UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

Sport, statistika a sázky

Vedoucí diplomové práce:
Mgr. Jaroslav Marek, Ph.D.
Rok odevzdání: 2010

Vypracoval:
Petr Dokoupil
M-E, III. ročník

Prohlášení

Prohlašuji, že jsem vytvořil tuto diplomovou práci samostatně za vedení Mgr. Jaroslava Marka, Ph.D. a že jsem v seznamu použité literatury uvedl všechny zdroje použité při zpracování práce.

V Olomouci dne 15. dubna 2010

Poděkování

Rád bych na tomto místě poděkoval vedoucímu diplomové práce Mgr. Jaroslavovi Markovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále si zaslouží poděkování můj počítač, že vydržel moje pracovní tempo, a typografický systém TEX, kterým je práce vysázena.

Obsah

Úvod

V souvislosti se slovy sport a statistika nás může napadnout, že propojení poznatků a znalosti obojího by mohlo jít využít při kurzovém sázení na sportovní výsledky. A to je hlavním cílem této práce. Využití vhodných pravděpodobnostních modelů a matematické statistiky k analýze sportovních výkonu a výsledků sportovních utkání bude pro celou práci podstatné.

Při sledování vývoje ligových tabulek ve fotbale, v hokeji či v basketbale, nebo při sledování individuálních výkonů jednotlivců můžeme využít vhodný statistický model. Ten nám pak umožní určit pravděpodobnosti výhry onoho družstva či jednotlivce a získané odhady použít při kurzovém sázení. Bez použití vhodného modelu je to pouze naše domněnka, na jejímž podkladě vyplňujeme co možná nejlépe vyhrávající tiket. K modelování lze využít vhodné rozdělení pravděpodobností a odhadnuté parametry lze použít k predikci výsledků. Parametry těchto rozdělení jsou ovlivněny mnoha faktory a jsou odlišné pro různé týmy. K jejich nalezení lze užít různých statistických metod. Tato bakalářská práce může být také použita pro demonstrování aplikace různých instrumentů teorie pravděpodobnosti a matematické statistiky.

V práci budu predikovat vítězné výkony v atletice při nadcházejících Letních olympijských hrách v Londýně v r. 2012. Dále se v práci budu zabývat sledováním a porovnáváním individuálních výkonů hráčů v NHL.

Hlavním cílem práce bude určení pravděpodobností výhry jednotlivých týmů v ligových fotbalových soutěžích (Gambrinus liga, Premier League).

V první kapitole prozkoumám některé údaje z oblasti sportu, které nevyžadují žádné náročné statistické výpočty, ale takové, které může lehce každý sportovní fanoušek provést sám.

Ve druhé kapitole s využitím statistické inference budu analyzovat výsledky dosahované v některých disciplinách na všech Olympijských hrách a také výsledky fotbalových ligových zápasů. Ze statistických metod využiji lineární regresi, metodu maximální věrohodnosti a momentovou metodu. Hlavně se zaměřím na logistickou regresi, kterou využiji při konstruování mého stěžejního experimentu.

Ve třetí kapitole blíže specifikuji pojmy z kurzového sázení a přiblížím sázkařské pojmy laikům. Kapitola je zaměřena zejména na internetové sázení, které v posledních letech zaznamenává velký rozmach.

Čtvrtá kapitola je věnována experimentu, který využívá výsledků logistické regrese pro strategii sázení a sestavení sázkového tiketu. Pokusil jsem se také simulovat, jak dopadnou tabulky české a anglické fotbalové ligy na konci ligové soutěže 2009/2010.

1 Sledování sportovních výkonů

Nejen sportovní veřejnost sleduje nejrůznější základní charakteristiky daného sportu. Ale jsou to hlavně manažeři a trenéři týmů, jejichž popisem práce je takovéto statistiky sledovat. Na jejich základě poté vybírají hráče, který bude reprezentovat jejich klub či hráče, který nejvíce zapadá do trenérovi herní koncepce. V následujících podkapitolách jsou vybrány různé charakteristiky, které se dají sledovat.

1.1 Rozložení gólů hráčů NHL

Zde se nabízí možnost sledovat náhodnou veličinu počtu gólů v jednotlivých zápasech zámořské hokejové ligy NHL u jednoho hráče. Později v sekci 2.4 na str.21 ověřím, zda počet gólů má Poissonovo rozdělení.

Vybral jsem si dva nejlepší střelce loňské sezóny a to Alexe Ovechkina z Pittsburgh Penguins a Zacha Parise z New Jersey Devils. Pro srovnání jsem vybral i českého zástupce, Jaromíra Jágra, a jeho nejpovedenější sezónu v dresu New York Rangers v roce 2005/2006. Jako posledního jsem vybral nejproduktivnějšího obránce Mika Greena z Washington Capitals.

Statistiky těchto hráčů jsou následující:

Hráč	Počet	zápasy s počtem gólů X					střední hodnota
	zápasů	0	1	2	3	4	EX
Ovechkin	80	40	27	10	3	0	
pravděpodobnost $\hat{p}_i$		0,50	0,34	0,13	0,04	0	0,72
Parise	82	42	35	5	0	0	
pravděpodobnost $\hat{p}_i$		0,51	0,43	0,06	0	0	0,55
Jágr	82	42	29	8	3	0	
pravděpodobnost $\hat{p}_i$		0,51	0,35	0,10	0,04	0	0,67
Green	68	42	21	5	0	0	
pravděpodobnost $\hat{p}_i$		0,62	0,31	0,07	0	0	0,45

Hodnotu EX jsem vypočítal jako střední hodnotu diskrétního rozdělení, tedy $EX = \sum_{i=0}^{n} X_i p_i$. Můžeme vidět, že střední hodnota gólů v zápasech se u všech hráčů liší.

1.2 Úspěšnost brankářů v NHL

Prozkoumáme, jak se liší úspěšnost dvou vybraných brankářů v jednotlivých letech a ověříme, zda jejich úspěšnost má normální rozdělení. Vybral jsem si našeho brankáře Tomáše Vokouna a gólmanskou legendu Martina Brodeura.

Tomáš Vokoun má procentuální úspěšnost zákroků v období 1998–2010

tým	úspěšnost v %
Predators	0,908
Predators	0,904
Predators	0,910
Predators	0,903
Predators	0,918
Predators	0,909
Predators	0,919
Predators	0,920
Panthers	0,919
Panthers	0,926
Panthers	0,927

Pro výpočet odhadu střední hodnoty jsem použil výběrový průměr $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0{,}915$.

Pro jednoduchost budeme uvažovat, že brankáři odehráli v každé sezóně stejný počet zápasů. Jinak bychom museli použít vážený průměr.

A pro výpočet odhadu rozptylu jsem použil výběrový rozptyl $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = 0{,}0003$.

Martin Brodeur má procentuální úspěšnost zákroků v období 1998–2010

tým	úspěšnost v %
Devils	0,906
Devils	0,910
Devils	0,910
Devils	0,906
Devils	0,906
Devils	0,914
Devils	0,917
Devils	0,911
Devils	0,922
Devils	0,920
Devils	0,916

Pro výpočet parametru střední hodnoty jsem opět použil výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = 0{,}913$.

A pro výpočet parametru rozptylu jsem také použil výběrový rozptyl $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = 0{,}00003$.

Můžeme vidět, že náš brankář je v průměru lepším brankářem, ale to hlavně díky posledním 2 sezónám, ve kterých chytal výborně. V poslední sezóně u Brodeura může mít na zhoršení procentuální úspěšnosti také vliv to, že patří mezi nejstarší hráče NHL vůbec.

1.3 Červené a žluté karty ve fotbale

Najdeme odhad střední hodnoty počtu žlutých a červených karet v jednotlivých kolech Gambrinus fotbalové ligy. Poté spočítáme jaká je pravděpodobnost, že průměrně trestaný hráč dostane žlutou či červenou kartu. Pro zajímavost to porovnáme s pravděpodobností vyloučení nejtvrdšího obránce letošní fotbalové ligy Davida Limberského z týmu Viktoria Plzeň.

Ligové kolo	ŽK	ČK
1.	42	2
2.	29	2
3.	35	1
4.	40	3
5.	40	0
6.	37	1
7.	27	2
8.	32	2
9.	31	0
10.	39	0
11.	27	2
12.	34	2
13.	17	0
14.	40	3
15.	32	3
16.	26	3
17.	30	0
18.	39	1
19.	32	1
20.	31	1

21.	26	0
22.	35	0
23.	42	1
24.	35	0

Počet žlutých karet na jedno ligové kolo je tedy 33,25, na jeden zápas jsou to potom 4 karty. Pravděpodobnost, že průměrně trestaný hráč dostane žlutou kartu je při uvážení, že do hry zasáhne 22 hráčů 19 %.

Počet červených karet na ligové kolo je 1,25, na jeden zápas 0,156 karet. Pravděpodobnost, že bude hráč vyloučen je při 22 hráčích 0,7 %.

