I. Leíró statisztika

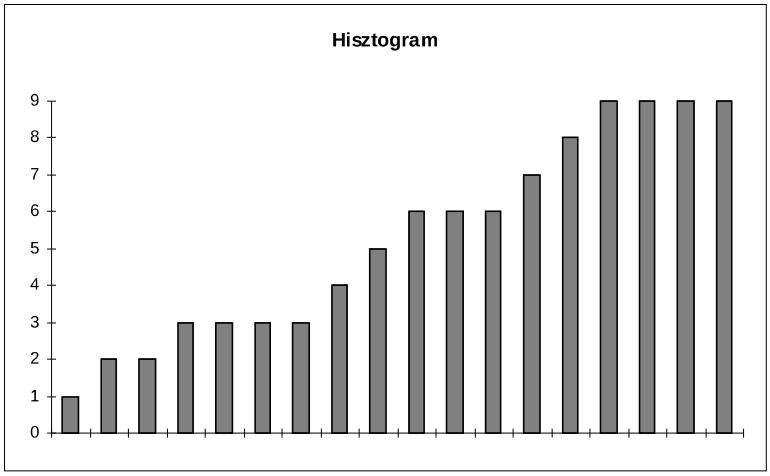
A leíró statisztika azzal foglalkozik, hogy egy adott, meghatározott elemekből álló információhalmazt kiértékeljen. Ezek az információk legtöbbször persze számokat jelentenek, hiszen ezek a matematikai módszerekkel legkönnyebben kezelhető objektumok, azonban nem kell szigorúan ezekhez ragaszkodnunk, olyan adathalmazokat is kiértékelhetünk, melyek nem számokból állnak. Természetesen a kiértékelés mikéntjét tekintve ez utóbbi esetben a lehetőségeink korlátozottabbak.

I.1. Az adatok grafikus ábrázolása

Az adatok kiértékelése legegyszerűbben valamilyen grafikus formában történő megjelenítéssel történhet. A rajzok nagyon sokfélék lehetnek, itt most a gyakrabban használt típusokat mutatjuk be. Az adatok ábrázolásánál legtöbb esetben az adathalmazban való előfordulási arányt szokták ábrázolni (relatív gyakoriság). Ennek nagyságát egy adott adat esetén úgy kaphatjuk meg, hogy az adat előfordulási számát osztjuk az adathalmazban levő elemek számával. Ezt a relatív gyakoriságot meg lehet adni százalékos formában is.

Ezen kívül ha sok adat van, és ezek esetleg mind különbözőek, de minket az adatok nagysága csak bizonyos pontossággal érdekel, akkor szokás az adatokat „adatsávokba”, osztályokba osztani, tehát nem az egyes értékeket vesszük figyelembe, hanem csak bizonyos, pl. 10-es vagy 100-as pontossággal ábrázoljuk őket. Ekkor az egyes adatsávokban (osztályokban) található elemek számát jelenítjük meg.

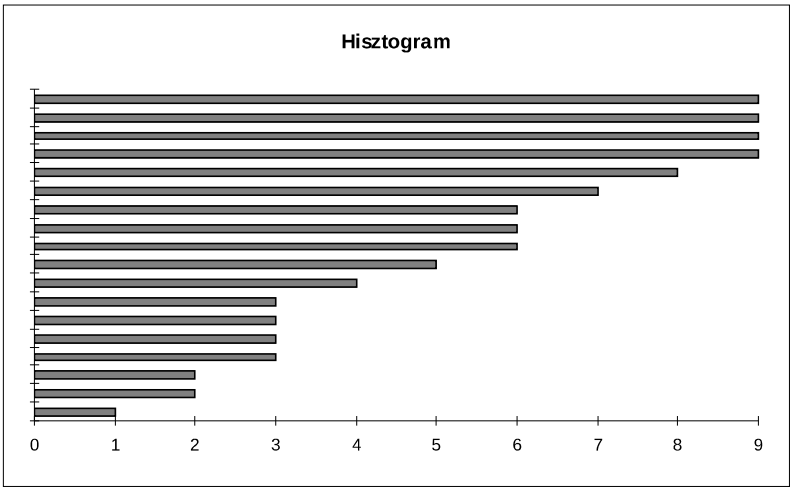
I.1.1. Oszlopdiagram (Hisztogram)



Ebben az ábrázolási módban az adatokat mint kis pálcikákat jelenítjük meg. A pálcikák magassága arányos az adat nagyságával. (A negatív adatokat lehet lefelé rajzolni.)

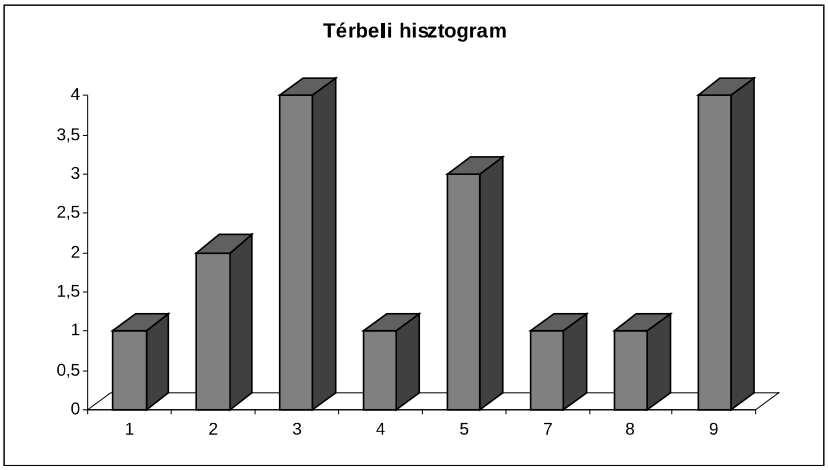
Nagyon gyakran előfordul, hogy nem magukat az adatokat ábrázolják a hisztogramon, hanem a gyakoriságukat. Ebben az esetben az oszlopok alatt fel kell tüntetni, hogy melyik adathoz tartozó relatív gyakoriságot mutatják.

A hisztogramok esetében alkalmazható vízszintes oszlop-elhelyezkedés is.



Ennek hátránya, hogy kevésbé követhető az adatok egymáshoz való viszonya.

Hisztogramok esetében lehet térbeli ábrákat is készíteni:



Ebben az esetben a térbeli ábrázolás teljesen felesleges, hiszen semmivel sem mond többet, mint a síkbeli ábra, sőt, a térhatás eléréséhez alkalmazott technika kissé torzítja az arányokat.

A bevezetőben mondottaknak megfelelően az adatok ábrázolása nem csak a relatív gyakoriság alapján, hanem adatsávokba osztva, az egyes tartományokba eső adatok számának ábrázolásával is történhet.

I.1.2. Szár-levél diagram

Százezresek	Tízezresek
0	6 4 5 8 8
1	3 0 0 1 6 8 3 3 6 4 2 5 7 5
2	4 3 2 2 1 6
3	1 2 1 1 6
4	7 8 4 5 2 0 3
5	1 1 2 2
6	3 5 7 8 9 9 9 9 9 2 3
7	1 2 4 2 3
8	4 7 8 2 3 4 5 6
9	2 3 9 9 1 2

A szár-levél diagram lényege, hogy az adatsávoknak megfelelő tartományokra osztva ábrázoljuk az adatokat, megtartva azok számértékét, és ezeket egy sajátos táblázatban tüntetjük fel. A baloldalon pl. a százezreseket jelöljük a táblázatban, a jobboldalra pedig azon adatok tízezreseit írjuk, amelyekben az adott százezres szerepel.

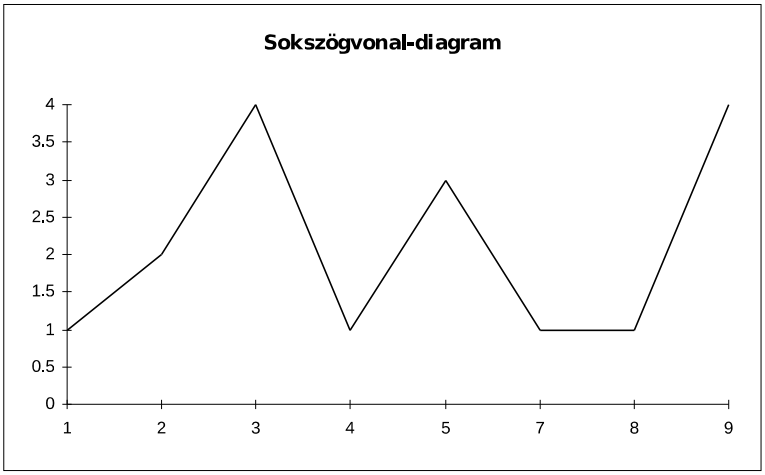
A fenti táblázat első sorában pl. a 62456, 42312, 51897, 82451, 89236 adatok vannak ábrázolva. Ezekben a százezres helyiértéken 0 áll, a tízezres helyiértéken pedig 6, 4, 5, 8, 8, amint az a táblázatból is leolvasható. A táblázat negyedik sorában pl. a 315682, 326897, 318642, 315259, 368970 adatok vannak ábrázolva. Ezekben a százezres helyiértéken 3 áll, a tízezres helyiértéken pedig 1, 2, 1, 1, 6.

A szár-levél diagram 90 fokkal elforgatva hisztogramot ad, természetesen itt az oszlopok magasságát a számsorok hosszúsága fogja adni. Azonban a szár-levél diagramnak a hisztogrammal szemben több előnye is van. Egyrészt nem veszik el az egyes adatok számszerűségében hordozott információ és az adatok eredeti sorrendjét is (egy-egy adatsávon belül) meg lehet tartani, másrészt pedig a hisztogram készítésekor az ábrázolás finomítása (tehát az adatsávok szűkítése) sokkal bonyolultabb. Itt egyszerűen csak az adatokat kell átrendezni annak megfelelően, hogy milyen új adatsávokat akarunk létrehozni. Például a fenti szár-levél diagram 50000-es finomítása a következő:

Százezresek	Tízezresek
0	4 5
	6 8 8
1	3 0 0 1 3 3 4 2
	6 8 6 5 7 5
2	4 3 2 2 1
	6
3	1 2 1 1
	6
4	0 3 4 2
	7 8 5
5	1 1 2 2
6	3 2 3
	5 7 8 9 9 9 9 9
7	1 2 4 2 3
8	4 2 3 4
	7 8 5 6
9	2 3 1 2
	9 9

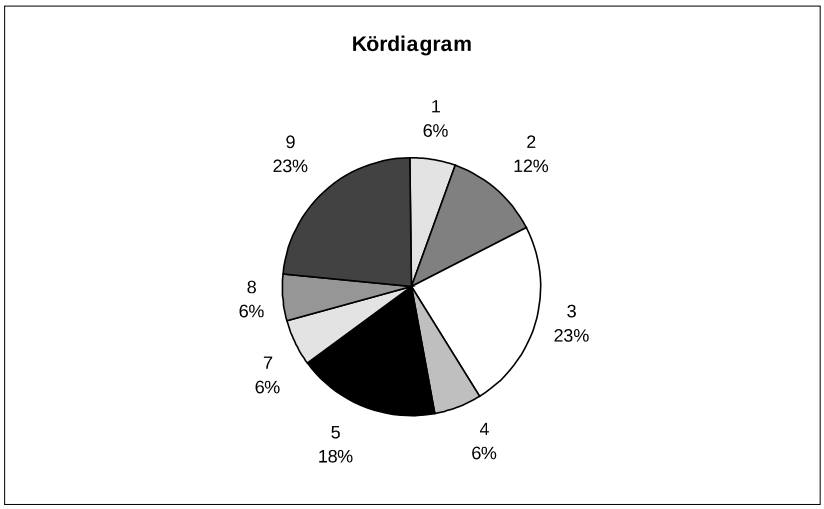
Természetesen a szár-levél diagramnak is megvannak a hátrányai, például ha nagyon sok adatból áll az adathalmaz, akkor ebben az ábrázolási módban nagyon sokat kell írni.

I.1.3. Sokszögvonal diagram (Vonaldiagram)



A sokszögvonalas ábrázolási mód hasonlít az oszlopdiagramos módszerhez. Koordináta-rendszerben ábrázoljuk az adatsávoknak és a benne levő adatok számának (illetve az adatok nagyságának) megfelelő pontokat, majd ezeket egy töröttvonallal összekötjük. Ennek az ábrázolási módnak az az előnye, hogy kiemeli a változások mértékét, mert az összekötő vonalak meredeksége dominál a rajzban, éppen ezért ezt valamely adat változásának szemléltetésére használják leginkább. Hátránya viszont az, hogy a vonalak folytonossága valamiféle folytonosságérzetet kelt a szemlélőben, tehát azt gondolhatja, hogy a változás folyamatos volt, pl. egy kisebb értékről egy nagyobb értékre folyamatos növekedéssel jutottunk el. Ez lehet csalóka, hiszen ha pl. minden második évben szerzett adatokat ábrázolunk, akkor a közbenső években az előző adatsor növekedésétől függetlenül lehet relatív csökkenés a korábbi évek eredményéhez képest.

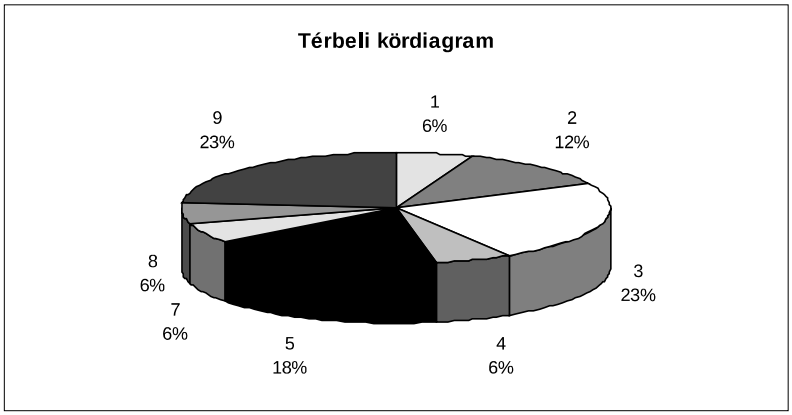
I.1.4. Kördiagram



A kördiagramot általában az adatok relatív gyakoriságának ábrázolására használják. A teljes kör jelképezi a 100%-ot, és az egyes adatok relatív gyakoriságát ábrázoló körcikkhez tartozó középponti szög arányos a relatív gyakorisággal.

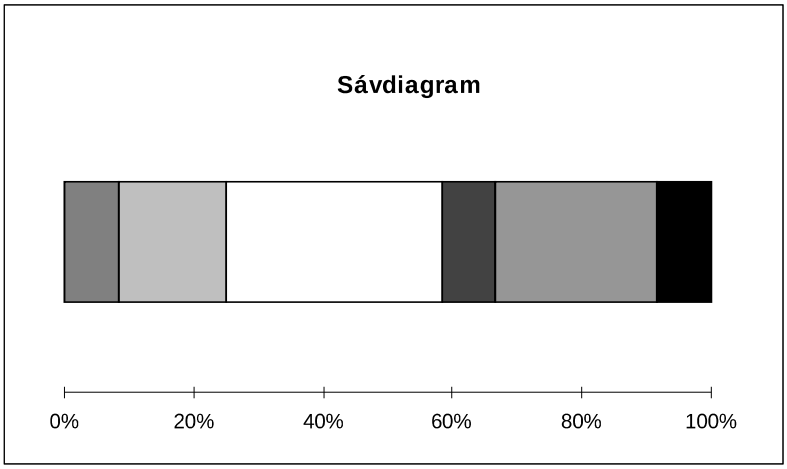
Ezen a kördiagramon az előző adathalmazban szereplő adatok relatív gyakoriságát jelenítettük meg. Az egyes értékek előfordulási gyakoriságát is szokás feltüntetni. A kördiagram előnye, hogy a rész és egész, valamint az egyes részek egymáshoz való viszonya jól látható, viszont ha nem tüntetjük fel a szeletek mellett az előfordulási arányokat, akkor nehéz pontosan megbecsülni az egyes adatok nagyságát.

Kördiagramok esetén szoktak térbeli ábrát is készíteni.



Ez ugyan látványos, és szeretik az egyszerű emberek, de a perspektíva miatt nagyon nagy mértékben eltorzíthatja az arányokat, ráadásul szintén a perspektíva miatt manipulációra ad lehetőséget.

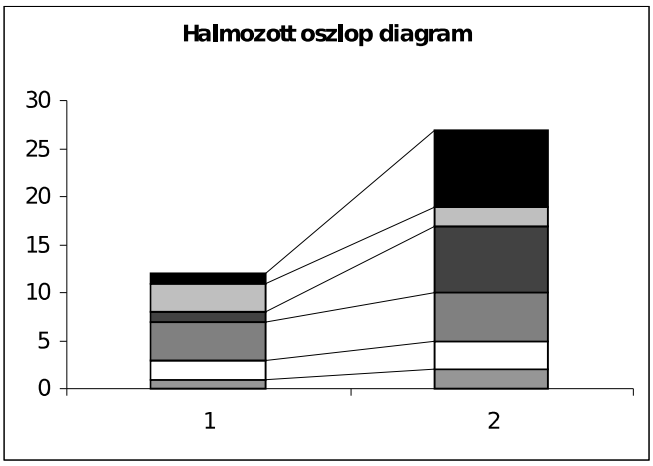
I.1.5. Sávdiagram



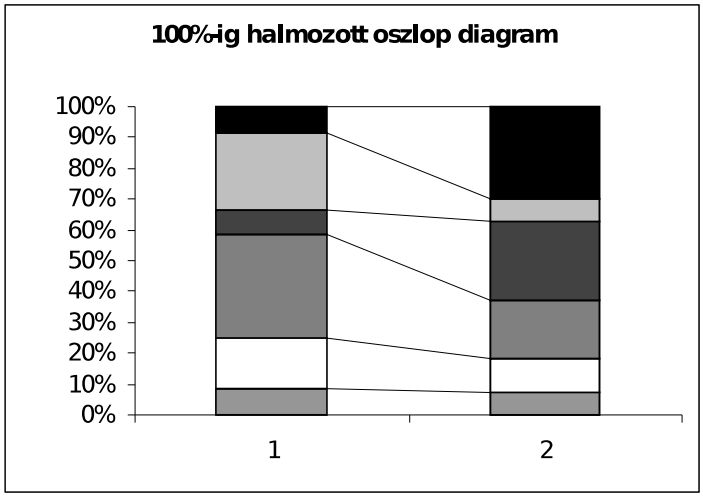
A sávdiagramban az adatok relatív gyakoriságát egy sávon ábrázoljuk. A megjelenítésnél az egyes részsávok hosszúsága arányos a megjelenített adat relatív gyakoriságának nagyságával.

A sávdiagram előnye, hogy a rész és egész viszonya jól látható, azonban az egyes részek egymáshoz való viszonya nem igazán szemléletes.

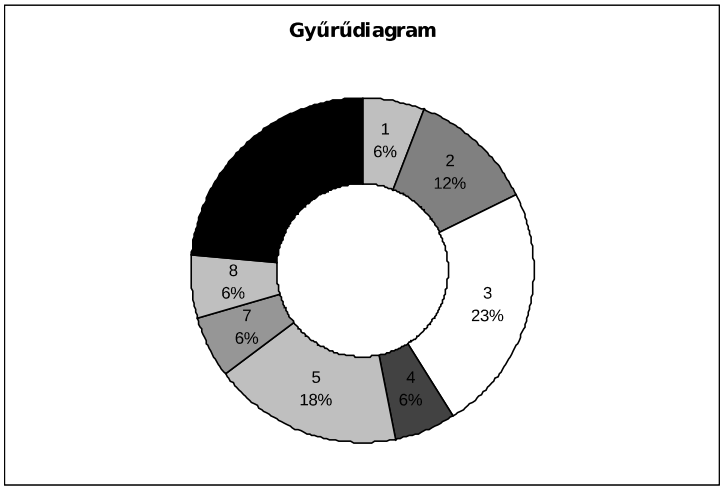
Létezik összehasonlító változatban is, amikor az egyes adatok változását egymás mellé helyezett, függőleges helyzetű sávdiagramokon (úgynevezett halmozott oszlopdiagramokon) szemléltetik. Ha az oszlopok magassága arányos az összmennyiséggel, akkor a szemléletesség torzul: sem az nem látható jól, hogy az egyes mennyiségek aránya az egészhez hogyan változott, sem az, hogy a mennyiségek abszolút nagysága hogyan változott.



