



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta informačních technologií

Projekt z MSP

Spracoval: xorsak02

Číslo zadání: 14, 8

Cvičení – skupina: čtvrtek, 8:00

Datum: 23.11.2020

Zadání projektu z předmětu MSP

Každý student obdrží na cvičení konkrétní data (čísla ze seznamu), pro které vypracuje projekt. K vypracování můžete použít libovolné statistické programy.

1. Při kontrole výrobků byla sledována odchylka X [mm] jejich rozměru od požadované velikosti.

Naměřené hodnoty tvoří statistický soubor v listu Data_př. 1.

- Provedte roztřídění statistického souboru, vytvořte tabulku četností a nakreslete histogramy pro relativní četnosti a relativní kumulativní četnosti.
- Vypočtěte aritmetický průměr, medián, modus, rozptyl a směrodatnou odchylku.
- Vypočtěte bodové odhady střední hodnoty, rozptylu a směrodatné odchylky.
- Testujte předpoklad o výběru z normálního rozdělení Pearsonovým (chí-kvadrát) testem na hladině významnosti 0,05.
- Za předpokladu (bez ohledu na výsledek části d)), že statistický soubor byl získán náhodným výběrem z normálního rozdělení, určete intervalové odhady střední hodnoty, rozptylu a směrodatné odchylky se spolehlivostí 0,95 a 0,99.
- Testujte hypotézu optimálního seřízení stroje, tj. že střední hodnota odchylky je nulová, proti dvoustranné alternativní hypotéze, že střední hodnota odchylky je různá od nuly, a to na hladině významnosti 0,05.
- Ověřte statistickým testem na hladině významnosti 0,05, zda seřízení stroje ovlivnilo kvalitu výroby, víte-li, že výše uvedený statistický soubor 50-ti hodnot vznikl spojením dvou dílčích statistických souborů tak, že po naměření prvních 20-ti hodnot bylo provedeno nové seřízení stroje a pak bylo naměřeno zbývajících 30 hodnot.
 - Návod: Oba soubory zpracujte neroztříděné. Testujte nejprve rovnost rozptylů odchylek před a po seřízení stroje. Podle výsledku pak zvolte vhodný postup pro testování rovnosti středních hodnot odchylek před a po seřízení stroje.

2. Měřením dvojice (Výška[cm], Váha[kg]) u vybraných studentů z FIT byl získán dvourozměrný statistický soubor zapsaný po dvojicích v řádcích v listu Data_př. 2.

- Vypočtěte bodový odhad koeficientu korelace.
- Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny Výška a Váha jsou lineárně nezávislé.
- Regresní analýza - data proložte přímkou: $Váha = \beta_0 + \beta_1 \cdot Výška$
 - Bodově odhadněte β_0 , β_1 a rozptyl s^2 .
 - Na hladině významnosti 0,05 otestujte hypotézy:
 $H : \beta_0 = -100, H_A : \beta_0 \neq -100,$
 $H : \beta_1 = 1, H_A : \beta_1 \neq 1,$
 - Vytvořte graf bodů spolu s regresní přímkou a pásem spolehlivosti pro individuální hodnotu výšky.

Termín pro odevzdání práce je 11 týden výuky zimního semestru ve cvičení.

Vypracování:

1. Při kontrole výrobků byla sledována odchylka X [mm] jejich rozměru od požadované velikosti. Naměřené hodnoty tvoří statistický soubor v listu Data_př. 1.

Statistický soubor	Usporadany statistický soubor
0,99	-0,49
0,52	-0,36
-0,49	-0,26
1,18	-0,12
1,17	0,07
0,88	0,11
0,27	0,16
0,07	0,21
0,41	0,27
0,16	0,39
0,88	0,41
1,48	0,52
0,93	0,59
0,75	0,69
0,21	0,75
1,22	0,79
-0,26	0,88
1,24	0,88
0,59	0,88
1,63	0,88
1,77	0,9
1,1	0,93
1,5	0,93
1,66	0,99
1	1
0,88	1,07
1,11	1,08
0,39	1,08
1,28	1,1
0,11	1,11
1,25	1,17
1,08	1,18
0,69	1,22
1,07	1,22
1,75	1,24
-0,36	1,25
1,3	1,25
1,39	1,28
1,55	1,28
0,88	1,3
1,25	1,33
1,33	1,39
0,9	1,48

