

BUDAPESTI MŰSZAKI EGYETEM
ÉPÍTŐMÉRNÖKI KAR

Dr. Reimann József

MATEMATIKA
II./2.

Műegyetemi Kiadó, 1993.

Jegyzet azonosító: **91257**

A J9-1257 számú KÉZIRAT (1991) változatlan utánnyomása

Kiadja a **Műegyetemi Kiadó**

a **Budapesti Műszaki Egyetem Építőmérnöki Kar** megrendelése alapján.
Felelős vezető: Veress János

Terjedelem: 26,3 (A/5) ív

Készül az **Új Élet Mg.Tsz. Magyaralmás**
Nyomdaipari Önelszámoló részlegében

Felelős vezető: Frigy Ottóné

Munkaszám: **93/80**

Tartalom

1. Fejezet	7
1.1. Valószínűségelméleti alapok.....	7
1.1.1. A véletlen esemény fogalma.....	7
1.1.2. A valószínűség fogalma. Valószínűségszámítási axiómák.....	12
1.1.3. Kombinatorikai alapok.....	18
1.1.4. Példák valószínűségek kombinatorikus kiszámítására.....	24
1.1.5. Feltételes valószínűség és függetlenség.	30
2. Fejezet	
2.1. Valószínűségi változók és valószínűségeloszlások.....	38
2.1.1. A valószínűségi változó és a valószínűségeloszlás fogalma.....	38
2.1.2. A többdimenziós eloszlás fogalma.....	48
2.1.3. Feltételes eloszlás- és sűrűségfüggvény.	52
2.1.4. Valószínűségi változó monoton függvényének eloszlása	57
2.1.5. Két valószínűségi változó összegének eloszlása	58
2.1.6. Két független valószínűségi változó szorzatának és hányadosának eloszlása.....	60
2.1.7. Valószínűségeloszlások jellemző adatai..	60
a) A várható érték fogalma és tulajdonságai	61
b) A feltételes várható érték.....	66
c) A szórás fogalma és tulajdonságai....	68
d) A valószínűségi változó momentumai...	71
e) A korrelációs együttható.....	74
2.1.8. A generátorfüggvény és a karakterisztikus függvény.....	78
a) A generátorfüggvény.....	78
b) A karakterisztikus függvény.....	80
2.2. Nevezetes valószínűség eloszlások.....	83
a) Diszkrét valószínűség eloszlások.....	83
2.2.1. Egy esemény indikátor-változója.....	83
2.2.2. A binomiális eloszlás.....	84
2.2.3. A polinomális eloszlás.....	88
2.2.4. A geometriai eloszlás.....	88
2.2.5. A Poisson-eloszlás.....	90
b) Folytonos eloszlások.....	93

2.2.6.	A normális eloszlás.....	93
2.2.7.	A binomiális eloszlás közelítése normális eloszlással.....	101
2.3.	A nagyszámok törvényei.....	117
2.3.1.	A nagy számok törvényének Bernoulli-féle alakja.....	117
2.3.2.	A központi határeloszlástétel.....	121
3.	Fejezet.....	125
3.1.	A matematikai statisztika elemei.....	125
3.1.1.	A matematikai statisztika tárgya és módszerei.....	125
3.1.2.	A statisztikai minta fogalma.....	128
3.1.3.	A tapasztalati (empirikus) eloszlásfüggvény.....	130
3.1.4.	A gyakorisági- és sűrűséghisztogram.....	132
3.1.5.	A statisztikai függvény fogalma.....	
	A legfontosabb tapasztalati jellemzők...	135
3.1.6.	A rendezett minták elmélete.....	142
3.2.	Becslésemélet.....	147
3.2.1.	A becslés fogalma.....	147
3.2.2.	A becslés tulajdonságai.....	149
	a) Torzítatlan becslések.....	149
	b) Konzisztens becslések.....	151
	c) A becslés hatásfoka.....	153
	d) A Cramér-Rao egyenlőtlenség.....	155
	e) Elégséges becslések.....	161
3.2.3.	A becslés módszerei.....	163
	a) A legnagyobb valószínűség (maximum likelihood) elve.....	163
	b) A momentumok módszere.....	170
3.2.4.	A becslés megbízhatósága. Konfidenciaintervallumok.....	171
	a) Konfidenciaintervallum a normális eloszlás várható értékére.....	171
	b) Konfidenciaintervallum ismeretlen valószínűségre.....	173
	c) Konfidenciaintervallum a normális eloszlás szórásnégyzetére.....	175
3.3.	A statisztikai hipotézisek (feltevések) vizsgálata.....	176
3.3.1.	Elvi megjegyzések.....	176
3.3.2.	A statisztikai próba fogalma. Paraméteres próbák.....	179
3.3.3.	Az U-próba.....	180
3.3.4.	A próba ereje. Erőfüggvény.....	187
3.3.5.	A t-próba (Student-próba).....	189
3.3.6.	Az F-próba.....	192
3.3.7.	A Welch-próba.....	194
3.3.8.	A Bartlett-próba.....	195
3.3.9.	Szekvenciális módszer hipotézisvizsgálata.....	196

3.4. Késztermékek minőségellenőrzésének matematikai statisztikai módszerei.....	205
3.5. Nem paraméteres próbák.....	212
3.5.1. Illeszkedésvizsgálat.....	214
a) χ^2 -próba.....	214
b) A Kolmogorov-próba.....	221
c) Grafikus módszer normalitásvizsgálata.....	224
3.5.2. Homogenitásvizsgálat.....	226
a) Homogenitásvizsgálat χ^2 -próbával...	227
b) A Kolmogorov-Szmirnov-Féle kétmin-tás próba.....	227
c) Homogenitásvizsgálat a Gnyegyenko-Koroljuk próbával.....	230
d) A Wilcoxon-próba.....	235
3.5.3. Függetlenségvizsgálat χ^2 próbával.....	241
3.6. A döntésfüggvények elméletének elemei.....	243
3.6.1. A statisztikai döntési eljárás.....	244
3.6.2. Veszteségfüggvény és kockázatfüggvény..	245
3.6.3. A Bayes-féle döntési elv.....	247
3.7. Korreláció- és regresszióelmélet.....	252
3.7.1. A valószínűségi változók közötti sztochasztikus kapcsolatokról.....	252
3.7.2. Korreláció- és regresszióelmélet két valószínűségi változó esetén.....	252
3.7.3. A regressziós egyenes meghatározása a legkisebb négyzetek módszerével.....	259
3.7.4. A regressziós egyenes paramétereinek becslése.....	260
Korrelációs-együttható becslése.....	263
3.7.5. Regressziós parabolák.....	267
3.7.6. Korreláció- és regresszió kétdimenziós normális eloszlás esetében.....	269
3.7.7. Korrelációs és regressziós több változó esetén.....	275
3.7.8. Parciális korreláció.....	277

1.1 Valószínűségelméleti alapok

1.1.1 A véletlen esemény fogalma

A valószínűségelmélet tárgya a véletlen tömegjelenségek objektív törvényszerűségeinek feltárása, matematikai megfogalmazása és vizsgálata.

Röviden megvilágítjuk, mit értünk véletlen tömegjelenségen. Mindenekelőtt vizsgáljuk meg, hogy mit értünk véletlen jelenségen.

A természeti jelenségek lefolyását általában számos körülmény, ok határozza meg. Vannak a természetben olyan jelenségek, amelyekre vonatkozólag mindazokat a feltételeket, körülményeket számba tudjuk venni, amelyek a jelenség lefolyását befolyásolják. Ismeretes pl., hogy a desztillált víz 760 mm légnyomásnál 100°C -on felforr. Jelöljük A-val azt az eseményt, hogy a víz felforr. Ha fennáll az a három feltétel, hogy a víz kémiailag tiszta, a légnyomás 760 mm és a víz hőmérséklete 100°C , akkor az A esemény szükségszerűen bekövetkezik.

Azokat a jelenségeket, amelyek olyan természetűek, hogy ha a körülmények bizonyos figyelembe vett együttese fennáll, akkor ezek a jelenség lefolyását egyértelműen meghatározzák, determinisztikus sémával leírható jelenségeknek nevezzük. A természeti jelenségek között nagy számban vannak olyanok, amelyek lefolyását a számításba vett (számításba vehető) feltételek, okok együttese határozza meg egyértelműen. Ez azt jelenti, hogy az ilyen jelenségeknél a számításba vett feltételek mellett lehet, hogy bekövetkezik valamely A esemény, de az is lehet, hogy nem következik be (más események következnek be). Az ilyen jelenségeket sztochasztikus sémával leírható jelenségeknek vagy más néven véletlen jelenségeknek nevezzük.

A determinisztikus sémával leírható jelenségek és a sztochasztikus sémával leírható jelenségek között nincs antagonsztikus ellentét. Ha egy determinisztikus sémával leírható jelenség lefolyását egyértelműen meghatározó feltételek valamelyike hiányzik, akkor a jelenség sztochasztikus sémával írható le. Ha pl. valahol egy folyóból kimerünk egy liter vizet és 100°C -ra hevítjük, akkor 760 mm légnyomás mellett is vagy bekövetkezik az A esemény (felforr a víz) vagy nem, attól függően, hogy mennyire szennyezett a víz. Másrészt sztochasztikus sémával leírható - tehát véletlen - jelenség is determinisztikussá válhat, ha mindazokat az okokat fel-

tárja a kutatás, amelyek a jelenség lefolyását meghatározzák és ha módunk van ezeket az okokat (feltételeket) számításba venni (pl. mérni tudjuk őket). Egy folyó vízállásának alakulása egy adott folyó adott vízmércéje meghatározott időintervallumban - egyelőre csak sztochasztikus sémával írható le, tehát véletlen jelenség, mert nem tudjuk számításba venni mindazon okokat és azok bonyolult láncolatát, amelyek a vízállás alakulását befolyásolják. Ha pl. szeretnénk előre látni, hogy a Tisza folyón Szegednél az év első három hónapjában lesz-e árvíz vagy nem, akkor pontosan számba kellene vennünk a vízgyűjtőre lehullott hó és eső csapadékot, az olvadás ütemét, amely a hőmérséklet alakulásától függ, a mellékfolyókból érkező víz mennyiségét, amihez a mellékfolyók vízgyűjtőire lehullott csapadék mennyiségét is tudni kellene stb. Ezen felül még azt is tudni kellene, hogy a Tisza és a mellékfolyók nagyvizei összetalálkoznak-e, tehát a nagyvíztömegek pontos levonulási ütemét is előre kellene látni. Ha mindezeket az árvizet okozó tényezőket egyes években pontosan számításba is tudnánk venni, nem sokra mennénk vele, mert ezek a feltételek évről évre változnak. Ilyenformán az árvizek alakulását (egyelőre) sztochasztikus sémával írhatjuk le.

Mindez nem jelenti azt, hogy a sztochasztikus sémával leírható jelenségek okozatilag kevésbé meghatározottak, mint a determinisztikus sémával leírhatók. A véletlen jelenségeknek éppúgy megvannak az okai, mint a determinisztikus jelenségeknek, csak éppen nem tudjuk számbavenni mindazokat az okokat (vagy az okok bonyolult láncolatát), amelyek a jelenség lefolyását meghatározzák. A "véletlen" és az "okozatilag meghatározott" jelzők tehát egyáltalán nem ellentétesek. A "véletlen" a "szükségszerűség" ellentéte, ami azt jelenti, hogy A esemény nem szükségszerűen következik be, lehet, hogy bekövetkezik, de az is lehet, hogy nem következik be. Tehát olyan jelenséget nevezünk véletlen jelenségnek, amelynek lefolyását a figyelembe vett feltételek, okok nem határozzák meg egyértelműleg.

A tömegjelenség kifejezése olyan folyamatokra vonatkozik, amelyek (nagyjából) azonos körülmények között igen sokszor (elvileg akárhányszor) megfigyelhetők.

A valószínűségszámítás csak olyan események vizsgálatával foglalkozik, amelyek bekövetkezése vagy be nem következése azonos körülmények között sokszor megfigyelhető. Az ilyen véletlen tömegjelenségekben statisztikai törvényszerűségek mutatkoznak és éppen az ilyen törvényszerűségek feltárása a valószínűségszámítás feladata. A statisztikai törvényszerűségek feltárásával nem küszöbölhetjük ki a véletlen játékát a természeti folyamatokból, mégis e törvényszerűségek megismerése és figyelembevétele jelentős előre látást tesz lehetővé nagy tömegű véletlen jelenségek kimene-teleinek megoszlását illetőleg.

A valószínűségelmélet és a matematikai statisztika a véletlen elleni harc jelentős eszköze. Egyszeri, nem ismétlődő

véletlen jelenségekkel kapcsolatban a tudomány nem tehet mást, mint megállapítja azok véletlen jellegét. Ezért egy-szeri véletlen jelenségekkel a valószínűségelmélet nem fog-lalkozik.

A véletlen eseményekkel kapcsolatos eddigi fejtegetése-ink jobbára filozófiai jellegűek voltak. Rátérünk most a véletlen események matematikai megközelítésére.

Tárgyalásaink során a véletlen jelenségekre vonatkozó minden megfigyelést, mérést közös névvel véletlen kísérlet-nek fogunk nevezni.^x

A megfigyelés eredménye a kísérlet kimenetele. A kísér-let kimenete a maga egészében általában igen bonyolultan le-írható objektum, ezért úgy járunk el, hogy kiválasztunk a folyamattal kapcsolatban egy vagy több jellemző ismérvet (pl. az árhullámmal kapcsolatban a tetőzési értéket, időtartamot, vízhozamot) és ezek numerikus értékét tekintjük a kísérlet kimenetelének. Ha például a kísérlet egy folyó vízállásának megmérése adott helyen és időben, akkor a kísérlet kimenete-le egy adat, a mért vízállásérték. Ennek a kísérletnek annyi lehetséges kimenetele van, ahány különböző vízállása az adott helyen előfordulhat 0 és egy ésszerűen választott k felső-határ között. Annak ellenére, hogy a vízállást a gyakorlat-ban cm-es pontossággal mérjük, valójában ezen kísérlet összes lehetséges kimeneteleinek halmaza a $(0, k)$ intervallum. A víz-állás mérésének mint véletlen kísérletnek tehát végtelen sok (kontinuum-sok) kimenetele van. A $(0, k)$ intervallum tetsző-leges pontját a kísérlet elemi kimenetelének vagy szokásos terminológiával elemi eseménynek nevezzük. A kísérlet összes elemi eseményeinek halmazát eseménytérnek nevezzük. Van bi-zonyos önkény abban, hogy egy kísérlettel kapcsolatban mit tekintünk elemi eseménynek (és ennek következtében a kísér-let eseményterének), ezért ezt minden kísérlettel kapcsolat-ban pontosan meg kell mondanunk. Az eseménytér a vizsgált jelenséggel kapcsolatban a valóságnak egyszerűsített modell-je, vetülete. A lényeges itt az, hogy ez a modell a gyakor-lat számára érdekességgel bíró szempontból, a vizsgált prob-léma szempontjából a valóságot helyesen tükrözze vissza.

Ha a kísérletünk abból állt, hogy megfigyeljük egy adott év március hónapjában az esős napok számát egy egy adott he-lyen (megállapodván pl. hogy esős napnak számítunk minden olyan napot, amikor legalább 5 mm csapadék hullott), akkor ennek a kísérletnek a lehetséges kimenetelei, elemei esemé-nyei a $0, 1, 2, \dots, 31$ számok. E számok együttese alkotja most az eseményteret.

Valamely kísérlettel kapcsolatban az elemi eseményeken kívül más eseményeket is megfogalmazhatunk. A Duna folyó jég-

^xA "véletlen" jelzöt a kísérlet szó mellől elhagyjuk a to-vábbiakban, mivel könyvünkben csak véletlen kísérletekkel foglalkozunk.

mentes nagyvízi vízállására vonatkozólag Pozsonynál hosszú évek során megfigyeléseket végeztek és azt tapasztalták, hogy a vízállás 4 m és 10 m között változott. A vízállás ingadozásának vizsgálatakor rendszerint nem az érdekel bennünket, hogy milyen gyakran volt a vízállás pl. pontosan 637 cm (lehet, hogy egyszer sem mértek pontosan ennyit), hanem sokkal érdekesebb pl., hogy milyen gyakran volt a vízállás nagyobb mint 7 m, milyen gyakran esett a vízállás értéke 8 m és 10 m közé stb.

Ennél a kísérletnél az eseménytér a számegyenes valamely $(0, k)$ intervalluma (ahol valamely ésszerűen választott felső határ). Az eseményteret az $\mathcal{X}$ szimbólummal jelöljük, amely jelen esetben egy intervallum, vagyis az

$$\mathcal{X} = \{x: x \in [0, k]\} \quad \text{halmaz.}$$

Az $\mathcal{X}$ halmaz minden $x \in \mathcal{X}$ pontja egy elemi esemény. Az eseménytérnek tetszőleges részhalmazát eseménynek nevezzük.

Az eseményeket latin $A, B, C, \dots$ betűkkel fogjuk jelölni.

Magát az $\mathcal{X}$ halmazt biztos eseménynek nevezzük és I -vel jelöljük, mivel adott kísérletnél valamelyik x elemi esemény bekövetkezik. Amikor a mért vízállás x értéke a $[8 \text{ m}, 10 \text{ m}]$ intervallumba esik, akkor azt mondjuk, hogy bekövetkezik az

$$A = \{x: x \in [8 \text{ m}, 10 \text{ m}]\} \quad \text{esemény.}$$

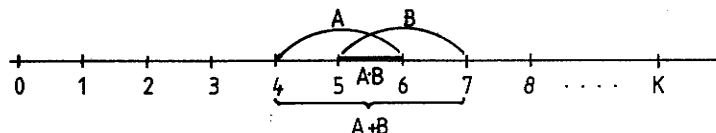
Minthogy az esemény matematikai modellje a halmaz, beszélhetünk események összegéről, szorzatáról, egy esemény ellentétéről (komplementer eseményéről), amely fogalmak mindegyikének egy-egy halmazalgebrai (Boole-algebrai) fogalom felel meg.

$$\text{Ha pl. } A = \{x: x \in [4 \text{ m}; 6 \text{ m}]\}, B = \{x: x \in [5 \text{ m}; 7 \text{ m}]\}$$

akkor az $A+B$ esemény mindannyiszor bekövetkezik, valahányszor az x vízállás-érték a $[4 \text{ m}, 7 \text{ m}]$ intervallumba esik, azaz:

$$A+B = \{x: x \in [4 \text{ m}; 7 \text{ m}]\} = \{x: x \in [4 \text{ m}; 6 \text{ m}]\} \cup \{x: x \in [5 \text{ m}; 7 \text{ m}]\}$$

$$A \cdot B = \{x: x \in [4 \text{ m}; 6 \text{ m}]\} \cap \{x: x \in [5 \text{ m}; 7 \text{ m}]\} = \{x: x \in [5 \text{ m}; 6 \text{ m}]\}$$



1. ábra

Két esemény szorzatán tehát a két esemény egyidejű bekövetkezését értjük, aminek halmazalgebrai szempontból az A és B halmazok közös része (metszete) felel meg.

Az A esemény ellentétes vagy komplementer eseményének azt az $\bar{A}$ eseményt nevezzük, amely az $\bar{X}$ alaphalmaz mindazon s pontjaiból áll, amelyek nem tartoznak A-ba:

$$\bar{A} = \{x: x \notin A\} = \bar{X} - A$$

Ha tehát pl. mint az előző példában $A = \{x: x \in [4 \text{ m}; 6 \text{ m}]\}$, akkor

$$\bar{A} = \{x: x \in [0, 4 \text{ m}] \cup [6 \text{ m}; K]\}$$

Természetesen $\bar{\bar{A}} = A$, vagyis egy A esemény ellentétének ellentéte maga az A esemény.

Az I biztos esemény ellentétét lehetetlen eseménynek nevezzük és a $\emptyset$ szimbólummal jelöljük:

$$\bar{I} = \emptyset; \quad \bar{\emptyset} = I$$

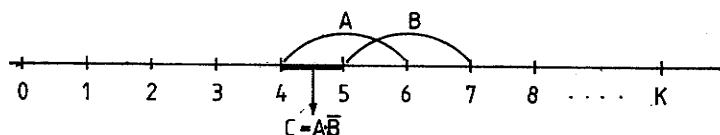
$$\text{Ha } A = \{x: x \in [4 \text{ m}; 6 \text{ m}]\} \quad \text{és} \quad B = \{x: x \in [5 \text{ m}; 7 \text{ m}]\}$$

és a vízállás mérés során pl. $x = 450 \text{ cm}$, akkor az A esemény bekövetkezett, B pedig nem. Ilyenkor azt mondjuk, hogy a

$$C = A - B = A\bar{B}$$

esemény következett be.

Könnyű belátni, hogy $A+B = B+A\bar{B} = B+C$



2. ábra

a B és C események egymást kizárják, mert egy kísérlet során ha B bekövetkezett, C nem következhet be és viszont. Az egymást kizáró eseményeknek diszjunkt (közös pont nélküli) halmazok felelnek meg. Két esemény összege mindig előállítható két egymást kizáró esemény összegeként. Azt a tényt, hogy B és C egymást kizáró események a $B \cdot C = \emptyset$ szimbólummal fejezzük ki.

A fenti példa kapcsán azt is könnyű belátni, hogy

$$A+B = A+A\bar{B} \quad \text{továbbá, hogy} \quad B = AB+A\bar{B}$$

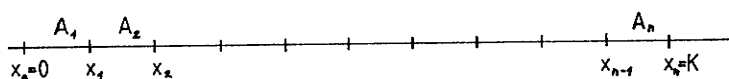
A példánkban szereplő $C = \{x: x \in [4 \text{ m}; 5 \text{ m}]\}$

esemény benne van az $A = \{x: x \in [4 \text{ m}; 6 \text{ m}]\}$

eseményben, tehát valahányszor a C esemény bekövetkezik, bekövetkezik az A esemény is. Azt mondjuk, hogy a C esemény bekövetkezése maga után vonja az A esemény bekövetkezését, amit a $C \subset A$ szimbólummal jelölünk.

Könnyű belátni, hogy ha $C \subset A$, akkor $AC = C$, továbbá $A+C=A$.

Osszuk fel a példaként említett $[0;K]$ intervallumot, az $\mathbb{R}$ alaphalmazt az $0=x_0 < x_1 < \dots < x_m = K$ 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:

$$A_i \cdot A_j = \emptyset, \text{ ha } i \neq j$$

és közülük egy kísérlet során egy és csakis egy feltétlenül bekövetkezik:

$$A_1 + A_2 + \dots + A_n = I \quad \text{Az } A_1, A_2, \dots, A_n$$

eseményekből álló eseményrendszert teljes eseményrendszernek nevezzük.

A teljes eseményrendszert röviden az eseménytér particiójának nevezzük.

1.1.2 A valószínűség fogalma

Valószínűségyszámítási axiómák

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. Valamely kísérlettel kapcsolatban a különböző bennünket érdeklő események valószínűségét igyekszünk számszerű formában kifejezni, egy mérőskálát bevezetni a valószínűségre vonatkozólag, mint ahogyan a hőmérsékletet, súlyt stb. mérjük. Erre a törekvésre elsősorban nyilván az indít bennünket, hogy a

jövőre vonatkozólag bizonyos előrelátásunk legyen, hogy a bennünket érdeklő esemény kb. a megfigyelések (mérések) hány százalékában fog bekövetkezni.

Valamely esemény valószínűségének fogalmát egy másik fogalom, az illető esemény relatív gyakoriságának fogalma segítségével közelítjük meg.

Tekintsünk egy kísérletet, amely nagyjából azonos körülmények között sokszor megismételhető, s amely kísérlettel kapcsolatban az érdekel bennünket, hogy valamely A esemény bekövetkezik-e vagy sem. Tegyük fel, hogy a kísérletet n -szer megismételve az A esemény K_n -szer következik be. Feltetelezzük, hogy az a körülmény, hogy a kísérlet végrehajtása során az A esemény bekövetkezett, nem befolyásolja a kísérlet következő végrehajtásának kimenetelét. A K_n számot az A esemény gyakoriságának, a $\frac{K_n}{n}$ hányadost az A esemény relatív gyakoriságának nevezzük n kísérlet során.

Valamely A esemény $\frac{K_n}{n}$ relatív gyakorisága n nagy értékei esetén meglepő stabilitást mutat. Ha pl. sokszor feldobunk egy szabályos érmét, az eseteknek kb. a felében "fej"-et dobunk. Erre vonatkozólag megemlítjük, hogy Buffon 4040 dobás során 2048 esetben kapott "fej"-et, Pearson 24000 dobásnál a "fej" dobások számát 12012-nek találta. A-val jelölve azt az eseményt, hogy az érmét feldobva "fej"-et kapunk, azt mondhatjuk, hogy a fenti példában szereplő A esemény relatív gyakorisága az $\frac{1}{2}$ szám körül ingadozik és nagy n esetén attól kevésbé tér el.

Azt a törvényszerűséget, hogy bizonyos kísérletet igen sokszor - azonos körülmények között - elvégezve, valamely esemény relatív gyakorisága stabilitást mutat, a nagy számok törvényének nevezzük. (Lásd: 2.3 pont) Ez a stabilitás abban mutatkozik meg, hogy valamely A esemény hosszú kísérletsorozatból számított relatív gyakoriságai gyakorlatilag ugyanakkorák, ugyanazon meghatározott szám körül ingadoznak, amely számot az illető A esemény valószínűségének nevezzük és P/A -val jelöljük.

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

Azt a tényt, hogy az A esemény $\frac{K_n}{n}$ relatív gyakorisága, ha n igen nagy szám közelítőleg megegyezik az A esemény P/A valószínűségével, a

$$g/A/ = \frac{K_n}{n} \approx P/A/$$

szimbóllumal jelöljük.

A valószínűség numerikus értékének a relatív gyakoriság segítségével történő közelítését ("mérését") illetőleg meg kell jegyeznünk, hogy egy esemény relatív gyakorisága tehát csak közelíti meg az illető esemény valószínűségét, ha az egyes kísérletek kimenetelei nem befolyásolják egymást (ezt úgy mondjuk, hogy a kísérletet sokszor végrehajtva az egyes kimenetek "függetlenek" egymástól), továbbá ha a kísérletet azonos körülmények között hajtjuk végre. A Duna évi maximális vízállásai általában függetleneknek tekinthetők, hiszen a jövő évi maximális vízállás aligha függ attól, hogy mekkora a jelen évben vagy az elmúlt években mért évi maximális vízállás. A közölt példában nehezebb annak megítélése, hogy a Duna vízállásának alakulását befolyásoló tényezőkben 1901 és 1970 között történt-e lényeges változás. Ennek eldöntéséhez igen sok körülményt kellene figyelembe vennünk. Ha nincs tudomásunk jelentősebb folyószabályozásról, a környezeti hatások lényeges változásáról, akkor magának az évi maximális vízállásadatsornak a statisztikai elemzése ún. "egyöntetűségvizsgálat" útján válaszolhatunk a kérdésre. (Lásd: 3.5.2 pont.)

Megjegyezzük, hogy egy A esemény $\frac{K_n}{n}$ relatív gyakorisága véletlen mennyiség. Ha sok n hosszúságú kísérletsorozatot végzünk és minden egyes sorozatban megszámláljuk az A esemény K/A gyakoriságát, valamint kiszámítjuk a $\frac{K_1}{n}, \frac{K_2}{n}, \dots, \frac{K_i}{n}, \dots$ relatív gyakoriságot, akkor ezek stabilitásuk ellenére is különböznek egymástól. A relatív gyakoriság tehát véletlen mennyiség, amelynek értéke kísérletsorozatonként változhat, míg az A esemény P/A valószínűsége objektív mértékszám, tehát egy esemény relatív gyakorisága és valószínűsége között ilyenformán elvi különbség van.

Mivel elég nagyszámú kísérletnél a relatív gyakoriság igen kevésbé tér el a valószínűségtől, a valószínűség tulajdonságait a relatív gyakoriság tulajdonságaiból származtatják.

A relatív gyakoriságra vonatkozólag mindig teljesülnek az alábbi összefüggések:

1. Tetszőleges A esemény g/A relatív gyakoriságára

$$0 \leq g/A \leq 1$$

2. Az I biztos esemény (amely minden kísérletnél be következik) relatív gyakorisága:

$$g(I) = 1$$

3. Ha A és B egymást kizáró események, akkor

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

Ez a tulajdonság tetszőleges véges sok egymást kizáró esemény egyesítésére igaz.

A relatív gyakoriság ezen tulajdonságait alapul véve valamely A esemény $P(A)$ valószínűségén 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 $A_1, A_2, \dots, A_n, \dots$ egymást kizáró

események, azaz $i \neq j$ esetén $A_i A_j = \emptyset$

akkor

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

Megemlítünk most néhány olyan összefüggést, amelyek az axiómák közvetlen következményei s amelyek a továbbiakban hasznosnak fognak bizonyulni, mert bizonyos események valószínűségeit ismerve lehetővé teszik más - esetleg összetettebb - események valószínűségének meghatározását.

1. Tétel:

Ha az A esemény valószínűsége $P(A)$, akkor az $\bar{A}$ ellentétes esemény valószínűsége:

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

A tétel belátásához vegyük figyelembe, hogy

$$A + \bar{A} = I, \quad A \cdot \bar{A} = \emptyset$$

így

a II. és III. axiómák alapján

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

amiből már adódik a tétel állítása.

Az 1. tétel alapján következik, hogy a $\emptyset$ lehetetlen esemény valószínűsége 0.²

² Megjegyezzük: abból a tényből, hogy valamely A esemény valószínűsége 0, nem következik, hogy A lehetetlen esemény!

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

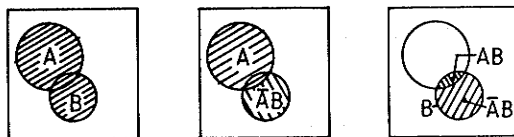
2. Tétel:

Ha A és B tetszőleges események, akkor annak valószínűsége, hogy közülük legalább egyik bekövetkezik:

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

Az állítás belátásához fejezzük ki mind az $(A+B)$ eseményt, mind a B eseményt két diszjunkt esemény összegeként. Az 5. ábra alapján is könnyű belátni, hogy

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



4. ábra

A II. axióma alapján:

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

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

A két egyenletből kivonással nyerjük a

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

amiből a 2. tétel állítása következik.

A 2. tételből következik, hogy

$$P(A+B) \leq P(A) + P(B) \quad \text{továbbá, ha}$$

$$P(A) + P(B) > 1, \quad \text{akkor}$$

$$P(AB) \geq P(A) + P(B) - 1$$

3. Tétel:

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

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) - \sum_{i_1 < i_2} P(A_{i_1} A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} A_{i_2} A_{i_3}) - \dots + (-1)^{n+1} \sum_{i_1 < i_2 < \dots < i_n} P(A_{i_1} A_{i_2} \dots A_{i_n})$$

Ezt az összefüggést teljes indukcióval láthatjuk be. Az állítás $n=2$ esetén igaz a 2. tétel alapján. Tételezzük fel, hogy

$$P(A_2 + A_3 + \dots + A_n) = \sum_{i=2}^n (P(A_i) - \sum_{2 \leq i_1 < i_2} P(A_{i_1} A_{i_2}) + \sum_{2 \leq i_1 < i_2 < i_3} P(A_{i_1} A_{i_2} A_{i_3}) - \dots)$$

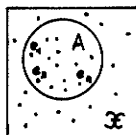
Továbbá:

$$P(A_1 A_2 + A_1 A_3 + \dots + A_1 A_n) = \sum_{i=2}^n P(A_1 A_i) - \sum_{2 \leq i_1 < i_2} P(A_1 A_{i_1} A_{i_2}) + \sum_{2 \leq i_1 < i_2 < i_3} P(A_1 A_{i_1} A_{i_2} A_{i_3}) - \dots$$

Alkalmazva a 2. tételt:

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2 + \dots + A_n) - P(A_1 A_2 + A_1 A_3 + \dots + A_1 A_n) = \\ = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i_1 < i_2} P(A_{i_1} A_{i_2}) + \sum_{1 \leq i_1 < i_2 < i_3} P(A_{i_1} A_{i_2} A_{i_3}) - \dots$$

A következő tétel ismertetése előtt egy fontos - megjegyzést teszünk, amely véges - vagy megszámlálható - sok - elemi eseményből álló eseményterekre vonatkozólag érvényes: bármely esemény valószínűsége az őt alkotó elemi események valószínűségeinek összegével egyenlő. Álljon az A esemény az $e_1, e_2, \dots, e_m$ elemi eseményekből. Az elemi események nyilván kizárják egymást, hiszen egy kísérlet során csak egyetlen elemi esemény következhet be.



5. ábra

$$P(A) = P(e_1 + e_2 + \dots + e_m) = \sum_{i=1}^m P(e_i)$$

Tekintsünk most egy olyan kísérletet, amelynek véges n számú lehetséges kimenetele van, legyenek ezek $e_1, e_2, \dots, e_n$ és legyen mindegyik kimenete egyforma valószínűségű:

$$P(e_i) = \frac{1}{n} \quad (i = 1, 2, \dots, n)$$

Ha most B a kísérlettel kapcsolatos olyan esemény, amely k számú elemei eseményből áll, akkor megjegyzésünk alapján:

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

Ezzel a következő tételt nyertük:

4. Tétel:

Ha egy kísérletben n elemi esemény valamelyike következhet be és mindegyik elemi esemény egyforma valószínűségű, akkor k - számú - elemi eseményt tartalmazó B esemény valószínűsége:

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

Megjegyezzük, hogy a valószínűségszámítás megszületésének kezdeti időszakában főleg olyan kísérleteket vizsgáltak, amelyeknek véges sok kimenetele van és minden eset egyforma valószínűségű.

A következő tétel a valószínűségnek a monotonitását fejezi ki, amely szerint valamely halmaz valószínűségi mértéke nem kisebb bármely részhalmazának valószínűségi mértékénél.

5. Tétel:

Ha $A \supset B$, akkor $P(A) \geq P(B)$

A tétel belátáshoz vegyük figyelembe, hogy ha $A \supset B$, akkor $A = B + \bar{A}B$
tehát:

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

$$\text{ahol } P(\bar{A}B) \geq 0$$

1.1.3 Kombinatorikai alapok

A valószínűségszámításban és a modern matematikai statisztikában nagy szerepet játszanak egyszerű kombinatorikus megfontolások. Ezért röviden összefoglaljuk a legegyszerűbb kombinatorikai alapfogalmakat, amelyekre a továbbiakban számmost esetben hivatkozni fogunk.

A kombinatorika véges halmazokra vonatkozó megszámlálási problémákkal foglalkozik. Legyen az A halmaz az első n természetes szám halmaza:

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

a) Permutációk.

Az első probléma, amelyet halmazzal kapcsolatban vizsgálunk, hogy hányféle sorrendben tudjuk felírni az A halmaz elemeit, az $1, 2, \dots, n$ számokat úgy, hogy mindegyik szám csak egyszer szerepelhet. Az első helyre az n darab szám bármelyikét írhatjuk, a második helyre a fennmaradó

Faktorálisok táblázata

n	n!	n	n!
1	1	26	40 329 146 10^{19}
2	2	27	10 888 869 10^{21}
3	6	28	30 488 834 10^{22}
4	24	29	88 417 620 10^{23}
5	120	30	26 525 286 10^{25}
6	720	31	82 228 387 10^{26}
7	5 040	32	26 313 084 10^{28}
8	40 320	33	86 833 176 10^{29}
9	362 880	34	29 525 280 10^{31}
10	3 628 800	35	10 333 148 10^{33}
11	39 916 800	36	37 199 333 10^{34}
12	47 900 160 10	37	13 763 753 10^{36}
13	62 270 280 10^2	38	52 302 262 10^{37}
14	87 178 291 10^3	39	20 397 882 10^{39}
15	13 076 774 10^5	40	81 591 528 10^{40}
16	20 922 790 10^6	41	33 452 527 10^{42}
17	35 568 743 10^7	42	14 050 061 10^{44}
18	64 023 737 10^8	43	60 415 263 10^{45}
19	12 164 510 10^{10}	44	26 582 716 10^{47}
20	24 329 020 10^{11}	45	11 962 222 10^{49}
21	51 090 942 10^{12}	46	55 026 222 10^{50}
22	11 240 007 10^{14}	47	25 862 324 10^{52}
23	25 852 017 10^{15}	48	12 413 916 10^{54}
24	62 044 840 10^{16}	49	60 828 186 10^{55}
25	15 511 210 10^{18}	50	30 414 093 10^{57}

n-1 darab szám bármelyikét, harmadik helyre a még nem írt
n-2 darab szám bármelyike kerülhet stb.

Az összes lehetséges sorrendek száma:

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

Az 1, 2, ..., n számok minden egyes sorrendjét egy permutációnak nevezzük.

Jelöljük $\mathcal{P}_n$ -nel az $1, 2, \dots, n$ számból alkotható összes lehetséges permutációk halmazát.

A $\mathcal{P}_n$ elemeinek számát P_n -nel jelöljük. Ekkor:

$$P_n = n(n-1) \dots 3 \cdot 2 \cdot 1 = n! \quad (n \text{ faktoriális})$$

Tehát $n!$ értékét úgy számítjuk ki, hogy 1-től n -ig a természetes számokat összeszorozzuk.
($0! = 1$. Ez megállapodás, amely később célszerűnek fog mutatkozni.)

Az összes n -elemű permutációk $n!$ száma igen gyorsan növekvő függvénye n -nek.

A permutációk számának értékeit a 1.1 táblázat tartalmazza.

Ha n értéke nagy, az $n!$ értéke rendkívül nagy, olyanira, hogy pontos értéke rendszerint nem túl érdekes számunkra, inkább csak a nagyságrendje. Felmerült ezért annak szükségessége, hogyan lehet az $n!$ értékét jó közelítéssel becsülni. 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).

$$Pl. n = 10 \text{ esetén } \left(\frac{10}{e}\right)^{10} \sqrt{20\pi} = 3\,598\,600$$

Ezt összehasonlítva $10! = 3\,628\,800$ pontos értékével, könnyen kiszámítható, hogy a relatív hiba: 0,8 %. Minél nagyobb az n szám, annál jobb közelítést eredményez $n!$ értékére vonatkozólag a Stirling-formula, azaz annál kisebb a közelítés relatív hibája.

b) Ismétléses permutációk

Tegyük fel, hogy az $A = \{a_1, a_2, \dots, a_n\}$ halmaz elemei között rendre $k_1, k_2, \dots, k_r$ számú elem egyforma, ahol $k_1 + k_2 + \dots + k_r = n$.

Ha most az A halmaz elemeit minden lehetséges sorrendben felírjuk, akkor az összes megkülönböztethető sorrendek száma:

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

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

Az összes megkülönböztethető sorrendek száma:

$$\frac{10!}{2! \cdot 3! \cdot 2! \cdot 1! \cdot 1!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 6 \cdot 2} = 151200$$

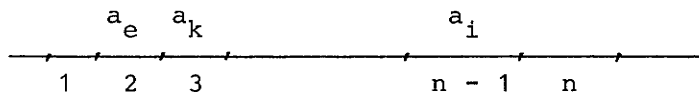
c) Variációk

Az $A = \{1, 2, \dots, n\}$ halmaz elemeiből válasszunk ki egymás után k darab elemet és írjuk fel őket a kiválasztás sorrendjében. Az így nyert szám k -ast variációnak nevezzük. Hány k -elemű variáció képezhető az A halmaz elemeiből? Az első helyre n db elem bármelyikét választhatjuk, a második helyre $n-1$ darab elem bármelyikét, a harmadik helyre $(n-2)$ darab elem (a megmaradt) bármelyikét stb. A k -adik helyre $(n-k+1)$ számú elem bármelyikét választhatjuk. Ilyenformán az A halmazból képezhető k -elemű variációk száma:

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

Tekintsük most a következő elhelyezési problémát: legyen adva k darab számozott elem: $a_1, a_2, \dots, a_k$.

Helyezzük el őket n db. számozott cellába ($n \geq k$), de úgy, hogy egy cellába legfeljebb egy tárgy kerülhet. Hányféle elhelyezkedés lehetséges?



Az a_1 elemet n cella bármelyikébe tehetjük, az a_2 elemet $(n-1)$ db cella bármelyikébe stb. Az összes lehetséges elhelyezkedések száma:

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

d) Ismétléses variációk

Tekintsük az előző elhelyezési problémát, de úgy, hogy nem kötjük ki, hogy egy cellába csak egy elem kerülhet.

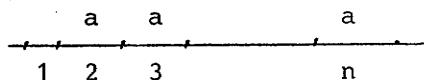
(Akár mindegyik elem is ugyanazon cellába kerülhet), tehát az $a_1, a_2, \dots, a_k$ elemek bármelyike részére bármely cellát választhatjuk. Hányféle elhelyezkedés lehetséges ily módon?

Az a_1 elem számára n cella bármelyikét, az a_2 elem számára ugyancsak az n db cella bármelyikét választhatjuk stb. Az a_1 és a_2 elemek elhelyezésére tehát n^2 lehetőségünk, az a_1, a_2, a_3 elemek elhelyezésére n^3 lehetőségünk van stb. Az $a_1, a_2, \dots, a_k$ elemek összes elhelyezkedési lehetősége: n^k .

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

e) Kombinációk

Tekintsük ismét azt az elhelyezési problémát, hogy az $a_1, a_2, \dots, a_k$ elemeket n cellába helyezzük el úgy, hogy egy cellába legfeljebb egy elem kerülhet, de feltesszük, hogy az $a_1, a_2, \dots, a_k$ elemek mindegyike azonos: $a_1 = a_2 = \dots = a_k$. Úgy is mondhatjuk, hogy k db megkülönböztethetetlen tárgyat helyezünk el n cellába. Hányféle módon lehetséges ez?



A k db megkülönböztethetetlen tárgy valamely elhelyezése n db cellában ekvivalens azzal, hogy n db számozott cella közül k db különböző cellát kiválasztunk, ezek sorrendje azonban közömbös, mivel nem tudjuk megkülönböztetni, hogy milyen sorrendben választottuk ki őket. Minden ilyen elhelyezkedést kombinációnak nevezünk. Az összes megkülönböztethető kombinációk száma most:

$$C_k^n = \frac{n(n-1) \dots (n-k+1)}{1 \cdot 2 \cdot \dots \cdot k} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

A $C_k^n = \binom{n}{k}$ számosság az $A = \{1, 2, \dots, n\}$ halmaz összes k elemű részhalmazainak számát adja meg. Ez azt jelenti, hogy az $1, 2, \dots, n$ számok közül $\binom{n}{k}$ féle módon lehet k db különböző számot kiválasztani.

Alkalmazásként megemlíti a binomiális tételt, amely arra vonatkozik, hogyan kell egy kéttagú kifejezést akárhányszor emelni:

$$(p + q)^n = (p + q) \overset{1}{(p + q)} \overset{2}{(p + q)} \dots \overset{n}{(p + q)}$$

A szorzást úgy végezzük el, hogy minden tényezőtől kiválasztunk egy tagot, ezeket összeszorozzuk, majd az így nyert szorzatokat összeadjuk. Ha minden tényezőtől a q -t választjuk, akkor q^n -t kapjuk. Ilyen kifejezést csak egyféle módon nyerhetünk. Ha k számú tényezőtől a p -t választjuk s a többi $n-k$ tényezőtől a q -t, akkor $p^k q^{n-k}$ típusú tagot kapunk, s ezt annyiféle módon nyerhetjük, ahányféleképpen n tényező közül k különböző kiválasztható, a kiválasztott k számú tényezőtől p -t vesszük, vagyis $\binom{n}{k}$ -féle módon. Tehát

$$(1.1) \quad (p+q)^n = \binom{n}{0} p^0 q^n + \binom{n}{1} p^1 q^{n-1} + \dots + \binom{n}{k} p^k q^{n-k} + \dots + \binom{n}{n} p^n q^0$$

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 $\binom{n}{k}$ értékét nagy n és k értékeire becsülni. 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:

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

A valószínűségelméleti alkalmazások során erre a becslésre többször szükségünk lesz.

Megemlítjük a következő szimmetria-tulajdonságot:

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

Ennek belátása a faktoriálisok segítségével igen egyszerű:

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

Rámutatunk még, hogy az (1.1) összefüggés speciális esetenként a

$$(1.2) \quad (1+1)^n = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = 2^n$$

összefüggést nyerjük.

Az (1.2) összefüggés alapján könnyen válaszolhatunk arra a kérdésre, hogy mennyi az $A = \{1, 2, \dots, n\}$ véges halmaz összes részhalmazainak száma:

- A -nak $\binom{n}{1}$ 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 A részhalmazai közé számítjuk az üres halmazt: $\emptyset$ -t és magát az $A = \{1, 2, \dots, n\}$ halmazt is, akkor az A összes részhalmazainak száma:

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

f) Ismétléses kombinációk.

Az ismétléses kombináció fogalma legegyszerűbben a következő elhelyezési problémával kapcsolatban érthető meg. Van n cellánk 1-től n -ig megszámozva, és k db egyforma (megkülönböztethetetlen) tárgyat helyezünk el bennük. Hányféle módon helyezhető el a k számú megkülönböztethetetlen tárgy az n különböző cellában?

1) Pl. $n = 4$ $k = 5$ esetén egy elhelyezkedés:

· || · · · | ·

6. ábra

Mivel n cellát $n-1$ vonás segítségével reprezentálhatunk, ezt az $n-1$ vonást kell elhelyezni minden lehető (megkülönböztethető) módon az egyforma tárgyakat jelképező k pont között. Minden elhelyezkedés $n-1+k$ számú jel egy konfigurációja, ahol a jelek közül $n-1$, illetve k egyforma. Az összes megkülönböztethető esetek száma nyilván ismétléses permutáció:

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

1.1.4 Példák valószínűségek kombinatorikus kiszámítására

a) Legyen a kísérletünk, hogy egy szabályos pénzérmét (pl. 1 Ft-ot) n -szer feldobunk. Mi a valószínűsége, hogy k -szor dobunk írást?

Gondoljuk el, hogy elvégezzük az n dobást és minden dobás után 1-est írunk, ha írást dobtunk és 0-t írunk ha fejet dobtunk. Minthogy minden dobás eredménye kétféle lehet, az összes esetek száma nyilván 2^n (ismétléses variáció). A kedvező esetek számának meghatározásához azt kell megállapítanunk, hogy a 0 és 1 jegyekből hány olyan n hosszúságú sorozat képezhető, amely pontosan k darab 1-est tartalmaz.

Az ilyen sorozatok száma $\binom{n}{k}$ hiszen ennyi módon tudunk kiválasztani n hely közül k -helyet, a kiválasztott helyekre 1-est írunk, az összes többi $(n-k)$ helyére 0-t.

A keresett valószínűség:

$$P_k = \frac{\binom{n}{k}}{2^n}$$

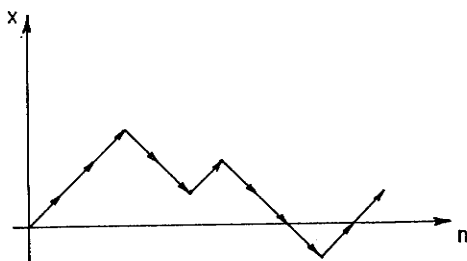
Látható, hogy a keresett valószínűség k -tól való függését a számláló fejezi ki. Minthogy a binomális együtthatók sorozata

1 Most tehát egy cellába több tárgy is kerülhet (akár mindegyik tárgy is egy cellába kerülhet).

zata k -val együtt eleninté nő, majd csökken és legnagyobb értékét $k = \left\lceil \frac{n}{2} \right\rceil$ esetén veszi fel, legvalószínűbb az az eset, hogy a dobások fele eredményez írást.

b) Bolyongás a számegyenesen

Mozogjon egy bábú a számegyenes egész koordinátájú pontjain úgy, hogy az origóból indulva jobbra vagy balra lép egy egységgel $\frac{1}{2} - \frac{1}{2}$ valószínűséggel. Jobbra lépés esetén $(+1)$ -gyel, balra lépés esetén (-1) -gyel változik a bábú koordinátája. A bábú útján szemléltethetjük úgy, hogy a vízszintes tengelyen a lépésszámot, a függőleges tengelyen pedig az elmozdulást tüntetjük fel, s jobbra lépése esetén felfelé, balra lépés esetén lefelé (45° -os szög alatti) vektort rajzolunk.



7. ábra

Mi a valószínűsége, hogy n lépés után a bábút az $x = \ell$ pontban találjuk?

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

Ekkor:

$$a + b = n$$

$$a - b = \ell$$

$$\frac{a + b}{2} = \frac{n}{2}, \quad \frac{a - b}{2} = \frac{\ell}{2}$$

A bábút tehát akkor találjuk az $x = \ell$ pontban, ha $\frac{n + \ell}{2}$ - szer lépett jobbra és $\frac{n - \ell}{2}$ - szer balra. Ez az eset

$\left(\frac{n + \ell}{2}\right)$ -féle módon következhet be. A keresett valószínűség:

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

Mivel mind a jobbra, mind a balra lépések száma csak egész lehet, $\frac{n+1}{2}$ pedig csak akkor egész szám, ha n és 1 vagy mindkettő páros vagy mindkettő páratlan szám, következik, hogy páros számú lépés után csak páratlan számon állhat. Speciálisan az az eset, hogy a bábú visszatér az origóba, azaz kiindulási helyére, csak páros számú lépésben következhet be.

Számítsuk ki annak valószínűségét, hogy a bábú $2n$ számú lépés után visszatér a kiindulási helyére!

$$P_0 = \frac{\binom{2n}{n}}{2^{2n}} = \frac{1}{2^{2n}} \frac{(2n)!}{(n!)^2}$$

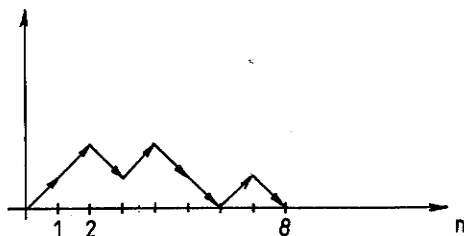
Alkalmazva a Stirling-formulát:

$$P_0 \approx \frac{1}{2^{2n}} \frac{\left(-\frac{2n}{e}\right)^{2n} \sqrt{2\pi \cdot 2n}}{\left(-\frac{n}{e}\right)^{2n} 2\pi n} = \frac{1}{\sqrt{\pi \cdot n}}$$

A következő példa a modern matematikai statisztikai eljárások szempontjából igen hasznos, ezért különösen az olvasó figyelmébe ajánljuk.

c) Az előző példához kapcsolódva, tételezzük fel, hogy a bábú $2n$ lépésben visszatért az origóba. Mi a valószínűsége, hogy a bábú bolyongása során nem érte el vagy lépte túl az $x = k$ pontot? A kérdést úgy is fogalmazhatjuk: mi a valószínűsége, hogy n db $(+1)$ és n db (-1) -ből $2n$ hosszúságú sorozatnak nincs olyan részletösszege, amely nagyobb mint k , ha minden lehetséges sorrend egyforma valószínűségű? A kérdés megválaszolása céljából vezessük be a következő szemléletes interpretációt. Tegyük fel, hogy $2n=8$ esetén a lépések a következők voltak:

$+1, +1, -1, +1, -1, -1, +1, -1$

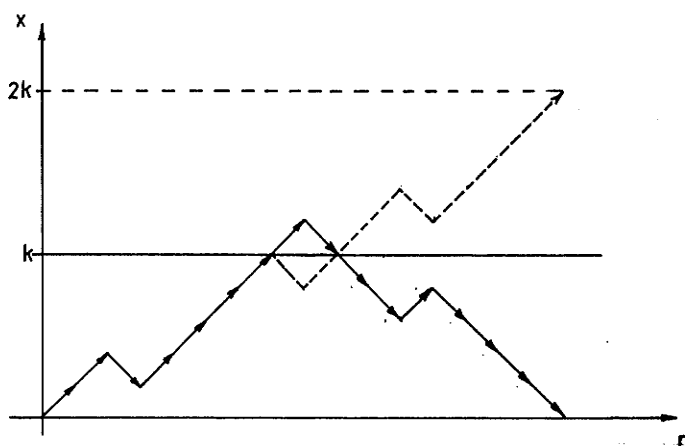


8. ábra

Feleltessünk meg a lépéssorozatnak miként az előző példában, egy vektorsorozatot. Ily módon egy utat, ún. trajektóriát nyerünk. Az összes lehetséges trajektóriák száma n db $+1$

és n db (-1) esetén $\binom{2n}{n}$

A kérdés most az, hogy hány olyan $2n$ vektorból álló trajektória rajzolható (amely a $(0,0)$ ponttól indul és a $(2n,0)$ pontban végződik), amely nem éri el az $x = k$ magasságú egyenest.



9. ábra

Könnyebb meghatározni azon trajektóriák számát, amelyek elérik vagy metszik az $x = k$ magasságú egyenest. Ha ezek számát ismerjük, akkor ezt a számot kivonva $\binom{2n}{n}$ -ből, az összes trajektóriák számából, megkapjuk a k -magasságot el nem érő trajektóriák számát. Az $x = k$ egyenest elérő (vagy metsző) trajektóriák számát a következő egyszerű fogással határozzuk meg. Ha egy trajektóriát, amely az $x = k$ egyenest eléri, az első elérési pontból kezdve tükrözzük az $x = k$ egyenesre, akkor a tükrözött trajektória $(2n, 2k)$ pontban végződik. Ez minden olyan trajektóriára igaz, amely az $x = k$ egyenest eléri. Minden olyan tükrözött út megfelel egy bábú bolyongásának, amely bábú $2n$ lépés után a $2k$ pontban található.

Ez $\binom{2n}{n+k} = \binom{2n}{n-k}$ féle módon lehetséges.

Ezek alapján annak valószínűsége, hogy n db $(+1)$ -ből és n db (-1) -ből álló $2n$ hosszúságú sorozatnak megfelelő trajektória elérje vagy meghaladja az $x = k$ magasságú egyenest:

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

Ezek alapján annak valószínűsége, hogy egy trajektória nem éri el az $x = k$ egyenest:

$$(1.3) \quad \frac{\binom{2n}{n} - \binom{2n}{n+k}}{\binom{2n}{n}} = 1 - \frac{\binom{2n}{n+k}}{\binom{2n}{n}}$$

Megjegyezzük, hogy az (1.3) formula gyakorlati célokra aligha használható, mivel a jobboldalon álló binomiális együtthatók kiszámítása igen nehézkes, viszont nagy n esetén igen aszimptotikus közelítést nyerhetünk a következő módon:

$$\begin{aligned} \frac{\binom{2n}{n+k}}{\binom{2n}{n}} &= \frac{(2n)!}{(n+k)! (n-k)!} \cdot \frac{(n!)^2}{(2n)!} \approx \frac{\left(\frac{n}{e}\right)^{2n}}{2^n n} = \\ &= \frac{1}{\left(1+\frac{k}{n}\right)^{n+k} \left(1-\frac{k}{n}\right)^{n-k} \left(1+\frac{k}{n}\right) \left(1-\frac{k}{n}\right)} \end{aligned}$$

Felhasználva az $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$ sorfejtést, egyszerű számolással kimutathatjuk, hogy

$$(1.4) \quad \frac{\binom{2n}{n+k}}{\binom{2n}{n}} \approx e^{-\frac{k^2}{n}}$$

ami statisztikai alkalmazások szempontjából könnyen kezelhető eredmény és a 6. fejezet (6.4.3) pontjában fel fogjuk használni.

d) Futamok

Az előző bolyongási példával kapcsolatban számos érdekes statisztikai vizsgálatot végezhetünk, amelyeknek igen fontos gyakorlati alkalmazásai vannak (lásd: 6.4 pont). Ilyen vizsgálatok egyike a "futamok" vizsgálata. Futamon egyforma jegyek megszakítatlan sorozatát értjük. Pl. az alábbi

28 +1 +1 -1 -1 +1 -1 -1 +1 +1 +1 -1 +1 -1

sorozatban négy darab (+1)-es futam (2,1,3,1 hosszúságának) és négy darab (-1)-es futam található (2,2,1,1 hosszúságúak). (A bolyongást trajektóriával ábrázolva egy futam a trajektória egy töretlen szakasza).

Tételezzük fel, hogy egy N hosszúságú sorozat n darab 1-est és m darab (-1)-est tartalmaz. Számítsuk ki mi a valószínűsége, hogy az összes futamok száma $R = 2k$ legyen.

Jelöljük R_{+1} -gyel a (+1)-es jegyekből álló futamok számát, R_{-1} -gyel a (-1)-es futamokét.

A sorozat vagy (+1)-gyel vagy (-1)-gyel kezdődik. Tegyük fel, hogy a sorozat (+1)-gyel kezdődik, akkor (-1)-gyel kell végződnie ahhoz, hogy $R = 2k$ lehessen. Mivel a sorozatban n db +1-es szerepel és a +1-es jegyekből álló futamok szempontjából a -1-es futamok csak elválasztó szerepet játszanak, $R_{+1} = k$ annyiféle módon következhet be, ahányféle módon n db (+1)-es jegyből álló sorozat k szakaszra tördelhető. Egy ilyen szakaszokra tördelést $k-1$ számú vonással elvégezhetünk:

pl. $n = 8, k = 3$ esetén egy lehetséges eset:

111 | 111 | 11

Mivel n db (+1)-es között $(n - 1)$ hely található a vonások száma, az $\{R_{+1} = k\}$ esemény $\binom{n-1}{k-1}$ -féle módon következhet be. Hasonlóan látható be, hogy az $\{R_{-1} = k\}$ esemény $\binom{m-1}{k-1}$ -féle módon állhat elő.

