Change point

Erik Mendroš, Marek Bedřich

PMSE - Matematická statistika
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

28.3.2024

Overview

1. Definitions and Notation

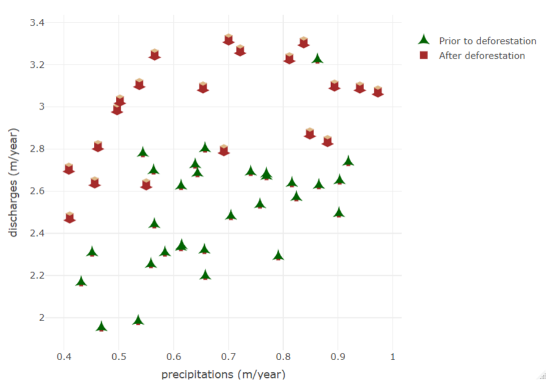
2. Statistical Testing

3. Permutation Test Procedures

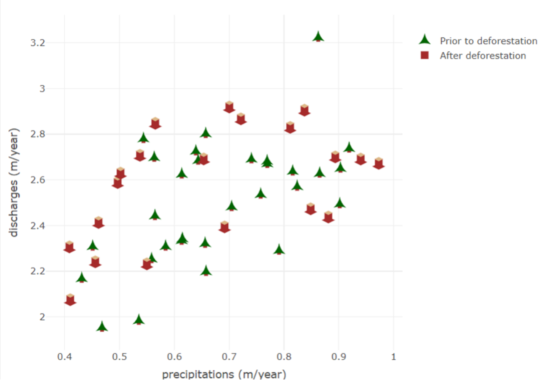
4. Simulations

Malá Ráztoka

Change?



No change?



The model

Modified regression model

By modified regression model we will understand model

$$Y_i = \mathbf{x}_i^T \beta + \mathbf{x}_i^T \delta \cdot \mathbb{1}\{i > m\} + \varepsilon_i, \quad i = 1, ..., n$$

where $m \leq n$, $\beta = (\beta_1, ..., \beta_p)$, $\delta = (\delta_1, ..., \delta_p) \neq \mathbf{0}$ and $\varepsilon_1, \ldots, \varepsilon_n$ are iid random errors with zero mean, nonzero variance σ^2 and finite moment $\mathrm{E}\left[|\varepsilon_i|^{2+\Delta}\right]$ with some $\Delta > 0$.

Hypothesis for the **change point** parameter m:

$$H_0 : m = n \ \text{(no change)} \qquad \text{against} \qquad H_1 : m < n$$

Important formulas

Partial sums

$$\mathbf{S}_k = \sum_{i=1}^{k} \mathbf{x}_i \left(Y_i - \mathbf{x}_i^T \widehat{\beta}_n \right) = \sum_{i=1}^{k} \mathbf{x}_i u_i, \quad k = 1, \dots, n$$

$$S_k^* = \sum_{i=1}^{k} \left(Y_i - \mathbf{x}_i^T \widehat{\beta}_n \right) = \sum_{i=1}^{k} u_i, \quad k = 1, \dots, n$$

where

$$\widehat{\beta}_n = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \quad u_i = Y_i - \mathbf{x}_i^T \widehat{\beta}_n.$$

is the LSE of β in the modified regression model with $m = n$ i.e. no change.

Notation

Let us denote the **partial regression matrices** by

$$\mathbf{X}_k = \begin{pmatrix} \mathbf{x}_1^T \\ \hline \vdots \\ \hline \mathbf{x}_k^T \end{pmatrix}, \quad \mathbf{X}_k^o = \begin{pmatrix} \mathbf{x}_{k+1}^T \\ \hline \vdots \\ \hline \mathbf{x}_n^T \end{pmatrix}.$$

Clearly $\mathbf{X}_n = \mathbf{X}$.

Test statistics

Statistic based on $\mathbf{S}_k$:

$$T_n = \frac{1}{\hat{\sigma}_n^2} \max_{p<k<n-p} \left\{ \mathbf{S}_k^T \left(\mathbf{X}_k{}^T \mathbf{X}_k \right)^{-1} \left(\mathbf{X}^T \mathbf{X} \right) \left(\mathbf{X}_k^o{}^T \mathbf{X}_k^o \right)^{-1} \mathbf{S}_k \right\}$$

Statistic based on S_k^*:

$$T_n^* = \max_{1\leq k<n} \left\{ \sqrt{\frac{n}{k(n-k)}} \cdot \frac{|S_k^*|}{\hat{\sigma}_n} \right\}$$

We require

$$\hat{\sigma}_n^2 - \sigma^2 = o_p \left(\frac{1}{\sqrt{\log \log n}} \right) \text{ as } n \to \infty.$$

Estimator of variance

The condition

$$\hat{\sigma}_n^2 - \sigma^2 = o_p\left(\frac{1}{\sqrt{\log\log n}}\right) \text{ as } n \to \infty$$

is satisfied by e.g.

$$\hat{\sigma}_n^2 = \frac{1}{n-p} \min_{p<k<n-p} \left\{ \sum_{i=1}^{k} (Y_i - \mathbf{x}_i^T \widehat{\beta}_k)^2 + \sum_{i=k+1}^{n} (Y_i - \mathbf{x}_i^T \widehat{\beta}_k^0)^2 \right\},$$

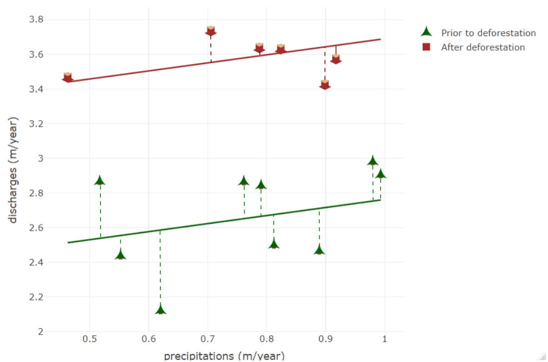
where $\widehat{\beta}_k$ and $\widehat{\beta}_k^0$ are the LSE based on $Y_1, \dots, Y_k$ and $Y_{k+1}, \dots, Y_n$, respectively. It can be shown that $\hat{\sigma}_n^2$ can be rewritten as

$$\hat{\sigma}_n^2 = \frac{1}{n-p} \left\{ \sum_{i=1}^{n} u_i^2 - \max_{p<k<n-p} \left\{ \mathbf{S}_k^T \left(\mathbf{X}_k{}^T \mathbf{X}_k\right)^{-1} \left(\mathbf{X}^T \mathbf{X}\right) \left(\mathbf{X}_k^{o\,T} \mathbf{X}_k^o\right)^{-1} \mathbf{S}_k \right\} \right\}$$

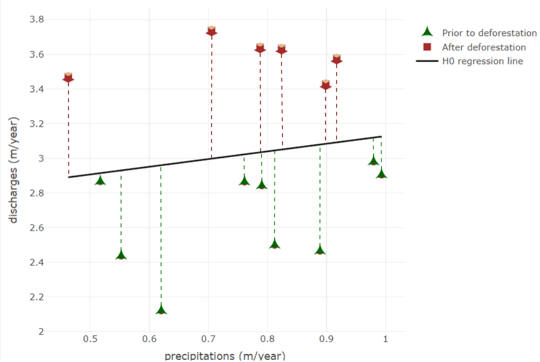
Interpretation of T_n

Large values of T_n speaks against H_0.