0,79	1,5
0,93	1,55
1,08	1,63
-0,12	1,66
1,22	1,74
1,74	1,75
1,28	1,77

a) Proved'te roztřídění statistického souboru, vytvořte tabulku četností a nakreslete histogramy pro relativní četnosti a relativní kumulativní četnosti.

$$x_{(1)} = \min_i x_i = -0,49$$

$$x_{(n)} = \max_i x_i = 1,77$$

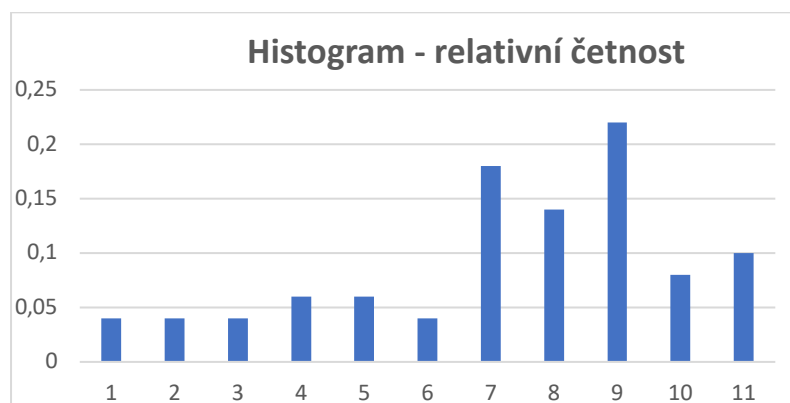
Variační obor: $\langle x_{(1)}, x_{(n)} \rangle = \langle -0,49, 1,77 \rangle$

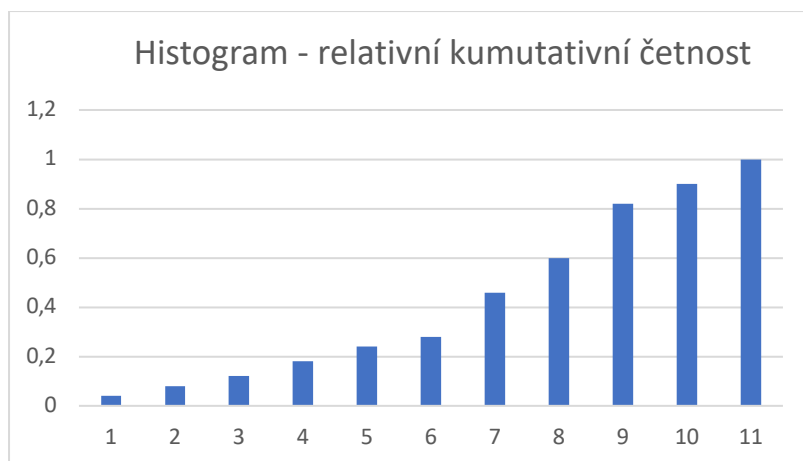
Rozpětí: $x_{(n)} - x_{(1)} = 2,26$

Počet tříd $m = 11$ (zvoleno)

Délka třídy $= \frac{x_{(n)} - x_{(1)}}{m} = 0,205454545$

trieda	xi-	xi+	stred triedy	kumulat. četnosť	četnosť	relat. četnosť	relat. kum. Čet
1	-0,49	-0,28454545	-0,387272727	2	2	0,04	0,04
2	-0,284545455	-0,07909091	-0,181818182	4	2	0,04	0,08
3	-0,079090909	0,12636364	0,023636364	6	2	0,04	0,12
4	0,126363636	0,33181818	0,229090909	9	3	0,06	0,18
5	0,331818182	0,53727273	0,434545455	12	3	0,06	0,24
6	0,537272727	0,74272727	0,64	14	2	0,04	0,28
7	0,742727273	0,94818182	0,845454545	23	9	0,18	0,46
8	0,948181818	1,15363636	1,050909091	30	7	0,14	0,6
9	1,153636364	1,35909091	1,256363636	41	11	0,22	0,82
10	1,359090909	1,56454545	1,461818182	45	4	0,08	0,9
11	1,564545455	1,77	1,667272727	50	5	0,1	1





b) Vypočítejte aritmetický průměr, medián, modus, rozptyl a směrodatnou odchylku.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0,9126$$

Medián: 1,035

Modus: 1,256363636

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = 0,31309924$$

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = 0,559552714$$

c) Vypočítejte bodové odhady střední hodnoty, rozptylu a směrodatné odchylky.

