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

BAKALÁŘSKÁ PRÁCE

Analýza sportovních statistik NHL

Vedoucí bakalářské práce:
Mgr. Ondřej Vencálek Ph.D.
Rok odevzdání: 2013

Vypracoval:
Dan Šafařík
AST, III. ročník

Prohlášení

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně pod vedením Mgr. Ondřeje Vencálka Ph.D. a výhradně s použitím uvedené literatury.

V Olomouci dne 21. dubna 2013

Poděkování

Na tomto místě bych velice rád poděkoval mému vedoucímu Mgr. Ondřeji Vencálkovi Ph.D. za všechen čas, který mi věnoval během konzultací a za všechny rady a připomínky, díky kterým se povedlo tuto práci dovést ke zdárnému konci.

Obsah

Úvod **4**

1 NHL **5**
1.1 Jak se boduje 5
1.2 Představení dat 6

2 Samotná analýza **8**
2.1 Základní statistiky 8
2.2 Test výhody domácího prostředí pomocí t-testu 11
2.3 Odds ratio - šance na play off 14
2.4 Bradley-Terry model 19

Závěr **25**

Literatura **26**

Úvod

Cílem mojí práce je ukázat použití statistických metod na reálných datech National Hockey League (NHL), kde se zaměřím na statistiky jednotlivých týmů, ne hráčů. Hlavním smyslem práce bylo tato data analyzovat a především se podívat na vliv domácího prostředí. Tyto analýzy budu provádět výhradně za využití statistického softwaru *R*.

V první kapitole stručně popíši Národní hokejovou ligu, abych čtenáře uvedl do tohoto tématu a měli tak alespoň základní znalosti především o bodování jednotlivých zápasů. Následně představím data, se kterými jsem pracoval. Stěžejní část práce se pak nachází ve třetí kapitole, kde už budou popsány jednotlivé provedené analýzy a na jejich základě vyvozeny různé závěry.

1. NHL

Národní hokejová liga (NHL) je nejprestižnější hokejovou ligou světa, které se účastní týmy z USA a Kanady. V současné době v ní hraje 30 mužstev, které jsou rozděleny do 2 konferencí, východní a západní. Tyto konference jsou nadále rozděleny na celkem 6 divizí, z nichž každá obsahuje 5 týmů.

Sezóna je rozdělena na dvě části. Tou první je základní část, ve které každý tým odehraje 82 zápasů a za každý z nich je náležitě bodově odměněn. Druhou částí je play off, do kterého postupuje 8 nejlepších týmů z každé konference.
V play off se hrají série na 4 vítězná utkání, kdy hraje 1. tým z konference s 8. týmem, 2. se 7. a tak dále. Na konci zbudou dvě družstva (jedno z východní a druhé ze západní konference) a sehrají mezi sebou finálovou sérii opět na 4 vítězné zápasy. Vítězné družstvo získá trofej pro vítěze - Stanley Cup.

Jednotlivé zápasy jsou rozděleny na 3 třetiny po 20 minutách. Pokud i po této době je stav nerozhodný, tak se hraje 5 minutové prodloužení, kde tým, který vstřelí branku, vyhrává. Pokud ani v prodloužení nepadne branka, je čas na nájezdy.

1.1. Jak se boduje

V mé analýze jsem se zaměřil na základní část a proto je velmi důležité pochopit, jak se jednotlivé zápasy v základní části bodují.

Tým získává:

- 2 body za jakoukoliv výhru - v základní hrací době, v prodloužení, po nájezdech
- 1 bod za prohru v prodloužení nebo po nájezdech
- 0 bodů za prohru v základní hrací době

Právě z toho důvodu, že týmy jsou bodově ohodnoceny i v případě prohry, jsem se rozhodl provést analýzy právě s počtem získaných bodů, namísto s počtem výher.

1.2. Představení dat

Data, se kterými jsem pracoval, byla posbírána za posledních 7 sezón od ročníku 2004/2005 do 2011/2012. V sezóně 2003/2004 se totiž NHL kvůli výluce a jednání o nové kolektivní smlouvě nehrála. V tomto roce také došlo k zavedení několika nových pravidel a hlavně platového stropu, který výrazně ovlivnil některé týmy. Z tohoto důvodu jsem se rozhodl data brát až po výluce, kdy se všechny zápasy odehrály za stejných podmínek jen s pár menšími zásahy do pravidel.

Pro většinu provedených analýz byla použita jedna data. Tato datová množina obsahuje 210 prvků a je rozdělena do 3 sloupců. V prvním sloupci jsou názvy jednotlivých týmů, kde se každý opakuje celkem sedmkrát. Druhý sloupec udává počet bodů získaný týmem doma za jednu sezónu a třetí sloupec udává počet bodů získaných týmem venku za jednu sezónu. Z důvodů popsaných výše jsem se rozhodl, že je vhodnější se zaměřit na počet získaných bodů, než na počet výher. Ukázka datové množiny:

NHL	HOME	ROAD
Anaheim	44	36
Anaheim	57	41
Anaheim	54	45
Anaheim	55	34
Anaheim	43	48
Anaheim	60	42
Anaheim	61	49
Atlanta	41	35
Atlanta	52	45
Atlanta	52	38

Druhá datová množina byla vytvořena přímo pro potřeby Bradley - Terryho modelu (viz kapitola 2.4.). Tato množina obsahuje data jen ze sezóny 2011/2012 a figuruje v ní pouze 5 týmů (Severovýchodní divize). Tyto týmy sehrály každý s každým celkem 6 zápasů (3 doma, 3 venku). V prvním sloupci tedy vždy máme domácí tým a ve druhém ten hostující, třetí sloupec pak udává, kolik z těchto vzájemných zápasů vyhrál tým domácí, a poslední sloupec udává, kolik jich vyhrál tým, který hrál venku.

