Cvičení 7: Binární logistická regrese

Příklad: V roce 2014 konalo státní závěrečné zkoušky bakalářského studia na jisté fakultě 167 studentů. U každého studenta bylo zaznamenáno jeho pohlaví (0 – žena, 1 – muž), občanství (1 – ČR, 2 – SR), studijní průměr za celou dobu studia, typ absolvované střední školy (1 – gymnázium, 2 – střední průmyslová škola či obchodní akademie, 3 – ostatní typy středních škol s maturitou) a úspěch u SZZ (1 – uspěl, 2 – neuspěl).

1. Vytvořte četnostní tabulky a nakreslete vhodné grafy pro kategoriální proměnné pohlaví, občanství, typ_SŠ, úspěch.

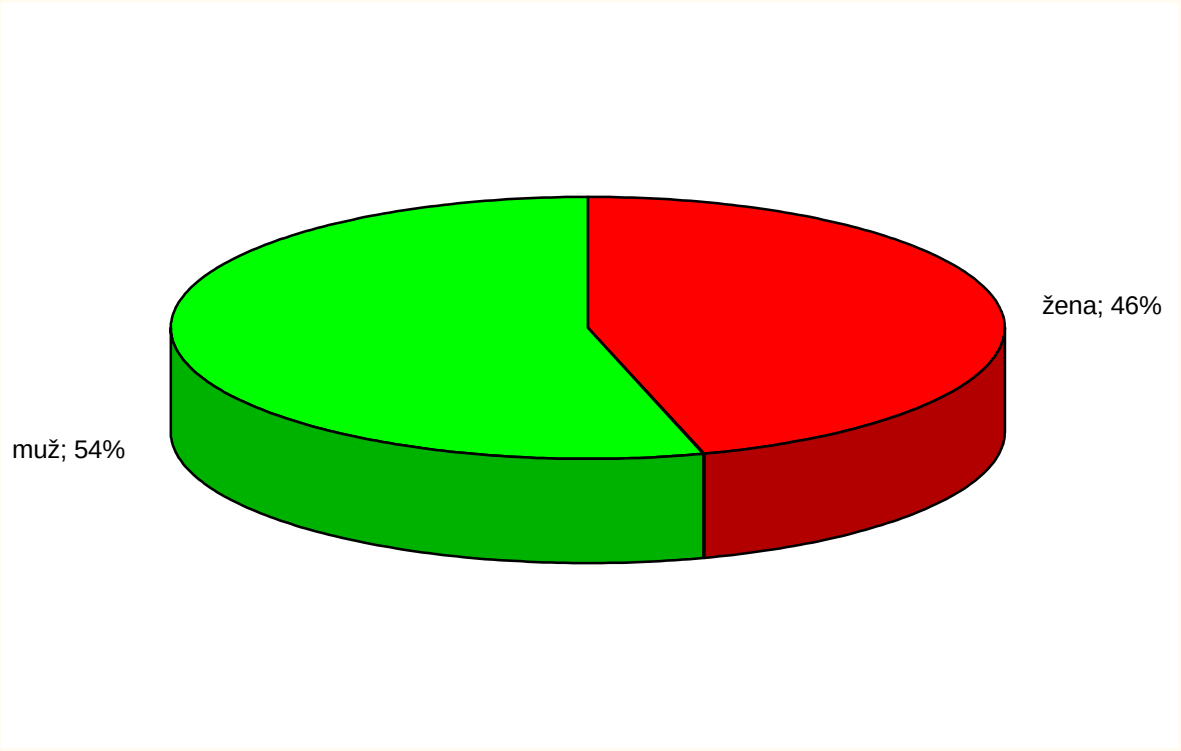
	Tabulka četností:pohlavi (SZZ.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
žena	76	76	45,50898	45,5090
muž	91	167	54,49102	100,0000

	Tabulka četností:obcanstvi (SZZ.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
Česká republika	136	136	81,43713	81,4371
Slovensko	31	167	18,56287	100,0000

	Tabulka četností:typ SS (SZZ.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
gymnázium	138	138	82,63473	82,6347
SPŠ+OA	13	151	7,78443	90,4192
ostatní	16	167	9,58084	100,0000

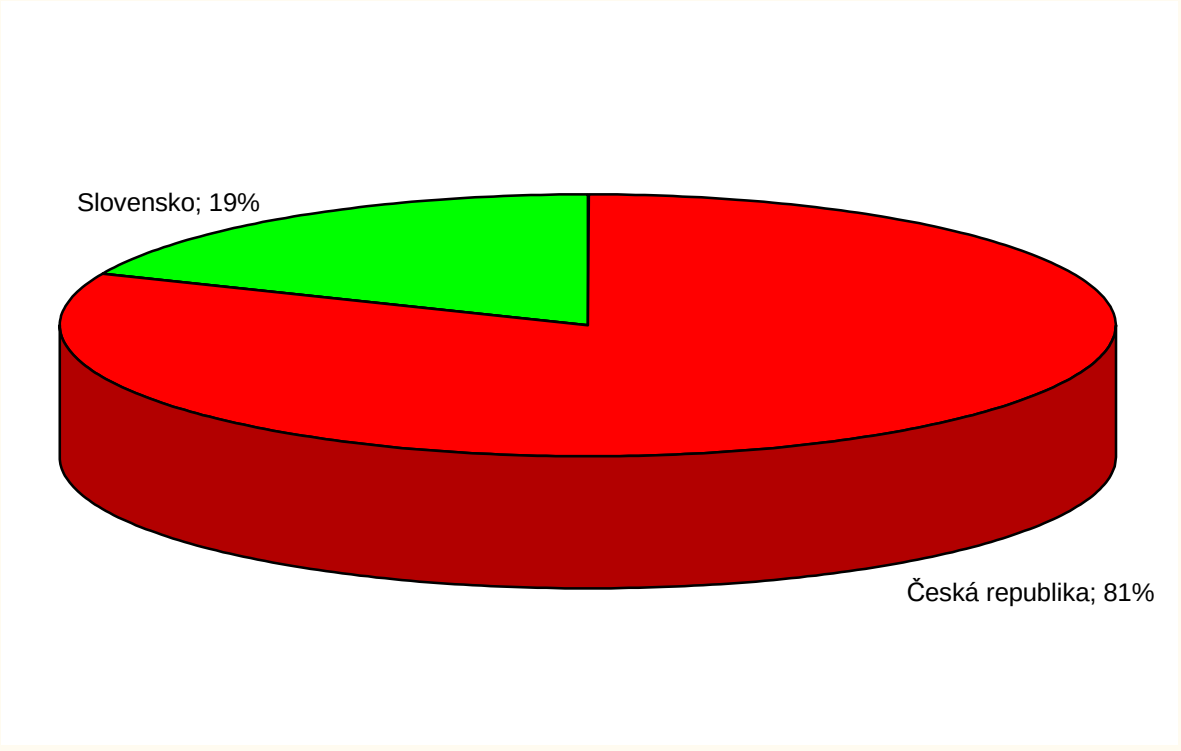
	Tabulka četností:uspech (SZZ.sta)			
Kategorie	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
uspěl	78	78	46,70659	46,7066
neuspěl	89	167	53,29341	100,0000