Ha most az n db (+1)-es jegyből álló sorozatot valahogyan felbontottuk $k - 1$ vonás segítségével k szakaszra, akkor a vonások helyére beépíthetjük az m db (-1)-es jegyből álló sorozat k -szakaszra történt minden felbontását. Ha tehát a sorozat (+1)-gyel kezdődik és (-1)-gyel végződik, akkor az $\{R = 2k\}$ esemény $\binom{n-1}{k-1} \binom{m-1}{k-1}$ -féle módon következhet be. Világos, hogy ugyanennyi módon következhet be az $\{R = 2k\}$ esemény, ha a sorozat (-1)-gyel kezdődik és (+1)-gyel végződik. Mivel N dobás során az n db (+1)-es és m db (-1)-es $\binom{N}{n}$ -féle módon helyezkedett, a keresett valószínűség:

$$P(R = 2k) = \frac{2 \binom{n-1}{k-1} \binom{m-1}{k-1}}{\binom{N}{n}}$$

Hasonló megfontolásokkal látható be, hogy az $\{R = 2k + 1\}$ esemény bekövetkezésének valószínűsége:

$$P(R = 2k + 1) = \frac{\binom{n-1}{k} \binom{m-1}{k-1} + \binom{n-1}{k-1} \binom{m-1}{k}}{\binom{N}{n}}$$

Ha n és m nagy számok (következik N is az, hiszen $N = n + m$). akkor a fenti valószínűségek kiszámítása rendkívül munkaigényes (Közelítő kiszámításukat lásd: 6.5.1 pont).

1.1.5 Feltételes valószínűség és függetlenség

A feltételes valószínűség fogalma igen fontos és alapvető fogalom elméleti és gyakorlati szempontból egyaránt. E fogalom jelentőségét elsősorban az adja meg, hogy a természetben, így a műszaki folyamatokban a különböző jelenségek általában nem függetlenek egymástól. Bizonyos jelenségek bekövetkezése együtt jár más jelenségek bekövetkezésével. Pl. ha valamely folyó vízgyűjtőjére nagyobb csapadék hullott, akkor általában nagyobb vízállásokra és vízhozamokra számíthatunk. Bizonyos események bekövetkezésének valószínűségét tehát igen jelentős mértékben befolyásolhatja más események bekövetkezése.

Amint egy esemény valószínűségének megvilágításakor a relatív gyakoriság fogalmából indultunk ki, úgy a feltételes valószínűség megközelítéséhez a feltételes relatív gyakoriság fogalmát világítjuk meg.

Az 1.2 táblázatban feltűntettük a Tisza folyó árhullámainak Tokajnál mért túllépéseinek nagyságát a 600 cm-es elsőfokú árvízvédelmi készültségi szint felett, valamint az árhullámok időtartamát (napokig kifejezve).

Az adatok minden év II. negyedévére (ápr. 1-jétől- jún. 30-ig) vonatkoznak 1901-1970 időszakra.

Legyen A az az esemény, hogy az árhullám időtartama meghaladja a 14 napot. A táblázatból olvasható, hogy az A esemény a kérdéses helyen és időszakokban 11-szer következett be 41 túllépés során, tehát az A esemény relatív gyakorisága:

$$g(A) = \frac{K(A)}{n} = \frac{11}{41} \approx \frac{1}{4}$$

Legyen most B az az esemény, hogy a túllépés nagysága meghaladja a 150 cm-t. Példánkban a B esemény gyakorisága, mint az az ábráról könnyen megszámlálható $K(B) = 12$. Számoljuk meg, hogy a 150 cm-nél nagyobb túllépések közül hány tartott 14 napnál tovább, azaz hányszor következett be az AB esemény. Ez a szám a

$$K(AB) = 9$$

1.2 TÁBLÁZAT

Tokaj [II. negyedév (ápr.1. - jan. 30.)] árvízi adatok
1900-1971.

Túllépések adatai.

Év:	X(cm)	Y(nap)	Év:	X(cm)	Y(nap)
1907	159	021	1942	024	003
	056	014	1944	051	011
1909	021	006	1945	065	008
1912	125	013		026	005
1913	027	002	1951	013	003
1914	072	007	1952	164	023
1915	058	007	1955	088	005
1916	052	003	1956	071	014
1922	160	017	1958	045	005
	008	001		006	002
	024	003		002	003
1924	202	013	1962	194	026
	128	013	1964	257	015
1932	256	020	1965	123	010
1933	028	001	1967	132	023
1935	006	001	1968	055	012
1937	017	002	1970	151	020
1940	217	012		258	021
	042	004		153	014
1941	133	017			
	185	023			
	116	010			

Az A eseménynek a B eseményre vonatkozó feltételes relatív gyakoriságán a

$$g(A|B) = \frac{K(AB)}{K(B)}$$

hányadost értjük.

Példánkban:

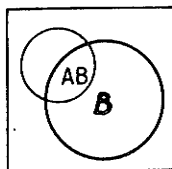
$$g(A|B) = \frac{9}{12} = \frac{3}{4}$$

Láthatjuk, hogy ez esetben az A esemény $g(A|B)$ feltételes relatív gyakorisága kb. háromszor nagyobb, mint az A esemény $g(A)$ relatív gyakorisága, azaz a 150 cm-nél nagyobb túllépések háromnegyed része két hétnél hosszabb ideig tartott.

Minthogy egy árhullám tetőzése általában kb. az időtartam felénél következik be, adott árhullám esetében már a tetőzés

időpontjában nagyjából előre látható a feltételes relatív gyakoriság ismeretében, hogy adott valószínűséggel mennyi idő alatt vonul le az árhullám.

A $g(A|B)$ feltételes relatív gyakoriság tehát azt jelenti, hogy csak azokat a kísérleteket tekintve, amikor a B esemény bekövetkezett, az esetek hány százalékában következett be az A esemény is B-vel együtt.



10. ábra

Hosszú kísérletsorozatoknál a feltételes relatív gyakoriság is stabilitást mutat (csakúgy, mint a relatív gyakoriság). Azt a számot, amely körül a $g(A|B)$ feltételes relatív gyakoriság ingadozik, az A esemény B-re vonatkozó feltételes valószínűségének nevezzük és $P(A|B)$ -vel jelöljük.

Mivel

$$g(A|B) = \frac{K(AB)}{K(B)} = \frac{\frac{K(AB)}{n}}{\frac{K(B)}{n}} \approx \frac{P(AB)}{P(B)}$$

a $P(A|B)$ feltételes valószínűséget a következő összefüggéssel definiáljuk:

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

Innen:

$$(1.5) \quad P(AB) = P(A|B) P(B)$$

Az (1.5) összefüggést a valószínűségek szorzási szabályának nevezzük.

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

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

Ekkor

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

vagyis

$$(1.6) \quad P(AB) = P(A) P(B)$$

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

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

$$(1.7) \quad P(B|A) = \frac{P(AB)}{P(A)}$$

Az (1.5) és (1.7) összefüggésekből adódik, hogy

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

azaz

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

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

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

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

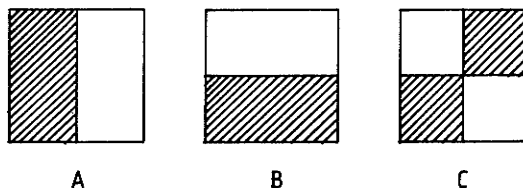
ekkor ugyanis (1.7) alapján

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

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.

A mondottak alapján az (1.6) összefüggés az A és B események függetlenségének definíciójaként tekinthető. Hasonlóképpen definiálhatjuk kettőnél több esemény függetlenségét. Mielőtt erre rátérnénk, rámutatunk arra, hogy már három esemény A, B, C függetlenségének definiálásakor bizonyos óvatosságra van szükség.

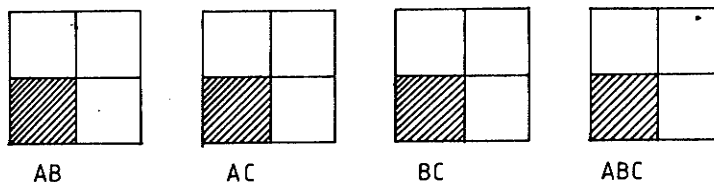
Tekintsük pl. a következő három eseményt:



11. ábra

Legyen

$$P(A) = P(B) = P(C) = \frac{1}{2}$$



12. ábra

Könnyű belátni, hogy

$$P(AB) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A) P(B)$$

$$P(AC) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A) P(C)$$

$$P(BC) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(B) P(C)$$

Ezek alapján az A, B és C események páronként függetlenek. Ugyanakkor

$$P(ABC) = \frac{1}{4} \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = P(A) P(B) P(C)$$

tehát a három esemény együttesen már nem független. Éppen ezért $n > 2$ esetén az $A_1, A_2, \dots, A_n$ események függetlenségét a következő módon definiáljuk:

Az $A_1, A_2, \dots, A_n$ eseményeket teljesen függetleneknek vagy röviden függetleneknek nevezzük, ha a

$$P(A_i A_j) = P(A_i) P(A_j) \quad i < j$$

$$P(A_i A_j A_k) = P(A_i) P(A_j) P(A_k) \quad i < j < k$$

...

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

relációk teljesülnek. Ezek a relációk.

$$\binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - n - 1$$

feltétel teljesülését követelik meg. (Ugyanis a binomiális tétel alapján:

$$\sum_{k=0}^n \binom{n}{k} = 2^n, \quad \binom{n}{0} = 1, \quad \binom{n}{1} = n$$

Kimutatható, hogy ha az $A_1, A_2, \dots, A_n$ események függetlenek, akkor ezek bármelyikét (vagy akár mindegyiket) az ellentétével, A_i -val helyettesítve ugyancsak független eseményrendszert kapunk.

Az (1.5) formulával kifejezve szorzási szabály általánosítható n eseményre. Általános szorzási szabály:

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

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

Itt feltételezzük, hogy a vonások mögött álló, vagy feltételként szereplő eseménysorozatok 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 akkor tekintünk függetlennek, ha kimeneteleik nincsenek egymásra semmiféle befolyással. Ez azt jelenti,

hogy az egyik kísérletnél bekövetkező bármely esemény független a másik kísérletnél bekövetkező bármely eseménytől. A valószínűségszámítási könyvekben általában bizonyos triviális példák találhatók, pl. többször ismételt kockadobás, érmedobás, stb. független kísérletek sorozatát képezik. Ugyanúgy független kísérletsorozatról van szó, ha egy urnában számozott golyók vannak, és találomra kivesszünk egy golyót, feljegyezzük a számát, majd visszatesszük 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.

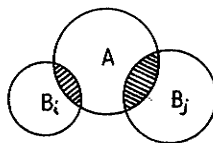
Alkossanak a $B_1, B_2, \dots, B_n$ események teljes eseményrendszert, azaz legyen $B_1 + B_2 + \dots + B_n = I$ és $B_i \cdot B_j = \emptyset$ ha $i \neq j$. Ez azt jelenti, hogy kísérletünk, megfigyelésünk során a B_i események közül egy és csakis egy feltétlenül bekövetkezik. Legyen most A egy tetszőleges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

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

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

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

ugyanis ha B_i és B_j diszjunkt halmazok, akkor AB_i és AB_j mint részhalmazok ugyancsak diszjunktak, s mint események kizárják egymást.



13. ábra

Az (1.5) formula alapján

$$P(AB_i) = P(A|B_i) P(B_i)$$

tehát

$$\begin{aligned} (1.8) \quad P(A) &= P(A|B_1)P(B_1) + \dots + P(A|B_n)P(B_n) = \\ &= \sum_{j=1}^n P(A|B_j)P(B_j) \end{aligned}$$

Az (1.8) összefüggést a teljes valószínűségek tételének nevezzük. Ha ismerjük az A eseménynek mindegyik B_j eseményre vonatkozó feltételes valószínűségét és a B_j események valószínűségét, $(j=1,2,\dots,n)$, akkor kiszámíthatjuk valamely B_i eseménynek az A eseményre vonatkozó feltételes valószínűségét.

Az (1.7) formula alapján:

$$(1.9) \quad P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i) P(B_i)}{P(A)}$$

Behelyettesítve az (1.9) formula nevezőjébe az (1.8) kifejezést:

$$(1.10) \quad P(B_i|A) = \frac{P(A|B_i) P(B_i)}{\sum_{j=1}^n P(A|B_j) P(B_j)}$$

Az (1.10) összefüggést Bayes-féle formulának nevezzük. A Bayes-formula a statisztikai gyakorlatban, különösen a döntéshozatalban igen fontos szerepet játszik.

E fejezet befejezéséül egy megjegyzést teszünk a teljes eseményrendszerekre vonatkozóan. Valamely $B_1, B_2, \dots, B_n$ teljes eseményrendszer adott I alaphalmaz diszjunkt részhalmazokra való felosztását jelenti úgy, hogy I minden pontja benne van valamelyik B_j részhalmazban (és csak abban). Felmerül a kérdés, hogy adott I alaphalmaz esetén hány teljes eseményrendszer képezhető. Ha I végtelen sok elemből (pontból) áll, akkor a belőle képezhető teljes eseményrendszerek (halmazpartíciók) száma természetesen végtelen. Ha I véges halmaz, pl. $I_n = \{1, 2, \dots, n\}$, akkor a belőle képezhető teljes eseményrendszerek száma rekurzióval kiszámítható. Jelöljük az I_n halmazból képezhető teljes eseményrendszerek számát $T(n)$ -nél. Definíáljuk: $T_0 = 1$. Ekkor fennáll a következő rekurzív reláció:

$$T_{n+1} = \sum_{k=0}^n \binom{n}{k} T_{n-k} = \sum_{k=0}^n \binom{n}{k} T_k$$

Ez az összefüggés abból következik, hogy I_{n+1} összes partíciót osztályozhatjuk aszerint, hogy az $(n+1)$ -es szám hány más számmal van egy osztályban. Ha ez a szám k és k értékét rögzítjük, az így kiválasztott partíciókat osztályozhatjuk aszerint, hogy melyik k darab számról van szó az $1, 2, \dots$, számok közül. Ha ezeket is rögzítettük, nyilván annyi ilyen partíció adható meg, ahány partíciója van a többi $(n-k)$ számnak.

2. Fejezet

2.1 Valószínűségi változók és valószínűségeloszlások

2.1.1 A valószínűségi változó és a valószínűségeloszlás fogalma

A műszaki gyakorlatban egy kísérlet kimenetelét igyekszünk numerikus formában jellemezni. Legtöbbször magának a kísérletnek eredménye numerikus formában jelentkezik. Ha pl. a vízállás méréséről van szó egy adott helyen, akkor a kísérlet eredményeként egy számértéket kapunk, amely számérték, ha a méréseket különböző időpontokban végezzük, a véletlentől függő ingadozást mutat, hiszen a vízállás alakulásában számba nem vehető számos tényező játszik szerepet. A Duna vízszintje adott helyen különböző időpontokban mérve, az esős napok száma májusban (különböző években figyelve) mind olyan számadat formájában jelentkezik, amelynek értéke véletlenszerű ingadozást mutat. Hasonlóan numerikusan fejezzük ki egy betonkocka törési szilárdságát, aszfaltkeverék bitumentartalmát stb. Ezek ugyancsak véletlen ingadozást mutatnak kisebb-nagyobb mértékben.

Az olyan mennyiségeket, amelyek értéke a véletlentől függ, véletlen mennyiségeknek vagy valószínűségi változónak nevezzük. Azokat a számértékeket, amelyek a kísérlet eredményeként felléphetnek (a kísérlet összes lehetséges számszerű kimeneteleit) a valószínűségi változó lehetséges értékeinek vagy értékkeszletének nevezzük.

Ha az X valószínűségi változó egy folyó adott helyen mért vízállását jelöli, akkor X értéke egy ésszerűen választott intervallum tetszőleges pontja lehet, értékkeszlete tehát egy intervallumon kontinuumnyi sok pontja.

Ha az Y valószínűségi változó a májusi csapadékos napok számát jelöli, akkor Y lehetséges értékei a $0, 1, 2, \dots, 31$ 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 nem negatív egész számok lehetnek, ezeket egész értékű való-

színűségi változónak szokás nevezni. Előfordulnak olyan valószínűségi változók is, amelyek sem a diszkrét sem a folytonos eloszlásúak családjába nem sorolhatók, ezeket kevert eloszlásúaknak nevezzük.

Matematikai szempontból tekintve az X valószínűségi változó az elemi események $\mathcal{X}$ halmazán - az eseménytéren - értelmezett függvény, amelynek értéke attól függ, hogy melyik e elemei esemény következett be: $X = X(e)$

Az $\mathcal{X}$ eseménytér lényegében a kísérlet matematikai modellje. Álljon a kísérlet egy érme n -szeri feldobásból. Állapodjunk meg abban, hogy 1-est írunk, ha a dobás eredménye "írás" és 0-át írunk, ha a dobás eredménye fej. Ekkor a kísérlet lehetséges eredményei - az elemi események:

$$e_1 = 0 \ 0 \ 0 \ \dots \ 0 \ 0$$

$$e_2 = 0 \ 0 \ 0 \ \dots \ 0 \ 1$$

$$e_3 = 0 \ 0 \ 0 \ \dots \ 1 \ 0$$

.

.

.

$$e_{n+1} = 1 \ 0 \ 0 \ \dots \ 0 \ 0$$

.

.

.

$$e_{2^n} = 1 \ 1 \ 1 \ \dots \ 1 \ 1$$

$$\mathcal{X} = \{e_1, e_2, \dots, e_{2^n}\}$$

Jelölje az $X(e)$ valószínűségi változó az "írást" eredményező dobások számát az n dobás során. Ekkor $X(e_1) = 0$, $X(e_2) = 1$, $X(e_3) = 1$, $\dots$, $X(e_{2^n}) = n$. Ha e_i olyan elemi esemény, amely, k számú 1-est (és $(n-k)$ darab 0-t) tartalmaz, akkor $X(e_i) = k$. Az X valószínűségi változó lehetséges értékei a $0, 1, 2, \dots, n$ számok. Jelöljük A_k -val mindazon e_i elemi események halmazát, amelyek pontosan k darab 1-est tartalmaznak (ilyen nyilván $\binom{n}{k}$ darab van), akkor

$$P(X = k) = P(A_k) \quad (k = 0, 1, 2, \dots, n)$$

Tételezzük fel, hogy a dobásokat homogén, szabályos érmével végezzük, ekkor mind az "írás", mind a "fej" $\frac{1}{2}$ valószínűséggel következik be, s mivel az egyes dobások eredményei függetlenek, bármelyik e_i elemi esemény valószínűsége:

$$P(e_i) = \frac{1}{2^n} \quad (i=1, 2, \dots, 2^n)$$

Ezek alapján:

$$P(X=k) = P(A_k) = \sum_{e_i \in A_k} P(e_i) = \frac{\binom{n}{k}}{2^n}$$

(Példánkból látható, hogy az $X(e)$ valószínűségi változó mint az eseménytéren értelmezett függvény különböző értékeket rendel. A fenti A_k esemény $\binom{n}{k}$ számi e_i pontjához $X(e_i)$ ugyanazt a k értéket rendel. Annak valószínűsége, hogy az $X(e)$ valószínűségi változó egy adott k értéket vegyen fel, megegyezik mindazon e elemi események A halmazának $P(A)$ valószínűségi mértékével, amelyekre

$$X(e) = k.$$

Az n számú érmédobási kísérlethez rendelt X valószínűségi változó véges sok értéket vehet fel. X lehetséges értékei a $0, 1, 2, \dots, n$ számok. Az ilyen valószínűségi változónál ha ismerjük, hogy az X valószínűségi változó a lehetséges értékeit milyen valószínűséggel veszi fel, akkor azt mondjuk, hogy ismerjük a valószínűségi változó eloszlását. Példánkban

$$P(X=k) = \binom{n}{k} \frac{1}{2^n} \quad (k=0, 1, \dots, n)$$

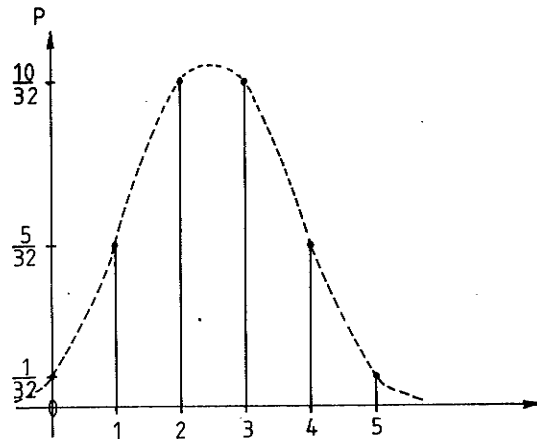
Ha pl. a kísérlet egy szabályos érme 5-szöri feldobásából áll, akkor

$$\begin{aligned} P(X=0) &= \binom{5}{0} \frac{1}{2^5} = \frac{1}{32} \\ P(X=1) &= \binom{5}{1} \frac{1}{2^5} = \frac{5}{32} \\ P(X=2) &= \binom{5}{2} \frac{1}{2^5} = \frac{10}{32} \\ P(X=3) &= \binom{5}{3} \frac{1}{2^5} = \frac{10}{32} \\ P(X=4) &= \binom{5}{4} \frac{1}{2^5} = \frac{5}{32} \\ P(X=5) &= \binom{5}{5} \frac{1}{2^5} = \frac{1}{32} \end{aligned}$$

(Nyilván:

$$\sum_{k=0}^5 P(X = k) = 1$$

Ezt a valószínűségeloszlást grafikusan is szemléltethetjük ún. valószínűség-diagram segítségével:



14. ábra

Ha kísérletünk egy játékkocka egymás utáni dobásából áll és mindaddig folytatjuk a dobássorozatot, amíg a dobás 6-ost nem eredményez, akkor 1-gyel jelölve azt az eseményt, hogy 6-ost dobtunk, 0-val pedig azt az eseményt, hogy a dobás eredménye nem 6-os, a kísérlet elemi eseményei a következők:

$$e_1 = 1$$

$$e_2 = 0 \ 1$$

$$e_3 = 0 \ 0 \ 1$$

.

.

$$e_k = \underbrace{0 \ 0 \ 0 \ \dots \ 0}_{k-1} \ 1$$

.

.

$$e_{\infty} = 0 \ 0 \ 0 \ \dots \ 0 \ 0 \ \dots$$

Jelölje az Y valószínűségi változó az első 6-os eléréséhez szükséges dobások számát, ekkor Y lehetséges értékei az $1, 2, \dots$ számok. Az Y valószínűségi változónak megszámlálható sok értéke van. Ha ennél a kísérletnél $P(1) = p$, $P(0) =$

$$= 1 - p = q. \text{ (szabályos kockánál } p = \frac{1}{6}, \quad q = \frac{5}{6} \text{ ,)}$$

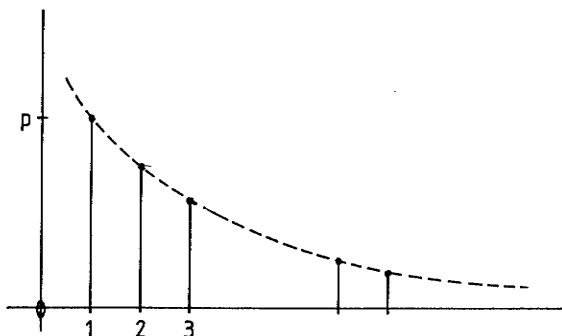
akkor annak a valószínűsége, hogy az Y valószínűségi változó a k értéket veszi fel:

$$P(Y=k) = q^{k-1} \cdot p \quad (k=1, 2, \dots)$$

Természetesen

$$P(Y=k) = \sum_{k=1}^{\infty} q^{k-1} \cdot p = p(1 + q + q^2 + \dots) = \frac{p}{1 - q} = 1$$

Az Y valószínűségi változó eloszlását a következő valószínűségi diagram szemlélteti:



15. ábra

Definíció:

Az olyan X valószínűségi változót, amelynek véges vagy megszámlálható sok lehetséges értéke van, diszkrét eloszlású valószínűségi változónak (vagy röviden diszkrét valószínűségi változónak) nevezzük. A diszkrét valószínűségi változó eloszlását úgy adjuk meg, hogy megadjuk minden lehetséges értékének a valószínűségét.

Az eddig példánkban szereplő diszkrét valószínűségi változók mindegyikének a lehetséges értékei nem negatív egész számok voltak. Az ilyen valószínűségi változókat egész értékű valószínűségi változóknak nevezzük. A gyakorlatban, különösen a megfigyelési eredmények statisztikai feldolgozása során vizsgálunk olyan diszkrét eloszlású valószínűségi változókat is, amelyek nem feltétlenül egész-értékűek.

Ha X diszkrét valószínűségi változó, amelynek lehetséges értékei az $x_1, x_2, \dots, x_n, \dots$ számok (véges vagy megszámlálható sok) s ezeket az értékeket rendre adott

$$P(X=x_1) = p_1, P(X=x_2) = p_2, \dots, P(X=x_k) = p_k, \dots$$

valószínűségekkel veszi fel, $(\sum_k p_k = 1)$ és annak való-

színűsége érdekel bennünket, hogy X értéke adott a és b határok közé esik, akkor összegezni kell mindazon x_i értékek valószínűségeit, amelyekre $a \leq x_i \leq b$ teljesül:

$$P(a \leq X < b) = \sum_{x_i \in [a, b)} p_k$$

A műszaki gyakorlatban a vizsgálat igen sokszor olyan valószínűségi változóra vonatkozik, amelynek értéke a számegyenes valamely intervallumában bárhová eshet, azaz kontinuum sok lehetséges értéke van. Mint említettük, az ilyen változókat folytonos valószínűségi változóknak nevezzük. Ha pl. az X valószínűségi változó a Tisza folyó vízállása Szolnokonál adott időpontban, akkor rendszerint annak valószínűsége érdekel, hogy a vízállás 600 cm és 700 cm közé esik, vagy hogy a vízállás kisebb mint 850 cm stb. A példánkban szereplő X valószínűségi változó eloszlását ismertnek tekintjük, ha tetszőleges (a, b) intervallumra meg tudjuk mondani az $\{a \leq X < b\}$ esemény

$$P(a \leq X < b)$$

valószínűségét.

Definíció:

Az X valószínűségi változó eloszlását folytonosnak nevezzük, ha megadható hozzá a számegyenesen értelmezett olyan nemnegatív, integrálható $f(x)$ függvény, amelyre:

$$P(a \leq X < b) = \int_a^b f(x) dx \quad \text{és} \\ \int_{-\infty}^{\infty} f(x) dx = 1$$

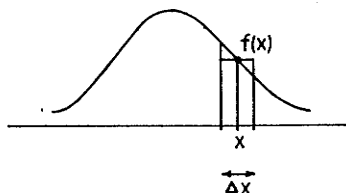
Az ilyen $f(x)$ függvényt az X valószínűségi változó sűrűségfüggvényének nevezzük.

A folytonos valószínűségi változó fenti definíciójából következik, hogy

$$P(X = x) = 0$$

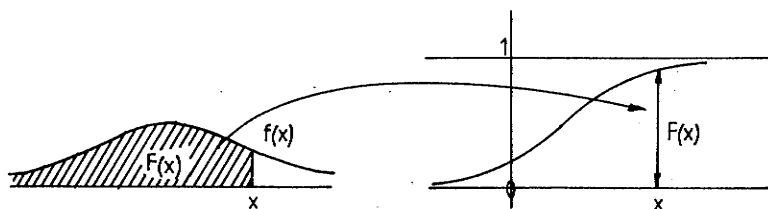
továbbá annak valószínűsége, hogy X értéke az x szám kicsiny Δx hosszúságú környezetébe esik:

$$P\left(x - \frac{\Delta x}{2} \leq X < x + \frac{\Delta x}{2}\right) \approx f(x) \Delta x$$



16. ábra

A $P(a \leq X < b)$ valószínűség kiszámításához elegendő ha tudjuk, hogy a számegyenes tetszőleges



17. ábra

x pontjára vonatkozólag mi a valószínűsége, hogy X értéke kisebb, mint x . A $P(X < x)$ valószínűség az x változó függvénye, jelöljük ezt a függvényt $F(x)$ -szel. Nyilván

$$F(x) = \int_{-\infty}^x f(t) dt.$$

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 X 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, hiszen $F(x)$ valószínűséget fejez ki, az $\{X < x\}$ esemény valószínűségét:

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

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 az $\{X < b\}$ esemény mindig bekövetkezik, ha a $\{X < a\}$ esemény bekövetkezik, de az $\{X < b\}$ esemény még akkor is bekövetkezik, ha $a \leq X < b$, tehát:

$$\{X < b\} = \{X < a\} \cup \{a \leq X < b\}$$

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

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

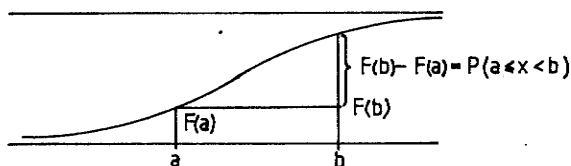
azaz

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

Az utóbbi összefüggésből:

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

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



18. ábra

Mutatunk egy példát, amely jól érzékelteti, mennyire természetes tulajdonsága az $F(x)$ eloszlásfüggvénynek a monoton nem csökkenő volta. Tegyük fel, hogy kísérletünk egy folyó vízállásának megméréseiből áll adott helyen. A vízállás nagysága legyen az X valószínűségi változó. Legyen $a = 5$ m, $b = 7$ m. Az $\{X < 5\}$ esemény akkor következik be, ha a vízállás alacsonyabb, mint 5 m. Az $\{X < 5\}$ esemény maga után vonja az $\{X < 7\}$ 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 $\{5 \leq X < 7\}$ esemény gyakorisága az $\{X < 7\}$ és $\{X < 5\}$ események gyakoriságainak különbségével egyenlő. Természetesen ugyanez igaz az illető események relatív gyakoriságára is. Pl. a táblázat adatai szerint az $\{X < 5\}$

2.1 Táblázat

Duna, Pozsony évi jégmentes nagyvízi vízállása
1892 - 1961

	Vízállás (cm)	k_i gyakoriság
Δ_1	450 - 499	1
Δ_2	500 - 549	1
Δ_3	550 - 599	7
Δ_4	600 - 649	11
Δ_5	650 - 699	15
Δ_6	700 - 749	12
Δ_7	750 - 799	15
Δ_8	800 - 849	2
Δ_9	850 - 899	3
Δ_{10}	900 - 949	1
Δ_{11}	950 - 999	2
	Összesen:	$\sum_{i=1}^n k_i = 70$

esemény relatív gyakorisága $\frac{1}{70}$ az $\{5 \leq X < 7\}$ esemény relatív gyakorisága $\frac{34}{70}$, az $\{X < 7\}$ esemény relatív gyakorisága $\frac{35}{70}$ és mivel a relatív gyakoriságokra érvényes összefüggések a valószínűségekre is teljesülnek, fennáll, hogy

$$P(X < 7) = P(X < 5) + P(5 \leq X < 7)$$

3. Az előforduló valószínűségi változók csak véges értéket vesznek fel; az $\{X < +\infty\}$ eseményt biztos eseménynek, az $\{X < -\infty\}$ eseményt lehetetlen eseménynek tekintjük, tehát:

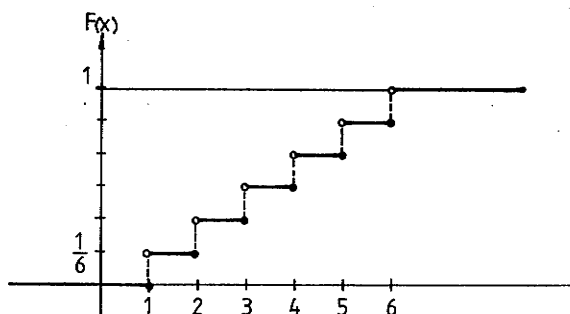
$$F(+\infty) = \lim_{x \rightarrow \infty} F(x) = 1$$

$$F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$$

A gyakorlatban előforduló valószínűségi változók legtöbbször nemcsak végesek, hanem korlátosak is; ez azt jelenti, hogy léteznek olyan k és K számok: $-\infty < k < K < +\infty$, amelyekre $F(k) = 0$, $F(K) = 1$. Ha pl. X a kockával dobott pontszámot jelölő valószínűségi változó, akkor

$$F(1) = 0, \quad F(7) = 1$$

A kockadobás pontszámát jelölő X valószínűségi változó eloszlásfüggvényét könnyen megkonstruálhatjuk. Ha az 1, 2, 3, 4, 5, 6 értékek mindegyikét $\frac{1}{6}$ valószínűséggel veszi fel, akkor az $\{X \leq x\}$ 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



19. ábra

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

$$F(a + 0) - F(a) = P(X = a)$$

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

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

eredményre jutunk.

4. Az utóbbi összefüggés is mutatja, hogy az eloszlásfüggvény balról folytonos x minden értékére. Az eloszlásfüggvény jobbról nem folytonos minden olyan $x = a$ pontban, amelyre $P(X = a) \neq 0$.

Ilyen pontokban az eloszlásfüggvénynek $P(X=a)$ nagyságú ugrása van. A kockadobás esetében az $a = 1, 2, 3, 4, 5, 6$, pontokban

$P(X=a) = \frac{1}{6}$, ezekben a pontokban, mint a 19. ábrán látható, az eloszlásfüggvénynek $\frac{1}{6}$ nagyságú ugrása van.

Hasonlóan lépcsős (szakadásos) függvény az eloszlásfüggvény, ha X egy adott hónap vagy év csapadékos napjainak számát jelenti, amelynek értéke $k = 0, 1, 2, \dots$ lehet. Ebben az esetben az ugrások nagysága természetesen általában nem egyforma. Az $F(x)$ eloszlásfüggvény ugrása az $x = k$ pontokban az $\{X = k\}$ esemény valószínűségével egyenlő.

A valószínűségeloszlást szemléletessé tehetjük az alábbi analógiával. Képzeljük el az x tengelyt végtelenül vékony drótnak, amelyen egységnyi 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 horodói lehetnek. Ha $F(x)$ az x ponttól balra jutó tömeg mennyiségét, akkor $F(x)$ az eloszlásfüggvény tulajdonságai-
val rendelkezik. Az (a, b) intervallumba jutó tömeg mennyisége $F(b) - F(a)$, ami annak valószínűségével ekvivalens, hogy az X valószínűségi változó értéke az (a, b) intervallumba esik, csak tömeg helyett valószínűséget kell gondolnunk.

Diszkrét eloszlás esetén csupán azokban az x_i pontokban helyezünk el tömeget, amelyekre $P(X=x_i)=p_i > 0$'s a tömeg nagysága éppen p_i . Ilyen értelemben beszélhetünk valószínűségtömegekről. Látni fogjuk, hogy ez a mechanikai szóhasználat a valószínűségelmélet más fogalmaira is kiterjeszthető.

2.1.2 A többdimenziós eloszlás fogalma

Sok esetben 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 X -szel, a talajvízszintet Y -nal jelölve vizsgáljuk az (X, Y) valószínűségi változópár vagy másként nevezve a $\vec{Z} = (X, Y)$ 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 az

$$\vec{X} = (X_1, X_2, X_3)$$

véletlen vektor eloszlásáról, amelyet az X_1, X_2 és X_3 valószínűségi változók együttes eloszlásának is szokás nevezni.

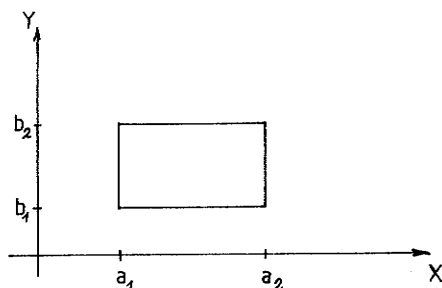
Ha egy X mennyiségre vonatkozólag n számú mérést végzünk, amelynek eredményei az $X_1, X_2, \dots, X_n$ mennyiségek, akkor vizsgálhatjuk ezen mérési eredmények együttes eloszlását, azaz az

$$\vec{X} = (X_1, X_2, \dots, X_n)$$

véletlen vektor vagy valószínűségi vektorváltozó eloszlását. Még gyakoribb, hogy az $(X_1, X_2, \dots, X_n)$ véletlen vektor helyett egy $Y = f(X_1, X_2, \dots, X_n)$ valószínűségi változó eloszlása érdekel bennünket, ahol f egy n -változós függvény. A matematikai statisztika legfontosabb területein: a becsléelméletben és a statisztikai hipotézisek vizsgálatánál ilyen típusú, véletlen argumentumú függvények eloszlása elsődleges fontosságú.

Foglalkozzunk először a kétdimenziós véletlen vektorok eloszlásával. Legyenek X és Y valószínűségi változók és tekintsük az (X, Y) véletlen pont síkbeli eloszlását.

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



20. ábra

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

Az X és Y változók együttes eloszlásfüggvényét a

$$H(x, y) = P(X < x, Y < y)$$

reláció definiálja.

Könnyű belátni, hogy

$$\begin{aligned} (2.1) \quad & P(a_1 \leq X < a_2, b_1 \leq Y < b_2) = \\ & = H(a_2, b_2) - H(a_1, b_2) - H(a_2, b_1) + H(a_1, b_1) \geq 0 \end{aligned}$$

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

$$(2.2) \quad H(+\infty, +\infty) = 1, \quad H(x, -\infty) = H(-\infty, y) = 0$$

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

Ha ismerjük az (X,Y) pár (x,y) együttes eloszlásfüggvényét, ebből könnyen megkaphatjuk az X , ill. Y valószínűségi változók $F(x)$, ill. $G(y)$ eloszlásfüggvényeit, az ún. vetület vagy peremeloszlásokat:

$$(2.3) \quad H(x, +\infty) = P(X < x, Y < +\infty) = P(X < x) = F(x)$$

$$H(+\infty, y) = P(X < +\infty, Y < y) = P(Y < y) = G(y)$$

Ha a (2.1) formulában $a_1 = x_1$, $b_1 = y$, $a_2 = +\infty$, $b_2 = +\infty$ értékekekt helyettesítünk, akkor a (2.2) és (2.3) formulák alapján a

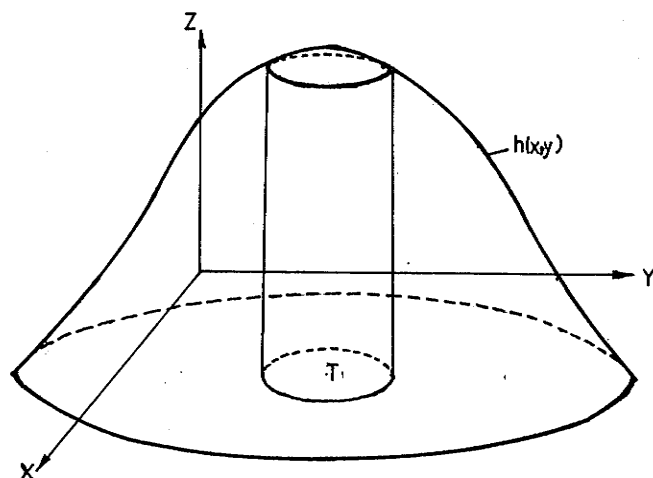
$$(2.4) \quad P(X > x, Y > y) = 1 - F(x) - G(y) + H(x,y)$$

összefüggés adódik.

Számukra az alkalmazások szempontjából legfontosabb az az eset, amikor mind X mind Y folytonos eloszlásúak $f(x)$, ill. $g(y)$ sűrűségfüggvénnyel és a $Z = (X,Y)$ véletlen vektor folytonos síkbeli eloszlású.

Ez esetben, a folytonos tömegeloszlás analógiájára az (x,y) pontbeli valószínűség sűrűségét egy kétváltozós $h(x,y)$ függvény segítségével adjuk meg, amelynek geometriai képe egy kétváltozós felület. Ha létezik olyan kétváltozós $h(x,y)$ függvény, amelyre $h(x,y) \geq 0$ továbbá $\iint_{-\infty}^{+\infty} h(x,y) dx dy = 1$, akkor $h(x,y)$ kétváltozós sűrűségfüggvény és annak valószínűségét, hogy az (X,y) véletlen számpár a sík egy adott T tartományába esik, a $h(x,y)$ függvénynek a T tartományra számított kettős integrálja szolgáltatja:

$$P((X,Y) \in T) = \iint_T h(x,y) dx dy$$



21. ábra

Ha a tartomány az $(X < x, Y < y)$ síknegyed, akkor

$$(2.5) \quad P(X < x, Y < y) = \int_{-\infty}^x \int_{-\infty}^y h(u, v) du dv = H(x, y)$$

vagyis a kétváltozós eloszlásfüggvényhez jutunk.

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

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

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

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

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

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

Ezekből az összefüggésekből differenciálás után nyerjük az X , ill. Y változók sűrűségfüggvényeit:

$$(2.6) \quad f(x) = F'(x) = \int_{-\infty}^{\infty} h(x, v) dv$$

$$(2.6') \quad g(y) = G'(y) = \int_{-\infty}^{\infty} h(u, y) du$$

2.1.3 Feltételes eloszlás- és sűrűségfüggvény

Az eseményekre vonatkozólag értelmeztük a feltételes valószínűséget; ennek alkalmazásával meghatározhatjuk az egyik változó feltételes eloszlásfüggvényét a mellett a feltétel mellett, hogy a másik változó bizonyos értékeket vesz fel:

$$P(X < x \mid Y < y) = \frac{P(X < x, Y < y)}{P(Y < y)} = \frac{H(x, y)}{G(y)}$$

$$P(Y < y \mid x_1 \leq X < x_2) = \frac{P(x_1 \leq X < x_2, Y < y)}{P(x_1 \leq X < x_2)} = \frac{H(x_2, y) - H(x_1, y)}{F(x_2) - F(x_1)}$$

$$P(Y < y \mid X \geq x) = \frac{P(X \geq x, Y < y)}{P(X \geq x)} = \frac{G(y) - H(x, y)}{1 - F(x)}$$

Ez utóbbi formulánál közvetlenül belátható, hogy az $\{X \geq x, Y < y\}$ esemény az $\{Y < y\}$ és $\{X < x, Y < y\}$ események különbsége, a második eredmény viszont része az elsőnek és így a valószínűségük különbsége képezhető.

A nevezőről mindig fel kell tenni, hogy a zérustól különbözik.

Az $\{X = x\}$ eseménynek folytonos változó esetén ugyan 0 a valószínűsége, azonban szüksége mutatkozik annak, hogy e feltétel mellett is meghatározzuk Y eloszlását. Bizonyos megszorítások mellett ez sikerül is, és pedig a középső formulából. Írjunk abban $x_1 = x$ és $x_2 = x + \Delta x$ értékeket, majd osszunk mind a számlálóban, mind pedig a nevezőben Δx -szel:

$$P(Y < y \mid x \leq X < x + \Delta x) = \frac{H(x + \Delta x, y) - H(x, y)}{\frac{F(x + \Delta x) - F(x)}{\Delta x}} = \frac{\Delta H(x, y)}{\Delta F(x)}$$

Ha $\Delta x \rightarrow 0$ esetén a nevezőben álló kifejezés az $f(x) \neq 0$ határértékhez tart, a keresett feltételes eloszlást kapjuk:

$$(2.7) \quad G(y \mid x) = P(Y < y \mid X = x) = \frac{\partial H(x, y)}{\partial x} = \frac{\int_{-\infty}^y h(x, v) dv}{\int_{-\infty}^{\infty} h(x, v) dv}$$

Ez utóbbi előállítás mutatja, hogy valóban eloszlásfüggvényhez jutottunk: $G(y \mid x)$ monoton nem csökkenő függvény, amelynek értéke a $-\infty$ -ben 0, a $+\infty$ -ben 1. A feltételes sűrűség-

függvényt differenciálással kapjuk:

$$(2.8) \quad g(y | x) = \frac{\partial G(y | x)}{\partial y} = \frac{h(x, y)}{f(x)}$$

Hasonlóan X eloszlás- és sűrűségfüggvénye az $Y = y$ feltétel mellett

$$(2.9) \quad F(x | y) = \frac{\int_{-\infty}^x h(u, y) du}{\int_{-\infty}^{\infty} h(u, y) du}$$

és

$$(2.10) \quad F(x | y) = \frac{h(x, y)}{\int_{-\infty}^{\infty} h(u, y) du}$$

a nevező mindkét esetben az Y változó sűrűségfüggvénye, azaz $g(y)$.

Az eddigiekben azzal az esettel foglalkoztunk, amikor mindkét változó folytonos. Egyszerűbb a helyzet, ha mindkét változó diszkrét. Erre nézve nincs egyöntetűen kialakult jelölésmód, hanem a feladattól függően választhatjuk azt meg.

Legyen mind X mind Y diszkrét eloszlású és lehetséges értékeik legyenek a nemnegatív egész számok. Szokásos az értékpárok valószínűségeire a következő jelölés:

$$P(X = k, Y = l) = r_{kl} \quad k, l = 0, 1, 2, \dots$$

$$\sum_k \sum_l r_{kl} = 1$$

Ezekből a mennyiségekből a gyakorlatban felmerülő események valószínűsége meghatározhatók. A peremeloszlásokra és az eloszlásfüggvényre a következőket nyerjük:

$$(2.11) \quad P(X = k) = p_k = \sum_l r_{kl} \quad (\sum_k p_k = 1)$$

$$(2.12) \quad P(Y = l) = q_l = \sum_k r_{kl} \quad (\sum_l q_l = 1)$$

$$(2.13) \quad H(k, l) = P(X < k, Y < l) = \sum_{i=0}^{k-1} \sum_{j=0}^{l-1} r_{ij} \quad (k, l = 1, 2, \dots)$$

Néhány feltételes valószínűség:

$$P(Y = l | X = k) = \frac{r_{kl}}{p_k} \quad ; \quad P(X = k | Y = l) = \frac{r_{kl}}{q_l}$$

$$P(X < k \mid l_1 \leq Y < l_2) = \frac{\sum_{i=0}^{k-1} \sum_{j=l_1}^{l_2-1} r_{ij}}{\sum_{j=l_1}^{l_2-1} q_j}$$

A matematikai statisztikában szükségünk lesz kettőnél több dimenziós valószínűségi vektorváltozók vizsgálatára. Ha $X_1, X_2, \dots, X_n$ valamely véletlen mennyiségre vonatkozó n számú mérés vagy megfigyelés eredménye, akkor ezeket a változókat egy

$$\vec{X} = (X_1, X_2, \dots, X_n)$$

n -dimenziós véletlen vektor komponenseinek tekintjük. $\vec{X}$ most az n -dimenziós tér véletlen helyzetű pontja. Az $\vec{X}$ vektor eloszlásán az $X_1, X_2, \dots, X_n$ változók együttes eloszlását értjük. Vizsgáljuk azt az esetet, amikor a szóban forgó valószínűségi változók folytonos eloszlásúak. Az $X_1, X_2, \dots, X_n$ valószínűségi változók együttes eloszlását a

$$H(x_1, x_2, \dots, x_n) = P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n)$$

n -dimenziós eloszlásfüggvény, vagy - ha létezik - a

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

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

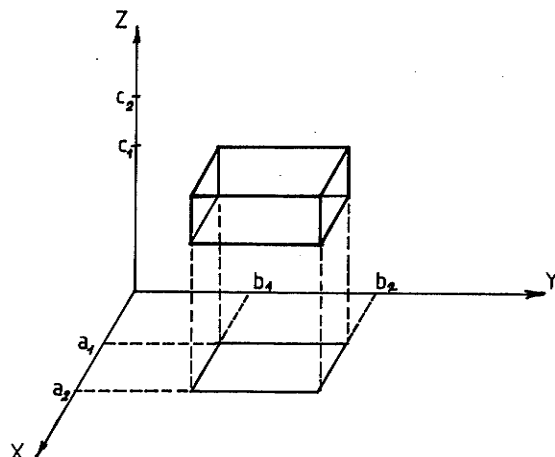
Az n -dimenziós $H(x_1, x_2, \dots, x_n)$ eloszlásfüggvény mind az n változójának monoton nem csökkenő függvénye és minden változójában balról folytonos, $H(x_1, x_2, \dots, x_n) = 0$ ha bármely változója helyébe $-\infty$ -t írunk (ha ugyanis x_i helyébe $-\infty$ -t írunk, akkor az $\{X_i < -\infty\}$ esemény is be kellene, hogy következzen, ami lehetetlen esemény), továbbá $H(+\infty, +\infty, \dots, +\infty) = 1$ (H értéke tehát csak akkor, ha mindegyik változója helyébe $+\infty$ -t írunk).

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

$$P(a_1 \leq X_1 < b_1, a_2 \leq X_2 < b_2, \dots, a_n \leq X_n < b_n)$$

valószínűség kifejezhető a $H(x_1, x_2, \dots, x_n)$ eloszlásfüggvény segítségével. Ez azt a kézenfekvő kikötést is jelenti a H függvénnyel szemben, hogy a fenti valószínűségek H értékei-

vel való kifejezése nemnegatív kell legyen. Pl. $n = 3$ esetén a H függvényre teljesülnie kell a következő feltételnek:



22. ábra

$$H(a_2, b_2, c_2) - H(a_1, b_2, c_2) - H(a_2, b_1, c_2) - H(a_2, b_2, c_1) + \\ + H(a_1, b_1, c_2) + H(a_1, b_2, c_1) + H(a_2, b_1, c_1) - H(a_1, b_1, c_1) \geq 0$$

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

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

akkor $\vec{X}$ folytonos eloszlású és a $H(x_1, x_2, \dots, x_n)$ és $h(x_1, x_2, \dots, x_n)$ függvényekre teljesül a (2.2) és (2.3) reláció, valamint a (2.5) formulával ekvivalens:

$$(2.14) \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} h(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n = \\ = H(x_1, x_2, \dots, x_n)$$

és

$$(2.15) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = 1$$

A (2.7) ill. (2.8) formulák könnyen általánosíthatók több-, dimenziós eloszlások esetére, csak X ill. Y helyébe véletlen vektorokat kell helyettesíteni és természetesen az $f(x)$, $g(y)$ $f(x, y)$, $g(y, x)$ függvények mindegyike $\vec{X}$ ill. $\vec{Y}$ dimenziójának megfelelő számú argumentummal bíró többváltozós függvény.

Ha pl. az $X_1, X_2, \dots, X_n$ változók együttes eloszlása folytonos eloszlás, akkor az $X_1, X_2, \dots, X_k$ változók együttes feltételes eloszlásfüggvénye azon feltétel mellett, hogy $X_{k+1} = x_{k+1}, \dots, X_n = x_n$

$$H(x_1, x_2, \dots, x_k \mid x_{k+1}, \dots, x_n)$$

a következő határértékkel értelmezhető:

$$\begin{aligned} H(x_1, x_2, \dots, x_k \mid x_{k+1}, \dots, x_n) &= \\ &= \lim_{\Delta x_{k+1} \rightarrow 0} \dots \lim_{\Delta x_n \rightarrow 0} P(X_1 < x_1, \dots, X_k < x_k \mid \\ &\quad x_{k+1} \leq X_{k+1} < x_{k+1} + \Delta x_{k+1}, \dots, x_n \leq X_n < x_n + \Delta x_n) \end{aligned}$$

A megfelelő feltételes sűrűségfüggvény:

$$\begin{aligned} \frac{\partial H(x_1, \dots, x_k \mid x_{k+1}, \dots, x_n)}{\partial x_1, \dots, \partial x_k} &= h(x_1, \dots, x_n \mid x_{k+1}, \dots, x_n) = \\ &= \frac{h(x_1, \dots, x_n)}{g(x_{k+1}, \dots, x_n)} \end{aligned}$$

ahol az $(n-k)$ változó g függvény az $X_{k+1}, X_{k+2}, \dots, X_n$ változók együttes sűrűségfüggvényét jelöli.

A többdimenziós feltételes sűrűségek fontos szerepet játszanak a regresszióanalízisben.

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

Definíció.

Az $X_1, X_2, \dots, X_n$ valószínűségi változókat teljesen függetleneknek nevezzük, ha tetszőleges $a_i < b_i$ ($i = 1, 2, \dots, n$) számok esetén

$$\begin{aligned} P(a_1 \leq X_1 < b_1, a_2 \leq X_2 < b_2, \dots, a_n \leq X_n < b_n) &= \\ &= P(a_1 \leq X_1 < b_1) P(a_2 \leq X_2 < b_2) \dots P(a_n \leq X_n < b_n) \end{aligned}$$

Speciálisan $a_i = -\infty$, $b_i = x_i$ választással ($i=1,2,\dots,n$)

$$\begin{aligned} P(X_1 < x_1, X_2 < x_2, \dots, X_n < x_n) &= \\ &= P(X_1 < x_1) P(X_2 < x_2) \dots P(X_n < x_n) \end{aligned}$$

azaz:

$$(2.16) H(x_1, x_2, \dots, x_n) = F(x_1)F_2(x_2) \dots F_n(x_n)$$

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, s akkor a függetlenségből következik, hogy együttes sűrűségfüggvényük is az egyes változók sűrűségfüggvényeinek szorzata:

$$(2.17) h(x_1, x_2, \dots, x_n) = f_1(x_1) f_2(x_2) \dots f_n(x_n)$$

A függetlenség fogalma számunkra azért is különös fontosságú, mert a matematikai jól kidolgozott módszerek túlnyomó részben független valószínűségi változókra vonatkoznak. Statisztikai módszerek kidolgozása nem független valószínűségi változók esetére egyike a matematikai statisztika legsürgősebb feladatainak, amely irányban jelenleg is intenzív kutatások folynak.

Annak tapasztalati úton való eldöntése, hogy adott $X_1, X_2, \dots, X_n$ valószínűségi változók valóban függetlenek-e, a matematikai statisztika feladata. Erre vonatkozó módszereket a 6. fejezetben ismertetünk.

2.1.4 Valószínűségi változó monoton függvényének eloszlása

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

$$G(y) = P(Y < y) = P[\varphi(X) < y] = P[X < \varphi^{-1}(y)] = F[\varphi^{-1}(y)]$$

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

$$(2.18) \quad g(y) = \frac{d G(y)}{dy} = \frac{d [F[\varphi^{-1}(y)]]}{\frac{\partial}{\partial y}} = \\ = f[\varphi^{-1}(y)] \frac{d \varphi^{-1}(y)}{dy}$$

Ha pl. $Y = aX + b$, azaz $\varphi(x) = ax + b$ és X sűrűségfüggvénye $f(x)$, akkor

$$(2.19) \quad x = \varphi^{-1}(y) = \frac{y - b}{a} \\ g(y) = \frac{1}{|a|} f\left(\frac{y - b}{a}\right)$$

Ha $Y = e^X$

azaz $y = \varphi(x) = e^x$; $x = \varphi^{-1}(y) = \ln y$

$$(2.20) \quad g(y) = \frac{1}{|y|} f(\ln y)$$

2.1.5 Két valószínűségi változó összegének eloszlása

Előbb a folytonos eloszlású változók esetét tárgyaljuk. Legyen az X és Y folytonos valószínűségi változók együttes sűrűségfüggvénye $h(x,y)$. A $Z = X + Y$ eloszlásfüggvényének meghatározásához az

$$\{Z < z\} = \{X + Y < z\}$$

esemény valószínűségét kell meghatározni minden z -re. Ez azt jelenti, hogy a $h(x,y)$ sűrűségfüggvényt integrálni kell a síknak ama T_z^+ tartományára, amelynek pontjaira az $x + y < z$ feltétel teljesül. Tehát ha z eloszlásfüggvényét $K(z)$ -vel jelöljük, akkor

$$K(z) = P(Z < z) = \iint_{T_z^+} h(x,y) \, dx \, dy$$

Az integrált úgy is kiszámíthatjuk, hogy rögzített x -re először y szerint integrálunk a $(-\infty, z-x)$ intervallumban, majd x szerint a $(-\infty, +\infty)$ intervallumon, vagyis a kétszeres integrált kétszeres integrálra vezetjük vissza:

$$(2.21) \quad K(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-x} h(x,y) \, dy \right) dx$$

Ha X és Y független valószínűségi változók, vagyis

$$h(x, y) = f(x) \cdot g(y) \quad \text{akkor}$$

$$(2.22) \quad K(z) = \int_{-\infty}^{\infty} f(x) \left(\int_{-\infty}^{z-x} g(y) dy \right) dx$$

Itt azonban a belső integrál az Y változó eloszlásfüggvényének értéke a $z-x$ helyen, vagyis

$$(2.23) \quad K(z) = \int_{-\infty}^{\infty} G(z-x) f(x) dx$$

Ebből a sűrűségfüggvényt z szerint differenciálással kapjuk:

$$(2.24) \quad K'(z) = k(z) = \int_{-\infty}^{\infty} g(z-x) f(x) dx$$

Azonos eloszlású, független változók esetén, vagyis, ha $f(x) = g(x)$, akkor $X + Y$ sűrűségfüggvénye

$$(2.25) \quad k(z) = \int_{-\infty}^{\infty} f(z-x) f(x) dx$$

Az összeg eloszlását illetően a diszkrét esetben csak független változókra jegyezzük meg a következőket:

Legyenek X lehetséges értékei $x_1, x_2, \dots$, az X -től független Y lehetséges értékei $y_1, y_2, \dots$, és jelöljük az értékek valószínűségét a következő módon:

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

$$P(Y = y_j) = q_j \quad (j=1, 2, \dots)$$

Ekkor a $Z = X + Y$ változóra érvényes

$$(2.26) \quad P(Z = z) = \sum_{x_i + y_j = z} p_i q_j$$

vagyis az összegezést mindazon (i, j) indexpárookra kell kiterjeszteni, amelyekre $x_i + y_j = z$.

Ha z nem állítható elő ilyen összeg alakjában, akkor a valószínűsége 0.

