

**BUDAPESTI MŰSZAKI ÉS
GAZDASÁGTUDOMÁNYI EGYETEM
TERMÉSZETTUDOMÁNYI KAR**

Reimann József
VALÓSZÍNŰSÉGELMÉLET
ÉS
MATEMATIKAI STATISZTIKA



Műegyetemi Kiadó, 2000.

Azonosító: 090978

A Budapesti Műszaki és Gazdaságtudományi Egyetem

Természettudományi Kar

megrendelése alapján kiadja a

Műegyetemi Kiadó

Felelős vezető: Veress János

Terjedelem: 27,3 (A/5) ív

Nyomtatta és kötötte: **Műegyetemi Nyomda**

Felelős vezető: Frigy Ottó

Munkaszám: 244/00

ALAPFOGALMAK

1.1 A véletlen esemény matematikai fogalma

Tárgyalásaink során valamilyen mérés végrehajtását, valamely természeti jelenség megfigyelését vagy egy fizikai kísérlet elvégzését közös névvel kísérletnek fogjuk nevezni. A "kísérlet" szót tehát a szokásosnál tágabb értelemben használjuk. Így pl. kísérletnek nevezzük egy folyó víz-állásának megfigyelését adott helyen és időpontban, bizonyos számú termék megvizsgálását a selejtes darabok számának megállapítása céljából, egy telefonközpontba adott időtartam alatt beérkező hívások megszámlálását, egy játékkocka feldobását stb. Valószínűségszámítási megfontolásaink során csak olyan kísérletekkel foglalkozunk, amelyeknek kimenetelét a számba vehető feltételek nem határozzák meg egyértelműen, hanem többféle eredményük lehetséges. Az ilyen kísérleteket véletlen kísérleteknek nevezzük, a "véletlen" jelzõt azonban a továbbiakban nem fogjuk feltüntetni.

Feltételezzük, hogy a vizsgált kísérleteket többször (elvéleg akárhányszor) megismételhetjük, mert ha ez a feltétel nem teljesül, semmiféle valószínűségi jellegű megállapításnak nincs helye.

A kísérlet kimenetele a maga egészében általában igen bonyolultan leírható objektum, ezért a kísérlet kimenetele rendszerint egy vagy néhány kiválasztott jellemző adata vonatkozólag érdekel bennünket. A kísérlet különböző kimeneteleit tehát az általunk kiválasztott ismérv megkülönböztethető értékei képezik. Pl. egy folyó víz-állásának mérése esetén a kísérletnek annyi lehetséges kimenetele van, ahány különböző víz-állás elképzelhető 0 és egy ésszerűen választott M felső határ között. Annak ellenére, hogy a víz-állást cm-ben mérjük és ennek törtrészeit nem szokás feltüntetni, mégis azt mondjuk, hogy a folyó víz-állásának értéke adott helyen és időben egy $(0, M)$ intervallum tetszőleges pontja, tehát a víz-állásmérésnek mint kísérletnek végtelen sok (kontinuum-sok) kimenetele van. Ilyenkor azt mondjuk, hogy a $(0, M)$ intervallum minden egyes pontja egy elemi kimenetele a kísérletnek vagy más néven elemi esemény. Hasonlóan kontinuum-sok elemi kimenetele van a kísérletnek, minden ún. folytonos fizikai mennyiség (hosszuság, súly, hőmérséklet, vízhozam stb.) mérése során.

Ha kísérletünk abból áll, hogy megfigyeljük adott időponttól kezdve egy telefonközpontban 10 perc alatt befutó hívások számát, akkor ennek a kísérletnek annyi elemi kimenetele van, ahány nemnegatív egész szám, hiszen a befutó hívások száma $0, 1, 2, \dots, k, \dots$ lehet. (A valóságban a befutó hívások száma természetesen csak véges szám lehet, azonban

megkönnyíti a probléma matematikai tárgyalását, ha azt mondjuk, hogy a kísérletnek elvileg megszámlálható sok kimenetele van. Ez nem jelent a valóságtól való nagy elszakadást, mert bizonyos korlátszám fölött érkező hívások számának olyan csekély gyakoriságot fogunk tulajdonítani, hogy a gyakorlati szempontból érdekes eredményeket nem zavarják.) Vannak olyan kísérletek, amelyeknek véges számú elemi kimenetele van. Ha pl. kísérletünk abból áll, hogy a 90 golyót tartalmazó lottó-urnából 5 golyót kiemelünk, akkor a különböző kimenetek, elemi események száma:

$$\binom{90}{5} = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 86}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 43\,949\,268$$

Ha kísérletünk abból áll, hogy 100 db gyártmányt megvizsgálunk abból a szempontból, hogy hány selejtest (valamilyen kritérium alapján) találunk köztük, akkor ennél a kísérletnél 101 elemi eseményünk van, ti. a selejtes darabok száma lehet 0, 1, 2, ..., 100. Kockadobás esetében hat elemi eseményünk van (a dobott pontok száma).

Ha kísérletünk abból áll, hogy egy kőzetdarabot megvizsgálunk abból a szempontból, hogy tartalmaz-e vasat vagy sem, akkor ennél a kísérletnél csak két elemi eseményünk van, ugyanis vagy van vas a kőzetben vagy nincs. A két elemi kimenetelt numerikusan is jellemezhetjük, pl. úgy, hogy ha van benne vas 1-et, ha nincs benne, akkor 0-t írunk.

Röviden szólva valamely véletlen kísérlet összes lehetséges kimeneteleit elemi eseményeknek nevezzük. Az összes elemi események halmazát az elemi események terének vagy röviden eseménytérnek nevezzük. Ezt a halmazt a továbbiakban X -szel fogjuk jelölni. Magukat az elemi eseményeket, az eseménytér pontjait x -szel jelöljük.

Meg kell jegyeznünk, hogy abban, hogy egy kísérlettel kapcsolatban mit tekintünk elemi eseményeknek - és ezek összességének, az eseménytérnek - bizonyos foku önkény van. Önkény van már abban, hogy a véletlen kísérlettel kapcsolatban milyen számunkra érdekes ismérvet, méretet vagy tulajdonságot választunk, amelynek megkülönböztethető - vagy megkülönböztetésre érdemes - értékei alkotják a kísérlet különböző kimeneteleit, az eseményteret. Az eseménytér tehát a valóságnak egyszerűsített modellje, vetülete. A lényeges csak az, hogy ez a modell a gyakorlat számára érdeklődésre számot tartó szempontból helyesen tükrözze vissza a valóságot. A továbbiakban rendszerint egyetlen méretet vagy ismérvet vizsgálunk és kimenetelen, elemi eseményen a vizsgált méretre vonatkozó adatot értjük.

Egy véletlen kísérlettel kapcsolatban az elemi eseményeken kívül más eseményeket is megfogalmazhatunk. Tegyük fel pl., hogy egy folyó jégmentes nagyvízi állására vonatkozólag egy vízmércén hosszú évek során megfigyeléseket végeztek és ezt tapasztalták, hogy a vízállás 4 m és 10 m között változott. A vízállások ingadozásának tanulmányozása során rendszerint nem az érdekel bennünket, hogy milyen gyakran volt a vízállás

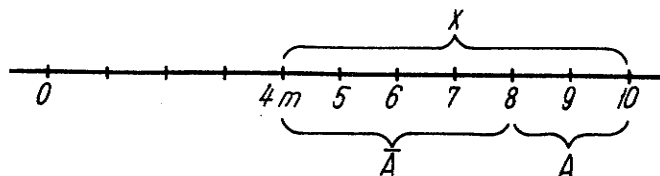
pontosan 637 cm, (lehet, hogy egyszer sem észlelték pontosan ennyit), sokkal érdekesebb pl. „hogy milyen gyakran esett a vízálási érték mondjuk 8 és 10 m közé, milyen gyakran volt a vízálás nagyobb mint 6 m stb.

A fenti kísérletnek megfelelő eseménytér a számegyenes 4 m-től 10 m-ig terjedő intervalluma, s ennek minden pontja egy elemi esemény. Az $x = 637$ cm elemi esemény gyakorisága helyett tehát most pl. mindazon x elemi események halmazának gyakoriságára vagyunk kíváncsiak, amely x pontok a (8, 10) intervallum pontjai, vagyis az

$$A = \{X : 8 \text{ m} \leq x < 10 \text{ m}\}$$

halmaz vagy esemény gyakorisága érdekes számunkra.

Valószínűségszámítási megfontolásainkban az összes elemi események X halmaza az eseménytér, az alaphalmaz, s valahányszor véletlen eseményről beszélünk, az alaphalmaz valamely részhalmazáról van szó. Minden elemi eseménynek egyetlen pontból álló halmaz felel meg.



1. ábra

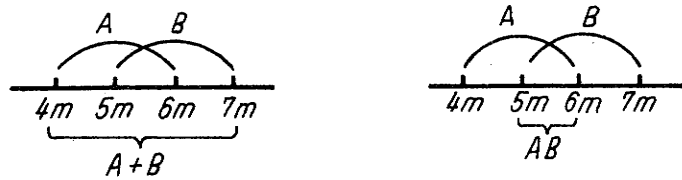
A példánkban felhozott A esemény, vagyis az az esemény, hogy a vízálásra vonatkozó mérési eredmény 8 m és 10 m közé esik - mindannyiszor bekövetkezik, valahányszor olyan elemi esemény következik be, azaz olyan számértéket mérünk, amely a (8 m, 10 m) intervallumba esik. Abban az esetben, ha a mérési eredmény olyan X érték, amely nem esik a (8 m, 10 m) intervallumnak, pl., ha a vízálás $x = 5, 48$ m, akkor az A esemény nem következik be, vagyis az A esemény ellentéte az $\bar{A}$ esemény következik be. Esetünkben:

$$\bar{A} = \{x : 4 \text{ m} \leq x < 8 \text{ m}\}$$

Különösen folytonos mennyiségekkel kapcsolatos eseménytér esetében általában nem egyes elemi események, hanem elemi események bizonyos összességének, halmazának gyakorisága érdekel bennünket. Az elemi események bizonyos összességét, azaz az X eseménytér valamely részhalmazát összetett eseménynek vagy röviden eseménynek nevezzük. Az eseményeket nagy betűvel jelöljük, $A, B, C, \dots, A_1, A_2, \dots, A_n, \dots$ vagy pedig kapcsos zárójelen belül jelezzük, hogy mely elemi eseményről van szó.

Az esemény matematikai szempontból nem más, mint egy halmaz,

az X eseménytér részhalmaza, ezért beszélhetünk események összegéről - amely fogalomnak halmazok összege vagy egyesítése felel meg -, események szorzatáról, ami lényegében két esemény egyidejű bekövetkezését jelenti és a halmazok szorzatának vagy más néven közös részének felel meg. Ha pl. A -val jelöljük azt az eseményt, hogy egy folyó vizállását egy adott helyen megmérve a mért érték a $(4\text{ m}, 6\text{ m})$ intervallumba esik, B -vel pedig azt az eseményt, hogy a mérési eredmény az $(5\text{ m}, 7\text{ m})$ intervallumba esik, akkor az $A + B$ esemény mindannyiszor bekövetkezik, valahányszor a mért vizállás 4 , és 7 m közé esik, vagyis mérési eredményünk eleme a $(4\text{ m}, 7\text{ m})$ intervallumnak, az



2. ábra

$A \cdot B$ esemény pedig mindannyiszor bekövetkezik, valahányszor a mért érték az $(5\text{ m}, 6\text{ m})$ intervallumba esik. Az $A - B = A\bar{B}$ esemény akkor következik be, ha az A esemény bekövetkezik, B pedig nem, azaz ha a mért érték a $(4\text{ m}, 5\text{ m})$ intervallumba esik. Ha a mért érték kívül esik a $(4\text{ m}, 7\text{ m})$ intervallumon, akkor sem az A esemény sem a B esemény nem következik be, vagyis egyidejűleg az $\bar{A}$ és $\bar{B}$ események, azaz az $\bar{A} \cdot \bar{B}$ esemény következik be.

Ha a folyó vizállásán a tényleges vízmélységet értjük (azaz a skála alsó határát 0 -nak választjuk) és K olyan ésszerűen választott felső határ, amelynél nagyobb vizállás sohasem következik be, akkor az az esemény, hogy a mért vizállási érték a $(0, K)$ intervallumba esik, minden mérésnél bekövetkezik. Ezt az eseményt biztos eseménynek nevezzük és a továbbiakban I -vel jelöljük, ami halmazelméleti szempontból az X alaphalmaznak, az eseménytérnek felel meg. Az az esemény, hogy a mért vizállási érték kívül esik a $(0, K)$ intervallumon, lehetetlen esemény, ezt $\emptyset$ -val jelöljük. $\emptyset = \bar{I}$, $I = \emptyset$, azaz a biztos eseményt és a lehetetlen eseményt ellentétes eseménynek tekintjük.

Osszuk fel a példaként említett $X = (0, K)$ intervallumot, az alaphalmazt, a $0 = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = K$ osztópontokkal n részre



3. ábra

és legyen A_i az az esemény, hogy az (x_{i-1}, x_i) intervallumba eső elemi

esemény következett be. Ekkor az $A_1, A_2, \dots, A_n$ események egymást kölcsönösen kizárják és közülük egy kísérlet során egy és csakis egy feltétlenül bekövetkezik. Az $A_1, A_2, \dots, A_n$ eseményekből álló eseményrendszert teljes eseményrendszernek nevezzük.

Az "esemény" és "halmaz" kifejezések alternatív használatának megkönnyítésére az alábbi szótárt állítjuk össze:

A kísérlet összes kimenetelei	$\sim X$	alaphalmaz
az elemi események tere		
kísérlet egy kimenetele	$\sim$	az X alaphalmaz egy x pontja
véletlen esemény	$\sim$	az X alaphalmaz A részhalmaza
$\bar{A}$ az A esemény ellentéte	$\sim A$	komplementer halmaza az $\bar{A} = X - A$ halmaz
I biztos esemény	$\sim I = X$	
$\emptyset$ lehetetlen esemény	$\sim \emptyset$	üres halmaz
az A vagy B esemény következik be	$\sim x \in A + B$	
az A és B események egyidejűleg bekövetkeznek	$\sim x \in A \cdot B$	
az A esemény bekövetkezése maga után vonja a B esemény bekövetkezését	$\sim A \subset B$	(A részhalmaza B -nek)
az A és B események kizárják egymást	$\sim A \cdot B$	(A és B diszjunkt halmazok)

1.2 A valószínűség fogalma

Az olvasó valamilyen formában már bizonyára találkozott a valószínűség fogalmával, sőt valamely véletlen eseményről az esetek nagy részében minden további nélkül meg tudja mondani, hogy nagyon valószínű, vagy nagyon valószínűtlen.

Pl. nagyon valószínű, hogy reggel otthonról elindulva épségben megérkezünk munkahelyünkre. A közlekedési balesetek statisztikája azt bizonyítja, hogy ezt csak nagy valószínűséggel állíthatjuk. Valószínűtlen viszont, hogy (amennyiben lottózunk) a lottó következő húzásán 5-ös találatot érjünk el, sőt négy találat elérése is meglehetősen valószínűtlen, de teljes bizonyosság azon a téren sincs bennünk, mert ha meg lennénk győződve arról, hogy még négy találatunk sem lesz, akkor igen kevesen lottoznánk.

Az élet minden területén az a helyzet, hogy a bennünket érdeklő ese-

mény csak bizonyos valószínűséggel következik be. A különböző események valószínűségét igyekszünk számszerű formában kifejezni, egy mérőskálát bevezetni, mint ahogyan a hőmérsékletet, súlyt stb. mérjük.

Ha valakitől megkérdezzük, hogy mi a valószínűsége, hogy egy szabályos pénzérmét pl. 1 forintot feldobva "fej"-et dobunk, bizonyára azt a választ kapjuk, hogy ez a valószínűség $1/2$. Ennek az állításnak az alapja az a tapasztalat, hogy ha sokszor feldobjuk az érmét, az eseteknek kb. a felében dobunk fejet. Erre vonatkozóan megemlítjük, hogy Buffon 4040 dobás során 2048 esetben kapott "fej"-et, Pearson 24 000 dobásnál a "fej" dobások számát 12 012-nek találta. Ha a "fej"-et eredményező dobások számát elosztjuk az összes dobások számával, akkor a "fej"-dobás relatív gyakoriságát Buffon kísérletsorozatában $\frac{2048}{4040} = 0,5069$ -nek, Pearson kísérletsorozatában $\frac{12\ 012}{24\ 000} = 0,5005$ -nek találjuk. Ugy érezzük, hogy ha még 24 000-nél is több dobást végeznénk, akkor a "fej" és vele együtt az "írás" relatív gyakorisága még jobban megközelítené az $\frac{1}{2}$ -et. Annyit mindenestre állíthatunk, hogy hosszú kísérletsorozatoknál a "fej" illetve "írás" relatív gyakorisága bizonyos stabilitást mutat, az $\frac{1}{2}$ érték körül ingadozik.

Ha nagyszámu hosszú dobássorozatot végzünk egy szabályos játékkockával, akkor pl. az 5-ös szám dobásának relatív gyakorisága csak igen ritkán fog jelentősen eltérni $\frac{1}{6}$ -tól.

Ha egy folyó vizállására vonatkozólag n számu mérési eredményünk van valamely szelvényen és ebből k számu mérési eredmény kisebb mint 700 cm, akkor az $A = \{x < 700\}$ cm esemény relatív gyakorisága ebben a mérési sorozatban $\frac{k}{n}$. (Pl. az 1. táblázatban található mérési adatokra vonatkozólag az $\{A = x < 700\}$ esemény relatív gyakorisága $\frac{35}{70} = \frac{1}{2}$.)

A véletlen események relatív gyakorisága nagyszámu, azonos körülmények között megismételt kísérlet esetén stabilitást mutat, egy meghatározott szám körül ingadozik. Ezt a számot az illető esemény valószínűségének nevezzük.

Valamely esemény valószínűsége objektív mértékszám, amely éppen úgy "megmérhető" mint bizonyos fizikai mennyiségek, mint pl. hőmérséklet, súly stb. Valamely véletlen esemény valószínűségének meghatározása, "megmérése" statisztikai uton, a relatív gyakoriság segítségével történik.

(Megjegyezzük, hogy tulajdonképpen a különböző fizikai mennyiségek megmérése is statisztikai uton történik, hiszen ha nagyobb pontosságot akarunk elérni, akkor általában sok mérést végzünk és számítással határozzuk meg a keresett mennyiséget. Egy objektív mértékszám elvi okoskodással soha nem határozható meg, hanem csakis valamely tapasztalati eljárás segítségével.)

I. táblázat

Duna, Pozsony évi jégmentes nagyvizi állás
1892-1961

	Vizállás (cm)	k_i gyakoriság
$\triangle_1$	450-499	1
$\triangle_2$	500-549	1
$\triangle_3$	550-599	7
$\triangle_4$	600-649	11
$\triangle_5$	650-699	15
$\triangle_6$	700-749	12
$\triangle_7$	750-799	15
$\triangle_8$	800-849	2
$\triangle_9$	850-899	3
$\triangle_{10}$	900-949	1
$\triangle_{11}$	950-999	2
Összesen: 70, $\sum_{i=1}^n k_i = 70.$		

Nagyszámu kísérlet esetén a bennünket érdeklő esemény relatív gyakorisága az illető esemény valószínűségét közelíti meg. Később pontosabban meg fogjuk vizsgálni a közelítés mértékét.

Mivel elég nagyszámu kísérletnél a relatív gyakoriság igen kevéssel tér el a valószínűségtől, a valószínűség tulajdonságait, a különböző események valószínűségei közötti összefüggéseket a relatív gyakoriság tulajdonságaitól származtathatjuk.

A relatív gyakoriság stabilitásának ténye, valamint a relatív gyakoriság egyszerűen belátható tulajdonságai szolgálnak számunkra alapul a valószínűségelmélet kiépítéséhez. Azt a törvényszerűséget, hogy bizonyos kísérletet igen sokszor azonos körülmények között elvégezve, valamely esemény relatív gyakorisága stabilitást mutat, a nagy számok törvényének nevezzük. E törvény matematikai leírására később visszatérünk. (Lásd: 5. 15.)

1.3 Valószínűségszámítási axiómák

Mint már említettük, a valószínűség objektív mértékszám, amely kifejezi, hogy egy kísérletet azonos körülmények között sokszor megismételve a bennünket érdeklő A esemény a kísérleteknek körülbelül mekkora hányadában fog előfordulni. Amikor valamely A eseményre vonatkozólag megadjuk, hogy mekkora az $\delta P(A)$ valószínűsége, (pl. A -val jelölve azt az eseményt, hogy egy kockával 2-est dobunk azt mondjuk, hogy $P(A) = \frac{1}{6}$),

akkor az illető A eseményhez egy $P(A)$ mértékszámot rendelünk. Arra vonatkozólag, hogy mekkora számot rendelünk az A eseményhez az δ valószínűségként, támpontot ad a relatív gyakoriság. Ha a kísérletet n -szer ismételjük meg, s ennek során az A esemény k -szor következett be, azaz az A esemény relatív gyakorisága $g(A) = \frac{k}{n}$, akkor $P(A)$ nem térhet el nagyon $\frac{k}{n}$ -tól. Ha az n számú kísérlet során az A esemény egyszer sem következett be, azaz $k = 0$, akkor $\frac{k}{n} = 0$, míg ha az A esemény minden egyes kísérletnél bekövetkezett, azaz ha $k = n$, akkor $\frac{k}{n} = \frac{n}{n} = 1$. Valamely esemény relatív gyakorisága 0 és 1 között változhat, azaz a $\frac{k}{n}$ relatív gyakoriságra fennáll a

$$0 \leq \frac{k}{n} \leq 1$$

összefüggés.

Ha igen nagyszámú kísérlet során az A esemény egyszer sem következett be, akkor A -t gyakorlatilag lehetetlen eseménynek tekintjük, azaz $A = \emptyset$ és azt mondjuk

$$P(\emptyset) = 0.$$

Ha viszont valamely kísérletet n -szer megismételve az A esemény minden esetben bekövetkezett és n nagy szám, akkor A -t biztos eseménynek tekintjük, $A = I$ és megállapodunk abban, hogy

$$P(I) = 1.$$

A fenti megállapodásokkal máris egy valószínűségi mértékskálát vettünk be egy kísérlettel kapcsolatban számba jöhető véletlen események bekövetkezését illetően.

Legyenek most A és B véletlen események valamely kísérlettel kapcsolatban. Pl. legyen A az az esemény, hogy a Duna vizállására vonatkozó mérési adatsorban a mért x vizállási érték 500 cm és 650 cm közé esik, B pedig az az esemény, hogy a mért x vizállási érték 650 cm és 750 cm közé esik. Az 1. táblázat adatai alapján pl. az

A : $\{ 500 \text{ cm} \leq x < 650 \text{ cm} \}$ esemény relatív gyakorisága $\frac{19}{70}$, a

B : $\{ 650 \text{ cm} \leq x < 750 \text{ cm} \}$ esemény relatív gyakorisága $\frac{27}{70}$,

Az A+B : $\{ 500 \text{ cm} \leq x < 750 \text{ cm} \}$ esemény relatív gyakorisága $\frac{46}{70} = \frac{19}{70} + \frac{27}{70}$.

A : $\{ 500 \text{ cm} \leq x < 650 \text{ cm} \}$ esemény relatív gyakorisága $g(A) = \frac{19}{70}$,

a B : $\{ 650 \text{ cm} \leq x < 750 \text{ cm} \}$ esemény relatív gyakorisága $g(B) = \frac{27}{70}$,

az A+B : $\{ 500 \text{ cm} \leq x < 750 \text{ cm} \}$ esemény relatív gyakorisága:

$$g(A+B) = \frac{46}{70} = \frac{19}{70} + \frac{27}{70} = g(A) + g(B).$$

A relatív gyakoriságra vonatkozólag mindig teljesülnek az alábbi összefüggések:

1. Tetszőleges A esemény $g(A)$ relatív gyakoriságára:

$$0 \leq g(A) \leq 1$$

2. $g(A) = 1$, $g(\emptyset) = 0$.

3. Ha A és B egymást kizáró események, akkor

$$g(A+B) = g(A) + g(B).$$

A relatív gyakoriság ezen tulajdonságait használjuk fel a valószínűségszámítás axiometikus felépítésére.

Valamely A esemény valószínűségén a továbbiakban olyan mértékszámot értünk, amely kielégíti az alábbi axiómákat:

I. axióma: Bármely A esemény $P(A)$ valószínűségére fennáll, hogy

$$0 \leq P(A) \leq 1$$

II. axióma: A biztos esemény valószínűsége: $P(I) = 1$.

III. axióma: Ha A és B egymást kizáró események, azaz $AB = \emptyset$, akkor

$$P(A + B) = P(A) + P(B).$$

A valószínűségelmélet feladata, hogy bizonyos események valószínűségét ismerve vagy feltételezve más, esetleg összetettebb jellegű esemé-

nyek valószínűségeit meghatározza. Megemlítünk most néhány olyan tételt, amelyek axiómarendszerünkben kiindulva könnyen származtathatók és a továbbiakban hasznosnak fognak bizonyulni.

1.4 Valószínűségek közötti összefüggések

1. Tétel: Ha az A esemény valószínűsége $P(A)$, akkor az A ellentétes esemény valószínűsége:

$$P(\bar{A}) = 1 - P(A)$$

Bizonyítás: Mivel $A + \bar{A} = I$ és $A\bar{A} = \emptyset$, a II. és III. axióma alapján

$$P(I) = P(A + \bar{A}) = P(A) + P(\bar{A}) = 1$$

Ebből már adódik a tétel állítása.

A tétel alapján következik, hogy a lehetetlen esemény valószínűsége 0.

Ugyanis:

$$P(\emptyset) = P(\bar{I}) = 1 - P(I) = 0.$$

Példa: Ha egy szabályos játékkockával dobva annak valószínűsége, hogy 3-ast dobjunk, $\frac{1}{6}$ akkor annak valószínűsége, hogy a dobás eredménye nem 3-as lesz: $\frac{5}{6}$.

2. Tétel: Ha az $A_1, A_2, \dots, A_n$ események páronként kizárják egymást, azaz $A_i A_j = \emptyset$ ha $i \neq j$ ($i, j = 1, 2, \dots, n$) akkor annak valószínűsége, hogy az $A_1, A_2, \dots, A_n$ események valamelyike bekövetkezik, ezen események valószínűségeinek összegével egyenlő, azaz:

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Bizonyítás: Ez a tétel a III. axiómából teljes indukcióval következik, $n = 2$ esetén a III. axióma alapján

$$P(A_1 + A_2) = P(A_1) + P(A_2)$$

Tegyük fel, hogy a tétel állítása $n-1$ esemény összegére igaz:

$$P(A_1 + A_2 + \dots + A_{n-1}) = P(A_1) + P(A_2) + \dots + P(A_{n-1}) = \sum_{i=1}^{n-1} P(A_i)$$

$$\begin{aligned} \text{Ekkor } P(A_1 + A_2 + \dots + A_{n-1}) &= P(A_1 + A_2 + \dots + A_{n-1} + A_n) = \\ &= P \sum_{i=1}^{n-1} A_i + A_n = \sum_{i=1}^{n-1} P(A_i) + P(A_n) = \sum_{i=1}^n P(A_i) \end{aligned}$$

Mint az 1.1 pontban említettük, az $A_1, A_2, \dots, A_n$ események összességét teljes eseményrendszernek nevezzük, ha a kísérlet során közülük egy és csakis egy feltétel feltétlenül bekövetkezik, más szóval, ha

$$A_1 + A_2 + \dots + A_n = I \text{ és } A_i A_j = \Phi, \\ \text{ha} \quad i \neq j \quad (i, j = 1, 2, \dots, n)$$

Teljes eseményrendszert alkotó események valószínűségeinek összege 1. Ugyanis:

$$1 = P(I) = P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

Megjegyezzük, hogy valamely kísérlettel kapcsolatban az összes elemi események mindig teljes eseményrendszert alkotnak. Ugyancsak teljes eseményrendszert alkotnak pl. az A és $\bar{A}$ események.

A 2. tétel alapján az alábbi fontos megjegyzést tehetjük:*) bármely esemény valószínűsége az őt alkotó elemi események valószínűségeinek összegével egyenlő. Ha ugyanis az A esemény az $A_1, A_2, \dots, A_k$ elemi események összege, azaz $A = A_1 + A_2 + \dots + A_k$ akkor mivel az elemi események kölcsönösen kizárják egymást,

$$P(A) = P(A_1 + A_2 + \dots + A_k) = \sum_{i=1}^k P(A_i)$$

Tekintsünk most egy olyan kísérletet, amelynek véges n számú kimenete van. Ez azt jelenti, hogy a kísérlet eseménytere n számú elemi

*) Megjegyzésünk véges vagy megszámlálható sok pontból álló eseményterekre vonatkozik!

eseményből áll, jelöljük ezeket $A_1, A_2, \dots, A_n$ -nel. Tegyük fel, hogy a kísérlet minden kimenetele egyformán valószínű, azaz

$$P(A_1) = P(A_2) = \dots = P(A_n) = \frac{1}{n}.$$

Ha most B a kísérlettel kapcsolatos olyan esemény, amely k számú elemi eseményt tartalmaz, akkor megjegyzésünk alapján

$$P(B) = \frac{k}{n}.$$

Érvényes tehát a következő

3. Tétel: Ha egy kísérletnél n elemi esemény következhet be, és mindegyik elemi esemény egyformán valószínű, akkor egy k számú elemi eseményt tartalmazó B esemény valószínűsége

$$P(B) = \frac{k}{n}$$

E tétel alkalmazását mindjárt néhány példával illusztráljuk:

Egy szabályos játékkockát dobálva jelölje a B esemény a páros szám dobását, azaz $B = \{2, 4, 6\}$. Ekkor, mivel összesen 6 elemi eseményünk van, s a B esemény 3 elemi eseményt tartalmaz:

$$P(B) = \frac{3}{6} = \frac{1}{2}.$$

Álljon a kockadobásnál a C esemény abból, hogy a dobás eredménye kisebb, mint 4, azaz $C = \{1, 2, 3\}$, ekkor $P(C) = \frac{3}{6} = \frac{1}{2}$.

Álljon a D esemény abból, hogy a dobás eredménye nagyobb, mint 3, azaz $D = \{4, 5, 6\}$, ekkor $P(D) = \frac{3}{6} = \frac{1}{2}$.

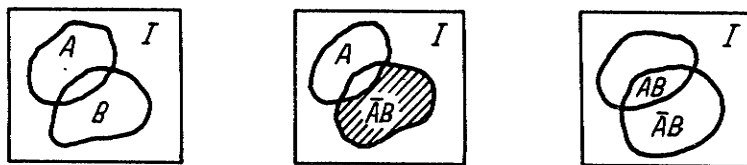
4. Tétel: Ha A és B tetszőleges események, (tehát nem kötjük ki, hogy kizárják egymást) akkor annak valószínűsége, hogy közülük legalább egyik bekövetkezik:

$$P(A + B) = P(A) + P(B) - P(AB)$$

Bizonyítás: Könnyű belátni, hogy

$$A + B = A + \bar{A}B \text{ és } B = AB + \bar{A}B.$$

(Ennek belátását megkönnyítik a következő ábrák is):



4. ábra

Mivel az A és $\bar{A}B$, továbbá AB és $\bar{A}B$ események kizárják egymást, a II. axióma alapján:

$$P(A + B) = P(A) + P(\bar{A}B),$$

$$P(B) = P(AB) + P(\bar{A}B),$$

$$P(A + B) - P(B) = P(A) - P(AB),$$

innen a 4. tétel állítása következik.

1.5 Valószínűségek kiszámítása kombinatorikai módszerekkel

Amikor egy kísérlet minden egyes kimenetele, azaz minden elemi esemény egyforma valószínűségű és véges számú elemi eseményt tartalmaz az eseménytér, a különböző lehetséges események valószínűségének kiszámítása kombinatorikai megfontolások segítségével egyszerűen elvégezhető.

A valószínűségnek az a klasszikus definíciója, hogy "egy esemény valószínűsége egyenlő a kedvező esetek és az összes esetek hányadosával", feltéve, hogy ezek az "esetek" egyenlően valószínűek, speciális esete a 3. tételnek; "eseteken" ugyanis elemi eseményeket kell érteni.

A klasszikus valószínűségszámításban az "eseteket" az "elegendő okok hiánya" elve alapján tekintették egyenlően valószínűnek. Azt mondták, hogy nincs elég okunk feltételezni, hogy az esetek nem egyenlően valószínűek, ami persze helytelen, idealista álláspont. Az egyes elemi események valószínűségeit mindig objektív körülmények, nem pedig az okok nem tudása határozza meg. Általában statisztikai úton ellenőrizni kell, hogy az elemi események egyforma valószínűségűek-e.

Ha valóban az a helyzet, hogy az eseménytér minden pontjához ugyanakkora valószínűség tartozik, akkor véges eseményteret feltételezve, valamely A esemény valószínűségének kiszámítási módja:

$$\frac{\text{az A eseményben foglalt elemi események száma}}{\text{az összes elemi események száma}} = \frac{\text{kedvező esetek száma}}{\text{összes esetek száma}}$$

Amikor egy véges eseménytér minden elemi eseménye egyforma valószínűségű, azt mondhatjuk - mivel az összes elemi esemény valószínűségének összege 1-, hogy az egységnyi valószínűség egyenletesen oszlik el az egyes elemi eseményekre, s ilyenkor diszkrét egyenletes valószínűségeloszlásról beszélünk. A diszkrét egyenletes eloszlás számos gyakorlati problémának reális modellje vagy legalábbis jó közelítéssel alkalmazható, ezért a valószínűség kiszámításának klasszikus egyszerű szabálya sokszor hasznos szolgálatot tesz. A szabály egyszerű, az alkalmazásáról azonban nem mindig mondható ugyanez. Sokszor egy kísérlettel összefüggő elemi események számának meghatározása is számottevő logikai munkát követel. Egy példán érzékeltetjük ezt. Tegyük fel, hogy kísérletünk abból áll, hogy egy, 90 számozott golyót tartalmazó lottóurnából találomra kiemelünk öt golyót. Az összes elemi események száma most annyi, ahányféle módon 90 szám közül öt számot ki tudunk választani (a sorrendre való tekintet nélkül). Hányféleképpen tehetjük ezt meg? Mint korábban említettük, 43 949 268 eset lehetséges! Hogyan lehet ezt kiszámítani? Ilyenféle megszámlálási problémákkal foglalkozik a matematikának egyik érdekes ága az un. kombinatorika, amelynek néhány alapfogalmát röviden ismertetjük.

1. Kombinatorikai alapok

A kombinatorika lényegében véges halmazokra vonatkozó megszámlálási problémák megoldásával foglalkozik. Legyen A véges halmaz, amelynek elemei az 1, 2, ..., n természetes számok:

$$A = \{ 1, 2, \dots, n \}$$

1. Permutációk. Az első probléma, amelyet az A halmazzal kapcsolatban vizsgálunk, hogy hányféle sorrendben tudjuk felírni az 1, 2, ..., n számokat. Legyen pl. $n = 3$, vagyis az A halmaz mindössze három elemet tartalmaz:

$$A = \{ 1, 2, 3 \}$$

Hányféle sorrendben tudjuk felírni az 1, 2, és 3 számjegyeket?

Az összes lehetséges sorrendeket a következő táblázat tartalmazza:

1	2	3
1	3	2
2	1	3
2	3	1
3	1	2
3	2	1

Látható, hogy a felírt sorrendek mindegyike különböző, és összesen hatféle sorrend lehetséges. Mindegyik sorrendet az 1, 2, és 3 számok egy permutációjának nevezzük. Hogyan lehetne előre megmondani három elem összes permutációinak számát az összes sorrend felírása nélkül?

Nyilvánvaló előttünk, hogy két elem, mondjuk az 1-es és 2-es számjegy összesen kétféle sorrendben írható fel:

1	2
2	1

Két elem permutációinak száma tehát: $P_2 = 2$

Ha most a három elem sorrendjeit tartalmazó táblázatot megfigyeljük, akkor láthatjuk, hogy az 1, 2, és 3 számjegyek mindegyike annyiszor szerepel az első helyen, ahányféle módon a mögötte álló másik két számjegy egymás között cserélgethető, azaz három elem összes permutációinak száma:

$$P_3 = 3 \cdot P_2 = 3 \cdot 2 = 3 \cdot 2 \cdot 1 = 6$$

Három elem összes permutációinak számát tehát aképpen határozhatjuk meg, hogy egytől háromig minden számot összeszorozunk. Az 1. 2. 3 szorzat jelölésére a $3!$ szimbólumot használjuk, és három faktoriálisnak olvassuk.

Tehát: $P_3 = 1 \cdot 2 \cdot 3 = 3! = 6.$

Ha az 1, 2, 3, 4 számok összes permutációinak számát kell meghatározunk, akkor elegendő meggondolnunk, hogy e számok mindegyike annyiszor szerepelhet első helyen, ahányféle módon a többi három szám egymás között permutálható cserélgethető, azaz:

$$P_4 = 4 \cdot P_3 = 4 \cdot 3 \cdot 2 \cdot 1 = 4! = 24.$$

Ha pl. egy urnában négy számozott golyó van: 1, 2, 3, 4, majd a golyókat egymás után kihuzzuk és sorba rakjuk a kihuzás sorrendjében, 24 különböző sorrendet nyerhetünk.

Általában, ha az 1, 2, ..., n számok összes permutációit akarjuk meghatározni, akkor hasonlóan gondolkodunk. Az első helyre a fenti számok bármelyikét írhatjuk, mégpedig annyiszor, ahányféle módon a többi $n-1$ elem egymás között permutálható, mert annyiszor az előzőektől különböző sorrendet kapunk.

Tehát:

$$P_n = n \cdot P_{n-1}$$

Analóg gondolatmenettel:

$$P_{n-1} = (n-1) P_{n-2} = (n-1) (n-2) P_{n-3} = \dots =$$

$$= (n-1)(n-2)\dots 3 \cdot 2 \cdot 1.$$

Ezek alapján:

$$P_n = n(n-1)\dots 3 \cdot 2 \cdot 1 = n!$$

Tehát n elem összes permutációjának száma $n!$, amit úgy számítottunk ki, hogy 1-től n -ig az összes számot összeszorozzuk. Amikor pl. egy futóversenyen 8 futó indul, akkor az összes lehetséges befutási sorrendek száma: $P_8 = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 8! = 40\,320$. A faktoriálisok számértékét az 1-50-ig terjedő számokra a 19. oldalon levő táblázat tartalmazza.

Látható tehát, hogy a faktoriálisok értéke az n növekedésével meglepően gyorsan nő. Ha az n értéke nagy, az $n!$ értéke rendkívül nagy, olyanira, hogy pontos értéke általában nem is tulságosan érdekes számunkra, inkább csak a nagyságrendje. Felmerült ezért annak szükségessége, hogyan lehet az $n!$ értékét jó közelítéssel megbecsülni.

Igen jó becslés érhető el az ún. Stirling-formula segítségével:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

(ahol $e = 2,718281\dots$ a természetes logaritmusrendszer alapszáma).

Pl. $n = 10$ esetén $10! = 3\,628\,800$, míg a Stirling-formulával nyert becslés értéke $3\,598\,600$, a hiba $0,8\%$. Minél nagyobb az n értéke, annál kisebb a relatív hiba.

2. Ismétléses permutációk. Tegyük fel, hogy egy urnában n számozott golyó van, de nem mindegyik visel különböző számot. Legyen k_1 számú 1-es, k_2 számú 2-es, $\dots$, k_r számú r -es számjeggyel megjelölt golyó, és $k_1 + k_2 + \dots + k_r = n$.

Egymásután kihuzzuk a golyókat visszatevés nélkül. Kérdés, hányféle különböző sorrend lehetséges?

Ha minden golyó különböző lenne, akkor az összes sorrendek száma $n!$. Most azonban nem tudjuk megkülönböztetni azokat a sorrendeket, amikor azonos számmal ellátott golyókat egymás között cserélünk fel. Az összes megkülönböztethető sorrendek száma tehát nyilván kevesebb lesz, mint ha mind az n golyóra különböző számot írnánk. Az 1-es jegyű golyók $k_1!$ -féle módon cserélgethetők egymás között, a 2-es jegyű golyók $k_2!$ -féle módon stb.

FAKTORIÁLISOK TÁBLÁZATA

n	nl
1	1
2	2
3	6
4	24
5	120
6	720
7	5 040
8	40 320
9	362 880
10	3 628 800
11	39 916 800
12	47 900 160 10
13	62 270 208 10 ²
14	87 178 291 10 ³
15	13 076 774 10 ⁵
16	20 922 790 10 ⁶
17	35 568 743 10 ⁷
18	64 023 737 10 ⁸
19	12 164 510 10 ¹⁰
20	24 329 020 10 ¹¹
21	51 090 942 10 ¹²
22	11 240 007 10 ¹⁴
23	25 852 017 10 ¹⁵
24	62 044 840 10 ¹⁶
25	15 511 210 10 ¹⁸

n	nl
26	40 329 146 10 ¹⁹
27	10 888 869 10 ²¹
28	30 488 834 10 ²²
29	88 417 620 10 ²³
30	26 525 286 10 ²⁵
31	82 228 387 10 ²⁶
32	26 313 084 10 ²⁸
33	86 833 176 10 ²⁹
34	29 523 280 10 ³¹
35	10 333 148 10 ³³
36	37 199 333 10 ³⁴
37	13 763 753 10 ³⁶
38	52 302 262 10 ³⁷
39	20 397 882 10 ³⁹
40	81 591 528 10 ⁴⁰
41	33 452 527 10 ⁴²
42	14 050 061 10 ⁴⁴
43	60 415 263 10 ⁴⁵
44	26 582 716 10 ⁴⁷
45	11 962 222 10 ⁴⁹
46	55 026 222 10 ⁵⁰
47	25 862 324 10 ⁵²
48	12 413 916 10 ⁵⁴
49	60 828 186 10 ⁵⁵
50	30 414 093 10 ⁵⁷

(0! = 1. Ez megállapodás, amely később célszerűnek fog mutatkozni.)

Az összes megkülönböztethető sorrendek, ismétléses permutációk száma:

$$P_n^{(ism)} = \frac{n!}{k_1! k_2! \dots k_r!}$$

Ha pl. n golyó közül k számú golyóra 1-et, $n-k$ golyóra 0-t írunk, akkor az összes megkülönböztethető sorrendek száma:

$$\frac{n!}{k!(n-k)!}$$

Az ismétléses permutációra példaként vizsgáljuk meg, hányféle sorrendben írhatjuk fel a MATEMATIKA szó betűit.

Az összes sorrendek száma:

$$\frac{10!}{2! 3! 2! 1! 1! 1!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 6 \cdot 2} =$$

$$= 10 \cdot 9 \cdot 8 \cdot 7 \cdot 5 \cdot 3 \cdot 2 = 151\,200.$$

3. Variációk. Legyen az urnában most is n különböző golyó: 1, 2, ..., n . Hányféle módon tudunk k különböző golyót kihuzni visszatevés nélkül úgy, hogy a kihuzott k számú golyó különböző permutációi különböző eseteknek számítanak?

Első helyre az n golyó bármelyikét huzhatjuk, a második helyre $n-1$ golyó bármelyikét, a harmadik helyre a megmaradt $n-2$ golyó bármelyikét stb., a k -edik helyre a megmaradt $n-k+1$ számú golyó bármelyike kerülhet, tehát az összes esetek száma, amelyet V_k^n -val jelölünk és n elem-ből alkotható k -ad osztályú variációknak nevezünk:

$$V_k^n = n(n-1) \dots (n-k+1).$$

a) Példa: Egy rejtvénypályázaton három különböző díjat sorsolnak ki a helyes megfejtők között. 35 helyes megfejtés érkezik be. Hányféle eredményt hozhat a sorsolás?

$$V_3^{35} = 35 \cdot 34 \cdot 33 = 39\,270.$$

b) Példa: Hány olyan hatjegyű telefonszám képezhető a 0, 1, 2, ..., 9 számjegyekből, amelyeknek minden jegye különböző? Kikötjük, hogy egyik ilyen telefonszám sem kezdődhet 0-val. Oldjuk meg először a feladatot ezen kikötés nélkül

$$V_6^{10} = 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 = 151\,200$$

Ezen számok között 0-val kezdődő annyi van, ahányféle módon a többi 5 helyre 9 különböző szám közül választani tudunk, ez pedig:

$$9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 = 15\,120$$

A keresett megoldás: $151\,200 - 15\,120 = 136\,080$.

4. Ismétléses variációk. Tételezzük fel most is, hogy egy urnában n különböző golyó van: 1, 2, ..., n . k -szor huzunk ki egymás után egy-egy golyót, de minden kihuzás után a számot feljegyezve, a golyót visszatevesszük az urnába. (A statisztikai alkalmazásoknál ez a modellje az ún. visszatevéses mintavételnek.) Hányféle eredményt nyerhetünk?

Mivel minden huzáskor az n számú golyó bármelyikét huzhatjuk, két huzás után $n \cdot n = n^2$, három huzás után $n \cdot n \cdot n = n^3$ stb. a különböző esetek száma, így k huzás után a különböző lehetséges esetek száma: n^k . Visszatevéses huzások esetén természetesen ugyanaz a szám többször is előfordulhat az eredményben elvileg az is lehetséges, hogy mindegyik huzásra ugyanazt a számot huzzuk. Az n elemből k -szor visszatevéssel végzett huzás eredményét ismétléses variációnak nevezzük:

$$V_{k(\text{ism})}^n = n^k.$$

Az ismétléses variációk száma adott n és k esetén természetesen több, mint az ismétlés nélküli variációk száma. De ha az n igen nagy és a k rögzített, akkor ugyyszólván mindegy, hogy visszatevéssel vagy visszatevés nélkül huzunk, ugyanis az összes esetek hányadosa ekkor:

$$\begin{aligned} \frac{V_k^n}{V_{k(\text{ism})}^n} &= \frac{n(n-1) \dots (n-k+1)}{n^k} = \\ &= \frac{n}{n} \cdot \frac{(n-1)}{n} \cdot \frac{(n-2)}{n} \dots \frac{(n-k+1)}{n} = \\ &= 1 \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \rightarrow 1, \quad \text{ha } n \rightarrow \infty \end{aligned}$$

Példaként vizsgáljuk meg, hány totószelvényt kellene kitöltenünk akkor, hogy biztos 13 találatot érjünk el. Egy mérkőzés eredménye 1, 2 vagy x lehet. Ez ugyanaz az eset, mint amikor egy urnában három golyó van: 1, 2 és x jegyekkel ellátva, és 13-szor huzunk egymás után visszatevéssel.

Az összes esetek száma:

$$V_{13(ism)}^3 = 3^{13} = 1\,594\,323.$$

Független kísérleteknél az összes esetek számát az ismétléses variációk segítségével számítjuk ki. Pl. egy forintos érmét 100-szor feldobva, az összes esetek száma: 2^{100} . Egy szabályos játékkockát 4-szer feldobva, az összes esetek száma: 6^4 . Két kockával dobva 24 dobás összes lehetséges eredményeinek száma: 36^{24} .

5. Kombinációk. Legyen most is egy urnában n darab számozott golyó: 1, 2, ..., n . Belenyulunk az urnába és egyszerre kiveszünk k darab golyót. Hányféle eset lehetséges? Két esetet különböztetünk meg, ha csak egyetlen elemben is eltér a kihuzott k -elemű számcsoporthoz. A kombináció csak abban tér el a variációtól, hogy a kihuzott elemek sorrendjére nem vagyunk tekintettel. (Típusos példa erre a lottó, ahol teljesen mindegy számunkra, hogy milyen sorrendben húzták ki az eredményül nyert öt darab golyót.)

Mivel a kihuzott k számú golyó összes lehetséges sorrendjeinek száma $k!$ (azaz $k!$ -féle módon nyerhetjük ugyanazt az eredményt), a kombinációk száma $k!$ -szor kevesebb, mint a variációk száma:

$$C_k^n = \frac{V_k^n}{k!} = \frac{n(n-1) \dots (n-k+1)}{k!} = \frac{n!}{(n-k)!k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

A $C_k^n = \binom{n}{k}$ (olvasd: n alatt a k) szimbólumot az n -elemből alkotott k -ad osztályú kombinációk számának nevezzük.

Példaként vizsgáljuk meg, hogy a 90 számozott golyót tartalmazó lottóurnából hányféle módon tudunk 5 különböző golyót kihuzni! Gyakorlatilag mindegy, hogy a golyókat egymás után visszatevés nélkül emeljük ki, vagy egyszerre huzunk 5 golyót. A golyók jobb összekeveredése miatt huzzuk egymás után a számokat.

Az összes esetek száma most:

$$C_5^{90} = \binom{90}{5} = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 86}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 43\,949\,268.$$

Tehát ennyi szelvény kitöltése eredményezne biztos 5-ös találatot.

Másik alkalmazásként megemlítjük a binomiális tételt; amely arra vonatkozik, hogyan kell egy kéttagú kifejezést akárhányszor emelni.

$$(a+b)^n = (a+b)^{\overset{1}{\cup}} (a+b)^{\overset{2}{\cup}} \dots (a+b)^{\overset{n}{\cup}} .$$

A szorzást úgy végezzük, hogy minden tényezőből kiválasztunk egy tagot, ezeket összeszorozzuk, majd az így nyert szorzatokat összeadjuk. Ha minden tényezőből a b-t választjuk, akkor b^n -t kapjuk. Ilyen kifejezést csak egyféle módon nyerhetünk. Ha k számú tényezőből az a-t választ-

juk, s a többi $n-k$ tényezőből a b-t, akkor $a^k b^{n-k}$ típusu tagot kapunk, s ezt annyiféle módon nyerhetjük, ahányféleképpen n tényező közül k külön-

böző kiválasztható a kiválasztott k számú tényezőből a-t veszünk, vagyis $\binom{n}{k}$ féle módon.

Tehát:

$$\begin{aligned} (a+b)^n &= \sum_{k=0}^n \binom{n}{k} a^k b^{n-k} = \binom{n}{0} a^0 b^n + \binom{n}{1} a b^{n-1} + \dots + \\ &+ \binom{n}{n} a^n b^0 = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k . \end{aligned}$$

Figyeljük meg, hogy a fenti formulában szereplő kombinációs együtthatókban a felső szám mindig n, az alsó szám 0-tól n-ig növekszik. Az a és b számok kitevőinek összege mindig n, és az a kitevője megegyezik a kombinációs együttható alsó számával.

Írjuk fel a binomiális tételt n különböző értékeire!

$$\begin{aligned} (a+b)^0 &= 1 &= \binom{0}{0} \\ (a+b)^1 &= 1 \cdot a + 1 \cdot b &= \binom{1}{0} a + \binom{1}{1} b \\ (a+b)^2 &= 1 \cdot a^2 + 2ab + 1 \cdot b^2 &= \binom{2}{0} a^2 + \binom{2}{1} ab + \binom{2}{2} b^2 \\ &\vdots \\ (a+b)^k &= \binom{k}{0} a^0 b^k + \binom{k}{1} a b^{k-1} + \dots + \binom{k}{k} a^k b^0 . \end{aligned}$$

Ha most csak az együtthatókat tekintjük és azok számértékét kiszámítjuk, a következő elrendezést kapjuk:

$$\begin{array}{ccccccccc}
 & & & & \binom{0}{0} & & & & & & 1 \\
 & & & & & & & & & & & \\
 & & & \binom{1}{0} & & \binom{1}{1} & & & & 1 & & 1 \\
 & & & & & & & & & & & \\
 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & 1 & 2 & 1 \\
 & & & & & & & & & & & \\
 \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & & 1 & 3 & 3 & 1 \\
 & & & & & & & & & & & & \\
 \binom{4}{0} & \binom{4}{1} & \binom{4}{2} & \binom{4}{3} & \binom{4}{4} & & & & & 1 & 4 & 6 & 4 & 1
 \end{array}$$

Ez az ún. Pascal-féle háromszög, amelynek rendkívül egyszerű képzési szabálya van. A szabály különösen könnyen megfigyelhető a jobb oldali háromszögön: minden sor első és utolsó eleme 1, egyébként bármely szám a fölötte álló két szám összege. Ugyanis a kombinációkra érvényes a következő összefüggés:

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

Ezen összefüggés helyességét könnyen beláthatjuk, ha a benne szereplő kombinációkat az formula alapján a faktoriálisok segítségével kifejezzük:

$$\frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k-1)!} = \frac{(n+1)!}{(k+1)!(n-k)!}$$

Ha az egyenlet mindkét oldalát $n!$ -sal osztjuk, kapjuk:

$$\frac{1}{k!(n-k)!} + \frac{1}{(k+1)!(n-k-1)!} = \frac{n+1}{(k+1)!(n-k)!}$$

Szorozzuk most meg az egyenlet mindkét oldalát a jobb oldal nevezőjével: $(k+1)!(n-k)!$ -sal, ekkor a:

$$k + 1 + n - k = n + 1$$

azonosságra jutunk.

Megjegyezzük, hogy az $\binom{n}{k}$ kombináció kiszámítása az n és a k nagy értékeire igen nehézkes, éppen a faktoriálisok nagy értékei miatt. Felmerül tehát a probléma, hogyan lehet a Pascal-háromszög elemeit nagy n és k értékeire megbecsülni. Az esetben, ha az n páros szám, azaz $2m$ alakú, a Stirling-formula felhasználásával egyszerű alakú becslést kapunk a háromszög $2m$ -edik sorának középső (legnagyobb) elemére:

$$\binom{2m}{m} + \frac{(2m)!}{(m!)^2} = \frac{\left(\frac{2m}{e}\right)^{2m} \sqrt{2\pi 2m}}{\left(\frac{m}{e}\right)^{2m} \cdot 2\pi m} = \frac{2^{2m}}{\sqrt{\pi m}}.$$

A valószínűségelméleti alkalmazások során erre a becslésre lesz a legtöbbször szükségünk.

Megemlítjük még, hogy a Pascal-háromszög képzeletbeli középvonára szimmetrikus, vagyis:

$$\binom{n}{k} = \binom{n}{n-k}.$$

Ennek belátása a faktoriálisok segítségével igen egyszerű:

$$\frac{n!}{k!(n-k)!} = \frac{n!}{(n-k)!k!}$$

Rámutatunk még, hogy a Pascal-féle háromszög n -edik sora elemeinek összege: 2^n .

Ugyanis

$$\binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = (1+1)^n.$$

Az 1.4.4 összefüggés alapján könnyen válaszolhatunk arra a kérdésre, hogy mennyi az $X = \{1, 2, \dots, n\}$ véges halmaz összes részhalmazainak száma:

- X -nek $\binom{n}{1}$ db egyelemű részhalmaza van;
- $\binom{n}{2}$ db kételemű részhalmaza;
- $\binom{n}{3}$ db háromelemű részhalmaza;
- $\binom{n}{k}$ db k -elemű részhalmaza stb.

Ha az X részhalmazai közé számítjuk az üres halmazt: O -t és magát az $X = \{1, 2, \dots, n\}$ halmazt is, akkor az X összes részhalmazainak száma:

$$\binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = 2^n.$$

Ismétléses kombinációk. Az ismétléses kombináció fogalma legegyszerűbben a következő elhelyezési problémával kapcsolatban érthető meg.

• | • • • | •

• | • | • • | | • | • • • | | • • • | | • | • •

$$\frac{(n+k-1)!}{k!(n-1)!} = \binom{n+k-1}{k} = \binom{n+k-1}{n-k}.$$
$$\frac{a}{a+b}$$

- 26 -

$$n(n-1) \dots (n-m+1)$$

A keresett valószínűség:

$$\frac{n(n-1) \dots (n-m+1)}{n^m}$$

3. Tegyük fel, hogy n urnánk van egymás mellett elhelyezve és mindegyikben a fehér és b fekete golyó van. Mindegyik urnából kiveszünk egy golyót. Mi a valószínűsége, hogy k fehér és $n-k$ fekete golyót húzunk ki? Jelöljük ezt a valószínűséget P_k -val.

Az összes esetek száma $(a+b)^n$, mivel minden urnából az $a+b$ számú golyó bármelyikét húzhatjuk. Kedvezőek azok az esetek, amelyeknél k számú urnából fehér golyót húztunk és a többi $n-k$ urnából feketét.

Ilyen összetétel $a^k b^{n-k}$ -féle módon jöhet létre, hiszen minden olyan urnából, amelyből fehér golyót húztunk, az a számú fehér golyó bármelyikét húzhatjuk, ugyanígy minden olyan urnából, amelyből feketét húztunk, a b számú fekete golyó bármelyikét húzhatjuk.

Mivel k számú olyan urna, amelyből fehéret húztunk az n urna között $\binom{n}{k}$ -féle módon helyezkedhet el, a kedvező esetek száma:

$$\binom{n}{k} a^k b^{n-k}$$

A keresett P_k valószínűség a klasszikus definíció alapján:

$$P_k = \frac{\binom{n}{k} a^k b^{n-k}}{(a+b)^n} = \binom{n}{k} \left(\frac{a}{a+b}\right)^k \left(\frac{b}{a+b}\right)^{n-k}$$

$\frac{a}{a+b}$ éppen annak valószínűsége, hogy egy húzás eredménye fehér golyó, míg $\frac{b}{a+b}$ annak valószínűsége, hogy egy húzás fekete golyót eredményez.

Bevezetve a

$$p = \frac{a}{a+b}, \quad q = 1 - p = \frac{b}{a+b}$$

jelölést, a keresett valószínűség:

$$(1.5.1) \quad P_k = \binom{n}{k} p^k q^{n-k}$$

Ugyanerre az eredményre jutunk, ha csak egy urnánk van, amelyben a fehér és b fekete golyót tettünk és n -szer húzunk, de úgy, hogy a kihúzott golyót a következő húzás előtt mindig visszatesszük (és a golyókat jól összekverjük).

Tekintsük most a fenti problémának azt a speciális esetet, amikor egy urnában 1 fehér és 1 fekete golyó van, és n -szer huzunk visszatevés-sel.

Mi a valószínűsége, hogy n huzás során k -szor fehéret huzunk? Most mind a fehér, mind a fekete golyó kihuzásának valószínűsége egy huzás során $\frac{1}{2}$. Így

$$P_k = \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = \frac{\binom{n}{k}}{2^n}$$

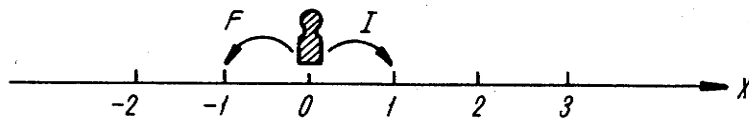
Ugyanennyi a valószínűsége, hogy n -szer feldobva egy szabályos érmét (1 forintot) k -szor írást, $n-k$ -szor fejet dobunk. (Itt az urnából való huzást a forint feldobása helyettesíti, s a fehér golyó huzásának az írást a fekete golyó huzásának a fejet feleltetjük meg. Nyilvánvaló, hogy a két modell teljesen ekvivalens.)

Vizsgáljuk meg kissé közelebbről, hogy ennek a kísérletnek, amely egy érme n -szeri feldobásából áll, mik az elemi eseményei. Tegyük fel, hogy elvégezzük az n érmedobást, s valahányszor írást dobunk, egy I betűt írunk, valahányszor fej eredményez a dobás, F betűt írunk. A kísérlet egy kimenetele n betű eredményez, I és F betűkből álló n elemű sorozatot:

I I F I F F F . . . F I F I. Egy ilyen sorozat most az eseménytér egy "pontja". Az I és F betűkből 2^n számú n -elemű sorozat képezhető, mert minden helyre akár I-t, akár F-et írhatunk. Az összes elemi események száma tehát 2^n , s mindegyik elemi esemény egyenlő valószínűségű, bármelyiknek $\frac{1}{2^n}$ a valószínűsége. Olyan n elemű sorozat az

I és F betűkből, amelyben k számú I betű és $n-k$ számú F betű szerepel nyilván $\binom{n}{k}$ -féle van, mert ennyiféle módon választhatunk ki k helyet az n hely közül az I betűk számára. A többi helyre F betűt írunk. Az az esemény, hogy k -szor dobtunk írást, $\binom{n}{k}$ számú elemi eseményt tartalmaz.

4. Bolyongás a számegyenesen. Tekintsük ismét az 1 forintos érme n -szer egymás után történő feldobásából álló kísérletet. Az egyes dobások eredményeit most is írásban rögzítjük I és F betű segítségével, és ezen felül egy bábút mozgatunk a számegyenes egész koordinátájú pontjain, az origóból indítva a bábút.



5. ábra

Valahányszor írást dobtunk, a bábu jobbra lép egy egységgel, valahányszor fejet dobunk, balra toljuk a bábút egy egységgel. Mi a valószínűsége, hogy n lépés után a bábút az $x = \ell$ pontban találjuk?

Tegyük fel, hogy n lépés során a bábu a -szor lépett jobbra és b -szer balra. Ekkor

$$\begin{aligned} a + b &= n \\ a - b &= \ell \\ \hline a &= \frac{n + \ell}{2} \\ b &= \frac{n - \ell}{2} \end{aligned}$$

A bábút tehát akkor találjuk n lépés után az $x = \ell$ pontban, ha $\frac{n+\ell}{2}$ -szer lépett jobbra, $n - \frac{n+\ell}{2} = \frac{n-\ell}{2}$ -szer pedig balra. Ez az eset $\binom{n}{\frac{n+\ell}{2}}$ -féleképpen következhet be. A keresett valószínűség

$$P = \frac{\binom{n}{\frac{n+\ell}{2}}}{2^n}.$$

Mivel mind a jobbra, mind a balra lépések száma csak egész szám lehet, $\frac{n+\ell}{2}$ pedig csak akkor egész, ha n és ℓ vagy mindkettő páros, vagy mindkettő páratlan, páros számú lépés után a bábu csak páros számon, páratlan számú lépés után csak páratlan számon állhat. Speciálisan pl. az az eset, hogy a bábu visszatér az origóba, azaz kiindulási helyére, csak páros számú lépésben következhet be.

Számítsuk ki annak valószínűségét, hogy a bábu $2n$ számú lépésben visszatér a kiindulási helyére!

Ez az eset nyilván csak akkor következhet be, ha a bábu ugyanannyiszor lépett jobbra, mint balra, vagyis n -szer. A keresett valószínűség:

$$P_0 = \frac{\binom{2n}{n}}{2^{2n}}$$

E valószínűség közelítő értéke nagy n esetén:

$$\frac{\binom{2n}{n}}{2^{2n}} \approx \frac{1}{\sqrt{\pi n}}.$$

Az olvasó előtt bizonyára ismeretes, hogy az ún. Brown-mozgás jelensége abban áll, hogy a gáztérben vagy folyadékokban szuszpendált kicsiny részecskék állandó mozgásban vannak. Ha képzeletben kiválasztunk egy részecskét, amelyet a fenti példában a bábu helyettesített, s figyeljük a részecske elmozdulásainak vetületét, az x-tengely mentén (a részecske mozgását persze nem érmedobás, hanem a többi részecskékkal való véletlen ütközések vezérlik), akkor látható, hogy a fenti bolyongási modell egy fontos természeti jelenség közelítő modellje. A bolyongási modell alkalmazható minden olyan kísérletnél, amelynél egy-egy megfigyelés során csak kétféle kimenetele lehetséges.

5. Mintavétel visszatevés nélkül. Tegyük fel, hogy egy urnában N számú golyó van, M fehér és $N-M$ fekete. Egyszerre kiveszünk az urnából $n < N$ golyót. Milyen a valószínűsége, hogy a kihuzott n golyó között k számú fehér golyó van?

Az összes lehetséges esetek száma most $\binom{N}{n}$, ezek közül kedvező minden olyan eset, amelynél az M számú fehér golyó közül k -számut, az $N-M$ számú fekete golyó közül $n-k$ számut huzzunk. A keresett valószínűség:

$$(1.5.2) \quad P_k = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$$

Általában a fenti modell ipari minőségellenőrzési problémáknál is alkalmazásra talál.

6. Elhelyezési problémák. Kategóriákba sorolás. N dobozban helyezünk el n számú golyót úgy, hogy minden egyes elhelyezkedés egyenlően valószínű. A dobozok 1- N -ig számozva vannak. A golyókra vonatkozólag az alábbi háromféle esetet vizsgáljuk:

a) Tegyük fel, hogy a golyók számozva vannak, azaz két különböző dobozba került golyó felcserélése más elhelyezkedést jelent. Milyen a valószínűsége, hogy egy kiválasztott dobozba k számú golyó kerül?

Az összes elhelyezkedések száma N^n , mert mindegyik golyó bármelyik dobozba kerülhet. (Ugy is képzelhetjük, hogy minden egyes golyónak kisorsolunk egy dobozt.) Kedvezőek azok az esetek, amikor a kiválasztott dobozba k számú golyó kerül, $n-k$ golyó pedig az $N-1$ dobozba kerül tetszőleges eloszlásban. A keresett valószínűség ekkor

$$P_k = \frac{\binom{n}{k} (N-1)^{n-k}}{N^n}$$

b) Tegyük fel, hogy a golyók nincsenek számozva, azaz megkülönböztethetetlenek, úgy két különböző dobozba került golyó felcserélése nem jelent új elhelyezkedést. Ekkor az összes esetek száma $\binom{N+n-1}{N}$ ismétléses kombináció. Ha egy kiválasztott dobozban k számú golyó van, akkor a többi $n-k$ golyó az $N-1$ számú dobozban annyiféle módon helyezkedhet el, amennyi $N-1$ elem $n-k$ -ad osztályú ismétléses kombinációjának száma:

$$\binom{N+n-k-2}{n-k}$$

A keresett valószínűség ekkor:

$$P_k = \frac{\binom{N+n-k-2}{n-k}}{\binom{N+n-1}{n}}$$

c) Tegyük fel, hogy a golyók megkülönböztethetetlenek, de egy dobozba legfeljebb egy golyó juthat, tehát egy dobozban vagy 1 vagy 0 a golyók száma.

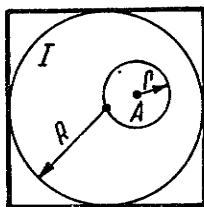
Ekkor az összes esetek száma $\binom{N}{n}$, a kedvező esetek száma $k=0$ esetén $\binom{N-1}{n}$, $k=1$ esetén $\binom{N-1}{n-1}$, tehát:

$$P_0 = \frac{\binom{N-1}{n}}{\binom{N}{n}}, \quad P_1 = \frac{\binom{N-1}{n-1}}{\binom{N}{n}}.$$

1.6 Valószínűségek meghatározása geometriai módszerekkel

A valószínűség fogalmának megvilágításakor vázoltuk, hogy a valószínűség speciális mérték. Az összes elemi események I halmazának, az eseménytérnek $P(I)$ mértékét 1-nek vesszük, s valamely $A \subset I$ esemény $P(A)$ mértéke mindig 0 és 1 közé eső szám. Ha az I véges számú elemi eseményből (pontból) áll, s minden elemi esemény egyforma valószínűségű, akkor valamely $A \subset I$ esemény valószínűsége egyszerűen az A eseményben foglalt elemi események száma osztva az összes elemi események számával.

Bizonyos kísérleteknél hasonló megfontolás alkalmazható a valószínűség meghatározására abban az esetben is, ha az I-eseménytér végtelen sok (még hozzá kontinuum számosságú) elemi eseményből (pontból) áll.



6. ábra

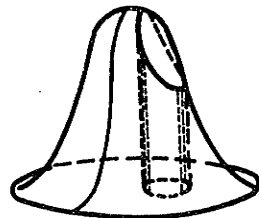
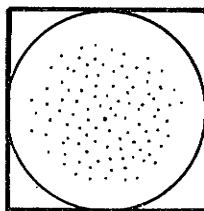
Tekintsük pl. a következő kísérletet:

Egy I céltáblára lövést adunk le. Jelöljük ki a kör alakú céltáblán egy r sugarú A kört.

Mi a valószínűsége, hogy a lövés az r sugarú A körbe talál? A kérdésre egy adott lövés esetében úgy válaszolhatnánk, hogy sok lövést adatunk le vele, s a lövések közül csak azokat vesszük figyelembe, amelyek az I táblára csapódtak (vagyis az I céltábla eltalálását biztos eseménynek tekintjük). Az A körbe eső

találatok és az összes becsapódások számának hányadosa, az A kör eltalálásának relatív gyakorisága, és sok lövés leadása esetén jó közelítőleg megadja az A kör eltalálásának $P(A)$ valószínűségét.

Általában annál nagyobb az A kör eltalálásának valószínűsége, minél nagyobbra választjuk az A kör sugarát, vagy ami ugyanaz, minél nagyobb az A kör területe. A $P(A)$ valószínűség rögzített nagyságú kör esetén általában függ az A kör középpontjának helyzetétől is, pl. jó lövész és pontos fegyver esetén a szóráskép (7. ábra) olyan, hogy a találatok java része az I céltábla középpontja körül sűrűsödik.



7. ábra

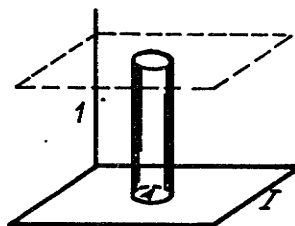
A becsapódási sűrűséget grafikusán a 7. ábra jobb oldalán látható sűrűségi felülettel jellemezhetnénk, amelyet úgy konstruálunk, hogy a felület alatti térfogat az egész I céltáblán egységnyi legyen, s ekkor az A körbe eső találati gyakoriság nagyjából az A körrel, mint alappal vett görbült fedőlapu henger térfogatával egyenlő. (A hengert felülről a sűrűségi felület határolja.)

Bizonyos esetekben (távolságról leadott lövés, kis céltábla stb.) feltehető, hogy a becsapódások egyenletesen helyezkednek el az I céltáblán, amin azt értjük, hogy az I céltáblán egy sűrű négyzethálós beosztást készítünk, akkor bármelyik kis négyzetbe egyforma valószínűséggel esik a találat.

Ez esetben annak valószínűsége, hogy a találat az I tábla egy A résztartományába essék, az A tartomány területével arányos. Az ilyen találati eloszlást nevezzük egyenletes eloszlásnak. Pontosabb fogalmazással, ha a becsapódási sűrűséget egy $f(x, y)$ kétváltozós függvénnyel jellemezzük, amelyet folytonosnak veszünk, tehát geometriailag egy felület reprezentál, és

$$\iint_I f(x, y) \, dx \, dy = 1, \text{ akkor}$$

$$P(A) = \iint_A f(x, y) \, dx \, dy.$$



8. ábra

Egyenletes becsapódási sűrűség esetén $f(x, y) = c$ (állandó), ha $(x, y) \in I$, egyébként

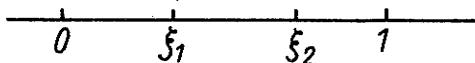
0; és az $\iint_I c \, dx \, dy = 1$ feltételből, ha az I

négyszeg területére egységnyi, akkor $c = 1$. Ez esetben $P(A) = \iint_A dx \, dy$ az A tartomány területével egyenlő ($A \subset I$).

A következőkben egy-két példát mutatunk be az egyenletes eloszlásra. Mindegyik példában feltételezzük, hogy annak valószínűsége, hogy egy véletlen pont az eseménytér egy kiválasztott részhalmazába esik, az illető részhalmaz geometriai méretével (intervallumnál az intervallum hosszát, síkbeli tartománynál annak területével) arányos:

1. Példa

Legyen az eseménytér a számsíkon $(0, 1)$ intervalluma.



9. ábra

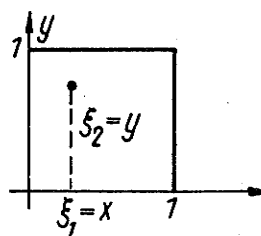
Mi a valószínűsége, hogy a $(0, 1)$ intervallum egy találatra kiválasztott ξ pontja (nevezzük véletlen pontnak) a $(0, x)$ szakaszra esik? Ez a valószínűség:

$$P(0 < \xi < x) = x, \text{ [a } (0, x) \text{ szakasz hossza]}.$$

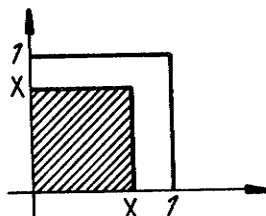
Helyezzünk most el két pontot véletlenszerűen a $(0, 1)$ intervallumra. Legyenek a két pont koordinátái ξ_1 és ξ_2 .

A (ξ_1, ξ_2) pontpár egy síkbeli (x, y) pontot képvisel, amely az egységnyi négyzet tetszőleges pontja lehet:

Tegyük fel, hogy a (ξ_1, ξ_2) véletlen pont az egységnyi négyzeten egyenletes eloszlású.



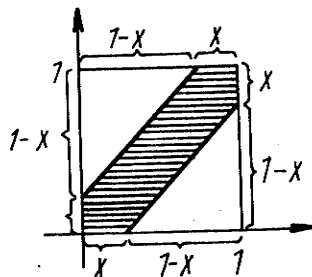
10. ábra



11. ábra

b) Mi a valószínűsége, hogy a két pont egymástól való távolsága kisebb, mint x ?

A $|\xi_1 - \xi_2| < x$ esemény akkor következik be, ha a pont a 11. ábra szerinti bevonalkozott tartományba esik:

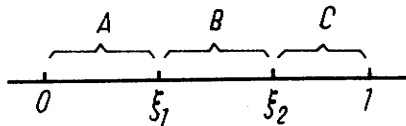


12. ábra

A keresett tartomány területe nyilván:

$$1 - (1 - x)^2 = 2x - x^2.$$

c) Mi a valószínűsége, hogy a ξ_1 és ξ_2 pontokkal találomra három részre osztott egységnyi hosszúságú szakasz A, B, C darabjaiból háromszög szerkeszthető?



13. ábra

Ahhoz, hogy az A, B és C szakaszokból háromszöget lehessen szerkeszteni, bármelyik két szakasz hosszúság összegének nagyobbának kell lennie a harmadik szakasz hosszánál:

$$\begin{aligned} A + B &> C \\ A + C &> B \\ B + C &> A \end{aligned}$$

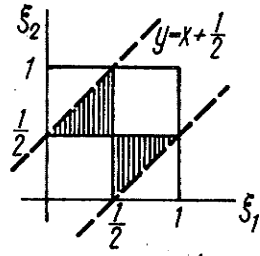
A $\xi_1 < \xi_2$ feltételezésnél ezekből az egyenlőtlenségekből a következők adódnak:

1. $\xi_2 > 1 - \xi_2$;
2. $\xi_1 + (1 - \xi_2) < \xi_2 - \xi_1$;
3. $1 - \xi_1 > \xi_1$.

Az 1. egyenlőtlenségből: $\xi_2 > \frac{1}{2}$;

a 2. egyenlőtlenségből: $\xi_2 < 1 + \frac{1}{2}$;

a 3. egyenlőtlenségből: $\xi_1 < \frac{1}{2}$ adódik.



14. ábra

A $\xi_2 < \xi_1$ esetén viszont a $\xi_1 > \frac{1}{2}$, $\xi_1 < \xi_2 + \frac{1}{2}$, $\xi_2 < \frac{1}{2}$ relációkat nyerjük. (Egyszerűen a ξ_1 és a ξ_2 szerepét felcseréljük.)

A keresett valószínűség tehát az ábrán látható bevonalkázott idom területével egyenlő, ami $\frac{1}{4}$.

2. Példa

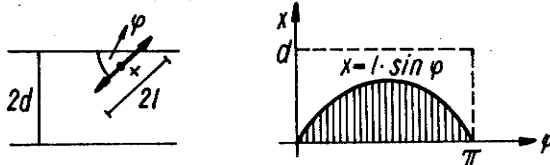
A Buffon-féle tüprobléma. Huzzunk egy papírlapon két párhuzamos egyenest, amelyek egymástól $2d$ távolságra vannak. Dobjunk a lapra egy tűt, s ennek hossza legyen 2ℓ . Feltesszük, hogy $\ell < d$.

Mi a valószínűsége, hogy a tű metszi valamelyik egyenest? A tű helyzetét két adattal jellemezhetjük: középpontjának távolsága a közelebbi egyenestől legyen x , és az egyenessel bezárt szöge legyen φ .

Nyilván: $0 \leq x \leq d$

$0 \leq \varphi < \pi$ lehet.

A tű akkor metszi az egyenest, ha $x \leq \ell \cdot \sin \varphi$.



15. ábra

A keresett valószínűség az ábrán látható bevonalkázott terület osztva $d\pi$ -vel, az összes lehetséges (x, φ) pontok halmazának területével:

$$P = \frac{1}{d\pi} \int_0^{\pi} \ell \cdot \sin \varphi \, d\varphi = \frac{2\ell}{d\pi} .$$

3. Példa

Egy felosztási probléma. Egy R sugaru I céltáblát 9 körrel úgy akarunk 10 részre bontani, hogy mindegyik körre $\frac{1}{10}$ legyen a találati valószínűség. Feltételezve, hogy a találat a tábla bármely pontjába egyforma valószínűséggel esik, és csak azokat a lövéseket számítjuk, amelyek a táblába találnak, hogyan kell az egyes körök $R_1, R_2, \dots, R_9$ sugarait megválasztani.

Jelöljük a k -adik kör sugarát R_k -val ($k = 1, 2, 3, \dots, 9$) akkor teljesülnie kell az

$$\frac{R_k^2 \pi}{R^2 \pi} = \frac{k}{10} \quad \text{összefüggésnek.}$$

Innen:

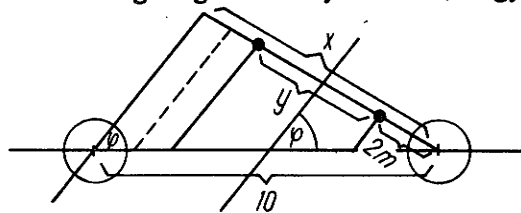
$$R_k^2 = \frac{R^2 k}{10}$$

$$R_k = R \sqrt{\frac{k}{10}} \quad (k = 1, 2, \dots, 9)$$

4. Példa

Gyakorlati szempontból hasznos lehet a következő példa átgondolása. Egy mező a következőképpen van aláaknázva. A mezőn átvezető egyenes mentén egymástól 10 méteres távolságra aknák vannak elhelyezve. Ha egy harckocsi bármely pontja 1 méternél közelebb jut egy akna középpontjához, akkor az akna robban. Ha egy 2 méter szélességű harckocsi az aknák középpontjait összekötő egyenessel φ szöget bezáró egyenes mentén halad át a mezőn, mi a valószínűsége, hogy sértetlenül áthalad?

A következő ábra segítségével könnyű belátni, hogy



16. ábra

a keresett valószínűség $\frac{y}{x}$, hiszen a harckocsi akkor halad át sértetlenül, ha középpontja az áthaladás során az y szakaszra esik. Mivel:

$$\frac{x}{10} = \sin \varphi;$$

$$x = 10 \sin \varphi;$$

$$y = x - 4.$$

Tehát a keresett valószínűség:

$$\frac{y}{x} = \frac{x - 4}{x} = 1 - \frac{4}{10 \sin \varphi} = 1 - \frac{2}{5 \sin \varphi}, \quad \text{ha } \sin \varphi > \frac{2}{5}.$$

Ha $\sin \varphi \leq \frac{2}{5}$, akkor a sértetlen áthaladás valószínűsége 0.

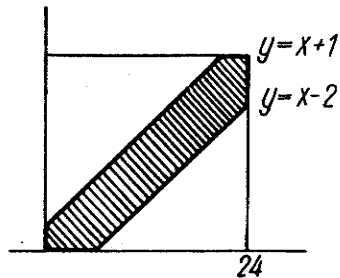
5. Példa

A következő példa abból a szempontból tanulságos, hogy bizonyos időtartamokra vonatkozó feladatot geometriai problémává lehet átfogalmazni, így a megoldást a geometriai szemléletünk is segíti.

Egy kikötőhöz 24 órás időtartamon belül véletlen időpontokban két hajó érkezik. Az előbb érkező hajón rögtön megkezdik a rakodást. A rakodás az egyik hajón egy órát, a másikon két órát vesz igénybe. Ha a második hajó akkor érkezik, amikor az elsőnek befutottra még rakodnak, úgy várnia kell a rakodás befejeztéig. Mi a valószínűsége, hogy valamelyik hajónak várnia kell a rakodásra?

Jelöljük azt az eseményt, hogy valamelyik hajónak várakoznia kell, A-val. A 24 órás időtartam kezdetétől az egyórás rakodóidejű hajó érkezéséig eltelt idő legyen x , a másik hajó érkezéséig eltelt idő legyen y (órákban). A 24 órás időintervallumnak feleltessünk meg egy 24 egységnyi oldalakkal bíró négyzetet, és legyen (x, y) ennek egy pontja. A négyzet területe $T = 24^2 = 576$.

Az A eseménynek megfelelő (x, y) pontok halmazát kell meghatároznunk. Tegyük fel először, hogy $x < y$, azaz a rövidebb rakodási idejű hajó érkezik először. A második hajónak nyilván akkor kell várakoznia, ha $y < x + 1$. Ha a kétórás rakodási idejű hajó érkezik először, azaz $x < y$, az esetben a másik hajónak akkor kell várakoznia, ha



17. ábra

$$x < y + 2, \text{ vagyis } y > x - 2.$$

Az A esemény tehát akkor következik be, ha az (x, y) pont az $y = x + 1$ és az $y = x - 2$ egyenesek közé esik. L. 17. ábra.

Az A esemény bekövetkezése szempontjából "kedvező" bevonalkázott terület:

$$576 - \left(\frac{22^2}{3} + \frac{23^2}{2} \right) = 69,5.$$

Ezek alapján az A esemény valószínűsége:

$$P(A) = \frac{69,5}{576} \approx 0,12.$$

1.7 Feltételes valószínűség és függetlenség

Legyen A és B két esemény, amelyek ugyanazon kísérlet eredményeként bekövetkezhetnek. Tételezzük fel, hogy mindkét esemény pozitív valószínűségű, azaz $P(A) > 0$ és $P(B) > 0$. Az A és B eseményeket egymástól függetlennek nevezzük, ha az A esemény bekövetkezése vagy be nem következése semmiféle hatást nem gyakorol a B esemény bekövetkezésére vagy be nem következésére. Mit jelent ez statisztikailag?

Végezzük el n -szer egymás után a kísérletet és számláljuk meg hányszor következett be az A esemény. Legyen ez a szám k_A , ekkor

$$g(A) = \frac{k_A}{n}$$

az A esemény relatív gyakorisága.

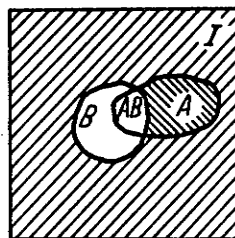
Tekintsük most csak azokat a kísérleteket, amelyek eredményeként a B esemény bekövetkezett, legyen ezek száma k_B és számláljuk meg, hogy ezen kísérletek során hányszor következett be a B eseménnyel együtt az A esemény is, vagyis az AB esemény. Legyen ez a szám k_{AB} .

A $\frac{k_{AB}}{k_B}$ hányadost az A esemény B-re vonatkozó feltételes relatív gyakoriságának nevezzük és $g(A|B)$ -vel jelöljük.

