Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra matematiky

Diplomová práce

Statistické a rozhodovací postupy při sázení

Plzeň, 2015 Tomáš Le Van

Místo tohoto listu bude vloženo zadání.

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně pod vedením vedoucího a výhradně s použitím citovaných pramenů.

V Plzni dne 8. 5. 2015 .....

Tomáš Le Van

Poděkování

Velmi rád bych chtěl poděkovat vedoucímu mé diplomové práce panu doc. Ing. Františku Vávrovi, CSc., za jeho odborné rady, vstřícný přístup a čas věnovaný při konzultacích během vytváření této práce.

Abstrakt

Cílem této práce je využití poznatků z teorie sázek při sázení jakožto jednoho z možných investičních nástrojů. Práce je zaměřena na kurzové sázení na nejvyšší fotbalové soutěže Česka, Anglie a Německa. Zkoumaným typem sázek je sázení na výhru domácích, remízu či výhru hostů. Problematika je řešena členěním zápasů do skupin podle vybraných kritérií a následným odhadováním parametrů multinomického rozdělení pro tyto tři základní jevy. Skupiny jsou vytvářeny především na základě pořadí v tabulce, počtu vstřelených gólů a jiných průběžných výsledků v soutěži. I přesto, že nebyl nalezen přesvědčivý model přinášející zisk, je v práci ukázáno, že je možné využít multinomického rozdělení pro odhadování výsledků zápasů. Jak bylo zjištěno, klíčovou překážkou v nalezení strategie přinášející dlouhodobý zisk jsou marže sázkových kanceláří. V případě sázení na kurzy upravené právě o marži by bylo dosaženo zisku na datech Česka a Anglie za posledních 5 (resp. 7) sezón, přičemž by bylo provedeno 894 (resp. 1444) sázek.

Klíčová slova: kurzové sázení, fotbal, proporcionální sázení, investování, multinomické rozdělení, odhadování parametrů

Abstract

The purpose of this thesis is to apply the knowledge of betting theory as one of the possible investment instruments. The thesis deals with fixed odds betting on the highest Czech, English and German football leagues. The type of betting described is betting on home-win, draw or away-win. The problem is solved by sorting matches into groups according to the selected criteria and then estimating parameters of multinomial distribution for these three basic elements. The criteria are based on the ranking in table, the number of scored goals and other results in the competition. Even though we have not been able to find convincing model, which would make a profit, it was shown that it is possible to use multinomial distribution for estimating the results of matches. As was discovered, the key obstacle in finding a strategy making long-term profit is bookmaker's margin. In the case of „fair odds“ (adjusted without margin) the presented strategy would make a profit on Czech (or English) data for the last 5 (or 7) seasons while placing 894 (or 1444) bets.

Keywords: fixed odds betting, football, proportional betting, investing, multinomial distribution, parameters estimation

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Základní myšlenkou sázení, stejně jako u všech ostatních hazardních her, je umožnit sázejícím získat s určitou pravděpodobností finanční odměnu. Výše této odměny je závislá na výši částky, se kterou hráč do hry vstupoval, kterou vsadil. Tato práce se zabývá sázením na sportovní kurzové sázky, konkrétně pak sázením na fotbal. V tomto případě hráči sází u provozovatele sázkové kanceláře na některý z možných sportovních výsledků nebo na určitou událost, která během utkání může nastat.

V dnešní době existuje řada sázkových kanceláří, které vypisují kurzy na různé sportovní sázky. Nutno dodat, že většina sázkových kanceláří funguje pro sázejícího na stejném principu a liší se jen v konkrétních detailech (výše kurzů, typy vypsaných sázek, aj.). Mezi nejčastější sportovní kurzové sázky patří jednoduché „1-0-2“ sázení, kdy sázející mohou sázet na jeden ze tří možných výsledků zápasu, a sice na výhru domácích, remízu či výhru hostů. Ukázku možného vypsání kurzů na takový typ sázky je možné najít v následující tabulce.

Utkání	1	0	2
Slavia - Sparta	4.6	3.65	1.75
Liberec - Ml. Boleslav	2.03	3.3	3.5

Tabulka: Vypsané kurzy na dva zápasy nejvyšší české fotbalové ligy

Tento typ sázky je typický pro dlouhodobé soutěže v kolektivních sportech (např. fotbal nebo hokej). Pro individuální sporty bývá někdy typičtější „1-2“ sázení (prostá výhra či prostá prohra). Kromě sázky na výsledek jednoho utkání je možné vsadit například na vítěze nějaké dlouhodobé soutěže ještě před jejím začátkem (typicky sázky na vítěze celé soutěže či turnaje).

Kromě zmíněných základních sázek a jiných speciálních sázek[1] existují sázky odvozené, kdy lze například vsadit na „neprohru“ nějakého týmu. Tyto sázky jsou specifické zejména proto, že jejich kurzy bývají pevně svázány a odvozeny od původních základních kurzů. Těchto sázek lze tedy dosáhnout vhodnou kombinací sázek základních. Tato práce se zabývá odhadováním právě základních výsledků.

Důležitým faktem při sázení je hodnota kurzu vypsaného sázkovou kanceláří. Jak je ukázáno a vysvětleno dále v této práci, vypsané kurzy nejsou konstantní v čase a mohou se tedy lišit v závislosti na tom, kdy je sázka uzavírána. V momentě rozhodování o sázce však sázející nemá žádnou informaci o budoucím vývoji těchto kurzů a jeho rozhodování je dáno situací v ten daný okamžik, pokud nevolí možnost odložit rozhodnutí na později. Uzavře-li sázející sázku s daným kurzem, je tento kurz již pro sázku pevný a není možné jej v rámci této sázky změnit. V literatuře se tento druh sázení někdy nazývá jako *fixed odds bet*.

Tato práce uvádí a rozvíjí techniky, které lze využít pro sázení a které by měly pomoci sázkaři k zisku.

[1] Speciálními sázkami se tato práce nebude zabývat. Jedná se například o sázky, zda první gól padne z penalty, zda daný hráč nastoupí do zápasu, nebo v které části zápasu padne první žlutá karta a mnoho dalších.

Souvislost sázení a investování

Z ekonomického hlediska lze pojem investice chápat jako prozatímní odložení prostředků či statků k pozdější spotřebě s tím, že za tuto dobu jejich hodnota neklesne nebo ještě lépe – vzroste. V současném běžném životě jsou tradičními investičními nástroji především různé typy bankovních účtů, cenné papíry, nemovitosti, zlato a jiné drahé kovy či diamanty. Nejen do těchto statků lze ukládat peníze a později je opět směnit.

Takto je možné postupovat v případě jednorázové investice, u které lze na konci investičního období změřit její výnosnost. Existují ale také jiné druhy investic a sice takové, které přinášejí pravidelný příjem. V případě akcií se jedná o pravidelně vyplácenou dividendu. V případě pronájmu nemovitosti to může být část nájemného placeného nájemníkem. Investiční rozhodování pak ohodnocuje jednotlivé peněžní toky daného projektu, a tím měří jeho výhodnost. Investor, který svou investicí dosáhne požadovaného zisku, je úspěšný investor. Naopak v případě, že instrumenty během průběhu dané investice ztratí svou původní hodnotu či nedosáhnou hodnoty, která byla očekávána (požadována), je taková investice chápána jako neúspěšná.

Jaká je tedy souvislost mezi sázením a investováním? Kurzová sázka je uzavření dohody mezi dvěma stranami, kdy jedna ze stran přijímá vklad (vsazenou částku) od druhé, a v případě, že nastane situace, na kterou protistrana sázela, vyplatí vsazenou částku násobenou kurzem. V celé této práci je příjemcem vkladu provozovatel sázkové kanceláře, jenž také vypisuje kurzy. Sázející se tedy může sám rozhodnout, na které sázky přistoupí. Zatímco investice je zpravidla dlouhodobá záležitost, v případě sázky může být výsledek znám okamžitě. Dalším rozdílem je, že v případě tradiční investice není běžné, aby hodnota investovaných prostředků spadla na nulu, avšak v případě neúspěšné sázky ztrácí investor okamžitě celou vsazenou částku[2].

Sázka na jeden výsledek je tedy na rozdíl od tradiční investice velmi riziková a krátkodobá záležitost. Aby bylo možné brát sázení jako investici a nikoliv nepodložený hazard, je třeba nejen umět správně odhadovat výsledky vybraných událostí, ale také řídit zásobu peněz určenou k sázení. Tím se pohled na sázení mění z nepodloženého hazardu na dlouhodobou investici. Riziko této investice je spojeno s rozhodovacími postupy investora, jeho vztahem k riziku a požadovaným ziskem. Celkový objem prostředků určených k sázení je tedy investovaným kapitálem a očekávané peněžní toky jsou dány výsledkem jednotlivých sázek. V případě úspěšné investice musí výhry převládat nad vsazenými částkami. V následujících kapitolách jsou představeny a využity některé statistické metody a postupy, které mohou pomoci nalézt takovou strategii sázení, jež by mohla být využita jako úspěšný investiční projekt.

Pro zpracování praktické části práce bylo využito výpočetního software MATLAB společnosti *MathWorks* a tabulkového procesoru Microsoft Excel od společnosti *Microsoft*.

[2]Existují i speciální sázky, které v případě neúspěchu či pouze částečného úspěchu vrací část prostředků zpět.

1 Některé teoretické prostředky

V této části jsou představeny a shrnuty poznatky z matematické statistiky, teorie pravděpodobnosti a teorie informace, které budou dále využívány a které jsou potřebné pro sestavení sázejících modelů (strategií). Všechny tyto poznatky využívané za účelem sázení lze shrnout do oblasti nazývané *teorie sázek*, která je podrobněji rozepsána v následující podkapitole.

1.1 Úvod do teorie sázek

V této podkapitole je stručně představena klasická teorie sázek. Většina poznatků je čerpána především z [1] a [2]. Názorně je úvod odvozen pro sázení na dostihové závody.

Přestože je celá tato kapitola odvozena na příkladě pro dostihové závody koní, lze ji využít také v jakémkoli jiném (sportovním) sázení. Analogicky se místo vítězství jednoho ze závodících koní uvažuje například jeden z možných výsledků sportovního utkání. V případě sázky na výhru, remízu či prohru nějakého týmu v zápase se opět jedná o situaci, kdy právě a pouze jeden z možných výsledků musí nastat (vítězství koně X v závodě). Je tedy zřejmé, že výstavba teorie sázek na dostihových závodech neubírá v této práci na obecnosti.

Definice 1.1.1 *Předpokládejme dostihový závod s $m \geq 2$ koňmi, kde pravděpodobnost vítězství i-tého koně je $p_i \geq 0$ a $\sum_{i=1}^{m} p_i = 1$. Označme o_i koeficient výše výplaty (kurz) při výhře i-tého koně a $b_i \geq 0$ částku, kterou sázkař vsadí na i-tého koně. Potom pro výši výplaty $S(i)$ v případě výhry i-tého koně platí*

$$S(i) = b_i o_i.$$

Kurz o_i tedy v případě výhry udává počet vyplacených jednotek za každou vsazenou jednotku. Aby mělo pro sázkaře smysl sázet na i-tý kurz, musí platit $o_i > 1$. Čistý zisk z takové sázky je potom roven $o_i - 1$. V případě prohry sázkař prohrává celou svoji vsazenou částku.

Uvažujme nyní případ, kdy sázkař rozloží své prostředky v rámci jedné sázky mezi více koní, tj. $\forall i$ platí $b_i \geq 0$ a $\sum b_i = 1$. Vsazenou částkou je tedy jedna celá jednotka a b_i značí její jednotlivé podíly.

Definice 1.1.2 *Nechť sázkař provádí opakované sázky na stejný dostihový závod se stále stejnými podmínkami. Označme X_i vítězného koně v i-tém kole sázky, tzn. $S(X_i) = b_{X_i} o_{X_i}$ značí výši výplaty po i-tém kole, potom pro sázkařovy prostředky po n dostizích S_n platí*

$$S_n = \prod_{i=1}^{n} S(X_i).$$

Prostředky po n dostizích jsou tedy dány součinem všech dílčích faktorů $S(X_i)$. V tomto případě tedy sázkař v každém kole sází všechny své prostředky do následujícího dostihu. Částka S_n je však náhodnou veličinou a sázkař se ji v některém smyslu snaží „maximalizovat“.

Definice 1.1.3 *Nechť $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ je vektor všech možných výsledků dostihového závodu s pravděpodobnostní funkcí $p(\boldsymbol{X})$, kde m značí počet koní (počet možných výsledků pro jednu sázku) v dostihu, a nechť $S(\boldsymbol{X})$ je vektor jejich násobících faktorů. Potom míra zdvojení $W(\boldsymbol{b}, p)$ pro rozložení prostředků $\boldsymbol{b}$ je v takovém dostihu definována*

$$W(\boldsymbol{b}, p) = E(\log_2 S(\boldsymbol{X})).$$

Míru zdvojení lze tedy z definice střední hodnoty spočíst jako

$$W(\boldsymbol{b}, p) = \sum_{i=1}^{m} p_i \log_2(b_i o_i). \tag{1.1.1}$$

Věta 1.1.1 *Pokud $W(\boldsymbol{b}, p)$ je míra zdvojení pro vektor $\boldsymbol{b}$, pak pro stav prostředků po n krocích S_n platí*

$$S_n \doteq 2^{nW(\boldsymbol{b},p)}.$$

Důkaz této věty lze nalézt v [1]. Míra zdvojení je jakýsi koeficient (exponent) zdvojnásobení prostředků. Zjednodušeně řečeno, měří tedy rychlost růstu střední hodnoty prostředků při daném typu sázení. V případě kladné míry zdvojení roste sázkařův kapitál exponenciálně, zatímco v případě záporné $W(\boldsymbol{b}, p)$ jde exponenciálně k nule.

Definice 1.1.4 *Optimální míra zdvojení $W^*(p)$ je definována jako*

$$W^*(p) = \max_{\boldsymbol{b}} W(\boldsymbol{b}, p),$$

kde maximum je přes všechny možné strategie $\boldsymbol{b}$.

Pro nalezení $\boldsymbol{b}$, které maximalizuje $W(\boldsymbol{b}, p)$, sestavíme Lagrangeovu funkci

$$\Phi(\boldsymbol{b}, \lambda) = \sum_{i=1}^{m} p_i \log_2(b_i o_i) + \lambda \left(\sum_{i=1}^{m} (b_i) - 1 \right). \tag{1.1.2}$$

Derivacemi podle b_i získáme

$$\frac{\partial \Phi}{\partial b_i} = \frac{p_i}{b_i} + \lambda, \, \forall i. \tag{1.1.3}$$

Položením parciálních derivací rovno nule získáváme

$$b_i = -\frac{p_i}{\lambda}. \tag{1.1.4}$$

Protože řešení musí splňovat podmínku $\sum b_i = 1$, dostáváme stacionární bod, pokud

$$\boldsymbol{b} = \boldsymbol{p}. \tag{1.1.5}$$

Že se jedná o maximum, lze snadno ukázat druhými parciálními derivacemi. Optimální rozložení prostředků je tedy ekvivalentní rozdělení pravděpodobností výher jednotlivých koní neboli všech možných výsledků toho, na co je sázeno. Takovéto sázení se nazývá proporcionální sázení nebo také sázení dle Kellyho. Podrobnosti k tomuto typu sázení jsou sepsány v následujícím odstavci, neboť je nutné uvést další související teoretické pojmy.

1.1.1 Sázení dle Kellyho

V tomto odstavci jsou rozvedeny poznatky z teorie sázek. Konkrétně je zde představena metoda proporcionálního sázení, která je často nazývána jako sázení dle Kellyho.

Definice 1.1.5 *Entropie diskrétní náhodné veličiny X je definována vztahem*

$$H(X) = -\sum_{x\in\Omega} p(x)\log_2(p(x)),$$

kde Ω značí množinu všech elementárních jevů.

Symbol Ω je pro množinu všech elementárních jevů používán i dále. Entropie je tedy definována jako střední hodnota náhodné veličiny $\frac{1}{\log_2 p(x)}$ a její jednotky se nazývají bity.

Definice 1.1.6 *Nechť $p(X)$ a $q(X)$ jsou pravděpodobnostní funkce dvou rozdělení. Relativní entropie mezi rozdělením odpovídající $p(X)$ a rozdělením s $q(X)$ je definována jako*

$$D(p||q) = E_p \log_2 \frac{p(X)}{q(X)} = \sum_{x\in\Omega} p(x)\log_2 \frac{p(x)}{q(x)}.$$

Relativní entropie bývá někdy nazývána *Kullback-Leiblerova vzdálenost* nebo *Kullback-Leiblerova divergence*. Ačkoliv nejde přímo o metriku, tak nulová je pouze pokud $p = q$. Pokud existuje x takové, že $p(x) > 0$ a $q(x) = 0$, potom se dodefinovává $D(p||q) = \infty$. Nezápornost a další vlastnosti jsou dokázány v [1].

Věta 1.1.2 *Optimální míru zdvojení lze vyjádřit vztahem*

$$W^*(\boldsymbol{p}) = \sum_{i=1}^{m} p_i \log_2 o_i - H(\boldsymbol{p})$$

a je jí dosaženo proporcionálním sázením $\boldsymbol{b} = \boldsymbol{p}$.

Důkaz: Přepíšeme tedy předpis pro míru zdvojení:

$$W(\boldsymbol{b},\boldsymbol{p}) = \sum_i p_i \log_2 b_i o_i, \tag{1.1.6}$$

$$= \sum_i p_i \log_2 \left(\frac{b_i}{p_i} p_i o_i\right), \tag{1.1.7}$$

$$= \sum_i p_i \log_2 o_i - H(\boldsymbol{p}) - D(\boldsymbol{p}||\boldsymbol{b}), \tag{1.1.8}$$

$$\leq \sum_i p_i \log_2 o_i - H(\boldsymbol{p}), \tag{1.1.9}$$

kdy rovnost nastává pouze v případě $\boldsymbol{p} = \boldsymbol{b}$. Tím je také dokázáno, že řešení (1.1.5) je opravdu maximem. Říkáme tedy, že proporcionální sázení je *log-optimální*.

Uvažujme nyní speciální případ, kdy sázková kancelář vypíše vzhledem k odhadnutým pravděpodobnostem spravedlivé kurzy (tzn. $\sum \frac{1}{o_i} = 1$). V tomto případě jsou tedy pravděpodobnosti odhadnuté sázkovou kanceláří na jednotlivé výsledky rovny $r_i = \frac{1}{o_i}$. Potom pro míru zdvojení platí

$$W(\boldsymbol{b}, \boldsymbol{p}) = \sum_i p_i \log_2 b_i o_i, \tag{1.1.10}$$

$$= \sum_i p_i \log_2 \left(\frac{b_i p_i}{p_i r_i} \right), \tag{1.1.11}$$

$$= D(\boldsymbol{p}||\boldsymbol{r}) - D(\boldsymbol{p}||\boldsymbol{b}). \tag{1.1.12}$$

Tento vztah ukazuje, že na míru zdvojení lze nahlížet jako na rozdíl „vzdálenosti“ mezi skutečnou pravděpodobností a odhadem sázkové kanceláře $D(\boldsymbol{p}||\boldsymbol{r})$ a „vzdálenosti“ mezi skutečnou pravděpodobností a odhadem sázejícího $D(\boldsymbol{p}||\boldsymbol{b})$. Pokud tedy odhad sázejícího bude skutečnosti „blíže“ nežli odhad sázkové kanceláře, bude míra zdvojení kladná a ve střední hodnotě může sázejícímu přinášet zisk. V opačném případě půjde hodnota sázkařových prostředků k nule.

Důležitou vlastností sázení dle Kellyho je, že optimální rozložení prostředků $\boldsymbol{b}$ není dáno hodnotami stanovených kurzů o_i, nýbrž skutečnými pravděpodobnostmi $\boldsymbol{b}$. Další vlastností je, že jakákoliv odchylka sázkové kanceláře od skutečné pravděpodobnosti pomáhá sázejícímu.

Problém u tohoto přístupu k sázení nastává v případě, kdy odhady sázkové kanceláře a potažmo i vypsané kurzy jsou přesné. Další komplikací je existence marží sázkových kanceláří. Více o sázkách s marží je uvedeno dále v následující podkapitole 1.2.