Estimated variance based on $\widehat{\sigma}^2$

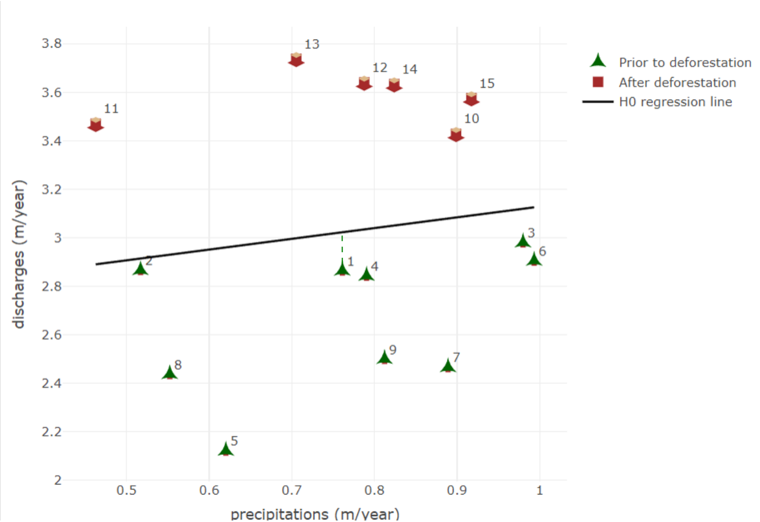


Estimated variance based on $\tilde{\sigma}^2$



Interpretation of T_n^*

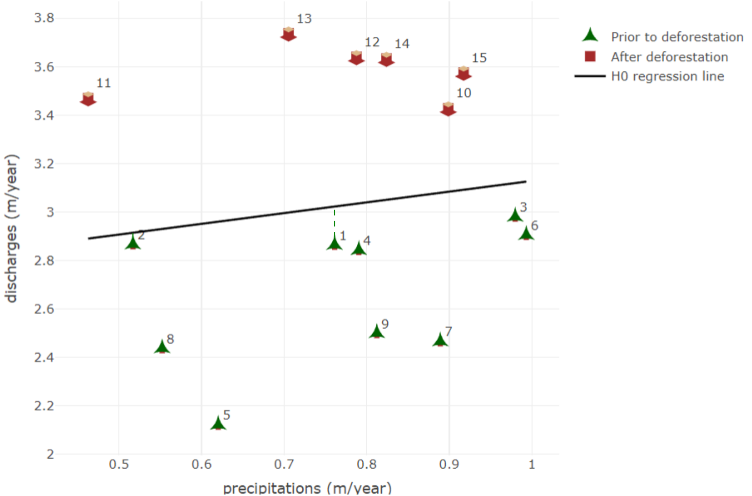
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$

Interpretation of T_n^*

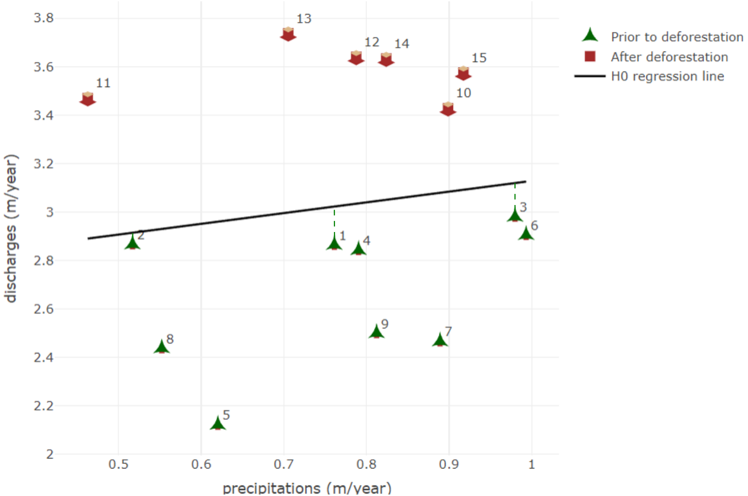
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$

Interpretation of T_n^*

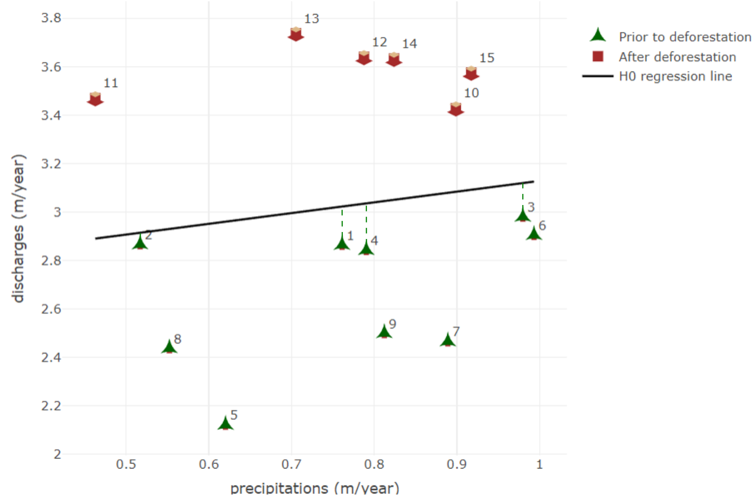
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$

Interpretation of T_n^*

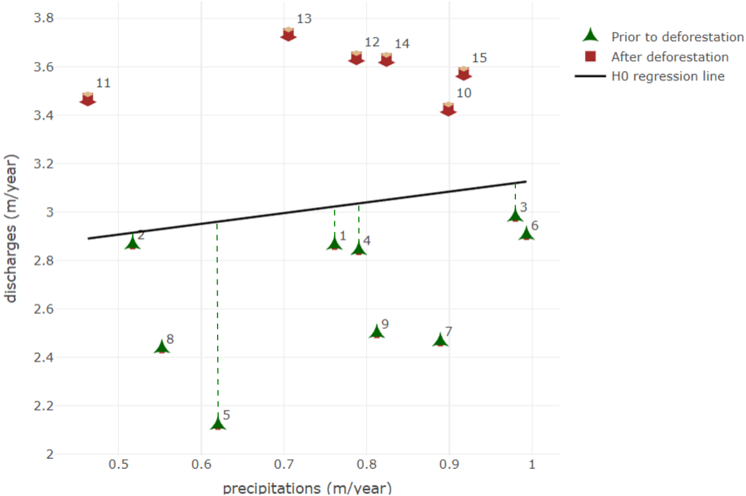
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$
$$\vdots$$