Bodový odhad střední hodnoty:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 0,9126$$

Bodový odhad rozptylu:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 0,31948902$$

Bodový odhad směrodatné odchylky:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 0,565233598$$

d) Testujte předpoklad o výběru z normálního rozdělení Pearsonovým (chí-kvadrát) testem na hladině významnosti 0,05.

Testovací kritérium:

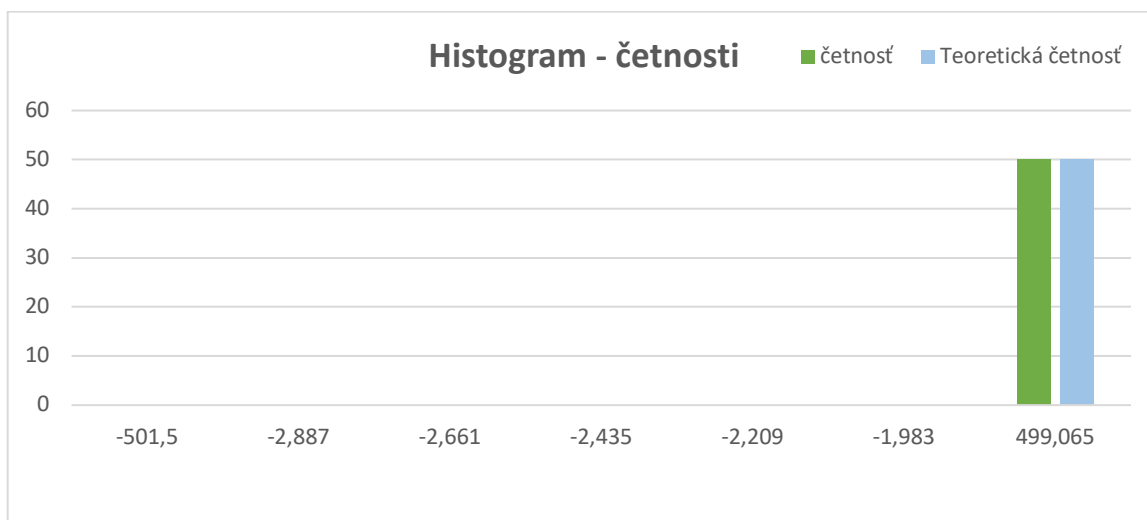
trieda	xi-	xi+	střed triedy	kumulat. četnost	četnost	Teoretická četnost	roz ² /teor čet
1	-1000	-3	-501,5	0	0	1,11254E-10	1,11254E-10
2	-3	-2,774	-2,887	0	0	1,62015E-09	1,62015E-09
3	-2,774	-2,548	-2,661	0	0	2,13091E-08	2,13091E-08
4	-2,548	-2,322	-2,435	0	0	2,3924E-07	2,3924E-07
5	-2,322	-2,096	-2,209	0	0	2,2929E-06	2,2929E-06
6	-2,096	-1,87	-1,983	0	0	1,87604E-05	1,87604E-05
7	-1,87	1000	499,065	50	50	49,99997868	9,08705E-12

$$t = \sum_{j=1}^m \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j} = 2,13155E-05 = \mathbf{0,0000213155}$$

$\chi^2_{1-\alpha}$ pre k = 7 - 2 - 1 stupňu volnosti: **9,487729037**

Doplňek kritického oboru: $\bar{W}_\alpha = \langle 0, \chi^2_{1-\alpha} \rangle = \langle \mathbf{0}, \mathbf{9,487729037} \rangle$

Protože $t \in \bar{W}_\alpha$, tedy hypotéza: $X \sim N(\mathbf{0,9126}; \mathbf{0,31948902})$ se **nezamítá**



e) Za předpokladu (bez ohledu na výsledek části d)), že statistický soubor byl získán náhodným výběrem z normálního rozdělení, určete intervalové odhady střední hodnoty, rozptylu a směrodatné odchylky se spolehlivostí 0,95 a 0,99.

