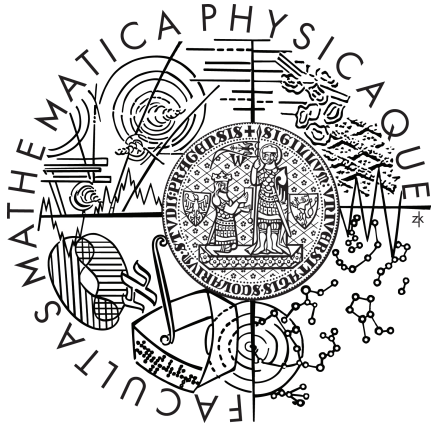
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Tomáš Protivínský

Vliv indikátorů na dosažené skóre v testu z anglického jazyka MANA 2005

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Ing. Marek Omelka

Studijní program: Matematika

Studijní obor: Obecná matematika

Praha 2006

Děkuji Ing. Marku Omelkovi za vedení mé bakalářské práce a cenné rady, které mi poskytl.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 8.3.2006 Tomáš Protivínský

Obsah

Název práce: Vliv indikátorů na dosažené skóre v testu z anglického jazyka MANA 2005
Autor: Tomáš Protivínský
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Ing. Marek Omelka
E-mail vedoucího: omelka@karlin.mff.cuni.cz

Abstrakt: Práce analyzuje data získaná testy Maturita nanečisto 2005. Zabývá se vlivem jednotlivých faktorů na dosažené skóre z anglického jazyka, hodnotí je z hlediska jejich významnosti a snaží se odhalit vztahy mezi nimi. Nejdříve uvažuji jednoduché modely s jedním vysvětlujícím faktorem. Zde jsem se zaměřil především na typ školy (gymnázium, SOŠ, SOU) a její polohu (dle příslušnosti ke kraji) a pohlaví žáka. V práci také krátce diskutuji možné příčiny pozorovaných rozdílů. Tyto dílčí výsledky pak na závěr porovnávám s celkovou analýzou, která do modelu zahrnuje všechny sledované faktory. Zahrnul jsem i stručný popis zaměření a organizace testů Maturita nanečisto.

Klíčová slova: Lineární regrese, analýza rozptylu

Title: Effect of indicators on adjusted score in English language test MANA 2005
Author: Tomáš Protivínský
Department: Department of Probability and Mathematical Statistics
Supervisor: Ing. Marek Omelka
Supervisor's e-mail address: omelka@karlin.mff.cuni.cz

Abstract: The thesis analyzes a data acquired in the test Maturita nanečisto 2005. It is engaged in an effect of separate factors on a final score in the English language test, rates them from a view of the significance and aims to expose a relation among them. First of all I consider simply models with one factor. In this part, I study especially a type of the school and its location (according to districts) and a student's sex. In the thesis, I discuss shortly possible causes of the difference. In the end I compare these partial results with a total analysis, which includes all the factors into a model. In thesis, a brief description of the Maturita nanečisto test is also included.

Keywords: Linear regression, analysis of variance

Kapitola 1

Úvod

Předmětem práce je zpracování výsledků testu z anglického jazyka MANA 2005 (MAturita NAnečisto), zejména porovnání vlivů různých indikátorů na dosažené skóre žáka, například působení typu školy, pohlaví a některých dalších faktorů, vyšetření, které vlivy jsou statisticky významné, a jejich znázornění pomocí několika grafů. Dále se pokusím odhadnout možné příčiny vlivů.

Ke zpracování dat jsem použil statistický software R, rovněž i ke grafickým výstupům.

V následující kapitole se zabývám původem dat, kde a jak byla získána, k čemu testy MANA 2005 jsou a jak jsou koncipovány.

Základní statistické charakteristice se věnuje třetí kapitola.

Kapitola čtvrtá vyšetřuje a znázorňuje vliv typu školy na dosažené skóre. Zvažuje významnost rozdílů a pokouší se analyzovat jejich příčiny.

Pátá a šestá kapitola rozebírají podobným způsobem jako kapitola čtvrtá vliv pohlaví a kraje, v němž se škola nachází, popřípadě vyšetřují faktory dále a hledají možný původ rozdílů.

Sedmá kapitola analyzuje vlivy ostatních indikátorů, dosaženou známku či maturitu z anglického jazyka a budoucí vysokoškolské zaměření studenta.

Osmá kapitola obsahuje celkovou analýzu všech faktorů pomocí lineárního modelu.

Devátá kapitola práci stručně uzavírá.

Několik posledních stránek tvoří stručně okomentované příkazy jazyka R a seznam použité literatury.

Kapitola 2

O projektu MANA

2.1 Co jsou testy MANA 2005

Testy Maturita nanečisto 2005 jsou jednou částí projektu Krok za krokem k nové maturitě, realizovaným Centrem pro zjišťování výsledků vzdělávání, zkráceně CERMAT. Účelem celého projektu je připravit žáky na novou podobu maturitních zkoušek, jež je vymezena takzvaným „školským zákonem“ – zákonem č. 561/2004 Sb., o předškolním, základním, středním, vyšším odborném a jiném vzdělávání, v platnost vstupuje školním rokem 2007/2008.

Nová maturitní zkouška se bude skládat ze dvou částí, společné a profilové. Žáci musí úspěšně vykonat obě, přitom společná část bude zadávána celorepublikově, ve stejném termínu a za stejných podmínek Ministerstvem školství, mládeže a tělovýchovy. Pro všechny maturanty bude mít stejnou úroveň. Stanovení obsahu, formy, témat a termínů zkoušek v profilové části maturity zůstává v pravomoci ředitele školy. Strukturu nových maturitních zkoušek ukazuje tabulka 2.1.

MANA 2005 je koncipována podle plánovaných požadavků MŠMT na novou maturitní zkoušku, je rovněž zadávána celorepublikově a hodnocena za stejných podmínek, dává tedy možnost žákům i pedagogům lépe se na nový typ maturit připravit.

Tab. 2.1: Nová struktura maturitní zkoušky

Společná část	**Profilová část**
Český jazyk	1. Profilová zkouška
Cizí jazyk *(angličtina, francouzština, italština, němčina, ruština, španělština)*	2. Profilová zkouška
Volitelná zkouška *(matematika, občanský základ, přírodovědně technický základ, informačně technologický základ)*	3. Profilová zkouška
	Nepovinné zkoušky (nejvýše 4)

2.2 Organizace testů

Stejně jako v ostatních částech projektu Krok za krokem k nové maturitě, i v MANA 2005 je účast škol zcela dobrovolná a bezplatná. Celý projekt byl zahájen v roce 2001 programem Seznamte se: Nová maturita, pokračoval programy Maturita po internetu (2002), Maturita nanečisto (2003) a Maturita nanečisto 2004. Dále navazuje letošní program Maturita nanečisto 2006. Tabulka 2.2 ukazuje počty zúčastněných žáku a škol v celém programu v letech 2004 a 2005.

Tab. 2.2: Účast na testech v letech 2004, 2005

	Podíl zúčastněných škol v %	**Počet zúčastněných škol**	**Počet žáků**
2004	63,7 %	961	50 189
2005	74,5 %	1 173	59 105

Přihlašování žáků na jednotlivé předměty prostřednictvím ředitelů škol probíhalo od 1. 11. do 10. 12. 2004 na webové stránce CERMATu. Pro každého žáka byl vygenerován identifikační kód ve formě samolepicího štítku s čárovým kódem, který sloužil k označení záznamových archů pro odpovědi žáka a zároveň k jeho identifikaci při zpracování výsledků.

Přihlášené školy obdržely začátkem února 2005 zásilky s testovými materiály, které obsahovaly: testový sešit se soubory úloh od každého předmětu, záznamové archy pro každého přihlášeného žáka, metodické listy, pokyny pro zadavatele, manuály pro rozbor vybraných testových úloh a učitelské dotazníky. Všechny testové materiály byly zaslány školám také na CD nosiči. Školám byly soubory testových úloh i doprovodný materiál, stejně tak jako zpracování výsledků, poskytnuty zdarma.

Testování na školách mělo probíhat v období od 3. 2. do 28. 2. 2005, vzhledem k chřipkovým prázdninám, které byly vyhlášeny v některých částech republiky v době testování, byl termín ukončení o týden posunut. Zároveň byl i o týden změněn termín, do kdy musely školy odeslat vyplněné záznamové archy zpět CERMATu.