Interpretation of T_n^*

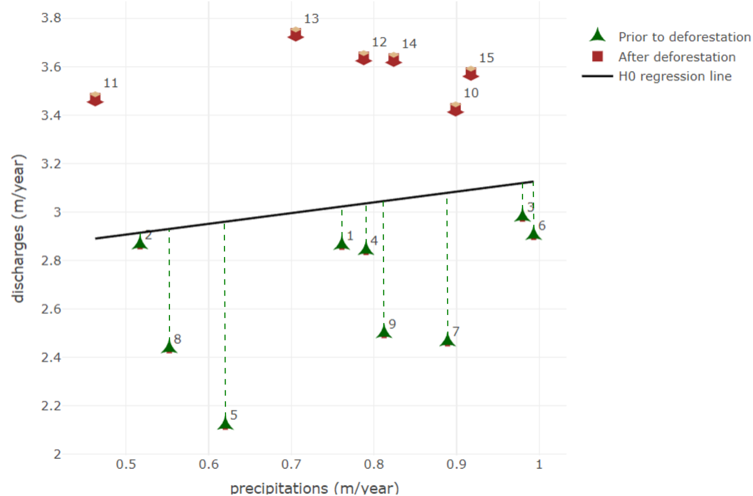
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$
$$\vdots$$

Interpretation of T_n^*

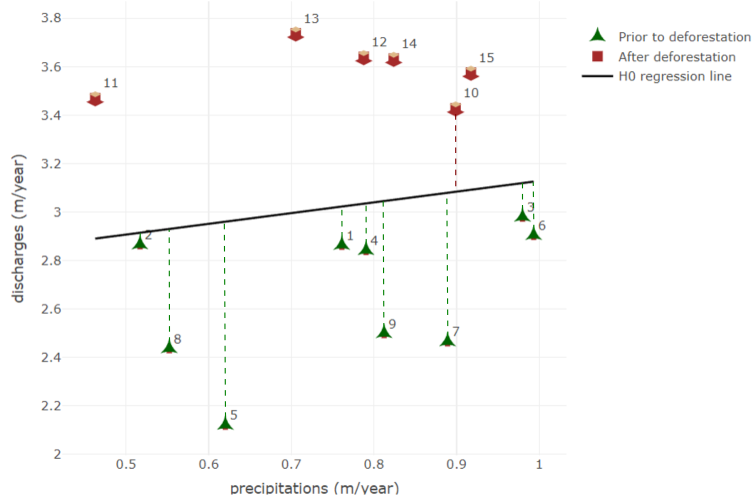
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$
$$\vdots$$

Interpretation of T_n^*

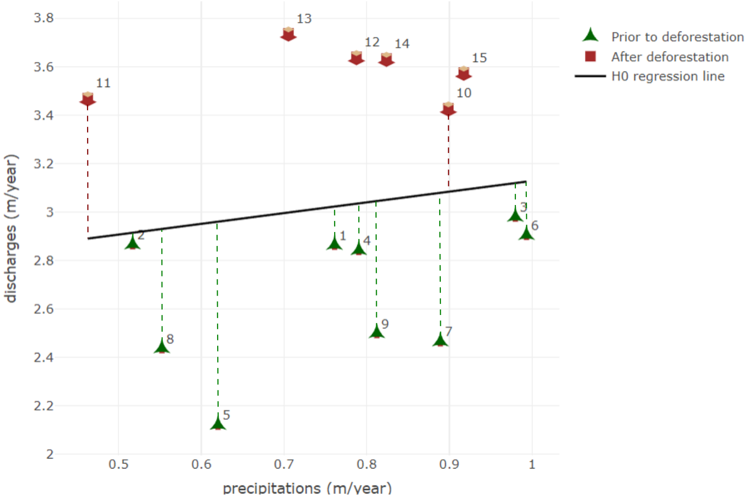
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$
$$\vdots$$

Interpretation of T_n^*

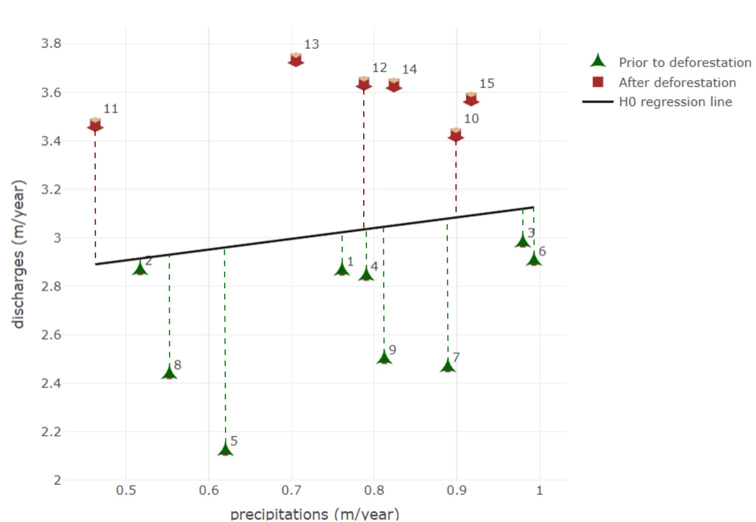
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$
$$\vdots$$

Interpretation of T_n^*

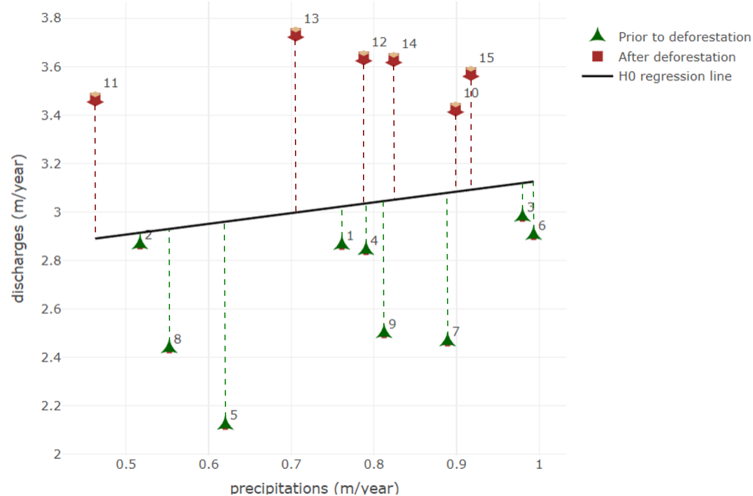
Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$
$$S_2^* = -0.18 + (-0.04)$$
$$S_3^* = -0.18 + (-0.04) + (-0.12)$$
$$\vdots$$

Interpretation of T_n^*

Large values of T_n^* speaks against H_0. Clearly, H_0 is violated on the figure below.