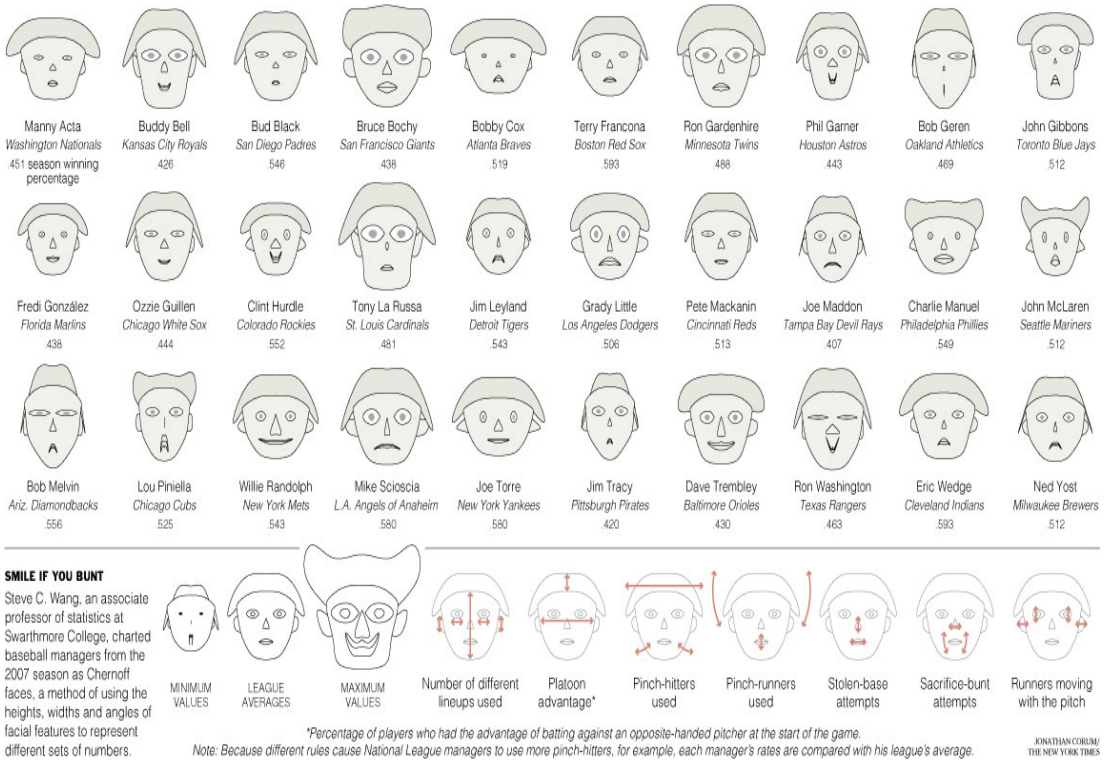
Ve skutečnosti David Limberský obdržel žlutou kartu v 10 z 21 zápasů, tedy s pravděpodobností 48 % a byl v 1 z 21 zápasů, tedy s pravděpodobností 5 % vyloučen.

1.4 Nejhorší sportovci

Většinou se sportovní fanoušci zajímají o nejlepší výkony. Přesto se stává, že předmětem zájmu se stává např. nejhorší skokan na lyžích nebo nejhorší tenista. V poslední době se několik novin omluvilo tenistovi Robertu Dee před hrozbou soudních sporů za to, že ho označili za nejhoršího tenistu světa. Důvodem byla skutečnost, že prohrál 54 zápasů v řadě po sobě, viz [12].

1.5 Trenéři

Nejen trenér může sledovat statistiky hráčů, ale i my můžeme sledovat statistiky trenérů jak často točí sestavou, jestli preferují útok či obranu atd. Tyto statistiky trenérů sledoval profesor Steve C.Wang. Porovnával manažery basebalových týmů v sezóně 2007 a toto porovnání zakreslil do tzv. Chernofových obličejů, kde šířka, výška a úhel parametrů obličeje reprezentuje různá čísla u trenérských statistik. Článek poté zveřejnil v deníku New York Times, viz [10].



Obrázek 1: Z článku zveřejněného v deníku New York Times.

2 Statistické metody

2.1 Regrese

V oblasti sportu můžeme hledat závislosti například vlivu počtu domácích fanoušků při utkání na konečný výsledek zápasu, vlivu ceny všech hráčů v týmu na dosahované výsledky nebo vlivu volby typu povrchu na výsledek tenisového utkání.

Uvažujme nejjednodušší případ, kdy závisle proměnná (tzv. vysvětlovaná) je určena nezávisle proměnnou (tzv. vysvětlující). Vysvětlující proměnnou jsou v našich případech počet fanoušků, cena hráčů a volby typu povrchu. Závisle proměnnou můžeme pokládat za náhodnou veličinu Y, která má při dané hodnotě (nenáhodné) vysvětlující veličiny x určité rozdělení pravděpodobnosti. Předpokládejme, že střední hodnota veličiny Y při dané hodnotě x, což značíme $E(Y/x)$, je rovna hodnotě známé funkce g v bodě x,

$$E(Y|x) = g(x; \beta_0, \beta_1, \ldots, \beta_k), \tag{1}$$

kde β_0, $\beta_1, \ldots, \beta_k, k \geq 1$ jsou neznámé konstanty, na kterých funkce g závisí. Funkce g se nazývá regresní funkce, konstanty β_0, $\beta_1, \ldots, \beta_k$ se nazývají regresní parametry. Regresí tedy rozumíme závislost mezi střední hodnotou náhodné veličiny Y a proměnnou x jsou neznámé konstanty.

2.1.1 Přímková regrese

V našich příkladech jsem použil tzv. ***přímkovou (lineární) regresi***, kdy pro náhodné veličiny $Y_1, \ldots, Y_n$ a čísla $x_1, \ldots, x_n$ platí

$$E(Y_i|x_i) = g(x_i) = \beta_0 + \beta_1 x_i, \ i = 1, \ldots, n. \tag{2}$$

tj.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ i = 1, \ldots, n, n \geq 2, \tag{3}$$

kde $x_1, \ldots, x_n$ jsou známá reálná čísla taková, že aspoň dvě z nich jsou různá a $\varepsilon_1, \ldots, \varepsilon_n$ jsou náhodné veličiny (náhodné odchylky, chyby měření).

2.1.2 Metoda nejmenších čtverců

Odhady $\hat{\beta}_0, \hat{\beta}_1$ neznámých parametrů β_0, β_1 určíme tzv. *metodou nejmenších čtverců*. V ní požadujeme, aby součet čtverců odchylek pozorovaných hodnot Y_i a odhadnutých hodnot $\hat{\beta}_0 + \hat{\beta}_1 x_i$ byl minimální, což je matematicky přesně popsáno v následující definici.

Definice 2.1. *Náhodné veličiny $\hat{\beta}_0, \hat{\beta}_1$, které pro daná $Y_1 \ldots, Y_n$ minimalizují výraz*

$$S_e(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2, \tag{4}$$

nazýváme odhady parametrů β_0, β_1 určené metodou nejmenších čtverců.

Věta 2.1. *V modelu lineární regrese (3) minimalizující kritérium (4) platí*

$$\hat{\beta}_0 = \frac{(\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} {x_i}^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i Y_i)}{n \sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} {x_i}^2)}, \tag{5}$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i Y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} Y_i)}{n \sum_{i=1}^{n} x^2 - (\sum_{i=1}^{n} {x_i}^2)}, \tag{6}$$

Reziduální součet lze určit pomocí vztahu

$$S_e(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} {Y_i}^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i Y_i, \tag{7}$$

Odhad disperze veličin ε_i, $i = 1, \ldots, n$ je dán vzorcem

$$S^2 = \frac{S_e}{n-2}. \tag{8}$$

Důkaz: viz [3], str. 157.

2.1.3 Pás spolehlivosti

V regresní analýze pracujeme s tzv. *pásem spolehlivosti*. Ten získáme, pokud budeme dosazovat do krajních bodů oboustranného intervalového odhadu hodnoty parametrické funkce $\beta_0 + \beta_1 x_i$ různé hodnoty x ležící v uzavřeném intervalu

$\langle \min x_i, \max x_i \rangle$. Tento pás, jak můžeme vidět na obrázcích, má nejmenší šířku pro $x = \frac{1}{n}\sum_{i=1}^{n} x_i$, vzdaluje-li se x od výběrového průměru, šířka pásu roste.

Při aplikacích se mnohdy také zajímáme o hodnotu $\beta_0 + \beta_1 x$, kde x je nějaké dané číslo, $x \in \langle \min x_i, \max x_i \rangle$.

Lze ukázat, že

$$P(T_d \leq \beta_0 + \beta_1 x \leq T_h) = 1 - \alpha,$$

kde

$$T_d = \hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}}, \tag{9}$$

$$T_h = \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}}. \tag{10}$$

Hodnota $t_{n-2,1-\frac{\alpha}{2}}$ je $(1 - \frac{\alpha}{2})$ kvantil studentova rozdělení.

Oboustranný intervalový odhad hodnoty parametrické funkce $\beta_0 + \beta_1 x$ pro dané x tvoří uspořádaná dvojice statistik (T_d, T_h).

Dosazujeme-li do T_d, T_h různé hodnoty $x \in \langle \min x_i, \max x_i \rangle$, dostaneme při spojitě se měnícím x tzv. *pás spolehlivosti kolem regresní přímky.* Tento pás má nejmenší šířku pro $x = \overline{x}$, vzdaluje-li se x od $\overline{x}$, šířka pásu roste.

