Univerzita Karlova

Pedagogická fakulta

Katedra matematiky a didaktiky matematiky

BAKALÁŘSKÁ PRÁCE

Využití sportovní tématiky ve statistických úlohách pro střední školy

Use of sport contexts in statistical problems for upper secondary schools

Kateřina Koštejnová

Vedoucí bakalářské práce: prof. RNDr. Jarmila Novotná, CSc

Studijní program: Specializace v pedagogice (B7507)

Studijní obor: Bakalářské M jednoobor od 2017 (ODBAMA17)

2022

Odevzdáním této bakalářské práce na téma Využití sportovní tématiky ve statistických úlohách pro střední školy potvrzuji, že jsem ji vypracovala pod vedením vedoucího práce samostatně za použití v práci uvedených pramenů a literatury. Dále potvrzuji, že tato práce nebyla využita k získání jiného nebo stejného titulu.

Svijany, 12. 4. 2022

Zde bych ráda poděkovala své vedoucí práce, prof. RNDr. Jarmila Novotná, CSc, za nápomocné rady, velkou trpělivost, a veškerý čas, který mi po dobu psaní bakalářské práce věnovala.

ABSTRAKT

Práce se zabývá využitím sportovní tématiky v úlohách ze statistiky pro střední školy. V práci jsou shrnuty základní statistické pojmy pro různé stupně škol. Hlavní částí práce je sbírka autorských řešených úloh se sportovní tématikou. Úlohy jsou rozděleny do tří kapitol na základě teoretických podkladů k jednotlivým statistickým oblastem. V první kapitole se zabýváme charakteristikami polohy a také čtením z grafů a jejich konstrukcí, dále v ní využíváme zadávání úloh pomocí tabulek; ve druhé kapitole se zaměřujeme na charakteristiky variability a v poslední kapitole je představena korelace. Úlohy v první kapitole jsou s fotbalovou tématikou; ve druhé kapitole využíváme hokejové tématiky a třetí kapitola nabízí více různých sportovních odvětví. Na začátku každé kapitoly je zařazen přehled pojmů ze statistiky, které jsou pro řešení úloh nezbytné. V práci jsou zařazeny také některé úlohy, jejichž řešení vyžahuje teorii, pro jejíž zavedení je nutná znalost matematiky na úrovni vysoké školy. I tyto úlohy jsou doplněny o teoretický podklad pro jejich řešení. V teoretické části příslušné k těmto úlohám předpokládáme znalost integrálního počtu. Tato sbírka úloh s teoretickým podkladem pro jejich řešení může být použita učiteli matematiky na středních školách i žáky středních škol. Poslední část práce obsahuje ukázku použití statistiky při práci s daty v profesionálním sportu. Představujeme χ^2 test nezávislosti a jeho užití na datech, která se týkají návštěvnosti zápasů a gólových rozdílů z pohledu domácího týmu. Využíváme k tomu programovacího jazyka R v programu R studio.

KLÍČOVÁ SLOVA

statistika, sport, řešené úlohy, momenty, kvantily, test, hypotéza

ABSTRACT
This thesis explores the use of sports-related themes in secondary school level statistical problems and exercises. It summarizes basic statistical terminology appropriate for various education levels. The core then consists of a set of original statistical sports-themed problems and their solutions. These problems are split into three chapters based on the theoretical foundation they require. The first chapter deals with location parameters as well as reading from graphs and their construction, and it utilizes tables in the problems' formulation. The next chapter then focuses on the properties of dispersion, and the last one introduces correlation. The problems presented in the first chapter make use of football themes, the ones from the second chapter are ice-hockey based, and lastly, the problems from the third chapter mix themes from various sports. Each presented exercise is preceded with a list of definition necessary for finding the solution. The thesis also includes questions, whose solutions require university level Mathematics. These are augmented with the necessary theoretical foundation as well but it assumes knowledge of integral calculus. This problem collection may be used by secondary school Mathematics educators. The last section demonstrates use of statistics, particularly hypothesis testing on data from professional sports. It introduces the χ^2 independence test and its applications on match attendance and the home team's advantage in terms of goal difference. We utilize the R programming language and its R studio editor.

KEYWORDS
Statistics, sport, solved problems, moments, quantiles, test, hypothesis

Obsah

Úvod

Moderní statistika, taková jakou ji známe dnes, se vyvíjí od 17. století. Počátek má v hazardních hrách. Nyní statistiku nevyužíváme jen k počítání v hazardních hrách. Hlavní význam přináší jako nástroj pro sběr, analýzu či interpretaci dat, které se týkají určitého jevu (společenského, přírodního, technického).

Podle mých zkušeností a analýz učebnic bývá statistika představována žákům ve formě vzorců, do kterých je třeba dosadit, místo toho, aby žáci pronikli do podstaty problémů. Pokud je statistická úloha zadána slovně (nejen hodnotami, ze kterých mají žáci vypočítat například aritmetický průměr), obvykle se zabývá jevy, které pro žáky nejsou příliš atraktivní (teplota, plat, věk, ...). Řešení úloh bývají pouze se stručnou (číselnou) odpovědí.

Cílem mé bakalářské práce je ukázat na úlohách se sportovní tématikou, co lze z výsledků statistických šetření zjistit, proč jsou různé statistické pojmy vhodné v různých situacích. Ukázat, jak se odlišují podobné statistické pojmy (např. aritmetický a harmonický průměr).

Sport jako téma statistických úloh jsem volila proto, že je pro žáky středních škol častou náplní volného času. Z mé zkušenosti žáky více zaujme úloha motivovaná aktivitou jim blízkou, než například rozdílností teplot v různých krajích České republiky. Volené sporty jsou všeobecně známé.

Pokud se v úlohách v práci objevuje nějaký pojem, který nemusí být známý, je vysvětlen na začátku dané úlohy. Největší prostor je věnován fotbalu a hokeji, protože jsou mezi žáky obecně nejznámější.

V částech práce, které jsou zaměřeny na statistiku na vysokých školách, předpokládáme základní znalost matematické analýzy (integrální počet).

Bakalářská práce je rozdělena do čtyř samostatných kapitol. První tři kapitoly obsahují úlohy a s nimi spojenou teorii pro žáky či studenty středních a vysokých škol. Čtvrtá kapitola představuje základní teorii k testování hypotéz a ukazuje využití konkrétního statistického testu na sportovních datech.

Na začátku každé kapitoly je nejprve zařazen přehled pojmů ze statistiky, které se v úlohách v kapitole využívají. Zavedené pojmy se objevují i v dalších kapitolách.

V první kapitole se čtenář seznámí s charakteristikami polohy. Důraz je kla-

den na průměry, modus či medián. Jednotlivé charakteristiky nemají stejně četné zastoupení. Je to proto, že například aritmetický průměr je používanější než průměr harmonický, s jehož použitím se ve sportu často nesetkáváme. Dále je v první kapitole zažazena práce s grafy a tabulkami. První kapitola obsahuje úlohy s fotbalovou tématikou.

Druhá kapitola je věnována charakteristikám variability (rozptýlenosti). Jak již název napovídá, základní charakteristiky jsou rozptyl a směrodatná odchylka. Druhá kapitola je doplněna o ukázku úloh, ke kterým jsou nutné znalosti na úrovni vysoké školy. Složitější úlohy jsou koncipované tak, aby žák výsledek interpretoval, popřípadě vypočítané charakteristiky porovnal. Pro tuto kapitolu je zvolena hokejová tématika.

Ve třetí kapitole je věnována pozornost popisu statistických dat pomocí korelace (závislosti). Čtenářovi je představen koeficient korelace a rozdíl mezi nezávislostí či přímou a nepřímou závislostí. I tato kapitola je doplněna o vysokoškolský pohled na korelaci s ukázkou řešení dvou úloh. Třetí kapitola obsahuje úlohy z prostředí biatlonu, košíkové a dalších sportů.

Čtvrtá kapitola je věnována ukázce využití statistiky jako nástroje pro analýzu reálných sportovních dat. Nejprve je představena teorie pro testování hypotéz, následně konkrétní χ^2 test a ukázka jeho použití. V ukázce je využit k testování fotbalových dat, která se týkají vlivu počtu diváků v hledišti na rozdíl inkasovaných a vstřelených gólů domácího týmu.

Všechny úlohy prezentované v této práci jsou autorské. Při jejich tvorbě jsem se inspirovala úlohami, které jsem počítala jako žákyně střední školy.

Ve všech úlohách (první, druhé i třetí kapitoly), v nichž počítáme charakteristiky (aritmetický průměr, rozptyl, korelační koeficient, ...) pro více hráčů (týmů, atd), budeme jednotlivé charakteristiky odlišovat různými indexy, na základě toho, k jakým datům se vztahují. Například aritmetický průměr gólů Jaromíra Jágra budeme značit $\bar{x_J}$. Výsledky mezivýpočtů se snažíme zapisovat desetinným číslem bez zaokrouhlení (pokud to není možné, necháváme číslo ve tvaru zlomku). Konečný výsledek zapisujeme celým, nebo desetinným číslem zaokrouhleným na dvě desetinná místa.

V kapitolách jsou použita data z https://www.livesport.cz/. Grafy v celé práci jsou tvořeny v aplikaci Excel. Pro testování dat ve čtvrté kapitole je využito programovacího jazyku R v programu R studio.

Kapitola 1

Charakteristiky polohy

V této části se zabýváme základním popisem statistických souborů pomocí úrovně (polohy). Pojmenování charakteristika polohy získaly proto, že díky nim je možné pomocí jedné číslovky popsat umístění znaku na číselné ose. Tuto hodnotu určujeme pomocí středních hodnot. Středními hodnotami nahrazujeme a zobecňujeme hodnoty souboru. Pokud počítáme střední hodnoty z celého souboru, nazýváme je průměry. Nejznámější a pro nás podstatné průměry jsou: aritmetický, geometrický a harmonický. Následně představujeme teorii kvantilů. Kvantil je hodnota v souboru rozdělující statistický soubor na dvě (a více) částí. Běžně hovoříme o polovinách, čtvrtinách, desetinách a setinách. Zaměřujeme se především na modus a medián, zmiňujeme však i kvartily, decily a percentily.

1.1 Základní pojmy

V částech práce, které se zabývají statistikou pro střední školy, pracujeme se statistickými jednotkami. Statistická jednotka je elementární prvek statistického zkoumání. Počet statistických jednotek statistického zkoumání značíme n. Vlastnost statistické jednotky popisuje statistický znak.

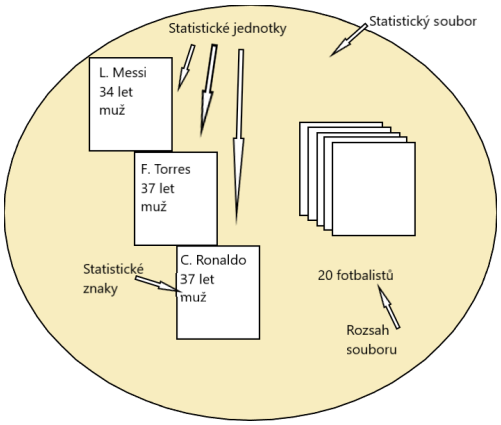
Poznámka (Kvantitativní znak, kvalitativní znak Calda a Dupač, 1993, str. 131). Znak, který určuje množství, nazýváme kvantitativní. Znak kvalitativní označuje znak, který není možné popsat číslem, označuje kvalitu.

Příklad. Kvantitativní znak: věk, výška, cena.
Kvalitativní znak: prospěl x neprospěl.

Definice 1 (Statistický soubor Hindsl a kol., 2018, str. 16)**.** *Množinu všech statistických jednotek, u nichž zkoumáme příslušné statistické znaky, nazýváme statistickým souborem.*

Značení. $x_1, x_2, ..., x_n$ jsou hodnoty znaku x zjištěné u jednotek $1, 2, ..., n$.
$x_1^*, x_2^*, ..., x_r^*$ jsou všechny možné různé hodnoty znaku x.

V diagramu na obrázku 1.1 vidíme vztah statistického znaku, jednotky a souboru.



Obrázek 1.1: Diagram statistických pojmů

V následujících definicích se věnujeme četnosti, relativní četnosti a rozdělení četností. V těchto definicích se zaměřujeme na případy, ve kterých u statistických jednotek pozorujeme pouze jeden statistický znak.

Definice 2 (Četnost Calda a Dupač, 1993, str. 156)**.** *Četnost n_j hodnoty x_j^* udává, u kolika jednotek byla tato hodnota zaznamenána.*

Definice 3 (Relativní četnost Calda a Dupač, 1993, str. 133)**.** *Relativní četnost označuje, jaká část souboru má hodnotu x_j^*. Značíme ji v_j. Platí:*

$$v_j = \frac{n_j}{n}.$$

Poznámka. Součet relativních četností v_j hodnot x_j^* se rovná 1.

$$\sum_{j=1}^{r} v_j = 1.$$

Relativní četnost je možné vyjádřit i v procentech. V takovém případě je součet relativních četností roven 100 %.

Definice 4 (Rozdělení četností znaku x Calda a Dupač, 1993, str. 156)**.** *Rozdělení četností znaku x přiřazuje hodnotám x_r^* jejich četnosti n_r. Obvykle se zapisuje do tabulky (tabulka 1.1):*

Hodnota znaku x	x_1^*	x_2^*	...	x_r^*
Četnost	n_1	n_2	...	n_r

Tabulka 1.1: Rozdělení četností znaku x

V následujících definicích uvažujeme pouze kvantitativní znak. Základní informaci o znaku x získáváme z čísla, které určuje polohu znaku na číselné ose. Významnou část měr polohy tvoří průměry, definujeme jich několik typů.

Definice 5 (Aritmetický průměr Hindsl a kol., 2018, str. 32)**.** *Nechť je dán statistický soubor o rozsahu n pozorování $x_1, x_2, ..., x_n$. Pak je aritmetický průměr $\bar{x}$ definován vzorcem:*

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Definice 6 (Vážený průměr Hindsl a kol., 2018, str. 32)**.** *Nechť je dán statistický soubor o rozsahu n pozorování $x_1, x_2, ..., x_n$ a hodnoty statistického znaku uspořádány do tabulky rozdělení četností. Pak je vážený průměr definován vzorcem:*

$$\bar{x} = \frac{x_1\ n_1 + x_2\ n_2 + ... + x_r\ n_r}{n_1 + n_2 + ... + n_r} = \frac{\sum_{i=1}^{k} x_i\ n_i}{\sum_{i=1}^{k} n_i}.$$

Následují definice prostého geometrického a harmonického průměru. U průměru geometrického musíme uvažovat hodnoty znaku nezáporné (ve vzorci se objevuje odmocnina). Ani jeden průměr není využíván tak často jako průměr aritmetický. Geometrický průměr se obvykle využívá v národohospodářství a harmonický průměr k vyjádření průměrné délky času potřebné k vykonání nějaké činnosti, kterou vykonává několik osob (strojů, ...) v jednu chvíli společně.

Definice 7 (Prostý geometrický průměr, Hindsl a kol., 2018, str. 35)**.** *Nechť je dán statistický soubor o rozsahu n pozorování $x_1, x_2, ..., x_n$. Pak je prostý geometrický průměr definován vzorem:*

$$\bar{x_G} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}.$$

Definice 8 (Prostý harmonický průměr Hindsl a kol., 2018, str. 34)**.** *Prostým harmonickým průměrem rozumíme převrácenou hodnotu aritmetického průměru převrácených hodnot. Pro výpočet využíváme vzorce:*

$$\bar{x_H} = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n})}.$$

Hodnoty statistického znaku ve statistickém souboru jsou rozděleny kvantily. Ty nejznámější jsou medián, kvartil, decil a percentil. Také zavádíme modus.

Definice 9 (Kvantil Hindsl a kol., 2018, str. 29)**.** *Kvantil je hodnota, která rozděluje uspořádané hodnoty statistického znaku v souboru na dvě části.*

Definice 10 (Modus Calda a Dupač, 1993, str. 142)**.** *Modus znaku x, značíme $Mod(x)$, je hodnota x s největší četností.*

Definice 11 (Medián Calda a Dupač, 1993, str. 142)**.** *Jsou-li $x_{(1)} \leq x_{(2)} \leq \leq .... \leq x_{(n)}$ hodnoty $x_1, x_2, ..., x_n$ uspořádané podle velikosti, pak medián znaku x, značí se $Med(x)$, je:*

$$Med(x) = x_{(\frac{n+1}{2})}, \qquad \textit{je-li n liché.}$$

$$Med(x) = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), \qquad \textit{je-li n sudé.}$$

Pro medián je také používán název prostřední hodnota.

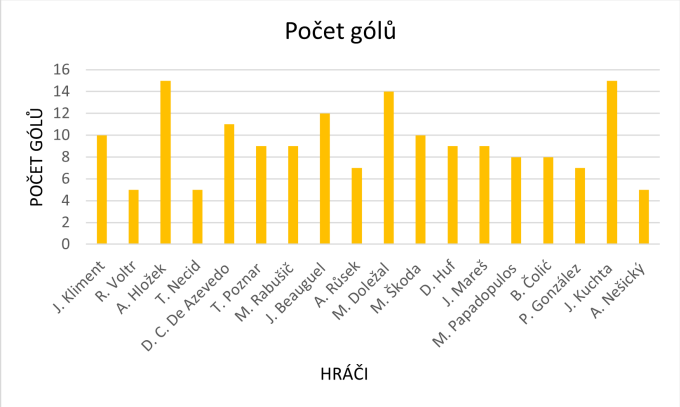
Definice 12 (Kvartily, decily, percentily Hindsl a kol., 2018, str. 30)**.** *Kvartily jsou hodnoty, které dělí uspořádaný statistický soubor na čtyři části, přičemž každá část obsahuje* 25 % *jednotek.*
Decily jsou hodnoty, které dělí uspořádaný statistický soubor na deset částí, přičemž každá část obsahuje 10 % *jednotek.*
Percentily jsou hodnoty, které dělí uspořádaný statistický soubor na sto částí, přičemž každá obsahuje 1 % *jednotek.*

1.2 Řešené úlohy

V této části uvádíme řešené úlohy, které jsou zaměřené na charakteristiky poloh (četnosti, průměry a kvantily) a také na čtení z grafu či tabulky. První dvě úlohy jsou jednodušší, zaměřené na dosazení do vhodně vybraného vzorce. Následuje osm složitějších, ve kterých je nutné odůvodnění a reprezentace výsledku. Úlohy jsou zaměřené na praktické použití vzorců z předchozí teoretické podkapitoly a jsou volené pro úroveň střední školy (gymnázia).