Výsečový graf z pohlavi
SZZ.sta 5v*167c



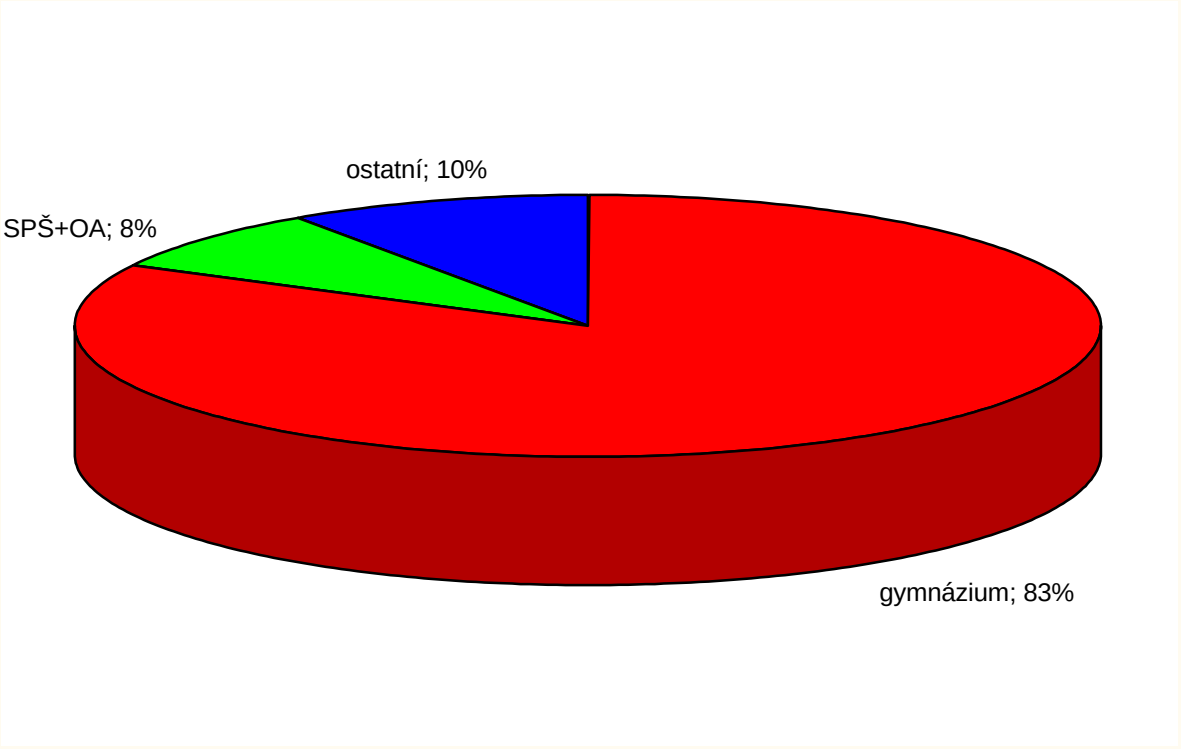
pohlavi

Výsečový graf z obcanstvi
SZZ.sta 5v*167c



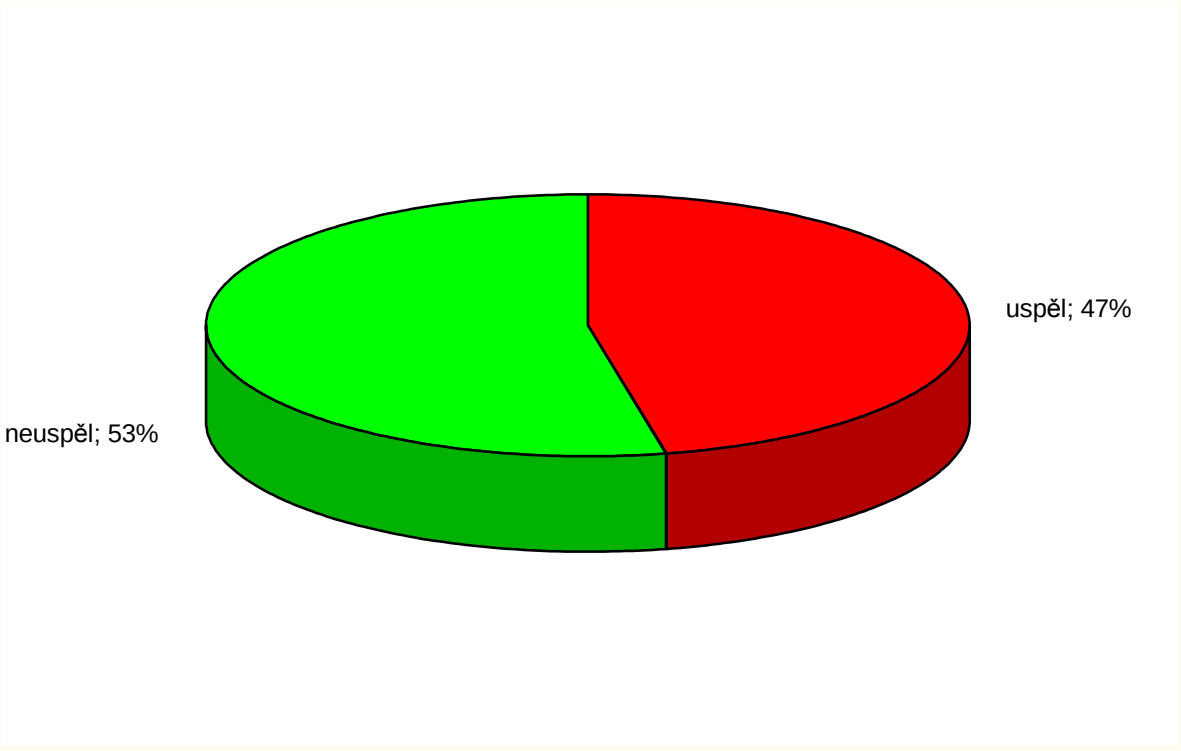
obcanstvi

Výsečový graf z typ SS
SZZ.sta 5v*167c



typ SS

Výsečový graf z uspech
SZZ.sta 5v*167c



uspech

2. Vypočtěte číselné charakteristiky proměnné průměr. A to pro celý soubor a pak pro studenty roztříděné podle pohlaví, občanství, typu SŠ a úspěchu u SZZ. Výpočty doplňte krabicovými diagramy. Vždy na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty (resp. mediány) studijního průměru jsou stejné v různých skupinách studentů. Ověřte normalitu proměnné průměr v daných skupinách studentů. Výpočty doplňte krabicovými diagramy.

Výsledky pro všechny studenty:

	Popisné statistiky (SZZ.sta)					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	167	2,912216	2,940000	1,060000	4,000000	0,838585

Výsledky pro ženy:

	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v1=0					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	76	2,889079	2,855000	1,060000	4,000000	0,819669

Výsledky pro muže:

	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v1=1					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	91	2,931538	3,100000	1,130000	4,000000	0,858108

Výsledky pro občany ČR:

	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v2=1					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	136	3,015735	3,170000	1,060000	4,000000	0,859049

Výsledky pro občany SR:

	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v2=2					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	31	2,458065	2,490000	1,130000	3,560000	0,555538

Výsledky pro absolventy gymnázií:

	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v4=1					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	138	2,841377	2,845000	1,060000	4,000000	0,850857

Výsledky pro absolventy středních průmyslových škol či obchodních akademií:

	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v4=2					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	13	2,961538	2,940000	1,360000	4,000000	0,776261

Výsledky pro absolventy jiných typů středních škol:

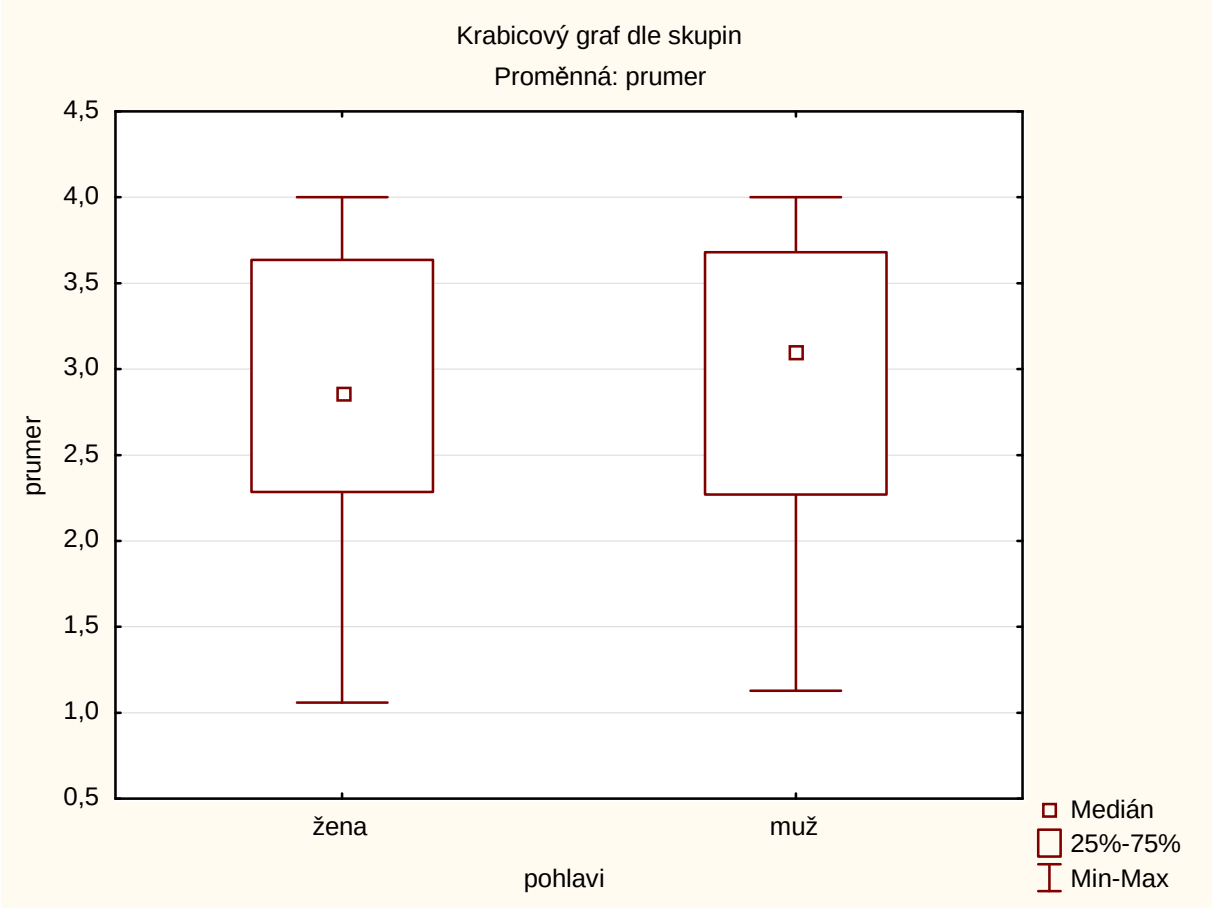
	Popisné statistiky (SZZ.sta) Zhrnout podmínku: v4=3					
Proměnná	N platných	Průměr	Medián	Minimum	Maximum	Sm.odch.
prumer	16	3,483125	3,685000	2,440000	4,000000	0,540552

