Dr. Sándor Zsolt – Tánczos Levente József

Gazdasági statisztika jegyzet

T3 Kiadó, Sepsiszentgyörgy

T3 Kiadó, Sepsiszentgyörgy
ISBN: 978-973-1962-87-0

Dr. Sándor Zsolt – Tánczos Levente József

Gazdasági statisztika jegyzet

A kötetet lektorálta:

Dr. Fejér-Király Gergely

Dr. Oláh-Gál Róbert

Tartalomjegyzék

Bevezetés

Ez a jegyzet a madridi Carlos III Egyetemen és a Sapientia Erdélyi Magyar Tudományegyetemen oktatott alap szintű Gazdasági statisztika előadásra készült. A bevezetésben vázoljuk, hogy miről lesz szó a jegyzetben. Szó lesz röviden arról, hogy mi is tulajdonképpen és mivel foglalkozik a Gazdasági statisztika.

A Gazdasági statisztika a gazdasági jelenségek elemzése gazdasági adatok (megfigyelések) alapján. A gazdasági adatok a gazdasági mennyiségekkel kapcsolatos mérések. Megemlíthetjük például az inflációt, a munkanélküliségi arányt, nettó átlagfizetést, fogyasztói keresletet, termékek jellemzőit, fogyasztók jellemzőit, stb.

A gazdasági statisztika a gazdasági mennyiségek bizonyos lényeges vonásait könnyen feldolgozható információvá sűríti. Ezért a gazdasági statisztika nem szentel figyelmet az egyedi vonásoknak. Például, két személyt hasonlónak tekint, ha jövedelmük és családi helyzetük hasonló, vagy két termék hasonló, ha a lényeges jellemzőik hasonlók. Továbbá, két ország hasonló, ha a főbb makroökonómiai jellemzőik hasonlók. Ez a szemlélet hasznos, mert, például, két hasonló személy valószínűleg hasonló termékeket vásárol, vagy két hasonló országban hasonlóan térülnek meg a befektetések.

A jegyzet keretén belül a statisztika és a valószínűségszámítás alapjait fogjuk tanulmányozni. Az I. rész a statisztika, pontosabban leíró statisztika, míg a II. rész a valószínűségszámítás. A tárgyalandó példák és gyakorlatok gazdasági mennyiségekhez kapcsolódnak.

A **leíró statisztika** egy hasznos eszköztár a különböző gazdasági vállalatok és szervezetek számára. A gazdasági vállalatok és szervezetek a statisztika módszereit alkalmazzák gazdasági adatok (megfigyelések) gyűjtésére és felhasználására. Ennek célja:

- új ismeretek szerzése,
- előrejelzések,
- a döntéshozatal megkönnyítése.

Konkrétabban, bemutatjuk, hogy:

- Hogyan írjunk le gazdasági adatokat
  - hogyan **ábrázoljunk** adatokat úgy, hogy fontos dolgokat tudjunk meg róluk,
  - hogyan **foglaljuk össze** egy nagyszámú adatbázis lényegét néhány számszerű adatban,
  - hogyan végezzük el a fenti műveleteket számítógép segítségével.
- Hogyan írjunk le két különböző gazdasági mennyiség közötti összefüggést.
- Sok mennyiség egyetlen számadatba összegzése (például árindexek).
- A gazdasági adatok időbeni változásának tanulmányozása (például infláció).
- Hogyan végezzük el ezeket számítógép segítségével.

A **valószínűségszámítás** a véletlen és a bizonytalan tudománya, ami fontos szerepet játszik az életünkben. A legtöbb gazdasági mennyiség, amely előfordul a példákban, jövőbeni értéke ismeretlen, ezért bizonyos értékekre „tippelünk”, amikor döntéseket hozunk. Ezért a statisztikai módszerek a valószínűségszámítás elméletére épülnek. Valószínűségszámítás

segítségével meg lehet határozni, hogy egy bizonyos előrejelzés mekkora valószínűséggel következik be.

A valószínűségszámítás keretén belül szó lesz véletlenszerű eseményekről és ezek valószínűségeiről, valószínűségi változókról és az ezekhez kapcsolódó fontosabb törvényszerűségekről. Szó lesz még a legfontosabb diszkrét és folytonos valószínűségi változókról, amelyek segítségével gyakorlati szempontból fontos helyzeteket lehet modellezni.

I. Leíró statisztika

Miért hasznos a leíró statisztika és az ezekhez kapcsolódó ábrák?

- Amint a tárgyalt példák alapján meg fogjuk látni, a gazdasági megfigyelések számokkal vagy karakterekkel megadott listákban jelennek meg.
- Ezekből a listákból nehéz hasznos információt kiszűrni. A leíró statisztika olyan mennyiségeket határoz meg, amelyek hasznos információvá sűrítik ezeket a listákat. Ezeket a mennyiségeket mutatóknak nevezzük.
- Az ábrák a mutatók segítségével vizualizálják a listákban rejlő információt, és megkönnyítik azok felhasználását.

Ebben a részben kilenc fejezetben foglalkozunk a leíró statisztika legfontosabb fogalmaival és módszereivel. Az 1. fejezetben bemutatjuk a statisztika alapfogalmait, amelyek a sokaság és a minta, valamint tárgyaljuk a változók típusait. A 2. fejezetben a nominális és diszkrét változók leíró statisztikáját tárgyaljuk, míg a 3. fejezetben az ilyen típusú változók grafikus ábrázolását. A 4. fejezet átülteti ezeket az eljárásokat folytonos változókra, és ennek keretén belül bemutatjuk a hisztogram szerkesztését. Az 5. fejezet a legfontosabb statisztikai mutatókkal foglalkozik, és ezen belül több helyzet- és szóródási mutatóról és ezek leglényegesebb tulajdonságairól is szó lesz; itt mutatjuk be a doboz ábra szerkesztését is. A 6. fejezet két változó közötti összefüggés leíró statisztikáját tárgyalja, ami kiterjeszti az egy változóra tanult eljárásokat, és kibővíti olyan fogalmakkal és mutatókkal, amelyek az összefügésekre derítenek fényt. A 7. fejezet a regressziós egyenest mutatja be és ennek kapcsolatát az előző fejezet néhány fogalmával. A 8. fejezet az index-számokat tárgyalja az egyszerű eladási indextől az éves infláció kiszámításáig. A 9. fejezet idősor-elemzéssel foglalkozik, melynek keretén belül elsősorban a trend és szezonalitás grafikus elemzését mutatja be.

I.1. Sokaság, minta, változók

Ebben a fejezetben bevezetjük a statisztika legfontosabb alapfogalmait és a változók típusait.

A s**okaság** (vagy **populáció**) a tanulmányozandó mennyiségekhez tartozó megfigyelések összessége. Például, az összes fogyasztó jellemzői (havi fizetés, gyerekek száma, stb.), az összes termék jellemzői (autók ára, fogyasztása, lóereje, stb.). Egy sokaság rengeteg megfigyelésből áll, ezért nehéz tanulmányozni (költséges, számítógép-memória nem elég). Képzeljük el azt, például, hogy mennyire költséges lenne, ha a ruhagyártók az összes ember testméretét számon tartanák, és nyomon követnék. A sokaság helyett sok szempontból alkalmasabb csak egy részét tanulmányozni. Ez a **minta**: egy sokaság része amelyet tanulmányozásra szánunk. Példaként megemlítjük, hogy gyakran végeznek közvélemény-kutatást különböző témákról, ahol az embereknek csak egy részét kérdezik meg, ami a mintát képezi ebben az esetben.

Egy **változó** a sokaság egy bizonyos jellemzője (havi fizetés, gyerekek száma, autók ára, fogyasztása, lóereje, stb.). További példák változókra:

- Fogyasztók neme, családi állapota,
- Autók márkája, ajtók száma.

A **megfigyelés** (vagy **adat**) egy változó értéke a minta egyik elemére. Példaként tekintsük a következő megfigyeléseket egy bizonyos fogyasztóra:

- havi fizetés: 1550 lej
- neme: nő
- gyerekek száma: 1.

A fenti példákban szereplő változók bizonyos lényeges szempontból különböznek.

A változók típusai

Statisztikából a változóknak egy hasznos osztályozása aszerint történik, hogy mennyiségi szempontból van-e értelmük. Két típust különböztetünk meg: kvalitatív és kvantitatív változókat. A **kvalitatív** változók nem mennyiségre vonatkozó változók. Két csoportba soroljuk:

- **Nominális**: olyan változó amelynek értékei nem hasonlíthatók össze (semmilyen mértékegység segítségével). *Példák*: márka, nem, családi állapot.
- **Ordinális**: olyan változó amelynek értékei semmilyen mértékegységgel nem mérhetők, de sorrendbe állíthatók. *Példa*: termékek rangsorolása marketing-kérdőíveknél.

A **kvantitatív** változók bizonyos mennyiséget fejeznek ki. Kétféle kvantitatív változót különböztetünk meg: diszkrét vagy folytonost.

- **Diszkrét**: egy bizonyos számolási eljárás eredménye, tehát értéke 0,1,2,… lehet. *Példák*: gyerekek száma egy családban, autó ajtóinak a száma, egy bizonyos autótípust vásárló fogyasztók száma.
- **Folytonos**: sok esetben valamilyen mértékegységgel mérhető, értékei egy intervallum bármelyik eleme lehetnek. *Példák*: évi jövedelem lejben [0,500000], motor ereje lóerőben [60,500].

I.2 Nominális és diszkrét változók leíró statisztikája

Ebben a fejezetben együtt tárgyaljuk a nominális és diszkrét változók leíró statisztikáját. Vegyünk egy példát, amelyhez kilenc autótípus jellemzőit tüntetjük fel az alábbi táblázatban. Hogyan tudjuk legjobban összefoglalni a márka változót? Megszámoljuk, hogy hányszor fordul elő a megfigyelések között mindegyik márka.

Kategóriák

A márka nominális változónak 3 értéke van: *Citroen, Renault, Peugeot.* Az ár folytonos változó, a légzsák diszkrét változó: 1-től 5-ig vesz fel értékeket. Az automata vezérlés nominális változónak a *nem* és az *igen* az értékei. Egy nominális változó értékeit **kategóriáknak** (**osztályoknak**) nevezzük.

Táblázat 1.2.1. – Példa különböző típusú adatokra

Márka	Ár (€ x1000)	Légzsák	Automata vezérlés
Citroen	10	1	nem
Citroen	13.5	4	nem
Citroen	20	5	igen
Renault	16.5	3	igen
Renault	15	2	nem
Peugeot	15.5	4	nem
Peugeot	14.5	3	nem
Peugeot	19.5	5	igen
Peugeot	13	2	nem

Gyakoriságok

A táblázatban megfigyelhetjük, hogy a márka változó

- *Citroen* kategóriája 3-szor jelenik meg,
- *Renault* kategóriája 2-szer jelenik meg,
- *Peugeot* kategóriája 4-szer jelenik meg.

Azt, hogy egy bizonyos kategória hányszor jelenik meg a kategória **abszolút gyakoriságának** nevezzük. *Tehát a Citroen kategória abszolút gyakorisága 3.* A márka változó kategóriáinak gyakorisága a változó értékeiről az **összes információt** tartalmazza.

A gyakoriság-eloszlás

Egy kategória gyakoriságának arányát az összes megfigyelés gyakoriságához viszonyítva **relatív gyakoriságnak** nevezzük. *Mivel 9 autótípus megfigyelésünk van, a Citroen relatív gyakorisága 3/9=1/3.* A **gyakoriság-eloszlás** az összes kategória relatív gyakoriságainak összessége. A márka gyakoriság-eloszlása a Citroen, Renault, Peugot kategóriákra (1/3, 2/9, 4/9).

A halmozott gyakoriság-eloszlás

Vegyük az első kategória abszolút gyakoriságát, és az első két kategória abszolút gyakoriságát adjuk össze, majd ehhez adjuk hozzá a harmadik gyakoriságot, és így tovább. Az így kapott számokat **halmozott abszolút gyakoriságoknak** nevezzük. A márkák esetében a Citroen, Renault, Peugot kategóriákra a halmozott abszolút gyakoriságok (3, 5, 9).

Ha ugyanezt elvégezzük a relatív gyakoriságokra, a kapott számokat **halmozott relatív gyakoriságoknak** nevezzük. A márkák esetében a Citroen, Renault, Peugot kategóriákra a halmozott relatív gyakoriságok: (1/3, 5/9, 1).

A **halmozott gyakoriság-eloszlás** az összes kategória halmozott relatív gyakoriságainak összessége. A márka halmozott gyakoriság-eloszlása a Citroen, Renault, Peugot kategóriákra (1/3, 5/9, 1).

I.3 Nominális és diszkrét változók grafikus ábrázolása

Ez a fejezet a nominális és diszkrét változók leggyakrabban használt diagramjait mutatja be. Ezek az oszlopdiagram, Pareto diagram és a kördiagram.

Az oszlopdiagram

Mindegyik osztálynak rajzolunk egy oszlopot. Az oszlop magassága megegyezik az abszolút vagy a relatív gyakorisággal.

A Pareto diagram

Készítünk egy oszlopdiagramot az abszolút gyakoriságokra úgy, hogy a gyakoriságok csökkenjenek az elsőtől az utolsóig. Utána húzzunk egy vonalat az első (leggyakoribb) kategóriától kezdve, amely összeköti az összes halmozott abszolút gyakoriságot. Figyeljünk arra, hogy a halmozott abszolút gyakoriságokat ugyanolyan sorrendben számítsuk ki mint az abszolút gyakoriságokat.

A kördiagram

Ez egy cikkekre osztott kör. A körcikkek területe arányos az abszolút gyakoriságokkal.

A módusz

Egy fontos információ, amit könnyedén leolvashatunk a diagramokról a **módusz**, vagyis a legnagyobb gyakoriságú kategória. Például, az autótípusokra a márka változó módusza a Peugeot.

Gyakorlat. A légzsák változóra készítsünk

- oszlopdiagramot,
- Pareto diagramot.

I.4 Folytonos változók leíró statisztikája

Ebben a fejezetben a folytonos változók leíró statisztikáját tárgyaljuk a gyakoriságok segítségével. A gyakoriságokat felhasználva bemutatjuk a hisztogramot és a gyakorisági poligont, amelyeket a leggyakrabban használnak a folytonos változók grafikus ábrázolására.

Folytonos változók gyakoriság-eloszlása

A fenti autó-jellemzők példában az ár folytonos változó. Folytonos változók esetén a különböző értékek általában **csak egyszer** fordulnak elő. Ezért egy oszlopdiagram nem nyújt érdekes információt a folytonos változókról, mert az oszlopok magassága a legtöbb kategóriára 1. Lényegében csak azt mutatja meg, hogy milyen értékek fordulnak elő.

Ezért a legkisebb és legnagyobb érték közötti intervallumot felosztjuk olyan kisebb intervallumokra, amelyek nem fedik egymást. Az intervallumokat kategóriáknak tekintjük, és kiszámítjuk az abszolút és relatív gyakoriságokat. A relatív gyakoriságok összességét a folytonos változó **gyakoriság-eloszlásának** nevezzük. Az intervallumok száma és hossza tetszőleges, de javasolt, hogy az intervallumok egyenlő hosszúságúak legyenek és számuk megközelítőleg a megfigyelések számának négyzetgyöke legyen.

Példa: 30 diák magassága. A magasságok méterben a következők:

1.69, 1.55, 1.74, 1.76, 1.62, 1.85, 1.85, 1.73, 1.76, 1.75, 1.71, 1.80, 1.67, 1.95, 1.72, 1.74, 1.84, 1.74, 1.72, 1.65, 1.76, 1.60, 1.80, 1.89, 1.66, 1.82, 1.86, 1.61, 1.59, 1.79.

(Megjegyezzük, hogy a pontok tizedes pontok. Például, 1.69 egy egész hatvankilenc század.)

A legkisebb érték: 1.55 m, míg a legnagyobb 1.95 m. Ezt a terjedelmet felosztjuk kisebb intervallumokra:

$1.55 \leq \text{magasság} \leq 1.60$

$1.60 < \text{magasság} \leq 1.65$

$1.65 < \text{magasság} \leq 1.70$

$1.70 < \text{magasság} \leq 1.75$

$1.75 < \text{magasság} \leq 1.80$

$1.80 < \text{magasság} \leq 1.85$

$1.85 < \text{magasság} \leq 1.90$

$1.90 < \text{magasság} \leq 1.95$

8 intervallumot kapunk, mindegyik hossza 0.05. 30 négyzetgyöke megközelítőleg 5.5, tehát az intervallumok javasolt száma 5 vagy 6. A könnyebb számolás végett veszünk mi 8 intervallumot.

A gyakoriságok a következő táblázatban szerepelnek:

Táblázat 1.4.1. – A diákok magassága alapján számolt gyakoriságok

Intervallum	Abszolút gyakoriság	Halmozott abszolút gyakoriság	Relatív gyakoriság	Halmozott relatív gyakoriság
[1.55,1.60]	3	3	3/30= 0.1	0.1
(1.60,1.65]	3	6	0.1	0.2
(1.65,1.70]	3	9	0.1	0.3
(1.70,1.75]	8	17	0.27	0.57
(1.75,1.80]	6	23	0.2	0.77
(1.80,1.85]	4	27	0.13	0.9
(1.85,1.90]	2	29	0.07	0.97
(1.90,1.95]	1	30	0.03	1

A hisztogram

A **hisztogram** az oszlopdiagramhoz hasonló ábra, ahol az intervallumok felelnek meg a különböző kategóriáknak. Megmutatja, hogy a megfigyelések hol helyezkednek el. Felhívjuk a figyelmet arra, hogy egy lényeges különbség az oszlopdiagramhoz képest az, hogy az oszlopok területe és nem a magassága egyenlő a gyakoriságokkal. Vagyis:

gyakoriság = oszlop területe = oszlop magassága x intervallum hossza.

Tehát:

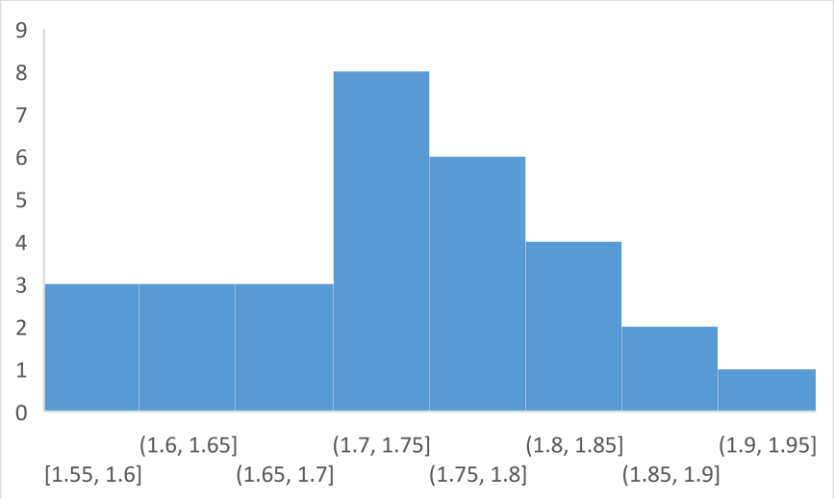
oszlop magassága = gyakoriság / intervallum hossza.

Példa: 30 diák magassága (folytatás). A hisztogram és a halmozott hisztogram oszlopainak magassága:

Táblázat 1.4.2. – A diákok magassága alapján számolt gyakoriságok (folytatás)

Intervallum	Intervallum hossza	Abszolút gyakoriság	Oszlop magassága	Halmozott gyakoriság	Oszlop magassága
[1.55,1.60]	0.05	3	3 / 0.05 = 60	3	3 / 0.05 = 60
(1.60,1.65]	0.05	3	3 / 0.05 = 60	6	6 / 0.05 = 120
(1.65,1.70]	0.05	3	3 / 0.05 = 60	9	9 / 0.05 = 180
(1.70,1.75]	0.05	8	8 / 0.05 = 160	17	17 / 0.05 = 340
(1.75,1.80]	0.05	6	6 / 0.05 = 120	23	23 / 0.05 = 460
(1.80,1.85]	0.05	4	4 / 0.05 = 80	27	27 / 0.05 = 540
(1.85,1.90]	0.05	2	2 / 0.05 = 40	29	29 / 0.05 = 580
(1.90,1.95]	0.05	1	1 / 0.05 = 20	30	30 / 0.05 = 600

Ábra 1.4.1. - Hisztogram: diákok magassága



A halmozott hisztogram

A **halmozott hisztogramot** hasonlóan tudjuk megszerkeszteni, viszont az oszlopok magasságait a halmozott gyakoriságok segítségével számítjuk ki, lásd az előző táblázatot.

Hisztogramok alakja

Egy hisztogram alakja nagyvonalú információt nyújt a megfigyelésekről, éspedig konkrétan arról, hogy az intervallumokban mennyire gyakoriak a megfigyelések. Gyakran találkozunk a következő formákkal:

- egymóduszú vagy többmóduszú,
- szimmetrikus vagy asszimmetrikus (ferde).

Az egymóduszú hisztogramnak egyetlen legmagasabb oszlopa van", míg egy többmóduszúnak több legmagasabb oszlopa van. Ezenkívül, az egymóduszú formák között gyakoriak a jobbra elnyúló és balra elnyúló hisztogramok. A szimmetrikus hisztogram közepetől jobbra és balra egyforma a hisztogram alakja. A fenti hisztogram példa egy asszimmetrikus alakú hisztogramra.

A gyakorisági poligon

A **gyakorisági poligon** tulajdonképpen ugyanazt az információt nyújtja mint a hisztogram, viszont az oszlopok tetejének középpontjait összekötő vonal határozza meg.

I.5 Statisztikai mutatók

Ez a fejezet a statisztikai mutatókkal foglalkozik. Tárgyaljuk a legfontosabb helyzetmutatókat és szóródási mutatókat és ezek alapvető tulajdonságait. A **statisztikai mutatók** olyan mennyiségek amelyeket a megfigyelésekből számítunk ki, és egyetlen számadatba sűrítik a megfigyelések valamilyen fontos vonását. Ezeknek **csak kvantitatív** változókra van értelmük. A megfigyelések fontos tulajdonságait mutatják meg, mint például:

- hol van a megfigyelések „közepe": helyzetmutatók (középértékek),
- mennyire vannak szétszóródva a megfigyelések: szóródási mutatók,
- mennyire asszimmetrikusak a megfigyelések: asszimmetria-mutatók.

Az átlag

Az egyik legismertebb statisztikai mutató. Jelöljünk n megfigyelést x_1, x_2, ..., x_n -nel. Ekkor az **átlaguk**:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

A jele általában: $\bar{x}$. Például, a 3.4, 1.3, 6.5, 4.0, 2.8 megfigyelések átlaga 18.0/5 = 3.6.

Az átlag tulajdonságai

A átlag egy helyzetmutató. A megfigyelések átlagtól való eltérései a megfigyelések és az átlag közötti különbségek: $x_1 - \bar{x},\ x_2 - \bar{x}, \ldots,\ x_n - \bar{x}$. Az átlagtól való eltérések egy tulajdonsága, hogy az összegük mindig nulla: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Két mintára, $x_1, x_2, \ldots, x_n$ *és* $y_1, y_2, \ldots, y_n$-re fennáll, hogy az átlagok összege egyenlő az összegek átlagával: $\bar{x} + \bar{y} = \overline{x + y}$.

A szórás

A szórás egy szóródási mutató, ami azt méri, hogy a megfigyelések mennyire esnek távol az átlagtól. Két minta lehet lényegesen különböző, még ha az átlaguk egyforma is, ugyanis az egyik minta megfigyelései eshetnek jóval közelebb az átlaghoz mint a másik minta megfigyelései. A szóródás méréséhez az eltérések **négyzetét** veszi figyelembe, mert ezáltal fejezhető ki az összes megfigyelés távolsága az átlagtól.

A **szórás** kiszámításához négyzetgyököt vonunk az eltérések négyzetének átlagából:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Ezt a képletet át lehet írni egy másik formába (a két képlet ugyanazt az eredményt adja) :

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}.$$

Például, a 3.4, 1.3, 6.5, 4.0, 2.8 megfigyelések átlaga 3.6. A szórás:

$$\sqrt{\frac{1}{5}(3.4^2 + 1.3^2 + 6.5^2 + 4.0^2 + 2.8^2) - 3.6^2} = \sqrt{15.868 - 12.96} = 1.705.$$

A variancia (szorásnégyzet)

Egy másik fontos szóródási mutató a **variancia**, ami egyszerűen a szórás négyzete, vagyis az átlagtól való eltérések négyzeteinek az átlaga:

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Példa. Könnyen meggyőződhetünk arról, hogy a variancia vagy a szórás valóban szóródási mutató, ha vesszük azt a szélsőséges esetet, amikor az összes megfigyelés egyforma. Ekkor nincs szóródás, és szórás = variancia = 0.

A szórás tulajdonságai

A szórás segítségével meg lehet állapítani, hogy a megfigyelések hányad része helyezkedik el az átlagtól egy adott távolságra. Ezt Csebisev szabálya határozza meg.

Csebisev szabálya: Azoknak a megfigyeléseknek az aránya, amelyek kevesebb mint m szórásra vannak az átlagtól legalább $1 - \frac{1}{m^2}$.

Eszerint, a megfigyelések legalább

- 75%-a közelebb van az átlaghoz mint $2s_x$,
- 89%-a közelebb van az átlaghoz mint $3s_x$, stb.

Az átlag és a szórás tulajdonságai lineáris transzformáció esetén

Egy y számot az x szám lineáris transzformáltjának tekintjük, ha bizonyos a és b számokra $y = ax+b$. Az $x_1, x_2, ..., x_n$ megfigyelések lineáris transzformáltjai az $y_1, y_2, ...,y_n$ megfigyelések amelyekre $y_1 = ax_1+b$, $y_2 = ax_2+b$,..., $y_n = ax_n+b$. Ez utóbbiak átlaga kiszámítható úgy mint az $\bar{x}$ átlag lineáris transzformációja:

$$\bar{y} = a\bar{x} + b.$$

Az $y_1, y_2, ...,y_n$ megfigyelések

varianciája: $s_y^2 = a^2 s_x^2$ és szórása: $s_y = |a|s_x$,

vagyis kiszámíthatók az $x_1, x_2, ..., x_n$ megfigyelések varianciája és szórása alapján. A variancia és a szórás azt mutatják, hogy ha a megfigyeléseknek egy a számmal való szorzatához hozzádjuk ugyanazt a b számot, ez a megfigyelések szóródását az átlag körül nem befolyásolja. Ezért nem függ b-től a variancia és a szórás.

A relatív szórás

A fenti képlet alapján a szórás függ a mértékegységtől. Ezt a következő példa is illusztrálja. A következő árak a Dacia Duster árai különböző kereskedőknél:

€12400; €11950; €12280; €12310; €11930: €12600; €12500; €11890; €11990; €12050; €12100; €12000

Az átlag €-ban 12167 és a szórás 231.53. Az árak lejben €1 = 4.3 lej árfolyam esetén:

53320, 51385, 52804, 52933, 51299, 54180, 53750, 51127, 51557, 51815, 52030, 51600

Az átlag 52316.67 és a szórás 1039.843.

Kérdés: lejben nagyobb az árak szóródása mint euróban? Válasz: nem, mert ugyanazokról az árakról van szó, csak a mértékegységek (pénznemek) különböznek.

Sok esetben a szóródás mérésekor, ami fontos, az a szórás nagysága az átlaghoz viszonyítva. Ezért a mértékegység nem kellene szerepet játsszon.

A relatív szórás egy ilyen szóródási mutató. A szórás és az átlag abszolút értékének hányadosa (**c**oefficient of **v**ariation):

$$CV_x = \frac{s_x}{|\bar{x}|}$$

A relatív szórás a szórást tükrözi az átlag mértékéhez viszonyítva. Leginkább két minta szórásainak összehasonlítására használják.

A Dacia Duster-árakra:

- Euróban: CV = 231.53 / 12167 = 0.019,
- Lejben: CV = 1039.843 / 52316.67 = 0.019.

A relatív szórás nem változik, ha a mértékegységet megváltoztatjuk (pl. € lejre, kilométert mérföldre, stb.). Az átlaggal és a szórással ellentétben a relatív szórásnak nincs mértékegysége.

Az aszimmetria-mutató

Egy olyan mutató, amelyik méri, hogy a gyakoriság-eloszlás (vagy hisztogram) szimmetrikus-e. Képlete:

$$CA_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\, s_x^3}.$$

Láthatjuk, hogy az aszimmetria-mutatót, akárcsak a szórást vagy varianciát az átlagtól való eltéréseket felhasználva számítjuk ki. Egy tulajdonsága az, hogy nincs mértékegysége.

Az aszimmetria-mutatót a következőképpen használjuk. Ha a gyakoriság-eloszlás

- tökéletesen szimmetrikus, akkor $CA_x = 0$,
- jobbra elnyúló, akkor $CA_x > 0$,
- balra elnyúló, akkor $CA_x < 0$.

A medián

A korábban tárgyalt statisztikai mutatók átlagokon alapulnak (átlag, variancia, ami az eltérések négyzeteinek az átlaga). Most egy olyan helyzetmutatót tárgyalunk, amely a megfigyelések sorrendjén alapul; meg fogjuk látni, hogy ez miért hasznos.

Jelöljük $x_1, x_2, \ldots, x_n$ –nel a megfigyeléseket, és jelöljük $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ –nel ugyanezeket a megfigyeléseket növekvő sorrendben. Tehát, $x_{(1)}$ a legkisebb, $x_{(2)}$ a következő, ... és $x_{(n)}$ legnagyobb:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

A medián a sorrendbe állított $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ megfigyelések középső megfigyelése. Ha n páratlan, akkor egészen egyszerű; a medián a (n+1)/2 sorrendű megfigyelés:

$$x_{\left(\frac{n+1}{2}\right)}.$$

Ha n páros, akkor 2 középső megfigyelés van, n/2 és (n/2)+1, tehát a medián:

$$\frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}.$$

Példa: a 3.4, 1.3, 6.5, 4.0, 2.8 megfigyelések mediánja **3.4** (mivel 1.3, 2.8, **3.4**, 4.0, 6.5 a sorrendbe állított megfigyelések; n = 5 páratlan, tehát a középső megfigyelés a 3.: (n+1)/2 = 3).

Példa: a 102, 93, 105, 110, 87, 91 megfigyelések mediánja (93+102)/2 = **97.5**. (mivel 87, 91, **93**, **102**, 105, 110 a sorrendbe állított megfigyelések; n = 6 tehát a n/2 = 3. és a (n/2)+1 = 4. megfigyeléseket vesszük, amelyek 93 és 102.

Az abszolút eltérések mediánja

A mediánnak megfelelő szóródási mutató hasonlóan ahhoz, ahogy a szórás az átlagnak megfelelő szóródási mutató. Akkor használják, amikor helyzetmutatóként a mediánt használják.