1.2.1 Řešené úlohy - jednodušší

Úloha 1. V grafu na obrázku 1.2 je znázorněn počet gólů nejlepších střelců jednotlivých týmů v minulé sezóně (2020/2021) ve Fortuna lize. Určete aritmetický průměr, medián a modus počtu gólů všech střelců v grafu na obrázku 1.2.



Obrázek 1.2: Počet gólů nejlepších střelců v minulé sezóně

Řešení. Aritmetický průměr počítáme pomocí vzorce:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Do vzorce dosadíme a dostáváme:

$$\bar{x} = \frac{168}{18} \doteq 9{,}33 \text{ (gólů)}.$$

Dále počítáme modus, což je hodnota s největší četností. Jednotlivé počty gólů a příslušné četnosti vidíme v tabulce 1.2.

Počet gólů	Četnost
15	2
14	1
12	1
11	1
10	2
9	4
8	2
7	2
5	3

Tabulka 1.2: Počet gólů nejlepších střelců

Tedy:

$$Mod(x) = 9 \text{ (gólů)}.$$

Pro výpočet $Med(x)$ využijeme vzorce:

$$Med(x) = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}),$$

protože počet všech prvků je sudý. Po dosazení do vzorce dostáváme:

$$Med(x) = 9 \text{ (gólů)}.$$

Aritmetický průměr je 9,33 vstřelených gólů. Medián je stejný jako modus, tedy 9 vstřelených gólů.

Úloha 2. V tabulce 1.3 vidíte počet vstřelených gólů anglického týmu FC Manchester city v Premier League za 19 zápasů na přelomu sezón 2020/21 a 2021/22 (od 3. dubna 2021 do 6. listopadu 2021). Uspořádejte počet gólů vzestupně a následně statistický soubor rozdělte pomocí kvartilů na příslušný počet částí.

Zápas	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Počet gólů	2	0	4	2	2	1	0	1	5	5
Zápas	11.	12.	13.	14.	15.	16.	17.	18.	19.	-
Počet gólů	1	5	2	4	2	2	2	1	2	-

Tabulka 1.3: Počet vstřelených gólů týmu Manchester city v Premier League

Řešení. Nejprve seřadíme jednotlivé počty gólů vzestupně:
0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 4, 4, 5, 5, 5.
Nyní rozdělíme statistický soubor pomocí kvartilů.
Hledáme hodnotu, která soubor rozděluje na poloviny:
0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 4, 4, 5, 5, 5;
a nyní hodnoty, která ho rozdělují na čtvrtiny:
0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 4, 4, 5, 5, 5.

Kvartily, které rozdělují soubor na čtvrtiny, jsou čísla 1 a 4 zvýrazněná zeleně a 2 zvýrazněná červeně.

1.2.2 Řešené úlohy - obtížnější

Úloha 3. V grafu na obrázku 1.3 jsou zaznamenány počty hráčů, kteří v ligové sezóně 2020/2021 Premier League vstřelili 2 – 3 góly, 4 – 6 gólů, 7 – 9 gólů, 10 – 11 gólů , 12 – 17 gólů a 18 – 23 gólů. Určete četnost a relativní četnost jednotlivých gólových rozmezí. Relativní četnost zakreslete do grafu a rozhodněte, zda se od sebe graf na obrázku 1.3 a vámi vytvořený graf strukturou liší (popřípadě jak), a zdůvodněte, proč tomu tak je.

Obrázek 1.3: Vztah množství vstřelených gólů a počtu hráčů

Řešení. V prvním kroku zapíšeme jednotlivé četnosti (pro přehlednost) do tabulky 1.4. Četnosti vidíme přímo z grafu na obrázku 1.3.

Počet gólů	Četnosti
2 - 3	64
4 - 6	32
7 - 9	22
10 - 11	13
12 - 17	11
18 - 23	5

Tabulka 1.4: Tabulka četností počtu gólů

Ve druhém kroku určíme relativní četnosti. Relativní četnosti vypočítáme jako podíl jednotlivých četností a součtu všech četností.
Relativní četnost počtu gólů 2 - 3: $\frac{64}{64+32+22+13+11+5} = \frac{64}{147}$,

relativní četnost počtu gólů 4 - 6: $\frac{32}{147}$,

relativní četnost počtu gólů 7 - 9: $\frac{22}{147}$,

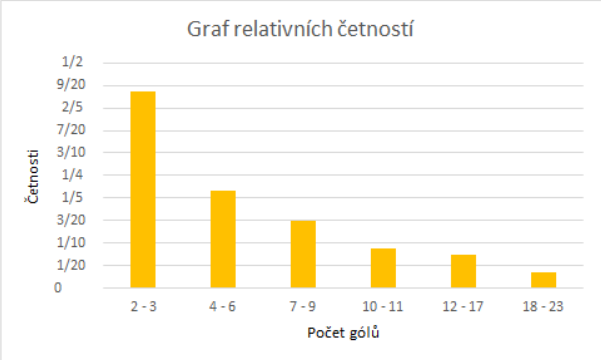
relativní četnost počtu gólů 10 - 11: $\frac{13}{147}$,

relativní četnost počtu gólů 12 - 17: $\frac{11}{147}$,

relativní četnost počtu gólů 18 - 23: $\frac{5}{147}$.

Pro přehlednost necháme relativní četnosti ve tvaru zlomku. Následná kontrola: součet všech relativních četností je roven 1.

Nyní vytvoříme graf relativních četností na obrázku 1.4. Na svislé ose jsou zaznamenány četnosti a na vodorovné ose gólová rozmezí.
Výšky sloupců v grafu v zadání i v grafu v řešení jsou ve stejném poměru. Koeficient k, kterým jsou hodnoty z grafu v zadání vynásobeny, je $k = \frac{1}{147}$. Oba grafy mají stejnou výpovědní hodnotu.



Obrázek 1.4: Graf relativních četností

Úloha 4. Jürgen Klopp, trenér Anglického týmu Liverpool, vypsal 20 nejmladších hráčů svého týmu. Na týdenní soustředění však chce vyvést přesně 2 hráče z každého věkového rozmezí, které vidíte v tabulce 1.5.

Věk	17 - 18	19 - 20	21 - 22	23 - 24	25 - 26	27 - 28	29
Četnost							

Tabulka 1.5: Tabulka četností

Potenciálními hráči jsou: Gordon Kaide 17 let, Elliott Harvey 18 let, Morton Tyler 19 let, Jones Curtis 21 let, Konaté Ibrahima 22 let, Kelleher Caoimhin 23 let, Alexander-Arnold Trent 23 let, Gomez Joe 24 let, Tsimikas Konstantinos 25 let, Diogo Jota 25 let, Díaz Luis 25 let, Origi Divock 26 let, Robertson Andrew 27 let, Keita Naby 27 let, Minamino Takumi 27 let, Fabinho 28 let, Oxlade-Chamberlain Alex 28 let, Alisson 29 let, Mane Sadio 29 let a Salah Mohamed 29 let. Doplňte tabulku četností 1.5 a popište, co vyjadřuje. Vysvětlete, zda má trenér potřebný počet hráčů, či zda je nutné někoho vyřadit, či přidat.

Řešení. Pro doplnění tabulky četností musíme spočítat, kolik hráčů patří do jednotlivých věkových rozmezí.

- Do prvního rozmezí patří 2 hráči Gordon Kaide 17 let, Elliott Harvey 18 let.
- Do druhého rozmezí patří 1 hráč Morton Tyler 19 let, žádný 20 letý hráč v týmu není.
- Do třetího rozmezí patří 2 hráči Jones Curtis 21 let, Konaté Ibrahima 22 let.
- Do čtvrtého rozmezí patří 3 hráči Kelleher Caoimhin 23 let, Alexander-Arnold Trent 23 let a Gomez Joe 24 let.
- Do pátého rozmězí patří 4 hráči Tsimikas Konstantinos 25 let, Diogo Jota 25 let, Díaz Luis 25 let a Origi Divock 26 let.
- Do šestého rozmezí patří 5 hráčů Robertson Andrew 27 let, Keita Naby 27 let Minamino Takumi 27 let, Fabinho 28 let, Oxlade-Chamberlain Alex 28 let.
- Do posledního rozmezí patří 3 hráči Alisson 29 let, Mane Sadio 29 let a Salah Mohamed 29 let.

Nyní dosazujeme do tabulky četností 1.6:

Věk	17 - 18	19 - 20	21 - 22	23 - 24	25 - 26	27 - 28	29
Četnost	2	1	2	3	4	5	3

Tabulka 1.6: Tabulka četností - doplněná

Kontolu provedme sečtením všech četností. Součet musí být roven počtu všech hráčů.

Trenér by chtěl, aby všechny četnosti byly rovny dvěma. Do věkové kategorie 19 - 20 je nutné jednoho hráče přidat přidat, z věkové kategorie 23 - 24 jednoho hráče vyřadit, z kategorie 25 - 26 vyřadit dva hráče, z kategorie 27 - 28 vyřadit tři hráče a z poslední kategorie vyřadit 1 hráče.

Že trenér bude muset některé hráče vyřadit, je zřejmé již ze zadání úlohy, které říká, že v tabulce je 20 hráčů, ovšem na soustředění mají jet jen 2 z každé věkové kategire a kategorií je 7; trenér by tak vybral pouze $2 \cdot 7 = 14$ hráčů. Zda musí nějaké hráče do seznamu přidat, přímo ze zadání vidět není.

Úloha 5. Brankář Bohemians 1905 J. Bačkovský a FK Teplice J. Čtvrtečka se hádali, který z nich byl v zápasech v září a v říjnu roku 2021 úspěšnější. Počet obdržených gólů v jednotlivých zápasech vidíme v tabulce 1.7. Brankář J. Bačkovský si vzpomněl na výsledky 8 zápasů, brankář J. Čtvrtečka pouze na výsledky 5 zápasů. Určete aritmetický průměr počtu obdržených gólů obou brankářů. Následně rozhodněte, který brankář byl podle dat úspěšnější (nejprve pouze na základě počtu inkasovaných gólů a následně podle aritmetických průměrů inkasovaných gólů). Diskutujte o nebezpečích použití aritmetického průměru při porovnávání statistických jednotek.

Zápas	1.	2.	3.	4.	5.	6.	7.	8.
Bačkovský	0	2	1	5	0	4	0	4
Čtvrtečka	3	0	2	2	4			

Tabulka 1.7: Počet obdržených gólů brankářů Teplic a Bohemians 1905

Řešení. Pro výpočet aritmetického průměru máme vše připraveno, a proto stačí dosadit do vzorce pro výpočet aritmetického průměru

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Nejprve spočítáme aritmetický průměr inkasovaných gólů brankáře Bačkovského, který označíme $\bar{x_B}$.

$$\bar{x_B} = \frac{0 + 2 + 1 + 5 + 0 + 4 + 0 + 4}{8} = 2 \text{ (góly)}.$$

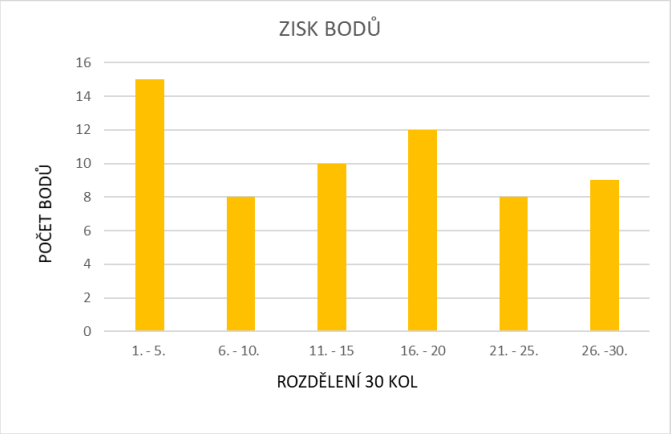
Nyní spočítáme aritmetický průměr inkasovaných gólů brankáře Čtvrtečky, který označíme $\bar{x_C}$, stejně jako brankáře Bačkovského.

$$\bar{x_C} = 2{,}2 \text{ (gólu)}.$$

Podle dat z tabulky 1.7 byl úspěšnější brakář J. Čtvrtečka (11 gólů) než brankář J. Bačkovský (15 gólů).
Naopak podle vypočítaných aritmetických průměrů z počtu inkasovaných gólů byl úspěšnější brankář J. Bačkovský (2 góly) než brankář J. Čtvrtečka (2,2 gólu).
Nemůžeme říci, že by jeden z brankářů byl lepším v obou porovnáních.

Rizikem použití aritmetického průměru je jeho citlivost při počítání s daty, ve kterých se jedna hodnota výrazně odlišuje od průměru ostatních (např. 20 hodnot se pohybuje mezi čísly 5 - 10 a 21. hodnota bude větší než 50).
Aritmetický průměr je ve sportovním odvětví často využívaným nástrojem pro popis počtu gólů či jiných dat. Pro veřejnost je jednoduše interpretovatelný.

Úloha 6. Vedení týmu AC Sparta Praha si přálo, aby došlo ke zvýšení průměrného počtu bodů za zápas oproti minulé sezóně, v níž byl průměrný počet bodů 1,73 bodů na zápas. Viz graf na obrázku 1.5. Určete aritmetický průměr bodů za zápas po odehrání 30 kol týmu AC Sparta Praha ve Fortuna lize za sezónu 2020/2021. Splnilo se přání vedení týmu, pokud ano, o kolik bodů?



Obrázek 1.5: Počet bodů v průběhu 30 kol týmu Sparta Praha

Řešení. Pro výpočet aritmetického průměru využijeme vzorce:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Do vzorce dosadíme a dostáváme:

$$\bar{x} = \frac{62}{30} \doteq 2{,}07 \text{ (bodu)}.$$

Aritmetický průměr bodů za zápas je 2,07 bodu.

Ano, přání vedení týmu se splnilo, protože v sezóně 2020/2021 získal tým v průměru 2,07 bodu za zápas a v sezóně předešlé pouze 1,73 bodu za zápas. Rozdíl v aritmetických průměrech bodů je 0,34 bodu. **Průměrný počet získaných bodů se zvýšil o 0,34 bodu na zápas.**

Úloha 7. V tabulce 1.8 vidíme počet vstřelených gólů v Premier League 10 nejlepších střelců za sezónu 2021/2022. Určete, o kolik gólů více by musel nastřílet hráč s největším počtem gólů, aby se aritmetický průměr zvýšil o 1,4 gólu.

Jméno	Salah M.	Diogo Jota	Mane S.	Sterling R.	Fernandes B.
Počet gólů	19	12	11	10	9
Jméno	Dennis E.	Son Heung-Min	Ronaldo C.	Raphinha	Smith Rowe E.
Počet gólů	9	9	9	9	9

Tabulka 1.8: Vstřelené góly v Premier League

Řešení. Začneme výpočtem aritmetického průměru $\bar{x_1}$ vstřelených gólů na základě tabulky 1.8. Využijeme k tomu vzorce:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n},$$

do něhož dosadíme.

$$\bar{x_1} = \frac{106}{10} = 10{,}6 \text{ (gólu)}.$$

Aritmetický průměr $\bar{x_1}$ je 10,6 gólu za zápas.
Nyní zjistíme, o kolik gólů více musí dát nejlepší střelec, aby se celkový průměr zvýšil o 1,4 gólu.
Konečný průměr $\bar{x_2}$ je roven součtu aritmetického průměru $\bar{x_1}$ (10,6 gólu) a hodnotě, o níž se má $\bar{x_1}$ (1,4 gólu) zvýšit. Dostáváme tedy:

$$\bar{x_2} = 12 \text{ (gólů)}.$$

Vzorec pro výpočet průměru $\bar{x_2}$ je

$$\bar{x_2} = \frac{s_2}{n}, \tag{1.1}$$

kde s_2 je součet všech gólů po zvýšení počtu gólů nejlepšího střelce a n zůstává stejné.
Dále ze vzorce (1.1) vyjádříme s_2:

$$s_2 = \bar{x_2} \cdot n$$

a dosadíme známé hodnoty $\bar{x_2}$ a n. Po dosazení dostáváme:

$$s_2 = 120 \text{ (gólů)}.$$

Také víme, že zvětšením hodnoty s_1 (součet všech gólů z tabulky 1.8) o nějakou neznámou hodnotu x dostaneme hodnotu s_2. Tedy

$$s_2 = s_1 + x, \tag{1.2}$$

Dosazením do vzorce (1.2) zjistíme hledanou hodnotu x. Dostáváme:

$$x = 14 \text{ (gólů)}.$$

Nejlepší hráč by musel nastřílet o 14 gólů více, aby se aritmetický průměr zvýšil o 1,4 gólu za zápas.