$$S_1^* = -0.18$$

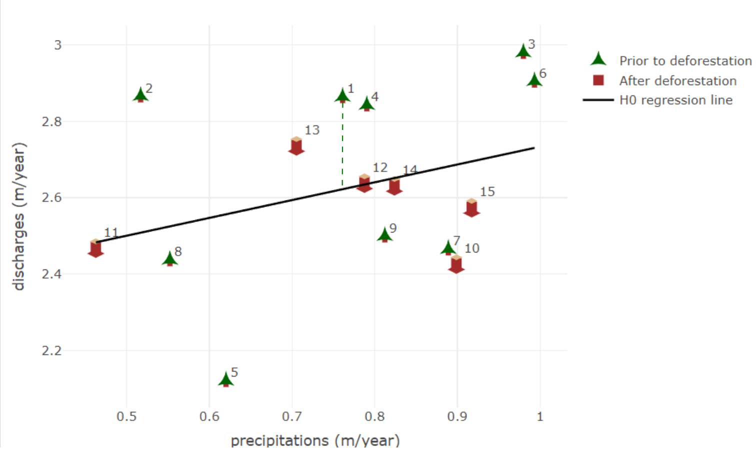
$$S_2^* = -0.18 + (-0.04)$$

$$S_3^* = -0.18 + (-0.04) + (-0.12)$$

$$\vdots$$

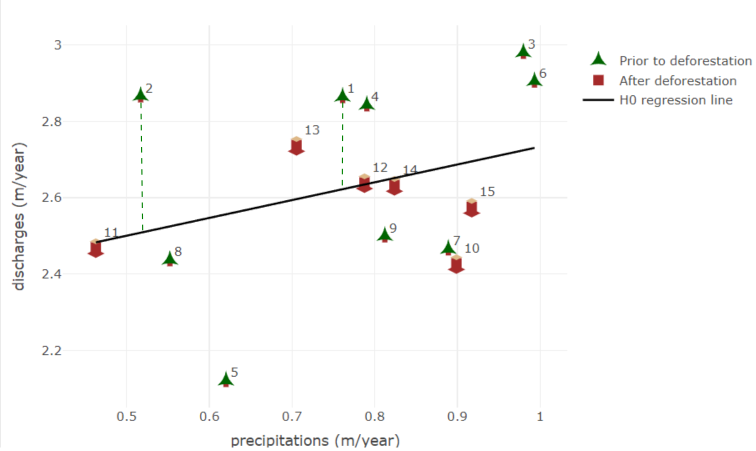
Interpretation of T_n^*

More probable case under H_0.



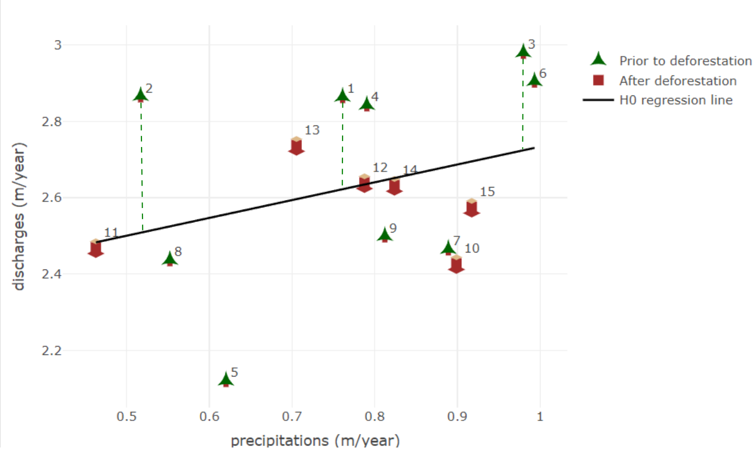
Interpretation of T_n^*

More probable case under H_0.



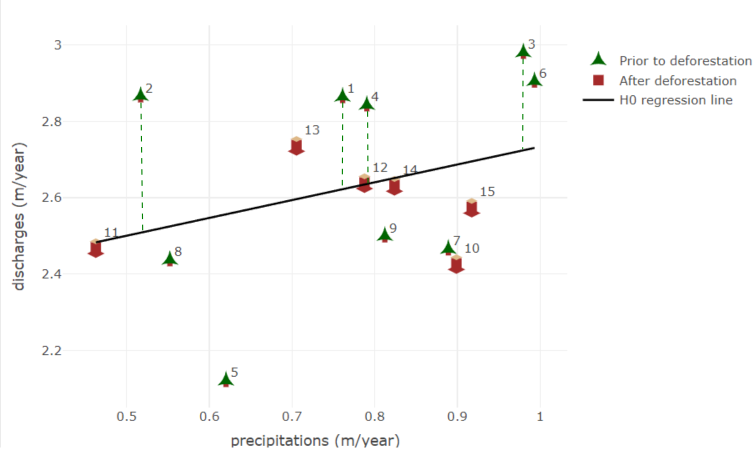
Interpretation of T_n^*

More probable case under H_0.



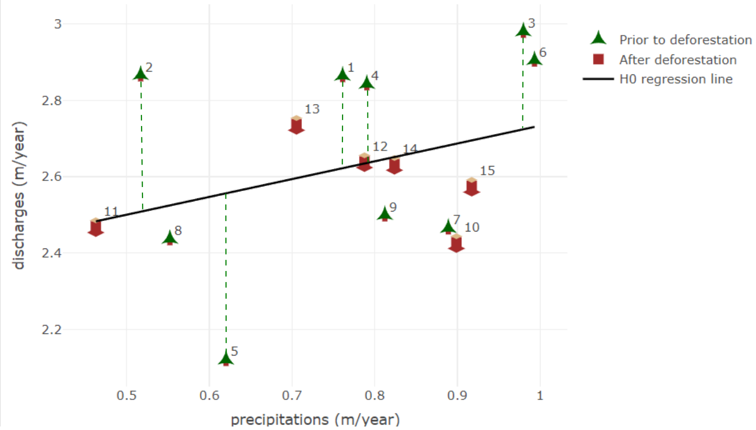
Interpretation of T_n^*

More probable case under H_0.



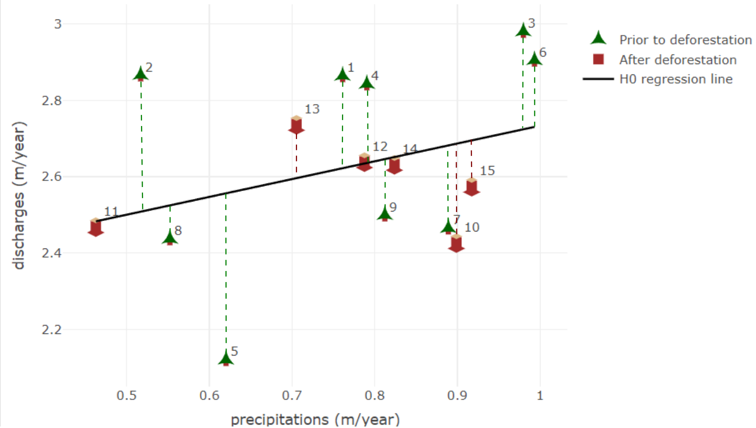
Interpretation of T_n^*

More probable case under H_0.



Interpretation of T_n^*

More probable case under H_0.