Výsledky testů si žáci i školy mohli po jejich vyhodnocení prohlédnout na webových stránkách CERMATu.

Programu Maturita nanečisto 2005 se mohli zúčastnit i žáci se speciálními vzdělávacími potřebami. Jejich začleněním do přípravy reformované maturitní zkoušky se CERMAT snaží přispívat k vyrovnávání příležitostí v oblasti vzdělávání. Jak soubory testových úloh, tak organizační a technické podmínky zadávání byly ve spolupráci se školou individuálně přizpůsobovány podle konkrétních vzdělávacích potřeb každého žáka.

Žáci si mohli vybírat z 25 souborů testových úloh, přičemž názvy jednotlivých testových sešitů byly ve srovnání s předchozími ročníky změněny tak, aby odpovídaly schválenému školskému zákonu. Konstrukce souborů testových úloh částečně předjímala změny, ke kterým bude docházet v souvislosti se zaváděním rámcových vzdělávacích programů a s tím souvisejících aktualizací katalogů požadavků z roku 2000, resp. 2001. Časový limit pro vypracování jednoho souboru testových úloh byl 60 minut. Tabulky 2.3 a 2.4 zobrazují výběr jednotlivých předmětů žáky.

2.3 Test z anglického jazyka

Skládal se z pěti částí, obsahujících dohromady 35 otázek, na jejichž vypracování žáci měli 60 minut (ačkoli u nových maturitních zkoušek se předpokládá délka testu 90 minut). Tři části obsahovaly úlohy s výběrem odpovědi ze tří nebo čtyř alternativ, kde pouze jedna je správná, celkový bodový zisk z těchto částí byl 35 bodů. Zbylé dvě části tvořily úlohy dichotomické (tj. s dvoučlennou volbou, rozhodování o pravdivosti či ne-

Tab. 2.3: Společná část

	Přišlo ze škol ke zpracování	
Společná část	**2004**	**2005**
Český jazyk	23 529	35 245
Český jazyk v komunikaci neslyšících	16	12
Polský jazyk	144	59
Matematika	15 507	18 122
Občanský základ	11 540	16 427
Anglický jazyk	13 388	20 123
Německý jazyk	9 496	12 996
Francouzský jazyk	442	587
Ruský jazyk	114	118
Španělský jazyk	216	176
Italský jazyk	39	14
CELKEM	74 431	103 869

pravdivosti tvrzení) a úlohy přiřazovací, s možným ziskem 9 a 8 bodů. Většina úloh byla zaměřena na práci s nepříliš složitým textem, pochopení hlavní myšlenky, porozumění závěrům, vyhledání informací nebo používání základních gramatických prostředků.

Žáci během testů nesměli používat slovníky, své odpovědi vyznačovali do záznamových archů. Za nezodpovězené či špatně zodpovězené otázky se body neodečítají. Záznamové archy obsahují i krátký žákovský dotazník, jehož vyplnění však není zahrnuto v časovém limitu 60 minut.

Učitelům bylo doporučeno nezahrnovat výsledky testů MANA 2005 do klasifikace žáků, což samozřejmě mohlo zkreslit výsledky.

Další informace o projektu, včetně testových souborů, je možné získat na internetové adrese www.cermat.cz/nanecisto2005.

Tab. 2.4: Profilová část

	Přišlo ze škol ke zpracování	
Profilová část	**2004**	**2005**
Český jazyk a literatura	26 302	18 660
Polský jazyk – rozšířená úroveň	39	45
Matematika – rozšířená úroveň	12 781	9 465
Občanský a společenskovědní základ	10 316	8 280
Anglický jazyk – rozšířená úroveň	15 923	11 492
Německý jazyk – rozšířená úroveň	8 076	4 997
Francouzský jazyk – rozšířená úroveň	714	501
Ruský jazyk – rozšířená úroveň	74	45
Španělský jazyk – rozšířená úroveň	256	153
Italský jazyk – rozšířená úroveň	45	9
Biologie	3 895	4 140
Chemie	2 529	2 537
Fyzika	2 709	2 705
Dějepis	3 244	3 306
Zeměpis	3 372	3 383
CELKEM	90 275	69 718

Kapitola 3

Vyšetřovaná data

3.1 Charakteristika dat

Zabýval jsem se pouze daty z předmětu Anglický jazyk (základní úroveň), krátký výběr z dat zobrazuje tabulka 3.1.

Sloupec „test“ označuje kód testu, u všech žáků tedy shodný, AJ09. „typ“ udává z jakého typu školy žák pochází, zdali z gymnázia, ze střední odborné školy či ze středního odborného učiliště, „kraj“ je kraj, kde se škola nachází, a „skupina“ upřesňuje její zaměření. „pohlavi“ rozlišuje muže a ženy, „znamka“ značí žákovo hodnocení z anglického jazyka na posledním vysvědčení a „matur“ skládá-li z anglického jazyka maturitní zkoušku. Sloupec „navs“ udává, hlásí-li se žák na VŠ, „vsobor“ její obor a „vsprij“ zdali koná přijímací zkoušku z anglického jazyka. Poslední dva sloupce se týkají přímo žákových výsledků v testu, „skore“ značí dosažené body na stupnici 0 až 52, „uspesnost“ je skóre přepočtené na škálu 0 až 100.

Chybějící data jsou označena symbolem <NA>, jejich množství nebylo nijak výrazné, výsledky tedy příliš nezkreslily. Kde bylo možné, zahrnoval jsem do šetření i žáky, od nichž jsem neměl kompletní údaje – například při vyšetřování vlivu stupně hodnocení na vysvědčení, vynechal jsem pouze výsledky studentů s chybějící známkou (skóre bylo u všech údajů), ale výsledky studentů s chybějícími údaji o pohlaví či maturitě z AJ jsem do této konkrétní analýzy zahrnul.

Tab. 3.1: Výběr z dat

	test	typ	kraj	skupina	pohlavi	znamka	...
1	AJ09	sos	Pardubicky	vseobecne	muz	1	...
2	AJ09	sos	Pardubicky	vseobecne	zena	2	...
3	AJ09	sos	Pardubicky	vseobecne	muz	3	...
4	AJ09	sos	Pardubicky	vseobecne	zena	3	...
5	AJ09	sos	Pardubicky	vseobecne	muz	3	...
6	AJ09	sos	Pardubicky	vseobecne	muz	2	...
...	...	...	...	...	...	...	...

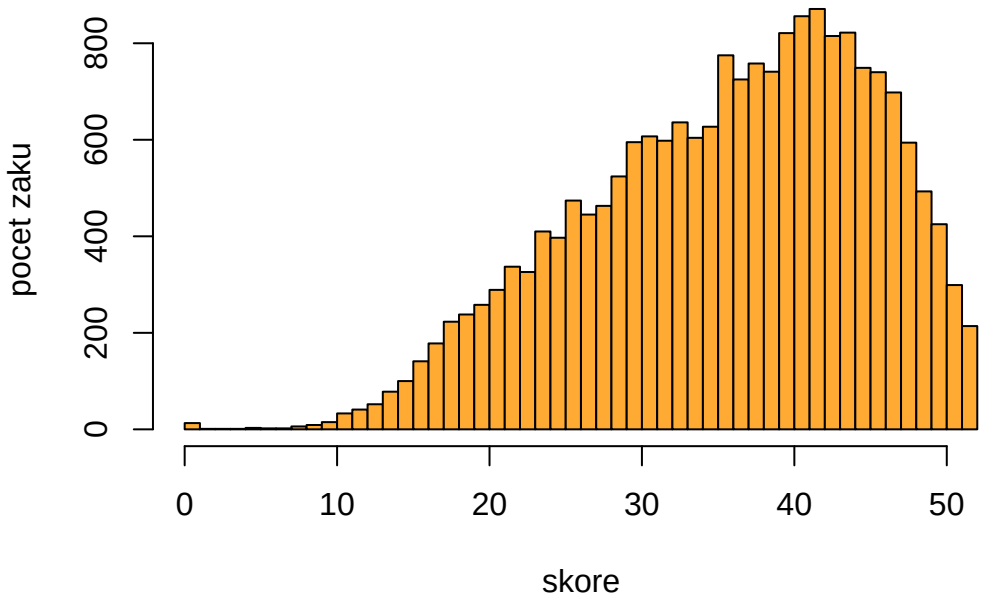
...	matur	navs	vsobor	vsprij	uspesnost	skore
...	ano	ano	ekonomicke	ano	84.61539	44
...	ano	ano	technicke	ano	65.38461	34
...	ano	ano	technicke	ano	78.84615	41
...	ano	ano	ekonomicke	ano	71.15385	37
...	ano	ano	ekonomicke	ano	84.61539	44
...	ano	ne	nehlasi	ne	57.69231	30
...	...	...	...	...	...	...

