A SZTOCHASZTIKA KÖZÉPISKOLAI OKTATÁSA

doktori (PhD) értekezés

Szászné Simon Judit

Debreceni Egyetem
Debrecen, 2005.

Ezen értekezést a Debreceni Egyetem TTK .Matematika Doktori Iskola matematika-didaktika alprogramja keretében készítettem a Debreceni Egyetem TTK doktori (PhD) fokozatának elnyerése céljából.

Debrecen, 2005. augusztus 15.

_____

Szászné Simon Judit
jelölt

T É M A V E Z E T Ő I N Y I L A T K O Z A T

Tanúsítom, hogy . Szászné Simon Judit.. doktorjelölt 2000.-2005. között a fent megnevezett Doktori Iskola matematika-didaktika alprogramja keretében irányításommal végezte munkáját. Az értekezésben foglaltak a jelölt önálló munkáján alapulnak, az eredményekhez önálló alkotó tevékenységével meghatározóan hozzájárult. Az értekezés elfogadását javasolom.

Debrecen, 2005. augusztus 15.

_____

Nemetz Tibor egyetemi tanár

TARTALOMJEGYZÉK

I. Bevezetés

A dolgozatban foglaltak kidolgozása során alapvető célunk volt valószínűségszámítási, statisztikai fogalmak olyan kialakítása, bevezetése, amely az osztályteremben lefolytatott megbeszélések, viták során alakul ki. Konkrét feladatok megoldására vártunk végrehajtható ötleteket a tanulóktól. Központi helyet igyekeztünk adni statisztikai döntéshozatal lehetőségének. Egyik fő célunk annak demonstrálása volt, hogy a statisztikai alkalmazások rendkívül hatékony döntéshozatalt tesznek lehetővé, de korántsem mindenhatók. Alkalmazásukhoz fárasztó adatgyűjtésre lehet szükség.

1.1. A kezdetek

1975-ben Herczeg János a Magyar Rádió Ifjúsági osztályán egy műsorsorozatot szerkesztett középiskolások valószínűségi és statisztikai fogalmakról alkotott véleményéről. A műsor szakértői Nemetz Tibor és Tusnády Gábor voltak. A megbeszélések alapjául Nemetz Tibor [N70] cikke szolgált. Idézzük még Tusnády Gáborné: „A valószínűségszámítás középiskolai tanítása” című cikkét [T73].

A műsor résztvevői a Budapesti Berzsenyi Dániel Gimnázium tanulói voltak. Az adások dinamikus lefolyása érdekében szakkörön előzetesen foglalkoztunk az anyaggal. A tanulók érdeklődése, aktivitása, a foglalkozások után a sztochasztika kérdéseihez való hozzáállásuk változása mély benyomást tett rám. Azóta is rendszeresen felhasználom ezt a témát gimnáziumi óráimon a valószínűségszámítás témakörében, és a foglalkozásokba beépített anyagot állandóan gazdagítom. Tapasztalataimat a Berzsenyi Dániel Gimnáziumon kívül a Móricz Zsigmond Gimnáziumban és a Fazekas Mihály Fővárosi Gyakorló Általános Iskola és Gimnáziumban gyűjtöttem össze.

A rádió egyik adását az egyszerű helyettesítéses titkosírás megfejtésének szentelte. A helyettesítő ábécé a rádió lehetőségeinek kihasználásával hangjegyekből állt, és a rádióadás fogadtatása nagyon élénk volt. Ennek megfelelően az oktatás elején én is foglalkozom az egyszerű helyettesítéses titkosírás megfejtésével. Kiderült, hogy ez rendkívüli mértékben motiválja a tanulókat statisztikai felmérések tervezésére és konkrét végrehajtására. Ezen a területen nagy gazdagsággal áll rendelkezésre statisztikai forrás, statisztikai nyersanyag írott nyelv formájában.

A dolgozatban beszámolok ezen a területen kifejlesztett módszereimről, összeállított oktatási anyagaimról, és az órákon szerzett közel 30 éves tapasztalataimról.

1.2. Az elmélet és a gyakorlat kapcsolata

A valószínűségszámítás és statisztika alkalmazásához szükséges szemléletmód kialakításához elengedhetetlenül szükséges, hogy alkalmas elméleti modelleket válasszunk, és azokat az oktatásra alkalmas formában tálaljuk. Ahhoz, hogy maga az oktatás is valóban hasznos lehessen, ugyancsak elengedhetetlenül szükséges, hogy az elméleti modelleket összehasonlítsuk a gyakorlati tapasztalatokkal, amelyeket a konkrét kísérletek végrehajtása során szereztünk. Fontos azt is megfigyelni, hogy ugyanazon jelenség, kísérlet többszöri végrehajtása során szerzett megfigyelések mennyire egyeznek meg egymással. A lehetséges felmerülő eltérésekre fel kell készíteni a tanárokat. Ezért rendkívül fontos számbavételük, rendszerezett összegyűjtésük, magyarázatuk. Ennek a feladatnak az elvégzése is a dolgozat célja.

1.3. Tanári hozzáállás, továbbképzés

A valószínűségszámítás és statisztika alkalmazásához elengedhetetlenül szükséges, hogy a tanárok rendelkezzenek a megfelelő ismeretekkel, és hajlandóak legyenek az anyagot tanítani. Az 1990-ben Budapesten tartott „Training Teachers to Teach Statistics” témájú konferencia részletesen foglakozott ezzel a területtel. Idézzük Fischbein professzor megállapítását a tanárképzésben megoldandó feladatról: [F90] "A tanárok valószínűségszámítási és statisztikai oktatásának tartalmaznia kell a megfelelő intuitív háttér kialakítására alkalmas területek és módszerek megismertetését.” Ugyanezt a kívánalmat hangsúlyozták más résztvevők is (lásd [H89]).

A hazai gyakorlatban az egymást váltó „oktatási reformok” egyik jellemzője az, hogy a valószínűségszámítás az új tananyagban szerepelt, ha a régiben nem volt, és kimaradt, ha előzőleg volt. Mindezt többnyire a tanári hozzáállással indokolták. Valójában a tanári gyakorlatot egy csendes ellenállás jellemezte. Ennek alapvető oka a biztonságérzet hiánya. A tananyag oktatásához kísérletek elvégzésére van szükség. Ahogyan azt maga a „véletlen” szó is mutatja, a kísérletek kimeneteleit a tanár nem tudja pontosan előre megmondani, és nincs felkészítve a lehetséges eltérések magyarázására, értelmezésére. Ezért nekik is szükségük van egy magyarázatokkal ellátott oktatási anyag összeállítására. Dolgozatom elkészítésekor egy ilyen anyag összeállítása is célom volt. Jelen dolgozaton kívül az összeállított anyagot az utóbbi néhány évben tanári továbbképzéseken is rendszeresen ismertetem. Ezek a tanári továbbképzések rendkívül pozitív visszhangra találtak.

1.4. Nemzetközi kitekintés

1996-ban a VIII. Nemzetközi Matematika Oktatási Konferencia egyik témacsoportja a valószínűségszámítás és matematikai statisztika oktatásának nemzetközi helyzetével foglalkozott. Kiderült, hogy a nemzetközi gyakorlatban az egyes országok ilyen irányú oktatáspolitikája jelentősen eltér egymástól. A témával rendszeresen foglalkozik a "Newsletter of the international study group for research on learning probability and statistics" Departamento de Didactica de las Matematicas Facultad de Ciencias de la Educacion, Universidad de Granada)

A különböző országokban lényegesen eltérő mélységben tanítanak valószínűségszámítást, és a tanított anyag is lényegesen eltér. A fenti konferencia kiadványa alapján öt fő hozzáállást lehet megkülönböztetni. [N97] Ezek a következők:

- Axiomatikus halmazelméleti hozzáállás
- Alapjaiban kombinatorika, ahol a valószínűségszámítás csak álruhaként szolgál
- Szerepel ugyan valószínűségszámítás, de ez urna-modellekre és egyszerű játékokra korlátozódik
- A valószínűségszámítási fogalmak modell-építés útján épülnek ki
- A valószínűségszámítás fogalmi kiépítését kísérletek, megfigyelések, projekt tevékenység támogatja.

A felsorolás ismertetését azért tartottuk itt szükségesnek, hogy a kifejlesztett anyagot beilleszthessük a hozzáállásoknak ebbe a nemzetközi sorába. Mi az utolsó két viszonyulást követjük, tapasztalatszerzés segítségével építjük ki a matematikai modell elemeit.

1.5. Módszertani kérdések

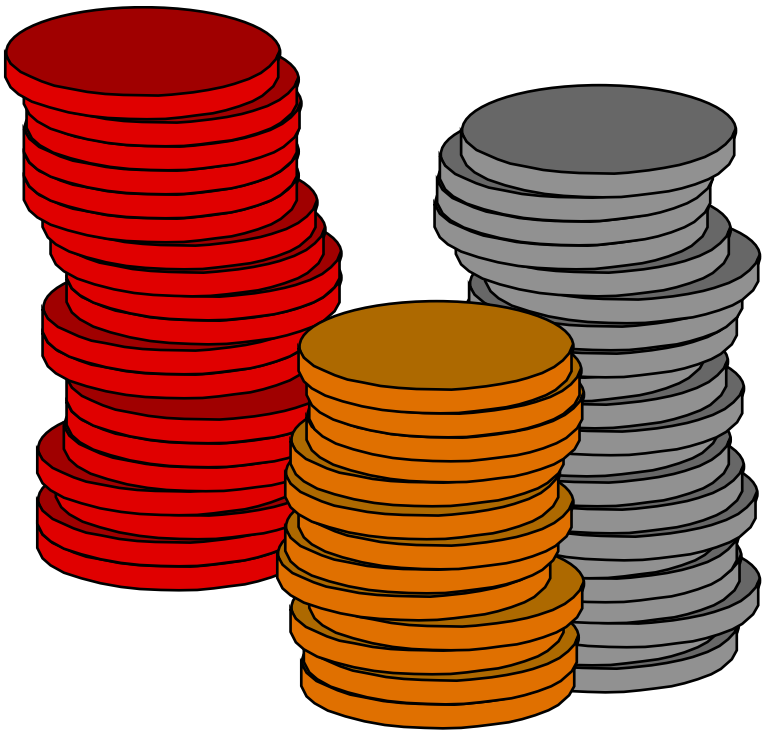
A tananyag megválasztása mellett ugyancsak fontosak a módszertani kérdések, tehát az oktatás formájának, metodikájának megválasztása. Ezen a területen is széles a nemzetközi skála.

1. A valószínűség matematikai szigorúsággal követett axiomatikus tárgyalása, formális levezetésekkel.
2. A valószínűség és a statisztika párhuzamos, vagy élesen elkülönített oktatása.
3. Alkalmazott matematikai problémák motivációs hasznosítása, szemben az alkalmazásokat teljesen figyelmen kívül hagyó levezetésekkel
4. Tapasztalatokra támaszkodó intuitív fogalomra támaszkodó felépítés, illetve az intuíció figyelmen kívül hagyása
5. Felfedeztető kísérletek végzése, osztályon belüli vita hasznosítása, illetve ezek teljes mellőzése

6. Egyes országokban nagyon szűk körben használnak számítógépes szimulációkat és más (többnyire előre gyártott) számítógépes alkalmazásokat, videókat és filmeket.

Az összeállított anyag megtervezése során a hatodik lehetőséggel nem foglalkoztunk. A valószínűségszámítás és a matematikai statisztika egymást támogató, párhuzamos tárgyalását követtük. Elsősorban felfedeztető kísérleteket, az ezekből leszűrt tapasztalatokat használtunk motivációul.

II. A legegyszerűbb véletlen sorozatok:
fej-írás sorozatok



2.1. A II. fejezet témája

„Mennyire véletlen a véletlen?” teszi fel a kérdést Révész Pál akadémiai székfoglalójában, és válaszként felsorol számos mély eredményt. Ez a kérdés tipikusnak és rendkívül alapvetőnek tekintendő a téma tudományos vizsgálata esetén, valamint a gyakorlati alkalmazások számára. Az iskolai oktatás számára éppen ilyen fontos azt megismerni, hogy a tanulók milyen kialakult véleménnyel rendelkeznek a témáról.

Varga Tamásnak tulajdonítandó az ötlet, hogy a véletlenről alkotott tanulói vélemények megismerését a legismertebbnek feltételezhető területtel kell kezdeni, és ezzel a területtel kell a foglalkozásokat is kezdeni. Ez a terület szerinte a szabályos érmével történő fej-írás sorozatok világa. Tapasztalataink megegyeznek vele, és állítható, hogy a tanulók nagyon határozott elképzelésekkel rendelkeznek ezen a területen.

Varga Tamás bevezetésként azt kérte III-IV. osztályos általános iskolai tanulóktól, hogy írjanak le két fej-írás sorozatot, az egyiket egy érme feldobálása segítségével, mégpedig írjanak F-et, ha fejet dobtak, és I-t ha írást. A másik sorozatot „fejből” írják, tehát úgy, ahogy egy lehetséges valódi fej-írás sorozat szerintük keletkezhet. A két sorozatot olyan sorrendben mutassák meg, hogy a sorrend ne árulja el, melyiket írták fejből, és melyik származott valódi dobássorozatból. Megvizsgálva a két leírt sorozatot, arra vállalkozott, hogy megmondja, melyiket írták, és melyiket dobták.

A döntés alapja a fej, illetve írás sorozatokból álló *futamok hossza* volt: amelyik sorozatban ez hosszabb volt, azt választotta dobott sorozatnak. (Futamnak nevezünk egy azonos elemekből álló olyan sorozatot, amely vagy elől kezdődik, vagy előtte a másik jel szerepel, továbbá vagy a sorozat végén van, vagy a másik jel követi.)

Ez a döntési eljárás rendkívül tekintélyt parancsoló eljárás: Varga Tamás az esetek több mint 80%-ban helyesen adta meg a dobott sorozatot.

Saját tanítási gyakorlatomban közel 30 éve én is ezt a bevezetést követem azzal a különbséggel, hogy mindegy, az első vagy a második sorba írt sorozatot dobják vagy írják, de először kell írni a sorozatot, és csak azután dobhatják. Tapasztalataim rendkívül pozitívak, és csak javasolni tudom másoknak is ezt a bevezetést. A 2.2 pontban összegezem a tapasztalataimat, míg az 2.3 pontban elméleti számítást hívok a jelenség elemzéséhez segítségül.

2.2. Fej-írás sorozatok a tanulók elképzelésében

A szóban forgó sorozatokat egyértelmű hivatkozás miatt nevezzük
- *véletlen sorozatnak*, ha érmedobás alapján keletkezett,
- *elképzelt sorozatnak*, ha a tanuló „fejből" írta.

Az egyszerűbb összehasonlíthatóság miatt tételezzük fel, hogy az összehasonlítandó sorozatok N hossza megegyezik. (N értékének megválasztásáról később lesz szó.)

Amint azt a fentiekben említettük, a tárgyalás megindításakor megkérjük a tanulókat, hogy írjanak le „fejből" egy N hosszúságú fej-írás sorozatot egy üres papírlapra, majd egy másik üres lapra egy valóban érmedobás segítségével nyert fej-írás sorozatot. Megkérjük őket, hogy önmaguk számára jegyezzék meg, melyik papírlap melyik sorozatot tartalmazza, de egy kívülálló (például a tanár) számára ez maradjon felismerhetetlen.

A tanár arra vállalkozik, hogy megmondja, melyik papírlapon van a dobott sorozat, és melyiken az elképzelt. A döntés során Varga Tamás módszerét kell követni: azt a sorozatot kell dobott sorozatnak választani, amelyben a „tiszta fej" (tiszta írás) sorozat hosszabb. Varga Tamás N=200 dobásból álló sorozatokat használt, amikor is a döntési eljárás rendkívül „tekintély megalapozónak" bizonyult, osztályonként legfeljebb egyszer hibázott. Az alábbiakban saját tapasztalatainkat foglaljuk össze.

Varga Tamás a témával általános iskola alsó tagozatában, III-IV. osztályos tanulókkal foglalkozott. Ebben a korosztályban az N=200 hosszúság nem jelentett problémát. A mi első tapasztalatunk viszont az volt, hogy gimnáziumban ez a hosszúság a megbízhatóság rovására ment: a „valódi" sorozat generálása során a tanulók egy része a dobás helyett is képzelt kimeneteleket írt le. Tapasztalataink alapján mi az N=60 hosszúságot választottuk, és a továbbiakban ismertetésünk erre az értékre vonatkozik.

Pusztán a tájékozódás megkönnyítése miatt megadunk öt képzelt sorozatot, kiemelve bennük a leghosszabb fej-írás sorozatot. A sorozatokat követő L a sorozat leghosszabb futamának a hosszát tünteti fel.

FF***IIII***FIII	FIFFFIFIFI	IFIFFIFFFF	IIFFIIFIIF	IIIIFFIFIF	IIFFFIIIFI	*L=4*
FIFFFFIFII	IFIFIFIIIF	FIIFFIFIFF	FIFFFIIFIF	FIIFIFIFIF	IIF***IIIII***FF	*L=5*
I***FFFF***IFIF***I***	***III***FFIFFII	***FFFF***IFFFII	IFFIIIFFFI	IF***IIIIFFFF***	III***FFFF***IFF	*L=4*
FIFIIIFIFF	FI***FFFF***IFFI	FFFIF***IIII***F	IFIIIFIFIF	IIFFIFIFII	FIFIIF***IIII***	*L=4*
FIFIIF***IIII***	FFIIFIIIFF	IIIFIFIIFF	FFIFFFFIIF	IF***IIII***FFII	FIFFFFIFFI	*L=4*

2.1.a. táblázat. Tanulók által elképzelt dobás-sorozatok

Összehasonlíthatóság miatt megadjuk ennek az öt sorozatnak a dobott párját is:

FFTTFFIFFF	FFIIIFFFIF	IIIIIIFIIF	FFIFIFFIFI	FI***FFFFFFFF***	IIIFFFIFIF	*L=8*
IFIFIIIIFF	FFIIIIFIFI	FFIFFFIIII	IFFII***FFFFF***	***FF***IIIIFIFF	IFIIIIFFFF	*L=7*
IFIIIIIFII	FFIFFIIFII	IIIIFFIIIF	I***FFFFFFFF***I	FFIFFFFFFI	FIFFFIFFFF	*L=8*
IFIFFFIIIF	FFIFFIIFIF	IFFIFIIFFI	FFF***IIIII***FF	FIFFIIFFIF	FFIFIFFFIF	*L=5*
FIIIIFFFFF	IFFIFFFFIF	IIFIFFIIFF	FIFIF***IIIII***	***II***FFFIIIFF	IIFIIFIIII	*L=7*

2.1.b. táblázat. Tanulók által dobott sorozatok

Megkérdeztük a tanulókat, miért éppen ilyen dobássorozatot képzelnek el. A három leginkább tipikus vélemény a következő:

- A véletlennek van egy kiegyensúlyozási szokása, ezért hosszabb sorozatokban hosszú *F*ejsorozatok és hosszú *Í*rássorozatok nem fordulnak elő.
- Valódi véletlen dobássorozatokban a fejek és írások száma csak nagyon kicsit térhet el egymástól.
- Fejről írásra, és írásról fejre való változás az esetek felében következik be.

Ezek a vélemények és a fenti öt sorozat támogatják Varga Tamás döntési elvét. Azt tapasztaltuk, hogy néhány „elvetemült”, a „csak azért is”-t szerető tanuló kivételével a tanulók következetesen alkalmazták ezt a hozzáállást a képzeletbeli sorozatok leírása során.

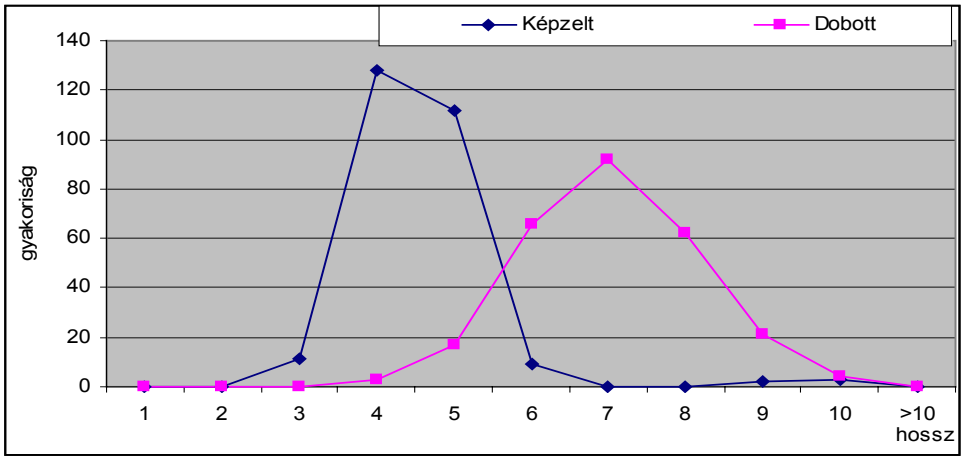
Az alábbi táblázat 265 tanuló által készített 60 hosszúságú „elképzelt” és dobott sorozatban a leghosszabb futam hosszának statisztikai eloszlását mutatja be:

	Hossz:	1	2	3	4	5	6	7	8	9	10	>10
Képzelt	Gyak:	0	0	11	128	112	9	0	0	2	3	0
Dobott	Gyak:	0	0	0	3	17	66	92	62	21	4	0

2.2. táblázat. 265 elképzelt és dobott 60 hosszúságú sorozat legnagyobb futamhosszai

2.3. Egy gyakorló feladat

Az oktatási anyag megszerkesztésének szerves része gyakorló feladatok megtervezése is. Egy ilyen feladatot mutatunk be illusztrációként. Ez a feladat csakúgy, mint társai, nagyon „barátságos” fogadtatásra talált a tanulók körében. Nemcsak saját véleményüket alakították ki elemző vizsgálódás alapján, hanem általában órán kívüli vitát is folytattak egymás között.



2.1. ábra: 265 elképzelt és dobott sorozat legnagyobb futamhosszai

A feladat alapját egy tíz soros lista képezte. Ezt a listát a tanár készítette a következőképpen: Korábbi években a tanulók által írt fej-írás sorozatokból készítettünk 10-10 darab 60 hosszúságú ***1***. segédlistát, majd valódi érmedobálással egy ugyanekkora ***2.*** segédlistát. Ezután egy érmét feldobtunk 10-szer egymás után. Ha a dobás fejet mutatott, akkor az első (képzelt) listáról írtuk át a sorozatot a feladatra készített listára, ha a dobás írás volt, akkor a második (dobott) listáról. Ezzel a módszerrel a tanulók számára elfogadhatóan a lista minden sora egyforma valószínűséggel vagy dobott, vagy képzelt sorozatként jelent meg. A feladat éppen az volt, hogy eldöntsék mindenegyes sorról, hogy abban elképzelt, vagy dobott sorozat szerepel-e.

1	FIIFFFIIIF	FFIFFIIFIF	IFFIFIIFFI	FFFIIIIIFF	FIFFIIFFIF	FFIFIFFFIF
2	FIIIIFFFFF	IFFIFFFFIF	IIFIFFIIFF	FIFIFIIIII	IIFFFIIIFF	IIFIIFIIII
3	IIFFFFIFFF	FIFFFIFIFI	FIFIIFIIII	FFIIFFIFFI	FFFFIIFIFI	FFIIIFFFIF
4	IFIIIIFIFF	FIFIFIFFFI	IFFIIFIFII	IFIIIFFIFI	IFFIFIFIFI	FFIFFFFFII
5	IFIFFFIIIF	FIFIIFIFFF	FFFIIIFIII	FFFIFIIIII	FIIFFIFIFI	FIIIIIIIIF
6	FIIIIFIFIF	FFFIIFIIFF	IIIIFIIIFF	FIIFFFIIIF	FIFFFFIIII	FFFIIIIFII
7	IFIIIFFFFI	FFIIIIFFII	IIFIIFIIFF	FIFIIFFFFF	FFIFFFFIFI	IIFIIIIIFI
8	IFIFFFIFII	IFIIIIFIIF	IIIFIFFFFI	FIFFFIFIFI	FFIIFIFIFF	IFIFFIFFFF
9	FIFFFFIIII	IIFFIIIIFI	FFIIFIFIIF	FFFIIIIFFI	IFIIIIFIIF	FIFFFIIIFF
X	FFIIIFFIIF	IIIIFIIFII	IFFIFIIFII	IFFIFIFFFI	IIFFIIIIFF	FFIFFIFFFF

2.3.táblázat. Képzelt, vagy dobott sorozat?

2.4. A leghosszabb futam elméleti vizsgálata

A fej-írás sorozatok a ***véletlen számsorozatok***, tehát a független, azonos eloszlású sorozatok legegyszerűbb megjelenését adják. Többnyire ez szolgál példaként, „etalonként" akkor, amikor a tanulók, de akár a felnőttek is, általában a véletlenre gondolnak.

A véletlen számsorozatokat általában rendkívül hamisan interpretálják. Ez a nagy számok törvényének helytelen alkalmazásából adódik, miközben a nagy számok törvényének még a heurisztikus ismerete is hiányos. Ez a hamis interpretálás vezet a fenti tanulói vélemények közül az első kettőhöz: a fejek és írások száma bármely pillanatban „majdnem" azonos. A valóság azonban az, hogy a véletlennek nincs semmilyen „kiegyenlítő" mechanizmusa.

Be fogjuk látni, hogy egy valódi (dobott) fej-írás sorozatban viszonylag hosszú „tiszta fej", illetve „tiszta írás" sorozatok fordulnak elő. Az ilyen sorozatok hossza $log_2(n)$ körül van egy n hosszúságú sorozatban, ami a fentiekben adott döntési elv helyességére utal. (A továbbiakban mindig kettes alapú logaritmussal dolgozunk, és csak az ettől különböző alapot írjuk ki.)

Ismertetésünkben a N=60 sorozat hosszra koncentrálunk, mivel sokéves tapasztalataink szerint ekkora véletlen sorozat generálására hajlandók a tanulók. (Természetesen nagyobb N mellett a döntés megbízhatósága is nagyobb, így N=200 tényleg csodálatos eredményre vezethetett általános iskolás gyerekek esetében.)

Mivel $\log_2 60 \approx 6$, ellenőrizzük, hogy mekkora valószínűséggel szerepel egy legalább hat hosszúságú futam egy 60 dobásból álló fej-írás sorozatban. Azt a valószínűséget akarjuk meghatározni, hogy ***nincs*** 6 hosszúságú, vagy ennél hosszabb tiszta fej sorozat. E célból tekintsük a 10 egymás után következő 6 hosszúságú részsorozatot. A keresett valószínűség nyilván nagyobb, mint az a valószínűség, hogy ennek a 10 blokknak egyike sem áll azonos dobásokból. Ennek a valószínűségnek a meghatározására a gimnáziumi oktatásban elérhető mélységű klasszikus valószínűségszámítási módszer alkalmazható.

Annak a valószínűsége, hogy a 10 blokk egyike sem áll azonos elemekből:

$$(62/64)^{10} = (1 - 1/32)^{10}$$

A komplementer esemény azt jelenti, hogy létezik legalább egy legalább 6 hosszúságú részsorozat. Ennek valószínűsége

$$1 - (1 - 1/32)^{10} \approx 0{,}728$$

Az általános, tetszőleges N-nek megfelelő esetben is éppen ilyen egyszerű „jó” alsó és felső becslés levezetése annak az $A_r(n)$ eseménynek a valószínűségére, hogy n hosszúságú fej-írás sorozatban van egy legalább r hosszúságú azonos elemekből álló futam.

A szóban forgó valószínűségek levezetéséhez jelölje $X_1, X_2, \ldots, X_n$ az érmedobás kimeneteleit, és legyen B_i az az esemény, hogy az $X_i, X_{i+1}, \ldots, X_{i+r-1}$ események azonosak. Nyilvánvalóan

$$A_r(n) = B_1 + B_2 + \ldots + B_{n-r+1}$$

Az $A_r(n)$ esemény $P(A_r(n))$ valószínűségére felső becslést eredményez a szubadditivitás:

$$P(A_r(n)) \leq P(B_1) + P(B_2) + \ldots + P(B_{n-r+1}) = (n - r + 1)\, P(B_1),$$

ahol felhasználtuk, hogy a B események egyenlő valószínűségűek.

Másrészről

$$P(A_r(n)) \geq P(B_1 + B_{1+r} + \ldots + B_{1+(k-1)\,r}) =$$

$$= P(B_1) + P(B_{1+r}) + \ldots + P(B_{1+(k-1)\,r}) = k\, P(B_1),$$

Itt k az n/r tört egész része.

Azt kaptuk tehát, hogy

$$\mathbf{[int(n/r)]\ (1/2)^{(r-1)} \leq P(A_r(n)) \leq (n - r + 1)/\ 2^{(r-1)}}$$

Speciálisan, n=60 és r=6 esetén a felső korlát triviális, mivel az 1-nél nagyobb, a rendkívül durva 10/32 alsó becslés azt mutatja, hogy minden harmadik sorozat várhatóan tartalmaz egy 6 hosszúságú futamot.

Mindamellett jól mutatja, hogy n=60-nál pl. egy r=10 hosszúságú (vagy hosszabb) futam nagyon valószínűtlen, hiszen ekkor a felső becslés értéke 51/512< 0,1.

Az egyenlőtlenségek elemzése azt mutatja, hogy $log(n) + 9$ olyan felső hosszat jelent, amelynél hosszabb futam 1/1024 –nél nagyobb valószínűséggel nem létezik. A második egyenlőtlenségből az ellenkező irányú becslés adódik: Legalább (1 - 1/1024) valószínűséggel létezik

$$log(n) - loglog(n)$$

hosszúságú futam. A két egyenlőtlenség együtt bizonyítja, hogy a leghosszabb futam hossza valóban $log(n)$ körül van.

Tekintsük mindkét egyenlőtlenséget $n>15$ esetén. A fenti heurisztikánál kissé pontosabb gondolatmenetet követve keressük meg azt az n-től függő legkisebb R értéket, amelyre teljesül, hogy

$$(n - R + 1)/ 2^{R-1} \leq 1/1024$$

tehát az R hosszú, vagy hosszabb futamok nagyon valószínűtlenek.

Másrészt, legyen S az a legnagyobb egész, amelyre teljesül, hogy

$$\frac{\left\lceil \frac{n}{S} \right\rceil}{2^{S-1}} \geq \frac{1}{2}$$

Legalább ekkora ugyanis annak a valószínűsége, hogy előáll egy legalább ekkora futam. A két egyenlőtlenség megoldásai nagyon közeli értékek. Az első egyenlőtlenség teljesül, ha

$$R = log\ (n) + 9,$$

ahogyan az az eddigiek alapján is várható volt.

A második egyenlőtlenségből $n>16$ esetén

$$log\ log\ (n) - 2 > 0,$$

és így az a legnagyobb S egész, amely kielégíti

$$S \leq log(n) - log\ log\ (n) + 2$$

egyenlőtlenséget, alkalmas választás lesz. A két megoldás együttesen bizonyítja, hogy **a leghosszabb futam hossza valóban *log (n)* körüli.**

Kívánatos, hogy a leghosszabb futamok hosszának a 2.1 táblázatban szereplő empirikus (tanulói) eloszlását az elméleti eloszlással összehasonlítsuk. A következőkben ezt az eloszlást fogjuk meghatározni. A fenti levezetés szerint a szóban forgó valószínűségek többsége rendkívül kicsi. A valószínűségek meghatározásához egy rekurziót adunk meg.

2.5. Rekurzió a leghosszabb futam eloszlására

A következőkben annak a valószínűségét határozzuk meg, hogy egy n hosszúságú fej-írás sorozatban a leghosszabb azonos elemekből álló futam hossza **pontosan** r.

A valószínűségek meghatározása előtt azon n hosszúságú fej-írás sorozatok számát határozzuk meg, amelyekben a leghosszabb azonos elemekből álló futam hossza **pontosan** r. Legyen az ilyen sorozatok száma $g(n,r)$, tehát**:**

legyen $g(n,r)$ azon n hosszúságú fej-írás sorozatok száma, amelyekben a leghosszabb azonos elemekből álló futam hossza **pontosan** r.

Peremfeltételként a $g(n,1)$ és $g(n,n)$ valószínűségek szolgálhatnak, amelyek meghatározása triviális. A leghosszabb azonos elemekből álló futam hossza akkor és csak akkor pontosan 1, ha a fej-írás sorozat alternál, míg a leghosszabb azonos elemekből álló futam hossza pontosan n csak úgy lehetséges, ha minden elem azonos. Ezért

$$g(n,1) = g(n.n) = 2$$

Ugyancsak nyilvánvaló, hogy $r > n$ esetén $g(n.r) = 0$.

Az általános eset vizsgálata előtt, a levezetés gondolatmenetének ismertetése miatt külön vizsgáljuk az r=2 és r=3 eseteket.

Nyilvánvaló, hogy r=2 esetén az első futam vagy 1, vagy 2 hosszúságú. Ha az első futam hossza 1, akkor a megmaradó $n - 1$ elemű sorozatban a leghosszabb futam pontosan 2 hosszúságú, tehát ezek száma $g(n-1,2)$. Ha az első futam hossza 2, akkor a maradék sorozattal szemben az a követelmény, hogy benne nem lehet 2-nél hosszabb futam, tehát a benne levő leghosszabb futam hossza vagy 1 vagy 2, és ez a két eset kizárja egymást. Ezért ekkor a kedvező sorozatok száma

$$g(n-2,1) + g(n-2,2) = 2 + g(n-2,2).$$

Összegezve:

$$g(n.2) = g(n-1,2) + g(n-2,2) + 2$$

Ezt a rekurziót lépésről lépésre haladva bármilyen n mellett meg lehet mondani a pontosan 2 leghosszabb futamú sorozatok számát anélkül, hogy a nagyobb r értékekről bármit kellene mondanunk. A $g(n,2)$ értékek sorozata n = 1 –gyel kezdve:

$$g(1,2)=0;\quad g(2,2)= 2;\quad g(3,2)= 4;\quad g(4,2)= 8;\quad g(5,2)= 14;$$

$$g(6,2)=24;\quad g(7,2)=40;\quad g(8,2)=66;\quad g(9,2)=108;\quad g(10,2)=176$$

A rekurziót feloldva explicit alakban is meg lehet adni a $g(n,2)$ gyakoriságokat. Legyen $x(n) = g(n,2) + 2$. Ezzel a jelöléssel az $x(n)$ sorozatra „tiszta”, konstans nélküli rekurziót kapunk:

$$x(n) = x(n-1) + x(n-2)$$

Feloldásához az $x^2 - x - 1 = 0$ egyenletet kell megoldani. Ennek gyökei

$$x_{1,2} = \frac{1 \pm \sqrt{1 + 4}}{2}$$

Mint ismeretes, a rekurzió feloldása a

$$x(n) = a_1 x_1^n + a_2 x_2^n$$

képlet, ahol az a_i konstansokat az $x(1) = 2$ *és* $x(2) = 4$ kezdeti értékek határozzák meg.

A megfelelő valószínűségeket a $g(n,2)/2^n$ hatványok adják meg, amik gyorsan tartanak nullához. Például n =60 mellett a valószínűség első hat számjegye 0.

Tekintsük most az r =3 esetet. Ekkor az első futam hosszára három lehetőség létezik: 1, 2, vagy 3. Az első két esetben a megmaradó részsorozatban a leghosszabb futam hossza pontosan 3, tehát

$$g(n-1,3) + g(n-2,3)$$

ilyen sorozat létezik. Ha az első futam hossza 3, akkor a maradék sorozatban a leghosszabb futam hossza legfeljebb 3, tehát 1, 2, vagy 3. Ezért ezek száma

$$g(n-3,1)+g(n-3,2)+g(n-3,3)$$

Összegezve:

$$g(n,3)= g(n-1,3)+g(n-2,3)+g(n-3,1)+g(n-3,2)+g(n-3,3)$$

Átrendezve:

$$g(n,3)= g(n-1,3)+g(n-2,3)+ g(n-3,3)+g(n-3,2)+2$$

Kezdeti feltételként szolgálhatnak $g(0,3)= g(1,3)=g(2,3)=0$ értékek. Ebből azonnal adódik a triviális $g(3,3)=2$ kezdőérték. További értékek:

n =4*: g(4,3)=g(3,3)+g(2,3)+g(1,3) +g(1,2)+2 = 2+0+0+0+2= 4*
n =5*: g(5,3)=g(4,3)+g(3,3)+g(2,3)+g(2,2)+2 = 4+2+0+2+2= 10*
n =6*: g(6,3)=g(5,3)+g(4,3)+g(3,3) +g(3,2)+2 =10+ 4+ 2+ 4+ 2= 22*
n =7*: g(7,3)=g(6,3)+g(5,3)+g(4,3) +g(4,2)+2 = 22+10+4+8+2= 46*
n =8*: g(8,3)=g(7,3)+g(6,3)+g(5,3) +g(5,2)+2 =46+22+10+14+2= 94*
n =9*: g(9,3)=g(8,3)+g(7,3)+g(6,3) +g(6,2)+2 =94+46+22+24+2=188*
n =10*: g(10,3)=g(9,3)+g(8,3)+g(7,3) +g(7,2)+2 =188+94+46+40+2 = =370*

2.5.1. Tétel: Azon *n* hosszúságú fej-írás sorozatok *g(n, r)* száma, amelyekben a leghosszabb azonos elemekből álló futam hossza **pontosan** *r*, kielégíti a

$$g(n,r) = \sum_{h=1}^{r-1} g(n-h, r) + \sum_{i=1}^{r} g(n-r, i)$$

rekurziót.

