
  - szórást $\Longrightarrow b_{opt} = C \cdot n^{1/3}$
  - egyoldali eloszlást $\Longrightarrow b_{opt} = C \cdot n^{1/4}$

De ha tudjuk is a nagyságrendet, C=?

Blokkméret kiválasztása

Több tényezőtől is függ

- az adatgeneráló folyamat - van-e összefüggőség:
  - térbeli
  - időbeli
- a bennünket érdeklő statisztika, például
  - átlag
  - medián
  - lag(k) korreláció
- a statisztikából mit szeretnénk számolni, ha például
  - szórást $\Longrightarrow b_{opt} = C \cdot n^{1/3}$
  - egyoldali eloszlást $\Longrightarrow b_{opt} = C \cdot n^{1/4}$

De ha tudjuk is a nagyságrendet, C=?

Blokkméret kiválasztása

Több tényezőtől is függ

- az adatgeneráló folyamat - van-e összefüggőség:
  - térbeli
  - időbeli
- a bennünket érdeklő statisztika, például
  - átlag
  - medián
  - lag(k) korreláció
- a statisztikából mit szeretnénk számolni, ha például
  - szórást $\Longrightarrow b_{opt} = C \cdot n^{1/3}$
  - egyoldali eloszlást $\Longrightarrow b_{opt} = C \cdot n^{1/4}$

De ha tudjuk is a nagyságrendet, C=?

Blokkméret kiválasztása

Több tényezőtől is függ

- az adatgeneráló folyamat - van-e összefüggőség:
  - térbeli
  - időbeli
- a bennünket érdeklő statisztika, például
  - átlag
  - medián
  - lag(k) korreláció
- a statisztikából mit szeretnénk számolni, ha például
  - szórást $\Longrightarrow b_{opt} = C \cdot n^{1/3}$
  - egyoldali eloszlást $\Longrightarrow b_{opt} = C \cdot n^{1/4}$

De ha tudjuk is a nagyságrendet, C=?

Blokkméret kiválasztása (Politis & White)

Jel. $\mathcal{F}_{-\infty}^{0} = \{X_n : n \leq 0\}$, $\mathcal{F}_{k}^{\infty} = \{X_n : n \geq k\}$

Def.: $\{X_t$: t$\in \mathbb{Z}\}$ erősen keverő, ha $\alpha_X(k) \longrightarrow 0$ $(k \to \infty)$, ahol

$\alpha_X(k) = sup\{|P(A \cap B - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{k}^{\infty}\}$

Tétel :

Tegyük fel, hogy $E|X_t|^{6+\delta} < \infty$, $\sum_{k=1}^{\infty} k^2(\alpha_X(k))^{\frac{\delta}{6+\delta}} < \infty$

valamely δ>0-ra

Legyen $b = o(n^{1/2})$, n $\to \infty$ esetén b $\to \infty$.

Ekkor $MSE(\sigma^2_{b,\overline{X}}) = \frac{G^2}{b^2} + D\frac{b}{n} + o(b^{-2}) + o(\frac{b}{n})$

ahol D=$\frac{4}{3}g^2(0)$ és $G = \sum_{k=-\infty}^{\infty} |k|R(k)$

g(·): spektrális sűrűségfüggvény

R(·): autokovariancia függvény

Blokkméret kiválasztása (Politis & White)

Jel. $\mathcal{F}_{-\infty}^{0} = \{X_n : n \leq 0\}$, $\mathcal{F}_{k}^{\infty} = \{X_n : n \geq k\}$

Def.: $\{X_t$: t$\in \mathbb{Z}\}$ erősen keverő, ha $\alpha_X(k) \longrightarrow 0$ $(k \to \infty)$, ahol

$\alpha_X(k) = sup\{|P(A \cap B - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{k}^{\infty}\}$

Tétel :

Tegyük fel, hogy $E|X_t|^{6+\delta} < \infty$, $\sum_{k=1}^{\infty} k^2(\alpha_X(k))^{\frac{\delta}{6+\delta}} < \infty$

valamely δ>0-ra

Legyen $b = o(n^{1/2})$, n $\to \infty$ esetén b $\to \infty$.

Ekkor $MSE(\sigma^2_{b,\overline{X}}) = \frac{G^2}{b^2} + D\frac{b}{n} + o(b^{-2}) + o(\frac{b}{n})$

ahol D=$\frac{4}{3}g^2(0)$ és $G = \sum_{k=-\infty}^{\infty} |k|R(k)$

g(·): spektrális sűrűségfüggvény

R(·): autokovariancia függvény

Blokkméret kiválasztása (Politis & White)

Jel. $\mathcal{F}_{-\infty}^{0} = \{X_n : n \leq 0\}$, $\mathcal{F}_{k}^{\infty} = \{X_n : n \geq k\}$

Def.: $\{X_t$: t$\in \mathbb{Z}\}$ erősen keverő, ha $\alpha_X(k) \longrightarrow 0$ $(k \to \infty)$, ahol

$\alpha_X(k) = sup\{|P(A \cap B - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{k}^{\infty}\}$

Tétel :

Tegyük fel, hogy $E|X_t|^{6+\delta} < \infty$, $\sum_{k=1}^{\infty} k^2(\alpha_X(k))^{\frac{\delta}{6+\delta}} < \infty$

valamely δ>0-ra

Legyen $b = o(n^{1/2})$, n $\to \infty$ esetén b $\to \infty$.