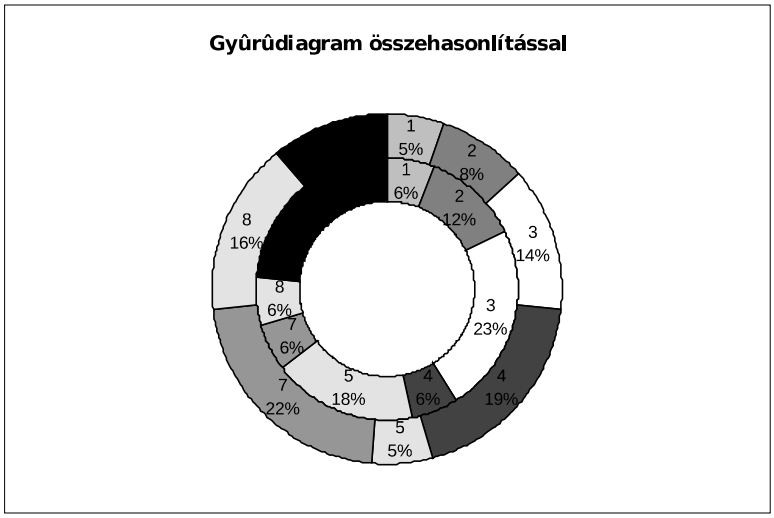
Az oszlopok magassága lehet egyenlő (100%-ig halmozott oszlop), ebben az esetben a százalékos arány változása nyomon követhető.



6. Gyűrűdiagram

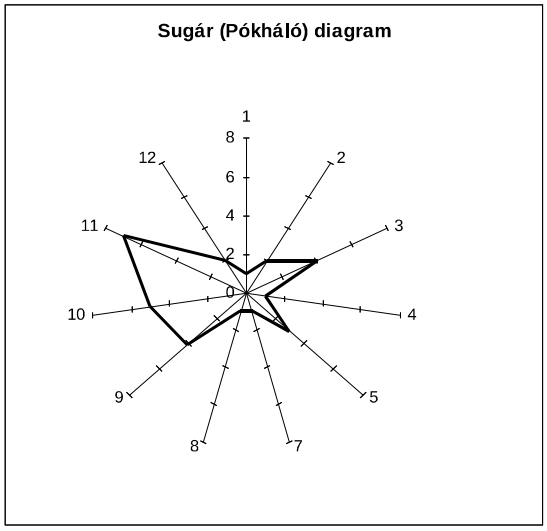


A gyűrűdiagram tulajdonképpen a kördiagram egy részlete, egy körgyűrű-sávot levágunk a kördiagramból. Akkor szokás használni, ha több adathalmaz összehasonlítását akarjuk megtenni, és az egyes gyűrűkben ábrázolhatjuk a különböző adathalmazokat.



Ennél az összehasonlításnál nagyon zavaró lehet az, hogy az arányok változása miatt a a két adathalmaz egy típusú adatait ábrázoló szeletkék elcsúszhatnak egymás mellől, és ez nehezen teszi láthatóvá a változások nagyságrendjét.

7. Sugár (Pókháló) diagram



Ezen a diagram típuson az adatok ábrázolása úgy történik, hogy ahány adat szerepel az adathalmazban, annyi egy pontból kiinduló félegyenest veszünk fel, melyek közül a szomszédosok egyforma szöget zárnak be. Minden egyes adatot a neki megfelelő félegyenesen ábrázolunk, és utána a kapott pontokat egy törött vonallal összekötjük.

Szintén az adatok változásának szemléltetésére alkalmas, de kézzel elkészíteni kissé nehézkes.

Az adatok változásának nagysága az egyenesek meredekségéből olvasható le: minél jobban az origó felé tart egy szakasz, annál jobban csökken az adat nagysága, és fordítva.

I.2. Az adatok időbeli változásának megjelenítése

Az adatok időbeli változását az egyszerű adatábrázolások segítségével is nyomon lehet követni. Sokkal inkább alkalmas azonban az úgynevezett bázisidőszakhoz viszonyított ábrázolásmód. Ez lehet kétféle: a bázisidőszak mindig az előző időegység, ekkor az adatokat a megelőző időegységhez képesti, %-ban kifejezett nagyságával ábrázoljuk. A második lehetőség: a bázisidőszak mindig egy rögzített időszak, és így minden adatot ehhez képesti, %-ban kifejezett nagyságával jelenítünk meg. A következő fejezetben látunk erre példát.

I.3. Manipulációs lehetőségek az adatok grafikus megjelenítésével

Az adatok grafikus megjelenítésekor az adatsor ábrázolójának nagy lehetősége van manipulatív módszerek kiválasztására: pusztán a megjelenítés során sugallni tud valamit az adatsorról. Szokták

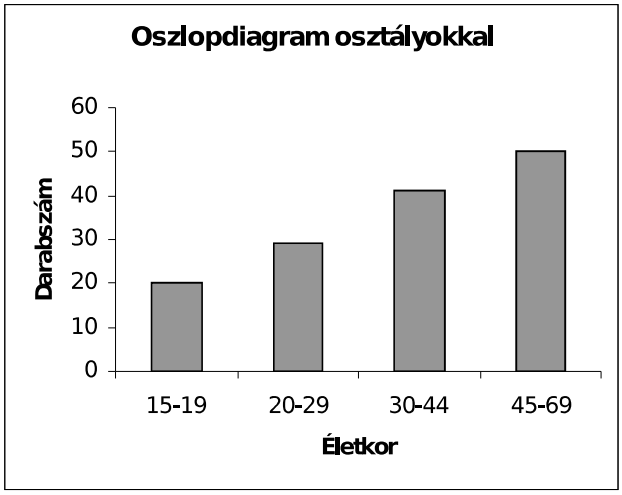
mondani: statisztikai adatokkal minden be lehet bizonyítani, és az ellenkezőjét is. Nézzünk erre néhány példát!

1. példa: A politika és az életkor

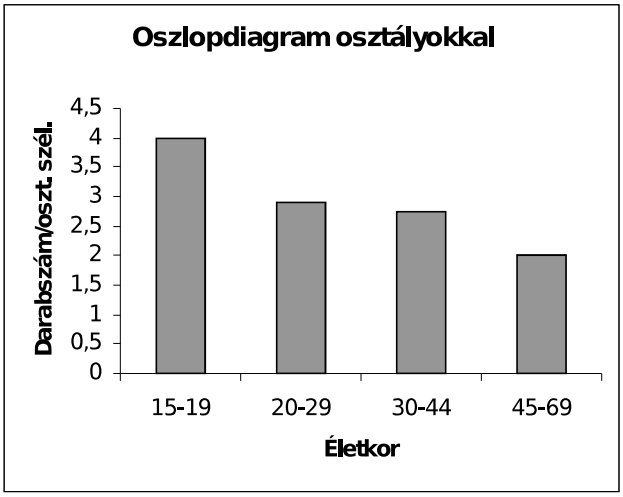
Egy vidéki városban tartott politikai rendezvényre 140 ember ment el. A résztvevők életkorát nagyság szerint közzétették (a jobb követhetőség érdekében összesítve közöljük, hogy az egyes életkorú emberekből hány volt jelen a rendezvényen, a zárójel előtti szám az életkor, a zárójelben álló szám a létszám):

15 (2), 16 (3), 17 (4), 18 (5), 19 (6), 20 (6), 21 (5), 22 (4), 23 (3), 24 (2), 25 (3), 26 (3), 27 (2), 28 (1), 29 (0), 30 (1), 31 (0), 32 (1), 33 (1), 34 (0), 35 (1), 36 (0), 37 (1), 38 (2), 39 (4), 40 (4), 41 (5), 42 (10), 43 (5), 44 (6), 45 (5), 46 (6), 47 (3), 48 (4), 49 (4), 50 (3), 51 (0), 52 (4), 53 (2), 54 (3), 55 (0), 56 (2), 57 (1), 58 (2), 59 (1), 60 (2), 61 (1), 62 (0), 63 (0), 64 (1), 65 (1), 66 (0), 67 (2), 68 (2), 69 (1)

Az első statisztikus azt az eredményt kapta, hogy a fiatalokat kevésbé érdekli a politika, és az időseket a legjobban. Az osztályokba sorolás alapján elkészítette az életkor szerinti részvételi létszám oszlopdiagramját:

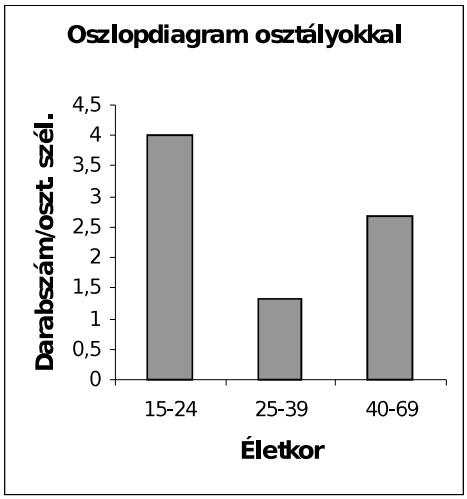


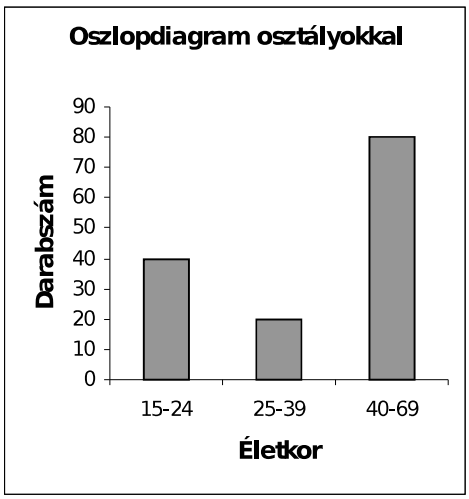
A második statisztikus fejét csóválva azt mondta: Nem jó, hiszen az osztályok nem egyforma életkori létszámról szólnak. Tehát figyelembe kell vennünk, hogy egy osztály hány évet ölel fel, és ezzel el kell osztanunk az adatokat. Így az első osztály létszámát 5-tel, a másodikét 10-zel, a harmadikét 15-tel, a negyedikét 25-tel osztjuk. A kapott értékeket ábrázoljuk oszlopdiagramon:



Ebből éppen az jött ki, hogy a fiatalokat érdekli a legjobban a politika, és az időseket legkevésbé.

A harmadik statisztikus azt mondta: Egyiknek sincs igaza, hiszen a grafikonokból származó kétféle, egymásnak ellentmondó eredmény azt mutatja, hogy rossz az osztálybasorolás. Olyan osztályokat kell keresni, ahol mindkét féle grafikonból ugyanazt az eredményt kapjuk. Mutatott is egy példát:





Azt ugyan nem lehet eldönteni a grafikonok alapján, hogy melyik korosztályt érdekli legjobban a politika, de az biztos, hogy a középkorúakat a legkevésbé, hiszen mindkét grafikon ezt alátámasztja.

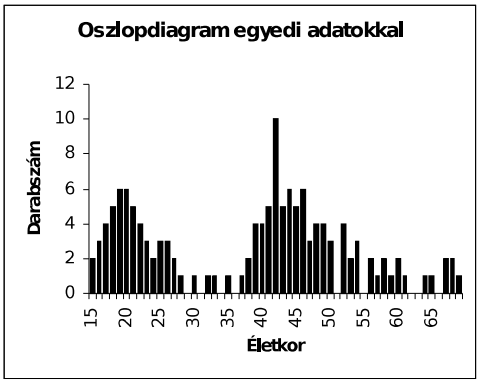
A negyedik statisztikus azt mondta: Ne önállóan nézzük az adatokat, hanem próbáljuk meg a város lakosságához viszonyítani. El is kérte a nyilvántartásból a lakossági létszámokat, és a következőket kapta (az első esetben figyelembe vett osztálybasorolással dolgozott):

Életkor	**15-19**	**20-29**	**30-44**	**45-69**

Lakosok összes száma	359	518	735	894
Rendezvényen részt vett	20	29	41	50
Lakosok számához viszonyított %-os arány	5,57%	5,6%	5,58%	5,59%

Ebből viszont látszik, hogy érdeklődésben nincs jelentős különbség a korosztályok között.

Megjegyzésként hozzáfűzzük a feladathoz, hogy az adatok osztálybasorolás nélkül is ábrázolhatóak, érdemes elgondolkodni, hogy ebből milyen következtetés vonható le:

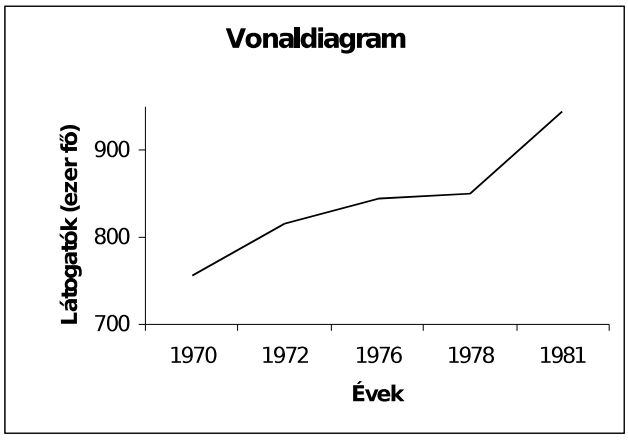


2. példa: A városi önkormányzat és a fejlesztés

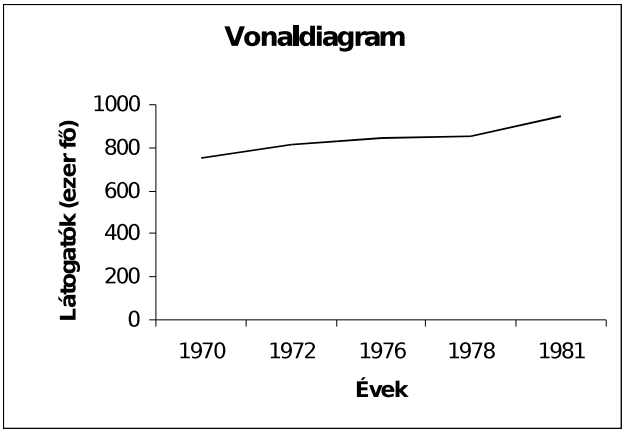
Egy népszerű fürdőhely fürdőjének látogatottsági adatait tartalmazza az alábbi táblázat:

Év	**1970**	**1972**	**1976**	**1978**	**1981**
Látogatószám	755 000	815 000	845 000	850 000	945 000

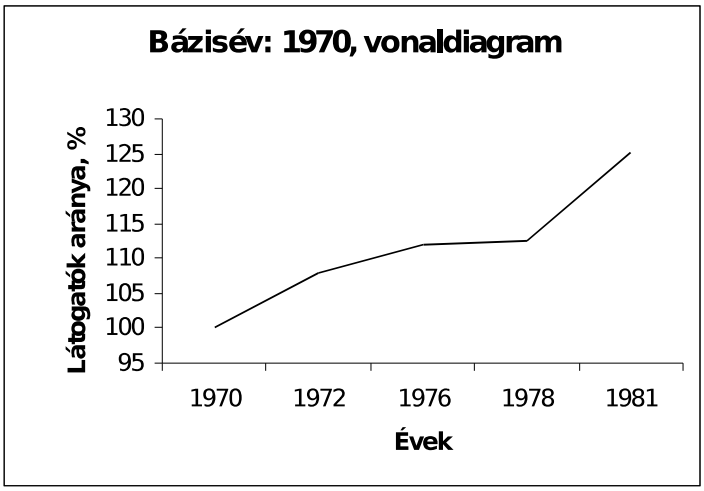
A városi tanácsban az ellenzék képviselője felszólal: Botrányos, hogy miközben ugrásszerűen nő a látogatók száma, a város vezetése nem tesz semmit a komoly fejlesztések érdekében. Bizonyítékként az alábbi grafikont mutatja be:



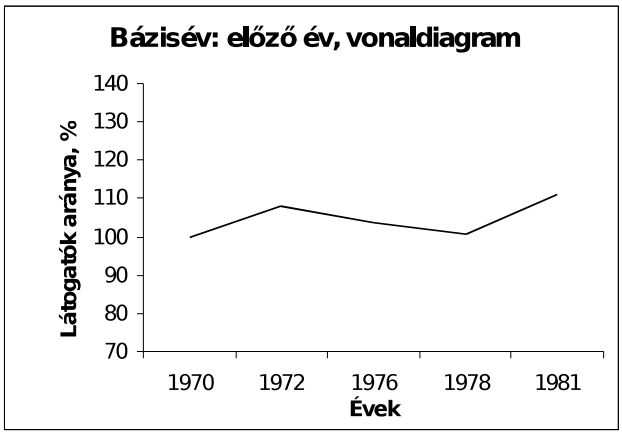
Erre a város vezetése reagál: Ismerjük az adatokat, de meg kell mondjuk, nem látunk semmiféle ugrásszerű változást, egy megfontolt, lassú növekedés érzékelhető, ami nem teszi indokolttá a nagyarányú fejlesztéseket.



Az ellenzék képviselője ismét szót kap: Nézzük meg, hogy 1970-hez képest milyen nagyarányú növekedés volt tapasztalható a látogatók létszámában %-ban kifejezve:



Erre a városi vezetés képviselője válaszol ismét: Nézzük meg, hogy az előző évhez képest mekkora növekedés volt az egyes években! Látható, hogy a növekedés nagyon ingadozó, sőt, csökkenő tendenciát is mutatott sokáig, tehát ismételten nem indokolt a fejlesztés.



Megjegyzés: Nagyon jó gyakorlóterep a különböző újságokban megjelenő grafikonok elemzése, rengeteg rossz grafikonnal, és ezek alapján félreértelmezett következtetéssel lehet találkozni a mindennapokban is.

I.4. Középértékek

Sok esetben valami számszerű jellemzőt keresünk, amely valahogy jellemzi a sokaságot. Milyen adatok segítségével tehetjük ezt meg?

Jellemezhetjük a leggyakrabban előforduló elemével, ezt *módusz*nak nevezzük. (Ha több olyan szám van, ami egyforma gyakorisággal fordul elő, akkor ezek a móduszok halmazát alkotják.) Ennek megadása valamit elárul a sokaságról, de ha minden elem csak egyszer-kétszer fordul elő benne, akkor a móduszok halmazának megadásával elég kevés, és viszonylag rosszul kezelhető információhoz jutunk.

Bizonyos sokaságokról valamivel többet mond a sokaság középső értéke (természetesen ez megkívánja, hogy az adatok rendezhetőek legyenek.) Vagyis rendezzük nagyságrendi sorrendbe az adatokat, és válasszuk ki a középső elemet; ha nincs középső elem, mert páros számú adatunk van, akkor a középső kettő számtani közepét vegyük. Az így kapott számot *medián*nak nevezzük. (Azaz ha az adathalmaz $2k+1$ elemből áll, akkor a sorbarendezés után a k-adik elem a medián, ha pedig $2k$ elemből áll, akkor a medián a sorrendbe állított elemek közül a k-adik és $k+1$-edik elem összegének fele.) A medián már egyértelműen meghatározott, de még mindig viszonylag kevés információt hordoz a sokaságról, hiszen az elemek sorának elején és végén a mediántól nagyon különböző elemek is állhatnak.