Předpokládejme nyní tedy, že odhady sázkové kanceláře jsou přesné, tj. $r_i = p_i$. Dále předpokládejme, že vypsané kurzy jsou snížené o marži ξ, tzn. $o_i = \frac{(1-\xi)}{r_i}$. Za zmíněných předpokladů je pak střední hodnota výhry ze sázky při libovolném rozložení prostředků $\boldsymbol{b}$ rovna $(1-\xi)\sum_i b_i$, neboť platí, že

$$E(w) = \sum_{i=1}^{m} r_i b_i o_i. \tag{1.1.13}$$

Za o_i lze dosadit dle předpokladu $\frac{(1-\xi)}{r_i}$ a dostáváme tvar

$$E(w) = (1-\xi) \sum_{i=1}^{m} b_i. \tag{1.1.14}$$

Sázkař tedy bez ohledu na to, jak rozložil své prostředky, prodělává z každé sázky ve střední hodnotě marži.

Tento fakt však neznamená, že sázkař nemůže při jedné sázce (nebo jejich konečném počtu) vyhrát. Často je pro sázejícího důležitá okamžitá výhra, nikoliv dlouhodobý výsledek. V každé sázce se tedy sázkař může rozhodovat na základě okamžité očekávané výhry ze sázky. Střední hodnota výhry ze sázky je však velmi důležitá pro sázkovou kancelář, která si marží a přibližováním odhadů ke skutečné pravděpodobnosti zajišťuje svoji dlouhodobou ekonomickou udržitelnost.

Férová a sub/super-férová sázka. V tomto odstavci jsou stručně představeny tři „krajní“ možnosti stanovení kurzů vzhledem k nějakému pravděpodobnostnímu rozdělení výsledků.

1. **Spravedlivé sázky:** Musí být splněna podmínka $\sum \frac{1}{o_i} = 1$. V tomto případě je sázka spravedlivá, protože pokud by sázkař rozděloval své prostředky proporcionálně (jinak řečeno, míra zdvojení by byla nulová), nezáleželo by na střední hodnotě výsledku dostihu a sázkařovy prostředky by po každém dostihu zůstávaly stejné.

2. **Super-spravedlivé sázky:** Musí být splněna podmínka $\sum \frac{1}{o_i} < 1$. V tomto případě jsou kurzy vyšší než spravedlivé, a v případě, že sázkař sází opět proporcionálně, jsou faktory, kterými se násobí jeho prostředky, $S(X) = 1/\sum \frac{1}{o_i} > 1$, a tudíž taková strategie vede k bezrizikovému zisku. V praxi se sázkové kanceláře snaží upravovat své kurzy právě takovým způsobem, aby sázkař nemohl získat takovéto příležitosti například využitím kurzů u různých sázkových kanceláří.

3. **Sub-spravedlivé sázky:** U takové sázky pro kurzy platí $\sum \frac{1}{o_i} > 1$. Tento případ je v reálném životě nejčastější. Aby sázkař dosáhl stejné výplatní částky jako u spravedlivých kurzů, musí vsadit vyšší částku, než je hodnota „celku“. Sázková kancelář si v tomto případě do kurzů zahrnuje marži, která jí zajišťuje zisk. Proporcionální sázení už tedy není log-optimální. Lze ukázat, že v tomto případě je vypočtená hodnota výrazu $D(\boldsymbol{p}||\boldsymbol{r})$ ve vztahu (1.1.12) nižší. To je dáno tím, že r_i již nejsou pravděpodobnosti ($\sum_i r_i > 1$), a pro sázkaře je tedy nutně těžší (či dokonce nemožné) dosáhnout tímto způsobem zisku.

1.2 Kurzové sázky z pohledu sázkové kanceláře

Základní myšlenkou sázkové kanceláře je dát sázkaři možnost s určitou nenulovou pravděpodobností získat finanční výhru. Za tuto možnost přitom sázkař musí zaplatit vkladem (sázkou).

Podobně jako v předchozím odstavci je uvažován dostihový závod s m koňmi. I v případě, že bude sázet více než jeden sázkař, lze z pohledu sázkové kanceláře objemy jejich sázek spojit a uvažovat je jako jednu sázku. Přejdeme tedy k definici očekávané výplaty.

Definice 1.2.1 *Nechť pro dostihový závod s m koňmi platí, že b_i je úhrn všech sázek na i-tého koně. Nechť dále r_i značí odhady pravděpodobností sázkové kanceláře na výhru i-tého koně, kterým odpovídají kurzy o_i. Potom střední očekávaná hodnota výplaty ze sázky $E(w)$ je rovna*

$$E(w) = \sum_{i=1}^{m} r_i b_i o_i.$$

Střední hodnotu výplaty lze tedy chápat jako hodnotou všech možných sázkovou kanceláří očekávaných výplat vážených pravděpodobnostmi, které odhaduje sázková kancelář.

Pokud sázková kancelář stanoví spravedlivé kurzy, je hodnota $E(w)$ rovna $\sum_i^m b_i$, což je objem vložených prostředků sázkařů. Ve střední hodnotě by tedy sázková kancelář nevydělávala ani neprodělávala. Z tohoto pohledu je nutné, aby sázková kancelář vypisovala sub-spravedlivé kurzy, pro které lze psát

$$r_i o_i = 1 - \xi, \tag{1.2.1}$$

kde ξ značí hrubou ziskovou marži.

Tento přístup také umožňuje odhadovat[1] z vypsaných kurzů hrubou ziskovou marži sázkové kanceláře. Odhad lze jednoduše vyjádřit vztahem

$$\xi = 1 - r_i o_i. \tag{1.2.2}$$

Pravděpodobnosti r_i ovšem sázkaři nejsou známé a je potřeba je odhadnout[2] vztahem

$$\hat{r_i} = \frac{1}{o_i}. \tag{1.2.3}$$

Protože však sázky nejsou spravedlivé a $\sum_i^m \hat{r_i} \neq 1$, je nutné tyto „pravděpodobnosti“ znormovat konstantou $\sum_i^m \hat{r_i}$. Pro výpočet marže dostáváme

$$\xi = 1 - \left(\frac{1}{o_i} \frac{1}{\sum_i^m \frac{1}{o_i}} \right) o_i, \tag{1.2.4}$$

$$\xi = 1 - \left(\sum_i^m \frac{1}{o_i} \right)^{-1}. \tag{1.2.5}$$

Právě vztah (1.2.5) je dále používán při výpočtech marží sázkových kanceláří.

Jak bylo zmíněno v úvodu textu, tato práce se blíže zabývá sázkami s pevnými kurzy. Během této práce se neřeší problematika kolísání kurzů v čase, které ve skutečnosti nastává. Ukázku takového kolísání zobrazují tabulky, které lze nalézt v příloze A.1. Z tabulek lze přečíst rozdíl v kurzech za 4 dny u jedné sázkové kanceláře.

Jedním z možných vysvětlení kolísání kurzů může být zajišťování kanceláře proti příliš vysokým výplatám (třeba i méně pravděpodobným). Kolísání kurzů v čase může být tedy způsobeno interakcí mezi sázejícími a sázkovými kancelářemi. Dalším možným vysvětlením může být proces získávání nových informací (o zraněných hráčích, výměnách trenérů, atd.). Zkoumání těchto příčin však přesahuje rámec této práce. Dále se tedy pracuje s kurzy dostupnými v okamžiku sázení.

[1] Jedná se pouze o odhad hrubé ziskové marže, neboť tento výpočet předpokládá, že sázková kancelář snižuje kurzy pro všechna i stejným poměrem, to však nemusí být pravda.

[2] Opět se jedná pouze o odhad (náhradu), kvůli neznámým skutečným maržím a odhadům pravděpodobností sázkových kanceláří.

Dalším problémem, který ovlivňuje stanovení kurzu sázkovou kanceláří, je existence jiných sázkových kanceláří, které vypisují sázky na stejná utkání. V tomto případě se kurzy sázkových kanceláří musí stanovovat takovým způsobem, jenž zamezí sázkařům získat jejich kombinacemi super-spravedlivé sázky.

Dalším vlivem při stanovování kurzů může být existence konkurenčního prostředí, kdy se sázkové kanceláře snaží získat klienty například nižšími maržemi. V této práci je při získávání informací ze stanovených kurzů využíváno vždy kurzů jedné sázkové kanceláře. Těch je využíváno i při odvozování či ověřování modelů. U některých datových zdrojů nejsou dostupné kurzy jedné konkrétní sázkové kanceláře, ale průměrné kurzy z více sázkových kanceláří. V takovém případě se předpokládá, že se jedná o jednu konkrétní sázkovou kancelář.

1.3 Sázení z pohledu sázejících

Z pohledu sázejícího je vypsaná sázka šancí vyhrát odměnu. Výše odměny je daná kurzem a měla by klesat se zvětšující se pravděpodobností výhry. Pokud by se úspěšnost sázkaře měřila poměrem výher ku vsazeným prostředkům, lze očekávat, že budou existovat sázkaři s různou úspěšností. Do jaké míry je tato úspěšnost výsledkem náhody a do jaké míry se jedná o skutečný rozdíl v úspěšnosti, rozhodují především strategie sázejících.

Jistě je možné, aby sázkař využíval známé poznatky z teorie sázek. Bez jakéhokoli důkazu se zde předpokládá, že významná část sázkařů se nerozhoduje na základě známé teorie sázek, ale na základě jiného (intuitivního) přístupu k sázení. Odpovídá to tedy situaci, kdy tito sázkaři sází „pro zábavu“ a nikoliv za účelem porazit sázkovou kancelář.

Některé základní či intuitivní přístupy jsou popsány dále a část z nich lze využívat při porovnání s dále navrženými modely. Tyto naivní (intuitivní) modely jsou pouhou domněnkou toho, jak je možné se rozhodovat při sázení. Jejich existence či četnost v populaci sázkařů není zkoumána. Výše částek, které sázkaři obecně sází, je bez datových zdrojů velmi obtížné zkoumat a jistě budou záviset na mnoha zde nepodstatných faktorech, jako je movitost sázkaře, racionálnost či averze k riziku.

Fundamentální přístup. Jedná se o přístup na základě některých (často i zákulisních) informací vycházejících například z médií. Často se diskutují personální obsazení v klubech (zranění, přestupy, apod.). Lze sledovat také strategii, motivaci či nasazení hráčů (někdy může jednomu z týmů stačit remíza apod.).

Subjektivní odhad. Sázkař se snaží subjektivně odhadnout nejpravděpodobnější variantu a na tu vsadí obnos dle uvážení.

Náhodný tip. Sázkař sází nesystematicky a bez strategie.

Oblíbený tým. Jistě se najde mnoho fanoušků, kteří sází na svůj oblíbený tým. Tento fakt pak může podporovat jejich zájem při fandění během utkání.

Jednoduché pravidlo. Mezi jednoduchá pravidla mohou patřit různá pravidla typu:

- sázka na favorita ($min(o_i)$),
- sázka na outsidera ($max(o_i)$),
- sázka na kurz splňující nějakou podmínku (např. $o_i < 1.5$),
- sázka na domácí tým (příp. hosty či remízu).

Všechny tyto přístupy je možné samozřejmě kombinovat, případně nějak modifikovat či vymýšlet podobné.

1.4 Vybrané statistické pojmy a definice

Tato podkapitola stručně uvádí některé definice a vztahy z teorie pravděpodobnosti a matematické statistiky, které jsou dále využívány. Hlavními prameny ke zpracování této části byly [3], [4], [5] a [6].

Definice 1.4.1 *Nechť A a B značí náhodné jevy náležící do σ-algebry množiny elementárních jevů Ω a nechť $P(A)$ a $P(B) > 0$ značí pravděpodobnost, že při realizaci náhodného pokusu tyto jevy nastanou. Potom podmíněnou pravděpodobností označíme hodnotu $P(A|B)$ a platí pro ni vztah*

$$P(A|B) = \frac{P(A,B)}{P(B)},$$

kde $P(A,B)$ značí pravděpodobnost, že jevy A a B nastanou současně.

V případě, kdy $P(B) = 0$, nemá podmíněná pravděpodobnost smysl. Z definice podmíněné pravděpodobnosti vychází velmi známá Bayesova věta o podmíněné pravděpodobnosti, kterou lze pomocí pravděpodobností pro systém nejvýše spočetného počtu jevů zapsat v následujícím znění.

Věta 1.4.1 *Nechť A a B_k jsou náhodné jevy a nechť $\{B_i\}$ je úplná soustava vzájemně disjunktních jevů s $P(B_i) > 0$ pro všechna i. Potom pro pravděpodobnost jevu B_k za podmínky jevu A platí vztah*

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_i P(A|B_i)P(B_i)}.$$

Bayesovu větu lze psát také pro vícerozměrné náhodné vektory spojitých náhodných veličin (podrobnosti a důkaz lze nalézt např. v [5]).

Věta 1.4.2 *Nechť $f(\boldsymbol{y})$ je marginální hustota náhodného vektoru $\boldsymbol{Y}$ a $f(\boldsymbol{z}|\boldsymbol{y})$ podmíněná hustota vektoru $\boldsymbol{Z}$ při daném $\boldsymbol{Y}$, potom podmíněná hustota $f(\boldsymbol{y}|\boldsymbol{z})$ je rovna*

$$f(\boldsymbol{y}|\boldsymbol{z}) = \frac{f(\boldsymbol{y})f(\boldsymbol{z}|\boldsymbol{y})}{\int_D f(\boldsymbol{y})f(\boldsymbol{z}|\boldsymbol{y})d\lambda(\boldsymbol{y})},$$

kde λ je čítací míra. a $\int_D f(\boldsymbol{y})f(\boldsymbol{z}|\boldsymbol{y})d\lambda(\boldsymbol{y}) \neq 0$.

Funkce $f(\boldsymbol{y})$ bývá nazývána apriorní hustota, zatímco podmíněná hustota $f(\boldsymbol{y}|\boldsymbol{z})$ se v bayesovských přístupech nazývá aposteriorní. Bayesovy věty se dále bude využívat ve zjednodušené formě $f(\boldsymbol{y}|\boldsymbol{z}) = \frac{f(\boldsymbol{z}|\boldsymbol{y})f(\boldsymbol{y})}{f(\boldsymbol{z})}$.

V následujících podkapitolách je o apriorních a aposteriorních hustotách jednáno v souvislosti se *systémem konjugovaných hustot*. Definici tohoto systému zde není třeba uvádět (lze ji nalézt například v [6]). Zjednodušeně řečeno, konjugovanými rozděleními nazveme takovou dvojici rozdělení, pro kterou apriorní i aposteriorní rozdělení jsou ze stejné rodiny pravděpodobnostních rozdělení.

Protože se dále v práci podrobněji pracuje s multinomickým rozdělením, je zde uvedeno jeho apriorní konjugované rozdělení, a sice Dirichletovo rozdělení pravděpodobnosti.

Definice 1.4.2 *Dirichletovým rozdělením pravděpodobnosti s kladnými parametry $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ (zn. Dir($\boldsymbol{\alpha}$)), nazveme rozdělení náhodného vektoru $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ s $k \geq 2$, jestliže pro jeho hustotu pravděpodobnosti platí*

$$f(x_1, x_2, \ldots, x_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k} x_i^{\alpha_i - 1},$$

kde $0 \leq x_i \leq 1$ a $\sum_{i=1}^{k} x_i = 1$.

Zde použitou funkci $B(\boldsymbol{x})$ lze chápat jako zobecnění Eulerovy beta funkce pro více proměnných. Funkce $\Gamma(x)$ a $B(a, b)$ bývají někdy označované jako Eulerovy integrály[3]. Speciálním případem Dirichletova rozdělení je rozdělení s $k = 2$, které je známé jako *beta rozdělení*.

Definice 1.4.3 *Beta rozdělením pravděpodobnosti s kladnými parametry a, b (zn. Be(a, b)), nazveme rozdělení náhodné veličiny X, jestliže pro její hustotu pravděpodobnosti platí*

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1},$$

kde $0 \leq x \leq 1$.

Na Dirichletovo rozdělení je možné tedy nahlížet jako na zobecněné beta rozdělení [7].

[3]Definice těchto integrálů lze najít v příloze A.2.

Definice 1.4.4 *Momentová vytvořující funkce $m_X(z)$ diskrétní náhodné veličiny X je definována jako střední hodnota funkce e^{zX}, tj.*

$$m_X(z) = E\left(e^{zX}\right) = \sum_x e^{zx} P(x),$$

kde z je pomocná proměnná.

Momentová vytvořující funkce pomáhá při výpočtu momentů, kdy výpočty z definice mohou být poměrně pracné. Jak je uvedeno v [3], postupnými derivacemi $m_X(z)$ podle z dostáváme

$$m'_X(z) = E\left(Xe^{zX}\right),$$
$$m''_X(z) = E\left(X^2e^{zX}\right),$$

a pro k-tou derivaci získáváme obecný vztah

$$m_X^{(k)}(z) = E\left(X^k e^{zX}\right). \tag{1.4.1}$$

Speciálně pro $z = 0$ platí

$$m_X^{(k)}(0) = E\left(X^k\right) = \mu'_k(X), \tag{1.4.2}$$

kde $\mu'_k(X)$ značí k-tý obecný moment.

Analogicky lze zadefinovat momentovou vytvořující funkci k-rozměrné diskrétní náhodné veličiny.

Definice 1.4.5 *Momentová vytvořující funkce $m_{\boldsymbol{X}}(z_1, z_2, \ldots, z_k)$ k-rozměrné diskrétní náhodné veličiny $\boldsymbol{X}$ je definována jako střední hodnota funkce $e^{\boldsymbol{z^T X}}$, tj.*

$$m_X(z_1, z_2, \ldots, z_k) = E\left(e^{\boldsymbol{z^T X}}\right) = E\left(e^{\sum_{i=1}^k z_i X_i}\right),$$

kde $\boldsymbol{X}$ a $\boldsymbol{z}$ jsou odpovídající sloupcové vektory.

Pro momentovou vytvořující funkci vícerozměrné diskrétní náhodné veličiny platí podobné vztahy jako pro jednorozměrné veličiny. Pro obecný moment r-tého řádu i-té složky či odpovídající dvojici náhodných veličin platí

$$E\left(X_i^r\right) = \frac{\partial^r m_{\boldsymbol{X}}(z_1, z_2, \ldots, z_k)}{\partial z_i^r}\Big|_{z_1, z_2, \ldots, z_k = 0}, \tag{1.4.3}$$

$$E\left(X_i, X_j\right) = \frac{\partial^2 m_{\boldsymbol{X}}(z_1, z_2, \ldots, z_k)}{\partial z_i \partial z_j}\Big|_{z_1, z_2, \ldots, z_k = 0}. \tag{1.4.4}$$

1.5 Multinomické rozdělení

V případě „1-0-2“ sázení, které bylo popsáno hned v úvodu této práce, se výsledkem může stát právě jedna ze tří možností. Je-li pro nás daný zápas náhodnou událostí, existují pravděpodobnosti, že jeden z těchto tří jevů nastane, přičemž nějaký nutně nastat musí[4]. Tuto situaci, jak je uvedeno dále, přesně popisuje trinomické rozdělení pravděpodobnosti.

Trinomické rozdělení je speciálním případem rozdělení multinomického, avšak namísto obecně k pracuje pouze se třemi možnými jevy. V této podkapitole je snaha uvést vztahy platné pro obecné multinomické rozdělení. V odstavci odhadů parametrů se však přejde konkrétně k trinomickému rozdělení, které je klíčové z hlediska praktické náplně této práce.

Definice 1.5.1 *Uvažujme sérii n nezávislých pokusů, kdy v každém pokusu musí nastat jeden z k disjunktních jevů $A_1, A_2, \ldots, A_k$ a pravděpodobnost jevu A_i je rovna p_i $\forall i$. Nechť $X_1, X_2, \ldots, X_k$ jsou náhodné veličiny označující počty, kolikrát nastaly jevy $A_1, A_2, \ldots, A_k$, potom pro sdruženou pravděpodobnostní funkci multinomického rozdělení $Mu(n, p_1, p_2, \ldots, p_k)$ platí:*

$$P\left[\bigcap_{i=1}^{k}(X_i = n_i)\right] = n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} = P(n_1, n_2, \ldots, n_k),$$

kde $\forall\ n_i \geq 0$, $X_i \geq 0$ a $p_i \geq 0$ platí $\sum n_i = n$, $\sum p_i = 1$. □

Volně řečeno, multinomické rozdělení je tedy možno chápat jako rozdělení „součtu“ n vzájemně nezávislých náhodných veličin, které se řídí obecně diskrétním rozdělením na množině vzájemně disjunktních jevů $\{A_1, A_2, \ldots, A_k\}$, kdy i-tá složka nabývá hodnoty 1, právě když nastane jev A_i.