Modified test statistics

Let $q(\cdot) : [0,1] \to \mathbb{R}^+$ be a positive weight function.
Statistic based on $\mathbf{S}_k$:

$$T_n(q) = \sup_{0<t<1} \left\{ q^{-2}(t)\hat{\sigma}_n^{-2} \mathbf{S}^T_{\lfloor (n+1)t \rfloor n} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{S}_{\lfloor (n+1)t \rfloor n} \right\}$$

Statistic based on S_k^*:

$$T_n^*(q) = \sup_{0<t<1} \left\{ \frac{\left| S^*_{\lfloor (n+1)t \rfloor n} \right|}{\sqrt{n} q(t) \hat{\sigma}_n} \right\}$$

We will not focus on those.

Assumptions

- A.1 - Intercept is included in the model and the covariates are centered.

$$x_{i1} = 1,\ i = 1, \dots, n \quad \text{and} \quad \sum_{i=1}^{n} x_{ij} = 0,\ j = 2, \dots, p.$$

- A.2 - There exists a positive definite $p \times p$ matrix $\mathbf{C}$ such that for any sequence $\{\ell_n\}, \lim_{n\to\infty} \ell_n = \infty, \ell_n \leq n$, it holds that

$$\left\| \frac{1}{\ell_n} (\mathbf{X}_{k+\ell_n}^T \mathbf{X}_{k+\ell_n} - \mathbf{X}_k^T \mathbf{X}_k) - \mathbf{C} \right\|_2 = o\left(\frac{1}{\log \ell_n} \right)$$

  uniformly for $1 \leq k \leq n - \ell_n$.

Assumptions

- A.3 - It holds as $n \to \infty$, that

$$\max_{1 \le k < n} \left(\frac{1}{k} \sum_{i=1}^{k} \|\mathbf{x}_i\|^4 + \frac{1}{n-k} \sum_{i=k+1}^{n} \|\mathbf{x}_i\|^4 \right) = O(1).$$

The condition A.3 is implied by a more interpretable condition

$$\max_{1 \le i < n} \|\mathbf{x}_i\|^4 = O(1)$$

Limit Theorem

Asymptotic distribution of T_n^* and T_n

Let assumptions A.1 – A.3 be satisfied and H_0 hold. Then

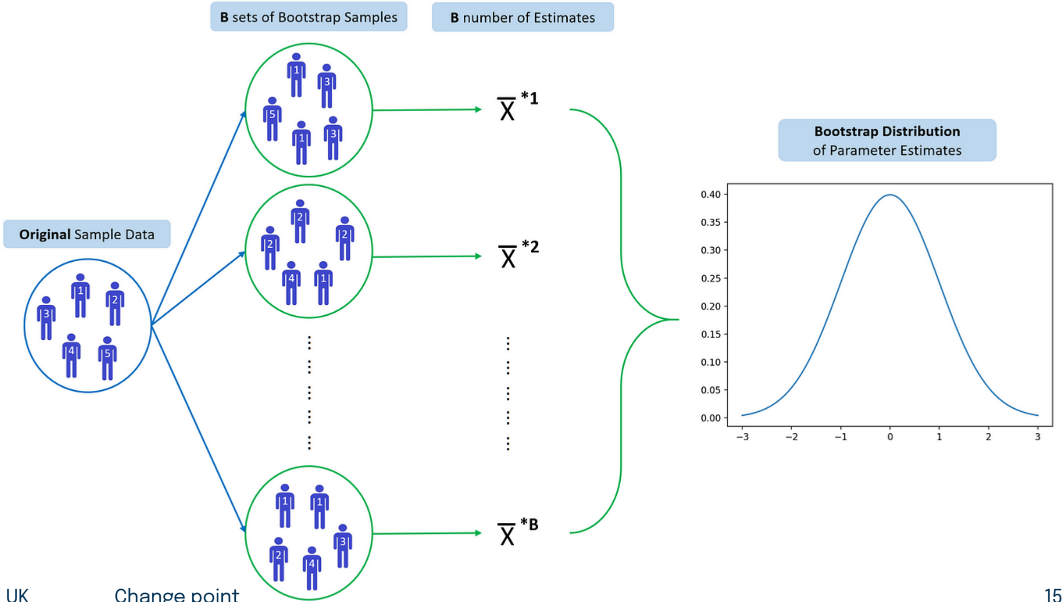
$$g(\log n) T_n^* - h_1(\log n) \xrightarrow[n \to \infty]{d} Z$$

$$g(\log n) \sqrt{T_n} - h_p(\log n) \xrightarrow[n \to \infty]{d} Z$$

where $Z \sim Gumbel(\log 2, 1)$, $g(y) = \sqrt{2 \log y}$, $h_p(y) = 2 \log y + \frac{p}{2} \log \log y - \log\left(\Gamma\left(\frac{p}{2}\right)\right)$

Remark: The assertions of the theorem remain true also for random design.

Bootstrap



B sets of Bootstrap Samples
B number of Estimates
Original Sample Data
Bootstrap Distribution
of Parameter Estimates
$\overline{X}^{*1}$
$\overline{X}^{*2}$
$\overline{X}^{*B}$
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−3
−2
−1
0
1
2
3

Permutation

Let $\mathbf{R} = (R_1, ..., R_n)$ be a random permutation on $1, ..., n$. Define

$$\mathbf{S}_k(\mathbf{R}) = \sum_{i=1}^{k} \mathbf{x}_i u_{R_i} - \mathbf{X}_k^T \mathbf{X}_k (\mathbf{X}^T \mathbf{X})^{-1} \sum_{j=1}^{n} \mathbf{x}_j u_{R_j}$$

$$S_k^*(\mathbf{R}) = \sum_{i=1}^{k} u_{R_i}$$

$$\hat{\sigma}_n^2(\mathbf{R}) = \frac{1}{n-p} \left\{ \sum_{i=1}^{n} u_i^2 - \max_{p<k<n-p} \left\{ \mathbf{S}_k^T(\mathbf{R}) \left(\mathbf{X}_k{}^T \mathbf{X}_k \right)^{-1} \left(\mathbf{X}^T \mathbf{X} \right) \left(\mathbf{X}_k^o{}^T \mathbf{X}_k^o \right)^{-1} \mathbf{S}_k(\mathbf{R}) \right\} \right\}$$

Permutational versions $T_n(\mathbf{R}), T_n^*(\mathbf{R})$ of T_n, T_n^* are defined by replacing $\mathbf{S}_k, S_k^*$ and $\hat{\sigma}_n^2$ by their permutational counterparts.

Limit Theorem - permutation

Asymptotic distribution of T_n^* and T_n

Let assumptions A.1 - A.3 be satisfied. Then

$$g(\log n)T_n^*(\mathbf{R}) - h_1(\log n) \mid \mathbf{Y}_n \xrightarrow[n\to\infty]{d} Z \quad \text{in probability,}$$

$$g(\log n)\sqrt{T_n(\mathbf{R})} - h_p(\log n) \mid \mathbf{Y}_n \xrightarrow[n\to\infty]{d} Z \quad \text{in probability,}$$

where $Z \sim \mathit{Gumbel}(\log 2, 1)$, $g(y) = \sqrt{2\log y}$, $h_p(y) = 2\log y + \frac{p}{2}\log\log y - \log\left(\Gamma\left(\frac{p}{2}\right)\right)$.

