VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA STROJNÍHO INŽENÝRSTVÍ
FACULTY OF MECHANICAL ENGINEERING

ÚSTAV MATEMATIKY
INSTITUTE OF MATHEMATICS

ANALÝZA SPORTOVNÍCH DAT
SPORTS DATA ANALYSIS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — **Patrik Horák**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Pavel Hrabec, Ph.D.**
SUPERVISOR

BRNO 2021

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA STROJNÍHO INŽENÝRSTVÍ

Zadaní bakalářské práce

Ústav:	Ústav matematiky
Student:	**Patrik Horák**
Studijní program:	Aplikované vědy v inženýrství
Studijní obor:	Matematické inženýrství
Vedoucí práce:	**Ing. Pavel Hrabec, Ph.D.**
Akademický rok:	2020/21

Ředitel ústavu Vám v souladu se zákonem č.111/1998 o vysokých školách a se Studijním a zkušebním řádem VUT v Brně určuje následující téma bakalářské práce:

Analýza sportovních dat

Stručná charakteristika problematiky úkolu:

Rostoucí dostupnost dat, vyšší výpočetní výkon a všeobecný posun k rozhodování na základě dat se v posledních letech promítá i do oblasti sportu. Vzhledem ke zmíněným možnostem je ale snaha postihnout i ty aspekty hry, které nejsou běžnými statistikami (ve fotbale např. počet střel, střel na branku, podíl držení míče, …) popsatelné. Komplexnější otázky jako “Jak kvalitní šance si tým tvoří?”, “Jaký přínos mají pro tým jednotliví hráči?”,”Dělají hráči dobrá rozhodnutí?” apod. se snaží zodpovědět tzv. pokročilé statistiky (expected goals, expected threat, VAEP, ...). Pokročilé statistiky (resp. jejich modely) vznikají na základě “událostních” či automaticky sbíraných dat, přičemž zejména automaticky sbíraná data jsou záležitostí pouze posledních několika let. Toto omezení společně s nedostupností dat pro širokou veřejnost bylo problémem v rozvoji datové analytiky ve sportu. V nedávné době však došlo k uvolnění vzorků takových dat různými společnostmi, což otevírá prostor pro jejich analyzování.

Cíle bakalářské práce:

1) Rešerše modelů ze zvolené oblasti.
2) Osvojení základních postupů analýzy dat (průzkumová analýza, vizualizace).
3) Seznámení studenta s vhodnými nástroji pro analýzu dat (R, Python, …).
4) Analýza dostupných dat z oblasti sportu.

Seznam doporučené literatury:

ANDĚL, Jiří. Základy matematické statistiky. Vyd. 3. Praha: Matfyzpress, 2011. ISBN 978-80-73-8-162-0.

JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. An introduction to statistical learning: with applications in R. New York: Springer, 2013. Springer texts in statistics. ISBN 978-1461471370.

EFRON, Bradley a Trevor HASTIE. Computer age statistical inference: algorithms, evidence, and data science. New York, NY: Cambridge University Press, 2016. ISBN 978-1107149892.

SHEA, Stephen a Christopher BAKER. Hockey analytics: A game-changing perspective. United States: Advanced Metrics, 2017. ISBN 978-1977533494.

Online kurz “Mathematical Modelling of Football” (University of Uppsala)

Termín odevzdání bakalářské práce je stanoven časovým plánem akademického roku 2020/21

V Brně, dne

L. S.

prof. RNDr. Josef Šlapal, CSc. ředitel ústavu	doc. Ing. Jaroslav Katolický, Ph.D. děkan fakulty

Abstrakt

Táto bakalárska práca sa zaoberá štatistickou analýzou dát z hokejových zápasov. Prvá časť popisuje základné a pokročilé hokejové štatistiky. Druhá časť obsahuje vybrané porovnateľné štúdie, pracujúce s podobnými hypotézami. Tretia časť zoznamuje s matematickými aparátmi použitými v tejto práci, ako Dvojvýberový t-test, korelácia, Dvojfaktorová analýza rozptylu bez interakcií a ich neparamtrické ekvivalenty. Posledná časť sa zaoberá aplikáciou popísaného aparátu na dáta z hokeja a interpretáciou výsledkov.

Abstract

The main preoccupation of this thesis is analysis of ice hockey statistics. The first section describes widely used elementary and advanced ice hockey statistics. The second part contains selected comparable studies working with similar hypotheses. The third section introduces necessary mathematical tools, such as, Two-sample t-test, correlation, Two-way analysis of variance with no interaction and their nonparametric equivalents. In the last section described statistical hypotheses tests are applied to ice hockey dataset. Results interpretation is also included.

kľúčové slová

hokej, hokejové štatistiky, testovanie hypotéz, parametrické štatistické metódy, neparametrické štatistické metódy

keywords

ice hockey, ice hockey statistics, hypothesis testing, parametric statistical methods, nonparametric statistical method

Horák, P. *Analýza sportových dat.* Brno: Vysoké učení technické v Brně, Fakulta strojního inženírství, 2021. 41 s. Vedúci bakalárskej práce Ing. Pavel Hrabec, Ph.D.

Prehlasujem, že som bakalársku prácu *Analýza sportovních dat* vypracoval samostatne pod vedením Ing. Pavla Hrabca, Ph.D. a Ing. Martina Roseckého s použitím materiálov uvedených v zozname literatúry.

Patrik Horák

Chcel by som poďakovať Ing. Pavlovi Hrabcovi Ph.D. a Ing. Martinovi Roseckému za odborné vedenie mojej bakalárskej práce, za ich ochotu a rady, ktoré mi pri písaní veľmi pomohli. Ďalej by som sa chcel poďakovať svojím kolegom z časomiery HK Skalica, menovite Jaroslavovi Bednárikovi za predané skúsenosti a vedomosti v oblasti hokejovej štatistiky. Taktiež by som sa chcel poďakovať svojím blízkym spolužiakom za nezabudnuteľné spomienky a zážitky popri vysokoškolskom stúdiu.

Patrik Horák

Obsah

Úvod

V dnešnej modernej pretechnizovanej dobe má významné miesto vo všetkých dôležitých odvetviach analýza. Z analýzy sa čerpajú informácie pre ďalší vývoj a postupy. Nevyhol sa jej ani šport, kde sa stala významným prvkom pri dosahovaní lepších výkonnostných výsledkov. Vývoju a pokroku sa nevyhol ani hokej. Od zábavného koníčka sa z hokeja stal profesionálny šprot. Hráči sa začali zdokonalovať, venovať sa tréningom, pilovať svoje zručnosti a pracovať na svojich výkonoch. Výrazne k tomu prispelo založenie prvých profesionálnych hokejových líg najmä v Kanade a USA. S rastúcou konkurenciou trénery začali viacej analyzovať správanie svojich zverencov čo viedlo k zapisovaniu a následnemu skúmaniu rôznych štatistík. Konkrétne v *National Hockey League* (NHL) sa začali štatistiky postupne zbierať v 60. rokoch minulého storočia. [15]

Ľadový hokej naberal na popularite medzi širokou verejnosťou. Dopomohlo k tomu aj živé komentovanie zápasov v rádiách a prenosy v televíziách. Fanúšikovia sa začali viacej zaujímať štatistikami svojich obľúbených tímov, ale aj ich súperov. Hlavným zdrojom informácií boli novinári, ktorí zapisovali štatistiky do športových časopisov a článkov. S príchodom kurzových stávok sa tento záujem rapídne zvýšil. Fanúšikovia bedlivo pozorovali štatistiky tímov, a snažili sa viacej analyzovať a predpovedať výsledky zápasov. Dnes sú všetky informácie dostupné pre každého na inernete.

S príchodom väčších peňazí do hokeja, sponzory začali klásť vyššie nároky ohľadom umiestnenia v tabulke a hlavne na výsledky tímov. Vďaka vyšším nárokom sa začalo ešte podrobnejšie skúmanie individuálnych ale aj tímových štatistík. K týmto úkonom už si začali tímy pozývať odborníkov a najmä štatistikov. Tí na základe výsledkov z ich pozorovaní, pomáhali tímom pri ich výbere hráčov, ale aj napríklad pri voľbe taktiky na ďalší zápas. Štatistici začali s dátami od tímov postupne pracovať, a rozhodli sa otestovať rôzne faktory hry, ktoré by mohli ovplyvňovať výsledky a výkony tímov.

Táto práca poukáže na zopár zásadných faktorov ovplyvňujúcich hokejové zápasy, vzťahy rozličných štatistík a v neposlednom rade na vplyv faktorov na aktivitu tímov.

1 Hokejové štatistiky

V tejto časti budú najprv predstavené a popísané základné pojmy ako sú: góly, strely, vhadzovania a vylúčenia. Ďalej bude poukázané aj na pokročilé štatistiky v hokeji ako je Corsiho a Fenwickovo číslo, ktoré budú v tejto práci využité.

1.1 Základné hokejové štatistiky

1.1.1 Gól

Gól (***G***) je najdôležitejšia merná jednotka v hokeji, ktorou sa určuje výsledok a následne aj víťaz zápasu. Tím skóruje (strelí gól) vtedy, ak v súlade s pravidlami ľadového hokeja o strelení gólu viď str. 64 [6], dostane puk celým objemom za bránkovú čiaru v rámci koštrukcie brány.

1.1.2 Strela na bránu

Pod pojmom *strela na bránu* (***S***) sa rozumie také ohrozenie brány, kde puk vyslaný hráčom smeruje medzi konštrukciu brány. Strela následne končí gólom alebo zákrokom brankára. Ak strela vyslaná na bránu zmení svoju trajektóriu v dôsledku stretu s hráčom, alebo jeho výstrojou, hovorí sa o takzvanej *zblokovanej strele* (***BS***). V prípade, ak strela nemieri do brány, ale do konštrukcie brány alebo mimo nej, hovorí sa o *strele mimo brány* (***MS***).

1.1.3 Vhadzovanie

Metóda využívaná na zahájenie hry pri začiatku zápasu, tretiny alebo prerušenej hry sa nazýva *vhadzovanie* (***FO***). Uskutočňuje sa tak, že tímy sa postavia na vhadzovanie oproti sebe. Z každého tímu jeden hráč ide do stredu k bodu na vhadzovanie, kde následne rozhodca vhodí medzi hokejky puk. Hráči sa pokúšajú dostať puk na svoju stranu, a tým vyhrať vhadzovanie pre svoj tím.

1.1.4 Vylúčenia

V prípade, že sa hráč previní proti pravidlám ľadového hokeja viď str. 70-113 [6], rozhodca udeľuje hráčovi trest na základe závažnosti priestupku, a vylučuje ho mimo hry. Podľa toho delíme tresty do troch kategórií: nižšie, vyššie a osobné. Nižší trest je udelený hráčovy na dobu 2 minút, vyšší na dobu 5 minút a osobný trest je vo výške 10 minút. V prípade vylúčenia hráča do konca zápasu sa do štatistiky uvádza dĺžka trestu 20 minút. V tejto práci budú použité *vylúčenia* (***P***), ako súčet všetkých trestných minút udelených pre daný tím za jeden zápas.

1.2 Pokročilé hokejové štatistiky

1.2.1 Corsiho číslo

Corsiho číslo (***C***) je pokročilá hokejová štatistika, popísaná Timom Barnesom, ktorá nesie meno po Talianskom brankárovi Jimovi Corsimu. Je používaná na meranie rozdielu streleckých pokusov, ak je na ľadovej ploche rovnaký počet hráčov. Existujú dva druhy *Corsiho čísla* jeden je individuálny a druhý tímový. Práca sa bude zaoberať len tímovým. ***C*** v sebe zahŕňa všetky druhy striel vyslané na súperovu bránku, mínus všetky pokusy

vyslané na vlastnú bránku. Popisuje aktivitu tímu v zápase. Záporné $\boldsymbol{C}$ indikuje, že daný tím sa v zápase viac bránil, a teda strávil viacej času v obranom pásme. Kladné $\boldsymbol{C}$ hovorí, že tím viac útočil, a tak strávil viacej času v útočnom pásame. V reči matematiky to vyzerá nasledovne: pozri tabuľku 1 nižšie. [7]

Názov	**Značka**	**Vzorec**
Corsi pre	$\boldsymbol{CF}$	$\boldsymbol{S + BS + MS}$
Corsi proti	$\boldsymbol{CA}$	$\boldsymbol{S + BS + MS}$
Corsi	$\boldsymbol{C}$	$\boldsymbol{CF - CA}$
Corsi pre v %	$\boldsymbol{CF\%}$	$\frac{\boldsymbol{CF}}{\boldsymbol{CF+CA}}.100\%$

Tabuľka 1: Corsiho štatistika

1.2.2 Fenwickovo číslo

Obdobou *Corsiho čísla* je *Fenwickovo číslo* ($\boldsymbol{F}$), ktoré je popísané blogerom Mattom Fenwickom z Alberty. Vo svojom blogu popísal, že zblokované strely nie sú reálne ohrozenie brány a preto ich z *Corsiho čísla* vynechal. Táto štatistika tiež popisuje aktivity tímov. Podobne ako u *Corsiho čísla* ju zapíšeme do tabuľky 2. [8]

Názov	**Značka**	**Vzorec**
Fenwick pre	$\boldsymbol{FF}$	$\boldsymbol{S + MS}$
Fenwick proti	$\boldsymbol{FA}$	$\boldsymbol{S + MS}$
Fenwick	$\boldsymbol{F}$	$\boldsymbol{FF - FA}$
Fenwick pre v %	$\boldsymbol{FF\%}$	$\frac{\boldsymbol{FF}}{\boldsymbol{FF+FA}}.100\%$

Tabuľka 2: Fenwickova štatistika

2 Hokejové štúdie

V hokeji sa robí veľa zaujímavých štúdií, ktoré sledujú rôzne aspekty hry. V tejto kapitole budú predstavené vybrané hokejové štúdie, a následne budú rozdelené do dvoch častí. Prvá časť bude obsahovať rešerše s pokročilými hokejovými štúdiami, ktoré sa zaoberajú podrobnými dátami a pokročilými modelmi. Avšak takto podrobné dáta niesú zvyčajne verejnosti dostupné, a preto v druhej časti budú popísané štúdie, ktoré používajú základné štatistiky a dáta. Údaje sú verejnosti dostupné a vyskytujú sa aj v tejto práci.