Legyenek mind X mind Y lehetséges értékei a nemnegatív egész számok. Akkor a $Z = X + Y$ változó értékészlete szintén a nemnegatív egész számokra terjed, és a valószínűségek a következők:

$$(2.27) \quad P(Z = k) = \sum_{i=0}^k p_i q_{k-i}$$

ahol most

$$p_i = P(X = i), \quad q_j = P(Y = j)$$

2.1.6 Két független valószínűségi változó szorzatának és hányadosának eloszlása

Legyen X sűrűségfüggvénye $f(x)$ és az X -től független Y sűrűségfüggvénye $g(y)$. Jelölje T_z^x az (x, y) sík azon tartományát, amelynek pontjaira $xy < z$ míg T_z^y azt a tartományt, amelynek pontjaira $\frac{x}{y} < z$; ekkor nyilvánvalóak a következő összefüggések:

$$R(z) = P(XY < z) = \iint_{T_z^x} f(x) g(y) dx dy$$

$$S(z) = P\left(\frac{X}{Y} < z\right) = \iint_{T_z^y} f(x) g(y) dx dy$$

Ha Y csak nemnegatív lehet: $Y \geq 0$, vagyis $g(y) = 0$, ha $y \leq 0$, akkor az összeg eloszlásának meghatározására alkalmazott megfontolással az eloszlás-, ill. sűrűségfüggvényekre a következő kifejezéseket kapjuk:

$$(2.28) R(z) = \int_0^{\infty} g(y) \left[\int_{-\infty}^{z/y} f(x) dx \right] dy = \int_0^{\infty} F\left(\frac{z}{y}\right) g(y) dy$$

$$(2.29) S(z) = \int_0^{\infty} g(y) \left[\int_{-\infty}^{zy} f(x) dx \right] dy = \int_0^{\infty} F(zy) g(y) dy$$

$$(2.30) r(z) = R'(z) = \int_0^{\infty} f\left(\frac{z}{y}\right) g(y) \frac{1}{y} dy$$

$$(2.31) s(z) = S'(z) = \int_0^{\infty} y f(zy) g(y) dy$$

Ha Y negatív értéket is felvehet, akkor a következőképpen alakulnak a sűrűségfüggvények:

$$(2.32) r(z) = \int_{-\infty}^{\infty} f\left(-\frac{z}{y}\right) g(y) \frac{1}{|y|} dy$$

$$(2.33) s(z) = \int_{-\infty}^{\infty} |y| f(zy) g(y) dy$$

2.1.7 Valószínűségeloszlások jellemző adatai

Valamely valószínűségi változó ingadozásának 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 tö-

megeloszlással kapcsolatban rendszerint elősorban 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 - 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 korlátszámot, 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.

a) A várható érték fogalma és tulajdonságai

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

$$P(X = x_k) = p_k \quad (k=1, 2, \dots)$$

Definíció:

A diszkrét eloszlású X valószínűségi változó várható értékén, amelyet $E(X)$ -szel jelölünk, a következő kifejezést értjük:

$$(2.34) \quad E(X) = \sum_k x_k p_k$$

amennyiben a $\sum_k |x_k| p_k$ összeg véges érték, azaz ha az $|X|$ -nek is létezik a várható értéke.

Tekintettel arra, hogy $p_k = 1$, láthatjuk, hogy diszkrét eloszlású X valószínűségi változó várható értéke az X lehetséges értékeinek súlyozott számtani középpontja, ahol a súlyok az egyes értékekhez tartozó valószínűségek.

Definíció:

Ha X folytonos eloszlású valószínűségi változó $f(x)$ sűrűségfüggvénnyel, akkor X várható értékét a folytonos tö-

megeloszlás súlypontjának analógiájára az

$$(2.35) E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

kifejezés értelmezi, feltéve, hogy az improprius integrál abszolút konvergens,

$$\text{azaz } \int_{-\infty}^{\infty} |x| f(x) dx < +\infty$$

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

1. Tétel:

Állandó várható értéke önmaga:

$$E(c) = c$$

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

$$E(c) = c \cdot 1 = c.$$

2. Tétel:

Ha X korlátos valószínűségi változó, azaz

$$a \leq X \leq b$$

akkor létezik a várható értéke és

$$a \leq E(X) \leq b$$

Bizonyítás:

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

$$a = ap_1 + a_2p_2 + \dots \leq x_1p_1 + x_2p_2 + \dots \leq bp_1 + bp_2 + \dots$$

$$\text{azaz } a = a \sum p_i \leq \sum x_i p_i \leq b \sum p_i = b$$

Ha X folytonos eloszlású, akkor

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

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

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

3. Tétel:

Ha $Y = aX + b$, azaz az Y valószínűségi változó az X valószínűségi változónak lineáris függvénye, akkor

$$(2.36) E(Y) = E(aX + b) = a E(X) + b$$

Bizonyítás:

Ha X diszkrét eloszlású $x_1, x_2, \dots$ lehetséges értékekkel, amelyeket rendre $p_1, p_2, \dots$ valószínűséggel vesz fel, akkor az Y 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

$$E(Y) = \sum (ax_k + b)p_k = a \sum x_k p_k + b \sum p_k = a E(X) + b$$

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

Ha X folytonos eloszlású $f(x)$ sűrűségfüggvénnyel, akkor $X = x$ esetén $Y = ax + b$ és Y sűrűségfüggvénye az (2.20) formula alapján:

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

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

$$\frac{y-b}{a} = u \text{ helyettesítéssel}$$

$$E(Y) = \frac{a}{|a|} \int_{-\infty}^{\infty} (au + b) f(u) du = a \int_{-\infty}^{\infty} u f(u) du + b \int_{-\infty}^{\infty} f(u) du = a E(X) + b$$

Speciális esetenként megemlítjük, hogy ha $Y = X - E(X)$ akkor

$$E(Y) = E(X) - E E(X) = E(X) - E(X) = 0$$

Definíció:

Legyen az Y valószínűségi változó az X valószínűségi változó valamely folytonos függvénye, $Y = \psi(X)$. Ekkor az Y valószínűségi változó várható értéké diszkrét eloszlás esetén az

$$(2.37) E(Y) = \sum_k \psi(x_k) p_k$$

folytonos eloszlás esetén az

$$(2.38) E(Y) = \int_{-\infty}^{\infty} \psi(x) f(x) dx$$

formulákkal definiáljuk, feltéve, hogy az összeg ill. integrál abszolút konvergens.

Analóg módon definiáljuk valószínűségi vektorváltozók függvényének várható értékét. Pl. folytonos eloszlású valószínűségi változók esetén, ha $Y = (X_1, X_2, \dots, X_n)$ és az $X_1, X_2, \dots, X_n$ valószínűségi változók együttes sűrűségfüggvénye $h(x_1, x_2, \dots, x_n)$, akkor

$$(2.39) \quad E(Y) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \varphi(x_1, \dots, x_n) h(x_1, \dots, x_n) dx_1, \dots, dx_n =$$

amennyiben az improprius integrál abszolút konvergens.

4. Tétel:

Legyenek X és Y tetszőleges eloszlású valószínűségi változók, amelyek várható értékei léteznek, ekkor

$$(2.40) \quad E(X + Y) = E(X) + E(Y)$$

A bizonyítást először diszkrét eloszlású változók eseteire mutatjuk meg. Legyenek X lehetséges értékei az $x_1, x_2, \dots, x_n$ számok, Y lehetséges értékei az $y_1, y_2, \dots, y_n$ számok. Legyenek továbbá

$$P(X = x_i) = p_i, \quad P(Y = y_j) = q_j, \quad P(X = x_i, Y = y_j) = r_{ij} \\ (i, j = 1, 2, \dots)$$

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

$$E(X + Y) = \sum_i \sum_j (x_i + y_j) r_{ij} = \\ = \sum_i \sum_j x_i r_{ij} + \sum_i \sum_j y_j r_{ij} = \\ = \sum_i x_i \left(\sum_j r_{ij} \right) + \sum_j y_j \left(\sum_i r_{ij} \right)$$

A (2.11) és (2.12) összefüggések alapján:

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

Tehát:

$$E(X + Y) = \sum_i x_i p_i + \sum_j y_j q_j = E(X) + E(Y)$$

[Folytonos eloszlású valószínűségi változók esetén, ha X és Y együttes sűrűsége $h(x,y)$, akkor az (2.39) összefüggés alapján $\varphi(x,y) = x + y$ helyettesítéssel

$$\begin{aligned} E(X + Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) h(x,y) dx dy = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x h(x,y) dx dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y h(x,y) dx dy = \\ &= \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} h(x,y) dy \right) dx + \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} h(x,y) dx \right) dy = \\ &= \int_{-\infty}^{\infty} x f(x) dx + \int_{-\infty}^{\infty} y g(y) dy = E(X) + E(Y) \end{aligned}$$

amiből az állítás leolvasható.

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

$$(2.41) \quad E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$$

Legyen ugyanis $X = X_1 + X_2 + \dots + X_{n-1}$; $Y = X_n$

és tegyük fel, hogy a tétel $n-1$ összeadandóra érvényes, az-az igaz, hogy

$$E(X) = E(X_1 + \dots + X_{n-1}) = E(X_1) + E(X_2) + \dots + E(X_{n-1})$$

akkor a bizonyított tétel alapján következik, hogy

$$\begin{aligned} E(X_1 + \dots + X_{n-1} + X_n) &= E(X) + E(X_n) = E(X) + E(Y) = \\ &= \sum_{i=1}^{n-1} E(X_i) + E(X_n) = \sum_{i=1}^n E(X_i) \end{aligned}$$

5. Tétel:

Független valószínűségi változók szorzatának várható értéke várható értékeik szorzatával egyenlő, azaz ha X és Y független valószínűségi változók, akkor

$$(2.42) \quad E(X \cdot Y) = E(X) \cdot E(Y)$$

A bizonyítást először folytonos eloszlású valószínűségi változók esetére végezzük el. Az X és Y változók függetlensége miatt együttes sűrűségfüggvényük az egyes változók sűrűségfüggvényeinek szorzatával egyenlő (lásd:

2.1.3 pont) azaz

$$h(x,y) = f(x)g(y)$$

A (2.39) formula alapján $\varphi(x,y) = x \cdot y$ helyettesítéssel:

$$\begin{aligned} E(X \cdot Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x)g(y) dx dy = \int_{-\infty}^{\infty} xf(x) \left(\int_{-\infty}^{\infty} y g(y) dy \right) dx = \\ &= \int_{-\infty}^{\infty} xf(x) dx \cdot \int_{-\infty}^{\infty} yg(y) dy \end{aligned}$$

Ha X és Y diszkrét eloszlásúak és $P(X = x_i) = p_i$, $P(Y = y_j) = q_j$, továbbá együttes eloszlásúak:

$$P(X = x_i, Y = y_j) = r_{ij}$$

akkor az X és Y változók függetlensége folytán

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

Ekkor

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j r_{ij} = \sum_i \sum_j x_i y_j p_i q_j = \\ &= \sum_i x_i p_i \left(\sum_j y_j q_j \right) = \sum_i x_i p_i \cdot \sum_j y_j q_j = E(X) E(Y) \end{aligned}$$

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

$$(2.43) \quad E(X_1 \cdot X_2 \dots X_n) = E(X_1) E(X_2) \dots E(X_n)$$

feltéve, hogy az egyes tényezők várható értéke létezik.

b) A feltételes várható érték

Legyenek X és Y diszkrét valószínűségi változók az alábbi eloszlásokkal:

$$X: \begin{pmatrix} x_1, x_2, \dots, x_n, \dots \\ p_1, p_2, \dots, p_n, \dots \end{pmatrix} \quad Y: \begin{pmatrix} y_1, y_2, \dots, y_n, \dots \\ q_1, q_2, \dots, q_n, \dots \end{pmatrix}$$

Legyen továbbá

$$P(X = x_i, Y = y_j) = r_{ij} \quad (i, j = 1, 2, \dots)$$

Definíció:

Az Y valószínűségi változónak a feltételes várható értékén azon feltétel mellett, hogy a X változó adott x_i értéket vesz fel az:

$$(2.44) \quad E(Y | X = x_i) = \sum_j y_j P(Y = y_j | X = x_i)$$

kifejezést értjük.

Formulánk segítségével az Y változó feltételes várható értékét az X változó minden aktuális értékére - mint feltételre - kiszámíthatjuk. Az így nyert mennyiség értéke attól függ,

hogyan X melyik x_i értéket veszi fel, az pedig a véletlentől függ, $E(Y|X)$ tehát valószínűségi változó, amelynek lehetséges értékei az

$$E(Y|X = x_1), E(Y|X = x_2), \dots, E(Y|X = x_n), \dots$$

Könnyű belátni, hogy $E(Y|X)$ az $E(Y|X = x_1)$ értékét éppen p_1 valószínűséggel veszi fel.

Számítsuk ki az $E(Y|X)$ valószínűségi változó várható értékét!

$$E[E(Y|X)] = \sum_i E(Y|X = x_i) p_i$$

Felhasználva a (2.44) összefüggést, az alábbi érdekes eredményt kapjuk:

$$\begin{aligned} (2.45) \quad E[E(Y|X)] &= \sum_i \left(\sum_j y_j P(Y = y_j | X = x_i) p_i \right) = \\ &= \sum_i \sum_j y_j r_{ij} = \sum_j y_j \sum_i r_{ij} = \\ &= \sum_j y_j q_j = E(Y) \end{aligned}$$

(Itt felhasználtuk, hogy

$$P(Y = y_j | X = x_i) P(X = x_i) = P(Y = y_j, X = x_i) = r_{ij}$$

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

Ha X és Y folytonos valószínűségi változók $f(x)$, ill. $g(y)$ sűrűségfüggvényekkel, akkor az Y valószínűségi változó X -re vonatkozó feltételes várható értékét az

$$E(Y|X = x) = \int_{-\infty}^{\infty} y \frac{h(x, y)}{f(x)} dy = \int_{-\infty}^{\infty} y g(y|x) dy$$

formulával értelmezzük. Ebben az esetben is érvényes az

$$\begin{aligned} (2.46) \quad E[E(Y|X)] &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y \frac{h(x, y)}{f(x)} dy \right) f(x) dx = \\ &= \int_{-\infty}^{\infty} y \left[\int_{-\infty}^{\infty} h(x, y) dx \right] dy = \int_{-\infty}^{\infty} y g(y) dy = E(Y) \end{aligned}$$

összefüggés.

(Feltéve természetesen, hogy a fenti improprius integrál konvergens.)

A $\varphi(x) = E(Y|X = x)$ függvényt, amely az x -változó függvénye, az Y - változó X -re vonatkozó regressziós görbéjének nevezzük.

Teljesen analóg módon definiáljuk az X valószínűségi változó Y -ra vonatkozó feltételes várható értékét.

Definiálhatjuk az $E(Y|X_1, X_2, \dots, X_n)$ feltételes várható értékét, amely az $X_1, X_2, \dots, X_n$ változók függvénye, tehát n -változós függvény. Ennek a függvénynek az alábbi érdekes minimum-tulajdonsága van: ha $g(X_1, X_2, \dots, X_n)$ az $X_1, X_2, \dots, X_n$ változók tetszőleges függvénye, akkor

$$(2.47) \quad E\{[Y - g(X_1, X_2, \dots, X_n)]^2\} \geq E\{[Y - E(Y|X_1, X_2, \dots, X_n)]^2\}$$

c) A szórás fogalma és tulajdonságai

Mint említettük, a várható érték a valószínűségeloszlásnak csak egyik számszerű jellemzője, amely 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

$$(2.48) \quad D(X) = + \sqrt{E[X - E(X)]^2}$$

kifejezéssel szokás mérni. A gyökjel alatt lényegében a valószínűségi változó értékeinek a várható értéktől való négyzetes átlageltérése szerepel. A szórásról akkor beszélhetünk, ha létezik $E(X)$ várható érték, továbbá létezik az $[X - E(X)]^2$ valószínűségi változó várható értéke.

Definíció:

A valószínűségi változó szórásnégyzete:

$$(2.49) \quad D^2(X) = E[X - E(X)]^2$$

A szórásnégyzet kiszámítását megkönnyíti az alábbi összefüggés:

$$(2.50) \quad \begin{aligned} D^2(X) &= E[X^2 - 2X E(X) + E^2(X)] = \\ &= E(X^2) - 2 E(X) E(X) + E^2(X) = \\ &= E(X^2) - E^2(X) \end{aligned}$$

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

$$E(X^2) \geq E^2(X)$$

A szórásnégyzet kiszámítása diszkrét valószínűségi változó esetén:

$$D^2(X) = \sum_i [X_i - E(X)]^2 p_i = \sum_i X_i^2 p_i - E^2(X)$$

Folytonos valószínűségi változó esetén

$$D^2(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - E^2(X)$$

A szórásnégyzet fogalma segítségével az $E(X) = m$ várható értékének egy újabb igen fontos tulajdonságát fejezhetjük ki, amelyet Steiner-tételének nevezünk.

Tétel: Ha a tetszőleges valós szám, akkor

$$(2.51) E[(X - a)^2] \geq E[(X - m)^2]$$

Bizonyítás:

$$\begin{aligned} E[(X - a)^2] &= E[(X - m + m - a)^2] = \\ &= E[(X - m)^2] + 2(m - a)E(X - m) + (m - a)^2 \end{aligned}$$

Tekintettel arra, hogy $E(X - m) - m = 0$ következik, hogy

$$\begin{aligned} E[(X - a)^2] &= E[(X - m)^2] + (m - a)^2 = \\ &= D^2(X) + (m - a)^2 \geq D^2(X) \end{aligned}$$

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

A (2.51) összefüggést Steiner-formulának nevezzük.

Tételek a szórásnégyzetre vonatkozólag.

1. Tétel:

Ha az Y valószínűségi változó az X valószínűségi változónak lineáris függvénye, $Y = aX + b$, akkor

$$(2.52) D^2(Y) = a^2 D^2(X)$$

Bizonyítás:

A szórásnégyzet definíciója alapján:

$$\begin{aligned} D^2(Y) &= E\{[(aX + b) - E(aX + b)]^2\} = \\ &= E\{[aX - E(aX)]^2\} = a^2 E\{[X - E(X)]^2\} = a^2 D^2(X) \end{aligned}$$

2. Tétel:

Ha X és Y független valószínűségi változók és $Z = X + Y$, akkor

$$(2.53) D^2(Z) = D^2(X) + D^2(Y)$$

Bizonyítás: Ha X és Y függetlenek, akkor $X - E(X)$ és $Y - E(Y)$ is függetlenek. Figyelembe véve, hogy független valószínűségi változók szorzatának várható értéke a várható

értékeik szórásával egyenlő, továbbá, hogy $E[X - E(X)] = 0$, $E[Y - E(Y)] = 0$, a szórásnégyzet definíciója alapján:

$$\begin{aligned} D^2(Z) &= D^2(X + Y) = E\{[(X + Y) - E(X + Y)]^2\} = \\ &= E\{[X - E(X) + Y - E(Y)]^2\} = \\ &= E\{[X - E(X)]^2\} + 2 E[X - E(X)] \cdot E[Y - E(Y)] + \\ &\quad + E\{[Y - E(Y)]^2\} = D^2(X) + D^2(Y) \end{aligned}$$

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

Ha $X_1, X_2, \dots, X_n$ független valószínűségi változók, akkor

$$(2.54) \quad D^2(X_1 + X_2 + \dots + X_n) = D^2(X_1) + D^2(X_2) + \dots + D^2(X_n)$$

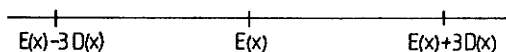
A valószínűségi változó várható értékének és szórásának ismeretében a valószínűségi változó ingadozását a várható értéke körül jól jellemezhetjük. Ezt fejezi ki a Csebisev-féle egyenlőtlenség.

3. Tétel:

Csebisev-egyenlőtlenség. Ha az X valószínűségi változó értéke és szórása létezik, akkor

$$(2.55) \quad P[|X - E(X)| > \lambda D(X)] \leq \frac{1}{\lambda^2}$$

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



23. ábra

ha X -re vonatkozólag nagyszámú megfigyelést végzünk, akkor a megfigyelt értékek kb. 90 %-a ebbe az intervallumba esik, feltéve, hogy X várható értéke és szórása létezik. Ha a

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

$$(2.56) \quad P(|X - E(X)| > \varepsilon) \leq \frac{D^2(X)}{\varepsilon^2}$$

Bizonyítás:

Egyszerűség kedvéért vezessük be az $E(X)=m$, $D(X) = \sigma^2$ jelölést. Ha X diszkrét eloszlású $x_1, x_2, \dots$ lehetséges értékekkel, amelyeket rendre $p_1, p_2, \dots$ valószínűségekkel vesz fel, akkor

$$\begin{aligned}\sigma^2 &= \sum_i (x_i - m)^2 p_i \geq \sum_{i: |x_i - m| \geq \varepsilon} (x_i - m)^2 p_i \geq \varepsilon^2 \sum_{i: |x_i - m| \geq \varepsilon} p_i = \\ &= \varepsilon^2 P(|X - m| \geq \varepsilon) .\end{aligned}$$

Innen:

$$P(|X - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

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

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

amiből a tétel állítása leolvasható.

Megjegyzés: A Csebisev-egyenlőtlenség igen általános érvényű, minden olyan valószínűségi változó ingadozásának mértékére vonatkozik, amelynek várható értéke és szórása létezik. Éppen mert a tétel rendkívül általános érvényű, nagy pontosság az adott nagyságú 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. az $|X - m| > 2$ eltérés valószínűsége, ha X normális eloszlású, kisebb, mint 0,05. A tétel jelentősége mégis abban áll, hogy olyan esetben is tájékoztatást nyújt az eloszlásra, amikor csak a várható értéket és szórást ismerjük.

d) A valószínűségi változó momentumai

Az előző pontban a (2.37) és (2.38) összefüggések segítségével definiáltuk egy X valószínűségi változó valamely $\varphi(X)$ függvényének a várható értékét. A $\varphi(X)$ függvény speciális megválasztásával nyerjük a valószínűségi változó momentumait.

Definíció:

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

$$(2.57) \quad \alpha_k = E(X^k)$$

várható értéket¹ az X valószínűségi változó k -adik momentumának nevezzük. A várható érték X első momentuma.

Definíció:

Legyen $\varphi(X) = [X - E(X)]^k$ ekkor

$$(2.58) \quad \mu_k = E \{ [X - E(X)]^k \}$$

az X valószínűségi változó k -adik centrális momentuma.

A valószínűségi változó szórásnégyzete nem más, mint a második centrális momentum:

$$D^2(X) = \mu_2 = E \{ [X - E(X)]^2 \} = E(X^2) - E^2(X) = \alpha_2 - \alpha_1^2$$

Egyszerű számolással belátható, hogy a többi, magasabbrendű centrális momentumok is kifejezhetők a nem centrális momentumok segítségével.

Pl.

$$(2.59) \quad \mu_3 = E \{ (X - \alpha_1)^3 \} = E(X^3) - 3 \alpha_1 E(X^2) + 3 \alpha_1^2 E(X) - \alpha_1^3 = \alpha_3 - 3 \alpha_1 \alpha_2 + 2 \alpha_1^3$$

$$(2.60) \quad \mu_4 = E \{ (X - \alpha_1)^4 \} = \alpha_4 - 4 \alpha_1 \alpha_3 + 6 \alpha_1^2 \alpha_2 - 3 \alpha_1^4$$

A momentumok kiszámítása diszkrét eloszlású valószínűségi változók esetében az

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

ill.

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

formulák segítségével történik,² míg folytonos eloszlású valószínűségi változó esetén, ha X sűrűségfüggvénye $f(x)$,

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

ill.

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

1 Amennyiben ez a várható érték létezik!

2 Feltételezzük, hogy a szóban forgó összegek és integrálok abszolút konvergensek!

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

$$(2.61) \quad \gamma_1 = \frac{\mu_3}{\sigma^3}$$

menyiséget használják. (Általában folytonos eloszlás esetében használatos).

A sűrűségfüggvény lapultságának mérése használatos a

$$(2.62) \quad \gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

menyiség. Normális eloszlás esetében $\gamma_2 = 0$. Ha valamely folytonos eloszlásra vonatkozólag $\gamma_2 > 0$, akkor az illető eloszlás sűrűségfüggvénye magasabbra ugrik ki, "csúcsosabb", mint a normális eloszlás sűrűségfüggvénye (lásd: 2.2.7. pont), míg $\gamma_2 < 0$ esetén lapultabb annál.

Folytonos eloszlások számszerű jellemzésére használatos mennyiség még a medián, jele M_e .

Definíció:

A medián az az érték, amelynél a valószínűségi változó $\frac{1}{2}$ valószínűséggel vesz fel nagyobb értéket, vagyis amelyre

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

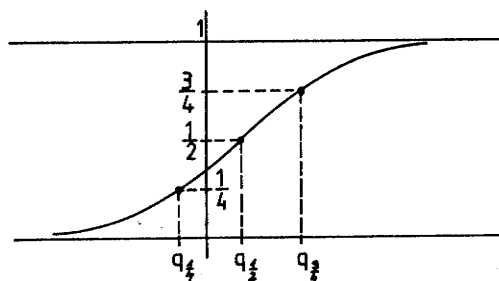
Abban az esetben, ha az $F(X) = \frac{1}{2}$ egyenletnek több megoldása van, vagyis ha az $F(X)$ eloszlásfüggvény az x_0 pontban eléri az $\frac{1}{2}$ értéket és állandó marad valamely x_1 pontig és csak ettől kezdve emelkedik, akkor

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

Definíció:

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

Nyilván $q_{\frac{1}{2}} = M_e$



24. ábra

A $q_{\frac{1}{4}}$ ill. $q_{\frac{3}{4}}$ kvantiliseket alsó-, ill. felső kvantiliseknek nevezzük.

e) 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, X -et és Y -t egyidejűleg kell tekintetbe vennünk, akkor elsősorban az szokott érdekelni bennünket, hogy függetlenek-e egymástól a kérdéses valószínűségi változók, vagy pedig értékeik között valamilyen függvénykapcsolat áll fenn. Gyakori eset, hogy X és Y értékei között bizonyos tendencia mutatkozik, pl. hogy X nagy értékei Y nagy értékeivel, míg X kisebb értékei általában Y kisebb értékeivel szoktak együtt jelentkezni.

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

Legyenek X és Y valószínűségi változók. Ha X és Y függetlenek, akkor a (2.) összefüggés alapján

$$E\{[X - E(X)][Y - E(Y)]\} = E[X - E(X)] \cdot E[Y - E(Y)] = 0$$

Ha X és Y nem függetlenek, akkor a fenti várható érték valamilyen c számmal egyenlő

Definíció: A

$$(2.63) \quad c = E\{[X - E(X)][Y - E(Y)]\}$$

mennyiséget az X és Y valószínűségi változók kovarianciájának nevezzük. Ha $Y = X$, akkor $c = E[X - E(X)]^2 =$ az X változó szórásnégyzete.

A c kovariancia értéke, attól függően, hogy milyen eloszlású valószínűségi változókról 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.

Definíció:

Az X és Y változók közötti korrelációs együtthatót a

$$(2.64) \quad \rho = \frac{E \{ [X - E(X)] [Y - E(Y)] \}}{D(X) D(Y)}$$

formulával értelmezzük.

A korrelációs együttható tehát az X és Y változók kovarianciája osztva a szórásaik szorzatával:

Tétel:

A ρ korrelációs együttható értéke -1 és +1 között változik.

Bizonyítás:

Ha X és Y olyan valószínűségi változók, amelyek négyzetének várható értéke létezik, akkor

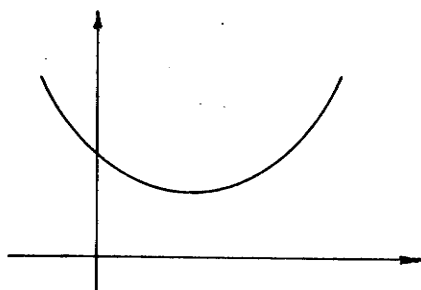
$$(2.65) \quad |E(X \cdot Y)| \leq \sqrt{E(X^2) E(Y^2)}$$

Ugyanis tetszőleges valós λ -ra

$$E(\lambda X - Y)^2 \geq 0$$

azaz

$$\lambda^2 E(X^2) - 2\lambda E(XY) + E(Y^2) \geq 0$$



25. ábra

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

$$4 E(X \cdot Y) - 4 E(X^2)E(Y^2) \leq 0$$

amiből az állítás már leolvasható.

Ha most a (2.65) összefüggésben X helyébe $X - E(X)$ -et, Y helyébe $Y - E(Y)$ -t írunk, akkor az

$$\begin{aligned} E \{ [X - E(X)] [Y - E(Y)] \} &\leq \sqrt{E [X - E(X)]^2 E [Y - E(Y)]^2} = \\ &= D(X) D(Y) \end{aligned}$$

összefüggést nyerjük, amiből következik, hogy

$$|\rho| \leq 1$$

Abban az esetben, ha az Y valószínűségi változó az X változónak lineáris függvénye:

$$Y = a X + b$$

akkor $\rho = 1$ vagy $\rho = -1$ aszerint, hogy $a > 0$ vagy $a < 0$. Ennek belátása egyszerű:

$$\begin{aligned} \rho &= \frac{E \{ [X - E(X)] [a X + b - a E(X) - b] \}}{\sqrt{D^2(X) D^2(a X + b)}} = \\ &= \frac{a E [X - E(X)]^2}{|a| D^2(X)} = \frac{a}{|a|} = \pm 1 \end{aligned}$$

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

Gyakorlati számítás céljára mind a kovariancia, mind a korrelációs együttható képlete némileg egyszerűsíthető:

$$\begin{aligned} (2.65) \quad c &= E \{ [X - E(X)] [Y - E(Y)] \} = \\ &= E [XY - E(X)Y - E(Y)X + E(X)E(Y)] = \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

innen:

$$(2.66) \quad \rho = \frac{E(X \cdot Y) - E(X)E(Y)}{D(X) D(Y)}$$

Ahhoz, hogy az X és Y valószínűségi változók közötti kovarianciát és korrelációt ki tudjuk számítani, ismernünk kell az együttes eloszlásukat.

Ha X és Y diszkrét eloszlásúak:

$$P(X = x_i) = p_i, \quad P(Y = y_k) = q_k \quad \text{és} \quad P(X = x_i, Y = y_k) = r_{ik} \\ (i, k = 1, 2, \dots)$$

akkor

$$c = \sum_i \sum_k [x_i - E(X)] [y_k - E(Y)] r_{ik} = \\ = \sum_i \sum_k x_i y_k \cdot r_{ik} - E(X) E(Y)$$

Ha X és Y folytonos eloszlásúak $h(x, y)$ együttes sűrűségfüggvényen, akkor

$$c = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x y h(x, y) dx dy - E(X) E(Y)$$

itt alkalmaztuk a (2.65) összefüggést.

Abban az esetben, ha X és Y függetlenek, a közöttük lévő kovariancia zérus. Ugyanis mivel független valószínűségi változók szorzatának várható értéke várható értékek szorzatával egyenlő

$$c = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

Ennek következtében, ha X és Y függetlenek

$$\rho = \frac{c}{D(X) D(Y)} = 0$$

Az állítás megfordítása általában nem igaz. Abból, hogy az X és Y közötti korreláció zérus, általában nem következik, hogy függetlenek. A korrelációs együttható nem annyira a sztochasztikus kapcsolat szorosságát, mint a kapcsolat linearitását méri!

(Példaként említjük, hogy ha pl. X egyenletes eloszlású a

$[-1, +1]$ intervallumban és $Y = 5X^3 - 3X$, akkor $\rho = 0$, jóllehet az X és Y valószínűségi változók között függvénykapcsolat van.)

Ha az X és Y valószínűségi változók közötti korrelációs együttható: $\rho = 0$, akkor az X és Y változókat korrelátlannak mondjuk. Mint említettük, két valószínűségi változó korrelátlansága általában nem jelent függetlenséget, ha azonban X és Y együttes eloszlás, akkor a korrelátlanságból következik a változók függetlensége. Ugyancsak maga után vonja a korrelátlanság a függetlenséget, ha X és Y két különböző esemény indikátor változója. (lásd: 2.2.1 pont)

2.1.8 A generátorfüggvény és a karakterisztikus függvény

a) A generátorfüggvény

Ismertetünk két igen hasznos analízisbeli segédeszközt, 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.

Definíció:

Legyen X valószínűségi változó, amelynek lehetséges értékei a

$0, 1, 2, \dots, n, \dots$ számok, amelyeket rendre

$p_0, p_1, p_2, \dots, p_n, \dots$ valószínűségekkel vesz fel.

Ekkor a

$$(2.67) \quad G(X) = \sum_k p_k x^k$$

függvényt, amelyet az X 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 egyenesen 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

$$G'(1) = \sum_k k p_k = E(X)$$

$$G''(x) = \sum_k k(k-1) p_k x^{k-2}$$

Ezen összefüggések alapján érvényes a következő.

Tétel:

Az X egész értékű valószínűségi változó várható értéke:

$$(2.68) \quad E(X) = G'(1)$$

szórásnégyzete:

$$(2.69) D^2(X) = G''(1) + G'(1) - [G'(1)]^2$$

Vegyük észre, hogy a

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

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

$$(2.70) E(x^X) = \sum_k x^k p_k = G(x)$$

Ebből az észrevételből fontos tételt származtathatunk független valószínűségi változók összegének generátorfüggvényére vonatkozólag.

Tétel:

Legyenek $X_1, X_2, \dots, X_n$ függetlenek, egész értékű valószínűségi változók és legyen $Y = X_1 + X_2 + \dots + X_n$. Ekkor az Y valószínűségi változó generátorfüggvénye:

$$\begin{aligned} G_Y(x) &= E(x^Y) = E(x^{X_1 + X_2 + \dots + X_n}) = \\ &= E(x^{X_1} \cdot x^{X_2} \cdot \dots \cdot x^{X_n}) = E(x^{X_1}) \cdot E(x^{X_2}) \cdot \dots \\ &\dots E(x^{X_n}) = G_{X_1}(x) \cdot G_{X_2}(x) \cdot \dots G_{X_n}(x) \end{aligned}$$

Bizonyítás:

Az $X_1, X_2, \dots, X_n$ változók függetlensége miatt az $x^{X_1}, x^{X_2}, \dots, x^{X_n}$ változók is függetlenek és független valószínűségi változók szorzatának várható értéke a várható érték 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 $G_Y(x)$ a következő alakot ölti (ha n tagú összegről van szó)

$$(2.71) \quad G_Y(x) = [G_{X_1}(x)]^n$$

A $G_Y(x)$ generátorfüggvény explicit kifejezésében x^k együtthatója adja meg a $P(Y=k)$ valószínűséget.

b) 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.

Definíció:

Az X valószínűségi változó karakterisztikus függvényen az e^{itX} komplex értékű valószínűségi változó várható értékét értjük.

Jelöljük az X valószínűségi változó karakterisztikus függvényét $\varphi_X(t)$ -vel, ekkor:

$$\varphi_X(t) = E(e^{itX}) = E(\cos tX) + i E(\sin tX) \quad (i^2 = -1)$$

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

$$(2.72) \quad \varphi_X(t) = \int_{-\infty}^{\infty} e^{itX} f(x) dx$$

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

A karakterisztikus függvény néhány fontosabb tulajdonsága:

$$(2.73) \quad |\varphi_X(t)| = \left| \int_{-\infty}^{\infty} e^{itX} f(x) dx \right| \leq \int_{-\infty}^{\infty} |e^{itX}| f(x) dx = 1$$

mivel

$$|e^{itX}| = \sqrt{\cos^2 tX + \sin^2 tX} = \sqrt{1} = 1$$

Megjegyezzük, hogy

$$\begin{aligned} \varphi_X(-t) &= E(e^{-itX}) = E[\cos(-tX) + i \sin(-tX)] = \\ &= E(\cos tX) - iE(\sin tX) = \overline{\varphi_X(t)} \end{aligned}$$

A Fourier-transzformáció elmélete szerint, ha

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

akkor

$$(2.74) \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

azaz a karakterisztikus függvény ismeretében a sűrűségfüggvény meghatározható. A karakterisztikus függvény egyértelműen meghatározza a valószínűségeloszlást. Tárgyalásaink során a (2.74) 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 $X_1, X_2, \dots, X_n$ független valószínűségi változók és legyen $Y = X_1 + X_2 + \dots + X_n$. Ekkor az Y valószínűségi változó karakterisztikus függvények az összeadandók karakterisztikus függvényeinek szorzatával egyenlő, azaz

$$\begin{aligned} (2.75) \quad \varphi_Y(t) &= E(e^{itY}) = E(e^{it(X_1 + \dots + X_n)}) = \\ &= E(e^{itX_1} \cdot e^{itX_2} \cdot \dots \cdot e^{itX_n}) = \\ &= E(e^{itX_1}) \cdot E(e^{itX_2}) \cdot \dots \cdot E(e^{itX_n}) = \\ &= \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) \cdot \dots \cdot \varphi_{X_n}(t) \end{aligned}$$

Ugyanis az $X_1, X_2, \dots, X_n$ változók függetlensége folytán az $e^{itX_1}, e^{itX_2}, \dots, e^{itX_n}$ valószínűségi változók is függetlenek és szorzatuk várható értéke várható értékeik szorzatával egyenlő.

Tétel:

Ha az Y valószínűségi változó egy másik X valószínűségi változónak lineáris függvénye: $Y = aX + b$, akkor

$$\begin{aligned} (2.76) \quad \varphi_Y(t) &= E(e^{itY}) = E[e^{it(aX+b)}] = \\ &= e^{itb} E[e^{i(at)X}] = e^{itb} \varphi_X(at) \end{aligned}$$

Megmutatjuk most, hogyan lehet a karakterisztikus függvény segítségével meghatározni, egy folytonos eloszlású valószínűségi változó momentumait, amennyiben azok léteznek.

A

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itX} f(x) dx$$

kifejezést t -szerint differenciálva:

$$\varphi'_X(t) = i \int_{-\infty}^{\infty} x e^{itX} f(x) dx$$

Az $Y = e^{itX} = \cos X + i \sin X = X_1 + i X_2$ komplex értékű valószínűségi változó várható értékét az

$$E(Y) = E(X_1) + i E(X_2)$$

szórásnégyzetét a $D^2(Y) = E(|Y - E(Y)|^2)$ formulákkal értelmezzük, és

$$(2.77) \varphi'_X(0) = i \int_{-\infty}^{\infty} x f(x) dx = i \alpha_1 = i E(X)$$

Továbbá:

$$\varphi''_X(t) = \int_{-\infty}^{\infty} (ix)^2 e^{itX} f(x) dx$$

és

$$\varphi''_X(0) = i^2 \int_{-\infty}^{\infty} x^2 f(x) dx = -E(X^2)$$

Innen:

$$E(X) = \frac{1}{i} \varphi'_X(0) = -i \varphi'_X(0)$$

$$D^2(X) = E(X^2) - [E(X)]^2 = \frac{1}{i^2} \varphi''_X(0) - \left[\frac{1}{i} \varphi'_X(0) \right]^2 =$$

$$(2.78) D^2(X) = [\varphi'_X(0)]^2 - \varphi''_X(0)$$

Általában:

$$\varphi_X^{(k)}(t) = i^k \int_{-\infty}^{\infty} x^k e^{itX} f(x) dx$$

$$(2.79) \varphi_X^{(k)}(0) = i^k \int_{-\infty}^{\infty} x^k f(x) dx = i^k \alpha_k$$

ahol α_k az X valószínűségi változó k -adik momentuma. Ennek alapján a $\varphi_X(t)$ függvény Taylor-sorfejtése a következő alakban írható:

$$(2.80) \varphi_X(t) = \varphi_X(0) + \frac{\varphi'_X(0)}{1!} t + \frac{\varphi''_X(0)}{2!} t^2 + \dots = 1 + \alpha_1 i t + \alpha_2 \frac{(it)^2}{2!} + \dots$$

2.2 Nevezetes valószínűségeloszlások

a) Diszkrét valószínűségeloszlások Egyszerű alternatíva

2.2.1 Egy esemény indikátor-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 folyó vízállásának mérése és csak az érdekel bennünket, hogy a vízállás a mérés 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 nagyszámú mérés alapján ismerjük a $P(A) = p$, $P(\bar{A}) = 1 - p = q$ valószínűséget.

Rendeljük most hozzá kísérletünkhöz az X valószínűségi változót, amely az 1 értéket veszi fel, ha az A esemény következik be és vegye fel X és 0 értéket, ha az $\bar{A}$ esemény következett be. Ekkor tehát $P(X = 1) = p$, $P(X = 0) = q = 1 - p$. Az X változó eloszlását a következő sémával szokás megadni:

$X:$	1	0
	p	q

Hasonló egyszerű alternatívával állunk szemben, ha a 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).

Definíció:

Az olyan X valószínűségi változót, amely az 1 értéket veszi fel, hogy egy kísérlet során a számunkra érdekes A esemény következett, míg $X = 0$ ha az $\bar{A}$ esemény következett be, az A esemény indikátor változójának nevezzük. Ekkor tehát:

$$P(A) = P(X = 1) = p, P(\bar{A}) = P(X = 0) = 1 - p = q$$

Számítsuk ki az indikátor változó várható értékét és szórásnégyzetét. A várható érték definíciója és X eloszlása alapján:

$$(2.81) E(X) = 1 \cdot p + 0 \cdot q = p$$

$$(2.82) E(X^2) = 1^2 \cdot p + 0^2 \cdot q = p$$

$$D^2(X) = E(X^2) - E^2(X) = p - p^2 = p(1 - p) = pq$$

Az indikátor változó generátorfüggvénye:

$$(2.83) \quad G(X) = \sum_k p_k x^k = q x^0 + p x = px + q$$

2.2.2 A binomiális eloszlás

A műszaki gyakorlatban mind elméleti, mind gyakorlati szempontból a diszkrét valószínűségeloszlások közül egyik legfontosabbnak a binomiális-, vagy Bernoulli-eloszlást tekinthetjük. Nagyon gyakori eset, hogy valamely kísérletnél csak az érdekel bennünket, hogy bizonyos A esemény bekövetkezik-e vagy sem (tehát az $\bar{A}$ esemény következik be) és a szóbanforgó kísérlet sokszor egymásután ismétlődik.

Ha olyan kísérletsorozatot tekintünk, amelyben minden kísérlet egyszerű alternatíva és az egyes kísérletek kimenetelei függetlenek egymástól, akkor az ilyen alternatíva sorozatot Bernoulli-féle kísérletsorozatnak nevezzük. Ilyen Bernoulli-féle kísérletsorozat pl. a fej-, vagy írás játék az érmedobással. Mutatunk most egy példát az alternatíva-sorozatra a hidrológiai gyakorlatból. Tekintsük pl. a Tisza folyóra vonatkozólag Szegednél az elmúlt 100 év során észlelt évi maximális vízállásokat.

Jelöljük most A-val azt az eseményt, hogy Szegednél az évi maximális vízállás nagyobb mint 700 m.

Legyen az A esemény valószínűsége $P(A) = p$.

Számítsuk ki mennyi a valószínűsége, hogy n-év során az A esemény k-szor következik be! Jelöljük X-szel az A esemény bekövetkezéseinek számát n-év során. Ekkor X lehetséges értékei a 0, 1, 2, ..., n nemnegatív számok. Keressük a $p_k = P(X = k)$ valószínűséget. (Az évi maximális vízállásértékeket függetleneknek tekinthetjük, lásd: 6. fej. 5.1. pontja).

Ha minden évben regisztráljuk, hogy az évi maximális vízállást illetőleg az A vagy az $\bar{A}$ esemény következett-e be, pl. úgy, hogy A bekövetkezésekor 1-et, A esetén 0-t írunk, akkor az n megfigyelésből álló kísérletsorozat eseménytere az összes-n hosszúságú 0 és 1 jegyekből álló sorozatok halmaza.

$$e_1 = 0 \ 0 \ 0 \ \dots \ 0 \ 0$$

$$e_2 = 0 \ 0 \ 0 \ \dots \ 0 \ 1$$

$$e_3 = 0 \ 0 \ 0 \ \dots \ 1 \ 0$$

$$\dots\dots\dots$$

$$e_{n+1} = 1 \ 0 \ 0 \ \dots \ 0 \ 0$$

$$\dots\dots\dots$$

$$e_{2n} = 1 \ 1 \ 1 \ \dots \ 1 \ 1$$

Ilyen eseményteret már a 2. fejezet 1.1 pontjában vizsgáltunk. Most azonban nem mindegyik elemi esemény egyforma valószínűségű.

$$\text{pl. } P(e_1) = q^n$$

$$P(e_2) = P(e_3) = \dots = P(e_{mn}) = pq^{n-1}$$

⋮

$$P(e_{2n}) = p^n$$

Ha valamelyik e_i elemi esemény k darab 1-est (és $n-k$ darab 0-t) tartalmaz, akkor

$$P(e_i) = p^k q^{n-k}$$

Az $\{X = k\}$ esemény akkor következik be, ha a kísérlet (n db megfigyelésből álló sorozat) eredménye olyan e_i elemi esemény, amely k -darab 1-est és $(n-k)$ darab 0-t tartalmaz. Ilyen e_i elemi esemény $\binom{n}{k}$ számú van és mindegyik $p^k q^{n-k}$ valószínűséget. Következésképpen:

$$(2.84) \quad P(X = k) = \binom{n}{k} p^k q^{n-k}$$

Amikor az A esemény bekövetkezésekor 1-est, $\bar{A}$ esetén pedig 0-t írunk, akkor minden egyes megfigyeléshez az A esemény indikátor-változóját rendeljük hozzá. Nyilván:

$$X = X_1 + X_2 + \dots + X_n$$

ahol X_i a sorozat i -edik megfigyelésének indikátor változója. Az indikátor változók felvett értékeinek összege éppen az A eseménybekövetkezéseinek számát adja meg, hiszen $\{X = k\}$ akkor következik be, ha az $X_1, X_2, \dots, X_n$ indikátor változók között k számú az 1 értéket $(n-k)$ számú a 0 értéket veszi fel.

Hasonlóan, ha pl. X egy adott év májusában az esős napok számát jelöli és $X_1, X_2, \dots, X_n$ az egyes napok karakterisztikus változói, amelyek tehát az 1 vagy 0 értéket veszik fel aszerint, hogy az illető napon esett az eső vagy sem, akkor X értéke nyilván annyi, ahány 1-es van az összeadandók között:

$$X = X_1 + X_2 + \dots + X_n = 1 + 0 + 0 + 1 + \dots + 0$$

A binomális eloszlású X valószínűségi változó várható értéke:

$$(2.85) E(X) = E(X_1) + E(X_2) + \dots + E(X_n) = n p$$

mivel az X_i karakterisztikus változók mindegyikének várható értéke p . A binomiális eloszlású X valószínűségi változó szórásnégyzete az (1.66) összefüggés alapján:

$$(2.86) D^2(X) = D^2(X_1) + \dots + D^2(X_n) = \\ = p q + p q + \dots + p q = n p q$$

Tehát a binomális eloszlás szórása:

$$(2.87) D(X) = \sqrt{n p q}$$

Az X változó generátorfüggvénye a (2.71) összefüggés szerint:

$$(2.88) G(X) = \sum_k p_k x^k = \prod_{i=1}^n G_i(x) = (p x + q)^n$$

ahol $G_i(x) = p x + q$ az X_i karakterisztikus változó generátorfüggvénye. A $G(x)$ generátorfüggvényben x^k együtthatója éppen a $p_k = P(X = k)$ valószínűséget szolgáltatja, ez pedig

$$p_k = \binom{n}{k} p^k q^{n-k}$$

Mind a várható érték, mind a szórásnégyzet egyszerűen kiszámítható a

$$G(x) = (p x + q)^n$$

generátorfüggvény segítségével is, ugyanis a (2.68) és (2.69) formulák alapján:

$$(2.89) E(X) = G'(1) = [n(p x + q)^{n-1} p]_{x=1} = n p$$

$$(2.90) D^2(X) = G''(1) - [G'(1)]^2 = \\ = [n(n-1)p^2(p x + q)^{n-2}]_{x=1} + n p - n^2 p^2 = \\ = n^2 p^2 - n p^2 + n p - n^2 p^2 = n p (1-p) = n p q$$

A binomális eloszlású X valószínűségi változó eloszlásának sémája:

$X:$	0	1	...	k	...
p_k	$p_0 = \binom{n}{0} p^0 q^n$	$p_1 = \binom{n}{1} p q^{n-1}$		$p_k = \binom{n}{k} p^k q^{n-k}$	
	...	n			
		$p_n = \binom{n}{n} p^n q^0$			

Felmerül a kérdés, hogy adott p esetén az X 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 $(n+1)p$ p -ben foglalt legnagyobb egész számmal egyenlő. Ha n nagy szám, akkor mivel $p < 1$, jórformán 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 binomális eloszlású X valószínűségi változó p_k 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
csapadékos nap	7,6	6,8	7,3	7,4	8,5	8,0	6,4	6,3	6,2
hó	X	XI	XII						
csapadékos nap	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

$$p_7 = \binom{30}{7} \left(\frac{1}{4}\right)^7 \left(\frac{3}{4}\right)^{23} = 2035\,800 \frac{3^{23}}{4^{30}} \approx 0,16$$

Az ilyen valószínűségek kiszámítása a binomális együtthatók miatt kellemetlen, ezért közelíteni szoktunk. (lásd: 2.2.5 és 2.2.8) Jelen példánkban $\binom{30}{7}$ értéke megtalálható a binomális 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:

$$P = \sum_{k=5}^{11} \binom{30}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{30-k}$$

E valószínűség kiszámítása igen munkaigényes lenne, ezért közelítjük. (lásd: 2.2.8 pont).

2.2.3 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

$$P(A_1) = p_1, \quad P(A_2) = p_2, \dots, \quad P(A_s) = p_s$$

($\sum p_i = 1$), akkor a kísérletet n -szer egymás után megismételve, az ún. polinomiális eloszláshoz jutunk. Ez a fogalom lényegében a binomiális eloszlás általánosítása és $s = 2$ esetén megegyezik vele.

Annak valószínűsége, hogy az A_1 esemény K_1 -szer, az A_2 esemény K_2 -ször, az A_s esemény K_s -szer következik be:

$$P_{K_1, K_2, \dots, K_s} = \frac{n!}{K_1! K_2! \dots K_s!} p_1^{K_1} \cdot p_2^{K_2} \dots p_s^{K_s}$$

2.2.4 A geometriai eloszlás

Tekintsünk egy olyan kísérletet, amelynek két lehetséges kimenetele van, tehát egyszerű alternatíva, s jelöljük az egyik kimenetelt A -val, a másikat $\bar{A}$ -sal. Legyen

$$P(A) = p, \quad P(\bar{A}) = 1 - p = q$$

Tegyük fel, hogy a kísérletet többször egymás után megismételjük mindaddig, amíg az A esemény először bekövetkezik.

Annak valószínűsége, hogy az A esemény először a k -adik kísérletnél következik be:

$$(2.91) \quad p_k = q^{k-1} \cdot p$$

Könnyű belátni, hogy $\sum_k p_k = 1$ ugyanis

$$p_k = p \sum_k q^{k-1} = \frac{p}{1 - q} = \frac{p}{p} = 1$$

A geometriai eloszlás generátorfüggvénye:

$$(2.92) \quad G(x) = \sum_k p q^{k-1} x^k = p \times \sum_k (qx)^{k-1} = \frac{px}{1 - qx}$$

Mivel

$$G'(x) = \frac{p(1 - qx) + p q x}{(1 - qx)^2} = \frac{p}{(1 - qx)^2}$$

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:

$$(2.93) E(X) = G'(1) = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

$$(2.94) D^2(X) = G''(1) + G'(1) - [G'(1)]^2 = \\ = \frac{2p^2q}{p^4} + \frac{1}{p} - \frac{1}{p^2} = \frac{q}{p^2}$$

2.2.5 A Poisson-eloszlás

A diszkrét valószínűségi 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 $[0, t)$ időintervallumban (mondjuk adott negyedév során) fellépő árhullámok száma (mint azt a (2.2.6) pontban látni fogjuk) Poisson-eloszlást követ. Ugyancsak Poisson-eloszlást követ adott időegység alatt egy útszakaszon áthaladó gépkocsik száma, adott időintervallumban elbomló rádióaktív atomok száma, stb.

Származtassuk először a Poisson-eloszlást a binomiális eloszlás határeseteként. Legyen $P(A) = p$, ekkor $P(\bar{A}) = 1 - p$. Itt egyszerű alternatívával állunk szemben, tehát annak valószínűsége, hogy n év során az A esemény k -szor következik be: (az évi maximális vízállásokat függetleneknek tételezzük fel)

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

Ezt a formulát n nagy értéke esetén rendkívül nehézkes lenne kiszámítani, ezért átalakítjuk a következő alakra:

$$P_k^{(n)} = \frac{1}{k!} \frac{n(n-1) \dots (n-k+1)}{n^k} (np)^k \left(1 - \frac{np}{n}\right)^{n-k}$$

Ha most $n \rightarrow \infty$, $p \rightarrow 0$ úgy, hogy $np = \lambda$ állandó akkor a fenti formulában az np helyébe λ -t írva:

$$p_k^{(p)} = \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \lambda^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Elvégezve az $n \rightarrow \infty$ határátmenetet:

$$(2.95) \quad p_k = \frac{\lambda^k}{k!} e^{-\lambda}$$

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

$$\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e \quad \text{és} \quad \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n}\right)^n = e^a$$

Ugyanis:

$$\lim_{\frac{n}{a} \rightarrow \infty} \left[\left(1 + \frac{1}{\frac{n}{a}}\right)^{\frac{n}{a}} \right]^a = e^a$$

tetszőleges valós számra. Az $a = -\lambda$ esetén kapjuk a (2.95) formulában lévő határértéket.

Tehát ha az Y valószínűségi változó jelöli az A esemény bekövetkezéseinek számát, akkor Y eloszlása:

K						
$p_k = P(Y=k)$	$\frac{\lambda^0}{0!} e^{-\lambda}$	$\frac{\lambda^1}{1!} e^{-\lambda}$	$\frac{\lambda^2}{2!} e^{-\lambda}$	$\dots$	$\frac{\lambda^k}{k!} e^{-\lambda}$	$\dots$

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. Ezt formálisan is beláthatjuk, ugyanis a Poisson-eloszlású valószínűségi változó várható értéke:

$$\begin{aligned} (2.96) \quad E(Y) &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \\ &= \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda \end{aligned}$$

1.) Megjegyzés: A gyakorlatban p rendszerint adott. Ha

$p < \frac{0,637}{\sqrt{n}}$, akkor a Poisson-eloszlással való közelítés indokolt.

Emlékeztetünk, hogy az e^x függvény Taylor-sora:

$$e = 1 + x + \frac{x^2}{2!} + \dots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Nyilván az $x = \lambda$ esetén

$$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$$

Kiszámítjuk a Poisson-eloszlású Y valószínűségi változó szórásnégyzetét is.

$$\begin{aligned} (2.97) \quad D^2(Y) &= E(Y^2) - [E(Y)]^2 = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 = \\ &= \lambda e^{-\lambda} \left[\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} \right] - \lambda^2 = \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} - \lambda^2 = \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} - \lambda^2 = \\ &= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} - \lambda^2 = \lambda \end{aligned}$$

Tehát a Poisson-eloszlású Y valószínűségi változó szórása:

$$(2.98) \quad D(Y) = + \sqrt{\lambda}$$

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: 5.1.4 pont). A λ paraméter ismeretében a P_k valószínűségek értéke általában táblázatból kikereshető.

A Poisson-eloszlás P_k valószínűségeire egyébként érvényes a következő rekurzió:

$$P_k = \frac{\lambda}{k} P_{k-1} \quad (k=1, 2, \dots)$$

b) Folytonos eloszlások

2.2.6 A normális eloszlás (Gauss-eloszlás)

Amikor egy kísérlet kimenetelét nagyszámú, egymástól csak kevéssé 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 tényezők 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.

Definíció:

Az X valószínűségi változót akkor nevezzük normális eloszlásúnak, ha sűrűségfüggvénye:

$$(2.99) \quad f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

ahol m és σ állandók, $\sigma > 0$. A formulában szereplő m és σ mennyiségeket a normális eloszlás paramétereinek nevezzük. A (2.99) formulát szemügyre véve könnyű belátni, hogy bármely 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:

$$1/\exp\left(\frac{x-m}{\sigma\sqrt{2}}\right)^2$$

mindig pozitív. A kitevő jelen esetben ugyancsak pozitív vagy zérus. Ebből az is kiolvasható, hogy $f(x)$ értéke akkor maximális, ha a kitevő zérus, azaz $x = m$. Ekkor

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}}$$

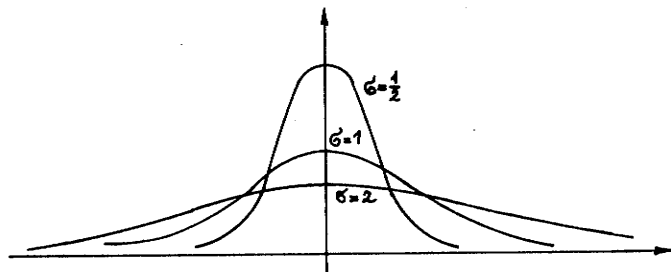
Minél kisebb a σ paraméter értéke, annál nagyobb $f(x)$ maximális értéke. $\sigma = 1$ esetén a maximum értéke

$$f_{\max}(x) = \frac{1}{\sqrt{2\pi}} \approx 0,4$$

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, $f(x)$ az $x_0 = m$ pontra szimmetrikus.

Az $f(x)$ függvény képe ún. haranggörbe. A függvény alakját különböző σ értékeknél az ábra szemlélteti.

Ahhoz, hogy az (2.99) alatti kifejezés sűrűségfüggvény legyen, az kell, hogy a görbe alatti terület egységnyi legyen.



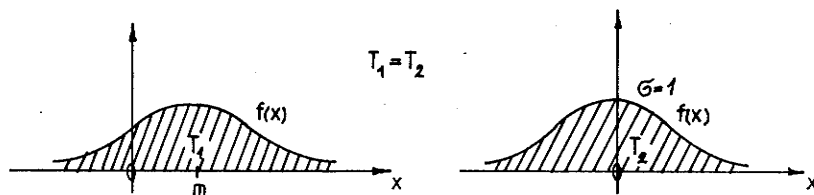
26. ábra

$$(2.100) \quad I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

Nem nehéz kimutatni, hogy ez valóban így van. Vezessük be az (2.100) integrálban az) $u = \frac{x-m}{\sigma}$ helyettesítést, ekkor $x = \sigma u + m$, $dx = \sigma du$ tehát az

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du$$

integrált nyerjük, amelynek természetesen ugyanaz az értéke, mint a (2.100) 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



27. ábra

a normális sűrűségfüggvény alatti terület nagysága nem függ az m és σ paraméterektől. Elegendő tehát kimutatni, hogy

$$(2.101) \quad I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = 1$$

Ennek igazolását a legegyszerűbben úgy végezhetjük el, hogy kimutatjuk az $I^2 = 1$ reláció fennállását.

$$\begin{aligned} I^2 &= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du \right)^2 = \\ &= \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \right) \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \right) = \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dx dy = \frac{1}{2\pi} \int_0^{2\pi} \left[\int_0^{\infty} r e^{-\frac{r^2}{2}} dr \right] d\varphi \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\varphi = 1 \end{aligned}$$

Itt alkalmaztuk az

$$x = r \cos \varphi$$

$$y = r \sin \varphi$$

polárkoordinátás transzformációt, ekkor

$$dx dy = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \varphi} \end{vmatrix} dr d\varphi = \begin{vmatrix} \cos & -r \sin \\ \sin & r \cos \end{vmatrix} dr d\varphi = r dr d\varphi$$

Ezzel kimutattuk, hogy az (2.100) alatti folytonos pozitív értékű $f(x)$ görbe alatti terület egységnyi, azaz $f(x)$ sűrűségfüggvény.

