UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

EXPERIMENTY A VLASTNÉ DÁTA PRI VYUČOVANÍ PRAVDEPODOBNOSTI A ŠTATISTIKY

BAKALÁRSKA PRÁCA

2015 Katarína KOCSISOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

EXPERIMENTY A VLASTNÉ DÁTA PRI VYUČOVANÍ PRAVDEPODOBNOSTI A ŠTATISTIKY

BAKALÁRSKA PRÁCA

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	9.1.9. Aplikovaná matematika (1114)
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Beáta Stehlíková, PhD.

Bratislava 2015 **Katarína Kocsisová**

33055394

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Katarína Kocsisová
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Experimenty a vlastné dáta pri vyučovaní pravdepodobnosti a štatistiky
Experiments and own data in teaching probability and statistics

Cieľ: Cieľom práce je spracovať "pokusy", ktoré sa dajú využiť pri vyučovaní pravdepodobnosti a štatistiky: experimenty motivujúce výpočty určitých pravdepodobností, získanie vlastných dát na testovanie hypotéz a pod. Konkrétne, práca sa bude zaoberať nasledovnými témami, ku ktorým sa pridá niekoľko podobných podľa vlastného výberu študenta:
(a) Vizualizácia distribučnej funkcie pomocou slovníka a testovanie zhody dvoch distribučných funkcií [1]
(b) Odhad veľkosti populácie (základná myšlienka: máme N guľôčok očíslovaných 1, 2, ..., N, vytiahneme n z nich a chceme na základe tohto výberu odhadnúť N) [2], [3]
(c) Hádzanie šípok a testovanie vplyvu dominantnej ruky [4]
(d) Hra s kockami Hog [5]
(e) Chí kvadrát test dobrej zhody a farby M&M's [6]
(f) Základné princípy testovania štatistických hypotéz a rozlišovanie Pepsi a Coke [7].
Súčasťou každej témy bude jej vysvetlenie, zrealizovanie pokusov a následný výpočet pravdepodobnosti, výsledku štatistického testovania a pod. Vybraná téma sa pripraví na praktické použitie pri výučbe a študent ju zrealizuje na predmete Metódy riešenia úloh z pravdepodobnosti a štatistiky.

Literatúra:
[1] Jernigan, R. W. (2008). A photographic view of cumulative distribution functions. Journal of Statistics, Education Volume, 16.
[2] Vännman, K. (1983). How to Convince a Student that an Estimator is a Random Variable. Teaching Statistics, 5 (2), 49-54.
[3] Johnson, R. W. (1994). Estimating the Size of a Population. Teaching Statistics, 16(2), 50-52.
[4] Nordmoe, E. (1998). Lawn Toss: Producing Data On-the-Fly. Teaching Statistics, 20, 66-67.
[5] Feldman, L., & Morgan, F. (2003). The pedagogy and probability of the dice game hog. Journal of Statistics Education, 11(2).
[6] Johnson, R. W. (1993). Testing colour proportions of M&Ms. Teaching Statistics, 15(1), 2-4.
[7] Levine, M., & Rolwing, R. H. (1993). Coke or Pepsi?. Teaching Statistics, 15(1), 4-5.
Ďalšia literatúra podľa vlastného výberu.

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Vedúci: RNDr. Beáta Stehlíková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 23.10.2014

Dátum schválenia: 23.11.2014

doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Úprimná vďaka patrí vedúcej bakalárskej práce doc. RNDr. Beáte Stehlíkovej, PhD. za odborné rady, ochotu a pomoc, ktoré mi pri písaní pomohli. Takisto ďakujem aj svojej rodine, kamarátom a známym za podporu a zapájanie do experimentov, bez ktorých by táto práca nemohla vzniknúť.

Abstrakt

KOCSISOVÁ, Katarína: Experimenty a vlastné dáta na vyučovanie pravdepodobnosti a štatistiky [Bakalárska práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2015, 53s.

Cieľom práce je zrozumiteľne spracovať pokusy použiteľné pri výučbe pravdepodobnosti a štatistiky. Zaoberáme sa v nej nielen experimentami na testovanie štatistických hypotéz, ale aj príkladmi motivujúcimi k výpočtu pravdepodobností, hľadaniu distribučných funkcií v bežnom živote či odhadovaniu celkového množstva súboru na základe malého počtu známych dát. Každá téma obsahuje vysvetlenie, realizáciu pokusu, príslušný výpočet a záver z experimentu.

Kľúčové slová: testovanie hypotéz, distribučná funkcia, Kolmogorovov-Smirnovov test, chí kvadrát test dobrej zhody, párový t-test, odhad

Abstract

KOCSISOVÁ, Katarína: Experiments and own data in teaching probability and statistics [Bachelor Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2015, 53p.

The aim of this thesis is to comprehensibly carry out experiments useable for teaching probability and statistics. The thesis is not only dealing with the experiments of statistical hypotheses testing, but also with mathematical problems motivating to probability calculations, searching distribution functions in real life and estimating the total population size, based on a small amount of the known data. Each thesis topic consists of an explanation, realization of the experiment, its respective calculation and conclusion of the experiment.

Keywords: hypothesis testing, distribution function, Kolmogorov-Smirnov test, Pearson's chi-squared test, paired t-test, estimator

Obsah

Úvod

Štatistika ako vedný odbor je čoraz viac populárna. Využíva sa asi vo všetkých odvetviach ľudskej činnosti, čoho dôkazom je jej výučba i na školách s iným než matematickým, ekonomickým či informatickým zameraním. Mnohokrát si však študenti nevedia celkom predstaviť aplikáciu naučených poznatkov v reálnom živote.

Experimentom alebo pokusom rozumieme zámerne vyvolanú situáciu s úmyslom sledovania určitých javov za špecifických podmienok. Naším cieľom je spracovať a podrobne vysvetliť takéto experimenty. Práca obsahuje prierez celým testovaním krok za krokom, od stanovenia množstva potrebných dát až po závery plynúce z testovania. V každej kapitole sa budeme venovať inému pokusu. Testovanie bude predvedené na nami zozbieraných dátach.

V 1. kapitole spracujeme teóriu ohľadom testovania štatistických hypotéz, uvedieme pojmy ako hypotéza H_0 a H_1, chyba I. a II. druhu. Predstavíme si aj Neymann-Pearsonovu lemu a ukážeme si ekvivalenciu jej zamietacieho kritéria s kritériom vyplývajúcim z chyby I. a II. druhu. Vychádzajúc z článku [5] overíme funkčnosť chuťových pohárikov ochotných degustátorov. Namiesto známych nealkoholických nápojov Coca Coly a Pepsi však budeme skúmať rozdiel medzi cukríkmi M&M's a Lentilkami.

Ďalšia kapitola bude venovaná distribučnej funkcii a jej vysvetleniu prostedníctvom slovníka sledujúc článok [3]. Následne budeme testovať zhodu medzi nimi v dvoch anglických slovníkoch za pomoci dvojvýberového Kolmogorovovho-Smirnovovho testu.

Kapitola porovnávajúca distribučné funkcie doby čakania v dvoch študentských jedálňach Eat and Meet a Venza situovaných v Mlynskej doline bude akýmsi doplnením predchádzajúcej kapitoly.

Vo 4. kapitole si predstavíme pravidlá dvoch hier s kockami. Ako prvú rozoberieme hru HOG z článku [7]. Vypočítame nielen očakávané výplaty pre hry s rôznymi počtami kociek, ale aj obmeny tejto hry s uvedením optimálnej stratégie. V druhej časti sa pozrieme na nami vymyslenú hru. Oslovíme niekoľkých hráčov, aby sme zistili, či sa správajú optimálne.

V ďalšej kapitole spracujeme dáta o obľúbených čokoládových cukríkoch M&M's. Budeme porovnávať výsledky nášho merania so starším vyjadrením výrobcu v článku [1] i s novšími pomermi farieb spomenutými v [2]. Na overenie hypotézy s pravdepo-

dobnosťami rozdelení farieb cukríkov použijeme chí kvadrát test dobrej zhody.

Kapitola 6 bude obsahovať experiment, pri ktorom sa bude hádzať šípkami na terč. Účastníci pokusu sa budú snažiť trafiť šípku čo najbližšie k stredu najskôr dominantnou, následne predvedú hody nedominantnou rukou. Podľa článku [6] by sa tieto hody nemali od seba významne líšiť. Na overenie použijeme párový t-test.

Posledná kapitola bude zameraná na odhady. Uvedieme si niekoľko z nich z článkov [22] a [25], ukážeme, že sú nevychýlené alebo vychýlené. Následne predstavíme odhady získané od študentov navštevujúcich predmet Metódy riešenia úloh z pravdepodobnosti a štatistiky.

V celej práci budeme používať predovšetkým základné poznatky z pravdepodobnosti a štatistiky spracované v [8] a [9].

1 Základné princípy testovania štatistických hypotéz a rozlišovanie M&M's a Lentiliek

Mnohí ľudia tvrdia, že medzi Coca Colou a Pepsi Colou nie je žiaden rozdiel. Iní zas vyhlasujú, že dokážu spomínané nápoje rozlíšiť. Článok [5] sa zaoberá testovaním tvrdenia z predchádzajúcej vety. Autori v ňom vysvetľujú základné štatistické princípy zaujímavou formou na stávke. Stávka sa odohráva medzi učiteľom a študentom, kde študent tvrdí, že dokáže tieto dva nealkoholické nápoje rozlíšiť s určitou pravdepodobnosťou. V hre je symbolická čiastka peňazí, ktoré stavili obe zúčastnené strany.

V tejto práci sa budeme venovať mierne upravenej verzii daného experimentu. Testovať budeme, či dobrovoľník na ochutnávanie cukríkov vie rozlíšiť Lentilky a M&M's s konkrétnou pravdepodobnosťou p, ktorú označí za charakteristiku svojich schopností. V prípade, ak sa bude často mýliť, bude to prakticky to isté ako hádanie, teda určovanie cukríkov s pravdepodobnosťou $p = 0,5$.

1.1 Teória

V nasledujúcej časti si uvedieme vysvetlenia pojmov z kníh [9], [14] a [18].

Hypotézou nazývame tvrdenie týkajúce sa rozdelenia pravdepodobnosti náhodných premenných alebo hodnôt parametrov týchto rozdelení. Overovanie správnosti príslušného tvrdenia nazývame testovaním štatistických hypotéz. V štatistike formulujeme nulovú hypotézu H_0 a alternatívnu hypotézu H_1, ktorá je obvykle negáciou nulovej. Pri testovaní môžu nastať dve možnosti, buď nulovú hypotézu zamietneme alebo ju nezamietneme. Keďže sa rozhodujeme vzhľadom na realizáciu náhodného vektora, nemôžeme zaručiť bezchybné rozhodnutie.

Pri rozhodovaní sa môžeme dopustiť chyby I. alebo II. druhu. Ak zamietame hypotézu H_0 napriek tomu, že platí, ide o chybu I. druhu. Pravdepodobnosť tejto chyby nazývame hladinou významnosti a označujeme α, pričom $\alpha \in (0, 1)$. V prípade, že nezamietame nulovú hypotézu, aj keď neplatí, dopúšťame sa chyby II. druhu a jej pravdepodobnosť označíme β. Hodnota $1 - \beta$ predstavuje pravdepodobnosť správneho zamietnutia nulovej hypotézy, teda ak neplatí. Nazývame ju sila testu.

Prirodzene chceme, aby pravdepodobnosti oboch druhov chýb boli malé. Podľa [19]

a [20] zvyčajne fixujeme hodnotu α a hľadáme test, ktorý minimalizuje β. Takýto test voláme najsilnejší test na hladine α a hovorí o ňom nasledujúca lema z [19].

Lema 1.1 (Neymann-Pearsonova lema). *Nech $X_1, X_2, \ldots, X_n$ je náhodný výber z rozdelenia s parametrom θ a nech $L(\theta)$ je funkcia vierohodnosti prislúchajúca tomuto výberu. Zoberme kritickú hodnotu c takú, aby platilo $P(\Lambda \leq c|H_0) = \alpha$, kde $\Lambda = \frac{L(\theta_0)}{L(\theta_1)}$. Potom test, ktorý zamieta $H_0 : \theta = \theta_0$ v prospech $H_1 : \theta = \theta_1$ v prípade, že $\Lambda \leq c$ je najsilnejším testom na hladine α.*

V nasledujúcej časti budeme postupovať podľa príkladu uvedeného v [19, str. 25], v ktorom je ukázané, že pre test $H_0 : p = p_0$ a $H_1 : p = p_1$, kde $p_1 > p_0$, sú kritéria zamietnutia $\Lambda \leq c$ a $Y \geq k$ ekvivalentné. V našom prípade však bude $p_0 > p_1$.

