Eötvös Loránd Tudományegyetem

Természettudományi Kar

Az egyenletes eloszláshoz kapcsolódó érdekességek

Gáyer Nóra

BSc szakdolgozat

Témavezető:

Csiszár Villő

Adjunktus
Valószínűségelméleti és Statisztika Tanszék

Budapest, 2016

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Csiszár Villőnek, hogy elvállalta a konzulensi teendőket, mindig a rendelkezésemre állt, ötleteivel végig segítette a szakdolgozatom készítését. Tanácsai és útmutatása nélkül a dolgozat nem jöhetett volna létre.

Hálás köszönettel tartozom továbbá a családomnak, akik a tanulmányaim alatt végig mellettem álltak, támogattak, minden nehézség ellenére.

Tartalomjegyzék

1. Bevezetés **4**

2. Az egyenletes eloszlásról általánosan **6**

3. Véletlenszám generátorok **8**

3.1. A véletlenszámok generálásának története 8

3.2. Nem egyenletes eloszlású mintából véletlen bitek generálása 9

3.3. Álvéletlen számok generálása számítógépeken 11

3.4. Fizikai véletlenszám generátorok 12

4. Az első számjegy probléma **13**

4.1. A témakör ismertetése 13

4.2. Szemléltetés példákon keresztül 14

4.3. Elméleti háttér elemzése 20

5. Bertrand-féle paradoxon **23**

5.1. A paradoxon bemutatása 23

5.2. Magyarázat az eltérésekre 25

6. Születésnapok eloszlása **26**

6.1. Adatok elemzése 26

6.2. Befolyásoló tényezők 30

7. Egyenletes eloszlás a permutációkon **31**

8. Összegzés **35**

9. Mellékletek **40**

1. Bevezetés

Szakdolgozatomban az egyenletes eloszláshoz kapcsolódó érdekességek, paradoxonok kerülnek középpontba. Életünkben több olyan esemény is megjelenik, melyről azt gondolhatjuk, hogy teljesen a véletlennek köszönhető, egyenletes eloszlás szerint zajlik. Viszont ha elkezdünk utánajárni, megvizsgálni, számításokat végezni, hogy mi áll a hátterében, meglepő eredményekre juthatunk. A dolgozat fejezetei egymástól különböző problémákat, érdekességeket mutatnak be, melyek mind az egyenletes eloszláshoz kapcsolódnak valamilyen formában.

A második fejezet egy rövid felvezetés, áttekintés, amely a matematikai ismertetést foglalja magában. Bemutatom ebben a fejezetben, hogy milyen módon, hogy írhatjuk le ezt az eloszlást, hogy definiálhatjuk.

Ezt követően a véletlenszám generátorokkal foglalkozok a harmadik fejezetben, amely magában foglalja a történelmi áttekintést, valamint azt, hogy a technológia fejlődése mennyiben járult hozzá a véletlen számok előállításához. Több módszert ismertetek az előállításra, valamint egyet ki is próbálok, amelynek alapjául egy könyv szövege szolgál. Erre alkalmazom majd az egyik bemutatásra kerülő módszert, és várom, hogy sikerül-e egyenletes eloszlású számsorozatot előállítani. Az eredmények kiértékelése után, egy újabb fejezetben, az úgynevezett kezdőszámjegy problémával foglalkozom.

Ennek lényege, hogy különböző adathalmazokat vizsgálok, melyekben az adatok kezdőszámjegyei szolgálnak a vizsgálódás alapjául. A feltételezés az, hogy ezek eloszlása egyenletes lesz, ebből kiindulva végzem az elemzéseket. Az eredményeimet különböző grafikonokon szemléltetem, melyekből jól látható a tényleges eredmény.

Az ötödik fejezetben a Bertrand paradoxonra térek át, amely szintén az egyenletes eloszláshoz kapcsolódik. A probléma lényegében abban rejlik, hogy nem mindegy, hogy milyen alakzaton választunk ki véletlenszerűen egy pontot, mivel ez nagyban befolyásolhat több eredményt is. Ebben az esetben arra keresem a választ, hogy mekkora annak a valószínűsége, hogy egy kör véletlenszerűen kiválasztott húrja nagyobb, mint a körbe írható szabályos háromszög egy oldala.

A hatodik fejezetben arra a kérdésre keresem a választ, hogy mekkora hibát követünk el, ha az emberek születésnapjait egyenletes eloszlásúnak tekintjük. Február 29-től elte-

kintve azt gondolhatnánk, hogy az év minden egyes napján ugyanakkora gyakorisággal születnek az emberek. A fejezetben példákon keresztül mutatom be, hogy a valóságban pontosan mi is zajlik.

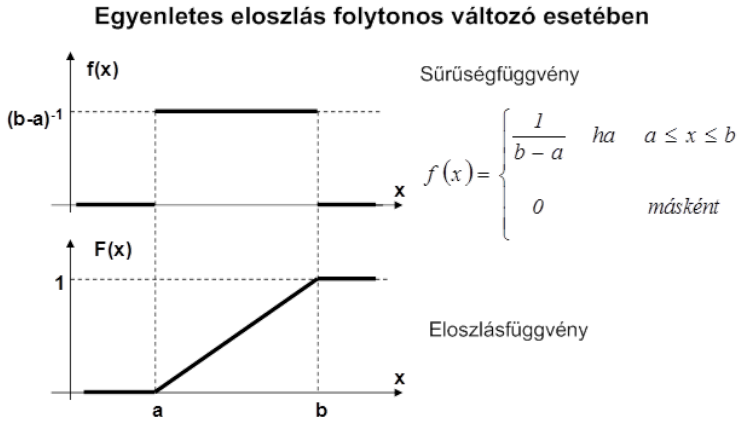
Végül pedig a permutációk kerülnek előtérbe. Mindenki abban bízik, hogy a különböző sorsolások kimenetelei, például a lottósorsolás eredménye véletlenszerű. Ebben a fejezetben egy ellenpéldát mutatok be, amely 1969-ben történt. Ebben az évben úgy próbálták igazságosan eldönteni, hogy melyik férfinak mikor kell bevonulnia a hadseregbe, hogy egy lottósorsoláshoz hasonló eseményen húztak sorszámokat, amelyekhez a születésnapjuk alapján voltak rendelve az emberek. A sorsolás végén, elemezve a kapott permutációt azonban érdekes dolgok derültek ki, melyekre a fejezetben derül majd fény.

2. Az egyenletes eloszlásról általánosan

Ahhoz, hogy megismerhessük az egyenletes eloszlást, először magukkal a véletlen eseményekkel ismerkedjünk meg. Véletlen eseményeknek nevezzük azokat az eseményeket, amelyek eredményét nem tudjuk előre megjósolni. Például ha dobunk egy szabályos dobókockával, nem tudjuk előre, hogy mi lesz az eredménye, vagy a szerencsejátékok kapcsán nem tudjuk előre például, hogy melyik számokat húzzák ki egy lottósorsolás során, vagy a rulettkerék mely számjegyéhez érkezik a golyó végül. Tulajdonképpen az egész valószínűségszámítás, mint tudományág kialakulása mögött a szerencsejátékok állnak, azok vizsgálatából indultak el a különfél kutatások. Tehát ezeknek a véletlen eseményeknek megfigyelhetünk különböző törvényszerűségeit is, amelyek már nem csak adatok számunkra, hanem információkat is hordoznak, majd ezekből építünk fel különböző modelleket.

Az egyenletes eloszlás a nevezetes eloszlások közé tartozik, folytonos esetben jelölésére az $E(a,b)$-t használjuk, ahol (a,b) jelöli a lehetséges értékek kezdő és végpontjait. Továbbá a következő módon definiálhatjuk az egyenletes eloszlást: legyen X folytonos valószínűségi változó, ezt egyenletes eloszlásúnak nevezzük az $[a,b]$ intervallumon, ha eloszlásfüggvénye:

$$F(x) = \begin{cases} 0 & \text{ha } x < a \\ \frac{x-a}{b-a} & \text{ha } a \leq x \leq b \\ 1 & \text{ha } x > b \end{cases}$$



1. ábra. Az egyenletes eloszlás eloszlás- és sűrűségfüggvénye (Huba & Lipovszki, 2014)

Sűrűségfüggvénye pedig a következő alakú:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{ha } a \leq x \leq b \\ 0 & \text{ha } x < a \text{ vagy } x > b \end{cases}$$

Szemléletesen az 1. ábrán láthatjuk az eloszlások megjelenését. A modell szavakban kifejtve pedig azt jelenti, hogy egy szakasz valószínűsége a hosszával arányos, azaz az (a, b) intervallumból "véletlenszerűen" választunk egy pontot. (Csiszár, 2009)

Abban az esetben, ha az $[a, b]$ intervallumot leszűkítjük a $[0, 1]$ intervallumra, az egyenletes eloszlás standardizált esetéről beszélhetünk. Ekkor elmondható, hogy a valószínűségek megegyeznek a hosszúságokkal a $[0, 1]$ intervallumban. (Feller, 1965)

A folytonos eset mellett a diszkrét esetet is meg kell említeni, hiszen a valóságban sokszor íly módon zajlanak le az események. Ebben az esetben van egy véges halmaz, amely a kimeneteleket írja le, a valószínűségi változó mindegyiket egyforma valószínűséggel veszi fel.

Láthattuk már az egyenletes eloszlás folytonos, diszkrét valamint standardizált esetét. Most pedig ezek bővítéseként nézzük meg a többdimenziós estet, ahol a terület többdimenziós kiterjesztéséhez is használatos mértékeket használjuk. Legyen G egy véges mérhető halmaz. Azt mondjuk, hogy X egyenletes eloszlású valószínűségi változó G-n, ha G bármely mérhető részhalmazára annak mértékével arányos valószínűséggel esik. G mértékét pedig jelöljük $\lambda(G)$-vel. Ekkor az X valószínűségi változó sűrűségfüggvénye a következőképp adható meg:

$$f(x) = \begin{cases} \frac{1}{\lambda(G)} & \text{ha } x \in G \\ 0 & \text{egyébként} \end{cases}$$

Az egyenletes eloszlás több különböző típusának bemutatása után vizsgáljunk meg néhány paradoxont, esetet, ahol ez az eloszlás megjelenhet, kérdéseket vethet fel. (TDM, 2016)

3. Véletlenszám generátorok

3.1. A véletlenszámok generálásának története

Manapság leggyakrabban a számítógépeket hívjuk segítségül amikor véletlen számok generálására van szükségünk. Azonban ez nem mindig volt így, véletlenszámokat már jóval korábban is próbáltak generálni különböző eszközök segítségével. Ezen eszközök közé tartozott például a kocka, a kártyák, az érmék, valamint a kerék is. A kockák megjelenése már időszámításunk előtt 2-3 évezredre is visszatehető, amikor még kezdetleges jelöléseket használva, kövekből, csontokból készítették őket, de a véletlen előidézésére használták. Később, Kínában kezdték használni a kártyákat ugyanezen célból, amely csak az 1300-as évek elején jutott el Európába, mintegy 700 évvel lemaradva a kínai megjelenéstől. Az ókori Rómában érméket használtak véletlen számok előállítására, de az érméknek a későbbiekben is nagy szerepe lett a valószínűségszámítás kialakulása, fejlődése során. Utolsó megemlített eszköz a kerék volt, melyet a görögök használtak, mégpedig azon az elven, hogy megadott beosztás szerint skálázták a kereket, megpörgették, és a kapott eredmény adta a véletlen számot. Mindezen módszereket a mai napig alkalmazzuk például a szerencsejátékok kapcsán, valóban véletlen eseményeket tudunk létrehozni segítségükkel, azonban az igazi áttörést csak később sikerült elérni. (Bob de Vivo, 2005)