Úloha 8. Trenér týmu FC Chelsea Thomas Tuchel se v zítřejším tréninku zaměří na trénování hlaviček a hlavičkových soubojů. Rozhodl se, že nebude trénovat s celým týmem, ale pouze s $\frac{1}{2}$ nejvyšších hráčů. Všechny hráče si zapsal do tabulky 1.9. Vypočítejte medián a aritmetický průměr z výšek hráčů a rozhodněte, zda jedna z vypočítaných hodnot (obě, či žádná) odpovídají hodnotě, která rozděluje soupisku na poloviny (obě poloviny obsahují stejný počet prvků (výšek hráčů) a výšky jsou uspořádané sestupně). Pokud zjistíte, že aritmetický průměr či medián nerozdělují soubor na poloviny, pokuste se vysvětlit, proč tomu tak je.

Pozice	Hráč	Výška
Obrana	Chalobah Trevoh	190 cm
	Chilwell Ben	178 cm
	Christensen Andreas	188 cm
	James Reece	182 cm
	Rüdiger Antonio	190 cm
	Sarr Malang	182 cm
	Silva Thiago	183 cm
Záloha	Barkley Ross	189 cm
	Havertz Kai	189 cm
	Hudson-Odoi Callum	177 cm
	Jorginho	180 cm
	Kovačić Mateo	176 cm
	Kante N'Golo	168 cm
	Loftus-Cheek Rube	191 cm
	Mount Mason	178 cm
	Pulišić Christian	173 cm
	Ziyech Hakim	181 cm
	Ñíguez Saúl	184 cm
Útok	Lukaku Romelu	190 cm
	Werner Timo	180 cm

Tabulka 1.9: Výšky hráčů týmu Chelsea 1

Řešení. Při řešení této úlohy začneme tím, že celou tabulku 1.9 přepíšeme do tabulky 1.10, ve které jsou hráči seřazeni podle velikosti sestupně. Následně hráče v tabulce 1.10 rozdělíme na polovinu.

Hráč	Výška
Loftus-Cheek Rube	191 cm
Chalobah Trevoh	190 cm
Rüdiger Antonio	190 cm
Lukaku Romelu	190 cm
Barkley Ross	189 cm
Havertz Kai	189 cm
Christensen Andreas	188 cm
Ñíguez Saúl	184 cm
Silva Thiago	183 cm
James Reece	182 cm
Sarr Malang	182 cm
Ziyech Hakim	181 cm
Jorginho	180 cm
Werner Timo	180 cm
Chilwell Ben	178 cm
Mount Mason	178 cm
Hudson-Odoi Callum	177 cm
Kovačić Mateo	176 cm
Pulišić Christian	173 cm
Kante N'Golo	168 cm

Tabulka 1.10: Výšky hráčů týmu Chelsea 2

Pro výpočet mediánu výšek hráčů využíváme vzorce:

$$Med(x) = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}),$$

protože počet všech prvků je sudý. Dosazením do vzorce pro výpočet mediánu dostáváme:

$$Med(x) = \frac{1}{2}(x_{10} + x_{11}) = 182 \text{ (cm)}.$$

Mediánem výšek hráčů je 182 cm. Aritmetický průměr výšek hráčů počítáme pomocí vzorce:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Do vzorce dosadíme a dosteneme:

$$\bar{x} = 182{,}45 \text{ (cm)}.$$

Průměrná výška hráčů FC Chelsea je 182,45 cm.

V tabulce 1.10 vidíme dvojitou čáru, která rozděluje uspořádaný soubor výšek hráčů na dvě poloviny, mezi hráči s výškami 182 cm.
Medián je 182 cm a rozděluje soubor přesně na poloviny. To jsme mohli rozhodnout již ze zadání, protože podle definice 11 je medián označován jako prostřední hodnota.

Aritmetický průměr je 182,45 cm, což není hodnota, která by soubor rozdělovala na dvě poloviny. To je v pořádku, protože pro aritmetický průměr obecně neplatí, že rozděluje soubor hodnot na poloviny.

Úloha 9. Před každým zápasem je nutné upravit trávník na hřišti (posekat, posbírat listí či odpadky atp.). Pro tuto práci jsou v garáži dva univerzální stroje. Stroji Amazone profihopper z roku 2020 trvá upravit celý trávník 2 hodiny a 30 minut. Druhý stroj je také Amazone profihhoper, ovšem s rokem výroby 2015. Druhému stroji trvá důkladná příprava trávníku 3 hodiny a 15 minut. Určete, jak dlouho průměrně trvá upravit jedno celé hřiště. Uveďte na příkladu, kdy je dále možné použít harmonický průměr.

Řešení. Protože naším úkolem je vypočítat průměrný čas nutný k nějaké práci, využijeme vzorce pro harmonický průměr:

$$\bar{x_H} = \frac{1}{\frac{1}{n}(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n})}.$$

Čas vyjádříme v minutách a do vzorce dosadíme. Dostáváme:

$$\bar{x_H} = \frac{3900}{23} \doteq 169{,}56 \text{ min} = 2 \text{ h a } 49{,}56 \text{ min}.$$

Průměrný čas potřebný pro posekání celé hrací plochy je 2 hodiny a 49,56 minuty.

Harmonický průměr dále používáme při počítání průměrné rychlosti (pokud jsou zadané rychlosti na úsecích o stejné délce).

Kapitola 2

Charakteristiky variability

Místo slova variabilita můžeme použít slovo rozptýlení, nebo proměnlivost, z čehož můžeme říci, že charakteristiky variability popisují pomocí číselného vyjádření, jak se jednotlivé hodnoty znaku od sebe navzájem liší. Základními ukazateli variability jsou rozptyl, směrodatná a mezikvartilová odchylka a variační koeficient.

2.1 Základní statistické pojmy ve středoškolské matematice

Následující definice a poznámky jsou věnovány termínům popisujícím charakteristiky variability ve středoškolské matematice.

Definice 13 (Rozptyl Hindsl a kol., 2018, str. 38)**.** *Nechť je dán statistický soubor o rozsahu n pozorování $x_1, x_2, ..., x_n$. Pak je rozptyl s_x^2 definován vzorcem:*

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Poznámka (Hindsl a kol., 2018, str. 38). Předchozí vzorec není příliš vhodný pro výpočty, při nichž nemáme k dispozici počítač. Jednoduchou úpravou ho převedeme do formy vhodnější pro praktické výpočty. Čitatel ve vzorci vyjádříme takto:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n(\bar{x})^2 =$$
$$= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \bar{x}\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_i =$$
$$= \sum_{i=1}^{n}x_i^2 - n(\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\frac{1}{n}\sum_{i=1}^{n}x_i)^2.$$

Poznámka (Výpočtový tvar rozptylu Hindsl a kol., 2018, str. 38). Můžeme proto používat tzv. výpočtový tvar rozptylu.

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 = (\bar{x^2}) - (\bar{x})^2.$$

Definice 14 (Směrodatná odchylka Hindsl a kol., 2018, str. 39)**.** *Směrodatná odchylka je definována jako kladná druhá odmocnina z rozptylu. Značíme ji s_x. Pro její výpočet používáme vzorec:*

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Definice 15 (Variační koeficient Calda a Dupač, 1993, str. 144)**.** *Nechť znak x nabývá jen nezáporných hodnot. Variační koeficient definujeme jako podíl směrodatné odchylky a aritmetického průměru. Značíme ho v_x. Pro výpočet používáme vzorec:*

$$v_x = \frac{s_x}{\bar{x}} \cdot 100\%.$$

Definice 16 (První a třetí kvartil znaku x Calda a Dupač, 1993, str. 145)**.** *První kvartil Q_1 a třetí kvartil Q_3 jsou definovány takto:*

$$\begin{gathered} Q_1 \text{ je medián z hodnot } x_{(1)} \leq x_{(2)} \leq ... \leq Med(x) \\ Q_3 \text{ je medián z hodnot } x_{(n)} \geq x_{(n-1)} \geq ... \geq Med(x), \\ \text{kde } Med(x) \text{ je medián.} \end{gathered}$$

Definice 17 (Mezikvartilová odchylka Calda a Dupač, 1993, str. 144)**.** *Mezikvartilovou odchylku znaku x, značíme ji $Q(x)$, definujeme jako:*

$$Q(x) = \frac{1}{2}(Q_3 - Q_1).$$

2.2 Základní statistické pojmy ve vysokoškolské matematice

V následujících definicích jsou definovány charakteristiky variability ve vysokoškolské matematice. Většinu definic i poznámek budeme využívat v této i následující kapitole.

První dvě definice σ-algebry a Kolmogorovy definice pravděpodobnosti jsou potřebné pro zbylou část bakalářské práce.

Nechť je dána libovolná množina Ω.

Definice 18 (σ-algebra Kulich, 2018, str. 8)**.** *Systém $\mathcal{A}$ podmnožin množiny Ω nazveme σ-algebrou, pokud platí:*

1. $\emptyset \in \mathcal{A}$;
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$, *kde A^c je doplňek množiny A v množině Ω;*
3. $A_1, A_2, ... \in \mathcal{A} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Definice 19 (Kolmogorova definice pravděpodobnosti Dupač a Hušková, 1999, str. 8)**.** *Nechť Ω je množina a $\mathcal{A}$ σ-algebra jejích podmnožin. Pak funkci $P : \mathcal{A} \to \langle 0; 1\rangle$ nazveme pravděpodobností, právě když splňuje následující podmínky:*

1. $P(A) \geq 0, A \in \mathcal{A}, P(\Omega) = 1$;
2. $A_1, A_2, \ldots, A_n \in \mathcal{A}$ *a* $A_i \cap A_j = \emptyset \, \forall i \neq j \Rightarrow P(\bigcup_1^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Nyní interpretujeme množinu Ω z předchozích definic a prvky σ-algebry. Spojením Ω, σ-algebry a pravděpodobnosti P definujeme pravděpodobnostní prostor.

Definice 20 (Prostor elementárních jevů, náhodné jevy, pravděpodobnostní prostor Kulich, 2018, str. 8)**.** *Množinu Ω nazýváme prostor elementárních jevů, její prvky $\omega \in \Omega$ nazýváme elementární jevy. Prvky σ-algebry $\mathcal{A}$ nazýváme náhodné jevy. Trojici $(\Omega, \mathcal{A}, P)$ nazýváme pravděpodobnostní prostor.*

Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$. Na tomto prostoru definujeme náhodnou veličinu jako základní objekt, který využíváme ve všech dalších statistických termínech vysokoškolské matematiky. Dále zavádíme rozdělení náhodné veličiny, které popisuje vlastnosti náhodné veličiny.

Poznámka. Nechť jsou dány σ-algebry $\mathcal{A}$ na množině Ω a $\mathcal{B}$ na množině $\mathcal{X}$. Zobrazení $X : \Omega \to \mathcal{X}$ se nazývá měřitelné vzhledem k σ-algebrám $\mathcal{A}$ a $\mathcal{B}$, právě když $\forall B \in \mathcal{B}$ platí $\{\omega \in \Omega : X(\omega) \in \mathcal{B}\} \in \mathcal{A}$.

Definice 21 (Náhodná veličina, výběrový prostor Kulich, 2018, str. 9)**.** *Měřitelné zobrazení $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$, kde $\mathcal{X}$ je nějaká množina a $\mathcal{B}$ nějaká σ-algebra na $\mathcal{X}$, nazýváme náhodnou veličinou. Množinu $\mathcal{X}$ nazýváme výběrový prostor.*

Definice 22 (Rozdělení náhodné veličiny Kulich, 2018, str. 9)**.** *Nechť X je náhodná veličina, Ω a $\mathcal{X}$ jsou prostory elementárních jevů, $\mathcal{A}$ je σ-algebra na množině Ω a $\mathcal{B}$ na množině $\mathcal{X}$. Pak rozdělením náhodné veličiny $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ rozumíme indukovanou (vytvářenou)*[1] *pravděpodobnostní míru P_X na $(\mathcal{X}, \mathcal{B})$ definovanou vztahem:*

$$P_X(B) = P[X \in B], B \in \mathcal{B},$$

kde $[X \in B] = \{\omega \in \Omega : X(\omega) \in B\}$.

[1] Pravděpodobnostní míru vytváří náhodná veličina X.

Dále představujeme základní vlastnost náhodných veličin, nezávislost.

Definice 23 (Nezávislost náhodných veličin Kulich, 2018, str. 19)**.** *Nechť jsou dány náhodné veličiny $X_1, \ldots, X_n$. Pak náhodné veličiny $X_1, \ldots, X_n$ nazveme nezávislé, právě když pro všechny $B_1, \ldots, B_n \in \mathcal{B}$ platí*

$$\mathsf{P}[X_1 \in B_1, \ldots, X_n \in B_n] = \mathsf{P}[X_1 \in B_1] \cdot \ \ldots \ \cdot \mathsf{P}[X_n \in B_n],$$

kde $[X_1 \in B_1, \ldots, X_n \in B_n] = [X_1 \in B_1] \cap \cdots \cap [X_n \in B_n]$.

Poznámka (Spojitá a diskrétní náhodná veličina Kulich, 2018, str. 11). Když je P_X absolutně spojitá vzhledem k Lebesgueově míře λ, pak X je spojitá náhodná veličina [náhodná veličina se spojitým rozdělením].
Když je P_X absolutně spojitá vzhledem k čítací míře [2] μ_S (S je nejvýše spočetná množina v $\mathbb{R}$), pak X je diskrétní náhodná veličina [náhodná veličina s diskrétním rozdělením].

Poznámka (σ-konečnost, absolutní spojitost Kulich, 2018, str. 10). Míra μ na $(\mathcal{X},\mathcal{B})$ je σ-konečná, právě když existují množiny $B_1, B_2, ... \in \mathcal{B}$ takové, že
$\forall i \in \mathbb{N}\ \mu(B_i) < \infty$ a $\bigcup_{i=1}^{\infty} B_i = \mathcal{X}$.
Míra P_X je absolutně spojitá vzhledem k míře μ na $(\mathcal{X}, \mathcal{B})$, právě když
$\forall B \in \mathcal{B}\ \mu(B) = 0 \Rightarrow \mathsf{P}_X(B) = 0$.

Další definice jsou přímo využívány v úlohách v podkapitole 2.3.3. Hustota i distribuční funkce jednoznačně určují rozdělení náhodné veličiny.

Definice 24 (Hustota Kulich, 2018, str. 10)**.** *Nechť $X : (\Omega, \mathcal{A}) \to (\mathcal{X},\mathcal{B})$ je náhodná veličina, nechť μ je σ-konečná míra na $\mathcal{X}$ a nechť P_X je absolutně spojitá vzhledem k μ. Pak hustota náhodné veličiny X je reálná měřitelná nezáporná funkce $f_X(x)$ určená jednoznačně μ-skoro všude (až na množinu míry 0) a platí pro ni:*

$$\int_{\mathcal{X}} h(x)\,\mathrm{d}\,\mathsf{P}_X(x) = \int_{\mathcal{X}} h(x) f_X(x)\,\mathrm{d}\mu(x)^3,$$

kde h je libovolná měřitelná funkce $h : (\mathcal{X}, \mathcal{B}) \to (\mathbb{R}, \mathcal{B}_0)$.

Definice 25 (Distribuční funkce Kulich, 2018, str. 11)**.** *Funkci $F_X : \mathbb{R} \to \mathbb{R}$ definovanou vztahem $F_X(x) = \mathsf{P}[X \leq x]$ nazýváme distribuční funkcí náhodné veličiny X.*

V následujících definicích věnujeme pozornost některým pojmům, které jsme v podkapitole 2.1 zavedly středoškolským způsobem. Začneme vysokoškolským ekvivalentem aritmetického průměru, střední hodnotou.

Definice 26 (Střední hodnota Kulich, 2018, str. 12)**.** *Nechť X je náhodná veličina, Ω je prostor elementárních jevů ω, P je pravděpodobnostní míra na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$. Pak střední hodnotou EX (reálné) náhodné veličiny X rozumíme reálné číslo EX definované takto:*

$$EX = \int_{\Omega} X(\omega)\,\mathrm{d}P(\omega),$$

[2] Čítací (aritmetická míra) je měřitelné zobrazení, přiřazující nejvýše spočetné podmnožině $\mathbb{R}$ počet jejích prvků.

[3] Jedná se o Lebesgueův integrál.

pokud integrál na pravé straně existuje.
P je pravděpodobnostní míra na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$.

Poznámka (Střední hodnota pro diskrétní a spojitou náhodnou veličinu).

$$EX = \sum_k k\,\mathsf{P}(X = k), \tag{2.1}$$

je-li náhodná veličina X diskrétní s hodnotami $k = 0, 1, 2, \ldots$.

$$EX = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x, \tag{2.2}$$

je-li náhodná veličina spojitá a $f(x)$ je hustota rozdělení náhodné veličiny X.

Poznámka (Střední hodnota reálné měřitelné funkce). Nechť h je reálná měřitelná funkce. Pak střední hodnota reálné měřitelné funkce je:

$$Eh(X) = \int_\Omega h(x)\, f_X(x)\mathrm{d}x. \tag{2.3}$$

Nyní definujeme k-tý centrální moment, který využijeme při zavedení rozptylu jako druhého centrálního momentu.

Definice 27 (k-tý centrální moment Kulich, 2018, str. 13)**.** *Nechť X je náhodná veličina. Pak k-tý centrální moment μ_k náhodné veličiny X je definována takto:*

$$\mu_k = E(X - EX)^k.$$

Definice 28 (Rozptyl Kulich, 2018, str. 13)**.** *Rozptyl $var(X)$ náhodné veličiny X je její druhý centrální moment, tj. $var(X) = E(X - EX)^2$.*

Poznámka. V literatuře je rozptyl značen také s_x^2.

