

Chí kvadrát test dobrej zhody

Beáta Stehlíková

Metódy riešenia úloh z pravdepodobnosti a štatistiky

FMFI UK v Bratislave

Chí kvadrát test dobrej zhody

- Chceme overiť, **či naše dáta pochádzajú z konkrétneho pravdepodobnostného rozdelenia.**
- Napríklad:
 - Je kocka pravidelná? Padajú na nej čísla 1, 2, 3, 4, 5, 6 rovnako často?
 - Je pravdepodobnosť výhry jednotlivých hráčov hry *mang kung* z domácej úlohy rovnaká?
 - Dostávame simuláciami v domácej úlohe náhodné čísla z exponenciálneho rozdelenia s parametrom 1?

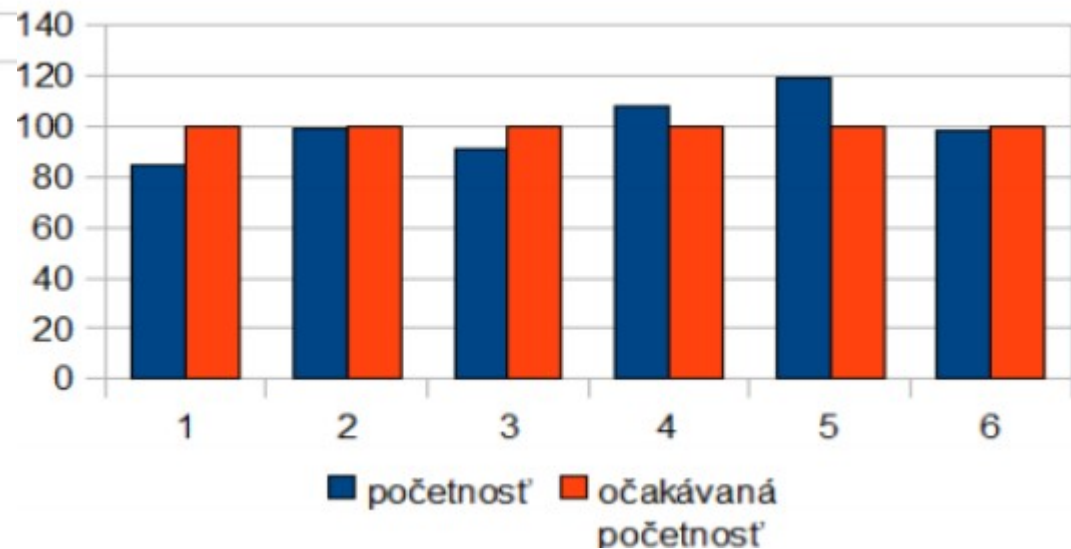
- **Príklad:** Môžeme kocku považovať za pravidelnú?
- Výsledky 600 hodov:

padnuté číslo	početnosť
1	85
2	99
3	91
4	108
5	119
6	98

- Očakávané početnosti sú všetky 100 – nakoľko veľké sú rozdiely?

- Pozorované a očekávané počtnosti:

	A	B	C	D
1	padnuté číslo	početnost	teoretická pravd.	očekávaná početnost
2	1	85	0,1667	=+C2*\$B\$9
3	2	99	0,1667	100
4	3	91	0,1667	100
5	4	108	0,1667	100
6	5	119	0,1667	100
7	6	98	0,1667	100
8				
9	spolu	600		



- Rozdiely:

početnosť	teoretická pravd.	očakávaná početnosť	početnosť – očakávaná
85	0,1667	100	-15
99	0,1667	100	-1
91	0,1667	100	-9
108	0,1667	100	8
119	0,1667	100	19
98	0,1667	100	-2

- Na čo treba myslieť:

- Kladné a záporné odchýlky sa nemôžu navzájom zrušiť
- Rozdiel -15 nemá rovnakú váhu, keď je očakávaný počet 100 a keď je 1000

	D	E	F
1	očakávaná početnosť	početnosť – očakávaná	$(\text{početnosť} - \text{očakávaná})^2 / (\text{očakávaná})$
2	100	-15	$=+(E2^2)/D2$
3	100	-1	0,01
4	100	-9	0,81
5	100	8	0,64
6	100	19	3,61
7	100	-2	0,04