Megjegyzés: A medián által megadott információt kiegészíti az ún. *alsó és felső kvartilis* értékének megadása. Ezek a mediánhoz hasonló középértékek, de nem a felező értéke az adathalmaznak, hanem az alsó kvartilis a „negyedelő”, a felső kvartilis pedig a „háromnegyedelő” érték.

A medián esetében fellép az a probléma, hogy a sokaság többi tagjának csak sorrendje határozza meg, de a nagyságrendjük nem szerepel benne. Ebből a szempontból még több információt kaphatunk a sokaságról akkor, ha minden benne szereplő számot figyelembe veszünk, tehát a számok összegét osztjuk a darabszámukkal. Az így kapott értéket nevezzük a *sokaság átlagának vagy számtani közepének*. Ez azonban megint csalóka lehet: ha van egy, a többiekhez nagyon nagy vagy nagyon kicsi szám a sokaságban, akkor az adatok jelentős része döntően eltérhet az átlagként kapott adattól.

A fentiekből látható, hogy a fent meghatározott középértékek más-más jellegű információt adnak a sokaságról, de egyik sem kielégítő önmagában. Nézzünk néhány példát ezek alkalmazására!

1/1. példa: Egy osztályban felmérjük azt, hogy a gyerekek közül kinek mi a kedvenc étele. Milyen adattal jellemezhetjük a kapott adathalmazt?

Egyértelmű, hogy ennél a feladatnál csak a módusz jöhet szóba, ugyanis nem számszerű, és nem is rendezhető adatokból álló adathalmazról van szó, így a medián és az átlag nem létezik. A módusz megadása azt jelenti, hogy az osztályban melyik ételt szeretik a legjobban.

Az olyan adathalmazoknál, melyek kvalitatív (minőséget kifejező) és összehasonlíthatatlan adatokból állnak, a jellemzés csak a módusszal történhet.

1/2. példa: Egy munkahelyen összeírjuk mindenkinek az iskolai végzettségét. Arra a kérdésre szeretnénk választ kapni, hogy ez egy magasan kvalifikált emberekből álló hely-e vagy pedig csupa alacsony iskolai végzettségű ember dolgozik itt. Milyen adattal jellemezhetjük a kapott adathalmazt?

Nyilvánvaló, hogy mivel nem számszerű adatokról van szó, az átlag nem kiszámítható, de az adatok rendezhetőek, tehát a medián illetve a módusz is meghatározható. A módusz megadása nem mond el semmit, hiszen lehet, hogy eggyel több nyolc általánost végzett takarító van, mint diplomás vagy doktori végzettségű, vagy érettségizett alkalmazott, akkor azt kapjuk, hogy ez egy eléggé alacsonyan kvalifikált társaság, holott nem az. Alkalmasabb a medián ezen adathalmaz jellemzésére, mert itt azt tudjuk megmondani, hogy a dolgozók középvégzettsége mekkora. Ha ez alacsony, akkor itt nem túl sok magas iskolai végzettségű ember van, ha magas, akkor sokan dolgoznak itt pl. diplomával.

2/1. példa: Felmérjük, hogy az iskolában a tanulók melyik kerületben laknak. Milyen adattal jellemezhetjük a kapott adathalmazt?

A kapott adatok számok, tehát kiszámítható az átlaguk, azonban eléggé egyértelmű, hogy ez itt semmiféle információt nem ad. Nyilván nincs értelme az olyan típusú kijelentéseknek, hogy „A gyerekek iskolánkba átlagosan a 12 és feledik kerületből érkeztek.” Hasonlóan nem sok információt hordoz a medián megadása, és itt a módusz sem olyan nagyon informatív, bár ez a legalkalmasabb a sokaság jellemzésére.

2/2. példa: Egy táborban részt vevő gyerekek iskolai osztályát tudjuk, 7. osztályosoktól 12. osztályosokig voltak jelen. Olyan adatot szeretnénk, melynek segítségével eldönthetjük, hogy a tábor inkább kisebbeknek szólt vagy inkább az idősebb korosztálynak. Milyen adattal jellemezhetjük a kapott adathalmazt?

Ha a móduszt adjuk meg, akkor ez csak abban az esetben ad információt, ha a módusz által meghatározott évfolyamból közel a tábor létszámával megegyező számú gyerek érkezett. A medián eléggé informatív, hiszen a közepe felett és alatt egyforma mennyiségű adat található. Ha a medián nagy, akkor a nagyobbak voltak nagyobb arányban, ha kicsi, akkor a kisebbek (esetleg az alsó és felső kvartilis értékét is hozzá lehet venni). Az átlag lehet nagyon félrevezető, például ha sok 7-9-es van, és elég sok 12-es, akkor az átlag lehet 10 körül, amire azt mondanánk, hogy az életkor nagyon vegyes volt, holott zömmel kisebbek voltak a táborban. Természetesen itt megint problémákba ütközhet az átlagosan 9 egész egyharmadik osztály értelmezése.

A 2/1 példában azt láthattuk, hogy attól, hogy valamilyen adathalmaz számokból áll, még nem biztos, hogy van értelme az átlagot kiszámolni. A 2/2 példában az „átlagosztály” még valamiféleképpen értelmezhető lenne, de mondhatjuk, hogy nem foglalkozunk ilyen jellegű értékkel.

3/1. példa: Egy kis cipőkészítő üzemnek csak arra van lehetősége, hogy egyféle méretű cipőt készítsen. A tulajdonosának milyen cipőméretet kell kiválasztania?

Nyilvánvaló, hogy itt azt a cipőméretet célszerű választani, ami a legtöbbször szerepel az emberek lábméretei között, hiszen ekkor lehet abból az adott méretből a lehető legtöbbet eladni, tehát a móduszát kell meghatározni a cipőméretek halmazának. Az is nyilvánvaló, hogy nem érdemes átlagot számolni, hiszen lehet, hogy nem is egész szám jön ki erre, a medián meghatározása pedig szintén nem ad megfelelő információt, hiszen lehet, hogy az ott lakók egyik felének 38-as, a másik felének 46-os lába van, és egyvalakinek 42-es; ebben az esetben a medián 42-es, de csak egy ember fog ekkora cipőt venni, és ebből nem nagyon lehet megélni.

3/2. példa: Valaki átlagos képesítéssel egy céghez akar menni dolgozni, és szeretné megtudni, hogy várhatóan mennyit fog keresni. Milyen adatot kell kérnie a fizetésekről?

Ha a móduszt kéri, akkor nagyon rosszul is járhat, hiszen lehet, hogy a cégnél elég sok alacsony fizetésű pl. takarító van, és ő ezeknél mindenképpen többet fog kapni. Az átlag megint csak nem lenne jó, hiszen a főnökség magas fizetése nagyon eltorzíthatja a ténylegesen kapható összeg nagyságát. A medián megadása tűnik a legjobb megoldásnak, hiszen ő, mint kezdő annál a cégnél feltehetően a közepes fizetés környékén fog kapni.

Természetesen előfordulhat, hogy ugyanannál az adathalmaznál más-más kérdésfelvetéshez más-más középértéket érdemes megadni. Például a 3/1. példában ha azt kérdeznénk, hogy általában mennyire vannak az emberek jól megfizetve ennél a cégnél, akkor a móduszt kéne megnézni, azaz a leggyakoribb fizetést. Ha viszont az adóhivatal érdeklődik a cégnél kifizetett jövedelmek után, akkor az átlagot kell szolgáltatni, illetve a dolgozói létszámot.

I.5. A középértékek „jóságának” mérőszámai

A fentiekben láthattuk, hogy a leggyakrabban használatos középértékek a módusz, a medián és az átlag. Természetesen az adathalmazt bármilyen más, egyéb módon definiált középértékkel lehet jellemezni, hiszen nagyon sokféle szempont dominálhat ennek megadásában.

Felmerül viszont az a kérdés, hogy egy adott középérték mennyire jellemzi jól az adathalmazt, mennyire nagy az egyes elemektől való eltérése. Ennek megadására újabb mérőszámot vagy mérőszámokat kell bevezetnünk.

Elsőként megadhatjuk az adathalmaz terjedelmét, azaz a legnagyobb és legkisebb elem különbségét. Ha ez kicsi, akkor gyakorlatilag bármelyik középérték jól jellemzi az adathalmazt, ha pedig nagy, akkor nem lehet eldönteni, hogy mi mennyi információt szolgáltat. A terjedelem másik nagy problémája, hogy egy-egy adatra nagyon érzékeny, tehát nagyon nagy lehet, ha van egy kiugró adat a többi között, amely a többihez képest nagyon nagy, vagy nagyon kicsi, holott az adatok lényegében egy szám környékén tömörülhetnek. Ezt szokták úgy kiküszöbölni pl. fizikai kísérletek eredményének kiértékelésekor, hogy a legnagyobb és legkisebb adatot kihagyják az értékelésből, azonban ez nem minden esetben tehető meg. (Természetesen ez a módszer a többi középérték esetén is javítja az értékelés jóságát.)

Vehetnénk azt is, hogy átlagosan mekkora eltérései vannak az adathalmaz elemeinek a megadott középértéktől, ezt nevezhetnénk átlagos eltérésnek. Azaz az átlagos eltérés képlettel megadva: ($\widetilde{X}$ jelöli az adott középértéket)

$$\frac{\left(x_1-\tilde{X}\right)+\left(x_2-\tilde{X}\right)+\ldots+\left(x_n-\tilde{X}\right)}{n}$$

Ennek azonban van egy óriási hátránya: mivel a megadott középértéknél feltehetően vannak nagyobb és kisebb adatok is az adathalmazban, az összegben szerepelnek pozitív és negatív tagok is, ezek viszont összességében eredményezhetnek nagyon kicsi számot, holott ők maguk abszolút értékben lehetnek nagyok. Például könnyen belátható, hogy a fenti kifejezés az átlag esetén mindig 0, függetlenül attól, hogy mik az adathalmaz tagjai.

Ki kell tehát küszöbölni az előjelproblémát az átlagos eltérésből. Erre a legegyszerűbb módszer, ha az eltérések abszolút értékét átlagoljuk, ennek neve átlagos abszolút eltérés. Kiszámítási módja tehát:

$$\frac{\left|x_1-\tilde{X}\right|+\left|x_2-\tilde{X}\right|+\ldots+\left|x_n-\tilde{X}\right|}{n}$$

Ha valaki kicsit is jártas az abszolút értékes függvények ábrázolásában, akkor láthatja, hogy ez a kifejezés a medián esetén lesz minimális. Ennek grafikus bizonyítása a függelékben található.

A másik módszer az előjel kiküszöbölésére a négyzetre emelés. Tehát megadhatjuk az átlagos négyzetes eltérést, ami az eltérések négyzetének átlaga.

$$\sigma^2\left(\tilde{X}\right)=\frac{\left(x_1-\tilde{X}\right)^2+\left(x_2-\tilde{X}\right)^2+\ldots+\left(x_n-\tilde{X}\right)^2}{n}$$

Ezzel azonban főleg mértékegység problémák vannak. Ha az adatok valamilyen mértékegységgel rendelkeznek, akkor az átlagos négyzetes eltérés mérőszáma ennek négyzete, tehát szokás ennek négyzetgyökét venni.

Szintén a matematikai jártassággal rendelkezők be tudják bizonyítani, hogy az átlagos négyzetes eltérés a számtani közép esetén lesz minimális, ehhez a másodfokú függvények ismerete szükséges. A bizonyítás természetesen nem nagyon bonyolult, részletesen a függelékben található.

Az átlagos négyzetes eltérést a számtani középre felírva empirikus szórásnégyzetnek nevezzük, a négyzetgyökét empirikus szórásnak, jelölése σ_n.

Felmerülhet az a kérdés, hogy ezen mérőszámok közül melyiket érdemes használni a gyakorlatban, és erre a kérdésre a későbbiekben tárgyalandó Csebisev-egyenlőtlenség adja meg a választ. Eszerint ugyanis az átlagtól az adatok legfeljebb 25%-a térhet el a szórás kétszeresénél jobban, legfeljebb 10-11%-a térhet el a szórás háromszorosánál jobban, és 5-6%-a a szórás négyszeresénél jobban. Természetesen maga a Csebisev-egyenlőtlenség durva becslésen alapszik,

ezért gyakorlatilag a szórás négyszeresénél jobban nem térnek el az adatok az átlagtól, de az 5-6%-ot biztosan mondhatjuk bármilyen adathalmaz esetén.

(Ez egyébként konkrét adathalmazokra vonatkoztatva az úgynevezett empirikus Csebisev-törvény, vagy az átlag körüli szórás empirikus törvénye.)

I.6. A középértékek és a grafikus ábrázolás kapcsolata

A hisztogramon ábrázolt adatok esetén a lehető legegyszerűbb megkapni a medián értékét, hiszen csak meg kell keresni az oszlopok közül a középsőt, és az az érték lesz a medián. Ha páros számú adat van, akkor a két középsőt kell átlagolni.

Szár-levél diagram esetén a mediánt megtalálni szintén egyszerű, de ehhez célszerű az egyes adatsávokon belül az egyes adatokat nagyságrendi sorrendben feltüntetni.

II. Valószínűségszámítási alapismeretek

II.1. A véletlen esemény

A véletlen felfogása alapvetően kétféle lehet. Az ún. determinisztikus világképben a véletlen egyszerűen a tudásunk hiányát jelenti. Mivel nem tudunk minden adatot, ezért nem tudjuk meghatározni, mi fog történni, de ha ismernénk minden információt, akkor a véletlen megszűnne. A másik felfogás szerint a véletlen alapvető része a világnak, ezért akkor sem tudnánk megmondani, hogy mi történik, ha ismernénk minden adatot egy adott problémával kapcsolatban.

Hasonló a helyzet a valószínűséggel. Nem tudjuk pontosan eldönteni, hogy egy adott helyzetben, pl. kockával dobva a dobott értékek bekövetkezésének valószínűsége csak a kockától függ-e, tehát ún. objektív valószínűsége van-e, vagy pedig valami más tényező is befolyásolja ezt. Sokszor ebben a problémában az „elegendő ok hiánya” dönt: nincs okunk azt feltételezni, hogy egy adott kísérlet kimenetelei nem szimmetrikusak, tehát objektív valószínűséget tételezünk fel. Persze más kérdés az, hogy az objektív valószínűségről hogy lehet eldönteni, hogy mekkora az értéke?

Próbálkozhatunk modellalkotással is. Laplace (1812) az ún. klasszikus modellt alkalmazta: egy esemény bekövetkezésének valószínűsége a kedvező esetek száma osztva az összes esetek számával. Eszerint az úgynevezett elemi események (ami az „összes esetekből” egy darabot jelent) valószínűsége meg kell egyezzen. De milyen jogon mondhatjuk, hogy az egyes elemi események valószínűsége megegyezik? Ugyanoda jutottunk vissza, ahonnan elindultunk.

Kolgomorov (1933) tovább lépett. Azt mondta: axiómákat, tehát nem bizonyítandó (és nem is bizonyítható) állításokat kell felállítani a valószínűségről, és ezek segítségével fel kell építeni a valószínűség-fogalmat. [Tömören: Metrikának nevezzük azt a hozzárendelést, amelyben egy alaphalmaz részhalmazaihoz számértékeket rendelünk. A hozzárendelt értékeket az adott részhalmaz mértékszámának nevezzük. Egy metrika akkor valószínűség, ha az alaphalmazhoz rendelt értéke 1, nem negatív és additív (diszjunkt részhalmazok mértékének összege az uniójukhoz rendelt mérték).] Ezzel azonban szintén nem jutottunk közelebb ahhoz a problémához, hogy a kockával mekkora valószínűséggel dobunk hatost.

Szeretnénk tehát egy olyan állapotot elérni, amiben az a helyzet áll fenn, hogy bár nem tudunk mindent (ez tény), de a nemtudásunknak ne legyen nagy kockázata.

Kezdjük tehát elölről, építsünk fel egy modellt, aztán próbáljuk meg megvizsgálni, hogy ez a modell mennyire jó.

II.2. Klasszikus valószínűségszámítási modell

A valószínűségszámításban használni fogunk néhány fogalmat, ismerkedjünk meg ezekkel! Kísérletnek nevezünk valamely folyamatot, melyben a véletlen dönti el, hogy mi fog történni. A kísérletek általában megismételhetőek (mi itt ilyen kísérletekkel foglalkozunk), de kimenetelük nem feltétlenül lesz ugyanaz, mint egy korábbi kísérletben. Egy kísérlet lehetséges kimeneteleiből képezhetünk halmazokat; ezeket nevezzük eseményeknek. Például dobókockával egyszer dobva a kockadobás lehetséges kimenetelei 1, 2, 3, 4, 5, 6; esemény például, hogy páros számot dobtunk, vagy hogy 3-mal osztható számot dobtunk, de az is esemény, hogy pozitív számot dobtunk. *Lehetetlen eseménynek* nevezzük azokat az eseményeket, melyekhez nem tartozik egyetlen kimenet sem (pl. dobókockával nem egész számot dobunk), *biztos eseménynek* nevezzük azokat az eseményeket, melyekhez az összes lehetséges kimenetel hozzátartozik (pl. dobókockával dobva egész számot dobunk).

A leíró statisztikában szerepeltek az alábbi fogalmak: relatív gyakoriság, módusz, átlag, medián, szórás. Próbáljunk meg ezeknek valamiféle megfelelőt találni.

A relatív gyakoriság azt jelenti, hogy egy adott adathalmazban egy adat hányszor fordul elő. Ha a valószínűségről az a képünk van, hogy adott számú kísérletből az esemény bekövetkezéseinek száma arányos a bekövetkezésének valószínűségével, akkor a valószínűségre a relatív gyakoriságnak megfelelő értéket kell adnunk. Tekintsünk egy adott kísérletsorozatot, amely annyi kísérletből áll, ahány lehetséges kimenetele van a kísérleteknek. Tételezzük fel, hogy minden lehetséges kimenetel pontosan egyszer következik be a kísérletsorozatban (ez persze meglehetősen hihetetlennek tűnik). Egy adott esemény pontosan annyiszor következett be, ahányféle lehetséges kimenetel tartozik hozzá. Ha sok ilyen kísérletsorozatot végzünk el, akkor tapasztalataink (és modellünk) szerint egy adott esemény átlagosan annyiszor következik be, ahányféle lehetséges kimenetel tartozik hozzá. Ez vezet a Laplace-féle klasszikus modellhez:

$$p(A\ bekövetkezik) = \frac{jó\ esetek\ száma}{összes\ esetek\ száma}$$

Itt p a valószínűséget, A pedig egy adott eseményt jelent. Természetesen ügyelnünk kell arra, hogy az „összes esetek száma” olyan eseteket tartalmazzon, melyek bekövetkezése egyformán valószínű, ezek az úgynevezett elemi események. Annak eldöntésére, hogy mely események vehetők elemi eseményeknek, megint csak a modellhez kell nyúlnunk: ha sok kísérletet végezve lényegében egyforma számban következnek be ezek az elemi események, akkor mondhatjuk, hogy ezek valószínűsége megegyezik.

A biztos esemény valószínűsége 1, mivel itt minden eset jó eset, a lehetetlen esemény valószínűsége 0, mivel itt egyetlen jó eset sincs. (De vigyázat! Az, hogy egy esemény valószínűsége 1, nem feltétlenül jelenti azt, hogy a biztos eseményről van szó; ugyanígy az, hogy egy esemény valószínűsége 0, nem feltétlenül jelenti azt, hogy a lehetetlen eseményről van szó. A klasszikus valószínűségszámítási modellben igen, amikor véges sok kísérletet hajtunk végre, és az esetek száma véges sok. Azonban lehet végtelen kísérletsorozatokat is végrehajtani, illetve a geometriai modellben is lehet „végtelen sok eset” problémával találkozni; és itt már a fenti megállapításokat konkrét példákkal is illusztrálni lehet.)