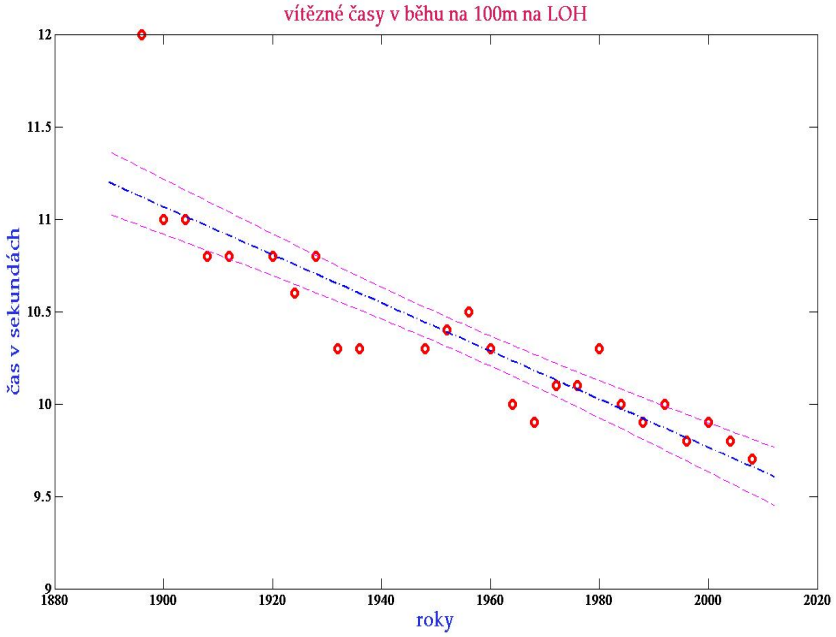
Příklad 2.1. *Jako příklad na lineární regresi jsem si vybral rozbor výsledků Letních olympijských her od počátku až do současnosti a to ve 3 disciplínách: běh mužů na 100m, běh žen na 100m a skok do výšky žen.* [1] *Sledoval jsem časy běhů a skoky, se kterými atleti vyhráli konkrétní olympijské hry. Pro tato nashromážděná data jsem sestrojil regresní přímku. Pomocí ní můžeme predikovat výsledky vítězů na příštích LOH 2012 v Londýně.*

Tabulka vítězných časů na LOH v běhu mužů na 100m

LOH	*místo*	*jméno vítěze*	*národnost*	*vítězný čas v s*
1986	*Athény*	*Tom Burke*	*USA*	*12,00*
1900	*Paříž*	*Frank Jarvis*	*USA*	*11,00*
1904	*St. Louis*	*Archie Hahn*	*USA*	*11,00*

[1] Poslední 2 disciplíny jsou až od LOH 1928 v Amsterodamu, od kterých ženy měly povoleno startovat na olympijských hrách.

1908	*Londýn*	*Reggie Walker*	*SAF*	*10,80*
1912	*Stockholm*	*Ralph Craig*	*USA*	*10,80*
1920	*Antverpy*	*Charles Paddock*	*USA*	*10,80*
1924	*Paříž*	*Harold Abrahams*	*GBR*	*10,60*
1928	*Amsterdam*	*Percy Williams*	*CAN*	*10,80*
1932	*Los Angeles*	*Eddie Tolan*	*USA*	*10,30*
1936	*Berlín*	*Jesse Owens*	*USA*	*10,30*
1948	*Londýn*	*Harrison Dillard*	*USA*	*10,30*
1952	*Helsinki*	*Lindy Remigino*	*USA*	*10,40*
1956	*Melbourne*	*Bobby Morrow*	*USA*	*10,50*
1960	*Řím*	*Armin Hary*	*GER*	*10,30*
1964	*Tokyo*	*Bob Hayes*	*USA*	*10,00*
1968	*Mexico City*	*Jim Hines*	*USA*	*9,95*
1972	*Mnichov*	*Valery Borzov*	*URS*	*10,14*
1976	*Montreal*	*Hasely Crawford*	*TRI*	*10,06*
1980	*Moskva*	*Allan Wells*	*GBR*	*10,25*
1984	*Los Angeles*	*Carl Lewis*	*USA*	*9,99*
1988	*Soul*	*Carl Lewis*	*USA*	*9,92*
1992	*Barcelona*	*Linford Christie*	*GBR*	*9,96*
1996	*Atlanta*	*Donovan Bailey*	*CAN*	*9,84*
2000	*Sydney*	*Maurice Greene*	*USA*	*9,87*
2004	*Athény*	*Justin Gatlin*	*USA*	*9,85*
2008	*Peking*	*Usain Bolt*	*JAM*	*9,69*



Obrázek 2: Výsledky v běhu mužů na 100 m.

Podle použité regrese vyšlo, že odhadovaný vítězný čas na LOH v roce 2012 v běhu na 100m mužů by měl být v intervalu 9,45 s až 9,77 s s očekávanou hodnotou 9,61 s.

Odhadnutá regresní přímka má tvar $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$. *Odhad očekávaného vítězného času jsem získal dosazením* $x = 2012$ *do tohoto vztahu.*

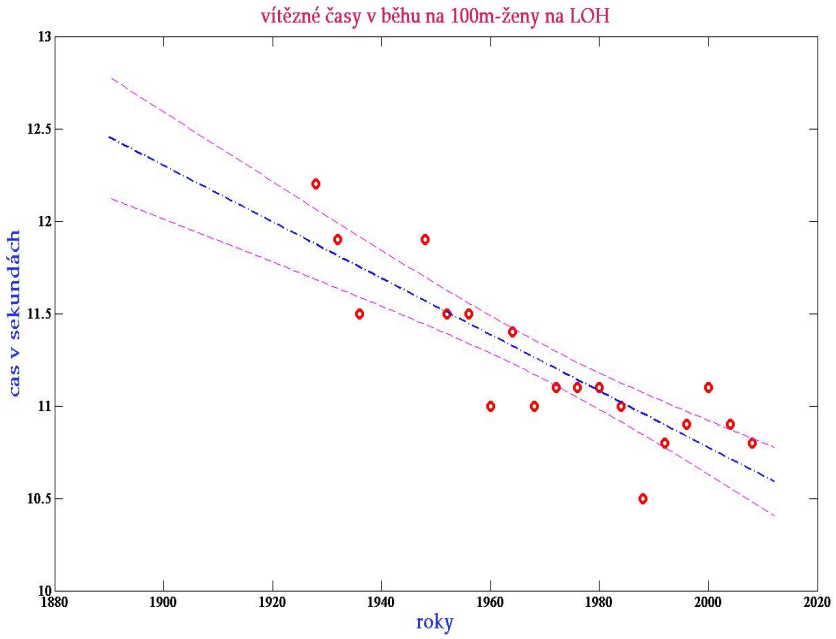
Obdobně po dosazení $x = 2012$ *do vztahů (9) a (10) jsem dostal interval spolehlivosti pro mou předpověď:*

$$T_d = \hat{\beta}_0 + \hat{\beta}_1 x - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}} = 9{,}45$$

$$T_h = \hat{\beta}_0 + \hat{\beta}_1 x + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}} = 9{,}77$$

Tabulka vítězných časů na LOH v běhu žen na 100m

LOH	*místo*	*jméno vítěze*	*národnost*	*vítězný čas v s*
1928	*Amsterdam*	*Elizabeth Robinson*	*USA*	*12,20*
1932	*Los Angeles*	*Stanislawa Walasiewicz*	*POL*	*11,90*
1936	*Berlín*	*Helen Stephens*	*USA*	*11,50*
1948	*Londýn*	*Fanny Blankers-Koen*	*NED*	*11,90*
1952	*Helsinki*	*Marjorie Jackson*	*AUS*	*11,50*
1956	*Melbourne*	*Betty Cuthbert*	*AUS*	*11,50*
1960	*Řím*	*Wilma Rudolph*	*USA*	*11,00*
1964	*Tokyo*	*Wyomia Tyus*	*USA*	*11,40*
1968	*Mexico City*	*Wyomia Tyus*	*USA*	*11,00*
1972	*Mnichov*	*Renate Stecher*	*GDR*	*11,07*
1976	*Montreal*	*Annegret Richter*	*FRG*	*11,08*
1980	*Moskva*	*Lyudmila Kondratyeva*	*URS*	*11,06*
1984	*Los Angeles*	*Evelyn Ashford*	*USA*	*10,97*
1988	*Soul*	*Florence Griffith-Joyner*	*USA*	*10,54*
1992	*Barcelona*	*Gail Devers*	*USA*	*10,82*
1996	*Atlanta*	*Gail Devers*	*USA*	*10,94*
2000	*Sydney*	*Ekaterini Thanou*	*GRE*	*11,12*
2004	*Athény*	*Yuliya Nesterenko*	*BLR*	*10,93*
2008	*Peking*	*Shelly-Ann Fraser*	*JAM*	*10,78*



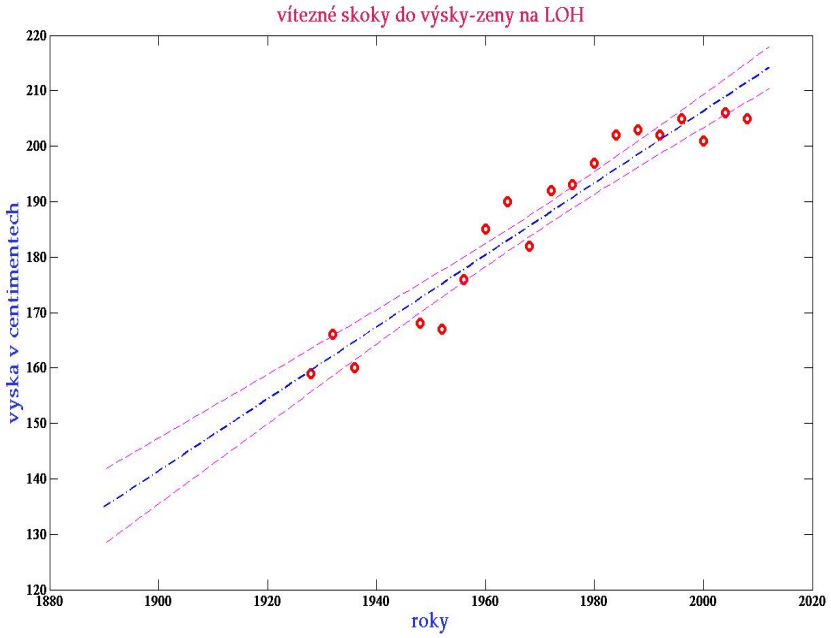
Obrázek 3: Výsledky v běhu žen na 100 m.

