Zvolte data pro

STATISTIKA

Seminární práce

Květen 2017

Pavel Majer

UNICORN COLLEGE

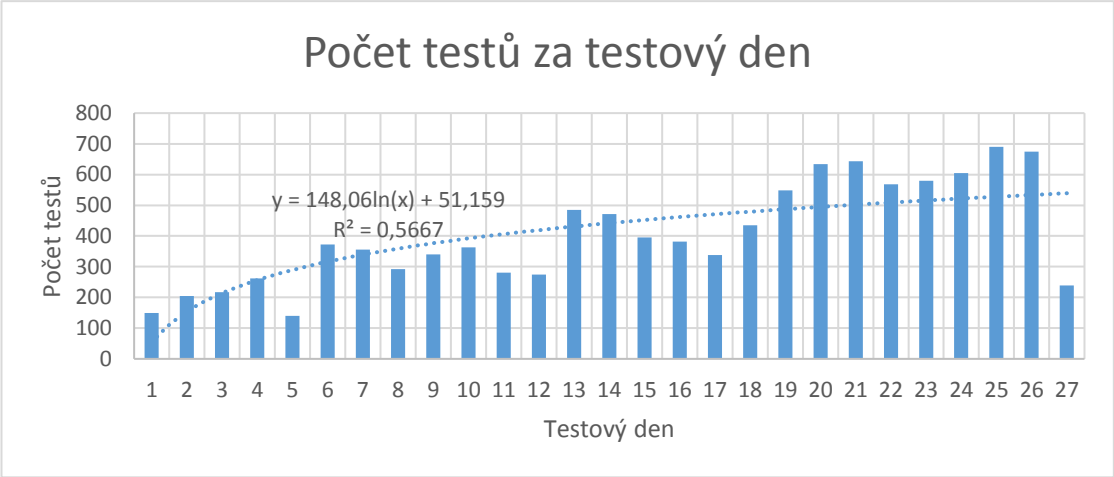
UNICORN COLLEGE

zpracování

Pro tento úkol jsem zvolil údaje, které zveřejňuje gymnázium Na Zatlance. Tato škola pomáhá studentům základních škol s přípravou na jednotné zkoušky Cermat.
Z dat je možné vyčíst průměrné výsledky žáků pátých tříd z dvaceti sedmi testových dnů. Hodnoty jsou rozdělené podle předmětů, lze rovněž vyčíst počet účastníků z každého testového dne.
http://www.zkousky-nanecisto.cz/modules.php?name=Tabule_nanecisto&rocnik=2&tabule=1&ak=195&menuzvol=pata

Úkol č.1: Všechny proměnné zpracujte pomocí popisné statistiky (včetně využití grafů).

Z dvaceti sedmi měření (kde každé jedno měření odpovídá průměrům z jednoho testovacího dne) jsme získali výběrový datový soubor. Tento datový soubor obsahuje data o průměrných výsledcích testů z matematiky, českého jazyka a všeobecného přehledu.

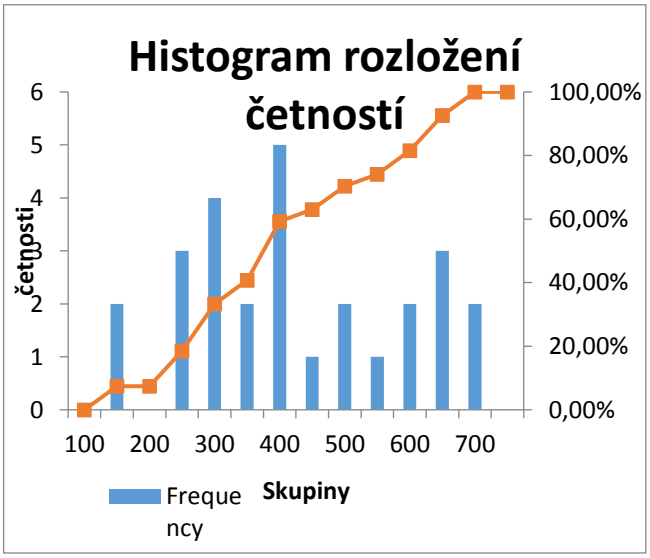


V různých testových dnech se účastnil různý počet studentů. Tato informace je pak dále uvažována při výpočtu celkového průměru napříč všemi měřeními. Z rozložení množství testů po jednotlivých dnech je zřejmé, že se nejedná o Normální (Gaussovo) rozložení.