A megfigyelések és a medián közti abszolút eltérések:

$$|x_1 - \text{medián}|,\ |x_2 - \text{medián}|, ..., |x_n - \text{medián}|.$$

Az abszolút eltérések mediánja:

$$MADx = \text{medián}\ (|x_1 - \text{medián}|,\ |x_2 - \text{medián}|, ..., |x_n - \text{medián}|).$$

Az ötlet hasonló a varianciához, ami az abszolút eltérések négyzeteinek átlaga.

Példa: a 3.4, 1.3, 6.5, 4.0, 2.8 megfigyelések mediánja 3.4. Az abszolút eltérések:

$$|3.4 - 3.4|, |1.3 - 3.4|, |6.5 - 3.4|, |4.0 - 3.4|, |2.8 - 3.4| = 0, 2.1, 3.1, 0.6, 0.6.$$

Ez alapján az abszolút eltérések mediánja **0.6**, mivel a sorrendbe állított abszolút eltérések 0, 0.6, 0.6, 2.1, 3.1.

A minta (szóródás) terjedelme

A minta (szóródás) terjedelme a legnagyobb és a legkisebb megfigyelés különbsége (R = range, angolul):

$$R = x_{(n)} - x_{(1)}.$$

Szóródási mutató, de kevésbé informatív mint a szórás vagy az abszolút eltérések mediánja.

Kvartilisek

Amint a fogalom elnevezése is mutatja (kvart = negyed), ezek a helyzetmutatók a sorrendbe állított megfigyelések első és utolsó negyedét határozzák meg. Két kvartilis létezik: az alsó kvartilis, jelölése Q1, és a felső kvartilis, jelölése Q3.

Az alsó kvartilis a megfigyelések legkisebb negyedét választja el a többitől. Azzal a megfigyeléssel egyenlő, amelynek sorrendszáma

$$0.25n + 0.5 = n/4 + 0.5.$$

Ha ez nem egész szám, akkor a sorrendben a két mellette lévő megfigyelés átlagát vesszük (ugyanúgy mint a medián meghatározásánál).

MEGJEGYZÉS. A kvartilisek meghatározására több képlet létezik, amelyek itt szerepelnek nem az egyedüliek. Viszont az eltérések a képletek eredményei között nem nagyok.

A felső kvartilis a megfigyelések legnagyobb negyedét választja el a többitől. Azzal a megfigyeléssel egyenlő, amelynek sorrendszáma

$$0.75n + 0.5 = 3n/4 + 0.5.$$

Megint, ha ez nem egész szám, akkor a sorrendben a két mellette lévő megfigyelés átlagát vesszük. Még megjegyezzük, hogy a ,,középső kvartilis" tulajdonképpen a medián.

Példa: A 5.3, 2.8, 3.4, 7.2, 8.3, 1.7, 6.2, 9.3, 3.2, 5.9 (n=10) megfigyelések növekvő sorrendben: 1.7, 2.8, 3.2, 3.4, 5.3, 5.9, 6.2, 7.2, 8.3, 9.3.

A medián = (5.3 + 5.9)/2 = **5.6**.

Az alsó kvartilis a 0.25x10+0.5 = 3 sorrendű megfigyelés, tehát Q1 = **3.2**.

Az felső kvartilis a 0.75x10+0.5 = 8 sorrendű megfigyelés, tehát Q3 = **7.2**.

Példa: A 5.3, 2.8, 3.4, 7.2, 8.3, 1.7, 6.2, 9.3, 3.2 (n=**9**) megfigyelések növekvő sorrendben: 1.7, 2.8, 3.2, 3.4, **5.3**, 6.2, 7.2, 8.3, 9.3

A medián = **5.3** (5. megfigyelés).

Az alsó kvartilishez 0.25x9+0.5 = 2.75, $\mathbf{Q_1}$ = (2.8+3.2)/2 = **3**, vagyis a 2. és 3. megfigyelés átlaga.

A felső kvartilishez 0.75x9+0.5 =7.25, tehát a 7. és 8. megfigyelés átlaga $\mathbf{Q_3 = 7.75}$.

Az interkvartilis terjedelem-mutató

Egy szóródási mutató, ami a felső és az alsó kvartilis közötti különbség:

$$IQR_x = Q_3 - Q_1$$

A megfigyelések középső felének a terjedelmét mutatja.

Tulajdonságok lineáris transzformáció esetén

Legyenek $y_1 = ax_1 + b$, $y_2 = ax_2 + b$,..., $ax_n + b$ az $x_1, x_2,..., x_n$ megfigyelésekből kapott lineárisan transzformált megfigyelések a és b számokra. Akkor a medián az eredeti medián lineáris transzformáltja:

$$\text{medián}(y_1,...,y_n) = a\ \text{medián}(x_1,...,x_n) + b,$$

míg az abszolút eltérések mediánja és az interkvartilis terjedelem

$$MAD_y = |a|\, MAD_x,$$

és

$$IQR_y = |a|\, IQR_x.$$

Az utóbbi két szóródási mutatót a lineárisan transzformált megfigyelésekre a szóráshoz hasonlóan számítjuk ki.

A doboz-ábra

A doboz ábra a mediánra és a kvartilisekre épülő diagram. A következő elemekből tevődik össze:

- egy doboz (téglalap) amelynek függőleges oldalai a Q_1 és Q_3 –nál találhatóak,
- húzzunk egy függőleges vonalat a doboz belsejében a mediánnál,
- húzzunk szaggatott függőleges vonalakat a doboz két oldalán 1.5IQR és 3IQR távolságra a doboz oldalaitól,
- rajzoljunk • jelt mindegyik megfigyelésnek, amelyik az 1.5IQR és 3IQR közé esik,

- rajzoljunk x jelt mindegyik megfigyelésnek amelyik a 3IQR-n kívül esik.

Gyakorlat. A következő megfigyelésekre

5.3, 2.8, 3.4, 7.2, 14.1, 1.7, 6.2, 20.5, 3.2, 5.9

készítsük el a doboz-ábrát felhasználva, hogy

medián = 5.6; Q1 = 32, Q3 = 7.2; IQR = 4.0; 1.5 IQR = 6.0, 3 IQR = 12.0.

Kiugró értékek

Azokat a megfigyeléseket, amelyek több mint 1.5 IQR-rel kisebbek mint Q1, vagy több mint 1.5 IQR-rel nagyobbak mint Q3 kiugró értékeknek nevezzük (azok az értékek amelyeket •-tal vagy ×-tel jelölünk a doboz-ábrán).

Azokat a megfigyeléseket, amelyek több mint 3 IQR-rel kisebbek mint Q1, vagy több mint 3 IQR-rel nagyobbak mint Q3 szélsőségesen kiugró értékeknek nevezzük (azok az értékek amelyeket ×-tel jelölünk a doboz-ábrán).

A kiugró értékek külön figyelmet igényelnek. Ezek szokatlanul kicsik vagy nagyok. Előfordulhat, hogy a megfigyelések gyűjtése során hiba csúszott be, viszont bizonyos esetekben fontos információval szolgálnak az illető változóról.

Az átlag és a medián összehasonlítása

Az átlag és a medián nagyon eltérhet egymástól, ha a megfigyelések között kiugró értékek vannak.

Példa: Tekintsük a következő megfigyeléseket: 5.3, 2.8, 3.4, 7.2, 8.3, 1.7, 6.2, 9.3, 3.2. Az átlag **5.27**, a medián **5.3**. Tegyük fel, hogy adatgyűjtéskor a 7.2 megfigyelést tévedésből 72-nek írták. Ez szélsőségesen kiugró érték. Az átlag **12.47**, de a medián még mindig **5.3**.

A medián robusztussága

Az előbbi példa alapján láthatjuk, hogy az átlag nagyon megváltozhat, ha kiugró értékek kerülnek a megfigyelések közé, míg a medián nem. Tehát, ilyen esetben a medián megbízhatóbb helyzetmutató mint az átlag.

Azt a tulajdonságot, hogy a medián kevésbé változik, ha kiugró értékek kerülnek a megfigyelések közé, robusztusságnak nevezik. A doboz-ábra egy olyan eszköz, amelyik megmutatja, hogy vannak-e kiugró értékek.

A vágott átlag

Az α%-os vágott átlag azoknak a megfigyeléseknek az átlaga, amelyeket úgy nyerünk, hogy az eredeti megfigyelésekből a legnagyobb és a legkisebb α %-ot eltávolítjuk. Robusztusabb mint az átlag, mert a kiugró értékeket így valószínűleg eltávolítjuk.

Példa: A 5.3, 2.8, 3.4, 72, 8.3, 1.7, 6.2, 9.3, 3.2 megfigyelésekre a 10%-os vágott átlag kiszámításához vegyük figyelembe, hogy 9 megfigyelés van, 10%-a 0.9, ezért eltávolítjuk a legkisebb és a legnagyobb értékeket, és a következőt kapjuk: 5.3, 2.8, 3.4, ~~72~~, 8.3, ~~1.7~~, 6.2, 9.3, 3.2. A 10%-os vágott átlag a megmaradt 7 megfigyelés átlaga, vagyis **5.5**.

A módusz

A módusz nominális és diszkrét változókra, amint fennebb láthattuk, a legnagyobb gyakorisággal előforduló kategória. Folytonos változókra intervallumokban vizsgáljuk a gyakoriságokat. A modális kategória (intervallum) a legnagyobb gyakoriságú intervallum, folytonos változóknál ez játssza a módusz szerepét.

Statisztikai mutatók kategorizált megfigyelésekre

Sok esetben nincsenek pontos megfigyeléseink, hanem a megfigyelések kategóriái vannak megadva. Például, a fizetések általában a fizetéskategóriák gyakoriságaival vannak megadva mint az alábbi havi fizetések példában:

Osztályok (lej)	Gyakoriság
600-800	1000
801-1000	750
1001-1300	500
1301-1500	250
1500-5000	200
5000-50000	75

Felmerül a kérdés, hogy mennyi az így megadott havi fizetések átlaga, és általában, hogy hogyan számítsunk statisztikai mutatókat ilyen esetben.

Az átlagra vegyük észre, hogy a képletét $\bar{x} = \sum_{i=1}^{n} x_i \frac{1}{n}$ alakban is írhatjuk, ahol x_i a megfigyelés és $1/n$ a relatív gyakorisága. Kategorizált megfigyelésekre ugyanezt az ötletet alkalmazzuk, csak a megfigyelések helyett az osztályközepeket (intervallumok közepét) vesszük figyelembe. Jelöljük az osztályközepeket c_1, c_2, ..., c_k –val és az osztályok relatív gyakoriságait f_1, f_2, ..., f_k -val. Ekkor az átlag képlete kategorizált megfigyelésekre:

$$\bar{x}_c = \sum_{i=1}^{k} c_i f_i.$$

A többi statisztikai mutatót ugyanez az ötlet alapján számítjuk ki. A szórás kategorizált megfigyelésekre:

$$s_c = \sqrt{\sum_{i=1}^{k} (c_i - \bar{x}_c)^2 f_i}.$$

Az asszimetria-mutató kategorizált megfigyelésekre:

$$CA_c = \frac{\sum_{i=1}^{k} (c_i - \bar{x}_c)^3 f_i}{s_c^3}.$$

A medián és a kvartilisek kategorizált megfigyelésekre

Kategorizált megfigyelésekre tudjuk az intervallumokra az abszolút gyakoriságokat, ezért tudunk hisztogramot készíteni. Mivel az abszolút gyakoriságok egyenlőek a hisztogramban szereplő oszlopok területével, a medián az az érték lesz, amely az egész hisztogram területét pontosan kettéosztja. Hasonlóan, az alsó kvartilis Q1 az az érték lesz, amely a hisztogram

területének ¼ -ét, míg a felső kvartilis Q3 az az érték lesz, amely a hisztogram területének ¾ -ét választja el a többitől.

Megfigyelések transzformációi

A megfigyelések transzformációja olyan megfigyeléseket eredményez, amelyeket **ugyanazon** függvény alkalmazásával kapunk. Beszéltünk már a megfigyelések lineáris transzformációjáról. Ez például akkor hasznos, ha megváltoztatjuk a mértékegységet.

Bizonyos esetekben a nem lineáris transzformációk is hasznosak. Sok esetben a megfigyelések transzformációja megkönnyíti a statisztikai elemzést: például a transzformáció eredményeként nulla átlagot vagy szimmetrikus gyakoriság-eloszlást kapunk.

Standardizált megfigyelések

Láttuk, hogy lineáris transzformáció esetén ($y_i = ax_i + b$):

$$\bar{y} = a\bar{x} + b; \quad s_y = |a|s_x.$$

Ha a transzformáció az eredeti megfigyelések eltéréseit elosztja a szórásukkal, akkor:

$$y_i = \frac{x_i - \bar{x}}{s_x} = \frac{1}{s_x}x_i - \frac{\bar{x}}{s_x}.$$

Az y_1 , y_2 ,…, y_n megfigyelések átlaga 0 és szórása 1. Ezeket a megfigyeléseket standardizált megfigyeléseknek nevezzük. A standardizált megfigyelések megmutatják, hogy a megfigyelések hány szórásra vannak az átlagtól. Fontos szerepet játszanak két változó közötti összefüggés meghatározásában, amiről a következő fejezetben lesz szó.

Példa: Vegyük a következő mintát: 5.3, 2.8, 3.4, 7.2, 8.3, 1.7, 6.2, 9.3, 3.2. A standardizált megfigyelések:

0.012, -0.927, -0.701, 0.727, 1.14, -1.340, 0.351, 1.516, -0.777.

A medián által standardizált megfigyelések

Standardizálhatunk a medián és az abszolút eltérések mediánja segítségével:

$$y_i = \frac{x_i - \text{median}(x_1, \ldots x_n)}{MAD_x}.$$

Az így kapott y_1 , y_2 ,…, y_n megfigyelések mediánja 0 és abszolút eltéréseinek mediánja 1. Akkor hasznos, amikor az átlag helyett a mediánt használjuk helyzetmutatóként. Viszont a legtöbb esetben az átlag és a szórás által standardizált megfigyeléseket használják.

Nem lineáris transzformációk

A leggyakrabban használt nem lineáris transzformációk a következők:

$$y_i = log(x_i),\ y_i = \sqrt{x_i},\ y_i = exp(x_i).$$

Elsősorban arra használjuk, hogy a gyakoriság-eloszlást vagy a hisztogramot szimmetrikussá tegyük. A lineáris transzformáció nem változtatja meg ezek szimmetria tulajdonságait, tehát a standardizálás sem.

Szimmetrikusabbá tevő transzformációk

A szimmetrikus gyakoriság-eloszlással rendelkező megfigyelések sok esetben könnyebben elemezhetők. Az átlag és a szórás elegendő, és nincs szükség asszimetria-mutatóra. Jobban hasonlítanak a normális eloszlásból (II. Rész) nyert megfigyelésekre, ami megkönnyíti a magasabb szintű elemzést is.

A leggyakrabban használt transzformációk, amelyek a gyakoriság-eloszlást szimmetrikusabbá teszik: logaritmus, négyzetgyök, reciprok (vagy inverz), négyzet, más hatvány. A transzformációt aszerint választjuk meg, hogy milyen típusú a gyakoriság-eloszlás aszimmetriája.

Ha a gyakoriság-eloszlás **jobbra elnyúló**, akkor a logaritmus és a négyzetgyök a gyakoriság-eloszlást szimmetrikusabbá teszik.

- A logaritmus ,,erősebb" mint a négyzetgyök.

Ha a gyakoriság-eloszlás **balra elnyúló**, akkor a négyzet, köb vagy nagyobb hatványok teszik a gyakoriság-eloszlást szimmetrikusabbá.

- Minél nagyobb a hatvány, annál ,,erősebb" a transzformáció.

Kategorizált megfigyelések transzformációja

Említettük, hogy a fizetések általában a fizetéskategóriák gyakoriságaival vannak megadva. Ugyanakkor, a fizetések gyakoriság-eloszlása általában jobbra elnyúló, amit bizonyos esetekben hasznos szimmetrikusabbá tenni. Nagyjából hasonlóan járhatunk el mint eddig:

- transzformáljuk az intervallumok végpontjait,
- a kapott intervallumokban a gyakoriságok ugyanazok maradnak,
- az új intervallumhosszúságokkal kiszámítjuk az új hisztogram oszlopainak a magasságát,
- a transzformációt az előző oldalon tárgyaltak alapján választjuk ki.

I.6 Két változó közötti összefüggés leíró statisztikája

Ez a fejezet két változó egyidejű tanulmányozásával foglalkozik. Az egyváltozóra bevezetett fogalmak továbbfejlesztésével tanulmányozzuk a leíró statisztikát, két változó függetlenségét, és két változó közötti lineáris összefüggés mérésére szolgáló legfontosabb statisztikai mutatókat.

Eddig azt tanulmányoztuk, hogyan írjuk le statisztikailag az **egyes** változókat. Számos gazdasági jelenség vizsgálatához több változót kell figyelembe venni. Például:

- A gazdasági növekedéshez a GDP változásán kívül a munkanélküliség változását is figyelembe veszik. Minden országra vannak megfigyelések a (GDP, munkanéküliségi arány) változópárra.
- Az autóknak több jellemzőjük van (ár, lóerő, stb.), amelyek a keresletet határozzák meg.

Változók közötti összefüggések

Gyakran fontos tudni, hogy két változó között milyen összefüggés van. Például, fontos az, hogy van-e összefüggés a GDP és a munkanélküliségi arány között vagy az autók ára és az eladások között.

Ugyanúgy mint egy változó esetében, a kvalitatív és a kvantitatív változókat külön fogjuk tárgyalni. Először két kvalitatív nominális változó közti összefüggést vizsgálunk.

Két nominális változó egyidejű gyakoriság-eloszlása

Legyen X és Y a két nominális változó. X-nek *k* különböző kategóriája van $c_1,c_2,...,c_k$; Y-nak *m* különböző kategóriája van d_1, $d_2,...,d_m$. A megfigyelések száma *N*, és legyenek a megfigyegyelések: ezek (x_i ,y_i), $i=1,...,N$.

Példa: 115 személy szakmai helyzete és képzése Spanyolországban egy kérdőíves felmérés alapján. A változók (zárójelben a kategóriáik):

- X = szakmai helyzet (szakképzett/menedzser/ hivatalnok/ részmunkaidős),
- Y = képzés (8 osztály/ középiskola/ egyetem/ főiskola).

Ebben az esetben $N = 115$, $k = 4$, $m = 4$ (az, hogy *k* egyenlő legyen *m*-mel nem kötelező). Mindkét változó (kvalitatív) nominális. Példák megfigyelésekre:

- (szakképzett, középiskola) vagy
- (hivatalnok, egyetem).

A kategóriáknak $k \times m$ különböző kombinációjuk van. Legyen n_{ij} azoknak a megfigyeléseknek a száma amelyekre $X = c_i$ $(i = 1,...,k)$ és $Y = d_j$ $(j = 1,...,m)$. Ekkor azt mondjuk, hogy n_{ij} a (c_i,d_j) kategória egyidejű abszolút gyakorisága. Két változóra az egyidejű abszolút gyakoriságokat táblázatba lehet foglalni:

Táblázat 1.6.1. – Minta táblázat két nominális változó egyidejű gyakoriság eloszlására

	$Y = d_1$	$Y = d_2$	...	$Y = d_m$
$X = c_1$	n_{11}	n_{12}	...	n_{1m}
$X = c_2$	n_{21}	n_{22}	...	n_{2m}
⋮	⋮	⋮		⋮
$X = c_k$	n_{k1}	n_{k2}	...	n_{km}

Egyidejű relatív gyakoriságok

Az egyidejű relatív gyakoriságokat úgy kapjuk, hogy az egyidejű abszolút gyakoriságokat elosztjuk a megfigyelések számával:

$$f_{ij} = \frac{n_{ij}}{N}, \quad i = 1, \dots k; \; j = 1, \dots m.$$

Az egyidejű gyakoriság-eloszlás az egyidejű relatív gyakoriságok táblázata.

Táblázat 1.6.2. - Példa: szakmai helyzet és képzés. Az egyidejű abszolút gyakoriságok táblázata:

	egyidejű abszolút gyakoriságok (megfigyelések száma = 115)			
	8 osztály	Közép-iskola	Egyetem	Főiskola
Szakképzett	0	2	2	12
Menedzser	1	8	14	14
Hivatalnok	5	21	10	2
Részmunka-idős	3	15	3	3

Az egyidejű relatív gyakoriságok táblázata, vagyis a gyakoriság-eloszlás:

Táblázat 1.6.3. – Példa (folytatás): egyidejű relatív gyakoriságok táblázata

	egyidejű relatív gyakoriságok (megfigyelések száma = 115)			
	8 osztály	Közép-iskola	Egyetem	Főiskola
Szakképzett	0.000	0.017	0.017	0.104
Menedzser	0.009	0.070	0.122	0.122
Hivatalnok	0.043	0.183	0.087	0.017
Részmunka-idős	0.026	0.130	0.026	0.026

Például, a 115 személy 8.7 százaléka egyetemet végzett és hivatalnokként dolgozik.

Peremeloszlások

Az i –edik sor elemeinek az összege az $X = c_i$ megfigyelések száma, tehát a c_i kategória abszolút/relatív gyakorisága:

$$n_{i\bullet} = \sum_{j=1}^{m} n_{ij}\,, \quad f_{i\bullet} = \sum_{j=1}^{m} f_{ij}.$$

A j –edik oszlop elemeinek az összege az $Y = d_j$ megfigyelések száma, tehát a c_i kategória abszolút/relatív gyakorisága:

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij}\,, \quad f_{\bullet j} = \sum_{i=1}^{k} f_{ij}.$$

Az $f_{i\bullet}$ összességét az X **peremeloszlás**ának nevezzük. Az $f_{\bullet j}$ összességét az Y **peremeloszlás**ának nevezzük. Az alábbi táblázat feltünteti a perem-gyakoriságokat:

Táblázat 1.6.4. – Példa (folytatás): Az abszolút gyakoriságok és a perem eloszlás

	8 osztály	Közép-iskola	Egyetem	Főiskola	**perem**
Szakképzett	0	2	2	12	**16**
Menedzser	1	8	14	14	**37**
Hivatalnok	5	21	10	2	**38**
Részmunka-idős	3	15	3	3	**24**
perem	**9**	**46**	**29**	**31**	**115**

A szakmai helyzet és képzés peremeloszlásai:

Táblázat 1.6.5. – Példa (folytatás): A relatív gyakoriságok és a perem eloszlás

	8 osztály	Közép-iskola	Egyetem	Főiskola	**perem**
Szakképzett	0.000	0.017	0.017	0.104	**0.139**
Menedzser	0.009	0.070	0.122	0.122	**0.322**
Hivatalnok	0.043	0.183	0.087	0.017	**0.330**
Részmunka-idős	0.026	0.130	0.026	0.026	**0.209**
perem	**0.078**	**0.400**	**0.252**	**0.270**	**1.000**

Például, a vizsgált 115 személy 13.9%-a szakképzettként dolgozik, míg 7.8%-a 8 osztályt végzett.

Feltételes eloszlások

Az X feltételes eloszlása ha (tudjuk, hogy) $Y = d_j$:

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}, \quad i = 1, \ldots, k.$$

Ez tulajdonképpen az X gyakoriság-eloszlása azokra a megfigyelésekre amelyekre $Y = d_j$. Hasonlóan, az Y feltételes eloszlása ha (tudjuk, hogy) $X = c_i$:

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}}, \quad j = 1, \ldots, m.$$

Tegyük fel, hogy egyedül azok szakmai helyzetére vagyunk kíváncsiak, akik képzettségi szintje középiskola: 46 ilyen személy van (relatív gyakoriságuk 0.4).

A **szakmai helyzet** feltételes eloszlása, ha a képzettségi szint:

- 8 osztály: 0/9, 1/9, 5/9, 3/9,
- Középiskola: 2/46, 8/46, 21/46, 15/46,
- Egyetem: 2/29, 14/29, 10/29, 3/29,
- Főiskola: 12/31, 14/31, 2/31, 3/31.

A **képzés** feltételes eloszlása, ha a személyek szakmai helyzete:

- Szakképzett: 0/16, 2/16, 2/16, 12/16,
- Menedzser: 1/37, 8/37, 14/37, 14/37,
- Hivatalnok: 5/38, 21/38, 10/38, 2/38,
- Részmunkaidős: 3/24, 15/24, 3/24, 3/24.

A feltételes eloszlások nagyon fontosak két változó közti összefüggés tanulmányozásánál. Ha az összes feltételes eloszlás egyenlő a peremeloszlással, akkor azt mondjuk, hogy a két változó független.

A **képzés** peremeloszlása: 0.078, 0.400, 0.252, 0.270.

A **képzés** feltételes eloszlása, ha a személyek szakmai helyzete

- Szakképzett: 0.000, 0.125, 0.125, 0.750
- Menedzser: 0.027, 0.216, 0.378, 0.378
- hivatalnok: 0.132, 0.553, 0.263, 0.053
- Részmunkaidős: 0.125, 0.625, 0.125, 0.125

Tehát a feltételes eloszlások különböznek egymástól és a peremeloszlástól. Ezért értelmezés szerint a képzés és a szakmai helyzet változók nem függetlenek.

Mi ennek a jelentése gyakorlati szempontból? Ha a feltételes eloszlások egyenlőek lennének, akkor a képzés feltételes eloszlása ugyanaz lenne a szakmai helyzettől **függetlenül**, tehát valóban azt mondhatnánk, hogy a változók nem függnek egymástól.

Két diszkrét változó egyidejű gyakoriság-eloszlása

Az egyidejű gyakoriságok táblázatát ugyanúgy készítjük el mint a kvalitatív nominális változókra. Viszont diszkrét változókra tudunk statisztikai mutatókat kiszámolni a peremeloszlásokra (átlag, szórás, stb.). Ezenkívül, a feltételes eloszlásokra is tudunk statisztikai mutatókat kiszámolni. Ez fontos, mert így könnyen ellenőrizhetjük a két változó függetlenségét abban az esetben, ha az átlag vagy a szórás eltér egymástól.

Példa: diszkrét változók

A következő oldalon láthatjuk két változó egyidejű gyakoriság-eloszlását. Az egyik változó a vizsgált személyek hitelkártyáinak a száma, míg a másik az, hogy a vizsgált személyek hányszor vásárolnak kártyával egy hét alatt. Láthatjuk a peremeloszlásokat is:

Táblázat 1.6.6. – Példa: A gyakoriságok és perem-gyakoriságok táblázata

	Vásárlások száma egy hét alatt					
Kártyák száma	0	1	2	3	4	perem
1	10	18	7	5	0	40
2	1	4	10	7	2	24
3	1	2	8	11	4	26
perem	12	24	25	23	6	90

Táblázat 1.6.7. – Példa (folytatás): A gyakoriság-eloszlás és a peremeloszlások

	Vásárlások száma egy hét alatt					
Kártyák száma	0	1	2	3	4	perem
1	0.111	0.200	0.078	0.056	0.000	0.444
2	0.011	0.044	0.111	0.078	0.022	0.267
3	0.011	0.022	0.089	0.122	0.044	0.289
perem	0.133	0.267	0.278	0.256	0.067	1.000

Például 4.4% azok aránya akik 2 kártyával rendelkeznek és egy hét alatt egyszer vásároltak.

A hitelkártyák számának feltételes eloszlása:

- 0 vásárlás esetén: (10 / 12, 1/12, 1/12) = (0.833, 0.083, 0.083),
- 1 vásárlás esetén: (18 / 24, 4 / 24, 2 / 24) = (0.750, 0.167, 0.083),
- 2 vásárlás esetén: (7 / 25, 10 / 25, 8 / 25) = (0.280, 0.400, 0.320).

A vásárlások számának feltételes eloszlása

- 1 kártya esetén: (10 / 40, 18 / 40, 7 / 40, 5 / 40, 0 / 40)

  = (0.250, 0.450, 0.175, 0.125, 0.000),

- 2 kártya esetén: (1 / 24, 4 / 24, 10 / 24, 7 / 24, 2 / 24)

  = (0.042, 0.167, 0.417, 0.292, 0.083).

A feltételes eloszlás átlaga diszkrét változókra

Ha az X változó értékei 0,1,...,k és az Y változó értékei 0,1,...,m, az X átlaga ha Y = j :

$$\sum_{i=1}^{k} i \; f_{i|j}.$$

Példa. A vásárlások számának átlaga 1 valamint 2 kártya esetén:

- 1 kártya esetén: 0×10/40 + 1×18/40 + 2×7/40 + 3×5/40 + 4×0/40 = 47/40 = 1.175,
- 2 kártya esetén: 0×1/24 + 1×4/24 + 2×10/24 + 3×7/24 + 4×2/24 = 56/24 = 2.1333.

Független-e a két változó? Nem, mert a feltételes átlagok különböznek egymástól.

Két folytonos változó egyidejű gyakoriság-eloszlása

Valamennyire hasonló helyzetben vagyunk mint egy folytonos változó esetén – a megfigyelések mindegyike általában egyszer fordul elő. Ezért hasonlóan járunk el mint egy változó esetén. Mindkét változó terjedelmét kisebb intervallumokra osztjuk. Megszámoljuk a megfigyeléseket az intervallumok mindegyik páronkénti kombinációjában.

Táblázat 1.6.8. – Példa: 40 megfigyelés (Hollandiában 1996-ban eladott) autó árakra (ezer euróban) és a motorjaik teljesítményére (kW-ban):

	Ár	Teljesít-mény		Ár	Teljesít-mény		Ár	Teljesít-mény		Ár	Teljesít-mény
1:	35	105	11 :	23	86	21 :	22	92	31 :	23	110
2:	30	80	12 :	23	75	22 :	23	100	32 :	35	110
3:	15	66	13 :	24	125	23 :	37	110	33 :	22	110
4:	38	142	14 :	35	55	24 :	20	66	34 :	27	107
5:	16	60	15 :	23	66	25 :	22	66	35 :	19	85
6:	30	125	16 :	28	107	26 :	14	55	36 :	19	74
7:	21	85	17 :	36	110	27 :	21	74	37 :	29	108
8:	21	85	18 :	31	85	28 :	21	66	38 :	38	128
9 :	20	55	19 :	18	66	29 :	26	77	39 :	19	66
10 :	19	66	20 :	22	51	30 :	27	110	40 :	20	85

Az árak terjedelme [14,38]. Ezt felosztjuk a következő intervallumokra:

[14,18], (18,23], (23,28], (28,33], (33,38].

A teljesítmény terjedelme [51,142]. Ezt a következő intervallumokra osztjuk:

[51,75], (75,100], (100,125], (125,142].

Megszerkesztjük a táblázatot az intervallum-kategóriákra, és meghatározzuk a gyakoriságokat. Az intervallumok számát mi határozzuk meg, akárcsak egy változó esetén, de itt jobb valamennyivel kevesebb intervallumot venni mint egy változó esetén.