Upozornění: Normalita proměnné průměr je ve většině případů porušena závažnějším způsobem, proto použijeme neparametrické testy.

Výsledky dvouvýběrového Wilcoxonova testu pro muže a ženy:

	Mann-Whitneyův U Test (w/ oprava na spojitost) (SZZ.sta) Dle proměn. pohlavi Označené testy jsou významné na hladině p <,05000								
Proměnná	Sčt poř. žena	Sčt poř. muž	U	Z	p-hodn.	Z upravené	p-hodn.	N platn. žena	N platn. muž
prumer	6273,500	7754,500	3347,500	-0,353510	0,723707	-0,354191	0,723196	76	91

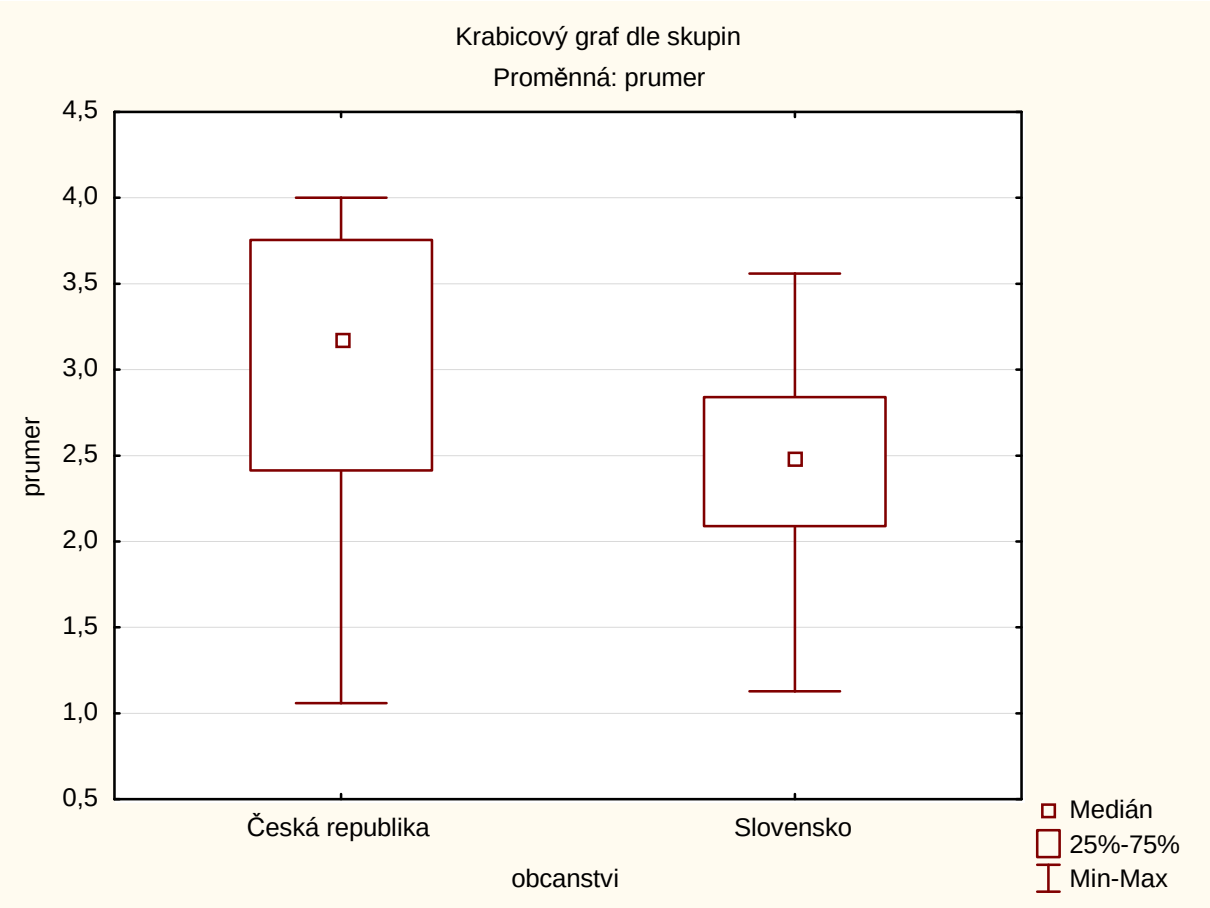
Na hladině významnosti 0,05 se neprokázal rozdíl v průměrném prospěchu mezi muži a ženami.