2.1 Pokročilé štúdie

2.1.1 Analýza hokejových vhadzovaní

Neodlúčiteľnou súčasťou hokeja sú vhadzovania. Vyhraté vhadzovanie je pre tím veľkou výhodou, avšak nie každé vyhraté vhadzovanie má pre tím takú hodnotu, ako by si mnohí mysleli. Touto témou sa autori zaoberali v štúdii [9] na vzorke vhadzovaní zo sezóny NHL 2017/2018.

Autori sa v nej nezamerali len na percentuálnu výhernosť vhadzovaní, ale skôr na pozičnú výhodu po vyhratí vhadzovania do vysoko hodnotných oblastí na ľade. Na základe toho rozdelili vyhraté vhadzovania na 2 časti. Prvé nazvali *čisté* vhadzovania, to boli také, ktoré hráč vyhral priamo na spoluhráča. Druhé boli *nečisté*, kde puk nešiel priamo na spoluhráča, ktorý stál v stacionárnej pozícii, ale musel si pre puk prikorčulovať alebo oňho zvádzať boj. Sledovali tiež silné a slabé strany hráčoc a aj oblasti kde hráči vyhrávali svoje vhadzovania. Prišli k zaujímavým zisteniam. Aj keď sa hodnota jednotlivých vhadzovaní pre tím zdá malá, objavili značnú variabilitu v hodnote založenej na kontexte zápasu, tj. kde sa buly odohráva, akú techniku na neho hráč použije atď. Preukázali tiež, že pri tých istých kontextoch zápasu nie všetky vyhrané vhadzovania hráči prevedú rovnako. Niektorí hráči robia pridanú hodnotu tým, ako *čisté* vhadzovanie vyhrajú a hlavne do akej oblasti na ľade. Preto sa *čisté* vhadzovanie preukázalo ako značne hodnotnejšie hlavne v útočnom pásme, pričom silné a slabé strany centrov ukázali lepší odhad na miesto, kam center puk vyhrá. Z toho im vyplynulo, že hráči, ktorí dokážu *čisto* vhadzovanie vyhrať a zároveň nasmerovať puk do hodnotnej oblasti, sú najhodnotnejší pre tím.

Na základe toho spravili model, ktorým počítajú hodnotu vhadzovania pre tím. Model spojil percentuálnu uspešnosť vhadzovania a hodnotu vyhraného vhadzovania pre výpočet celkového vplyvu generovaného počas vhadzovania. Tento model zmenil pohľad na vhadzovania, a miesto jednoduchej percentuálnej hodnoty výhernosti, ktorá sa počítala doposial, pomáha merať a hlavne chápať plnú hodnotu, ktorú hráč pre tím získaním puku vytvoril.

2.1.2 Hodnotenie tempa hry

Hokej patrí medzi najrýchlejšie športy na našej planéte. Taktiež zmeny v taktikách tímov posunuli tempo hry za posledných 15 rokov na ešte vyšší level. Touto témou sa zaoberá štúdia [10].

Autori sa v nej zaoberali hlavne ako tými a hráči ovplyvňujú tempo hry. Porovnávali tempo hry v rôznych oblastiach klziska, medzi tretinami, v početných výhodách a nevýhodách, taktiež naprieč rôznymi profesionálnymi ligami a aj medzi sezónami.

Takisto vyhodnocovali ako kľúčové udalosti (strely, vstup do pásma, nahrávky) sú ovplyvnené tempom hry. Prišli k nasledujúcim výsledkom. Zvýšené tímové tempo hry je prínosom, ale len v určitých častiach hry. Hlavne pri útočení zvýšené tempo dokáže narušiť obrannú štruktúru súpera, a môže prísť k nebezpečnejším prienikom do tretiny a následným zlepšním streleckých príležitostí. Na druhej strane prihrávky s vyššou rýchlosťou viedli k vyšším stratám pukov. Zlepšili aj +/- model, ktorý obohatili o účinky spoluhráčov a výkony hráča. Metriky, ktorými merali tempo hry v rôznych oblastiach klziska sa preukázali, ako veľmi užitočný nástroj, ktorý možno implemenovať aj do iných športov. Definovaním tempa hry ako rýchlosti pohybov puku namiesto trajektórií pohybov hráčov sa preukázalo ako ešte lepšie zachytenie kontextu hry. Použitím tejto definície sa môže skúmať tempo hry aj u iných športov, u ktorých niesú dostupné dáta o pohybe hráčov.

2.2 Základné štúdie

2.2.1 Výhoda domáceho prostredia

Výhoda domáceho prostredia v kolektívnom športe je často diskutované téma športových odborníkov ale aj fanúškov či stávkarov. Niektorí sú toho názoru, že domáce prostredie nemá na výsledok zápasu žiadny vplyv, nakoľko hráči sú profesionály a je im jedno kde hrajú. Iní zastávajú názor, že doma sa hrá podstatne lepšie kvôli známemu prostrediu, domácemu publiku atď. Tejto problematike sa venuje aj štúdia [11], ktorá porovnáva domácu výhodu pri rôznych ukončiniach zápasov.

Autori v nej skúmali výhody domáceho prostredia na vzorke 9 sezón NHL spolu 10 534 zápasov základných častí NHL. Do výpočtov zahrnuli aj kvalitu domáceho a hosťujúceho tímu, ktorú počítali cez *Pythagorejskú metódu* [12] pre každý zápas v danej sezóne zvlášť. Správnosť tejto metódy si overili pomocou Pearsnovej korelácie, medzi predpovedanými hodnotami a reálnymi vysledkami zápasov, ktorá preukázala vysokú prestnosť. Zápasy rozdelili do 3 kategórii podľa toho, či zápas skončil v riadnom hracom čase (REG), predĺžení (OT) alebo až na samostatné nájazdy (SO). Potom medzi týmito ukončeniami zápasov sledovali, či domáce tímy viac vyhrávajú alebo preháravjú. Sledovali to s prihliadnutím aj na kvalitu tímu aj bez nej a taktiež s interakciou kvality a ukončenia. Výsledky rozdelili do 3 blokov.

Blok 1 neprihliadal na žiadnu tímovú kvalitu a výsledky v rôznych ukončeniach boli nasledované: v prípade dvojice REG a OT sa vyššia alebo nižšia pravdepodobnosť na výhru domáceho tímu zvoleným testom nepreukázala, a teda zásadne sa nelíši. U dvojice SO a OT pravdepodobnosť na výhru domáceho tímu klesla pri SO o 30 % a u REG a SO klesla pre SO až o 44 %.

Blok 2 obsahuje prihliadnutie na kvalitu tímu. Pri pozovnaní REG a OT sa podobne ako pri bloku 1 nepreukázala rozdielna pravdepodobnosť. Výsledky ukázali, že kvalita tímu je dôležitý faktor. V prípade, ak nastúpi proti sebe lepší domáci tím proti horšiemu hosťujúcemu tímu, pravdepodobnosť výhry domáceho tímu sa zvýši o 5 %. Pri dvojici SO a OT sa pravdepodobnosť výhry pri SO opäť znížila a to o 33 %. Na druhú stranu, pri dvojici silný domáci tím a slabý hosťujúci tím, faktor kvality je opäť dôležitý a tým pravdepodobnosť výhry narastie o 5 %. U REG a SO sa preukázla rozdielna pravdepodobnosť, ktorá sa znížila o 41 % u SO. Aj u poslednej dvojice v tomto bloku je preukázaný faktor kvality a podobne ako pri dvojiciach vyšla, pravdepodobnosť výhry narastie o 5 %.

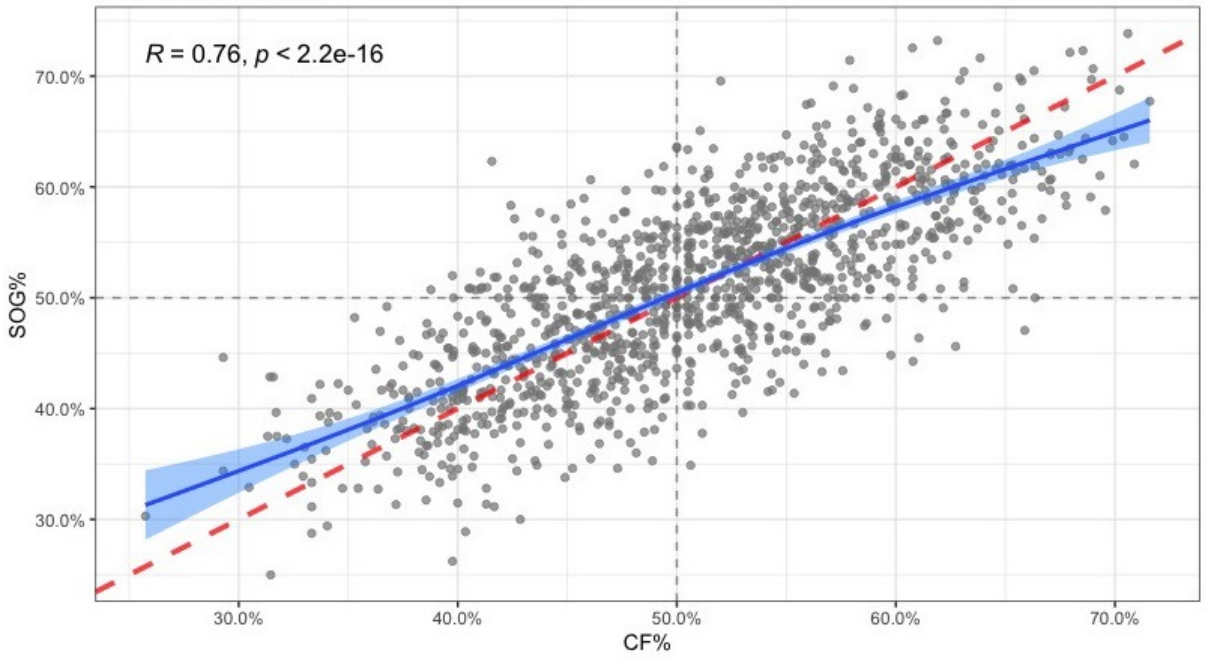
Blok 3 zahŕňa interakciu medzi kvalitou tímov a ukončením zápasu. Teraz sa oproti Bloku 1 a Bloku 2 pri dvojici REG a OT preukázala zvoleným testom pravdepodobnosť

výhry domáceho tímu o 19 % nižšia pri OT. Taktiež sa ukázalo, že v prípade kvalitnejšieho domáceho tímu závislosť na pravdepodobnosti výhry v REG je v porovnaní s OT o 3 % vyššia. Medzi SO a OT je pravdepodobnosť výhry domáceho tímu pri SO nižšia o 23 %, na druhú stranu kvalita tímu výrazne neovplyvnila výsledok domáceho tímu a ani trvanie zápasu. U dvojice REG a SO sa testom nepreukázal vplyv prostredia domáceho tímu. Avšak výsledky prukázali značnú závislosť, ktorá silnejším domácim tímom znižuje pravdepodobnosť výhry o 4 % keď sa zápas dostane do SO.

2.2.2 Vplyv striel na výsledok

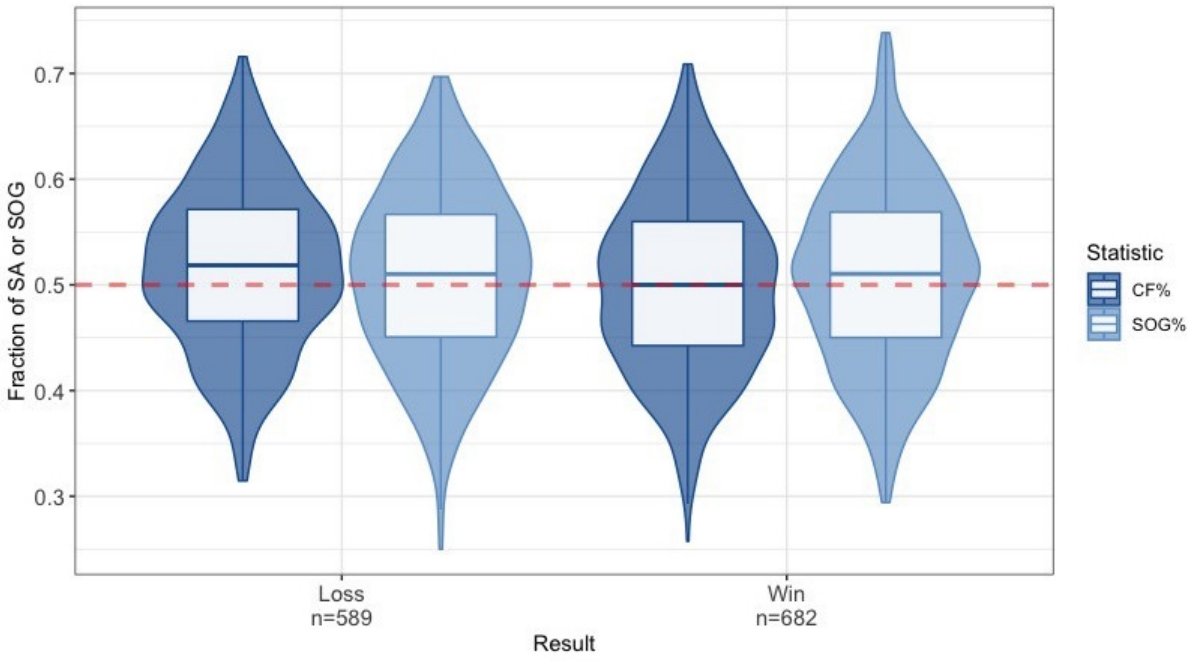
Každý hokejový hráč si aspoň raz za svoju kariéru vypočul od svojho trénera: „Ak nevystrelíš, nemôžeš dať gól!“ V štúdii od autora Roba Founda [13] sa uvádza, že štatisticky v NHL je potrebných na dosiahnutie jedného góla až 10 striel. Súvisia spolu strelecké pokusy (***C***) a strely na bránu (***SOG***)? Dá sa nimi aj predpovedať víťaz zápasu? Týmto otázkam sa venuje štúdia [14].