A sűrűségfüggvény és eloszlásfüggvény közötti összefüggés alapján a normális eloszlás eloszlásfüggvénye:

$$(2.101) \quad F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

A (2.101) összefüggésben ismét elvégezve az $u = \frac{t - m}{\sigma}$ helyettesítést, az

$$(2.102) F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-m}{\sigma}} e^{-\frac{u^2}{2}} du = \Phi\left(\frac{x-m}{\sigma}\right)$$

relációt nyerjük, ahol $\Phi(x)$ az $m = 0$, $\sigma = 1$ paraméterekkel bíró normális eloszlás eloszlásfüggvényét jelöli. A $\Phi(x)$ függvény értékeit az I. táblázat tartalmazza.

A (2.102) összefüggés értelmét kissé részletesebben megvilágítjuk. Legyen X normális eloszlású valószínűségi változó m és $\sigma > 0$ paraméterekkel (rövid jelöléssel: $X : N(m, \sigma)$), akkor X eloszlásfüggvénye az

$$F(x) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

adja annak a valószínűségét, hogy X értéke kisebb, mint x .

Az X valószínűségi változó felvett értékei (a kísérlet kimenetelei) helyett tekintsük az $X^* = \frac{X - m}{\sigma}$ értékeket. X^* természetesen ugyancsak valószínűségi változó, amelynek értékeit X felvett értékei egyértelműen meghatározzák. Az X^* valószínűségi változót standardizált változónak nevezzük. Ha az X valószínűségi változó egy kísérletnél

az $X = x_0$ értéket veszi fel, akkor az X^* változó ugyanannál a kísérletnél az $\frac{x_0 - m}{\sigma}$ értéket képviseli. Az X^* változó eloszlásfüggvényét jelöljük $\Phi(x)$ -szel. Nyilvánvaló, hogy az $\{X < x\}$ esemény valószínűsége ugyanaz, mint az $\{X^* < \frac{x-m}{\sigma}\}$ esemény valószínűsége,

$$P(X < x) = P(\sigma X^* + m < x) = P(X^* < \frac{x-m}{\sigma})$$

azaz

$$F(x) = \Phi\left(\frac{x-m}{\sigma}\right)$$

Annak valószínűsége, hogy az m és $\sigma > 0$ paraméterekkel bíró normális eloszlású változó az (a, b) intervallumba essék:

$$\begin{aligned} (2.103) F(b) - F(a) &= P(a \leq X < b) = \\ &= P(a \leq \sigma X^* + m < b) = \\ &= P\left(\frac{a-m}{\sigma} \leq X^* < \frac{b-m}{\sigma}\right) = \\ &= \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right) \end{aligned}$$

Az $\phi(x)$ függvény értéktáblázatát általában pozitív x -értékekre adják meg. Megmutatjuk, hogy igen egyszerű módon nyerjük $\phi(x)$ értékeit negatív x értékekre is.

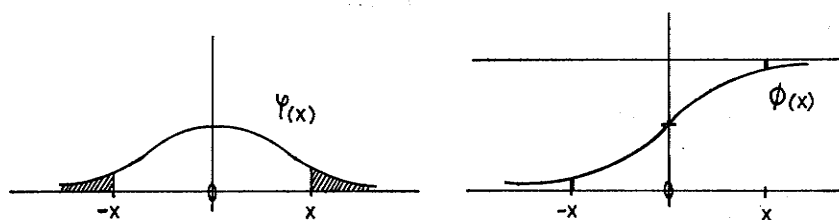
Az X^* standardizált normális valószínűségi változó sűrűségfüggvénye az (2.99) sűrűségfüggvényből $m=0$, $\sigma=1$ helyettesítéssel nyerhető:

$$(2.104) \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

A $\varphi(x)$ függvény az $x = 0$ pontra szimmetrikus, így

$$(2.15) \quad \phi(-x) = \int_{-\infty}^{-x} \varphi(t) dt = \int_x^{\infty} \varphi(t) dt = 1 - \int_{-\infty}^x \varphi(t) dt = 1 - \phi(x)$$

amely összefüggés a $\phi(x)$ eloszlásfüggvény grafikonjáról is leolvasható.

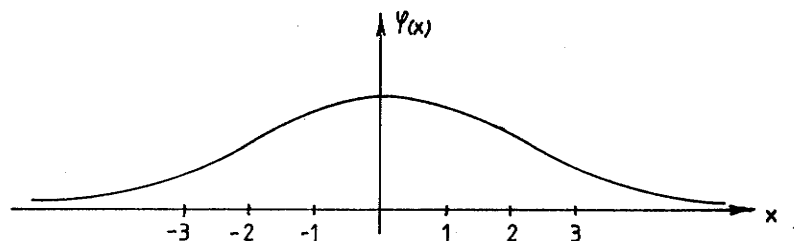


28. ábra

Annak valószínűsége, hogy az X^* valószínűségi változó értéke az $(-x, x)$ intervallumba esik:

$$(2.106) \quad \phi(x) - \phi(-x) = \phi(x) - [1 - \phi(x)] = 2\phi(x) - 1$$

Az (2.103) összefüggés alapján könnyen belátható, hogy a következő két táblázat teljesen ekvivalens:



29. ábra

$X^{\#}: N(0, 1)$	$X : N(m, \sigma^2)$
$P(0,67 \leq X^{\#} < 0,67) = 0,5$	$P(m-0,67\sigma \leq X < m+0,67\sigma) = 0,5$
$P(-1 \leq X^{\#} < 1) = 0,683$	$P(m-\sigma \leq X < m+\sigma) = 0,683$
$P(-2 \leq X^{\#} < 2) = 0,954$	$P(m-2\sigma \leq X < m+2\sigma) = 0,954$
$P(-3 \leq X^{\#} < 3) = 0,997$	$P(m-3\sigma \leq X < m+3\sigma) = 0,997$

E táblázatból pl. kiolvasható, hogy igen kicsi, kisebb mint 0,05 annak valószínűsége, hogy egy m és σ paraméterekkel bíró normális eloszlású X valószínűségi változó megfigyelt értéke az m ponttól σ -nál nagyobb távolságra essék. Ezt a tényt a matematikai statisztikában gyakran fel fogjuk használni, ez az ún. 2σ -szabály. Az pedig gyakorlatilag úgyszólván biztos, hogy X megfigyelt értéke az $(m-3\sigma, m+3\sigma)$ intervallumba esik. A fenti táblázat adatai azt sugallják, hogy m paraméter az X valószínűségi változó várható értéke, amely körül X megfigyelt értékei tömörülnek, a σ paraméter pedig az eloszlás szórása.

Valóban:

$$\begin{aligned}
 E(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m+m) e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \\
 &= \frac{\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{x-m}{\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx + \frac{m}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx
 \end{aligned}$$

Ha most alkalmazzuk az $U = \frac{x-m}{\sigma}$ helyettesítést, akkor azt kapjuk, hogy

$$(2.107) \quad E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du + \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = m$$

Továbbá

$$\begin{aligned}
 (2.108) \quad D^2(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u^2 e^{-\frac{u^2}{2}} du = \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot u e^{-\frac{u^2}{2}} du = \frac{\sigma^2}{\sqrt{2\pi}} \left[-u e^{-\frac{u^2}{2}} \right]_{-\infty}^{\infty} + \\
 &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du = \sigma^2
 \end{aligned}$$

A továbbiakban többször fel fogjuk használni a normális eloszlás karakterisztikus függvényét, amelyet először az $N(0,1)$ standard normális eloszlású X^* valószínűségi változóra származtatunk.

$$(2.109) \quad \varphi(t) = E(e^{itX^*}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx = \\ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{(x-it)^2}{2}} e^{-\frac{t^2}{2}} dx = e^{-\frac{t^2}{2}}$$

Itt alkalmaztuk 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 az X valószínűségi változó $N(m, \sigma^2)$ eloszlású, azaz

$$X = \sigma X^* + m$$

akkor a (2.109) összefüggés alapján

$$(2.110) \quad \varphi_X(t) = E(e^{it(\sigma X^* + m)}) = e^{itm} E(e^{it\sigma X^*}) = \\ = e^{itm} e^{-\frac{\sigma^2 t^2}{2}} = e^{itm - \frac{\sigma^2 t^2}{2}}$$

Természetesen a $\varphi_X(t)$ karakterisztikus függvény $t = 0$ helyen vett deriváltjai segítségével igen egyszerűen származtathatók a normális eloszlás momentumai. (Ennek elvégzését az olvasóra bízjuk).

Legyenek most az X_1 és X_2 független normális eloszlású valószínűségi változók $N(m_1; \sigma_1^2)$, illetve $N(m_2; \sigma_2^2)$ eloszlásokkal. Legyen $X = X_1 + X_2$. Határozzuk meg az X eloszlását. Mivel a karakterisztikus függvény az eloszlást egyértelműen meghatározza, elegendő számunkra X karakterisztikus függvényének meghatározása, ennek képletéből az eredmény kiolvasható. A (2.110) formula alapján:

$$\varphi_X(t) = E[e^{it(X_1 + X_2)}] = E(e^{itX_1} \cdot e^{itX_2}) = \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) = \\ = e^{itm_1 - \frac{\sigma_1^2 t^2}{2}} \cdot e^{itm_2 - \frac{\sigma_2^2 t^2}{2}} = e^{it(m_1 + m_2) - \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}}$$

Ebből látható, hogy X ugyancsak normális eloszlású és

$$(2.111) \quad E(X) = m_1 + m_2 \\ D^2(X) = \sigma_1^2 + \sigma_2^2$$

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 tulajdonságra hívjuk fel a figyelmet, amelyek a gyakorlati alkalmazások szempontjából rendkívül fontosak.

Ha X normális eloszlású, $E(X) = m$ várható értékkel és $D(X) = \sigma^2$ szórással rendelkezik és $Y = aX + b$ az X változó lineáris függvénye, akkor Y ugyancsak normális eloszlású, $E(Y) = am + b$ várható értékkel és $D(Y) = a^2 \sigma^2$ szórással bír.

Állításunkat az (2.76) összefüggés segítségével könnyen bebizonyíthatjuk.

Mivel az X valószínűségi változó sűrűségfüggvénye:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

és a (2.76) formula szerint az Y változó sűrűségfüggvénye:

$$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right) = \frac{1}{|a| \sigma \sqrt{2\pi}} e^{-\frac{\left(\frac{y-b}{a} - m\right)^2}{2\sigma^2}}$$

azaz

$$(2.112) \quad g(y) = \frac{1}{|a| \sigma \sqrt{2\pi}} e^{-\frac{(y-am-b)^2}{2a^2 \sigma^2}}$$

Az utóbbi formulából állításunk máris kiolvasható, hiszen a (2.112) formulában szereplő függvény normális sűrűségfüggvény, amelyben $am + b$ a várható érték, és a σ a szóráss.

Mivel a különböző normális eloszlásokat az m és $\sigma > 0$ paraméterek értéke egyértelműen meghatározza, az előbbiekből az is következik, hogy ha két normális eloszlású változóval van dolgunk - $X : N(m_1, \sigma_1)$ eloszlású és $Y : N(m_2, \sigma_2)$ eloszlású -, akkor mindig található olyan $a > 0$ és $b > 0$ számok, hogy az $Y = aX + b$ eloszlása $N(m_2, \sigma_2)$.

Más szóval egy normális eloszlású valószínűségi változó lineáris transzformációval másik előírt normális eloszlásba vihető át. Az $a > 0$ és $b > 0$ 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

$$a \sigma_1 = \sigma_2 \text{ azaz} \\ a = \frac{\sigma_2}{\sigma_1} \text{ és } b = m_2 - \frac{\sigma_2}{\sigma_1} m_1$$

teljesül.

2.2.7 A binomiális eloszlás közelítése normális eloszlással

A Moivre-Laplace tétel

A binomiális eloszlás tárgyalásánál említettük, hogy a

$$(2.13) p_k = \binom{n}{k} p^k (1-p)^{n-k}$$

valószínűségek kiszámítása nagy n értéknél igen fáradságos, szinte reménytelen, s ezért arra törekszünk, hogy az (2.113) 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:

$$\frac{0,637}{3\sqrt{n}} < p < 1 - \frac{0,637}{3\sqrt{n}}$$

(ahol n a kísérletek száma), akkor a célszerű közelítés:

$$(2.114) p_k = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}$$

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, $\sqrt{npq}$ szórással 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.

A (2.114) formulával történő közelítés anélkül 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 a (2.114) formulában bevezetjük az $x = \frac{k-np}{\sqrt{npq}}$ helyettesítést, akkor

$$(2.115) \quad p_k \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{npq}} = \varphi(x) \frac{1}{\sqrt{npq}}$$

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.

A (2.115) összefüggés matematikai származtatása az analízis eszközeivel egyszerű, nem igényel mást, csupán a Stirling-formula alkalmazását és az $y = \ln(1+n)$ függvény Taylor-sorának ismeretét.

A Stirling-formula alkalmazásával:

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{k!(n-k)!} \approx \frac{\frac{n^n}{e^n} \sqrt{2\pi n}}{\left(\frac{k^k}{e^k}\right) \left(\frac{(n-k)^{n-k}}{e^{n-k}}\right) \sqrt{2\pi k} \sqrt{2\pi(n-k)}} = \\ &= \frac{1}{\sqrt{2\pi \frac{k}{n} \left(1 - \frac{k}{n}\right)}} \frac{n^n}{k^k (n-k)^{n-k}} \end{aligned}$$

Ennek alapján a p_k valószínűség a következő alakban írható:

$$\begin{aligned} p_k &= \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi n \frac{k}{n} \left(1 - \frac{k}{n}\right)}} \left(\frac{np}{k}\right)^k \\ &\quad \left(\frac{nq}{n-k}\right)^{n-k} \end{aligned}$$

Alkalmazzuk az $x = k - np$ helyettesítést, azaz legyen

$$k = np + x$$

$$n - k = nq - x$$

akkor

$$\begin{aligned} (2.116) \quad p_k &\approx \frac{1}{\sqrt{2\pi n \left(p + \frac{x}{n}\right) \left(q - \frac{x}{n}\right)}} \\ &\quad \frac{1}{\left(1 + \frac{x}{np}\right)^{np+x} \left(1 - \frac{x}{nq}\right)^{nq-x}} \end{aligned}$$

Könnyű belátni, hogy n növekedésével $\frac{x}{n} \rightarrow 0$ így az első tényező $\frac{1}{\sqrt{2\pi npq}}$ -hoz tart, amint az a (2.114) formulában szerepel. A második tényező nevezőjének logaritmus:

$$\begin{aligned} & (np+x) \ln\left(1 + \frac{x}{np}\right) + (nq-x) \ln\left(1 - \frac{x}{nq}\right) = \\ & = (np+x) \left[\frac{x}{np} - \frac{x^2}{2n^2 p^2} + \frac{x^3}{3n^3 p^3} - \dots \right] - \\ & - (nq-x) \left[\frac{x}{nq} + \frac{x^2}{2n^2 q^2} + \frac{x^3}{3n^3 q^3} - \dots \right] = \\ & = x - \frac{x^2}{2np} + \frac{x^3}{3n^2 p^2} - \dots + \frac{x^2}{np} - \frac{x^3}{2n^2 p^2} + \dots - \\ & - x - \frac{x^2}{2nq} - \frac{x^3}{3n^2 q^2} - \dots + \frac{x^2}{nq} + \frac{x^3}{2n^2 q^2} + \dots = \\ & = -\frac{x^2}{2n} \left(\frac{1}{p} + \frac{1}{q} \right) - \frac{x^3}{6n^2} \left(\frac{1}{p^2} - \frac{1}{q^2} \right) - \dots \approx -\frac{x^2}{2npq} \end{aligned}$$

A második tagtól kezdve a tagokat elhanyagolhatjuk, mert már $\frac{x^2}{n^2} \rightarrow 0$, 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. Következésképpen

$$\frac{x^2}{n^2} \approx \frac{x^{3/2}}{n^2} = \frac{1}{\sqrt{n}} \rightarrow 0$$

Az (2.116) formula jobb oldalának második tagjában szereplő

nevező tehát $\approx e^{-\frac{x^2}{2npq}}$ így a p_k valószínűség:

$$p_k \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}}$$

2.2.8 Kétdimenziós (síkbeli) normális eloszlás

A többdimenziós valószínűségeloszlások közül részletesen csak a gyakorlatban legnagyobb jelentőségű síkbeli normális eloszlást vizsgáljuk.

A $\bar{Z} = (X, Y)$ véletlen vektort kétdimenziós normális eloszlásúnak nevezzük, ha sűrűségfüggvénye:

$$(2.117) \quad h(x, y) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho \frac{(x-m_1)(y-m_2)}{\sigma_1 \sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2} \right]}$$

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 a (2.117) formulában szereplő paraméterek jelentését.

Kimutatjuk, hogy m_1 , ill. σ_1 az X változó várható értékét és szórását, m_2 , ill. σ_2 az Y változó várható értékét és szórását, ρ pedig az X és Y változók közötti korrelációs együtthatót jelenti. Ennek belátását a következő módon végezzük. A kétdimenziós eloszlásról mondottak alapján (lásd: a (1.15) formulát) az X változó sűrűségfüggvénye:

$$f(x) = \int_{-\infty}^{\infty} h(x, y) dy$$

A (2.117) formulában vezessük be az

$$u = \frac{x - m_1}{\sigma_1}, \quad v = \frac{y - m_2}{\sigma_2}$$

helyettesítést, s ekkor

$$f(x) = \int_{-\infty}^{\infty} h(x, y) dy = \frac{1}{2\pi \sigma_1 \sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2(1-\rho^2)} [u^2 - 2\rho uv + v^2]} dv = \frac{e^{-\frac{u^2}{2}}}{2\pi \sigma_1 \sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \frac{e^{-\frac{(v-u)^2}{2(1-\rho^2)}}}{2(1-\rho^2)} dv$$

Alkalmazva a $z = \frac{v - \rho u}{\sqrt{1 - \rho^2}}$ változót, $dv = \sqrt{1 - \rho^2} dz$ és így:

$$f(x) = \frac{e^{-\frac{u^2}{2}}}{2\pi\sigma_1} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz =$$

$$= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

azaz:

$$f(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x - m_1)^2}{2\sigma_1^2}}$$

amely eredmény mutatja, hogy X normális eloszlású, m_1 várható értékkel és σ_1 szórással. Mivel a (2.117) formula az X és Y változóiban szimmetrikus, teljesen analóg meggondolással nyerjük az Y változó $g(y)$ sűrűségfüggvényét:

$$g(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y - m_2)^2}{2\sigma_2^2}}$$

tehát Y is normális eloszlású, m_2 várható értékkel és σ_2 szórással.

Meggondolásunk eredményeként nemcsak az m_1 , m_2 és σ_1 , σ_2 paraméterek jelentése vált világossá, hanem egyben azt is láthatjuk, hogy ha X és Y valószínűségi változók együttes eloszlása kétdimenziós normális eloszlás, akkor az egyes komponensek külön-külön is normális eloszlást követnek, vagyis a kétdimenziós normális eloszlás vetületi egydimenziós normális eloszlások.

Kimutatjuk, hogy a (2.117) formulában szereplő paraméter az X és Y változók közötti korrelációs együttható. Az X és Y változók közötti kovariancia az (1.46) összefüggés alapján:

$$\begin{aligned} c &= E(X - m_1)(Y - m_2) = \iint_{-\infty}^{\infty} (x - m_1)(y - m_2) h(x, y) dx dy = \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}} dy \int_{-\infty}^{\infty} (x - m_1)(y - m_2) \cdot \\ &\quad \cdot e^{-\frac{1}{2(1-\rho^2)} \left(\frac{x-m_1}{\sigma_1} - \rho \frac{y-m_2}{\sigma_2} \right)^2} dx \end{aligned}$$

Ha most bevezetjük az

$$u = \frac{1}{\sqrt{1-\rho^2}} \left(\frac{x-m_1}{\sigma_1} - \frac{y-m_2}{\sigma_2} \right), \quad v = \frac{y-m_2}{\sigma_2}$$

helyettesítéseket, akkor

$$\begin{aligned} c &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\sigma_1\sigma_2\sqrt{1-\rho^2}uv + \rho\sigma_1\sigma_2v^2) e^{-\frac{u^2}{2}} du dv = \\ &= \frac{\rho\sigma_1\sigma_2}{2\pi} \int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du + \\ &+ \frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{2\pi} \int_{-\infty}^{\infty} v e^{-\frac{v^2}{2}} dv \int_{-\infty}^{\infty} u e^{-\frac{u^2}{2}} du = \rho\sigma_1\sigma_2 \end{aligned}$$

vagyis

$$\rho = \frac{c}{\sigma_1\sigma_2} = \frac{E[(X - m_1)(Y - m_2)]}{\sigma_1\sigma_2}$$

amely éppen a korrelációs együttható definíciója.

A (2.117) összefüggést szemügyre véve láthatjuk, hogy ha X és Y korrelálatlanok, azaz $\rho = 0$ a

$$\begin{aligned}
 (2.118) \quad h(x, y) &= \frac{1}{2\pi \sigma_1 \sigma_2} e^{-\frac{1}{2} \left[\frac{(x - m_1)^2}{\sigma_1^2} + \frac{(y - m_2)^2}{\sigma_2^2} \right]} = \\
 &= \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x - m_1)^2}{2 \sigma_1^2}} \cdot \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(y - m_2)^2}{2 \sigma_2^2}}
 \end{aligned}$$

Ekkor tehát az X és Y változók együttes sűrűségfüggvénye az egyes változók sűrűségfüggvényeinek szorzata. Ez azonban azt jelenti, hogy X és Y függetlenek. Azt az eredményt kaptuk tehát, hogy ha két valószínűségi változó együttes eloszlása kétdimenziós normális eloszlás és a változók korrelálatlanok, akkor egyben függetlenek is.

2.2.9 A lognormális eloszlás

Az Y valószínűségi változót logaritmikus normális eloszlásúnak vagy röviden lognormális eloszlásúnak nevezzük, ha $\ln Y$ normális eloszlású.

Legyen $X = \ln Y$ és legyen X normális eloszlású valamely m és σ paraméterekkel, vagyis X 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 Y sűrűségfüggvénye:

$$\begin{aligned}
 (2.119) \quad g(x) &= \frac{1}{|x|} f(\ln x) = \frac{1}{\sigma \sqrt{2\pi}} \cdot \frac{1}{x} e^{-\frac{(\ln x - m)^2}{2 \sigma^2}} \\
 (x > 0)
 \end{aligned}$$

Meghatározzuk az Y valószínűségi változó várható értékét és szórásnégyzetét.

$$E(Y) = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{(\ln x - m)^2}{2\sigma^2}} dx$$

Alkalmazzuk az

$$\frac{\ln x - m}{\sigma} = u$$

helyettesítést, akkor

$$x = e^{\sigma u + m}, \quad dx = \sigma e^{\sigma u + m} du$$

Ennek alapján:

$$\begin{aligned} (2.120) \quad E(Y) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\sigma u + m} \cdot e^{-\frac{u^2}{2}} du = \\ &= e^{m + \frac{\sigma^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(u-\sigma)^2}{2}} du = e^{m + \frac{\sigma^2}{2}} \\ E(Y^2) &= \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(\ln x - m)^2}{2\sigma^2}} dx = e^{2m + 2\sigma^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(u-2\sigma)^2}{2}} du = \\ &= e^{2m + 2\sigma^2} \end{aligned}$$

$$\begin{aligned} (2.121) \quad D^2(Y) &= E(Y^2) - [E(Y)]^2 = e^{2m + 2\sigma^2} - e^{2m + \sigma^2} = \\ &= e^{2m + \sigma^2} (e^{\sigma^2} - 1) \end{aligned}$$

Lognormális eloszlás általában aprítási, osztódási, mállási folyamatoknál lép fel, mint a végtermék súlyának, köbtartalmának vagy valamilyen méretének eloszlása.

2.2.10 Az egyenletes eloszlás

Az érmedobás, kockadobás stb. esetében a kísérlethez rendelt X valószínűségi változó mindegyik lehetséges értékét egyforma valószínűséggel veszi fel. Az ilyen valószínűségi változót egyenletes eloszlásúnak nevezzük. Megkülönböztetünk diszkrét és folytonos egyenletes eloszlást. Legyen X diszkrét valószínűségi változó $X_1, X_2, \dots, X_n$ (véges sok) lehetséges értékkel, és legyen

$$(2.122) \quad P(X = x_i) = \frac{1}{n} \quad (i=1, 2, \dots, n)$$

Ekkor az X 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: 1.1.4 pont).

A (2.122) eloszlású valószínűségi változó várható értékét és szórásnégyzetét a következő formulák szolgáltatják:

$$E(X) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

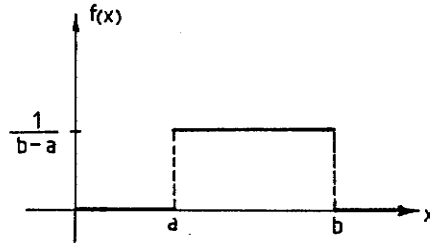
ami nem más, mint a lehetséges értékek számtani közepe.

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

Egy folytonos X valószínűségi változót egyenletes eloszlásúnak nevezünk az $[a, b)$ intervallumon, ha sűrűségfüggvénye a következő:

$$(2.123) \quad f(x) = \begin{cases} \frac{1}{b-a} & \text{ha } a \leq x < b \\ 0, & \text{egyébként} \end{cases}$$

A (2.123) sűrűségfüggvénnyel bíró valószínűségi változó sűrűségfüggvénye: (1.10.3. $f(x; \alpha, p)$)



30. ábra

$$(2.124) \quad F(x) = \begin{cases} 0 & , \text{ ha } x < a \\ \frac{x-a}{b-a} & , \text{ ha } a \leq x < b \\ 1 & , \text{ ha } x \geq b \end{cases}$$

A várható értéket és a szórásnégyzetet egyszerűen kiszámíthatjuk:

$$(2.125) \quad E(X) = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \frac{b^2 - a^2}{2} = \frac{a+b}{2}$$

$$(2.126) \quad D^2(X) = \int_a^b \frac{x^2}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

2.2.11 A gamma-eloszláscsalád: Gamma-eloszlás, exponenciális eloszlás χ^2 -eloszlás

Definíció: Legyen a folytonos eloszlású X valószínűségi változó sűrűségfüggvénye:

$$(2.127) \quad f(x; \alpha, p) = \begin{cases} \frac{\alpha^p}{\Gamma(p)} x^{p-1} e^{-\alpha x}, & \text{ha } x > 0 \\ 0, & \text{egyébként} \end{cases}$$

Ekkor az X valószínűségi változót gamma-eloszlásúnak nevezzük.

Itt a formulában szereplő $\Gamma(p)$ függvény az ún. p-paraméterű gamma függvény, amelynek definíciója

$$(2.128) \quad \Gamma(p) = \int_0^{\infty} x^{p-1} e^{-x} dx$$

Könnyű belátni, hogy $f(x; \alpha, p)$ valóban sűrűségfüggvény, mivel

$$(2.128') \quad \int_0^{\infty} x^{p-1} e^{-\alpha x} dx = \frac{1}{\alpha^p} \int_0^{\infty} (\alpha x)^{p-1} e^{-\alpha x} d(\alpha x) = \frac{\Gamma(p)}{\alpha^p}$$

tehát:

$$\int_0^{\infty} f(x; \alpha, p) dx = 1$$

Az $f(x; \alpha, p)$ sűrűségfüggvénnyel bíró X valószínűségi változó p-paraméterű gamma eloszlásúnak nevezzük.

Számítsuk ki a p-paraméterű gamma eloszlás karakterisztikus függvényét!

$$(2.129) \quad \varphi(t) = \int_0^{\infty} e^{itx} f(x; \alpha, p) dx = \frac{\alpha^p}{\Gamma(p)} \int_0^{\infty} x^{p-1} e^{-(\alpha - it)x} dx = \frac{1}{(1 - \frac{it}{\alpha})^p}$$

Itt felhasználtuk a (2.128') 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.

Definíció: A (2.17) sűrűségfüggvényben $p = 1$ választása mellett

$$(2.130) \quad f(x; \alpha, 1) = \alpha e^{-\alpha x}$$

az ún. exponenciális eloszlás, sűrűségfüggvényét nyerjük.

Az exponenciális eloszlás eloszlásfüggvénye:

$$(2.131) \quad F(x) = 1 - e^{-\alpha x}$$

Az exponenciális eloszlás karakterisztikus függvénye:

$$(2.132) \quad \varphi(t) = (1 - \frac{it}{\alpha})^{-1}$$

A karakterisztikus függvény segítségével kiszámíthatjuk az exponenciális eloszlás várható értékét és szórásnégyzetét a (2.77) és (2.78) formulák alapján.

$$(2.133) \quad E(X) = \frac{1}{\alpha} \quad \varphi'(0) = \frac{1}{\alpha}$$

$$(2.134) \quad D^2(X) = [\varphi'(0)]^2 - \varphi''(0) = \frac{1}{\alpha^2},$$

tehát $D(X) = \frac{1}{\alpha}$

Látható, hogy az exponenciális eloszlás várható értéke és szórása numerikusan megegyezik.

Definíció: Ha a (2.127) sűrűségfüggvényben $\alpha = \frac{1}{2}$, $p = \frac{n}{2}$ értéket választunk, akkor az ún. χ^2 -eloszlás függvényét nyerjük:

$$(2.135) \quad k_n(x) = f(x; \frac{1}{2}, \frac{n}{2}) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{y}{2}-1} e^{-\frac{x}{2}}$$

$$(x > 0)$$

Ez az eloszlás igen jelentős a statisztikai alkalmazásokban.

Tétel:

Független normális eloszlású valószínűségi változók négyzetösszege χ^2 -eloszlású, azaz ha $X_1, X_2, \dots, X_n$ független $N(0;1)$ standard normális eloszlású valószínűségi változók és

$$X = X_1^2 + X_2^2 + \dots + X_n^2$$

akkor az X változó sűrűségfüggvénye a (2.135) formulával megadott függvény.

Bizonyítás:

Ha $X \sim N(0;1)$ normális eloszlású és $Y = X^2$, akkor Y eloszlásfüggvénye:

$$(2.136) \quad G(x) = P(Y < x) = P(X^2 < x) = P(-\sqrt{x} \leq X < \sqrt{x}) = 2 \Phi(\sqrt{x}) - 1$$

Az Y változó sűrűségfüggvénye

$$(2.137) \quad g(x) = 2 \vartheta'(\sqrt{x}) \cdot \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} =$$

$$= \frac{\frac{1}{2}}{\Gamma(\frac{1}{2})} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

Tehát az $Y = X^2$ valószínűségi változó $\alpha = \frac{1}{2}$, $p = \frac{1}{2}$ paraméterekkel bíró gamma eloszlása. Az Y változó karakterisztikus függvénye a (2.129) formula alapján:

$$(2.138) \quad \varphi_Y(t) = \frac{1}{(1 - 2it)^{\frac{1}{2}}}$$

Tekintettel arra, hogy az $X = \sum_{i=1}^n X_i^2$ összegben szereplő

mindegyik X_i^2 ($i=1,2,\dots,n$) valószínűségi változó eloszlása megegyezik az Y változó eloszlásával, az X változó karakterisztikus függvénye

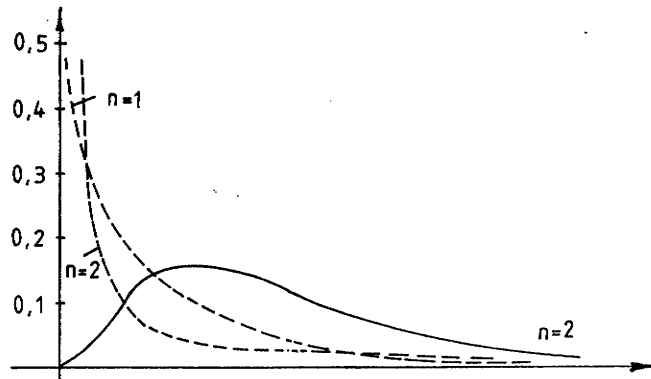
$$(2.139) \quad \varphi_X(t) = \sum_{i=1}^n \varphi_{X_i^2}(t) = \left[\frac{1}{(1 - 2it)^{\frac{1}{2}}} \right]^n =$$

$$= \frac{1}{(1 - 2it)^{\frac{n}{2}}}$$

A (2.139) formulát összehasonlítva a (2.129) függvény-nyel, láthatjuk, hogy az utóbbi függvény az

$$\alpha = \frac{1}{2}, \quad p = \frac{n}{2}$$

paraméterekkel bíró gamma eloszlás karakterisztikus függvénye, vagyis X sűrűségfüggvénye a (2.135) formulával megadott függvény, amelynek alakját $n=1$, $n=2$ és $n=6$ esetében a 31. ábra szemlélteti.



31. ábra

A (2.135) formulával kifejezett $k_n(x)$ sűrűségfüggvényt, amelyet n darab független $N(0;1)$ eloszlású valószínűségi változó négyzetösszegének eloszlásaként nyertünk, szokás n szabadságfokú eloszlásnak is nevezni. A szabadságfok értelmét a 3.5.1 pontban világítjuk meg. A (2.139) karakterisztikus függvény segítségével könnyen származtathatjuk az n -szabadságfokú χ^2 -eloszlás momentumait. A várható érték, illetve a szórásnégyzet kiszámításához elegendő tekintetbe venni, hogy

$$\varphi'_X(0) = -\frac{n}{2} \left[(1 - 2it)^{-\frac{n}{2} - 1} (-2i) \right]_{t=0} = i n$$

$$\begin{aligned} \varphi''_X(0) &= -\frac{n}{2} \left[-\left(\frac{n}{2}\right) + 1 \right] \left[(1 - 2it)^{-\frac{n}{2} - 2} \right]_{t=0} \\ &= i^2 (n^2 + 2n) \end{aligned}$$

Ezek alapján a (2.77) és (2.78) formulák figyelembevételével

$$\begin{aligned} (2.140) \quad E(X) &= n; \quad D^2(X) = -n^2 + n^2 + 2n = 2n; \\ D(X) &= \sqrt{2n} \end{aligned}$$

Ugyancsak a karakterisztikus függvény felhasználásával igazolható a χ^2 -eloszlás ún. additív tulajdonsága, amely abból áll, hogy két független χ^2 -eloszlású valószínűségi

változó összege ugyancsak χ^2 -eloszlású. Ha ugyanis X_1, n_1 szabadságfokú χ^2 -eloszlású és X_1 és X_2 függetlenek, akkor az $X = X_1 + X_2$ összeg karakterisztikus függvénye a megfelelő karakterisztikus függvények szorzata, vagyis a (2.139) formulát figyelembevéve:

$$(2.141) \quad \varphi_X(t) = \varphi_{X_1}(t) \cdot \varphi_{X_2}(t) = (1 - 2it)^{-\frac{n_1}{2}} \cdot (1 - 2it)^{-\frac{n_2}{2}} = (1 - 2it)^{-\frac{n_1+n_2}{2}}$$

vagyis az X 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, 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!

2.2.12 A Student-féle t-eloszlás

A matematikai statisztika néhány fontos problémája az alábbi típusú hányados eloszlásának vizsgálatát követeli meg:

$$(2.142) \quad t = \frac{\sqrt{n} \cdot \bar{X}}{\sqrt{X_1^2 + X_2^2 + \dots + X_n^2}}$$

ahol $X, X_1, X_2, \dots, X_n$ független $N(0;1)$ normális eloszlású valószínűségi változók.

A t valószínűségi változó sűrűségfüggvényének meghatározásához ki kell számítanunk mind a számlálót, mind a nevező sűrűségfüggvényét, majd alkalmaznunk kell a független valószínűségi változók hányadosának sűrűségfüggvényére vonatkozó formulát.

A (2.1.4) pont alapján ismeretes, hogy ha az Y valószínűségi változó egy X valószínűségi változónak folytonos függvénye:

$$Y = \varphi(X)$$

és az X változó sűrűségfüggvénye $f(x)$ akkor Y sűrűségfüggvénye a 2.1.4 pontban mondottak alapján:

$$h(y) = f[\varphi^{-1}(y)] \cdot \frac{d \varphi^{-1}(y)}{dy}$$

Számítsuk ki ennek alapján az (2.142) tört számlálójának sűrűségfüggvényét. Most

$$Y = \sqrt{n} X, \quad \varphi^{-1}(y) = \frac{1}{\sqrt{n}} y, \quad \text{így a } \sqrt{n} X$$

változó sűrűsége:

$$h(x) = \frac{1}{\sqrt{2\pi n}} e^{-\frac{x^2}{2n}}$$

A tört nevezőjének sűrűségfüggvénye az n szabadságfokú χ^2 -eloszlás sűrűségfüggvényének ismeretében könnyen adódik, ugyanis a nevező X ahol X n -szabadságfokú χ^2 -eloszlású, így a nevező sűrűsége:

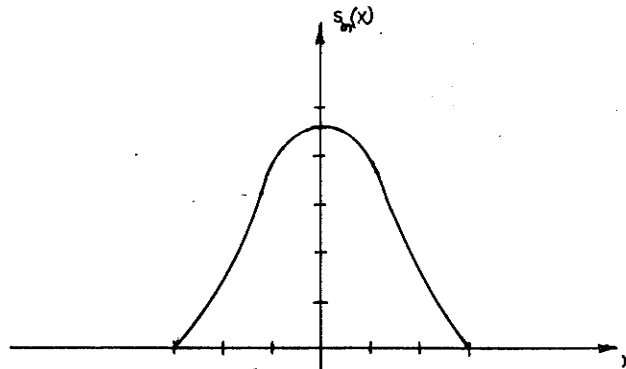
$$g(y) = \frac{2y^{n-1} e^{-\frac{y}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$$

Ha most a t a valószínűségi változó sűrűségfüggvényét $s_n(x)$ -szel jelöljük, akkor

$$s_n(x) = y h(x, y) g(y) dy = \frac{2}{\sqrt{2\pi n} 2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} \int_0^\infty y^n e^{-\left(\frac{x^2}{n} + 1\right)y} dy$$

Alkalmazva az $u = \left(\frac{x^2}{n} + 1\right) \frac{y^2}{2}$ helyettesítést, az

$$\begin{aligned}
 (2.143) \quad s_n(x) &= \frac{\left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}}}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \int_0^\infty u^{\frac{n-1}{2}} e^{-u} du = \\
 &= \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}
 \end{aligned}$$



32. ábra

összefüggésre jutunk, amely a t változó sűrűségfüggvénye, s amely sűrűséggel bíró eloszlást Student-féle t -eloszlásnak nevezzük.

2.3 A nagyszámok törvényei

2.3.1 A nagy számok törvényének Bernulli-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 nagy n ese-

tén a $\frac{k}{n}$ relatív gyakorisága tetszőleges kevéssel tér el a p valószínűségtől, meggyőz bennünket a nagy szám törvényének Bernoullitól származó megfogalmazása.

Alkalmazzuk a Csebisev-egyenlőtlenséget a binomiális eloszlásra.

A (2.55) formulával kifejezett

$$P(|X - E(X)| > \lambda \sigma) \leq \frac{1}{\lambda^2}$$

összefüggésben $X = k$, $E(X) = np$, $\sigma = \sqrt{npq}$ helyettesítéssel:

$$P(|k - np| \geq \lambda \sqrt{npq}) \leq \frac{1}{\lambda^2}$$

A zárójelen belüli egyenlőtlenséget n -nel osztva:

$$P\left(\left|\frac{k}{n} - p\right| \geq \lambda \sqrt{\frac{pq}{n}}\right) \leq \frac{1}{\lambda^2}$$

Vezessük be a $\lambda \sqrt{\frac{pq}{n}} = \varepsilon$ jelölést, ekkor

$$\frac{1}{\lambda^2} = \frac{qp}{n \varepsilon^2} \leq \frac{1}{4n \varepsilon^2}$$

Tehát:

$$(2.145) \quad P\left(\left|\frac{k}{n} - p\right| \geq \varepsilon\right) \leq \frac{1}{4n \varepsilon^2}$$

A (2.145) összefüggés a nagy számok törvényének Bernoulli-féle alakja. A (2.145) egyenlőtlenség 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 a (2.145) összefüggés alapján is válaszolhatnánk, ekkor azonban a szűkegésnél nagyobb n értéket nyernénk. Ennek okát a 2.2.8 pont végén tett megjegyzésünk világítja meg.

Láttuk, hogy a binomiális eloszlás a 2.2.8 pontban említett feltételek között normális eloszlással közelíthető. A 29. ábra alatti táblázatból látható, hogy egy normális eloszlású X valószínűségi változó igen nagy valószínűséggel tér el a várható értéktől, mint szórásának a háromszorosa, azaz

$$P(|X - m| \leq 3\sigma) \approx 0,997$$

Tehát közelítőleg ugyancsak 0,997 annak a valószínűsége, hogy az A esemény k gyakorisága az np várható értéktől kevesebbel térjen el, mint szórásának a háromszorosa, tehát

$$(2.146) \quad P(|k - np| \leq 3\sqrt{npq}) \approx 0,997$$

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.

A (2.146) képletben szabad az egyenlőtlenség mindkét oldalát n -nel osztani, anélkül, hogy a valószínűség megváltozzék, azaz

$$(2.147) \quad P\left(\left|\frac{k}{n} - p\right| \leq 3\sqrt{\frac{pq}{n}}\right) \approx 0,997$$

Ez azt mutatja, hogy igen nagy valószínűséggel a 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

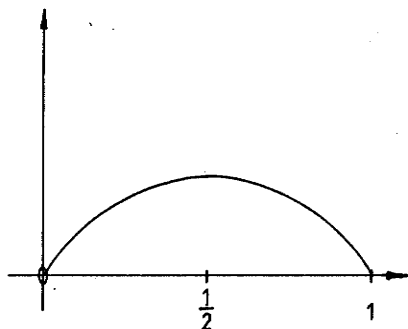
$$(2.148) \quad \left|\frac{k}{n} - p\right| \leq \frac{3}{2\sqrt{n}} = \frac{1,5}{\sqrt{n}}$$

azaz csaknem biztos, hogy a p valószínűség a

$$\frac{k}{n} - \frac{1,5}{\sqrt{n}} \quad \text{és} \quad \frac{k}{n} + \frac{1,5}{\sqrt{n}}$$

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 $\frac{1}{4}$, (1). tehát a relatív gyakoriság szórása $\sqrt{\frac{pq}{n}} \leq \frac{1}{2\sqrt{n}}$



33. ábra.

Ilyen esetben a p valószínűség helyett a $\frac{k}{n}$ relatív gyakoriság értékét írjuk és $q = (1-p)$ helyébe $1 - \frac{k}{n} = \frac{n-k}{n}$ -et írunk, így

$$\sqrt{\frac{pq}{n}} \approx \frac{1}{n} \sqrt{\frac{k(n-k)}{n}}$$

Még jobb becslést kapunk a szórásra, ha a gyök alatti kifejezés nevezőjében n helyett $n-1$ -et írunk.

Igen nagy valószínűséggel, gyakorlatilag úgyszólván biztosan igaz tehát, hogy

$$\left| \frac{k}{n} - p \right| \leq \frac{3}{n} \sqrt{\frac{k(n-k)}{n-1}}$$

Ha arra akarunk válaszolni, mekkorára kell n -et, a kísérletek számát választani, hogy a $\frac{k}{n}$ relatív gyakoriság

- (1). Ugyanis az $f(p) = p(1-p)$ függvény maximumát csak ott veheti fel, ahol $f'(p) = 1 - 2p = 0$ vagyis a $p = \frac{1}{2}$ helyen. Mivel $f''(p) = -2$, valóban maximumról van szó.

az ismeretlen p valószínűségtől csaknem biztosan kevesebb térjen el, mint pl. $\varepsilon = 0,1$ akkor a (2.148) összefüggés alapján csak az

$$\frac{1,5}{\sqrt{n}} = 0,1$$

egyenletet kell megoldnunk, 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, a (2.148)

$$n = 900$$

adódik, míg $\varepsilon = 0,01$ esetén (ez már igen nagy pontosság, amire a gyakorlatban ritkán van szükség) a szükséges kísérletek száma:

$$n = 22\,500$$

2.3.2 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 X valószínűségi változó véletlen ingadozása sok egymástól független véletlen komponens eredménye, amely komponensek külön-külön is mértékben hatnak, de hatásuk összegeződik, akkor x normális eloszlású.

A binomiális eloszlás vizsgálatánál láttuk, hogy a binomiális eloszlású X valószínűségi változó n darab független karakterisztikus változó összege:

$$X = X_1 + X_2 + \dots + X_n$$

ahol az egyes X_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 változók nem olyan természetűek, hogy csaknem mindig egyik értéküket veszik fel. (Úgy is mondhatjuk, hogy ha az X_i változók eloszlása nem túlságosan elfajult).

A binomiális eloszlásnak normális eloszlással való közelíthetősége, csak speciális esete, bár nagyon fontos speciális eset, egy sokkal általánosabb törvényszerűségnek,

amelyet úgy fogalmazhatunk meg, hogy ha igen nagyszámú 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 $X_1, X_2, \dots, X_n$ azonos eloszlású független és véges szórású valószínűségi változók,

$$E(X_k) = m, D^2(X_k) = \sigma^2 \quad (k=1, 2, \dots, n), \text{ akkor}$$

$$(2.149) \quad \lim_n P\left(\frac{X_1 + X_2 + \dots + X_n - nm}{\sigma \sqrt{n}} < x\right) =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$

A tétel bizonyítása karakterisztikus függvény alkalmazásával végezhető el legegyszerűbben.

Jelöljük $\varphi(t)$ -vel az X_k -m valószínűségi változó karakterisztikus függvényét. Ekkor az $\frac{X_k - m}{\sigma \sqrt{n}}$ változó karakterisztikus függvénye:

$$\varphi\left(\frac{t}{\sigma \sqrt{n}}\right)$$

Legyen:

$$Y = \frac{X_1 + X_2 + \dots + X_n - nm}{\sigma \sqrt{n}} = \sum_{k=1}^n \frac{X_k - m}{\sigma \sqrt{n}}$$

A komponensek függetlensége miatt az összeg karakterisztikus függvénye az összeadandók karakterisztikus függvényeinek szorzata, tehát:

$$\varphi_Y(t) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right) \right]^n$$

Vezessük be a $\frac{t}{\sigma\sqrt{n}} = u$ jelölést és fejtsük ki 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 az (2.80) formula alapján a momentumok segítségével:

$$\begin{aligned} \varphi(u) = 1 + \alpha_1 iu + \alpha_2 \frac{(iu)^2}{2!} + \dots + K! \\ + \alpha_k \frac{(iu)^k}{k!} + o(u^k) \end{aligned}$$

ahol α_j az $\frac{X_{k-m}}{\sigma\sqrt{n}}$ változó j -edik momentuma.

$$\alpha_1 = E\left(\frac{X_k - m}{\sigma\sqrt{n}}\right) = 0$$

$$\alpha_2 = E\left(\frac{X_k - m}{\sigma\sqrt{n}}\right)^2 = \frac{1}{n\sigma^2} E(X_k - m)^2 = \frac{1}{n}$$

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:

$$\begin{aligned} (2.150) \quad \lim_{n \rightarrow \infty} \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right) \right]^n &= \lim_{n \rightarrow \infty} \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \right]^n = \\ &= e^{-\frac{t^2}{2}} \end{aligned}$$

Mint a normális eloszlás tárgyalásánál láttuk, ez a függvény 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 $X_1, X_2, \dots, X_n$ független valószínűségi változók, amelyek harmadik és következőleg alacsonyabb rendű momentumai léteznek, $E(X_k) = 0$, $(k=1, 2, \dots, n)$ továbbá

$$\lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \beta_k}{S_n^3} = 0$$

ahol

$$\beta_k = E(|X_k|^3), \quad S_n = \sqrt{\sum_{k=1}^n \sigma_k^2}, \quad \sigma_k^2 = D^2(X_k),$$

akkor

$$(2.151) \quad \lim_{n \rightarrow \infty} P\left(\frac{X_1 + X_2 + \dots + X_n}{S_n} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$$

A centrális határeloszlástétel igen nagy gyakorlati jelentőségű. Egyrészt, ha a tétel feltételei teljesülnek legalábbis jó közelítéssel, akkor jogossá teszi azt a feltételezést, hogy normális eloszlással állunk szemben, amelyet a megfigyelési adatok alapján a matematikai statisztika módszereivel ellenőrizni tudunk. Másrészt a matematikai statisztikában szereplő ún. statisztikai függvényekben az esetek nagy részében független valószínűségi változók összege szerepel és nagyszámú 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 meghatározatalában alapvető segédeszközünk lehet.

3. Fejezet

3.1 A matematikai statisztika elemei

3.1.1 A matematikai statisztika tárgya és módszerei

A matematikai statisztika a valószínűségelmélet önálló fejezete, amely a gyakorlat számára igen nagy jelentőségű.

A matematikai statisztika tárgya a véletlen tömegjelenségekben rejlő statisztikai törvényszerűségek vizsgálata. Problémái, és ennek következtében módszerei sajátosak, a valószínűségelmélet más fejezeteihez képest.

A valószínűségelméletben ismertnek tételezzük fel valamely valószínűségi változó eloszlásfüggvényét és annak paramétereit, ezek birtokában válaszoltunk valószínűségi jellegű kérdésekre.

A valószínűségelméleti részben a feladatokat pl. így fogalmaztuk: Legyen X normális eloszlású ismert m várható értékkel és ismert σ szórással. Számítsuk ki annak valószínűségét, hogy X megfigyel. értéke adott $[a, b)$ intervallumba esik! A válasz a következő volt:

$$P(a \leq X < b) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{(x-m)^2}{2\sigma^2}} dx .$$

Konkrét m , σ , a és b értékek esetén ezt a valószínűséget táblázat segítségével pontosan meg tudtuk mondani. A gyakorlatban (és a matematikai statisztikában) azonban arra a kérdésre keressük a választ, hogy pl. bizonyos típusú házigyári panel esetében mi a valószínűsége, hogy a hosszabbik oldal hosszát jelölő X valószínűségi változó értéke a $[399, 401)$ intervallumba esik. Tapasztalatból tudjuk, hogy a panel hossza

- X - normális eloszlású valószínűségi változó. Nem tudjuk azonban, hogy az adott gyártási technológia alkalmazásakor milyen m várható értékű és σ szórással normális eloszlású valószínűségi változóval számolhatunk. A paraméterek ismerete nélkül a kérdéses valószínűséget nem tudjuk kiszámítani. Ahhoz, hogy ezeket a paramétereket meghatározhassuk, meg kell

mérnünk az adott típusú panelből minél több darabnak a hosszabik oldalát. Legyenek a mérési eredmények $x_1, x_2, \dots, x_n$.

A matematikai statisztika foglalkozik azzal a kérdéssel, hogy a kapott mérési eredményekből hogyan kaphatjuk meg pl. a várható értéket. Látni fogjuk, hogy ha kiszámítjuk a mérési eredmények számtani közepét, az

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

mennyiséget, akkor az - elég nagy n esetén - "jó" közelíti az ismeretlen m várható értéket (lásd: 8.3.1. pont). (A gyakorlatban a n mindig véges szám).

Ha azonban további n panel hosszát megmérnénk, és ismét kiszámítanánk a számtani közepet, akkor az más számszerű eredményre vezetne. Ez azt mutatja, hogy a várható érték pontos meghatározása helyett meg kell elégednünk annak közelítő meghatározásával, becslésével.

Az említett becslési eljárás az m várható érték számára egyetlen számmértékhez $\bar{x}$ -hez vezet. Ilyen esetben a paraméter a pontbecsléséről beszélünk. Mivel a becsléssel kapott érték általában nem azonos a keresett elméleti értékkel, ezért szokás, hogy a műszaki biztonság érdekében az ismeretlen paraméter számára alsó és felső határt állapítsunk meg, tehát intervallumbecslést adjunk. Olyan intervallumot kell meghatározunk, amely nagy biztonsággal tartalmazza a szóbanforgó elméleti paraméter ismeretlen értékét.

Valamely elméleti érték becslésének, a becslés módszerének, megbízhatóságának (általában tulajdonságainak) vizsgálatával a matematikai statisztika becsléelmélet című fejezete foglalkozik, melyet a 10. pontban részletesebben ismertetünk.

A műszaki gyakorlatban igen sokszor nem ismerjük a szóbanforgó valószínűségi változó eloszlását, sőt annak típusát sem. Pl. az árvizekkel kapcsolatos valószínűségi kérdések megválaszolásához ismernünk kell valamely folyó, mondjuk a Tisza adott helyen, pl. Szegednél mért évi maximális vízállás értékeinek valószínűségi változó jelöli az évi maximális vízállását, akkor eloszlásfüggvényét, a $V \in R$ esetén $P(X < x) = F(x)$ eloszlásfüggvényt kell meghatározunk (becsülnünk) az eddig feljegyzett évi maximális vízállásértékek alapján.

A műszaki gyakorlatban sokszor előfordul a következő típusú feladat is: valamely kísérlettel, gyártási folyamattal kapcsolatban ismerjük az azt jellemző valószínűségi változó eloszlásfüggvényét. Bizonyos idő eltelte után a körülmények, gyártási feltételek megváltoznak. Ilyenkor mérési eredményekből ellenőriznünk kell, hogy megváltozott-e vagy sem a kérdéses valószínűségi változó eloszlása. A folyamatos gyártásközi ellenőrzésben rendszeresen vizsgáljuk, hogy pl. megnövekedett-e valamely gyártmány mérete, aszfaltkeverékek esetében megváltozott-e a keverék összetétele valamely komponens szempontjából stb. A hidrológiában érdekelhet bennünket, hogy valamely tározó megépítése befolyásolja-e a folyó vízjárását és így tovább. Ilyen esetekben korábbi ismereteink alapján a vizsgált valószínűségi változó eloszlását illetőleg feltevessel élünk, statisztikai hipotézist állítunk fel, és a matematikai statisztika módszereivel döntünk, hogy elfogadjuk vagy elvetjük hipotézisünket. Valamely hipotézis helyességének eldöntésével a matematikai statisztika hipotézis vizsgálatáról szóló fejezete foglalkozik (lásd: 11. fejezet). A hipotézis vizsgálat a matematikai statisztika másik nagy problémaköre, amelynek alapjául ugyancsak megfigyelési (mérési) adatok szolgálnak, miként a becsléelméletnek.

Minthogy a gyakorlatban mindig véges számú minta alapján számolunk, a helyes döntés 100 %-os biztonsággal nem tudjuk megadni. A matematikai statisztika megállapítja a hibás döntés meghozatalának valószínűségét, és olyan módszereket szolgáltat, amelyek alkalmazása esetén hibás döntés viszonylag ritkán fordul elő.

A becsléelmélet és a statisztikai hipotézis vizsgálat a matematikai statisztika két alapvető - immár klasszikus - fejezete, amelyeket a műszaki alkalmazások szempontjából is elsődleges jelentőségűeknek tekintünk.

A mondottak alapján a matematikai statisztikai alapvető feladatát nagy általánosságban így fogalmazhatjuk meg: Következtetés tapasztalatai (megfigyelési, mérési) adatokból események ismeretlen valószínűségeire, valószínűségi változó ismeretlen eloszlásfüggvényére vagy azok paramétereire. A matematikai statisztika ezen alapvető feladatát a statisztikusok olyan módszerek kidolgozásával igyekeznek megoldani, amelyek segítségével a megfigyelési adatokból a keresett elméleti értékekre, ill. a felvetett problémákra a lehető legtöbb információt nyerjük.

A matematikai statisztika fenti alapvető feladatának megjelölése egyben rámutat a matematikai statisztika nagy gyakorlati jelentőségére, hiszen a valószínűségelmélet tételeit, eredményeit csak akkor alkalmazhatjuk, ha a valószínűségeloszlások kielégítően meghatározottak.

3.1.2 A statisztikai minta fogalma

A műszaki gyakorlatban valamely kísérlet elvégzésekor egy vagy néhány jellemző (valószínűségi változó) értékét mérjük. Ha pl. házigyári panelt vizsgálunk, akkor annak geometriai méretei, törési szilárdsága stb., lehet érdekes számunkra a felhasználás szempontjából. Árvíz elleni védekezésnél az árhullám tetőzési értéke, a vízhozam alakulása, az árvíz időtartama stb., érdekel bennünket. Tekintsük először azt az esetet, amikor egyetlen valószínűségi változó (pl. egy panel hosszabbik oldalának hossza) értékeire irányul a figyelmünk. Ha n panelt megmérünk, akkor minden egyes panel hossza véletlentől függő mennyiség, valószínűségi változó. Ha a megmérendő paneleket találomra választjuk, akkor a mérési eredményeket általában függetleneknek tekinthetjük. A statisztika elmélete szempontjából n független kísérlet kimeneteléről beszélünk.

Definíció:

A statisztikai minta valamely valószínűségi változóra vonatkozó véges számú független kísérlet vagy megfigyelés (mérés) eredménye: véges sok, azonos eloszlású független valószínűségi változó együttese. Az egyes megfigyelési eredményeket a minta elemeinek, a megfigyelések számát a minta nagyságának vagy elemszámának nevezzük.

Jelöljük X -szel a megfigyelandő valószínűségi változót, ekkor az X -re vonatkozó n elemű minta:

$$X_1, X_2, \dots, X_n.$$

A mintaelemek sorrendje az észlelések sorrendjének, véletlen sorrendnek felel meg. Az n számú kísérlet tényleges elvégzése során mindegyik X_i mintaelem egy konkrét számértéket vesz fel,

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n.$$

Az $(x_1, x_2, \dots, x_n)$ értékhalmozat vagy más néven szám- n -est a továbbiakban ugyancsak statisztikai mintának fogjuk nevezni, ami a jelölés különbözősége miatt nem fog félreértésre vezetni.

Ha több elemű mintát veszünk X -re vonatkozólag, akkor mind a X_1 első mintaelem, mind a X_2 második mintaelem, ..., mind a X_n -edik mintaelem általában mindegyik mintában különböző értéket vesznek fel, így a $X_1, X_2, \dots, X_n$ mintaelemek maguk is valószínűségi változók.