A $g(A|B)$ feltételes relatív gyakoriság lényegében igen egyszerű fogalom. Arról van szó, hogy csak azokat a kísérleteket vesszük számításba, amelyek eredményeként a B esemény bekövetkezett és azt számoljuk, hogy ezen kísérletek során hányszor következett be az A esemény. Ez azt jelenti, hogy az eseményteret leszűkítjük a B halmaz elemi eseményeire.

Ha az A esemény független a B eseménytől, akkor az A esemény relatív gyakorisága és feltételes relatív gyakorisága megegyezik egymással, azaz

$$\frac{k_A}{n} = \frac{k_{AB}}{k_B} = \frac{\frac{k_{AB}}{n}}{\frac{k_B}{n}} \approx \frac{P(AB)}{P(B)}$$



18. ábra

Hosszu kísérletsorozatoknál azt tapasztaljuk, hogy csakugy, mint a relatív gyakoriság, a feltételes relatív gyakoriság is stabilitást mutat.

A $\frac{P(AB)}{P(B)}$ hányadost az A eseménynek a B eseményre vonatkozó feltételes valószínűségének nevezzük és az alábbi módon jelöljük:

$$(1.7.1) \quad P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\text{Innen: (1.7.2)} \quad P(AB) = P(A|B) P(B)$$

Az (1.7.2) összefüggést a valószínűségek szorzási szabályának nevezzük.

Azt mondjuk, hogy az A esemény független a B eseménytől, ha

$$P(A|B) = P(A)$$

$$\text{Ekkor} \quad P(A) = \frac{P(AB)}{P(B)}$$

$$\text{vagyis (1.7.3)} \quad P(AB) = P(A) P(B)$$

Ez azt jelenti, hogy független események szorzatának valószínűsége a valószínűségeik szorzatával egyenlő. Ez a szabály csak független eseményekre igaz.

Analóg módon definiáljuk a B eseménynek A-ra vonatkozó feltételes valószínűségét:

$$(1.7.4) \quad P(B|A) = \frac{P(AB)}{P(A)}$$

Az (1.7.2) és (1.7.4) összefüggésekből adódik, hogy

$$P(A|B) P(B) = P(B|A) P(A)$$

$$\text{azaz} \quad \frac{P(A|B)}{P(B|A)} = \frac{P(A)}{P(B)}$$

Könnyű belátni, hogy ha az A esemény független a B eseménytől, akkor B is független A-tól, azaz

$$P(A|B) = P(A) \quad \text{-ből következik}$$

$$P(B|A) = P(B) ,$$

ekkor ugyanis (1. 7. 4) alapján:

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

Ugyancsak könnyű belátni, hogy ha A független B-től, akkor A független a $\bar{B}$ eseménytől, valamint $\bar{A}$ független B-től és $\bar{A}$ független $\bar{B}$ -től. Ezek bizonyítása:

$$P(A|\bar{B}) = \frac{P(A\bar{B})}{P(\bar{B})} = \frac{P(A) - P(AB)}{1 - P(B)} = \frac{P(A) [1 - P(B)]}{1 - P(B)} = P(A)$$

Itt felhasználtuk a

$$P(A) = P(A|) = P[A(B + \bar{B})] = P(AB) + P(A\bar{B})$$

összefüggést, amelyből

$$P(A\bar{B}) = P(A) - P(AB)$$

adódik.

$$P(\bar{A}|B) = \frac{P(\bar{A}B)}{P(B)} = \frac{P(B) - P(AB)}{P(B)} = \frac{P(B) - P(A)P(B)}{P(B)} = 1 - P(A) = P(\bar{A})$$

Itt alkalmaztuk a

$$P(B) = P(B|) = P[(A + \bar{A})B] = P(AB) + P(\bar{A}B)$$

összefüggést, amelyből

$$P(\bar{A}B) = P(B) - P(AB)$$

adódik.

Az (1. 7. 2) formulával kifejezett szorzási szabály általánosítható n eseményre. Általános szorzási szabály:

Ha $A_1, A_2, \dots, A_n$ tetszőleges események, akkor

$$(1. 7. 5) \quad P(A_1 A_2 \dots A_n) = P(A_n | A_1 A_2 \dots A_{n-1}) P(A_{n-1} | A_1 A_2 \dots A_{n-2}) \dots \\ \dots P(A_2 | A_1) P(A_1)$$

Itt feltételezzük, hogy a vonások mögött álló vagyis feltételként szereplő eseményszorzatok valószínűségei pozitívak.

A tétel bizonyítása triviális.

$$P(A_1 A_2 \dots A_{n-1} A_n) = P(A_n | A_1 A_2 \dots A_{n-1}) P(A_1 A_2 \dots A_{n-1})$$

$$P(A_1 A_2 \dots A_{n-1}) = P(A_{n-1} | A_1 A_2 \dots A_{n-2}) P(A_1 A_2 \dots A_{n-2})$$

$$P(A_1 A_2) = P(A_2 | A_1) P(A_1)$$

Az események függetlenségén kívül használni fogjuk a kísérletek függetlenségének fogalmát, ami egy kicsit általánosabb, mint az események függetlensége. Lényegében két kísérletet is akkor tekintünk függetlennek, ha nincsenek egymásra semmiféle befolyással. Ez azt jelenti, hogy az egyik kísérletnél bekövetkező bármely esemény független a másik kísérletnél bekövetkező bármely eseménytől. Pl. többször ismételt kockadobás, érmedobás stb. független kísérletek sorozatát képezik. Ugyanugy független kísérletsorozatról van szó, ha egy urnában számozott golyók vannak, és találomra kiveszünk egy golyót, feljegyezzük a számát, majd visszateszszük az urnába, s a következő húzás előtt jól összekeverjük a golyókat. Ilyen urna segítségével történhet pl. egy statisztikai sokaság tagjainak véletlenszerű kiválasztása. (Sorsolás.)

Levezetünk most egy fontos összefüggést, amelyet a teljes valószínűség tételének neveznek, s amelyre a továbbiakban többször fogunk hivatkozni.

Alkossanak a $B_1, B_2, \dots, B_n$ események teljes eseményrendszert, azaz legyen $B_1 + B_2 + \dots + B_n = I$ és $B_i B_j = 0$, ha $i \neq j$. Ez azt jelenti, hogy kísérletünk megfigyelésünk során a B_i események közül egy és csakis egy feltétlenül bekövetkezik. Legyen most A egy tetszőleges esemény, amely a $B_1, B_2, \dots, B_n$ események mindegyikétől különbözik. Ekkor:

$$\begin{aligned} P(A) &= P(AI) = P[A(B_1 + B_2 + \dots + B_n)] = \\ &= P(AB_1 + AB_2 + \dots + AB_n) \end{aligned}$$

Itt a zárójelen belüli összeadandók ugyancsak egymást kizáró események, ezért:

$$P(A) = P(AB_1) + P(AB_2) + \dots + P(AB_n)$$



19. ábra

ugyanis ha B_i és B_j diszjunkt halmazok, ak-

kor $AB_1 \subset B_1$, $AB_j \subset B_j$, mint részhalmazok ugyancsak diszjunktak, s mint események kizárják egymást.

Az (1. 7. 2) formula alapján:

$$P(AB_1) = P(A|B_1) P(B_1) , \quad (i = 1, 2, \dots, n)$$

tehát

$$\begin{aligned} (1. 7. 6) \quad P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_n)P(B_n) = \\ &= \sum_{j=1}^n P(A|B_j) P(B_j) \end{aligned}$$

Az (1. 7. 6) összefüggést a teljes valószínűségek tételének nevezzük. Ha ismerjük az A eseménynek mindegyik B_j eseményre vonatkozó feltételes valószínűségét és a B_j események valószínűségét, akkor kereshetjük valamelyik B_j eseménynek az A eseményre vonatkozó feltételes valószínűségét.

Az (1. 7. 4) formula alapján:

$$(1. 7. 7) \quad P(B_1|A) = \frac{P(AB_1)}{P(A)} = \frac{P(A|B_1) P(B_1)}{P(A)}$$

Behelyettesítve a (1. 7. 7) formula nevezőjébe az (1. 7. 6) kifejezést:

$$(1. 7. 8) \quad P(B_1|A) = \frac{P(A|B_1) P(B_1)}{\sum_{j=1}^n P(A|B_j) P(B_j)}$$

Az (1. 7. 8) összefüggést Bayes-féle formulának nevezzük.

VALÓSZÍNŰSÉGI VÁLTOZÓK ÉS VALÓSZÍNŰSÉGELOSZLÁSOK

2.1 A valószínűségi változó és a valószínűségeloszlás fogalma

A véletlen kísérlet fogalmát az I. fejezetben megismertük. Egy kísérlet kimenetelét igyekszünk numerikus formában jellemezni. Legtöbbször maga a kísérlet olyan természetű, hogy eredménye numerikus formában jelentkezik. Ha pl. a kísérlet a talajvízszint megmérése egy adott helyen, akkor a kísérlet eredményeként egy számértéket kapunk, amely számérték, ha a méréseket különböző időpontokban végezzük, igen különböző lehet, a véletlentől függ, hiszen a talajvízszint alakulásában számos véletlen komponens játszik szerepet. A Duna vízszintje adott helyen különböző időpontokban mérve, az esős napok száma májusban (különböző években megfigyelve) mind olyan számadat formájában jelentkezik, amelynek értéke függ a véletlentől.

Az olyan mennyiségeket, amelyek értéke a véletlentől függ, véletlen mennyiségeknek vagy valószínűségi változóknak nevezzük. Azokat a számértékeket, amelyek a kísérlet eredményeként felléphetnek (a kísérlet összes lehetséges számszerű kimeneteleit) a valószínűségi változó lehetséges értékeinek vagy értékkészletének nevezzük. A valószínűségi változókat a görög ξ , η , ζ betűkkel fogjuk jelölni.

Ha a ξ valószínűségi változó egy folyó adott helyen mért vizállását jelöli, akkor ξ értéke egy ésszerűen választott intervallum tetszőleges pontja lehet, ξ értékkészlete tehát egy intervallum kontinuumnyi sok pontja.

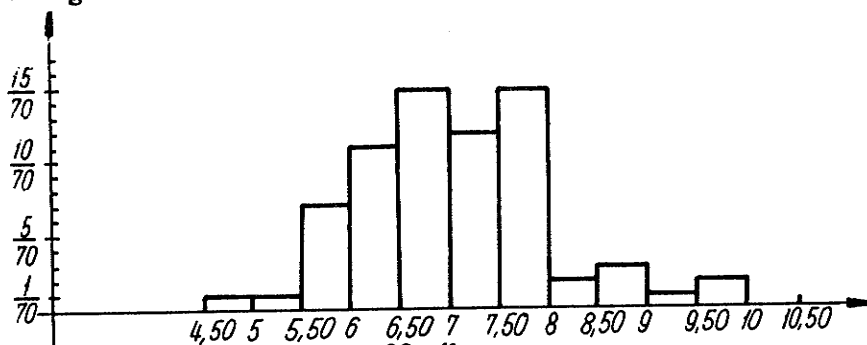
Ha az η valószínűségi változó a májusi csapadékos napok számát jelöli, akkor η lehetséges értékei a 0, 1, 2, ..., 31 számok.

Ez a két példa is mutatja, hogy a valószínűségi változók különböző típusúak lehetnek. Az olyan valószínűségi változók, amelyeknek értéke a számegyenes egy intervallumának (vagy akár az egész számegyenesnek) tetszőleges pontjába eshet, folytonos valószínűségi változónak nevezzük. A mérési eredmények általában folytonos valószínűségi változók. Az olyan valószínűségi változót, amelynek csak véges sok vagy megszámlálható sok értéke lehet diszkrét valószínűségi változónak nevezzük. A gyakorlatban szerepet játszó diszkrét valószínűségi változók legtöbbször értékei csak nemnegatív egész számok lehetnek, ezeket egész értékű valószínűségi változóknak szokás nevezni. Előfordulnak olyan valószínűségi változók is, amelyek sem a diszkrét sem a folytonos eloszlásuak családjába nem sorolhatók, ezeket kevert eloszlásuaknak nevezzük. A későbbiekben mutatunk példát kevert eloszlásu valószínűségi változóra.

A valószínűségi változóval, mint a kísérlet kimenetelének numerikus jellemzőjével kapcsolatban elsősorban az érdekel bennünket, hogy értékei milyen sűrűséggel esnek a számegyenes különböző intervallumaiba (diszkrét eloszlás esetén adott pontjaiba). Amikor pl. azt kérdezzük, mi a valószínűsége, hogy egy folyó vizállása egy adott hónapban 650 és 700 cm közé esik, akkor lényegében az iránt érdeklődünk, hogy ha igen sok éven át ugyanazon hónapban megmérjük a folyó vizállását, akkor az észlelt értékek hány százaléka esik a (650 cm, 700 cm) intervallumba.

Ha ezt a százalékos megoszlást a számegyenes tetszőleges intervallumára ismerjük, akkor azt mondjuk, ismerjük a szóban forgó valószínűségi változó eloszlását. Az 1. táblázat a Duna pozsonyi vizállására vonatkozó ξ valószínűségi változó értékeinek k_1 gyakoriságot tartalmazza.

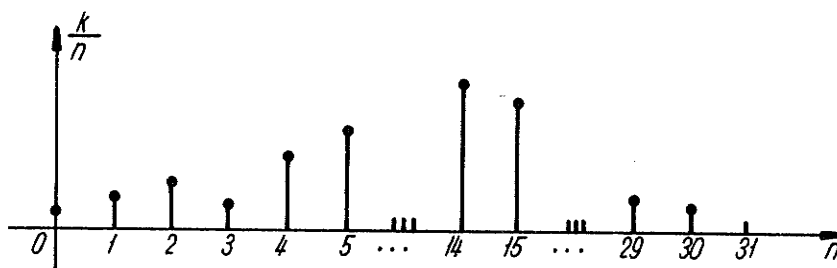
Ha az egyes intervallumokra vonatkozólag kiszámítjuk a $\frac{k_1}{70}$, $\frac{k_2}{70}$, ..., $\frac{k_{11}}{70}$ relatív gyakoriságokat és az egyes Δ_1 intervallumokra $\frac{k_1}{70}$ magasságú téglalapokat rajzolunk, akkor közelítő szemléletes képet nyerünk ξ valószínűségeloszlásáról.



20. ábra

Az így kapott grafikont hisztogramnak nevezzük. Abban az esetben, ha a ξ valószínűségi változó diszkrét eloszlású, pl. ξ a májusi csapadékos napok számát jelenti, akkor ξ valószínűségeloszlásának szemléltetésére eljárhatunk úgy, hogy a $\xi = 0, 1, 2, \dots, 31$ értéket feltüntetjük az abszcisszatengelyen és sokévi megfigyelési adatokból kiszámítjuk ezen számok relatív gyakoriságát, majd a relatív gyakoriságokkal arányos hosszúságú ordinátákat rajzolunk úgy, hogy az ordináták hosszúságainak összege 1 legyen.

Az így kapott ábrát a relatív gyakoriságok diagramjának nevezzük. Mind a hisztogram, mind a diagram csak közelítő képet ad az eloszlásról. Más megfigyelési adatsor, több megfigyelés vagy szűkebb osztályközök esetében más-más ábrát kapunk. Ennek ellenére első közelítés céljából igen hasznos az ilyen jellegű ábrázolás. Egy ξ valószínűségi változó valószínűségeloszlását akkor tekintjük ismertnek, ha a számegyenes tetsző-



21. ábra

leges (a, b) intervallumára vonatkozólag meg tudjuk mondani mi a valószínűsége, hogy ξ értéke az (a, b) intervallumba esik, jelöljük ezt a valószínűséget $P(a \leq \xi < b)$ -vel. Ehhez elegendő, ha tudjuk, hogy a számegyenes tetszőleges x pontjára vonatkozólag tudjuk, mi a valószínűsége, hogy ξ értéke kisebb mint x . Ezt a valószínűséget $P(\xi < x)$ -szel jelöljük. A $P(\xi < x)$ valószínűség az x változó függvénye, jelöljük ezt a függvényt $F(x)$ -szel. Az $F(x)$ függvényt a valószínűségi változó eloszlásfüggvényének nevezzük. Az $F(x)$ eloszlásfüggvény értéke tetszőleges valós x esetén megadja annak valószínűségét, hogy a kísérlet végrehajtása során ξ megfigyelt értékét x -nél kisebbnek találjuk.

Vegyük szemügyre az $F(x)$ eloszlásfüggvény tulajdonságait.

1. Az $F(x)$ eloszlásfüggvény értéke mindig 0 és 1 közé esik (vagy ezen értékek egyike) hiszen $F(x)$ valószínűséget fejez ki, a $\{\xi < x\}$ esemény valószínűségét:

$$(2.1.1) \quad 0 \leq F(x) \leq 1$$

2. Az $F(x)$ eloszlásfüggvény az x mennyiség monoton nem csökkenő függvénye, azaz $b > a$ esetén $F(b) \geq F(a)$. E tulajdonság teljesülését könnyű belátni, mivel $b > a$ esetén a $\{\xi < b\}$ esemény mindig bekövetkezik, ha a $\{\xi < a\}$ esemény bekövetkezik, de a $\{\xi < b\}$ esemény még akkor is bekövetkezik, ha $a \leq \xi < b$ tehát:

$$\{\xi < b\} = \{\xi < a\} + \{a \leq \xi < b\}$$

Itt a jobb oldalon egymást kizáró események állnak, ezért a III. axióma alapján

$$P(\xi < b) = P(\xi < a) + P(a \leq \xi < b)$$

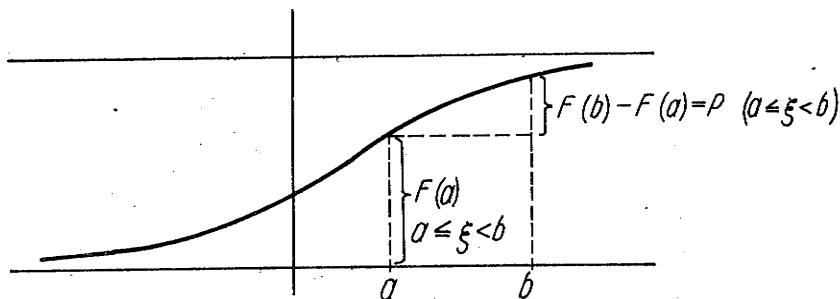
azaz

$$F(b) = F(a) + P(a \leq \xi < b)$$

Az utóbbi összefüggésből:

$$(2.1.2) \quad P(a \leq \xi < b) = F(b) - F(a).$$

Ez az összefüggés mutatja, hogy $F(x)$ értékeinek ismeretében tetszőleges (a, b) intervallum esetében meg tudjuk mondani annak valószínűségét, hogy ξ értéke az (a, b) intervallumba esik:



22. ábra

Mutatunk egy példát, amely jól érzékelteti, mennyire természetes tulajdonsága az $F(x)$ eloszlásfüggvénynek a monoton növekvő volta. Tegyük fel, hogy kísérletünk egy folyó vízállásának megméréseiből áll adott helyen. A vízállás nagysága legyen a ξ valószínűségi változó. Legyen $a = 5$ m, $b = 7$ m. A $\{\xi < a\}$ esemény akkor következik be, ha a vízállás alacsonyabb mint 5 m. A $\{\xi < 5\}$ esemény maga után vonja a $\{\xi < 7\}$ esemény bekövetkezését is, de ez utóbbi esemény még akkor is bekövetkezik, ha a vízállás 5 m és 7 m közé esik. Könnyű belátni, hogy az $\{5 \leq \xi < 7\}$ esemény gyakorisága a $\{\xi < 7\}$ és $\{\xi < 5\}$ események gyakoriságainak különbségével egyenlő. Természetesen ugyanez igaz az illető események relatív gyakoriságaira is. Pl. az 1. táblázat adatai szerint a $\{\xi < 5\}$ esemény relatív gyakorisága $\frac{1}{70}$, a $\{\xi < 7\}$ esemény relatív gyakorisága $\frac{34}{70}$, a $\{5 \leq \xi < 7\}$ esemény relatív gyakorisága $\frac{35}{70}$, és mivel a relatív gyakoriságokra érvényes összefüggések a valószínűségekre is teljesülnek, fennáll, hogy

$$P(\xi < 7 \text{ m}) = P(\xi < 5 \text{ m}) + P(5 \text{ m} \leq \xi < 7 \text{ m})$$

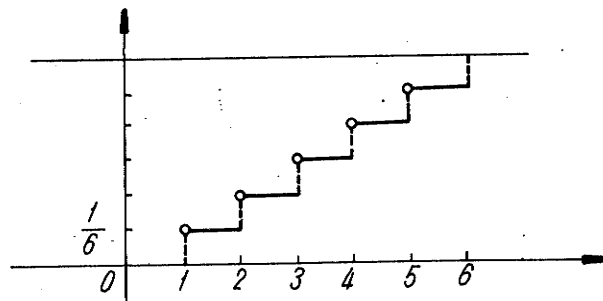
A gyakorlatban csak olyan valószínűségi változókkal találkozunk, amelyek valamilyen véges értéket vesznek fel, a $\xi < +\infty$ eseményt biztos eseménynek, a $\xi < -\infty$ eseményt lehetetlen eseménynek tekintjük, tehát:

$$(2.1.3) \quad \begin{cases} F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = 1 \\ F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0 \end{cases}$$

A (2.1.3) összefüggések egyáltalán nem jelentik azt, hogy x -nek nagy pozitív vagy nagy negatív értéket kell adnunk ahhoz, hogy $F(x)$ értéke 1, ill. 0 legyen. Ha pl. ξ a kockával dobott pontszámot jelölő valószínűségi változó, akkor

$$F(1) = 0 \quad \text{és} \quad F(7) = 1$$

Egyébként a kockadobás pontszámát jelölő ξ valószínűségi változó eloszlásfüggvényét könnyen megkonstruálhatjuk. Ha ξ az 1, 2, 3, 4, 5, 6 értékek mindegyikét $\frac{1}{6}$ valószínűséggel veszi fel, akkor a $F(\xi < x)$ esemény valószínűsége, azaz $F(x)$ értéke annyszor $\frac{1}{6}$, ahány szám esik x -től balra az 1, 2, 3, 4, 5, 6 számok közül. Az eloszlásfüggvényt ekkor az alábbi lépcsős függvény szemlélteti.



23. ábra

Ha a $P(a \leq \xi < b) = F(b) - F(a)$ kifejezésben a -t rögzítjük, b -t pedig a -hoz közelítjük, kapjuk:

$$F(a + 0) - F(a) = P(\xi = a)$$

ha viszont b -t rögzítjük és a -t közelítjük b -hez, akkor az

$$F(b) - F(b-0) = 0$$

eredményre jutunk.

Az utóbbi összefüggés azt mutatja, hogy az eloszlásfüggvény balról folytonos x minden értékére. Az eloszlásfüggvény jobbról nem folytonos minden olyan $x = a$ pontban, amelyre $P(\xi = a) > 0$. Ilyen pontokban az

eloszlásfüggvénynek $P(\xi = a)$ nagyságu ugrása van. A kockadobás esetében az $a = 1, 2, 3, 4, 5, 6$ pontokban $P(\xi = a) = \frac{1}{6}$, ezekben a pontokban, mint a 23. ábrán látható, az eloszlásfüggvénynek $\frac{1}{6}$ nagyságu ugrása van.

Hasonlóan lépcsős szakadós függvény az eloszlásfüggvény, ha egy adott hónap vagy év csapadékos napjainak számát jelenti. Ebben az esetben az ugrások nagysága természetesen általában nem egyforma. Az $F(x)$ eloszlásfüggvény ugrása az $x = x_1$ pontban a $P(\xi = x_1)$ esemény valószínűségével egyenlő.

A valószínűségeloszlást szemléletessé tehetjük az alábbi analógiával. Képzeljük el az x tengelyt végtelenül vékony drótnak, amelyen egy-egynyi mennyiségű tömeget oszlatunk el. A drót különböző szakaszaira jutó tömegmennyiség igen változó lehet. Lehetnek olyan szakaszok, amelyekre egyáltalán nem jut tömeg, egyes izolált pontok pedig pozitív mennyiségű tömegek hordozói lehetnek. Ha $F(x)$ jelenti az x ponttól balra jutó tömeg mennyiségét, akkor $F(x)$ az eloszlásfüggvény tulajdonságaival rendelkezik. Az (a, b) intervallumba jutó tömeg mennyisége $F(b) - F(a)$, ami annak valószínűségével ekvivalens, hogy a ξ valószínűségi változó értéke az (a, b) intervallumba esik, csak tömeg helyett valószínűséget kell gondolnunk. A továbbiakban szemléletesség kedvéért olykor valószínűség-tömegről fogunk beszélni. Az elemi mechanika más fogalmait is fel fogjuk használni valószínűségelméleti fogalmak interpretálására.

2.2 Diszkrét valószínűségi változók

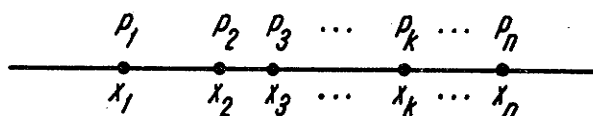
Eddigi tárgyalásaink során több példát láttunk olyan kísérletekre, amelyekhez rendelt ξ valószínűségi változó csak véges sok értéket vehet fel. Pl. az érmedobáshoz rendelt változó csak 0 vagy 1 értéket, a kockadobás pontszámát jelölő ξ változó az 1, 2, ..., 6 értékek valamelyikét, egy adott hónap, pl. május csapadékos napjainak számát jelölő ξ változó a 0, 1, 2, ..., 31 értékek valamelyikét veheti fel. Az ilyen valószínűségi változókat diszkrét változóknak nevezzük. Ha a kísérlet egy telefonközpontba érkező hívások megszámlálásából áll bizonyos időtartamon keresztül, akkor a hívások számát jelölő ξ változó a 0, 1, 2, ... értékeket veheti fel (ha a megfigyelési intervallumot képzeletben nagyon hosszúra nyújtjuk, akkor a hívások számát elvileg végtelennek is vehetjük).

Általában: egy ξ valószínűségi változót diszkrét eloszlásúnak nevezünk, ha lehetséges értékei véges (vagy megszámlálható) halmazt alkotnak.

A diszkrét eloszlás fogalmának megvilágítását igen egyszerűvé teszi a tömegeloszlás analógiája.

Egy ξ valószínűségi változót és a neki megfelelő eloszlást diszkrétnek nevezzük, ha az egységnyi valószínűség-tömeget az x tengely bizonyos izolált pontjaiba koncentráljuk. Ezen felül kikötjük azt a feltételt, hogy bármely véges intervallumba csak véges számú ilyen tömegpont esik.

Legyen $x_1, x_2, \dots$ tömegpontok véges vagy megszámlálható sorozata, $p_1, p_2, \dots$ az illető pontokba elhelyezett tömegek súlyok



24. ábra

Ekkor az $x_1, x_2, \dots$ koordináták a diszkrét eloszlású ξ változó lehetséges értékei, $p_1, p_2, \dots$ a megfelelő valószínűségek analogonjai.

Itt p_i tehát annak a valószínűsége, hogy a ξ változó az x_i értéket veszi fel:

$$P(\xi = x_i) = p_i \quad (i = 1, 2, \dots)$$

Mivel az egész eloszlott tömeg egységnyi

$$\sum_i p_i = 1.$$

A diszkrét eloszlású valószínűségi változó eloszlásfüggvénye:

$$F(x) = \sum_{x_i < x} p_i$$

ahol az összegzést minden olyan i -re kiterjesztjük, amelyre

$$x_i < x$$

Az ilyen típusú eloszlásfüggvény lépcsőfüggvény, amely állandó minden olyan intervallum fölött, amely nem tartalmaz valamely x_i tömegpontot. Az eloszlásfüggvénynek minden x_i pontban p_i nagyságú ugrása van.

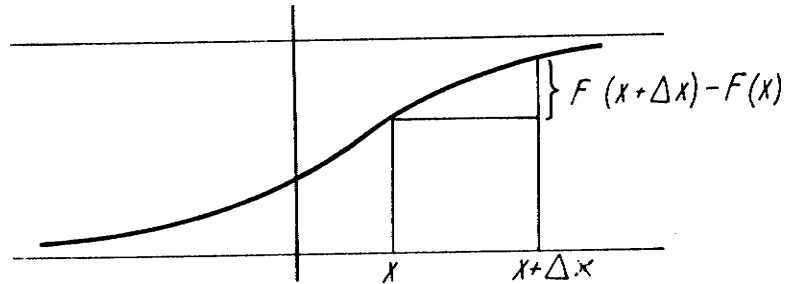
2.3 Folytonos valószínűségi változók

Egy valószínűségi változót akkor nevezünk folytonosnak, ha értékei egy (vagy több) intervallumban bárhova eshetnek. Ilyen valószínűségi változók pl. a mérési eredmények.

Pontosabban szólva a ξ valószínűségi változót folytonos eloszlásúnak nevezük, ha eloszlásfüggvénye $F(x)$ differenciálható, azaz létezik az $F'(x) = f(x)$ függvény és $f(x)$ folytonos.

Az eloszlásfüggvény értékeinek ismeretében (ami rendszerint táblázat formájában rendelkezésünkre áll), tehát minden egyes valós x -re tudjuk, mi a valószínűsége, hogy a ξ valószínűségi változó értéke a számegyenesen az x ponttól balra esik. A gyakorlatban rendszerint az érdekel bennünket, hogy mi a valószínűsége, hogy a szóban forgó valószínűségi változó értéke bizonyos ésszerű határokkal bíró intervallumba esik. Látuk, hogy annak valószínűségét, hogy a ξ valószínűségi változó értéke az (a, b) intervallumba esik, az eloszlásfüggvénynek az intervallum végpontjaiban felvett értékeinek a különbsége: $F(b) - F(a)$ adja meg. Ennek alapján azon esemény valószínűsége, hogy a ξ valószínűségi változó értéke egy kicsiny Δx hosszúságú $[x, x + \Delta x)$ intervallumba esik:

$$P(x \leq \xi < x + \Delta x) = F(x + \Delta x) - F(x)$$



25. ábra

Ha a ξ valószínűségi változó folytonos eloszlású, akkor $F(x)$ differenciálható függvény, azaz létezik az

$$(2.3.1) \quad F'(x) = \lim_{\Delta x \rightarrow 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = f(x)$$

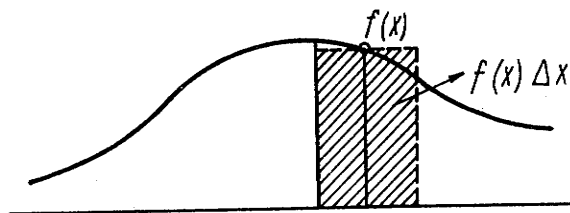
derivált függvény.

A differenciálszámítás középértéktétele alapján

$$P(x \leq \xi < x + \Delta x) \approx f(x) \Delta x$$

Az $f(x)$ függvényt a ξ valószínűségi változó sűrűségfüggvényének nevezzük.

A tömegeloszlás analógiájával élve, a folytonos valószínűségeloszlásnak olyan tömegeloszlás felel meg, amikor az egységnyi tömeget nem egyes diszkrét pontokba koncentráljuk, hanem folytonosan osztjuk el az x -tengelyen vagy annak egy intervallumában, s ekkor az $f(x)$ függvény reprezentálja az egyes pontokba eső tömegsűrűséget, az $f(x) \Delta x$ szorzat pedig a kicsiny Δx intervallumra eső tömeg mennyiségét.



26. ábra

Folytonos eloszlások esetében rendszerint előnyösebb a sűrűségfüggvény-nel dolgozni, különösen elméleti jellegű megfontolásaink során.

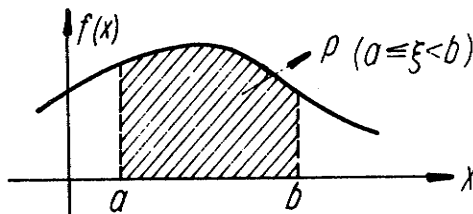
Összevetve a

$$P(a \leq \xi < b) = F(b) - F(a)$$

valamint az $F'(x) = f(x)$ összefüggéseket, könnyű belátni, hogy

$$(2.3.2) \quad f(b) - F(a) = P(a \leq \xi < b) = \int_a^b f(x) dx$$

Tehát annak valószínűségét, hogy a ξ valószínűségi változó értéke az $[a, b)$ intervallumba esik, a sűrűségfüggvénynek az illető intervallumon vett integrálja szolgáltatja.



27. ábra

Figyelembe véve, hogy

$$F(+\infty) = 1, \quad F(-\infty) = 0,$$

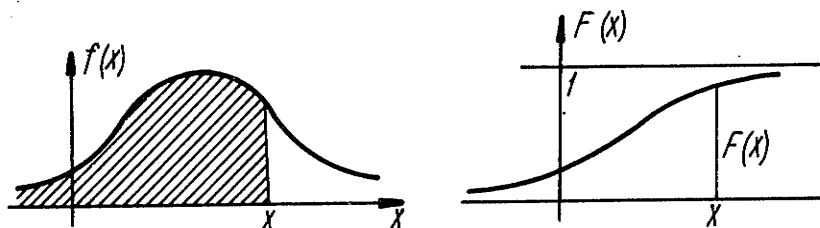
$$(2.3.3) \quad 1 = F(+\infty) - F(-\infty) = \int_{-\infty}^{\infty} f(x) dx$$

relációból látható, hogy a sűrűségfüggvénynek az egész számsíkon vett integrálja 1.

Az

$$(2.3.4) \quad F(x) = F(x) - F(-\infty) = P(-\infty \leq \xi < x) = \int_{-\infty}^x f(t) dt$$

összefüggés mutatja, hogyan nyerjük a sűrűségfüggvényből az eloszlásfüggvényt, ami egyébként a sűrűségfüggvény definíciójából is következik. Az $F(x)$ eloszlásfüggvény tehát a sűrűségfüggvény alatti területet adja a $(-\infty, x)$ intervallum fölött



27/a ábra

E terület mérőszámát olvashatjuk le az $F(x)$ eloszlásfüggvény grafikonjáról, vagy praktikusán $F(x)$ táblázatából.

Példa kevert eloszlásra

Említettük, hogy a gyakorlatban fellépő legfontosabb valószínűség-eloszlások vagy a diszkrét vagy a folytonos eloszlástípusba tartoznak. Előfordul azonban kevert eloszlás is, erre mutatunk most egy példát.

Tegyük fel, hogy egy üres víztározót egy folyó vizével töltünk fel az év egy adott hónapjában. A folyó vízhozama az illető hónapban legyen η m³, ahol η valószínűségi változó, eloszlását folyamatosnak tételezzük fel és legyen η sűrűségfüggvénye $g(x)$, ahol $g(x) = 0$ ha $x < 0$. A tározó kapacitása legyen k m³.

Ha most $\eta < k$, akkor a tározó az adott hónap végén η m³ vizet tartalmaz. Ha $\eta > k$, akkor a tározó a hónap végén k m³ vizet tartalmaz, a többi víz túlfolyik.

Jelöljük most a hónap végén a tározóban levő víz mennyiségét a ξ valószínűségi változóval. Amennyiben az $\eta \leq k$ esemény pozitív valószínűségi, azaz

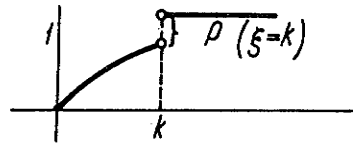
$$\int_k^{\infty} g(x) dx > 0,$$

akkor ξ kevert eloszlású, eloszlásfüggvénye folytonos minden $x < k$ értékre és ilyen x értékekre

$$P(\xi < x) = P(\eta < x) = \int_{-\infty}^x g(t) dt$$

Annak valószínűsége, hogy $\xi = k$, annak valószínűségével egyenlő, hogy a vízhozam eléri vagy meghaladja a k m³-t, azaz

$$\begin{aligned} (*) \quad P(\xi = k) &= P(\eta \leq k) = \\ &= \int_k^{\infty} g(x) dx > 0. \end{aligned}$$



28. ábra

A ξ valószínűségi változó $F(x)$ eloszlásfüggvényének az $x = k$ helyén ugrása van és az ugrás nagysága éppen a (*) alatti mennyiség.

2.4 A többdimenziós valószínűségeloszlás fogalma

Számos esetben valamely kísérlet, megfigyelés kimenetele nem jellemezhető egyetlen számadattal, hanem valamely jelenséggel kapcsolatban két vagy több valószínűségi változót egyidejűleg kell vizsgálnunk. Ha pl. arra vagyunk kíváncsiak, hogyan függ a Duna vízállásától valamely partmenti területen a talajvízszint, akkor a Duna vízállását ξ -vel, a talajvízszintet η -val jelölve vizsgáljuk a (ξ, η) valószínűségi változópár vagy másként nevezve a $\tilde{\xi} = (\xi, \eta)$ véletlen vektor eloszlását, mert ez szolgál kiindulási alapként az összefüggés felderítésére.

Ha egyidejűleg egy folyó vízállását, a lehullott csapadék mennyiségét és a talajvízszintet vesszük szemügyre, akkor beszélhetünk a

$$\tilde{\xi} = (\xi_1, \xi_2, \xi_3)$$

véletlen vektor eloszlásáról, amelyet a ξ_1 , ξ_2 és ξ_3 valószínűségi változók együttes eloszlásának is szokás nevezni.

Ha egy ξ mennyiségre vonatkozólag n számú mérést végzünk, amelyek eredményei a $\xi_1, \xi_2, \dots, \xi_n$ mennyiségek, akkor vizsgálhatjuk

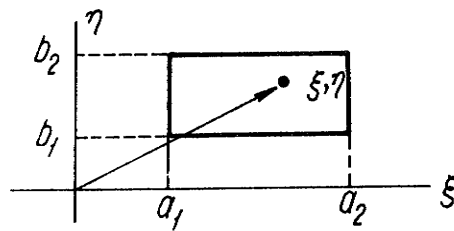
ezen mérési eredmények együttes eloszlását, azaz a

$$\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$$

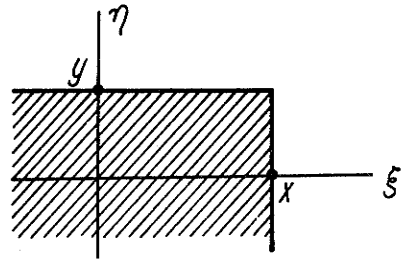
véletlen vektor vagy valószínűségi vektorváltozó eloszlását. Még gyakoribb, hogy a $(\xi_1, \xi_2, \dots, \xi_n)$ véletlen vektor helyett egy $\eta = f(\xi_1, \xi_2, \dots, \xi_n) = f(\vec{\xi})$ valószínűségi változó eloszlása érdekel bennünket, ahol f egy n -változós függvény. A matematikai statisztika legfontosabb területein: a becsléelméletben és a statisztikai hipotézisek vizsgálatánál ilyen típusu véletlen argumentumu függvények eloszlása elsődleges fontosságú.

Foglalkozzunk először a kétdimenziós véletlen vektorok eloszlásával. Legyenek ξ és η valószínűségi változók és tekintsük a (ξ, η) véletlen pont síkbeli eloszlását.

Jelöljük $P(a_1 \leq \xi < a_2, b_1 \leq \eta < b_2)$ -vel annak valószínűségét, hogy a véletlen pont ($a \vec{\xi} = (\xi, \eta)$ véletlen vektor) a fent szereplő számokkal megjelölt téglalapba esik:



29. ábra



30. ábra

Ha ismerjük ezt a valószínűséget minden a_1, a_2, b_1, b_2 szám esetére, akkor azt mondjuk, hogy ismerjük a ξ és η valószínűségi változók együttes eloszlását.

A ξ és η változók együttes eloszlásfüggvényét a

$$(2.4.1) \quad H(x, y) = P(\xi \leq x, \eta \leq y)$$

reláció definiálja.

Könnyű belátni, hogy

$$(2.4.2) \quad \begin{aligned} &P(a_1 \leq \xi < a_2, b_1 \leq \eta < b_2) = \\ &= H(a_2, b_2) - H(a_1, b_2) - H(a_2, b_1) + H(a_1, b_1) \geq 0, \end{aligned}$$

amely összefüggés mutatja, hogy a $H(x, y)$ kétváltozós eloszlásfüggvény ismeretében tetszőleges síkbeli téglalap valószínűségi mértékét meg tudjuk határozni.

Bizonyítás nélkül megemlítjük, hogy $H(x, y)$ mindkét változójának monoton nemcsökkenő függvénye, továbbá

$$H(+\infty, +\infty) = 1, \quad H(-\infty, y) = H(x, -\infty) = 0.$$

A kétdimenziós valószínűségeloszlás matematikai tárgyalása lényegében teljesen analóg a síkbeli tömegeloszlás vizsgálatával, ha egységnyi tömeget képzelünk elosztatva az (x, y) síkon. Egy adott téglalapba eső tömeg mennyisége azon esemény valószínűségének felel meg, hogy a (ξ, η) véletlen pont az illető téglalapba esik. A $H(x, y)$ függvény adott (x_0, y_0) pontbeli értéke most azt a tömegmennyiséget reprezentálja, amely az $x < x_0, y < y_0$ végtelen síknegyedbe esik.

Ha ismerjük a (ξ, η) pár $H(x, y)$ együttes eloszlásfüggvényét ebből könnyen megkaphatjuk a ξ ill. η valószínűségi változók $F(x)$ ill. $G(y)$ eloszlásfüggvényeit.

Ugyanis:

$$(2.4.3) \quad H(x, +\infty) = P(\xi < x, \eta < +\infty) = P(\xi < x) = F(x)$$

$$(2.4.4) \quad H(+\infty, y) = P(\xi < +\infty, \eta < y) = P(\eta < y) = G(y)$$

Az így nyert $F(x)$ ill. $G(y)$ eloszlásfüggvényeket vetület- vagy peremeloszlás-függvényeknek is nevezzük.

Kétdimenziós eloszlásoknál is megkülönböztetünk diszkrét és folytonos eloszlásokat.

Ha mind ξ , mind η diszkrét valószínűségi változó, akkor együttes eloszlásuk diszkrét síkbeli eloszlás. Egyszerűség kedvéért legyenek ξ és η egész értékű valószínűségi változók, azaz lehetséges értékeik legyenek nemnegatív egész számok. A (ξ, η) értékpárok valószínűségeire vezessük be a

$$P(\xi = k, \eta = \ell) = r_{k\ell} \quad \text{jelölést. } (k, \ell = 0, 1, 2, \dots)$$

Természetesen

$$\sum_k \sum_\ell r_{k\ell} = 1$$

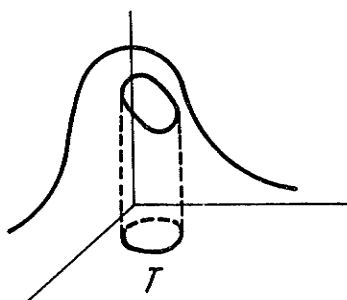
Ha az $r_{k\ell}$ együttes valószínűségeket egyik indexük szerint összegezzük, nyerjük a peremeloszlásokat.

$$(2.4.5) \quad \sum_{\ell} r_{k\ell} = \sum_{\ell} P(\xi = k, \eta = \ell) = P(\xi = k) = p_k,$$

$$(2.4.6) \quad \sum_k r_{k\ell} = \sum_k P(\xi = k, \eta = \ell) = P(\eta = \ell) = q_{\ell}.$$

Ezeket az összefüggéseket a teljes valószínűség tétele alapján egyszerűen beláthatjuk.

Számunkra az alkalmazások szempontjából legfontosabb az az eset, amikor mind ξ mind η folytonos eloszlásúak $f(x)$ ill. $g(y)$ sűrűségfüggvénnyel és a $\zeta = (\xi, \eta)$ véletlen vektor folytonos síkbeli eloszlású. Ez



31. ábra

esetben, a folytonos tömegeloszlás analógiájára az (x, y) pontbeli valószínűség sűrűségét egy kétváltozós $h(x, y)$ függvény segítségével adjuk meg, amelynek geometriai értelme egy kétváltozós felület. Ha létezik ilyen kétváltozós sűrűségfüggvény, amelyre

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h(x, y) dx dy = 1, \quad \text{akkor}$$

annak valószínűségét, hogy a $\zeta = (\xi, \eta)$ véletlen számpár a sík egy adott T tartományába esik, a $h(x, y)$ függvénynek a T tartományra számított kettős integrálja szolgáltatja:

$$P((\xi, \eta) \in T) = \iint_T h(x, y) dx dy$$

Ha a T tartomány a $(\xi < x, \eta < y)$ síknegyed, akkor

$$(2.4.7) \quad P(\xi < x, \eta < y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv = H(x, y),$$

vagyis a kétváltozós eloszlásfüggvényhez jutunk.

Fennáll tehát a következő összefüggés a kétváltozós eloszlásfüggvény és a sűrűségfüggvény között:

$$(2.4.8) \quad h(x, y) = \frac{\partial^2 H(x, y)}{\partial x \partial y}$$

Vagyis a kétváltozós sűrűségfüggvény (ha létezik) a kétváltozós eloszlásfüggvény vegyes parciális deriváltja.

A $h(x, y)$ kétváltozós sűrűségfüggvény ismeretében meghatározhatjuk a § ill. η változók perem eloszlásfüggvényeit, illetve a peremsűrűségeket az alábbi összefüggések alapján:

$$(2.4.9) \quad F(x) = H(x+\infty) = \int_{-\infty}^x \int_{-\infty}^{\infty} f(u, v) \, du \, dv = \int_{-\infty}^{\infty} \left(\int_{-\infty}^x h(u, v) \, du \right) dv$$

$$(2.4.10) \quad G(y) = H(+\infty, y) = \int_{-\infty}^{\infty} \int_{-\infty}^y f(u, v) \, du \, dv = \int_{-\infty}^{\infty} \left(\int_{-\infty}^y h(u, v) \, dv \right) du$$

Ezekből az összefüggésekből differenciálás útján nyerjük a § ill. η változók sűrűségfüggvényeit:

$$(2.4.11) \quad f(x) = F'(x) = \int_{-\infty}^{\infty} h(x, v) \, dv$$

$$(2.4.12) \quad g(y) = G'(y) = \int_{-\infty}^{\infty} h(u, y) \, du$$

2.5 Feltételes eloszlásfüggvény és sűrűségfüggvény

Az eseményekre vonatkozólag értelmeztük a feltételes valószínűséget. Ennek alkalmazásával meghatározhatjuk egy § valószínűségi változó feltételes eloszlásfüggvényét azon feltétel mellett, hogy egy másik η valószínűségi változó adott értéket vesz fel.

Jelöljük A-val a $\{\xi < x\}$ eseményt és B-vel az $\{\eta < y\}$ eseményt, ekkor a feltételes valószínűség definíciója alapján (lásd: 1.7. pont)

$$(2.5.1) \quad P(\xi < x | \eta < y) = P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(\xi < x, \eta < y)}{P(\eta < y)} = \frac{H(x, y)}{G(y)}$$

Ha

$$A: \{\xi < x\}, \quad B: \{y \leq \eta < y + \Delta y\},$$

akkor

$$P(\xi < x | y \leq \eta < y + \Delta y) = \frac{P(\xi < x, y \leq \eta < y + \Delta y)}{P(y \leq \eta < y + \Delta y)} = \frac{H(x, y + \Delta y) - H(x, y)}{G(y + \Delta y) - G(y)}$$

A § valószínűségi változó feltételes eloszlásfüggvénye azon feltétel mellett, hogy $\eta = y$ a következő határértékkel értelmezhető:

$$\begin{aligned}
 (2.5.2) \quad F(x|y) &= \lim_{\Delta y \rightarrow 0} \frac{\frac{H(x, y+\Delta y) - H(x, y)}{\Delta y}}{\frac{G(y+\Delta y) - G(y)}{\Delta y}} = \\
 &= \frac{\frac{\partial H(x, y)}{\partial y}}{g(y)} = \frac{\int_{-\infty}^x h(u, y) du}{\int_{-\infty}^{\infty} h(u, y) du}
 \end{aligned}$$

Hasonlóan:

$$(2.5.3) \quad G(y|x) = \frac{\frac{\partial H(x, y)}{\partial x}}{f(x)} = \frac{\int_{-\infty}^y h(x, v) dv}{\int_{-\infty}^{\infty} h(x, v) dv}$$

Ezekből az összefüggésekből differenciálás útján nyerjük a feltételes sűrűségfüggvényeket:

$$(2.5.4) \quad f(x|y) = \frac{\partial H(x|y)}{\partial x} = \frac{h(x, y)}{g(y)}$$

$$(2.5.5) \quad g(y|x) = \frac{\partial H(y|x)}{\partial y} = \frac{h(x, y)}{f(x)}$$

Ezekből az összefüggésekből

$$(2.5.6) \quad h(x, y) = f(x|y) g(y) = g(y|x) f(x)$$

adódik.

Felhasználva a (2.4.11) és (2.5.6) formulákat, az

$$(2.5.7) \quad f(x) = \int_{-\infty}^{\infty} f(x|y) g(y) dy, \quad \text{ill.}$$

$$(2.5.8) \quad g(y) = \int_{-\infty}^{\infty} g(y|x) f(x) dx$$

összefüggéseket kapjuk, amelyek a teljes valószínűség tételének folytonos analogonjai. Behelyettesítve a 2.4.10 és 2.4.11 formulákat a 2.4.9 képletbe, az

$$(2.5.9) \quad f(x|y) = \frac{g(y|x) f(x)}{\int_{-\infty}^{\infty} g(y|x) f(x) dx}$$

képletet nyerjük, amely a Bayes-tétel folytonos változata.

Abban az esetben, ha ξ és η egyaránt diszkrét eloszlású valószínűségi változók, akkor a (ξ, η) változópár is diszkrét eloszlású. Számunkra elegendő annak az esetnek a tárgyalása, amikor mind ξ mind η értékei csak nemnegatív egész számok lehetnek. Legyen $P(\xi = i) = p_i$, $P(\eta = j) = q_j$,

$$P(\xi = i, \eta = j) = r_{ij}, \quad (i, j = 1, 2, \dots).$$

Ebben az esetben a feltételes eloszlásokat az alábbi formulák értelmezik:

$$(2.5.10) \quad P(\xi = i | \eta = j) = \frac{P(\xi = i, \eta = j)}{P(\eta = j)} = \frac{r_{ij}}{q_j}$$

$$(2.5.11) \quad P(\eta = j | \xi = i) = \frac{P(\xi = i, \eta = j)}{P(\xi = i)} = \frac{r_{ij}}{p_i}$$

A matematikai statisztikában szükségünk lesz kettőnél több dimenziós valószínűségi vektorváltozók vizsgálatára. Ha $\xi_1, \xi_2, \dots, \xi_n$ valamely mennyiségre vonatkozó n számú mérés vagy megfigyelés eredménye, ekkor ezeket a változókat egy

$$\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$$

n -dimenziós véletlen vektor komponenseinek tekintjük. $\vec{\xi}$ értéke most az n -dimenziós tér véletlen helyzetű pontja. A $\vec{\xi}$ vektor eloszlását szokás a $\xi_1, \xi_2, \dots, \xi_n$ változók együttes eloszlásának is nevezni. Vizsgáljuk azt az esetet, amikor a szóban forgó valószínűségi változók folytonos eloszlásúak. A $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók együttes eloszlását a

$$(2.5.12) \quad H(x_1, x_2, \dots, x_n) = P(\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n)$$

n -dimenziós eloszlásfüggvény, vagy - ha létezik - a

$$(2.5.13) \quad h(x_1, x_2, \dots, x_n) = \frac{\partial^n H(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

n -dimenziós sűrűségfüggvény segítségével jellemezhetjük. Az n számú valószínűségi változó együttes eloszlásfüggvénye tehát a $\{\xi_i < x_i\} (i = 1, 2, \dots, n)$ n -számu esemény egyidejű bekövetkezésének valószínűségét adja meg.

Az n -dimenziós $H(x_1, x_2, \dots, x_n)$ mind az n változójának monoton nem csökkenő függvénye és minden változójában balról folytonos, $H(x_1, x_2, \dots, x_n) = 0$, ha bármelyik változója helyébe $-\infty$ -t írunk, (ha ugyanis, x_i helyébe $-\infty$ -t írunk, akkor a $\xi_i < -\infty$ esemény is belene, hogy következék, ami lehetetlen esemény), továbbá $H(+\infty, +\infty, \dots, +\infty) = 1$. (H értéke tehát csak akkor 1, ha mindegyik változója helyébe $+\infty$ -t írunk.)

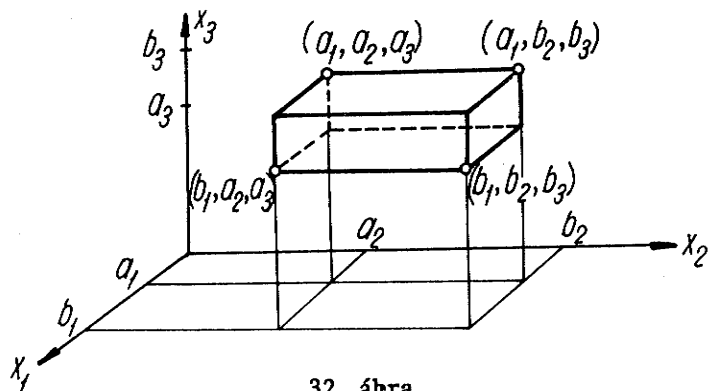
Annak valószínűsége, hogy a $\vec{\xi}$ n -dimenziós véletlen vektor végpontja az n -dimenziós tér egy előírt téglájába essék, azaz a

$$P(a_1 \leq \xi_1 < b_1, a_2 \leq \xi_2 < b_2, \dots, a_n \leq \xi_n < b_n) \geq 0$$

valószínűség kifejezhető a $H(x_1, x_2, \dots, x_n)$ eloszlásfüggvény segítségével, ami azt a kézenfekvő kikötést jelenti a

$$H(x_1, x_2, \dots, x_n)$$

függvénnyel szemben, hogy a fenti valószínűségnek H értékeivel való kifejezése nemnegatív kell legyen. Pl. $n = 3$ esetén a H függvényre teljesülnie kell a következő feltételnek:



32. ábra

$$H(b_1, b_2, b_3) - H(a_1, b_2, b_3) - H(b_1, a_2, b_3) - H(b_1, b_2, a_3) + \\ + H(a_1, a_2, b_3) + H(a_1, b_2, a_3) + H(b_1, a_2, a_3) - H(a_1, a_2, a_3) \geq 0.$$

Ha van olyan $h(x_1, x_2, \dots, x_n) \geq 0$ n -változós függvény, hogy az n -dimenziós tér tetszőleges E tartománya esetén

$$P(\vec{\xi} \in E) = \iiint_E h(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

akkor $\vec{\xi}$ folytonos eloszlású és a $H(x_1, \dots, x_n)$ és $h(x_1, \dots, x_n)$ függvényekre teljesül a reláció, valamint a vele ekvivalens:

$$(2.5.4) \quad \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} h(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n = \\ = H(x_1, x_2, \dots, x_n)$$

és

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = 1.$$

A (2.4.9)-(2.5.9) formulák könnyen általánosíthatók többdimenziós eloszlások esetére, csak ξ ill. η helyébe véletlen vektorokat kell helyettesíteni és természetesen az $f(x)$, $g(y)$, $f(x|y)$, $g(y|x)$ függvények mindegyike ξ ill. η dimenziójának megfelelő számú argumentummal bíró többváltozós függvény.

Ha pl. a $\xi_1, \xi_2, \dots, \xi_n$ változók együttes eloszlása folytonos eloszlás, akkor a $\xi_1, \xi_2, \dots, \xi_k$ változók együttes feltételes eloszlásfüggvénye azon feltétel mellett, hogy

$$\xi_{k+1} = x_{k+1}, \dots, \xi_n = x_n$$

$$F(x_1, x_2, \dots, x_k | x_{k+1}, \dots, x_n) =$$

$$= P(\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_k < x_k | \xi_{k+1} = x_{k+1}, \dots, \xi_n = x_n)$$

a következő határértékkel értelmezhető:

$$\begin{aligned}
 & F(x_1, x_2, \dots, x_k | x_{k+1}, \dots, x_n) = \\
 & = \lim_{\Delta x_{k+1} \rightarrow 0} \dots \lim_{\Delta x_n \rightarrow 0} P(\xi_1 < x_1, \dots, \xi_k < x_k | x_{k+1} \leq \xi_{k+1} < x_{k+1} + \Delta x_{k+1}, \dots, \\
 & \dots, x_n \leq \xi_n < x_n + \Delta x_n)
 \end{aligned}$$

A megfelelő feltételes sűrűségfüggvény:

$$\begin{aligned}
 (2.5.15) \quad & \frac{\partial F(x_1, \dots, x_k | x_{k+1}, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_k} = f(x_1, \dots, x_k | x_{k+1}, \dots, x_n) = \\
 & = \frac{h(x_1, x_2, \dots, x_n)}{g(x_{k+1}, \dots, x_n)}
 \end{aligned}$$

A többdimenziós feltételes sűrűségek fontos szerepet játszanak a regresszióanalízisben lásd: (12. fejezet).

Több valószínűségi változó együttes eloszlásának ismeretében most egy igen fontos fogalmat, a valószínűségi változók függetlenségét értelmezzük.

A $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változókat teljesen függetleneknek nevezzük, ha tetszőleges $a_i < b_i$ ($i = 1, 2, \dots, n$) számok esetén

$$\begin{aligned}
 & P(a_1 \leq \xi_1 < b_1, a_2 \leq \xi_2 < b_2, \dots, a_n \leq \xi_n < b_n) = \\
 & = P(a_1 \leq \xi_1 < b_1) P(a_2 \leq \xi_2 < b_2) \dots P(a_n \leq \xi_n < b_n)
 \end{aligned}$$

Speciálisan $a_i = -\infty$, $b_i = x_i$ választással

$$\begin{aligned}
 & P(\xi_1 < x_1, \dots, \xi_n < x_n) = \\
 & = P(\xi_1 < x_1) P(\xi_2 < x_2) \dots P(\xi_n < x_n), \text{ azaz} \\
 (2.5.16) \quad & H(x_1, x_2, \dots, x_n) = F_1(x_1) F_2(x_2) \dots F_n(x_n)
 \end{aligned}$$

Más szóval n számú valószínűségi változót akkor nevezünk teljesen függetlennek, ha együttes eloszlásfüggvényük az egyes változók eloszlásfüggvényeinek szorzatával egyenlő. Ha a változók folytonosak együttes eloszlásuk is folytonos eloszlás, akkor a függetlenségből következik, hogy

együttes sűrűségfüggvényük is az egyes változók sűrűségfüggvényeinek szorzata:

$$(2.5.17) \quad h(x_1, x_2, \dots, x_n) = f_1(x_1) f_2(x_2) \dots f_n(x_n).$$

A függetlenség fogalma számunkra azért is különös fontosságu, mert a matematikai statisztika jól kidolgozott módszerei szinte kizárólag független valószínűségi változókra vonatkoznak. Statisztikai módszerek kidolgozása nem független valószínűségi változók esetére egyike a matematikai statisztika legsürgősebb feladatainak. Jelenleg intenzív kutatások folynak ebben az irányban, de még csak kezdeti eredmények ismeretesek.

Annak eldöntése, hogy adott $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók valóban függetlenek-e, a matematikai statisztika feladata. Erre vonatkozó módszereket a 10.4 pontban ismertetünk.

2.6 Valószínűségi változó függvényének eloszlása

a) Valószínűségi változó monoton függvényének eloszlása

A gyakorlatban előfordul, hogy ismerjük valamely ξ valószínűségi változó eloszlását és szükségünk van a $\eta = \psi(\xi)$ függvényének eloszlására. Az η változó természetesen ugyancsak valószínűségi változó. Az η változó eloszlását egyszerűen meghatározhatjuk ξ eloszlása alapján, ha az η és ξ közötti függvénykapcsolatot kifejező $y = \psi(x)$ függvény monoton növekvő vagy csökkenő és differenciálható függvény, tehát létezik az inverz függvénye. Ez esetben ugyanis, ha eloszlásfüggvénye $F(x)$, akkor az η változó eloszlásfüggvényét $G(y)$ -nal jelölve:

$$(2.6.1) \quad G(y) = P(\eta < y) = P(\psi(\xi) < y) = P(\xi < \psi^{-1}(y)) = F(\psi^{-1}(y))$$

Ha a ξ változónak létezik a sűrűségfüggvénye és az $f(x)$, akkor az η változó sűrűségfüggvénye:

$$(2.6.2) \quad g(y) = \frac{dG(y)}{dy} = \frac{dF[\psi^{-1}(y)]}{dy} = f[\psi^{-1}(y)] \left| \frac{d\psi^{-1}(y)}{dy} \right|.$$

Ha pl. $\eta = a\xi + b$, azaz $y = \psi(x) = ax + b$ és ξ sűrűségfüggvénye $f(x)$, akkor

$$x = \psi^{-1}(y) = \frac{y-b}{a}$$

$$(2.6.3) \quad g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$$

$$\text{Ha } \eta = e^{\xi}$$

$$\text{az } y = \psi(x) = e^x, \quad x = \psi^{-1}(y) = \ln y$$

$$(2.6.4) \quad g(y) = \frac{1}{|y|} f(\ln y).$$

b) Független valószínűségi változók összegének, szorzatának és hányadosának eloszlása

Független valószínűségi változók összegének eloszlása
Valószínűségeloszlások kompozíciója

Gyakran előfordul, hogy a megfigyelés tárgyát képező mennyiség véletlen ingadozása két vagy több véletlen mennyiség összegezett hatásának eredménye, amely hatásokról feltételezzük, hogy egymástól függetlenek. Ilyen esetben az egyes komponensek eloszlásának ismeretében összegük eloszlása meghatározható. Matematikailag kifejezve legyen a vizsgált ζ valószínűségi változó az egymástól független ξ és η valószínűségi változók összege:

$$\zeta = \xi + \eta$$

Határozzuk meg ζ valószínűségeloszlását, ha ismerjük ξ és η eloszlását. Vizsgáljuk először azt az esetet, amikor mind ξ mind η egész értékű valószínűségi változók

$$P(\xi = i) = p_i, \quad P(\eta = j) = q_j, \quad (i, j = 0, 1, 2, \dots)$$

eloszlásokkal. Ez esetben ζ is egész értékű valószínűségi változó 1, 2, 3, ... lehetséges értékekkel.

A $\{\zeta = n\}$ esemény a következő egymást kizáró módokon jöhet létre:

$$\begin{array}{ll}
\xi = 0, & \eta = n \\
\xi = 1, & \eta = n-1 \\
\vdots & \vdots \\
\xi = k, & \eta = n-k, \\
\vdots & \vdots \\
\xi = n, & \eta = 0.
\end{array}$$

A ξ és η változók feltételezett függetlensége folytán:

$$\begin{aligned}
(2.6.5) \quad P(\xi = n) &= \sum_{k=0}^n P(\xi = k, \eta = n-k) = \sum_{k=0}^n P(\xi = k)P(\eta = n-k) = \\
&= \sum_{k=0}^n p_k q_{n-k} = \sum_{k=0}^n p_{n-k} q_k
\end{aligned}$$

A (2.6.5) összefüggés a p_i ill. q_j valószínűségekből alkotott eloszlásokból egy újabb $P(\xi = n) = r_n$ valószínűségeloszlást származtat

($n = 0, 1, 2, \dots$), amely $\{r_n\}$ eloszlást a $\{p_i\}$ és $\{q_j\}$

eloszlások kompozíciójának nevezzük. A kompozíció művelete igen fontos számunkra, mivel független valószínűségi változók összegének eloszlására a gyakorlatban sokszor van szükségünk.

Ha a fent szereplő eloszlásokra bevezetjük a

$$P = \begin{pmatrix} 0 & 1 & 2 & \dots & k & \dots \\ p_0 & p_1 & p_2 & \dots & p_k & \dots \end{pmatrix}, \quad Q = \begin{pmatrix} 0 & 1 & 2 & \dots & j & \dots \\ q_0 & q_1 & q_2 & \dots & q_j & \dots \end{pmatrix},$$

$$R = \begin{pmatrix} 0 & 1 & 2 & \dots & n & \dots \\ r_0 & r_1 & r_2 & \dots & r_n & \dots \end{pmatrix}$$

jelölést, akkor a kompozíciót szokás az

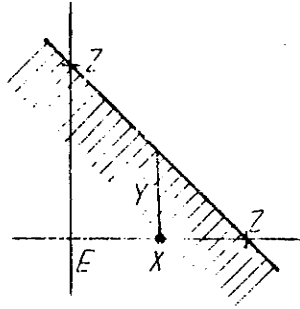
$$R = P * Q$$

szimbólummal jelölni. Képezhető kettőnél több független egész értékű valószínűségi változó eloszlásának kompozíciója is. Ha $\xi_1, \xi_2, \dots, \xi_n$ független egész értékű valószínűségi változók $P_1, P_2, \dots, P_n$ eloszlá-

sokkal, akkor a $\xi = \xi_1 + \xi_2 + \dots + \xi_n$ összeg P eloszlása az egyes komponensek eloszlásának kompozíciója:

$$P = P_1 * P_2 * P_3 * \dots * P_n$$

Több összeadandó kompozíciójának tényleges kiszámítása általában bonyolult, ilyenkor rendszerint ún. generátorfüggvényekkel dolgozunk, amelyek ismertetését a 4.1 pontban adjuk meg.



33. ábra

Legyenek most ξ és η független folytonos valószínűségi változók $f(x)$ ill. $g(y)$ sűrűségfüggvényekkel és $h(x, y)$ együttes sűrűségfüggvénnyel, a $\zeta = \xi + \eta$ valószínűségi változó eloszlásfüggvényét jelöljük $R(z)$ -vel, sűrűségfüggvényét pedig $r(z)$ -vel.

A $\{\zeta < z\} = \{\xi + \eta < z\}$ esemény most is azt jelenti, hogy a (ξ, η) véletlen helyzetű pont az $x + y \leq z$ összefüggés által meghatározott E félsíkra esik: (33. ábra)

Tehát:

$$\begin{aligned} (2.6.6) \quad R(z) &= P(\zeta < z) = P(\xi + \eta < z) = \iint_E h(x, y) \, dx \, dy = \\ &= \iint_E f(x) g(y) \, dx \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x) g(y) \, dx \, dy \end{aligned}$$

Bevezetve az $u = x + y$, $v = y$ helyettesítést

$$(2.6.7) \quad R(z) = \int_{-\infty}^z \int_{-\infty}^{\infty} f(u-v) g(v) \, du \, dv$$

Innen z -szerinti deriválással nyerjük az $r(z)$ sűrűségfüggvényt:

$$(2.6.8) \quad r(z) = \int_{-\infty}^{\infty} f(z-v) g(v) \, dv = \int_{-\infty}^{\infty} f(z-y) g(y) \, dy$$

amely összefüggést folytonos eloszlások kompozíciójának nevezzük. Szokásos az

$$r(z) = f(x) * g(y)$$

jelölés is.

Ha mind $f(x)$ mind $g(x)$ csak nemnegatív x értékekre különbözik zérustól, akkor a kompozíció az

$$(2.6.9) \quad r(z) = \int_0^z f(z-y) g(y) dy$$

formulára redukálódik.

c) Szorzat és hányados sűrűségfüggvénye

Legyenek ξ és η folytonos valószínűségi változók $f(x)$ ill. $g(x)$ sűrűségfüggvényekkel és legyen $\zeta = \xi \eta$. Határozzuk meg ζ feltételes sűrűségfüggvényét azon feltétel mellett, hogy $\eta = y$. Jelöljük ezt a feltételes sűrűségfüggvényt $h(z|\eta = y)$. Ha $\eta = y$, akkor $\xi \cdot \eta = \xi \cdot y$, ahol y rögzített szám, így $\xi \cdot y$ sűrűségfüggvénye a (2.5.2) formula alapján:

$$\frac{1}{|y|} f\left(\frac{x}{y}\right)$$

A ζ változó sűrűségfüggvényét $h(z)$ -vel jelölve a (2.4.11) képlet alapján:

$$(2.6.10) \quad h(z) = \int_{-\infty}^{\infty} h(z|\eta = y) g(y) dy = \int_{-\infty}^{\infty} \frac{1}{|y|} f\left(\frac{x}{y}\right) g(y) dy$$

A feltételes sűrűségfüggvény segítségével ugyancsak egyszerűen határozható meg a $\zeta = \frac{\xi}{\eta}$ valószínűségi változó sűrűségfüggvénye. Számítsuk ki először a $\frac{\xi}{\eta}$ változó $k(z|\eta = y)$ feltételes sűrűségfüggvényét az $\eta = y$ feltétel mellett. Ugyancsak a (2.5.2) formula alapján az $\frac{1}{y} \cdot \xi$ sűrűségfüggvénye:

$$|y| f(xy)$$

Tehát:

$$(2.6.11) \quad k(z) = \int_{-\infty}^{\infty} k(z|\eta = y) g(y) dy = \int_{-\infty}^{\infty} |y| f(xy) g(y) dy$$

A VALÓSZÍNŰSÉGELOSZLÁSOK JELLEMZŐ ADATAI

Valamely valószínűségi változó eloszlásának törvényszerűségeit az eloszlásfüggvény vagy (ha létezik) a sűrűségfüggvény segítségével tudjuk pontosan megfogalmazni. A gyakorlat számára sokszor elegendő (és szükséges), hogy a valószínűségeloszlásról néhány jellemző számadat segítségével szemléletes áttekintő képet nyerjünk. Utaltunk már a valószínűségeloszlás és a tömegeloszlás analógiájára. A tömegeloszlással kapcsolatban rendszerint elsősorban az érdekel bennünket, hogy hol van az eloszlás súlypontja, centruma, másrészt, hogy milyen szorosan tömörül a tömeg a súlypont körül, amire az inercianyomaték utal. Hasonló módon, a valószínűségeloszlással kapcsolatban is elsődleges érdekességű, hogy mely érték körül tömörülnek a valószínűségi változó felvett értékei, hol van a valószínűségeloszlás súlypontja, amelyet várható értéknek nevezünk. Ugyancsak érdekel bennünket, hogy a valószínűségi változó értékei mennyire szorosan tömörülnek a várható érték körül, hogy a várható érték - mint középpont - körül rövid vagy hosszú intervallum fogja tartalmazni a valószínűségi változó felvett értékeinek nagy százalékát (mondjuk 70-90 százalékát). A valószínűségeloszlással kapcsolatban az inercianyomatéknak megfelelő fogalom a szórás, ami lényegében a valószínűségi változó értékeinek a várható értéktől való négyzetes átlageltérése.

A várható értéken és a szóráson kívül egyéb numerikus jellemzőkkel is meg fogunk ismerkedni. Megadunk korlát-számokat, amelyek közre fogják a valószínűségi változó értékeinek 25, 50, 75%-át, továbbá olyan mérőszámot, amely az eloszlás szimmetrikus vagy aszimmetrikus voltát jellemzi. Vizsgálni fogjuk több valószínűségi változó értékei közötti sztochasztikus kapcsolat mérőszámait.

3.1 A várható érték fogalma és tulajdonságai

Valamely valószínűségeloszlás várható értékének vagy centrumának kiszámítása a tömegeloszlás-súlypontjának kiszámításával analóg módon történik. Legyen ξ diszkrét eloszlású valószínűségi változó, amelynek lehetséges értékeit jelöljük az $x_1, x_2, \dots, x_n, \dots$ számokkal és legyen

$$P(\xi = x_i) = p_i \quad (i = 1, 2, \dots)$$

Ekkor a ξ valószínűségi változó várható értékén, amelyet $M(\xi)$ -vel jelölünk, a következő kifejezést értjük:

$$(3.1.1) \quad M(\xi) = \sum_k x_k p_k$$

Tekintettel arra, hogy $\sum_k p_k = 1$, látjuk, hogy diszkrét eloszlású ξ

valószínűségi változó várható értéke a ξ lehetséges értékeinek súlyozott számtani középáránysa, ahol a súlyok az egyes értékekhez tartozó valószínűségek.

Ha ξ folytonos eloszlású valószínűségi változó $f(x)$ sűrűségfüggvényével, akkor ξ várható értékét a folytonos tömegeloszlás súlypontjának analógiájára az

$$(3.1.2) \quad M(\xi) = \int_{-\infty}^{\infty} x f(x) dx$$

kifejezés értelmezi.

A várható érték néhány fontos tulajdonságát fejezik ki az alábbi tételek.

1. Tétel: Állandó várható értéke önmaga.

$$(3.1.3) \quad M(c) = c \quad (c = \text{const.})$$

Bizonyítás: A c állandó úgy tekinthető, mint olyan valószínűségi változó, amely az egyetlen c értéket 1 valószínűséggel veszi fel, tehát a várható érték definíciója alapján

$$M(c) = c \cdot 1 = c$$

2. Tétel: Ha ξ korlátos valószínűségi változó azaz

$$a \leq \xi \leq b,$$

akkor létezik a várható értéke és

$$(3.1.4) \quad a \leq M(\xi) \leq b.$$

Bizonyítás: Ha ξ diszkrét valószínűségi változó $x_1, x_2, \dots$ lehetséges értékekkel, s az egyes értékeket rendre $p_1, p_2, \dots$ valószínűségekkel veszi fel, akkor mivel az $x_1, x_2, \dots$ értékek mindegyike az a és b határok közé esik

$$a = ap_1 + ap_2 + \dots \leq x_1 p_1 + x_2 p_2 + \dots \leq bp_1 + bp_2 + \dots$$

azaz

$$a = a \sum p_i \leq \sum x_i p_i \leq b \sum p_i = b.$$

Ha ξ folytonos eloszlású, akkor

$$\begin{aligned} a &= a \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^{\infty} af(x) dx \leq \int_{-\infty}^{\infty} xf(x) dx \leq \int_{-\infty}^{\infty} bf(x) dx = \\ &= b \int_{-\infty}^{\infty} f(x) dx = b. \end{aligned}$$

Megjegyezzük, hogy $a \leq \xi \leq b$ esetén

$$\int_a^b f(x) dx = 1.$$

3. Tétel: Ha $\eta = a\xi + b$, azaz az η valószínűségi változó a ξ valószínűségi változónak lineáris függvénye, akkor

$$(3.1.5) \quad M(\eta) = M(a\xi + b) = a M(\xi) + b$$

Bizonyítás: Ha ξ diszkrét eloszlású $x_1, x_2, \dots$ lehetséges értékekkel, amelyeket rendre $p_1, p_2, \dots$ valószínűségekkel vesz fel, akkor az η valószínűségi változó lehetséges értékei az $ax_1 + b, ax_2 + b, \dots$ számok, amelyeket ugyancsak rendre $p_1, p_2, \dots$ valószínűségekkel vesz fel, tehát

$$M(\eta) = \sum_k (ax_k + b)p_k = a \sum_k x_k p_k + b \sum_k p_k = aM(\xi) + b$$

feltéve természetesen, hogy a $\sum_k x_k p_k$ sor konvergens, azaz ξ -nek létezik várható értéke.

Ha ξ folytonos eloszlású $f(x)$ sűrűségfüggvénnyel, akkor $\xi = x$ esetén $\eta = ax + b$ és η sűrűségfüggvénye a (2.5.3) formula alapján:

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right)$$

és

$$M(\eta) = \int_{-\infty}^{\infty} y g(y) dy = \frac{1}{|a|} \int_{-\infty}^{\infty} y f\left(\frac{y-b}{a}\right) dy$$

$$\begin{aligned} \frac{y-b}{a} &= u \text{ helyettesítéssel} \\ M(\eta) &= \frac{a}{|a|} \int_{-\infty}^{\infty} (au+b) f(u) dy = \\ &= a \int_{-\infty}^{\infty} u f(u) du + b \int_{-\infty}^{\infty} f(u) du = aM(\xi) + b. \end{aligned}$$

Speciális esetenként megemlítjük, hogy ha $\eta = \xi - M(\xi)$, akkor

$$M(\eta) = M(\xi) - M(\xi) = 0.$$

Definíció: Legyen az η valószínűségi változó a ξ valószínűségi változó valamely függvénye, $\eta = \varphi(\xi)$. Ekkor az η valószínűségi változó értékét diszkrét eloszlás esetén az

$$(3.1.6) \quad M(\eta) = \sum_k \varphi(x_k) p_k$$

folytonos eloszlás esetén az

$$(3.1.7) \quad M(\eta) = \int_{-\infty}^{\infty} \varphi(x) f(x) dx$$

formulákkal definiáljuk.

Analóg módon definiáljuk valószínűségi vektorváltozók függvényének várható értékét. Pl. folytonos eloszlású valószínűségi változók esetén, ha $\eta = \varphi(\xi_1, \xi_2, \dots, \xi_n)$ és a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók együttes sűrűségfüggvénye $f(x_1, x_2, \dots, x_n)$ akkor

$$(3.1.8) \quad M(\eta) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \varphi(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n$$

4. Tétel: Legyenek ξ és η tetszőleges eloszlású valószínűségi változók, amelyek várható értékei léteznek, ekkor

$$(3.1.9) \quad M(\xi + \eta) = M(\xi) + M(\eta).$$

A bizonyítást először diszkrét eloszlású változók esetére mutatjuk meg. Legyenek ξ lehetséges értékei az $x_1, x_2, \dots$ számok, η lehetséges értékei az $y_1, y_2, \dots$ számok, ekkor $\xi + \eta$ lehetséges értékei az $x_1 + y_1, x_2 + y_2, \dots$ számok. Legyen továbbá

$$P(\xi = x_i) = p_i, \quad P(\eta = y_j) = q_j, \quad P(\xi = x_i, \eta = y_j) = r_{ij},$$

itt tehát az r_{ij} valószínűségek a (ξ, η) változó pár együttes eloszlását adják meg. A várható érték definíciója alapján

$$\begin{aligned} M(\xi + \eta) &= \sum_i \sum_j (x_i + y_j) r_{ij} = \\ &= \sum_i \sum_j x_i r_{ij} + \sum_i \sum_j y_j r_{ij} = \\ &= \sum_i x_i \left(\sum_j r_{ij} \right) + \sum_j y_j \left(\sum_i r_{ij} \right). \end{aligned}$$

A (2.4.5) és (2.4.6) összefüggések alapján:

$$\sum_i r_{ij} = p_i, \quad \sum_i r_{ij} = q_j.$$

Tehát:

$$M(\xi + \eta) = \sum_i x_i p_i + \sum_j y_j q_j = M(\xi) + M(\eta).$$

Folytonos eloszlású valószínűségi változók esetén, ha ξ és η együttes sűrűsége $f(x, y)$, akkor a (3.1.8) összefüggés alapján $\varphi(x, y) = x + y$ helyettesítéssel

$$\begin{aligned} M(\xi + \eta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f(x, y) dx dy = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dx dy = \\ &= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x, y) dy \right) dx + \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f(x, y) dx \right) dy = \\ &= \int_{-\infty}^{\infty} x f(x) dx + \int_{-\infty}^{\infty} y g(y) dy, \end{aligned}$$

amiből az állítás leolvasható.

A valószínűségi változók összegének várható értékére vonatkozó tétel teljes indukcióval tetszőleges véges számú összeadandóra általánosítható:

$$(3.1.10) \quad M(\xi_1 + \xi_2 + \dots + \xi_n) = M(\xi_1) + M(\xi_2) + \dots + M(\xi_n).$$

Legyen ugyanis $\xi = \xi_1 + \xi_2 + \dots + \xi_{n-1}$, $\eta = \xi_n$ és tegyük fel, hogy a tétel $n-1$ összeadandóra érvényes, azaz igaz, hogy

$$M(\xi) = M(\xi_1) + M(\xi_2) + \dots + M(\xi_{n-1}),$$

akkor a bizonyított tétel alapján következik, hogy

$$(3.1.11) \quad M(\xi_1 + \dots + \xi_n) = M(\xi + \eta) = M(\xi) + M(\eta) = \\ = \sum_{i=1}^{n-1} M(\xi_i) + M(\xi_n).$$

5. Tétel: Független valószínűségi változók szorzatának várható értéke várható értékeik szorzatával egyenlő, azaz ha ξ és η független valószínűségi változók, akkor

$$(3.1.12) \quad M(\xi \eta) = M(\xi) M(\eta)$$

A bizonyítást először folytonos eloszlású valószínűségi változók esetére végezzük el. A ξ és η változók függetlensége miatt együttes sűrűségfüggvényük az egyes változók sűrűségfüggvényeinek szorzatával egyenlő (lásd: (2.4.19) formula), azaz $h(x, y) = f(x) g(y)$, a (3.1.11) formula alapján $\varphi(x, y) = x y$ helyettesítéssel:

$$M(\xi \eta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x) g(y) dx dy = \int_{-\infty}^{\infty} x f(x) dx \int_{-\infty}^{\infty} y g(y) dy = M(\xi) M(\eta)$$

Ha ξ és η diszkrét eloszlásúak és $P(\xi = x_i) = p_i$, $P(\eta = y_j) = q_j$, továbbá együttes eloszlásuk: $P(\xi = x_i, \eta = y_j) = r_{ij}$, akkor az η és ξ változók függetlensége folytán

$$r_{ij} = p_i q_j \quad (i, j = 1, 2, \dots)$$

Ekkor

$$M(\xi \eta) = \sum_i \sum_j x_i y_j r_{ij} = \sum_i \sum_j x_i y_j p_i q_j = \\ = \sum_i x_i p_i \sum_j y_j q_j = M(\xi) M(\eta).$$

Utóbbi tételünk ugyancsak könnyen általánosítható tetszőleges véges számú egymástól független valószínűségi változó szorzatok várható értékének kiszámítására. A bizonyítást teljes indukcióval végezhetjük miként az összeg esetében. Igaz tehát, hogy függetlenség esetén

$$(3.1.12) \quad M(\xi_1 \xi_2 \dots \xi_n) = M(\xi_1) M(\xi_2) \dots M(\xi_n)$$

feltéve, hogy az egyes tényezők várható értéke létezik.

3.2 A feltételes várható érték

Legyenek ξ és η diszkrét valószínűségi változók az alábbi eloszlásokkal:

$$\begin{array}{ll} \xi : x_1, x_2, \dots, x_n, \dots & \eta : y_1, y_2, \dots, y_n, \dots \\ p_1, p_2, \dots, p_n, \dots & q_1, q_2, \dots, q_n, \dots \end{array}$$

Legyen továbbá $P(\xi = x_i, \eta = y_k) = r_{ik}$ ($i = 1, 2, \dots, n; k = 1, 2, \dots, m$)

Értelmezhetjük az η valószínűségi változónak a feltételes várható értékét azon feltétel mellett, hogy a ξ változó adott x_i értéket vett fel.

$$(3.1.1) \quad M(\eta | \xi = x_i) = \sum_k y_k P(\eta = y_k | \xi = x_i)$$

Formulánk segítségével az η változó feltételes várható értékét a ξ változó minden aktuális értékére - mint feltételre - kiszámíthatjuk. Az így nyert mennyiség értéke attól függ, hogy ξ melyik x_i értéket veszi fel, ez pedig a véletlentől függ, $M(\eta | \xi)$ tehát valószínűségi változó, amelynek lehetséges értékei az

$$M(\eta | \xi = x_1), \quad M(\eta | \xi = x_2), \quad \dots, \quad M(\eta | \xi = x_n).$$

Könnyű belátni, hogy $M(\eta | \xi)$ az $M(\eta | \xi = x_k)$ értékét éppen p_k valószínűséggel veszi fel.