Felmerült az a kérdés az előzőekben, hogy a klasszikus modell alkalmazásakor milyen eseményeket tekinthetünk elemi eseményeknek. Ez a kérdés már régen felmerült, és az úgynevezett „három kocka” problémájaként volt ismeretes. Ez a következőképpen hangzott:

Ha három kockával dobunk, akkor ugyanannyiféleképpen dobhatunk összesen 9-et, mint 10-et:

A dobott számok összege 9		A dobott számok összege 10	
1 + 2 + 6	1 + 3 + 5	1 + 3 + 6	1 + 4 + 5
1 + 4 + 4	2 + 2 + 5	2 + 2 + 6	2 + 4 + 4
2 + 3 + 4	3 + 3 + 3	2 + 3 + 5	3 + 3 + 4

Ha azonban elkezdünk dobálni három kockával, akkor azt tapasztaljuk, hogy az összeg gyakrabban lesz 10, mint 9. Mi lehet ennek az oka? Amikor azt mondtuk, hogy ugyanannyiféleképpen dobhatunk a három kockával 10-et és 9-et, akkor az elemi eseményeknek tulajdonképpen azt tekintettük, hogy a dobott számokat nagyságrendi sorrendbe rakva hogyan kaphatjuk meg a 10-et illetve a 9-et, azaz elemi eseményeknek az 1, 2, ..., 6 számokból összeállítható rendezetlen hármasokat vettük. De van-e jogunk ehhez? Ha a sorbarendezést nem engedjük meg (tehát az elemi eseményeket a rendezett számhármasok jelentik), akkor a 10-et már nem ugyanannyiféleképpen kaphatjuk meg, mint a 9-et. A táblázatban szereplő számhármasokat többször kell számolni a sorbarendezéseknek megfelelően, így a 10 esetén 27 lehetőségünk van, míg a 9 esetén csak 25. Hogyan lehet igazságot tenni? Mivel azt szeretnénk, hogy a gyakorlatban lejátszódó véletlen eseményeket tudjuk modellezni, ezért végre kell hajtanunk a három kocka dobásának kísérletét sokszor, és meg kell nézni, hogy melyik modell írja le jobban a megfigyelt jelenséget. A tapasztalat azt mutatja, hogy az a modell áll közelebb a valósághoz, amelyben nem tekintünk el a kockák dobási sorrendjétől.

Erre persze mondhatja valaki, hogy a három teljesen egyforma kinézetű, azonos anyagból készült, minden szempontból egyforma kockát mi különbözteti meg egymástól? Honnan tudhatjuk, hogy melyik melyik?

Végezzük el a következő gondolatkísérletet: Vegyünk három teljesen egyforma dobókockát, és fessük be őket pirosra, zöldre, kékre. Adjuk őket oda egy normális embernek, és dobáltassuk fel a kockákat. Ő meg tudja különböztetni a kockákat egymástól, hisz azok különböző színűek. Sok kísérlet végrehajtása után kap valamiféle relatív gyakoriságot a lehetséges értékekre. Ezután adjuk oda a kockát egy színvaknak, aki számára a három kocka teljesen egyforma, hiszen nem látja a színüket. Ha ő dobál a kockákkal, akkor nyilvánvalóan ugyanazt az eredményt kell kapja az egyes értékek relatív gyakoriságára, hiszen a kockák nem tudják, hogy most éppen egy színvak dobál velük, tehát nyilvánvalóan ugyanúgy viselkednek, mint eddig. Vegyük le ezután a színezést a kockákról, és adjuk vissza a normális embernek. Az ő kezében a kockák ugyanúgy kell viselkedjenek, mint a színvak ember kezében, hiszen a kockák arról sem tudnak, hogy be vannak-e színezve. Ebből a gondolatmenetből viszont az következik, hogy a megkülönböztető jellel ellátott kockák ugyanúgy viselkednek, mint a nem megkülönböztethető kockák. Tehát a sorrendet figyelembe kell vennünk a dobott értékeknél.

Hasonlóan kiváló terep az elemi események előfordulásának vizsgálatára az úgynevezett kockapóker játék. Ebben 5 kockával dobunk, és a klasszikus pókerszabályoknak megfelelően értékeljük a kapott számötöst (azaz a kiértékelésnél nem számít a dobási sorrend). A lehetőségek:

egy pár:	2 egyforma + 1 + 1 + 1 szám (pl. 3, 4, 5, 1, 3)
két pár:	2 egyforma + 2 egyforma + 1 szám (pl. 3, 5, 4, 3, 5)
terc vagy drill:	3 egyforma + 1 + 1 szám (pl. 2, 3, 4, 2, 2)
sor:	5 egymást követő szám, tetszőleges sorrendben (pl. 2, 3, 1, 4, 5)
full:	3 egyforma + 2 egyforma szám (pl. 3, 2, 3, 3, 2)
póker:	4 egyforma + 1 szám (pl. 3, 4, 3, 3, 3)
royal póker:	5 egyforma szám

Kérdés, hogy mekkora valószínűsége van az egyes lehetőségek bekövetkeztének? A kérdés megválaszolása szempontjából nem érdektelen, hogy a dobott számok sorrendjét is figyelembe vesszük az egyes lehetőségekhez tartozó „jó esetek” megszámolásánál, vagy pedig a dobási sorrendtől eltekintve határozzuk meg az esetek számát. A függelékben részletezett számítások

szerint az egyes lehetőségekhez az alábbi „jó eset”-számok tartoznak, attól függően, hogy a sorrendet figyelembe vettük-e avagy sem:

Lehetséges eredmény	Sorrend számít	Sorrend nem számít
Egy pár	3600	60
Két pár	1800	60
Terc	1200	60
Sor	240	2
Full	300	30
Póker	150	30
Royal póker	6	6

Látható, hogy a táblázat két oszlopa elég jelentősen eltérő adatokat tartalmaz. Hogy lehet eldönteni tehát, hogy a táblázat melyik oszlopát tekintsük érvényesnek a gyakorlati problémákban?

Egy osztálynyi tanuló (kb. 30 fő) mindegyike dobjon 50-szer az 5 db kockával, és jegyezze fel a kapott eredményeket a táblázatba. Utána összesítsék a tanulók a kapott eredményeket kimenetelenként, és ha ezen számok aránya az 1. oszlop arányainak megfelel, akkor a sorrendet figyelembe kell vennünk, ha a 2. oszlop arányainak felel meg, akkor a sorrendet nem kell figyelembe vennünk.

A tapasztalat azt mutatja már ilyen viszonylag kevés dobás esetén is, hogy a számok az 1. oszlop arányait mutatják, tehát a sorrendet figyelembe kell vennünk.

Jó-jó, mondhatja még mindig a kételkedő, de mi van azokkal a problémákkal, ahol nyilvánvalóan nem kell megkülönböztetnünk a sorrendet? Vegyük például a lottóhúzást! Ott már tényleg nem kell foglalkozni a sorrendi problémákkal! Valóban így van ez? Járjuk körbe ezt a kérdést alaposabban!

A következő kijelentéssel elég gyakran lehet találkozni: A lottó ötös kihúzásának valószínűsége $\frac{1}{\binom{90}{5}}$, (az összes esetek száma $\binom{90}{5}$(eltekintve a húzási sorrendtől, hiszen az nem számít), és a jó esetek száma 1, feltéve, hogy csak 1 szelvény töltöttünk ki).

Ha a klasszikus modellt a húzási sorrend figyelembevételével alkalmazzuk, akkor az összes esetek száma $90 \cdot 89 \cdot 88 \cdot 87 \cdot 86$, a jó esetek száma pedig 1, hiszen továbbra is csak 1 lottót töltöttünk ki, tehát az eredmény nem ugyanaz, mint az előbb. Igen ám, csakhogy ha a sorrendet

figyelembe vesszük a húzásnál, akkor figyelembe kell vennünk a lottószelvényünk kitöltésénél is! Az egy darab lottószelvényt $5\cdot 4\cdot 3\cdot 2\cdot 1$-féleképpen tölthetjük ki ugyanazokkal a számokkal, hiszen a beikszelést annyiféleképpen végezhetjük, ahányféleképpen az öt általunk kiválasztott számot sorbarakhatjuk. Tehát az összes esetek száma $\frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{90\cdot 89\cdot 88\cdot 87\cdot 86}$, amit átalakítva, és felhasználva a $\binom{90}{5}=\frac{90\cdot 89\cdot 88\cdot 87\cdot 86}{5\cdot 4\cdot 3\cdot 2\cdot 1}$ egyenlőséget az

$$\frac{\frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{5\cdot 4\cdot 3\cdot 2\cdot 1}}{\frac{90\cdot 89\cdot 88\cdot 87\cdot 86}{5\cdot 4\cdot 3\cdot 2\cdot 1}}=\frac{1}{\binom{90}{5}}$$

összefüggést kapjuk, tehát a kétféleképpen számolt valószínűség megegyezik.

Ha az átalakítást részletesebben megnézzük, akkor ki is derül ennek oka: amikor nem vesszük figyelembe a sorrendet, akkor tulajdonképpen az történik, hogy csoportosítjuk az elemi eseményeket: azokat a húzásokat, amelyekben ugyanazok a számok szerepelnek, egy kupacba tesszük, és „újfajta” elemi eseménynek tekintjük. Minden „újfajta” elemi esemény ugyanannyi régi elemi eseményből tevődik össze, ezért a sorrend figyelembe vételekor kapott hányadosban a számlálót és a nevezőt ugyanazzal a számmal kell leosztanunk, hogy a sorrend nélküli eset hányadosát kapjuk, így az eredmény nyilvánvalóan ugyanaz lesz.

Ez egyben azt is mutatja, hogy a sorrendet figyelembe véve mindig helyes eredményt kaphatunk, de számolásunkat leegyszerűsítheti a sorrendtől való eltekintés. Mikor lehet ezt megtenni? Erre is választ kaptunk: akkor, ha az elemi események beoszthatóak egyforma elemszámú csoportokba olyan módon, hogy minden csoportban olyan elemi események kerüljenek, melyek egymástól csak sorrendben térnek el, és az összes ilyen esemény bekerült az adott csoportba. Az is látszik, hogy ez a csoportba osztás nem tehető meg pl. a visszatevéses húzásnál a húzási sorrend alapján megkülönböztetett események esetén, hiszen ekkor a különböző elemeket tartalmazó húzásokat a sorrend szerint nem ugyanannyiszor számoljuk meg, mint az egyforma elemeket is tartalmazó húzásokat.

A legtöbb véletlen esemény modellezésére alkalmas az úgynevezett urna-modell. Legyen egy adott esemény bekövetkezésének valószínűsége $p=\frac{K}{N}$. Tegyünk egy urnába K db fehér, és $N-K$

db fekete golyót, és húzzunk ki egy golyót az urnából. Ekkor az esemény bekövetkezte megfelel a fehér golyó húzásnak, az esemény be nem következte a fekete golyó húzásnak.

Ennek segítségével ki tudjuk számítani, hogy pl. *n* db kísérletből hányszor következik be egy adott esemény. Az előzőekben ismertetett urnából húzzunk *n*-szer, visszatevéssel (tehát a kihúzott golyót mindig visszatesszük, azaz mindig $\frac{K}{N}$ eséllyel húzunk fehéret). Mi a valószínűsége, hogy *k*-szor húztunk fehér golyót? (Azaz mi a valószínűsége, hogy az A esemény *k*-szor következett be?)

Használjuk a klasszikus modellt. Az összes húzások száma N^n, a jó esetek száma, amikor *k* db fehéret és *n*-*k* db feketét szerepeltetünk: $\binom{n}{k} \cdot K^k \cdot (N-K)^{n-k}$ (*), így a keresett valószínűség

$$p(\text{A esemény } k\text{-szor következik be}) = \frac{\binom{n}{k} K^k \cdot (N-K)^{n-k}}{N^n} = \binom{n}{k} \cdot \frac{K^k}{N^k} \cdot \frac{(N-K)^{n-k}}{N^{n-k}} =$$

$$= \binom{n}{k} \cdot \frac{K^k}{N^k} \cdot \left(1 - \frac{K^{n-k}}{N^{n-k}}\right) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

(*) ***Megjegyzés:*** Először ki kell jelölnünk a *k* db fehér golyó helyét, ezt $\binom{n}{k}$-féleképpen tehetjük meg, majd kiválasztjuk a *k* fehéret visszatevéssel a *K* golyó közül, amelyeket a korábbiakban mondottak miatt megkülönbözethetőnek veszünk, K^k-féleképpen, és az $n-k$ feketét az $N-K$ fekete közül $(N-K)^{n-k}$-féleképpen.

II.3. Műveletek eseményekkel; kapcsolat a valószínűségek között

Az eseményekkel való műveletek között két legfontosabbat különböztetünk meg:

a) *események szorzata:* Az A esemény és B esemény szorzatán azon eseményt értjük, melyben A és B esemény együttesen bekövetkezik. Jelölése: AB. [Úgy is megfogalmazhatnánk, hogy az AB esemény minden esetben bekövetkezik, amikor A és B együttesen bekövetkezik.]

Az AB esemény bekövetkezésének valószínűségét úgy tudjuk meghatározni, hogy az együttes bekövetkezéshez tartozó elemi események számát elosztjuk az összes esetek számával.

Az AB esemény valószínűségének lehetséges kiszámításáról később még, az Események függetlensége című fejezetben ejtünk szót.

b) *események összege:* Az A esemény és B esemény összegén azon eseményt értjük, melyben A vagy B esemény bekövetkezik (nem kizáró vagy; A és B külön-külön, de együtt is bekövetkezhet). Jelölése: A + B. [Úgy is megfogalmazhatnánk, hogy az A + B esemény minden esetben bekövetkezik, amikor A vagy B bekövetkezik.]

Ha ismerjük az A és B események bekövetkezését jelentő elemi eseményeket, akkor az (A + B) esemény bekövetkezését jelentő elemi események száma úgy alakul, hogy A-hoz és B-hez tartozó elemi események darabszámának összegéből ki kell vonnunk az A és B mindegyikénél szereplő elemi események darabszámát, hiszen ezeket az összegben kétszer számoltuk, de csak egyszer kellett volna. Ha ezt valószínűségekre fordítjuk le, akkor $p(A+B)=p(A)+p(B)-p(AB)$.

Ha A és B diszjunkt események, azaz együttes bekövetkezésük nem lehetséges (tehát AB lehetetlen esemény), akkor a fenti összefüggés a $p(A+B)=p(A)+p(B)$ alakot ölti.

II.4. Feltételes valószínűség

Nevezzük az A esemény B eseményre vonatkozó feltételes valószínűségének annak a valószínűségét, hogy A esemény bekövetkezik, feltéve, hogy a B esemény mindenképpen bekövetkezik. Jelölje ezt a valószínűséget $p(A\mid B)$. Ebben az esetben a klasszikus valószínűségszámítási modellt alkalmazva felmerül a kérdés: mit is jelent az „összes esetek” illetve a „jó esetek” száma a problémára vetítve? Az „összes esetek” száma nyilvánvalóan azokra az esetekre korlátozódik, amikor a B esemény bekövetkezik; ezek közül azok számítanak majd „jó eseteknek”, amikor az A esemény is bekövetkezik. Tehát valamiféle összefüggés gyártható ennek segítségével:

$$p(A\mid B)=\frac{\textit{azon esetek száma, amikor A és B bekövetkezik}}{\textit{azon esetek száma, amikor B bekövetkezik}}$$

Ha a kapott törtnek a számlálóját és a nevezőjét is elosztjuk a kísérlet összes lehetséges végkimeneteleinek számával (összes eset), akkor a számlálóban A és B együttes bekövetkezésének valószínűsége, a nevezőben B bekövetkezésének valószínűsége jelenik meg. Tehát $p(A\mid B)=\frac{p(AB)}{p(B)}$. Természetesen ehhez a kiszámítási módhoz nem kell ragaszkodni konkrét számítási feladatokban, lehet a korábban vázolt esetszámok hányadosával is dolgozni.

A kapott képlet lehetőséget ad arra, hogy $p(A\mid B)$ és $p(B)$ ismeretében $p(AB)$ értékét meghatározzuk.

Példa: Két urnában fehér és fekete golyók vannak. Az első urnában 3 fehér és 2 fekete, a második urnában 2 fehér és 3 fekete. Dobókockával döntjük el, hogy melyik urnából veszünk ki egyet: ha a kockával 5-öst vagy 6-ost dobunk, akkor az első urnából, egyéb esetben a második urnából húzunk ki egy golyót. Mi annak a valószínűsége, hogy a második urnából húzunk fekete golyót?

Megoldás: Legyen a B esemény az, hogy a második urnából húzunk, az A esemény pedig az, hogy feketét húzunk. Ekkor annak valószínűsége, hogy a második urnából húzunk: $\frac{4}{6}$ (összes esetek száma 6, ebből 4-ben húzunk a második urnából), annak valószínűsége, hogy feketét húzunk, feltéve, hogy a második urnából húzunk: $\frac{3}{5}$. Tehát annak a valószínűsége, hogy a második urnából húzunk, és ez a golyó fekete lesz: $p(A \mid B) \cdot p(B) = p(AB) = \frac{2}{3} \cdot \frac{3}{5} = \frac{2}{5}$.

Példa: Két urnában fehér és fekete golyók vannak. Az első urnában 3 fehér és 2 fekete, a második urnában 2 fehér és 3 fekete. Dobókockával döntjük el, hogy melyik urnából veszünk ki egyet: ha a kockával 5-öst vagy 6-ost dobunk, akkor az első urnából, egyéb esetben a második urnából húzunk ki egy golyót. Mi annak a valószínűsége, hogy fekete golyót húzunk?

Megoldás: A „fekete golyó húzása” esemény két diszjunkt eseményre bontható: az első urnából húzunk fekete golyót vagy a második urnából húzunk fekete golyót. Az előző példában kiszámoltuk, hogy a második urnából való fekete golyó húzásának valószínűsége $\frac{2}{5}$, hasonlóan kiszámítható az első urnából történő fekete golyó húzásának valószínűsége $\frac{1}{3} \cdot \frac{2}{5} = \frac{2}{15}$. Összesen tehát $\frac{2}{5} + \frac{1}{15} = \frac{7}{15}$.

Megjegyzés: A feladat megoldása során a következő sémát használtuk: Legyen B_1 és B_2 két diszjunkt esemény úgy, hogy összegük kiadja a biztos eseményt, és A egy tetszőleges esemény. Ekkor az A esemény valószínűsége a következőképpen írható fel:

$$p(A) = p(B_1) \cdot p(A \mid B_1) + p(B_2) \cdot p(A \mid B_2)$$

A mondott azonosság formálisan is egyszerűen belátható: $p(B_1)\cdot p(A\mid B_1)=p(AB_1)$, $p(B_2)\cdot p(A\mid B_2)=p(AB_2)$, ennek megfelelően, mivel B_1 és B_2 események összege a biztos esemény: $p(AB_1)+p(AB_2)=p(A(B_1+B_2))=p(A)$. Ennek az állításnak az általánosan megfogalmazott alakját a *teljes valószínűség tételének* nevezik:

Legyenek B_1, B_2, …, B_n páronként diszjunkt események, melyek összege a biztos eseményt adja ki (úgynevezett teljes eseményrendszer). Ekkor egy tetszőleges A esemény valószínűsége úgy határozható meg, hogy $p(A)=p(B_1)\cdot p(A\mid B_1)+p(B_2)\cdot p(A\mid B_2)+...+p(B_n)\cdot p(A\mid B_n)$

A teljes valószínűség tételét nem feltétlenül kell kimondanunk, hiszen a mintafeladatok megoldásánál vázolt gondolatmenet alapján mindig végiggondolható az egyes események valószínűségének feltételes valószínűségekkel történő kiszámítása, azonban a teljesség kedvéért megemlítjük itt.