A definícióban két alapvető tulajdonságot kötöttünk ki a mintára vonatkozólag:

a) A mintaelemek eloszlása megegyezik a vizsgált X valószínűségi változó eloszlásával ← hiszen mindegyik kísérletnél X értékét figyeljük meg - vagyis

$$Vx \in R \text{ esetén } P(X_i < x) = F(x) \quad (i=1, 2, \dots, n).$$

A mintának ezt a tulajdonságát a minta reprezentativitásának nevezzük, azaz mindegyik mintaelem "reprezentálja" a vizsgálat valószínűségi változót.

b) A mintaelemeket összességükben független valószínűségi változóknak tekintjük. A kísérleteket ennek érdekében egymástól függetlenül végezzük. (ha pl. acélrudak törési szilárdságát vizsgáljuk, akkor az azonos technológiával gyártott nagyszámú rúd közül pl. sorsolással választunk ki n darabot.)

A függetlenség feltétele bizonyos esetekben csak közelítőleg teljesül. Pl. valamely folyó évi középvízállásainak adatsora jó közelítéssel független sorozatnak, tehát a folyó középvízállására vonatkozólag statisztikai mintának tekinthető. A havi középvízhozamok sorozatára ez már általában nem érvényes, még kevésbé a napi középvízhozamok sorozatára, hiszen egyik napi vízhozam befolyásolja a következő napi vízhozamot.

Nézzünk példákat a statisztikai mintára:

1. Példa. Jelölje az X valószínűségi változó a Duna évi maximális vízállását Budapestnél. Ekkor az X -re vonatkozó n elemű minta: $X_1, X_2, \dots, X_n$. E minta konkrét értékeit a vízrajzi évkönyvből kikereshetjük, pl. a legutóbbi n évre vonatkozólag. Ekkor egy $(x_1, x_2, \dots, x_n)$ szám-n-est kapunk. Ha a másik n évet választunk, akkor általában más szám-n-est lesz a minta konkrét értéke.

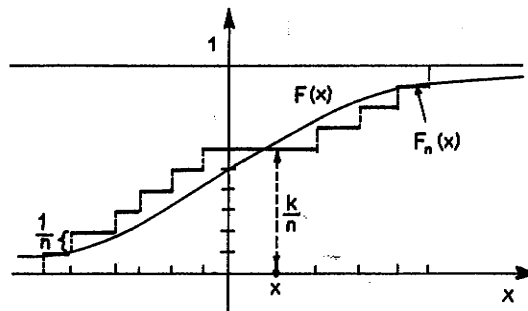
2. Példa. Aszfaltút építéséhez keverő telepen állítják elő az aszfaltkeveréket. Minden keverékből naponta néhányszor adott mennyiségű keverékadagot vesznek ki, és laboratórium-ban meghatározzák az aszfaltkeverék különböző jellemzőinek numerikus értékét. Jelölje az Y valószínűségi változó az egy keverék adagban levő bitumen tartalmát. Ekkor n adag mindegyikében levő bitumen tartalom rendre $Y_1, Y_2, \dots, Y_n$, statisztikai minta. A mintaelemek megfigyelt értékei egy $Y_1, Y_2, \dots, Y_n$ szám-n-est adnak. Más n adagban mért bituméntartalom általában más szám-n-eshez vezet. A kapott statisztikai mintából következtetünk az Y valószínűségi változó eloszlására, illetve az eloszlás paramétereire.

3.1.3 A tapasztalati (empirikus) eloszlásfüggvény

Mint már az előző pontban vázoltuk, a matematikai statisztika feladata, hogy a statisztikai mintából - tapasztalati adatokból - a vizsgált valószínűségi változó eloszlására és az eloszlás jellemzőire (paramétereire) következtessen.

Legyen a vizsgált X valószínűségi változó ismeretlen eloszlásfüggvény $V_x \in R$ esetén $P(X < x) = F(x)$. Az X valószínűségi változóra vonatkozó statisztikai minta $X_1, X_2, \dots, X_n$ elemeiben elrejtve benne van az eloszlásra és annak paramétereire vonatkozó információ. Abból a célból, hogy a mintából ezt kiolvashassuk, az információt célirányosan kell tömörítenünk a feltett statisztikai kérdések szempontjából.

Hogyan olvashatjuk ki például a statisztikai mintából valamely X valószínűségi változó $F(x)$ eloszlásfüggvényére vonatkozó információt? Mint tudjuk, az $F(x)$ eloszlásfüggvény értéke adott x helyen megadja az $\{X < x\}$ esemény valószínűségét. Egy-egy mintaelem értéke az $\{X < x\}$ esemény valószínűségére vonatkozólag nem sokat mond. Tekintsük az $X_1, X_2, \dots, X_n$ mintaelem megfigyelt értékeinek összességét. Tetszőleges x értékre megkapjuk az $\{X < x\}$ esemény relatív gyakoriságát, ha az x -nél kisebb mintaelemek számát elosztjuk n -nel, a minta nagyságával. Ez a relatív gyakoriság az $\{X < x\}$ esemény valószínűségét, tehát $F(x)$ értékét közelíti adott x helyen.



34. ábra

Az $\{X < x\}$ esemény relatív gyakoriságát közvetlenül szemléltetjük azáltal, hogy a mintaelemeket a számegyenesen ábrázoljuk, majd egy lépcsős függvényt rajzolunk, amelynek $\frac{1}{n}$

nagyságú ugrása van minden egyes mintaelemnél. (34. ábra). Jelöljük ezt a lépcsős függvényt $F_n(x)$ -szel, és nevezzük el az X valószínűségi változó tapasztalati (empirikus) eloszlásfüggvényének.

Definíció:

Az $F_n(x)$ empirikus eloszlásfüggvény valamely x helyen az x -nél kisebb megfigyelések számának relatív gyakorisága:

$$F_n(x) = \frac{1}{n} \sum_{x_i < x} 1.$$

Mivel a nagy számok törvénye szerint elég nagy minta esetén az $\{X < x\}$ esemény relatív gyakorisága $Vx \in R$ -re ezen esemény valószínűségét közelíti, ezért

$$P(|F_n(x) - F(x)| \geq \varepsilon) \leq \frac{1}{4n\varepsilon}.$$

Az $F_n(x)$ empirikus eloszlásfüggvényt a minta eloszlásfüggvényének is szokás nevezni. A minta eloszlásfüggvénye a minta elemszámának növelésével minden x -re egyenletesen konvergál az $F(x)$ elméleti eloszlásfüggvényhez. Ezt a tényt fejezi ki szabatosan V. J. Glivenkó tétele:

Tétel:

Legyen $X_1, X_2, \dots, X_n$ az X valószínűségi változóra vonatkozó n elemű statisztikai minta, és legyen eloszlásfüggvénye $F(x)$. Legyen $F_n(x)$ a minta empirikus eloszlásfüggvénye, továbbá legyen Δ_n az elméleti és az empirikus eloszlásfüggvények közötti abszolút eltérés maximuma, azaz

$$\Delta_n = \max_x |F_n(x) - F(x)|,$$

akkor Δ_n 1-valószínűséggel egyenletesen 0-hoz konvergál.

A tétel bizonyításával itt nem foglalkozunk, rámutatunk azonban a tétel nagy gyakorlati jelentőségére, amely-nél fogva Glivenkó tételét a matematikai statisztika alap-tételének nevezzük.

Az a tény, hogy az $F_n(x)$ tapasztalati eloszlásfüggvény és az $F(x)$ elméleti eloszlásfüggvény közötti abszolút eltérés maximuma szinte biztosan egyenletesen 0-hoz konvergál, azt jelenti, hogy elég nagy minta esetén $F_n(x)$ értéke minden x -re tetszőlegesen közel van $F(x)$ értékéhez, és n -et növelve

mindenütt annak közelében marad. Amikor valamely valószínűségi változó eloszlására következtetünk mintavétel útján, akkor eljárásunk jogosultsága Glivenkó tételén alapszik.

3.1.4 A gyakorisági és sűrűség-hisztogram

Folytonos valószínűségeloszlások esetében az eloszlásról szemléletesebb képet kapunk a sűrűségfüggvény segítségével, mint az eloszlásfüggvény alapján. A sűrűségfüggvénynek tapasztalati adatokkal való közelítése céljából az $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ megfigyelt mintaelemeket felmérjük a számegyenesre. Az ábrázolás során kapott értékek valamely $[a, b]$ intervallumban helyezkednek el. Osszuk fel az $[a, b]$ intervallumot r része az

$$a = d_0 < d_1 \dots < d_r = b$$

osztópontok segítségével.

A felosztás módjára nincs általános szabály. A részintervallumok r száma a mintanagyságtól függ általában 6-12 részintervallumot képezünk; igen nagy mintánál ennél több is lehet. A részintervallumok hosszát választhatjuk egyenlőnek, esetleg a szélső intervallumokat, ahová kevesebb mintaelem esik, hosszabbnak választjuk.

Adjuk meg az egyes $\Delta_i = [d_{i-1}, d_i)$ részintervallumba eső mintaelemek k_i számát ($i=1, 2, \dots, r$), és mindegyik részintervallumra rajzoljunk az oda eső mintaelemek gyakoriságával arányos magasságú téglalapot. Az i -edik részintervallumra (jelöljük ennek hosszát Δ_i -vel) rajzolt téglalap magassága legyen.

$$\frac{k_i}{d_i - d_{i-1}} = \frac{k_i}{\Delta_i}$$

Ekkor a téglalapok területeinek összege:

$$\sum_{i=1}^r \frac{k_i}{d_i - d_{i-1}} (d_i - d_{i-1}) = \sum_{i=1}^r k_i = n.$$

Az így kapott ábra a gyakorisági hisztogram. (35. ábra).

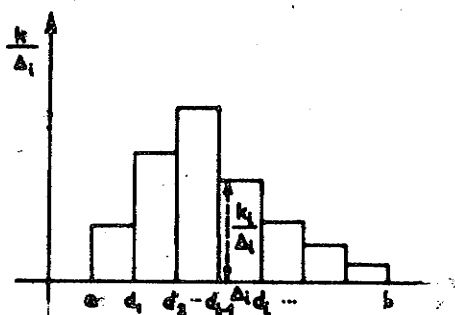
Ha az egyes részintervallumokra rajzolt téglalapok magasságát az oda eső mintaelemek relatív gyakoriságai segítségével számítjuk ki, akkor az i -edik részintervallumra rajzolt téglalap magassága legyen

$$\frac{k_i}{n(d_i - d_{i-1})} \quad (i=1, 2, \dots, r).$$

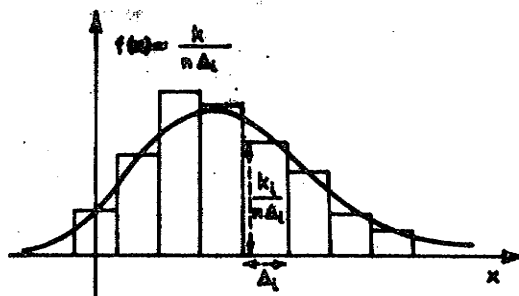
Ez esetben a téglalapok területeinek összege 1 lesz:

$$\sum_{i=1}^r \frac{k_i}{n(d_i - d_{i-1})} (d_i - d_{i-1}) = \sum_{i=1}^r \frac{k_i}{n} = 1.$$

Az így kapott ábra a sűrűséghisztogram, lásd 36. ábra.



35. ábra



36. ábra

• Kimutatható, hogy elég nagy n és jól választott részintervallumok esetén (úgy választunk, hogy mindegyik részintervallumba kb. $\sqrt[3]{n}$ számú mintaelem jusson) az $[a, b)$ inter-

vallum tetszőleges $[a', b')$ részintervallumára a sűrűség-hisztogram alatti terület közelítőleg a sűrűségfüggvény alatti területtel egyezik meg; azaz a közelítőleg a $P(a' \leq X < b')$ valószínűség értékét adja.

Megjegyezzük, hogy a sűrűség-hisztogramot az empirikus sűrűségfüggvény grafikonjának is szokás nevezni. Ennek alapján a sűrűség-hisztogramot $f_n(x)$ -szel jelöljük. Az i -edik részintervallumba eső minden x értékre

$$f_n(x) = \frac{\frac{k_i}{n}}{d_i - d_{i-1}} = \frac{\Delta F_n(x)}{\Delta_i},$$

$$d_{i-1} \leq x < d_i \quad (i=1, 2, \dots, r).$$

ahol $\Delta F_n(x)$ az $F_n(x)$ tapasztalati (empirikus) eloszlásfüggvény növekménye a $[d_{i-1}, d_i)$ részintervallumon. A két függvény közti kapcsolat analóg az elméleti sűrűségfüggvényre ismert $f(x) = F'(x)$ relációval, csak itt differenciahányadost kell venni.

Ha csak szemléletes képet akarunk nyerni az eloszlásról, akkor a gyakorisági hisztogramot használjuk, mivel a sűrűségfüggvény alakját ez is szemlélteti.

Példa:

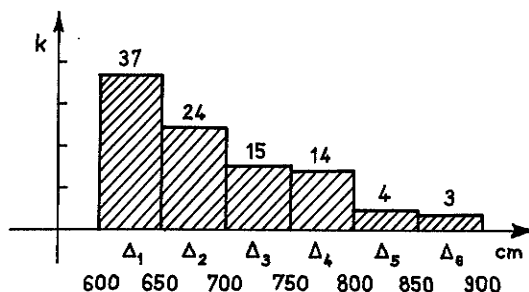
A Tisza folyó árhullámaira vonatkozólag Tokajnál az 1903-1971 időszakban a tetőzési értékek az alábbi gyakoriságokkal estek a feltüntetett intervallumokba (mindegyik év első félévében bekövetkezett árvizeket vizsgálva):

Δ_1	= 600 - 650 cm	37 árhullám
Δ_2	= 650 - 700 cm	24 árhullám
Δ_3	= 700 - 750 cm	15 árhullám
Δ_4	= 750 - 800 cm	14 árhullám
Δ_5	= 800 - 850 cm	4 árhullám
Δ_6	= 850 - 900 cm	3 árhullám

Összesen: 97 árhullám

Készítsük el a tetőzési értékek gyakorisági hisztogramját!
Megoldás. Minthogy egyenletes feloszlást használunk, az

egyes részintervallumok hossza $\Delta_i = d_i - d_{i-1} = 0,5$ m.
A gyakorisági hisztogram a 37. ábrán látható.



37. ábra

A gyakorisági hisztogram alatti terület most:

$$\sum_{i=1}^6 \frac{k_i}{d_i - d_{i-1}} (d_i - d_{i-1}) = (74 + 48 + 30 + 28 + 8 + 6) \cdot \frac{1}{2} = 97.$$

azaz a terület a mintaelemek számával egyenlő. Ha a gyakorisági hisztogramban minden ordinátát 97-tel osztunk, a sűrűséghisztogramot kapjuk.

3.1.5 A statisztikai függvény fogalma. A legfontosabb tapasztalati jellemzők

A valószínűségszámítás című részben láttuk, hogy a valószínűségi változó eloszlását néhány számadattal, a várható értékkel, mediánnal, szórással és más jellemzőkkel a gyakorlat számára sokszor kielégítően tudjuk jellemezni. A várható érték az eloszlás "súlypontjáról", a szórás a változó értékeinek szétszórtságáról ad felvilágosítást. Ezekre az elméleti jellemzőkre is a mintaelemekből igyekszünk következtetni a matematikai statisztika módszereivel oly módon, hogy az $X_1, X_2, \dots, X_n$ mintán különböző (n változós) függvényeket értelmezünk. Valamely $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ n -változós függvény egyetlen számadatba tömöríti a mintaelemekben rejlő információt a kérdéses elméleti jellemzőkre vonatkozólag. A kérdés tehát az, hogy milyen függvényelírás szerint számoljunk a mintaelemekkel ahhoz, hogy jó közelítéssel megkapjuk az eloszlás várható értékét, és milyen függvény szolgáltatja a szórásnégyzet vagy más elméleti jellemző közelítő értékét. Ezekre a kérdésekre válaszolunk a következő pontban, a függvények közelítésének jóságát a 3.5.2 pontban vizsgáljuk.

Definíció:

A mintaelemek valamely $\alpha_n(X_1, X_2, \dots, X_n)$ függvényét statisztikai függvénynek, vagy röviden statisztikának nevezzük. Minthogy a mintaelemek valószínűségi változók, ezek függvényei - a statisztikák - is valószínűségi változók. Így beszélhetünk ezek eloszlásáról, várható értékéről, szórásáról, stb. Ezek ismeretére szükségünk lesz ahhoz, hogy a statisztikát a gyakorlatban fel tudjuk használni.

Az alábbiakban áttekintést adunk a leggyakrabban használatos statisztikákról, amelyek egy ismeretlen, a vizsgálat tárgyát képező eloszlás legfontosabb elméleti jellemzőinek statisztikai közelítései.

Definíció:

Az $\bar{X}$ valószínűségi változóra vonatkozó $X_1, X_2, \dots, X_n$ statisztikai minta elemeinek számtani közepét, a

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

statisztikát mintaközépnek nevezzük.

A mintaközép az elméleti várható érték közelítésére szolgál, amint az alábbi tételből látható.

Tétel:

Ha az X valószínűségi változó várható értéke $M(X) = m$, szórásnégyzete $D^2(X) = \sigma^2$, akkor az X -re vonatkozó $X_1, X_2, \dots, X_n$ statisztikai mintából számított $\bar{X}$ mintaközép várható értéke:

$$M(\bar{X}) = M(X) = m$$

szórásnégyzete:

$$D^2(\bar{X}) = \frac{1}{n} D^2(X) = \frac{\sigma^2}{n}$$

szórása:

$$D(\bar{X}) = \frac{\sigma}{n}$$

Bizonyítás.

A várható érték és a szórásnégyzet ismert tulajdonságait felhasználva (lásd: 2.1.7 pont)

$$M(\bar{X}) = \frac{1}{n} [M(X_1) + M(X_2) + \dots + M(X_n)] = \frac{1}{n} \cdot nM(X) = m,$$

$$\begin{aligned}
 D^2(\bar{X}) &= \frac{1}{n^2} [D^2(X_1) + D^2(X_2) + \dots + D^2(X_n)] = \\
 &= \frac{1}{n^2} \cdot nD^2(X) = \frac{\sigma^2}{n}, \\
 D(\bar{X}) &= \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.
 \end{aligned}$$

E formulák mutatják, hogy az $\bar{X}$ mintaközép az m várható érték körül ingadozik, és szórása n növekedésével zérushoz konvergál.

A mintaközépnek az a tulajdonsága, hogy nagy n esetén kicsi a szórása, a gyakorlatban rendkívül jelentős. Ha pl. az előzőekben példaként említett házigyári panel hosszabbik oldalát jelentő X valószínűségi változó normális eloszlású $M(X) = 300$ cm) várható értékkel és $D(X) = 1$ (cm) szórással, akkor 100 találomra választott panel hosszát megmérve és az adatok számtani közepét képezve, az $\bar{X}$ mintaközép szórása: $D(\bar{X}) = 0.1$ (cm).

Megjegyezzük, hogy valamely A esemény relatív gyakorisága a mintaközép speciális esete. Ha ugyanis n független kísérletet végzünk, és X_i az A esemény indikátorváltozója, vagyis,

$$X_i = \begin{cases} 1, & \text{ha } A \text{ következik be} \\ 0, & \text{ha } \bar{A} \text{ következik be} \end{cases} \quad (i=1, 2, \dots, n).$$

akkor

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{k}{n},$$

ahol k az A esemény bekövetkezéseinek száma. Az

$$X = X_1 + X_2 + \dots + X_n = k$$

valószínűségi változó binomiális eloszlású $M(k) = np$ várható értékkel és $D^2(k) = npq$ szórásnégyzettel. Továbbá

$$M\left(\frac{k}{n}\right) = \frac{1}{n} M(k) = p, \quad D^2\left(\frac{k}{n}\right) = \frac{1}{n^2} npq = \frac{pq}{n},$$

vagyis

$$D\left(\frac{k}{n}\right) = \sqrt{\frac{pq}{n}}.$$

Eredményünk azt mutatja, hogy valamely A esemény $\frac{k}{n}$ relatív gyakorisága olyan valószínűségi változó, amelynek várható értéke a $P(A) = p$ (ismeretlen) valószínűséggel egyezik meg, szórása pedig $\frac{1}{\sqrt{n}}$ nagyságrendű.

Definíció:

Empirikus szórásnégyzetnek nevezzük a mintaelemeknek az $\bar{X}$ mintaközéptől való négyzetes átlageltérését, az

$$S_n^2 = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{n} =$$

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

statisztikát.

Az empirikus szórásnégyzet (pozitív előjelű) négyzetgyöke az empirikus szórás:

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

Az empirikus szórásnégyzet az elméleti $D^2(X) = \sigma^2$ szórásnégyzet közelítésére szolgál, mivel az S_n^2 valószínűségi változó σ^2 körül ingadozik. Erre vonatkozik az alábbi tétel.

Tétel:

Az S_n^2 empirikus szórásnégyzet várható értéke:

$$M(S_n^2) = \frac{n-1}{n} \sigma^2.$$

Bizonyítás.

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - m) - (\bar{X} - m)]^2.$$

ahol m -mel az $M(X)$ elméleti várható értéket jelöljük.
Továbbá:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n [(X_i - m) - (\bar{X} - m)]^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - \\ &- 2 \frac{1}{n} (\bar{X} - m) \cdot \sum_{i=1}^n (X_i - m) + \frac{1}{n} (\bar{X} - m)^2 . \end{aligned}$$

Mivel

$$\sum_{i=1}^n (X_i - m) = n\bar{X} - nm = n(\bar{X} - m) ,$$

következik, hogy

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - (\bar{X} - m)^2 .$$

Felhasználva a várható érték additív tulajdonságát (lásd: 2.1.7 pont) adódik, hogy:

$$M(S_n^2) = \frac{1}{n} \sum_{i=1}^n M(X_i - m)^2 - M(\bar{X} - m)^2 .$$

Mivel:

$$\frac{1}{n} \sum_{i=1}^n M(X_i - m)^2 = \frac{1}{n} n \sigma^2 = \sigma^2$$

elméleti szórásnégyzet, és

$$\begin{aligned} M(\bar{X} - m)^2 &= D^2(\bar{X}) = D^2\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \\ &= \frac{1}{n^2} \sum_{i=1}^n D^2(X_i) = \frac{\sigma^2}{n} , \end{aligned}$$

ezért

$$M(S_n^2) = \sigma^2 - \frac{\sigma^2}{n} = \left(1 - \frac{1}{n}\right) \sigma^2 = \frac{n-1}{n} \sigma^2.$$

Minthogy nagy n esetén $1 - \frac{1}{n}$ kevésbé tér el 1-től, nagy elemszámú mintából számított empirikus szórásnégyzet várható értéke közelítőleg az elméleti szórásnégyzettel egyenlő. Kisebb minta esetében célszerűbb σ^2 közelítésére az ún. korrigált empirikus szórásnégyzetet használni.

Definíció:

Korrigált empirikus szórásnégyzetnek nevezzük a

$$S_n^{*2} = \frac{n}{n-1} S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

statisztikát.

Tétel:

Az S_n^{*2} korrigált empirikus szórásnégyzet várható értéke az elméleti szórásnégyzettel egyezik meg:

$$M(S_n^{*2}) = \sigma^2.$$

Bizonyítás.

$$\begin{aligned} M(S_n^{*2}) &= M\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} M(S_n^2) = \\ &= \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2. \end{aligned}$$

A fenti tulajdonsága folytán az S_n^{*2} korrigált empirikus szórásnégyzetet előnyösebb használni a σ^2 elméleti szórásnégyzet statisztikai közelítésére, mint S_n^2 -et.

Mind az empirikus szórásnégyzet, mint a korrigált empirikus szórásnégyzet kiszámítását egyszerűbbé teszi az ún. Steiner-formula alkalmazása, amely szerint

$$\begin{aligned} nS_n^2 &= (n-1)S_n^{*2} = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n [(X_i - a) - (\bar{X} - a)]^2 = \\ &= \sum_{i=1}^n (X_i - a)^2 - 2(\bar{X} - a) \sum_{i=1}^n (X_i - a) + n(\bar{X} - a)^2 = \\ &= \sum_{i=1}^n (X_i - a)^2 - n(\bar{X} - a)^2. \end{aligned}$$

Ha itt $a = 0$ értéket választunk, akkor

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

ez a szórásnégyzetre vonatkozó tétel empirikus megfelelője (lásd: 2.1.7 pont).

A Steiner-formula általános alakjából az is látható, hogy

$$S_n^2 \leq \frac{1}{n} \sum_{i=1}^n (x_i - a)^2 .$$

azaz az $\bar{x}$ mintaközép az a szám, amelytől a mintaelemek négyzetes átlageltérése a legkisebb.

Definíció:

Az empirikus szórásnak és a mintaközépnek a hányadosát, a

$$c_v = \frac{S_n}{\bar{x}} .$$

statisztikát variációs tényezőnek nevezzük.

A variációs tényező azt fejezi ki, hogy a szórás hány százaléka az átlagértéknek. Ezt a tapasztalati mérőszámot csak pozitív értékű valószínűségi változó esetében van értelme használni. A variációs tényezőt relatív szórásnak is szokás nevezni.

Az elméleti momentumok statisztikai közelítésére használjuk a megfelelő tapasztalati momentumokat.

Definíció:

Az $X_1, X_2, \dots, X_n$ statisztikai mintából számított k -adik tapasztalati momentum a mintaelemek k -adik hatványainak számtani közepe:

$$m_k = \frac{1}{n} \sum_{i=1}^n x_i^k$$

Az $\bar{x}$ mintaközép az első tapasztalati momentum:

$$m_1 = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} .$$

A második tapasztalati momentum:

$$m_2 = s_n^2 + \bar{x}^2 .$$

Az

$$m_3 = \frac{1}{n} \sum_{i=1}^n x_i^3$$

harmadik tapasztalati momentumot az eloszlás ferdeségének megítélésére használjuk. Az empirikus ferdeségi együttható:

$$\gamma_1 = \frac{m_3}{s_n^3} .$$

A negyedik tapasztalati momentum az eloszlás lapultságának jellemzésében játszik szerepet. A lapultság szokásos tapasztalati mérőszáma:

$$\gamma_2 = \frac{m_4}{s_n^4} - 3 .$$

3.1.6 A rendezett minták elmélete

A folytonos eloszlású X valószínűségi változó eloszlásfüggvényét jelöljük $F(x)$ -szel, továbbá legyen a változóra vonatkozó n elemű minta: $X_1, X_2, \dots, X_n$. A mintaelemek indexe itt az észlelés sorrendjének felel meg. Rendezzük a mintaelemek megfigyelt értékeit nagyság szerint növekvő sorrendben! Ezt úgy végezzük el, hogy az n darab mintaelem közül kiválasztjuk a legkisebbet - jelöljük ezt x_1^* -gal -, majd a fennmaradó $(n-1)$ darab közül ismét a legkisebbet választjuk, legyen ez x_2^* stb. Ily módon az

$$x_1^* < x_2^* < \dots < x_n^*$$

rendezett mintát nyerjük. Mivel feltevésünk szerint X folytonos eloszlású, ezért a mintaelemek 1 valószínűséggel mind különbözők, nagyság szerinti rendezésük egyértelműen végrehajtható.

A rendezett mintának a matematikai statisztikában igen fontos szerepe van, mint azt a későbbiekben látni fogjuk (lásd: 3.5.1-3.5.2 pontok). A rendezett mintából nyerhető tapasztalati jellemzők igen egyszerűek, ugyanakkor szemléletes képet adnak az eloszlásról.

Definíció:

A nagyság szerint rendezett mintaelemek közül a középső mintaelemet tapasztalati mediánnak nevezzük. Ha a mintaelemek száma páratlan szám: $n = 2m + 1$, akkor a tapasztalati medián: $\hat{X}_{0,5} = X_{m+1}$; míg ha $n = 2m$, akkor a tapasztalati medián:

$$\hat{X}_{0,5} = \frac{X_m^* + X_{m+1}^*}{2}$$

A tapasztalati medián az eloszlás elméleti $\tilde{x}_{0,5}$ (más jelöléssel: m_e) mediánja körül ingadozik. Az $\tilde{x}_{0,5}$ az a szám, amelyre $F(\tilde{x}_{0,5}) = 0.5$ (Ez a mennyiség csak akkor egyértelműen meghatározott, ha a folytonos eloszlásfüggvény szigorúan monoton.) A rendezett mintából a tapasztalati medián megállapítása számolás nélkül, egyszerű megszámolással történik. A tapasztalati medián eloszlására vonatkozik az alábbi tétel, amelyet bizonyítás nélkül közlünk.

Tétel:

Ha a valószínűségi változó $F(x)$ eloszlásfüggvénye szigorúan növekvő és differenciálható függvény, akkor a minta $\hat{X}_{0,5}$ tapasztalati mediánja nagy n esetén közelítőleg normális eloszlású, és várható értéke megegyezik az eloszlás $\tilde{x}_{0,5}$ elméleti mediánjával.

$$M(\hat{X}_{0,5}) = \tilde{x}_{0,5}$$

szórása pedig:

$$D(\hat{X}_{0,5}) = \frac{1}{2f(\tilde{x}_{0,5})\sqrt{n}}$$

ahol $f(\tilde{x}_{0,5})$ az X valószínűségi változó sűrűségfüggvényének értéke az elméleti mediánpontban. Ha $n = 2m + 1$ és n nagy szám, akkor

$$P\left(|X_{m+1}^* - \tilde{x}_{0,5}|\right) < \frac{1}{f(\tilde{x}_{0,5})\sqrt{n}} \approx 0.95 ;$$

vagyis a tapasztalati medián az elméleti mediántól nagy valószínűséggel kevesebbel tér el, mint

$$\frac{1}{f(x_{0,5}) \sqrt{n}}.$$

Normális eloszlás esetében az elméleti medián megegyezik a várható értékkel. Mivel

$$\forall x \in \mathbb{R} \text{ esetén } f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

$$f(\tilde{x}_{0,5}) = f(m) = \frac{1}{\sigma \sqrt{2\pi}}; \quad \frac{1}{f(\tilde{x}_{0,5}) \sqrt{n}} = \sigma \sqrt{\frac{2\pi}{n}},$$

ezér ekkor

$$P \left(|X_{m+1}^* - M(X)| < \sigma \sqrt{\frac{2\pi}{n}} \right) \approx 0,95;$$

tehát normális eloszlás esetén, nagy mintánál a tapasztalati medián jól közelíti az eloszlás várható értékét. (Amennyiben a σ szórás nem ismerjük, a fenti formulában a korrigált empirikus szórás értékét helyettesítjük.)

Mint már vázoltuk, a statisztikai minta elemei valószínűségi változók, mégpedig független és egyforma eloszlású változók. A rendezett minta elemei természetesen ugyancsak valószínűségi változók, hiszen értékük és sorrendjük a véletlentől függ. Az $X_1^*, X_2^*, \dots, X_n^*$ rendezett mintaelemek azonban már nem függetlenek, ami abból is látható, hogy pl. X_1^* nem lehet nagyobb, mint X_2^* stb. Az egyes rendezett mintaelemek valószínűségeloszlása sem azonos. Ezt fejezi ki az alábbi tétel.

Tétel:

A rendezett mintaelemek eloszlásfüggvényei között a nagyságrend a következő:

$$\forall x \in \mathbb{R} \text{ esetén } F_{n,1}(x) \geq F_{n,2}(x) \geq \dots \geq F_{n,n}(x),$$

ahol

$$F_{n,k}(x) = P(X_k^* < x) \quad (k=1, 2, \dots, n).$$

Bizonyítás.

Az $\{X_{k+1}^* < x\}$ esemény maga után vonja az $\{X_k^* < x\}$ eseményt: $\{X_{k+1}^* < x\} \subset \{X_k^* < x\}$ hiszen $X_{k+1}^* > X_k^*$ x ,

következik, hogy $X_k^* < x$.

Ennek következtében:

$$P(X_{k+1}^* < x) \leq P(X_k^* < x),$$

azaz

$$\forall x \in R \text{ esetén } F_{n,k+1}(x) \leq F_{n,k}(x) \quad (k=1,2,\dots,n).$$

Az X valószínűségi változó $F(x)$ eloszlásfüggvénye ismeretében az $X_1^* < X_2^* < \dots < X_n^*$ rendezett mintaelemek eloszlásfüggvényei kifejezhetők. Az alábbi tétel ezt fejezi ki.

Tétel:

Ha a folytonos eloszlású X valószínűségi változó eloszlásfüggvénye $F(x)$, akkor az X -re vonatkozó rendezett minta nagysága szerinti k -adik elemének, X_k^* -nak az eloszlásfüggvénye:

$$F_{n,k}(x) = P(X_k^* < x) = \sum_{i=k}^n \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i}$$

$$(k=1,2,\dots,n).$$

Bizonyítás.

Az $\{X_k^* < x\}$ esemény akkor következik be, ha legalább k számú megfigyelés kisebb, mint x . Ez az alábbi egymást kizáró esetek valamelyikével valósulhat meg: pontosan k darab mintaelem kisebb x -nél, a többi $(n-k)$ darab pedig nagyobb annál; vagy $(k+1)$ darab mintaelem kisebb, mint x , a többi $(n-k-1)$ darab pedig nagyobb, mint x ; ...; vagy mind az n darab megfigyelés, mint x . Annak valószínűsége, hogy n független mintaelem közül k darab kisebb, mint x , a többi pedig nagyobb annál - binomiális eloszlásról lévén szó - $\binom{n}{k} F(x)^k [1 - F(x)]^{n-k}$.

Az $\{X_k^* < x\}$ esemény valószínűsége tehát, mivel egymást kizáró eseményekről van szó:

$$\begin{aligned} \forall x \in R \text{ esetén } P(X_k^* < x) &= F_{n,k}(x) = \\ &= \sum_{i=k}^n \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i} \end{aligned}$$

ami a tétel állítása.

Ha $F_{n,k}(x)$ képletében $k = 1$ -et helyettesítünk, kapjuk az X_1^* legkisebb rendezett mintaelem eloszlásfüggvényét:

$$\begin{aligned} P(X_1^* < x) &= F_{n,1}(x) = \sum_{i=1}^n \binom{n}{i} [1 - F(x)]^{n-i} = \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

$$\begin{aligned} \text{(Itt felhasználjuk, hogy } \sum_{i=0}^n \binom{n}{i} [F(x)]^i [1 - F(x)]^{n-i} &= \\ &= [F(x) + 1 - F(x)]^n = 1 \text{.)} \end{aligned}$$

Az $F_{n,k}(x)$ -et meghatározó összeg k növekedésével csökken, hiszen az összeadandók nem negatív mennyiségek, így a rendezett mintaelemek eloszlásfüggvényei közötti kapcsolat ennek alapján is belátható; sőt az is kiolvasható a formulából, hogy

$$F_{n,k}(x) - F_{n,k+1}(x) = \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}$$

Az $F_{n,k}(x)$ konkrét meghatározásához a binomiális eloszlás táblázata szükséges (Függelék I. táblázat.) Meghatározható a rendezett mintaelemek párjainak, hármasainak stb. együttes eloszlása is.

A statisztikai elemzésben, különösen a minőségellenőrzésben, hasznos számunkra a rendezett minta terjedelmének, az

$$R_n = X_n^* - X_1$$

valószínűségi változónak az eloszlása.

Kimutatható, hogy az R_n változó eloszlásfüggvénye:

$$\forall r \in R \text{ esetén } P(R_n < r) = W_n(r) = n \int_{-\infty}^{\infty} [F(r+x) - F(x)]^{n-1} f(x) dx.$$

Megjegyezzük, hogy a rendezett minták nemcsak a matematikai statisztika elméleti kérdéseinek vizsgálatában, hanem az ipari gyakorlatban, a gyártásközi minőségellenőrzésnél is fontos szerepet játszanak.

3.2 Becsléelmélet

3.2.1 A becslés fogalma

A műszaki gyakorlatban számos esetben előfordul, hogy a vizsgált X valószínűségi változó eloszlásának típusát ismerjük. Tudjuk pl., hogy a Duna évi középvízállása Budapestnél jó közelítéssel normális eloszlású, azaz sűrűségfüggvénye:

$$\forall x \in \mathbb{R} \text{ esetén } f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x - m)^2}{2\sigma^2}}$$

Ugyanilyen alakú a Tisza évi középvízállásának sűrűségfüggvénye pl. Szolnoknál, csak éppen más m és σ paraméterekkel. Általában nem ismerjük az m és σ paramétereket, tehát a várható érték és szórás részleges értékét.

Az m és σ paramétereket az X -re vonatkozó statisztikai mintából kell közelítőleg meghatároznunk, becsülnünk, mert csak ezek ismeretében tudunk az évi középvízállásra vonatkozó valószínűségi kérdésekre válaszolni.

Ha m és $\sigma > 0$ értékét ismerjük, akkor annak valószínűsége, hogy a fenti normális eloszlású X valószínűségi változó adott a, b intervallumba esik:

$$\begin{aligned} P(a \leq X < b) &= P\left(\frac{a - m}{\sigma} \leq \frac{X - m}{\sigma} < \frac{b - m}{\sigma}\right) = \\ &= \Phi\left(\frac{b - m}{\sigma}\right) - \Phi\left(\frac{a - m}{\sigma}\right), \end{aligned}$$

ahol Φ a standard normális eloszlásfüggvény, amelynek értékei a Függelék IV. táblázatában találhatók.

Ha valamely nagy darabszámú tétel, pl. téglatömeg selejtarányára kell következtetnünk n elemű minta alapján, akkor a mintában található selejtes darabok ν száma binomiális eloszlást követ, tehát

$$P(\nu = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (k=1, 2, \dots, n).$$

Feladatunk ekkor a binomiális eloszlás p paraméterének (a selejtarányának az egész tételben) becslése statisztikai mintából.

A tapasztalat azt mutatja, hogy több folyó (pl. Tisza, Körös, Maros, Dráva) esetében egy év során fellépő árhullámok száma Poisson-eloszlást követ.

Jelöljük ν -vel az egy év során bekövetkező árvizek számát, ekkor

$$P(\nu = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Különböző folyókra a λ paraméter értéke különböző, ezért minden folyóra külön kell becsülni a λ paraméter értékét az illető folyó árvizeire vonatkozó statisztikai mintából. (A vízrajzi évkönyvből kiolvassuk az illető folyóra vonatkozólag értékeit mindazon évekre, amelyekről a feljegyzés van.)

Mindegyik valószínűségeloszlás tartalmaz ismeretlen paramétereket, amelyeket statisztikai módszerekkel kell minél jobb közelítéssel meghatározni. Az egyik alapvető statisztikai probléma tehát a következő: A vizsgált valószínűségi változókról tegyük fel először, hogy folytonos eloszlású és sűrűségfüggvénye

$$\forall x \in R \text{ esetén } f(x; a),$$

ahol a ismeretlen paraméter (több paraméter esetén paramétervektor). Az X valószínűségi változóra vonatkozó n elemű mintából kell becsülnünk az a paramétert (vagy annak valamely függvényét). Az a paraméter lehetséges értékeinek halmazát paraméterternek nevezzük.

Normális eloszlás esetén $a = m$ vagy $a = \sigma^2$ vagy $a(m, \sigma)$; exponenciális (Poisson-eloszlás) esetében $a = \lambda$, s.i.t.

Az a paraméter becslése az X -re vonatkozó $X_1, X_2, \dots, X_n$ statisztikai mintából úgy történik, hogy képezzük a mintaelemeknek valamely $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ statisztikai függvényét, és ennek értékét tekintjük az a paraméter közelítő értékének. Az a paraméter konkrét esetben egy meghatározott érték - állandó- α_n viszont mint már a becslés, valószínűségi változó. Nem arról van tehát szó, hogy a mintából kiszámítjuk az a paraméter értékét. A mintából kiszámítunk egy α_n számot, amelynek értéke függ a véletlentől. Annak megítélése, hogy az α_n statisztika mikor tekinthető az a elméleti jellemző (paraméter) "jó" becslésének, többféle szempontból történhet.

Az alábbiakban felsoroljuk azokat a legfontosabb tulajdonságokat, amelyekkel pontbecslés esetében a "jó" becslésnek rendelkeznie kell.

Először is megköveteljük, hogy az α_n statisztikának a várható értéke megegyezzen a becsült paraméterrel, azaz $M(\alpha_n) = a$ legyen. Az ilyen becslést torzítatlan becslésnek nevezzük. Másodszor megköveteljük, hogy α_n értékei lehetőleg

kismértékben ingadozzanak az α paraméter körül, azaz torzítatlan becslés esetében α_n szórása minél kisebb legyen.

Ugyanazt az elméleti paramétert a mintából általában többféle módon becsülhetjük. Két különböző becslés (statisztikai függvény) közül azt tekintjük efficiensebbnek (hatékonyabbnak), amelyiknek kisebb a szórása.

Minél több elemű mintát veszünk, általában annál megbízhatóbb következtetést tudunk levonni az ismeretlen paraméterre vonatkozólag. A jó becsléstől megköveteljük, hogy a minta elemszámának növekedésével az α_n statisztika sztochasztikusan konvergáljon az α paraméterhez. Ezen azt értjük, hogy tetszőleges $\varepsilon > 0$ megadása esetén

$$\lim_{n \rightarrow \infty} P(|\alpha_n - \alpha| > \varepsilon) = 0$$

legyen.

Végül arra törekszünk, hogy az α_n statisztika a mintában rejlő minden információt magában foglaljon az α paraméterre vonatkozólag. Az ilyen statisztikát elégséges statisztikának nevezzük.

Ezeket a követelményeket a továbbiakban részletesen tárgyaljuk.

3.2.2 A becslés tulajdonságai

a) Torzítatlan becslések

Ha az X valószínűségi változó valamely elméleti jellemzője az α paraméter és a statisztikai mintából az $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ statisztikával kívánjuk értékét becsülni, akkor elvárjuk, hogy α_n értékei az α szám körül ingadozzanak; azaz $M(\alpha_n) = \alpha$ legyen.

Definíció.

Ha az $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ statisztika az α elméleti paraméter becslése és α_n várható értéke megegyezik a becsült paraméter értékével, azaz

$$M(\alpha_n) = \alpha,$$

akkor α_n az α paraméternek torzítatlan becslése. Ha $M(\alpha_n) \neq \alpha$, akkor a becslés torzított.

A mintaközép az elméleti várható értéknek torzítatlan becslése, mivel $M(\bar{X}) = M(X) = \alpha$ (ahogy ezt már 3.1.5 pontban bizonyítottuk). A várható értéknek más torzítatlan becslése is létezik (sőt végtelen sok ilyen van).

Tétel:

Az $X_1, X_2, \dots, X_n$ minta-elemeknek valamely

$$\alpha'_n = \tilde{X} = p_1 X_1 + p_2 X_2 + \dots + p_n X_n$$

súlyozott számtani közepe, ahol $\sum_{i=1}^n p_i = 1$, várható értéknek torzítatlan becslése.

Bizonyítás.

$$\begin{aligned} M(\alpha'_n) &= M(\tilde{X}) = M(p_1 X_1 + p_2 X_2 + \dots + p_n X_n) = \\ &= p_1 M(X_1) + p_2 M(X_2) + \dots + p_n M(X_n) = \\ &= M(X) (p_1 + p_2 + \dots + p_n) = M(X). \end{aligned}$$

Tétel:

A $\frac{k}{n}$ relatív gyakoriság torzítatlan becslése valamely A esemény $P(A) = p$ valószínűségének.

Bizonyítás.

$$M\left(\frac{k}{n}\right) = \frac{1}{n} np = p.$$

Tétel.

Ha az X valószínűségi változó eloszlásában szereplő σ^2 paraméter a szórásnégyzetet jelöli $D^2(X) = \sigma^2$, akkor az

$$S_n^* = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

korrigált empirikus szórásnégyzet torzítatlan becslése az elméleti szórásnégyzetnek. A tétel bizonyítását nem részletezzük, mivel az könnyen belátható a 3.1.5 pontban tanultak felhasználásával.

A 3.1.5 pontban megmutattuk azt is, hogy az S_n^2 empirikus szórásnégyzet várható értéke:

$$M(S_n^2) = \frac{n-1}{n} \sigma^2 = \left(1 - \frac{1}{n}\right) \sigma^2.$$

Az empirikus szórásnégyzet tehát σ^2 -nek torzítatlan becslése.

Látható, hogy az $M(S_n^2)$ értéke n -től függ. Általában valamely $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ becslésnél az α_n statisztika értéke függ n -től, azaz $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ becsléssorozat.

Definíció.

Az a paraméternek egy $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ becsléssorozatát aszimptotikusan torzítatlan becsléssorozatnak nevezük, ha

$$\lim_{n \rightarrow \infty} M(\alpha_n) = a.$$

Tétel.

Az S_n^2 empirikus szórásnégyzet aszimptotikusan torzítatlan becslése σ^2 -nek, az elméleti szórásnégyzetnek.

Bizonyítás.

$$\lim_{n \rightarrow \infty} M(S_n^2) = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \sigma^2 = \sigma^2.$$

b.) Konzisztens becslések

Definíció.

Ha az $\alpha_1, \alpha_2, \dots, \alpha_n, \dots$ becsléssorozat sztochasztikusan konvergál az a paraméterhez, azaz bármely tetszőlegesen kicsi $\varepsilon > 0$ szám esetén

$$\lim_{n \rightarrow \infty} P(|\alpha_n - a| > \varepsilon) = 0,$$

akkor az $\{\alpha_n\}$ becsléssorozatot az a paraméter konzisztens becslésének nevezzük.

Tétel.

Ha α_n torzítatlan becslése az a paraméternek és

$$\lim_{n \rightarrow \infty} D^2(\alpha_n) = 0,$$

akkor α_n konzisztens becslés.

Bizonyítás.

A Csebisev-egyenlőtlenség alapján (lásd: 2.3.1 pont).

$$P(|\alpha_n - M(\alpha_n)| \geq \varepsilon) = P(|\alpha_n - a| \geq \varepsilon) \leq \frac{D^2(\alpha_n)}{2}.$$

Mivel $D^2(\alpha_n) \rightarrow 0$ esetén a bal oldalon álló valószínűség is 0-hoz tart, ezért

$$\lim_{n \rightarrow \infty} P(|\alpha_n - a| \geq \varepsilon) = 0.$$

1. Példa.

A 3.1.5 pontban láttuk, hogy az $\bar{X}$ mintaközép torzítatlan becslése az $M(X)$ várható értéknek, továbbá, ha létezik az eloszlás második momentuma, akkor

$$D^2(\bar{X}) = \frac{D^2(X)}{n}.$$

Nyilvánvaló, hogy $n \rightarrow \infty$ esetén:

$$\lim_{n \rightarrow \infty} \frac{D^2(X)}{n} = 0.$$

vagyis ekkor az $\bar{X}$ mintaközép konzisztens becslése az $M(X)$ várható értéknek.

2. Példa.

A $\frac{k}{n}$ relatív gyakoriság (mint láttuk) minden n -re torzítatlan becslése valamely A esemény $P(A) = p$ valószínűségének, és mivel

$$D^2\left(\frac{k}{n}\right) = \frac{pq}{n} \leq \frac{1}{4n} \rightarrow 0, \text{ ha } n \rightarrow \infty,$$

ezért a $\frac{k}{n}$ relatív gyakoriság egyben konzisztens becslése is az ismeretlen p valószínűségnek.

Kimutatható, hogy ha X valószínűségi változó negyedik momentuma létezik, akkor az S_n^2 és S_n^{*2} statisztikák konzisztens becslései a $D^2(X)$ szórásnégyzetnek. Speciálisan, ha X normális eloszlású, 0 várható értékű és σ^2 szórásnégyzetű valószínűségi változó, akkor

$$D^2(S_n^{*2}) = \frac{2\sigma^4}{n-1} \rightarrow 0, \text{ ha } n \rightarrow \infty.$$

Előfordulhat, hogy valamely X valószínűségi változó várható értéke, szórása stb. nem létezik. Ez a helyzet, ha pl. az X valószínűségi változó Cauchy-eloszlású, azaz sűrűségfüggvénye:

$$f(x) = \frac{1}{\pi(1+x^2)}$$

Könnyű belátni, hogy az X várható értékét definiáló

$$M(X) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x}{1+x^2} dx$$

formula jobb oldalán álló integrál nem abszolút konvergens, tehát X várható értéke nem létezik. Hasonló a helyzet X eloszlásának magasabb momentumaival. Ilyen esetben a mintából számított $\bar{X}$, S_n^2 stb. se nem torzítatlan, se nem konzisztens becslések.

c.) A becslés hatásfoka (efficiencia)

A műszaki gyakorlat szempontjából nyilván fontos követelmény, hogy a becslésül szolgáló statisztika lehetőleg minél kisebb ingadozást mutasson. Valamely a paraméter becslésekor az α_n torzítatlan becslést nagyobb hatásfokú, más szóval efficiensebb, becslésnek tekintjük, mint egy másik α'_n torzítatlan becslést, ha $D^2(\alpha_n) \leq D^2(\alpha'_n)$.

Bizonyos esetekben meg lehet találni a számításba jövő becslések között a legkisebb szórásnégyzetűt.

Definíció:

Ha az a paraméternek van olyan α_n^* torzítatlan becslése, amelynek más α_n torzítatlan becslésnél kisebb a szórásnégyzete:

$$D^2(\alpha_n^*) \leq D^2(\alpha_n),$$

akkor α_n^* a legjobb hatásfokú torzítatlan becslése az a paraméternek. Az ilyen becslést efficiens becslésnek nevezzük.

Ki lehet mutatni, hogy ha valamely paraméterre vonatkozólag létezik efficiens becslés, akkor csak egy ilyen van. Akár van efficiens becslés, akár nincs, valamely a paraméterre vonatkozólag, mindig létezik az a paraméter összes lehetséges α_n becslései szórásnégyzetének alsó határa, legyen ez D^2 .

Definíció:

Az a paraméter valamely α_n becslése hatásfokának a $\frac{D^2}{D^2(\alpha_n)}$ hányadost nevezzük, amely mindig 0 és 1 közé esik.

A becslés hatásfokát nem mindig vonatkoztatjuk az összes torzítatlan becslés szórásnégyzetének a legkisebbikére, hanem esetleg a becslések bizonyos szűkebb osztályaira. Ekkor a becslések adott osztályára vonatkoztatott hatásfokról beszélünk. Tekinthejtük pl. a valószínűségi változó $M(X)$ várható értékének összes lineáris torzítatlan becslését a

$$\tilde{X} = \sum_{i=1}^n p_i X_i$$

alakú becsléseket, ahol $\sum_{i=1}^n p_i = 1$. Ekkor érvényes a következő állítás.

Tétel.

Az $M(X)$ várható érték összes $\tilde{X} = \sum_{i=1}^n p_i X_i$ ($\sum_{i=1}^n p_i = 1$) alakú lineáris becslései közül az

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

számtani közép a legkisebb szórásnégyzetű.

Bizonyítás. Már láttuk, hogy

$$D^2(\bar{X}) = \frac{D^2(X)}{n} .$$

Továbbá:

$$\begin{aligned} D^2(\tilde{X}) &= D^2 \left(\sum_{i=1}^n p_i X_i \right) = p_1^2 D^2(X_1) + p_2^2 D^2(X_2) + \dots \\ &= p_n^2 D^2(X_n) = (p_1^2 + p_2^2 + \dots + p_n^2) D^2(X) . \end{aligned}$$

Az $\bar{X}$ mintaközép kisebb szórású, mint $\tilde{X}$, ha

$$\frac{1}{n} \leq p_1^2 + p_2^2 + \dots + p_n^2 .$$

Ez az egyenlőtlenség teljesül, amit az analízisből ismert Cauchy-féle egyenlőtlenség alapján mutathatunk ki. Eszerint, ha $a_1, a_2, \dots, a_n$ és $b_1, b_2, \dots, b_n$ valós számok, akkor mindig fennáll a következő egyenlőtlenség:

$$\left(\sum_{k=1}^n a_k b_k \right)^2 \leq \sum_{k=1}^n a_k^2 \sum_{k=1}^n b_k^2,$$

és egyenlőség akkor és csak is akkor áll fenn, ha $a_k = c b_k$, $k=1, 2, \dots, n$.

Legyen ebben az összefüggésben $a_k = p_k$, $b_k = 1$ ($k=1, 2, \dots, n$) ekkor

$$\left(\sum_{k=1}^n p_k \cdot 1 \right)^2 = 1 = \sum_{k=1}^n p_k^2 \cdot n,$$

azaz

$$\frac{1}{n} \leq \sum_{k=1}^n p_k^2,$$

és egyenlőség csakis a $p_1 = p_2 = \dots = p_n = \frac{1}{n}$ esetben áll fenn.

A 3.2.2 pontban láttuk, hogy a $\frac{k}{n}$ relatív gyakoriság torzítatlan becslése a p valószínűségnek. Bebizonyítható, hogy a relatív gyakoriság a legkisebb szórású becslése a valószínűségnek az összes lehetséges torzítatlan becslések közül.

d.) A Cramér-Rao egyenlőtlenség

Ha az α_n statisztika torzítatlan és konzisztens becslése valamely paraméternek, akkor elég nagy n esetén α_n nagy valószínűséggel tetszőleges közel van a becsült paraméterhez.

Felmerül a kérdés, hogy valamely α paraméterre vonatkozólag létezik-e olyan torzítatlan becslés, amelynek szórásnégyzete már adott véges n mintaelemszám esetén tetszőlegesen közel van zérushoz.

Az olyan $f(x; a)$ sűrűségfüggvényű eloszlások összességére vonatkozólag, amelyeknél a számegyenes azon szakasza, ahol $f(x; a) > 0$ nem függ az a paramétertől, a választ az ún. Cramér-Rao egyenlőtlenség adja meg.

Tétel.

(H. Cramér-C.R.Rao). Ha az $\alpha_n (X_1, X_2, \dots, X_n)$ statisztika torzítatlan becslés a -ra vagy általánosabban annak valamely $g(a)$ függvényére, akkor

$$D^2(\alpha_n) \geq \frac{[g'(a)]^2}{I_n}$$

A nevezőben lévő I_n mennyiség (n a minta elemszáma utal) a következő várható érték:

$$I_n = M \left[\left(\frac{1}{f} \frac{df}{da} \right)^2 \right] = M \left[\left(\frac{d \ln f}{da} \right)^2 \right]$$

ahol a várható érték úgy értendő, hogy a mintaelemek

$f(x_1, x_2, \dots, x_n; a) = \prod_{i=1}^n f(x_i; a)$ együttes sűrűségfüggvényének képletébe az $X_1, X_2, \dots, X_n$ változókat helyettesítjük.

Ha α magának az a paraméternek torzítatlan becslése, akkor

$$D^2(\alpha_n) \geq \frac{1}{I_n}$$

Definíció.

A Cramér-Rao egyenlőtlenségben a $\frac{[g'(a)]^2}{I_n}$ mennyiséget információs határnak nevezzük.

Ha az α_n statisztika szórásnégyzete az információs határt eléri, akkor tehát nincs nála kisebb szórásnégyzetű torzítatlan becslése $g(a)$ -nak. Előfordul, hogy van ugyan legkisebb szórásnégyzetű torzítatlan becslés valamely paraméterre, de az információs határt ez nem éri el. Megemlítjük, hogy ha a szóban forgó paraméterre vonatkozólag létezik "elégséges" becslés (lásd:e) pont), akkor ennek szórásnégyzete eléri az információs határt.

Nem foglalkozunk itt a Cramér-Rao egyenlőtlenség bizonyításával, mindössze annak megjegyzésére szorítkozunk, hogy ez az egyenlőtlenség olyan folytonos eloszlásokra vonatkozik, amelyeknél az $f(x; a)$ sűrűségfüggvény értelmezési tartományának minden pontjában pozitív - mint pl. a normális és az exponenciális eloszlás -, továbbá olyan diszkrét eloszlásokra, amelyeknél a valószínűségi változó minden lehetséges értékét pozitív valószínűséggel veszi fel, mint pl. a binomiális és a Poisson-eloszlás. Nem vonatkozik a Cramér-Rao egyenlőtlenség pl. az egyenletes eloszlásra.

Ugyancsak rámutatunk arra, hogy a Cramér-Rao egyenlőtlenség n elemű minta esetén általában $\frac{1}{n}$ nagyságrendű határt ad a torzítatlan becslés szórásnégyzetére, miként azt a következő példákban látni fogjuk.

Bizonyítás nélkül megemlítjük, hogy olyan eloszlások esetében, amelyekre a Cramér-Rao egyenlőtlenség nem érvényes, valamely paraméterre vonatkozólag létezhet lényegesen kisebb szórásnégyzetű becslés, mint az információs határ. Pl. egyenletes eloszlás esetében az

$M(X) = \frac{a+b}{2}$ várható értéknek torzítatlan becslése a legkisebb és legnagyobb rendezett mintaelem számtani közepe:

$$\frac{x_1^* + x_n^*}{2}, \text{ és ennek szórásnégyzete}$$

$$D^2 \left(\frac{x_1^* + x_n^*}{2} \right) = \frac{(b-a)^2}{2(n+1)(n+2)},$$

amely $\frac{1}{n^2}$ nagyságrendű.

A következőkben néhány fontos eloszlásra vonatkozólag szemléltetjük az információs határ kiszámításának módját.

1. Példa.

Legyen X normális eloszlású ismeretlen $M(X) = a$ várható értékkel és ismert $D(X) = \sigma_0^2$ szórással. Számítsuk ki a torzítatlan becslései szórásnégyzeteinek alsó határát, az $\frac{1}{I}$ mennyiséget!

Megoldás.