Táblázat 1.6.9. – Példa (folytatás): A gyakoriságok táblázata

Ár-kategóriák	Teljesítmény-kategóriák			
	[51,75]	(75,100]	(100,125]	(125,142]
[14,18]	4	0	0	0
(18,23]	10	6	2	0
(23,28]	1	2	4	0
(28,33]	0	2	2	0
(33,38]	1	0	4	2

Relatív gyakoriságok

A gyakoriságokat elosztva a megfigyelések számával kapjuk a relatív gyakoriságokat:

Táblázat 1.6.10. – Példa (folytatás): A relatív gyakoriságok táblázata

Ár-kategóriák	Teljesítmény-kategóriák			
	[51,75]	(75,100]	(100,125]	(125,142]
[14,18]	0.100	0	0	0
(18,23]	0.250	0.150	0.050	0
(23,28]	0.025	0.050	0.100	0
(28,33]	0	0.050	0.050	0
(33,38]	0.025	0	0.100	0.050

Például 15% azok aránya ahol a teljesítmény 75 és 100 kW közötti, míg az ár 18 és 23 ezer euró közötti.

Perem- és feltételes gyakoriság-eloszlások

Mindkettőt ugyanúgy számítjuk ki mint nominális és diszkrét változókra. A perem-eloszlások az előző táblázat sorainak és oszlopainak az összege. A feltételes eloszlást egy bizonyos változó bizonyos intervallumára a megfigyelések relatív gyakoriságai adják, ha a többi intervallumban lévő megfigyelésektől eltekintünk.

Perem-eloszlások folytonos változókra

A következő táblázat tartalmazza a perem-gyakoriságokat:

Táblázat 1.6.11.- Példa (folytatás): A perem-gyakoriságok táblázata

Ár-kategóriák	Teljesítmény-kategóriák				
	[51,75]	(75,100]	(100,125]	(125,142]	**Perem**
[14,18]	4	0	0	0	4
(18,23]	10	6	2	0	18
(23,28]	1	2	4	0	7
(28,33]	0	2	2	0	4
(33,38]	1	0	4	2	7
Perem	16	10	12	2	40

Feltételes eloszlások folytonos változókra

A teljesítmény feltételes eloszlása a (18,23] árkategóriára: (10/18, 6/18, 2/18, 0).

A teljesítmény feltételes eloszlása a (33,38] árkategóriára: (1/7, 0, 4/7, 2/7).

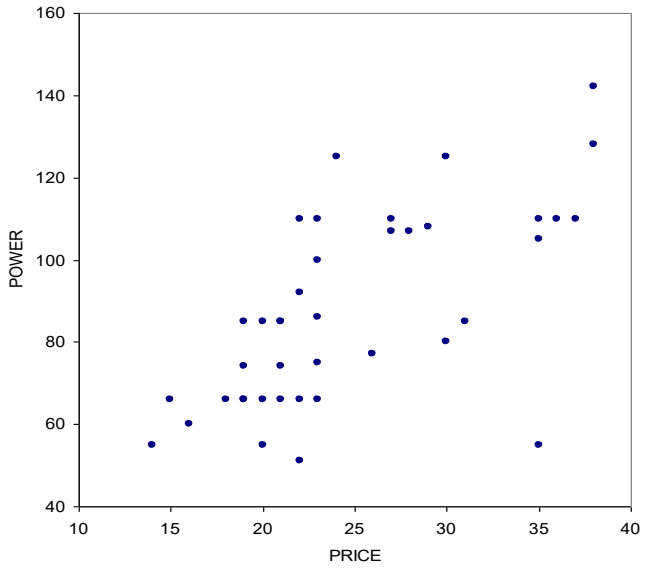
Az ár feltételes eloszlása a (75,100] teljesítménykategóriára: (0, 6/10, 2/10, 2/10, 0).

Az ár feltételes eloszlása a (100,125] teljesítménykategóriára: (0, 2/12, 4/12, 2/12, 4/12).

Változók közötti összefüggések ábrázolása: a szóródási kép

Minden megfigyelést a változók koordináta-rendszerében egy ponttal ábrázolunk. A következő oldal az autóárak és teljesítmények szóródási képét mutatja. Láthatjuk, hogy az árak megközelítőleg lineárisan változnak az autók teljesítményének függvényében, vagyis a ponthalmazon áthúzható egy egyenes. Azt mondjuk, hogy az autók ára és teljesítménye között pozitív lineáris összefüggés van.

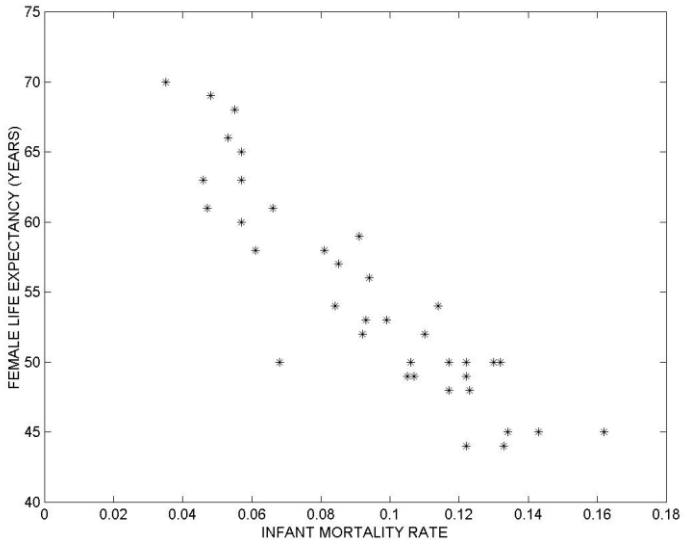
Ábra 1.6.1. - Példa: szóródási kép az autók árára és teljesítményére



Másfajta összefüggések is lehetségesek. A következő oldalakon példákat láthatunk a következőkre:

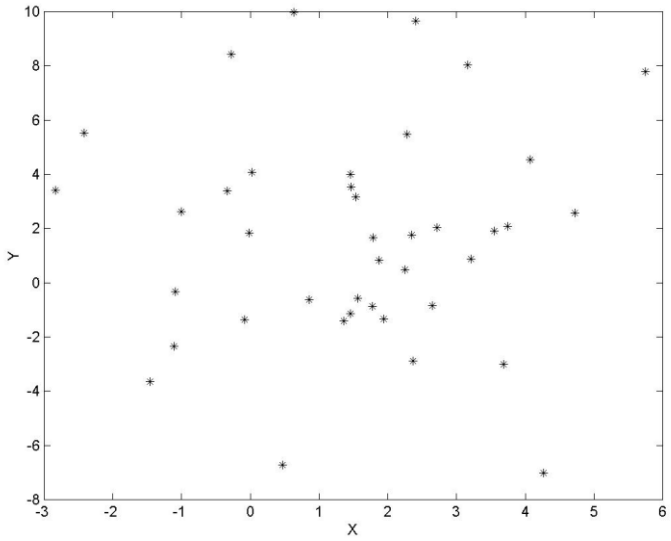
1. Negatív lineáris összefüggés
2. Nincs összefüggés
3. Nemlineáris összefüggés

Ábra 1.6.2. - Példa: szóródási kép – negatív lineáris összefüggés



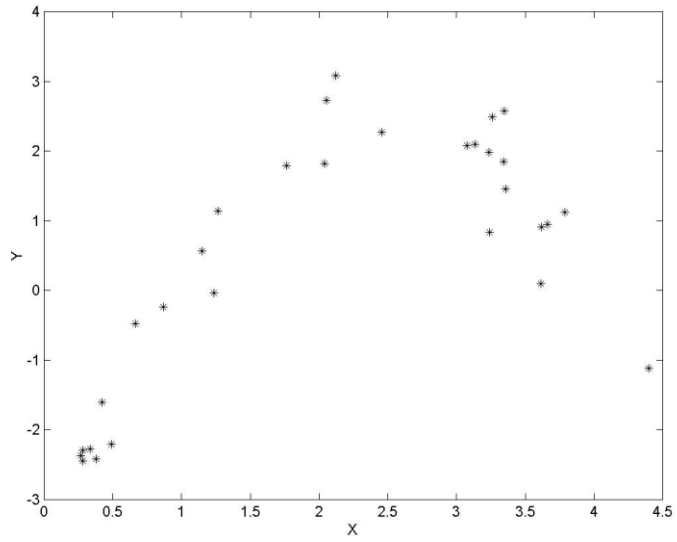
Az összefüggés negatív, mert a vízszintes tengelyen lévő változó növekedése a függőleges tengelyen lévő változó csökkenésével jár együtt. Az összefüggés lineáris, mert a pontokon áthúzható egy egyenes.

Ábra 1.6.3. - Példa: szóródási kép – nincs összefüggés



Egy ilyen szóródási kép nem mutat összefüggést a két változó között, mert az x változó kicsi, közepes és nagy értékeire is az y lehet kicsi, közepes vagy nagy.

Ábra 1.6.4. - Példa: szóródási kép – nemlineáris összefüggés



Ezen a szóródási képen lehet egy görbe vonalat húzni a pontok között, viszont egyenest kevésbé. A képzeletbeli görbe vonal azt mutatja, hogy van összefüggés a változók között, de az nem lineáris.

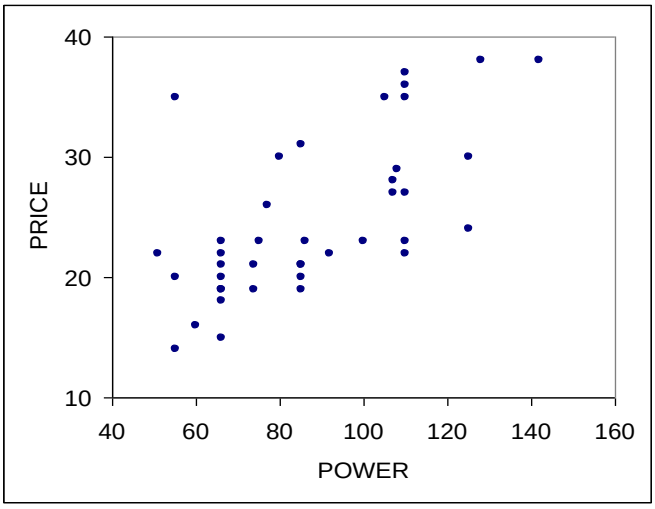
Abban az esetben amikor az egyik változó határozza meg a másikat, de ez fordítva nem áll fenn, akkor az előbbit az **x**-tengelyen míg az utóbbit az **y**-tengelyen ábrázoljuk. A meghatározó változót <u>független változó</u>nak (ez a fogalom más mint két változó függetlensége) és a meghatározott változót <u>függővá</u>ltozónak nevezzük.

Példák:

- apák (független) és fiúk (függő) magassága,
- tanulási évek száma és fizetés,
- autók ára és teljesítménye.

Az utóbbi példa szóródási képe az alapján, hogy a motor teljesítménye határozza meg az árat:

Ábra 1.6.5. – Az ár és motor teljesítmény összefüggését ábrázoló szóródási kép



Két folytonos változó függetlensége

Két folytonos változó X, Y független ha

- X összes feltételes eloszlása egyenlő az X peremeloszlásával

vagy

- Y összes feltételes eloszlása egyenlő az Y peremeloszlásával

Két egymástól független változó szóródási képe azt mutatja, hogy nincs összefüggés köztük. Az autóár és teljesítmény példában a megfigyelésektől eltekintve a két változó nem független, mert nagyobb teljesítményű autók drágábbak. Ezt láthattuk a szóródási képből is, ami lineáris összefüggést mutat. Ugyanezt láthatjuk a feltételes eloszlásokból is. Például, az ár feltételes eloszlása:

- ha a teljesítmény (50,75] között van: (4/16, 10/16, 1/16, 0, 1/16),
- ha a teljesítmény (100,125] között van: (0, 2/12, 4/12, 2/12, 4/12).

A kettő nem egyenlő, ezért a teljesítmény és az ár nem függetlenek egymástól.

Ha az ár és a teljesítmény függetlenek volnának (vagy nem lenne köztük semmilyen összefüggés), akkor mindegy lenne, hogy melyik teljesítmény-kategóriát vesszük. Vagyis, bármelyik teljesítmény-kategóriára, az árak feltételes eloszlása ugyanaz lenne. Következésképpen, az árak feltételes eloszlása ugyanaz lenne mint a peremeloszlásuk

A kovariancia

A kovariancia egy statisztikai mutató, ami mutatja, hogy két változó között van-e lineáris összefüggés. Kvantitatív változókra van értelme.

Az $(x_1,y_1),\ldots,(x_n,y_n)$ megfigyelésekre a kovariancia:

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- Pozitív lineáris összefüggés esetén cov(x,y) > 0.
- Negatív lineáris összefüggés esetén cov(x,y) < 0.
- Ha nincs lineáris összefüggés, akkor cov(x,y) =0.

A korrelációs együttható

A korrelációs együttható a kovarianciához hasonlóan a lineáris összefüggést méri, de megadja annak mértékét is. A kovariancia függ a mértékegységtől, ezért a mértékegység változásával (pl. lej-euró) a kovariancia is változik, noha a lineáris összefüggés mértéke ugyanaz marad.

A korrelációs együttható

$$r_{xy} = \frac{cov(x,y)}{s_x\ s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}}.$$

nem változik ilyenkor. A korrelációs együttható fontos tulajdonságai:

- r_{xy} mindig egy –1 és +1 közötti szám.
- Negatív, ha negatív lineáris összefüggés van a változók között.
- Pozitív, ha pozitív lineáris összefüggés van.
- Ha az összes megfigyelés egy felfele menő egyenesen van, akkor $r_{xy} = 1$.
- Ha az összes megfigyelés egy lefele menő egyenesen van, akkor $r_{xy} = -1$.
- Minél közelebb van r_{xy} a –1-hez vagy a +1-hez, annál erősebb a lineáris összefüggés.

Példa. Apák (X) és fiúk (Y) magassága:

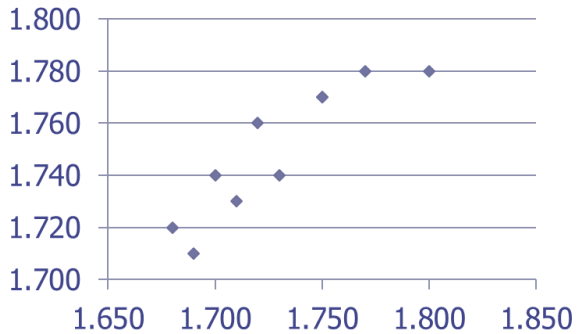
X: 1.70, 1.77, 1.68, 1.75, 1.80, 1.75, 1.69, 1.72, 1.71, 1.73

Y: 1.74, 1.78, 1.72, 1.77, 1.78, 1.77, 1.71, 1.76, 1.73, 1.74

Ekkor $\bar{x} = 1.73$, $\bar{y}=1.75$, $s_x=0.036$, $s_y=0.024$, $cov(x,y) =0.00078$ és

$$r_{xy} = \frac{cov(x,y)}{s_x\ s_y} = \frac{0.00078}{0.036 \times 0.024} = 0.903.$$

Ábra 1.6.6. - A szóródási kép



I.7 A regressziós egyenes

Ebben a fejezetben továbbra is két változó közötti összefüggést tanulmányozzuk, viszont ezúttal úgy, hogy előre feltételezzük, hogy lineáris összefüggés van közöttük. A regressziós egyenes tulajdonképpen a pontokon áthúzható egyik egyenes, amely függőleges távolságokban mérve a lehető legközelebb van a pontokhoz. Alább meg fogjuk látni, hogy a regressziós egyenest meghatározó x együtthatója kapcsolódik a korrelációs együtthatóhoz.

A regressziós egyenes megmutatja, hogy az egyik változó (Y) hogy változik a másik függvényében (X). Az X független változó, míg az Y függő változó.

Tegyük fel, hogy a megfigyeléseink $(x_1,y_1),\ldots,(x_n,y_n)$. A regressziós egyenes a ,,legjobb" olyan $y = a + bx$ egyenes ami leírja az Y megfigyeléseinek a változását az X megfigyelései függvényében. Az i megfigyelés reziduuma az y_i megfigyelés értékének és a neki megfelelő pont függőleges koordinátájának különbsége:

$$\text{i reziduuma} = e_i = y_i - (a + bx_i).$$

A legkisebb négyzetek kritérium

Egy olyan egyenes, amely jól leírja az X és Y közötti összefüggést, úgy határozza meg az *a*-t és *b*-t hogy a reziduumok a lehető legkisebbek legyenek. Ezt pédául úgy érhetjük el, hogy a reziduumok négyzetének összegét minimalizáljuk:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Ez a legkisebb négyzetek kritérium. Ahhoz, hogy megkapjuk az egyenest, meg kell határozzuk az *a* és *b* paraméterek értékét, amely minimalizálja a reziduumok négyzetének összegét. Ezek a következők:

$$b = \frac{cov(x,y)}{s_x^2} = r_{xy}\frac{s_y}{s_x},$$
$$a = \bar{y} - b\bar{x}.$$

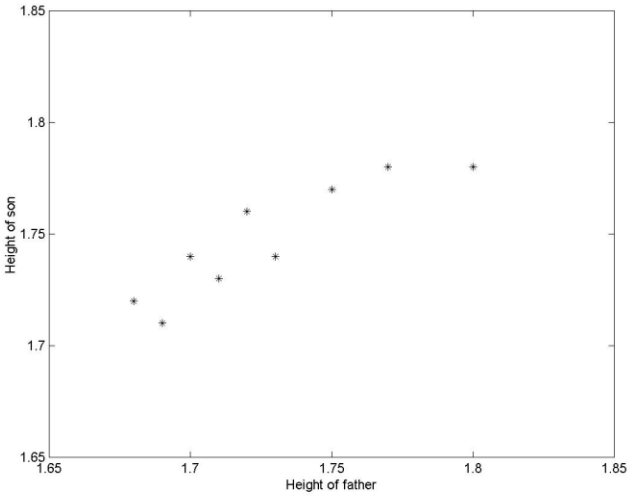
A regressziós egyenes az $y = a + bx$ egyenes erre az *a* és *b* értékre.

Példa: apák és fiúk magassága. Az alábbi táblázat tartalmazza 10 apa és legnagyobb fiának magasságát

Táblázat 1.7.1. – Példa: apák és fiúk magassága

Apa magassága (m)	Fiú magassága (m)	Apa magassága (m)	Fiú magassága (m)
1.70	1.74	1.75	1.77
1.77	1.78	1.69	1.71
1.68	1.72	1.72	1.76
1.75	1.77	1.71	1.73
1.80	1.78	1.73	1.74

Ábra 1.7.1. - A szóródási kép:



A regressziós egyenes paramétereinek kiszámításához a következő mennyiségekre van szükség:

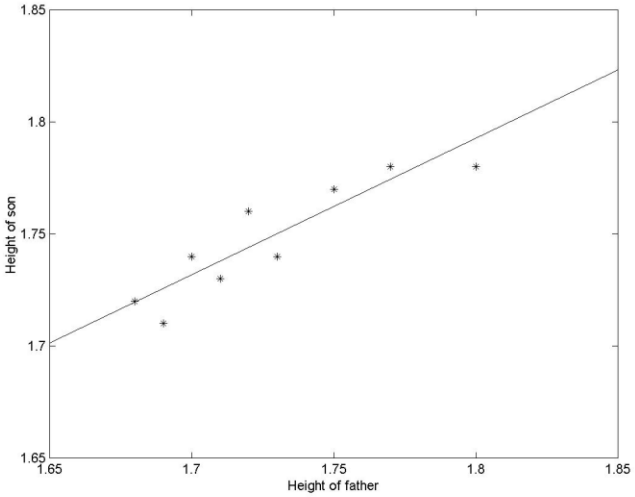
$$\bar{x} = 1.73, \quad s_x = 0.036, \quad \bar{y} = 1.75, \quad s_y = 0.024, \quad r_{xy} = 0.903.$$

Tehát a paraméterek:

$$b = r_{xy}\frac{s_y}{s_x} = 0.903\frac{0.024}{0.036} = 0.602,$$
$$a = \bar{y} - b\bar{x} = 1.75 - 0.602 \times 1.73 = 0.708.$$

A következő szóródási kép feltünteti a regressziós egyenest is. Láthatjuk, hogy az egyenes valóban nagyon közel van a szóródási kép mindegyik pontjához.

Ábra 1.7.2. – A szóródási kép és a regressziós egyenes



Előrejelzés

Ha nem lennének megfigyeléseink az apák magasságáról, akkor egy fiú magasságát a többi fiú átlagmagasságával becsülnénk. Ha vannak megfigyeléseink az apák magasságáról is, akkor előre tudjuk jelezni egy fiú magaságát az apja magassága alapján:

$$\hat{y} = a + bx$$

Ez az előrejelzés pontosabb mint az átlagmagassággal történő becslés.

Regresszió és korreláció

Amint láttuk, mind a regresszió, mind a korreláció két változó közötti összefüggést vizsgálja. A korrelációs együttható a két változót **szimmetrikusan** kezeli, vagyis $r_{xy} = r_{yx}$. Regresszió esetén megkülönböztetünk függő és független változót, amelyek különbözőképpen jelennek meg a számítások és a grafikus ábrázolás során. Az Y regressziós egyenese X függvényében **nem ugyanaz** mint az X regressziós egyenese Y függvényében!

I.8 Index-számok

Ez a fejezet egy időrendben mért változó leírásával foglalkozik. Egy egyszerű módszert mutat be, amelyik mindegyik megfigyeléshez egy számot rendel úgy, hogy egy adott viszonyítási időpontban ez a szám pontosan 100. Meg fogjuk látni, hogy ez az egyszerű ötlet megkönnyíti az időrendben mért változó leírását.

Az időrendi mérések vonatkozhatnak napokra, hetekre, hónapokra, negyedévekre, évekre, stb. Időrendben mért megfigyeléseket idősornak nevezzük. Fogjuk tanulmányozni például, hogy:

- a különböző árindexeket hogy számítják ki,
- Az inflációt milyen módszerrel számítják ki.

Eladási index

Eladási megfigyelések változásának leírására használják. Példaként tekintsük a következő táblázatot, ami 7 eladási megfigyelést tartalmaz:

Táblázat 1.8.1. - Példa

Idő (hetek)	Eladások
1	11.5
2	14.3
3	12.6
4	10.7
5	13.2
6	10.9
7	13.4

Az eladási megfigyelések változásának egy egyszerű leírási módja az, hogy egy bizonyos periódushoz viszonyítva számítsuk ki a megfigyeléseket. Ehhez először kiválasztunk egy alapperiódust. Gyakran az alapperiódus az első periódus, de ez nem kötelező.

A t-edik periódus eladási indexe:

$$\frac{\text{eladások t-ben}}{\text{eladások az alapperiódusban}} \times 100.$$

A következő táblázat tartalmazza az eladási indexeket:

Táblázat 1.8.2. – Az eladási index kiszámítása

Idő	Eladások	Eladási index
1	11.5	(11.5 / 11.5) x 100 = **100.0**
2	14.3	(14.3 / 11.5) x 100 = **124.3**
3	12.6	(12.6 / 11.5) x 100 = **109.6**
4	10.7	(10.7 / 11.5) x 100 = **93.0**
5	13.2	(13.2 / 11.5) x 100 = **114.8**
6	10.9	(10.9 / 11.5) x 100 = **94.8**
7	13.4	(13.4 / 11.5) x 100 = **116.5**

Miért hasznosak az index-számok? Az index-számok megmutatják az idősor változását az alapperiódushoz viszonyítva. Könnyen kiszámítható az idősor változása más periódusokhoz képest is. Megkönnyítik több idősor változásának összehasonlítását.

Árindexek

Amikor a megfigyelések árak (€, RON, stb.), az indexet árindexnek nevezzük.

Táblázat 1.8.3. – Példa árindexre

Idő	1995	1996	1997	1998	1999	2000
1 liter benzin ára	€0.70	€0.74	€0.73	€0.77	€0.81	€0.82
Árindex (alapév=1995)	100	105.7	104.3	110.0	115.7	117.1

Az árindexek azonnal mutatják, hogy az alapév után az árak nőttek, és azt is, hogy hány százalékkal nőttek.

A alábbiakban egy példa segítségével bemutatjuk, hogy hogyan végezhetjük idősorok összehasonlítását az árindexek segítségével.

Táblázat 1.8.4. - Példa: Három idősor, év végi lakásárak mediánja 1000 euróban

Év	Reykjavik	London	Lagos (Nigéria)
1996	85	130	10
1997	84	132	13
1998	86	140	19
1999	87	155	26
2000	85	180	30
2001	88	220	36
2002	90	265	45

Ha kiszámítjuk mindhárom árindexet, láthatjuk, hogyan változtak az árak egymáshoz viszonyítva:

Táblázat 1.8.5. – Példa (folytatás): A kiszámított árindexek:

	Lakásárak árindexe (alapév = 1996)		
Év	Reykjavik	London	Lagos
1996	100.0	100.0	100.0
1997	98.8	101.5	130.0
1998	101.2	107.7	190.0
1999	102.4	119.2	260.0
2000	100.0	138.5	300.0
2001	103.5	169.2	360.0
2002	105.9	203.8	450.0

Alapperiódushoz viszonyított százalékos változás

Az alapperiódushoz viszonyított százalékos változás a t-edik periódusban:

$$\frac{\text{t-beli érték} - \text{alapperiódusbeli érték}}{\text{alapperiódusbeli érték}} \times 100.$$

Az index segítségével a százalékos változás:

index – 100.

A fenti példában 1996 és 1999 között Lagosban az árak 160%-kal nőttek (= (26-10)/10 × 100%).

Az idősor százalékos változása bármilyen periódushoz képest

Az idősor százalékos változása a t-edik periódusban az s-edik periódushoz képest:

$$\frac{\text{t-beli érték} - \text{s-beli érték}}{\text{s-beli érték}} \times 100.$$

Index-számok segítségével ugyanezt így számíthatjuk ki:

$$\left(\frac{\text{t-beli index}}{\text{s-beli index}} - 1\right) \times 100.$$

A fenti példában 1998 és 2001 között Reykjavikban a lakásárak (103.5/101.2 - 1) × 100%-kal változtak.

Táblázat 1.8.6. - Gyakorlat. Az A és B vállalat részvényeinek az árai (euróban) így alakultak:

Év	1998	1999	2000	2001	2002
A vállalat	0.50	0.60	0.75	0.65	0.55
B vállalat	10.00	12.30	12.50	12.60	12.60

Számítsuk ki az árindexeket 1998–at véve alapul. Ha 1998-ban vásároltunk részvényt €100-ra mindkét vállalat részvényéből, mennyi a nyereségünk 2002-ben? Mennyi a B vállalat részvényeinek a százalékos változása 1999 és 2001 között?

Indexek egyszerű átlagolása

Tegyük fel, hogy a különböző lakásárakat együtt szeretnénk tanulmányozni, vagyis egy indexet meghatározni a három árindexből. Ez például az infláció tanulmányozásához lehet hasznos. Ehhez átlagoljuk a három lakásárat.

Többféleképpen lehet átlagolni az indexeket. Az átlagolás legegyszerűbb módja az egyszerű átlagolás ami az indexek átlagának kiszámítása. Ha K indexünk van, az indexek átlaga a t periódusban:

$$\frac{1}{K}\sum_{k=1}^{K} k \text{ index t-ben} = \frac{1}{K}\sum_{k=1}^{K} \frac{k \text{ értéke t-ben}}{k \text{ értéke az alapperiódusban}} \times 100.$$

A következő táblázat mutatja a három árindex egyszerű átlagát:

Táblázat 1.8.7. – Példa (folytatás): Az árindexek átlagának az alakulása

Év	Reykjavik	London	Lagos	Átlag index
1996	100.0	100.0	100.0	**100.0**
1997	98.8	101.5	130.0	**110.1**
1998	101.2	107.7	190.0	**133.0**
1999	102.4	119.2	260.0	**160.5**
2000	100.0	138.5	300.0	**179.5**
2001	103.5	169.2	360.0	**210.9**
2002	105.9	203.8	450.0	**253.2**

Indexek súlyozott átlagolása

Az egyszerű átlagolás ugyanolyan fontosságot tulajdonít mindegyik indexnek. Ez sokszor nem ésszerű, mert például előfordulhat, hogy Londonban és Lagosban sokkal több lakást adnak el mint Reykjavikban.

Egy ésszerűbben kiszámított átlag figyelembe veszi az eladott lakások számát. Az indexek súlyozott átlaga a t-edik periódusban

$$\sum_{k=1}^{K} (k \text{ index súlya}) \times (k \text{ index } t\text{-ben}).$$

Megjegyezzük, hogy az egyszerű átlagolás a súlyozott átlagolás speciális esete, amikor mindegyik index súlya egyenlő 1/K -val. Amikor az indexek bizonyos termékek árindexei, a súlyok kiszámításának egy ésszerű módja a kiadások aránya az összes kiadáshoz viszonyítva az alapperiódusban, vagyis

$$k \text{ index súlya} = \frac{\text{kiadások } k\text{-ra az alapperiódusban}}{\text{összes kiadás az alapperiódusban}}.$$

Amikor az indexek súlyát a fenti módszerrel számítjuk ki, a súlyozott indexet Laspeyres indexnek nevezzük.

Laspeyres index

Tegyük fel, hogy az alapévben (1996) 250 lakást adtak el Reykjavikban, 13500-at Londonban és 8600-at Lagosban. Ezért az összes kiadás ebben az évben (1000€ -ban):

(250 x 85) + (13500 x 130) + (8600 x 10) = **1 862 250**.

A Reykjavik-i árindex súlya (250 x 85)/1 862 250 = **0.012**, a londonié (13500 x 130) / 1862250 = **0.942** és a Lagos-ié (8600 x 10) / 1862250 = **0.046**.

Táblázat 1.8.8. – Példa (folytatás): A Laspeyres indexek

Év	Árindexek			
Súlyok:	Reykjavik *0.012*	London *0.942*	Lagos *0.046*	**Laspeyres index**
1996	100.0	100.0	100.0	**100.0**
1997	98.8	101.5	130.0	**102.8**
1998	101.2	107.7	190.0	**114.6**
1999	102.4	119.2	260.0	**125.5**
...	...	...	...	**...**

Gyakorlat. Vegyük most újra a vállatok példáját. Ha az alapévben (1998-ban) az A vállalat 100 000 részvényét és a B vállalat 25 000 részvényét adták el, számítsuk ki a részvények Laspeyres árindexét 1999-re a két részvényre.

Paasche index

Láttuk, hogy a Laspeyres index az alapperiódus kiadásait veszi figyelembe a súlyok kiszámításánál. Lehet venni más periódust is erre a célra; várhatóan más értékeket kapunk a súlyoknak. Ha az utolsó periódust vesszük alapperiódusnak, vagyis,

$$k \text{ index súlya} = \frac{\text{kiadások } k\text{-ra az utolsó periódusban}}{\text{összes kiadás az utolsó periódusban}},$$

akkor az így kapott súlyozott indexet Paasche indexnek nevezzük.