Bizonyítás:

Az $n > r > 1$ esetben n-re vonatkozó indukció segítségével vezetjük le a rekurziót. Az *n* hosszúságú fej-írás sorozatokat az első futam hossza szerint három csoportra osztjuk fel.

Jelölje *h* az első futam hosszát.

a) Legyen $h < r$. Ezek között a fej-írás sorozatok között azok száma, amelyekben a leghosszabb futam pontosan *r* hosszúságú, megegyezik azon *n–h* hosszúságú sorozatok számával, amelyekben a leghosszabb futam hossza pontosan *r*. Rekurziós feltételünk szerint ez a szám

$$\sum_{h=1}^{r-1} g(n-h, r)$$

b) Legyen $h = r$. Ekkor a leghosszabb futam akkor és csak akkor *r*, ha a sorozat további *n-h* elemű részében a leghosszabb futam hossza legfeljebb *r*. A maradékban nincs *r*-nél hosszabb futam, tehát ezek között a részsorozatok között a leghosszabb sorozat hossza 1 és *r* között változik. Ez a szám tehát

$$\sum_{i=1}^{r} g(n-r, i)$$

Ha $h > r$, akkor nyilván nincs *r* hosszúságú futam.
Az a)és b) eseteket összegezve adódik a megkívánt rekurzió:

$$g(n,r) = \sum_{h=1}^{r-1} g(n-h, r) + \sum_{i=1}^{r} g(n-r, i)$$

A *g(n,r)* számok kiszámítását megkönnyíti az a felismerés, hogy köztük sok azonos szám van. Nevezetesen igaz a következő :

2.5.2. Tétel: A *g(n,n-t)* számok azonos *t* mellett (tehát egy „ferde” diagonális mentén) egy *n*-től függő küszöb-szám után megegyeznek.

Bizonyítás:

Az állítás $t = 0$ mellett kiinduló észrevételünk szerint igaz: *g(n,n)*=2 minden *n*-re. Az $r = n-1$, $r= n-2$ és $r= n-3$ eseteket részletesen is kiírjuk

$$g(n,n-1)= g(n-1,n-1)+g(1,1) = 2+2 = 4, \quad \text{ha} \quad n>2.$$

$$g(n,n-2)= g(n-1,n-2)+g(n-2,n-2)+g(2,1)+g(2,2) = 4+2+2+2= 10,$$

ha $n>4$.

$$g(n,n-3)= g(n-1,n-3) + g(n-2,n-3) + g(n-3,n-3) + g(3,1) + g(3,2) +g(3,3)=$$

$$=6+10+2+4+2=24 , \quad \text{ha} \quad n > 6$$

Általánosan

$$g(n,n-t) = \sum_{h=1}^{n-t-1} g(n-h, n-t) + \sum_{i=1}^{n-t} g(t,i)$$

Az indukciós feltétel szerint mindkét összeadandóban egy indexszám után konstans összeadandók szerepelnek, ami állításunkat bizonyítja.

A *g(n,r)* értékeket $n \le 10$ esetén a 2.4. táblázat tartalmazza

n \ r	1	2	3	4	5	6	7	8	9	10	össz.
2	2	2	0	0	0	0	0	0	0	0	4
3	2	4	2	0	0	0	0	0	0	0	8
4	2	8	4	2	0	0	0	0	0	0	16
5	2	14	10	4	2	0	0	0	0	0	32
6	2	24	22	10	4	2	0	0	0	0	64
7	2	40	46	24	10	4	2	0	0	0	128
8	2	66	94	54	24	10	4	2	0	0	256
9	2	108	188	118	56	24	10	4	2	0	512
10	2	176	370	254	126	56	24	10	4	2	1024

2.4. táblázat. *g(n.r)* értékek $n \le 10$ esetén

A *g(n,r)* gyakoriságok helyett sokkal áttekinthetőbbek a

$$p(n,r) = \frac{g(n,r)}{2^n}$$

valószínűségek értékei:

$$p(n,r)=\frac{g(n,r)}{2^n}=\frac{1}{2^n}\cdot\sum_{h=1}^{r-1}g(n-h,r)+\frac{1}{2^n}\cdot\sum_{i=1}^{r}g(n-r,i)=$$

$$=\cdot\sum_{h=1}^{r-1}\frac{1}{2^h}p(n-h,r)+\frac{1}{2^r}\cdot\sum_{i=1}^{r}p(n-r,i)$$

Pl: ha $n=1$; $g(n,1)=2$; $p(n,1)=1/2^{n-1}$; $p(n,n)=p(n,1)$.

2.5.3. Tétel: A $g(n,2)$ sorozat szigorúan monoton nő, és $g(n,2)$ nagyságrendje megegyezik a Fibonacci sorozat nagyságrendjével.

Bizonyítás :

A $g(n,r)$ rekurziót $r=2$ –re kiírva adódik:

$$g(n,2) = g(n-1,2) + g(n-2,1) + g(n-2,2)$$

Kihasználva, hogy $g(n,1)=2$, ha $n>0$, adódik:

$$g(n,2) > g(n-1,2), \quad \text{ha } n>1,$$

ami bizonyítja a monotonitást. A Fibonacci típust a

$$g(n,2) = g(n-1, 2) + g(n-2, 2) + 2$$

rekurzió igazolja.

A $p(n,r)$ valószínűségeket $n < 11$ esetén a 2.5. táblázatban összegeztük.

r \ n	1	2	3	4	5	6	7	8	9	10
2	0,5	0,5	0	0	0	0	0	0	0	0
3	0,25	0,5	0,25	0	0	0	0	0	0	0
4	0,125	0,5	0,25	0,125	0	0	0	0	0	0
5	0,063	0,438	0,312	0,125	0,063	0	0	0	0	0
6	0,031	0,375	0,344	0,156	0,063	0,031	0	0	0	0
7	0,016	0,312	0,359	0,187	0,078	0,004	0,016	0	0	0
8	0,008	0,258	0,367	0,211	0,094	0,039	0,004	0,008	0	0
9	0,004	0,211	0,367	0,230	0,109	0,047	0,019	0,004	0,004	0
10	0,002	0,172	0,361	0,248	0,123	0,055	0,023	0,010	0,004	0,002

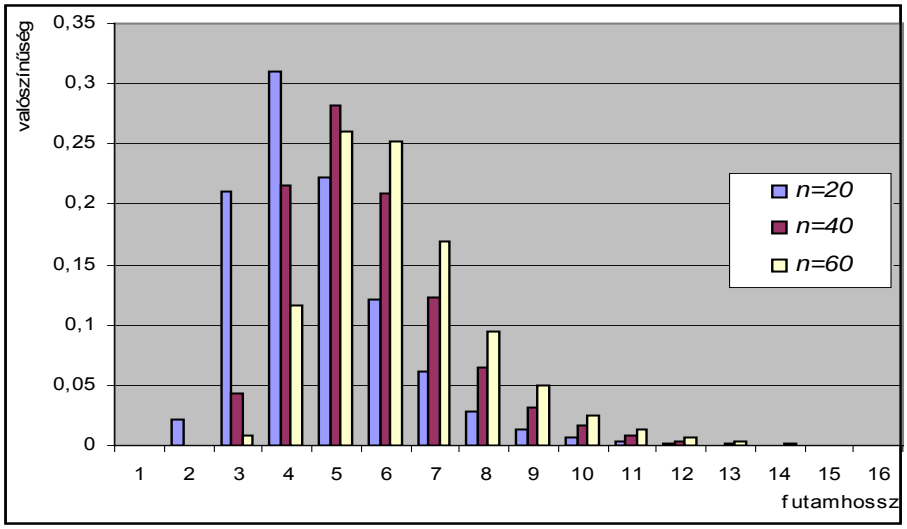
2.5. táblázat: $p(n.r)$ valószínűségek $n\leq 10$ esetén

A $p(n,r)$, $n<60$, $r<16$ valószínűségeket a rekurzió alapján egy egyszerű BASIC program segítségével határoztuk meg.

Az *n*=10, 20, 30, 40, 50, 60 melletti értékeket a 2.6. táblázat összegezi.

r \ n	10	20	30	40	50	60
1	0,0019	0,0000	0,0000	0,0000	0,0000	0,0000
2	0,1719	0,0209	0,0025	0,0003	0,0000	0,0000
3	0,3613	0,2107	0,0977	0,0431	0,0187	0,0081
4	0,2480	0,3100	0,2743	0,2156	0,1603	0,1157
5	0,1230	0,2215	0,2682	0,2823	0,2768	0,2601
6	0,0547	0,1218	0,1722	0,2092	0,2352	0,2525
7	0,0234	0,0607	0,0934	0,1220	0,1470	0,1686
8	0,0098	0,0290	0,0472	0,0642	0,0802	0,0952
9	0,0039	0,0136	0,0231	0,0323	0,0412	0,0499
10	0,0019	0,0063	0,0112	0,0159	0,0206	0,0252
11	0,0000	0,0029	0,0054	0,0078	0,0102	0,0125
12	0,0000	0,0013	0,0026	0,0038	0,0050	0,0062
13	0,0000	0,0006	0,0012	0,0018	0,0024	0,0030
14	0,0000	0,0003	0,0006	0,0008	0,0012	0,0015
15	0,0000	0,0001	0,0003	0,0004	0,0006	0,0007
>15	0,0000	0,0000	0,0002	0,0004	0,0005	0,0008

2.6. táblázat: *n* érmedobás leghosszabb futamainak eloszlása



2.2.ábra: *n*=20, 40, 60 melletti leghosszabb futamok hisztogramja

III. Hiányos sorozatok helyreállítása: statisztikai szemléletformálás

A	V	E	L		T	L		N	J	E		E	N	S	E	G	E		M
E		I	T	E	L		S	E	S	O	R		N	A	Z	E		B	E
R	E		K	U	L	O		B	O	Z	O	S		O	K		S	O	K
	T	B	E	I		E	G	Z	O		E		E	K	E	T		O	V
E		N	E		A	M	E		Y	E	K	M	E		S	A	J		
K	O		A	B	B	I		A	P	A		Z	T	A		A	T	A	
	K	A	L		E	M		I	N	D	I		E	G		E	Z		E
T		E	T	O	K		S	S		E	V	I	S	Z	O		T	K	I
F	E		E	Z	I	K	A	Z	O		A	T		Z	E	L	V	A	
A		O	K	A		A	M	E		Y	E	K		Z	A	M	U		R

3.1. A III. fejezet témája

A mindennapi életben gyakran fordul elő az a feladat, hogy valahány feltételezhető lehetőség közül ki kell választani az igazit. Ilyenkor megfigyeléseket végzünk. Az adatokat összegyűjtjük, majd azokat kiértékeljük. Az adatok általában függnek a véletlentől, így abszolút biztos döntés meghozatalára nincs mód. Ha azonban az ugyanazon lehetőségek között megfigyelhető adatok száma nő, és ezek képesek újabb és újabb információt közvetíteni, akkor a helyes döntés esélye is növekszik.

Ebben a fejezetben ezzel a témakörrel foglalkozunk. Egyszerű, könnyen áttekinthető feladatokkal foglalkozunk, amelyek mindazonáltal egy érdekes tanulói hozzáállásra mutatnak rá: Ha ugyanaz a döntési feladat ismétlődően fordul elő, a tanulók hajlamosak eltérő döntési eljárást követni. A hiányos sorozatok kiegészítése alkalmasnak mutatkozik ennek a tévhitnek az eloszlatására.

3.2. Három kiegészítendő hiányos sorozat

Egészítsük ki az alábbi három táblázatot!

3.2.*a*. Értelmes magyar szöveg

Kitöltöttük az alábbi ábrát egy értelmes magyar szöveggel, a szóközt a + jellel jelölve. Ezután minden kis kockához feldobtunk 3 érmét, és ha mind fejre esett, a kockában levő betűt kitöröltük. A feladat az üres kockák kitöltése.

	Z	Ó	R		K	O	Z	Á	S	+	M	I	N		E	N	K	I		E	K	+	M	I
N	D	E	N	F	E	L	É	+	K			Á	L	K	O	Z	O	T	T		P	E	R	
Z		+	C		A	K	+	A	N	N	A	K	+	A	K		N	E	K	+	P	É	N	Z
E		V	O	L		+	V	E	N		É	G	L	Ő	B	E		L	E		E	T	E	T
T	+	M	E	N		I	+	A	H		L	+	N			Á	N	Y	+	K	O	R	O	N
Á	É	R	T		K	I	T	Ű	N	Ő	+	V	A	C	S	O	R	Á	T	+	A	D		A
K	+	E	G	Y	+	K	O		O	N	Á	B	A	+	S	Z	Á	M	Í	T	O	T	T	Á
K	+		+	B	É	C	S	I	S	Z	E	L	E		E	T		F	I	N		M	+	B
O	R	J	Ú	H	Ú	S		K		S	Ü	T	V	E	+	P	R		Z	L		B	E	N
+	É	S	+	T	O	J	Á		B	A	N	+	A			O	R	A	+		O	G	Y	+

3.1.*a*. táblázat: Értelmes szöveg

3.2.*b*. Tizedik betűk

Kitöltöttük az alábbi ábrát egy értelmes magyar szöveg minden 10. betűjével (a többit figyelmen kívül hagyva), a szóközt a + jellel jelölve. Ezután ugyanazon kockákból kitöröltük a betűket, mint az előbb. A feladat az üres kockák kitöltése.

	K	É	E		D	R	Z	E	K	L	É	K	É		T	T	C	Ö		Ő	N	R	Ó	É
A	I	A	G	K	V	K	Á	N	I			+	K	S	I	V	L	Y	A		I	R	+	
E		G	M		S	T	+	E	+	+	+	R	Ő	E	E		S	I	É	+	J	Ü	K	+
K		N	É	A		B	T	O	K		K	Á	Z	R	Ö	T		O	V		O	S	+	+
O	+	K	Á	A		É	R	A	E		S	+	T	K			S	S	M	Á	R	Z	S	R
G	N	L	É		E	Y	V	T	B	Ö	T	L	S	R	+	A	S	G	E	N	L	+		U
Á	V	S	R	+	Á	J	É		I	A	M	+	R	K	+	G	E	D	K	Ö	I	Ö	T	S
L	E		A	Z	M	T	E	Y	Y	K	A	Ó	J		S	A		Á	C	S		+	Ü	E
T	U	R	É	E	N	E		E		S	+	K	S	A	+	A	S		K	É		E	A	+
+	N	T	E	R	A	O	Ü		A	U	N	A	T			+	G	I	E		A	S	+	+

3.1.*b*. táblázat: Tizedik betűk

3.2.c. Kockadobás

Kitöltöttük az alábbi ábrát egy kockadobás sorozat egymás utáni kimeneteleivel. Ezután ugyanazon kockákból kitöröltük a betűket, mint az előbb. A feladat az üres kockák kitöltése.

	5	1	3		4	2	6	4	5	5	3	5	3		4	4	3	3		6	4	3	5	2
3	2	5	5	2	3	3	2	2	5			4	4	5	1	6	3	5	6		4	6	2	
4		4	3		3	6	2	6	2	2	6	5	3	4	1		4	1	4	5	6	6	1	1
6		2	3	1	1		4	2	6		4	4	6	6	4	4		2	6		6	3	6	4
6	1	6	4	1		5	6	4	2		3	3	5	5			5	6	4	3	1	3	6	6
4	4	5	4		5	1	6	1	3	2	6	3	6	1	4	4	3	2	4	4	3	5		1
5	3		4	5	3	5	3		4	2	6	3	2	2	3	6	6	4	4	4	1	2	2	2
5	5	5	4	6	1	1	1	4	2	5	6	4	1		2	2		3	2	6		6	3	1
6	4	1	5	4	4	1		1		2	5	6	4	1	4	2	2		3	2		2	6	4
4	3	4	3	1	2	4	3		5	6	5	5	6			4	1	6	2		1	3	3	4

3.1.*c*. táblázat: Egymás utáni kockadobások

Hogyan egészítsük ki a táblázatokat?

A 3.1.*a*. táblázat esetén a válasz triviális: a kiegészítésnek értelmes szöveget kell eredményezni, így a beírandó betűk egyértelműen meghatározottak. A második táblázatban különböző mértékig ugyan, de bizonytalan, mit kell beírni. A beírandó betű függ a véletlentől, de nagyon erősen függ a környezetétől. A harmadik táblázat esetén a véletlen teljesen

egyeduralkodóvá válik: az 1,…,6 számok bármelyike bárhol egyenlő eséllyel fordulhat elő, nem lehet a hiányzó számokat „okosan” kitölteni.

A három táblázat a véletlen három arcát segít megmutatni. Ezek:

- A környezetétől nagyon függő, csaknem determinisztikus véletlen karakterekre
- A környezettől alig függő, mindazonáltal függő véletlen jelekre
- A környezettől teljesen független véletlen jelenségekre.

Kívánatos, hogy a tanulók több hasonló példa segítségével maguk alakítsák ki igényüket az esetek megkülönböztetésére. Erre mutatunk további lehetőségeket**.**

3.3. Találgatás a soron következőre

Ezek a tantermi foglalkozások a tanár játékvezetői tevékenységét igénylik. [S78]

3.3.a. Találgatás egy szöveg tizedik betűire

A játékvezető kiválaszt egy értelmes magyar nyelvű szöveget, és annak egy adott betűjétől kezdve kiírja minden tizedik betűjét, miközben a szóközt egy speciális jellel, például „”+” jellel jelöli, és semmilyen más jelet nem vesz figyelembe. A feladat soron következő betű kitalálása rákérdezéssel. Mindig csak egyetlen betűt lehet kérdezni, tehát a felteendő kérdés

„A soron következő betű ez-és-ez?”

típusú.

3.3.b. Találgatás egy szöveg minden tizedik betűjét követő betűire

A játékvezető most is kiválaszt egy értelmes magyar nyelvű szöveget, és annak egy adott betűjétől kezdve kiírja ugyanúgy minden tizedik betűjét, mint előbb, de vele együtt kiírja a rákövetkező betűt is. Megadja a soron következő betűt, és az erre következő betűt kell kitalálni rákérdezéssel ugyanúgy, mint a 3.3.a. feladatban.

3.3.c. Találgatás tíz betűt követő betűre

Most a játékvezető a teljes szöveget kiírja tízes blokkokba csoportosítva. Megad egy teljes tízes blokkot, és a feladat a tízes blokkot követő betű kitalálása rákérdezéssel, ugyanúgy, mint a 3.3.*a* találgatás esetén.

Hogyan találgassunk?

A hiányos táblázatok kitöltése során már egyértelművé vált a tanulók számára, hogy a találgatást a kitalálandó betűk bekövetkezésének esélyei szerint kell folytatni. A helyes kérdezési stratégia megválasztásához ennek megfelelően a kitalálandó jel gyakorisági sorrendjét kellene ismerni. Ennek megismeréséhez várhatunk, és kapunk is javaslatokat az osztályban. A tanulók számára nem kérdés, hogy valamennyi betű esetén ugyanazt a stratégiát kell folytatni. Ez a kérdezési stratégia a magyar nyelvben a betűk gyakorisági sorrendjével egyezik meg. Az ismeretlen gyakorisági sorrend meghatározására a tanulók hosszú szöveg megvizsgálását javasolják.

3.4. Betűstatisztikák tanulói meghatározása

Jelen fejezetben ismertetjük a gimnáziumi kilenc-tizedik osztályban több éven át folytatott gyakorlatunk tapasztalatait. Ezek szerint a tanulócsoportok maguk javasolják a magyar nyelv betűgyakoriságainak leszámlálással történő meghatározását. Természetes számukra a következő (durva) állítás:

3.1.tanulói állítás: Hosszú magyar nyelvű szövegekben az egyes betűk lényegében ugyanakkora gyakorisággal fordulnak elő. Az egyes gyakoriságok szövegről szövegre kicsit változnak.

Arról, hogy mekkorák ezek a gyakoriságok, vagy egyáltalán milyen a gyakorisági sorrendjük, még egyetlen osztályomban sem tudtak a tanulók megegyezni. Viszont a cél (megoldás) érdekében általában hajlandók voltak hosszabb szövegekből statisztikák manuális meghatározására. Az utóbbi években a statisztikákat már nem kell manuálisan elkészíteni. Megváltozott a nyersanyag beszerzésének módja is: A korábbi fárasztó begépelés helyett az Internetről rendkívül hosszú szöveget lehet letölteni.

A számítógépben rendelkezésre álló szövegekből programozás nélkül, szövegszerkesztővel is készíthető betűstatisztika, ahogy azt a Fazekas gimnáziumban az egyik tanuló csinálta: Ehhez a szövegszerkesztő „csere” módját úgy kell alkalmazni, hogy legyen azonos a kicserélendő betű és az, amire kicseréljük. A szövegszerkesztő megadja a kicserélt betűk számát. Így például az A betű gyakoriságát megkapjuk, ha az A betűt A betűre cseréljük, és kiolvassuk, hogy hány csere történt.

A 3.2. táblázatban megadtunk 40 tanuló által meghatározott betűstatisztikákat, míg a 3.3. táblázat számítógéppel saját magunk által készített nyelvstatisztikát ad meg. Ez utóbbi mutatja, hogy a tanulók valóban elkészítették a statisztikákat.

	köz	A,Á	B	C	D	E,É	F	G	H	I,Í	J	K	L	M	N
1	3049	2083	356	130	477	2430	123	611	423	736	205	727	1266	692	864
2	4990	2883	714	306	539	3165	245	953	415	1228	405	1574	2128	1031	1628
3	9948	2412	375	146	299	2482	143	5582	268	1791	136	8218	2279	6956	1037
4	5933	1933	429	125	172	3444	724	729	923	2560	23	215	1232	728	2619
5	4059	3027	670	203	405	3484	387	760	502	940	388	2100		738	1633
6	5720	3617	563	243	742	4051	347	1127	572	1164	329	1347	1948	1756	1739
7	1507	1246	193	429	186	1557	107	3280	676	5671	961	5854	5831	2957	4490
8	1366	1016	179	304	133	1206	766	3048	137	4132	138	5466	4840	2484	4774
9	792	720	110	19	94	637	30	185	102	259	59	231	317	135	291
10	5172	3207	501	219	660	3165	342	1101	381	1263	219	861	1824	900	1200
11	3899	2316	604	276	504	2316	240	795	360	1115	320	1152		820	
12	5447	3329	546	253	551	3906	421	867	327	1241	369	1374	1575	831	1435
13	464	357	52	38	42	331	19	81	43	140	25	168	136	101	200
14	3960	2567	389	171	464	2814	240	7661	401	8915	261	9994	1331	8890	1378
15	3395	2644	286	316	478	2150	353	418	360	1079	278	938	1270	815	878
16	4273	2845	404	185	519	3095	240	8546	461	1059	251	1048	1494	8892	1471
17	1038	898	105	42	127	709	91	192	75	3300	82	341	358	228	396
18	5785	3642	560	230	767	4365	323	1088	553	1252	347	1366	1976	1776	1777
19	4246	2920	494	228	483	3311	182	895	360	1219	272	1125	1389	1033	1595
20	4553	3415	502	213	330	2749	226	883		1298	153	939	1322	603	895
21	7821	6433	708	117	104	7001	1	1240	731	2962	713	2132	2987	1663	2336
22	4231	2441	435	108	584	3067	121	863	398	1225	257	1190	1227	947	1591
23	4223	3365	503	132	388	3263	211	860	336	1133	237	1971	1482	620	1601
24	479	273	49		163	430	16	0	0	109	27	342	120	154	273
25	4563	3117	303	228	267	5055	490	1026	260	1215	153	417	1254	1482	1329
26	2735	1909	290	86	282	2150	120	509	282	680	214	848	954	601	805
27	3859	3343	529	160	317	3427	231	789	412	1167	243	1935	1678	701	1681
28	6382	4634	722	276	724	5540	311	1284	508	1990	529	1709	2511	1451	2114
29	3663	2814	401	132	489	3205	221	812	364	1125	222	1032	1517	814	1389
30	4014	3287	493	141	455	3122	200	850	308	1133	237	1524	1425	699	1470
31	5262	5307	267	108	132	4702	156	528	315	1386	108	687	795	399	819
32	3858	3516	428	165	476	3200	175	936	427	1247	314	1203	1646	855	1348
33	4233	2113	535	287	610	3843	402	870	458	1103	234	1143	1921	830	1311
34	5601	5064	762	270	672	5757	415	1408	414	2002	452	1917	2514	1119	1980
35	2123	836	182	103	153	1468	83	374	161	437	82	432	671	365	515
36	3376	2787	311	181	341	2788	191	660	421	856	270	1147	1585	637	1238
37	7387	4017	719	389	789	5270	323	1081	481	2100	430	1840	2349	1459	2156
38	4757	2611	549	348	556	3098	298	832	366	1304	321	1587	2022	908	1631
39	5668	4203	711	239	711	5062	301	1123	474	1649	415	1612	2387	1199	1845
40	1900	1423	250	731	204	1668	130	3770	188	5047	110	5991	8466	5915	6591

3.2.a. táblázat: 40 tanuló által készített magyar betűgyakoriságok: 1.rész

	O,Ó	Ö,Ő	P	Q	R	S	T	U,Ú	Ü,Ű	V	W	X	Y	Z	Össz:
1	657	394	284	0	937	1237	159	153	105	468	0	8	330	764	21 018
2	169	0	432	0	112	1550	239	458	0	494	14	0	493	1256	27 097
3	860	3273	243	0	706	1272	146	214	122	408	0	75	414	8692	27903
4	235	0	650	37	164	1988	233	844	0	299	97	36	344	35	25866
5	101	681	415	3	117	1312	180	250	101	602	1	15	542	1216	23192
6	149	586	281	0	132	1587	278	529	249	567	0	2	647	1123	14892
7	496	2054	162	1	386	7626	853	127	468	218	2	15	183	5246	74491
8	418	1562	902	7	324	5064	676	734	562	134	47	6	198	3202	4274
9	293	72	33	0	185	356	429	64	33	106	0	0	128	266	22116
10	111	501	120	0	111	1002	213	300	144	342	0	0	219	942	15711
11	994	340	268	0	101	...	...	376	152	352	0	4	408	980	23840
12	138	585	283	0	121	1579	206	275	126	624	17	8	591	1096	2318
13	121	57	30	0	117	163	190	60	24	52	1	2	49	120	18715
14	119	2543	216	6	903	1177	134	329	150	380	0	79	504	8467	16630
15	972	333	353	15	106	1094	166	179	56	300	5	19	351	680	20387
16	126	4489	195	7	101	1346	200	336	173	458	20	52	547	9115	8319
17	337	132	96	0	313	353	353	86	39	124	0	0	118	288	27324
18	157	580	273	0	125	1534	272	508	239	582	0	3	689	1190	21055
19	133	335	270	1	961	1402	215	398	168	419	0	1	529	959	18543
20	462	1193	423	0	112	1554	213	231	161	465	0	2	436	967	42131
21	312	669	154	26	254	4001	456	683	312	103	54	18	647	1891	19659
22	974	214	115	0	742	1263	162	284	202	370	0	1	521	927	21569
23	124	580	260	0	117	1299	172	350	189	412	2	0	430	1396	2595
24	160	0	43	0	76	288	188	0	0	7	0	0	53	17	28728
25	759	942	264	0	158	1329	261	153	267	456	0	0	267	798	13284
26	819	190	115	0	659	820	131	261	113	283	0	1	326	646	21586
27	113	610	317	1	129	1594	208	237	197	431	2	2	452	1418	32337
28	162	665	364	1	168	2368	287	505	262	674	8	3	699	1613	19323
29	113	436	214	0	897	1211	179	298	138	440	7	15	512	978	56695
30	127	571	257	0	129	1381	191	401	197	483	3	1	515	1407	23575
31	264	699	723	10	858	1242	435	108	338	36	0	12	699	236	23375
32	138	295	267	0	952	1537	181	301	103	479	25	2	539	1050	20997
33	114	342	226	12	115	1754	194	285	129	498	33	16	501	932	32326
34	199	9425	397	64	176	2798	320	312	377	344	42	32	799	1924	6176
35	314	249	153	0	490	747	632	...	...	186	1	9	254	425	17860
36	101	423	214	0	677	1280	146	298	139	413	1	3	512	912	32373
37	153	693	317	1	141	2180	265	391	223	781	1	41	919	1452	22557
38	139	2	303	2	116	1892	232	338	32	705	13	6	625	1233	29060
39	141	745	313	0	141	2200	267	469	258	629	1	10	595	1505	10209
40	655	2797	105	1	428	6452	118	143	892	228	24	14	284	4226	58751

3.2.*b*. táblázat: 40 tanuló által készített magyar betűgyakoriságok: 2. rész

A 3.3. táblázatban szereplő gyakoriság-eloszlás Internetről levett magyar nyelvű szövegek feldolgozásával készült. A 278 432 betűs szöveg alapján készített betűgyakoriságok a következő gyakorisági sorrendet eredményezték:

+ E A T L N S O I K R Z M G Y D V Ö H B U F J P C Ü W X Q

A tizedik betűkre való találgatás során tehát ebben a gyakorisági sorrendben célszerű kérdezgetni.

A leggyakoribb 7 betű *gyakoriságai*:

42703, 30925, 28405, 20230, 14954, 14719, 13486

Kumulatív gyakoriságok:

42703, 71108, 99513, 119743, 137217, 151936, 165422

Relatív gyakoriságok:

0,1532 0,2551 0,3570 0,4295 0,4922 0,5450 0,5934

Ezek a relatív gyakoriságok azt mutatják, hogy átlagosan az esetek felében öt kérdéssel sikerülhet a kitalálás. Ez a megállapítás fontos a tanár számára, hiszen gyorsan meg kell ítélnie a tanulók teljesítményét.

A 3.3.*b* esetben ismert egy betű, és a rákövetkezőre kell találgatni. Teljesen nyilvánvaló a tanulók előtt, hogy most nincs univerzálisan legjobb kérdezési stratégia, a kérdések sorrendje függ az adott betűtől. Egy *e* vagy *a* betű után például nem célszerű ismét *e*-t vagy *a*-t kérdezni. Most is a gyakorisági sorrendben kell kérdezni, de ez a gyakorisági sorrend függ a megadott (előző) betűtől. Minden betű esetén ugyanolyan gyakorisági sorrendet kell meghatározni, mint azt a betűgyakoriságok esetén tettük.

	GYAKORISÁGOK				RELATÍV GYAKORISÁGOK			
KÖZ	3 049	39 610	3 395	42 703	0,145	0,159	0,150	0,153
A,Á	2 083	25 627	2 644	28 405	0,099	0,103	0,116	0,102
B	356	3 889	286	4 074	0,017	0,016	0,013	0,015
C	130	1 781	316	1 895	0,006	0,007	0,014	0,007
D	477	4 604	478	5 169	0,023	0,019	0,021	0,019
E,É	2 430	28 154	2 150	30 925	0,116	0,113	0,095	0,111
F	123	2 400	353	2 460	0,006	0,010	0,016	0,009
G	611	7 661	418	8 546	0,029	0,031	0,018	0,031
H	423	4 091	360	4 681	0,020	0,016	0,016	0,017
I,Í	736	8 915	1 079	10 569	0,035	0,036	0,048	0,038
J	205	2 661	278	2 581	0,010	0,011	0,012	0,009
K	727	9 994	938	10 498	0,035	0,040	0,041	0,038
L	1 266	13 341	1 270	14 954	0,060	0,054	0,056	0,054
M	692	8 890	815	8 892	0,033	0,036	0,036	0,032
N	864	13 738	878	14 719	0,041	0,055	0,039	0,053
O.Ó	657	11 809	972	12 776	0,031	0,047	0,043	0,046
Ö,Ő	394	2 543	333	4 489	0,019	0,010	0,015	0,016
P	284	2 146	353	1 935	0,014	0,009	0,016	0,007
Q	0	6	15	7	0,000	0,000	0,001	0,000
R	937	9 023	1 006	10 321	0,045	0,036	0,044	0,037
S	1 237	11 737	1 094	13 486	0,059	0,047	0,048	0,048
T	1 509	13 904	1 686	20 230	0,072	0,056	0,074	0,073
U,Ú	153	3 249	179	3 336	0,007	0,013	0,008	0,012
Ü,Ű	105	1 530	56	1 723	0,005	0,006	0,002	0,006
V	468	3 860	300	4 598	0,022	0,016	0,013	0,016
W	0	0	5	230	0,000	0,000	0,000	0,001
X	8	79	19	52	0,000	0,000	0,001	0,000
Y	330	5 074	351	5 407	0,016	0,020	0,015	0,019
Z	764	8 467	680	9 115	0,036	0,034	0,030	0,033
ÖSSZ:	21018	248 783	22 707	278 776	1	1	1	1

3.3. táblázat: Tanár által készített négy gyakoriság-eloszlás

Ez a munka azonban lényegesen nagyobb. Azonnal látszik ugyanis, hogy a tizedik betűk ismeretében nem a betűket, hanem az egymás utáni betűpárokat kell megvizsgálni.

A tanulói javaslat most is az, hogy készítsünk a betűpárokból gyakorisági eloszlást. A „nagyon hosszú szöveg” most azonban jóval hosszabbat jelent, mint az előbb.

Tanári tapasztalat:

A tanár számára lényeges tapasztalat, hogy ezt a leszámlálást házi feladatként a tanulók általában nem hajtják végre. Ezért mi magunk készítettünk betűpár gyakoriság-eloszlást egy 40 000 betűs szövegből. Ennek során a szokásos előkészítő átírást azzal egészítettük ki, hogy a sorvégi „Enter” jelet is szóközzel helyettesítettük.

A gyakoriságok pontos ismerete nem is annyira fontos egy jó stratégia kialakításához, mint inkább a „rákövetkezési” sorrendek. Elegendő ezeket meghatározni.

A magyar betűk feltételes gyakorisági sorrendjét a 3.4. táblázatban adtuk meg.

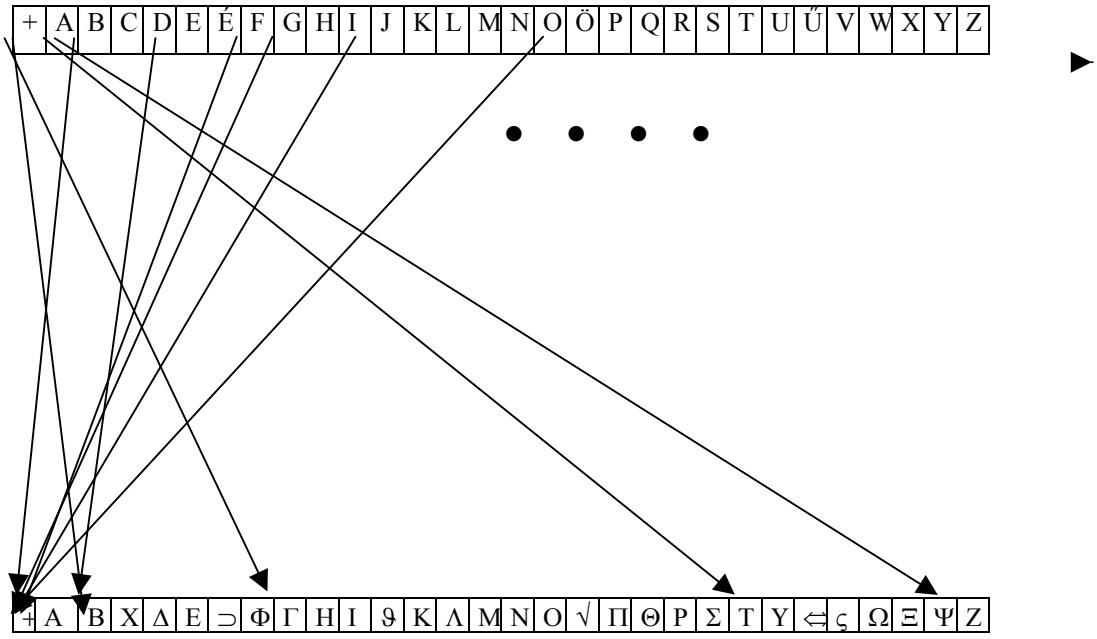
A táblázat első oszlopa az *ABC*. Ezek után az *ABC* betűit abban a gyakorisági sorrendben adtuk meg, amilyenben azok az elöl álló betűt értelmes szövegekben követik. Azokat a betűket, amelyek a sorok elején lévő betű után a vizsgált szövegben nem fordultak elő, alfabetikus sorrendben az előfordult betűk után soroltuk be.