Ukázka datové množiny:

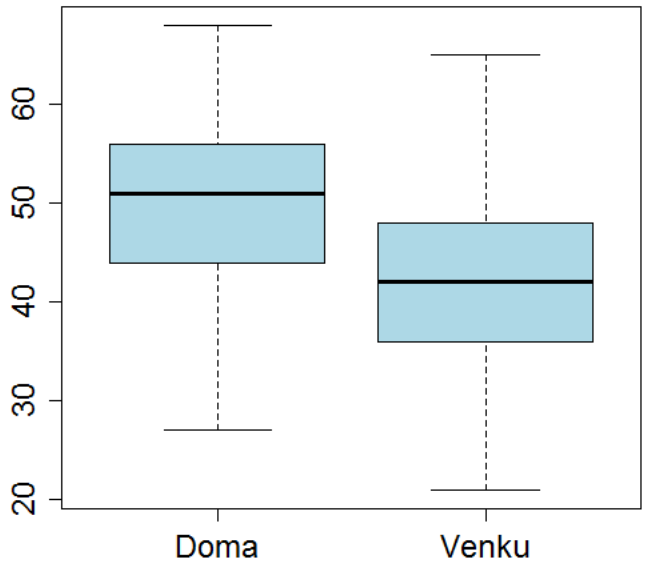
home.team	away.team	home.wins	away.wins
Toronto	Montreal	1	2
Toronto	Boston	0	3
Toronto	Buffalo	3	0
Toronto	Ottawa	1	2
Montreal	Toronto	1	2
Montreal	Boston	1	2
Montreal	Buffalo	0	3
Montreal	Ottawa	2	1
Boston	Toronto	3	0
Boston	Montreal	2	1

2. Samotná analýza

2.1. Základní statistiky

V této úvodní kapitole ještě nebudu aplikovat modely nebo testovat hypotézy, ale podívám se jen na základní statistiky, abych získal obecný přehled o datech. K tomu využiji především grafického znázornění dat. Nejdříve vytvořím boxplot pro počet získaných bodů doma a venku a následně je porovnám. Boxplot vypadá následovně:

Obrázek 1: Boxplot - Počet bodů získaných doma/venku



Vidím, že počty bodů získaných doma obecně nabývají větších hodnot. V následující tabulce jsou pak vidět konkrétní čísla.

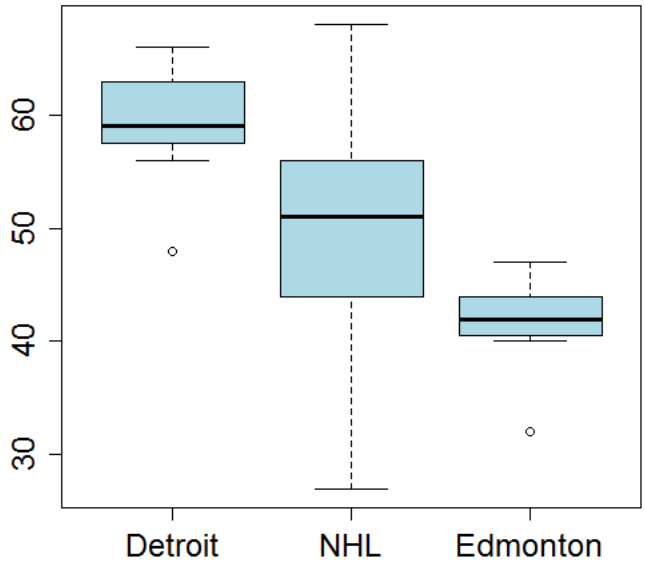
	Doma	Venku
Minimum	27	21
1. kvartil	44	36
Medián	51	42
Průměr	49,96	41,63
3. kvartil	56	48
Maximum	68	65

Pro zajímavost se také můžu podívat, jak si jednotlivé týmy v posledních 7 sezónách vedly. Který tým například získal nejvíc bodů doma a venku. Tyto údaje zapíši pro přehlednost do tabulky. Počty bodů uvedené v tabulce jsou průměry za všech 7 sezón. Nejprve se podívám, jaký tým byl nejlepší a nejhorší na domácím ledě.

	Tým	Počet bodů
Nejlepší doma	Detroit	59,14
Nejhorší doma	Edmonton	41,43

Pro porovnání teď vytvořím 3 boxploty pro počet bodů získaných doma (Detroit, NHL, Edmonton).

Obrázek 2: Boxplot - Počet bodů získaných doma

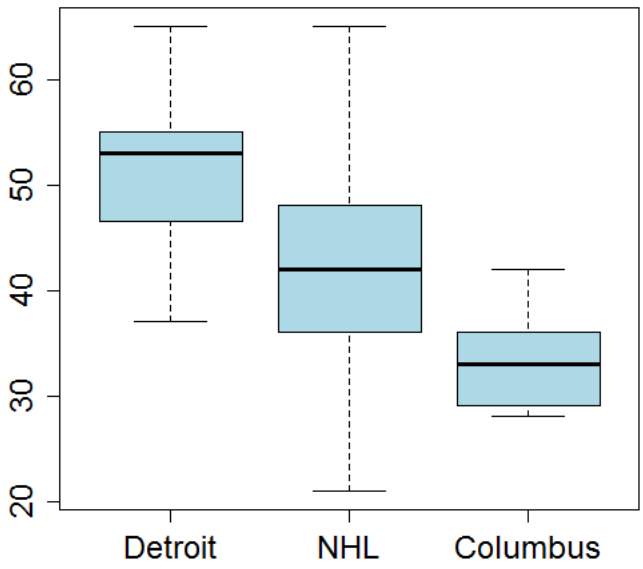


Na boxplotu také můžeme vidět, že Detroit hrál po celých 7 sezón nadprůměrně, jen v jedné sezóně spadl do ligového průměru, Edmonton se na druhou stranu držel celou dobu u ligového dna.