Számítsuk ki az $M(\eta | \xi)$ valószínűségi változó várható értékét!

$$M[M(\eta | \xi)] = \sum_i M(\eta | \xi = x_i) p_i$$

Felhasználva a (3.2.1) összefüggést, az alábbi érdekes eredményt kapjuk:

$$\begin{aligned}
 (3.2.2) \quad M[M(\eta|\xi)] &= \sum_i \left[\sum_k y_k P(\eta = y_k | \xi = x_i) \right] p_i = \\
 &= \sum_i \sum_k y_k r_{ik} = \sum_k y_k \sum_i r_{ik} = \sum_k y_k q_k = M(\eta)
 \end{aligned}$$

(Itt felhasználtuk, hogy

$$P(\eta = y_k | \xi = x_i) P(\xi = x_i) = P(\eta = y_k, \xi = x_i) = r_{ik}.)$$

Azt az eredményt kaptuk tehát, hogy az η változó ξ -re vonatkozó feltételes várható értékének a várható értéke az η változó feltétel nélküli várható értékével egyenlő.

Ha ξ és η folytonos valószínűségi változók $f(x)$ ill. $g(y)$ sűrűségfüggvényekkel és $h(xy)$ együttes sűrűségfüggvénnyel, akkor az valószínűségi változó ξ -re vonatkozó feltételes várható értékét az

$$(3.2.3) \quad M(\eta | \xi = x) = \int_{-\infty}^{\infty} y \frac{h(x, y)}{f(x)} dy = \int_{-\infty}^{\infty} y g(y | x) dy$$

formulával értelmezzük. Ebben az esetben is érvényes az

$$\begin{aligned}
 M[M(\eta | \xi)] &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y \frac{h(x, y)}{f(x)} dy \right) f(x) dx = \\
 &= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} h(x, y) dx \right) dy = \int_{-\infty}^{\infty} y g(y) dy = M(\eta)
 \end{aligned}$$

összefüggés.

(Feltéve természetesen, hogy a fenti Improprius-integrál konvergens.)

A $h(x) = M(\eta | \xi = x)$ függvényt, amely az x -változó függvénye, az η -változó ξ -re vonatkozó regressziós görbéjének nevezzük.

Teljesen analóg módon definiáljuk a ξ valószínűségi változó η -ra vonatkozó feltételes várható értékét.

Definiálhatjuk az $M(\eta | \xi_1, \xi_2, \dots, \xi_n)$ feltételes várható értéket, amely a $\xi_1, \xi_2, \dots, \xi_n$ változók függvénye, tehát n -változós függvény. Ennek a függvénynek az alábbi érdekes minimum-tulajdonsága van: ha $g(\xi_1, \xi_2, \dots, \xi_n)$ a $\xi_1, \xi_2, \dots, \xi_n$ változók tetszőleges függvénye, akkor

$$(3.2.4) \quad M \left\{ \left[\eta - g(\xi_1, \dots, \xi_n) \right]^2 \right\} \geq M \left\{ \left[\eta - M(\eta | \xi_1, \dots, \xi_n) \right]^2 \right\}$$

3.3 A szórás fogalma és tulajdonságai

Mint említettük, a várható érték a valószínűségeloszlásnak csak egyik számszerű jellemzője, amely csak azt mondja meg, hogy melyik az a centrális pont, amely körül a valószínűségi változó értékei ingadoznak, az ingadozás mértékére vonatkozólag azonban nem mond semmit. A valószínűségi változó értékeinek a centrális pont körüli szóródását a

$$(3.3.1) \quad D(\xi) = + \sqrt{M[\xi - M(\xi)]^2}$$

kifejezés méri. A gyökjel alatt lényegében a valószínűségi változó értékeinek a várható értéktől való négyzetes átlageltérése szerepel. A szórásról akkor beszélhetünk, ha létezik az $M(\xi)$ várható érték továbbá létezik a $[\xi - M(\xi)]^2$ valószínűségi változó várható értéke.

A valószínűségi változó szórásnégyzete:

$$(3.3.2) \quad D^2(\xi) = M \left\{ [\xi - M(\xi)]^2 \right\}$$

A szórásnégyzet kiszámítását megkönnyíti az alábbi összefüggés:

$$\begin{aligned} D^2(\xi) &= M \left\{ \xi^2 - 2\xi M(\xi) + M^2(\xi) \right\} = \\ (3.3.3) \quad &= M(\xi^2) - 2 M(\xi) M(\xi) + M[M^2(\xi)] = \\ &= M(\xi^2) - M^2(\xi) \end{aligned}$$

Mivel $D^2(\xi) \geq 0$, következik, hogy

$$(3.3.4) \quad M(\xi^2) \geq [M(\xi)]^2$$

A szórásnégyzet kiszámítása diszkrét valószínűségi változó esetén:

$$(3.3.5) \quad D^2(\xi) = \sum_i [x_i - M(\xi)]^2 p_i = \sum_i x_i^2 p_i - M^2(\xi)$$

Folytonos valószínűségi változó esetén

$$(3.3.6) \quad D^2(\xi) = \int_{-\infty}^{\infty} [x - M(\xi)]^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - M^2(\xi).$$

A szórásnégyzet fogalma segítségével az $M(\xi) = m$ várható értéknek egy újabb igen fontos tulajdonságát fejezhetjük ki. Ha a tetszőleges valós szám, akkor

$$(3.3.7) \quad M[(\xi - a)^2] \geq M[(\xi - m)^2].$$

Ugyanis

$$M(\xi - a)^2 = M(\xi - m + m - a)^2 = M(\xi - m)^2 + 2(m - a)M(\xi - m) + (m - a)^2$$

Tekintettel arra, hogy $M(\xi - m) = M(\xi) - m = m - m = 0$,

$$M(\xi - a)^2 = M(\xi - m)^2 + (m - a)^2 = D^2(\xi) + (m - a)^2 \geq D^2(\xi)$$

Tehát az $M(\xi) = m$ várható érték a számegyenesnek az a pontja, amelytől a valószínűségi változó értékeinek négyzetes átlageltérése kisebb, mint bármely más ponttól számított négyzetes átlageltérése.

A (3.3.7) összefüggést Steiner-formulának nevezzük.

Tételek a szórásnégyzetre vonatkozólag.

1. Tétel: Ha az η valószínűségi változó a ξ valószínűségi változónak lineáris függvénye: $\eta = a\xi + b$, akkor

$$(3.3.8) \quad D^2(\eta) = a^2 D^2(\xi)$$

Bizonyítás: A szórásnégyzet definíciója alapján:

$$\begin{aligned} D^2(\eta) &= M\{[(a\xi + b) - M(a\xi + b)]^2\} = M[a^2(\xi - M(\xi))^2] = \\ &= a^2 M(\xi - M(\xi))^2 = a^2 D^2(\xi) \end{aligned}$$

2. Tétel: Ha ξ és η független valószínűségi változók és $\zeta = \xi + \eta$ akkor

$$(3.3.9) \quad D^2(\zeta) = D^2(\xi) + D^2(\eta).$$

Bizonyítás: Ha ξ és η függetlenek akkor $\xi - M(\xi)$ és $\eta - M(\eta)$ is függetlenek. Figyelembe véve, hogy független valószínűségi változók szorzatának várható értéke a várható értékek szorzásnégyzetére, továbbá, hogy $M[\xi - M(\xi)] = 0$, $M[\eta - M(\eta)] = 0$, a szórásnégyzet definíciója alapján:

$$\begin{aligned}
 D^2(\xi) &= D^2(\xi + \eta) = M[(\xi + \eta) - (M(\xi) + M(\eta))]^2 = \\
 &= M[(\xi - M(\xi)) + (\eta - M(\eta))]^2 = M(\xi - M(\xi))^2 + \\
 &+ 2M[(\xi - M(\xi))(\eta - M(\eta))] = M(\xi - M(\xi))^2 + \\
 &+ 2M[(\xi - M(\xi))M(\eta - M(\eta))] = M(\xi - M(\xi))^2 + \\
 &+ 2M(\xi - M(\xi))M(\eta - M(\eta)) = D^2(\xi) + D^2(\eta)
 \end{aligned}$$

A 2. tétel tetszőleges véges számú független valószínűségi változó esetére általánosítható teljes indukcióval:

Ha $\xi_1, \xi_2, \dots, \xi_n$ független valószínűségi változók, akkor

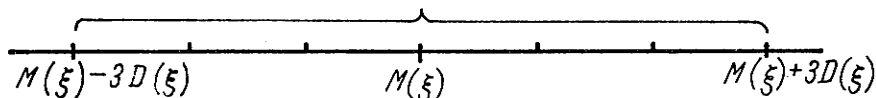
$$(3.3.10) \quad D^2(\xi_1 + \xi_2 + \dots + \xi_n) = D^2(\xi_1) + D^2(\xi_2) + \dots + D^2(\xi_n)$$

A valószínűségi változó várható értékének és szórásának ismeretében a valószínűségi változó ingadozását a várható értéke körül jól jellemezhetjük. Ezt fejezi ki a Csebisev-féle egyenlőtlenség.

3. Tétel: Csebisev-egyenlőtlenség. Ha a ξ valószínűségi változó várható értéke és szórása létezik, akkor

$$(3.3.11) \quad P(|\xi - M(\xi)| > \lambda D(\xi)) \leq \frac{1}{\lambda^2}$$

A Csebisev egyenlőtlenség azt jelenti, hogy pl. annak valószínűsége, hogy egy valószínűségi változó értéke a várható értékétől abszolútértékre nézve jobban eltérjen, mint $2 D(\xi)$ kisebb mint $\frac{1}{4}$, míg annak valószínűsége, hogy ξ megfigyelt értéke az $(M(\xi) - 3 D(\xi), M(\xi) + 3 D(\xi))$ intervallumon kívül essék, kisebb mint $\frac{1}{9}$. Más szóval, a számegyenesen felrajzolva a fenti intervallumot



34. ábra

ha ξ -re vonatkozólag nagyszámú megfigyelést végzünk, akkor a megfigyelt értékek kb. 90%-a ebbe az intervallumba esik, feltéve, hogy ξ várható értéke és szórása létezik. Ha a (3.3.11) összefüggésben $\lambda D(\xi) = \varepsilon$

helyettesítést alkalmazunk, akkor $\frac{1}{\lambda^2} = \frac{D^2(\xi)}{\varepsilon^2}$, így a (3.3.11) összefüggés ekvivalens a következő állítással:

$$(3.3.12) \quad P(|\xi - M(\xi)| > \varepsilon) \leq \frac{D^2(\xi)}{\varepsilon^2}$$

Bizonyítás: Egyszerűség kedvéért vezessük be az $M(\xi) = m$, $D(\xi) = \sigma$ jelölést. Ha ξ diszkrét eloszlású $x_1, x_2, \dots$ lehetséges értékekkel, amelyeket rendre $p_1, p_2, \dots$ valószínűségekkel vesz fel:

$$\begin{aligned} \sigma^2 &= \sum_k (x_k - m)^2 p_k \geq \sum_{k: |x_k - m| > \varepsilon} (x_k - m)^2 p_k \geq \varepsilon^2 \sum_{k: |x_k - m| > \varepsilon} p_k = \\ &= \varepsilon^2 P(|\xi - m| > \varepsilon). \end{aligned}$$

Innen:
$$P(|\xi - m| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Ha ξ folytonos eloszlású $f(x)$ sűrűséggel, akkor

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} (x-m)^2 f(x) dx \geq \int_{-\infty}^{m-\varepsilon} (x-m)^2 f(x) dx + \int_{m+\varepsilon}^{\infty} (x-m)^2 f(x) dx \geq \\ &\geq \varepsilon^2 \left(\int_{-\infty}^{m-\varepsilon} f(x) dx + \int_{m+\varepsilon}^{\infty} f(x) dx \right) = \varepsilon^2 P(|\xi - m| > \varepsilon), \end{aligned}$$

amiből a tétel állítása leolvasható.

Megjegyzés: A Csebisev-egyenlőtlenség igen általános érvényű, minden olyan valószínűségi változó ingadozási mértékére vonatkozik, amelyek várható értéke és szórása létezik. Éppen mert a tétel rendkívül általános érvényű, nagy pontosság az adott nagyságu eltérés valószínűségére vonatkozólag nem is várható tőle. Látni fogjuk, hogy pl. a gyakorlatban igen fontos normális eloszlás esetében a valószínűségi változó értékei sokkal szűkebb intervallumba tömörülnek a várható érték körül és pl. a $|\xi - M(\xi)| > 2\sigma$ eltérés valószínűsége, ha ξ normális eloszlású, kisebb, mint 0,05.

3.4 A valószínűségi változó momentumai

Az előző pontban a (3. 1. 6) és (3. 1. 7) összefüggések segítségével definiáltuk egy ξ valószínűségi változó valamely $\varphi(\xi)$ függvényének a várható értékét. A $\varphi(\xi)$ függvény speciális megválasztásával nyerjük a valószínűségi változó momentumait.

Legyen $\varphi(\xi) = \xi^k$, akkor az

$$(3.4.1) \quad \alpha_k = M(\xi^k)$$

várható értéket a ξ valószínűségi változó k -adik momentumának nevezzük. A várható érték ξ első momentuma. Legyen $\varphi(\xi) = [\xi - M(\xi)]^k$ ekkor

$$(3.4.2) \quad \mu_k = M\{[\xi - M(\xi)]^k\}$$

a ξ valószínűségi változó k -adik centrális momentuma. A valószínűségi változó szórásnégyzete nem más, mint a második centrális momentum:

$$D^2(\xi) = \mu_2 = M(\xi - M(\xi))^2 = M(\xi^2) - [M(\xi)]^2 = \alpha_2 - \alpha_1^2$$

Egyszerű számolással belátható, hogy a többi, magasabb rendű centrális momentumok is kifejezhetők a nem centrális momentumok segítségével. Pl.:

$$\begin{aligned} \mu_3 &= M(\xi - \alpha_1)^3 = M(\xi^3) - 3\alpha_1 M(\xi^2) + 3\alpha_1^2 M(\xi) - \alpha_1^3 = \\ &= \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3, \end{aligned}$$

$$\mu_4 = M(\xi - \alpha_1)^4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^2\alpha_2 - 3\alpha_1^4$$

A momentumok kiszámítása diszkrét eloszlású valószínűségi változók esetében az

$$\alpha_k = \sum_i x_i^k p_i$$

ill.

$$\mu_k = \sum_i (x_i - \alpha_1)^k p_i$$

formulák segítségével történik, míg folytonos eloszlású ξ valószínűségi változó esetén, ha ξ sűrűségfüggvénye $f(x)$,

$$\alpha_k = \int_{-\infty}^{\infty} x^k f(x) dx$$

ill.

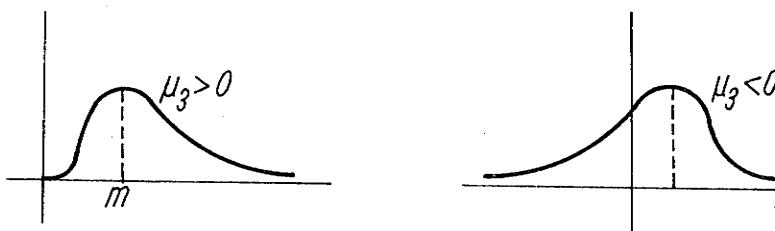
$$\mu_k = \int_{-\infty}^{\infty} (x - \alpha_1)^k f(x) dx$$

ahol $\alpha_1 = M(\xi)$.

Megemlítjük, hogy ha egy ξ valószínűségi változó eloszlása a várható értékére nézve szimmetrikus, akkor páratlan rendű centrális momentumai zérusok. Előfordulhat, hogy ξ eloszlása a várható értékre vonatkozólag aszimmetrikus, ferde. A ferdeség mérőszámául a

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

mennyiséget használják. (Általában folytonos eloszlás esetében használatos.)



35. ábra

A sűrűségfüggvény lapultságának mérésére használatos a

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

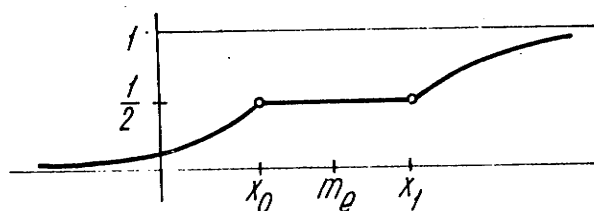
mennyiség. Normális eloszlás esetében $\gamma_2 = 0$. Ha valamely folytonos eloszlásra vonatkozólag $\gamma_2 > 0$, akkor az ill. eloszlás sűrűségfüggvénye magasabbra ugrik ki, "csucsosabb" mint a normális eloszlás sűrűségfüggvénye (lásd: 5.7 pont) míg $\gamma_2 < 0$ esetén lapultabb annál.

Folytonos eloszlások számszerű jellemzésére használatos mennyiség még a medián, jele m_e . A medián az az érték, amelynél a valószínűségi változó $\frac{1}{2}$ valószínűséggel vesz fel kisebb értéket és ugyancsak $\frac{1}{2}$ valószínűséggel vesz fel nagyobb értéket, vagyis amelyre

$$F(m_e) = \frac{1}{2}$$

Abban az esetben, ha az $F(x) = \frac{1}{2}$ egyenletnek több megoldása van, vagyis ha az $F(x)$ eloszlásfüggvény az x_0 pontban eléri az $\frac{1}{2}$ értéket és állandó marad valamely pontig és csak ettől kezdve emelkedik, akkor

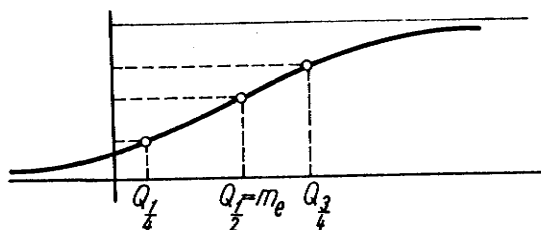
$$m_e = \frac{x_0 + x_1}{2}$$



36. ábra

Az $F(x) = p$ egyenlet megoldását (tételezzük fel, hogy egy megoldása van), a valószínűségeloszlás p -kvantilisének nevezzük és Q_p -vel jelöljük.
Nyilván

$$Q_{\frac{1}{2}} = m_e$$



37. ábra

A $Q_{1/2}$ ill. $Q_{3/4}$ kvantiliseket alsó- ill. felső kvartiliseknek nevezzük.

3.5 A korrelációs együttható

Két vagy több valószínűségi változó együttes eloszlásával kapcsolatban is szokás különböző számszerű jellemzőket alkalmazni, amelyek közül kiemeljük a kovariancia és a korrelációs együttható fogalmát. Ha valamely kísérlettel kapcsolatban két véletlen mennyiséget, ξ -t és η -t egyidejűleg kell tekintetbe vennünk, akkor elsősorban az szokott érdekelni bennünket, hogy függetlenek-e egymástól a kérdéses valószínűségi változók, vagy pedig értékeik között valamiféle függvénykapcsolat áll fenn. Gyakori eset, hogy ξ és η értékei között nincs szigorú függvénykapcsolat, de felvett értékeik között bizonyos tendencia mutatkozik, pl. hogy ξ nagy értékei η nagy értékeivel, míg ξ kisebb értékei általában η kisebb értékeivel szoktak együtt jelentkezni.

Ha pl. ξ a Duna vízállása adott pontban mérve, η pedig a talajvízszint egy Dunához közeli pontban, akkor ha ξ és η között nincs is szigorú függvénykapcsolat, de közös tendencia, sztochasztikus kapcsolat tapasztalható közöttük. Ezét a sztochasztikus kapcsolatot igyekszünk numerikus formában kifejezni a kovariancia, ill. a korrelációs együttható segítségével.

Legyenek ξ és η valószínűségi változók. Ha ξ és η függetlenek, akkor a (3. 1. 12) összefüggés alapján

$$M[(\xi - M(\xi)) \cdot (\eta - M(\eta))] = M[\xi - M(\xi)] \cdot M[\eta - M(\eta)] = 0.$$

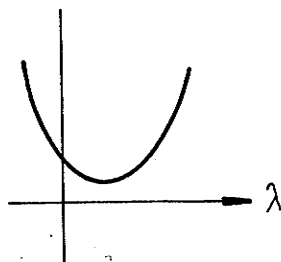
Ha ξ és η nem függetlenek, akkor a fenti várható érték valamilyen c számmal egyenlő. A

$$(3.5.1) \quad c = M[(\xi - M(\xi))(\eta - M(\eta))]$$

mennyiséget a ξ és η valószínűségi változók kovarianciájának nevezzük. Ha $\xi = \eta$, akkor $c = M[\xi - M(\xi)]^2 = D^2(\xi)$ a ξ változó szórásnégyzete. A c kovariancia értéke, attól függően, hogy milyen eloszlású valószínűségi változókra van szó, tetszőleges valós szám lehet, tehát c numerikus értékéből a két változó közötti sztochasztikus kapcsolat szorosságára vonatkozólag nehéz lenne következtetést levonni. Célszerűbb olyan mértékszámot választani, amely csak valamilyen rögzített határok közé eshet, s amelynek numerikus értéke utalni fog a két változó közötti kapcsolat jellegére, szoros vagy laza voltára. Ilyen mérőszám a korrelációs együttható, amelyet a

$$(3.5.2) \quad \varrho = \frac{M[(\xi - M(\xi))(\eta - M(\eta))]}{D(\xi) D(\eta)}$$

formulával értelmezzük. A korrelációs együttható tehát a ξ és η változók kovarianciája osztva a szórásaik szorzatával. A ϱ korrelációs együttható



38. ábra

értéke -1 és $+1$ között változik, mivel a számlálóban szereplő kovariancia nem lehet nagyobb, mint a szórások szorzata. Ezt az állítást az alábbi tétel felhasználásával bizonyítjuk:

Ha ξ és η olyan valószínűségi változók, amelyek négyzetének várható értéke létezik, akkor

$$(3.5.3) \quad |M(\xi\eta)| \leq \sqrt{M(\xi^2) M(\eta^2)}$$

Ugyanis tetszőleges valós λ -ra $M(\lambda\xi - \eta)^2 \geq 0$

azaz

$$\lambda^2 M(\xi^2) - 2\lambda M(\xi\eta) + M(\eta^2) \geq 0$$

Ez a kifejezés λ -ban másodfokú polinom, amely legfeljebb érinti a λ -tengelyt, azaz diszkriminánsa nem pozitív, tehát

$$4[M(\xi\eta)]^2 - 4M(\xi^2)M(\eta^2) \leq 0,$$

amiből az állítás már leolvasható.

Ha most a (3.5.3) összefüggésben ξ helyébe $\xi - M(\xi)$ -t, η helyébe $\eta - M(\eta)$ -t írunk, akkor az

$$\begin{aligned} M[(\xi - M(\xi))(\eta - M(\eta))] &\leq \sqrt{M[\xi - M(\xi)]^2} \sqrt{M[\eta - M(\eta)]^2} = \\ &= D(\xi) D(\eta) \end{aligned}$$

összefüggést nyerjük, amiből következik, hogy $|\varrho| \leq 1$.

Abban az esetben, ha az η valószínűségi változó a ξ változónak lineáris függvénye: $\eta = a\xi + b$, akkor $\varrho = 1$ vagy $\varrho = -1$, aszerint, hogy $a > 0$ vagy $a < 0$. Ennek belátása egyszerű:

$$\begin{aligned} \varrho &= \frac{M[(\xi - M(\xi))(a\xi + b - aM(\xi) - b)]}{\sqrt{D^2(\xi) D^2(a\xi + b)}} = \\ &= \frac{a M[\xi - M(\xi)]^2}{|a| D^2(\xi)} = \frac{a}{|a|} = \pm 1. \end{aligned}$$

Az állítás megfordítása is igaz, ha $|\varrho| = 1$, akkor egyik valószínűségi változó a másik változónak lineáris függvénye.

Gyakorlati számítás céljára mind a kovariancia, mind a korreláció együtttható képlete némileg egyszerűsíthető:

$$\begin{aligned}
 (3.5.4) \quad c &= M[(\xi - M(\xi)) (\eta - M(\eta))] = \\
 &= M[\xi \eta - M(\xi) \eta - M(\eta) \xi + M(\xi) M(\eta)] = \\
 &= M(\xi \eta) - M(\xi) M(\eta)
 \end{aligned}$$

innen:

$$(3.5.5) \quad \varrho = \frac{M(\xi \eta) - M(\xi) M(\eta)}{D(\xi) D(\eta)}$$

Ahhoz, hogy a ξ és η valószínűségi változók közötti kovarianciát és korrelációt ki tudjuk számítani, ismernünk kell az együttes eloszlásukat.

Ha ξ és η diszkrét eloszlásúak:

$$\begin{aligned}
 (3.5.6) \quad c &= \sum_i \sum_k [x_i - M(\xi)] [y_k - M(\eta)] r_{ik} = \\
 &= \sum_i \sum_k x_i y_k r_{ik} - M(\xi) M(\eta).
 \end{aligned}$$

Ha ξ és η folytonos eloszlásúak $h(x, y)$ együttes sűrűségfüggvénnyel, akkor

$$(3.5.7) \quad c = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy h(x, y) dx dy - M(\xi) M(\eta)$$

itt alkalmaztuk a (3.5.4) összefüggést.

Abban az esetben, ha ξ és η függetlenek, a közöttük levő kovariancia zérus. Ugyanis mivel független valószínűségi változók szorzatának várható értéke várható értékek szorzatával egyenlő

$$c = M(\xi \eta) - M(\xi) M(\eta) = M(\xi) M(\eta) - M(\xi) M(\eta) = 0.$$

Ennek következtében ha ξ és η függetlenek

$$\varrho = \frac{c}{D(\xi) D(\eta)} = 0.$$

Az állítás megfordítása általában nem igaz, abból, hogy a ξ és η közötti korreláció zérus, általában nem következik, hogy függetlenek. A korrelációs együttjáró nem annyira a sztochasztikus kapcsolat szorosságát, mint a kapcsolat linearitását méri!

Ha ξ és η valószínűségi változók közötti korrelációs együttható: $\rho = 0$, akkor a ξ és η változókat korrelátlannak mondjuk. Mint említettük két valószínűségi változó korrelátlansága általában nem jelent függetlenséget, ha azonban ξ és η együttes eloszlása kétváltozós normális eloszlás, akkor a korrelátlanságból következik a változók függetlensége. (Lásd: 5.11. pont.)

A GENERÁTORFÜGGVÉNY ÉS A KARAKTERISZTIKUS FÜGGVÉNY

4.1 A generátorfüggvény

Ismertetünk két igen hasznos analízisbeli segédesszközt, a generátorfüggvényt és a karakterisztikus függvényt, amelyek segítségével a valószínűségi változók momentumai egyszerűen kiszámíthatók, továbbá független valószínűségi változók összegének eloszlása egyszerűen meghatározható. Ugyancsak hasznosnak bizonyulnak, sőt szinte nélkülözhetetlenek ezek a függvények különböző határeloszlások meghatározásánál.

A generátorfüggvényt egész értékű valószínűségi változók eloszlásának vizsgálatára használjuk.

Legyen ξ valószínűségi változó, amelynek lehetséges értékei a

0, 1, 2, ..., n, ... számok, amelyeket rendre
 $p_0, p_1, p_2, \dots, p_n, \dots$ valószínűségekkel vesz fel.

Képezzük a

$$(4.1.1) \quad G(x) = \sum_k p_k x^k$$

függvényt, amelyet a ξ valószínűségi változó generátorfüggvényének nevezünk. Mivel a $\{p_k\}$ számok valószínűségeloszlást képeznek azaz

$\sum_k p_k = 1$, a generátorfüggvényt értelmező hatványsor (amennyiben k megszámlálható sok értéket fut be) $x = 1$ esetén konvergens és $G(1) = 1$. Ebből következik, hogy $|x| < 1$ esetén a $G(x)$ függvény akárhányszor differenciálható és

$$G'(x) = \sum_k k p_k x^{k-1}$$

Ha ez a hatványsor az $x = 1$ helyen is konvergens, akkor

$$(4.1.2) \quad \begin{aligned} G'(1) &= \sum_k k p_k = M(\xi) \\ G''(x) &= \sum_k k(k-1) p_k x^{k-2} \\ G''(1) &= \sum_k k^2 p_k - \sum_k k p_k = M(\xi^2) - M(\xi) \text{ stb.} \end{aligned}$$

Innen látható, hogy a ξ egész értékű valószínűségi változó várható értéke:

$$M(\xi) = G'(1)$$

szórásnégyzete:

$$(4.1.3) \quad D^2(\xi) = G''(1) + G'(1) - [G'(1)]^2$$

Vegyük észre, hogy a

$$G(x) = \sum_k x^k p_k$$

generátorfüggvény maga is egy várható érték, mégpedig az x^ξ valószínűségi változó várható értéke a várható érték definíciója szerint, azaz

$$M(x^\xi) = \sum_k x^k p_k = G(x).$$

Ebből az észrevételből fontos tételt származtatunk független valószínűségi változók összegének generátorfüggvényére vonatkozólag. Legyenek $\xi_1, \xi_2, \dots, \xi_n$ függetlenek, egész értékű valószínűségi változók és legyen $\eta = \xi_1 + \xi_2 + \dots + \xi_n$. Ekkor az η valószínűségi változó generátorfüggvénye:

$$\begin{aligned} (4.1.4) \quad G_\eta(x) &= M(x^\eta) = M(x^{\xi_1 + \xi_2 + \dots + \xi_n}) = \\ &= M(x^{\xi_1}) M(x^{\xi_2}) \dots M(x^{\xi_n}) = \\ &= G_{\xi_1}(x) \cdot G_{\xi_2}(x) \dots G_{\xi_n}(x) \end{aligned}$$

Ugyanis a $\xi_1, \xi_2, \dots, \xi_n$ változók függetlensége miatt az $x^{\xi_1}, x^{\xi_2}, \dots, x^{\xi_n}$ változók is függetlenek és független valószínűségi változók szorzatának várható értéke a várható értékek szorzatával egyenlő.

Látjuk tehát, hogy független valószínűségi változók összegének generátorfüggvénye az egyes változók generátorfüggvényeinek szorzatával egyenlő. Ezt a tételt az egyes nevezetesebb valószínűségeloszlások tárgyalása során alkalmazni fogjuk független változók összege eloszlásának meghatározására, mert generátorfüggvények segítségével általában sokkal egyszerűbb az összeg eloszlását meghatározni, mint a konvolúciós szabály alkalmazásával, amely inkább elméleti jelentőségű.

Az alkalmazásokban gyakran előfordul, hogy független, azonos eloszlású valószínűségi változók összegének eloszlását kell meghatároznunk.

Ekkor a $G_{\eta}(x)$ a következő alakot ölti (ha n tagú összegről van szó)

$$(4.1.5) \quad G_{\eta}(x) = G_{\xi_1 + \xi_2 + \dots + \xi_n}(x) = \left[G_{\xi_1}(x) \right]^n$$

A $G_{\eta}(x)$ generátorfüggvény explicit kifejtésében x^k együtthatója adja meg a $P(\eta = k)$ valószínűséget.

4.2 A karakterisztikus függvény

A generátorfüggvény csak nemnegatív egész értékű valószínűségi változók esetére van értelmezve. Tetszőleges valószínűségi változók esetében a generátorfüggvényhez hasonló szerepet játszik az ún. karakterisztikus függvény. Egy ξ valószínűségi változó karakterisztikus függvényén az $e^{it\xi}$ komplex értékű valószínűségi változó várható értékét értjük. Jelöljük a ξ valószínűségi változó karakterisztikus függvényét $\varphi_{\xi}(t)$ -vel, ekkor:

$$\varphi_{\xi}(t) = M(e^{it\xi}) = M(\cos t\xi) + i M(\sin t\xi) \quad (i = \sqrt{-1}).$$

Tárgyalásaink során a karakterisztikus függvényt csak folytonos eloszlású valószínűségi változókkal kapcsolatban fogjuk használni. Ez esetben:

$$(4.2.1) \quad \varphi_{\xi}(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

ahol $f(x)$ a ξ valószínűségi változó sűrűségfüggvénye. A karakterisztikus függvényt az analízisben Fourier-transzformációnak nevezik.

A karakterisztikus függvény néhány fontosabb tulajdonsága:

$$(4.2.2) \quad |\varphi_{\xi}(t)| = \left| \int_{-\infty}^{\infty} e^{itx} f(x) dx \right| \leq \int_{-\infty}^{\infty} |e^{itx}| f(x) dx = 1$$

(mivel $|e^{itx}| = \sqrt{\cos^2 tx + \sin^2 tx} = \sqrt{1} = 1$).

Megjegyezzük, hogy

$$(4.2.3) \quad \begin{aligned} \varphi(-t) &= M(e^{-it\xi}) = M(\cos(-t\xi)) + i M(\sin(-t\xi)) = \\ &= M(\cos t\xi) - i M(\sin t\xi) = \overline{\varphi(t)}. \end{aligned}$$

A Fourier-transzformáció elmélete szerint ha

$$\int_{-\infty}^{\infty} |\varphi(t)| dt < \infty,$$

akkor

$$(4.2.4) \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt,$$

azaz a karakterisztikus függvény ismeretében a sűrűségfüggvény meghatározható. A karakterisztikus függvény egyértelműen meghatározza a valószínűségeloszlást. Tárgyalásaink során a (4.2.4) inverziós formula alkalmazására sem lesz szükségünk, mert csak néhány folytonos eloszlás karakterisztikus függvényét fogjuk meghatározni, s az ismert karakterisztikus függvényből közvetlenül tudni fogjuk, hogy milyen eloszlásról van szó.

Ismertetünk most a karakterisztikus függvényre vonatkozó fontos tételt, amelynek alapján független valószínűségi változók összegének valószínűségeloszlása meghatározható.

Tétel: Legyenek $\xi_1, \xi_2, \dots, \xi_n$ független valószínűségi változók és legyen $\eta = \xi_1 + \xi_2 + \dots + \xi_n$. Ekkor az η valószínűségi változó karakterisztikus függvények az összeadandók karakterisztikus függvényeinek szorzatával egyenlő, azaz

$$\begin{aligned} (4.2.5) \quad \varphi_{\eta}(t) &= M(e^{it\eta}) = M(e^{it(\xi_1 + \xi_2 + \dots + \xi_n)}) = \\ &= M(e^{it\xi_1} e^{it\xi_2} \dots e^{it\xi_n}) = \\ &= M(e^{it\xi_1}) M(e^{it\xi_2}) \dots M(e^{it\xi_n}) = \\ &= \varphi_{\xi_1}(t) \varphi_{\xi_2}(t) \dots \varphi_{\xi_n}(t). \end{aligned}$$

Ugyanis a $\xi_1, \xi_2, \dots, \xi_n$ változók függetlensége folytán $e^{it\xi_1}, e^{it\xi_2}, \dots, e^{it\xi_n}$ valószínűségi változók is függetlenek és szorzatuk várható értéke várható értékeik szorzatával egyenlő.

Megemlítjük még, hogy ha egy η valószínűségi változó egy másik valószínűségi változónak lineáris függvénye,

$$\eta = a\xi + b$$

akkor

$$(4.2.6) \quad \varphi_{\eta}(t) = M(e^{it\eta}) = M(e^{it(a\xi + b)}) = \\ = e^{itb} M(e^{i(at)\xi}) = e^{itb} \varphi_{\xi}(at).$$

Megmutatjuk most, hogyan lehet a karakterisztikus függvény segítségével meghatározni egy folytonos eloszlású valószínűségi változó momentumait, amennyiben azok léteznek.

A

$$\varphi_{\xi}(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

kifejezést t -szerint differenciálva:

$$\varphi'_{\xi}(t) = \int_{-\infty}^{\infty} i x e^{itx} f(x) dx$$

Az $\eta = e^{it\xi} = \cos t\xi + i \sin t\xi = \xi_1 + i \xi_2$ komplex értékű valószínűségi változó várható értékét az $M(\eta) = M(\xi_1) + i M(\xi_2)$, szórásnégyzetét a $D^2(\eta) = M(|\eta - M(\eta)|^2)$ formulával értelmezzük,

és

$$\varphi'_{\xi}(0) = i \int_{-\infty}^{\infty} f(x) x dx = i \alpha_1 = i M(\xi)$$

Továbbá:

$$\varphi''_{\xi}(t) = \int_{-\infty}^{\infty} (ix)^2 e^{itx} f(x) dx$$

és

$$\varphi''_{\xi}(0) = i^2 \int_{-\infty}^{\infty} x^2 f(x) dx = i^2 \alpha_2 = i^2 M(\xi^2)$$

Innen:

$$(4.2.7) \quad M(\xi) = \frac{1}{i} \varphi'_{\xi}(0)$$

$$(4.2.8) \quad D^2(\xi) = M(\xi^2) - [M(\xi)]^2 = \\ = \frac{1}{i^2} \varphi''_{\xi}(0) - \left[\frac{1}{i} \varphi'_{\xi}(0) \right]^2 = [\varphi'_{\xi}(0)]^2 - \varphi''_{\xi}(0).$$

Általában:

$$\varphi_{\xi}^{(k)}(t) = i^k \int_{-\infty}^{\infty} x^k e^{itx} f(x) dx$$

$$(4.2.9) \quad \varphi_{\xi}^{(k)}(0) = i^k \int_{-\infty}^{\infty} x^k f(x) dx = i^k \alpha_k$$

ahol α_k a ξ valószínűségi változó k -adik momentuma. Ennek alapján a $\varphi_{\xi}(t)$ függvény Taylor-sorfejtése a következő alakban írható:

$$(4.2.10) \quad \varphi_{\xi}(t) = \varphi_{\xi}(0) + \frac{\varphi'_{\xi}(0)}{1!} t + \frac{\varphi''_{\xi}(0)}{2!} t^2 + \dots = 1 + \alpha_1 it + \alpha_2 \frac{(it)^2}{2!} t \dots$$

FONTOSABB VALÓSZÍNŰSÉGELOSZLÁSOK

A) DISZKRÉT VALÓSZÍNŰSÉGELOSZLÁSOK

5.1 Egyszerű alternatíva

Egy esemény karakterisztikus változója. A legegyszerűbb diszkrét eloszlás az egyszerű alternatíva. Ilyen eloszlás lép fel, valahányszor egy kísérlettel kapcsolatban csak az érdekel bennünket, hogy valamely A esemény bekövetkezik-e vagy sem. Ha pl. a kísérlet egy adott helyen a talajvízszint mérése és csak az érdekel bennünket, hogy a talajvízszint mérése helyén kisebb-e vagy nagyobb mint 3 m, akkor jelöljük A-val azt az eseményt, hogy a vízszint kisebb mint 3 m, $\bar{A}$ -sal pedig az, hogy 3 m-nél nagyobb a vízszint. A kísérlet eredményeként vagy A vagy $\bar{A}$ következik be. Tegyük fel, hogy nagyszámu mérés alapján ismerjük a $P(A) = p$, $P(\bar{A}) = 1 - p = q$ valószínűségeket.

Rendeljük most hozzá kísérletünkhöz a ξ valószínűségi változót, amely az 1 értéket veszi fel, ha az A esemény következik be és vegye fel ξ a 0 értéket, ha az $\bar{A}$ esemény következett be. Ekkor tehát $P(\xi = 1) = p$, $P(\xi = 0) = q = 1 - p$. A ξ változó eloszlását a következő sémával szokás megadni:

$$\xi : \begin{array}{c|c} 1 & 0 \\ \hline p & q \end{array}$$

Hasonlóan egyszerű alternatívával állunk szemben, ha kísérlet abban áll, hogy megfigyeljük és regisztráljuk, hogy egy adott napon adott helyen esett az eső vagy sem (vagy hogy a lehullott csapadék meghaladott egy adott szintet vagy sem).

Alternatíva, hogy egy megvizsgált ércdarab tartalmaz vasat vagy sem, hogy egy megvizsgált gyártmány selejtes vagy sem, hogy egy forintot feldobva irást dobunk vagy sem stb.

Az olyan ξ valószínűségi változót, amely az 1 értéket veszi fel, ha egy kísérlet során a számunkra érdekes A esemény bekövetkezett, míg $\xi = 0$, ha az $\bar{A}$ esemény következett be, az esemény karakterisztikus változójának vagy indikátorának nevezzük. Ekkor tehát:

$$P(A) = P(\xi = 1) = p, \quad P(\bar{A}) = P(\xi = 0) = 1 - p = q.$$

Számítsuk ki a karakterisztikus változó várható értékét és szórás-négyzetét. A várható érték definíciója és ξ eloszlása alapján:

$$(5.1.1) \quad M(\xi) = 1 \cdot p + 0 \cdot q = p$$

$$M(\xi^2) = 1^2 \cdot p + 0^2 \cdot q = p$$

$$(5.1.2) \quad D^2(\xi) = M(\xi^2) - [M(\xi)]^2 = p - p^2 = p(1-p) = p \cdot q.$$

A karakterisztikus változó generátorfüggvénye:

$$(5.1.3) \quad G(x) = \sum_k p_k x^k = q x^0 + p x = p x + q.$$

5.2 A binomiális vagy Bernoulli-eloszlás

Legyen egy egyszerű alternatíva kísérletnél a bennünket érdeklő A esemény valószínűsége $P(A) = p$, (ekkor nyilván $P(\bar{A}) = 1 - p = q$).

Ismételjük meg n -szer a kísérletet (azonos körülmények között) és jelölje a ξ valószínűségi változó az A esemény bekövetkezéseinek számát az n kísérlet során. ξ lehetséges értékei a $0, 1, 2, \dots, n$ számok.

Mi a valószínűsége, hogy a ξ változó éppen a k értéket veszi fel? ($k = 0, 1, 2, \dots, n$). Az egyes kísérleteket függetleneknek tekintjük. A problémát úgy is interpretálhatjuk, hogy egy urnában csak fehér és fekete golyók vannak, a fehér golyók aránya az összes golyók számához p , a fekete golyóké q . Annak valószínűsége, hogy n -szer huzva visszatevéssel k -szor huzunk fehér golyót, az (1. 5. 1) formula alapján:

$$(5.2.1) \quad p_k = P(\xi = k) = \binom{n}{k} p^k q^{n-k}.$$

Ezt a valószínűséget a következő módon is megkaphatjuk. A valószínűségi változó nyilván

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n$$

ahol ξ_i az i -edik kísérlet karakterisztikus változója.

Há pl. ξ egy adott év májusában az esős napok számát jelöli és $\xi_1, \xi_2, \dots, \xi_n$ az egyes napok karakterisztikus változói, amelyek tehát az 1 vagy 0 értéket vesznek fel aszerint, hogy az illető napon esett az eső vagy sem, akkor értéke nyilván annyi, ahány 1-es van az összeadandók között:

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n = 1 + 0 + 1 + \dots + 0.$$

A binomiális eloszlású ξ valószínűségi változó várható értéke:

$$(5.2.2) \quad M(\xi) = M(\xi_1) + M(\xi_2) + \dots + M(\xi_n) = np,$$

mivel a ξ_i karakterisztikus változók mindegyikének várható értéke p .
A binomiális eloszlású ξ valószínűségi változó szórásnégyzete a (3.3.10) összefüggés alapján:

$$(5.2.3) \quad D^2(\xi) = D^2(\xi_1) + D^2(\xi_2) + \dots + D^2(\xi_n) = \\ = pq + pq + \dots + pq = npq.$$

Tehát a binomiális eloszlás szórása:

$$(5.2.4) \quad D(\xi) = \sqrt{npq}.$$

A ξ változó generátorfüggvénye a (4.1.5) összefüggés szerint:

$$(5.2.5) \quad G_\xi(x) = \sum_{k=0}^n P_k x^k = \prod_{i=1}^n G_{\xi_i}(x) = (px+q)^n = \\ = \sum_{k=0}^n \binom{n}{k} (px)^k q^{n-k} = (px+q)^n$$

A $G_\xi(x)$ generátorfüggvényben x^k együtthatója éppen a $P_k = P(\xi = k)$ valószínűséget szolgáltatja, ez pedig

$$P_k = \binom{n}{k} p^k q^{n-k}$$

Mind a várható érték, mind a szórásnégyzet egyszerűen kiszámítható a

$$G_\xi(x) = (px+q)^n$$

generátorfüggvény segítségével is, ugyanis a és formulák alapján:

$$M(\xi) = G'_\xi(1) = \left[n(px+q)^{n-1} p \right]_{x=1} = np$$

$$\begin{aligned}
 D^2(\xi) &= G_{\xi}''(1) + G_{\xi}'(1) - [G_{\xi}'(1)]^2 = \\
 &= \left[n(n-1) p^2 (px+q)^{n-2} \right]_{x=1} + np - n^2 p^2 = n^2 p^2 - np^2 + np - n^2 p^2 = \\
 &= np(1-p) = npq.
 \end{aligned}$$

A binomiális eloszlású ξ valószínűségi változó eloszlásának sémája:

0	1	2	...	k	...	n
$P_0 = \binom{n}{0} p^0 q^n$	$P_1 = \binom{n}{1} p q^{n-1}$	$P_2 = \binom{n}{2} p^2 q^{n-2}$	...	$P_k = \binom{n}{k} p^k q^{n-k}$	...	$P_n = \binom{n}{n} p^n q^0$

Felmerül a kérdés, hogy adott p esetén a ξ valószínűségi változó, melyik k értéket veszi fel legnagyobb valószínűséggel, azaz melyik P_k valószínűség a legnagyobb?

A P_k számok sorozata addig nő, amíg

$$\frac{P_k}{P_{k-1}} \geq 1,$$

azaz amíg

$$\frac{\binom{n}{k} p^k q^{n-k}}{\binom{n}{k-1} p^{k-1} q^{n-k+1}} = \frac{n-k+1}{k} \cdot \frac{p}{1-p} \geq 1,$$

vagyis amíg

$$(n+1)p \geq k.$$

Amennyiben $(n+1)p$ egész szám, az eloszlás maximális tagja a $k = (n+1)p$ indexű P_k valószínűség. Ebben az esetben $P_k = P_{k-1}$, azaz két egyenlő nagyságú maximális valószínűség létezik. Ha $(n+1)p$ nem egész szám, akkor a P_k valószínűség olyan k -ra maximális, amely az $(n+1)p$ -ben foglalt legnagyobb egész számmal egyenlő. Ha n nagy szám, akkor mivel $p \leq 1$, jóformán mindegy, hogy $(n+1)p$ -t vagy np -t veszünk. Azt mondhatjuk tehát, hogy ha egy kísérlet során valamely A esemény p valószínűséggel következik be, akkor n kísérlet során az A esemény legnagyobb valószínűséggel np -szer, azaz a kísérletek p -ed részében következik be. Ez azt jelenti, hogy binomiális eloszlású ξ valószínűségi változó legnagyobb valószínűségi értéke nagyjából a várható értékével egyezik meg.

Egy példával szemléltetjük a mondottakat.

Budapesten a csapadékos napok megoszlása az utóbbi 50 év megfigyelési adatainak átlaga alapján (csapadékos napnak számítva azon napokat, amikor legalább 1 mm csapadék volt):

Hó	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
Csapadékos nap	7,6	6,8	7,3	7,4	8,5	8,0	6,5	6,3	6,2	7,5	8,8	9,1

Ennek alapján pl. annak valószínűsége, hogy egy adott év áprilisában

7 csapadékos napot észlelünk $p \approx \frac{1}{4}$ mellett:

$$p_7 = \binom{30}{7} \left(\frac{1}{4}\right)^7 \left(\frac{3}{4}\right)^{23} = 2\,035\,800 \cdot \frac{3^{23}}{4^{30}} \approx 0,16$$

Az ilyen valószínűségek kiszámítása a binomiális együtthatók miatt kellemtelen, ezért közelíteni szoktuk. Lásd: (Jelen példánkban $\binom{30}{7}$ értéke megtalálható a binomiális együtthatók táblázatában, a $3^{23} \cdot 4^{-30}$ értékét pedig logaritmussal határozhatjuk meg.)

A fentiekből látható, hogy a binomiális eloszlású valószínűségi változó még a legnagyobb valószínűségi értékét is viszonylag kis valószínűséggel veszi fel. Más a helyzet, ha azt kérdezzük mennyi a valószínűsége, hogy az áprilisi esős napok száma mondjuk 5 és 11 közé esik. Ennek valószínűsége:

$$P = \sum_{k=5}^{10} \binom{30}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{30-k}$$

E valószínűség kiszámítása igen munkaigényes lenne, ezért közelítjük. Lásd: (5.10) pontot.

5.3 A polinomiális eloszlás

Ha egy kísérletnek S számú különböző kimenetele lehet, amelyeket jelöljünk $A_1, A_2, \dots, A_s$ -sel (más szóval az $A_1, A_2, \dots, A_s$ események teljes eseményrendszert alkotnak), továbbá $P(A_1) = p_1$,

$P(A_2) = p_2, \dots, P(A_s) = p_s, \left(\sum_{i=1}^s p_i = 1\right)$, akkor a kísérletet n -szer

egymás után megismételve az un. polinomiális eloszláshoz jutunk. Ez a fogalom lényegében a binomiális eloszlás általánosítása és $k = 2$ esetén megegyezik vele.

Annak valószínűsége, hogy az A_1 esemény k_1 -szer, az A_2 esemény k_2 -szer, az A_3 esemény k_3 -szor következik be ($k_1 + k_2 + \dots + k_s = n$):

$$(5.3.1) \quad P_{k_1, k_2, \dots, k_s} = \frac{n!}{k_1! k_2! \dots k_s!} p_1^{k_1} p_2^{k_2} \dots p_s^{k_s}.$$

Ha pl. a kísérlet egy 10 részre osztott közös céltáblára leadott lövés, és $p_1, p_2, \dots, p_{10}$ jelöli annak valószínűségét, hogy a találat az 1-es, 2-es, ..., 10-es körbe esik, akkor annak valószínűségét, hogy pontosan k_1 számú 1-es, k_2 számú 2-es, ..., k_{10} számú 10-es találatot érjünk el n lövésből ($k_1 + k_2 + \dots + k_{10} = n$), az (5.3.1) formula szolgáltatja.

5.4 A geometriai eloszlás

Tekintsünk egy olyan kísérletet, amelynek két lehetséges kimenetele van, tehát egyszerű alternatíva, s jelöljük az egyik kimenetet A -val, a másikat $\bar{A}$ -sal. Legyen $P(A) = p$, $P(\bar{A}) = q = 1 - p$.

Tegyük fel, hogy a kísérletet többször egymás után megismételjük mindaddig, amíg az A esemény először bekövetkezik. Könnyebb megértés kedvéért képzeljük el, hogy ipari minőségellenőrzést végzünk. A -val jelöljük azt az eseményt, hogy egy-egy gyártmányt selejtesnek, $\bar{A}$ -sal, hogy megfelelőnek találjuk. Az egész gyártmánytömegben a selejtarány p . Mi a valószínűsége, hogy először a k -adik megvizsgált gyártmány bizonyul selejtesnek? Az első selejtes darab megjelenéséig megvizsgálandó gyártmányok száma nyilván valószínűségi változó, s jelöljük ξ -vel. A $\{\xi = k\}$ esemény akkor következik be, ha az első $k-1$ megvizsgált gyártmány nem selejtes és a k -adik selejtes, ezért

$$(5.4.1) \quad P_k = P(\xi = k) = q^{k-1} \cdot p.$$

Könnyű belátni, hogy $\sum_{k=1}^{\infty} P_k = 1$, ugyanis

$$\sum_{k=1}^{\infty} P_k = p \sum_{k=1}^{\infty} q^{k-1} = \frac{p}{1-q} = 1.$$

Ugyancsak az (5.4.1) formulával számítható ki annak valószínűsége, hogy ha egy célra lövéseket adunk le p találati valószínűséggel, akkor a k -adik lövésre érjük el az első találatot.

A geometriai eloszlás generátorfüggvénye:

$$(5.4.2) \quad G(x) = \sum_{k=1}^{\infty} pq^{k-1} x^k = px \sum_{k=1}^{\infty} (qx)^{k-1} = \frac{px}{1-qx}.$$

Mivel

$$G'(x) = \frac{p(1-qx) + pqx}{(1-qx)^2} = \frac{p}{(1-qx)^2},$$

és

$$G''(x) = \frac{2pq(1-qx)}{(1-qx)^4} = \frac{2pq - 2pq^2x}{(1-qx)^4},$$

a geometriai eloszlás várható értékére és szórásnégyzetére a következő formulák adódnak:

$$(5.4.3) \quad M(\xi) = G'(1) = \frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p};$$

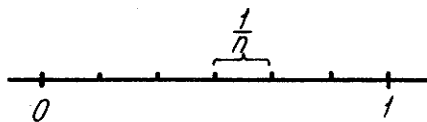
$$(5.4.4) \quad D^2(\xi) = G''(1) + G'(1) - [G'(1)]^2 = \frac{2pq}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}.$$

5.5 A Poisson-eloszlás

A diszkrét valószínűség eloszlások között a binomiális eloszlás mellett kiemelkedő elméleti és gyakorlati jelentősége van az ún. Poisson-eloszlásnak, amely legegyszerűbben a binomiális eloszlásból származtatható, annak határeseteként adódik, ha a kísérletek száma, az n nagy szám, a p pedig (a bennünket érdeklő esemény bekövetkezésének valószínűsége egy kísérlet során) kicsi. A Poisson-eloszlást a kis valószínűségű, vagyis ritka események eloszlási törvényének is szokás nevezni. A Poisson-eloszlás jelentősége azonban korántsem csak az, hogy vele a binomiális eloszlás közelíthető.

Számos természeti folyamatnál a Poisson-eloszlás a véletlen ingadozások törvényszerűségeként lép fel. Az ún. véletlen eseményfolyamatok, pl. adott időintervallumban egy telefonközpontba befutó hívások száma, bizonyos idő alatt elbomlott rádióaktív atomok mennyisége, egy meghatározott időtartam alatt bekövetkezett közlekedési balesetek száma stb., mind a Poisson-eloszlást követik. Ugyancsak Poisson-eloszlással találkozunk a véletlen pontelhelyezkedések esetében bizonyos körülmények között, amit példákon is szemléltetni fogunk.

Álljon a kísérletünk abból, hogy megszabott időn át megfigyeljük, vajon hány jel érkezik egy regisztráló berendezésbe. Válasszunk egy bizonyos megfigyelési időintervallumot egységnyi hosszúságúnak, és osszuk fel n egyenlő részre:



39. ábra

Jelöljük p -vel annak a valószínűségét, hogy egy $\frac{1}{n}$ hosszúságú részintervallumba érkezik egy jel. Feltesszük, hogy bármelyik részintervallumba egyforma a valószínűsége egy jel érkezésének. További feltételezésünk: n elég nagy értéke miatt nagyon kicsi annak a valószínűsége, hogy egy $\frac{1}{n}$ hosszúságú kis intervallumba egynél több jel érkezzék. Nyilván minél több részre osztjuk az egységül választott intervallumot, vagyis minél nagyobb az n értéke, annál kisebb annak a valószínűsége, hogy egy $\frac{1}{n}$ hosszúságú intervallumba érkezzék jel, azaz annál kisebb a p értéke. Ha tehát n értéke nagy, akkor p értéke kicsi.

Tegyük fel, hogy p értékének csökkenése arányos n növekedésével, azaz $np = \lambda$ állandó.

Számítsuk ki mennyi a valószínűsége, hogy az egységül választott intervallumba pontosan k számú jel érkezzék, ha n értéke nagy, p értéke pedig kicsi.

Az, hogy egy $\frac{1}{n}$ hosszúságú időintervallumba érkezik-e jel vagy sem, egyszerű alternatíva, s mivel n darab $\frac{1}{n}$ hosszúságú részintervallumra bontottuk az egységül választott intervallumot, n -szer ismételt alternatívával, vagyis binomiális eloszlással állunk szemben:

$$P_k^{(n)} = \binom{n}{p} p^k (1-p)^{n-k}.$$

Ezt a formulát n nagy értéke esetén rendkívül nehézkes lenne kiszámítani, ezért átalakítjuk a következő alakra:

$$P_k^{(n)} = \frac{1}{k!} \cdot \frac{n(n-1) \dots (n-k+1)}{n^k} (np)^k \left(1 - \frac{np}{n}\right)^{n-k}.$$

Ha most $n \rightarrow \infty$, $p \rightarrow 0$ úgy, hogy $np = \lambda$ állandó, akkor a fenti formulában az np helyébe λ -t írva:

$$P_k^{(n)} = \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \lambda^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Elvégezve az $n \rightarrow \infty$ határátmenetet:

$$(5.5.1) \quad P_k^{(n)} \rightarrow P_k = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Gondoljunk arra, hogy az $e = 2,718281\dots$ szám nem más, mint az $\left(1 + \frac{1}{n}\right)^n$ ($n = 1, 2, \dots$), sorozat határértéke, azaz

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n &= e \quad \text{és} \quad \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = \\ &= \lim_{\frac{n}{a} \rightarrow \infty} \left[\left(1 + \frac{1}{\frac{n}{a}}\right)^{\frac{n}{a}} \right]^a = e^a \end{aligned}$$

tetszőleges valós a számra. Az $a = -\lambda$ esetén kapjuk az (5.5.1) formulában levő határértéket.

Ha a megfigyelt egységnyi időintervallumba érkező jelek számát a ξ valószínűségi változó jelöli, akkor ξ lehetséges értékei $0, 1, 2, \dots$. Annak valószínűsége, hogy egy egységnyi intervallumba pontosan k számú jel érkezzék:

$$P_k = P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Tehát ξ eloszlása

$\xi:$	0	1	2	3	...	k
	$\frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda}$	$\frac{\lambda}{1!} e^{-\lambda}$	$\frac{\lambda^2}{2!} e^{-\lambda}$	$\frac{\lambda^3}{3!} e^{-\lambda}$	...	$\frac{\lambda^k}{k!} e^{-\lambda}$

A P_k valószínűség értéke adott k esetén a λ paramétertől függ. Mivel $\lambda = np$ és np a binomiális eloszlás várható értéke - ha n kísérletet végzünk és p valószínűséggel következik be egy kísérlet során a bennünket érdeklő esemény -, azért λ is várható értéket jelent, jelfolyamat esetében az egységnyi intervallumra eső átlagos jelsűrűséget. Ezt formálisan is beláthatjuk, ugyanis a Poisson-eloszlású valószínűségi változó várható értéke:

$$(5.5.2) \quad M(\xi) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{\ell=0}^{\infty} \frac{\lambda^{\ell}}{\ell!} =$$

$$= \lambda e^{-\lambda} e^{\lambda} = \lambda$$

Emlékeztetünk, hogy az e^x függvény Taylor-sora:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Nyilván a $x = \lambda$ esetén

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}.$$

Kiszámítjuk a Poisson-eloszlású ξ valószínűségi változó szórásnégyzetét is.

$$(5.5.3) \quad D^2(\xi) = M(\xi^2) - [M(\xi)]^2 = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 =$$

$$= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2 =$$

$$= \lambda e^{-\lambda} \left(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2 =$$

$$= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} - \lambda^2 = \lambda.$$

Tehát a Poisson-eloszlású ξ valószínűségi változó szórása:

$$(5.5.4) \quad D(\xi) = \sqrt{\lambda}.$$

Ahhoz, hogy a Poisson-eloszlás táblázatában szereplő P_k valószínűségeket ki tudjuk számítani, ismerni kell a λ paraméter értékét. A gyakorlatban λ értékét rendszerint statisztikai úton kell becsülni. Lásd:

(8.2) pont. A λ paraméter ismeretében a P_k valószínűségek értéke általában táblázatból kikereshető. Lásd: VII. táblázat.

Ime egy példa a táblázat használatára.

Tegyük fel, hogy egy telefonközpontba egy perc alatt átlagosan 4 hívás fut be, azaz $\lambda = 4$. Mennyi annak valószínűsége, hogy 1 percig figyelve a hívásokat, 6 hívás fut be?

A táblázat $\lambda = 4$ oszlopában a $k = 6$ sorban a következő értéket találjuk:

$$P_6 = \frac{\lambda^6}{6!} e^{-\lambda} = \frac{4^6}{6!} e^{-4} = 0,10420 \approx 0,10$$

Ez azt jelenti, hogy 100 megfigyelt perc során átlag 10 olyan perc lesz, amelyben pontosan 6 hívást észlelünk. Egyébként a táblázat $\lambda = 4$ oszlopát szemügyre véve a következőket olvashatjuk ki belőle:

100 megfigyelt perc során átlag egy olyan perc lesz, amelyben nem jön hívás, 7 olyan perc, amelyben 1 hívás érkezik, 15 olyan perc, amelyben 2 hívás szerepel, 20 olyan perc, amelyben 3 hívás történik, ugyancsak 20 olyan perc, amelyben 4 hívás, 15 olyan perc, amelyben 5 hívás fut be stb. Ezek a számok utmutatást adnak arra, hogy $\lambda = 4$ esetén mekkora lesz a központ maximális terhelése, milyen nagyra kell a központ tároló kapacitását méretezni. Pl. $\lambda = 4$ értéknél 10-nél több hívásra gyakorlatilag nem kell számítanunk, mert ennek valószínűsége már csak ezrelékben fejezhető ki.

Megemlítettünk egy híres kísérletet, amelyet a radioaktív bomlás α -részecskéinek számlálásával kapcsolatban végeztek. Egységnyi időintervallumnak 7,5 sec-ot választva, $N = 2608$ egységnyi intervallumban megszámlálták a beérkező részecskék számát. Ezután összeszámolták, hogy hány olyan intervallum volt, amelyben egyetlen részecskét sem tudtak megfigyelni, hány olyan, amelyben 1, 2, ..., 9 részecske érkezett, s végül hány olyan, amelyben 10 vagy annál több részecskét számláltak, majd az eredményt táblázatban foglalták össze:

k	N	NP_k
0	57	54,399
1	203	210,523
2	383	407,361
3	525	525,496
4	532	508,418
5	408	393,515
6	273	253,817
7	139	140,325
8	45	67,882
9	27	29,189
k-10	16	17,075
Össze- sen	2608	2608,000

A közelítést itt a $\lambda = \frac{\sum k N_k}{2608}$ paraméterű Poisson-eloszlással végezték.

A táblázat adatai jól szemléltetik, hogy a beérkező részecskék számának véletlen ingadozása Poisson-eloszlást követ.

5.6 Véletlen eseményfolyamatok A Poisson-folyamat alkalmazása

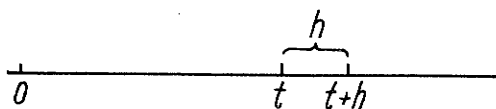
A továbbiakban az időben véletlenszerűen lefolyó jelenségeknél közös néven jelnek fogjuk nevezni egy α részecske érkezését a számlálócsőbe, egy hívás beérkezését a telefonközpontba, egy alkatrész meghibásodását egy sok alkatrészt tartalmazó készülékben stb. Az ilyen jellegű események elméleti szempontból közös természetűek, azonos valószínűségeloszlási törvényszerűségeknek tesznek eleget, ezért egységesen tárgyalhatók. Feltesszük, hogy kísérletünk ilyen jelek beérkezésének megfigyeléséből áll, és megfigyelésünket a $t = 0$ időpontban kezdjük. Keressük annak valószínűségét, hogy a $(0, t)$ időtartam alatt pontosan k számú jelet észlelünk. Ezt a valószínűséget $P_k(t)$ -val fogjuk jelölni. E valószínűség meghatározásához a következő két egyszerű feltételből indulunk ki.

1. Feltesszük, hogy bármely időintervallumban érkező jelek száma független az intervallum kezdete előtt befutott jelek számától. Másként fogalmazva: feltesszük, hogy diszjunkt intervallumokba eső jelek számai független valószínűségi változók.

2. Ugyancsak feltesszük, hogy annak valószínűsége, hogy egy kicsiny h hosszúságú intervallumba egy jel érkezzék, arányos az intervallum hosszával, azaz annak valószínűsége, hogy egy $(t, t+h)$ intervallumba egyetlen jel érkezzék, λh -val egyenlő (ahol λ az arányossági tényező, amely az egységnyi intervallumba átlagosan érkező jelek számát jelenti). Annak valószínűsége, hogy a $(t, t+h)$ időközben egynél több jel érkezzék, legyen elhanyagolhatóan kicsi, ha h elég kicsiny.

Ezen feltevések mellett először határozzuk meg annak valószínűségét, hogy a $(0, t)$ időközben egyetlen jel sem érkezik. Jelöljük ezt a valószínűséget $P_0(t)$ -vel.

Tekintsük a $(0, t+h)$ intervallumot, amelyet két összetevőre bontunk, a $(0, t)$ és $(t, t+h)$ időközökre.



40. ábra

Annak valószínűsége, hogy a $(0, t)$ időközben nem jön jel, $P_0(t)$, míg annak valószínűsége, hogy a $(t, t+h)$ időközben nem jön jel; $1 - \lambda h$. Így annak valószínűsége, hogy a $(0, t+h)$ időközben nem érkezik jel:

$$P_0(t+h) = P_0(t) (1 - \lambda h) .$$

Mivel a $(0, t)$ és $(t, t+h)$ intervallumok diszjunktak, tehát feltevésünk szerint ezen intervallumokba befutó jelek száma független valószínűségi változók, így a valószínűségek szorzási szabálya alkalmazható.

Az összefüggésből

$$\frac{P_0(t+h) - P_0(t)}{h} = -\lambda P_0(t) .$$

Elvégezve a $h \rightarrow 0$ határmenetet, a bal oldali differenciáhányados differenciáhányadosba megy át,

$$(5.6.1) \quad P'_0(t) = -\lambda P_0(t) ,$$

azaz:

$$\frac{P'_0(t)}{P_0(t)} = -\lambda .$$

Integrálva a kapott differenciálegyenletet:

$$\ln P_0(t) = -\lambda t + \ln C_0 ,$$

antiból:

$$P_0(t) = C_0 e^{-\lambda t}.$$

Mivel $P(0) = 1$ -nek vehető, hiszen annak valószínűsége, hogy 0 időtartam alatt nem jön jel, nyilván biztos esemény, az összefüggésből $t = 0$ helyettesítéssel

$$P_0(0) = C_0 = 1 \text{ adódik, tehát:}$$

(5.6.2)

$$P_0(t) = e^{-\lambda t}.$$

Itt most egy fontos megjegyzést teszünk. Annak valószínűsége, hogy a $(0, t)$ időközben nem jön el, más szóval, hogy az első jel megérkezéséig t -nél hosszabb ideig kell várakoznunk, $e^{-\lambda t}$ -vel egyenlő, ami nyilván a t idő függvénye. Minél nagyobb a t értéke, annál kisebb ez a valószínűség, mégpedig t növekedésével exponenciálisan csökken.

Annak valószínűsége, hogy a $(0, t)$ időközben legalább egy jel jön, más szóval annak valószínűsége, hogy az első jel megérkezéséig t -nél rövidebb ideig kell várakoznunk: $1 - e^{-\lambda t}$. Ez a valószínűség ugyancsak t függvénye, s jelöljük ezt a függvényt $F(t)$ -vel. Ha tehát η -val jelöljük az első jel érkezéséig eltelt várakozási időt, akkor

$$(5.6.3) \quad P(\eta < t) = F(t) = 1 - e^{-\lambda t}.$$

Könnyű belátni, hogy az $F(t)$ folytonos függvény t növekedésével állandóan nő, $F(0) = 0$, $F(+\infty) = 1$, tehát $F(t)$ eloszlásfüggvény. Az $F(t)$ függvényvel jellemzett valószínűségeloszlást exponenciális eloszlásnak nevezzük, amelynek vizsgálatára még visszatérünk.

Folytatva a jelek időbeli eloszlásának vizsgálatát, számítsuk most ki annak valószínűségét, hogy a $(0, t)$ időintervallumban pontosan k számú jel jön, tehát a $P_k(t)$ valószínűséget.

Először felírjuk annak a valószínűségét, hogy a $(0, t+h)$ időközben k számú jel jön, vagyis a $P_k(t+h)$ valószínűséget. Ehhez gondoljuk meg a következőt. Az az esemény, hogy a $(0, t+h)$ időközben k számú jel érkezik, két egymást kizáró módon jöhet létre: * 1. k számú jel jön a $(0, t)$ intervallumban és nem jön jel a $(t, t+h)$ intervallumban, 2. $k-1$ jel jön $(0, t)$ -ben és egy jel jön a $(t, t+h)$ időközben.

* Feltételezve, hogy kicsiny h hosszúságú időközben az egynél több jel érkezésének valószínűsége elhanyagolhatóan csekély, az egy jel érkezésének valószínűségéhez képest.

A valószínűségek összeadási és szorzási szabályát alkalmazva:

$$P_k(t+h) = P_k(t)(1-\lambda h) + P_{k-1}(t)\lambda h.$$

Innen:

$$\frac{P_k(t+h) - P_k(t)}{h} = -\lambda P_k(t) + \lambda P_{k-1}(t).$$

Elvégezve a $h \rightarrow 0$ határmenetet, a következő differenciálegyenlethez jutunk:

$$(5.6.4) \quad P'_k(t) = -\lambda P_k(t) + \lambda P_{k-1}(t), \quad k = 1, 2, \dots$$

Az (5.6.4) egyenlet lényegében egy végtelen differenciálegyenlet-rendszer, ez azonban semmi nehézséget sem jelent, mert rekurzív megoldható. rögzített k -ra a lineáris elsőrendű differenciálegyenletet kell integrálnunk.

Vizsgáljuk először a $k = 1$ esetet. Ekkor $P_{k-1}(t) = P_0(t) = e^{-\lambda t}$, tehát

$$P'_1(t) = -\lambda P_1(t) + \lambda e^{-\lambda t}.$$

Oldjuk meg most a

$$P'_1(t) = -\lambda P_1(t) \quad \text{homogén egyenletet.}$$

$$\frac{P'_1(t)}{P_1(t)} = -\lambda,$$

$$\ln P_1(t) = -\lambda t + \ln C_1,$$

$$(5.6.5) \quad P_1(t) = C_1 e^{-\lambda t}.$$

Tekintsük most a C_1 állandót a t változó függvényének, és jelöljük $C_1(t)$ -vel. Ezt a differenciálegyenletek elméletében szokásos fogást az állandó variálásának nevezzük. Differenciáljuk t szerint az egyenletet és hasonlítsuk össze az egyenlet jobb oldalával:

$$P'_1(t) = C'_1(t) e^{-\lambda t} - \lambda C_1(t) e^{-\lambda t} = -\lambda P_1(t) + \lambda e^{-\lambda t}.$$

Ha most a $P_1(t)$ helyébe az (5.6.5.) összefüggés jobb oldalán álló mennyiséget írjuk, azt kapjuk, hogy

$$C'_t(t) e^{-\lambda t} - \lambda C_1(t) e^{-\lambda t} = -\lambda C_1(t) e^{-\lambda t} + \lambda e^{-\lambda t}.$$

Ebből az összefüggésből

$$C'_1(t) = \lambda \quad \text{adódik, vagyis}$$

$$C_1(t) = \lambda t.$$

A kapott értéket az (5. 6. 5) formulába helyettesítve

$$(5. 6. 6) \quad P_1(t) = \lambda t e^{-\lambda t} = \frac{(\lambda t)^1}{1!} e^{-\lambda t}$$

Hasonló eljárással kaphatjuk, hogy

$$P_2(t) = \frac{(\lambda t)^2}{2!} e^{-\lambda t},$$

$$P_3(t) = \frac{(\lambda t)^3}{3!} e^{-\lambda t} \text{ stb.}$$

Most teljes indukcióval kimutatjuk, hogy tetszőleges k-ra

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

Tételezzük fel, hogy

$$P_{k-1}(t) = \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t},$$

és írjuk be ezt az értéket az (5. 6. 4) összefüggésbe:

$$P'_k(t) = -\lambda P_k(t) + \lambda \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t}.$$

Az egyenlet homogén részét megoldva:

$$(5. 6. 7) \quad P_k(t) = C_k e^{-\lambda t} \text{ adódik.}$$

Ismét az állandó variálását alkalmazva:

$$P'_k(t) = C'_k(t) e^{-\lambda t} - \lambda C_k(t) e^{-\lambda t} = -\lambda C_k(t) e^{-\lambda t} + \lambda \frac{t^{k-1}}{(k-1)!} e^{-\lambda t}$$

Ebből láthatjuk, hogy $C'_k(t) = \lambda \frac{t^{k-1}}{(k-1)!}$, melyből integrálással:

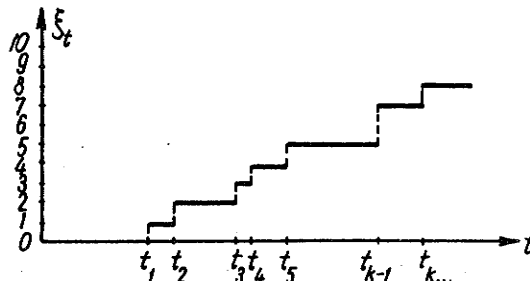
$$C_k(t) = \frac{(\lambda t)^k}{k!}.$$

A nyert $C_k(t)$ értéket az (5. 6. 7) egyenletbe helyettesítve:

$$(5. 6. 8) \quad P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Ezzel kimutatjuk, hogy a tett feltételek mellett a véletlen jelfolyamat λt paraméterű Poisson-eloszlást követ. Az (5. 6. 8) formula mutatja, hogy $P_k(t)$ valószínűsége a t változó függvénye. A t változó most az időt jelenti. Rögzített t érték esetén a $(0, t)$ időközben befutó jelek száma egy véletlen szám, valószínűségi változó, amelyet ξ_t -vel jelölünk.

Ha az egyes hívások időpontjait a t tengelyen feltüntetjük, akkor ξ_t képét a 41. ábra lépcsős függvénye szemlélteti:



41. ábra

A ξ_t valószínűségi változó t értékétől függően más-más valószínűségekkel veszi fel a $1, 1, 2, \dots, k, \dots$ értékeket, tehát itt lényegében nem egyetlen valószínűségi változóval van dolgunk, hanem valószínűségi változóknak a t időparamétertől függő seregével. Ezért, ha t értékét nem rögzítjük, hanem úgy tekintjük, hogy t az origóból indulva befutja a számegyeneset, akkor a ξ_t szimbólum nem egy valószínűségi változót, hanem egy véletlen folyamatot jelöl. Hogy erről a véletlen folyamatról konkrétan szemléletünk legyen, tegyük fel, hogy egy telefonközpontba befutó hívásokat figyeljük, és megfigyelésünket a $t = 0$ időpontban kezdjük. Tegyük fel továbbá, hogy az egyes hívások a $t_1, t_2, \dots, t_n, \dots$ időpontokban érkeznek. Jelöljük ξ_t -vel a t időpontig befutott hívások számát, ekkor

$$\xi_t = \begin{cases} 0, & \text{ha } t < t_1; \\ 1, & \text{ha } t_1 \leq t < t_2; \\ 2, & \text{ha } t_2 \leq t < t_3; \\ \vdots & \\ k, & \text{ha } t_k \leq t < t_{k+1} \\ \vdots & \end{cases}$$

Az így kapott lépcsős függvény a ξ_t hívásfolyamat egy realizációja. Ha az egyes hívások más időpontokban futnak be, akkor ξ_t -nek másik lépcsős függvény felel meg. Minden olyan nem csökkenő lépcsős függvény, amely csak a 0, 1, 2, ... értékeket veheti fel, egy telefonhívási folyamat lehetséges realizációjának tekinthető. A ξ_t folyamat most az összes ilyen típusú lépcsős függvények összessége, halmaza és minden ilyen függvény a folyamat egy lehetséges realizációja. Ha mármost a kísérletünk egy telefonközpontba befutó hívások (hosszu ideig tartó) megfigyelése, akkor kísérletünk minden egyes kimenetele egy fenti típusú függvény. Kísérletünk kimenetele szempontjából most az ilyen függvény egy elemi eseménynek felel meg, s az elemi események tere az összes lehetséges realizációk halmaza. Annak az eseménynek, hogy egy rögzített $t = T_0$ ideig befutott hívások száma mondjuk 100 és 120 közé esik, mindama lépcsős függvények halmaza felel meg, amely függvények a T_0 időpontban 100 és 120 magasságok közé eső ordinátáival rendelkeznek.

Ha most nagyon sokszor, N -szer megfigyelünk a befutó hívásokat egy rögzített $t = T_0$ ideig, továbbá minden esetben ábrázolnánk a megfelelő realizációt és ezek között M számú olyan függvényt találnánk, amely a $t = T_0$ időpontban a 100 és 120 magasságok között fut, akkor igazolva látnánk, hogy

$$\frac{M}{N} \approx \sum_{k=100}^{120} \frac{(\lambda T_0)^k}{k!} e^{-\lambda T_0},$$

ahol λ az egységnyi időintervallumra eső hívások átlagos számát jelenti.

5.7 Egy sorbanállási probléma

Számos, a gyakorlatban gyakran fellépő helyzetnek elméleti modellje a következő:

Adva van egy kiszolgáló berendezés (pl. benzinkut), amelyhez véletlenszerűen érkeznek a kiszorgálandók, és egy-egy személy kiszorgálása a véletlentől függően több-kevesebb ideig tart. Amíg egy személy kiszorgálása folyik, a többi várakozik, sorban áll. Feltesszük, hogy összesen N várakozó hely van, vagyis a sor maximális hossza N . Tegyük fel továbbá, hogy a kiszorgálandó személyek érkezési időpontjai λ paraméterű Poisson-folyamatot alkotnak, s ez egyes kiszorgálási időtartamok egymástól és az érkezési időpontoktól független valószínűségi változók és azonos paraméterű exponenciális eloszlásuak.

Ezen feltevések mellett annak a valószínűsége, hogy egy kicsiny h időintervallumban egy személy érkezék: λh azaz arányos az intervallum hosszával, míg annak a valószínűsége, hogy h időtartam alatt egynél több személy érkezék, legyen elhanyagolhatóan kicsi, ha h elég kicsi. Annak valószínűsége, hogy h időtartam alatt egy személy kiszorgálása befejeződik: μh , míg annak valószínűsége, hogy elég kicsi h időközben egynél több személy kiszorgálása fejeződik be, legyen ugyancsak elhanyagolhatóan kicsi.

Keressük annak a valószínűségét, hogy a t időpontban pontosan n személy várakozik ($n \leq N$). Jelöljük ezt a valószínűséget $P_n(t)$ -vel. Kiszámítjuk először annak a valószínűségét, hogy a $t+h$ időpontban n személy várakozik. Ez az esemény a tett feltevések mellett a következő három egymást kizáró módon jöhet létre:

1. A t időpontban n személy várakozik, és h időtartama alatt sem új kiszorgálandó személy nem érkezett, sem a kiszorgálás nem fejeződött be. Ennek valószínűsége:

$$P_n(t) (1 - \lambda h) (1 - \mu h) .$$

2. A t időpontban $n-1$ személy várakozott, és h időtartam alatt egy személy érkezett, de a kiszorgálás nem fejeződött be. Ennek valószínűsége:

$$P_{n-1}(t) \lambda h (1 - \mu h) .$$

3. A t időpontban $n+1$ személy várakozott, és h idő alatt új személy nem érkezett, egy személy kiszorgálása viszont befejeződött. Ennek valószínűsége:

$$P_{n+1}(t) (1 - \lambda h) \mu h .$$

Ezek alapján:

$$(5.7.1) \quad P_n(t+h) = P_n(t) (1-\lambda h) (1-\mu h) + P_{n-1}(t) \lambda h (1-\mu h) + P_{n+1}(t) \mu h (1-\lambda h)$$

A fenti gondolatmenet $n = 1, 2, \dots, N-1$ -re alkalmazható.

Az $n = 0$ esetben a

$$(5.7.2) \quad P_0(t+h) = P_0(t)(1-\lambda h) + P_1(t)\mu h(1-\lambda h)$$

egyenlet adódik, mivel az az esemény, hogy a $t+h$ időpontban ne legyen várakozó, csak úgy következhet be, ha a t időpontban nem volt várakozó és a n időtartam alatt nem érkezett új személy, vagy a t időpontban 1 várakozó volt és h idő alatt egy kiszolgálás befejeződött, új kiszolgálandó személy pedig nem érkezett. (Várakozó személyen olyan személyt értünk, aki sorban áll vagy éppen kiszolgálják, de kiszolgálása még nem fejeződött be.)

Az $n = N$ esetben a fentiek analógiájára könnyen belátható, hogy a következő egyenletre jutunk:

$$(5.7.3) \quad P_N(t+h) = P_N(t) (1-\mu h) + P_{N-1}(t) \lambda h (1-\mu h).$$

Ha az (5.7.1), (5.7.2) és (5.7.3) egyenleteket átrendezzük és elvégezzük a határátmenetet, akkor elhanyagolva a h^2 nagyságrendű tagokat a következő differenciálegyenlet-rendszert nyerjük:

$$(5.7.4) \quad \begin{cases} P'_0(t) = -\lambda P_0(t) + \mu P_1(t); \\ P'_n(t) = -(\lambda + \mu) P_n(t) + \lambda P_{n-1}(t); \\ P'_N(t) = -\mu P_N(t) + \lambda P_{N-1}(t). \end{cases}$$

Az (5.7.4) egyenletben szereplő állandó együtthatóju lineáris differenciálegyenlet-rendszer explicite megoldható. Megoldása:

$$(5.7.5) \quad P_n(t) = P_n + C_{n1} e^{-r_1 t} + C_{n2} e^{-r_2 t} + \dots + C_{nN} e^{-r_N t}$$

alakú, ahol a C_{nk} együtthatókat a $P_n(0)$ kezdőfeltételek alapján kell meghatározni ($n = 0, 1, 2, \dots, N$).

Gyakorlati célokra általában kielégítő a

$$\lim_{t \rightarrow \infty} P_n(t) = P_n$$

határvalószínűségek ismerete. Ezek meghatározása sokkal egyszerűbb, mint az (5. 7. 4) egyenletrendszer megoldásának végigvitele. Ugyanis, mint az (5. 7. 5) formulából könnyen belátható:

$$\lim_{t \rightarrow \infty} P'_n(t) = 0; \quad n = 0, 1, 2, \dots, N,$$

így az (5. 7. 4) egyenletrendszerből az egyes P_n valószínűségek rekurzív uton meghatározhatók:

$$P_1 = \frac{\lambda}{\mu} P_0;$$

$$(\lambda + \mu) P_1 = \lambda P_0 + \mu P_2;$$

$$\left(\frac{\lambda^2}{\mu} + \lambda\right) P_0 = \lambda P_0 + \mu P_2,$$

azaz

$$P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0;$$

$$P_3 = \left(\frac{\lambda}{\mu}\right)^3 P_0;$$

$\vdots$

$$P_N = \left(\frac{\lambda}{\mu}\right)^N P_0.$$

Figyelembe véve, hogy $P_0 + P_1 + \dots + P_N = 1$, kapjuk:

$$1 = P_0 \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \dots + \left(\frac{\lambda}{\mu}\right)^N \right],$$

amiből a véges geometriai haladvány összegezésével:

$$(5. 7. 6) \quad P_0 = \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}},$$

tehát:

$$(5.7.7) \quad P_n = \left(\frac{\lambda}{\mu}\right)^n \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}}, \quad n = 0, 1, 2, \dots, N.$$

Ha a kiszolgáló helyek száma α , és egyébként a fenti feltevések érvényesek, akkor a megoldandó egyenletrendszer $n \geq \alpha$ esetén

$$P'_n(t) = -(\lambda + \alpha\mu)P_n(t) + \lambda P_{n-1}(t) + \alpha\mu P_{n+1}(t);$$

$n < \alpha$ esetén pedig:

$$P'_n(t) = -(\lambda + n\mu)P_n(t) + \lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t).$$

(Itt most is $P_n(t)$ annak a valószínűsége, hogy a t időpontban n személy áll a sorban vagy éppen kiszolgálják őket.)