A mintaelemek együttes sűrűségfüggvénye:

$$f = f(x_1, x_2, \dots, x_n; a, \sigma_0) = \frac{1}{(\sigma_0 \sqrt{2\pi})^n} e^{-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - a)^2}$$

$$\ln f = n \ln \frac{1}{\sigma_0 \sqrt{2\pi}} - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - a)^2;$$

$$\frac{d \ln f}{da} = -\frac{1}{\sigma_0^2} \sum_{i=1}^n (x_i - a).$$

Ha itt az x_i változók helyébe az X_i mintaelemeket helyettesítjük:

$$\begin{aligned} \varphi(x_1, x_2, \dots, x_n; a, \sigma) &= \left(-\frac{d \ln f}{da} \right)^2 = \\ &= \left[-\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - a) \right]^2 = \frac{1}{\sigma^4} \sum_{i=1}^n (x_i - a)^2 = \\ &= \frac{1}{\sigma_0^4} (n \bar{x} - na)^2 = \frac{n^2}{\sigma_0^4} (\bar{x} - a)^2. \end{aligned}$$

Tehát

$$\begin{aligned} I_n &= M \left(\left[-\frac{d \ln f}{da} \right]^2 \right) = \frac{n^2}{\sigma_0^4} M([\bar{x} - a]^2) = \\ &= \frac{n^2}{\sigma_0^4} D^2(\bar{x}) = \frac{n^2}{\sigma_0^4} = \frac{\sigma_0^2}{n} = \frac{n}{\sigma_0^2}. \end{aligned}$$

Következésképpen, ha α_n torzítatlan becslése a -nak, akkor

$$D^2(\alpha_n) = \frac{1}{I_n} = \frac{\sigma_0^2}{n}.$$

Mivel $D^2(\bar{x}) = \frac{\sigma_0^2}{n}$, tehát a mintaközép olyan torzítatlan becslése a normális eloszlás várható értékének, amelynek szórásnégyzete éppen az alsó határral egyező, vagyis a normális eloszlás várható értékére ennél kisebb szórásnégyzetű torzítatlan becslés nem létezhet.

2. Példa.

Legyen X exponenciális eloszlású, ekkor sűrűségfüggvénye:

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad x \in \mathbb{R}^+.$$

Ismeretes, hogy az exponenciális eloszlás várható értéke:

$$M(X) = \frac{1}{\lambda}.$$

Keressük a $g(\lambda) = \frac{1}{\lambda}$ torzítatlan becsléseire a szórásnégyzetek alsó határát.

Megoldás.

Az $X_1, X_2, \dots, X_n$ mintaelemek együttes sűrűségfüggvénye:

$$f = f(x_1, x_2, \dots, x_n; \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i};$$

$$\ln f = \ln \lambda - \lambda \sum_{i=1}^n x_i;$$

$$\frac{d \ln f}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = \frac{n}{\lambda} - n\bar{x};$$

$$\begin{aligned} I_n &= M \left(\left[\frac{d \ln f}{d \lambda} \right]^2 \right) = M \left(\left[\frac{n}{\lambda} - n\bar{x} \right]^2 \right) = \\ &= \frac{n^2}{\lambda^2} - \frac{2n^2}{\lambda} M(\bar{x}) + n^2 M(\bar{x}^2). \end{aligned}$$

Mivel exponenciális eloszlás esetén:

$$M(X) = \frac{1}{\lambda}, \quad D^2(X) = \frac{1}{\lambda^2};$$

$$M(\bar{x}) = \frac{1}{\lambda},$$

$$D^2(\bar{x}) = \frac{1}{n \lambda^2} = M(\bar{x}^2) - M^2(\bar{x}) = M(\bar{x}^2) - \frac{1}{\lambda^2};$$

ezért

$$I_n = -\frac{n^2}{\lambda^2} + n^2 \left(\frac{1}{n \lambda} + \frac{1}{\lambda^2} \right) = \frac{n}{\lambda^2}.$$

Figyelembe véve, hogy $[g'(\lambda)]^2 = -\left[\frac{1}{\lambda^2}\right]^2$, kapjuk, hogy

$$D^2(\alpha_n) = \frac{\left(-\frac{1}{\lambda^2}\right)^2}{\frac{n}{\lambda^2}} = \frac{1}{n \lambda} = D^2(\bar{x}).$$

Eredményünk azt mutatja, hogy exponenciális eloszlás esetében a várható értékre a legjobb hatásfokú torzítatlan becslés, a mintaközép.

Diszkrét eloszlás esetében a mintaelemek együttes sűrűsége helyett azok együttes valószínűségével kell számolnunk az információs határ kiszámításához.

Tétel.

Valamely A esemény $P(A) = p$ valószínűségére a legjobb hatásfokú torzítatlan becslés a relatív gyakoriság.

Bizonyítás.

Legyen n független kísérletünk, és mindegyiknél regisztráljuk, hogy az A vagy az $\bar{A}$ esemény következett be. Az i -edik kísérletnél legyen

$$X_i = \begin{cases} 1, & \text{ha A bekövetkezett} \\ 0, & \text{ha } \bar{A} \text{ következett be} \end{cases} \quad (i=1, 2, \dots, n).$$

Tehát X_i az A esemény indikátorváltozója az i -edik kísérletnél.

Ha n kísérlet során az A esemény k -szor következett be, akkor

$$\begin{aligned} f &= P(X_1 = 1, X_2 = 0, X_3 = 1, \dots, X_{n-1} = 1, X_n = 0) = \\ &= p^k (1 - p)^{n-k} \end{aligned}$$

(függetlenül attól, hogy melyik k darab kísérlet eredményezett 1-est).

$$\ln f = k \ln p + (n - k) \ln (1 - p);$$

$$\frac{d \ln f}{dp} = \frac{k}{p} - \frac{n - k}{1 - p} = \frac{k - np}{p(1 - p)};$$

$$\begin{aligned} I_n &= M \left(\left[\frac{k - np}{p(1 - p)} \right]^2 \right) = \frac{1}{p^2 (1 - p)^2} M([k - np]^2) = \\ &= \frac{np(1 - p)}{p^2 (1 - p)^2} \cdot \frac{n}{p(1 - p)}. \end{aligned}$$

Az információs határ tehát:

$$\frac{1}{I_n} = \frac{p(1 - p)}{n}.$$

Ismeretes, hogy a relatív gyakoriság szórásnégyzete:

$$D^2\left(\frac{k}{n}\right) = \frac{1}{n^2} D^2(k) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

A relatív gyakoriság szórásnégyzete tehát eléri az információs határt, vagyis a relatív gyakoriság az ismeretlen p valószínűség legjobb hatásfokú torzítatlan becslése.

e.) Elégséges becslések

A matematikai statisztikában arra törekszünk, hogy a vizsgált X valószínűségi változó eloszlására és annak paramétereire vonatkozólag az $X_1, X_2, \dots, X_n$ mintaelemekből minden információt kiolvassunk, amit a mintaelemek az eloszlást illetőleg tartalmaznak. A becsléelmélet szempontjából ezt a törekvést a következőképpen fogalmazhatjuk meg:

Ha az $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ statisztikai függvény (becslés) a becsült α paraméterre vonatkozólag minden információt magában foglal, ami a mintaelemekben volt, akkor az α_n statisztikát elégséges statisztikának vagy elégséges becslésnek nevezzük. Ez a "definíció" a lényegét fejezi ki, de egyrészt nem pontos, másrészt matematikai szempontból még nem ad támpontot arra, hogyan döntsük el adott eloszlásösszesség esetén, hogy létezik-e egyáltalán elégséges statisztika, ill., hogy egy adott statisztika elégséges-e. A kérdést a következő megfontolással világítjuk meg.

Példa.

Legyen az X valószínűségi változó normális eloszlású $M(X) = m$ várható értékkel és $D(X) = \sigma^2$ szórással; vagyis $N(m, \sigma^2)$ eloszlású.

Tegyük fel, hogy az m paramétert az

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$$

statisztikai mintából az

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \Rightarrow \frac{x_1 + x_2 + \dots + x_n}{n} = \bar{x}$$

mintaközéppel becsüljük. Kimutatható, hogy az $\bar{X}$ valószínűségi változó ugyancsak normális eloszlású, $M(\bar{X}) = m$, $D(\bar{X}) = \frac{\sigma^2}{n}$ paraméterekkel; azaz $\bar{X}$ sűrűségfüggvénye:

$$f_n(\bar{x}) = \frac{1}{\sigma} \sqrt{\frac{n}{2\pi}} e^{-\frac{n(x-m)^2}{2\sigma^2}}$$

Minthogy az egyes X_i mintaelemek függetlenek, és egyenként $N(m, \sigma^2)$ eloszlásúak, a mintaelemek együttes sűrűségfüggvénye az egyes változók sűrűségfüggvényeinek szorzata:

$$f(x_1, x_2, \dots, x_n; m) = \frac{1}{(\sigma \sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2}$$

Figyelembe véve, hogy

$$\begin{aligned} \sum_{i=1}^n (x_i - m)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - m)^2 = \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - m) \sum_{i=1}^n (x_i - \bar{x}) + n(\bar{x} - m)^2 = \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - m)^2 \end{aligned}$$

$$(\text{mivel } \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = 0);$$

a mintaelemek együttes sűrűségfüggvénye a következő alakban írható:

$$\begin{aligned} f(x_1, x_2, \dots, x_n; m) &= \frac{1}{\sigma^n} \sqrt{\frac{n}{2\pi}} e^{-\frac{n(x - m)^2}{2\sigma^2}} \\ &= \frac{1}{(\sigma \sqrt{2\pi})^{n-1} \sqrt{n}} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Vegyük észre, hogy a mintaelemek együttes sűrűségfüggvénye két tényezőre bomlik. Az első tényező éppen a mintaközép f_n sűrűségfüggvénye, a második tényező pedig olyan függvény, amelyben az m paraméter szerepel.

Az $X_1, X_2, \dots, X_n$ mintaelemek együttes feltételes sűrűségfüggvénye azon feltétel mellett, hogy az $\bar{X}$ mintaközép adott $\bar{x}$ értékkel egyenlő:

$$f(x_1, x_2, \dots, x_n \mid \bar{X} = \bar{x}) = \frac{f(x_1, x_2, \dots, x_n; \bar{x})}{f_n(\bar{x})} =$$

$$= \begin{cases} \frac{1}{(\sigma \sqrt{2\pi})^{n-1} \sqrt{n}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2} & , \text{ ha } \bar{X} = \bar{x} ; \\ 0, & \text{ egyébként.} \end{cases}$$

A feltételes sűrűségfüggvény értéke nem függ m -től. Ez azt jelenti, hogy amikor az $X_1 = x_1, X_2 = x_2, \dots, X_n = \bar{x}_n$ mintaelemekből az $\bar{X}$ mintaközépet meghatároztuk, ezzel az m várható értékre vonatkozólag a mintából minden információt kimerítettünk, az egyes X_i mintaelemek megvalósult értékei nem adnak további információt m -re vonatkozólag. Az $\bar{X}$ mintaközép tehát elégséges becslés a normális eloszlás m várható értékére vonatkozólag. Az elégséges statisztika a következő módon jellemezhető matematikailag:

Definíció.

Az $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ statisztika elégséges statisztika az a paraméterre vonatkozólag, ha az $X_1, X_2, \dots, X_n$ mintaelemek együttes sűrűségfüggvénye két olyan tényezőre bontható, amelyek egyike az α_n statisztika sűrűségfüggvénye, a másik tényező pedig nem függ az a paramétertől, azaz

$$f(x_1, x_2, \dots, x_n; a) = f_n(\alpha_n; a) f(x_1, x_2, \dots, x_n \mid \alpha_n = \alpha).$$

Megjegyezzük, hogy a mintaközép pl. a Poisson-eloszlás paraméterére, valamint az exponenciális eloszlásnál $\frac{1}{\lambda}$ -ra elégséges becslés. Ugyancsak elégséges becslés a relatív gyakoriság valamely A esemény $P(A) = p$ valószínűségére.

3.2.3 A becslés módszerei

a) A legnagyobb valószínűség (maximum-likelihood) elve

Megismertük a becslés különböző "jó" tulajdonságait, és egy-két statisztikai függvényről megmutattuk, hogy rendelkezik ezekkel a tulajdonságokkal. Több helyen említettük, hogy

valamely valószínűségeloszlás várható értékének becslésére a mintaközepet, a szórásnégyzet becslésére a korrigált empirikus szórásnégyzetet, valamely esemény ismeretlen valószínűségének becslésére a relatív gyakoriságot használjuk. A gyakorlatban legtöbbször ezek azok az ismeretlen paraméterek, amelyeket becsülnünk kell.

Nyitva hagytuk eddig azt a kérdést, hogy valamely adott eloszlás paramétereire vonatkozólag hogyan "található" jó becslés. Más szóval van-e olyan általános matematikai módszer, amely megmutatja, hogy milyen $\alpha_n = \alpha_n(X_1, X_2, \dots, X_n)$ statisztikát számítsunk a mintaelemekből ahhoz, hogy adott a paraméterre jó becslést kapjunk.

A legfontosabb általános módszer a becslésre az ún. maximum likelihood módszer, amely magyarra fordítva a legnagyobb valószínűség módszerének nevezhető. A módszer R.A.Fishertől származik, lényegét az alábbiakban ismertetjük.

Tegyük fel, hogy X folytonos eloszlású valószínűségi változó, sűrűségfüggvénye $f(x; a)$, ahol a ismeretlen paraméter. Legyen az X -re vonatkozó statisztikai minta:

$$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n,$$

ahol az x_i számok a megfigyelések elvégzése során a mintaelemek észlelt értékei.

Minthogy a mintaelemek függetlenek és azonos eloszlásúak, együttes sűrűségfüggvényük:

$$L(x_1, x_2, \dots, x_n; a) = f(x_1; a) f(x_2; a) \dots f(x_n; a).$$

Az $L(x_1, x_2, \dots, x_n; a)$ függvényt likelihood-függvénynek nevezzük. A likelihood-függvény tehát a mintaelemek együttes sűrűségfüggvénye, a változók helyében azonban beírtuk a mintaelemek megvalósult értékeit, az észlelt (mért) számokat. Ennek következtében a likelihood-függvény csak az a ismeretlen paraméter függvénye, amelyet most változónak tekintünk, és a -val jelöljük. Másrészt az $L(x_1, x_2, \dots, x_n; a)$ függvény az $X_1, X_2, \dots, X_n$ minta sűrűségfüggvényének értéke az n dimenziós tér $(x_1, x_2, \dots, x_n)$ pontjában.

Ha az $X_1, X_2, \dots, X_n$ statisztikai mintát n dimenziós valószínűségi változónak tekintjük, akkor annak valószínűsége, hogy e vektorváltozó értéke az n dimenziós tér $(x_1, x_2, \dots, x_n)$ középpontú, kicsiny $\Delta x_1, \Delta x_2, \dots, \Delta x_n$ élek által meghatározott téglájába essék, közelítőleg:

$$f(x_1; a) f(x_2; a) \dots f(x_n; a) \Delta x_1 \Delta x_2 \dots \Delta x_n.$$

Ha a $\Delta x_1, \Delta x_2, \dots, \Delta x_n$ számokat, az n dimenziós "téglala" éleit rögzítettük, akkor ez a valószínűség akkor maximális, ha a likelihood-függvény a maximumát veszi fel.

Definíció.

A folytonos eloszlású X valószínűségi változó $f(x; a)$ sűrűségfüggvényében szereplő a paraméter maximum likelihood becslésének azt az a statisztikát nevezzük, amelyekre a mintaelemek együttes sűrűségfüggvénye maximális értéket vesz fel, azaz

$$L(x_1, x_2, \dots, x_n; a) = f(x_1; a) f(x_2; a) \dots f(x_n; a) : \text{maximum.}$$

Mivel egy függvény ugyanott veszi fel a maximumát, ahol a függvény logaritmusa, általában egyszerűbb a értékét az

$$\ln L(x_1, x_2, \dots, x_n; a) = \sum_{i=1}^n \ln f(x_i; a) = \text{maximum}$$

egyenletből meghatározni.

Mintegy a likelihood-függvény esetünkben csak az a paraméter függvénye, szélsőértéke csak ott lehet, ahol

$$\frac{d \ln L}{da} = 0.$$

Ha ennek az egyenletnek több gyöke van, akkor a lokális szélsőértékek közül kell kiválasztani az abszolút maximumhelyet.

Példa.

Legyen X exponenciális eloszlású és sűrűségfüggvénye:

$$f(x; \lambda) = \lambda e^{-\lambda x}.$$

Keressük a λ paraméter $\hat{\lambda}$ maximum likelihood becslését.

Megoldás.

A likelihood-függvény:

$$L(x_1, x_2, \dots, x_n; \hat{\lambda}) = \hat{\lambda} e^{-\hat{\lambda} x_1} \hat{\lambda} e^{-\hat{\lambda} x_2} \dots \hat{\lambda} e^{-\hat{\lambda} x_n} =$$

$$= \hat{\lambda}^n e^{-\hat{\lambda} \sum_{i=1}^n x_i};$$

$$\ln L = n \ln \hat{\lambda} - \hat{\lambda} \sum_{i=1}^n x_i;$$

$$\frac{d \ln L}{d \hat{\lambda}} = \frac{n}{\hat{\lambda}} - \sum_{i=1}^n x_i = \frac{n}{\hat{\lambda}} - n\bar{x} = 0.$$

Inne

$$\hat{\lambda} = \frac{1}{\bar{x}}.$$

A λ paraméter maximum likelihood becslése tehát a mintaközép reciprok értéke. Mivel exponenciális eloszlás esetén

$M(X) = \frac{1}{\lambda}$, következik, hogy a várható érték maximum likelihood becslése ez esetben az $\bar{x}$ mintaközép.

Abban az esetben, ha az X valószínűségi változó eloszlása több ismeretlen paramétert tartalmaz, akkor a likelihood-függvény mindegyik paraméter szerinti parciális deriváltjait kell zérussal egyenlővé tenni. Nézzünk egy példát erre az esetre is.

Példa.

Legyen $X \sim N(m, \sigma^2)$ eloszlású. Számítsuk ki az m és σ paraméterek $\hat{m}$ és $\hat{\sigma}$ maximum likelihood becsléseit!

$$L(x_1, x_2, \dots, x_n; \hat{m}, \hat{\sigma}) = \frac{1}{(\hat{\sigma} \sqrt{2\pi})^n} e^{-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (x_i - \hat{m})^2}$$

$$\ln L = n \ln \frac{1}{\hat{\sigma} \sqrt{2\pi}} - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (x_i - \hat{m})^2.$$

Mivel egy függvény logaritmusának ugyanazon a helyen van maximuma, ahol parciális deriváltjai nullával egyenlők, nézzük először a $L(x_1, x_2, \dots, x_n; \hat{m}, \hat{\sigma})$ függvény $\hat{m}$ szerinti parciális deriváltját:

$$\frac{\partial \ln L}{\partial \hat{m}} = \sum_{i=1}^n \frac{x_i - \hat{m}}{\hat{\sigma}^2} = 0,$$

innen

$$\sum_{i=1}^n x_i - n\hat{m} = 0 ,$$

azaz

$$n\bar{x} - n\hat{m} = 0 ,$$

$$\hat{m} = \bar{x} .$$

Tekintsük most az L függvény $\hat{\sigma}$ szerinti deriváltját:

$$\frac{\partial \ln L}{\partial \hat{\sigma}} = - \frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}} \sum_{i=1}^n (x_i - \hat{m})^2 = 0 .$$

m helyébe az $\bar{x}$ értéket helyettesítve:

$$- n \hat{\sigma}^2 \sum_{i=1}^n (x_i - \bar{x})^2 = 0 ;$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = S_n^2$$

A maximum likelihood módszer tehát normális eloszlású esetében a várható érték becslésére a mintaközepet, a szórásnégyzet becslésére az empirikus szórásnégyzetet szolgáltatja. Ez utóbbi becslés, mint a (3.2.2) pontban már láttuk, asszimptotikus torzítatlan.

A maximum likelihood módszer diszkrét valószínűségeloszlások esetében lényegében ugyanúgy alkalmazható, mint folytonos eloszlások esetében, ekkor azonban a likelihood-függvény nem a mintaelemek együttes sűrűségfüggvénye, hanem azok együttes valószínűsége.

Ha X diszkrét eloszlású és eloszlása az α ismeretlen paraméter maximum likelihood becslését $P(X = x_i) = p_i(x_i; \alpha)$, $i=1,2,\dots$, akkor az α paraméter maximum likelihood becslését a következő módon nyerjük. Ha az

$$x_1, x_2, \dots, x_r$$

mintabeli értékek gyakoriságai rendre

$$k_1, k_2, \dots, k_r$$

akkor a likelihood-függvény:

$$L(x_1, x_2, \dots, x_n; a) = p(x_1, a)^{k_1} \cdot p(x_2, a)^{k_2} \dots p(x_r, a)^{k_r}$$

Példa.

Legyen X Poisson-eloszlású ismeretlen λ paraméterrel, amelyet az $X_1, X_2, \dots, X_n$ statisztikai mintából a maximum likelihood módszerrel kívánunk becsülni.

Megoldás.

A mintaelemek mindegyike elvileg a $0, 1, 2, \dots$ nem negatív egész számokbármelyikét felveheti, a mintában azonban csak véges számú érték fordul elő.

Jelölje r a legnagyobb megfigyelt értéket, és legyenek a $0, 1, 2, \dots, r$ számok gyakoriságai rendre $k_0, k_1, k_2, \dots, k_r$,

ahol $\sum_{i=0}^n k_i = n$, mivel n elemű mintánk van.

Annak valószínűsége, hogy valamelyik mintaelem az i értéket veszi fel:

$$P(X_j = i) = \frac{\lambda^i}{i!} e^{-\lambda}$$

A likelihood-függvény tehát

$$L = \prod_{i=0}^r \left(\frac{\hat{\lambda}^i}{i!} e^{-\hat{\lambda}} \right)^{k_i};$$

$$\ln L = \sum_{i=0}^r k_i \ln \left(\frac{\hat{\lambda}^i}{i!} e^{-\hat{\lambda}} \right) = \sum_{i=0}^r k_i (i \ln \hat{\lambda} - \ln i!) - n \hat{\lambda};$$

$$\frac{d \ln L}{d \hat{\lambda}} = \frac{1}{\hat{\lambda}} \sum_{i=1}^r i k_i - n.$$

Maximum csak ott lehet, ahol ez a derivált 0, azaz

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^r i k_i = \frac{1}{n} \sum_{j=1}^n x_j = \bar{x}.$$

(Azt, hogy a $\hat{\lambda} = \bar{X}$ helyen L-nek valóban maximuma van a

$$\frac{d^2 \ln L}{d \hat{\lambda}^2} = - \frac{1}{\hat{\lambda}^2} \sum_{i=1}^n i k_i \text{ második derivált mutatja.)}$$

A maximum likelihood módszer tehát a λ paraméter becsléseként a mintaközepet szolgáltatja, amely a várható érték torzítatlan becslése.

Vizsgáljuk most meg, milyen statisztikához vezet a maximum likelihood módszer valamely A esemény ismeretlen $P(A)=p$ valószínűségének becslésekor!

Tegyük fel, hogy n kísérlet során az A esemény k -szor következett be. Legyen X_i az i -edik kísérletnél az A esemény indikátorváltozója, azaz

$$X_i = \begin{cases} 1, & \text{ha A bekövetkezett} \\ 0, & \text{ha } \bar{A} \text{ következett be.} \end{cases}$$

Az $X_1, X_2, \dots, X_n$ minta most indikátorváltozók sorozata.

Ha az A esemény k -szor következett be, akkor a mintában k darab 1-es és $(n-k)$ darab 0 szerepel. A minta valószínűsége:

$$L = \hat{p}^k (1 - \hat{p})^{n-k},$$

$$\ln L = k \ln \hat{p} + (n-k) \ln (1 - \hat{p});$$

$$\frac{d \ln L}{d \hat{p}} = \frac{k}{\hat{p}} - \frac{n-k}{1-\hat{p}} = \frac{k - n\hat{p}}{\hat{p}(1-\hat{p})} = 0$$

Tehát

$$k - n\hat{p} = 0,$$

$$\hat{p} = \frac{k}{n}.$$

A valószínűség maximum likelihood becslése a relatív gyakoriság, amely mint a 3.2.2. a)-e) pontokban láttuk, egyben torzítatlan, konzisztens, efficiens és elégséges becslés.

A maximum likelihood módszer elég általános feltételek mellett "jó" becsléshez vezet: konzisztens becsléshez, amely nagy n esetén közelítőleg normális eloszlású. A maximum likelihood becslés torzítatlan vagy aszimptotikus torzítatlan (miként a szórásnégyzet becslése esetében). Ha a becsült

paraméterre létezik elégséges statisztika, akkor a maximum likelihood módszer ezt szolgáltatja. Ezek azok a tulajdonságok, amelyek a maximum likelihood módszert a többi becslési módszerek közül kitüntetik.

b) A momentumok módszere

A momentumok módszere abban áll, hogy ha k számú paramétert akarunk becsülni, akkor az eloszlás első k számú elméleti momentumát egyenlővé tesszük a mintából számított tapasztalati momentumokkal; ily módon az ismeretlen paraméterekre egyenletrendszert nyerünk, amely kedvező esetben megoldható. A momentumok módszerének elméleti alapját Glivenkó tétele adja, amely szerint az $F_n(x)$ empirikus eloszlásfüggvény - a minta eloszlásfüggvénye - n növekedésével konvergál az $F(x)$ elméleti eloszlásfüggvényhez. Ebből következik, hogy a mintából számított empirikus momentumok az elméleti momentumokhoz konvergálnak.

Ha az X valószínűségi változóra vonatkozó n elemű minta $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, akkor a tapasztalati momentumok:

$$m_1 = \frac{\sum_{i=1}^n x_i}{n}, m_2 = \frac{\sum_{i=1}^n x_i^2}{n}, \dots, m_k = \frac{\sum_{i=1}^n x_i^k}{n}$$

Példa.

Legyen $X \sim N(m, \sigma^2)$ eloszlású, ahol $m = M(X) = \mu_1$, $M(X^2) = \mu_2$. Ekkor $\sigma^2 = D^2(X) = M(X^2) - M^2(X) = \mu_2 - \mu_1^2$.
Legyen

$$\mu_1 = m_1 = \bar{x}, \mu_2 = m_2 = \frac{\sum_{i=1}^n x_i^2}{n}$$

Ekkor az m és σ^2 paraméterek becslésére az alábbiakat kapjuk:

$$\hat{m} = \bar{x}, \hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = s_n^2$$

A maximum likelihood módszerrel ugyanezt az eredményt kaptuk.

3.2.4 A becslés megbízhatósága. Konfidenciaintervallumok

A becslésről szóló eddigi fejtegetéseink során az eloszlás valamely ismeretlen a paraméterét egyetlen mennyiséggel, a mintaelemekből számított α_n statisztika numerikus értékével, tehát egyetlen számadattal becsültük, azaz pontbecslést alkalmaztunk. Mivel α_n valószínűségi változó, aktuális értéke általában eltér a becsült a paramétertől. Ha sokszor (sok n -es mintából) végezzük a becslést, akkor α_n értékei - torzítatlan becslés esetén - az a paraméter körül szóródnak. A szóródás mértéke természetesen függ a minta nagyságától. Bár egyetlen $X_1, X_2, \dots, X_n$ mintából nem tudjuk megmondani a pontos értékét, az α_n statisztika eloszlásának ismeretében sokszor meg tudunk adni egy olyan (c_1, c_2) intervallumot, amely az ismeretlen paramétert nagy - pl. 0,95 - valószínűséggel tartalmazza. Az ilyen intervallumot az a paraméterre vonatkozó 95 %-os konfidenciaintervallumnak (megbízhatósági intervallumnak) nevezzük. A továbbiakban az intervallumbecsléssel foglalkozunk.

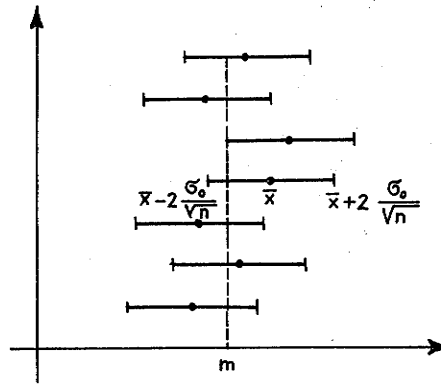
a) Konfidenciaintervallum a normális eloszlás várható értékére

Tegyük fel, hogy az X valószínűségi változó $N(m, \sigma_0)$ eloszlású, ahol a σ_0 szórás ismert. Az m paramétert az $X_1, X_2, \dots, X_n$ statisztikai mintából becsüljük az $\bar{X}$ mintaközéppel. Kimutatható, hogy az $\bar{X}$ statisztika ugyancsak normális eloszlású $M(\bar{X}) = m$ várható értékkel és $D(\bar{X}) = \frac{\sigma_0}{\sqrt{n}}$ szórással. A normális eloszlás ismert tulajdonsága az ún. 2 - szabály alapján $\bar{X}$ értéke kb. 0,95 valószínűséggel az $(m - 2 \frac{\sigma_0}{\sqrt{n}}, m + 2 \frac{\sigma_0}{\sqrt{n}})$ intervallumba esik, azaz

$$P(m - 2 \frac{\sigma_0}{\sqrt{n}} \leq \bar{X} < m + 2 \frac{\sigma_0}{\sqrt{n}}) \approx 0,95$$

A formula valószínűségelméleti értelmezése világos. Ha ismerjük az m várható értéket, és a számegyenesen megrajzolnánk az $(m - 2 \frac{\sigma_0}{\sqrt{n}}, m + 2 \frac{\sigma_0}{\sqrt{n}})$ intervallumot, akkor az X -re vonatkozó sok, pl. 100 darab n elemű minta mindegyi-

kének számtani közepét kiszámítva kb. 95 db mintaközép ebbe az intervallumba esik. (38. ábra)



38. ábra

Mint hogy m értékét nem ismerjük, a fenti intervallumot nem tudjuk megrajzolni.

Rendezzük át ezért összefüggésünket a következő formára:

$$P\left(\bar{X} - 2 \frac{\sigma_0}{\sqrt{n}} < m < \bar{X} + 2 \frac{\sigma_0}{\sqrt{n}}\right) \approx 0,95 .$$

Ezen összefüggés valószínűségelméleti értelme a következő: Az m ismeretlen paraméter nem valószínűségi változó, hanem állandó, a számegyenes egy pontja. Viszont valószínűségi változó az $(\bar{X} - 2 \frac{\sigma_0}{\sqrt{n}}, \bar{X} + 2 \frac{\sigma_0}{\sqrt{n}})$ intervallum közép-

pontja. Annak valószínűsége, hogy ez a véletlen helyzetű intervallum tartalmazza (lefedi) az m pontot, közelítőleg 0,95.

A $(\bar{X} - 2 \frac{\sigma_0}{\sqrt{n}}, \bar{X} + 2 \frac{\sigma_0}{\sqrt{n}})$ véletlen helyzetű intervallumot a normális eloszlás m várható értékére vonatkozó 95 %-os konfidenciaintervallumnak nevezzük. Ez az intervallum annál rövidebb, minél nagyobb n értéke, a minta nagysága.

Természetesen nemcsak 95 %-os konfidenciaintervallumot tudunk konstruálni az $N(m, \sigma_0)$ eloszlású X valószínűségi változó várható értékére. Ha σ_0 ismert, akkor mint az kimutatható, az $\bar{X}$ mintaközép $N(m, \frac{\sigma_0}{\sqrt{n}})$ eloszlású, így tetsző-

leges kicsiny $\varepsilon > 0$ számhoz meghatározható olyan u_ε mennyiség, hogy a

$$P\left(m - u_\varepsilon \frac{\sigma_0}{\sqrt{n}} \leq \bar{X} < m + u_\varepsilon \frac{\sigma_0}{\sqrt{n}}\right) = 1 - \varepsilon$$

teljesül.

Ez az összefüggés ekvivalens a

$$P\left(-u_\varepsilon \leq \sqrt{n} \frac{\bar{X} - m}{\sigma_0} < u_\varepsilon\right) = 2 \Phi(u_\varepsilon) - 1 = 1 - \varepsilon$$

összefüggéssel, amiből látható, hogy a Φ standard normális eloszlásfüggvény táblázata segítségével (Függelék IV. táblázat) u_ε meghatározható.

Normális eloszlás esetén tehát az $(\bar{X} \pm u_\varepsilon \frac{\sigma_0}{\sqrt{n}})$,

$\bar{X} \pm u_\varepsilon \frac{\sigma_0}{\sqrt{n}}$ intervallum $(1 - \varepsilon)$ szintű konfidenciaintervallum az m várható értéke. Adott eloszlás esetén minél nagyobb az $(1 - \varepsilon)$ szint, annál hosszabb konfidenciaintervallumot nyerünk. Nagy biztonsággal csak viszonylag hosszabb intervallumról állíthatjuk, hogy valóban tartalmazza az ismeretlen paramétert.

b) Konfidenciaintervallum ismeretlen valószínűségre

Valamely A esemény $P(A) = p$ valószínűségére kétféle módon szerkeszthetünk konfidenciaintervallumot annak megfelelően, hogy nagy elemszámú minta vagy csupán kis számú megfigyelés áll-e rendelkezésünkre.

a.) Ha a kísérletek száma, az n nagy szám és az A esemény gyakorisága k , akkor (mivel k binomiális eloszlású) $M(k) = np$, és $D(k) = np(1 - p)$, a

$$\frac{k - np}{\sqrt{npq}}$$

ún. standardizált binomiális eloszlású valószínűségi változó közelítőleg standard normális eloszlású, ezért:

$$P\left(-u_\varepsilon \leq \frac{k - np}{np(1 - p)} < u_\varepsilon\right) \approx 2 \Phi(u_\varepsilon) - 1.$$

Innen

$$P\left(-\frac{u_\varepsilon}{\sqrt{n}} \sqrt{p(1 - p)} \leq \frac{k}{n} - p < \frac{u_\varepsilon}{\sqrt{n}} \sqrt{p(1 - p)}\right) \approx 2 \Phi(u_\varepsilon) - 1.$$

$(1 - \varepsilon)$ valószínűséggel fennáll tehát az

$$n \left(\frac{k}{n} - p \right)^2 < u_{\varepsilon}^2 p(1-p)$$

összefüggés. Ebből némi számolás után az alábbi konfidenciaintervallumot kapjuk:

$$\left(\frac{\frac{k}{n} + \frac{u_{\varepsilon}^2}{2n} - \frac{u_{\varepsilon}}{\sqrt{n}} \sqrt{\frac{k}{n} \left(1 - \frac{k}{n}\right) + \frac{u_{\varepsilon}^2}{4n}}}{1 + \frac{u_{\varepsilon}^2}{n}}; \frac{\frac{k}{n} + \frac{u_{\varepsilon}^2}{2n} + \frac{u_{\varepsilon}}{\sqrt{n}} \sqrt{\frac{k}{n} \left(1 - \frac{k}{n}\right) + \frac{u_{\varepsilon}^2}{4n}}}{1 + \frac{u_{\varepsilon}^2}{n}} \right)$$

Az intervallum megbízhatósági szintje közelítőleg $(1 - \varepsilon)$.

Ha n nagy, úgy hogy $\frac{1}{n}$ már elhanyagolható $\frac{1}{\sqrt{n}}$ -hez képest, akkor a konfidenciaintervallum egyszerűbb alakot ölt:

$$\left(\frac{k}{n} - \frac{u_{\varepsilon}}{\sqrt{n}} \sqrt{\frac{k}{n} \left(1 - \frac{k}{n}\right)}; \frac{k}{n} + \frac{u_{\varepsilon}}{\sqrt{n}} \sqrt{\frac{k}{n} \left(1 - \frac{k}{n}\right)} \right)$$

b) Ha n nem nagy szám, akkor a binomiális eloszlás táblázat segítségével konstruálhatunk konfidenciaintervallumot. Tegyük fel, hogy n kísérlet során az A esemény k -szor következett be. A k valószínűségi változó binomiális eloszlású ismeretlen $P(A) = p$ paraméterrel.

Ha p a paraméter igazi értéke és ezt ismernénk, akkor adott $(1 - \varepsilon)$ -hoz a binomiális eloszlás táblázata segítségével találnánk olyan c_1 és c_2 számokat, hogy teljesüljön (jó közelítéssel) a

$$P(c_1 \leq k < c_2) = \sum_{i=c_1}^{c_2-1} \binom{n}{i} p^i (1-p)^{n-i} \approx 1 - \varepsilon.$$

Itt természetesen mind c_1 , mind c_2 függ ε -től és p -től:

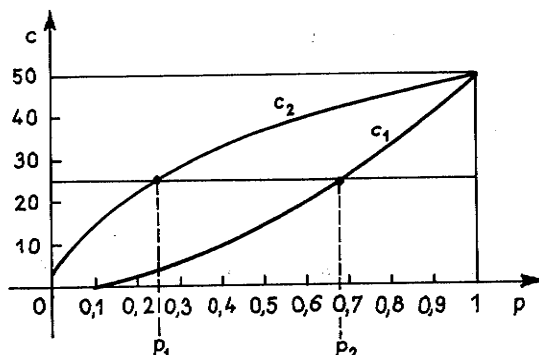
$$c_1 = c_1(\varepsilon, p); \quad c_2 = c_2(\varepsilon, p).$$

Adott p esetén a c_1 és c_2 számokat úgy választjuk, hogy c_1 a legnagyobb és c_2 a legkisebb olyan egész szám legyen, amelyekre

$$\sum_{i=0}^{c_1-1} \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{\varepsilon}{2} \quad \text{és} \quad \sum_{i=c_2}^n \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{\varepsilon}{2}$$

teljesüljön. Adott p esetén nyilván $c_1 < c_2$.

Mivel p akutális értékét nem ismerjük, elég sűrűn egymásra következő $p_1, p_2, \dots, p_s$ értékek mindegyikére meghatározzuk a $c_1^{(i)}, c_2^{(i)}$ határokat. Ezeket ábrázolva egy $(p; c)$ koordináta-rendszerben két monoton függvényt kapunk, amely p_1 . $n = 50$ és $\varepsilon = 0,01$ esetén sematikusán az alábbi módon ábrázolható (39. ábra).



39. ábra

A c tengelyen lévő $c = k$ helyhez a p tengelyen vetítésével egy (p_1, p_2) intervallumot kapunk ($p_1 < p_2$), amely az ismeretlen p valószínűségre közelítőleg $(1 - \varepsilon)$ szintű konfidenciaintervallum.

c.) Konfidenciaintervallum a normális eloszlás szórásnégyzetére

Ha az X valószínűségi változó $N(m, \sigma^2)$ eloszlású és a σ^2 szórásnégyzetre akarunk adott $(1 - \varepsilon)$ szintű konfidenciaintervallumot konstruálni, akkor adott n elemű mintából először kiszámítjuk az S_n^2 empirikus szórásnégyzetet. Az S_n^2 statisztika természetesen valószínűségi változó. Ugyancsak valószínűségi változó az $n \frac{S_n^2}{\sigma^2}$ statisztika, amelyről ki-

mutatható, hogy $(n - 1)$ szabadságfokú χ^2 -eloszlású. A χ^2 -eloszlás táblázata (Függelék V. táblázat) segítségével meg tudunk határozni olyan c_1 és c_2 állandókat, amelyekre

$$P\left(n \frac{s_n^2}{\sigma^2} < c_1\right) = \frac{\varepsilon}{2} \quad \text{és} \quad P\left(n \frac{s_n^2}{\sigma^2} < c_2\right) = 1 - \frac{\varepsilon}{2},$$

azaz

$$P(c_1 \leq n \frac{s_n^2}{\sigma^2} < c_2) = 1 - \varepsilon.$$

Ebben az egyenletben a zárójelen belüli egyenlőtlenség ekvivalens a

$$\sqrt{n} \frac{s_n}{\sqrt{c_2}} \leq \sigma < \sqrt{n} \frac{s_n}{\sqrt{c_1}}$$

relációval, amely σ -ra $(1 - \varepsilon)$ szintű konfidenciaintervallum.

3.3 A statisztikai hipotézisek (feltevések) vizsgálata

3.3.1 Elvi megjegyzések

Definíció.

Statisztikai hipotézisen egy vagy több valószínűség eloszlásra vonatkozó valamilyen feltevést értünk.

Ha a vizsgált valószínűségi változó eloszlásának függvény típusa ismert és ismeretlen paramétereket tartalmaz, akkor a statisztikai hipotézis a paraméterekre vonatkozik. Ha az eloszlás függvény típusa nem ismert, akkor a feltevés magára az eloszlásra vonatkozik. Gyakran az érdekel bennünket, hogy két vagy több valószínűségi változó eloszlása megegyezik-e (homogentásvizsgálat). Olykor azt kell eldöntenünk, hogy két (vagy több) valószínűségi változó várható értéke (vagy szórása) megegyezik-e vagy sem.

Az alábbiakban felsorolunk néhány gyakorlati problémát, amelyek érzékeltetik a statisztikai hipotézis fogalmát.

Példák:

1. Valamely folyó esetében adott c vízállásértékre vonatkozólag el kell döntenünk, hogy annak valószínűsége, hogy az évi maximális vízállás meghaladja a c értéket, kisebb-e mint 2 %.

2. Adott technológiával gyártott panel mérete normális eloszlást követ, ismert szórással. Kérdés, hogy az adott méret az előírt 300 cm mint várható érték körül ingadozik-e.

3. Két különböző keverőgéppel készült aszfaltkeverékben a bitumentartalom szórása megegyezik-e?

4. Két különböző gyárban készült azonos típusú beton. Megegyezik-e ezek törési szilárdságának eloszlása?

5. Adott folyó évi középvízhozama mint valószínűségi változó normális eloszlást követ-e?

A felsorolt kérdések mindegyikében szerepel egy vagy két valószínűségi változó. A kérdések a valószínűségi változók ismeretlen eloszlására vagy azok valamely paraméterére vonatkoznak. A kérdések mindegyikére igenlő vagy tagadó választ kell adnunk, döntenünk kell.

A feltett kérdések a valószínűségelmélet terminológiájával a következő módon fogalmazhatók meg:

1. Ha valamely folyó évi maximális vízállását az X valószínűségi változóval jelöljük és c adott érték, igaz-e, hogy

$$P(X \geq c) < 0,02.$$

2. Ha $X \sim N(m, \sigma_0)$ eloszlású ismert σ_0 szórással; igaz-e, hogy

$$M(X) = 300.$$

3. Jelölje X , ill. Y az első, ill. a második aszfaltkeverő gépen előállított aszfaltkeverék bitumentartalmát. Igaz-e, hogy

$$D(X) = D(Y)?$$

4. Legyen X , ill. Y a két különböző gyárban előállított beton törési szilárdsága. Legyen X eloszlásfüggvénye $F(x)$; Y eloszlásfüggvénye $G(x)$. Megy-e a két eloszlásfüggvény vagyis igaz-e hogy

$$F(x) \equiv G(x).$$

5. Jelölje az X valószínűségi változó a szóban forgó folyó évi középvízhozamát. Igaz-e, hogy

$$P(X < x) = \Phi(x; m, \sigma)$$

valamely (m, σ) értékpárra.

Az említett problémákkal kapcsolatban a statisztika hipotézisvizsgálat abból indul ki, hogy a fenti módon megfogalmazott állítást igaznak tételezi fel.

Ez a feltételezés az ún. nullhipotézis, amelyet H_0 -val jelölünk. Az eloszlás vagy paraméter számára a nullhipotézistől eltérő más lehetőségek bizonyos halmazát, esetleg az összes más lehetőségek együttesét ellenhipotézisnek vagy alternatív hipotézisnek nevezzük, és H_1 -gyel jelöljük.

A feltevésvizsgálatban megkülönböztetünk egyszerű és összetett hipotézist. Egyszerű hipotézisre példa a 2. probléma, ahol

$$H_0: M(X) = m_0 = 300.$$

Ez egyszerű hipotézis, mert a szórást ismerjük, és így m_0 értéke az eloszlást egyértelműen meghatározza, normális eloszlásról lévén szó. Ha ellenben úgy tettük volna fel a kérdést, igaz-e, hogy

$$H'_0: 299 < M(X) = m < 301$$

akkor összetett hipotézisről van szó, mert a H'_0 hipotézist mindazon σ_0 szórású normális eloszlások kielégítik, amelyek várható értéke a 299 és 301 határok közé esik. Ekkor tehát az eloszlások bizonyos összességére vonatkozik a nullhipotézis.

A matematikai statisztika elméletében megkülönböztetünk paraméteres és nemparaméteres problémákat.

Ha a vizsgált X valószínűségi változó $F(x; \theta_1, \theta_2, \dots, \theta_k)$ eloszlásfüggvényének típusa ismert, akkor - ha a $\theta_1, \theta_2, \dots, \theta_k$ véges sok paraméter értékét ismernénk - teljes egészében ismert lenne az eloszlás. Általában, ha eloszlását véges sok paraméter egyértelműen meghatározza, amit úgy fejezünk ki, hogy a paraméter valamely véges (k) dimenziós tér vagy annak egy részhalmaza, akkor paraméteres problémáról beszélünk.

Vannak statisztikai problémák, amelyeknél véges számú paraméter értékének ismerete nem teszi egyértelműen meghatározottá az eloszlást. Ha a 4. példában a beton törési szilárdságáról csak annyit tudunk, hogy az mint valószínűségeloszlás folytonos, akkor szóba jöhet az összes olyan folytonos eloszlás, amely a $(0, \infty)$ intervallumon értelmezett. A $H_0: F = G$ feltevés ekkor sokkal többet jelent, mint az F , ill. G eloszlásfüggvényben szereplő egy-két paraméter értékének megegyezése. A $(0, \infty)$ intervallumon értelmezett folytonos eloszlások halmazának bármely eleme szóba jöhet, csak arra vagyunk kíváncsiak, hogy a két beton esetében a fenti halmaz ugyanazon eleméről van-e szó. Ilyen esetben nem paraméteres problémáról beszélünk. A paraméteres problémákra vonatkozó statisztikai hipotézisek ellenőrzésére az ún. nem paraméteres módszerek szolgálnak, amelyeket a 3.5 pontban ismertetünk.

3.3.2 A statisztikai próba fogalma. Paraméteres próbák

Az előző pontban felsorolt 1-5. gyakorlati problémák mindegyikével kapcsolatban feltettünk egy kérdést, amelyre igenlő vagy tagadó válasz adható. Mindegyik feltett kérdéssel kapcsolatban megfogalmaztunk egy feltevést. Amikor pl. azt kérdezzük a panel hosszát jelölő X valószínűségi változóval kapcsolatban: igaz-e, hogy $M(X) = 300$ (cm) akkor ezt így is fogalmazhatjuk: igaz-e, helytálló-e a

$$H_0: M(X) = 300 \text{ (cm)}$$

hipotézis, az ún. nullhipotézis? Amikor a kérdésre választ adunk, minthogy válaszunk itt is statisztikai minta alapján történik, nem várhatunk biztos választ, döntünk tehát arról, hogy a H_0 hipotézist elfogadjuk, vagy visszautasítjuk. (Megjegyezzük, hogy a kérdéssel kapcsolatban abból indultunk ki: tapasztalatból tudjuk, hogy X normális eloszlású, ismert σ_0 szórással. Ez, mint látni fogjuk, a válaszadás szempontjából igen lényeges).

Az adott H_0 hipotézisről, annak elfogadásáról vagy elutasításáról láthatatlanban nem tudunk dönteni. Meg kell vizsgálnunk, hogy feltevésünket mennyire igazolja a tapasztalat. A példában szereplő hipotézis ellenőrzésére úgy járunk el, hogy megmérjük minél több találomra választott panel hosszát, statisztikai terminológiát használva X -re vonatkozó n elemű mintát veszünk. Legyen ennek kimenetele $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$. A mintából kiszámítjuk az $\bar{x}$ mintaközepet, amely az $M(X) = m$ várható érték torzítatlan becslése, és $\bar{x}$ numerikus értéke, valamint az $\bar{X}$ valószínűségi változó eloszlásának ismerete alapján döntünk arról, hogy elfogadjuk a H_0 hipotézist, vagy pedig elvetjük azt.

Nyilván inkább elfogadjuk H_0 -t, ha $\bar{x}$ értéke közel van az $M(X) = m$ értékhez, mintha távol esik tőle.

Az adott példában döntésünk tehát attól függ, hogy a H_0 hipotézisben szereplő $M(X) = m$ értékhez közel van vagy jelentősen (szignifikánsan) eltér az m várható érték becslésére használt $\bar{X}$ statisztika aktuális értéke.

Minden olyan esetben, amikor a hipotézis az eloszlás valamely paraméterére vonatkozik, a döntésünket annak alapján hozzuk meg, hogy az illető paraméter becslésére használt statisztika aktuális értéke közel van vagy távol esik a H_0 hipotézisben szereplő értéktől. Azt, hogy a statisztika értéke a feltételezett értékez elég közel van-e, vagy távol esik tőle, a szóban forgó statisztika eloszlása (elsősorban a szórása) alapján ítéltetjük meg, miként azt a továbbiakban részletesen látni fogjuk.

Definíció.

Azt az eljárást, amelynek alapján egy statisztikai hipotézisről döntünk, statisztikai próbának nevezzük. Amennyiben a H_0 hipotézist véges sok paraméter határozza meg, és így a kérdés az eloszlás valamely paraméterére vonatkozik, akkor ennek ellenőrzését paraméteres próbának nevezzük.

Minthogy a statisztikai próba alkalmazásakor döntésünket az $X_1, X_2, \dots, X_n$ mintára, illetve a mintából konstruált valamely $\alpha = \alpha(X_1, X_2, \dots, X_n)$ függvényre, tehát valószínűségi változóra alapozzuk, a véletlen játéka folytán hibás döntést is hozhatunk. Döntésünk során kétféle hibát követhetünk el.

Definíció.

Elsőfajú hibát követünk el, ha elvetjük a H_0 hipotézist, holott H_0 igaz (a felhasznált statisztika értéke a véletlen folytán távol esik az elméleti paraméter valódi értékétől). Másodfajú hibát követünk el, ha elfogadjuk a H_0 hipotézist, holott nem igaz (a H_1 alternatív hipotézis igaz, azonban statisztikánk értéke közel esik a H_0 -ban szereplő paraméterértékhez). A logikailag lehetséges döntéseket az alábbi táblázat tartalmazza:

	Ha a H_0 hipotézist	
	elfogadjuk	elutasítjuk
H_0 fennáll	helyes döntés	elsőfajú hiba
H_0 nem áll fenn	másodfajú hiba	helyes döntés

A statisztikai próbát úgy konstruáljuk, hogy sokszori alkalmazáskor ritkán eredményezzen téves döntést.

A továbbiakban néhány konkrét paraméteres próbát ismertetünk, amelyek kapcsán szemléltetjük a statisztikai próba konstruálásának és konkrét végrehajtásának módszerét. A próbákkal kapcsolatos elvi, elméleti kérdéseket főleg az elsőnek ismertetendő u-próba tárgyalása során fejtjük ki.

3.3.3 Az u-próba

Azt a H_0 hipotézist vizsgáljuk, hogy a normális eloszlású X valószínűségi változó $M(X)$ várható értéke adott m_0 számmal egyenlő, vagyis nullhipotézisünk:

$$H_0: M(X) = m_0.$$

Itt most az ún. kétoldali ellenhipotézissel szemben fogjuk H_0 -t vizsgálni, amikor is az ellenhipotézis (alternatíva)

$$H_1 : M(X) \neq m_0 ,$$

tehát vagy $m < m_0$, vagy $m > m_0$.

Tekintsük az X valószínűségi változóra vonatkozó, n független megfigyelésből álló

$$X_1, X_2, \dots, X_n$$

statisztikai mintát és számítsuk ki az $\bar{X}$ mintaközepet, amely a várható érték torzítatlan becslése. A H_0 hipotézis fennállása esetén $M(\bar{X}) = m_0$ és $D(\bar{X}) = \frac{\sigma_0^2}{n}$, ahol σ_0^2 az X valószínűségi változó ismertnek tekintett szórása. Mint már említettük, X ugyancsak normális eloszlású.

A normális eloszlás ismert tulajdonsága, az ún. 2 σ szabály alapján:

$$P(m_0 - 2 \frac{\sigma_0}{\sqrt{n}} < \bar{X} < m_0 + 2 \frac{\sigma_0}{\sqrt{n}}) \approx 0,95 .$$

Ezt átrendezve:

$$P(-2 < \sqrt{n} \frac{\bar{X} - m_0}{\sigma_0} < 2) \approx 0,95 .$$

Az $u = \sqrt{n} \frac{\bar{X} - m_0}{\sigma_0}$ valószínűségi változó (statisztika),

amelyet u -statisztikának nevezünk, az $\bar{X}$ mintaközép standardizáltja. Amennyiben H_0 igaz, az u -statisztika $N(0;1)$ eloszlású, tehát a $(-2,2)$ intervallumon 95 %-os konfidenciaintervallum számára. Ez azt jelenti, hogy H_0 fennállása esetében

u értéke nagy - 95 %-os - valószínűséggel a $(-2,2)$ intervallumba esik; és csak kicsiny - 5 %-os - valószínűséggel esik az intervallumon kívülre. Ezért kézenfekvő, hogy ha u akutális értékét a $(-2,2)$ intervallumon kívül találjuk, akkor a H_0 hipotézist elvetjük. Úgy is kifejezzük magunkat,

hogy ez esetben az $\bar{X}$ mintaközép megfigyelt értéke ($\bar{x}$) 95 %-os szinten szignifikánsan (lényegesen) eltér a H_0 hipotézisben szereplő m_0 értéktől. Ha pedig a $(-2 < u < 2)$ esemény következik be, nincs semmi indok a H_0 hipotézis elvetésére, nincs ellentmondás a H_0 hipotézis és a konkrét minta adatai között.

Az u-próba alapja tehát az

$$u = \sqrt{n} \frac{\bar{X} - m_0}{\sigma_0}$$

statisztika, amelyet próbastatisztikának nevezünk.

A $H_0 : M(X) = m_0$ hipotézis u-próba segítségével történő ellenőrzésekor természetesen nemcsak 95 %-os, hanem pl. 0,99 vagy akár 0,999 szinteken is dönthetünk. Általában tetszőleges $(1 - \alpha)$ szinthez ($\alpha > 0$) meghatározhatunk a normális eloszlásfüggvény táblázata alapján olyan u_α számot, amelyre:

$$P(-u_\alpha \leq u < u_\alpha | H_0) = 2\Phi(u_\alpha) - 1 = 1 - \alpha,$$

ahol Φ a standard normális eloszlásfüggvény.

Definíció.

A döntés szintjét jelentő $(1 - \alpha)$ számot a próba szignifikanciaszintjének, vagy röviden a próba szintjének nevezzük. A próba szintjét már a mintavétel előtt meghatározzuk, majd a mintából kiszámítjuk a megfigyelt értékek alapján az u-statisztika aktuális értékét.

Ezek után két eset van. Vagy $|u| > u_\alpha$, vagy $|u| \leq u_\alpha$.
a.) Az $|u| > u_\alpha$ esemény valószínűsége H_0 fennállása esetén

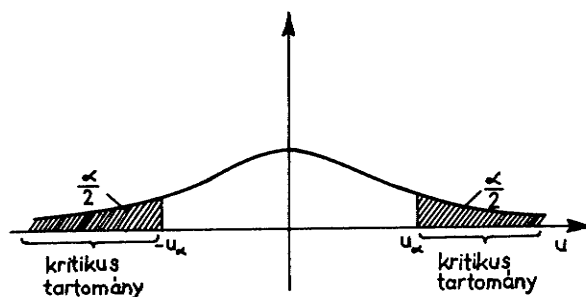
mindössze α (amely kicsiny, pl. 0,05; 0,01; 0,001).

Ha mégis az $|u| > u_\alpha$ esemény következik be, akkor arra következtethetünk, hogy helytelen a kiinduló H_0 hipotézis, tehát elvetjük a H_0 -t, mégpedig $(1 - \alpha)$ szignifikanciaszinten.

Definíció.

Valamely statisztikai próba kritikus tartománya a próbastatisztika mindazon értékeinek halmaza, amely értékek esetén a vizsgált H_0 hipotézist elvetjük.

Az u-próba esetében a kritikus tartomány - a $H_1 : m \neq m_0$ alternatíva - a számegyenesnek az $-u_\alpha$ számtól balra eső és az u_α számtól jobbra eső részei, vagyis a $\{u : |u| > u_\alpha\}$ halmaz (lásd: 40. ábra)



40. ábra

A statisztikai próba lényege a kritikus tartomány megválasztása. Ha ez megtörtént, és az u -statisztika aktuális értéke a kritikus tartományba esik, akkor H_0 -t elvetjük.

Amennyiben H_0 igaz, és az u -statisztika aktuális értéke mégis kritikus tartományba esik, aminek következtében H_0 -t elvetjük, akkor elsőfajú hibát követünk el. Az u_α érték fenti megválasztásából következik, hogy kétoldali u -próba esetében az elsőfajú hiba valószínűsége α -val egyenlő.
b.) Amennyiben az $|u| \leq u_\alpha$ esemény következik be, akkor a H_0 hipotézist elfogadjuk $(1 - \alpha)$ szignifikanciaszinten.

Definíció.

A próbastatisztika mindazon értékeinek halmazát, amelyek a vizsgált H_0 hipotézis elfogadását eredményezik, elfogadási tartománynak nevezzük.

Az u -próba alkalmazásakor előfordulhat, hogy az X valószínűségi változó várható értéke $M(X) = m_1 \neq m_0$, mégis az $|u| \leq u_\alpha$ esemény következik be, azaz a próbastatisztika értéke az elfogadási tartományba esik. Ekkor H_0 -t elfogadjuk, holott nem igaz, hanem a $H_1 : M(X) = m_1$ alternatív hipotézis igaz.

H_0 elfogadásával ekkor másodfajú hibát követünk el, amelynek valószínűségét β -val jelöljük. A következő tétel a másodfajú hiba valószínűségének kiszámítására vonatkozólag ad útmutatást.

Tétel.

Ha a $H_0 : M(X) = m_0$ hipotézis ellenőrzésére α elsőfajú hibájú u -próbát alkalmazunk, és $H_1 : M(X) = m_1$ alternatív hipotézis igaz, akkor a másodfajú hiba valószínűsége:

$$\beta = P(|u| < u_\alpha \mid H_1) = \Phi\left(\sqrt{n} \frac{m_0 - m_1}{\sigma_0} + u_\alpha\right) - \Phi\left(\sqrt{n} \frac{m_0 - m_1}{\sigma_0} - u_\alpha\right).$$

ahol Φ a standard normális eloszlásfüggvényt jelöli.