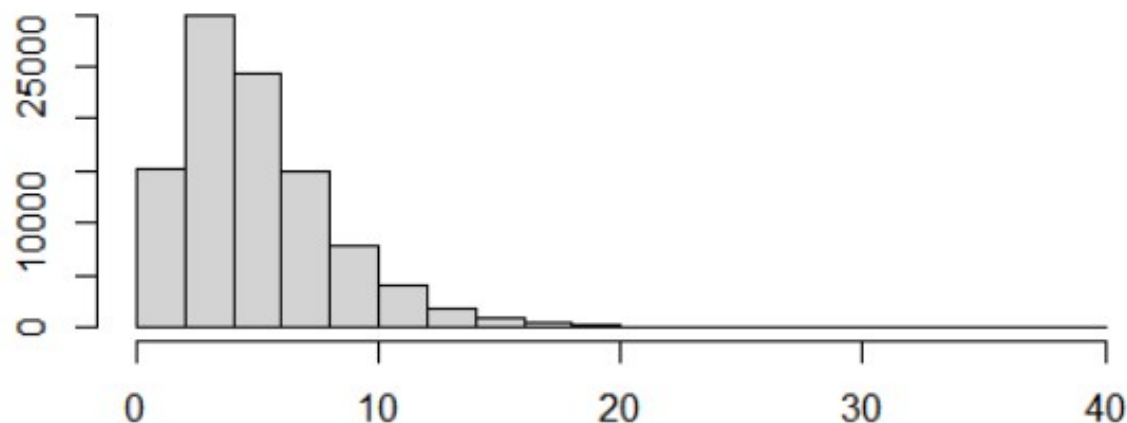
- „Mierou rozdielu“ medzi dátami a testovaným rozdelením bude súčet týchto rozdielov
- Ak bude veľký, hypotézu zamietneme

- Čo znamená „veľký rozdiel“:

```
statistika <- function(){  
  data <- sample(1:6, size = 600, replace = TRUE)  
  stat <- sum((table(factor(data, levels = 1:6)) - 100)^2/100)  
  return(stat)  
}
```

```
hodnoty <- replicate(10^5, statistika())
```

Histogram of hodnoty



```
> quantile(hodnoty, 0.95)  
95%  
11.1
```

- Čo znamená „veľký rozdiel“:
 - Väčší ako kvantil chí kvadrát rozdelenia s $q-1$ stupňami voľnosti, kde q je počet skupín, do ktorých sú dáta zaradené (teraz je teda $q=6$)
 - Chí kvadrát rozdelenie je aproximácie, ktorá sa dá použiť, ak je očakávaný počet dát v každej skupine aspoň 5



```
> qchisq(0.95, df = 5)
[1] 11.0705
```

- V našom prípade:

f
$(\text{početnosť} - \text{očakávaná})^2 / (\text{očakávaná})$
2,25
0,01
0,81
0,64
3,61
0,04
=SUM(F2:F7)
7,36

- Pripomeňme si:

```
> qchisq(0.95, df = 5)
```

```
[1] 11.0705
```

- Záver: **Pravidelnosť našej kocky nezamietame.**

Príklad 1: Výhry v hre *mang kung*

◆ DESCRIPTION OF THE GAME ◆

We first describe the game with three players. Each player contributes 7 units of stake into a “pool”. Thus, the total initial amount in the pool is 21 units. A player is randomly selected to start the first game. Each time six *Mang Kung Dice* are thrown simultaneously. The total face value of these dice is counted. This total face value (F) is then compared with the total amount in the pool (P). The possible outcomes and the corresponding player's actions are listed in Table 1.



Table 1. The three possible situations in the Mang Kung Dice Game.

Outcomes	Player's Actions
$F < P$	The Player can take the amount F from the pool. He then passes the dice to the next player on his left to continue the game.
$F > P$	The Player has to top up the pool by the difference, $(F - P)$. He then passes the dice to the next player to continue the game.
$F = P$	The Player wins the game. He receives the amount P from the pool. Moreover, each of the other players has to pay him an amount F as “bonus”. He is also the starter of the next new game.

```
> table(vitazi)
vitazi
      1      2      3
3335 3359 3306
```

Naprogramujte priebeh hry a odhadnite pravdepodobnosť výhry prvého, druhého a tretieho hráča. Uveďte aj počet simulácií a presné počty víťazstiev jednotlivých hráčov (využijeme neskôr).

