Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra matematiky

Bakalářská práce

Výsledky vstupních testů z matematiky a úspěšnost studia

Plzeň, 2014 Zuzana Rábová

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů.

V Plzni dne 29. 5. 2014

Zuzana Rábová

Poděkování

V první řadě bych chtěla velmi poděkovat svému vedoucímu bakalářské práce, Mgr. Michalu Frieslovi, Ph.D., za odborné vedení, ochotný přístup a přínosné rady během zpracování této práce. Poděkování patří i těm, kteří mě během mého dosavadního studia podporovali.

Abstrakt

Bakalářská práce se zabývá statistickým zpracováním dat výsledků ze vstupních testů z matematiky z akademického roku 2012/2013. Testy ze znalostí středoškolské matematiky se na Západočeské univerzitě v Plzni zavedly v roce 2006. Hlavním cílem práce je vyšetřit souvislosti mezi výsledky vstupních testů z matematiky pro rok 2012 a následnými výsledky studenta v matematických předmětech.

Klíčová slova: vstupní testy, kontingenční tabulka, t-test, logistická regrese

Abstackt

The thesis deals with statistical data processing of results of entrance tests in mathematics from the academic year 2012/2013. Tests of knowledge of high school mathematics at the University of West Bohemia in Pilsen was introduced in 2006. Primary intention of this work is to investigate the relation between the results of entrance tests in mathematics in 2012 and next outcomes of students in math subjects.

Keywords: entrance tests, contingency table, t-test, logistic regression

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Cílem této bakalářské práce je zpracování vstupních testů z matematiky na Západočeské univerzitě v Plzni, které se psaly v akademickém roce 2012/2013. Práce se zaměřuje na zkoumání závislosti výsledku u vstupního testu a následné úspěšnosti u zkoušky z matematického předmětu v zimním semestru.

Práce je členěna do několika hlavních kapitol. Ve druhé kapitole je provedeno seznámení se vstupními testy, s jejich strukturou a účelem. V kapitole jsou také popsány použitý software a charakter samotných dat.

Obsahem třetí kapitoly je, pro obecnou představu o datech, základní statistické zpracování. V bakalářské práci byly použity základní prvky popisné statistiky. Statistické zpracování doprovází také vhodné grafické výstupy.

Další kapitoly se již zabývají samotným zkoumáním závislosti mezi výsledkem u vstupního testu z matematiky a úspěchem u zkoušky v zimním semestru. Čtvrtá kapitola je věnována testu nezávislosti v kontingenčních tabulkách. Do pozorování byly zahrnuty i další možné ovlivňující faktory.

V následující kapitole pro zjištění vztahu mezi výsledky ze vstupního testu u studentů, kteří uspěli a neuspěli u zkoušky, byl použit t-test pro dva nezávislé výběry.

V závěrečné kapitole je sestaven model logistické regrese s více ovlivňujícími faktory, kapitola se mimo jiné zabývá testováním podmodelu.

2 Vstupní testy z matematiky a příslušná data

2.1 Vstupní testy z matematiky

Vstupní testy z matematiky se na Západočeské univerzitě píší od roku 2006. Test se v roce 2012 psal na osmi fakultách Západočeské univerzity, Fakultě aplikovaných věd, Fakultě elektrotechnické, Fakultě strojní, Fakultě ekonomické, Fakultě filozofické, Fakultě pedagogické, Fakultě zdravotnických studií a Ústavu umění a designu, a na jedné fakultě Jihočeské university v Českých Budějovicích, konkrétně Přírodovědecké fakultě. Hromadný test je zaměřen na základní znalosti středoškolské matematiky studentů přihlášených k bakalářskému studiu. Matematický test se skládá z devíti otázek, které mají čtyři možné odpovědi, z nichž právě jedna je správná, a jedné otevřené otázky. Celkový čas na vypracování je 20 minut. Za každou správnou odpověď může student získat právě jeden bod. Kromě odpovědí student vyplňuje základní osobní údaje jako fakultu, název a typ střední školy, rok ukončení střední školy a zda skládal maturitu z matematiky. Vstupní znalosti středoškoláků jsou velmi rozdílné a cílem testování je identifikace oblastí matematiky, ve kterých si studenti nejsou jisti a upravit tak možný obsah vyučovaných předmětů, aby se nestávaly důvodem k ukončení studia na ZČU.

2.2 Vstupní data a použitý software

U poskytnutých dat bylo u každého studenta několik informací, jako například studovaný předmět, fakulta, ročník, obor, typ maturity, typ střední školy, pohlaví, kraj a další. Bylo nutné se nejprve ve vstupních datech zorientovat a částečně je upravit. Hlavním cílem bylo zjistit závislost mezi výsledky testů a bezprostřední úspěšností studia, proto bylo nutné vybrat data u studentů, kteří se dále věnovali matematickým předmětům zakončeným zkouškou. V roce 2012 se testování zúčastnilo 2008 studentů, z nichž 1499 si zapsalo v zimním semestru jeden z předmětů M1, M1S, M1E, MA1, ZM1. Studenti těchto předmětů mají zastoupení na Fakultě aplikovaných věd o počtu 339 studentů, Fakultě ekonomické 530, Fakultě elektrotechnické 382, Fakultě pedagogické 22 a Fakultě strojní o 226 studentech. Tabulky, histogramy a základní prvky popisné statistiky byly v práci

vypracovány v MS Office Excel. Výpočty kontingenčních tabulek, vykreslování grafů a testování statistických hypotéz bylo provedeno v softwaru MATLAB.

2.3 Doplnění dat

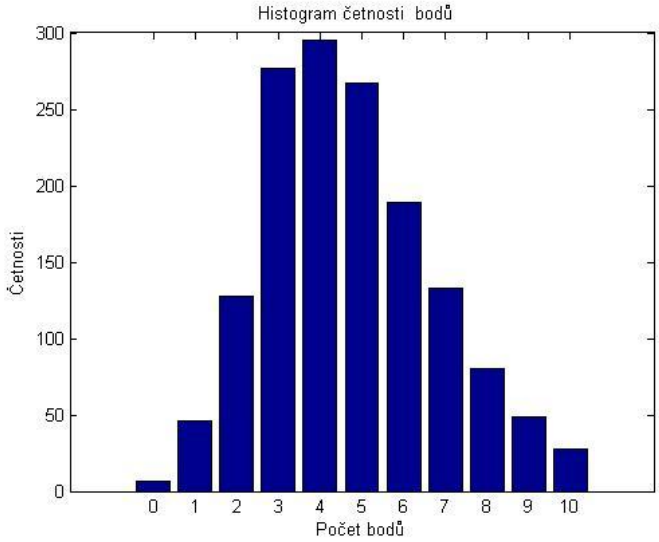
U vybraných studentů, kteří měli v akademickém roce 2012/2013 zapsaný matematický předmět zakončený zkouškou, jsme měli k dispozici známku ze zkoušky. Avšak ne u všech studentů se dalo rozlišit, zda u zkoušky neuspěli, tedy měli výslednou známku čtyři, nebo se zkoušky vůbec neúčastnili. Těmto studentům byla přiřazena známka čtyři. V dalším textu jsou brána jako základní soubor data u předmětů M1, M1E, M1S, MA1 a ZM1.