Ekkor $MSE(\sigma^2_{b,\overline{X}}) = \frac{G^2}{b^2} + D\frac{b}{n} + o(b^{-2}) + o(\frac{b}{n})$

ahol D=$\frac{4}{3}g^2(0)$ és $G = \sum_{k=-\infty}^{\infty} |k|R(k)$

g(·): spektrális sűrűségfüggvény

R(·): autokovariancia függvény

Blokkméret kiválasztása (Politis & White)

Optimális blokkméret: $b_{opt} = [(\frac{2G^2}{D})n^{1/3}]$

Kérdés: hogyan becsüljük G-t és D-t

$$\hat{D} = \frac{4}{3}\hat{g}^2(0)$$

$$\hat{G} = \sum_{k=-M}^{M} \lambda(\frac{k}{M})|k|\hat{R}(k))$$

ahol $\hat{R}(k) = n^{-1} \sum_{k=1}^{n-|k|} (X_i - \overline{X}_n)(X_{i+|k|} - \overline{X}_n)$

$$\lambda(t) = \begin{cases} 1 & \text{ha } |t| \in [0, 1/2] \\ 2(1-|t|) & \text{ha } |t| \in [1/2, 1] \\ 0 & \text{különben} \end{cases}$$

$M = 2\hat{m}$, ahol $\hat{m}$: ahonnan a korrelogram "lényegében" 0

Blokkméret kiválasztása (Politis & White)

Optimális blokkméret: $b_{opt} = [(\frac{2G^2}{D})n^{1/3}]$

Kérdés: hogyan becsüljük G-t és D-t

$$\hat{D} = \frac{4}{3}\hat{g}^2(0)$$

$$\hat{G} = \sum_{k=-M}^{M} \lambda(\tfrac{k}{M})|k|\hat{R}(k))$$

ahol $\hat{R}(k) = n^{-1}\sum_{k=1}^{n-|k|}(X_i - \overline{X}_n)(X_{i+|k|} - \overline{X}_n)$

$$\lambda(t) = \begin{cases} 1 & \text{ha } |t| \in [0, 1/2] \\ 2(1-|t|) & \text{ha } |t| \in [1/2, 1] \\ 0 & \text{különben} \end{cases}$$

$M = 2\hat{m}$, ahol $\hat{m}$: ahonnan a korrelogram "lényegében" 0

Blokkméret kiválasztása (Politis & White)

Optimális blokkméret: $b_{opt} = [(\frac{2G^2}{D})n^{1/3}]$

Kérdés: hogyan becsüljük G-t és D-t

$$\hat{D} = \frac{4}{3}\hat{g}^2(0)$$

$$\hat{G} = \sum_{k=-M}^{M} \lambda(\tfrac{k}{M})|k|\hat{R}(k))$$

ahol $\hat{R}(k) = n^{-1} \sum_{k=1}^{n-|k|} (X_i - \overline{X}_n)(X_{i+|k|} - \overline{X}_n)$

$$\lambda(t) = \begin{cases} 1 & \text{ha } |t| \in [0, 1/2] \\ 2(1-|t|) & \text{ha } |t| \in [1/2, 1] \\ 0 & \text{különben} \end{cases}$$

$M = 2\hat{m}$, ahol $\hat{m}$: ahonnan a korrelogram "lényegében" 0

Sorozat összefüggőség

- Legyenek $X_1, X_2, \ldots, X_n$ egyváltozós stacionárius megfigyelések; $EX_i = \mu$, Var(X_i)=σ^2
- Ha $X_1, X_2, \ldots, X_n$ i.i.d., akkor $Var(\overline{X}) = \frac{\sigma^2}{n}$
- Sorozat összefüggőség $\longrightarrow$ nagyobb variancia
- Effektív mintaméret (n_e):

$$n_e = \frac{\sigma^2}{Var^*(\overline{X})}$$

ahol $Var^*(\overline{X})$: becsült variancia $\longleftarrow$ bootstrap

Sorozat összefüggőség

- Legyenek $X_1, X_2, \ldots, X_n$ egyváltozós stacionárius megfigyelések; $EX_i = \mu$, Var(X_i)=σ^2
- Ha $X_1, X_2, \ldots, X_n$ i.i.d., akkor $Var(\overline{X}) = \frac{\sigma^2}{n}$
- Sorozat összefüggőség $\longrightarrow$ nagyobb variancia
- Effektív mintaméret (n_e):

$$n_e = \frac{\sigma^2}{Var^*(\overline{X})}$$

ahol $Var^*(\overline{X})$: becsült variancia $\longleftarrow$ bootstrap

Sorozat összefüggőség

- Legyenek $X_1, X_2, \ldots, X_n$ egyváltozós stacionárius megfigyelések; $EX_i = \mu$, Var(X_i)=σ^2
- Ha $X_1, X_2, \ldots, X_n$ i.i.d., akkor $Var(\overline{X}) = \frac{\sigma^2}{n}$
- Sorozat összefüggőség $\longrightarrow$ nagyobb variancia
- Effektív mintaméret (n_e):

$$n_e = \frac{\sigma^2}{Var^*(\overline{X})}$$

ahol $Var^*(\overline{X})$: becsült variancia $\longleftarrow$ bootstrap

Sorozat összefüggőség

- Legyenek $X_1, X_2, \ldots, X_n$ egyváltozós stacionárius megfigyelések; $EX_i = \mu$, Var(X_i)=σ^2
- Ha $X_1, X_2, \ldots, X_n$ i.i.d., akkor $Var(\overline{X}) = \frac{\sigma^2}{n}$
- Sorozat összefüggőség $\longrightarrow$ nagyobb variancia
- Effektív mintaméret (n_e):

$$n_e = \frac{\sigma^2}{Var^*(\overline{X})}$$

ahol $Var^*(\overline{X})$: becsült variancia $\longleftarrow$ bootstrap

Politis & White algoritmusával

- n=2591 megfigyelés (szélsebességi adatok heti maximumai) 5 német településre
- Automatikus blokkméret-kiválasztó algoritmus eredménye:

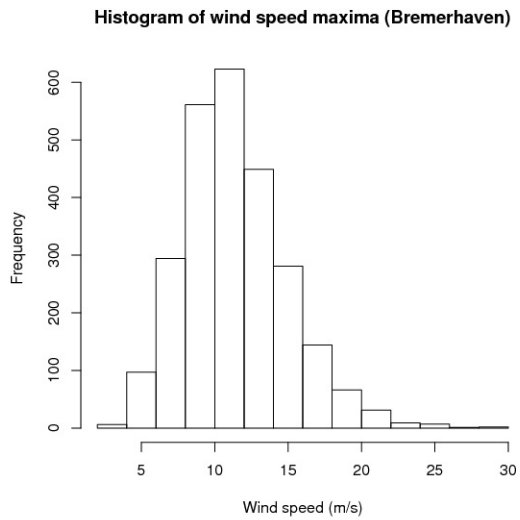
Település	Optimális blokkméret
Hamburg	**31**
Hannover	11
Bremerhaven	**28**
Fehmarn	**31**
Schleswig	15