Z níže histogramu rovněž nelze vysledovat znaky Normálního (Gaussova) rozložení. Je zřejmé, že naměřené míry mají spíše stoupající tendenci. Rovněž lze z grafu vysledovat informaci, že nejčastější množství provedených testů za jeden testový den je v rozmezí mezi 350 a 400 a maximální množství účastníků bylo posledních 6 týdnů před zkouškami (když nepočítáme poslední zkouškový termín)

Histogram počtů testů (rozdělení po 50)

Skupina	*Četnosti*	*Kumul. %*
100	0	0,00%
150	2	7,41%
200	0	7,41%
250	3	18,52%
300	4	33,33%
350	2	40,74%
400	5	59,26%
450	1	62,96%
500	2	70,37%
550	1	74,07%
600	2	81,48%
650	3	92,59%
700	2	100,00%
Více	0	100,00%



UNICORN COLLEGE

Výběrový soubor dat má následující parametry

Proměnná	Způsob výpočtu	Počty	MAT	ČJ	VP
Modus ($\hat{x}$) - nejčastější (zaokrouhlená) průměrná míra pro den	V xls: MODE(<sloupec se zaokrouhlenými hodnotami>)	300	19	21	19
Medián ($\tilde{x}$) – střední hodnota	V xls: MEDIAN(<sloupec>) (TBD OK bez uvažování počtů testů?)	372	19,46	20,74	20,4
Aritmetický průměr ($\bar{x}$)	V xls: součet všech bodů z každého předmětu vyděleno počtem testů Vzorec: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	405,18	21,32	22,26	20,77
Variační rozpětí – rozdíl nejvyšší hodnoty od nejnižší	V xls: (MAX(<sloupce>)) - MIN(<sloupce>) R= max(x) – min (x)	550	22,00	25,95	4,81
Výběrový rozptyl - průměr druhých mocnin vzdáleností od průměru	V xls: Data analysis-> descriptive statistics výběrový rozptyl $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	27377,23	31,75	29,49	2,18
Směrodatná odchylka	V xls: Data analysis-> descriptive statistics $\sigma = \sqrt{\mathrm{Var}(X)}$	165,45	5,63	5,43	1,47

Úkol č.2: Navrhněte hypotézu pro testování (střední hodnota, relativní četnost) a tuto hypotézu otestujte pomocí vhodně zvoleného statistického testu.

Dá se předpokládat, že průměrné výsledky žáků, kteří se připravují na přijímací zkoušky na osmiletá gymnázia budou mít lepší výsledky z matematiky než průměrné výsledky z českého jazyka. Stejní žáci se v každém testovém dni vždy účastní obou testů.

HO: Průměrné hodnoty výsledků z MAT jsou vyšší než průměrné výsledky testů z ČJ.
H1: non H0

V Různých dnech se účastnil jiný počet žáků. Vždy se však titíž žáci účastnili obou testů. V následující tabulce jsou znázorněny průměry z obou předmětů rozdělené po dnech, ale i celkové součty se zohledněním počtu testů.