3 Základní zpracování dat

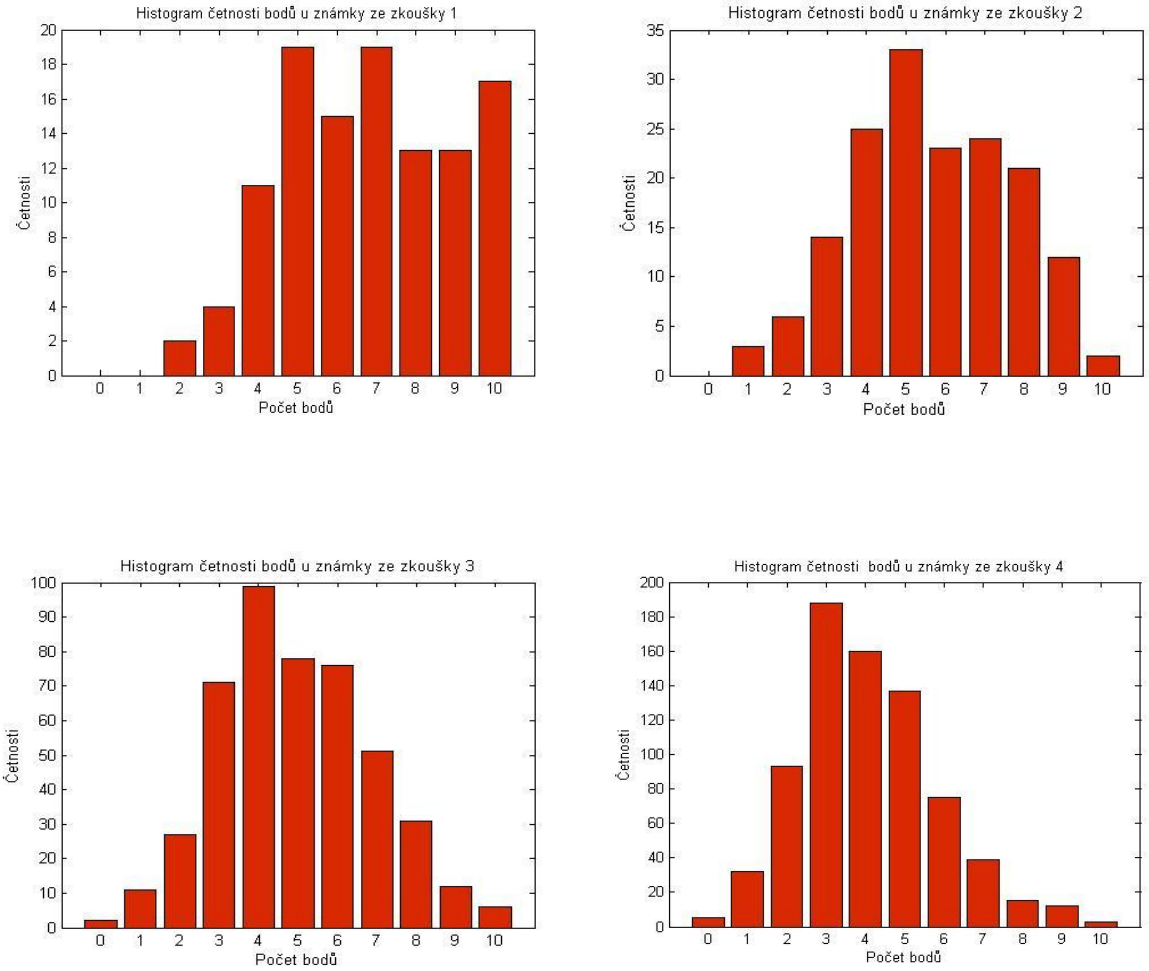
Pro další pozorování bylo nejprve nutné podívat se na dostupná data jako na celek. Proto se v této kapitole budeme zabývat základní popisnou statistikou. Informace o studentech matematických předmětů byly zpracovány ve formě tabulek a grafů. Byly vypočítány číselné charakteristiky jako průměr, medián, modus, rozptyl, směrodatná odchylka a kvantily. Všechny vypočtené údaje jsou zaznamenány v Tabulce 3.1. Pozorovaný soubor má rozsah 1499, ten byl rozdělen do čtyř skupin podle získané známky u zkoušky v zimním semestru. Na Obrázku 3.2 a 3.3 jsou zachyceny histogramy četností počtu bodů ze vstupního testu a následné známky u zkoušky z matematického předmětu. Pro vykreslení grafů byl použit v softwaru MATLAB příkaz *hist()*. Z Tabulky 3.1 můžeme vypozorovat, že v celkovém souboru bylo dosaženo nejčastěji celkového počtu bodů ze vstupního testu čtyři z možných deseti, stejně jako u studentů se známkou ze zkoušky čtyři. Průměrně nejlepších výsledků z testu dosáhli ti studenti, kteří měli poté jedničku u zkoušky.

		Uspěl			*Neuspěl*
	Celkem	**ZK 1**	**ZK 2**	**ZK 3**	**ZK 4**
počet studentů	1499	113	163	464	759
průměrný počet bodů	4,72	6,81	5,67	4,94	4,07
medián	4	7	6	5	4
modus	4	7	5	4	3
rozptyl	4,18	4,58	4,05	3,77	3,11
směr. odchylka	2,05	2,14	2,01	1,94	1,76
dolní kvartil	3	5	4	4	3
horní kvartil	6	9	7	6	5

