Two-sample gradual change analysis

Autoři článku Zdeněk Hlávka a Marie Hušková

Martin Hruška

April 18, 2023

Úvod

Článek "Two-sample gradual change analysis" autorů Zdeňka Hlávky a Marie Huškové představuje metodu pro detekci postupných změn v rozdělení dvou nezávislých výběrů. Příklad: rozdíly v rychlosti skoku mezi pohlavími - 432 dívek a 364 chlapců ve věku 6 až 19 let.

- Metoda vícenásobného testování
- Change point analýza
- Srovnání
- Simulace

Základní charakteristiky dat

Age cat.	girls		boys	
	$\overline{Y_1}$ ($\hat{\sigma}_1$)	n_1	$\overline{Y_2}$ ($\hat{\sigma}_2$)	n_2
6–7	1.89 (0.17)	33	1.87 (0.18)	19
7–8	2.00 (0.21)	43	1.98 (0.20)	38
8–9	2.01 (0.21)	33	2.06 (0.21)	38
9–10	2.06 (0.18)	42	2.14 (0.18)	29
10–11	2.19 (0.22)	42	2.17 (0.19)	45
11–12	2.23 (0.15)	30	2.31 (0.23)	37
12–13	2.26 (0.13)	41	2.35 (0.23)	40
13–14	2.30 (0.22)	32	2.53 (0.21)	36
14–15	2.28 (0.23)	31	2.66 (0.19)	20
15–16	2.37 (0.17)	29	2.72 (0.22)	26
16–17	2.33 (0.19)	17	2.83 (0.28)	9
17–18	2.35 (0.18)	25	2.76 (0.16)	13
18–19	2.33 (0.17)	34	2.87 (0.10)	14

Metoda vícenásobného testování

- chceme odhad neznámého bodu změny $\rightarrow$13 t-testů
- rychlost skoků u chlapců a u dívek jsou od 6 do 10 let přibližně stejné
- zřetelně vyšší u chlapců od 13 let
- korekce na vícenásobné testování
  - Bonferonni - kontroluje chybu 1. druhu
  - Benjamini–Hochberg (BH) - kontroluje chybu 2. druhu

Age cat.	girls		boys		p-values		
	$\overline{Y_1}$ ($\widehat{\sigma}_1$)	n_1	$\overline{Y_2}$ ($\widehat{\sigma}_2$)	n_2	t-test	Bonferroni	BH
6–7	1.89 (0.17)	33	1.87 (0.18)	19	0.780	1.000	0.780
7–8	2.00 (0.21)	43	1.98 (0.20)	38	0.646	1.000	0.763
8–9	2.01 (0.21)	33	2.06 (0.21)	38	0.369	1.000	0.479
9–10	2.06 (0.18)	42	2.14 (0.18)	29	0.081.	1.000	0.117
10–11	2.19 (0.22)	42	2.17 (0.19)	45	0.713	1.000	0.773
11–12	2.23 (0.15)	30	2.31 (0.23)	37	0.062.	0.800	0.100
12–13	2.26 (0.13)	41	2.35 (0.23)	40	0.047*	0.615	0.088.
13–14	2.30 (0.22)	32	2.53 (0.21)	36	0.000***	0.001***	0.000***
14–15	2.28 (0.23)	31	2.66 (0.19)	20	0.000***	0.000***	0.000***
15–16	2.37 (0.17)	29	2.72 (0.22)	26	0.000***	0.000***	0.000***
16–17	2.33 (0.19)	17	2.83 (0.28)	9	0.001***	0.006**	0.001**
17–18	2.35 (0.18)	25	2.76 (0.16)	13	0.000***	0.000***	0.000***
18–19	2.33 (0.17)	34	2.87 (0.10)	14	0.000***	0.000***	0.000***

Předpoklady

(A1) Pozorování $Y_{jik}(j = 1, 2; k = 1, ..., n_{ji})$ jsou získána v čase $i(i = 1, ..., n)$.