Den	Počet testů	Průměr MA	Průměr ČJ	Celkem bodů MA	Celkem bodů ČJ	MA/ČJ
1	149	23,17	17,98	3452,33	2679,02	28,87%
2	205	18,84	20,74	3862,2	4251,7	-9,16%
3	217	18,56	21,18	4027,52	4596,06	-12,37%
4	262	20,19	19,92	5289,78	5219,04	1,36%
5	140	18,85	20,42	2639	2858,8	-7,69%
6	372	18,30	18,17	6807,6	6759,24	0,72%
7	356	21,69	19,33	7721,64	6881,48	12,21%
8	292	21,70	15,70	6336,4	4584,4	38,22%
9	340	20,62	19,48	7010,8	6623,2	5,85%
10	363	21,68	19,97	7869,84	7249,11	8,56%
11	281	19,46	22,90	5468,26	6434,9	-15,02%
12	274	19,27	19,23	5279,98	5269,02	0,21%
13	485	23,43	16,53	11363,55	8017,05	41,74%
14	471	16,74	21,25	7884,54	10008,75	-21,22%
15	395	21,27	18,27	8401,65	7216,65	16,42%
16	382	18,47	21,25	7055,54	8117,5	-13,08%
17	338	19,82	21,82	6699,16	7375,16	-9,17%
18	435	16,02	21,17	6968,7	9208,95	-24,33%
19	549	16,00	19,61	8784	10765,89	-18,41%
20	634	17,07	23,55	10822,38	14930,7	-27,52%
21	643	16,84	19,59	10828,12	12596,37	-14,04%
22	568	18,58	21,30	10553,44	12098,4	-12,77%
23	580	16,54	22,55	9593,2	13079	-26,65%
24	605	21,07	21,05	12747,35	12735,25	0,10%
25	690	31,59	31,11	21797,1	21465,9	1,54%
26	675	36,81	33,47	24846,75	22592,25	9,98%
27	239	38,00	41,65	9082	9954,35	-8,76%
Celkem				**233192,83**	**243568,14**	**-4,26%**

Jelikož jde o velký výběr, lze podle Centrální limitní věty použít test pro obecné rozdělení. Proto bude pro ověření hypotézy použit následující vzorec:

Střední hodnota, obecné rozdělení, velký výběr

H_0	H_1	Testové kritérium	Kritický obor
$E(X)=\mu_0$	$E(X)>\mu_0$ $E(X)<\mu_0$ $E(X)\neq\mu_0$	σ^2 neznámé $(n>30)$ $U=\frac{\bar{x}-\mu_0}{s'_x}\sqrt{n}$ $U\approx N[0;1]$	$W_\alpha=\{U\geq u_{1-\alpha}\}$ $W_\alpha=\{U\leq -u_{1-\alpha}\}$ $W_\alpha=\{\|U\|\geq u_{1-\alpha/2}\}$

(průměr z ČJ) μ_0=22,26; (průměr z MA) $\bar{X}$=21,32; (směrodatná odch x; stdev) Sx=5,6; n=27

H0 : E(X) = μ_0 H1 : E(X) < μ_0 => u_α = { U ≤ $u_{1-\alpha}$ }

α = 0,05 => 1- α = 0,95

$$U=\frac{\bar{x}-\mu_0}{s'_x}\sqrt{n}$$ U= (21.32-22.26) / 5.63 * 27^1/2 = -0.87

-0.87 ≤ 1.645 => platí H1, hypotézu H0 zamítáme

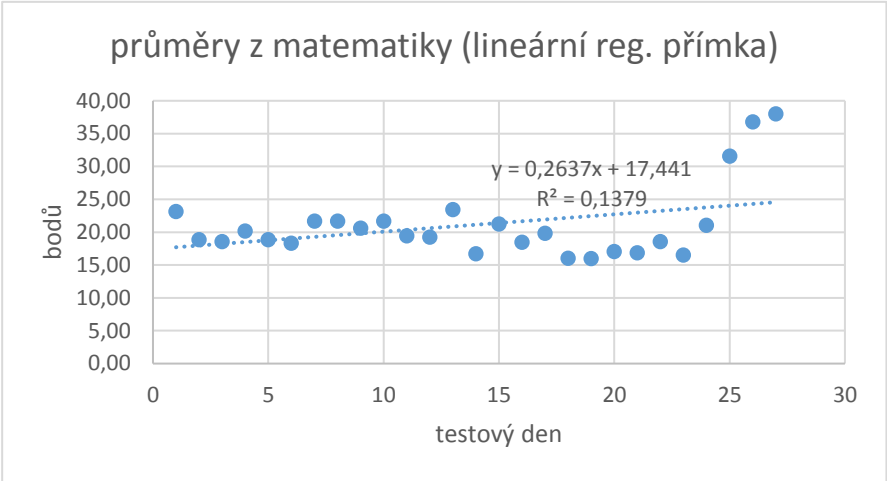
Výsledky z matematiky nejsou lepší než výsledky s českého jazyka s pravděpodobností 95%.