Uvažujme nyní náhodný vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$, který se řídí multinomickým zákonem rozdělení pravděpodobnosti. Při určování jednorozměrného marginálního rozdělení pravděpodobnosti pro X_i lze uvažovat nezávislé opakování náhodné veličiny X_i^*, pro kterou platí

$$X_i^* = \begin{cases} 1, & \text{s } p_i \text{ nastane jev } A_i, \\ 0, & \text{s } 1 - p_i \text{ nenastane jev } A_i. \end{cases}$$

To je ovšem opakování alternativního rozdělení, jehož součtem je rozdělení binomické. Zřejmě tedy platí následující věta.

Věta 1.5.1 *Nechť se vektor náhodných veličin $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ řídí multinomickým rozdělením $Mu(n, p_1, p_2, \ldots, p_k)$. Potom pro každou složku tohoto vektoru platí*

$$X_i \sim Bi(n, p_i).$$

[4] Zde se neřeší problematika odložených, nedohraných nebo kontumovaných utkání. V těchto případech se většinou uzavřené sázky ruší a vklady se vracejí sázkařům.

Obecně platí, že všechna dílčích rozdělení multinomického rozdělení jsou opět multinomická. Konkrétně v tomto případě se jedná o multinomické rozdělení s $k = 2$, což je však známé jako rozdělení binomické. Důkaz lze nalézt například v [5].

Momentovou vytvořující funkci multinomického rozdělení lze spočítat podle definice 1.4.5 a zapast ve tvaru

$$m_x(z_1, z_2, \ldots, z_k) = (p_1 e^{z_1} + p_2 e^{z_2} + \ldots + p_k e^{z_k})^n. \tag{1.5.1}$$

Využitím vztahů (1.4.3) a (1.5.1) a dosazením do příslušných derivací lze najít předpis pro střední hodnotu a rozptyl i-té složky.

$$E(X_i) = np_i, \tag{1.5.2}$$
$$D(X_i) = np_i(1 - p_i). \tag{1.5.3}$$

To odpovídá zmíněné skutečnosti, že každá i-tá složka multinomického rozdělení se řídí rozdělením $Bi(n, p_i)$. Jednotlivé složky vektoru však nejsou nezávislé a s využitím vztahu (1.4.4) a definičního vztahu pro kovarianci lze spočíst, že kovariance mezi složkami vektoru $\boldsymbol{X}$ je rovna

$$\mathrm{cov}(X_i, X_j) = -np_ip_j. \tag{1.5.4}$$

Je zřejmé, že kovariance je vždy záporná, a tudíž je záporný i korelační koeficient. Za předpokladu pevného a daného rozsahu náhodného výběru pak tento fakt odpovídá skutečnosti, že pokud jev A_i pozorujeme vícekrát, vyskytují se jevy A_j méně často. Jednoduše řečeno, četnosti pozorování jednotlivých možností jdou „proti sobě“.

1.5.1 Bodové odhady parametrů

Problematika odhadu parametrů multinomického rozdělení je značně široká a v česky psané literatuře ne příliš často zpracovávána. V tomto odstavci jsou představeny základní postupy bodových odhadů parametrů multinomického rozdělení. Problematice intervalových odhadů je věnován následující odstavec.

Máme-li k dispozici výběr z multinomického rozdělení velikosti n a známe-li[5] počet tříd k, jsou pro nás neznámými parametry pravděpodobnosti $\{p_1, p_2 \ldots, p_k\}$. Vzhledem k podmínce $\sum_i p_i = 1$ je počet neznámých parametrů ve skutečnosti roven $k - 1$.

Jak je uvedeno například v [8], jsou-li $n_1, n_2, \ldots, n_k$ pozorované četnosti příslušných jevů, pak maximálně věrohodnými odhady pravděpodobností p_i jsou relativní frekvence

$$\hat{p_i} = \frac{n_i}{n}, \qquad \forall i. \tag{1.5.5}$$

[5]Zde je důležitou podmínkou informace o počtu tříd multinomického rozdělení, neboť v obecném případě nemusí být počet jevů k známý. To by vedlo k dalším teoretickým problémům.

Tento odhad je odhadem nestranným a vydatným a z (1.5.2) a (1.5.3) plyne

$$E(\hat{p_i}) = p_i, \tag{1.5.6}$$
$$D(\hat{p_i}) = p_i(1 - p_i)/n. \tag{1.5.7}$$

Nevýhodou tohoto odhadu je, že pokud $n_i = 0$, potom $\hat{p_i} = 0$. V praktické části práce by toto mohlo být nevýhodou. Nulová četnost může být způsobena „malým“ počtem pozorování[6]. Tento problém je možné řešit například odhadem pomocí tzv. bayesovského přístupu.

Bayesovský přístup: Základní myšlenkou bayesovského odhadu je připuštění náhodnosti odhadovaných parametrů pravděpodobnostních rozdělení. Tato náhodnost je pak formována apriorní informací o hledaném parametru bez ohledu na pozorované hodnoty. Tato informace bývá vyjadřována pomocí pravděpodobnostního rozdělení. Volba apriorního rozdělení je jedním ze základních problémů tohoto přístupu a více se uvádí například v [6].

V definici 1.5.1 je značením $P(n_1, n_2, \ldots, n_k)$ vyjádřena pravděpodobnost, že veličiny X_i nabudou hodnot n_i. To ovšem platí v případě, že existují jednoznačné (byť neznámé) pravděpodobnosti p_i. Tuto skutečnost budeme v tomto případě značit dále $P(n_1, n_2, \ldots, n_k|p_1, p_2, \ldots, p_k)$.

Dále tedy předpokládejme, že p_i jsou nyní náhodné veličiny a řídí se nějakým rozdělením pravděpodobnosti. Apriorně bude předpokládáno, že p_i jsou rovnoměrně rozloženy na příslušné množině. Jinak řečeno, apriorní hustota rozdělení p_i je konstantní. Protože musí platit podmínka $p_i \geq 0$ a $\sum_i p_i = 1$ pro všechna i, vychází apriorní hustota rovna $\Gamma(k)$.

Jelikož apriorní informace nemusí být popsána skutečnou hustotou pravděpodobnosti, bude dále uvažováno apriorní rozdělení p_i ve tvaru

$$f(p_1, p_2, \ldots, p_k) = 1.$$

Pro sdruženou pravděpodobnost dle Bayesovy věty dále platí

$$P(n_1, n_2, \ldots, n_k, p_1, p_2, \ldots, p_k) = n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} = \frac{\Gamma(n+1)}{\prod_{i=1}^{k} \Gamma(n_i + 1)} \prod_{i=1}^{k} p_i^{n_i}. \tag{1.5.8}$$

Integrováním přes všechny pravděpodobnosti p_i lze s využitím vztahu pro výpočet Dirichletova integrálu[7] spočíst, že platí

$$P(n_1, n_2, \ldots, n_k) = \frac{\Gamma(n+1)}{\Gamma(n+k)}. \tag{1.5.9}$$

Z předchozích rovnic a opět s využitím Bayesovy věty o podmíněné pravděpodobnosti dostáváme

[6] Problém nízkých frekvencí je obecně znám jako „Zero-frequency problem“.
[7] Definice zde využitého integrálu je uvedena v příloze A.2.

$$f(p_1, p_2, \ldots, p_k | n_1, n_2, \ldots, n_k) = \frac{P(n_1, n_2, \ldots, n_k, p_1, p_2, \ldots, p_k)}{P(n_1, n_2, \ldots, n_k)},$$
$$= \frac{\Gamma(n+k)}{\prod_{i=1}^{k} \Gamma(n_i + 1)} \prod_{i=1}^{k} p_i^{n_i}. \tag{1.5.10}$$

To znamená, že

$$f(p_1, p_2, \ldots, p_k | n_1, n_2, \ldots, n_k) \sim \mathrm{Dir}(n_1 + 1, n_2 + 1, \ldots, n_k + 1).$$

Předpis (1.5.10) je tedy pravděpodobnostní funkcí popisující rozložení parametrů na daném definičním oboru za předpokladu, že jsou pozorvány hodnoty n_i a apriorně bylo předpokládáno rovnoměrné rozložení p_i.

Lze spočíst, že za zmíněných předpokladů pro bodový odhad (ozn. $\hat{p_i}^b$) pak platí

$$\hat{p_i}^b = E(p_i | n_1, n_2, \ldots, n_k) = \frac{n_i + 1}{n + k}, \tag{1.5.11}$$
$$\sigma(\hat{p_i}^b)^2 = \sigma(p_i | n_1, n_2, \ldots, n_k)^2 = \frac{(n_i + 1)(n - n_r + k - 1)}{(n + k)^2 (n + +k + 1)}. \tag{1.5.12}$$

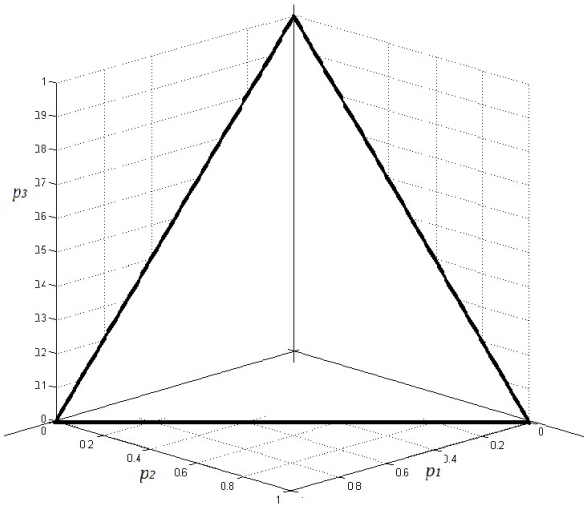
Protože v tomto případě je rozdělení pravděpodobností p_i známo, lze počítat intervalové odhady pro parametry p_i, čemuž se věnuje následující odstavec.

1.5.2 Intervalové odhady parametrů trinomického rozdělení - obecně

Tato část odhadů navazuje na předchozí bodové odhady a pokouší se najít vyjádření intervalových odhadů pro p_i. Protože je v praktické části využíváno konkrétně trinomického rozdělení, přejdeme nyní od obecně multinomického ke $k = 3$. Obecně se pod pojmem intervalový odhad chápe interval, ve kterém s určitou pravděpodobností leží reálná hodnota parametru. V rámci této práce bude často užíváno pojmů konfidenční oblasti či oblasti nebo množiny spolehlivosti jako zobecnění jednorozměrného intervalu do více dimenzí. Pro práci dále není stěžejní teoretická problematika spojená s rozdíly mezi konfidenčními a věrohodnostními množinami.

Je zřejmé, že parametrickým prostorem Θ trinomického rozdělení je část prostoru R^3, konkrétně $[0, 1]^3$ s podmínkou $p_1 + p_2 + p_3 = 1$. To je obecně trojúhelník znázorněný na obrázku 1.5.1.

Oblastí spolehlivosti se tak již nestává pouze interval, ale podoblast tohoto trojúhelníku. Smyslem intervalových odhadů je pak nalézt takovou oblast, která s předem danou $100(1 - \alpha)\%$ pravděpodobností pokrývá skutečnou hodnotu parametru. Je zřejmé, že oblast lze definovat libovolným průmětem do jedné ze tří základních rovin, neboť jakákoliv dvojice parametrů $p_i \times p_j$ jednoznačně určuje hodnotu zbývajícího parametru. Také je zjevné, že oblastí spolehlivosti může



Obrázek 1.5.1: Parametricky prostor $p1 \times p2 \times p3$.

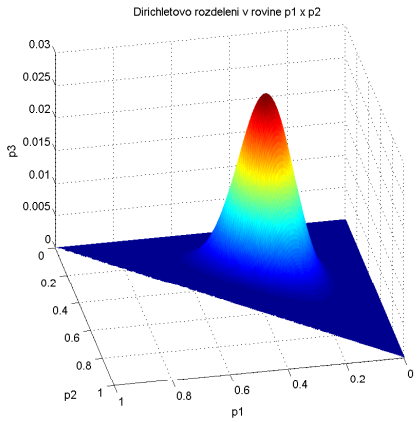
býti libovolný geometrický rovinný útvar (mnohoúhelník, kružnice, elipsa,...). Při určování oblasti spolehlivosti bude snaha nalézt takovou oblast, jejíž míra (obsah) je co nejmenší při zachování hladiny α.

V souvislosti s praktickou částí práce je odhadování parametrů p_i založeno na reálně pozorovaných hodnotách. Nahlížíme-li na sportovní utkání jako na náhodnou událost, je největším problémem při tomto odhadování parametrů fakt, že nelze opakovat utkání s pravděpodobnostmi p_i za stejných podmínek.

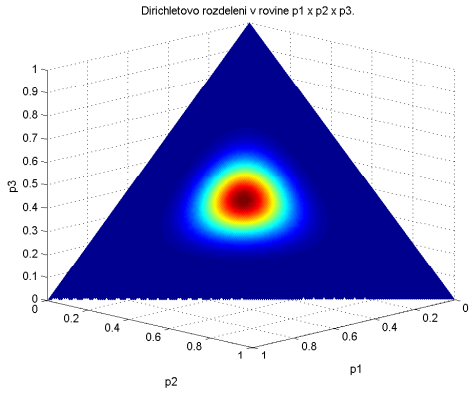
Jedním z úkolů praktické části je nalézt vhodné kritérium či pravidlo, podle kterého lze kategorizovat zápasy tak, aby pravděpodobnosti v dané kategorii byly vždy stejné. Na základě takto vytvořených kategorií se získají pozorování (náhodný výběr), ze kterých lze odhadovat neznámé parametry p_i.

Jak bylo uvedeno v předchozím odstavci (rovnice (1.5.10)), tak za předpokladu naměřených četností n_i se sdružené rozdělení p_i řídí Dirichletovým rozdělením. Protože trojici pravděpodobností $\{p_1, p_2, p_3\}$ lze jednoznačně popsat dvojicí $\{p_1, p_2\}$, jsou grafické obrázky dále zpracovávány převážně v rovině $p_1 \times p_2$. Ukázka hustoty Dirichletova rozdělení na parametrickém prostoru a v rovině $p_1 \times p_2$ je zobrazena na následujících obrázcích 1.5.2a a 1.5.2b.

Díky znalosti rozložení pravděpodobnosti nad parametrickým prostorem, lze vytvářet oblasti spolehlivosti, které udávají, s jakou pravděpodobností pokrývají námi hledaný bodový parametr. Při hledání takových oblastí lze postupovat mnoha způsoby, přičemž dva z nich jsou popsány dále. Konkrétně je představen numerický postup, který numerickými aproximacemi hledá „přesnou“ oblast spolehlivosti. Dále je představen analytický postup, založený na „transformaci proměnných“, jehož výstupem je aproximace té přesné oblasti.



(a) Hustota Dirichletova rozdělení v rovině $p1 \times p2$.



(b) Rozložení hustoty Dirichletova rozložení na simplexovém trojúhelníku.

Obrázek 1.5.2: Dirichletovo rozdělení pravděpodobnosti.

1.5.3 Numerické hledání konfidenčních oblastí

Nejzákladnější metodou hledání konfidenční oblasti (při bayesovském přístupu) pro neznámý parametr trinomického rozdělení je numerický postup. Základní myšlenkou je stále nalézt takovou oblast v simplexovém trojúhelníku, která daný parametr (dle přístupu i jeho odhad) pokryje s $100(1-\alpha)\%$ pravděpodobností.

Podobně, jako je tomu u jednorozměrných parametrů, lze spolehlivost oblasti spočíst integrálem z hustoty pravděpodobnostního rozdělení parametrů nad danou oblastí. Jinak zapsáno chceme, aby platilo

$$\iint_D f(p_1, p_2)\, dp_1\, dp_2 \geq 1 - \alpha, \tag{1.5.13}$$

kde D značí vybranou konfidenční oblast definovanou v rovině $p_1 \times p_2$. V případě numerického postupu lze díky znalosti hustoty $f(p_1, p_2)$ takový integrál aproximovat s takřka libovolnou přesností[8].

Problém konfidenční oblasti se v tomto případě tedy přesouvá na nalezení optimální (s nejmenší mírou, s nejmenším obsahem) oblasti D, která pro zadané α nerovnost (1.5.13) splňuje. Je zřejmé, že tuto nerovnost splňuje více než jedna oblast, a proto je zde snaha vybrat tu nejmenší. O způsobu, jak takovou oblast vybrat, napovídá následující tvrzení.

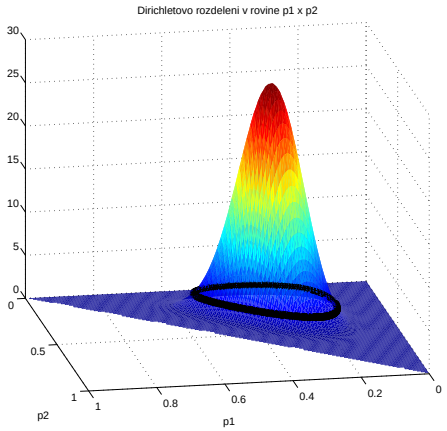
Tvrzení 1.5.1 *Nechť $f(p_1, p_2)$ je funkcí hustoty Dirichletova rozdělení v rovině $p_1 \times p_2$ a nechť $\alpha \in (0, 1)$ je reálné číslo, potom pro oblast D roviny $p_1 \times p_2$ nejmenšího obsahu pro kterou platí nerovnost* (1.5.13)*, existuje číslo $c \in R$ tak, že platí*

$$D = \{(p_1, p_2) : f(p_1, p_2) \geq c\}.$$

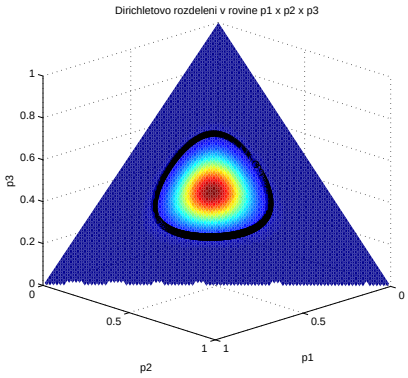
[8] V praktické části je tento integrál počítán obdobou obdélníkového pravidla pro aproximaci jednorozměrných integrálů.

Tvrzení je převzato z [6]. Tvrzení tedy říká, že konfidenční množinu nejmenšího obsahu s předem zadanou spolehlivostí lze určit vrstevnicí hustoty Dirichletova rozdělení, kde hodnota vrstevnice určuje koeficient spolehlivosti. Alternativně je možné postupovat při maximalizaci věrohodnosti množiny s předem zadanou mírou této množiny.

V případě numerického řešení konfidenční oblasti lze c hledat například metodou půlení intervalu, dokud nebude požadované spolehlivosti dosaženo. Ukázka výstupu takového numerického algoritmu je zobrazena na obrázcích 1.5.3a a 1.5.3b, kde je ve dvou variantách zobrazena oblast spolehlivosti.



(a) Konfidenční oblast na hustotě Dirichletova rozdělení v rovině $\{p_1 \times p_2\}$.



(b) Konfidenční oblast na simplexovém trojúhelníku.

Obrázek 1.5.3: Černými značkami je zobrazena 95% konfidenční oblast.

1.5.4 Konfidenční oblasti na základě transformace proměnných

Jak bylo dříve uvedeno v rovnici (1.5.10), rozdělení námi hledaných parametrů je Dirichletovo s danými parametry. Pro toto rozdělení však také platí, pokud $f(p_1, \ldots, p_k | n_1, \ldots, n_k) \sim \mathrm{Dir}(n_1 + 1, \ldots, n_k + 1)$ a $h < k$, potom

$$f(p_1, \ldots p_h, (p_{h+1}, \ldots, p_k)) \sim \mathrm{Dir}(n_1 + 1, \ldots, n_h + 1, (n_{h+1} + 1, \ldots + n_k + 1).$$

Na základě této vlastnosti lze dále odvodit (převzato z [7]), že podíly $p_1, \frac{p_2}{1-p_1}, \frac{p_3}{1-p_1-p_2}, \ldots, \frac{p_{k-1}}{1-p_1-p_2-\ldots-p_{k-2}}$ jsou vzájemně nezávislé náhodné veličiny a platí pro ně

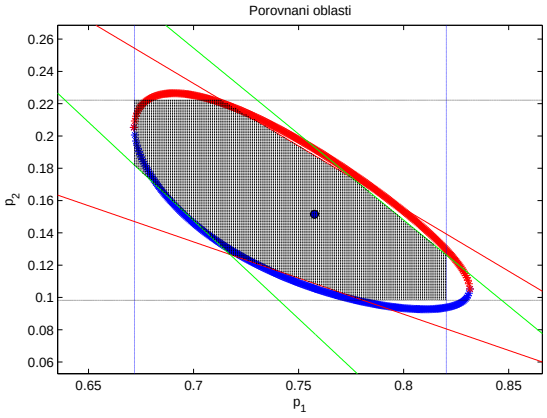
$$\frac{p_j}{1 - \sum_{r=1}^{j-1} p_r} \sim \mathrm{Be}(n_j + 1, \sum_{r=j+1}^{k} (n_r + 1)), \tag{1.5.14}$$

pro $j = 1, \ldots, k-1$ a kde $\mathrm{Be}(\cdot, \cdot)$ značí beta rozdělení.