- Spočítame najskôr test ručne.
- V R-ku funkcia **chisq.test**:

```
> chisq.test(table(vitazi), p = c(1/3, 1/3, 1/3))
```

Chi-squared test for given probabilities

```
data: table(vitazi)  
X-squared = 0.4226, df = 2, p-value = 0.8095
```

- Rovnaké pravdepodobnosti sú defaultné, takže stačí:

```
> chisq.test(table(vitazi))
```

Chi-squared test for given probabilities

```
data: table(vitazi)  
X-squared = 0.4226, df = 2, p-value = 0.8095
```

Príklad 2: Hypotéza o Exp(1) rozdelení

Príklad 5 (20 bodov)

Pomocou simulácií odhadnite, aké známe rozdelenie (teda typ rozdelenia a jeho parameter, resp. parametre) má náhodná premenná X , ktorá vzniká nasledujúcim algoritmom, pričom Y sú nezávislé náhodné premenné s rovnomerným rozdelením na $(0, 1)$:

Select a random number Y_1 , then another one Y_2 , etc., continuing as long as the numbers so far chosen form a decreasing sequence, i.e.,

$$Y_1 > Y_2 > \dots > Y_n > Y_{n+1}$$

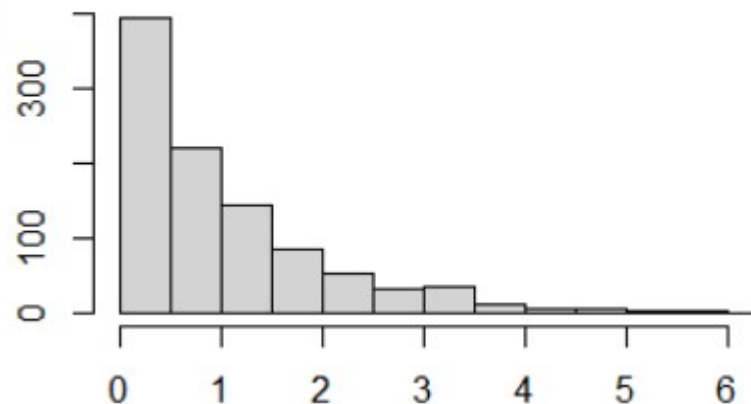
As soon as a random number is drawn that exceeds its immediate predecessor, the process halts.

If n is even, forget the entire sequence of random numbers, call this experiment a failure, and repeat from the beginning.

If, however, n is odd then we're ready for output. The number X that will be output is just this: the integer part of X is the number of failures so far seen, and the fractional part of X is Y_1 , the largest (and first) member of the last decreasing sequence chosen.

For example, if we choose random numbers .63, .19, .37 we would record a failure and start again. If the second sequence is .44, .91 then we are ready for output, and the number that is selected by these events is $X = 1.44$.

Histogram of x



```
> mean(x)
[1] 1.042542
> var(x)
[1] 1.086234
```

- Pozorované početnosti hodnôt v jednotlivých intervaloch (spolu 1000):

```
> pozorovane <- table(cut(x, breaks = c(0:4, Inf)))  
> pozorovane
```

(0,1]	(1,2]	(2,3]	(3,4]	(4,Inf]
617	230	87	46	20

- Teoretické pravdepodobnosti **prob** z domácej úlohy:

```
> ocakavane <- prob * 1000
```

```
> ocakavane
```

```
[1] 632.12056 232.54416 85.54821 31.47143 18.31564
```

- Chí kvadrát test:

```
> chisq.test(pozorovane, p = prob)
```

Chi-squared test for given probabilities

data: pozorovane

X-squared = 7.2761, df = 4, p-value = 0.122

Príklad 3: Testovanie kocky II.

- Walter F. L. Weldon (1860 – 1906)
 - Hodil 12 kockami, zopakoval to 26 306 krát
 - Úspech – padnutie 5 alebo 6
 - Teoretická pravdepodobnosť $1/3$
 - Prakticky vyššia početnosť, pripisoval to konštrukcii bodiek, ktorá mala podľa neho vplyv na váhu stien
- Budeme testovať hypotézu, že počet úspechov má rozdelenie $\text{Bin}(12, 1/3)$



- Weldonove výsledky v tabuľke
- Niekoľko známych mien zo štatistiky:
 - Weldon poslal dáta v liste Galtonovi
 - Analyzoval ich aj Pearson, ktorý je autorom aj tohto chí kvadrát testu dobrej zhody

úspechov	početnosť
0	185
1	1149
2	3265
3	5475
4	6114
5	5194
6	3067
7	1331
8	403
9	105
10	14
11	4
12	0
spolu	26306