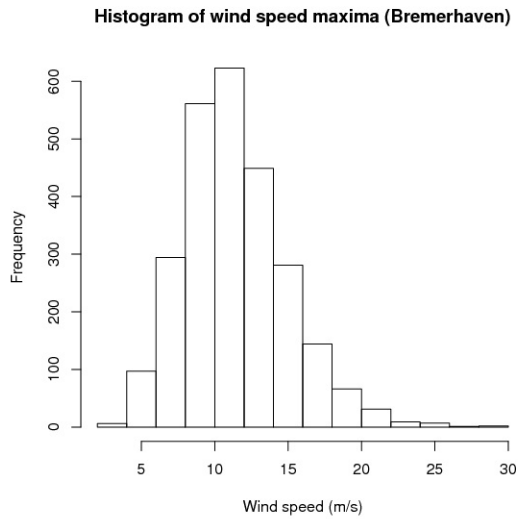
- 30 hét $\approx$ 0,58 év $\longrightarrow$ meteorológiailag értelmetlen

Egy másik modell

Módszer:

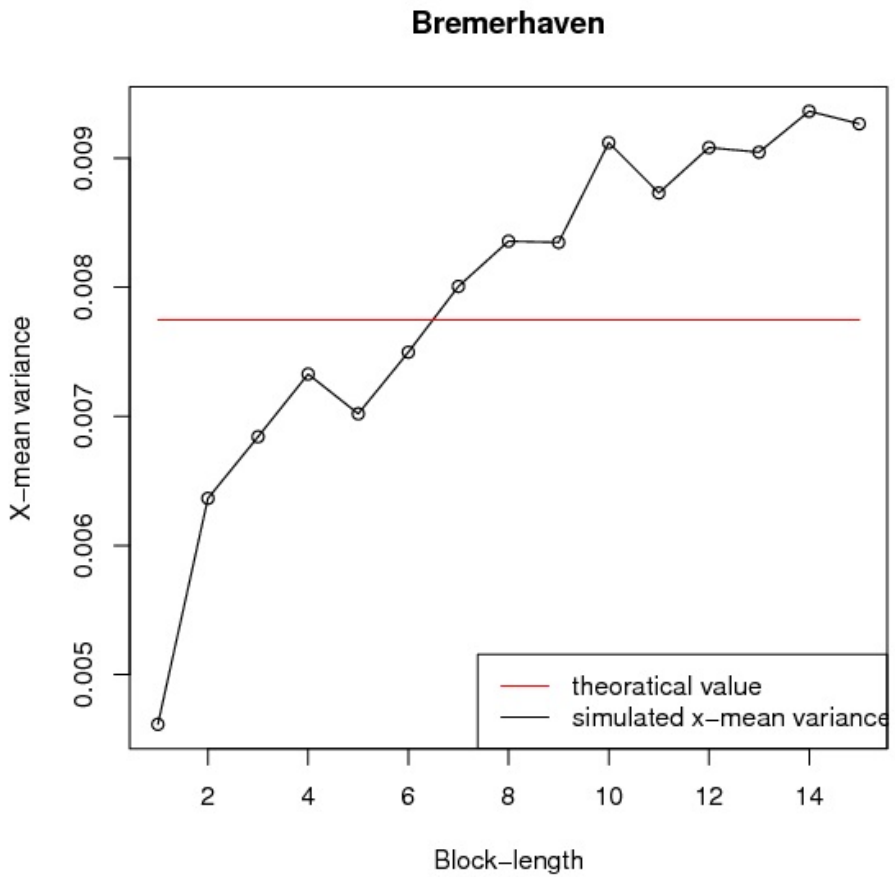
1. AR(1) modell illesztése az adatokra: $X_t = \mu + \phi X_{t-1} + Z_t$, $Z_t \sim$ extrém érték eloszlás
2. Az elméleti $Var(\overline{X}_n)$ kiszámolása az AR(1) paraméterekből:
$$Var(\overline{X}_n) = \frac{\hat{\sigma}^2}{n(1-\hat{\phi}^2)} \frac{2\hat{\phi}^{n+1} - n\hat{\phi}^2 - 2\hat{\phi} + n}{n(\hat{\phi}-1)^2}$$
3. b^* optimális blokkméret: ahol az átlag szimulált varianciája először elmetszi az elméleti értéket

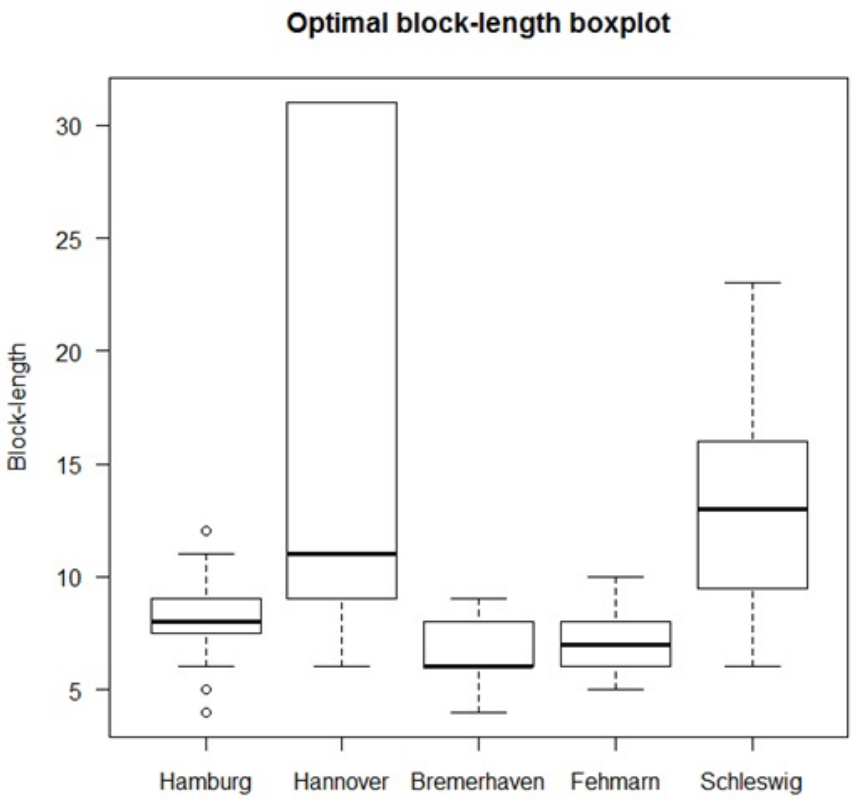




Alkalmazás szélsebességi adatokra

A bootstrap szimuláció eredményei





$b^* = 6$

Alkalmazás szélsebességi adatokra

Bootstrap szimulációs eredmények:

Település	Optimális blokkméret	Bootstrap $Var(\overline{X})$	Elméleti $Var(\overline{X})$	i.i.d. $Var(\overline{X})$	Mintaméret csökkenése
Hamburg	8-9	0,0038	0,0034	0,0020	1,85
Hannover	7	0,0067	0,0071	0,0042	1,59
Bremerhaven	6	0,0073	0,0077	0,0043	1,71
Fehmarn	7	0,0035	0,0034	0,0020	1,74
Schleswig	13	0,0037	0,0030	0,0018	2,09

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig

     $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig

     $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X^*_{-u}, \ldots, X^*_{-u+\hat{p}-1}) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X^*_{t-j} - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig

     $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig

     $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

AR-sieve bootstrap

- Feltétel: a folyamat stacionárius és jól becsülhető AR(p) modellel:

$$X_t - \mu_X = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu_X) + \varepsilon_t, \quad t \in \mathbb{Z}$$

ahol $\mu_X = EX_t$

$(\varepsilon_t)_{t\in\mathbb{Z}}$ i.i.d., E(ε_t)=0 és ε_t független $\{ X_s; s < t \}$-től

- Paraméterek és hibák becslése:
  1. $\hat{p}$=? $\longrightarrow$ AIC
  2. $\hat{\mu}_X = n^{-1} \sum_{t=1}^{n} X_t$
  3. $\hat{\phi}_1, \ldots, \hat{\phi}_{\hat{p}}$=? $\longrightarrow$ Yule-Walker módszer
  4. $R_t = X_t - \sum_{j=1}^{\hat{p}} \hat{\phi}_j X_{t-j}$, ahol $t = \hat{p}+1, \ldots, n$ ebből pedig

     $\hat{\varepsilon}_t = R_t - \overline{R}_t$, ahol $t = \hat{p}+1, \ldots, n$
- Bootstrap minta konstruálásának lépései:
  - ε_t^*: véletlen elem $\{ \hat{\varepsilon}_{\hat{p}+1}, \ldots, \hat{\varepsilon}_n \}$ halmazból
  - Nagy u-ra $(X_{-u}^*, \ldots, X_{-u+\hat{p}-1}^*) = (\hat{\mu}_X, \ldots, \hat{\mu}_X)$ (a folyamat indítása)
  - $X_t^* = \mu_X + \sum_{j=1}^{p} \phi_j (X_{t-j}^* - \mu_X) + \varepsilon_t^* \quad t \in \mathbb{Z}$
- Ebből a bootstrap minta: $\{ X_1^*, \ldots, X_n^* \}$

Lag(1) korreláció becslése

3 módszert próbáltam ki:

1. "Naiv" (sima) blokk bootstrap (b=10)
2. Dupla blokk bootstrap ($b_1 = 2$ és b_2=10)
3. AR sieve (szita) bootstrap

Futtatási eredmények lag(1) korrelációra:

Módszer	Hamburg	Hannover	Bremerhaven	Fehmarn	Schleswig
Mintából	0,2755	0,2528	0,2901	0,2548	0,2588
Naiv blokk bootstrap	0,2485	0,2277	0,2598	0,2281	0,2306
Dupla blokk bootstrap	0,2755	0,252	0,2907	0,255	0,2565
AR-sieve bootstrap	0,274	0,2514	0,2887	0,2531	0,2588

Lag(1) korreláció becslése

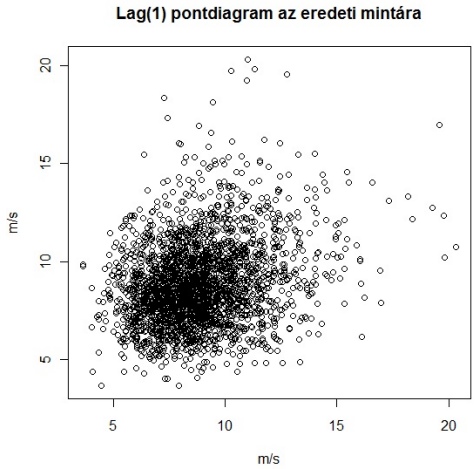
3 módszert próbáltam ki:

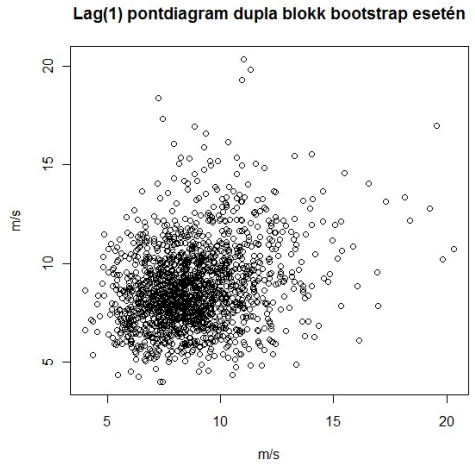
1. "Naiv" (sima) blokk bootstrap (b=10)
2. Dupla blokk bootstrap ($b_1 = 2$ és b_2=10)
3. AR sieve (szita) bootstrap