Statisztikai sokaságot 1924-ben L. Tipett hozott létre először, melyhez népszámlálási adatokat használt fel, segítségükkel pedig 40000 véletlenszámot publikált. Ezt követően, 1939-től már célgépeket is készítettek véletlenszámok generálására, mely segíségével 1955-ben a RAND Corporation egymillió számjegyet állított elő. Ezt a későbbiekben is sokáig használták, úgynevezett véletlenszám táblázatként, azaz amikor szükség volt véletlenszámra, ebből az adatbázisból dolgoztak. Az igazi áttörést viszont a számítógépek megjelenése hozta, amikor is Neumann János ötlete alapján elkezdték a gépeket véletlenszám generálására használni, módszereket keresni. Neumann János javasolt is egy ilyen módszert, amelynek lényege, hogy a számokat négyzetre kell emelni és a középső számjegyeit kiválasztani, ezáltal mindig újabb és újabb véletlenszámokhoz jutunk. A módszer ugyan nem terjedt el, de helyette másik módokat találtak, ezek közül mutatok be néhányat a következőkben. (Nemetz & Wintsche, 1999)

3.2. Nem egyenletes eloszlású mintából véletlen bitek generálása

Ebben az esetben egy olyan adathalmaz lesz a kiindulópont, amelyben az adatok eloszlása nem egyenletes. Viszont az itt meghatározott módszerrel, végül mégis olyan véletlen számokat fogunk kapni, melyek eloszlása már egyenletes lesz. Ehhez először vegyünk egy független bitekből (olyan változók, amelyek értéke vagy 0, vagy 1) álló X_n sorozatot, amelyben $p = P(X_n = 0)$, és p se nem egyenlő 0-val, se nem egyenlő 1-gyel. Vizsgáljuk ezen sorozat részletösszegeit, amiket a későbbiekben segítségül fogok majd hívni a bizonyításhoz. Amennyiben az első n tag összege páros, akkor a részletsorozat elemei helyett írjunk 0-t, ha pedig páratlan, akkor 1-et. Az így kapott számokat S_n-nel jelöljük, valamint számítsuk ki a következő $p_n = P(S_n = 0)$ valószínűségeket.

Vizsgáljuk meg a következő feltételes valószínűséget $P(S_{n+1} = 0|S_n = 0)$. Ha $S_n = 0$, akkor S_{n+1} pontosan akkor lesz 0, ha $X_{n+1} = 0$, tehát a valószínűség n-től független, értéke pontosan p lesz. További feltételes valószínűségeket is fel tudunk írni az előzőhöz hasonlóan, ekkor ezek az esetek keletkeznek: $P(S_{n+1} = 0|S_n = 0)$; $P(S_{n+1} = 1|S_n = 0)$; $P(S_{n+1} = 0|S_n = 1)$; $P(S_{n+1} = 1|S_n = 1)$. Majd ezeket felhasználva, a keresett p_n valószínűséget egy rekurzió kapcsán meg tudjuk határozni, az alábbiak szerint:

$$p_{n+1} = P(S_{n+1} = 0) = P(S_{n+1} = 0|S_n = 0) \cdot P(S_n = 0) + P(S_{n+1} = 0|S_n = 1) \cdot P(S_n = 1)$$

$$p_{n+1} = p \cdot p_n + (1 - p) \cdot (1 - p_n)$$

Tovább rendezve az egyenletet a következőt kapjuk:

$$p_{n+1} = (2 \cdot p - 1) \cdot p_n + (1 - p)$$

Feltételezzük, hogy ez a kapott p_n számsorozat 1/2-edhez tart, azaz a $p_n - 1/2$ nullához. Ennek bizonyításához az alábbi lépéseket hajthatjuk végre:

$$p_{n+1} - 1/2 = (2 \cdot p - 1) \cdot p_n + (1/2 - p)$$

$$[p_{n+1} - 1/2] = (2 \cdot p - 1) \cdot [p_n - 1/2]$$

Ebből pedig már láthatjuk is, hogy egyértelműen következik a konvergencia, sőt a sebessége elég gyors lesz.

$$[p_{n+1} - 1/2] = (2 \cdot p - 1)^n \cdot [p_1 - 1/2]$$

További általánosítás is megfogalmazható erre az előbbi állításra. Legyenek $X_1, X_2, \ldots$ egyforma eloszlású, független valószínűségi változók, amelyek a $0, 1, \ldots, m-1$ egész számok valamelyikét veszik fel, pozitív valószínűséggel. Ekkor el tudjuk készíteni szintén a részletösszegeiket, amelyeket modulo-m számítunk. Legyen $S_n = X_1 + X_2 + \cdots + X_n$. Azt állapíthatjuk meg, hogy ezeknek a részletösszegeknek a valószínűség-eloszlása konvergál az egyenletes eloszláshoz. (Nemetz & Wintsche, 1999)

Tehát összességében pontosan azt kaptuk, amit eredetileg kerestünk. Egy tetszőleges eloszlású mintából egyenletes eloszlású elemeket állítottunk elő. A valóságban ilyen kiinduló halmaz lehet például egy szöveg, amelyet átkonvertálhatunk számok sorozatává. Minden karakterhez rendeljünk egy sorszámot, például az ABC-ben elfoglalt helyük alapján, valamint a szóközt és minden más speciális karaktert is egy-egy adott számmal helyettesíthünk. Amennyiben egy elég nagy szöveget konvertálunk át, majd meghatározzuk az egyes S_n összegeket, valóban egyenletes eloszlást fogunk kapni.

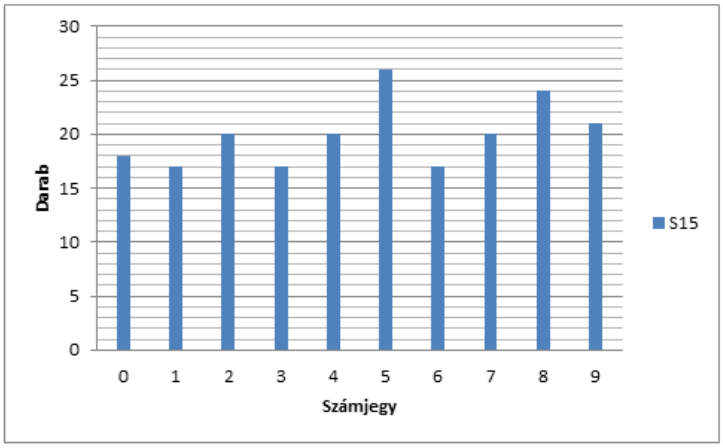
Nézzünk meg egy konkrét példát, hogy ott milyen értékeket kapunk. Első lépésben tehát szükség van egy adathalmazra, esetemben ez Az arany ember című regény első oldaláról származó 3000 darab karakter. Valamint el kellett készíteni egy kódolást az ABC minden karakterére vonatkozóan. Ezt úgy készítettem el, hogy felsoroltam az összes betűt, számot, sorszámoztam őket, majd ezen sorszámok utolsó számjegyével dolgoztam tovább. A speciális karakterekhez 1-est, a szóközhöz pedig 3-as számot rendeltem hozzá, de választhattam volna bármely másik számjegyet is, mivel bármilyen eloszlásból kiindulhatunk a feladat kiírása alapján.

Ezt követően összepárosítottam a szöveg karaktereit, valamint a karakterekhez rendelt számokat. Összesen 3000 darab ilyen szám keletkezett. Ezután már csak az összegzés maradt, melyet most 15 elemenként végeztem. Tehát így összesen $3000/15 = 200$ darab összeg keletkezett. Ezeknek szintén az utolsó számjegyére van szükségem, azokat meghatároztam, végül pedig egy táblázatba kigyűjtöttem, hogy melyik számjegy milyen gyakorisággal szerepel (2. melléklet). Ezt után ábrázoltam is, a 2. ábrán látható a végeredmény.

Ahogy az ábrán látszik is, ekkor a számjegyek eloszlása már tényleg közelít az egyenleteshez, egymáshoz közeli értékeket kaptunk, megállapíthatjuk, hogy valóban egy tetszőleges eloszlású szám (szöveg) mintából indultunk ki, majd egy algoritmus végrehajtása

S_{15} mod 10 összegzés

Számjegy	Darab
0	18
1	17
2	20
3	17
4	20
5	26
6	17
7	20
8	24
9	21
Összesen:	200
Átlag:	20



2. ábra. Az arany ember első 3000 karaktere (Saját ábra)

után egy egyenletes eloszlású adathalmazt kaptunk.

3.3. Álvéletlen számok generálása számítógépeken

Ebben a fejezetben úgynevezett álvéletlen számok előállításával foglalkozunk, amely azt foglalja magában, hogy kezdetben egy olyan sorozatunk van, amelyben az elemek a $[0, 1]$ intervallumon egyenletes eloszlású, független véletlen számok, ezeket jelöljük U_n-nel. Ekkor azt állítjuk, hogy ezekből kiindulva tetszőleges véges eloszlású, független számsorozatot létre tudunk hozni.

Legyen X valószínűségi változó, értékkészlete $x_1, x_2, \ldots, x_k$, ezek jelöljék az előállítandó elemeket. Minden x_i-hez rendeljük a következő valószínűségeket: $p_i = P(X = x_i)$, ahol $i = 1, \ldots, k$. Hasonlóan az előző fejezetben használt jelöléshez, vezessünk be a valószínűségek összegeire jelölést a következő módon:

$$u_0 = 0, u_1 = p_1, u_i = u_{i-1} + p_i, i = 1, 2, \ldots, k$$

Ekkor $X_n = x_i$, akkor és csak akkor, ha $u_{i-1} \leq U_n \leq u_i$, ami azt jelenti, hogy tényleg elegendő az egyenletes eloszlású U_n sorozatok előállításával foglalkozni. A számítógépek véges számokkal operálnak, ezért olyan véletlenszámokkal tudunk dolgozni, amelyek a $[0, 1]$ intervallum valamely nagyon sűrű ponthalmazán egyenletes eloszlásúak. Ha a ponthalmaz i/K, ahol $i = 0, 1, \ldots, K - 1$, akkor az egész számok $0, 1, \ldots, K - 1$ halmazán

egyenletes eloszlású számokat kell generálnunk, amelyhez egy V_n kongruencia sorozatot használhatunk. Ennek a sorozatnak négy paramétere van: a K, a V_0 kezdőérték, az a, mint multiplikatív konstans, valamint a c, mint additív konstans. Ezekből a sorozatot a következő rekurzióval tudjuk előállítani:

$$V_{n+1} = (aV_n + c) \mod K, 0 \leq n$$

Ebből $U_n = V_n/K$. Amire figyelni kell a konstansok megválasztásánál, ha például a, c és K közül az összes páros, akkor V_n is mindig páros lesz. Továbbá elmondható erről a sorozatról, hogy periodikus lesz, amint egyik szám ismétlődik, onnantól kezdve minden ismétlődik. Ezért olyan K-t érdemes választani, ami kellően hosszú periódust eredményez. Akkor lesz maximális a periódus, ha K kettőhatvány, c páratlan, a pedig 4-gyel osztva egy maradékot ad. (Nemetz & Wintsche, 1999)

Egy rövid példán nézzük is meg, hogy áll elő ez a sorozat. Szeretnék maximális periódust, ezért az értékek legyenek: $K = 16$, $a = 5$, $c = 7$. Ekkor a sorozat:

$$0, 7, 10, 9, 4, 11, 14, 13, 8, 15, 2, 1, 12, 3, 6, 5, 0, 7, \ldots$$

Természetesen a számítógépek ennél nagyobb kettőhatvánnyal képesek gyorsan dolgozni, amellyel ténylegesen gyorsan elő lehet állítani álvéletlen számokat.