ÁBC	Feltételes gyakorisági sorrend	*ÁBC*	Feltételes gyakorisági sorrend
A	-lzntkrdgi+pjmsbhvaceuéfoxyüáwq	**O**	lrnskmtzg+dpbvcfiéhjauoeáyöüwxq
Á	rlsntgkbmzdj+cviuaéhfpáeoöyüxwq	**Ö**	+rlsztkdnvbhgjeimaopáöuféüycxwq
B	eab+oiruöáélüsfnkhpctzmgdyvjxwq	**P**	oieráajlé+vctkupgzösnübhmdyfxwq
C	sih+éráaobenkdftlzmgöyvujpcüxwq	**Q**	u+eatlsniokrzámégdöybvjhfpcüxwq
D	eoaié+öusntjdrvlbhküámzyfcgpxwq	**R**	ea+toiásöméxdrüzknbjcvhulgfpywq
E	ltnrg+kmszjidahpbuáfvéeocüxüywq	**S**	z+aetéáosiömbunkürlhvgpcdfyjxwq
É	snrgltkpbzmv+djihaéefácoöyuüxwq	**T**	+eatáoiöéujshrlkvbnümycfpgdzxwq
F	eoiöéauürháj+lstnkzmgdybvfpcxwq	**U**	sltndrgkj+mhceazpboáviéöyufüxwq
G	y+eaéoáitnlsjgözuhkrfümbvdcpxwq	**Ü**	lkn+ztgsrvefámaioédöybujhpcüxwq
H	aoeá+iötéulürknshgzbjmpcvdyfxwq	**V**	aeáiéoöhütu+lnrvskzmgdybjfpcxwq
I	+stknazlgrámdevbfhcopéiuüjöyxwq	**W**	e+atlsniokrzámégdöybvujhfpcüxwq
J	aeoáltdu+ézöüsirbnkfmgyvjhpcxwq	**X**	+étpealsnokrzámgdöybvujhfcüxwq
K	+oaeiöéuákütrbmnslhpvdjfczgyxwq	**Y**	+eoaiásétüzöbmvuhrkfnjlgdypcxwq
L	e+atiloyámökésdnjgvhurfcübpzxwq	**Z**	+eotaiámésödzüunkvlbghrjpcyfxwq
M	eai+áuébomtülnzöspcrkfhygvjdxwq	**köz**	Amekstfvéhinrlpbjuáocdögzüywqx+
N	+aáetéödiugjnsockbüyhzvlmfprwxq		

3.4. táblázat: magyar betűk feltételes gyakorisági sorrendje

A 3.3.c. esetben a magyar nyelv alapos ismerete segíthet bennünket. A gyakoriságok tapasztalati meghatározása most szóba sem jöhet, olyan nagy minta kellene.

IV. Az egyszerű helyettesítéses titkosírás fejtése



4.1. Bevezetés

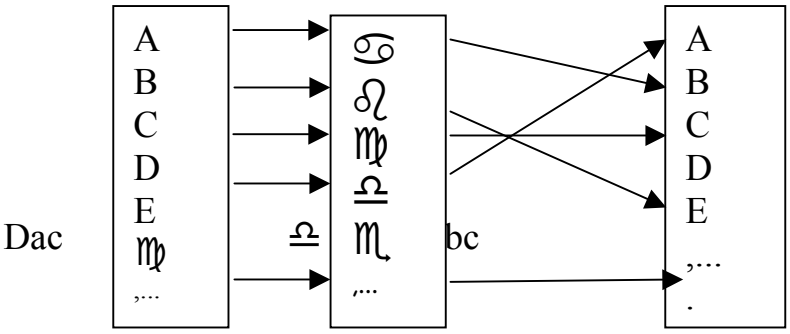
A sztochasztika oktatása során rendkívül fontos feladat annak elérése, hogy a tanulók felfedezzék a statisztika felhasználhatóságát gyakorlati feladatok megoldására. Ennek egyik hatékony, egyben hatásos módszere a titkosított szövegek megfejtésében való felhasználhatóság. A legegyszerűbb titkosírás az **egyszerű helyettesítés**. Itt egy szöveg *titkosítása során a szöveg ábécéjének minden betűjét a titkosításra használt ábécé valamelyik betűjével helyettesíti, különböző szövegbetűket különböző kód betűkkel, ugyanazokat a betűket pedig mindig ugyanazzal a kódbetűvel.*

Ez a titkosírás már az általános iskola III-IV. osztályos tanulói számára ismert, megszokott titkosítási mód. Ők általában kódábécéként egy önállóan tervezett, különböző jelekből, ábrákból álló halmazt használnak. Pusztán megemlítjük, hogy a tanár tanulói tiszteletét nagymértékben növelheti, ha demonstrálja, hogy az így készített titkos szövegeket ő képes elolvasni. Felső tagozatos és középiskolás tanulók számára még a megoldhatóság ténye sem jelent meglepetést.

A titkosítás folyamatának egyszerűsítése és az áttekinthetőség miatt jelen ismertetésben a diplomáciában használt átírást követtük. Ennek során a magyar nyelv hosszú magánhangzóit rövid párjukkal helyettesítettük, a számokat betűkkel írtuk ki és az írásjeleket elhagytuk. Végül valamennyi betűt nagybetűvel írtunk, és a szóközt + jellel jelöltük.

4.2. A kódábécé megválasztása

Különösen az alsó tagozatos tanulók rendkívül leleményesek a kódábécé halmazának megválasztásában. Megdöbbentő számukra annak megmutatása, hogy a kódábécé megválasztása lényegtelen. Ezt **formális matematikai tétel-bizonyítás** formájában bizonyíthatjuk be. Így egy *korai, első matematikai bizonyítást* tudunk bemutatni már alsó tagozatos tanulóknak.



4.1. ábra: Kódábécé átírása

4.1. Tétel. Egy egyszerű helyettesítéses titkosírás megfejthető bármilyen kódábécé esetén; megfejthető akkor is, ha a kódábécé megegyezik a forrás (szöveg) ábécéjével.

Bizonyítás. Helyettesítsük a titkos szövegben megjelenő első „betűt" a forrásábécé első betűjével, majd a titkos szövegben megjelenő második „betűt" a forrásábécé második betűjével. Folytassuk ezt mindaddig, amíg valamennyi titkos „betűt" át nem írtuk. Így ismét az eredeti szöveg egy egyszerű helyettesítéssel átírt titkosítását kapjuk, amelyben valamennyi betű a kódábécéből kerül ki. Nyilvánvalóan a két titkos szöveg egyszerre fejthető, vagy nem fejthető meg.

4.3. Betűstatisztikák tanulói meghatározása

Jelen fejezetben ismertetjük a gimnáziumi 9.-10. osztályban több éven át folytatott gyakorlatunk tapasztalatait. Ezek szerint az osztályközösségek maguk javasolják a nyelv betűstatisztikájának alkalmazását. Az előző fejezetben bemutatott betűgyakoriságok meghatározásának igénye, és azok meghatározása természetes. Nyilvánvaló számukra a következő (durva) állítás:

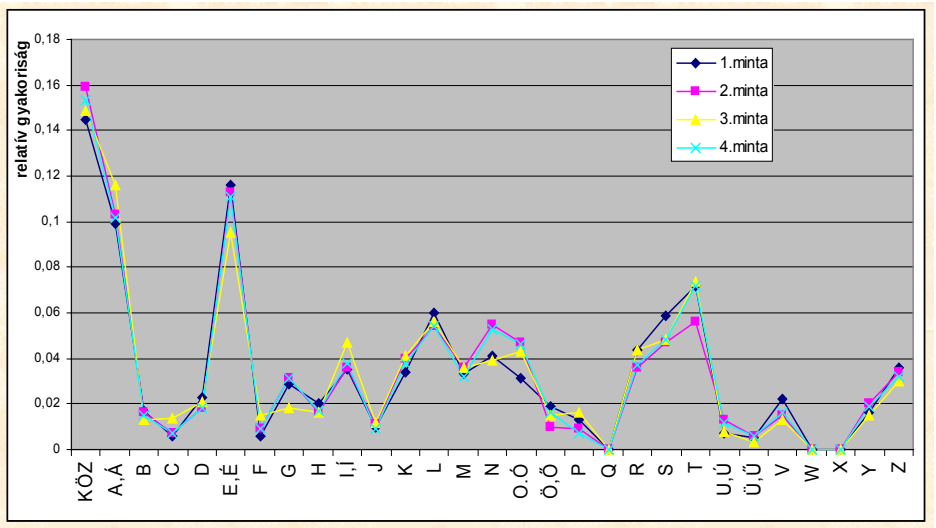
4.1. tanulói állítás: Hosszú magyarnyelvű szövegekben az egyes betűk lényegében ugyanakkora gyakorisággal fordulnak elő. Az egyes gyakoriságok szövegről szövegre kicsit változnak.

Arról, hogy mekkorák ezek a gyakoriságok, vagy egyáltalán milyen a gyakorisági sorrendjük, még egyetlen osztályomban sem tudtak a tanulók megegyezni. Viszont a cél (megoldás) érdekében általában ugyanúgy hajlandók hosszabb szövegekből statisztikák manuális meghatározására is, mint a hiányos szövegek kiegészítésével kapcsolatos feladatok során. Megjegyezzük, hogy már ekkor javasolták a betűstatisztikán kívül más típusú statisztikák felhasználását és ehhez szükséges meghatározásukat. Mindenütt, ahol a jelen fejezetben ez szükségesnek látszik, visszautalunk az előző fejezet gyakorisági táblázataira.

4.4. Grafikus ábrázolás

A valószínűségszámítás és a matematikai statisztika eddigi oktatásában jelentős időt kellett eltölteni a grafikus ábrázolás lehetőségeinek bemutatásával. A személyi számítógépek elterjedése, a családokban való megjelenése, valamint a PC-k alkalmazásra való minél tökéletesebb felkészítése ezt a feladatot rendkívül leegyszerűsítette. A PC programok tartalmaznak kész programokat grafikonok rajzolására, illetve a megfelelő típus kiválasztására.

Az utóbbi években minden osztályban található néhány olyan tanuló, aki a grafikon-rajzolásban már megfelelő ismerettel rendelkezik. Az ő kezdeményezésükre szinte mindig felmerül a betűgyakoriságok grafikus ábrázolásának igénye. Bevonásukkal a tanári munka hatékonysága megsokszorozható, gyakorlatilag egyetlen tanóra alatt átadható minden szükséges ismeret. A 4.2. ábra betűgyakorisági poligonját tanulók készítették számítógépes program segítségével.



4.2. ábra: Magyar betűk gyakoriság-poligonja

4.5. Példa az egyszerű helyettesítésre

Szemléltetésként bemutatunk egy példát, amely egyik lapunkban megjelent két viccet tartalmaz a fentebb ismertetett átírással, tízes csoportosításban.

AZ+UJSAG+R	OVATSZERKE	SZTOJE+LEV	ELET+MUTAT	+A+FOSZERK
ESZTONEK+U	RAM+A+SZAH	ARAI+TUDOS	ITONK+MEGI	NT+VIZHIAN
YROL+PANAS	ZKODIK+UGY	AN+HISZEN+	ALLANDOAN+	EZT+A+NOTA
T+FUJJA+IG	EN+DE+MOST	+AZ+EGYSZE	R+KOMOLYNA	K+LATSZIK+
A+DOLOG+ME	RT+A+BELYE	GET+BIZTOS	ITOTUVEL+E	ROSITETTE+
A+BORITEKR	A+EGY+MASI	K+VICC+REP	ULOGEP+KOZ	ELEDIK+AZ+
EJSZAKA+KO	ZEPEN+A+RE	PULOTER+FE	LE+A+PILOT	A+ELHATARO
ZZA+HOGY+M	EGTREFALJA	+A+TORNYOT	+EZERT+ELT	ORZITOTT+H
ANGON+SZOL	+A+MIKROFO	NBA+TALALD	+KI+HOGY+K	I+VAGYOK+E
RRE+AZ+IRA	NYITOTORON	Y+DISZPECS	ERE+LEKAPC	SOLJA+A+LE
SZALLOPALY	A+FENYEIT+	ES+VISSZAS	ZOL+TE+MEG	+TALALD+KI
+HOGY+HOL+	VAGY+A+VIC	CEK+AZ+UGY	ES+HAHOTA+	ROVATABOL+

4.1 táblázat: Az eredeti szöveg

Ezt a szöveget rejtjeleztük egyszerű helyettesítéssel. A kapott rejtjeles szöveget szintén tízes csoportosításban adjuk meg:

+UZRAX+JZE	FK+NXUOELO	XUNFAOZGOK	OGONZCRN+N	+ZSFXUOEL
OXUNFBOLZR	E+CZ+ZXU+I	+E+QZNRWFX	QNFBLZCOJQ	BNZKQUIQ+B
TEFGZM+B+X	ULFWQLZRJT	+BZIQXUOBZ	+GG+BWF+BZ	OUNZ+ZBFN+
NZSRAA+ZQJ	OBZWOZCFXN	Z+UZOJTXUO	EZLFCFGTB+	LZG+NXUQLZ
+ZWFGFJZCO	ENZ+ZPOGTO	JONZPQUNFX	QNFNRKOGZO	EFXQNONNOZ
+ZPFEQNOLE	+ZOJTZC+XQ	LZKQDDZEOM	RGFJOMZLFU	OGOWQLZ+UZ
OAXU+L+ZLF	UOMOBZ+ZEO	MRGFNOEZSO	GOZ+ZMQGFN	+ZOGI+N+EF
UU+ZIFJTZC	OJNEOS+GA+	Z+ZNFEBTFN	ZOUOENZOGN	FEUQNFNNZI
+BJFBZXUFG	Z+ZCQLEFSF	BP+ZN+G+GW	ZLQZIFJTZL	QZK+JTFLZO
EEOZ+UZQE+	BTQNFNFEFB	TZWQXUMODX	OEOZGOL+MD	XFGA+Z+ZGO
XU+GGFM+GT	+ZSOBTOQNZ	OXZKQXXU+X	UFGZNOZCOJ	ZN+G+GWZLQ
ZIFJTZIFGZ	K+JTZ+ZKQD	DOLZ+UZRJT	OXZI+IFN+Z	EFK+N+PFGZ

4.2. táblázat: A rejtjelzett szöveg

4.6. Tanulói javaslatok a fejtésre

A statisztika alkalmazhatóságára való figyelemfelkeltés a jelen témakör oktatásának egyik alapvető célja. Erre vonatkozóan megnyugtató megfigyelni, hogy valamennyi olyan osztályban, ahol az egyszerű titkosírás megfejtését kezdeményeztük, a *tanulók maguk javasolták betűstatisztikák alkalmazását.* Ennek során megállapodtak abban, hogy az egyes betűk a magyar nyelvben különböző gyakorisággal fordulnak elő, és az előfordulásuk szerinti gyakorisági sorrend hosszabb szövegek esetén lényegében ugyanaz. A „lényegében ugyanaz” vélemény általában hosszas vitákat vált ki. Ebből a szempontból hasznos a 3.1. táblázat elemzése, a 40 tanuló által készített gyakoriság eloszlások összehasonlítása. Ezek világosan mutatják, hogy a gyakorisági sorrendek eltérnek az egyes mintákban. A tanulók viszont ezt az állítást önmaguk képesek szűkíteni:

4.2. tanulói állítás: Nagyon hosszú magyar nyelvű szövegekben mindig ugyanaz a leggyakoribb 4-5 betű. Speciálisan a leggyakoribb magyar betű a szóköz.

Különböző, - főként rövidebb- rejtjeles szövegek megoldása során a tanulók hamar tapasztalják, hogy csak a rejtjeles szövegek néhány gyakoribb betűje feleltethető meg az eredeti szövegekben előforduló gyakori betűk valamelyikének. Ezért javasolják a leggyakoribb betűk önálló vizsgálatát. Egyik tipikus javaslat az egymást követő azonos betűk közti távolság

vizsgálata, ami a leggyakoribb rejtjeles betűkhöz a leggyakoribb magyar betűk hozzárendelésében segíthet.

A következőkben az a, e, t betűkre nyert tanulói adatokat mutatunk be. A három leggyakoribbnak tartott betű egymás közti megkülönböztethetőségére javasolták az egymás utáni betűk közti távolságok eloszlásának kihasználását.

A gyakoriság-eloszlásokat egy 21 tanulós csoportban nem kötelező házi feladatként készítették el. Érdekességként rámutatunk, hogy az *a* betűk távolság-eloszását valamennyi tanuló elvégezte, az *e* betűét 14-en, a *t* betűét viszont csak kilencen határozták meg. Az eredményeket a 4.3.a,b,c táblázatok ismertetik.

További érdekes (és hasznos) javaslat a *szóköz-a-e-t* jelnégyesből alkotott rendezett betűpárok nyílt szövegben előforduló relatív gyakoriságainak meghatározása. Indoklása az, hogy két szóköz soha nem fordul elő egymás mellett, két *a*, két e egymás mellett gyakorlatilag soha nem fordul elő, közülük két t egymás mellett viszonylag gyakori, a szóköz-a, szóköz-e párok (mind ebben, mind a fordított sorrendben) szintén viszonylag gyakoriak.

Távolság:	0	1	2	3	4	5	6	7	8	9	10	>10	>20
1	0	2	8	0	1	4	2	1	1	0	2	7	2
2	1	7	6	2	3	1	2	0	2	0	0	1	5
3	0	5	5	1	1	1	1	0	2	2	3	7	2
4	0	3	4	2	1	1	2	2	1	2	1	7	4
5	0	1	8	1	2	2	0	2	0	2	1	8	3
6	0	0	4	6	5	3	6	0	0	0	1	2	3
7	0	4	5	0	2	1	2	1	0	2	0	3	10
8	0	1	8	1	2	2	0	2	0	2	1	8	3
9	0	1	3	2	0	1	4	1	0	2	1	5	10
10	0	3	3	2	3	0	1	0	1	1	1	7	8
11	1	4	4	4	3	0	1	0	0	1	0	6	6
12	1	7	6	1	1	2	2	0	1	2	0	4	3
13	0	1	6	3	2	1	2	0	1	0	1	3	10
14	0	5	3	3	1	2	1	1	0	0	0	6	8
15	0	3	4	2	1	3	2	3	0	1	0	6	5
16	0	1	6	2	3	2	1	2	2	2	0	6	3
17	0	5	2	2	2	0	1	2	2	0	1	4	9
18	1	3	2	0	0	0	1	0	2	5	1	11	4
19	0	3	7	2	1	2	1	1	0	2	0	6	5
20	0	2	5	0	2	2	1	3	2	0	1	3	9
21	1	4	1	3	2	0	1	0	3	0	2	4	9
Összes	5	65	97	39	38	30	34	21	20	26	17	114	121

4.3.a táblázat: Egymás utáni ***a*** betűk közti távolság-eloszlás

Távolság:	0	1	2	3	4	5	6	7	8	9	10	>10	>20
1	0	2	10	2	0	0	2	1	2	0	0	7	4
2	0	7	6	5	0	2	0	1	1	0	0	5	3
3	0	3	8	2	4	0	0	1	3	1	1	4	3
4	0	8	7	1	2	3	0	1	1	1	2	1	3
5	0	8	7	0	0	0	2	2	0	1	0	5	5
6	1	6	4	3	1	3	2	0	1	3	1	1	4
7	0	5	6	1	2	1	2	2	0	0	1	6	4
8	0	3	3	2	1	2	3	4	0	1	1	2	8
9	0	8	5	4	1	2	1	1	1	1	1	4	1
10	0	5	10	3	1	0	0	1	1	0	3	3	3
11	0	4	4	1	3	0	2	2	0	1	1	6	6
12	0	8	3	3	1	2	2	1	0	0	1	3	6
13	1	3	4	4	1	3	3	1	0	0	1	3	6
14	0	7	3	0	3	0	1	0	0	0	1	8	7
Össz:	2	77	80	31	20	18	20	18	10	9	14	58	63

4.3.b táblázat: Egymás utáni ***e*** betűk távolság-eloszlása

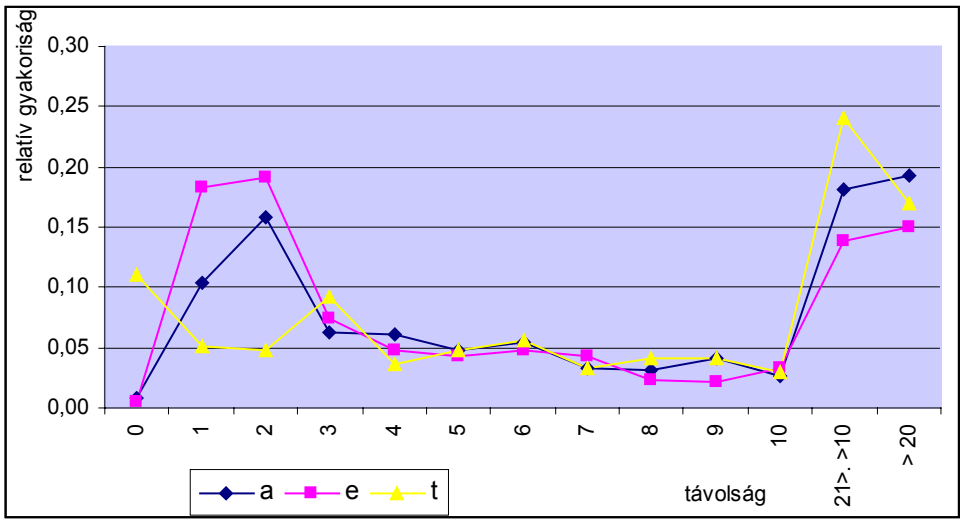
Táv:	0	1	2	3	4	5	6	7	8	9	10	>10	>20
1	4	1	0	2	1	0	1	1	2	0	3	6	9
2	4	0	4	1	2	2	1	0	3	1	1	9	2
3	3	2	2	6	3	2	4	1	1	2	0	2	2
4	2	1	2	3	0	3	0	1	1	2	2	9	4
5	2	4	1	4	0	1	2	2	1	3	0	7	3
6	4	3	1	0	1	2	4	3	0	1	1	5	5
7	3	1	0	5	2	1	0	0	1	0	1	11	5
8	4	1	3	1	1	2	1	0	1	1	0	9	6
9	4	1	0	3	0	0	2	1	1	1	0	7	10
Össz:	30	14	13	25	10	13	15	9	11	11	8	65	46

4.3.c táblázat: Egymás utáni **t** betűk közti távolságok

A 4.3. ábrán az ezekből nyert egyesített eloszlásoknak megfelelő relatív gyakorisági poligonokat készítettük el.

Táv.	0	1	2	3	4	5	6	7	8	9	10	10<.<20	.≥20
a:	5	65	100	39	38	30	34	21	20	26	17	114	121
e:	2	77	80	31	20	18	20	18	10	9	14	58	63
t:	30	14	13	25	10	13	15	9	11	11	8	65	46

4.4. táblázat: Az összesített eloszlásnak megfelelő gyakoriságok



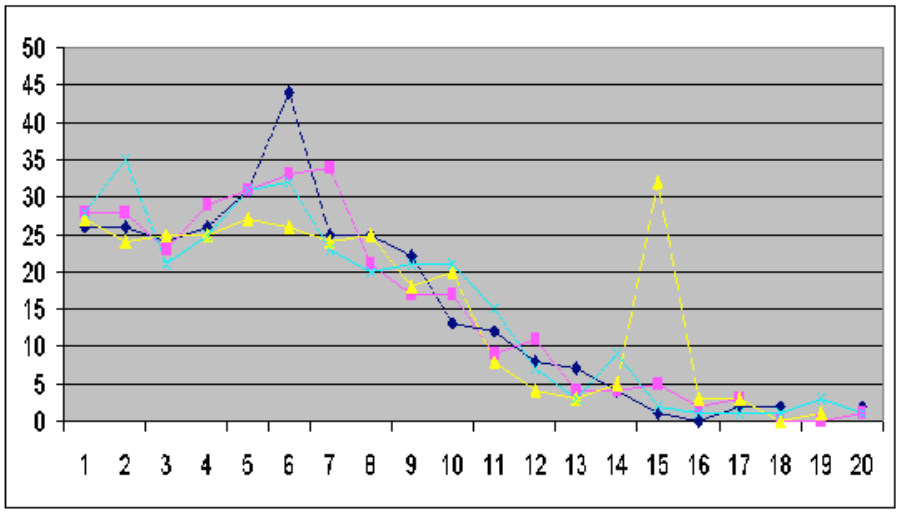
4.3. ábra: egymás utáni a,e,t betűk közti távolságok relatív gyakorisági poligonja

Szintén fontos jellemzője minden nyelvnek, hogy milyen hosszú szavakból áll. (Ebből egyébként a szóközök közti távolságokat tudjuk meg.)

A szóhosszak gyakoriságára mutat be a 4.5.a. táblázat négy darab 300 elemű mintát, amit a 4.4.a.ábrán szemléltettünk.

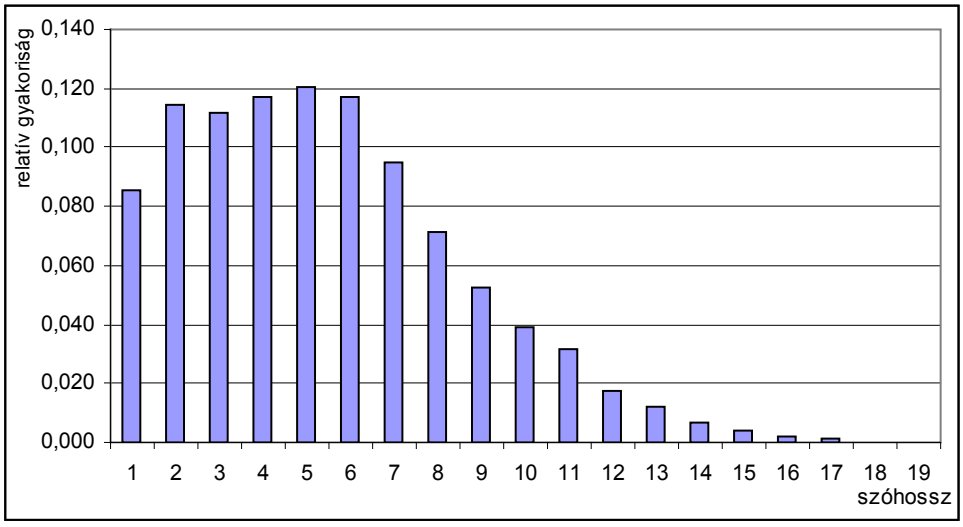
Hossz:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1. minta	26	26	24	26	31	44	25	25	22	13
2. minta	28	28	23	29	31	33	34	21	17	17
3. minta	27	24	25	25	27	26	24	25	18	20
4. minta	28	35	21	25	31	32	23	20	21	21
Össz:	109	113	93	105	120	135	106	91	78	71
Hossz:	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1. minta	12	8	7	4	1	0	2	2	0	2
2. minta	9	11	4	4	5	2	3	0	0	1
3. minta	8	4	3	5	32	3	3	0	1	0
4. minta	15	7	3	9	2	1	1	1	3	1
Össz:	44	30	17	22	40	6	9	3	4	4

4.5. táblázat: A szóhosszak gyakorisága négy mintában



4.4.a ábra: A szóhossz-eloszlás gyakorisági poligonja négy mintában

A szóhosszak eloszlására megadjuk egy másik 4000 szóból készült eloszlás hisztogramját is a 4.3.b. ábrán.



4.4.b ábra: A szóhossz-eloszlás gyakorisági hisztogramja

Megoldási ötletek a rejtjelzett szöveg fejtéséhez:

Mindenekelőtt próbáljuk azonosítani a magyar nyelv négy leggyakoribb betűjét, a szóközt, az *e, a, t* betűket. Próbáljuk ezeket rejtjeles szöveg négy leggyakoribb betűjével azonosítani.

Az egyszerű helyettesítéssel nyert *rejtjeles* szöveg gyakoriságai:

+	A	B	C	D	E	F	G	H	I	J	K	L	M
64	7	20	10	6	29	47	23	0	1	18	11	22	9
N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
42	57	5	28	10	0	20	29	22	11	17	42	89	

4.6. táblázat: A rejtjeles szöveg gyakoriságai

A hét legnagyobb rejtjeles gyakoriság:

Z:89	+:64	O:57	N:42	E:29	U:29	Q:28

4.7. táblázat: A hét legnagyobb rejtjeles gyakoriság

Helyettesítsük a titkos szöveg leggyakoribb jelét a szóközzel, és ennek megfelelően szabdaljuk fel a titkos szöveget a feltételezett szavakra. Mielőtt ezt megtennénk, megszüntetjük a tízes csoportosítást biztosító szeparálást (tehát kiírtjuk a nem-valódi szóközöket). Az Enter jelet figyelmen kívül hagyjuk.

```
+U RAX+J EFK+NXUOELOXUNFAO GOKOGON CRN+N
 + SFXUOELOXUNFBOL RE+C + XU+I+E+Q NRWFX
QNFBL COJQBN KQUIQ+BTEFG M+B+XULFWQL RJT
+B IQXUOB +GG+BWF+B OUN + BFN+N SRAA+ QJ
 OB WO CFXN +U OJTXUOE LFCFGTB+ G+NXUQL
+ WFGFJ COEN + POGTOJON PQUNFXQNFNRKOG O
EFXQNONNO + PFEQNOLE+ OJT C+XQL KQDD EOM
RGFJOM LFUOGOWQL +U OAXU+L+ LFUOMOB + EO
MRGFNOE SOGO + MQGFN+ OGI+N+EFUU+ IFJT C
OJNEOS+GA+ + NFEBTFN OUOEN OGNFEUQNFNN I
+BJFB XUFG + CQLEFSFBP+ N+G+GW LQ IFJT L
Q K+JTFL OEEO +U QE+BTQNFNFEFBT WQXUMODX
 OEO GOL+MDXFGA+ + GOXU+GGFM+GT+ SOBTOQN
 OX KQXXU+XUFG NO COJ N+G+GW LQ IFJT IFG
 K+JT + KQDDOL +U RJTOX I+IFN+ EFK+N+PFG
```

Amennyiben a szóköz feltételezése helyes, a többször is előforduló ***szóköz - +jel – szóköz*** betűhármas alapján az *A* betűt a titkos szövegben a + jelnek kell jelölnie, és a + jel valóban gyakori, 64-es gyakoriságával a második leggyakoribb jel. Bár tökéletes egyezés esetén ez a jel az *E* betűt helyettesítené, de hasznosabbnak látszik a + jelet az *A* betű megfelelőjének

tekinteni. Ekkor viszont a harmadik leggyakoribb rejtjeles betűt tételezhetjük fel az *E* betű képének. Ez a rejtjeles *O* betű.

Egy további helyettesítés feltételezése is jogosnak tűnik. Mivel a *szóköz—**+U**—szóköz* jelnégyes háromszor is előfordul, feltehetjük, hogy a *+U* pár az „az" névelőt jelöli. A következő leggyakoribb rejtjeles betű az *N*, 42-es gyakorisággal. Próbáljuk ezt a nyílt (értelmes) szöveg következő leggyakoribb betűjével, a *T*-vel kicserélni. Most ezt a négy betűt próbáljuk a szövegbe illeszteni. Ezeket kis betűvel írjuk az esetleges félreértések elkerülése miatt.

```
az RAXaJ EFKatXzeELeXztFAe GeKeGet CRtat
 a SFXzeELeXztFBeL REaC a XzaIaEaQ tRWFX
QtFBL CeJQBt KQzIQaBTEFG MaBaXzLFWQL RJT
aB IQXzeB aGGaBWFaB ezt a BFtat SRAAa QJ
 eB We CFXt az eJTXzeE LFCFGTBa GatXzQL
a WFGFJ CeEt a PeGTeJet PQztFXQtFtRKeG e
EFXQtette a PFEQteLEa eJT CaXQL KQDD EeM
RGFJeM LFzeGeWQL az eAXzaLa LFzeMeB a Ee
MRGFteE SeGe a MQGFta eGIataEFzza IFJT C
eJtEeSaGAa a tFEBTFt ezeEt eGtFEzQtFtt I
aBJFB XzFG a CQLEFSFBPa taGaGW LQ IFJT L
Q KaJTFL eEEe az QEaBTQtFtFEFBT WQXzMeDX
 eEe GeLaMDXFGAa a GeXzaGGFMaGTa SeBTeQt
 eX KQXXzaXzFG te CeJ taGaGW LQ IFJT IFG
 KaJT a KQDDeL az RJTeX IaIFta EFKataPFG
```

A kialakuló „*ezt*", és „*te*" szavak igazolni látszanak a *t*-re vonatkozó hipotézist. Újabb azonosítást tesz lehetővé a 10. sorban kialakult *ezeEt* csoport. Az *E* betű helyébe csak az *r*-t írva kapunk értelmes szót: *ezeEt = ezért*. Megerősíti ezt az „*eEEe*" betűnégyes is: *eEEe = erre*. Ugyancsak gyakori a „*Xz*" betűpár, ami az *X=S* helyettesítést valószínűsíti. A 4. Sorban található „*CeEt*" → *Cert* szó a *C* → *M* helyettesítést indikálja. Hajtsuk végre ezeket a helyettesítéseket is!

az RAsaJ rFKatszerLesztFAe GeKeGet mRtat
a SFszerLesztFBeL Rram a szaIaraQ tRWFs
QtFBL meJQBt KQzIQaBTrFG MaBaszLFWQL RJT
aB IQszeB aGGaBWFaB ezt a BFtat SRAAa QJ
eB We mFrt az eJTszer LFmFGTBa GatszQL
a WFGFJ mert a PeGTeJet PQztFsQtFtRKeG e
EFsQtette a PFrQteLra eJT masQL KQDD reM
RGFJeM LFzeGeWQL az eAszaLa LFzeMeB a re
MRGFter SeGe a MQGFta eGIatarFzza IFJT m
eJtreSaGAa a tFrBTFt ezert eGtFrzQtFtt I
aBJFB szFG a CQLrFSFBPa taGaGW LQ IFJT L
Q KaJTFL erre az QraBTQtFtFrFBT WQszMeDs
ere GeLaMDsFGAa a GeszaGGFMaGTa SeBTeQt
es KQsszaszFG te meJ taGaGW LQ IFJT IFG
KaJT a KQDDeL az RJTes IaIFta rFKataPFG

Néhány további ötlet:

mRtat, Rram, (mutat, uram) : *R → u*
meJ (még), *eJT* (egy): *JT → gy*
szerLeszt (szerkeszt), *GatszQL* (látszik) stb.

Nem folytatjuk tovább a lehetőségek számba vételét. Meg kell azonban jegyezni, hogy számos, a fentiekhez hasonló ötlet merül fel minden osztályban, és a tanulók nagy kedvvel oldják meg az ilyen típusú feladatokat.

Az ellenőrzés lehetőségének biztosítására ugyancsak megadjuk a rejtjeles betűk nyílt megfelelőit.