Majme $X_1, X_2, \ldots, X_n$, ktoré sú nezávislé rovnako rozdelené z alternatívneho rozdelenia s parametrom p a ich súčet označme Y. Uvažujme test s hypotézami $H_0 : p = p_0$ a $H_1 : p = p_1$, kde $p_0 > p_1$. Keďže funkcia vierohodnosti pre alternatívne rozdelenie má tvar

$$L(p) = p^Y (1-p)^{n-Y},$$

testovacia štatistika vyzerá nasledovne:

$$\Lambda = \frac{L(p_0)}{L(p_1)} = \frac{p_0^Y (1-p_0)^{n-Y}}{p_1^Y (1-p_1)^{n-Y}} = \left(\frac{p_0}{p_1}\right)^Y \left(\frac{1-p_0}{1-p_1}\right)^{n-Y} = \left(\frac{1-p_0}{1-p_1}\right)^n \left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right)^Y.$$

Ukážeme, že test zamietajúci H_0 v prospech H_1 pri $\Lambda \leq c$ je ekvivalentný s takým, čo zamieta H_0 pre $Y \leq k$, kde c a k spĺňajú vzťah $P(\Lambda \leq c|H_0) = P(Y \leq k|H_0) = \alpha$.

$$\begin{aligned}
\Lambda \leq c &\iff \log(\Lambda) \leq \log(c) \\
&\iff Y \log\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right) + n \log\left(\frac{1-p_0}{1-p_1}\right) \leq \log(c) \\
&\iff Y \log\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right) \leq \log(c) - n \log\left(\frac{1-p_0}{1-p_1}\right) \\
&\iff Y \leq k,
\end{aligned}$$

kde $k = \left(\log(c) - n \log\left(\frac{1-p_0}{1-p_1}\right)\right) / \log\left(\frac{p_0(1-p_1)}{p_1(1-p_0)}\right)$. Podľa Neymann-Pearsonovej lemy je teda test založený na Y najsilnejší.

1.2 Testovanie

Predstavme si situáciu, že nám niekto povie, že dokáže rozoznať Lentilky a M&M's podľa chuti, teda ich vie určiť s pravdepodobnosťou $p > 0,5$. Jeho tvrdenie by sme chceli porovnať s tvrdením, že cukríky rozlišuje s pravdepodobnosťou $p = 0,5$. Experiment prebehne so šatkou previazanou cez oči, pretože cukríky M&M's sú sýtejšej farby s nápisom „m“. Keďže majú aj viacej plnky, sú hrubšie, vložia sa degustátorovi do úst pomocou lyžičky. Po každej ochutnávke musí povedať, ktorý mal ako posledný, pričom má možnosť sa medzi jednotlivými ochutnávkami napiť čistej vody na neutralizáciu chute. V prípade súboru dát pozostávajúceho iba z jedného cukríku by správne určenie nemalo výpovednú hodnotu. Výstupom každého z n pokusov je buď úspešné určenie s pravdepodobnosťou p alebo neúspech s pravdepodobnosťou $1 - p$. Pokusy sú všetky rovnaké a nezávislé. Podľa knihy [17] v prípade n opakovaní pokusu je možné zapísať pravdepodobnosť k úspešných označení pomocou Bernoulliho schémy:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

pre $k = 0, \ldots, n$.

Obr. 1: Predmety potrebné na experiment

Budeme testovať na hladine významnosti $\alpha = 0,05$ a postupovať podľa článku [5]. Naša hypotéza H_0 teda hovorí, že ochutnávač vie rozoznať cukríky a hypotéza H_1 tvrdí, že iba tipuje. Následne vyvstáva otázka, s akou pravdepodobnosťou ich určí. Keďže náš degustátor rozmýšľa logicky, nechá si istú rezervu a nebude tvrdiť, že ich rozlíši so 100% istotou. Test by to totiž veľmi jednoducho zamietol pre všetky úspešné pokusy okrem prípadu, keď $n = k$. Ochutnávač tvrdí, že správne označí aspoň 85% cukríkov.

Minimálnu hodnotu k, pre ktorú nulovú hypotézu nezamietneme, vyberáme tak, aby pravdepodobnosť chyby I. druhu nepresiahla číslo α. Asi prvým nápadom pri hľadaní kritického oboru je vyčíslenie pravdepodobnosti pre $k = np$. V prípade, že mu dáme ochutnať 20 cukríkov, 85% z toho je 17. Teda ak by mal bezchybne označiť aspoň 17 cukríkov, pravdepodobnosť, že jeho tvrdenie je pravdivé a test hypotézu nezamietne, vypočítame ako:

$$\sum_{k=17}^{20} \binom{20}{k} 0,85^k (1-0,85)^{20-k} = 0,648.$$

Z toho jednoducho vyčíslime pravdepodobnosť chyby I. druhu, čiže $1 - 0,648 = 0,352$. Keďže toto číslo je oveľa väčšie než hladina významnosti, musíme zobrať menšie k. Ako vidíme v tabuľke 1, chyba I. druhu pre prípad, keď $k = 14$, je rovná

$$1 - \sum_{k=14}^{20} \binom{20}{k} 0,85^k (1-0,85)^{20-k} = 0,022,$$

teda menšia než α, čo sme aj chceli. Pre $k = 15$ je už chyba väčšia od hladiny významnosti, preto budeme zamietať už pri $k = 14$.

Tabuľka 1: Chyby I. druhu vyčíslené pre vybrané k

k	chyba I. druhu
13	0,006
14	0,022
15	0,067

Na zistenie sily testu, potrebujeme najprv vypočítať chybu II. druhu. Hľadáme teda pravdepodobnosť, že test hypotézu nezamietne, pričom neplatí. Chyba II. druhu je rovná

$$\sum_{k=14}^{20} \binom{20}{k} 0,5^{20} = 0,058.$$

Z toho vyplýva, že sila testu je $1 - 0,058 = 0,942$, čiže s touto pravdepodobnosťou dôjde k odhaleniu neplatnosti nulovej hypotézy.

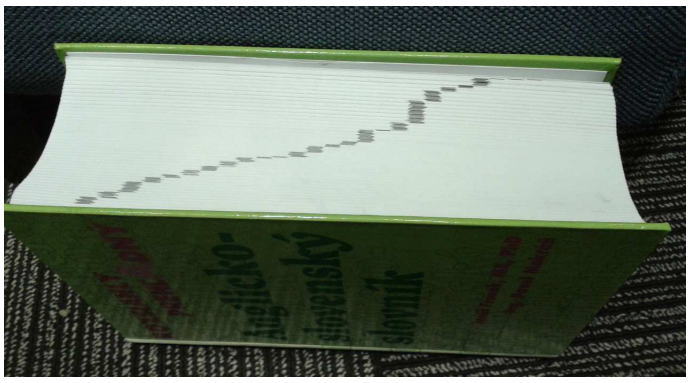
Pred testovaním sme si pripravili do nádoby rovnaké množstvá z oboch druhov cukríkov. Náhodne sme vytiahli jeden a dali ochutnať degustátorovi. Následne sme doplnili cukrík vytiahnutej značky. Napriek tomu, že si degustátor zahryzol počas experimentu do líca, podarilo sa mu správne určiť 18 cukríkov. Test teda hypotézu H_0 nezamietol.

Autori v článku [5], ktorý bol inšpiráciou pre vznik tohto experimentu, uvádzajú, že z mnohých testovaní o rozlišovaní Coca Coly a Pepsi Coly zamietli nulovú hypotézu iba v dvoch prípadoch. Raz sa dokonca stalo, že si študent ráno pred experimentom popálil jazyk pri pití kávy. Výsledok tohto testovania, žiaľ, z článku známy nie je.

2 Vizualizácia distribučnej funkcie pomocou slovníka a testovanie zhody dvoch distribučných funkcií

V tejto kapitole sa budeme venovať slovníkom podľa vzoru článku [3]. Konkrétnym objektom nášho záujmu budú tie, ktoré majú „záhadné obrazce“ viditeľné pri pohľade zboku.

Na obrázku 2 je odfotografovaný slovník [12] z bočnej strany, na ktorej sú vytvorené tmavé plôšky. Tieto plôšky reprezentujú časti knihy venované jednotlivým písmenám. Idú od vrchu strany pre písmená zo začiatku abecedy až po spodok strany pre tie, ktoré sú na konci. Umožňujú efektívnejšie vyhľadávanie slov a definícií.



Obr. 2: Bočná strana slovníka [12]

2.1 Teória

V nasledujúcej časti uvedieme definície spracované v knihe [8].

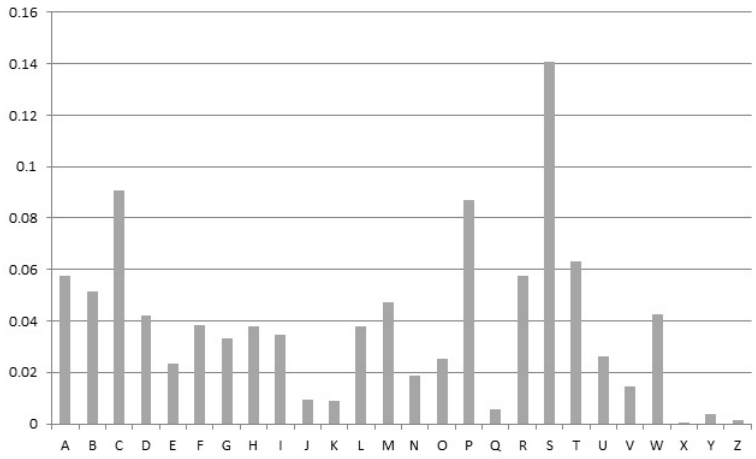
Definícia 2.1. *Nech Ω je neprázdna množina elementárnych udalostí. Náhodná veličina je ľubovoľná funkcia $X : \Omega \to R$ taká, že pre každé*

$$x \in R : \{\omega \in \Omega : X(\omega) < x\} \in S.$$

Definícia 2.2. *Distribučná funkcia náhodnej veličiny $X : \Omega \to R$ je funkcia $F : R \to < 0, 1 >$ definovaná predpisom $F(x) = P(X < x)$.*

2.2 Testovanie

Zamerajme sa na slovník [12] postupujúc podľa článku [3]. V tabuľke 2 sú spracované počty strán i percentuálne zastúpenie pre jednotlivé písmená abecedy. Pozrime sa napríklad na písmeno „A“. V slovníku je preň vyčlenených 79 strán na boku označených štítkom „A“. Posledná strana so štítkom ľubovoľného písmena udáva počet strán venovaných slovám, ktoré začínajú s písmenami nachádzajúcimi sa v abecednom poradí pred týmto štítkom. Nech $G(x)$ označuje číslo strany pre každé písmeno $x \in \{A, B, \ldots, Z\}$. Maximum M funkcie $G(x)$ je číslo poslednej strany slovníka pri písmene „Z“. Definujme $F(x)$ ako $G(x)/M$, $F(x)$ potom reprezentuje kumulatívnu relatívnu početnosť písmen abecedy, čo je distribučná funkcia abecedy.



Obr. 3: Relatívne početnosti jednotlivých písmen slovníka [12]

Z obrázku 3 a tabuľky 2 pozorujeme, že písmená na začiatku abecedy sú obsiahnuté na relatívne veľkom počte strán. Písmená „J“ a „K“ zo stredu abecedy nemajú v tomto slovníku výrazné zastúpenie. Na druhej strane, pri písmenách „P“ a „S“ je zreteľný vzrast prislúchajúcich stĺpikov oproti ostatným. Koncové písmená abecedy, teda „Q“, „X“, „Y“ a „Z“, sa vyskytujú na začiatku slov pomerne zriedkavo.

Ako vidíme z obrázku 3, písmenu „S“ je venovaných najviac strán. Počet strán prislúchajúcich tomuto písmenu vypočítame jednoducho využitím informácií z tabuľky 2, čiže

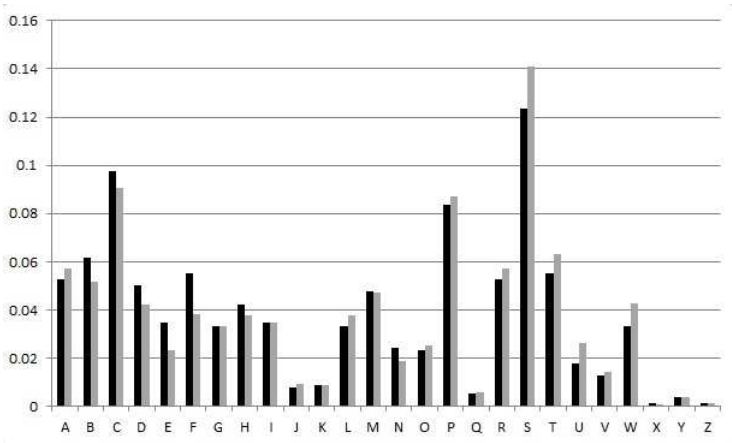
$$G(S) - G(R) = 1168 - 974 = 194.$$

Aby sme získali relatívnu početnosť, musíme predeliť rozdiel maximom M = 1378, teda

194/1378 = 0,141. Početnosť i relatívna početnosť pre jednotlivé písmená sú zobrazené vo vyššie spomínanej tabuľke.