Podle použité regrese vyšlo, že odhadovaný vítězný čas na LOH v roce 2012 v běhu na 100m žen by měl být v intervalu 10,41 s až 10,78 s s očekávanou hodnotou 10,60 s.

Tabulka vítězných skoků na LOH ve skoku do výšky žen

LOH	*místo*	*jméno vítěze*	*národnost*	*vítězný skok v m*
1928	*Amsterdam*	*Ethel Catherwood*	*CAN*	*1,59*
1932	*Los Angeles*	*Jean Shiley*	*USA*	*1,66*
1936	*Berlín*	*Ibolya Csák*	*HUN*	*1,60*
1948	*Londýn*	*Alice Coachman*	*USA*	*1,68*
1952	*Helsinki*	*Esther Brand*	*SAF*	*1,67*
1956	*Melbourne*	*Millie McDaniel*	*USA*	*1,76*
1960	*Řím*	*Iolanda Balas*	*ROM*	*1,85*
1964	*Tokyo*	*Iolanda Balas*	*ROM*	*1,90*
1968	*Mexico City*	*Miloslava Rezková*	*CZE*	*1,82*
1972	*Mnichov*	*Ulrike Meyfarth*	*FRG*	*1,92*
1976	*Montreal*	*Rosemarie Ackermann*	*GDR*	*1.,93*
1980	*Moskva*	*Sara Simeoni*	*ITA*	*1,97*
1984	*Los Angeles*	*Ulrike Meyfarth*	*FRG*	*2,02*
1988	*Soul*	*Louise Ritter*	*USA*	*2,03*
1992	*Barcelona*	*Heike Henkel*	*GER*	*2,02*
1996	*Atlanta*	*Stefka Kostadinova*	*BUL*	*2,05*

2000	*Sydney*	*Yelena Yelesina*	*RUS*	*2,01*
2004	*Athény*	*Yelena Slesarenko*	*RUS*	*2,06*
2008	*Peking*	*Tia Hellebaut*	*BEL*	*2,05*



Obrázek 4: Výsledky ve skoku do výšky žen.

Podle použité regrese vyšlo, že odhadovaný vítězný skok na LOH v roce 2012 ve skoku do výšky žen by měl být v intervalu 2,10 m až 2,18 m s očekávanou hodnotou 2,14 m.

2.2 Momentová metoda

Následující metoda je nejstarší metodou odhadu parametrů, je poměrně jednoduchá a používá se hlavně v případech, kdy jiné metody odhadu jsou numericky nebo z jiných důvodů těžko zvládnutelné. Odhady získané touto metodou se někdy používají jako počáteční aproximace. Na druhé straně je zpravidla použitelná jen tehdy, když jsou výchozí náhodné veličiny nezávislé a stejně rozdělené. Pokud se však jedná a rozdělení, které nemají konečné momenty, tak se tato metoda nedá aplikovat.

Nechť zkoumaná náhodná veličina X má rozdělení závislé na parametru

$\Theta = (\theta_1, \ldots, \theta_k)$. Předpokládejme, že existují všeobecné momenty

$$\mu_k^{'} = E(X_i^k),$$

kde $k = 1, \ldots, m$.

Je zřejmé, že tyto momenty $\mu_1^{'}, \ldots, \mu_k^{'}$ jsou funkcemi parametru θ a tedy

$$\mu_k^{'} = \mu_k^{'}(\theta_1, \ldots, \theta_k),$$

kde $k = 1, \ldots, m$.

Označme $M_1^{'}, \ldots, M_k^{'}$ všeobecné výběrové momenty

$$M_k^{'} = \frac{1}{n} \sum_{i=1}^{n} (X_i)^k, \ k = 1, \ldots, m.$$

Neznámý parametr $\Theta = (\theta_1, \ldots, \theta_k)$ získáme řešením soustavy rovnic

$$\mu_k^{'}(\theta) = M_k^{'}(\theta), \tag{11}$$

kde $k = 1, \ldots, m$.

Může se stát, že vyřešení rovnice nepovede k jednoznačnému určení neznámého parametru. V takovém případě se většinou připojují další rovnice pro $k = m + 1, \ldots$ až se získá potřebný počet rovnic.

Jestliže při libovolných hodnotách $M_1^{'}, \ldots, M_k^{'}$ má soustava

$$\mu_k^{'}(\theta) = M_k^{'}(\theta),$$

jediné řešení, dává metoda momentů jednoznačně určené odhady parametru $\Theta = (\theta_1, \ldots, \theta_k)$.

Poznámka 2.1. *Namísto všeobecných momentů je možné porovnat centrální momenty zkoumané náhodné veličiny s výběrovými centrálními momenty.*

2.2.1 Odhad parametru Poissonova rozdělení

Nechť $X \sim \mathrm{Po}(\lambda)$. Víme, že pro Poissonovo rozdělení platí

$$p(x, \lambda) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \ E(X) = \lambda$$

Řešíme jednu rovnici

$$\mu_1^{'} = M_1^{'},$$

kde střední hodnota je

$$E(X) = \mu_1^{'} = \lambda$$

a

$$M_1^{'} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$$

t.j. odhad parametru λ z Poissonova rodělení určený momentovou metodou je

$$\hat{\lambda} = \overline{X},$$

tedy stejný jako v metodě maximální věrohodnosti.

Poznámka 2.2. *Ne vždy dostaneme explicitní vyjádření momentových odhadů (hlavně při soustavách rovnic). V takovém případě se řešení hledá numericky.*

2.3 Metoda maximální věrohodnosti

Nejpoužívanější metodou, která je ve všeobecnosti považována za „nejsprávnější“, je metoda maximální věrohodnosti. Je založená na tom, že odhady neznámých parametrů uvažované náhodné veličiny se vyberou tak, aby hodnoty hustoty v bodech náhodného výběru byli maximální. Tuto metodu můžeme používat ve velmi rozmanitých situacích a odhady získané tímto způsobem mají velmi dobré vlastnosti (např. odhad je asymptoticky nezkreslený).

Nechť $\mathbf{X} = (X_1, \ldots, X_n)$ je náhodný výběr z diskrétního rozdělení s pravděpodobnostní funkcí $p(\mathbf{x}, \boldsymbol{\Theta})$, resp. ze spojitého rozdělení s hustotou $f(\mathbf{x}, \boldsymbol{\Theta})$, kde

$$\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_n)' \in \underline{\Theta}$$

je vektor neznámých parametrů.

Omezme se na jednorozměrný parametr. Pro každé pevné $\mathbf{x}$ se však dá $p(\mathbf{x}, \boldsymbol{\Theta})$, resp. $f(\mathbf{x}, \boldsymbol{\Theta})$, chápat i jako funkce proměnné $\boldsymbol{\Theta}$. Pro tuto funkci budeme používat označení $L(\mathbf{x}, \boldsymbol{\Theta})$ (z angl. likelihood - věrohodnost) a budeme ji nazývat ***věrohodnostní funkce***. Pro libovolnou dvojici $(\mathbf{x}, \boldsymbol{\Theta})$ samozdřejmě platí

$$L(\boldsymbol{\Theta}) = \prod p(x_i, \boldsymbol{\Theta}),\ resp.\ L(\boldsymbol{\Theta}) = \prod f(x_i, \boldsymbol{\Theta}). \tag{12}$$

Jde jen o to, že použitím symbolu L poukazujeme na tuto funkci jako na funkci proměnné θ při pevně daném $\mathbf{x}$.

Jestliže existuje takový bod $\hat{\boldsymbol{\Theta}} \in \Omega$, že pro všechny $\boldsymbol{\Theta} \in \Omega$ platí,

$$L(\mathbf{x}, \boldsymbol{\Theta}) \leq L(\mathbf{X}, \boldsymbol{\Theta}),$$

potom hovoříme, že $\hat{\boldsymbol{\Theta}}$ je *odhad parametru* $\boldsymbol{\Theta}$ *získaný metodou maximální věrohodnosti*.

Často je výhodnější maximalizovat místo funkce $L(\boldsymbol{\Theta})$ její logaritmus $\ln L(\boldsymbol{\Theta})$ (*logaritmická funkce věrohodnosti*). Když si uvědomíme, že logaritmus součinu je rovný součtu logaritmů, maximálně věrohodný odhad $\boldsymbol{\Theta}$ se zpravidla (jak má funkce $L(\boldsymbol{\Theta})$ parciální derivaci) stanoví řešením rovnice

$$\frac{\partial \ln L(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = 0,$$

které říkáme *věrohodnostní rovnice*.

2.3.1 Odhad parametru Poissonova rozdělení

Nechť $X \sim \mathrm{Po}(\lambda)$. Věrohodnostní funkce je

$$L(\lambda) = \prod_{i=1}^{n} P(X = x_i) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} e^{-n\lambda}.$$

Teda

$$\ln L(\lambda) = \left(\sum_{i=1}^{n} x_i \right) \ln \lambda - n\lambda - \ln \prod_{i=1}^{n} x_i!.$$

Věrohodnostní rovnice je

$$\frac{\partial \ln L(\lambda)}{\partial \lambda} = \left(\sum_{i=1}^{n} x_i \right) \frac{1}{\lambda} - n = 0.$$

jejímž řešením je

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}.$$

Maximálně věrohodným odhadem parametru λ Poissonova rozdělení je tedy $\hat{\lambda} = \overline{X}$.

viz příklad 1.1

Najdeme odhad parametru λ Poissonova rozdělení pro počty gólů v jednotlivých zápasech sezóny. Tento odhad je podle momentové metody i podle maximální věrohodnosti roven $\overline{X}$. Ovechkin má $\hat{\lambda} = 0{,}7$, Parise má $\hat{\lambda} = 0{,}55$, Jágr má $\hat{\lambda} = 0{,}66$ a Green má $\hat{\lambda} = 0{,}46$, viz tabulka strana 6.