Bizonyítás.

$$\begin{aligned} P(|u| < u_{\alpha} \mid H_1) &= P(-u_{\alpha} < u < u_{\alpha} \mid H_1) = \\ &= P(-u_{\alpha} < \sqrt{n} \frac{\bar{X} - m_0}{\sigma_0} < u_{\alpha} \mid H_1) = \\ &= P(m_0 - u_{\alpha} \frac{\sigma_0}{\sqrt{n}} < \bar{X} < m_0 + u_{\alpha} \frac{\sigma_0}{\sqrt{n}} \mid H_1). \end{aligned}$$

Ez a valószínűség nem változik, ha az utolsó zárójelen belül az egyenlőtlenség minden tagjából levonunk m_1 -et, és minden tagot elosztunk $\frac{\sigma_0}{\sqrt{n}}$ -nel, amelynek végrehajtásával:

$$\begin{aligned} P(|u| < u_{\alpha} \mid H_1) &= P\left(\sqrt{n} \frac{m_0 - m_1}{\sigma_0} - u_{\alpha} < \right. \\ &< \left. \sqrt{n} \frac{\bar{X} - m_1}{\sigma_0} < \sqrt{n} \frac{m_0 - m_1}{\sigma_0} + u_{\alpha} \mid H_1\right). \end{aligned}$$

Mivel a jobb oldalon lévő egyenlőtlenség középső tagja H_1 fennállása esetén $N(0;1)$ eloszlású, állításunkat igazoltuk.

Ha $n \rightarrow \infty$, akkor $\sqrt{n} \rightarrow \infty$, és $m_0 - m_1 < 0$ esetén

$$\lim_{n \rightarrow \infty} P(|u| < u_{\alpha} \mid H_1) = \Phi(-\infty) - \Phi(-\infty) = 0 - 0 = 0;$$

$m_0 - m_1 > 0$ esetén

$$\lim_{n \rightarrow \infty} P(|u| < u_{\alpha} \mid H_1) = \Phi(\infty) - \Phi(\infty) = 1 - 1 = 0;$$

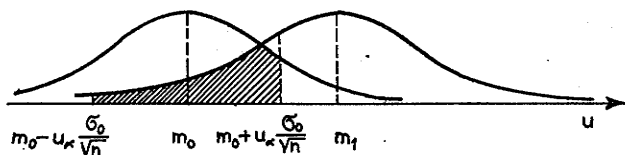
Tehát akár $m_1 > m_0$, akár $m_1 < m_0$, ha $n \rightarrow \infty$, akkor a másodfajú hiba valószínűsége nullához tart.

Definíció.

Az olyan próbát, amelynél a minta elemszámának minden határon túl történő növekedésével a másodfajú hiba minden rögzített alternatív hipotézis esetén 0-hoz tart, konzisztens próbának nevezzük.

Az u-próba tehát konzisztens próba. Ugyanezzel a tulajdonsággal rendelkezik a könyvünkben tárgyalt többi paraméteres próba is.

Megjegyezzük, hogy az $\bar{X}$ statisztika sűrűségfüggvénye segítségével adott $H_1 : M(X) = m_1$ alternatíva hipotézis esetében a másodfajú hiba jól szemléltethető, miként azt a 41. ábra mutatja.

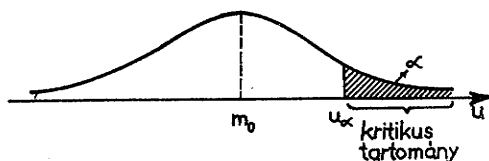


41. ábra

Az eddigiekben az u -próbát egyszerű nullhipotézis és egyszerű alternatív hipotézis esetében vizsgáltuk. Vizsgáljuk most az összetett hipotézisek esetét. Legyen X normális eloszlású valószínűségi változó, ismert σ_0 szórással. Állítsuk fel a

$H_0 : M(X) = m \leq m_0$ hipotézist, és legyen $H_1 : m > m_0$.

Ekkor az u kritikus érték megválasztásakor a legrosszabb esetet kell figyelembe venni, vagyis az $m = m_0$ esethez szabjuk az $\{u > u_\alpha\}$ kritikus tartományt (lásd. 42. ábra).



42. ábra

Ez az adott esetre azt fogja eredményezni, hogy $m < m_0$ fennállása esetén az elsőfajú hiba valószínűsége, vagyis H_0 téves elutasításának valószínűsége, kisebb lesz, mint α . Összetett hipotézis esetében tehát az elsőfajú hiba valószínűsége más és más, aszerint, hogy mennyi $M(X)$ igazi értéke.

Definíció.

Összetett hipotézis esetén az elsőfajú hiba elkövetésének legnagyobb valószínűségét a próba terjedelmének nevezzük.

A fenti összetett hipotézispárra az ún. egyoldali u -próba esetében kritikus tartománynak az u tengely (számegyenes) $\{u : u > u_\alpha\}$ tartományát választjuk.

Meg lehet mutatni, hogy az α -terjedelmű kritikus tartomány ily módon történő választása mellett legkisebb a másodfajú hiba elkövetésének valószínűsége, bármely $m > m_0$ paraméterértékre, vagyis a fenti hipotézispárra az összes

α -terjedelmű u-próbák közül ez a "legjobb" próba.

Az u-próba kétmintás változatát akkor alkalmazzuk, ha tudjuk, hogy az X és Y valószínűségi változók normális eloszlásúak $D(X) = \sigma_1^2$ és $D(Y) = \sigma_2^2$ ismert szórásokkal, és vizsgáljuk a

$$H_0 : M(X) = M(Y)$$

hipotézist a

$$H_1 : M(X) \neq M(Y)$$

alternatív hipotézissel szemben.

Tegyük fel, hogy az X változóra vonatkozólag a $X_1, X_2, \dots, X_n$ az Y-ra az $Y_1, Y_2, \dots, Y_m$ független mintákkal rendelkezünk. (A két minta elemszáma különböző lehet). Mindkét mintából kiszámítjuk a mintaközepet és megalkotjuk az

$$u = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

próbastatisztikát. Ebben az esetben a fenti u-statisztika ún. kétmintás u-statisztika.

Ha a H_0 hipotézis igaz, akkor próbastatisztikánk $N(0,1)$ eloszlású. Az $(1 - \alpha)$ szinten történő döntéshez a normális eloszlásfüggvény táblázata alapján meghatározzuk azt az u_α számot, amelyre

$$P(-u_\alpha \leq u < u_\alpha) = 2\Phi(u_\alpha) - 1 = 1 - \alpha.$$

Megmutatható, hogy ha kritikus tartománynak az $[-u_\alpha, u_\alpha]$ intervallum komplementer halmazát választjuk, akkor ez az origóra szimmetrikus helyzetű tartomány adja a legjobb $(1 - \alpha)$ szintű próbát.

Példa.

Tegyük fel, hogy házgyárban gyártanak azonos típusú panelt, melyről tudjuk, hogy bizonyos mérete - mondjuk a hosszabbik oldalának a hossza - normális eloszlású. Az egyik gyárban gyártott panel hosszabbik oldalának hosszát jelöljük X-szel, a másik gyárban gyártottakét Y-nal. Tudjuk, hogy $D(X) = \sigma_1^2 = 0,6$ (cm), $D(Y) = \sigma_2^2 = 0,8$ (cm).

Vizsgáljuk a $H_0 : M(X) = M(Y)$ hipotézist a $H_1 : M(X) \neq M(Y) \neq M(Y)$ ellenhipotézissel szemben, $1 - \alpha = 0,95$ szignifikanciaszinten. Minthogy $\alpha = 0,05$; a normális eloszlás táblázata alapján $u_{0,05} = 1,96$ adódik.

Megmérjük az első gyárban - találomra választva - 100 db panel hosszát, és kiszámítjuk a mintaközepet, amelyre $\bar{x} = 300,1$ (cm) adódik. A második gyárban is kiválasztunk 100 db panelt, ezek hosszának mintaközepe legyen $\bar{y} = 299,8$ (cm).

Kiszámítjuk az

$$u = \frac{300,1 - 299,8}{\sqrt{\frac{0,36}{100} + \frac{0,64}{100}}} = \frac{0,3}{\sqrt{\frac{1}{100}}} = 3$$

statisztikát. Minthogy az u-statisztika numerikus értéke kívül esik a $[-1,96, 1,96]$ intervallumon, vagyis a kritikus tartományba esik, H_0 -t 95 %-os szinten el kell vetnünk.

3.3.4 A próba ereje. Az erőfüggvény

Az előzőekben láttuk, hogy a másodfajú hiba β valószínűségét nem tudjuk egyszerűen meghatározni, mert ez a valószínűség

$$= P(-u_\alpha \leq u \leq u_\alpha \mid H_1),$$

a kritikus tartomány megválasztása után attól függ, hogy milyen H_1 alternatív hipotézis igaz, más szóval, ha megmondjuk, hogy mennyi a szóban forgó normális eloszlású valószínűségi változó $M(X) = m$ tényleges várható értéke. Vizsgálhatjuk a másodfajú hiba helyett a kiegészítő esemény valószínűségét, vagyis annak valószínűségét, hogy az u-statisztika értéke a kritikus tartományba esik - tehát nem követünk el másodfajú hibát, feltéve, hogy a $H_1 : M(X) = m$ hipotézis igaz. Ez a valószínűség

$$1 - \beta = P(|u| \geq u_\alpha \mid H_1) = P(|u| \geq u_\alpha \mid m),$$

amely nyilván ugyancsak m-től, a tényleges várható értéktől függ.

Definíció.

Az $1 - \beta = P(|u| > u_\alpha \mid m)$ valószínűséget mint m függvényét a próba erőfüggvényének nevezzük. Az erőfüggvény értéke mindig rögzített m-re megadja a próba erejét.

Jelöljük az erőfüggvényt $E(m)$ -mel. Ezek szerint:

$$E(m) = 1 - \beta = P(|u| > u_{\alpha} \mid m).$$

Könnyen belátható, hogy ha α terjedelmű próbát alkalmazunk, akkor

$$E(m_0) = P(|u| \geq u_{\alpha} \mid m_0) = \alpha,$$

hiszen annak valószínűsége, hogy az u -statisztika értéke a kritikus tartományba esik, vagyis H_0 -t elutasítjuk, feltéve, hogy $H_0 : M(X) = m_0$ igaz, éppen az elsőfajú hiba valószínűsége.

Valamely próba alkalmazásakor természetesen az lenne az ideális, ha mind az elsőfajú hiba α valószínűsége, mind a másodfajú hiba β valószínűsége nagyon kicsi lenne. Sajnos ilyen ideális próbák általában csak igen nagy mintanagyságra várhatók.

Az elsőfajú hiba valószínűségét, a próba terjedelmét α -nak választva nyilván az a kívánatos, hogy β minél kisebb, illetve az $1 - \beta = E(m)$ valószínűség minél nagyobb legyen, hiszen ha $m \neq m_0$, akkor $E(m)$ éppen a helyes döntés valószínűsége.

Láttuk, hogy a próba α terjedelmének megválasztása egyben meghatározza az erőfüggvény értékét az $m = m_0$ pontban, mivel $E(m_0) = \alpha$, ezt pedig általában kicsinek (pl. $\alpha = 0,05$) választjuk.

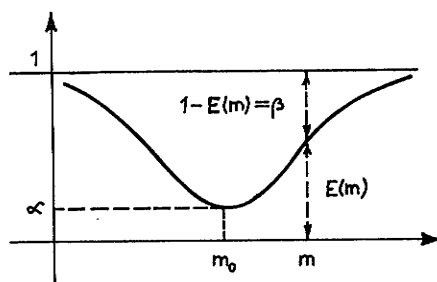
Az u -próba esetében az $E(m)$ erőfüggvény explicit alakban megadható:

$$\begin{aligned} \beta &= P(|u| \leq u_{\alpha} \mid m) = \Phi \left(\sqrt{n} \frac{m_0 - m}{\sigma_0} + u_{\alpha} \right) - \\ &= \Phi \left(\sqrt{n} \frac{m_0 - m}{\sigma_0} - u_{\alpha} \right), \end{aligned}$$

tehát

$$\begin{aligned} E(m) &= 1 - \beta = 1 - \left[\Phi \left(\sqrt{n} \frac{m_0 - m}{\sigma_0} + u_{\alpha} \right) - \right. \\ &\quad \left. - \Phi \left(\sqrt{n} \frac{m_0 - m}{\sigma_0} - u_{\alpha} \right) \right]. \end{aligned}$$

A formulából látható, hogy E az m -nek folytonos függvénye, ennél fogva értéke az m_0 -nak nem nagy környezetében általában kicsi (1.43. ábra). Az erőfüggvénnyel szemben azt a követelményt szoktuk támasztani, hogy értéke valamely $(m_0 - \Delta ; m_0 + \Delta)$ intervallumon kívül érjen el a konkrét feladatnak megfelelő elég nagy értéket. Ezt n elég nagy értékben való megválasztásával érhetjük el vagy σ_0 csökkenésével.



43. ábra

3.3.5 A t-próba (Student-próba)

A 3.3.3 pontban ismertetett u-próba abban az esetben alkalmazható, amikor a normális eloszlású valószínűségi változó vagy változók szórása ismert, és a várható értékre vonatkozó hipotézist akarjuk ellenőrizni. Minthogy a gyakorlatban a szórást legtöbbször nem ismerjük, ezért az u-próba alkalmazására viszonylag ritkán kerülhet sor. Ha a várható értékre vonatkozó egyszerű hipotézist akarjuk ellenőrizni, és csak annyit tudunk, hogy az X valószínűségi változó normális eloszlású, de szórását nem ismerjük, akkor a szórásnégyzetet az X -re vonatkozó $X_1, X_2, \dots, X_n$ statisztikai mintából becsüljük az

$$s_n^{x2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

korrigált empirikus szórásnégyzettel.

A $H_0 : M(X) = m_0$ nullhipotézis vizsgálatára konstruáljuk a

$$t_{n-1} = \frac{\bar{X} - m_0}{S_n^*} \sqrt{n}$$

próbastatisztikát, amelyet t-statisztikának nevezünk.

Kimutatható, hogy a t-statisztika eloszlásfüggvénye:

$$\forall x \in R, S_{n-1}(x) = P(t_{n-1} < x) =$$

$$= \frac{1}{\sqrt{(n-1)\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \int_{-\infty}^x \frac{du}{(1 + \frac{u^2}{n-1})^{\frac{n}{2}}};$$

sűrűségfüggvénye pedig:

$$S_{n-1}(x) = \frac{1}{\sqrt{(n-1)\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} (1 + \frac{x^2}{n-1})^{-\frac{n}{2}}$$

A $S_{n-1}(x)$ eloszlásfüggvényű, ill. $S_{n-1}(x)$ sűrűségfüggvényű valószínűségeloszlást $(n-1)$ paraméterű Student-eloszlásnak nevezzük. Ha ellenhipotézisünk a

$$H_1 : M(X) = m \neq m_0,$$

akkor a t-próbánál is szimmetrikus tartományt alkalmazunk. A Függelék VI. táblázatából tetszőleges $\varepsilon > 0$ számhoz található olyan t_ε érték, amelyre:

$$P(-t_\varepsilon < t_{n-1} < t_\varepsilon | H_0) = 1 - \varepsilon.$$

(A gyakorlatban $\varepsilon = 0,05$, $\varepsilon = 0,01$ vagy $\varepsilon = 0,001$).

Ha pl. a mintaelemszám $n = 11$, vagyis $n - 1 = 10$ és $\varepsilon = 0,05$, akkor a Függelék VI. táblázata alapján $t_\varepsilon = 2,228$,

$$P(-2,228 < t_{10} < 2,228 | H_0) = 0,95$$

Ha viszont a mintaelemszám $n = 31$, tehát $n - 1 = 30$ és $\varepsilon = 0,05$, akkor a VI. táblázat alapján

$$P(-2,042 < t_{30} < 2,042 | H_0) = 0,95.$$

A t_{n-1} statisztika $n - 1$ paraméterét (eggyel csökkentett mintaelemszámot) szabadságfoknak is szokás nevezni.

A t -eloszlás sűrűségfüggvényének grafikonja nagy n -re a normális eloszlás sűrűségfüggvényét közelíti, és $n \rightarrow \infty$ esetén meg is egyezik vele. Kimutatható ugyanis, hogy

$$\lim_{n \rightarrow \infty} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \frac{1}{\sqrt{n-1}} = \frac{1}{\sqrt{2}} ;$$

továbbá

$$\lim_{n \rightarrow \infty} (1 + \frac{x^2}{n-1}) = e^{-\frac{x^2}{2}} ,$$

tehát

$$\lim_{n \rightarrow \infty} s_{n-1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} ,$$

vagyis a végtelen szabadságfokú Student-eloszlás a normális eloszlással egyezik meg. Ezért ha $n > 80$, akkor a Student-eloszlás táblázata helyett már a normális eloszlás táblázata használható a t -próba; ekkor tehát már nem t -próba, hanem u -próba, de már $n > 30$ esetén sem nagy az eltérés.

A $H_0 : M(X) = m_0$ nullhipotézissel szembeni egyoldali alternatíva:

$$H'_1 : M(X) = m > m_0 .$$

Erre az alternatív hipotézisre vonatkozólag a kritikus tartomány a $\{t_{n-1} \geq t_\varepsilon\}$ részhalmaza a számegyenesnek, ahol t_ε értékét a

$$P(t_{n-1} < t_\varepsilon \mid H_0) = 1 - \varepsilon$$

összefüggés adja. A t_ε értéket adott $\varepsilon > 0$ választás mellett ugyancsak a VI. táblázatból határozhatjuk meg.

Ha két normális eloszlású valószínűségi változó várható értékét akarjuk összehasonlítani, és feltehetjük, hogy a két változó szórása megegyezik (de egyébként a szórás ismeretlen számunkra), akkor az ún. kétmintás t -próba alkalmazható. (Ha a két szórás egyezésére vonatkozólag nincs előzetes tapasztalatunk, akkor a szórások egyezését F -próbával kell ellenőriznünk, amelyet a 3.3.6 pontban ismertetünk).

Legyen X és Y normális eloszlású független valószínűségi változók, amelyek egyforma szórásúak, azaz $D(X) = D(Y)$ és vizsgáljuk a

$$H_0 : M(X) = M(Y)$$

nullhipotézist.

Legyenek az X , ill. Y változókra vonatkozó minták:

$$X_1, X_2, \dots, X_n \quad \text{ill.} \quad Y_1, Y_2, \dots, Y_n.$$

Képezzük az alábbi statisztikát:

$$t_{n+m-2} = \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_n^2 + (m-1)S_m^2}} \sqrt{\frac{nm(n+m-1)}{n+m}}$$

A t_{n+m-2} statisztika eloszlására vonatkozólag kimutatható, hogy amennyiben a H_0 hipotézis igaz, vagyis a két normális eloszlású valószínűségi változó egyforma várható értékű, akkor a t_{n+m-2} statisztika $n+m-2$ paraméterű (szabadságfokú) Student-eloszlást követ.

Ha az ellenhipotézisünk

$$H_1 : M(X) \neq M(Y),$$

akkor - az egymintás esethez hasonlóan - szimmetrikus kritikus tartományt választunk, azaz a VI. táblázatból megkeresük az adott ε -hoz azt a t_ε értéket, amelyre

$$P(-t_\varepsilon \leq t_{n+m-2} \leq t_\varepsilon) = 1 - \varepsilon$$

Ha a kiszámított t_{n+m-2} statisztika értéke nem esik a $[-t_\varepsilon, t_\varepsilon]$ intervallumba, akkor a H_0 hipotézist $1 - \varepsilon$ szinten elvetjük. Ha ellenhipotézisünk a $H'_1 : M(X) > M(Y)$ alternatív hipotézis, akkor a kritikus tartomány az egymintás t -próbához hasonlóan a $\{t_{n+m-2} \geq t_\varepsilon\}$ halmaz, ahol t_ε az a szám, amelyre $P(t_{n+m-2} < t_\varepsilon \mid H_0) = 1 - \varepsilon$.

3.3.6 Az F-próba

Annak eldöntésére, hogy két normális eloszlású valószínűségi változó szórása megegyezik-e, az F-próbát alkalmazzuk.

Legyenek X és Y független normális eloszlású valószínűségi változók, amelyek várható értékét most figyelmen kívül hagyjuk. A várható értékek tehát különbözők lehetnek.

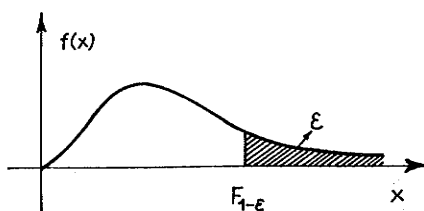
Vizsgáljuk a

$$H_0 : D(X) = D(Y)$$

nullhipotézist a két változóra vonatkozó $X_1, X_2, \dots, X_n$, ill. $Y_1, Y_2, \dots, Y_m$ független minták alapján. A próbatisztika, amelynek alapján döntést hozunk, az empirikus (korrigált) szórásnégyzetek hányadosa:

$$F = F_{f_1, f_2} = \frac{S_n^2}{S_m^2}$$

ahol $f_1 = n-1$, $f_2 = m-1$. Az F_{f_1, f_2} statisztika ún. F -eloszlást követ $(n-1)$, $(m-1)$ szabadságfokkal. Az eloszlás sűrűségfüggvényét a 44. ábra érzékelteti.



44. ábra

Egyoldali ellenhipotézisünk:

$$H_1 : D(X) > D(Y).$$

Ha $S_n^2 < S_m^2$, akkor a H_1 ellenhipotézist már elutasítjuk, hiszen az X -re vonatkozó minta empirikus szórásnégyzete kisebb, mint az Y -ra vonatkozó minta empirikus szórásnégyzete, ez tehát nem támasztja alá H_1 -et.

Ha $S_n^2 > S_m^2$, vagyis $F > 1$, akkor az $1 - \epsilon$ szint megválasztása után az F -eloszlás táblázatából (lásd. Függelék VII. táblázatát) az $f_1 = n - 1$, $f_2 = m - 1$ szabadságfokhoz tartozó $F_{1-\epsilon}$ értékkel hasonlítjuk össze F értékét, és $F > F_{1-\epsilon}$ esetén elutasítjuk H_0 -t, ellenkező esetben H_0 elvetésére nincs okunk (de csak nagy minta esetében ajánlatos elfogadni).

Kétoldali ellenhipotézis esetén, amikor tehát az alternatíva:

$$H_1 : D(X) \neq D(Y) ,$$

úgy járunk el, hogy a nagyobb empirikus szórásnégyzetet osztjuk a kisebbel, vagyis képezzük az

$$F^* = \max \left(\frac{s_n^2}{s_m^2} , \frac{s_m^2}{s_n^2} \right) > 1$$

statisztikát. Az $1 - \varepsilon$ szint megválasztása után F^* értékét hasonlítjuk össze az $F_{1-\varepsilon}$ értékkel. A táblázat használata során ügyelni kell arra, hogy f_1 mindig a számlálóhoz, f_2 pedig a nevezőhöz (kisebb empirikus szórással) tartozó minta elemszámának eggyel csökkentett értéke. $F^* > F_{1-\varepsilon}$ esetén H_0 -t elutasítjuk ($1 - \varepsilon$) szinten. Ellenkező esetben és nagy minta esetén H_0 -t elfogadjuk.

3.3.7 A Welch-próba

Ha két normális eloszlású független valószínűségi változó, X és Y várható értékét akarjuk összehasonlítani, amelyek ismeretlen, de különböző szórással (pl. Az F -próba a $D(X) \neq D(Y)$ eredményre vezetett), akkor kétmintás t -próba nem alkalmazható.

A $H_0 : M(X) = M(Y)$ hipotézis ellenőrzésére B.L. Welch a következő próbastatisztikát javasolta:

$$t_f = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_n^2}{n} + \frac{s_m^2}{m}}}$$

ahol $\bar{X}$ és $\bar{Y}$ az $X_1, X_2, \dots, X_n$, ill. $Y_1, Y_2, \dots, Y_m$ mintából számított mintaközépek; s_n^2 , ill. s_m^2 az empirikus szórásnégyzetek.

Az f paraméter értékét a következő módon számítjuk ki.
Legyen

$$c = \frac{\frac{S_m^2}{m}}{\frac{S_n^2}{n} + \frac{S_m^2}{m}}$$

akkor

$$f = \frac{(n-1)(m-1)}{(n-1)c^2 + (m-1)(1-c^2)}$$

Ha a H_0 hipotézis igaz, akkor a t_f statisztika közelítőleg f szabadságfokú Student-eloszlású, vagyis adott $\varepsilon > 0$ szinthez a VI. táblázat f -edik sorában található a megfelelő t_ε kritikus érték.

Ha mindkét minta elemszáma elég nagy (pl. $n, m \geq 40$), akkor a t_f statisztika közelítőleg normális eloszlású, vagyis a megfelelő t_ε értékek normális eloszlás táblázata alapján kereshetők ki.

3.3.8 A Bartlett-próba

Mint a 3.3.6 pontban láttuk, azt a kérdést, hogy két normális eloszlású valószínűségi változó szórása megegyezik-e az F-próba segítségével dönthetjük el. A gyakorlatban sokszor felmerül annak szükségessége, hogy egyidejűleg kettőnél több normális eloszlású valószínűségi változó szórását hasonlítsuk össze, erre használjuk a Bartlett-próbát.

Legyenek $X^{(1)}, X^{(2)}, \dots, X^{(k)}$ összesen k darab normális eloszlású valószínűségi változó, és vizsgáljuk a

$$H_0 : D(X^{(1)}) = D(X^{(2)}) = \dots = D(X^{(k)})$$

nullhipotézist a

$$X^{(1)} : x_{11}, x_{12}, \dots, x_{1n_1}$$

$$X^{(2)} : x_{21}, x_{22}, \dots, x_{2n_2}$$

⋮

$$X^{(k)} : x_{k1}, x_{k2}, \dots, x_{kn_k}$$

statisztikai minták alapján.

Az egyes valószínűségi változókra vonatkozó nagysága tehát különböző lehet, de kikötjük, hogy $n_i \geq 4$ legyen, és $(i=1,2,\dots,k)$.

A H_1 ellenhipotézis most abból áll, hogy nem mindegyik szóban forgó változó szórása egyezik meg.

A próba végrehajtását azzal kezdjük, hogy mindegyik mintából kiszámítjuk a korrigált empirikus szórásnégyzetet, legyenek ezek:

$$s_{n_1}^{*2}, s_{n_2}^{*2}, \dots, s_{n_k}^{*2}.$$

Ezután kiszámítjuk a nyert empirikus szórásnégyzetek súlyozott számtani közepét a következő módon: Vezessük be az

$f_i = n_i - 1$ jelölést $(i=1,2,\dots,k)$, és képezzük a $\sum_{i=1}^k f_i = f$ összeget. Számítsuk ki az $s^{*2} = \frac{1}{f} \sum_{i=1}^k f_i s_{n_i}^{*2}$ súlyozott közepet.

M.S.Bartlett kimutatta, hogy a

$$K^2 = \frac{2,3026}{c} (f \lg s^{*2} - \sum_{i=1}^k f_i \lg s_{n_i}^{*2})$$

statisztika közelítőleg χ^2 -eloszlást követ $k - 1$ paraméterrel. A c állandót az alábbi módon számítjuk:

$$c = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{f_i} - \frac{1}{f} \right).$$

Ezután választunk egy $\varepsilon > 0$ szintet (pl. $\varepsilon = 0,05$) és amennyiben a K^2 mennyiség nagyobb, mint a χ^2 -eloszlás táblázatának (Függelék V. táblázat) $k - 1$ -edik sorában az $(1 - \varepsilon)$ szinthez tartozó χ_{k-1}^2 kritikus érték, akkor a H_0 hipotézist $(1 - \varepsilon)$ szinten elvetjük.

3.3.9 Szekvenciális módszer hipotézisvizsgálatra

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.

Wald Ábrahám által kidolgozott szekvenciális módszer lehetővé teszi, hogy amikor becslésméleti vagy hipotézisvizsgálati problémákról van szó, lényegesen kevesebb mintaelemről ugyanolyan biztonsággal vonhassunk le a 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 hagyunk 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 X 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 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 B számok a szekvenciális próba szempontjából alapvető határmennyiségek lesznek, nevezetesen ezek a számok a számegyenesét három szakaszra, tartományra osztják.

elfogadási tartomány	indifferens tart.	kritikus tart.
$\xleftarrow{\quad}$	$\xleftarrow{\quad}$	$\xleftarrow{\quad}$
$\xrightarrow{\quad}$	$\xrightarrow{\quad}$	$\xrightarrow{\quad}$
$\xrightarrow{\quad}$	$\xrightarrow{\quad}$	$\xrightarrow{\quad}$

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; az X valószínűségi változóra egymás után végzünk megfigyeléseket. Legyen az első megfigyelés eredménye, az első mintaelem x_1 . Kiszámítjuk az $f(x_1; \theta_0)$ és az $f(x_1; \theta_1)$ értéket és képezzük az

$$l_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)}$$

hányadost.

Ha pl. X normális eloszlású valószínűségi változó ismert szórással és $H_0 : M(X) = 0$, $H_1 : M(X) = 1$, akkor

$$l_1 = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1 - 1)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}}}$$

Amennyiben

$$l_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)} \geq A,$$

a H_0 hipotézist máris elvetjük, mert l_1 értéke az elvetési tartományba esett. Ha

$$l_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)} \leq B,$$

akkor a H_0 -t elfogadjuk, mert az l_1 értéke az elfogadási tartományba esett. Ha viszont

$$B < l_1 = \frac{f(x_1; \theta_1)}{f(x_1; \theta_0)} < A$$

aakor 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

$$l_2 = \frac{f(x_2; \theta_1)}{f(x_2; \theta_0)}$$

hányadost, majd képezzük az

$$L = l_1 l_2 = \frac{f(x_1; \theta_1) f(x_2; \theta_1)}{f(x_1; \theta_0) f(x_2; \theta_0)}$$

menyiséget és a döntést illetően ugyanúgy 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

$$L = l_1 l_2 l_3 = \frac{f(x_1; \theta_1) f(x_2; \theta_1) f(x_3; \theta_1)}{f(x_1; \theta_0) f(x_2; \theta_0) f(x_3; \theta_0)} =$$

$$= \frac{\sum_{i=1}^3 f(x_i; \theta_1)}{\sum_{i=1}^3 f(x_i; \theta_0)}$$

hányadost, amelyet likelihood-hányadosnak nevezünk.

Általában a k -edik mintaelem megfigyelése után (ha az első $k-1$ számú megfigyelés nem vezetett végső döntéshez), az

$$L = l_1 l_2 \dots l_k = \frac{\prod_{i=1}^k f(x_i; \theta_1)}{\prod_{i=1}^k f(x_i; \theta_0)} = \frac{p_{1k}}{p_{0k}}$$

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 hányadost.

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 X diszkrét eloszlású, akkor az $f(x_i; \theta_1)$ jelentése a $P(X = x_i | H_1)$, az $f(x_i; \theta_0)$ pedig a $P(X = x_i | H_0)$ valószínűséget.

Ekkor $\prod_{i=1}^k f(x_i; \theta_1)$ az egyes mintaelemek együttes bekövetkezésének, vagyis az adott mintának a valószínűsége azzal a felté-

tellel, 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 az X folytonos eloszlású. Ez esetben a kérdéses valószínűségek hányadosát (mivel $P(X = x_i | H_1) \approx f(x_i; \theta_1) \Delta x_i$)

$$\frac{\prod f(x_i; \theta_1) \Delta x_i}{\prod f(x_i; \theta_0) \Delta x_i}$$

alakban kellene írunk, azonban a Δx_i tényezőktől el szoktunk tekinteni. Az L jelentése alapján a fent leírt próbát valószínűségi hányadospróbának nevezzük.

A próba végrehajtása alkalmával az egyes $l_i = \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)}$

mennyiségek helyett ezek logaritmusait használjuk, mert ekkor szorzás helyett összeadással dolgozhatunk.

A $z_i = \log l_i = \log \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)}$ 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 hoztunk. 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 az X normális eloszlású változó $\sigma^2 = 1$ szórásnégyzettel, ismeretlen θ várható értékkel. Legyen továbbá

$$H_0 = \theta = \theta_0, \quad H_1 : \theta = \theta_1$$

(ahol $\theta_1 > \theta_0$).

Ekkor

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$$

$$L = \frac{p_{1k}}{p_{0k}} = \frac{\prod_{i=1}^k e^{-\frac{1}{2}(x_i - \theta)^2}}{\prod_{i=1}^k e^{-\frac{1}{2}(x_i - \theta_0)^2}}$$

$$z_i = \log \frac{e^{-\frac{1}{2}(x_i - \theta_1)^2}}{e^{-\frac{1}{2}(x_i - \theta_0)^2}} = \log e^{\frac{1}{2}[(x_i - \theta_0)^2 - (x_i - \theta_1)^2]}$$

$$\begin{aligned} z_i &= \log e^{x_i(\theta_1 - \theta_0) + \frac{1}{2}(\theta_0^2 - \theta_1^2)} = \\ &= \frac{1}{2}(\theta_0^2 - \theta_1^2) + (\theta_1 - \theta_0)x_i \end{aligned}$$

A z_i változók összegét képezve:

$$z_1 + z_2 + \dots + z_k = (\theta_1 - \theta_0)[(x_1 + x_2 + \dots + x_k) - \frac{k}{2}(\theta_1 + \theta_0)]$$

$$\begin{aligned} \log B &= \log \frac{\beta}{1 - \alpha} z_1 + z_2 + \dots + z_k < \log \frac{1 - \beta}{\alpha} = \\ &= \log A \end{aligned}$$

egyenlőtlenség tehát most ekvivalens az

$$\frac{1}{\theta_1 - \theta_0} \log \frac{\beta}{1 - \alpha} + \frac{k}{2} (\theta_1 + \theta_0) < \sum_1^k x_i < \frac{1}{\theta_1 - \theta_0} \log \frac{1 - \beta}{\alpha} + \frac{k}{2} (\theta_1 + \theta_0)$$

Konkrét esetben tegyük fel, hogy $\alpha = 0,05$, $\beta = 0,10$ első-, illetve másodfajú hibát engedtünk meg (gyakorlati megfontolások alapján, a vizsgált ismerv természetétől és a különféle fajtajú 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 : $\theta = 10$ feltevést fogadjuk el. Ha az (I) vagy a (II)

esetek egyike sem fordul elő, akkor folytatjuk a megfigyelés-sorozatot.

Másik példaként nézzük meg a binomiális eloszlás esetét. Az X legyen binomiális eloszlású valószínűségi változó ismeretlen paraméterrel. Itt tehát a kísérletek olyan sorozatáról van szó, amelyben minden egyes kísérletnek 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

$$P(X = 1) = f(1, p) = p, \quad P(X = 0) = f(0, p) = 1 - p$$

Legyen a H_0 hipotézis: $p = p_0$, a H_1 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

$$\begin{aligned} z_1 + z_2 + \dots + z_n &= K_n \log \frac{p_1}{p_0} + (n - K_n) \log \frac{1 - p_1}{1 - p_0} = \\ &= n \log \frac{1 - p_1}{1 - p_0} + K_n \left(\log \frac{p_1}{p_0} - \log \frac{1 - p_1}{1 - p_0} \right) = \\ &= n \log \frac{1 - p_1}{1 - p_0} + K_n \left(\log \frac{p_1}{p_0} \cdot \frac{1 - p_0}{1 - p_1} \right) \end{aligned}$$

A kísérletsorozatot mindaddig folytathatjuk, amíg fennáll, hogy

$$\log B \quad z_1 + z_2 + \dots + z_n < \log A$$

más szóval amíg

$$\begin{aligned} & \frac{\log B}{\log \frac{p_1}{p_0} \cdot \frac{1 - p_0}{1 - p_1}} + n \frac{\log \frac{1 - p_0}{1 - p_1}}{\log \frac{p_1}{p_0} \cdot \frac{1 - p_0}{1 - p_1}} < K_n < \\ & < \frac{\log A}{\log \frac{p_1}{p_0} \cdot \frac{1 - p_0}{1 - p_1}} + n \frac{\log \frac{1 - p_0}{1 - p_1}}{\log \frac{p_1}{p_0} \cdot \frac{1 - p_0}{1 - p_1}} \end{aligned}$$

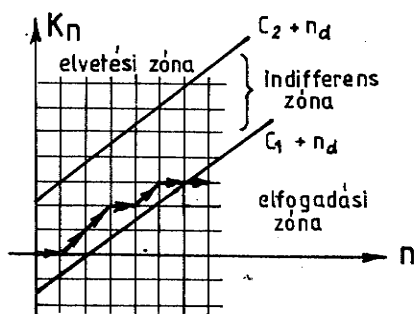
Ez az egyenlőtlenség

$$c_1 + nd < K_n < c_2 + nd$$

alakban írható, ahol c_1 , c_2 és d jelentése a fenti relációk összehasonlítása alapján nyilvánvaló. Az utolsó egyenlőtlenség első és utolsó tagja egy-egy egyenes egyenletre, amelyek párhuzamosak, a d iránytangens pedig pozitív előjelű.

Ezek után az (n, K_n) értékpárokat egy derékszögű koordináta-rendszerben ábrázolva, az utóbbi 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 (45. 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 $c_1 + nd$ egyenest metszi át, akkor a H_0 -t elfogadjuk, ha pedig a felső $c_2 + nd$ egyenest metszi, a H_0 -t elutasítjuk.



45. ábra

3.4 Késztermékek minőségellenőrzésének matematikai statisztikai módszerei

A statisztikai hipotézisvizsgálat tárgyalás során megismert fogalmak gyakorlati jelentőségéről képet kapunk, ha megismerkedünk ezeknek a fogalmaknak alkalmazásával készáruk minőségének ellenőrzésére.

Tömegcikkék minőségét nem mindig lehet minden egyes darab megvizsgálásával ellenőrizni. Roncsolással járó minőségellenőrzésnél, élettartam-vizsgálatoknál és más esetekben ez elvileg is lehetetlen, máskor gazdaságossági okok szólnak ellene. Ilyenkor adott N darabszámú tétel minőségellenőrzését n elemű véletlen minta alapján végezzük. A kiválasztott n mintaelem között talált selejtes darabok számából következtethetünk a teljes tétel selejtarányára. (A tétel selejtaránya a tételben levő selejtes darabok száma osztva N -nel, a tétel elemszámával).

A mintavétel útján megvizsgált darabok n száma - a minta terjedelme - általában csak néhány (5-10) százaléka a tétel darabszámának, N -nek.

A mintavételes minőségellenőrzésre szolgáló matematikai statisztikai eljárást mintavételi tervnek nevezzük. Az ilyen terveket a mintavétel módja, a minta nagysága, a minőségi követelmények és a statisztikai jellemzők adta szempontok szerint tervtáblázatokba szokás foglalni, amelyek a különböző szabványokban találhatók meg. Az alábbiakban azokat a matematikai statisztikai alapelveket ismertetjük, amelyek a mintavételi tervek konstrukciójánál alkalmazásra kerülnek.

A mintavételi terv elkészítésekor rögzíteni kell, hogy mi a tétellel szembeni minőségi követelmény. A minőségellenőrzés során mindenekelőtt feltesszük, hogy a tétel minden darabjáról eldönthető, hogy megfelelő vagy selejtes.

Definíció.

Valamely tétellel szembeni minőségi követelményen olyan p^* selejtarány megadását értjük, amelynél alacsonyabb selejtarányú tételt elfogadhatónak, a nála nagyobb selejtarányú tételt meg nem felelőnek tartunk.

A fenti minőségi követelmény a következő formában fogalmazható meg hipotézisvizsgálati feladatként. Vizsgáljuk a

$$H_0^* : p < p^*$$

hipotézist a

$$H_1^* : p \geq p^*$$

alternatív hipotézissel szemben.

E két hipotézis közötti megnyugtató döntéshez igen nagy mintaelemszámra lenne szükség. A minta alapján történő döntés a következő ismert hibalehetőségekkel jár: elsőfajú hibát követünk el, ha "jó" tételt p^* -nál (kisebb selejtarányú tételt) visszautasítunk, másodfajú hibát vétünk, ha "rossz" tételt (p^* -nál nagyobb selejtarányú tételt) elfogadunk.

Tekintsük át a szemben álló felek érdekeit.

A minőségellenőrzés során az átadó (gyártó) és az átvevő (felhasználó) érdekei ellentétesek. Az átadó érdeke, hogy a kis selejtarányú tételeket nagy valószínűséggel elfogadjuk (azaz csak kis valószínűséggel utasítsuk el), az átvevő érdeke, hogy a nagy selejtarányú tételeket nagy valószínűséggel elutasítsuk (azaz csak kis valószínűséggel fogadjuk el). Ezért saját érdekének védelmére az átvevő kiköti: a mintavételi terv olyan legyen, hogy a legalább p_1 ($> p^*$) selejtarányú tételeket kicsiny - legfeljebb ε_1 - valószínűséggel kelljen elfogadnia, az átadó viszont kiköti, hogy a legfeljebb p_0 ($< p^*$) selejtarányú tételeket az átvevő legfeljebb ε_0 valószínűséggel utasíthasson el.

A fenti kikötéseknek megfelelően a következő hipotéziseket vizsgáljuk:

$$H_0 : p \leq p_0$$

$$H_1 : p \geq p_1$$

Definíció.

A mintavételi terv olyan statisztikai próba, amelynek alkalmazásakor az n elemű mintában talált selejtes darabok k száma alapján átvesszük a tételt (a H_0 hipotézist fogadjuk el), ha k nem nagyobb, mint bizonyos c küszöbszám; ha viszont a k (próbastatisztika) értéke nagyobb, mint c , akkor a tételt elutasítjuk (a H_1 hipotézist fogadjuk el). Az n és c számokat úgy kell meghatározni, hogy a $(p_0, p_1; \varepsilon_0, \varepsilon_1)$ számokkal jellemezhető követelmény teljesüljön. Az ilyen mintavételi tervet (n, c) -tervnek nevezzük.

Példa.

Egy csavargyárból a csavarok 2000 db-os tételekben érkeznek. Legyen $p_0 = 0,02$; $p_1 = 0,08$; $\varepsilon = 0,05$; $\bar{\varepsilon} = 0,10$.

Tehát azt kívánjuk, hogy a 2 % vagy kevesebb selejtes darabot tartalmazó tételeknek legalább 95 %-a elfogadásra kerüljön, viszont a 8 %-nál nagyobb selejtarányú tételeknek legfeljebb 10 %-a menjen át az ellenőrzésen. Kérdés: Mekkora legyen n , a megvizsgálandó csavarok száma, minden egyes tételből, és mekkora legyen c , az elfogadáskor megengedhető selejtes darabok minimális száma a mintában?

Megoldás.

A kérdésre válaszolandó, vizsgáljuk meg először elméletileg, hogyan kell az (n, c) -tervet megkonstruálni.

Ha az N darabszámú tételben M selejtes darab van (azaz a teljes selejtarány $\frac{M}{N} = p$), akkor annak valószínűsége, hogy egy n elemű mintában k selejtes darabot találjunk:

$$P_k = \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}}$$

tehát a mintában lévő selejtes darabok száma hipergeometrikus eloszlást követ.

Kimutatható, hogy ha N igen nagy n -hez képest ($N > 10n$), akkor a hipergeometrikus eloszlás binomiális eloszlással jól közelíthető, azaz

$$P_k \approx \binom{n}{k} p^k (1-p)^{n-k}.$$

Ennek alapján annak valószínűsége, hogy a definícióban vázolt eljárás alkalmazásakor az átvevő egy p selejtarányú tételt elfogadjon:

$$L(p; n, c) = \sum_{k=0}^c \binom{n}{k} p^k (1-p)^{n-k}$$

Sorozatos parciális integrálással igazolható, hogy

$$L(p; n, c) = (n-c) \binom{n}{c} \int_p^1 t^c (1-t)^{n-c-1} dt$$

Ez utóbbi formulából látható, hogy $L(p; n, c)$ a p selejtarány monoton csökkenő folytonos függvénye.

Az átvétel valószínűsége $p = p_0$ esetén (közelítőleg) $1 - \varepsilon_0$ kell legyen, azaz

$$L(p_0; n, c) \approx 1 - \varepsilon_0,$$

továbbá, ha $p = p_1$, akkor az átvétel valószínűségére

$$L(p_1; n, c) \approx \varepsilon_1$$

kell, hogy teljesüljön.

Ha a $(p_0, p_1, \varepsilon_0, \varepsilon_1)$ követelmény adott, akkor e két egyenlet az ismeretlen n és c értékeket egyértelműen meghatározza.

Ha p kicsiny, akkor a binomiális eloszlás Poisson-eloszlással közelíthető (lásd: 2.2.5). A minőségellenőrzésnél mint p_0 , mind p_1 kis valószínűségek, ezért jó közelítéssel:

$$L(p; n, c) = \sum_{k=0}^c \frac{(np)^k}{k!} e^{-np}.$$

A megoldandó egyenletrendszer tehát:

$$L(p; n, c) = \sum_{k=0}^c \frac{(np_0)^k}{k!} e^{-np_0} \leq 1 - \varepsilon_0,$$

$$1 - L(p_1; n, c) = 1 - \sum_{k=0}^c \frac{(np_1)^k}{k!} e^{-np_1} \leq \varepsilon_1.$$

A példában feltett kérdésre válaszolandó, eljárásunk a következő. Ha $n = 100$, $\lambda_0 = np_0 = 100 \cdot 0,02 = 2$, $\lambda_1 = np_1 = 100 \cdot 0,08 = 8$, akkor a Poisson-eloszlás táblázatából megállapíthatjuk:

$$\sum_{k=0}^4 \frac{2^k}{k!} e^{-2} = 0,947 \approx 0,95 = 1 - \varepsilon_0$$

$$\sum_{k=0}^4 \frac{8^k}{k!} e^{-8} = 0,099 \approx 0,1 = \varepsilon_1.$$

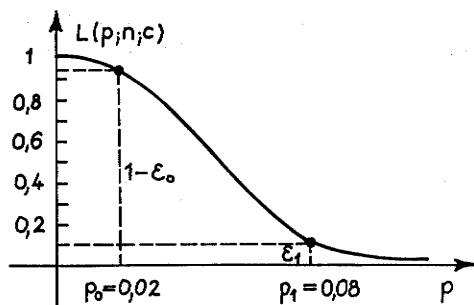
Ha tehát $n = 100$ és $c = 4$, akkor a $(0,02; 0,08; 0,05; 0,1)$ követelmény teljesül:

Definíció.

Az $L(p; n, c)$ függvényt a minőségellenőrzési terv karakterisztikájának nevezzük.

Nyilvánvaló, hogy $L(0; n, c) = 1$, $L(1; n, c) = 0$; továbbá, mint láttuk, $L(p; n, c)$ p -nek monoton csökkenő függvénye. A mintavételi terv követelményei alapján $L(p_0, n, c) = 1 - \varepsilon_0$

$L(p; n, c) = \varepsilon_1$ kell legyen. Ezek alapján a terv karakterisztikájának görbéje megrajzolható. A példában vizsgált terv karakterisztikájának görbáját a 46. ábrán vázoltuk.



46. ábra

Megjegyezzük, hogy $1 - L(p; n, c)$ mint p függvénye a próba erőfüggvénye arra a hipotézisre nézve, hogy a tétel selejtaránya p .

Definíció.

A mintavételi tervnél az elsőfajú hiba valószínűségét (H_0 -t elvetjük, holott igaz), azaz elutasítjuk a tételt, jóllehet a tétel selejtaránya nem nagyobb p_0 -nál, az átadó kockázatának nevezzük. A másodfajú hiba valószínűségét (elfogadjuk H_0 -t holott H_1 igaz), azaz annak valószínűségét, hogy p_1 -nél nagyobb selejtarányú tétel átcsúszik az ellenőrzésen, az átvevő kockázatának nevezzük.

A mintavétellel történő minőségellenőrzés esetén általános gyakorlat, hogy a visszautasított tételeket teljes egészében átvizsgálják, és a selejtes darabokat megfelelő minőségű darabokkal pótolják. A mintavételi terv alkalmazása során átvett tételek azonban általában bizonyos mennyiségű selejtet tartalmaznak. Sorozatosan alkalmazott mintavételi eljárás folyamán kialakul egy bizonyos p_s átlagos selejtarány, amelyet átvettünk. Ha a teljes gyártmánytömeg selejtaránya p , akkor $p = 0$ értékre $p_s = 0$, továbbá $p = 1$ értékre ugyancsak $p_s = 0$, hiszen ekkor a tételt visszautasítjuk, és minden selejtes darabot kicserélnek megfelelőre.

Vizsgáljuk meg, hogy a teljes gyártmánytömeg milyen p selejtaránya esetében veszi fel a p_s átlagosan átengedett selejtarány maximális értékét $p_{\max}$ -ot. Ha ezt meg tudjuk állapítani adott mintavételi terv esetében, akkor gyakorlati szempontból ez nagyon fontos tájékoztatást nyújt számunkra, hiszen ha semmit sem tudunk a gyártmánytömeg valódi p selejt-

arányáról, akkor is tudjuk, hogy az átvett gyártmányok átlagos selejtaránya ennél a $p_{\max}$ -nál nem nagyobb. Megmutatjuk, hogy $p_{\max}$ értéke közelítőleg meghatározható grafikus úton a tervkarakterisztika görbéje alapján. Ehhez gondoljuk meg a következőket.

Ha a teljes gyártmánytömeg selejtaránya p , akkor N elemű tételben átlagosan N_p selejtes darab van. Mivel minden tételt $L(p)$ valószínűséggel veszünk át (a selejttel együtt), $1 - L(p)$ valószínűséggel pedig selejtmentes tételt kapunk (mert ilyen valószínűséggel a tételt elutasítjuk, és a selejtet megszüntetjük), az átvett selejtes darabok várható száma: $NpL(p)$. A keresett átlagos selejtarány:

$$p_s = pL(p) .$$

A $pL(p)$ függvény maximumát csak ott veheti fel, ahol p szerinti deriváltja zérus:

$$\frac{dp_s}{dp} = L(p) + pL'(p) = 0 .$$

Az egyenlet gyöke, grafikusan (durva) közelítéssel, az alábbi módon határozható meg: $pL(p)$ annak a téglalapnak a területe, amelyet a $(p; L(p))$ koordinátarendszerben a $(p; L(p))$ pont koordinátái a koordinátatengelyekkel bezárnak. A $p_{\max}$ érték azaz a p -vel egyenlő, amelyre a $pL(p)$ terület maximális. Ezen a helyen a görbe érintőjének iránytangense:

$$L'(p) = - \frac{L(p)}{p} ,$$

amely megegyezik a kérdéses téglalap negatív lejtésű átlójával.

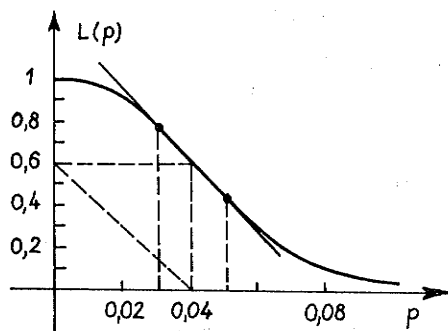
Példa.

Az előző példában ismertetett ($p_0 = 0,02$; $p_1 = 0,08$; $\varepsilon_0 = 0,05$; $\varepsilon_1 = 0,1$) követelményt kielégítő $(n, c) = (100; 4)$ terv esetében a Poisson-eloszlás táblázata alapján kiszámítjuk a p , $L(p)$, $p = pL(p)$ értékeket az alábbi selejtarányokra:

p	$L(p)$	$p_s = pL(p)$
0,02	0,95	0,0190
0,03	0,82	0,0246
0,04	0,63	0,0252
0,05	0,44	0,0220
0,06	0,29	0,0174
0,07	0,17	0,0119
0,08	0,10	0,0080

Az adatokból látható, hogy 4 %-os selejtarány esetében kapjuk a maximális átengedett selejtarányt, és ez 2,52 %. Ez azt jelenti, hogy ha minden tétel 4 % selejtet tartalmazna, akkor átlagosan 2,52 % selejtet veszünk át a terv alapján történő ellenőrzésnél. Minden más esetben ennél kisebb lesz az átengedett átlagos selejtarány.

A táblázat adatai alapján heurisztikusan megrajzolt terv-karakterisztika görbéje és a grafikus szerkesztés menete a 47. ábrán látható.



47. ábra

3.5 Nem paraméteres próbák

Az eddigiekben ismertetett statisztikai próbák normális eloszlású valószínűségi változók várható értékére vagy szóráására vonatkozó hipotézisek ellenőrzésére szolgáltak. Mindegyik próbánál feltételeztük, hogy a szóban forgó valószínűségi változó normális eloszlású.

A műszaki gyakorlatban sokszor előfordul, hogy nem ismerjük a vizsgált valószínűségi változó eloszlását, nem tudjuk, hogy normális eloszlással vagy valamely más eloszlással állunk szemben. Ilyenkor azt kell eldöntetünk, hogy milyen eloszlásról van szó. Azt általában könnyű eldönteni - a műszaki háttér alapján - , hogy a vizsgált változó diszkrét vagy folytonos eloszlású.

Diszkrét eloszlás esetén, ha valamely esemény bekövetkezése vagy be nem következése érdekel bennünket, amely esemény valószínűsége állandó lehet, binomiális eloszlásra gondolunk. Bizonyos megszámlálási problémáknál, mint pl. adott időintervallumban érkező telefonhívások száma, kiválasztott útszakaszon meghatározott idő alatt áthaladó gépkocsik száma, szövegépen adott időtartam alatti fonalszakadások száma, bizonyos időtartamban bekövetkező árvizek száma stb. Poisson-eloszlás léphet fel.

Ha folytonos valószínűségi változó eloszlását vizsgáljuk, akkor a kérdés műszaki háttérének vizsgálata alapján lehetőleg elméleti úton igyekszünk az eloszlásra következtetni. Ha a vizsgált változó ingadozását sok kisebb, függetlenül ható tényező hozza létre, indokolt lehet a normalitás feltételezése. Ha semmiféle támpontunk nincs az eloszlás jellegére vonatkozólag, akkor viszonylag nagy (150-300) elemszámú mintából kiindulva ajánlatos először a sűrűség-hisztogramot megszerkeszteni. Ennek alakja utalhat valamilyen elméleti eloszlásra, amely a mintához való illeszkedését alkalmas statisztikai eljárással (próbával) ellenőrizzük.

Definíció.

Az olyan statisztikai próbát, amelynek alapján arról döntünk, hogy valamely X valószínűségi változó $F(x)$ eloszlása lehet-e adott $F_0(x)$ eloszlásfüggvénnyel jellemzett eloszlás, illeszkedésvizsgálatnak nevezzük. Ilyenkor tehát a H_0 nullhipotézis:

$$H_0 : F(x) = F_0(x) .$$

Ha a nullhipotézis az eloszlás paramétereinek ismeretét is feltételezi, akkor tiszta illeszkedésvizsgálatról beszélünk. Ha viszont hipotézisünk csak az eloszlás jellegét (normalitás, exponencialitás, stb.) tételezi fel, és a paramétereket a mintából kell becsülnünk, akkor becslési illeszkedésvizsgálatot végzünk.

Az illeszkedésvizsgálatra szolgáló próbák alkotják a nem paraméteres próbák egyik nagy csoportját. E próbák közül legelterjedtebb a χ^2 -próba és a Kolmogorov-próba. A χ^2 -próba mind diszkrét, mind folytonos eloszlások esetében alkalmazható, de nagy mintaelemszámot igényel. A Kolmogorov-próba folytonos eloszlások esetében alkalmazható illeszkedésvizsgálatra. Lényegében ez a próba is nagy mintákra alkalmas, bár vannak táblázatok kis mintás esetre is. Elsődleges tájékozódásra normális eloszlás feltételezése esetén használatos grafikus illeszkedésvizsgálat is.

A nem paraméteres eljárások másik nagy osztályát az ún. homogenitásvizsgálati módszerek képezik.

Definíció.

A homogenitásvizsgálat tárgya annak eldöntése, hogy két valószínűségi változó, X és Y egyforma eloszlású-e vagy sem.

Ilyenkor tehát a nullhipotézis:

$$H_0 : P(X < x) = P(Y < x).$$

A közösnek feltételezett eloszlásfüggvény nem is szerepel a próbában.

Mind diszkrét, mind folytonos eloszlások esetében használatos homogenitásvizsgálatra a χ^2 -próba; folytonos eloszlások esetére vonatkozólag a Kolmogorov-Szmirnov-féle kétmintás próbát (nagy mintákra), a Gnyegyenko-Koroljuk próbát, valamint a Wilcoxon-próbát ismertetjük.

3.5.1 Illeszkedésvizsgálat

a.) A χ^2 -próba

A χ^2 -próba mind diszkrét-, mind folytonos eloszlás esetében alkalmazható illeszkedésvizsgálatra, ha elég nagy statisztikai mintával rendelkezünk. A próba elvi alapját az alábbiakban vázoljuk.

Tekintsünk egy olyan kísérletet, amelynek során az $A_1, A_2, \dots, A_n$ események közül egy és csakis egy következik be, tehát az A_i események egymást kizárják, és $\sum_{i=1}^n A_i = \Omega$. Ez azt jelenti, hogy $A_1, A_2, \dots, A_n$ események teljes eseményrendszert alkotnak, amelyre vonatkozólag felállítjuk a

$$H_0 : P(A_i) = p_i \quad (i=1, 2, \dots, n; \sum_{i=1}^n p_i = 1)$$

hipotézist.

A kísérletet N -szer megismételve (feltételezve, hogy az egyes kísérletek eredményei függetlenek egymástól), az A_1 esemény v_1 -szer, az A_2 esemény v_2 -ször, ..., az A_n

esemény v_n -szer következik be ($v_i \geq 0, \sum_{i=1}^n v_i = N$).

Mindegyik v_i valószínűségi változó binomiális eloszlást követ, hiszen egy kísérletnél az A_i esemény vagy bekövetkezik, vagy nem, míg a $(v_1, v_2, \dots, v_n)$ változók együttes eloszlása multinomiális eloszlás. Ha H_0 igaz, akkor $M(v_i) = Np_i$ ($i=1, 2, \dots, n$). Az χ^2 -próba alkalmazásakor N olyan nagy legyen, hogy $v_i > 5-10$, vagy egyesítsünk eseményeket.