Tabulka 3.1: Základní statistické údaje

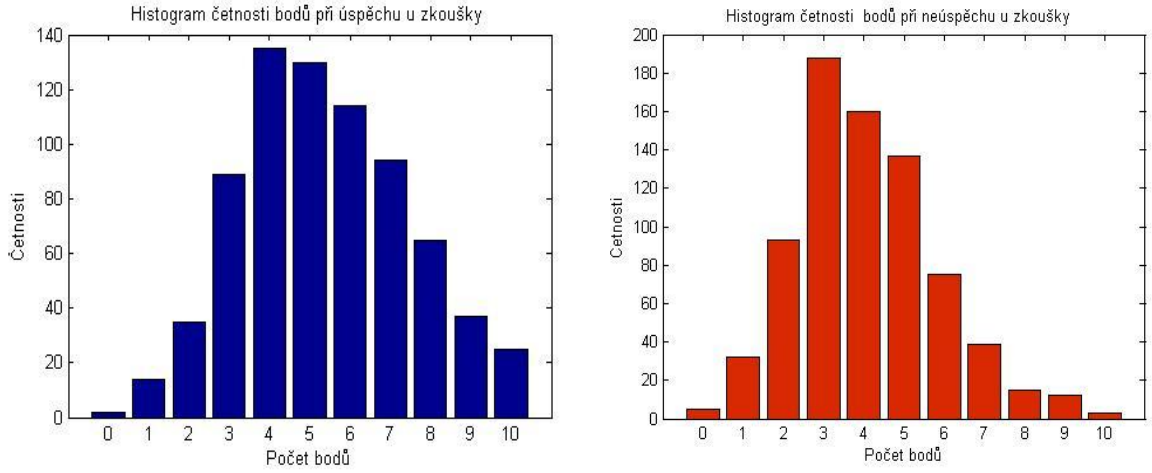


Obrázek 3.1: Histogram četností pro soubor o rozsahu 1499

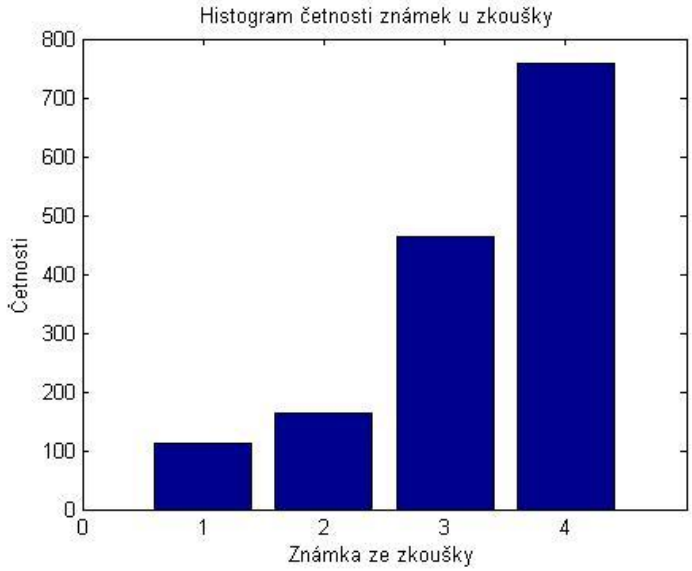


Obrázek 3.2: Zobrazení histogramů podle dosažené známky v zimním semestru

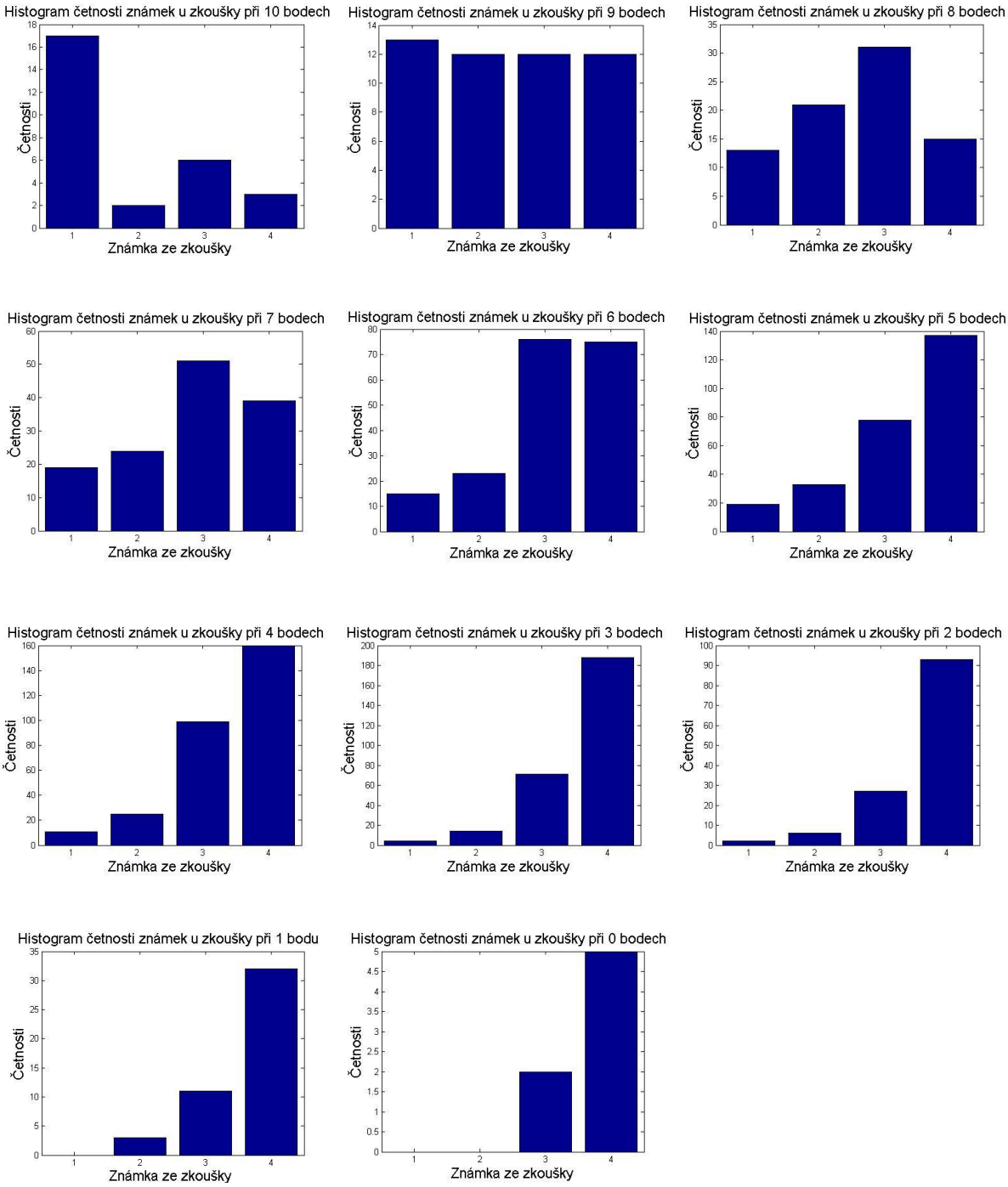
Zobrazené histogramy v Obrázku 3.3 jsou rozděleny podle toho, zda student v zimním semestru 2012 u zkoušky uspěl nebo neuspěl. Vlevo je histogram pro studenty, kteří uspěli a vpravo pro ty, co neuspěli.



Obrázek 3.3: Histogramy podle úspěchu u zkoušky.



Obrázek 3.4: Histogram četností získaných známek ze zkoušky v zimním semestru z matematického předmětu



Obrázek 3.5: Histogramy rozdělení známek v závislosti na dosaženém počtu bodů ze vstupního testu

V Obrázku 3.5 jsou vykresleny histogramy četností známek ze zkoušky, které jsou rozděleny podle získaného počtu bodů ze vstupního testu z matematiky. Počty studentů, kteří dosáhli určitého zisku bodů, jsou zaznamenány v Tabulce 3.2.

Počet bodů	10	9	8	7	6	5	4	3	2	1	0
Počet studentů	28	49	80	133	189	267	295	277	128	46	7

Tabulka 3.2: Četnosti celkového počtu bodů ze vstupního testu

4 Kontingenční tabulky

V této kapitole se zabýváme vyšetřením závislosti mezi získaným počtem bodů ze vstupního testu z matematiky a získanou známkou u zkoušky v zimním semestru. Mezi sledované znaky byla zahrnuta fakulta, kterou student navštěvuje. Jako nástroj pro zjištění závislosti byly použity kontingenční tabulky. V praxi vzniká dvourozměrná kontingenční tabulka sledováním dvou znaků, např. Y, Z. Sledované znaky mohou nabývat buď diskrétních hodnot, nebo spojitý charakter sami převedeme na konečně mnoho tříd. Obvykle bývá vhodnější pro spojitá data využít metody korelační analýzy. Teoretickým základem pro tvorbu tabulek jsou matice pravděpodobností pro dvourozměrné náhodné vektory. Hodnoty $n_{ij.}$ jsou počty těch případů, kdy se ve výběru vyskytla dvojice (i, j). Náhodné veličiny $n_{ij.}$ pak mají sdružené multinomické rozdělení s parametrem n a s pravděpodobnostmi $p_{ij.}$ Čísla $n_{r.}$ a $n_{.c}$ se nazývají marginální četnosti, viz Tabulka 4.2. Podrobněji je vyloženo v [1] nebo [2].

	Z			
Y	1	...	c	Σ
1	p_{11}	...	p_{1c}	$p_{1.}$
...	...	...	...	...
r	p_{r1}	...	p_{rc}	$p_{r.}$
Σ	$p_{.1}$	...	$p_{.c}$	1

Tabulka 4.1: Matice pravděpodobností

	Z			
Y	1	...	c	Σ
1	n_{11}	...	n_{1c}	$n_{1.}$
...	...	...	...	...
r	n_{r1}	...	n_{rc}	$n_{r.}$
Σ	$n_{.1}$	...	$n_{.c}$	n

Tabulka 4.2: Kontingenční tabulka o r řádcích a c sloupcích

4.1 Test nezávislosti

Nejčastějším typem úlohy pro rozbor kontingenčních tabulek je test hypotézy, že dva sledované znaky jsou na sobě nezávislé, kdy jeden znak odpovídá sloupci a druhý řádku. Testem nezávislosti se zabývá například [1].

Věta 4.3: Nechť Y a Z jsou diskrétní veličiny. Y a Z jsou nezávislé tehdy a jen tehdy, platí-li $p_{ij} = p_{i.}\ p_{.j}$ pro všechny dvojice (i, j). Čísla $p_{i.}$ a $p_{.j}$ se nazývají marginální pravděpodobnosti.

Nyní můžeme formulovat hypotézy o nezávislosti ve tvaru

$$H0: p_{ij} = p_{i.}\, p_{.j}, \quad i = 1, \ldots, r; j = 1, \ldots, c,$$
$$H1: p_{ij} \neq p_{i.}\, p_{.j}, \quad i = 1, \ldots, r; j = 1, \ldots, c.$$

Veličina (testovací statistika)

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{4.1}$$

má při n $\rightarrow \infty$ asymptoticky rozdělení chí-kvadrát χ^2 s počtem stupňů volnosti $rc - (r + c - 2) - 1 = (r - 1)(c - 1)$, kde r je počet řádků kontingenční tabulky a c počet sloupců. Pokud vyjde, že testovací statistika je větší nebo rovna kvantilu chí-kvadrát rozdělení se stupni volnosti $(r - 1)(c - 1)$, tedy $\chi^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)$, tak nulovou hypotézu o nezávislosti veličin Y a Z s chybou 1. druhu α zamítáme. Naopak pokud platí, že $\chi^2 < \chi^2_{(r-1)(c-1)}(1-\alpha)$, tak nulovou hypotézu H_0 nezamítáme. Pro shodu s limitním rozdělením se vyžaduje, aby očekávané četnosti $\frac{n_{i.}n_{.j}}{n}$ byly větší než pět, pokud tato podmínka neplatí, tak obvykle dochází ke slučování některých sloupců nebo řádků, jak bude uvedeno v dalším textu.