Otázky autor testoval na základnej hracej časti NHL sezóny 2018/2019, čo bolo dokopy 1271 zápasov. V prvej časti skúmal závislosť ***C*** a ***SOG***. Podotkol, že ***C*** sa počíta pri rovnakom počte hráčov na ľadovej ploche, a ***SOG*** zahrňajú aj situácie ako presilovky a oslabenia, kde môže byť početnosť striel vyššia. Nekoreloval však medzi sebou samotné ***C*** a ***SOG***, ale ich percentuálne ekvivalenty. Percentuálny počet ***SOG%*** sa podobne ako u ***CF%*** počíta tak, že počet striel tímu je podelený celkovým počtom striel v zápase vynásobených 100 %. Autor predpovedal vysokú mieru korelácie medzi týmito štatistikami, ale prišiel aj k zaujímavým zisteniem. Zistená pozitívna korelácia je na krajoch slabšia tj. pri zvýšení ***CF%*** z 60 % na 70 % obecne bude znamenať vyššie ***SOG%***, ale nárast bude menší, ako v prípade keď to bude z 50 % na 60 % pri ***CF%***. U znižovnia naopak ***SOG%*** na druhej strane stúpa. Tento opačný trend zobrazený modrou krivkou v obr. 1. autor vysvetluje tým, že ***CF%*** môže byť viacej extrémnejšie ako ***SOG%***.



Obr. 1: Korelácia medzi SOG% a CF% [14]

V druhej časti autor venuje pozornosť otázke vzťahu ***SOG*** a ***C*** s výhrou či prehrou. Pomocou husľového grafu (violin plot) prišiel k zaujímavým výsledkom. V prípade, že domáci tím prehral, mal tendenciu mať o niečo vyššie ***C*** a ***SOG*** v zápase, ako jeho protivník. U hosťujících tímov tomu bolo naopak, keď vyhrali mali o trocha nižšie ***C*** a ***SOG***. Medián ***SOG%*** medzi prehrami a výhrami sa pohybuje jemne nad 50 %. Priemer ***CF%*** je podstatne vyšší pri prehrách ako pri výhrach. Obe rozdelenia sa približne pohybujú okolo stredu viď obr. 2, preto pre autora model nemá až takú výpovednú hodnotu.



Obr. 2: C a SOG verzus výsledky [14]

V poslednej časti sa zo získaných skutočností autor pokúsil o zostavenie modelu na predikcie výhier, ktorý mal len 56,6% úspešnosť.

3 Matematická teória

V tejto kapitole budú predstavené štatistické nástroje použité pri testovaní. Budú zadefinované pojmy ako parametrické a neparametrické t-testy, korelácia a ANOVA. Pre tieto testy je hlavný predpoklad normálneho rozdelenia, ktoré sa zadefinuje ako prvé. Všetky informácie napísané v tejto kapitole budú čerpané zo zdrojov [1] [2] [3] [4] [5].

3.1 Normálne rozdelenie

Náhodná veličina $\boldsymbol{X}$ má *normálne rozdelenie* s parametrami $\mu \in \mathbb{R}$ a $\sigma^2 > 0$, ak je jej hustota rovná

$$\boldsymbol{f}(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}.$$

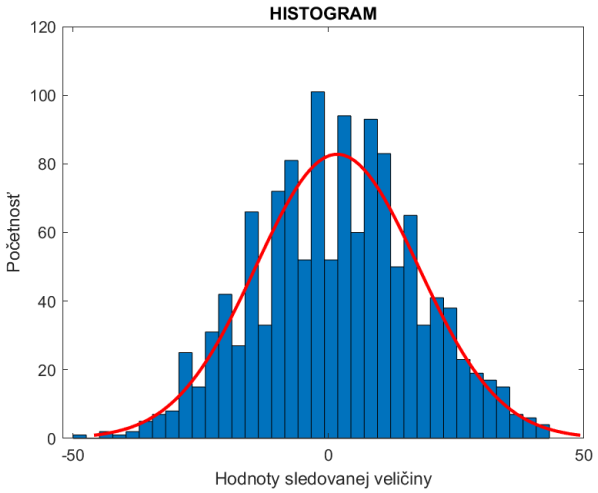
Zapisovať budeme $\boldsymbol{X} \sim N(\mu, \sigma^2)$ a bude platiť $EX = \mu$ a $\texttt{var}\ X = \sigma^2$, kde EX bude označenie pre strednú hodnotu a $\texttt{var}\ X$ bude označenie pre rozptyl.

3.2 Testy normality

Pri testovaní vybraných hypotéz budú použité také nástroje, ktoré majú ako hlavný predpoklad *normálne rozdelenie*. Preto je nutné tento predpoklad otestovať. Existuje veľa nástrojov na otestovanie normality. Pre použité dáta bol vybraný jeden grafický *Histogram* a jeden výpočetný *Andersonov-Darlingov test*, ktoré budú popísané nižšie.

3.2.1 Histogram

Graf, ktorý sa vytvorí pomocou hodnôt sledovanej veličiny a jej početnosti, bude nazývaný *Histogram*. Pri väčšom počte testovaných dát alebo pri spojitej veličine rozdelíme osu na intervaly. Četnosti budú určené ako počty pozorovaných hodnôt v intervaloch, ktoré budú stanovené napríklad pomocou tzv. Sturgesovho pravidla. Veľkými výhodami histogramu sú, že je pekne vidieť ako sú dáta rozložené, aké majú odchýlky a či sú vôbec symetrické. Ak je dané rozdelenie normálne, mal by *Histogram* pripominať Gaussovu krivku, ktorú v obr. 3 reprezentuje červená krivka.



Obr. 3: Histogram s normálnym rozdelením

3.2.2 Andersonov-Darlingov test

Andersonov-Darlingov (AD) test patrí do rodiny testov pomocou kvadratických štatistík empirickej distribučnej funkcie. Meria vzdialenosť medzi predpokladaným normálnym rozdelením $\boldsymbol{F}(x)$, a empirickou distribučnou funkciou $\boldsymbol{F}_n(x)$ ako

$$n \int_{-\infty}^{\infty} (\boldsymbol{F}_n(x) - \boldsymbol{F}(x))^2 \boldsymbol{w}(x) \boldsymbol{d}\boldsymbol{F}(x),$$

nad zoradenými hodnotami výberu $x_1 < x_2 < ... < x_n$ kde $\boldsymbol{w}(x)$ je váhová funkcia a n je počet dát vo výbere. Kde predpokladané $\boldsymbol{F}(x)$ má z dát odhadnutú strednú hodnotu a smerodatnú odchýľku.

Váhová funkcia pre *Andersonov-Darlingov* test je

$$\boldsymbol{w}(x) = [\boldsymbol{F}(x)(1 - \boldsymbol{F}(x))]^{-1},$$

ktorá kladie väčšiu váhu na pozorovanie v koncových častiach rozdelenia. Test sa stáva citlivejším na odľahlé hodnoty, a lepšie detekuje odchýľku od normality v koncových častiach rozdelenia.

Testovacia štatistika pre *Andersonov-Darlingov* test je

$$\boldsymbol{A}_n^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} [\ln(\boldsymbol{F}(\boldsymbol{X}_i)) + 1 - \ln(\boldsymbol{F}(\boldsymbol{X}_{n+1-i}))],$$

kde $X_1 < X_2 < ... < X_n$ sú zoradené body dát a n je ich počet.

Pri tomto teste bude zamietnutá alebo nezamietnutá nulová hypotéza, ktorou je testovaná normalita dát tak, že sa porovnáva *p-hodnotu* pre test hypotézy so stanovenou hladinou významnosti, a neporovnáva štatistiku testu s kritickou hodnotou.

3.3 Parametrické testy

3.3.1 Dvojvýberový t-test

Prvý z testov, ktorý bude v tejto práci pužítý, je *Dvojvýberový t-test*. V ňom bude predpokladaná zhoda rozptylov oboch výberov, ktorá bude overená pomocou *F-testu*.

Pri *F-teste* bude predpokladané, že $X_1, X_2, ..., X_n$ je výber z $N(\mu_1, \sigma_1^2)$ a že $Y_1, Y_2, ..., Y_m$ je výber z $N(\mu_2, \sigma_2^2)$, pričom $n \geq 2$, $m \geq 2$, $\sigma_1^2 > 0$ a $\sigma_2^2 > 0$. Charakteristiky týchto výberov budú postupne označené ako

$$\overline{X_n} = \frac{1}{n} \sum_{k=1}^{n} X_k,$$

$$\overline{Y_m} = \frac{1}{m} \sum_{k=1}^{m} Y_k,$$

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2,$$

$$S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \overline{Y_m})^2,$$

a nech tieto dva výbery sú na sebe nezávislé a označenie dvoch výberov bude také, aby platilo $S_X^2 \geq S_Y^2$. Bude testovaná hypotéza $H_0 : \sigma_1^2 = \sigma_2^2$ oproti altrnatívnej hypotéze $H_1 : \sigma_1^2 \neq \sigma_2^2$. Test bude prevedený tak, že H_0 sa zamietne, keď bude platiť $S_X^2/S_Y^2 \geq F_{n-1,m-1}(\frac{\alpha}{2})$, kde $F_{n-1,m-1}$ sú kritické hodnoty F-rozdelenia.

V prípade, že rozptyly sú rovnaké, bude postup nasledovný. Nech je $X_1, X_2, ..., X_n$ výber z $N(\mu_1, \sigma^2)$ a nech je $Y_1, Y_2, ..., Y_m$ výber z $N(\mu_2, \sigma^2)$. Nech tieto dva výbery sú na sebe nezávislé. Pri predpoklade, že $n \geq 2$, $m \geq 2$, $\sigma^2 > 0$. Označia sa $\overline{X}_n, \overline{Y}_m, S_X^2, S_Y^2$ ako charakteristiky týchto výberov. Potom

$$T = \frac{\overline{X}_n - \overline{Y}_m - (\mu_1 - \mu_2)}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t_{n+m-2}.$$

Dôkaz viď str. 75 [1]

Za predpokladov uvedených vyššie bude testovaná hypotéza $H_0 : \mu_1 - \mu_2 = \Delta$ proti $H_1 : \mu_1 - \mu_2 \neq \Delta$ vypočíta sa T, ktoré je kvantilom Studentova rozdelenia, a keď $|T| \geq t_{n+m-2}(\alpha)$, hypotéza bude zamietnutá H_0 na hladine α.

Ak rozptyly sú rôzne, bude použitý *Welchov test*. Nech je $X_1, X_2, ..., X_n$ výber z $N(\mu_1, \sigma_1^2)$ a nech je $Y_1, Y_2, ..., Y_n$ výber z $N(\mu_2, \sigma_2^2)$. Nech tieto dva výbery sú na sebe nezávislé. Bude predpokladané, že $n \geq 2$, $m \geq 2$, $\sigma_1^2 > 0$ a $\sigma_2^2 > 0$. Označia sa $\overline{X}_n, \overline{Y}_n, S_X^2, S_Y^2$ ako charakteristiky týchto výberov. Potom testovacia charakteristika je

$$T = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t_\gamma,$$

kde γ je počet stupňov voľnosti, pre ktorú platí vzťah

$$\gamma = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)}{\frac{\left(\frac{S_X^2}{n}\right)}{n-1} + \frac{\left(\frac{S_Y^2}{m}\right)}{m-1}}.$$

Ak bude za predpokladov *Welchovho testu* testovaná hypotéza $H_0 : \mu_1 - \mu_2 = \Delta$ proti $H_1 : \mu_1 - \mu_2 \neq \Delta$ vypočita sa T a keď $|T| \geq t_\gamma(\alpha)$, hypotéza H_0 bude na hladine α zamietnutá.

3.3.2 Pearsonov korelačný koeficient

Ďalej v tejto práci budú pozorované rôzne páry štatistík, a či majú medzi sebou nejakú lineárnu závislosť. K tomu bude použitý *Pearsonov korelačný koeficient*, ktorý je pre N párových premenných definovaný ako

$$\rho(X,\ Y) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right),$$

kde μ_X je stredná hodnota a σ_X je smerodatná odchýľka pre náhodný výber X, a podobne μ_Y je stredná hodnota a σ_Y je smerodatná odchýľka pre náhodný výber Y.