(A2) Všechna pozorování jsou nezávislá.

(A3) $E(\overline{Y_{1i}} - \overline{Y_{2i}}) = \mu + \delta((i - k_0)/n)_+ (i = 1, ..., n)$, kde μ, δ jsou neznámé parametry a $k_0 = n\theta_0$ pro nějaké $\theta_0 \in (0, 1)$.

(A4) $Var(Y_{jik}) = \sigma^2_{ji} > 0 (j = 1, 2; i = 1, ..., n; k = 1, ..., n_{ji})$.

Homoskedastický případ

(A4*) $Var(\overline{Y_{1i}} - \overline{Y_{2i}}) = \sigma^2/m(i = 1, ..., n)$, kde σ^2 je neznámý parametr a m může záviset na n.

LSE $\hat{\mu}, \hat{\delta}, \hat{k_\mu}$ minimalizují $\sum_{i=1}^{n}\{\overline{Y_{1i}} - \overline{Y_{2i}} - a - d((i-k)/n)_+\}^2$.

$$\widehat{k}_\mu = \arg\max_{k\in(1,n)} \left[\frac{\left\{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)(\overline{Y}_{1i} - \overline{Y}_{2i})\right\}^2}{\sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2} \right],$$

$$\widehat{\delta}_\mu = \frac{\sum_{i=1}^{n}(x_{i\widehat{k}} - \overline{x}_{\widehat{k}})(\overline{Y}_{1i} - \overline{Y}_{2i})}{\sum_{i=1}^{n}(x_{i\widehat{k}} - \overline{x}_{\widehat{k}})^2},$$

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(\overline{Y}_{1i} - \overline{Y}_{2i}) - \widehat{\delta}_\mu \overline{x}_{\widehat{k}}\,.$$

LSE v našem případě

$$\widehat{k}_0 = \arg \max_{k \in (1,n)} \left[\frac{\left\{ \sum_{i=1}^{n} x_{ik} (\overline{Y}_{1i} - \overline{Y}_{2i}) \right\}^2}{\sum_{i=1}^{n} x_{ik}^2} \right],$$

$$\widehat{\delta}_0 = \frac{\sum_{i=1}^{n} x_{i\widehat{k}} (\overline{Y}_{1i} - \overline{Y}_{2i})}{\sum_{i=1}^{n} x_{i\widehat{k}}^2}.$$

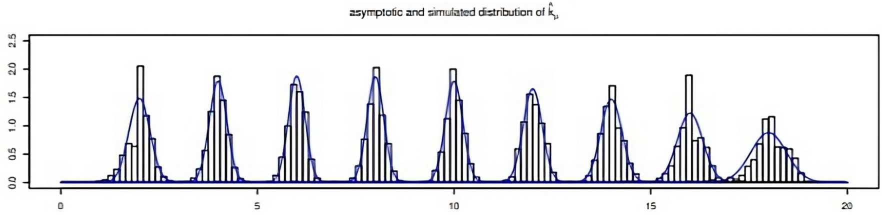
Vlastnosti odhadů

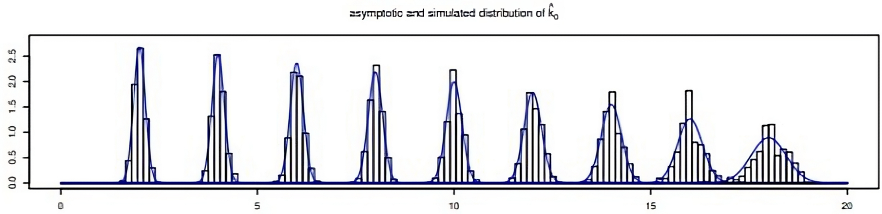
$$(nm)^{1/2} \frac{\delta}{\sigma} \left\{ \frac{\theta_0(1-\theta_0)}{1+3\theta_0} \right\}^{1/2} \frac{\widehat{k}_\mu - k_0}{n} \xrightarrow{\mathcal{D}} N(0,1)$$

$$(nm)^{1/2} \frac{(1-\theta_0)^{3/2}}{\sigma} \left(\frac{1+3\theta_0}{12} \right)^{1/2} (\widehat{\delta}_\mu - \delta) \xrightarrow{\mathcal{D}} N(0,1)$$