4.2 Test nezávislosti počtu bodů z testu a známky u zkoušky

Aby se mohla porovnat shoda s limitním rozdělením, je nutné mít očekávané četnosti větší než pět. Z tohoto důvodu byl použit obvyklý postup slučování tříd. Počty bodů z testu byly rozděleny podle následující tabulky.

špatný test	0-3 body
průměrný test	4-6 bodů
dobrý test	7-10 bodů

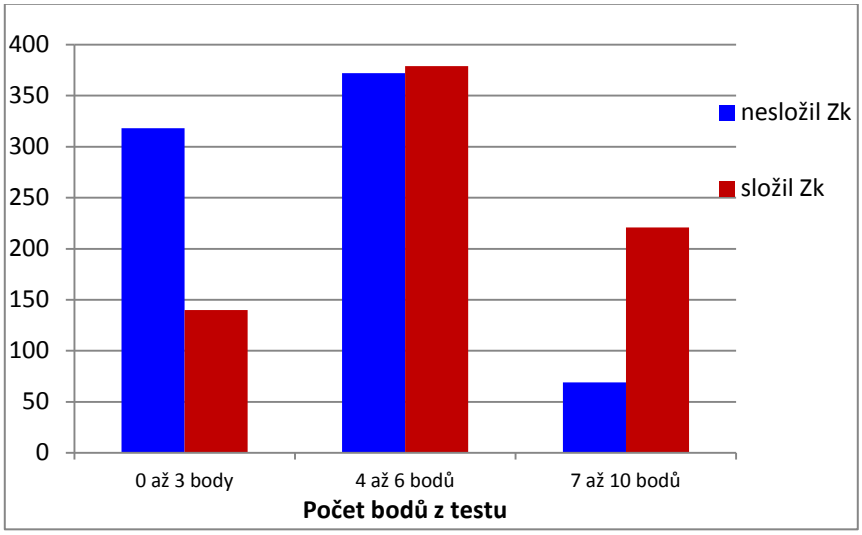
Tabulka 4.3: Počty bodů rozdělené do tří tříd

Úspěch u zkoušky byl rozdělen na dvě kategorie uspěl/ neuspěl.

Výsledky

Kontingenční tabulky, testovací statistiky a p-hodnoty byly vypočítané pomocí softwaru Matlab pomocí příkazu *[table,chi2,p] = crosstab(x2,x1)*, kontingenční grafy byly vykresleny v MS Excel.

Nejprve byla testována nezávislost počtu bodů z testu a úspěšnost studenta u zkoušky. Na základě analýzy sestavené kontingenční tabulky byla získána testovací statistika $\chi^2 = 148{,}70$ a příslušná p-hodnota $p = 5{,}14*10^{-33}$. Z porovnání kritické hodnoty a testovací statistiky $\chi^2 > \chi^2_{(2)}(0{,}95) = 5{,}99$ na hladině významnosti 5% byla zamítnuta nulová hypotéza o nezávislosti počtu bodů z testu a úspěšnosti u zkoušky.



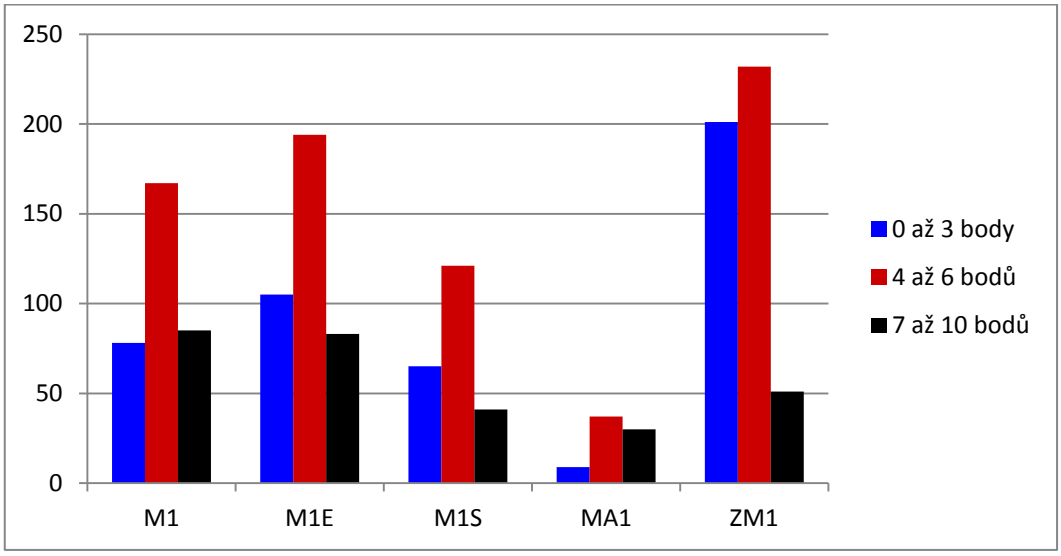
Obrázek 4.1: Kontingenční graf, porovnání výsledků u zkoušky

Do úvah byly zahrnuty i další možné faktory, jako je například předmět, na který je student zapsán, případně fakulta, kterou student navštěvuje. Byl proveden test nezávislosti mezi počtem bodů a zapsaným předmětem, resp. fakultou. Dále byl soubor rozdělen podle fakult. Pro každou fakultu byl vyhodnocen test nezávislosti mezi počtem bodů z testu a úspěchem u zkoušky. Opět došlo ke slučování tříd podle Tabulky 4.3.

Výsledky

Nyní bylo cílem testování zjistit závislost mezi zapsaným předmětem a získaným počtem bodů ze vstupního testu z matematiky. Testovací statistika vyšla $\chi^2 = 79{,}06$ a příslušná p-hodnota $p = 7{,}55*10^{-14}$.

Z porovnání kritické hodnoty a testovací statistiky $\chi^2 > \chi^2_{(8)}(0{,}95) = 15{,}5$ na hladině významnosti 5% zamítáme nulovou hypotézu o nezávislosti mezi zapsaným předmětem a získaným počtem bodů.



Obrázek 4.2: Kontingenční graf, porovnání výsledků z testu a zapsaného předmětu

Pro testování závislosti druhu fakulty na počtech bodů se mohly očekávat podobné výsledky jako u testu závislosti předmětu a získaným počtem bodů, neboť předměty z větší části vypovídají i o druhu fakulty, kterou student navštěvuje, jak je doloženo v Tabulce 4.4.

	M1	**M1E**	**M1S**	**MA1**	**ZM1**
FAV	285	-	1	56	-
FEK	46	-	-	-	486
FEL	-	385	-	-	-
FPE	23	-	-	-	-
FST	-	-	230	-	-

Tabulka 4.4: Rozdělení studentů podle fakult a zapsaných předmětů

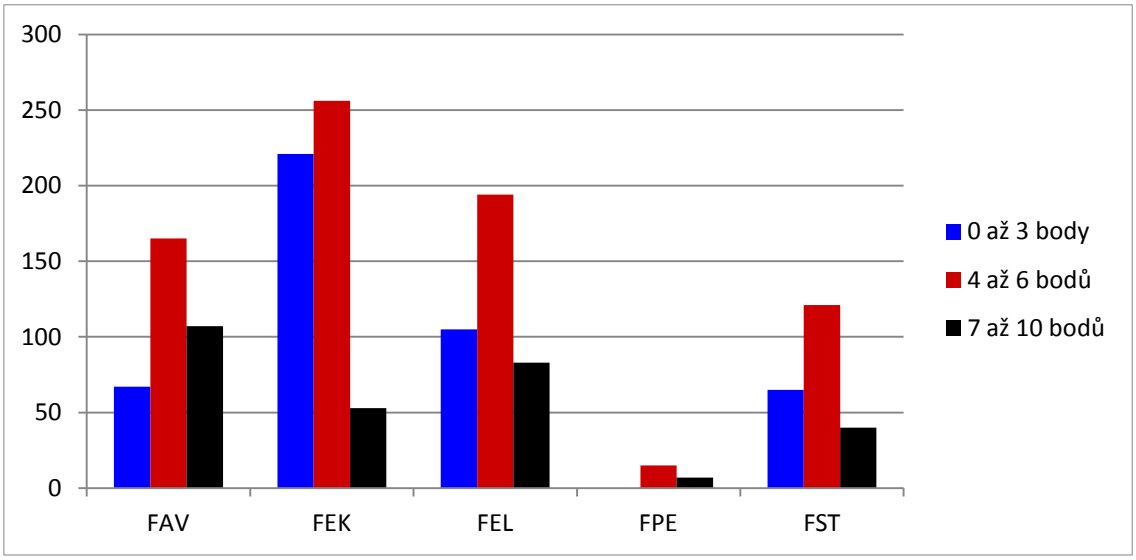
Při testování závislosti mezi počtem bodů a fakultou byla opět zamítnuta nulová hypotéza o nezávislosti na základě těchto výsledků $\chi^2 = 98{,}37 > 15{,}5$ a $p = 9{,}18*10^{-18}$.