Taktiež sa môže zadefinovať korelačný koeficient z hľadiska kovariácie X a Y nasledovne

$$\rho(X,Y) = \frac{\mathtt{cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Maticu korelačných koeficientov pre všetky kombinácie párových premenných sa zapíše ako

$$R = \begin{pmatrix} \rho(X,X) & \rho(X,Y) \\ \rho(Y,X) & \rho(Y,Y) \end{pmatrix}.$$

Pretože X a Y korelujú vždy priamo, na diagonálu sa dá 1 a výsledná matica vyzerá takto

$$R = \begin{pmatrix} 1 & \rho(X,Y) \\ \rho(Y,X) & 1 \end{pmatrix}.$$

Ďalej sa zavedie test *Pearsnovej korelácie*. Nech $(X_1, Y_1)', (X_2, Y_2)', ..., (X_n, Y_n)'$ je výber z dvojrozmerného normálneho rozdelenia, ktoré má kladné rozptyli a korelčný koeficient $\rho = 0$. Potom náhodná veličina

$$T = \frac{R}{\sqrt{1 - R^2}} \sqrt{n-2}, \tag{1.1}$$

má rozdelenie t_{n-2}. Dôkaz viď str. 196 [2].

Potom sa pomocou 1.1 testuje hypotéza $H_0 : \rho = 0$ proti alternatívnej $H_1 : \rho \neq 0$. Vypočíta sa najprv korelačný koeficient R, pre ktorý musí platiť $R \neq \pm 1$, potom veličinu T podľa 1.1, a ak $|T| \geq t_{n-2}(\alpha)$, sa zamietne hypotéza H_0 na hladine významnosti α. Pritom je veľmi podstatný predpoklad, že výber $(X_1, Y_1)', (X_2, Y_2)', ..., (X_n, Y_n)'$ pochádza z normálneho rozdelenia.

3.3.3 Dvojfaktorová analýza rozptylu bez interakcií

Posledným parametrickým nástrojom, ktorý bude použitý, je *Dvojfaktorová analýza rozptylu bez interakcií*. Budú pozorované dva triediace znaky a ich súčtom sa môže prejaviť vplyv vektorov na strednú hodnotu. Model sa zapíše ako

$$Y_{ijp} = \mu + \alpha_i - \beta_j + e_{ipj}, \tag{1.2}$$

pre $i = 1, 2, ..., I; j = 1, 2, ..., J; p = 1, 2, ..., P$, kde μ, α_i, β_j sú neznáme parametre a e_{ijp} sú nezávislé veličiny s rozdelením $N(0, \sigma^2)$. Parameter $\sigma^2 > 0$ je tiež neznámy. Poznamená sa, že α_i sú tzv. *riadkové efekty* a β_j sú tzv. *stĺpcové efekty*. Označí sa $n = IJP$ ako celkový počet veličín Y_{ijp} a ďalej sa položí

$$Y_{ij\bullet} = \sum_{p=1}^{P} Y_{ijp}, \qquad Y_{i\bullet\bullet} = \sum_{j=1}^{J}\sum_{p=1}^{P} Y_{ijp}, \qquad Y_{\bullet\bullet\bullet} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{p=1}^{P} Y_{ijp},$$

$$y_{ij\bullet} = \frac{Y_{ij\bullet}}{P}, \qquad y_{i\bullet\bullet} = \frac{Y_{i\bullet\bullet}}{JP}, \qquad y_{\bullet\bullet\bullet} = \frac{Y_{\bullet\bullet\bullet}}{n}.$$

Normálne rovnice sa vypočítajú tak že výraz

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{p=1}^{P}(Y_{ijp} - \mu - \alpha_i - \beta_j)^2,$$

postupne parciálne sa derivuje podľa μ, α_i, β_j a tieto derivácie budú položené rovno nule. Získa sa tým sústava rovníc viď str. 217 [1].

Matica sústavy má hodnosť $r = I+J-1$. Je $I+J+1$ rovníc teda je zrejmá prítomnosť nadbytočných parametrov, preto treba zaviesť reparametrizačné rovnice

$$\sum \alpha_i = 0, \qquad \sum \beta_j = 0.$$

Ako ďalšie sa zavedú sučty štvorcov

$$S_{X_1} = JP \sum_{i=1}^{I} y_{i\bullet\bullet}^2 - n y_{\bullet\bullet\bullet}^2,$$

$$S_{X_2} = IP \sum_{j=1}^{J} y_{\bullet j\bullet}^2 - n y_{\bullet\bullet\bullet}^2,$$

$$S_T = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{p=1}^{P} Y_{ijp}^2 - n y_{\bullet\bullet\bullet}^2,$$

$$S_e = S_T - S_{X_1} - S_{X_2},$$

$$f_{X_1} = I-1, \qquad f_{X_2} = J-1, \qquad f_T = n-1, \qquad f_e = n-I-J+1$$

Ak by na riadkovom triediacom znaku nezáležalo, bude to $\alpha_1 = \alpha_2 = ... = \alpha_I = 0$. Vznikne tým podmodel

$$Y_{ijp} = \mu + \beta_j + e_{ipj}, \tag{1.3}$$

ktorý zodpovedá jednoduchému triedeniu.

Položí sa $s^2 = S_e/f_e$. Hypotézu o možnosti redukcie modelu z (1.2) na podmodel (1.3) sa pripúšťa, ak

$$F_{X_1} = \frac{S_{X_1}}{f_{X_1}} \frac{1}{s^2} < F_{f_{X_1}, f_e}(1-\alpha).$$

Ak by sa chcelo overiť, či záleží na stĺpcovom triediacom znaku, model (1.2) bude redukovaný na podmodel

$$Y_{ijp} = \mu + \alpha_i + e_{ipj},$$

pri podmienke

$$F_{X_2} = \frac{S_{X_2}}{f_{X_2}} \frac{1}{s^2} < F_{f_{X_2}, f_e}(1-\alpha).$$

Výsledky sa zapíšu do tabuľky 3 nižšie.

Variabilita	**Súčet štvorcov SS**	**Počet stupňov volnosti** *df*	**Podiel SS/*df***	**Testová štatistika**
Riadková	S_{X_1}	f_{X_1}	$\frac{S_{X_1}}{f_{X_1}}$	F_{X_1}
Stĺpcová	S_{X_2}	f_{X_2}	$\frac{S_{X_2}}{f_{X_2}}$	F_{X_2}
Reziduálna	S_e	f_e	$s^2 = \frac{S_e}{f_e}$	-
Celkový	S_T	f_T	-	-

Tabuľka 3: Analýza rozptylu dvojného triedenia bez interakcií

Keď je F_{X_1} štatisticky významná, z dvojice riadkov bude nutné zistiť či sa od seba významne odlišujú. Podľa Tukeyovy metódy sa zamieta rovnosť $\alpha_i = \alpha_t$, ak

$$|y_{i\bullet\bullet} - y_{\bullet t\bullet}| > \frac{s}{\sqrt{JP}} q_{I,n-I-J+1}(\alpha),$$

kde $q_{I,n-I-J+1}(\alpha)$ sú tabelované kritické hodnoty studentizovaného rozpätia.

Obdobne pri významnej hodnote F_{X_2} sa zamieta rovnosť $\beta_j = \beta_t$ Tukeyovou metódou, ak

$$|y_{\bullet j\bullet} - y_{\bullet t\bullet}| > \frac{s}{\sqrt{IP}} q_{I,n-I-J+1}(\alpha).$$

3.4 Neparametrické testy

3.4.1 Dvojvýberový Wilcoxonov test (modifikovaný na Mannov-Whitneyov test)

V prípade, že použité dáta nebudú spĺňať predpoklad normality, bude použitý *Dvojvýberový Wilcoxonov test (modifikovaný na Mannov-Whitneyov test)* namiesto *Dvojvýberového t-testu.*

Nech $X_1, X_2, ..., X_n$ je náhodný výber zo spojitého rozdelenia s distribučnou funkciou $\boldsymbol{F}$ a nech $Y_1, Y_2, ..., Y_m$ je ňom nezávislý náhodný výber zo spojitého rozdelenia s distribučnou funkciou $\boldsymbol{G}$. Potom bude testovaná hypotéza $H_0 : \boldsymbol{F} = \boldsymbol{G}$ proti alternatívnej hypotéze $H_1 : \boldsymbol{F} \neq \boldsymbol{G}$

Všetky $m + n$ veličiny $X_1, X_2, ..., X_n,\ Y_1, Y_2, ..., Y_m$(označované aj *združený výber*) budú usporiadané vzostupne podľa veľkosti. T_1 označia sa ako súčet poradí hodnôt $X_1, X_2, ..., X_m$ a T_2 ako súčet poradí hodnôt $Y_1, Y_2, ..., Y_n$, platí

$$T_1 + T_2 = \frac{(m+n)(m+n+1)}{2}.$$

Častejšie sa však používa Mannov-Whitneyova modifikácia, ktorá je modifikáciou Wilcoxonovho dvojvýberového testu. V nej sa spočíta hodnota U_1 pomocou hodnoty T_1 nasledovne

$$U_1 = mn + \frac{m(m+1)}{2} - T_1,$$

obdobne bude spočítaná hodnota U_2

$$U_2 = mn + \frac{n(n+1)}{2} - T_2,$$

pritom platí $U_1 + U_2 = mn$. Ak $\min(U_1, U_2)$ je menšie alebo rovné ako tabelovaná kritická hodnota, sa zamietne nulová hypotéza. Označenie výberu sa volí tak, aby platilo $m \geq n$.

Pre veľké m a n bude použitý nasledujúci postup. Z vety o empirickej distribučnej funkcii viď str. 243 [1] vyplýva že

$$\mathtt{E}U_1 = \frac{mn}{2},$$

$$\mathtt{var}\, U_1 = \frac{mn(m+n+1)}{12}.$$

Keďže $U_1 = mn - U_2$, z toho vyplýva $\mathrm{E}U_1 = \mathrm{E}U_2$ a $\mathtt{var}\, U_1 = \mathtt{var}\, U_2$. Je tiež dokázané, že pri $m \to \infty$ a $n \to \infty$ má veličina U_1 (tiež T_1) asymptoticky normálne rozdelenie. Vypočíta sa teda ako

$$U_{MW} = \frac{U_1 - \mathrm{E}U_1}{\sqrt{\mathtt{var}\, U_1}}. \tag{1.3}$$

Keď $|U_{MW}| \geq u(\frac{\alpha}{2})$, bude zamietnutá hypotéza H_0 na hladine významnosti blížiacej sa α. Test založený na (1.3) sa dá použiť pri $n > 10$ a $m > 10$.

3.4.2 Spearmanov korelačný koeficient

Podobne ako pri *Dvojvýberovom t-teste*, akonáhle bude porušená normalita nebude použitý *Pearsonov korelačný koeficient*, ale namiesto neho *Spearmanov korelačný koeficient*.

Predpokladá sa, že $(X_1, Y_1)', (X_2, Y_2)', ..., (X_n, Y_n)'$ je výber zo spojitého dvojrozmerného rozdelenia. Nech $R_1, R_2, ..., R_n$ sú poradia veličín $X_1, X_2, ..., X_n$ a nech $Q_1, Q_2, ..., Q_n$ sú poradia veličín $Y_1, Y_2, ..., Y_n$. Dvojice $(X_1, Y_1)', (X_2, Y_2)', ..., (X_n, Y_n)'$ sa často už vopred usporiadavajú podľa rastúcich hodnôt $X_1, X_2, ..., X_n$. V takom prípade je $R_i = i, i = 1, 2, ..., n$.

Spearmanov korelačný koeficient r_S sa definuje ako výberový korelačný koeficient počítaný z dvojíc $(R_1, Q_1)', (R_2, Q_2)', ..., (R_n, Q_n)'$ a je rovný

$$r_S = \frac{\sum R_i Q_i - n\overline{RQ}}{\sqrt{(\sum R_i^2 - n\overline{R}^2)(\sum Q_i^2 - n\overline{Q}^2)}}.$$

Kritické hodnoty $r_S(\alpha)$ budú hľadané v tabulkách a pre $n > 30$ sa využije asymptotická normalita koeficientu r_S, ktorá sa vypočíta ako

$$r_S^*(\alpha) = \frac{u(\frac{\alpha}{2})}{\sqrt{n-1}}.$$

Hypotéza neprítomnosti monotónej záavislosti bude zamietnutá ak $|r_S| \geq r_S^*(\alpha)$. Jednostranné varianty tohoto testu sa odvodia obdobne.

4 Testované hypotézy

Posledná kapitola sa bude venovať testovaniu hypotéz na dostupných hokejových štatistikách. Preto bude zvolená ako testovacia vzorka najznámejšia hokejová liga NHL. Bude na nej skúmaná výhoda domáceho prostredia, vzťahy niektorých štatistík a v neposlednom rade bude preskúmaná aktivita tímu v závislosti na dvoch faktoroch. Ak nebude uvedené inak, sú všetky výpočty, testy a vizualizácie vytvorené prostredníctvom softvéru Matlab, a budú počítané na hladine významnosti $\alpha = 0,05$.

4.1 Dáta

Ako bolo spomínané v úvode tejto kapitole, sú vybrané dáta NHL na základe toho, že sú verejne dostupné pre verejnosť. V neposlednom rade sú vybrané štatistiky NHL aj kvôli tomu, že liga NHL je najkvalitnejšia a najvyššia hokejová liga vo svete. Konkrétne je zvolená základná časť sezóny 2018/2019, pretože to bola posledná sezóna odohraná klasickým štýlom bez reštrikcií spojených s Covid-19. Počet tímov, ktoré hrali základnú časť, je 31 a spolu odohrali 1271 zápasov. Z týchto zápasov boli zozbierané dáta pomocou webovej stránky [15]. Ku každému zápasu boli pridelené tieto štatistiky a pre každý tím: $\boldsymbol{G}$, $\boldsymbol{S}$, $\boldsymbol{FO}$, $\boldsymbol{P}$, $\boldsymbol{C}$, $\boldsymbol{F}$ a návštevnosť divákov ($\boldsymbol{Att}$).