A fogyasztói árindex

Az országos statisztikai hivatalok általában havonta kiszámítják és közzéteszik. Az árak változását méri az átlagos háztartásokra. A fogyasztói árindex egy **Laspeyres index**, amelyet 8 termékcsoport árindexének súlyozott átlagaként számítanak ki.

Példaként tekintsük a következő termékcsoportokat a nekik megfelelő árindexek súlyaival, ahol az alapév 1992:

Táblázat 1.8.9. – Példa: Különböző termékcsoportok árindex súlyai

Termékcsoport	Súly (alapév =1992)
Élelem, ital és dohány	0.294
Szállítás és kommunikáció	0.165
Más kiadások	0.152
Ruhák és cipők	0.115
Lakás	0.103
Szabadidő, kultúra és képzés	0.073
Lakásfenntartás és más szolgáltatások	0.067
Egészség	0.031

Mindegyik termékcsoportra a statisztikai hivatal bizonyos fontos termékeket vesz figyelembe, amit termékkosárnak neveznek. A termékkosárban szereplő termékekre havonta lemérik az árakat és árindexet számolnak az alapperióduhoz képest. A fogyasztói árindex ezek Laspeyres indexe, amit az előző oldalon szereplő súlyokkal számítanak ki. Bizonyos idő után a statisztikai hivatal megváltoztatja az alapperiódust azért, hogy a súlyok jobban tükrözzék a kiadásokat a különböző termékekre.

Az éves infláció

Az éves infláció egy adott hónapban a fogyasztói árindex százalékos változása az egy évalószínűségi változóel korábbi fogyasztói árindex értékéhez viszonyítva. Ezt a számot közlik a sajtóban minden hónapban.

Tehát ha FAI_t jelöli a fogyasztói árindexet a t-edik hónapban, akkor az

$$\text{éves infláció } t\text{-ben} = \left(\frac{FAI_t}{FAI_{t-12}} - 1\right) \times 100.$$

Táblázat 1.8.10. – Az árindexek idősora

Hónap	Jan 99	Feb 99	Már 99	Apr 99	Máj 99	Jún 99	Júl 99
FAI	100.0	100.5	101.2	101.3	102.1	103.4	104.8
Hónap	Aug 99	Szep 99	Okt 99	Nov 99	Dec 99	Jan 00	Feb 00
FAI	104.9	106.1	108.0	109.3	110.1	111.9	112.6

A táblázatban levő árindexek alapján:

- 2000. januárban az infláció: ((111.9 / 100) – 1) x 100 = 11.9%.
- 2000. februárban az infláció: ((112.6 / 100.5) –1) x 100 = 12.0%.

I.9 Idősorok elemzése

Ebben a fejezetben egy idősor megfigyeléseinek a változását grafikus szempontból tanulmányozzuk. A két tulajdonság, amelyek grafikus szempontból a legfontosabbak, a trend és a szezonalitás. Az előbbi azt vizsgálja, hogy a megfigyelések hosszú távon milyen tendenciát mutatnak. Az utóbbi azt vizsgálja, hogy a megfigyelések mutatnak-e ismétlődést.

Egy idősor olyan megfigyelések összessége, amelyeket szabályos időközönként mérnek. Egy idősor jelölése $x_1, x_2, x_3,..., x_t,...$. Példák idősorokra:

- negyedéves GDP,
- hónapos infláció,
- éves GDP növekedés az Európai Únióban.

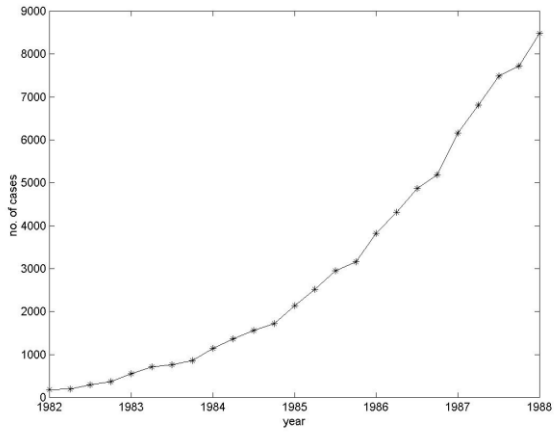
Az idősorokat általában vonalgrafikonnal ábrázoljuk, ahol az Ox tengelyen az időt, míg az Oy tengelyen a megfigyelések értékeit tüntetjük fel. Az így kapott pontokat egyenes vonalakkal kötjük össze. Alább három vonalgrafikonra adunk példát.

Az első negyedéves AIDS megbetegedések mutat 1982 – 1987 között az Egyesült Királyságban. A megfigyeléseket 1986 második negyedévéig a következő táblázat tartalmazza:

Táblázat 1.9.1. – Példa: Az AIDS megbetegedések negyed éves alakulása az AEÁ-ban

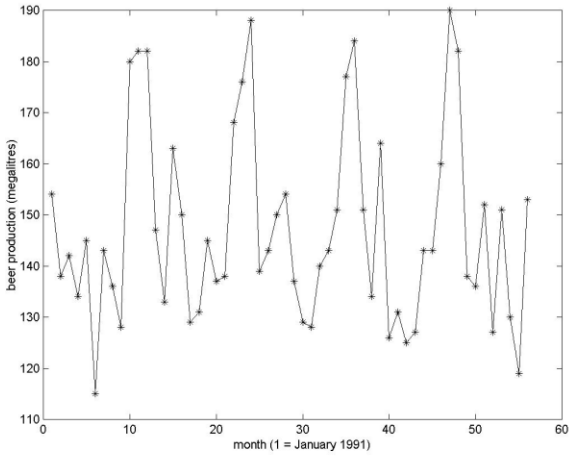
Negyedév	AIDS meg-betegedések	Negyedév	AIDS meg-betegedések
1982 Q1	185	1984 Q2	1369
1982 Q2	200	1984 Q3	1563
1982 Q3	293	1984 Q4	1726
1982 Q4	374	1985 Q1	2142
1983 Q1	554	1985 Q2	2525
1983 Q2	713	1985 Q3	2951
1983 Q3	763	1985 Q4	3160
1983 Q4	857	1986 Q1	3819
1984 Q1	1147	1986 Q2	4321

Ábra 1.9.1. - Az idősor vonalgrafikonja 1982 – 1987 között



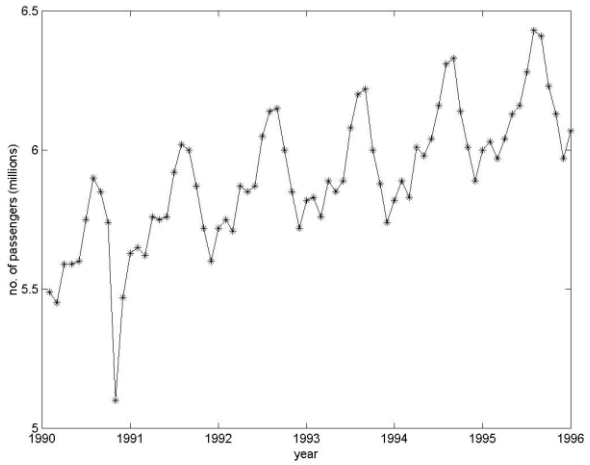
Egy másik vonalgrafikon a hónapos sörtermelést mutatja Ausztráliában 1991. Jan. és 1995. Aug. között:

Ábra 1.9.2. – Példa (folytatás): Az Ausztrália sörtermelés alakulása 1991 és 1995 között



A harmadik példa a nemzetközi légi utasok hónaponkénti számának vonalgarfikonja Spanyolországban 1990 – 1995 között:

Ábra 1.9.3. – A légi utasok számának alakulása 1990 és 1995 között



Idősorok osztályozása

Az idősorokat 2 fő osztályba sorolhatjuk:

- Stacionárius. Az átlag, a szórás és a korrelációk változatlanok maradnak az idő függvényében. Általában olyan gazdasági idősorok, amelyek különbséget vagy változást fejeznek ki (pl. GDP növekedési aránya bizonyos országokra bizonyos időszakokra).
- Nem stacionárius. Az átlag, a szórás vagy a korrelációk változnak az idő függvényében. Például, GDP, ipari termelés, hónapos sörtermelés, stb.

Trend

A trend egy idősor értékeinek a rendszeres hosszú távú változása az idő függvényében. Az AIDS idősor növekvő trendet mutat. A hónapos sörtermelés megfigyelések nem mutatnak semmilyen trendet. Kétfajta trend létezik:

- Determinisztikus trend: a trend ugyanaz marad az idő függvényében (pl. nemzetközi légi utasok száma).
- Sztochasztikus trend: a trend nagysága és iránya véletlenszerűen (előre ismeretlen módon) változik (pl. AIDS megbetegedések).

Szezonalitás

A hónapos sörtermelésnél láthatjuk, hogy a vonalgrafikon formája 12 hónaponként nagyjából ismétlődik. Vagyis, a legnagyobb értékek mindig október, november és decemberben vannak, a legkisebbek május, június és júliusban. Azt mondjuk, hogy egy idősor **szezonalitással rendelkezik** (vagy **szezonalitást mutat**) amikor az idősor vonalgrafikonja szisztematikusan ismétlődik. Gazdasági megfigyeléseknél a szezonalitás általában éves, ami azt jelenti, például, hogy negyedéves megfigyeléseknél négy periódusonként, míg hónapos megfigyeléseknél 12 periódusonként történik az ismétlődés.

Idősorok felbontása

A legtöbb idősor felbontható 3 komponensre:

- trend,
- szezonalitás,
- irreguláris vagy véletlenszerű komponens.

A felbontás során a komponensek összeadódnak:

$$\text{Megfigyelés} = \text{trend} + \text{szezonalitás} + \text{irreguláris komponens:}$$
$$x_t = T_t + S_t + I_t.$$

Ahhoz, hogy megértsük a megfigyelések változását (pl. hónapos ipari termelés), fontos a trend és a szezonalitás meghatározása. Ezt az eljárást az idősorok grafikus elemzésének nevezzük.

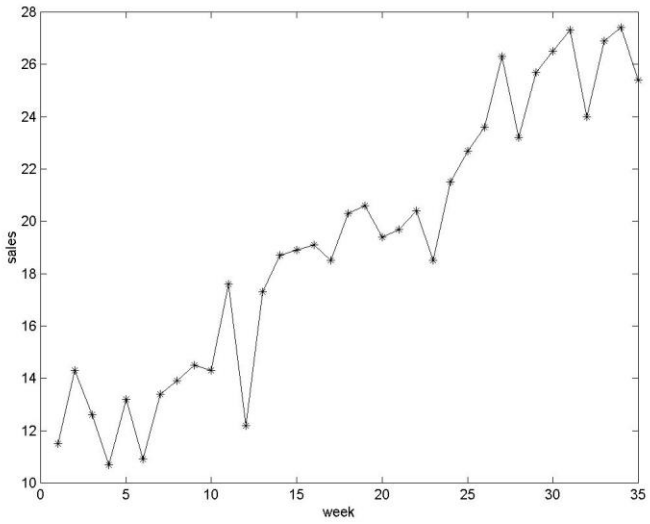
A trend elemzése

Ez az az eljárás, amellyel a trend komponenst meghatározzuk. Kezdjük egy olyan esettel, amelyben azt feltételezzük, hogy *nincs szezonalitás* (pl. AIDS megbetegedések), vagyis:

$$X_t = T_t + I_t.$$

Példa: egy termék heti eladásai. (szezonalitás nélküli idősor). Láthatjuk, hogy a megfigyelések hosszú távon (vagyis az első és az utolsó megfigyelés között) növekednek, ezért a vonalgrafikon alapján azt mondjuk, hogy az idősor növekvő trended mutat.

Ábra 1.9.4. – Példa: egy termék heti eladásai



Determinisztikus trend

A trend legegyszerűbb formája az egyenes, vagyis egy lineáris függvény (az idő függvényében):

$$T_t = a + bt.$$

Az *a* és *b* meghatározására használhatjuk a regressziós egyenest.

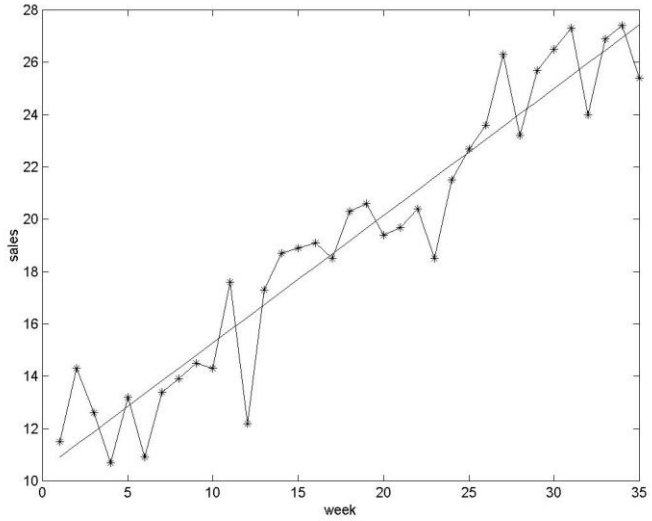
A függőváltozó az idősor $x_1,...,x_n$ megfigyelései, míg a független változó az idő t=1,2,...,35.

Tehát a regressziós egyenes paramétereit a (1,x_1=11.5), (2,x_2=14.3), ..., (35,x_{35}=25.4) megfigyelések alapján határozzuk meg a

$$b = \frac{cov(x,t)}{s_t^2}, a = \bar{x} - b\bar{t}$$

képletekkel. Azt kapjuk, hogy a regressziós egyenes T_t = 10.4 + 0.49 t, amit az alábbi grafikon tüntet fel:

Ábra 1.9.5. – Példa (folytatás): egy termék heti eladásai és a regressziós egyenes



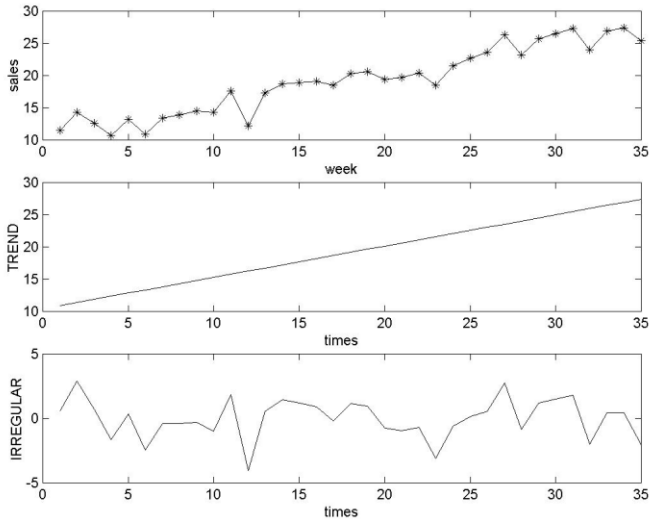
Ekkor az irreguláris komponens a megfigyelések és a trend közötti különbség, mivel nincs szezonális komponens, vagyis 0:

$$I_t = x_t - T_t = x_t - (a + bt).$$

Alább láthatjuk az idősor felbontását a trendre és az irreguláris komponensre. Tehát, a trendelemzés FONTOS ELVE: ha az idősor vonalgrafikonja lineáris trendet mutat, akkor ezt a regressziós egyenessel határozzuk meg.

Ha az ilyen módon meghatározott irreguláris komponens nem mutat semmilyen trendet mint az alábbi grafikonon, akkor ez azt jelenti, hogy helyesen végeztük a trendelemzést.

Ábra 1.9.6. – Példa (folytatás): egy termék értékesítése, trend és irreguláris komponens

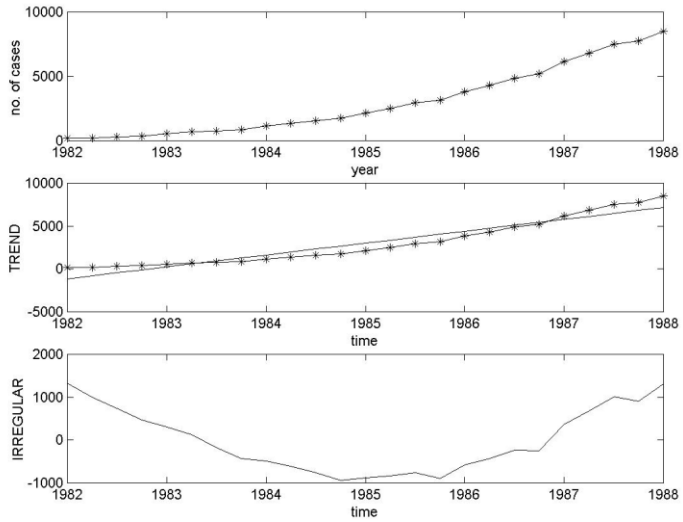


Sztochasztikus trend

Sok esetben a trend nem lineáris és a vonalgrafikon véletlenszerű változást mutat az idő függvényében. Ebben az esetben egy egyenes nem tudja meghatározni a trendet.

Példa: AIDS megbetegedések, ahol a trend növekszik. Az alábbi ábra mutatja, hogy a megfigyelések valóban jobban nőnek mint egy egyenes pontjai. Láthatjuk a regressziós egyenest és az irreguláris komponents is. Láthatjuk, hogy az irreguláris komponens nem váltakozik véletlenszerűen, ugyanis előbb csökkenő majd növekvő trendet mutat.

Ábra 1.9.7. – AIDS megbetegedések száma, trend és az irreguláris komponens



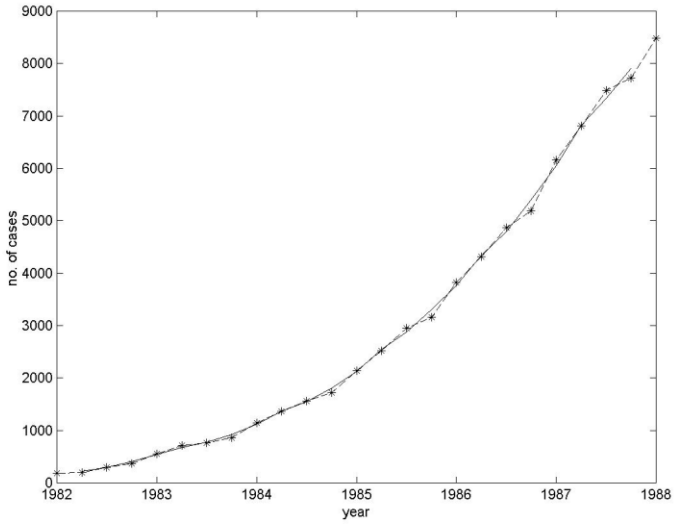
Ezen a ponton jobb ötletnek tűnik egy mozgó átlag alkalmazása. Vegyük az előző és utána következő megfigyelésektől függő mozgó átlagot, vagyis m_1, m_2, ..., m_t, ... , ahol

m_t az x_{t-1}, x_t és x_{t+1} átlaga.

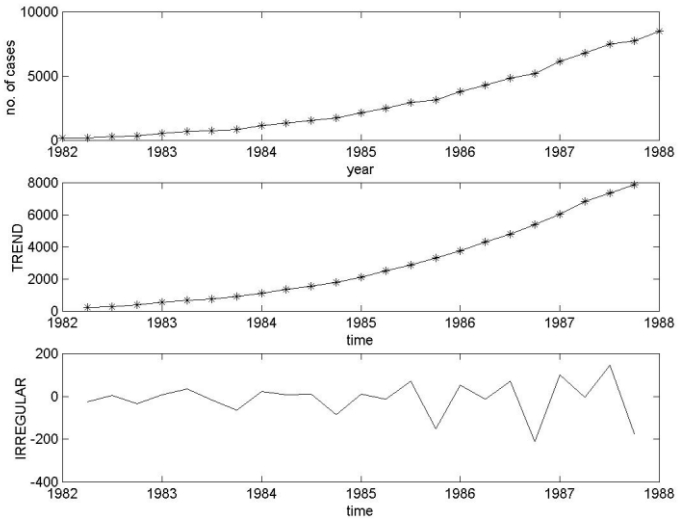
Az m_t mozgó átlagot a T_t trendjének tekintjük (tehát $x_t = m_t + It$). Az AIDS megbetegedések első 7 megfigyelésére a mozgó átlagok:

Táblázat 1.9.2. – Mozgó átlagok

x_t	185	200	293	374	554	713	763
m_t	—	226	289	407	547	676.7	777.7



Ábra 1.9.8. - A mozgó átlagok által meghatározott trend és irreguláris komponens:



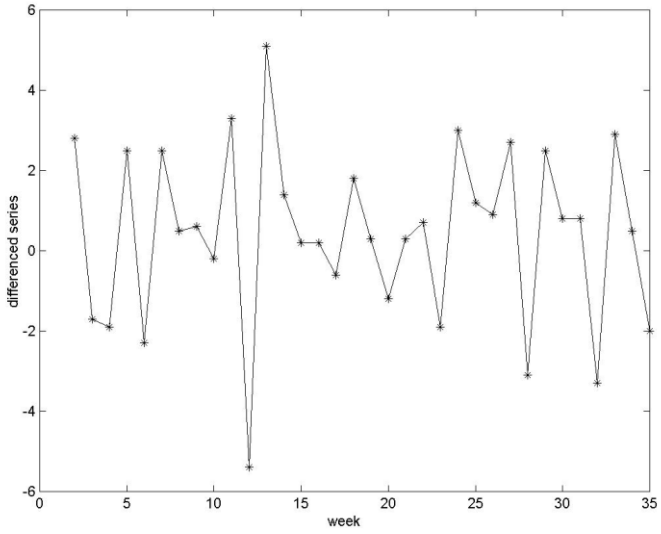
Sztochasztikus trend esetén akkor is trend nélküli idősort kaphatunk, ha az egymás utáni megfigyelések különbségét határozzuk meg. Az így kapott idősor $y_t = x_t - x_{t-1}$. Ez az eljárás használható lineáris trend esetén is. A heti termékeladások első 7 megfigyelésére a különbségek:

Táblázat 1.9.3. – A megfigyelések különbségei

x_t	11.5	14.3	12.6	10.7	13.2	10.9	13.4
y_t	—	2.8	–1.7	–1.9	2.5	–2.3	2.5

Megfigyelhetjük, hogy a különbségekből kapott idősor vonalgrafikonja nem mutat semmilyen trendet.

Ábra 1.9.9. – A különbségekből kapott idősor ábrája



Az egymás utáni megfigyelések különbségéből kapott idősor csak irreguláris komponenssel rendelkezik (vagyis nincs trend). A megfigyelések különbsége eltávolítja a trendet. Ezt az eljárást gyakran használják, mivel determinisztikus és sztochasztikus trend esetén is alkalmazható. A megfigyelések értelmezésénél figyelembe kell venni, hogy a megfigyelések különbségével van dolgunk.

Szezonalitás elemzése

Amikor egy idősor szezonalitással rendelkezik, az ismétlődések közötti időszakot ciklusnak nevezik. Ez általában 1 év, vagyis 12 megfigyelés (hónapos megfigyelésekre) és 4 megfigyelés (negyedéves megfigyelésekre). A ciklusokon belüli ugyanannak az időpontnak megfelelő megfigyelések hasonlóan viselkednek.

A szezonalitás elemzése szempontjából lényeges, hogy a ciklusokon belüli ugyanannak az időpontnak megfelelő megfigyelések átlagai mennyire különböznek az összes megfigyelés átlagától. A különböző átlagok adják a szezonális együtthatókat. Vagyis:

szezonális együttható = átlag a ciklusban – átlag.

Az alábbi táblázat tartalmazza a szezonális együtthatókat mindegyik hónapra:

Táblázat 1.9.4. – A szezonalitási együtthatók

	1991	1992	1993	1994	1995	Átlag	**szezonális együttható**
Jan	154	147	139	151	138	145.8	**–1.4**
Feb	138	133	143	134	136	136.8	**–10.4**
Már	142	163	150	164	152	154.2	**7.0**
Ápr	134	150	154	126	127	138.2	**–9.0**
Máj	145	129	137	131	151	138.6	**–8.6**
Jún	115	131	129	125	130	126.0	**–21.2**
...	...	...	...	...	...		
				Összes átlaga		147.2	

	1991	1992	1993	1994	1995	Átlag	**szezonális együttható**
...	...	...	...	...	...		
Júl	143	145	128	127	119	132.4	**–14.8**
Aug	136	137	140	143	153	141.8	**–5.4**
Szep	128	138	143	143		138.0	**–9.2**
Okt	180	168	151	160		164.8	**17.6**
Nov	182	176	177	190		181.3	**24.1**
Dec	182	188	184	182		184.0	**26.8**
				Összes átlaga		147.2	

Az idősor szezonális komponense a megfelelő S_t szezonális együttható, vagyis, ha a ciklus 12 megfigyelésből áll, S_1, S_2, ..., S_{12}, $S_{13} = S_1$, $S_{14} = S_2$,..., stb.

Egy szezonalitás-mentesített idősor az az idősor, amelyet úgy kapunk, hogy az eredeti idősorból kivonjuk a megfelelő szezonális együtthatót: $x_t - S_t$. A szezonalitás-mentesített idősornak csak trend és irreguláris komponense van. Egy ilyen idősort trendelemzésnek vetünk alá (mint előbb) így megkapjuk a trend, szezonális és irreguláris komponenseket.

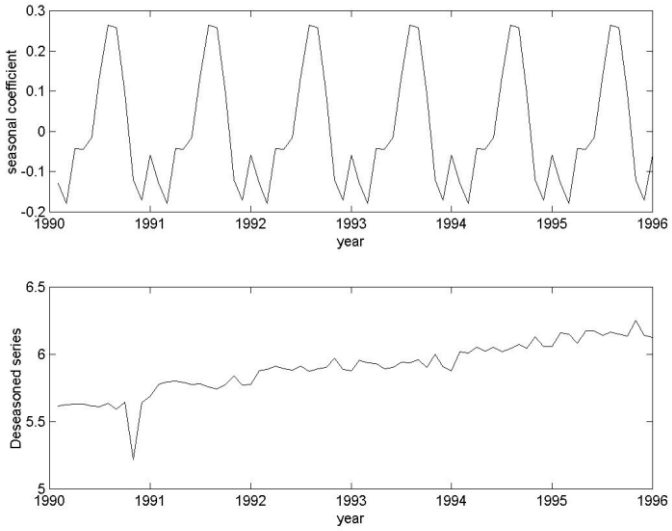
Példa: légi utasok. A vonalgrafikon alapján ennek az idősornak van trend komponense és szezonális komponense. A szezonális komponens ciklusa 12 megfigyelés (= 1 év). Először kiszámítjuk a szezonális együtthatókat:

Táblázat 1.9.5. – Példa: légi utasok adathalmazra számolt szezonalitási együtthatók

Hónap	Jan	Feb	Már	Ápr	Máj	Jún
Szezonális együttható	–0.13	–0.18	–0.04	–0.04	–0.02	0.14
Hónap	Júl	Aug	Szep	Okt	Nov	Dec
Szezonális együttható	0.26	0.26	0.09	–0.12	–0.17	–0.06

Az alábbi két ábra közül a felső a szezonális komponenst mutatja, míg az alsó a szezonalitás-mentesített idősort:

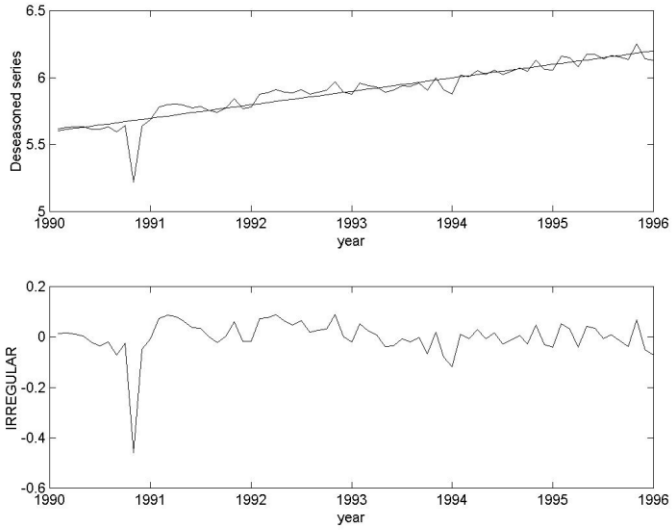
Ábra 1.9.10. – Szezonális komponens és a szezonalitás mentesített idősor



A szezonalitás-mentesített idősor trendjének meghatározása:

- A szezonalitás-mentesített idősor többé-kevésbé lineáris trenddel rendelkezik.
- Tehát meghatározzuk a szezonalitás-mentesített idősor $x_t - S_t$ regressziós egyenesét.
- A következő ábrák a szezonalitás-mentesített idősor regressziós egyenesét és a szezonalitás-mentesített idősor és a trend különbségét $x_t - S_t - T_t$ mutatják.
- A felbontás alapján ez az irreguláris komponens I_t.

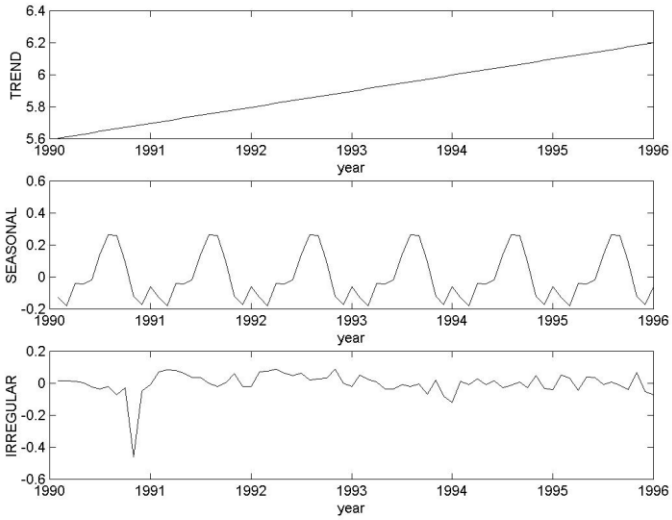
Ábra 1.9.11. - A regressziós egyenes és az irreguláris komponens



Tehát, az idősort felbontottuk (lásd az alábbi három vonalgrafikont):

- szezonális komponensre (középső vonalgrafikon),
- trendre, ami a szezonalitás-mentesített idősor regressziós egyenese (felső vonalgrafikon),
- irreguláris komponensre, ami az idősor mínusz a szezonális komponens és a trend (alsó vonalgrafikon).