Výsledky dvouvýběrového Wilcoxonova testu pro Čechy a Slováky:

	Mann-Whitneyův U Test (w/ oprava na spojitost) (SZZ.sta) Dle proměn. obcanstvi Označené testy jsou významné na hladině p <,05000									
Proměnná	Sčt poř. Česká republika	Sčt poř. Slovensko	U	Z	p-hodn.	Z upravené	p-hodn.	N platn. Česká republika	N platn. Slovensko	2*1str. přesné p
prumer	12320,50	1707,500	1211,500	3,688024	0,000226	3,695133	0,000220	136	31	0,000168

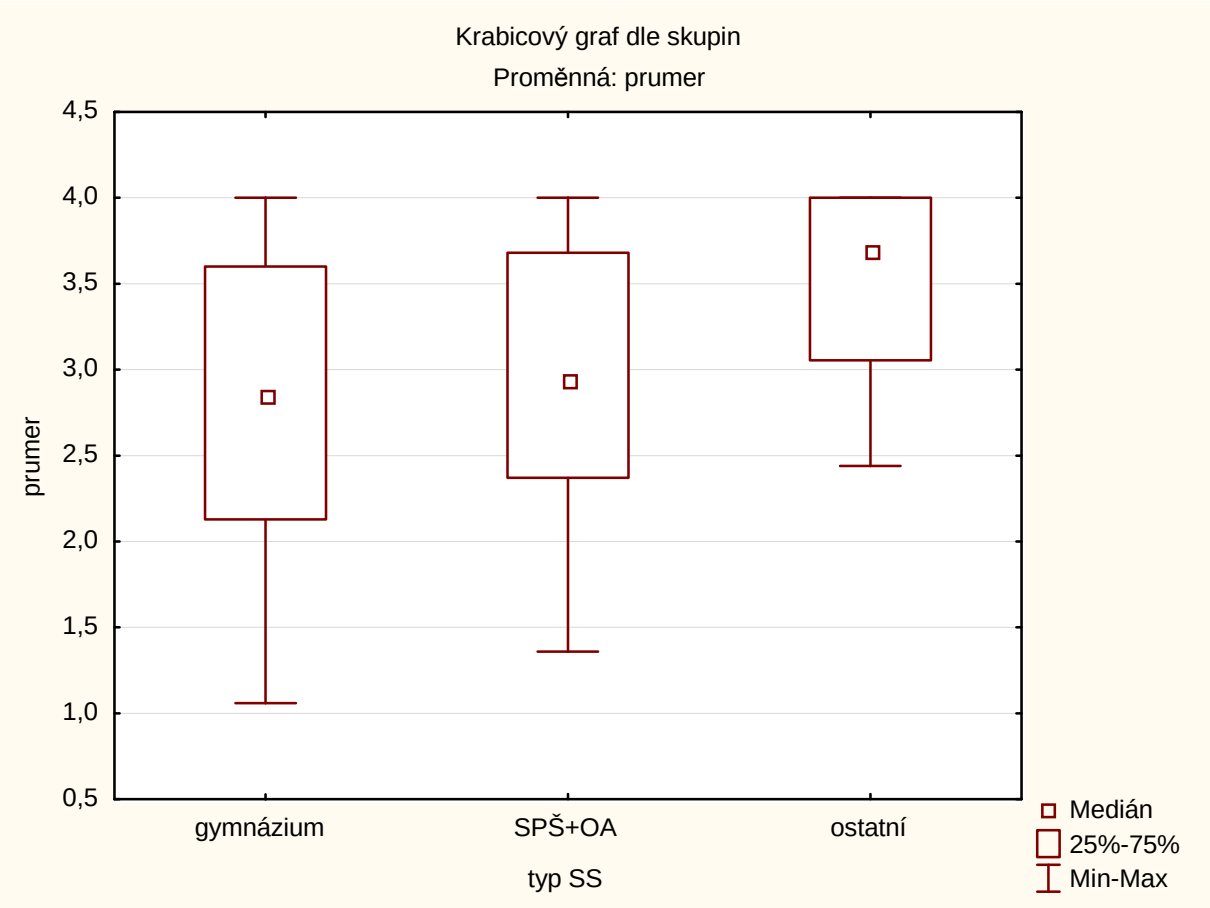
Na hladině významnosti 0,05 se prokázal rozdíl v průměrném prospěchu mezi Čechy a Slováky.



Výsledky Kruskalova – Wallisova testu pro absolventy různých typů středních škol:

	Kruskal-Wallisova ANOVA založ. na poř.; prumer (SZZ.sta) Nezávislá (grupovací) proměnná : typ SS Kruskal-Wallisův test: H (2, N= 167) =8,793145 p =,0123			
Závislá: prumer	Kód	Počet platných	Součet pořadí	Prům. Pořadí
gymnázium	1	138	11033,50	79,9529
SPŠ+OA	2	13	1111,00	85,4615
ostatní	3	16	1883,50	117,7188

Na hladině významnosti 0,05 se prokázal rozdíl v průměrném prospěchu mezi absolventy různých typů středních škol.



Výsledky metody mnohonásobného porovnávání:

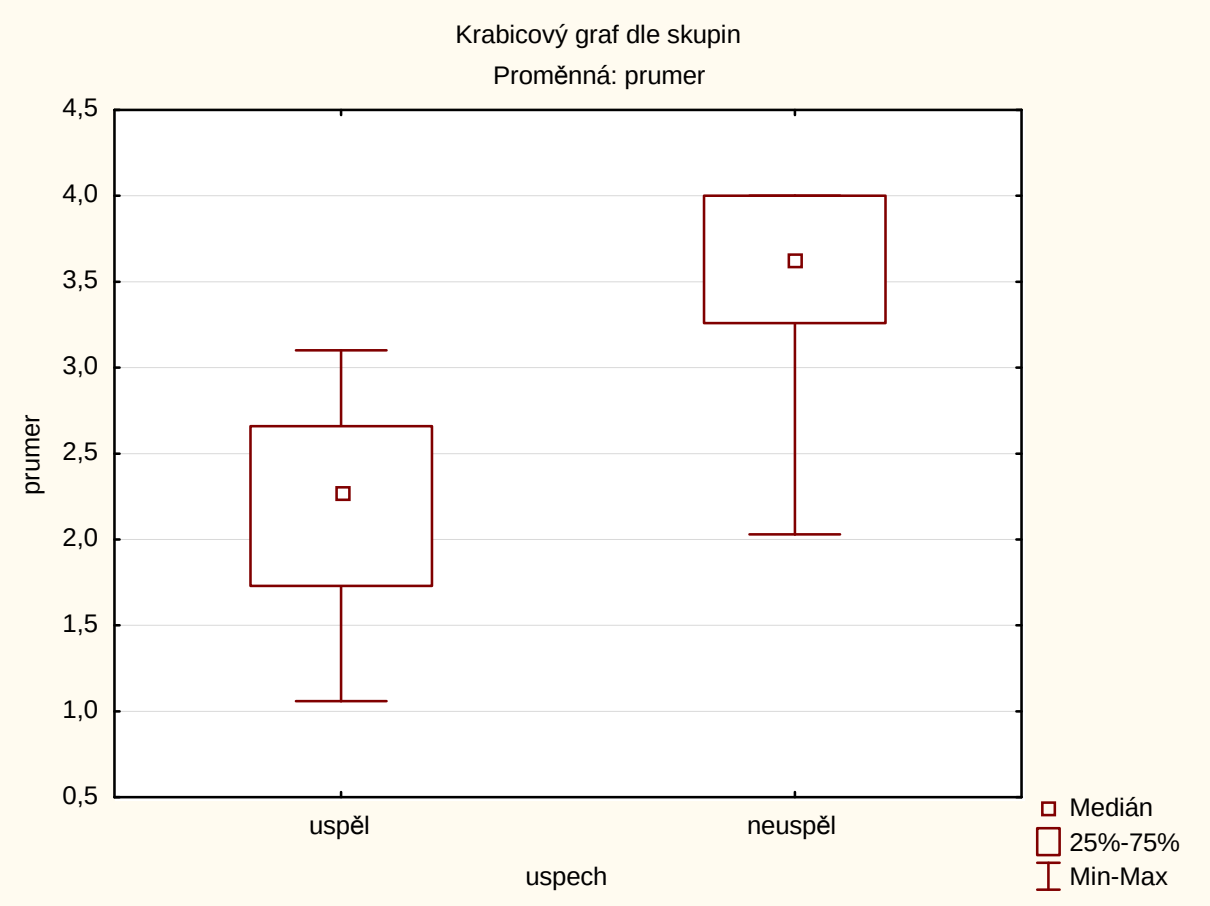
Závislá: prumer	Vícenásobné porovnání p hodnot (oboustr.); prumer (SZZ.sta) Nezávislá (grupovací) proměnná : typ SS Kruskal-Wallisův test: H (2, N= 167) =8,793145 p =,0123		
	gymnázium R:79,953	SPŠ+OA R:85,462	ostatní R:117,72
gymnázium		1,000000	0,009306
SPŠ+OA	1,000000		0,221987
ostatní	0,009306	0,221987	

Na hladině významnosti 0,05 se prokázal rozdíl v průměrném prospěchu absolventů gymnázií a absolventů středních škol odlišných od středních průmyslových škol a obchodních akademií.

Výsledky dvouvýběrového Wilcoxonova testu pro úspěšné a neúspěšné studenty:

Proměnná	Mann-Whitneyův U Test (w/ oprava na spojitost) (SZZ.sta) Dle proměn. uspech Označené testy jsou významné na hladině p <,05000								
	Sčt poř. uspěl	Sčt poř. neuspěl	U	Z	p-hodn.	Z upravené	p-hodn.	N platn. uspěl	N platn. neuspěl
prumer	3396,000	10632,00	315,0000	-10,1219	0,000000	-10,1414	0,000000	78	89

Na hladině významnosti 0,05 se prokázal rozdíl v průměrném prospěchu mezi úspěšnými a neúspěšnými studenty.



3. Vytvořte kontingenční tabulky absolutních četností a sloupcově podmíněných relativních četností dvojic kategoriálních proměnných (úspěch, pohlaví), (úspěch, občanství), (úspěch, typ SŠ) a na hladině významnosti 0,05 testujte hypotézu o nezávislosti úspěchu na příslušné kategoriální proměnné. Nezapomeňte ověřovat splnění podmínek dobré aproximace pro Pearsonův chí- kvadrát test nezávislosti.

Výsledky pro pohlaví:

	Kontingenční tabulka (SZZ.sta) Tab. :			
	uspech	pohlavi žena	pohlavi muž	Řádk. součty
Četnost	uspěl	42	36	78
Sloupc. četn.		55,26%	39,56%	
Četnost	neuspěl	34	55	89
Sloupc. četn.		44,74%	60,44%	
Četnost	Vš.skup.	76	91	167

Souhrnná tab.: Očekávané četnosti (SZZ.sta) Pearsonův chí-kv. : 4,10239, sv=1, p=,042823			
uspech	pohlavi žena	pohlavi muž	Řádk. součty
uspěl	35,49701	42,50299	78,0000
neuspěl	40,50299	48,49701	89,0000
Vš.skup.	76,00000	91,00000	167,0000

Na hladině významnosti 0,05 zamítáme hypotézu, že úspěch a pohlaví jsou nezávislé veličiny.