3.4. Fizikai véletlenszám generátorok

A véletlenszám generátorok megvalósításának egyik leggyakoribb módszere ez, ezért mindenképp meg kell említenünk. Ennek lényege, hogy a véletlenszám kiszámításának alapját egyfajta természeti jelenség adja. Ez a természeti jelenség lehet valamilyen elektromos zaj, vagy rádiótechnikai eszköz is, a legtöbbször ezeket alkalmazzák a gyakorlatban. Ahogy azt az előző fejezetben láthattuk, egy tetszőleges eloszlású adatsorból kiindulva tudunk egyenletes eloszlású elemeket generálni, a módszer hasonló, a gépek ezen természeti jelenségek ingadozásait vizsgálják. (Nemetz & Wintsche, 1999)

4. Az első számjegy probléma

4.1. A témakör ismertetése

Az első számjegy problémája, amely a későbbiekben kerül részletes bemutatásra, Benford-törvény néven vált ismertté. A problémára eredetileg Simon Newcomb hívta fel a figyelmet egy 1881-ben megjelent folyóiratcikkben. Ezután, 57 évvel később Frank Benford, a General Electric Company fizikusa Newcombtól függetlenül ugyanazokra a megállapításokra jutott. Tehát a probléma nem az eredeti feltalálója nevét viseli, azonban nem ez az első ilyen eset, a történelem során többször előfordult már hasonló igazságtalanság. Newcomb felismerésének alapjául egy nagy logaritmustáblázat szolgált, amelyben észrevette, hogy az oldalak kinézete, kopottsága nem egyezik a könyv elején és végén. Sokkal használtabbak voltak az első oldalak, mint a hátsók, tehát feltételezte, hogy többet keresik fel a kisebb számokhoz tartozó értékeket, mint a nagyobbakhoz. Azonban Newcomb nem adott pontos értékeket, csak a feltételezést. (Raimi, 1976)

Benford azonban elkezdett vizsgálódni, kezdetben több különböző adattáblázatot vizsgált meg, összeszámolta a különböző kezdőértékekhez tartozó mennyiségeket. Vizsgálódásához felhasznált táblázatok között volt többek között a fizikai állandók táblázata, néhány természetes szám egész hatványainak táblázata, valamint népességstatisztikai táblázatok is. Megfigyelései során azt tapasztalta, hogy a táblázatokban szereplő számok nem egyforma valószínűséggel veszik fel az $1, 2, \ldots, 9$ értékeket, hanem a legelső számjegyet veszik fel a legnagyobb gyakorisággal, az utána következőket pedig egyre kevésbé. Ezek után összegezte az eredményeit, amelyben a következő (1. táblázat) gyakoriságok jöttek ki az egyes számjegyekre (a 0-t nem tekintjük kezdőszámjegynek).

Első szám	1	2	3	4	5	6	7	8	9
Gyakorisága	0,306	0,185	0,124	0,094	0,080	0,064	0,051	0,049	0,047

1. táblázat. Benford adatai (Raimi, 1976)

Tehát a megérzés, hogy nem egyenletesen oszlanak el a számjegyek, igazolódni látszik. Viszont felmerül a kérdés, hogy akkor milyen arányban jelennek meg az egyes számjegyek. Benford erre is megadott egy képletet, amelyet Benford-törvényként is említenek:

$\log(n+1) - \log n$, ahol a logaritmus 10-es alapú. Ezen képlet alapján pedig a 2. táblázatban szereplő értékek mutatják az egyes számjegyek elméleti gyakoriságát.

n	1	2	3	4	5	6	7	8	9
Törvény	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

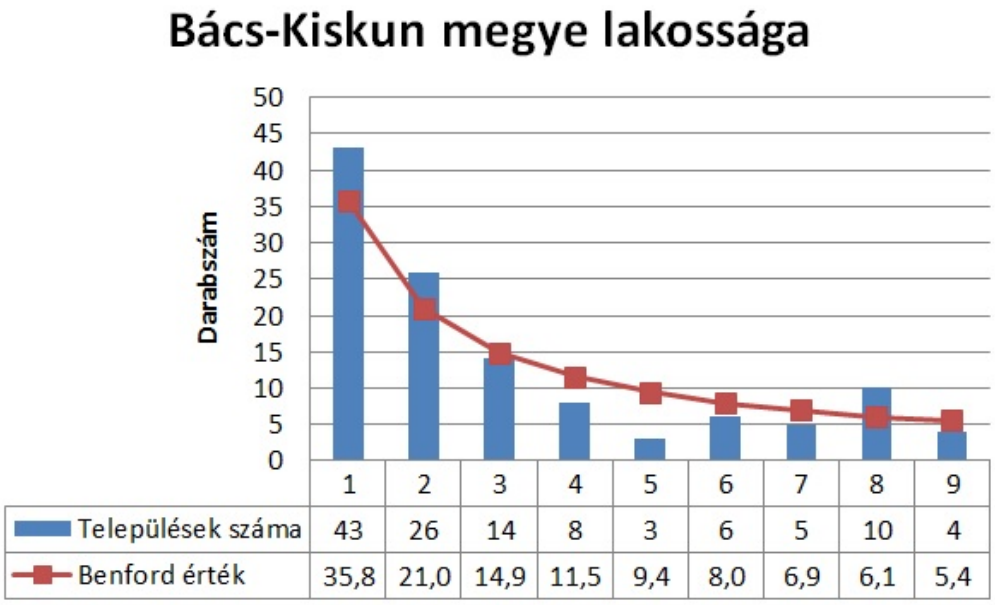
2. táblázat. Benford-törvény (Raimi, 1976)

Fontos hangsúlyozni, hogy ez nem azt jelenti, hogy minden táblázatban ilyen gyakorisággal szerepelnek a kezdőszámjegyek, hiszen olyan táblázatot bárki készíthet amiben például egyáltalán nincs 1-es, vagy 2-es számjegy. Annyit jelent, hogy tipikusan ilyenek a táblázatok. (Székely, 2004) A paradoxon tehát ott jelenik meg ebben a problémában, hogy először sokan érezhetjük úgy, hogy a számszerűsített adatok kezdőszámjegyei egyenletesen oszlanak meg egy táblázatban. Azonban a gyakorlat nem ezt mutatja, több esetet elemezve, több matematikus jutott arra az eredményre, hogy egy másfajta összefüggésnek kell megjelennie. További érdekességnek mondható, hogy ha az első számjegy után elkezdjük a második, harmadik, stb. számjegyeket vizsgálni, akkor azokban már kevésbé érvényesül ez a törvény, ott a számjegyek eloszlása egyre jobban közelít az egyenleteshez. Mielőtt kitérnék a matematikai háttér bemutatására, nézzünk meg két konkrét elemzést, melyben arra keresem a választ, hogy az adatok valóban a Benford által felállított elmélet szerinti gyakorisággal jelennek-e meg.

4.2. Szemléltetés példákon keresztül

Az első példa alapjául Bács-Kiskun megye településeinek állandó népesség száma szolgál, ahol a megye összes települését figyelembe veszem. Ehhez az adatokat a Központi Statisztikai Hivatal honlapjáról értem el (KSH,2011), amely alapján megállapítom, hogy a megye 22 darab városból, 97 darab községből áll, azaz összesen 119 darab település szolgál az elemzés alapjául. Ez nem egy nagy érték, azonban átlátható, így a szemléltetésre megfelelő. A települések lakosságának száma a 2011-es népszámlálsból származik, az adatokat Excelben kezelem, melyekből egy részlet megtekinthető az 1. számú mellékletben.

Egy egyszerű BAL elnevezésű függvény segítségével gyorsan megkapjuk minden település lakosság-számának kezdő értékét. Ezután csak egy DARABTELI függvénnyel összegezni kell a kapott értékeket, melyek eredményét a következő grafikon segítségével mutatom be. A grafikonon piros pontokkal és egy azokat összekötő vonallal szombolizáltam a Benford által várt értékeket. Emellett pedig kék oszlopok jelölik a ténylegesen megjelenő értékeket. A diagram alatt lévő sorban láthatjuk, hogy melyik számjegyhez tartozó értékeket jelölik az egyes oszlopok, valamint alatta a pontos értékek kerültek kiírásra.



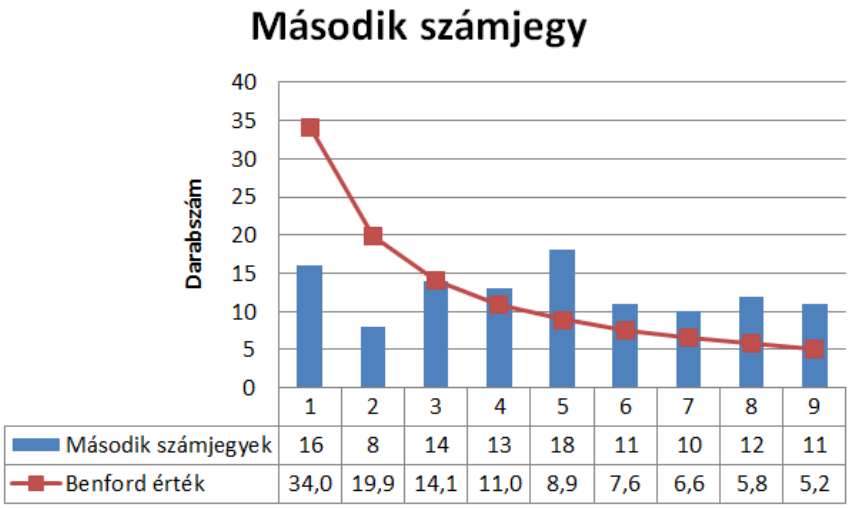
	1	2	3	4	5	6	7	8	9
Települések száma	43	26	14	8	3	6	5	10	4
Benford érték	35,8	21,0	14,9	11,5	9,4	8,0	6,9	6,1	5,4

3. ábra. Bács-Kiskun megye (Saját ábra)

Láthatjuk, hogy csak egyetlen esetben van pontos egyezés a várt és a tényleges érték között, azonban a többi érték is elég közelinek tekintető. A felhasznált adatok mennyisége nem volt számottevő, ezért nem is várhatjuk el, hogy tökéletesen illeszkedjen a modellünkhöz, azonban arányaiban már ennyi adat mellett is látszik, hogy közelítenek az értékek.

A feladat felvezetésében említettem, hogy ez a törvény csak az első számjegy esetében igaz, a további számjegyek vizsgálata során az egyenletes eloszláshoz jobban közelítő értékekkel találkozunk. Azt fogom vizsgálni, hogy ebben a példában a második és a harmadik számjegy esetén hogy alakul az eloszlás. Mindegyiket egy-egy grafikonon szemléltetem az

átláthatóság kedvéért.

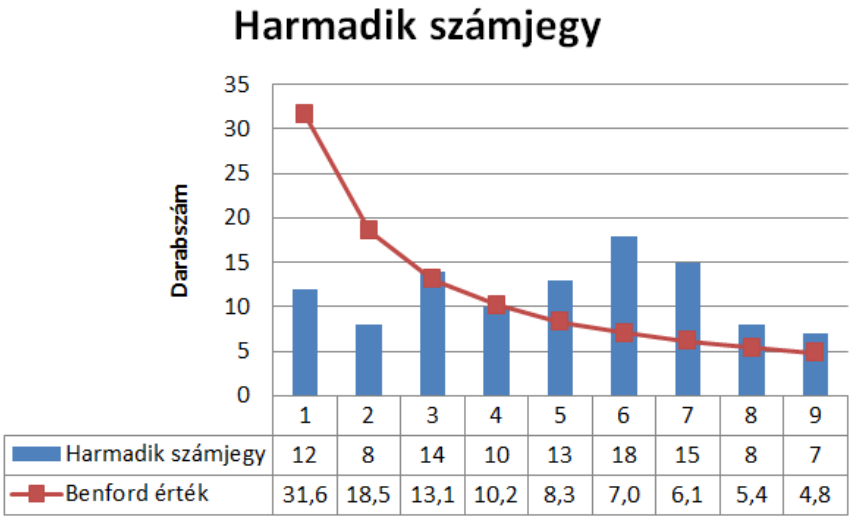


4. ábra. Benford, második számjegyre (Saját ábra)

Az első lényeges különbség, hogy a második számjegynél már nem csak kilenc számjegy jöhet szóba, hanem a 0 is, hiszen itt már nem kirtérium, hogy az ne szerepeljen. Azonban a Benford-törvényt nem tudjuk erre kiszámítani, mivel a $\log 0$-t nem értelmezzük. Ezért ebben az esetben is csak az egytől kilencig terjedő értékeket vizsgálom. Emiatt kevesebb adatunk is lesz, már csak 113 darab került ábrázolásra. Egyértelműen látszik, hogy már nem követik az adatok a törvény által meghatározott vonalat, sokkal egyenletesebbek, az első számjegynél nem szerepel jelentős kiugrás, az utolsók pedig szintén elég gyakoriak. Ebből is láthatjuk, hogy valóban csak az első számjegyre teljesül, a továbbiakra már kevésbé igaz.