V ďalšej časti tejto kapitoly overíme zhodu distribučných funkcií dvoch rôznych anglických slovníkov ([12], [13]). Testovanie urobíme pomocou Kolmogorovovho-Smirnovovho testu a budeme postupovať ako v článku [3].

Najprv uvedieme charakteristiku testu z knihy [14]. Nech $X_1, \ldots, X_m$ je náhodný výber z rozdelenia so spojitou distribučnou funkciou F a nech $Y_1, \ldots, Y_n$ je od neho nezávislý náhodný výber so spojitou distribučnou funkciou G. Budeme sa zaoberať testom hypotézy $H_0 : F = G$ proti alternatíve $H_1 : F \neq G$. Empirickú funkciu prvého výberu označíme F_m a druhého výberu G_n. Funkcie F_m a G_n sa pri rastúcich m a n blížia k skutočným distribučným funkciám F, G (vyplýva z viet uvedených v [14, str. 241]). Hypotézu H_0 na hladine významnosti α zamietame, ak $D_{m,n} \geq D_{m,n}(\alpha)$, kde $D_{m,n} = \sup_x |F_m(x) - G_n(x)|$. Pre veľké m a n $(m + n > 35)$ môžeme kritickú hodnotu aproximovať pomocou vzorca: $D_{m,n}(\alpha) = \sqrt{\frac{1}{2M} \ln \frac{2}{\alpha}}$, pričom $M = \frac{mn}{m+n}$.



Obr. 4: Porovnanie relatívnych početností slovníkov [12] (sivou farbou) a [13] (čiernou farbou)

Z obrázku 4 sa nám zdá, že jednotlivé početnosti vyzerajú podobne, preto sme sa rozhodli to štatisticky otestovať postupom z [3]. Pri testovaní predpokladáme, že na stranách je približne rovnaký počet slovníkových hesiel, teda počítame strany a nie heslá. Množina Ω predstavuje strany v slovníku, funkcia F priradí písmenu číslo podľa schémy: $1 = A, 2 = B, \ldots, 26 = Z$. Následne sme vypočítali empirické distribučné funkcie F_m a G_n a výsledky zapísali do tabuľky 3, $F_m(x)$ je distribučná funkcia slovníka

Tabuľka 2: Rozsiahly prehľadný Anglicko-slovenský slovník

písmeno	kumulatívna početnosť	kumulatívna relatívna početnosť	početnosť	relatívna početnosť
A	79	0,057	79	0,057
B	150	0,109	71	0,052
C	275	0,200	125	0,091
D	333	0,242	58	0,042
E	365	0,265	32	0,023
F	418	0,303	53	0,038
G	464	0,337	46	0,033
H	516	0,374	52	0,038
I	564	0,409	48	0,035
J	577	0,419	13	0,009
K	589	0,427	12	0,009
L	641	0,465	52	0,038
M	706	0,512	65	0,047
N	732	0,531	26	0,019
O	767	0,557	35	0,025
P	887	0,644	120	0,087
Q	895	0,649	8	0,006
R	974	0,707	79	0,057
S	1168	0,848	194	0,141
T	1255	0,911	87	0,063
U	1291	0,937	36	0,026
V	1311	0,951	20	0,015
W	1370	0,994	59	0,043
X	1371	0,995	1	0,001
Y	1376	0,999	5	0,004
Z	1378	1,000	2	0,001

Tabuľka 3: Porovnanie dvoch anglických slovníkov

písmeno	kumulatívna početnosť [12]	kumulatívna relatívna početnosť	kumulatívna početnosť [13]	kumulatívna relatívna početnosť	rozdiel kum. relatívnych početností
A	79	0,057	41	0,053	0,004
B	150	0,109	89	0,115	0,006
C	275	0,200	165	0,212	0,012
D	333	0,242	204	0,263	0,021
E	365	0,265	231	0,297	0,032
F	418	0,303	274	0,353	0,050
G	464	0,337	300	0,386	0,049
H	516	0,374	333	0,429	0,055
I	564	0,409	360	0,463	0,054
J	577	0,419	366	0,471	0,052
K	589	0,427	373	0,480	0,053
L	641	0,465	399	0,514	0,049
M	706	0,512	436	0,561	0,049
N	732	0,531	455	0,586	0,055
O	767	0,557	473	0,609	0,052
P	887	0,644	538	0,692	0,048
Q	895	0,649	542	0,698	0,049
R	974	0,707	583	0,750	0,043
S	1168	0,848	679	0,874	0,026
T	1255	0,911	722	0,929	0,018
U	1291	0,937	736	0,947	0,010
V	1311	0,951	746	0,960	0,009
W	1370	0,994	772	0,994	0,000
X	1371	0,995	773	0,995	0,000
Y	1376	0,999	776	0,999	0,000
Z	1378	1,000	777	1,000	0,000

[12] a $G_n(x)$ prislúcha slovníku [13]. Keďže sú výberové distribučné funkcie F_m a G_n obe konštantné na intervaloch $(-\infty, 1), (1, 2), \ldots, (26, \infty)$, hodnotu $D_{m,n}$ vieme nájsť v tabuľke 3. Hypotézu H_0 otestujeme na hladine významnosti $\alpha = 0,05$. Vypočítame kritickú hodnotu pre $m = n = 26$ a $M = \frac{26^2}{26+26} = 13$:

$$D_{m,n}(\alpha) = \sqrt{\frac{1}{26} \ln \frac{2}{0,05}} = 0,377.$$

Odchýlky pre jednotlivé písmená $A, B, \ldots, Z$ sú zapísané v poslednom stĺpci tabuľky 3. Nakoľko $D_{m,n} = \sup_x |F_m(x) - G_n(x)| = 0,055 < D_{m,n}(\alpha) = 0,377$, hypotézu H_0 nezamietame. Z toho vyplýva, že distribučné funkcie sa od seba významne nelíšia.

2.3 Komentáre

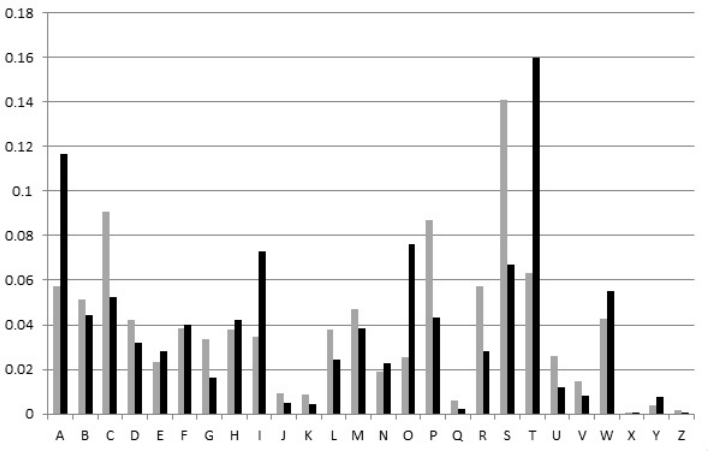
Pri použití Kolmogorovovho-Smirnovovho testu v predošlej podkapitole sa vyskytli dva problémy. Jedným z nich je, že slová zapísané v slovníku nie sú tak celkom náhodným výberom. Ide totiž o výber najčastejšie sa vyskytujúcich slov v danom jazyku. Ďalším problémom je, že Kolmogorovov-Smirnovov test predpokladá spojité distribučné funkcie. Distribučná funkcia slovníka je však diskrétna. Vhodnejší príklad na testovanie zhody dvoch distribučných funkcií uvedieme v ďalšej kapitole.

Text Petra Norviga [15] sa zaoberá analyzovaním slov v knihách naskenovaných spoločnosťou Google. Autor najskôr stiahol 23 GB textu. V nich vyhľadal slová, ktoré boli uvedené aspoň 100 000-krát, pričom nezáležalo na veľkosti písmen. V článku spracoval rôzne tabuľky pre výskyt slov, písmen a skupín písmen založených na dĺžke slov a mieste, kde sa nachádzajú. Uvedené sú aj počty pre písmená vyskytujúce sa na prvom mieste. Táto pozícia bola najviac obsadzovaná písmenami „T“, „A“, „O“, „I“ a „S“.

Ako sa v diskusii k článku uvádza, veľke množstvo z týchto textov je technického typu ako tabuľky a zoznamy, čo však nereprezentuje každodenne používaný jazyk. Ďalší účastník píše o tom, ako robil analýzy spracujúc viacero druhov textu a prišiel k podobným výsledkom. Podľa neho sú veľké rozdiely medzi bežným jazykom a slovnou zásobou používanou napríklad v spravodajstve spôsobené zhusťovaním textu či roku, v ktorom bol text napísaný. Lingvistka zapojená do diskusie by použila dané štatistiky na určenie vzorca, podľa ktorého sú tvorené zložené slovesá. Keby vedela, že je štatisticky viac slovies začínajúcich na „A“ než na „W“, musela by hľadať ďalšie na „A“, napríklad

v slangových a regionálnych slovníkoch.

V blogu na internetovej stránke firmy SAS, ktorého autorom je Rick Wicklin [16], sa analyzovali frekvencie písmen v textoch na základe [15]. Autor píše o možnosti využiť dané výsledky pri šifrovaní. V diskusii sa objavil názor ohľadom spojitosti medzi výskytom písmen a hodnotami písmen v spoločenskej hre Scrabble. Autor odpovedá, že v tejto hre sa dá z písmen vytvoriť ľubovoľné slovo nachádzajúce sa v slovníku. Avšak v písanej angličtine sa často používajú napríklad písmená „T“, „H“ a „E“, pretože je slovo „the“ veľmi bežné. Preto sa pomery písmen v slovníku, kde sa každé slovo objavuje raz, líšia od tých v textoch, kde sa bežné slová spomínajú niekoľkokrát. Táto odlišnosť je zobrazená na obrázku 5.



Obr. 5: Porovnanie relatívnych početností písmen zo slovníka [12] (sivou farbou) a naskenovaných textov podľa [15] (čiernou farbou)

3 Porovnávanie distribučných funkcií doby čakania v dvoch jedálňach

V kapitole 2 sme porovnávali zhodu dvoch distribučných funkcií na anglických slovníkoch. Keďže nebol splnený predpoklad Kolmogorovovho-Smirnovovho testu na náhodný výber a spojitú distribučnú funkciu, rozhodli sme sa pre testovanie vhodnejšieho príkladu. Myšlienky sa odvíjali od faktu, že funkcia závisiaca od času je spojitá. Zamýšľali sme sa nad rôznymi situáciami na porovnávanie. Prvým nápadom bolo overenie, či dĺžka čakania na autobus má rovnomerné rozdelenie. To sme však zamietli, keďže príchod na zastávku môže byť ovplyvnený napríklad koncom vyučovania alebo pracovnej doby. Hlavným impulzom pre vznik tohto experimentu však bolo uvažovanie nad porovnávaním času stráveného v ambulancii dvoch doktorov. Takáto téma však nie je mladým ľuďom až tak blízka, preto sme sa radšej rozhodli pozrieť na dĺžku čakania v rade v dvoch študentmi navštevovaných jedálňach v Mlynskej doline, Eat and Meet a Venza.

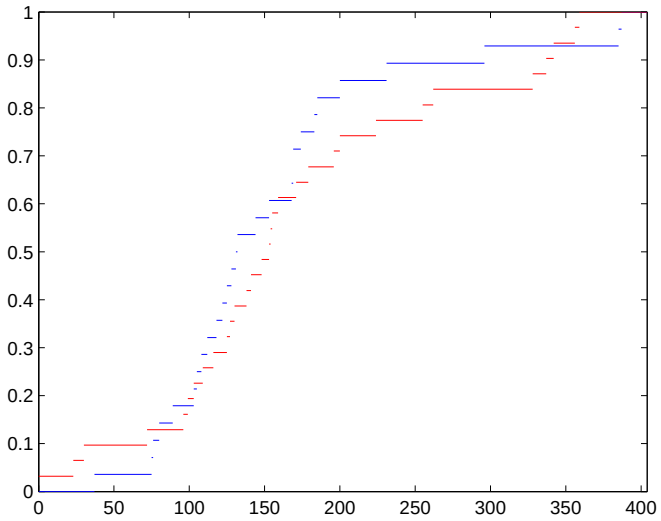
Oslovili sme niekoľko ľudí a požiadali ich, aby si počas jedného týždňa v čase obeda (11:00 - 14:00) merali na stopkách čas od momentu zaradenia sa do radu až po zaplatenie pri kase. V prípade, ak v jedálni rad utvorený nebol, časomiera sa spustila pri uchopení tácky či prvej súčasti príboru. Oslovením viacerých ľudí sme chceli dosiahnuť, aby namerané dáta boli nezávislé.

