

Domácí úkol ze statistiky (A)

1.

Jako statistický soubor jsem zvolil záznam o činnosti počítačového programu Think běžícího na mém počítači, který funguje jako platforma pro distribuovaný computing projektu [Find-A-Drug](#). Program porovnává množinu molekul s různými obměnami s cílovým proteinem a snaží se nalézt možnou interakci. V záznamu aplikace ukládá pro každou úlohu počet možných interakcí, cílový protein, množinu vyzkoušených molekul a počet bodů, vypočítaný na základě rychlosti procesoru a tráveného času.

Soubor obsahuje 66 řádků a několik znaků: pořadové číslo úlohy (Job Numer), cílový protein (Target), počet zásahů (Hits), počet zpracovaných molekul (Numer of Molecules) a počet bodů (Points). Jako sledované znaky jsem vybral atributy Target a Points, v případě Target jde o kvalitativní nominální znak, v případě Points jde o znak kvantitativní, s velkým počtem obměn.

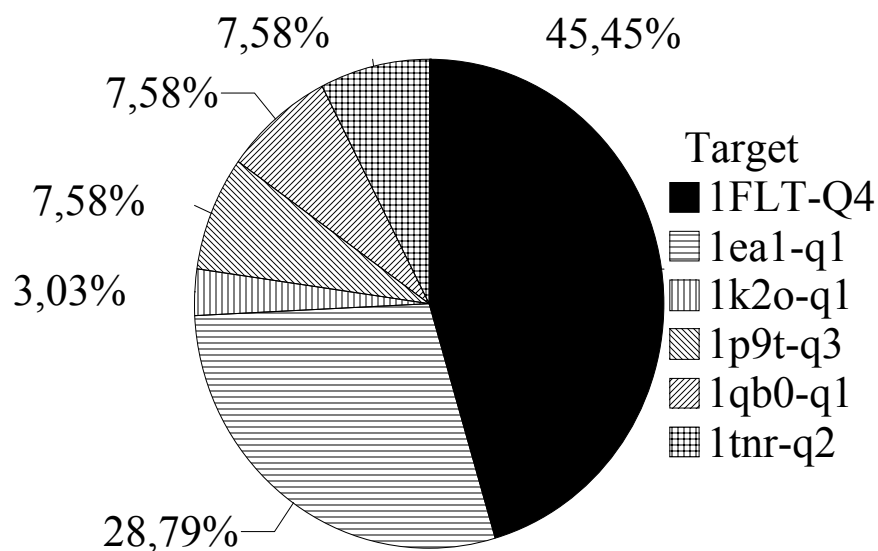
(Tabulka se zdrojovými daty se nachází v [Příloze 1.](#))

2.

Tabulka rozdělení hodnot pro Targets (cílový protein):

Var. Č.	Cíl (Target)	Frekvence	Relativní Frekvence	Kumulativní Frekvence	Kum. Rel. Frekvence
1	1FLT-Q4	30	0,4545	30	0,4545
2	1ea1-q1	19	0,2879	49	0,7424
3	1k2o-q1	2	0,0303	51	0,7727
4	1p9t-q3	5	0,0758	56	0,8485
5	1qb0-q1	5	0,0758	61	0,9242
6	1tnr-q2	5	0,0758	66	1,0000

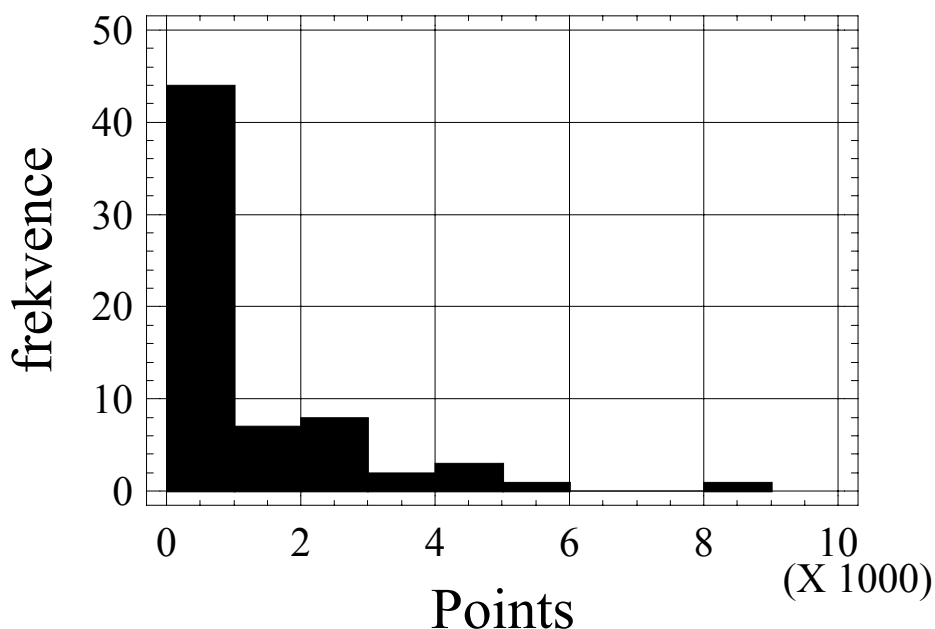
Výšečový graf atributu Targets:



Tabulka rozdělení hodnot pro atribut Points (body):

rozpětí Č.	spodní limit	horní limit	střed	frekvence	relativní frekvence	kumulativní frekvence	kum. rel. frekvence
na nebo pod		0,0		0	0,0000	0	0,0000
1	0,0	1000,0	500,0	44	0,6667	44	0,6667
2	1000,0	2000,0	1500,0	7	0,1061	51	0,7727
3	2000,0	3000,0	2500,0	8	0,1212	59	0,8939
4	3000,0	4000,0	3500,0	2	0,0303	61	0,9242
5	4000,0	5000,0	4500,0	3	0,0455	64	0,9697
6	5000,0	6000,0	5500,0	1	0,0152	65	0,9848
7	6000,0	7000,0	6500,0	0	0,0000	65	0,9848
8	7000,0	8000,0	7500,0	0	0,0000	65	0,9848
9	8000,0	9000,0	8500,0	1	0,0152	66	1,0000
nad	9000,0			0	0,0000	66	1,0000

Histogram pro Points (body):



Následující výpočty a rozbor se týkají atributu Points (body), tedy kvantitativního znaku s velkým počtem obměn.

3.

Jako vhodné charakteristiky úrovně se jeví modus, medián a z průměrů pak vzhledem k smyslu součtu bodů průměr aritmetický. Vzhledem k rozložení hodnot jsou zajímavé taky spodní a horní kvartily.

aritmetický průměr = 1259,47

medián = 598,5

modus = 599,0

spodní kvartil = 343,0

horní kvartil = 1345,0

Zajímavými z pohledu variability jsou tyto hodnoty:

variační rozpětí = 8202,0

kvartilové rozpětí = 1002,0

rozptyl = 2397000

směrodatná odchylka = 1548,23

variační koeficient = 122,927%

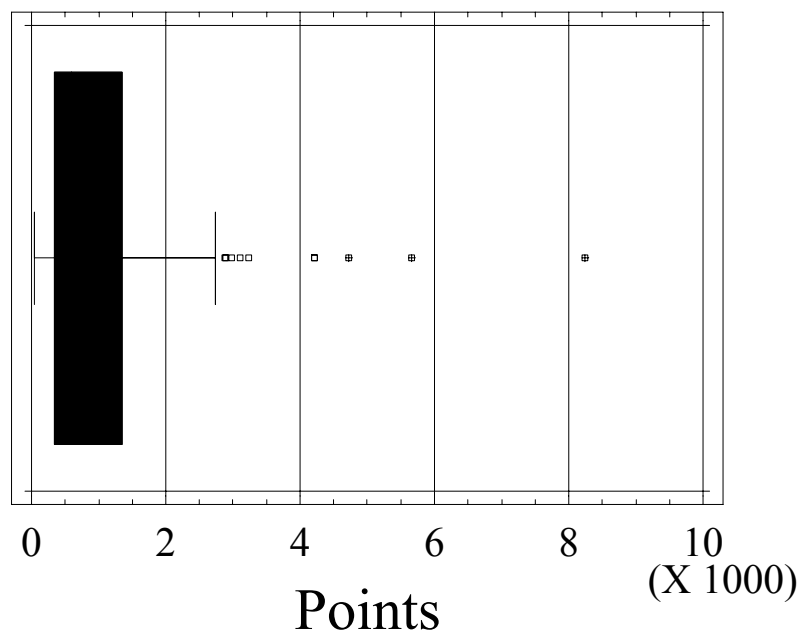
4.

Soubor vykazuje pro zkoumaný znak vysokou variabilitu, většina položek se nachází v intervalu od 0 do 1000 bodů, o tom vypovídají také hodnoty kvartilů a jejich rozpětí. Také

Aritmetický průměr je silně zkreslen horními extrémními hodnotami, které dosahují až takřka 8500 bodů. Jak již bylo naznačeno, je rozdělení hodnot nesymetrické, s výraznou převahou nízkých hodnot. Kolísání hodnot je veliké, o čemž vypovídá také spočítaný rozptyl, směrodatná odchylka a variační koeficient. Také vzhledem k tomu nemá aritmetický průměr velkou vypovídací hodnotu.