Ugyancsak a határvalószínűséget határozzuk meg.

A $\lim_{t \rightarrow \infty} P_n(t) = 0$ érvényessége folytán

$$(\lambda + \alpha\mu) P_n = \lambda P_{n-1} + \alpha\mu P_{n+1},$$

és rekurzív uton azt kapjuk, hogy

$$(5.7.8) \quad P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{\alpha! \alpha^{n-1}} P_0, \text{ ha } n \geq \alpha, \text{ míg } n < \alpha\text{-nál}$$

$$(5.7.9) \quad P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0.$$

Megjegyezzük, hogy a P_n határvalószínűségek a $\frac{\lambda}{\mu} < \alpha$ esetén léteznek.

5.8 Készülékek karbantartási problémáinak matematikai modellje

A sorbanállási problémával csaknem teljesen megegyezik a következő modell, amelyet az iparban több sikerrel alkalmaznak.

Tegyük fel, hogy m darab automatikusan működő készülékünk van, amelyek időnként meghibásodnak, s ilyenkor emberi beavatkozás, javítás szükséges. Mindegyik készülék kétféle állapotban lehet: vagy működik, vagy hibás. Feltesszük, hogy a t időpontban a készülék működik, s akkor annak a valószínűsége, hogy kicsiny h időtartam alatt meghibásodik, λh , míg ha egy készülék a t időpontban hibás (azaz nem működik), akkor annak valószínűsége, hogy egy kicsiny h időtartam alatt a javítása befejeződik, μh . (Annak a valószínűsége, hogy h időtartam alatt egynél több készülék hibásodik meg vagy egynél több készülék javítása fejeződik be, legyen elhanyagolhatóan kicsi.)

Természetesen a készülék akkor jó hatásfoku, ha λ viszonylag kicsiny, μ pedig viszonylag nagy. A $\frac{\lambda}{\mu}$ hányadost karbantartási faktornak nevezzük.

Feltesszük, hogy az m számu készülék mindegyike a többitől függetlenül működik, továbbá a λ és μ paraméterek értéke mindegyik készüléknél ugyanaz, és egyetlen karbantartó szerelő áll rendelkezésre.

Keressük annak $P_n(t)$ valószínűségét, hogy a t időpontban n készülék nem működik (ennyi vár javításra).

Ismét annak a valószínűségét számítjuk ki először, hogy a $t+h$ időpontban n készülék vár javításra. Ez az esemény feltevéseink mellett a következő három egymást kizáró módon jöhet létre:

a) A t időpontban n készülék várt javításra, h időtartam alatt a működő $(m-n)$ készülék egyike sem hibásodott meg, és egyetlen készülék javítása sem fejeződött be.

Ennek valószínűsége:

$$P_n(t) [1 - (m-n) \lambda h - \mu h].$$

b) A t időpontban $n-1$ készülék volt hibás, h idő alatt egy készülék elromlott (a működő $m-n+1$ közül), viszont egyetlen készülék javítása sem fejeződött be.

Ennek valószínűsége:

$$P_{n-1}(t) (m-n+1) \lambda h (1 - \mu h).$$

c) A t időpontban $n+1$ készülék volt hibás, h idő alatt nem romlott el készülék, és egy készülék javítása h idő alatt befejeződött.

Ennek valószínűsége:

$$P_{n+1}(t) (1-\lambda h)^{m-n-1} \mu h .$$

Elhanyagolva a h^2 és annál kisebb nagyságrendű tagokat, a következő differenciálegyenlet-rendszert nyerjük:

$$(5.8.1) \quad \begin{cases} P'_n(t) = -[(m-n)\lambda + \mu]P_n(t) + (m-n+1)\lambda P_{n-1}(t) + \mu P_{n+1}(t) , \\ P'_0(t) = -m\lambda P_0(t) + \mu P_1(t) , \\ P'_m(t) = -\mu P_m(t) + \lambda P_{m-1}(t) . \end{cases}$$

Ki lehet mutatni, hogy a $\lim_{t \rightarrow \infty} P_n(t) = P_n$ határvalószínűségek léteznek, és kiszámításuk a következő rekurzíóval történik:

$$(5.8.2) \quad \begin{cases} m\lambda P_0 = \mu P_1 , \\ (m-1)\lambda P_1 = \mu P_2 , \\ \vdots \\ (m-n)\lambda P_n = \mu P_{n+1} , \\ \vdots \\ \lambda P_{m-1} = \mu P_m , \end{cases}$$

azaz

$$\begin{aligned} P_{m-1} &= \frac{\mu}{\lambda} P_m , \\ P_{m-2} &= \frac{1}{2!} \frac{\mu}{\lambda} P_{m-1} = \frac{1}{2!} \left(\frac{\mu}{\lambda}\right)^2 P_m , \\ P_{m-3} &= \frac{1}{3!} \left(\frac{\mu}{\lambda}\right)^3 P_m , \\ &\vdots \\ P_1 &= \frac{1}{(m-1)!} \left(\frac{\mu}{\lambda}\right)^{m-1} P_m , \\ P_0 &= \frac{1}{m!} \left(\frac{\mu}{\lambda}\right)^m P_m . \end{aligned}$$

Figyelembe véve, hogy $P_0 + P_1 + \dots + P_m = 1$, azt kapjuk, hogy

$$P_m = \frac{1}{1 + \frac{\mu}{\lambda} + \frac{1}{2!} \left(\frac{\mu}{\lambda}\right)^2 + \dots + \frac{1}{m!} \left(\frac{\mu}{\lambda}\right)^m}.$$

Tehát:

$$(5.8.3) \quad P_{m-k} = \frac{\frac{1}{k!} \left(\frac{\mu}{\lambda}\right)^k}{1 + \frac{\mu}{\lambda} + \frac{1}{2!} \left(\frac{\mu}{\lambda}\right)^2 + \frac{1}{3!} \left(\frac{\mu}{\lambda}\right)^3 + \dots + \frac{1}{m!} \left(\frac{\mu}{\lambda}\right)^m}.$$

Ez az ún. Erlang-formula, amely a mérnöki gyakorlatban jól ismert a telefonközpontokkal kapcsolatban.

Elméleti modellünket úgy is megfogalmazhattuk volna, hogy van egy telefonközpontunk m kapcsolási vonallal, minden vonal szabad vagy foglalt állapotban lehet, s a befutó hívás vagy szabad vonalat talál, vagy tárolódik egy tároló vonalon, s keressük annak valószínűségét, hogy egy adott időpontban n hívás várakozik kapcsolásra a tárolóvonalon. Az $n = m-k$ esetre az (5.8.3) formula adja a keresett valószínűséget, míg az (5.8.4) formula a várakozó hívások átlagos számát mutatja.

A javításra váró készülékek várható száma (nem számítva azt, amelyet a szerelő éppen javít):

$$M = \sum_{k=1}^m (k-1) P_k = \sum_{k=1}^m k P_k - (1 - P_0).$$

M értékének kiszámítása az (5.8.3) egyenletrendszer segítségével elvégezhető. Írjuk át az egyenletrendszert a következő alakba, és adjuk össze az egyenleteket:

$$m \lambda P_0 = \mu P_1,$$

$$m \lambda P_1 - \lambda P_1 = \mu P_2,$$

$$m \lambda P_2 - 2 \lambda P_2 = \mu P_3,$$

$$\vdots$$

$$m \lambda P_n = \mu P_{n+1},$$

$$\begin{aligned}
 m\lambda P_{m-1} - (m-1)\lambda P_{m-1} &= \mu P_{n+1} , \\
 m\lambda P_m - m\lambda P_m &= 0 \\
 \hline
 m\lambda - \lambda M - \lambda(1-P_0) &= \mu(1-P_0) .
 \end{aligned}$$

(Az összeadás során felhasználtuk, hogy $P_0 + P_1 + \dots + P_m = 1$.)

A kapott összegből M -et kifejezve az

$$(5.8.4) \quad M = m - \frac{\lambda + \mu}{\lambda} (1 - P_0)$$

összefüggést kapjuk.

A következő numerikus példa bizonyos számszerű képet ad a tárgyalt határvalószínűségekről és a javításra váró készülék átlagos számáról.

Tegyük fel, hogy a készülékek száma: $m = 6$, és a karbantartási faktor: $\frac{\lambda}{\mu} = 1$. Ekkor $n = 0, 1, 2, \dots, 6$ hibás készülék előfordulásának a valószínűsége rendre:

n	P	Javításra váró készülék
0	0,4845	0
1	0,2907	0
2	0,1454	1
3	0,0582	2
4	0,0175	3
5	0,0035	4
6	0,0003	5

$$M = 0,3294$$

Modellünket általánosíthatjuk arra az esetre, amikor m készülékhez $r < m$ számú karbantartó szerelőt alkalmazunk. (A többi feltevésünket változtatlanul hagyjuk.) Ha mármost egy bizonyos időpontban n készülék hibás, akkor $n < r$ esetén $r-n$ szerelő tétlenkedik, $n > r$ esetén pedig $n-r$ készülék várakozik javításra. Az előzőekkel hasonló gondolatmenetet követve, a határvalószínűségek meghatározására az alábbi rekurziót nyerjük:

$$\begin{aligned}
 n < r \text{ esetén } (m-n)\lambda P_n &= (n+1) P_{n+1} , \\
 \text{míg} \quad n \geq r \text{ esetén } (m-n)\lambda P_n &= \mu P_{n+1} .
 \end{aligned}$$

Ezekből az egyenletekből a $\frac{P_n}{P_0}$ hányados $n = 1, 2, \dots$ -re egymás után meghatározható, végül a $\sum_{k=0}^m P_k = 1$ összefüggés alapján P_0 kiszámítható.

A táblázat $\frac{\lambda}{\mu} = 0,1$ karbantartási arányt, $m = 20$ készüléket és $r = 3$ karbantartó szerelőt alapul véve mutatja a numerikus viszonyokat.

n	Karbantartás alatt álló ké- szülékek száma	Javításra váró készülékek száma	Tétlen szerelők száma	P_n
0	0	0	3	0,13625
1	1	0	2	0,27250
2	2	0	1	0,25888
3	3	0	0	0,15533
4	3	1	0	0,08802
5	3	2	0	0,04694
6	3	3	0	0,02347
7	3	4	0	0,01095
8	3	5	0	0,00475
9	3	6	0	0,00190
10	3	7	0	0,00070
11	3	8	0	0,00023
12	3	9	0	0,00007

Az ismertetett eredmények felhasználásával különböző gazdaságossági számítások végezhetők. Kiderül pl., hogy 0,1 karbantartási aránynál gazdaságosabb 20 készülékhez 3 szerelőt alkalmazni, mint 6 készülékhez egy szerelőt, ami meglepő eredmény.

B) FOLYTONOS ELOSZLÁSOK

5.9 A normális eloszlás

Amikor egy kísérlet kimenetelét nagyszámu, egymástól csak kevésbé vagy egyáltalán nem függő véletlen tényező határozza meg, mégpedig úgy, hogy az egyes tényezők külön-külön csak igen kis mértékben járulnak hozzá az összes véletlen hatásból eredő ingadozásokhoz, továbbá az egyes ténye-

zők hatásai egyszerűen összeadódnak, akkor ún. normális valószínűségeloszlás lép fel. Az ilyen körülményekkel a gyakorlatban nagyon sokszor találkozunk, s innen ered a normális vagy más néven Gauss-eloszlás központi szerepe a valószínűségelméletben. A normális eloszlás leggyakoribb alkalmazási területe a statisztikai sokaságok vizsgálata, a hibaszámítás és a binomiális eloszlás közelítése.

Pontosabb, matematikai fogalmazással, egy ξ valószínűségi változót akkor nevezünk normális eloszlásúnak, ha sűrűségfüggvénye:

$$(5.9.1) \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

ahol m és σ állandók, s $\sigma > 0$. A formulában szereplő m és σ mennyiségeket a normális eloszlás paramétereinek nevezzük. Az (5.9.1) formulát szemügyre véve, könnyű belátni, hogy bármilyen véges szám is az m , és $\sigma > 0$, az $f(x)$ függvény értéke minden x értékre pozitív, az első tényező a $\sigma > 0$ kikötés miatt, a második tényező pedig: $1/\exp\left[\frac{x-m}{\sqrt{2\sigma^2}}\right]^2$ mindig pozitív, akármekkora is e kitevője, amely kitevő jelen esetben ugyancsak pozitív vagy zérus. Ebből az is kiolvasható, hogy $f(x)$ értéke akkor maximális, ha e kitevője zérus, azaz $x = m$. Ekkor

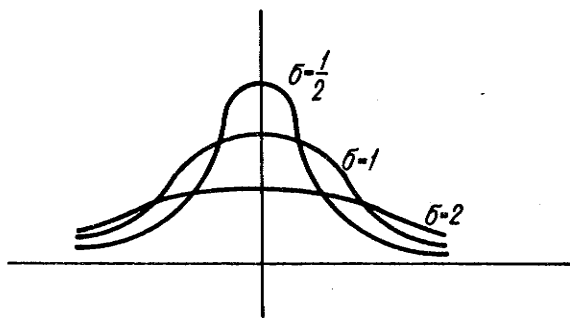
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}.$$

Minél kisebb a σ paraméter értéke, annál nagyobb $f(x)$ maximális értéke. $\sigma = 1$ esetén a maximum értéke

$$\frac{1}{\sqrt{2\pi}} \approx 0,4$$

Ha a σ paraméter értékét rögzítjük, akkor $f(x)$ értéke csak $(x-m)^2$ -től függ, ez az érték pedig ugyanakkora pozitív szám, akár pozitív, akár negatív az $x-m$ különbség. $f(x)$ értékét tehát az x változónak az $x_0 = m$ ponttól való távolsága határozza meg, s $f(x)$ az $x_0 = m$ pontra szimmetrikus. Az $f(x)$ függvény képe ún. haranggörbe. A függvény alakját különböző σ értékeknél a 42. ábra szemlélteti.

Ahhoz, hogy a (5.9.1) alatti kifejezés sűrűségfüggvény legyen, az kell, hogy a görbe alatti terület



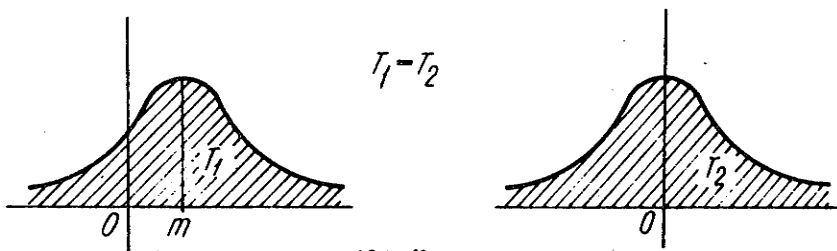
42. ábra

$$(5.9.2) \quad I = \int_{-\infty}^{\infty} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

egységnyi legyen. Nem nehéz kimutatni, hogy ez valóban így van. Vezessük be az (5.9.2) integrálban az $u = \frac{x-m}{\sigma}$ helyettesítést, ekkor $x = \sigma u + m$, $dx = \sigma du$, tehát az

$$(5.9.3) \quad I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du$$

integrált nyerjük, amelynek természetesen ugyanaz az értéke, mint a (5.9.2) integrálnak. Ebből látható, hogy bármilyen m és $\sigma > 0$ paraméterekkel bíró normális sűrűséggörbe alatti terület ugyanakkora, mint az $m = 0$, $\sigma = 1$ paraméterű sűrűséggörbe alatti terület, vagyis



43. ábra

a normális sűrűségfüggvény alatti terület nagysága nem függ az m és $\sigma > 0$ paraméterektől. Elegendő tehát kimutatni, hogy

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = 1.$$

Ennek igazolását a legegyszerűbben úgy végezhetjük el, hogy kimutatjuk az $I^2 = 1$ reláció fennállását.

$$\begin{aligned} I^2 &= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du \right)^2 = \\ &= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \right) \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \right) = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx dy = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} \cdot r dr d\varphi = \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left[-e^{-\frac{r^2}{2}} \right]_0^{\infty} d\varphi = 1. \end{aligned}$$

Itt alkalmaztuk az

$$\begin{aligned} x &= r \cdot \cos \varphi \\ y &= r \cdot \sin \varphi \end{aligned}$$

polárkoordinátás transzformációt, akkor

$$dx dy = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{vmatrix} dr d\varphi = \begin{vmatrix} \cos \varphi & -r \sin \varphi \\ \sin \varphi & r \cos \varphi \end{vmatrix} dr d\varphi = r dr d\varphi.$$

Ezzel kimutattuk, hogy a (5.9.1) alatti folytonos pozitív értékű $f(x)$ görbe alatti terület egységnyi, azaz $f(x)$ sűrűségfüggvény.

A sűrűségfüggvény és az eloszlásfüggvény közötti összefüggés alapján a normális eloszlás eloszlásfüggvénye:

$$(5.9.4) \quad F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt.$$

Az (5.9.4) összefüggésben ismét elvégezve az $u = \frac{t-m}{\sigma}$ helyettesítést, az

$$(5.9.5) \quad F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{u^2}{2}} du = \Phi\left(\frac{x-m}{\sigma}\right)$$

relációt nyerjük, ahol $\Phi(x)$ az $m = 0$, $\sigma = 1$ paraméterekkel bíró normális eloszlás eloszlásfüggvényét jelöli. A $\Phi(x)$ függvény értékeit a IV. táblázat tartalmazza.

Az (5.9.5) összefüggés értelmét kissé részletesebben megvilágítjuk. Legyen ξ normális eloszlású valószínűségi változó m és $\sigma > 0$ paraméterekkel (rövid jelöléssel: $N(m; \sigma)$) akkor ξ eloszlásfüggvénye

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

annak a valószínűsége, hogy ξ értéke kisebb, mint x .

A ξ valószínűségi változó felvett értékei (a kísérlet kimenetelei) helyett tekintsük a $\xi^* = \frac{\xi - m}{\sigma}$ értékeket. ξ^* természetesen ugyancsak valószínűségi változó, amelynek értékeit ξ felvett értékei egyértelműen meghatározzák. A ξ^* valószínűségi változót standardizált változónak nevezzük. Ha a ξ valószínűségi változó egy kísérletnél a $\xi = x_0$ értéket veszi fel, akkor a ξ^* változó ugyanannál a kísérletnél a

$\xi^* = \frac{x_0 - m}{\sigma}$ értéket képviseli. A ξ^* változó eloszlásfüggvényét jelöljük $\Phi(x)$ -szel. Nyilvánvaló, hogy a $\xi < x$ esemény valószínűsége ugyanaz, mint a $\xi^* = \frac{x-m}{\sigma}$ esemény valószínűsége,

$$P(\xi < x) = P(\sigma\xi^* + m < x) = P\left(\xi^* < \frac{x-m}{\sigma}\right),$$

azaz

$$(5.9.6) \quad F(x) = \Phi\left[\frac{x-m}{\sigma}\right].$$

Annak valószínűsége, hogy az m és $\sigma > 0$ paraméterekkel bíró normális eloszlású változó az (a, b) intervallumba essék:

$$(5.9.7) \quad F(b) - F(a) = P(a \leq \xi < b) = F(a \leq \sigma \xi^* + m < b) =$$

$$= P\left[\frac{a-m}{\sigma} \leq \xi^* < \frac{b-m}{\sigma}\right] = \Phi\left[\frac{b-m}{\sigma}\right] - \Phi\left[\frac{a-m}{\sigma}\right].$$

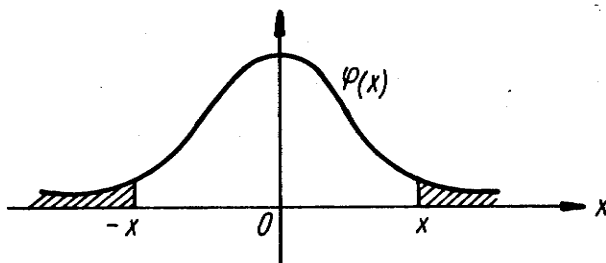
A $\Phi(x)$ függvény értéktáblázatát általában pozitív x -értékekre adják meg. Az (5.9.7) összefüggés segítségével azonban könnyen nyerjük a $\Phi(x)$ értékeit negatív x értékekre is.

A ξ^* standardizált normális valószínűségi változó sűrűségfüggvénye $\varphi(x)$ az (5.9.1) sűrűségfüggvényből $m = 0$, $\sigma = 1$ helyettesítéssel nyerhető:

$$(5.9.8) \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

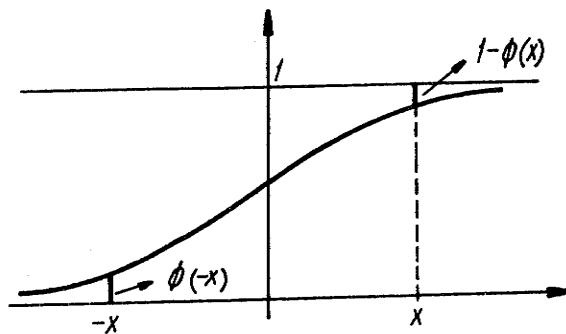
A $\varphi(x)$ függvény az $x = 0$ pontra szimmetrikus, így

$$(5.9.9) \quad \Phi(-x) = \int_{-\infty}^{-x} \varphi(t) dt = 1 - \int_{-x}^{\infty} \varphi(t) dt = 1 - \int_{-\infty}^x \varphi(t) dt = 1 - \Phi(x),$$



44. ábra

amely összefüggés a $\Phi(x)$ eloszlásfüggvény grafikonjáról is leolvasható



45. ábra

Annak valószínűsége, hogy a ξ^* valószínűségi változó értéke a $(-x, x)$ intervallumba esik:

$$(5.9.10) \quad \Phi(x) - \Phi(-x) = \Phi(x) - [1 - \Phi(x)] = 2 \Phi(x) - 1.$$

Az elmondottak gyakorlati alkalmazását szemlélteti a következő példa. Tegyük fel, hogy egy löveggel lövéseket adunk le egy 1500 m távolságra levő célra. A becsapódási távolságot jelölje a ξ valószínűségi változó $M(\xi) = 1500$ m várható értékkel és $D(\xi) = 40$ m szórással. Számítsuk ki, mi a valószínűsége annak, hogy a becsapódások a céltől legfeljebb 20 m távolságra történnek, azaz ξ értéke az (1480, 1520) intervallumba esik.

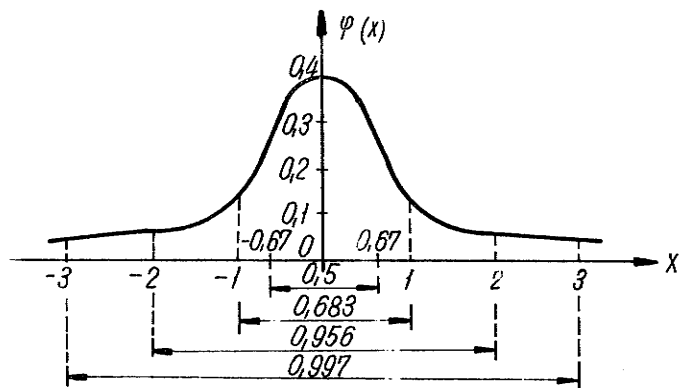
$$P(1480 \leq \xi < 1520) = P\left(\frac{1480-1500}{40} \leq \xi^* < \frac{1520-1500}{40}\right) =$$

$$= P(-0,5 \leq \xi^* < 0,5) = 2 \Phi(0,5) - 1 = 2 \cdot 0,6915 - 1 =$$

$$= 1,3830 - 1 = 0,3830.$$

A $\Phi(x)$ eloszlásfüggvény értéktáblázatát használva látjuk, hogy a ξ^* standardizált változó a 0 pont körül megjelölt intervallumokba a 46. ábrán látható valószínűségekkel esik.

Az (5.9.6) összefüggés alapján könnyen belátható, hogy a következő két táblázat teljesen ekvivalens:



46. ábra

$\xi^* : N(0; 1)$	$\xi : N(m; \sigma)$
$P(-0,67 \leq \xi^* < 0,67) = 0,5$	$P(m-0,67\sigma \leq \xi < m+0,67\sigma) = 0,5$
$P(-1 \leq \xi^* < 1) = 0,684$	$P(m-\sigma \leq \xi < m+\sigma) = 0,683$
$P(-2 \leq \xi^* < 2) = 0,956$	$P(m-2\sigma \leq \xi < m+2\sigma) = 0,956$
$P(-3 \leq \xi^* < 3) = 0,997$	$P(m-3\sigma \leq \xi < m+3\sigma) = 0,997$

E táblázatból pl. kiolvasható, hogy igen kicsi, kisebb mint 0,05 annak valószínűsége, hogy egy m és σ paraméterekkel bíró normális eloszlású ξ valószínűségi változó megfigyelt értéke az $x = m$ ponttól 2σ -nál nagyobb távolságra essék. Ezt a tényt, a matematikai statisztikában gyakran fel fogjuk használni, ez az ún. 2σ -szabály. Az pedig gyakorlatilag ugyiszólván biztos, hogy ξ megfigyelt értéke az $(m-3\sigma, m+3\sigma)$ intervallumba esik. A fenti táblázat adatai azt sugallják, hogy m paraméter valószínűség-számítási jelentése: m az eloszlás centruma, várható értéke, amely körül ξ megfigyelt értékei tömörülnek, a σ paraméter pedig az eloszlás szórása.

Valóban:

$$\begin{aligned}
 M(\xi) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m+m) e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \\
 &= \frac{\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{x-m}{\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx + \frac{m}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx =
 \end{aligned}$$

Ha most alkalmazzuk az $\frac{x-m}{\sigma} = u$ helyettesítést, akkor azt kapjuk, hogy

$$(5.9.11) \quad M(\xi) = \frac{\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du + \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = m.$$

Ugyancsak a fenti helyettesítés alkalmazásával, parciális integrálással, egyszerűen származtatható a normális eloszlás szórásnégyzete is:

$$(5.9.12) \quad D^2(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}} du =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot u e^{-\frac{u^2}{2}} du = \frac{\sigma^2}{\sqrt{2\pi}} \left[-u e^{-\frac{u^2}{2}} \right]_{-\infty}^{\infty} +$$

$$+ \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = \sigma^2.$$

A továbbiakban többször fel fogjuk használni a normális eloszlás karakterisztikus függvényét, amelyet először az $N(0, 1)$ standard normális eloszlású ξ^* valószínűségi változóra származtatunk.

$$(5.9.13) \quad \varphi(t) = M(e^{it\xi^*}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-it)^2}{2}} e^{-\frac{t^2}{2}} dx = e^{-\frac{t^2}{2}}.$$

Itt alkalmazzuk az $x-it = z$ helyettesítést, és mint az analízisben kimutatják:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty-it}^{\infty-it} e^{-\frac{z^2}{2}} dz = 1.$$

Ha most a ξ valószínűségi változó $N(m; \sigma)$ eloszlású, azaz

$$\xi = \sigma \xi^* + m,$$

akkor a (4. 2. 6) összefüggés alapján

$$\begin{aligned} (5.9.14) \quad \varphi_{\xi}(t) &= M \left[e^{it(\sigma \xi^* + m)} \right] = e^{itm} \left[e^{i\sigma t \xi^*} \right] = \\ &= e^{itm} e^{-\frac{\sigma^2 t^2}{2}} = e^{itm - \frac{\sigma^2 t^2}{2}}. \end{aligned}$$

Természetesen a $\varphi_{\xi}(t)$ karakterisztikus függvény $t = 0$ helyen vett deriváltjai segítségével igen egyszerűen származtathatók a normális eloszlás momentumai. (Ennek elvégzését az olvasóra bizzuk.)

Legyenek most a ξ_1 és ξ_2 független normális eloszlású valószínűségi változók $N(m_1; \sigma_1^2)$, illetve $N(m_2; \sigma_2^2)$ eloszlásokkal. Legyen $\zeta = \xi_1 + \xi_2$. Határozzuk meg a ζ eloszlását. Mivel a karakterisztikus függvény az eloszlást egyértelműen meghatározza, elegendő számunkra ζ karakterisztikus függvényének meghatározása, ennek képletéből az eredmény kiolvasható. A (4. 2. 5) formula alapján:

$$\begin{aligned} (5.9.15) \quad \varphi_{\zeta}(t) &= M \left[e^{it(\xi_1 + \xi_2)} \right] = M(e^{it\xi_1}) M(e^{it\xi_2}) = \\ &= \varphi_{\xi_1}(t) \varphi_{\xi_2}(t) = e^{itm_1 - \frac{\sigma_1^2 t^2}{2}} e^{itm_2 - \frac{\sigma_2^2 t^2}{2}} = \\ &= e^{it(m_1 + m_2) - \frac{(\sigma_1^2 + \sigma_2^2) t^2}{2}}. \end{aligned}$$

Ebből látható, hogy ζ ugyancsak normális eloszlású és

$$(5.9.16) \quad M(\zeta) = m_1 + m_2,$$

$$(5.9.17) \quad D^2(\zeta) = \sigma_1^2 + \sigma_2^2.$$

Azt az eredményt kaptuk tehát, hogy független normális eloszlású valószínűségi változók összege ugyancsak normális eloszlású. Ez az eredmény, amint az a karakterisztikus függvény alkalmazásával közvetlenül látható, tetszőleges véges számú független normális eloszlású valószínűségi változó összegének eloszlására igaz.

A továbbiakban a normális eloszlás néhány olyan tulajdonságára hívjuk fel a figyelmet, amelyek a gyakorlati alkalmazások szempontjából rendkívül fontosak.

1. Ha ξ normális eloszlású, $M(\xi) = m$ várható értékkel és $D(\xi) = \sigma^2$ szórással rendelkezik és $\eta = a\xi + b$, a ξ változó lineáris függvénye, akkor η ugyancsak normális eloszlású, $M(\eta) = am + b$ várható értékkel és $D(\eta) = a^2 \sigma^2$ szórással bír.

Állításunkat a (2. 6. 3) összefüggés segítségével könnyen bebizonyíthatjuk. Mivel a ξ változó sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}},$$

és a (2. 6. 3) formula szerint az η változó sűrűségfüggvénye:

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right) = \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{\left(\frac{y-b}{a} - m\right)^2}{2\sigma^2}},$$

$$(5. 9. 18) \quad g(y) = \frac{1}{a\sigma\sqrt{2\pi}} e^{-\frac{(y-am-b)^2}{2a^2\sigma^2}}.$$

Az utóbbi formulából állításunk máris kiolvasható, hiszen az (5. 9. 18) formulában szereplő függvény normális sűrűségfüggvény, amelyben $m_1 = am + b$ a várható érték, és $\sigma_1^2 = a^2 \sigma^2$ a szórással.

Mivel a különböző normális eloszlásokat az m és σ paraméterek értéke egyértelműen meghatározza, az előbbiekből az is következik, hogy ha két normális eloszlású változóval van dolgunk - $\xi: N(m_1, \sigma_1^2)$ eloszlású és $\eta: N(m_2, \sigma_2^2)$ eloszlású -, akkor mindig találhatók olyan $a > 0$ és $b > 0$ számok, hogy $a\xi + b$ eloszlása $N(m_2, \sigma_2^2)$. Más szóval egy normális eloszlású valószínűségi változó lineáris transzformációval másik előírt normális eloszlásba vihető át. Az a és b számokat úgy kell csak megválasztani, hogy $am_1 + b = m_2$ és $a^2 \sigma_1^2 = \sigma_2^2$ legyen. Mivel $a > 0$, elegendő, ha

$$a \sigma_1 = \sigma_2 ,$$

$$a = \frac{\sigma_2}{\sigma_1} \quad \text{és}$$

$$b = m_2 - \frac{\sigma_2}{\sigma_1} m_1 \quad \text{teljesül.}$$

2. A valószínű közepes eltérésnek és a szórásnak a hányadosa minden normális eloszlásra ugyanaz.

Legyen ξ normális eloszlású σ_1 szórással és T_1 valószínű közepes eltéréssel, η ugyancsak normális eloszlású σ_2 szórással és T_2 valószínű közepes eltéréssel. Ekkor az előzőekben mondottak alapján az η változó a ξ változónak egy lineáris transzformáltja, azaz $\eta = a\xi + b$ alakban írható. A valószínű közepes eltérés definíciója szerint:

$$P(|(a\xi + b) - [aM(\xi) + b]| < T_2) = \frac{1}{2} .$$

Ebből következik, hogy

$$P(a|\xi - M(\xi)| < T_2) = \frac{1}{2} ,$$

vagyis

$$P(|\xi - M(\xi)| < \frac{T_2}{a}) = \frac{1}{2} .$$

Viszont ugyancsak a valószínű közepes eltérés definíciója szerint:

$$P(|\xi - M(\xi)| < T_1) = \frac{1}{2} ,$$

$$\text{tehát } \frac{T_2}{a} = T_1 \quad \text{vagyis} \quad \frac{T_2}{T_1} = a .$$

A szórás tulajdonságainál kimutattuk, hogy ha $\eta = a\xi + b$, akkor

$$\sigma_2^2 = D^2(\eta) = a^2 \sigma_1^2 \quad (\text{lásd (3.3.8) formulát,})$$

vagyis

$$\frac{\sigma_2}{\sigma_1} = a .$$

Ezekből az összefüggésekből következik, hogy

$$\frac{T_2}{T_1} = \frac{\sigma_2}{\sigma_1}, \quad \text{azaz} \quad \frac{T_2}{\sigma_2} = \frac{T_1}{\sigma_1}.$$

Tehát normális eloszlásnál a $\frac{T}{\sigma}$ hányados állandó. E hányadost λ -val fogjuk jelölni, s numerikus értéke:

$$\frac{T}{\sigma} = \lambda = \sqrt{\frac{2}{\pi}} \approx 0,674 \text{ lesz.}$$

Mínt hogy tetszőleges normális eloszlás esetében $T = \sqrt{\frac{2}{\pi}} \sigma$, abból következően az ilyen fajtaú eloszlásnál teljesen mindegy, hogy az ingadozás jellemzésére a valószínű közepes eltérést vagy a szórásat használjuk. Már említettük, hogy számítástechnikai szempontból a szórás használata előnyösebb; a valószínű közepes eltérés használata a tűzéségi gyakorlatban terjedt el, ahol általában normális eloszlás fordul elő. Éppen ezért itt az utóbbi mérőszám használata semmi zavart nem okoz, s mivel a két mérőszám csak egy numerikus faktorban különbözik (skálaparaméterben), a szórás legfontosabb tulajdonsága - hogy ti. független valószínűségi változók összegének szórásnégyzete a szórásnégyzetek összegével egyenlő - változatlanul érvényes a valószínű közepes eltérésekre is!

5.10 A binomiális eloszlás közelítése normális eloszlással A Moivre – Laplace határeloszlás tétel

A binomiális eloszlás tárgyalásánál (5.2 pont) említettük, hogy a

$$(5.10.1) \quad P_k = \binom{n}{k} p^k (1-p)^{n-k}$$

valószínűségek kiszámítása nagy n értéknél igen fáradságos, szinte reménytelen, s ezért arra törekszünk, hogy az (5.10.1) formulát valamilyen egyszerűbb, könnyen kezelhető közelítő formulával helyettesítsük. A p paraméter értékétől függően kétféle eljárás használatos a binomiális valószínűségek közelítő kiszámítására. Ha p értéke a következő határok közé esik:

$$(5.10.2) \quad \frac{0,637}{3\sqrt{n}} < p < 1 - \frac{0,637}{3\sqrt{n}}$$

(ahol n a kísérletek száma), akkor a célszerű közelítés:

$$(5.10.3) \quad P_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}$$

Ez a közelítő összefüggés azt jelenti, hogy ha n számú kísérletnél a p értéke a már ismertetett határok közé esik, binomiális eloszlás helyett np várható értékű, $\sqrt{npq}$ szórású normális eloszlással célszerű számolni. Amennyiben rögzített n esetén p az említett határokon kívül esik (azaz p a 0-hoz, vagy 1-hez esik közel), a binomiális eloszlást Poisson-eloszlással célszerű közelíteni. Az (5.10.3) formulával történő közelítés annál jobb, minél nagyobb az n értéke, és minél közelebb áll a p értéke az $\frac{1}{2}$ -hez, vagyis amikor $p \approx q = 1-p$.

Ha az (5.10.3) formulában bevezetjük az $x = \frac{k-np}{\sqrt{npq}}$ helyettesítést, akkor

$$P_k \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{npq}} = \varphi(x) \frac{1}{\sqrt{npq}},$$

ahol $\varphi(x)$ a standard normális sűrűségfüggvény.

A gyakorlatban ritkán van szükségünk a binomiális eloszlás egyes tagjainak értékére, sokkal többször bizonyos tagok összegére vagyunk kíváncsiak. Ha pl. egy célra leadott lövéseknél a találati valószínűség 0,6, és 100 lövést adunk le, akkor rendszerint nem az érdekel bennünket, hogy mi a valószínűsége - mondjuk - pontosan 57 találat elérésének, hanem pl. mi a valószínűsége annak, hogy a találatok száma 50 és 70 közé esik. Számítsuk ki ezt a valószínűséget a normális eloszlással való közelítés alkalmazásával.

$$\text{Most } M(\xi) = 100 \cdot 0,6 = 60, \quad D(\xi) = \sqrt{100 \cdot 0,6 \cdot 0,4} = \sqrt{24} \approx 5;$$

$$P(50 \leq \xi < 70) = \sum_{k=50}^{70} \binom{100}{k} 0,6^k 0,4^{100-k} \approx \sum_{k=50}^{70} \frac{1}{\sqrt{2\pi \cdot 24}} e^{-\frac{(k-60)^2}{2 \cdot 24}} \frac{1}{\sqrt{24}} \approx$$

$$\approx \frac{1}{\sqrt{2\pi}} \int_{\frac{50-60}{\sqrt{24}}}^{\frac{70-60}{\sqrt{24}}} e^{-\frac{x^2}{2}} dx = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 = 0,9544.$$

A példából látható, hogy a normális eloszlással történő közelítés során a binomiális valószínűségekre vonatkozó kérdések rendkívül könnyen, egész minimális számolással megválaszolhatók. Erre a közelítésre először Moivre és Laplace jutott.

Az (5. 10. 3) összefüggés matematikai származtatása az analízis eszközeivel egyszerű, nem igényel mást, csupán a Stirling-formula alkalmazását és a $\log(1+x)$ függvény Taylor-sorának ismeretét.

A Stirling-formula alkalmazásával:

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{k!(n-k)!} \approx \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\frac{k}{e}\right)^k \left(\frac{n-k}{e}\right)^{n-k} \sqrt{2\pi k} \sqrt{2\pi(n-k)}} = \\ &= \frac{1}{\sqrt{2\pi n \frac{k}{n} \left(1 - \frac{k}{n}\right)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}}. \end{aligned}$$

Ennek alapján a P_k valószínűség a következő alakban írható:

$$P_k = \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi n \frac{k}{n} \left(1 - \frac{k}{n}\right)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}.$$

Alkalmazzuk az $x = k - np$ helyettesítést, azaz legyen

$$k = np + x,$$

$$n - k = nq - x, \text{ akkor}$$

$$(5. 10. 4) \quad P_k \approx \frac{1}{\sqrt{2\pi n \left(p + \frac{x}{n}\right) \left(q - \frac{x}{n}\right)}} \cdot \frac{1}{\left(1 + \frac{x}{np}\right)^{np+x} \left(1 - \frac{x}{nq}\right)^{nq-x}}.$$

Könnyű belátni, hogy n növekedésével $\frac{x}{n} \rightarrow 0$, így az első tényező

$\frac{1}{\sqrt{2\pi npq}}$ -hoz tart, amint az az (5. 10. 3) formulában is szerepel. A második

tényező nevezőjének logaritmusai:

$$\begin{aligned}
 & (np + x) \log \left(1 + \frac{x}{np} \right) + (nq - x) \log \left(1 - \frac{x}{nq} \right) = \\
 & = (np + x) \left[\frac{x}{np} - \frac{x^2}{2n^2 p^2} + \frac{x^3}{3n^3 p^3} - \dots \right] - \\
 & - (nq - x) \left[\frac{x}{nq} + \frac{x^2}{2n^2 q^2} + \frac{x^3}{3n^3 q^3} + \dots \right] = x - \frac{x^2}{2np} + \frac{x^3}{3n^2 p^2} - \\
 & - \dots + \frac{x^2}{np} - \frac{x^3}{2n^2 p^2} + \dots - x - \frac{x^2}{2nq} - \frac{x^3}{3n^2 q^2} - \dots + \frac{x^2}{nq} + \\
 & + \frac{x^3}{2n^2 q^2} + \dots = \frac{x^2}{2n} \left(\frac{1}{p} + \frac{1}{q} \right) - \frac{x^3}{6n^2} \left(\frac{1}{p^2} - \frac{1}{q^2} \right) - \dots \approx \frac{x^2}{2npq}.
 \end{aligned}$$

A második tagtól kezdve a tagokat elhanyagolhatjuk, mert már $\frac{x^2}{n} \rightarrow 0$,

ha $n \rightarrow +\infty$. Ennek belátására gondoljuk meg, hogy $x = k - np$, vagyis a k valószínűségi változónak a várható értékétől való eltérése $\sqrt{n}$ nagyságrendű, ez a Csebisev-egyenlőtlenségből adódik. Következésképp $\frac{x^3}{n^2} \sim \frac{n^{3/2}}{n^2} = \frac{1}{\sqrt{n}} \rightarrow 0$, ha $n \rightarrow +\infty$. Az (5. 10. 4) formula jobb oldalának második tagjában szereplő nevező tehát $\sim e^{\frac{x^2}{2npq}}$, így a P_k valószínűség:

$$P_k \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}}.$$

5.11 Kétdimenziós (síkbeli) normális eloszlás

A többdimenziós valószínűségeloszlások közül részletesen csak a gyakorlatban legnagyobb jelentőségű síkbeli normális eloszlást vizsgáljuk.

A $\xi = (\xi, \eta)$ véletlen vektort kétdimenziós normális eloszlásnak nevezzük, ha sűrűségfüggvénye:

$$(5.11.1) \quad h(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho}} e^{-\frac{1}{2(1-\varrho)^2} \left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\varrho \frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2} \right]}$$

E kétváltozós sűrűségfüggvény geometriai képe egy folytonos felület, amelyről közelítő képet kapunk, ha egy harangot képzelünk magunk elé.

A formulában öt paraméter (állandó) szerepel, amelyek értékétől függően a felület vagy valóban harang alakú, vagy pedig két oldalról összenyomott haranghoz hasonlít. Ennek a hozzávetőleges hasonlatnak pontosabb tartalmat fogunk adni a továbbiakban végzendő geometriai jellegű vizsgálatok alapján. Előzőleg azonban megvizsgáljuk az (5.11.1) formulában szereplő paraméterek jelentését.

Kimutatjuk, hogy m_1 , ill. σ_1 a ξ változó várható értékét és szórását, m_2 , ill. σ_2 az η változó várható értékét és szórását, ϱ pedig a ξ és η változók közötti korrelációs együtthatót jelenti. Ennek belátását a következő módon végezzük. A kétdimenziós eloszlásról mondottak alapján (lásd a ξ változó sűrűségfüggvénye:

$$f(x) = \int_{-\infty}^{\infty} h(x,y) dy$$

Az (5.11.1) formulában vezessük be az

$$u = \frac{x-m_1}{\sigma_1}, \quad v = \frac{y-m_2}{\sigma_2} \text{ helyettesítést, s ekkor}$$

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} h(x,y) dy = \frac{1}{2\pi\sigma_1\sqrt{1-\varrho^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\varrho^2)} [u^2 - 2\varrho uv + v^2]} dv = \\ &= \frac{e^{-\frac{u^2}{2}}}{2\pi\sigma_1\sqrt{1-\varrho^2}} \int_{-\infty}^{\infty} e^{-\frac{(v-\varrho u)^2}{2(1-\varrho^2)}} dv. \end{aligned}$$

Alkalmazva az $z = \frac{v-\varrho u}{\sqrt{1-\varrho^2}}$ változót a $dv = \sqrt{1-\varrho^2} dz$, és így:

$$\begin{aligned}
 f(x) &= \frac{e^{-\frac{u^2}{2}}}{2\pi\sigma_1} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \\
 &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \text{ azaz} \\
 &\quad -\frac{(x-m_1)^2}{2\sigma_1^2} \\
 (5.11.2) \quad f(x) &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}},
 \end{aligned}$$

amely eredmény mutatja, hogy ξ normális eloszlású, m_1 várható értékkel és σ_1 szórással. Mivel a (5.11.1) formula az x és y változóiban szimmetrikus, teljesen analóg megfontolással nyerjük az η változó $g(y)$ sűrűségfüggvényét:

$$(5.11.3) \quad g(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}},$$

tehát η is normális eloszlású, m_2 várható értékkel és σ_2 szórással. Megfontolásunk eredményeként nem csak az m_1 , m_2 , σ_1 és σ_2 paraméterek jelentése vált világossá, hanem egyben azt is láthatjuk, hogy ha a ξ és η valószínűségi változók együttes eloszlása kétdimenziós normális eloszlás, akkor az egyes komponensek külön-külön is normális eloszlást követnek, vagyis a kétdimenziós normális eloszlás vetületei egydimenziós normális eloszlások.

Kimutatjuk, hogy az (5.11.1) formulában szereplő ρ paraméter a ξ és η változók közötti korrelációs együttható. A ξ és η változók közötti kovariancia a (3.5.1) összefüggés alapján:

$$\begin{aligned}
 C = M(\xi - m_1)(\eta - m_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-m_1)(y-m_2) h(x,y) dx dy = \\
 &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}} dy.
 \end{aligned}$$

$$\int_{-\infty}^{\infty} (x-m_1) (y-m_2) e^{-\frac{1}{2(1-\varrho^2)} \left(\frac{x-m_1}{\sigma_1} - \varrho \frac{y-m_2}{\sigma_2} \right)^2} dx.$$

Ha most bevezetjük az

$$u = \frac{1}{1-\varrho^2} \left(\frac{x-m_1}{\sigma_1} - \frac{y-m_2}{\sigma_2} \right), \quad v = \frac{y-m_2}{\sigma_2} \text{ helyettesítéseket,}$$

akkor:

$$\begin{aligned} C &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\sigma_1 \sigma_2 \sqrt{1-\varrho^2} uv + \varrho \sigma_1 \sigma_2 v^2) e^{-\frac{u^2}{2} - \frac{v^2}{2}} du dv = \\ &= \frac{\varrho \sigma_1 \sigma_2}{2\pi} \int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du + \\ &+ \frac{\sigma_1 \sigma_2 \sqrt{1-\varrho^2}}{2\pi} \int_{-\infty}^{\infty} v e^{-\frac{v^2}{2}} dv \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du = \varrho \sigma_1 \sigma_2, \end{aligned}$$

vagyis

$$\varrho = \frac{M[(\xi - m_1)(\eta - m_2)]}{\sigma_1 \sigma_2}$$

amely éppen a korrelációs együttható definíciója.

Az (5. 11. 1) összefüggést szemügyre véve láthatjuk, hogy ha ξ és η korrelálatlanok, azaz $\varrho = 0$, a

$$\begin{aligned} h(x, y) &= \frac{1}{2\pi \sigma_1 \sigma_2} e^{-\frac{1}{2} \left[\frac{(x-m_1)^2}{\sigma_1^2} + \frac{(y-m_2)^2}{\sigma_2^2} \right]} = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} \cdot \\ &\cdot \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}} = f(x) g(y). \end{aligned}$$

Ekkor tehát a ξ és η változók együttes sűrűségfüggvénye az egyes változók sűrűségfüggvényeinek szorzata. Ez azonban azt jelenti, hogy ξ és η függetlenek. Azt az eredményt kaptuk tehát, hogy ha két valószínűségi változó együttes eloszlása kétdimenziós normális eloszlás és a változók korrelálatlanok, akkor egyben függetlenek is.

5.12 A logaritmikus normális eloszlás

Az η valószínűségi változót logaritmikus normális eloszlásúnak vagy röviden lognormális eloszlásúnak nevezzük, ha $\ln \eta$ normális eloszlású.

Legyen $\xi = \ln \eta$, ekkor tehát ξ normális eloszlású valamely m és σ paraméterekkel, vagyis ξ sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

A valószínűségi változó függvényének eloszlására vonatkozó formula alapján η sűrűségfüggvénye:

$$g(x) = \frac{1}{|x|} f(\ln x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-\frac{(\ln x - m)^2}{2\sigma^2}}, \quad (x > 0).$$

Meghatározzuk az η valószínűségi változó várható értékét és szórásnégyzetét.

$$M(\eta) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{(\ln x - m)^2}{2\sigma^2}} dx.$$

Alkalmazzuk az

$$\frac{\ln x - m}{\sigma} = u$$

helyettesítést, akkor

$$x = e^{\sigma u + m}, \quad dx = e^{\sigma u + m} du.$$

Ennek alapján:

$$M(\eta) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{\sigma v + m} e^{-\frac{v^2}{2}} dv = e^{m + \frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(u-\sigma)^2} du =$$

$$= \frac{1}{2} e^{m + \frac{\sigma^2}{2}},$$

$$M(\eta^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^{\infty} x e^{-\frac{(\ln x - m)^2}{2\sigma^2}} dx =$$

$$= e^{2m + 2\sigma^2} \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{1}{2}(v - 2\sigma)^2} dv = e^{2m + 2\sigma^2},$$

$$D^2(\eta) = M(\eta^2) - [M(\eta)]^2 = e^{2m + 2\sigma^2} - e^{2m + \sigma^2} = e^{2m + \sigma^2} (e^{\sigma^2} - 1).$$

Lognormális eloszlás általában aprítási, osztódási, mállási folyamatoknál lép fel, mint a végtermék súlyának, köbtartalmának vagy valamilyen méretének eloszlása.

5.13 Az egyenletes eloszlás

Az érmedobás, kockadobás stb. esetében a kísérlethez rendelt ξ valószínűségi változó mindegyik lehetséges értékét egyforma valószínűséggel veszi fel. Az ilyen valószínűségi változót egyenletes eloszlású változónak nevezzük. Megkülönböztetünk diszkrét és folytonos egyenletes eloszlást. Legyen ξ diszkrét valószínűségi változó $x_1, x_2, \dots, x_n$ (véges sok) lehetséges értékkel, és legyen

$$(5.13.1) \quad P(\xi = x_i) = \frac{1}{n}, \quad i = 1, 2, \dots, n.$$

Ekkor a ξ valószínűségi változót egyenletes eloszlásúnak nevezzük az $x_1, x_2, \dots, x_n$ számokon. A valószínűségszámítás klasszikus korszakában csak ilyen valószínűségi változókra vonatkozó problémákat vizsgáltak.

Diszkrét egyenletes eloszlás csak véges sok különböző kimenetel esetében értelmezhető. Diszkrét egyenletes eloszlású valószínűségi változóra vonatkozó valószínűségi kérdések kombinatorikus módszerekkel válaszolhatók meg. (Lásd: 1.5. pont.)

Az (5.13.1) eloszlású valószínűségi változó várható értékét és szórásnégyzetét a következő formulák szolgáltatják:

$$(5.13.2) \quad M(\xi) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x},$$

ami nem más, mint a lehetséges értékek számtani közepe.

$$(5.13.3) \quad D^2(\xi) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{\sum_{i=1}^n x_i}{n} + \frac{1}{n} \bar{x}^2 n =$$

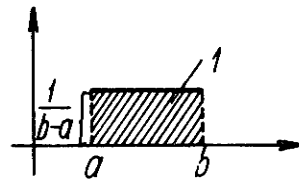
$$= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2.$$

Egy folytonos ξ valószínűségi változót egyenletes eloszlásúnak nevezünk az $[a, b]$ intervallumon, ha sűrűségfüggvénye a következő:

$$(5.13.4) \quad f(x) = \begin{cases} \frac{1}{b-a}, & \text{ha } a < x < b, \\ 0, & \text{egyébként.} \end{cases}$$

Az (5.13.4) sűrűségfüggvénnyel bíró valószínűségi változó eloszlásfüggvénye:

$$(5.13.5) \quad F(x) = \begin{cases} 0 & \text{ha } x < a \\ \frac{x-a}{b-a} & \text{ha } a \leq x < b \\ 1 & \text{ha } x \geq b. \end{cases}$$



47. ábra

A várható értéket és a szórásnégyzetet egyszerűen kiszámíthatjuk:

$$(5.13.6) \quad M(\xi) = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{a+b}{2}$$

$$(5.13.7) \quad D^2(\xi) = \int_a^b \frac{x^2}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

5.14 Néhány valószínűségeloszlás, amely a matematikai statisztikában fontos szerepet játszik

a) A χ^2 -eloszlás

A statisztikai mintával kapcsolatban ismertettük, hogy a mintaelemeket független, azonos eloszlású valószínűségi változóknak tekintjük. Amikor egy eloszlás ismeretlen paramétereinek becslése céljából a mintaelemekből különböző statisztikai mennyiségeket számítunk, ezek ugyancsak véletlen mennyiségek lesznek, hiszen értékük mintáról-mintára változik. A számunkra érdekes statisztikai mennyiségek tehát valószínűségi változók különböző függvényei, amelyek maguk is valószínűségi változók. A különböző statisztikai mennyiségek eloszlásának tárgyalása előtt szükséges megismernünk néhány eloszlás-típust, amelyeket a valószínűségeloszlások tárgyalása során nem részleteztünk.

Legyen a folytonos eloszlású ξ valószínűségi változó sűrűségfüggvénye:

$$(5.14.1) \quad f(x; \alpha, p) = \begin{cases} \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x} & \text{ha } x > 0 \\ 0 & \text{egyébként} \end{cases}$$

Itt a formulában szereplő $\Gamma(p)$ függvény az ún. paraméterű gamma függvény, amelynek definíciója

$$(5.14.2) \quad \Gamma(p) = \int_0^{\infty} x^{p-1} e^{-x} dx$$

Könnyű belátni, hogy $f(x; \alpha, p)$ valóban sűrűségfüggvény, mivel

$$(5.14.3) \int_0^{\infty} x^{p-1} e^{-\alpha x} dx = \frac{1}{\alpha^p} \int_0^{\infty} (\alpha x)^{p-1} e^{-\alpha x} d(\alpha x) = \frac{\Gamma(p)}{\alpha^p}, \text{ így}$$

$$\int_0^{\infty} f(x; \alpha, p) dx = 1.$$

Az $f(x; \alpha, p)$ sűrűségfüggvénnyel bíró ξ valószínűségi változót p -paraméterű gamma eloszlásnak nevezzük.

Számítsuk ki a p -paraméterű gamma eloszlás karakterisztikus függvényét!

$$(5.14.4) \varphi(t) = \int_0^{\infty} e^{itx} f(x; \alpha, p) dx = \frac{\alpha^p}{\Gamma(p)} \int_0^{\infty} x^{p-1} e^{-(\alpha - it)x} dx = \frac{1}{(1 - \frac{it}{\alpha})^p}$$

Itt felhasználtuk az (5.14.3) összefüggést, amely komplex α esetén is érvényes, feltéve, hogy annak valós része pozitív előjelű.

Az $f(x; \alpha, p)$ sűrűségfüggvényben az α és p paraméterek értékének különböző megválasztásával különböző, a gyakorlatban fontos eloszlásokat nyerünk. Pl. $p = 1$ választása mellett

$$(5.14.5) \quad f(x; \alpha, 1) = \alpha e^{-\alpha x}$$

a jól ismert exponenciális eloszlás, amely különböző élettartam vizsgálatokkal vagy a várakozási idővel kapcsolatban játszik szerepet.

Ha az (5.14.1) sűrűségfüggvényben $\alpha = \frac{1}{2}$, $p = \frac{n}{2}$ értékeket választunk, akkor az ún. χ^2 eloszlás sűrűségfüggvényét nyerjük:

$$(5.14.6) \quad k_n(x) = f(x; \frac{1}{2}, \frac{n}{2}) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \quad (x > 0)$$

Ez az eloszlás igen jelentős a statisztikai alkalmazásokban. Független normális eloszlású valószínűségi változók négyzetösszege χ^2 eloszlású. Pontosabban fogalmazva, ha $\xi_1, \xi_2, \dots, \xi_n$ független $N(0,1)$ standard - normális eloszlású valószínűségi változók és

$$(5.14.7) \quad \zeta = \xi_1^2 + \xi_2^2 + \dots + \xi_n^2$$

akkor a ζ változó sűrűségfüggvénye az (5.14.6) formulával megadott függvény.

Ezen állítás belátása céljából vizsgáljuk meg először az egyes összeadandók eloszlását. Ha $\xi \sim N(0,1)$ normális eloszlású és $\eta = \xi^2$, akkor η eloszlásfüggvénye:

$$(5.14.8) \quad G(x) = P(\eta < x) = P(\xi^2 < x) = P(-\sqrt{x} < \xi < \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1$$

Az η változó sűrűségfüggvénye

$$(5.14.9) \quad g(x) = 2\Phi'(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{x}} \varphi(x) = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

$$\left(\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\right)$$

Tehát az $\eta = \xi^2$ valószínűségi változó $\alpha = \frac{1}{2}$, $p = \frac{1}{2}$ paraméterekkel bíró gamma eloszlás. Az η változó karakterisztikus függvénye az (5.14.4) formula alapján:

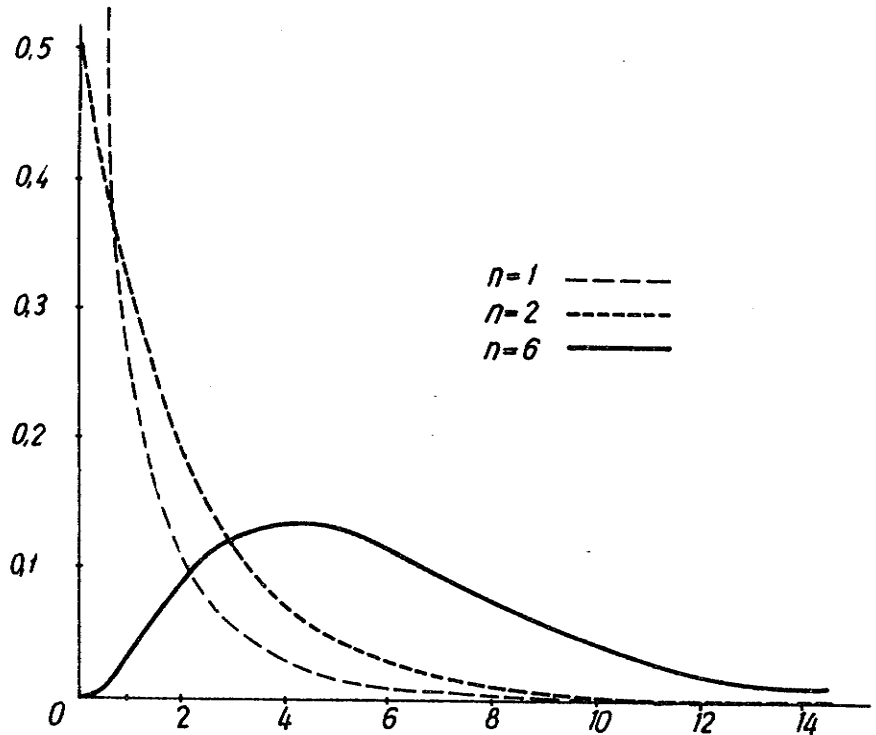
$$(5.14.10) \quad \varphi_{\eta}(t) = \frac{1}{(1 - 2it)^{\frac{1}{2}}}$$

Tekintettel arra, hogy az (5.14.7) összegben szereplő mindegyik ξ_i^2 ($i = 1, 2, \dots$) valószínűségi változó eloszlása megegyezik az η változó eloszlásával, a ζ változó karakterisztikus függvénye

$$5.14.11) \quad \varphi_{\zeta}(t) = \prod_{i=1}^n \varphi_{\xi_i^2}(t) = \left[\frac{1}{(1 - 2it)^{\frac{1}{2}}} \right]^n = \frac{1}{(1 - 2it)^{\frac{n}{2}}}$$

A (5.14.11) formulát összehasonlítva az (5.14.4) függvénnyel, láthatjuk, hogy ez utóbbi függvény az $\alpha = \frac{1}{2}$, $p = \frac{n}{2}$ paraméterekkel bíró gamma eloszlás karakterisztikus függvénye, vagyis ζ sűrűségfüggvénye az

(5.14.6) formulával megadott függvény, amelynek alakját $n = 1$, $n = 2$ és $n = 6$ esetében a 48. ábra szemlélteti.



48. ábra

Az (5.14.6) formulával kifejezett $k_n(x)$ sűrűségfüggvényt, amelyet n darab független $N(0,1)$ eloszlású valószínűségi változó négyzetösszegének eloszlásaként nyertünk, szokás n szabadságfokú χ^2 -eloszlásnak is nevezni. A szabadságfok értelmét a pontban világítjuk meg. Az (5.14.11) karakterisztikus függvény segítségével könnyen származtathatjuk az n -szabadságfokú χ^2 eloszlás momentumait. A várható érték, illetve a szórásnégyzet kiszámításához elegendő tekintetbe venni, hogy

$$\varphi'_\xi(0) = -\frac{n}{2} \left[(1-2it)^{-\frac{n}{2}-1} \right] (-2i) = in$$

$$\varphi''_\xi(0) = -\frac{n}{2} \left[-\left(\frac{n}{2}+1\right) \right] \left\{ \left[1-2it \right]^{-\frac{n}{2}-2} \right\}_{t=0} = i^2 (n^2+2n)$$

Ezek alapján a (4.2.7) és (4.2.8) formulák figyelembevételével

$$(5.14.12) \quad M(\zeta) = n; \quad D^2(\zeta) = -n^2 + n^2 + 2n = 2n$$

Ugyancsak a karakterisztikus függvény felhasználásával igazolható a χ^2 eloszlás un. additív tulajdonsága, amely abból áll, hogy két független χ^2 -eloszlású valószínűségi változó összege ugyancsak χ^2 -eloszlású. Ha ugyanis ζ_1 n_1 szabadságfokú χ^2 -eloszlású, és ζ_2 n_2 szabadságfokú χ^2 -eloszlású és ζ_1 és ζ_2 függetlenek, akkor a $\zeta = \zeta_1 + \zeta_2$ összeg karakterisztikus függvénye a megfelelő karakterisztikus függvények szorzata, vagyis az (5.14.11) formulát figyelembe véve:

$$\varphi_{\zeta}(t) = \varphi_{\zeta_1}(t) \cdot \varphi_{\zeta_2}(t) = (1 - 2it)^{-\frac{n_1}{2}} \cdot (1 - 2it)^{-\frac{n_2}{2}} = (1 - 2it)^{-\frac{n_1 + n_2}{2}}$$

vagyis a ζ változó ugyancsak χ^2 -eloszlású $n_1 + n_2$ szabadságfokkal.

Tehát független χ^2 -eloszlású valószínűségi változók összeadásakor a szabadságfokok összegeződnek, az eloszlás típusa nem változik. Meg kell azonban jegyeznünk, hogy ha sok χ^2 eloszlású valószínűségi változót adunk össze, amikor is a szabadságfok már elég nagy lesz, vagyis nagyszámú független $N(0,1)$ eloszlású valószínűségi változó négyzetösszegével állunk szemben, akkor a centrális határeloszlástétel érvényesül és normális eloszláshoz jutunk. Már pl. $n = 30$ esetén a χ^2 -eloszlás sűrűségfüggvénye alig különbözik egy normális eloszlás sűrűségétől!

b) A Student-féle t-eloszlás

A matematikai statisztika néhány fontos problémája az alábbi típusú hányados eloszlásának vizsgálatát követeli meg:

$$(5.14.13) \quad t = \frac{\sqrt{n} \bar{\xi}}{\sqrt{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}}$$

ahol $\xi_1, \xi_2, \dots, \xi_n$ független $N(0,1)$ normális eloszlású valószínűségi változók.

A t valószínűségi változó sűrűségfüggvényének meghatározásához ki kell számítanunk mind a számláló, mind a nevező sűrűségfüggvényét,

majd alkalmaznunk kell a független valószínűségi változók hányadosának sűrűségfüggvényére vonatkozó formulát.

A 2.6 pont alapján ismeretes, hogy ha az η valószínűségi változó egy ξ valószínűségi változónak folytonos függvénye:

$$\eta = \varphi(\xi)$$

és a ξ változó sűrűségfüggvénye $f(x)$, akkor η sűrűségfüggvénye:

$$h(y) = f[\varphi^{-1}(y)] \left| \frac{d\varphi^{-1}(y)}{dy} \right|$$

Számítsuk ki ennek alapján az (5.14.13) tört számlálójának sűrűségfüggvényét.

Most $\eta = \sqrt{n} \xi$, $\varphi^{-1}(y) = \frac{1}{\sqrt{n}} y$, így a $\sqrt{n} \xi$ változó sűrűsége:

$$h(x) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{x^2}{2n}}$$

Az (5.14.13) tört nevezőjének sűrűségfüggvénye az n szabadságfokú χ^2 -eloszlás sűrűségfüggvényének ismeretében könnyen adódik, ugyanis a nevező $\sqrt{\xi}$, ahol ξ n -szabadságfokú χ^2 -eloszlású, így a nevező sűrűsége:

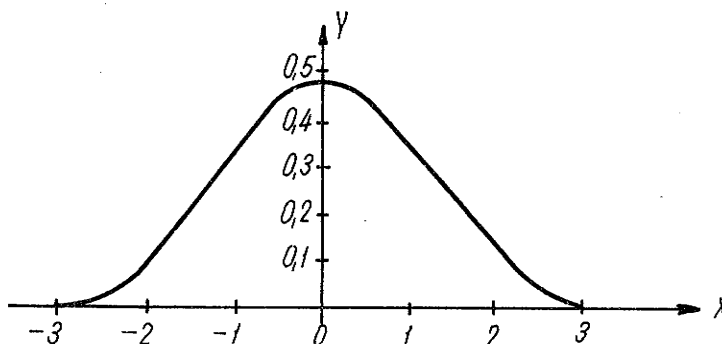
$$g(y) = \frac{2 y^{n-1} e^{-\frac{y^2}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}$$

Ha most a t valószínűségi változó sűrűségfüggvényét $s_n(x)$ -szel jelöljük, akkor a (2.6.11) formula alapján:

$$s_n(x) = \int_0^{\infty} y h(x, y) g(y) dy = \frac{2}{\sqrt{2\pi n} 2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \int_0^{\infty} y^n e^{-\left(\frac{x^2}{n} + 1\right) \frac{y^2}{2}} dy$$

Alkalmazva az $n = \left(\frac{x^2}{n} + 1\right) \frac{y^2}{2}$ helyettesítést, az

$$(5.14.14) s_n = \frac{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}{\sqrt{\pi n} \Gamma(\frac{n}{2})} \int_0^\infty u^{\frac{n-1}{2}} e^{-u} du = \frac{1}{\sqrt{\pi n}} \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}$$



49. ábra

összefüggésre jutunk, amely a t változó sűrűségfüggvénye, s amely sűrűséggel bíró eloszlást Student-féle t -eloszlásnak nevezünk. Ezt az eloszlást először W.S. Gosset angol matematikus használta, aki "Student" álnéven publikálta eredményeit. A χ^2 eloszláshoz hasonlóan az (5.14.14) sűrűségű eloszlást szokás a szabadságfokú Student-eloszlásnak is nevezni. Ezen eloszlás alkalmazásaival a 9.4. pontban foglalkozunk. A Student-eloszlás sűrűségfüggvényét a 49. ábra szemlélteti.

c) χ^2 -eloszlású valószínűségi változók hányadosának eloszlása

A matematikai statisztikában, elsősorban a szórásanalízisben felmerül a következő típusú probléma: adva vannak a $\xi_1, \xi_2, \dots, \xi_{n+m}$ független $N(0,1)$ normális eloszlású valószínűségi változók, meghatározandó a

$$(5.14.15) \quad \zeta = \frac{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}{\xi_{n+1}^2 + \xi_{n+2}^2 + \dots + \xi_{n+m}^2}$$

hányados sűrűségfüggvénye. Az előzőek (5.14/a pont) alapján látható, hogy a tört számlálója n szabadságfokú χ^2 -eloszlású, nevezője pedig ugyancsak χ^2 -eloszlású m szabadságfokkal, s a feltételezett függetlenség következtében a számláló független a nevezőtől. Ha a ζ változó sűrűségfüggvényét $h(z)$ -vel jelöljük, akkor a független valószínűségi változók hányadosának sűrűségfüggvényére vonatkozó (2.6.11) formula alapján

$$(5.14.16) \quad h(z) = \frac{z^{\frac{n}{2}-1}}{2^{\frac{n+m}{2}} \Gamma(\frac{n}{2}) \Gamma(\frac{m}{2})} \int_0^{\infty} y^{\frac{n+m}{2}-1} e^{-\frac{y(1+z)}{2}} dy =$$

$$= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2}) \Gamma(\frac{m}{2})} \frac{z^{\frac{n}{2}-1}}{(1+z)^{\frac{n+m}{2}}}$$

d) A béta - eloszlás

Több statisztikai alkalmazásnál, elsősorban a rendezett minták elméletében játszik szerepet az un. béta eloszlás, amely a (5.14.15) formulával definiált ζ valószínűségi változó egy függvényének a

$\kappa = \frac{\zeta}{1+\zeta}$ valószínűségi változónak az eloszlása.

Ha $k(x)$ -szel jelöljük a κ változó eloszlásfüggvényét, akkor

$$\kappa(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2}) \Gamma(\frac{m}{2})} \int_0^x \frac{z^{\frac{n}{2}-1}}{(1+z)^{\frac{n+m}{2}}} dz$$

Az integrandusban $\frac{z}{1+z} = t$ helyettesítést alkalmazva a

$$(5.14.7) \quad k(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2}) \Gamma(\frac{m}{2})} \int_0^x t^{\frac{n}{2}-1} (1-t)^{\frac{m}{2}-1} dt \quad (0 \leq x \leq 1)$$

Ez az un. Euler-féle béta integrál 0-tól x -ig, amelyet szokás nem teljes béta-függvénynek is nevezni.

A $k(x)$ függvény speciális esete az un. béta-eloszlásoknak. Általánosabb alakban (α, β) rendű béta eloszlásnak nevezzük azt az eloszlást, amelynek eloszlásfüggvénye

$$B_{\alpha, \beta}(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \int_0^x t^{\alpha-1} (1-t)^{\beta-1} dt \quad (0 \leq x \leq 1)$$

Láthatjuk, hogy a $k(x)$ függvény $B_{\frac{n}{2}, \frac{m}{2}}(x)$ alakban írható, tehát $(\frac{n}{2}, \frac{m}{2})$ paraméterű béta eloszlás.

5.15 A nagy számok törvényének Bernoulli-féle alakja

A gyakorlatban igen sokszor lép fel az a probléma, hogy valamely bennünket érdeklő A esemény ismeretlen $P(A) = p$ valószínűségét n számú kísérlet alapján a $\frac{k}{n}$ relatív gyakorisággal közelítve, mennyire tér el a relatív gyakoriság az ismeretlen valószínűségtől. Arról a tényről, hogy n esetén a $\frac{k}{n}$ relatív gyakoriság tetszőleges kevéssel tér a p valószínűségtől, meggyőz bennünket a nagy számok törvényének Bernoullitól származó megformulázása.

Alkalmazzuk a Csebisev egyenlőtlenséget a binomiális eloszlásra. A (3.3.11) formulával kifejezett

$$P(|\xi - M(\xi)| > \lambda \sigma) \leq \frac{1}{\lambda^2}$$

összefüggésben $\xi = k$, $M(\xi) = np$, $\sigma = \sqrt{npq}$ helyettesítéssel:

$$P(|k - np| > \lambda \sqrt{npq}) \leq \frac{1}{\lambda^2}$$

A zárójelen belüli egyenlőtlenséget n -nel osztva:

$$P\left(\left|\frac{k}{n} - p\right| > \lambda \sqrt{\frac{pq}{n}}\right) \leq \frac{1}{\lambda^2}$$

Vezessük be a $\lambda \sqrt{\frac{pq}{n}} = \varepsilon$ jelölést, ekkor

$$\frac{1}{\lambda^2} = \frac{q p}{n \varepsilon^2} \leq \frac{1}{4 n \varepsilon^2}$$

Tehát:

$$(5.15.1) \quad P\left(\left|\frac{k}{n} - p\right| > \varepsilon\right) \leq \frac{1}{4 n \varepsilon^2}$$

Az (5.15.1) összefüggés a nagy számok törvényének Bernoulli-féle alakja. (5.15.1) jobb oldala $n \rightarrow +\infty$ esetén nyilván zérushoz tart, ami azt jelenti, hogy a kísérletek számát korlátlanul növelve, a relatív gyakoriságnak a valószínűségtől való eltérése tetszőleges nagy biztonsággal előírt kicsiny lesz. A gyakorlatban nincs módunk a kísérletek számát korlátlanul növelni, ezért elsődleges érdekességű számunkra, hogy rögzített n esetén mekkora lesz nagy valószínűséggel a $\frac{k}{n}$ relatív gyakoriságnak a p valószínűségtől való eltérése, illetve milyen nagyra kell n -et, a kísérletek számát választani ahhoz, hogy a $\left|\frac{k}{n} - p\right|$ eltérés előírt ε -nál kisebb legyen. Erre a kérdésre az (5.15.1) összefüggés alapján is válaszolhatnánk, ekkor azonban a szükségesnél nagyobb n értéket nyernénk. Ennek okát a (3.3) pont végén tett megjegyzésünk világítja meg.

Láttuk, hogy a binominális eloszlás az 5.10 pontban említett feltételek között normális eloszlással közelíthető. A 2. táblázatból látható, hogy egy normális eloszlású ξ valószínűségi változó igen nagy valószínűséggel kevesebbel tér el a várható értékétől, mint szórásának a háromszorosa, azaz

$$P(|\xi - m| < 3\sigma) \approx 0,997$$

Tehát közelítőleg ugyancsak 0,997 annak a valószínűsége, hogy az A esemény k gyakorisága az np várható értékétől kevesebbel térjen el, mint szórásának a háromszorosa, tehát

$$(5.15.2) \quad P(|k - np| < 3\sqrt{npq}) \approx 0,997$$

Más szavakkal olyan kevésbé valószínű, hogy a $|k - np|$ nagyobb legyen, mint $3\sqrt{npq}$, hogy gyakorlatilag alig kell vele számolnunk. Ez az ún. 3 σ -szabály.

Az (5.15.2) képletben szabad az egyenlőtlenség mindkét oldalát n -nel osztani, anélkül, hogy a valószínűség megváltozzék, azaz

$$(5.15.3) \quad P\left(\left|\frac{k}{n} - p\right| < 3\sqrt{\frac{pq}{n}}\right) \approx 0,997$$

Ez azt mutatja, hogy igen nagy valószínűséggel a $\frac{k}{n}$ relatív gyakoriság kevesebbel tér el az ismeretlen p valószínűségtől, mint $3 \sqrt{\frac{pq}{n}}$.

Figyelembe véve, hogy a $pq = p(1-p)$ kifejezés értéke nem lehet nagyobb, mint $\frac{1}{4}$ ^{*}, gyakorlatilag biztosra vehetjük, hogy

$$\left| \frac{k}{n} - p \right| < \frac{3}{2\sqrt{n}} = \frac{1,5}{\sqrt{n}}$$

azaz csaknem biztos, hogy a p valószínűség a $\frac{k}{n} - \frac{1,5}{\sqrt{n}}$ és $\frac{k}{n} + \frac{1,5}{\sqrt{n}}$ számok közé esik.

Ez a hibahatár akkor tekinthető jónak, ha az ismeretlen valószínűség $\frac{1}{2}$ közelében van. Ha a p valószínűség lényegesen eltér $\frac{1}{2}$ -től, akkor pq is kisebb mint $\frac{1}{4}$, tehát a relatív gyakoriság szórása $\sqrt{\frac{pq}{n}}$ kisebb, mint $\frac{1}{2\sqrt{n}}$.

Ilyen esetben a p valószínűség he-

lyett a $\frac{k}{n}$ relatív gyakoriság érté-

két írjuk és $q = 1-p$ helyébe

$1 - \frac{k}{n} = \frac{n-k}{n}$ -et írunk, így

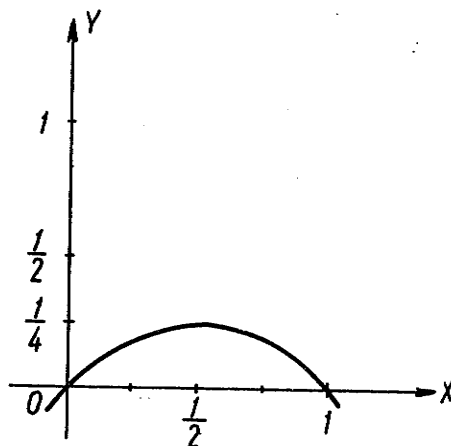
$$\sqrt{\frac{pq}{n}} \approx \frac{1}{n} \sqrt{k(n-k)}$$

Még jobb becslést kapunk a szórásra, ha a gyök alatti kifejezés nevezőjében n helyett $n-1$ -et írunk.

Igen nagy valószínűséggel gyakorlatilag ugyszólván biztosan igaz tehát, hogy

$$\left| \frac{k}{n} - p \right| < \frac{3}{n} \sqrt{\frac{k(n-k)}{n-1}}$$

Ha arra akarunk válaszolni, mekkorára kell n -et, a kísérletek számát választani, hogy a $\frac{k}{n}$ relatív gyakoriság az ismeretlen p valószínűség-



50. ábra

^{*}) Ugyanis az $y = x(1-x)$ függvény maximumát csak ott veheti fel, ahol $y' = 1 - 2x = 0$, vagyis az $x = \frac{1}{2}$ helyen. Mivel $y'' = -2 < 0$, valóban maximumról van szó.

től csaknem biztosan kevesebbel térjen el, mint pl. $\varepsilon = 0,1$, akkor az (5.15.3) összefüggés alapján csak az

$$\frac{1,5}{\sqrt{n}} = 0,1$$

egyenletet kell megoldanunk, amelyből $n = 225$ adódik.

Az $\varepsilon = 0,1$ érték a gyakorlatban sokszor nem kielégítő pontosság. Ha nagyobb pontosságra törekszünk, akkor viszont nagyon megnövekszik a szükséges kísérletek száma. Ha pl. $\varepsilon = 0,05$ -öt választunk, az (5.15.3) egyenletből

$$n = 900$$

adódik, míg $\varepsilon = 0,01$ esetén, ez már igen nagy pontosság, amire a gyakorlatban ritkán van szükség a szükséges kísérletek száma:

$$n = 22\ 500.$$

5.16 A központi határeloszlástétel

A központi vagy centrális határeloszlástétel magyarázza meg annak elméleti hátterét, hogy miért lép fel a természeti jelenségek vizsgálatánál olyan gyakran normális eloszlás. A tétel tartalmának lényege abban áll, hogy ha a vizsgált ξ valószínűségi változó véletlen ingadozása sok egymástól független véletlen komponens eredménye, amely komponensek külön-külön kis mértékben hatnak, de hatásuk összegeződik, akkor ξ normális eloszlású.

A binomiális eloszlás vizsgálatánál láttuk, hogy a binomiális eloszlású ξ valószínűségi változó n darab független karakterisztikus változó összege:

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n$$

ahol az egyes ξ_i változók egymástól függetlenül veszik fel a 0 vagy 1 értéket, és mint kimutattuk, a binomiális eloszlás normális eloszlással közelíthető, ha csak a p paraméter értéke nem túl kicsi, vagy nem túl nagy; más szóval, ha az egyes ξ_i változók nem olyan természetűek, hogy csaknem mindig egyik értéküket veszik fel. (Ugy is mondhatjuk, hogy ha a ξ_i változók eloszlása nem túlságosan elfajult.)

A binomiális eloszlásnak normális eloszlással való közelíthetősége csak speciális esete, bár nagyon fontos speciális eset, egy sokkal általá-

nosabb törvényszerűségnek, amelyet úgy fogalmazhatunk meg, hogy ha igen nagyszámu egymástól független valószínűségi változót adunk össze és az egyes komponensek véges szórásnégyzettel bírnak, akkor az összegük normális eloszlású lesz, függetlenül attól, hogy az egyes komponensek milyen eloszlásúak.

A centrális határeloszlás tételének többféle matematikai megfogalmazása ismeretes. Mivel a matematikai statisztikában rendszerint független, egyforma eloszlású komponensekkel állunk szemben, először a tételnek azt a változatát ismertetjük, amikor a független komponensek azonos eloszlásúak.

Tétel: Ha $\xi_1, \xi_2, \dots$, azonos eloszlású, független és véges szórásu valószínűségi változók, $M(\xi_k) = m$, $D^2(\xi_k) = \sigma^2$, $k = 1, 2, \dots$, akkor

$$(5.16.1) \lim_{n \rightarrow \infty} P\left(\frac{\xi_1 + \xi_2 + \dots + \xi_n - nm}{\sigma \sqrt{n}} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$

A tétel bizonyítása karakterisztikus függvény alkalmazásával végezhető el legegyszerűbben.

Jelöljük $\varphi(t)$ -vel a $\xi_k - m$ valószínűségi változó karakterisztikus függvényét. Ekkor a $\frac{\xi_k - m}{\sigma \sqrt{n}}$ változó karakterisztikus függvénye

$$\varphi\left(\frac{t}{\sigma \sqrt{n}}\right); (k = 1, 2, \dots)$$

Legyen:

$$\eta = \frac{\xi_1 + \xi_2 + \dots + \xi_n - nm}{\sigma \sqrt{n}} = \sum_{k=1}^n \frac{\xi_k - m}{\sigma \sqrt{n}}$$

A komponensek függetlensége miatt az összeg karakterisztikus függvénye az összeadandók karakterisztikus függvényeinek szorzata, tehát:

$$\varphi_{\eta}(t) = \left[\varphi\left(\frac{t}{\sigma \sqrt{n}}\right) \right]^n$$

Vezessük be a $\frac{t}{\sigma \sqrt{n}} = u$ jelölést és fejtjük Taylor-sorba a $\varphi(u)$ függvényt:

$$\varphi(u) = \varphi(0) + \frac{\varphi'(0)}{1!} u + \frac{\varphi''(0)}{2!} u^2 + \dots$$

A karakterisztikus függvény $u = 0$ helyen vett deriváltjai kifejezhetők a (4.2.10) formula alapján a momentumok segítségével:

$$\varphi(u) = 1 + \alpha_1 i u + \alpha_2 \frac{(iu)^2}{2!} + \dots + \alpha_k \frac{(iu)^k}{k!} + o(u^k)$$

ahol α_j a $\frac{\xi_k - m}{\sigma \sqrt{n}}$ változó j -edik momentuma.

$$\alpha_1 = M \left(\frac{\xi_k - m}{\sigma \sqrt{n}} \right) = 0$$

$$\alpha_2 = M \left(\frac{\xi_k - m}{\sigma \sqrt{n}} \right)^2 = \frac{1}{n \sigma^2} M (\xi_k - m)^2 = \frac{1}{n}; \dots$$

Ezek alapján:

$$\varphi\left(\frac{t}{\sigma \sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n \sigma^2}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

Tehát

$$\lim_{n \rightarrow \infty} \left[\varphi\left(\frac{t}{\sigma \sqrt{n}}\right) \right]^n = \lim_{n \rightarrow \infty} \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right]^n = e^{-\frac{t^2}{2}}$$

Mint a normális eloszlás tárgyalásánál láttuk (lásd (5.9.13) formula), ez éppen a standard normális eloszlás karakterisztikus függvénye.