+	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	J	N	M	C	R	O	L	W	H	G	V	K	P	T	E	B	I	U	F	Y	Z	Q	D	S	X	+

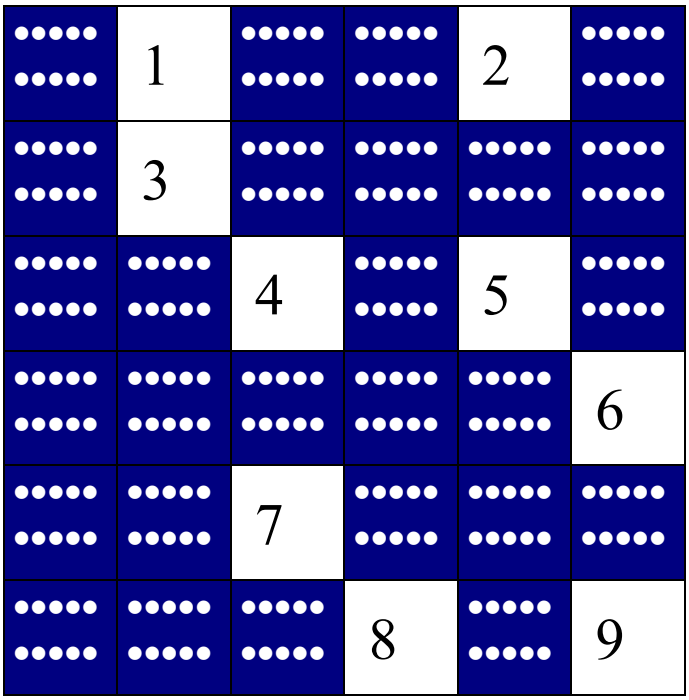
A teljesség kedvéért ugyancsak a kódoláshoz (rejtjelzéshez) használt **NYÍLT → REJTJELES** megfeleltetést.

+; A; B; C; D; E; F; G; H; I; J; K; L; M;N; O; P; Q; R; S; T; U; V; W; X; Y;Z
Z; +; P; D; W; O; S; J; I; Q; A; L; G; C; B; F; M;V; E; X;N; R; K; H; Y; T; U

A bemutatott tanulói megoldás járhatóságának ellenőrizhetősége céljából megadunk egy rejtjeles szöveget, amelyben a rejtjeles szöveget ötös betűcsoportokba osztottuk:

```
pe cd bcmgs vbepm Wevxq och"s WWmco r ,dq cjh"p
h y"W cdqqd ocdcv byov, khvv, khlrk ;cdc" hm Wc
krqgr "dWrW ch" h NWv,o cdc,h cqhpc dzdWr WWdoc
vbepe  d, W lhWhW "hqcv bszhk clrpe ";e,d cNh"h
q"hkm cmvph  hWhm qocvb h mqW cdWmW orvmW evroc
ldvbq e"dWd cdbcr or ,d cq;x" moczm vvbdc dcomq
dmcdc th bv dcdbm qgmdm c,vch k;,,c fmzm" mbefm
rocm" ;hqcW xgrpe q;e r "cvdN qrvcf vdocm khqoh
zhvhW cWxgxq
```

V. Poziciócserés titkosító eljárás fejtése



5.1. Bevezetés

Ebben a fejezetben a pozíciócserés titkosírás megfejtésének egy lehetséges módját mutatjuk be. Célunk változatlan: a statisztika alkalmazására szeretnénk egy középiskolai szinten könnyen követhető eljárást mutatni, amely bizonyos statisztikai adatgyűjtés szükségességét is mutatja. A valószínűségi fogalmak közül lehetőség nyílik az együttes eloszlás, feltételes eloszlás szemléletes bevezetésére.

Alapvető számítógépes programozásra is szükség lehet. Kirándulásra kínálkozik alkalom a permutációk világába, valamint gráfelméleti területekre is.

5.2. Keverés: Pozíciók felcserélésén alapuló titkosító eljárás

5.1. Definíció: A pozíciócserés eljárás: Ez az ókor egyik gyakori rejtjelző eljárása. A nyílt szöveget rögzített hosszúságú blokkokra osztják, ezeken a blokkokon belül az egyes pozíciókban levő betűket egymással kicserélik, minden blokkban ugyanazon szabály szerint. (Az eljárást a magyar gyakorlatban keverésnek, angolul: transposition-nak nevezik). A blokkokon belül az egyes pozíciókban levő betűket tehát egy előre megválasztott, feladó-vevő által ismert permutáció szerint összekeverik.

Az utolsó blokk általában nem teljes. A kiegészítésre többféle módszert alkalmaznak, amelyek nem játszanak lényeges szerepet jelenlegi vizsgálatainkban. Mivel mind a küldő, mind a vevő pontosan ismeri az alkalmazott permutációt, így a kódolás, dekódolás nem jelent problémát. Az ilyen típusú titkosítások még a század elején is divatosak voltak, és általánosan használták őket.

Ez az eljárás betűpárokra vonatkozó statisztikák segítségével fejthető, tehát az egyszerű helyettesítéshez hasonló, bár kissé bonyolultabb statisztikai módszer alkalmazható a fejtéshez. Ez az alkalmazás a statisztikai hipotézis vizsgálatok területére vezet el bennünket.

5.3. Sándor Mátyás rácsos levele

Érdekességként említjük meg, hogy Verne "Sándor Mátyás" című regényében alkalmazott "rácsos" rejtjelzés is ebbe a kategóriába tartozik. [N97] (A 6x6-os Verne rácshoz hasonlót mutat be a fejezet bevezető oldala.) Ennek a megállapításnak a megvizsgálására számos tanuló önként áldozza szabadidejét, tehát még a fejtés lehetőségének megismerése előtt sikerül motiválni a tanulókat.

A regényben egy postagalamb vitt egy titkos üzenetet, ami balszerencse folytán Sárkány kezébe került. Az üzenet csak tizennyolc szóból állt, három oszlopos leírásba rendezve:

R H G A A Y	L K A E E N	S L Ö Ö Ü Y
š Z G G R Ü	N Z E Ö L M	Y K B N E T
A F X S G M	S N E Ö O L	E A K Y L D
N T L Ź R Ü	T K E Y K Z	S R N L Ü E
E Y L Ź R Ü	G A G E A E	I Ź I M R L
S E R Ü O G	E N R G Ü L	N T E Ö Y N

A titkosításhoz és a visszaállításhoz egy 6x6-os „rostélyt” használtak. Ennek bizonyos ablakait kivágták. Az üres négyzeteket olyan ügyesen helyezték el, hogy ha a rostélyt négyszer egy-egy negyed fordulattal elforgatták, a papíron a 6x6=36 négyzet mindegyike egyszer, és csak egyszer maradt látható. A nyílt szöveget (üzenetet) ennek megfelelően egymás után a látható mezőkbe írták (és ugyanígy olvasták ki).

Egy rostélyforgatással 36 betűt lehet leírni, tehát elképzelhetjük, hogy az eredeti nyílt szöveget (üzenetet) első lépésként 36 betűs blokkokba csoportosították, majd a titkosítást külön-külön, de ugyanazon szabály szerint elvégezve, a rejtjeles szöveget újra folyamatosan leírhatjuk, soronként olvasva ki őket. Képzeljük el, hogy az eredeti és a titkosított szöveget egymás alá írjuk. Ekkor a két blokk ugyanazon betűket tartalmazza, mégpedig a második (rejtjeles) blokk az első blokk helyeinek (pozícióinak) cseréjével, permutációjával keletkezett.

Az átfogalmazás a beharangozásnak megfelelően egy általános titkosítási eljáráshoz, a pozíciócserés eljáráshoz vezetett. Bár Verne ezt a rostélyos eljárást elvileg is megfejthetetlennek tekintette, mi megadunk a továbbiakban az általánosabb rejtjelző eljárásra is egy megfejtési módszert. Ma Verne rostélyos eljárása számítógépen akár teljes kipróbálással is megfejthető lenne.

5.4. Példák

Mivel ez az eljárás blokkhossz = n mellett az 1,2,. . . , n számok egy permutációját alkalmazza, természetes lehet egy kirándulás a permutációk területére, így az inverz permutáció (inverz transzformáció) fogalmának bevezetéséhez.

Az alábbi példát B=9 blokkhosszúság esetére mutatjuk be.

5.1. Példa: Tekintsük a $B = 9$ blokkhosszúságot és legyen az alkalmazandó keverő permutáció

$$\mathbf{Perm} = \begin{Bmatrix} 1,2,3,4,5,6,7,8,9 \\ 7,3,4,1,5,8,9,2,6 \end{Bmatrix}.$$

Ekkor a dekódoláshoz ennek inverzét, a

$$\mathbf{Perm}^{-1} = \begin{Bmatrix} 1,2,3,4,5,6,7,8,9 \\ 4,8,2,3,5,9,1,6,7 \end{Bmatrix}$$

permutációt kell használni.

Az eljárás illusztrálására kiírjuk egy nyílt szöveg első 3 blokkját, ismétlődően alájuk írva a permutációt, majd a permutáció eredményeként adódó rejtjeles szöveget. Ilyen blokkméret mellett a titkos szöveg megfejtése nem túl bonyolult munka. Megjegyezzük, hogy tapasztalataink szerint középiskolások ekkora méretű (blokkhosz<10) keverést ad-hoc módszerekkel gyorsan tudtak rekonstruálni.

B	E	T	Ű	P	Á	R	O	K
1	2	3	4	5	6	7	8	9
7	3	4	1	5	8	9	2	6
Ű	O	E	T	P	K	B	Á	R
A	T	+	B	I	G	R	A	M
1	2	3	4	5	6	7	8	9
7	3	4	1	5	8	9	2	6
B	A	T	+	I	M	A	G	R
M	N	A	K	+	N	E	V	E
1	2	3	4	5	6	7	8	9
7	3	4	1	5	8	9	2	6
K	V	N	A	+	E	M	N	E

5.1. táblázat: Példa n=9 blokkhossz melletti keverésre

5.2. Példa: Egy további példában az n=16 blokkhosszat választottuk. Az alábbi nyílt szöveg Knuth [K87] számítástechnikai "bibliájából" származik. Szemléltető példánkban a rejtjeles (csere utáni) szöveg elkészítése során a 31 betűs magyar ABC-t használtuk, tehát a hosszú í, ó, ő, ú, ű betűket rövid párjukkal helyettesítettük. A szóköz jelölésére most a - jelet használtuk.

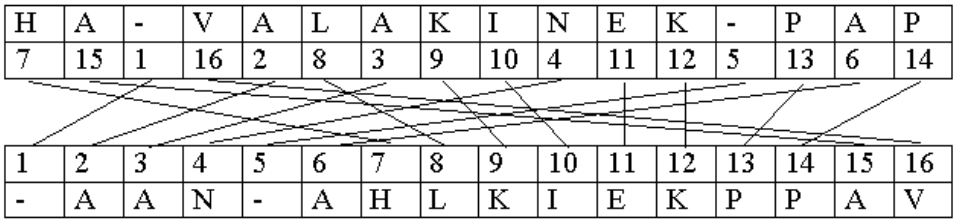
HA-VALAKINEK-PAP	IRT-ÉS-CERUZÁT-A	DNÁNK-A-KEZÉBE-É
S-MEGKÉRNÉNK-HOG	Y-IRJON-LE-SZÁZ-	VÉLETLEN-SZÁMJEG
YET-ALIGHA-TUDNA	-ELFOGADHATOAN-V	ÉLETLENSZERÜ-SOR
OZATOT-KÉSZITENI	-AZ-EMBEREK-SZER	ETIK-KIHAGYNI-AZ
-OLYAN-RÉSZEKET-	AMELYEK-NEM-TÜNN	EK-VÉLETLENNEK-P
ÉLDÁUL-EGYENLÖ-S	ZÁMJEGYEKET-NEM-	IRNAK-EGYMÁS-UTÁ
N-JOLLEHET-EGY-V	ÉLETLEN-SZÁMSORB	AN-ÁTLAGOSAN-MIN
DEN-TIZ-BETÜBÖL-	EGY-MEGEGYEZIK-A	-RÁ-KÖVETKEZÖVEL
-MÁSRÉSZT-HA-VAL	AKINEK-EGY-VALOD	I-VÉLETLEN-SZÁMJ
EGYEKBÖL-ÁLLO-T		

5.2. táblázat: A nyílt szöveg

A helycserét megvalósító permutáció példánkban:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
7	15	1	16	2	8	3	9	10	4	11	12	5	13	6	14

Magát a helyettesítést pusztán az első nyílt blokk esetén mutatjuk meg.



5.1. ábra: Rejtjelzést illusztráló graf

A permutáció alkalmazása után a következő rejtjeles szöveg keletkezett

-AAN-AHLKIEKPPAV	TÉ-R –ISCEUZTAR-	KAEB-D—KZÉEÉNN ?	MGÉÉ-OSKRNNKHG-E
IJNEZZYO-L-S --R	LTESMEVLN-Z JGÉE	TAIAUNYLGH-TDAE-	LOAAA--GDHTONVEF
ELNE-OÉESZRÜSRLT	AO-STNOTKÉZIEIZT	ZEBESE-MERK-ZRA-	I-IGIAEKHAYN-ZTK
LA-SKT-NRÉZEE-OY	EYKETNAE-NM-ÜNML	-ÉEEE-ELTLNNKPKV	DU-YL-ÉLEGENÖSL ?
MEYENMZGEKT-E- J	NKEM-TI-GZ SU RA	JLETG-NLHE-EYV-O	ELNZSRÉE-S MOBLT
-TAS-IALGOANMNN ?	NTZEBLDI-BTÜÖ-E-	YMGYI-EEEGEZKAG-	KVKÖE-ÖETEZVLR- ?
RS--A-ÉZTHAVLMS ?	IE-YAOAKEG-VLDKN	VLTNZMIELE-S J-É	YKÖÁOTEBL-LL- GE

5.3. táblázat: Az 5.2. példa keveréssel keletkezett rejtjeles szövege

Ezt a 28 blokkból álló rejtjeles üzenetet fogjuk a következő pontban egy lehetséges fejtési eljárás bemutatására felhasználni.

5.5. A transzpozíciós eljárás fejtése: hipotézisvizsgálatok láncolata

Ha a blokkhossz = *n*, akkor a lehetséges keveréses eljárások száma *n*!. Ez a szám még akkor is óriási, ha *n* kicsi, mondjuk *n* = 9. A teljes kipróbálás lehetetlen.

Természetes adódik az alkalom arra, hogy bemutassuk a matematika egy szokásos trükkjét. Ilyen esetekben a feladatot apróbb, áttekinthetőbb kis feladatok láncolatára bontjuk, miközben ezek megoldásainak egymás utáni kombinálása az eredeti feladat megoldását is eredményezi. Esetünkben ilyen apróbb kérdések a következők:

A feltett számos (*n*) kérdés mindegyikére csak *n* válasz lehetséges:

- vagy a blokk másik *n-1* betűje követi a nyílt blokkban ezt a betűt,
- vagy ez volt a nyílt blokk utolsó betűje.

(A válasz nyilván nem függ attól, hogy melyik blokkról van szó.)

Az is nyilvánvaló, hogy ha a választ ismerjük minden adott helyen levő rejtjeles (áthelyezett) betű esetén, akkor egy lehetséges eltolástól eltekintve az eredeti szöveg visszakapható. Maga az esetleges eltolás egyértelművé tehető annak az ellenőrzésével, hogy megállapítjuk, melyik eltolás ad értelmes szöveget.

A most következőkben ismertetjük az osztályközösségekben az egész osztályra kiterjedő megtárgyalás (vita) során kialakuló módszert. Ez az általánosabb, matematikailag megalapozott "legjobb" megoldás egy leegyszerűsített változata. Az általunk kidolgozott megoldást ezután külön fejezetben ismertetjük.

Első lépésként a rejtjeles blokkok első betűivel kezdve próbáljuk megválaszolni ezeket a kérdéseket szisztematikusan. Munkahipotézisként tegyük fel, hogy a nyílt szövegben a rájuk következő betűk a rejtjeles blokkok második betűi. Ennek megfelelően gyűjtsük ki ezt a 28 betűpárt! Ezek:

rejtjeles első—rejtjeles második

-A; TÉ; KA; MG; IJ; LT; TA; LO; EL; AO; ZE; T-; LA; EY;
-É; DU; ME; NK; JL; EL; -T; NT; YM; KV; RS; IE; VL; YK

Hasonlóan kigyűjthetjük a

rejtjeles első—rejtjeles harmadik

betűpárokat! Ezek:

-A; T-; KK; ME; IN; LE; TI; LA; EN; A-; ZB; II; L-; EK; -E;
D-; MY; NE; TE; JE; EN; -A; NZ; YG; YÖ; R-; I-; VT; YÖ

Ugyanilyen betűpár-halmazt gyűjthetünk ki az *első-negyedik*, ..., *első-tizenhatodik* rejtjeles párokról is. A párokat a 5.4. táblázat mutatja. A betűpárok első betűit a 5.4. táblázat első oszlopa adja meg, a többi oszlop rendre a rákövetkező első, második stb ***k*-adik** betűket mutatja.

Az eredeti kérdésünket átalakítva legyen az a kérdés, hogy ezek között melyik áll legnagyobb valószínűséggel egymást követő betűpárokból.?

k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
-	A	A	N	-	A	H	L	K	I	E	K	P	P	A	V
T	E	-	R	-	?	I	S	C	E	U	Z	T	A	R	-
K	A	E	B	-	D	-	-	K	Z	É	E	É	N	N	?
M	G	E	É	-	O	S	K	R	N	N	K	H	G	-	E
I	J	N	E	Z	Z	Y	O	-	L	-	S		-	R	
L	T	E	S	M	E	Y	L	N	-	Z		I	G	É	E
T	A	I	A	U	N	Y	L	G	H	-	T	D	A	E	-
L	O	A	A	A	-	-	G	D	H	T	O	N	V	E	F
E	L	N	E	-	O	E	E	S	Z	R	Ü	S	R	L	T
A	O	-	S	T	N	O	T	K	É	Z	I	E	I	Z	T
Z	E	B	E	S	E	-	M	E	R	K	-	Z	R	A	-
I	-	I	G	I	A	E	K	H	A	Y	N	-	Z	T	K
L	A	-	S	K	T	-	N	R	E	Z	E	E	-	O	Y
E	Y	K	E	T	N	A	E	-	N	M	-	U	N	M	L
-	É	E	E	E	-	E	L	T	L	N	N	K	P	K	V
D	U	-	Y	L	-	É	L	E	G	E	N	Ö	S	L	?
M	E	Y	E	N	M	Z	G	E	K	T	-	E	-		J
N	K	E	M	-	T	I	-	G	Z		S	U		R	A
J	L	E	T	G	-	N	L	H	E	-	E	Y	V	-	O
E	L	N	Z	S	R	E	E	-	S	M	O	B	L	T	?
-	T	A	S	-	I	A	L	G	O	A	N	M	N	N	?
N	T	Z	E	B	L	D	I	-	E	T	U	Ö	-	E	-
Y	M	G	Y	I	-	E	E	E	G	E	Z	K	A	G	-
Y	K	Ö	Á	O	T	E	B	L	-	L	L	-	G	E	?
R	S	-	-	A	-	E	Z	T	H	A	V	L	M	S	?
I	E	-	Y	A	O	A	K	E	G	-	V	L	D	K	N
V	L	T	N	Z	M	I	E	L	E	-	S	J	-	É	?
Y	K	Ö	Á	O	T	E	B	L	-	L	L	-		G	E

5.4. táblázat: A rejtjeles betűpárokban az elsőt *k*-adik lépésre követő rejtjeles betűk

Ugyanúgy, ahogyan azt az egyszerű helyettesítés esetén tettük, nyelvstatisztikákat hívhatunk segítségül. Most betűpárok gyakoriság-eloszlását kell meghatározni. (Szokásos betűpárok helyett bigramokról beszélni.) Ez a következő feladat megoldását várja el: A feladat megfogalmazása természetesen merül fel azokban az osztályokban, ahol már az egyszerű helyettesítés kapcsán meggyőződtek a betűstatisztikák alkalmazásának hasznosságáról. Különösen természetes azon tanulók számára, akik az egyszerű helyettesítéses titkosírás megfejtése során is "érveltek" gyakoribb betűpárok előfordulása mellett (pl. *szóköz-a* előfordulása valószínűbb az *e-a* betűpár előfordulásánál).

A feladat megoldása során a feldolgozandó nyílt szöveget a felhasználásnak megfelelően kissé átalakítjuk. Mivel lényegtelen, hogy az előforduló betűk kisbetűk, vagy nagybetűk, ezért minden betűt nagybetűvel írtunk. A "sorvége-új sor" karakter mindig egy szóközt is jelent, ennek megfelelően ezt a karaktert mindig egy szóközzel helyettesítettük. Valamennyi egyéb karaktert a gyakoriságok megállapítása előtt minden további nyom nélkül eltüntettük. A 5.5. fejezetben használt valamennyi gyakoriság ilyen táblázatból származik.

Meglepő tapasztalni, hogy a statisztikai feldolgozással nyert betűpár gyakorisági táblázat közvetlen felhasználása helyett a tanulók csak azt kívánják hasznosítani, hogy az egymást követő betűpárok között "melyikek szeretik egymást", és melyikek nem nagyon szeretik. Ezt a véleményt a statisztika nyelvére lefordítva az alkalmazni kívánt statisztika egy adott betűt az írott szövegben közvetlenül követő betűk gyakorisági sorrendjének megállapítását kéri. Ilyen statisztika a bigram statisztika birtokában azonnal nyerhető. A 5.5. táblázat ennek megfelelően előfordulásuk gyakorisága szerint listázza az egyes betűket közvetlenül követő betűket.

Ez a lista egy 70 000 betűs Internet szöveg alapján készült betűpár statisztikából származik.

***Megjegyzés**:* Ha nem az egymást követő betűpárokat keresnénk, hanem az egymástól pontosan *k*-adikként követő betűpárokat szeretnénk megtalálni, akkor a feladat nagyon hasonló lenne. Most a kiindulást a forrásszöveg egymástól *k* távolságra levő betűk párjaiból készített betűpár statisztika jelentené. Ennek megfelelő lehet a tanulók elé állítandó frekvencia leszámoló program:

A	-LZNTKRDGIPJMSBHVACEUFOÉÖÁXYÜWQ
Á	RLSNTGKBMZDJ-CVIUAÉHFPÁEOÖYÜXWQ
B	EAB-OIRUÖÁÉLÜSFNKHPCTZMGDYVJXWQ
C	SIH-ÉRÁAOBENKDFTLZMGÖYVUJPCÜXWQ
D	EOAIÁÉ-ÖUSNTJDRVLBHKÜMZYFCGPXWQ
E	LTNRG-KMSZJIDAHPBUÁFVÉEOCÖXÜYWQ
É	SNRGLTKPBZMV-DJIHAÉEFÁCOÖYUÜXWQ
F	EOIÖÉAUÜRHÁJ-LSTNKZMGDYBVFPCXWQ
G	Y-EAÉÜÁITNLSJGÖZUHKRFÜMBVDCPXWQ
H	AOEÁ-IÖTÉULÜRKNSHGZBJMPCVDYFXWQ
I,í	-STKNAZLGRÁMDEVBFHCOPÉIUÜJÖYXWQ
J	AEOÁLTDU-ÉZÖÜSIRBNKFMGYVJHPCXWQ
K	-OAEIÖÉUÁKÜTRBMNSLHPVDJFCZGYXWQ
L	E-ATILOYÁMÖKÉSDNJGVHURFCÜBPZXWQ
M	EAI-ÁUÉBOMTÜLNZÖSPCRKFHYGVJDXWQ
N	-AÁETÉÖDIUGJNSOCKBÜYHZVLMFPRWXQ
O,Ó	LRNSKMTZG-DPBVCFIÁÉHJAUOEYÖÜWXQ
Ö,Ő	-RLSZTKDNVBHGJEIMAOPÁÖUFÉÜYCXWQ
P	OIERÁAJLÉ-VCTKUPGZÖSNÜBHMDYFXWQ
Q	U-EATLSNIOKRZÁMÉGDÖYBVJHFPCÜXWQ
R	EA-TOIÁSÖMÉXDRÜZKNBJCVHULGFPYWQ
S	Z-AETÉÁOSIÖMBUNKÜRLHVGPCDFYJXWQ
T	-EATÁOIÖÉUJSHRLKVBNÜMYCFPGDZXWQ
U,Ú	SLTNDRGKJ-MHCEAZPBOÁVIÉÖYUFÜXWQ
Ü,Ű	LKN-ZTGSRVEFÁMAIOÉDÖYBUJHPCÜXWQ
V	AEÁIÉOÖHÜTU-LNRVSKZMGDYBJFPCXWQ
W	E-ATLSNIOKRZÁMÉGDÖYBVUJHFPCÜXWQ
X	-IÉTPEALSNOKRZÁMGDÖYBVUJHFCÜXWQ
Y	-EOAIÁSÉTÜZÖBMVUHRKFNJLGDYPCXWQ
Z	-EOTAIÁMÉSÖDZÜUNKVLBGHRJPCYFXWQ
Köz-	AMEKSTFVÉHINRLPBJUÁOCDÖGZÜYWQX-

5.5. táblázat: A rákövetkező betűk gyakorisági sorrendje

5.2. Feladat: Ölelkező bigram frekvencia helyett készítsük el az egymástól *k* betűnyi távolságra levő betűk betűpár statisztikáját. Ez a korábbi programok minimális megváltoztatását követeli. (Akár paraméterként is beépíthető *k* értéke a programba.)

Figyelésre méltó, hogy nagy *k* mellett a relatív gyakoriság-táblázatok sorai alig térnek el egymástól. Erre a célra már $k>5$ is elegendően nagy lehet. Az“alig térnek el egymástól” kijelentés a valószínűség-eloszlások eltérésére szolgáló mértékek alkalmazását követeli meg. Ezzel később kívánunk foglalkozni.

A 5.5. táblázat alkalmazásakor minden betűpár mellett megállapítható, hogy a második betű hányadikként követi az első betűt. Ezt a 28-28 betűpárra összegezve várhatjuk, hogy a legkisebb összeg felel meg az egymást követő betűpároknak.

Ehhez ki kell tehát számítani a "rákövetkezés-összegeket". Erre a célra jobban megfelel a 5.5. táblázat helyett a betűk sorszámait közvetlenül megadó 5.6.*a* és 5.6.*b*. táblázat.

	—	A	Á	B	C	D	E	É	F	G	H	I	J	K	L
—	31	1	19	16	21	22	3	9	7	24	10	11	17	4	14
A	1	18	25	15	19	8	20	24	21	9	16	10	12	6	2
Á	13	18	23	8	14	11	24	19	21	6	20	16	12	7	2
B	4	2	10	3	20	25	1	11	15	24	18	6	28	17	12
C	4	8	7	10	27	14	11	5	15	20	3	2	25	13	17
D	7	3	5	18		14	1	6	25	27	19	4	13	20	17
E	6	14	19	17	25	13	22	21	20	5	15	12	11	7	1
É	13	18	22	9	23	14	20	19	21	4	17	16	15	7	5
F	13	6	11	24	28	22	1	5	26	21	10	3	12	18	14
G	2	4	7	24	27	26	3	5	21	14	18	8	13	19	11
H	5	1	4	20	24	26	3	9	28	18	17	6	21	14	11
I,í	1	6	11	16	19	13	14	22	17	9	18	23	26	4	8
J	9	1	4	17	28	7	2	10	20	22	26	15	25	19	5
K	1	3	9	14	25	22	4	7	24	27	19	5	23	10	18
L	2	3	9	26	24	15	1	13	23	18	20	5	17	12	6
M	4	2	5	8	19	28	1	7	22	25	23	3	27	21	13
N	1	2	3	18	16	8	4	6	26	11	21	9	12	17	24
O,Ó	10	22	18	13	15	11	25	16	16	9	20	17	21	5	1
Ö,Ő	1	18	21	11		8	15	25	24	13	12	16	14	7	3
P	10	6	5	23	12	26	3	9	28	17	24	2	7	14	8
Q	2	4	14	21	27	18	3	16	25	17	24	9	23	11	6
R	3	2	7	19	21	13	1	11	27	26	23	6	20	17	25
S	2	3	7	13	24	25	4	6	26	22	20	10	28	16	19
T	1	3	5	18	23	27	2	9	24	26	13	7	11	16	15
U,Ú	10	15	20	18	13	5	14	23	27	7	12	22	9	8	2
Ü,Ű	4	15	13	22	27	19	11	18	12	7	25	16	24	2	1
V	12	1	3	24	28	22	2	5	26	21	8	4	25	18	13
W	2	3	13	20	27	17	1	15	25	16	24	8	23	10	5
X	1	7	15	21	27	18	6	3	26	17	25	2	24	12	7
Y	1	4	6	13	28	25	2	8	20	24	17	5	22	19	23
Z	1	5	7	20	26	12	2	9	28	21	22	6	24	17	19

5.6.a.táblázat: A rákövetkező betűk sorszáma: 1.rész

	M	N	Oó	Öő	P	Q	R	S	T	U	Ü,ű	V	W	X	Y	Z
—	2	12	20	23	15	29	13	5	6	18	20	8	28	30	27	25
A	12	4	22	24	11	31	7	13	5	20	29	17	30	25	26	3
Á	9	4	25	26	22	31	1	3	5	17	28	15	30	29	27	10
B	23	16	5	9	19	31	7	14	21	8	13	27	30	28	26	22
C	19	12	9	21	26	31	6	1	16	24	28	23	30	29	22	18
D	22	11	2	8	28	31	15	10	12	9	21	16	30	29	24	23
E	8	3	24	26	16	31	4	9	2	18	28	21	30	28	29	10
É	11	2	24	25	8	31	3	1	6	27	28	12	30	29	26	10
F	20	17	2	4	27	31	9	15	16	7	8	25	30	29	23	19
G	23	10	6	15	28	31	20	12	9	17	22	25	30	29	1	16
H	22	15	2	7	23	31	13	16	8	10	12	25	30	29	27	19
I,í	12	5	20	27	21	31	10	2	3	24	25	15	30	29	28	7
J	21	18	3	12	27	31	16	14	6	8	13	24	30	29	23	11
K	15	16	2	6	20	31	13	17	12	8	11	21	30	29	28	26
L	10	16	6	11	27	31	22	14	4	21	26	19	30	24	7	27
M	10	14	9	16	18	31	20	17	11	6	12	26	29	30	24	15
N	25	13	15	7	27	31	4	14	5	10	19	23	29	30	20	22
Oó	6	3	24	27	12	31	2	4	7	23	28	14	29	30	26	8
Öő	17	9	19	22	20	31	2	4	6	23	26	10	30	29	27	5
P	25	21	1	19	16	31	4	20	13	15	22	11	30	19	27	18
Q	15	8	10	09	26	31	12	7	5	1	28	22	30	29	21	13
R	10	18	5	9	28	31	14	8	4	24	15	22	30	12	29	16
S	12	15	8	11	23	31	18	9	5	14	17	21	30	29	27	1
T	20	19	6	8	25	31	14	12	4	10	20	17	30	29	22	28
U,Ú	11	4	19	24	17	31	6	1	3	26	28	21	30	29	25	16
Ü,Ű	14	3	17	20	26	31	9	8	6	23	28	10	30	29	21	5
V	20	14	6	7	27	31	15	17	10	11	9	16	30	29	23	19
W	14	7	9	18	26	31	11		64	22	28	21	30	29	19	12
X	16	10	11	19	5	31	13	9	4	23	28	22	30	29	20	14
Y	14	21	3	12	27	31	18	7	9	16	10	15	30	29	26	11
Z	8	16	3	11	25	31	23	10	4	15	14	18	30	29	27	13

5.6.b.táblázat: A rákövetkező betűk sorszáma: 2.rész

Határozzuk meg minden rejtjeles blokk esetén a sorszámokat, és adjuk össze a 28 blokknak megfelelő sorszámokat!

Kiindulásként az egyes betűpárokhoz tartozó számokat határozzuk meg. Ezt illusztráljuk pusztán, amikor négy betűpárra meghatározzuk a számértékeket.

- *a*: a "-" sorban az "a" betű az első helyen áll, tehát a sorszám = 1,
t é: a "t" sorban az "é" betű a 9. helyen áll, tehát a sorszám = 9,
- *k*: a "köz" sorban a "k" betű a 7. helyen áll, tehát a sorszám = 7,
m g: az "m" sorban a "g" betű a 25. helyen áll, tehát a sorszám = 25.

Az 5.5. táblázat alapján folytathatjuk a betűpárokhoz tartozó számok meghatározását. Összegezzük ezeket a számokat a 28 betűpárra!

Az így adódó eredményeket foglalja össze az 5.7. táblázat.

A 5.7. táblázat segít a valódi sorrend kialakításában. Látható, hogy a rejtjeles blokkok első betűihez tartozó összegek közül a 16. oszlop a legkisebb (166), ezért feltételezzük, hogy a nyílt blokkokban ez az -oszlop követi a rejtjeles első oszlopot. A 16. betűkhöz tartozó összegek közül a 2. oszlop a legkisebb (191), ezért feltételezzük, hogy a nyílt blokkokban ez az oszlop követi a rejtjeles 16. oszlopot.

Tovább haladva, a 2. oszlop esetén a legkisebb a 8. oszlop összege (160), a nyolcadik oszlop esetén a 3.oszlop, és ezt folytatva a 3., 9., 10., 4., 11., 12., 6., 14., 6. oszlopsorrend adódik. A legkisebb összegeket a 4.6. táblázatban könnyebb áttekinthetőség miatt kiemeltük, így azok részletezése nem látszik szükségesnek.

A legkisebb összegeket sorra kigyűjtve azt feltételezhetjük, hogy a nyílt szövegben a rejtjeles betűk a

7, 15, 1, 16, 2, 8, 3, 9, 10, 4, 11, 12, 5, 13, 6, 14

sorrendben követik egymás. Következtetésünket persze csak akkor fogadhatjuk el, ha ez a sorrend visszapermutálás után értelmes szöveghez vezet. Ennek eldöntéséhez lássuk az első blokkot!

0	262	230	348	323	289	314	372	343	335	302	321	349	290	221	***166***
302	0	346	358	291	270	340	***160***	345	321	308	346	356	380	327	252
270	322	0	309	349	351	318	192	***168***	344	323	316	333	344	320	401
339	340	279	0	346	228	334	321	328	254	***120***	339	385	345	317	261
366	368	293	303	0	274	376	243	295	337	318	221	***208***	368	254	277
378	311	370	297	305	0	330	321	314	320	346	331	260	***188***	331	366
377	349	289	322	321	244	0	337	320	307	304	269	369	336	***156***	367
395	209	***100***	336	278	242	323	0	371	293	362	284	356	363	256	212
322	332	233	349	286	289	272	359	0	***218***	401	295	418	388	271	298
297	318	320	***108***	220	225	303	250	267	0	358	322	413	365	311	323
328	270	322	249	317	273	349	357	377	305	0	***218***	400	335	292	362
249	297	297	310	***202***	208	302	325	278	345	206	0	407	391	333	312
328	344	252	372	237	***106***	352	311	310	349	293	343	0	375	314	248
328	298	330	319	307	***199***	386	342	324	308	360	339	373	0	327	314
139	400	300	311	296	260	240	224	302	346	351	343	396	395	0	286
314	***191***	387	281	286	347	351	247	286	346	350	364	392	358	278	0

5.7. táblázat: A 5.2. példához tartozó sorrend-összegek

-aan-ahlkiekppav --> ha-valakinek-pap

Ez egy értelmes szöveg része, tehát folytathatjuk a visszaállítást. Eredményül visszakapjuk a 5.2. Példa nyílt szövegét, tehát megfejtésünk sikeres volt.

5.6. Elméleti rész

Jelen részben a "rákövetkezés sorszámok" helyett magukat az előfordulási valószínűségeket használjuk fel a döntések meghozatalához. A fejezet végén az 5.8. táblázat közel 480 ezer karakterből készített betűpár gyakorisági táblázatot ad meg. Általában a két-három legjobb tanuló önmaga kezdeményez ilyen megoldást. A „találd ki a rákövetkezőt" játék kapcsán már elkészített betűpár gyakorisági táblázat különböző felhasználási lehetőségeire tesznek javaslatot. Az alábbiakban vázolt megoldás ugyan nem szerepelt ezek között, de ez is összerakható több osztályban elhangzottakból.

A korábbi megfogalmazást kissé megismételve egy kissé eltérő jelöléseket vezetünk be.

Felidézzük, hogy a rejtjelezendő nyílt szöveget adott hosszúságú blokkokra osztjuk fel. Legyen B a blokkhossz, és legyen a blokkok száma N. Jelölje $M(i)$ az i-edik blokkot, továbbá jelölje $M(i,j)$ az i-edik blokk j-edik karakterét, $i=1,2,\ldots,N$ és $j=1,2,\ldots,B$.