$\Rightarrow (nm)^{1/2}\delta(\hat{k}_\mu - k_0)/n = O_P(1)$ a $(nm)^{1/2}(\hat{\delta}_\mu - \delta)/n = O_P(1)$

MC simulace


asymptotic and simulated distribution of k̂
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20


asymptotic and simulated distribution of k̂
0.0
0.5
1.0
1.5
2.0
2.5
0
5
10
15
20

Heteroskedastický případ

$$\widehat{k}_0(\widehat{\tau}^2) = \arg\max_{k \in (1,n)} \left[\frac{\left\{ \sum_{i=1}^{n} x_{ik} (\overline{Y}_{1i} - \overline{Y}_{2i}) / \widehat{\tau}_i^2 \right\}^2}{\sum_{i=1}^{n} x_{ik}^2 / \widehat{\tau}_i^2} \right] = \arg\max_{k \in (1,n)} T_{2,\widehat{\tau}^2}(k)$$

Bootstrap algoritmus

- Odhad parametrů δ a k_0
- Vypočítejte vyrovnané hodnoty $\hat{D}_i = \hat{\delta}_0((i - \hat{k})/n)_+ (i = 1, ..., n)$
- For b = 1 to b = B:
  - Vygenerujte ${D_i}^* = \hat{D}_i + \hat{\tau}_i {\epsilon_i}^* (i = 1, ..., n)$, kde ${\epsilon_i}^* \sim N(0,1)$ jsou nezávislé
  - Vypočítejte odhad bodu změny $\hat{k_b}^*$ z bootstrapového výběru ${D_1}^*, \ldots, {D_n}^*$
- Vypočítejte empirický kvantil ${q_\alpha}^*$ z $\hat{k_1}^* - \hat{k}, \ldots, \hat{k_B}^* - \hat{k}$ pro $\alpha \in (0,1)$.

Simulace-homoskedastický případ

		θ_0	$\widehat{\sigma}^2_{\text{pooled}}$				$\widehat{\sigma}^2_{ji}$			
			$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}^{\text{corr}}_\mu$	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}^{\text{corr}}_\mu$	$\widehat{k}_0(\widehat{\tau}^2)$
$n = 10$	$n_{ji} = 10$	0.1	88.8	95.2	93.3	94.0	89.5	93.9	93.7	94.5
		0.2	92.4	95.9	92.8	94.9	91.7	94.8	94.3	95.6
		0.4	94.1	92.3	92.9	91.9	95.5	92.4	93.3	92.0
		0.6	92.8	93.2	92.6	92.4	93.9	89.9	92.2	90.2
		0.8	90.4	90.5	90.0	90.8	89.6	87.7	89.2	89.1
		0.9	78.1	78.3	79.2	78.4	80.1	74.8	76.4	76.0
	$n_{ji} = 20$	0.1	93.9	92.0	94.4	92.1	95.1	92.8	94.6	93.0
		0.2	96.3	92.8	95.6	92.4	95.4	92.7	95.8	94.7
		0.4	93.5	92.0	92.3	91.1	92.7	91.6	92.0	90.8
		0.6	87.6	90.1	90.1	89.8	89.3	87.1	88.9	88.4
		0.8	88.8	89.0	89.7	87.8	89.9	86.9	87.2	88.4
		0.9	72.1	70.4	74.9	70.1	72.4	70.9	72.6	70.3
$n = 20$	$n_{ji} = 10$	0.1	96.5	94.3	94.0	94.9	94.9	93.5	95.8	93.1
		0.2	97.1	94.1	95.0	93.7	96.9	93.0	95.0	93.6
		0.4	94.0	93.7	93.8	93.9	94.4	92.1	93.0	92.8
		0.6	93.2	90.9	92.7	92.7	91.9	92.1	91.7	91.8
		0.8	94.8	95.6	94.3	95.3	93.8	94.5	92.5	93.7
		0.9	84.1	84.3	84.8	84.0	83.1	81.8	84.4	80.9
	$n_{ji} = 20$	0.1	97.3	95.0	94.4	94.9	97.0	93.5	93.4	95.3
		0.2	95.1	94.3	94.1	94.3	93.5	93.9	94.1	94.0
		0.4	93.0	93.1	93.1	92.9	93.6	93.6	93.1	94.7
		0.6	91.9	90.7	92.8	92.8	91.7	93.6	92.0	91.4
		0.8	93.2	91.9	91.3	90.7	91.8	92.5	91.5	89.3
		0.9	79.5	81.4	83.4	79.3	82.3	80.4	82.4	82.4