Teď se podívám ještě na výkony týmů na ledech soupeřů.

	Tým	Počet bodů
Nejlepší venku	Detroit	51,14
Nejhorší venku	Columbus	33,29

Obrázek 3: Boxplot - Počet bodů získaných venku



Na první pohled je vidět, že data jsou zde mnohem více rozptýlena než u počtu bodů získaných doma. Opět tu dominuje Detroit, ale i ten se zde v některých sezónách pohyboval na nižších pozicích. Edmonton zde vystřídal Columbus.

V této kapitole jsem získal základní přehled o datové množině a ze zvědavosti se podíval i na to, jak si vedly některé týmy a teď už můžu přejít ke konkrétním analýzám těchto dat.

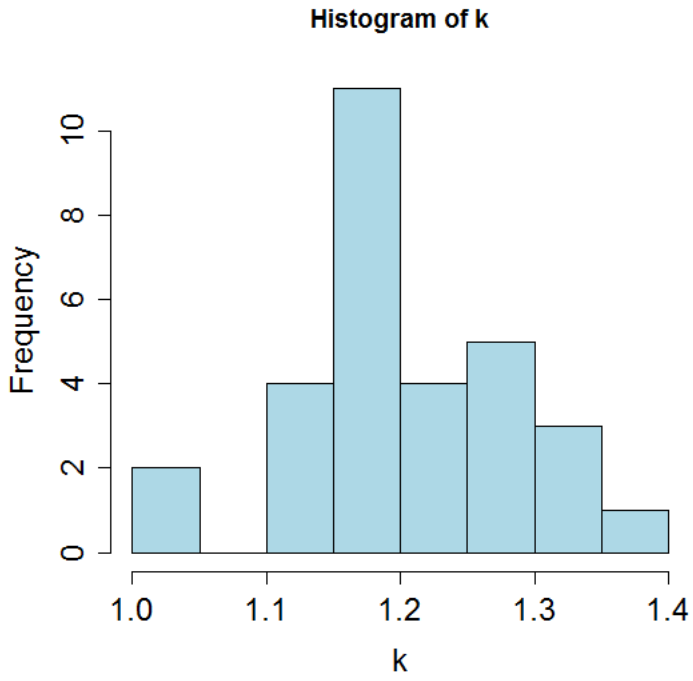
2.2. Test výhody domácího prostředí pomocí t-testu

V prvním příkladu se budu snažit ověřit, zda existuje statisticky významný rozdíl mezi počtem bodů získaných doma a venku. Kvůli závislosti jednotlivých řádků nemůžeme použít dvouvýběrový t-test, který předpokládá nezávislost. Uvažujeme podíl $k = \frac{\mathbb{E}_0}{\mathbb{E}_1}$, kde $\mathbb{E}_0$ je střední počet bodů nějakého týmu doma a $\mathbb{E}_1$ je střední počet bodů nějakého týmu venku. Hodnotu k odhadneme z jejich realizací $k_1, ..., k_{30}$, kde $k_i = \frac{\mathbb{E}_{i0}}{\mathbb{E}_{i1}}$, kde $\mathbb{E}_{i0}$ je průměrný počet bodů i-tého týmu doma a $\mathbb{E}_0$ je průměrný počet bodů i-tého týmu venku. Data nejdříve upravíme tak, že vypočítáme vektor k, který bude mít 30 prvků (30 týmů). Vektor k se vypočítá takto:

$$\frac{\mathbb{E}_{i0}}{\mathbb{E}_{i1}} = k_i \quad \text{pro } i = 1, ..., 30$$

Vektor k je tedy podílem těchto středních hodnot a jednotlivé prvky vektoru už můžu považovat za nezávislé. Hodnoty tohoto vektoru potom považuji za realizace náhodné veličiny, o jejímž rozdělení však nic nevím. Než provedu test, je tedy nutné ověřit, zda data pocházejí z normálního rozdělení, abych věděl, jaký test použít. Nejdřív vykreslím histogram a podívám se, jestli by data
z normálního rozdělení pocházet mohla a potom ještě provedu Shappirův test a normalitu přímo otestuji.

Obrázek 4: Histogram - vektor k



Z histogramu můžeme usoudit, že data normálně rozdělená jsou, abych si ale byl jistý, provedu ještě Shappirův test. Jeho p-value vyjde **0,58**. Na hladině testu $\alpha = 0{,}05$ tedy nelze zamítnout nulovou hypotézu, že data pochází z normálního rozdělení.

Teď už můžu přejít k samotnému testu výhody domácího prostředí, pro který použiji t-test, neboť podmínky pro jeho použití (normální rozdělení, nezávislost) jsou splněny. Hypotézy stanovíme takto.

$$\mathrm{H}_0 : k = 1$$
$$\mathrm{H}_1 : k > 1$$

Nulová hypotéza nám říká, že neexistuje rozdíl mezi počtem bodů získaným doma a venku. Alternativu stanovím pravostrannou, protože nepředpokládám, že by týmy získávaly více bodů venku než doma. Provedu t-test a jeho p-value vyjde $<$ **0,0001**, takže nulovou hypotézu, že týmy získávají stejně bodů doma

i venku, na hladině významnosti $\alpha = 0{,}05$ zamítáme. Tuto hypotézu zamítáme ve prospěch alternativy, že týmy získávají více bodů doma.