4.2 Vplyv domáceho prostredia na góly

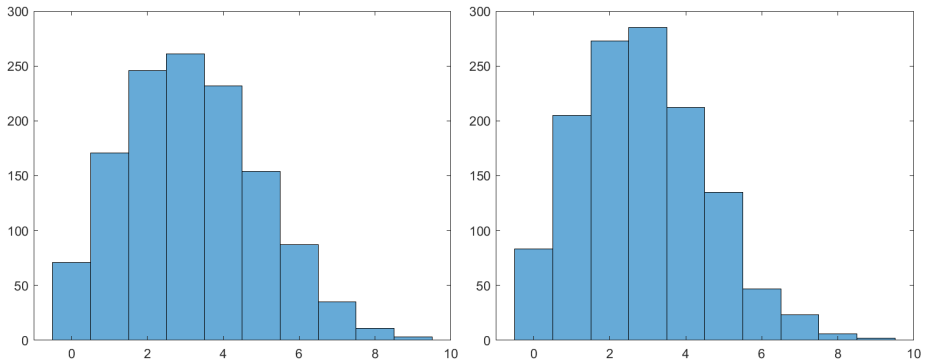
Výhoda domáceho prostredia v kolektívnych športoch je stále aktuálne téma plné rôznych názorov. Preto bude preverené, či naozaj domáce tímy s podporou fanúšikov, a s výhodou známeho prostredia, strielajú v zápasoch viacej gólov ako ich súperi.

Pokladá sa hypotézu $\boldsymbol{H}_0$: *hosťujúce tímy strielajú viacej gólov v zápasoch*, proti alternatívnej hypotéze $\boldsymbol{H}_1$: *domáce tímy strialajú viacej gólov v zápasoch*. Matematicky sú hypotézy zapísané nasledovne.

$$\boldsymbol{H}_0 : \mu_{GV} > \mu_{GH},$$

$$\boldsymbol{H}_1 : \mu_{GV} < \mu_{GH}.$$

Pre overenie hypotézy bude použitý *Dvojvýberový t-test* v prípade, že budú splnené predpoklady testu. Ako prvé sa pri parmatrických testoch overí *normalita* testovaných súborov. K tomu budú využité dva spôsoby prvý grafický a druhý výpočetný. Histogramy pre dáta sú nasledovné, kde góly domácich $\boldsymbol{GH}$ sú vľavo a góly hostí $\boldsymbol{GV}$ vpravo viď obrázok 4.



Obr. 4: Histogramy $\boldsymbol{GH}$ (vľavo) a $\boldsymbol{GV}$ (vpravo)

Ako je vidieť, dáta sú vychýlené do ľavej strany a veľmi sa nepodobajú Gaussovej krivke. Potvrdil to aj AD test, kde v oboch prípadoch vyšla *p-hodnota* menšia, ako najnižšia tabelovaná hodnota softvérom Matlab, ktorou je 0,0005. Preto sa normalitu dát na hladine významnosti zamieta. Pretože sa pracuje s veľkým počtom zápasov a *Dvojvýberový t-test* pre $n > 30$ je robustný aj pri porušenej normalite, môže mať pre závery výpovednú hodnotu. Ďalej bude testovaná pomocou *F-testu* zhoda rozptylov oboch súborov. Vyšla *p-hodnota* = 0,0146 a preto sa test na zhodnosť rozptylov na hladine významnosti zamieta, a tak bude použitý *Welchov test*. Po výpočte *Welchovho testu* je výsledok *p-hodnotu* $= 5,0655 \times 10^{-6}$, a teda hypotéza $\boldsymbol{H}_0$ sa na hladine významnosti zamieta.

Keďže v tomto teste neboli splnené predpoklady, urobí sa ešte jeho neparametrická alternatíva, ktorou je *Dvojvýberový Wilcoxonov test (modifikovaný na Mannov-Whitneyov test)*. Ten však nepracuje so strednými hodnotami ale s mediánmi. Preto sa matematické hypotézy prepíšu nasledovne

$$\boldsymbol{H}_0 : \widetilde{GV} > \widetilde{GH},$$
$$\boldsymbol{H}_1 : \widetilde{GV} < \widetilde{GH}.$$

U tohto testu vyšla *p-hodnota* $= 1,344 \times 10^{-5}$ a preto ako u parametrického testu sa hypotéza $\boldsymbol{H}_0$ na hladine významnosti zamieta.

Na základe výsledkov, ktoré vyšli, je možné vidieť značný vplyv domáceho prostredia na počet strelených gólov v zápase. NHL je vysoko profesionálna liga, no aj napriek tomu domáce tímy strielajú väčší počet gólov. Toto zistenie môže podporiť aj fakt, že hráči domáceho tímu viac oddychujú. V deň zápasu nemusia cestovať dlhé kilometre naprieč USA, ktoré môžu mať vplyv na ich výkon. Takisto veľkú úlohu pravdepodobne zohráva aj podpora domáceho obecenstva, ktoré býva, ako sa v hokejovej terminológii hovorí, "siedmym hráčom"na ľade.

4.3 Vplyv domáceho prostredia na strely

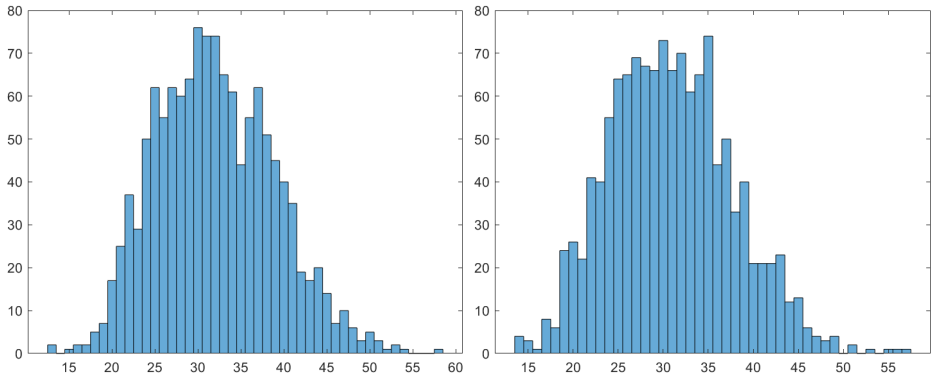
Výhoda domáceho prostredia bola pri strelených góloch potvrdená. Ďalej bude pozorované, či domáce tímy v zápasoch viacej strielajú na bránu, a tým viacej ohrozujú súperovu bránu.

Pokladá sa hypotéza $\boldsymbol{H}_0$: *hosťujúce tímy viacej strielajú na bránu v zápasoch*, proti alternatívnej hypotéze $\boldsymbol{H}_1$: *domáce tímy viacej strialajú na bránu v zápasoch*. Matematicky hypotézy budú zapísané nasledovne

$$\boldsymbol{H}_0 : \mu_{SV} > \mu_{SH},$$

$$\boldsymbol{H}_1 : \mu_{SV} < \mu_{SH}.$$

Pre overenie hypotézy bude opäť použitý *Dvojvýberový t-test* v prípade, že budú splnené predpoklady testu. Najprv sa pri parmatrických testoch overí *normalita* testovaných súborov. Bude posudzovaná podobne ako v predošlom prípade. Histogramy pre používané dáta vyzerajú nasledovne: strely na bránu domácich $\boldsymbol{SH}$ sú vľavo a strely na bránu hostí $\boldsymbol{SV}$ vpravo viď obrázok 5.



Obr. 5: Histogramy $\boldsymbol{SH}$ (vľavo) a $\boldsymbol{SV}$ (vpravo)

Dáta sú vychýlené jemne do ľavej strany, ale jemne pripomínajú Gaussovu krivku. Výsledky AD testov ale v oboch prípadoch vyšli s *p-hodnotov* menšou, ako najnižšia tabelovaná hodnota softvérom Matlab ktorou je 0,0005. Preto sa normalita dát na hladine významnosti zamieta. Keďže sa pracuje s veľkým počtom zápasov a *Dvojvýberový t-test* pre $n > 30$ je robustný aj pri porušenej normalite, môže mať pre závery výpovednú hodnotu. Ako ďalšie bude testovaná pomocou *F-testu* zhoda rozptylov oboch súborov. Vyšla *p-hodnota* = 0,7349 a preto sa test na zhodnosť rozptylov na hladine významnosti nezamieta, a tak bude použitý *Dvojvýberový t-test* pre rovnaké rozptyly. Po výpočte testu je výsledok *p-hodnotu* $= 1,0296 \times 10^{-5}$, a teda hypotéza $\boldsymbol{H}_0$ sa na hladine významnosti zamieta.

Keďže v tomto teste neboli splnené predpoklady, bude spravený ešte jeho neparametrická alternatíva, ktorou je *Dvojvýberový Wilcoxonov test (modifikovaný na Mannov-Whitneyov test)*. Ten však nepracuje so strednými hodnotami, ale s mediánmi. Preto matematické hypotézy budú prepísané nasledovne

$$\boldsymbol{H}_0 : \widetilde{SV} > \widetilde{SH},$$

$$\boldsymbol{H}_1 : \widetilde{SV} < \widetilde{SH}.$$

U tohto testu vyšla *p-hodnota* $= 2,4712 \times 10^{-5}$ a preto ako u parametrického testu sa hypotéza $\boldsymbol{H}_0$ na hladine významnosti zamieta.

Z výsledkov, ktoré vyšli, je vidieť vplyv domáceho prostredia na počet striel mierených na bránu v zápase. Aj keď sa hokej nehraje na strely, bez striel na bránu sa skórovať nedá. Výsledok môže zapríčiniť to, že domácemu tímu sedí viac domáce klzisko, na ktorom trénujú a tiež aj odvaha, ktorú im dodáva publikum.

4.4 Vplyv domáceho prostredia na úspešnosť vhadzovaní

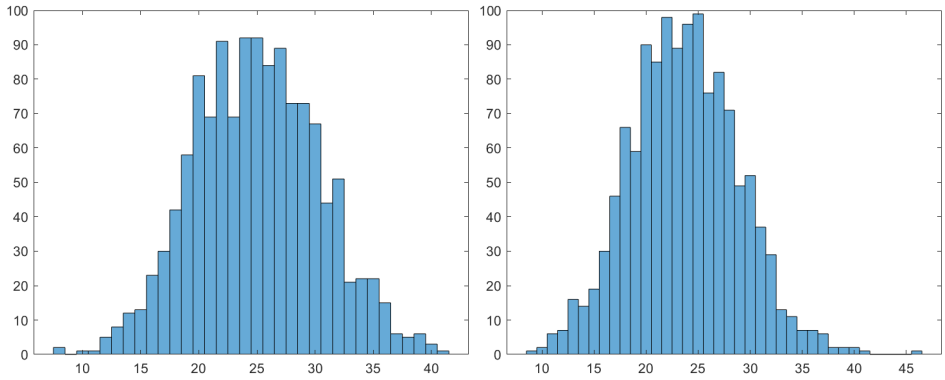
Výhoda domáceho prostredia bola pri strelených góloch a strelách na bránu potvrdená. Vyhrávajú aj hráči domácich tímov viacej vhadzovaní ako ich súperi?

Pokladá sa hypotéza $\boldsymbol{H}_0$: *hosťujúci hráči viacej vyhrávajú vhadzovaní v zápasoch*, proti alternatívnej hypotéze $\boldsymbol{H}_1$: *domáci hráči viacej vyhrávajú vhadzovaní v zápasoch*. Matematicky budú hypotézy zapísané nasledovne.

$$\boldsymbol{H}_0 : \mu_{FOV} > \mu_{FOH},$$

$$\boldsymbol{H}_1 : \mu_{FOV} < \mu_{FOH}.$$

Pre overenie hypotézy bude opäť použitý *Dvojvýberový t-test* v prípade, že budú splnené predpoklady testu. Najprv sa pri parmatrických testoch overí *normalita* testovaných súborov. Histogramy pre použité dáta vyzerajú nasledovne: vyhraté vhadzovania domácich $\boldsymbol{FOH}$ sú vľavo a vyhraté vhadzovania hostí $\boldsymbol{FOV}$ vpravo viď obrázok 6.



Obr. 6: Histogramy $\boldsymbol{FOH}$ (vľavo) a $\boldsymbol{FOV}$ (vpravo)

Dáta $\boldsymbol{FOH}$ viacej pripomínajú Gaussovu krivku ako $\boldsymbol{FOV}$. Avšak výsledky AD testov pri oboch prípadoch vyšli s *p-hodnotami* menšími, ako najnižšia tabelovaná hodnota softvérom Matlab, ktorou je 0,0005. Preto sa normalita dát na hladine významnosti zamieta. Keďže sa pracuje s veľkým počtom zápasov a *Dvojvýberový t-test* pre $n > 30$ je robustný aj pri porušenej normalite, môže mať pre závery výpovednú hodnotu. Ako ďalšie bude testovaná pomocou *F-testu* zhoda rozptylov oboch súborov. Vyšla *p-hodnota* $= 0{,}1775$ a preto sa test na zhodnosť rozptylov na hladine významnosti nezamieta, a tak bude použitý *Dvojvýberový t-test* pre rovnaké rozptyly. Po výpočte testu je výsledok *p-hodnotu* $= 1,2914 \times 10^{-10}$. A teda hypotéza $\boldsymbol{H}_0$ sa na hladine významnosti zamieta.