2.4 Testy dobré shody

Věta 2.2. *Nechť náhodný vektor* $\mathbf{X}$ *má multinomické rozdělení s parametry* $n, p_1, \ldots, p_k$. *Potom náhodná veličina*

$$\sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j} \tag{13}$$

má při $n \to \infty$ *asymptoticky rozdělení* χ^2_{k-1}.

Důkaz: viz [2], str. 270.

Ve vztahu z předchozí věty tak aproximujeme diskrétní rozdělení statistiky rozdělením spojitým. Tato aproximace se považuje za použitelnou, když je dostatečně velký rozsah n a je-li pro každé $j = 1, 2, \ldots, k$ splněna nerovnost $np_j \geq 5$.

Věta 2.2 je základem řady testů, v nichž konfrontujeme *empiricé četnosti*, tj. realizace náhodných veličin $X_1, \ldots, X_k$, se středními hodnotami těchto náhodných veličin (tzv. *očekávené* neboli *teoretické četnosti*) $np_1^0, \ldots, np_k^0$, v nichž jsou

pravděpodbonosti $p_1^0, \ldots, p_k^0$ určeny z platnosti nějakého pravděpodobnostního modelu. Nulová hypotéza pak tvrdí, že pravděpodobnosti $p_1, \ldots, p_k$ v modelu multinomického rozdělení jsou rovny $p_1^0, \ldots, p_k^0$ (známe tak rozdělení náhodného vektoru $\mathbf{X}$), jinak řečeno

$$H_0 : p_1 = p_1^0, \ldots, p_k = p_k^0.$$

Oproti alternativě H_1, že alespoň jedna z rovností neplatí.

Jako testové kritérium použijeme Větu 2.2, tj. za platnosti H_0 má statistika

$$Z = \sum_{j=1}^{k} \frac{(X_j - np_j^0)^2}{np_j^0} = \sum_{j=1}^{k} \frac{X_j^2}{np_j^0} - n$$

asymptoticky (pro velká n) rozdělení χ^2 o $k-1$ stupních volnosti. Nulovou hypotézu pak na hladině významnosti α zamítáme v případě, že $Z \geq \chi^2_{k-1}(1-\alpha)$. Z tvaru testové statistiky je zřejmé, že neshodě skutečnosti s hypotézou odpovídají právě velké hodnoty této statistiky.

Stejně jako v případě Věty 2.2, i zde pažadujeme splnění předpokladu $np_1^0 \geq 5$, $\forall j = 1, \ldots, k$. Tento lze ve většině praktických situací oslabit tzv. Yaroldovým kritériem, které požaduje pro všechna j splnění nerovnosti $np_1^0 \geq 5q$, kde q je podíl těchto skupin (tříd), ve kterých $np_1^0 < 5$.

Příklad 2.2. *Z tabulky na straně 6 jsem udělal tabulky uvedené níže, když teoretické pravděpodobnosti byly určeny pro $\widehat{\lambda} = \overline{X}$. Poté jsem vypočítal u každého hráče hodnotu testového kriteria χ^2. Tu jsem porovnal s kritickou hodnotou χ^2 o 3 stupních volnosti při hladině testu $\alpha = 5$ %, která se rovná 7.8147. U všech hráčů vyšla hodnota χ^2 menší než je kritická hodnota. U žádného hráče nemůžeme zamítnout nulovou hypotézu o tom, že rozložení počtu gólů má Poissonovo rozdělení.*

Tabulka četností gólů Ovechkina

počet gólů	empirická relativní četnost	teoretická četnost	hladina χ^2
0	0,5	0,5	0,72697
1	0,34	0,35	
2	0,13	0,12	
3	0,04	0,03	
4	0,00	0,00	

Tabulka četností gólů Parise

počet gólů	empirická relativní četnost	teoretická četnost	hladina χ^2
0	0,51	0,58	1,0582
1	0,43	0,32	
2	0,06	0,09	
3	0,00	0,02	
4	0,00	0,00	

Tabulka četností gólů Jágra

počet gólů	empirická relativní četnost	teoretická četnost	hladina χ^2
0	0,51	0,52	5,8712
1	0,35	0,34	
2	0,10	0,11	
3	0,04	0,03	
4	0,00	0,00	

Tabulka četností gólů Greena

počet gólů	empirická relativní četnost	teoretická četnost	hladina χ^2
0	0,62	0,63	0,94741
1	0,31	0,29	
2	0,07	0,07	
3	0,00	0,01	
4	0,00	0,00	

2.5 Logistická regrese

Modelování vztahů mezi vysvětlující a vysvětlovanou proměnnou patří k častým problémům, se kterými se ve statistice můžeme setkat. Obvykle předpokládáme, že je závisle proměnná náhodnou veličinou s normálním rozdělením. Pro odvození modelu pak používáme metodu nejmenších čtverců.

Problém však může nastat tehdy, není-li závisle proměnná statistickým znakem, ale znakem binárním. V takovém případě, již nelze použít k odhadu parametrů „klasickou“ regresní analýzu s odhadem regresních koeficientů prostřednictvím metody nejmenších čtverců. K odhadu těchto parametrů β_0 a β_1 používáme metodu maximální věrohodnosti.

Označme si pravděpodobnosti: $P(Y_i = 1) = \pi_i$, $P(Y_i = 0) = 1 - \pi_i$. Je zřejmé, že $Y_i \sim Bi(1, \pi_i)$. V případě nezávislosti jednotlivých pozorování můžeme věrohodnost zapsat jako součin pravděpodobností:

$$L(\beta_0, \beta_1, x, y) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \tag{14}$$

Použijme k modelování pravděpodobnosti již výše zmiňovanou logistickou funkci, tj.

$$\pi_i = \frac{e^{(\beta_0+\beta_1 x_i)}}{1 + e^{(\beta_0+\beta_1 x_i)}}. \tag{15}$$

pak tedy

$$1 - \pi_i = \frac{1}{1 + e^{(\beta_0+\beta_1 x_i)}}.$$

Po dosazení získáme tedy věrohodnostní funkci ve tvaru:

$$L(\beta_0, \beta_1, x, y) = \prod_{i=1}^{n} \left(\frac{e^{(\beta_0+\beta_1 x_i)}}{1 + e^{(\beta_0+\beta_1 x_i)}} \right)^{y_i} \left(\frac{1}{1 + e^{(\beta_0+\beta_1 x_i)}} \right)^{1-y_i},$$

což lze zjednodušit na

$$L(\beta_0, \beta_1, x, y) = \prod_{i=1}^{n} \frac{(e^{(\beta_0+\beta_1 x_i)})^{y_i}}{1 + e^{(\beta_0+\beta_1 x_i)}}.$$

Maximalizovat přímo věrohodnostní funkci $L(\beta_0, \beta_1, x, y)$ by nebylo příliš vhodné. Je lepší věrohodnostní funkci zlogaritmovat. Výpočet se pak znatelně zjednoduší. Po logaritmování získáme věrohodnostní funkce:

$$\ln L = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) + \sum_{i=1}^{n} \ln\left(1 + e^{(\beta_0+\beta_1 x_i)}\right).$$

Naše parametry najdeme pomocí tzv. Newtonovy metody

$$\Theta^{(k+1)} = \Theta^{(k)} - (\mathbf{H}(\ln L(\Theta^{(k)})))^{-1} \cdot \nabla(\ln L(\Theta^{(k)})), \tag{16}$$

kde $\mathbf{H}$ je Hessova matice a ∇ gradient.

Určeme si nyní první parcialní derivace, které budeme potřebovat pro výpočet gradientu:

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{i=1}^{n} \left(\frac{y_i - e^{\beta_0+\beta_1 x_i} + e^{\beta_0+\beta_1 x_i} y_i}{1 + e^{\beta_0+\beta_1 x_i}} \right) = \sum_{i=1}^{n} \left(y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \right),$$

$$\frac{\partial \ln L}{\partial \beta_1} = \sum_{i=1}^{n} \left(\frac{\left(y_i - e^{\beta_0+\beta_1 x_i} + e^{\beta_0+\beta_1 x_i} y_i\right) x_i}{1 + e^{\beta_0+\beta_1 x_i}} \right) = \sum_{i=1}^{n} \left(x_i y_i - \frac{x_i e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}} \right).$$

V našem případě tedy gradient $\nabla(\ln L(\Theta^{(k)}))$ bude vypadat:

$$\nabla(\ln L(\Theta^{(k)}) = \begin{bmatrix} \sum_{i=1}^{n} \left(y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) \\ \sum_{i=1}^{n} \left(x_i y_i - \frac{x_i e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \right) \end{bmatrix}.$$

Dále budeme potřebovat druhé parciální derivace pro výpočet Hessovy matice:

$$\frac{\partial \ln L}{\partial \beta_0^2} = -\frac{e^{\beta_0+\beta_1 x_i}}{(1 + e^{\beta_0+\beta_1 x_i})^2},$$

$$\frac{\partial \ln L}{\partial \beta_0 \partial \beta_1} = -\frac{x_i e^{\beta_0+\beta_1 x_i}}{(1 + e^{\beta_0+\beta_1 x_i})^2},$$

$$\frac{\partial \ln L}{\partial \beta_1 \partial \beta_0} = -\frac{x_i e^{\beta_0+\beta_1 x_i}}{(1 + e^{\beta_0+\beta_1 x_i})^2},$$

$$\frac{\partial \ln L}{\partial \beta_1^2} = -\frac{e^{\beta_0+\beta_1 x_i} {x_i}^2}{(1 + e^{\beta_0+\beta_1 x_i})^2}.$$