Poznámka (Upravený tvar rozptylu Kulich, 2018, tvrzení 2.4, str. 13). Vzorec pro výpočet rozptylu můžeme upravit do tvaru:

$$var(X) = E(X)^2 - (EX)^2.$$

Definice 29 (Směrodatná odchylka Kulich, 2018, str. 13)**.** *Směrodatná odchylka s_x náhodné veličiny X je rovna odmocnině z jejího rozptylu:*

$$s_x = \sqrt{varX}.$$

Nyní definujeme dva speciální typy rozdělení (jedno diskrétní a jedno spojité), které využijeme v úlohách podkapitoly 2.3.3.

Definice 30 (Poissonovo rozdělení DeGroot, 1975, str. 207)**.** *Nechť X je náhodná veličina s diskrétním rozdělením a X nabývá pouze nezáporných celočíselných hodnot. Náhodná veličina X má Poissonovo rozdělení se střední hodnotou $EX = \lambda, \lambda > 0$, pokud má X pravděpodobnostní funkci:*

$$f(x) = \mathsf{P}(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{pro } x = 0, 1, 2, \ldots, \\ 0 & \text{jindy.} \end{cases}$$

Definice 31 (Exponenciální rozdělení DeGroot, 1975, str. 238)**.** *Náhodná veličina X má exponenciální rozdělení s parametrem $\beta > 0$, pokud X má spojité rozdělení s funkcí hustoty:*

$$f(x) = \begin{cases} \beta\, e^{-\beta x} & pro\ x > 0, \\ 0 & x \leq 0. \end{cases}$$

Další typ spojitého rozdělení, χ^2 rozdělení, potřebujeme pro teorii i ukázku testování hypotéz v kapitole 4.

Definice 32 (χ^2 rozdělení DeGroot, 1975, str. 323)**.** *Náhodná veličina X má χ^2 rozdělení o n stupních volnosti*[4]*, pokud X má spojité rozdělení s funkcí hustoty:*

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1}\, e^{-x/2}\,,\ pro\ x > 0,$$

kde

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\ \mathrm{d}x.$$

2.3 Řešené úlohy

V této části uvádíme řešené úlohy, které jsou zaměřené na charakteristiky variability (rozptyl, odchylky a variační koeficient). První dvě úlohy jsou záměrně jednodušší, zaměřené na procvičení dosazení do správného vzorce. Následuje pět úloh složitějších, ve kterých je nutné nějaké odůvodnění či reprezentace výsledku. Na závěr jsou zařazeny dvě úlohy, jejichž řešení vyžaduje použití vysokoškolské matematiky, na kterých ukazujeme, jak se výpočetní metody a postupy odlišují od úrovně středoškolské. Úlohy jsou zaměřené na praktické použití vzorců z předchozí teoretické podkapitoly. V této podkapitole se zaměřujeme na hokejovou tématikou.
Úlohy i jejich řešení jsou, pokud není řečeno jinak, autorská.

2.3.1 Řešené úlohy - jednodušší

Úloha 10. Určete rozptyl počtu vstřelených gólů hokejového týmu HC Škoda Plzeň v prvních 11 zápasech sezóny 2021/2022. Počty vstřelených gólů: 0, 3, 1, 2, 3, 4, 5, 5, 4, 3, 8.

Řešení. Při řešení této úlohy si napíšeme vzorec výpočtového tvaru rozptylu:

$$s_x^2 = (\bar{x^2}) - (\bar{x})^2.$$

K tomu vypočítáme $(\bar{x^2})$ a $(\bar{x})^2$. Pro výpočet $(\bar{x})^2$ dosadíme do vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

[4]Název parametru.

a vypočítáme druhou mocninu z výsledku. Tím získáme

$$(\bar{x})^2 = \frac{1\ 444}{121}.$$

Při výpočtu $(\bar{x^2})$ počítáme aritmetický průměr druhých mocnin počtů vstřelených gólů, tedy

$$(\bar{x^2}) = \frac{178}{11}.$$

Nyní známe všechny hodnoty nutné k dosazení do vzorce pro výpočet rozptylu.

$$s_x^2 = \frac{178}{11} - \frac{1444}{121} = \frac{514}{121} \doteq 4{,}25.$$

Rozptyl počtu vstřelených gólů je 4,25.

Úloha 11. V tabulce 2.1 vidíte 20 nejlepších českých střelců a počty gólů, které vstřelili v historii NHL. Určete rozptyl, směrodatnou odchylku a variační koeficient počtu gólů.

Jméno	Počet gólů
J. Jágr	766
P. Eliáš	408
M. Hejduk	375
B. Holík	326
P. Sýkora	323
P. Klíma	313
P. Nedvěd	310
R. Vrbata	284
R. Lang	261
M. Straka	257
V. Prospal	255
R. Reichel	255
M. Havlát	242
M. Ručínský	241
T. Plekanec	233
R. Dvořák	227
J. Voráček*	216
D. Krejčí	215
M. Pivoňka	181
R. Hamrlík	155

Tabulka 2.1: Počet vstřelených gólů nejlepších českých střelců v historii NHL
*stále aktivní

Řešení. Nejprve vypočítáme rozptyl počtu vstřelených gólů podle vzorce:

$$s_x^2 = (\bar{x^2}) - (\bar{x})^2.$$

K tomu potřebujeme hodnotu $(\bar{x})^2$, kterou zjistíme dosazením do vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a vypočítáním druhé mocniny z výsledku $\bar{x}$. Získáme:

$$\bar{x} = 292{,}15 \text{ (gólu)},$$

$$(\bar{x})^2 = \frac{34\ 140\ 649}{400}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin počtů vstřelených gólů, tedy

$$(\bar{x^2}) = 100\ 776{,}25.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_x^2 = 100\ 776{,}25 - \frac{34\ 140\ 649}{400} = \frac{6\ 169\ 851}{400} \doteq 15\ 424{,}63.$$

Rozptyl je 15 454,63. Směrodatná odchylka s_x je rovna druhé odmocnině z rozptylu s_x^2.

$$s_x \doteq 124{,}2 \text{ (gólu)}.$$

Směrodatná odchylka je přibližně 124,2 gólu. Vzorec pro výpočet variačního koeficientu je:

$$v_x = \frac{s_x}{\bar{x}} \cdot 100\ \%.$$

Do vzorce dosadíme a dostaneme:

$$v_x \doteq 42{,}51\ \%$$

Variační koeficient je přibližně 42,51 %.

2.3.2 Řešené úlohy - obtížnější

Úloha 12. Luboš Jenáček, trenér hokejového týmu Berani Zlín, si všiml, že hráči v zápasech často prohrávají osobní souboje, přicházejí o puky a dostávají více gólů, než tomu bylo dříve. Napadlo ho, že by to mohlo být jejich hmostností. Potřeboval by hmotnostně vyvážený tým. Výčet hmostností některých hráčů: J. Dluhoš 78 kg, T. Žižka 96 kg, R. Černý 72 kg, O. Němec 90 kg, J. Karafiát 79 kg, J. Svoboda 70 kg, B. Köhler 105 kg, A. Zbořil 86 kg, M. Sebera 71 kg, D. Luža 73 kg a O. Flynn 71 kg. Určete rozptyl a směrodatnou odchylku a na jejich základě rozhodněte a odůvodněte, zda se jedná o hmotnostně vyvážený tým či nikoliv.

Řešení. Rozptyl hmotností hráčů vypočítáme pomocí vzorce:

$$s_x^2 = (\bar{x^2}) - (\bar{x})^2.$$

Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. Hodnotu $(\bar{x})^2$ zjistíme dosazením do vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a umocněním na druhou výsledku $\bar{x}$. Získáme:

$$\bar{x} = 81 \text{ (kg)},$$

$$(\bar{x})^2 = 6\ 561.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých hmotností, tedy

$$(\bar{x^2}) = 6\ 687.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_x^2 = 6\ 687 - 6\ 561 = 126.$$

Rozptyl je roven 126.

Směrodatnou odchylku hmotností hráčů zjistíme jako druhou odmocninu z rozptylu hmotností hráčů. Tedy

$$s_x \doteq 11{,}23 \text{ (kg)}.$$

Směrodatná odchylka je rovna přibližně hodnotě 11,23 kg. Směrodatná odchylka ukázala, že hmotnost se od průměrné hodnoty výrazně odlišuje. Hmotnostní souboje tento rozdíl může ovlivňovat. **Tým hmotnostně vyvážený není.**

Úloha 13. Při oslavě narozenin jednoho z českých hokejových hráčů se u jednoho stolu sešlo 10 hráčů NHL. Povídali si a vzájemně se předháněli, kolik v NHL za svou kariéru získali kanadských bodů (součet všech bodů za góly a asistence). Počet kanadských bodů jednotlivých hráčů byl: 178, 249, 183, 309, 188, 250, 140, 317, 73 a 82. Po chvíli se k nim připojil i Jaromír Jágr, který za svou kariéru v NHL získal 1 921 kanadských bodů. Určete medián a mezikvartilovou odchylku počtu kanadských bodů. Odůvodněte, proč v této úloze určujeme mezikvartilovou odchylku a ne směrodatnou odchylku.

Řešení. Začneme výpočtem mediánu bodů, který známe z kapitoly 1. Medián $Med(x)$ vypočítáme pomocí vzorce:

$$Med(x) = x_{(\frac{n+1}{2})},$$

protože počet všech prvků je lichý.

$$Med(x) = 188 \text{ (bodů)}.$$

Medián je 188 bodů. Dále počítáme mezikvartilovou odchylku pomocí vzorce:

$$Q(x) = \frac{1}{2}\,(Q(3) - Q(1)).$$

$Q(1)$ vypočítáme jako medián z prvků mezi první hodnotou a mediánem $Med(x)$ a $Q(3)$ jako medián hodnot mezi mediánem $Med(x)$ a poslední hodnotou.
Pro výpočet obou kvartilů využijeme vzorce $Med(x) = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$, protože počet prvků mezi první (poslední) hodnotou a $Med(x)$ (včetně obou) je sudý. Dostáváme:

$$Q(1) = 159 \text{ (bodu)},$$
$$Q(3) = 279{,}5 \text{ (bodu)}.$$

Na závěr dosadíme do vzorce pro $Q(x)$ a získáme

$$Q(x) = 60{,}25 \text{ (bodu)}.$$

Mezikvartilová odchylka je 60,25 bodu.
Důvodem, proč využíváme mezikvartilovou odchylku a ne směrodatnou odchylku, je ten, že směrodatná odchylka počítá s aritmetickým průměrem, do kterého se významě promítne i jediná velká hodnota (velká oproti ostatním v souboru). Proto v případě, že máme soubor hodnot s jednou výrazně odlišnou (velkou, či malou) hodnotou, volíme mezikvartilovou odchylku, která k výpočtu využívá medián, do jehož výpočtu se poslední velká hodnota nepromítne. Tato odchylka více odpovídá skutečnosti.

Úloha 14. Představte si, že jste hokejovým trenérem a rozhodujete se mezi dvěma brankáři, kterého si zvolíte do svého týmu. Rozhodněte se na základě aritmetického průměru a směrodatné odchylky inkasovaných gólů v Tipsport Extralize v sezóně 2021/2022 mezi daty 28.1.2022 - 27.2.2022 (oba odehráli stejné množství zápasů). Jedním z nich je Petr Kváča, který inkasoval v jednotlivých zápasech toto množství gólů: 5, 0, 1, 1, 2, 3, 1, 5, 3, 1, 2, 6. Druhým je Libor Kašík, který inkasoval následující počty gólů: 4, 4, 2, 2, 5, 4, 3, 4, 1, 3, 2, 3. Své rozhodnutí zkuste odůvodnit na základě vypočítaných hodnot.

Řešení. Řešení této úlohy začneme výpočtem jednotlivých rozptylů počtu gólů pomocí vzorce:

$$s_x^2 = (\bar{x^2}) - (\bar{x})^2.$$

Potřebujeme k tomu hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme dosazením do vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a následným vypočítáním druhé mocniny z jeho výsledku. Dostáváme:

$$\bar{x_P} = 2{,}5 \text{ (gólu)},$$
$$(\bar{x_P})^2 = 6{,}25,$$
$$\bar{x_L} = \frac{37}{12} \text{ (gólu)},$$
$$(\bar{x_L})^2 = \frac{1\ 369}{144}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých gólů, tedy

$$(\bar{x_P^2}) = \frac{29}{3},$$

$$(\bar{x_L^2}) = \frac{43}{4}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_P}^2 = \frac{29}{3} - \frac{25}{4} = \frac{41}{12} \doteq 3{,}42,$$

$$s_{x_L}^2 = \frac{43}{4} - \frac{1369}{144} \doteq 1{,}24.$$

Na závěr vypočítáme směrodatné odchylky počtu gólů jako druhé odmocniny z rozptylů $s_{x_P}^2$ a $s_{x_L}^2$:

$$s_{x_P} \doteq 1{,}85 \text{ gólu},$$

$$s_{x_L} \doteq 1{,}11 \text{ gólu}.$$

Shrnutí:
Při porovnání obou brankářů se zaměříme na aritmetické průměry a směrodatné odchylky.
Brankář Kváča má průměrný počet obdržených gólů 2,5 gólu za zápas a odchylku gólů přibližbě 1,85 gólu. Brankář Kašík má průměrný počet inkasovaných gólů přibližně 3,08 gólu a přibližnou směrodatnou odchylku 1,11 gólu.
Vidíme, že sice brankář Kváča dostává v průměru méně gólů, ale průměrná odchylka od průměru je vyšší než u brankáře Kašíka.
Každý může upřednostnit na základě těchto výpočtů a svých preferencí jiného brankáře. V tomto řešení zvolíme brankáře Kváču, protože sice odchylka od průměru je větší, ale průměrný počet gólů je menší, a to pro nás má větší význam. (Pokud bychom považovali za důležitější, že brankář dostává více gólů, ale nevychyluje se od průměru, tedy nestává se, že v jednom zápase inkasuje 2 góly a v dalším 7 gólů, zvolíme brankáře Kašíka.)

Úloha 15. Trenér týmu HC Motor České Budějovice dokončil 26.4.2021 přestup hokejového útočníka Milana Gulaše. Druhou potenciální posilou byl Oksanen Ahti. Rozhodoval se na základě počtu gólů v posledních 11 zápasech sezóny 2020/2021 Tipsport Extraligy a Finské národní ligy. Milan Gulaš byl úspěšný a v rozhodujících zápasech vstřelil 0, 0, 0, 0, 0, 2, 1, 0, 0, 2, 2, 2 góly. Útočník Oksanen Ahti za stejný počet zápasů vstřelil 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 2, 1 góly. Vypočtěte variační koeficient a na základě dostupných dat vycházejících ze zmíněného počtu gólů za 11 zápasů rozhodněte, zda byla volba Milana Gulaše správná. Svou odpověď odůvodněte. Vysvětlete, co nám o datech říká variační koeficient.

Řešení. Variační koeficient v_x počtu gólů vypočítáme pomocí vzorce:

$$v_x = \frac{s_x}{\bar{x}} \cdot 100\ \%.$$

K tomu potřebujeme směrodatnou odchylku s_x, kterou vypočítáme jako druhou odmocninu z rozptylu gólů s_x^2.
Vzorec pro výpočet rozptylu je:

$$s_x^2 = (\bar{x^2}) - (\bar{x})^2.$$

Nyní zjistíme $(\bar{x^2})$ a $(\bar{x})^2$ obou hráčů. $(\bar{x})^2$ vypočítáme dosazením do vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a následným vypočítáním druhé mocniny z jeho výsledku. Dostáváme:

$$\bar{x_G} = \frac{3}{4} \text{ (gólu)},$$

$$(\bar{x_G})^2 = \frac{9}{16},$$

$$\bar{x_A} = \frac{3}{4} \text{ (gólu)},$$

$$(\bar{x_A})^2 = \frac{9}{16}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých počtů gólů, tedy

$$(\bar{x_G^2}) = \frac{17}{12},$$

$$(\bar{x_A^2}) = \frac{11}{12}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s^2_{x_G} = \frac{17}{12} - \frac{9}{16} = \frac{41}{48},$$

$$s^2_{x_A} = \frac{11}{12} - \frac{9}{16} = \frac{17}{48}.$$

Na závěr vypočítáme směrodatné odchylky počtu gólů jako druhé odmocniny z rozptylů $s^2_{x_G}$ a $s^2_{x_A}$:

$$s_{x_G} = \frac{\sqrt{123}}{12} \text{ (gólu)},$$

$$s_{x_A} = \frac{\sqrt{51}}{12} \text{ (gólu)}.$$

Dále dosadíme do vzorce pro výpočet v_x a dostáváme:

$$v_{x_G} \doteq 123{,}23\ \%,$$

$$v_{x_A} \doteq 79{,}34\ \%.$$

Variační koeficient gólů Milana Gulaše je 123,23 % **a variační koeficient gólů Oksanena Ahti je** 79,34 %**.**

Variační koeficient využíváme, pokud je naším cílem pomocí procent vyjádřit relativní velikot rozptylu hodnot od průměru (jakou část (v procentech) průměru vyjadřuje směrodatná odchylka).

Oba útočníci nastříleli stejné množství gólů (9). Celkový počet vstřelených gólů hráče neodliší.

Můžeme porovnat hráče podle toho, zda dávají góly v zápasech pravidelně (např. v každém druhém zápase 1), nebo zda jsou úspěšní jen vyjímečně (např. dvakrát za 10 zápasů dají 4 góly).