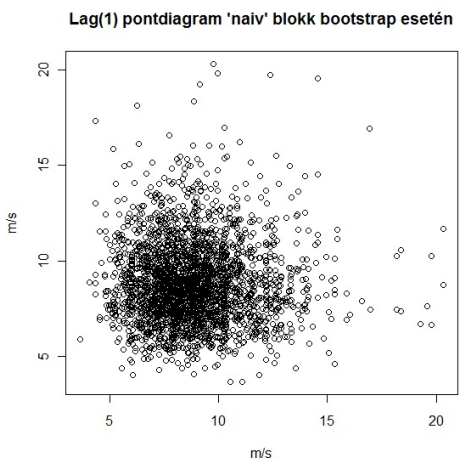
Futtatási eredmények lag(1) korrelációra:

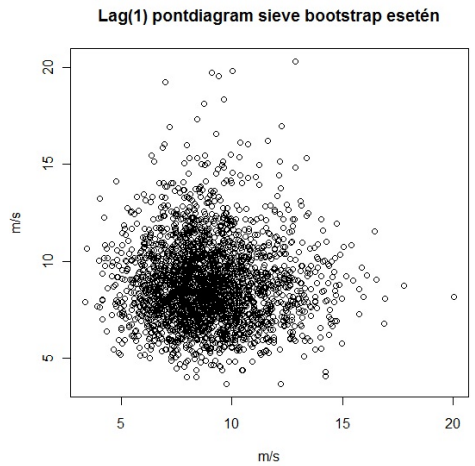
Módszer	Hamburg	Hannover	Bremerhaven	Fehmarn	Schleswig
Mintából	0,2755	0,2528	0,2901	0,2548	0,2588
Naiv blokk bootstrap	0,2485	0,2277	0,2598	0,2281	0,2306
Dupla blokk bootstrap	0,2755	0,252	0,2907	0,255	0,2565
AR-sieve bootstrap	0,274	0,2514	0,2887	0,2531	0,2588

Lag(1) pontdiagramok Schleswig szélsebességi adataira









Goodness-of-fit (GoF) tesztek

- A modell paramétereinek becslése
- Goodness-of-fit teszt: H_0: $C \in \mathbb{C}_0 = \{ C_\theta, \theta \in \Theta \}$
  - Cramér-von Mises tesztstatisztika:

    $$T = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \Phi(x) dF(x)$$

    - F: eloszlásfv.
    - F_n: empirikus eloszlásfv.
    - Φ: súlyfv. – Anderson-Darling: $\Phi(x) = \frac{1}{F(x)(1-F(x))}$
  - Kritikus érték - szimuláció
    1. Minta szimulálása a $\mathbb{C}_0$ copula modellből H_0 esetén
    2. $\hat{\theta}$ újbóli becslése ML-módszerrel
    3. A tesztstatisztika kiszámítása

       Ismétlések, p-érték becslése

Több dimenzióban:

- Probability integral transformation (PIT) - összehúzás a d-dimenziós egységkockára
- Kendall-transzformáció (*K-függvény*):
  $K(\theta, t) = P(C_\theta(F_1(X_1), \ldots, F_d(X_d)) \leq t)$

Goodness-of-fit (GoF) tesztek

- A modell paramétereinek becslése
- Goodness-of-fit teszt: H_0: $C \in \mathbb{C}_0 = \{ C_\theta, \theta \in \Theta \}$
  - Cramér-von Mises tesztstatisztika:
    $$T = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \Phi(x) dF(x)$$
    - F: eloszlásfv.
    - F_n: empirikus eloszlásfv.
    - Φ: súlyfv. – Anderson-Darling: $\Phi(x) = \frac{1}{F(x)(1-F(x))}$
  - Kritikus érték - szimuláció
    1. Minta szimulálása a $\mathbb{C}_0$ copula modellből H_0 esetén
    2. $\hat{\theta}$ újbóli becslése ML-módszerrel
    3. A tesztstatisztika kiszámítása

       Ismétlések, p-érték becslése

Több dimenzióban:

- Probability integral transformation (PIT) - összehúzás a d-dimenziós egységkockára
- Kendall-transzformáció (*K-függvény*):
  $K(\theta, t) = P(C_\theta(F_1(X_1), \ldots, F_d(X_d)) \leq t)$

Goodness-of-fit (GoF) tesztek

- A modell paramétereinek becslése
- Goodness-of-fit teszt: H_0: $C \in \mathbb{C}_0 = \{ C_\theta, \theta \in \Theta \}$
  - Cramér-von Mises tesztstatisztika:
    $$T = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \Phi(x) dF(x)$$
    - F: eloszlásfv.
    - F_n: empirikus eloszlásfv.
    - Φ: súlyfv. – Anderson-Darling: $\Phi(x) = \frac{1}{F(x)(1-F(x))}$
  - Kritikus érték - szimuláció
    1. Minta szimulálása a $\mathbb{C}_0$ copula modellből H_0 esetén
    2. $\hat{\theta}$ újbóli becslése ML-módszerrel
    3. A tesztstatisztika kiszámítása