Hessova matice $\mathbf{H}(\ln L(\Theta^{(k)}))$ tedy bude vypadat:

$$\mathbf{H}(\ln L(\Theta^{(k)})) = \begin{bmatrix} \sum_{i=1}^{n} \left(-\frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \right) & \sum_{i=1}^{n} \left(-\frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \right) \\ \sum_{i=1}^{n} \left(-\frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \right) & \sum_{i=1}^{n} \left(-\frac{e^{\beta_0+\beta_1 x_i} {x_i}^2}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \right) \end{bmatrix}.$$

Pro náš konkrétní případ využití Newtonovy metody bude vypadat následovně:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^0 - (\mathbf{H}(\ln L(\Theta^{(k)})))^{-1} \cdot \nabla(\ln L(\Theta^{(k)})) \tag{17}$$

kde matice $(\mathbf{H}(\ln L(\Theta^{(k)})))^{-1} =$

$$\begin{bmatrix} \sum_{i=1}^{n} -\frac{e^{\beta_0+\beta_1 x_i} {x_i}^2}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} A, & -\sum_{i=1}^{n} -\frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} A \\ -\sum_{i=1}^{n} -\frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} A, & \sum_{i=1}^{n} -\frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} A \end{bmatrix},$$

kde

$$A = \left(\sum_{i=1}^{n} -\frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \sum_{i=1}^{n} -\frac{e^{\beta_0+\beta_1 x_i} {x_i}^2}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} - \left(\sum_{i=1}^{n} -\frac{x_i e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \right)^2 \right)^{-1}.$$

Řešení dostaneme, když za vektor $\boldsymbol{\beta}^0$ nejdříve dosadíme vhodné počáteční řešení, tím dostaneme nějaké nový odhad řešení $\hat{\boldsymbol{\beta}}$. Ten použijeme jako nové počáteční řešení. Vzniklá posloupnost konverguje k hledanému řešení.

3 Kurzové sázení

Princip *kurzového sázení* spočívá v odhadování výsledku události sázkařem, které je za účelem výhry podloženo finančním vkladem — **sázkou**. Vypisovatel sázky (sázková kancelář) zvažuje *pravděpodobnost* jednotlivých možných výsledků události a podle ní stanovuje příslušné kurzy. Sázkař, který správně odhadne výsledek události, získává zpět vsazenou částku násobenou kurzem, který sázková kancelář k danému výsledku vypsala. V případě špatného odhadu si provozovatel vsazenou částku ponechává jako výdělek.

3.1 Formy kurzového sázení

Nejběžnější jsou kurzové sázky na sportovní utkání, následují politické a společenské události (vítěz voleb, volba Miss apod.). Kurzové sázení je realizováno v sázkových kancelářích nebo ve formě sázení po internetu, které v poslední době zaznamenává velkého vzrůstu. Některé sázkové kanceláře umožňují podat sázky pomocí telefonu.

3.1.1 Výhody sázení po internetu z pohledu hráče

1. Dostupnost a pohodlí

   — Hráč se ke svému hernímu účtu může připojit z jakéhokoliv počítače s dostupností internetu. Při sportovním sázení bývají na vybrané zápasy k dispozici i stovky možných sázek zohledňující všechny aspekty hry, tedy nejen obvyklý výsledek hry (počet gólů, rozdíl konečného skóre, počet faulů, počet rohů ve fotbale atd.).

2. Sázky v přímém přenosu

   — Mnoho internetových sázkových kanceláří umožňuje sázení v průběhu zápasu. Kurzy na jednotlivé události v utkání se mění v závislosti na jeho průběhu.

3. Výhodnější sázení

   — Díky tomu, že si zákazník může vybrat z více sázkových kanceláří, vzniká o něj mezi provozovateli online sázení přirozený boj. Jeho výsledkem jsou u kurzového sázení výhody ve formě konkurujících si kurzů a prakticky absence manipulačního poplatku, jak jej známe z většiny „kamenných sázkových kanceláří".

3.1.2 Právní hledisko sázení po intenetu v Evropě a ČR

V rámci Evropské unie má každý stát přidělování licencí na provozování sázkových her ve své vlastní kompetenci. Nemůže však překračovat principy Unie o podnikání poskytování služeb v rámci EU. To paradoxně znamená, že např. český zákon nemusí českým firmám umožnit získat licenci na provozování online sázkových her, ale Česká republika nemůže znemožnit zahraničním sázkovým kancelářím podnikat na svém území. Většina provozovatelů má navíc sídlo v tzv. daňových rájích (Gibraltar, Malta.), které poskytují těmto společnostem licence. Podle platných zákonů se pak sázka uskutečňuje v zemi kde má společnost sídlo. Díky tomu neporušuje zákon země, odkud sází její klient. Jednotlivé státy by tak mohly postihovat pouze vlastní občany - sázkaře.

Důležité je zabezpečení převodů financí proti praní špinavých peněz a nelegálním operacím. Jedním z bezpečnostních prvků identifikace je, že finanční převod od provozovatele k hráči musí být proveden pouze na účet nebo kreditní kartu vedenou pod jménem, které se shoduje se jménem majitele sázkového účtu. Finanční převody lze tak vždy vystopovat. Nevýhodou je, že anonymní účast není možná. Pozitivní je fakt, že výhra vždy skončí v rukou majitele.

Rozhodujícím faktorem pro účast v internetovém sázení je věk. Spodní hranice věku hráče je daná zákony země, jejíž je hráč občanem. Ta je nejčastěji stanovena na 18 nebo 21 let.

3.2 Stanovení kurzů

Kurzy na jednotlivé výsledky událostí jsou stanovovány na základě pravděpodobnosti, že nastanou. Čím více je pravděpodobné, že konkrétní výsledek nastane, tím nižší je na něj vypsaný kurz a potažmo i výhra ze sázky na tento výsledek. Analogicky, čím vyšší je kurz, tím menší je šance, že nastane výsledek, pro který byl daný kurz stanoven. Platí, že hodnota kurzu je vždy větší než jedna. Horní hranice není stanovena. Sázkové kanceláře také provádí korekci kurzů na základě aktuálně vsazených částek.

3.2.1 Skutečnosti ovlivňující výši kurzu

Výše kurzu je stanovena na základě aktuálních informací o dané události (vývoj soutěže, volební preference, stav utkání, zranění klíčových hráčů) a na základě historických předpokladů (výsledky předchozích utkání, umístění v tabulce, volební účast). Výsledná hodnota kurzu je dána mnoha faktory. Do úvahy připadají také fyzická připravenost hráčů, souhra týmu, zdravotní stav, místo utkání. Kurz se může časem vyvíjet s ohledem na nově zjištěné informace.

U sázkových kanceláří dále platí, že kurz musí být konkurenceschopný ve vztahu k nabídce jiných sázkových kanceláří. Rozdíl mezi kurzy od různých sázkových kanceláří je přímo úměrný výši kurzu. Čím vyšší je kurz, tím více se může lišit od kurzu v nabídce jiného pořadatele. Naopak kurzy na vysoce pravděpodobné výsledky budou téměř identické.

Další faktor ovlivňující kurzy je finanční výtěžnost z vypisované sázky. Cílem pořadatele je, aby vklady sázkařů, kteří sázku prohráli, přinejmenším pokryly výhry vyplácené sázkařům, kteří vyhráli. Rozdíl pak tvoří ztrátu, nebo častěji zisk.

Ve výjimečných situacích jsou kurzy úmyslně stanoveny tak, aby při libovolném výsledku pořadatel ze sázky zisk neměl. Cílem bývá nabídnout lepší kurzy než konkurenční společnosti a přetáhnout tak zákazníky. Děje se tak většinou při významných sportovních utkáních, kdy je vyšší šance oslovit nové zákazníky, kteří dříve nesázeli a vybírají si novou sázkovou společnost. Při výběru je pak

kurz sázky jeden z hlavních faktorů.

Kurzy stanovují tzv. ***bookmakeři*** – odborníci na oblast, pro kterou kurzy vypisují. Jejich úkolem je získání a vyhodnocení informací, které mají vliv na výsledek vypisované události. Jejich práce je prakticky shodná s prací profesionálních sázkařů.

3.3 Manipulační poplatek

Manipulační poplatek je částka (často v procentech), kterou si sázková kancelář může odečíst od sázkového vkladu. Některé sázkové kanceláře (online i kamenné) však poplatek požadovat nemusí. Na první pohled je zřejmé, že druhá zmíněná varianta je výhodnější. Avšak může nastat situace, kdy kurz sázkové kanceláře, která si účtuje poplatek, je o tolik vyšší, že zisk z případné výhry pokrývá „ztrátu“ způsobenou manipulačním poplatkem.