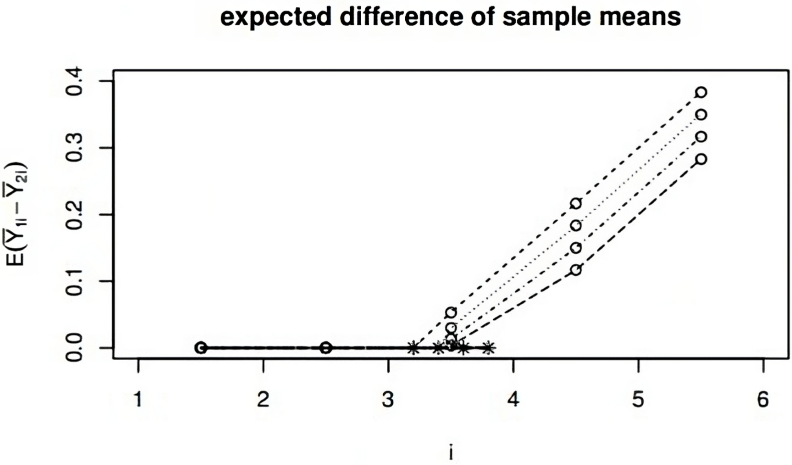
Simulace-heteroskedastický případ

		$n=10$								$n=20$			
		$\widehat{\sigma}^2_{\text{pooled}}$				$\widehat{\sigma}^2_{ji}$				$\widehat{\sigma}^2_{ji}$			
	θ_0	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_\mu$	$\widehat{k}_0$	$\widehat{k}_\mu^{\text{corr}}$	$\widehat{k}_0(\widehat{\tau}^2)$
H01	0.1	65.1	66.8	58.0	67.9	88.1	92.5	92.5	93.2	97.5	93.1	95.1	93.8
	0.4	69.9	68.9	67.7	70.7	96.3	94.7	95.7	92.2	94.1	93.3	94.4	95.1
	0.8	74.7	71.2	72.2	68.5	88.5	86.5	86.7	88.4	95.9	94.9	95.2	91.4
	0.9	84.1	82.8	77.9	80.5	78.6	77.4	78.8	78.1	77.5	80.4	80.8	78.9
H02	0.1	60.5	63.2	50.6	65.6	83.7	94.2	92.5	93.9	95.3	93.0	95.0	93.1
	0.4	65.9	65.4	63.9	69.6	90.6	88.4	91.0	93.0	93.6	92.5	92.8	93.2
	0.8	69.4	69.5	74.1	72.9	90.4	89.2	91.0	86.3	89.6	88.4	90.7	90.4
	0.9	80.0	78.2	77.2	83.7	76.8	71.9	75.6	75.8	82.6	84.2	82.3	81.9
H10	0.1	90.0	93.9	92.4	94.0	88.8	94.6	93.4	92.8	92.6	94.0	93.1	95.3
	0.4	97.1	99.6	99.7	99.7	92.7	91.5	94.3	91.1	94.6	94.9	94.9	93.3
	0.8	89.0	93.6	93.3	92.2	91.6	92.5	90.7	89.0	95.3	91.4	94.8	90.5
	0.9	94.1	87.5	87.5	68.5	92.4	88.8	86.1	73.4	87.5	86.6	89.1	82.3
H11	0.1	65.6	65.4	55.6	69.9	89.9	93.1	91.8	93.1	90.4	95.5	91.7	94.2
	0.4	71.0	67.2	67.4	70.8	92.4	94.8	93.2	91.9	93.7	92.4	93.1	93.3
	0.8	79.0	78.7	76.1	71.0	92.2	84.3	90.7	89.8	96.8	95.9	96.4	89.3
	0.9	95.5	92.5	84.6	73.1	93.1	90.3	87.5	66.2	88.4	86.8	86.3	80.1