- Požiadavka na očakávané počty:

úspechov	početnosť	očakávaná početnosť	
0	185	202,75	
1	1149	1216,5	
2	3265	3345,37	
3	5475	5575,61	
4	6114	6272,56	
5	5194	5018,05	
6	3067	2927,2	
7	1331	1254,51	
8	403	392,04	
9	105	87,12	
10	14	13,07	treba zlúčiť na "10 a viac" kvôli požiadavke na očakávaný počet
11	4	1,19	
12	0	0,05	
spolu	26306		

- Dokončíme testovanie

Príklad 4: Bertrandov paradox

- Bakalárska práca, Katarína Kocsisová, 2017:

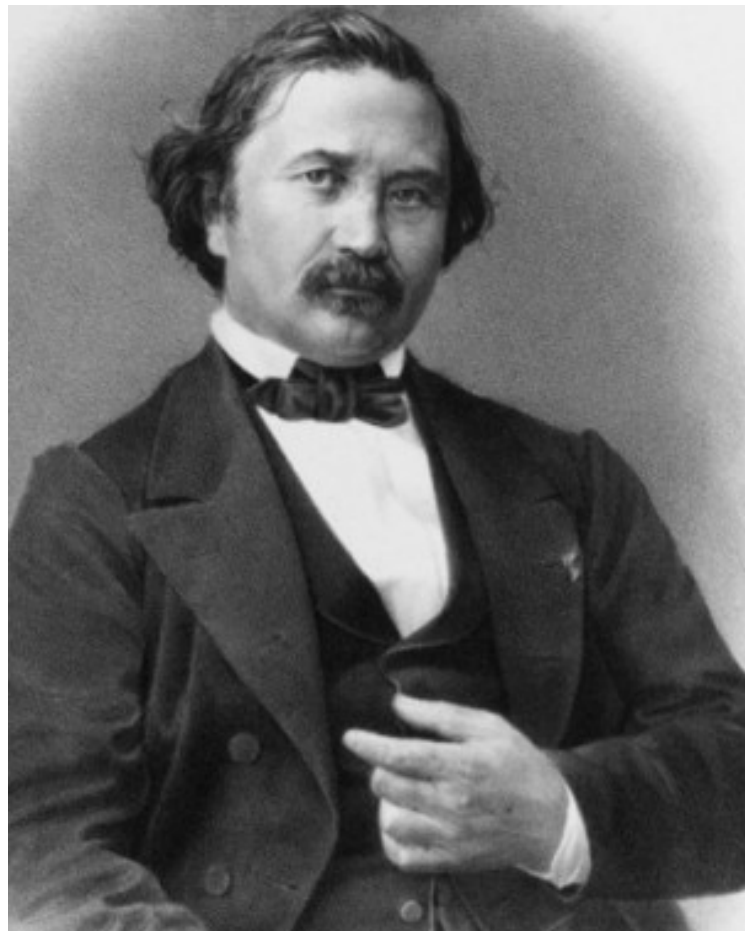
Názov: Don't just solve it; fight it! Príklady z pravdepodobnosti
Don't just solve it; fight it! Probability problems

Cieľ: Téma práce je malou obmenou citátu Paula Halmosa o dôkazoch: "Don't just read it; fight it! Ask your own questions, look for your own examples, ..." Jeho myšlienka sa dá vzťahovať aj na riešenie príkladov. To sa nemusí skončiť vyriešením určitého príkladu, mnohé príklady dávajú možnosť zamýšľať sa nad nimi ďalej, klásť si otázky a hľadať na ne odpovede. Cieľom práce je venovať sa takýmto spôsobom niekoľkým úlohám z oblasti pravdepodobnosti:

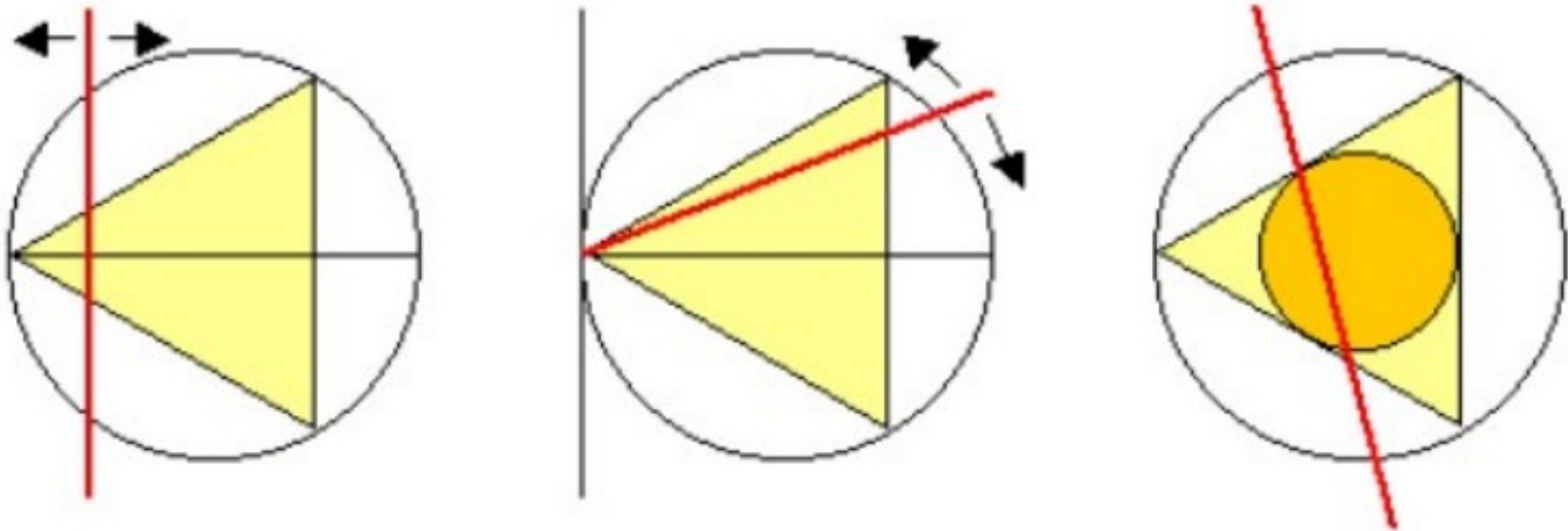
- (a) Príklad 293 zo zbierky [1] a súvislosť získaných pravdepodobností s pravdepodobnosťami získanými z ratingov typu [3]
- (b) Známy Bertrandov paradox - porovnanie s experimentom popísaným v článku [4]
- (c) Úloha o spojeniach kariet z [3] - výpočet pravdepodobnostného rozdelenia

- Bertrand (1889):

Majme kruh, v ktorom náhodne zvolíme tetivu. Aká je pp., že zvolená tetiva bude dlhšia ako strana rovnostranného trojuholníka vpísaného do daného kruhu?



- V závislosti od toho, čo znamená „náhodná tetiva“, dostaneme tri rôzne výsledky ($1/2$, $1/3$, $1/4$)



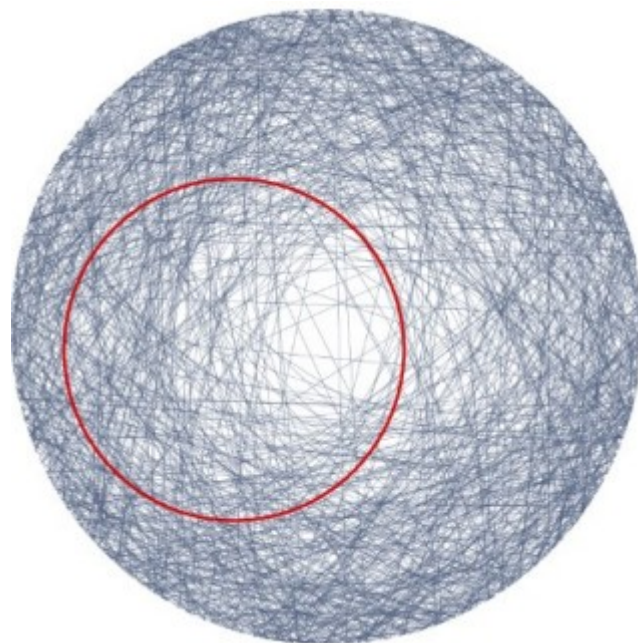
Obrázok: <http://utenti.quipo.it/base5/probabil/bertrand.htm>

Jaynes (1973)

- Rozdelenie dĺžky tetivy musí byť invariantné na polohu a veľkosť kruhu.
- Odvodil hustotu rozdelenia:

$$p(x) = \frac{x}{\sqrt{1-x^2}}, \quad (4.1.2)$$

pričom x predstavuje dĺžku normalizovanej tetivy, teda $x = \frac{L}{2 \cdot R}$ (L je dĺžka tetivy, R je polomer kruhu a $0 \leq x \leq 1$).