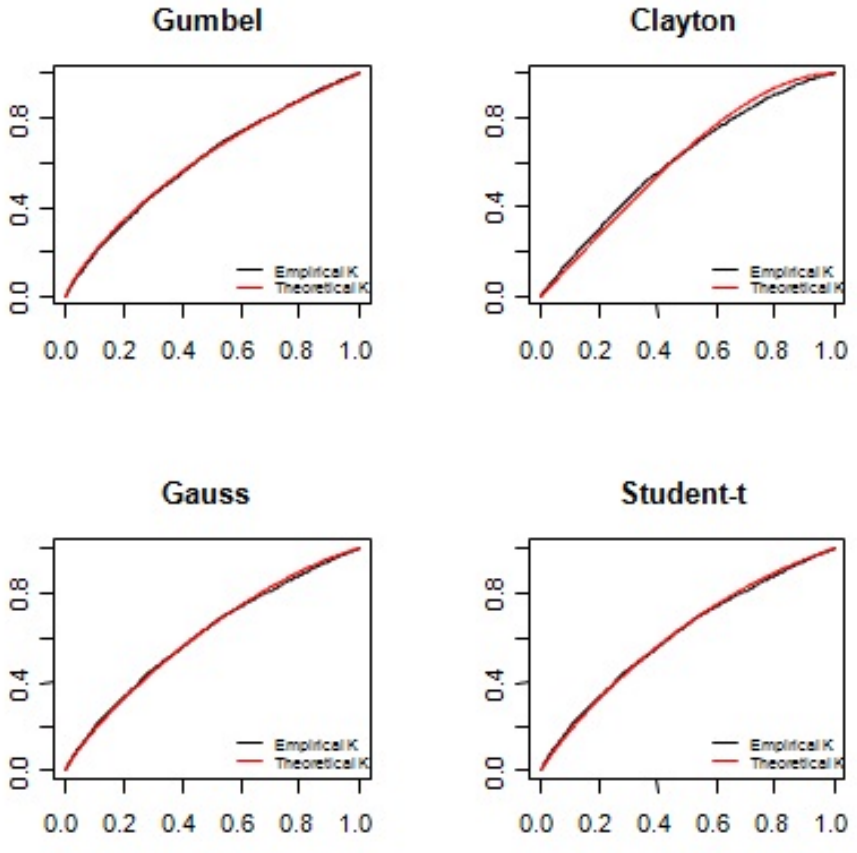
       Ismétlések, p-érték becslése

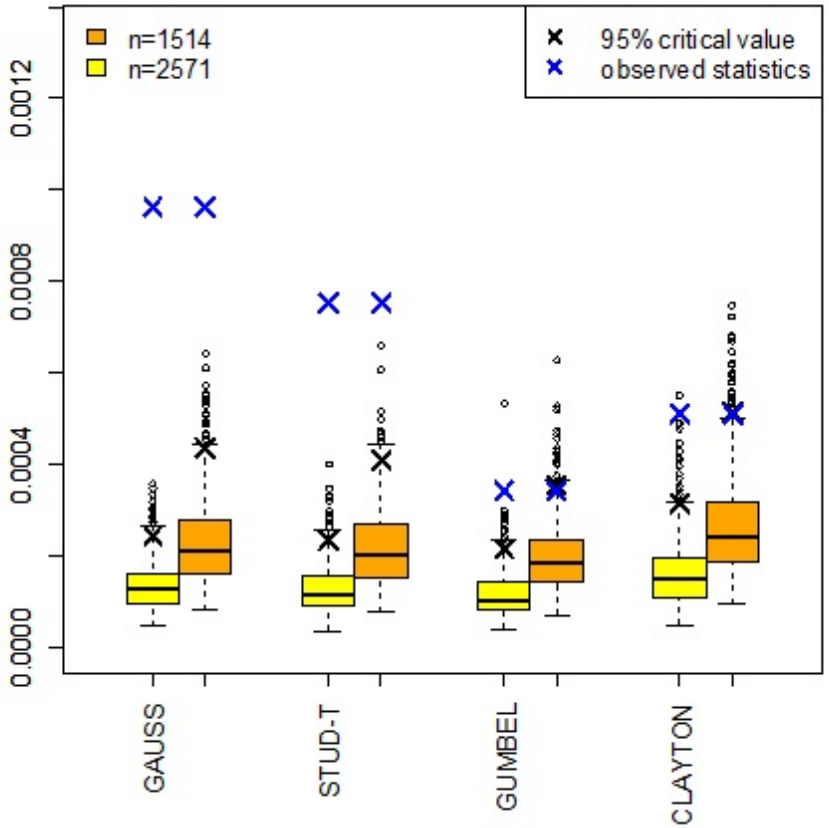
Több dimenzióban:

- Probability integral transformation (PIT) - összehúzás a d-dimenziós egységkockára
- Kendall-transzformáció (*K-függvény*):
  $K(\theta, t) = P(C_\theta(F_1(X_1), \ldots, F_d(X_d)) \leq t)$