Még egy megjegyzés: A feltételes valószínűségekkel kapcsolatos az úgynevezett Bayes-tétel is. Ha B_1, B_2, …, B_n páronként diszjunkt események, melyek összege a biztos eseményt adja ki, és ismerjük a $p(B_1)$, $p(B_2)$, … valószínűségeket, valamint a $p(A\mid B_1)$, $p(A\mid B_2)$ … feltételes valószínűségeket, akkor ki tudjuk számítani a $p(B_1\mid A)$, $p(B_2\mid A)$, … feltételes valószínűségeket. A Bayes-tétel azt az eljárást foglalja össze és mondja ki tétel formájában, melyben a $p(B_1\mid A)=\frac{p(B_1A)}{p(A)}$ típusú hányadosok számlálóját és nevezőjét határozzuk meg a korábban mondott módszerek szerint. A Bayes-tétel kimondását azért nem tartjuk itt szükségesnek, mert felesleges képlet, és alkalmat ad arra, hogy összezavarja a középiskolások gondolatait. Akik szükségesnek tartják, a rendelkezésre álló információk alapján összeállíthatják a tételt maguk is. A számolás menetét azonban nézzük meg egy gyakorlati példán:

Példa: Két urnában fehér és fekete golyók vannak. Az első urnában 3 fehér és 2 fekete, a második urnában 2 fehér és 3 fekete. Dobókockával döntjük el, hogy melyik urnából veszünk ki egyet: ha a kockával 5-öst vagy 6-ost dobunk, akkor az első urnából, egyéb esetben a második urnából húzunk ki egy golyót. Valaki elvégzi a kísérletet, és fekete golyót húz. Mi annak a valószínűsége, hogy a második urnából húzott?

Megoldás: Legyen a B esemény az, hogy a második urnából húzunk, az A esemény pedig az, hogy feketét húzunk. Ekkor a $p(B\mid A)$ feltételes valószínűséget keressük. Ennek kiszámítása a

$p(B \mid A) = \frac{p(BA)}{p(A)}$ képlet alapján történik. Határozzuk meg a $p(BA)$ [a második urnából húztunk és fekete golyót] és $p(A)$ [fekete golyót húztunk] valószínűségeket a korábban bemutatott módon!

$p(BA)$ kiszámítása: Annak valószínűsége, hogy a második urnából húzunk: $\frac{4}{6}$, annak valószínűsége, hogy feketét húzunk, feltéve, hogy a második urnából húzunk: $\frac{3}{5}$. Tehát annak a valószínűsége, hogy a második urnából húzunk, és ez a golyó fekete lesz: $p(A \mid B) \cdot p(B) = p(AB) = \frac{2}{3} \cdot \frac{3}{5} = \frac{2}{5}$.

$p(A)$ kiszámítása: annak a valószínűsége, hogy a második urnából húzunk, és ez a golyó fekete lesz: $\frac{2}{3} \cdot \frac{3}{5} = \frac{2}{5}$ annak a valószínűsége, hogy a második urnából húzunk, és ez a golyó fekete lesz: $\frac{1}{3} \cdot \frac{2}{5} = \frac{2}{15}$. Összesen tehát $p(A) = \frac{2}{5} + \frac{1}{15} = \frac{7}{15}$.

A keresett feltételes valószínűség: $p(B \mid A) = \frac{p(BA)}{p(A)} = \frac{\frac{2}{5}}{\frac{7}{15}} = \frac{2}{7}$.

II.5. Események függetlensége

Az A eseményt a B eseménytől függetlennek tekinthetjük akkor, ha a B esemény bekövetkezése vagy be nem következése nem befolyásolja az A esemény bekövetkezésének valószínűségét. Ezt úgy is megfogalmazhatjuk, hogy minden olyan esetben, amikor B esemény bekövetkezik, az A esemény ugyanolyan valószínűséggel következik be, mint azokban az esetekben, amikor a B esemény nem következik be (és ez a valószínűség nyilván megegyezik az A esemény mindenféle feltétel nélküli bekövetkezésének valószínűségével). Ha numerikus összefüggést keresünk, akkor a korábban kapott feltételes valószínűségre vonatkozó összefüggést alkalmazhatjuk: $p(A \mid B) = p(A)$, azaz $p(A \mid B) = \frac{p(AB)}{p(B)} = p(A)$. A második egyenletből azt kapjuk, hogy $p(AB) = p(A) \cdot p(B)$. Ez az összefüggés több dologra is alkalmas:

1. Ha ismerjük $p(A)$, $p(B)$, $p(AB)$ értékét, akkor el tudjuk dönteni, hogy az A és B események függetlenek-e.

2. Ha tudjuk, hogy A és B események függetlenek, akkor $p(A)$ és $p(B)$ értékének ismeretében ki tudjuk számítani $p(AB)$ értékét.

A probléma persze az, hogy mi van akkor, ha két esemény függetlenségét vagy összefüggését szeretnénk eldönteni, de nem tudjuk meghatározni $p(AB)$ értékét. Nyilván nem járható út, hogy számítsuk ki a fenti képletből, hiszen ez csak akkor alkalmazható, ha tudjuk, hogy A és B függetlenek; viszont amíg nem tudjuk, hogy A és B függetlenek, addig nem alkalmazhatjuk ezt a képletet. Ez egyfajta ördögi kör, amiből mindenképpen ki kell lépni. Úgy tudjuk feloldani ezt a látszólagos ellentmondást, hogy a függetlenség fogalmát kiterjesztjük: ha a két esemény olyan, hogy nyilvánvalóan nem befolyásolják egymás bekövetkezésének valószínűségét, akkor elfogadjuk, hogy a két esemény független, és az együttes bekövetkezésük valószínűségét a fenti összefüggéssel meghatározhatjuk.

II.6. Valószínűségi változó eloszlása

Valószínűségi változónak nevezzük azt a mennyiséget, melynek értékét valamely véletlen esemény határozza meg. A valószínűségi változó eloszlása azt adja meg, hogy a változó egy-egy értéket milyen valószínűséggel vesz fel. A mi mostani tárgyalásunkban az úgynevezett *binomiális eloszlású valószínűségi változók* játszanak nagy szerepet, ezek eloszlása

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

ahol *p* valamely esemény bekövetkeztének valószínűsége.

Ilyen változó pl. a egy adott *p* valószínűségű esemény *n* független kísérletből történő bekövetkezéseinek száma.

Legyen az adott esemény bekövetkezésének valószínűsége $p=\frac{K}{N}$. Tegyünk egy urnába *K* db fehér, és $N-K$ db fekete golyót, és húzzunk ki egy golyót az urnából. Ekkor az esemény bekövetkezte megfelel a fehér golyó húzásnak, az esemény be nem következte a fekete golyó húzásnak.

Az előzőekben ismertetett urnából húzzunk *n*-szer, visszatevéssel (tehát a kihúzott golyót mindig visszatesszük, azaz mindig $\frac{K}{N}$ eséllyel húzunk fehéret). Mi a valószínűsége, hogy *k*-szor húztunk fehér golyót? (Azaz mi a valószínűsége, hogy a vizsgált esemény pontosan *k*-szor következett be?)

Használjuk a klasszikus modellt. Az összes húzások száma N^n, a jó esetek száma, amikor k db fehéret és n-k db feketét szerepeltetünk: $\binom{n}{k} \cdot K^k \cdot (N-K)^{n-k}$ (*), így a keresett valószínűség

$$p(A \text{ esemény } k-\text{szor következik be}) = \frac{\binom{n}{k} K^k \cdot (N-K)^{n-k}}{N^n} = \binom{n}{k} \cdot \frac{K^k}{N^k} \cdot \frac{(N-K)^{n-k}}{N^{n-k}} =$$

$$= \binom{n}{k} \cdot \frac{K^k}{N^k} \cdot \left(1 - \frac{K^{n-k}}{N^{n-k}}\right) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

(*) ***Megjegyzés:*** Először ki kell jelölnünk a k db fehér golyó helyét, ezt $\binom{n}{k}$-féleképpen tehetjük meg, majd kiválasztjuk a k fehéret visszatevéssel a K golyó közül, amelyeket a korábbiakban mondottak miatt megkülönbözethetőnek veszünk, K^k-féleképpen, és az $n-k$ feketét az $N-K$ fekete közül $(N-K)^{n-k}$-féleképpen.

A gyakorlati életben előforduló problémákban nem mindig visszatevéses mintavételt alkalmazunk, hanem sokszor visszatevés nélkülit, és ekkor az úgynevezett *hipergeometrikus eloszlást* kapjuk. (Pl. ha a közvéleménykutatásnál megkérdezünk embereket, akkor ügyelünk arra, hogy kétszer ne ugyanazt az embert kérdezzük meg.) A hipergeometrikus eloszlás azonban nagy elemszámú halmazokban kis elemszámú mintavétel esetén közelíthető a binomiális eloszlással, nevezetesen:

$$p(X=k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \approx \binom{n}{k} \cdot \left(\frac{K}{N}\right)^k \cdot \left(\frac{N-K}{N}\right)^{n-k}$$

Itt N a halmaz elemszáma, K a kitüntetett elemek száma, n a minta elemszáma, és X azt jelöli, hogy a kihúzott n elem közül hány esik a kitüntetett elemek közé.

A fenti egyenlőség szemléletesen is megmagyarázható: ha sok elem van az urnában, és keveset húzok, akkor egy kihúzott elemet kicsi valószínűséggel húznék ki még egyszer az urnából, tehát nem jelentős eltérést okoz, ha nem is teszem vissza; másrészt pedig egy kihúzott elemmel nem csökken lényegesen az elemek száma, tehát egy elem kihúzásának valószínűsége csak nagyon kicsivel változik a visszatevéses esethez képest. A fenti közelítés csak a mondott feltételek teljesülése esetén áll fenn (N „nagy”, n „kicsi”), a tényleges valószínűségtől való eltérés a fenti binomiális eloszlással való közelítés esetén tetszőleges k-t választva kisebb $\left(\frac{n}{N}\right)^n$-nél.

II.5. A valószínűségi változókat jellemző adatok

Az adathalmazoknál már láthattuk, hogy az átlag nagy szerepet játszott az adatok jellemzésében. Mi felelne meg a mi modellünkben az átlagnak? Az átlagban a számok összege szerepelt, darabszámukkal osztva. Ha csoportosítjuk az adatokat, akkor a számok összegében minden adatnak annyiszorosa szerepel, ahányszor előfordul az adathalmazban. Tehát

$$\frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{k_1 x_1 + k_2 x_2 + \ldots + k_m x_m}{n} = \frac{k_1}{n} x_1 + \frac{k_2}{n} x_2 + \ldots + \frac{k_m}{n} x_m,$$

ahol $k_1 + k_2 + \ldots + k_m = n$, ezért $\frac{k_1}{n} + \frac{k_2}{n} + \ldots + \frac{k_m}{n} = 1$.

A fenti képletben az egyes adatok relatív gyakoriságai szerepelnek, amiket a valószínűségnek feleltettünk meg, az adatok pedig a valószínűségi változó lehetséges értékeit jelölik. Ez a kapcsolat lehetőséget ad arra, hogy az átlaggal analóg várható érték definícióját megadjuk:

Az X valószínűségi változó várható értéke

$$E(X) = p_1 x_1 + p_2 x_2 + \ldots + p_n x_n,$$

ahol a valószínűségi változó eloszlása $p(X = x_n) = p_n$

A várható érték jelentése nem az, hogy ha pl. kockával dobunk, akkor 3,5-et fogunk dobni, hiszen ez meglehetősen furcsa lenne, hanem az, hogy elegendően sok kísérletet végezve a kapott adatok átlaga a várható érték környékén lesz. Megelőlegeztük már az átlag és a szórás kapcsolatában az adatok elhelyezkedését, és itt is lehetőséget kapunk majd arra, hogy tippeljünk előre arra, hogy mi lesz a kimenetele egy adott véletlen kísérletnek.

A várható érték tulajdonságai:

– Ha $E(X)$ létezik, akkor létezik $E(cX)$ is, és $E(cX) = cE(X)$, ahol c tetszőleges állandó
– Ha $E(X)$ és $E(Y)$ létezik, akkor $E(X+Y)$ is, és $E(X + Y) = E(X) + E(Y)$
– Ha X és Y független valószínűségi változók, és létezik a várható értékük, akkor létezik $E(X \cdot Y)$ is, és $E(X \cdot Y) = E(X) \cdot E(Y)$

A binomiális eloszlású valószínűségi változók várható értékének meghatározására használjuk fel a várható érték fent jelzett második tulajdonságát!

Vezessünk be egy úgynevezett indikátor-változót, melynek értéke 1, ha az A esemény bekövetkezik, és 0, ha az A esemény nem következik be. Nyilvánvaló a definícióból, hogy az

indikátor-változó várható értéke $p(A)$. Vizsgáljuk most n számú független kísérletben az A esemény bekövetkezéseinek számát. Ez nyilván binomiális eloszlású valószínűségi változót jelent. Vezessünk be n darab indikátor változót: x_i értéke 0, ha az i-edik kísérletben az A esemény nem következett be, és 1, ha bekövetkezett. Nyilvánvalóan az A esemény bekövetkezéseinek számát az indikátor változók összege jelenti, tehát várható értéke az indikátor változók várható értékének összege, azaz $n \cdot p(A)$. Ha tehát egy p, n paraméterű binomiális eloszlást vizsgálunk, akkor annak várható értéke np.

Annak mérésére, hogy a várható érték mennyire jó mérőszám (akárcsak arra, hogy az átlag mennyire jó), többféle módszer is választható. A leíró statisztikához hasonlóan nézzük végig az ott már definiált mérőszámok megfelelőit.

Az átlagos abszolút eltérésre adott képlet szerint $\frac{\left|x_1 - \tilde{X}\right| + \left|x_2 - \tilde{X}\right| + \ldots + \left|x_n - \tilde{X}\right|}{n}$, ha az x_i adat k_i-szer szerepel a felsoroltak között, akkor a fenti képlet az

$$\frac{\left|x_1 - \tilde{X}\right| \cdot k_1 + \left|x_2 - \tilde{X}\right| \cdot k_2 + \ldots + \left|x_l - \tilde{X}\right| \cdot k_n}{n}$$

alakot ölti, aminek a valószínűségi változók esetén a

$$|x_1 - E(X)| \cdot p_1 + |x_2 - E(X)| \cdot p_2 + \ldots + |x_l - E(X)| \cdot p_n$$

érték felel meg, ami nem más, mint az $|X - E(X)|$ valószínűségi változó várható értéke. Ezt a kifejezést a valószínűségi változó *várható abszolút eltérésének* nevezzük.

A leíró statisztikában már láthattuk, hogy a szórásnak nagyobb szerepe van az adathalmazok leírásában, ezért az empirikus szórásnégyzetre kapott

$$\sigma^2\left(\tilde{X}\right) = \frac{\left(x_1 - \tilde{X}\right)^2 + \left(x_2 - \tilde{X}\right)^2 + \ldots + \left(x_n - \tilde{X}\right)^2}{n}$$

képlet alapján valószínűségi változók esetén a szórásnégyzetet (a fenti gondolatmenethez hasonlóan) az $(X - E(X))^2$ várható értékeként definiálhatjuk, azaz

$$D^2(X) = \left(x_1 - E(X)\right)^2 \cdot p_1 + \left(x_2 - E(X)\right)^2 \cdot p_2 + \ldots + \left(x_n - E(X)\right)^2 \cdot p_n$$

A valószínűségi változó szórása a szórásnégyzet négyzetgyöke, és formálisan is $D(X)$-szel jelöljük.

A várható érték alaptulajdonságait felhasználva a szórásnégyzet a

$$D^2(X) = E\left[(X - E(X))^2\right] = E\left[X^2 - 2X \cdot E(X) + E(X)^2\right] =$$
$$= E(X^2) - 2E(X) \cdot E(X) + E(X)^2 = E(X^2) - E(X)^2$$

alakban írható.

A szórásnégyzet tulajdonságai közül bizonyítás nélkül közöljük az alábbit:

Ha X és Y független valószínűségi változók, akkor $D^2(X+Y)=D^2(X)+D^2(Y)$.

Ezt a tulajdonságot felhasználhatjuk a binomiális eloszlású valószínűségi változó szórásának meghatározására. Használjuk a korábban bevezetett, független indikátor változókat. Ezek eloszlása $p(X_i=1)=p$, $p(X_i=0)=1-p$, így négyzetük eloszlása $p(X_i^2=1)=p$, $p(X_i^2=0)=1-p$ ugyanez, tehát várható értékük *p*, négyzetük várható értéke *p*. Ekkor viszont szórásnégyzetük:

$D^2(X_i)=E(X_i^2)-E(X_i)^2=p-p^2=p(1-p)$

Mivel a binomiális eloszlású valószínűségi változó *n* db ilyen független indikátorváltozó összege, ezért a szórásnégyzet említett tulajdonságát felhasználva a binomiális eloszlású valószínűségi változó szórásnégyzete: $D^2(X)=np(1-p)$, ezért a binomiális eloszlás szórása $D(X)=\sqrt{np(1-p)}$

A leíró statisztikában használt többi középértéknek is keressük meg a valószínűségszámításbeli megfelelőjét!

A módusz azt mutatta meg, hogy melyik adat szerepel a legtöbbször az adathalmazban. Ennek itt a *legnagyobb valószínűségű érték* felel meg (az az érték, amelyhez a legtöbb elemi esemény tartozik).

A medián az adatok felezője volt, tehát az adatok fele nála kisebb, az adatok fele nála nagyobb. Ennek az a *k* érték felel meg, melyre teljesül, hogy $p(X<k)=\frac{1}{2}$. Ha nincs ilyen érték, akkor az ezt legjobban teljesítő számot tekintsük. Hasonlóan az alsó és felső kvartilisnek azok az *m* illetve *n* értékek felelnek meg, melyekre $p(X<m)=\frac{1}{4}$ illetve $p(X<n)=\frac{3}{4}$.

III. Statisztikai becslések

III.1. Egyszerű becslések

A gyakorlati életben felmerülő problémák között sokszor előfordul, hogy szeretnénk egy adott sokaságról valamit megtudni, pl. hogy az emberek melyik mosóport vásárolják szívesebben, vagy hogy egy gyár termékei között hány selejtes van. Ha azonban minden elemét végig kéne nézni a sokaságnak, akkor az egyrészt nagyon sok ideig tartana, másrészt nagyon sok pénzbe kerülne. Természetesen bármilyen, a sokaságnak csak egy részét vizsgáló megoldás nem ad pontos eredményt, de ha jóval olcsóbban és gyorsabban jutunk egy kevésbé pontos eredményhez, akkor lehet, hogy inkább a pontosságból engedünk szívesebben.

Tehát: adott egy N elemből álló adathalmaz, ahol N elég nagy, és van benne K darab bizonyos típusú (nevezzük őket „megjelölt” adatoknak), a mi érdeklődésünkre számot tartó adat. Szeretnénk megtudni, hogy K mennyi. Ehhez kiválasztunk n darab adatot, és megnézzük, hogy közöttük hány „megjelölt” van, ez legyen k darab. Véleményünk szerint a $\frac{k}{n}$ mennyiség valamennyire megközelíti a $\frac{K}{N}$ mennyiséget. Jó lenne tudni, hogy mennyire közelíti meg, azaz $\frac{k}{n}$ értékének ismeretében $\frac{K}{N}$ milyen határok közé eshet; továbbá arról is jó lenne információt kapni, hogy hány elemet kell kiválasztanunk a sokaságból, hogy jól (azaz általunk megadott pontosságon belül) megközelítse.