Pro porovnání hráčů můžeme z našich výpočtů využít aritmetických průměrů a směrodatných odchylek, které udávají, o kolik se liší množství gólů od průměru. Směrodatná odchylka Milana Gulaše je 0,92 gólu a Oksanena Ahti je 0,59 gólu. O úspěšnosti Oksanena Ahti můžeme říci, že góly dává vcelku pravidelně, narozdíl od Milana Gulaše.
Tento ukazatel říká, že volba Milana Gulaše může být chybná. V řešení neuvažujeme, v jakých zápasech branky padaly (Bylo to v důležitých zápasech?). I to by ovšem při rozhodování trenéra mělo mít výrazný vliv.

Úloha 16. Trenéři amerických hokejových týmů Rod Brind'Amour (Carolina Hurricanes) a Bruce Cassidy (Boston Bruins) porovnávali své týmy. Trenér Bruce Cassidy má pocit, že hráči jeho týmu v zápasech bývají častěji vyloučení (počty vyloučení: 4, 4, 6, 3, 3, 3, 2) než hráči týmu Carolina Hurricanes (počty vyloučení: 4, 3, 4, 2, 3, 4, 4), a proto je v tabulce tým hůře umístěný. Na základě počtu vyloučení obou týmů v 7 zápasech od 27.02.2022 do 12.03.2022 vypočítejte aritmetické průměry a směrodatné odchylky počtu vyloučení. Rozhodněte, zda měl trenér Bruce Cassidy správný pocit, že hráči jeho týmu bývají v zápasech průměrně vícekrát vyloučeni než hráči týmu Roda Brind'Amoura.

Řešení. Začneme výpočtem jednotlivých rozptylů počtu vyloučení, které potřebujeme k výpočtu směrodatných odchylek. Rozptyl počítáme pomocí vzorce:

$$s_x^2 = (\bar{x^2}) - (\bar{x})^2.$$

K tomu potřebujem hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme dosazením do vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a následným vypočítáním druhé mocniny z jeho výsledku. Dostáváme:

$$\bar{x_{BB}} = \frac{25}{7}(\text{vyloučení}),$$

$$(\bar{x_{BB}})^2 = \frac{625}{49},$$

$$\bar{x_{CH}} = \frac{24}{7}\ (\text{vyloučení}),$$

$$(\bar{x_{CH}})^2 = \frac{576}{49}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin vyloučení, tedy

$$(\bar{x_{BB}^2}) = \frac{99}{7},$$

$$(\bar{x_{CH}^2}) = \frac{86}{7}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_{BB}}^2 = \frac{99}{7} - \frac{625}{49} = \frac{68}{49},$$

$$s^2_{x_{CH}} = \frac{86}{7} - \frac{576}{49} = \frac{26}{49}.$$

Na závěr vypočítáme směrodatné odchylky počtu gólů jako druhé odmocniny z rozptylů $s^2_{x_C}$ a $s^2_{x_B}$:

$$s_{x_{BB}} \doteq 1{,}17 \text{ (vyloučení)},$$
$$s_{x_{CH}} \doteq 0{,}73 \text{ (vyloučení)}.$$

Směrodatná odchylka počtu vyloučení hráčů týmu Carolina Hurricanes je 1,17 vyloučení a směrodatná odchylka počtu vyloučení hráčů týmu Boston Bruins je 0,73 vyloučení.
Porovnání v tabulce 2.2 (s přibližnými hodnotami):

	$\bar{x} - s_x$	$\bar{x}$	$\bar{x} + s_x$
Boston Bruins	2,39	3,57	4,75
Carolina Hurricanes	2,7	3,43	4,16

Tabulka 2.2: Tabulka pro porovnání obou týmů

Jak vidíme v tabulce 2.2, **trenér týmu Boston Bruins měl správný pocit**, protože hráči jeho týmu bývají v průměru méně vyloučeni než hráči týmu Carolina Hurricanes. Avšak směrodatná odchylka je větší u týmu Carolina Hurricanes. To říká, že se počet vyloučení tohoto týmu více vychyluje od průměru, než u týmu Boston Bruins.

2.3.3 Řešené úlohy - vysokoškolského charakteru

Úloha 17. Nechť náhodná veličina X udává počet všech faulů během jednoho hokejového zápasu (obou týmů dohromady). Pro to můžeme předpokládat, že náhodná veličina X má Poissonovo rozdělení s parametrem $\lambda = 6$. Určete střední hodnotu, rozptyl a směrodatnou odchylku.

Řešení. Z definice 30 víme, že střední hodnota je $\lambda = 6$ (faulů).
Střední hodnota je 6 faulů.
Pokračujeme výpočtem rozptylu. Vzorec pro jeho výpočet je

$$var(X) = E(X)^2 - \ (EX)^2.$$

Nyní vypočítáme EX^2. S využitím vzorců (2.3) ($h(X) = X^2$) a (2.1) píšeme

$$EX^2 = \sum_{k=0}^{\infty} k^2 \,\mathsf{P}(X = k) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}\,\mathsf{e}^{-\lambda} = \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!}\,\mathsf{e}^{-\lambda}\,.$$

V dalším kroku úpravy navrátíme meze sumy zpět na k od 0, přeznačíme a sumu rozdělíme na součet dvou sum.

$$\lambda \sum_{k=0}^{\infty} (k+1) \frac{\lambda^k}{k!}\,\mathsf{e}^{-\lambda} = \lambda \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!}\,\mathsf{e}^{-\lambda} + \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\,\mathsf{e}^{-\lambda}\,.$$

První suma je ze vzorce (2.1) rovna λ a druhá suma je součet všech pravděpodobností, tedy z definice 19 1. Dostáváme tak

$$EX^2 = \lambda^2 + \lambda.$$

Zjištěné hodnoty dosazujeme do vzorce pro rozptyl a získáváme

$$var(X) = (36 + 6) - 36 = 6.$$

Rozptyl je 6.
Pro dopočítání směrodatné odchylky využijeme vztahu z definice 14:

$$s_x = \sqrt{6} \text{ (faulů)}.$$

Směrodatná odchylka je $\sqrt{6}$ faulů.

Úloha 18. Nechť náhodná veličina X udává počet minut mezi vstřelenými góly týmu. Pro to můžeme předpokládat, že náhodná veličina X má exponenciální rozdělení dané parametrem $\beta = \frac{1}{10}$. Určete střední hodnotu, rozptyl a směrodatnou odchylku.

Řešení. Začneme výpočtem střední hodnoty EX. Využijeme vzorce (2.2) a definice 31. Jejich kombinací dostáváme

$$EX = \int_0^\infty x \frac{1}{10} \mathrm{e}^{-\frac{x}{10}} \, \mathrm{d}x.$$

Využitím integrační metody per partes upravíme integrál do tvaru:

$$EX = \frac{1}{10} \left(-10 \left[x \, \mathrm{e}^{-\frac{x}{10}} \right]_0^\infty + 10 \int_0^\infty \mathrm{e}^{-\frac{x}{10}} \, \mathrm{d}x \right) = -10 \left[\mathrm{e}^{-\frac{x}{10}} \right]_0^\infty = 10 \text{ (minut)}. \tag{2.4}$$

Střední hodnota je 10 minut.
Při výpočtu rozptylu využíváme vzorce $var(X) = E(X)^2 - (EX)^2$. Začneme výpočtem EX^2. S využitím vzorce (2.3) ($h(X) = X^2$) a (2.2) získáváme

$$EX^2 = \int_0^\infty x^2 \frac{1}{10} \mathrm{e}^{-\frac{x}{10}} \, \mathrm{d}x.$$

Opět využijeme integrační techniky per partes a dostaneme upravený integrál

$$EX^2 = \frac{1}{10} \left(-10 \left[x^2 \, \mathrm{e}^{-\frac{x}{10}} \right]_0^\infty + 20 \int_0^\infty x \, \mathrm{e}^{-\frac{x}{10}} \, \mathrm{d}x \right) = 20 \int_0^\infty x \frac{1}{10} \mathrm{e}^{-\frac{x}{10}} \, \mathrm{d}x.$$

Hodnotu integrálu již známe z integrálu ve výpočtu střední hodnoty 2.4.

$$EX^2 = 20 \cdot 10 = 200.$$

Pro výslednou hodnotu rozptylu počtu minut mezi vstřelenými góly zbývá dosadit vypočítané hodnoty do vzorce rozptylu

$$var(X) = 100.$$

Rozptyl je 100.
Směrodatnou odchylku počtu minut mezi vstřelenými góly vypočítáme s využitím vzorce z definice 14.

$$s_x = 10 \text{ (minut)}.$$

Směrodatná odchylka je 10 minut.

Kapitola 3

Korelace

V této kapitole se zabýváme popisem statistických souborů pomocí korelace. Slovo korelace můžeme chápat jako nějakou závislost. V této kapitole se narozdíl od obou předchozích kapitol, v nichž jsme uvažovali jeden statistický znak (popřípadě více znaků, ale každý samostatně), zabýváme dvojicí statistických znaků a jejich vzájemnou závislostí (mírou statistické závislosti obou znaků). Pokud za charakteristiku polohy považujeme aritmetické průměry a za charakteristiku variability směrodatné odchylky, tak za charakteristiku korelace označujeme korelační koeficient.

3.1 Základní statistické pojmy ve středoškolské matematice

Nyní popíšeme metodu výpočtu koeficientu korelace.

Definice 33 (Koeficient korelce Calda a Dupač, 1993, str. 149)**.** *Nechť jsou dány dvojice statistických znaků* (x, y)*. Pak je koeficicent korelace* r_{xy} *definován vzorcem:*

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y},$$

s_x *a* s_y *jsou směrodatné odchylky,* $\bar{x}$ *a* $\bar{y}$ *aritmetické průměry znaku* x *a znaku* y*.*

Poznámka. Stejně jako u rozptylu, i tady není předchozí vzorec příliš vhodný pro výpočty, při nichž nemáme k dispozici počítač. Jednoduchou úpravou ho opět převedeme do formy vhodnější pro praktické výpočty. Čitatel ve vzorci vyjádříme takto: nejprve postupně roznásobíme závorky, do celého výrazu přičteme a odečteme $\bar{x}\bar{y}$ a následně vytkneme $\bar{x}$ a $\bar{y}$.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i + \bar{x}\bar{y} - \bar{x}y_i + \bar{x}\bar{y} - \bar{y}x_i - \bar{x}\bar{y}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i y_i + \bar{x}(\bar{y} - y_i) + \bar{y}(\bar{x} - x_i) - \bar{x}\bar{y}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}(\bar{y} - y_i) + \frac{1}{n}\sum_{i=1}^{n}\bar{y}(\bar{x} - x_i) - \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y}. \tag{3.1}$$

Upravíme jednotlivé sumy a opět si sumy rozdělíme. Tak dostaneme rozdíl totožných sčítanců:

$$\frac{1}{n}\sum_{i=1}^{n}\bar{x}(\bar{y} - y_i) = \frac{\bar{x}}{n}\sum_{i=1}^{n}(\bar{y} - y_i) = \frac{\bar{x}\bar{y}}{n}\sum_{i=1}^{n}1 - \frac{\bar{x}}{n}\sum_{i=1}^{n} y_i = 0.$$

$$\frac{1}{n}\sum_{i=1}^{n}\bar{y}(\bar{x} - x_i) = \frac{\bar{y}}{n}\sum_{i=1}^{n}(\bar{x} - x_i) = \frac{\bar{y}\bar{x}}{n}\sum_{i=1}^{n}1 - \frac{\bar{y}}{n}\sum_{i=1}^{n} x_i = 0.$$

$$\frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y} = \frac{n\bar{x}\bar{y}}{n} = \bar{x}\bar{y}.$$

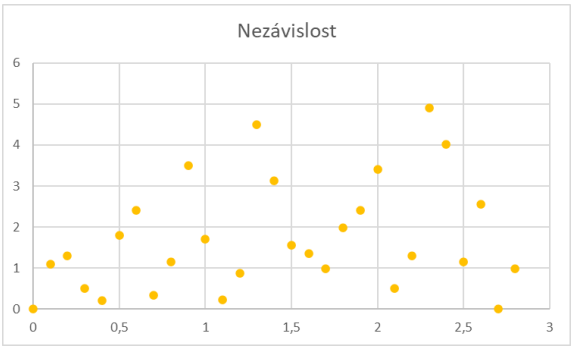
Upravené sumy dosadíme zpět do posledního kroku (3.1) a dostáváme výraz

$$\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}.$$

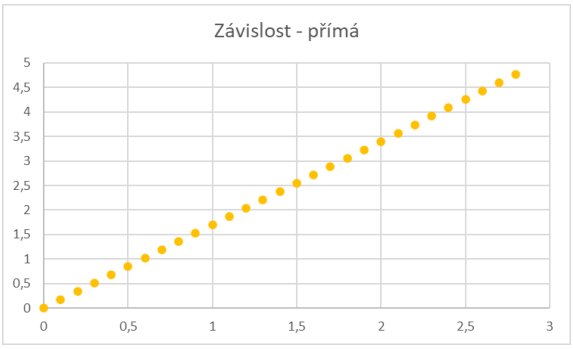
Dosadíme tento výraz do čitatele vzorce z definice 33. Pro r_{xy} dostaneme

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{s_x \cdot s_y}.$$

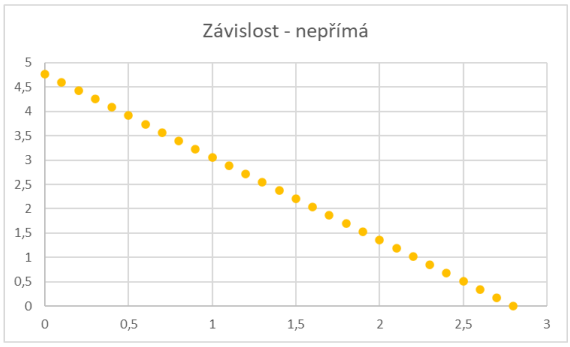
Poznámka (Výsledný koeficient korelace). Koeficient korelace je vždy reálné číslo z intervalu $\langle -1; 1\rangle$.
Pokud jsou jednotlivé znaky nezávislé, je koeficient korelace nula (nebo poblíž nuly). Tento vztah vidíme na obrázku 3.1
Pokud je koeficient korelace kladný a blízký 1, pak je mezi znaky vztah „čím více, tím více“ a označujeme ho přímá závislost. Na obrázku 3.2 vidíme konkrétně korelační koeficient roven 1.
Pokud je koeficient korelace záporný a blízky -1, pak je mezi nimi vztah „čím více, tím méně“ a označujeme ho nepřímá závislost. Na obrázku 3.3 vidíme konkrétně korelační koeficient roven -1.



Obrázek 3.1: Graf nezávislosti



Obrázek 3.2: Graf přímé závislosti



Obrázek 3.3: Graf nepřímé závislosti

3.2 Základní statistické pojmy ve vysokoškolské matematice

V následujících definicích definujeme pojmy, které jsou nutné k zavedení koeficientu korelace; ten jsme v podkapitole 3.1 definovali středoškolským způsobem. Nejprve definujeme náhodný vektor.

Definice 34 (Náhodný vektor Kulich, 2018, str. 15)**.** *Náhodný vektor je (do sloupce) uspořádaná n-tice náhodných veličin, tj.*

$$\mathbf{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))^T .$$

Poznámka. Pro rozlišení náhodného vektoru a náhodné veličiny v textu používáme pro náhodné vektory tučná písmena.

Nyní definujeme sdružené rozdělení, které popisuje rozdělení celého náhodného vektoru, a marginální rozdělení, které popisuje rozdělení jednotlivých náhodných veličin.

Definice 35 (Sdružené rozdělení, sdružená hustota Kulich, 2018, str. 16)**.** *Rozdělení celého náhodného vektoru* $\mathbf{X} = (X_1, \ldots, X_n)$ *se říká sdružené rozdělení. Jeho hustota se nazývá sdružená hustota.*

Definice 36 (Marginální rozdělení, marginální hustota Kulich, 2018, str. 16)**.** *Rozdělením jednotlivých náhodných veličin* $X_1, \ldots, X_n$ *se říká marginální rozdělení.*
Jejich hustoty se nazývají marginální hustoty.

Dále zavádíme charakteristiky, které popisují závislost dvou náhodných veličin.

Definice 37 (Kovariance Kulich, 2018, str. 9)**.** *Nechť jsou dány náhodné veličiny X, Y. Pak kovarianci $cov(XY)$ definujeme pomocí vzorce:*

$$cov(XY) = E((X - EX)(Y - EY)).$$

Poznámka. Vzorec pro výpočet kovariance z definice 37 lze upravit do tvaru:

$$cov(XY) = E(XY) - EXEY.$$

Definice 38 (Korelační koeficient Dupač a Hušková, 1999, str. 47)**.** *Nechť X, Y jsou náhodné veličiny s kladnými a konečnými rozptyly. Korelační koeficient veličin X, Y se značí $\rho(XY)$ a je definován vztahem:*

$$\rho(XY) = \frac{cov(XY)}{\sqrt{varX \cdot varY}}.$$

3.3 Řešené úlohy

V této části uvádíme řešené úlohy, které jsou zaměřené na korelaci. První dvě úlohy jsou záměrně jednodušší, zaměřené na volbu a následné dosazení do správného vzorce. Následuje pět úloh složitějších, ve kterých je nutné nějaké odůvodnění či reprezentace vlastního výsledku, a na závěr dvě úlohy na úrovni vysoké školy, na kterých ukazujeme, jak se výpočetní metody a postupy odlišují od úrovně střední školy. Úlohy jsou zaměřené na praktické použití vzorců z předchozí teoretické podkapitoly. Jsou volené pro úroveň střední školy (gymnázia) doplněné o dvě úlohy na úrovni vysoké školy. U úloh budeme často rozhodovat, zda mezi sebou jednotlivé znaky mají závislost. Pro toto rozhodnutí na základě hodnot koeficientu korelace budeme používat tabulku 3.1[1]. Tato tabulka byla vytvořena autorkou pro potřeby této bakalářské práce. Úlohy v této podkapitole jsou s různorodou tématikou zahrnující více různých sportů.
Úlohy i jejich řešení v této části jsou, pokud není řečeno jinak, autorská.