H12	0.1	58.9	63.9	49.9	64.0	88.3	95.6	92.0	94.2	88.8	94.4	92.4	93.1
	0.4	70.3	68.3	65.3	69.5	90.0	87.1	88.6	91.5	94.5	93.0	94.7	92.9
	0.8	79.6	77.5	78.8	72.2	91.1	90.1	92.7	87.9	93.9	88.5	91.3	88.8
	0.9	93.2	88.3	81.7	78.0	91.3	85.2	85.7	74.3	88.6	87.3	90.9	82.2
H20	0.1	80.9	92.4	91.0	99.4	82.1	91.5	88.0	98.5	97.0	97.2	98.9	99.3
	0.4	93.8	94.6	93.2	99.8	93.0	89.6	89.9	93.7	97.4	95.4	96.6	99.2
	0.8	78.0	75.5	79.1	74.1	78.2	77.4	76.2	73.8	93.2	90.6	91.1	94.1
	0.9	90.6	86.2	84.8	78.3	90.0	86.5	83.3	77.5	79.0	78.4	83.7	74.4
H21	0.1	58.8	63.8	48.8	66.9	76.6	88.7	80.7	94.5	87.8	87.4	87.0	93.3
	0.4	60.6	62.8	61.9	65.2	83.9	82.0	83.3	88.4	85.0	85.4	84.2	90.8
	0.8	73.3	70.7	68.0	69.3	79.9	79.4	76.9	79.8	84.7	84.8	84.3	88.1
	0.9	93.3	88.7	81.8	82.5	87.6	85.4	81.3	78.9	81.1	81.9	79.7	77.4
H22	0.1	49.8	58.4	39.4	69.1	61.6	84.4	73.7	94.8	79.0	83.2	80.4	93.0
	0.4	57.1	61.2	56.0	72.5	74.1	68.2	75.6	90.2	78.0	81.9	81.0	93.3
	0.8	72.6	67.5	72.1	67.6	82.6	79.0	74.9	73.5	76.7	75.4	75.5	89.4
	0.9	92.2	86.6	79.2	79.9	90.3	83.7	83.2	82.1	83.9	78.7	81.9	76.8

Případy v simulaci

	Nr. of observations (n_{ji})		
	$n_{ji} = m$	$m\{1 + 3I(i \text{ odd})\}/2$	$m\{1 + 3I(i > n/2)\}/2$
$\bar{\sigma}_{ji}$ constant ($\sigma_{ji} = \sigma$)		H01	H02
$\sigma_{ji} = \sigma(1 + 2I(i > k_0))$	H10	H11	H12
$\sigma_{ji} = \sigma(1 + 2I(i \text{ even}))$	H20	H21	H22