Ábra 1.9.12. – Trend, szezonális komponens és irreguláris komponens



I.10 Gyakorlatok

1. A következő megfigyelések egy tisztítószer használati idejére vonatkoznak hetekben kifejezve 20 háztartásra:

7.4 4.5 8.0 9.5 8.4 8.2 7.8 8.2 7.5 7.6
8.1 8.5 7.5 8.9 7.6 7.1 5.5 8.5 7.7 8.7

a. Határozzuk meg a használati idő átlagát és szórását. (E: átlag 7.76; szórás 1.09)
b. A megfigyelések terjedelmét 4.5-től kezdve osszuk öt 1 hosszúságú intervallumra, és ezekre határozzuk meg a megfigyelések gyakoriság-eloszlását. (E: 2/20; 0/20; 4/20; 11/20; 3/20)
c. Határozzuk meg a megfigyelések mediánját és kvartiliseit. Van-e kiugró érték a megfigyelések között? (E: medián 7.9; alsó kvartilis 7.5; felső kvartilis 8.45; van kiugró érték a 4.5 és az 5.5)

2. Az alábbi megfigyelések 1980-ban és 1981-ben a Toyota márkájú autók havi összeladásai százezer dollárban kifejezve:
18, 20, 20, 21, 21, 20, 22, 14, 20, 20, 17, 17, 16, 18, 21, 21, 19, 20, 19, 12, 19, 20, 19, 18.

a. Határozzuk meg az eladások átlagát és szórását. (E: átlag 18.83; szórás 2.29)
b. A megfigyelések terjedelmét 12-től kezdve osszuk öt 2 hosszúságú intervallumra, és ezekre határozzuk meg a megfigyelések gyakoriság-eloszlását. (E: 2/24; 1/24; 5/24; 11/24; 5/24)
c. Határozzuk meg a megfigyelések mediánját és kvartiliseit. Van-e kiugró érték a megfigyelések között? (E: medián 19.5; alsó kvartilis 18.0; felső kvartilis 20.0; van kiugró érték a 12.0 és a 14.0)

3. 2013-ban a kormányzati adósságok a következőképpen alakultak a GDP-hez viszonyítva:

Amerikai Egyesült Államok	104.50%
Ausztria	74.20%
Egyesült Arab Emírségek	12.30%
Egyesült Királyság	90.10%
Franciaország	93.90%
Görögország	173.80%
Japán	243.20%
Kína	22.40%
Kuwait	5.30%
Lengyelország	57.50%
Magyarország	79.20%
Németország	78.10%
Olaszország	132.50%
Oroszország	13.40%
Románia	38.20%
Spanyolország	93.90%

Forrás: knoema.com

a. Határozzuk meg a megfigyelések átlagát és szórását! (E: átlag 82.03 %; szórás 60.94%)
b. A megfigyelések terjedelmét osszuk 4 egyenlő hosszúságú intervallumra, és ezekre határozzuk meg a megfigyelések gyakoriság-eloszlását! (E: 6/16; 7/16; 2/16; 1/16)
c. Határozzuk meg a megfigyelések mediánját és kvartiliseit. Van-e kiugró érték a megfigyelések között? (E: medián 78.65%; alsó kvartilis 30.3%; felső kvartilis 99.2%; van kiugró érték a 243.2%)

4. Megkérdeztünk 180 embert, hogy mennyit költenek havonta a telefonjukra és a hozzá köthető vásárlásokra (számla, részlet, applikációk, játékok). A kérdés válaszlehetőségeként 6 kategóriát adtunk meg, a következő táblázatban a kapott válaszokat láthatjátok:

Költség kategória (lej)	Válasz gyakoriság
0-50	45
50-100	65
100-200	30
200-300	25
300-500	10
500-1.000	5

a. Határozzuk meg a válaszok alapján a havi átlagos költséget és a szórást. (E: átlag 136.11; szórás 143.55)
b. Határozzuk meg a költés mediánját, az alsó és a felső kvartilist. (E: medián 85; alsó kvartilis 50.38; felső kvartilis 185)
c. Hogyan változna az átlag és medián, ha a válaszok alsó és felső 10%-t is kivennénk a vizsgálatból? (E: átlag 107.99; medián 85)

5. Megkérdeztünk 120 embert, hogy mennyit költenek havonta élelmiszerre. A kérdés válaszlehetőségeként 6 kategóriát adtunk meg, a következő táblázatban a kapott válaszokat láthatjátok:

Költség kategória (lej)	Válasz gyakoriság
0-150	15
150-300	30
300-500	45
500-1.000	20
1.000-2.500	5
2.500-4.000	5

a. Határozzuk meg a válaszok alapján a havi átlagos költséget és a szórást. (E: átlag 548.96; szórás 658.38)
b. Határozzuk meg az árak mediánját, az alsó és a felső kvartilist. (E: medián 368.89; alsó kvartilis 227.5; felső kvartilis 512.5)
c. Hogyan változna az átlag és medián, ha a válaszok alsó és felső 10%-t kivennénk a vizsgálatból? (E: átlag 400.78; medián 368.89)

6. A tavaly végzett egyetemisták körében felmérést végeztek, hogy tanulmányozzák a munkaerőpiaci helyzetüket. Eszerint a megkérdezettek közül 30-an találtak munkát, és az ők nettó fizetéseik eloszlása a következő:

Kategoriák (lej)	Válasz gyakoriság
800-1000	3
1000-1200	6
1200-1400	9
1400-1600	7
1600-1800	5

a. Határozzuk meg a nettó fizetések átlagát és szórását. (E: átlag 1333.3; szórás 242.67)
b. Határozzuk meg a nettó fizetések mediánját. (E: medián 1333.33)
c. Határozzuk meg azt a fizetésintervallumot, amelybe a közepes nettó fizetéseknek a fele tartozik. (E: intervallum [1150;1528.6])

7. Egy felmérésben 450 urat kérdeztek meg, hogy átlagosan hány sört fogyaszt egy világbajnoki mérkőzés megnézése közben (X változó) és mennyi pénzt költenek átlagosan egy átlagos sörözőben történő mérkőzés nézés közben (Y változó). Az alábbi táblázat tartalmazza a megfigyeléseket.

		X		
		2	**3**	**4**
Y	**30**	90	57	32
	40	36	35	40
	50	33	25	25
	60	19	22	36

Például, a táblázat alapján 36 úr mondta azt, hogy 2 sört fogyaszt el és 40 lejt költ átlagosan.

a. Számoljátok ki a perem eloszlásokat. A megkérdezett urak hány százaléka fogyaszt el 3 sört egy mérkőzés alatt? (E: X peremeloszlásai: 178/450; 139/450; 133/450; Y peremoszlásai: 179/450; 111/450; 83/450; 77/450; A megkérdezett urak 30.9%-a fogyaszt el 3 sört egy mérkőzés alatt.)
b. Átlagosan mennyi pénzt költenek a megkérdezett urak egy mérkőzés nézéssel összekötött sörözés közben? (E: Átlagosan 41.29 lejt költenek.)
c. Határozzátok meg a sör fogyasztás mértékének a szórását (X változó)? (E: S(x) = 0.83)

8. Egy felmérésben 360 hölgyet kérdeztek meg, hogy hány magas sarkú (X változó) és hány lapos talpú (Y változó) cipő van a birtokában jelenleg. Az alábbi táblázat tartalmazza a megfigyeléseket.

		X		
		2	**3**	**4**
Y	**5**	83	50	25
	6	29	32	22
	7	25	18	18
	8	11	14	33

Például, a táblázat alapján annak a 18 hölgy mondta azt, hogy 3 magas sarkú és 7 lapos talpú cipője van jelenleg.

a. Számoljátok ki a perem eloszlásokat. A hölgyek hány százalékának van 7 lapostalpú cipője? (E: x(2)=41.1%; x(3)=31.7%; x(4)=27.2%; y(5)=43.9%; y(6)=23.1%; **y(7)=16.9%**; y(8)=16.1%)
b. Átlagosan hány lapostalpú cipője van a megkérdezett hölgyeknek? (E: E(y)=6.05)
c. Határozzátok meg a lapostalpú cipők számának a szórását (Y változó)? (E: S(y)=1.12)

9. A 2011-s romániai népszámlálási adatok alapján, a következő táblázat Hargita, Kovászna és Maros megyékben a lakosság számát mutatja (ezer főben) életkor szerinti megoszlásban.

	14 év alattiak	15 és 34 év közöttiek	35 és 64 év közöttiek	65 év fölöttiek
Hargita megye	53	84	125	48
Kovászna megye	37	56	85	32
Maros megye	93	143	224	90

Forrás: http://www.recensamantromania.ro/

a. Határozzuk meg a megye és az életkor szerinti peremeloszlásokat! Az összlakosság hány százaléka 14 év alatti? (E: HR-0.29; KV-0.2; MS-0.51; 0-14 0.17; 15-34 0.26; 35-64 0.41; 65+ 0.16; A lakosság 17%-a 14 év alatti.)
b. Határozzuk meg a lakosság életkor szerinti feltételes eloszlását Hargita és Maros megyében! Ez alapján független-e a két változó egymástól? (E: HR feltételes eloszlás 0-14 0.17; 15-34 0.27; 35-64 0.4 és 65+ 0.16; MS feltételes eloszlás 0-14 0.17; 15-34 0.26; 35-64 0.41 és 65+ 0.16; A két feltételes eloszlás megközelítőleg egyforma, ezért ennyi információ alapján a két változó nem függ egymástól. Viszont ahhoz, hogy ezt kijelenthessük, meg kellene vizsgáljuk a Kovászna megyei lakosság életkor szerinti eloszlását is.)
c. Határozzuk meg az lakosság életkorának átlagát összesen, és a három megyében külön-külön! (Az életkor határok legyenek 0-14; 14-34; 34-64; 64-90). (Teljes lakosság átlaga 39.65; HR 39.38; KV 39.2; MS 39.98)

10. A következő táblázatban különböző országok mezőgazdasági területeinek az arányát (a teljes területhez mérten %-ban) és a mezőgazdasági termelés GDP-ben való részarányát (szintén %-ban) láthatjátok.

	Afganisztán	Argentína	Franciaország	Ghána	India	Mexikó	Románia	Ruanda	Spanyolország
Mezőgazdasági terület (összes terület %-a)	58.10	54.50	52.50	69.00	60.60	54.90	60.40	74.70	53.90
Mezőgazdaság aránya a GDP-ben (%)	23.90	7.19	1.63	23.10	18.30	3.52	6.13	33.40	2.82

Forrás: gapminder.org

a. Számítsuk ki a terület és a GDP arány közötti kovarianciát és a korrelációt! Milyen összefüggés van a két változó között? A korreláció kiszámításához használjuk fel, hogy a terület szórása 7.1, míg a GDP arány szórása 10.9. (E: kovariancia 66.52; korreláció 0.86; pozitív erősnek mondható összefüggés)
b. Számítsuk ki a regressziós egyenes paramétereit. (A terület legyen a független változó) (E: y = -65.63 + 1.32x; a = -65.63; b = 1.32)
c. Ha Magyarország területének 59%-a mezőgazdasági terület, akkor mit becsültök, mennyi a mezőgazdasági termelés aránya a GDP-ben? (E: 12.22%)

11. A következő táblázatban különböző országok egy főre jutó GDP-jét (ezer dollárban kifejezve) és az országra vonatkozó várható élettartamot (évben kifejezve) láthatjátok.

	Afganisztán	India	Kína	Kongó	Luxemburg	Oroszország	Románia	USA	Vietnám
GDP/fő (dollár)	1	2.5	6.4	0.3	75.9	14	10.8	42.8	2.5
Várható élettartam (év)	58	65	74	48	79	68	73	78	75

Forrás: gapminder.org

a. Számítsuk ki a GDP/fő és a várható élettartam közötti kovarianciát és a korrelációt! Milyen összefüggés van a két változó között? A korreláció kiszámításához használjuk fel, hogy a GDP/fő szórása 25.6, míg a várható élettartam szórása 10.2. (E: kovariancia 138.35; korreláció 0.53; Nem túl erős, pozitív)
b. Számítsuk ki a regressziós egyenes paramétereit. (E: y = 65 + 0.21x; b=0.21; a=65)
c. Ha Magyarországon a GDP/fő értéke 17.8, akkor mit becsültök, mennyi a várható élettartam? (E: y=68.76 év)

12. Egy fagylaltgyáros tanulmányozza a hőmérséklet és az eladott fagylaltmennyiség közötti összefüggést, ezért egy hét minden napján feljegyezte a déli hőmérsékletet és az aznap eladott fagylalt mennyiségét literben. A következő megfigyeléseket kapta:

Hőmérséklet	23	27	29	31	33	25	23
Liter fagylalt	51	67	72	80	101	67	58

a. Számítsuk ki a hőmérséklet és az eladott fagylaltmennyiség közötti kovarianciát és a korrelációt. Milyen összefüggés van a két változó között? A korreláció kiszámításához használjuk fel, hogy a hőmérsékletek szórása 3.6, míg az eladott fagylaltmennyiség szórása 15. (E: kovariancia 51.18; korreláció 0.95, pozitív és erős a kapcsolat)
b. Számítsuk ki a regressziós egyenes paramétereit. (E: y = -36.9 + 3.95*x; b=3.95; a=-36.9)
c. Egy olyan napon, amikor délben 27 fok van, várhatóan hány liter fagylaltot fognak eladni? (E: y=69.73 liter)

13. Az alábbi táblázat egy vállalat 4 különböző termékének értékesítését tartalmazza 2017-ben és 2018-ban. A vállalat vezetője szeretné bizonyítani, hogy a termékeik árszintje az inflációnál kisebb mértékben növekedett, ezért egy kimutatást kért.

	2017		2018	
	Ár (RON)	Mennyiség	Ár (RON)	Mennyiség
DN-1	63	270	55	500
DN-2	40	400	54	45
DN-3	54	180	58	150
DN-4	90	200	90	270

a. Határozzuk meg a 4 termék egyszerű árindexét 2017-ben és 2018-ban. . (E: DN-1 87.3; DN-2 135.0; DN-3 107.41; DN-4 100)
b. Határozzuk meg a Laspreyes-indexet a vizsgált években 2017-et véve alapévnek. (E: 106.85)
c. Határozzuk meg a Paasche-indexet a vizsgált években. Ahhoz, hogy a vezető jobban tudja demonstrálni, hogy a vállalat termékeinek az ár növekedése az infláció alatt marad, melyik index a megfelelőbb? (E: 96.83; a Paasche index)

14. Az alábbi táblázat egy vállalat gyümölcs befőttjeinek értékesítését tartalmazza 2016-ben és 2017-ben. A vállalat vezetője szeretné bizonyítani, hogy a termékeik árszintje a konkurenciánál kisebb mértékben növekedett, ezért egy kimutatást kért.

	2016		2017	
	Ár (RON)	Mennyiség	Ár (RON)	Mennyiség
Alma	12	450	13	425
Körte	15	270	14	330
Szilva	21	180	23	160
Barack	18	270	20	250

a. Határozzuk meg a 4 termék egyszerű árindexét 2016-ben és 2017-ben. (E: alma 108.33; körte 93.33; szilva 109.52; barack 111.11)
b. Határozzuk meg a Laspreyes-indexet a vizsgált években 2016-et véve alapévnek. (E: 105.97)
c. Határozzuk meg a Paasche-indexet a vizsgált években. Ahhoz, hogy a vezető jobban tudja demonstrálni, hogy a vállalat termékeinek az ár növekedése a konkurenciáénál kisebb, melyik index a megfelelőbb? (E: 105.62; a Paasche index)

15. Az alábbi táblázat egy vállalat 4 különböző termékének értékesítését tartalmazza 2014-ben és 2015-ben. A vállalat vezetője szeretné bizonyítani, hogy a termékeik árszintje az inflációnál kisebb mértékben növekedett, ezért egy kimutatást kért.

	2014		2015	
	Ár (RON)	Mennyiség	Ár (RON)	Mennyiség
IB-1	60	300	55	500
IB-2	40	400	50	350
IB-3	55	150	58	150
IB-4	90	200	90	270

a. Határozzuk meg a 4 termék egyszerű árindexét 2014-ben és 2015-ben. (E: IB-1 91.67; IB-2 125.00; IB-3 105.45; IB-4 100)
b. Határozzuk meg a Laspreyes-indexet a vizsgált években 2014-et véve alapévnek. (E: 105.12)
c. Határozzuk meg a Paasche-indexet a vizsgált években. Ahhoz, hogy a vezető jobban tudja demonstrálni, hogy a vállalat termékeinek az ár növekedése az infláció alatt marad, melyik index a megfelelőbb? (E: 98.08; a Paasche index)

II. Valószínűségszámítás

A jegyzet második része a valószínűségszámítás alapfogalmait tárgyalja öt fejezetben. Az 1. fejezet bevezeti a legalapvetőbb fogalmakat, és bemutatja a különböző valószínűségek kiszámításához használható szabályokat, mint a teljes valószínűség törvénye és Bayes törvénye. A 2. fejezet a valószínűségi változókat tárgyalja. Ezen belül bemutatjuk a diszkrét és folytonos valószínűségi változókkal kapcsolatos alapfogalmakat, statisztikai mutatókat és ezek tulajdonságait. A 3. fejezet a legfontosabb diszkrét valószínűségi változókat mutatja be, mint a Bernoulli, binomiális és Poisson. A 4. fejezet a legfontosabb folytonos valószínűségi változókat mutatja be, mint az egyenletes, exponenciális és normálisis eloszlás. Az 5. fejezet két valószínűségi változó egyidejű eloszlását, ezek függetlenségét és összefüggéséhez használt mutatókat tanulmányozza.

II.1 Események és valószínűségek

A véletlen fontos szerepet játszik az életünkben. A véletlen azokra a helyzetekre vonatkozik amikor:

- a helyzetnek több lehetséges kimenetele (eredménye) van,
- a helyzet kimenetele előre nem ismert.

Példák: A következők olyan helyzetek, amelyekre érvényes a fenti két feltétel:

- a lottó nyerőszámai jövő hét végén,
- gurítunk egy kockát – mi lesz felül?
- egy diák statisztika vizsga előtt,
- a euró-lej árfolyam a következő negyedévben,
- a benzin árának változása a következő félévben.

Fontos fogalmak

Egy (véletlenszerű) kísérlet egy olyan helyzet, ahol a véletlen jelen van. Egy kísérlet egy eredménye a kísérlet egy lehetséges kimenetele. A kísérlet összes lehetséges eredményét eseménytérnek nevezzük. Az eredmények egy halmaza eseményt alkot. Az egyedi eredményeket elemi eseményeknek nevezzük.

Tehát az esemény és az eredmény fogalmak nem ugyanazt jelentik, mivel egy esemény lehet több eredmény valamelyike.

Példa. Vegyük a következő kísérletet: gurítunk egy 6 oldalú kockát és megnézzük mi van felül. Az eredmények az 1, 2, 3, 4, 5, 6 számok.

Eseménytér = {1, 2, 3, 4, 5, 6}.

Egy esemény az eredmények bármilyen halmaza, példák:

- "páros szám" = {2, 4, 6},
- "kisebb mint 4" = {1, 2, 3},
- "az 1,2,3,4,5,6 közül bármelyik" = {1, 2, 3, 4, 5, 6},
- "4-es" = {4} (egy elemi esemény).

Az eseménytér bármelyik részhalmaza egy lehetséges esemény. Az az esemény amelyik az egész eseményteret magába foglalja a biztos esemény (mert ez az esemény biztos bekövetkezik). Az az esemény amelyik egyik elemi eseményt sem tartalmazza a lehetetlen esemény. Az A ,,nem következik be" eseményt az A komplementer eseményének nevezzük.

Például,

- ha A = ,,páros szám" = {2, 4, 6} akkor az A komplementer eseménye {1,3,5}, vagyis, ,,páratlan szám",
- ,,az 1,2,3,4,5,6 nem következik be" a lehetetlen esemény.

Azt mondjuk, hogy az ,,**A és B**" esemény bekövetkezett, ha mindkét esemény bekövetkezett. Például, ha A=,,páros szám" és B=,,kisebb mint 4", akkor ,,A és B" = {2}, vagyis, az A és B események akkor következnek be ha a kísérlet során 2-t kapunk.

Az ,,**A vagy B**" esemény akkor következik be, ha valamelyik esemény bekövetkezik, ezért a példában ,,A vagy B" = {1,2,3,4,6}. Két eseményt amelyek nem következhetnek be egyszerre kölcsönösen exkluzív (kizáró) eseményeknek nevezzük. Ebben az esetben ,,A és B" a lehetetlen esemény.

Példa: A = ,,nagyobb mint 4" és B = ,,1 vagy 3" kölcsönösen exkluzív események.

Venn diagram

Halmazokkal végzett műveletekre használjuk. Hasznos eseményekkel végzett műveletekre is. ,,A és B" az A és B metszete, míg ,,A vagy B" az A és B egyesítése.

Gyakoriság és valószínűség

Tegyük fel, hogy egy kockát n-szer gurítunk. Legyen n_i azoknak a gurításoknak a száma amikor i-t gurítunk, i = 1, 2, 3, 4, 5 , 6. Tehát n_i egy abszolút gyakoriság. A relatív gyakoriság $f_i = n_i/n$.

Mi történik, ha n $\rightarrow\infty$? Azt figyelhetjük meg, hogy f_i konvergál egy számhoz. Mindegyik f_i relatív gyakoriság az 1/6-hoz konvergál ha n $\rightarrow \infty$. Ez az eredménynek a valószínűsége. A valószínűség egy 0 és 1 közötti szám, mert a relatív gyakoriságok is 0 és 1 között vannak)

Egy eredmény valószínűsége az eredmény relatív gyakoriságának az értéke amikor a kísérletet végtelen sokszor ismételjük meg.

Formálisan, egy x eredményre az x valószínűsége a $\lim_{n\rightarrow\infty} n_x / n$

határértékkel egyenlő, ahol n_x az a szám ahányszor *x*-et kapunk az *n* kísérlet során.

Példa: Egy pénzt dobunk fel sokszor. Az eredmények fej (F) és írás (Í). Az első 30 dobás:

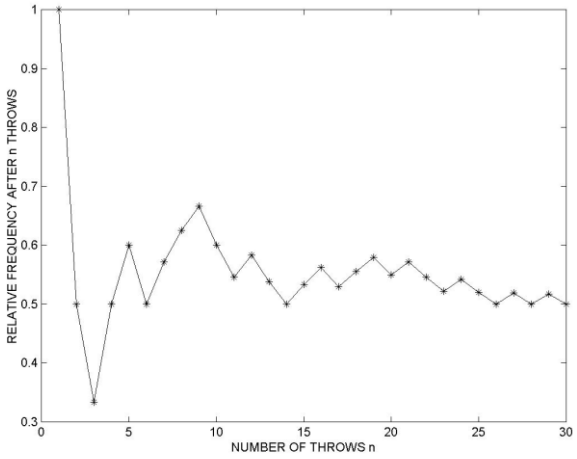
FÍÍFF ÍFFFÍ ÍFÍÍF FÍFFÍ FÍÍFÍ ÍFÍFÍ

Az F relatív gyakoriságai mindegyik dobás után:

1, 1/2, 1/3, 2/4, 3/5, 3/6, 4/7, ..., 14/28, 15/29, 15/30

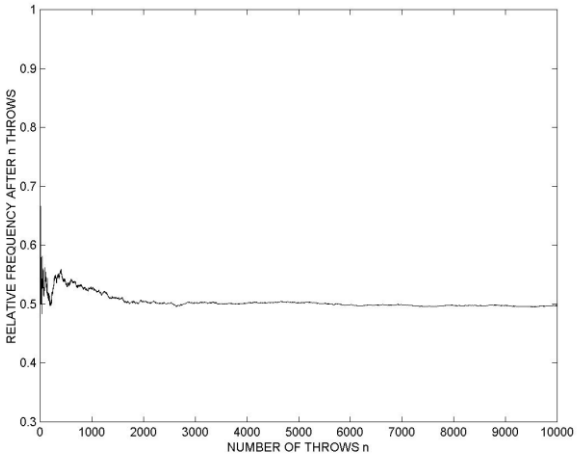
A következő vonalgrafikon a relatív gyakoriságokat mutatja minden dobás után a dobások számának függvényében:

Ábra 2.1.1. – A relatív gyakoriságok változása a dobások számának függvényében (1)



Most nézzük meg mi történik, ha még többször dobjuk fel a pénzt. A következő vonalgrafikon a relatív gyakoriságokat mutatja a dobások számának függvényében, amikor 10000-szer dobjuk fel a pénzt:

Ábra 2.1.2. – A relatív gyakoriságok változása a dobások számának függvényében (2)



Persze, ha újra elvégezzük a kísérletet nem pontosan ezt a vonalgrafikont kapjuk, viszont a relatív gyakoriságok minden esetben ½-hez konvergálnak. Tehát a ,,fej" valószínűsége ½, amit úgy jelölünk, hogy **P(F) = ½**.

A valószínűség tulajdonságai

Bármelyik A eseményre $\mathbf{0 \leq P(A) \leq 1}$.

A biztos esemény valószínűsége **1**, mert a relatív gyakorisága mindig 1 lesz.

Ha A és B kölcsönösen kizáró (exkluzív) események, akkor **P(A vagy B) = P(A) + P(B)**, mert az ,,A vagy B" kölcsönösen kizáró (exkluzív) esemény relatív gyakorisága az A és B relatív gyakoriságainak az összege.

Példa: Az A=,,1 vagy 3" esemény valószínűsége P(1 vagy 3) = P(1)+P(3) = 1/6+1/6 = 1/3. A B=,,páros szám" esemény valószínűsége P(A) = P(2 vagy 4 vagy 6) = 1/6+1/6+1/6 = ½.

Emlékezzünk vissza, hogy a relatív gyakoriságok összege mindig 1. Ezért, **az összes eredmény valószínűségének összege is 1.** Tehát ha egy kísérlet eredményei x_1, x_2,..., x_n akkor

$$\sum_{i=1}^{n} P(\{x_i\}) = 1.$$

Az A és „nem A” (az A komplementer eseménye) mindig **kölcsönösen exkluzívak**.

Továbbá, az „A vagy nem A” esemény a biztos esemény, tehát a valószínűsége 1:

$$P(A) + P(\text{nem } A) = 1.$$

Ezért bármelyik A eseményre:

$$\mathbf{P(\text{nem } A) = 1 - P(A).}$$

„A vagy B” valószínűsége amikor nem kölcsönösen exkluzívak

Bármely A és B eseményekre (a Venn diagram alapján):

- „A vagy B” = „A és nem B” vagy B, tehát

$$P(A \text{ vagy } B) = P(\text{„A és nem B” vagy } B) \quad (1)$$

- „A és nem B” és B kölcsönösen exkluzívak, tehát

$$P(\text{„A és nem B” vagy } B) = P(A \text{ és nem } B) + P(B) \quad (2)$$

- A = „A és nem B” vagy „A és B”, melyek kölcsönösen exkluzívak, tehát

$$P(A) = P(A \text{ és nem } B) + P(A \text{ és } B) \quad (3)$$

Az (1), (2) és (3) alapján:

$$\mathbf{P(A \text{ vagy } B) = P(A) + P(B) - P(A \text{ és } B).}$$

Egyenlő valószínűségű eseményterek által értelmezett valószínűség

Amikor egy pénzt dobunk fel vagy egy kockát gurítunk, (vagy lottózunk), általában mindegyik lehetséges eredmény egyenlő valószínűséggel következik be. Egy ilyen kísérlet eseményterét egyenlő valószínűségű eseménytérnek nevezzük. Mindegyik eredmény valószínűsége ugyanaz, és mivel a valószínűségek összege 1, ezért

$$\mathbf{P(\text{eredmény}) = 1 / \text{eredmények száma.}}$$

Egy esemény valószínűsége

$$P(\text{esemény}) = \frac{\text{az eseményben szereplő eredmények száma}}{\text{eredmények száma}}.$$

Ez volt a valószínűség első értelmezése (Laplace által a 18. században). **Vigyázat: csak** egyenlő valószínűségű eseményterekben használható.

Példa: Dobjunk fel egy pénzt 3-szor. Ebben az esetben az eseménytér = {FFF,FFÍ,FÍF,ÍFF,FÍÍ,ÍFÍ,ÍÍF,ÍÍÍ} míg az eredmények száma = 8.

A = “2 vagy több fej” = {FFF,FFÍ,FÍF,ÍFF}, B = “3 írás” = {ÍÍÍ}, az A és B kölcsönösen exkluzívak események.

Legyen C = {FFF,FÍF,ÍÍF,FFÍ}, D = {FÍF,ÍÍÍ,ÍFÍ}. A C és D események nem kölcsönösen exkluzívak. Meghatározzuk a következő eseményeket:

„nem C” = {ÍFF,FÍÍ,ÍFÍ,ÍÍÍ}, „C és D” = {FÍF}, „C vagy D” = {FFF,FÍF,ÍÍF,FFÍ,ÍÍÍ,ÍFÍ}.

Meghatározzuk a következő valószínűségeket: P(2 fej) = 3/8, P(nem 2 fej) = 5/8, P(a 3. dobás fej) = 4/8 = 1/2, P(2 fej és a 3. dobás fej) = 2/8 = 1/4, P(2 fej vagy a 3. dobás fej) = 5/8.

Feltételes valószínűség

Képzeljük el, hogy gurítunk egy kockát. Tudjuk, hogy P(6-os) = 1/6. Ha most gurítok egy kockát, és azt mondom, hogy „páros számot gurítottam”, mi lesz annak a valószínűsége, hogy 6-os? Az információ amit megadtam megváltoztatja a valószínűséget.

Ez a feltételes valószínűség alapötlete: egy olyan valószínűség amelyre tudjuk, hogy egy másik esemény bekövetkezett. A feltételes valószínűség alapötlete kapcsolódik a feltételes eloszláshoz.

Az A esemény **feltételes valószínűségét**, ha tudjuk hogy a B esemény bekövetkezett **P(A | B)**-vel jelöljük. A P(A | B) feltételes valószínűséget a következő képlettel számítjuk ki:

$$P(A|B) = \frac{P(A \text{ és } B)}{P(B)}.$$

Ezt a képletet gyakran írják szorzat formájában, amit a **valószínűség szorzási szabály**ának neveznek:

$$P(A \text{ és } B) = P(A|B)P(B) = P(B|A)P(A).$$

Példa: Vegyünk két dobozt. Az elsőben 3 fehér és 2 fekete, a másodikban 1 fehér és 6 fekete golyó van. Végezzük a következő kísérletet: egy pénzt feldobunk, ha fej akkor az első dobozból, másképp a másodikból húzunk egy golyót. Ekkor kiszámíthatjuk a következő valószínűségeket:

- P(fehér | fej) = 3/5, mivel „fej” esetén tudjuk, hogy az első dobozból húzunk, ahol 5 golyóból 3 fehér.
- P(fehér| írás) = 1/7, mivel „írás” esetén a második dobozból húzunk, ahol 7 golyóból 1 fehér.
- P(fej és fehér) = P(fehér | fej) × P(fej) = 3/5 × ½ = 3/10.

A teljes valószínűség törvénye

Legyenek B_1, B_2, ..., B_n események amelyek bekövetkezhetnek egy kísérlet során úgy, hogy pontosan egy következik be közülük. Másszóval, páronként kölcsönösen exkluzívak (kizáró események) és a „B_1 vagy B_2 vagy ... B_n” esemény a biztos esemény.