Absolutní hodnota korelačního koeficnetu	Interpretace
0 - 0,09	žádná
0,1 - 0,25	nízká
0,26 - 0,39	střední
0,4 - 0,65	podstatná
0,66 - 0,85	silná
0,86 - 1	významná

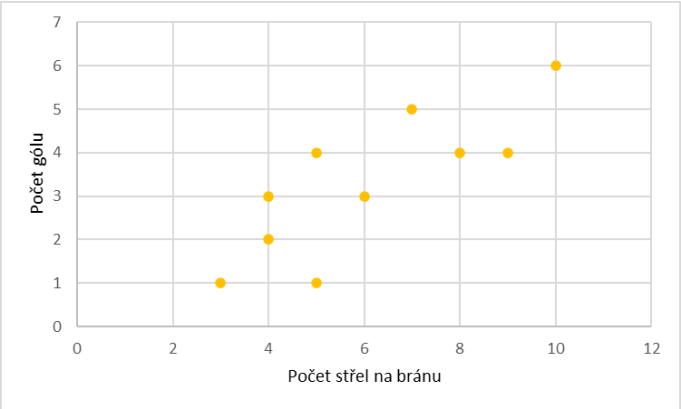
Tabulka 3.1: Interpretace koeficientu korelace

[1]Tabulka 3.1 je vytvořená na základě tabulky, kterou mi interpretoval učitel matematiky na SŠ

3.3.1 Řešené úlohy - jednodušší

Úloha 19. V této úloze budeme pracovat s daty:
Počet střel na bránu: 7, 10, 4, 5, 9, 5, 4, 6, 8, 3.
Počet vstřelenách gólů: 5, 6, 3, 1, 4, 4, 2, 3, 4, 1.
Aritmetický průměr střel: 6,1.
Rozptyl střel: 4,89.
Aritmetický průměr gólů: 3,3.
Rozptyl gólů: 2,41.
Na základě dat v úvodu úlohy vytvořte graf závislosti počtu střel na bránu a počtu vstřelených gólů v 10 fotbalových zápasech Premier League 2017/2018 na přelomu 3. a 4. kola. Na základě grafu zkuste rozhodnout, zda bude koeficient korelace blíže 0, 1, nebo -1. Podle zadaných aritmetických průměrů a rozptylu vypočítejte koeficient korelace a popište výslednou závislost podle tabulky 3.1.

Řešení. Začneme tím, že sestrojíme graf na obrázku 3.4, který bude mít na vodorovné ose data týkající se počtu střel na bránu a na svislé ose týkající se počtu gólů.
Z grafu na obrázku 3.4 odhadujeme, že koeficient korelace bude blízko 1.



Obrázek 3.4: Graf závislosti počtu střel a počtu gólů

Koeficient korelace počítáme podle vzorce: $r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{s_x \cdot s_y}$. Směrodatnou odchylku vypočítáme jako druhou odmocninu z rozptylu.

$$s_{x_S} = \frac{\sqrt{489}}{10} \text{ (střel)},$$

$$s_{x_G} = \frac{\sqrt{241}}{10} \text{ (gólů)}.$$

Nyní dosadíme do vzorce pro výpočet koeficientu korelace a dostaneme:

$$r_{(x_S x_G)} \doteq 0{,}81$$

Koeficient korelace je přibližně 0,81, tedy závislost je významná a přímá.

Úloha 20. V tabulce 3.2 vidíte v prvním sloupci počet vyloučení a ve druhém sloupci počet inkasovaných gólů týmů, které 9.3.2022 prohrály své zápasy v americké hokejové soutěži NHL. Spočítejte koeficient korelace počtu vyloučení a počtu inkasovaných gólů. Na základě tabulky 3.1 popište, zda mezi znaky je nějaká závislost.

Počet vyloučení	Počet inkasovaných gólů
3	5
3	8
1	5
3	3
2	4
5	7
4	9
7	10
2	5
2	3
2	4

Tabulka 3.2: Počet vyloučení a k němu příslušící počet inkasovyných gólů

Řešení. Koeficent korelace počítáme pomocí vzorce

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{s_x \cdot s_y}.$$

K tomu potřebujeme znát $\bar{x}$ a s_x obou znaků. Vzorec pro výpočet aritmetického průměru je

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Do vzorce dosadíme a dostaneme:

$$\bar{x_V} = \frac{34}{11} \text{ (vyloučení)},$$

$$\bar{x_I} = \frac{63}{11} \text{ (inkasovaných gólů)}.$$

Směrodatná odchylka je rovna druhé odmocnině z rozptylu. Rozptyl vypočítáme pomocí vzorce pro výpočet s_x^2. Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme pomocí vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a vypočítáním druhé mocniny z výsledku $\bar{x}$ získáme:

$$(\bar{x_V})^2 = \frac{1\ 156}{121},$$

$$(\bar{x_I})^2 = \frac{3\ 969}{121}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých počtu vyloučení a inkasovaných gólů, tedy

$$(\bar{x_V^2}) = \frac{134}{11},$$

$$(\bar{x_I^2}) = \frac{419}{11}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_V}^2 = \frac{318}{121},$$

$$s_{x_I}^2 = \frac{640}{121}.$$

Směrodatné odchylky s_{x_V} a s_{x_V} jsou druhé odmocniny z příslušných rozptylů $s_{x_V}^2$ a $s_{x_I}^2$

$$s_{x_V} = \frac{\sqrt{318}}{11} \text{ (vyloučení)},$$

$$s_{x_I} = \frac{8\sqrt{10}}{11} \text{ (inkasovaných gólů)}.$$

Na závěr dosadíme do vzorce pro výpočet r_{xy} a dostáváme:

$$r_{(x_V x_I)} \doteq 0{,}76.$$

Koeficient korelace je přibližne 0,76, což znamená, že mezi znaky existuje silná závislost.

3.3.2 Řešené úlohy - obtížnější

Úloha 21. Nebylo možné si nepovšimnout, že se českým a slovenským biatlonistům a biatlonistkám ve sprintu na ZOH 2022 nedařilo podle představ, především umístění nebyla u většiny z nich uspokojující. Jedním z možných důvodů je neúspěšnost střelby a s ní spojený počet trestných kol. Na základě tabulky 3.3 vypočítejte koeficient korelace, vysvětlete, co říká o souvislosti mezi trestnými koly a umístěním závodníků. Pokud by koeficient korelace závislost neprokázal, zkuste vymyslet jiné možné důvody špatného umístění závodníků.

Závodník(ce)	Umístění	Počet trestných kol
Fialková P.	14	2
Charvátová	25	1
Jislová	31	1
Davidová	41	4
Fialková I.	41	4
Voborníková	58	2
Horvátová	72	1
Machyniaková	87	3
Krčmář	16	1
Karlík	28	3
Štvrtecký	58	3
Václavík	59	3
Bartko	65	3
Šíma	68	1
Sklenárik	87	3

Tabulka 3.3: Úspěšnost biatlonistů(tek) ve sprintu na ZOH 2022

Řešení. Vzorec pro výpočet koeficientu korelace je:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{s_x \cdot s_y}.$$

K jeho využití potřebujeme $\bar{x}$ a s_x obou znaků. Vzorec pro výpočet aritmetického průměru je

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Do vzorce dosadíme a dostaneme:

$$\bar{x_U} = 50 \text{ (50. pozice)},$$

$$\bar{x_T} = \frac{7}{3} \text{ (trestných kol)}.$$

Směrodatná odchylka je rovna druhé odmocnině z rozptylu. Rozptyl vypočítáme pomocí vzorce pro výpočet s_x^2, do kterého dosadíme a výsledek umocníme na druhou mocninu. Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme pomocí vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a vypočítáním druhé mocniny z výsledku $\bar{x}$. Získáme:

$$(\bar{x_U})^2 = 2\ 500,$$

$$(\bar{x_T})^2 = \frac{49}{9}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých umístění a počtu trestných kol, tedy

$$(\bar{x_U^2}) = 3\ 037{,}6,$$

$$(\bar{x_T^2}) = \frac{99}{15}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_U}^2 = 537{,}6,$$

$$s_{x_T}^2 = \frac{52}{45}.$$

Směrodatné odchylky s_{x_V} a s_{x_V} jsou druhé odmocniny z příslušných rozptylů $s_{x_V}^2$ a $s_{x_I}^2$

$$s_{x_U} = \frac{8\sqrt{210}}{5} \text{ (umístění)},$$

$$s_{x_T} = \frac{2\sqrt{13}}{7} \text{ (trestných kol)}.$$

Na závěr dosadíme do vzorce pro výpočet r_{xy} a dostaneme:

$$r_{(x_U x_T)} \doteq 0{,}23.$$

Koeficient korelace je 0,23, což vyjadřuje nízkou závislost.
Dalšími možnými důvody, proč se závodníkům nedařilo, může být například špatná volba lyží, u některých biatlonistů únava z předchozího závodu smíšených štafet, nedostatečná psychická příprava nebo souhra různých okolností (protivítr, zlomená hůl, únava, špatný den, ...). Také podle výsledků nemůžeme říct, že se všem nedařilo, například biatlonistka Fialková byla úspěšná.

Úloha 22. Data k této úloze:
Počet úspěšných es: 2, 5, 14, 7, 20, 14, 3.
Počet her se ztraceným podáním: 5, 4, 5, 5, 3, 3, 6.
Rafael Nadal se zúčastnil tenisového turnaje ATP Australian Open, který vyhrál. V celém turnaji odehrál 7 zápasů a ve všech se stal vítězem. Jedno z vysvětlení, proč se mu v turnaji tolik dařilo, se může skrývat v počtu úspěšných es a počtu her se ztraceným podáním. Počty jednotlivých es v zápasech a počty her se ztraceným podáním ve stejném pořadí vidíte v úvodu této úlohy. Očekáváte, že počty her se ztraceným podáním budou mít souvislost s počtem es? Proč? Spočítejte koeficient korelace a na základě jeho výsledku rozhodněte, zda na sobě závisí, či nikoli.

Řešení. **Před začátkem výpočetního postupu se zamysleme nad tím, zda spolu mohou mít jednotlivá data nějakou spojitost. Očekáváním může být, že při vysokém počtu es podávajícího hráče není dovoleno protihráči sebrat (podávajícímu hráči) podání, a tedy jednotlivé veličiny na sobě nějak závisí.**
Koeficient korelace vypočítáme pomocí vzorce:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})}{s_x \cdot s_y}.$$

Potřebujeme vypočítat $\bar{x}$ a s_x obou znaků. Vzorec pro výpočet aritmetického průměru je

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Do vzorce dosadíme a dostaneme:

$$\bar{x_E} = \frac{65}{7} \text{ (es)},$$

$$\bar{x_H} = \frac{7}{3} \text{ (her)}.$$

Směrodatná odchylka je rovna druhé odmocnině z rozptylu. Rozptyl vypočítáme pomocí vzorce pro výpočet s_x^2, do kterého dosadíme a výsledek umocníme na druhou mocninu. Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme pomocí vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a vypočítáním druhé mocniny z výsledku $\bar{x}$. Získáme:

$$(\bar{x_E})^2 = \frac{4\ 225}{49},$$

$$(\bar{x_H})^2 = \frac{961}{49}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých počtu es a her se ztraceným podáním, tedy

$$(\bar{x_E^2}) = \frac{879}{7},$$

$$(\bar{x_H^2}) = \frac{145}{7}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_E}^2 = \frac{1\ 928}{49},$$

$$s_{x_H}^2 = \frac{54}{49}.$$

Směrodatné odchylky s_{x_E} a s_{x_B} jsou druhé odmocniny z příslušných rozptylů $s_{x_E}^2$ a $s_{x_B}^2$

$$s_{x_E} = \frac{2\sqrt{482}}{7} \text{ (es)},$$

$$s_{x_H} = \frac{3\sqrt{6}}{7} \text{ (her)}.$$

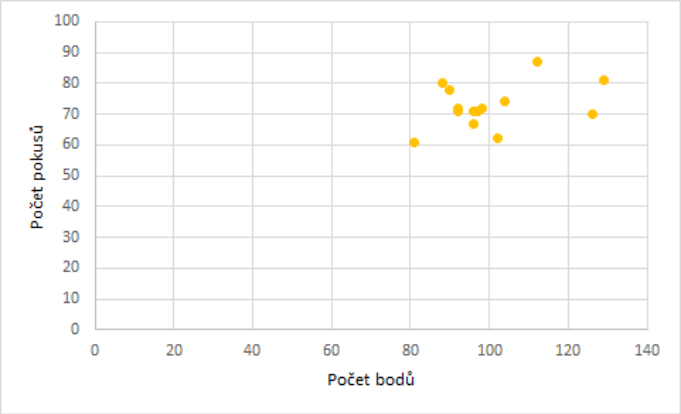
Na závěr dosadíme do vzorce pro výpočet r_{xy} a dostaneme:

$$r_{(x_E x_H)} \doteq -0{,}71.$$

Koeficient korelace je -0,71, takže počet es a počet her se ztraceným podáním spolu spíše souvisí a závislost je silná. U Rafaela Nadala se na tomto turnaji v větším počtem es pojí méně ztracených podání.

Úloha 23. Tým Basketball Nymburk v soutěži NBL (Národní basketbalová liga) - část vítězů 2022 (označení části soutěže) získal ve 14 zápasech tyto počty bodů: 102, 104, 129, 92, 90, 92, 81, 112, 97, 98, 96, 126, 88 a 96 za tyto počty střeleckých pokusů: 62, 74, 81, 72, 78, 71, 61, 87, 71, 72, 67, 70, 80 a 71. Na základě těchto dat sestrojte graf závislosti (viz 3.1) počtu získaných bodů a počtu střeleckých pokusů, popište ho a rozhodněte, zda se jedná o graf závislosti (přímé, či nepřímé), nebo nezávislosti. Na základě grafu odhadněte, jaký by přibližně mohl být koeficient korelace. Na závěr vypočítejte skutečnou hodnotu koeficientu korelace a porovnejte ji s vaší odhadovanou hodnotou.

Řešení. Začneme tím, že sestrojíme graf na obrázku 3.5. Na vodorovnou osu vynášíme hodnoty popisující získané body a na svislou osu vynášíme data o střeleckých pokusech. Popis grafu:



Obrázek 3.5: Graf závislosti počtu pokusů a počtu bodů

Graf má na vodorovné ose hodnoty počtu získaných bodů, které se pohybují přibližně od 80 do 130 bodů. Na svislé ose vidíme počet všech střel. Jejich hodnoty se pohybují přibližně od 60 do 90 pokusů.
Graf působí spíše jako graf závislosti (nízké, nebo střední). Zdá se, že některé body náleží přímce (ose prvního a třetího kvadrantu), která je rostoucí, a proto odhaduji koeficient korelace na hodnotu 0,3, tedy střední závislost.
Skutečnou hodnotu koeficientu korelace vypočítáme pomocí vzorce:

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})}{s_x \cdot s_y}.$$

Potřebujeme vypočítat $\bar{x}$ a s_x obou znaků. Vzorec pro výpočet aritmetického průměru je

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Do vzorce dosadíme a dostaneme:

$$\bar{x_B} = \frac{1\ 403}{14}\ \text{(bodu)},$$

$$\bar{x_S} = \frac{1\ 017}{14}\ \text{(střel)}.$$

Směrodatnou odchylku vypočítáme jako druhou odmocninu z rozptylu a rozptyl vypočítáme pomocí vzorce pro výpočet s_x^2, do kterého dosadíme a výsledek umocníme na druhou. Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$.

$$\bar{x_B}^2 = \frac{1\ 968\ 409}{196},$$

$$\bar{x_S}^2 = \frac{1\ 034\ 289}{196}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých počtu bodů a střel, tedy

$$(\bar{x_B^2}) = \frac{20\ 437}{2},$$

$$(\bar{x_S^2}) = \frac{74\ 535}{14}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_B}^2 = \frac{34\ 417}{196},$$

$$s_{x_S}^2 = \frac{9\ 201}{196}.$$

Směrodatné odchylky s_{x_V} a s_{x_V} jsou druhé odmocniny z příslušných rozptylů $s_{x_V}^2$ a $s_{x_I}^2$

$$s_{x_B} = \frac{\sqrt{34\ 417}}{14} \text{ (bodu)},$$

$$s_{x_S} = \frac{\sqrt{9\ 201}}{14} \text{ (střel)}.$$

Nyní dosadíme do vzorce pro výpočet r_{xy} a dostáváme:

$$r_{(x_B x_S)} \doteq 0{,}36.$$

Koeficient korelace je přibližně 0,36. Odhad nebyl zcela přesný, ale teoretický podklad k odhadu ano. Skutečně se jedná o střední závislost. Některé body leží na rostoucí přímce, která je osou prvního a třetího kvadrantu.