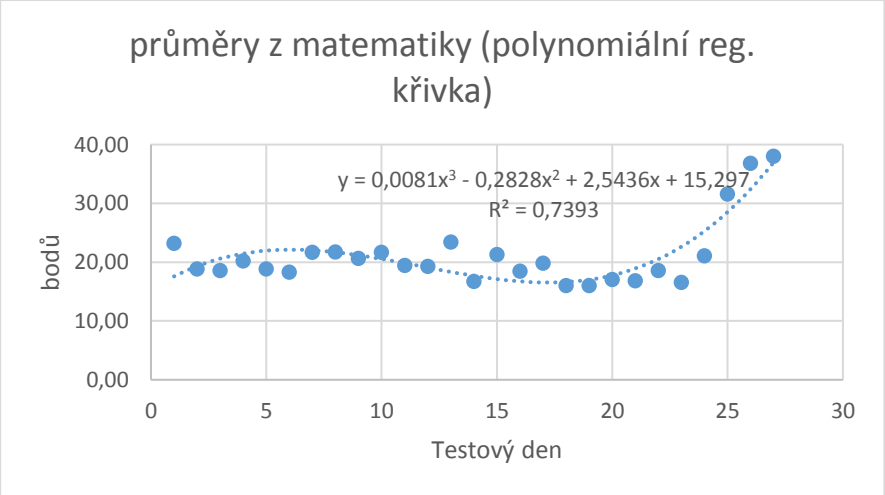
Úkol č.3: Zhodnoťe závislost dvou vybraných proměnných, kdy jedna musí mít podobu kategoriální proměnné. Využijte vhodných metod.

Testový den je kategoriální proměnná, nabývá hodnot 1 až 27 (ve skutečnosti pořadí reprezentuje den, kdy k testu došlo), průměrné výsledky testů z MA se dá považovat za kvantitativní proměnnou, nabývá hodnot 0-50

Cílem je ověřit, zda se jedná se o regresi v závislosti na čase, kde vysvětlovaná proměnná je pořadí testu a vysvětlující proměnná je výsledek z matematiky.

Den	testů	průměr MA
1	149	23,17
2	205	18,84
3	217	18,56
4	262	20,19
5	140	18,85
6	372	18,30
7	356	21,69
8	292	21,70
9	340	20,62
10	363	21,68
11	281	19,46
12	274	19,27
13	485	23,43
14	471	16,74
15	395	21,27
16	382	18,47
17	338	19,82
18	435	16,02
19	549	16,00
20	634	17,07
21	643	16,84
22	568	18,58
23	580	16,54
24	605	21,07
25	690	31,59
26	675	36,81
27	239	38,00
Celkem	**10940**	





Za použití statistických funkcí v MS Excel lze potvrdit lineární závislost výsledků z matematiky na čase. Přímka má rovnici y = 0,2637x + 17,441 s významností modelu (index determinace R^2=0,1379)