3.2 Základní statistický popis zkoumaných veličin

Data obsahují údaje od 20 123 žáků. Předmětem mého výzkumu je dosažený výsledek. Závislost úspěšnosti na skóre je zadaná a jednoznačná (úspěšnost = skóre x 100 / 52), tudíž i všechny charakteristiky úspěšnosti lze odvodit ze skóre, proto se budu zabývat pouze jím. Rozdělení skóre naznačuje histogram (obr. 3.1).

Průměrné dosažené skóre bylo 36.03 a výběrový rozptyl 88.38. Následující tabulka 3.2 shrnuje kvantily po 10 %.

Rozborem jednotlivých částí testu lze poměrně snadno zjistit, že průměrný výsledek, který dosáhne žák vyplňující test pouze náhodně, je přibližně 13 bodů. Proto je pochopitelné, že hodnoty skóre nižší než deset téměř nejsou zastoupeny. Výraznější zastoupení nulového skóre svědčí spíše o lenosti žáků než o jejich skutečných znalostech z anglického jazyka.



Obr. 3.1: Histogram skóre

Tab. 3.2: Kvantily

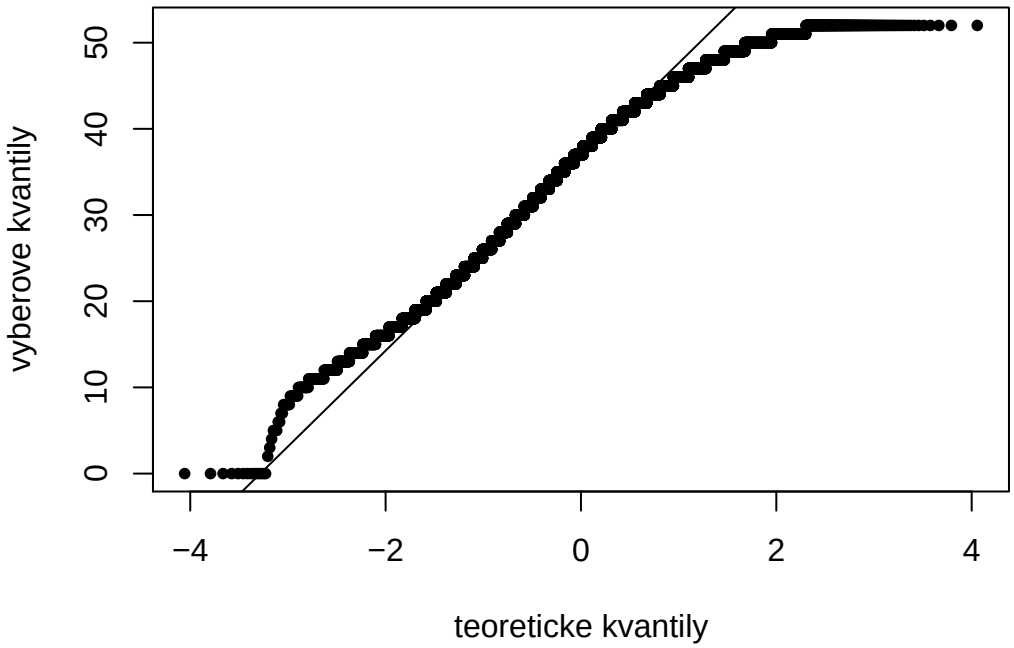
0 %	10 %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	100 %
0	22	27	31	34	37	40	42	45	48	52

3.3 Rozdělení skóre

V dalších kapitolách užívám některé statistické prostředky určené pro nezávislý výběr z normálního rozdělení, oprávněnost jejich užití musím prodiskutovat.

Graf 3.2 porovnává kvantily normálního rozdělení s kvantily skóre. V místech, kde se vypočtené hodnoty přibližují přímce, je skóre téměř normálně rozdělené. Na obou koncích dochází k odklonům.

Jak ukazují grafy 3.1 a 3.2, skóre zřejmě normálně rozdělené není, nicméně jeho rozdělení též není nijak „divoké“, není výrazně asymetrické, ani se nemusím obávat odlehlých pozorování (neboť tato veličina nabývá



Obr. 3.2: Porovnání s normálním rozdělením

pouze hodnot 0 až 52). Mnou používáné statistické metody (t-test, analýza rozptylu, lineární regresní model) jsou sice odvozené pro normální rozdělení, avšak pro korektní analýzu stačí, budou-li používané statistiky správně rozdělené, což mi zaručuje velký počet nezávislých pozorování (20 123).

Kapitola 4

Vliv typu školy na skóre

4.1 Výsledky na jednotlivých typech škol

Dá se očekávat, že různé typy škol dosáhnou různých výsledků. Nejlépe by měly dopadnout gymnázia, ať už z důvodu kvalitnější výuky či nejlepších žáků, kteří prošli náročnějšími přijímacími zkouškami. Potvrzením tohoto očekávání se zabývá tato kapitola.

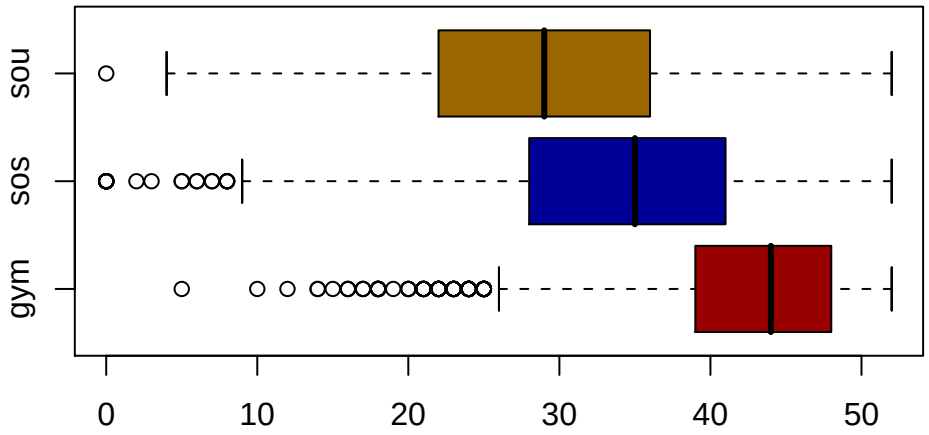
Pro účely testování byly školy rozděleny do tří základních typů – gymnázia, střední odborné školy a střední odborná učiliště. Rozložení žáků mezi tyto školy a základní popis těchto tří skupin shrnuje tabulka 4.1.

Tab. 4.1: Popis výsledků jednotlivých typů škol

	Počet žáků	V %	Průměr	Směr. odchylka	Medián
Gymnázia	5515	27,41	42,90	6,33	44
SOŠ	12453	61,88	34,14	8,82	35
SOU	2155	10,71	29,32	9,32	29

Lepší představu umožňuje krabicový diagram (obr. 4.1), který zachycuje medián (prostřední tlustší čára) a kvartily (hranice obdélníku). Tečkou jsou znázorněna odlehlá pozorování, která jsou vzdálena od bližšího kvartilu dále než 1,5násobek rozdílu kvartilů. Graf 4.2 pak znázorňuje relativní četnost dosaženého skóre na daném typu školy. Stejně jako v tabulce 4.1, i zde vychází gymnázia výrazně lépe. To potvrzuje i formální test shody