Obrázok:

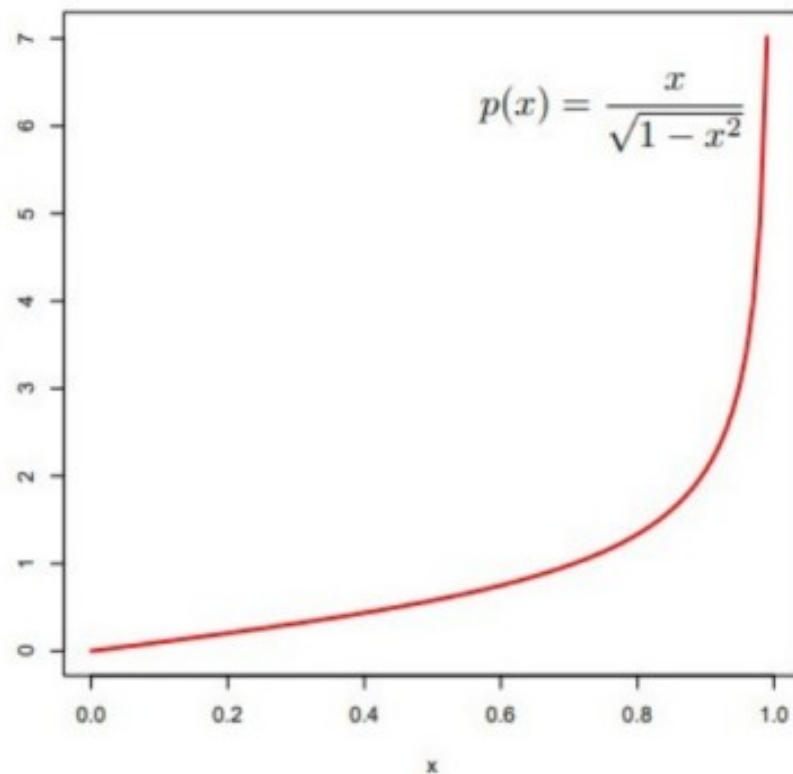
[https://en.wikipedia.org/wiki/Bertrand_paradox_\(probability\)#/media/File:Bertrand3-translate_ru.svg](https://en.wikipedia.org/wiki/Bertrand_paradox_(probability)#/media/File:Bertrand3-translate_ru.svg)

- Pravdepodobnosti z domácej úlohy:

Príklad 2 (20 bodov)

Vypočítajte pravdepodobnosti, že náhodná premenná nadobúda hodnoty uvedené v tabuľke, ak jej hustota je nasledovná (mimo intervalu (0,1) je hustota nulová):