Jen ze zvědavosti provedu také Wilcoxonův test, který bych byl nucen použít, pokud bych zamítl hypotézu, že data pocházejí z normálního rozdělení.
I u toho testu vyjde p-value $<$ **0,0001**, takže i zde nulovou hypotézu jednoznačně zamítáme.

Tento výsledek se dal jistě očekávat a není v rozporu s naším očekáváním, že výhoda domácího ledu opravdu existuje.

Pro zajímavost se ještě podívám, jak by dopadly testy zvlášť pro jednotlivé sezóny. V tomto případě už mohu použít přímo párový t-test, protože jednotlivá pozorování jsou v rámci jedné sezóny nezávislá. Data je však potřeba nejprve upravit, protože se v nich údaj o tom, k jaké sezóně daný záznam spadá, neobjevuje. Data jsou však seřazena abecedně podle týmů a pro každou tuto sedmičku i chronologicky, vytvořím tedy nový sloupec, ve kterém se třicetkrát zopakuje sekvence 1,2,...,7.

NHL	HOME	ROAD	SEZ
Anaheim	44	36	1
Anaheim	57	41	2
Anaheim	54	45	3
Anaheim	55	34	4
Anaheim	43	48	5
Anaheim	60	42	6
Anaheim	61	49	7
Atlanta	41	35	1
Atlanta	52	45	2
Atlanta	52	38	3

Nulovou hypotézu pro tyto párové t-testy stanovím takto:

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$
$$\mathrm{H}_1 : \mu_1 - \mu_2 > 0$$

Kde μ_1 je střední hodnota počtu bodů získaných doma a μ_2 udává střední hodnotu počtu bodů získaných venku. Pravostrannou alternativu volím, protože předpokládám, že střední hodnota počtu bodů získaných doma bude větší než střední

hodnota počtu bodů získaných venku, a ne naopak. Hladinu významnosti pak volím $\alpha = 0{,}05$. Výsledky jednotlivých testů a také náležité střední hodnoty zapíši do tabulky:

Sezóna	μ_1	μ_2	p-value	závěr
1	50,33	41,60	$< 0,0001$	zamítám
2	48,23	43,93	$0,0008$	zamítám
3	51,67	41,87	$< 0,0001$	zamítám
4	51,17	40,77	$< 0,0001$	zamítám
5	47,93	41,97	$< 0,0001$	zamítám
6	49,37	41,93	$< 0,0001$	zamítám
7	51,03	39,33	$< 0,0001$	zamítám

Ve všech případech případech nulovou hypotézu výrazně zamítáme a výhoda domácího prostředí se tak projevila ve všech sezónách.

2.3. Odds ratio - šance na play off

V tomto příkladu pro mě bude opět důležitá výhoda domácího prostředí. Tentokrát se však podívám, jak výkony týmů doma ovlivňují jejich šanci na postup do play off. K tomuto využiji kontingenčních tabulek a odhadu poměru šancí (Odds ratio).

Data o tom, jestli tým doopravdy postoupil do play off či ne, jsem k dispozici ve svých datech neměl, takže jsem se rozhodl pro jiný přístup. Z vlastních zkušeností vím, že týmy, které získají za sezónu více jak 90 bodů, mají slušnou šanci dostat se do postupové osmičky. Jednu kategorii tedy stanovenou máme a to, že tým získá více jak 90 bodů oproti tomu, že tým získá do 90 bodů. Druhou kategorií bude, zda tým hrál lépe doma než venku, což v mých datech znamená, že tým v jedné sezóně získal více bodů doma než venku. Kontingenční tabulka pro celou datovou množinu tady vypadá takto:

	v play off	mimo play off
lepší doma	95	80
horší doma	24	11

Pomocí vzorce spočítáme Odds ratio:

$$\widehat{OR} = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}.$$

Po dosazení

$$\widehat{OR} = \frac{95 \cdot 11}{80 \cdot 24} = \mathbf{0{,}54}.$$

Z tohoto výsledku vidím, že týmy, které hrají lépe venku, mají téměř o polovinu menší šanci, že nepostoupí do play off oproti týmům, které hrály lépe doma. Než se však pokusím tento výsledek nějak interpretovat, je potřeba si uvědomit, že tento postup není zrovna vhodný. Jak jsme si řekli v prvním příkladu, jednotlivá pozorování nejsou nezávislá. Mnohem vhodnějším postupem tedy je spočítat Mantel-Haenszel Odds ratio.

Vzorec pro Mantel-Haenszelův odhad odds ratia vypadá takto:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k} \frac{a_i d_i}{n_i}}{\sum_{i=1}^{k} \frac{b_i c_i}{n_i}}.$$

Kde k udává počet jednotlivých kontingenčních tabulek, v mém případě to je 7, protože mám 7 sezón. Informaci o tom, k jaké sezóně daný záznam spadá už máme díky úpravě dat v příkladu 2.2..

NHL	HOME	ROAD	SEZ
Anaheim	44	36	1
Anaheim	57	41	2
Anaheim	54	45	3
Anaheim	55	34	4
Anaheim	43	48	5
Anaheim	60	42	6
Anaheim	61	49	7
Atlanta	41	35	1
Atlanta	52	45	2
Atlanta	52	38	3

Můžu tedy vytvořit 7 kontingenčních tabulek, pro každou sezónu jednu, podle stejných pravidel jako předtím. Tímto dostaneme jednotlivé hodnoty pro výpočet Mantel-Haneszelova odds ratia. Vzorec ještě můžeme upravit pro snazší výpočet tak, že ze sumy vytkneme n_i a pokrátíme, neboť n_i je ve všech kontingenčních tabulkách stejné, a to 30.