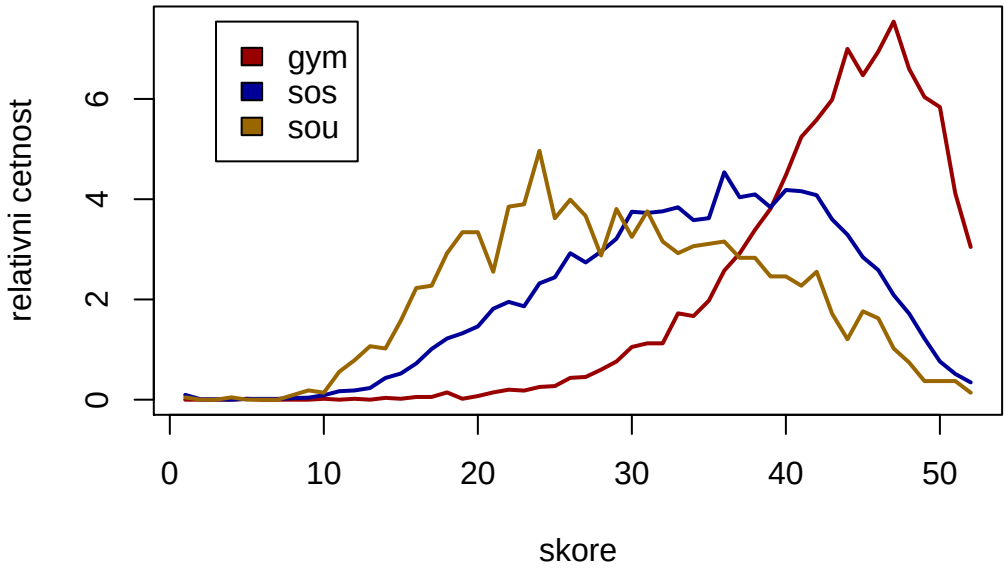
středních hodnot. Tukeyova metoda pak navíc prokazuje různost všech dvojic středních hodnot.



Obr. 4.1: Krabicové diagramy jednotlivých typů škol

4.2 Shrnutí

Nejlépe dopadly gymnázia, za nimi jsou s poměrně dost velkým odstupem střední školy a na posledním místě střední odborná učiliště. Příčin může být několik, zároveň i jejich vliv se může prolínat a doplňovat – výuka na gymnáziích je pravděpodobně kvalitnější, rovněž i domácí samostudium lze očekávat spíše od gymnazijních žáků. Také náročnější přijímací zkoušky přispěli k výběru pilnějších a talentovanějších studentů do gymnazijních tříd. Naopak na střední odborná učiliště se pravděpodobně budou hlásit žáci s nejnižší ochotou zabývat se svým vzděláváním, tedy od nich lze očekávat i horší výsledky.



Obr. 4.2: Relativní četnost dosažené hladiny skóre

Kapitola 5

Vliv pohlaví na skóre

5.1 Výsledky žen a mužů

Kapitola se svým obsahem značně podobá kapitole předchozí, pouze nedělí studenty do skupin podle typu školy, z níž pochází, ale podle jejich pohlaví. Základní popis je shrnut v tabulce 5.1.

Tab. 5.1: Výsledky mužů a žen

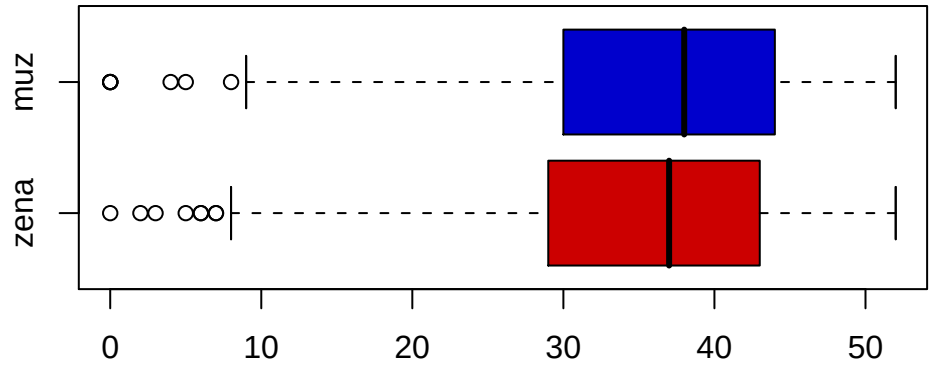
	Počet žáků	V %	Průměr	Směr. odchylka	Medián
Ženy	11359	56,45	35,59	9,45	37
Muži	8724	43,35	36,65	9,23	38

Názornější představu o výsledcích opět poskytují dva grafy, krabicový diagram (obr. 5.1) a relativní četnost dosaženého skóre u žen a mužů (obr. 5.2).

5.2 Porovnání výsledků žen a mužů

Přesto, že jsou výsledky velmi podobné, zdá se, že muži dopadli nepatrně lépe. K dokázání či vyvrácení tohoto tvrzení použiju dvouvýběrový t-test, přesněji jeho modifikaci známou jako Welchův test.

P-hodnota při hladině testu 5% vyšla $1,217.10^{-15}$, hypotézu o rovnosti středních hodnot skóre dosaženého ženami a muži mohu zamítnout.



Obr. 5.1: Krabicový diagram mužů a žen

Zjištěný rozdíl průměrů 1,064 je statisticky významný, 95%ní interval spolehlivosti pro tento rozdíl je (0,803; 1,324).

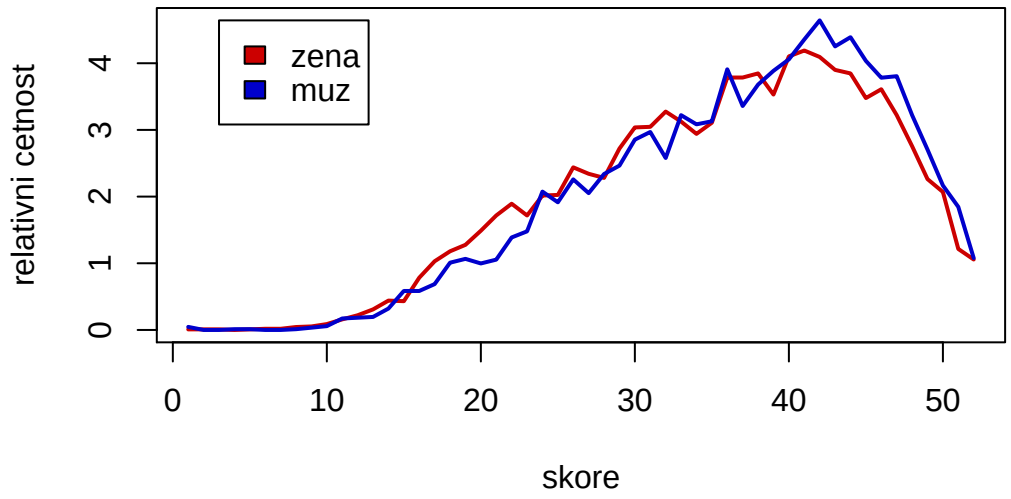
5.3 Shrnutí

Daly se očekávat velmi vyrovnané výsledky, toto očekávání se ukázalo pravdivým, ale i rozdíl jednoho bodu mezi průměrným skórem mužů a žen by měl mít nějakou příčinu. Variant je samozřejmě několik, je možné, že muži jsou obecně mírně talentovanější, co se týče jazykových schopností. Další příčinou by mohly být různé poměry v zastoupení žen a mužů na jednotlivých typech škol, například větší procento mužů na gymnáziích a středních odborných školách oproti vyššímu relativnímu zastoupení žen na středních odborných učilištích.

Následující tabulka 5.2 ukazuje zastoupení jednotlivých pohlaví na různých typech škol.