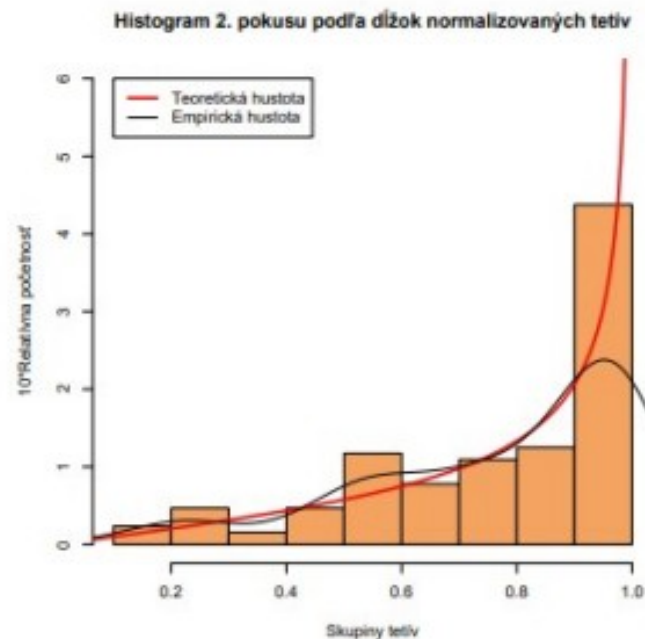
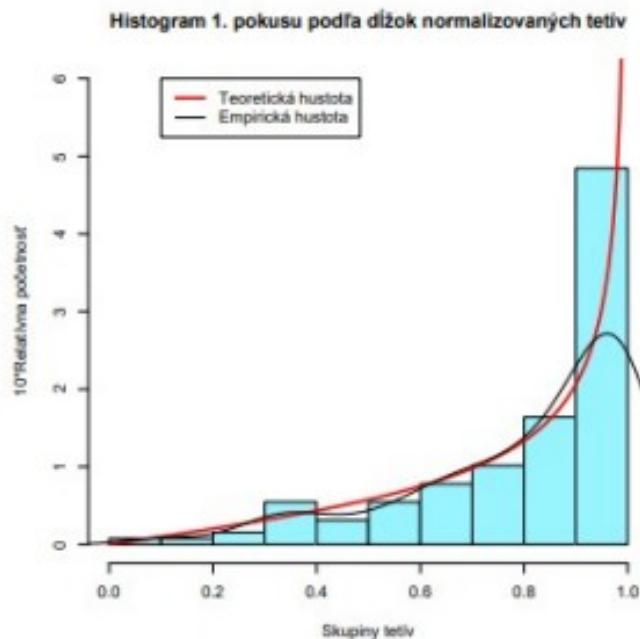
<i>hodnota</i>	<i>pravdepodobnosť</i>
medzi 0 a 0,5	
medzi 0,5 a 0,6	
medzi 0,6 a 0,7	
medzi 0,7 a 0,8	
medzi 0,8 a 0,85	
medzi 0,85 a 0,9	
medzi 0,9 a 0,925	
medzi 0,925 a 0,95	
medzi 0,95 a 0,975	
medzi 0,975 a 1	



- Experiment v bakalárke:

model. My sme sa rozhodli realizovať experiment popísaný v [15], teda hádzať slamku (v našom prípade špajdl'u) na kruh s priemerom 5 palcov = 12.7 cm. Výsledky podľa informácií z článku potvrdili nízkymi testovými štatistikami (v tomto prípade s veľkou jednoznačnosťou) rozdelenie tetív podľa T. Jaynesa, čo si teraz overíme realizáciou rovnakého experimentu. Tohto experimentu sa zúčastnili 3 ľudia, ktorí hádzali slamku zostojia na kruh nakreslený na papieri a položený na zemi. Výsledkom boli 4 pokusy so 128 meraniami. Získané merania sme následne normalizovali, teda predelili ich priemerom kruhu a rozdelili sme ich do 10 skupín.

- Výsledky:



Obr. 4.2.1: Histogramy vykreslené z dát spolu s funkciou očakávanej (teoretickej) a vypočítanej (empirickej) hustoty. Na osi x sú rovnomerne rozdelené skupiny tetív a na osi y početnosť.

- Testovanie chí kvadrát testom:

Číslo skupiny	Skupina tetív	Očak. počet
1	$(L/2R \leq 0.5)$	17.1487
2	$(0.5 < L/2R \leq 0.6)$	8.45125
3	$(0.6 < L/2R \leq 0.7)$	10.9897
4	$(0.7 < L/2R \leq 0.8)$	14.6103
5	$(0.8 < L/2R \leq 0.85)$	9.37182
6	$(0.85 < L/2R \leq 0.9)$	11.6343
7	$(0.9 < L/2R \leq 0.925)$	7.15812
8	$(0.925 < L/2R \leq 0.95)$	8.6678
9	$(0.95 < L/2R \leq 0.975)$	11.5258
10	$(0.975 < L/2R \leq 1)$	28.4422

Tabuľka 4.2.1: Druhé rozdelenie normalizovaných tetív do skupín na testovanie s ich očakávanými početnosťami (červenou) podľa hustoty 4.1.4.

- Testovanie chí kvadrát testom:
- Vypočítame kritickú hodnotu a dokončíme testovanie

	Pokus č.1		Pokus č.2	
Skupiny	Početnosť	TS	Početnosť	TS
1	15	0.2692	17	0.0012
2	7	0.2492	15	5.0745
3	10	0.0891	10	0.0891
4	13	0.1774	14	0.0254
5	10	0.0421	5	2.0393
6	11	0.0345	11	0.0345
7	6	0.1873	7	0.0034
8	6	0.8211	9	0.0127
9	10	0.2019	10	0.2019
10	40	4.6966	30	0.0853
Súčet	128	6.7688	128	7.5679