Upravený vzorec:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k} a_i d_i}{\sum_{i=1}^{k} b_i c_i}.$$

Dosadím do vzorce a výsledné odds ratio vyjde **0,584**. Tento výsledek je hodně podobný tomu, který vyšel v prvním případě.

Pro přesnost tohoto odhadu je potřeba určit ještě intervalový odhad. Protože by tento intervalový odhad byl pro Mantel-Heanszlovo odds ratio poměrně složitý, tak ho vypočítám pro původní odhad, jelikož výsledky si jsou dosti podobné.

Vzorec pro intervalový odhad vypadá takto:

$$\left(\widehat{OR} \cdot \exp\left\{ -u_{1-\alpha/2} \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right\} ; \widehat{OR} \cdot \exp\left\{ u_{1-\alpha/2} \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right\} \right)$$

$\widehat{OR} = 0,54\ldots$ bodový odhad odds ratia

$u_{1-\alpha/2}\ldots$ $1-\alpha/2$ kvantil normálního normovaného rozdělení; v mém případě volím $\alpha = 0{,}05$, takže kvantil $u_{0,975}$ je roven **1,96**

$a, b, c, d\ldots$ jednotlivé hodnoty z původní kontingenční tabulky pro všech 7 sezón; $a = 95, b = 80, c = 24, d = 11$

Po dosazení těchto hodnot dostávám:

$$\left(0,54 \cdot \exp\left\{ -1,96 \cdot \sqrt{\frac{1}{95} + \frac{1}{80} + \frac{1}{24} + \frac{1}{11}} \right\} ; 0,54 \cdot \exp\left\{ 1,96 \cdot \sqrt{\frac{1}{95} + \frac{1}{80} + \frac{1}{24} + \frac{1}{11}} \right\} \right)$$

$$(0{,}2512; 1{,}1792)$$

Pokud bych nevypočítal intervalový odhad, tak bych mohl usoudit, že týmy, které hrají lépe venku, mají výrazně větší šanci na postup do play off. V předchozím příkladě jsem ověřil výhodu domácího prostředí a tedy tendenci týmů získávat více bodů doma. K postupu do play off je však důležité vyhrávat i ty těžké zápasy na ledě soupeřů. Nepodařilo se prokázat, že by tento závěr byl správný, neboť nám intervalový odhad pokrývá 1, takže nemůžeme říct, že by lepší hra venku zvyšovala šanci na postup do play off.

Intervalový odhad OR je však poměrně široký, tak se pro zajímavost zkusím ještě podívat, jak by vypadaly odhady OR pro jednotlivé sezóny, abych zjistil, jak se hodnoty odhadu poměru šancí v sezónách vyvíjely a jestli v některém případě nevyšel tento odhad 0. Výsledky zapíšu do tabulky:

Sezóna	$\widehat{OR}$
1	0,00
2	0,52
3	2,80
4	0,00
5	2,30
6	1,41
7	0,00

Dokonce ve 3 případech vyšel odhad Odds ratia 0, ve 2 případech nad 2 a pouze v jednom případě se podobá celkovému odhadu Odds ratia pro všech 7 sezón. Problémovou hodnotu v těchto kontingenčních tabulkách je kolonka d, která udává, kolik týmů hrálo lépe venku a přitom by i tak skončily mimo play off (celkový počet bodů menší jak 90). Například pro případ první sezóny vypadá kontingenční tabulka takhle:

	v play off	mimo play off
lepší doma	14	13
horší doma	3	0

Tyto nulové hodnoty ve 3 kontingenčních tabulkách sice při výpočtu Mantel-Haenszleova poměru šancí tolik nevadí, protože to je postup, který se v případě výskytu nulových hodnot používá, ale jistě ten odhad nějak ovlivňují, protože

součiny $a_i \cdot d_i$ pro $i = 1, 4, 7$ budou nulové. Alternativním přístupem, jak si poradit s nulovými buňkami, je ke všem buňkám přičíst určitou konstantu a následně vypočítat odhad adjustovaného odhadu šancí (adjusted odds ratio). Rozhodl jsem se tedy ke všem buňkám přičíst hodnotu 1 a $\widehat{aOR}$ vypočítat. Vzorec vypadá následovně:

$$\widehat{aOR} = \exp\left\{\sum_{i=1}^{k} \frac{w_i}{\sum_{j=1}^{k} w_j} \ln\left(\widehat{OR_i}\right)\right\},$$
$$\text{kde } w_i = \frac{1}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}.$$

Zde je vidět důvod, proč je nutné ke všem buňkám přičíst nějakou hodnotu, jinak bych totiž měl ve vzorci pro jednotlivé váhy w_i ve jmenovateli nulu. Já se rozhodl ke všem buňkám d_i přičíst hodnotu 1 a s takto upravenými tabulkami vypočítat $\widehat{aOR}$. Po dosazení do vzorce dostanu výsledek **0,97**. Tento výsledek se výrazně liší od předem vypočtených odhadů poměrů šancí, a to jsem ke každé buňce přičetl pouze hodnotu 1. Tento výsledek mi říká, že šance na play off lepší hra doma nijak neovlivňuje. Zkusím teď $\widehat{aOR}$ vypočítat ještě jednou, ale tentokrát přičtu jen hodnotu 0,5, což je hodnota, která se v takové situaci přičítá nejčastěji. Vzorec zůstává stejný a výsledek je **0,81**. Tato hodnota sice už je blíže předtím vypočítaným odhadům, ale i tak se stále dost liší. Stačí jen malý zásah do dat a výsledky se hned výrazně změní. To mi naznačuje, že výsledky jsou dosti nespolehlivé, což jsem mohl vyčíst i z vypočítaného intervalového odhadu, který je dosti široký.