V dalším pozorování se tedy zaměříme na rozdělení podle fakult. U každé fakulty byly dány do kontingenční tabulky dva znaky, počet bodů (opět proběhlo sloučení tříd, viz Tabulka 4.3) a uspěl, resp. neuspěl u zkoušky. Pomocí software Matlab byly získány výsledky uvedené v Tabulce 4.5.

	χ^2	p
FAV	63,02	<0,0001
FEK	35.35	<0,0001
FEL	42,09	<0,0001
FPE	5,46	0,0200
FST	17.90	0,0001

Tabulka 4.5: Hodnoty testovací statistiky a příslušné p-hodnoty pro test nezávislosti počtu bodů a úspěchu u zkoušky, rozdělení podle fakult

U pedagogické fakulty bylo rozdělení tříd odlišné, testu se zúčastnilo pouze 22 studentů. Rozdělení bodů tedy bylo do dvou tříd 4 až 6 bodů a 7 až 10 bodů. Tímto postupem vznikla tabulka 2x2, u které však očekávané četnosti byly pod hranicí 5. U většiny provedených testů je testovací statistika výrazně větší než kritická hodnota $\chi^2_{(2)}(0{,}95) = 5{,}99$.



Obrázek 4.3: Kontingenční graf, znázornění rozdělení četností bodů podle druhu fakulty

5 T-test

Pro další analýzu našich dat byl zvolen t-test. V páté kapitole se tedy budeme zabývat testováním statistických hypotéz pomocí t-testu. Soubor výsledných známek u zkoušky z matematického předmětu v zimním semestru byl rozřazen podle celkového počtu bodů ze vstupního testu na dva výběry.

Veličiny $X_1, X_2, \ldots, X_n$ *a* $Y_1, Y_2, \ldots, Y_m$ jsou získané známky ze zkoušky z matematického předmětu v zimním semestru pro rozpětí bodů ze vstupního testu 0 až 5 a 6 až 10. Navíc předpokládáme, že veličiny jsou navzájem nezávislé.

5.1 t-test pro dva libovolné nezávislé výběry

Pro otestování shody středních hodnot na hladině významnosti α využijeme t-test pro dva libovolné nezávislé výběry. Rozptyly těchto veličin jsou neznámé reálné hodnoty, tyto hodnoty nemusí být rovny a rozdělení veličin X a Y nemusí být obecně normální. Rozsahy jednotlivých veličin se mohou lišit. [2] V softwaru Matlab je pro tento test definována funkce *ttest2(X, Y, [], [], 'unequal').* Testovací statistika pro tento test má následující tvar

$$t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}, \tag{5.1}$$

kde $\bar{x}, \bar{y}$ jsou výběrové průměry a s_x^2, s_y^2 výběrové rozptyly. Statistika t má za obecných předpokladů asymptoticky normální normované rozdělení (pro n, m $\rightarrow \infty$).

Hypotézy o středních hodnotách budeme formulovat ve tvaru

$$H_0: \mu_x = \mu_y,$$
$$H_1: \mu_x \neq \mu_y,$$

kde μ_x je střední hodnota veličiny X a μ_y je střední hodnota veličiny Y.

Nulovou hypotézu o shodě středních hodnot zamítneme, pokud $|t| > u_{1-\alpha/2}$, kde $u_{1-\alpha/2}$ je kvantil normálního rozdělení.

Výsledky

Rozsah výběru X je 1020. To jsou známky studentů, kteří získali 0 až 5 bodů ze vstupního testu. Výběr Y má přibližně rozsah o polovinu nižší 479, to jsou studenti se ziskem 6 až 10 bodů.

Testovací statistika t vyšla 11,99 a p-hodnota t-testu $2{,}00*10^{-30}$.

Na základě tohoto testu na hladině významnosti $\alpha = 0{,}05$ zamítáme nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_1.
Protože $11{,}99 > u_{0{,}975} = 1{,}960$.

6 Logistická regrese

Cílem kapitoly je pomocí vhodného logistického modelu vyjádřit závislost úspěchu u zkoušky v zimním semestru akademického roku 2012/2013 na různých ovlivňujících faktorech. V logistické regresi, oproti klasické lineární regresi, se setkáváme se situací, kdy závislá proměnná Y_i má diskrétní charakter, v našem případě závislá proměnná může nabývat dvou hodnot *uspěl*, resp. *neuspěl* u zkoušky, proto jí budeme nazývat binární proměnnou.

Nejprve se krátce seznámíme se základním vymezením pojmů logistické regrese.

Nezávislé náhodné veličiny $Y_1, \ldots, Y_n$ mají alternativní rozdělení s parametrem μ_i, který je shodný s pravděpodobnostmi jedniček a pro střední hodnotu platí $EY_i = \mu_i$.

Pravděpodobnost dvou možných hodnot $Y_i = 1$ a $Yi = 0$ lze psát dohromady jako

$$P(Y_i = j) = \mu_i^j (1-\mu_i)^{1-j},\ j=0, 1. \tag{6.1}$$

Pak věrohodnostní funkce má tvar

$$L(\mu) = \prod_{i=1}^{n} \mu_i^{Y_i} (1-\mu_i)^{1-Y_i} \tag{6.2}$$

a logaritmická věrohodnostní funkce lze zapsat jako

$$l(\mu) = \sum_{i=1}^{n} Y_i ln(\frac{\mu_i}{1-\mu_i}) + \sum_{i=1}^{n} ln(1-\mu_i). \tag{6.3}$$

Pozorované veličiny $Y_1, Y_2, \ldots, Y_n$ se v logaritmické věrohodnostní funkci projevují pouze v součinech $Y_i ln(\frac{\mu_i}{1-\mu_i})$, výraz uvnitř logaritmu lze interpretovat jako podíl pravděpodobnosti jedničky a nuly, tj. $\frac{P(Y_i=1)}{P(Y_i=0)}$. Anglický název pro tento podíl je *odds* a v češtině se užívá označení *šance*. Logaritmické transformaci šance se říká *logit (*značení

$\eta_i = ln\left(\frac{P(y_i=1)}{P(y_i=0)}\right)$ a v zobecněných lineárních modelech (GLM) se využívá jako případ linkovací funkce.

V logistickém regresním modelu předpokládáme lineární závislost *logitu* η_i na vysvětlujících proměnných $x_{1i}, x_{2i}, \ldots, x_{ki}$.

Dostáváme tedy model ve tvaru

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \qquad i = 1, \ldots, n. \tag{6.4}$$

Střední hodnotu (vztah pro logistickou regresi), můžeme zapsat ve tvaru

$$\mu_i(\beta) = \frac{1}{1+exp(-\beta_0-\beta_1 x_{1i}-\cdots-\beta_k x_{ki})} \qquad i = 1, \ldots, n. \tag{6.5}$$

Pro odhad parametrů lze užít metodu maximální věrohodnosti, odvození lze nalézt například v [3] nebo [4].
Pro logistickou analýzu byl využit opět software Matlab konkrétně funkce *glmval()* a g*lmfit(X,y,'binomial','link', 'logit')*. Logistickou regresí se blíže zabývají například [3] a [4].