A tétel következő megfogalmazása nem köti ki, hogy az összeadandók azonos eloszlásúak legyenek, de feltételezi azok harmadik abszolút momentumainak létezését.

Laplace-Ljapunov tétel: Ha $\xi_1, \xi_2, \dots, \xi_n, \dots$ független valószínűségi változók, amelyek harmadik és következőleg alacsonyabb rendű momentumai léteznek, $M(\xi_k) = 0$, $(k = 1, 2, \dots)$ továbbá

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \beta_k}{S_n^3} = 0$$

ahol

$$\beta_k = M(|\xi_k|^3), S_n = \sqrt{\sum_{k=1}^n \sigma_k^2}; \sigma_k^2 = D^2(\xi_k), \text{ akkor:}$$

$$\lim_{n \rightarrow \infty} P\left(\frac{\xi_1 + \xi_2 + \dots + \xi_n}{S_n} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$

A centrális határeloszlástétel igen nagy gyakorlati jelentőségű. Egyrészt ha a tétel feltételei teljesülnek legalábbis jó közelítéssel, akkor jogosná teszi azt a feltételezésünket, hogy normális eloszlással állunk szemben, amelyet a megfigyelési adatok alapján a matematikai statisztika módszereivel ellenőrizni tudunk. Másrészt a matematikai statisztikában szereplő un. statisztikai függvényekben az esetek nagy részében független valószínűségi változók összege szerepel, így nagyszámu megfigyelés esetén maguk a statisztikai függvények mint valószínűségi változók ugyancsak normális eloszlásúak, így a normális eloszlás sajátosságai és a normális eloszlás függvénytáblázata a statisztikai döntések meghozatalában alapvető segédeszközünk lesz.

MARKOV-LÁNCOK

6.1 A Markov-lánc fogalma

Több gyakorlati probléma megoldásában sikerrel alkalmazható a következő modell: adva van egy fizikai rendszer, amelynek véges sok állapota van, s egymás után következő állapotait véletlenszerűen változtatja. A rendszer állapotait valamely $t_0, t_1, t_2, \dots$ időpontokban figyeljük meg.

(Feltételezzük, hogy a megfigyeléseket szabályos időközökben végezzük.) Az időegység megválasztása önkényes. Példa ilyen fizikai rendszerre egy víztározó, amelynek mondjuk $K \text{ m}^3$ a köbtartalma és azt mondjuk a tározó a t időpontban az i állapotban van, ha $i \text{ m}^3$ víz van benne. A tározót egy folyó vize táplálja, s a víz hozzáfolyása következtében a tározó "magasabb állapotba lép", növekszik a benne levő víz mennyisége, míg az időnkénti leeresztés (vízkivétel ipari célra vagy a túlfolyás elkerülésére) következtében a tározó víztartalma csökken, a rendszer "alacsonyabb állapotba lép".

A vizsgált fizikai rendszerünk állapota bármely időpontban valószínűségi változó. Ha a rendszer a $t = t_n$ időpontban az i állapotban van, akkor $\xi_{t_n} = i$.

A $\xi_{t_0}, \xi_{t_1}, \dots, \xi_{t_n}, \dots$ valószínűségi változók felvett értékei megadják a rendszer mozgásának, élettörténetének leírását. ξ_{t_0} a kiinduló állapot, amely adott esetben lehet véletlenszerű állapot is.

Ha $\xi_{t_k} = i$ és $\xi_{t_{k+1}} = j$, akkor azt mondjuk, hogy a rendszer egy lépésben az i állapotból a j állapotba ment át ($i, j = 0, 1, \dots, n$), vagy röviden $i \rightarrow j$ átmenet történt. Az $i \rightarrow j$ egy lépéses átmenet valószínűségét P_{ij} -vel jelöljük, ami lényegében a következő feltételes valószínűség:

$$P_{ij} = P(\xi_{t_{k+1}} = j \mid \xi_{t_k} = i)$$

Ha az $i \rightarrow j$ átmenet valószínűsége nem függ attól, hogy mely időpontban találjuk a rendszert az i állapotban, azaz a $P(\xi_{t_{e+1}} = j \mid P_{t_e} = i)$ feltételes valószínűség ugyanakkor a minden $\ell = 0, 1, 2, \dots$ értékre, akkor a

rendszer állapotváltozásait időben homogénnek vagy röviden homogén állapotváltozásoknak nevezzük. A továbbiakban csak homogén állapotváltozásokkal foglalkozunk. (Ez olykor erős egyszerűsítést jelenthet gyakorlati szempontból, azonban időben inhomogén állapotváltozások tárgyalása matematikailag sokkal bonyolultabb.)

A rendszer állapotváltozásainak sorozatát, vagy ami ugyanaz a $\xi_{t_0}, \xi_{t_1}, \dots, \xi_{t_n}, \dots$ valószínűségi változók sorozatát homogén Markov-láncnak nevezzük, ha tetszőleges pozitív értékű $t_1 < t_2 < \dots < t_{n+1}$ időpontok és $k_1, k_2, \dots, k_{n+1}$ állapotok esetén ($k_i = 0, 1, 2, \dots, k$; $i = 1, 2, \dots, n+1$)

$$(6.1.1) P(\xi_{t_{n+1}} = k_{n+1} \mid \xi_{t_1} = k_1, \xi_{t_2} = k_2, \dots, \xi_{t_n} = k_n) = \\ = P(\xi_{t_{n+1}} = k_{n+1} \mid \xi_{t_n} = k_n)$$

A (6.1.1) összefüggés azt jelenti, hogy a rendszernek a t_n időpont utáni élettörténete csak attól függ, hogy a t_n időpontban mely állapotban van a rendszer és közvetlenül nem függ attól, hogy korábbi időpontokban milyen állapotsorozatot futott be, mely állapotokon keresztül jutott a t_n időpontbeli k_n állapotba. Ugy is mondhatnánk, hogy a rendszer lépésről-lépésre elfelejti korábbi multját.

A már említett víztározók esetében pl. ez a feltétel nyilván teljesül, hiszen ha a tározó a t_n időpontban k_n m³ vizet tartalmaz, akkor a t_{n+1} időpontban $k_n + s$ állapotban lesz, ha egy időegység alatt s m³ víz folyik bele, vagy $k_n - \ell$ állapotba kerül, ha a tározót t_n időpontban lezárjuk és ℓ m³ vizet leeresztünk, függetlenül attól, hogy korábban milyen ütemű volt a feltöltés vagy hogy hányszor és mennyi vizet eresztettünk le belőle.

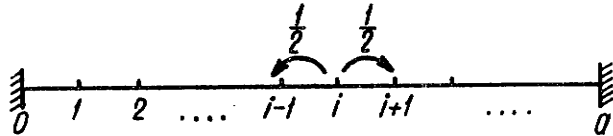
Mutatunk most egy másik egyszerű példát homogén Markov-láncre. Tegyük fel, hogy egy részecske bolyongást végez a számegyenes egész értékű pontjain, pl. az $x = 0$ pontból indulva és időegységenként egyet lép jobbra vagy balra $\frac{1}{2} \leftrightarrow \frac{1}{2}$ valószínűséggel. Az egyes lépésekhez hozzárendelhetjük az $\eta_1, \eta_2, \dots, \eta_n, \dots$ független valószínűségi változókat a következő eloszlással:

$$P(\eta_1 = 1) = \frac{1}{2}, \quad P(\eta_1 = -1) = \frac{1}{2}, \quad (i = 1, 2, \dots)$$

A bolyongó részecske helyét $t = n$ időpontban jelölje a ξ_n valószínűségi változó. Ekkor

$$\xi_n = \eta_1 + \eta_2 + \dots + \eta_n, \quad n \geq 1.$$

Ugyancsak homogén Markov-láncot kapunk, ha a részecske útjába a számegyenes adott szakaszának két végpontjába elnyelő vagy visszaverő falakat helyezünk:



51. ábra

Átmeneti valószínűségek:

A Markov-láncokra szokásos terminológiával azt mondjuk, hogy a vizsgált fizikai rendszer időegységeként "lép" egyet, azaz változtat állapotán, lépésnek tekintve az $i \rightarrow i$ átmenetet is, amikor tehát nem történik állapotváltozás, azaz a rendszer "helyben marad".

Jelöljük $P_{ij}^{(r)}$ -rel annak valószínűségét, hogy a rendszer az i állapotból r lépésben megy át a j állapotba:

$$P_{ij}^{(r)} = P(\xi_{t_{k+r}} = j \mid \xi_{t_k} = i)$$

$r = 1$ esetén a már bevezetett egylépéses átmenetvalószínűségeket kapjuk, amelyeket egy mátrixba rendezhetünk el, amelyet egylépéses átmenetmatrixnak nevezünk és $\underline{P}$ -vel fogunk jelölni.

$$(I) \quad \underline{P} = \begin{pmatrix} P_{00} & P_{01} & \dots & P_{0k} \\ P_{10} & P_{11} & \dots & P_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & & P_{kk} \end{pmatrix}$$

Az r -lépéses átmenetvalószínűségeket természetesen ugyancsak elrendezhetjük matrixba, jelöljük ezt a matrixot $\underline{P}_r$ -rel és nevezzük r -lépéses átmenetmatrixnak.

$$\underline{P}_r = \begin{pmatrix} P_{00}^{(r)} & P_{01}^{(r)} & \dots & P_{0k}^{(r)} \\ P_{10}^{(r)} & P_{11}^{(r)} & \dots & P_{1k}^{(r)} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1}^{(r)} & P_{k2}^{(r)} & \dots & P_{kk}^{(r)} \end{pmatrix}$$

Az egylépéses és r -lépéses átmenetmatrixok között igen egyszerű összefüggés áll fenn, ugyanis

$$\underline{P}_r = \underline{P}^r$$

azaz az r -lépés átmenetmatrix egyszerűen az egylépéses átmenetmatrix r -edik hatványával egyenlő.

Ennek az állításnak a bizonyítása egyszerű. Számítsuk ki annak valószínűségét, hogy a rendszer az i állapotból két lépésben kerül a j állapotba. Az $i \rightarrow j$ átmenet két lépésben úgy következhet be, hogy a rendszer az i állapotból egy lépésben valamely v állapotba megy át, majd a v állapotból ismét egy lépésben a j állapotba. A v állapot a $0, 1, \dots, k$ állapotok bármelyike lehet, de közülük csakis egy, tehát

$$(6.1.2) \quad P_{ij}^{(2)} = \sum_{v=0}^k P_{iv} P_{vj}$$

A (6.1.2) összefüggés jobb oldala a $\underline{P}$ matrix i -edik sorának ugyanezen matrix j -edik oszlopával való komponáltja a sor-oszlop szorzási szabály alapján, vagyis $P_{ij}^{(2)}$ a $\underline{P}^{(2)}$ matrix i -edik sorának j -edik eleme. Állítunk tehát $r = 2$ esetére igaz.

Tételezzük most fel, hogy

$$\underline{P}_{r-1} = \underline{P}^{r-1}$$

Mivel az $i \rightarrow j$ átmenet r lépésben $k+1$ egymást kizáró módon történhet, mégpedig úgy, hogy a rendszer az i állapotból $r-1$ lépésben va-

lamely ν állapotba megy át, ahol ν a $0, 1, \dots, k$ állapotok valamelyike, azután a ν állapotból egy lépésben a j állapotba lép, ezért:

$$(6.1.3) \quad p_{ij}^{(r)} = \sum_{\nu=0}^k p_{i\nu}^{(r-1)} p_{\nu j}$$

ahol a (6.1.3) összefüggés jobb oldala a $\underline{P}^{r-1}$ matrix i -edik sorának és a $\underline{P}$ matrix j -edik oszlopának komponáltja, vagyis

$$(6.1.4) \quad \underline{P}_r = \underline{P}^{r-1} \cdot \underline{P} = \underline{P}^r.$$

Ebből az is következik, hogy

$$\underline{P}_r = \underline{P}_\ell \cdot \underline{P}_{r-\ell},$$

hiszen nyilvánvaló, hogy

$$\underline{P}^r = \underline{P}^\ell \cdot \underline{P}^{r-\ell},$$

ami részletesen kiírva azt jelenti, hogy a $\underline{P}_r$ matrix (i, j) eleme az r lépésben történő $i \rightarrow j$ átmenet valószínűsége felírható

$$(6.1.5) \quad p_{ij}^{(r)} = \sum_{\nu=0}^k p_{i\nu}^{(\ell)} p_{\nu j}^{(r-\ell)}$$

alakban. Ez utóbbi összefüggést Markov-féle egyenlőségnek nevezzük. Az átmenetmatrixokra vonatkozólag megjegyezzük, hogy un. sztochasztikus matrixok, ami azt jelenti, hogy elemeik nem negatív számok, és bármely sor elemeinek összege 1, hiszen pl. az (1) egy lépéses átmenetmatrix i -edik sora elemeinek összege éppen annak valószínűségét fejezi ki, hogy a rendszer az i -edik állapotból indulva vagy a 0 vagy az 1 ... vagy az i ... vagy a k állapotba megy át egy lépésben, ez pedig biztos esemény. Egyszerű számolással az is belátható, hogy sztochasztikus matrix.

Láttuk tehát a (6.1.4) összefüggésből, hogy az r -lépéses átmenetvalószínűségek meghatározása egyszerű matrixhatványozási probléma (amit általában elektronikus számítógéppel célszerű végezni, különösen nagy matrix esetén). Abból a tényből, hogy az (1) matrix sorai elemei nem negatívak és összegük 1, következik, hogy amikor a $\underline{P}$ matrixot önmagával szo-

rozzuk, akkor a $\underline{P}^2$ matrix j-edik oszlopának elemei a $\underline{P}$ matrix j-edik oszlopa elemeinek súlyozott számtani középértékei az egyes sorok elemei, mint súlyok segítségével. ($j = 0, 1, 2, \dots, k$ -ra.)

$$P_{0j}^{(2)} = P_{00} P_{0j} + P_{01} P_{1j} + P_{02} P_{2j} + \dots + P_{0k} P_{kj}$$

$$P_{1j}^{(2)} = P_{10} P_{0j} + P_{11} P_{1j} + P_{12} P_{2j} + \dots + P_{1k} P_{kj}$$

⋮

⋮

⋮

$$P_{ij}^{(2)} = P_{i0} P_{0j} + P_{i1} P_{1j} + P_{i2} P_{2j} + \dots + P_{ik} P_{kj}$$

⋮

⋮

⋮

$$P_{kj}^{(2)} = P_{k0} P_{0j} + P_{k1} P_{1j} + P_{k2} P_{2j} + \dots + P_{kk} P_{kj}$$

($j = 0, 1, 2, \dots, k$)

Mivel a $k+1$ szám súlyozott számtani közepe a legnagyobb és legkisebb érték közé esik (feltéve, hogy 1-nél kisebb súlyok is szerepelnek) általában a

$\underline{P}^{(2)}$ matrix j-edik oszlopának legnagyobb és legkisebb eleme közötti különbség ($j = 0, 1, 2, \dots, k$ -ra). Méginkább igaz ez a $\underline{P}$ matrix magasabb hatványainak oszlopaira az ismételt súlyozott közepelések következtében. A matrix hatványozása során tehát az oszlopok elemei között bizonyos kiegyenlítődés történik. Bizonyos esetekben elég nagy hatványokra a j-edik oszlop legnagyobb és legkisebb eleme közötti különbség eltűnik, az oszlop belül mindegyik elem egyformává válik, azaz a $P_{ij}^{(n)}$ átmenetvalószínűség minden i -re azonos lesz, azaz az n lépéses átmenet-valószínűség i -től, az induló állapottól nem függ. Az ilyen tulajdonságu Markov-lánccokat ergodikusként nevezzük.

Matematikailag fogalmazva: egy Markov-láncot ergodikusként nevezünk, ha léteznek a

$$\lim_{n \rightarrow \infty} P_{ij}^{(n)} = P_j$$

határértékek, amelyek i -től függetlenek és

$$\sum_{j=0}^k P_j = 1.$$

Nehezebb kérdés eldönteni, hogy adott átmenet-matrix esetén léteznek-e határvalószínűségek. Erre vonatkozik az alábbi tétel.

Markov-tétele: Egy véges sok állapotú homogén Markov-lánc akkor, és csak akkor ergodikus, ha az átmenetvalószínűségek

$$\underline{P} = \begin{pmatrix} P_{00} & P_{01} & P_{0k} \\ P_{10} & P_{11} & P_{1k} \\ \vdots & \vdots & \vdots \\ P_{k0} & P_{k1} & P_{kk} \end{pmatrix}$$

matrixának van olyan P^s hatványa, amelynek legalább egy oszlopában minden elem pozitív. A P_j határértékekhez való konvergencia sebességét illetően utmutatást ad az alábbi összefüggés:

$$P_{ij}^{(n)} - P_j \leq (1 - N_1 \delta)^{\frac{n}{s} - 1}$$

ahol N_1 a $\underline{P}^s$ matrix azon oszlopainak száma, amelyek csupa pozitív elemekből állnak, δ pedig az ezen oszlopokban álló elemek közül a legkisebb.

A tétel bizonyítását illetően az irodalomra utalunk, lásd pl.: [15]
A határmatrix

$$\lim_{n \rightarrow \infty} \underline{P}^n = \underline{P}^* = \begin{pmatrix} P_0 & P_1 & P_k \\ P_0 & P_1 & P_k \\ \vdots & \vdots & \vdots \\ P_0 & P_1 & P_k \end{pmatrix}$$

minden sora egyforma. A határvalószínűségek által alkotott valószínűség-eloszlást stacionárius eloszlásnak is nevezik. A $P_0, P_1, \dots, P_k$ határvalószínűségek azt fejezik ki, hogy mekkora valószínűséggel találjuk a rendszer hosszú állapotváltozás sorozat után az egyes $0, 1, 2, \dots, k$ állapot-

tokban. A $\{P_0, P_1, \dots, P_k\}$ határeloszlás kielégíti a következő egyenlet-rendszert:

$$(6.1.6) \quad P_j = \sum_{\nu=0}^k P_{\nu} P_{\nu j} \quad (j = 0, 1, \dots, k) \text{ és } \sum_{j=0}^k P_j = 1.$$

Ugyanis a Markov egyenlőség szerint (6.1.3) formula

$$P_{ij}^{(n+1)} = \sum_{\nu=0}^k P_{i\nu}^{(n)} P_{\nu j}$$

Ha most $n \rightarrow \infty$, akkor $P_{ij}^{(n+1)} \rightarrow P_j$, $P_{i\nu}^{(n)} = P_{\nu}$, vagyis a (6.1.6) összefüggést nyerjük. A $\sum_{j=0}^k P_j = 1$ reláció következik abból a tényből, hogy

sztochasztikus matrix minden hatványa ismét sztochasztikus matrix.

A (6.1.6) összefüggést a következő alakban is felírhatjuk:

$$(P_0, P_1, \dots, P_k) \begin{pmatrix} P_{00} & P_{01} & P_{0k} \\ P_{10} & P_{11} & P_{1k} \\ \vdots & \vdots & \vdots \\ P_{k0} & P_{k1} & P_{kk} \end{pmatrix} = (P_0, P_1, \dots, P_k)$$

Ha most ezen egyenlet mindkét oldalát jobbról ismét megszorozzuk a $\underline{P}$ matrix-szal, akkor

$$(P_0, P_1, \dots, P_k) \begin{pmatrix} P_{00} & P_{01} & P_{0k} \\ P_{10} & P_{11} & P_{1k} \\ \vdots & \vdots & \vdots \\ P_{k0} & P_{k1} & P_{kk} \end{pmatrix} \begin{pmatrix} P_{00} & P_{01} & P_{0k} \\ P_{10} & P_{11} & P_{1k} \\ \vdots & \vdots & \vdots \\ P_{k0} & P_{k1} & P_{kk} \end{pmatrix} = (P_0, P_1, \dots, P_k) \begin{pmatrix} P_{00} & P_{01} & P_{0k} \\ P_{10} & P_{11} & P_{1k} \\ \vdots & \vdots & \vdots \\ P_{k0} & P_{k1} & P_{kk} \end{pmatrix}$$

$$\begin{bmatrix} P^2 \end{bmatrix} = \begin{bmatrix} P \end{bmatrix}$$

Ezt az eljárást akárhányszor megismételhetjük, s azt az eredményt kapjuk, hogy tetszőleges n -re:

$$(P_0, P_1, \dots, P_k) \begin{bmatrix} P^n \\ \vdots \end{bmatrix} = (P_0, P_1, \dots, P_k) .$$

A Markov-láncokkal kapcsolatban eddig csak átmenetvalószínűségekről beszéltünk. Kérdezhetjük természetesen, hogy mi a valószínűsége, hogy rendszerünket a $t = n$ időpontban a j állapotban találjuk, azaz kereshetjük a $\{\xi_n = j\}$ esemény valószínűségét. ($j = 0, 1, \dots, k$.) Ezt a valószínűséget abszolút valószínűségnek nevezzük.

A $P(\xi_0 = j)$, ($j = 0, 1, \dots, k$) valószínűségeloszlásnak nevezzük.

Jelöljük a kezdeti valószínűségeloszlást

$$(P_0^{(0)}, P_1^{(0)}, \dots, P_k^{(0)}) \text{-al}$$

Annak valószínűsége, hogy a rendszer egy lépés után a j állapotban lesz a teljes valószínűség tétele alapján

$$P(\xi_1 = j) = \sum_{i=0}^k P(\xi_1 = j | \xi_0 = i) P(\xi_0 = i) = \sum_{i=0}^k P_i^{(0)} P_{ij},$$

amelyet matrixos jelöléssel a következő alakban írhatunk:

$$(P_0^{(0)}, P_1^{(0)}, \dots, P_k^{(0)}) \begin{pmatrix} P_{00} & \dots & P_{0k} \\ P_{10} & \dots & P_{1k} \\ \vdots & & \vdots \\ P_{k0} & & P_{kk} \end{pmatrix} = (P_0^{(1)}, P_1^{(1)}, \dots, P_k^{(1)})$$

Annak valószínűsége, hogy a rendszer a $t = n$ időpontban a j állapotban lesz, ugyancsak a teljes valószínűség tétele szerint:

$$P(\xi_n = j) = \sum_{i=0}^k P(\xi_n = j | \xi_0 = i) P(\xi_0 = i) = \sum_{i=0}^k P_i^{(0)} P_{kj}^{(n)},$$

vagy matrixos jelöléssel:

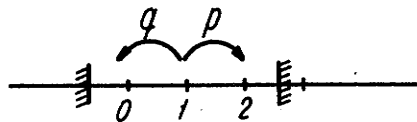
$$(P_0^{(0)}, P_1^{(0)}, \dots, P_k^{(0)}) \begin{bmatrix} P^n \end{bmatrix} = (P_0^{(n)}, P_1^{(n)}, \dots, P_k^{(n)})$$

A $t = n$ időpontbeli abszolút valószínűségeket tehát egyszerűen úgy számítjuk ki, hogy a kezdeti eloszlást (mint vektort) megszorozzuk az n -lépéses átmenet matrix-szal. Ha van olyan $[Q_0, Q_1, \dots, Q_k]$ kezdeti eloszlás, amelyre teljesül a

$$(6.1.7) \quad (Q_0, Q_1, \dots, Q_k) \begin{bmatrix} P^n \end{bmatrix} = (Q_0, Q_1, \dots, Q_k)$$

reláció minden n -re, akkor az ilyen eloszlást stacionárius eloszlásnak nevezzük. Láttuk, hogy ha a Markov-lánc ergodikus, akkor a nyert határel-oszlás stacionárius eloszlás.

Megjegyezzük, hogy ha a $\underline{P}$ egy-lépéses átmenet matrix kétszeresen sztochasztikus, azaz nemcsak a sorok, hanem az oszlopok elemeinek összege is 1 és a lánc ergodikus, akkor a



52. ábra

$\lim_{n \rightarrow \infty} \underline{P}^n$ matrix is kétszeresen sztochasztikus, így a határmatrix minden eleme

$\frac{1}{k+1}$, vagyis hosszú átmenetsorozat után a fizikai rendszert bármely állapotában egyforma valószínűséggel találjuk.

Mutatunk egy egyszerű példát a stacionárius eloszlás meghatározására a (6.1.7) összefüggés segítségével.

Tegyük fel, hogy egy részecske bolyong visszaverő falak között a számegyenes 52. ábra szerinti szakaszán. A részecske időegységenként 1-et lép jobbra vagy balra, a jobbra lépés valószínűsége legyen p , a balra lépése $q = 1 - p$. Ha a részecske az $x = 0$ ponttól balra lép, akkor a falon visszaverődik, ezért helyben marad. Ugyanez a helyzet, ha a részecske az $x = 2$ pontból jobbra lép. Az egylépéses átmenet-matrix most a következő:

$$\underline{P} = \begin{pmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{pmatrix} = \begin{pmatrix} q & p & 0 \\ q & 0 & p \\ 0 & q & p \end{pmatrix}$$

Keressük a összefüggést kielégítő (P_0, P_1, P_2) stacionárius határeloszlást. Teljesülnie kell a

$$(P_0, P_1, P_2) \begin{pmatrix} q & p & 0 \\ q & 0 & p \\ 0 & q & p \end{pmatrix} = (P_0, P_1, P_2) \text{ összefüggésnek, azaz}$$

$$P_0 \cdot q + P_1 \cdot q + P_2 \cdot 0 = P_0$$

$$P_0 \cdot p + P_1 \cdot 0 + P_2 \cdot q = P_1$$

$$P_0 \cdot 0 + P_1 \cdot p + P_2 \cdot p = P_2$$

Az első egyenletből $P_1 \cdot q = P_0 (1 - q) = P_0 \cdot p$, vagyis

$$\frac{P_1}{P_0} = \frac{p}{q}$$

A második egyenletből

$$P_2 \cdot q = P_1 - P_0 \cdot p = P_1 - P_1 \cdot q = P_1 (1 - q) = P_1 \cdot p = \frac{p^2}{q} \cdot P_0$$

$$\text{azaz } P_2 = \left(\frac{p}{q}\right)^2 \cdot P_0$$

Mivel $P_0 + P_1 + P_2 = 1$ kell legyen,

$$P_0 + \frac{p}{q} P_0 + \left(\frac{p}{q}\right)^2 P_0 = 1$$

$$P_0 = \frac{1}{1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2}$$

$$P_1 = \frac{\frac{p}{q}}{1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2}$$

$$P_2 = \frac{\left(\frac{p}{q}\right)^2}{1 + \frac{p}{q} + \left(\frac{p}{q}\right)^2}$$

Könnyű belátni, hogy ha $p = q = \frac{1}{2}$, akkor $P_0 = P_1 = P_2 = \frac{1}{3}$. Ez esetben a (I) alatti $\underline{P}$ matrix duplán sztochasztikus, így az erre az esetre vonatkozó megjegyzésünket is igazoltuk példánkban.

6.2 A Markov-láncok alkalmazása víz tározókkal kapcsolatos problémákra

Tegyük fel, hogy van egy véges $k \text{ m}^3$ kapacitású tározó. A t időpontban a tározóban levő vízmennyiséget jelöljük ξ_t -vel. A valószínűségi változó értéket diszkrét ($t = 0, 1, 2, \dots$) időpontokban figyeljük meg. A $(t, t+1)$ intervallumban befolyó víz mennyiségét nevezzük hozzáfolyásnak és jelöljük ζ_t -vel. Feltesszük, hogy a $\zeta_0, \zeta_1, \dots, \zeta_k, \dots$ valószínűségi változók függetlenek és azonos eloszlásúak. Ha $\xi_t = \xi_t > k$, akkor a $(t, t+1)$ intervallumban túlfolyás történik, azaz

$$\xi_t + \zeta_t - k$$

mennyiségű víz túlfolyik.

Feltesszük, hogy minden egységnyi időintervallum végén azonos $M \text{ m}^3$ vizet leeresztünk a tározóból. (M ésszerűen választott állandó, amelyet rendszerint a szükséglet határoz meg és természetesen $M < k$. Általában M -et egységnyiinek választjuk.)

A $\xi_0, \xi_1, \dots$ valószínűségi változók sorozata homogén Markov-láncot képez, amelynek stacionárius határeloszlása létezik. Attól függően, hogy a bevitel, a ξ_i valószínűségi változók milyen eloszlásúak, a stacionárius határeloszlás más és más lesz.

Legyen

$$P(\xi_t = j) = q_j \quad (j = 0, 1, 2, \dots)$$

a hozzáfolyás valószínűség-eloszlása.

Feltesszük, hogy $q_j > 0$ minden j -re. Képezzük most a hozzáfolyás eloszlás generátorfüggvényét a

$$G(z) = \sum_{j=0}^{\infty} q_j z^j \quad (|z| < 1)$$

függvényt. A generátorfüggvény segítségével könnyen kiszámíthatjuk az időegységenként átlagosan befolyó víz mennyiségét, az átlagos hozzáfolyást. A (4.1.2) formula alapján:

$$M(\xi_t) = m = G'(1) = \sum_{j=0}^{\infty} j q_j$$

A $\{\xi_t\}$ Markov-lánc egy lépéses átmenetvalószínűségeinek matrixa most a következő:

$$\underline{P} = \begin{pmatrix} P_{00} & P_{01} & \dots & P_{0,k-1} \\ P_{10} & P_{11} & \dots & P_{1,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k-1,0} & P_{k-1,1} & \dots & P_{k-1,k-1} \end{pmatrix} = \begin{pmatrix} q_0 + q_1 & q_2 & q_3 & \dots & q_{k-1} h_k \\ q_0 & q_1 & q_2 & & q_{k-2} h_{k-1} \\ 0 & q_0 & q_1 & & q_{k-3} h_{k-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & q_0 h_1 \end{pmatrix}$$

(A jelölés megvilágítása céljából még egyszer megemlítjük, hogy a tározó állapotát mindig a $(t, t+1)$ intervallum végén regisztráljuk.) $t = 0, 1, 2, \dots$ a t időpontbeli ξ állapothoz hozzájön ξ_t a $(t, t+1)$ egységnyi intervallumban befolyt víz mennyisége, de ugyanezen idő alatt $M = 1$ egységnyi vizet is kivettünk (ha van benne). A tározó tehát a $0, 1, \dots, k-1$ állapotokban lehet. A tározó a $\xi_t = 0$ állapotban marad, ha $\xi_t = 0$, ekkor a $(t, t+1)$ intervallumban nem jött víz és nincs mit kivenni, ennek valószínűsége q_0 , vagy ha $\xi_t = 1$, ekkor a $(t, t+1)$ intervallumban egységnyi mennyiségű bevétel

van, aminek q_1 a valószínűsége, viszont ugyanennyi vizet ki is veszünk a $(t, t+1)$ intervallumban. Általában a tározó a fenti modellnél az i állapotból a j állapotba lép egy lépésben (ezen átmenet valószínűségét az i -edik sor j -edik eleme szolgáltatja, ha $\zeta_t = i$ és $\xi_t = j+1$, ennek valószínűsége q_{j+1} . Itt is figyelembe kell venni, hogy a $(t, t+1)$ intervallumban egységnyi mennyiségű vizet kiveszünk).

A fenti matrix utolsó oszlopában

$$h_i = \sum_{j=1}^{\infty} q_j \quad (i = 1, 2, \dots, k)$$

Mivel feltételeztük, hogy $q_j > 0$ ($j = 0, 1, 2, \dots$) a fenti egy lépéses átmenet-matrixnak van olyan oszlopa (a két utolsó oszlop), amely csupa pozitív elemből áll, tehát a Markov-lánc ergodikus, létezik a stacionárius határeloszlás. Ennek a stacionárius határeloszlásnak az explicit meghatározása bizonyos bevitel-eloszlások esetében a szukcesszív matrix hatványozás tényleges végrehajtása nélkül is elérhető. Ehhez nyújt segítséget az alábbi tétel:

Moran tétele: Ha egy k kapacitású tározóban tárolt vízmennyiség stacionárius eloszlása a

$$\begin{pmatrix} 0 & 1 & 2 & \dots & k-1 \\ P_0 & P_1 & P_2 & \dots & P_{k-1} \end{pmatrix}$$

eloszlás, akkor a

$$\psi_i = \frac{P_i}{P_0} \quad (i = 1, 2, \dots, k-1)$$

hányadosok k -tól függetlenek és a ψ_i mennyisége a

$$V(z) = \frac{q_0 (1-z)}{G(z) - z}$$

generátorfüggvényben a z^i hatvány együtthatója.

A fenti összefüggésben q_0 a bevitel-eloszlásban a $\{\xi_t = 0\}$ esemény valószínűsége, $G(z)$ pedig a hozzáfolyás eloszlásának generátorfüggvénye.

Eltekintünk itt ezen hasznos tétel bizonyításától, lásd: [10], csak változni kívánjuk, hogyan alkalmazható a tétel bizonyos speciális bevitel-eloszlások esetén.

Megjegyezzük, hogy ha a φ_i számokat sikerül meghatározni, akkor a

$\{P_0, P_1, \dots, P_{k-1}\}$ stacionárius eloszlást is meghatároztuk,

$$P_0 + P_1 + \dots + P_{k-1} = 1$$

lévén

$$1 + \frac{P_1}{P_0} + \dots + \frac{P_{k-1}}{P_0} = \frac{1}{P_0},$$

azaz

$$\sum_{i=1}^{k-1} \varphi_i = \frac{1}{P_0} - 1$$

$$P_0 = \frac{1}{1 + \sum_{i=1}^{k-1} \varphi_i}$$

$$P_i = P_0 \varphi_i, \quad i = 1, 2, \dots, k-1$$

A stacionárius eloszlásra vonatkozólag diszkrét véges modell esetében explicit eredmények ismeretesek geometriai-, negatív binomiális- és Poisson eloszlású hozzáfolyás esetére.

Meghaladja a jegyzetünk kereteit e problémakör részletes tárgyalása, csupán a geometriai eloszlású hozzáfolyás esetére mutatjuk meg, hogyan alkalmazható Moran tétele, majd néhány megjegyzést és irodalmi utalást teszünk.

Tételezzük fel, hogy a tározóba a vízbevitel geometriai eloszlású, azaz legyen $P(\xi_t = j+1) = pq^j$, ($j = 0, 1, 2, \dots$), ahol $0 < p < 1$ és $q = 1-p$.

A geometriai eloszlás generátorfüggvénye

$$G(z) = \sum_{j=0}^{\infty} p q^j z^j = p \sum_{j=0}^{\infty} (gz)^j = \frac{p}{1-gz}$$

A geometriai eloszlás várható értéke:

$$\mu = G'(1) = \left(\frac{p}{1-qz} \right)' z$$

A $V(z)$ generátorfüggvény:

$$V(z) = \frac{\frac{p(1-z)}{1-qz}}{\frac{p}{1-qz}} = \frac{1-qz}{1-\mu z} = (1-qz) \sum_{i=0}^{\infty} (\mu z)^i$$

Innen

$$\psi_0 = 1, \psi_i = \mu^i - q\mu^{i-1} = \mu^i \left(1 - \frac{q}{\mu}\right) = q\mu^i,$$

$$\sum_{i=0}^{k-1} \psi_i = 1 + q \sum_{i=1}^{k-1} \mu^i = p \frac{1-\mu^{k+1}}{1-\mu}$$

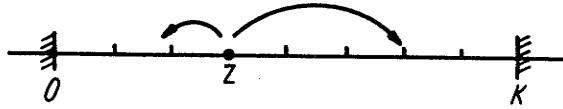
A keresett stacionárius eloszlás tehát:

$$P_0 = \frac{1-\mu}{p(1-\mu^{k+1})}, P_i = \frac{\mu^{i+1}(1-\mu)}{1-\mu^{k+1}}, (i = 1, 2, \dots, k-1)$$

Gyakorlati szempontból igen fontos az a kérdés, hogy adott eloszlásu hozzáfolyás és leeresztés esetén mi a valószínűsége, hogy a tározó kiürül, mielőtt tulfolya. Ez a probléma lényegében visszavezethető egy bolyongási problémára a következő módon: A tározó lehetséges állapotai a $0, 1, 2, \dots$, állapotok.

Tegyük fel, hogy az induló állapot z . Ha most egy időegység alatt ℓ m^3 víz jön bele, a tározó a $z + \ell$ állapotba lép. Ha megállapodunk abban, hogy időegységenként μm^3 vizet kiveszünk, akkor ℓm^3 hozzáfolyás esetén a tározó a z állapotból egy lépésben a $z + \ell - \mu$ állapotba lép. Attól függően, hogy $\ell - \mu > 0$ vagy $\ell - \mu < 0$, a tározó z -nél nagyobb vagy kisebb állapotba lépett.

Tegyük fel most, hogy egy részecske bolyongást végez a számegyes egész koordinátájú pontjain a $(0, k)$ intervallumban, az $x = 0$ és $x = k$ pontokban elnyelő falakat helyezünk el, a részecske az $x = z$ pontból indult és időegységenként jobbra vagy balra lép akkora távolságot, amennyi a víz bevitel és kivétel különbsége egy időegység alatt (egész m^3 -ben számolva), azaz $\ell - \mu$ nagyságú lépést tesz. ($\ell - \mu$ lehetséges értékei: $-\mu, -\mu+1, \dots, -1, 0, 1, 2, \dots$)



53. ábra

Ilyen típusú bolyongási probléma megoldásával többen foglalkoztak, más-más módszerrel. Különösen említésre méltó Feller [3]; Wald [25], továbbá Girsick módszere [25].

Ha a folyó vízhozama, a hozzáfolyás olyan eloszlású, hogy elenyésző kicsi annak valószínűsége, hogy bizonyos véges ν mennyiség esetén $\mu + \nu$ m^3 -nél több víz jön egy időegység alatt, azaz a részecske egylépéses ugrása a $-\mu, -\mu+1, \dots, -1, 0, 1, 2, \dots, \nu$ értékeket veheti fel és az ugráshossz várható értéke 0, azaz a hozam és a leeresztés kiegyenlíti egymást, akkor annak P_z valószínűségére, hogy a z kezdeti állapotú tározó előbb kiürül, mint megtelne, a következő Fellertől eredő becslés ad-

ható: $\frac{k - z}{k + \nu - 1} \leq P_z \leq \frac{k + \mu - z - 1}{k + \mu - 1}$, ahol k a tározó kapacitása.

Ha k nagy $\mu + \nu$ -höz képest, akkor ez igen jó becslés a fenti valószínűségre.

A MATEMATIKAI STATISZTIKA ELEMEI

7.1 A matematikai statisztika tárgya
és alkalmazási területei

A matematikai statisztika a valószínűségszámításnak igen fontos, a gyakorlat számára nélkülözhetetlen fejezete, amelynek tárgyát ugyancsak a véletlen tömegjelenségek vizsgálata képezi, azonban problémái és ennek következtében legtöbbször a módszerei is sajátosak.

A valószínűségszámítás előző fejezetekben bizonyos események valószínűségeit ismerteknek tekintettük, s azt vizsgáltuk, hogy ezen események valószínűségeinek ismeretében hogyan lehet más, összetettebb események valószínűségét kiszámítani. Ha pl. ismerjük, hogy egy A esemény valószínűsége $P(A) = p_1$, egy másik B esemény valószínűsége $P(B) = p_2$ és az A és B események függetlenek, akkor az $A + B$ esemény valószínűsége:

$$P(A + B) = p_1 + p_2 - p_1 \cdot p_2$$

A gyakorlatban azonban sokszor előfordul, hogy sem az A , sem a B esemény valószínűségét nem ismerjük, ezért a fenti formulát nem tudjuk alkalmazni. Ilyenkor első feladatunk az A , illetve a B események valószínűségének kielégítő pontosságu meghatározása. Ezzel a problémával a matematikai statisztika foglalkozik.

A korábbi fejezetekben úgy foglaltunk, hogy legyen ξ normális eloszlású valószínűségi változó m várható értékkel és σ szórással, azaz legyen ξ sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

s ennek alapján ki tudjuk számítani, mennyi a valószínűsége, hogy ξ értéke adott határok közé essék. A gyakorlatban sokszor ismeretlenek számunkra az m és σ paraméterek, ezek meghatározása ekkor elsődleges feladatunk. Ha pl. ξ valamely gyártmány adott méretét jelenti, amely méretet a szabvány szerint m_0 kell, hogy legyen, s az illető gyártmány csak akkor hozható forgalomba, ha a kérdéses mérete az $m_0 - 2\sigma_0$ és $m_0 + 2\sigma_0$

határok közé esik, akkor el kell döntenünk, hogy a kérdéses méretet jelölő ξ valószínűségi változó várható értéke valóban m_0 -e, s a σ szórás értéke valóban σ_0 -al egyezik-e meg. Ha ugyanazt a gyártmányt több gépen állítjuk elő, s az egyes gépeken gyártott alkatrészek kérdéses méretét a $\xi_1, \xi_2, \dots, \xi_m$ valószínűségi változók jelölik, akkor ellenőriznünk kell, hogy mindegyik gép azonos méretet produkál-e, azaz fennáll-e, hogy $M(\xi_1) = M(\xi_2) = \dots = M(\xi_m)$. Ilyen jellegű kérdések vizsgálata a matematikai statisztika körébe tartozik. Már ezekből a példákban is látható a matematikai statisztika nagy gyakorlati jelentősége: a valószínűségszámítás eredményes gyakorlati felhasználásának előfeltétele a valószínűségek és eloszlások kielégítő meghatározottsága, ami csak a matematikai statisztika módszereinek ismeretében érhető el.

Abban az esetben, ha egy vizsgált ξ valószínűségi változóra vonatkozólag nagyszámu (néhány száz) megfigyelési adat áll rendelkezésünkre, akkor ξ eloszlása és a vele kapcsolatban bennünket érdeklő események valószínűsége a gyakorlat számára kellő pontossággal meghatározható. A gyakorlatban azonban a különböző fizikai, kémiai, technológiai kísérletek elvégzése tetemes mennyiségű időt és anyagi ráfordítást igényel, az ipari minőségellenőrzésben pedig vannak vizsgálatok, amelyek a gyártmány tönkretétele vagy rongálása útján végezhetők csak el, ezért gyakran viszonylag kevés megfigyelési adatból kell megbízható következtetéseket levonnunk. A matematikai statisztika feladata olyan módszerek kidolgozása, amelyek segítségével a megfigyelt tapasztalati adatokból a kérdéses véletlen mennyiségre vonatkozólag a lehető legtöbb információt nyerhetjük. Feladata a matematikai statisztikának maguknak a kísérleteknek a megtervezése és a kísérletek optimális számának meghatározása, továbbá a selejtvizsgálat viszonylag kis mintán alapuló gazdaságos módszereinek kidolgozása.

7.2 A matematikai statisztika fő fejezetei

A matematikai statisztikai módszerek természetének megvilágítása céljából vizsgáljuk meg, hogy valószínűségszámítási ismeretek birtokában milyen előrejelzés érhető el valamely véletlen jelenség jövőbeli lefolyására vonatkozólag. Ha tudjuk, hogy pl. egy A esemény valószínűsége $P(A) = p$ ($0 < p < 1$), akkor arra vonatkozólag, hogy egyetlen kísérlet során az A esemény bekövetkezik-e vagy sem, nem sokat tudunk mondani, hiszen lehet, hogy A bekövetkezik, de az is lehet, hogy nem következik be. Ha azonban a kísérletet - amellyel kapcsolatban az A esemény érdekel bennünket - sokszor megismételjük, akkor p értéke már sokat mond szá-

munkra, mert előre tudjuk, hogy kb. az esetek p százalékában fog bekövetkezni az A esemény. Más szóval n számú kísérlet során az A esemény relatív gyakorisága

$$\frac{k}{n} \approx p.$$

A $\frac{k}{n}$ relatív gyakoriság az ismeretlen p valószínűség közelítése, becslése.

Az A esemény p valószínűségéből - ha p értékét ismerjük - következtetni tudunk az A esemény relatív és abszolút gyakoriságára, ha n elég nagy. A gyakorlatban azonban $P(A) = p$ értékét általában nem ismerjük,

módunk van azonban a $\frac{k}{n}$ relatív gyakoriság meghatározására. Adott A esemény p valószínűsége egy állandó szám, míg a $\frac{k}{n}$ relatív gyakoriság rögzített n mellett is valószínűségi változó, ami azt jelenti, hogy ha több n hosszúságú kísérletsorozatot végzünk, a relatív gyakoriság viszonylagos stabilitása ellenére is különböző értékeket vehet fel. A nagy számok erős

törvénye értelmében a $\frac{k}{n}$ relatív gyakoriság 1 valószínűséggel p -hez tart, ha $n \rightarrow \infty$, a gyakorlatban azonban csak véges számú kísérletet tudunk végezni, nyilvánvaló tehát, hogy p értékét csak közelíteni, "becsülni"

tudjuk rögzített véges n esetében a $\frac{k}{n}$ relatív gyakoriság segítségével.

Amikor az ismeretlen p valószínűséget, amely a $[0,1]$ intervallum egy pontja a $\frac{k}{n}$ valószínűségi változóval becsüljük, pontbecsléssel állunk szemben. Felmerül természetesen az a kérdés, hogy rögzített n esetén mennyire tér el a $\frac{k}{n}$ relatív gyakoriság az ismeretlen p valószínűségtől. E

kérdés megválaszolása során valószínűségszámítási ismereteinkre támaszkodunk. Tudjuk, hogy a $\frac{k}{n}$ relatív gyakoriság mint valószínűségi változó binomiális eloszlást követ $M(\frac{k}{n}) = p$ várható értékkel és $D(\frac{k}{n}) = \sqrt{\frac{pq}{n}}$

szórással. Ha n elég nagy, akkor a binomiális eloszlás normális eloszlással közelíthető (Moire - Laplace határeloszlás tétel).

Tudjuk, hogy egy normális eloszlású ξ valószínűségi változóra $M(\xi) = m$, $D(\xi) = \sigma$ esetén érvényes a

$$P(|\xi - m| < 2\sigma) \approx 0,95 \quad \text{összefüggés.}$$

Ennek alapján:

$$P\left(\left|\frac{k}{n} - p\right| < 2\sqrt{\frac{pq}{n}}\right) \approx 0,95$$

Mivel $pq \leq \frac{1}{4}$ érvényes a

$$P\left(\left|\frac{k}{n} - p\right| < \frac{1}{\sqrt{n}}\right) \approx 0,95$$

Minden esetre a $\left(\frac{k}{n} - \frac{1}{\sqrt{n}}, \frac{k}{n} + \frac{1}{\sqrt{n}}\right)$ intervallum 95%-os valószínűséggel tartalmazza az ismeretlen p valószínűséget. Ez a hibahatár akkor tekinthető jónak, ha az ismeretlen p valószínűség (és vele együtt $\frac{k}{n}$) $\frac{1}{2}$ közelében van. Ha p értéke lényegesen eltér $\frac{1}{2}$ -től, $pq < \frac{1}{4}$, ezért szűkebb intervallumot nyerünk, ha a zárójelen belüli egyenlőtlenség jobb oldalán p helyébe $\frac{k}{n}$ -et írunk, ekkor

$$P\left(\left|\frac{k}{n} - p\right| < \frac{2}{n} \sqrt{\frac{k(n-k)}{n}}\right) \approx 0,95$$

Azt kaptuk tehát, hogy a $\left(\frac{k}{n} - \frac{2}{n} \sqrt{\frac{k(n-k)}{n}}, \frac{k}{n} + \frac{2}{n} \sqrt{\frac{k(n-k)}{n}}\right)$ intervallum

95%-os, valószínűséggel tartalmazza az ismeretlen p valószínűséget. Ez azt jelenti, hogy ha sok kísérlet-sorozatot végzünk (amely sorozatok mindegyike n kísérlethől áll) mindegyik sorozatra kiszámítjuk a $\frac{k}{n}$ relatív gyakoriságot s ezek köré mint középpont köré megszerkesztjük a fenti intervallumot, akkor ezen intervallumoknak mindössze 5%-a lesz olyan, amely nem tartalmazza a belsejében a p pontot.

Az ily módon nyert intervallumot 95%-os konfidencia-intervallumnak nevezzük. A becslésnek ez a módja az ún. intervallum-becslés, amely J. Neyman lengyel származású matematikustól ered.

Statisztikai becslés szükségessége nemcsak az ismeretlen p valószínűséggel kapcsolatban merül fel. A gyakorlatban sokszor előfordul, hogy a kísérlet eredményét jelölő ξ valószínűségi változó eloszlás-típusát ismerjük. Tudjuk pl., hogy bizonyos időintervallumban érkező hívások száma egy telefonközpontba Poisson-eloszlást követ, csak éppen nem ismerjük ezen eloszlás λ paraméterét, tehát becsülnünk kell. Igen sokszor tudjuk, hogy bizonyos mérési eredmény normális eloszlású (sokszor elméleti megfontolások útján, a centrális - határelosztástétel érvényesülése folytán következtetünk erre), de nem ismerjük sem az m várható értékét, sem a σ szórás értékét, ezeket tehát tapasztalati úton kell becsülnünk. Valamely valószínűségeloszlással kapcsolatos állandók, paraméterek statisztikai meghatározásával a matematikai statisztika becslésmélet című fejezete foglalkozik.

Bonyolultabb a feladatunk, ha a vizsgált ξ valószínűségi változó $F(x)$ eloszlásfüggvénye teljesen ismeretlen számunkra, más szóval, ha az eloszlás típusa, függvényalakja sem ismeretes, vagy legalábbis nem vagyunk biztosak abban, hogy ξ eloszlásfüggvénye valóban az $F(x)$ eloszlásfüggvény, legfeljebb gyanítjuk elméleti meggondolás vagy előző tapasztalatok alapján. Ilyenkor ξ eloszlására nézve feltevessel élünk. Felteszszük, hogy ξ eloszlásfüggvénye adott $F(x)$ függvény, majd megvizsgáljuk, hogy a ξ változóra vonatkozó megfigyelési, mérési eredmények mennyire vannak összhangban hipotézisünkkel, majd döntünk, hogy elfogadhatjuk-e hipotézisünket, vagy el kell azt vetnünk. A feltevésvizsgálatnak ezt az esetét illeszkedésvizsgálatnak nevezzük, hiszen azt vizsgáljuk, hogy ξ megfigyelt értékei, illetve az azokból számított statisztikai mennyiségek mennyire illeszkednek az $F(x)$ elméleti eloszláshoz.

Gyakran nincs is szükség a konkrét eloszlás, illetve annak paramétereinek meghatározására, hanem csak azt a kérdést kell eldöntenünk, hogy két sokaság (pl. két különböző gépen gyártott alkatrészek bizonyos mérete) azonos valószínűségeloszlást követ-e, illetve megegyeznek-e az adott sokaságok, amelyeket egy-egy valószínűségi változó reprezentál bizonyos paraméterei. Ekkor úgy járunk el, hogy feltételezzük az eloszlások, illetve a kérdéses paraméterek megegyezését, s megvizsgáljuk, hogy tapasztalati adataink valószínűségi számítási szempontból összeegyeztethetők-e vagy sem a feltételezésünkkel, tehát ismét döntenünk kell hipotézisünk igaz vagy hamis voltáról. Minthogy megfigyelési adataink mindkét sokaságra vonatkozólag véletlen, ingadozó értékek, előfordulhat, hogy tévesen döntünk.

A matematikai statisztika feladata olyan módszereket kidolgozni, amelyek segítségével - nagy valószínűséggel - helyesen tudjuk eldönteni hipotézisünk elfogadható, vagy elfogadhatatlan voltát.

Azokat az eljárásokat, amelyek segítségével hipotéziseink helyességét ellenőrizzük, statisztikai próbáknak nevezzük.

A becslélmélet és a statisztikai hipotézisek vizsgálata című fejezetek alkotják a matematikai statisztika alapvető problémakörét. Wald Ábrahám magyar származású matematikus a döntésfüggvények elméletében egybefoglalta ezt a két fejezetet is.

E két legfontosabb problémakörön kívül röviden tárgyalni fogunk néhány egyéb kérdéskört is, mint pl. a mintavétel elmélete, a szórásanalízis, a regresszió és korreláció statisztikai problémái. Természetesen mindezen kérdésköröknek csupán legáltalánosabb, főként elvi problémáit tárgyaljuk, s bizonyos kérdésekkel kapcsolatban irodalmi utalással kell megelégednünk.

7.3 A mintavétel elmélete

A statisztikai minta

Mint már említettük, a matematikai statisztika feladata tapasztalati adatokból valamely valószínűségi változó eloszlására, annak ismeretlen paramétereire való következtetés. A tapasztalati adatokat valamely ξ véletlen mennyiségre, valószínűségi változóra vonatkozó megfigyelések vagy mérések útján nyerjük, s ezeket számadatok formájában jegyezzük fel. Az így nyert adatokat statisztikai mintának nevezzük. Statisztikai mintán vagy röviden mintán tehát egy valószínűségi változóra vonatkozó véges számú megfigyelés (mérés) eredményét értjük. Ha pl. a ξ valószínűségi változó egy üzemben adott napon előállított villanykörték élettartamát (égési idejét) jelöli, s a napi terméktömegből találomra kiemelünk n darab égőt és megvizsgáljuk az élettartamát, akkor valamely $x_1, x_2, \dots, x_n$ mintát nyerünk. Itt n -et a minta elemszámának, az $x_1, x_2, \dots$ adatokat a minta elemeinek nevezzük. Ujabb n darab égőt kiemelve és azok élettartamát megfigyelve, valamely másik $y_1, y_2, \dots, y_n$ mintát kapunk eredményül.

Jóllehet a minta elemei adott esetben konkrét számok, ezek értéke azonban minden megfigyeléssorozatban más és más lehet, ezért a mintaelemeket valószínűségi változóknak tekintjük, s a mintára a

$$\xi_1, \xi_2, \dots, \xi_n$$

jelölést alkalmazzuk.

Lényegében mintavételről van szó akkor is, amikor egy A esemény ismeretlen p valószínűségére akarunk következtetni. Ekkor n számú független kísérletet végezve 1 -est írhatunk valahányszor az A esemény bekövetkezik, amikor pedig A nem következik be (vagyis $\bar{A}$ következik be), akkor 0 -t írunk. Ez esetben mindegyik ξ_i mintaelem az A esemény indikátor-változója, s a

$$\xi_1, \xi_2, \dots, \xi_n$$

minta minden eleme az 1 vagy 0 értéket veheti fel.

A minta iránt támasztott legfontosabb gyakorlati követelmény, hogy hiven tükrözze azt a statisztikai sokaságot, amelyből a minta származik, más szóval, hogy a minta reprezentatív legyen.

A statisztikai sokaság fogalmát a következő példán érzékeltetjük: alkossa a mintát egy láda csavarból találomra kiemelt 50 db csavar egyenként lemért hosszúsága. Ebben az esetben a sokaság a ládában lévő ösz-

szes csavarok hosszmereteit értjük. Általában statisztikai sokaságon a megfigyelés tárgyát képező ξ valószínűségi változó összes számba jöhető értékeinek halmazát értjük.

Tekintettel arra, hogy a mintából az egész sokaság eloszlására (tehát a meg nem figyelt értékek viselkedésére is) következtethetünk, érthető, hogy miért olyan fontos a minta reprezentatív volta.

Avégből, hogy ez a követelmény teljesüljön, a mintavételt alkalmas módon kell végeznünk. Ha pl. egy bizonyos fizikai, kémiai vagy műszaki kísérlet eredményét kifejező ξ valószínűségi változóról van szó, akkor ügyelnünk kell arra, hogy a kísérletek (mérések) függetlenek legyenek (azaz egyik kísérlet eredménye se legyen befolyással a többi kísérlet kimenetelére). A mintavétel elméletében feltesszük, hogy olyan jellegű kísérletekről van szó, amelyek nagyjából azonos körülmények között sokszor - elvileg akárhányszor - megismételhetők. A mintaelemek sorrendje az észlelés sorrendjével egyezik meg, ami egy véletlen sorrendet jelent. Ha a minta valóban reprezentatív, akkor a mintaelemek egyforma eloszlású valószínűségi változók, s mindegyik mintaelem eloszlása megegyezik az egész sokaság eloszlásával, azaz

$$P(\xi_i < x) = P(\xi < x) \quad i = 1, 2, \dots, n\text{-re teljesül.}$$

Azt a tényt, hogy a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek összességükben független és azonos eloszlású valószínűségi változók, röviden úgy szoktuk kifejezni, hogy a mintaelemek véletlen sorozatot alkotnak. Hogy az így értelmezett véletlenség valóban teljesül-e adott minta esetében, azt statisztikai próba un. véletlenségvizsgálat segítségével ellenőrizhetjük. Erre vonatkozó módszert a 10.4 pontban ismertetjük.

Nem foglalkozunk részletesen a mintavétel technikájával, ez egyébként függhet a vizsgált probléma természetétől, nevezetesen, hogy technológiai, mezőgazdasági, közgazdasági, biológiai kísérletről van-e szó, ezért itt csak egy-két megjegyzést teszünk.

Ha a mintavétel előre megadott véges N elemű sokaságból történik (pl. tömegcikkek minőségellenőrzéskor) ez esetben a sokaság elemeit valamilyen módon 0-tól N -ig megszámozzuk és sorsolással (pl. véletlen számok táblázata segítségével) döntjük el, mely elemek kerülnek a mintába.

Igen gazdaságos módja a mintavételnek az ún. szekvenciális (sorozatos) módszer, különösen olyankor, amidőn az egyes mintadarabok vizsgálata nem igényel túl hosszú időt. Itt a mintaelemek számát nem rögzítjük előre, hanem az egyes mintaelemeket véletlenszerűen választva (független kísérleteket végezve) minden egyes lépés után annak eredményétől függően választunk két lehetőség között, nevezetesen, hogy végleges döntést hozunk-e (elfogadjuk vagy elvetjük a felállított hipotézisünket), vagy pedig további mintaelemet veszünk. Ezt az eljárást, amely számottevő idő- és

költségmegtakarítással járhat, Wald Ábrahám dolgozta ki. A szekvenciális módszert a 9.10. pontban ismertetjük.

Megjegyezzük még, hogy a matematikai statisztikában szokás kis és nagy mintáról beszélni, jöllehet a két kategóriát nehéz élesen elhatárolni. Általában kis mintának tekintjük az 5-30 mintaelemszám esetét, míg 50-200 elemű minta már nagynak számít. A statisztikai becslés általában nagy mintaelemszámot követel, míg a statisztikai hipotézisek ellenőrzése, különösen ha a kísérletek elvégzése költséges (pl. a termék megsemmisítésével járó vagy bonyolult és drága kísérletről van szó) rendszerint kis mintára vagyunk utalva.

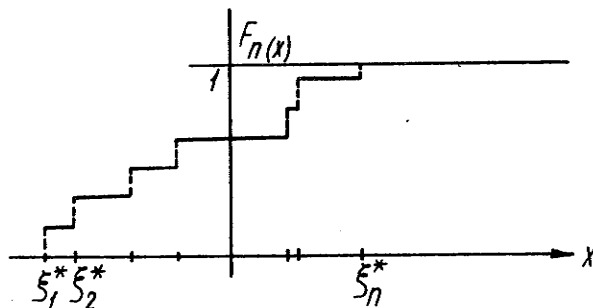
7.4 A statisztikai minta jellemzői

Ha a megfigyelésünk, mérésünk tárgyát képező ξ valószínűségi változóra vonatkozólag n számú megfigyelésünk, a $\xi_1, \xi_2, \dots, \xi_n$ minta rendelkezésünkre áll, akkor az a kérdés merül fel előttünk, hogy a mintaelemek segítségével hogyan közelíthetjük meg a ξ változó ismeretlen $F(x)$ eloszlás függvényét vagy annak egy vagy több állandóját, paraméterét. Más szóval az a problémánk, hogy milyen függvényt konstruáljunk a mintaelemekből, amely függvény valamilyen értelemben közelíti az $F(x)$ eloszlásfüggvényt, vagy milyen számadatot számítsunk ki az n darab tapasztalt számértékből, amely számadat információt ad arra nézve, hogy pl. mennyi a ξ változó várható értéke vagy szórása. Abban az esetben, ha az ismeretlen $F(x)$ eloszlásfüggvény folytonos, célszerű a következő módon eljárni. A mintaelemeket nagyság szerint növekvő sorrendben, vagyis kiválasztjuk a $\xi_1, \xi_2, \dots, \xi_n$ számok közül a legkisebbet, ezt ξ_1^* -gal jelöljük, az utána következőt (vagyis a fennmaradó $n-1$ számú mintaelem közül a legkisebbet) ξ_2^* -gal stb. Ezáltal a

$$\xi_1^* < \xi_2^* < \dots < \xi_n^*$$

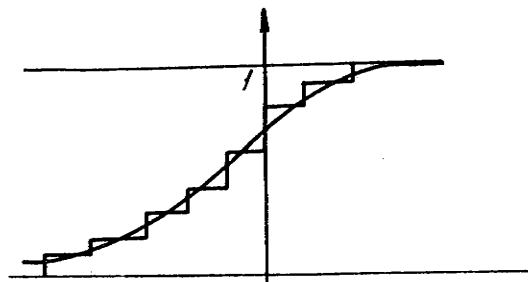
un. rendezett mintát kapjuk. Itt ξ_k^* -ot a k -adik rendezett mintaelemnek nevezzük. Ábrázoljuk a $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ számokat a számegyenesen és alkossuk meg a következő $F_n(x)$ függvényt:

$$(7.3.1) \quad F_n(x) = \begin{cases} 0 & \text{ha } x < \xi_1^* \\ \frac{k}{n} & \text{ha } \xi_k^* \leq x < \xi_{k+1}^* \\ 1 & \text{ha } \xi_n^* < x \end{cases}$$



54. ábra

Mint látható, $F_n(x)$ monoton növekvő lépcsős függvény, amely balról folytonos és az egyes ξ_i^* helyeken $\frac{1}{n}$ nagyságu ugrása van ($i = 1, 2, \dots, n$). Tehát az $F_n(x)$ függvény az eloszlásfüggvények tulajdonságaival rendelkezik, ezt tapasztalati vagy empirikus eloszlásfüggvénynek nevezzük. Mivel a mérendő ξ mennyiségről feltettük, hogy folytonos eloszlású, ezért igen kicsiny (gyakorlatilag 0) annak a valószínűsége, hogy két vagy több mérési eredmény azonos legyen. Ha ez mégis előfordulna, akkor az $F_n(x)$ függvénynek abban a pontban, ahova a megegyező nagyságu mérési eredmények esnek, nem $\frac{1}{n}$, hanem annyiszor $\frac{1}{n}$ nagyságu ugrása van, ahány mérési eredmény egybe esik. Ha a mintaelemek száma n nagy, akkor $F_n(x)$ ugrásai kicsinyek, s ez esetben szemléltetést $F_n(x)$ valamely nyújtott S-alakú $F(x)$ függvényt közelít, ekörül ingadozik.



55. ábra

Vizsgáljuk meg, hogy milyen $F(x)$ függvény az, amelyhez az $F_n(x)$ tapasztalati eloszlásfüggvény illeszkedik. Erre a kérdésre nem lesz nehéz válaszolni, ha figyelembe vesszük a következőket. Az $F_n(x)$ tapasztalati eloszlásfüggvény értéke az x pontban $\frac{k}{n}$, ha n számú mintaelem közül pontosan k számú mintaelem értéke kisebb x -nél. Vagyis az $F_n(x)$ értéke az x pontban a $\{\xi < x\}$ esemény relatív gyakorisága. A $\{\xi < x\}$ esemény valószínűségét az $F(x)$ függvény, a valószínűségi változó elméleti eloszlásfüggvénye adja. Mivel nagy n esetén a relatív gyakoriság erősen megközelíti a valószínűséget, az $F_n(x)$ tapasztalati eloszlásfüggvény nagy mintaelemszámát a ξ valószínűségi $F(x)$ elméleti eloszlásfüggvényét közelíti.

Glivenko szovjet matematikus mutatta ki, hogy tetszőleges kis pozitív ϵ és δ számokhoz megadható olyan N_0 küszöbszám, hogy ha a minta elemszáma $n \geq N$, akkor annak valószínűsége, hogy a tapasztalati és az elméleti eloszlásfüggvények közötti maximális eltérés kisebb legyen ϵ -nál, nagyobb mint $1 - \delta$.

$$P\left\{\sup_x |F_n(x) - F(x)| \leq \epsilon\right\} \geq 1 - \delta \quad \text{hacsak } n \geq N_0.$$

Glivenko tételét a matematikai statisztika alaptételének szokás nevezni.

Minden diszkrét valószínűségi változó eloszlásfüggvénye lépcsős függvény, amelynek a ξ valószínűségi változó lehetséges értékeit képviselő pontokban akkora ugrása van, amekkora valószínűséggel a ξ változó az illető értéket felveszi. Ha most η olyan valószínűségi változó, amelynek

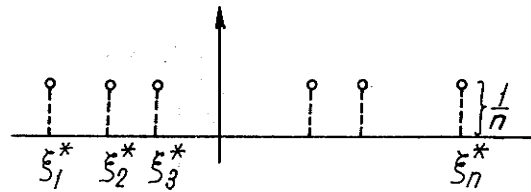
lehetséges értékei éppen a $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ számok, a rendezett minta

megfigyelt elemei, s az η változó a fenti értékei mindegyikét $\frac{1}{n}$ valószínűséggel veszi fel, az eloszlását az alábbi táblázat mutatja:

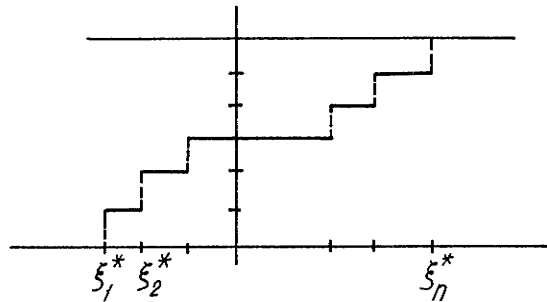
$$\eta :$$

ξ_1^*	ξ_2^*	ξ_3^*		ξ_n^*
$\frac{1}{n}$	$\frac{1}{n}$	$\frac{1}{n}$		$\frac{1}{n}$

akkor η eloszlásfüggvénye éppen az $F_n(x)$ függvény, az empirikus eloszlásfüggvény.



56/a. ábra



56/b. ábra

Mivel az η valószínűségi változó $F_n(x)$ eloszlásfüggvénye közel van a ξ valószínűségi változó $F(x)$ eloszlásfüggvényéhez, kézenfekvő arra gondolnunk, hogy az η változó várható értéke, szórása és egyéb számszerű jellemzői a ξ változó várható értékét, szórását stb. közelítik, más szóval

$$M(\xi) \approx (\eta), \quad D^2(\xi) \approx D^2(\eta) \text{ stb.}$$

Az η valószínűségi változó várható értéke:

$$(7.3.2) \quad (\eta) = \sum_{k=1}^n \frac{1}{n} \xi_k^* = \frac{\xi_1^* + \xi_2^* + \dots + \xi_n^*}{n} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} = \bar{\xi}$$

vagyis η várható értéke a mintaelemek számtani középértéke. Az η valószínűségi változó szórásnégyzete,

$$(7.3.3) \quad D^2(\eta) = \sum_{k=1}^n \frac{1}{n} (\xi_k^* - \bar{\xi})^2 = \frac{(\xi_1 - \bar{\xi})^2 + (\xi_2 - \bar{\xi})^2 + \dots + (\xi_n - \bar{\xi})^2}{n} = S_n^2$$

Az S_n^2 mennyiséget, amely a mintaelemeknek a minta középértékétől való négyzetes átlageltérése, tapasztalati szórásnégyzetnek nevezzük.

A mondottak értelmében $\bar{\xi}$ a ξ valószínűségi változó várható értékét S_n^2 pedig a ξ változó szórásnégyzetét közelíti:

$$\bar{\xi} \sim M(\xi)$$

$$S_n^2 \sim D^2(\xi)$$

Eddigi fejtegetésünkkel pusztán azt kívántuk érzékeltetni, hogy valamely mérendő (megfigyelendő) ξ véletlen mennyiség ismeretlen $F(x)$ eloszlásfüggvényének megközelítése céljából kiindulásul mindig a ξ értékeire vonatkozó n számú megfigyelés a minta szolgál.

A $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ rendezett minta alapján konstruálható $F(x)$ tapasztalati eloszlásfüggvény, a minta eloszlásfüggvénye az $F(x)$ ismeretlen eloszlásfüggvénynek, $\bar{\xi}$ a mintaközép az $M(\xi)$ várható értéknek, az S_n^2 tapasztalati szórásnégyzet a $D^2(\xi)$ elméleti szórásnégyzetnek a közelítése, becslése.

Mínthogy az egyes ξ_i mintaelemek valószínűségi változók, természetesen a belőlük konstruált $\bar{\xi}$, ill. S_n^2 statisztikai függvények is valószínűségi változók. Arra a kérdésre vonatkozólag, hogy $\bar{\xi}$ mennyire közelíti az $M(\xi)$ elméleti szórásnégyzetet, csak akkor adható kielégítő válasz, ha ismerjük $\bar{\xi}$, ill. S_n^2 eloszlását, vagy legalábbis a várható értéküket és a szórásukat.

Mivel mind $\bar{\xi}$, mind S_n^2 független valószínűségi változók összege, ezért a centrális határeloszlástétel érvényessége folytán - nagy mintaelemszám esetén - közelítőleg normális eloszlásuak.

Pl. ha $n > 30$, akkor a $\bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i$ mintaközép már normális elosz-

lásúnak vehető, függetlenül attól, hogy milyen eloszlásu a vizsgált ξ valószínűségi változó, amely azt a statisztikai sokaságot reprezentálja, amelyből a mintát vettük.

A $\bar{\xi}$ valószínűségi változó várható értéke:

$$M(\bar{\xi}) = M\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n} \sum_{i=1}^n M(\xi_i) = \frac{1}{n} \cdot n \cdot M(\xi) = M(\xi),$$

azaz a mintaközép várható értéke megegyezik a vizsgált ξ valószínűségi változó várható értékével.

A $\bar{\xi}$ mintaközép szórásnégyzete:

$$D^2(\bar{\xi}) = D^2\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n D^2(\xi_i) = \frac{1}{n^2} \cdot n D^2(\xi) = \frac{D^2(\xi)}{n}$$

tehát a mintaközép szórása: $D(\bar{\xi}) = \frac{D(\xi)}{\sqrt{n}}$, lényegesen kisebb, mint a ξ

változó szórása, s a mintaelemszám növelésével $\bar{\xi}$ szórása tetszőleges kicsinnyé tehető. Elég nagy ($n > 30$) minta esetén tehát $\bar{\xi}$ közelítőleg normális eloszlású, $m = M(\xi)$, $\sigma = \frac{D(\xi)}{\sqrt{n}}$ paraméterekkel.

Sok esetben előfordul, hogy ξ -re vonatkozólag nem tudunk elég sok megfigyelést végezni, hanem kicsiny (pl. 8-10 elemű) mintából kell következtetnünk a szóbanforgó eloszlás várható értékére, ill. szórására. Ilyenkor szükségünk van $\bar{\xi}$, ill. S_n^2 pontos eloszlásának ismeretére. A gyakorlat szempontjából különösen fontos az az eset, amikor ξ normális eloszlású, ezért fogjuk most megvizsgálni, hogy milyen az eloszlása egy normális eloszlású sokaságból származó minta esetében a $\bar{\xi}$ mintaközépnek és a S_n^2 tapasztalati szórásnégyzetnek.

A normális eloszlású additív tulajdonságából következik, hogy ha ξ normális eloszlású m várható értékkel és σ szórással, akkor a $\bar{\xi}$ mintaközép ugyancsak normális eloszlású m várható értékkel és $\frac{\sigma}{\sqrt{n}}$ szórással, n bármilyen értékére (tehát nemcsak nagy n -re).

Vizsgáljuk most az S_n^2 tapasztalati szórásnégyzetét, eloszlását.

Az általánosság csorbitása nélkül feltehetjük, hogy $\xi \sim N(0, \sigma)$ eloszlású. (Ha ugyanis ξ várható értéke $m \neq 0$, akkor minden megfigyelt értékből levonhatunk m -et, így 0 várható értékű valószínűségi változót vizsgálunk.) Ekkor a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók függetlenek és mindegyikük $N(0, \sigma)$ eloszlású.

Tekintsük a $\xi_1, \xi_2, \dots, \xi_n$ mintát egy n -dimenziós véletlen vektornak és válasszunk egy olyan $\underline{A} = [a_{ij}]$ n -edrendű ortogonális matrixot, amelyben az utolsó sor minden eleme $\frac{1}{\sqrt{n}}$. Ilyen matrix nyilván található, hiszen geometriailag arról van szó, hogy az $(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ vektor-

hoz választanunk kell $n-1$ darab rá merőleges vektort, ezek lesznek a matrix sorai. Ilyen alkalmas matrix pl. a következő:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{2}{\sqrt{2}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{2.3}} & \frac{1}{\sqrt{2.3}} - \frac{2}{\sqrt{2.3}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{3.4}} & \frac{1}{\sqrt{3.4}} & \frac{1}{\sqrt{3.4}} - \frac{3}{\sqrt{3.4}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \dots & \dots & \frac{1}{\sqrt{n}} \end{pmatrix}$$

Végezzük el az

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}$$

ortogonális transzformációt, amely az n -dimenziós térben az origó körüli elforgatást jelent, s ennek során valamely pontnak az origótól való távolsága változatlan marad, tehát

$$\sum_{i=1}^n \xi_i^2 = \sum_{i=1}^n \eta_i^2$$

ahol, mint a transzformáló egyenletből látható:

$$\eta_i = \sum_{j=1}^n a_{ij} \xi_j$$

Mivel $\underline{A}$ ortogonális matrix, $\underline{A} \cdot \underline{A}^* = \underline{E}$ (itt $\underline{A}$ az $\underline{A}^*$ matrix transzponáltját, $\underline{E}$ pedig egységmatrixot jelöl), azaz

$$\sum_{k=1}^n a_{ik} a_{kj} = \begin{cases} 1 & \text{ha } i = j \\ 0 & \text{ha } i \neq j \end{cases}$$

$$\sum_{k=1}^n a_{ki} a_{kj} = \begin{cases} 1 & \text{ha } i = j \\ 0 & \text{ha } i \neq j \end{cases}$$

Az $\eta_1, \eta_2, \dots, \eta_n$ valószínűségi változók származtatásából kézenfekvő, hogy:

$$M(\eta_i) = M\left(\sum_{j=1}^n a_{ij} \xi_j\right) = \sum_{j=1}^n a_{ij} M(\xi_j) = 0$$

mivel:

$$M(\xi_j) = 0 \quad j = 1, 2, \dots, n$$

Továbbá:

$$\begin{aligned} (7.3.4) \quad M(\eta_i \eta_j) &= M\left(\sum_k a_{ik} \xi_k\right) \left(\sum_n a_{jk} \xi_n\right) = \\ &= \sum_k \sum_n a_{ik} a_{jn} M(\xi_k \xi_n) = \sigma^2 \sum_k a_{ik} a_{jk} = 0 \end{aligned}$$

$$(7.3.5) \quad \text{ha } i \neq j \quad \text{és } M(\eta_i^2) = \sigma^2$$

Kimutatható, hogy ha a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók együttes eloszlása, vagyis a $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ véletlen vektor eloszlása normális eloszlás, akkor ezen vektor bármely lineáris transzformáltja ugyancsak normális eloszlású, feltéve, hogy a transzformáció determinánsa zérustól különbözik. Esetünkben, mivel ortogonális transzformációt alkalmaztunk, amelynek determinánsa 1, ez a feltétel teljesül, tehát az $\eta_1, \eta_2, \dots, \eta_n$ változók együttes eloszlása normális eloszlás, az egyes η_i változók $N(0, \sigma)$ eloszlásúak és a (7.3.4) összefüggés alapján korrelálatlanok, ami ez esetben a függetlenséget is jelenti.

Az $\frac{\eta_1}{\sigma}, \frac{\eta_2}{\sigma}, \dots, \frac{\eta_n}{\sigma}$ valószínűségi változók független $N(0,1)$ eloszlásúak.

Mint a transzformáló egyenletből látható

$$\eta_n = \bar{\xi} \sqrt{n}$$

$$n S_n^2 = \sum_{i=1}^n (\xi_i - \bar{\xi})^2 = \sum_{i=1}^n \xi_i^2 - n \bar{\xi}^2 = \sum_{i=1}^{n-1} \eta_i^2$$

Tehát:

$$\frac{n S_n^2}{\sigma^2} = \sum_{i=1}^{n-1} \left(\frac{\eta_i}{\sigma} \right)^2$$

Az $\frac{n S_n^2}{\sigma^2}$ változó $n-1$ számú $N(0,1)$ normális eloszlású valószínűségi változó összege, tehát χ^2 eloszlású $n-1$ szabadságfokkal. Továbbá, mivel az $\eta_1, \eta_2, \dots, \eta_{n-1}$ változók függetlenek az $\eta_n = \bar{\xi} \sqrt{n}$ változóktól, következik, hogy $\bar{\xi}$ és S_n^2 függetlenek egymástól. Ezt az eredményünket a 8.1 pontban fogjuk felhasználni.

Összegezve a kapott eredményeket: ha a vizsgált ξ valószínűségi változó $N(m; \sigma)$ normális eloszlású, akkor a ξ -re vonatkozó n elemű

mintából számított $\bar{\xi}$ mintaközép $N\left(m; \frac{\sigma}{\sqrt{n}}\right)$ normális eloszlású, az

$\frac{n S_n^2}{2}$ statisztikai függvény pedig χ^2 eloszlású $n-1$ szabadságfokkal.

BECSLÉSELMÉLET

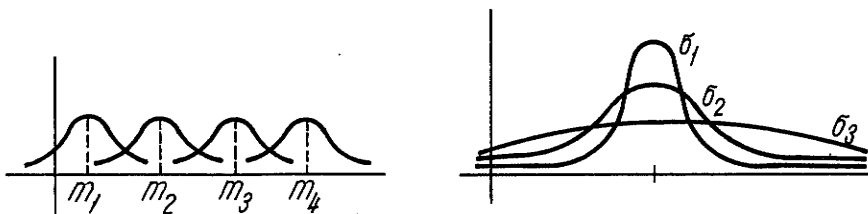
8.1 A becslés fogalma és tulajdonságai

A statisztikai gyakorlatban számos esetben előfordul, hogy a vizsgált ξ valószínűségi változó eloszlásának függvényalakját az $F(x)$ eloszlásfüggvényt, ill. az $f(x)$ sűrűségfüggvényt megadó formulát ismerjük, az eloszlás azonban bizonyos ismeretlen paramétereket tartalmaz, azaz $f(x) = f(x; \theta_1, \theta_2, \dots, \theta_k)$ alakban írható, ahol $\theta_1, \theta_2, \dots, \theta_k$ az ismeretlen paraméterek. Gyakori eset, hogy bizonyos mérési eredményről tudjuk, hogy normális eloszlást követ, azaz sűrűségfüggvénye.

$$f(x; \theta_1, \theta_2) = \frac{1}{\theta_2 \sqrt{2\pi}} e^{-\frac{(x - \theta_1)^2}{2\theta_2^2}} = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

itt $\theta_1 = m$, $\theta_2 = \sigma$ az esetleg ismeretlen paraméterek.

A sűrűségfüggvény formulájában szereplő paraméterek a gyakorlat számára legfontosabb esetekben az eloszlás várható értékét, szórását esetleg korrelációs együttható többdimenziós eloszlás esetén jelentik, amelyek ismerete rendkívül fontos, nélkülözhetetlen ahhoz, hogy a vizsgált véletlen mennyiségre vonatkozólag valószínűségi jellegű megállapításokat tehesünk, hiszen pl. a normális eloszlás esetében annak valószínűsége, hogy a ξ változó megfigyelt értéke adott $[a, b]$ intervallumba essék, igen különböző lehet, attól függően, hogy mekkora az m és σ paraméterek értéke, amint azt az alábbi ábra is mutatja:



57. ábra

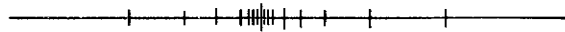
Tudjuk, hogy egy telefonközpontba befutó hívások száma Poisson-eloszlást követ, így pl. annak valószínűsége, hogy egy percig számlálva a hívásokat, pontosan 12 hívást észleljünk:

$$P_{12} = \frac{\lambda^{12}}{12!} e^{-\lambda}$$

Nyilvánvaló, hogy a λ paraméter ismerete nélkül ezzel a formulával semmit nem érünk, λ értékét tehát megfigyelési adatokból feltétlenül meg kell becsülnünk ahhoz, hogy a kérdéses valószínűséget akár közelítőleg is meg tudjuk mondani egy adott központ esetében.

Az egyik alapvető statisztikai probléma tehát a következő: a vizsgált ξ valószínűségi változó sűrűségfüggvénye legyen $f(x; \Theta)$, ahol Θ ismeretlen paramétert vagy paraméter-vektort jelöl. Hogyan lehet meghatározni (legalábbis jó közelítőleg) a Θ ismeretlen paraméter értékét a ξ -re vonatkozó n számú megfigyelés, a $\xi_1, \xi_2, \dots, \xi_n$ minta alapján?

Mint hogy a mintaelemek valószínűségi változók, s mindegyik minta-elem eloszlása megegyezik a vizsgált ξ változó eloszlásával, a minta nyilván tájékoztatást, információt rejt magában ξ eloszlására és az eloszlás állandóira, paramétereire vonatkozólag. Ha a mintaelemeket a számegyenesen ábrázoljuk, s e mintaelemek valamely pont körül tömörülnek, ott gyanítjuk az eloszlás súlypontját, várható értékét.



58. ábra

A mintaelemek szétszórtságából vagy tömörültségéből az eloszlás szórására igyekszünk következtetni. A probléma elsősorban az, hogy miféle számadatokat konstruáljunk a mintaelemekből, milyen "függvényt" képezzünk a $\xi_1, \xi_2, \dots, \xi_n$ változóknak, hogy a konstruált számadat, amelyet statisztikának is szokás nevezni, valamilyen értelemben jól közelítse az ismeretlen Θ paraméter értékét.

Ha a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemeknek valamilyen

$$\hat{\Theta} = \hat{\Theta}(\xi_1, \xi_2, \dots, \xi_n)$$

függvényét képezzük, ez a függvény ugyancsak valószínűségi változó. A ξ valószínűségi változó eloszlása meghatározott függvénytípusú $f(x; \Theta)$ sűrűségfüggvény esetében a Θ paraméter értékétől függ, ugyanez érvényes minden egyes ξ_i mintaelem eloszlására és ugyanugy a Θ paraméter értékétől

függ a $\hat{\Theta} = \hat{\Theta}(\xi_1, \xi_2, \dots, \xi_n)$ statisztikai eloszlása is. Ha $\hat{\Theta}$ kiszámításának az a célja, hogy vele a Θ paraméter tényleges értékét becsüljük, akkor azt kívánjuk meg $\hat{\Theta}$ -tól, hogy értéke minél közelebb essék Θ tényleges értékéhez, azaz $\hat{\Theta}$ eloszlása a Θ pont mint súlypont körül koncentráldjék. Ez alapvető kívánsága egy "jó" becslésnek.