Díky známým rozdělením těchto nezávislých veličin je možné např. u trinomického rozdělení se dvěma parametry zkonstruovat intervalové odhady p_1 a $\frac{p_2}{1-p_1}$ se zadanou spolehlivostí. Intervalový odhad pro parametrickou funkci podílu lze však psát jako funkci jedné proměnné p_1, a tak lze díky nezávislosti získat průnikem oblast se spolehlivostí $(1-\alpha_1)(1-\alpha_2)$. V rovině $p_1 \times p_2$ budou tedy nalezeny meze ve tvaru $\{d_1 < p_1 < h_1\}$ a $\{d_2 < \frac{p_2}{1-p_1} < h_2\}$. Jedná se tak o čtyři přímky vytvářející čtyřúhelník.

Analogicky je možné postupovat s parametry p_1 a p_2 v opačném pořadí a najít intervaly spolehlivosti pro proměnné p_2 a $\frac{p_1}{1-p_2}$. Výsledkem tohoto a předchozího postupu jsou dvě různé konfidenční množiny (dva čtyřúhelníky), z nichž každá má předem danou spolehlivost $(1-\alpha_1)(1-\alpha_2)$. Pro průnik těchto množin lze však spolehlivost pouze odhadnout, a to zdola například Bonferroniho[9] nerovností. Spolehlivost takto zkonstruované oblasti je pak tedy vyšší než $2(1-\alpha_1)(1-\alpha_2)-1$.

Ukázka výsledku takového postupu v porovnání s „optimální“ konfidenční množinou je zobrazena na obrázku 1.5.4. Numericky spočtená „přesná“ množina připomíná svým tvarem elipsu, zatímco aproximace pomocí „nezávislých podílů“ je zobrazena vybarveným osmiúhelníkem.



Obrázek 1.5.4: „Optimální“ konfidenční oblast s 95% spolehlivostí a „aproximační“ oblast se spolehlivostí alespoň 90 % odvozena ze dvou 95% oblastí.

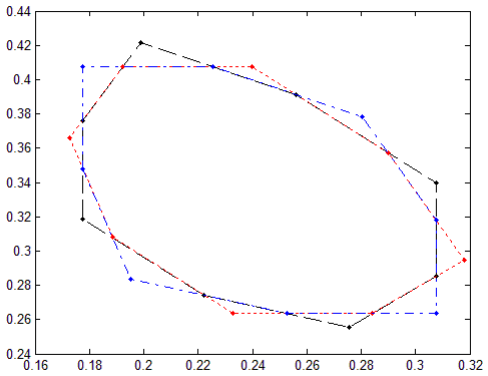
Algoritmus hledání konfidenční oblasti spočívá v tomto případě nejen ve spočtení horních a dolních mezí, které zde jsou reprezentovány přímkami, ale také v následném spočtení průsečíků těchto přímek.

Problém výběru takových bodů, které vytvoří vnitřní oblast, je zde značně usnadněn, neboť stačí procházet jednotlivé přímky a odebírat ty průsečíky, které leží v nesprávné polorovině. Po tomto postupu tedy zbývá osm[10] průsečíků, které vytváří vnitřní oblast spolehlivosti.

[9] Znění této nerovnosti lze nalézt v příloze A.3.

[10] Tato práce nezkoumá, zda zbývajících osm bodů je pravidlem. Během zpracování však nebyl nalezen případ, kdy by byl výsledkem tohoto postupu jiný počet.

Jednoznačnost konfidenční množiny. Množina vytvořená tímto postupem však není jednoznačná a závisí na dvojici pozorování, která je vybrána. Celkem lze tedy získat tři různé osmiúhelníky se stejnou zdola omezenou spolehlivostí. Náčrt obrázku, jak takové tři množiny mohou vypadat, je na obrázku 1.5.5. Dále v této práci již nebudou zkoumány vlastnosti těchto oblastí. Vždy bude vybrána jedna oblast dle předcházejícího postupu s pozorováním n_1 a n_2.



Obrázek 1.5.5: Náčrt tří různých zčásti se překrývajících konfidenčních oblastí ze stejných hodnot pozorování.

1.5.5 Porovnání konfidenčních oblastí

Tato část práce shrne dva předchozí přístupy v hledání konfidenčních oblastí. Numerický přístup zde bude chápán jako „přesný“. Zdůvodnění tohoto chápání je opřeno o numericky libovolně nastavitelnou přesnost při hledání konfidenční množiny a současně o numericky libovolnou přesnost při vyčíslování integrálů.

V praxi tento přístup není možný, neboť numerické řešitele jsou výpočetně náročné a v prakticky proveditelném čase lze nastavit pouze omezenou přesnost. V rámci zpracování této práce byl krok při hledání konfidenční množiny a integrování nastaven na hodnotu 5×10^{-4}. Spolehlivostní množina byla hledána se spolehlivostí 95 % $\pm$ 1 %. Dále se tato varianta nazývá čistě jako „numerická“.

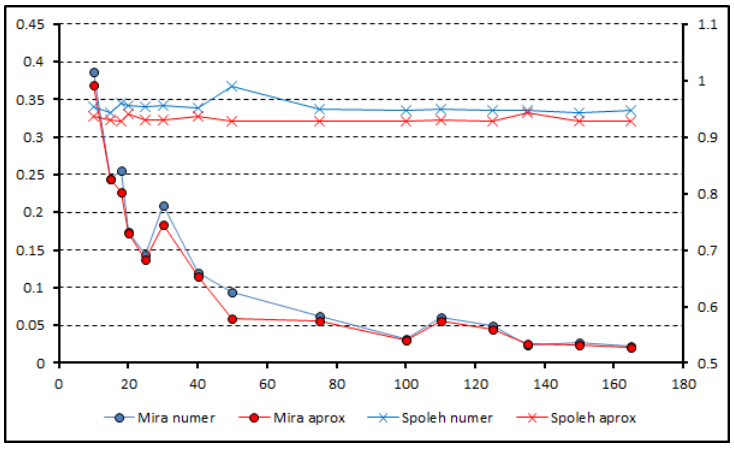
Druhým přístupem je hledání konfidenčních oblastí na základě analytického odvození nezávislých podílů (viz subsekce 1.5.4). Tato metoda dokáže najít pomocí osmi přímek oblast se spolehlivostí alespoň $(1 - \alpha)$ %. Skutečná, vyšší než garantovaná, spolehlivost však není známa, neboť se jedná o oblast vytvořenou průnikem dvou oblastí. Pro toto porovnání byla požadována spolehlivost alespoň 90 %, přičemž se jedná o průnik dvou 95% oblastí. Výsledná oblast je dále nazývána aproximační oblastí.

Pro porovnání bylo libovolně zvoleno 15 různých kombinací, pro které jsou spočteny numerické konfidenční oblasti a aproximační „přímkové“ oblasti. Pro obě oblasti jsou pak numericky spočteny jejich spolehlivosti a plošné míry v rovině $p_1 \times p_2$ (program DP_conf02.m). Plošné míry jsou pak vyjádřeny procentuálním poměrem vzhledem k celému parametrickému prostoru. Získané hodnoty jsou uvedeny v tabulce 1.5.1.

ID	Počty pozorování				Konfidenční oblast				Počet iterací
					Numerická		Aproximační		
	n	n_1	n_2	n_3					
1	10	2	2	6	95.4%	38.6%	93.7%	36.8%	4
2	15	2	10	3	94.3%	24.4%	93.0%	24.4%	4
3	18	9	7	2	96.0%	25.6%	92.9%	22.7%	4
4	20	2	3	15	95.6%	17.3%	94.2%	17.3%	3
5	25	2	18	5	95.4%	14.4%	93.1%	13.7%	3
6	30	10	10	10	95.6%	20.9%	93.1%	18.3%	3
7	40	4	16	20	95.2%	12.1%	93.7%	11.5%	3
8	50	5	40	5	98.9%	9.4%	92.8%	6.0%	15 ($\approx max$)
9	75	50	15	10	94.9%	6.2%	92.8%	5.7%	3
10	100	80	9	11	94.7%	3.2%	92.8%	3.0%	3
11	110	40	40	30	94.9%	6.1%	93.0%	5.5%	3
12	125	65	30	30	94.8%	4.9%	92.9%	4.5%	3
13	135	10	20	105	94.7%	2.5%	94.4%	2.6%	3
14	150	110	25	15	94.2%	2.8%	92.8%	2.4%	3
15	165	125	25	15	94.7%	2.2%	92.9%	2.1%	3

Tabulka 1.5.1: Tabulka s výstupem z porovnání oblastí.

Z naměřených dat je možné zjistit, že s výjimkou jednoho pozorování, kdy algoritmus numerického hledání oblasti nedošel ke spolehlivosti 95 % ± 1 %, byla spolehlivost získaná druhou variantou vždy alespoň 92.8%, přičemž v případě množiny s nižší spolehlivostí je až na jedno pozorování vždy míra této množiny nižší. Vykreslení naměřených hodnot je zobrazeno na grafu 1.5.6.



Obrázek 1.5.6: Porovnání numerické a aproximační oblasti.

Ze získaných výsledků lze tedy podloženě pracovat s domněnkou, že aproximační oblast získanou na základě transformace proměnných lze opravdu považovat za jakousi aproximaci skutečné oblasti spolehlivosti. Z výsledků plyne, že nastavení hranice spolehlivosti na alespoň 90 % přináší v průměru 93% oblast spolehlivosti, která je průměrně o 0.95 procentních bodů menší než spočtená numerická oblast.

Vzhledem k nestabilitě numerických řešitelů a výpočetní náročnosti hledání „přesných“ oblastí se bude dále pracovat s aproximační variantou, která také usnadní praktickou část této práce, kde je hledání konfidenčních oblastí důležitou součástí rozhodovacích postupů.

1.5.6 Testování parametrů typu $p_i > p_j$

Tento odstavec věnovaný testování parametrů je zpracován pro multinomické rozdělení $Mu(n, p_i, p_j, p_z)$ se třemi parametry (pro trinomické rozdělení). Uveďme, že rozhodnutí, že $p_i > p_j$, bude přijato se spolehlivostí $1 - \alpha$, pokud $P(p_i > p_j | n_i, n_j, n_z) > 1 - \alpha$. Nutno zde dodat, že se ve skutečnosti jedná pouze o rozhodovací pravidlo, podle kterého se bude rozhodovat. Toto pravidlo je založené na pravděpodobnosti $p_i > p_j$. Dále se o tomto testování bude mluvit jako o testování hypotézy a jejím přijetí (zde stejné jako nezamítnutí), či zamítnutí. V tomto odstavci není řešena teoretická problematika „bayesovského pojetí“ testování hypotéz.

Již v odstavci s bodovými odhady multinomického rozdělení bylo uvedeno, že při bayesovském pojetí mají parametry multinomického rozdělení za předpokladu apriorního rovnoměrného rozdělení rozdělení $\mathrm{Dir}(n_i+1, n_j+1, n_z+1)$ (viz rovnice (1.5.10)).

Mějme tedy neznámý parametr $\boldsymbol{p}$ popsán hustotou Dirichletova rozdělení na parametrickém prostoru trinomického rozdělení (obrázek 1.5.3b). Na základě pozorování n_i, n_j, n_z bude rozhodováno, že $p_i > p_j$, pokud bude splněna následující nerovnost.

$$\iiint\limits_{\substack{p_i+p_j+p_z=1 \\ 0<p_i,p_j,p_z<1 \\ p_i>p_j}} f(p_i, p_j, p_z | n_i, n_j, n_z)\, dp_i\, dp_j\, dp_z \geq 1 - \alpha. \tag{1.5.15}$$

„Hypotéza“ je tedy přijata, pokud „obsah“ (objem) pod grafem hustoty na té části definičního oboru, kde $p_i > p_j$, je větší než $1 - \alpha$.

Jednoduchou úpravou s využitím (1.5.10) lze zde uvedený trojný integrál nahradit dvojným integrálem. Rozhodovací pravidlo pak vypadá následovně[11].

$$\iint\limits_{\substack{1-p_i-p_j>0 \\ 0<p_i,p_j,p_z<1 \\ p_i>p_j}} f(p_i, p_j, 1 - p_i - p_j | n_i, n_j, n_z)\, dp_i\, dp_j \quad \geq 1 - \alpha, \tag{1.5.16}$$

$$\iint\limits_{\substack{1-p_i-p_j>0 \\ 0<p_i,p_j,p_z<1 \\ p_i>p_j}} \frac{\Gamma(n+3)}{\prod\limits_{k\in\{i,j,z\}} \Gamma(n_k+1)} p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{n_z}\, dp_i\, dp_j \quad \geq 1 - \alpha. \tag{1.5.17}$$

[11] Stále za předpokladu, že $\sum_{i=1}^{3} p_i = 1$ a $\sum_{i=1}^{3} n_i = n$.

Funkce uvnitř integrálu (1.5.17) je však pravděpodobnostní funkcí dvourozměrného beta rozdělení[12]. Protože se stále jedná o spolehlivost[13], lze dvojný integrál zapsat jednoduchým (viz [9]), čímž dostáváme nerovnost

$$\int_{1/2}^{1} \frac{1}{\mathrm{Be}(n_i+1, n_j+1)} x^{n_i}(1-x)^{n_j}\,dx \geq 1-\alpha. \tag{1.5.18}$$

Výsledek testu tedy nezávisí na počtu pozorování n_z (resp. n_3). Snadno lze vidět, že funkce uvnitř získaného integrálu je pravděpodobnostní funkcí beta rozdělení (definice 1.4.3) s parametry $n_i+1,\ n_j+1$. Získanou nerovnost (rozhodovací pravidlo) lze tedy přepsat do následující podoby, kde F_{Be} je distribuční funkce beta rozdělení.

$$1 - F_{\mathrm{Be}}(0.5, n_i+1, n_j+1) \geq 1-\alpha, \tag{1.5.19}$$
$$F_{\mathrm{Be}}(0.5, n_i+1, n_j+1) \leq \alpha. \tag{1.5.20}$$

Bude-li dále předpokládáno, že $\alpha < 0.5$, a současně bude rozhodnuto, že $p_i > p_j$, tj. nerovnost (1.5.20), pak s využitím symetrie[14] platí

$$1 - F_{\mathrm{Be}}(0.5, n_j+1, n_i+1) \leq \alpha, \tag{1.5.21}$$
$$F_{\mathrm{Be}}(0.5, n_j+1, n_i+1) \geq 1-\alpha. \tag{1.5.22}$$

Protože bylo předpokládáno, že $\alpha < 0.5$, platí, že

$$F_{\mathrm{Be}}(0.5, n_j+1, n_i+1) \geq \alpha. \tag{1.5.23}$$

Vzhledem k nerovnostem (1.5.20) a (1.5.23) je zjevné, že pro $\alpha \neq 0.5$ nemůže být rozhodnuto o platnosti $p_i > p_j$ a současně $p_j > p_i$. Tento výsledek dává v praxi návod, jak postupovat při testování:

1. Do n_1 dosadíme nejvyšší hodnotu pozorování.
2. Za n_2 dosadíme druhou nejvyšší hodnotu pozorování.
3. Pokud je splněna nerovnost (1.5.20), je učiněno rozhodnutí $p_1 > p_2$.
4. Pokud nerovnost není splněna, nelze prohlásit ani jedno pozorování za nejpravděpodobnější.

Pokud by v tomto postupu byla nejvyšší a druhá nejvyšší hodnota stejná, byla by „hypotéza“ (samozřejmě) zamítnuta. Postupu však tato situace nijak nevadí. S nejnižší hodnotou se zde vůbec nepracuje, neboť v případě nezamítnutí mezi dvěma největšími se předpokládá, že by se nezamítalo ani mezi největší a nejmenší.

[12] Definici tohoto rozdělení dle [9] lze najít v příloze A.2.

[13] Zde je spolehlivostí (ang. reliability) myšleno $P(p_i > p_j)$.

[14] Pro beta rozdělení platí následující symetrie: $F_{\mathrm{Be}}(x, n_i+1, n_j+1) = 1 - F_{\mathrm{Be}}(1-x, n_j+1, n_i+1)$, podrobnosti k tomuto vztahu jsou v příloze A.3.

2 Zpracování historických dat

Nalezení vhodné strategie sázení, která by sázkaři přinášela zisk a mohla být považována za investici, je hlavním tématem celé práce. Hledání takové strategie je založeno na zpracování historických dat. Tato část práce se věnuje právě výběru a základnímu zpracování historických dat, která slouží nejen ke „kalibraci“ vybraných strategií, ale také k jejich následnému ověření.

Pro potvrzení správného nastavení konkrétních modelů sázení se pak ověřování provádí na sadě kontrolních dat, která není součástí podkladových dat pro proces „kalibrace“.

2.1 Popis dat

Pro zpracování praktické části práce jsou využita historická data nejvyšších fotbalových soutěží Česka, Anglie a Německa. Zahraniční země byly vybrány na základě subjektivního vnímání těchto soutěží jako kvalitních tradičních soutěží. Bez újmy na obecnosti by bylo možné vybrat jakékoliv jiné země případně i jiné úrovně soutěží. Jediným kritériem je v tento moment možnost „1-0-2“ sázení na utkání dlouhodobých (sezónních) soutěží.

Potřebnými historickými daty se v této práci rozumí data, kde ke každému odehranému zápasu dané soutěže je k dispozici datum utkání, pořadové číslo kola soutěže, výsledek utkání (ve tvaru *Počet branek domácí : Počet branek hosté*) a vypsané výše kurzů pro všechny tři možnosti výsledku. Ukázka několika takovýchto záznamů je v následující tabulce 2.1.1.

Kolo	Datum	Zápas	Výsledek	Kurz D	Kurz R	Kurz H
28	11.5.2014	Brno - Slavia Praha	2:0	2.49	2.94	2.88
28	11.5.2014	Plzeň - Mladá Boleslav	2:0	1.71	3.68	4.45
28	11.5.2014	Sparta - Olomouc	5:0	1.3	5.13	8.35
28	10.5.2014	Liberec - Teplice	2:1	2.61	3.2	2.57

Tabulka 2.1.1: Ukázka několika záznamů historických dat, kde D = vítězství domácích, R = remíza, H = vítězství hostů.

Vzhledem k veřejné bezplatné dostupnosti dat jsou pro tuto práci k dispozici kurzová data buďto z několika sázkových kanceláří (Anglie a Německo dostupné z [12]), nebo jako průměr více sázkových kanceláří (Česko dostupné z [13]). V případě zahraničních soutěží se vybere jedna sázková kancelář. V případě českých dat se pak pracuje, jako by se jednalo o jednu sázkovou kancelář.

Jedním z problémů konzistence dat jsou kontumace a nedohrané zápasy, které ale nakonec mají připsaný výsledek nebo různé mediálně známé „korupční skandály“. Dalším problémem je existence dvoubodového systému[1] v nejvyšší

[1]Dvoubodový systém je pravidlo, kdy se do žebříčku soutěže za výhru zápasu připisují dva body. Dnes je standardní systém tříbodový. Za remízu byl v obou případech vždy jen jeden bod.

německé soutěži během prvních dvou zde zpracovaných sezón ([14]). Ze statistického hlediska jsou takové zápasy pro tuto práci chybnými pozorováními, která by měla být odstraněna. Faktem ale je, že i kontumovaný výsledek se do soutěže počítá, a proto bude brán v úvahu například při počítání veličin, které jsou představeny dále. Zpracovávaná data tedy mohou být v některých ohledech nekonzistentní, ale i přesto jsou zpracovávána a vzniklá chyba zanedbána.