6.1 Základní model logistické regrese

Nejprve budeme uvažovat základní model s pouze jednou nezávislou proměnnou a to získaným počtem bodů ze vstupního testu.
Jako výchozí byl zvolen model

$$\eta_i = \beta_0 + \beta_1 počet\_bodů_i \qquad i = 1, \ldots n, \tag{6.6}$$

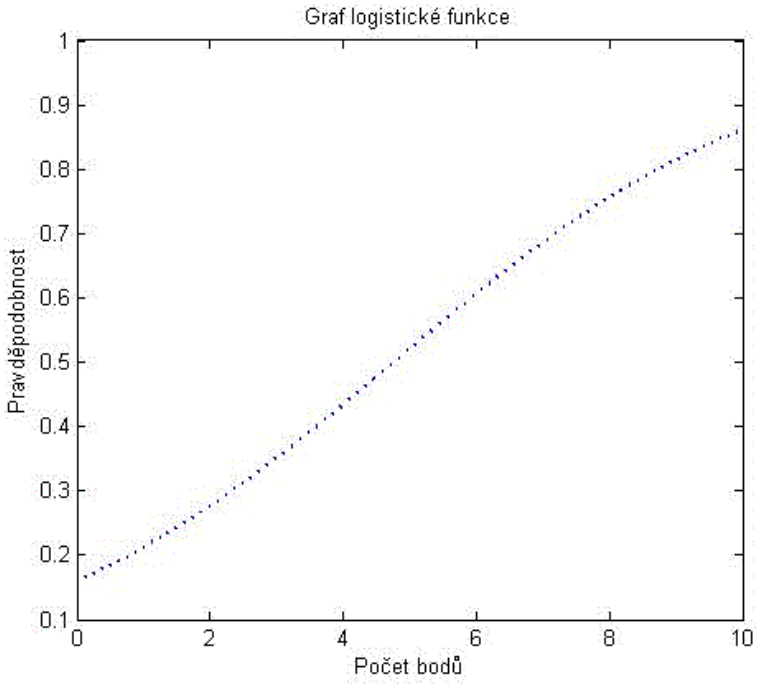
kde proměnná *počet_bodů*$_i$ nabývá hodnot 0, 1, …, 10.

Výsledky

Získané odhady regresních koeficientů a příslušné p-hodnoty t-testu pro otestování významnosti odhadů koeficientů jsou zaznamenány v Tabulce 6.1.

Regresní koeficienty	**Odhady regresních koeficientů**	**p-hodnota**
absolutní člen $\boldsymbol{\beta_0}$	-1,66	<0,0001
$\boldsymbol{\beta_1}$	0,35	<0,0001

Tabulka 6.1: Odhady koeficientů a p-hodnoty t-testu pro logistickou regresi



Obrázek 6.1: Logistická funkce

V Obrázku 6.1 je odhadnutá závislost na počtu bodů pro pravděpodobnost, že student u zkoušky z matematiky v zimním semestru uspěl (tečkovaně jsou znázorněny odhady pravděpodobnosti pro získaný počet bodů). Absolutní člen β_0 určuje posunutí S-křivky podél osy x a koeficient β_1 pak míru „strmosti“ v okolí bodu $[-\frac{\beta_0}{\beta_1}; 0{,}5]$.

6.2 Zahrnutí dalších ovlivňujících faktorů

Mezi další ovlivňující faktory byly zahrnuty fakulta, typ střední školy, maturita z matematiky, pohlaví a kraj.

Jednotlivé kategoriální proměnné zahrnují následující varianty. Mezi fakulty patří Fakulta aplikovaných věd, Fakulta elektrotechnická, Fakulta strojní, Fakulta ekonomická a Pedagogická fakulta.

V dotazníku v rámci vstupního testu vyplňovali studenti typ střední školy, mohli také doplnit slovně jiný typ. Kvůli velkému množství variant byly již v datech možnosti rozřazeny do 6 skupin a to na gymnázia, lycea, obchodní akademie, střední odborné školy, střední odborná učiliště a vyšší odborné školy.

Proměnná maturita z matematiky je binární, může tedy nabývat pouze dvou hodnot, skládal maturitu z matematiky, resp. neskládat maturitu z matematiky. Proměnná pohlaví je opět binární, nabývá hodnot muž a žena.

Studenti Západočeské univerzity pocházejí ze všech 14 krajů České republiky. Ne všechny kraje však mají velké zastoupení. Proto byly v dalším pozorování sloučeny ty kraje, které měly malé četnosti. Mezi ně patří Jihomoravský kraj (4), Královéhradecký kraj (2), Moravskoslezský kraj (1), Olomoucký kraj (1), kraj Hlavní město Praha (22), kraj Vysočina (3), Zlínský kraj (2), Liberecký kraj (14) a Pardubický kraj (1). V závorkách za jménem kraje je uveden počet studentů z daného kraje.

Datový soubor pro logistickou regresi s 6 proměnnými byl „očištěn“ o chybějící hodnoty, viz Tabulka 6.2. Celkový rozsah souboru je tedy 1367.

proměnná	**počet chybějících a nezahrnutých údajů**	**proměnná**	**počet chybějících a nezahrnutých údajů**
počet bodů	0	***maturita z matematiky***	17
druh fakulty	0	***pohlaví***	2
typ střední školy	63	***kraj***	90

Tabulka 6.2: Četnost chybějících hodnot v datech

Uvažujeme následující model

$$(\eta_{jklmn})_i = \beta_0 + \beta_1(počet\_bodů_{jklm})_i + \beta_{F_j} + \beta_{SŠ_k} + \beta_{M_l} + \beta_{P_m} + \beta_{K_n}, \quad (6.7)$$

kde jednotlivé proměnné jsou vysvětleny v Tabulce č. 6.3. Závislá proměnná $(\eta_{jklmn})_i$ je pro *i*-tého studenta ze skupiny studentů s *j*-tou fakultou, *k*-tou střední školou, *l*-tým pohlavím a *n*-tým krajem. Každému parametru odpovídají regresní koeficienty.

Proměnná	**Hodnota**	**Proměnná**	**Hodnota**	**Proměnná**	**Hodnota**
$\boldsymbol{F_1}$	FAV	$\boldsymbol{SŠ_3}$	OA	$\boldsymbol{P_2}$	žena
$\boldsymbol{F_2}$	FEK	$\boldsymbol{SŠ_4}$	SOS	$\boldsymbol{K_1}$	Jihočeský
$\boldsymbol{F_3}$	FEL	$\boldsymbol{SŠ_5}$	SOU	$\boldsymbol{K_2}$	Karlovarský
$\boldsymbol{F_4}$	FPE	$\boldsymbol{SŠ_6}$	VOŠ	$\boldsymbol{K_3}$	Plzeňský
$\boldsymbol{F_5}$	FST	$\boldsymbol{M_1}$	skládal	$\boldsymbol{K_4}$	Středočeský
$\boldsymbol{SŠ_1}$	G	$\boldsymbol{M_2}$	neskládal	$\boldsymbol{K_5}$	Ústecký
$\boldsymbol{SŠ_2}$	LYC	$\boldsymbol{P_1}$	muž	$\boldsymbol{K_6}$	ostatní kraje

Tabulka 6.3: Hodnoty proměnných logistické regrese

Hodnoty koeficientů pro *FST, VOŠ, neskládal, žena* a *ostatní kraje* byly parametrizovány jako

$$\beta_{F_5} = \beta_{SŠ_6} = \beta_{M_2} = \beta_{P_2} = \beta_{K_6} = 0. \quad (6.8)$$

Výsledky

Některé odhady regresních koeficientů pro nezávislé proměnné mohou být statisticky nevýznamné, viz Tabulka 6.3. Proto se v dalším textu budou uvažovat různé kombinace parametrů v modelu.