A becslés "jóságának" megítélése igen különböző szempontokból lehetséges. Általában arra törekszünk, hogy statisztikánk, becslésünk az alábbi tulajdonságokkal rendelkezzen:

1. A becslés legyen torzítatlan. A $\hat{\Theta} = \hat{\Theta}(\xi_1, \xi_2, \dots, \xi_n)$ statisztikát a Θ paraméter torzítatlan becslésének nevezzük, ha a $\hat{\Theta}$ várható értéke Θ -val, a becsült állandóval egyezik meg, az

$$M(\hat{\Theta}) = \Theta.$$

Mutatunk egy példát torzítatlan becslésre. Tegyük fel, hogy egy normális eloszlású ξ valószínűségi változó $\Theta = m$ várható értékét akarjuk becsülni a ξ -re vonatkozó $\xi_1, \xi_2, \dots, \xi_n$ minta segítségével. A pontban vázolt megfontolások alapján az $M(\xi) = m$ várható érték becslésére a mintaelemek számtani közepét használjuk, azaz képezzük a

$$\hat{\Theta} = \bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} \text{ mintaközepet.}$$

A $\bar{\xi}$ mintaközép torzítatlan becslése $M(\xi)$ -nek, mert a mintaközép várható értéke

$$M(\bar{\xi}) = \frac{1}{n} \sum_{i=1}^n M(\xi_i) = \frac{1}{n} \cdot n M(\xi) = M(\xi) = m.$$

Megjegyezzük, hogy ha a $\Theta = m$ várható érték becslésére a mintaelemeknek valamely súlyozott számtani közepét használjuk, azaz képezzük a

$$\tilde{\Theta} = p_1 \xi_1 + p_2 \xi_2 + \dots + p_n \xi_n \text{ statisztikát, ahol}$$

a p_i számok mindegyike nemnegatív és $\sum_{i=1}^n p_i = 1$, akkor ugyancsak torzítatlan becslést nyerünk m -re nézve, ugyanis:

$$M(\tilde{\Theta}) = p_1 M(\xi_1) + p_2 M(\xi_2) + \dots + p_n M(\xi_n) = M(\xi)(p_1 + \dots + p_n) = m$$

Ennek ellenére a $\Theta = m$ várható érték becslésére előnyösebb a mintaelemek közönséges számtani közepét használni, mint bármely más súlyozott számtani közepet, mint azt mindjárt látni fogjuk.

Ha a ξ valószínűségi változó $\Theta = \sigma^2$ szórásnégyzetének becslésére a

$$\hat{\Theta} = S_n^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2 \text{ statisztikát használjuk, ez a becslés nem tesz}$$

eleget a torzítatlanság követelményének, ez a becslés torzított, mert

$M(S_n^2) \neq \sigma^2$. Az S_n^2 statisztika ugyanis a következő alakban írható:

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n [(\xi_i - m) - (\bar{\xi} - m)]^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - m)^2 - (\bar{\xi} - m)^2$$

$$M(\xi_i - m)^2 = D^2(\xi_i) = \sigma^2$$

$$M(\bar{\xi} - m)^2 = D^2(\bar{\xi}) = D^2\left(\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}\right) = \frac{1}{n^2} \cdot n D^2(\xi) = \frac{\sigma^2}{n}$$

Ezek alapján

$$M(S_n^2) = \frac{1}{n} \cdot n \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \left(1 - \frac{1}{n}\right) = \frac{n-1}{n} \sigma^2$$

Tehát az S_n^2 valószínűségi változó értékei nem a σ^2 , hanem a $\sigma^2 - \frac{\sigma^2}{n}$

érték körül ingadoznak. A torzítás könnyen kiküszöbölhető, mert az

$S_n^{*2} = \frac{n}{n-1} S_n^2$ statisztika már torzítatlan becslése σ^2 -nek:

$$M(S_n^{*2}) = M \frac{(\xi_1 - \bar{\xi})^2 + (\xi_2 - \bar{\xi})^2 + \dots + (\xi_n - \bar{\xi})^2}{n-1} = \sigma^2.$$

Egyébként $n \rightarrow \infty$ esetén $(1 - \frac{1}{n})\sigma^2 \rightarrow \sigma^2$, tehát nagy n esetén a torzítás

jelentéktelen. Az olyan becslést, amely $n \rightarrow \infty$ esetén torzítatlanná válik, aszimptotikusan torzítatlannak nevezzük.

2. Arra törekszünk, hogy a becslés minél hatásosabb legyen. Ezen azt értjük, hogy a $\hat{\Theta}$ statisztika értékei nagy valószínűséggel a Θ paraméter értékének minél kisebb környezetébe essenek, azaz.

$M[(\hat{\Theta} - \Theta)^2]$ minél kisebb legyen. Torzítatlan becslés esetében ez a követelmény azt jelenti, hogy $\hat{\Theta}$ minél kisebb szórásnégyzetű legyen. Ha $\hat{\Theta}_1$ és $\hat{\Theta}_2$ a Θ paraméter két becslése, akkor közülük azt tekintjük hatásosabbnak, amelyre $M[(\hat{\Theta}_1 - \Theta)^2]$ kisebb. Az $\hat{\Theta}_2$ becslés relatív hatásfokát a $\hat{\Theta}_1$ becslésre vonatkozólag az

$$\frac{M[(\hat{\Theta}_1 - \Theta)^2]}{M[(\hat{\Theta}_2 - \Theta)^2]}$$

hányados definiálja. Ha e hányados értéke kisebb, mint 1, akkor $\hat{\Theta}_1$ hatásosabb becslése Θ -nak, mint $\hat{\Theta}_2$.

Ha pl. a $\Theta = m$ várható érték becslésére valamely

$$\hat{\Theta}_2 = p_1 \xi_1 + p_2 \xi_2 + \dots + p_n \xi_n$$

statisztikát használunk

$$(p_i \geq 0, \sum_1^n p_i = 1)$$

akkor

$$D^2(\hat{\Theta}_2) = \sum_1^n p_i^2 D^2(\xi_i) = \sigma^2 \sum_1^n p_i^2 \geq \frac{\sigma^2}{n}$$

Ugyanis a Cauchy-féle egyenlőtlenség alapján

$$(\sum a_i b_i)^2 \leq \sum a_i^2 \sum b_i^2$$

legyen most $a_i = p_i$, $b_i = 1$ ($i = 1, 2, \dots, n$), ekkor

$$1 = \left(\sum p_i \cdot 1 \right)^2 \leq \sum p_i^2 \cdot n$$

vagyis

$$\sum p_i \geq \frac{1}{n}$$

Tehát a $\Theta = m$ várható érték becslésére a $\bar{x}$ számtani közép a legjobb hatásfoku lineáris becslés, mivel a mintaelemek bármely konvex lineáris kombinációjának nagyobb a szórása, mint a közönséges számtani középnek.

3. Minél több megfigyelést végzünk valamely ξ valószínűségi változóra vonatkozólag, azaz minél nagyobb a minta elemszáma, annál pontosabban tudjuk becsülni a kérdéses Θ paramétert.

Legyen $\hat{\Theta}_n = \hat{\Theta}_n(\xi_1, \xi_2, \dots, \xi_n)$ torzítatlan becslése Θ -nak. Ha $M[(\hat{\Theta}_n - \Theta)^2] \rightarrow 0$ ha $n \rightarrow \infty$, akkor a Csebisev-egyenlőtlenségből következik, hogy

$$(*) P(|\hat{\Theta}_n - \Theta| > \varepsilon) \rightarrow 0 \text{ ha } n \rightarrow \infty$$

Ugyanis torzítatlan becslés esetében:

$$P(|\hat{\Theta}_n - \Theta| > \varepsilon) < \frac{M[(\hat{\Theta}_n - \Theta)^2]}{\varepsilon^2}$$

amiből a (*) reláció már adódik. Az olyan $\hat{\Theta}_n$ becslést, amelyre a (*) összefüggés teljesül a Θ paraméter konzisztens becslésének nevezzük.

Felmerül a kérdés, hogy rögzített véges n elemű minta esetén a legjobb hatásfoku $\hat{\Theta}_n$ becslés szórása tetszőlegesen kicsi lehet-e, vagy pedig megadható olyan alsó határ, amelynél $\hat{\Theta}_n$ szórása nem lehet kisebb.

Erre ad választ az ún. Rao-Cramér egyenlőtlenség, amely azt mondja ki, hogy véges rögzített n elemű mintából számított bármely becslés szórása nem lehet kisebb, mint egy bizonyos előre kiszámítható mennyiség.

Legyen ugyanis $f(x_1, x_2, \dots, x_n; \Theta)$ a ξ valószínűségi változóra vonatkozó $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek együttes sűrűségfüggvénye, amely egy ismeretlen Θ paramétert tartalmaz. Legyen továbbá $\hat{\Theta} = \hat{\Theta}(\xi_1, \xi_2, \dots, \xi_n)$ a Θ paraméter torzítatlan becslése. Nyilván fennáll, hogy

$$(8.1.1) \quad \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n; \Theta) dx_1 dx_2 \dots dx_n = \int_{R_n} f dv = 1$$

Feltesszük, hogy szabad a Θ paraméter szerint az integrál jel alatt deriválni, azaz a (1) összefüggés mindkét oldalát Θ szerint deriválva:

$$(8.1.2) \quad \frac{d}{d\Theta} \int_{R_n} f dv = \int_{R_n} \frac{df}{d\Theta} dv = 0$$

A (8.1.2) összefüggés az alábbi alakban is írható:

$$(8.1.3) \quad \int_{R_n} \frac{1}{f} \frac{df}{d\Theta} f dv = \int_{R_n} \frac{1}{f} \frac{df}{d\Theta} df = 0$$

(ahol F jelöli a mintaelemek együttes eloszlásfüggvényét).

Mivel $\hat{\Theta}$ a Θ paraméter torzítatlan becslése,

$$\int_{R_n} \hat{\Theta} f dv = \Theta$$

Innen:

$$(8.1.4) \quad \int_{R_n} \hat{\Theta} \frac{df}{d\Theta} f dv = 1.$$

A (8.1.3) összefüggésből következik, hogy

$$(8.1.5) \quad \int_{R_n} \Theta \cdot \frac{1}{f} \frac{df}{d\Theta} f dv = 0.$$

A (n) és (5) összefüggéseket kivonva egymásból, kapjuk, hogy

$$(8.1.6) \quad \int_{R_n} (\hat{\Theta} - \Theta) \frac{1}{f} \frac{df}{d\Theta} f dv = \int_{R_n} (\hat{\Theta} - \Theta) \left(\frac{1}{f} \cdot \frac{df}{d\Theta} \right) dF = 1.$$

A (8.1.6) összefüggésre alkalmazva a Schwartz-féle egyenlőtlenséget:

$$I = \int_{R_n} (\hat{\Theta} - \Theta) \left(\frac{1}{f} \frac{df}{d\Theta} \right) dF \leq \int_{R_n} (\hat{\Theta} - \Theta)^2 dF \cdot \int_{R_n} \left(\frac{1}{f} \frac{df}{d\Theta} \right)^2 dF$$

Innen:

$$(8.1.7) \quad \int_{R_n} (\hat{\Theta} - \Theta)^2 dF \geq \frac{1}{\int_{R_n} \left(\frac{1}{f} \frac{df}{d\Theta} \right)^2 dF}$$

A (8.1.7) összefüggésben a baloldalon a $\hat{\Theta}$ statisztikai függvény szórása áll. A jobboldal nevezője az ún. Fischer-féle információs mennyiség:

$$(8.1.8) \quad I = \int_{R_n} \left(\frac{1}{f} \frac{df}{d\Theta} \right)^2 dF = \int_{R_n} \left(\frac{d \ln f}{d\Theta} \right)^2 dF$$

Ha az $f(x_1, x_2, \dots, x_n; \Theta)$ függvényben az $x_1, x_2, \dots, x_n$ változók helyébe a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek megfigyelt értékeit helyettesítjük, akkor $f(\xi_1, \xi_2, \dots, \xi_n; \Theta)$ valószínűségi változó, következésképpen a

(8.1.8) összefüggés jobboldala a $\frac{(d \ln f)}{d\Theta}$ valószínűségi változónak a várható értéke, azaz

$$I = M \left(\frac{d \ln f}{d\Theta} \right)^2$$

A (8.1.7) egyenlőtlenség, a Rao-Cramér egyenlőtlenség, amelyet a matematikai statisztika határozatlansági relációjának nevezhetünk, azt mondja ki, hogy a Θ paraméter tetszőleges $\hat{\Theta}$ torzítatlan becslésének szórásnégyzete nem lehet kisebb, mint a Fischer-féle információs mennyiség reciprokok értéke:

$$D^2(\hat{\Theta}) \geq \frac{1}{I}.$$

Bizonyos becslésekre az alsó határ elérik, azaz ilyenkor

$$D^2(\hat{\Theta}) = \frac{1}{I}.$$

Mutatunk erre két példát.

a) Legyen ξ binomiális eloszlású $\Theta = p$ paraméterrel.
A $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek együttes sűrűségfüggvénye most:

$$f(\xi_1, \xi_2, \dots, \xi_n; p) = p^k (1-p)^{n-k}$$

(ha az n számú megfigyelés során k -szor következett be a bennünket érdeklő esemény).

Mivel

$$\ell_n f = k \ell_n p + (n-k) \ell_n (1-p)$$

$$\frac{d \ell_n f}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-np}{p(1-p)}$$

$$I = M \left[\frac{k-np}{p(1-p)} \right]^2 = \frac{1}{p^2 (1-p)^2} \cdot M(k-np)^2 = \frac{D^2(k)}{p^2 (1-p)^2} = \frac{np(1-p)}{p^2 (1-p)^2} = \frac{n}{p(1-p)}$$

Ha most a p paraméter becslésére a $\hat{p} = \frac{k}{n}$ relatív gyakoriságot használjuk, akkor

$$M\left(\frac{k}{n}\right) = \frac{1}{n} M(k) = \frac{np}{n} = p$$

$$D^2\left(\frac{k}{n}\right) = \frac{p(1-p)}{n} = \frac{1}{I}$$

Ez az összefüggés mutatja, hogy a p valószínűség becslésére nincs jobb hatásfoku becslés, mint a relatív gyakoriság.

b) Legyen ξ Poisson-eloszlású λ paraméterrel, s tegyük fel, hogy a ξ -re vonatkozó n számú megfigyelés során a $\xi_1 = k_1, \xi_2 = k_2, \dots, \xi_n = k_n$ értékeket észleltük.

Ekkor

$$f(\xi_1, \xi_2, \dots, \xi_n; \lambda) = \frac{\lambda^{k_1}}{k_1!} \cdot \frac{\lambda^{k_2}}{k_2!} \cdots \frac{\lambda^{k_n}}{k_n!} e^{-n\lambda}$$

$$\frac{1}{f} \frac{df}{d\lambda} = -n + \frac{k_1}{\lambda} + \frac{k_2}{\lambda} + \dots + \frac{k_n}{\lambda}$$

$$I = M \left(\frac{1}{f} \frac{df}{d\lambda} \right)^2 = \frac{1}{\lambda^2} \sum_{i=1}^n D^2(k_i) = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

Ha most a λ paraméter becslésére a $\hat{\lambda} = \frac{k_1 + k_2 + \dots + k_n}{n} = \bar{\xi}$ statisztikai függvényt használjuk, akkor

$$D^2(\hat{\lambda}) = \frac{1}{n^2} \sum_{i=1}^n D^2(k_i) = \frac{n\lambda}{n^2} = \frac{\lambda}{n} = \frac{1}{I}$$

Eredményünk azt mutatja, hogy Poisson eloszlás esetében a λ várható érték becslésére nincs jobb hatásfoku becslés, mint a számtani közép.

4. A matematikai statisztikában arra törekszünk, hogy a vizsgált ξ valószínűségi változó eloszlására vonatkozólag a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemekből minden információt kimerítsünk, amit a mintaelemek az eloszlást illetőleg tartalmaznak. A becsléelmélet szempontjából ez a törekvés úgy fogalmazható, hogy a mintaelemekből lehetőleg olyan statisztikai függvényt igyekszünk konstruálni, amely mindazt az információt magában foglalja, amit a minta a becsült paraméterre vonatkozólag tartalmaz. Az ilyen tulajdonsággal bíró statisztikai függvényt elégséges becslésnek nevezzük.

Bizonyos esetekben a kérdéses paramétert illetőleg lehet elégséges becslést találni. Mutatunk erre néhány példát.

Tegyük fel, hogy $\xi \sim N(m; \sigma)$ normális eloszlású valószínűségi változó, amelynek m várható értékét a

$$\xi_1, \xi_2, \dots, \xi_n$$

mintából a $\bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i$ mintaközéppel becsüljük.

Mint láttuk, $\bar{\xi}$ eloszlása $N(m; \frac{\sigma}{\sqrt{n}})$ normális eloszlás, tehát a sűrűségfüggvénye:

$$f_n(x) = \frac{1}{\sigma} \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\bar{\xi} - m)^2}{2\sigma^2}}$$

Mint hogy a ξ_1 mintaelemek függetlenek és $N(m; \sigma)$ normális eloszlásúak, együttes sűrűségfüggvényük az egyes mintaelemek sűrűségfüggvényeinek szorzata:

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - m)^2}$$

Figyelembe véve az alábbi egyszerű számolással belátható összefüggést:

$$\sum_{k=1}^n (x_k - \bar{x})^2 = \sum_{k=1}^n (x_k - m)^2 - n(\bar{x} - m)^2,$$

a mintaelemek együttes sűrűségfüggvénye a következő alakban írható:

$$f(x_1, x_2, \dots, x_n) = \frac{1}{\sigma} \sqrt{\frac{n}{2\pi}} e^{-\frac{n(\bar{x} - m)^2}{2\sigma^2}} \cdot \frac{1}{(\sigma\sqrt{2\pi})^{n-1} \sqrt{n}} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \bar{x})^2}$$

a $\xi_1, \xi_2, \dots, \xi_n$ változók együttes feltételes sűrűségfüggvénye, azon feltétel mellett, hogy a $\bar{\xi}$ mintaközép egy adott $\bar{x}$ értékkel egyenlő:

$$f(x_1, x_2, \dots, x_n | \bar{\xi} = \bar{x}) = \begin{cases} \frac{1}{(\sigma\sqrt{2\pi})^{n-1} \sqrt{n}} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \bar{x})^2} \\ 0 \text{ egyébként} \quad \text{ha } \bar{\xi} \neq \bar{x} \end{cases}$$

Ez a feltételes sűrűség többé nem függ m értékétől. Ez azt jelenti, hogy amikor a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemekből a $\bar{\xi} = \bar{x}$ mintaközepet meghatároztuk, ezzel az m várható értéket illetően a mintaelemekből minden információt kimerítettünk, tehát $\bar{\xi}$ elégséges becslés az $M(\xi) = m$ várható értékre vonatkozólag.

Megjegyezzük, hogy a mintaközép Poisson-eloszlás, valamint az exponenciális eloszlás esetében is elégséges becslése a λ paraméternek. Ugyancsak elégséges becslés a relatív gyakoriság valamely A esemény

ismeretlen $P(A) = p$ valószínűségére nézve, ugyanis, ha az A esemény ξ karakterisztikus változójára vonatkozó

$\xi_1, \xi_2, \dots, \xi_n$ minta k -számu 1-est (és $n-k$

számu 0-t) tartalmaz, akkor mivel binomiális eloszlásról van szó:

$$P(\xi_1 + \xi_2 + \dots + \xi_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

A mintaelemek együttes eloszlása:

$$P(\xi_1 = \varepsilon_1, \xi_2 = \varepsilon_2, \dots, \xi_n = \varepsilon_n) = p^k (1-p)^{n-k} \quad (\varepsilon_i = \begin{cases} 0 \\ 1 \end{cases} \sum_{i=1}^n \varepsilon_i = k)$$

Ez annak a valószínűsége, hogy az A esemény éppen olyan sorrendben következzen be k -szor, ahogyan a mintaelemek mutatják. (Tudjuk, hogy k -számu 1-es n helyen $\binom{n}{k}$ féle módon helyezkedhet el, az adott mintán ezen elhelyezkedések egyike.)

8.2 A becslés módszere, maximum-likelihood módszer

Megismertük a becslés különböző "jó" tulajdonságait és egy-két statisztikai függvényről megmutattuk, hogy rendelkezik a vázolt tulajdonságokkal vagy azok egy részével. Azt is több helyen említettük, hogy valamilyen valószínűségeloszlás várható értékének becslésére a korrigált empirikus szórásnégyzetet valamely esemény ismeretlen valószínűségének becslésére az illető esemény relatív gyakoriságát használjuk. A gyakorlatban legtöbbször ezek az ismeretlen paraméterek, amelyeket becsülnünk kell.

A matematikai érdeklődésű olvasóban természetesen felmerül az a kérdés, hogy nincs-e valami olyan általános matematikai módszer, amelynek segítségével adott $f(x; \Theta)$ eloszlás esetében a Θ paraméterre vonatkozólag "jó" becslés nyerhető.

A legfontosabb általános módszer a becslésre az ún. maximum likelihood (maximális valószínűség) módszere, amelyet R. A. Fisher vezetett be, s amelyet az alábbiakban ismertetünk.

Tegyük fel, hogy egydimenziós folytonos eloszlású ξ valószínűségi változóval állunk szemben, amelynek eloszlása egy ismeretlen paramétert tartalmaz, tehát sűrűségfüggvénye $f(x; \Theta)$ alakban írható.

Legyen a ξ változóra vonatkozó minta:

$$\xi_1, \xi_2, \dots, \xi_n$$

Definiáljunk most egy függvényt, az un. likelihood-függvényt az alábbi relációval:

$$L(x_1, x_2, \dots, x_n; \Theta) = f(x_1; \Theta) \cdot f(x_2; \Theta) \dots f(x_n; \Theta)$$

Mint látható, az $L(x_1, x_2, \dots, x_n; \Theta)$ függvény a $\xi_1, \xi_2, \dots, \xi_n$ változók együttes sűrűsége az n -dim. tér $(x_1, x_2, \dots, x_n)$ pontjában.

Ha a vizsgált statisztikai sokaság, illetve a sokaságot reprezentáló ξ valószínűségi változó diszkrét eloszlású és $P(\xi = x_i) = p_i(\Theta)$, s az

$x_1, x_2, \dots, x_r$ mintaértékek gyakorisága rendre

$k_1, k_2, \dots, k_r$, akkor a likelihood-függvényt az

$$L(x_1, x_2, \dots, x_n; \Theta) = [p_1(\Theta)]^{k_1} \cdot [p_2(\Theta)]^{k_2} \dots [p_r(\Theta)]^{k_r}$$

reláció definiálja.

A likelihood-függvény tehát folytonos eloszlás esetében a mintaelemek együttes sűrűségét, diszkrét eloszlás esetében a mintaelemek együttes valószínűségét jelenti. (A mintaelemek független valószínűségi változók!)

Ha a mintaértékek már adottak (az n számú megfigyelést elvégeztük), akkor a mintaelemek értékét a likelihood függvényt definiáló egyenletbe helyettesítve, a függvény értéke csak a Θ paramétertől függ, mind folytonos, mind diszkrét eloszlás esetében.

A maximum-likelihood módszer abból áll, hogy meghatározzuk azt az $\hat{\Theta}$ értéket, amelyre $L(x_1, \dots, x_n; \Theta)$ mint a Θ függvénye maximális értéket vesz fel és ezt a $\hat{\Theta}$ értéket az ismeretlen Θ paraméter becslésének fogadjuk el.

Az $L = L(x_1, x_2, \dots, x_n; \Theta)$ függvény maximumának meghatározása céljából az a differenciál-számításból ismeretes a

$$\frac{\partial L}{\partial \Theta} = 0$$

egyenlet megoldásait kell vizsgálnunk.

Mivel $\ln L$ ugyanott veszi fel maximumát, mint az L likelihood függvény, vizsgálhatjuk a

$$\frac{\partial \ln L}{\partial \Theta} = 0$$

egyenlet megoldásait, ez rendszerint egyszerűbb számolást igényel. Ezen egyenlet valamennyi gyöke a mintaelemeknek egy

$$\hat{\Theta} = \hat{\Theta}(\xi_1, \xi_2, \dots, \xi_n)$$

függvénye.

Ha egy $\hat{\Theta}$ gyöknek a likelihood függvény maximuma felel meg, akkor $\hat{\Theta}$ -t a Θ paraméter maximum-likelihood becslésének nevezzük.

Mutatunk egy példát a maximum-likelihood módszer alkalmazására.

Tegyük fel, hogy ξ Poisson-eloszlású valószínűségi változó ismeretlen λ paraméterrel, amelyet a $\xi_1, \xi_2, \dots, \xi_n$ mintaértékekből akarunk becsülni.

A $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek mindegyikének elvileg lehetséges értékei a 0, 1, 2, ..., végtelen sorozat elemei, a mintában azonban csak véges értékek vannak képviselve. Jelölje r a legnagyobb megfigyelt értéket és legyenek a 0, 1, 2, ..., r számok gyakoriságai $k_0, k_1, \dots, k_r$, ahol természetesen $\sum_{i=0}^r k_i = n$, hiszen n elemű mintánk van. Annak valószínűsége, hogy valamelyik mintaelem az i értéket vegye fel

$$P(\xi_j = i) = \frac{\lambda^i}{i!} e^{-\lambda}$$

A likelihood függvény tehát:

$$L = \prod_{i=0}^r \left(\frac{\lambda^i}{i!} e^{-\lambda} \right)^{k_i}$$

$$\ln L = \sum_{i=0}^r k_i \ln \left(\frac{\lambda^i}{i!} e^{-\lambda} \right) = \sum_{i=1}^r k_i \ln \frac{\lambda^i}{i!} - n\lambda$$

figyelembe véve, hogy $\sum_{i=0}^r k_i = n$.

A λ paraméter szerinti deriváltat képezve,

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{i=1}^r k_i \frac{i!}{\lambda^i} \cdot \frac{\lambda^{i-1}}{(i-1)!} - n = \frac{1}{\lambda} \sum_{i=1}^r i k_i - n$$

A kapott eredményből:

$$\lambda = \frac{1}{n} \sum_{i=1}^r i k_i = \frac{1}{n} \sum_{j=1}^n \xi_j = \bar{\xi}$$

A maximum-likelihood módszer tehát a λ paraméter becslésként a $\bar{\xi}$ mintaközepet szolgáltatja. Mint tudjuk* $M(\bar{\xi}) = \lambda$ vagyis a becslés torzítatlan.

Vizsgáljuk most meg, milyen statisztikához vezet a maximum-likelihood módszer valamely A esemény ismeretlen p valószínűségének becslésekor.

Tegyük fel, hogy a p valószínűségű A esemény n kísérlet során k-szor következett be.

A likelihood-függvény most:

$$L = p^k (1 - p)^{n-k}$$

$$\log L = k \log p + (n - k) \log (1 - p)$$

A p paraméter szerint deriválunk és a deriváltat 0-sal tesszük egyenlővé:

$$\frac{\partial \log L}{\partial p} = \frac{k}{p} - \frac{n-k}{1-p} = \frac{k - kp - np + kp}{p(1-p)} = \frac{k - np}{p(1-p)} = 0$$

Tehát

$$k - np = 0$$

összefüggésre jutunk, amelyből

$$\hat{p} = \frac{k}{n}$$

tehát az ismeretlen p valószínűség becslésére a maximum-likelihood módszer a relatív gyakorisághoz vezetett, amely ugyancsak torzítatlan becslés.

Abban az esetben, ha a ξ valószínűségi változó eloszlása több ismeretlen paramétert tartalmaz, azaz sűrűségfüggvénye $f(x; \Theta_1, \Theta_2)$ alakban

írható, akkor a Θ_1 és Θ_2 paraméterek $\hat{\Theta}_1$ és $\hat{\Theta}_2$ maximum likelihood becsléseit a

$$\frac{\partial \log L}{\partial \Theta_1} = 0 \quad \text{és} \quad \frac{\partial \log L}{\partial \Theta_2} = 0$$

egyenletekből nyerjük.

* Lásd: (5.5.2) formulát!

Minden olyan $(\hat{\theta}_1, \hat{\theta}_2)$ gyökpár, amely mellett L -nek maximuma van, a $\hat{\theta}_1$ és $\hat{\theta}_2$ paraméterek együttes maximum-likelihood becslése.

Legyen pl. a ξ valószínűségi változó normális eloszlású ismeretlen m és σ paraméterekkel.

Tegyük fel, hogy a kapott mintaértékek az

$$x_1, x_2, \dots, x_n$$

számok.

A likelihood-függvény

$$L = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - m)^2}{2\sigma^2}} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^n \frac{(x_i - m)^2}{2\sigma^2}}$$

A paraméterek szerinti deriváltakat kiszámítva

$$\frac{\partial \ln L}{\partial m} = \sum_{i=1}^n \frac{x_i - m}{\sigma^2} = 0$$

$$\frac{\partial \ln L}{\partial \sigma} = \sum_{i=1}^n \left(-\frac{1}{\sigma} + \frac{(x_i - m)^2}{\sigma^3} \right) = 0$$

egyenletekből,

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

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = S_n^2$$

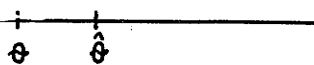
adódik. A maximum-likelihood módszer tehát a várható érték becsléseként a mintaközepet, a szórásnégyzet becsléseként az empirikus szórásnégyzetet szolgáltatja. Az utóbbi becslés, mint tudjuk, torzított, jöllehet a torzítás nagy n esetén nem jelentős.

Megjegyezzük, hogy a maximum-likelihood módszerrel meglehetősen általános feltételek mellett konzisztens és aszimptotikusan normális elosz-

lásu becsléseket nyerünk, amelyek azonban, mint láttuk, nem mindig torzítatlanok. Ki lehet mutatni azt is, hogy a maximum likelihood-becslések között a legkisebb szórásnégyzetűek, és amennyiben valamely paraméter becslésére létezik elégséges statisztika, a mintaelemekben rejlő információt legjobban kimerítő statisztika, akkor a maximum-likelihood módszer az illető paraméter becsléseként ilyen statisztikához vezet. Ezek azok a tulajdonságok, amelyek a maximum-likelihood módszert a többi becslési módszer közül kitüntetik.

8.3 A becslés megbízhatósága

A becslésről szóló eddigi fejtegetéseink során az eloszlás valamely ismeretlen Θ paraméterét egyetlen mennyiséggel, a mintaelemekből konstruált egyetlen $\hat{\Theta}$ számadattal becsültük. A becslésnek ezt a módját szokás pontbecslésnek is nevezni, mivel a Θ paraméter valódi értéke a számegeyes valamely pontja, s ezt a pontot a mintából számitott $\hat{\Theta}$ ponttal igyekszünk "eltalálni", vagy legalábbis jól közelíteni (59. ábra).



59. ábra

A $\hat{\Theta}$ statisztika - melyről tegyük fel, hogy torzítatlan becslés-valószínűségi változó, amelynek értékei szóródnak a becsült Θ paraméter valódi értéke körül. A gyakorlati munkában, amikor egy n elemű minta alapján valamely $\hat{\Theta} = \hat{\Theta}(\xi_1, \xi_2, \dots, \xi_n)$ statisztikai függvénnyel becsüljük a Θ paramétert, tájékozódni kívánunk becslésünk megbízhatóságáról, tudni akarjuk, milyen közel van $\hat{\Theta}$ a becsült Θ értékhez. Hogyan tudhatnánk ezt meg, kérdezheti valaki, amikor Θ igazi értékét nem ismerjük és mindössze egyetlen n elemű mintával rendelkezünk.

Valószínűségelméleti megfontolások segítségével meg tudunk adni olyan σ hibahatárt, hogy a $\hat{\Theta}$ becslés a Θ paraméter értékétől nagy valószínűséggel σ -nál kevesebbel tér el.

Egy példán vázoljuk, mit jelent ez az állítás.

Tegyük fel, hogy egy normális eloszlású ξ valószínűségi változó ismeretlen m várható értékét becsüljük a mintaközéppel, amely a várható érték torzítatlan becslése. Feltesszük, hogy az eloszlás szórását ismerjük: $\sigma = 4$.

ξ -re vonatkozólag 100 elemű mintát veszünk

$$\xi_1, \xi_2, \dots, \xi_{100}$$

és kiszámítjuk a $\bar{\xi}$ mintaközepet.

Tudjuk, hogy $\bar{\xi}$ ugyancsak normális eloszlású valószínűségi változó, m várható értékkel és $\sigma_{\bar{\xi}} = \frac{\sigma}{\sqrt{100}} = 0,4$ szórással.

A normális eloszlás tárgyalásánál mondtuk szerint annak valószínűsége, hogy $|\bar{\xi} - m| < 2 \frac{\sigma}{\sqrt{n}}$ legyen, 0,95-tel egyenlő, azaz

$$P\left(m - 2 \frac{\sigma}{\sqrt{n}} < \bar{\xi} < m + 2 \frac{\sigma}{\sqrt{n}}\right) = P(m - 0,8 < \bar{\xi} < m + 0,8) = 0,95$$

$$A \quad |\bar{\xi} - m| < 2 \frac{\sigma}{\sqrt{n}} \text{ egyenlőtlenség azonban}$$

$$\bar{\xi} - 2 \frac{\sigma}{\sqrt{n}} < m < \bar{\xi} + 2 \frac{\sigma}{\sqrt{n}}$$

alakban is írható, tehát

$$P\left(\bar{\xi} - 2 \frac{\sigma}{\sqrt{n}} < m < \bar{\xi} + 2 \frac{\sigma}{\sqrt{n}}\right) \approx 0,95$$

Ez azt jelenti, hogy 0,95 annak valószínűsége, hogy

$$\left(\bar{\xi} - 2 \frac{\sigma}{\sqrt{n}}, \bar{\xi} + 2 \frac{\sigma}{\sqrt{n}}\right)$$

intervallum tartalmazza az ismeretlen m paramétert. Ebben az esetben természetesen $\bar{\xi}$, azaz az intervallum középpontja, vagyis az intervallum helyzete a valószínűségi változó, az m paraméter rögzített szám. A fenti kijelentés azt jelenti, hogyha $\bar{\xi}$ -re vonatkozólag sok 100 elemű mintát veszünk, mindegyikből kiszámítjuk a $\bar{\xi}$ mintaközepet, s mindegyik mintaközép mint középpont köré $2 \frac{\sigma}{\sqrt{n}}$ sugarú intervallumot rajzolunk, akkor

ezen intervallumok 95%-a olyan, hogy tartalmazza az ismeretlen m paramétert.

$$\begin{aligned}
 (1) & \xi_{11}, \xi_{12}, \dots, \xi_{1n}; \bar{\xi}^{(1)} \\
 (2) & \xi_{21}, \xi_{22}, \dots, \xi_{2n}; \bar{\xi}^{(2)} \\
 & \vdots \\
 (N) & \xi_{N1}, \xi_{N2}, \dots, \xi_{N,n}; \bar{\xi}^{(N)}
 \end{aligned}$$

(A nyert intervallumokat áttekinthetőség céljából egymás fölé rajzoltuk.)

A gyakorlatban egyetlen n elemű mintánk van,

$$\xi_1, \xi_2, \dots, \xi_n$$

Normális eloszlás esetében mint látjuk, megadható egy olyan $\sigma = \frac{2\sigma}{\sqrt{n}}$ hi-

bahatár, hogy a mintából számított $\bar{\xi}$ mintaközép 0,95%-os valószínűséggel σ -nál kevesebbel tér az ismeretlen m paramétertől.

$$A \left(\bar{\xi} - \frac{2\sigma}{\sqrt{n}}, \bar{\xi} + \frac{2\sigma}{\sqrt{n}} \right) \text{ vé-}$$

letlen helyzetű intervallumot 95%-os konfidencia (megbízhatósági) intervallumnak nevezzük az m várható értékre vonatkozólag.

Ilyen intervallum grafikus megszerkesztését $\frac{\sigma}{\sqrt{n}} = 1$ esetén

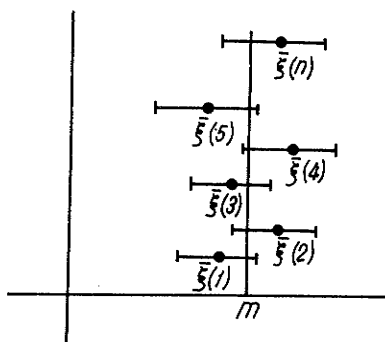
a 61. ábra szemlélteti.

Abban az esetben, ha $\xi \sim N(m; \sigma)$ normális eloszlású és mind m mind σ ismeretlenek, tekintjük a

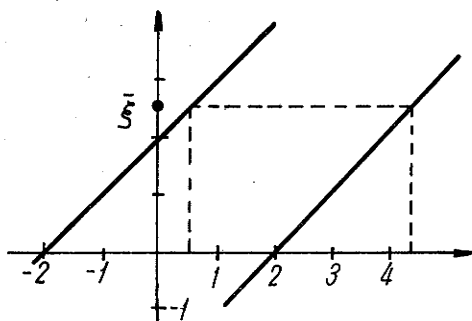
$$t = \sqrt{n-1} \frac{\bar{\xi} - m}{S_n}$$

valószínűségi változót, amely $n-1$ szabadságfokú Student-eloszlású.

Itt az előző esettől lényegében csak annyi az eltérés, hogy σ^2 -et az S_n^2 korrigált empirikus szórásnégyzettel helyettesítettük.



60. ábra



61. ábra

A III. táblázatból meghatározható olyan t_0 érték, hogy annak valószínűsége, hogy t értéke a $(-t_0, t_0)$ intervallumba essék, 0,95 legyen, azaz teljesüljön

$$P(-t_0 < \sqrt{n-1} \frac{\bar{\xi} - m}{S_n} < t_0) \approx 0,95$$

reláció, vagy ami ezzel ekvivalens, fennálljon a

$$P(\bar{\xi} - t_0 \frac{S_n}{\sqrt{n-1}} < m < \bar{\xi} + t_0 \frac{S_n}{\sqrt{n-1}}) = 0,95$$

összefüggés.

Ez esetben a

$$(\bar{\xi} - t_0 \frac{S_n}{\sqrt{n-1}}, \bar{\xi} + t_0 \frac{S_n}{\sqrt{n-1}})$$

intervallum 95%-os konfidencia intervallum az ismeretlen m várható értéke.

A konfidencia intervallum alkalmazását szokás intervallumbecslésnek is nevezni. Mint említettük, amikor a Θ paraméterpontot egy $\hat{\Theta}$ statisztikával becsüljük és a kiszámított $\hat{\Theta}$ pontot is felvisszük a számegyenesre, akkor szinte biztos, hogy $\hat{\Theta}$ többé-kevésbé eltér Θ -tól, telitalálat igen ritka $\hat{\Theta}$ szórása miatt. Azonban $\hat{\Theta}$ eloszlását ismerve, meg tudunk adni olyan δ -t, hogy a $\hat{\Theta}$ középpontú 2δ szélességű intervallum szinte biztosan el fogja fedni az ismeretlen Θ paramétert.

STATISZTIKAI HIPOTÉZISEK VIZSGÁLATA

9.1 Elvi megjegyzések

Az előzőekben már említettük, hogy a gyakorlatban sokszor előre tudjuk, hogy milyen típusu eloszlással állunk szemben. Ha a kísérlet kimenetelét sok véletlen tényező befolyásolja és ezek a tényezők külön-külön csekély mértékben hatnak, de hatásuk összegeződik, akkor általában normális eloszlás lép fel, mint arról a centrális határeloszlástétel kapcsán szoltunk. Amikor a kísérletnek csak kétféle kimenetele lehet, más szóval egyszerű alternatívával van dolgunk, s a kísérletet többször megismételjük, akkor általában Poisson-eloszlás lép fel, valamely jel megjelenéséig eltelt várakozási idő pedig exponenciális eloszlású stb.

Igen gyakran azt tudjuk, hogy a vizsgált ξ valószínűségi változó sűrűségfüggvénye egy adott $f(x; \Theta)$ függvény, és ezzel kapcsolatban csak a Θ paraméter értékét nem ismerjük pontosan. Sokszor fordul elő az is, hogy a Θ értékét már meghatároztuk, s azt Θ_0 -nak találtuk, a kísérlet egy további szakaszában pedig arra vagyunk kíváncsiak, vajon nem változott-e meg a Θ értéke, vagyis fennáll-e, hogy $\Theta = \Theta_0$. Ilyenkor felállítjuk a $H_0: \Theta = \Theta_0$ ugynevezett null-hipotézist, és a ξ változóra vonatkozó megfigyelési adatok alapján ellenőrizzük, hogy elfogadható-e ez a feltevésünk vagy sem.

Ha a $H_0: \Theta = \Theta_0$ hipotézis elvetése mellett döntöttünk, akkor ezzel máris a $H_1: \Theta \neq \Theta_0$ ellenhipotézist vagy alternatív feltevést fogadtuk el.

Előfordulhat, hogy gyakorlati tapasztalat alapján biztosak lehetünk arról, hogy ha $\Theta = \Theta_0$ nem áll fenn, akkor az értéke csakis valamely (Θ_1, Θ_2) intervallumba eshet, tehát $H_1: \Theta \in (\Theta_1, \Theta_2)$, sőt az is lehetséges, hogy $\Theta \neq \Theta_0$ esetén csakis $\Theta = \Theta_1$ lehet, és ekkor $H_1: \Theta = \Theta_1$.

Az eloszlás általában több paramétert tartalmaz. Tegyük fel, hogy a ξ sűrűségfüggvénye $f(x; \Theta_1, \Theta_2, \dots, \Theta_k)$ alakú. Ha a H_0 hipotézisben valamennyi paraméter értékét specifikáljuk: $H_0: \Theta_1 = \alpha_1, \Theta_2 = \alpha_2, \dots, \Theta_k = \alpha_k$, akkor azt mondjuk, hogy ez egyszerű hipotézis, ha pedig valamelyik paraméterre több lehetséges értéket engedünk meg (pl. azt

tételezzük fel róla, hogy az értéke valamely intervallumba esik), vagy pedig semmit sem teszünk fel némelyik paraméterről, olyan esetben összetett hipotézisről beszélünk.

Ha pl. tudjuk, hogy ξ normális $N(m; \sigma)$ eloszlású, akkor a

$$H_0 : m = 0, \sigma = 1 \quad \text{egyszerű, a}$$

$$H_0 : m = 0 \quad \text{pedig összetett}$$

hipotézis, mert ekkor a σ paraméter értéke a $(0, +\infty)$ intervallum bármely pontja lehet.

A gyakorlatban igen sokszor a H_0 feltevés arra vonatkozik, hogy két valószínűségi változó, a ξ és a η azonos eloszlásúak. Jelöljük a ξ eloszlásfüggvényét $F(x)$ -szel, az η eloszlásfüggvényét pedig $G(x)$ -szel. A $H_0 : F(x) \equiv G(x)$ nyilván összetett hipotézis, hiszen a paraméterek a számokra lehetséges értékek bármelyikét felvehetik, csupán azzal a feltevessel élünk, hogy mindkét eloszlásban ugyanazt az értéket veszik fel.

Meglehet az is, hogy még az eloszlásfüggvény típusában sem vagyunk biztosak, legfeljebb azt gyanítjuk, hogy milyen eloszlással állunk szemben. Ilyenkor a H_0 feltevés abból áll, hogy a ξ eloszlásnak adott $F(x)$ eloszlásfüggvénye vagy $f(x)$ sűrűségfüggvénye van. Ez esetben ellenőrizzük, hogy a ξ -re vonatkozó megfigyelési eredmények mennyire vannak összhangban a H_0 hipotézissel, a kapott értékek mennyire "illeszkednek" az $F(x)$ eloszlásfüggvényhez, illetve az $f(x)$ sűrűségfüggvényhez, vagyis un. illeszkedésvizsgálatot végzünk. Ha az $F(x)$ eloszlást paramétereivel együtt, tehát a maga teljességében tételezzük fel, akkor tiszta illeszkedésvizsgálatról, ha pedig az eloszlás paramétereit a mintából kell becsülnünk, becsléses illeszkedésvizsgálatról beszélünk. Az illeszkedésvizsgálat leggyakoribb módja az un. normalitásvizsgálat, amelyről részletesebben a 10.3 pontban szólnunk.

Statisztikai hipotézisen mindig egy ξ valószínűségi változó $f(x; \theta)$ eloszlására vonatkozó valamilyen feltevést értjük. E hipotézis ellenőrzésére szolgáló eljárást statisztikai próbának, másnéven tesztnek nevezzük. A próba célja, hogy dönteni tudjunk arról: elfogadjuk-e vagy pedig elvetjük a szóbanforgó feltevést.

Ezzel kapcsolatban megvizsgáljuk, hogy mi a különféle statisztikai próbák közös vonása, mi a próba lényege?

A hipotézis ellenőrzéséhez, tehát a próba elvégzéséhez alapul a ξ valószínűségi változóra vonatkozó n számú megfigyelés, más szóval a

$$\xi_1, \xi_2, \dots, \xi_n$$

adatokból álló minta szolgál. Mint tudjuk, a mintaelemek valószínűségi változót, tehát a minta egy $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ n-dimenziós véletlen vektorként fogható fel, vagyis a minta egy véletlen pont az n-dimenziós térben.

Az érdekel bennünket, hogy a minta, tehát a $(\xi_1, \xi_2, \dots, \xi_n)$ véletlen pont helyzete igazolja-e feltevésünk helyességét, vagy ellentmond-e annak. Egyszerűen oly módon járunk el, hogy az összes lehetséges

$$\xi_1, \xi_2, \dots, \xi_n$$

minták terében az ún. mintatérben az n-dimenzió térben, vagy annak bizonyos részhalmazában, valamilyen megfontolás alapján kijelölünk egy K tartományt, amelyet kritikus tartománynak nevezünk. Ha az észlelt

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

mintapont a K tartományba esik, akkor elvetjük a H_0 feltevést, ha pedig a K tartományon kívül esik, akkor elfogadjuk.

Lássunk szemléletesség kedvéért egy egyszerű példát. El akarjuk dönteni, hogy bizonyos gyártmány-tömeg, pl. valamilyen egyforma alkatrészekből álló tétel selejtaránya bizonyos korlát alatt marad vagy sem. Evégből n számu gyártmányt minőségellenőrzési vizsgálatnak vetünk alá, s ha selejtes darabok g relatív gyakorisága nem nagyobb egy bizonyos p_0 ($0 < p_0 < 1$) számnál, akkor

azt mondjuk, hogy a tétel elfogadható. Minden ellenőrzött darab vizsgálati eredménye egy ξ valószínűségi változó, amely vegye fel a 0 értéket, ha adott gyártmány megfelelő és az 1 értéket, ha selejtes. Itt most az a H_0 hipotézis, hogy az összes szóban forgó gyártmányok selejtaránya $p < p_0$.

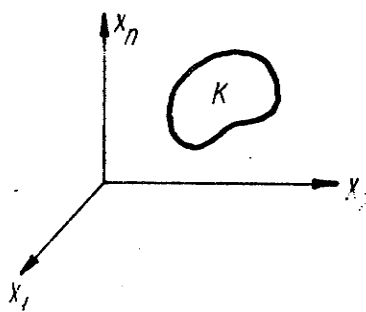
A

$$\xi_1, \xi_2, \dots, \xi_n$$

minta most a 0 és az 1 jegyekből

$$0, 1, \dots, 1$$

álló sorozat.



62. ábra

Mivel a $\xi_1, \xi_2, \dots, \xi_n$ változó mindegyike a 0 vagy az 1 értéket egyaránt felveheti, az összes lehetséges minták száma 2^n , ennyi lévén a 0 és az 1 jegyek segítségével képezhető n elemű sorozatok száma. Minden ilyen sorozatot az n -dimenziós tér egy pontjának tekintve, a mintatér most 2^n számú pontból áll.

Tegyük fel, hogy k a legnagyobb olyan egész szám, amelyre $\frac{k}{n} < p_0$ de $\frac{k+1}{n} \geq p_0$. Ekkor a K kritikus tartomány mindazokból a

$\xi_1, \xi_2, \dots, \xi_n$ mintapontokból áll, amelyekben legalább $k+1$ számú 1-es található, mert minden ilyen mintára selejtes darabok relatív gyakorisága a p_0 -nál nagyobb.

Az n -dimenziós tér e pontjainak halmazát, a K kritikus tartomány természetesen nem tudja szemléltetni, még elképzelni sem, hiszen szemléletünk nem terjed túl a háromdimenziós téren, amelynek fizikai realitása van, de erre nincs is szükségünk. Mindazokon a mintákon, amelyekben legalább $k+1$ darab 1-es szerepel, fennáll a

$$\bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} \geq p_0$$

összefüggés, hiszen itt a mintaközép a relatív gyakoriságnak felel meg.

Valahányszor a $\bar{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$ mintapont a K kritikus tartományba esik, az adott mintából számított $\bar{\xi}$ mintaközép a számegyenes p_0 pontjától jobbra van. Annak a valószínűsége tehát, hogy a $\bar{\xi}$ valószínűségi változóra vonatkozó mintapont a K kritikus tartományba essék, megegyezik annak a valószínűségével, hogy a számegyenes $(p_0, +\infty)$ intervallumában van a $\bar{\xi}$ mintaközép. Azt is mondhatjuk tehát, hogy a $H_0 = \bar{\xi} < p_0$ feltevés ellenőrzése céljából kritikus tartománynak a számegyenes $x = p_0$ ponttól eső részét, vagyis a $(p_0, +\infty)$ intervallumot választjuk.

A gyakorlatban mindig hasonló módon járunk el. Valahányszor egy n elemű mintánk van, az összes lehetséges mintapontok tere, a mintatér, az n -dimenziós tér, vagy annak valamely részhalmaza. A kritikus tartományt gyakorlatilag rendkívül nehézkes lenne az n -dimenziós tér részhalmazaként megválasztani, ehelyett vesszük a mintaelemeknek valamely

$$u = u(\xi_1, \xi_2, \dots, \xi_n)$$

statisztikai függvényét, amelynek értékei a számegyenesen vannak. Azt is mondhatjuk, hogy az u függvény a mintatér egy

$\vec{\xi}_0 = (\xi_1^{(0)}, \xi_2^{(0)}, \dots, \xi_n^{(0)})$ pontjának a számegyenesre egy

$u_0 = u_0(\xi_1^{(0)}, \xi_2^{(0)}, \dots, \xi_n^{(0)})$ pontját felelteti meg.

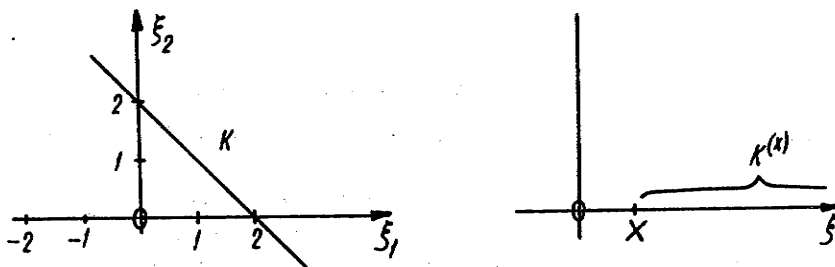
Kritikus tartománynak ekkor a számegyenes alkalmas részhalmazát választjuk, vagyis a mintatér K kritikus tartománya helyett a számegyenes $K^{(x)}$ kritikus tartományával dolgozunk. Azt mondjuk, hogy legyen a K kritikus tartomány mindazoknak a

$$\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_n)$$

mintapontoknak a halmaza, amelyekre az

$$u = u(\xi_1, \xi_2, \dots, \xi_n)$$

függvény értéke a számegyenesen megválasztott $K^{(x)}$ halmazába esik.



63. ábra

A számegyenes $K^{(x)}$ kritikus tartományának megválasztása után a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemekből kiszámítjuk az $u = u(\xi_1, \xi_2, \dots, \xi_n)$ függvény értékét, s ha az u értéke a $K^{(x)}$ kritikus tartományába esik, akkor elvetjük H_0 -t, ha pedig u értéke nem a kritikus tartományba esik, akkor elfogadjuk H_0 feltevését. A kritikus tartomány megválasztása után tehát a minta, illetve a mintából számított u függvény értéke egyértelműen eldönti a H_0 hipotézis sorsát. Valamely statisztikai próba megválasztása tehát lényegében a kritikus tartomány megválasztását jelenti.

A kritikus tartomány megválasztásának módját, s a választás következményeit néhány konkrét statisztikai próba tárgyalása során ismertetjük.

9.2 Az u-próba

Tételezzük fel, hogy a ξ normális eloszlású ismert $\sigma = \sigma_0$ szórással. Felállítjuk azt a H_0 hipotézist, hogy a ξ várható értéke $m = m_0$. Ezek után az a kérdés, hogy elfogadjuk-e a

$$H_0 : M(\xi) = m_0$$

feltevést a

$$H_1 : M(\xi) \neq m_0$$

alternatív feltevással szemben.

Tekintsük a ξ -re vonatkozó n számú független megfigyelésből álló $\xi_1, \xi_2, \dots, \xi_n$ mintát, és alkossuk meg a következő statisztikát:

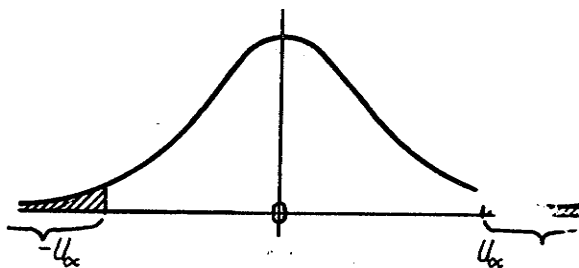
$$u = \frac{\bar{\xi} - m_0}{\sigma_0} \sqrt{n}$$

Ha fennáll a H_0 hipotézis, vagyis a ξ várható értéke valóban m_0 , akkor tudvalevően a várható értéke is m_0 , szórása $\frac{\sigma_0}{\sqrt{n}}$, ezért az u

statisztika standard normális eloszlású, tehát u eloszlása $N(0, 1)$. Ekkor egy megadott valószínűségi szinthez normális eloszlás táblázatából meghatározható $1 - \alpha$ (pl. $\alpha = 0,05$ esetén 95%-os) szintű konfidencia-intervallum, vagyis

$$P(-u_\alpha < u_0 < u_\alpha | H_0) = 1 - \alpha$$

Ilyenformán tehát, ha felteesszük, hogy a H_0 igaz, akkor $1 - \alpha$ annak a valószínűsége, hogy az u értéke a $(-u_\alpha, u_\alpha)$ intervallumba esik, annak a valószínűsége pedig, hogy az u valószínűségi változó értéke kívül van ezen az intervallumon, α -val egyenlő.



64. ábra

Az u - próbánál már a mintavétel előtt megadjuk az α valószínűségi szintet és megválasztjuk az ezen szinthez tartozó u_α értéket. Ha az u valószínűségi változó értéke a $(-u_\alpha, u_\alpha)$ intervallumon kívül esik, akkor a H_0 hipotézist elvetjük és a H_1 -et fogadjuk el. Ezt a döntésünket arra alapozzuk, hogyha a H_0 valóban igaz, akkor annak a valószínűsége, hogy az u változó értéke a $(-u_\alpha, u_\alpha)$ intervallumon kívül essék, mindössze α , amelyet rendszerint kicsire választunk. Ennélfogva tehát, ha az észlelt értéke mégis a fenti intervallumon kívül esett, akkor olyan kis valószínűségi esemény következett be, mely nem erősíti meg feltevésünket.

Ebben a próbában a számegyenesnek a $-u_\alpha$ számtól balra és az u_α számtól jobbra eső részét választottuk kritikus tartománynak, a $(-u_\alpha, u_\alpha)$ intervallumot pedig elfogadási tartománynak nevezzük. Konkrét esetben legyen pl. $\alpha = 0,05$. Ekkor az u valószínűségi változóra, mivel $N(0,1)$ eloszlású a $(-2,2)$ intervallum 95%-os konfidencia-intervallum, a $(-u_\alpha, u_\alpha) = (-2,2)$ tehát a kritikus tartomány és a $(-2,2)$ intervallum kiegészítő halmaza a számegyenesen.

Előfordulhat, hogy a H_0 igaz, azaz $M(\xi) = m_0$ és az u változó értéke a kritikus tartományba esik, vagyis a H_0 -t elvetjük. Ebben az esetben tehát hibás döntést hozunk. Az olyan hibát, amely onnan áll elő, hogy a helytálló hipotézist elvetettük, elsőfajú hibának nevezzük. Az elsőfajú hiba elkövetésének valószínűsége

$$P(|u| > u_\alpha | H_0) = \alpha$$

és az esetünkben 5%.

A mondottak szerint megválasztott kritikus tartomány tehát igen nagy valószínűséggel biztosítja, hogy a helyes feltevést nem fogjuk elvetni, nevezetesen 100 esetből átlagosan 5 ízben tévedünk.

Mint mondtuk, a kritikus tartomány megválasztása után teljesen mechanikusan járunk el, ugyanis ha az $u = u(\xi_1, \xi_2, \dots, \xi_n)$ értéke a kritikus tartományba esik, elvetjük a H_0 -t, ha pedig az elfogadási tartományba esik, akkor elfogadjuk. Így tehát a H_0 -t akkor is elfogadjuk, ha nem igaz (amit persze nem tudunk róla), de ugyanakkor az u értéke az elfogadási tartományba esik. Megeshet, hogy a ξ várható értéke nem az m_0 szám, hanem egy másik m_1 érték:

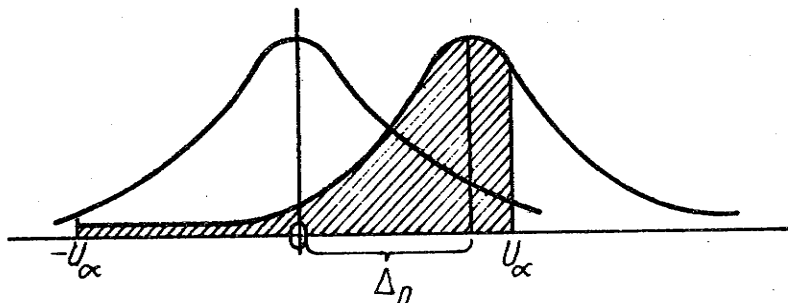
$$M(\xi) = m_1 \neq m_0$$

vagyis a H_1 feltevés igaz, s az u változó értéke mégis az $(-u_\alpha, u_\alpha)$ intervallumba esik. Ez esetben a H_0 -t hibásan fogadjuk el. Ezt a hibát másodfajú hibának nevezzük.

Vizsgáljuk meg, hogy mekkora a másodfajú hiba valószínűsége? Feltéve, hogy $M(\xi) = m_1$, az u valószínűségi változó már nem 0 várható értékű, hiszen ekkor $M(\bar{\xi}) = m_1$ és

$$M\left(\frac{\bar{\xi} - m_0}{\frac{\sigma_0}{\sqrt{n}}}\right) = \frac{m_1 - m_0}{\sigma_0} \sqrt{n} = \Delta_n$$

Adott $M(\xi) = m_1$ esetében a másodfajú hiba elkövetésének valószínűségét a 65. ábra vonalkázott része mutatja:



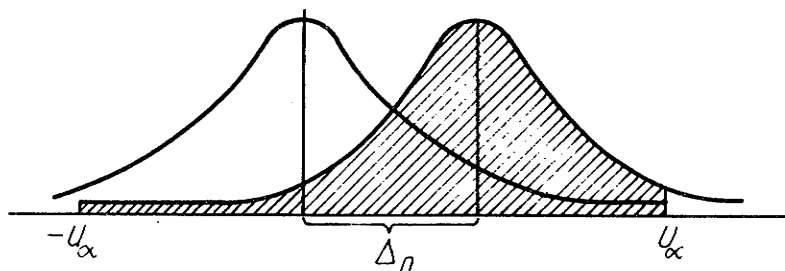
65. ábra

Jelöljük másodfajú hiba elkövetésének valószínűségét β -val.

$$\beta = P(-u_\alpha < u < u_\alpha | m_1)$$

A másodfajú hiba valószínűsége nyilvánvalóan függ az m_1 -től, vagyis attól, hogy mekkora a $\bar{x}$ változó várható értéke. Az ábra alapján könnyű belátni, hogyha az m_1 értéke közel esik az m_0 -hoz, akkor nagy valószínűséggel elfogadjuk a H_0 -t, holott a H_1 igaz, más szóval másodfajú hibát követünk el. Kedvező körülmény azonban, hogyha az m_1 közel esik az m_0 -hoz, akkor a H_0 feltevés majdnem teljesül, s ezért az elfogadása a gyakorlatban nem okoz nagy kárt.

Láttuk tehát, hogy a kritikus tartomány fenti módon történő megválasztásával biztosítani tudjuk, hogy az elsőfajú hiba valószínűsége gyakorlati szempontból kielégítő biztonságu α legyen, ahol az α -t tetszőleges kicsiny értékűnek (általában 0,05-nek) választhatjuk. Kellemetlen, hogy minél kisebbre választjuk az α -t, általában annál nagyobb lesz a β , a másodfajú hiba valószínűsége, amint az a 66. ábrából is kitűnik.



66. ábra

Másrészről azonban az is igaz, hogyha a hipotézis helyes elvetését illetően kisebb biztonsággal is megelégszünk (tehát az α értékét nagyobbra választjuk), akkor általában csökken a másodfajú hiba. A nagyobb biztonság tehát ára van, a kisebb elsőfajú hibát nagyobb másodfajú hibával kell megfizetnünk.

Amikor a kritikus tartományt oly módon választjuk meg, hogy az elsőfajú hiba valószínűsége α legyen, olyankor a próbát α terjedelműnek nevezzük. A probléma konkrét természetétől függ, hogy egy adott H_0 feltevés vizsgálatára mekkora terjedelmű próbát alkalmazunk, nevezetesen azt kell meggondolnunk, hogy az első-, illetve a másodfajú hiba elkövetése milyen terjedelmű szakmai vagy anyagi kárt okoz.

9.3 Az erőfüggvény

Az előzőekben láttuk, hogy a másodfajú hiba β valószínűségét nem lehet egyszerűen meghatározni, mert ez a valószínűség

$$\beta = P(-u_\alpha < u < u_\alpha \mid H_1)$$

attól függ, hogy milyen H_1 alternatív hipotézis igaz, más szóval az u -próba esetében mekkora az m várható érték. A másodfajú hiba helyett a kiegészítő esemény valószínűségét szoktuk vizsgálni, vagyis annak a valószínűségét, hogy az $u = u(\xi_1, \xi_2, \dots, \xi_n)$ statisztika értéke a kritikus tartományba esik, feltéve, hogy a H_1 hipotézis igaz. Ez a valószínűség nyilvánvalóan

$$1 - \beta = P(|u| > u_\alpha \mid H_1) = P(|u| > u_\alpha \mid m)$$

A $P(|u| > u_\alpha \mid m)$ valószínűséget, mint m függvényét a próba erőfüggvényé-nek nevezzük. Az erőfüggvény m minden értékére megadja annak a feltételes valószínűségét, hogy az u -statisztika a kritikus tartományba esik, tehát a H_0 -t elutasítjuk, feltéve, hogy m a tényleges várható érték.

Jelöljük az erőfüggvényt $E(m)$ -mel. Ezek szerint tehát:

$$(9.3.1) \quad E(m) = 1 - \beta = P(|u| > u_\alpha \mid m)$$

Könnyen belátható, hogyha α terjedelmű próbát alkalmazunk, akkor

$$E(m_0) = \alpha$$

hiszen annak valószínűsége, hogy H_0 -t elutasítjuk, feltéve, hogy a tényleges várható érték m_0 , vagyis H_0 igaz, éppen az elsőfajú hiba valószínűségével egyenlő.

Valamely próba alkalmazásakor természetesen az lenne az ideális, ha mind az elsőfajú hiba α valószínűsége, mind a másodfajú hiba β valószínűsége nagyon kicsiny lenne. Sajnos nincsenek ilyen ideális próbák. Az elsőfajú hiba valószínűségét, a próba terjedelmét α -nak választva nyilvánvalóan az a kívánatos, hogy a β minél kisebb, illetve az $1 - \beta = E(m)$ minél nagyobb legyen, más szóval, hogy minél erősebb próbát alkalmazzunk.

Láttuk, hogy a próba α terjedelmének megválasztása egyben determinálja az erőfüggvény értékét az $m = m_0$ pontban, azaz

$$E(m_0) = \alpha$$

ez pedig általában kicsiny (pl. $\alpha = 0,05$). Mivel az erőfüggvény az m -nek folytonos függvénye, ennél fogva értéke az m_0 -nak nem nagy környezetében általában kicsiny. Gyakorlati szempontból azonban az m_0 -hoz közeli értékek nem jelentenek lényeges eltérést, ezért az erőfüggvény iránt azt a követelményt támasztjuk, hogy értéke valamilyen $(-m_1 < m_0 < m_1)$ környezetén kívül érjen el a konkrét feladatnak megfelelő eléggé nagy értéket.

Az u -próba esetében az erőfüggvényt nem nehéz explicit alakban megadni:

$$E(m) = P(|u| > u_\alpha | m) = 1 - P(-u_\alpha < u < u_\alpha | m) =$$

$$= 1 - P\left(-u_\alpha < \frac{\bar{x} - m_0}{\sigma_0} \sqrt{n} < u_\alpha | m\right).$$

Bevezetve a $\Delta_n = \frac{(m - m_0) \sqrt{n}}{\sigma_0}$ jelölést, $E(m)$ -re az alábbi kifejezést kapjuk:

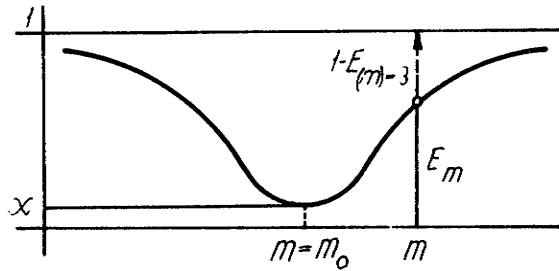
$$E(m) = 1 - P\left(-u_\alpha - \Delta_n < \frac{\bar{x} - m}{\sigma_0} < u_\alpha - \Delta_n | m\right)$$

Mint hogy az $M(\bar{x}) = m$ feltétellel $M(\bar{x}) = m$, azaz $\frac{\bar{x} - m}{\sigma_0} \sqrt{n}$ változó standard normális eloszlású, és ennek az eloszlásfüggvénye

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt \quad \text{az } E(m) \text{ erőfüggvény explicit alakja}$$

$$(9.3.2) \quad E(m) = 1 - \Phi(u_\alpha - \Delta_n) + \Phi(-u_\alpha - \Delta_n)$$

A (9.3.2). képletben szereplő $E(m)$ függvény alakját a 67. ábra szemlélteti.



67. ábra

A (9.3.2) képlet kissé részletesebben kiírva az

$$E(m) = 1 - \Phi\left(u_{\alpha} - \frac{m - m_0}{\sigma_0} \sqrt{n}\right) + \Phi\left(-u_{\alpha} - \frac{m - m_0}{\sigma_0} \sqrt{n}\right)$$

összefüggésből kitűnik, hogy

$$\lim_{m \rightarrow -\infty} E(m) = \lim_{m \rightarrow \infty} E(m) = 1$$

(ugyanis $\Phi(-\infty) = 0$).

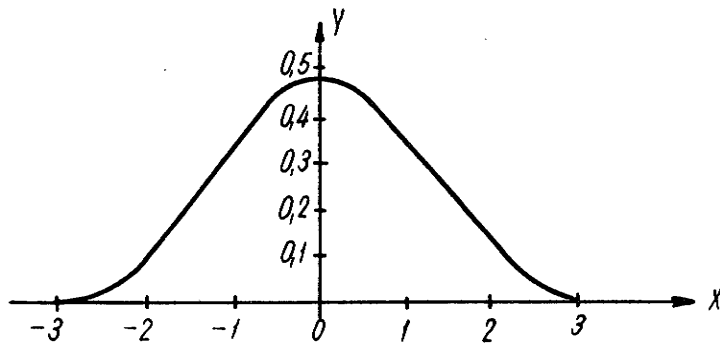
Az $E(m)$ erőfüggvény értékeit ismerve minden egyes m értékre a másodfajú hiba valószínűsége: $\beta = 1 - E(m)$ is ismert.

9.4 A Student-féle t-próba

Legyen $\xi \sim N(m; \sigma)$ normális eloszlású valószínűségi változó, ahol az m és a σ ismeretlenek. Legyen a H_0 hipotézisünk, hogy ξ várható értéke $m = m_0$. Végezzünk a ξ -re n számú megfigyelést, s a

$$\xi_1, \xi_2, \dots, \xi_n$$

mintából számítsuk ki a $t = \sqrt{n-1} \frac{\bar{\xi} - m_0}{S}$ statisztikát, amely un. $n-1$ szabadságfoku Student-eloszlású, s amelynek a sűrűségfüggvényét a 68. ábra tünteti fel.



68. ábra

Jelöljön a t_0 olyan értéket, hogy

$$P(-t_0 < \sqrt{n-1} \frac{\bar{\xi} - m}{S} < t_0) = 0,95$$

legyen, $n-1$ szabadságfoku Student-eloszlás esetén. Ilyen t_0 érték a III. táblázatból közvetlenül kiolvasható.

A

$$-t_0 < \sqrt{n-1} \frac{\bar{\xi} - m}{S} < t_0$$

egyenlőtlenség ekvivalens a

$$\bar{\xi} - t_0 \frac{S}{\sqrt{n-1}} < m < \bar{\xi} + t_0 \frac{S}{\sqrt{n-1}}$$

egyenlőtlenséggel. Ez azt jelenti, hogy

$$P\left(\bar{\xi} - t_0 \frac{S}{\sqrt{n-1}} < m < \bar{\xi} + t_0 \frac{S}{\sqrt{n-1}}\right) = 0,95,$$

azaz 95% annak a valószínűsége, hogy a $\left(\bar{\xi} - t_0 \frac{S}{\sqrt{n-1}}, \bar{\xi} + t_0 \frac{S}{\sqrt{n-1}}\right)$

intervallum tartalmazza az ismeretlen m várható értéket. Más szóval, ha a ξ -re sok n -elemű mintát vennénk,

$$\begin{array}{ccc} \xi_1^{(1)}, & \xi_2^{(1)}, \dots, \xi_n^{(1)} \\ \xi_1^{(2)}, & \xi_2^{(2)}, \dots, \xi_n^{(2)} \\ \vdots & \vdots & \vdots \\ \xi_1^{(N)}, & \xi_2^{(N)}, \dots, \xi_n^{(N)} \end{array}$$

továbbá mindegyik mintából kiszámíthatnánk a $\bar{\xi}^{(i)}$ mintaközepet és az s tapasztalati szórást, majd mindegyik mintaközép mint felezőpont köré felmérnénk a

$$\left(\bar{\xi}^{(i)} - t_0 \frac{s}{\sqrt{n-1}}, \bar{\xi}^{(i)} + t_0 \frac{s}{\sqrt{n-1}} \right)$$

intervallumot, akkor az N számú intervallum 95%-a olyan lenne, amely magába foglalja az ismeretlen m várható értéket.

A

$$\left(\bar{\xi} - t_0 \frac{s}{\sqrt{n-1}}, \bar{\xi} + t_0 \frac{s}{\sqrt{n-1}} \right)$$

intervallumot 95%-os megbízhatósági intervallumnak nevezzük. Ha a feltételezett $m = m_0$ érték nem esik bele ebbe az intervallumba, akkor alávétjük a H_0 hipotézist, mondván hogyha H_0 mégis igaz lenne, akkor csak az esetek 5%-ában fordulna elő, hogy m_0 nincs benne, a fenti módon konstruált intervallumban. Más szavakkal tehát olyan kis valószínűségű esemény következett be, hogy a hipotézis aligha lehet helytálló.

Kétmintás eset

Legyenek ξ és η normális eloszlású, független valószínűségű változók, amelyek egyforma szórásúak, azaz $D(\xi) = D(\eta)$ és vizsgáljuk a $H_0: M(\xi) = M(\eta)$ nullhipotézist.

Legyenek a ξ , ill. η változókra vonatkozó minták:

$$\xi: \xi_1, \xi_2, \dots, \xi_n$$

$$\eta: \eta_1, \eta_2, \dots, \eta_n$$

Képezzük az alábbi statisztikát:

$$t_{n+m-2} = \frac{\bar{\xi} - \bar{\eta}}{\sqrt{(n-1)S_n^2 + (m-1)S_m^2}} \sqrt{\frac{n \cdot m \cdot (n+m-2)}{n+m}}$$

Amennyiben a H_0 hipotézis igaz, vagyis a két valószínűségi változó egyforma várható értékű, akkor t_{n+m-2} statisztika $n + m - 2$ paraméterű Student-eloszlást követ.

Ha ellenhipotézisünk $H_1: M(\xi) \neq M(\eta)$, akkor az egymintás esethez hasonlóan szimmetrikus kritikus tartományt választunk, azaz a Student-eloszlás táblázatból megkeressük azt a t_ε értéket, amelyre

$$P(-t_\varepsilon \leq t_{n+m-2} < t_\varepsilon) = 1 - \varepsilon.$$

Ha a kiszámított t_{n+m-2} statisztikánk nem esik a $(-t_\varepsilon, t_\varepsilon)$ intervallumba, akkor a H_0 hipotézist ε -szinten elvetjük.

9.5 Welch-próba

Két normális eloszlású valószínűségi változó várható értékének összehasonlítása ismeretlen és különböző szórások esetén.

Legyenek a ξ és η valószínűségi változók normális eloszlásúak és vizsgáljuk a $H_0: M(\xi) = M(\eta)$ hipotézist, ha a szórások nem ismeretesek és az F-próba a $D(\xi) \neq D(\eta)$ eredményre vezet. Ez esetben a kétmintás t-próba nem alkalmazható, sőt egzakt próba nem is adható meg. A H_0 hipotézis helyességének eldöntésére Welch a következő statisztikát adta meg:

$$t_f = \frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{S_n^2}{n} + \frac{S_m^2}{m}}}$$

ahol $\bar{\xi}$ és $\bar{\eta}$ a mintaközépek, S_n^2 , ill. S_m^2 a tapasztalati szórásnégyzetek a megfelelő

$$\xi: \xi_1, \xi_2, \dots, \xi_n$$

$$\eta: \eta_1, \eta_2, \dots, \eta_m$$

mintából számítva.

A H_0 hipotézis fennállása esetén a t_f statisztika közelítőleg Student-eloszlású f paraméterre, ahol f értékét a következő módon határozzuk meg:

$$\frac{1}{f} = \frac{1}{m-1} \left[\frac{\frac{S_m^2}{m}}{\frac{S_n^2}{n} + \frac{S_m^2}{m}} \right]^2 + \frac{1}{n-1} \left[\frac{\frac{S_n^2}{n}}{\frac{S_n^2}{n} + \frac{S_m^2}{m}} \right]^2$$

9.6 Az F-próba

Ha azt akarjuk eldönteni, hogy két normális eloszlású valószínűségi változó szórása megegyezik-e vagy sem, akkor az F-próbát alkalmazzuk.

Legyenek ξ és η normális eloszlású valószínűségi változók, amelyek várható értéke különböző lehet, s legyen a H_0 hipotézisünk a szórások megegyezése, azaz:

$$H_0 : D(\xi) = D(\eta)$$

Legyenek a ξ , ill. η változóra vonatkozó minták:

$$\xi : \xi_1, \xi_2, \dots, \xi_n$$

$$\eta : \eta_1, \eta_2, \dots, \eta_m$$

A szórások egyforma vagy különböző voltát az

$$F = F_{f_1, f_2} = \frac{\sigma_n^2}{\sigma_m^2}$$

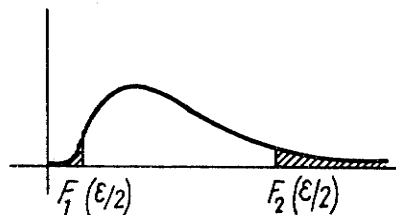
statisztika alapján döntjük el, ahol $f_1 = n-1$ és $f_2 = m-1$. (Azt, hogy melyik minta tapasztalati szórása szerepeljen a számlálóban, a mintavétel előtt döntjük el.)

Az F_{f_1, f_2} valószínűségi változó másodrendű béta eloszlást követ, azaz eloszlásfüggvénye:

$$P(F_{f_1, f_2} < x) = \frac{f_1}{f_2} \frac{\Gamma(f_1 + f_2)}{\Gamma(\frac{f_1}{2})\Gamma(\frac{f_2}{2})} \int_0^x \frac{(\frac{f_1}{f_2} u)^{\frac{f_2}{2} - 1}}{(1 + \frac{f_1}{f_2} u)^{\frac{f_1 + f_2}{2}}} du \quad x \geq 0$$

A $H_1 : D(\xi) \neq D(\eta)$, ellenhipotézissel szemben kritikus tartományként $1 - \varepsilon$ szinten a 69. ábrán látható tartományt szokás választani.

A gyakorlatban rendszerint úgy járunk el, hogy a nagyobb tapasztalati szórásnégyzetet osztjuk a kisebbel, vagyis az



69. ábra

$$F^* = \max \left(\frac{S_n^{*2}}{S_m^{*2}}, \frac{S_m^{*2}}{S_n^{*2}} \right) > 1$$

statisztikát tekintjük és F^* értékét hasonlítjuk össze a táblázatbeli $F_{1-\varepsilon}$ értékkel. (A táblázaton szereplő f_1 mindig a számlálóhoz, f_2 a nevezőhöz tartozó mintadarabszám eggyel csökkentett értéke. $F^* < F_{1-\varepsilon}$ esetén elfogadjuk, $F^* \geq F_{1-\varepsilon}$ esetén elutasítjuk H_0 -t.

9.7 Több szórás megegyezésének vizsgálata; a Bartlett-próba

Mint a 9.6 pontban láttuk, azt a kérdést, hogy két normális eloszlású valószínűségi változó szórása megegyezik-e, az F-próba segítségével döntjük el. Legyen most $\xi^{(1)}, \xi^{(2)}, \dots, \xi^{(k)}$ összesen k -darab valószínűségi változó, amelyek normális eloszlásúak, s azt akarjuk megvizsgálni, hogy ezek szórása egyenlő-e.

Legyen tehát

$$H_0 : D(\xi^{(1)}) = D(\xi^{(2)}) = \dots = D(\xi^{(k)})$$

Legyenek az egyes változókra vonatkozó minták:

$$\xi^{(1)}: \xi_{11}, \xi_{12}, \dots, \xi_{1,m}$$

$$\xi^{(2)}: \xi_{21}, \xi_{22}, \dots, \xi_{2,m}$$

⋮

$$\xi^{(k)}: \xi_{k1}, \xi_{k2}, \dots, \xi_{k,n}$$

Számítsuk ki mindegyik mintából a tapasztalati szórásnégyzetet S_1^{*2} , $S_2^{*2}, \dots, S_k^{*2}$ -et, valamint ezek számtani közepét

$$S^{*2} = \frac{1}{k} \sum_{i=1}^k S_i^{*2} \text{-et.}$$

Bartlett kimutatta, hogy a következő statisztika:

$$k^2 = \frac{2,3026}{C} k(n-1) \left(\log S^{*2} - \frac{1}{k} \sum_{i=1}^k \log S_i^{*2} \right)$$

ahol $C = 1 + \frac{k+1}{3k(n-1)}$, közelítőleg χ^2 -eloszlású $k-1$ paraméterrel. A közelítés akkor kielégítő, ha egyes minták elemszáma > 4 . (A log itt 10-es alapú logaritmust jelent.)

9.8 Pitman-próba

A gyakorlatban a statisztikai vizsgálatok során igen gyakran felmerül az a kérdés, hogy a két valószínűségi változónak, ξ -nek és η -nak azonos-e a várható értéke vagy sem. Ha mind ξ , mind η normális eloszlásúak, akkor a Student-féle t-próba kétmintás változatának alkalmazásával dönthető

el a kérdés. Igen gyakran azonban nem normális eloszlású változókkal van dolgunk, így a t -próba nem alkalmazható. Ha a valószínűségi változókra vonatkozólag a normális eloszlás nem feltételezhető, de eloszlástípusuk azonos, akkor a $H_0: M(\xi) \equiv M(\eta)$ hipotézis ellenőrzésére Pittman a következő eljárást ajánlotta.

Legyen a ξ -re vonatkozó minta: $x_1, x_2, \dots, x_n$ (I)

az η -ra vonatkozó minta: $y_1, y_2, \dots, y_m$ (II)

Tételezzük fel, hogy ξ és η egyforma eloszlásúak, ekkor lényegében egy $n+m$ elemű mintánk van valamely eloszlásból. Így $(n+m)$ elemű mintát $\binom{n+m}{m}$ féle módon oszthatunk be két csoportba úgy, hogy egyik n -elemű, a másik m elemű legyen. Ugy tekintjük, hogy a rendelkezésünkre álló két megfigyelés sorozat a lehetséges csoportosítások egyike.

Számítsuk ki az adott két mintára vonatkozólag a két minta számtani közepéi közötti eltérés abszolút értékét, azaz képezzük a

$$(*) \quad d = |\bar{x} - \bar{y}| = \frac{n+m}{nm} \left(\sum_{i=1}^n x_i - n\bar{z} \right)$$

statisztikát, ahol $\bar{z}$ a két minta egyesítéséből nyert (III) $z_1, z_2, \dots, z_{n+m}$ minta számtani közepe.

Kis minták esetében a (III) mintából elkészíthetjük mind az $\binom{n+m}{m}$ számú csoportokra bontást és minden felbontásra kiszámíthatjuk az $|\bar{x} - \bar{y}|$ statisztikát. Ez a (*) formula alkalmazásával jár legkevesebb számításal. Jelölje a nyert értékeket $\sigma_1, \sigma_2, \dots, \sigma_{\binom{n+m}{m}}$ nagyság szerint csökkenő sorrendben felsorolva.

Legyen most k olyan egész szám, hogy $\frac{k}{n+m} = \alpha$ (pl. $\alpha = 0,05$). Ha $\sigma_1, \sigma_2, \dots, \sigma_{\binom{n+m}{m}}$ számok között kevesebb mint k számú olyan van, amely nagyobb, mint d , azaz $\sigma_k < d$, akkor a H_0 hipotézist $1 - \alpha$ szinten elvetjük.

Nagy minták esetén ez az eljárás nem vihető keresztül, de nincs is rá szükség. A H_0 hipotézis fennállása esetén a

$$d = |\bar{x} - \bar{y}| = \frac{\sqrt{nm(n+m-2)}}{\sqrt{(n+m)(nS_1^2 + mS_2^2)}}$$

statisztika Student-féle t-eloszlást követ, így lényegében a kétmintás t-próba alkalmazható akkor is, ha az alapsokaság nem normális eloszlású.

9.9 Több normális eloszlású valószínűségi változó várható értékének összehasonlítása

(Szórásanalízis)

Mint a 9.4 pontban láttuk, két normális eloszlású, azonos szórású valószínűségi változó várható értékének összehasonlítására a Student-féle kétmintás t-próba szolgál. Vizsgáljuk most, hogyan lehet eldönteni, hogy több független, azonos de ismeretlen szórású valószínűségi változó szórása megegyezik-e vagy sem.

Az ismertetendő matematikai módszer az R. A. Fisher által kezdeményezett ún. szórásanalízis, vagy idegen szóval variancia-analízis többek között ennek a problémának eldöntésére is alkalmas. Maga a szórásanalízis számos területen alkalmazást nyert, a kérdéskör részleteit itt nem tárgyalhatjuk, utalunk erre vonatkozólag a [23] könyvre.