A $\frac{K}{N}$ mennyiséget jelölje p, ez annak a valószínűsége, hogy az első húzásnál „megjelölt” adattal találkozunk. Természetesen a második húzásnál már nem ekkora a valószínűsége, hiszen akár „megjelölt” volt az első adat, akár nem, N értéke megváltozik (hiszen nem tesszük vissza a kiválasztott adatokat), és az új valószínűség $\frac{K}{N-1}$ vagy $\frac{K-1}{N-1}$. Azonban ha sok adat van, és összességében keveset veszünk ki közülük, akkor ez a változás nem lényeges. Nagy pontossággal mondhatjuk azt, hogy a további esetekben is mindig p a valószínűsége annak, hogy megjelölt adatot választunk ki. Ez persze némi pontatlanságot eredményez, de az általunk vett tizedesjegyek által generált hibahatáron belül maradunk ezzel a közelítéssel. Ezzel az úgynevezett binomiális eloszlást kapjuk, melynél n független kísérletben vizsgáljuk egy adott p valószínűségű esemény bekövetkezéseinek darabszámát.

A becslések végrehajtásához szükségünk lesz két egyenlőtlenségre:

Markov-egyenlőtlenség:

Legyen az *x* nemnegatív értékű valószínűségi változó. Ekkor

$$p[x>\mu\cdot E(x)]<\frac{1}{\mu}$$

A Markov-egyenlőtlenség előnye, hogy bármilyen valószínűségi változó esetén igaz. Ez természetesen hátránya is, hiszen emiatt a becslés meglehetősen durva; ennek ellenére használható eredményt fog adni számunkra.

A Markov-egyenlőtlenség segítségével levezethető a **Csebisev-egyenlőtlenség**:

$$p\left(\left|x-E(x)\right|\geq\lambda\cdot D(x)\right)<\frac{1}{\lambda^2}$$

A Csebisev-egyenlőtlenség szintén minden valószínűségi változó esetén igaz, és örökli a Markov-egyenlőtlenség pontatlanságát.

(A Markov-és Csebisev-egyenlőtlenség levezetésével most nem foglalkozunk, az érdeklődők a függelékben utánaolvashatnak.)

A binomiális eloszlású valószínűségi változó a következő tulajdonságokkal rendelkezik:

1. Annak valószínűsége, hogy egy adott, *n* kísérletből álló kísérletsorozatban az általunk vizsgált *p* valószínűségű esemény *k*-szor következik be: $p(X=k)=p_k=\binom{n}{k}p^k(1-p)^{n-k}$.

2. Várható értéke $E(X)=np$.

3. Szórása $D(X)=\sqrt{np(1-p)}$.

Írjuk fel a Csebisev-egyenlőtlenséget binomiális eloszlású valószínűségi változó esetén, felhasználva a fenti tulajdonságokat:

$$p\left(\left|x-np\right|\geq\lambda\cdot\sqrt{np(1-p)}\right)<\frac{1}{\lambda^2}$$

Térjünk át az *x* értéke helyett a relatív gyakoriságra, azaz $\frac{x}{n}$-re. A zárójelben levő egyenlőtlenséget ehhez végig kell osztanunk *n*-nel:

$$p\left(\left|\frac{x}{n}-p\right|\geq\lambda\cdot\sqrt{\frac{p(1-p)}{n}}\right)<\frac{1}{\lambda^2}$$

Vezessük be a $\delta=\lambda\cdot\sqrt{\frac{p(1-p)}{n}}$ jelölést! Ekkor a fenti egyenlőtlenség a

$$p\left(\left|\frac{x}{n}-p\right|\geq\delta\right)<\frac{p(1-p)}{n\cdot\delta^2}$$

alakot ölti.

Ezt az alakot még a *p*-től is függetlenné tudjuk tenni, ha felhasználjuk a $p(1-p)\leq\frac{1}{4}$ egyenlőtlenséget (ez pl. a számtani és mértani közép közti egyenlőtlenséggel vagy rendezéssel és teljes négyzetté alakítással bizonyítható):

$$p\left(\left|\frac{x}{n}-p\right|\geq\delta\right)<\frac{p(1-p)}{n\cdot\delta^2}\leq\frac{1}{4n\cdot\delta^2}\text{, tehát}$$

$$p\left(\left|\frac{x}{n}-p\right|\geq\delta\right)<\frac{1}{4n\cdot\delta^2}$$

Felhasználva a valószínűség azon tulajdonságát, hogy egy eseménynek és komplementer eseményének valószínűsége összesen 1, a kapott egyenlőtlenséget a következő formába írhatjuk:

$$p\left(\left|\frac{x}{n}-p\right|<\delta\right)\geq1-\frac{1}{4n\cdot\delta^2}$$

Vizsgáljuk meg a most kapott egyenlőtlenséget közelebbről! $\frac{x}{n}$ egy adott kísérletsorozatban az általunk figyelemmel kísért, *p* valószínűségű esemény bekövetkezésének relatív gyakorisága. δ értékét mi választhatjuk meg, tehát a relatív gyakoriságnak a valószínűségtől való eltérését tetszőlegesen előírhatjuk (azaz tetszőlegesen kicsinek választhatjuk). A jobboldalon viszont tetszőlegesen nagy értéket szerepeltethetünk, ha *n*-et elég nagynak választjuk. Ez azt jelenti, hogy elegendően nagy számú kísérlet esetén a vizsgált esemény bekövetkezésének relatív gyakorisága tetszőlegesen nagy valószínűséggel, tetszőleges pontossággal megközelítheti a vizsgált esemény bekövetkezésének valószínűségét. (Ennek általánosabb alakját szokás a nagy számok törvényének is nevezni.) Ez viszont lehetőséget ad nekünk arra, hogy megbecsüljük egy ismeretlen valószínűségű, de binomiális eloszlású (vagy annak tekinthető) esemény valószínűségét.

Próbáljuk meg a kapott egyenlőtlenség segítségével megvizsgálni, hogy mit kell tennünk ahhoz, hogy egy érméről eldöntsük: elhisszük-e, hogy szabályos, illetve hogyan határozhatnánk meg pl. a fejdobás valószínűségét!

Hányszor kell feldobni egy pénzérmét, hogy 99%-os biztonsággal az „objektív” valószínűsége a fejdobásnak ne térjen el a relatív gyakoriságtól 0,01-nál jobban? Adjunk becslési módszert a fejdobás valószínűségére a kapott eredmény alapján!

A feladatban megfogalmazott feltétel azt jelenti, hogy annak valószínűsége, hogy a fejdobás relatív gyakorisága a fejdobás valószínűségétől legalább 0,01-ra eltér, kisebb, mint 1%, azaz 0,01. Ebben az esetben $\delta = 0{,}01$, így teljesülnie kell $\frac{1}{4n \cdot \delta^2} \le 0{,}01$-nek, azaz $n \ge 25000$. Tehát legalább ennyi dobást kell végrehajtanunk a becsléshez. Ha csak egy dobássorozatot végzünk, akkor a fejdobások relatív gyakorisága 99%-os valószínűséggel nem tér el 0,01-nál jobban a fejdobás tényleges valószínűségétől. Tehát a fejdobás tényleges valószínűségére azokat az értékeket hisszük el, melyek az $\left[\frac{x}{n} - 0{,}01; \frac{x}{n} + 0{,}01\right]$ intervallumba fognak esni. Ez a módszer lehetőséget ad úgynevezett hipotézisvizsgálatra: a hipotézis a kezdeti feltételezésünk (pl. hogy a pénzérme szabályos, azaz a fejdobás valószínűsége 0,5), és a kapott eredmény függvényében döntjük el, hogy a hipotézist elfogadjuk vagy elvetjük.

Ha viszont pontosabb eredményt szeretnénk kapni az általunk teljesen ismeretlennek tekintett fejdobás-valószínűségről, akkor a következő lehetőség kínálkozik: végezzünk sok ilyen 250000 dobásos kísérletet, nézzük meg, hogy ezekben mekkora a relatív gyakorisága a fejdobásnak. Mivel ez a fejdobás tényleges valószínűségétől nagy valószínűséggel nem tér el 0,01-nél jobban, ezért a kapott relatív gyakoriságokat fedjük le egy 0,02 hosszúságú intervallummal úgy, hogy minél kevesebb „lógjon ki” az adatok közül, és ennek a lefedő intervallumnak a közepét tudjuk használni a fejdobás valószínűségének becsléséhez.

Mi a hátránya ennek a módszernek? Borzasztóan sokat kell dobálni a vizsgált pénzérmével, de ez nem is olyan nagy csoda, hiszen a felhasznált becslések nagyon durvák voltak. Úgy is mondhatnánk, hogy kis harc árán csak kis győzelmet tudunk elérni ebben a küzdelemben.

Az imént vizsgált problémában egy dobássorozat esetén hipotézist állítottunk fel, melyet elfogadtunk vagy elvetettünk az eredménytől függően. Milyen módszer alkalmazható azokban az esetekben, amikor egy kísérletsorozat eredménye alapján akarunk közvetlen információt kapni egy adott valószínűségről, kezdeti feltételezések nélkül?

III.2. Konfidencia intervallum

Népszavazáshoz gyűjtöttek aláírást egy országban. Az első kontroll után (kiválogatták azokat az aláírásokat, melyek mellől valamely a hitelesítéshez szükséges adat hiányzott, illetve azokat, melyek többször szerepeltek) maradt 41000 aláírás. Ebből kiválasztottak 3000-et véletlenszerűen, és ezek hitelességét ellenőrizték. Azt találták, hogy ezek közül 343 nem volt hiteles. Mit mondhatunk a 41000 aláírás között levő hiteles aláírások számáról, 95%-os biztonsággal?

Ebben a példában szintén nem ismerjük a 41000 aláírás között levő hiteles aláírások számát, tehát nem ismerjük annak valószínűségét, hogy közülük egyet véletlenszerűen kiválasztva az hiteles lesz. Ha valaki azt mondja, hogy a hiteles aláírások száma 15000, akkor azt elég nehezen fogjuk elhinni, hogy mi a 3000 kiválasztottból csak egytizedét húztuk a nem hitelesek közül, holott abból jóval több van. Azt sem fogjuk elhinni, hogy 40500 hiteles aláírás van, hiszen ez azt jelentené, hogy a kihúzott 3000 aláírás között van az összes nem hiteles több, mint fele.

Van tehát egy alsó és egy felső határa annak, hogy milyen valószínűséget „hiszünk el” egy adott esemény bekövetkeztére; az ezek között levő számok halmazát konfidencia (megbízhatósági) intervallumnak nevezzük

Megjegyzés: Be kéne bizonyítanunk természetesen azt, hogy tényleg minden szám „hihető” valószínűséget ad a kapott szélső értékek között, tehát hogy a megfelelő („hihető”) valószínűségek halmaza valóban egy intervallum. Ezt nem fogjuk külön bizonyítani, hanem a folytonossági tulajdonságokra alapozva szemléletesen elfogadjuk.

Feladatunkban tehát azt kell meghatározni, hogy mennyi a konfidencia intervallum az aláírások esetén, sőt, igazából csak az alsó határára vagyunk kíváncsiak (tehát hogy legalább hány hiteles aláírás van közöttük).

A megvizsgált aláírásokról feltételezzük, hogy kiválasztásuk véletlenszerű volt. Mivel a 41000 darabhoz képest csak keveset vettünk ki hitelesítésre, ezért egy aláírás kivétele a halmazból nem lényegesen változtatta meg a nem hiteles aláírás kiválasztásának valószínűségét egy-egy lépésre vonatkoztatva. Emiatt jó közelítéssel kezelhetjük úgy a problémát, hogy minden egyes lépésben a nem hiteles aláírás választásának valószínűsége p, ahol $p=\frac{\text{nem hiteles aláírások száma}}{41000}$. Tehát a közelítésünkben a kiválasztott nem hiteles aláírások számát binomiális eloszlásúnak tekintjük. A

kiválasztott nem hiteles aláírások száma $x = 343$, a kiválasztott összes aláírások („kísérletek”) száma: $n = 3000$.

A Csebisev-egyenlőtlenségből binomiális eloszlás esetén kapott $P\left(\left|\frac{x}{n}-p\right|<\delta\right)\geq 1-\frac{1}{4n\cdot\delta^2}$ egyenlőtlenséget fogjuk használni. A vizsgálatból $n = 3000$, $x = 343$, $1-\frac{1}{4n\cdot\delta^2}=0{,}95$, és innen δ értéke meghatározható: $\delta\approx 0{,}041$. Tehát 95%-os biztonsággal állíthatjuk, hogy $\left|\frac{x}{n}-p\right|<0{,}041$, azaz $\frac{x}{n}-0{,}041<p<\frac{x}{n}+0{,}041$. Figyelembe véve a megadott értékeket, $0{,}073<p<0{,}155$, azaz a nem hiteles aláírások száma 2993 és 6355 közé esik. Tehát a hiteles aláírások száma legalább 34645. A kapott eredmény természetesen 95%-os biztonsági küszöb mellett érvényes, azaz 95% a valószínűsége annak, hogy a hiteles aláírások száma ebbe az intervallumba esik.

Nézzük meg, mi történik akkor, ha nagyobb biztonsági küszöbbel számolunk!

Állapítsuk meg 99%-os biztonsággal a konfidencia intervallumot!

Ekkor $1-\frac{1}{4n\cdot\delta^2}=0{,}99$, innen $\delta\approx 0{,}0913$. Tehát 99%-os biztonsággal azt állíthatjuk, hogy $\left|\frac{x}{n}-p\right|<0{,}0913$, azaz $\frac{x}{n}-0{,}0913<p<\frac{x}{n}+0{,}0913$. Figyelembe véve a megadott értékeket, $0{,}023<p<0{,}2056$, azaz a nem hiteles aláírások száma 943 és 8429 közé esik. Tehát a hiteles aláírások száma legalább 32571.

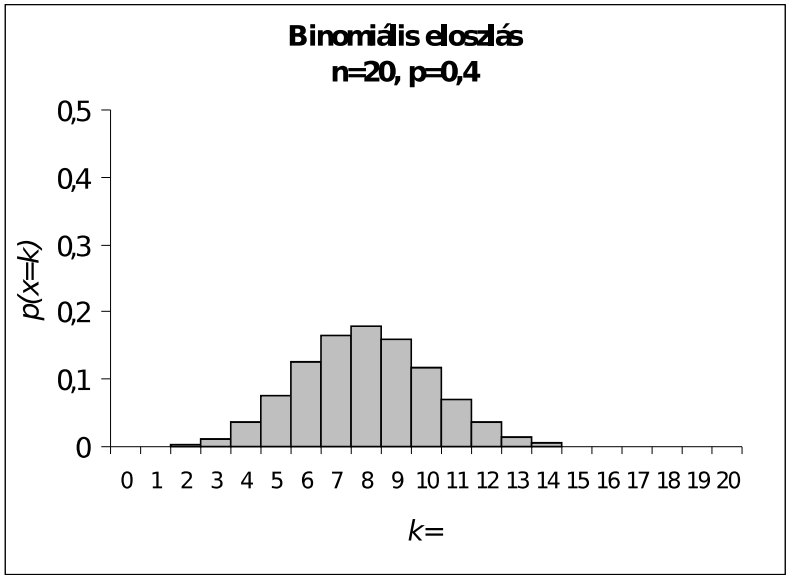
Látható tehát, hogy a biztonság növelésével a konfidencia-intervallum is növekszik, ami érthető is, hiszen ha nagyobb valószínűséggel, nagyobb biztonsággal akarjuk meghatározni egy adott mennyiség értékét, akkor szélesebb tartományt kell ehhez biztosítanunk.

Az is látható, hogy ez a módszerünk meglehetősen pontatlan, hiszen a második esetben p értékére csak egy 0,1826 hosszúságú intervallumot tudtunk biztosítani. Felmerülhet a kérdés, hogy van-e más számítási eljárás, ami jobban használható, vagyis pontosabb eredményt ad? Van, de ennek elméleti háttere kicsit bonyolultabb, mint a fent ismertetett egyszerű becslési eljárás. A továbbiakban erről lesz szó.

III.3. A binomiális eloszlású változó és módosítása (standardizálás)

Ha egy binomiális eloszlású valószínűségi változó egyes felvehető értékeihez tartozó valószínűségeket grafikonon próbáljuk ábrázolni, akkor azt tapasztalhatjuk, hogy a valószínűségek

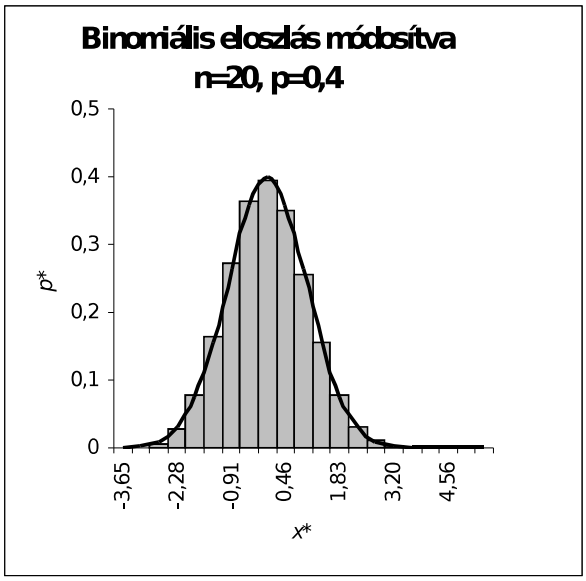
maximuma a várható érték környékén van, és minél nagyobb a valószínűségi változó szórása, annál jobban „szétfolyik” a grafikon. Az illusztrációban $n = 20$ és $p = 0{,}4$ esetén látható x értékének függvényében a $p(x = k)$ valószínűség.



A grafikonon oszlopdiagram formájában jelenítettük meg az egyes valószínűség-értékeket. Mivel minden oszlop szélessége egységnyi, ezért egy oszlop területe éppen annak valószínűségét adja meg, hogy a valószínűségi változó értéke az adott, x tengelyen levő értékkel egyezik meg. Emiatt, ha pl. arra vagyunk kíváncsiak, hogy a valószínűségi változó értéke milyen valószínűséggel esik két adott érték közé, akkor semmi egyebet nem kell tennünk, mint a két adott érték közé eső oszlopok területét összeadni. (Természetesen ez egyelőre semmi egyebet nem jelent, mint hogy a megfelelő valószínűségeket kell összeadnunk.)

Végezzük el a következő átalakítást: x helyett térjünk át az $x^* = \frac{x - E(x)}{D(x)}$ alakra, a $p(x = k)$ valószínűség helyett pedig a $p^*_k = D(x) \cdot p(x = k)$ értékre. Ezzel tulajdonképpen a koordinátarendszerünkben a téglalapok szélességét összenyomtuk $\frac{1}{D(x)}$-szeresükre, magasságukat pedig $D(x)$-szeresükre változtattuk, tehát a téglalapok területe nem változott. ($E(x)$ levonása a koordináta-rendszerben csak eltolást jelent.) Tehát továbbra is fennáll, hogy az x valószínűségi változó két adott szám közé esésének valószínűsége a megfelelő téglalapok területének összege. Természetesen figyelembe kell venni, hogy x módosítása miatt azon értékek, melyekhez tartozó téglalapokat figyelembe kell vennünk az új grafikonon, x-hez hasonlóan változnak. Ha pl. az eredeti grafikonon az a és b értékek közti téglalapok területének összegére volt

szükségünk, akkor a módosított grafikonon az $\frac{a-E(x)}{D(x)}$ és $\frac{b-E(x)}{D(x)}$ értékek közti téglalapok területének összegét kell kiszámítanunk.