Tab. 5.2: Absolutní i relativní zastoupení

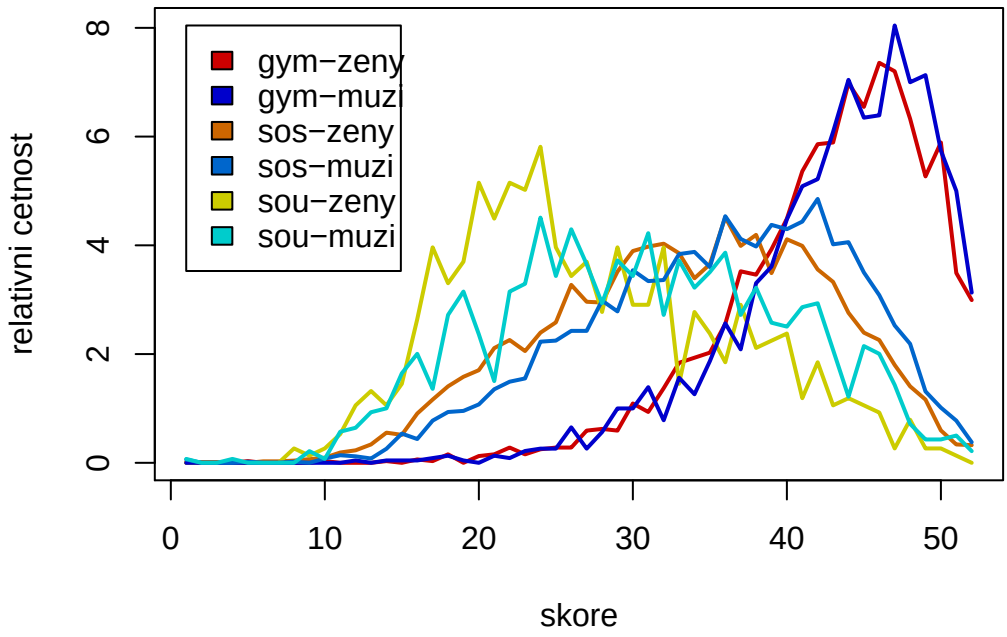
	Celkem (v %)	Gymnázia	SOŠ	SOU
Ženy	11359 (56,45%)	3208 (58,17%)	7394 (59,38%)	757 (35,13%)
Muži	8724 (43,35%)	2300 (41,70%)	5027 (40,37%)	1397 (64,83%)



Obr. 5.2: Relativní četnosti dosaženého skóre u žen a mužů

Ze zastoupení pohlaví na různých typech škol by vyplývalo spíše opačné tvrzení, tudíž by se dal očekávat lepší výsledek u žen. Podrobnější závěry ukazuje následující graf (obr. 5.3), zachycující relativní četnost jednotlivých hodnot skóre podle typů škol a pohlaví.

Rozdíly mezi gymazijními studenty obou pohlaví jsou velmi malé, na středních odborných školách dosahovali muži mírně lepších výsledků a na středních odborných učilištích již poměrně značně lepších výsledků. Proto dopadli muži v celkovém porovnání s ženami lépe, ačkoli mnohem více z nich navštěvuje školy s obecně nižší úrovní.



Obr. 5.3: Relativní četnosti dosaženého skóre

Kapitola 6

Vliv kraje na skóre

6.1 Výsledky v jednotlivých krajích

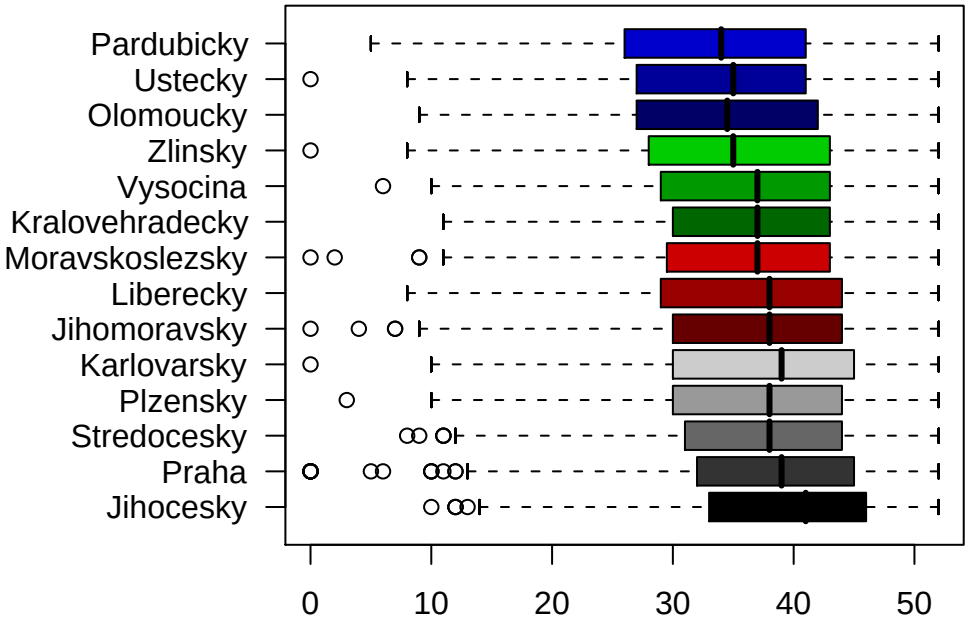
Třetí rozdělení žáků vychází z polohy školy, dělí je do 14 krajů. Základní popis je shrnut do tabulky 6.1.

Tab. 6.1: Výsledky podle krajů

	Počet žáků	V %	Průměr	Směr. odchylka	Medián
Jihočeský	1111	5,52	38,62	9,49	41
Praha	2210	10,98	37,56	9,22	39
Středočeský	2048	10,18	36,80	9,38	38
Plzeňský	659	3,27	36,61	9,19	38
Karlovarský	501	2,49	36,51	9,93	39
Jihomoravský	2283	11,34	36,44	9,23	38
Liberecký	1686	8,38	36,29	9,91	38
Moravskoslezský	3064	15,23	35,86	9,11	37
Královéhradecký	1016	5,05	35,82	8,88	37
Vysočina	769	3,82	35,67	9,18	37
Zlínský	1467	7,29	34,74	9,51	35
Olomoucký	1304	6,48	34,08	9,22	34,5
Ústecký	1115	5,54	33,91	8,95	35
Pardubický	890	4,42	33,69	9,61	34

Kraje nebyly vyplňovány žáky, ale školou, jsou tedy u všech dat.

Krabicový diagram (obr. 6.1) umožňuje lepší představu. Druhý graf, znázorňující relativní četnosti dosaženého skóre pro jednotlivé kraje, neuvádím, neboť tentokrát neukazoval žádné názorné výsledky.



Obr. 6.1: Krabicové diagramy krajů

6.2 Shrnutí

Mezi jednotlivými kraji vychází poměrně velké rozdíly, snadno lze nahlédnout, že moravské kraje dopadly hůře než české. Jednou z příčin by opět mohlo být odlišné zastoupení různých typů škol, proto je toto zastoupení shrnuto do tabulky 6.2.

Tabulka je seřazena podle průměrného skóre. Je vidět, že lepší kraje mají opravdu vyšší zastoupení gymnázií (s výjimkou Plzeňského kraje; některé z nich mají i poměrně výrazné zastoupení středních odborných učilišť, ty však mají na výsledné skóre menší vliv než SOŠ a gymnázia, neboť se týkají pouze asi 10% z celkového počtu žáků) než kraje v dolní

Tab. 6.2: Relativní zastoupení

	Celkem	Gymnázia	SOŠ	SOU
Jihočeský	5,52	8,90	3,97	5,85
Praha	10,98	11,78	11,20	7,76
Středočeský	10,18	12,63	9,57	7,38
Plzeňský	3,27	2,67	3,54	3,29
Karlovarský	2,49	3,19	2,55	0,37
Jihomoravský	11,35	12,48	10,08	15,78
Liberecký	8,38	10,03	7,52	9,10
Moravskoslezský	15,23	17,64	14,94	10,67
Královéhradecký	5,05	2,99	6,19	3,71
Vysočina	3,82	3,74	3,81	4,13
Zlínský	7,29	4,19	7,73	12,67
Olomoucký	6,48	3,05	7,74	7,98
Ústecký	5,54	3,39	6,39	6,13
Pardubický	4,42	3,32	4,76	5,29

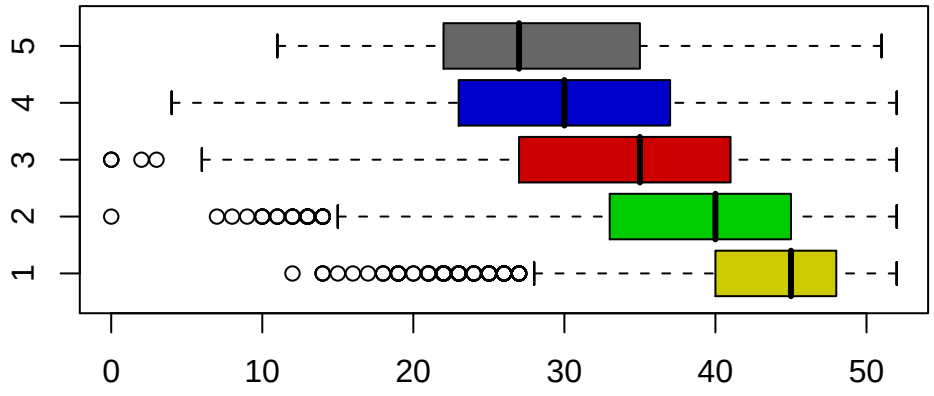
části tabulky, u nichž převažují střední odborné školy nebo učiliště. Rozdíly mezi jednotlivými kraji tedy budou pravděpodobně způsobené typy škol. To částečně potvrzuje i závěrečná analýza.