Legyenek $\xi_1, \xi_2, \dots, \xi_r$ független, normális eloszlású valószínűségi változók egyforma, de ismeretlen σ szórással (ezt a Bartlett-próbával ellenőrizni kell). Jelöljük az egyes változók várható értékeit $M(\xi_1) = m_1, M(\xi_2) = m_2, \dots, M(\xi_r) = m_r$ -rel és legyen a H_0 hipotézis $m_1 = m_2 = \dots = m_r = m$.

Legyenek, az egyes valószínűségi változókra vonatkozó minták, amelyek nem feltétlen egyforma elemszámúak:

$$\xi_1: x_{11}, x_{12}, \dots, x_{1n_1}$$

$$\xi_2: x_{21}, x_{22}, \dots, x_{2n_2}$$

.

.

.

$$\xi_r: x_{r1}, x_{r2}, \dots, x_{rn_r}$$

$$\sum_{i=1}^r n_i = n$$

Amennyiben a H_0 hipotézis igaz, valamennyi ξ_{ij} mintaelem $N(m, \sigma)$ eloszlású.

Képezzük most az egyes minták számtani közepét az $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

statisztikákat, valamint az összes megfigyelés számtani közepét, amely egyben az egyes minták számtani közepének számtani közepével egyezik meg.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^r n_i \bar{x}_i$$

Könnyű belátni, hogy

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) (\bar{x}_i - \bar{x}) = \sum_{i=1}^r \left[(\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) \right] = 0$$

Ugyanis

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = n_i \bar{x}_i - n_i \bar{x}_i = 0$$

Ennek alapján:

$$\begin{aligned} (9.9.1) \quad \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 &= \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 = \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2 \end{aligned}$$

Az utóbbi formula mutatja, hogy az összes megfigyelés négyzetes eltéréseinek összege két összegre bomlik. A formula jobb oldalán álló első összeg az egyes minták elemeinek az illető mintaközéptől való négyzetes eltéréseinek összege (nevezzük ezt csoporton belüli négyzetösszegek össz-

szegének) a második a csoportátlagoknak az összes megfigyelések átlagától való négyzetes eltéréseinek súlyozott összege.

Ha a $H_0: m_1 = m_2 = \dots = m_r$ hipotézis igaz, akkor mivel mindegyik x_{ij} változó egyforma eloszlású és függetlenek

$$(9.9.2) \quad \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = (n-1) S^2$$

χ^2 eloszlású $(n-1)$ szabadságfokkal. (Lásd: 5.14. a pont). Hasonlóan a

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

menyiség is χ^2 -eloszlású $(n-1)$ szabadságfokkal. A χ^2 -eloszlás addíciós tétele alapján (5.14. a) a

$$(9.9.3) \quad \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = (n-r) S_2^2$$

ugyancsak χ^2 -eloszlású $\sum_{i=1}^r (n_i - 1) = n-r$ szabadságfokkal.

Az $(n-r) S_2^2$ tagot kvadratikus maradékösszegnek nevezzük.

Amennyiben a H_0 hipotézis igaz, akkor érvényes a következő tétel:

A

$$\sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2 = (r-1) S_1^2$$

menyiség független az $(n-r) S_2^2$ mennyiségtől és χ^2 eloszlású $(r-1)$ szabadságfokkal.

A tétel helyességét beláthatjuk az alábbi, Fishertől eredő lemma segítségével (lásd [22]), amelyet bizonyítás nélkül közlünk.

Lemma: Legyenek $S_1, S_2, \dots, S_k$ független valószínűségi változók egyforma $N(0, \sigma^2)$ eloszlással. Legyenek továbbá $\eta_1, \eta_2, \dots, \eta_p$ ($p < k$) lineáris függvényei a $S_1, S_2, \dots, S_k$ változóknak:

$$\eta_i = c_{i1} S_1 + c_{i2} S_2 + \dots + c_{ik} S_k \quad (i=1, 2, \dots, p)$$

ahol a c_{ij} együtthatók az alábbi ortogonalitási relációknak tesznek eleget:

$$\sum_{j=1}^k c_{ij} c_{\ell j} = \begin{cases} 1 & \text{ha } i = \ell \\ 0 & \text{ha } i \neq \ell \end{cases} \quad (i, \ell = 1, 2, \dots, p)$$

Ekkor a

$$\sum_{j=1}^k \zeta_j^2 - \sum_{i=1}^p \eta_i^2$$

különbség független az $\eta_1, \eta_2, \dots, \eta_p$ változóktól, tehát független a

$\sum_{i=1}^p \eta_i^2$ négyzetösszegétől is, és χ^2 -eloszlású $k-p$ szabadságfokkal.

A lemmára támaszkodva nem nehéz kimutatni, hogy az $(r-1)$ kifejezés független $(n-r)$ S_2^2 -től és χ^2 -eloszlású $(r-1)$ szabadságfokkal.

Ennek következtében

$$(9.9.4) \quad S_1^2 = \frac{\sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2}{r-1} \quad \text{és} \quad S_2^2 = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n-r}$$

változók is függetlenek, közös várható értékük: $M(S_1^2) = M(S_2^2) = \sigma^2$ és

hányadosuk - amennyiben H_0 igaz - F-eloszlású (lásd: 5.14.c pont)

$(r-1; n-r)$ paraméterrel. A H_0 hipotézis ellenőrzésére tehát az F próba alkalmazhatsz.

Amennyiben az

$$F_{r-1, n-r} = \frac{S_1^2}{S_2^2} > F(\varepsilon)$$

(ε a választott valószínűségi szint), akkor a H_0 hipotézist ε -szinten elvetjük. A $F(\varepsilon)$ kritikus értékek a VII. táblázatból olvashatók ki.

A szórásanalízisben szokásos az alábbi ún. szórásfelbontó táblázat készítése:

A szóródás oka	Négyzetösszeg	Szabadsági fok	Empirikus szórásnégyzet
Csoportok között	$(r-1)S_1^2 = \sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2$	$r-1$	$S_1^2 = \frac{\sum_{i=1}^r n_i (\bar{x}_i - \bar{x})^2}{r-1}$
Csoportokon belül	$(n-r)S_2^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$n-r$	$S_2^2 = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{n-r}$
Teljes	$(n-1)S^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$n-1$	$S^2 = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2}{n-1}$

Megjegyezzük, hogy tulajdonképpen nem az S^2 szórásnégyzetet bontottuk összeadandókra, hiszen $S_1^2 + S_2^2 \neq S^2$, hanem a szórásnégyzetek a szabadságfokkal szorozva összegeződnek, azaz

$$(r-1) S_1^2 + (n-r) S_2^2 = (n-1) S^2$$

mint ez a (9.9.1) formulából látható.

9.10 Szekvenciális módszer hipotézisvizsgálata

A becslésről és hipotézisvizsgálatról szóló eddigi fejtegetéseinkben általában feltételeztük, hogy a mintaelemek n száma rögzített. Az egyes becslési és hipotézisvizsgálati módszerek általában annál pontosabbak, minél több elemű mintával dolgozhatunk. Az idő és a költségek azonban majdnem mindig korlátot szabnak a megfigyelések száma növelésének. A Wald Ábrahám által kidolgozott szekvenciális módszer lehetővé teszi, hogy amikor becslésméleti vagy hipotézisvizsgálati problémákról van szó, lényegesen kevesebb mintaelemből ugyanolyan biztonsággal vonhassunk le

statisztikai következtetéseket, mint a legjobb fix mintaelemszámon alapuló eljárások alkalmával.

Az eddigi eljárások folyamán egy $x_1, x_2, \dots, x_n$ mintából mint számadatokból készítettünk valamilyen statisztikát, néhány számadatba tömörítettük a mintaelemek számértékeit és csakis ezeket a számértékeket vettük figyelembe. Nem érdekelt bennünket, hogyan változtak ezek az értékek időrendben, s így a mintában tükröződő valószínűségi változót tekintve értékes információt hagytunk kihasználatlanul. Megjegyezzük, hogy különböző megfigyelések, kísérletek során sokszor igen fontos lehet a mintaelemek értékén kívül az értékek időbeli alakulása, váltakozása is.

Az alábbiakban a szekvenciális módszert hipotézisvizsgálat esetére ismertetjük, amikor mind a H_0 null-hipotézis, mind a H_1 alternatív hipotézis egyszerű feltevés.

Tételezzük fel, hogy valamely ξ valószínűségi változó eloszlását az $f(x; \Theta)$ sűrűségfüggvény jellemzi.

Legyen a H_0 null-hipotézis: $\Theta = \Theta_0$; továbbá a

H_1 alternatív hipotézis: $\Theta = \Theta_1$.

A szekvenciális módszer lényege, hogy egymást követő megfigyeléseket végezve, minden egyes mintaelem megfigyelése után a következő döntések valamelyikét hozzuk:

- a) a H_0 hipotézist elfogadjuk
- b) a H_0 hipotézist elvetjük
- c) folytatjuk a kísérletet, újabb megfigyelést végzünk.

Az a) és a b) döntéseket, vagyis a H_0 hipotézis elfogadását vagy elvetését végső döntésnek nevezzük.

A szekvenciális hipotézisvizsgálat alkalmával előzetesen megadjuk, hogy mekkora első-, illetve másodfajú hibát (tévedési valószínűséget) engedünk meg. Legyen az elsőfajú hiba valószínűsége α , a másodfajú hibáé β . Számítsuk ki az

$$A = \frac{1 - \beta}{\alpha} \quad \text{és} \quad B = \frac{\beta}{1 - \alpha}$$

mennyiségeket. Az A és a B számok a szekvenciális próba szempontjából alapvető határmennyiségek lesznek, nevezetesen ezek a számok a számegyenesen három szakaszra, tartományra osztják.

$$\frac{\text{elfogadási tart.} \quad \text{indifferens tart.} \quad \text{kritikus tart.}}{B \quad \quad \quad A}$$

A számegyenesnek az A ponttól jobbra eső részét kritikus vagy elutasítási tartománynak, míg a számegyenesnek B ponttól balra eső részét elfogadási tartománynak nevezzük. A B és A pontok közé eső szakasz az indifferens tartomány.

Magát a szekvenciális próbát az alábbi módon hajtjuk végre; a ξ valószínűségi változóra egymás után végzünk megfigyeléseket. Legyen az első megfigyelés eredménye, az első mintaelem x_1 . Kiszámítjuk az $f(x_1; \theta_0)$ és $f(x_1; \theta_1)$ értéket és képezzük az

$$L = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)}$$

hányadost.

Ha pl. ξ normális eloszlású valószínűségi változó ismert szórással és $H_0: M(\xi) = 0$, $H_1: M(\xi) = 1$, akkor

$$\ell_1 = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1 - 1)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}}}$$

Amennyiben

$$\ell_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)} \geq A,$$

a H_0 hipotézist már elvetjük, mert ℓ_1 értéke az elvetési tartományba esett. Ha

$$\ell_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)} \leq B,$$

akkor a H_0 -t elfogadjuk, mert az ℓ_1 értéke az elfogadási tartományba esett. Ha viszont

$$B < \ell_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)} < A$$

akkor újabb megfigyelést végzünk, mert az ℓ_1 értéke az indifferens tartományba esett. Tegyük fel, hogy az utóbbi eredményt kaptuk, vagyis az első mintaelem x_1 , nem eredményezett végső döntést. Ekkor újabb megfigyelést végzünk, legyen ez x_2 , és kiszámítjuk az

$$\ell_2 = \frac{f(x_2; \Theta_1)}{f(x_2; \Theta_0)}$$

hányadost, majd képezzük az

$$L = \ell_1 \ell_2 = \frac{f(x_1; \Theta_1) f(x_2; \Theta_1)}{f(x_1; \Theta_0) f(x_2; \Theta_0)}$$

mennyiségeket és a döntést illetően ugyanugy járunk el, mint az előbb. Ha az L értéke ismét az indifferens tartományba esik, akkor újabb megfigyelést végzünk, legyen ez x_3 és ismét kiszámítjuk az

$$L = \ell_1 \ell_2 \ell_3 = \frac{f(x_1; \Theta_1) f(x_2; \Theta_1) f(x_3; \Theta_1)}{f(x_1; \Theta_0) f(x_2; \Theta_0) f(x_3; \Theta_0)} = \frac{\prod_{i=1}^3 f(x_i; \Theta_1)}{\prod_{i=1}^3 f(x_i; \Theta_0)}$$

hányadost, amelyet likelihood-hányadosnak nevezünk.

Általában a k -adik mintaelem megfigyelése után (ha az első $k-1$ számú megfigyelés nem vezetett végső döntéshez), az

$$L = \ell_1 \ell_2 \dots \ell_k = \frac{\prod_{i=1}^k f(x_i; \Theta_1)}{\prod_{i=1}^k f(x_i; \Theta_0)} = \frac{p_{1k}}{p_{0k}}$$

likelihood alapján

$$L \geq A$$

esetén elvetjük a H_0 -t,

$$L \leq B$$

esetén elfogadjuk a H_0 -t,

$$B < L < A$$

esetén pedig újabb $k+1$ -edik mintaelemet veszünk. A mintavételt mindaddig folytatjuk, amíg L értéke az A vagy a B határokat nem lépi át.

Vegyük kissé szemügyre az

$$L = \frac{p_{1k}}{p_{0k}} = \frac{\prod_{i=1}^k f(x_i; \Theta_1)}{\prod_{i=1}^k f(x_i; \Theta_0)}$$

likelihood-hányadosot.

Ezt az első k megfigyelés eredményeként kapott

$$x_1, x_2, \dots, x_k$$

minta alapján számítottuk, ahol az x_i mintaelemek független, azonos eloszlású valószínűségi változók.

Ha a ξ diszkrét eloszlású, akkor az $f(x_i; \Theta_1)$ jelentse a $P(\xi = x_i | H_1)$, az $f(x_i; \Theta_0)$ pedig a $P(\xi = x_i | H_0)$ valószínűséget. Ekkor $\prod_{i=1}^k f(x_i; \Theta_1)$ az egyes mintaelemek együttes bekövetkezésének, vagyis az adott mintának a valószínűsége azzal a feltétellel, hogy a H_1 feltevés igaz, a $\prod_{i=1}^k f(x_i; \Theta_0)$ pedig a minta valószínűsége azzal, hogy a H_0 feltevés igaz. Az $L = \frac{p_{1k}}{p_{0k}}$ hányados tehát azt fejezi ki, hányszorosa valószínűbb az adott minta előfordulása a H_1 , mint a H_0 feltevessel.

Ugyanez a helyzet akkor is, amikor a ξ folytonos eloszlású. Ez esetben a kérdéses valószínűségek hányadosát (mivel $P(\xi = x_i | H_1) \approx f(x_i; \Theta_1) \Delta x_i$)

$$\frac{\prod_{i=1}^k f(x_i; \Theta_1) \Delta x_i}{\prod_{i=1}^k f(x_i; \Theta_0) \Delta x_i}$$

alakban kellene írunk, azonban a Δx_i tényezőktől el szoktunk tekinteni.

Az L jelentése alapján a fent leírt próbát valószínűségi hányadospróbanak nevezzük.

A próba végrehajtása alkalmával az egyes $\ell_i = \frac{f(x_i; \Theta_1)}{f(x_i; \Theta_0)}$

mennyiségek helyett ezek logaritmusait használjuk, mert ekkor szorzás

helyett összeadással dolgozhatunk. A $z_i = \log \ell_i = \log \frac{f(x_i; \Theta_1)}{f(x_i; \Theta_0)}$

változók bevezetésével

$$\log L = \log \frac{\prod_{i=1}^k f(x_i; \Theta_1)}{\prod_{i=1}^k f(x_i; \Theta_0)} = z_1 + z_2 + \dots + z_k.$$

Minden egyes mintaelemre kiszámítjuk a

$$z_i = \log \frac{f(x_i; \Theta_1)}{f(x_i; \Theta_0)}$$

mennyiséget, és ha a $\log B < z_1 + z_2 + \dots + z_k < \log A$, akkor a mintavételt folytatjuk, viszont

$$z_1 + z_2 + \dots + z_k \geq \log A \text{ esetén elvetjük a } H_0\text{-t,}$$

$$z_1 + z_2 + \dots + z_k \leq \log B \text{ esetén elfogadjuk a } H_0\text{-t.}$$

E két utóbbi esetben tehát végső döntést hozunk. Az egyes mennyiségek pozitív vagy negatív értéket vesznek fel aszerint, hogy az $\frac{f(x_i; \Theta_1)}{f(x_i; \Theta_0)}$ értéke 1-nél nagyobb vagy kisebb.

Az elmondottak jobb megvilágítására lássuk a normális eloszlás esetét. Tegyük fel, hogy a ξ normális eloszlású változó $\sigma^2 = 1$ szórásnégyzettel, ismeretlen $\bar{\Theta}$ várható értékkel. Legyen továbbá

$$H_0: \bar{\Theta} = \bar{\Theta}_0, \quad H_1: \bar{\Theta} = \bar{\Theta}_1$$

(ahol $\bar{\Theta}_1 > \bar{\Theta}_0$).

Ekkor

$$f(x; \Theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\Theta)^2}{2}}$$

$$L = \frac{p_{1k}}{p_{0k}} = \frac{\prod_{i=1}^k e^{-\frac{1}{2} (x_i - \Theta_1)^2}}{\prod_{i=1}^k e^{-\frac{1}{2} (x_i - \Theta_0)^2}}$$

$$Z_1 = \log \frac{e^{-\frac{1}{2} (x_1 - \Theta_1)^2}}{e^{-\frac{1}{2} (x_1 - \Theta_0)^2}} = \log e^{\frac{1}{2} [(x_1 - \Theta_0)^2 - (x_1 - \Theta_1)^2]}$$

$$Z_1 = \log e^{x_1 (\Theta_1 - \Theta_0) + \frac{1}{2} (\Theta_0^2 - \Theta_1^2)} = \frac{1}{2} (\Theta_0^2 - \Theta_1^2) + (\Theta_1 - \Theta_0) x_1$$

A z_1 változók összegét képezve:

$$z_1 + z_2 + \dots + z_k = (\Theta_1 - \Theta_0) [(x_1 + x_2 + \dots + x_k) - \frac{k}{2} (\Theta_1 + \Theta_0)]$$

$$\log B = \log \frac{\beta}{1-\alpha} < z_1 + z_2 + \dots + z_k < \log \frac{1-\beta}{\alpha} = \log A$$

egyenlőtlenség tehát most ekvivalens az

$$\frac{1}{\Theta_1 - \Theta_0} \log \frac{\beta}{1-\alpha} + \frac{k}{2} (\Theta_1 + \Theta_0) < \sum_{i=1}^k x_i < \frac{1}{\Theta_1 - \Theta_0} \log \frac{1-\beta}{\alpha} + \frac{k}{2} (\Theta_1 + \Theta_0)$$

Konkrét esetben tegyük fel, hogy $\alpha = 0,05$, $\beta = 0,10$ első-, illetve másodfajú hibát engedünk meg (gyakorlati megfontolások alapján, a vizsgált ismerv természetétől és a különféle fajtájú hibák okozta kártól függően) és legyen

$$\Theta_0 = 9,5 \quad \Theta_1 = 10$$

Ekkor a

$$-4,50 + 9,75 k < \sum x_i < 5,78 + 9,75 k$$

egyenlőtlenséghez jutunk.

Amennyiben $\sum x_i \leq -4,50 + 9,75 k$, (I.) elfogadjuk a H_0 -t: $\Theta = 9,5$,
ha viszont $\sum x_i \geq 5,78 + 9,75 k$, (II.) elvetjük a H_0 -t és a
 $H_1 : \Theta = 10$ feltevését fogadjuk el. Ha az (I) vagy a (II) esetek egyike sem
fordul elő, akkor folytatjuk a megfigyeléssorozatot.

Másik példaként nézzük meg a binomiális eloszlás esetét. A legyen
binomiális eloszlású valószínűségi változó ismeretlen paraméterrel. Itt
tehát a kísérletek olyan sorozatáról van szó, amelyben minden egyes ki-
sérletnek csak kétféle kimenetele lehet. Jelöljük az egyik kimenetelt 1-
gyel, a másikat 0-val, vagyis az x_i mintaelemek mindegyike az 1 vagy a 0
értéket veheti csak fel. Ekkor

$$P(\xi = 1) = f(1, p) = p, \quad P(\xi = 0) = f(0, p) = 1 - p$$

Legyen a H_0 hipotézis: $p = p_0$, a H_1 alternatív hipotézis: $p = p_1$.

Erre az esetre

$$z_i = \log \frac{f(x_i; p_1)}{f(x_i; p_0)} = \begin{cases} \log \frac{p_1}{p_0}, & \text{ha } x_i = 1 \\ \log \frac{1 - p_1}{1 - p_0}, & \text{ha } x_i = 0 \end{cases}$$

Ha K_n -nel jelöljük az első n számú mintaelem közül az 1-esek számát,
akkor

$$z_1 + z_2 + \dots + z_n = K_n \log \frac{p_1}{p_0} + (n - K_n) \log \frac{1 - p_1}{1 - p_0} =$$

$$\begin{aligned}
&= n \log \frac{1-p_1}{1-p_0} + K_n \left(\log \frac{p_1}{p_0} - \log \frac{1-p_1}{1-p_0} \right) = \\
&= n \log \frac{1-p_1}{1-p_0} + K_n \left(\log \frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1} \right)
\end{aligned}$$

A kísérletsorozatot mindaddig folytathatjuk, amíg fennáll, hogy

$$\log B < z_1 + z_2 + \dots + z_n < \log A$$

más szóval amíg

$$\begin{aligned}
(9.10.1) \quad & \frac{\log B}{\log \frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1}} + n \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1}} < K_n < \frac{\log A}{\log \frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1}} + \\
& + n \frac{\log \frac{1-p_0}{1-p_1}}{\log \frac{p_1}{p_0} \cdot \frac{1-p_0}{1-p_1}}
\end{aligned}$$

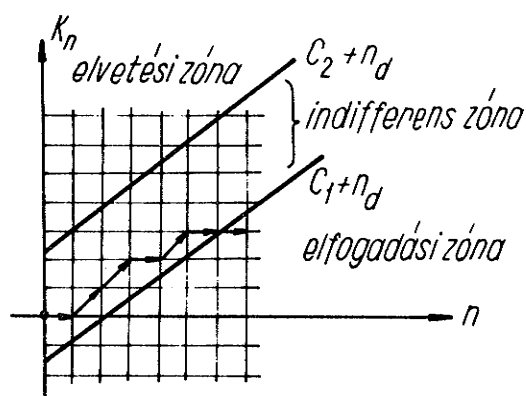
Ez az egyenlőtlenség

$$(9.10.2) \quad c_1 + nd < K_n < c_2 + nd$$

alakban írható, ahol c_1 , c_2 és d jelentése a (9.10.1) és a (9.10.2) relációk összehasonlítása alapján nyilvánvaló. A (9.10.2) egyenlőtlenség első és utolsó tagja egy-egy egyenes egyenlete, amelyek párhuzamosak, a d iránytangens pedig pozitív előjellű.

Ezek után az (n, K_n) értékpárokat egy derékszögű koordináta-rendszerben ábrázolva, a (9.10.2) egyenlőtlenség által meghatározott tartomány a sík mindazon egész koordinátájú pontjaiból ún. rácspontjaiból áll, amelyek a $c_1 + nd$ és $c_2 + nd$ párhuzamos egyenesek között vannak (70. ábra).

A szekvenciális próba lefolyását olyan bolyongásként is értelmezhetjük, amelyben a mozgó pont a P valószínűségtől függően P valószínűséggel lép egyet függőleges irányban. Ha a mozgó pont útja az alsó $c_1 + n_d$ egyenest metszi át, akkor a H_0 -t elfogadjuk, ha pedig a felső $c_2 + n_d$ egyenest metszi, a H_0 -t elutasítjuk.



70. ábra

NEMPARAMÉTERES PRÓBÁK

Az eddigiekben ismertetett statisztikai próbákat - mint azt megjegyeztük - akkor alkalmazzuk, ha a vizsgált eloszlás függvényalakja ismert és feladatunk az eloszlás paramétereire vonatkozó döntés meghozatala.

Gyakran előfordul a gyakorlatban, hogy nem vagyunk biztosak az eloszlás függvényalakjában sem, legfeljebb gyanítjuk, hogy pl. normális vagy lognormális, vagy egyenletes eloszlásról van szó. Az is előfordulhat, hogy semmiféle támpontunk nincs a szóban forgó $\mathcal{X}$ valószínűségi változó eloszlását illetően. Ilyenkor nagy mintából kell kiindulnunk (ha van rá lehetőségünk 150-300 megfigyelési értékből). Először célszerű a gyakorisági-hisztogramot megalkotni, ennek alakja utalhat valamilyen elméleti eloszlás-típusra, majd statisztikai próbával un. illeszkedésvizsgálattal ellenőrizzük, hogy az észlelt megfigyelési értékek, a minta elég jól "illeszkedik-e" a feltételezett $F(x)$ elméleti eloszláshoz. Abban az esetben, ha tapasztalati alapon módunkban áll az $F(x)$ eloszlás paramétereit is feltételezni (pl. ha feltehetjük, hogy adott $m = m_0$ várható értékű és $\sigma = \sigma_0$ szórásu normális eloszlással állunk szemben), akkor un. tiszta

illeszkedésvizsgálatot végezhetünk. Ha viszont az eloszlás paramétereit illetően nincs támpontunk, csak az eloszlás típusát sejtjük, akkor a paramétereket a mintából kell becsülnünk és un. becsléses illeszkedésvizsgálatot végeznünk, ami az eljárás hatásfokát némileg csökkenti.

Az illeszkedésvizsgálatra használatos próbák között a két legelterjedtebb a Kolmogorov-próba és a χ^2 -próba. A Kolmogorov-próba csak folytonos eloszlású $\mathcal{X}$ valószínűségi változó eloszlásának vizsgálatára alkalmas, a χ^2 -próba mind diszkrét, mind folytonos eloszlások esetében alkalmazható. Az illeszkedésvizsgálat leggyakoribb esete az un. normalitásvizsgálat, amelyre egyéb egyszerű grafikus eljárás is használatos, leginkább elsődleges tájékozódás céljára. Sajnos mind a Kolmogorov-próba, mind a χ^2 próba elsősorban nagy mintaelemszám mellett megbízható. Kisebb mintaelemszámok esetén még normalitásvizsgálatnál is bonyolultabb számításokhoz, a változók transzformációjához kell folyamodnunk. Lásd: [17]

Igen gyakran felmerülő statisztikai probléma, hogy két valószínűségi változó, ξ és η azonos $F(x)$ eloszlásúak-e vagy sem. Ha mind ξ , mind η folytonos eloszlásúak, akkor a probléma megválaszolására nagy minták esetén Szmirnov-tétele és az azon alapuló statisztikai próba a leghasználatosabb. Ugyancsak alkalmazható erre az esetre is a χ^2 próba egyik változata. Kisebb mintaelemszámok esetén a két minta összehasonlítására a Gnedenko-Koroljuk-féle tétel és annak általánosítása szolgálhat alapul. Röviden ismertetünk egyéb módszereket is két minta összehasonlítására. Az ismertetendő próbák jelentős része kevés számítást igényel, főleg a rendezett minták elméletén alapuló eljárások esetében. A próbák ismertetése előtt célszerűnek látjuk röviden vázolni a rendezett minták elméletének legfontosabb elemeit.

10.1 A rendezett minták elméletének elemei

A rendezett minta fogalmát már említettük a tapasztalati eloszlásfüggvénnyel kapcsolatban. A mintaelemek nagyság szerinti rendezése azonban korántsem csak a tapasztalati eloszlásfüggvény megalkotásához szükséges "fogás", hanem a nagyság szerint rendezett megfigyelési értékek, mintaelemek eloszlásával kapcsolatos vizsgálatok a matematikai statisztikának gazdag fejezetévé fejlődtek és statisztikai következtetések levonásának fontos és hasznos eszközeivé lettek.

Röviden ismertetjük a rendezett minták elméletének néhány elemét.

Legyen ξ folytonos eloszlású valószínűségi változó $F(x)$ eloszlásfüggvénnyel és legyen $\xi_1, \xi_2, \dots, \xi_n$ a ξ -re vonatkozó n számú megfigyelés, a minta. Keressük ki a mintaelemek közül a legkisebbet ξ_1^* -gal stb. Az eljárást folytatva adatsorunk nagyság szerint növekvő sorrendben áll előttünk:

$$\xi_1^*, \xi_2^*, \dots, \xi_n^*$$

Ez a rendezett minta, ahol most már az indexek nincsenek összefüggésben az észlelés sorrendjével csak a nagyságrenddel. Ha pl. egy folyó vízállása egy adott pl. június hónapban és $\xi_1, \xi_2, \dots, \xi_{30}$ a naponta mért vízállásokat jelölik, akkor ha június 10-én észleltük a legalacsonyabb vízállást abban a hónapban, akkor $\xi_{10} = \xi_1^*$, míg ha a legmagasabb júniusi vízállást június 1-én észleltük, akkor $\xi_1 = \xi_{30}^*$ stb.

Az $F(x)$ eloszlásfüggvénnyel bíró folytonos eloszlású ξ valószínűségi változóra vonatkozó minta elemei, a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók mindegyikének $F(x)$ az eloszlásfüggvénye, azaz

$$P(\xi_i < x) = F(x), \quad (i = 1, 2, \dots, n)$$

Azáltal, hogy a mintaelemeket nagyság szerint rendeztük, közöttük bizonyos relációk keletkeztek, így a $\{\xi_k^* < x\}$ esemény valószínűségét nem adja meg az $F(x)$ eloszlásfüggvény. Az ξ_k^* valószínűségi változó eloszlásának meghatározása némileg bonyolultabb, de annál hasznosabb, mivel ennek ismeretében pl. a folyó legkisebb, vagy legnagyobb vizállásának, azaz ξ_1^* -nak és ξ_n^* -nak eloszlását is ismerjük.

Határozzuk meg a nagyság szerinti k -adik mintaelemnek, ξ_k^* -nak az eloszlásfüggvényét.

Tekintsük a (7.3.1) formulával definiált tapasztalati eloszlásfüggvényt, $F_n(x)$ -et. Említettük, hogy az $F_n(x)$ tapasztalati eloszlásfüggvény a $\{\xi < x\}$ esemény relatív gyakoriságát adja meg.

Az $F_n(x)$ függvény rögzített x pontban $\frac{m}{n}$ -nel egyenlő, ha pontosan m -számu mintaelem kisebb, mint x , a többi $n-m$ számu mintaelem pedig nagyobb, mint x , azaz alapján, mivel binomiális eloszlással állunk szemben:

$$(10.1.1) \quad P(F_n(x) = \frac{m}{n}) = \binom{n}{m} [F(x)]^m [1 - F(x)]^{n-m}$$

A $\{\xi_k^* < x\}$ esemény akkor következik be, ha legalább k számu mintaelem kisebb mint x , azaz

$$(10.1.2) \quad G_{k,n}(x) = P(\xi_k^* < x) = P(F_n(x) \geq \frac{k}{n}) = \\ = \sum_{m=k}^n \binom{n}{m} [F(x)]^m [1 - F(x)]^{n-m}$$

Könnyű belátni, hogy a $G_{k,n}(x)$ eloszlásfüggvény integrálalakban is felírható a következő módon:

$$(10.1.3) \quad G_{k,n}(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} t^{k-1} (1-t)^{n-k} dt.$$

A (10.1.2) és (10.1.3) ekvivalens voltát parciális integrálással nyerhetjük.

Amennyiben ξ eloszlásának sűrűségfüggvénye létezik és ez $f(x)$, akkor ξ_k^* sűrűségfüggvénye:

$$(10.1.4) \quad g_{k,n}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} \cdot f(x)$$

Ha pl. ξ a $[0,1]$ intervallumban egyenletes eloszlású, akkor sűrűségfüggvénye:

$$(10.1.5) \quad g_{k,n}(x) = \begin{cases} \frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{n-k}, & \text{ha } 0 \leq x \leq 1, \\ 0 & \text{egyébként.} \end{cases}$$

Ez esetben un. béta-eloszlásról van szó.

A (10.1.3) formula alapján adódik $k=1$ helyettesítéssel, hogy a legkisebb rendezett mintaelem, ξ_1^* eloszlásfüggvénye:

$$(10.1.6) \quad G_{1,n}(x) = P(\xi_1^* < x) = n \int_0^{F(x)} (1-t)^{n-1} dt =$$

$$= \left[-n \frac{(1-t)^n}{n} \right]_0^{F(x)} = 1 - [1 - F(x)]^n$$

A (10.1.3) formulából $k=n$ helyettesítéssel nyerjük a legnagyobb mintaelem, azaz ξ_n^* eloszlásfüggvényét:

$$(10.1.7) \quad G_{n,n}(x) = P(\xi_n^* < x) = n \int_0^{F(x)} t^{n-1} dt = [F(x)]^n$$

Ez utóbbi fontos eredmény egyébként a rendezett minták elmélete nélkül is egyszerűen származtatható, hiszen a legnagyobb mintaelem akkor kisebb, mint x , ha mindegyik mintaelem kisebb, mint x , azaz mivel független eseményekről van szó:

$$P(\xi_n^* < x) = P(\xi_1 < x) P(\xi_2 < x) \dots P(\xi_n < x) = [F(x)]^n.$$

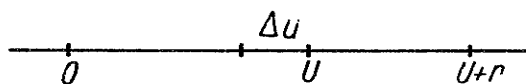
A minta terjedelmének eloszlása

Az $R_n = \xi_n^* - \xi_1^*$ mennyiséget a minta terjedelmének nevezzük. A minta terjedelme egyben ξ értékeinek szóródására jellemző mennyiség, R_n nyilván ugyancsak valószínűségi változó, hiszen mintáról mintára változhat. Jelöljük a minta terjedelmének eloszlásfüggvényét $W_n(r)$ -re. Ezt az eloszlásfüggvényt a következő formula szolgáltatja:

$$(10.1.8) \quad W_n(r) = n \int_{-\infty}^{\infty} [F(r+u) - F(u)]^{n-1} f(u) du.$$

A (10.1.8) formula származtatása egyszerű. Annak valószínűsége, hogy az n számú mintaelemből valamelyik az $(u - \Delta u, u)$ intervallumban esik, közelítőleg

$$n f(u) \Delta u.$$



71. ábra

Annak valószínűsége, hogy egy további mintaelem az $(u, u+r)$ intervallumba esik: $F(u+r) - F(u)$. Annak valószínűsége, hogy ez utóbbi esemény $n-1$ számú mintaelemre teljesül:

$$[F(u+r) - F(u)]^{n-1}$$

Figyelembe véve, hogy a legkisebb mintaelem $(-\infty, +\infty)$ -ben bárhova eshet, nyerjük a (10.1.8) integrál kifejezést.

10.2 Kolmogorov-Szmirnov típusú határeloszlások és statisztikai alkalmazásuk

A rendezett minták elméletének nem csak abban áll a jelentősége, hogy ezen elmélet segítségével a legnagyobb, illetve legkisebb mintaelemek (pl. vizállás maximumok, ill. minimumok) eloszlása meghatározható, bár a gyakorlat számára ez sokszor fontos lehet. A most ismertetésre kerülő eredmények segítségével, az alábbi fontos statisztikai problémákra válaszol.

1. Valamely folytonos eloszlású ξ valószínűségi változóra vonatkozó megfigyelési adatsor, a $\xi_1, \xi_2, \dots, \xi_n$ minta származhat-e adott $F(x)$ eloszlású sokaságból?

2. Két valószínűségi változó ξ és η amelyekről feltesszük, hogy folytonos eloszlásúak, azonos eloszlással bírnak-e, azaz a ξ -re vonatkozó

$$\xi_1, \xi_2, \dots, \xi_n$$

és az η -ra vonatkozó

$$\eta_1, \eta_2, \dots, \eta_m$$

minták azonos eloszlásból származnak-e? (Ilyen problémával állunk szemben, ha pl. egy folyó vizállására vonatkozólag $n+m$ számú mérési adatunk van egy adott szelvényen és arra vagyunk kíváncsiak, hogy az utolsó m adat alapján következtethetünk-e mederfeltöltődésre.)

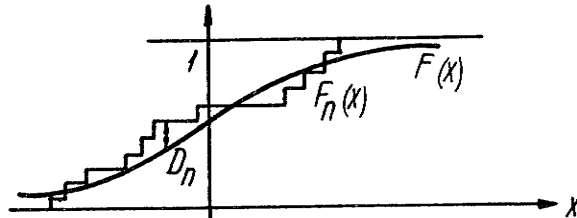
Ismertetjük most a rendezett minták elméletén alapuló eredményeket, amelyek a fenti kérdések megválaszolásához elméleti alapot nyújtanak.

Az előzőekben láttuk, hogyha a megfigyelésünk tárgyát képező folytonos ξ valószínűségi változó eloszlásfüggvénye $F(x)$, akkor a ξ -re vonatkozó n számú megfigyelés eredményét nagyság szerint rendezve, a

$$\xi_1^*, \xi_2^*, \dots, \xi_n^*$$

rendezett minta alapján konstruált $F_n(x)$ tapasztalati eloszlásfüggvény nagy n esetén közelíti az $F(x)$ elméleti eloszlásfüggvényt. Ezt a tényt fogalmazza meg szabatosan Glivenko tétele. Ez a fontos tétel azonban semmit sem mond arról, hogy milyen valószínűséggel lép fel az $F(x)$ elméleti és $F_n(x)$ tapasztalati eloszlásfüggvény között adott nagyságú eltérés. Ahhoz, hogy

el tudjuk dönteni azon hipotézisünk helytállását, hogy a ξ valószínűségi változó eloszlásfüggvénye valóban az $F(x)$ függvény, ismernünk kell a ξ -re vonatkozó megfigyelési eredmények alapján nyert $F_n(x)$ tapasztalati eloszlásfüggvény és az $F(x)$ elméleti eloszlásfüggvény közötti maximális eltérés mekkora valószínűséggel marad adott korlát alatt.



72. ábra

Jelölje D_n az $|F_n(x) - F(x)|$ legnagyobb értékét,* azaz legyen

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|$$

Mivel az $F_n(x)$ tapasztalati eloszlásfüggvény alakulása a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek nagyságától, vagyis a véletlentől függ, a D_n mennyiség valószínűségi változó, beszélhetünk tehát a D_n mennyiség eloszlásáról.

Abban az esetben, ha ξ eloszlásfüggvénye valóban az $F(x)$ függvény, amelyről feltesszük, hogy folytonos, továbbá ha a mintaelemek száma n igen nagy, akkor azt várjuk, hogy D_n értéke kicsi lesz. Ha meg tudunk adni olyan z_0 határt (korlát-számot), amelynél nagyobb eltérés $F_n(x)$ és $F(x)$ között már nagyon kicsi valószínűségű, akkor dönteni tudunk arról, hogy elfogadható-e az a H_0 hipotézis, hogy ξ eloszlásfüggvénye valóban $F(x)$. Ha az észlelt D_n mennyiség értéke eléri vagy meghaladja a megadott korlátot, akkor a hipotézist elvetjük, míg $D_n < z_0$ esetén azt mondjuk, hogy a kísérlet nem mond ellent hipotézisünknek.

A D_n valószínűségi változó eloszlását Kolmogorov szovjet matematikus határozta meg, aki kimutatta, hogy amennyiben $F(x)$ folytonos, akkor $n \rightarrow \infty$ esetén**

* A legnagyobb eltérést supremumnak nevezzük és a sup. rövidítéssel jelöljük.

** A $D_n > \lambda$ esemény valószínűsége $n \rightarrow \infty$ esetén Glivenko tétele értelmében 0-hoz tart, ezért szerepel a fenti határeloszlásban a $\sqrt{n} D_n$ mennyiség.

$$P(\sqrt{n} D_n < z) \rightarrow K(z) = \sum_{\lambda=-\infty}^{\infty} (-1)^\lambda e^{-2\lambda^2 z^2}$$

Nem foglalkozunk itt a fenti reláció bizonyításával, csak arra mutatunk rá, hogyan használható fel a gyakorlatban az említett hipotézis ellenőrzésére. Számunkra most csak az a lényeges, hogy ismeretes egy olyan $K(z)$ eloszlásfüggvény, amelynek értékei táblázatba vannak foglalva, (lásd: II. táblázat), amely minden egyes $z = z_0$ értékre megadja, mi a valószínűsége, hogy

$$\sqrt{n} D_n < z_0$$

legyen.

Gyakorlatilag már most úgy járunk el, hogy a ξ valószínűségi változóra vonatkozólag n számú megfigyelést végzünk, a nyert mintaelemeket nagyság szerint rendezzük, megkonstruáljuk az $F_n(x)$ tapasztalati eloszlásfüggvényt és megnézzük, hogy mekkora a maximális eltérés az $F_n(x)$ és az $F(x)$ görbék között abszolút értékre nézve. Tegyük fel, hogy ez az eltérés D_n^0 . Ezután kiszámítjuk a $z_0 = \sqrt{n} D_n^0$ értéket és megnézzük, mennyi a $K(z)$ eloszlásfüggvény értéke a z_0 helyen, azaz a II. táblázatból kikeressük $K(z_0)$ -t. Ha a kísérletek száma n nagy, akkor

$$P(\sqrt{n} D_n < z_0) \approx K(z_0)$$

és

$$1 - K(z_0) \approx P(\sqrt{n} D_n \geq z_0).$$

Ha $1 - K(z_0)$ értéke kicsi, mondjuk kisebb, mint 0,05, akkor olyan kis valószínűségi esemény következett be, hogy az $F_n(x)$ és $F(x)$ közötti ekkora eltérés nem igen magyarázható véletlen ingadozásként, tehát elvetjük azt a H_0 hipotézist, hogy ξ eloszlásfüggvénye $F(x)$.

Megjegyezzük, hogy az elméleti és tapasztalati eloszlásfüggvény közötti maximális eltérés helyett tekinthető a két függvény közötti százalékos eltérés maximuma. Erre vonatkozólag Rényi Alfréd végzett vizsgálatokat és a következő határeloszlásokat nyerte:

$$\lim_{n \rightarrow \infty} P(\sqrt{n} \sup_x \frac{F_n(x) - F(x)}{F(x)} < y) = \sqrt{\frac{2}{\pi}} \int_0^{\frac{y \sqrt{a}}{1-a}} e^{-\frac{t^2}{2}} dt, (t > 0)$$

és

$$\begin{aligned} & \lim_{n \rightarrow \infty} P(\sqrt{n} \sup_{a \leq F(x)} \left| \frac{F_n(x) - F(x)}{F(x)} \right| < y) = \\ & = \frac{4}{\pi} \sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1} \exp \left[-\frac{(2j+1)^2 \pi^2}{8} \cdot \frac{1-a}{ay^2} \right], (y > 0). \end{aligned}$$

A részletek iránt érdeklődő olvasó számára utalunk a könyvre. A Kolmogorov-próba alapján a felvetett első kérdésünkre tudunk válaszolni, az elméleti és tapasztalati eloszlásfüggvények közötti maximális eltérés, mint a két függvény közötti "távolság" figyelembevételével.

A második kérdésre, a Szmirnov-féle határeloszlás és az erre épülő statisztikai próba adja meg a választ.

Azt a problémát vizsgáljuk tehát, mondhatjuk-e, hogy két minta:

$$(I) \quad x_1, x_2, \dots, x_n$$

$$(II) \quad y_1, y_2, \dots, y_n$$

ugyanazon folytonos $F(x)$ eloszlású statisztikai sokaságból származik.

Tekintsük az (I) mintát egy folytonos $F(x)$ eloszlású ξ valószínűségi változó megfigyelt értékeinek, a (II) mintát egy ugyancsak folytonos $G(y)$ eloszlású η változó megfigyelt értékeinek.

A probléma megoldása céljából nagyság szerint növekvő sorrendbe rendezzük az (I) és (II) minták elemeit. Az (I) és (II) mintákban a minta-elemeket megfigyelésük sorrendjében tüntettük fel.

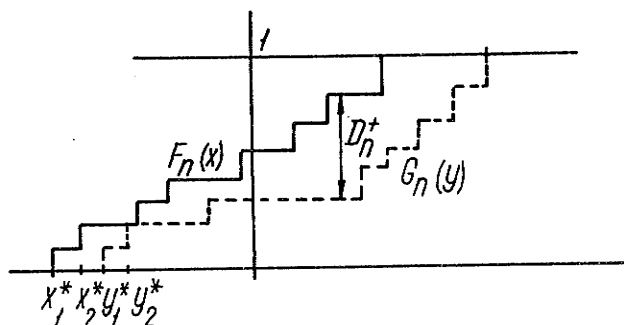
A rendezés után az

$$(I^*) \quad x_1^*, x_2^*, \dots, x_n^*$$

$$(II^*) \quad y_1^*, y_2^*, \dots, y_n^*$$

rendezett mintákat nyerjük.

A két rendezett minta alapján konstruáljuk meg az $F_n(x)$, ill. $G_n(y)$ tapasztalati eloszlásfüggvényeket.



73. ábra

Jelölje D_n^+ az $F_n(x)$ és $G_n(y)$ tapasztalati eloszlásfüggvények közötti legnagyobb pozitív eltérést, azaz legyen

$$D_n^+ = \sup_x [F_n(x) - G_n(y)]$$

(legnagyobb eltérés $F_n(x)$ javára)

Az a kérdés merül fel, hogy milyen nagy lehet a két tapasztalati eloszlásfüggvény közötti D_n^+ maximális eltérés ahhoz, hogy a H_0 hipotézis még elfogadható legyen. Ha a mintaelemszám, n igen nagy szám, és a H_0 hipotézis igaz, akkor mind az $F_n(x)$, mind a $G_n(y)$ ugyanazt az $F(x)$ elméleti eloszlásfüggvényt közelítvén, egymástól sem térhet el nagyon. A D_n^+ mennyiség nyilván valószínűségi változó, ha ennek eloszlását ismernénk, akkor meg tudnánk határozni olyan z_0 korlátot, hogy a $D_n^+ > z_0$ esemény valószínűsége kisebb legyen egy előírt α (pl. 0,05) mennyiségnél.

Szmírnov kimutatta, hogy nagy n érték esetén

$$P \left\{ \sqrt{\frac{n}{2}} \sup_x [F_n(x) - G_n(y)] < z \right\} \approx 1 - e^{-z^2}.$$

Más szóval annak valószínűsége, hogy a tapasztalt D_n^+ mennyiség $\sqrt{\frac{n}{2}}$ -szerese nagyobb legyen, mint egy z szám:

$$P\left(\sqrt{\frac{n}{2}} D_n^+ > z\right) \approx e^{-2z^2}$$

Nyilvánvaló, hogy ez a valószínűség annál kisebb, minél nagyobb z értéke. Ha tehát azt kívánjuk meghatározni, hogy mekkora z_0 esetén lesz ez a valószínűség kisebb, mint α , akkor az

$$e^{-2z_0^2} = \alpha$$

egyenletet kell megoldanunk, amiből

$$z_0 = \sqrt{-\frac{1}{2} \ln \alpha}$$

adódik.

Ha tehát a tapasztalt D_n^+ mennyiség nagyobb mint

$$\sqrt{\frac{2}{n}} \sqrt{-\frac{1}{2} \ln \alpha} = \sqrt{-\frac{\ln \alpha}{n}},$$

akkor a H_0 hipotézist α valószínűségi szinten elvetjük.

A D_n^+ mennyiség pontos eloszlását Gnedenko és Koroljuk szovjet matematikusok határozták meg, és eredményük a következő: Ha az és megfigyeléssorozatok azonos eloszlásból származnak, akkor

$$P\left(\sqrt{\frac{n}{2}} D_n^+ < z_0\right) = 1 - \frac{\binom{2n}{n-c}}{\binom{2n}{n}}$$

ahol $c = [z_0 \sqrt{2n}]$ (a $z_0 \sqrt{2n}$ számban foglalt legnagyobb egész szám.)

Ismertetjük a gondolatmenetet, amelynek segítségével a fenti eredmény származtatható.

$$\text{Az (I)} \quad x_1^*, x_2^*, \dots, x_n^*$$

$$\text{és (II)} \quad y_1^*, y_2^*, \dots, y_n^*$$

A rendezett mintából készítsünk most egyetlen

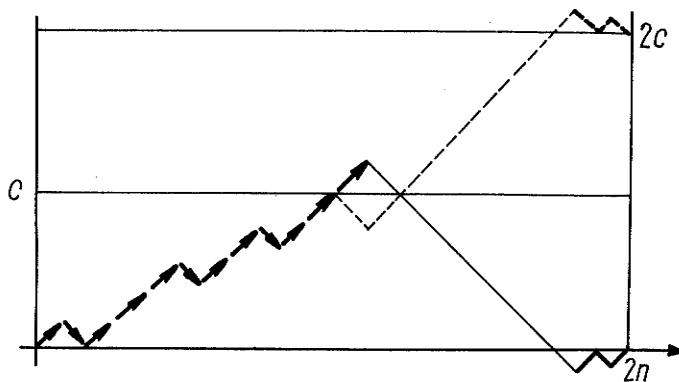
$$(III) z_1^*, z_2^*, \dots, z_{2n}^*$$

rendezett mintát, amelyben mindkét minta elemei együttesen szerepelnek, szigorúan nagyság szerint növekvő sorrendben. A (III) minta minden eleme alá írjuk oda, hogy melyik mintából származik, mégpedig úgy, hogy ha z_k^* az (I) mintából ered, írjunk alá $+1$ -et, ha a (II) mintából való, akkor írjunk alá (-1) -et. Ugyanigy járunk el mindegyik z_k^* mintaelemnél a (3) sorozatban.

Ezáltal egy $+1$ és -1 -ekből álló 2_n elemű sorozatot kaptunk. Jelöljük ezt a sorozatot

$$(IV) \psi_1, \psi_2, \dots, \psi_{2n} \text{ -nel, ahol } \psi_1 = \pm 1.$$

Egy ilyen $+1$ és -1 -ekből álló sorozatot grafikusán ábrázolhatunk a következő módon: egy derékszögű koordinátarendszerben az origóból indulva $+1$ esetén jobbra felfelé, (-1) esetén jobbra lefelé rajzolunk egy $\sqrt{2}$ egységnyi hosszúságú vektort:



74. ábra

A $\psi_1, \psi_2, \dots, \psi_{2n}$ sorozat elemei között n számú $+1$ és n számú -1 szerepel, mivel a (III) sorozat elemeinek fele az (I) fele pedig a (II) mintából származik.

Képezzük most a (IV) sorozat részletösszegeit:

$$S_0 = 0, S_1 = \psi_1, \dots, S_k = \psi_1 + \psi_2 + \dots + \psi_k$$

E részletösszegek közül a legnagyobb éppen a két tapasztalati eloszlásfüggvény közti legnagyobb eltérés n -szerese, azaz

$$D_n^+ = \sup_{0 \leq k \leq 2n} S_k$$

Tehát az általunk keresett valószínűség

$$P(D_n^+ < z \sqrt{2n}) = P(\sup S_k < c).$$

(Ugyanis, ha a tapasztalati eloszlásfüggvényt, $F_n(x)$ -et úgy rajzolnánk, hogy minden $x = \xi_k^*$ pontban $\frac{1}{n}$ helyett 1 ugrása van, akkor a D_n^+ mennyiség éppen $\sup S_k$ -val lenne egyenlő.)

Az (1) és (2) minták elemeinek véletlenszerű alakulásától függően a fenti eljárás mellett különböző $\varphi_1, \varphi_2, \dots, \varphi_{2n}$ sorozatokat kapnánk, amelyek mindegyikében persze csak +1 és -1 jegyek szerepelnek, és mindegyikben pontosan n számú +1 és -1 van, a legkülönbözőbb sorrendekben.

Mindegyik ilyen sorozatra igaz, hogy $S_{2n} = 0$, vagyis mindegyik lehetséges (n) sorozatnak olyan trajektória felel meg, amely a $(2n, 0)$ pontban végződik. Az összes lehetséges (n) sorozatok és nekik a megfelelő trajektoriák száma most $\binom{2n}{n}$, mivel n számú +1-es ennyiféle módon helyezkedhet el $2n$ számú helyen. Ha meg tudjuk határozni, hány olyan (n) sorozat van, amelyre $\sup S_k < c$, azaz hány olyan trajektória van, amely a $(2n, 0)$ pontban végződik és lefutása során nem éri el az $y = c$ egyenest, akkor egyszerűen megkapjuk a kérdéses valószínűséget.

Számítsuk ki, hogy a kérdéses trajektoriák közül hány olyan van, amely eléri az $y = c$ egyenest! Ha egy olyan trajektoriát, amely az $y = c$ egyenest eléri, az első elérési ponttól kezdve tükrözzük az $y = c$ egyenesre, olyan trajektoriát nyerünk, amely a $(2n, 2c)$ pontban végződik. Ha most ez a trajektória x -szer lépett felfelé és y -szor lefelé, akkor

$$x + y = 2n$$

$$y - x = 2c$$

$$x = n + c$$

$$y = n - c$$

Azon trajektoriák száma, amelyek az $y = c$ egyenest elérik:

$$\binom{2n}{n-c} = \binom{2n}{n+c}$$

Nyilván azon trajektoriák száma, amelyek az $y=c$ egyenest nem érik el:

$$\binom{2n}{n} - \binom{2n}{n-c}$$

Ezek alapján:

$$P(\sup S_k < c) = P(n D_n^+ < z_0 \sqrt{2n}) = \frac{\binom{2n}{n} - \binom{2n}{n-c}}{\binom{2n}{n}} = 1 - \frac{\binom{2n}{n-c}}{\binom{2n}{n}}$$

Azon hipotézisünk ellenőrzése céljából, hogy az $F_n(x)$ és $G_n(x)$ tapasztalati eloszlásfüggvények tekinthetők-e ugyanazon folytonos $F(x)$ eloszlásfüggvény közelítésének, azaz a két minta azonos eloszlásból származik-e, nem szükséges az $F_n(x)$ és $G_n(x)$ függvényeket ténylegesen megkonstruálnunk, hanem elkészítve a +1 és -1-ekből álló (IV) sorozatot és megrajzolva a neki megfelelő és a 74. ábrán látható típusú trajektoriát, azt kell csak megnéznünk, hogy a trajektoria metszi-e az alkalmasan választott $y = c_0$ egyenest, ahol $c = [z_0 \sqrt{2n}]$. Ha metszi, akkor a hipotézist elvetjük.

A probléma most már csak az, hogyan válasszuk a c számot ahhoz, hogy pl. 0,05-nél kisebb legyen a metszés valószínűsége.

A Stirling formula alkalmazásával belátható, hogy a

$$P(\sup S_k > c) = \frac{\binom{2n}{n-c}}{\binom{2n}{n}}$$

kifejezésben nagy n esetén

$$\frac{\binom{2n}{n-c}}{\binom{2n}{n}} \approx e^{-2z^2} \quad \text{ahol } [z \sqrt{2n}] = c.$$

Igy egyrészt látható, hogy $n \rightarrow \infty$ esetén

$$P(\sup S_k > c) = P(n D_n^+ > z \sqrt{2n}) = P\left(\sqrt{\frac{n}{2}} D_n^+ > z\right) = e^{-2z^2},$$

vagyis a Gnedenko-Koroljuk-féle eloszlás a Szmirnov-féle eloszlásba megy

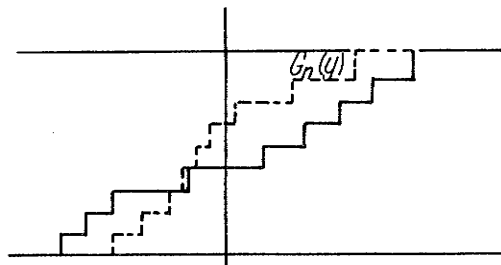
át, másrészt, mint láttuk, ha hipotézisünket α valószínűségi szinten akarjuk ellenőrizni, akkor Szmirnov eredménye alapján $z = \sqrt{-\frac{1}{2} \ln \alpha}$ választható.

Ha tehát hipotézisünket grafikus uton, a trajektoria megrajzolásával akarjuk ellenőrizni, akkor a

$$C_0 = [z_0 \sqrt{2n}] = [\sqrt{-n \ln \alpha}] \text{ magasságban huzott}$$

egyenes alkalmas kritikus vonal, amennyiben a trajektoria a c magasságu egyenest metszi, a hipotézist α szinten elvetjük.

A D_n^+ mennyiség az $F_n(x)$ és $G_n(y)$ tapasztalati eloszlásfüggvények között az $F_n(x)$ függvény javára mutatkozó maximális eltérést méri. Az



75. ábra

olvasóban felmerülhet a probléma, mit csinálunk, ha a két rendszerezett mintából készített lépcsős függvények közül $F_n(x)$ mindvégig $G_n(y)$ alatt fut. Ez a H_0 hipotézis fennállása mellett 0 valószínűségi esemény, ha tehát ezt tapasztalnánk, akkor H_0 -t elvetjük.

Megtörténhet viszont, hogy a két tapasztalati eloszlásfüggvény között nem mutatkozik nagy eltérés $F_n(x)$ javára, de nagy eltérés mutatkozik $G_n(y)$ javára, azaz pl. a két tapasztalati eloszlásfüggvény a 75. ábra szerinti.

Arra gondolhatunk, hogy ekkor egyszerűen felcseréljük a két minta szerepét, hiszen mindegy, hogy melyik eloszlásfüggvényt nevezem $F_n(x)$ -nek és melyiket $G_n(y)$ -nak, bennünket csak az érdekel, hogy egyáltalán milyen nagy az eltérés a két tapasztalati eloszlásfüggvény közötti maximális eltérés abszolút értékét, a $\sup_x |F_n(x) - G_n(y)|$ mennyiséget vizsgáljuk.

Jelöljük ezt a mennyiséget D_n -nel, tehát

$$D_n = \sup_x |F_n(x) - G_n(y)|$$

A D_n valószínűségi változó eloszlása természetesen nem ugyanaz, mint D_n^+ eloszlás, bonyolultabban írható fel, gyakorlati szempontból azonban ez nem sok nehézséget jelent.

Szmirnov kimutatta, hogy nagy n érték esetén D_n eloszlása aszimptotikusan az alábbi alakban írható:

$$P(n D_n > z \sqrt{2n}) \approx 2 e^{-2z^2} - 2 e^{-2^2 \cdot 2z^2} + 2 e^{-3^2 \cdot 2z^2} - \dots$$

A sor gyorsan konvergál, így gyakorlati célra elegendő az első néhány tagját figyelembe venni.

Megjegyezzük, hogy

$$\lim_{n \rightarrow \infty} P\left(\sqrt{\frac{n}{2}} D_n < z\right) = K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}, \text{ ha } z > 0,$$

ami nem más, mint a Kolmogorov-féle $K(z)$ függvény, így nagy n érték, pl. $n > 30$ esetén a két minta összehasonlítására a Kolmogorov-féle próba alkalmazható.

Megjegyezzük még, hogy abban az esetben, ha a két minta elemszáma nem azonos, mondjuk ξ -re vonatkozólag n , η -ra vonatkozólag m számu megfigyelésünk van, akkor megtehetjük ugyan, hogy $m > n$ esetén $m-n$ számu megfigyelési adatot figyelmen kívül hagyunk, hogy egyenlő legyen a minták elemszáma, ezáltal azonban esetleg értékes információt veszünk el. Célszerűbb úgy eljárni, hogy képezzük a

$$B_{n,m}^+ = \max_x [n F_n(x) - m G_m(y)]$$

statisztikát, ahol $F_n(x)$ a ξ -re vonatkozó, $G_m(y)$ az η -ra vonatkozó tapasztalati eloszlásfüggvényeket jelöli.

A $B_{n,m}^+$ statisztika pontos eloszlása:

$$P(B_{n,m}^+ = k) = \frac{2k+1+m-n}{m+k+1} \cdot \frac{\binom{m+n}{n-k}}{\binom{m+n}{n}}, \quad k = 0, 1, 2, \dots, n$$

Ez esetben határeloszlásként $n \rightarrow \infty$, $m \rightarrow \infty$, $\frac{m-n}{\sqrt{m+n}} \rightarrow 2c$ ($m \geq n$, $c \geq 0$), a következő összefüggést nyerjük:

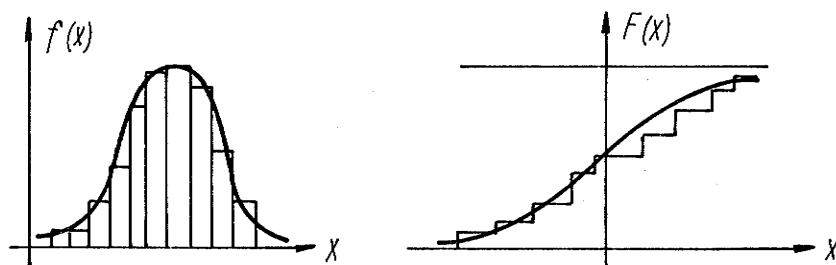
$$\lim_{\substack{m \rightarrow \infty \\ n \rightarrow \infty}} P(B_{n,m}^+ < z) = 1 - e^{-2z^2 - 4cz}$$

Ugyancsak ismeretes különböző mintaelemszámok esetén a kétoldalu maximális eltérés eloszlása. A kérdéskör részletei iránt érdeklődő olvasó számára utalunk a [14] dolgozatra.

10.3 További módszerek illeszkedésvizsgálata

a) Normalitás vizsgálat

Folytonos eloszlású ξ valószínűségi változó esetében, ha nagy mintával rendelkezünk, a legmegbízhatóbb illeszkedésvizsgálati módszer a Kolmogorov-próba. A gyakorlatban legtöbbször normális eloszlással állunk szemben, s annak eldöntése, hogy adatsorunk valóban normális eloszlásból származik-e, igen egyszerű grafikus módszereket is kidolgoztak, amelyek ugyan kevésbé megbízhatóak, de elsődleges tájékozódásra igen praktikusak. Ismertetjük ezért röviden a normalitásvizsgálatnak ezeket az egyszerű módszereit. A normális eloszlást vagy a sűrűségfüggvénye, vagy az eloszlásfüggvénye segítségével reprezentálhatjuk, ennek megfelelően vagy azt vizsgáljuk, mennyire illeszkedik a mintaelemekből konstruált hisztogram a normális eloszlás sűrűségfüggvényéhez, vagy azt, hogy mennyire illeszkedik a mintaelemekből készített tapasztalati eloszlásfüggvény a normális eloszlásfüggvényhez.



76. ábra

A legegyszerűbb és leggyakrabban alkalmazott módszer a tapasztalati eloszlás hisztogramjának összehasonlítása az elméleti eloszlás sűrűségfüggvényével. Először megnézhetjük, hogy a hisztogram eléggé szimmetrikus-e és hasonlít-e az alakja a normális eloszlás ismert sűrűségfüggvényének alakjához.

Pontosabb képet kapunk az illeszkedésről, ha a megfelelő normális eloszlás sűrűségfüggvényét rárajzoljuk a hisztogramra. Ilyenkor úgy járunk el, hogy elkészítjük a minta osztályba sorolt értékeinek relatív gya-

korisági hisztogramját, majd a vízszintes tengelyen felmérjük a minta szám-
tani közepét és ettől számítva mindkét irányban a korrigált empirikus szó-
rás többszöröseit, azaz feltüntetjük az

$$\bar{x}, \bar{x} \pm \frac{1}{2} s, \quad \bar{x} \pm s, \quad \bar{x} \pm 2s, \quad \bar{x} \pm 3s$$

abszcissza értékeket. Ezen abszcissza pontokban függőleges irányban a
következő ordinátaértékeket mérjük fel:

x:	y:	
$\bar{x}$	$0,40 \frac{h}{s}$	
$\bar{x} \pm \frac{1}{2} s$	$0,35 \frac{h}{s}$	(h az osztályköz hosszát, s a korrigált empirikus szórást jelöli.)
$\bar{x} \pm s$	$0,24 \frac{h}{s}$	
$\bar{x} \pm 2 s$	$0,05 \frac{h}{s}$	
$\bar{x} \pm 3 s$	$0,004 \frac{h}{s}$	

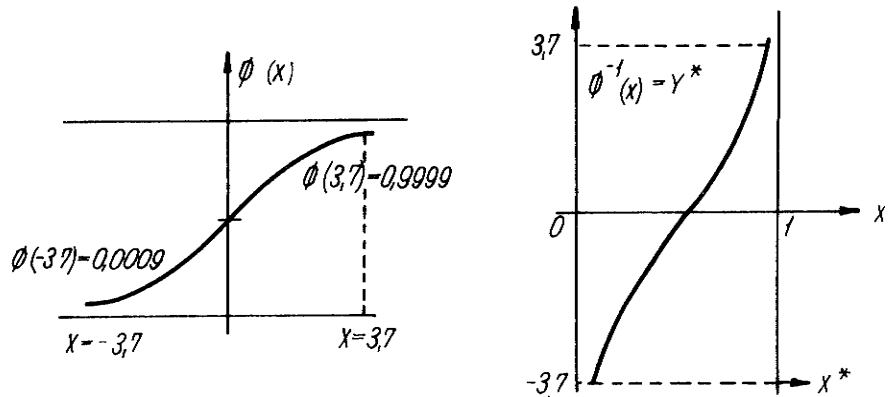
Ezek az ordináta értékek az elméleti sűrűségfüggvény értékei a meg-
felelő abszcissza pontokban. Az így nyert 9 pontot görbével összekötve a
normális eloszlás sűrűség-görbéjének közelítő képét kapjuk.

Arra vonatkozólag, hogy a sűrűség-görbe és a hisztogram között
mekkora eltérés engedhető meg, nehéz lenne gyakorlati támpontot adni, ez
erősen függ a minta elemszámától, az intervallumok számától stb. Ez az
eljárás csak elsődleges, közelítő tájékozódásra alkalmas. Megjegyezzük,
hogy ha más-más mintából sűrűség-görbe közti eltérésre vonatkozólag va-
lami alkalmas mérőszámot tudunk konstruálni, akkor ez a mérőszám egy
valószínűségi változó lesz, amelynek eloszlását kell ismernünk ahhoz, hogy
az észlelt eltérés alapján az illeszkedést (a normalitási hipotézist) elfogad-
hatónak vagy elvetendőnek ítélhessük. Később látni fogjuk (a χ^2 próba c.
pontban), hogyan kell eljárunk ahhoz, hogy adott valószínűségi szinten
dönthessünk hipotézisünk elfogadását vagy elvetését illetően.

Másik grafikus eljárás normalitás vizsgálatra, a tapasztalati elosz-
lásfüggvény és a normális eloszlás eloszlásfüggvénye közti illeszkedés
vizsgálatára irányul.

Ezt a célt szolgálja az ún. Gauss-papír (lásd: Melléklet), amely el-
sődleges tájékozódásra igen alkalmas. A Gauss-papír függőleges skálája a
standard normális eloszlás $\Phi(x)$ eloszlásfüggvényének inverze, a $\Phi^{-1}(x)$

függvény értékei szerint van skálázva, 0,0001-től 0,9999-ig, azonban ezen értékek 100-zal vannak szorozva, így 0,01-től 99,99-ig vannak a számok feltüntetve.



77. ábra

A vízszintes tengely egyenletes beosztású, de nincs skálázva, ide a ténylegesen megfigyelt mintaértékeket visszük fel. (Az origó a $(0, -3,7)$ pontban van elhelyezve.)

Ha Gauss-papíron egy $N(m; \sigma)$ normális eloszlású ξ valószínűségi változó $F(x)$ eloszlásfüggvényét ábrázoljuk, akkor egyenes vonalat kapunk.

Ugyanis a Gauss-papíron valamely x^* értékhez az $y = F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$ érték helyett az

$$y^* = \Phi^{-1}[F(x)] + 3,7 = \Phi^{-1}\left[\Phi\left(\frac{x-m}{\sigma}\right)\right] + 3,7 = \frac{x-m}{\sigma} + 3,7$$

ordinátát mérjük, ahol $x = \lambda x^* - 3,7$ (A λ számot úgy választjuk, hogy az $x^* = \frac{1}{\lambda}(x + 3,7)$ transzformációt a $(-3,7, 3,7)$ intervallumot a $(0,1)$ intervallumba vigye át.

$$\text{Ezek szerint } y^* = \frac{\lambda x^* - 3,7 - m}{\sigma} + 3,7 = \frac{\lambda}{\sigma} x^* + \left(3,7 - \frac{3,7+m}{\sigma}\right),$$

ami nyilván egyenes egyenlete.

Ha most a ξ valószínűségi változóra, amelyről feltesszük, hogy normális eloszlású, n számú megfigyelést végzünk:

$$x_1, x_2, \dots, x_n$$

és minden x_i mintaelemhez kiszámítjuk az $F_n(x_i)$ tapasztalati eloszlásfüggvény-értéket, akkor mivel $F_n(x_i) \approx F(x)$, az $[x_i, F_n(x_i)]$ pontok a Gauss-papíron közelítőleg egyenes vonalon helyezkednek el.

Megjegyezzük, hogy ez a grafikus módszer is elsősorban nagy mintaelemszám esetén alkalmazható (mert ekkor lesz kicsi az eltérés az $F(x)$ elméleti és $F_n(x)$ tapasztalati eloszlásfüggvény között).

Amennyiben az egyeneshez való illeszkedés nem látszik megnyugtatónak, finomabb módszerekhez fordulunk. (Kolmogorov-próba, χ^2 -próba).

b) A χ^2 próba

Abban az esetben, ha a vizsgált ξ valószínűségi változó diszkrét eloszlású valamely $x_1, x_2, \dots, x_n, \dots$ lehetséges értékekkel, amelyeket rendre $p_1, p_2, \dots, p_n, \dots$ valószínűségekkel vesz fel, (pl. diszkrét egyenletes eloszlást, vagy pl. Poisson-eloszlást feltételezhetünk), akkor az eddig tárgyalt módszerekkel nem tudunk illeszkedésvizsgálatot végezni, mert azok csak folytonos eloszlások esetében alkalmazhatók. Ilyenkor az illeszkedés-vizsgálatot ún. χ^2 próbával végezzük. Megjegyezzük, hogy a χ^2 próba folytonos eloszlás esetén is alkalmazható illeszkedésvizsgálatra. Ennek módját is ismertetjük a normális eloszlás esetére.

Vizsgáljuk először, hogy végzünk illeszkedésvizsgálatot diszkrét egyenletes eloszlás esetén. Ekkor a ξ valószínűségi változónak csak véges n számú lehetséges értéke van és mindegyik értékét $\frac{1}{n}$ valószínűséggel veszi fel. Könnyebb megértés kedvéért az $n = 6$ esetet választjuk, amelynek szemléletes modellje a kockadobás. Gondolatmenetünk tetszőleges véges n kimenetel esetére alkalmazható, a megfelelő formulákat fel fogjuk írni.

Tegyük fel, hogy egy játékkockát dobálunk. Minden egyes kísérletnek (dobásnak) hat lehetséges kimenetele van: $x_1 = 1, x_2 = 2, \dots, x_6 = 6$.

Abban az esetben, ha a kocka anyagában homogén és szabályos, akkor bármelyik szám egyforma valószínűséggel következik be minden egyes dobás során. Más szóval a kockadobáshoz rendelt ξ valószínűségi változó lehetséges értékei: 1, 2, 3, 4, 5, 6, s ezeket rendre $p_1 = p_2 = \dots = p_6 = \frac{1}{6}$ valószínűséggel veszi fel. N dobás során mindegyik szám körülbelül $\frac{N}{6}$ -szor fog bekövetkezni. (Amikor ugyanis azt vizsgáljuk, hogy pl. az 5-ös szám előfordulásának mennyi a várható értéke, akkor A-val jelölve azt

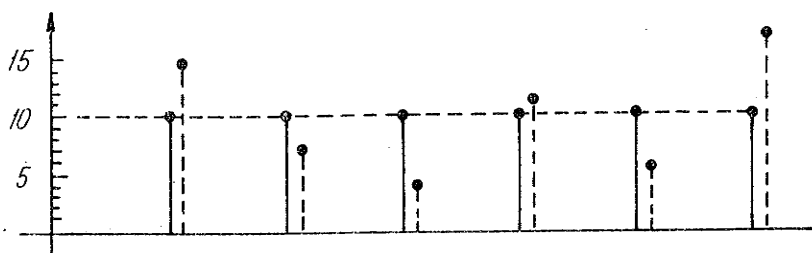
az eseményt, hogy a dobás eredménye nem 5-ös $P(A) = \frac{1}{6} = p$, $P(\bar{A}) = \frac{5}{6} = q$,
s mivel binomiális eloszlással állunk szemben, a várható érték:
 $Np = \frac{N}{6}$. Ugyanez érvényes az 1, 2, 3, 4, 6 számok mindegyikére.)

A kocka szabályos voltára vonatkozó H_0 hipotézisünket statisztikailag
ugy ellenőrizhetjük (mondhatnánk úgy is, hogy a kocka szabályos és homo-
gén voltát statisztikailag mérjük), hogy nagyszámu dobást végzünk vele és
megvizsgáljuk, hogy az egyes számok előfordulási gyakorisága elég egyen-
letes eloszlású-e. Tegyük fel, hogy valaki 60 dobást végezve a kockával,
az egyes számokra az alábbi gyakoriságokat kapta:

	1	2	3	4	5	6
Megfigyelt gyakoriság:	x_1 15	x_2 7	x_3 4	x_4 11	x_5 6	x_6 17
elméleti vár- ható érték:	np_1 10	np_2 10	np_3 10	np_4 10	np_5 10	np_6 10
eltérés	5	-3	-6	1	-4	7

Megfelelnek-e a kapott gyakoriságok várakozásunknak? Az a tény,
hogy az 1, 2, ..., 6 számok bármelyike nem pontosan 10-szer következik
be, egészen természetes a véletlen ingadozások miatt. Kérdés azonban,
hogy a ténylegesen tapasztalt eltérések összességükben tulajdoníthatók-e
pusztán a véletlen szerepének.

Ha felrajzoljuk a kockadobáshoz rendelt ξ valószínűségi változó el-
méleti gyakorisági diagramját, valamint a megfigyelt gyakoriságok diagram-
ját (78. ábra), s összehasonlítjuk őket,



78. ábra

azt kell megállapítanunk, hogy a két diagram nem nagyon hasonlít egymás-
hoz, s a kocka szabályos voltát illetően kétségeink támadnak.

Mennyire térhet el a két diagram egymástól ahhoz, hogy még elfogadjuk H_0 hipotézisünket?

Hogy erre a kérdésre válaszolhassunk, alkalmas mérőszámot kell konstruálnunk a két diagram eltérésére vonatkozólag. Egy ilyen mértékszám, amely természetesen ugyancsak a véletlentől függő mennyiség, különféle jó statisztikai tulajdonsággal kell rendelkeznie. Megkivánjuk tőle, hogy viszonylag könnyen kiszámítható legyen, ismert eloszlással bírjon stb.

K. Pearson az elméletileg várható Np_i gyakoriságok és a megfigyelt x_i gyakoriságok közötti eltérés mérésére az alábbi mennyiséget vezette be:

$$(10.3.1) \quad \chi^2 = \sum_{i=1}^n \frac{(x_i - Np_i)^2}{Np_i}$$

Példánkban ez a mennyiség:

$$\begin{aligned} \chi^2 &= \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(4-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} = \\ &= \frac{1}{10} (5^2 + 3^2 + 6^2 + 1^2 + 4^2 + 7^2) = \frac{1}{10} (25 + 9 + 36 + 1 + 16 + 49) = 13,6 \end{aligned}$$

A fenti χ^2 mennyiségben az egyes összeadandók azt fejezik ki, hogy a megfigyelt és az elméletileg várható gyakoriságok közötti különbség négyzete mekkora hányada a várható gyakoriságnak. Ha az így konstruált χ^2 mennyiség nem túlságosan nagy, akkor a kísérlet nem mond ellent a H_0 hipotézisnek.

A kapott $\chi^2 = 13,6$ eredmény értékeléséhez szükséges lenne tudnunk, hogy a ténylegesen szabályos kockával nagyon sokszor elvégeznénk a 60 dobásból álló kísérletsorozatot, s minden kísérletsorozathoz kiszámítanánk a nyert gyakoriságok alapján a megfelelő χ^2 mennyiséget, akkor a sorozatok hány százalékánál kapnánk χ^2 -re 13,6 vagy ennél nagyobb értéket. Ha ez a százalék kicsiny, elvetnénk a H_0 hipotézist. Ismernünk kellene tehát a $P\{\chi^2 \geq 13,6\}$ esemény valószínűségét, vagy másként kifejezve a fenti χ^2 mennyiség valószínűségeloszlását.

A (10.3.1) összefüggéssel definiált χ^2 mennyiség természetesen diszkrét eloszlású valószínűségi változó, eloszlása azonban jól közelíthető egy folytonos eloszlással az ún. χ^2 -eloszlással, hasonlóan, mint ahogy a binomiális eloszlást nagy N esetén közelíteni tudtuk normális eloszlással.

Vegyük tekintetbe, hogy kockadobásnál az 1, 2, ..., 6 számok gyakoriságai az $x_1, x_2, \dots, x_6$ mennyiségek binomiális eloszlásuak.

Legyen most

$$z_i = \frac{x_i - Np_i}{\sqrt{Np_i}} \quad (i=1, 2, \dots, 6)$$

Ekkor a z_i változók (ugyancsak binomiális eloszlásuak) nagy számú kísérlet esetén közelítőleg normális eloszlásuak.

A (10.3.1) összefüggés alapján

$$\chi^2 = \sum_{i=1}^6 \frac{(x_i - Np_i)^2}{Np_i} = \sum_{i=1}^6 z_i^2$$

A χ^2 mennyiség tehát felfogható, mint normális eloszlásu valószínűségi változók négyzetösszege, ahol a z_i változók között egy esetünkben igen egyszerű lineáris összefüggés áll fenn:

$$\sum_{i=1}^6 z_i = 0.$$

Ugyanis:

$$z_1 + z_2 + \dots + z_6 = \frac{1}{\sqrt{Np}} \sum_{i=1}^6 (x_i - Np) = \frac{1}{\sqrt{Np}} (N - N) = 0$$

mert

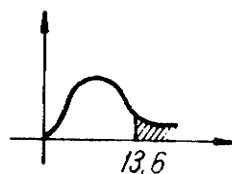
$$x_1 + x_2 + \dots + x_6 = N \quad \text{és} \quad \sum_{i=1}^6 Np = 6 \frac{N}{6} = N.$$

A χ^2 eloszlás tárgyalásánál mondottak alapján (lásd 5.14/a pont) a (10.3.1) összefüggésben szereplő χ^2 mennyiség eloszlása közelíthető az 5. szabadságfokú χ^2 eloszlással, amelynek sűrűségfüggvénye:

$$k_5(x) = \frac{x^{\frac{5}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{5}{2}} \Gamma(\frac{5}{2})}$$

E sűrűségfüggvény görbéje a 79. ábrán látható. Annak valószínűsége, hogy kockadobásnál $\chi^2 \geq 13,6$ legyen a függvény alatt 13,6-tól jobbra eső terület mértékszámaival egyenlő, azaz

$$P(\chi^2 \geq 13,6) = \int_{13,6}^{\infty} k_3(x) dx = 1 - k_5(13,6)$$



79. ábra

ahol

$$k_5(x) = \int_0^x k_5(t) dt$$

A χ^2 eloszlás táblázatát használva (I. táblázat) megállapítható, hogy 5 szabadságfoku χ^2 -eloszlás esetén annak valószínűsége, hogy $\chi^2 \geq 13,6$ legyen, kisebb, mint 0,02. Mivel szabályos kocka mellett $\chi^2 \geq 13,6$ igen kicsiny valószínűségű, a H_0 hipotézist elvetjük.

A tárgyalt példában követett gondolatmenetet alkalmazzuk tetszőleges diszkrét eloszlású véletlen mennyiségre vonatkozó tiszta illeszkedésvizgálatra.

Ha egy N elemű megfigyeléssorozatban a ($\sum x_i = N$),

$$k_1, k_2, \dots, k_n$$

értékek rendre

$$x_1, x_2, \dots, x_n$$

gyakoriságokkal fordultak elő, s az a H_0 hipotézisünk, hogy a sokaság, amelyből megfigyeléseink (kísérletünk) által mintát vettünk, olyan diszkrét eloszlású, amelyben a k_1 érték bekövetkezésének p_1 , a k_2 értéknek $p_2, \dots$, a k_n -értéknek p_n a valószínűsége, akkor a H_0 hipotézisünk helyességének ellenőrzése céljából a kísérleti adatok alapján kiszámítjuk a

$$\chi^2 = \frac{(x_1 - Np_1)^2}{Np_1} + \dots + \frac{(x_n - Np_n)^2}{Np_n} = \sum_{i=1}^n \frac{(x_i - Np_i)^2}{Np_i}$$

mennyiséget.

(Megjegyezzük, hogy számítási célra a χ^2 mennyiség egyszerűbb alakra is hozható):

$$\sum_1^n \frac{(x_i - Np_i)^2}{Np_i} = \sum_1^n \frac{x_i^2 - 2x_i Np_i + N^2 p_i^2}{Np_i} =$$

$$= \sum_1^n \frac{x_i^2}{Np_i} - 2 \sum_1^n x_i + N \sum_1^n p_i = \sum_1^n \frac{x_i^2}{Np_i} - N$$

A (10.3.1) alatti χ^2 mennyiség eloszlását $n-1$ szabadságfoku χ^2 -eloszlással közelítve a χ^2 eloszlás táblázatából megkeressük azt az x mennyiséget, amelyre

$$P(\chi^2 \geq x_\varepsilon) = 1 - k_{n-1}(x_\varepsilon) = \varepsilon,$$

ahol ε egy megadott kis szám, $\varepsilon > 0$. Ha a számított χ^2 mennyiség értéke nagyobb, mint x_ε , akkor ez azt jelenti, hogyha hipotézisünk igaz lenne, akkor olyan esemény következne be, amelynek valószínűsége kisebb, mint ε . Mivel igen "valószínűtlen" esemény következett be, elvetjük a H_0 hipotézist. Ha viszont a (10.3.1) alatti χ^2 mennyiség értéke kisebb, mint x_ε , akkor a kísérlet nem szól a H_0 hipotézis ellen.

Becsléses illeszkedésvizsgálat : χ^2 -próbával

Abban az esetben, ha ismerjük azon elméleti eloszlás paramétereit, amelyre H_0 hipotézisünk vonatkozik, akkor ezeket a minta (megfigyelés sorozat) adataiból becsüljük, s ezután becsléses illeszkedésvizsgálatot végezhetünk.

Ha pl. feltételezzük, hogy megfigyeléssorozatunk adatai egy normális eloszlású sokaságból származnak, s megfigyelésünk eredményei a $\xi_1, \xi_2, \dots, \xi_n$ számok, akkor a feltételezett normális eloszlás várható értékét a mintaközéppel, szórásnégyzettel becsüljük, azaz az eloszlás paramétereinek az

$$m \sim \bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_N}{N}$$

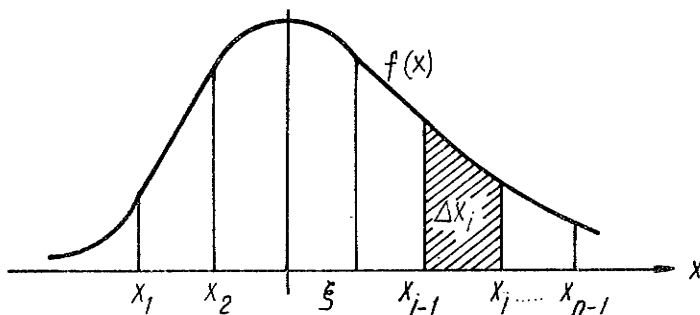
$$\sigma^2 \sim s_{N-1}^2 = \frac{1}{N-1} \sum_{i=1}^N (\xi_i - \bar{\xi})^2$$

menntiségeket tekintjük, s felállítjuk azt a H_0 hipotézist, hogy mintánk az

$$(10.3.2) \quad f(x) = \frac{1}{\sqrt{2} S_{N-1}} e^{-\frac{(x - \xi)^2}{2 S_{N-1}^2}}$$

sűrűségfüggvénnyel bíró normális eloszlású sokaságból származik. Jelöljük ezt a sokaságot reprezentáló valószínűségi változót ξ -vel.