Ještě pro zajímavost vypočítám intervalový odhad i pro $\widehat{aOR}$ s přičtenou hodnotou 0,5, protože mě zajímá, jestli i zde bude pokrývat 1, můj předpoklad však je, že i zde tomu tak bude. Vzorec pro intervalový odhad $\widehat{aOR}$ vypadá takto:

$$\left(\widehat{aOR}\ \exp\left\{-u_{1-\alpha/2}\frac{1}{\sqrt{\sum w_j}}\right\}; \widehat{aOR}\ \exp\left\{-u_{1-\alpha/2}\frac{1}{\sqrt{\sum w_j}}\right\}\right)$$
$$(0{,}37; 1{,}78)$$

Jak jsem předpokládal, i zde intervalový odhad pokrývá hodnotu 1. Žádným z použitých postupů se mi tedy nepovedlo ověřit asociaci toho, že by lepší výkony doma, či venku, dávaly týmům větší šanci na postup do play off.

2.4. Bradley-Terry model

Bradley-Terryho model patří k těm méně známým, a proto jsem se ho rozhodl v úvodu téhle kapitoly stručně popsat. Tento model předpokládá, že v nějaké "soutěži" mezi dvěma "hráči", řekněme, že s hráči i a j $(i, j \in \{1, , ..., K\})$, je šance, že hráč i porazí hráče j α_i/α_j, kde α_i a α_j jsou kladné parametry, které mohou být chápany jako určitá "schopnost". Model se dá také napsat v alternativním tvaru:

$$logit\left[(i \ \text{ porazí } j)\right] = \lambda_i - \lambda_j,$$

kde $\lambda_i = \log \alpha_i \quad \forall i$.

V mém případě, kdy budu odhadovat efekt domácího prostředí, je potřeba uvažovat ještě upravený model, který zahrnuje efekt pořadí (order effect), který v mém případě je právě zmiňovaná výhoda. Po zahrnutí efektu pořadí vypadá model následovně:

$$logit\left[(i \ \text{ porazí } j)\right] = \lambda_i - \lambda_j + \delta_z,$$

kde $z = 1$ pokud i má danou výhodu a $z = -1$ pokud ji má j. (Pokud "výhoda" je ve skutečnosti nevýhoda, tak bude δ záporné.) Hodnoty λ_i potom vyjadřují schopnost při absenci jakékoliv výhody.

V mém případě se opět podívám na výhodu domácího prostředí, tentokrát však s využitím výše popsaného Bradley-Terryho modelu. Pro tento příklad využiji jinou datovou množinu, ve které mám informaci o vzájemných zápasech týmů a jejich počtu výher doma a venku. Data pocházejí ze sezóny 2011/2012 ze Severovýchodní divize. Data už jsem představil dříve a vypadají tedy takto:

home.team	away.team	home.wins	away.wins
Toronto	Montreal	1	2
Toronto	Boston	0	3
Toronto	Buffalo	3	0
Toronto	Ottawa	1	2
Montreal	Toronto	1	2
Montreal	Boston	1	2
Montreal	Buffalo	0	3
Montreal	Ottawa	2	1
Boston	Toronto	3	0
Boston	Montreal	2	1

Vidím, že například Toronto odehrálo 3 domácí zápasy proti Bostonu, z nichž se jim však nepodařilo ani jeden vyhrát a všechny 3 výhry tak vybojoval hostující tým z Bostonu, proto máme v tabulce pro Toronto doma a Boston venku skóre 0-3. I Boston hrál na svém domácím ledě proti Torontu 3 zápasy, ten ale dokázal všechna 3 utkání proměnit ve vítězství a proto skóre 3-0.

Po aplikaci Bradley-Terryho modelu, který v sobě zahrnuje efekt domácího prostředí, mi koeficient pro tento efekt vyjde at.home $= -7{,}890e^{-13}$. Po přepočtení $\exp(-7{,}890e^{-13}) = \mathbf{1}$. Toto je poměrně zvláštní výsledek a je naprosto v rozporu se závěrem, který jsem učinil v prvním příkladu. Podle Bradley-Terryho modelu totiž žádná výhoda domácího prostředí neexistuje. Tento výsledek ukazuje, že šance na výhru pro domácí i hostující tým je stejná. Abych pochopil, co stojí za tímto zvláštním výsledkem, tak data upravím do přehlednější tabulky a více je prozkoumám. Po úpravě vypadají data takto:

	away.team				
home.team	Toronto	Montreal	Boston	Buffalo	Ottawa
Toronto	-	1-2	0-3	3-0	1-2
Montreal	1-2	-	1-2	0-3	2-1
Boston	3-0	2-1	-	3-0	2-1
Buffalo	3-0	2-1	2-1	-	1-2
Ottawa	1-2	1-2	0-3	1-2	-

Teď provedu součet pro jednotlivé řádky a sloupce. Pokud má například Toronto doma proti Montrealu výsledek $1-2$, dostaneme z toho pro Toronto hodnotu -1 doma. Takhle sečteme hodnoty v řádku pro všechny týmy. Stejným způsobem

sečteme pro všechny týmy i hodnoty ve sloupcích, jen výsledek vynásobíme -1, abychom dostali jejich výsledky pro venkovní zápasy. Všechny hodnoty v původní tabulce jsou totiž brány z pohledu domácího týmu.