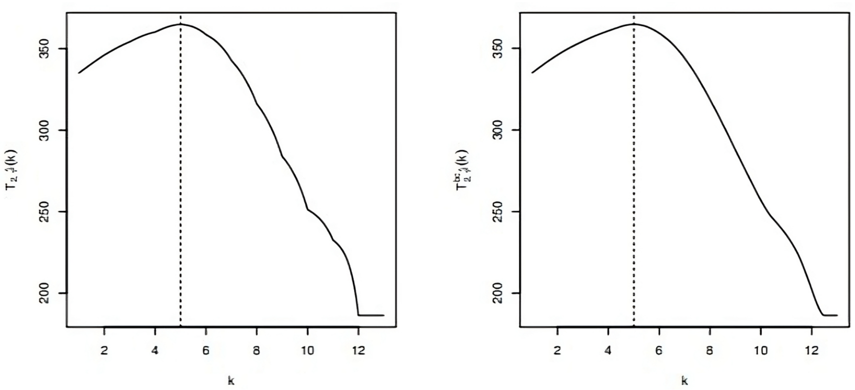
Zaokrouhlování



Simulace

		θ_0	$\widehat{k}_0$ MSE	$\widehat{k}_0$ bias	$\widehat{k}_0$ coverage	$\widehat{k}_0^{bc}$ MSE	$\widehat{k}_0^{bc}$ bias	$\widehat{k}_0^{bc}$ coverage
$n = 10$	$n_{ji} \equiv 10$	0.20	0.124	0.003	93.6%	0.112	−0.010	94.6%
		0.22	0.113	−0.001	95.4%	0.113	−0.016	95.3%
		0.25	0.125	−0.018	91.5%	0.123	−0.010	92.9%
		0.28	0.122	0.012	91.2%	0.133	0.011	90.3%
		0.30	0.116	0.008	93.9%	0.131	0.001	95.0%
		0.70	0.432	−0.109	92.7%	0.365	−0.041	93.3%
		0.72	0.466	−0.099	94.0%	0.498	−0.065	91.7%
		0.75	0.678	−0.151	89.6%	0.726	−0.138	88.4%
		0.78	0.936	−0.226	86.5%	0.971	−0.123	91.5%
		0.80	1.160	−0.263	91.3%	1.255	−0.224	91.8%
	$n_{ji} \equiv 20$	0.20	0.053	−0.011	94.1%	0.052	0.002	93.2%
		0.22	0.054	−0.013	95.4%	0.050	0.003	98.4%
		0.25	0.053	−0.011	95.8%	0.060	−0.017	95.4%
		0.28	0.059	0.001	92.8%	0.054	−0.009	95.2%
		0.30	0.063	0.000	92.6%	0.060	0.011	92.8%
		0.70	0.177	−0.056	86.9%	0.173	0.004	96.3%
		0.72	0.194	−0.073	94.1%	0.191	−0.024	95.6%
		0.75	0.246	−0.072	94.6%	0.230	−0.052	91.7%
		0.78	0.319	−0.116	88.1%	0.302	−0.034	93.2%
		0.80	0.399	−0.097	88.0%	0.506	−0.092	96.9%
$n = 20$	$n_{ji} \equiv 10$	0.20	0.054	−0.005	93.6%	0.050	−0.004	95.7%
		0.22	0.056	−0.005	94.5%	0.057	−0.003	95.3%
		0.25	0.056	−0.016	94.6%	0.050	0.003	95.4%
		0.28	0.061	0.002	94.5%	0.055	−0.009	94.6%
		0.30	0.063	−0.004	93.1%	0.060	−0.012	93.9%
		0.70	0.150	−0.040	93.7%	0.158	0.001	94.7%
		0.72	0.175	−0.035	93.9%	0.163	−0.033	94.5%
		0.75	0.200	−0.051	93.3%	0.178	−0.023	96.7%
		0.78	0.220	−0.043	94.0%	0.229	−0.026	90.8%
		0.80	0.263	−0.050	94.7%	0.276	−0.022	93.8%
	$n_{ji} \equiv 20$	0.20	0.027	−0.006	94.0%	0.026	−0.004	93.3%
		0.22	0.026	−0.008	94.3%	0.024	−0.010	98.0%
		0.25	0.030	−0.006	93.3%	0.029	0.005	93.6%
		0.28	0.026	0.005	95.8%	0.030	0.001	95.9%
		0.30	0.033	−0.015	93.9%	0.031	−0.004	94.1%
		0.70	0.078	−0.024	90.7%	0.072	−0.000	94.1%
		0.72	0.089	−0.032	96.7%	0.075	−0.001	95.9%
		0.75	0.093	−0.037	90.6%	0.095	−0.006	93.4%
		0.78	0.112	−0.046	95.3%	0.103	−0.012	94.3%
		0.80	0.114	−0.040	90.0%	0.111	−0.012	95.7%