Legyen P a helycseréhez alkalmazott permutáció inverze (az eredeti permutáció helyett; - a jelölések egyszerűsítése miatt használjuk ezt a jelölést az inverzre).

Jelölje $C(i)$ az i-edik kódolt (rejtjeles) blokkot, és legyen ennek j-edik betűje $C(i, j)$.

A megoldandó problémát a hipotézisvizsgálatok elméletében tipikus feladattá alakítva a bevezetett jelölések segítségével fogalmazzuk meg:

***D(K)* Döntésfeladat:** $K=1,2,\ldots,B$.

- *Vagy a rejtjeles (kódolt) blokkok **k**-adik betűjét ugyanennek a blokknak a **j**-edik betűje követi az eredeti nyílt blokkban* $\boldsymbol{j}=1,2...,B,$ $\boldsymbol{j\neq k}$.
- *Vagy a **k**-adik betű az eredeti nyílt blokk utolsó betűje*

A tanulók által javasolt döntési eljáráshoz hasonló az az eljárás is, amely a "lehető legjobb" döntési eljárás:

1.lépés: A nyílt szöveg N-hosszú forrásszövege alapján készítsünk $F(i,j)$ betűpár gyakorisági táblázatot! (Egy ilyen táblázat már bevetésre készen állhat, például a tanár egy korábbi években készített táblázatot ajánlhat, vagy az irodalomban is találhatunk ilyet.)

Technikiai megjegyzés: Esetünkben a forrásszöveg az írott magyar nyelven készített szöveg. Ebben néhány karakter érdeklődésünkön kívül esik

(például számok, egyes írásjelek). Ezeket egy speciális karakterrel felülírjuk (például a "+" jellel). Az első feladat előkészítéséhez hasonlóan minden kisbetűt nagybetűs párjával, a „Enter" jelet szóközzel, a hosszú magánhangzókat rövid párjukkal helyettesítjük. Így készült a fejezet végén levő 5.8. táblázat.

***Megjegyzés**:* az **együttes eloszlás** bevezetése triviálissá válik a betűpárokban szereplő betűkre való hivatkozással. Ugyancsak természetessé válik a **"feltételes gyakoriság", "feltételes valószínűség"** kifejezésekre vonatkozó szóhasználat, ha a betűpárok első betűjét követő betűről beszélünk.

Ugyanúgy, mint a "rákövetkezés-szám" felhasználása esetén, most magukat a bigram valószínűségeket fogjuk felhasználni. A rejtjeles blokkok első, második, stb. betűihez tartozó párok együttes valószínűségei közül a legnagyobb alapján döntünk arról, hogy melyik a nyílt szövegben rákövetkező betű. Az egyes betűpárok $B>10$ blokkhossz esetén eléggé távol vannak egymástól a nyílt szövegben ahhoz, hogy együttes valószínűségeket az egyes valószínűségek szorzatával közelítsük meg. Ez a szorzat akkor lesz legnagyobb, ha a sokkal könnyebben számítható logaritmusuk összege a legnagyobb. Ezért előállítjuk a logaritmikus valószínűségek táblázatát.

2.a.lépés: Készítsük el a betűpár gyakorisági tábla alapján készített együttes valószínűség-eloszlások logaritmusait tartalmazó táblázatot.

2.b.lépés: Határozzuk meg ezek összegeinek a maximumát minden rögzített K esetén.

***Megjegyzés*:** Ezt a táblázatot szokás "log-likelihood" táblázatnak nevezni. A gyakorisági táblázatból való elkészítése problémamentes. A logaritmusokat nem lehet kiszámítani akkor, ha a megfelelő gyakoriság nulla. Ezért erre már a gyakorisági tábla előállításakor figyelemmel kell lenni.

Ennek megfelelően erre a jelenségre fel kell készülni. Igazán akkor van probléma, ha a rejtjelzett szövegben a nulla gyakoriságú helyeknek megfelelő betűpárok között van nulla gyakoriságú betűpár. Ezt a nehézséget el lehet kerülni, ha valamennyi gyakorisághoz 1-et hozzáadunk. Ha a gyakorisági táblázatot nagyon hosszú szöveg alapján készítjük el, tehát ha N nagyon nagy, akkor reménykedhetünk abban, hogy ez a jelenség nem fordul elő. Ez a helyzet jelenlegi példánkban is.

A döntés Bayes-hibája exponenciálisan tart nullához valamennyi k, $D(k)$ esetén, ha N tart végtelenhez.

Gyakorlatilag egyszerűbb a második lépésben szereplő összeg helyett annak ellentettjével számolni, ami pozitív. Ekkor viszont maximum helyett a minimumot kell keresnünk.

3.lépés: Definiálunk egy irányított gráfot, amelynek a pontjai a blokkok pozíciói, és a *k*-adik és *j*-edik pozíció akkor van irányított éllel összekötve, ha a *j*-edik pozíció adja az adott *k* mellett a likelihood függvény maximumát.

4.lépés: Határozzunk meg egy maximális hosszúságú Hamilton vonalat ebben a gráfban, és ezt használjuk az eredeti szöveget visszaállító *P* permutáció meghatározására. Ha ilyen Hamilton vonal nem létezik, akkor önkényesen járhatunk el.

	A	Á	B	C	D	E	É	F	G	H	I	J	K	L	M	N
A	39	19	754	204	2576	56	17	106	3310	469	843	1375	6116	6611	2377	6152
Á	14	4	1226	183	392	4	34	19	1478	151	232	609	1080	3261	829	3374
B	2691	714	1434	0	7	3604	260	2	0	12	777	26	17	214	4	25
C	213	309	44	72	0	187	90	4	0	653	306	5	103	13	3	15
D	1393	605	95	18	251	3218	645	23	4	134	1588	665	64	242	78	593
E	49	304	958	229	2200	82	24	156	8598	900	673	638	5072	11041	6205	8651
É	0	0	656	66	426	3	13	47	2674	137	27	340	1077	2096	262	1970
F	407	217	3	0	0	2493	645	26	1	2	645	126	0	17	0	4
G	1819	1066	317	62	40	1945	1003	138	370	349	734	352	170	306	139	408
H	4599	2069	2	0	1	1565	213	2	7	38	938	14	0	8	13	14
I	690	449	253	872	1023	351	97	252	2083	207	28	134	2787	1127	398	4496
J	1705	1607	30	26	257	1324	454	38	12	33	49	170	82	26	3	50
K	3018	860	403	20	6	4663	1850	26	4	107	3874	73	734	156	23	415
L	3575	2837	205	166	872	6747	1598	418	597	459	966	717	832	2960	856	1332
M	2861	2186	894	17	10	6387	1541	32	4	34	5167	28	66	278	319	102
N	4288	941	133	699	2993	6740	1204	144	1030	194	3274	154	1603	81	28	2054
O	51	9	718	259	1311	5	17	92	3935	415	100	113	2934	6375	1956	4887
Ö	24	7	398	65	552	18	5	51	287	87	94	129	932	2318	299	2029
P	1142	598	17	70	5	1190	249	16	2	50	586	142	37	199	5	68
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	3628	1494	251	728	1446	4293	625	177	457	237	1843	396	479	145	629	745
S	2611	1260	218	74	85	2751	1184	35	38	81	1561	0	383	59	428	338
T	5871	3109	203	150	5	7042	1914	55	13	561	1272	543	308	469	54	803
U	16	3	84	77	921	12	1	23	906	126	55	405	525	1973	146	639
Ü	0	0	25	2	115	16	2	3	112	50	5	32	425	2189	8	558
V	4459	1609	20	3	1	3934	1373	1	0	6	1558	24	5	7	2	19
W	18	0	0	0	0	11	0	0	0	1	11	0	1	0	0	0
X	0	0	0	12	0	2	1	0	0	0	7	0	0	0	0	1
Y	1379	636	193	4	25	2357	312	30	60	94	862	219	103	32	257	281
Z	2047	1726	154	12	396	4662	1696	45	52	93	1263	1	301	222	126	560
+	22779	1255	3780	2155	1681	10492	3750	5600	1487	9043	4158	1954	8996	4506	13887	6689

5.8.a. táblázat: 479 234 betűs magyar betűpár gyakoriságok 1.rész

	O	Ö	P	Q	R	S	T	U	Ü	V	W	X	Y	Z	köz
A	6	6	1688	6	3681	2316	4517	116	14	392	1	1	190	7265	20163
Á	5	1	86	0	4465	2619	3585	18	3	453	0	0	0	790	978
B	1388	412	2	0	187	31	31	632	159	11	0	0	0	7	821
C	200	10	1	0	34	3651	22	62	2	19	0	0	0	86	141
D	1689	1262	14	0	178	187	1074	431	189	484	0	0	20	30	2423
E	134	8	343	0	5110	4300	7894	75	35	896	2	39	4	3305	8228
É	0	0	1000	0	2779	3337	1923	0	4	354	0	0	0	519	1107
F	1637	722	0	0	126	4	10	263	191	1	0	0	0	0	21
G	1556	444	49	0	327	330	607	324	63	312	0	0	10072	141	4078
H	4249	307	0	0	7	20	10	408	51	98	2	0	8	0	81
I	189	14	200	0	1235	3645	3794	182	20	875	1	2	1	940	7151
J	1021	500	0	0	67	78	707	510	120	4	0	0	0	22	489
K	3059	1772	4	0	377	258	277	324	602	91	0	0	3	12	12221
L	2297	1612	48	0	258	521	4606	283	521	940	0	0	2706	31	8307
M	1992	127	188	0	135	125	91	485	116	26	0	0	0	115	6007
N	622	726	7	1	70	415	2882	270	211	147	5	0	4364	126	11856
O	0	1	224	1	3962	2821	3608	20	2	637	1	2	1	1633	1945
Ö	0	11	85	0	1641	878	1848	21	21	499	0	1	0	1213	1887
P	835	67	291	0	211	52	177	255	62	34	0	0	1	36	609
Q	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0
R	2443	828	53	0	697	766	2941	552	553	361	1	1	100	209	3857
S	1817	594	77	1	179	2168	2369	375	280	51	0	0	26	12441	8279
T	3741	1835	21	0	664	708	7817	1185	804	687	1	0	303	41	22923
U	6	0	53	0	1083	819	1838	0	1	45	0	3	0	210	451
Ü	0	0	6	0	193	202	386	0	2	143	0	1	0	340	343
V	1988	243	0	0	17	10	31	34	96	11	0	0	6	0	98
W	0	0	0	0	0	0	0	0	0	0	0	0	0	?	?
X	3	0	8	0	0	0	8	1	0	1	0	1	1	0	8
Y	1293	367	38	0	157	525	455	453	421	185	0	0	0	79	6991
Z	3276	858	23	0	198	400	3728	565	373	298	0	0	0	642	7321
+	2589	2673	2497	13	2897	8577	5866	2555	577	7500	18	3	2	805	0

5.8.b. táblázat: 479 234 betűs magyar betűpár gyakoriságok 2.rész

Ahogyan azt fentebb már említettük, ha a betűpárok közti távolság elég nagy, akkor a feltételes valószínűségek nagyon közel kerülnek az egydimenziós gyakorisági eloszláshoz. (Esetünkben $k>10$ már "elég nagy".) Ez azt is jelenti, hogy ha j nem a valódi hipotézisnek felel meg, akkor a megfelelő várható érték nagyobb, mint a "valódi" hipotézisnek megfelelő várható érték.

Statisztikai vizsgálataink azt mutatják, hogy $B=20$ esetén $N=50$ blokknyi vizsgált (nyílt) szöveg elegendően hosszú; a neki megfelelően elérhető hibavalószínűség 1% alatt marad.

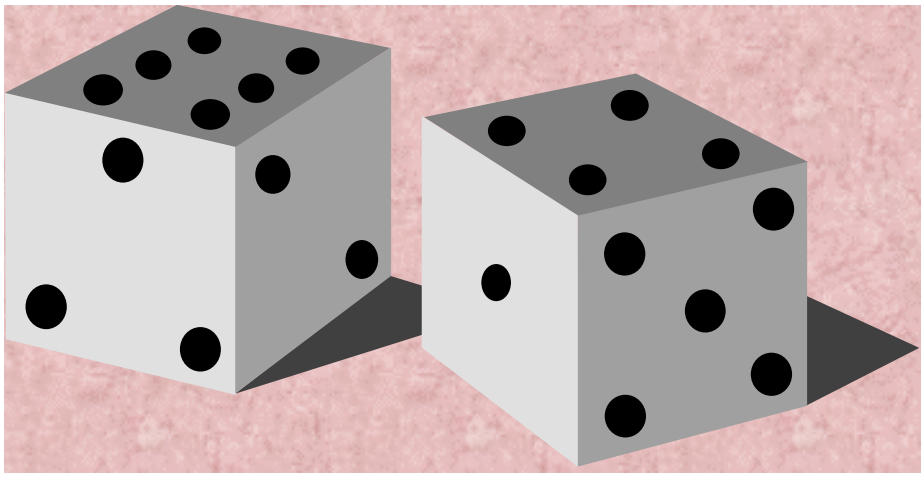
Az elmúlt évek gyakorlatában 50.000 vagy hosszabb nyílt szöveg alapján készítettük el a gyakorisági eloszlásokat. Az 5.8. táblázat azonban 480 ezer hosszúságú szövegből készült.

Nyelvstatisztikák igénye és elkészítésük egy évszázados múltra tekint vissza, lásd [K98].

Célszerű elkészíteni a forrásszöveg k-lépéses bigram statisztikáját. Itt a "k-lépéses" arra utal, hogy a bigram első és második betűje között a nyílt szöveg $k - 1$ betűje helyezkedik el. Ha k elég nagy - és ebben az esetben $k > 5$ már elég nagy-, akkor az eloszlások statisztikailag nem különböztethetők meg egymástól.

A felvázolt programon módszeresen végighaladva a korábbiakban nyert megoldáshoz jutunk. Ennek részletezése semmi újat nem tartalmaz, így arra nem térünk ki.

VI. Első lépések a valószínűség számszerű jellemzéséhez



6.1. Lehet-e mérni a véletlent?

A címben szereplő kérdés ismét és ismét provokál tantermi vitákat, amelyek szinte mindig azzal zárulnak, hogy véletlen események, kimenetelek bekövetkezésének esélyét nemcsak lehet mérni, de egy erre a célra szolgáló mérőszámra feltétlenül szükség is van. Általános az a vélemény is, hogy egy elméleti mérőszám megválasztásához az is elengedhetetlen, hogy az elméleti modelleket a gyakorlati tapasztatok, a konkrét kísérletek végrehajtása során szerzett megfigyelések figyelembevételével alakítsuk ki.

A modell kiépítését a korábban szerzett tapasztalatokra támaszkodva, néhány véletlen számpélda segítségével mutatjuk be. A követett utat több éves iskolai tapasztalatunkra támaszkodva alakítottuk ki. A következő 6.1. Példa és annak részletes ismertetése iskolai szinten követhető elemzést mutat be.

6.2. Klasszikus valószínűség hozzárendelés

6.1. Példa: Három kocka - legkisebb dobás
A játékvezető feldob három kockát, kimenetelüket a tanulók előtt titokban tartja. A tanulóknak rákérdezéssel (találgatással) ki kell találni, hogy melyik volt a legkisebb dobott szám. A tanulók célja, hogy a véletlentől függő valódi kimenetelt minél előbb kitalálják.

Milyen stratégia szerint kell találgatni? Ahogy azt már korábban megállapítottuk, célszerű először arra a kimenetelre rákérdezni, amelyik a legnagyobb eséllyel következik be. Ha ez volt a valódi, megfigyelt kimenetel, akkor a játék befejeződött. Ha nem, akkor a további lehetőségek közül kell választani, mégpedig megint azt, amelyiknek legnagyobb az esélye. Ezt a stratégiát kell folytatni a valódi kimenetel kitalálásáig.

Ahhoz azonban, hogy ezt a stratégiát követni lehessen, meg kell állapítani, hogy a lehetséges kimenetelek esélyei hogyan viszonyulnak egymáshoz. Ennek követhető iskolai tárgyalását fogjuk most megvizsgálni. A módszeres tárgyaláshoz a kísérletre vonatkozó megfigyelésekre kell szert tennünk. Az oktatási idő korlátossága azonban nem engedi meg, hogy egy teljes órát áldozzunk 3 kocka dobálására. Helyette most itt megadunk egy kísérleti adatokat tartalmazó listát. Számítógépes szimulációra minden osztályból többen is képesek, és szívesen el is végzik, célszerű ezzel a lehetőséggel élni.

2	1	2	4	3	1	2	4	2	5	3	2	3	2	4	1	2	2	1	2
1	2	3	1	1	4	1	2	1	2	3	4	4	2	3	1	1	1	2	1
1	1	1	3	4	4	2	1	2	1	1	5	4	1	3	5	3	2	2	1
1	4	1	4	1	1	1	1	2	2	3	1	2	1	1	1	1	1	2	3
4	2	1	3	2	2	2	1	2	3	3	2	3	1	3	3	2	1	4	2
1	1	1	1	1	1	2	1	1	2	5	3	2	5	3	3	1	1	6	2
4	2	1	2	1	1	3	5	1	3	1	3	6	3	5	1	1	1	2	2
1	1	4	1	1	1	1	2	2	1	1	4	1	2	5	1	2	1	2	2
4	3	1	1	1	3	1	2	1	1	2	4	2	1	1	4	2	2	5	1
1	1	1	1	1	3	3	4	1	3	2	4	3	4	2	3	5	1	3	1
1	5	1	2	1	1	2	1	2	3	3	3	3	1	1	1	1	2	2	2
3	4	1	1	2	4	2	4	2	1	3	3	1	3	1	1	2	1	4	2
1	1	3	2	1	2	1	4	3	3	2	1	3	1	2	2	2	3	1	2
5	4	2	3	3	1	1	1	2	1	1	3	1	2	1	2	2	1	1	4
1	2	3	5	4	1	1	1	1	1	3	5	2	1	5	2	1	1	1	2
1	2	1	2	3	2	3	3	2	1	2	1	3	1	1	2	1	1	3	4
1	1	2	4	2	3	1	1	2	6	2	1	4	3	6	3	2	1	3	4
1	1	1	2	1	1	3	1	1	2	1	3	4	1	1	3	3	1	2	2
4	3	4	1	1	3	4	1	2	2	3	1	2	4	1	1	1	2	4	2
1	3	3	3	1	2	1	2	2	1	1	2	2	2	1	3	1	2	2	4

6.1. táblázat: 3 kocka minimuma. 400 megfigyelés listája

Megfigyeléseink lényegét összegzi a következő 6.2.táblázat.

kimenetel	1	2	3	4	5	6
gyakoriság	164	106	70	41	15	4
Relatív gyakoriság	0,4100	0,2650	0,1750	0,1025	0,0375	0,0100

6.2.táblázat : 3 kocka minimuma. Gyakoriságok.

A tanulók számára nyilvánvaló, hogy a leíró elemzésnek a lehetséges kimenetelek összegyűjtésével kell kezdődnie. Esetünkben ezek az 1, 2, 3, 4, 5 és 6 egészek.

Ugyancsak nyilvánvalóan fogadják el (többnyire ismerik már) a tanulók a **klasszikus valószínűségszámítás** alapját:

Az egyes kimenetelek valószínűsége arányos a lehetséges különböző bekövetkezéseik számával.

A lehetséges kimenetek megszámolásánál egy általánosan előforduló probléma, hogy „**megkülönböztethető, vagy nem megkülönböztethető**?" A tanulók számára elfogadható válasz: „*Nem lehet más a lehetőségek száma, ha a kockák különböző színűek, mint akkor, ha a színükről semmit sem tudunk*."

Különböző színű kockákat feltételezve az osztály számára azonnal elfogadott, hogy az összes lehetséges esetek száma $6^3 = 216$.

Minden ilyen lehetőség csak egyféleképen következhet be. Így juthatunk el az **elemi esemény** fogalmához. Hasonlóan a konkrét helyzet engedi meg a **kedvező esemény** fogalmának bevezetését.

Az összes lehetséges esemény közül azokat, amelyek bekövetkezése esetén a legkisebb dobott szám 1, az 1 kimenetelre kedvező eseményeknek nevezik. Hasonlóan beszélhetünk a 2, 3, stb kimenetelekre kedvező eseményekről is.

Nyilvánvaló, hogy egyetlen lehetséges elemi esemény sem lehet kedvező esemény különböző kimenetelre, és az is, hogy minden lehetséges elemi esemény kedvező valamelyik kimenetelre. Így a lehetséges elemi eseményeket most hat olyan osztályba soroltuk, amelyeknek nincs közös eleme. Az egyes kimenetelek egyidejűleg nem következhetnek be, ezért az ilyen eseményeket **egymást kizáró (diszjunkt) eseményeknek** nevezik. Ezek az osztályok jelenítik meg az **összetett eseményeket**.

A következő lépés azon elemi események megszámlálása, amelyek bekövetkezése esetén a legkisebb dobott szám n, mégpedig n=6–tal kezdve. A leszámlálást itt nem részletezzük.

Amikor az egyes kimenetelek esélyeit számokkal akarjuk jellemezni, akkor kiindulhatunk abból, hogy az elemi események között egyik sincs kitüntetve, azok egyenlő eséllyel következhetnek be. A kedvező elemi események száma megmutatja, hogy az adott kimenetel hány ilyen egyenlő esélyű elemi esemény mellett következik be. Ezért *az egyes kimenetelek esélyei úgy aránylanak egymáshoz, mint a rájuk nézve kedvező elemi események számai*. Ez a meggondolás adja **a valószínűségek** tanulók által elfogadható meghatározásának **klasszikus elvét**.

Ha véges sok elemi esemény van, és feltételezhető, hogy azok egyenlően valószínűek, akkor egy tetszőleges esemény valószínűségét a

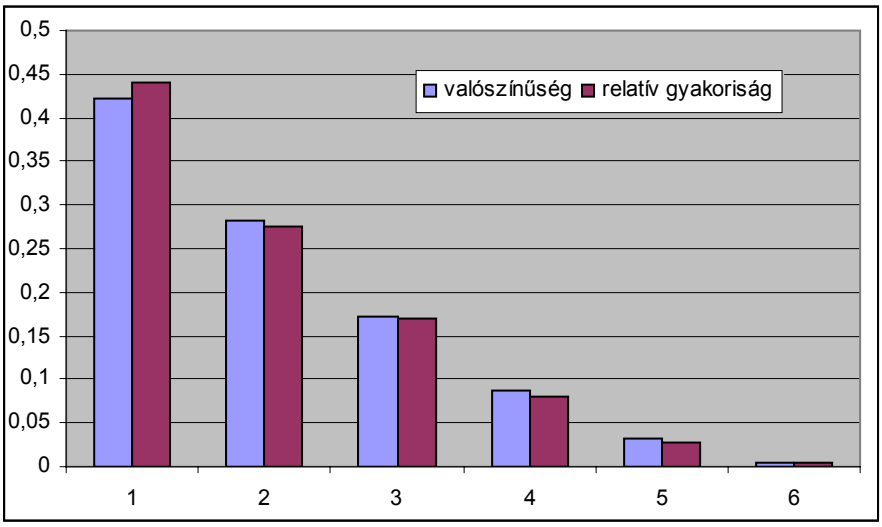
$$\frac{\textbf{a kedvező elemi események száma}}{\textbf{az összes elemi események száma}}$$

tört adja meg.

A kapott eredményeket a 6.3. táblázat 2. és 3. sora összegezi. A 3. sorban szereplő arányok adják meg a **valószínűségeket**. A nyert valószínűséget a tanteremben kapott 2400 játék kimeneteleivel szembesítettük (20 tanuló fejenként 120 játék eredményeit összegezte). A gyakoriságokat tartalmazza a 6.3. táblázat negyedik sora.

Legkisebb szám:	1	2	3	4	5	6
Kedvező esetek száma:	91	61	37	19	7	1
Kedvező esetek aránya:	0,4213	0,2824	0,1713	0,0880	0,0324	0,0046
Gyakoriság:	1057	659	410	195	67	12
Relatív gyakoriság:	0,4404	0,2746	0,1708	0,0813	0,0279	0,0050

6.3. táblázat: Három kocka minimuma. Összefoglaló adatok.



6.1. ábra: 3 kocka minimuma a 6.3. táblázat szerint

6.3. A binomiális eloszlás

Példa: 20 kockadobás

Egy játékvezető kiválasztja véletlenszerűen az alábbi két kísérlet egyikét. A kiválasztott kísérletet ezután 20-szor végrehajtja, és a kimeneteleket feljegyzi. A 20 adatból kell arra következtetni, hogy a játékvezető melyik változatot hajtotta végre.

A. változat: A játékvezető feldob egy kockát, és
- 0-t ír le, ha a dobott szám 1, 2 vagy 3;
- 1-et ír, ha a dobott szám 4, 5 vagy 6.

B. változat: A játékvezető feldob egy kockát, és
- 0-t ír le, ha a dobott szám 1 vagy 2;
- 1-et ír, ha a dobott szám 3, 4, 5 vagy 6.

Legyen például a kapott sorozat:
1,0,1,1,0, 0,1,1,0,1, 1,1,1,0,1, 1,1,1,0,0.

Ebben az esetben a heurisztika helyes eredményt ad meg. Több osztályban tettük fel a kérdést és kaptuk ugyanazt a választ: „Ez a sorozat a második változattal készült, mert több benne az 1, mint a 0. Az első változat esetén lényegében ugyanannyinak kellene lenni mind a két számból, a második esetén a számoknak körülbelül a harmada lenne 0. (Valóban a második változattal keletkezett az adott sorozat.)

Számunkra maga a gondolatmenet a lényeges. A megadott sorozat gyakoriságait vizsgálták a tanulók. Mivel a kísérleteket egymás után hajtották végre, azoknak egymásra semmilyen hatása nem volt, így az sem lényeges, hogy az egyes kimeneteleket milyen sorrendben kaptuk meg. Arra koncentrálunk, hogy melyik kimenetel hányszor következett be. Ezeket a gyakoriságokat hasonlítjuk össze azokkal a heurisztikusan elképzelt valószínűségekkel, amelyek az egyes változatoknak felelnek meg.

Határozzuk meg mindkét változat szerint annak a valószínűségét, hogy az $n = 20$ dobás pontosan k-szor eredményezett 0-t!

Most már ismerjük a klasszikus valószínűség hozzárendelési szabályát. Ez segít a keresett valószínűség kiszámításában. Feltételezzük a binomiális együttható fogalmának ismeretét. A klasszikus valószínűségszámítási szabály szerint meg kell határozni az összes esetek számát. Ez 6^{20} . A kedvező esetek számát akkor kapjuk meg, ha összeszorozzuk a n helyből kiválasztható k hely összes lehetőségének számát azon lehetőségek számával, ahányféleképpen éppen k darab 0-t és $n-k$ darab 1-est választottunk ki, ami az

A változat szerint $g^k = 3^k$ és $(g')^{20-k} = 3^{20-k}$;
B változat szerint $g^k = 2^k$ és $(g')^{20-k} = 4^{20-k}$.

Ennek megfelelően a keresett valószínűség:

$$P(n = 20 - ból\ \ k - szor\ \ lett\ \ 0) =$$

$$= \frac{\binom{20}{k} \cdot g^k \cdot g'^{20-k}}{6^{20}} = \binom{20}{k} \frac{g^k}{6^k} \cdot \frac{g'^{20-k}}{6^{20-k}} = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$, ha $p=g/6$ és $(1-p)=g'/6$. ($k=0,1,\ldots,n$.)

Ezzel a példa speciális esetére levezettük a binomiális eloszlás képletét.

A levezetett képlet egészen általános helyzetben is érvényes marad. Legyen A egy véletlen kísérlet $p=Prob[A]$ valószínűségű eseménye. Tegyük fel, hogy a kísérletet n -szer végrehajtják, és ennek során az A esemény pontosan $X(n,p)$ alkalommal következik be. A véletlen $X(n,p)$ szám lehetséges értékei a $k = 0,1,...,n$ pozitív egész számok. $X(n,p)$ ezeket az

értékeket rendre a fenti valószínűséggel veszi fel. *X(n,p)* valószínűség-eloszlását binomiális eloszlásnak nevezik. Szokásos jelölése: *B(n,p)*.

Az eloszlás a valószínűségszámítás fontos fogalma.

6.1.Definíció: Nemnegatív számok egy $\{p_1, p_2, \ldots, p_n\}$ összességét valószínűség-eloszlásnak nevezik, ha összegük 1, azaz $\sum_{i=0}^{n} p_i = 1$.

Számítógépen gyorsan kiszámíthatjuk a binomiális eloszlás értékeit. Ehhez az szolgáltatja az ötletet, hogy egy kísérlet (n+1)-szeri megismétlése során ha egy *A* esemény pontosan *k*-szor következik be, akkor az első *n* kísérletben vagy *k*-szor, vagy (*k*-1)-szer kellett előfordulnia. Tanári tapasztalatom szerint érdemes ezt az ötletet közölni, és ezután már maguk a tanulók megadják a következő előállítási algoritmus valamilyen variánsát.

***B(n,p)* binomiális eloszlást előállító rekurziós algoritmus**

1.*lépés* (inicializálás): három tárolót kinullázunk: Legyen
A(i)=B(i)=C(i)=0, ha *i*=0,1,...,100.

2.*lépés* (Megadjuk a két paramétert): *N* az ismétlések száma, *P* a megfigyelt esemény valószínűsége.

3.*lépés* (értékadás):
A(0)=1-P; A(1)=P; B(0)=0; B(1)=1-P; B(2)=P; I=2

4.*lépés* (az *I,P* paraméterű binomiális eloszlás kiszámítása):
*C(k)=(1-P)*A(k)+P*B(k), k=0,1,...,I. I=I+1.*

5.*lépés* (Ellenőrzés: készen vagyunk?)
Ha *I>N*, akkor vége. Különben folytatjuk.

6.*lépés* (Módosítás):
A(k)=C(k); B(k+1)=C(k), k=0,1,...,I. Folytassuk a 4. lépéssel.

A téma lezárásaként megadunk négy gyakorló feladatot. Ezek a tanárok számára szolgálhatnak mintául. Hasonlókat könnyen tudunk előállítani.

BGy1 gyakorló feladat: érmék
Feldobunk egy érmét *n* –szer. Legyen *k* a dobott fejek véletlen száma. Meghatározandó *k* eloszlása.

BGy2 gyakorló feladat: kockák
Feldobunk *d* kockát *n* –szer. Legyen *k* a dobott hatosok véletlen száma. Meghatározandó *k* eloszlása.

BGy3 gyakorló feladat: sorsolás visszatevéssel
Egy dobozva beleteszünk f fehér és s sárga golyót. Ebből a dobozból húzunk visszatevéssel egymás után n –szer. Legyen k a kihúzott fehér golyók véletlen száma. Meghatározandó k eloszlása.

BGy4 gyakorló feladat: kártyák
Egy csomag magyar kártyából alapos keverés után felütünk egy lapot. Ezt a húzást n –szer megismételjük, mindig alapos keverés után a teljes pakliból. Legyen k a kihúzott makk felsők száma. Meghatározandó k eloszlása.

6.4. Geometriai eloszlás

6.3. példa: kockadobás az első hatosig

Egy kockát addig dobunk fel, amíg először 6-ost dobunk. Történjen ez az n-edik dobásra. Ez nyilván függ a véletlentől. Mit mondhatunk erről a véletlen n számról, mekkora a valószínűsége annak, hogy egy konkrét értéket vesz fel. Alkalmazható-e a klasszikus valószínűségszámítási képlet? A választ minimális tanári irányítással megadhatja az alábbi tanulói levezetés.

Nyilvánvaló, hogy

$$P(\text{ az első dobás hatos})] = 1/6,$$

$$P(\text{ az első dobás nem hatos }) = 5/6.$$

Az utóbbi esetek 1/6 részében másodikra hatost dobunk, tehát

$$P(\text{ először a második dobás a hatos }) = \frac{5}{6}\cdot\frac{1}{6},$$

$$P(\text{az első két dobás egyike sem hatos}) = \frac{1}{6}\cdot\frac{1}{6} = \left(\frac{1}{6}\right)^2$$

Ugyancsak nyilvánvaló, hogy

$$P(\text{az első } (n\text{-}1) \text{ dobás egyike sem hatos}) = \left(\frac{5}{6}\right)^{n-1}.$$

Bármely ilyen eseményt követően mindig az esetek 1/6 részében dobunk hatost.

$$P(\text{először pontosan az } n\text{-edik dobás lesz hatos }) = \left(\frac{5}{6}\right)^{n-1}\cdot\frac{1}{6}.$$

Ennek a végtelen geometriai sornak az összege 1. Ez általánosan is igaz lesz.

Általános eset.

Tekintsük a véletlen kísérleteknek azt az általános esetét, amikor egy véletlen kísérletet mindaddig ismételünk, amíg egy p valószínűségű A esemény (először) be nem következik. Jelöljük az összes végrehajtott kísérletek számát $X(p)$-vel. $X(p)$ lehetséges értékei a pozitív egész számok, tehát az eddigiektől eltérően elvben végtelen sok különböző kimenetel fordulhat elő.

Az $X(p)$ véletlen szám eloszlását egy mértani sor adja meg, melynek az első tagja p, és kvociense pedig $(1-p)$:

$P[X=k]$ = Prob[A először a n-edik kísérletnél következik be] = $p(1-p)^{(n-1)}$

6.2.Definíció: Azokat a valószínűség-eloszlásokat, amelyek a fenti alakúak, p paraméterű geometriai eloszlásoknak nevezik.

A mértani sor összegképlete a gimnáziumi emelt szintű matematika tananyag részét képezi. Ezek alapján megállapítható, hogy a geometriai eloszlás tagjainak összege p minden értékére 1, tehát tagjai valóban eloszlást alkotnak. Gyakorlásként még középszinten is meghatározhatók véges részletösszegek.

Az eloszlás tagjai monoton csökkenő sorozatot alkotnak, tehát köztük az első a legnagyobb. Ez a megállapítás ebben a formális formában teljesen nyilvánvaló. Ha azonban azt kérdezzük, hogy melyik kimenetelt fogjuk a legtöbbször megfigyelni, ha egy kockát rendkívül sokszor az első hatosig dobunk fel, akkor a fenti elméleti „első" válasz többnyire a legjobb tanulók számára is hihetetlen lesz.

Amint a binomiális rész után, most is megadunk két gyakorló feladatot. Célunk most is a tanári munka könnyítése

GGy1 feladat. Három szabályos érmét addig dobunk fel, amíg mindhárom érme ugyanarra az oldalra nem esik. Mi a szükséges dobások számának az eloszlása?

GGy2 feladat. Egy csomag magyar kártyából alapos keverés után felütünk egy lapot. Ha az babás, (alsó, felső, király, ász), akkor a húzást befejezzük, különben ismételt keverés után újra húzunk. Ezt az első babás lap húzásáig ismételjük. Mi a kihúzott lapok számának az eloszlása?

6.5. Összefüggő megfigyelések esete

Ebben a részben az egymás utáni megfigyelések szorosan összefüggnek. Az oktatási cél éppen az, hogy a gyakorlatban tipikus, összefüggő véletlen jelenségeket modellezzünk.

6.4. Példa: Három doboz

Példában három doboz szerepel, amelyek a 0, 1, 2 számokkal vannak megszámozva. Mindháromban 10 golyó van, ezekből húzunk visszatevéssel.

- A nulla sorszámú blokkban 7 golyót 0-val jelöltek meg, 2 golyót 2-vel, és egyet 1-gyel.
- Az egyes sorszámú dobozban egyaránt öt-öt golyót jelöl 0 és 2 .
- A kettes sorszámú dobozban hét golyót jelöl 2, kettőt 0, és egyet 1.

A húzást az 1 sorszámú dobozzal kezdjük. Feljegyezzük a golyó sorszámát, és ennek a sorszámnak megfelelő dobozzal folytatjuk a húzást. Ezt összesen 400-szor végeztük el.

A húzások eredményét tízesével csoportosítva a 6.4. táblázat tartalmazza. A táblázatban minden tízes csoportból véletlenszerűen kitöröltünk három számot, amelyeknek a helyét „_” vonal mutatja.

A feladat az, hogy a hiányzó „_” helyeken szereplő számokat a legtöbb találat reményében megtippeljük. A korábbi hasonló feladatokkal ellentétben most az egymás után következő betűk a tanulói számára is nyilvánvalóan nem függetlenek, ahogy arra a fentiekben utaltunk.