Végül pedig a harmadik számjegyet is megvizsgálom, hogy ott milyen lesz az eloszlás. Ahogy az 5. ábrán látszik, ez sem közelít a Bennford-törvény által várt eredményre, ugyan ingadozó az eremény, de annyi biztosan látszik, hogy az egyenletes eloszláshoz nagyobb mértékben közelít, mint a Benford-törvény által meghatározott értékekhez. Számításokkal pontosítani is lehet, hogy milyen mértékű a kapcsolat.

Ahhoz, hogy össze tudjuk veti a második és a harmadik esetet, meg kell vizsgálni, hogy melyiknél nagyobb mértékű az eltérés. Ehhez Khi-négyzet próbát alkalmazok, mivel egy



5. ábra. Benford, harmadik számjegyre (Saját ábra)

diszkrét valószínűségi változó eloszlását szeretném vizsgálni. Eredményül a 3. táblázatban szereplő p-értékeket kaptam. A táblázatban azt látjuk, hogy a rendelkezésre álló adataink

	Benford-törvény	Egyenletes eloszlás
Első számjegy	$0,148$	$2,91 * 10^{-19}$
Második számjegy	$9,42 * 10^{-7}$	$0,639$
Harmadik számjegy	$9,56 * 10^{-9}$	$0,307$

3. táblázat. Khi-négyzet próba (Saját táblázat)

összegzése milyen szoros kapcsolatban áll a Benford által meghatározott számokkal, valamint az egyenletes eloszlással. Amint láthatjuk, valóban az első számjegy egyértelműen a Benfordhoz, míg a másik kettő az egyenletes eloszláshoz áll közelebb, tehát valóban Benford állítása csak az első számjegyre vonatkozik.

Az előző példa után még egy esetben megvizsgáltam az első számjegyek előfordulásának gyakoriságát. Ebben a példában saját magam gyűjtöttem össze az adatokat, én készítettem el a táblázatot. Ehhez először egy adatforrást kellett keresnem, amely megfelelő mennyiségű adattal rendelkezik, valamint alapvetően azzal az elképzeléssel élnénk vele szemben, hogy a vizsgált témakörben az első számjegyek előfordulásának gyakorisága

egyenletesen oszlik el. Egy napjainkban sokak által használt közösségi oldalt választottam ehhez, amely rendelkezik egy olyan funkcióval, hogy nem csak a napjainkban élő társainkkal vehetjük fel a kapcsolatot, hanem a rendszer segítségével kifejezhetjük szimpátiánkat elérhetetlenebb, híresebb személyekkel szemben is. Ez a közösségi oldal a Facebook, amelyben többek közt híres, $20-21$. századi közéleti-, sport-, politikai-szereplők adatlapjait vizsgáltam, az alapján, hogy hány ember kedveli/követi őket. A híres személyek kiválasztásához egy amerikai portál által elkészített toplistát használtam, amelyben 127 személy van feltűntetve. Minden személyhez kikerestem a pontos értékeket, majd az előző példához hasonlóan elkészítettem a grafikont, a számításokat. Ez alapján a 6 áb-

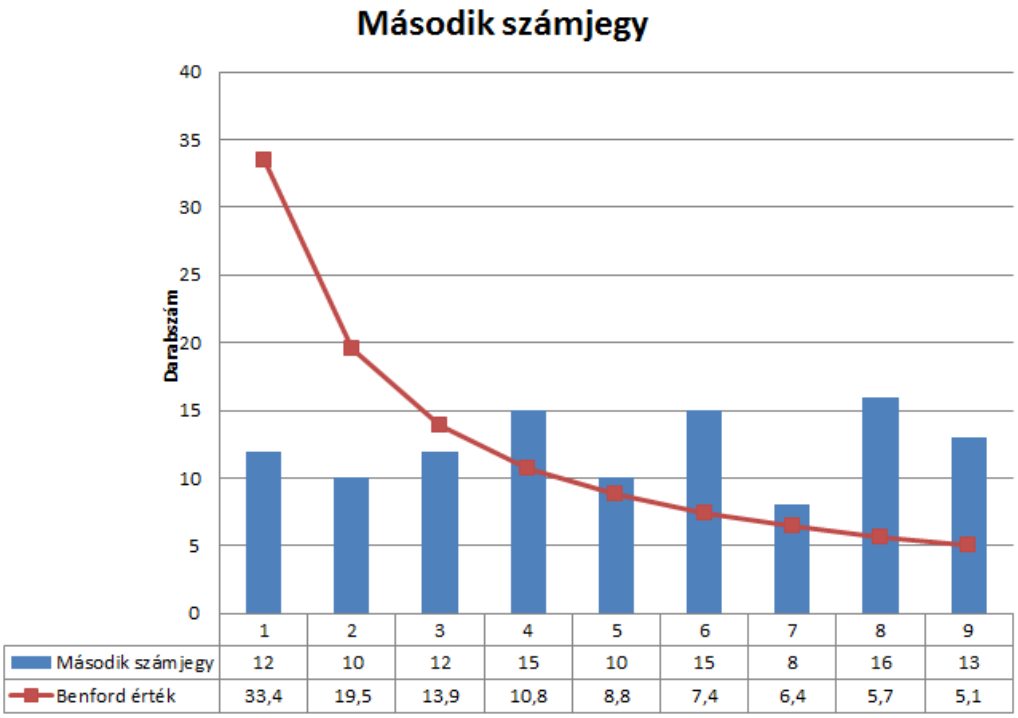
	1	2	3	4	5	6	7	8	9
Híresség esete	44	25	11	12	11	9	9	4	2
Benford érték	38,2	22,4	15,9	12,3	10,1	8,5	7,4	6,5	5,8

6. ábra. Top 100 híres ember (Saját ábra)

rán szereplő eredmények születtek. Ebben a példában már elmondható, hogy hasonlóan pontos értékeket kaptunk, mint az első esetben, egyáltalán nem az egyenletes eloszláshoz közelítenek az értékek.

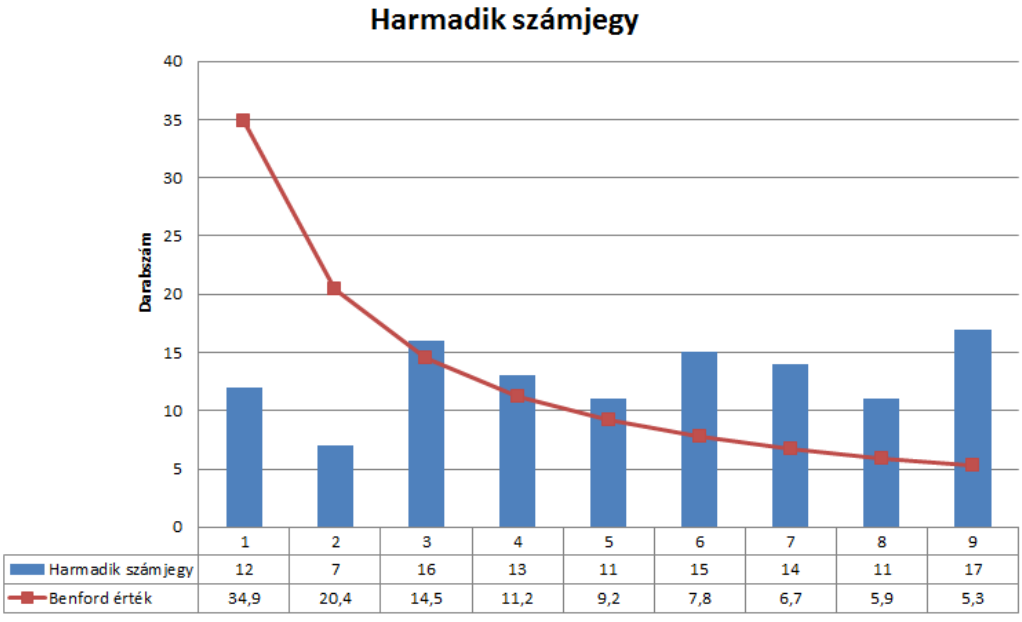
Ebben a példában is megvizsgálom a második (7. ábra), harmadik (8. ábra) számjegyeket, hasonlóan az előző példához, majd kiszámítom Khi-négyzet próba segítségével az eloszlások illeszkedésének jóságát.

A Khi-négyzet próbára kapott eredményeket szintén táblázatban (4. táblázat) foglaltam össze. Hasonló eredmények születtek, mint az első példában, azt láthatjuk, hogy



	1	2	3	4	5	6	7	8	9
Második számjegy	12	10	12	15	10	15	8	16	13
Benford érték	33,4	19,5	13,9	10,8	8,8	7,4	6,4	5,7	5,1

7. ábra. Benford, második számjegyre (Saját ábra)



	1	2	3	4	5	6	7	8	9
Harmadik számjegy	12	7	16	13	11	15	14	11	17
Benford érték	34,9	20,4	14,5	11,2	9,2	7,8	6,7	5,9	5,3

8. ábra. Benford, harmadik számjegyre (Saját ábra)

itt az első számjegy már sokkal jobban közelít a Benford értékekhez, mint az előző esetben, valamint a második és a harmadik számjegyek is jóval közelebb állnak az egyenletes eloszláshoz.

	Benford-törvény	Egyenletes eloszlás
Első számjegy	$0,5779$	$5,053 * 10^{-17}$
Második számjegy	$5,37 * 10^{-10}$	$0,7888$
Harmadik számjegy	$6,64 * 10^{-12}$	$0,6685$

4. táblázat. Khi-négyzet próba (Saját táblázat)

4.3. Elméleti háttér elemzése

A probléma matematikai hátterének vizsgálata nem egy egyszerű feladat, különböző matematikusok különböző magyarázatokat adtak az elmúlt évtizedekben, melyek között több áttér egyfajta filozófiai megközelítésbe. Benford például azt a magyarázatot adta, hogy egyfajta ellentét van az emberi számolás és a természeti jelenségek között, hiszen az emberek aritmetikusan számolnak, míg a természet logaritmikus alapokon nyugszik. Ebből kifolyólag azt állítja, hogy a természetben előforduló adatok geometriai sorozatok keverékei, amelyekre teljesülnie kell az általa készített törvénynek.

Ezt a filozófikus következtetést megelőzően azonban különböző számításokat is végezhetünk, hogy megtudjuk miért született ilyen magyarázat. Hívjuk segítségül a 2-es szám hatványait. Ha az első n hatványát vizsgáljuk, akkor a 2^n szám első számjegye potosan akkor lesz 1-es, ha létezik s egész szám, hogy

$$10^s < 2^n < 2 \cdot 10^s$$

(Székely, 2004) Vegyük ennek az egyenlőtlenségnek a 10-es alapú logaritmusát, majd alakítsuk át a következők szeint:

$$s < n \cdot \log 2 < \log 2 + s$$

$$(n-1)\log 2 < s < n \cdot \log 2$$

Ebből láthatjuk, hogy ez akkor teljesül, ha a $\log 2$-nek $n-1$ és n-szerese között szerepel egész szám, ami pont minden $\log 2$-ik esetben teljesül. Ez azt jelenti, hogy minden $\log 2$-edik esetben kezdődik 1-es számjeggyel a 2 hatvány. Hasonlóan látható, hogy a legfeljebb k-val kezdődő kettőhatványok aránya körülbelül $\log(k+1)$.