	away.team					
home.team	Toronto	Montreal	Boston	Buffalo	Ottawa	Součet
Toronto	-	-1	-3	3	-1	-2
Montreal	-1	-	-1	-3	1	-4
Boston	3	1	-	3	1	8
Buffalo	3	1	1	-	-1	4
Ottawa	-1	-1	-3	-1	-	-6
Součet	-4	0	6	-2	0	

Hned na první pohled je vidět, že pouze 2 týmy, Boston a Buffalo, získaly v domácích zápasech více jak polovinu možných bodů, zbylé 3 týmy hrály doma podprůměrně. Pokud se teď podívám na výsledky doma a venku jednotlivých týmů, tak dostanu:

$$\text{Doma: } -2 - 4 + 8 + 4 - 6 = 0$$
$$\text{Venku: } -4 + 0 + 6 - 2 + 0 = 0$$

Zde vidím, že rozdíl mezi mezi počtem bodů získaných doma a venku, doopravdy není žádný rozdíl, a proto mi efekt pro domácí prostředí v Bradley-Terryho modelu vyšel 1.

Zkusím se teď podívat, jak by vypadaly tyto upravené tabulky pro 2 extrémní situace, v jednom případě budu uvažovat, že týmy vyhrály všechny domácí zápasy a v druhém případě, že týmy nevyhrály jediný domácí zápas. V prvním případě bude upravená tabulka vypadat takto:

	away.team					
home.team	Toronto	Montreal	Boston	Buffalo	Ottawa	Součet
Toronto	-	3	3	3	3	12
Montreal	3	-	3	3	3	12
Boston	3	3	-	3	3	12
Buffalo	3	3	3	-	3	12
Ottawa	3	3	3	3	-	12
Součet	-12	-12	-12	-12	-12	

Pokud hodnoty v součtu vynásobím 2, tak dostanu počet bodů, které jednotlivé týmy získaly doma/venku. V tomto případě každý tým získal 24 z 24 možných bodů doma, zato venku nezískaly jediný bod. Pokud na tento ilustrativní případ aplikuji Bradley-Terryho model, tak ten odhadne efekt domácího prostředí na $2{,}3 \cdot 10^{11}$.

Podívám se teď tedy ještě na opačný případ, kdy týmy nevyhrály jediný zápas doma. Tabulka bude v této situaci vypadat takto:

	away.team					
home.team	Toronto	Montreal	Boston	Buffalo	Ottawa	Součet
Toronto	-	-3	-3	-3	-3	-12
Montreal	-3	-	-3	-3	-3	-12
Boston	-3	-3	-	-3	-3	-12
Buffalo	-3	-3	-3	-	-3	-12
Ottawa	-3	-3	-3	-3	-	-12
Součet	12	12	12	12	12	

Zde všechny týmy získaly z domácích zápasů 0 bodů a v těch venkovních jich získaly 24. Ve výsledku má každý tým stejný bodový zisk jako v předchozím případě, jen zde pocházejí všechny získané body ze zápasů na ledech soupeřů. Bradley-Terryho model tu odhadne efekt domácího prostředí na $-2{,}3 \cdot 10^{11}$.

Uvedl jsem zde tedy 3 ilustrativní příklady, z nichž 2 jsem pro představu, jak by tato situace vypadala, nasimuloval a ten jeden, kdy týmy hrály stejně dobře doma i venku, se objevil na reálných datech z NHL. Za tento případ může buď náhoda, nebo také to, že jsem měl data pouze za jednu sezónu, což je poměrně malý vzorek. Z tohoto důvodu zde uvedu ještě další 2 příklady. V jednom z nich se podívám opět na data z té samé sezóny, ale použiji jinou divizi. V druhém příkladě potom doplním současná data ze severovýchodní divize o další 2 předešlé sezóny, tím budu mít mnohem více pozorování a výsledek tak bude důvěryhodnější.

Podívám se tedy nejdřív na data ze stejné sezóny 2011-2012, ale tentokrát vezmu jinou divizi, a to Centrální.

Tabulka vypadá takto:

	away.team				
home.team	Chicago	St.Louis	Detroit	Columbus	Nashville
Chicago	-	3-0	2-1	3-0	1-2
St.Louis	2-1	-	2-1	2-1	1-2
Detroit	1-2	3-0	-	3-0	2-1
Columbus	1-2	1-2	2-1	-	0-3
Nashville	2-1	2-1	2-1	2-1	-

Po úpravě tabulky a součtu hodnot v řádcích a sloupcích dostanu:

	away.team					
home.team	Chicago	St.Louis	Detroit	Columbus	Nashville	Součet
Chicago	-	3	1	3	-1	6
St.Louis	1	-	1	1	-1	2
Detroit	-1	3	-	3	1	6
Columbus	-1	-1	1	-	-3	-4
Nashville	1	1	1	1	-	4
Součet	0	-6	-4	-8	4	

Opět je zde hned na první pohled vidět výrazný rozdíl oproti Severovýchodní divizi. V Centrální divizi totiž hned 4 týmy hrály dobře na domácím ledě a jen Nashville hrál hůře. Z toho můžu usoudit, že zde se už efekt domácího ledu projeví a podívám se tedy, jak ho odhadne Bradley-Terryho model. Ten ho odhaduje na **1,76**. Tento výsledek mi potvrzuje závěr z příkladu 2.2., že výhoda domácího ledu opravdu existuje a výsledek ze Severovýchodní divize je pouze anomálie, tu se teď pokusím odstranit a získat tak přesnější výsledek tím, že k původním datům přidám ještě data z předchozích dvou sezón a opět se podívám na tabulku a na Bradley-Terryho odhad domácího efektu.