Hypotéza H_0 hovorí, že distribučné funkcie týchto jedální sú rovnaké. Zozbierané dáta sú zapísané v tabuľke 4, kde hodnoty v stĺpcoch 1. jedáleň a 2. jedáleň sú časy uvedené v sekundách. Namerané hodnoty sú zoradené od najmenšieho po najväčšie. Uvedené distribučné funkcie sú vykreslené na obrázku 6. Pri testovaní postupujeme analogicky ako v kapitole 2. Vyjadríme kumulatívne relatívne početnosti pre každé x, kde x je čas, pri ktorom boli stopky zastavené. Následne vypočítame absolútne hodnoty rozdielov distribučných funkcií pre dané jedálne. Tieto výsledky sú zaznamenané v tabuľke 5. Vyčítame z nej aj maximálny rozdiel, ktorý je testovacou štatistikou Kolmogorovovho-Smirnovovho testu a má hodnotu $D_{m,n} = 0,144$. Kritická hodnota na hladine $\alpha = 0,05$ pre $m = 28$, $n = 31$ a $M = \frac{28\times 31}{28+31} = 14,712$ je rovná

$$D_{m,n}(\alpha) = \sqrt{\frac{1}{29,424} \ln \frac{2}{0,05}} = 0,354.$$

Tabuľka 4: Namerané časy zoradené vzostupne

	1. jedáleň	2. jedáleň
1	72	23
2	76	30
3	80	37
4	89	75
5	96	99
6	105	103
7	108	105
8	109	112
9	116	118
10	122	127
11	125	128
12	127	131
13	130	141
14	132	144
15	138	153
16	148	154
17	154	155
18	169	159
19	171	168
20	171	174
21	179	183
22	185	200
23	196	224
24	224	231
25	255	262
26	328	296
27	387	337
28	404	342
29		356
30		359
31		385



Obr. 6: Distribučné funkcie 1. jedálne (modrou farbou) a 2. jedálne (červenou farbou)

Testovacia štatistika je teda menšia než kritická hodnota, preto hypotézu H_0 nezamietame.

Tabuľka 5: Kumulatívne relatívne početnosti oboch jedální

interval	1. jedáleň	2. jedáleň	rozdiel	interval	1. jedáleň	2. jedáleň	rozdiel
$(0, 23]$	0	0.032	0.032	$(144, 148]$	0.571	0.452	0.119
$(23, 30]$	0	0.065	0.065	$(148, 153]$	0.571	0.484	0.087
$(30, 37]$	0	0.097	0.097	$(153, 154]$	0.607	0.516	0.091
$(37, 72]$	0.036	0.097	0.061	$(154, 155]$	0.607	0.548	0.059
$(72, 75]$	0.036	0.129	0.093	$(155, 159]$	0.607	0.581	0.026
$(75, 76]$	0.071	0.129	0.058	$(159, 168]$	0.607	0.613	0.006
$(76, 80]$	0.107	0.129	0.022	$(168, 169]$	0.643	0.613	0.03
$(80, 89]$	0.143	0.129	0.014	$(169, 171]$	0.714	0.613	0.101
$(89, 96]$	0.179	0.129	0.05	$(171, 174]$	0.714	0.645	0.069
$(96, 99]$	0.179	0.161	0.018	$(174, 179]$	0.75	0.645	0.105
$(99, 103]$	0.179	0.194	0.015	$(179, 183]$	0.75	0.677	0.073
$(103, 105]$	0.214	0.226	0.012	$(183, 185]$	0.786	0.677	0.109
$(105, 108]$	0.25	0.226	0.024	$(185, 196]$	0.821	0.677	0.144
$(108, 109]$	0.286	0.226	0.06	$(196, 200]$	0.821	0.71	0.111
$(109, 112]$	0.286	0.258	0.028	$(200, 224]$	0.857	0.742	0.115
$(112, 116]$	0.321	0.258	0.063	$(224, 231]$	0.857	0.774	0.083
$(116, 118]$	0.321	0.29	0.031	$(231, 255]$	0.893	0.774	0.119
$(118, 122]$	0.357	0.29	0.067	$(255, 262]$	0.893	0.806	0.087
$(122, 125]$	0.393	0.29	0.103	$(262, 296]$	0.893	0.839	0.054
$(125, 127]$	0.429	0.323	0.106	$(296, 328]$	0.929	0.839	0.09
$(127, 128]$	0.429	0.355	0.074	$(328, 337]$	0.929	0.871	0.058
$(128, 130]$	0.464	0.355	0.109	$(337, 342]$	0.929	0.903	0.026
$(130, 131]$	0.464	0.387	0.077	$(342, 356]$	0.929	0.935	0.006
$(131, 132]$	0.5	0.387	0.113	$(356, 359]$	0.929	0.968	0.039
$(132, 138]$	0.536	0.387	0.149	$(359, 385]$	0.929	1	0.071
$(138, 141]$	0.536	0.419	0.117	$(385, 387]$	0.964	1	0.036
$(141, 144]$	0.536	0.452	0.084	$(387, 404]$	1	1	0

4 Hry s kockami

Kocky sú dobrým nástojom na vysvetlenie mnohých matematických a štatistických javov. Píše sa o tom aj v článku [7], kde sa konkrétne spomína hra s kockami HOG. Ide o aktivitu nie len s edukačným charakterom v oblasti pravdepodobnosti a štatistiky, ale taktiež v študentoch vzbudzuje veľký záujem. Hra vznikla v roku 1994 a od vtedy si ju zahrali absolventi matematiky, ale aj žiaci základných a stredných škôl.

V tejto kapitole si uvedieme pravidlá hry HOG, ako aj najlepšiu stratégiu z pohľadu téorie pravdepodobnosti z článku [7] pre hru s normálnou šesťstrannou kockou, na ktorej padajú všetky hodnoty s pravdepodobnosťou $\frac{1}{6}$. Následne uvedieme aj nami vymyslenú hru s optimálnou stratégiou a pozrieme sa, či sa hráči rozhodovali na jej základe.

4.1 Teória

Definícia 4.1. *Nech X je diskrétna náhodná veličina, ktorá nadobúda hodnoty x_i, $i = 1, 2, \ldots$ s pravdepodobnosťami $p_i = P(X = x_i)$. Strednou hodnotou náhodnej veličiny X nazveme číslo $E(X) = \sum_{i=1}^{\infty} x_i p_i$, ak tento rad konverguje absolútne. Ak rad nekonverguje absolútne budeme hovoriť, že náhodná veličina nemá strednú hodnotu.*

4.2 Hra HOG

Hra s kockami HOG je podľa článku [7] hrou, pri ktorej sa hráč môže rozhodnúť koľkými kockami bude hádzať, odporúča sa mať k dispozícií aspoň 10 kociek, pričom ich počet sa môže v každom kole meniť. Hráčovi sa započíta počet bodov rovný súčtu hodnôt na kockách okrem prípadu, keď aspoň na jednej padne číslo 1, kedy sa mu nič nezapočíta. Ten hráč, ktorý ako prvý dosiahne vopred dohodnuté skóre, napríklad 100, vyhráva. V prípade, ak sa viacerí hráči dostanú na túto hranicu v tom istom kole, vyhráva ten, ktorý má na svojom konte najvyšší počet bodov.

Na začiatku hľadania optimálnej stratégie si treba uvedomiť, že pri zvyšujúcom sa počte kociek je väčšia pravdepodobnosť, že sa aspoň na jednej kocke objaví číslo 1. Na druhej strane, čím viac kociek sa použije, tým vyšší súčet môže padnúť na kockách.

Jeden z autorov článku nechal študentov hrať túto hru, až kým nedosiahli skóre 100.

Následne sa ich spýtal na počet kociek, ktorým sa podľa nich oplatí hádzať najviac. Väčšina uvádzala hodnoty 3 a 4, čo bolo zrejme zapríčinené prehnaným strachom z padnutia čísla 1 v jednotlivých kolách. Znova ich nechal hrať hru, avšak každej skupine určil, s koľkými kockami majú hádzať. Po 10 hodoch mali za úlohu vypočítať priemerné skóre na jeden hod. Zistili, že čím viac kociek použili, tým vyššie priemerné skóre dosiahli. Preto sa pri rovnakej otázke ako v predošlom prípade skorej uchyľovali k číslam 4, 5 a 6. Potom s nimi spravil experiment, počas ktorého musel každý hádzať vlastným výberom zvoleným fixným počtom kociek, až kým nedosiahli požadovanú hranicu. Výsledkom boli víťazi s 5 kockami v dvoch skupinách, so 6 v ďalších dvoch skupinách a s 10 v jednej skupine.

V článku sú vyčíslené očakávané hodnoty pri hode 1 až 12 kockami. Z tabuľky 7 a obrázku 1, obe z článku [7], je zjavné, že najvyššie skóre by sa malo dať nahrať s 5 a 6 kockami, čo súhlasí so spomínaným experimentom. Pre porovnanie, pri hode 5 alebo 6 kockami je očakávané skóre približne rovné 8, zatiaľ čo pri hode 40 kockami je už blízke 0. Uvádza sa taktiež rozšírenie optimálnej stratégie na s-strannú kocku.

Tabuľka 6: Možné výplaty pre hru s dvoma kockami

	1	2	3	4	5	6
1	2	3	4	5	6	0
2	3	4	5	6	7	0
3	4	5	6	7	8	0
4	5	6	7	8	9	0
5	6	7	8	9	10	0
6	0	0	0	0	0	0

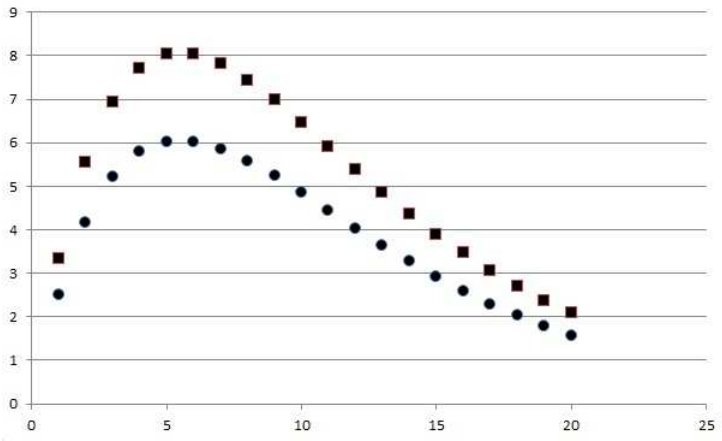
Rozhodli sme sa pozrieť na mierne pozmenený prípad, kedy nežiaducim číslom je číslo 6. V prípade, ak by sme chceli hádzať iba jednou kockou, mali by sme možnosť získať $1, 2, 3, 4, 5$ alebo 0 bodov s pravdepodobnosťou $\frac{1}{6}$, teda očakávané skóre by bolo

$$E(X_1) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 0 = \frac{5}{2},$$

kde X_1 značí hod 1 kockou. Stredná hodnota nenulového skóre, teda $E(X_1|X_1 > 0)$, je 3. Ako možno vypočítať z tabuľky 6, pri 2 kockách je táto hodnota 6. Teda očakávané

skóre z hry 2 kockami je $\left(\frac{5}{6}\right)^2 \times 6 = \frac{25}{6}$. Všeobecne pre hru s n kockami to môžeme zapísať v tvare $\left(\frac{5}{6}\right)^n \times 3 \times n$, kde $\frac{5}{6}$ je pravdepodobnosť a 3 je stredná hodnota nenulového skóre pri hode 1 kockou.

Z obrázku 7 vidíme, že rovnako ako hra s nulovým súčtom pri hode čísla 1, tak aj pozmenená s nulovým súčtom pri padnutí čísla 6 majú najvyššie očakávané skóre pri výbere hry s 5 a 6 kockami.