Egyelőre még csak egy példán keresztül láttuk az 1-es számjegy problémájának előfordulását, ezt tovább általánosíthatjuk. Legyen X pozitív valószínűségi változó, ekkor ennek az első számjegye akkor lesz legfeljebb k, ha létezik s, hogy

$$10^s \leq X < (k+1)10^s$$

amelyet átalakítva a következő alakra jutunk:

$$s \leq \log X < \log(k+1) + s$$

Benford törvénye akkor teljesülne, ha $\log(k+1)$ annak a valószínűsége, hogy $\log X$ törtrésze legfeljebb $\log(k+1)$. Ennek elégséges feltétele, hogy a $[0,1]$ intervallumon egyenletes eloszlású legyen $\log X$ törtrésze. Innentől kezdve pedig azt kéne vizsgálni, hogy melyek azok a tulajdonságai X-nek, amelykből következik, hogy logaritmusának törtrésze egyenletes eloszlású, valamint, hogy az adattáblázatok miért rendelkeznek ezekkel a tulajdonságokkal. Ezen kérdések megválaszolására részben vannak matematikai megoldások, részben pedig csak filozofálgatások a természet rendjéből kiindulva, amelyek a korábbiakban kerültek említésre.

További érdekességeket vehetünk észre a feladat különböző vizsgálatai során. Tegyük fel, hogy az adattáblázatunk értékei valamilyen pénznemben vannak megadva. Ekkor felmerülhet a kérdés, hogy mi történik, ha áttérünk egy másik pénznemre. Azt gondolnánk, hogy ekkor teljesen megváltozik az eddigi rendszer és a kezdőszámjegyek arányai megváltoznak. Azonban gondoljunk bele, hogy pontosan mi is történik. Legyen e, euróban megadott tőkénk, ekkor azokban az években kezdődik legfeljebb k-val a pénzünk, ha van olyan s egész szám, melyre az

$$n \cdot \log e - \log(k+1) < s < n \cdot \log e$$

egyenlőtlenség érvényes. Ez egy $\log(k+1)$ hosszú intervallum, tehát az ilyen n-ek aránya kb. $log(k+1)$ lesz. Ha most átváltjuk a pénzünket, akkor y-nal szorzunk, ekkor azokban

az években kezdődik legfeljebb k-val a pénzünk, ha van olyan s egész szám, melyre az

$$n \cdot \log e - \log(k+1) + \log d < s < n \cdot \log e + \log d$$

egyenlőtlenség teljesül. Ez is egy $log(k+1)$ hosszú intervallum (csak eltoltuk az intervallumot $\log y$ -nal), tehát az ilyen n-ek aránya szintén kb. $log(k+1)$ lesz. Az érdekesség tehát, hogy a kezdőszámjegyeknek ez egy olyan eloszlása, amely mértékegység átváltás után is érvényben marad.

További érdekesség is megfigyelhető ezen példa kapcsán. Tegyük fel, hogy választunk egy természetes számot 1 és n között, mindegyiket ugyanakkora, $1/n$ eséllyel. Azoknak a számoknak az arányát, amelyek 1-essel kezdődnek jelöljük p_n-el. Így, ha létezik ezen p_n számoknak p határértéke ($n \to \infty$), akkor azt mondanánk, hogy "egy véletlenül kiválasztott természetes szám kezdőszámjegye p valószínűséggel lesz 1-es". Azonban jobban megfigyelve ezt a sorozatot, azt láthatjuk, hogy nem lesz konvergens, hanem egyfajta hullámzást fog követni, ahol a hullámok egyre szélesebbek lesznek. Kiszámíthatjuk ezeknek a hullámhegyeknek a magasságát és a mélységét is. Akkor lesz a legmagasabb, olyan n-ek esetén, hogy $n = 2 * 10^s - 1$. Ilyen számok például az $1, 19, 199, 1999$,stb. Ezekre $p_n = \frac{10^{s+1}-1}{9\cdot(2\cdot 10^s-1)} \approx \frac{5}{9} = 0,555$. A mélysége pedig az $n = 10^{s+1} - 1$ számok esetén lesz a legnagyobb, amikor $p_n = \frac{1}{9} = 0,1111$. Innentől pedig a p_n számlálója és nevezője is egyesével növekszik a következő csúcsig. Ha vesszük ezeknek az értékeknek az átlagát, az pont 30% lesz, amely pedig megyegyezik a Benford-törvényben meghatározott értékkel. (Csiszár)

5. Bertrand-féle paradoxon

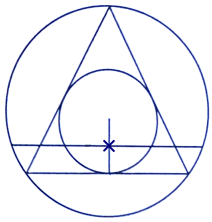
5.1. A paradoxon bemutatása

Ebben a fejezetben a korábbiakkal ellentétben nem kombinatorikai, hanem geometriai módszerekkel dolgozunk a valószínűségek meghatározásánál. Ez egy másfajta szemléletmódot igényel, máshogy kell megfogalmazni az egyes problémákat. Először Georges Buffon 1777-ben megjelent "tűproblémával" foglalkozó cikkében találkozhattunk a valószínűségszámítási problémák geometriai megfogalmazásával. Az ilyen esetekben azzal a feltételezéssel élünk, hogy a véletlen pontok, amelyeket vizsgálni szeretnénk, egy adott tartományban véletlenszerűen oszlanak el. Az elmúlt évszázadok során több probléma volt már a geometriai valószínűségszámításhoz köthető, mivel viszonylag új irányzatnak számított a matematika történelmében. Kiemelve egy ilyen problémát, vegyünk egy céltáblát, amelynél azt vizsgáljuk, hogy mennyi a valószínűsége, hogy pontosan a közepébe találunk bele. Mivel a középponton kívül még végtelen sok pont van a táblán, ezért annak a valószínűsége, hogy pont a középsőt találjuk el 0. Nem csak egyetlen pont eltalálásának a valószínűsége 0, hanem tetszőleges mennyiségű (de véges sok) pontot is pont ugyanannyi, azaz 0 valószínűséggel lehet eltalálni, hiszen rajtuk kívűl még szintén végtelen sok pont van a táblán. Tehát a véges ponthalmaz 0 mértékű. (Székely, 2004)

Nézzük, hogy most mi is jelenti a paradoxont. Először is válasszuk ki véletlenszerűen egy adott kör valamelyik húrját. Ezt követően pedig arra a kérdésre keressük a választ, hogy mekkora annak a valószínűsége, hogy ez a húr nagyobb lesz, mint az eredeti körbe írható szabályos háromszög egy oldala. A kérdésre három különböző megközelítésből három különböző válasz is adható, melyekben a közös, hogy az egyenletes eloszlás alapján választjuk a pontokat. Nézzük mi ez a három módszer.

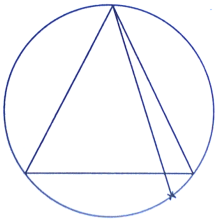
Első esetben a körben választunk ki egy pontot véletlenszerűen. Ekkor ez a pont egyértelműen meg fogja határozni a kör egy húrját, mégpedig úgy, hogy a választott pont legyen a húr középpontja (9. ábra). A következő kérdés az, hogy mikor lesz ez a húr hosszabb, mint a körbe írható szabályos háromszög oldala. Erre a válasz könnyen megadható, akkor lesz hosszabb a húr, ha a középpontja a háromszögbe írható körbe esik. Tehát már csak azt kell meghatározni, hogy a keletkezett két kör területének aránya mekkora. A kisebb

körnek pontosan fele akkora a sugara, mint a nagyobbnak, tehát a területe negyedakkora lesz. Ebből következik, hogy az eredeti nagy körben kiválasztott pont $\frac{1}{4}$-ed valószínűséggel esik a kis körbe. Tehát a valószínűség amit szererttünk volna kiszámítani, ezzel a módszerrel $\frac{1}{4}$-ed lesz.



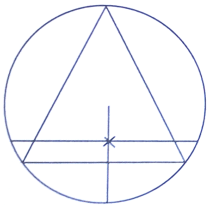
9. ábra. Véletlen pont a körlapon (Székely, 2004)

Nézzünk egy másik megoldási módot. Most ne a húr középpontját határozzuk meg véletlenszerűen, hanem az egyik végpontját. Mivel szimmetrikus az alakzat, az egyik végpontját rögzíthetjük a húrnak, a másikat pedig véletlenszerűen választhatjuk ki a köríven. Feltehetjük, hogy a rögzített végpont pontosan a háromszög egyik csúcsában található, ekkor pontosan akkor lesz hosszabb a húr az oldalaknál, ha a háromszögön belül halad (10. ábra). Mivel az egyértelmű, hogy három egyenlő részre osztják a kört a háromszög csúcsai, így a keresett valószínűség ebben az esetben $\frac{1}{3}$-ad lesz.



10. ábra. Véletlen pont a körvonalon (Székely, 2004)

A harmadik esetben pedig vegyük a kör egy sugarát és ezen válasszunk véletlenszerűen egy pontot. Ezután azt a húrt vizsgáljuk, amelyik áthalad ezen a ponton és merőleges a sugárra. Ekkor a húr pontosan akkor lesz hosszabb a háromszög oldalánál, ha a sugárnak a külső felére esik (11. ábra). Mivel szimmetrikus az alakzat, a kör bármely sugarára igaz ez az állítás. Tehát ebben az estben a kérdéses valószínűség már $\frac{1}{2}$-edre nő. (Székely, 2004)



11. ábra. Véletlen pont a kör sugarán (Székely, 2004)

5.2. Magyarázat az eltérésekre

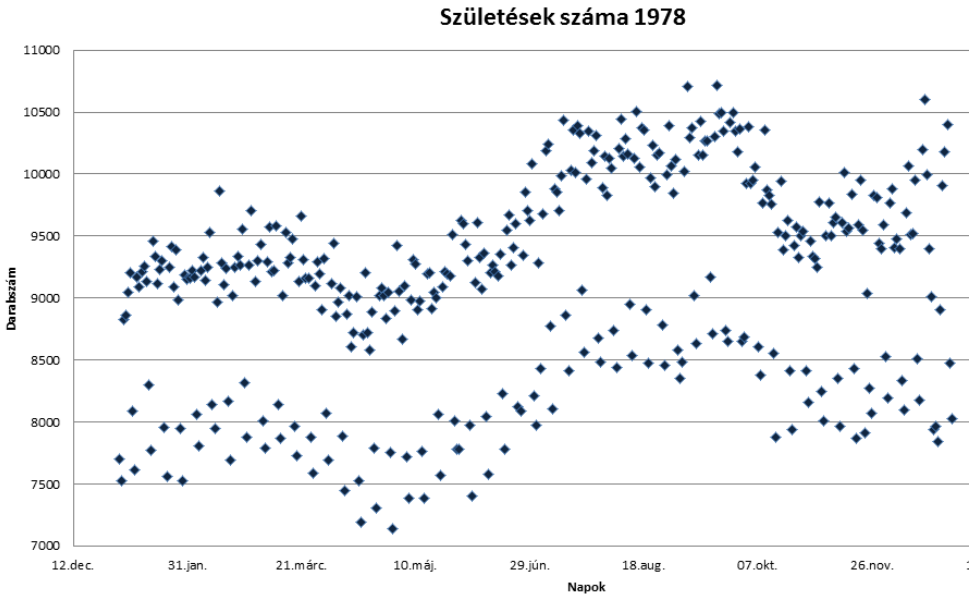
A fenti példákban látszólag minden esetben egyenletes eloszlás alapján választottuk ki a húrt, azonban mégis eltérő eredményeket kaptunk végeredményben. Mi állhat ennek hátterében, melyik a helyes? Attól, hogy találomra választunk egy pontot, még nem jelenti azt, hogy utána egyforma valószínűségeket kapunk, mivel az is fontos, hogy hol választjuk ki ezeket a pontokat. Egyik esetben a körlapon, másik esetben a körvonalon, végül pedig a kör egyik sugarán választottuk ki a pontot, ezek azok a különbségek amelyek a végeredményt meghatározzák, az nem elegendő, hogy egyenletes eloszlás alapján választunk. Tehát mindegyik kiszámított eredmény helyes a felállított rendszerben. Ennek ellenére a harmadik felfogás elfogadása tűnik a legjobb választásnak, mivel itt jelenik meg a mozgásinvariáns geometriai valószínűség, azaz, ha a húroknak két halmaza egybevágó geometriailag, akkor ugyanakkora a valószínűsége annak, hogy az egyik halmazból választunk ki egy húrt, mint annak, hogy a másikból.