Alkalmazás szélsebességi adatokra

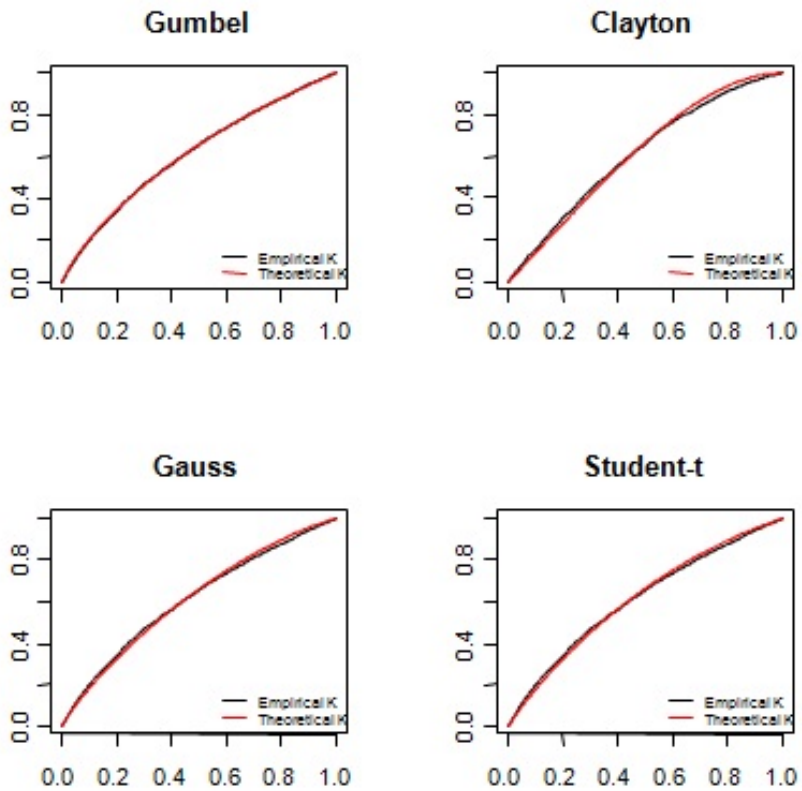
Bremerhaven & Fehmarn

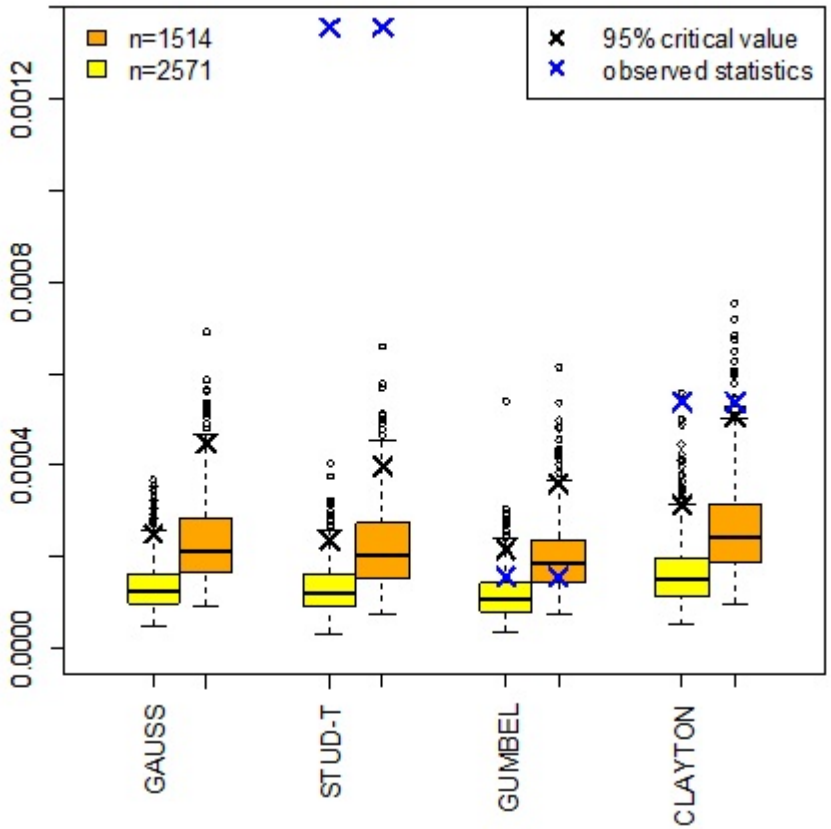




Alkalmazás szélsebességi adatokra

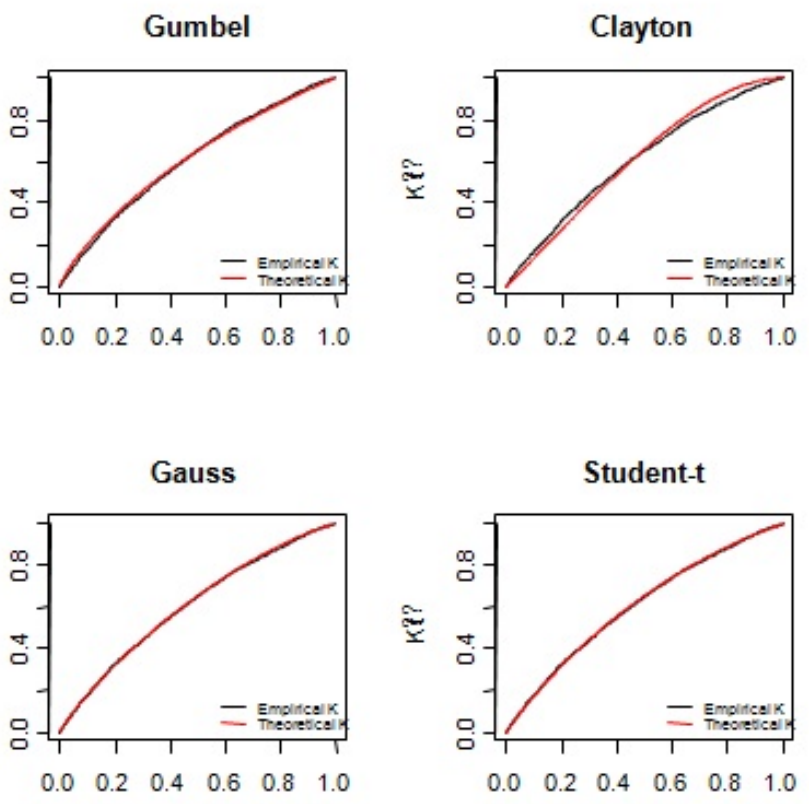
Bremerhaven & Schleswig

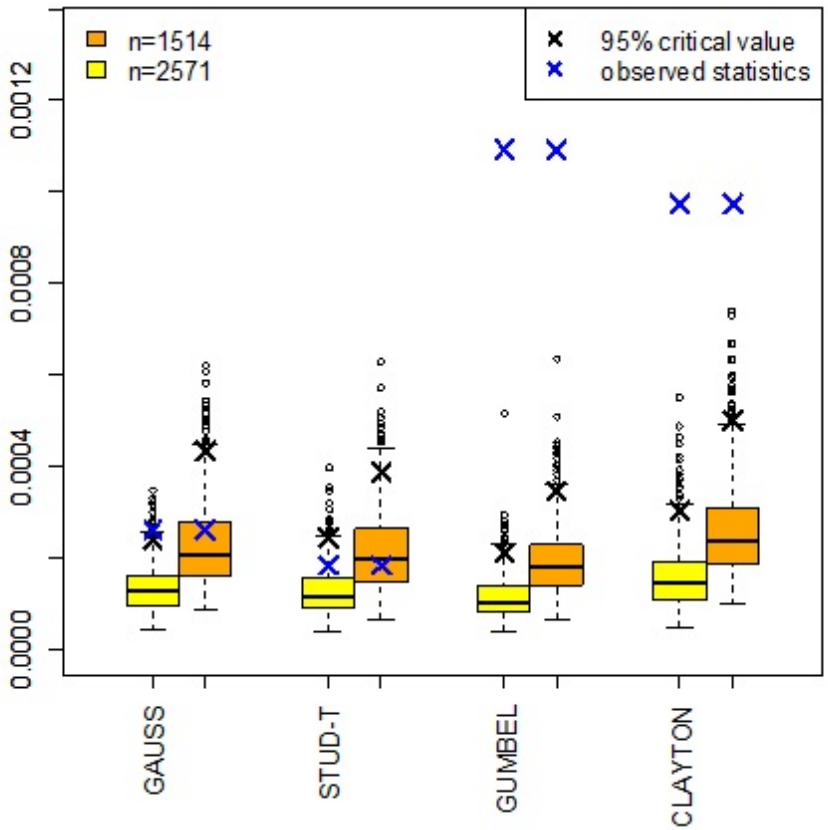




Alkalmazás szélsebességi adatokra

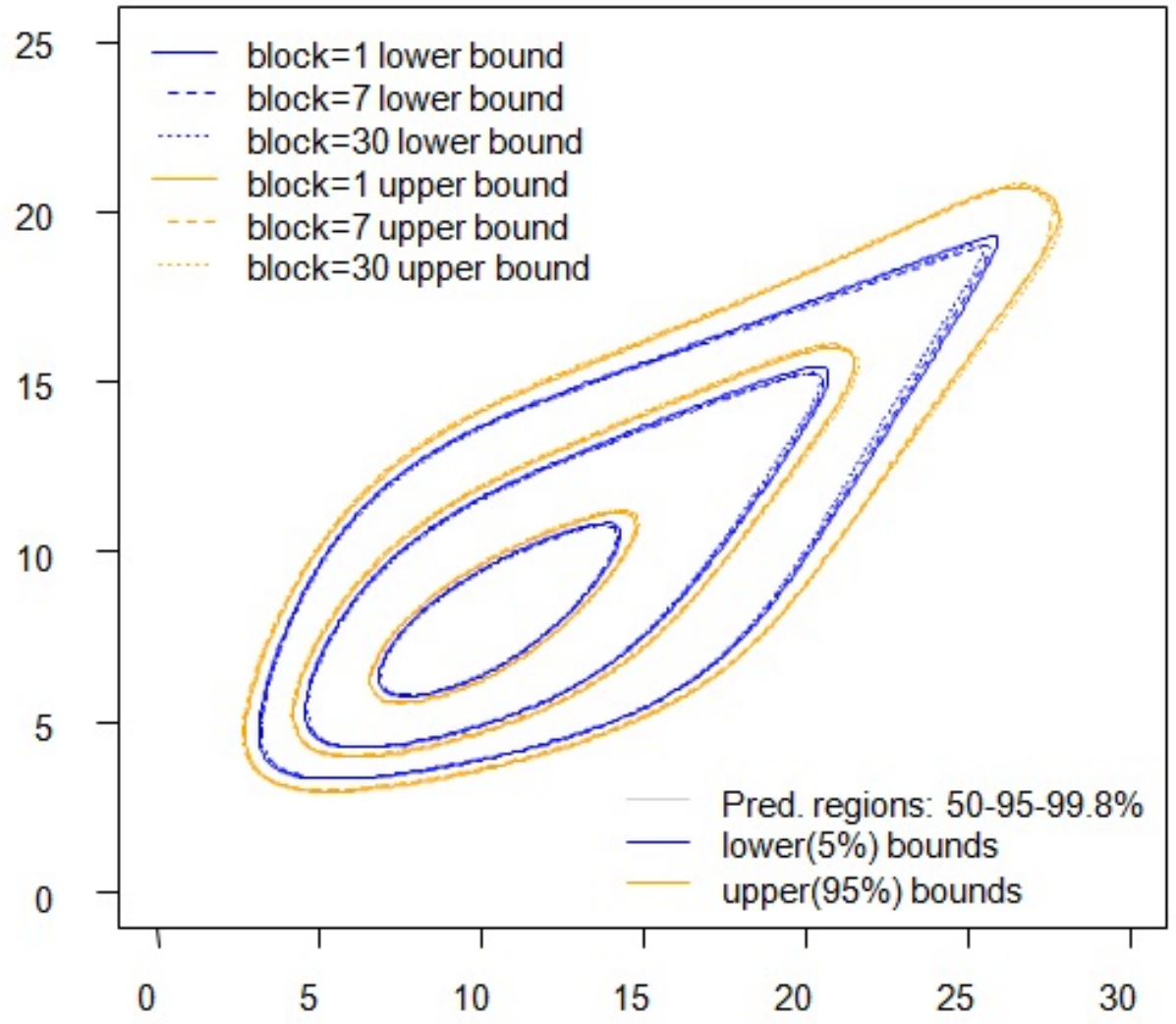
Fehmarn & Schleswig





Alkalmazás szélsebességi adatokra

Előrejelzési régiók (Bremerhaven & Fehmarn)



Összefoglalás

- A bootstrap módszer kiválasztásánál figyelembe kell venni:
  - milyen az adatok struktúrája
  - mi a bennünket érdeklő statisztika
  - mire akarjuk felhasználni
- nincs feltétlen legjobb módszer - egy (fél)paraméteres vagy egy paraméter nélküli is jó lehet

Köszönetnyilvánítás

- Zempléni András
- Rakonczai Pál

Irodalomjegyzék I


S.N. Lahiri:
Resampling Methods for Dependent Data.
Springer, 2003.


L. Kish:
Survey Sampling.
J. Wiley, 1965.


P. Rakonczai, L. Varga, A. Zempléni:
Copula Fitting to Autocorrelated Data, with Applications to Wind Speed Modelling.
Working paper, november 11., 2010.


P. Rakonczai, A. Zempléni:
Copulas and goodness of fit tests. Recent advances in stochastic modeling and data analysis.
World Scientific, pp. 198-206, 2007.

Irodalomjegyzék II

P. Bühlmann:
Bootstraps for Time Series.
Statistical Science, Vol. 17, pp. 52-72, 2002.

P. Bühlmann, H.R. Künsch:
Block length selection in the bootstrap for time series.
Computational Statistics Data Analysis, pp. 295-310, 1999.

D.N. Politis, H. White:
Automatic Block-Length Selection for the Dependent Bootstrap.
Econometric Reviews, Vol. 23, pp. 53-70, 2004. 2000.

C. Genest, B. Rémillard:
Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models.
Annales de IInstitut Henri Poincaré - Probabilités et Statistiques, Vol. 44, No. 6, 10961127, 2008. 2000.