Obr. 7: Porovnanie očakávaných výhier pôvodnej hry (štvorček) a pozmenenej hry (gulička) pre počet kociek od 1 do 20

Porovnáme ešte pôvodnú hru s takou, v ktorej pád čísel 1 alebo 6 značí súčet 0. Prirodzene, pri vyššom počte neželaných čísel sa zvyšuje pravdepodobnosť ich padnutia. Pozrime sa teda, ako sa zmení stratégia vedúca k najvyššej očakávanej výhre v prípade hry so 6 strannými kockami. Pri hode jednej kocky je očakávaná výhra rovná

$$E(X_1) = \frac{2}{6} \times 0 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 = \frac{7}{3}$$

a za pomoci tabuľky 6 ju vyrátame aj pre hod dvoma kockami

$$E(X_2) = \frac{1}{6} \times 4 + \frac{2}{6} \times 5 + \frac{3}{6} \times 6 + \frac{4}{6} \times 7 + \frac{3}{6} \times 8 + \frac{2}{6} \times 9 + \frac{1}{6} \times 10 = \frac{28}{9}.$$

Očakávanú hodnotu môžeme vo všeobecnosti určiť ako strednú hodnotu premennej $E(XY)$, kde X_i je náhodná premenná, ktorá predstavuje číslo na i-tej kocke nadobúdajúca hodnoty $1, 2, \ldots, 6$, avšak pri čísle 1 a 6 máme súčet 0, preto ich neberieme do úvahy a premenná Y dosahuje hodnoty 1 pre prípad hodu bez čísel 1 a 6 a 0 v opačnom

prípade. Pravdepodobnosť pádu nenulového výsledku pri hode k kockami je $\left(\frac{4}{6}\right)^k$.

$$E(XY) = 0\Big(1 - \Big(\frac{4}{6}\Big)^k\Big) + 1\sum_{i=1}^{k} E(X_i)\Big(\frac{4}{6}\Big)^k = k\frac{2+3+4+5}{6-2}\Big(\frac{4}{6}\Big)^k = \frac{7}{2}k\Big(\frac{2}{3}\Big)^k.$$

Počet kociek, ktorým budeme v priemere dosahovať najvyššie skóre, určíme ako k, pre ktoré má postupnosť $\frac{7}{2}k\Big(\frac{2}{3}\Big)^k$ maximálnu hodnotu. Definujeme $f(k) = \frac{7}{2}k\Big(\frac{2}{3}\Big)^k$ ako funkciu premennej $k \in R^+$. Potom podmienka extrému je:

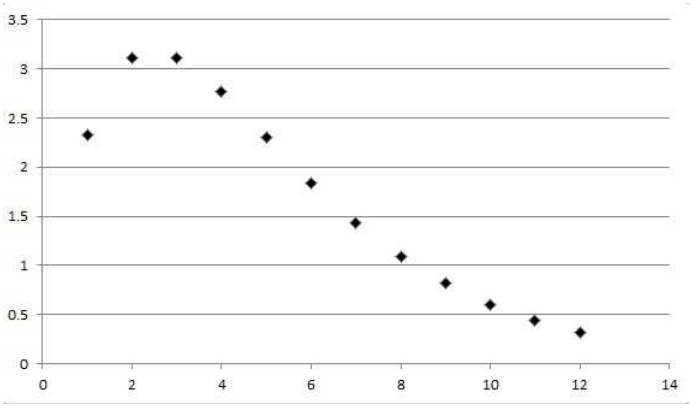
$$f'(k) = (\frac{7}{2}k\Big(\frac{2}{3}\Big)^k)' = \frac{7}{2}\Big(\frac{2}{3}\Big)^k + \frac{7}{2}k\Big(\frac{2}{3}\Big)^k \ln\Big(\frac{2}{3}\Big) = \frac{7}{2}\Big(\frac{2}{3}\Big)^k(1 + k\ln\Big(\frac{2}{3}\Big)) = 0.$$

Výraz $\frac{7}{2}\Big(\frac{2}{3}\Big)^k$ je vždy kladný, preto musí byť

$$(1 + k\ln\Big(\frac{2}{3}\Big)) = 0,$$

$$k = \frac{-1}{\ln\Big(\frac{2}{3}\Big)} = \frac{1}{\ln\Big(\frac{3}{2}\Big)} \doteq 2,466.$$

Súčasne vidíme, že pre $k < \frac{1}{\ln\Big(\frac{3}{2}\Big)}$ je $f'(k) > 0$ a pre $k > \frac{1}{\ln\Big(\frac{3}{2}\Big)}$ je $f'(k) < 0$, takže v bode $k = \frac{1}{\ln\Big(\frac{3}{2}\Big)}$ je maximum. Ak zúžime definičný obor na $k \in N$, maximum sa môže nadobúdať pre $k = 2$ alebo $k = 3$. Vyčíslime: $f(2) = \frac{28}{9} = f(3)$. Najvhodnejšie je teda pri tejto hre zvoliť 2 alebo 3 kocky, čo odzrkadľuje aj graf z obrázku 8.



Obr. 8: Očakávaná výplata pri hre 1 až 12 kockami, kde padnutie čísel 1 a 6 značí za daný hod skóre 0

4.3 Druhá hra

Aj v ďalšej hre je kladený dôraz na jednu zvolenú hodnotu. Hráč má k dispozícii 100 žetónov, bodov. Pred hodom si môže kúpiť 1 až 7 kociek. V prípade, že mu padne aspoň jedna 6, započítajú sa mu body. Ak je na kockách nepárny počet hodnôt 6, získa 6 žetónov. Ak je však počet kociek s číslom 6 párny, pripočíta sa mu $6k$ bodov, kde k značí počet kociek s týmto číslom. Cieľom hry je počas 10 kôl nazbierať čo najviac žetónov.

Vypočítame strednú hodnotu hodu napríklad pre hody s 1 a 7 kockami. Možné výplaty pri hre 1 kockou sú 0 a 6. Za túto kocku však musíme aj zaplatiť, teda

$$E(X_1) = 0 \times \frac{5}{6} + 6 \times \frac{1}{6} - 1 = 0.$$

Pri hre so 7 kockami už so zarátanou sumou za kocky môžeme získať -7, -1, 5, 17 alebo 29 bodov, z čoho

$$\begin{aligned} E(X_7) &= (-7) \times \binom{7}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^7 + (-1) \times \binom{7}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^6 + (-1) \times \binom{7}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^4 \\ &+ (-1) \times \binom{7}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^2 + (-1) \times \binom{7}{7}\left(\frac{1}{6}\right)^7\left(\frac{5}{6}\right)^0 + 5 \times \binom{7}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^5 \\ &+ 17 \times \binom{7}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^3 + 29 \times \binom{7}{6}\left(\frac{1}{6}\right)^6\left(\frac{5}{6}\right)^1 \doteq -0,983. \end{aligned}$$

V tabuľke 7 sú uvedené očakávané výplaty pre hru s 1 až 7 kockami. Ako si môžeme všimnúť, tieto výplaty dosahujú hodnotu rovnú nanajvýš 0, z čoho vyplýva, že najlepšiou reakciou na ponuku hrať túto hru je odmietnuť ju.

Tabuľka 7: Očakávaná výplata v jednom kole hry

počet kociek	očakávaná výhra
1	0
2	0
3	-0,056
4	-0,185
5	-0,389
6	-0,658
7	-0,983

Hru sme hrali s piatimi hráčmi, ktorí boli motivovaní sladkou odmenou pre toho s najvyšším počtom žetónov. Väčšie množstvo kociek zrejme evokuje vidinu vyššej výhry, keďže sa zväčša rozhodli hádzať 6 alebo 7 kockami. Prvý hráč si ani raz nezvolil 6 kociek, ale hral s 1, 2 alebo 3 a skončil s 91 žetónmi. Paradoxne, iný hráč, ktorý stále hádzal so 7 kockami, nazbieral 114 žetónov. Najvyššie skóre, 119, sa podarilo dosiahnuť taktikou zahrňujúcou počet kociek 3 a 7. Najmenší počet žetónov po 10 kolách bol 76. Nakoľko hráči do hry nič nestavili a nemohli teda nič stratiť, nemali možno dostatočnú motiváciu zamýšľať sa nad počtom kociek prinášajúcim najvyššiu výplatu.

5 Chí kvadrát test dobrej zhody a farby $M\&M's$

Farebné čokoládové dražé M&M's sú veľmi obľúbené u konzumentov z celého sveta. Málokto by sa však zrejme zamýšľal, či existuje nejaké pravidlo na rozdelenie cukríkov do jednotlivých balíčkov. Ešte donedávna zverejňovala spoločnosť Mars na svojej internetovej stránke percentuálne zastúpenie každej farby. Podľa článku [1], ktorý sa venuje testovaniu pomerov farieb $M\&M's$, balenie v roku 1991 obsahovalo: 30% hnedej, 20% žltej, 20% červenej a oranžovú, zelenú a žltohnedú po 10% z každej farby. Ako uvádza zdroj [4], prieskum spoločnosti Mars ukázal, že posledná menovaná farba je najmenej obľúbená, preto oslovili verejnosť, aby hlasovala za novú. Z vyše 10 miliónov ľudí sa takmer 55% vyjadrilo v prospech modrej. Žlhohnedú nahradila s totožným percentuálnym vyjadrením.

Článok [2] popisuje novšie tvrdenie spoločnosti Mars z roku 2007, ktoré uvádza nasledujúce pravdepodobnostné rozdelenie cukríkov: 13% hnedej, 14% žltej, 13% červenej, 24% modrej, 20% oranžovej a 16% zelenej.[1]

Naším cieľom bude štatisticky testovať, že pomery farieb v balíčkoch sú naozaj také, o akých hovoria vyjadrenia z rokov 1991 a 2007 spomínané vyššie.

Obr. 9: Testované cukríky

[1] V samotnom tele článku sa spomína, že v balíčku má byť 14% hnedých, 14% žltých, 13% červených, 24% modrých, 20% oranžových a 16% zelených cukríkov, čo dokopy dáva 101%. Ako je však možné vidieť v tabuľke, pri analýze svojich dát použil vyššie uvedené percentá. Totožné údaje sú uvedené aj v [11].

5.1 Teória

Chí kvadrát test dobrej zhody je používaný na overovanie hypotézy, že náhodný výber pochádza z určitého vopred daného rozdelenia pravdepodobnosti.

Podľa vzoru knihy [8] pre diskrétne rozdelenie platí: Nech $y_1, \ldots, y_N$ je realizáciou náhodného výberu. Nech $I_1, \ldots, I_n, n < N$ sú disjunktné intervaly na R a nech

$$p_i = \sum_{I_i} P(X = y_i)$$

sú predpokladané pravdepodobnosti jednotlivých intervalov I_i, pričom $\sum_{i=1}^{n} p_i = 1$. Označme x_i počet bodov výberu $y_1, \ldots, y_N$, ktorý padne do intervalu I_i pre $i = 1, \ldots, n$. Vektor $(x_1, \ldots, x_n)^T$ je realizácia náhodného vektora $(X_1, \ldots, X_n)^T$, o ktorom vieme, že má multinomické rozdelenie so strednou hodnotou $(Np_1, \ldots, Np_n)^T$.

Pre veľké $N = \sum_{i=1}^{n} X_i$ platí:

$$\sum_{i=1}^{n} \frac{(X_i - Np_i)^2}{Np_i} \sim \chi^2_{n-1},$$

čoho dôkaz je možné nájsť v knihe [9].

Hypotézu H_0 na hladine významnosti α zamietame, ak $\chi^2 \geq \chi^2_{n-1}(\alpha)$, kde

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - Np_i)^2}{Np_i}. \tag{1}$$

Na začiatku každého testovania je prirodzene nutné, stanoviť si potrebný počet dát. V prípade chi kvadrát testu dobrej zhody o tom rozhoduje Cochranovo pravidlo, ktoré je uvedené v [10]. Hovorí, že aby sme mohli použiť daný test, musia dáta spĺňať podmienku $Np_i \geq 5$ pre všetky $i = 1, \ldots, n$.

5.2 Testovanie

V nasledujúcej časti budeme skúmať tvrdenie o farbách a ich počtoch v balíčkoch na zozbieraných dátach. Tvrdenie spoločnosti Mars uvedené na začiatku kapitoly hovorí, že balíček obsahuje 30% hnedých, 20% žltých a červených, po 10% z modrých, oranžových a zelených cukríkov. Hypotézu H_0 teda zapíšeme: $p_1 = 0{,}3$, $p_2 = p_3 = 0{,}2$ a $p_4 = p_5 = p_6 = 0{,}1$.

Na začiatku sme riešili otázku počtu potrebných dát. Podľa spomínaného Cochranovho pravidla má byť splnené kritérium $Np_i \geq 5$, kde N je počet dát-cukríkov a p_i pre $i = 1, 2, \ldots, n$ sú testované pravdepodobnosti. Pre splnenie tohto pravidla stačí dosadiť najmenšiu, teda $0,1N \geq 5$, z čoho dostávame $N \geq 50$. V tomto experimente sme dohromady spracovali 443 cukríkov z dvoch balíčkov. Naše výsledky sú zapísané v tabuľke 8. Získané množstvá budeme porovnávať s očakávanými. Prirodzene predpokladáme, že ak sú skutočné dáta príliš odlišné od očakávaných, hypotézu H_0 zamietneme.