22221_ _2_0	0_ _22020_2	2_22_01_00	0000_10_2_	22_022_2_0
0_0002_22_	0_2221_21_	2_0_0212_0	02_ _12_000	_0000_00_2
20000_2_2_	22_ _20120_	2_ _212_000	000_022_2_	_ _ _2221222
2222_ _222_	2000000_ _ _	000222_ _ _1	0_ _122_000	000_ _00_02
2_0_22_222	_2_22_2000	222_2_2_01	_22_ _22222	22_20_ _221
2_2_22_220	0222_1_ _00	_00_0100_2	2_212_ _222	_000_000_0
_ _ _00*2*1222	2_222_2_11	2_00_2_120	22_ _0_0222	00022_1_ _2
_0_0000_22	2_2022_22_	22_0_1_010	0222_ _2_22	2_22022_ _2

6.4.táblázat: A 6.4. példa 3 dobozából húzott számok

Kimenetel :	0	1	2
Gyakoriság:	98	33	149

6.5.táblázat: A 6.4. példa megfigyeléseinek összegzése

A 6.5. táblázat szerint egy jó kérdezési sorrend lenne az 2-0-1. Általános tanulói javaslat, hogy ennél jobb stratégia is elképzelhető, ha a megelőző jeltől való függést is figyelembe vesszük. Ehhez kiindulásként határozzuk meg azokat a valószínűségeket, amelyek attól függően mérik az {„–„ jel = adott szám} esemény bekövetkezésének esélyét, hogy mi a megelőző betű. Ugyancsak általános a megegyezés abban, hogy a megelőző szám ismeretében teljesen lényegtelen, hogy a korábbi számok micsodák. Ez az a pillanat, amikor a feltételes valószínűséget definiáljuk, és a szokásos jelölést is bevezetjük. Ezt a jelölést használva, a

$$P(\text{az „-„ jel} = K \mid \text{megelőző szám} = J) \ , \ K; J = 0, 1, 2$$

feltételes valószínűségeket kell meghatározni. Meghatározásuk is természetes a tanulók számára, ezért itt csak megadjuk (mátrix formában) az értéküket.

J \ **K**	**0**	**1**	**2**
0	0,7	0,1	0,2
1	0,5	0	0,5
2	0,2	0,1	0,7

6.6.táblázat: 6.4. példa átmenet mátrixa

Látható tehát, hogy nem mindig a 2 a célszerű kérdés, 0 után inkább 0-t kell először kérdezni.

A tanulók általában nem tesznek kísérletet arra, hogy a „rákövetkező” jelet bevonják a döntési szabály megfogalmazásába.

6.5. példa: Sorsolás visszatevés nélkül

Egy dobozban 1 fehér és 7 piros golyó van. Ebből a dobozból húzunk a fehér kihúzásáig visszatevés nélkül. Kimenetel: az összes húzások X száma.

A matematikai modell építése az összes lehetőség összegyűjtésével kezdődik. A lehetséges kimenetelek az 1, 2, 3, 4, 5, 6, 7, 8 számok. A klasszikus valószínűségszámítási modell most sem működik. A szóban forgó valószínűségeket X növekvő értékei szerint haladva egymás után lehet kiszámítani. Részletezés nélkül megadjuk ennek levezetését.

$P[X=1] = 1/8 = 0{,}125; \qquad P[X>1] = 7/8$
$P[X=2] = 1/7 \cdot 7/8 = 1/8 = 0{,}125$
$P[X>2] = 1 - P[X \leq 2] = 6/8$
$P[X=3] = 1/6 \cdot 6/8 = 1/8 = 0{,}125;$
$P[X>3] = 1 - P[X \leq 3] = 5/8$
$P[X=4] = 1/5 \cdot 5/8 = 1/8 = 0{,}125;$
$P[X>4] = 1 - P[X \leq 4] = 4/8$
$P[X=5] = 1/4 \cdot 4/8 = 1/8 = 0{,}125;$
$P[X>5] = 1 - P[X \leq 5] = 3/8$
$P[X=6] = 1/3 \cdot 3/8 = 1/8 = 0{,}125;$
$P[X>6] = 1 - P[X \leq 6] = 2/8$
$P[X=7] = 1/2 \cdot 2/8 = 1/8 = 0{,}125;$
$P[X>7] = 1 - P[X \leq 7] = 1/8$
$P[X=8] = 1/1 \cdot 1/8 = 1/8 = 0{,}125;$

Az eredményként adódó valószínűség-eloszlás:

1	2	3	4	5	6	7	8
0,125	0,125	0,125	0,125	0,125	0,125	0,125	0,125

Meglepőnek látszik, hogy ezek a valószínűségek egyenlők.

Javasoljuk, hogy a tanulók hajtsák végre sokszor a megfigyelést (legalább 120-150-szer).

6.3.Definíció: Legyen A egy tetszőleges véges halmaz. Ha egy X véletlen szám értékei ebbe a halmazba esnek, és ott valamennyi értéket egyenlő valószínűséggel vehet fel, akkor X-et A-n egyenletes eloszlású véletlen számnak nevezik.

Példánkban A elemszáma 8. Fej-írás sorozatok esetén A elemszáma 2 (bináris eset). Szabályos kocka esetén a paraméter értéke 6. A példa alapján tetszőleges k esetén megadhatunk egy módszert az $1/k$ valószínűségű egyenletes eloszlás generálására.

6.6. A józan szemlélet csalhat.

Ebben a részben egy olyan típuspéldával foglalkozunk, ami a „józan észnek” ellentmond. Időről-időre találkozhatunk ilyen esettel, ezért a tanulókkal való megismertetésük fontos oktatási cél. Lényegesen kevésbé volt meglepő, de ilyen eset volt az „első hatos” leggyakoribb kijövetelének esete is.

6.6. Példa: 2 golyó vagy 5 golyó?

Véletlenszerűen kiválasztottuk a következő két változat egyikét.

A változat: Beleteszünk 2 fehér és 2 fekete golyót egy dobozba. Ebből a dobozból kihúzunk két golyót, és 0-t írunk le, ha azonos színűek a golyók, 1-et, ha különbözőek.
B változat: Beleteszünk 5 fehér és 5 fekete golyót egy dobozba. Ebből a dobozból kihúzunk két golyót, és 0-t írunk le, ha azonos színűek a golyók, 1-et, ha különbözőek.

A kisorsolt változatot hússzor végrehajtottuk. Így a következő sorozatot kaptuk:

1,0,0,0,0, 0,1,1,1,0, 1,1,0,0,0, 1,0,0,1,0

Ennek a sorozatnak az alapján kell eldönteni, hogy melyik változatot választottuk ki.

Ennél a döntési feladatnál az *első reakció általában az, hogy nem lehet igazi döntést hozni*, hiszen mindkét esetben a fehér és a fekete golyók száma ugyanakkora, tehát az azonos színűek húzásának az esélye is megegyezik. Bármilyen tetszetősen hangzik is ez az okoskodás, ellenőrizni kell.

Kérdés, hogy egy N hosszúságú megfigyelés-sorozat esetén milyen eséllyel számíthatunk általában a helyes döntésre? Mekkora esélyünk van adott relatív gyakoriság mellett az *A változat*ra, és mekkora a *B változat*ra? Ehhez mindkét esetben ki kell számítani, hogy mekkora valószínűséggel lesz a két kihúzott golyó azonos színű.

A valószínűséget mindjárt általánosan arra az esetre számítjuk ki, amikor a dobozban k fehér és k fekete golyó van, $k>1$.

P (a két kihúzott golyó azonos színű) =

$=P$ (a két golyó fehér vagy a két golyó fekete) $= 2\dfrac{k(k-1)}{2k(2k-1)} = \dfrac{k-1}{2k-1} \neq \dfrac{1}{2}$

Az *A esetben* tehát a valószínűség 1/3=3/9, míg a *B esetben* 4/9. A két valószínűség tehát eltér, hosszú megfigyelés-sorozat esetén megbízható döntésre számíthatunk. „Jó” döntést jelenthet, ha aszerint döntünk, hogy a 0 relatív gyakorisága kisebb, vagy nagyobb, mint (1/3 + 4/9) / 2 = 7/18. Esetünkben a 0 relatív gyakorisága 12/20 = 3/5 =54/90 > 35/90 = 7/18. Döntési elvünk szerint a minta az *A változattal* készült.

6.7. Csak statisztikai modell létezik

6.7. Példa: betűk közti távolság

Egyenlő valószínűséggel kiválasztottuk az „e”, „t” és „szóköz” írásjelek egyikét.

A változat: Az „e” betűt választottuk ki.

B változat: A „t” betűt választottuk ki.

C változat: A szóközt választottuk ki.

Ezután egy magyar nyelvű szépirodalmi könyvet véletlenszerűen kinyitottunk. Megszámláltuk, hogy hány betű volt az első kiválasztott jelig a lapon, majd innen kezdve hány betű fordult elő a másodikig, aztán innen a harmadikig, és így tovább a huszadikig.

Megfigyeléseink:

1, 5, 2, 3, 4, 2, 3, 14, 3, 2, 2, 4, 3, 14, 3, 6, 3, 3, 17, 25

Ebben a feladatban nem tudunk egzakt matematikai modellt levezetni. Valószínűségek helyett hatalmas mintából készített relatív gyakoriságokat használhatunk. Ilyeneket már használtunk a korábbiakban, amikor az egyszerű helyettesítéses titkosírással foglalkoztunk. A 4. fejezetben meg is adtunk a gyakoriság-eloszlásokat. Itt most egy újabbat azért adunk meg, hogy a tanár láthassa, mire számíthat 200-300 betűs szöveg frekventálása alapján.

Hossz	1	2	3	4	5	6	7	8	9-20	≥21	Össz:
Mintában	1	4	6	3	1	1	0	0	3	1	20
Szóköz	0	27	34	28	35	40	21	13	65	0	263
„e”	1	28	31	21	18	14	11	4	45	31	204
„t”	35	16	12	17	11	11	8	12	69	98	289

6.7. táblázat: Gyakoriságok

A megoldandó feladat alapján azt kell eldöntenünk, hogy a *minta gyakoriság-eloszlása* a másik három gyakoriság-eloszlás közül *melyikre hasonlít legjobban*. A tanulók tipikusan kifejtik, hogy így nem lehet őket összehasonlítani, először „közös nevezőre” kell őket hozni. Ez azt jelenti, hogy gyakoriságok helyett relatív gyakoriságokat kell összehasonlítani.

Hossz	1	2	3	4	5	6	7	8	9-20	≥21
Minta	0,05	0,2	0,3	0,153	0,05	0,05	0	0	0,15	0,05
Szóköz	0	0,103	0,129	0,106	0,133	0,152	0,08	0,049	0,247	0
„e”	0,005	0,137	0,152	0,103	0,088	0,069	0,054	0,020	0,221	0,152
„t”	0,121	0,055	0,041	0,059	0,038	0,038	0,028	0,042	0,239	0,339

6.8. táblázat: Relatív gyakoriságok

Most már csak azt kell eldönteni, hogy a mintának megfelelő gyakoriság-eloszlás a másik három közül melyikre hasonlít legjobban? Erre vonatkozóan idézzük a Berzsenyi Dániel Gimnázium egyik tanulóját. „Nyilván arra hasonlít legjobban, amelyikhez legközelebb van. Mi az, hogy melyikhez van legközelebb? Hát ezek a 10 dimenziós euklideszi tér pontjai, így csak ki kell számítani az euklideszi távolságokat.”

Ezt most nem végezzük el, bár ránézésre is látszik, hogy az „*e*” betűt választottuk.

6.8. Szóhasználat

Egy véletlen jelenség esetén azt fogjuk **esemény**nek nevezni, amelyekről a jelenség megfigyelése során egyértelműen eldönthető, hogy bekövetkezett-e vagy sem. Események jelölésére nyomtatott nagy betűket használunk.

Azokat az eseményeket, amelyeknek semmilyen részeseményét nem lehet megfigyelni, csak magát az eseményt, **elemi esemény**nek nevezik. Egy adott kísérlet esetén az összes elemi esemény együttesét **összes esemény**nek nevezzük.

Egy véletlen jelenség valamely kimenetelére, vagy eseményére vonatkozóan **kedvező esemény**nek nevezzük azokat az elemi eseményeket, amelyeknek a bekövetkezése maga után vonja az adott kimenetel vagy esemény bekövetkezését.

Szokás beszélni egy esemény **ellentett esemény**éről, röviden ellentett eseményről is. Ez akkor és csak akkor következik be, ha a szóban forgó esemény nem következik be.

Ha azt akarjuk kifejezni, hogy két esemény közül legalább az egyik bekövetkezik, akkor a **két esemény összeg**éről beszélünk. Az A és B események összegét $A+B$ jelöli. Ha két esemény mindegyike egyszerre bekövetkezik, akkor a **két esemény szorzatá**ról beszélünk. Az A és B események szorzatát AB jelöli,

Ha az A esemény bekövetkezése esetén a B esemény mindig bekövetkezik, akkor azt mondjuk, hogy az A esemény maga után vonja a B eseményt.

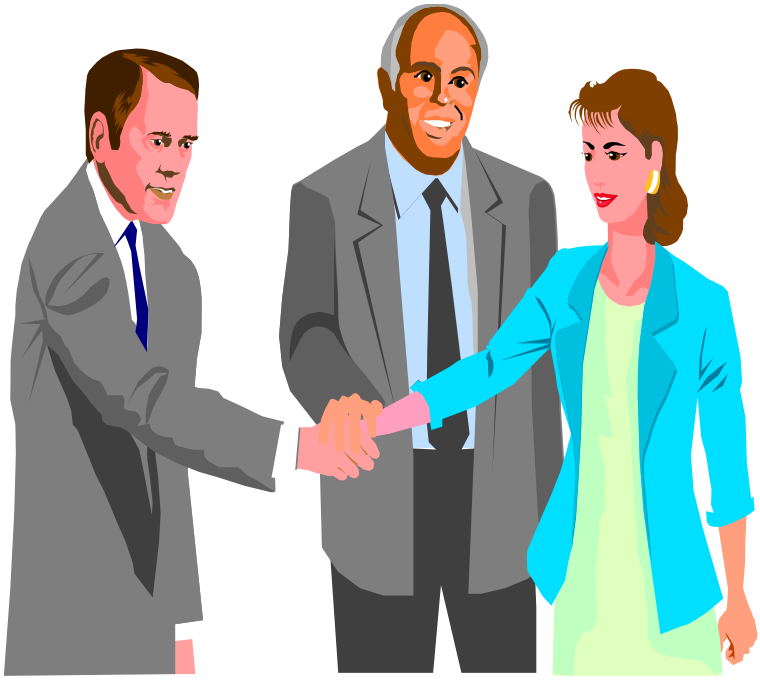
Van két speciális esemény. Az egyik az, amelyik soha nem következik be, ez a **lehetetlen esemény**, és a másik az, amelyik mindig bekövetkezik, ez a **biztos esemény**.

Két eseményt **egymást kizáró esemény**nek nevezünk, ha egyszerre soha nem következhetnek be.

Egy megfigyelés-sorozatban egy adott esemény **gyakoriság**ának nevezik azt a számot, ahányszor az esemény bekövetkezett. Ha ezt a számot elosztjuk a megfigyelések összes számával, akkor az esemény **relatív gyakoriság**át kapjuk.

Egy véletlen jelenség modelljét akkor adjuk meg, ha felsoroljuk valamennyi lehetséges kimenetelét, és megadjuk ezeknek a kimeneteleknek a valószínűségeit.

VII. Közvélemény-kutatás: statisztikai mintavétel



7.1. Bizonyítás helyett meggyőzés

Napjaink politikai életében egyre gyakrabban esik szó a közvélemény kutatók „jóslatairól", amit sokszor kemény kritikával illetnek. Ilyenkor mondják a felmérést végzők, hogy a mi felmérésünk „*reprezentatív mintavételen*" alapszik. Jelen fejezetben azt mutatjuk meg, hogy milyen egyszerűen lehet a tanulói gondolatvilágot helyes mederbe terelni. Ha nem akarunk szigorú és nehéz matematikai bizonyításokkal operálni, akkor lehetőség van olyan statisztikai feladatok elvégzésére, amelyek meggyőzően mutathatják be, hogy a megfelelő statisztikai eljárások valóban a valós közvélemény megismerését teszik lehetővé. A vélemény kialakításának kulcsmondata: *bizonyítás helyett meggyőzés.* [S04]

A magyar nyelvvel kapcsolatos statisztikai vizsgálatainkra való hivatkozással azt kívánjuk megmutatni, hogy a tanulók meggyőzhetők arról, hogy statisztikai mintavétel segítségével megbízhatóan lehet a közvéleményt megismerni. Iskolai óra keretén belül illusztráljuk a vélemények összegyűjtésének folyamatát, és tartalommal töltjük meg a „reprezentatív" szót is.

7.2. A mintavételről általában

A statisztikát tanító tanár számára a **mintavétel szerepe és fontossága** nyilvánvaló. Ennek számos oka lehet. Így például:

- Az adott feladatban lehetetlen lehet a teljes statisztikai sokaságot megszámolni (például vadon élő állatok számát megállapítani)
- Lehetséges, hogy a megvizsgált objektumok tönkremennek, megsemmisülnek a vizsgálat során. (Például villanyégők élettartamának vizsgálatakor)
- Az is lehetséges, hogy nagyszámú egyed vizsgálata lehetetlen vagy megbízhatatlan (például a közvélemény megismeréséhez)

A tanár számára az is nyilvánvaló, hogy a mintavételnek hogyan kell történnie, és hogy miért lehet kis mintából az egészre következtetni. Ami odafigyelést, tervezést igényel, az a mód, ahogyan ezt az ismeretet az osztályközösség számára elérhetővé lehet tenni. A hagyományos iskolai oktatásban erre két lehetőség is kínálkozik:

- Követhetjük a matematika „állítás – bizonyítás" típusú oktatását,
- Megfelelően választott mintavételezési feladat segítségével demonstrálhatjuk az eljárás hatékonyságát.

Az átlagtanuló számára az első módozat kevésbé hatékony. Ebben a fejezetben ezért a második lehetőséggel foglalkozunk. Ehhez alkalmas alkalmazási területet kell találni, ahol a mintavételezés osztálykörülmények között elvégezhető, hasonló körülmények között megismételhető. Magának a feladatnak is értelmesnek kell lennie, amelyekről nem jut a tanulók eszébe, hogy ezt a "tanár találta ki számunkra".

Erre a feladatra kívánunk most egy példát mutatni. Példánkat a Fazekas Mihály Gyakorlógimnáziumban több éven át teszteltük eredményesen.

7.3. A statisztikai feladat megfogalmazása

Egy nagyobb statisztikai sokaságot, amelyet vizsgálni akarunk, a statisztikában szokásos módon **populáció**nak ("népességnek"), elemeit **egyed**nek nevezzük. Ennek bármilyen módon választott bármilyen részhalmazát a köznyelv *mintának* nevezi. A statisztikában általában ez a részhalmaz kicsi, viszont belőle nagyravágyó következtetéseket szeretnénk levonni. Ennek érdekében a részhalmaz kiválasztására vonatkozó szabályokat állítanak fel. Az ennek megfelelő kiválasztási folyamatot **mintavételezés**nek nevezik. A mintavételezés folyamatát, annak paramétereit kissé formálisan ugyan, de egyértelműség miatt megfogalmazzuk:

1. Legyen a populáció egyedeinek száma ***N*.** Ezt a **populáció-méretnek** nevezzük.
2. Tegyük fel, hogy a populációban ***M*** egyed valahogyan meg van jelölve ("kiválasztottak" halmaza).
3. Választunk egy ***n*** elemű részhalmazt, amit **mintának** neveznek. Az *n* elemszám a minta mérete (terjedelme).
4. Legyen az *n* elemű mintában **pontosan *m* megjelölt** egyed.
5. A mintát **reprezentatív mintának** nevezik, ha a populáció valamennyi *n* elemű részhalmaza közül egyenlő valószínűséggel választunk egyet.

A matematikai statisztikában mindig feltesszük, hogy a minta kiválasztása reprezentatív módon történik. Ezen feltétel mellett ki tudjuk számítani, hogy mekkora a másik négy paraméter egyidejű megjelenésének a valószínűsége. Ezt klasszikus valószínűségszámítási módszerrel kaphatjuk meg. Ez az a pont, ameddig a tanár narratív, előadói szerepe nélkülözhetetlen. Ettől a ponttól számíthatunk a tanulók aktív, vagy legalább interaktív közreműködésére.

A továbblépés a tanulói előképzettség szerint elágazik. A mi gyakorlatunkban erre az egységre olyankor kerül sor, amikor a tanulók már rendelkeznek minimális kombinatorikai ismerettel. Alapvetően elegendő a

binomiális együtthatók és értelmezésük ismerete. (Ennél többre van szükség az előző fejezetben szereplő binomiális eloszlás tárgyalásához.)

A szóban forgó valószínűség meghatározásához meg kell mondani az összes események számát, ami a binomiális együttható definíciója szerint triviális. Ugyancsak meg kell határozni a kedvező esetek számát. Ez sem jelent problémát a legalább átlagos tanulóknak.

A kettő hányadosa a keresett valószínűség.

$$H(N,n,M,m)=\frac{\binom{M}{m}\cdot\binom{N-M}{n-m}}{\binom{N}{n}}$$

A mintavételtől egy **statisztikai döntési feladat** megoldását várjuk. Ez tipikusan abban áll, hogy a négy paraméterből valamelyik három ismert (mérhető, leszámolható), míg a negyedik ismeretlen. Erre a negyedik paraméterre kell "jó" becslést találni a három ismert paraméter függvényében. Bárhogyan vannak is megfogalmazva, bármilyen ruhába legyenek is öltöztetve, a konkrét megjelenési formától levetkőztetve őket, mindegyik a következő négy alapeset egyikét jeleníti meg. Ezt a négy esetet egyértelműen meghatározza az ismeretlen (hiányzó) paraméter. Most felsoroljuk a négy esetet, önkényesen egy-egy fantázia nevet adva nekik.

1) ***Közvéleménykutatás***: Tegyük fel, hogy *M* ismeretlen, míg a másik három paraméter értéke ismert. A feladat *M*-re egy "jó" becslést adni. Egy ilyen helyzet például az, amikor a szeretnénk tudni, hány ember vallja magát hazánkban materialistának. Ekkor *N* az ország népessége, *n*-et viszont mi magunk választjuk meg.

2) ***Madárles:*** Tegyük fel, hogy *N* ismeretlen, míg a másik három paraméter értéke ismert. A feladat *N*-re egy "jó" becslést adni. Egy ilyen helyzet áll elő például akkor, ha egy madárfaj példányainak a száma érdekel bennünket. Ekkor megválasztjuk *M*-et, (pl. meggyűrűzünk *M* madarat,) majd egy hosszabb időt kivárva elfogunk *n* madarat és megszámoljuk a meggyűrűzöttek számát. Az *n* és *M* paramétereket mi magunk választjuk meg.

3) ***Placebo*** (gyógyszer): Tegyük fel, hogy *n* ismeretlen, míg a másik három paraméter értéke ismert. A feladat *n*-re egy "jó" becslést adni. Egy ilyen helyzet fordulhat elő egy új gyógyszer hatékonyságának a bevizsgálásakor. Ekkor *N* a vizsgálatba bevont diagnosztizált betegek száma. Köztük véletlenszerűen osztjuk szét a valódi és a placebo

gyógyszert. Legyen *M* a gyógyultak száma, és *m* ezek között azok száma, akik a valódi gyógyszert kapták.

4) ***Minőségellenőrzés***: Tegyük fel, hogy *m* ismeretlen, míg a másik három paraméter értéke ismert. A feladat *m*-re egy "jó" becslést adni. Egy ilyen helyzet fordulhat elő szállítási szerződésekkel kapcsolatban. Ekkor a leszállítandó termékek száma *N*, a szállítótól akkor veszik át a leszállított terméket, ha ezek között legfeljebb *M* a hibásak száma. Szerződés szerint a vevő egy *n* elemű mintát vesz, és ha abban a hibásak *m* száma egy küszöbszámnál nagyobb, megtagadhatja a szállítmány átvételét.

7.4. A legnagyobb valószínűség elve

A fenti statisztikai döntésfeladatok közös jellemzője, hogy egy ismeretlen paraméterre keresünk "jó" becslést. Az ismeretlen paraméter minden értékére ismert annak a *H(N,n,M,m)* valószínűsége, hogy az *N, n, M* paraméterekkel jellemzett mintában pontosan *m* megjelölt elem van. Egy jó becslésnek ezt a valószínűséget kell maximalizálnia. Ez az elv vezet el bennünket a döntéshez. A kifejtett vezérelv a statisztikában általánosan elfogadott és alkalmazott. Neve a ***legnagyobb valószínűség elve***. Az elv elfogadása után a döntésfeladat egy matematikai szélsőérték feladattá egyszerűsödik. (Persze általában semmi sem garantálja, hogy a matematikai feladat egyszerűen megoldható.)

Kedvező **tanári tapasztalat**, hogy a témakört a tanulók érdeklődéssel követik, és számukra természetesként jelenik meg a legnagyobb valószínűség elve is. Ijesztőek viszont számukra a matematikai szélsőérték feladatok, főként a binomiális együtthatókkal kapcsolatos járatlanságuk miatt. Javasoljuk, hogy megnyugtatásul a tanár egyik szélsőérték feladatot interaktív módon oldja meg a tanteremben, a többi esetén egyszerűen közölje a megoldást. (Az "időprés" többre aligha hagy időt.)

Egyszerűsége miatt a **"közvélemény kutatási" feladatot javasoljuk frontális tárgyalásra**. A szélsőérték meghatározásához próbáljuk ki szerencsénket. Szerencsésnek azt neveznénk, ha a *H(N,n,M,m)* valószínűség *M* függvényében valameddig nőne, aztán csökkenne. (Szokásos matematikai megoldás - próbálkozási módszer.) Ezt ellenőrizhetjük két egymás utáni paraméteres *M* értékhez tartozó valószínűség összehasonlításával. Az összehasonlítás során a nevezőben szereplő összes mintaszám nem függ az ismeretlen *M*-től, ezért az összehasonlítás lényegesen egyszerűsödik. A

$$H(N, n, M+1, m) > H(N, n, M, m)$$

egyenlőtlenség akkor és csak akkor teljesül, ha teljesül az

$$\binom{N-M-1}{n-m}\cdot\binom{M+1}{m}>\binom{M}{m}\cdot\binom{N-M}{n-m}$$

egyenlőtlenség is teljesül.

A binomiális együtthatókkal a szokásos manipulációkat elvégezve egyenlőtlenségünk lineárissá válik. Itt a levezetést abbahagyjuk, interaktív befejezése semmi gondot sem jelent. A négy szélsőérték feladat megoldását összegezve kapjuk a közel optimális becsléseket.

Közvéleménykutatás: Az ismeretlen *M* paraméter közel optimális becslése

$$\widetilde{M} \approx N\cdot\frac{m}{n}$$

Madárles: Az ismeretlen *N* paraméter közel optimális becslése *M*n/m* egész része,

$$\widetilde{N}\approx M\cdot\frac{n}{m}$$

Placebo (gyógyszer):. Az ismeretlen *m* paraméter közel optimális becslése

$$\widetilde{n}\approx N\cdot\frac{m}{M}$$

Minőségellenőrzés: Az ismeretlen *n* paraméter közel optimális becslése

$$\widetilde{m}\approx M\cdot\frac{n}{N}$$

A tanárnak minden körülmények között készen kell lennie arra, hogy a négy alapesetnek megfelelő "csontvázat" a katedráról kapásból felöltöztessen. A következő négy történetből álló olvasmányszerű esetgyűjtemény ehhez kíván segítséget nyújtani.

7.5. Kiegészítés, olvasmány: történetek

1. történet: A "madárles" témához kapcsolódik a következő történet. A II. Világháborút követően a Nemzetközi Bálnaügyi Bizottság szerint a különböző vadásztársaságok profitéhsége miatt a kékbálnák száma rohamosan csökkent. Ennek a ténynek az elfogadtatásához megbízhatóan meg kellett számlálni a bálnákat. Megszámlálásukra következő módszert dolgozták ki.

Egy bizonyos időn át kb. 30 cm hosszú fémhengereket lőttek be a bálnák zsírpárnájába. Feljegyezték, hogy hány bálnát jelöltek meg fémhengerrel. Bizonyos idő eltelte után, amikor már feltételezhető volt, hogy a bálnák

rendszeres mozgásuk révén elkeveredtek, felszólították (nemzetközi rendelettel kötelezték) a bálnavadászokat, hogy rendszeresen jelentsék az általuk kifogott kékbálnák számát, és azt, hogy ezek között hány volt megjelölve.

Nyilvánvaló a „madárles" absztrakt változatával való kapcsolat.

- A kékbálnák ismeretlen N száma a populáció mérete
- A belövéssel megjelöltek száma az M paraméter
- A kihalászottak képezik a mintát, számuk n.
- A kihalászottak között megjelöltek száma a m paraméter.
- A reprezentatív mintavétel csak megközelítően teljesül.

2. történet: „Közvélemény-kutatás". A magyar jog szerint 200 ezer aláírás benyújtásával bármely olyan kérdésben lehet népszavazást kezdeményezni, amely nem alkotmányellenes. 1997-ben a földszerzéshez fűződő jogokkal kapcsolatban mintegy 300 ezer aláírást nyújtottak be. Ezeket az Országos Választási Bizottságnak ellenőriznie kellett, hogy van-e köztük legalább 200 ezer hiteles. Valamennyi érdekeltet megkérdezni nyilvánvalóan gyakorlatilag kivitelezhetetlen lenne. Ezért a közvélemény-kutatási mintavételt alkalmazták.

3. történet: „Minőség-ellenőrzés". Mezei András "Magyar kocka, avagy még mindig ilyen gazdagok vagyunk" c. könyvében arról számol be, hogy az USA-ba leszállított egymillió Rubik kocka között 800 000 selejtes volt. Egy átvételi terv szerint a vevőnek akkor kell átvennie a tételt, ha 1000 véletlenszerűen kiválasztott kocka között legfeljebb 33 hibásat találnak. Szereposztás:

- A populáció N mérete: a leszállított kockák száma, ismert
- A megjelöltek M száma: hibásak száma, ismert, M=800 000
- A minta n nagysága: szerződésben meghatározva, ismert n=1000
- A mintában szereplő m megjelöltek száma: hibásak a mintában, ismeretlen. Kérdés: $m > 33$?
- A mintavétel az átvételi terv szerint reprezentatív.

Megjegyzés: az átvétel $P(m<34)$valószínűségének első 14 tizedesjegye nulla.

4. történet. „Minőség-ellenőrzés" KENO játék. Ebben a játékban az 1,2,...,80 egész számok közül a játékos tetszése szerint kiválaszthat $1 \le M \le 10$ számot. Minden sorsolás alkalmával véletlenszerűen kisorsolnak 20 számot. A játék nyereménye a találatok m számától függ. Ez a szerencsejáték a felsorolt négy eset közül a negyedik.

Szereposztás:

- A populáció N mérete: a teljes számkészlet, ismert, N=80
- A megjelöltek M száma: száma, kiválasztott számok, választás után ismert
- A minta n nagysága: a húzott számok száma, ismert $n = 20$
- A mintában szereplő megjelöltek száma: a kihúzott tippek száma, ismeretlen
- A mintavétel a sorsolás szerint reprezentatív

7.6. Elméleti számítás helyett szimuláció

Természetesen merül fel a tanulók részéről a kérdés (és mindig valóban fel is merül), hogy "***mit jelent az, hogy egy becslés jó?"*** A válasz egyszerű: ***egy jó becsléstől azt várjuk el, hogy értéke a paraméter valódi, bár ismeretlen értéke közelébe essen***. Sajnos iskolai szinten nem tudjuk ezt az elvárást szabatosan megfogalmazni, még kevésbé matematikai szabatossággal bizonyítani, hogy a legnagyobb valószínűség elve szerint kapott fenti becslések rendelkeznek ezzel a tulajdonsággal. .

Mit lehet ilyenkor tenni? Hasonló helyzetek tömegét kellene megvizsgálni (szimulálni), ahol össze tudjuk hasonlítani a legnagyobb valószínűség elve szerint adódó becsléseket a paraméter valódi értékével. Ha ezek az esetek többségében közel vannak egymáshoz, akkor elhihetjük, hogy elvünk tényleg jó becslést eredményez. A statisztikai alkalmazások területén az olyan következtetésre alkalmas megoldásokat, amelyek nagyszámú, mesterségesen konstruált körülmények során tapasztalt megfigyelésekre támaszkodnak, szimulációs eljárásoknak neveznek. Mi most ilyen megoldással foglalkozunk.

Mindenekelőtt a vizsgálandó populáció típusát kell ügyesen megválasztani.

Elvárjuk, hogy

1) Nagyszámú, jellegében megegyező, de megjelenésében eltérő populáció-kópiát lehessen megfigyelni, és köztük véletlenszerűen lehessen választani.
2) Könnyen lehessen a populációt előállítani és megfigyelni.
3) Ne legyen "ezt a populációt a demonstráció kedvéért készítették" érzete az adatsokaságnak.
4) Létezzenek olyan statisztikai ismérvek, amelyek egyedeket véletlenszerűen jelölnek meg.
5) A megjelöltek száma gyorsan megállapítható legyen.

6) Könnyen lehessen adott elemszámú reprezentatív mintát generálni. Egyszerű és gyors legyen a mintában levő megjelöltek számának a meghatározása.
7) Már legalább 20 éve kacérkodunk azzal a gondolattal, hogy a megfelelő populáció-sokaságot írott értelmes szövegek jelentsék. Bár voltak biztató kezdeti próbálkozásaink, valóban átütő sikerről csak a gyors szövegszerkesztővel rendelkező PC-k megjelenése óta számolhatunk be.

Szinte utolsó akadályként a 2) elvárással kapcsolatban az adatbevitel fáradságos munkáját az Internetről letölthető szépirodalmi anyagok, újságcikkek és egyéb szövegek szüntették meg.

Ismérvnek önkényesen választhatjuk az írásjegyek bármely halmazát. Azok az írásjegyek lesznek a megjelölt egyedek, amelyek ebbe a halmazba tartoznak. Leszámolásukhoz még programozni sem kell tudni: a betűgyakoriságok leszámolásakor ismertetett tanulói ötlet "cseréld ki önmagára" most is gyors megoldást tesz lehetővé. Az egyetlen viszonylag bonyolultabb feladat a reprezentatív minta kiválasztása. Ehhez egyelőre csak programozási megoldást ismerünk. Maga a program viszont nagyon egyszerű, minden osztályban több tanuló is el tudja készíteni. A mintavétellel a következő fejezetben foglalkozunk.

7.7. Véletlen minta generálása

Frontális foglalkozásként (lehetőleg interaktív módon) célszerű indítani a tárgyalást.

Ennek a résznek a témája, hogyan lehet véletlenszerűen kiválasztani N elem közül n elemet.

Mindenek előtt megállapítjuk, hogy ebben a feladatban lényegtelen az, hogy mik a konkrét elemek. Ezeket a konkrét elemeket ugyanis önkényesen megszámozhatjuk a 1, 2,...,N számokkal, és így közülük minden n-es kiválasztása esetén automatikusan adódik egy szám n-es is. Az is nyilvánvaló, hogy az 1,2,...,N számok közül kiválasztva egy szám n-est, az adott megszámozás mellett egyértelműen kapunk az eredeti elemek közül is n-es kiválasztást. Ezért elegendő foglalkozni azzal az esettel, amikor az alappopuláció az 1, 2,...,N számokból áll.

Első lépésként állapítsuk meg, hogy reprezentatív minta esetén bármely számnak egyenlő esélye van arra, hogy bekerüljön a mintába. Ennek az esélynek a valószínűsége tehát $1/N$. Automatikusan elhangzik az interaktív tanulói javaslat.