Beru tedy data ze 3 sezón od roku 2009 do roku 2012 ze Severovýchodní divize. Tabulka bude vypadat takto:

	away.team				
home.team	Toronto	Montreal	Boston	Buffalo	Ottawa
Toronto	-	4-5	4-5	6-3	4-5
Montreal	4-5	-	6-3	2-7	4-5
Boston	7-2	4-5	-	6-3	5-4
Buffalo	8-1	5-4	5-4	-	3-6
Ottawa	3-6	4-5	0-9	3-6	-

Po úpravě:

	away.team					
home.team	Toronto	Montreal	Boston	Buffalo	Ottawa	Součet
Toronto	-	-1	-1	3	-1	0
Montreal	-1	-	3	-5	-1	-4
Boston	5	-1	-	3	1	8
Buffalo	7	1	1	-	-3	6
Ottawa	-3	-1	-9	-3	-	-16
Součet	-8	2	6	2	4	

Je zde vidět, že jsou opět jen 2 týmy, které hrály nadprůměrně doma (Boston, Buffalo). Pokud se podívám, jak efekt domácího prostředí tentokrát odhadne Bradley-Terryho model, tak dostanu hodnotu **0,93**, což je ještě menší hodnota, než mi vyšla pro jednu sezónu ze Severovýchodní divize. Tento výsledek dokonce ukazuje, že týmy mají tendenci hrát lépe venku (v této divizi). Pokud se však podrobněji podívám na tato data, tak vidím zajímavé výsledky Ottawy. Ta má totiž doma skóre 10-26 a nedokázala tak doma získat ani proti jednomu z týmů alespoň polovinu možných bodů. Venku však Ottawa hrála nad očekávání dobře s výsledným skóre 20-16. Můžu tedy usoudit, že právě Ottawa má na výsledek velký vliv. Zkusím ji teď z dat odebrat a provést analýzu pouze pro zbylou čtveřici týmů.

	away.team			
home.team	Toronto	Montreal	Boston	Buffalo
Toronto	-	4-5	4-5	6-3
Montreal	4-5	-	6-3	2-7
Boston	7-2	4-5	-	6-3
Buffalo	8-1	5-4	5-4	-

Pokud teď provedu odhad efektu domácího prostředí pomocí Bradley-Terryho modelu, tak vyjde hodnota **1,31**, což je výsledek, který mnohem lépe odpovídá předpokladu, že domácí výhoda existuje. Výsledky Ottawy tedy výslednou hodnotu efektu výrazně ovlivnily.

Závěr

V úvodu práce jsem čtenáře seznámil se základními pojmy a znalostmi o NHL tak, aby byl schopen lépe porozumět další práci s daty a významu jednotlivých analýz a testů. Nadále jsem představil použité datové množiny, aby měl čtenář lepší představu, s jakými daty jsem pracoval a jaké proměnné v nich vystupují.

Hlavní část mé práce se pak odehrává ve druhé kapitole, kde jsou postupně popsány všechny provedené analýzy. V úvodu této kapitoly jsou představeny základní statistiky dat i s grafickým znázorněním pro snazší přehled a představu o datech. Hlavním cílem bylo podívat se na vliv domácího prostředí, jehož pozitivní vliv se mi povedlo hned v prvním příkladu potvrdit pomocí t-testu pro všechna data a také pomocí párových t-testů pro jednotlivé sezóny zvlášť. Když jsem ověřil výhodu domácího prostředí, tak jsem se chtěl v dalším příkladě podívat na to, jestli má lepší hra doma nějaký vliv na šanci postupu do play off, k tomuto jsem využil odhad poměru šancí. Zde byly výsledky hodně rozporuplné a odhady nejisté. Ve výsledku však můžu říct, že mezi lepší hrou doma a tím rostoucí šancí na play off neexistuje statisticky významná asociace. Nejrozsáhlejším příkladem pak byla analýza pomocí Bradley-Terryho modelu. Tento model není tak známý, a tak byl na úvod nejdříve stručně popsán. Tento model byl nadále aplikován na řadě reálných dat z NHL a také pár ilustrovaných příkladů. Cílem bylo opět ověřit efekt domácího prostředí, na jedné datové množině nebyl problém domácí výhody ověřit, na další se však zprvu jevilo, že mezi hrou doma ani venku žádný rozdíl není, po doplnění této datové množiny o další data z předešlých let pak Bradley-Terry model dokonce odhadl, že hrát doma je nevýhodou a větší šanci na výhru tak mají týmy hrající venku.

Cílem práce bylo ukázat aplikování statistických metod na reálných datech a jejich práci s nimi, kdy mám stanovenou nějakou otázku, či problém, který chci zodpovědět a musím tak hledat správný postup a metodu k jeho vyřešení.

Literatura

[1] Anděl, J.: *Matematická statistika*, SNTL/Alfa, 1978

[2] Kunderová, P.: *Základy pravděpodobnosti a matematické statistiky*, Univerzita Palackého v Olomouci, 2004

[3] Bradley-Terry Models in R: The BradleyTerry2 Package, http://cran.r-project.org/web/packages/BradleyTerry2/vignettes/BradleyTerry.pdf

[4] National hockey league, http://www.nhl.com

[5] ESPN NHL, http://espn.go.com/nhl/standings