Předpoklad: $X \sim N(\mu, \sigma^2)$, σ^2 - **neznáme**

Bodový odhad střední hodnoty:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \mathbf{0,9126}$$

Bodový odhad rozptylu:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \mathbf{0,31948902}$$

Bodový odhad směrodatné odchylky:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \mathbf{0,565233598}$$

Intervalový odhad parametru μ :

0,975 kvantil Studentova rozdělení $t_{1-\alpha/2}$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **2,009575237**

0,995 kvantil Studentova rozdělení $t_{1-\alpha/2}$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **2,679951974**

$$\alpha = 0,05: \left\langle \bar{x} - t_{1-\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2} \frac{s}{\sqrt{n}} \right\rangle = \langle \mathbf{0,751962389}; \mathbf{1,07323761} \rangle$$

$$\alpha = 0,01: \left\langle \bar{x} - t_{1-\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2} \frac{s}{\sqrt{n}} \right\rangle = \langle \mathbf{0,698375086}; \mathbf{1,12682491} \rangle$$

Intervalový odhad parametru σ^2 :

0,975 kvantil Pearsova rozdělení $\chi_{\alpha/2}^2$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **31,55492**

0,975 kvantil Pearsova rozdělení $\chi_{1-\alpha/2}^2$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **70,22241**

0,995 kvantil Pearsova rozdělení $\chi_{\alpha/2}^2$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **27,24935**

0,995 kvantil Pearsova rozdělení $\chi_{1-\alpha/2}^2$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **78,23071**

$$\alpha = 0,05: \left\langle \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}; \frac{(n-1)s^2}{\chi_{\alpha/2}^2} \right\rangle = \langle \mathbf{0,222934}; \mathbf{0,496117999} \rangle$$

$$\alpha = 0,01: \left\langle \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}; \frac{(n-1)s^2}{\chi_{\alpha/2}^2} \right\rangle = \langle \mathbf{0,200113}; \mathbf{0,574507742} \rangle$$

Intervalový odhad parametru σ :

$$\alpha = 0,05: \left\langle \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}}; \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}} \right\rangle = \langle \mathbf{0,472158848}; \mathbf{0,704356} \rangle$$

$$\alpha = 0,01: \left\langle \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}}; \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}} \right\rangle = \langle \mathbf{0,447339634}; \mathbf{0,757963} \rangle$$

- f) Testujte hypotézu optimálního seřízení stroje, tj. že střední hodnota odchylky je nulová, proti dvoustranné alternativní hypotéze, že střední hodnota odchylky je různá od nuly, a to na hladině významnosti 0,05.

Studentův jednovýběrový test:

Testujeme hypotézu $H_0: \mu = 0$:

$$\text{testovací kritérium: } t = \frac{\bar{x} - \mu_0}{s} \sqrt{n} = \frac{\bar{x} - 0}{s} \sqrt{n} = \mathbf{11,41661873}$$

doplňěk kritického oboru: $\bar{W}_\alpha = \langle -t_{1-\alpha/2}, t_{1-\alpha/2} \rangle$ pre alternativnú hypotézu: $H_A: \mu \neq \mu_0, 0,975$
kvantil Studentova rozdelení $t_{1-\alpha/2}$ s $k = n - 1 = 50 - 1 = 49$ stupni volnosti = **2,009575237**

$$\bar{W}_\alpha = \langle -t_{1-\alpha/2}, t_{1-\alpha/2} \rangle = \langle -2,0095752, 2,0095752 \rangle$$

Protože $t \notin \bar{W}_\alpha$, tedy hypotéza $H_0: \mu = 0$ se **zamítá** a alternativna hypotéza $H_A: \mu \neq 0$ sa ne-zamieta.

- g) Ověřte statistickým testem na hladině významnosti 0,05, zda seřízení stroje ovlivnilo kvalitu výroby, víte-li, že výše uvedený statistický soubor 50-ti hodnot vznikl spojením dvou dílčích statistických souborů tak, že po naměření prvních 20-ti hodnot bylo provedeno nové seřízení stroje a pak bylo naměřeno zbývajících 30 hodnot.

1	0,99
2	0,52
3	-0,49
4	1,18
5	1,17
6	0,88
7	0,27
8	0,07
9	0,41
10	0,16
11	0,88
12	1,48
13	0,93
14	0,75
15	0,21
16	1,22
17	-0,26
18	1,24
19	0,59
20	1,63

21	1,77
22	1,1
23	1,5
24	1,66
25	1
26	0,88
27	1,11
28	0,39
29	1,28
30	0,11
31	1,25
32	1,08
33	0,69
34	1,07
35	1,75
36	-0,36
37	1,3
38	1,39
39	1,55
40	0,88
41	1,25
42	1,33
43	0,9
44	0,79
45	0,93
46	1,08
47	-0,12
48	1,22
49	1,74
50	1,28