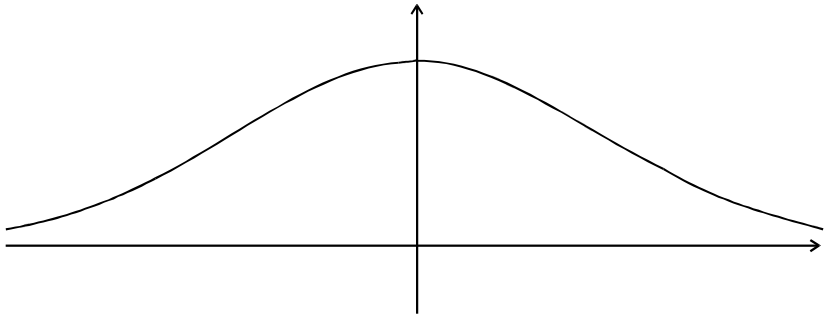


Az így kapott grafikon azonban a megfigyelések szerint nagyon jól illeszkedik a $\varphi(x)=\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2}}$ függvény grafikonjához (ez utóbbit a koordinátarendszerben vastag vonallal jelöltük). Mire jó ez a megfigyelés? Azt mondtuk, hogy az *x* valószínűségi változó két adott szám közé esésének valószínűsége a megfelelő téglalapok területének összege. Itt a téglalapok területének összegét helyettesíthetjük a $\varphi(x)$ függvény grafikonja alatti terület megfelelő értékek közt vett nagyságával. Természetesen felmerül a kérdés, hogy ez milyen hibát okoz. A részletesebb vizsgálatok úgy találták, hogy ha teljesül az úgynevezett Laplace-feltétel, mely szerint $n\cdot p\cdot(1-p)>9$, akkor az *n, p* paraméterű binomiális eloszlás módosított értékei helyett a számolásban használhatjuk fent említett függvényt. Ez sokkal pontosabb eredményt ad, mint az eddigi számításaink.

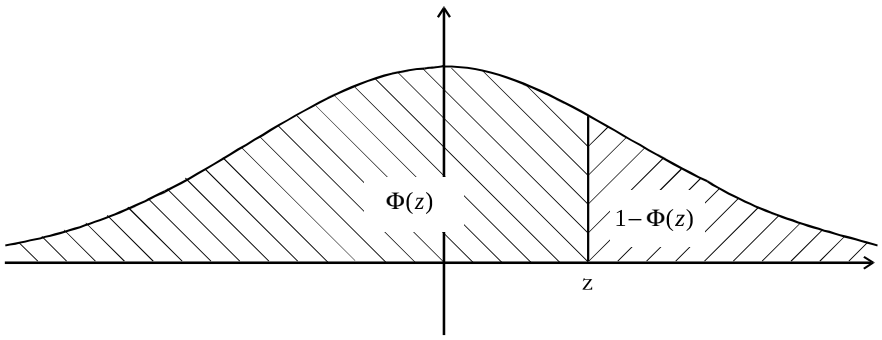
(Megjegyzésként megemlítjük, hogy az illusztrációként használt esetben még nem teljesül a Laplace-feltétel, de a szemléltetés eredményessége miatt célszerűbb volt kis *n* értéket választani, és már itt is jól látható a közelítés pontossága.)

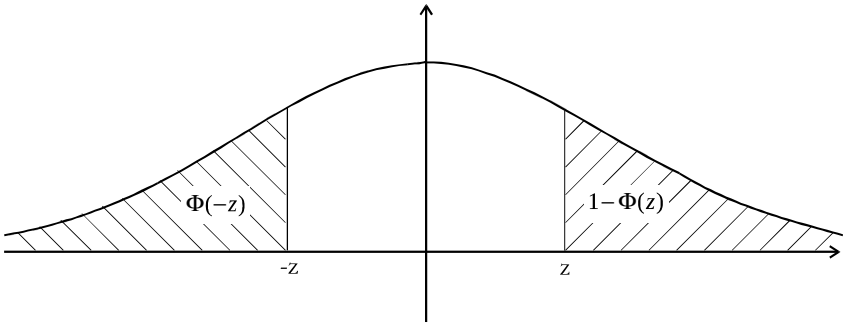
III.4. A φ(*x*) függvény grafikonja alatti terület számolása

A $\varphi(x)$ függvény grafikonja (az úgynevezett „haranggörbe”) alatti terület meghatározása sajnos nem megy teljesen simán, néhány tulajdonságát felhasználva azonban a számolás leegyszerűsíthető. A grafikon alatti területet $-\infty$-től egy adott z értékig táblázatba foglalva adjuk meg. (Sajnos nincs olyan elemi függvény, amely ennek a területnek a nagyságát z függvényében meg tudná adni.) A táblázatban feltüntetett értékek adják a $\Phi(z)$ függvényt, a továbbiakban ezzel fogunk dolgozni. Milyen értékeket kell megadni? Célszerű ehhez megismerkedni a $\varphi(x)$ függvény grafikonjával (az *x* és *y* tengelyen az egységek nem egyeznek, a jobb láthatóság kedvéért):



Látható a képlet alapján is, hogy ez a grafikon szimmetrikus az *y* tengelyre. Az egész görbe alatti terület 1 (ezt integrálszámítással is lehetne igazolni, de szemléletesen is látszik: az egész grafikon alatti terület nagysága azt jelenti, hogy minden lehetséges valószínűséget összeadunk a valószínűségi változónk vizsgálatakor; ennek 1-nek kell lennie, hiszen a valószínűségi változó minden lehetséges értékének együttese a biztos eseményt adja), ezért elég csak pozitív *z* esetén megadni $\Phi(z)$ értékét. Ekkor ugyanis az ábrán vonalkázott területek jelölt nagyságviszonya illetve egyenlősége miatt $\Phi(-z)=1-\Phi(z)$, ha $z\geq 0$.





A táblázatban szerencsére nem kell túl sok értéket feltüntetni, hiszen a görbe viszonylag meredek lefutása miatt $\Phi(4)=0{,}9999$, ami már gyakorlatilag 1.

Könnyen látható továbbá az is, hogy a grafikon alatti terület nagysága az *x* tengelyen vett *a* és *b* értékek között $\Phi(b)-\Phi(a)$ (feltéve, hogy $b \geq a$).

Térjünk vissza arra a két példára, amellyel korábban a Csebisev-egyenlőtlenséggel való becslésnél foglalkoztunk.

Hányszor kell feldobni egy pénzérmét, hogy 99%-os biztonsággal az „objektív” valószínűsége a fejdobásnak ne térjen el a relatív gyakoriságtól 0,01-nál jobban? Adjunk becslési módszert a fejdobás valószínűségére a kapott eredmény alapján!

Itt binomiális eloszlásról van szó, hiszen *n* független kísérletet vizsgálunk, ahol minden kísérletben a fejdobás azonos valószínűséggel következik be. A feladatban megfogalmazott feltétel azt jelenti, hogy annak valószínűsége, hogy a fejdobás relatív gyakorisága a fejdobás valószínűségétől legfeljebb 0,01-ra eltér, legalább 99%, azaz 0,99. Azaz keressük, hogy milyen *n*-re teljesül a $\frac{k}{n}$ relatív gyakoriságra

$$p\left(\left|\frac{k}{n}-p\right| \leq 0{,}01\right) \geq 0{,}99$$

Alakítsuk át a zárójelen belüli egyenlőtlenséget!

$$\left|\frac{k}{n}-p\right| \leq 0{,}01$$

$|k-np| \leq 0{,}01n$, azaz

$$np-0{,}01n \leq k \leq np+0{,}01n$$

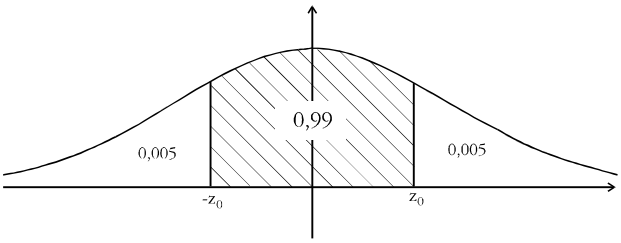
Itt *k* binomiális eloszlású valószínűségi változó, melynek várható értéke éppen *np*, és arra vagyunk kíváncsiak, hogy milyen értékek közé esik 99%-os valószínűséggel. Így a korábban mondott átalakítás segítségével éppen olyan alakot kaphatunk, melyet a $\Phi(z)$ értékek segítségével becsülhetünk!

Ha $np-0{,}01n \le k \le np+0{,}01n$ 99%-os valószínűséggel teljesül, akkor a binomiális eloszlása grafikonján az $np-0{,}01n$ és $np+0{,}01n$ közti oszlopok területének összege 0,99. Ennek megfelelően viszont az átalakított

$$-\frac{0{,}01n}{\sqrt{np(1-p)}} \le \frac{k-np}{\sqrt{np(1-p)}} \le \frac{0{,}01n}{\sqrt{np(1-p)}}$$

egyenlőtlenségben a két szélén szereplő érték közti, $\varphi(z)$ grafikonja alatti terület értéke 0,99.

Úgy kell tehát megválasztanunk az egyenlőtlenség két szélén levő számok értékét, hogy a haranggörbe alatti terület e két érték között legalább 0,99 legyen. Legyen $z_0 = \frac{0{,}01n}{\sqrt{np(1-p)}}$. Rajzoljuk fel a görbét! A szimmetria miatt a keresett tartomány két oldalán levő területek megegyeznek, tehát ezek nagysága $\frac{1-0{,}99}{2} = 0{,}005$.



Az ábra alapján tehát a legkisebb olyan z_0 értéket kell megkeresnünk a $\Phi(z)$ táblázatból, melyre teljesül, hogy $\Phi(z_0) \ge 0{,}995$. Ez két tizedesjegy pontosságra $z_0 = 2{,}58$. Tehát $\frac{0{,}01n}{\sqrt{np(1-p)}} \ge 2{,}58$, azaz négyzetre emelve és a nevezővel beszorozva $10^{-4}n^2 \ge 6{,}6564np(1-p)$, majd rendezve $n \ge 66564p(1-p)$. Korábban láttuk már, hogy $p(1-p)$ maximuma 0,25, tehát ha teljesül a $n \ge 66564 \cdot 0{,}25$ egyenlőtlenség, akkor biztosan teljesül a mi általunk kívánt feltétel is. Ez azt jelenti, hogy $n \ge 16641$, tehát legalább ennyi dobást kell végrehajtanunk a becsléshez.

Innen kezdve ugyanazt az eljárást lehet alkalmazni, mint ahogy azt korábban tettük: Végezzünk sok ilyen kísérletet, nézzük meg, hogy ezekben mekkora a relatív gyakorisága a fejdobásnak. Mivel ez a fejdobás tényleges valószínűségétől nagy valószínűséggel nem tér el 0,01-nél jobban, ezért a kapott relatív gyakoriságokat fedjük le egy 0,02 hosszúságú intervallummal úgy, hogy minél

kevesebb „lógjon ki” az adatok közül, és ennek a lefedő intervallumnak a közepét tudjuk használni a fejdobás valószínűségének becsléséhez.

Látható, hogy a most alkalmazott becslésnél sokkal kisebb számot kaptunk a kísérletek számára, mint a Csebisev-egyenlőtlenségnél, ezért úgy tűnik, hogy a gyakorlati életben sokkal jobban lehet használni ezt a módszert, mint a másikat.

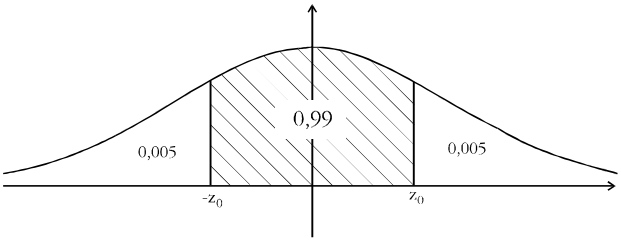
III.5. Konfidencia intervallum haranggörbés becsléssel

Népszavazáshoz gyűjtöttek aláírást egy országban. Az első kontroll után (kiválogatták azokat az aláírásokat, melyek mellől valamely a hitelesítéshez szükséges adat hiányzott, illetve azokat, melyek többször szerepeltek) maradt 41000 aláírás. Ebből kiválasztottak 3000-et véletlenszerűen, és ezek hitelességét ellenőrizték. Azt találták, hogy ezek közül 343 nem volt hiteles. Mit mondhatunk a 41000 aláírás között levő hiteles aláírások számáról 99%-os biztonsággal?

A feladat a Csebisev-egyenlőtlenségnél megfogalmazott problémát veti fel, megoldásához használjuk a haranggörbe alatti terület segítségével történő valószínűségi becslést. (Mivel ez itt egy visszatevés nélküli mintavétel, ezért ez hipergeometrikus eloszlású, de a korábban mondottak alapján a binomiális eloszlással közelíthető.)

A kiválasztott nem hiteles aláírások száma $x = 343$, a kiválasztott összes aláírások („kísérletek”) száma: $n = 3000$. Mivel az aláírásoknak csak kis hányadát vizsgáljuk, jó közelítéssel kezelhetjük úgy a problémát, hogy minden egyes lépésben a nem hiteles aláírás választásának valószínűsége p, ahol $p = \frac{\text{nem hiteles aláírások száma}}{41000}$. Ha 99%-os biztonságot adunk meg, akkor az $\frac{x}{n}$ relatív gyakoriság ekkora valószínűséggel a $[p-\delta, p+\delta]$ intervallumba fog esni, azaz $p-\delta \le \frac{x}{n} \le p+\delta$, vagyis $np-n\delta \le x \le np+n\delta$ 99%-os valószínűséggel teljesül. Mivel x binomiális eloszlású valószínűségi változó, ezért a megfelelő téglalapok területének összege $np-n\delta$ és $np+n\delta$ értékek között 0,99 a binomiális eloszlás oszlopdiagramján. Ez azt jelenti, hogy a transzformált változónál a $\frac{-n\delta}{\sqrt{np(1-p)}} \le \frac{x-np}{\sqrt{np(1-p)}} \le \frac{n\delta}{\sqrt{np(1-p)}}$ egyenlőtlenség áll fenn, vagyis a haranggörbe alatti terület $\frac{-n\delta}{\sqrt{np(1-p)}}$ és $\frac{n\delta}{\sqrt{np(1-p)}}$ értékek között 0,99 nagyságú.

Úgy kell tehát megválasztanunk az egyenlőtlenség két szélén levő számok értékét, hogy a haranggörbe alatti terület e két érték között 0,99 legyen. Legyen $z_0 = \frac{n\delta}{\sqrt{np(1-p)}}$. Rajzoljuk fel a görbét! A szimmetria miatt a keresett tartomány két oldalán levő területek megegyeznek, tehát ezek nagysága $\frac{1-0{,}99}{2} = 0{,}005$.



Az ábra alapján tehát a legkisebb olyan z_0 értéket kell megkeresnünk a $\Phi(z)$ táblázatból, melyre teljesül, hogy $\Phi(z_0) = 0{,}995$. Ez két tizedesjegy pontosságra $z_0 = 2{,}58$. Tehát $\frac{n\delta}{\sqrt{np(1-p)}} \geq 2{,}58$, rendezve $\delta \geq 2{,}58\sqrt{\frac{p(1-p)}{n}}$. Korábban láttuk már, hogy $p(1-p)$ maximuma 0,25, és használjuk fel, hogy *n* értéke 3000, tehát ha teljesül a $\delta = 2{,}58\sqrt{\frac{0{,}25}{3000}}$ egyenlőtlenség, akkor a mi általunk kívánt feltétel is. Ez azt jelenti, hogy $\delta \approx 0{,}0235$.

Azokat a *p* értékeket tartjuk tehát 99%-os valószínűséggel „hihetőnek” (azaz azok a *p* értékek alkotják a konfidencia-intervallumot), melyekre $p - 0{,}02355 \leq \frac{x}{n} \leq p + 0{,}0235$, azaz $\frac{x}{n} - 0{,}02355 \leq p \leq \frac{x}{n} + 0{,}0235$. Mivel $\frac{x}{n} \approx 0{,}1143$, ezért $0{,}0908 \leq p \leq 0{,}1378$, azaz a nem hiteles aláírások száma 3723 és 5654 közé esik, tehát a hiteleseké legalább 35346 (99%-os biztonsággal).

A korábban a Csebisev-egyenlőtlenségből kapott eredményünk az volt (99%-os biztonsággal), hogy a nem hiteles aláírások száma 943 és 8429 közé esik. Itt láthatóan pontosabb eredményt értünk el, pedig még itt is az ismeretlen *p* értéket egy ponton becsültük, a $p(1-p) \leq \frac{1}{4}$ felhasználásakor. Sőt, még a Csebisev-egyenlőtlenség felhasználásával kisebb, 95%-os biztonsággal kapott eredményünknél is kisebb intervallumot eredményezett a mostani számolásunk. Ez rámutat arra, hogy ez a módszer még nagyobb biztonság előírása esetén is pontosabb értéket ad, mint a korábbi

módszerünk alacsonyabb biztonsági küszöb mellett. (Ezt annak tudatában kell értékelni, hogy a biztonsági küszöb növelése az esetek nagy többségében a konfidencia intervallum szélesedésével jár.)

Ha még pontosabban akarunk számolni, akkor a kapott $\delta \geq z_0\sqrt{\frac{p(1-p)}{n}}$ ($z_0 = 2{,}58$) egyenlőtlenségből vegyük a még legkisebb megfelelő δ értéket, $\delta = z_0\sqrt{\frac{p(1-p)}{n}}$, és ezt helyettesítsük vissza az eredetileg felírt $p-\delta \leq \frac{x}{n} \leq p+\delta$, azaz $\left|\frac{x}{n}-p\right| \leq \delta$ egyenlőtlenségbe:

$$\left|\frac{x}{n}-p\right| \leq z_0\sqrt{\frac{p(1-p)}{n}}$$

Négyzetre emelés és rendezés után egy másodfokú egyenlőtlenséget kapunk:

$$\left(1+\frac{z_0^2}{n}\right)\cdot p^2 - \frac{2x+z_0^2}{n}\cdot p + \frac{x^2}{n^2} \leq 0$$

A szereplő másodfokú kifejezés főegyütthatója garantáltan pozitív, tehát ennek az egyenlőtlenségnek a megoldása mindig egy intervallum; mivel *p*-nek ide kell esnie, ezért ez éppen a keresett konfidencia-intervallum.

Megoldva a konkrét esetben a másodfokú egyenlőtlenséget, azt kapjuk, hogy $0{,}1 \leq p \leq 0{,}130$, azaz a nem hiteles aláírások száma 4100 és 5342 közé esik, tehát legalább 35658 hiteles szavazat van a 41000 szavazat között 99%-os biztonsággal. Látszik, hogy az eredmény pontosabb lett, de talán annyival nem, mint amennyivel többet kellett érte számolni. Ha gyorsabban, de elég pontosan akarunk számolni, érdemesebb az első módszert választani, a beépített becsléssel; ha mindenképpen a pontosabb eredményre van szükségünk, akkor érdemes csak a második módszert választani.

Megjegyzésként megemlítjük, hogy éppen az illusztrációként használt példában csak egy meghatározott minimum fölé kell mennie a hiteles aláírásoknak, hogy az aláírásgyűjtés értelmet nyerjen. Ennek megfelelően ha ezt a szintet már eléri a becslés a durvább módszerekkel is, akkor felesleges még egyszer finomítani rajta.