Remark: Notice that, contrary to the previous theorem, we do not require H_0 to hold.

Application – Malá Ráztoka

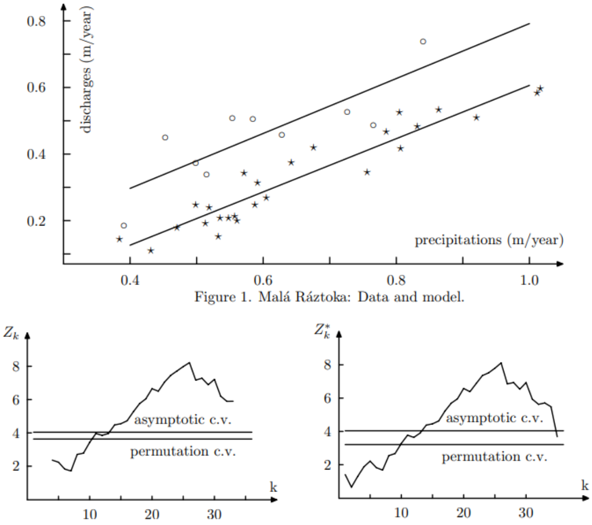


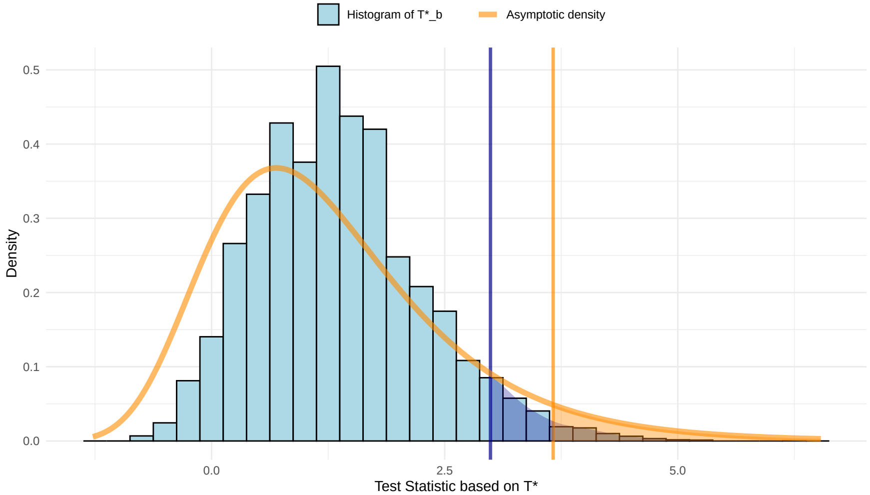
Figure 1. Malá Ráztoka: Data and model.

Figure 2a. Statistics Z_k.

Figure 2b. Statistics Z_k^*.

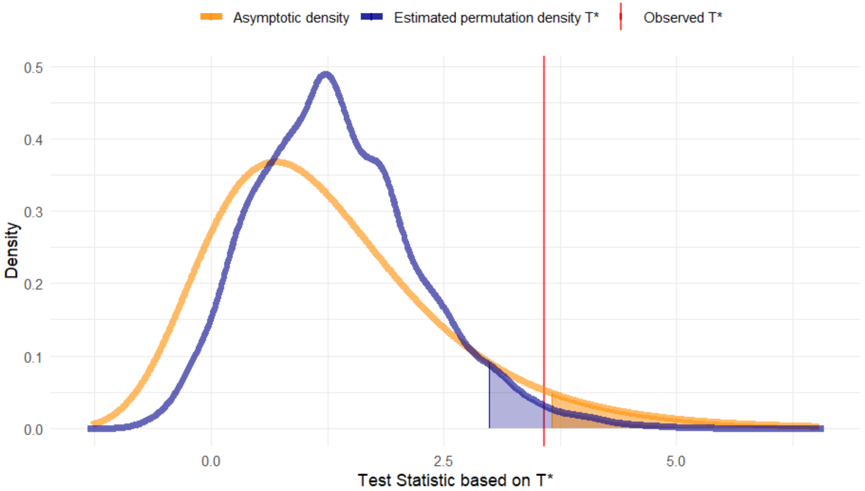
Semi-simulation study

Comparison of critical regions: Gumbel density (orange) versus kernel density estimation from permutation resamples based on T_n^* (blue).



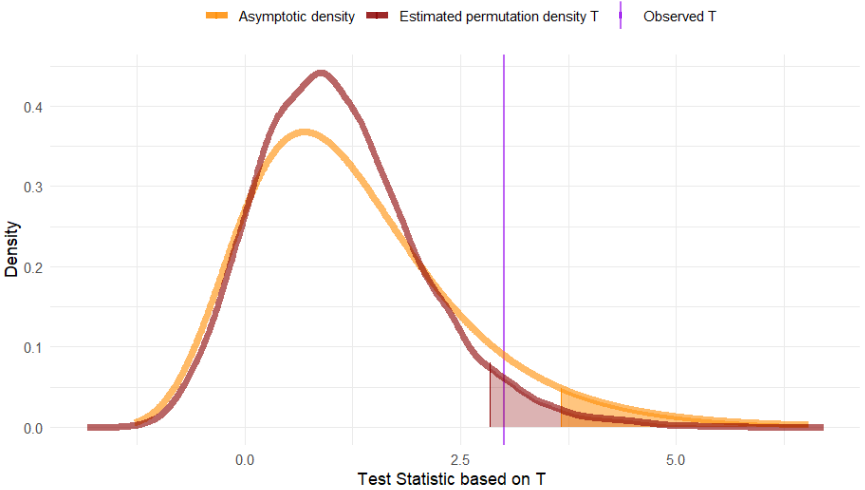
Semi-simulation study

Comparison of critical regions: Gumbel density (orange) versus kernel density estimation from permutation resamples based on T_n^* (blue).



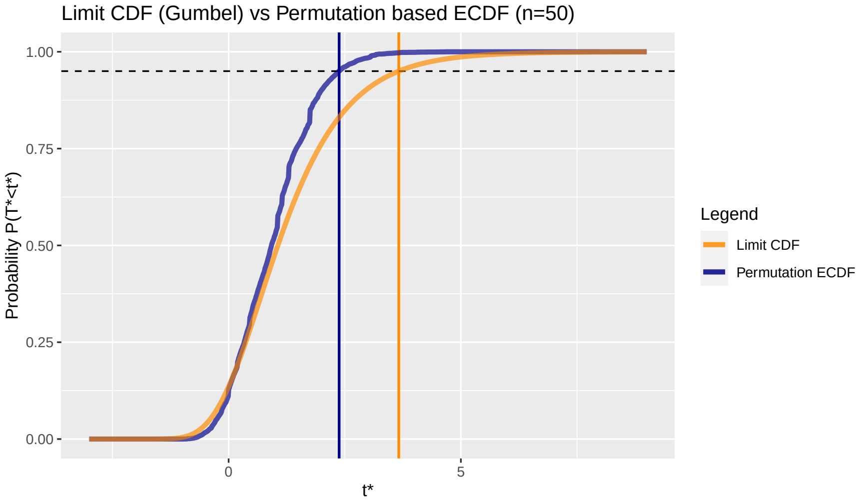
Semi-simulation study

Comparison of critical regions: Gumbel density (orange) versus kernel density estimation from permutation resamples based on T_n (red).