Tabuľka 8: Dáta v balíčkoch podľa farieb

	hnedá	žltá	červená	modrá	oranžová	zelená	spolu
balíček 1	12	52	35	46	18	59	222
balíček 2	15	48	41	44	18	55	221
spolu	27	100	76	90	36	114	443

Postupne podosádzame do vzorca (1), kde X_i sú skutočné a Np_i očakávané počty cukríkov v balíčkoch. V tabuľke 9 sú prehľadne zapísané množstvá jednotlivých farieb. Hypotézu H_0 na hladine významnosti $\alpha = 0,05$ zamietame, ak $\chi^2 \geq \chi^2_{6-1}(0,05) = 11,07$.

$$\begin{aligned}\chi^2 = &\frac{(27-132,9)^2}{132,9} + \frac{(100-88,6)^2}{88,6} + \frac{(76-88,6)^2}{88,6} \\ &+\frac{(90-44,3)^2}{44,3} + \frac{(36-44,3)^2}{44,3} + \frac{(114-44,3)^2}{44,3} \quad = \quad 246,01 \end{aligned} \tag{2}$$

Tabuľka 9: χ^2 test (1991)

farby	hnedá	žltá	červená	modrá	oranžová	zelená	spolu
očakávané	132,9	88,6	88,6	44,3	44,3	44,3	443
skutočné	27	100	76	90	36	114	443
$\frac{(X_i-Np_i)^2}{Np_i}$	84,39	1,47	1,79	47,14	1,56	109,66	246,01

Ako sme ukázali, $\chi^2 = 246,01 > \chi^2_{6-1}(0,05) = 11,07$, čiže hypotézu H_0 zamietame. V poslednom riadku tabuľky 9 vidíme príspevok sčítancov zodpovedajúcich jednotlivým farbám k výslednej hodnote štatistiky. Najväčšie odchýlky pri meraní vyšli u hnedej, modrej a zelenej.

Následne overíme aj pravdivosť novšieho tvrdenia spomínaného v úvode kapitoly, kde hypotéza H_0 hovorí: $p_1 = 0{,}13$, $p_2 = 0{,}14$, $p_3 = 0{,}13$, $p_4 = 0{,}24$, $p_5 = 0{,}2$ a $p_6 = 0{,}16$. Postupujeme analogicky ako pri prvom testovaní. Potrebné údaje sú uvedené v tabuľke 10.

Tabuľka 10: χ^2 test (2007)

farby	hnedá	žltá	červená	modrá	oranžová	zelená	spolu
očakávané	57,59	62,02	57,59	106,32	88,6	70,88	443
skutočné	27	100	76	90	36	114	443
$\frac{(X_i - Np_i)^2}{Np_i}$	15,79	23,29	5,59	2,42	31,56	26,04	104,69

Ako môžeme vidieť z posledného riadku tabuľky, testovacia štatistika χ^2 má hodnotu 104,69, z čoho vyplýva, že hypotéza H_0 sa rovnako ako v predošlom prípade zamieta.

Autor článku [2] pri svojom analyzovaní spracoval spolu 48 balíčkov. Vytvoril si pre jednotlivé balenia tabuľky, ktoré sú dostupné na jeho internetovej stránke a postupne do nich vpisoval množstvá cukríkov podľa farieb. Bolo ich dokopy 2620, priemerne teda jeden balíčkek obsahoval 55 cukríkov.

Rozhodli sme sa overiť hypotézu H_0 z predošlého príkladu na jeho dátach zobrazených v tabuľke 11.

Tabuľka 11: χ^2 test s dátami z článku [2]

farby	hnedá	žltá	červená	modrá	oranžová	zelená	spolu
očakávané	340,6	366,8	340,6	628,8	524	419,2	2620
skutočné	371	369	372	481	544	483	2620
$\frac{(X_i - Np_i)^2}{Np_i}$	2,71	0,01	2,89	34,74	0,76	9,71	50,82

Ako môžeme vidieť v tabuľke vyššie, aj napriek tomu, že u hnedej, žltej, červenej a oranžovej ide o pomerne malé odchýlky, test hypotézu H_0 i v tomto prípade zamietol.

6 Hádzanie šípok a testovanie vplyvu dominantnej ruky

Pekné počasie a žiadosti od študentov o vyučovanie konané v prírode sú pre väčšinu pedagógov matematiky „nočnou morou“, nakoľko pre tento predmet je tabuľa výsostne dôležitá. Takáto situácia podnietila autora článku [6] k zamysleniu sa nad aktivitou, pomocou ktorej by bolo možné vysvetliť naplánované učivo a zároveň by sa mohla konať mimo priestorov učebne. Ako v článku spomína, keďže vyrastal na stredozápade Spojených štátov amerických, spomenul si na hranie hry podobnej šípkam odohrávajúcej sa na dvore.

6.1 Teória

Na začiatok uvedieme charakteristiky párového t-testu z kníh [14] a [18] a centrálnu limitnú vetu z [23].

Majme n nezávislých dvojíc náhodných veličín $(X_1, Y_1), \ldots, (X_n, Y_n)$ z dvojrozmerného normálneho rozdelenia, kde vektor (μ_1, μ_2) je vektorom stredných hodnôt. Budeme testovať hypotézu $H_0 : \mu_1 = \mu_2$ proti alternatívnej hypotéze $H_1 : \mu_1 \neq \mu_2$. Označme $Z_i = X_i - Y_i$ pre $i = 1, \ldots, n$. Potom premenné $Z_1, \ldots, Z_n$ sú nezávislé a rovnako rozdelené. Predpokladajme, že pochádzajú z rozdelenia $N(\mu, \sigma^2)$, kde $\mu = \mu_1 - \mu_2$. Hypotézy pre tento test majú tvar $H_0 : \mu = 0$ a $H_1 : \mu \neq 0$. Takýto test sa nazýva párový t-test. Testovacia štatistika má tvar

$$T = \frac{\bar{Z} - 0}{s}\sqrt{n},$$

pričom T je zo Studentovho rozdelenia s $(n-1)$ stupňami voľnosti a

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2.$$

Hypotézu H_0 na hladine významnosti α zamietame, ak $|T| > t_{n-1}\left(\frac{\alpha}{2}\right)$, kde $t_{n-1}\left(\frac{\alpha}{2}\right)$ je tabelovaná hodnota.

Veta 6.1 (Centrálna limitná veta)**.** *Nech $\bar{X}$ je priemer náhodného výberu $X_1, X_2, \ldots, X_n$ o veľkosti n z populácie so strednou hodnotou μ a smerodajnou odchýlkou σ. Centrálna*

limitná veta hovorí, že rozdelenie veličiny $\bar{X}$ je približne normálne so strednou hodnotou μ a smerodajnou odchýlkou $\frac{\sigma}{\sqrt{n}}$ pre dostatočne veľké n.

6.2 Testovanie

V nasledujúcej časti budeme testovať podľa vzoru článku [6], či má pri hádzaní na terč významný vplyv dominantná ruka. Hypotéza H_0 bude tvrdiť, že hod šípkou dominantnou rukou je rovnako presný ako nedominantnou, čiže $H_0 : \mu = 0$.

Experiment sme predviedli na vzorke 18 dobrovoľníkov v interiéri, keďže väčšina z nich sa mohla zúčastniť iba podvečer, kedy sa vonku stmieva, a preto by nemali rovnaké svetelné podmienky. Vyskúšaním hádzania na terč z viacerých vzdialeností sme určili vhodnú vzdialenosť na hádzanie 2 metre, aby bolo možné z tejto vzdialenosti úspešne trafiť i nedominantnou rukou. Každý zo zúčastnených sa pokúšal trafiť čo najbližšie k stredu po 10 razy jednotlivými rukami postupne. Pre nedominantnú ruku boli k dispozícii úvodné 3 skúšobné hody. Po každom pokuse bola odmeraná vzdialenosť šípky od stredu terča. V prípade, ak šípka neskončila zapichnutá v terči, bol umožnený dobrovoľníkovi ďalší hod.

Priemerná úspešnosť jednotlivých dobrovoľníkov na jeden hod v centimetroch je uvedená v tabuľke 12, kde hodnoty v stĺpci rozdiel sú vypočítané ako rozdiel priemerných vzdialeností od stredu hodu nedominantnou rukou a dominantnou, čo podľa centrálnej limitnej vety spĺňa predpoklad normality.

Podľa predošlej podkapitoly má štatistika tvar

$$T = \frac{\bar{Z} - 0}{s}\sqrt{n}.$$

Hodnota $\bar{Z}$ vypočítaná z rozdielov uvedených v tabuľke 12 je 1,056. Následne vypočítame smerodajnú odchýlku

$$s = \sqrt{\frac{1}{18-1}\sum_{i=1}^{18}(Z_i - \bar{Z})^2} = 2,742.$$

Dosadením dostávame

$$T = \frac{1,056 - 0}{2,742}\sqrt{18} = 1,634.$$

Keďže testujeme na hladine $\alpha = 0,05$,

$$T = 1,634 < t_{18-1}\left(\frac{0,05}{2}\right) = 2,110,$$

Tabuľka 12: Priemer vzdialeností všetkých pokusov jednotlivých hráčov od stredu terča

ruka	dominantná	nedominantná	rozdiel
1. človek	8,02	11,71	3,69
2. človek	6,67	10,55	3,88
3. človek	10,45	10,76	0,31
4. človek	7,15	12,68	5,53
5. človek	10,2	11,5	1,3
6. človek	9,69	8,55	-1,14
7. človek	11,05	11,98	0,93
8. človek	11,21	8,29	-2,92
9. človek	10,56	9,33	-1,23
10. človek	8,48	8,65	0,17
11. človek	10,24	11,14	0,9
12. človek	7,28	13,25	5,97
13. človek	10,38	7,19	-3,19
14. človek	4,43	7,54	3,11
15. človek	7,8	9,7	1,9
16. človek	7,43	6,62	-0,81
17. človek	10,49	8,16	-2,33
18. človek	7,12	10,06	2,94
priemer			1,056

z čoho vyplýva, že hypotézu H_0 nezamietame a hod nedominantnou rukou nie je významne odlišný od hodu dominantnou rukou.

Na záver však treba poznamenať, že viacej hodov nedominantnou rukou bolo neúspešných a skončilo nezapichnutých v terči. Keďže nemožno takéto hody merať, mohlo to značne ovplyvniť výsledky testovania. Za spomínané hody by sa však mohlo napríklad pripočítavať nejaké dostatočne veľké číslo, ktoré by reprezentovalo tento neúspech.

V článku [6] testovali nielen rozdielne ruky, ale aj dve rôzne vzdialenosti. Študenti hádzali z miesta vzdialeného približne 4,5 a 9 metra od terča. Hypotéza H_0 hovorila, že priemerný rozdiel je rovný 0, čiže na výsledky nemá vplyv efekt dominantnej ruky.

Výstupom z testovania bola p hodnota o veľkosti 0,109, čo značí žiaden alebo takmer žiadny dôvod na zamietnutie nulovej hypotézy. Ako sa v článku spomína, študenti považovali zber dát za zaujímavý a zábavný. Na druhej strane, podľa počtu študentov na vyučovaní je experiment aj časovo náročný, s čím musíme súhlasiť, keďže priemerný čas na jedného hráča u nás bol 10 minút. Autor uvádza aj rôzne varianty tejto hry, napríklad použítím konskej podkovy, lietajúceho taniera či snehových gúľ v prípade konania v zime.

Na telovýchovnej fakulte iránskej univerzity robili podľa [24] podobný experiment. Ich cieľom bolo zistiť vplyv dominantnej kombinácie ruky a oka na úspešnosť hodu šipkou na terč. Najskôr vybrali náhodne 100 študentov, ktorí mali za úlohu vyplniť dotazník a následne im boli urobené testy na určenie dominantného oka. Z tejto skupiny boli vylúčení tí, ktorí nosili dioptrické okuliare a šošovky, skúsení strelci šípkami a takí, ktorí nevideli na terč iba jedným okom. Zo študentov s vyhovujúcimi výsledkami testov bolo vybratých 20 mužov vo veku od 20 do 23 rokov, ktorých rozdelili po 10 do dvoch skupín. V jednej boli takí, u ktorých sa prejavila jednostranná dominancia, teda dominancia pravého oka a pravej ruky alebo ľavého oka a ľavej ruky. Do druhej skupiny boli zaradení študenti s dominantným pravým okom a ľavou rukou a naopak.