2.1.1 Rozdělení na sady dat a proměnná STYP

Protože jsou historická data potřebná nejen ke správnému zvolení a kalibraci modelů, ale také k následnému ověření jejich vlastností, byla ve zpracovávaných datech vytvořena proměnná *STYP*, která nabývá hodnoty 1 pro zápasy určené k výběru strategií a nastavení modelů (*trénovací / kalibrační data*) a hodnoty 2 pro zápasy určené k ověřování zkoumaných modelů (*ověřovací / kontrolní data*). Pro účel kalibrace modelů byly vybrány zhruba 2/3 všech dostupných dat (konkrétně 14 z 21 sezón u zahraničních soutěží a 11 ze 16 sezón u české fotbalové ligy), přičemž zbylá data poslouží k ověřování.

2.2 Základní přehledové statistiky

Tato podkapitola stručně představí souhrny o zpracovávaných datech. Souhrny slouží k rychlému přehledu o množství dat a jejich základních charakteristikách. V tabulce 2.2.1 lze nalézt počty dat, která jsou pro tuto práci k dispozici a která byla zpracovávána. Tato tabulka, další základní přehledy a veškerá vstupní data jsou součástí sešitu DP_Prehledy.xlsx (seznam všech přiložených souborů je k dispozici na konci textu v příloze A.4).

Základní údaje	CZE	ENG	GER
Počet sezón	16	21	21
Počet zápasů	3840	8144	6426
Počet kurzů	3756	4560	3671
Průměr gólů domácích	1.506	1.524	1.671
Průměr gólů hostí	0.961	1.114	1.216
Průměr gólů na zápas	2.467	2.638	2.887
Počet týmů v soutěži	16	22/20*	18
Počet kol za sezónu	30	42/38*	34
Počet zápasů STYP = 1	2640	5484	4284
Výhry domácích celkem	49.53%	46.32%	46.97%
Remízy celkem	27.55%	26.56%	25.74%
Výhry hostů celkem	22.92%	27.12%	27.30%

Tabulka 2.2.1: Tabulka základního přehledu o počtu dat, kde * značí změnu počtu týmů v anglické soutěži.

První dvě sezony anglické nejvyšší ligy, které jsou v těchto datech k dispozici měly 22 týmů. Později byl počet týmů v této lize snížen na 20.

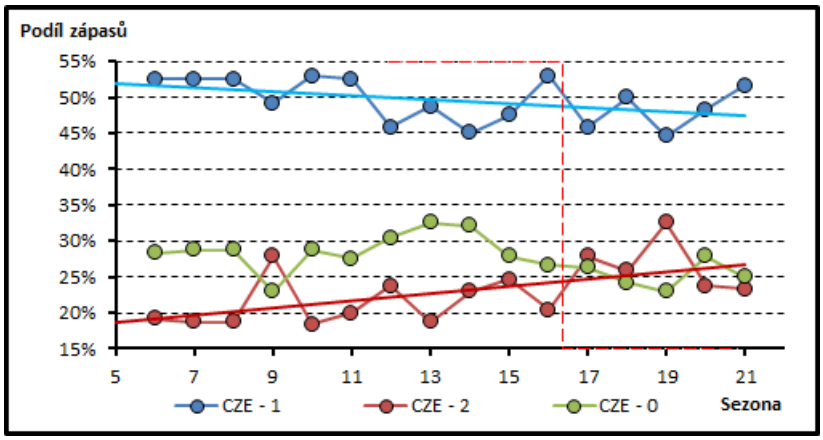
Pokud se jedním zápasem bude rozumět realizace náhodné veličiny s multinomickým rozdělením pravděpodobnosti, lze opakováním takového pokusu získat náhodný výběr jistého rozsahu.

Problém získání náhodného výběru spočívá v tom, že nelze označit všechny zápasy dané soutěže za opakování jednoho pokusu se stále stejnými neměnnými podmínkami. Ze základních vlastností hry je totiž patrné, že „každý zápas je jiný“ už jen proto, že v každém zápase proti sobě hrají jiné týmy. Pokud spolu však dané týmy hrají opakovaně, hrají proti sobě často jiné sestavy hráčů.

Základním problémem je tedy samotná existence určité „heterogenity“ fotbalových zápasů, která vyplývá ze samotné podstaty hry. K určování strategií jak sázet nebo pouze odhadovat pravděpodobnosti výsledků zápasů lze přistupovat mnoha způsoby. Jedním z přístupů, jak modelovat výsledky zápasů, je přístup založený na statistickém rozdělení počtu vstřelených branek (např. Poissonovo). Pro hrající týmy lze pak pomocí různých typů dvourozměrných Poissonových rozdělení modelovat „útočné a obranné síly“ týmů na domácím a hostujícím hřišti. Takové přístupy lze nalézt například v [10] nebo [11].

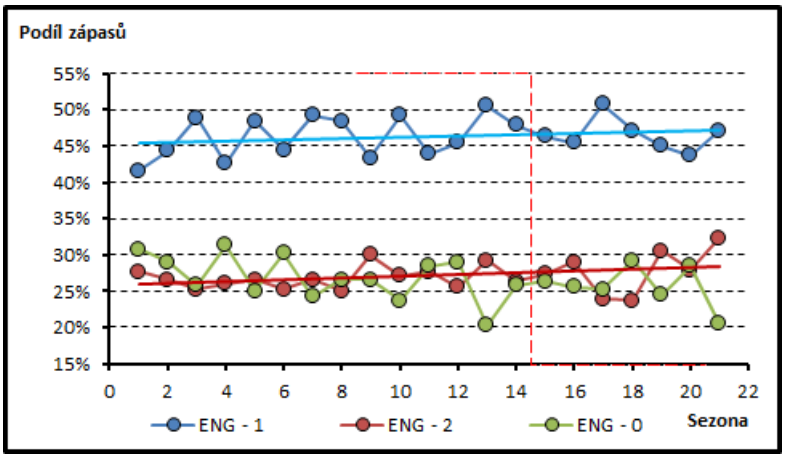
Právě domácí prostředí hraje nezanedbatelnou roli v pravděpodobnostech výsledků zápasů. Tento fenomén výhody domácího prostředí jako takový zde není zkoumán, bude však dále předpokládáno, že určitá „výhoda“ domácího prostředí existuje.

Podíly výsledků během všech zpracovávaných sezón lze vidět na obrázcích 2.2.1, 2.2.2 a 2.2.3, kde je možné vždy pozorovat vyšší pravděpodobnost výhry domácích. Pouze pro ilustraci jsou ve všech obrázcích lineárními čarami proložené[2] výhry domácích a výhry hostů, ze kterých by bylo možné usuzovat o existenci systematické změny v čase (u českých a německých dat lze vidět pokles 1 - výher domácích a růst 2 - výher hostů). Podrobné zkoumání těchto jevů však není zahrnuto v této práci. Součástí všech obrázků je také červená pomocná přerušovaná čára, zobrazující zlom dělení pro proměnnou $STYP$.

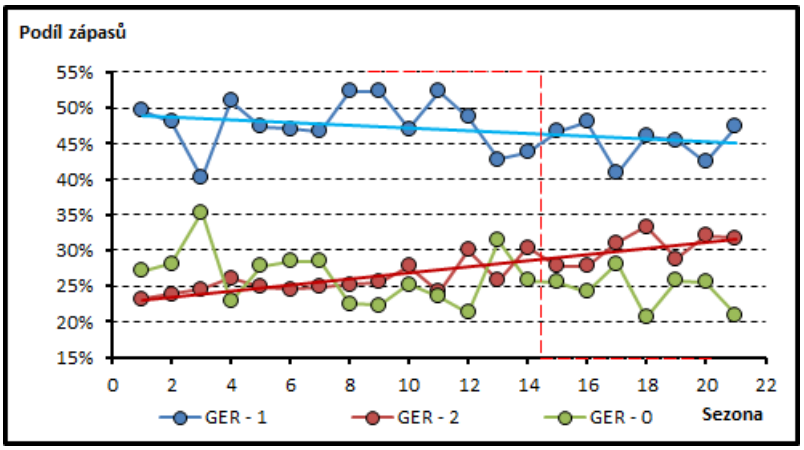


Obrázek 2.2.1: Historická data po sezónách. Poměry výher domácích, remíz a výher hostů v nejvyšší české fotbalové lize.

[2]Jedná se o minimalizaci kvadrátů odchylek.



Obrázek 2.2.2: Historická data po sezónách. Poměry výher domácích, remíz a výher hostů v nejvyšší anglické fotbalové lize.



Obrázek 2.2.3: Historická data po sezónách. Poměry výher domácích, remíz a výher hostů v nejvyšší německé fotbalové lize.

Pro získání homogenního náhodného výběru, či alespoň přiblížení se takovému výběru, jsou zavedeny postupy založené na hodnotách veličin, kterým se věnují následující podkapitoly.

2.3 Charakteristické veličiny hrajících týmů

Základní myšlenkou této podkapitoly je nalezení rozhodujících faktorů, které budou dále popisovat hrající týmy v daném zápase. Kvůli omezené dostupnosti dat jsou za takové faktory považovány spočtené veličiny, které vycházejí z dosavadních výsledků týmu.

V následujícím výčtu je uveden výpis veličin, které jsou dále zkoumány a uvažovány. Každá veličina je na konci svého značení opatřena písmenem „D“, nebo „H“ podle toho, zda se jedná o hodnotu domácího týmu, nebo o hodnotu týmu hostů.

Veškeré dále uvedené veličiny byly spočteny algoritmicky (zdrojové kódy DP_001.m - DP_004.m). V důsledku tohoto kompletního algoritmického zpracování se mohou vyskytnout nepřesnosti, které však během zpracování nepůsobily zásadní problémy a budou dále v této práci zanedbány. Jednou z nepřesností je pořadí týmů na stejném místě tabulky (viz výčet dále).

- *Pořadí v tabulce (PD, PH).* Může nabývat hodnot od jedné až do počtu týmů v sezóně. Pořadí je vždy určeno nejprve počtem bodů, pak rozdílem skóre a poté větším počtem vstřelených branek. V případě shodného umístění dvou či více týmů na jednom místě je určeno pořadí týmů lexikograficky[3].

- *Pořadí v dílčí tabulce (PDD, PHH).* Veličina je podobná předchozímu *pořadí v tabulce* s rozdílem, že pořadí domácího týmu se počítá z tabulky domácích zápasů, kam se výsledky týmů započítávají ze zápasů na domácím hřišti. Podobně pak pořadí hostů, které se počítá z tabulky hostujících zápasů. Pro každý tým jsou tedy současně k dispozici dvě tabulky a pro každý zápas se pořadí určuje podle toho, na jakém hřišti se hraje.

- *Síla týmu (SD, SH).* Veličina je odvozená z pořadí v tabulce. Jde o hrubší dělení pořadí v tabulce. Veličina tedy vyjadřuje pozici týmu v některé části tabulky (třetina, čtvrtina apod.). Veličina může být odvozena z obou typů pořadí (tabulek). Zde je však uvažována pouze pro běžnou (celkovou) tabulku.

- *Forma týmu (FkD,FkH).* Jedná se o počet bodů za posledních k zápasů. Výběr k bude z množiny $\{4, 5, 6, 7\}$ proveden tak, aby bylo dosaženo co nejlepších výsledků. Po odehrané půlce sezóny se nuluje forma týmů z důvodu zimní pauzy.

- *Rozdíl skóre (RD,RH).* Popisuje sílu týmu nikoliv prostřednictvím pořadí v tabulce, ale podle rozdílu vstřelených a obdržených branek. Problémem tohoto kritéria je vliv času na hodnotu tohoto kritéria. S přibývajícím množstvím zápasů se z podstaty věci zvětšuje variační rozpětí[4] této hodnoty v soutěži.

[3] Problém více týmů na stejném místě nastává především po 1. odehraném kole sezony, kdy se stává, že dva či více zápasů skončí stejným výsledkem.

[4] Variační rozpětí je zde počítáno jako rozdíl maximální a minimální hodnoty.

- *Podíl bez obdrženého (BOD, BOH).* Jedná se o procento zápasů (ze všech odehraných), kdy tým neinkasoval žádnou branku. Z veličiny je dále odvozena veličina *obrana (OD, OH)*, která nabývá hodnot:

$$OD(resp.OH) = \begin{cases} 1, & BOD(\text{resp.}BOH) > \text{median}\{BOD(\text{resp.}BOH)\}, \\ 0, & \text{jinak}, \end{cases}$$

  kde medián se počítá z množiny odehraných zápasů takových, kde kolo zápasu je větší než 5. Při výpočtu se do této množiny zahrnují jak týmy domácích, tak týmy hostů. Veličina tedy popisuje, zda tým patří do té poloviny týmů, která „lépe brání“, či nikoliv.

- *Podíl bez vstřeleného (BVD, BVH).* Je analogická předchozí veličině. Počítá se jako procento zápasů, kdy tým nevstřelil žádnou branku. Podobně jako v předchozím případě je na tuto hodnotu navázána veličina *útok (UD,UH):*

$$UD(resp.UH) = \begin{cases} 1, & BVD(\text{resp.}BVH) < \text{median}(BVD(\text{resp.}BVH)), \\ 0, & \text{jinak}, \end{cases}$$

  kde medián se počítá analogicky jako v předchozím případě. Veličina tedy popisuje, zda procento zápasů bez vstřelené branky je menší než medián, či nikoliv. Velmi zjednodušeně řečeno dělí veličinu BVD na dvě poloviny.

Libovolně by bylo možné vytvářet další veličiny nebo dělit veličiny zvláště pro domácí a hostující zápasy. Nadbytečné množství veličin by však způsobilo dělení na tolik skupin, že by počty pozorování ve skupinách byly příliš nízké.

2.4 Charakteristické veličiny zápasu

Veličiny zmíněné v předchozí podkapitole umožňují vytvářet veličiny charakterizující přímo konkrétní zápasy. Pomocí těchto charakteristik je pak možné dělit zápasy do skupin, které by měly být blízké homogenním souborům.

- *Ukazatel pořadí (ip).* Tento ukazatel je definován rozdílem pořadí v tabulce $ip = PD - PH$. Pro nedostatečný počet četností u krajních hodnot ip mohou být pro konkrétní soutěže hodnoty ip upravovány spojováním více hodnot do jedné. V případě českých dat je počítán již od druhého kola soutěže, jinak je počítán až po odehrání alespoň 4 zápasů každým týmem.

- *Ukazatel pořadí 2 (ipDH).* Tato jiná verze ukazatele pořadí v tabulce je analogickou verzí předchozího ip, kde se rozdíl počítá jako $ipDH = PDD - PHH$. V případě českých dat je počítán již od třetího kola soutěže, jinak je počítán až po odehrání alespoň 4 zápasů každým týmem.

- *Ukazatel síly (is).* V tomto případě se vytváří všechny možné kombinace veličiny *síla týmu* tedy $is \in \{(SD \times SH)\}$.

- *Ukazatel formy týmů (ifk).* Podobně jako u rozdílu pořadí, se počítá rozdíl veličiny formy týmů $ifk = FkD - FkH$. Pro zvýšení četností u krajních hodnot mohou být opět spojována pozorování některých skupin do jedné. Ukazatel se počítá při odehrání alespoň tří zápasů. Po nulování tohoto ukazatele v polovině sezóny se veličina počítá po odehrání alespoň dvou zápasů.

- *Ukazatel rozdílu (ir).* Ukazatel rozdílu *ir* je definován způsobem: $ir = RD - RH$. Jelikož rozpětí hodnot *ir* může být příliš velké, byl dále tento ukazatel upravován pro každou soutěž konkrétními mezemi tak, aby bylo vytvořeno přibližně 10 skupin. Pro vytvoření mezí bylo využito ekvidistantních kvantilů z dostupných hodnot *ir*. Hodnota *ir* se počítá pouze pokud jsou odehrány alespoň 4 zápasy.

- *Ukazatel obrany a útoku OU.* Tento ukazatel je kombinací všech možných hodnot veličin *UD, UH* a *OD, OH*. Počítán je po odehraných alespoň 5 zápasech v dané sezoně.

Všechny zde představené ukazatele vychází z veličin z předchozí podkapitoly. Hodnoty ukazatelů byly v některých případech upraveny tak, aby byl k dispozici dostatečný počet (60) četností (podrobnosti lze nalézt v přiloženém souboru DP_AC_ADN.xlsx). Veličiny byly vždy spočteny tak, aby jejich hodnoty byly známé ještě před zahájením zápasu. Jedinou výjimkou je ukazatel *ir*, pro který byly meze spočteny příslušnými kvantily ze všech dostupných trénovacích dat. Většina ukazatelů bohužel nerozlišuje všechny možné kombinace, neboť by jich bylo příliš mnoho. Rozdílem se však tyto veličiny snaží vyjádřit rozdíl výkonnosti mezi hrajícími týmy.

2.5 Kritéria kvality a úspěšnost odhadů SK

Jedním z problémů vhodnosti zkoumaných modelů je určení toho, jak moc jsou modely kvalitní v odhadování pravděpodobnosti[5].

V pojetí této práce je jeden fotbalový zápas náhodným pokusem, jehož výsledkem může být jeden z množiny základních jevů ozn. $\{D, R, H\}$. Tyto jevy mohou nastat zřejmě s nějakými nenulovými pravděpodobnostmi. Problémem je však pouze jedno dostupné pozorování výsledného jevu a nemožnost opakování tohoto náhodného pokusu se stejnými podmínkami.

Tato podkapitola tedy uvede dvě kritéria kvality odhadů. Jedním kritériem je *průměrná doplňková pravděpodobnost* a druhým je *relativní rozdíl četností*. Tato kritéria tedy poslouží nejen k hodnocení modelů, ale také k měření úspěšnosti sázkové kanceláře.

[5]Zde se mínění kvality odhadu dotýká pouze bodových odhadů pravděpodobnosti.

2.5.1 Průměrná doplňková pravděpodobnost *AC*

Průměrnou doplňkovou pravděpodobností AC pro vybranou množinu zápasů $X = \{x_1, x_2, \ldots, x_n\}$ nazveme hodnotu spočtenou aritmetickým průměrem z takových hodnot c_i, které pro i-tý odehraný zápas z dané množiny nabývají hodnoty

$$c_i(A) = 1 - r_i(A), \tag{2.5.1}$$

kde $r_i(A)$ byl odhad pravděpodobnosti, že nastane událost A. Jevem A je zde označen skutečný výsledek utkání. Pro každý zápas z množiny, pro kterou se hodnota AC počítá, je tedy spočten doplněk odhadu pravděpodobnosti do jedné.

Při porovnávání různých technik odhadů se bude považovat nižší hodnota AC za lepší. Nutno dodat, že nulová hodnota kritéria AC při $n \to +\infty$ je dosažitelná pouze teoreticky a to v případě, že systém odhadování výsledků je naprosto přesný a již se nejedná o událost náhodnou, ale deterministickou. To by však odporovalo základnímu přístupu k této práci.

Další vlastností tohoto kritéria je, že nezávisí pouze na kvalitě odhadů, ale také na hodnotách skutečných pravděpodobností. Tuto vlastnost ilustruje následující příklad.

Nechť skutečné a neznámé pravděpodobnosti p_{ij} jsou vždy rovnoměrně rozdělené na množině základních jevů. Pro střední hodnotu $E\{c_i\}$ pak platí

$$E\{c_i\} = \sum_{j=1}^{3} p_{ij} \cdot (1 - r_{ij}), \tag{2.5.2}$$

$$E\{c_i\} = \sum_{j=1}^{3} \frac{1}{3} \cdot (1 - r_{ij}), \tag{2.5.3}$$

$$E\{c_i\} = \frac{2}{3}, \tag{2.5.4}$$

kde j indexuje všechny tři základní jevy a o r_{ij} se předpokládá, že jsou pravděpodobnostmi ($\forall i : \sum_j r_{ij} = 1$ a $\forall j : r_{ij} \in [0, 1]$).

Za předchozích předpokladů tedy nelze tímto kritériem odlišit úspěšnost modelů při odhadování. Navzdory těmto vlastnostem bude dále sloužit k porovnání technik odhadů.

2.5.2 Relativní rozdíl četností *ADN*

Relativní rozdíl četností ADN se spočte poměrem celkového absolutního rozdílu četností DN a počtu zápasů n, kde DN se spočte dle vzorce

$$DN = \sum_{j=1}^{j=3} |n_j - r_j|, \tag{2.5.5}$$

kde j indexuje přes jevy množiny $\{D, R, H\}$. Vektor $\boldsymbol{n}$ je pro množinu zápasů $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ určen četnostmi jednotlivých výsledků z dané množiny $\boldsymbol{X}$ a prvky vektoru $\boldsymbol{r}$ jsou počítány dle vzorce

$$r_j = \sum_{i=1}^{n} r_{ij}, \tag{2.5.6}$$

kde i značí zápasy z celé množiny $\boldsymbol{X}$ a j nabývá hodnot $\{1, 2, 3\}$ pro všechny tři možné výsledky zápasu.