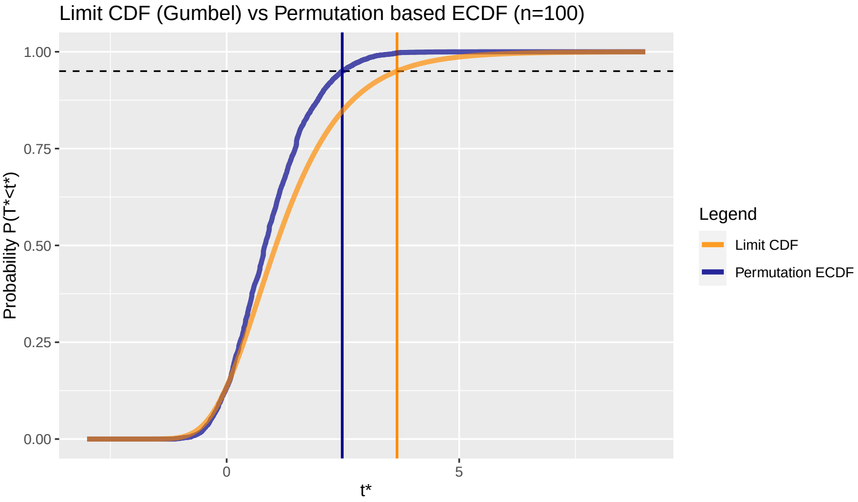
The Problem with Asymptotic Distribution

Critical region based on *Gumbel* distribution is very **conservative**.



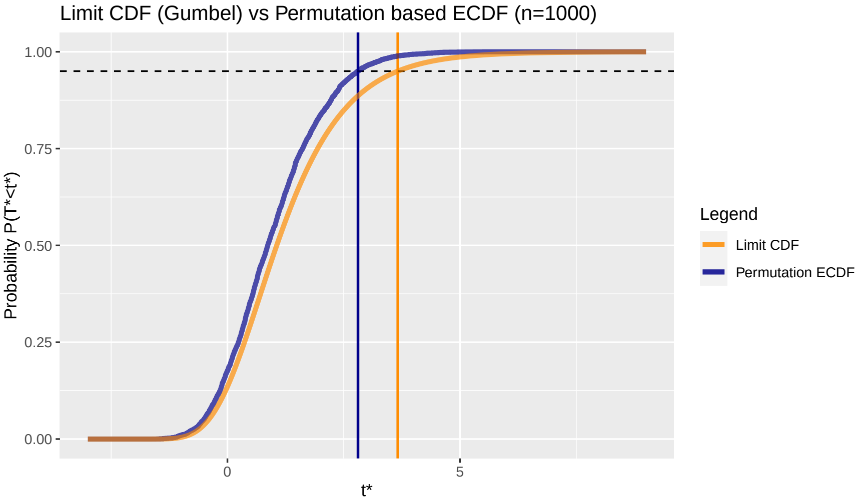
The Problem with Asymptotic Distribution

Critical region based on *Gumbel* distribution is very **conservative**.



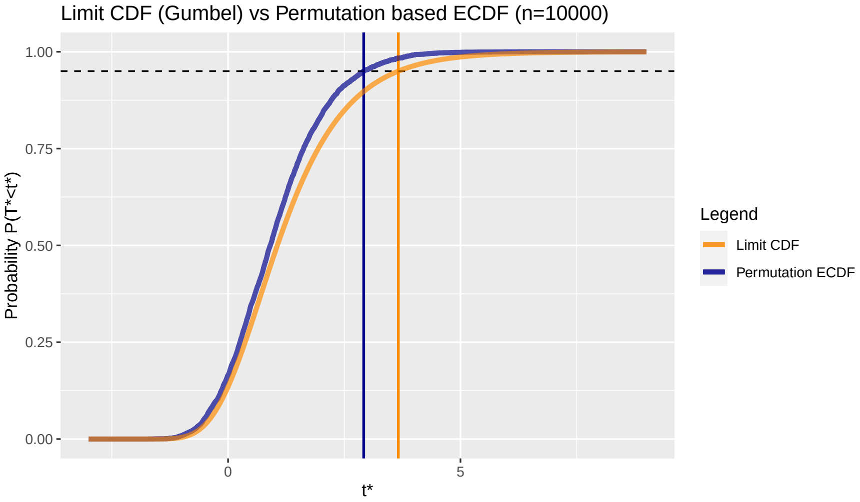
The Problem with Asymptotic Distribution

Critical region based on *Gumbel* distribution is very **conservative**.



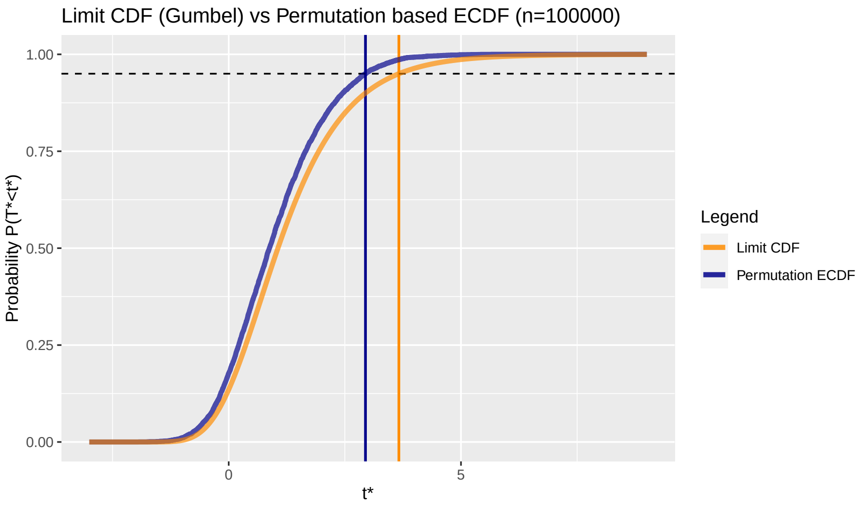
The Problem with Asymptotic Distribution

Critical region based on *Gumbel* distribution is very **conservative**.