Keďže v tomto teste neboli splnené predpoklady bude spravená ešte jeho neparametrická alternatíva, ktorou je *Dvojvýberový Wilcoxonov test (modifikovaný na Mannov-Whitneyov test)*. Ten však nepracuje so strednými hodnotami ale s mediánmi. Preto matematické hypotézy budú prepísané nasledovne

$$\boldsymbol{H}_0 : \widetilde{FOV} > \widetilde{FOH},$$

$$\boldsymbol{H}_1 : \widetilde{FOV} < \widetilde{FOH}.$$

U tohto testu vyšla *p-hodnota* $= 3,3597 \times 10^{-10}$ a preto ako u parametrického testu sa hypotéza $\boldsymbol{H}_0$ na hladine významnosti zamieta.

Ako je vidieť z výsledkov, aj na vyhrané vhadzovania má vplyv domáce prostredie. Z toho plynie, že aj hráči domácich tímov majú vačší kľud na hokejkách pri vyhrávaní vhadzovaní doma, ako vonku. Ale závažný dopad na výsledok zápasu to podľa štúdie [9] nemá. Podľa jej autorov, vzťah medzi gólmi a vhadzovaniami je tak malý, že je potrebných približne 75 vyhraných vhadzovaní na zvýšenie skóre o jeden gól. Je to však zaujímavý poznatok, že aj na vhadzovania vplýva výhoda domáceho prostredia.

4.5 Vplyv domáceho prostredia na počet vylúčení

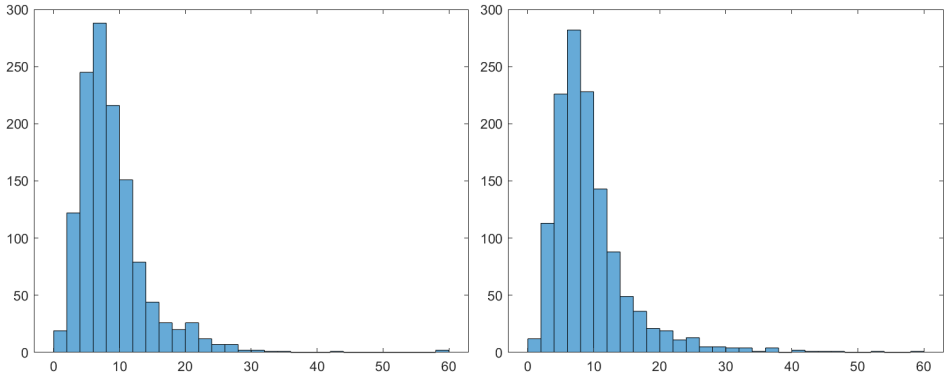
Hokej má jasne stanovené a definované pravidlá viď [6]. Organizátory líg a zápasov neustále zlepšujú VAR systémy a video rozhodcov, ale stále pri rozhodovaní a najmä

u vylúčení záleží na ľudskom faktore. Nižšie bude preverené, či sú vylučovaný viacej hostia ako domáci.

Pokladá sa hypotéza $\boldsymbol{H}_0$: *domáci sú viacej vylučovaní v zápasoch*, proti alternatívnej hypotéze $\boldsymbol{H}_1$: *hostia sú viacej vylučovaní v zápasoch*. Matematicky hypotézy budú zapísané nasledovne

$$\boldsymbol{H}_0 : \mu_{PH} > \mu_{PV},$$
$$\boldsymbol{H}_1 : \mu_{PH} < \mu_{PV}.$$

Pre overenie hypotézy bude opäť použitý *Dvojvýberový t-test* v prípade, že budú splnené predpoklady testu. Najprv sa pri parmatrických testoch overí *normalita* testovaných súborov. Histogramy pre použité dáta vyzerajú nasledovne: trestné minúty domácich $\boldsymbol{PH}$ sú vľavo a trestné minúty hostí $\boldsymbol{PV}$ vpravo viď obrázok 7.



Obr. 7: Histogramy $\boldsymbol{PH}$ (vľavo) a $\boldsymbol{PV}$ (vpravo)

Dáta $\boldsymbol{PH}$ a ani $\boldsymbol{PV}$ nepripomínajú Gaussovu krivku, a aj výsledky AD testov pri oboch prípadoch vyšli s *p-hodnotami* menšími, ako najnižšia tabelovaná hodnota softvérom Matlab, ktorou je 0,0005. Preto sa normalita dát na hladine významnosti zamieta. Keďže sa pracuje s veľkým počtom zápasov a *Dvojvýberový t-test* pre $n > 30$ je robustný aj pri porušenej normalite, môže mať pre závery výpovednú hodnotu. Ako ďalšie bude otestovaná pomocou *F-testu* zhoda rozptylov oboch súborov. Výpočtom vyšla *p-hodnota* $= 7,699 \times 10^{-6}$ a preto sa test na zhodnosť rozptylov na hladine významnosti zamieta, a použije sa *Welchov test*. Po výpočte *Welchovho testu* je výsledok *p-hodnoty* = 0,008 a tým pádom sa hypotéza $\boldsymbol{H}_0$ na hladine významnosti zamieta.

Keďže v tomto teste neboli splnené predpoklady, bude spravený ešte jeho neparametrická alternatíva, ktorou je *Dvojvýberový Wilcoxonov test (modifikovaný na Mannov-Whitneyov test)*. Ten však nepracuje so strednými hodnotami, ale s mediánmi. Preto matemtické hypotézy budú prepísané nasledovne

$$\boldsymbol{H}_0 : \widetilde{PH} > \widetilde{PV},$$
$$\boldsymbol{H}_1 : \widetilde{PH} < \widetilde{PV}.$$

U tohto testu vyšla *p-hodnota* = 0,0274 a tým, podobne ako u paramtrického testu, sa hypotéza $\boldsymbol{H}_0$ na hladine významnosti zamieta.

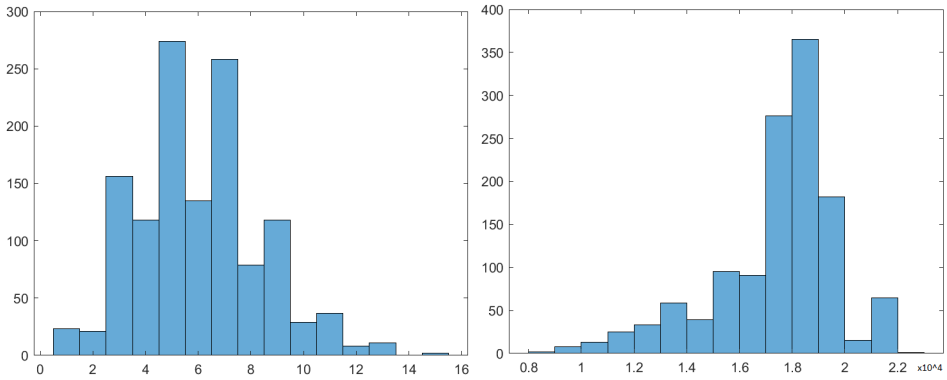
Výsledky ukázali, že hostia viac faulujú alebo sú viac vylučovaní, ako domáci. NHL má profesionálnych hráčov a takisto profesionálnych rozhodcov. Preto sa prikláňam k tomu, že rozhodcovia používajú na domáci rovnako i na hosťujúci tím rovnaký meter. Výsledky, ktoré boli zistené, poukazujú skôr na väčšiu nervozitu hosťujúceho tímu a aj na väčší počet

faulov na jeho strane. Ako aj v štúdii [11] bola preukázaná výhoda domáceho prostredia, tak aj u všetkých mnou skúmaných herných činností.

4.6 Vzťah počtu divákov a gólov

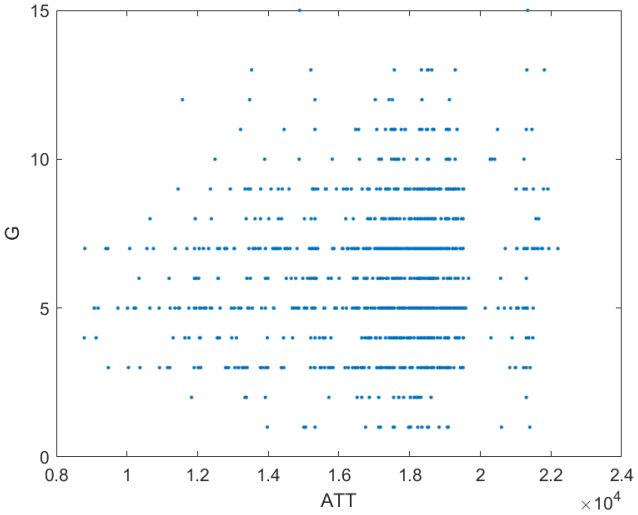
Výhoda domáceho prostredia sa preukázala tým, že domáce tímy strielajú viac gólov, ako ich súperi. Ďalej bude preverené, či celkový počet divákov na stretnutí ovplyňuje počet strelených gólov v zápase, a či je medzi týmito veličinami nejaký vzťah.

Lineárnu závislosť týchto dvoch veličín bude otestovaná pomocou *Pearsonovho korelačného koeficientu* v prípade splnených predpokladov. Ako prvé sa otestuje *normalitu* oboch súborov. Najprv bude preverená pomocou histogramov, ktoré vyzerajú nasledovne: celkový počet strelených gólov za zápas $\boldsymbol{G}$ je vľavo a celková návšteva na stretnutí $\boldsymbol{Att}$ vpravo viď obrázok 8.



Obr. 8: Histogramy $\boldsymbol{G}$ (vľavo) a $\boldsymbol{Att}$ (vpravo)

Ako je vidieť, dáta sa ani v jednom prípade nepodobajú na Gaussovu krivku. Potvrdil to aj AD test, ktorý v oboch prípadoch vyšiel s *p-hodnotou* menšou, ako najnižšie tabelovaná hodnota Matlabu 0,0005. Keďže sa pracuje s veľkým počtom zápasov, môže mať korelačný koeficient nejakú výpovednú hodnotu. Normalita dát na hladine významnosti sa teda v oboch prípadoch zamieta. Dáta sú vidieť spolu v grafe na obr. 9 nižšie.



Obr. 9: Bodový graf $\boldsymbol{Att}$ a $\boldsymbol{G}$

Pearsonov korelačný koeficient bol spočítaný softvérom Matlab. Pre koeficient vyšla *p-hodnota* = 0,0091, ktorá je menšia ako α a to značí, že korelácia sa podstatne líši od 0. *Pearsonov korelačný koeficient* je $R_{Pear} = 0,0732$.

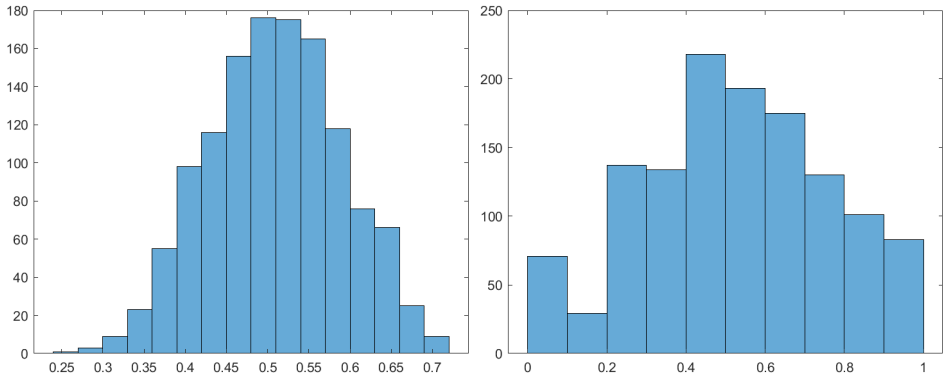
Keďže nebola potvrdená normalita dát, v oboch prípadoch bude použitý aj *Spearmanov korelačný koeficient*, pre ktorý vyšla *p-hodnota* = 0,0358. Aj tu je *p-hodnota* menšia ako α a značí to, že korelácia sa podstatne líši od 0. Výsledný *Spearmanov korelačný koeficient* je $R_{Spear} = 0,0589$.

Z výsledkov je vidieť len malú závislosť od počtu divákov na počet strelených gólov. Pretože korelácia vyšla blízka 0, dá sa povedať, že hráči hrajú skoro rovnako a nijak zásadne ich neovplyvňuje rozlišný počet divákov v publiku.

4.7 Vzťah aktivity tímu (Corsiho čislo) a gólov

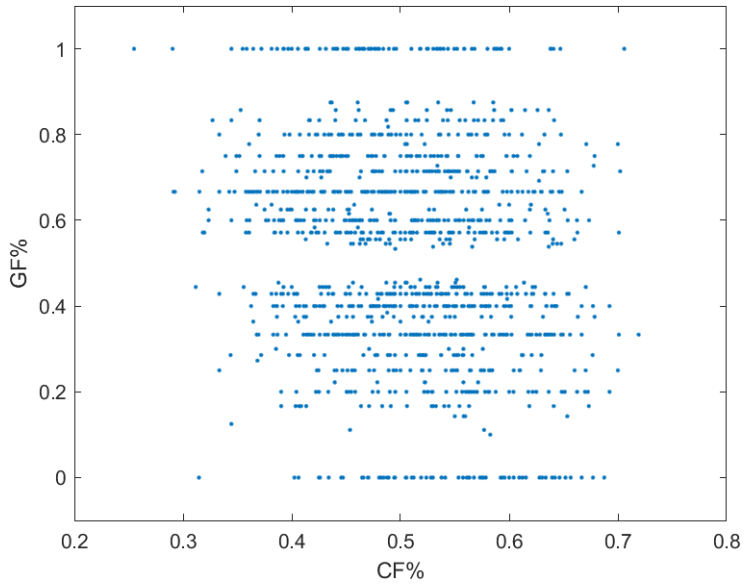
Ďalšie dve štatistiky, ktoré budú testované koreláciou, by mali mať spolu úzky vzťah. ***C*** počíta aktivitu tímu v danom zápase, a či daný tím viacej útočí alebo bráni. Tímy, ktoré viacej útočia, by mali v zápasoch strielať viacej gólov než tímy, ktoré väčšinu času strávili v obranom pásme. Bude tomu tak ?