Az illeszkedésvizsgálatot ezután a következő módon végezzük. ξ értéktartományát valamilyen ésszerű módon n számú intervallumra osztjuk (a gyakorlatban n értékének 10-15 körüli számot célszerű venni), s a normális eloszlás táblázata segítségével megállapítjuk, hogy mi a valószínűsége annak, hogy ξ értéke az egyes intervallumokba essék.



80. ábra

Pl. annak valószínűsége, hogy ξ értéke a Δx_i intervallumba essék (az $f(x)$ sűrűségfüggvény alatti területnek a Δx_i intervallum fölé eső része):

$$F(x) - F(x_{i-1}) = \tilde{p}_i.$$

Ha a sokaság, amelyből mintánk származik, valóban a (10.3.2) alatti sűrűségfüggvényű normális eloszlással bír, azaz ha a H_0 hipotézis igaz, akkor annak valószínűsége, hogy ξ értéke a Δx_i intervallumba essék, éppen $\tilde{p}_i$, és annak várható száma, ahányszor N kísérlet során ξ értéke Δx_i -be fog esni: $N\tilde{p}_i$.

Most megszámoljuk, hogy a megfigyelt $\xi_1, \xi_2, \dots, \xi_n$ értékek

közül hány darab esik az egyes intervallumokba. Tegyük fel, hogy a Δx_i intervallumba ezen értékek közül m_i számú esik ($i = 1, 2, \dots, n$; $x_0 = -\infty, x_n = \infty$) és megalkotjuk a

$$\chi^2 = \sum_{i=1}^n \frac{(m_i - N\tilde{p}_i)^2}{Np_i}$$

mennyiséget.

Az így számított χ^2 mennyiség most is χ^2 eloszlású, a szabadságfokok száma azonban kevesebb, mint a tiszta illeszkedésvizsgálatnál, mert a $\tilde{p}_i$ valószínűségek most $\tilde{\xi}$ -től és S_{n-1}^2 -től függenek. A szabadságfokok száma ezért $n - 1 - 2 = n - 3$. (Elméletileg kimutatható, hogy becsléses illeszkedésvizsgálatnál, amennyiben a számított χ^2 mennyiség n -számu komponensből áll, a közelítő χ^2 eloszlás szabadságfokainak száma $n-1$ -nél annnyival kevesebb, amennyi a becsült paraméterek száma.)

Megjegyezzük, hogy az intervallumok hosszának megválasztásánál arra kell ügyelnünk, hogy $N\tilde{p}_i \geq 10$ legyen.

A fent leírtak során egyben igyekeztünk érzékelteni, hogy folytonos eloszlás esetében is használható egy alkalmasan számított χ^2 mennyiség illeszkedésvizsgálatra. Meg kell azonban jegyeznünk, hogy folytonos eloszlások esetén eljárásunk az osztópontok önkényes megválasztása miatt kevésbé megbízható, mint diszkrét eloszlás esetében.

Folytonos eloszlásra vonatkozó hipotézisek ellenőrzésére a 10. fejezetben található módszerek megbízhatóbbak.

10.4 A minta véletlenségének vizsgálata

Mint az előző fejezetekben láttuk, a matematikai statisztikai vizsgálatokhoz kiindulásul a vizsgált ξ valószínűségi változóra vonatkozó n számú megfigyelési adat, a $\xi_1, \xi_2, \dots, \xi_n$ minta szolgál. A mintaelemokről mindvégig feltettük, hogy függetlenek és azonos eloszlásúak. Röviden ismertetünk most egy eljárást, amelynek segítségével ellenőrizhetjük, hogy teljesül-e ez a feltevésünk. Az eljárást Wald és Wolfowitz dolgozták ki. Feltételezzük, hogy a vizsgált ξ valószínűségi változó folytonos $F(x)$ eloszlású.

A $\xi_1, \xi_2, \dots, \xi_n$ mintáról azt mondjuk, hogy véletlen minta, ha a

mintaelemek teljesen függetlenek és azonos eloszlásuak. Azt, hogy a véletlenség feltétele teljesül-e, a következő módon ellenőrizhetjük. Képezzük a mintaelemek számtani közepét, a

$$\bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}$$

menyiséget és vonjuk le a mintaközepet mindegyik mintaelemből, azaz alkossuk meg a

$$\xi'_1 = \xi_1 - \bar{\xi}, \xi'_2 = \xi_2 - \bar{\xi}, \dots, \xi'_n = \xi_n - \bar{\xi}$$

valószínűségi változókat, ami lényegében csak az origó speciális megválasztását jelenti. (Az origót a $\bar{\xi}$ pontba helyezzük.)

Képezzük most a következő statisztikai függvényt:

$$R = \sum_{i=1}^{n-1} \xi'_i \xi'_{i+1} + \xi'_n \xi'_1.$$

R nyilván valószínűségi változó, mint minden statisztikai függvény. Wald és Wolfowitz kimutatták, hogy

$$M(R) = - \frac{\sum_{i=1}^n \xi_i^2}{n-1},$$

szórásnégyzete:

$$D^2(R) = \frac{\left(\sum_{i=1}^n \xi_i^2\right)^2}{n-1} - \sum_{i=1}^n \xi_i^4 + \frac{\left(\sum_{i=1}^n \xi_i^2\right)^2}{(n-1)(n-2)} - 2 \sum_{i=1}^n \xi_i^4 - \frac{\left(\sum_{i=1}^n \xi_i^2\right)^2}{(n-1)^2}$$

Az említett szerzők ugyancsak kimutatták, hogy az R valószínűségi változó szimptotikusan normális eloszlású, azaz $n \rightarrow \infty$ esetén:

$$P\left(\frac{R - M(R)}{D(R)} < x\right) \Rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

Ennek alapján, ha n elég nagy ahhoz, hogy határeloszlást használhassunk, ($n \gg 30$), akkor ha

$$R^* = \frac{R - M(R)}{D(R)}$$

értéke kívül esik a $(-2, +2)$ intervallumon, el kell vetnünk a véletlenségre vonatkozó hipotézisünket.

10.5 Cramér-Mises próba

A Kolmogorov-próba esetében az $F_n(x)$ empirikus eloszlásfüggvénynek az $F(x)$ elméleti eloszlásfüggvénytől való távolságát a két függvény közötti maximális eltéréssel (illetve annak abszolút értékével) mértük, s a H_0 hipotézist, amely szerint a vizsgált ξ valószínűségi változó eloszlásfüggvénye $F(x)$, akkor fogadtuk el, ha a fenti módon mért távolság $\sqrt{n}$ -szere elég kicsinek adódott.

Két függvény távolságának egy másik szokásos mérőszáma, a két függvény közötti különbség négyzetének integrálja. Tekintettel arra, hogy az $F_n(x)$ empirikus eloszlásfüggvény értéke minden rögzített x -re valószínűségi változó, következésképp valószínűségi változó az $F_n(x)$ mennyiség is. Tételezzük fel, hogy a bennünket érdeklő ξ valószínűségi változó eloszlásfüggvénye $F(x)$. Legyen ez a H_0 hipotézis. Defináljuk az η valószínűségi változót az

$$\eta = n \left[F_n(x) - F(x) \right]^2$$

mennyiséggel, vagyis az empirikus és elméleti eloszlásfüggvények négyzetes eltérésének n -szerezésével. A Cramér - Mises próba alkalmazásakor az

$$(10.5.1) \quad M(\eta) = W_n^2 = n \int_{-\infty}^{\infty} \left[F_n(x) - F(x) \right]^2 dF(x)$$

mennyiséget (a két függvény közötti négyzetes eltérés n -szerezésének várható értékét) használjuk annak mérésére, hogy mennyire tér el az empirikus eloszlásfüggvény az elméleti eloszlásfüggvénytől. Minél kisebb $M(\eta)$, annál jobban illeszkedik a minta a feltételezett elméleti eloszláshoz.

A (10.5.1) formula gyakorlati számítások céljára a

$\xi_1^* = x_1, \xi_2^* = x_2, \dots, \xi_n^* = x_n$ rendezett mintaelemek figyelembevételével a következő alakban írható:

$$(10.5.2) \quad W_n^2 \approx \frac{1}{12n} + \sum_{i=1}^n \left[F(x_i) - \frac{i-1}{2n} \right]^2$$

Ha választunk egy $0 < \alpha \leq \frac{1}{2}$ mennyiséget (a gyakorlatban $\alpha = 0,05$ vagy $\alpha = 0,001$ szokásos), akkor minden n -re meghatározható egy $w(n, \alpha)$ szám úgy, hogy a H_0 hipotézis fennállása esetén

$$P(W_n^2 > w(n, \alpha)) \leq \alpha$$

teljesüljön.

A H_0 hipotézist $w_n^2 > w(n, \alpha)$ esetén $1 - \alpha$ szinten elvetjük. A megfelelő $w(n, \alpha)$ számokat Walsh határozta meg.

A w_n^2 valószínűségi változó határeloszlását Owen, Anderson és Darling határozták meg. A Cramér-Mises próba ereje lényegesen nagyobb, mint a χ^2 próbáé.

Kétmintás Cramér-Mises próba

A Szmirnov-próbához hasonlóan a Cramér-Mises próba is használható annak eldöntésére, hogy két minta azonos folytonos eloszlású statisztikai sokaságból származik-e.

Tegyük fel, hogy $\xi_1, \xi_2, \dots, \xi_n$ minta a folytonos $F(x)$ eloszlásfüggvénnyel bíró ξ valószínűségi változóra vonatkozó megfigyeléssorozat, az $\eta_1, \eta_2, \dots, \eta_m$ minta a folytonos $F(x)$ eloszlásfüggvényű η valószínűségi változóra vonatkozik.

A $H_0 : F(x) \equiv G(x)$ hipotézis ellenőrzésére most a

$$(10.5.3) \quad W_N = \frac{m \cdot n}{m+n} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 dK_N(x)$$

statisztikát használjuk, ahol $N = m + n$, és

$$K_N(x) = \frac{1}{N} [n F_n(x) + m G_m(x)]$$

Mivel $F_n(x)$ és $G_m(x)$ lépcsős függvények, $dK_N(x)$ minden olyan x pontban, hol egyik lépcsőfüggvénynek sincs ugrása, így a (10.5.3) formulában szereplő integrál összegezéssel számítható.

A W_N statisztika határeloszlását (nagy n és m értékek esetén használható) Kiefer határozta meg és táblázta. Abban az esetben, ha W_N bizonyos kritikus mennyiségnél nagyobbak adódik, elvetjük a H_0 hipotézist.

HIBAELEMÉLET

11.1 Közvetlen megfigyelések

Valamely fizikai mennyiség meghatározása, mérése folyamán minden egyes mérési eredményt (megfigyelést) mérési hiba terheli. Ha ugyanarra a mennyiségre nézve ismételt méréseket végzünk, akkor az eredmény általában mérésről-mérésre változik. A hibaelmélet - amelynek eredete Gausshoz és Laplacehoz nyulik vissza - olyan módszereket dolgoz ki, amelyek segítségével a vizsgált fizikai mennyiség numerikus értéke megbecsülhető a kapott mérési eredményekből és amelyek lehetővé teszik számunkra, hogy a kapott becslés pontosságára bizonyos hibahatárokat adjunk meg.

A mérések alkalmával fellépő hibákat általában két csoportra szokták osztani, a szisztematikus rendszeres hibák és a véletlen hibák csoportjába.

Szisztematikus hibának általában a mérőberendezésben vagy a megfigyelési rendszerben rejlő hibákat nevezik. Ilyen pl., ha egy fémdarab súlyát akarjuk meghatározni egy mérleg segítségével, s a használt súlyszorozat egyes súlyai a szabályos súlyoknál könnyebbek. Ilyenkor minden mérési eredmény torzított, s a torzítottság arányos a súlyok hibájával. (A fémdarab súlyát a valóságosnál nagyobbak mérjük.) A szisztematikus hibák a mérőberendezés vagy a megfigyelési módszer gondos elemzésével, valamint ellenőrzésével általában kiküszöbölhetők, s a továbbiakban felteszszük, hogy méréseinket szisztematikus hibák nem terhelik.

Akármilyen mérési, ill. megfigyelési módszerhez folyamodunk, mindig számolnunk kell véletlen hibák fellépésével, amelyek okait nem tudjuk pontosan számításba venni. Ilyen okok lehetnek többek között a berendezés előre nem látható rezgései (pl. villamos megy el az épület előtt), a hőmérséklet ellenőrizhetetlen változásai, leolvasási hibák szemünk elfáradása következtében, gyenge légáramlás stb. A véletlen hibák általában számos olyan tényező következtében állnak elő, amelyek egyenként csekély mértékben hatnak ugyan, de hatásuk összegeződik. Az így fellépő véletlen hibákat valószínűségi változóknak tekintjük, s valószínűségelméleti módszerekkel tárgyaljuk őket.

A centrális határeloszlás-tételből tudjuk, hogy normális eloszlás lép fel minden olyan esetben, amikor nagyszámu, külön-külön kis mértékben ható véletlen tényező összegezett hatásáról van szó, és ezek a ténye-

zők egymástól függetleneknek tekinthetők. Ugy fogjuk fel, hogy az egyes véletlen tényezők okozta hibák $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók, amely ξ_1 változók összességüket tekintve (legalábbis jó közelítésben) függetlenek, s az együttes hatás eredményeként fellépő hiba, amelyet a ξ valószínűségi változóval jelölünk, a $\xi_1, \xi_2, \dots, \xi_n$ változók összege, azaz:

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n$$

Itt a ξ normális eloszlású valószínűségi változó. A mérendő mennyiség tényleges nagyságát a ξ valószínűségi változó m várható értékével, a mérőeszköz pontosságát pedig a ξ változó σ szórásával azonosítjuk.

Ezek alapján a mért fizikai mennyiség m nagyságát (numerikus értékét) n számú mérés elvégzése után a kapott

$$\xi_1, \xi_2, \dots, \xi_n$$

mérési eredményekből becsüljük.

$$\text{Kiszámítjuk a } \bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i, \text{ valamint az } S^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$$

mennyiségeket és mint az előzőekben () láttuk, t_0 alkalmas választásával

$$\bar{\xi} \pm t_0 \frac{s}{\sqrt{n-1}}$$

az m -re kívánt pontosságu konfidencia-határokat szolgáltat, ahol a t_0 értéket a Student-eloszlás táblázatból keressük ki (III. táblázat) az $n-1$ szabadságfoknak megfelelő sorból.

Ha a nagy, pl. $n > 30$, akkor adott pl. 95%-os megbízhatósági szintet véve követelménynek, a konfidencia határokat a normális eloszlás táblázata segítségével is megállapíthatjuk.

$$\left(\bar{\xi} - 2 \frac{s}{\sqrt{n}}, \bar{\xi} + 2 \frac{s}{\sqrt{n}} \right)$$

intervallum 95%-os megbízhatósági intervallum.

Ha ugyanis a vizsgált ξ valószínűségi változó $N(m; \sigma)$ eloszlású, ahol m értéke (a mért fizikai mennyiség numerikus értéke) ismeretlen, a

σ szórását tételezzük fel ismertnek (a gyakorlatban rendszerint ez a helyzet) és akkor a $\bar{x}$ mintaközép $N(m, \frac{\sigma}{\sqrt{n}})$ eloszlású, tehát

$$P(m - \lambda \frac{\sigma}{\sqrt{n}} < \bar{x} < m + \lambda \frac{\sigma}{\sqrt{n}}) = 2\Phi(\lambda) - 1$$

A zárójelben álló esemény mindannyiszor bekövetkezik, valahányszor a $\bar{x}$ középpontu $(\bar{x} - \lambda \frac{\sigma}{\sqrt{n}}, \bar{x} + \lambda \frac{\sigma}{\sqrt{n}})$ intervallum lefedi az m ismeretlen (egyébként fix állandó) értéket, azaz

$$P(\bar{x} - \lambda \frac{\sigma}{\sqrt{n}} < m < \bar{x} + \lambda \frac{\sigma}{\sqrt{n}}) = 2\Phi(\lambda) - 1$$

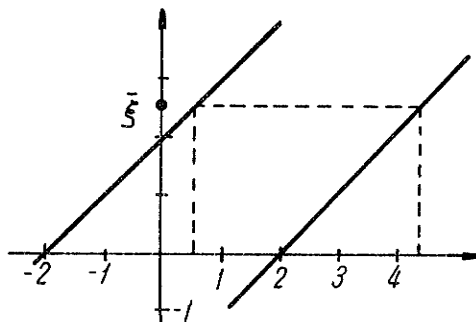
$\lambda = 2$ esetén $2\Phi(2) - 1 \approx 0,95$, tehát a fenti intervallum 95%-os konfidencia intervallum. Ennek grafi-

kus megszerkesztését ($\frac{\sigma}{\sqrt{n}} = 1$

egységet véve) a 82. ábra mutatja.

Ha σ értéke ismeretlen és σ^2 -et az $\frac{n}{n-1} S^2$ tapasztalati szórásnégyzettel becsüljük, akkor mivel a $\bar{x}$ szórása

$\sqrt{\frac{\sigma^2}{n}}$ ennek becsléseként az $\frac{s}{\sqrt{n-1}}$ mennyiségeket kapjuk, ez pedig nagy n esetében nem tér el lényegesen $\frac{s}{\sqrt{n}}$ -tól.



81. ábra

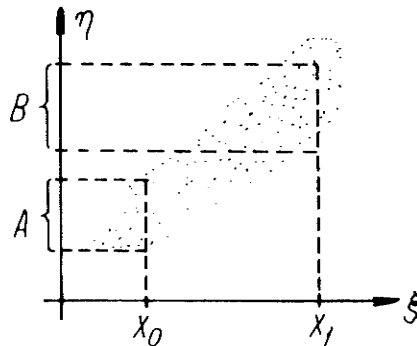
KORRELÁCIÓ- ÉS REGRESSZIÓ ANALÍZIS

12.1 Valószínűségi változók közötti sztochasztikus kapcsolatokról

A természettudományokban és a technikában gyakran találkozunk funkcionális összefüggéssel, amelynél egy bizonyos fizikai mennyiség egy vagy több másik mennyiség egyértékű függvényeként jelenik meg, azaz

$$y = f(x_1, x_2, \dots, x_n)$$

alakú függvénykapcsolattal állunk szemben, tehát az $x_1, x_2, \dots, x_n$ változók értékeinek ismeretében y értéke egyértelműen meghatározható. Ilyen funkcionális összefüggés valószínűségi változók között is fennállhat. Ha a ξ és η valószínűségi változók között funkcionális kapcsolat van: $\eta = f(\xi)$, akkor tehát ξ értékének megismerése egyben η értékének ismeretét is jelenti, ξ értéke egyértelműen meghatározza η értékét. Gyakran két (vagy több) valószínűségi változó között nem ennyire szoros az összefüggés. Előfordul, hogy egyik változó pl. ξ értékének ismeretében nem tudjuk ugyan egyértelműen megmondani η értékét, de mégis egy szűkebb intervallumot tudunk megadni, ahova η értéke nagy valószínűséggel esik, mintha ξ értékét nem ismernénk.



82. ábra

Képzeld el pl., hogy valamely (ξ, η) valószínűségi változópárra sok megfigyelést végzünk, s az észlelt (x_i, y_i) ($i=1, 2, \dots, n$) értékeket mint véletlen pontokat egy koordináta-rendszerben ábrázoljuk, amelynek eredménye a 82. ábrán látható alaku pontfelhő:

Ekkor ha $\xi = x_0$, az η változó értéke nagy valószínűséggel az A intervallumba esik, míg ha $\xi = x_1$, akkor nagy a valószínűsége, hogy η megfigyelt értéke a B intervallumba esik. Ilyen esetben, amikor tehát ξ értéke nem határozza meg egyértel-

műen η értékét, de ξ különböző értékeihez η -nak más eloszlása (legalábbis más paraméterekkel bíró eloszlása) tartozik, azt mondjuk, hogy ξ és η között sztochasztikus kapcsolat van.

Ha pl. mind ξ , mind η több véletlen komponens függvénye,

$$\xi = f(\zeta_1, \zeta_2, \dots, \zeta_i)$$

$$\eta = f(\zeta_1, \zeta_2, \dots, \zeta_u)$$

ahol az argumentumok ξ ill. η értékét meghatározzák, ξ és η között általában nem fogunk függvénykapcsolatot észlelni, de sztochasztikus kapcsolat általában mutatkozni fog közöttük. Ilyenkor ξ nem argumentuma η -nak, legfeljebb argumentumainak egyike, de lehet, hogy az sem, mégis úgy tűnik, hogy ξ értéke befolyásolja η értéket, mert mindkét változó nagyjából ugyanazon véletlen hatások eredménye, legalábbis a lényegesebb hatások közösek.

A matematikai statisztikában különböző módszereket dolgoztak ki a sztochasztikus kapcsolatok vizsgálatára. E módszerek egyike az ún. korreláció-elmélet. Két valószínűségi változó közötti korrelációs együttható fogalmával már a pontban megismerkedtünk. Előfordul, hogy ξ és η pontos eloszlását nem ismerjük, ilyenkor a

$$\rho = \frac{M[(\xi - M(\xi))(\eta - M(\eta))]}{D(\xi) D(\eta)}$$

korrelációs együtthatót sem tudjuk meghatározni, hanem a (ξ, η) változó-párra vonatkozó (x_i, y_i) ($i = 1, 2, \dots, n$) megfigyelt értékpárok, a minta alapján kell becsülnünk, megalkotnunk a "tapasztalati" vagy "empirikus" korrelációs együtthatót.

A továbbiakban vizsgálni fogjuk a sztochasztikus függés néhány alakját és megmutatjuk, hogy bizonyos elméleti mennyiségeket hogyan célszerű becsülni.

A korreláció-elmélet egyik legfontosabb feladata prognózis felállítása az η valószínűségi változó értékére, ha a ξ valószínűségi változó, amellyel η sztochasztikusan összefügg, adott értéket vett fel. Ez a megfogalmazás a gyakorlatban felmerülő számos probléma általánosítása. Egyszerű példa erre: jelölje ξ a Duna vízgyűjtő területén bizonyos időtartam alatt lehullott csapadék mennyiségét, η pedig a Duna vízállását adott szelvényen. Nyilván ξ és η között sztochasztikus kapcsolat van. A korreláció-elmélet segítségével tudunk bizonyos előrejelzést adni a vízállás alakulására vonatkozólag. Ha tudjuk, hogy $\xi = x$, akkor az η változó értékére vonat-

közö "legjobb jóslatunk" az η változó feltételes várható értéke az $M(\eta | \xi = x)$ számérték lesz.

12.2 Regressziós-görbék. Feltételes szórásnégyzet

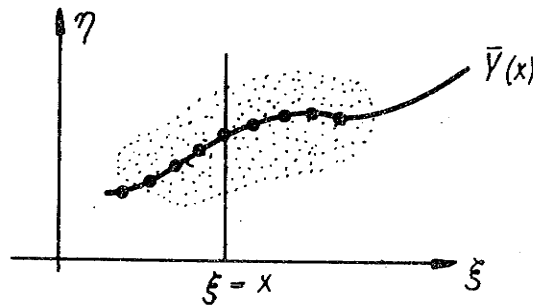
Két valószínűségi változó közötti sztochasztikus kapcsolat egyik legfontosabb megnyilatkozása, hogy egyik változó feltételes eloszlása a másik változó értékeire, mint feltételre vonatkoztatva, értékről értékre változik. A feltételes eloszlás változása a paraméterek változását is jelenti, ezért szükségünk van a feltételes várható érték alkalmazására:

$$M(\eta | \xi = x) = \bar{y}(x)$$

$$M(\xi | \eta = y) = \bar{x}(y)$$

Ezeket a feltételes várható értékeket, mint függvényeket, regressziós görbéknek nevezzük.

Ha ξ és η között funkcionális függés van, akkor $\bar{y}(x)$ természetesen azt a függést fejezi ki: $\eta = \bar{y}(\xi)$. Hasonló a helyzet a másik regressziós görbével. Ez esetben η szórása $\xi = x$ esetén zérussal egyenlő. Ha ξ és η között nincs funkcionális összefüggés, csak ennél gyengébb sztochasztikus kapcsolat, akkor η értékei szóródnak az $\bar{y}(x)$ görbe körül minden egyes $\xi = x$ rögzített érték esetén (ezt a szóródást természetesen csak akkor észleljük, ha elég sok megfigyelést végzünk a (ξ, η) értékpárra vonatkozólag).



83. ábra

E szóródás különböző x érték esetén különböző lehet, tehát maga a szóródás is x függvénye.

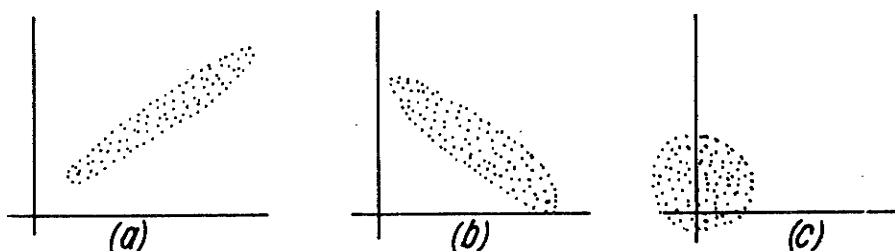
Ez a feltételes szórásnégyzet-függvény diszkrét változók esetén

$$(12.2.1) \quad \sigma^2(\eta | x) = D^2(\eta | x) = \sum_{i=1}^m [y_i - y(x)]^2 P(y_i | x)$$

folytonos változók esetén a

$$(12.2.2) \quad \sigma^2(\eta | x) = D^2(\eta | x) = \int_{-\infty}^{\infty} [y - y(x)]^2 f(y | x) dy$$

Két valószínűségi változó közötti sztochasztikus kapcsolat szorosságaról elsődleges képet nyerünk az (x_i, y_i) pontokat ábrázoló diagramok segítségével:



84. ábra

A korreláció-elméletben elsődleges szerepet a feltételes várható érték görbe és a feltételes szórás játszik.

Ha az $\bar{y}(x)$ regressziós görbe ismert és ha a ξ valószínűségi változó adott x_0 értéket vett fel és az η változó leendő y_0 értékére vonatkozólag jósolunk, akkor $D(\eta | x)$ a jóslás átlagos kvadrativ hibájaként fogható fel. Mivel $D(\eta | x)$ az x változótól függ, a jóslás pontossága is x -től függ, az-az attól, hogy ξ melyik értékét veszi fel. Ha a jóslás átlagos pontosságáról akarunk beszélni, akkor az átlagos feltételes szórásnégyzetet használjuk:

$$(12.2.3) \quad M \{ D^2(\eta | x) \} = \sum_{i=1}^m \sigma^2(\eta | x_i) p(x_i)$$

$$(12.2.4) \quad M \{ D^2(\eta | x) \} = \int_{-\infty}^{\infty} \sigma^2(\eta | x) f(x) dx$$

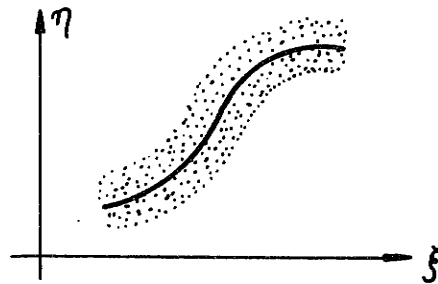
A (12.2.2) formula alapján könnyű belátni, hogy

$$(12.2.5) \quad M \{ D^2 (\eta | x) \} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [y - \bar{y}(x)]^2 f(y | x) f(x) dy dx =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [y - \bar{y}(x)]^2 f(x, y) dx dy = M [\eta - \bar{y}(\xi)]^2$$

12.3 Regresszió a legkisebb négyzetek elve alapján

Tegyük fel, hogy megfigyelésünk tárgyát $a(\xi, \eta)$ valószínűségi vektorváltozó képezi, amely folytonos eloszlású, $h(x, y)$ kétdimenziós szögfüggvénnyel. A (ξ, η) véletlen pontra vonatkozólag nagyszámu megfigyelést végezve és a nyert pontokat koordináta-rendszerben ábrázolva egy pontfelhőt nyerünk (85. ábra).



85. ábra

Ha a ξ és η változók közötti sztochasztikus függést közelítőleg függvénykapcsolat formájában kívánjuk kifejezni, akkor keressük azt az $y = \varphi(x)$ függvényt, amely a fenti pontfelhőhöz a legjobban illeszkedik a legkisebb négyzetek elve alapján, azaz amelyre

$$M [\eta - \varphi(\xi)]^2 = \text{minimum}$$

Nem nehéz kimutatni, hogy a minimum feltételt kielégítő függvény éppen az

$$M(\eta | \xi = x) = \bar{y}(x)$$

függvény, az η változó ξ -re vonatkozó regressziós görbéje.

Ugyanis, ha $g(x)$ tetszőleges másik függvénye x -nek, akkor rögzített $\xi = x$ helyhez tartozó η változó értékét η_x -szel jelölve a Steiner-formula szerint (lásd: 3.3 pont)

$$(12.3.1) \quad M [\eta_x - M(\eta_x)]^2 \leq M [\eta_x - g(x)]^2$$

Mivel a (12.3.1) egyenlőtlenség minden x -érték mellett igaz, igaz az $\bar{y}(x)$ függvényre.

Azt az eredményt kaptuk tehát, hogy a (ξ, η) pontokból álló ponthalmazhoz a legkisebb négyzetek elve értelmében legjobban illeszkedő $y = \varphi(x)$ függvény az η változó ξ -re vonatkozó regressziós görbéje: $y = \bar{y}(x)$.

Abban az esetben, ha a (ξ, η) vektorváltozó eloszlása kétdimenziós normális eloszlás, akkor az $\bar{y}(x)$ regressziós görbe egyenes. (Ugyanígy $x = \bar{x}(y)$.)

Ekkor ugyanis

$$h(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho \frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2} \right]}$$

Továbbá az η változó feltételes sűrűségfüggvénye a (2.5.5) formula szerint

$$g(y|x) = \frac{h(x, y)}{f(x)} = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_2^2} \left[y - m_2 - \rho \frac{\sigma_2}{\sigma_1} (x - m_1) \right]^2}$$

Rögzített x esetén ez $N(m_2 + \rho \frac{\sigma_2}{\sigma_1} (x - m_1); \sigma_2 \sqrt{1 - \rho^2})$ normális eloszlás sűrűségfüggvénye, vagyis:

$$\bar{y}(x) = M(\eta | \xi = x) = m_2 + \rho \frac{\sigma_2}{\sigma_1} (x - m_1)$$

Tehát kétdimenziós normális eloszlás esetén az $\bar{y}(x)$ regressziós görbe $\rho \frac{\sigma_2}{\sigma_1}$ iránytangensű egyenes, amely átmegy az (m_1, m_2) ponton, a két-

dimenziós normális eloszlás esetén a korrelációs együttható geometriai jelentésére is fény derült.

12.4 Lineáris regresszió

Ha a (ξ, η) valószínűségi vektorváltozó eloszlását nem ismerjük, akkor az $\bar{y}(x)$ regressziós görbe egyenletét nem tudjuk meghatározni. Az (x_1, y_1) értékek halmaza, a megfigyelt pontfelhő alakja sokszor lehetővé teszi, hogy a ξ és η változók közötti függést lineáris függvénnyel közelítsük és ez a gyakorlat számára legtöbbször kielégítő. Ilyenkor tehát feltesszük, hogy az η valószínűségi változó a ξ változó lineáris függvénye, azaz közelítőleg $\eta = a\xi + b$.

Keressük tehát azt az $a\xi + b$ lineáris függvényt, amelyre

$$M[\eta - a\xi - b]^2 = \text{minimum.}$$

A zárójelen belüli kifejezést átalakítva, keressük azon a és b számokat, amelyekre

$$M[\eta - m_2 - a(\xi - m_1) + m_2 - a m_1 - b]^2 = \text{minimum}$$

ahol

$$M(\xi) = m_1,$$

$$M(\eta) = m_2$$

Elvégezve a négyzetre emelést:

$$M \left\{ (\eta - m_2)^2 - 2a(\eta - m_2)(\xi - m_1) + a^2(\xi - m_1)^2 + \right. \\ \left. + 2[m_2 - am_1 - b](\eta - m_2 - a(\xi - m_1)) + (m_2 - am_1 - b)^2 \right\} = \text{minimum}$$

Figyelembe véve, hogy

$$M[(\eta - m_2) - (\xi - m_1)] = M(\eta - m_2) - M(\xi - m_1) = 0$$

$$M(\eta - m_2)^2 = D^2(\eta) = \sigma_2^2$$

$$M(\xi - m_1)^2 = D^2(\xi) = \sigma_1^2$$

$$M[(\xi - m_1)(\eta - m_2)] = \rho \sigma_1 \sigma_2$$

a minimalizálandó függvény:

$$(12.4.1) \quad f(a,b) = \sigma_2^2 - 2 a \varrho \sigma_1 \sigma_2 + a^2 \sigma_1^2 + [m_2 - a m_1 - b]^2$$

Az $f(a,b)$ függvénynek csak ott lehet szélső értéke, ahol mind az a mind b szerinti parciális deriváltja zérus.

$$(12.4.2) \quad \frac{\partial f}{\partial a} = -2 \varrho \sigma_1 \sigma_2 + 2a \sigma_1^2 - 2 [m_2 - a m_1 - b] m_1 = 0$$

$$(12.4.3) \quad \frac{\partial f}{\partial b} = -2 [m_2 - a m_1 - b] = 0.$$

A (12.4.3) összefüggésből

$$b = m_2 - a m_1$$

Ezt az értéket (12.4.2)-be helyettesítve:

$$a = \varrho \frac{\sigma_2}{\sigma_1}$$

A keresett lineáris függvény tehát:

$$(12.4.4) \quad y = \varrho \frac{\sigma_2}{\sigma_1} x + m_2 - \frac{\sigma_2}{\sigma_1} m_1$$

melyet a következő alakra hozhatunk:

$$(12.4.5) \quad \frac{y - m_2}{\sigma_2} = \varrho \frac{x - m_1}{\sigma_1}$$

Elméleti vizsgálatunk tehát azt mutatja, hogyha a ξ és η valószínűségi változókat standardizáljuk, azaz a

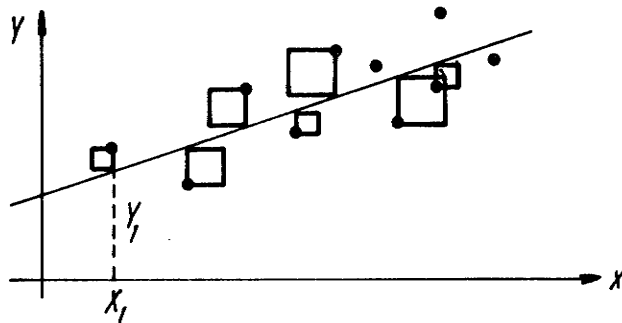
$$\xi^* = \frac{\xi - m_1}{\sigma_1} \quad ; \quad \eta^* = \frac{\eta - m_2}{\sigma_2}$$

értékeket tüntetjük fel a koordináta-rendszerben, akkor az így nyert pont-

halmazhoz a legkisebb négyzetek elve értelmében legjobban illeszkedő egyenes átmegy az origón és iránytangense éppen a ξ és η közötti korrelációs együttható.

Eddigi vizsgálatunk a regresszióval kapcsolatban elméleti jellegű volt, mivel vagy azt tételeztük fel, hogy a (ξ, η) valószínűségi vektorváltozó kétdimenziós sűrűségfüggvényét, $h(x, y)$ -t ismerjük, vagy pedig legalább a ρ korrelációs együttható, továbbá az $M(\xi) = m_1$, $D^2(\xi) = \sigma_1^2$,

$M(\eta) = m_2$, $D^2(\eta) = \sigma_2^2$ mennyiségek ismereteseek. A gyakorlatban azonban igen sokszor az a helyzet, hogy mindezeket a mennyiségeket nem ismerjük. Nem tudunk tehát mást tenni, minthogy a (ξ, η) véletlen vektorra vonatkozólag nagyszámu megfigyelést végzünk, s az észlelt (x_1, y_1) , $(x_2, y_2), \dots, (x_n, y_n)$ pontokat egy koordináta-rendszerben ábrázoljuk.



86. ábra

Ezután megkeressük azt a $y = ax + b$ egyenest, amely a nyert pont-halmazhoz a legjobban illeszkedik a legkisebb négyzetek elve értelmében, azaz meghatározzuk azt az a és b számokat, amelyekre

$$f(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2 = \text{minimum}$$

A megoldást a differenciálszámítás segítségével nyerjük:

$$(12.4.6) \quad \frac{\partial f}{\partial a} = -2 \sum_{i=1}^n (y_i - ax_i - b) x_i = 0$$

$$(12.4.7) \quad \frac{\partial f}{\partial b} = -2 \sum_{i=1}^m (y_i - a x_i - b) = 0$$

A (12.4.6) és (12.4.7) összefüggésekből az a és b ismeretlenekre az alábbi egyenletrendszert nyerjük:

$$\left(\sum x_i^2 \right) a + \left(\sum x_i \right) b = \sum x_i y_i$$

$$\left(\sum x_i \right) a + n b = \sum y_i$$

A Cramér-szabály alkalmazásával:

$$a = \frac{\begin{vmatrix} \sum x_i y_i & \sum x_i \\ \sum y_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2}$$

$$b = \frac{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}}{\begin{vmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{vmatrix}} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2}$$

Ezzel gyakorlatilag feladatunkat megoldottuk. A legjobban illeszkedő egyenes iránytangense

$$a = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

alakban is írható, amely kifejezés az elméletileg nyert irányhatározó becslése:

$$a \approx \varrho \frac{\sigma_2}{\sigma_1}$$

Tekintettel arra, hogy

$$\sigma_2 \approx \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = s_2; \quad \sigma_1 \approx \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = s_1$$

$$(12.4.8) \quad \varrho \approx a \frac{s_1}{s_2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Az így nyert r statisztikai függvényt tapasztalati korrelációs együttható-nak nevezzük, amely a ϱ elméleti korrelációs együttható becslése.

A bevezetett jelölésekkel az (x_i, y_i) ($i = 1, 2, \dots, n$) pontokból álló halmazhoz legjobban illeszkedő egyenes egyenlete:

$$(12.4.9) \quad y = r \frac{s_2}{s_1} (x - \bar{x}) + \bar{y}$$

alakban írható. Ennek alapján ha $\xi = x$ a megfigyelt érték, akkor η értékére a legjobb lineáris "jóslásunk" a összefüggés szerinti y érték. A (12.4.9) összefüggést könnyebb megjegyzés kedvéért

$$(12.4.10) \quad \frac{y - \bar{y}}{s_2} = r \frac{x - \bar{x}}{s_1}$$

alakban írhatjuk, teljes összhangban az elméletileg nyert (12.4.5) formulával.

Felmerül természetesen az a kérdés, hogy adott n számú megfigyelési adatból számított r tapasztalati korrelációs együttható mennyire közelíti a ξ és η közötti ϱ elméleti korrelációs együtthatót. Eerre vonatkozólag akkor tudnánk pontosabb választ adni, ha ismernénk az r statisztikai függvénynek (amely nyilván valószínűségi változó) az eloszlását. Ha a $\zeta = (\xi, \eta)$ vektor eloszlása kétváltozós normális eloszlás, akkor r sűrűségfüggvénye

$$(12.4.11) \quad f_n(z; \varrho) = \frac{n-2}{\pi} (1 - \varrho^2)^{\frac{n-1}{2}} (1 - z^2)^{\frac{n-4}{2}} \int_0^1 \frac{x^{n-2}}{(1 - \varrho x z)^{n-1} \sqrt{1-x^2}} dx; \quad -1 \leq z \leq 1.$$

ahol ϱ az elméleti korrelációs együttható.

Ha ζ nem kétdimenziós normális eloszlású, akkor r eloszlásának meghatározása bonyolult feladat. Ha ζ eloszlásának első négy momentuma létezik, akkor r várható értékére és szórásnégyzetére érvényesek a következő aszimptotikus formulák:

$$(12.4.12) \quad M(r) = \varrho + 0 \left(\frac{1}{n} \right); \quad D^2(r) = \frac{(1 + \varrho^2)^2}{n} + 0 \left(\frac{1}{n^{3/2}} \right)$$

A (12.4.12) összefüggés mutatja, hogy az r tapasztalati korrelációs együttható a ϱ elméleti korrelációs együtthatónak aszimptotikusan torzítatlan becslése.

R. A. Fisher bevezette a

$$(12.4.13) \quad z = \frac{1}{2} \ln n \frac{1+r}{1-r}$$

statisztikai függvényt, amelyre vonatkozólag, ha n nagy, akkor:

$$(12.4.14) \quad M(z) \approx \frac{1}{2} \ln n \frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)} \quad \text{és} \quad D^2(z) \approx \frac{1}{n-3}$$

Látható, hogy z szórásnégyzete nem függ a korrelációs együttható értékétől. z elég általános feltételek mellett közelítőleg normális eloszlású a (12.4.14) formulával kifejezett paraméterekkel.

Megjegyezzük még, hogyha a (ξ, η) vektor eloszlása kétdimenziós normális eloszlás, akkor a

$$(12.4.15) \quad t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$$

statisztikai függvény Student-eloszlású $n-2$ paraméterrel. Abban az esetben, ha ξ és η függetlenek, a t statisztika értéke közel van a zérushoz. A (12.4.15) statisztika felhasználható a $H_0: P(\xi < x, \eta < y) = P(\xi < x) P(\eta < y)$ függetlenségre vonatkozó hipotézis ellenőrzésére.

12.5 Korreláció- és regresszió több változó esetén

Regressziós felület és sík

A műszaki gyakorlatban sokszor tisztáznunk kell, hogy bizonyos mennyiségek hogyan függnek más mennyiségektől, ezek közül melyektől függnek erősen és melyektől lényegtelenül. Sokszor vizsgálnunk kell olyan kérdést is, hogy két mennyiség erős összefüggése valóban okszerű-e vagy csak további mennyiségek okozzák a szoros kapcsolat látszatát.

Tegyük fel, hogy egy adott gyakorlati vizsgálatnál n számú mennyiség játszik szerepet, ezek mindegyike valószínűségi változó és együttesük egy n -dimenziós vektor ugyancsak valószínűségi változó.

Az n számú változó bármelyikét vizsgálhatjuk a többi $n-1$ változó függvényeként, esetleg 2 (vagy három) kiválasztott változót a többi $n-2$ vagy $n-3$ változó függvényeként. (Természetesen n megválasztása bizonyos fókuszra önkényes, mindenestre igyekszünk a vizsgált jelenség alakulásában lényeges szerepet játszó valamennyi komponens figyelembe venni.)

Jelöljük a szóban forgó valószínűségi változók együttesét a $(\xi_1, \xi_2, \dots, \xi_n)$ véletlen vektorral, s ennek n -dimenziós sűrűségfüggvénye legyen $f(x_1, x_2, \dots, x_n)$.

Vizsgálhatjuk, hogy a ξ_1 változó értékeit a $\xi_1, \xi_2, \dots, \xi_n$ változók mely függvényének értékei közelítik legjobban. A legkisebb négyzetek elve most is a feltételes várható értékhez vezet, amely most $n-1$ változós függvény:

$$M(\xi_1 | x_2, x_3, \dots, x_n) = g(x_2, x_3, \dots, x_n) = \frac{\int_{-\infty}^{\infty} x_1 f(x_1, x_2, \dots, x_n) dx_1}{\int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1}$$

A $g(x_2, \dots, x_n)$ függvényt ξ_1 -nek, a $(\xi_2, \dots, \xi_n)$ -re vonatkoztatott regressziós felületének nevezzük. Adott rögzített (megfigyelt) $x_2, x_3, \dots, x_n$ értékrendszer mellett $g(x_2, x_3, \dots, x_n)$ képviseli a legjobb jóslást ξ_1 megfigyelendő értékére vonatkozólag.

Abban az esetben, ha a $(\xi_1, \xi_2, \dots, \xi_n)$ változók együttes eloszlása n -dimenziós normális eloszlás, akkor a regressziós felület sík, vagyis

$$x_1 = M(\xi_1 | \xi_2 = x_2, \dots, \xi_n = x_n) = a_{12} x_2 + a_{13} x_3 + \dots + a_{1n} x_n + a_1$$

A ξ_1 változó helyett természetesen a ξ_i változók bármelyikét kifejezhetjük a többi változó függvényeként, esetleg nem is mindegyik változót vesszük a $\xi_2, \dots, \xi_n$ változók közül. Annak kifejezésre juttatása, hogy melyik változót emeljük ki és mely többi változó függvényeként közelítjük, meg lehetőségen bonyolult jelölésmóddal lehetséges, amelyet nem kívánunk bevezetni, inkább megállapodunk, hogy mindig az aktuális kiválasztott változót jelöljük ξ_1 -gyel.

Hasonlóan a két valószínűségi változó közötti regressziós görbe egyenessel való közelítéséhez, több valószínűségi változó esetében is gyakran megelégszünk egy kiválasztott ξ_1 változó értékeinek lineáris függvénnyel való közelítésével, meghatározzuk azt a regressziós "síkot" (valójában $n-1$ dimenziós ún. hipersíkot), amelyről ξ_1 értékei a legkisebb négyzetek elve értelmében a legkevésbé térnek el.

Egyszerűség kedvéért feltesszük, hogy

$$M(\xi_i) = 0 \quad (i = 1, 2, \dots, n)$$

Ezt mindig elérhetjük azáltal, hogy mindegyik valószínűségi változóból levonjuk a várható értékét.

Keressük tehát azt a

$$\xi_1 = a_{12} \xi_2 + a_{13} \xi_3 + \dots + a_{1n} \xi_n$$

$n-1$ változós lineáris függvényt, amelyre

$$(12.5.1) \quad M(\xi_1 - a_{12} \xi_2 - a_{13} \xi_3 - \dots - a_{1n} \xi_n)^2 = \text{minimum}$$

Feltételezve, hogy a szóban forgó változók folytonos eloszlásúak, a (12.5.1) összefüggést integrál alakban is írhatjuk:

$$(12.5.2) \quad \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (x_1 - a_{12} x_2 - \dots - a_{1n} x_n)^2 f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = \text{minimum}$$

A vegyes momentumok jelölésére (kovarianciák) bevezetjük a

$$b_{ij} = b_{ji} = \int_{-\infty}^{\infty} x_i x_j f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

jelölést. Speciálisan

$$b_{ii} = M(\xi_i^2) = D^2(\xi_i) \quad (i = 1, 2, \dots, n)$$

Differenciálással az ismeretlen a_{ij} együtthatókra a következő lineáris egyenletrendszert kapjuk:

$$b_{i2} a_{12} + b_{i3} a_{13} + \dots + b_{in} a_{1n} = b_{i1} \quad (i = 2, 3, \dots, n)$$

Feltéve, hogy

$$(12.5.3) \quad \underline{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix} \quad \text{kovariancia-matrix nem szin-} \\ \text{guláris (azaz determinánsa nem} \\ \text{zérus),}$$

megoldásként az

$$(12.5.4) \quad a_{1k} = - \frac{B_{1k}}{B_{11}} \quad (k = 1, 2, \dots, n)$$

kifejezést nyerjük, (ahol B_{1k} , ill. B_{11} a b_{1k} , ill. b_{11} elemekhez tartozó aldeterminánsokat jelölik).

A ξ_i és ξ_j valószínűségi változók korrelációs együtthatója (lásd: pont)

$$\rho_{ij} = \frac{M(\xi_i \xi_j)}{D(\xi_i)D(\xi_j)} = \frac{b_{ij}}{\sqrt{b_{ii} b_{jj}}} \quad (i, j = 1, 2, \dots, n)$$

Képezzük most az

$$\eta_1 = \xi_1 - a_{12} \xi_2 - a_{13} \xi_3 - \dots - a_{1n} \xi_n = \frac{1}{B_{11}} \sum_{j=1}^n B_{1j} \xi_j$$

(itt figyelembe vettük (12.5.4)-et!)

különbséget az un. maradékot, amelyet tehát úgy nyerünk, hogy ξ_1 értékéből levonjuk a lineáris közelítést. Könnyen beláthatjuk, hogy η_1 korrelálatlan a $\xi_2, \xi_3, \dots, \xi_n$ változókkal, viszont ξ_1 -gyel pozitívan korrelált.

Ugyanis

$$M(\eta_1, \xi_i) = M(\eta_1, \xi_i) - \sum_{j=1}^n a_{ij} M(\xi_j, \xi_i) =$$

$$= \frac{1}{B_{11}} \sum_{j=1}^n B_{ij} b_{ij} = \begin{cases} \frac{|B|}{B_{11}} & \text{ha } i = j \\ 0 & \text{ha } i \neq j \end{cases}$$

(Itt $|B|$ a B kovariancia-matrix determinánsát jelöli.) Mínt hogy a $\xi_1, \xi_2, \dots, \xi_n$ változók mindegyikének várható értéke zérus, a (3.1.4) és (3.1.5) összefüggések alapján következik, hogy $M(\eta_1) = 0$.

Megjegyezzük, hogy abból a tényből, hogy a B kovariancia-matrix szimmetrikus, következik, hogy $|B|$ és B_{11} nemnegatív értékek, s abban az esetben, ha a $\xi_1, \xi_2, \dots, \xi_n$ változók között nem áll fenn lineáris függés, a fenti determinánsok értéke pozitív. Az η_1 és ξ_1 változók közötti korrelációs együttható:

$$\rho(\eta_1, \xi_1) = \frac{M(\xi_1, \eta_1)}{D(\xi_1) D(\eta_1)} = \sqrt{\frac{|B|}{b_{11} B_{11}}}$$

Ha a $(\xi_1, \xi_2, \dots, \xi_n)$ véletlen vektor-eloszlása n -dimenziós normális eloszlás, akkor η_1 független a $\xi_2, \dots, \xi_n$ változóktól, η_1 -gyel pedig pozitívan korrelált. Ebben az esetben azt mondhatjuk, hogy η_1 -ben a $\xi_2, \xi_3, \dots, \xi_n$ változók hatását teljesen kiküszöböltük.

12.6 Parciális korreláció

Tekintsük ismét a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változókat. Ha kiszámítjuk a ξ_i és ξ_k változók közötti

$$R(\xi_i, \xi_k) = \rho_{ik} = \frac{M\{[\xi_i - M(\xi_i)] [\xi_k - M(\xi_k)]\}}{D(\xi_i) D(\xi_k)}$$

korrelációs együtthatót, - amelyet szokás totális korrelációs együtthatónak is nevezni, - akkor mint a 3.5 pontban kifejtettük, ez a korrelációs együttható -1 és +1 között tetszőleges értéket felvehet, s ha abszolút értéke közel van 1-hez, akkor azt mondjuk, hogy ξ_1 és ξ_k között szoros korreláció áll fenn. Szoros korreláció mutatkozhat annak következtében is, hogy mind ξ_1 , mind ξ_k értékeinek alakulását a többi valószínűségi változók értékeitől való függés irányítja, egymással pedig nincsenek okozati összefüggésben és talán ha a többi változó hatását kiküszöbölnénk, nem is találnánk közöttük szoros korrelációt. A többi változó hatásának kiküszöbölése céljából úgy járunk el, hogy mind ξ_1 -et, mind ξ_2 -t kifejezzük a $\xi_3, \dots, \xi_s$ változókra vonatkozó legjobb közelítő regressziós felületekkel

$$\xi_1 \approx \sum_{j=3}^n a_{1j} \xi_j; \quad \xi_2 \approx \sum_{j=3}^n a_{2j} \xi_j$$

Ezután képezzük a

$$\eta_1 = \xi_1 - \sum_{j=3}^n a_{1j} \xi_j \quad \text{és} \quad \eta_2 = \xi_2 - \sum_{j=3}^n a_{2j} \xi_j$$

maradékokat, amelyekről tudjuk, hogy a $\xi_3, \dots, \xi_n$ változókkal korrelálatlanok, együttes normális eloszlás esetén pedig függetlenek is tőlük, viszont η_1 pozitívan korrelált ξ_1 -gyel, η_2 pedig pozitívan korrelált ξ_2 -vel. Az η_1 és η_2 maradékok közötti korrelációs együtthatót a ξ_1 és ξ_2 változók közötti parciális korrelációs együtthatónak nevezzük. Ugy tekintjük, hogy a parciális korrelációs együttható ξ_1 és ξ_2 kapcsolatát tükrözi, megtisztítva ezt a kapcsolatot a $\xi_3, \dots, \xi_n$ valószínűségi változók hatásától.

A parciális korrelációs együtthatóra bevezetjük a következő jelölést:

$$\rho_{3,4,\dots,n}^{(1,2)} = \frac{M(\eta_1 \eta_2)}{D(\eta_1)D(\eta_2)} = - \frac{B_{12}}{\sqrt{B_{11} \cdot B_{22}}}$$

ahol B_{ik} a B korrelációs matrix (lásd 12.5.3) b_{ik} eleméhez tartozó algebrai aldeterminánst jelöli.

A korreláció és regresszió elmélettel kapcsolatos mondanivalónk befejezésül még egy megjegyzést teszünk. Fejtegetéseink során a lineáris modell játszott központi szerepet, mind a kétváltozós, mind a többváltozós

regresszió esetében, mivel ennek kezelése egyszerű, a számítások keresztülvihetők és első közelítésre, a tendenciák megállapítására a gyakorlatban legtöbbször kielégítő. Ennek ellenére előfordul, hogy két valószínűségi változó regressziós görbéje közelítőleg sem lineáris, a (ξ, η) változó-pár megfigyelt értékeiből álló pontthalmaz egyáltalán nem "illeszkedik" egyeneshez. Ilyenkor kereshetjük ξ -nek azt az

$$a_0 + a_1 \xi + a_2 \xi^2 + \dots + a_n \xi^n$$

n -edfoku polinomját, amely η -t legjobban közelíti a legkisebb négyzetek elve értelmében, azaz amelyre

$$M(\eta - a_0 - a_1 \xi - a_2 \xi^2 - \dots - a_n \xi^n)^2 = \text{minimum.}$$

Bevezetve $a \xi = \xi_1, \xi^2 = \xi_2, \dots, \xi^n = \xi_n$ jelölést, az ismeretlen együtthatók meghatározására most is alkalmazhatjuk a 12.5 pontban ismertett módszert, csak szem előtt kell tartanunk, hogy mit jelentenek most a $\xi_1, \xi_2, \dots, \xi_n$ változók.

A legkisebb négyzetek elve alapján legjobban közelítő polinomot szokás polinomiális regressziónak is nevezni. A közelítő polinom fokszámát n -et illetően a megválasztás önkényes, legtöbbször a pontthalmaz alakja, vagy az elméleti sűrűségfüggvény ismerete nyújt támpontot. A gyakorlatban általában harmad vagy negyedfoku polinommal megelégszünk, mert ennél magasabb fokszám esetén a számolás rendkívül munkaigényes.

A legkisebb négyzetek módszere alkalmazást nyer idősorok elemzése esetében is. Ilyenkor az argumentum az idő, a függvényértékek valószínűségi változók, s a tendenciák megállapítására a legkisebb négyzetek módszerre hasznos szolgálatot tesz. E kérdéskör iránt érdeklődő olvasó számára utalunk a [8] és [18] munkákra.

IRODALOM

- [1] Cramér, H.: Mathematical Methods of Statistics. Princeton, Princeton, Princeton Univ. Press, 1954.
- [2] Dunin-Barkovszkij, I. V. - Szmírnov, N. V.: Teorija verojatnosztej i matematiceszkaja sztatisztika v tehike.
- [3] Feller, W.: An Introduction to Probability Theory and its Applications. New York, Wiley, 1952.
- [4] Fisher, R. A.: Contributions to Mathematical Statistics. New York, Wiley, 1950.
- [5] Fis, M.: Wahrscheinlichkeitsrechnung und mathematische Statistik. Berlin. Dt. Verl. d. Wissenschaften, 1958.
- [6] Hald, A.: Statistical Theory with Engineering Applications. New York, Wiley, 1952.
- [7] Jordan K.: Mathetmatikai statisztika. Budapest, Athenaeum, 1927.
- [8] Kendall, M. G.: The Advanced Theory of Statistics. Vols 1-2. London, Friffin, 1952-1955.
- [9] Medgyessy P. - Takács L.: Valószínűségyszámítás. Műszaki Matematikai Gyakorlatok (szerk. Fazekas F.) Budapest, Tankönyvkiadó, 1957.
- [10] Prabhu, N. U.: Some Exact Result for the Finite Dam. The Annals of Mathematical Statistics. Vol. 29. No. 4. 1958.
- [11] Prékopa, A.: Valószínűségelmélet műszaki alkalmazásokkal. Budapest, Műszaki K., 1962.
- [12] Rao, C. R.: Advanced Statistical Methods in Biometric Research. New York, Wiley, 1965.
- [13] Reimann J.: Valószínűségyszámítás. (jegyzet) Budapest. Tankönyvkiadó. 1971.

- [14] Reimann J. - Vincze I.: On the Comparison of two Samples with Slightly Different Sizes. MTA Mat. Kut. Int. Közl., 1960.
- [15] Rényi A.: Valószínűesszámitás. Budapest, Tankönyvkiadó 1954.
- [16] Rényi A.: Valószínűesszámitás. Budapest, Tankönyvkiadó 1966.
- [17] Sarkadi K.: On Testing for Normality. MTA. Mat. Kut. Int. Közl., 1960.
- [18] Sweschnikow: Untersuchungsmethoden der Theorie der Zufallsfunktionen mit Praktischen Anwendungen. Leipzig, Teubner, 1965.
- [19] Statisztikai minőségellenőrzés (szerk. Vincze I.), Budapest, Közgazd. Jogi K., 1958.
- [20] Takács L: Tartózkodási idő problémákról. MTA III. oszt. Közleményei, VII. kötet. 3-4. szám, 1957.
- [21] Todorovics: On some problems involving random number of random variables. Annals of Math. Stat. Vol. 41. No.3, 1970.
- [22] Vincze I.: Matematikai statisztika (jegyzet). Budapest, Tankönyvkiadó, 1971.
- [23] Vincze I.: Matematikai statisztika ipari alkalmazásokkal. Budapest, Műszaki Könyvkiadó, 1968.
- [24] Vincze I.: Mathematische Statistik mit Industriellen Anwendungen. Budapest, Akadémiai Kiadó, 1971.
- [25] Wald A.: Sequential Analysis. New York, Wiley, 1947.
- [26] Wald A.: Statistical Decision Functions. New York, Wiley, 1950.
- [27] van der Waerden, B. L.: Mathematische Statistik. Berlin, Springer, 1957.
- [28] Zelenhasic E.: Theoretical Probability Distributions for Flood Peaks. Hidrology Papers Colorade State Univ. (1970).

I. táblázat

 χ^2 -eloszlás

P	n	P														
		0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	n
1	0,000	0,000	0,003	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827	1	
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815	2	
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,268	3	
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,465	4	
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,517	5	
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457	6	
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,282	9,803	12,017	14,067	16,622	18,475	24,322	7	
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125	8	
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877	9	
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588	10	
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264	11	
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909	12	
13	4,107	4,765	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528	13	
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123	14	
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697	15	
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252	16	
17	6,408	7,225	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790	17	
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312	18	
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820	19	
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315	20	
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797	21	
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268	22	
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728	23	
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,533	33,196	36,415	40,270	42,980	51,179	24	
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620	25	
26	12,198	13,409	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052	26	
27	12,879	14,125	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476	27	
28	13,565	14,847	16,928	18,939	21,558	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,793	28	
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302	29	
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703	30	

A $K(z)$ függvény

- 297 -

III. táblázat

A Student-eloszlás

$\frac{P}{n}$	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001	$\frac{P}{n}$
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619	1
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598	2
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941	3
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610	4
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,859	5
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959	6
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405	7
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041	8
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781	9
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587	10
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,756	2,201	2,718	3,106	4,437	11
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318	12
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221	13
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140	14
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073	15
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015	16
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965	17
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922	18
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883	19
20	0,127	0,257	0,391	0,533	0,687	0,860	1,054	1,325	1,725	2,086	2,528	2,845	3,850	20
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819	21
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792	22
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767	23
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745	24
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,458	2,787	3,725	25
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707	26
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690	27
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674	28
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659	29
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646	30
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551	40
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460	60
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373	120
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291	∞

A nem teljes Γ -függvény táblázata

n	$\lambda=0,001$	$\lambda=0,002$	$\lambda=0,003$	$\lambda=0,004$
1	0,000999	0,001988	0,002995	0,003992
2	000001	000001	000005	000008
n	$\lambda=0,005$	$\lambda=0,006$	$\lambda=0,007$	$\lambda=0,008$
1	0,004987	0,005982	0,006976	0,007968
2	000013	000018	000024	000032
n	$\lambda=0,009$	$\lambda=0,01$	$\lambda=0,02$	$\lambda=0,03$
1	0,008960	0,009950	0,019801	0,029555
2	000040	000050	000197	000441
3			000002	000004
n	$\lambda=0,04$	$\lambda=0,05$	$\lambda=0,06$	$\lambda=0,07$
1	0,039211	0,048771	0,058236	0,067606
2	000779	001209	001730	002339
3	000010	000020	000034	000054
4				000001
n	$\lambda=0,08$	$\lambda=0,09$	$\lambda=0,10$	$\lambda=0,11$
1	0,076884	0,086069	0,09563	0,104166
2	003034	003815	004679	005624
3	000080	000114	000155	000204
4	000002	000002	000003	000006
n	$\lambda=0,12$	$\lambda=0,13$	$\lambda=0,14$	$\lambda=0,15$
1	0,113080	0,121905	0,130642	0,139292
2	006649	007752	008932	010186
3	000263	000332	000412	000503
4	000008	000011	000014	000018
5				000001
n	$\lambda=0,16$	$\lambda=0,17$	$\lambda=0,18$	$\lambda=0,19$
1	0,147856	0,156335	0,164730	0,173041
2	011513	012912	014381	015919
3	000606	000721	000850	000992
4	000024	000031	000038	000047
5	000001	000001	000001	000001
n	$\lambda=0,20$	$\lambda=0,22$	$\lambda=0,24$	$\lambda=0,26$
1	0,181269	0,197481	0,213372	0,228948
2	017523	020927	024581	028475
3	001149	001506	001927	002414
4	000057	000082	000114	000154
5	000002	000004	000006	000008
6				000001
n	$\lambda=0,28$	$\lambda=0,30$	$\lambda=0,40$	$\lambda=0,50$
1	0,244216	0,259182	0,329680	0,393469
2	032597	036936	061551	090204
3	002970	003600	007926	014388
4	000205	000266	000776	001752
5	000011	000015	000061	000173
6	000001	000001	000004	000014
7			000001	000001

V. tábl. folyt.

n	$\lambda=1,0$	$\lambda=1,5$	$\lambda=2,0$	$\lambda=2,5$
1	0,63212	0,77687	0,86466	0,91792
2	26424	44218	59399	71270
3	08030	19116	32332	45619
4	01899	06564	14288	24242
5	00366	01858	05265	10882
6	00059	00446	01656	04202
7	00008	00093	00453	01419
8	00001	00017	00110	00425
9	00001	00002	00024	00114
10			00005	00028
11			00001	00006
12				00001
n	$\lambda=3,0$	$\lambda=3,5$	$\lambda=4,0$	$\lambda=4,5$
1	0,95021	0,96980	0,98168	0,98889
2	80085	86411	90842	93890
3	57681	67915	76190	82542
4	35277	46338	56653	65770
5	18474	27456	37116	46789
6	08392	14239	21487	29708
7	03351	06529	11067	16895
8	01191	02674	05113	08659
9	00380	00987	02137	04026
10	00110	00331	00813	01709
11	00029	00102	00284	00667
12	00007	00029	00092	00240
13	00002	00008	00027	00081
14		00001	00008	00025
15			00002	00007
16			00001	00002
17				00001
n	$\lambda=5,0$	$\lambda=5,5$	$\lambda=6,0$	$\lambda=6,5$
1	0,99326	0,99591	0,99752	0,99850
2	95957	97345	98265	98872
3	87535	91162	92804	95696
4	73497	79830	84880	88816
5	44951	64248	71494	77633
6	38404	47108	55433	63096
7	23782	21296	39370	47347
8	13337	19051	25603	32724
9	06809	10564	15276	20843
10	03183	05378	08392	12262
11	01369	02525	04262	06684
12	00545	01099	02009	03389
13	00202	00445	00883	01603
14	00070	00169	00363	00710
15	00023	00060	00140	00296
16	00007	00020	00051	00116
17	00002	00006	00017	00044
18	00001	00002	00006	00015
19		00001	00002	00005
20			00001	00001

n	$\lambda=7,0$	$\lambda=7,5$	$\lambda=8,0$	$\lambda=8,5$
1	0,99909	0,99945	0,99966	0,99980
2	99271	99530	99698	99807
3	97036	97975	98625	99072
4	91823	93085	95762	96589
5	82701	86794	90037	92564
6	69925	75565	80876	85040
7	55029	62184	68663	74382
8	40129	47536	54704	61440
9	27091	33803	40745	47689
10	16950	22359	28338	34703
11	09852	13776	18412	23664
12	05335	07924	11192	15134
13	02700	04267	06380	09092
14	01281	02157	03418	05141
15	00572	01026	01726	02743
16	00241	00461	00823	01383
17	00096	00196	00372	01111
18	00036	00079	00160	00300
19	00013	00031	00065	00130
20	00005	00011	00025	00054
21		00004	00010	00020
22		00001	00003	00008
23			00001	00003
24				00001

n	$\lambda=9,0$	$\lambda=9,5$	$\lambda=10,0$
1	0,99988	0,99993	0,99996
2	99877	99921	99950
3	99377	99584	99724
4	97877	98514	98966
5	94504	95974	97075
6	88431	91147	93291
7	79322	83505	86986
8	67610	73134	77978
9	54435	60818	66719
10	41259	47817	54207
11	29401	35467	41696
12	19699	24801	30322
13	12423	16338	20844
14	07385	10186	13554
15	04147	05999	08346
16	02204	03347	04874
17	01111	01773	02704
18	00533	00893	01428
19	00243	00428	00719
20	00106	00196	00345
21	00044	00086	00159
22	00018	00036	00070
23	00006	00015	00030
24	00003	00006	00012
25		00002	00004
26		00001	00001

VII. táblázat

A Poisson - féle eloszlás

$k \backslash \lambda$	1	2	3	4	5	6	7	8
0	0,36788	0,13534	0,04978	0,01831	0,00673	0,00247	0,00091	0,00033
1	0,36778	0,27067	0,14936	0,07326	0,03369	0,01478	0,00638	0,00268
2	0,18394	0,27067	0,22404	0,14653	0,08422	0,04461	0,02234	0,01073
3	0,06131	0,18045	0,22404	0,19537	0,14037	0,08923	0,05212	0,02862
4	0,01532	0,09022	0,16803	0,19537	0,17547	0,13385	0,09122	0,05725
5	0,00306	0,03609	0,10082	0,15629	0,17547	0,16062	0,12772	0,09160
6	0,00051	0,01203	0,05040	0,10420	0,14622	0,16062	0,14900	0,12214
7	0,00007	0,00343	0,02160	0,05954	0,10444	0,13768	0,14900	0,13959
8		0,00085	0,00810	0,02977	0,06527	0,10326	0,13038	0,13959
9		0,00019	0,00270	0,01322	0,03626	0,06883	0,10140	0,12408
10		0,00003	0,00081	0,00529	0,01813	0,04130	0,07098	0,09926
11			0,00022	0,00192	0,00824	0,02252	0,04517	0,07219
12			0,00005	0,00064	0,00343	0,01126	0,02635	0,04812
13			0,00001	0,00019	0,00132	0,00519	0,01418	0,02961
14				0,00005	0,00047	0,00222	0,00709	0,01692
15				0,00001	0,00015	0,00089	0,00331	0,00902
16					0,00004	0,00033	0,00144	0,00451
17					0,00001	0,00011	0,00059	0,00212
18						0,00003	0,00023	0,00094
19						0,00001	0,00008	0,00039
20							0,00003	0,00015
21								0,00005
22								0,00002

$k \backslash \lambda$	9	10	11	12	13	14	15
0	0,00012	0,00004	0,00001				
1	0,00111	0,00045	0,00018	0,00007	0,00002	0,00001	
2	0,00499	0,00227	0,00101	0,00044	0,00019	0,00008	0,00003
3	0,01499	0,00756	0,00370	0,00177	0,00082	0,00038	0,00017
4	0,03373	0,01891	0,01018	0,00530	0,00269	0,00133	0,00064
5	0,06072	0,03783	0,02241	0,01274	0,00699	0,00373	0,00193
6	0,09109	0,06305	0,04109	0,02548	0,01515	0,00869	0,00483
7	0,11712	0,09007	0,06457	0,04368	0,02814	0,02739	0,01037
8	0,13176	0,11260	0,08879	0,06552	0,04573	0,03043	0,01944
9	0,13176	0,12511	0,10853	0,08736	0,06605	0,04734	0,03240
10	0,11858	0,12511	0,11938	0,10484	0,08587	0,06628	0,04861
11	0,09702	0,11374	0,11938	0,11437	0,10148	0,08435	0,06628
12	0,07276	0,09478	0,10943	0,11437	0,10994	0,09841	0,08285
13	0,05037	0,07290	0,09259	0,10557	0,10994	0,10599	0,09560
14	0,03238	0,05207	0,07275	0,09048	0,10209	0,10599	0,10244
15	0,01943	0,03471	0,05335	0,07239	0,08847	0,09892	0,10244
16	0,01093	0,02169	0,03668	0,05429	0,07188	0,08655	0,09603
17	0,00578	0,01276	0,02373	0,03832	0,05497	0,07128	0,08373
18	0,00289	0,00709	0,01450	0,02555	0,03970	0,05544	0,07061
19	0,00137	0,00373	0,00839	0,01613	0,02716	0,04085	0,05574
20	0,00061	0,00183	0,00461	0,00968	0,01765	0,02859	0,04181
21	0,00026	0,00088	0,00241	0,00553	0,01093	0,01906	0,02986
22	0,00010	0,00040	0,00121	0,00301	0,00645	0,01213	0,02036
23	0,00004	0,00017	0,00057	0,00157	0,00365	0,00738	0,01328
24	0,00001	0,00007	0,00026	0,00078	0,00197	0,00430	0,00830
25		0,00002	0,00011	0,00037	0,00102	0,00241	0,00498
26		0,00001	0,00004	0,00017	0,00051	0,00129	0,00287
27			0,00002	0,00007	0,00024	0,00067	0,00159
28				0,00003	0,00011	0,00033	0,00085
29				0,00001	0,00005	0,00016	0,00044
30					0,00002	0,00007	0,00022
31						0,00003	0,00010
32						0,00001	0,00005
33							0,00002
34							0,00001

ibadási foka: f_1														
	1	2	3	10	45	50	60	80	100	200	500	∞		
1	161	200	216	228	251	252	252	252	253	254	254	254		
2	18,5	19,0	19,2	19,5	19,5	19,5	19,5	19,5	19,5	19,5	19,5	19,5		
3	10,2	9,55	9,28	9,59	8,59	8,58	8,57	8,56	8,55	8,54	8,53	8,53		
4	7,61	6,94	6,59	6,72	5,71	5,70	5,69	5,67	5,66	5,65	5,64	5,63		
5	6,61	5,79	5,41	5,46	4,45	4,44	4,43	4,41	4,41	4,39	4,37	4,37		
6	5,99	5,14	4,76	4,77	3,76	3,75	3,74	3,72	3,71	3,69	3,68	3,67		
7	5,59	4,74	4,35	4,34	3,33	3,32	3,30	3,29	3,27	3,25	3,24	3,23		
8	5,32	4,46	4,07	3,04	3,03	3,02	3,01	2,99	2,97	2,95	2,94	2,93		
9	5,12	4,26	3,86	3,83	2,81	2,80	2,79	2,77	2,76	2,73	2,72	2,71		
10	4,96	4,10	3,71	3,66	2,65	2,64	2,62	2,60	2,59	2,56	2,55	2,54		
11	4,84	3,98	3,59	3,53	2,52	2,51	2,49	2,47	2,46	2,43	2,42	2,40		
12	4,75	3,89	3,49	3,43	2,41	2,40	2,38	2,36	2,35	2,32	2,31	2,30		
13	4,67	3,81	3,41	3,34	2,33	2,31	2,30	2,27	2,26	2,23	2,22	2,21		
14	4,60	3,74	3,34	3,27	2,25	2,24	2,22	2,20	2,19	2,16	2,14	2,13		
15	4,54	3,68	3,29	3,20	2,19	2,18	2,16	2,14	2,12	2,10	2,08	2,07		
16	4,49	3,63	3,24	3,15	2,14	2,12	2,11	2,08	2,07	2,04	2,02	2,01		
17	4,45	3,59	3,20	2,10	2,09	2,08	2,06	2,03	2,02	1,99	1,97	1,96		
18	4,41	3,55	3,16	2,06	2,05	2,04	2,02	1,99	1,98	1,95	1,93	1,92		
19	4,38	3,52	3,13	2,03	2,01	2,00	1,98	1,96	1,94	1,91	1,89	1,88		
20	4,35	3,49	3,10	2,99	1,98	1,97	1,95	1,92	1,91	1,88	1,86	1,84		
21	4,32	3,47	3,07	2,96	1,95	1,94	1,92	1,89	1,88	1,84	1,82	1,81		
22	4,30	3,44	3,05	2,94	1,92	1,91	1,89	1,86	1,85	1,82	1,80	1,78		
23	4,28	3,42	3,03	2,91	1,90	1,88	1,86	1,84	1,82	1,79	1,77	1,76		
24	4,26	3,40	3,01	2,89	1,88	1,86	1,84	1,82	1,80	1,77	1,75	1,73		
25	4,24	3,39	2,99	2,87	1,86	1,84	1,82	1,80	1,78	1,75	1,73	1,71		
26	4,23	3,37	2,98	2,85	1,84	1,82	1,80	1,78	1,76	1,73	1,71	1,69		
27	4,21	3,35	2,96	2,84	1,82	1,81	1,79	1,76	1,74	1,71	1,69	1,67		
28	4,20	3,34	2,95	2,82	1,80	1,79	1,77	1,74	1,73	1,69	1,67	1,65		
29	4,18	3,33	2,93	2,81	1,79	1,77	1,75	1,73	1,71	1,67	1,65	1,64		
30	4,17	3,32	2,92	2,79	1,77	1,76	1,74	1,71	1,70	1,66	1,64	1,62		
32	4,15	3,29	2,90	2,77	1,75	1,74	1,71	1,69	1,67	1,63	1,61	1,59		
34	4,13	3,28	2,88	2,75	1,73	1,71	1,69	1,66	1,65	1,61	1,59	1,57		
36	4,11	3,26	2,87	2,73	1,71	1,69	1,67	1,64	1,62	1,59	1,56	1,55		
38	4,10	3,24	2,85	2,71	1,69	1,68	1,65	1,62	1,61	1,57	1,54	1,53		
40	4,08	3,23	2,84	2,69	1,67	1,66	1,64	1,61	1,59	1,55	1,53	1,51		
42	4,07	3,22	2,83	2,68	1,66	1,65	1,62	1,59	1,57	1,53	1,51	1,49		
44	4,06	3,21	2,82	2,67	1,65	1,63	1,61	1,58	1,56	1,52	1,49	1,48		
46	4,05	3,20	2,81	2,65	1,64	1,62	1,60	1,57	1,55	1,51	1,48	1,46		
48	4,04	3,19	2,80	2,64	1,62	1,61	1,59	1,56	1,54	1,49	1,47	1,45		
50	4,03	3,18	2,79	2,63	1,61	1,60	1,58	1,54	1,52	1,46	1,46	1,44		
55	4,02	3,16	2,77	2,61	1,59	1,58	1,55	1,52	1,50	1,46	1,43	1,41		
60	4,00	3,15	2,76	2,59	1,57	1,56	1,53	1,50	1,48	1,44	1,41	1,39		
65	3,99	3,14	2,75	2,58	1,56	1,54	1,52	1,49	1,46	1,42	1,39	1,37		
70	3,98	3,13	2,74	2,57	1,55	1,53	1,50	1,47	1,45	1,40	1,37	1,35		
80	3,96	3,11	2,72	2,54	1,52	1,51	1,48	1,45	1,43	1,38	1,35	1,21		
90	3,95	3,10	2,71	2,53	1,51	1,49	1,46	1,43	1,41	1,36	1,32	1,30		
100	3,94	3,09	2,70	2,52	1,49	1,48	1,45	1,41	1,39	1,34	1,31	1,28		
125	3,92	3,07	2,68	2,49	1,47	1,45	1,32	1,39	1,36	1,31	1,27	1,25		
150	3,90	3,06	2,66	2,48	1,45	1,44	1,41	1,37	1,34	1,29	1,25	1,22		
200	3,89	3,04	2,65	2,46	1,43	1,41	1,39	1,35	1,32	1,26	1,22	1,19		
300	3,87	3,03	2,63	2,43	1,41	1,39	1,36	1,32	1,30	1,23	1,19	1,15		
500	3,86	3,01	2,62	2,42	1,40	1,38	1,34	1,30	1,28	1,21	1,16	1,11		
1000	3,85	3,00	2,62	2,41	1,38	1,36	1,33	1,29	1,26	1,19	1,13	1,08		
∞	3,84	3,00	2,60	2,39	1,37	1,35	1,32	1,27	1,24	1,17	1,11	1,00		

A nevező szabadsági foka: f_2

A nevező szabadsági foka: f_2

TARTALOM

1. Fejezet

Alapfogalmak	3
1.1 A véletlen esemény matematikai fogalma	3
1.2 A valószínűség fogalma	7
1.3 Valószínűség-számítási axiómák	10
1.4 Valószínűségek közötti összefüggések és tételek.	12
1.5 A valószínűségek klasszikus kombinatorikus kiszámítási .. módja	15
1.6 Valószínűségek meghatározása geometriai módszerekkel. .	31
1.7 Feltételes valószínűség és függetlenség	38

2. Fejezet

Valószínűségi változók és valószínűségeloszlások	43
2.1 A valószínűségi változó és a valószínűségeloszlás fogalma	43
2.2 Diszkrét valószínűségi változók	48
2.3 Folytonos valószínűségi változók	50
2.4 A többdimenziós valószínűségeloszlás fogalma és valószí- nőségi változók függetlensége	53
2.5 Feltételes eloszlásfüggvény és sűrűségfüggvény	57
2.6 Valószínűségi változók függvényeinek eloszlása	63
a) Valószínűségi változó monoton függvényének eloszlása. .	63
b) Független valószínűségi változók összegének, szorza- tának és hányadosának eloszlása. Kompozíció	64

3. Fejezet

A valószínűségeloszlások jellemző adatai	68
3.1 A várható érték fogalma és tulajdonságai	68
3.2 A feltételes várható érték	74
3.3 A szórás fogalma és tulajdonságai	76
3.4 A valószínűségi változó momentumai	80
3.5 A korrelációs együttható	83

4. Fejezet

A generátorfüggvény és a karakterisztikus függvény	87
4.1 A generátorfüggvény	87
4.2 A karakterisztikus függvény	89

5. Fejezet

Fontosabb valószínűségeloszlások	93
A) Diszkrét valószínűségeloszlások	93
5.1 Egyszerű alternatíva	93
5.2 Binomális vagy Bernoulli-eloszlás	94
5.3 A polinomiális - eloszlás	97
5.4 A geometriai - eloszlás	98
5.5 Poisson - eloszlás	99
5.6 Véletlen eseményfolyamatok. A Poisson-folyamat alkalmazása	104
5.7 Egy sorbanállási probléma	111
5.8 Készülékek karbantartási problémáinak matematikai modellje	115
B) Folytonos eloszlások	119
5.9 A normális eloszlás	119
5.10 A binomiális eloszlás közelítése. A Moivre-Laplace határeloszlás-tétel	131
5.11 A többdimenziós normális eloszlás	134
5.12 A logaritmikus normális eloszlás	138
5.13 Az egyenletes eloszlás	139
5.14 Néhány valószínűségeloszlás, amely a matematikai statisztikában fontos szerepet játszik	141
a) A χ^2 eloszlás	141
b) A Student-féle t-eloszlás	145
c) χ^2 -eloszlású változók hányadosának eloszlása	147
d) A béta-eloszlás	148
5.15 A nagy számok törvényének Bernoulli-féle alakja	149
5.16 A központi határeloszlástétel	152

6. Fejezet

Markov-láncok	156
6.1 A Markov-lánc fogalma	156
6.2 A Markov-láncok alkalmazása viztározókkal kapcsolatos problémákra	167

7. Fejezet

A matematikai statisztika elemei	173
7.1 A matematikai statisztika tárgya és alkalmazási területei	173
7.2 A matematikai statisztika fő fejezetei	174
7.3 A mintavétel elmélete	178
7.4 A statisztikai minta jellemzői	180

8. Fejezet

Becslésmélelet.	189
8.1 A becslés fogalma és tulajdonságai	189
8.2 A becslés módszere; maximum-likelihood-módszer	200
8.3 A becslés megbízhatósága	205

9. Fejezet

Statisztikai hipotézisek vizsgálata	209
9.1 Elvi megjegyzések	209
9.2 Az u-próba. Első- és másodfajú hiba	214
9.3 Az erőfüggvény	218
9.4 A Student-féle t-próba	220
9.5 Welch-próba	223
9.6 Az F-próba	224
9.7 A Bartlett-próba	225
9.8 A Pitman-próba	226
9.9 Több normális eloszlású valószínűségi változó várható értékének összehasonlítása (Szórásanalízis).	228
9.10 Szekvenciális módszer	232

10. Fejezet

Nemparaméteres próbák	242
10.1 A rendezett minták elméletének elemei	243
10.2 Kolmogorov-Szmirnov típusú határeloszlások és statisztikai alkalmazásuk	247
10.3 További módszerek illeszkedésvizsgálata	258
a) Normalitásvizsgálat	258
b) A χ^2 -próba	261
10.4 A minta véletlenségének vizsgálata.	268
10.5 Cramér-Mises próba	270

11. Fejezet

Hibaelmélet	273
11.1 Közvetlen megfigyelések.	273

12. Fejezet

Korreláció- és regresszió analízis	276
12.1 Valószínűségi változók közötti sztochasztikus kapcsolatokról	276
12.2 Feltételes várható érték és feltételes szórásnégyzet. Regressziós-görbék.	278
12.3 Regresszió a legkisebb négyzetek elve alapján. Kétváltozós normális eloszlás esete	280
12.4 Lineáris regresszió	282
12.5 Korreláció és regresszió több változó esetén.	288
12.6 Parciális-korreláció	291
Irodalom	294
Táblázatok	296
Tartalom	309