Alkalmazzuk a matematika szokásos visszavezetési trükkjét: Sorsoljuk ki $1/N$ valószínűséggel, hogy N-et beválasztjuk-e a mintába. Ha igen, akkor a maradék N-1 elemű populációból n-1 elemű reprezentatív mintát, ha nem akkor n elemű mintát kell generálni. A trükk eredménye:

Mintavételi algoritmus: Legyen K és L két nem-negatív egész értékű paraméter, S egy n elemű vektor (a kiválasztandó minta)

1. Lépés (Inicializálás): $K=N$, $L=n$, $S=\{0,0,...,0\}$, ahol S-ben pontosan n darab 0 szerepel, L/K.

2. Lépés (Véletlen választás) Generálunk egy olyan A eseményt, ami $P(A):= L/K$ valószínűséggel következik be.

3. Lépés: Ha A nem következik be, akkor legyen $K:= K$-1, L változatlan, és visszatérünk a 2. Lépésre. (Ekkor még mindig L/K, és az algoritmusnak nincs vége.)

4. Lépés: Ha A nem következett be, akkor cseréljük ki S -ben az első 0 -t K -val, legyen $K=K$-1, $L:=L$-1. Ha ekkor $L=0$ (tehát ha már nincs nem-nulla szám S -ben), akkor befejeződik az algoritmus, különben visszatérünk a 2. lépésre.

5. Lépés: Ha az algoritmus befejeződött, S megadja a kiválasztott mintát.

Megjegyzés: az algoritmus alkalmazhatóságához szükséges, hogy racionális valószínűségű véletlen eseményeket tudjunk generálni. Mivel rengeteg ilyen alkalmazásra készülünk fel, manuális generálás szóba sem jöhet. Tanári tapasztalat, hogy a tanulók többsége tud bánni a PC -kbe épített véletlenszám generátorral, és nem okoz nekik problémát a kívánt véletlen eseményt generáló $RND < L/K$ utasítás alkalmazása. A számítógépes véletlenszám generálás kérdéseivel az időprés miatt nem javasoljuk a részletesebb iskolai foglalkozást. A érdeklődő kollégák számára Knuth "bibliáját" javasoljuk olvasmányként.[K87]

A véletlen minta generálására léteznek természetesen más eljárások is. Ezek közül csak egy olyat említünk, amelyik egyszerűbb programozást tételez fel. Ennek az az ára, hogy az előre meghatározott pontos n mintanagyság helyett csak megközelítőleg ekkora lesz a mintanagyság. Az algoritmushoz megválasztjuk a $p = n/N$ valószínűséget, és a populáció valamennyi egyedéről p valószínűséggel döntjük el, hogy bekerül-e a mintába.

7.8. Szimulációs eredmények

Az Interneten esetlegesen megjelenő politika elkerülése miatt populációként az angol nyelvet választottuk.

1. demonstráció. A Times 2000 augusztusi – októberi számaiból választottunk egy összesen ***N*=50 000** karakterből álló szövegrészeket. A magánhangzókat tekintettük megjelölt egyedeknek. A szövegrészben összesen ***M*=17 050** magánhangzó (megjelölt elem) szerepelt.

Az egyszerűbb programozás kedvéért a mintanagyságot csak megközelítően megadó algoritmust választottuk, különböző *p* valószínűségek mellett.

A *p* valószínűséget 0,01 –től 0,15 –ig változtattuk 0,01 lépésközzel, és minden p mellett 10 mintát generáltunk. A minták várható nagysága **50.000*p,** aminek a várható értéke rendre 500, 1000, . . . , 7500. A 7.1. táblázat megadja generált minták valódi elemszámát. A táblázat első oszlopa mutatja a *p* valószínűséget, a következő 10 oszlop pedig a valódi elemszámokat.

p	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0.01	495	472	495	532	462	485	525	498	508	470
0.02	957	1017	1040	1016	990	1007	1021	1019	1000	1042
0.03	1568	1522	1550	1487	1475	1521	1474	1487	1545	1535
0.04	1960	2004	1994	2006	1969	1948	1951	2003	2062	2087
0.05	2376	2451	2516	2437	2582	2527	2537	2572	2486	2490
0.06	3029	2942	3013	3008	2965	2967	3008	3014	2971	2976
0.07	3472	3480	3437	3599	3556	3496	3570	3512	3534	3561
0.08	3993	3927	4076	4098	3995	4008	4035	3838	3957	3926
0.09	4519	4490	4566	4522	4471	4567	4452	4514	4365	4574
0.10	4943	4957	4980	5060	5022	4935	5023	5012	5041	4947
0.11	5489	5556	5620	5562	5418	5497	5519	5508	5589	5511
0.12	5946	6099	5878	5953	6133	5916	6118	6020	5987	5956
0.13	6355	6467	6498	6600	6546	6540	6576	6413	6506	6441
0.14	6978	6945	6853	7076	6924	7059	6901	7012	7080	7116
0.15	7449	7484	7351	7493	7422	7576	7545	7506	7497	7516

7.1. táblázat: *p* valószínűséggel generált minták nagysága

A mintaméretek 7.1. táblázatban szereplő véletlen nagyságának statisztikai jellemzőit adja meg számszerűen a 7.2. táblázat. Ezt a tanári munka megkönnyítése miatt adjuk meg, a szemléltetés elsősorban az ő számukra fontos. Hasonló felmérést egy iskolai foglalkozás tervezése során mindig el kell végezni. A konkrét eredmények azt mutatják, hogy a PC pszeudó-véletlenszám generátora nem teljesen megbízható.

100·p	Min	Max	$\bar{x}$	d^2	d	$\bar{x}$ -3d	$\bar{x}$ +3d
1	462	532	494,2	531,5	23,05	425	563,4
2	957	1042	1010,9	613,4	24,77	936,6	1085,2
3	1474	1568	1516,4	1138,7	33,74	1415,2	1617,6
4	1948	2087	1998,4	2127,8	46,13	1860,0	2136,8
5	2376	2582	2497,4	4026,3	63,45	2307,0	2687,8
6	2942	3029	2989,3	809,3	28,45	2904,0	3074,6
7	3437	3599	3521,7	2590,,9	50,90	3369,0	3674,4
8	3838	4098	3985,3	5942,2	77,09	3754,0	4216,6
9	4365	4574	4504,0	4074,7	63,83	4312,5	4695,5
10	4935	5060	4992,0	2036,7	45,13	4856,6	5127,4
11	5418	5620	5526,9	3273,9	57,22	5355,2	5698,6
12	5878	6133	6000,6	7886,7	88,81	5734,2	6267,0
13	6355	6600	6494,2	5817,7	76,27	6265,4	6723,0
14	6853	7116	6994,4	7737,6	87,96	6730,5	7258,3
15	7351	7576	7483,9	4077,9	63,86	7292,3	7675,5

7.2. táblázat: A 7.1. táblázat adatainak statisztikai jellemzői

A 7.3. táblázat megadja a mintában levő magánhangzók (megjelöltek) *m* számát

p	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0.01	203	177	210	216	190	206	210	211	198	209
0.02	405	411	407	399	407	405	409	402	413	424
0.03	638	604	626	563	621	634	610	601	627	635
0.04	798	809	826	802	782	820	777	825	827	881
0.05	972	1014	1025	1007	1059	1087	1001	1090	976	1032
0.06	1239	1159	1218	1223	1237	1214	1280	1237	1221	1185
0.07	1410	1418	1364	1501	1423	1451	1463	1423	1436	1461
0.08	1687	1614	1661	1666	1610	1596	1635	1612	1637	1585
0.09	1844	1874	1838	1814	1805	1828	1886	1823	1778	1902
0.10	2052	1968	2067	2063	1996	1984	2089	2014	2118	2066
0.11	2215	2263	2306	2266	2220	2236	2219	2236	2341	2327
0.12	2448	2520	2389	2386	2539	2437	2530	2437	2483	2460
0.13	2598	2680	2661	2748	2668	2653	2686	2655	2635	2697
0.14	2922	2832	2796	2934	2802	2915	2899	2877	2878	2913
0.15	3062	3016	2934	3084	2979	3063	3122	3016	3072	3051

7.3. táblázat: A mintában szereplő megjelöltek *m* száma

Ugyanúgy, mint a minták véletlen nagysága esetén, most is megadjuk a statisztikai jellemzőiket, lásd 7.4. táblázat.

100 ·p	Min	Max	$\overline{x}$	d^2	d	$\overline{x}$ -3d	$\overline{x}$ +3d
1	177	216	203.0	138.4	11.77	167.7	238.3
2	399	424	408.2	47.5	6.89	387.5	428.9
3	563	638	615.9	514.3	22.68	547.9	683.9
4	777	881	814.7	863.6	29.39	726.5	902.9
5	972	1090	1026.3	1720.9	41.48	901.8	1150.8
6	1159	1280	1221.3	1057.6	32.52	1123.7	1318.9
7	1364	1501	1435.0	1368.4	36.99	1324.0	1546.0
8	1585	1687	1630.3	1082.2	32.90	1531.6	1729.0
9	1778	1902	1839.2	1478.6	38.45	1723.8	1954.6
10	1968	2118	2041.7	2385.1	48.84	1895.2	2188.2
11	2215	2341	2262.9	2176.1	46.65	2123.0	2402.8
12	2386	2539	2462.9	2989.4	54.68	2298.9	2626.9
13	2598	2748	2668.1	1571.2	39.64	2549.2	2787.0
14	2796	2934	2876.8	2523.3	50.23	2726.1	3027.5
15	2934	3122	3039.9	2991.9	54.70	2875.8	3204.0

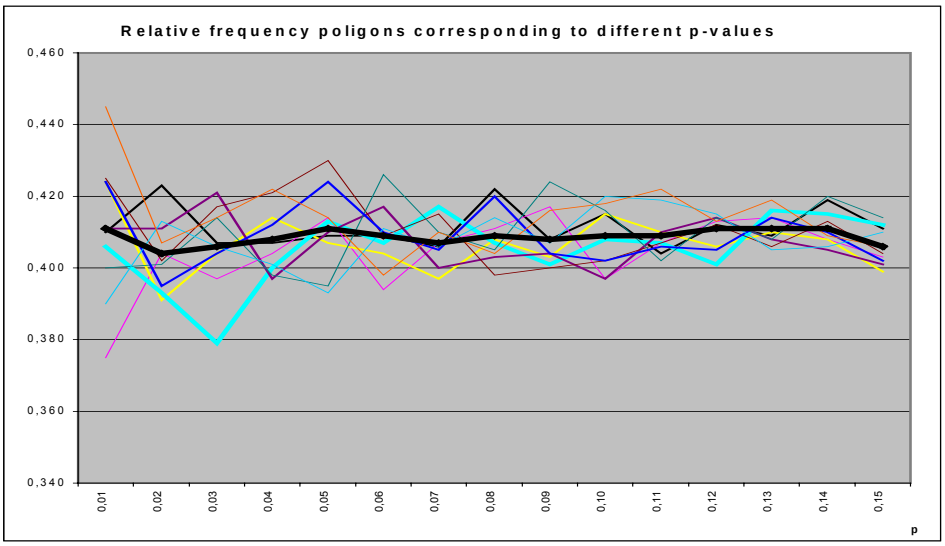
7.4. táblázat: 7.3 táblázat statisztikai jellemzői

Hatékonyabb összehasonlítást tesznek lehetővé a mintában talált megjelöltek számának relatív gyakoriságai. Ezeket összegezi a 7.5. táblázat. Érdemes megfigyelni, hogy a $p = 0{,}01$ valószínűség kivételével a minták közti eltérések, valamint az átlaguktól való eltérés kicsi.

P	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	átlag
0,01	0,410	0,375	0,424	0,406	0,411	0,425	0,400	0,424	0,390	0,445	0,411
0,02	0,423	0,404	0,391	0,393	0,411	0,402	0,401	0,395	0,413	0,407	0,404
0,03	0,407	0,397	0,404	0,379	0,421	0,417	0,414	0,404	0,406	0,414	0,406
0,04	0,407	0,404	0,414	0,400	0,397	0,421	0,398	0,412	0,401	0,422	0,408
0,05	0,409	0,414	0,407	0,413	0,410	0,430	0,395	0,424	0,393	0,414	0,411
0,06	0,409	0,394	0,404	0,407	0,417	0,409	0,426	0,410	0,411	0,398	0,409
0,07	0,406	0,407	0,397	0,417	0,400	0,415	0,410	0,405	0,406	0,410	0,407
0,08	0,422	0,411	0,408	0,407	0,403	0,398	0,405	0,420	0,414	0,404	0,409
0,09	0,408	0,417	0,403	0,401	0,404	0,400	0,424	0,404	0,407	0,416	0,408
0,10	0,415	0,397	0,415	0,408	0,397	0,402	0,416	0,402	0,420	0,418	0,409
0,11	0,404	0,407	0,410	0,407	0,410	0,407	0,402	0,406	0,419	0,422	0,409
0,12	0,412	0,413	0,406	0,401	0,414	0,412	0,414	0,405	0,415	0,413	0,411
0,13	0,409	0,414	0,410	0,416	0,408	0,406	0,408	0,414	0,405	0,419	0,411
0,14	0,419	0,408	0,408	0,415	0,405	0,413	0,420	0,410	0,406	0,409	0,411
0,15	0,411	0,403	0,399	0,412	0,401	0,404	0,414	0,402	0,410	0,406	0,406

7.5.táblázat: a mintában talált megjelöltek relatív gyakoriságai

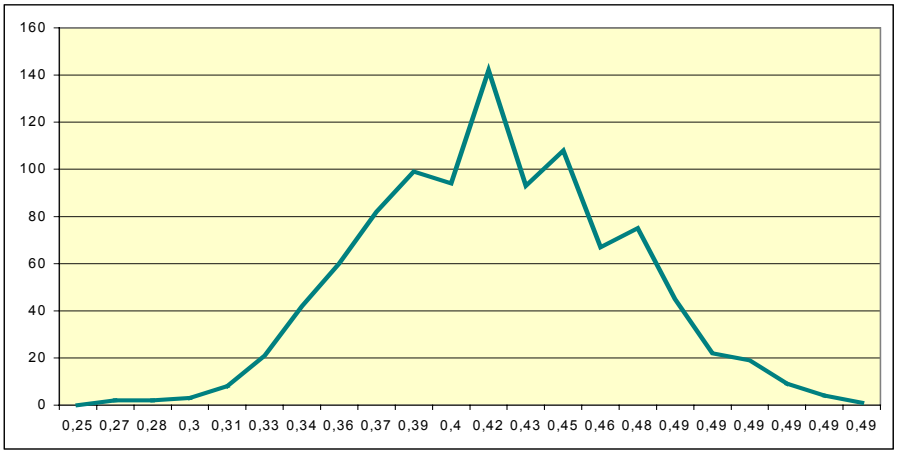
Természetesen a mintában talált megjelöltek statisztikai jellemzőit meghatározhatjuk az adott elemszámú mintavételi algoritmus esetén is. Nagyon tanulságos a kétféle úton nyert statisztikák összehasonlítása.



7.1. ábra: Különböző *p* értékekhez tartozó relatív gyakorisági poligon, az átlaggal

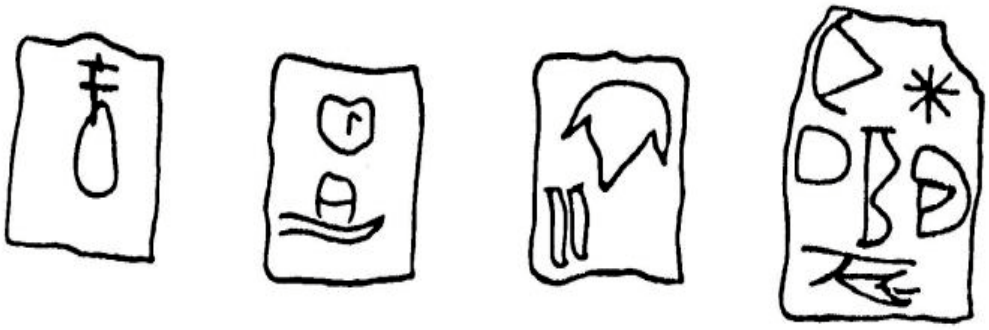
A 7.5. táblázatban szereplő eltéréseket szemléletesen mutatja be a 7.1. ábra.

Semmivel sem igényel több munkát sokkal több minta megvizsgálása. A *p*=0,02 valószínűség mellett generáltunk 1000 mintát. A kummulált relatív gyakoriságokat mutatja be 7.2. ábra.



7.2. ábra: 1000 mintából adódó relatív gyakorisági poligon

VIII. A magyar nyelv entrópiája



8.1. Bevezetés

Az írott szövegekben az előforduló helyesírási hibákat általában könnyen ki tudjuk javítani. Ezt többek közt az teszi lehetővé, hogy lényegesen több betűt használunk fel gondolataink rögzítésére, mint amennyi szükséges lenne. Az írásjegyeknek, ezen belül a betűknek ezt a bőkezű felhasználását a redundancia (terjengősség, bőség, létszámon felüli mennyiség) méri. A redundanciával ellentétes fogalom az entrópia. Ez azt jellemzi, hogy az írásjegyek optimális felhasználása esetén az írott szövegeket átlagosan hányadrészükre lehetne rövidíteni. Ebben az értelemben az entrópia egy nehezen meghatározható technikai mérőszám.

Az entrópia fogalmát C. E. Shannon [S49] vezette be és tárgyalta matematikai egzaktsággal egyik alapvető munkájában, amelyben a hírközlés egy matematikai modelljét adta meg. Shannon egy másik munkájában [S51] kidolgozott egy olyan módszert, amely a nyelvet magas fokon ismerő személyek szubjektív ítéleteinek felhasználásával az adott élő nyelv entrópiájának objektív becslését teszi lehetővé. Idézett munkájában Shannon módszerét az írott angol nyelvre alkalmazva azt találta, hogy az kb. 75%-os redundanciájú. Vizsgálatai során az írásjeleket figyelmen kívül hagyta, és csak számjegyeket nem tartalmazó szövegekre szorítkozott, így azok 27 írásjegyből (26 betű + szóköz) álltak.

Ettől az időtől kezdve általános eljárás, hogy egy adott írott nyelv entrópiájának, tömöríthetőségének vizsgálatában az adott nyelv egyjegyű betűire és a szóközre szorítkoznak. Természetesen az írásjelek elhagyásával keletkező szöveg nem minden esetben rekonstruálható egyértelműen. Klasszikus példa erre az az ókori kétértelmű jóslat, amely állítólag Krőzus lydiai király számára készült: "Ibis redibis numquam per bella peribis." A vesszők kitétele szerint jelentheti azt, hogy: „Elmégy, visszajössz, sohasem veszel el a háborúban", illetve azt, hogy: „Elmégy, sohasem jössz vissza., elveszel a háborúban".

Hasonló magyar vonatkozású példát Albericus trois-fontanesi francia ciszterci ta 1241-ig terjedő krónikájában jegyez fel. János esztergomi érsek állítólagos üzenete a Gertrúd királynő megölését tervező összeesküvőkhöz így hangzik magyar fordításban: A királynőt megö1ni nem kell félni jó lesz ha mind beleegyeznek én nem ellenzem. Tóth Béla [T96] tanúsága szerint efféle szintaktikus játékokat Párizsban a nominalisták iskoláiban az unalomig gyártottak. Ilyen mesterségesen konstruálható példák ellenére általánosan elfogadott, hogy hosszabb szöveg tartalmi összefüggései egyértelműen meghatározzák az írásjeleket.

Az entrópia nyilván nagymértékben függ a szövegek típusától. Így például a pilóták és az irányítótornyok közötti beszélgetésekben rendkívül magas redundancia tapasztalható [FS62]

Az ötvenes években különböző típusú szövegek entrópiáját vizsgálták a legkülönbözőbb célból. Közülük főként azok voltak figyelemre méltók, amelyek a fogalom eredeti jelentésének megfelelően az adattömörítés lehetőségét vizsgálták az adott technikai szinten.

Ebben a tekintetben a magyar nyelvre vonatkozó eredmények meglehetősen hiányosak. Elsősorban betűk és betűpárok gyakorisága alapján igyekeztek becslést adni a különböző típusú szövegek entrópiájára. A kutatások középpontjában nyelvstatisztikai vizsgálatok álltak. A magyar nyelvre vonatkozó nyelvstatisztikai munkákat részletesen tárgyalja Petőfi S. János:[P67]

Napjainkban ismét egyre több külföldi publikáció foglalkozik e kérdéskörrel. Ennek elsősorban az elektronikában bekövetkezett óriási fejlődés az oka. A mikroprocesszorok kifejlesztésével lehetőség nyílik olyan olcsó és kisméretű berendezések készítésére, amelyek akár százezres szókészlettel operálva is képesek adott típusú szövegeket mintegy harmadrészükre tömöríteni. Korábban ilyen úgynevezett kód-szótárak automatikus felhasználása aligha lett volna elképzelhető, jóllehet az a tény, hogy ily módon nagymérvű adattömörítés lehetséges, már több mint száz évvel ezelőtti gyakorlatból ismeretes. Az első, adattömörítésre szolgáló kódkönyvet MORSE ügyvédje. O. J. SMITH [S45] készítette. A kereskedelmi levelezésben használt könyve nagy anyagi sikert jelentett számára.

Vizsgálataink során Shannon módszere helyett a téma jellegéhez szemléletesen közelebb álló módszert választottunk, amely hiányos szövegek rekonstruálását követeli meg. A. A. Moles ugyancsak hiányos szövegek rekonstruálása útján igyekszik redundancia-becslést nyerni.[M73] Állítása szerint a francia nyelv redundanciája 50% körül van. Következtetése azonban a redundancia fogalmának félreértése miatt hibás. A kötet fordítói, Vajda András. Pléh Csaba és Kanyó Zoltán magyar szövegminták alapján is szemléltetik Moles eljárását. Négy hiányos szövegüket úgy állították össze, hogy azokból az egyes betűket egymástól függetlenül 0,1; 0,3; 0,4; 0,45 va1ószínűséggel hagyták ki.

8.2. A kísérletek 1eírása.

Vizsgálatainkhoz régebbi folyóiratokból és napilapokból 400 betűs, számjegyeket nem tarta1mazó cikkrészleteket választottunk. Ezekből elhagytuk az írásjelet., a hosszú *í, ó, ő, ú, ű* betűket rövid párjukkal helyettesítettük. A szóköz jelölésére pedig a + jelet használtuk. Végül a teljes szöveget csupa nagybetűvel írtuk le. Az így előkészített szövegekben az első néhány szót változatlanul hagytuk, azután betűket hagytunk ki bizonyos szabályok szerint. A kihagyás módja szerint öt feladatlapot állítottunk össze. Mindegyiken három szövegminta szerepelt, amelyekben a kihagyás módja és jelölése csak kismértékben tért el egymástól. A feladatlapokat római számmal, ezen belül az egyes szövegmintákat arab számokkal különböztettük meg egymástól. [S89]

A kihagyás szabályai az egyes feladatlapokon a következők voltak:

I/l. Elhagytuk az összes magánhangzót és a + jelet, és helyüket a – jellel jelöltük meg.
I/2. Elhagytuk a magánhangzókat, de nem jelöltük meg, hogy honnan.
I/3. A magánhangzókat és + jelet hagytuk el a helyük megjelölése nélkül.

II/1. A szöveget ötös csoportokba osztottuk, és minden csoportból kihagytuk a második és a negyedik írásjegyet.
II/2. A szöveg minden második írásjegyét elhagytuk.
II/3. A szöveget négyes csoportokba osztottuk, és minden csoport utolsó két írásjegyét hagytuk el.

III/1. A szöveget ötös csoportokba osztottuk, és minden csoportból két betűt hagytunk ki. Csoportonként sorsolással döntöttük el, hogy melyik kettőt. A kihagyott betű helyét megjelöltük.
III/2. Ismét csoportokba osztottuk a szöveget. Egy csoportba öt valódi betű került és annyi + jel, ahány szóköz fordult elő közben. Minden csoportból két valódi betűt hagytunk ki csoportonkénti sorolás útján. A + jelek az eredeti helyzetnek megfelelő helyen maradtak. A kihagyott betűk helyét nem jelöltük.
III/3. Ugyanúgy jártunk el, mint, a III/1.-nél, de a kihagyott betűk helyét nem jelöltük.

IV/1. Az újságcikktöredéket öt- és négybetűs csoportokba osztottuk váltakozva, azaz ötbetűs csoportot mindig négybetűs követett és fordítva. Csoportonkénti sorolással minden csoportból két betűt hagytunk ki. A szóközt is betűnek vettük a csoportosításnál. A kihagyott betűk helyét megjelöltük.
IV/2. Az előző csoportosítástól annyiban tér el, hogy a szóközt nem

tekintjük betűnek sem a csoportosításnál, sem a kihagyásnál, és az eredeti helyének megfelelő helyen jelöljük.

IV/3. Ugyanúgy hagytunk ki, mint IV/1.-nél, de a kihagyott betűk helyét nem jelöltük.

V/1-3. Ezen a feladatlapon öt- és hárombetűs csoportokba osztottuk a szöveget váltakozva. Az első és harmadik feladatban a szóközt betűnek tekintettük, a másodikban viszont nem. Minden csoportból csoportonkénti sorsolás útján két betűt hagytunk ki. Az első feladatban a kihagyott betűk helyét jelöltük, a második kettőben nem. Ezekben a feladatokban egymás mellé írva adtuk meg két egymás utáni csoport (egy ötös és egy hármas) megmaradó betűit.

Az így nyert hiányos szövegeket a függelék tartalmazza.

Az első két feladatlapon a kihagyás lényegében determinisztikusan történt, míg az utolsó hármon véletlen számok segítségével generáltuk a kihagyás helyét. Könnyű látni, hogy az utolsó három feladatlap nehézségi foka határozottan növekszik.

A szövegek rekonstruálására 15 gimnáziumi tanulót és 15 felnőttet kértünk fel. A gimnazisták a budapesti Berzsenyi Dániel Gimnázium 9.-10. osztályos tanulói voltak, a felnőttek között nyelvszakos tanárok, matematikusok, mérnökök, irodai dolgozók szerepeltek. Teljesítményüket az 1. sz. táblázatban összegeztük (D: diák; F: felnőtt). Az egyes szövegminták rekonstruálásának nehézsége jól tükröződik a táblázat adataiból.

Az I. és II. feladatlap megoldása nem jelentett különösebb problémát a kísérletben résztvevőknek. Az egy-két, értelmet nem zavaró hibás szó főként a következő típushibákból adódott: I/2: *újabb* helyett *jobb*. I/3: a) *szénkészleteink* helyett *szénkészletünk*. (Kérdéses, hogy hibának tekintendő-e ez a rekonstrukciós hiba…) b) A határozott névelő betoldása, illetve elhagyása az eredeti szöveghez viszonyítva. - II/I: A szövegben előforduló ’*e′* mutató névmás helyett ’*a*’ névelő. II/2: A MTESZ betűszó kitalálása okozott problémát. II/3: Egy-egy (általában különböző) szó maradt rekonstruálatlan.

A kísérletben részt vevők valamennyien arról számoltak be, hogy a. III-V. feladatlapok kitöltéséhez szükséges idő ugrásszerűen megnőtt. Különösen vonatkozik ez a feladatlapok 2. és 3. szövegmintájára. Ezzel magyarázható, hogy ez utóbbiak megoldására a diákok már nem vállalkoztak, és a felnőtt résztvevők száma is csökkent. A III.-V. feladatlapokkal kapcsolatban típushibákról nem beszélhetünk.

Az első két feladatlap kitöltése átlagosan 1-1 órát vett igénybe, a III-V

feladatlapok első szövegének visszaállítása maximálisan fél óráig tartott, míg a fennmaradó feladatok mindegyikével a résztvevők 1,5 - 2 órát foglalkoztak. A helyes vagy csaknem helyes megoldást adók általában több napon át, visszatérően foglalkoztak a feladattal.

Az entrópia fogalmában és definíciójában a bonyolultság és ehhez kapcsolódóan a kódoláshoz és dekódoláshoz szükséges idő nem szerepel. Ez volt az oka, hogy a feladatlapok megoldása során az időfaktort figyelmen kívül hagytuk, jóllehet pszichológiai fontosságát nem tartjuk elhanyagolhatónak.

Feladatlap	**Résztvevők**		**Hibátlan**	**Értelmet nem zavaró hibás szó**				**Értelmetlen rész**	**Nem teljes**
	Összesen			1	2	3	több		
I/1	15	D	14	1					
	15	F	14	1					
I/2	15	D	4	5	4		1	1	
	15	F	6	5	3		1		
I/3	15	D	4	5	2	1	1	1	1
	15	F	5	6	3	1			
II/1	10	D	8	2					
	15	F	7	6	1	1			
II/2	10	D	5	2	1	1	1		
	15	F	6	6			3		
II/3	10	D	2	2	1	2	1	1	1
	15	F	6	3	1	2	2		
III/1	10	D	6	2	1		1		
	13	F	5	6	1	1			
III/2	13	D	8	3	1	1			
III/3	11	F	4	4	1		1		1
IV/1	10	D	7	3					
	10	F	7	3					
IV/2	10	F	3	3	3				1
IV/3	9	F	2	2	2	1	1	1	
V/1	10	D	8	1	1				
	8	F	5	3					
V/2	7	F	2	3	2				
V/3	6	F	2	1	2				

8.1.táblázat A megoldók megoszlása és teljesítménye

8.3. Hiányos szövegek rekonstruálásával nyerhető entrópia – becslések

Ebben a részben azt vázoljuk fel, hogy értelmes írott magyar nyelvű szövegek entrópiájára miként adható felső becslés a feladatlapokon szereplő hiányos szövegtípusok rekonstruálhatóságának felhasználásával. Előrebocsátjuk azt a megjegyzést, hogy a becslést a legjobb teljesítményt nyújtó kísérleti személyek eredményei alapján kell kiszámítani. Ennek az az oka, hogy az elérhető legjobb tömörítés határát keressük, és így nyilvánvalóan nem az átlagos képességűek teljesítménye alapján kell következtetnünk. Természetesen tisztában vagyunk azzal, hogy a legjobb teljesítményt nyújtó kísérleti személyek rekonstruáló képességét nagyobb mintán (tehát több, véletlenszerűen választott szövegmintán) kellene ellenőrizni. Ezért eredményeink és következtetéseink elsősorban tájékoztató jellegűek.

Az entrópiabecslés módszerét nem kívánjuk matematikai szabatossággal tárgyalni, csupán lényegét akarjuk szemléletesen ismertetni. Állításaink azonban matematikai precízséggel bizonyíthatók (1ásd: [N77]).

A) Entrópiabecslés az első feladatlap alapján. - Lényegében bizonyítottnak tekinthetjük. hogy értelmes magyar nyelvű szöveget rekonstruálni lehet abban az esetben, ha a magánhangzókat és a szóközt nyomtalanul elhagyjuk. Ily módon általában mintegy a felére rövidíthetünk. Ugyanis ismeretes, hogy az írott magyar nyelv írásjegyeinek 0,331 része magánhangzó, 0,124 része szóköz, 0.035 része pedig írásjel. A kihagyott írásjegyek száma tehát átlagosan a szövegben szereplő összes írásjegy számának 49%. Ezeknek az adatoknak csakúgy, mint a dolgozatban szereplő további magyar nyelvre vonatkozó gyakorisági adatoknak a forrása [NSz79] A dolgozatban szereplő gyakoriságokat 25 véletlenszerűen választott folyóiratcikk-részletet tartalmazó 51 500 karakterből álló szöveg alapján állapították meg.

A magyar nyelv entrópiájára az előbbiekben adódó felső becslés nagyon durva lenne. Arról van szó, hogy az eredetileg használt 31 karakter (30 betű + szóköz) helyett hiányos szövegünk összesen 21 karaktert (egyjegyű mássalhangzót) tartalmazhat, s ezek előfordulási gyakorisága sem egyenletes. A továbbiakban a *H(. . .)* jelölést használjuk a zárójelben álló kifejezés entrópiájának jelölésére.

Jól ismert, hogy európai nyelvekben az írásjegyek statisztikai függősége 100 betűn túl már elhanyagolható. Így n > 400 esetén a

$$H(\text{írott magyar nyelv}) \approx \frac{H(n \text{ karakterből álló szöveg})}{n}$$

közelítés elfogadható pontosságú lesz.

A rekonstruálhatóság lényegében kölcsönösen egyértelmű kapcsolatot tételez fel az eredeti szöveg és az abból csak a mássalhangzók megtartásával keletkező hiányos szöveg között, ezért

$$H(\text{eredeti szöveg}) \approx H(\text{hiányos szöveg});$$

továbbá a fenti adatok szerint a hiányos szöveg átlagosan $0{,}51 \cdot n$ mássalhangzót tartalmaz, ha az eredeti n írásjegyből állt. Ez azt jelenti, hogy

$$H(\text{magyar nyelv}) \approx 0{,}51\, H(\text{mássalhangzókra redukált magyar nyelv}).$$

Nemetz-Szilléry említett cikkükben [NSz79] vizsgálták az összefüggő szövegből csak mássalhangzók megtartásával keletkező betűsorozat különböző rendű entrópiáit. Speciálisan a harmadrendű entrópiára nyert becslésünk alapján a

$$H(\text{mássalhangzókra redukált magyar nyelv}) \leq 3{,}23$$

adódik, így az előzőek szerint

$$H(\text{írott magyar nyelv}) \leq 1{,}65\,.$$

Ez a becslés azt jelenti, hogy optimális kódolás esetén 31 betűs *ABC* használatával a magyar nyelvű szövegek legalább harmad részükre rövidíthetőek.

B) Entrópiabecslés a második feladatlap alapján. - A II/l. szövegtípus melyben a betűk 40%-át hagytuk el, csak előkészítő jellegű volt. Azt a célt szolgálta, hogy a kísérleti személyek megszokják az ilyen jellegű hiányos szövegek rekonstruálását. A második és a harmadik szövegmintából a betűk felét hagytuk el. Jóllehet az I/3. típusnál is az írásjegyeknek kb. felét hagytuk el, mégis más entrópiabecslést kell most nyernünk, hiszen most a felhasználásra kerülő betűk számát nem csökkentettük, és a megmaradt betűk közötti statisztikai függőség is jellegében tér el a korábbitól. A statisztikai függőség fajtája megkülönbözteti a II/2. és a II/3. szövegmintákat is. A II/2. típusnál a szöveg egy betűjének és a rákövetkező második betűnek a függősége a döntő, míg a II/3. típusnál egyforma súllyal kell számításba venni a szomszédos betűk közötti és az egymástól három távolságra levő betűk közötti függőséget. A $H(i)$ jelölést használva az

egymástól *i* távolságra lévő betűk átlagos feltételes entrópiájára a II/2. típusú szövegek rekonstruálhatósága a

$$H(\text{magyar nyelv}) \le \frac{1}{2} H(2) \approx 2{,}05\,;$$

míg a II/3. típusú szövegek rekonstruálhatósága a

$$\text{H(magyar nyelv)} \le \frac{1}{2} \cdot \frac{H(1) + H(2)}{2} \approx 1{,}98$$

felső becslést eredményezi.

Ezek a becslések az előző pontban ismertetett módon nyerhetők. Tekintettel arra, hogy a korábbinál nagyobb korlátot eredményeznek, a módszer részletezésétől eltekintünk. Érdekes jelenség, hogy az I/3. feladatot a kísérletben résztvevők könnyebben és jobban tudták megoldani, mint a II/3. feladatot, jóllehet az utóbbi számottevően kisebb bizonytalanságot tartalmaz.

C) Entrópiabecslés véletlenszerű kihagyás esetén. - A. III-V. feladatlapokat az jellemzi, hogy a kihagyandó betűk helyét véletlenszerűen, sorsolás útján határoztuk meg. A szövegek rekonstruálása során a rekonstruáló információt nyer arra nézve is, hogy mi volt a sorsolásunk eredménye, így nemcsak a szöveget rekonstruálja, - hanem a szövegtő1 teljesen független véletlen számsorozatra nézve fennálló bizonytalanságot is jelentős mértékben csökkenteni képes.

Konkrét példaként vizsgáljuk azt az esetet, amikor a szöveget ötbetűs csoportokra tördeltük, s ebből véletlenszerűen hagytunk ki kettőt. A két kihagyandó betű helyének megválasztására 10 lehetőségünk van, tehát egy ilyen sorsolás eredménye $\log_2 10$ bizonytalanságot tartalmaz. Ha egy ötös csoport minden betűje különböző, akkor a rekonstruálás során ez a bizonytalanság teljesen megszűnik, különben nem feltétlenül egyértelműen állapítható meg a sorsolás eredménye. A sorsolás eredményére vonatkozó információt mutatja a 8.2. táblázat.