6. Születésnapok eloszlása

6.1. Adatok elemzése

Egy érdekes gyakorlati példa vizsgálata kerül előtérbe ebben a fejezetben. Azt fogom vizsgálni, hogy az emberek születésnapja mennyire egyenletes eloszlású, mekkora hiba annak tekinteni. Mindenféle vizsgálódás nélkül, a megfigyeléseink alapján azt feltételezhetjük, hogy néhány eltéréstól (pl. február 29) eltekintve közel egyenletesen oszlanak el a születésnapok, amennyiben megfelelő nagyságú az adathalmazunk. Nézzünk néhány konkrét elemzést meglévő adatokra.

A legtöbb rendelkezésre álló adat az amerikai lakosságra vonatkozóan érhető el, ezek közül is főként az $1978-1994$-es időszakra vonatkozóan. Az első két, számszerűen elérhető adathalmaz konkrétan az 1978-as és az 1994-es évekre vonatkozik. Ezek közül először vizsáljuk meg konkrétan, hogy az 1978-ban született emberek születésnapjai mely napokra milyen gyakorisággal esnek (12. ábra).

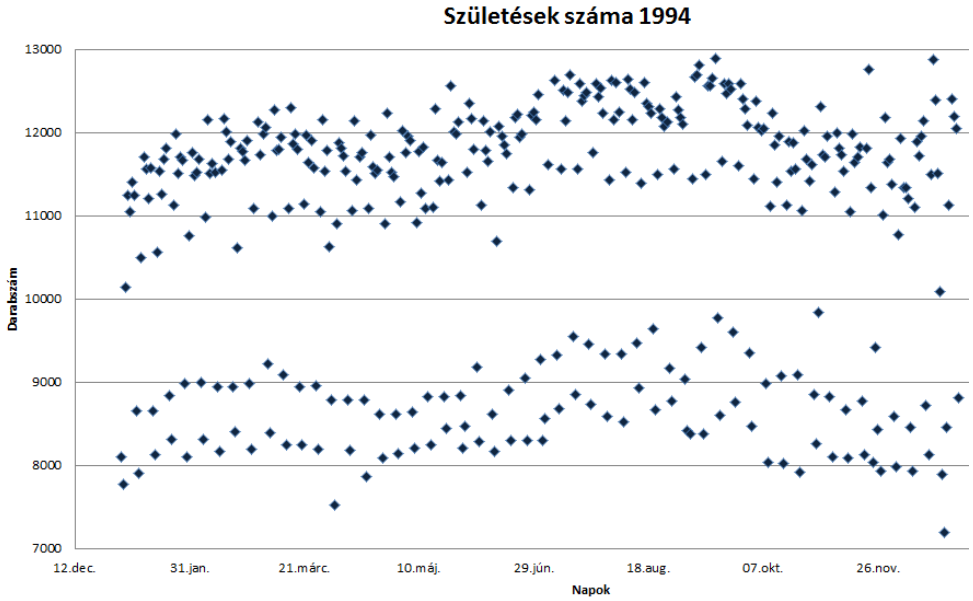


12. ábra. 1978-ben élve születettek száma (Saját ábra)

Az ábra első érdekessége, hogy két fő részre csoportosulnak az adatok. Ez nem azért van így, mert két külön adathalmazzal dolgoztam, hanem erre a magyarázat a hétköznapok és

a hétvégék közti eltérés. Érdekes módon a hétvégéken arányaiban sokkal kevesebb ember született, mint a hétköznapokon. Emellett az egész évet átfogóan tekintve észrevehetünk egy ingadozást is, láthatjuk, hogy nem egyenletes az eloszlás, júliustól szeptember végéig magasabbak az értékek, a tavaszi hónapokban pedig jóval alacsonyabbak. (Gleich, 2009)

Nézzünk meg azonban egy másik példát is, 16 évvel később vajon hasonlóan alakul-e az eloszlás. A 13. ábrán az 1994-ben születettek darabszámait láthatjuk napok szerint.

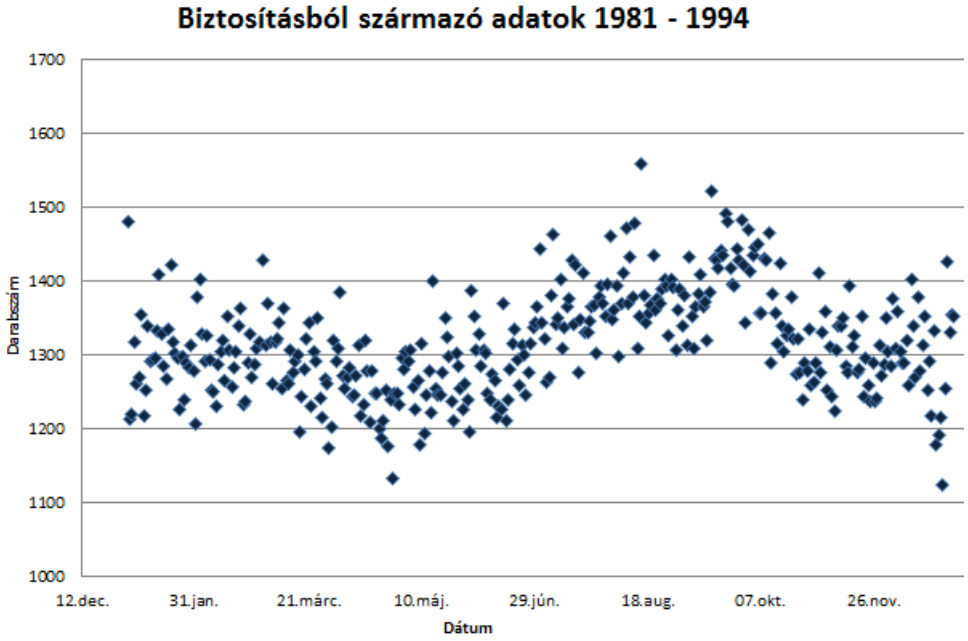


13. ábra. 1994-ben élve születettek száma (Saját ábra)

Amit ebben az esetben is elsőre észrevehetünk, az az, hogy két csoportra bomlanak a pontok, amelynek hátterében itt is szintén a hétvégék állnak. Ha az egész évet nézzük, akkor már nem találunk olyan nagy ingadozásokat, mint az előző esetben. Ebben az évben is észre vehetjük, hogy a nyár végi, ősz elején lévő hónapokban magasabbak az értékek, viszont nincs olyan jelentős eltérés a tavasziakhoz képest. Azonban egyenletesnek itt se mondhatjuk az eloszlást, már csak a hétvégi kiugró értékek miatt sem. (Gleich, 2009)

A következő vizsgált adathalmaz (Murphy,2016) nem egy konkrét évre vonatkozik, hanem olyan személyek születésnapjait tartalmazza, akik 1981 és 1994 között kötöttek egyfajta biztosítást. Így összesen 480040 darab adat áll rendelkezésünkre. Ha feltételezzük, hogy egyenletes a születésnapok eloszlása, akkor a valószínűsége, hogy egy konkrét kijelölt napra esik valakinek a születésnapja $1/365,25$ (kivéve február 29-ét, akkor csak $1/1461$).

A 14. ábrán láthatjuk a pontos eloszlást.



14. ábra. 1981 és 1994 közötti adatok alapján (Saját ábra)

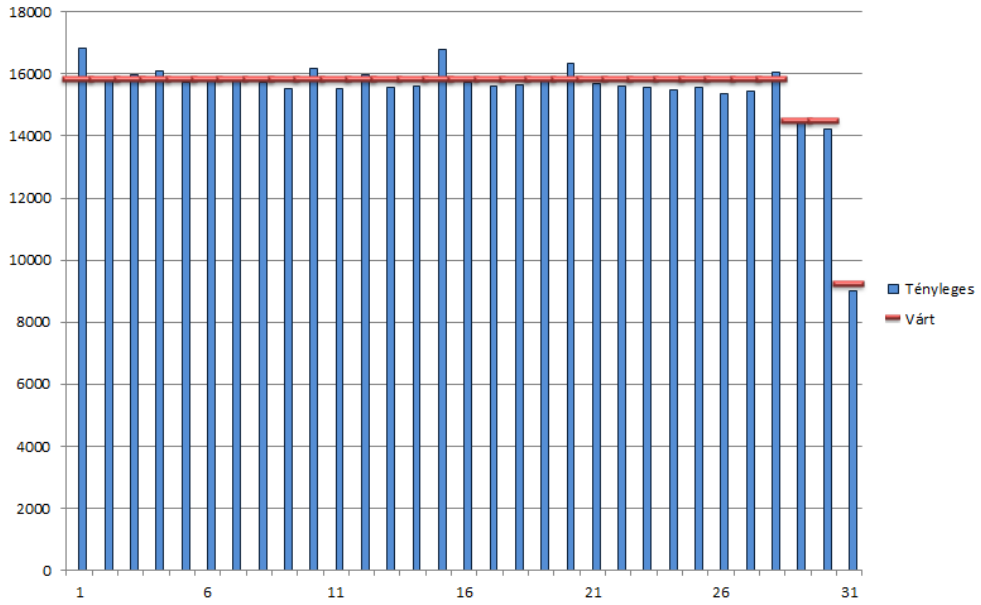
Összevetve az előző adatokkal, itt is nagyon hasonló eredményeket kaptunk, az évnek ugyanazon a részén látható növekedés/csökkenés. Ami itt különbségként jelenik meg, hogy a hétvégék nem válnak külön, mivel nem egy konkrét évet vizsgáltunk, hanem többet. Mivel összeadódnak az értékek, így átlagban minden naphoz magasabb értékek kerülnek, nem is válhatna szét két csoportra.

Statisztikai módszerekkel számszerűen is meg tudjuk határozni, hogy elfogadható-e az az állítás, hogy az adatok egyenletes eloszlásúak. Tehát a hipotézisünk az, hogy a születésnapok eloszlása egyenletes. Ekkor február 29-től eltekintve minden napra $\frac{1}{365,25}$ valószínűséggel esnek a születések, míg február 29-re $\frac{1}{1461}$ valószínűséggel esik valakinek a születésnapja. Ebből arra a következtetésre juthatunk, hogy az emberek számának eloszlása a születésnapok tekintetében binomális, várható értéke$= n \cdot p = 1314,28$ (február 29-nek $328,57$), szórásnégyzete$= n \cdot p \cdot q = 1313,41$ (feb. 29-nél $329,03$).

Illeszkedésvizsgálatot hajthatunk végre az adathalmazon, mellyel ténylegesen elfogadhatjuk, vagy elvethetjük a hipotézisünket, azaz, hogy egyenletes lesz az eloszlás, vagy sem. Khi-négyzet próbát alkalmaztam az adathalmazon, a próbastatisztika értéke $1356,4$ lett.