Počas 4 týždňov prebehlo 12 trénovacích stretnutí, na ktorých každý za rovnakých podmienok hádzal 60-krát na terč, 30 skúšobných a 30 meraných hodov. Experiment bol rozdelený na dve fázy, v jednej šlo o osvojenie si zručnosti, pričom sa do úvahy brali výsledky namerané na poslednom trénovacom stretnutí a v druhej o predvedenie naučeného, kde sa merali výsledky hádzania týždeň po ukončení tréningu, avšak k dispozícii už nebolo 30 skúšobných hodov. Tieto merania sa porovnávali podľa dosiahnutého skóre, ktoré záviselo od umiestnenia hodu. Získať sa dalo 0, 1, 3 a 5 bodov, kde hodnota 0 značila nezapichnutie šípky v terči.

Ako sa ukázalo, počas testovania sa pozorovaní študenti naučili lepšie hádzať. Medzi hodmi týchto dvoch skupín nebol významný rozdiel, teda interakcia medzi okom a rukou nemá významný vplyv.

7 Odhad veľkosti populácie

Odhady veľkostí populácií sa využívali už počas 2. svetovej vojny, kedy si podľa [21] Američania zaznamenávali značenia a sériové čísla uvedené na vybavení nepriateľov, napríklad na pneumatikách, tankoch, zbraniach a iných. Z nich získavali odhad nemeckej sily a vojenskej produkcie. Tieto informácie boli potrebné pre lepšie načasovanie útoku na protivníka. Článok je rozdelený na dve časti, prvá časť opisuje historický vývoj a problémy spojené s odhadovaním popísaným vyššie a v druhej časti sa analyzuje spoľahlivosť odhadov získaných touto metódou na základe doposiaľ dostupnej oficiálnej nemeckej produkcie.

V článku [22] sa zas spomína prípad, keď istú britskú univerzitu navštívil Kerstin Vännman. Spolu s profesorom Gottfriedom Noetherom sa vybrali na prechádzku po meste. Zrazu však Gottfried zastal, keďže si všimol číslo uvedené na taxíku a ihneď si ho zaznamenal. Zostali na mieste a všímajúc si autá taxi služby získal nasledovné čísla: 97, 234, 166, 7, 65, 17, 4. Následne položil otázku o počte taxíkov v meste.

Ukážeme si, ako možno veľkosti jednotlivých populácii odhadovať.

7.1 Teória

V tejto časti uvedieme definície pojmov z kníh [8] a [18].

Definícia 7.1. *Nech $X_1, X_2, \ldots, X_n$ je náhodný výber zo základného súboru s rozdelením $f(x, \theta)$. Nech T je merateľné zobrazenie z (R^n, B_n) do (R, B). Potom každú funkciu $T_n = T(X_1, X_2, \ldots, X_n)$, pomocou ktorej odhadujeme neznámy parameter θ, budeme nazývať výberový odhadom (alebo len odhadom) parametra θ.*

Definícia 7.2. *Odhad T parametra θ je nevychýlený, ak*

$$E_\theta(T(X_1, X_2, \ldots, X_n)) = \theta$$

pre každé θ.

Definícia 7.3. *Nevychýlený odhad T^* parametra θ je lepší ako nevychýlený odhad T parametra θ, ak platí*

$$var_\theta(T^*) \leq var_\theta(T)$$

pre všetky θ.

7.2 Testovanie

7.2.1 Odhady bez návratu

Majme súbor dát označených číslami $1, 2, \ldots, N$, kde N je neznáme. Z nich náhodným výberom bez návratu získame $X_1, X_2, \ldots, X_n$, pričom $X_1 < X_2 < \ldots < X_{n-1} < X_n$. Článok [22] predstavuje možnosť, ako vysvetliť študentom, že odhad je náhodná premenná. Uvádza sa v ňom situácia s taxíkmi zo začiatku kapitoly, pričom sa od študentov očakáva, že prídu s vlastnými nápadmi, ako možno množstvo taxíkov N odhadnúť. Rozoberajú sa tu nasledovné 3 odhady študentov:

1. $N = 2\bar{X} - 1$,
2. $N = X_1 + X_n - 1$,
3. $N = X_n + \frac{X_n}{n} - 1$.

Vysvetlenia uvádzame z článku [25].

Pri prvom odhade sa predpokladá, že poznáme prostredné z čísel $1, 2, \ldots, N$, nazvime ho m. Pred týmto číslom sa nachádza $m - 1$ čísel, rovnako ako za ním, čo dokopy dáva $N = 2(m-1)+1 = 2m-1$. Keďže pri rovnomernom rozdelení je stredná hodnota rovná mediánu, hodnotu m môžeme nahradiť odhadom strednej hodnoty zo vzorky $X_1, X_2, \ldots, X_n$. Pri tomto odhade však môže vyjsť odhad N menší než maximálna hodnota X_n.

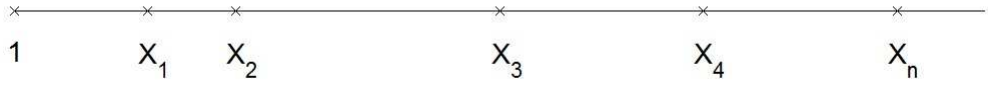
Druhý zas vzišiel zo symetrie, pretože sa predpokladá, že počet neznámych čísel väčších než X_n by mal byť približne rovný počtu nevytiahnutých čísel, ktoré sú menšie než X_1, teda $N - X_n = X_1 - 1$.

Tretí odhad vznikol pri načrtnutí hodnôt od 1 až po najväčšie známe číslo na číselnú os. Následne zistíme počet neznámych čísel na všetkých intervaloch, čiže menších než X_1, medzi X_1 a X_2, ..., medzi X_{n-1} a X_n. Z vysvetlenia uvedeného v predošlom odhade potom máme

$$N - X_n = \frac{\Big((X_1 - 1) + (X_2 - X_1 - 1) + \ldots + (X_n - X_{n-1} - 1)\Big)}{n} = \frac{X_n - n}{n} = \frac{X_n}{n} - 1.$$

Z toho

$$N = X_n + \frac{X_n}{n} - 1 = \frac{n+1}{n} X_n - 1.$$



Obr. 10: číselná os s vytiahnutými hodnotami

V ďalšej časti sa pozrieme na vychýlenosť či nevychýlenosť týchto odhadov. Najprv overíme túto vlastnosť pri poslednom odhade. Určíme ju z

$$E\Big(\frac{n+1}{n}X_n - 1\Big) = \frac{n+1}{n}E(X_n) - 1.$$

Potrebujeme zistiť, čomu sa rovná $E(X_n)$. Náhodne vybrať n čísel z množiny s N prvkami môžeme $\binom{N}{n}$ spôsobmi. Zoradíme ich podľa hodnoty vzostupne. Nech číslo X_n je maximom z vybratých čísel a platí $X_n = k$, potom čísla $X_1, X_2, \ldots, X_{n-1}$ nadobúdajú hodnoty od 1 do $k-1$. Z toho pravdepodobnosť, že $X_n = k$ vypočítame ako

$$P(X_n = k) = \frac{\binom{1}{1}\binom{k-1}{n-1}}{\binom{N}{n}},$$

kde $k = n, n+1, \ldots, N$. Z definície strednej hodnoty platí

$$\begin{aligned} E(X_n) &= \sum_{x=n}^{N} xP(X_n = x) = \sum_{x=n}^{N} x\frac{\binom{x-1}{n-1}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}}\sum_{x=n}^{N} x\frac{(x-1)!}{(n-1)!(x-n)!} \\ &= \frac{1}{\binom{N}{n}}\sum_{x=n}^{N}\frac{nx!}{n!(x-n)!} = \frac{n}{\binom{N}{n}}\sum_{x=n}^{N}\binom{x}{n}. \end{aligned} \tag{3}$$

Súčet pravdepodobností nastatia jednotlivých udalostí z pravdepodobnostného priestoru je vždy 1, teda

$$\sum_{k=n}^{N}\frac{\binom{k-1}{n-1}}{\binom{N}{n}} = 1$$

$$\sum_{k=n}^{N}\binom{k-1}{n-1} = \binom{N}{n}. \tag{4}$$

Dosadením do (3) dostávame

$$E(X_n) = \frac{n}{\binom{N}{n}}\binom{N+1}{n+1} = n\frac{(N+1)!n!(N-n)!}{(n+1)!(N-n)!N!} = \frac{n(N+1)}{n+1},$$

čiže

$$E\Big(\frac{n+1}{n}X_n - 1\Big) = \frac{n+1}{n}E(X_n) - 1 = \frac{n+1}{n}\frac{n(N+1)}{n+1} - 1 = N,$$

a preto je tento odhad nevychýlený.

Preskúmame spomínanú vlastnosť aj pre druhý odhad.

$$E(X_n + X_1 - 1) = E(X_n) + E(X_1) - 1,$$

kde $E(X_n)$ už poznáme z predošlého výpočtu a $E(X_1)$ je zatiaľ neznáma. Budeme postupovať analogicky ako pri hľadaní $E(X_n)$. Nech $X_1 = i$, potom ostatné vytiahnuté čísla sú rovné $i + 1$ až N. Vypočítame pravdepodobnosť

$$P(X_1 = i) = \frac{\binom{1}{1}\binom{N-i}{n-1}}{\binom{N}{n}},$$

kde $i = 1, \ldots, N - n + 1$.

$$E(X_1) = \sum_{x=1}^{N-n+1} xP(X_1 = x) = \sum_{x=1}^{N-n+1} x\frac{\binom{N-x}{n-1}}{\binom{N}{n}} = \frac{1}{\binom{N}{n}} \sum_{x=1}^{N-n+1} x\binom{N-x}{n-1}.$$

Spravíme substitúciu $y = N - x + 1$ a využitím (4) dostávame

$$\begin{aligned} E(X_1) &= \frac{1}{\binom{N}{n}} \sum_{y=n}^{N} (N+1-y)\binom{y-1}{n-1} \\ &= \frac{1}{\binom{N}{n}} \sum_{y=n}^{N} (N+1)\binom{y-1}{n-1} - \frac{1}{\binom{N}{n}} \sum_{y=n}^{N} y\binom{y-1}{n-1} \\ &= \frac{N+1}{\binom{N}{n}}\binom{N}{n} - E(X_n) = N + 1 - E(X_n). \end{aligned} \tag{5}$$

Vrátime sa na začiatok a dosadením $E(X_n)$ a $E(X_1)$ dostávame

$$E(X_n + X_1 - 1) = E(X_n) + N + 1 - E(X_n) - 1 = N,$$

čiže ďalší nevychýlený odhad je $X_n + X_1 - 1$.

Následne preskúmame odhad $N = 2\bar{X} - 1$.

$$E(2\bar{X} - 1) = E\Big(\frac{2}{n}\sum_{i=1}^{n} X_i - 1\Big) = \frac{2}{n}\sum_{i=1}^{n} E(X_i) - 1.$$

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n = Y_1 + Y_2 + \ldots Y_N,$$

kde

$$Y_k = \begin{cases} k, & \text{ak číslo } k \text{ je vo výbere} \\ 0, & \text{inak} \end{cases}$$

a platí

$$E(X_1 + X_2 + \ldots + X_n) = E(Y_1 + Y_2 + \ldots Y_N) = \sum_{k=1}^{N} E(Y_k).$$

$$E(Y_k) = kP(Y_k = k) = k\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = k\frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = k\frac{n}{N}$$

a dosadením dostávame

$$\sum_{k=1}^{N} E(Y_k) = \sum_{k=1}^{N} k\frac{n}{N} = \frac{n}{N}\frac{N(N+1)}{2} = \frac{n(N+1)}{2}.$$

Teda

$$E(2\bar{X} - 1) = \frac{2}{n}\frac{n(N+1)}{2} - 1 = N,$$

z čoho vyplýva, že odhad $2\bar{X} - 1$ je nevychýlený.

Variancie týchto odhadov sme získali z článku [25] a zapísali do tabuľky 13. Podľa tohto článku je posledný zo spomínaných najlepší vďaka najmenšej variancii.

Tabuľka 13: Variancie odhadov

odhad	variancia odhadu
$N = 2\bar{X} - 1$	$\frac{(N-n)(N+1)}{3n}$
$N = X_1 + X_n - 1$	$\frac{2(N-n)(N+1)}{(n+1)(n+2)}$
$N = X_n + \frac{X_n}{n} - 1$	$\frac{(N-n)(N+1)}{n(n+2)}$

7.2.2 Odhady s návratom

Podobný experiment sme vyskúšali so študentmi na predmete Metódy riešenia úloh z pravdepodobnosti a štatistiky. Papier sme postrihali na malé kúsky a na každý z nich sme napísali čísla od 1 do 67. Vhodili sme ich do nepriehľadého vrecka a premiešali. Ešte pred predvedením pred študentmi sme si situáciu 26-krát vyskúšali pre náhodný výber 4 až 9 papierikov s návratom. Pre jednotlivé počty a odhady z článku [22] sme vypočítali priemer a smerodajnú odchýlku, aby sme vedeli vybrať počet papierikov, ktorý by dával dobrú informáciu o ich celkovom množstve. Ako vidíme z tabuľky 14, už pri náhodnom výbere 6 papierikov je priemer približne rovný skutočnému množstvu a smerodajná odchýlka je porovnateľná s tými pre väčšie výbery, preto sme sa rozhodli vytiahnuť na cvičení práve pre tento počet papierikov.