Příklad 3.1. *Příklad, kdy se vyplatí vsadit u sázkové kanceláře, která si účtuje manipulační poplatek.*

- *První sázková kancelář vypisuje sázku bez menipulačního poplatku s kurzem 1,80.*
- *Druhá sázková kancelář vypisuje stejnou sázku s kurzem 2,00 ale s manipulačním poplatkem 10 % ze vsazené částky.*
- *V prvním případě vsadíme 110 Kč a potenciální výhra bude 198 Kč (110 × 1,80), zisk 88 Kč.*
- *V druhém případě vsadíme 100+10 Kč a výhra bude 200 Kč(100 × 2,0), zisk 90 Kč. Reálný kurz (bez poplatku) je 2,00/1,1 což je cca 1,82.*
- *Porovnáním reálných kurzů* $(1{,}82 > 1{,}80)$ *dospějeme k závěru, že je lepší vsadit na kurz 2,00 i navzdory 10 % poplatku z vkladu.*

4 Experiment

Můj stěžejní experiment spočívá ve sledování výsledků fotbalových utkání české fotbalové soutěže Gamrinus ligy a anglické nejvyšší soutěže Premier League. Všechny zápasy z uplynulé sezóny jsem zanamenával do tabulky. Konečné pořadí týmů, rozdíl ve vstřelených a obdržených gólech a celkový počet bodů jsem převzal z internetové databáze ligových soutěží. Poté jsem z naší tabulky výsledků všech zápasů udělal jednodušší tabulku, ve které bylo pouze zaznamenáno, zda tým vyhrál či nevyhrál (remizoval nebo prohrál). Dále jsem si musel udělat tabulku rozdílu bodů na konci sezóny. Ke každému rozdílu bodů jsem vypočítal četnost odehranyých zápasů a každému zápasu jsem přiřadil 1 (když tým vyhrál) nebo 0 (když tým nevyhrál).

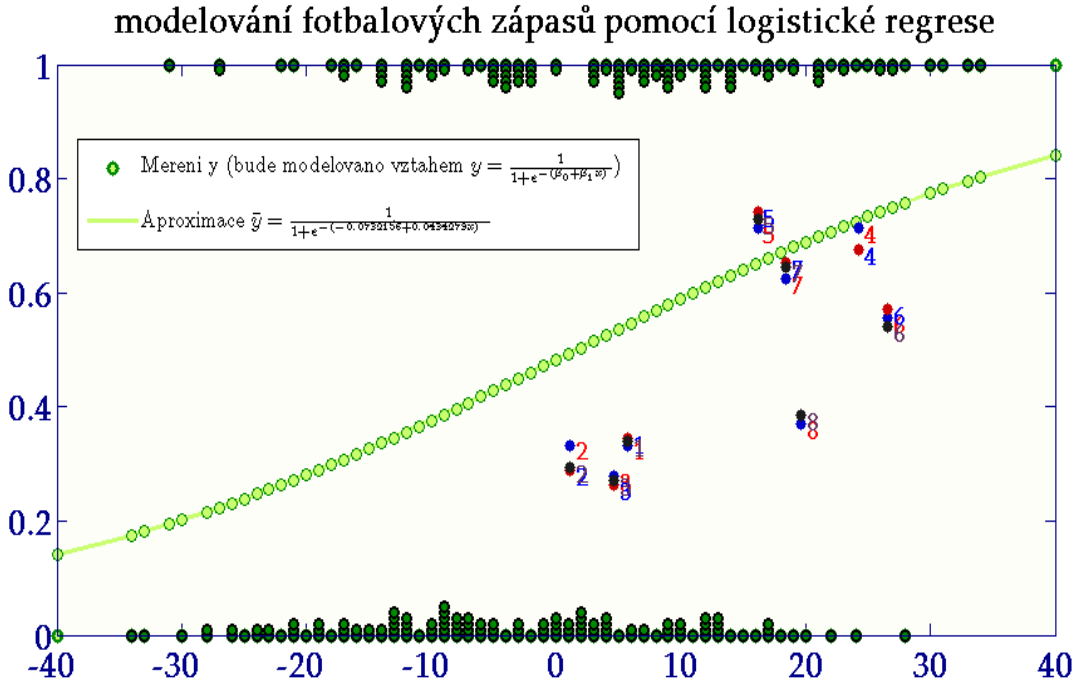
Na tyto data jsem aplikoval výše zmiňovanou logistickou regresi, která mi vykreslila křivku závislosti výhry na velikosti rozdílu počtu bodů, jak můžete vidět na obrázcích. Vše jsem programoval v matematickém softwaru Matlab.

Z této křivky jsem pak snadno určil pravděpodobnosti výhry lépe či hůře postaveného týmu v soutěžní tabulce. Tyto pravděpodobnosti jsem srovnal s pravděpodobnostmi vypočítanými z kurzů uváděných sázkovými kancelářemi. Ty se vypočítali tak, že jsme 1 podělili stanoveným kurzem. Samozřejmě každá sázková kancelář se liší ve výši nabízených kurzů na jednotlivý zápas. Pro tento příklad jsem si vybral 3 sázkové kanceláře. Z čehož jedna je normální „kamenná sázková kancelář“, která ovšem poskytuje i sázení po internetu a tou je Tipsport a.s. Zbylé dvě BET365 a Expect jsou pouze internetové sázkové společnosti.

Příklad 4.1. *GAMBRINUS LIGA*

Konečná tabulka Gambrinus ligy v sezóně 2008/2009

Číslo týmu	*Tým*	*Body*	*Skóre*	*Rozdíl skóre*
1	*Bohemians*	*34*	*33:46*	*-13*
2	*Brno*	*35*	*32:36*	*-4*
3	*Budějovice*	*36*	*30:37*	*-7*
4	*Jablonec*	*46*	*43:37*	*6*
5	*Kladno*	*31*	*21:41*	*-20*
6	*Liberec*	*52*	*41:28*	*13*
7	*M.Boleslav*	*46*	*39:38*	*1*
8	*Olomouc*	*48*	*39:36*	*3*
9	*Ostrava*	*39*	*38:36*	*2*
10	*Plzeň*	*43*	*45:38*	*7*
11	*Příbram*	*34*	*30:40*	*-10*
12	*Slavie*	*62*	*57:25*	*32*
13	*Sparta*	*56*	*48:25*	*23*
14	*Teplice*	*43*	*33:25*	*8*
15	*Zlín*	*29*	*26:49*	*-23*
16	*Žižkov*	*22*	*27:45*	*-18*



Obrázek 5: Použitá logistická regrese na výsledky Gambrinus ligy

V obrázku 5 jsou na ose x znázorněny pomocí tmavě zelených koleček četnosti sledovaných zápasů při daném rozdílu bodů v tabulce, červenou tečkou jsou znázorněny pravděpodobnosti sázkové společnosti BET365, černou tečkou je znázorněna sázková kancelář Tipsport a.s. a modrou tečkou sázková společnost Expect.

Koeficienty v použité logistické regresi jsou rovny $\beta_0 = -0{,}732156$ *a* $\beta_1 = 0{,}434279$

Z výše uvedeného obrázku, můžeme určit pravděpodobnosti výhry lépe či hůře postavených týmů v tabulce:

Rozdíl bodů v tabulce	*Pravděpodobnost výhry lépe postaveného*
30	*77 %*
20	*69 %*
15	*64 %*
10	*59 %*
5	*54 %*
-5	*43 %*
-10	*37 %*
-15	*32 %*
-20	*28 %*
-30	*20 %*

Vybrali jsme ty zápasy, kde se pravděpodobnost sázkové kanceláře nejvíce lišila od naší vypočítané pravděpodobnosti, tedy kde byl pro nás kurz nejvíce výhodný. Uvedu příklad na třech zápasech Gammrinus ligy:

Liberec - Ostrava: naše pravděpodobnost je 0,6854, Bet36 má pravděpodobnost 0,3846, diference je 0,3008

Teplice - Příbram: naše pravděpodobnost je 0,6745, Bet36 má pravděpodobnost 0,6667, diference je 0,0078

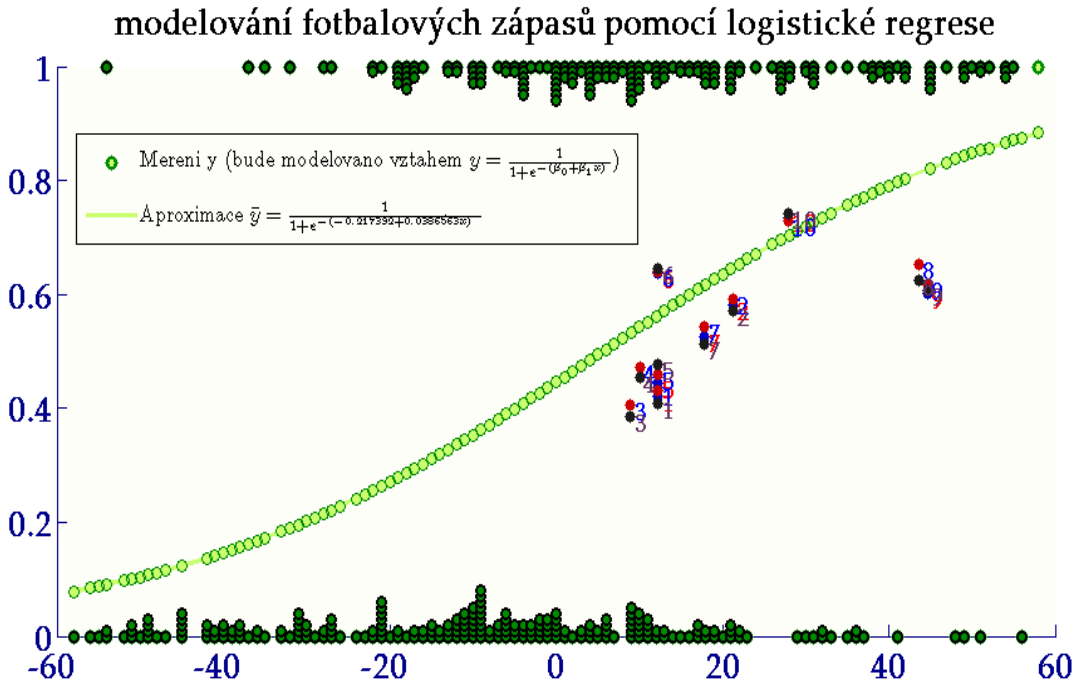
Plzeň - Jablonec: naše pravděpodobnost je 0,5442, Bet36 má pravděpodobnost 0,3448, diference je 0,1994

Z těchto zápasů jsem vybral zápas Liberec – Ostrava, protože u něj byla vypočítána největší diference a současně vyšla vysoká pravděpodobnost výhry Ostravy. Tím pádem podle naší vypočítané pravděpodobnosti nejvíce nadhodnocený kurz a pro nás nejvíce výhodný.