Výsledky pro občanství:

	Kontingenční tabulka (SZZ.sta) Tab. :			
	uspech	obcanstvi Česká republika	obcanstvi Slovensko	Řádk. součty
Četnost	uspěl	55	23	78
Sloupc. četn.		40,44%	74,19%	
Četnost	neuspěl	81	8	89
Sloupc. četn.		59,56%	25,81%	
Četnost	Vš.skup.	136	31	167

Souhrnná tab.: Očekávané četnosti (SZZ.sta) Pearsonův chí-kv. : 11,5542, sv=1, p=,000676			
uspech	obcanstvi Česká republika	obcanstvi Slovensko	Řádk. součty
uspěl	63,5210	14,47904	78,0000
neuspěl	72,4790	16,52096	89,0000
Vš.skup.	136,0000	31,00000	167,0000

Na hladině významnosti 0,05 zamítáme hypotézu, že úspěch a občanství jsou nezávislé veličiny.

Výsledky pro typ střední školy:

	Kontingenční tabulka (SZZ.sta) Tab. :				
	uspech	typ SS gymnázium	typ SS SPŠ+OA	typ SS ostatní	Řádk. součty
Četnost	uspěl	70	5	3	78
Sloupc. četn.		50,72%	38,46%	18,75%	
Četnost	neuspěl	68	8	13	89
Sloupc. četn.		49,28%	61,54%	81,25%	
Četnost	Vš.skup.	138	13	16	167

Souhrnná tab.: Očekávané četnosti (SZZ.sta) Pearsonův chí-kv. : 6,27396, sv=2, p=,043414				
uspech	typ SS gymnázium	typ SS SPŠ+OA	typ SS ostatní	Řádk. součty
uspěl	64,4551	6,07186	7,47305	78,0000
neuspěl	73,5449	6,92814	8,52695	89,0000
Vš.skup.	138,0000	13,00000	16,00000	167,0000

Na hladině významnosti 0,05 zamítáme hypotézu, že úspěch a typ střední školy jsou nezávislé veličiny.

4. Vytvořte model binární logistické regrese, který umožní predikovat pravděpodobnost úspěchu u státní závěrečné zkoušky bakalářského studia. Vzhledem k tomu, že jednorozměrné analýzy prokázaly závislost úspěchu na studijním průměru, pohlaví, občanství a typu absolvované střední školy, zahrňte nejprve do modelu všechny sledované nezávisle proměnné veličiny. Přitom u kategoriálních proměnných použijte kódování pomocí referenční kategorie.

a) Odhadněte regresní parametry a podíly šancí. Na hladině významnosti 0,05 proveďte dílčí testy významnosti regresních parametrů a celkový test významnosti.

Tabulky odhadů parametrů a odhadů podílů šancí společně s dílčími testy významnosti:

uspech - Odhady parametrů (SZZ.sta)
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT
Modelovaná pravděpodobnost, že uspech = uspěl

Efekt	Úroveň Efekt	Sloupec	Odhad	Standard chyba	Wald. Stat.	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1	11,79869	2,208899	28,53091	7,46932	16,12805	0,000000
prumer		2	-4,37536	0,720268	36,90100	-5,78706	-2,96366	0,000000
pohlavi	žena	3	1,67722	0,646821	6,72373	0,40947	2,94496	0,009514
obcanstvi	Česká republika	4	-0,37608	0,680065	0,30582	-1,70899	0,95682	0,580256
typ SS	gymnázium	5	0,25651	0,967538	0,07029	-1,63982	2,15285	0,790917
typ SS	SPŠ+OA	6	0,41652	1,342016	0,09633	-2,21379	3,04682	0,756282
Měřítko			1,00000	0,000000		1,00000	1,00000	

uspech - Poměry šancí (SZZ.sta)
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT
Modelovaná pravděpodobnost, že uspech = uspěl

Efekt	Úroveň Efekt	Sloupec	Šance Poměr	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1				
prumer		2	0,012584	0,003067	0,05163	0,000000
pohlavi	žena	3	5,350639	1,506021	19,00992	0,009514
obcanstvi	Česká republika	4	0,686545	0,181049	2,60340	0,580256
typ SS	gymnázium	5	1,292417	0,194014	8,60938	0,790917
typ SS	SPŠ+OA	6	1,516670	0,109286	21,04832	0,756282
Měřítko			1,000000			

Výsledek celkového testu významnosti:

Testování glonální nulové hypotézy: BETA=0 (SZZ.sta)
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT
Modelovaná pravděpodobnost, žeuspech = uspěl (Vzorek pro analýzu)

	Chí-kvadrát	SV	p
Poměr věrohodnos	147,338897	5	0,000000
Skóre	105,891705	5	0,000000
Wald.	40,548315	5	0,000000

Na hladině významnosti 0,05 zamítáme hypotézu, že dostačující je model konstanty.
Významné jsou však jen proměnné průměr a pohlaví, občanství a typ střední školy nikoliv.

Sestavíme nový model s nezávisle proměnnými průměr a pohlaví:

uspech - Odhady parametrů (SZZ.sta)
Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT
Modelovaná pravděpodobnost, že uspech = uspěl

Efekt	Úroveň Efekt	Sloupec	Odhad	Standard chyba	Wald. Stat.	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1	12,11923	1,977184	37,57126	8,24402	15,99444	0,000000
prumer		2	-4,48162	0,708794	39,97881	-5,87083	-3,09241	0,000000
pohlavi	žena	3	1,59031	0,597262	7,08977	0,41969	2,76092	0,007753
Měřítko			1,00000	0,000000		1,00000	1,00000	

V modelu se dvěma nezávisle proměnnými průměr a pohlaví jsou obě proměnné významné na hladině významnosti 0,05.