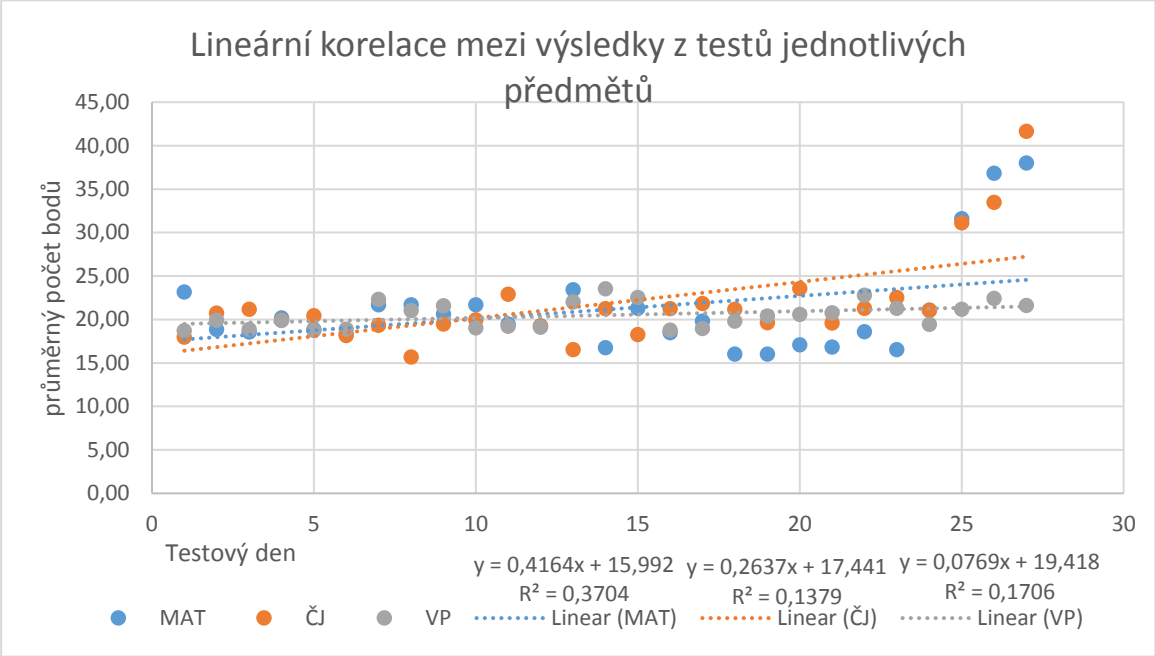
(Také i) s pomocí statistických funkcionalit v programu MS Excel lze nalézt přesnější model. Ten má polynomickou rovnici y=0.0081x3 – 0,2828x2 + 2,5436x + 15,29 a významnost R^2 je 0,7393 a pro výpočty odhadů bude vhodnější tato rovnice.

Výsledek: Směrnice výsledků z matematiky v závislosti na čase je lineární, daná kladná daná regresní přímkou y = 0,2637x + 17,441, přičemž kladná hodnota směrnice nám udává, že jde o přímou závislost a spíše volnější závislost (R^2 = 0,13)

Úkol č.4: Zhodnoťte závislost dvou vybraných kvantitativních proměnných.

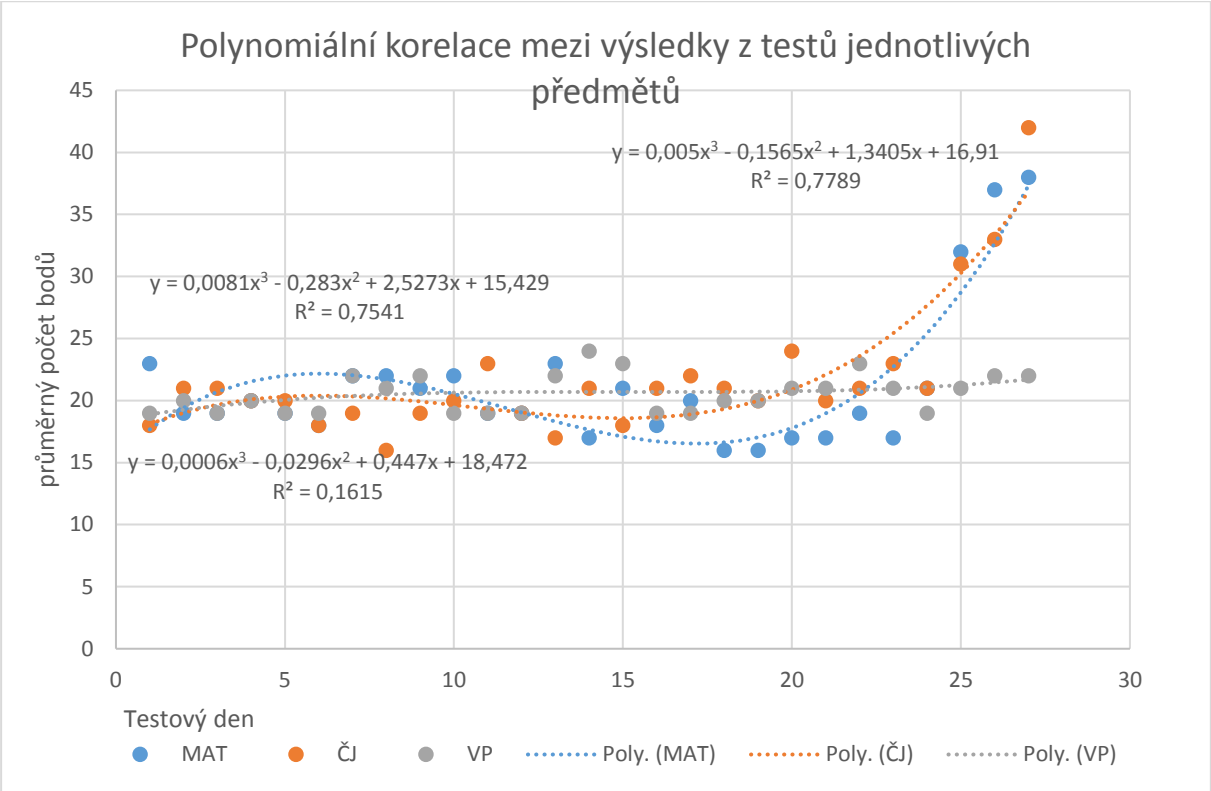
Průměrné výsledky testů z každého předmětu a dne jsou kvantitativní proměnné, nabývají hodnot 0 až 50. Máme k dispozici průměrné výsledky z třech předmětů za každý testový den. Dá se předpokládat, že postupem času se výsledky testů budou rovnoměrně zlepšovat a že tak budou spolu korelovat.