Kritérium tedy nabývá hodnoty $DN = 0$, pokud naměřené četnosti jednotlivých jevů odpovídají úhrnu odhadů pravděpodobností, že dané jevy nastanou.

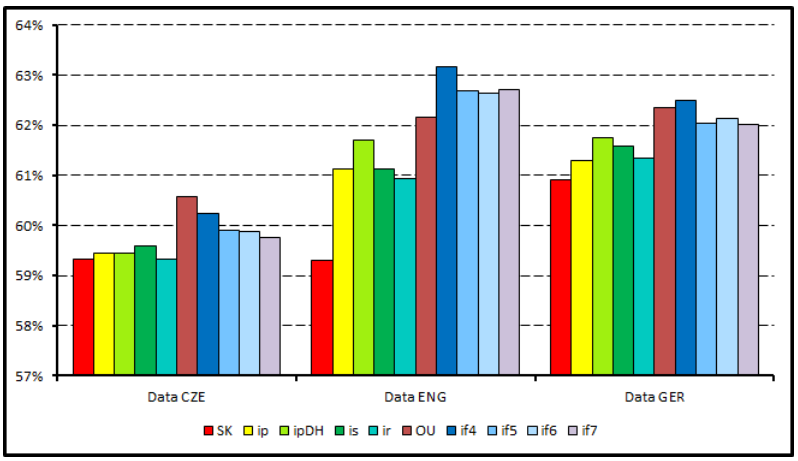
Zjevně nemá, zcela určitě, smysl počítat kritérium pro méně než 3 zápasy, neboť pozorovaných četností by bylo méně než možných jevů. Pokud je způsob odhadování pravděpodobností vytvořen správně, bude platit, že pro dostatečně velký počet zápasů se bude hodnota DN blížit nule. To však nemusí platit opačně. Nízká hodnota kritéria DN nezaručuje nutně správnost odhadu. To je názorně ukázáno na následujícím příkladu.

Příklad: V 13. sezóně nejvyšší anglické fotbalové ligy byly z celkových 380 zápasů pozorovány četnosti výsledků $\{192, 77, 111\}$[6]. Pokud by v každém zápase byly odhady pravděpodobností rovny $\{\frac{192}{380}, \frac{77}{380}, \frac{111}{380}\}$, byl by úhrn těchto odhadů za celou sezónu totožný s pozorovanými četnostmi. Zde se však předpokládá, že skutečné pravděpodobnosti mohou být (a v mnoha případech jsou) různé. Ani toto kritérium tedy není ideální.

2.5.3 Výběr charakteristických veličin

Kvůli následujícímu postupu hledání co nejlepší strategie sázení je vhodné snížit počet dříve zavedených charakteristik zápasů. Snížení bylo provedeno pomocí kritéria AC na množině dat, kde $STYP = 1$ (trénovací sadě). Výsledky těchto hodnot lze vidět na následujícím grafu 2.5.1.

Ve všech případech byly pomocí hodnot charakteristik vytvořeny skupiny zápasu a následně dle relativních četností spočteny bodové pravděpodobnosti. Tyto odhady sloužily ke spočtení hodnot AC.



Obrázek 2.5.1: Hodnota AC na sadách trénovacích dat. Význam zkratek: *SK* - sázková kancelář; *ip* - ukazatel pořadí; *ipDH* - ukazatel pořadí 2; *is* - ukazatel síly; *ifk* - ukazatel formy za posledních k zápasů; *ir* - ukazatel rozdílu.

[6]Viz zdrojová data Anglie.

Z výsledných hodnot zobrazených v grafu 2.5.1 lze shrnout, že ukazatel *ip* vykazuje lepší nebo stejné hodnoty na všech sadách dat než-li ukazatel založený na podobném principu *ipDH*, případně odvozená veličina *is*. Veličina *ir* dosahuje v případě českých a anglických dat nejlepších hodnot z vybraných ukazatelů, zatímco veličina *OU* vyšla v případě českých dat nejhůře. Za soubor veličin *ifk* byl vybrán ukazatel *if5*, neboť dosáhl vždy lepšího výsledku než *if4* a na rozdíl od podobných výsledků *if6* a *if7* stále vystihuje krátkodobý charakter soutěže. Dále jsou tedy k dalšímu postupu práce vybrány ukazatele *ip*, *ir* a *if5*.

Odhady sázkové kanceláře jsou dle kritéria AC (graf 2.5.1) ve všech případech lepší než odhady spočtené pomocí navržených veličin. Nejvýraznější rozdíl lze vidět v případě anglických dat, zatímco nejnižší rozdíl je patrně u sady českých dat.

Vzájemné rozdíly mezi soutěžemi, mohou být také způsobeny rozdíly ve „vyrovnanosti“ daných soutěží. Například v tabulce 2.2.1 lze vidět, že v Česku končí výhrou domácích 49.53% zápasů, zatímco v Anglii je to 46.32%. Naopak výhry hostů jsou častější pro Anglii.

Pro porovnání úspěšnosti sázkové kanceláře na jednotlivých soutěží bylo také využito kritérium ADN (odstavec 2.5.2). Výsledky z porovnání tohoto kritéria jsou shrnuty do tabulky 2.5.1.

Data	Počet dat	ADN (SK)	AC(SK)
Data CZE	2556	9.19%	59.32 %
Data ENG	1900	3.45%	59.31 %
Data GER	1528	2.57%	60.91 %

Tabulka 2.5.1: Hodnoty kritéria ADN pro všechny soutěže kde STYP = 1.

V tabulce lze pozorovat nejhorší hodnotu kritéria právě na sadě českých dat, kde dosahuje více než dvojnásobné hodnoty oproti anglické soutěži a více než trojnásobné oproti datům německé soutěže.

Nutno zde dodat, že počítání kritéria pro odvozené modely na stejné množině, na které proběhlo jejich odvození nemá smysl, neboť by kritérium dosahovalo nulových hodnot. Dále je pro úplnost zobrazena tabulka s hodnotami *ADN* pro zbylá kontrolní data. V této tabulce lze nalézt nižší hodnoty oproti předchozí tabulce. Zda jsou nižší hodnoty způsobeny systematicky lepšími postupy sázkové kanceláře při odhadování výsledků není testováno a jedná se tedy pouze o domněnku.

Data	Počet dat	ADN (SK)	AC(SK)
Data CZE	1200	5.87%	58.95%
Data ENG	2660	1.93%	57.36%
Data GER	2142	2.28%	59.68%

Tabulka 2.5.2: Hodnoty kritéria ADN pro data všech soutěží kde STYP = 2.

3 Analýza vybraných strategií sázení

V této části jsou uvedené poznatky z teorie sázek a matematické statistiky využity pro nalezení optimální strategie. Optimální strategií je zde myšlena taková strategie, která bude sázkaři přinášet zisk a bude možné ji tedy zvažovat jako investiční příležitost.

V následujících podkapitolách jsou představeny vybrané přístupy k sázení (strategie sázení). Každá ze zmíněných strategií může mít několik různých variant. Tyto varianty jsou dále nazývány modely sázení. Pojmem strategie je tedy dále rozuměn přístup k tvoření modelů. V dalších podkapitolách jsou pak představeny výsledky vybraných strategií na množině kalibračních dat.

3.1 Představení vybraných strategií

Představeny jsou strategie založené na proporcionálním sázení, sázení na nejpravděpodobnější variantu a jedna modifikace sázky na domácí. Součástí této podkapitoly je také seznam naivních modelů sázení, které poslouží k porovnání s ostatními modely.

3.1.1 Využití proporcionálního sázení

Strategie vychází z proporcionálního sázení dle Kellyho (odstavec 1.1.1), při kterém se prostředky rozdělují na všechny varianty dle nějakého bodového odhadu pravděpodobností. Postup této strategie lze shrnout v následujících krocích:

1. Vytvoření homogenních skupin zápasů na základě nějakého předem zvoleného klíče (ukazatele).
2. Spočtení bodového odhadu pravděpodobností pro každou skupinu.
3. Nalezení konfidenční množiny pro odhadované parametry (pravděpodobnosti výsledku).
4. Výběr takového bodu z konfidenční množiny, který bude při proporcionálním sázení v nějakém smyslu maximalizovat ziskovou funkci.
5. Rozhodnutí o provedení sázky.
6. Rozhodnutí o výši sázky.

Pro takovou strategii lze vytvářet různé varianty, které se mohou lišit v provedení konkrétních bodů zmíněného postupu. Modely se tedy mohou lišit v prvním bodě, kdy lze volit různé klíče pro vytváření homogenních skupin. Další neznámou, kterou lze volit je vhodný výběr ziskové funkce či proces rozhodování o přistoupení na sázku.

Zisková funkce proporcionálního sázení. [1] Hlavní idea při výběru bodu z konfidenční množiny spočívá v maximalizaci ziskové funkce. Ziskovou funkcí je zde uvažována funkce tvaru

$$E(w) = \sum_{i=1}^{3} p_i b_i o_i, \tag{3.1.1}$$

$$= \sum_{i=1}^{3} o_i b_i^2, \tag{3.1.2}$$

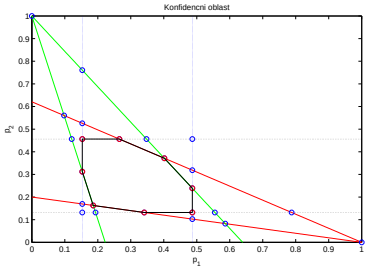
$$= o_1 b_1^2 + o_2 b_2^2 + o_3(1 - b_1 - b_2)^2, \tag{3.1.3}$$

kde $\sum_{i=1}^{3} b_i = 1$ a $\boldsymbol{p} = \boldsymbol{b}$ (podstata sázení dle Kellyho).

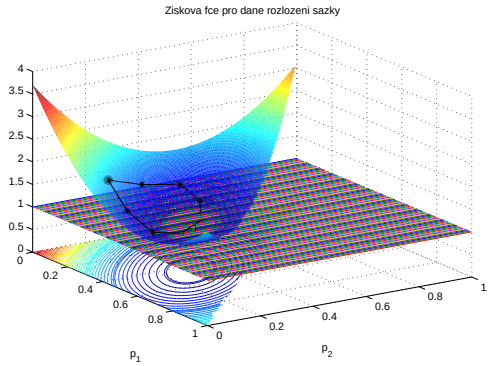
Protože kurzy o_i jsou předem známé konstanty, jedná se o funkci proměnné b_1 a b_2. Proměnné jsou tedy poměry vsazené na první dvě varianty a zbytek do jedné je poměr vsazený na třetí variantu. Jedná se tedy ve své podstatě o parabolickou funkci dvou proměnných.

Jako možní kandidáti na výběr bodu maximalizujícího ziskovou funkci, jsou zkoušeny postupně všechny „rohové“ body konfidenční množiny (odstavec 1.5.4). Myšlenka tedy spočívá ve výběru takového bodu, který s minimálně garantovanou (90%) pravděpodobností lze považovat za skutečnou pravděpodobnost. Dle této pravděpodobnosti je provedena sázka na všechny tři varianty a za předpokladu této pravděpodobnosti pak spočtena střední hodnota výhry. Po té se vybere bod, ve kterém dosahuje zisková funkce svého maxima a dle toho se provede sázka.

Obrázek jedné konfidenční množiny pro hodnoty pozorování $n = [10, 9, 14]$ a kurzy $o = [2.1, 3.1, 3.7]$ a ziskové funkce včetně zobrazení této množiny lze vidět na obrázcích 3.1.1a a 3.1.1b.



(a) Konfidenční množina pro parametry trinomického rozdělení s garantovanou alespoň 90% spolehlivostí v rovině $p_1 \times p_2$.



(b) Zisková funkce pro rozložení prostředků v každém bodě roviny $p_1 \times p_2$ a vyznačená konfidenční množina. Jednotková rovina zobrazuje řez ziskové funkce a pokrytí části prostoru marží.

Obrázek 3.1.1: Zisková funkce a konfidenční množina.

[1] Funkce je nazývána zisková i přesto, že popisuje hodnotu výhry, nikoliv zisku ze sázky.

3.1.2 Sázka na nejpravděpodobnější variantu

Tato strategie je podobná té předchozí v části odhadování pravděpodobností, ale sázka nebude prováděna proporcionálně na všechny varianty, nýbrž bude sázeno na nejpravděpodobnější variantu.

Nejpravděpodobnější varianta je zde vybrána dle postupu popsaném v 1.5.6. Strategii lze popsat následujícími kroky:

1. Vytvoření homogenních skupin zápasů na základě nějakého předem známého klíče.
2. Spočtení četností pro všechny možné výsledky a bodového odhadu pravděpodobností pro každý jev.
3. Statistické testování velikosti parametrů a výběr nejpravděpodobnější varianty.
4. Rozhodnutí o provedení sázky.
5. Rozhodnutí o výši vsazené částky.

Konkrétní modely této strategie se opět mohou lišit v prvním bodě tj. při vytváření homogenních skupin nebo při rozhodování o výši sázky. Všechny zde uvedené modely sází na nejpravděpodobnější variantu vždy, když je taková varianta určena.

3.1.3 Modifikace sázky na domácí

Tato strategie byla přidána k předchozím po prozkoumání naivních modelů. Sázení na domácí se ukázalo mezi naivními modely jako velmi efektivní.

Pro získání co nejlepších výsledků byl naivní model modifikován a postup při aplikaci této modifikované strategie již pouze spočívá ve volbě, zda v daném zápase vsadit na domácí a kolik vsadit.

Konkrétní podobu tohoto modelu lze volit libovolně (dle kurzů, dle různě spočtených ukazatelů, nebo dle marže v daném zápase). Ve zde zkoumané verzi se na sázku přistoupí vždy, pokud se testem prokáže, že vítězství domácích je nejpravděpodobnější variantou. Na rozdíl od předchozí strategie se tedy nesází na hosty ani remízy. Tuto strategie lze shrnout následujícím postupem:

1. Vytvoření homogenních skupin zápasů podle nějakého předem známého klíče.
2. Spočtení četností pro všechny možné výsledky a bodového odhadu pravděpodobností pro každý jev.
3. Statistické testování velikosti parametrů.
4. Na sázku se přistoupí, pokud je výhra domácích nejpravděpodobnější variantou.
5. Rozhodnutí o výši vsazené částky.

3.1.4 Metodika tvoření modelů

Problematika tvoření a kalibrování modelů spočívá především v počtu jejich možných variant. Tento odstavec tedy stručně uvede tyto varianty a jejich značení používané dále.

V této práci jsou celkem zkoumány tři strategie:

- proporcionální (odstavec 3.1.1) - zkr. *prop*,
- nejpravděpodobnější (odstavec 3.1.2) - zkr. *maxpi*,
- modifikované domácí (odstavec 3.1.3) - zkr. *modidom*.

Všechny tyto strategie byly zkoumány ve variantách dle různých využívaných ukazatelů (*ip, if5, ir*). Při rozhodování o provedení sázky v případě proporcionálního sázení je rozhodující hodnota výherní („ziskové“) funkce. V této práci jsou modely a hranice jejich požadovaných minimálních očekávaných výher voleny pouze na základě výsledků simulací. Dále je tedy vždy ke každé strategii uvedeno, jakého ukazatele (potažmo min. požadované výhry) využívá.

3.1.5 Naivní modely sázení

Zde je seznam naivních modelů, které slouží k porovnání s vybranými (odvozenými) modely.

- Sázka na favorita - vždy je vsazeno na nejnižší kurz (pokud jsou některé kurzy shodné, je vsazeno postupně na domácí, remízu nebo hosty).
- Sázka na outsidera - je analogické předchozímu postupu, avšak je sázeno na nejvyšší kurz.
- Sázka na domácí.
- Sázka na hosty.
- Sázka na remízu.
- Náhodné sázení - náhodně sázeno na jednu ze tří možných variant.

3.2 Strategie řízení kapitálu

Mezi klíčové problémy sázení nepatří jen správné odhadování výsledků, volba vhodné strategie či správná kalibrace konkrétního modelu. Dalším podstatným prvkem úspěšného sázení je také správné řízení kapitálu, který je k sázení určen.

V podkapitole věnované teorii sázek (1.1) bylo podrobně představeno proporcionální sázení, u kterého je dokázáno (viz [2]), že je log-optimálním. Název je odvozen z principu, kdy je maximalizována střední hodnota logaritmu výhry, nikoliv střední hodnota výhry samotná. Tato vlastnost však na úkor maximalizace držených prostředků či potenciální výhry zajišťuje nemožnost dostat sázkařovi prostředky do záporných hodnot. Dokonce mu vždy zajišťuje možnost vsadit na další sázku a přitom za jistých podmínek stále zajišťuje exponenciální růst prostředků. Exponenciální růst je v oboru investic běžným jevem. Setkat se s ním lze již například u složeného úrokování.

Veškeré předchozí výsledky a odvození v odstavci 1.1.1 však platí pouze za předpokladu neomezené dělitelnosti sázených částek a nulové marže. Především druhý předpoklad je v praxi téměř vždy nesplněný, a proto již není optimální sázet všechny prostředky [1]. Přitom právě reinvestování vyhraných prostředků jsou spolu se správným (lepším než sázková kancelář) odhadem výsledku základem pro zajištění exponenciálního růstu prostředků. Rozhodnutí, kolik sázet a kolik si nechávat stranou pro další sázení je tedy nyní ještě těžší.

Řešení tohoto problému je naznačeno v [2] a nebude zde pro jeho netriviálnost uváděno. Problematika však spočívá v maximalizaci střední hodnoty logaritmu sázkařova kapitálu při sázení podílu d. Kapitál pak po n sázkách splňují

$$S_n = (1-d)S_{n-1} + \sum_{i=1}^{m} o_i b_i S_{n-1} p_i, \tag{3.2.1}$$

kde $\sum_{i=1}^{m} b_i = d$ a S_0 je počáteční kapitál. Nejedná se tedy o míru zdvojení uvedenou v definici 1.1.3. Tato maximalizace však není triviální, a proto budou při zpracovávání zkoumány jen některé heuristické přístupy:

- Reinvestování s využitím proporcionálního sázení. Tento postup však kvůli existenci marže může vést k exponenciálnímu poklesu kapitálu.
- Proporcionální sázení bez reinvestice. V tomto případě se využívá rozložení prostředků jako v předchozím bodě, ale již se nesází všechny prostředky, avšak jen nějaká část.
- Sázka jedné jednotky v každém kole. V tomto případě lze předpokládat na počátku neomezené množství prostředků a sledovat pak počet vsazených jednotek a celkovou výhru.

Právě poslední zmíněná varianta bude nejčastěji využívána při měření efektivity zkoumaných strategií. Důvodem pro tuto variantu je snadno vyhodnotitelné kritérium průměrného zisku z provedené sázky. Skutečné řešení tohoto problému by zahrnovalo analýzu navržení vhodných kritérií, která by v nějakém smyslu optimalizovala parametry b_i. Pokud by probíhalo sázení jen nějaké části kapitálu, bylo by v praxi vhodné investovat zbývající kapitál nějakým jiným způsobem (úročení na bankovním účtu apod.), který zajistí minimální výnos.

3.3 Výsledky strategií na kalibračních datech

V této podkapitole jsou představeny výsledky získané modelováním všech představených strategií a jejich konkrétních modelů. Pro každou zemi (soutěž) zvlášť byly simulovány nejen vybrané strategie v různých variantách, ale také naivní modely.

3.3.1 Výsledky naivních modelů

Z výsledků, které jsou shrnuty v tabulce 3.3.1, vyplývá, že mezi nejhorší způsoby sázení patří sázení na hosty nebo outsidery (nejvyšší kurz). Naopak mezi nejlepší patří sázení na domácí nebo na favorita (nejnižší kurz). Náhodné sázení dosahovalo převážně výsledku „mezi“ naivním modelem pro domácí a modelem pro hosty. Je to dáno povahou tohoto náhodného sázení, kdy byla nastavena 33% pravděpodobnost na každou variantu.

Soutěž:	CZE	ENG	GER
Počet dat	2556	1900	1529
Prům. marže	9.39%	8.07%	9.55%
Model:	Průměrná výhra		
favorit	-4.41%	-4.89%	**-5.43**%
outsider	-29.03%	-16.19%	-11.76%
domácí	**-2.50**%	**-2.24**%	-8.20%
remíza	-6.64%	-11.92%	-13.85%
hosté	-30.01%	-19.76%	-10.65%
náhodné	-13.64%	-7.71%	-12.59%

Tabulka 3.3.1: Výsledky naivních modelů pro STYP=1. Na každý zápas, kde byl vypsán kurz, byla vsazena jedna jednotka.