Betűcsoport hossza	A különböző betűk számának eloszlása a csoportban	Az adott típus gyakorisága	A visszaállítás során nyert információ	Átlagos információ
1	1-1-1-1-1	0,641	$\log_2 10 = 3{,}322$	H(5 ; 2)= 3,11
	2-1-1-1	0,317	$\log_2 7 = 2{,}807$	
	2-2-1	0,031	$\log_2 5 = 3{,}324$	
	3-1-1	0,010	$\log_2 4 = 2{,}000$	
	3-2	0,001	$\log_2 3 = 1{,}585$	
	4-1	0,000	$\log_2 2 = 1{,}000$	
	5 azonos	0,000	0	
4	1-1-1-1	0,841	$\log_2 6 = 2{,}585$	H(4 ; 2)= 2,53
	2-1-1	0,177	$\log_2 4 = 2{,}000$	
	2-2	0,002	$\log_2 3 = 1{,}585$	
	3-1	0,000	$\log_2 2 = 1{,}000$	
	4 azonos	0,000	0	
3	1-1-1	0,888	$\log_2 3 = 1{,}585$	H(3 ; 2)= 1,52
	2-1	0,112	$\log_2 2 = 1{,}000$	
	3 azonos	0,000	0	
2	1-1	0,980	$\log_2 2 = 1{,}000$	H(2 ; 1)= 0,98
	2 azonos	0,000	0	

8.2. táblázat Entrópia-becslések

A táblázatban szereplő gyakoriságok Fekete Gyula „A kisbíró" című novellájának 5-5 ezer betűs részletei alapján adódtak.

Ennek alapján 5 betűs csoportokból két betű elhagyása esetén a rekonstruálás során átlagosan $H_{5,2} = 3.113$ bit/5 betű információt nyerünk.

A III/3. feladatnál ez azt jelenti, hogy

$$H(\text{magyar nyelv}) \le \frac{3}{5} H(\text{III/3. típusú szöveg}) - \frac{1}{5} H_{5,.2}$$

Némi számolás után belátható, hogy

$$H(\textit{III/3. típusú szöveg}) \le$$

$$\le \frac{1}{2} \cdot \frac{56H(1) + 54H(2) - 45H(3) - 30H(4) - 10H(5) - 4H(6) - H(7)}{100}$$

A Nemetz – Szilléri - féle dolgozatban [NSz79] közölt $H(i)$ értékek alapján adódik a

$$H\,(\text{magyar nyelv}) \le 1.81 \quad \text{becslés.}$$

Hasonló meggondolások alapján a IV/3. típusú szövegminták rekonstruálhatósága esetén:

$$H(\text{magyar nyelv}) \le \frac{5}{9}\, H(\text{IV/3. típusú szöveg}) - -\frac{1}{9}\left(H_{5,2} - H_{4,2}\right)$$

Ugyancsak némi számolás után adódik a

$$H(\text{IV/3 típusú szöveg}) \le$$

$$\le \frac{1}{2} \cdot \frac{39H(1) + 36H(2) + 27H(3) + 17H(4) + 4H(5) + H(6)}{60}$$

egyenlőtlenség.

Konkrét számokat behelyettesítve: $H(\text{magyar nyelv}) \le 1{,}61$.

.Az V/3. típusú szövegminták rekonstruálhatósága esetén hasonlóan nyerhető a legjobb becslés:

$$H(\text{magyar nyelv}) \le 1{,}42\,.$$

8.4. Záró megjegyzések

A jelen dolgozatban különböző módon "csonkított" írott magyar nyelvű szövegek rekonstruálhatóságára vonatkozó kísérleteinket ismertettük. Bár eredményeink (és így következtetéseink is) elsősorban tájékoztató jellegűek, mégis elfogadhatónak látszik a vizsgált hiányos szövegtípusok rekonstruálhatóságának feltételezése. Ezen feltétel mellett felső becsléseket nyertünk az írott magyar nyelv entrópiájára.

Becsléseink a "csonkítás" típusától függően 1,42 és 2,05 között váltakoznak. Az 1,42 becslés jól szemlélteti az entrópia vizsgálatának fontosságát. Ez a becslés azt jelenti, hogy az írásjegykészlet (az ABC írásjegyeinek) legjobb felhasználása esetén írott magyar nyelvű szövegek terjedelme kb. negyedrészükre lenne csökkenthető.

Összefoglalás- Summary

A dolgozatban foglaltak kidolgozása során alapvető célunk volt a modern valószínűségszámítási, statisztikai fogalmak iskolai órán tanítható anyagának kialakítása. A témakörre irányítják a tanulók figyelmét, rámutatva a matematikai statisztika által nyújtott megoldásokra. Tapasztalataink szerint a tanulók önmaguk is képesek szükséges statisztikai vizsgálatok kezdeményezésére, és hajlandók ilyen vizsgálatokat folytatni.

Az általam képviselt szemléletmód összhangban van a nemzetközileg elfogadott véleménnyel [F90], amely szerint: a leendő tanárok képzésének szükségszerűen tartalmaznia kell olyan anyagokat, amelyek a témakörbe vezető intuitív háttért szemléletesen, a tanításban megvalósítható módon mutatják be.

Ugyancsak nemzetközi megegyezés van abban, hogy a valószínűségszámítás és a statisztika tanításának párhuzamosan kell megvalósulnia.[M90]

A valószínűségszámítás és statisztika alkalmazásához szükséges szemléletmód kialakításához elengedhetetlen, hogy alkalmas elméleti modelleket válasszunk, és azokat az oktatásra alkalmas formában tálaljuk.

A tanári gyakorlatot a valószínűségszámítás és statisztika tanítását tekintve csendes ellenállás jellemzi, amelynek alapvető oka a biztonságérzet hiánya. A tanár nincs felkészítve a lehetséges eltérések megmagyarázására. Ezért rendkívül fontos számbavételük, rendszeres összegyűjtésük, magyarázatuk. Nagy szükségük van tehát egy magyarázatokkal ellátott oktatási anyag összeállítására. Dolgozatom elkészítéséhez egy ilyen anyag összeállítása volt a fő motivációs erő.

A második fejezetben a Fej-Irás sorozatokban előforduló leghosszabb futamokat (legtöbb egymást követő Fej illetve Írás száma egy n hosszú dobássorozatban) vizsgálom. Bebizonyítom, hogy **a leghosszabb futam hossza $\log(n)$ körüli.**

A harmadik fejezet hiányos sorozatok rekonstrukciójával a statisztikai szemléletformálást szolgálja, míg a negyedik fejezetben egyszerű helyettesítéses titkosírást akarunk betűstatisztika segítségével megfejteni.

A transzpozíciós titkosító eljárás fejtése hipotézisvizsgálatok láncolataként fogható fel, melyhez bigram-statisztikát készítünk és használunk fel az ötödik fejezetben.

A hatodik fejezetben vezetjük be az alapvető valószínűségszámítási fogalmakat, például az eseményalgebra fogalmait és összefüggéseit; néhány egyszerű eloszlást (klasszikus, binomiális – és geometriai eloszlás), és lényeges esetként vizsgálunk egy csak statisztikai modellből becsülhető eloszlást.

Az összefüggő és független megfigyelések vizsgálatánál az oktatási cél éppen az, hogy a gyakorlatban tipikus, összefüggő véletlen jelenségeket modellezzünk bármilyen típusú 9-12-es csoportban.

A hetedik fejezetben a reprezentatív mintavételről esik szó, és elméleti számítás helyett szimulációs eljárást mutatunk be a hipergeometriai eloszlás egy ismeretlen paramétere értékének becslésére, ha a négy paraméter közül hármat ismerünk.

Az utolsó fejezetben megmutatom, hogy miként adható felső becslés bizonyos szabályok szerint csonkított szövegtípusok rekonstruálhatóságának felhasználásával az értelmes írott magyar nyelvű szövegek entrópiájára.

A disszertáció a szerző következő publikációin alapul:

[S78]; [S88]; [S89]; [S02]; [S04].

A 2.4. és 2.5. szakaszok eredményei még nincsenek publikálva.

Summary

Writing this thesis, one of the major goals was to develop and introduce the methods for teaching the concepts of probability theories and statistics at schools, drawing the attention of the students to the subject and pointing out the solutions offered by mathematical statistics. Our experience shows that students are independently capable of initiating necessary statistic tests, and are willing to carry them out. Teachers should be prepared for the possible differences. This is why it is highly important to collect, systemize and explain them. To solve this task is also among the goals of the thesis.

My approach is in line with the internationally accepted opinion [F90], which states that the education of future teachers should necessarily include materials, which introduce the intuitive background of the topic in an expressive and teachable way.

It is also internationally agreed that statistics and probability theories should be taught parallel.[M90]

It is indispensable in creating the necessary approach to the application of statistics and probability theory, to choose appropriate theoretic models, and transform them to suitable for education.

In Hungary silent resistance characterizes the practice of teaching statistics and probability theories, which may be caused by the lack of confidence. Teachers are not prepared for the explanation of possible differences; they also need teaching materials, which contain teaching notes with explanation. This need was one of the major motivations for preparing the present study.

In the second chapter of the thesis, I examine the longest run in head-tail series (the most heads or tails following each other in an n-long series). I prove that **the length of the longest run is around log(n).**

The third chapter is intended to help the development of statistical approach, by the reconstruction of incomplete series. In the fourth chapter, we try to solve simple substitutional cryptograms by using letter frequencies.

In the fifth chapter I try to solve the decoding of the transpositional secret codes. Method can be regarded as a chain of hypothesis testing, for which bigram-statistics are used.

In the sixth chapter the basic concepts of probability theories are introduced, for example the concepts and relationships of algebra of events,

some simple distributions (classical probability field, binomial and geometrical distributions) and I examine the important case where only purely statistical approach works.

An emphasis is given to the study and modeling of dependent and independent events at secondary level.

The sevens chapter deals with representative sampling. Instead of complicated theoretical calculations we use a simple simulation for determining an unknown parameter of the hyper geometrical distribution in the knowledge of the other three parameters.

In the last chapter I suggest a method for estimating the entropy of the written Hungarian. This method is based on the reconstruction of purposefully truncated written texts.

Theses are based on the following publications of mine:

[S78]; [S88]; [S89]; [S02]; [S04].

The results of the 2.4. and the 2.5. chapters are not yet published .

IRODALOMJEGYZÉK

[A98] Ambrus A. (1998): „Bevezetés a matematika didaktikába” ELTE TTK egyetemi jegyzet

[B82] Barnett, V. (1982): "Teaching Statistics in Schools throughout the World", ISI TOTSAS, Voorburg

[BBT01] J. Bennett, W. Briggs, M. Triola: (2001): Statistical reasonong for everyday life Addison Wesley Longman

[BN77] Bognár,K. – Nemetz T. (1977): On the teaching of probability at secondary level, Educational Studies in Mathematics, 1977, 399-404

[BNT79] Bognár, J., Nemetz, T. Tusnády, G. (1979): Ismerkedés a véletlennel. Középiskolai szakköri füzet. Tankönyvkiadó, Budapest

[CN94] Cakiroglu E., Nemetz, T. (1994): "Secret Codes in Statistical Education". In: L.Brunelli, Cichitelli (ed): Proceedings of the first Scientific Meeting of IASE, Perugia, pp. 367-376.

[F68] Feller, W. (1968): Bevezetés a valószínűségelméletbe és alkalmazásaiba.

[F72] Freudenthal, H. (1972): The Empirical Law of Large Numbers or the Stability of Frequencies in: Educational Studies

[F73] Freudenthal, H.(1973): „Mathematics as an Educational Task” Reidel: Dordrecht

[F90] Fischbein, E. (1990): "Training Teachers for Teaching Statistics" In: Hawkins, A. (ed.): Training Teachers to Teach Statistics, ISI, Voorburg, 1990, pp. 48-57

[FS62]Frick F. C.-Sumby W. H. (1962),: Controll tower language. Journ. Acoust. Soc. Amer. XXIV 595-596.

[GV75] Glayman, T., Varga T., (1975): Zwischen Unmöglich und Sicher, Herder: Freiburg-Wien von Harten,G. - Steinbring,H. 1984]: Stochastik in der Sekundarstufe I, IDM 8 Band, Aulis Verlag Deubner Co.: Köln

[HNP92] Hajnal I.-Nemetz T.-Pintér L. (1992): Matematika III. Gimnázium, Fakultatív B változat, Tankönyvkiadó, 5. kiadás

[HNPU94] Hajnal I.-Nemetz T.-Pintér L.-Urbán János (1994): Matematika IV. Gimnázium, Fakultatív B változat, Tankönyvkiadó, 5. kiadás

[H89] Hawkins, A. (1989), The annual United Kingdom Statistical Prize, in: Morris, M. (editor), The teaching of statistics, Studies in Mathematics Education, Volume 7, Unesco

[K98] Kaeding, F. W. (1898), Haufigkeitswörterbuch der deutschen Sprache, Berlin

[K87] Knuth, D.E. (1987): "A számítógépes programozás művészete 2. Szeminumerikus algoritmusok". Műszaki Könyvkiadó, Budapest.

[KN92] Kusolitsch, N., Nemetz, T. (1992): "Übungen zum subjektiven Zugang zu Wahrscheinlichkeiten", Stochastik in der Schule, S. 42-47.

[MG99] Mignon, H.S.-D. Gamermann, D.(999): Statistical Inference An integrated approach, Arnold: London-New York-Sydney-Auckland

[M90] Moore, D. (1990): Uncertainty. In: L.A.Steen (ed): On the shoulder of Giants, National Academy Press, Washington, D.C., pp. 95-137.

[M73] Moles, A. A. (1973) Információelmélet és esztétikai élmény. Gondolat Budapest

[N78] Nemetz, T. (1978): Valószínűségszámítás, gimnáziumi tankönyv, Tankönyvkiadó, Budapest.

[N70] Nemetz, (1970): Egy lehetőség valószínűségszámítási fogalmak bevezető szemléltetésére, A matematika Tanítása, 143-149 old.

[N77] Nemetz, T(1977): Etropy Estimation Via Reconstruction of Mutilated Texts: Colloques Intemationaux du CNRS. No. 276. Theorie de l'Information. Developpements Récents et Applications. Cachan. 1977. 389-397.

[N82] Nemetz, T. (1982): "Pre-University Stochastics Teaching in Hungary", n: V. Barnett, ibid. cit., Ch. 4. pp. 85-112.

[N92] Nemetz, T. (1992): "Cryptology: a rich source of applications offering entertaining mathematical instructions." In: de Lange at al. (Eds): Teaching Mathematical Modelling and Applications. Ellis Vorwood, New York. 230--244.
[N97] Nemetz, T. (1997): An overview of the teaching probability in secondary schools. Papers on Statistical Education presented at ICME-8, Seville, July 14-21, 1996. Ed.: B. Phillips. Swisburne University Press, 1997, 75-86.
[N97] Nemetz, T. (1997): "Milyen matematikát rejt Verne postagalambja", Módszertani Lapok, Matematika, 3.évfolyam, 3. szám.
[N98] Nemetz, T. (1998): Írott nyelvek redundanciája, Uj alaplap, XVI/8
[N00] Nemetz, T. (2000): Véletlenszámok generálása és használata, Élet és Tudomány, Diákoldal, LV. 47.szám, (november 24)
[NSz79] Nemetz,T. – Szilléry,A. (1979): Nyelvstatisztikai táblázatok, Alkalmazott Matematikai Lapok, 69-87.
[NW99] Nemetz, T. – Wintsche, G. (1999): Valószínűség és statisztika mindenkinek, Polygon, Szeged, 1999
[P67] Petőfi S. János: (1967) A nyelvstatisztikai vizsgálatok néhány kérdése Nyelvfeldolgozás és dokumentáció. Szerk. Szépe György: A tudományos tájékoztatás elmélete és gyakorlata 11. sz. 117-140
[PV96] W. Peschek- Vancsó Ö., (1996): Sztochasztika tanítás: Centrális tevékenységek, lokális jelentések, globális célok, OKSZI Módszertani Lapok 2. Évf., 3-4. szám 1-16, 1-15.
[R78] Révay, Z. (1978): "Titkosírások". Zrínyi Katonai Kiadó, Budapest.
[S49] Shannon, C.E. (1949): "Communication theory of secrecy systems". Bell Syst. Techn. J., 28, 656-715.
[S51] Shannon, C.E.(1951): "Prediction and Entropy of Printed English", Bell Syst. Techn. J., 30, 50-64
[S45] Smith, O.J.(1845). The Secret Corresponding Vocabulary; Adapted for Use to Morse's Electro-Magnetic Telegraph. Thruston.
[S75] Sz. Simon Judit (1975): Self-information and optimal codes (Nemetz Tiborral), Colloquia Math. Soc. János Bolyai 16.: Topics in information Theory, Keszthely, 457-468.
[S78] Sz. Simon Judit (1978): Estimating the Entropy of Written Hungarian by Gambling Technique, Trans. Of the 8th Prague Conf. On Info. Theory, Vol B.
[S86] Sz. Simon Judit: Számítógép a matematikaórán – két budapesti gimnáziumban. Móricz Zsigmond Gimnázium, Köznevelés, XLII évf. 12. sz. 1986.03.21. 26-27
[S88] Sz. Simon Judit (1988): Az írott magyar nyelv entrópiája, Intenzív tanár-továbbképzés – matematika, ELTE TTK
[S89] Sz. Simon Judit (1989): Hiányos szövegek rekonstruálhatósága és a magyar nyelv entrópiája, Magyar Nyelv, LXXXV évfolyam, 427-438 (Nemetz T-ral közösen).
[S00] Sz. Simon Judit (2000): Felvételi feladatok matematikából (a matematika munkaközösséggel) Fazekas könyvek, Fazekas Mihály Gimnázium
[S02] Sz. Simon Judit (2002): Überzeugen statt beweisen -- der zentrale Grenzverteilungs-satz im Mittelschuleunterricht, Stochastik in der Schule, N. Kusolitsch társszerzővel
[S04] Sz. Simon Judit (2004): Statistical Inference in Schools. : "Teaching of Mathematics and Computer Sciences".
[S05] Sz. Simon Judit (2005): Matematikai feladatgyűjtemény (Gerőcs – Orosz – Paróczai – Sz. Simon) Előkészületben, Nemzeti Tankönyvkiadó Vállalat
[S64] Svékus Olivér (1964): Titkosírások. Móra Könyvkiadó, Budapest
[T96] Tóth Béla (1896) Mendemondák Bp.
[T73] Tusnády Gáborné (1973): A valószínűségszámítás középiskolai tanítása, A matematika tanítása, 10, 6-13

[V92] Vancsó, Ö. (1992): Valószínűségszámítás és statisztika a középiskolában, 1992. évi Varga Tamás Napok előadásai, Műszaki Kiadó, 1993 63-73
[V87] Varga, T. (1987): "New Maths since 1963: Its Implementation as a Historical Process". In: I. Wirszup (ed). Developments in School Mathematics Around the World, NCTM, Washington, D.C. 1987

Mellékletek

Táblázatok jegyzéke

Ábrák, grafikonok

Függelék.

A VIII. fejezet kísérletében felhasznált hiányos szövegek.

I. Feladatlap. Forrás: a Műszaki Élet 1973. november 9-i száma.
I/1. feladat („Budapest hőellátása"). A szöveg eleje:
A+MEGLEVŐ+FŰTŐERŐMŰVEK+KÖRZETÉN+KÍVÜL+TERVEZETT+.
Folytatás:

– J – L – K – T – L – P – K – – LL – T – S – R – – T – B B
N Y – R – – G – Z T – Z – L – S – – F – R R – V – Z K – Z –
N – S – F – T – – R – M – V – K – T – – P – T – T T – K – –
M – – V – S Z – N Y L – G – K – D V – Z – – B – R – H – Z –
S – – K – L T S – G G – L – B – Z T – S – T J – – – Z – – J
– L – K – S – K – K – R S Z – R – – H – – L L – T – S – T –
– – N – M Z – T K – Z – – G Y – K – R L – T – – G Y – N – K
K – R – – G Y – R T – L M – – N – – Z T – M – T – T J – – H
– G Y – – – M – G Y – R – R S Z – G – H – Z – H – S – N L –
– K L – M – T – K – S – V – S Z – N Y – K – K – Z – T T – K
– L L – – N – N – G Y – H – – G – N Y – K – – S – T – N – –
– V – R – S – K – H – – L L – T – S – T – F – T – – R – M –
V – K B – L – C – L S Z – R – – M – G – L D – N – –

I/2. feladat („Bondor József válaszából"). A szöveg eleje: *A + MINISZTÉRIUM + A + VÁROSRENDEZÉSI + TERVEK +*. Folytatás:

SZKSZ	*RSGNK*	*+ BZTS*	*TSR + L*	*LTV + L*	*LNRZS*	*R + MGS*
ZRVZT	*+ + VRS*	*RNDZS*	*+ TRVT*	*NCSKT*	*+ + GYK*	*RLT + B*
BZNYT	*TT + HG*	*Y + + VR*	*SRNDZ*	*S + TRV*	*K + MGF*	*LL + SZ*
NVNLN	*K + BZT*	*STS + N*	*M + LGS*	*GS + MR*	*T + + MG*	*VLSTS*
G+MG+	*GN + SK*	*+ LKLM*	*+ DDK +*	*+ TRVK*	*+ MGVL*	*TZTTS*
R+S++	*VRSRN*	*DZS + T*	*RKVSK*	*+ LKRC*	*SSTSR*	*+ NNK +*
+ FLSM	*RSNK +*	*+ HTSR*	*+ JBB +*	*NTZKD*	*SKT + H*	*ZTNK +*

I/3. feladat („A szén az energiaellátásban"). A szöveg eleje: *A + LEGUTOBBI + IDÖBEN + NEMCSAK + A + HAZAI + HANEM + A + NEMZETKÖZI +*. Folytatás:

SJTBN	*SGYRT*	*BBLYN*	*CKKJL*	*NKMGM*	*LYHSS*	*ZBBDG*
LBCSL	*TSZNF*	*NTSSG*	*RHVJF*	*LFGYL*	*MTMVL*	*HZNKN*
RGHRD	*ZKBNV*	*SZNYL*	*GSZGN*	*YRSZG*	*NKNKS*	*MGKLL*
VZSGL	*NNKHG*	*YSZNK*	*SZLTN*	*KMLYN*	*MRTKS*	*TMKHS*
ZNLSL	*TSZKK*	*VNTSN*	*KMGYR*	*SZNKN*	*CSFTR*	*TKVSZ*
NYLGC	*SKLYH*	*MTRTL*	*MNGYG*	*YBNYS*	*ZTSML*	*GTRML*

KNY

II. Feladatlap. Forrás: 1., 2. feladat: a Műszaki Élet fenti száma, 3. feladat: a Magyar Nemzet 1977. október 9-ig száma.

II/1. feladat („A védjegy"). A szöveg eleje: *A + SZAKEMBEREKEN + KIVÜL + A + NAGYKÖZÖNSÉG + ÉRDEKLÖDÉSÉRE + IS +*. Folytatás:

S – Á – O T – T – R T – A – + A – R – K L – M – K I – K – N
Y – T – R + – O – O Z – T – E L – Ö – K Ö – E – E + – + – É
D – E – Y + – + – É Z – K – N Y – + – + V – D – E G – Y – L
+ – A – C S – L – T O – + – A L – M – N N – I – K É – D – S
T – R – S Z – E – E S – N – T A – L – L J – + – Y A – O – L
A – I – T A – Á – S O – K – L + – Á – J A – E – + A – + – R
D – K – Ö D – K – T + – A – Z N – S – S E – É – K Ö – Y – +
M – N – + A – V – D J – G – E T – M – R + – L – A L – A – O
+ – Á – L A – A – O K – A – + M – N – – + P – D – G + – Z – K
N – K – A K – K – M O – T – I S – E – K E – N – K + – + – É
D – E – Y + – O – A L – Á – A L – J – L E – T – S É – É – E
L – É – + A – K – Z E – J – V Ö – E – + S – E – E T – É – E
K – M – G U – + – S + – É – J E – Y – T + – E – E Z – T – I +

II/2. feladat („Az MTESZ megalakulásának előzményei"). A szöveg eleje: *A – TUDOMÁNYOS – ÉS + MÉRNÖKI + SZERVEZETEKET + MINDENÜTT + AHOL –*. Folytatás:

I – Y – N – I – T – Z – É – Y – K – V – N – A – + – + – A –
D – S – G – É – + – + – Á – S – D – L – M – I – É – Y – I –
H – Z – Á – + – É – R – + – + – E – L – D – S – E – Y – B –
Z – N – O – + – O – Á – + – + – I – E – Y – S – L – T – I –
K – E – Y – R – S – E – I – + – G – N – H – S – Z – + – R –
D – É – Y – K – E – + – A – D – G – M – L – R – + – E – I –
T – V – S – Z – + – G – E – Ü – E – E – N – + – S – A – O –
+ – Z – V – T – É – E – S – K – R – K – N – V – N – S – + –
U – A – + – Á – + – R – Z – G – K – T – D – M – N – O – + –
S – M – R – Ö – S – E – V – Z – T – I – E – + – Z – K – A –
+ – É – J – I – É – + – U – K – M – D – Z – R – I – T – K –
N – E – É – E – + – S – A – + – T – S – + – É – I – + – L –
P – E – Ö – N – S – E – I – L – S – A – I – T – Z – É – Y –

II/3. feladat („Csoportérdek"). A szöveg eleje: *A + VÉDÖÜGYVÉD + AZZAL + ÉRVELT + HOGY + VÁLLALAT +*. Folytatás:

V E – – T Ö – – E K – – A Z – – S Á – – + T – – V É – – E I
– – + M – – S É – – É S – – Ö L – – M I – – L + – – D O – –
Á K – – K E – – S E – – I F – – E + – – E M – – Y E – – E L
– – Y Ü – – N E – – S Z – – M A – – T T – – R R – – M I – –
+ E – – H I – – + K – – Ü L – – N Y – – + H – – T A – – E L
– – + F – – Y E – – E T – – S A – – A + – L L – – A T – –
K O – – E K – – V Á – – K + – – A R – – K + – – T + – – + Ü
– – M + – – R N – – É N – – + M – – Ö R – – S É – – T + – –
T E – – E Z – – T + – – E R – – É G – – S + – – V Ö – – L E
– – B I – – O S – – Á S – – R T – – M I – – + M – – K A – –
Ö + – – G T – – T Á – – N A – – E G – – K + – – L T – – E L
– – L É – – E K – – + T – – T O – – + Ü – – A + – – T Á – –
S + – – U G – – S Z – – V + – – V E – – S + – – L T – – Z +

III. Feladatlap. Forrás: Népszabadság 1977. május 24.

III/1. feladat („Közös érdekből"). A szöveg eleje: *EZ + IS + SZÉP + PÉLDÁJA + A + TÁRSADALMI + ÖSSZEFOGÁSNAK + A +*. Folytatás:

T – – Á C S – K – É S – A – + – Z E – E K + – G – Ü – T – Ü
– Ö D É – – – E K + A – E – Y + – D D – G + – R – Z – G – Z
E R – – + S – – + S I K – – + – O R R – S A – V – – T – N E
M – S A – – A + – E L E – – L É – – E J L E – – T É S – – N
– – A N E M + – Á – + T – R – L – T E – E N – – S + E – – I
T – E – J Ü K + – – – D Á – L + A + – – P – S Ö – – T E – +
M U – K – – E Z D – – T – B U – A P – S T E – – É S + T – –
B + – Á – – V Á – O S B A – – A + T – – Á – S – K + K – Z –
E M – N – E – É S – R E – – Z + – – E M E – + Ö – – Z – B E
– Z É – T – K + E – – H – T E N E – + – + – – G G E L I – –
S + – – D É – – T Á N I – – S – C S – O – G – L O M + – – I
G – E L – I – E L – E T – – Z S U – – L – S Á G – N – R – G
I + – – N – + – O L T + A – – I – + H – – – + C S – P Á N –

III/2. feladat („A mezőgazdaság fejlesztése Pest megyében"). A szöveg eleje: *A – TERMELÉS – FEJLESZTÉSE + MÁS + TERÜLETEN + IS + TERVSZERŰ +* Folytatás:

TDA	+PÁR	THT	RAT	TAL++	GÉH	EZ+R	ÜTK
+A+KÁ	BBN+	+FÖR	OSH	+TAO	Z+SZ	K+EZ	NK+A
+TSZ	NK+U	YIS	+FOO	TSA	+FOY	TT++F	LJÜ
+A+ÖV	ÁRS+	ERS	KDÉ	SEL	+RÁN	B+ED	G+TÖ
B+VL	T+UA	PSI	+TS+G	ESÜ	+A+FÖ	DDL+	DEL

EŐ+P	ST+E	GIE	KL+E	+LHE	TV+T	ES+A	Z+ES
ÉES	BB+O	IIK	I+S+G	DAÁ	I+EJ	LEZ	É+A+S
VET	EZT	E+MN	ÁÁN	A+JO	B+ÖS	EHG	OLÁ
T++ER	MES+	THN	KA+				

III/3. feladat („A jubileumi munkaversenyről"). A szöveg eleje: *IDŐ + ANYAG + ENERGIA + A + TERMELÉS + ALAPVETŐ + FELTÉTELEI +*. Folytatás:

KÖZ	+ET	+HÁ	ATÁ	LLO	TTK	UNK	ERE
NFE	LAJ	SAI	KÖZ	PPN	TJA	+KI	SMT
ÉS+	ÉGY	ÁDO	LGZ	IA+	ÁAL	ATA	ELA
+KZ	TJÁ	GYÁ	TKN	AK+	ZIT	MOR	ÉS+
KSE	ELÉ	YEK	TTO	BÁ+	EÉB	KSZ	ÜLÉ
ETÉ	+AL	ARÉ	EET	AKÁ	SA+	ABU	AES
IÁL	LAT	K+A	YRÉ	Z+M	KAE	RHÁ	NNY
KÜZ	+ZÉ	T+A	ROT	KUG	Y+G	YR+	OLO
OIA	BRI	DEZ	TIT	ANC	KÁS	+A+	MSA
KGA	ZAS	GI+	FER	NIÁ	MJD		

IV. Feladatlap. Forrás: Magyar Nemzet 1976. szeptember 4.

IV/1. feladat („Az idei tanév feladatai és gondjai"). A szövegek eleje: *MA + AZ + ISKOLÁKBAN + MÁR + 120 + EZER + NEVELŐ + MŰKÖDIK + AKIK +*. Folytatás:

F E – A – – – A I K A – + – + T – – Á B – I – K – A –
+ C – A – – A – K O R + – – T – – T – – K + E L – M –
G – – L E – ü E – + H – + – I – E – E S – – ü L K – –
Z I – É – Ü – E – + – S – B I – – O S I – A – – + – E
H – T – É S + – – N K – + – – Z – E – – S E M + – A –
Y – – K + – A – U K R – – ü K E – – E Z – R – + – + –
U L – + – A – – V B E N – – O G – – – – – O Z O T – – A
Z – O K – – – Á S I + – – N I – – T – R I U – – – – + P
E D – – O – – S K É – Z É – S E – – – S + T – – – B B
– É – Z É – – E L + – – N E K + – – G Y – – E – + – A
V A L – – É S – – – + I D – – + A – – Y O – C + – A –
I T O – – – – ü + I N – É – E – – F – – S K O L – V –
+ – – A K U – – + I – L – T V E – – L A – – L + A – –

IV/2. feladat (5. lap: „A tokaji bor világhírének megőrzéséért"). A szöveg eleje: *A + HEGYALJAI + SZŐLŐTERMESZTÉS + FEJLESZTÉSÉRE + IGÉNYBE + VETTÉK +*. Folytatás:

A+IS	L+S	ÁTA	AU	HEL	+ÖV	ATO	+AL
+KÉ+F	LÉ+	+VSZ	OY	G+SI	+YN	EVZ	T+S
KNY	+TE	TEK	T+S	++TOV	BE	JES	ÉS
+MID	N+E	BE+Ő	SE	HNO	LK	++MUN	AE
ELZ	TE	+EZK	E+A+	RÜE	TK	+ZÁR	+EL
GÜ+A	YÜ	EI+S	ÖL	BLÁ	T+A	AKO	AK
+I+AZ	AL++	ÉLA	+GY	+A+MJ	+RM	ŐR+F	UL
O+SL	+ME	GFE	Ö+N	YAI+	ÁZ	T+DA	N+A
+HGO	LL	+MGS	AB+	SZI	+ÉG	REH	ND
O+ZL	RK	ONR	UK	(*vége*: OHOZ+)			

IV/3. feladat („Szolgálat — díjtalanul"). A szöveg eleje: *MERT – SAJNOS – EZ + A + HELYZET + OTT – AHOL + VALÓBAN + TEHETNÉNEK +*. Folytatás:

VAL	IA	+AO	SS	+SL	GA	TÁA	+É	AGY	OS
+A+	ZK	EN+	VY	HOG	Y+	TAÁ	BA	MOD	+S
OKT	K+	HÁN	SÁ	G+A	+N	HNY	PÉ	+AZ	EI
K+G	YÜ	L+K	IR	TÁB	+F	ÜNÖ	TÁ	+HI	DE
++N	LU	+VÁ	ÁL	+ZI	NE	TLE	IO	+DJ	AL
NUL	HO	Z+S	LI	JUU	GA	EZ+	+Á	IK+	RA
ATB	+Y	+HG	IK	A+M	VÁ	ROL	TA	OSO	YS
ZSI	IK	KKT	DI	ALN	U+	ZHO	+S	LTJ	U+

V. Feladatlap. Forrás: Esti Hírlap 1977. március 29.

V/1. feladat („. . . . Igaz ügyért, igaz eszközökkel harcoltam"). A szöveg eleje: *A + TÁRLAT + MELYNEK + ELSÖ + RÉSZE + KÉT + ÉVVEL + EZELÖTT +*. Folytatás:

N Y – – T – M – G – A + – A – – A – – M U – K – S – – Z
G – – O – + T – R – – N E T É – – – – – M U N – – S – U –
T U – A – – E – L – D – S – – – + V Á Z – – – – + – – O L
+ – + – O R – – L + – – – – U R – L – – D – + O – Z – Á –
Y O K – – – – – I – I K – J – – O L – A – + O – – Z – – G V
– – Z – J T Ö – – É – – – L – S R – – + A – – O R – + –
+ F – – É – T – – R O – + D – – – – – E N T – – A – – N Y
I T – – K – – E G – – + – – Z D A – Á – – + – Á L – – G
+ – D Ö – Z – – – T + A – N – – E T – – E – – Z Á – – Á
S – – Y O M – S – – O – H – – A P – A – – + – A J – + –
Z + O – – Z – – – – – S + B U – – – E – T + F – – S Z – U
– – I – – – Á T + – – É – – K + F – L – É R Z – – – E –
E – E N – – + – – – – D E Z Ö – –

V/2. feladat („Város az üveg alatt"). A szöveg eleje: *TAVALY + JUNIUSBAN + LÁTTUNK + AZ + ALAPOZÁSHOZ + AZ + ÜVEGEZÉS +*. Folytatás:

R+MEE	ZDÖ+S	+AMN+M	NNE++É	ITÖ+A+	VNYÁ
L+MEN	ULA++P	LÁTÁ	+H+OTE	RBN+Z+	+NEM+L
NE+H+Ö+	+HIG+I	+A+TÉ++E	M+TNÁ	K+TEI+	ÖNRM
NE+ID	+ÁUSA	+MÉI+K	ZDJ+A+	AÁTÁ	ST+A+E
S+ÉIZ	ÉS+A+V+	GR+BE	JEDK+	IAN++F	NY+VL
OK+TR	ZK+BD	ALS+S	+ZENE	RE+KÖ	+G+KIO
ÉER+O	N+NÖN	YÁZR	S+PL++A	TKTR	S+ÖVN+

V/3. feladat („Ma folytatódnak a szovjet—amerikai . . ."). A szöveg eleje: *A + PRAVDA + MAI + SZÁMÁBAN + AZ + ELSÖ + OLDALON + SZÁMOLT + BE + A +*. Folytatás:

SZV+	ARIA	+ÁGY	ÁK+E	KEDR	LLEN
I+BS	NYVE	AELÖ	ÜÉEV	ALOS	TELR
L+A+	NGYD	OLDÁ	+MMN	RNÉK	L+ÖÖ
+ ATN	A+ÉT	N+TR	ALSK	O+SL	ELET
SVAA	NTAI	LÁRG	IN+H	ANGT	ESÉK
KIOT	+A+T	YAÁK	MEGE	ÉSÉE	LS+A
LNAD	T+IA	+TÖ+	SZKA	KZPN	+LPI