Lineárnu závislosť týchto dvoch veličín sa otestuje pomocou *Pearsonovho korelačného koeficientu* v prípade splnených predpokladov. Nebudú medzi sebou porovnávané samotné štatistiky (***C***) a (***G***), ale ich percentuálne ekvivalenty (***CF%***) a (***GF%***). Tieto percentuálne ekvivalenty lepšie odzrkadlujú pomer aktivity tímov a výsledný pomer gólov. Ako prvé bude otestovaná *normalita* oboch súborov, podobne ako v predošlom prípade. Celkový percentuálny pomer aktivity ***CF%*** je vľavo a celkový percentuálny pomer gólov ***GF%*** vpravo viď obrázok 10.



Obr. 10: Histogramy ***CF%*** (vľavo) a ***GF%*** (vpravo)

GF% dáta sa nepodobajú na Gaussovu krivku. Potvrdil to aj AD test, ktorý vyšiel s *p-hodnotou* menšou ako najnižšie tabelovaná hodnota Matlabu 0,0005. Dáta však ***CF%*** Gaussovu krivku pripomínajú, ale normalitu *p-hodnota* = 0,0294 testu nepotvrdila. Tým pádom na hladine významnosti v oboch prípadoch je noramlita zamietnutá. Keďže sa pracuje s veľkým počtom zápasov, môže mať pre korelačný koeficient nejakú výpovednú hodnotu. Dáta je možné vidieť spolu v grafe na obr. 11 nižšie.



Obr. 11: Bodový graf $\boldsymbol{CF\%}$ a $\boldsymbol{GF\%}$

Pearsonov korelačný koeficient bol taktiež spočítaný softvérom Matlab. Pre koeficient vyšla *p-hodnota* $= 4,7624 \times 10^{-9}$, ktorá je menšia ako α a to značí, že korelácia sa podstatne líši od 0. *Pearsonov korelačný koeficient* je $R_{Pear} = -0,1633$.

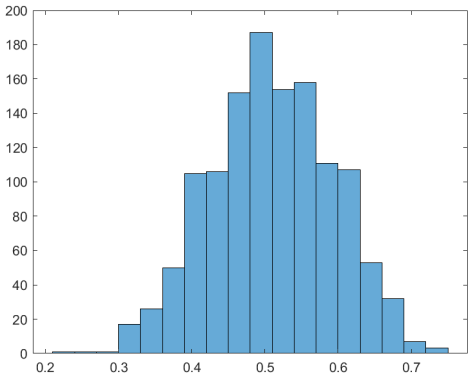
Keďže nebola potvrdená normalita dát v oboch prípadoch bude použitý aj *Spearmanov korelačný koeficient*, pre ktorý vyšla *p-hodnota* $= 3,2078 \times 10^{-8}$. Aj tu je *p-hodnota* menšia ako α a značí to, že korelácia sa podstatne líši od 0. Výsledný *Spearmanov korelačný koeficient* je $R_{Spear} = -0,1543$.

Z prekvapivých výsledkov je vidieť nepriamu závislosť pomeru aktivity tímov na pomere vystrelených gólov. Dá sa pozorovať, že aktivita tímov má opačný efekt na celkový výsledok zápasu. Avšak koeficient korelácie je v oboch prípadoch malý, a nedá sa jednoznačne preukázať nepriamu závislosť týchto veličín. Výsledky môžu mať ale logické vysvetlenie. Tímy, ktoré v zápasoch vedú, nemusia vyvíjať takú aktivitu smerom dopredu ako tímy, ktoré prehrávajú. Toto môže zapríčiniť mnou zistenú zápornú koreláciu. Zaujímavé by bolo pozorovať tento vzťah na podrobnejších dátach, rozkúskovaných minimálne na tretiny zápasu, poprípade po zmenách skóre.

4.8 Vzťah aktivity tímu (Fenwickovo čislo) a gólov

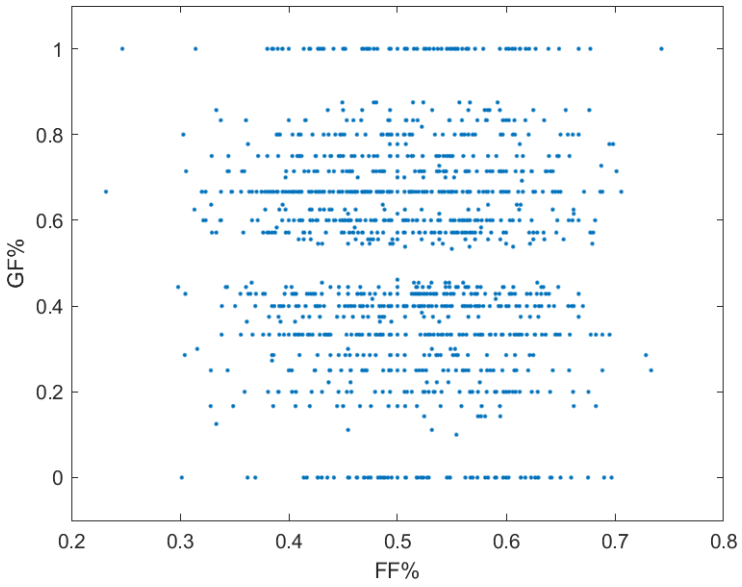
Po zaujímavých zisteniach s $\boldsymbol{C}$ budú očakávané podobné výsledky aj u $\boldsymbol{F}$. Uvidí sa, či obidve aktivity majú rovnakú výpovednú hodnotu.

Lineárnu závislosť týchto dvoch veličín bude ootestovaná pomocou *Pearsonovho korelačného koeficientu* v prípade splnených predpokladov. Podobne, ako pri predchádzajúcich hypotézach, nebudú medzi sebou porovnávané samotné štatistiky, ale ($\boldsymbol{F}$) a ($\boldsymbol{G}$) a taktiež ich percentuálne ekvivalenty ($\boldsymbol{FF\%}$) a ($\boldsymbol{GF\%}$). Ako prvé bude otestovaná *normalita* $\boldsymbol{FF\%}$ súboru, nakoľko $\boldsymbol{GF\%}$ je už otestovaný a testami neprešlo. Histogram celkového percentuálneho pomeru aktivity $\boldsymbol{FF\%}$ viď obrázok 12.



Obr. 12: Histogram ***FF%***

Dáta ***FF%*** Gaussovu krivku pripomínajú, avšak *p-hodnota* = 0,0203 normalitu dát pri AD teste zamietla. Keďže sa pracuje s veľkým počtom zápasov, môže mať korelačný koeficient nejakú výpovednú hodnotu. Dáta je možné vidieť spolu v grafe na obr. 13 nižšie.



Obr. 13: Bodový graf ***CF%*** a ***GF%***

Pre *Pearsonov korelačný koeficient* vyšla *p-hodnota* = 0,0481, ktorá je menšia ako α a to značí, že korelécia sa podstatne líši od 0. No pre nesplné predpoklady sa jej nedá veľmi veriť. *Pearsonov korelačný koeficient* je $R_{Pear} = -0,0554$.

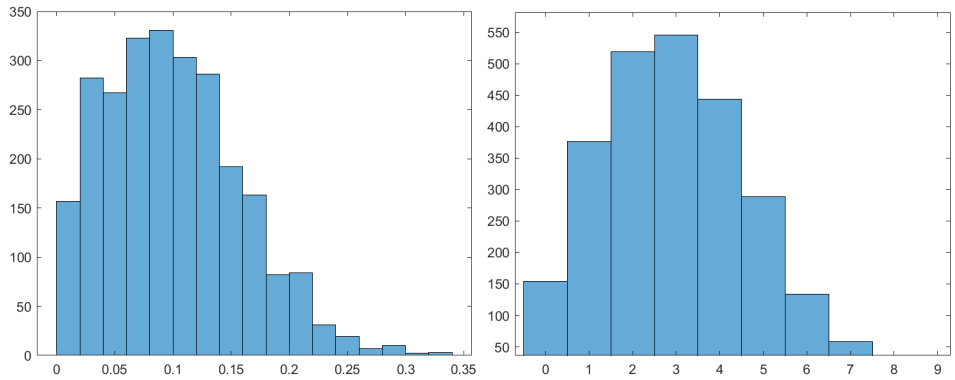
Keďže nebola potvrdená normalita dát, v oboch prípadoch bude použitý aj *Spearmanov korelačný koeficient*, pre ktorý vyšla *p-hodnota* = 0,0635. Tu je *p-hodnota* vyššia ako α a to značí, že korelácia sa podstatne nelíši od 0.

V tomto prípade len z parametrického testu vyšla malá negatívna závislosť a neparametrický test nepotvrdil žiadnu koreláciu. ***F*** aktivita nemá žiadny vplyv na počet strelených gólov. Nakoľko sa aj v tomto prípade jedná o dáta z celého zápasu, ich výpovedná hodnota nemusí byť veľmi presná.

4.9 Vzťah presnosti strelby a gólov

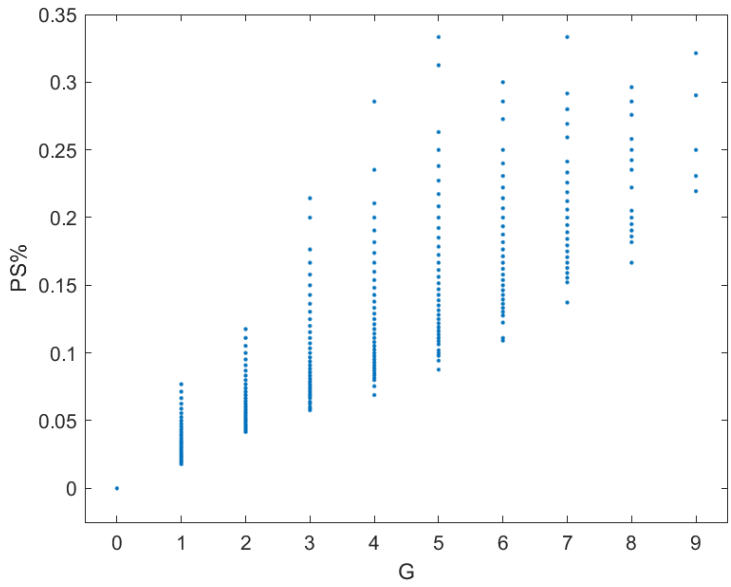
Ďalej bude pozorovaná presnosť strelby (***PS%***) ako súvisí s počtom gólov. Bude sa zisťovať, či tímy musia k vačšiemu počtu gólov vyslať na bránu viacej striel alebo záleží len na presnosti.

PS% sa vypočíta ako podiel gólov strelených tímom a počtu striel tímu za jeden zápas. Presnosť streľby bola napočítaná pre každý zápas dvakrát, preto počet testovaných súborov je dvojnásobný na rozdiel od hypotéz vyššie. Lineárnu závislosť týchto dvoch veličín bude otestovaná pomocou *Pearsonovho korelačného koeficientu* v prípade splnených predpokladov. Ako prvé sa otestuje *normalitu* oboch súborov. Presnosť streľby ***PS%*** je vľavo a počet gólov tímu za zápas ***GZ*** vpravo viď obrázok 14.



Obr. 14: Histogramy ***PS%*** (vľavo) a ***GZ*** (vpravo)

Ako je vidieť, dáta sú v oboch prípadoch vychýlené do ľavej strany a veľmi nepripomínajú Gaussovu krivku. To potvrdil aj AD test, ktorý v oboch prípadoch vyšiel s *p-hodnotou* menšou ako najnižšie tabelovaná hodnota Matlabu 0,0005. Keďže sa pracuje s ešte väčším počtom zápasov ako doposiaľ, môže mať korelačný koeficient tiež výpovednú hodnotu. Normalita dát sa na hladine významnosti teda v oboch prípadoch zamieta. Dáta sú vidieť spolu v grafe na obr. 15 nižšie.



Obr. 15: Bodový graf ***PS%*** a ***GZ***

Pre *Pearsonov korelačný koeficient* vyšla *p-hodnota* $= 0$, ktorá je menšia ako α a to značí, že korelécia sa podstatne líši od 0. *Pearsonov korelačný koeficient* je $R_{Pear} = 0,9071$.