Kapitola 7

Vlivy dalších indikátorů

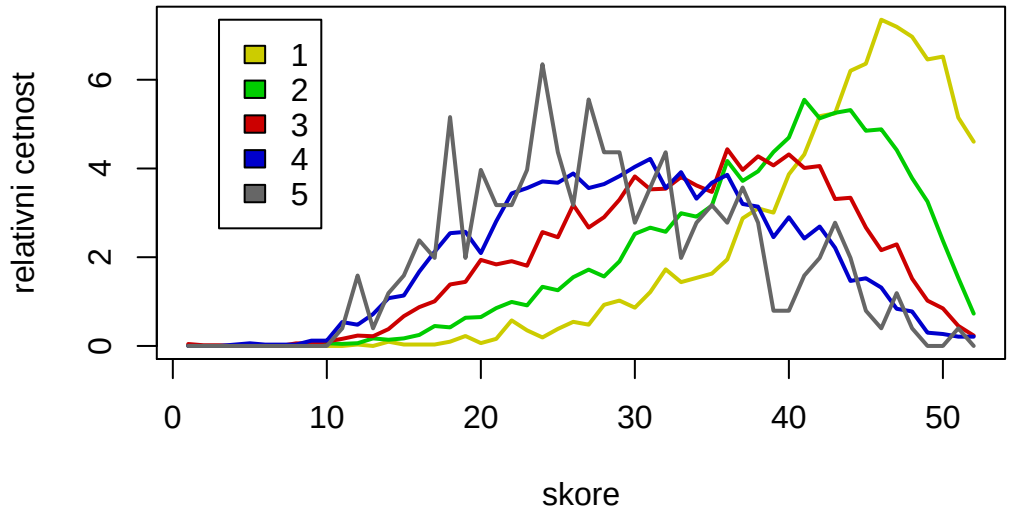
7.1 Vliv známky

Hodnotí-li školy žáky podle skutečných znalostí, pak se očekává, že známka na vysvědčení i výsledek v testu MANA 2005 spolu budou souviset. Jak si tyto veličiny navzájem odpovídají, pomáhají naznačit i následující grafy, krabicový diagram (obr. 7.1) a graf relativních četností jednotlivých hodnot skóre (obr. 7.2).



Obr. 7.1: Krabicové diagramy podle známek na vysvědčení

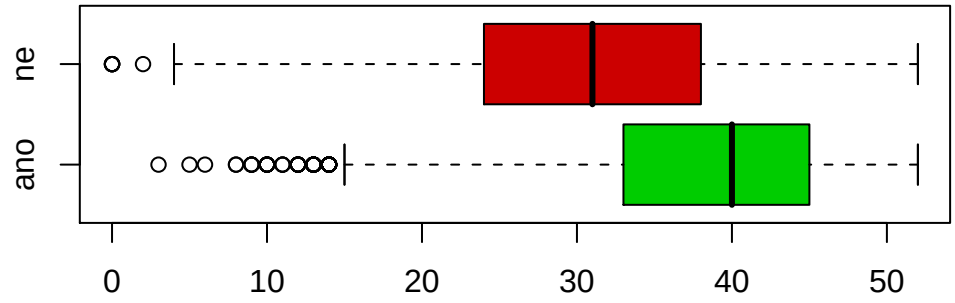
Oba grafy tedy jen potvrzují to, co bylo očekávané.



Obr. 7.2: Relativní četnosti dosaženého skóre podle známek

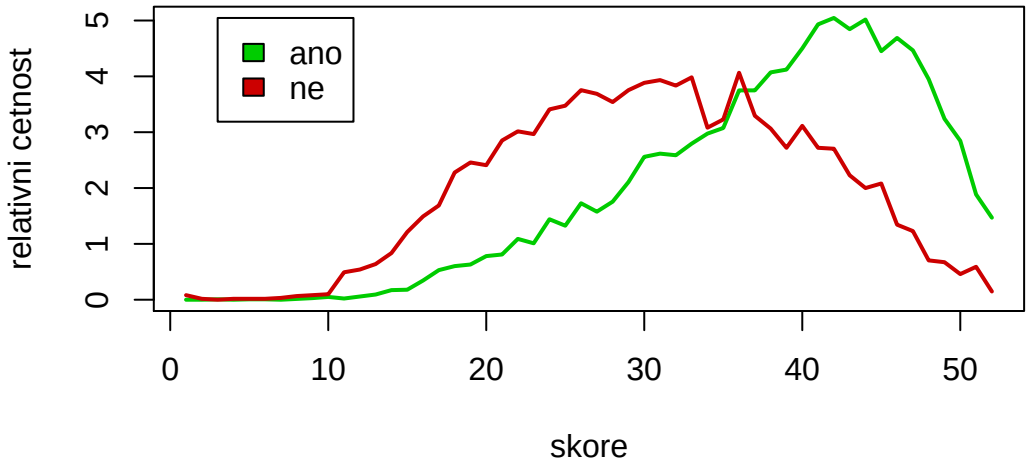
7.2 Vliv maturity z anglického jazyka

Žáci, které čeká maturita z anglického jazyka, by ve cvičných maturitách z anglického jazyka samozřejmě měli uspět lépe – je tomu skutečně tak?



Obr. 7.3: Krabicový diagram maturit

Oba grafy 7.3 a 7.4 potvrzují, že ano.



Obr. 7.4: Relativní četnosti dosaženého skóre podle maturity

7.3 Vliv zaměření žáka

Posledním vyšetřovaným indikátorem je obor vysoké školy, na kterou se žák hlásí. Tabulka 7.1 shrnuje základní statistické veličiny pro jednotlivá vysokoškolská zaměření.

Průměr mezi studenty hlásícími se na VŠ je obecně vyšší než celkový průměr (při t-testu vyšla p-hodnota shody středních hodnot prakticky nulová). Ale i přesto je průměr studentů hlásících se na VŠ výrazně nižší než průměr žáků na gymnáziích. Tabulka 7.2 ukazuje, kolik studentů z jakého typu školy se hlásí na VŠ.

Poměrně zajímavé je umístění studentů hlásících se na matematicko-fyzikální obory, dopadli dokonce lépe než studenti hlásící se ke studiu cizích jazyků, od kterých by se daly očekávat nejlepší výsledky. Proto v následujícím grafu (obr. 7.5) shrnuji zastoupení studentů z různých typů škol mezi jednotlivými VŠ zaměřeními.

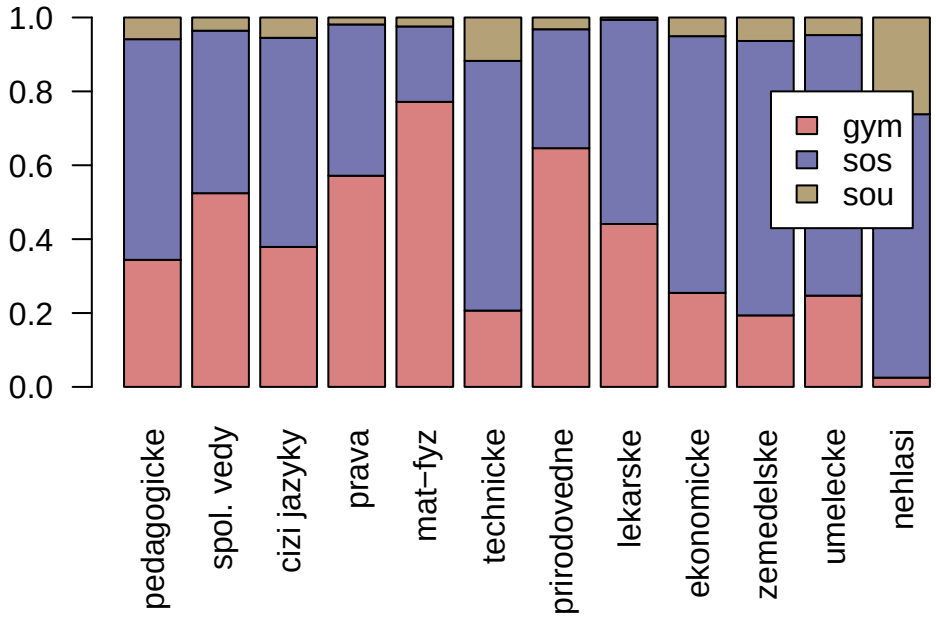
Skutečně, studenti hlásící se na matematicko-fyzikální obory mají nejvyšší zastoupení studentů z gymnázií, tím je jejich úspěch zdůvodněn. Oproti tomu studenti hlásící se ke studiu cizích jazyků mají zastoupení gymnázií mnohem nižší, než by odpovídalo jejich výsledku. To by mohlo naznačovat jejich talent ke studiu cizích jazyků, což je v souladu s celkovou analýzou. Poměrně překvapivým zjištením zůstává nízký výsledek pedagogického zaměření, i přesto, že zastoupení gymnázií nízké není.