A khi-négyzet próba kritikus értéke 99, 9%-os szignifikancia szint és 365-ös szabadsági fok mellett pedig 454, 2. Mivel a próbastatisztika értéke nagyobb, mint a kritikus érték, ezért a nullhipotézisünket elvetjük, azaz az adathalmazunk eloszlása nem egyenletes.

Egy újabb megközelítésben is megvizsgálom ezeket az adatokat. Most nem napi szinten ábrázolom az adatokat, hanem a hónap napjai alapján készítek egy összegzést. Ezt a 15. ábrán szemléltetem.



15. ábra. Napi bontásban a születésnapok eloszlása (Saját ábra)

Ezen már láthatjuk, hogy sokkal egyenletesebb az eloszlás. Két kiugró értéket fedezhetünk fel az ábrán, minden hónap elsejét és 15.-ét. Erre magyarázat lehet, hogy sokan, akik a századforduló előtt születtek, nem tudták pontosan, hogy hányadikán volt a születésnapjuk, ezért ilyenkor gyakran elseje, vagy tizenötödike lett bejegyezve.

Ha egyenletesnek feltételeznénk az emberek születésnapjának eloszlását, szinte biztosan hibát követnénk el, mivel egész évben ingadozást mutatnak a megvizsgált adatok alapján. Több külső kritérium határozza meg az eloszlását, amelyek közül néhány bemutatásra kerül a következő részben, ezeket mindenképp érdemes figyelembe venni a születésnapok eloszlásának meghatározásakor. Azonban ha másik bontásban figyeljük az adatokat, már láthatunk egyenletességre utaló jeleket.

6.2. Befolyásoló tényezők

Mivel emberi születésekről beszélünk, annak magyarázata, hogy miért nem egyenletes az eloszlás, kevésbé matematikai. Viszont ahhoz, hogy bármilyen modellt felállítsunk, ezeket a tényezőket is figyelembe kell venni. A Harvard egyetemen készült egy antropológiai kutatás, amely vizsgálja a befolyásoló tényezőket. Végeredményben három fő területet említenek, amelyek a legnagyobb mértékben, a legszélesebb körökben érvényesek. Ezek a kutatások is az amerikai kultúrát vizsgálták, alapvetően az ottani környezetet veszik figyelembe.

Az elsőként meghatározott faktor a társadalom, annak szokásai, hagyományai. Azt mondják, hogy hatással vannak a születésekre a különböző valláshoz kötődő ünnepek, például a karácsony, amit általában a szeretteik körében töltenek az emberek, vagy a nyári szabadságok, amikor szintén sokkal több ideje van foglalkozni az embereknek egymással. Emellett az esküvőknek a főszezonja is inkább a tavasztól őszig terjedő időszakra esik, statisztikák alapján gyakori, hogy a házasság után egyből gyermeket vállalnak a párok. Összevetve ezeket az állításokat az adatainkkal, máris magyarázatot kaptunk arra, hogy miért magasabb a születések száma szeptemberben.

A második tényező az időjárás, amellyel kapcsolatban arra az eredményre jutottak a kutatók, hogy nagy mértékben befolyásolja a születések eloszlását, hogy éppen napos, esős, hideg vagy meleg időjárás volt-e. Utolsó tényezőként pedig a női termékenységet jelölik meg. Azonban ezt a két tényezőt nem tudom összevetni az adataimmal, ezekhez kapcsolódóan nincs elég információ. (Ellison, Valeggia és Sherry, 2005)

7. Egyenletes eloszlás a permutációkon

Egy érdekes példa kerül előtérbe ebben a fejezetben, melynek során az év napjainak permutálása lesz a középpontban, valamint ehhez kapcsolódó vizsgálatok, hogy mennyire tekinthető egyenletes eloszlásúnak az így kapott sorrend. Ehhez kapcsolható egy történet, amely egészen 1969-ig nyúlik vissza, amikor az USA-ban egy lottóhoz hasonló rendszert hoztak létre, amely a fiatal férfiak hadseregbe történő besorozásának sorrendjét határozta meg a Vietnámi háború idején. Ez egy akkora esemény volt, hogy még a televízióban is közvetítették a sorsolást, különös érdeklődéssel figyelték, hiszen sokan érintettek voltak benne.

A sorsolás lebonyolítása azzal kezdődött, hogy az év minden egyes napjához egy sorszámot rendeltek, 1-től 366-ig. Majd utána a lottósorsoláshoz hasonlóan kihúzták véletlenszerűen a számokat, ezáltal permutálták azokat. Például az első kihúzott szám a 258-as volt, amely eredetileg szeptember 14-et jelentette, mivel az a 258. nap az évben, de ezáltal az új sorrendben az 1-es számjegyet kapta. Ez azt jelentette, hogy mindenki, aki ezen a napon született, szintén az egyes sorszámot nyerte el. Ezzel az volt a céljuk, hogy a nyilvánosság előtt is meg tudják mutatni, hogy valóban véletlenszerűen sorozzák be a fiatalokat, az első 195 kihúzott dátumhoz tartozó férfiakat „jutalmul" csak később hívták be, mint a többi számhoz tartozó társaikat. Azonban nem hagyott mindenkit nyugodni a rendszer, a statisztikusok például elkezdték vizsgálni a felállított rendszert, mivel nem voltak meggyőződve róla, hogy valóban véletlenszerűen alakult ki ez a sorrend. Először is megvizsgálták, hogy a kihúzott számoknak volt-e valamiféle rendszere, hogyan oszlottak el. Ezt ábrázolták is, de ebből semmilyen következtetést nem tudtak levonni, látszólag véletlenszerűen alakultak ki az eredmények. Egyedül azt lehetett észrevenni, hogy a kapott eredmények alapján az egyes hónapokhoz tartozó átlagok egy csökkenést mutattak. Tehát a novemberben és decemberben születetteket előbb húzták ki, mint az év elején születetteket. Több statisztikai vizsgálat is azt mutatta, hogy nem véletlenszerű az eloszlás, ezért tovább folytatták a vizsgálódást. Mi állhat ennek a hátterében?

Ehhez először azt kell megismerni pontosan, hogy fizikailag hogy zajlott a sorsolás. Azt már tudjuk, hogy minden dátum kapott egy sorszámot, amit egy-egy papír cetlire írtak, azokat pedig beletettek kapszulákba. Majd utána ezeket a sorszámokat elhelyezték egy-egy

dobozban, úgy, hogy minden doboz egy hónap számait tartalmazta, ezeket külön-külön összerázták, ezáltal keverve őket. A sorsolás előtt pedig ezekből a dobozokból öntötték bele a kapszulákat egy nagy urnába. Először a januárhoz tartozó számokat, majd havonta szépen sorba, decemberrel bezárólag kerültek bele a számok az urnába. Folyamatosan próbálták keverni közben a számokat, forgatták egyik oldalról a másikra az edényt, majd utána így hagyták a sorsolásig.

Mikor eljött a sorsolás ideje, véletlenszerűen húztak a kapszulákból, viszont a legtöbbször a tetejéhez közel esőket választották. Ha a keverés tökéletes lett volna, akkor ez ténylegesen egy véletlenszerű permutáció lett volna, azonban ebben az évben az történhetett, hogy mégsem keverték át elég alaposan a számokat, így a később beleöntöttek, azaz az év végéhez közelebb eső számok, sokkal közelebb voltak az urna tetejéhez, mint a többi. Így fordulhatott elő, hogy ténylegesen, akik az év vége felé születtek, jobban jártak, mint a többiek, mivel őket nagyobb eséllyel csak később sorozták be. Ahhoz, hogy ezt bebizonyítsák a statisztikusok, különböző metrikákat alkalmaztak, mellyel vizsgálták, hogy mennyire véletlenszerű a permutáció. Ezek közül most három távolságot (Spearman, Kendall, Hamming) veszek alapul, melyekkel vizsgálom a permutációkat. (Marden, 1995)

Mivel nem tudjuk, hogy az összekeverés után az elejét jelölő számok kerülnek-e feljebb az urnában, vagy az év végét jelölők, ezért legyen $e_m = (1, \ldots, m)$ és $e_m^c = (m, \ldots, 1)$, amelyek arra vonatkoznak, hogy az eredeti számozást fordított sorrendben is összevetjük majd a kapott eredménnyel. Jelölje Y a kihúzott számokat. Ekkor egy kétoldalú hipotézisük lesz, melyben az egyik hipotézis, hogy az Y-nal jelölt, kihúzott lottószámok egyenletes eloszlásúak, a másik pedig, hogy az e_m, e_m^c permutáció valamelyikéhez állnak közel. A hipotézisvizsgálathoz pedig használjuk az u-próbát, valamint a korábban említett távolságokat, azok várható értékével és szórásával, melyekre a statisztikai próba elkészítése során fogunk alkalmazni.

1. Definíció. *(Spearman távolság) Legyen $s, t \in S_n$ permutáció, ekkor a*

$$d_S(s,t) = \sum_{i=1}^{n} |s_i - t_i| \tag{1}$$

kifejezést az s és t permutációk Spearman távolságának nevezzük.

2. Definíció. *(Kendall távolság) Legyen $s, t \in S_n$ tetszőleges permutáció, ekkor a*

$$d_K(s,t) = \sum_{i,j} (s_i < s_j, t_i > t_j) \tag{2}$$

kifejezést az s és t permutációk Kendall távolságának nevezzük.

Ez a távolság tehát azt fejezi ki, hogy hány olyan számpár van, melyet a két permutáció fordított sorrendbe rak.

3. Definíció. *(Hamming távolság) Legyen $s, t \in S_n$ permutáció, ekkor a*

$$d_H(s,t) = \sum_{i=1}^{n} \delta(s_i, t_i) \tag{3}$$

kifejezést az s és t permutációk Hamming távolságának nevezzük.

Ez a távolság tehát azt mutatja meg, hogy a két permutáció hány helyen tér el egymástól. Kiszámítható, hogy ha Y egyenletes eloszlású az összes permutáció S_m halmazán, $s \in S_m$ pedig tetszőleges rögzített permutáció, akkor mennyi lesz a $d(s, Y)$ valószínűségi változó várható értéke és szórásnégyzete. Ezeket a most definiált három távolságra az alábbi (5.) táblázat tartalmazza:

	Várható érték (M)	Szórásnégyzet (D^2)
Spearman	$m(m^2-1)/6$	$m^2(m-1)(m+1)^2/36$
Kendall	$m(m-1)/4$	$m(m-1)(2m+5)/72$
Hamming	$m-1$	1

5. táblázat. Távolságok várható értéke és szórásnégyzete (Marden, 1995)

Az is megmutatható, hogy nagy m esetén a $d(s, Y)$ valószínűségi változó közelítőleg normális eloszlású. Tehát a fenti módokon definiált képletekkel kiszámíthatjuk az u-próba próbatatisztikáit. Most $m = 366$, s pedig legyen az $e^c_{366} = (366, 365, \ldots, 2, 1)$ permutáció (mivel az a gyanúnk, hogy a nagyobb sorszámokat húzták ki előbb, azaz a kihúzott sorrend túl közel van az e^c_{366} permutációhoz). A próbastatisztika kiszámítása az

$$U = \frac{d(e^c_{366}, y) - M}{D}$$

	$d(e_m, y)$	U	$d(e_m^c, y)$	U
Spearman	10015392	$4,49$	39276	$-3,64$
Kendall	38369	$4,31$	6327116	$-4,31$
Hamming	364	$-1,00$	364	$-1,00$

6. táblázat. Távolságok, próbastatisztika (Marden, 1995)

képlet alapján lehetséges, ahol y a megfigyelt permutáció. A távolságok, valamint a próbastatisztikák értékei a következőként (6. táblázat) alakultak:

A Spearman és a Kendall távolságok esetén a p-értéket megkaphaatjuk a Z-ből, mivel normális eloszlású, p-értéke elég kicsi, a Hamming-távolság esetén a p-érték pedig $0,26$. Ezért arra az eredményre jutottunk, hogy a Spearman és a Kendall távolság megmutatja, hogy van szignifikáns kapcsolat az eredeti számozás és a permutálást követő számozás között. Ezzel ellentétben a Hamming-távolság nem mutatott ki szignifikáns hasonlóságot. (Marden, 1995)

Tehát a vizsgált három távolság közül kettő azt mondja, hogy nem keverték meg elég alaposan a számokat a sorsolás előtt, így nem volt teljesen véletlenszerű a sorsolás. Azok a férfiak jobban jártak, akik novemberben, vagy decemberben születtek, mivel az ő születési dátumukhoz tartozó számok feljebb maradtak az urnában a sorsolás alatt.