Koeficient	Odhad koeficientu	p-hodnota	Koeficient	Odhad koeficientu	p-hodnota
$\boldsymbol{\beta_0}$	-1,1838	0,0486	$\boldsymbol{\beta_{sš_4}}$	0,4091	0,3743
$\boldsymbol{\beta_1}$	0,3679	<0,0001	$\boldsymbol{\beta_{sš_5}}$	-0,1877	0,7099
$\boldsymbol{\beta_{F_1}}$	-1,2997	<0,0001	$\boldsymbol{\beta_{M_1}}$	0,3620	0,0102
$\boldsymbol{\beta_{F_2}}$	-0,8509	0,0004	$\boldsymbol{\beta_{P_1}}$	-0,5909	0,0005
$\boldsymbol{\beta_{F_3}}$	-0,4274	0,0309	$\boldsymbol{\beta_{K_1}}$	0,3993	0,2627
$\boldsymbol{\beta_{F_4}}$	-1,6237	0,0145	$\boldsymbol{\beta_{K_2}}$	0,1353	0,7002
$\boldsymbol{\beta_{sš_1}}$	0,5599	0,2423	$\boldsymbol{\beta_{K_3}}$	-0,0017	0,9957
$\boldsymbol{\beta_{sš_2}}$	0,1427	0,7715	$\boldsymbol{\beta_{K_4}}$	-0,2430	0,5616
$\boldsymbol{\beta_{sš_3}}$	0,0245	0,9598	$\boldsymbol{\beta_{K_5}}$	0,0195	0,9591

Tabulka 6.4: Odhady regresních koeficientů a příslušné p-hodnoty pro logistickou regresi

6.3 Testování podmodelu

V této podkapitole je cílem vytvořit vhodný rozšířený regresní logistický model, který by nejlépe vystihoval vysvětlovanou proměnnou, tedy úspěch u zkoušky v zimním semestru. Testováním podmodelu se blíže zabývá například [4]. Nejprve je třeba zavést několik vymezujících pojmů, pro testování vytvořeného podmodelu.

K testování podmodelu lze použít test poměrem věrohodnosti. Logicistický regresní model si označíme jako M a jeho odhady regresních koeficientů jako b. Jeho podmodel si označíme $\widetilde{M}$ a k němu příslušné odhady koeficientů $\widetilde{b}$. Podmodel můžeme dostat například vynecháním některých nezávislých proměnných. Nás tedy zajímá, zda se M a $\widetilde{M}$ významně liší. Jeden z možných způsobů, jak ověřit, že daný model odpovídá studovaným datům, je použití tzv. deviance.

Nejprve budeme uvažovat nejbohatší možný model, který obsahuje tolik parametrů, kolik je různých regresorů. Model s větší hodnotou věrohodnostní funkce neexistuje. Tento model budeme nazývat jako saturovaný a jeho maximální hodnotu věrohodnostní funkce označíme l_{max}. Každý jiný model je podmodelem právě tohoto saturovaného modelu.

Přiléhavost běžného modelu lze posuzovat pomocí deviance, která je definována následujícím vztahem

$$D(b) = 2\big(l_{max} - l(b)\big). \tag{6.9}$$

Pokud všechny hodnoty x_i jsou různé, pak $l_{max} = 0$ a devianci v modelu logistické regrese můžeme definovat jako

$$D(b) = -2l(b) = -2\sum_{i=1}^{n}(Y_i ln\widehat{\mu_i} + (1 - Y_i)\ln\,(1 - \widehat{\mu_i})). \tag{6.10}$$

Nyní v našem běžném modelu *M* budeme uvažovat podmodel ($\widetilde{M}$), který vznikl například vynecháním nějakého regresoru z modelu. Pak testovou statistiku testu poměrem věrohodností lze vyjádřit ve tvaru

$$T = D\big(\tilde{b}\big) - D(b). \tag{6.11}$$

Za platnosti testovaného podmodelu má testovací statistika T asymptoticky χ_f^2 rozdělení s počtem stupňů volnosti *f*, který se rovná rozdílu počtu nezávislých proměnných v porovnávaných modelech.

Výsledky

Nejprve byl zvolen základní model pouze s proměnnou *počet bodů* a k němu byly postupně přidávány jednotlivé proměnné, získané hodnoty deviancí jednotlivých modelů a podmodelu jsou zaznamenány v Tabulce 6.5.

Jako podmodel tedy uvažujeme

$$(\eta_{jklmn})_i = \beta_0 + \beta_1 počet\_bodů_i. \tag{6.12}$$

Zahrnuté proměnné	Deviance	Deviance podmodelu	p-hodnota
Fakulta	1683,2	1726,6	<0,0001
Střední škola	1714,9	1726,6	0,0387
Maturita z matematiky	1716,8	1726,6	0,0018
Pohlaví	1725,1	1726,6	0,2216
Kraj	1715,1	1726,6	0,0417

Tabulka 6.5: Hodnoty deviancí a p-hodnoty testu poměrem věrohodností

Tabulka 6.5 dokazuje, že přidání jednotlivých proměnných *fakulta, střední škola*, *maturita z matematiky* a *kraj* vede ke zlepšení kvality modelu. Na hladině 5% zamítáme hypotézu, že proměnné *fakulta*, *střední škola*, *maturita z matematiky* a *kraj* nemají vliv na větší vysvětlení závislé proměnné. Kvalita modelu se nejvíce zlepšila po přidání proměnné *fakulta*.

Nyní budeme uvažovat model v nejbohatším tvaru

$$(\eta_{jklmn})_i = \beta_0 + \beta_1 počet\_bodů_i + \beta_{F_j} + \beta_{SŠ_k} + \beta_{M_l} + \beta_{P_m} + \beta_{K_n}. \qquad (6.13)$$

Model bez	Deviance	Deviance podmodelu	p-hodnota
střední škola	1644,5	1658,8	0,0140
Kraj	1644,5	1650,3	0.3254
kraj* a *střední škola	1644,5	1664,5	0.0297
maturita za matematiky	1644,5	1651,1	0.0102
Pohlaví	1644,5	1656,9	0,0004
Fakulta	1644,5	1689,9	<0,0001

Tabulka 6.6: Podmodely nejbohatšího modelu

Z Tabulky 6.6 můžeme vyvodit, že největší změnu kvality modelu způsobilo vynechání proměnné *fakulta*, kdy změna deviance byla nejznatelnější.
Při testování podmodelu jsme zjistili, že vynechání proměnné *kraj* významně nezhorší kvalitu modelu.

Uvažujeme tedy model bez proměnné *kraj*

$$(\eta_{jklmn})_i = \beta_0 + \beta_1 počet\_bodů_i + \beta_{F_j} + \beta_{SŠ_k} + \beta_{M_l} + \beta_{P_m}. \qquad (6.14)$$

Model bez	Deviance	Deviance podmodelu	p-hodnota
fakulta	1650,3	1701,6	<0,0001
maturita z m.	1650,3	1657,4	0.0077
střední škola	1650,3	1664,5	0.0147
pohlaví	1650,3	1661,7	0,0007

Tabulka 6.7: Hodnoty deviancí a p-hodnoty testu poměrem věrohodností

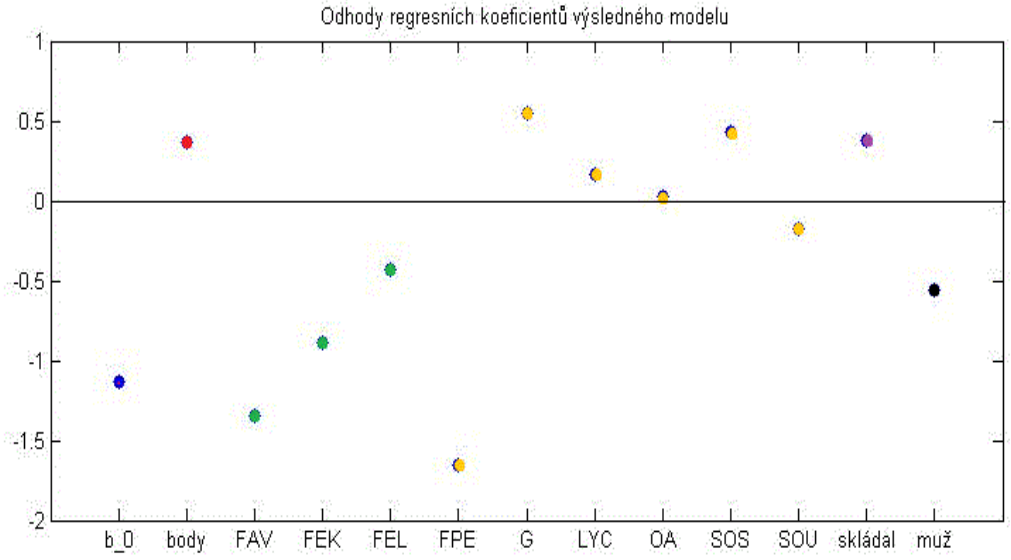
Proměnné fakulta, maturita, střední škola a pohlaví přispívají k vysvětlení závislé proměnné. Na základě testování podmodelu můžeme vyvodit, že vynechání kterékoli z těchto proměnných by významněji zhoršilo přiléhavost modelu, viz Tabulka 6.7. K největšímu zhoršení došlo při nezahrnutí proměnné *kraj*.