Tab. 7.1: Přehled podle VŠ zaměření

	Počet žáků	V %	Průměr	Směr. odchylka	Medián
Mat.-fyz.	289	1,44	42,19	8,11	44
Cizí jazyky	871	4,32	41,60	7,49	43
Práva	796	3,96	40,22	7,91	42
Společenské vědy	1680	8,35	39,55	8,20	41
Přírodovědné	879	4,37	39,14	8,86	41
Ekonomické	3382	16,81	37,85	8,28	39
Technické	3175	15,78	37,17	8,63	38
Umělecké	777	3,86	36,90	8,87	38
Lékařské	1297	6,45	35,95	9,86	37
Pedagogické	1515	7,53	35,07	8,75	36
Zemědělské	346	1,72	34,64	9,13	36
Celkem VŠ	15007	74,58	37,83	8,75	39
Celkem neVŠ	4786	23,78	30,64	9,04	31
Celkem	20123	100,00	36,03	9,40	37

Tab. 7.2: VŠ mezi různými typy škol

	Celkem	Na VŠ	V %
Gymnázia	5515	5337	96,77
SOŠ	12453	8881	71,32
SOU	2155	872	40,46



Obr. 7.5: Poměr typů škol mezi VŠ zaměřeními

Kapitola 8

Souhrnná analýza

Na závěr jsem data vyšetřil souhrnnou analýzou pomocí lineárního modelu, v němž se vlivy jednotlivých faktorů projeví očištěny od vlivu ostatních faktorů. Tabulka 8.1 shrnuje výsledky této analýzy – řádek s absolutním členem odpovídá dívce z jihočeského gymnázia se všeobecným zaměřením, jež se hlásí na VŠ pedagogického oboru, přitom z AJ skládá maturitu i přijímací zkoušku a na vysvědčení byla hodnocena 1 (první hodnoty všech faktorů). Další řádky znázorňují, jaký vliv má každá jednotlivá hodnota každého faktoru. Jednotlivé sloupce postupně značí odhadnutý průměr, p-hodnotu pro test nulovosti koeficientu a meze 95%ního konfidenčního intervalu.

Analýza potvrdila některé mnou zjištěné výsledky. Z tabulky lze například vidět, že pořadí krajů podle jejich odhadnutých vlivů již tak přesně neodpovídá pořadí jejich průměrů, i rozdíly mezi nimi jsou menší.

Další potvrzení mých výsledků se týká vlivu pohlaví – průměrné skóre mužů bylo přibližně o 1 bod vyšší než průměrné skóre žen, ačkoli muži obecně navštěvovali školy s nižší kvalitou výuky, tedy podle objektivních okolností (typ školy) by měli dopadnout hůře. Po odstranění tohoto ovlivnění rozdílnou úrovní výuky vychází odhadované skóre mužů dokonce o 2,16 bodů lépe.

Opět je zřetelný vliv známky na vysvědčení i maturity z anglického jazyka. Rozdíly mezi jednotlivými vysokoškolskými obory jsou nyní menší – velmi pravděpodobně tedy výsledek neovlivňují tolik, jak by naznačovaly průměry podle jednotlivých oborů, příčina rozdílů je spíše jinde, například v typu školy, jež žáci studují.

Analýza však naznačuje malý vliv typu školy, což je mírně v rozporu

s výsledky kapitoly 4. Porovnáním s výsledky jednotlivých středoškolských zaměření je tento rozpor vysvětlen – vliv typu školy je jakoby přenesen sem. Gymnázia jsou zaměřena téměř vždy všeobecně a všeobecné zaměření vychází zpravidla o několik bodů lépe než ostatní zaměření, mezi nimiž jsou opět nejhůře odhadované obory učilišť. Rozpor se tedy ukázal být pouze zdánlivým.

Podobné „přenesení“ vlivu je pravděpodobné i u přijímacích zkoušek, předpokládám, že student, jehož čeká zkouška z AJ, bude též maturovat z anglického jazyka, a u maturujících vyšel výsledek značně lepší. Matoucí je i porovnání budoucích vysokoškoláků s ostatními – neboť budoucí vysokoškolské studium by mělo výsledek ovlivnit negativně, ač jen velmi mírně. Toto se však týká pouze pedagogických oborů, ostatní obory vychází o další bod až dva lépe, navíc je pravděpodobné, že studenti hlásící se na VŠ budou navštěvovat kvalitnější školy a mít lepší prospěch. Proto „očištění“ jednoho faktoru od ostatních zde výsledek spíše zatemnilo.

Regresní model tedy poměrně dobře souhlasí s předchozími výsledky, ačkoli některé údaje se na první pohled mohly zdát zarážející. Toto je dáno zejména vzájemnou provázaností jednotlivých faktorů, pro přesné výsledky by bylo žádoucí provést mnohem hlubší analýzu, jež však svým rozsahem značně překračuje rámec této práce.

Tab. 8.1: Souhrnná analýza

	Odhad	p-hodnota	Spodní mez	Horní mez
Absolutní člen	48,68	$< 2.10^{-16}$	48,03	49,33
Typ – SOŠ	-2,63	$6,3.10^{-13}$	-3,34	-1,91
Typ – SOU	-1,22	0,08	-2,60	0,14
Kraj – Praha	-0,57	0,04	-1,10	-0,04
Kraj – Středočeský	-0,81	$3,0.10^{-3}$	-1,35	-0,28
Kraj – Plzeňský	-0,33	0,37	-1,05	0,39
Kraj – Karlovarský	-1,48	$2,5.10^{-4}$	-2,28	-0,69
Kraj – Jihomoravský	-1,29	$2,1.10^{-6}$	-1,82	-0,76
Kraj – Liberecký	-1,11	$8,1.10^{-5}$	-1,67	-0,56
Kraj – Moravskoslezský	-1,89	$2,5.10^{-13}$	-2,39	-1,38
Kraj – Královéhradecký	-0,16	0,63	-0,78	0,47
Kraj – Vysočina	-1,39	$6,8.10^{-5}$	-2,08	-0,71
Kraj – Zlínský	-1,44	$1,1.10^{-6}$	-2,01	-0,86
Kraj – Olomoucký	-1,81	$3,6.10^{-9}$	-2,41	-1,21
Kraj – Ústecký	-2,55	$1,3.10^{-15}$	-3,17	-1,92
Kraj – Pardubický	-2,48	$3,3.10^{-13}$	-3,14	-1,81
Zaměření – ekonomické	-1,92	$2,1.10^{-7}$	-2,64	-1,20
Zaměření – technické	-1,56	$8,7.10^{-5}$	-2,35	-0,78
Zaměření – technologické	-3,67	$< 2.10^{-16}$	-4,47	-2,88
Zaměření – zemědělské	-2,64	$6,8.10^{-8}$	-3,59	-1,68
Zaměření – humanitní	-4,56	$< 2.10^{-16}$	-5,42	-3,70
Zaměření – hotelové	-0,97	0,03	-1,84	-0,10
Zaměření – zdravotní	-5,00	$< 2.10^{-16}$	-5,88	-4,12
Zaměření – umělecké	-4,82	$< 2.10^{-16}$	-5,85	-3,79
Zaměření – učňák – tech.	-7,09	$< 2.10^{-16}$	-8,54	-5,64
Zaměření – učňák – netech.	-9,40	$< 2.10^{-16}$	-10,80	-8,01
Muž	2,16	$< 2.10^{-16}$	1,91	2,41
Známka (stupeň)	-2,98	$< 2.10^{-16}$	-3,10	-2,87
Nematuruje z AJ	-3,47	$< 2.10^{-16}$	-3,75	-3,19
Nehlásí se na VŠ	0,02	0,93	-0,44	0,49
Nedělají přij. zkoušky z AJ	-0,93	$2,7.10^{-10}$	-1,22	-0,64
VŠ obor – spol. vědy	1,93	$3,4.10^{-14}$	1,43	2,43
VŠ obor – cizí jazyky	2,76	$< 2.10^{-16}$	2,14	3,38
VŠ obor – práva	1,44	$4,5.10^{-6}$	0,82	2,05
VŠ obor – mat.-fyz.	2,21	$2,1.10^{-6}$	1,30	3,12
VŠ obor – technické	2,58	$< 2.10^{-16}$	2,09	3,07
VŠ obor – přírodovědné	1,80	$6,3.10^{-9}$	1,19	2,41
VŠ obor – lékařské	1,46	$8,7.10^{-7}$	0,88	2,05
VŠ obor – ekonomické	1,21	$4,8.10^{-7}$	0,74	1,67
VŠ obor – zemědělské	0,58	0,21	-0,33	1,48
VŠ obor – umělecké	1,99	$1,2.10^{-8}$	1,30	2,67