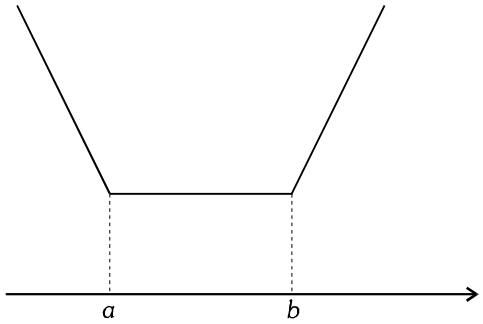
$\Phi(z)$ táblázat

x	0	1	2	3	4	5	6	7	8	9
0,00	0,50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0,10	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0,20	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0,30	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0,40	65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0,50	0,69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0,60	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0,70	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0,80	78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0,90	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1,00	0,84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1,10	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,20	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,30	90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1,40	91294	92073	92220	92364	92507	92647	92785	92922	93056	93189
1,50	0,93319	93448	93574	93699	93822	93493	94062	94179	94295	94408
1,60	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,70	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1,80	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,90	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2,00	0,97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,10	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,20	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,30	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,40	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,50	0,99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,60	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,70	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,80	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,90	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3,00	0,99865	99869	99874	99878	99882	99886	99889	99893	99897	99900
3,10	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,20	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,30	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,40	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3,50	0,99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3,60	99984	99985	99985	99986	99986	99987	99988	99988	99988	99989
3,70	99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3,80	99993	99993	99993	99994	99994	99994	99995	99995	99995	99995
3,90	99995	99996	99996	99996	99996	99996	99996	99996	99997	99997
4,00	0, 99997	99997	99997	99997	99997	99997	99998	99998	99998	99998

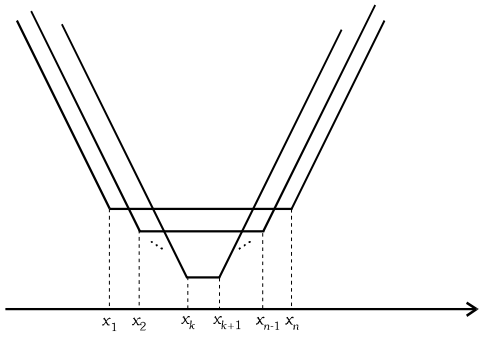
IV. Függelék

IV.1. Átlagos abszolút eltérés a mediánra minimális

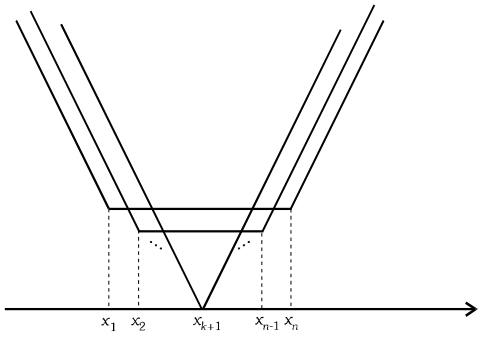
Ha ábrázoljuk az $|x-a|+|x-b|$ függvény grafikonját, akkor az alábbi jellegű ábrát kapjuk:



A „vödör” alja $b-a$ magasságban található. Ha most az $\frac{|x_1-\tilde{X}|+|x_2-\tilde{X}|+\ldots+|x_n-\tilde{X}|}{n}$ kifejezésben az adatok nagyságrendi sorrendbe rakva szerepelnek, és párosítjuk az első abszolút értékes kifejezést az utolsóval, a másodikat az utolsó előttivel, stb. és páronként ábrázoljuk a grafikonjukat, akkor a következő ábrát kapjuk:



Az adatok száma 2*k*



Az adatok száma 2*k*+1

Az egymásba csúsztatott „vödörforma” grafikonok alja azért kerül mindig lejjebb, mert a megfelelő *x* értékek különbsége egyre kisebb. Az ábráról leolvasható, hogy a minimális értéket minden pár esetén a középső adatnál kapjuk, ha az adatok száma páratlan, és a középső két adat közötti bármely érték minimumot ad, ha az adatok száma páros (ezért választották a két középső érték között félúton levő számot, azaz az átlagukat mediánnak).

IV.2. Átlagos négyzetes eltérés az átlagra minimális

Az átlagos négyzetes eltérést a $\sigma^2\left(\tilde{X}\right)=\frac{\left(x_1-\tilde{X}\right)^2+\left(x_2-\tilde{X}\right)^2+...+\left(x_n-\tilde{X}\right)^2}{n}$ képlettel adhatjuk meg. Végezzük el a számlálóban a négyzetre emelést, majd végezzük el az összevonásokat:

$$\sigma^2\left(\tilde{X}\right)=\frac{x_1^2-2x_1\tilde{X}+\tilde{X}^2+x_2^2-2x_2\tilde{X}+\tilde{X}^2+...+x_n^2-2x_n\tilde{X}+\tilde{X}^2}{n}=$$
$$=\tilde{X}^2-2\cdot\frac{x_1+x_2+...+x_n}{n}\cdot\tilde{X}+\frac{x_1^2+x_2^2+...+x_n^2}{n}=$$
$$=\left(\tilde{X}-\frac{x_1+x_2+...+x_n}{n}\right)^2-\left(\frac{x_1+x_2+...+x_n}{n}\right)^2+\frac{x_1^2+x_2^2+...+x_n^2}{n}$$

Mivel a kifejezés végén álló $-\left(\frac{x_1+x_2+...+x_n}{n}\right)^2+\frac{x_1^2+x_2^2+...+x_n^2}{n}$ mennyiség állandó (az adatok nagysága meghatározza), ezért a teljes kifejezés akkor lesz minimális, ha $\left(\tilde{X}-\frac{x_1+x_2+...+x_n}{n}\right)^2$ minimális, azaz ha $\tilde{X}=\frac{x_1+x_2+...+x_n}{n}$. Tehát az átlagos négyzetes eltérés valóban az átlag esetén minimális.

IV.3. Kockapóker eredményeinek valószínűsége

1 pár:

a) *számít a dobási sorrend:* 6-féle számból kerülhet ki a két egyező szám, ezeket $\binom{5}{2}$-féleképpen lehet elhelyezni, és mellé a maradék 5-féle számból kell 3 különbözőt kiválasztani a sorrend figyelembevételével: $6\cdot\binom{5}{2}\cdot5\cdot4\cdot3=3600$

b) *nem számít a dobási sorrend:* 6-féle számból kerülhet ki a két egyező szám, a maradék 5-ből pedig 3 különbözőt kell kiválasztani úgy, hogy nem számít a sorrend: $6\cdot\binom{5}{3}=60$

2 pár:

a) *számít a dobási sorrend:* $\binom{6}{2}$-féleképpen lehet kiválasztani a 2-2 egyező számot, az első párt

$\binom{5}{2}$-féleképpen lehet elhelyezni, a második párt $\binom{3}{2}$ helyre lehet elhelyezni, a maradék 4-ből pedig 1-et kell kiválasztani, ami a megmaradt helyre kerül.: $\binom{6}{2}\cdot\binom{5}{2}\cdot\binom{3}{2}\cdot 4 = 1800$

b) *nem számít a dobási sorrend:* $\binom{6}{2}$-féleképpen lehet kiválasztani a 2-2 egyező számot, a maradék 4-ből pedig 1-et kell kiválasztani: $\binom{6}{2}\cdot 4 = 60$

Terc:

a) *számít a dobási sorrend:* 6-féle számból kerülhet ki a 3 egyező szám, ezeket $\binom{5}{3}$-féleképpen lehet elhelyezni, és mellé a maradék 5-féle számból kell 2 különbözőt kiválasztani a sorrend figyelembevételével: $6\cdot\binom{5}{3}\cdot 5\cdot 4 = 1200$

b) *nem számít a dobási sorrend:* 6-féle számból kerülhet ki a 3 egyező szám, a maradék 5-ből pedig 2 különbözőt kell kiválasztani úgy, hogy nem számít a sorrend: $6\cdot\binom{5}{2} = 60$

Sor:

a) *számít a dobási sorrend:* 2-féle számötös lehet, ezeket kell sorbarendezni, ha számít a sorrend: $2\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1 = 240$

b) *nem számít a dobási sorrend:* 2-féle számötös lehet: 2

Full:

a) *számít a dobási sorrend:* 6-féle számból kerülhet ki a 3 egyező szám, ezeket $\binom{5}{3}$-féleképpen lehet elhelyezni, és mellé a maradék 5-féle számból kell a másik 2 egyezőt kiválasztani: $6\cdot\binom{5}{3}\cdot 5 = 300$

b) *nem számít a dobási sorrend:* 6-féle számból kerülhet ki a 3 egyező szám, és a maradék 5-ből kell kiválasztani a másik 2 egyezőt: $6\cdot 5 = 30$

Póker:

a) *számít a dobási sorrend:* 6-féle számból kerülhet ki a 4 egyező szám, ezeket $\binom{5}{4}$-féleképpen

lehet elhelyezni, és mellé a maradék 5-féle számból kell 1-et kiválasztani: $6 \cdot \binom{5}{4} \cdot 5 = 150$

b) *nem számít a dobási sorrend:* 6-féle számból kerülhet ki a 4 egyező szám, a maradék 5-ből pedig 1-et kell kiválasztani: $6 \cdot 5 = 30$

Royal póker:

a) *számít a dobási sorrend:* 6-féle számból kerülhet ki az 5 egyező szám: 6

b) *nem számít a dobási sorrend:* 6-féle számból kerülhet ki az 5 egyező szám: 6

IV.4. Binomiális eloszlás várható értéke formálisan

Számítsuk ki a binomiális eloszlású valószínűségi változó várható értékét a definíció alapján!

$E(x) = \sum_{k=0}^{n} k \cdot p(x=k) = \sum_{k=1}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$ (a $k = 0$ eset azért hagyható ki, mert az összegzésben ez egy 0-t jelent, ami nem változtatja meg az összeg értékét.)

$$E(x) = \sum_{k=1}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \sum_{k=1}^{n} k \cdot \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} =$$

$$= \sum_{k=1}^{n} n \cdot \frac{(n-1)!}{(k-1)! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} = n \cdot p \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)! \cdot ([n-1]-[k-1])!} \cdot p^{k-1} \cdot (1-p)^{[n-1]-[k-1]} =$$

$$= n \cdot p \cdot \sum_{k=1}^{n} \binom{n-1}{k-1} \cdot p^{k-1} \cdot (1-p)^{[n-1]-[k-1]} = n \cdot p \cdot \sum_{k'=0}^{n-1} \binom{n-1}{k'} \cdot p^{k'} \cdot (1-p)^{[n-1]-k'} =$$

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

IV.5. Binomiális eloszlás szórása formálisan

Számítsuk ki a binomiális eloszlású valószínűségi változó szórásnégyzetét a definíció alapján!

$D^2(x) = E(x^2) - [E(x)]^2$. Az előzőekben már láttuk, hogy $E(x) = np$, tehát $[E(x)]^2 = n^2 p^2$, így csak $E(x^2)$ értékét kell meghatároznunk. $E(x^2) = \sum_{k=0}^{n} k^2 \cdot p(x=k) = \sum_{k=1}^{n} k^2 \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$

(a $k = 0$ eset azért hagyható ki, mert az összegzésben ez egy 0-t jelent, ami nem változtatja meg az összeg értékét.)

$$E(x^2)=\sum_{k=1}^{n}k^2\cdot\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}=\sum_{k=1}^{n}k^2\cdot\frac{n!}{k!\cdot(n-k)!}\cdot p^k\cdot(1-p)^{n-k}=$$

$$=\sum_{k=1}^{n}n\cdot k\cdot\frac{(n-1)!}{(k-1)!\cdot(n-k)!}\cdot p^k\cdot(1-p)^{n-k}=n\cdot p\cdot\sum_{k=1}^{n}k\cdot\binom{n-1}{k-1}\cdot p^{k-1}\cdot(1-p)^{[n-1]-[k-1]}=$$

$$=n\cdot p\cdot\sum_{k'=0}^{n-1}(k'+1)\cdot\binom{n-1}{k'}\cdot p^{k'}\cdot(1-p)^{[n-1]-k'}=$$

$$=n\cdot p\cdot\sum_{k'=0}^{n-1}k'\cdot\binom{n-1}{k'}\cdot p^{k'}\cdot(1-p)^{[n-1]-k'}+n\cdot p\cdot\sum_{k'=0}^{n-1}\binom{n-1}{k'}\cdot p^{k'}\cdot(1-p)^{[n-1]-k'}$$

A kapott összeg második tagjá értékét már az előző pontban meghatároztuk, ez *np*, az első tag értéke pedig az előző pontban közölt számoláshoz hasonlóan számítható:

$$n\cdot p\cdot\sum_{k'=0}^{n-1}k'\cdot\binom{n-1}{k'}\cdot p^{k'}\cdot(1-p)^{[n-1]-k'}=n\cdot p\cdot\sum_{k'=1}^{n-1}k'\cdot\binom{n-1}{k'}\cdot p^{k'}\cdot(1-p)^{[n-1]-k'}=$$

$$=n\cdot p\cdot\sum_{k'=1}^{n-1}k'\cdot\frac{(n-1)!}{k'!\cdot([n-1]-k')!}\cdot p^{k'}\cdot(1-p)^{[n-1]-k'}=$$

$$=n\cdot p\cdot\sum_{k'=1}^{n-1}(n-1)\cdot\frac{(n-2)!}{(k'-1)!\cdot([n-2]-[k'-1])!}\cdot p^{k}\cdot(1-p)^{[n-1]-k'}=$$

$$=n\cdot p\cdot(n-1)\cdot p\cdot\sum_{k'=1}^{n-1}\binom{n-2}{k'-1}\cdot p^{k'-1}\cdot(1-p)^{[n-2]-[k'-1]}=n\cdot p\cdot(n-1)\cdot p\cdot\sum_{k''=0}^{n-2}\binom{n-2}{k''}\cdot p^{k''}\cdot(1-p)^{[n-2]-k''}=$$

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

Tehát $D^2(x)=E(x^2)-[E(x)]^2=n(n-1)p^2+np-(np)^2=np-np^2=np(1-p)$, tehát a binomiális eloszlású valószínűségi változó szórása $D(x)=\sqrt{np(1-p)}$.

IV.6. Markov-egyenlőtlenség bizonyítása

Markov-egyenlőtlenség:

Ha x nemnegatív értékű valószínűségi változó, és $\mu>1$, akkor

$$p[x>\mu\cdot E(x)]<\frac{1}{\mu}$$

Biz: Legyen A az az esemény, amikor a valószínűségi változó értéke nagyobb $\mu \cdot E(x)$-nél. Ekkor $p[x > \mu \cdot E(x)] = p(A)$. Másrészt $E(x) = x_1 p_1 + x_2 p_2 + ... + x_r p_r$, ahol $p_i = p(x = x_i)$. Helyettesítsük az összes, $\mu \cdot E(x)$-nél nem nagyobb x_i-t $\mu \cdot E(x)$-szel, a többit pedig 0-val. Ekkor az összefüggést átalakítva:

$$E(x) > \mu \cdot E(x) \cdot p_1 + \mu \cdot E(x) \cdot p_2 + ... + \mu \cdot E(x) \cdot p_s = \mu \cdot E(x) \cdot (p_1 + p_2 + ... + p_s) = \mu \cdot E(x) \cdot p(A)$$ In

nen $E(x) > \mu \cdot E(x) \cdot p(A)$, azaz $\frac{1}{\mu} > p(A)$, és ez volt a bizonyítandó állítás.

IV.7. A Csebisev-egyenlőtlenség bizonyítása

Csebisev egyenlőtlenség:

$$p\left(|x - E(x)| \geq \lambda \cdot D(x)\right) < \frac{1}{\lambda^2}$$

Biz: Írjuk fel a Markov-egyenlőtlenséget $(x - E(x))^2$-re!

$$p\left[(x - E(x))^2 > \mu \cdot E\left[(x - E(x))^2\right]\right] < \frac{1}{\mu}$$

mivel $E\left[(x - E(x))^2\right] = D^2(x)$, ezért a zárójelben levő egyenlőtlenségből gyökvonással

$$p\left(|x - E(x)| \geq \sqrt{\mu} \cdot D(x)\right) < \frac{1}{\mu},$$

és $\sqrt{\mu} = \lambda$ jelöléssel $p\left(|x - E(x)| \geq \lambda \cdot D(x)\right) < \frac{1}{\lambda^2}$.

IV.8. $\Phi(z)$ táblázat

x	0	1	2	3	4	5	6	7	8	9
0,00	0,50000	50399	50798	51197	51595	51994	52392	52790	53188	53586
0,10	53983	54380	54776	55172	55567	55962	56356	56749	57142	57535
0,20	57926	58317	58706	59095	59483	59871	60257	60642	61026	61409
0,30	61791	62172	62552	62930	63307	63683	64058	64431	64803	65173
0,40	65542	65910	66276	66640	67003	67364	67724	68082	68439	68793
0,50	0,69146	69497	69847	70194	70540	70884	71226	71566	71904	72240
0,60	72575	72907	73237	73565	73891	74215	74537	74857	75175	75490
0,70	75804	76115	76424	76730	77035	77337	77637	77935	78230	78524
0,80	78814	79103	79389	79673	79955	80234	80511	80785	81057	81327
0,90	81594	81859	82121	82381	82639	82894	83147	83398	83646	83891
1,00	0,84134	84375	84614	84849	85083	85314	85543	85769	85993	86214
1,10	86433	86650	86864	87076	87286	87493	87698	87900	88100	88298
1,20	88493	88686	88877	89065	89251	89435	89617	89796	89973	90147
1,30	90320	90490	90658	90824	90988	91149	91309	91466	91621	91774
1,40	91294	92073	92220	92364	92507	92647	92785	92922	93056	93189
1,50	0,93319	93448	93574	93699	93822	93493	94062	94179	94295	94408
1,60	94520	94630	94738	94845	94950	95053	95154	95254	95352	95449
1,70	95543	95637	95728	95818	95907	95994	96080	96164	96246	96327
1,80	96407	96485	96562	96638	96712	96784	96856	96926	96995	97062
1,90	97128	97193	97257	97320	97381	97441	97500	97558	97615	97670
2,00	0,97725	97778	97831	97882	97932	97982	98030	98077	98124	98169
2,10	98214	98257	98300	98341	98382	98422	98461	98500	98537	98574
2,20	98610	98645	98679	98713	98745	98778	98809	98840	98870	98899
2,30	98928	98956	98983	99010	99036	99061	99086	99111	99134	99158
2,40	99180	99202	99224	99245	99266	99286	99305	99324	99343	99361
2,50	0,99379	99396	99413	99430	99446	99461	99477	99492	99506	99520
2,60	99534	99547	99560	99573	99585	99598	99609	99621	99632	99643
2,70	99653	99664	99674	99683	99693	99702	99711	99720	99728	99736
2,80	99744	99752	99760	99767	99774	99781	99788	99795	99801	99807
2,90	99813	99819	99825	99831	99836	99841	99846	99851	99856	99861
3,00	0,99865	99869	99874	99878	99882	99886	99889	99893	99897	99900
3,10	99903	99906	99910	99913	99916	99918	99921	99924	99926	99929
3,20	99931	99934	99936	99938	99940	99942	99944	99946	99948	99950
3,30	99952	99953	99955	99957	99958	99960	99961	99962	99964	99965
3,40	99966	99968	99969	99970	99971	99972	99973	99974	99975	99976
3,50	0,99977	99978	99978	99979	99980	99981	99981	99982	99983	99983
3,60	99984	99985	99985	99986	99986	99987	99988	99988	99988	99989
3,70	99989	99990	99990	99990	99991	99991	99992	99992	99992	99992
3,80	99993	99993	99993	99994	99994	99994	99995	99995	99995	99995
3,90	99995	99996	99996	99996	99996	99996	99996	99996	99997	99997
4,00	0, 99997	99997	99997	99997	99997	99997	99998	99998	99998	99998