Celkový počet všech možných kombinací parametrů (bez proměnné *počet_bodů*, kterou do modelu zahrnujeme vždy) je 25. Při sledování jednotlivých přiléhavostí všech možných modelů, tedy pozorování příslušných deviancí, jsme zjistili, že nejnižší hodnotu deviance má model právě s kombinací proměnných *počet bodů*, *fakulta*, *střední škola*, *maturita z matematiky* a *pohlaví*. Hodnoty všech deivancí jsou uvedeny v tištěných přílohách.

Koeficient	Odhad koeficientu	p-hodnota	Koeficient	Odhad koeficientu	p-hodnota
β_0	-1,1339	0,0279	$\beta_{sš_2}$	0,1697	0,7307
β_1	0,3672	<0,0001	$\beta_{sš_3}$	0,0230	0,9626
β_{F_1}	-1,3444	<0,0001	$\beta_{sš_4}$	0,4262	0,3581
β_{F_2}	-0,8890	0,0001	$\beta_{sš_5}$	-0,1756	0,7292
β_{F_3}	-0,4344	0,0266	β_{M_1}	0,3737	0,0077
β_{F_4}	-1,6528	0,0115	β_{P_1}	-0,5620	0,0009
$\beta_{sš_1}$	0,5507	0,2531			

Tabulka 6.8: Odhady regresních koeficientů a příslušné p-hodnoty pro logistickou regresi

Odhady regresních koeficientů výsledného modelu logistické regrese pro vysvětlující proměnnou úspěch u zkoušky jsou zaznamenány v Tabulce 6.8. Pro srovnání hodnot koeficientů byly také vyneseny do grafu v Obrázku 6.2



Obrázek 6.2: Odhady regresních koeficientů výsledného modelu

V Obrázku 6.2 jsou barevně odlišeny odhady regresních koeficientů pro jednotlivé proměnné. Modře je vynesen odhad pro absolutní člen a červeně pro *počet bodů.* Zeleně jsou vyznačeny koeficienty proměnné *fakulta*, které jsou vůči nulové *FST* (Fakulta strojní), žlutě koeficientu pro *střední školu* vůči nulové *VOŠ* (vyšší odborná škola), fialově *maturita z matematiky* vůči *neskládal* maturitu z matematiky a černě *pohlaví* vůči *ženě*. V obrázku

je také vyznačena nulová osa, která odpovídá právě proměnným *FST, VOŠ, neskládal* a *žena*.

Pokud se zaměříme na srovnání odhadů regresních koeficientů mezi jednotlivými proměnnými ve výsledném modelu, tak do výsledného modelu, na základě testování podmodelu, byla sice zahrnuta proměnná *střední škola*, ale její vliv na vysvětlení závislé proměnné (úspěch u zkoušky) není takový jako například proměnná *fakulta*, viz Obrázek 6.2. Když se porovná hodnota regresního koeficientu u nezávislé proměnné *počet bodů* v modelu (6.11) a ve výsledném modelu, tak odhad regresního koeficientu u této proměnné se pro výsledný model zanedbatelně zvýšil o hodnotu 0,0172.

Závěr

Cílem této bakalářské práce bylo vyšetřit závislost výsledků vstupních testů z matematiky z akademického roku 2012/2013 a následného úspěchu u zkoušky z matematického předmětu v zimním semestru.

Nejprve bylo nutné se na data podívat jako celek a provést základní statistické zpracování. Byly vypočítány základní charakteristiky jako průměr, medián, modus a další. Pro základní představu byly vykresleny histogramy četností. Histogramy byly například rozděleny podle zisku bodů a známky u zkoušky.

V dalších kapitolách se již práce zaměřila na posuzování závislosti výsledků u vstupních testů a následným úspěchem u zkoušky z matematického předmětu. Byla testována hypotéza, zda jsou dvě veličiny nezávislé. Z testu nezávislosti v kontingenčních tabulkách jsme pokaždé hypotézu o nezávislosti zamítli. Následně byly všechny výsledky u zkoušky rozděleny do dvou výběrů podle toho, kolik student získal bodů u vstupního testu. Testovali jsme hypotézu o shodě středních hodnot těchto dvou výběrů. Na hladině významnosti 5 % jsme ji zamítli.

V závěru práce byl sestaven model logistické regrese. Nejprve do sestaveného modelu byla zahrnuta pouze proměnná vyjadřující počet bodů, příslušný odhadnutý regresní koeficient byl statistický významný. Jako výsledný byl vybrán model obsahující proměnné *počet bodů, střední škola, fakulta, maturita z matematiky* a *pohlaví*.

Na základě všech provedených metod byla zjištěna závislost mezi výsledkem u vstupního testu a následným úspěchem u zkoušky. Všechny závěry platí pouze pro studenty, kteří psali vstupní test a kteří měli v akademickém roce 2012/2013 v zimním semestru zapsaný matematický předmět.

Literatura

[1] Anděl, J. *Základy matematické statistiky*. MATFYZPRESS, 2007

[2] Reif, J. *Metody matematické statistiky*. Západočeská univerzita v Plzni, 2004

[3] Agresti, A. *Categorical Data Analysis*. Copyright, 2002

[4] Zvára, K. *Regrese*. MATFYZPRESS, 2008

A Přílohy

Přílohy tištěné

Proměnné	Deviance modelu	Deviance podmodelu	p-hodnota
BF	1683,20	1726,60	<0,0001
BS	1714,90	1726,60	0,0387
BM	1716,80	1726,60	0,0018
BP	1725,10	1726,60	0,2216
BK	1715,10	1726,60	0,0417
BMP	1713,70	1726,60	0,0015
BFP	1673,30	1726,60	<0,0001
BFM	1674,40	1726,60	<0,0001
BFS	1668,80	1726,60	<0,0001
BFK	1677,90	1726,60	<0,0001
BSM	1707,70	1726,60	0,0043
BSP	1710,00	1726,60	0,0110
BSK	1703,80	1726,60	0,0113
BMK	1707,60	1726,60	0,0041
BPK	1711,20	1726,60	0,0176
BFMP	1664,50	1726,60	<0,0001
BFSM	1661,70	1726,60	<0,0001
BFSP	1657,40	1726,60	<0,0001
BFSK	1663,50	1726,60	<0,0001
BFMK	1669,70	1726,60	<0,0001
BFPK	1667,00	1726,60	<0,0001
BSMP	1701,60	1726,60	0,0008
BSMK	1698,00	1726,60	0,0026
BSPK	1696,80	1726,60	0,0017
BMPK	1701,80	1726,60	0,0008
BFSMP	1650,30	1726,60	<0,0001
BFSMK	1656,90	1726,60	<0,0001
BFSPK	1651,10	1726,60	<0,0001
BFMPK	1658,80	1726,60	<0,0001
BSMPK	1689,90	1726,60	0,0002

Tabulka A.2: Hodnoty deviancí všech variant proměnných, ve srovnání s modelem s proměnnou *počet bodů*

Značení: B-počet bodů, F – fakulta, S – střední škola, M – maturita z matematiky, P - pohlaví, K - kraj

Přílohy na CD

- Bakalářská práce
- Obrázky vygenerované programem Matlab
- Vstupní data
- Zadání vstupního testu z matematiky pro akademický rok 2012/2013
- Zdrojové kódy v Matlabu
- Výpočty v MS Excel