Na následujícím grafu je zřejmý rostoucí trend průměrných výsledků, ale přesnost modelu s lineární tendenční křivkou.



R^2 (koeficient determinace) v modelu značí, jaká je statistická míra vzdálenosti bodů od tendenční křivky a určuje tak přesnost modelu. Při volbě testování lineární závislosti není model moc přesný (R^2 je blíže k 0, než k extrémům)

Pro rozložení hodnot se zdá být vhodnější polynomiálání model tendenční křivky, který pro testy z matematiky a českého jazyka zvýšil přesnost modelu na více než 0.75.



Korelační koeficienty

Korelační koeficient nabývá hodnot mezi -1 a 1. Čím více se blíží korelační koeficient k extrémním hodnotám, tím více spolu tyto proměnné na sobě statisticky závislé. Metoda rozměrového efektu zjednodušeně definuje tak, že pokud je r=0.1, pak se účinek vyhodnocuje jako malý, pokud je okolo 0.3, pak je střední, a pokud je nad 0.5, pak je velký.

Korelační koeficienty, vypočtené z průměrů všech tří předmětů pomocí statistických funkcí MS excel:

	MAT	*ČJ*	*VP*
MAT	1		
ČJ	0,792829448	1	
VP	0,287167578	0,230081988	1

Test: Otestujte na hladině významnosti p = 0,05, zda u výsledků z MA a ČJ, může jít o lineární závislost.

Hypotézy:

H0 : mezi proměnnými ČJ a MAT neexistuje lineární vztah
H1 : non H0

Hodnoty r=0,7928 a n=27 vložíme do testového kritéria.

korelační koeficient:

H_0	H_1	Testové kritérium	Kritický obor
$\rho_{yx} = 0$	$\rho_{yx} \neq 0$	$t = \frac{r_{yx}\sqrt{n-2}}{\sqrt{1-r_{yx}^2}}$ $t \sim t[n-2]$	$W_\alpha = \{\lvert t \rvert \geq t_{1-\alpha/2}\}$

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2}$$

t=0,792829448/((1-0,792829448^2)$^{1/2}$ * (25)$^{1/2}$ =

t =0,792829448/(SQRT(1-0,7928294482)*SQRT(25)) = **0,348374**

Kritický obor:

α = 0,05 => 1- α/2 = 0,975

$t_{0,05}$(27-2) =TINV(0,05;25) = **2,060**

$t < t_{1-\alpha/2}$

0,348374 < 2,060

Závěr testu:
Hodnota testovacího kritéria nepřekročila kritickou hodnotu.
Není nutno zamítnout hypotézu o lineární nezávislosti *výsledků z ČJ a MA (t.j. výsledky mohou být lineárně závislé)*

Tabulka VI. Kvantily rozdělení t

	P				
ν	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,553	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750

Vyhodnocení: Z tabulky korelačních koeficientů je zřejmé, že výsledky z matematiky a českého jazyka spolu korelují s velkým účinkem (0.78).