Példa: egy kockát gurítunk, B_1 = {2,4,6} („páros”) és B_2 = {1,3,5} („páratlan”).

Ki szeretnénk számítani egy A esemény valószínűségét amikor ismerjük mindegyik B_i valószínűségét és a P(A | B_i) feltételes valószínűségeket. Először vegyük észre, hogy

$$P(A) = P(A \text{ és } B_1) + P(A \text{ és } B_2) + \ldots + P(A \text{ és } B_n).$$

Továbbá, a szorzás szabály alapján minden i-re:

$$P(A \text{ és } B_i) = P(A \mid B_i)\, P(B_i).$$

Tehát, a **teljes valószínűség törvénye**:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\ P(B_i).$$

Példa. A két golyós doboz esetén mennyi P(fehér), vagyis annak a valószínűsége, hogy fehér golyót húzunk?

Kétféleképpen fordulhat ez elő:

- fejet dobunk és fehéret húzunk az első dobozból vagy
- írást dobunk és fehéret húzunk a második dobozból.

Tehát:

P(fehér) = P("fej és fehér az első dobozból" vagy "írás és fehér a második dobozból").
Mivel a valószínűségben szereplő két esemény kölcsönösen exkluzív:

= P(fej és fehér az első dobozból) +P(írás és fehér a második dobozból).

A szorzás szabálya alapján:

= P(fej) P(fehér |fej) + P(írás) P(fehér|írás).

Ezeket a valószínűségeket már kiszámítottuk, ezért: P(fehér) = ½ × 3/5 + ½ × 1/7 = 13/35.

Bayes törvénye

A két golyós doboz példában tegyük fel a következő kérdéseket:

- Ha tudjuk, hogy fehér golyót húztak valamelyik dobozból, mi a valószínűsége annak, hogy fejet dobtak, vagyis, hogy az első dobozból húzták a fehér golyót?
- Vagyis, mennyi a következő valószínűség: P(fej| fehér).

A szorzás szabályát használva:

$$P(Bi \mid A) = P(Bi \text{ és } A) / P(A) = P(A \mid Bi)\, P(Bi) / P(A).$$

A nevezőben szereplő P(A) valószínűséget kiszámíthatjuk a teljes valószínűség törvényéből, ha szükséges. A így kapott képlet Bayes törvénye:

Egy A eseményre és B_1, B_2, ..., B_n eseményekre (mint ezelőtt):

$$P(B_j|A) = \frac{P(A|B_j)\ P(B_j)}{P(A)} = \frac{P(A|B_j)\ P(B_j)}{\sum_{i=1}^{n} P(A|B_i)\ P(B_i)}.$$

A fenti példában P(fej| fehér) = P(fehér |fej) × P(fej) / P(fehér) = 3/5 × ½ / 13/35 = 21/26.

Független események

Két A és B esemény független ha: P(A | B) = P(A) és P(B |A) = P(B). Ha A és B független események, akkor, ha közülük egyik bekövetkezik, a másik esemény valószínűsége nem változik.

Az események függetlensége hasonló a változók függetlenségéhez. Az események függetlenségét gyakran használjuk. Például, sok esetben egy kísérlet megismétlése során az események függetlenek (ha kétszer dobjuk fel a pénzt a két dobás független egymástól).

II.2 Valószínűségi változók

Egy valószínűségi változó (rövídítve: valószínűségi változó) egy olyan változó amely olyan mennyiségi (kvantitatív) értékeket vesz fel, amelyek egy véletlenszerű kísérlet eredménye.

Nagy betűkkel, például, X, Y, stb. fogjuk jelölni a valószínűségi változókat. Kis betűkkel, pl. x, y, stb. fogjuk jelölni a valószínűségi változók megvalósult értékeit.

Példa: X = kocka gurítás utáni értéke. Az eredmény egy szám 1-től 6-ig. Ez egy valószínűségi változó.

Példa: Y = feldobunk egy pénzérmét. Az eredmény Fej vagy Írás. Ez nem valószínűségi változó, mert az értékei nem kvantitatívak.

Az első részben a kvantitatív változóknak két típusát különböztettük meg, diszkrétet és folytonost. Ezeknek külön tárgyaltuk a leíró statisztikáját a gyakoriságok alapján. Ehhez hasonlóan, a következőkben külön tárgyaljuk a diszkrét és folytonos valószínűségi változókat.

Diszkrét valószínűségi változók

Ugyanolyan értékeket vesznek fel mint a diszkrét változók, vagyis egy számlálási folyamat eredményeit: 0, 1, 2, ... Mindegyik x_i lehetséges értéknek van valószínűsége.

Jelölés: $P(X = x_i) = p(x_i) = p_i$.

A p_i valószínűségek teljesítik a

- $0 \leq p_i \leq 1$, vagyis, a valószínűségek 0 és 1 közötti számok,
- $\sum_i p_i = 1$, vagyis, a valószínűségek összege 1.

tulajdonságokat.

Tömegfüggvény

A $p(x_i)$ valószínűségeket mindegyik x_i-re az **X tömegfüggvény**ének nevezzük. Ezeket a valószínűségeket oszlopdiagrammal ábrázolhatjuk. Az oszlopdiagramból majd láthatjuk, hogy a tömegfüggvény úgy néz ki mint a relatív gyakoriságok oszlopdiagramja. Ugyanakkor a tömegfüggvény ugyanazokkal a tulajdonságokkal rendelkezik mint a relatív gyakoriságok. **A tömegfüggvény pontosan leírja egy diszkrét valószínűségi változó valószínűségeit.**

Példa: Gurítsunk 2 kockát, és legyen X a két gurított szám összege. Ekkor X egy diszkrét valószínűségi változó. Az X lehetséges értékei: 2, 3, 4, …, 12.

Ahhoz, hogy meghatározzuk ezek valószínűségeit, vegyük észre, hogy 36 lehetséges eredmény van 2 kocka gurításakor. Az összeg mindegyik lehetséges értékére a következő táblázat tartalmazza a 2 kocka gurításának lehetséges eredményeit.

x	A 2 kocka eredménye	*x*	A 2 kocka eredménye
2	(1,1)	**8**	(2,6), (3,5), (4,4), (5,3), (2,6)
3	(1,2), (2,1)	**9**	(3,6), (4,5), (5,4), (6,3)
4	(1,3), (2,2), (3,1)	**10**	(4,6), (5,5), (6,4)
5	(1,4), (2,3), (3,2), (4,1)	**11**	(5,6), (6,5)
6	(1,5), (2,4), (3,3), (4,2), (5,1)	**12**	(6,6)
7	(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)		

Mindegyik eredménynek a 36-ból ugyanaz a valószínűsége.

Tehát P(X = x) = [azoknak az eredményeknek a száma amelyre X = x] / 36.

Ezért a tömegfüggvény:

P(X = 2) = 1/36 P(X = 8) = 5/36

P(X = 3) = 2/36 P(X = 9) = 4/36

P(X = 4) = 3/36 P(X = 10) = 3/36

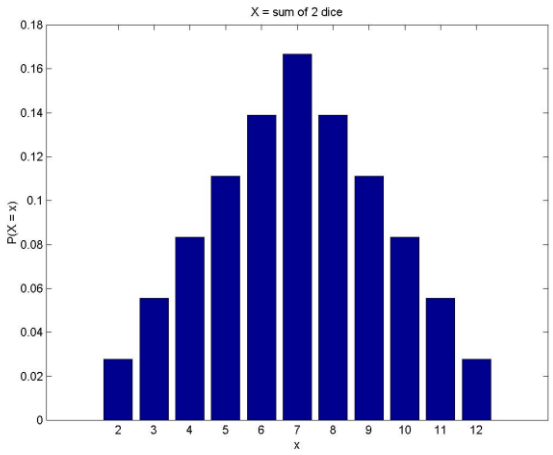
P(X = 5) = 4/36 P(X = 11) = 2/36

P(X = 6) = 5/36 P(X = 12) = 1/36

P(X = 7) = 6/36

Ellenőrizhetjük, hogy a valószínűségek összege 1.

Ábra 2.2.1. - A tömegfüggvény oszlopdiagramja



Példa: Egy pénzt dobunk fel 3-szor. X az a szám ahányszor fejet dobunk a 3 dobás során.

- X lehetséges értékei: 0, 1, 2, 3.
- A 3 dobás lehetséges eredményei: FFF,FFÍ,FÍF,ÍFF,FÍÍ,ÍFÍ,ÍÍF,ÍÍÍ.
- X tömegfüggvénye: P(X = 0) = 1/8, P(X = 1) = 3/8, P(X = 2) = 3/8, P(X = 3) = 1/8.

Gyakorlat. Ábrázoljuk a tömegfüggvényt oszlopdiagrammal.

A tömegfüggvény és a gyakoriság-eloszlás közötti összefüggés

Amikor olyan megfigyeléseink vannak, amelyek egy kísérlet eredményei, összefüggés van a tömegfüggvény és a gyakoriság-eloszlás között. Tegyük fel, hogy lejegyezzük 2 kocka gurításakor az összegüket: $x_1, x_2, ..., x_n$.

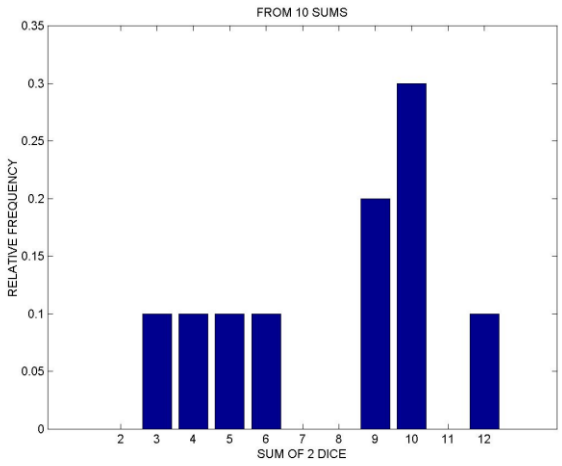
Ekkor mindegyik x_i egy 2 és 12 közötti szám. Legyen n_i azoknak a megfigyeléseknek a száma ahol az összeg i, i = 2,...,12 (vagyis, n_i az i abszolút gyakorisága).

A gyakoriság-eloszlás n_i / n, ahol i = 2,...,12. Az előző fejezetben az i valószínűségét a $P(X = i) = \lim_{n\to\infty} n_i / n$ képlettel értelmeztük. Tehát a p(i) = P(X = i), i=2,3,…,12 tömegfüggvény az i relatív gyakorisága amikor *n* nagyon nagy ($\to \infty$). Ez azt jelenti, hogy a relatív gyakoriságok oszlopdiagramja a tömegfüggvény oszlopdiagramjához konvergál, ha $n \to \infty$.

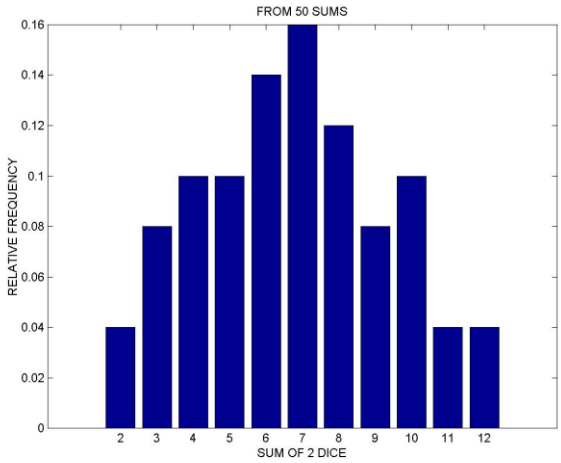
A gyakorlatban az történik, hogy a relatív gyakoriságok és a tömegfüggvény oszlopdiagramja hasonlít, ha *n* nem túl nagy akkor is. Ha egyre több megfigyelést gyűjtünk (vagyis *n* egyre nő), a két oszlopdiagram egyre jobban hasonlít egymásra.

Ezt láthatjuk a következőkben n = 10, 50, 200 és 1000-re

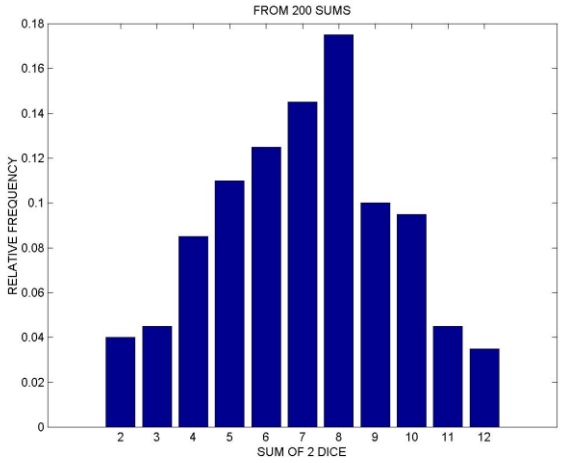
Ábra 2.2.2. - Példa: A tömegfüggvény oszlopdiagramja n=10 esetén



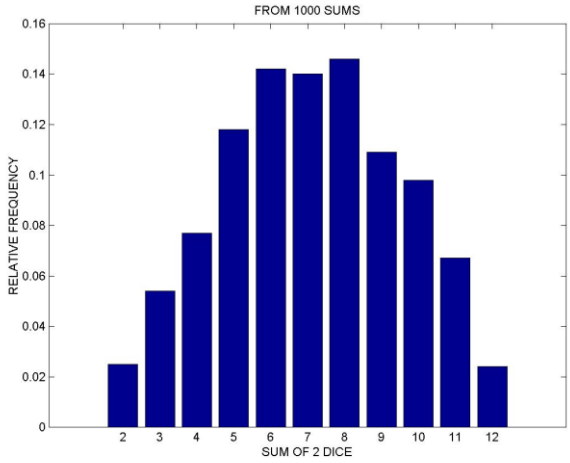
Ábra 2.2.3. - Példa: A tömegfüggvény oszlopdiagramja n=50 esetén



Ábra 2.2.4. - Példa: A tömegfüggvény oszlopdiagramja n=200 esetén



Ábra 2.2.5. - Példa: A tömegfüggvény oszlopdiagramja n=1000 esetén



Folytonos valószínűségi változók

Egy folytonos valószínűségi változó egy intervallumban veszi fel az értékeit.

Példák:

- X = amennyi idő alatt az előadóterembe érünk otthonról,
- Y = az utcán véletlenszerűen kiválasztott személy magassága.

A sűrűségfüggvény

Egy folytonos valószínűségi változó esetén nem lehet a tömegfüggvényt értelmezni mert a valószínűségi változó egy intervallumon belül minden értéket felvehet. Ez hasonló ahhoz, hogy egy folytonos változó minden értéket általában csak egyszer vesz fel, ezért a gyakoriság-eloszlásnak nincs értelme úgy mint diszkrét változó esetén. Folytonos változóknál annak van értelme, hogy a gyakoriság-eloszlást a terjedelem kisebb intervallumokra való felosztásával határozzuk meg.

Folytonos valószínűségi változók valószínűségeit a sűrűségfüggvény határozza meg, amit f(x)-szel jelölünk. Ez egy függvény amelyre annak a valószínűsége, hogy X a és b között van (vagyis P($a < X < b$)) az f(x) grafikonja alatti terület a és b között, vagyis:

$$P(a < X < b) = \int_a^b f(x)\ dx.$$

Később meglátjuk, hogy hogyan kapcsolódnak az így értelmezett valószínűségek az X folytonos változó megfigyeléseinek gyakoriság-eloszlásához.

A sűrűségfüggvény tulajdonságai

Egy sűrűségfüggvénynek a következő két fontos tulajdonsága van.

1. Tulajdonság: Egy folytonos valószínűségi változó biztosan -∞ és +∞ közötti értéket vesz fel, ezért ennek az eseménynek a valószínűsége 1, vagyis a sűrűségfüggvény grafikonja alatti terület 1-gyel egyenlő:

$$\int_{-\infty}^{\infty} f(x)\ dx = P(-\infty < X < \infty) = 1.$$

2. Tulajdonság: f(x) ≥ 0 minden x-re.

Ha ez nem így lenne, akkor előfordulhatna, hogy a sűrűségfüggvény grafikonja alatti terület egy bizonyos intervallumban negatív, aminek nincs értelme, mert ez azt jelentené, hogy az ennek megfelelő esemény valószínűsége negatív.

Példa sűrűségfüggvényre

Az alábbi sűrűségfüggvény egy pozitív folytonos valószínűségi változó sűrűségfüggvénye:

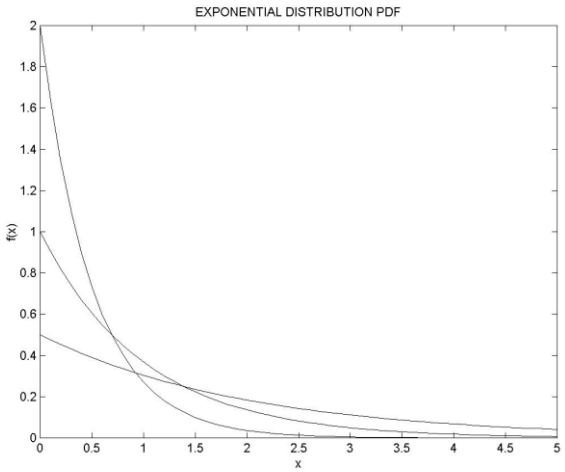
$$f(x) = \lambda e^{-\lambda x},\ x \geq 0,$$

ahol λ > 0 egy szám amit mi adunk meg. f(x) teljesíti a fenti két tulajdonságot.

Példa: ha λ = 0.5,

- P(1 < X < 2.5) = $\int_1^{2.5} 0.5e^{-0.5x}dx = -e^{-0.5x}|_1^{2.5} = e^{-0.5} - e^{-1.25} = 0.32.$
- P(X > 3) = $\int_3^{\infty} 0.5e^{-0.5x}dx = -e^{-0.5x}|_3^{\infty} = e^{-1.5} - 0 = 0.223.$

Ábra 2.2.6. - Különböző λ értékekre az f(x) grafikonja



<u>Folytonos változó sűrűségfüggvénye és gyakoriság-eloszlása közötti összefüggés</u>

Amikor a megfigyeléseink olyan kísérletből származnak amely folytonos valószínűségi változónak felel meg, összefüggés van a folytonos valószínűségi változó sűrűségfüggvénye és a megfigyelések gyakoriság-eloszlása között. Emlékezzünk vissza, hogy egy folytonos változó gyakoriság-eloszlásának a meghatározásához felosztottuk a megfigyelések terjedelmét kisebb intervallumokra.

Tegyük fel, hogy *n* megfigyelésünk van és n_i = az i-edik intervallumban szereplő megfigyelések száma. Ekkor a relatív gyakoriság n_i / n. A valószínűség értelmezése szerint:

$$P(\text{X az i-edik intervallumban van}) = \lim_{n \to \infty} n_i / n.$$

Valamint:

$$P(X \text{ az } i-edik \text{ intervallumban van}) = \int_{\text{i-edik intervallum}} f(x)\ dx.$$

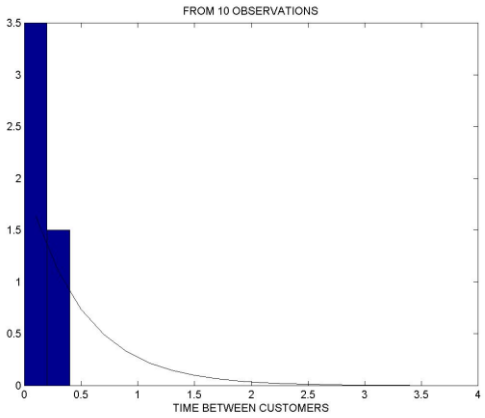
Tehát, az intervallum relatív gyakorisága konvergál a sűrűségfüggvény grafikonja alatti területhez az adott intervallumban, ha $n \to \infty$. Ha megszerkesztjük a relatív gyakoriságok hisztogramját, az oszlop területe az intervallum relatív gyakorisága. Tehát, ahogy a megfigyelések száma nő, a hisztogram oszlopainak a területe konvergál a sűrűségfüggvény grafikonja alatti területhez a megfelelő intervallumban.

Példa: tegyük fel, hogy egy bankban mérjük két egymás után érkező ügyfél megjelenése között eltelt időt; a megfigyelések x_1, x_2, ..., x_n. Tegyük fel, hogy a megfigyeléseknek megfelelő valószínűségi változó sűrűségfüggvénye

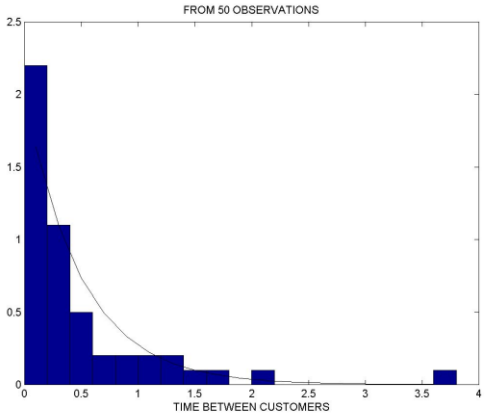
$$f(x) = 2e^{-2x},\ x \geq 0.$$

A következő oldalakon láthatjuk a megfigyelések hisztogramjait és sűrűségfüggvényét különböző n értékekre.

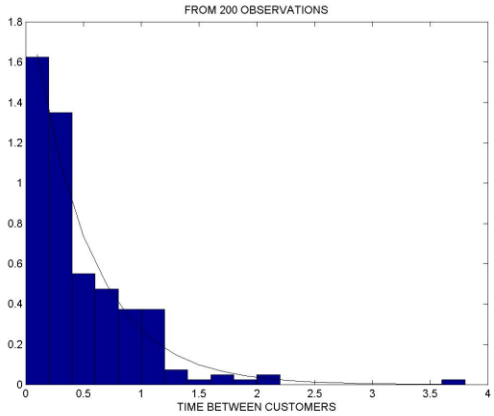
Ábra 2.2.7. - Hisztogram 10 megfigyelésre



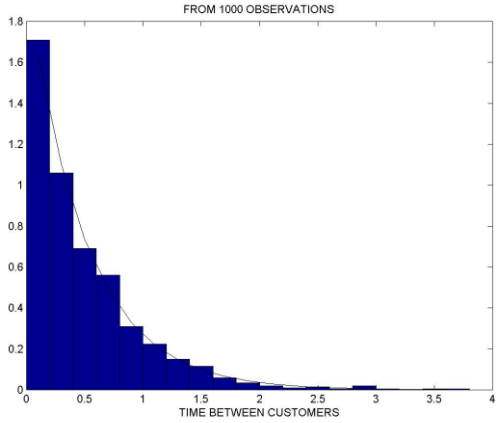
Ábra 2.2.8. - Hisztogram 50 megfigyelésre



Ábra 2.2.9. - Hisztogram 200 megfigyelésre



Ábra 2.2.10 - Hisztogram 1000 megfigyelésre



<u>Egy valószínűségi változó várható értéke</u>

Tegyük fel, hogy az $x_1,...,x_n$ egy diszkrét változó megfigyelései. Ekkor a megfigyelések átlaga:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Legyen n_x az x érték gyakorisága. Ekkor az x érték n_x-szer szerepel az átlag képletének az összegében. Ezért az átlag-képletet másképp is írhatjuk:

$$\bar{x} = \frac{1}{n}\sum_{x} x\ n_x = \sum_{x} x\ \frac{n_x}{n} = \sum_{x} x \times (x \text{ relatív gyakorisága}).$$

Példa: dobjunk fel egy pénzt. Legyen az X valószínűségi változó 0 ha írás és 1 ha fej. Tegyük fel, hogy n=10 dobás után a következő értékeket kapjuk:

$x_1=1$, $x_2=0$, $x_3=0$, $x_4=1$, $x_5=1$, $x_6=1$, $x_7=0$, $x_8=1$, $x_9=0$, $x_{10}=1$.

0 gyakorisága $n_0=4$ és 1 gyakorisága $n_1=6$

A megfigyelések átlaga

$$\bar{x} = \frac{1}{10}(1+0+0+1+1+1+0+1+0+1)$$
$$= \frac{1}{n}(0n_0 + 1n_1) = 0\frac{n_0}{n} + 1\frac{n_1}{n}.$$

Most tegyük fel, hogy a megfigyelések egy kísérletből származnak, tehát mindegyik megfigyelés egy X valószínűségi változó megvalósult értéke és a tömegfüggvényük ugyanaz a p(x). Tegyük fel, hogy $n \to \infty$; ezért minden x érték relatív gyakorisága konvergál a p(x)-hez.

Tehát

$$\bar{x} \to \sum_{x} x\ p(x).$$

A kapott képlet adja az X <u>diszkrét</u> valószínűségi változó <u>várható értékét</u>:

$$E(X) = \sum_x x\ p(x),$$

ahol az összeg az x összes lehetséges értékére vonatkozik. Az E jelölés az angol ,,expected value"-ból származik.

Folytonos valószínűségi változók esetén:

- a p(x) tömegfüggvényt az f(x) sűrűségfüggvénnyel helyettesítjük,
- az összeget integrállal helyettesítjük.

Egy X folytonos valószínűségi változó várható értéke

$$E(X) = \int_x xf(x)\ dx.$$

Példa: A fejek számának várható értéke, ha 3-szor dobunk fel egy pénzt:

$$0 \times 1/8 + 1 \times 3/8 + 2 \times 3/8 + 3 \times 1/8 = 1.5,$$

ahol felhasználtuk a korábban kiszámított tömegfüggvényt.

Az

$$f(x) = 1,\ 0 \le x \le 1.$$

sűrűségfüggvénnyel rendelkező valószínűségi változó (tulajdonképpen egyenletes eloszlású, lás alább) várható értéke:

$$\int_0^1 xf(x)\,dx = \int_0^1 x\,dx = \frac{x^2}{2}|_0^1 = 0.5.$$

Egy valószínűségi változó varianciája

Az $x_1,...,x_n$ megfigyelésekre a variancia

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

ami az $(x_i - \bar{x})^2$ (átlagtól való eltérések négyzetének) átlaga. Ezért megismételhetjük az eljárást, amit a várható értékre alkalmaztunk, és úgy értelmezzük egy valószínűségi változó varianciáját mint az s^2, amikor $n \to \infty$.

Példa: pénzfeldobás. Az $x_1,...,x_n$ megfigyelésekre a variancia

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}[(1-\bar{x})^2 + (0-\bar{x})^2 + (0-\bar{x})^2 + (1-\bar{x})^2 + (1-\bar{x})^2 \\ &\quad + (1-\bar{x})^2 + (0-\bar{x})^2 + (1-\bar{x})^2 + (0-\bar{x})^2 + (1-\bar{x})^2] \\ &= (0-\bar{x})^2\frac{n_0}{n} + (1-\bar{x})^2\frac{n_1}{n}. \end{aligned}$$

Ha n $\to \infty$ az átlag$\to$E(X) és az x relatív gyakorisága$\to$p(x).

Ezért egy X diszkrét valószínűségi változó varianciája

$$Var(X) = \sum_x (x - E(X))^2 \, p(x) = E[(X - E(X))^2].$$

Egy X folytonos valószínűségi változó varianciája

$$Var(X) = \int_x (x - E(X))^2 \; f(x) \; dx = E[(X - E(X))^2].$$

Példák. A fejek számának varianciája, ha 3-szor dobunk fel egy pénzt, felhasználva a már meghatározott tömegfüggvényt és várható értéket:

$$(0 - 1.5)^2 \times 1/8 + (1 - 1.5)^2 \times 3/8 + (2 - 1.5)^2 \times 3/8 + (3 - 1.5)^2 \times 1/8 = 0.75.$$

Az *egyenletes eloszlás* varianciája:

$$\int_0^1 (x - 0.5)^2 dx = \frac{(x - 0.5)^3}{3} |_0^1 = 1/12.$$

Egy valószínűségi változó mediánja

Egy X folytonos valószínűségi változóra a medián az az *m* érték, amelyre

$$P(X \leq m) = \int_{-\infty}^{m} f(s) \; ds = 0.5.$$

Hasonló a folytonos változók mediánjának kiszámításához, amikor kategorizált megfigyeléseink vannak, ugyanis mindkét esetben a medián felezi a grafikon (sűrűségfüggvény vagy hisztogram) alatti területet

Példa: Az egyenletes eloszlás mediánja 0.5, mert a sűrűségfüggvény alatti területet pontosan 0.5-nél felezi.

Lineáris transzformáció várható értéke

Legyen Y = aX + b az X egy lineáris transzformációja, ahol a és b konstansok (nem valószínűségi változók).

Ha a és b konstansok akkor E(aX+b) = a E(X) + b.

Ha c konstans akkor E(c) = c.

Két valószínűségi változó összegének várható értéke

Legyen X várható értéke E(X), és Y várható értéke E(Y). Ekkor:

E(X+Y) = E(X) + E(Y).

Példa: Két kocka összegének a várható értéke, ha a kockákat külön valószínűségi változóknak tekintjük várható értékek összege. Egy kocka várható értéke

Lineáris transzformáció varianciája

Legyen X varianciája Var(X) és Y = aX + b, ahol a és b konstansok.

Tehát a és b konstansokra $Var(aX+b) = a^2 Var(X)$.

Ha c konstans akkor Var(c) = 0.

II.3 Fontosabb diszkrét valószínűségi változók

Amint tárgyaltuk, a diszkrét valószínűségi változók egy számlálási folyamat eredményei, tehát az értékeik 0,1,2, … . A tömegfüggvény mindegyik lehetséges i értéknek megadja a valószínűségét: $P(X = i) = p(i) = p_i$. Ebben a fejezetben a legfontosabb diszkrét valószínűségi változókat tárgyaljuk. Ezek a valószínűségi változók gyakorlati szempontból fontos helyzeteket **modelleznek**.

Modellek és paraméterek

A valószínűségszámításban a valószínűségi változókat modelleknek is nevezik. Egy modell egy bizonyos helyzetet leíró változók közötti összefüggés, ami ebben az esetben egyszerűen egy valószínűségi változó. Általában egy modellnek vannak olyan változói is amiket mi adunk meg. Ezeket paramétereknek nevezzük.

Bernoulli eloszlás (vagy modell vagy valószínűségi változó)

Ez a legegyszerűbb valószínűségi változó. Egy Bernoulli eloszlású X valószínűségi változó p valószínűséggel 1 és 1-p valószínűséggel 0 értéket vesz fel:

$$P(X = 1) = p,$$

$$P(X = 0) = 1 – p.$$

Egy olyan kísérletet amelyik során egy Bernoulli valószínűségi változót kapunk Bernoulli kísérletnek nevezzük.

p-t a modell **paraméterének** nevezik.

Általában a valószínűségi változó 1 értékét „siker”-nek míg a 0 értékét „kudarc”-nak nevezik. A p valószínűséget siker-valószínűségnek nevezik.