Zajímavým výsledkem je, že ve všech třech případech lze najít naivní model s průměrným ziskem (ve skutečnosti se vždy jedná o ztrátu) lepším, než je průměrná marže spočtená dle vzorce (1.2.5). Právě tento výsledek byl impulzem pro zahrnutí strategie modifikace sázky na domácí v odstavci 3.1.3 do této práce.

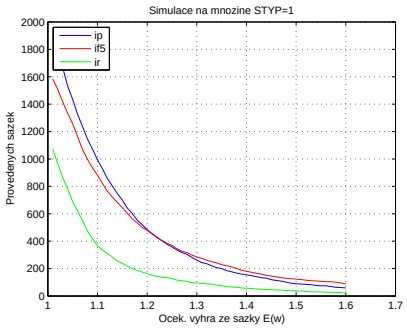
3.3.2 Závislost výhry na počtu provedených sázek

Během modelování se ukázalo, že všechny modely fungují na uvedených kalibračních datech tím lépe, čím je očekávaný zisk ze sázky vyšší. S rostoucím požadovaným ziskem však klesá dostupnost takových sázek. Výběr zápasů, na které má být sázeno, je tedy dán minimální očekávanou výhrou (ziskem) z dané sázky. Zřejmě nebudou uzavírány sázky, je-li očekávaný zisk záporný neboli očekávaná výhra menší než 1[2].

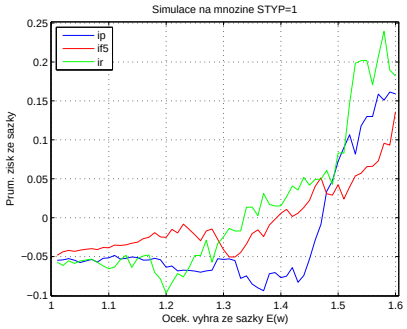
Názorně lze tento jev pozorovat na následujícím obrázku 3.3.1. Problémem je zde velká citlivost na hodnotu požadovaného zisku, kdy při malé množině

[2]V tomto případě je sázena jedna jednotka vždy, když se přistupuje na sázku.

zápasů velmi kolísá průměrný zisk ze vsazené jednotky. Je to dáno problematikou procent z malých čísel. Na obrázku je zobrazena strategie proporcionálního sázení při odhadech vytvářených pomocí všech tří vybraných ukazatelů (odstavec 2.5.3). Důvodem proč skutečný průměrný zisk neodpovídá požadovanému, může být dán nepřesnými odhady pravděpodobností.



(a) Počet zápasů na které bylo vsazeno.



(b) Průměrná výhra z jedné sázky (z jednoho zápasu).

Obrázek 3.3.1: Závislost průměrné výhry ze sázky na počtu vsazených zápasů. Strategie proporcionálního sázení, při sázení jedné jednotky pro modely založené na ukazatelích ip, $if5$ a ir. Simulováno na českých kalibračních datech.

Použitím jiných strategií, kde se počítá očekávaný zisk, lze získat podobné výsledky. V případě jiných zemí (soutěží) jsou však výsledky méně přesvědčivé. Problém u jiných zemí může být způsoben „horší“ schopností zde vytvářených modelů správně odhadovat výsledky nebo lepšími odhady sázkové kanceláře (případně obojím).

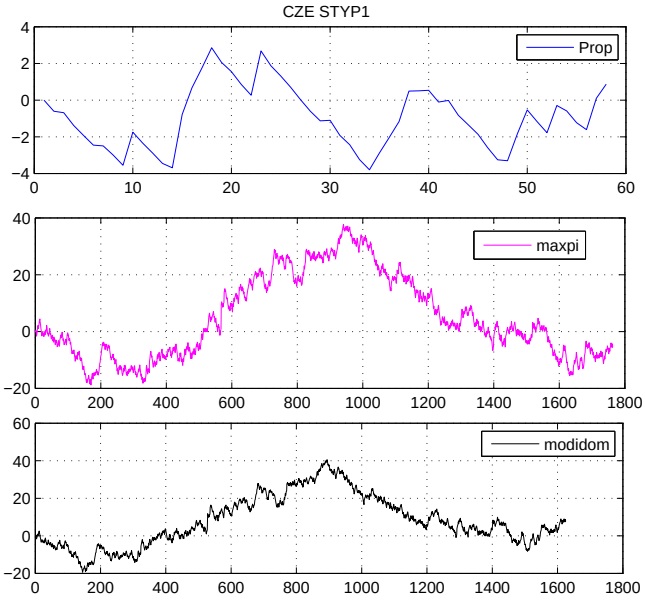
3.3.3 Výsledky pro vybrané soutěže

Tento odstavec představuje výsledky získané na kalibračních datech[3]. Modely, které se ukáží jako úspěšné na této sadě dat, jsou testovány dále na kontrolních datech. Pro každou soutěž a strategii byl nastaven jeden konkrétní model. Tyto modely jsou potom srovnávány z hlediska průměrného zisku ze sázky a z hlediska počtu provedených sázek, ze kterého byl tento průměr počítán. Ve všech případech je vždy vsazena pouze jedna jednotka.

Nejlepším modelem (měřeno průměrným ziskem ze sázky) pro sázení na nejvyšší českou fotbalovou soutěž se ukázala strategie proporcionálního sázení při očekávané výhře alespoň 1.4 z dané sázky a využitím ukazatele pořadí ip. Další strategií, která skončila kladným průměrným ziskem, je strategie modifikace sázky na domácí. Poslední testovaná strategie (sázka na nejpravděpodobnější variantu) skončila se záporným zůstatkem. Problémem prvně zmiňovaného modelu je hranice „požadované“ výhry 1.4, která zapříčinila velmi malý (57) počet provedených sázek. Naproti tomu strategie modifikovaného sázení na domácí s využitím ukazatele pořadí ip umožnila uzavřít 1624 sázek s průměrným ziskem 0.47 %.

[3]Rozdělení na kalibrační a kontrolní (ověřovací) data je popsáno v odstavci 2.1.1

Tyto výsledky jsou však pouze jednou náhodnou trajektorií, nikoliv reprezentativním průběhem daného postupu. Dalším problémem simulace je, že se sázející může během sázení dostat do takřka neomezené ztráty. Ukázka takového průběhu sázení je zobrazena na obrázku 3.3.2.



Obrázek 3.3.2: Zobrazení stavu kapitálu po každé provedené sázce. Simulovány jsou všechny strategie na českých datech STYP = 1. Význam zkratek: *prop* – proporcionální sázení; *maxpi* – sázení na nejpravděpodobnější variantu; *modidom* – modifikace sázky na domácí. Více k popisu je v podkapitole 3.1.

Celkové výsledky včetně výsledků anglické a německé soutěže jsou shrnuty v tabulce 3.3.2.

Soutěž	Strategie	Model	# Sázek	Konečný stav účtu	Průměrný zisk
CZE	Prop	ir (1.4)	57	0.87	1.52%
CZE	Maxpi	ip	1762	-5.33	-0.30%
CZE	Modidom	ip	1624	7.71	0.47%
ENG	Prop	ip (1.0)	1259	-83.13	-6.60%
ENG	Maxpi	ip	1000	0.86	0.09%
ENG	Modidom	ip	946	11.69	1.24%
GER	Prop	if5 (1.4)	91	-6.10	-6.70%
GER	Maxpi	if5	1056	-27.01	-2.56%
GER	Modidom	if5	1052	-26.28	-2.50%

Tabulka 3.3.2: Výsledky odvozených modelů na kalibračních datech. Pro každou strategii (označeno zkratkou) je uveden nejvhodnější model, počet jím provedených sázek, konečný stav účtu a průměrný zisk z jedné sázky.

V souhrnu lze ještě dodat, že sázení na německou soutěž se ve všech případech ukázalo jako „nevýhodné“. Nejlepším modelem tam byla modifikovaná strategie sázení na domácí, pokud se dle ukazatele *if5* ukázala varianta výhry domácích jako nejpravděpodobnější (strategie *modidom* - *if5*). I tak byl ale průměrný zisk z 1052 provedených sázek -2.50%. Naopak za nejlepší lze považovat strategii modifikovaného sázení na domácí použitou pro anglická data (*modidom* - *ip*), kde z 946 provedených sázek bylo dosaženo průměrného zisku 1.24 %. Právě sázení na anglickou a českou soutěž je hlavním předmětem následující kapitoly, kde se tyto modely ověřují.

Všechny tyto výsledky lze získat změnou parametrů (soutěž a $STYP$) ve zdrojovém kódu DP_sim01.m, který využívá zdrojová data ze sešitu Data_pro_matl2.xlsx.

4 Ověření použitých modelů

Tato část práce shrnuje konečné výsledky modelů nastavených dle podkapitoly 3.1. Během simulování těchto modelů na dosud neznámých datech byl řešen problém s měřením všech veličin a počtem četností pro vybrané ukazatele tak, že po každém odehraném zápase byly všechny modely upraveny aktualizováním o ten jeden příslušný záznam. Tímto postupem je pak neustále zajištěn nejvyšší počet dostupných pozorování během procesu simulace. V každém takovém kroku je pak zajišťováno (spojováním skupin „malého počtu“ hodnot ukazatelů do jedné), aby vždy bylo k dispozici alespoň 60 pozorovaných hodnot.

Pro porovnání jsou v tabulce 4.0.1. uvedeny výsledky naivních modelů (dostupné také v DP_Naivni_Sazeni.xlsx), počty dat a průměrné marže na dané množině dat. Jedná se o analogickou tabulku jako v případě tabulky 3.3.1.

Soutěž:	CZE	ENG	GER
Počet dat	1200	2660	2142
Prům. marže	8.37%	4.67%	6.08%
Model:	Průměrná výhra		
favorit	**-0.51%**	-2.45%	-6.33%
outsider	-20.89%	-7.56%	-2.80%
domácí	-3.70%	**-2.19%**	-5.20%
remíza	-13.55%	-4.96%	-11.21%
hosté	-18.92%	-12.99%	**-2.13%**
náhodné	-14.11%	-4.85%	-13.32%

Tabulka 4.0.1: Výsledky naivních modelů na sadě dat STYP = 2.

Zajímavé jsou především výsledky sázení na favorita v případě českých dat, které vzhledem k průměrné marži dopadlo velmi dobře, dokonce lépe než sázení na domácí, které bylo pro testovací množinu dat lepší. Dále lze oproti výsledkům v tabulce 3.3.1 pozorovat ve všech případech zlepšení u sázení na hosty a zhoršení sázky na domácí u českých dat.

Výsledky v této práci odvozených modelů jsou shrnuty v tabulce 4.0.2, kde lze vidět, že ve všech případech je průměrný konečný zisk ze sázky záporný. Také lze pozorovat, že oproti původním kalibračním výsledkům dosáhly skoro všechny modely horších výsledků. Shodu lze tak nalézt pouze ve vzájemném porovnání modelů, kdy se v ani jednom případě nepodařilo nalézt model pro sázení na německou soutěž. Nejlépe pak v obou případech dopadlo sázení strategií *modidom* na česká a anglická data.

Úspěchem může být tedy pouze fakt, že pro každou vybranou soutěž lze najít model, který dosahuje lepších výsledků (nižší ztráty) než je marže. Zjistit pravé důvody tohoto neúspěchu by bylo velmi zajímavé. Jedním z možných důvodů může být změna tvorby kurzů sázkovou kanceláří. Zlepšení „naivního“ modelu sázení na hosty by totiž mohlo znamenat zvyšování vypisovaných kurzů na hosty. Důkazem „výhodnějších“ kurzů jsou i nižší průměrné marže u kontrolní sady

Soutěž	Strategie	Model	# Sázek	Konečný stav	Průměrný zisk
CZE	Prop	ir (1.4)	29	-3.44	-11.88%
CZE	Maxpi	ip	959	-26.26	-2.74%
CZE	Modidom	ip	894	-21.98	-2.46%
ENG	Prop	ip (1.0)	2185	-132.92	-6.09%
ENG	Maxpi	ip	1544	-39.56	-2.56%
ENG	Modidom	ip	1444	-42.31	-2.93%
GER	Prop	if5 (1.4)	274	-23.23	-8.48%
GER	Maxpi	if5	1516	-103.17	-6.81%
GER	Modidom	if5	1510	-98.42	-6.52%

Tabulka 4.0.2: Výsledky vybraných modelů na datech STYP = 2.

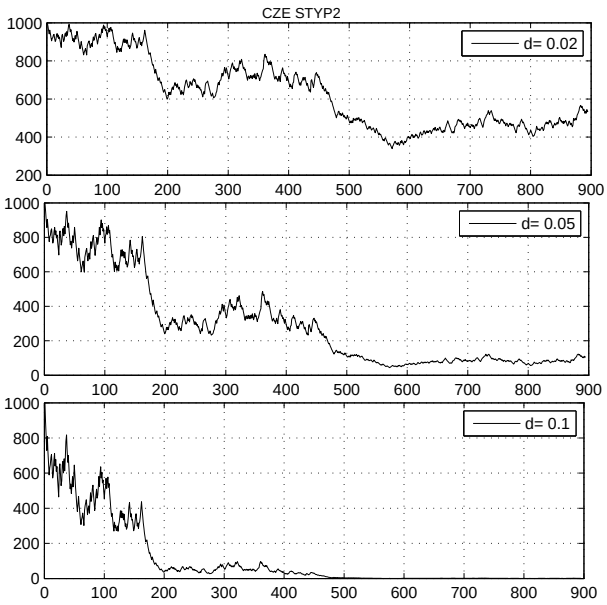
dat. Další možností je zvyšování kurzů u „vyrovnaných“ zápasů, které se více blíží rovnoměrnému rozdělení.

Jiným vysvětlením pak může být změna charakteru soutěží, kdy již na obrázcích v podkapitole 2.2) bylo ukázáno, že u dat $STYP = 2$ je zejména u české a německé soutěže vidět nižší podíl výher domácích a vyšší podíl výher hostů.

4.1 Reinvestování části kapitálu

Problematika řízení kapitálu již byla krátce představena v podkapitole 3.2. V předchozí části práce byly porovnávány modely z hlediska průměrného zisku, pokud byla sázena vždy jedna jednotka. Ve skutečnosti je však nevýhodné sázet fixní částku kapitálu, neboť nemůže docházet k jeho exponenciálnímu zhodnocování. Zde je tedy testována situace, kdy je v každé sázce investován určitý pevný podíl d aktuálně drženého kapitálu.

Volba parametru d není v práci určována žádným optimalizačním kritériem a pro ilustraci bylo voleno z množiny $d \in \{2\%, 5\%, 10\%\}$. Je zjevné, že vyšší d zaznamenává vyšší zhodnocování (v případě úspěšné strategie), ale také rychlejší možnost dostat se s prostředky na 0 (v případě neúspěšné strategie). Naopak nižší hodnoty d jsou „konzervativnější“. Volba d je tedy přirozenou volbou „rizika“, které chce investor podstupovat, přesto ji lze patrně najít sofistikovanějším způsobem. Pro ukázku je simulována strategie *modidom (ip)*, tzn. sázení na domácí, pokud je tato možnost dle ukazatele *ip* označena jako nejpravděpodobnější. Ukázková simulace je provedena na českých datech $STYP = 2$, kde pro různá d lze vidět průběžné stavy kapitálu na následujících grafech obrázku 4.1.1.



Obrázek 4.1.1: Simulace strategie *modidom* na českých datech STYP = 2 při reinvestici kapitálu pro různá d. Vývoj sázkařova kapitálu po n (vodorovná osa) sázkách.

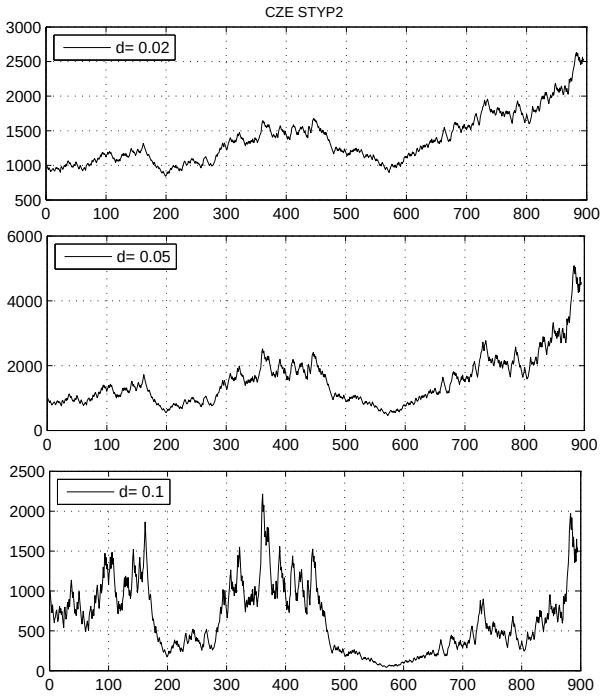
Z obrázku lze vidět, že ve všech případech je konečný stav prostředků nižší než počáteční. Strategie tedy skončila ve ztrátě. To není příliš překvapivé, neboť sázení jedné jednotky skončilo ve ztrátě taktéž. Snadno lze pak vidět, že vyšší d způsobuje v tomto případě rychlejší pokles k nule. Podobné obrázky je možné získat změnou soutěže pro Anglii a Německo ve zdrojovém kódu DP_sim02.m.

Důležité je zde zmínit, že sázkař nemusí vždy požadovat dlouhodobý zisk. Cílem sázkaře může být v určitém případě dosažení konkrétní hodnoty a následného ukončení sázení. Dlouhodobé sázení pak může zvyšovat pravděpodobnost jeho bankrotu (zvláště, pokud se mění charakter soutěže nebo způsobu vypisování kurzů).

4.2 Sázení při nulové marži

Stejně jako v předchozí části textu je zde simulována strategie *modidom – ip* na českých ověřovacích datech, avšak kurzy sázkové kanceláře jsou očištěné o marži sázkové kanceláře vynásobením původního kurzu zlomkem $\frac{1}{1-\xi}$, kde ξ je odhad marže dle vzorce (1.2.5). Simulována je tedy situace, kdy je sázeno na „spravedlivé“ kurzy bez marže.

Na obrázku 4.2.1 lze již pohledem vidět nižší „volatilitu“ v případě nižšího d a naopak. Ve smyslu konečného stavu kapitálu pak vychází nejlépe sázení pro $d = 5\%$. Stanovení hodnoty d je tedy důležité, avšak podstatnějším se jeví nalezení vhodné strategie. Pro simulování modelů při „spravedlivých“ kurzech slouží program DP_sim03.m, kde lze libovolně volit proměnou *list* (zdrojová data) a proměnnou *STYP*.



Obrázek 4.2.1: Simulace strategie *modidom* na českých datech STYP2 při reinvestici kapitálu a nulových maržích sázkové kanceláře pro různá d. Vývoj sázkařova kapitálu po n (vodorovná osa) sázkách.

Ukazuje se tedy, že existence marží je zcela zásadní problém pro sázkaře, pokud chce dlouhodobě „vyhrávat“. Důvodem pro toto simulování je však možnost získat „výhodnější“ kurzy kombinací sázení u více sázkových kanceláří.

4.3 Shrnutí a porovnání výsledků

Tato krátká podkapitola porovnává důležité výsledky předchozích podkapitol. V tabulce 4.3.1 jsou uvedeny výsledky dvou nejlepších modelů (strategie *modidom* – *ip* pro CZE a ENG) z kalibrační části. Ve výsledcích jsou zobrazeny přehledové hodnoty při sázení na obě sady STYP a to při existenci i absenci marže.

Liga	d	STYP	# sázek	Marže	Zůstatek	# sezón	% / rok
CZE	2%	1	1624	ANO	851.41	11	-1.45%
ENG	2%	1	946	ANO	1 060.62	14	0.42%
CZE	10%	1	1624	NE	4 503 834.78	11	114.85%
ENG	10%	1	946	NE	97 749.31	14	38.72%
CZE	2%	2	894	ANO	535.97	5	-11.73%
ENG	2%	2	1444	ANO	333.19	7	-14.53%
CZE	5%	2	894	NE	4 578.21	5	35.56%
ENG	2%	2	1444	NE	1 264.97	7	3.41%

Tabulka 4.3.1: Souhrn hlavních výsledků.