Pred študentmi bolo vytiahnutých 6 čísel: 50, 39, 19, 55, 13 a 29, o ktorých sa následne diskutovalo. Navrhovali možnosti ako odhadovať množstvo papierikov vo vrecúšku, napríklad podľa priemeru vytiahnutých. Taktiež začali rozmýšľať nad možnosťou, že by zoradili čísla vzostupne a pozreli sa na počet čísel pred minimom z vytiahnutých. Dostali možnosť týždeň rozmýšľať sa nad ďalšími nápadmi odhadov s informáciou rozšírenou o čísla 30, 31 a 29.

Tabuľka 14: Priemery a smerodajné odchýlky pre jednotlivé výbery

Počet čísel	$N = 2\bar{X} - 1$		$N = X_n + X_1 - 1$		$N = \frac{n+1}{n}X_n - 1$	
	$\bar{x}$[i]	s[ii]	$\bar{x}$	s	$\bar{x}$	s
4	65,635	22,058	63,192	18,1	63,904	13,802
5	68,338	17,935	65,5	14,465	66,062	10,73
6	67,654	15,833	66,269	13,163	66,308	8,523
7	67,637	15,38	66,462	11,669	66,561	7,216
8	68,356	15,007	66,462	10,638	66,543	6,811
9	67,487	14,437	66,154	10,155	66,009	6,792

Bolo odovzdaných 5 odhadov a vo väčšine z nich bol uvedený priamočiary odhad $N = 2\bar{X} = 2 \times 32,778 = 65,556$ alebo $N = 2\bar{X} - 1 = 64,556$, avšak iba jeden študent pri tomto čísle aj ostal. Medzi modelmi na odhadovanie celkovej populácie sa objavili aj také, ktoré po vyčíslení dali hodnotu menšiu než maximum z vytiahnutých čísel. Išlo o odhady $N = 2IQR$ a $N = \tilde{X} + IQR$, kde IQR je medzikvartilové rozpätie, teda rozdiel medzi 1. a 3. kvartilom a $\tilde{X}$ je medián výberu. V špecifických prípadoch, napríklad pri vytiahnutí čísel 1, 2, 3, 4, 21, 30 a 37, môžu vychádzať aj odhady ako $N = 2\bar{X} - 1$ a $N = 2\tilde{X} - 1$ nižšie než vytiahnuté maximum. Zaujímavý je odhad $N = \frac{n+1}{n}X_n - \frac{1}{n}$, ktorý vznikol z toho, že náhodný výber z množiny celých, po sebe idúcich čísel by mal byť rozložený rovnomerne pozdĺž číselné osi. Vieme, že prvé číslo je 1. Tvar odhadu dostaneme vyrátaním priemeru dĺžok intervalov.

Zo zozbieraných odhadov vidíme, že sa študenti zväčša uchyľovali k odhadovaniu populácie okrúhlymi číslami, napr. 60, 65 a 70. Ako dôvod uvádzali fakt, že počet

[i] priemer výberov

[ii] smerodajná odchýlka

čísel určoval človek, preto predpokladali, že najväčšie číslo na papierikoch bude nejakým násobkom čísla 5. Takéto názory sme očakávali, preto sme zvolili 67 lístkov, teda „nepekné“ prvočíslo.

Na cvičení bola opýtaná otázka ohľadom existencie správnej odpovede na otázku o odhadovanom počte. Správnu odpoveď ohľadom počtu poznáme, keďže sme si pripravovali papieriky vopred, avšak neexistuje odpoveď na to, ktorý odhad poskytne informáciu o správnom počte.

Študenti taktiež rozmýšľali aj nad myšlienkou, či pri odhadovaní čísla je dôležitá iná informácia než o maxime. Maximálna hodnota má pri odhadovaní významnú úlohu, keďže hovorí aké je minimálne najvyššie číslo vo vrecúšku. Ako vidíme z tabuľky 14, pri hľadaní vhodného počtu vytiahnutých papierikov mal tento odhad pri všetkých možnostiach najnižšiu smerodajnú odchýlku, čo podporuje túto myšlienku.

Ukázalo sa, že študentov téma zaujala a odpoveď hľadali aj pomocou internetu.

Záver

V práci sme sa venovali siedmim experimentom prevedeným na situáciach z každodenného života. Snahou bola aplikácia získaných teoretických vedomostí pri riešení konkrétnych situácií, ktoré boli spracované podrobne, jednoduchou a zrozumiteľnou formou.

V prvej kapitole sme sa zaoberali testovaním hypotéz podľa článku [5]. Princíp sme ozrejmili na jednoduchom prípade rozlišovania cukríkov M&M's a Lentiliek podľa chuti. Ako sa ukázalo, degustátor správne určil 18 z 20 ochutnávaných vzoriek, čiže sme nezamietli hypotézu, že vie cukríky naozaj rozlíšiť.

V ďalšej časti sme sledujúc článok [3] vysvetlili pojem distribučná funkcia na stĺpikoch z bočnej strany slovníka. Pri porovnaní distribučných funkcií slovníkov [12] a [13] pomocou Kolmogorovovho-Smirnovovho testu sme nezistili medzi nimi štatisticky významný rozdiel. Nakoľko tento test predpokladá spojité distribučné funkcie, rozhodli sme sa vymyslieť vhodnejší príklad, a preto sme v kapitole o jedálňach nadviazali na túto tému. Test však rozdiel medzi čakaním v jedálňach Eat and Meet a Venza opäť nepotvrdil.

Štvrtá kapitola bola venovaná hrám s kockami. Pre hru HOG z článku [7] sme uviedli očakávané výplaty z hry rôznymi počtami kociek. Pri rôznych variantoch tejto hry sme vyriešili otázku optimálnej stratégie. Následne sme sledovali hráčov pri nami vymyslenej hre. Napriek tomu, že sa nesprávali optimálne, dosahovali pomerne vysoké skóre. Zaujímavé by možno bolo, pozrieť sa na ich hru v prípade, že by do hry museli staviť niečo vlastné. Reakciou by zrejme bolo hlbšie zamyslenie nad pravidlami, po ktorom by pravdepodobne nasledovalo odmietnutie hry alebo hra s nižším počtom kociek.

V nasledujúcej časti sme predstavili chí kvadrát test dobrej zhody, ktorým sme testovali percentuálne rozdelenie farieb cukríkov M&M's v balíčkoch. Ukázalo sa, že naše dáta, rovnako ako dáta z [2], nezodpovedajú percentám z článkov [1] a [2], nakoľko test hypotézu o zhode zamietol.

Predposledná kapitola opisovala experiment, v ktorom šlo o hádzanie na terč. Hráči mali za úlohu triafať dominantnou a nedominantnou rukou čo najbližšie k stredu. Meralo sa 10 hodov. Výsledok experimentu použitím párového t-testu bol totožný s

výsledkom z článku [6], teda test hypotézu o irelevantnosti výberu ruky na hádzanie nezamietol.

V 7. kapitole sme uviedli rôzne príklady odhadov z článkov [22] a [25]. Odvodili sme nevychýlenosť troch spomínaných odhadov. Pre spracovanie tejto kapitoly sme sa zúčastnili predmetu Metódy riešenia úloh z pravdepodobnosti a štatistiky a pred študentmi vytiahli najskôr 6 papierikov, na záver cvičenia sme však vybrali ešte ďalšie 3. Študenti sa zamýšľali nad otázkou ohľadom počtu papierikov vo vrecúšku. V druhej časti tejto kapitoly sme predstavili odhady, ktoré študentom napadli počas cvičení i tie, nad ktorými mali možnosť doma pouvažovať.

Práca bola veľkým prínosom pre autorku, nakoľko získanými výsledkami z experimentov sa jej podarilo lepšie zorientovať a pochopiť opodstatnenie využívania štatistických metód. Uvedomila si, že javy v pozadí štatistiky, ktoré sú zdanlivo jednoduché, často podliehajú zložitému skúmaniu.

Zoznam použitej literatúry

[1] Johnson, R. W.: *Testing Colour Proportions of M&M's*. Teaching Statistics 15 (1993), 2-4

[2] Madison, J.: *M&M's color distribution analysis*, dostupné na internete (6.12.2014): http://joshmadison.com/2007/12/02/mms-color-distribution-analysis/

[3] Jernigan, R. W.: *A Photographic View of Cumulative Distribution Functions*. Journal of Statistics Education 16 (2008), dostupné na internete (6.12.2014): http://www.amstat.org/publications/jse/v16n1/jernigan.html

[4] Ficker, Jr., R. D.: *The Mysterious Case of the Blue M&M's*. Chance 9 (1996), 19-12

[5] Levine, M., Rolwing, R. H.: *Coke or Pepsi?*. Teaching Statistics 15 (1993), 4-5

[6] Nordmoe, E.: *Lawn Toss: Producing Data On-the-Fly*. Teaching Statistics 20 (1998), 66-67

[7] Feldman, L., Morgan, F.: *The Pedagogy and Probability of the Dice Game Hog*. Journal of Statistics Education 11 (2003), dostupné na internete (6.12.2014): http://www.amstat.org/publications/jse/v11n2/feldman.html

[8] Janková, K., Pázman, A.: *Pravdepodobnosť a štatistika*, Univerzita Komenského, Bratislava, 2011

[9] Štěpán, J., Zvára, K.: *Pravděpodobnost a matematická statistika*, Matfyzpress, Praha, 1997

[10] Ostertagová, E.: *Aplikácia štatistických testov dobrej zhody*, dostupné na internete (29.1.2015): http://www.sjf.tuke.sk/transferinovacii/pages/archiv/transfer/23-2012/pdf/072-074.pdf

[11] Schultz, L.: *M&Ms Quality Control: A Chi-Square Analysis*, dostupné na internete (31.1.2015): http://users.rowan.edu/~schultzl/Activities/M&MsExercise.pdf

[12] Fronek, J., Mokráň, P.: *Rozsiahly prehľadný Anglicko-slovenský slovník*, Nová práca, Bratislava, 2006

[13] *Active Study Dictionary - International Students Edition*, Longman, Paris, 2000

[14] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2005

[15] Norvig, P.: *English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU*, dostupné na internete (7.2.2015): http://norvig.com/mayzner.html

[16] Wicklin, R.: *The frequency of letters in an English corpus*, dostupné na internete (7.2.2015): http://blogs.sas.com/content/iml/2014/09/19/frequency-of-letters/

[17] Płocki, A., Tlustý, P.: *Pravděpodobnost a statistika pro začátečníky a mírně pokročilé*, Prometheus, Praha, 2007

[18] Lamoš, F., Potocký, R.: *Pravdepodobnosť a matematická štatistika - Štatistické analýzy*, Univerzita Komenského, Bratislava, 1998

[19] Baglivo, J.: *MT427: Mathematical Statistics, Course Notes*, Notebook 3, dostupné na internete (22.3.2015): https://www2.bc.edu/~baglivo/MT427/notebook03.pdf

[20] Reinert, G.: *Statistical Theory*, dostupné na internete (22.3.2015): http://www.stats.ox.ac.uk/~reinert/stattheory/theoryshort09.pdf

[21] Ruggles, R., Brodie, H.: *An Empirical Approach to Economic Intelligence in World War II*, dostupné na internete (4.4.2015): https://engineering.purdue.edu/~ipollak/ece302/FALL09/notes/An_Empirical_Approach_to_Economic_Intelligence_in_World_War_II_Ruggles_Brodie_1947.pdf

[22] Vännman, K.: *How to Convince a Student that an Estimator is a Random Variable*. Teaching Statistics 5 (1983), 49-54, dostupné na internete (11.4.2015): http://www.teachingstatistics.co.uk/bts/vannman/text.html

[23] Hystad, G.: *Chapter 5: Sampling Distributions and the Central Limit Theorem*, dostupné na internete (9.5.2015): http://math.arizona.edu/~ghystad/chapter5.pdf

[24] Maleki, F. a kol.: *Effect of interaction between eye-hand dominance on dart skill.* Journal of Neuroscience and Behavioral Health 4 (2012), 6-12, dostupné na internete (15.5.2015): http://www.academicjournals.org/article/article1379688177_Razeghi%20et%20al.pdf

[25] Johnson, R. W.: *Estimating the Size of a Population.* Teaching Statistics 16 (1994), 50-52