A Bernoulli eloszlást számos olyan helyzet szemléltesésére használják, ahol 2 lehetséges eredmény van. Példák:

- pénzfeldobás (Fej vagy Írás),
- vizsga (sikerül vagy nem),
- elnökválasztás két jelölt esetén (egyik jelölt vagy másik jelölt).

Ezekben az esetekben az egyik eredményt 1-nek míg a másikat 0-nak vesszük, hogy jól értelmezett valószínűségi változót kapjunk. Egy Bernoulli valószínűségi változó várható értéke:

$E(X) = 0 \text{ x}(1-p)+1 \text{ x } p = p.$

Egy Bernoulli valószínűségi változó variancája:

$Var(X) = (0-p)^2 \text{ x } (1-p)+(1-p)^2 \text{ x } p = p(1-p).$

Binomiális eloszlás

Vegyünk egy Bernoulli kísérletet melynek sikervalószínűsége p, és ismételjük meg n-szer. Tegyük fel, hogy mindegyik kísérlet független a többitől (vagyis egyik kísérlet eredménye sem függ az előző eredményektől).

Legyen X a "sikerek" száma az n kísérlet során (vagyis az 1-ek száma). Ekkor X lehetséges értékei 0, 1,..., n. Ekkor azt mondjuk, hogy X binomiális eloszlású.

A tömegfüggvénye:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\dots,n,$$

ahol $\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k! = 1 \cdot 2 \cdot \ldots \cdot k.$

A **paraméterek** *n* és *p*. A binomképlet segítségével kimutatható, hogy

$$\sum_{k=0}^{n} P(X = k) = 1.$$

A binomiális eloszlás sok helyzet leírására alkalmas:

- Dobjunk fel egy pénzt *10*-szer. Egyszeri dobáskor a Fej valószínűsége 0.5. Várhatóan hányszor dobunk Fejet?
- Egy számítógégyártó cég 1000 képernyőt gyárt hetente. Annak a valószínűsége, hogy egy képernyő hibás 0.025. Várhatóan hány hibás képernyőt gyártanak hetente?

Példa: tegyük fel, hogy egy pénzérmére P(Fej) = 0.4, és 5-ször dobjuk fel. Ekkor:

- P(k Fej) = $\binom{5}{k} 0.4^k 0.6^{5-k}$.
- P(2 Fej) = $\binom{5}{2} 0.4^2 0.6^3 = 0.346.$
- P(1-nél több Fej) = 1 – P(0 Fej) – P(1 Fej) = $1 - \binom{5}{0} 0.4^0 0.6^{5-0} - \binom{5}{1} 0.4^1 0.6^{5-1} = 0.663.$

Excelben kiszámíthatjuk a binomiális valószínűségi változó valószínűségeit nagyobb számokra is.

A binomiális eloszlás várható értéke és varianciája

Ezeket könnyen kiszámíthatjuk, ha észrevesszük, hogy egy X binomiális eloszlású valószínűségi változó felírható az egyes kísérletekből származó n független Bernoulli eloszlású valószínűségi változó $X_1, X_2,\ldots, X_n$ összegeként.

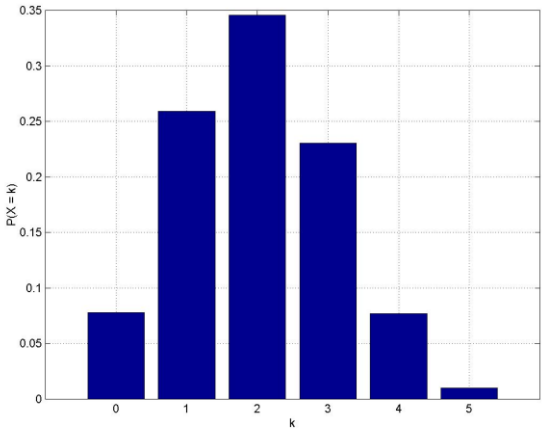
X várható értéke:

$$E(X) = E(X_1+X_2+\ldots+X_n) = E(X_1)+E(X_2)+\ldots+E(X_n) = np.$$

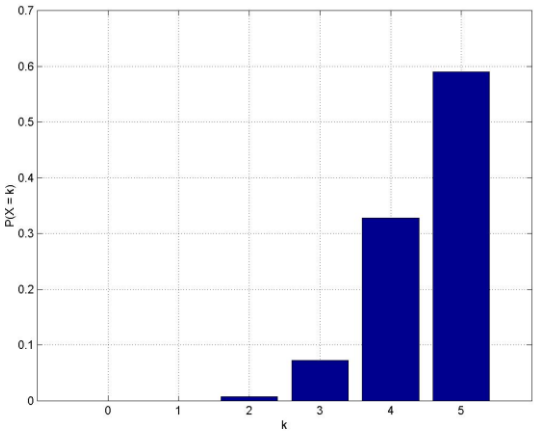
X varianciája:

$$Var(X) = Var(X_1+X_2+\ldots+X_n) = Var(X_1)+Var(X_2)+\ldots+Var(X_n) = np(1-p).$$

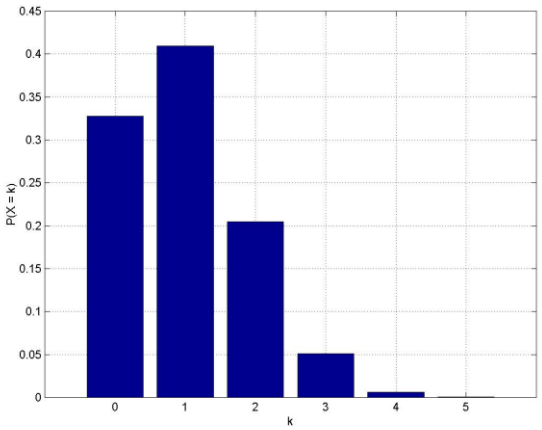
Ábra 2.3.1. - Binomiális valószínűségi változó tömegfüggvényének oszlopdiagramja, ha n = 5, p = 0.4



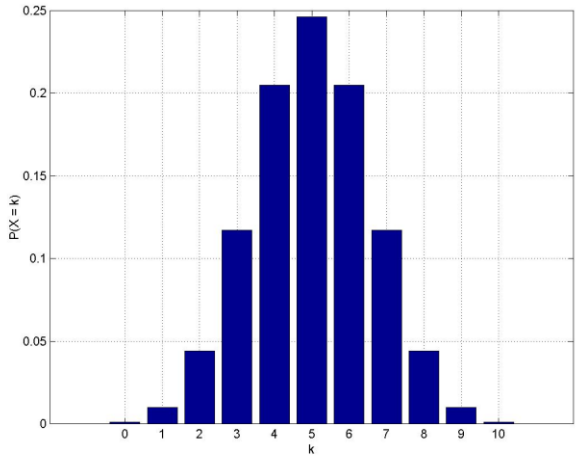
Ábra 2.3.2. - Binomiális valószínűségi változó tömegfüggvényének oszlopdiagramja, ha n = 5, p = 0.9



Ábra 2.3.3. - Binomiális valószínűségi változó tömegfüggvényének oszlopdiagramja, ha n = 5, p = 0.2



Ábra 2.3.4. - Binomiális valószínűségi változó tömegfüggvényének oszlopdiagramja, ha n = 10, p = 0.5



A binomiális eloszlás feltételei

Vigyázzunk arra, hogy a tanulmányozott helyzet teljesítse a binomiális eloszlás feltételeit:

- a Bernoulli kísérletek legyenek független egymástól,
- a Bernoulli kísérletek siker-valószínűségei legyenek egyenlőek.

Ha valamelyik ezek közül nem teljesül, akkor a binomiális eloszlás nem alkalmazható!

Példa: A binomiális eloszlás nem alkalmazható az egyetemen használt projektorok lámpáinak meghibásodásának modellezésére, ugyanis a projektorok különböző mértékben használtak, ezért a meghibásodási valószínűségek nem egyenlőek.

Poisson eloszlás

Egy Poisson valószínűségi változó a 0, 1, 2, ... értékeket veszi fel. Tömegfüggvénye:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0,1,2,\ldots,$$

$\lambda > 0$ a modell paramétere. A valószínűségek kiszámításához kell ismerjük a paraméter értékét. A Poisson eloszlást gyakorlati szempontból fontos helyzetek modellezésére használják.

Például:

- Azon ügyfelek száma, akik egy adott időintervallumban lépnek be egy bankba vagy üzletbe.
- Telefonhívások száma egy adott időintervallumban.
- Autóbalesetek száma egy év alatt.

A Poisson eloszlás várható értéke és varianciája

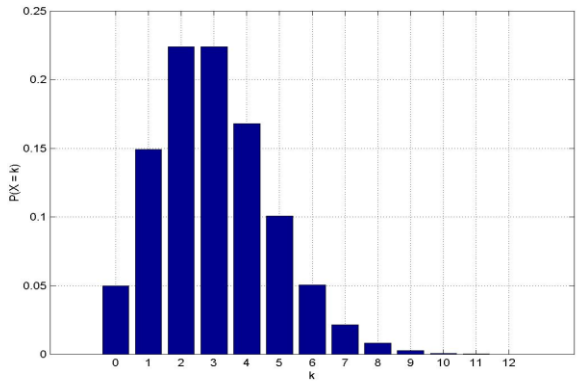
Mind a várható érték, mind a variancia a Poisson eloszlás paraméterével egyenlő:

$$X \text{ várható értéke: } E(X) = \lambda, \quad X \text{ varianciája: } \quad Var(X) = \lambda.$$

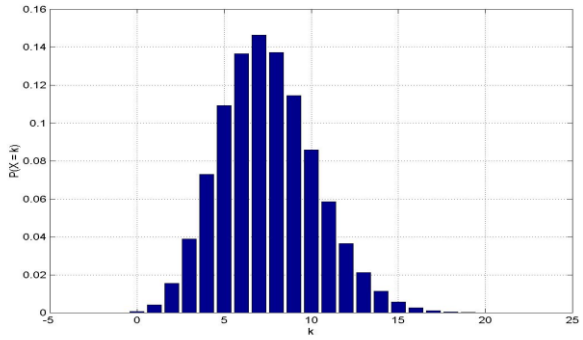
Példa. Egy bankban a percenként érkező ügyfelek számának várható értéke 3. Legyen X egy adott percben érkező ügyfelek száma. Ebben az esetben $\lambda = 3$, tehát:

- X tömegfüggvénye: $P(X = k) = \frac{3^k}{k!}e^{-3}, \quad k = 0,1,2,\ldots$
- $P(X = 2) = \frac{3^2}{2!}e^{-3} = 0.224.$
- $P(X > 2) = 1 - P(X = 0) - P(X = 1) = 1 - \frac{3^0}{0!}e^{-3} - \frac{3^1}{1!}e^{-3} = 0.8.$

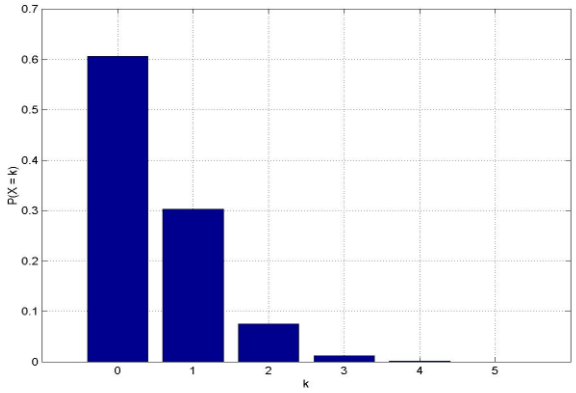
Ábra 2.3.5. - Poisson valószínűségi változó tömegfüggvényének oszlopdiagramja, ha λ = 3



Ábra 2.3.6. - Poisson valószínűségi változó tömegfüggvényének oszlopdiagramja, ha λ = 7.5



Ábra 2.3.7. - Poisson valószínűségi változó tömegfüggvényének oszlopdiagramja, ha λ = 0.5



Az értelmezése alapján a Poisson eloszlás olyan események **számát** modellezi, amelyek akkor következhetnek be, ha:

- az események valószínűsége kicsi,
- a kísérletek száma nagy.

Példák.

Egy bank ügyfelei

> A lehetséges ügyfelek száma nagyon nagy, és annak a valószínűsége, hogy egy bizonyos ügyfél elmegy a bankba egy bizonyos időintervallumban nagyon kicsi. Ebben az esetben az érkező ügyfelek száma Poisson eloszlású.

Autóbalesetek száma

> Biztosítótársaságok Poisson és továbbfejlesztett Poisson eloszlással modellezik az ügyfelek baleseteinek a számát egy bizonyos időintervallum (mondjuk egy év) alatt.

II.4 Fontosabb folytonos valószínűségi változók

Emlékezzünk vissza, hogy a folytonos valószínűségi változók egy intervallumban bármilyen értéket felvehetnek. Ezenkívül, egy folytonos valószínűségi változó eloszlását az f(x) sűrűségfüggvénye által értelmezzük. A sűrűségfüggvény segítségével valószínűségeket számíthatunk ki:

$$P(a < X < b) = \int_a^b f(x)\ dx.$$

Más típusú valószínűségek kiszámítása ugyanígy történik:

$$P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b) = \int_a^b f(x)\ dx.$$

Ez azért van, mert a vonal területe 0, vagyis,

$$P(X = a) = \int_a^a f(x)\ dx = 0.$$

<u>Az egyenletes eloszlás</u>

Tegyük fel, hogy X egy olyan valószínűségi változó amelyre:

- X az a és b számok között van,
- X bármelyik értéke a és b között ugyanannyira valószínű.

Ez azt jelenti, hogy annak a valószínűsége, hogy X egy adott intervallumban van <u>csak az intervallum hosszától függ</u>, és nem függ az intervallum helyétől.

Az egyedüli sűrűségfüggvény amely ezt a tulajdonságot teljesíti:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, \text{ ha } a \leq x \leq b, \\ 0, \quad \text{másképp.} \end{cases}$$

Az a és *b* számok az eloszlás paraméterei.

Példa: Az X sűrűségfüggvénye, amely egyenletes eloszlású 5 és 13 között: *f(x)* = 1/8, ha $5 \leq x \leq 13$, és *f(x)* = 0, másképp. Erre a valószínűségi változóra:

P(6 < X < 10) = (10-6) / 8 = 0.5,

P(X < 9) = (9-5) / 8 = 0.5.

Az egyenletes eloszlás várható értéke:

$$E(X) = \frac{a+b}{2}.$$

Az egyenletes eloszlás varianciája:

$$Var(X) = \frac{(b-a)^2}{12}.$$

Az exponenciális eloszlás

Az exponenciális eloszlással már találkoztunk egy példában. A sűrűségfüggvénye:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

$\lambda > 0$ az eloszlás paramétere.

Az exponenciális eloszlás várható értéke:

$$E(X) = 1/\lambda$$

Az exponenciális eloszlás varianciája:

$$Var(X) = 1/\lambda^2$$

Az exponenciális eloszlás gyakorlati szempontból hasznos, számos helyzetre alkalmazzák. Gyakran használják két esemény között eltelt idő modellezésére:

- két egymás után megjelenő autó az utcán,
- két egymás után érkező ügyfél egy bankban,
- különböző rendszerek meghibásodása között eltelt idő.

A normális eloszlás

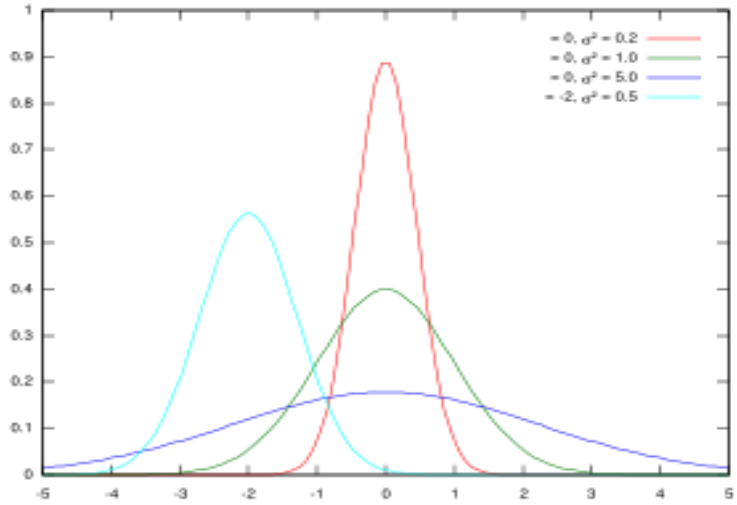
A statisztikában a legfontosabb eloszlás. Sok gyakorlati szempontból fontos helyzetre alkalmazzák. A két legfontosabb statisztikai mutató, a várható érték és a variancia, egyértelműen meghatározzák.

A sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \qquad -\infty < x < \infty.$$

Jelölés: $X \sim N(\mu, \sigma^2)$, ahol μ és σ^2 a normális eloszlás paraméterei.

Ábra 2.4.1. - A sűrűségfüggvény grafikonja a paraméterek különböző értékeire



Egy normálisis eloszlású valószínűségi változó várható értéke μ, varianciája σ^2, tehát a szórása σ. Amint a fenti ábra is mutatja, a sűrűségfüggvény változik a μ és σ^2 paraméterek különböző értékeire; ezért a valószínűségek is változnak.

A normálisis eloszlás sűrűségfüggvénye teljesíti a következő tulajdonságokat:

- A maximumát a m várható értékben éri el.
- Szimmetrikus μ körül, tehát a mediánja is μ .
- Annak a valószínűsége, hogy X μ - σ és μ + σ között van 0.6826.
- Annak a valószínűsége, hogy X μ - 2 σ és μ + 2 σ között van 0.9545.
- Annak a valószínűsége, hogy X μ - 3 σ és μ + 3 σ között van 0.9973.
- 0.5 valószínűséggel X μ - 0.675 σ és μ + 0.675 σ között van.

A standard normális eloszlás az a normális eloszlás amelynek várható értéke 0 és varianciája 1. Sűrűségfüggvénye:

$$f(x) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}x^2\right), \qquad -\infty < x < \infty.$$

A standard normális valószínűségi változót általában Z-vel jelöljük: Z ~ N(0, 1). A normális eloszlás valószínűségeit nem tudjuk könnyen kiszámítani, mert az f(x) integrál nem számítható ki egyszerű képlettel. Ezért a standard normális eloszlás

$$P(Z \leq z) = \int_{-\infty}^{z} f(x)\ dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\ dx$$

valószínűségeit megközelítőleg szokták kiszámítani. Ezeket a valószínűségeket egyféleképpen táblázatokból, másféleképpen Excel segítségével tudjuk kiszámítani.

A táblázatok általában a P(Z < z) vagy a P(0 < Z < z) értékeket tartalmazzák z ≥ 0-ra. Ebben az esetben:

P(Z < z) = P(Z < 0 vagy 0 < Z < z) = P(Z < 0) + P(0 < Z < z) = 0.5 + P(0 < Z < z).

Példa: P(Z < 1) = 0.5 + P(0 < Z < 1), ahol a P(0 < Z < 1) valószínűséget ki tudjuk nézni táblázatból. Ennek értéke 0.341, ezért P(Z < 1) = 0.841.

Mivel P(nem A) = 1 – P(A) bármilyen A eseményre, ezért

P(Z > z) = 1 – P(Z < z)

Mivel a standard normális sűrűségfüggvény **szimmetrikus 0 körül**, ezért:

P(Z < – z) = 1 – P(Z < z)

Példák:

- P(Z < – 1.5) = 0.0668,
- P(Z > 0.5) = 1 - P(Z < 0.5) = 0.6915,
- P(Z > – 1.25) = 1 - P(Z < – 1.25) = 1 - 0.1056 = 0.8944.

Egy (a, b) intervallumra

P(a < Z < b) = P(Z < b) – P(Z < a),

tehát használhatjuk a táblázatot, hogy meghatározzuk a P(a < Z < b) valószínűséget bármilyen intervallumra.

Példák:

- P(1 < Z < 2) = P(Z < 2) – P(Z < 1) = P(0 < Z < 2) – P(0 < Z < 1) = 0.477 – 0.341 = 0.136.
- P(–1.5 < Z < 0.2) = P(Z < 0.2) – P(Z < –1.5) = 0.5 + P(0 < Z < 0.2) – P(1.5 < Z) = 0.5 + P(0 < Z < 0.2) – [1 – P(Z < 1.5)] = 0.5 + P(0 < Z < 0.2) – [1 – 0.5 – P(0 < Z < 1.5)] = P(0 < Z < 0.2) + P(0 < Z < 1.5) = 0.079 + 0.433 = 0.512.

A táblázatok csak a standard normális eloszlás valószínűségeit tartalmazzák. A normális eloszlás valószínűségei különböző μ és σ^2 értékekre kiszámíthatók a standard normális eloszlás valószínűségeiből. Erre a célra a következő tulajdonságot használjuk: ha $X \sim N(\mu,\sigma^2)$ akkor

$$Z = \frac{X - \mu}{\sigma}.$$

standard normális eloszlású.

Példa: Ha $X \sim N(2, 9)$ akkor

$$\mathrm{P}(X < 3) \;=\; \mathrm{P}\left(\frac{X-2}{3} < \frac{3-2}{3}\right) = P\left(Z < \frac{1}{3}\right).$$

<u>Normális eloszlású valószínűségi változók lineáris kombinációja</u>

A normális eloszlásnak a következő tulajdonságai vannak:

- Ha $X \sim N(\mu,\sigma^2)$ és c konstans akkor:

  $cX \sim N(c\,\mu, c^2\,\sigma^2)$.

- Ha $X \sim N(\mu_1, \sigma_1^2)$ és $Y \sim N(\mu_2, \sigma_2^2)$, akkor X + Y is normális eloszlású. Ha X és Y függetlenek is, akkor:

  $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Példa: X ~ N(0, 4) és Y ~ N(2, 9), és X és Y függetlenek. Ekkor a P(–2X + Y < 4) valószínűség kiszámításához először meghatározzuk T = –2X eloszlását. A fenti tulajdonságok alapján T = –2X ~ N(0,16). Ezért P(–2X + Y < 4) = P(T + Y < 4), ahol W = T + Y ~ N(2, 25). Így a keresett valószínűséget visszavezettük olyan formára, amilyent már kiszámítottunk: P(W < 4), ahol W ~ N(2, 25).

Valószínűségek inverze

Legyen Z ~ N(0, 1). Mi a z értéke amelyre P(Z < z) = 0.81? A táblázatból kiolvashatjuk, hogy:

z = 0.88

Ezt a számot a valószínűség inverzének nevezzük.

Ha X ~ N(μ,σ^2), akkor tudjuk, hogy Z = (X – μ)/ σ ~ N(0,1); ezért X = σ Z + μ. Tehát, ha X ~ N(μ,σ^2) akkor az x-et amelyre P(X < x) = p úgy kapjuk, hogy **x = σ z + μ**, ahol z-t a P(Z < z) = p összefüggés határozza meg.

Példa: X ~ N(-4, 1). Mi az x értéke amelyre P(X < x) = 0.6? A fenti tulajdonság alapján először a z-t kell meghatározzuk, amelyre P(Z < z) = 0.6, tehát P(0 < Z < z) = 0.1. A táblázatból z = 0.25. Ez alapján: x = z – 4 = – 3.75.

A lognormális eloszlás

Gyakran használják gazdasági tanulmányokban. Például a háztartások éves jövedelmét gyakran modellezik a lognormális eloszlással. Ez abból látszik, hogy ha elkészítjük a háztartások éves jövedelmének a hisztogramját, ez hasonlít a lognormális eloszlás sűrűségfüggvényéhez. A nevét onnan kapta, hogy olyan valószínűségi változónak az eloszlása amelynek a logaritmusa normális eloszlású.

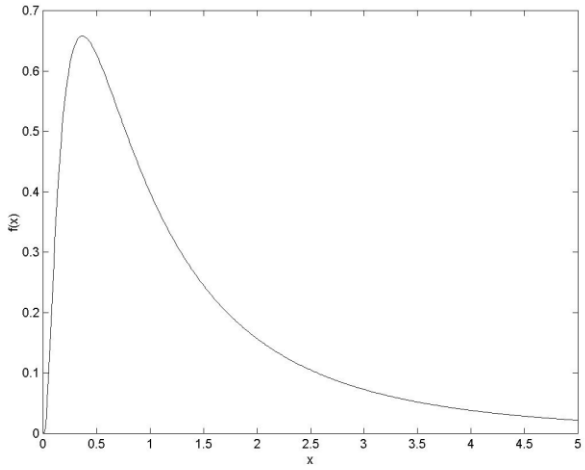
X lognormális eloszlású, ha ln(X) ~ N(μ,σ^2). Az X sűrűségfüggvénye:

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}x}\, exp\left(-\frac{1}{2\sigma^2}(log\, x - \mu)^2\right), & \text{ha } x > 0, \\ 0, & \text{ha } x \leq 0. \end{cases}$$

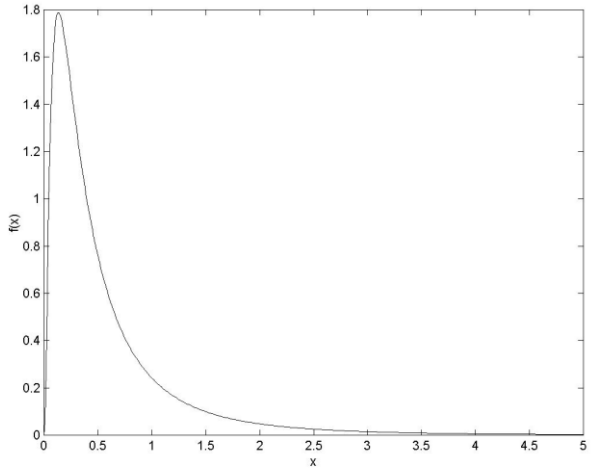
Ennek a sűrűségfüggvénynek nehéz az integrálját kiszámítani, hasonlóan a normális eloszláséhoz. Viszont a valószínűségek kiszámításához használhatjuk a normális eloszlást.

Példa: Ha Y = ln(X) ~ N(-1, 1) akkor P(X < 1) = P(ln(X) < ln(1)) = P(Y < 0), amit könnyen kiszámíthatunk felhasználva, hogy Y normális eloszlású.

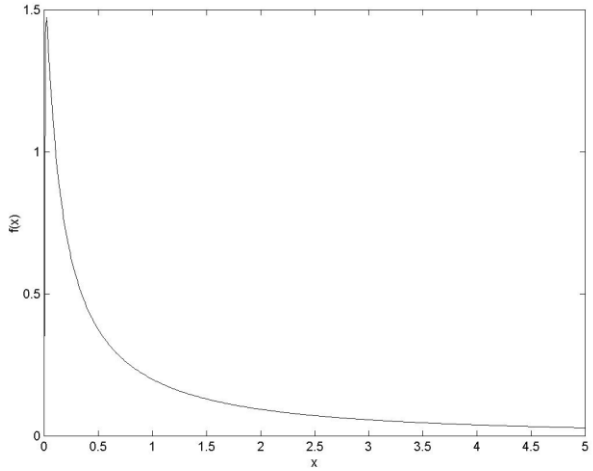
Ábra 2.4.2. - A lognormális sűrűségfüggvény ha $\mu = 0$, $\sigma^2 = 1$



Ábra 2.4.3. - A lognormális sűrűségfüggvény ha $\mu = -1$, $\sigma^2 = 1$



Ábra 2.4.4. - A lognormális sűrűségfüggvény ha $\mu = 0$, $\sigma^2 = 4$



II.5 Többváltozós eloszlások

Tegyük fel, hogy több valószínűségi változót szeretnénk tanulmányozni azért, hogy tanulmányozzuk a köztük lévő összefüggést. Ezeknek a valószínűségi változóknak értelmezhetjük az eloszlásukat együtt. Ezt egyidejű eloszlásnak nevezzük. Az egyidejű eloszlást különbözőképpen értelmezzük diszkrét és folytonos valószínűségi változókra. Részletesebben csak a diszkrét egyidejű eloszlást fogjuk tárgyalni két változóra.

Az egyidejű tömegfüggvény

Legyenek X és Y diszkrét valószínűségi változók. Jelöljük az X lehetséges értékeit $x_1,x_2,...,x_n$-nel és az Y-ét $y_1, y_2,..., y_m$-mel. Az X és Y egyidejű tömegfüggvénye:

$$p(x_i,y_j) = P(X = x_i \text{ és } Y = y_j).$$

Ezt másképp kétváltozós tömegfüggvénynek is nevezik.

Megjegyzés: általában, az X_1, X_2,..., X_n diszkrét valószínűségi változók egyidejű tömegfüggvénye:

$$p(x_1, x_2,..., x_n) = P(X_1 = x_1 \text{ és } X_2 = x_2 \text{ és } ... \text{ és } X_n = x_n).$$

Tömegfüggvény-táblázat

Két diszkrét valószínűségi változó egyidejű tömegfüggvényét megadhatjuk egy táblázattal, ugyanúgy ahogy az egyidejű gyakoriság- eloszlással tettük:

Táblázat 2.5.1. – Egyidejű gyakoriság eloszlás táblázat

	X				
Y	x_1	x_2	x_3	...	x_n
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$	$p(x_3, y_1)$	...	$p(x_n, y_1)$
y_2	$p(x_1, y_2)$	$p(x_2, y_2)$	$p(x_3, y_2)$	...	$p(x_n, y_2)$
...	...	...	...	...	...
y_m	$p(x_1, y_m)$	$p(x_2, y_m)$	$p(x_3, y_m)$	...	$p(x_n, y_m)$

Az egyidejű eloszlás tulajdonságai

Mivel valószínűségekről van szó, ezért teljesítik a következő tulajdonságokat:

$$0 \le p(x_i, y_j) \le 1,$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) = 1.$$

Példa: X az 1, 2, 3 és 4 értékeket veszi fel; Y a 0, 1 és 2-t.

Táblázat 2.5.2. – Példa: Az egyidejű eloszlás

	X			
Y	1	2	3	4
0	0.12	0.00	0.23	0.05
1	0.13	0.03	0.02	0.05
2	0.05	0.16	0.10	0.06

Tehát P(X = 2, Y=0) = 0.0, P(X = 3, Y = 2) = 0.1, stb. Megfigyelhetjük, hogy a táblázatban szereplő valószínűségek összege 1.

Peremeloszlások

Az X és Y változók eloszlásait peremeloszlásoknak nevezzük. A peremeloszlásokat az egyidejű eloszlásból a teljes valószínűség törvényével számítjuk ki (vagyis összeadjuk az x_i -nek megfelelő oszlop elemeit):

$$P(X = x_i) = \sum_{j=1}^{m} P(X = x_i \text{ és } Y = y_j) = \sum_{j=1}^{m} p(x_i, y_j).$$

Hasonlóan:

$$P(Y = y_j) = \sum_{i=1}^{n} p(x_i, y_j),$$

vagyis összeadjuk a táblázat y_j-nek megfelelő sorának elemeit.