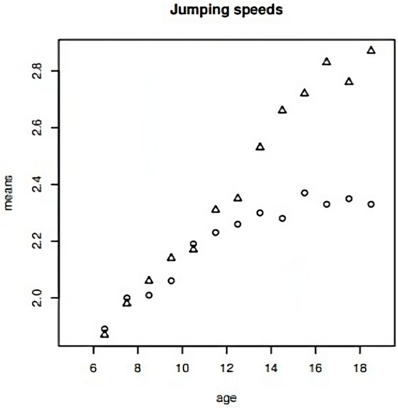
Funkce

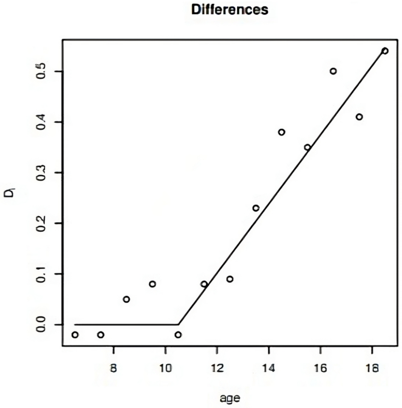


Porovnání

Age cat.	girls		boys		p-values					Age
	$\overline{Y_1}$ ($\widehat{\sigma}_1$)	n_1	$\overline{Y_2}$ ($\widehat{\sigma}_2$)	n_2	t-test	Bonferroni	BH	$\widehat{k}_0(\widehat{\tau}^2)$	$\widehat{k}_0^{bc}(\widehat{\tau}^2)$	
6–7	1.89 (0.17)	33	1.87 (0.18)	19	0.780	1.000	0.780	1.000	1.000	6
7–8	2.00 (0.21)	43	1.98 (0.20)	38	0.646	1.000	0.763	1.000	1.000	7
8–9	2.01 (0.21)	33	2.06 (0.21)	38	0.369	1.000	0.479	1.000	1.000	8
9–10	2.06 (0.18)	42	2.14 (0.18)	29	0.081.	1.000	0.117	0.999	0.997	9
10–11	2.19 (0.22)	42	2.17 (0.19)	45	0.713	1.000	0.773	0.861	0.846	10
11–12	2.23 (0.15)	30	2.31 (0.23)	37	0.062.	0.800	0.100	0.113	0.117	11
12–13	2.26 (0.13)	41	2.35 (0.23)	40	0.047*	0.615	0.088.	0.003**	0.003**	12
13–14	2.30 (0.22)	32	2.53 (0.21)	36	0.000***	0.001***	0.000***	0.000***	0.000***	13
14–15	2.28 (0.23)	31	2.66 (0.19)	20	0.000***	0.000***	0.000***	0.000***	0.000***	14
15–16	2.37 (0.17)	29	2.72 (0.22)	26	0.000***	0.000***	0.000***	0.000***	0.000***	15
16–17	2.33 (0.19)	17	2.83 (0.28)	9	0.001***	0.006**	0.001**	0.000***	0.000***	16
17–18	2.35 (0.18)	25	2.76 (0.16)	13	0.000***	0.000***	0.000***	0.000***	0.000***	17
18–19	2.33 (0.17)	34	2.87 (0.10)	14	0.000***	0.000***	0.000***	0.000***	0.000***	18

Verifikace (A3)





Zdroje

Hlávka, Z., & Hušková, M. (2017). Two-sample gradual change analysis. *Revstat - Statistical Journal*, 15(3), 355-372.