Zdroje

1. **Základní zdroj**
   zpracovaná data, které zveřejňuje gymnázium Na Zatlance:
   http://www.zkousky-nanecisto.cz/modules.php?name=Tabule_nanecisto&rocnik=2&tabule=1&ak=195&menuzvol=pata

			Průměrný počet bodů			Zaokrouhlené průměry				Bodů celkem		
Den	počet testů	%	MAT	ČJ	VP	testů	MAT	ČJ	VP	MAT	ČJ	VP
1	149	1,36%	23,17	17,98	18,71	100	23	18	19	3452	2679	2788
2	205	1,87%	18,84	20,74	19,93	200	19	21	20	3862	4252	4086
3	217	1,98%	18,56	21,18	18,85	200	19	21	19	4028	4596	4090
4	262	2,39%	20,19	19,92	19,87	300	20	20	20	5290	5219	5206
5	140	1,28%	18,85	20,42	18,74	100	19	20	19	2639	2859	2624
6	372	3,40%	18,3	18,17	18,89	400	18	18	19	6808	6759	7027
7	356	3,25%	21,69	19,33	22,31	400	22	19	22	7722	6881	7942
8	292	2,67%	21,7	15,7	21,01	300	22	16	21	6336	4584	6135
9	340	3,11%	20,62	19,48	21,59	300	21	19	22	7011	6623	7341
10	363	3,32%	21,68	19,97	19,03	400	22	20	19	7870	7249	6908
11	281	2,57%	19,46	22,9	19,21	300	19	23	19	5468	6435	5398
12	274	2,50%	19,27	19,23	19,1	300	19	19	19	5280	5269	5233
13	485	4,43%	23,43	16,53	22,03	500	23	17	22	11364	8017	10685
14	471	4,31%	16,74	21,25	23,52	500	17	21	24	7885	10009	11078
15	395	3,61%	21,27	18,27	22,53	400	21	18	23	8402	7217	8899
16	382	3,49%	18,47	21,25	18,78	400	18	21	19	7056	8118	7174
17	338	3,09%	19,82	21,82	18,97	300	20	22	19	6699	7375	6412
18	435	3,98%	16,02	21,17	19,81	400	16	21	20	6969	9209	8617
19	549	5,02%	16	19,61	20,4	500	16	20	20	8784	10766	11200
20	634	5,80%	17,07	23,55	20,59	600	17	24	21	10822	14931	13054
21	643	5,88%	16,84	19,59	20,76	600	17	20	21	10828	12596	13349
22	568	5,19%	18,58	21,3	22,8	600	19	21	23	10553	12098	12950
23	580	5,30%	16,54	22,55	21,27	600	17	23	21	9593	13079	12337
24	605	5,53%	21,07	21,05	19,44	600	21	21	19	12747	12735	11761
25	690	6,31%	31,59	31,11	21,16	700	32	31	21	21797	21466	14600
26	675	6,17%	36,81	33,47	22,44	700	37	33	22	24847	22592	15147
27	239	2,18%	38	41,65	21,61	200	38	42	22	9082	9954	5165
Celkem	10940	100,00%	-	-	-	-	-	-	-	233193	243568	227205
Průměr	-	-	-	-	-	-	-	-	-	21	22	21
Medián	-	-	19,46	20,74	20,4	-	-	-	-	-	-	-
Modus	-	-	-	-	-	300	19	21	19	-	-	-
Rozpětí (průměrů)			22	25,95	4,81	-	-	-	-	-	-	-

2. Další zdroje

Statistika pro flákače

http://www.volko.cz/new/chikvadrat_v_excelu.php

http://www.hazardni-hry.eu/statistika/anglictina.html

http://k101.unob.cz/~neubauer/pdf/regresni_analyza.pdf

https://homen.vsb.cz/~oti73/cdpast1/KAP11/KAP12.HTM

http://homel.vsb.cz/~dor028/Regrese.pdf

https://www.ncsu.edu/labwrite/res/gt/gt-reg-home.html