Zkoumaný soubor má velké variační rozpětí (8202,0), zatímco nejnižší hodnota je 46, nejvyšší hodnota činí 8248, přesto se většina hodnot nachází v intervalu do 1000 bodů, jak lze vidět na histogramu. Taktéž nejčastější hodnota proměnné, modus, a střední hodnota, medián, se nacházejí v tomto intervalu. O vysokém výskytu nízkého počtu bodů také svědčí hodnoty horního a dolního kvartilu a jejich rozpětí. Rozdělení hodnot názorně ilustruje také krabicový graf:

Krabicový graf



Lze vidět, že rozdělení hodnot je vysoce nesymetrické, především horní extrémní hodnoty se velmi odchyľují jak od průměru, tak také od střední hodnoty.

Kolísání hodnot je vzhledem k průměru vysoké, o čemž svědčí vysoká hodnota směrodatné odchylky a hlavně variačního koeficientu (122,927%). Vzhledem ke všem uvedeným skutečnostem nemá průměr velkou vypovídací hodnotu.

Příloha 1 (tabulka se zdrojovými daty):

Job Number	Target	Hits	CPU Time	Points
1	1tnr-q2	0	6688.34	336
2	1tnr-q2	0	5957.69	300
3	1tnr-q2	0	5182.19	303
5	1tnr-q2	0	7032.89	320
21	1FLT-Q4	113	14555.30	645
22	1FLT-Q4	155	63886.50	3104
23	1FLT-Q4	166	20121.10	984
24	1FLT-Q4	32	3677.77	193
25	1FLT-Q4	20	10212.30	499
26	1ea1-q1	0	12145.30	599
6	1tnr-q2	0	5358.09	254
27	1ea1-q1	0	14316.70	698
28	1ea1-q1	0	12690.20	612
29	1ea1-q1	1	9561.50	468
30	1ea1-q1	1	12236.40	616
31	1ea1-q1	0	8852.33	476
32	1ea1-q1	1	20906.50	1077
33	1ea1-q1	2	8358.56	447
34	1ea1-q1	1	7504.09	438
35	1ea1-q1	0	3867.50	192
36	1FLT-Q4	96	15479.20	726
37	1FLT-Q4	27	12521.40	599
38	1FLT-Q4	31	5492.00	282
39	1FLT-Q4	20	9385.63	450
40	1FLT-Q4	53	8342.11	393
41	1ea1-q1	3	4856.19	258
42	1ea1-q1	0	7943.67	400
43	1ea1-q1	0	22160.30	1113
44	1ea1-q1	5	27549.80	1342
45	1ea1-q1	0	12980.00	626
46	1ea1-q1	0	8564.20	455
47	1ea1-q1	0	5869.25	343
48	1ea1-q1	0	5507.80	346
49	1ea1-q1	0	6860.31	344
50	1p9t-q3	116	56343.30	2720
51	1p9t-q3	41	13767.50	598
52	1p9t-q3	128	106004.00	5663
53	1p9t-q3	204	164229.00	8248
54	1p9t-q3	147	97973.20	4722
55	1FLT-Q4	149	56850.80	2895
56	1FLT-Q4	23	3650.28	180
57	1FLT-Q4	153	12841.10	705
58	1FLT-Q4	174	45812.10	2457
59	1FLT-Q4	12	2414.39	113
61	1FLT-Q4	13	4637.00	217
62	1FLT-Q4	54	14851.20	712
63	1FLT-Q4	24	7010.38	347
64	1FLT-Q4	24	9529.53	434
60	1FLT-Q4	167	65130.00	3230
65	1FLT-Q4	146	23901.70	1175
66	1FLT-Q4	91	23415.10	1168
67	1FLT-Q4	12	5827.08	323
68	1FLT-Q4	49	9127.33	422
69	1FLT-Q4	172	56810.10	2877
70	1FLT-Q4	42	11030.30	579
71	1FLT-Q4	0	785.52	46
72	1FLT-Q4	38	4385.47	239
73	1FLT-Q4	10	2140.70	108
74	1FLT-Q4	146	27814.00	1345
84	1qb0-q1	96	80690.80	4212
85	1qb0-q1	66	35923.50	1914
86	1qb0-q1	138	49317.60	2742
87	1qb0-q1	66	77803.50	4215
88	1qb0-q1	131	55103.10	2975
89	1k2o-q1	103	56106.50	2724
90	1k2o-q1	74	51277.60	2582