Úloha 24. Vysvětlivky a data k úloze:
All yards: All yards je slovní spojení, které v americkém fotbalu popisuje počet všech přihrávek a průniků jednoho týmu.
Počet all yards: 346, 431, 369, 449, 298, 409, 314, 376, 404, 405, 318, 331, 325, 475, 390.
Rushing yards: Rushing yards popisuje počet všech průniků jednoho týmu,
Počet rushing yards: 174, 202, 195, 135, 195, 85, 79, 205, 99, 124, 101, 104, 226, 171, 135.
Na základě dat z 18. kola NFL amerického fotbalu (všechna data jsou týmů, které vyhrály) týkajících se all yards a rushing yards rozhodněte, zda je mezi znaky závislost, či nikoliv. V obou případech nakreslete graf korelace. Pokud na sobě znaky závisí, určete o jakou závislost se jedná.

Řešení. Koeficent korelace počítáme pomocí vzorce

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i) - \bar{x}\bar{y}}{s_x \cdot s_y}.$$

K tomu potřebujeme znát $\bar{x}$ a s_x obou znaků. Vzorec pro výpočet aritmetického průměru je

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Do vzorce dosadíme a dostaneme:

$$\bar{x_A} = 376 \text{ (all yards)},$$

$$\bar{x_R} = \frac{446}{3} \text{ (rushing yards)}.$$

Směrodatná odchylka je rovna druhé odmocnině z rozptylu. Rozptyl vypočítáme pomocí vzorce pro výpočet s_x^2. Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme pomocí vzorce

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

a vypočítáním druhé mocniny z výsledku $\bar{x}$. Získáme:

$$(\bar{x_A})^2 = 141\ 376,$$

$$(\bar{x_R})^2 = \frac{198\ 916}{9}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých all yards a rushing yards, tedy

$$(\bar{x_A^2}) = \frac{2\ 160\ 572}{15},$$

$$(\bar{x_R^2}) = \frac{365\ 582}{15}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s_{x_A}^2 = \frac{39\ 932}{15},$$

$$s_{x_R}^2 = \frac{102\ 166}{45}.$$

Směrodatné odchylky s_{x_V} a s_{x_V} jsou druhé odmocniny z příslušných rozptylů $s_{x_V}^2$ a $s_{x_I}^2$

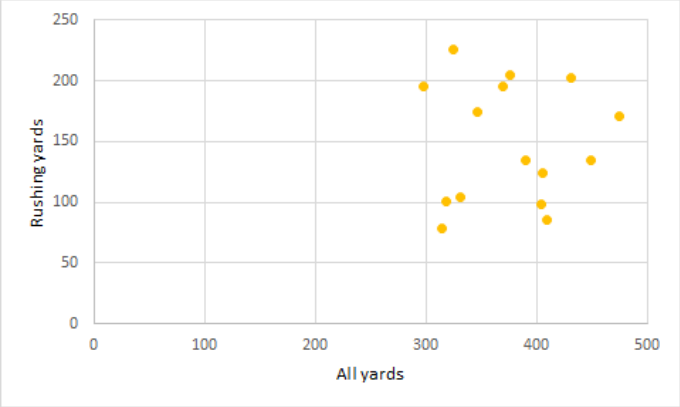
$$s_{x_A} = \frac{2\sqrt{149\ 745}}{15} \text{ (all yards)},$$

$$s_{x_R} = \frac{\sqrt{510\ 830}}{15} \text{ (rushing yards)}.$$

Na závěr dosadíme do vzorce pro výpočet r_{xy} a dostaneme:

$$r_{(x_A x_R)} \doteq 0{,}005$$

Koeficient korelace je přibližně 0,005, statistické znaky mezi sebou nemají žádnou závislost.
Graf korelace je na obrázku 3.6:



Obrázek 3.6: Graf závislosti all yards a rushing yards

Úloha 25. Říká se, že na fotbalových zápasech jsou pro tým, který hraje doma, fanoušci dvanáctým hráčem. V této úloze se pokusíme toto tvrzení potvrdit, či vyvrátit na základě výsledků 15 domácích zápasů fotbalového týmu SK Slavia Praha (v sezóně 2018/2019). Zjistěte, jestli je souvislost mezi počtem diváků na těchto zápasech a počtem vstřelených gólů.
Uvádíme dvojice: počet vstřelených gólů - počet diváků: 2 - 12 956, 1 - 19 370, 1 - 16 217, 4 - 14 123, 4 - 10 089, 2 - 10 847, 3 - 9 091, 3 - 10 107, 4 - 11 108, 4 - 10 747, 1 - 15 272, 4 - 17 338, 0 - 13 263, 3 - 13 127, 4 - 12 103. Dříve, než úlohu budete řešit, se zkuste zamyslet, zda očekáváte, že skutečně počet diváků má vliv na počet vstřelených gólů domácího týmu, či nikoli. Následně spočítejte koeficient korelace a popište, zda váš názor potvrzuje, či vyvrací. Navrhněte důvod, proč jsou (nebo nejsou) znaky závislé.

Řešení. **Předpokládejme například, že počet fanoušků na stadionu skutečně má vliv na to, kolik gólů tým vstřelí.** Jedním z důvodů, proč by tomu tak mohlo být, je ten, že větší množství diváků motivuje a povzbuzuje více, fotbalisté se (i pro ně) snaží hrát na pohled hezký fotbal a dát více gólů. Jednoduše řečeno, nejde jim jen o výhru.
Koeficient korelace počítáme podle vzorce $r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})}{s_x \cdot s_y}$. K jeho využití potřebujeme $\bar{x}$ a s_x obou znaků. Vzorec pro výpočet aritmetického průměru je

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Do vzorce dosadíme a dostaneme:

$$\bar{x_G} = \frac{8}{3} \text{ (gólu)},$$

$$\bar{x_D} = \frac{195\ 758}{15} \text{ (diváků)}.$$

Směrodatná odchylka je rovna druhé odmocnině z rozptylu. Rozptyl vypočítáme pomocí vzorce pro výpočet s_x^2, do kterého dosadíme a výsledek umocníme na druhou. Pro dosazení do něj vypočítáme hodnoty $(\bar{x})^2$ a $(\bar{x^2})$. $(\bar{x})^2$ vypočítáme pomocí vzorce

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

a vypočítáním druhé mocniny z výsledku $\bar{x}$. Získáme:

$$(\bar{x_G})^2 = \frac{64}{9},$$

$$(\bar{x_D})^2 = \frac{38\ 321\ 194\ 564}{225}.$$

$(\bar{x^2})$ vypočítáme jako aritmetický průměr druhých mocnin jednotlivých počtu gólů a diváků, tedy

$$(\bar{x_G^2}) = \frac{134}{15},$$

$$(\bar{x_D^2}) = \frac{99}{15}.$$

Nyní do vzorce pro výpočet rozptylu dosadíme:

$$s^2_{x_G} = \frac{82}{45},$$

$$s^2_{x_D} = \frac{1\ 836\ 534\ 266}{225}.$$

Směrodatné odchylky s_{x_V} a s_{x_V} jsou druhé odmocniny z příslušných rozptylů $s^2_{x_V}$ a $s^2_{x_I}$

$$s_{x_G} = \frac{\sqrt{410}}{15} \text{ (gólu)},$$

$$s_{x_D} = \frac{\sqrt{1\ 836\ 534\ 266}}{15} \text{ (diváků)}.$$

Na závěr dosadíme do vzorce pro výpočet r_{xy} a dostaneme:

$$r_{(x_G x_D)} \doteq -0{,}42.$$

Koeficient korelace je -0,42.
Závislost je významná, ale také pro ni platí, že „čím více, tím méně“, tedy čím je více diváků, tím padá méně gólů. Což náš předpoklad vyvrací.
Jedním vysvětlením může být to, že diváci navštěvují především zápasy mezi velkými týmy (například zápasy Sparty a Slavie), ve kterých padá menší množství gólů, protože oba týmy jsou na podobné úrovni.
Tedy více diváků vidí zápasy, ve kterých padne méně gólů.

3.3.3 Řešené úlohy vysokoškolského charakteru

Zadání těchto úloh jsou celá vymyšlená. Nejsou založená na reálných datech ani pravděpodobnostech.

Úloha 26. Jsou dány náhodně veličiny X a Y. Náhodná veličina X (v řádku) popisuje počet žlutých karet v jednom zápase jednoho týmu a náhodná veličina Y (ve sloupci) popisuje počet inkasovaných gólů stejného týmu v totožném zápase. Sdružené rozdělení náhodných veličin X, Y je zadáno pomocí tabulky 3.4. Vypočítejte korelaci náhodných veličin X a Y.

Žluté karty/inkasované góly	0	1	2	3	4
0	0,01	0,04	0,04	0,01	0
1	0,04	0,16	0,11	0,07	0,02
2	0,08	0,08	0,1	0,03	0,01
3	0,07	0,02	0,05	0,03	0,03

Tabulka 3.4: Tabulka udávající sdružené rozdělení náhodných veličin X a Y

Řešení. V řešení začneme rozšířením tabulky o součty pravděpodobností ve sloupcích a řádcích, čímž dostaneme marginální rozdělení X, Y. Vzorec pro výpočet

Žluté karty/inkasované góly	0	1	2	3	4	
0	0,01	0,04	0,04	0,01	0	0,1
1	0,04	0,16	0,11	0,07	0,02	0,4
2	0,08	0,08	0,1	0,03	0,01	0,3
3	0,07	0,02	0,05	0,03	0,03	0,2
	0,2	0,3	0,3	0,14	0,06	

Tabulka 3.5: Tabulka rozšířená o součty pravděpodobností

korelace v definici 38 použijeme ve tvaru, který má v čitateli $cov(XY)$ a ve jmenovateli odmocninu z $var(X)$ a $var(Y)$.

$$\rho(XY) = \frac{E(XY) - EXEY}{\sqrt{(EX^2 - (EX)^2)(EY^2 - (EY)^2)}}.$$

Neznámé hodnoty ze vzorce $\rho(XY)$ jsou: $E(XY)$, EX, EY, EX^2 a EY^2.
Začneme výpočtem $E(XY)$. Využitím vzorce (2.3) ($h : \mathbb{R}^2 \to \mathbb{R}^2, h(X, Y) = XY$) a vzorce (2.1) dostáváme tento vzorec:

$$E(XY) = \sum_{k=0}^{3} \sum_{j=0}^{4} kj \,\mathsf{P}(X = k, Y = j) =$$

$$= (0{,}16 + 2 \cdot 0{,}11 + 3 \cdot 0{,}07 + 4 \cdot 0{,}02) + (2 \cdot 0{,}08 + 4 \cdot 0{,}1 + 6 \cdot 0{,}03 + 8 \cdot 0{,}01) + \\ + (3 \cdot 0{,}02 + 6 \cdot 0{,}05 + 9 \cdot 0{,}03 + 12 \cdot 0{,}03).$$

$$E(XY) = 2{,}36.$$

Pokračujeme výpočty EX a EY.

$$EX = \sum_{k} k \,\mathsf{P}(X = k) = 1{,}6.$$

$$EY = \sum_{j} j \,\mathsf{P}(Y = j) = 1{,}56.$$

Dále vypočítáme EX^2 a EY^2.

$$EX^2 = \sum_{k=0}^{3} k^2 \,\mathsf{P}(X = k) = 3{,}4.$$

$$EY^2 = \sum_{j=0}^{4} j^2 \,\mathsf{P}(Y = j) = 3{,}72.$$

Pro kompletní vyřešení úlohy zbývá dosadit vypočítané hodnoty do vzorce pro výpočet korelace.

$$\rho(XY) = \frac{2{,}36 - 1{,}6 \cdot 1{,}56}{\sqrt{(3{,}4 - 2{,}56)(3{,}72 - 2{,}4\ 336)}} = -0{,}13.$$

Korelace je -0,13.

Úloha 27. Jsou dány náhodné veličiny X a Y. Náhodná veličina X (v řádku) popisuje počet vstřelených gólů v přesilovce v jednom zápase jednoho hokejového týmu a náhodná veličina Y (ve sloupci) popisuje počet vstřelených gólů v oslabení stejného týmu v témže zápase. Sdružené rozdělení náhodných veličin X, Y je zadáno pomocí tabulky 3.6. Vypočítejte korelaci náhodných veličin X a Y.

Přesilovka/oslabení	0	1	2
0	0,4	0,07	0,03
1	0,15	0,09	0,01
2	0,08	0,06	0,1
3	0,07	0,03	0

Tabulka 3.6: Tabulka udávající sdružené rozdělení náhodných veličin X a Y

Řešení. Na řešení této úlohy se vztahují stejné komentáře jako v předchozí úloze 26.
Začneme sestavením tabulky 3.7 doplněné o hodnoty součtu pravděpodobností ve sloupcích a řádcích, kterými získáme marginální rozdělení X, Y.

Přesilovka/oslabení	0	1	2	
0	0,4	0,07	0,03	0,5
1	0,15	0,09	0,01	0,25
2	0,08	0,06	0,1	0,15
3	0,07	0,03	0	0,1
	0.7	0.25	0.05	

Tabulka 3.7: Tabulka udávající sdružené rozdělení náhodných veličin X a Y

Vzorec pro výpočet korelace:

$$\rho(XY) = \frac{E(XY) - EXEY}{\sqrt{(EX^2 - (EX)^2)(EY^2 - (EY)^2)}}.$$

Nyní vypočítáme $E(XY)$.

$$E(XY) = \sum_{k=0}^{3} \sum_{j=0}^{2} kj \,\mathsf{P}(X = k, Y = j) =$$

$$= (0{,}09 + 2 \cdot 0{,}01) + (2 \cdot 0{,}06 + 4 \cdot 0{,}01) + 3 \cdot 0{,}03.$$

$$E(XY) = 0{,}36.$$

Pokračujeme výpočty EX a EY.

$$EX = \sum_k k \, \mathsf{P}(X = k) = 0{,}85.$$

$$EY = \sum_j j \, \mathsf{P}(Y = j) = 0{,}6.$$

Dále vypočítáme EX^2 a EY^2.

$$EX^2 = \sum_{k=0}^{3} k^2 \, \mathsf{P}(X = k) = 1{,}75.$$

$$EY^2 = \sum_{j=0}^{2} j^2 \, \mathsf{P}(Y = j) = 0{,}45.$$

Ny závěr dosadíme do vzorce pro výpočet $\rho(XY)$.

$$\rho(XY) = \frac{0{,}36 - 0{,}85 \cdot 0{,}6}{\sqrt{(1{,}75 - 0{,}7\ 225)(0{,}45 - 0{,}36)}} = -0{,}49.$$

Korelace je -0,49.

Kapitola 4

Využití statistiky při testování hypotéz ve sportu

V této kapitole se zabýváme testováním hypotéz. Nejprve definujeme základní pojmy, které využíváme při zavedení statistického χ^2 testu nezávislosti. V závěru kapitoly ukazujeme praktické použití testu na reálných datech se sportovní fotbalovou tématikou.
První tři definice (náhodný výběr, model pro pozorování a statistika) jsou základní objekty pro budování teorie testování hypotéz.

Definice 39 (Náhodný výběr Kulich, 2014, str. 3)**.** *Posloupnost $X_1, X_2, \ldots, X_n$ nezávislých stejně rozdělených náhodných veličin, z nichž každá má distribuční funkci F_0, nazýváme náhodný výběr z rozdělení F_0. Konstantu n nazýváme rozsah výběru.*

Definice 40 (Model Kulich, 2014, str. 3)**.** *Modelem pro náhodné veličiny $X_1, X_2, \ldots, X_n$ rozumíme předem stanovenou množinu rozdělení $\mathcal{F}$, do níž patří neznámé rozdělení F_0.*

Rozdělení F_0 veličin z náhodného výběru pocházející z modelu $\mathcal{F}$ je určeno různými charakteristikami (např. střední hodnota, rozptyl, distribuční funkce), které neznáme. Nazýváme je parametry. Účelem testování je testování parametru na základě pozorovaných hodnot náhodných veličin (označujeme jako pozorování)(Kulich, 2014, str. 3).

Definice 41 (Statistika Kulich, 2014, str. 4)**.** *Pojmem statistika nazýváme libovolnou měřitelnou funkci $S(\mathbf{X})$ veličin z náhodného výběru. Statistika je náhodná veličina.*

Pro účely teorie testování hypotéz statistiku $S(\mathbf{X})$ nazýváme testovou statistikou a $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ je náhodný vektor prvků náhodného výběru.

Po definicích základních pojmů definujeme hypotézu, kterou testujeme proti alternativě.

Definice 42 (Hypotéza a alternativa Kulich, 2014, str. 30)**.** *Nechť je dána množina* $\Theta = \{t(F)\ ^1,\ F \in \mathcal{F}\} \subseteq \mathbb{R}$*. Dále nechť jsou dány dvě neprázdné disjunktní podmnožiny množiny* Θ*, a to* Θ_0 *a* Θ_1*. Pak množinu* Θ_0 *nazýváme nulová hypotéza a množinu* Θ_1 *nazýváme alternativa.*
Hypotézu označujeme symbolem H_0 *a alternativu symbolem* H_1*.*

Testování hypotézy proti alternativě na základě testové statistiky probíhá na principu zkoumání, zda hodnota testové statistiky patří do předem určené množiny, kterou nazýváme kritický obor, či nikoliv.

Definice 43 (Kritický obor DeGroot, 1975, str. 370)**.** *Nechť* S *je množina všech možných hodnot náhodného výběru* $X_1, X_2, \ldots, X_n$*. Pak říkáme, že kritický obor* $\mathcal{C}$ *je podmnožina množiny* S *obsahující hodnoty náhodného výběru, při kterých zamítáme hypotézu* H_0*.*

Dále definujeme statistický test, popisujeme jeho základní interpretaci a definujeme základní vlastnost hladinu testu.