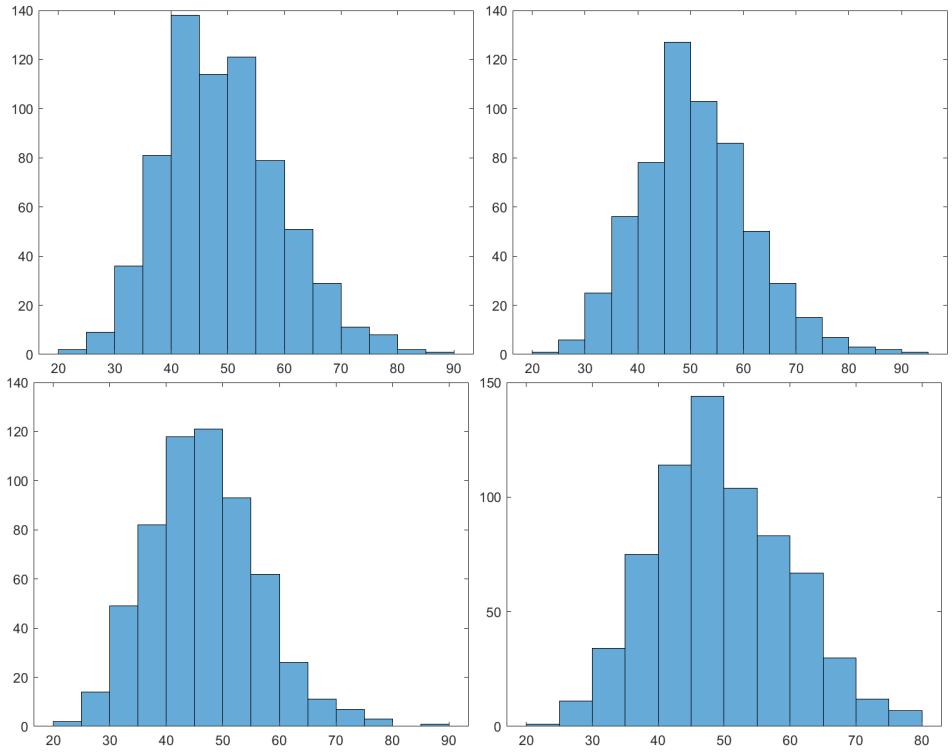
Keďže nebola potvrdená normalita dát, v oboch prípadoch bude použitý aj *Spearmanov korelačný koeficient*, pre ktorý vyšla tiež *p-hodnota* $= 0$. Aj tu je *p-hodnota* menšia ako α a značí to, že korelácia sa podstatne líši od 0. Výsledný *Spearmanov korelačný koeficient* je $R_{Spear} = 0,9282$.

Z výsledkov je jasne vidieť veľkú závislosť $\boldsymbol{PS\%}$ na počte strelených gólov. Poukazuje na to, že tímy ktoré v zápase strelili vyšší počet gólov, nemusia nutne vystreliť na bránu úmerný počet striel. Ale skôr záleží na lepšie vypracovaných šanciach s vyššou presnosťou streľby. Z grafu je vidieť, že tímy, ktoré strelili menší počet gólov (1,2), nedali viac gólov kvôli zle vypracovaným šanciam, a nie kvôli tomu že málo strielali na bránu.

4.10 Závislosť Corsiho čísla na prostredí a výsledku zápasu

Doteraz bol sledovaný vplyv jednej veličiny na jednu štatistiku. Teraz sa preverí, ako je ovplyvnená celková aktivita tímu $\boldsymbol{C}$, v závislosti na prostredí a výsledku zápasu.

K zisteniu týchto závislostí bude využitá *Dvojfaktorová analýza rozptylu bez interakcií*. Ako prvé sa overia predpoklady tohto testu, ktorými sú *normalita* a zhodnosť rozptylov. Najprv sa skontroluje normalita dát. Histogramy pre všetky skupiny vyzerajú nasledovne: aktivita tímu domáci vyhrá (DV) je vľavo hore, aktivita tímu domáci prehrá (DP) je vpravo hore, aktivita tímu hosť vyhrá (HV) je vľavo dole a aktivita tímu hosť prehrá (HP) je vpravo dole viď obrázok 16.



Obr. 16: Histo. DV (vľavo hore), DP (vpravo hore), HV (vľavo dole) a HP (vpravo dole)

Dáta sa pripomínajú Gausovu krivku v niektorých prípadoch viac v iných menej, ale sú aspoň približne symetrické. AD testy pre skupiny vyšli nasledovne : pre DV bola *p-hodnota* nižšia ako najnižšie tabelovaná hodnota Matlabom 0,0005, pre DP bola

p-hodnota $= 7,9316 \times 10^{-4}$, pre HV bola *p-hodnota* = 0,0089 a pre HP vyšla *p-hodnota* $= 5,0071 \times 10^{-4}$. Preto sa noramlita dát u všetkých skupín na hladine významnosti α zamieta.

Ďalší preverovaný predpoklad je zhodnosť rozptylov. Testovaná zhodnosť rozptylov všetkých skupín bola potvrdená pomocou Levenova testu viď [16]. Pre tento test vyšla *p-hodnota* = 0,1709, ktorá je vyššia ako mnou zvolená hladina významnosti α.

Keďže *dvojfaktorová analýza rozptylu bez interakcií* nieje až tak citlivá pre tento nesplnený predpoklad hlavne u početnejších dát, test bude mať výpovednú hodnotu. Napočítané výsledky *dvojfaktorovej analýzy rozptylu bez interakcií* popisuje tabuľka 4 nižšie.

Variabilita	**Súčet štvorcov SS**	**Počet stupňov volnosti** *df*	**Podiel SS/*df***	**Testová štatistika**
Prostredie	2198,1	1	2198,1	20,88
Výsledok	3771	1	3771	35,82
Chyba	267306	2539	105,28	-
Celkový	272883,6	2541	-	-

Tabuľka 4: Analýza rozptylu dvojného triedenia bez interakcií pre použité dáta

P-hodnota pre riadkový faktor prostredia je rovná $5,1275 \times 10^{-6}$ a pre stĺpcový faktor výsledku je $2,4714 \times 10^{-9}$. V oboch prípadoch je menšia ako α, čo indikuje, že ako prostredie tak výsledok zápasu ovplyvňuje celkovú aktivitu tímov.

Z výsledkov je vidieť vplyv oboch faktorov na výsledné $\boldsymbol{C}$. Vplyv výsledku, ako už bolo spomenuté pri testovaní vzťahu $\boldsymbol{C}$ a $\boldsymbol{G}$, môže byť zapríčinený tým, že tímy ktoré v zápasoch vedú, nemusia vyvíjať takú aktivitu smerom vpred ako tímy, ktoré v zápase prehrávajú. Tento vplyv nám vyšiel tiež ako viacej ovplyvňujúci celkové $\boldsymbol{C}$. Vplyv prostredia môže mať naopak psychický účinok na hráčov, ktorí doma hrajú uvoľnenejšie a sú zároveň hnaní domácim publikom. Hráči vyvíjajú väčšiu aktivitu než súperi, ktorí sú unavení z cesty a bez pomoci a podpory svojich divákov. Aj keď vyššiel ako menej ovplyvňujúci stále hraje svoju roľu pri celkovom $\boldsymbol{C}$.

Záver

V prvej kapitole boli zavedené základné a pokročilé hokejové štatistiky, ktoré sú zaznamenané pri každom zápase v NHL. Bol kladený dôraz hlavne na $\boldsymbol{C}$ a $\boldsymbol{F}$, pretože nie každý tieto štatistiky pozná.

Druhá kapitola sa venovala hokejovým štúdiám, ktoré boli rozčlenené do dvoch skupín. Prvá skupina využívala zložitejšie štatistické aparáty a dáta, ktoré nie sú vo väčšej miere verejnosti dostupné. To však viedlo k tomu, že výsledky z nich boli zaujímavejšie, a mali väčší prínos pre hokej. Druhá skupina sa zaoberala jednoduchými štatistikami, ktoré pracovali s podobnými dátami, ktoré boli použité v tejto práci.

V ďalšej kapitole bolo hneď v úvode zadefinované normálne rozdelenie, ktoré je ako jedno z hlavných predpokladov pri parametrických testoch použité v poslednej kapitole. Ďalej boli popísané testy, ktorými tento predpoklad na dátach bude zisťovaťovaný. Z parametrických testov boli uvedené: dvojvýberový t-test, korelácia a dvojvýberová analýzu rozptylu. Následne v závere tejto kapitole boli pridané aj ich neparametrické ekvivalenty.

Posledná kapitola obsahovala testovanie hypotéz na reálnych dátach zo základnej časti NHL sezóny 2018/19. Cieľom bolo nie len testovať hokejové hypotézy, ale následne ich aj interpretovať v reči hokeja. Bola zisťovaná výhoda domáceho prostredia, vzťah rôznych štatistík na počet strelených gólov a tímová aktivita v závislosti na dvoch faktoroch.

Výhoda domáceho prostredia bola potvrdená vo všetkých štyroch hypotézach s týmto zameraním. Prekvapením bolo najmä zistenie väčšieho počtu vylúčení na hosťujúcej strane. To mohlo zapríčiniť aj zistenie viac strelených gólov domácimi tímami, kde práve góly z presiloviek mohli byť rozhodujúcim faktorom pri tomto zistení.

Pri testovaní korelácie niektorých štatistík a gólov boli zistené zaujímavé skutočnosti. Prvé bolo zistenie zápornej korelácie medzi $\boldsymbol{C}$ a $\boldsymbol{G}$. Avšak výsledný korelačný koeficient nebol tak veľký, aby mohla byť s pokojným svedomím táto závislosť prehlásená za hodnotnú. Ďalšie bolo zistenie korelácie, medzi $\boldsymbol{PS}$ a $\boldsymbol{GZ}$, kde oba koeficienty boli väčšie ako 0,9. Tento výsledok poukazuje aj na fakt, že tímy v zápasoch nemajú až tak rozdielne počty striel, a skôr ako od kvantity záleží na kvalite pokusov vyslaných na bránu.

Na konci tejto kapitoly bol u analýzy rozptylu bez inerakcií potvrdený vplyv oboch faktorov na výslednú aktivitu tímov. Ako bolo spomenuté v závere sekcie **5.7**, aj tu by bolo lepšie do budúcna preveriť zistené skutočnosti na podrobnejších dátach. Vďaka tomu by sa eliminovali nepresnosti a výpovedná hodnota zistení bola vyššia.

Sledovanie a analyzovanie hokejových dát má veľký potenciál a prínos pre tento šport. Na základe zistení z analýzy, tímy dokážu pracovať na svojich slabinách a zlepšovať tak svoje taktiky a stratégie v budúcich zápasoch. V hokeji však existuje veľmi veľa faktorov, ktoré ani tie najdokonalejšie matemtické modely nedokážu predpovedať. Vďaka tomu hokej ostáva stále športom, ktorý je plný zvratov a prekvapení.

Literatúra

[1] ANDĚL, Jiří. Základy matematické statistiky. Praha: Matfyzpress, 2005. ISBN 80-86732-40-1.

[2] ANDĚL, Jiří. Statistické metody. Praha: Matfyzpress, 2005. ISBN 80-7378-003-8.

[3] MONTGOMERY, Douglas. Applied statistics and probability for engineers. New York: John Wiley, 2003. ISBN 0-471-20454-4.

[4] Testy normality. *WikiSkripta* [online]. 27. 3. 2017 [cit. 18-03-2021]. Dostupné z: https://www.wikiskripta.eu/w/Testynormality

[5] Adtest *MathWorks* [online]. © 1994-2021 [cit. 24-03-2021]. Dostupné z: https://ch.mathworks.com/help/stats/adtest.html

[6] Pravidla ľadového hokeja. *Slovenský zväz ľadového hokeja* [online]. 2018 [cit. 25-04-2021]. Dostupné z: https://www.hockeyslovakia.sk/sk/article/pravidla-ladoveho-hokeja

[7] Corsi (statistic). In: *Wikipedia: the free encyklopedia* [online]. San Francisko (CA): Wikimedia Foundation, 2001 [cit. 01-05-2021]. Dostupné z: https://en.wikipedia.org/wiki/Corsi_(statistic)

[8] Fenwick (statistic). In: *Wikipedia: the free encyklopedia* [online]. San Francisko (CA): Wikimedia Foundation, 2001 [cit. 01-05-2021]. Dostupné z: https://en.wikipedia.org/wiki/Fenwick_(statistic)

[9] CZUZOJ-SHULMAN, Nick, David YU, Christopher BOUCHER, Luke BORNN a Mehrsan JAVAN. Winning Isn't Everything – A contextual analysis of hockey face-offs [online]. 2019 [cit. 05-05-2021]. Dostupné z: https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5f6d38d2f51633204ddb9c07_Sloan_2019_Faceoffs2.pdf

[10] YU, David, Christopher BOUCHER, Luke BORNN a Mershan JAVAN. Playing Fast Not Loose: Evaluating team-level pace of play in ice hockey using spatio-temporal possession data [online]. 2019 [cit. 06-05-2021]. Dostupné z: https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5f6d343cc58cd38c4b12f664_HockeyPace.pdf

[11] HOFFMANN, D. Matt, Todd M. LOUGHEAD, Jess C. DIXON a Alyson J. CROZIER. Examining the home advantage in the National Hockey League: Comparisons among regulation, overtime, and the shootout [online]. 2017 [cit. 07-05-2021] Dostupné z: https://www.sciencedirect.com/science/article/pii/S1469029216301558

[12] COCHRAN J. James a Rob BLACKSTOCK. Pythagoras and the National Hockey League [online]. 2009 [cit. 08-05-2021] Dostupné z: https://www.degruyter.com/document/doi/10.2202/1559-0410.1181/html

[13] FOUND Rob. Goal-based Metrics Better Than Shot-based Metrics at Predicting Hockey Success [online]. 2019 [cit. 11-05-2021]. Dostupné z: https://thesportjournal.org/article/goal-based-metrics-better-than-shot-based-metrics-at-predicting-hockey-success/

[14] LEE Christian. Are Shot Attempts and Shots on Goal Meaningful Predictors of NHL Game Outcomes? Not Really [online]. 2019 [cit. 11-05-2021]. Dostupné z: https://medium.com/hockey-stats/are-shot-attempts-and-shots-on-goal-meaningful-predictors-of-nhl-game-outcomes-not-really-f8f8d16811bf

[15] 2018-19 NHL Schedule and Results. In: Hockey Reference [online]. Sports Reference LLC © 2000-2021 [cit. 05-02-2021] Dostupné z: `https://www.hockey-reference.com/leagues/NHL_2019_games.html`

[16] Vartestn *MathWorks* [online]. © 1994-2021 [cit. 18-05-2021]. Dostupné z: `https://ch.mathworks.com/help/stats/vartestn.html`