	X	Y
n =	20	30
průměr =	0,6915	1,06
rozptyl s ² =	0,31346275	0,25854
směr_odch =	0,5598774	0,508468288

Test rovnosti rozptylů – F-test:

Testujeme hypotézu $H_0: \sigma_x^2 = \sigma_y^2$:

testovací kritérium: $t = \frac{s^2(X)}{s^2(Y)} = \frac{0,313463}{0,25854} = 1,212434246$

doplňk kritického oboru: $\bar{W}_\alpha = \langle F_{\frac{\alpha}{2}}(n-1, m-1), F_{1-\frac{\alpha}{2}}(n-1, m-1) \rangle$ pre $H_0: \sigma_x^2 \neq \sigma_y^2$,

$F_{\frac{\alpha}{2}}(k_1, k_2), F_{1-\frac{\alpha}{2}}(k_1, k_2)$ jsou kvantily Fischerova-Snedecorova rozdělení s $k_1 = n - 1$ a $k_2 = m - 1$ stupni volnosti.

$F_{\frac{\alpha}{2}}(k_1, k_2) = 0,416329668$

$F_{1-\frac{\alpha}{2}}(k_1, k_2) = 2,231273833$

$\langle F_{\frac{\alpha}{2}}(n-1, m-1), F_{1-\frac{\alpha}{2}}(n-1, m-1) \rangle = \langle 0,416329668, 2,231273833 \rangle$

Protože $t \in \bar{W}_\alpha$, tedy hypotéza $H_0: \sigma_x^2 = \sigma_y^2$ se **ne-zamítá**.

Studentův dvouvýběrový test:

Testujeme hypotézu $H_0: \mu_x - \mu_y = 0$ za podmínky $\sigma_x^2 = \sigma_y^2$

testovací kritérium: $t = \frac{\bar{x} - \bar{y} - \mu_0}{\sqrt{\frac{(n-1)s^2(X) + (m-1)s^2(Y)}{n+m}}} \sqrt{\frac{n*m(n+m-2)}{n+m}} = -2,3615125$

doplňk kritického oboru: $\bar{W}_\alpha = \langle -t_{1-\alpha/2}, t_{1-\alpha/2} \rangle$ pre $H_A: \mu_x - \mu_y \neq 0$,

$t_{1-\alpha/2}$ – kvantil Studentova rozdělení s $k = n + m - 2 = 20 + 30 - 2 = 48$ stupni volnosti.

$t_{1-\alpha/2} = 2,010634758$

$\bar{W}_\alpha = \langle -t_{1-\alpha/2}, t_{1-\alpha/2} \rangle = \langle -2,010634758, 2,010634758 \rangle$

Protože $t \in \bar{W}_\alpha$, tedy hypotéza $H_0: \mu_x - \mu_y = 0$ se **ne-zamítá**.

2. Měřením dvojice (Výška[cm], Váha[kg]) u vybraných studentů z FIT byl získán dvourozměrný statistický soubor zapsaný po dvojicích v řádcích v listu Data_př. 2.

8	
Př. 2	
Výška [cm]	Váha [kg]
156	65
182	103
153	58
162	45
181	93
172	81
200	113
185	96
157	50
153	35
173	70
165	67
157	64
162	56
197	96
190	111
157	70
181	84
175	84
172	76

$$n = 20$$

$$\bar{x} = 171,5$$

$$\bar{y} = 75,85$$

$$\sum_{i=1}^n x_i^2 = 592316$$

$$\sum_{i=1}^n y_i^2 = 124089$$

$$\sum_{i=1}^n x_i^2 y_i^2 = 265633$$

a) Vypočtěte bodový odhad koeficientu korelace.

$$r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n \bar{x}^2)(\sum_{i=1}^n y_i^2 - n \bar{y}^2)}} = \mathbf{0,902039307}$$

b) Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny Výška a Váha jsou lineárně nezávislé.