Definice 44 (Statistický test Kulich, 2014, str. 31)**.** *Nechť je dána testová statistika* $S(\mathbf{X})$ *a kritický obor* $\mathcal{C}$*. Pak statistický test definujeme jako dvojici* $(S(\mathbf{X}), \mathcal{C})$*.*

Výsledek testu na základě testové statistiky a kritického oboru interpretujeme dvěma možnými způsoby:
Pokud testová statistika patří do kritického oboru, zamítáme hypotézu ve prospěch alterernativy.
Pokud do kritického oboru nepatří, říkáme, že hypotézu nemůžeme zamítnout.

Definice 45 (Hladina testu Kulich, 2014, str. 31)**.** *Nechť* $\alpha \in (0, 1)$ *je předem stanovené číslo. Jestliže kritický obor* $\mathcal{C}$ *splňuje podmínku:*

$$\sup_{\theta \in \Theta_0} P_\theta[S(\mathbf{X}) \in \mathcal{C}] = \alpha,^2$$

říkáme, že test $(S(\mathbf{X}), \mathcal{C})$ *má hladinu* α*.*
Jestliže kritický obor $\mathcal{C}$ *splňuje předchozí podmínku asymptoticky*[3] *pro* $n \to \infty$*, říkáme, že test* $(S(\mathbf{X}), \mathcal{C})$ *má asymptoticky hladinu* α*.*

Vyhodnocení testu na základě toho, zda $S(\mathbf{X}) \in \mathcal{C}$, je méně běžný způsob posouzení výsledku. Častěji k posouzení využíváme p-hodnoty neboli dosažené hodnoty testu ((Kulich, 2014), str. 34).

Definice 46 (P-hodnota Kulich, 2014, str. 34)**.** *Nechť je dán kritický obor* $\mathcal{C} = \mathbb{R}\backslash(c_L, c_U)$*, kde* $-\infty \leq c_L < c_U \leq \infty$*. Necht* $s_{\mathbf{x}}$ *je pozorovaná hodnota testové statistiky* $S_{\mathbf{X}}$*.*

---

[1] $t(F)$ značíme také θ

[2] $\mathsf{P}_\theta[S(\mathbf{X}) \in \mathcal{C}]$ je pravděpodobnost, že $S(\mathbf{X}) \in \mathcal{C}$, předpokládáme-li, že $\mathbf{X}$ má rozdělení θ

[3] $\lim_{n\to\infty} (\sup_{\theta\in\Theta_0} \mathsf{P}_\theta[S(X_1, X_2, \ldots, X_n) \in \mathcal{C}]) = \alpha$

Pak p-hodnotu neboli dosaženou hladinu testu definujeme jako:

$$p(x) = P_{\theta_0}[S(\mathbf{X}) \geq s_{\mathbf{x}}] = 1 - F_0(s_{\mathbf{x}}) \text{ pokud } c_L = -\infty;$$

$$p(x) = P_{\theta_0}[S(\mathbf{X}) \leq s_{\mathbf{x}}] = F_0(s_{\mathbf{x}}) \text{ pokud } c_U = \infty;$$

$$p(x) = 2\min\left(P_{\theta_0}[S(\mathbf{X}) \geq s_{\mathbf{x}}], P_{\theta_0}[S(\mathbf{X}) \leq s_{\mathbf{x}}] = \right.$$

$$= 2\min\{(1 - F_0(s_{\mathbf{x}}), F_0(s_{\mathbf{x}}))\} \text{ pokud } c_L \text{ a } c_U \text{ jsou konečné a } F_0(c_L) =$$

$$= 1 - F_0(c_U) = \alpha/2.$$

Věta 1 (Kulich, 2014, str. 35)**.** *Zamítáme-li hypotézu podle pravidla*

$$H_0 \text{ zamítáme, jestliže } p(x) \leq \alpha$$

$$H_0 \text{ nezamítáme, jestliže } p(x) > \alpha,$$

výsledný test má hodnotu α (přesně, nebo asymptoticky).

4.1 χ^2 test

Nyní obecně podle (DeGroot, 1975, str. 452) zavádíme χ^2 test nezávislosti. V testu používáme značení z (DeGroot, 1975, str. 453).

Je dána tabulka o R řádcích a C sloupcích (budeme psát tabulka R x C). V tabulce uvádíme pravděpodobnosti p_{ij} a četnosti $N_{ij}, i = 1, \ldots, R; j = 1, \ldots, C$. Symbolem p_{ij} označujeme pravděpodobnost jevu $[X = i, Y = j]$ a symbolem N_{ij} označujeme četnost tohoto jevu pro $i = 1, \ldots, R; j = 1, \ldots, C$. Nazýváme ji kontingenční tabulka.

Pravděpodobnosti p_{ij} udávají sdružené rozdělení (X, Y). Nyní představujeme pravděpodobnosti udávající marginální rozdělení (X, Y):

$$p_{i.} = \sum_{j=1}^{C} p_{ij} \text{ a } p_{.j} = \sum_{i=1}^{R} p_{ij},$$

a četnosti jevu $X = i, Y = j$:

$$N_{i.} = \sum_{j=1}^{C} N_{ij}, N_{.j} = \sum_{i=1}^{R} N_{ij}. \tag{4.1}$$

Součet všech četností označujeme n:

$$\sum_{i=1}^{R} \sum_{j=1}^{C} N_{ij} = n.$$

V dalším kroku zavádíme hypotézu testovanou proti alternativě.
H_0 je hypotéza testu : $p_{ij} = p_{i.}p_{.j}$ pro $i = 1, \ldots, R; j = 1, \ldots, C$.

Hypotézu interpretujeme jako nezávislost veličin X, Y. Z této interpretace vychází název testu χ^2 test **nezávislosti**.
H_1 je alternativa testu : Hypotéza H_0 není pravdivá.

Test je určen testovou statistikou a kritickým oborem. Testová statistika Q je ve tvaru:

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - \hat{\mathrm{E}}_{ij})^2}{\hat{\mathrm{E}}_{ij}}, \tag{4.2}$$

kde $\hat{\mathrm{E}}_{ij}$ představuje očekávanou (teoretickou) četnost (za platnosti hypotézy) zadanou vzorcem:

$$\hat{\mathrm{E}}_{ij} = \frac{N_{i.} N_{.j}}{n}. \tag{4.3}$$

Testová statistika má asymptoticky pro $(n \to \infty)$ χ^2 rozdělení o $(R-1)(C-1)$ stupních volnosti. Z tohoto asymptotického rozdělení lze zjistit kritický obor, do kterého spadá testová statistika Q, když hypotézu H_0 zamítáme. V tomto případě hypotézu zamíteme právě tehdy, když $Q \geq \chi^2_{(R-1)(C-1)}(1-\alpha)$, kde α je hladina testu a $\chi^2_{(R-1)(C-1)}(1-\alpha)$ označuje $(1-\alpha)$-tý kvantil χ^2 rozdělení o $(R-1)(C-1)$ stupních volnosti.

4.2 Ukázka χ^2 testu na reálných datech

Nyní ukazujeme použití χ^2 testu, který jsme v předchozí podkapitole 4.1 teoreticky zavedli, na konkrétním příkladu reálných dat.

Testovaná data jsou ze sezóny 2018/2019, ve které nebyl omezen divákům přístup na stadiony a, v sezóně 2020/2021, ve které byl omezen divákům přístup na stadiony, ze soutěže Fortuna liga. Počítáme se všemi zápasy, které se v této sezóně odehrály.

Testujeme dvě náhodné veličiny X a Y. Veličina X popisuje, zda byl počet diváků na stadionu omezen (kvantifikujeme jako $X = 0$), či byla možná návštěva stadionu bez omezení[4] (kvantifikujeme jako $X = 1$). Veličina Y označuje gólový rozdíl v zápasu z pohledu domácího týmu.[5] Veličina Y nabývá 9 možných hodnot (viz tabulka 4.1). Výsledky vzájemných zápasů jsou nezávislé a stejně rozdělené. Užitím testu chceme posoudit, zda má přítomnost diváků vliv na výsledek zápasu z pohledu domácího týmu.

Četnosti rozdílů vidíme v tabulce 4.1, kde $R = 2$ a $C = 9$, dle značení zavedeného v podkapitole 4.1.

[4]Diváci byli přítomni.

[5]Pokud zápas dopadl 2:1 (z pohledu domácího týmu), gólový rozdíl je 1. Pokud zápas skončil 1:2, gólový rozdíl je -1.

	Y $\leq$ -4	Y = -3	Y = -2	Y = -1	Y = 0
X = 0	6	12	33	48	82
X = 1	8	9	19	35	46
	Y = 1	Y = 2	Y = 3	Y $\geq$ 4	
X = 0	49	38	26	12	
X = 1	60	31	18	14	

Tabulka 4.1: Kontingenční tabulka 2 x 9 s četnostmi N_{ij}

Hladinu χ^2 testu volíme $\alpha = 5$ %. Dále uvádíme hypotézu testovanou proti alternativě.

H_0 : Přítomnost diváků neovlivňuje výsledek domácího týmu v zápasu (veličiny X, Y jsou nezávislé).
H_1 : Přítomnost diváků ovlivňuje výsledek domácího týmu (veličiny X, Y nejsou nezávislé).

Testová statistika Q ze vzorce (4.2) má při $R = 2$ a $C = 9$ (za platnosti hypotézy) asymptoticky χ^2 rozdělení o 8 stupních volnosti. Hypotézu zamítáme právě tehdy, když

$$Q \geq \chi^2_{(8)}(0{,}95). \tag{4.4}$$

Test provádíme v programovacím jazyce R. Komentovaný kód z programu Rstudio je v příloze A.

V tabulce 4.2 uvádíme základní výstupní hodnoty z provedeného testu χ^2 testu nezávislosti, provedeného v programovacím jazyce R.

Q	$\chi^2_{(8)}(0{,}95)$	p-hodnota
12,275	15,507	0,139

Tabulka 4.2: Výsledné hodnoty Q, příslušného kvantilu a p-hodnoty χ^2 testu nezávislosti

V tabulce 4.2 vidíme, že na základě kritického oboru ze vzorce (4.4) i na základě věty 1 (při $\alpha = 0{,}05$) hypotézu H_0 nemůžeme zamítnout.

Nyní interpretujeme výsledek χ^2 testu.

Očekávali jsme, že fanoušci domácího týmu budou mít na výsledek zápasu (z pohledu domácího týmu) vliv. V tabulce 4.2 vidíme, že p-hodnota je větší než 0,05. Hypotézu tedy nemůžeme zamítnout. Nemůžeme vyloučit, že při testování více dat, popřípadě jiné soutěže, například Premier League, v níž je návštěvnost větší než v českých soutěžích, bychom hypotézu zamítali.

Důvodem, proč tomu tak může být, je třeba, že každý hráč na počet fanoušků reaguje jinak (někteří jsou nervózní se zvyšujícím se počtem fanoušků, jiní tuto

změnu nevnímají a další jsou větším počtem fanoušků motivováni). To může vést k tomu, že týmy budou na nízký i vysoký počet fanoušků reagovat podobně.

Závěr

Cílem této bakalářské práce bylo vytvořit sbírku řešených statistických úloh vhodných pro střední školy se sportovní tématikou doplněnou o teorii, která je při jejich řešení nepostradatelná. V první kapitole jsme využili grafů a tabulek a na jejich základě zadávali a řešeili úlohy. V první sérii úloh jsme využili fotbalu, ve druhé kapitole hokeje a třetí kapitola obsahuje úlohy, které se týkají různých sportů. Postupně jsme si představili jednotlivá statistická odvětví, která se vyučují na střední škole. Postupovali jsme od základních termínů, jako jsou statistická jednotka či soubor, k pojmům složitějším, jako je směrodatná odchylka či koeficient korelace. V práci jsou představeny a využity v řešení úloh termíny, které se týkají statistiky tak, jak je běžně vykládána na střední škole. Tento cíl práce byl splněn.

Dalším z cílů této práce bylo představit pojmy spadající do statistiky učené na vysokých školách a ukázat příklady úloh se sportovní tématikou, které jsou založené na znalosti právě vysokoškolské matematiky a statistiky. Ve druhé a třetí kapitole jsou vyřešeny vždy dvě takové úlohy. Cíl představit statisiku vyučovanou na vysoké škole byl také splněn.

Posledním cílem bylo ukázat, jak využíváme statistiky při práci s daty v profesionálním sportu. Ve čtvrté kapitole jsme zavedli obecnou teorii k testování hypotéz, následně jsme definovali konkrétní χ^2 test nezávislosti a použili ho k testování jedné konkrétní hypotézy proti alternativě. Test jsme konstruovali pomocí programovacího jazyku R. I poslední cíl bakalářské práce byl tedy splněn.

Tato bakalářská práce může být využita jako sbírka úloh pro učitele i žáky s neobvyklými řešenými úlohami, materiál pro samostudium či jako příručka pro učitele.

Literatura

Calda, E. a Dupač, C. (1993). *Matematika pro gymnázia. Kombinatorika, pravděpodobnost, statistika.* Učebnice pro střední školy. Prometheus, Praha, 4., dotisk., vyd. edition. ISBN 80-7196-365-3.

DeGroot, M. H. (1975). *Probability and statistics.* Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company, Reading, Massachusetts ;. ISBN 0-201-01503-X.

Dupač, V. a Hušková, M. (1999). *Pravděpodobnost a matematická statistika.* Karolinum, Praha, 1. vydání edition. ISBN 80-246-0009-9.

Hindsl, R., Arltová, M., Hronová, S., Malá, I., Marek, L., Pecáková, I. a Řezaniková, H. (2018). *Statistika v ekonomii.* Učebnice pro střední školy. Professional Publishing, Průhonice, 1., vyd. edition. ISBN 80-88260-09-7.

Kulich, M. (2014). Přehledový větník. `https://www2.karlin.mff.cuni.cz/~pesta/NMFM301/statistika_fm.pdf`. Souhrn pro předmět statistika pro finanční matematiky.

Kulich, M. (2018). Základy teorie pravděpodobnosti. `https://www2.karlin.mff.cuni.cz/~zichova/FinMat/PravdepodobnostKulich.pdf`. Souhrn pro předmět Matematická statistika 1.

Příloha A

Kód s χ^2 testem nezávislosti

Zde uvádíme kód v programovacím jazyce R, ve kterém jsme naprogramovali ukázku χ^2 testu nezávislosti v kompilačním prostředí Rstudio. Vstupní data z kontingenční tabulky 4.1 ukládáme do proměnné *ft_data* ve struktuře *data.frame*. Ta umožňuje do sloupců uspořádat v proměnných *goal_dif* kategorie různých gólových rozdílů z pohledu domácího týmu, *freq_att* četnosti gólových rozdílů při návštěvě bez omezení a *freq_no_att* četnosti gólových rozdílů při omezené návštěvnosti.

```
ft_data <- data.frame(goal_dif = c("leq 4", -3:3, "geq 4"),
                      freq_att = c(8,9,19,35,46,60,31,18,14),
                      freq_no_att = c(6,12,33,48,82,49,38,26,12))
```

Poté vytvoříme matici *cetm* rozměru 2 x 9, kde v prvním řádku jsou četnosti v proměnné *freq_att* a v druhém četnosti v proměnné *freq_no_att*. Tato matice pozorovaných četností je jeden ze dvou nutných vstupů funkce *chisq.test* provádějící χ^2 test nezávislosti.

```
cetm = matrix(c(ft_data$freq_att, ft_data$freq_no_att),
              nrow = 2, ncol = 9, byrow = TRUE)
```

Druhý vstup je vektor teoretických pravděpodobností jevů $[X = i, Y = j]$, $i = 1, 2; j = 1, \ldots, 9$. Tyto pravděpodobnosti získáme z teoretických četností ve vzorci (4.3), a to vydělením celkovým součtem všech četností n. Potřebujeme tedy určit součty četností v řádcích a sloupcích podle vzorců v (4.1). V proměnné *sec* je uložen vektor devíti prvků udávající součty ve sloupcích a v proměnné *fir* je uložen vektor 2 prvků jako součtů v řádcích. Celkový součet n pak vypočítáme

sčítací funkcí *sum* a můžeme tak vytvořit v proměnné *psts* vektor teoretických pravděpodobností dle postupu výše. Bylo ověřeno sečtením všech pravděpodobností, že dávají 1 a že jsou tedy správně zavedeny.

```
sec = c(ft_data$freq_att+ft_data$freq_no_att)

fir = c(sum(ft_data$freq_att),sum(ft_data$freq_no_att))

psts = c((fir[1]*sec)/(sum(fir))^2,(fir[2]*sec)/(sum(fir))^2)
```

Dále je potřeba ověřit pro potřeby testu, že každý z 18 možných jevů v kontingenční tabulce dosahuje alespoň teoretické četnosti 5. Ty dostaneme vynásobením proměnné *psts* celkovým součtem n. Zde je uveden i výstup z naprogramovaného kódu (četnosti jsou zaokrouhleny na 3 desetinná místa). Je tak možné vidět, že podmínka o teoretických četnostech je zde splněna.

```
psts*sum(fir)

6.154;  9.231; 22.857; 36.484; 56.264; 47.912; 30.330; 19.341; 11.429;
7.846; 11.769; 29.143; 46.516; 71.736; 61.088; 38.670; 24.659; 14.571
```

Můžeme tak zadat χ^2 testu nezávislosti funkcí *chisq.test* se vstupy *cetm* a *psts*.

```
chisq.test(cetm, p = psts)
```

Výstup obsahující hodnotu testové statistiky Q, počet stupňů volnosti asymptotického χ^2 rozdělení a p-hodnotu testu je prezentován v tabulce 4.2, kde je doplněn o hodnotu příslušného 95% kvantilu daného rozdělení.