Do tabulky bylo vždy vybráno nejlepší d z testované množiny $\{2\%, 5\%, 10\%\}$. Zůstatek je hodnotou sázkařových prostředků po všech provedených sázkách a odpovídá počáteční hodnotě 1000. Hodnota v posledním sloupci uvádí odpovídající roční procentuální úrokovou sazbu, kterou by bylo nutné složeně každý rok úročit počáteční hodnotu 1000, aby konečné hodnoty byly stejné.

V tabulce lze například pozorovat, že v případě českých trénovacích dat a neexistenci marže by sázkařovy prostředky během testovaných 11 sezón rostly v průměru 115% ročním složeným úrokováním. Existence marže však na stejné množině zápasů způsobuje průběh zhruba 600 sázek s průměrným růstem a následný pokles. Zůstatek sázkařových prostředků pak odpovídá zhruba 1.45% roční ztrátě. Velmi podobný průběh sázkařových prostředků je zobrazen na obrázku 3.3.2.

Z tabulky je dále patrné, že v případě množiny dat $STYP = 2$ je při neexistenci marží možné stále dosáhnout zisku (nyní již nižšího) a naopak vyšší ztráty při zachování skutečných marží. Existence marže je tedy zcela zásadní faktor při zkoumání úspěchu zde odvozených modelů sázení.

Závěr

Hlavním tématem této práce je sportovní kurzové sázení a možnost jeho využití jako investice. Poznatky z teorie sázek jsou uplatněny především při tzv. proporcionálním sázení. Tento druh sázení je však výhodný hlavně při spravedlivých kurzech, což se v závěrech (závěru) ukázalo jako klíčové. Právě existence marží sázkových kanceláří znesnadňuje využití sázení jako efektivní investice.

Hlavním cílem této práce bylo nalezení strategie, kterou by bylo možné využít jako investiční nástroj. Práce se soustřeďuje hlavně na tzv. „1-0-2“ sázení, což je kurzové sázení na výhru domácích, remízu, či výhru hostů. Z veřejně dostupných datových zdrojů byla vybrána data nejvyšší české, anglické a německé fotbalové ligy. Na těchto datech jsou nejprve vytvořené konkrétní modely sázení a následně je ověřována jejich efektivita na kontrolní části dat.

Z pravděpodobnostního pohledu je základem některých strategií odhadování parametrů multinomického (konkrétně trinomického) rozdělení pravděpodobnosti. Problematika nalezení přesné konfidenční množiny popsaná v kapitole 1.5 zde byla nakonec vyřešena aproximací založené na rozdělení „nezávislých podílů“ (odstavec 1.5.4). Tato aproximace umožňuje jednoduchým algoritmem najít konfidenční množinu se sice přesně neznámou, avšak zdola omezenou spolehlivostí. Aproximací je tento postup nazýván proto, že takto získaná množina (osmiúhelník) je zdánlivě podobná přesné optimální oblasti (viz obrázek 1.5.4). Porovnání obou přístupů je pak zpracováno v odstavci 1.5.5.

Samotné hledání efektivní strategie je ve všech případech založené na ukazatelích, které jsou známé před každým zápasem. Jako nejlepší ukazatele se z vybraných ukázaly ukazatele pořadí v tabulce, výsledky za posledních 5 zápasů a rozdíl mezi vstřelenými a obdrženými brankami. Ostatní ukazatele a jejich výsledky jsou zpracovány jako součást kapitoly 2 o zpracování historických dat.

Pro množinu trénovacích dat byly simulovány strategie proporcionálního sázení a sázení na nejpravděpodobnější variantu. Pro porovnání byly také simulovány naivní modely (odstavec 3.1.5), po jejichž simulaci se jako velmi efektivní ukázalo sázení na domácí, které bylo podnětem pro vytvoření modifikace této strategie označené jako *modidom*. Modifikací je myšlena kombinace nejpravděpodobnější varianty a sázky na domácí, kdy se na domácí sází pouze v případě, že je tato varianta označena jako nejpravděpodobnější.

U strategií založených na principu proporcionálního sázení se ukázala důležitost volby požadovaného minimálního očekávaného zisku. Volba tohoto parametru je však problematická v případě, že bodové odhady pravděpodobností nejsou přesné. Bodové odhady u tohoto přístupu závisí na vytvoření homogenních (nebo jim podobných) skupin pozorování. Z dostupných výsledků se ukázalo, že tento přístup funguje nejlépe na množině českých dat. V celé práci se totiž sázení na nejvyšší českou fotbalovou ligu ukázalo (oproti anglické a německé) jako nejvhodnější volba. Lze se tedy domnívat, že ukazatele pořadí v tabulce či aktuální forma mají v české lize větší význam. V případě vybraných zahraničních soutěží však nemusí být tyto ukazatele rozhodující při určování pravděpodobnosti výhry

(vytváření skupin). Jedná se však o nezkoumanou domněnku a zcela jistě je možné tento jev vysvětlit i jiným způsobem.

Zdrojové kódy a algoritmy, které jsou součástí této práce, umožňují modelování všech odvozených strategií na kalibračních i kontrolních datech všech soutěží. Modelováním se ukázalo (viz kapitola 4.3), že z hlediska úspěšnosti strategií je zásadní právě existence marže sázkové kanceláře. Výsledkem je, že sázení na kurzy zvýšené o marži by patrně umožňovalo dlouhodobý zisk. Důležité je zde zmínit, že kombinací kurzů u různých sázkových kanceláří je možné získat výhodnější kurzy „s nižší marží“, což pomáhá právě sázejícímu.

U všech modelů se jevilo sázení na německou soutěž jako nejméně výhodné a nebyl nalezen žádný model s přesvědčivými výsledky. Dokonce lze pozorovat horší hodnoty kritéria *AC* u sázkové kanceláře (viz obrázek 2.5.1), což může vypovídat o vyšší „vyrovnanosti“ této soutěže. Vyšší vyrovnaností je zde myšleno rozdělení blízké rovnoměrnému rozdělení pravděpodobnosti všech výsledků. Naproti tomu nejlepších výsledků dosahuje sázková kancelář u anglické ligy. Dobré výsledky sázení na tuto soutěž však mohou být způsobeny nižší průměrnou marží u této ligy (tabulky 3.3.1 a 4.0.1).

V práci je tedy ukázán přístup, který předpovídá výsledky (pravděpodobnosti) fotbalových utkání a tyto předpovědi následně využívá k sázení jakožto jednoho z možných investičních nástrojů. Ačkoliv nebyl nalezen efektivní model, bylo navzdory existenci marží ukázáno, že je možné dosahovat výsledků lepších, než je ztráta průměrné velikosti marže. Součástí práce je také kapitola *Seznam nevyřešených problémů*, kde jsou shrnuty některé dílčí problémy, jejichž vyřešení by mohlo zlepšit představené modely a jejich výsledky. Důsledné řešení těchto problémů by však přesahovalo rozsah této práce.

Seznam nevyřešených problémů

Zpracovávání této práce s sebou přineslo problémy, které nebyly efektivně (nebo vůbec) vyřešeny. Z tohoto důvodu byla vytvořena samostatná kapitola se seznamem a krátkým popisem některých z těchto problémů. Mezi nimi jsou také některé neověřené domněnky či hypotézy.

- **Výběr sázkové kanceláře** u které sázet. Nejen při studiu této problematiky, ale i v reálné situaci lze vybírat z několika sázkových kanceláří. Přesto, že se sázkové kanceláře nemohou ve svých kurzech lišit příliš (vznikl by prostor pro super-férovou sázku), jsou ve vypsaných kurzech rozdíly. Sázejícímu nemusí být jasné, jakým způsobem kurzy vznikají a které lze využít při kalibraci modelů.

- **Rozhodnutí, jestli vsadit nebo jestli čekat na jiný kurz.** Bylo ukázáno, že kurzy se mohou měnit v čase, avšak není jisté, zda například později vypsané kurzy mají nižší marži nebo jestli jsou dříve vypsané kurzy méně přesné.

- **Kolik vsadit** a kolik si nechávat ve „fondu“. Bylo ukázáno, že při existenci marže se nevyplácí sázet veškeré prostředky. Problematika však spočívá v nalezení vhodného kritéria pro určení optimálního parametru d uvedeného v podkapitole 3.2 nebo 4.1.

- **Jak spojit hodnoty ukazatelů** zápasů s nízkými četnostmi. Pokud jsou požadovány minimální počty pozorování pro daný „výběr / skupinu“, není jednoznačné řešení jaké skupiny spojit. V práci je tento problém řešen algoritmem, který spojuje skupiny ve směru „ke kraji“, a pokud není dosaženo požadované četnosti, přidávají se skupiny „z druhé strany“.

- **Zrychlení algoritmů při hledání konfidenčních oblastí.** V rámci práce byl napsán algoritmus hledající konfidenční množinu pomocí řezů hustotou Dirichletova rozdělení. Tyto řezy jsou však hledány nepříliš efektivními numerickými postupy. V práci je pak tento problém vyřešen aproximací, založené na rozdělení nezávislých podílů (odstavec 1.5.4).

- **Zlepšování sázkových kanceláří v čase.** V odstavci 2.5.3 je dvěma kritérii ukázáno, že u druhé (pozdější) části dat je chyba sázkové kanceláře ve všech případech nižší než u první části. Skutečné testování úspěšnosti odhadů sázkových kanceláří však vyžaduje podrobnější zkoumání.

- **Analýza trendů sportovních soutěží v čase** Grafy v podkapitole 2.2 naznačují, že charakter soutěží se v čase může měnit. Nalezení závislosti na čase by mohlo sloužit k modifikaci zdejších strategií.

- **Problém nalezení největší vnitřní konvexní množiny** ze zadaných bodů (průsečíků). Při hledání konfidenčních množin podle postupu v odstavci 1.5.4 jsou výsledkem průsečíky přímek (horních a dolních mezí).

V této práci je řešení podstatně snadnější, neboť jsou předem známé přímky, které body vytvářejí. Obecně se jedná o netriviální úlohu.

- **Tvar výsledné konfidenční množiny.** V odstavci 1.5.4 je výsledná množina průnikem dvou čtyřúhelníků. Během zpracování se ukázalo, že výsledkem je vždy osmiúhelník. Není však zcela zřejmé, jestli v případě hledání konfidenční oblasti je výsledkem vždy a pouze osmiúhelník.

- **Jednoznačnost konfidenční množiny dle odstavce 1.5.4.** Již ve zmíněném odstavci je ukázáno, jak tvar konfidenční množiny závisí na vybrané dvojici pozorování. Není zde však zcela jasné, jaký je v těchto množinách rozdíl a kterou je nejlepší vybrat.

- **Zapomínání starých výsledků.** Problematika spočívá v postupném zapomínání starých výsledků, kvůli jejich nestálosti v čase. V této práci jsou uvažovány všechny dostupné výsledky bez zapomínání. Není zcela jisté, zda by postupné zapomínání starých výsledků přispělo k lepším výsledkům. Dalším problémem by bylo určení, jak moc staré výsledky vynechávat.

- **Nestálost tvoření kurzů sázkových kanceláří.** Výsledky v podkapitole 3.3 a v kapitole 4 mohou naznačovat, že sázkové kanceláře mění v průběhu své existence způsob vytváření kurzů. Je možné, že sázkové kanceláře nepřizpůsobují své kurzy jen sázejícím, ale také například systematicky snižují kurzy u méně pravděpodobných událostí.

- **Měření efektivity strategií sázení.** V práci je efektivita posuzována nejprve dle sázení jedné jednotky na každý zápas a až po té je simulováno sázení s efektivnějším řízením kapitálu. Není zjevné, zda lze strategii která má v případě „jednotkového sázení“ záporný průměrný zisk upravit vhodným řízením kapitálu tak, aby dosahovala zisku.

- **Volba parametrů modelů.** V podkapitole 3.3 byly hodnoty minimálních požadovaných zisků a konkrétní veličiny vybrány dle empirických výsledků. Není zde uvedeno žádné kritérium, které by optimální parametry dokázalo vybrat obecně.

- **Studium naivních modelů.** Během zpracování byl modifikován naivní model sázení na domácí, který se ukázal jako velmi efektivní. Následně se na českých datech ukázalo jako velmi efektivní sázení na favorita. Tyto naivní modely dosahují nižší ztráty než je průměrná marže a bylo vy velmi vhodné jim věnovat hlubší analýzu.

Literatura a zdroje

[1] Cover T. M., Thomas J. A.: *Elements of information Theory*, John Wiley & Sons, Inc., Hoboken, 2006, 2nd ed.

[2] Kelly J. L., JR: *A New Interpretation of Information Rate*, The Bell System Technical Journal Vol. 35, July 1956, pp. 917 - 926

[3] Hátle J, Likeš J.: *Základy počtu pravděpodobnosti a matematické statistiky*, SNTL, Praha, 1974

[4] Renyi A.:*Teorie pravděpodobnosti*, Academia, Praha, 1972

[5] Anděl J.:*Matematická statistika*, SNTL/ALFA, Praha, 1978

[6] Hušková M.: *Bayesovské metody*, Skriptum MFF UK, Praha, 1985

[7] Kotz S., Johnson L. N., Balakrishnan N.: *Continuous multivariate distributions*, John Wiley & Sons, Inc., New York, 2000

[8] Johnson L. N., Kotz S., Balakrishnan N.: *Discrete Multivariate Distributions*, John Wiley & Sons, Inc., New York, 1997

[9] Omar M. H., Joarder A.H.: *Some Mathematical Characteristics of the Beta Density Function of Two Variables*, Bulletin of the Malaysian Mathematical Sciences Society, Vol. 35.4, 2012, pp. 923-933

[10] Maher M. J.: *Modelling association football scores*, Statistica Neerlandica Vol. 36, September 1982, pp. 109 – 118

[11] Marek P., Šedivá B., Ťoupal T.: *Modeling and prediction of ice hockey match results*, Journal of Quantitative Analysis in Sports, Vol. 10.3, 2014, pp. 357–365

[12] Football-Data.co.uk: *Football Results, Statistics & Soccer Betting Odds Data* [online], [cit. 12. 2. 2015]. Dostupné z: `http://www.football-data.co.uk/data.php`

[13] Liga.cz: *Česko - přehled ligových soutěží a pohárů, archivní statistiky* [online], [cit. 12. 2. 2015]. Dostupné z `http://www.liga.cz/fotbal/cesko/`

[14] Bundesliga.de - die offizielle Webseite der Bundesliga: *DREI-PUNKTE-REGEL GILT SEIT 1995/96* [online], [cit. 2. 10. 2014]. Dostupné z: `http://www.bundesliga.de/de/historie/news/kurioses/drei-punkte-regel-gilt-seit-1995-96_0000247406.php`

A Přílohy

A.1 Rozdíly v kurzech

Zápas 9. kola	1	0	2
Teplice-Slovácko	1.95	3.35	3.95
Č.Buděj.-Příbram	2.3	3.15	3.1
Brno-Plzeň	4.35	3.5	1.82
Dukla-Jihlava	2.05	3.25	3.55
Hr.Králové-Jablonec	3	3.2	2.35
Ostrava-Boh. 1905	1.8	3.5	4.25
Slavia-Sparta	4.5	3.6	1.77
Liberec-Ml.Boleslav	2.05	3.3	3.5

Tabulka A.1.1: Kurzy na zápasy nejvyšší české fotbalové ligy z 22. 9. 2014. Zdroj: `www.ifortuna.cz`.

Zápas 9. kola	1	0	2
Teplice-Slovácko	2.05	3.35	3.45
Č.Buděj.-Příbram	2.2	3.25	3.2
Brno-Plzeň	4	3.5	1.85
Dukla-Jihlava	1.78	3.55	4.3
Hr.Králové-Jablonec	3.15	3.3	2.2
Ostrava-Boh. 1905	2.03	3.35	3.5
Slavia-Sparta	3.95	3.55	1.85
Liberec-Ml.Boleslav	2.28	3.3	3

Tabulka A.1.2: Kurzy na zápasy nejvyšší české fotbalové ligy z 26. 9. 2014. Zdroj: `www.ifortuna.cz`.

A.2 Definice a pojmy ze statistiky

Gamma funkce: Gamma funkce pro $x > 0$ je definována předpisem (zdroj: [5])

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt. \tag{A.2.1}$$

Gamma funkce bývá známá také jako zobecněný faktoriál pro $x > 0$, kdy pro přirozená čísla $n \geq 1$ platí $\Gamma(n) = (n-1)!$. Další často využívanou vlastností pak je $\Gamma(x+1) = x\Gamma(x)$.

Beta funkce: Beta funkce $B(a,b)$ s parametry $a,b > 0$ se definuje jako (zdroj: [5])

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx. \tag{A.2.2}$$

Často využívanou vlastností je

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Zobecněnou beta funkcí s parametry $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ se pak rozumí

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}. \tag{A.2.3}$$

Dirichletův integrál: Dirichletovým integrálem v této práci nazveme integrál

$$\int_{\forall x_i>0:\sum_{i=1}^{k}x_i=1} \prod_{i=1}^{k} x_i^{\alpha_i-1}dx_1\ldots dx_k = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}, \tag{A.2.4}$$

což je opět zobecněná beta funkce.

Dvourozměrné beta rozdělení: Řekneme, že náhodný vektor $\boldsymbol{Z} = (X,Y)$ se řídí dvourozměrným beta rozdělením, má-li pravděpodobnostní funkci ve tvaru:

$$f(x,y) = \frac{\Gamma(a+b+c)}{\Gamma(a)\Gamma(b)\Gamma(c)}x^{a-1}y^{b-1}(1-x-y)^{c-1}, \tag{A.2.5}$$

kde parametry $a,b,c > 0$, proměnné $x,y \geq 0$ a $x+y \leq 1$ ([9]). Jak již jednou bylo zmíněno v této práci, jedná se o speciální případ Dirichletova rozdělení.

A.3 Další užitečné pojmy a vztahy

Bonferroniho nerovnost: (někdy také Booleova nerovnost)

$$P\left(\bigcap_{i=1}^{n}A_i\right) \geq \sum_{i=1}^{n}(A_i) - (n-1). \tag{A.3.1}$$

Tato nerovnost vychází ze základního tvaru

$$P\left(\bigcup_{i=1}^{n}A_i\right) \leq \sum_{i=1}^{n}(A_i). \tag{A.3.2}$$

V různých dalších tvarech lze tuto nerovnost nalézt ve většině úvodů do pravděpodobnosti (např. [4] nebo [5]).

Symetrie beta rozdělení: Z definice beta rozdělení (Definice 1.4.3) zřejmě platí

$$f(x;a,b) = f(1-x;b,a). \tag{A.3.3}$$

Díky tomuto vztahu lze pro distribuční funkci beta rozdělení $F(x;a,b)$ pak psát

$$F(x,a,b) = \int_0^x f(1-t;b,a)dt. \tag{A.3.4}$$

S využitím substituce $u = 1-t$ pak získáváme

$$\begin{aligned} F(x,a,b) &= \int_1^{1-x} -f(u;b,a)du, & \text{(A.3.5)}\\ &= \int_{1-x}^1 f(u;b,a)du, & \text{(A.3.6)}\\ &= F(1;b,a) - F(1-x;b,a), & \text{(A.3.7)}\\ &= 1 - F(1-x;b,a). & \text{(A.3.8)} \end{aligned}$$

A.4 Zdrojová data a soubory přiložené na CD

A.4.1 Sešity MS Excel

1. DP_Prehledy.xlsx: Zdrojová data a základní přehledy.
2. DP_Naivni_Sazeni.xlsx: Simulování naivního sázení.
3. DP_AC_ADN.xlsx: Zpracování veličin (ukazatelů) zápasů.
4. Data_pro_matl.xlsx: Pomocný sešit pro přiložené zdrojové kódy.
5. Data_pro_matl2.xlsx: Pomocný sešit pro přiložené zdrojové kódy.

A.4.2 Zdrojové kódy pro MATLAB

1. DP_001.m - DP_004.m: Počítání veličin hrajících týmů.
2. DP_005.m: Program pro vytvoření ukázkové ziskové funkce.
3. DP_006.m: Simulování sázení pro různé hodnoty min. požadovaného zisku.
4. DP_conf01.m: Pro libovolně zadané hodnoty n_i, lze vytvořit obrázky s rozdělením parametrů a 95% konfidenčními množinami.
5. DP_conf02.m: Porovnává dvě možnosti vytváření konfidenčních množin.
6. DP_sim01.m - DP_sim03.m: Programy simulující sázení.
7. DP_fce01.m: Pomocný program, který vrací „krajní rohové“ body konfidenční množiny.