Testujeme hypotézu $H_0 = \rho = 0$:

$$\text{Testovací kritérium: } t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} = \mathbf{8,865965719}$$

Doplňek kritického oboru: $\bar{W}_\alpha = \langle 0, t_{1-\alpha/2} \rangle$ pre alternativní hypotézu $H_A: \rho \neq 0$,

$$t_{1-\alpha/2} (n-2) = t_{0,975} (20-2) = \mathbf{2,10092204}$$

Protože $t \notin \bar{W}_\alpha$, tedy hypotéza $H_0 = \rho = 0$ se **zamítá**.

b) **Regresní analýza** - data proložte přímkou: Váha = $\beta_0 + \beta_1 \cdot$ Výška

xi - Výška [cm]	yi - Váha [kg]	xi ^2	yi ^2	xi*yi
156	65	24336	4225	10140
182	103	33124	10609	18746
153	58	23409	3364	8874
162	45	26244	2025	7290
181	93	32761	8649	16833
172	81	29584	6561	13932
200	113	40000	12769	22600
185	96	34225	9216	17760
157	50	24649	2500	7850
153	35	23409	1225	5355
173	70	29929	4900	12110
165	67	27225	4489	11055
157	64	24649	4096	10048
162	56	26244	3136	9072
197	96	38809	9216	18912
190	111	36100	12321	21090
157	70	24649	4900	10990
181	84	32761	7056	15204
175	84	300625	7056	14700
172	76	29584	5776	13072
3430	1517	592316	124089	265633
171,5	75,85			

Tedy:

$$n = 20, \sum_{i=1}^n x_i = 3430, \sum_{i=1}^n y_i = 1517, \sum_{i=1}^n x_i^2 = 592316, \sum_{i=1}^n y_i^2 = 124089, \sum_{i=1}^n x_i y_i = 265633$$

$$\det(H) = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 = \mathbf{81420}$$

1. Bodově odhadněte β_0 , β_1 a rozptyl s^2 .

$$b_2 = \frac{1}{\det(H)} (n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i) = \mathbf{1,343036109}$$

$$b_1 = \bar{y} - b_2 \bar{x} = \mathbf{-154,4806927}$$

$$y = \mathbf{-154,4806927 + 1,343036109 x}$$

$$S_{\min}^* = \sum_{i=1}^n y_i^2 - b_1 \sum_{i=1}^n y_i - b_2 \sum_{i=1}^n x_i y_i = \mathbf{1681,500074}$$

$$s^2 = \frac{S_{\min}^*}{n-2} = \frac{S_{\min}^*}{20-2} = \mathbf{93,41667076}$$

2. Na hladině významnosti 0,05 otestujte hypotézy:

$$H: \beta_0 = -100, \quad H_A: \beta_0 \neq -100,$$

$$h^{11} = \frac{\sum_{i=1}^n x_i^2}{\det(H)} = \mathbf{7,274821911}$$

$$t = \frac{b_1 - (-100)}{s\sqrt{h^{11}}} = \mathbf{-2,089869871}$$

$$t_{1-\alpha/2}(n-2) = t_{0,975}(20-2) = \mathbf{2,10092204}$$

$t \in \bar{W} = \langle -2,10092204, 2,10092204 \rangle$, a tedy $H: \beta_1 = -100$ se **ne-zamítá**

$$H: \beta_1 = 1, \quad H_A: \beta_1 \neq 1,$$

$$h^{22} = \frac{n}{\det(H)} = \mathbf{0,00024564}$$

$$t = \frac{b_2 - 1}{s\sqrt{h^{22}}} = \mathbf{2,264530613}$$

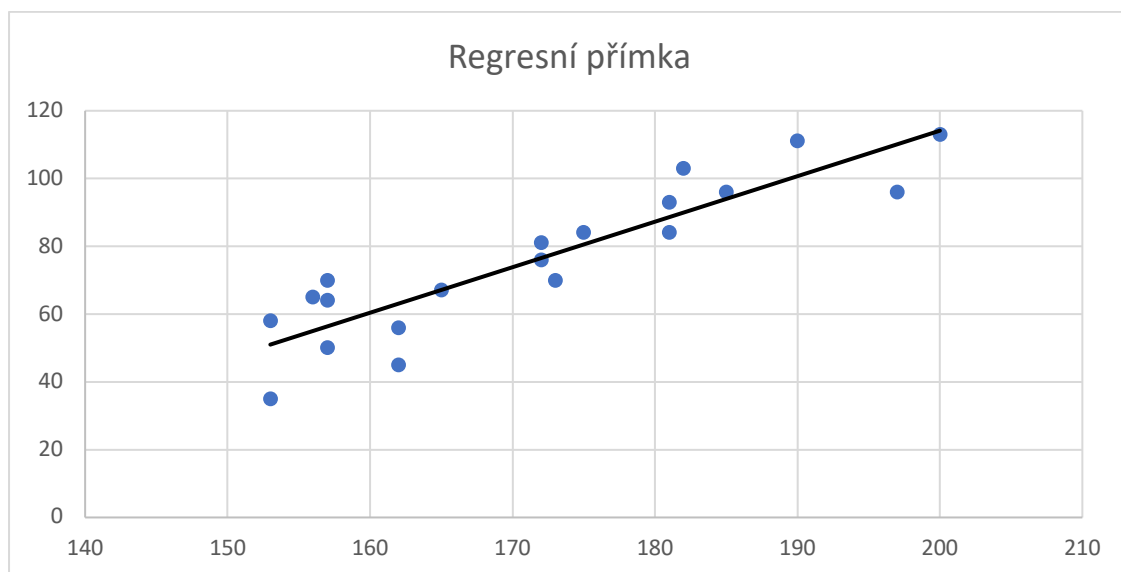
$$t_{1-\alpha/2}(n-2) = t_{0,975}(20-2) = \mathbf{2,10092204}$$