Kapitola 9

Závěr

Data jsem statisticky zpracoval, rozdělil je do skupin podle různých faktorů a snažil se odhalit jejich vliv na dosažené skóre. Téměř u všech faktorů jsem prokazatelný vliv nalezl a dále analyzoval jeho příčiny. Pomocí grafů a tabulek jsem se pokusil ukázat vzájemnou souvislost mezi jednotlivými proměnnými zkoumaných dat.

Na závěr jsem své výsledky porovnal s celkovou analýzou pomocí lineárního modelu.

Vzhledem k množství dat téměř všechny parametry vycházejí statisticky významně odlišné od nuly. Bylo by vhodné tedy dále prozkoumat, jak tyto jednotlivé faktory přispívají k vysvětlení celkové variability dat, což by poskytlo přesnější představu o důležitosti jednotlivých faktorů. Tato analýza by však vyžadovala také mnohem podrobnější zkoumání vztahů mezi faktory. To by však již bylo nad rámec této práce.

Kapitola 10

Vybrané zdrojové kódy v R

Zde uvádím výběr ze zdrojových kódů, jež jsem použil ke statistickému zpracování dat. Většinou jsou zkrácené, bez popisků či barevného rozlišení, některé jednodušší či podobné již uvedenému jsem vynechal.

10.1 Načtení a první zpracování dat

Načtení dat ze souboru `data.csv` do datové tabulky `mana`, s následnou úpravou veličin charakterizujících studenta na faktory.

```
mana <- read.csv2("data.csv");
mana$test <- factor(mana$test, labels="AJ09");
mana$typ <- factor(mana$typ, labels=c("gym", "sos", "sou"));
poradi.prumeru.kraju <- c(2, 3, 1, 4, 5, 13, 7, 9, 14, 10, 6,
   12, 11, 8);
mana$kraj <- poradi.prumeru.kraju[mana$kraj];
mana$kraj <- factor(mana$kraj, labels=c("Jihocesky", "Praha",
   "Stredocesky", "Plzensky", "Karlovarsky", "Jihomoravsky",
   "Liberecky", "Moravskoslezsky", "Kralovehradecky",
   "Vysocina", "Zlinsky", "Olomoucky", "Ustecky", "Pardubicky"));
mana$skupina <- factor(mana$skupina, labels=c("vseobecne",
   "ekonomicke", "technicke", "technologicke", "zemedelske",
   "humanitni", "hotelove", "zdravotni", "umelecke",
   "ucnak - tech", "ucnak - netech"));
mana$pohlavi <- factor(mana$pohlavi, labels=c("zena", "muz"));
mana$matur <- factor(mana$matur, labels=c("ano", "ne"));
mana$navs <- factor(mana$navs, labels=c("ano", "ne"));
mana$vsobor[mana$navs=="ne"] <- 12;
mana$vsobor <- factor(mana$vsobor, labels=c("pedagogicke",
```

```
   "spol. vedy", "cizi jazyky", "prava", "mat-fyz", "technicke",
   "prirodovedne", "lekarske", "ekonomicke", "zemedelske",
   "umelecke", "nehlasi"));
mana$vsprij <- factor(mana$vsprij, labels=c("ano", "ne"));
attach(mana);
```

10.2 Kapitola 3

Zdrojové kódy použité k vykreslení histogramu, výpočtu decilů a porovnání teoretických kvantilů normálního rozdělení s kvantily rozdělení skóre.

```
hist(skore, breaks=52);
quantile(skore, prob=seq(0, 1, 0.1));
qqnorm(skore);
qqline(skore);
```

10.3 Kapitola 4

Zde jsou nejčastěji používané příkazy, základní popis dat podle nějakého dělení (zde podle typu školy), krabicový diagram pro srovnání několika skupin, poměrně složitá konstrukce k vykreslení grafu relativní četnosti dosaženého skóre v jednotlivých zkoumaných skupinách a základní příkazy analýzy rozptylu.

```
by(skore, typ, length);
by(skore, typ, length)/length(skore);
by(skore, typ, mean);
by(skore, typ, sd);
by(skore, typ, median);

boxplot(skore~typ, horizontal=TRUE);

color.typ <- c("#990000", "#000099", "#996600");
names(color.typ) <- levels(typ);
skore.dle.typu <- table(skore, typ);
skore.dle.typu.prepoctene <- skore.dle.typu
for (i in colnames(skore.dle.typu)) for (j in 1:52)
   skore.dle.typu.prepoctene[j,i] <-
   100*skore.dle.typu[j,i]/sum(skore.dle.typu[,i]);
r.prepoctene=c(0, max(skore.dle.typu.prepoctene));
```

```
plot(skore.dle.typu.prepoctene[,"gym"], ylim=r.prepoctene,
   type="n", xlab="skore", ylab="relativni cetnost");
for (i in colnames(skore.dle.typu.prepoctene))
   points(skore.dle.typu.prepoctene[,i], col=color.typ[i],
   type="l", lwd=2);
legend(x=3, y=r.prepoctene[2], fill=color.typ,
   legend=colnames(skore.dle.typu.prepoctene));

model<-aov(skore~typ);
summary(model);
TukeyHSD(model);
by(skore, typ, t.test);
```

10.4 Kapitola 5

Ke zjištění vlivu pohlaví jsem používal stejné příkazy jako v minulé kapitole, navíc jsem vyšetřoval zastoupení jednotlivých pohlaví na různých typech škol a relativní četnost dosaženého skóre podle těchto šesti skupin, uvádím pouze použitý t-test a příkazy ke získání dat do tabulky 5.2, konstrukce grafu 5.3 byla v základních krocích totožná s ostatními grafy relativních četností dosaženého skóre.

```
mean(skore[pohlavi=="zena"], na.rm=T)
   - mean(skore[pohlavi=="muz"], na.rm=T)
t.test(skore~pohlavi);
table(pohlavi, typ);
table(pohlavi);
table(typ, pohlavi)/c(table(typ),table(typ));
table(pohlavi)/length(skore);
```

10.5 Kapitola 6

Zdrojový kód pro data v tabulce 6.2.

```
table(typ, kraj)/rep(table(typ), 14);
table(kraj)/length(skore);
```

10.6 Kapitola 7

K postupům z předchozích kapitol zde přibyl graf znázorňující zastoupení typů škol podle vysokoškolského zaměření studenta.

```
typy.skol.vs<-table(typ, vsobor);
barplot(prop.table(typy.skol.vs,2));
```

10.7 Kapitola 8

Zdrojový kód pro mnou použitý lineární model.

```
summary(l<-lm(skore~typ + kraj+ skupina + pohlavi + znamka
   + matur + navs + vsprij + vsobor));
confint(l);
```

Literatura

[1] Anděl J. (1998): Statistické metody. Matfyzpress, Praha.

[2] Anděl J. (2005): Základy matematické statistiky. Matfyzpress, Praha.

[3] Dalgaard P. (2002): Introductory Statistics with R. Springer-Verlag, New York.

[4] Zdroj dat a dalších informací: CERMAT, http://www.cermat.cz