8. Összegzés

Számos témakör került előtérbe a dolgozatban. A közös pont mindegyikben az egyenletes eloszlás, ebből is láthatjuk, hogy egy adott témakörhöz mennyi különböző érdekesség kapcsolható. A fejezetek mindegyikére elmondható, hogy bemutattam az adott feladatot, paradoxont, majd ehhez próbáltam szemléletes példákat keresni, készíteni, amelyekben vizsgáltam, hogy a valóságban hogyan jelenik meg az adott eset. A legtöbb esetben a várttól eltérő eredmény született, ez szolgált a dolgozat alapjául, olyan esetek feltárása, amelyekben a mindennapi életünkben megszokotthoz képest, valami más eredmény születik.

A bevezetést és az elméleti összefoglalást követő fejezetben a véletlenszám generátorok kerültek előtérbe, ahol azt vizsgáltam, hogyan tudunk a számítógépek segítségével véletlenszámokat generálni. Ehhez mind elméleti összefoglaló, mind gyakorlati példa tartozott. A példában egy könyvből vett karaktersorozat volt a kiinduló helyzet. Ezt konvertáltam át számsorozattá, majd ezek megfelelő összegzésével egy újabb sorozatot kaptunk, amely már ténylegesen egyenletes eloszlású lett.

A negyedik fejezet, a kezdőszámjegy probléma fejezete volt. Ebben az esetben két nagyobb példa is előtérbe került, melyek alapjául egyrészt népességstatisztikai adatok, másrészt a Facebookon szereplő, meghatározott személyek adatlapjai szolgáltak. Megvizsgáltam, hogy a kezdőszámjegyek eloszlása valóban nem volt egyenletes, míg a második és a harmadik számjegyeké már igen. A paradoxon ebben az esetben az volt, hogy arra számíthattunk, hogy az első számjegy eloszlása is egyenletes lesz. Ezt a problémát Benford vizsgálta többek között, akinek egy képlete is van arra vonatkozóan, hogy milyen gyakorisággal jelennek meg a kezdőszámjegyek. A példákban láthatjuk, hogy valóban közelítenek az értékek Benford számaihoz.

A következő fejezetben egy kicsit a geometria irányába tértem át. A Bertrand-paradoxont vizsgáltam, melyben egy meghatározott valószínűséget kerestem, mégpedig azt, hogy egy kör véletlenszerűen kiválasztott húrja mekkora eséllyel lesz nagyobb, mint a körbe írható szabályos háromszög egy oldala. Három különböző módszerrel három különböző eredmény született, melyek mindegyike a maga módján helyesnek tekinthető. Az eltérés pedig azzal magyarázható, hogy nem mindegy, hogy milyen alakzaton választunk ki véletlenszerűen

egy pontot. Attól függött az eredmény, hogy a pontot a körlapon, a köríven, vagy egy sugáron választottuk-e.

A hatodik fejezet az emberek születésnapjához kapcsolódik. Azt gondolhatnánk, ha elég nagy mintát vizsgálunk, a születésnapok eloszlása egyenletes lesz. Viszont több adathalmazt elemezve, különös érdekességek kerültek előtérbe. Például az, hogy hétvégéken jóval kevesebb ember születik, mint hétköznapokon, amennyiben napi szinten nézzük az eredményeket. Valamint további tényezők is befolyásolják a születéseket, az évben különböző ingadozások figyelhetők meg. Természetesen ezek Amerika népességére igazak ebben a formában, kultúránként eltérő eredmények születhetnek. Viszont ha más megközelítésben vizsgáljuk az eredményeket, például a napok sorszámai alapján összegezzük az értékeket, akkor már sokkal egyenletesebb eredményt kapunk. Tehát ebben az esetben is fontos figyelembe venni, hogy milyen adatokkal, milyen felbontásban vizsgáljuk azokat.

Végül az utolsó fejezet tulajdonképpen egy megtörtént eset bemutatását tartalmazza. Azt vizsgáltam, hogy egy számsorozat permutálása egyenletes eloszláshoz vezet-e. Általában elmondható, hogy ha megfelelő módon végzik a keverést, akkor például egy lottósorsolás során valóban egyenletes eloszlású lesz a kihúzás által előállított új sorrend. Azonban van ellenpélda is, amelyet bemutattam, mégpedig amikor 1969-ben Amerikában besorozták a férfiakat katonának. Utólag kiderült, hogy a sorsolás mégsem volt annyira véletlenszerű, mint amennyire tervezték. Ennek igazolásához különböző távolságok felhasználásával készítettem számításokat.

Hivatkozások

[1] Bob de Vivo (2005) *History of Random Number Generators*, Probability and Statistics előadás

[2] Csiszár Villő *Valószínűségszámítás speci I. éves matematika tanárszakos hallgatóknak*, URL: http://www.cs.elte.hu/~villo/speci/specijegyzet.pdf, Letöltve: 2016.02.12

[3] Csiszár Villő (2009) *Valószínűségszámítás (jegyzet)*, URL: http://www.cs.elte.hu/~villo/esti/valszam.pdf, Letöltve: 2015.11.20

[4] David Gleich (2009) *Birthday distribution* URL: http://web.stanford.edu/~dgleich/notebook/2009/04/birthday_distribution.html Letöltve: 2016.04.29

[5] Dr. Huba, Antal, Dr. Lipovszki, György (2014) *Méréselmélet* BME MOGI, Url: http://mogi.bme.hu/TAMOP/mereselmelet/math-ch03.html

[6] John I. Marden (1995) *Analyzing and Modeling Rank Data*, Chapman & Hall, London

[7] KSH (2011) *A népesség számának alakulása, terület, népsűrűség*, URL: http://www.ksh.hu/nepszamlalas/tablak_teruleti_00, Központi Statisztikai Hivatal honlapja, Letöltve: 2016.01.12

[8] Nemetz Tibor, Wintsche Gergely (1999) *Valószínűségszámítás és statisztika mindenkinek*, Polygon, Szeged

[9] Peter T. Ellison , Claudia R. Valeggia és Diana S. Sherry (2005) *Human birth seasonality*, URL: http://www.sas.upenn.edu/~valeggia/pdf%20papers/birth_seasonality.pdf Department of Anthropology, Harvard University, Letöltve: 2016.04.29

[10] R.A. Raimi (1976) *The first digit problem* The American Math. Monthly 83, pp. 521 – 538

[11] Roy Murphy *Birthday*, URL: http://www.panix.com/~murphy/bday.html, Letöltve: 2016.04.29

[12] Székely J. Gábor (2004) *Paradoxonok a véletlen matematikájában*, Typotex, Pécs

[13] TDM (2016) *Egyenletes eloszlás*
URL: https://hu.wikipedia.org/wiki/Egyenletes_eloszlas, Letöltve: 2015.11.14

[14] William Feller (1965) *An Introduction to Probability Theory and Its Applications* John Wiley & Sons, New York

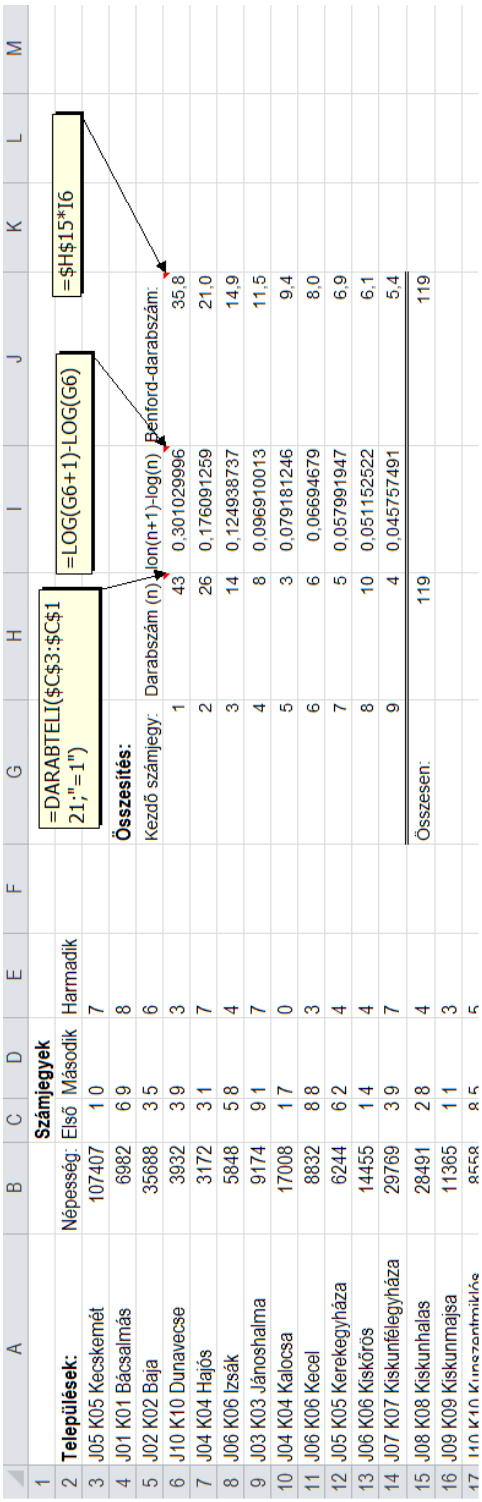
Ábrák jegyzéke

1. Az egyenletes eloszlás eloszlás- és sűrűségfüggvénye (Huba & Lipovszki, 2014) 6
2. Az arany ember első 3000 karaktere (Saját ábra) 11
3. Bács-Kiskun megye (Saját ábra) 15
4. Benford, második számjegyre (Saját ábra) 16
5. Benford, harmadik számjegyre (Saját ábra) 17
6. Top 100 híres ember (Saját ábra) 18
7. Benford, második számjegyre (Saját ábra) 19
8. Benford, harmadik számjegyre (Saját ábra) 19
9. Véletlen pont a körlapon (Székely, 2004) 24
10. Véletlen pont a körvonalon (Székely, 2004) 24
11. Véletlen pont a kör sugarán (Székely, 2004) 25
12. 1978-ben élve születettek száma (Saját ábra) 26
13. 1994-ben élve születettek száma (Saját ábra) 27
14. 1981 és 1994 közötti adatok alapján (Saját ábra) 28
15. Napi bontásban a születésnapok eloszlása (Saját ábra) 29
16. 1. melléklet, A népességszám elemzése (részlet) 40
17. 2. melléklet, Véletlenszám generálás (részlet) 41

Táblázatok jegyzéke

1. Benford adatai (Raimi, 1976) 13
2. Benford-törvény (Raimi, 1976) 14
3. Khi-négyzet próba (Saját táblázat) 17
4. Khi-négyzet próba (Saját táblázat) 20
5. Távolságok várható értéke és szórásnégyzete (Marden, 1995) 33
6. Távolságok, próbastatisztika (Marden, 1995) 34

9. Mellékletek



16. ábra. 1. melléklet, A népességszám elemzése (részlet)

17. ábra. 2. melléklet, Véletlenszám generálás (részlet)