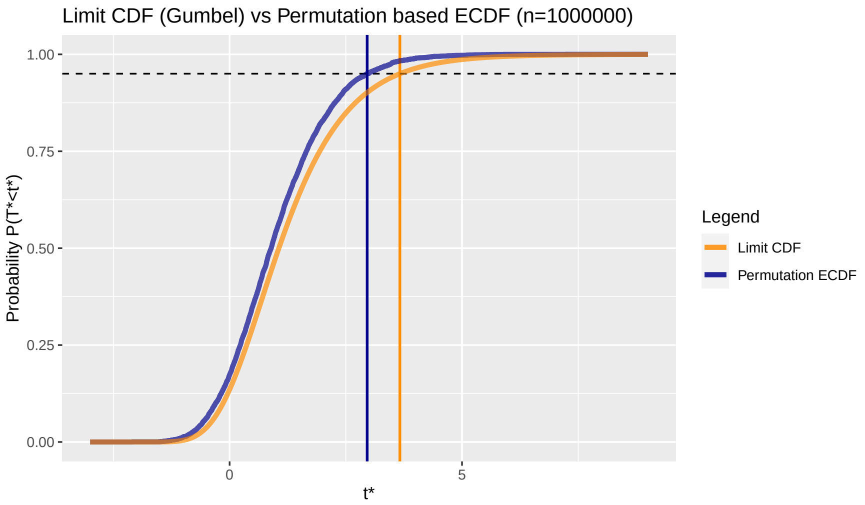
The Problem with Asymptotic Distribution

Critical region based on *Gumbel* distribution is very **conservative**.



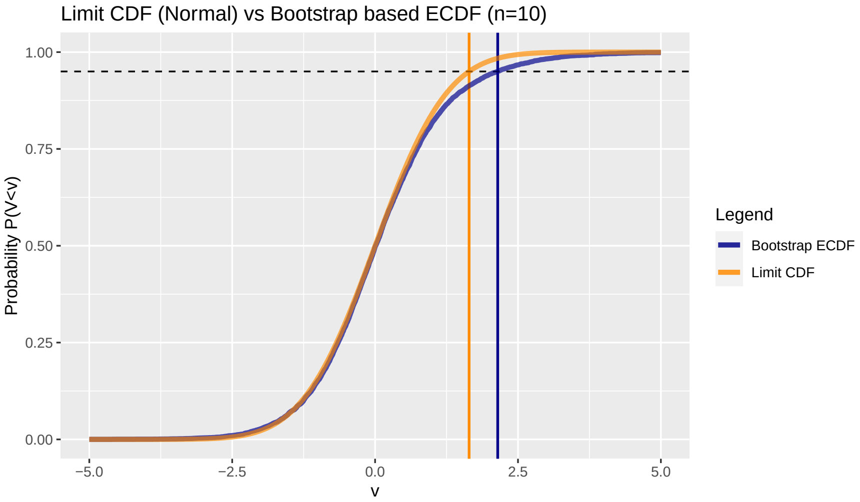
The Problem with Asymptotic Distribution

Critical region based on *Gumbel* distribution is very **conservative**.



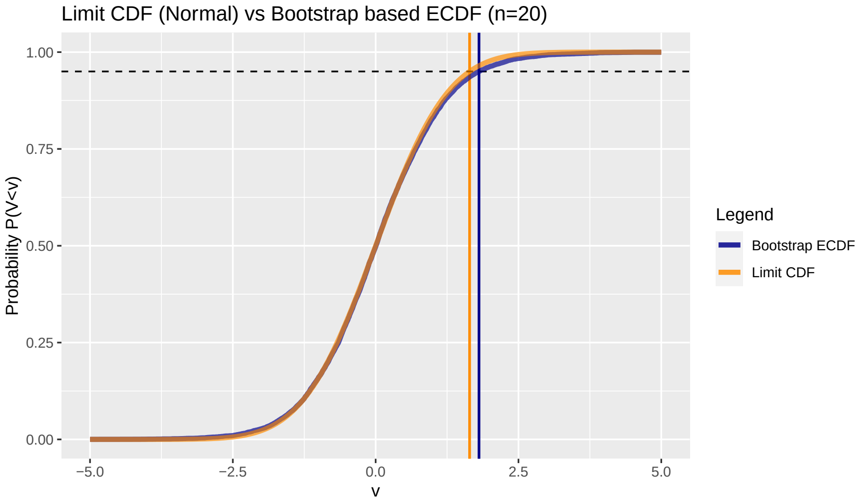
Comparison with Central Limit Theorem

Let $X_1, \ldots, X_n$ be random sample from $Unif(0,1)$. Denote $V = \frac{\sqrt{n}\left(\overline{X} - \mathrm{E}[X_1]\right)}{sd(X_1)}$.



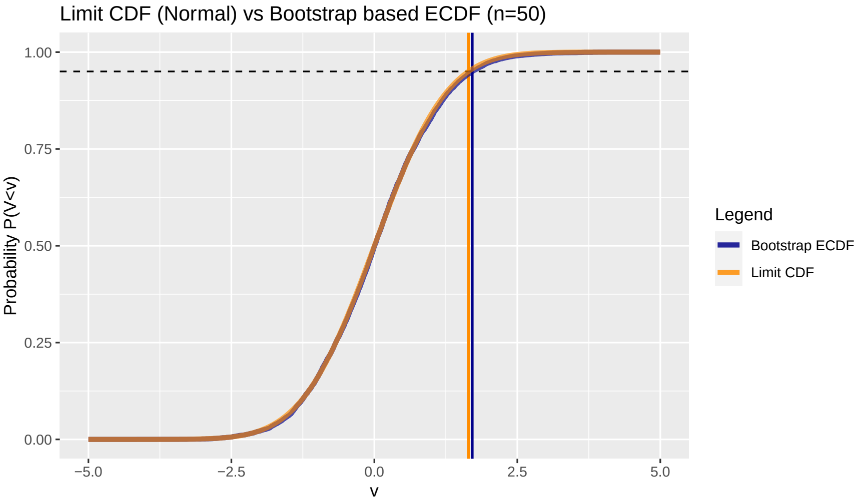
Comparison with Central Limit Theorem

Let $X_1, \ldots, X_n$ be random sample from $Unif(0,1)$. Denote $V = \frac{\sqrt{n}\left(\overline{X} - \mathrm{E}[X_1]\right)}{sd(X_1)}$.



Comparison with Central Limit Theorem

Let $X_1, \ldots, X_n$ be random sample from $Unif(0,1)$. Denote $V = \frac{\sqrt{n}\left(\overline{X} - \mathrm{E}[X_1]\right)}{sd(X_1)}$.



Application – Malá Ráztoka – Small effect

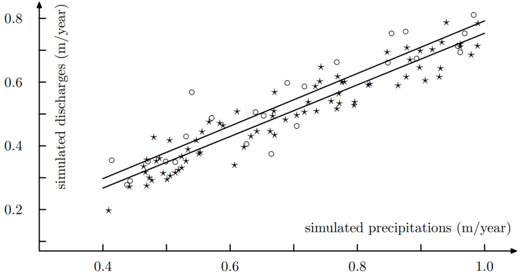


Figure 5. Simulated data and model.

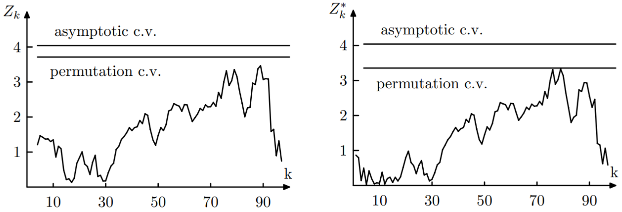


Figure 6a. Statistics Z_k.

Figure 6b. Statistics Z_k^*.