Příklad 4.2. *PREMIER LEAGUE*

Konečná tabulka Premier League v sezóně 2008/2009

Číslo týmu	*Tým*	*Body*	*Skóre*	*Rozdíl skóre*
1	*Manchester United*	*90*	*68:24*	*44*
2	*Liverpool*	*86*	*77:27*	*50*
3	*Chelsea*	*83*	*68:24*	*44*
4	*Arsenal*	*72*	*68:37*	*31*
5	*Everton*	*63*	*55:37*	*18*
6	*Aston Villa*	*62*	*54:48*	*6*
7	*Fulham*	*53*	*39:34*	*5*
8	*Tottenham*	*51*	*45:45*	*0*
9	*West Ham*	*51*	*42:45*	*-3*
10	*Manchester City*	*50*	*58:50*	*8*
11	*Wigan*	*45*	*34:45*	*-11*
12	*Stoke*	*45*	*38:55*	*-17*
13	*Bolton*	*41*	*41:53*	*-12*
14	*Portsmouth*	*41*	*38:57*	*-19*
15	*Blackburn*	*41*	*40:60*	*-20*
16	*Sunderland*	*63*	*34:54*	*-20*
17	*Hull*	*35*	*39:64*	*-25*
18	*Newcastle*	*34*	*40:59*	*-19*
19	*Middlesbrough*	*32*	*28:57*	*-29*
20	*West Bromwich*	*32*	*36:67*	*-31*



Obrázek 6: Použitá logistická regrese na výsledky anglické Premier League

V obrázku 6 jsou na ose x znázorněny pomocí tmavě zelených koleček četnosti sledovaných zápasů při daném rozdílu bodů v tabulce, červenou tečkou jsou znázorněny pravděpodobnosti sázkové společnosti BET365, černou tečkou je znázorněna sázkova kancelář Tipsport a.s. a modrou tečkou sázková společnost Expect.

Koeficienty v použité logistické regresi jsou rovny $\beta_0 = -0{,}2173924$ *a* $\beta_1 = 0{,}386563$.

Můžeme určit pravděpodobnosti výhry lépe či hůře postavených týmů v tabulce:

Rozdíl bodů v tabulce	*Pravděpodobnost výhry lépe postaveného*
30	*72 %*
20	*63 %*
15	*59 %*
10	*54 %*
5	*50 %*
-5	*40 %*
-10	*35 %*
-15	*31 %*
-20	*27 %*
-30	*20 %*

Na anglickou Premiere League jsem aplikoval stejnou metodu výběrů zápasů, tedy ty, které jsou pro nás nejvíce výhodné.

Na sázkový tiket jsem pak vsadil 4 nejvýhodnější zápasy z Gambrinus ligy a 5 zápasů z Premier League. Vsazené částka byla 100 Kč a možná výhra byla 158 027,95 Kč. Sázková společnost BET365 nemá žádný manipulační poplatek. Pokud vsázím na několik zápasů dohromady, vsazené kurzy se mezi sebou násobí a výsledným kurzem je potom vynásobem vklad. Proto je možná výhra tak velká.

Obrázek 7: Vsazený tiket u sázkové společnosti BET365

Ve druhé části experimentu jsem se věnoval predikci konečné podoby tabulky obou fotbalových soutěží. K tomu jsem využil opět logistickou regresi. Pomocí ní jsem si vypočítal pravděpodobnosti výhry, prohry a remízy všech týmů v daném kole. Ty jsem vždy porovnal s náhodně vygenerovaným číslem. Pokud bylo vygenerované číslo menší jak pravděpodobnost výhry domácích, tak jsem domácím připsal 3 body za výhru. Pokud bylo vygenerované číslo menší jak pravděpodobnost výhry domácích+pravděpodobnost výhry hostů, tak jsem připsal 3 body hostům a v opačném případě každému týmu 1 bod za remízu. Takhle jsem postupoval ve všech zbývajících kolech. Při každé simulaci dopadli výsledky jinak, proto jsem simulaci opakoval 10 krát a napsal jsem závěrečnou tabulku, která vyšla nejčastěji.

Předpokládaná konečná tabulka Gambrinus ligy v sezóně 2009/2010

Pořadí týmu	*Tým*	*Body*
1	*Ostrava*	*60*
2	*Sparta*	*57*
3	*Jablonec*	*57*
4	*Teplice*	*56*
5	*M.Boleslav*	*47*
6	*Olomouc*	*46*
7	*Plzeň*	*45*
8	*Slavia*	*44*
9	*Liberec*	*40*
10	*Bohemians 1905*	*35*
11	*Brno*	*34*
12	*Příbram*	*34*
13	*Budějovice*	*27*
14	*Slovácko*	*26*
15	*Kladno*	*21*
16	*Bohemians*	*17*

Předpokládaná konečná tabulka Premier League v sezóně 2009/2010

Pořadá týmu	*Tým*	*Body*
1	*Manchester United*	*85*
2	*Chelsea*	*83*
3	*Arsenal*	*78*
4	*Tottenham*	*73*
5	*Liverpool*	*68*
6	*Aston Villa*	*68*
7	*Manchester City*	*66*
8	*Everton*	*63*
9	*Birmingham*	*54*
10	*Fulham*	*49*
11	*Sunderland*	*45*
12	*Stoke*	*43*
13	*Blackburn*	*43*
14	*Bolton*	*38*
15	*Wolverhampton*	*38*
16	*Wigan*	*36*
17	*West Ham*	*34*
18	*Hull*	*29*
19	*Burnley*	*27*
20	*Portsmouth*	*18*

Závěr

V práci jsem ukázal, že lze téměř ve všech sportovních odvětvích aplikovat různé instrumenty teorie pravděpodobnosti a matematické statistiky. Pomocí nich můžeme dojít k zajímavým výsledkům a to nejen z pohledu sportovních expertů ale i z pohledu laiků.

Hlavním cílem práce bylo určení pravděpodobností výhry jednotlivých týmů v ligových soutěžích a v porovnání s kurzy sázkových kanceláří vybrat ty nejvýhodnější kurzy a ty vsadit. Vsazený tiket bohužel nevyšel. Ale to jen potvrdilo to, že fotbal je ovlivňován mnoha faktory. Ve své práci jsem zohledňoval pouze faktor postavení týmu v tabulce a rozdíl bodů mezi těmito týmy. Ze vsazených 9 zápasů, kde jsem vsázel pouze na čistou výhru lépe postaveného týmu v tabulce, vyšly 4 zápasy (44 %). Ale pokud bych vsadil i na remízu, tedy neprohru lépe postaveného týmu, tak by mi z 9 zápasů vyšlo 7 (78 %). I když na první pohled se může zdát, že experiment byl neúspěšný, protože jsem nevyhrál. Na druhou stranu 78 % není tak malé číslo a nebýt nečekaných dvou proher favoritů Premier League Chelsea a Arsenalu, tak jsem mohl být ještě úspěšnější.

Jak už jsem zmínil výsledky ve fotbale, ale i v jiných kolektivních sportech jsou ovlivňovány mnoha faktory, které je potřeba zohlednit, abychom byli při sázení úspěšní. V tomhle duchu by se dalo i dále pokračovat v této bakalářské práci.

Při zpracovávání bakalářské práce a při aplikaci uvedených postupů jsem si prohloubil znalosti ze statistiky. Naučil jsem se pracovat s programem Matlab i psát v systému TEX.

Věřím, že znalosti, dovednosti a zkušenosti získané při psaní této bakalářské práce využiji při dalším studiu a v praxi.

Literatura

[1] Albert, J., Bennett, J., Cochran J.J: Anthology of Statistics in Sports., SIAM, Philadelphia, 2005

[2] Anděl, J.: Základy matematické statistiky, Univerzita Karlova v Praze, Praha 2002

[3] Kunderová, P.: Základy pravděpodobnosti a matematické statistiky, Vydavatelství UP Olomouc, 2004

[4] Seber, G.A.F.: Nonlinear regression, John Wiley & sons, New Jersey, 2003

[5] Michalewicz, Z., Fogel, D.B.: How to Solve It: Modern Heuristics, Springer-Verlag, Berlin Heidelberg, 2000

[6] NHL, statistiky hráčů. Dostupné z:
http://www.nhl.com/ice/player.htm

[7] LOH, statistiky atletů. Dostupné z:
http://www.databasesports.com/olympics/

[8] Gambrinus liga, výsledky utkání. Dostupné z:
http://fotbal.sport.cz/fotbal/gambrinus-liga/

[9] Premier League, výsledky utkání. Dostupné z:
http://fotbal.sport.cz/fotbal/premier-league/

[10] New York Times, Professor Puts a Face on the Performance of Baseball Managers. Dostupné z:
http://www.nytimes.com/2008/04/01/science/01prof.html

[11] Kurzové sázení. Dostupné z :
http://cs.wikipedia.org/wiki/Kurzové_sázení

[12] Nejhorší tenista světa. Dostupné z :
http://www.telegraph.co.uk/sport/tennis/7644181/Worlds-worst-tennis-player-loses-again.html