Megfigyelhetjük, hogy a peremeloszlásokat ugyanúgy határozzuk meg mint a perem gyakoriság-eloszlásokat.

Táblázat 2.5.3. - Példa: Az alábbi táblázat tartalmazza a peremeloszlásokat a fenti példára

	X				
Y	1	2	3	4	**P(Y = y)**
0	0.12	0.00	0.23	0.05	**0.4**
1	0.13	0.03	0.02	0.05	**0.23**
2	0.05	0.16	0.10	0.06	**0.37**
P(X = x)	**0.30**	**0.19**	**0.35**	**0.16**	**1.00**

Feltételes eloszlások

Ha meg van adva egy $Y = y_j$ érték, az X feltételes eloszlása ha $Y = y_j$ a következő:

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i \text{ és } Y = y_i)}{P(Y = y_i)} = \frac{p(x_i, y_j)}{P(Y = y_i)}, \quad i = 1, \dots, n.$$

Ez nem más mint az **$Y = y_j$-nek megfelelő sor a táblázatban**, elosztva a sor elemeinek összegével (akárcsak a feltételes gyakoriság-eloszlásnál).

Ugyanígy, ha adva van az $X = x_i$, az Y feltételes eloszlása ha $X = x_i$:

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i \text{ és } Y = y_i)}{P(Y = y_i)} = \frac{p(x_i, y_j)}{P(X = x_j)}, \quad j = 1, \dots, m.$$

Ismét, ez az $X = x_i$–nek megfelelő oszlop a táblázatból elosztva az oszlop elemeinek összegével.

Az X feltételes eloszlása, ha Y = 1:

P(X = 1|Y = 1) = 0.13/0.23, P(X = 2|Y = 1) = 0.03/0.23,

P(X = 3|Y = 1) = 0.02/0.23, P(X = 4|Y = 1) = 0.05/0.23.

Az Y feltételes eloszlása, ha X =2:

P(Y = 0|X = 2) = 0.0/0.19, P(Y = 1| X = 2) = 0.03/0.19,

P(Y = 3| X = 2) = 0.16/0.19.

Független valószínűségi változók

Két X és Y valószínűségi változó független, ha

$$P(X = x \,|Y = y) = P(X = x) \text{ és } P(Y = y \mid X = x) = P(Y = y)$$

minden x és y értékre.

Ezért:

$$p(x,y) = P(X=x \text{ és } Y=y) = P(X =x)\, P(Y = y \,|X = x) = P(X = x)\, P(Y = y),$$

ami azt jelenti, hogy az egyidejű tömegfüggvény a perem tömegfüggvények szorzata.

Hogyan ellenőrizzük, hogy X és Y függetlenek-e?

Az egyidejű eloszlás táblázatából:

- kiszámítjuk az X és Y perem tömegfüggvényeit,
- ellenőrizzük, hogy p(x,y) = P(X = x) P(Y = y) **minden** x és y-ra.
- Ha igaz, akkor X és Y függetlenek,
- ha van olyan x és y amelyre nem igaz, akkor nem függetlenek.

Táblázat 2.5.4. - Példa: Az X és Y perem-tömegfüggvényei

X	1	2	3	4
$P(X = x)$	0.3	0.19	0.35	0.16

Y	0	1	2
$P(Y = y)$	0.4	0.23	0.37

Az egyidejű valószínűség p(1,0) = 0.12; a valószínűségek szorzata P(X = 1) P(Y = 0) = 0.3 × 0.4 = 0.12. Tehát teljesül a feltétel ebben az esetben.

De p(2,0) = 0 és P(X = 2) P(Y = 0) = 0.19 × 0.4 = 0.076, tehát p(2,0) nem egyenlő P(X = 2) × P(Y = 0)-val. Ezért X és Y **nem független.**

Példa független valószínűségi változókra. Tekintsük a következő egyidejű eloszlást:

Táblázat 2.5.5. – Egyidejű eloszlás táblázat

	X				
Y	1	2	3	4	**P(Y = y)**
1	0.15	0.075	0.03	0.045	0.3
2	0.15	0.075	0.03	0.045	0.3
3	0.2	0.1	0.04	0.06	0.4
P(X = x)	0.5	0.25	0.1	0.15	

Ellenőrizzük, hogy p(x, y) = P(X = x) P(Y = y) fennáll-e minden x és y-ra. Az első 3 eset:

p(1,1) = 0.15; P(X = 1) P(Y = 1) = 0.5 × 0.3 = 0.15,

p(2,1) = 0.075; P(X = 2) P(Y = 1) = 0.25 × 0.3 = 0.075,

p(3,1) = 0.03; P(X = 3) P(Y = 1) = 0.1 × 0.3 = 0.03.

Mindhárom esetben az egyidejű a valószínűségek egyenlők az egyenkénti valószínűségek szorzatával. A többi eset leellenőrzése után kapjuk, hogy X és Y valóban függetlenek.

Két valószínűségi változó kovarianciája

Két X és Y valószínűségi változó kovarianciája a következő:

$$\mathrm{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = \sum_x \sum_y (x - E(X))(y - E(Y))p(x,y).$$

Ugyanúgy mint a változóknál, a kovariancia két valószínűségi változó lineáris összefüggését méri. A várható érték tulajdonságai alapján, a kovariancia egy másik, az előzővel egyenértékű képlete:

Cov(X,Y) = E(XY) – E(X)E(Y), mivel

$$\begin{aligned} &E[(X - E(X))(Y - E(Y))] \\ &= E[XY - E(X)Y - XE(Y) + E(X)E(Y)] \\ &= E[XY] - E[E(X)Y] - E[XE(Y)] + E[E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y), \end{aligned}$$

ahol

$$E(XY) = \sum_x \sum_y xyp(x,y)\,.$$

Példa: Az alábbi táblázatban megadott egyidejű eloszlás alapán kiszámíthatjuk X, Y kovarianciáját a Cov(X,Y) = E(XY) – E(X) E(Y) képlettel. Előbb kiszámítjuk a következő várható értékeket: E(XY), E(X), E(Y)-t.

Táblázat 2.5.6. – Egyidejű eloszlás táblázat

	X				
Y	1	2	3	4	**P(*Y* = *y*)**
0	0.12	0.00	0.23	0.05	**0.4**
1	0.13	0.03	0.02	0.05	**0.23**
2	0.05	0.16	0.10	0.06	**0.37**
P(*X* = *x*)	**0.30**	**0.19**	**0.35**	**0.16**	**1.00**

E(XY) = 1 × 0.13 + 2 × 0.03 + 3 × 0.02 +4 × 0.05 + 2 × 0.05 + 4 × 0.16 + 6 × 0.1 + 8 × 0.06 = 2.27,

E(X) = 1 × 0.3 + 2 × 0.19 + 3 × 0.35 + 4 × 0.16 = 2.37,

E(Y) = 1 × 0.23 + 2 × 0.37 =0.97.

Tehát Cov(X,Y) = 2.27 – 2.37 × 0.97 = –0.029.

Ha X és Y függetlenek akkor Cov(X,Y) = 0

Ez azért van mert:

$$E(XY) = \sum_x \sum_y xyp(x,y) = \sum_x \sum_y xyp(x)p(y)$$
$$= \sum_x xp(x) \sum_y yp(y) = E(X)E(Y)$$

Tehát: Cov(X,Y) = E(XY) – E(X)E(Y) = E(X)E(Y) – E(X)E(Y) = 0.

Vagyis, ha X és Y függetlenek, akkor köztük lineáris összefüggés sincs. **Fontos** megjegyezni, hogy ennek a fordítottja **nem igaz**:

Ha X és Y között nincs lineáris összefüggés (Cov(X,Y) = 0) nem jelenti azt, hogy X és Y függetlenek.

<u>Két valószínűségi változó összegének varianciája</u>

Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)

Ez azért van, mert a Z = X + Y valószínűségi változó varianciája:

$\text{Var}(X+Y) = E[\ (X+Y)^2 - (E(X+Y))^2\ ]$

$= E[X^2 + Y^2 + 2XY - (E(X) + E(Y))^2\ ]$

$= E[X^2 + Y^2 + 2XY - (E(X))^2 - (E(Y))^2 - 2E(X)E(Y)]$

$= E[X^2 - (E(X))^2] + E[Y^2 - (E(Y))^2] + 2[E(XY) - E(X)E(Y)]$

$= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y).$

<u>Két független valószínűségi változó összegének varianciája</u>

Ha X és Y függetlenek akkor:

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

Ez azért van, mert két független valószínűségi változó kovarianciája 0:

$$\text{Cov}(X,Y) = 0.$$

Ez a tulajdonság érvényes több független valószínűségi változóra is: ha $X_1, X_2, \ldots, X_n$ független valószínűségi változók, akkor

$$\text{Var}(X_1 + X_2 + \ldots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n).$$

II.6. Gyakorlatok

1. Megfigyelték, hogy az X márkájú üdítőből a fogyasztók 1, 2, 3 vagy 4 míg az Y márkájú üdítőből a fogyasztók 0, 1 vagy 2 üveggel vásárolnak egy hét alatt. Azoknak az eseményeknek a valószínűségeit, hogy egy véletlenszerűen kiválasztott fogyasztó hányat vásárol a két típusú üdítőből a következő táblázat tartalmazza:

		X			
		1	**2**	**3**	**4**
Y	**0**	0.12	0.00	0.23	0.05
	1	0.13	0.03	0.02	0.05
	2	0.05	0.16	0.10	0.06

Például, a táblázat alapján annak a valószínűsége, hogy egy véletlenszerűen kiválasztott fogyasztó az X-ből 2-őt és az Y-ból egyet sem vásárol 0.

a. Mi annak a valószínűsége, hogy egy véletlenszerűen kiválasztott fogyasztó az Y-ból egyet vásárol? (E: P(Y=1)=0.23)
b. Mi egy véletlenszerűen kiválasztott fogyasztó által vásárolt X márkájú üdítők számának várható értéke?(E: E(X)=2.37)
c. Független-e egymástól a vásárolt X és Y márkájú üdítők száma? (E: Nem függetlenek egymástól)

2. A kedvenc csapatunk még versenyben van a kupában (ezt jelöli a X valószínűségi változó), ahol megszerezheti az első 4 hely valamelyikét, és a bajnokságban is (ezt jelöli az Y valószínűségi változó, ahol a 4. és 7. hely között végezhet. A helyezések valószínűségét tartalmazza a következő táblázat:

		X			
		1	**2**	**3**	**4**
Y	**4**	0.070	0.140	0.280	0.210
	5	0.015	0.030	0.060	0.045
	6	0.010	0.020	0.040	0.030
	7	0.005	0.010	0.020	0.015

Például, a táblázat alapján annak a valószínűsége, hogy a kupában második és a bajnokságban ötödik legyen az 3%.

a. Számoljátok ki a perem valószínűségeket. Mi a valószínűsége annak, hogy a csapat megnyeri a kupát? (E: P(X=1)=0.1; P(X=2)=0.2; P(X=3)=0.4; P(X=4)=0.3; P(Y=4)=0.7; P(Y=5)=0.15; P(Y=6)=0.1; P(Y=7)=0.05; 10% eséllyel nyeri meg a kupát)
b. Várhatóan a csapat hányadik helyen végez a bajnokságban? (E: E(Y)=4.5)
c. Határozzátok meg a bajnoki helyezés varianciáját (Y változó varianciája)? (E: Var(Y)=0.75)

3. A szeptember 12.-én megjelenő, várhatóan iPhone 99-nak nevezett új Apple telefon árát tekintsük egy valószínűségi változónak, amelynek a sűrűség függvénye a következő:

$$f(x) = \begin{cases} 0.4, & ha\ 8 \leq x < 9 \\ \frac{12-x}{7.5} & ha\ 9 \leq x \leq 12 \end{cases}$$

Az x értékei 100 euróban vannak kifejezve a függvényben, vagyis 800 euró megfelel a 8-nak.

a. Mi annak a valószínűsége, hogy a megjelenő telefon olcsóbb, mint 850 euró (8.5)? (E: P(x<=8.5)=0.2)
b. Mi annak a valószínűsége, hogy a megjelenő telefon ára 1 000 és 1 200 euró (10 és 12) között lesz? (E: P(10<=x<=12)=0.267)
c. Határozzuk meg a telefon árának várható értékét, mediánját! (E: E(X)=1.051 euró, medián 9.260 euró)

4. Tegyük fel, hogy egy nagy italüzletben egy véletlenszerűen kiválasztott üveg sör árát lejben egy valószínűségi változónak tekintjük, amelynek sűrűségfüggvénye

$$f(x) = \begin{cases} 0.5 & ha\ 2 \leq x < 3 \\ \frac{5-x}{4} & ha\ 3 \leq x \leq 5 \\ 0 & ha\ x < 2\ vagy\ x > 5 \end{cases}$$

a. Mi annak a valószínűsége, hogy egy véletlenszerűen kiválasztott sör olcsóbb mint 3 lej? (E: P(x<3)=0.5)
b. Mi annak a valószínűsége, hogy egy véletlenszerűen kiválasztott sör drágább mint 4 lej? (E: P(x>4)=0.125)
c. Határozzuk meg a sör árának várható értékét és mediánját! (E: E(x)=3.08; medián 3)

5. Egy biztosító társaság megfigyelte az évek során, hogy az ügyfeleinél két egymásutáni autóbaleset között eltelt idő hónapokban kifejezve egy exponenciális eloszlású valószínűségi változó, amelynek sűrűségfüggvénye f (x) = $2e^{-2x}$, ha x ≥ 0 és f (x) = 0, ha x < 0. Tegyük fel, hogy az egyik ügyfél most jelentett egy autóbalesetet.

a. Mennyi két egymásutáni autóbaleset között eltelt idő várható értéke és varianciája? (E: E(x)=1/2; Var(x)=1/4)
b. Mi annak a valószínűsége, hogy a következő baleset egy hónapon belül következik be? (E: P(X<1)=0.865)
c. Ha azt feltételezzük, hogy a következő baleset nem következik be egy hónapon belül, mi annak a valószínűsége, hogy az utána levő hónapban sem következik be? (E: P(X>2 | X>1)=0.135)

6. Az Egyesült Államok egyetemi férfi kosárlabda bajnokságát (NCAA), 2014-ben, a Connecticut Huskies nyerte. Az NCAA az NBA (Észak-Amerikai profi kosárlabda bajnokság) előszóbájának is számít, nagyon sok játékost beválasztanak (draftolnak) az NBA-be, ha jól játszanak az egyetemi bajnokságban. Az eddig 76 szezon alapján azt mondhatjuk el, hogy annak az esélye, hogy a nyertes csapat egy tagja bekerüljön az NBA-be 40%.

a. Ha a Connecticut Huskies csapatának játékoskerete 20 főből áll, várhatóan hányan játszhatnak majd az NBA-ben ebből a csapatból? Mi a varianciája az NBA-ben játszó játékosok számának? (E: E(x)=8; Var(x)=4.8.)
b. Mi a valószínűsége annak, hogy legfeljebb csak 2 játékost draftolnak? (E: P(X<=2)=0.36%)
c. Az eddigi rekord, hogy egy nyertes csapat 18 játékosát is draftolták. Mi az esélye annak, hogy a Connecticut beállítja vagy akár meg is dönti ezt a rekordot? (E: P(X>=18) = 0.0005%)

7. Telefónia országában két nemzetiségű ember él: androidák és ájfonok. Lehetőségetek van ellátogatni az országban és beszélni 9 ottani őslakossal. A lakosság 65% androida és 35%-a ájfon. A 9 ottani lakossal véletlenszerűen futtok össze, nem ti döntitek el kivel találkoztok.

a. A 9 lakos közül várhatóan hány lesz androida? Határozzátok meg a tömegfüggvényét, annak a valószínűségi változónak, hogy androidával találkoztok! (E: E(x)=5.85; P(X=k)=$C_9^k * 0.65^k * 0.35^{(9-k)}$)
b. Mi a valószínűsége annak, hogy több mint 2 (legalább 3) ájfonnal találkozhattok? (E: P(y>2)=0.6627=66.27%)
c. Ketten is elutaztok az országba, mi a valószínűsége, hogy egyiketek csak ájfonokkal, és a másikatok csak androidákkal találkozik?
(E: P(x=9 és y=9)=0.00000163=0.000163%)

8. Egy fagyi gép közelében ültök, amelynél csak szalonnás vagy tökmagos fagyit lehet vásárolni. Unalmatokban számoljátok, hogy az ügyfelek milyen fagyit vásárolnak, és azt tapasztaljátok, hogy 25%-a az ügyfeleknek szalonnást, míg a 75%-k tökmagost választ. Az elkövetkező fél órában 12-n fognak fagyit vásárolni.

a. A 12 vásárló közül várhatóan hányan vesznek szalonnás fagyit? Határozzátok meg a tömegfüggvényét, annak a valószínűségi változónak, hogy szalonnás fagyit választanak! (E: E(x) = 3; P(X=k)=$C_{12}^k * 0.25^k * 0.75^{(12-k)}$)
b. Mi a valószínűsége annak, hogy több mint 10 (legalább 11) vásárló választ tökmagos fagyit? (E: P(y>10) = 15.84%)
c. Ha két különböző nap is megvizsgáljátok 12-12 ember vásárlási szokását, mi a valószínűsége, hogy az egyik nap pontosan 6 ember, míg a másik nap csak 2 (pontosan 2) ember választja a szalonnást? (E: P(x=6 és x=2) = 0.93%)

9. A Startup+ program keretében vállalkozó palántákat képeznek különböző szervezetek, a képzési program végén pedig néhány tervet finanszíroznak is, egy pályázatos kiválasztási folyamatot következően. Egy véletlenszerűen kiválasztott jelentkező üzleti terve 10% valószínűséggel kap finanszírozást a folyamat végén. Az egyetemen is indul egy 15 fős képzési csoport a nyáron.

a. A 15 jelentkező közül várhatóan hány üzleti terv kap finanszírozást? Határozzátok meg a tömegfüggvényét, annak a valószínűségi változónak, hogy az adott jelentkező tervét finanszírozzák!
   (E: E(X) = 1.5; $P(X = k) = C_{15}^{k} * 0.1^{k} * 0.9^{(15-k)}$)
b. Mi a valószínűsége annak, hogy kevesebb mint 3 (legtöbb 2) jelentkező kap finanszírozást a projekt végén? (E: P(X<3) = 81.59%)
c. Ha két különböző képzési csoportban is vizsgáljátok a finanszírozott projektek számát (az egyik csoportban 15 fő volt, a másikban 12), mi annak a valószínűsége, hogy mindkét csoportban pontosan 2 (összesen tehát 4) üzleti terv kap finanszírozást? (E: P((X=2)és(Y=2)) = 6.14%)

10. A TFI nevű magániskola speciális képzési rendszerben oktatja a diákokat. Az oktatási profil arra sarkalja a hallgatóikat, hogy később jó vállalkozók legyenek. Az elmúlt évek vizsgálata alapján 75% a valószínsége annak, hogy egy diákjuk később vállalkozó legyen. A jelenlegi végzős évfolyamon 18-an tanulnak.

a. A végzős évfolyamból várhatóan hányan lesznek vállalkozók? Határozzuk meg a vállalkozóvá válás valószínűségének tömegfüggvényét!
   (E: E(x) = 13.5; $P(X = k) = C_{18}^{k} * 0.75^{k} * 0.25^{(18-k)}$)
b. Mi a valószínűsége annak, hogy ebből az évfolyamból csak 10 (pontosan 10) vállalkozó lesz? (E: P(X=10) = 3.8%)
c. Az eddigi „legvállalkozóbb”évfolyamból 15-ön is vállalkozók lettek. Mi a valószínűsége annak, hogy a jelenlegi évfolyamból többen lesznek vállalkozók mint 15? (E: P(X>15) = 13.5%)

11. Egy bankfiók igazgatója megfigyelte az évek során, hogy a megjelenő ügyfelek száma Poisson eloszlású és egy perc alatt átlagban két ügyfél lépik be a bankfiókba. Jelöljük X-szel egy véletlenszerűen kiválasztott perc alatt megjelenő ügyfelek számát.

a. Határozzuk meg az X tömegfüggvényét. (E: $P(x = k) = \frac{2^{k}}{k!} e^{-2}$)
b. Mi annak a valószínűsége, hogy X = 1? (E: P(x=1)=0.27)
c. Mi annak a valószínűsége, hogy X ≥ 2? (E: P(x≥2)=0.59)

12. Egy gyógyszertárban egy ügyfelet átlagosan 3 perc alatt szolgálnak ki, és az ügyfelek kiszolgálásának ideje percekben mérve exponenciális eloszlást követ. Az X valószínűségi változó legyen a kiszolgálási idő percben kifejezve.

a. Határozzuk meg az X valószínűségi változó sűrűségfüggvényét!
   (E: $f(x) = \frac{1}{3}e^{-\frac{x}{3}}, ha\ x \geq 0$)
b. Amikor megérkezünk a gyógyszertárba, van előttünk valaki, mi a valószínűsége annak, hogy több mint 5 percet kell sorba állnom (ha az előttem állót pont akkor kezdik kiszolgálni, amikor megérkezem)? (E: P(X>5)=0.19)
c. Mi a valószínűsége annak, hogy 4, egymás után érkező, ügyfelet is 2 percnél rövidebb idő alatt sikerül kiszolgálni (fejenként 2 percnél, a 4 kiszolgálási idő független egymástól)? (E: P(X<2)=0.49; 4 egymás után következő ügyfél esetén 0.058)

13. A Tibi Mix ABC vegyesüzletben egy vásárlót átlagosan 4 perc alatt szolgálnak ki, és az ügyfelek kiszolgálásának ideje percekben mérve exponenciális eloszlást követ. Az X valószínűségi változó legyen a kiszolgálási idő percben kifejezve.

a. Határozzuk meg az X valószínűségi változó sűrűségfüggvényét!
   (E: $f(x) = \begin{cases} \frac{1}{4}e^{-\frac{1}{4}x}, ha\ x \geq 0 \\ 0, ha\ x < 0 \end{cases}$)
b. Amikor megérkezünk a bótba, van előttünk valaki (aki ezelőtt 1 perccel kezdte meg a kiszolgálási folyamatot), mi a valószínűsége annak, hogy több mint 5 percet kell sorba állnom? (E: P(x>=6) = 22.31%)
c. Mi a valószínűsége annak, hogy 3, egymás után érkező, ügyfelet is 2 percnél rövidebb idő alatt sikerül kiszolgálni (fejenként 2 percnél)? (E: $(P(0<=X<=2))^3$ = 6.09%)

14. A Kurtoseria kürtőskalács bódéban egy vásárlót átlagosan 5 perc alatt szolgálnak ki, és az ügyfelek kiszolgálásának ideje percekben mérve exponenciális eloszlást követ. Az X valószínűségi változó legyen a kiszolgálási idő percben kifejezve.

a. Határozzuk meg az X valószínűségi változó sűrűségfüggvényét!
   (E: $f(x) = \begin{cases} \frac{1}{5}e^{-\frac{1}{5}x}, ha\ x \geq 0 \\ 0, ha\ x < 0 \end{cases}$)
b. Amikor megérkezünk a bódéhoz, van előttünk valaki (aki ezelőtt 3 perccel kezdte meg a kiszolgálási folyamatot), mi a valószínűsége annak, hogy több mint 3 percet kell sorba állnom? (E: P(x>=6) = 30%)
c. Mi a valószínűsége annak, hogy 5, egymás után érkező, ügyfelet is 3 percnél rövidebb idő alatt sikerül kiszolgálni (fejenként 3 percnél)? (E: $(P(0<=X<=3))^5$ = 1.87%)

15. Csíkszeredából szeretnénk Münchenbe utazni vonattal, úgy, hogy Budapesten váltunk. A megfelelő kényelem érdekében végig hálókocsis fülkében szeretnénk utazni. A Csíkszereda-Budapest járaton a maximális fekvőhelyek száma 80, amelyek kihasználtsága normális eloszlást követ X~N(55,100). A Budapest-München járaton a maximális fekvőhelyek száma 50, amelyek kihasználtsága szintén normál eloszlást követ Y~N(36,81).

a. Ha egy baráti társasággal (összesen 10 fő) szeretnénk utazni, de csak az utolsó nap szeretnénk jegyet venni, mi a valószínűsége, hogy tudunk venni ennyi fekvőhelyet a Budapest-Csíkszereda vonalon? (E: P(X<70) = 89%)
b. Mi a valószínűsége annak, hogy a teljes utazás alatt (Csíkszereda-München) hálókocsiban tudunk utazni? (E: P(X<70 és Y<40) = 45%)
c. Mi a valószínűsége annak, hogy a két szakaszra összesen legalább 90 jegyet adjanak el? (E: P(X+Y>90) = 84%)

16. Egy vizsgán három feladat van, amelyekre maximum 3-3 pontot lehet kapni. A megjelenésre jár még egy pont. Vizsga előtt egy véletlenszerűen kiválasztott diák pontszámai a három feladatra normál eloszlású egymástól független valószínűségi változók, amelynek eloszlásai X1 ~ N(2, 0.09), X2 ~ N(1.8, 0.16) és X3 ~ N(1.5, 0.75). Vegyünk egy véletlenszerűen kiválasztott diákot.

a. Mi annak a valószínűsége, hogy a diák az első feladatra legalább 2 pontot kap? (E: P(X1>=2)=0.5)
b. Mi annak a valószínűsége, hogy a diák az első és a második feladatra legalább 2-2 pontot kap? (E: P(X1>=2 és X2>=2)=0.155)
c. Mi annak a valószínűsége, hogy a diák átmegy a vizsgán (vagyis legalább 5-s végső jegyet kap a vizsgán)? (E: P(X1+X2+X3>=4)=0.9)

17. A kedvenc kocsmátokban Aranyka egy kedves aranyos, míg Hisz Téria egy általában búval bélelt pincér hölgy. Mindkettőjük által kapott havi borravaló értéke jól megközelíthető egy nomál eloszlású valószínűségű változóval. (Aranyka X~ N(9, 1); Hisz Téria Y ~ N(4, 9); az értékek 100 lejben vannak kifejezve). A két hölgy borravalója független egymástól.

a. Mi a valószínűsége annak, hogy Aranyka 1,000 (10) lejnél több borravalót kap egy hónapban? (E: P(X>10) = 15.87%)
b. Mi a valószínűsége annak, hogy mindkét pincér fejenként legtöbb 700 (7) lej borravalót kap egy hónapban? (E: P(X<7 és Y<7) = 1.92%)
c. Mi a valószínűsége annak, hogy a két pincér borravalóinak az összege 1,500 és 2,000 lej között lesz? (E: P(15< X+Y < 20) = 25.07%)

18. A kedvenc számítógépes játékotokban Kufli egy ügyes, talpraesett, míg Gángó elégé szerencsétlen karakter. Mindkettőjük által egy játék során összegyűjtött átlagos pontszám jól megközelíthető egy nomál eloszlású valószínűségű változóval. (Kufli X~ N(12, 4); Gángó Y ~ N(7, 9)) A két karakter pontszáma független egymástól.

a. Mi a valószínűsége annak, hogy Kufli 10 pontnál többet szerez egy játékban? (E: P(X>=10) = 84.13%)
b. Mi a valószínűsége annak, hogy mindkét karakter 9 pontnál többet gyűjt egy játék alatt? (E: P(X>=9 és Y>=9) = 23.46%)
c. Mi a valószínűsége annak, hogy a két karakter együttesen (ketten együtt) 15 és 20 pont között gyűjt egy játék során? (E: P(15<= X+Y <= 20) = 47.68%)

19. A sprint triatlon táv 750 méter úszás, 20 kilométer kerékpározás és 5 kilométer futás, tekintsük mindenik részverseny feladat végrehajtási idejét percben egy normál eloszlású valószínűségű változónak (úszás idő – X; kerékpározás idő – Y; futás idő – Z). A különböző szakaszok idejének eloszlása a következő: X ~ N(10, 4) (vagyis a versenyzők várhatóan 10 perc alatt úszák le a távot, az eredmények varianciája 4 perc), Y ~ N(30, 16) és Z ~ N(18, 9). Vegyünk egy véletlenszerűen kiválasztott versenyzőt.

a. Mi annak a valószínűsége, hogy a versenyző 9 percnél kisebb időt úszik az első távon? (E: P(X<9) = 30.85%)
b. Mi annak a valószínűsége, hogy a versenyző az úszást kevesebb, mint 9 perc alatt, a kerékpározást kevesebb, mint 28 perc alatt, míg a futást 20 percnél hosszabb idő alatt teljesíti? (E: P(X≤9 és Y≤28 és Z≥20) = 2.39%)
c. Mi annak a valószínűsége, hogy a versenyző kevesebb, mint 1 óra alatt teljesíti a teljes versenyt? (E: P(X+Y+Z < 60) = 64.43%)

20. Az OW busztársaság autóbuszainak késése normál eloszlást követ, ahol negatív késés azt jelenti, hogy korábban jön. A késés várható értéke 5 perc és a varianciája 4. A különböző járatok késése független egymástól.

a. Ádám otthonról késve indult el, úgy becsüli, hogy futva csak 4 perccel érkezik később a megállóba, mint a pontos indulási idő. Mi a valószínűsége annak, hogy a busz is késik legalább 4 percet és ő eléri a buszt? (E: P(X>4) = 69%)
b. Boti számít arra, hogy a busz késik, ezért ő úgy számol, hogy elér 2 perccel a pontos idő után érkezzen az állomásra. Mennyi esély van arra, hogy elkési a buszt? (E: P(X<2) = 6.7%)
c. Tünde Pálfalváról szeretne Csatószegre utazni a társaság buszaival, és szeret mindig pontosan érkezni a buszmegállóba. Pálfalváról csak 2 járattal tud Csatószegre menni, Csíkszeredában kell váltani. Mi a valószínűsége annak, hogy ha mindkét megállóba pontosan érkezik, a két megállóban összesen kevesebbet kell várjon mint 5 perc? (E: P(X+Y<5) = 3.8%)

Bibliográfia

- ANTALNÉ Takács Éva–HORVÁTH Jenőné dr.–dr. SZUNYOGH Zsuzsanna (2005) *Statisztikai Alapismeretek*. Budapest.
- BIJI, Mircea–BIJI, Maria Elena (2002) *Tratat de statistică*. Bucureşti, Editura Economică.
- HUNYADI László–MUNDRUCZÓ György–VITA László (2001) *Statisztika*. Budapest, Aula Kiadó.
- Gerald KELLER – Brian WARRACK (2003) *Statistics for Management and Economics*. Thomson - Brooks/Cole.
- KERÉKGYÁRTÓ Györgyné–MUNDRUCZÓ György–SUGÁR András (2001) *Statisztikai módszerek és alkalmazásuk a gazdasági, üzleti elemzésekben*. Budapest, Aula Kiadó.
- Daniel PEÑA – Juan ROMO (2003) *Introducción a la estadística para las ciencias sociales*. McGraw-Hill/Interamericana de España.