$t \notin \bar{W} = \langle -2,10092204, 2,10092204 \rangle$, a tedy $H: \beta_2 = -100$ se **zamítá**

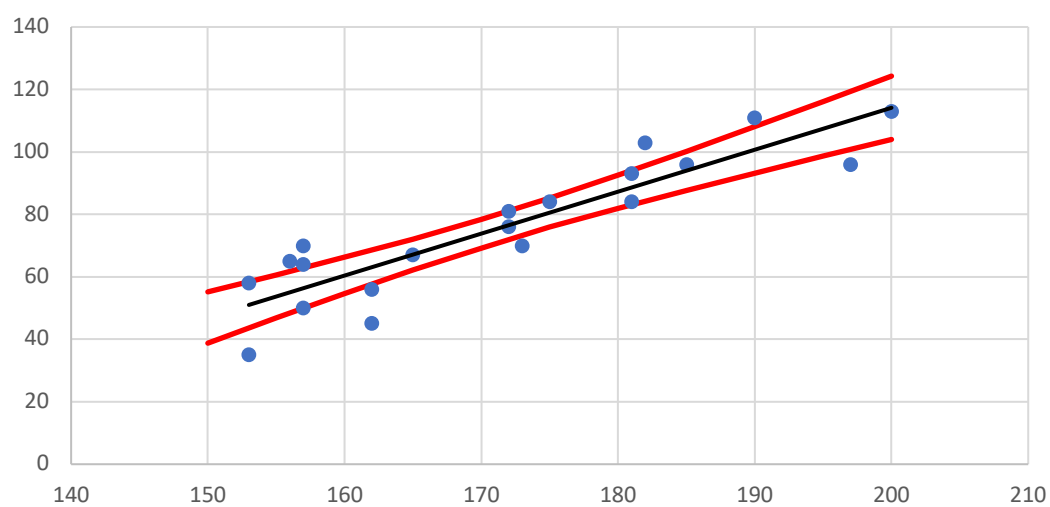
3. Vytvořte graf bodů spolu s regresní přímkou a pásem spolehlivosti pro individuální hodnotu výšky.

Výpočet pásu spolehlivosti

xi	yi	střední Y		individuální Y		h*
		dolní	horní	dolní	horní	
150	46,97472	38,76282702	55,18662029	25,07119411	68,8782532	0,163547
155	53,68990	46,74791718	60,63189123	32,2301618	75,1496466	0,116875
160	60,40508	54,57315982	66,23700967	39,2783077	81,53186179	0,082486
165	67,12027	62,13070221	72,10982838	46,21033703	88,03019355	0,060378
170	73,83545	69,26988364	78,40100804	53,02262003	94,64827165	0,050553
175	80,55063	75,87545755	85,22579521	59,71348254	101,3877702	0,053009
180	87,26581	81,98051829	92,55109556	66,28334673	108,2482671	0,067747
185	93,98099	87,72994154	100,2320334	72,7346952	115,2272797	0,094768
190	100,69617	93,26104182	108,1312942	79,07186545	122,3204706	0,134070
195	107,41135	98,66201402	116,1606831	85,30071278	129,5219843	0,185655
200	114,12653	103,9833129	124,2697453	91,42819888	136,8248593	0,249521



Pás spoľahlivosti pre strednú hodnotu



Pás spoľahlivosti pre individuálnu hodnotu