Pravděpodobnost, že student uspěje u SZZ, je vyjádřena rovnicí

$$P(\text{uspech} = 1 / \text{prumer} = x_1 \wedge \text{pohlavi} = x_2) = \frac{1}{1 + e^{-12,1192 + 4,4816 \cdot x_1 - 1,5903 \cdot x_2}}$$

Efekt	uspech - Poměry šancí (SZZ.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, že uspech = uspěl					
	Úroveň Efekt	Sloupec	Šance Poměr	Dolní LS 95,0%	Horní LS 95,0%	p
Abs.člen		1				
prumer		2	0,011315	0,002821	0,04539	0,000000
pohlavi	žena	3	4,905255	1,521497	15,81437	0,007753
Měřítko			1,000000			

Zvýší-li se studijní průměr o 1, má student 0,01x menší šanci na úspěch.
Je-li student žena, má 4,9x větší šanci na úspěch než muž.

c) Proveďte hodnocení kvality modelu.

Nagelkerkův koeficient a Pearsonův chí-kvadrát test dobré shody:

	uspech - Statistiky kvality modelu (SZZ.sta) Rozdělení : BINOMICKÉ, Linkující funkce: LOGIT Modelovaná pravděpodobnost, žeuspech = uspěl (Vzorek pro analýzu)		
	SV	Stat.	Stat/sv
Odchylka	164	83,853931	0,511304
Deviance v měřít	164	83,853931	0,511304
Pearsonovo Chi2	164	99,817713	0,608645
Scaled P. Chi2	164	99,817713	0,608645
AIC		89,853931	
BIC		99,207912	
Cox-Snell R2		0,585148	
Nagelkerke R2		0,781330	
Log-věrohodnost		-41,926965	

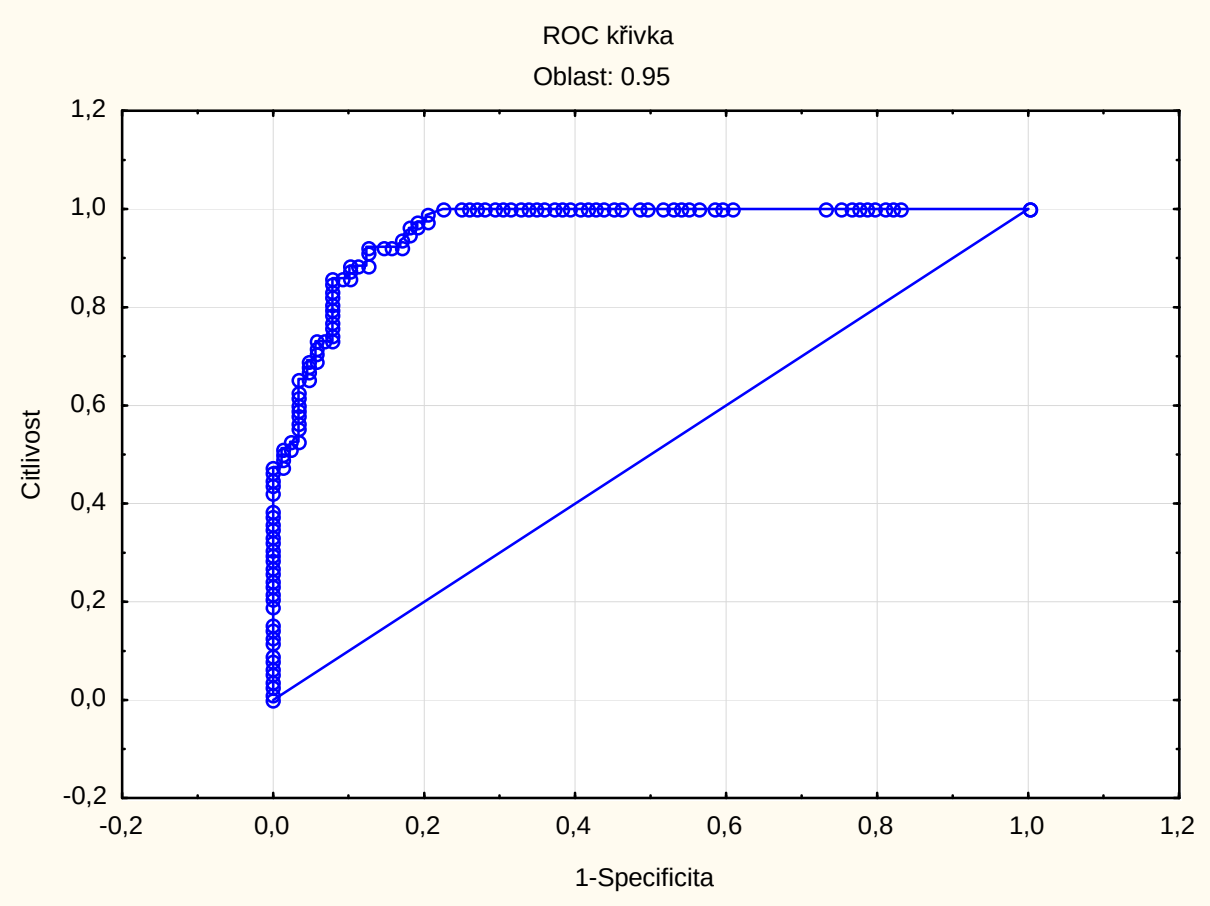
Nagelkerkův koeficient je 0,78, což svědčí o dobré kvalitě modelu. Pearsonův chí-kvadrát test dobré shody má testovou statistiku 99,8177, kritický obor
je $W = \langle \chi^2_{0,95}(164), \infty) = \langle 194,8825; \infty)$, tedy naše data jsou v souladu s modelem.

Klasifikační tabulka:

	Klasifikace případů (SZZ.sta) Odds ratio: 60,566667 Log odds ratio: 4,103745		
	Předpovězená: uspěl	Předpovězená: neuspěl	Procento správných
Pozorované: uspěl	69	9	88,4615385
Pozorované: neuspěl	10	79	88,7640449

Model správně zařadil 88,5 % úspěšných studentů a 88,8% neúspěšných studentů.

ROC křivka:



Naše ROC křivka se blíží ideální ROC křivce. Plocha pod ní je 0,95.