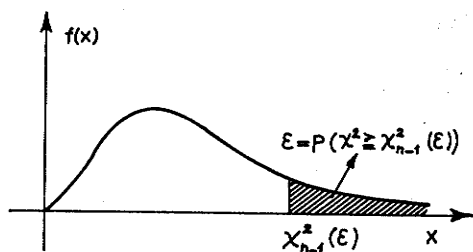
Képezzük a következő, ún. χ^2 -statisztikát:

$$\chi^2 = \sum_{i=1}^n \frac{(v_i - Np_i)^2}{Np_i}.$$

A H_0 hipotézis fennállása esetén $M(v_i - Np_i) = 0$, vagyis a számlálóban 0 várható értékű valószínűségi változók állnak; így ha H_0 igaz, a χ^2 -statisztika aktuális értéke nem lesz túl nagy. Kimutatható, hogy a fenti statisztika $N \rightarrow \infty$ esetén $(n-1)$ paraméterű χ^2 -eloszlást követ. A χ^2 -eloszlást követ. A χ^2 -eloszlás tárgyalása során láttuk, hogy $M(\chi_{n-1}^2) = n-1$, $D(\chi_{n-1}^2) = \sqrt{2(n-1)}$. Adott $\varepsilon > 0$ szint választása mellett a χ^2 -eloszlás táblázatából kikereshető olyan $\chi_{n-1}^2(\varepsilon)$ kritikus érték, amelyre

$$P(\chi^2 \geq \chi_{n-1}^2(\varepsilon)) = \varepsilon.$$

Kritikus tartományok a számegyenes $\chi_{n-1}^2(\varepsilon)$ -től jobbra eső részét választjuk (48. ábra). Amennyiben a számított χ^2 -statisztika értéke a kritikus tartományba esik, H_0 -t $(1 - \varepsilon)$ szinten elvetjük.



48. ábra

Illeszkedésvizsgálat χ^2 -próbával diszkrét eloszlás esetében. Először mutatunk egy példát tiszta illeszkedés-vizsgálatra diszkrét eloszlás esetében.

Példa.

Tegyük fel, hogy egy játékkockáról el akarjuk dönteni, 1 - $\varepsilon = 0,95$ szinten, hogy szabályos-e. A kockával 600 dobást végezve az alábbi eredményeket kaptuk:

A_1 : 1-es	$v_1 = 98$,
A_2 : 2-es	$v_2 = 109$,
A_3 : 3-as	$v_3 = 90$,
A_4 : 4-es	$v_4 = 102$,
A_5 : 5-ös	$v_5 = 103$,
A_6 : 6-os	$v_6 = 98$,
	600

A H_0 hipotézis most $P(A_i) = p_i = \frac{1}{6}$ ($i=1,2,\dots,6$); $Np_i = 100$.

$$\chi^2 = \frac{1}{100} (2^2 + 9^2 + 10^2 + 2^2 + 3^2 + 2^2) = \frac{202}{100} = 2,02.$$

Mint ahogy a χ^2 -eloszlás táblázatában (Függelék V. táblázat) az $n = 5$ sorban a 0,05 százalékos oszlopban található kritikus érték 11,07, és az adott statisztikának ennél jóval kisebb, nincs okunk a H_0 hipotézis elvetésére.

(A χ^2 -eloszlás táblázatának 5. sorát figyelembe véve az is látható, hogy annak valószínűsége, hogy H_0 fennállása, tehát szabályos kocka esetén a számított χ^2 -statisztika értéke a véletlen játéka folytán nagyobb legyen, mint a kapott 2,02, nagyobb mint 0,80.)

Mutatunk egy példát becsléses illeszkedésvizsgálatra χ^2 -próbával diszkrét eloszlás esetében.

Példa.

A Tisza folyóra vonatkozólag feljegyezték, hogy Tokajnál a $c = 600$ cm szintet meghaladó árhullámok száma az egyes évek első negyedéveiben (jan.-márc.) az 1903-1971 időszakban, tehát 68 év során a következő módon alakult:

- 30 olyan év volt, amikor nem volt árhullám (a kérdéses negyedévben);
- 25 olyan év volt, amikor 1 árhullám volt;
- 9 olyan év volt, amikor 2 árhullám jött;
- 4 olyan év volt, amikor 3 vagy több árhullám következett be.

Jelölje az X valószínűségi változó az árhullámok számát Tokajnál az első negyedévben. Mivel a kérdéses időszakban az első negyedévekben összesen 55 árhullám volt,

$$M(X) = \frac{55}{68} \approx 0,8.$$

Legyen a H_0 hipotézis, hogy X Poisson-eloszlású $\lambda = 0,8$ paraméterrel. Az 1. táblázat tartalmazza a megfelelő Poisson-valószínűségeket (P_k) is.

	k	ν_k	P_k	NP_k
$A_1 :$	0	30	0,4493	30,55
$A_2 :$	1	25	0,3595	24,45
$A_3 :$	2	9	0,1438	9,78
$A_4 :$	3	4	0,0474	3,22
	(vagy több)			

A χ^2 -statisztika aktuális értéke:

$$\chi^2_3 = \frac{0,55^2}{30,55} + \frac{0,35^2}{24,45} + \frac{0,78^2}{9,78} + \frac{0,78^2}{3,22} \approx 0,165$$

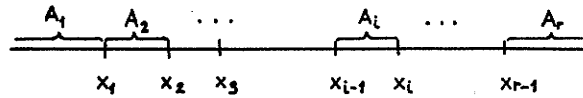
Mint hogy a λ paramétert a mintából becsültük, a χ^2 -eloszlás szabadságfokát 1-gyel csökkenteni kell, így a χ^2 -eloszlás táblázatának második sorában kell a megfelelő kritikus értéket keresnünk. Ha 0,95 szinten kívánunk dönteni, akkor a táblázat 2. sorának 0,05 oszlopában a kritikus érték: 5,99. Mint hogy statisztikánk aktuális értéke a kritikus értéknél sokkal kisebb, nincs okunk a H_0 hipotézis elvetésére.

Tiszta illeszkedésvizsgálat χ^2 -próbával folytonos eloszlás esetében. Tételezzük fel, hogy a folytonos eloszlás X valószínűségi változóra vonatkozólag a

$$H_0 : P(X < x) = F(x)$$

nullhipotézist kívánjuk vizsgálni, ahol $F(x)$ adott eloszlásfüggvény, amelyben a paramétereket is ismertnek tételezzük fel.

Legyen az X-re vonatkozó mintánk $X_1, X_2, \dots, X_n$, ahol n nagy szám (pl. $n \approx 100 - 300$). Osszuk fel X változási tartományát (pl. normális eloszlás feltételezése



49. ábra

esetén az egész számegyenest, exponenciális eloszlásnál a pozitív félegyenest) az $x_1, x_2, \dots, x_{r-1}$ osztópontokkal r számú részintervallumra (49. ábra).

Legyen $A_i = \{x_{i-1} \leq X < x_i\}$, $i=1, 2, \dots, r$ (ahol $x_r = \infty$). Ekkor H_0 fennállása esetén igaz a

$$H'_0 : P(A_i) = F(x_i) - F(x_{i-1}) \quad (i=1, 2, \dots, r)$$

hipotézis, ami az eredeti nullhipotézisnek kissé módosított alakja, az osztópontok önkényes megválasztása miatt.

Tegyük fel, hogy az A_1 esemény v_1 -szer, az A_2 esemény v_2 -szer, ..., az A_r esemény v_r -szer következett be, ahol

$$\sum_{i=1}^r v_i = n.$$

(Ez azt jelenti, hogy $[x_{i-1}, x_i)$ intervallumba v_i darab mintaelem esik ($i=1, 2, \dots, r$).)

Ezután az $F(x)$ eloszlásfüggvény táblázata alapján meghatározzuk a

$$P(A_i) = F(x_i) - F(x_{i-1}) = p_i$$

valószínűséget és az np_i várható értékeket, majd képezzük a

$$\chi^2_{r-1} = \sum_{i=1}^r \frac{(v_i - np_i)^2}{np_i}$$

statisztikát, amely $(r - 1)$ paraméterű χ^2 -eloszlást követ nagy n esetén. A próba végrehajtása ezután az előzőekben ismeretett módon történik. Nézzünk erre egy példát.

2. táblázat

0,9488;	1,5796;	-0,7292;	1,7438;	0,1512;	-0,5112;	-0,2078;	0,2759;	0,0327;	-0,6823;
-0,0006;	0,4747;	-0,4997;	1,6944;	-2,1880;	-0,4741;	-1,2876;	0,7262;	-1,5398;	-1,8755;
-2,5146;	-0,7356;	-1,2487;	0,7306;	1,0473;	1,6348;	-0,1345;	1,2403;	-0,8845;	-0,8129;
0,0744;	-0,3548;	0,3007;	-0,2221;	1,0965;	-0,4230;	-1,0516;	1,2233;	1,1832;	-1,1152;
-1,7919;	-0,0075;	0,7375;	-0,0025;	0,0469;	-0,0221;	0,4908;	-0,6121;	0,4330;	0,8381;
0,4955;	0,4648;	2,0358;	-0,3438;	0,7301;	-2,3824;	1,2653;	1,1960;	-0,5576;	0,5345;
1,9017;	1,3396;	1,3489;	0,4382;	-0,4175;	-0,3742;	1,3310;	0,2876;	0,4313;	-0,2746;
0,2739;	0,2858;	0,3336;	2,7574;	-0,1516;	-0,0576;	-0,0241;	0,8167;	-0,8946;	0,0832;
0,7703;	-0,4910;	0,0890;	1,1083;	1,1489;	1,2994;	1,1045;	1,9648;	0,5181;	0,2129;
0,5486;	1,2093;	0,6238;	0,8535;	0,9189;	0,0903;	0,4218;	-0,3885;	1,4228;	-0,7755;

Példa.

Azt a H_0 hipotézist akarjuk ellenőrizni $1 - \varepsilon = 0,95$ szinten, hogy adott technológiával gyártott házgyári panel hosszabbik oldalát jelölő X valószínűségi változó normális eloszlású $m = 300$ cm várható értékkel és $\sigma = 0,5$ cm szórással.

Megmértük találomra 100 db panel hosszát. Az i -ediknek mért X_i hosszmeretből levonjuk az $m = 300$ (cm) várható értéket, és osztjuk a $\sigma = 0,5$ (cm) szórással, azaz standardizáljuk az adatokat:

$$x_i^x = \frac{X_i - 300}{0,5}$$

A 100 db standardizált adatot a 2. táblázatban foglaltuk össze. Osszuk fel ezután a számegyeneset az $x_1 = -1$;

$x_2 = -0,5$; $x_3 = 0$; $x_4 = 0,5$; $x_5 = 1$ osztópontokkal hat részintervallumra; azaz legyen

$$A_1 : (-\infty, -1); \quad A_2 : [-1, -0,5); \quad A_3 : [-0,5, 0);$$

$$A_4 : [0, 0,5); \quad A_5 : [0,5, 1); \quad A_6 : [1, \infty).$$

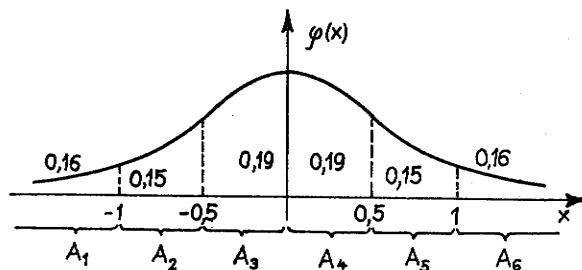
Az egyes részhalmazokba eső mintaelemek száma a következő:

$$v_1 = 10; v_2 = 10; v_3 = 20; v_4 = 22; v_5 = 14; v_6 = 24.$$

Az egyes részhalmazok valószínűségi mértékét a standard normális eloszlás $\phi(x)$ eloszlásfüggvénye táblázata (Függelék IV. táblázat) segítségével könnyen meghatározhatjuk. Szimmetria megfontolások alapján csak a

$$\phi(1) = 0,8413; \phi(0,5) = 0,6915 \text{ és } \phi(0) = 0,5$$

értékekre van szükség. Az egyes halmazok valószínűségei (2 tizedesre kerekítve) az 50. ábrán láthatók, amelyekből az np_i várható értékek rögtön adódnak, mivel $n = 100$.



50. ábra

Ezután kiszámítjuk a χ^2 -statisztikát:

$$\chi^2_6 = \frac{6^2}{16} + \frac{5^2}{15} + \frac{1^2}{19} + \frac{3^2}{19} + \frac{1^2}{15} + \frac{8^2}{16} = 8,49.$$

Legyen $\alpha = 0,05$; $0,95$ szinten kívánunk a H_0 hipotézis elfogadásáról vagy elutasításáról dönteni. A χ^2 -eloszlás táblázatának 5. sorában található $\chi^2_{0,05}$ kritikus érték: $11,07$ alapján nincs okunk arra, hogy a H_0 hipotézist elutasítsuk.

Becsléses illeszkedésvizsgálat esetén, amikor a paramétereket a mintából becsüljük, a próbát a fenti módon hajtjuk végre, de a χ^2 -eloszlás paraméterét a becsült paraméterek számával csökkentjük. Meg kell jegyeznünk, hogy becsléses illeszkedésvizsgálat esetén a χ^2 -próba megbízhatósága kisebb.

b) A Kolmogorov-próba

A Kolmogorov-próba is annak ellenőrzésére szolgál, hogy valamely folytonos eloszlású X valószínűségi változó eloszlásfüggvénye adott $F(x)$ függvény-e. A nullhipotézis tehát ebben az esetben is

$$H_0 : P(X < x) = F(x) .$$

A Kolmogorov-próba elméleti alapját az alábbiakban vázoljuk. Glivenkó tétele szerint az

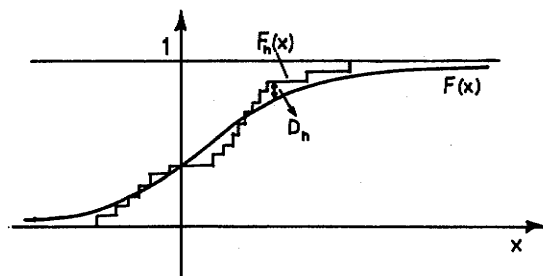
$$x_1^*, x_2^*, \dots, x_n^*$$

rendezett minta segítségével konstruált $F_n(x)$ tapasztalati eloszlásfüggvény nagy n esetén közelíti az $F(x)$ elméleti eloszlásfüggvényt. Ez a fontos tétel azonban semmit sem mond arról, hogy véges n esetén mekkora valószínűséggel lép fel az $F(x)$ elméleti és $F_n(x)$ tapasztalati eloszlásfüggvény között adott nagyságú eltérés. A Kolmogorov-próbával az $F_n(x)$ és $F(x)$ függvények közötti eltérést (távolságot) a két függvény közötti abszolút eltérés maximumával mérjük. A H_0 hipotézis ellenőrzéséhez először meghatározzuk a

$$D_n = \max_x |F_n(x) - F(x)|$$

mennyiséget. Tiszta illeszkedésvizsgálat esetén D_n meghatározása történhet grafikusán is, úgy, hogy megrajzoljuk az

$F(x)$ elméleti eloszlásfüggvény görbáját, és ugyanazon ábrán megrajzoljuk az $F_n(x)$ tapasztalati eloszlásfüggvényt (51. ábra).



51. ábra

Amennyiben D_n értéke nagy, akkor a feltételezett $F(x)$ eloszlásfüggvény nem illeszkedik a tapasztalati adatokhoz, tehát H_0 -t kell utasítanunk. A kérdés csak az, hogy mikor tekinthető D_n értéke nagynak vagy kicsinek. Mivel az elméleti eloszlásfüggvény értéke, $F(x)$ az $\{X < x\}$ esemény valószínűségét jelenti adott x helyen, az empirikus eloszlásfüggvényénél pedig $F_n(x)$ ugyanezen esemény relatív gyakoriságát adja meg, ezért a Csebisev-egyenlőtlenség alapján

$$P(|F_n(x) - F(x)| > \lambda \sqrt{\frac{F(x)(1 - F(x))}{n}}) \leq \frac{1}{\lambda^2}$$

Minthogy $\sqrt{F(x)(1 - F(x))} \leq \frac{1}{2}$, látható, hogy $|F_n(x) - F(x)|$ nagyságrendje $\frac{1}{\sqrt{n}}$ bármely x helyen. Ennek alapján a maximális eltérést is ilyen nagyságrendűnek várjuk. A D_n mennyiség maga is valószínűségi változó, hiszen értéke az $X_1, X_2, \dots, X_n$ mintaelemektől függ. Glivenkó tétele alapján 1 valószínűséggel $D_n \rightarrow 0$, ha $n \rightarrow \infty$, márpedig olyan gyorsan mint az $\left\{\frac{1}{\sqrt{n}}\right\}$ sorozat. Ha a D_n mennyiséget megszorozzuk n -nel, azaz képezzük a $\sqrt{n}D_n$ statisztikát, akkor ez már nem 0-hoz, ha $n \rightarrow \infty$ -hez. A $\sqrt{n}D_n$ statisztika - mivel valószínűségi változó - valamilyen eloszlással rendelkezik, ennek határeloszlását határozza meg, A.N.Kolmogorov.

Tétel.

(Kolmogorov). H_0 fennállása esetén tehát ha a szóban forgó eloszlás valóban $F(x)$ akkor

$$P(\sqrt{n}D_n < z) = K(z) = \sum_{\lambda = -\infty}^{\infty} (-1)^{\lambda} e^{-2\lambda^2 z^2}$$

A $K(z)$ függvény értékeit a Függelék VIII. táblázata tartalmazza. Ennek alapján a Kolmogorov-próbát a következőképpen hajthatjuk végre:

Ha H_0 hipotézisről $1 - \varepsilon = 0,95$ szinten kívánunk dönteni, akkor a VIII. táblázatból megkeressük azt a z_0 számot, amelyre $K(z_0) = 0,95$. Azt találjuk, hogy $z_0 = 1,35$. Tehát

$$P(\sqrt{n}D_n < 1,35) \approx 0,95$$

ha n nagy szám. Ez azt jelenti, hogy

$$P(\sqrt{n}D_n \geq 1,35) \approx 0,05.$$

Ha most az adott minta esetében a $\sqrt{n}D_n$ statisztika értéke nagyobb, mint 1,35, akkor H_0 fennállása esetén igen kis valószínűségű esemény következne be. Ekkora eltérést $F_n(x)$ és $F(x)$ között nem tekinthetünk a véletlen játéknak, inkább arra következtetünk, hogy a nagy eltérés annak következménye, hogy a sokaság tényleges eloszlásfüggvénye különbözik $F(x)$ -től, ezért H_0 -t elutasítjuk.

A Kolmogorov-próba tetszőleges eloszlás esetében alkalmazható, mert a $\sqrt{n}D_n$ statisztika eloszlása nem függ attól, hogy milyen eloszlásról van szó, csak attól, hogy a minta eloszlásfüggvényét, $F_n(x)$ -et annak a sokaságnak tényleges $F(x)$ eloszlásfüggvényéhez illesztjük-e, amelyből a minta származik.

Becsléses illeszkedésvizsgálatot Kolmogorov-próbával elvileg ugyanúgy hajthatunk végre, mint tiszta illeszkedésvizsgálatot, ez azonban bizonyos óvatosságot igényel. Ha pl. normalitásvizsgálatot végzünk, és feltesszük, hogy $F(x)$ valamely $N(m, \sigma^2)$ (normális) eloszlás eloszlásfüggvénye, ahol m és σ^2 ismeretlenek, akkor az m várható értékét az $\bar{X}$ mintaközépkel, a σ^2 szórását az S_n^2 empirikus szórással becsüljük. Az $F_n(x)$ tapasztalati eloszlásfüggvényt ezek után az $N(\bar{X}, S_n^2)$ normális eloszlásfüggvényéhez illesztjük.

A Kolmogorov-próba becsléses illeszkedésvizsgálatra történő alkalmazásakor célszerű alacsonyabb kritikus értéket alkalmazni, mint a tiszta illeszkedésvizsgálatnál. A 3. táblázatban néhány döntési szintre megadunk ilyen értékeket.

$1 - \varepsilon$	Kritikus érték tiszta illeszkedésvizsgálatnál	Kritikus érték becsléses illeszkedésvizsgálatnál
0,9	$z_0 = 1,23$	$z'_0 = 0,82$
0,95	$z_0 = 1,36$	$z'_0 = 0,90$
0,99	$z_0 = 1,63$	$z'_0 = 1,04$

Becsléses illeszkedésvizsgálatra megbízhatóbb módszerek a mintaelemek transzformációján alapuló próbák. Ezek elméleti megvilágítása és táblázati igénye meghaladja könyvünk kereteit.

c) Grafikus módszer normalitásvizsgálata

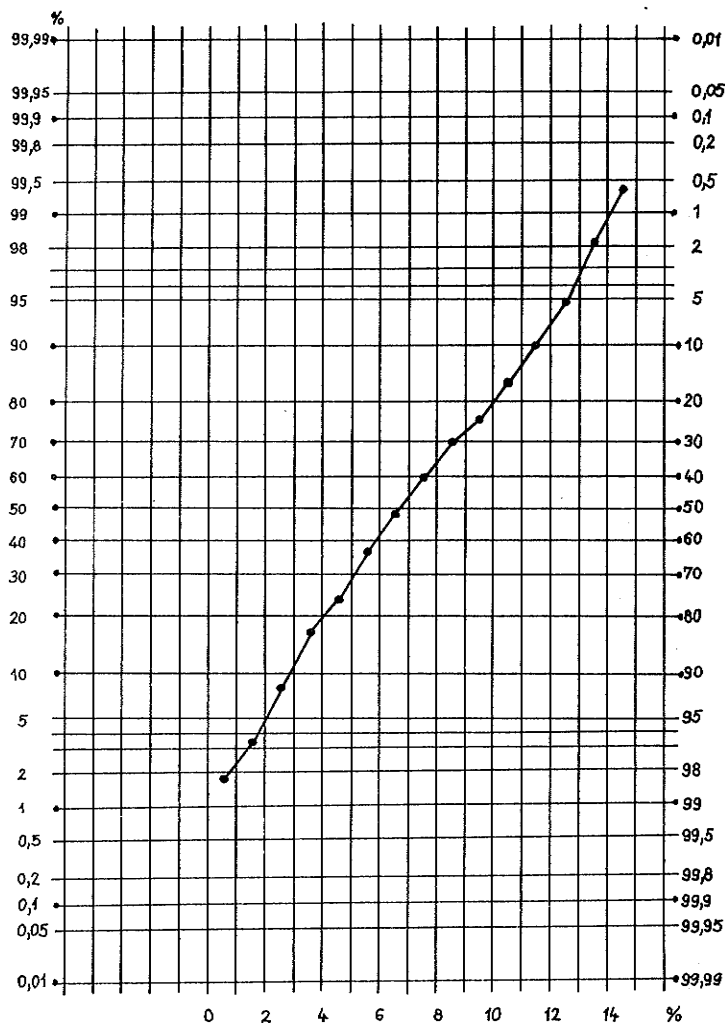
A műszaki gyakorlatban igen sokszor indokolt normális eloszlású sokaság feltételezése. Annak eldöntésére, hogy adott statisztikai minta valóban normális eloszlású sokaságból származik-e, igen egyszerű grafikus módszer ismeretes, amely ugyan kevésbé megbízható, mint az előzőekben ismertetett nem paraméteres próbák, de elsődleges tájékozódásra igen alkalmas.

Grafikus normalitásvizsgálatra az ún. Gauss-papír szolgál. A Gauss-papír függőleges skálája a standard normális eloszlás $\Phi(x)$ eloszlásfüggvényének inverze, a $\Phi^{-1}(x)$ 100-zal vannak szorozva, így 0,01-től 99,99-ig vannak a számok feltüntetve (52. ábra).

A vízszintes tengely egyenletes beosztású, de nincs skálázva. A Gauss-hálózat jobb szélén a függőleges skála közepére a $0,5 = 50\%$ értéket írjuk, ettől lefelé valamilyen egységül választott c távolságra a $\Phi(-1) = 0,1587 = 15,87\%$, felfelé a $\Phi(1) = 0,8413 = 84,13\%$ érték kerül. A középtől lefelé és felfelé $2c$ távolságra a $\Phi(-2) = 0,0228 = 2,28\%$, ill. a $\Phi(2) = 0,9772 = 97,72\%$ értékeket írjuk s.i.t.

Ha most az X valószínűségi változóra vonatkozólag - amelyről feltételezzük, hogy normális eloszlású - az $X_1, X_2, \dots, X_n$ mintánk van, akkor osszuk fel a mintát a

$$-\infty = x_0 < x_1 < x_2 < \dots < x_r = \infty$$



52. ábra

osztópontokkal r részre, és mérjük fel az osztópontokat a hálózat vízszintes tengelyére. Ezután az x_1 abszcisszához mérjük fel a nála kisebb mintaelemek $F_n(x_1)$ relatív gyakoriságát, az x_2 ponthoz $F_n(x_2)$ értékét, s.i.t. Amennyiben valóban normális eloszlású, az $(x_i; F_n(x_i))$ pontok közelítőleg

egyenes mentén helyezkednek el, ugyanis a Gauss-hálózaton a normális eloszlás

$$y = F(x) = \Phi\left(\frac{x - m}{\sigma}\right)$$

eloszlásfüggvényének képe egyenes, és nagy n esetén $F_n(x) \approx F(x)$.

Ha a kapott pontok jelentősen eltérnek az egyenestől, akkor elvetjük a normalitás hipotézisét, kétes esetben - próbával vagy Kolmogorov-próbával végzünk illeszkedésvizsgálatot.

3.5.2 Homogenitásvizsgálat

A homogenitásvizsgálat célja annak eldöntése, hogy két valószínűségi változó, X és Y egyforma eloszlásúak-e vagy sem. Ha az X változó eloszlásfüggvényét $F(x)$, az Y változó eloszlásfüggvényét $G(x)$ jelöli, akkor a nullhipotézis:

$$H_0 : F(x) \equiv G(x).$$

A H_0 hipotézis elfogadásáról vagy elvetéséről a szóban forgó vizsgálat tárgyát képező valószínűségi változókra vonatkozó

$$X_1, X_2, \dots, X_n$$

és

$$Y_1, Y_2, \dots, Y_m$$

statisztikai minták alapján döntünk. A H_0 hipotézis azt fejezi ki, hogy a két minta azonos eloszlású statisztikai sokaságból származik.

A műszaki gyakorlatban homogenitásvizsgálatra sokszor szükségünk lehet. Ha pl. két különböző keverőgépen állítanak elő aszfaltkeveréket, akkor homegintásvizsgálattal dönthetjük el, hogy pl. az A gépen előállított keverék X bitumentartalma és B gépből kikerülő keverék Y bitumentartalma egyforma eloszlásúak-e. Ugyancsak homogenitásvizsgálattal döntünk arról, hogy pl. adott folyó adott helyen mért évi maximális vízállása mondjuk 1940-1960 és 1960-1980 között egyforma eloszlású-e, vagy pl. meder feltöltődés lehet következtetni, vagy esetleg egy tározó megépítése után változott az árvízi vízállás stb.

Homogentisávizsgálatra nagy minták esetében a χ^2 -próba és a Kolmogorov-Szmirnov-féle kétmintás próba, kisebb minták esetére a Gnyegyenko-Koroljuk-próba, valamint a Wilcoxon-próba ajánlható. Ez utóbbi nagy minták esetében is igen célszerű.

a) Homegitásvizsgálat χ^2 -próbával.

χ^2 -próbával az $\{X_n\}$ és $\{Y_m\}$ statisztikai minták alapján a homogenitásvizsgálatot a következő módon végezzük. Osszuk fel a számegyeneset a $-\infty = z_0 < z_1 < \dots < z_r = \infty$ osztópontokkal r számú részintervallumra.

Jelölje ν_i a $[z_{i-1}, z_i)$ intervallumba eső X_i minta-elemek számát, továbbá legyen μ_i ugyanezen intervallumba eső Y_i értékek száma ($i=1, 2, \dots, r$). Ekkor

$$\sum_{i=1}^r \nu_i = n, \quad \sum_{i=1}^r \mu_i = m.$$

Képezzük a

$$\chi^2 = nm \sum_{i=1}^r \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\frac{\nu_i}{n} + \frac{\mu_i}{m}}$$

statisztikát, amely $n \rightarrow \infty, m \rightarrow \infty$ esetén $(r-1)$ paraméterű χ^2 -eloszlást követ. Tehát ha n és m elég nagyok, a χ^2 -próba alkalmazható, a kapott statisztika értékét $\varepsilon > 0$ előzetes megválasztása után a Függelék V. táblázatának $(r-1)$ -edik sorában az ε -nak megfelelő oszlopban álló kritikus értékkel hasonlítjuk össze.

b) A Kolmogorov-Szmirnov-féle kétmintás próba

Ha a folytonos eloszlású X valószínűségi változó eloszlásfüggvénye $F(x)$, az ugyancsak folytonos Y változó eloszlásfüggvénye $G(x)$, akkor a

$$H_0 : F(x) \equiv G(x)$$

nullhipotézis fennállása esetén nagy mintáknál - Glivenkó tétele alapján - a megfelelő empirikus eloszlásfüggvények sem nagyon térhetnek el egymástól.

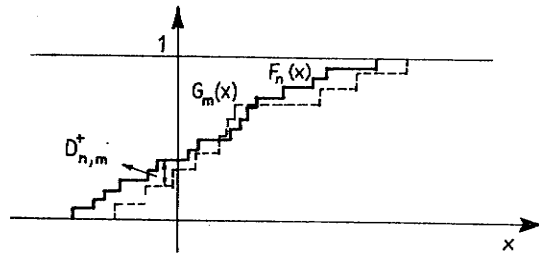
E próba végrehajtásához rendezzük az $\{x_i\}$ és $\{y_i\}$ statisztikai minták elemeit nagyság szerint növekvő sorrendben, azaz képezzük az

$$x_1^* < x_2^* < \dots < x_n^*$$

és az

$$y_1^* < y_2^* < \dots < y_m^*$$

rendezett mintákat.



53. ábra

A rendezett minták alapján rajzoljuk meg az $F_n(x)$ és $G_m(x)$ empirikus eloszlásfüggvényeket, és állapítsuk meg ezek maximális eltérését (53. ábra).

Ha a H_0 nullhipotézissel szemben a

$$H_1 : F(x) > G(x)$$

egyoldali alternatív hipotézist állítjuk, akkor próbastatisztikának a

$$D_{n,m}^+ = \max_x [F_n(x) - G_m(x)]$$

menyiséget választjuk maximális eltérés az $F_n(x)$ függvény javára.

Ha viszont alternatív hipotézisünk a

$$H_1' : F(x) \neq G(x)$$

kétoldali ellenhipotézis, akkor a megfelelő próbastatisztika

$$D_{n,m} = \max_x |F_n(x) - G_m(x)|$$

valószínűségi változó, amely tehát két empirikus eloszlásfüggvény közötti abszolút eltérés maximuma.

A $D_{n,m}^+$ és a $D_{n,m}$ statisztikák határeloszlását N.V. Szmirnov határozta meg.

Tétel.
(N.V.Szmirnov). A $H_0: F(x) \equiv G(x)$ nullhipotézis fennállása esetén

$$\lim_{\substack{n \rightarrow \infty \\ m \rightarrow \infty}} P\left(\sqrt{\frac{nm}{n+m}} D_{n,m}^+ < z\right) = 1 - e^{-2z^2}$$

$$\lim_{\substack{n \rightarrow \infty \\ m \rightarrow \infty}} P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} < z\right) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}$$

A próba végrehajtása a $H_1: F(x) > G(x)$ egyoldali ellenhipotézis esetében a következő módon történik: $D_{n,m}^+$ megállapítása után kiszámítjuk a $\sqrt{\frac{nm}{n+m}} D_{n,m}^+ = z^*$ aktuális értéket.

Ha most $1 - \varepsilon = 0,95$ százalékos szinten kívánunk dönteni H_0 elfogadásáról vagy elutasításáról, akkor meg kell határoznunk azt a z_0 kritikus értéket, amelyre - nagy n és m esetén - :

$$P\left(\sqrt{\frac{nm}{n+m}} D_{n,m}^+ < z_0\right) = e^{-2z_0^2} = 0,05.$$

Ez utóbbi összefüggésből:

$$-2z_0^2 = \ln 0,05 \approx -3.$$

$$z_0 \approx \sqrt{\frac{3}{2}} \approx 1,22$$

Amennyiben $z^* \geq 1,22$, a H_0 hipotézist 0,95 szinten elutasítjuk. Hasonló egyszerű számítással kapjuk a z_0 kritikus értéket tetszőleges $\varepsilon > 0$ választása esetén:

$$z_0 = \sqrt{-\frac{1}{2} \ln \varepsilon}.$$

Ha próbastatisztikának a

$$D_{n,m} = \max_x |F_n(x) - G_m(x)|$$

maximális abszolút eltérést választjuk, akkor Szmirnov tétele értelmében a statisztika határeloszlása a Kolmogorov-féle eloszlással egyezik meg; vagyis ez esetben a kritikus érték adott $\varepsilon > 0$ választása mellett a $K(z)$ függvény táblázatából kereshető ki (Függelék VIII. táblázat) a Kolmogorov-próbánál leírtakkal azonos módon.

c) Homogenitásvizsgálat a Gnyegyenko-Koroljuk próbával

Mivel a Kolmogorov-Szmirnov próba elméleti alapját határeloszlások képezik, ezért kisebb minták (pl. $n, m=30$) esetén a próba nem alkalmazható. Egyenlő mintaelemszám (azaz $n=m$) esetében véges n -re a

$$D_n^+ = \max_x [F_n(x) - G_n(x)], \text{ ill. } D_n = \max_x |F_n(x) - G_n(x)|$$

statisztikák pontos eloszlását B.V.Gnyegyenko és V.Sz.Koroljuk határozták meg. Eredményüket tétel formájában mondjuk ki.

Tétel.

(Gnyegyenko-Koroljuk). A $H_0 : F(x) \equiv G(x)$ hipotézis fennállása esetén

$$P(D_n^+ \geq \frac{k}{n}) = \frac{\binom{2n}{m+k}}{\binom{2n}{n}} \quad (k=0, 1, 2, \dots, n).$$

$$P(D_n \geq \frac{k}{n}) = \sum_{i=-\infty}^{\infty} (-1)^i \frac{\binom{2n}{n-ik}}{\binom{2n}{n}} \quad (k=1, 2, \dots, n)$$

Bizonyítás.

Ismertetjük az első formula származtatásának gondolatmenetét, miközben a próba végrehajtására egyszerű grafikus módszert nyerünk. Az

$$x_1^* < x_2^* < \dots < x_n^*$$

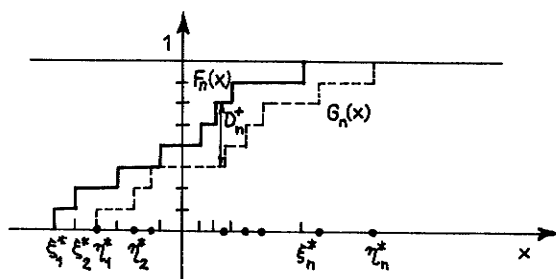
és az

$$y_1^* < y_2^* \dots < y_n^*$$

rendezett mintaelemek mindegyikét nagyság szerint mérjük fel a számegyenesre, ami által balról jobbra haladva egy

$$\zeta_1^* < \zeta_2^* < \dots < \zeta_{2n}^*$$

$2n$ elemű nagyság szerint rendezett mintát nyerünk (54. ábra). Az ily módon kapott egyesített, rendezett minta mindegyik eleme alá írjuk oda, hogy az melyik mintából származik, mégpedig úgy, hogy ha $\zeta_i^* = x_j^*$, akkor $(+1)$ -et írunk, ha $\zeta_i^* = y_k^*$, akkor pedig (-1) -et. Ezáltal egy $(+1)$ és (-1) -ekből álló $2n$ elemű sorozatot nyerünk, amelyben n db $(+1)$ -es és n db (-1) -es van.



54. ábra

Jelöljük ezt a sorozatot így:

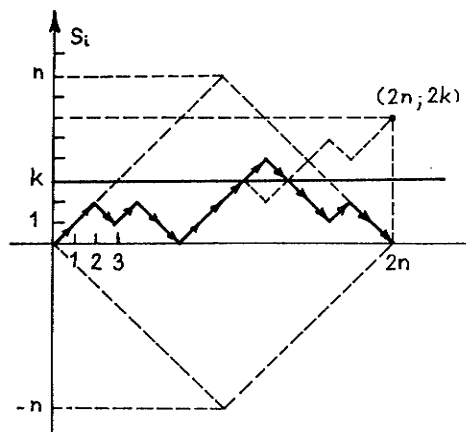
$$v_1^h, v_2^h, \dots, v_{2n}^h$$

ahol

$$v_i^h = \pm 1, \quad (i=1, 2, \dots, 2n) ,$$

$$\sum_{i=1}^{2n} v_i^h = 0 .$$

A v_n^h sorozatot grafikusán ábrázoljuk a következő módon: egy derékszögű koordináta-rendszerben a origóból indulva $(+1)$ esetén jobbra felfelé, (-1) esetén jobbra lefelé rajzolunk egy $\sqrt{2}$ hosszúságú vektort (55. ábra.)



55. ábra

Képezzük most a $\{v_n^h\}$ sorozat elemeinek részletösszegeit:

$$S_0 = 0, S_1 = v_1^h, S_2 = v_1^h + v_2^h, \dots, S_i = v_1^h + v_2^h + \dots + v_i^h, \dots, S_{2n} = 0.$$

Az i -edik részletösszeg, S_i éppen a megrajzolt trajektória ordinátája az i -edik lépésnél. Az $S_0, S_1, S_2, \dots, S_{2n}$ sorozat legnagyobb eleme a trajektória maximális ordinátája amely éppen az $F_n(x)$ és $G_n(x)$ függvények közötti maximális eltérés n -szeresével egyenlő [az $F_n(x)$ függvény javára számítottuk az eltérést]:

$$\max_i S_i = n D_{n,m}^+$$

Mivel a $\{v_n^h\}$ sorozat n db $(+1)$ -est és n db (-1) -est tartalmaz, az összes lehetséges ilyen sorozatok száma $\binom{2n}{n}$, hiszen $2n$ számú helyen n db $(+1)$ -es ennyiféle módon helyezhető el. Ugyaneannyi a sorozatnak megfelelő lehetséges trajektóriák száma.

Számítsuk ki, hogy a kérdéses trajektóriák közül néhány olyan van, amely eléri vagy meghaladja az $y = k$ egyenest!

Ha egy olyan trajektóriát, amely az $y = k$ egyenest eléri az első elérési ponttól kezdve tükrözzünk erre az egyenesre akkor olyan trajektóriát nyerünk, amely a $(2n; 2k)$ pontban végződik. Tehát azon trajektóriák száma, amelyek a sorozatnak felelnek meg, és az $y = k$ egyenest eléri, megegyezik az origóból induló azon trajektóriák számával, amelyek a $(2n; 2k)$ pontban végződnek. Az ilyen trajektóriák száma pedig

$$\binom{2n}{n+k}.$$

Ugyanis, ha egy ilyen (tükrözött) trajektória a -szor lépett felfelé és b -szer lefelé, akkor

$$\begin{aligned} a + b &= 2n, \\ a - b &= 2k, \end{aligned}$$

$$\begin{aligned} 2a &= 2(n + k) \\ a &= n + k \\ b &= n - k. \end{aligned}$$

Márpedig $(n+k)$ darab felfelé mutató nyíl a $2n$ lépés során $\binom{2n}{n+k}$ -féle módon helyezhető el (a többi nyíl pedig lefelé mutat). Tehát ha $F(x) \equiv G(x)$, akkor a két minta minden kölcsönös sorrendje, azaz minden 0-ákból és 1-esekből álló sorozat azonos $\frac{1}{2^n}$ valószínűségű, és így

$$P(\max_i S_i = k) = \frac{\binom{2n}{n+k}}{\binom{2n}{n}}$$

Ha a $H_0 : F(x) \equiv G(x)$ hipotézisről a Gnyegyenko-Koroljuk próbával $1 - \varepsilon$ szinten akarunk dönteni, akkor az $y = k$ egyenest úgy kell megválasztanunk, hogy a fenti valószínűség ε legyen. Minthogy a kombinatorikus formulák kezelése nehézkes, alkalmazzuk a

$$\frac{\binom{2n}{n+k}}{\binom{2n}{n}} \approx e^{-\frac{k^2}{n}}$$

aszimptotikus formulát. Ennek segítségével már könnyen meghatározhatjuk az adott ε szinthez tartozó $y = k$ értéket.
Ha pl. $\varepsilon = 0,05$, akkor

$$\frac{k^2}{e \cdot n} = 0,05$$

$$- \frac{k^2}{n} = \ln 0,05 \approx -3,$$

$$k \approx \sqrt{3n}.$$

Ha tehát mindkét elemszáma n , akkor $\sqrt{3n}$ magasságban megrajzolva az $y = k$ egyenest, a próbát a 46. ábra mintájára grafikusán végrehajthatjuk. Amennyiben a trajektória a $k = 3n$ magasságú egyenest eléri, a $H_0 : F(x) = G(x)$ hipotézist 0,95 szinten elutasítjuk.

A grafikus módszert a próba elméleti alapjainak könnyebb megértése céljából vázoltuk. Maga a próba grafikus ábrázolás nélkül is végrehajtható, hiszen az $S_0, S_1, \dots, S_n$ részletösszeg-sorozatot kiszámítva $\max_i S_i$ meghatározható.

Ez a módszer csak viszonylag kisebb minták esetén célszerű. Hasonlóan végezhető a Gnyegyenko-Koroljuk tételben szereplő másik egyenlőség igazolása. Ehhez a $\max_i S_i$ mennyiséget kell szem előtt tartani, és a tükrözési elvet sokszor egymás után kell alkalmazni. Ennek részletezésétől itt eltekintünk.

Tétel.

Nagy n esetén a Gnyegyenko-Koroljuk próba a Kolmogorov-Szmirnov-féle kétmintás próbával azonos.

Bizonyítás.

Ha a

$$\frac{\binom{2n}{n+k}}{\binom{2n}{n}} \approx e^{-\frac{k^2}{n}}$$

összefüggésben $k = z \sqrt{2n}$ helyettesítést alkalmazunk, akkor figyelembe véve, hogy $\max_i S_i = D_{n,n}^+$ kapjuk a

$$P(\max_i S_i \geq k) \rightarrow P(nD_{n,n}^+ \geq z \sqrt{2n}) \rightarrow e^{-2z^2}$$

összefüggést, tehát

$$\lim_{n \rightarrow \infty} P(nD_{n,n}^+ < z \sqrt{2n}) = \lim_{n \rightarrow \infty} P\left(\sqrt{\frac{n}{2}} D_{n,n}^+ < z\right) = e^{-2z^2} = 1 - e^{-2z^2}$$

ami Szmirnov formulájával egyezik meg.

d.) A Wilcoxon-próba

Ez a próba is a $H_0: F(x) = G(x)$ nullhipotézis ellenőrzésére szolgál, de főként a $G(x) = F(x - \Delta)$ ún. eltolási alternatívával szemben.

A próba a mintaelemek rangszámai segítségével hajtható végre. Ismerkedjünk meg a rangszám fogalmával. Tegyük fel, hogy az $F(x)$ eloszlásfüggvényű X valószínűségi változóra és a $G(x)$ eloszlásfüggvényű Y valószínűségi változóra vonatkozólag az

$$x_1^* < x_2^* < \dots < x_n^*$$

$$y_1^* < y_2^* < \dots < y_m^*$$

rendezett mintákkal rendelkezünk. Rendezzük a két minta elemeit együtt nagyság szerint növekvő sorrendbe (szemléleteség kedvéért mérjük fel a mintaelemeket a számegyenesre), ezután állapítsuk meg, hogy x_1^* hányadik helyre került az egyesített sorozatban. Legyen ez a szám r_1 , ekkor azt mondjuk, hogy x_1 rangja r_1 . Ezután rendre megállapítjuk, hogy $x_2^*, x_3^*, \dots, x_n^*$ hányadik helyre került az egyesített rendezett mintában. Ha x_2^* a r_2 -ik helyre került, akkor x_2 rangja r_2 , s.i.t. Ily módon az X -re vonatkozó rendezett minta elemeinek rangszámait az $r_1, r_2, \dots, r_n$ számok képviselik. Az egyesített mintában a fennmaradó $y_1^*, y_2^*, \dots, y_m^*$ elemek rangjai legyenek $s_1, s_2, \dots, s_m$. A két sorozat nyilván kiadja az egész számokat 1-től $(n+m)$ -ig.

Könnyen belátható, hogy $r_1 < r_2 < \dots < r_n$, továbbá $r_i \geq i$, valamint $s_1 < s_2 < \dots < s_m$ és $s_j \geq j$.

Az r_i és s_j számok valószínűségi változók. Ha igaz a H_0 hipotézis, hogy a két minta azonos eloszlásból származik, akkor a mintaelemek bármely egymás közti elhelyezkedése egyforma valószínűségű. Minthogy $(n+m)$ helyre n db X_i ($\binom{n+m}{n}$)-féle módon helyezhető el, ezért egy meghatározott sorrend valószínűsége $\frac{1}{\binom{n+m}{n}}$.

A Wilcoxon-próbánál próbastatisztikának a

$$W_{n,m} = \sum_{i=1}^n (r_i - i) = \sum_{i=1}^n r_i - \frac{n(n+1)}{2}$$

statisztikát használjuk, amely additív állandótól eltekintve az X_i^* mintaelemek rangszámainak összege. Látható, hogy az X_i^* mintaelemek rangszámainak összegéből azok lehetséges minimális rangszámösszegét vontuk le. Ha ugyanis az $X_1^*, X_2^*, \dots, X_n^*$ mintaelemek rangjai $r_1 = 1, r_2 = 2, \dots, r_n = n$ lennének - vagyis az $\{X_n^*\}$ minta minden eleme kisebb lenne, a $\{Y_m^*\}$ minta legkisebb eleme -, akkor

$$\sum_{i=1}^n r_i = 1 + 2 + \dots + n = \frac{n(n+1)}{2}$$

lenne.

A $W_{n,m}$ statisztika értéke 0 és nm között változhat, eloszlásáról az alábbi tétel informál bennünket, amelyet bizonyítás nélkül közlünk.

Tétel.

H_0 fennállása esetén $W_{n,m}$ várható értékét és szórásnégyzetét a

$$M(W_{n,m}) = \frac{nm}{2},$$

$$D^2(W_{n,m}) = \frac{nm(n+m+1)}{12}$$

formulák szolgáltatják. A $W_{n,m}$ határeloszlása normális eloszlás, azaz $n, m \rightarrow \infty$ esetén:

$$P\left(\frac{W_{n,m} - M(W_{n,m})}{D(W_{n,m})} < x \mid H_0\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

Ha most a H_0 hipotézis felől 95 %-os szinten akarunk dönteni, akkor $W_{n,m}$ meghatározása után kiszámítjuk annak standardizáltját, a

$$W_{n,m}^* = \frac{W_{n,m} - M(W_{n,m})}{D(W_{n,m})}$$

menyiséget, amelyre tehát nagy n és m esetén:

$$P(-2 \leq W_{n,m}^* < 2) \approx 0,95.$$

Ha $W_{n,m}^* \geq 2$, akkor az adott szinten elvetjük H_0 -t és a

$$H_1 : F(x) \neq G(x)$$

alternatív hipotézist fogadjuk el.

Példa.

Vizsgáljuk meg, hogy Szegednél a Tisza évi maximális vízállása növekedett-e az 1941-1978 időszakban az 1901-1940 időszakhoz képest!

Jelöljük X -szel az 1901-1940 időszakban az évi maximális vízállást, és X eloszlásfüggvénye legyen $F(x)$, továbbá jelöljük Y -szal az évi maximális vízállást az 1941-1978 időszakban, az Y eloszlásfüggvénye legyen $G(x)$.

Vizsgáljuk Wilcoxon-próbával a

$$H_0 : F(x) \equiv G(x)$$

hullhipotézist a

$$H_1 : G(x) = F(x - \Delta)$$

ellenhipotézissel szemben.

Az X -re vonatkozó $n = 40$ elemű mintát, valamint az Y -ra vonatkozó $m = 38$ elemű mintát a 4. táblázat tartalmazza.

Az X-re vonatkozó mintaelemek rangjainak összege:

$$\sum_{i=1}^{40} r_i = 1558.$$

A próbastatisztika:

$$\begin{aligned} W = W_{40,38} &= \sum_{i=1}^{40} r_i - \frac{n(n+1)}{2} = 1558 - \frac{40 \cdot 41}{2} = \\ &= 1558 - 820 = 738. \end{aligned}$$

$$M(W) = \frac{40 \cdot 38}{2} = 760.$$

$$D^2(W) = \frac{nm(n+m+1)}{12} = \frac{40 \cdot 38 \cdot 79}{12} = 1006.$$

$$D(W) = \sqrt{1006} = 32.$$

$$W^* = \frac{W - M(W)}{D(W)} = \frac{738 - 760}{32} = \frac{22}{32} \approx -0,7.$$

Mivel W^* aszimptotikus standard normális eloszlású:

$$P(-2 \leq W^* < 2) \approx 0,95,$$

ezért W^* akutális értéke alapján nincs okunk a H_0 hipotézis elvetésére. Úgy döntünk, hogy a Tisza évi maximális vízállása Szegednél a két vizsgált időszakban egyforma eloszlást követ.

Végezzük el ugyanezt a homogenitásvizsgálatot χ^2 próbával is! Ha a Tisza évi maximális vízállásának változási tartományát az $A_1 = [301 \text{ cm}, 600 \text{ cm}]$, $A_2 = [601 \text{ cm}, 700 \text{ cm}]$, $A_3 = [701 \text{ cm}, 800]$, $A_4 = [801 \text{ cm}, 1000 \text{ cm}]$ részintervallumokra osztjuk, akkor az X, ill., Y változók értékei alapján az A_1, A_2, A_3, A_4 teljes eseményrendszer tagjainak alábbi gyakoriságait kapjuk:

$$X : \quad \nu_1 = 17, \quad \nu_2 = 9; \quad \nu_3 = 9; \quad \nu_4 = 5; \quad \sum_{i=1}^4 \nu_i = 40.$$

3.1 táblázat

Év	(cm)	r	Év	n (cm)
1901	680	47	1941	855
1902	668	46	1942	780
1903	508	15	1943	366
1904	450	6	1944	654
1905	518	17	1945	560
1906	550	23	1946	525
1907	758	59	1947	602
1908	595	30	1948	714
1909	642	40	1949	495
1910	496	13	1950	517
1911	563	25	1951	550
1912	753	58	1952	648
1913	802	70	1953	706
1914	778	64	1954	454
1915	791	67	1955	657
1916	791	68	1956	689
1917	614	36	1957	604
1918	349	2	1958	730
1919	916	76	1959	436
1920	708	54	1960	582
1921	325	1	1961	394
1922	774	63	1962	820
1923	637	38	1963	587
1924	870	75	1964	764
1925	681	48	1965	748
1926	759	60	1966	799
1927	488	11	1967	790
1928	542	21	1968	600
1929	458	8	1969	626
1930	496	14	1970	961
1931	603	33	1971	521
1932	923	77	1972	606
1933	660	45	1972	475
1934	526	20	1974	807
1935	594	29	1975	692
1936	472	9	1976	654
1937	703	52	1977	767
1938	638	39	1978	684
1939	579	26		
1940	817	73		

$$Y : \mu_1 = 14; \mu_2 = 11; \mu_3 = 9; \mu_4 = 4; \sum_{i=1}^4 \mu_i = 38.$$

Képezzük a

$$\chi^2 = nm \sum_{i=1}^4 \frac{(\frac{y_i}{n} - \frac{y_i}{m})^2}{\frac{y_i}{n} + \frac{y_i}{m}}$$

statisztikát:

$$\begin{aligned} \chi^2 &= 40 \cdot 38 \frac{(\frac{17}{40} - \frac{14}{38})^2}{\frac{17}{40} + \frac{14}{38}} + \frac{(\frac{9}{40} - \frac{11}{38})^2}{\frac{9}{40} + \frac{11}{38}} + \\ &+ \frac{(\frac{9}{40} - \frac{9}{38})^2}{\frac{9}{40} + \frac{9}{38}} + \frac{(\frac{5}{40} - \frac{4}{38})^2}{\frac{5}{40} + \frac{4}{38}} = 0,26 \end{aligned}$$

A fenti χ^2 -statisztika 3 szabadságfokú χ^2 -eloszlást követ. Ha a H_0 hipotézisről 0,95 szinten kívánunk dönteni, akkor a Függelék V. táblázatának 3. sorában az $\varepsilon = 0,05$ oszlopban a kiritikus érték: 7,82. Nincs tehát okunk a H_2 hipotézis elvetésére.

Ugyanezt a hipotézisvizsgálatot elvégezhetjük a Gnyegyenko-Koroljuk-próbával is. Ha az első időszak adataiból találomra kettőt elhagyunk (pl. sorsolás eredményeként a 325 cm és 349 cm adatokat), akkor mind X-re, mind Y-ra vonatkozólag $n = m = 38$ adatunk van. Ezekből egyetlen 76 elemű nagyság szerint növekvő sorrendű rendezett mintát $\{\xi_n^x\}$ képezve, és +1-et írva ha $\xi_i^x = x_k^x$, (-1)-et írva ha $\xi_i^x = y_j$, az alábbi sorozatot nyerjük.

-1, -1, -1, 1, -1, 1, 1, -1, 1, -1, 1, 1, 1, -1, 1, -1, -1,
-1, -1, 1, 1, -1, 1, -1, 1, 1, -1, -1, 1, 1, -1, 1, -1, 1,
1, -1, -1, 1, -1, 1, 1, 1, -1, -1, -1, -1, -1, 1, 1, 1,
-1, -1, -1, 1, -1, 1, -1, -1, -1, 1, 1, 1, 1, -1, -1, -1,
1, 1, -1, -1, 1, 1, -1, 1, -1, -1, 1, -1,
1, 1, 1, -1.

A $S_0, S_1, \dots, S_{2n}$ részletösszegeket képezve azt kapjuk, hogy $\max_k S_k = 3$.

Ha $\xi = 0,05$ értéket választjuk, akkor a kritikus határ $\max_k S_k$ -ra a $k = \sqrt{3n} = \sqrt{114} \approx 10$, tehát H_0 -t nem vehetjük el.

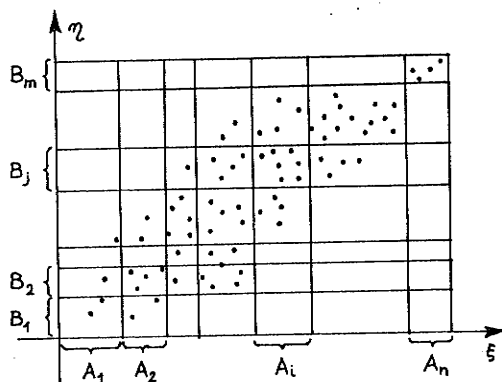
3.5.3 Függetlenségvizsgálat χ^2 -próbával

A műszaki gyakorlatban a függetlenségvizsgálat szükségessége sokszor felmerül. Miként a 2.1.3 pontban láttuk, ha két valószínűségi változó, X és Y függetlenek, akkor együttes eloszlásfüggvényük az egyes változók eloszlásfüggvényeinek szorzata:

$$H(x, y) = P(X < x, Y < y) = P(X < x) P(Y < y) = F(x)G(y).$$

Két valószínűségi változó függetlensége nagymértékben megkönnyíti együttes bekövetkezésük valószínűségének kiszámítását, ugyancsak egyszerűbb kiszámítani a két változó összegének, különbségének, szorzatának és hányadosának eloszlásfüggvényét. Ugyanakkor, ha két valószínűségi változó független egymástól – vagyis nincs közöttük sztochasztikus kapcsolat –, akkor egyik változó megfigyelt értékéből semmi információt nem nyerünk a másik változó aktuális értékére vonatkozólag.

Függetlenségvizsgálatot χ^2 -próbával a következő módon végezhetünk. A síknak az 56. ábrán látható felosztása az $A_1, A_2, \dots, A_n$ és $B_1, B_2, \dots, B_m$ teljes eseményrendszereket szolgáltatja.



56. ábra

A

$$H_0 : (PX < x; Y < y) = P(X < x) P(Y < y) = F(x) G(y)$$

nullhipotézis ellenőrzésénél tegyük fel először, hogy az $F(x)$ és $G(x)$ vetületeloszlásokat (X , ill. Y egydimenziós eloszlását) ismerjük. Legyen

$$P(x_{i-1} \leq X < x_i) = P(A_i) = p_i \quad (i=1,2,\dots,n),$$

$$P(y_{j-1} \leq Y < y_j) = P(B_j) = q_j \quad (j=1,2,\dots,m).$$

Mivel az $A_i B_j$ események ($1 \leq i \leq n$, $1 \leq j \leq m$) is teljes eseményrendszert alkotnak, H_0 fennállása esetén

$$P(A_i B_j) = p_i q_j$$

Tegyük fel, hogy az $(x_{i-1}; y_{j-1})$, $(x_{i-1}; y_j)$, $(x_i; y_{j-1})$ és $(x_i; y_j)$ csúcspontok által meghatározott téglalapba v_{ij} számú mintaelem esik.

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(v_{ij} - p_i q_j)^2}{N p_i q_j}$$

statisztikát, amely nagy N esetén (ahol N a minta elemszáma) közelítőleg χ^2 -eloszlást követ $(n-1)(m-1)$ szabadságfokkal. Az $(1-\varepsilon)$ döntési szint megválasztása után tehát a kritikus értéket a χ^2 -eloszlás táblázatában az $(n-1)(m-1)$ -edik sorban, ε oszlopában találjuk. Ha statisztikánk aktuális értéke nagyobb, mint a kritikus érték, a függetlenség hipotézisét $(1-\varepsilon)$ szinten elvetjük.

A gyakorlatban sokszor a p_i és q_j számokat sem ismerjük, tehát az egydimenziós eloszlások sem ismertek. Ha az A_i esemény összesen

$$v_{i\cdot} = \sum_{j=1}^m v_{ij} \text{-szer}$$

fordult elő, a B_j esemény gyakorisága pedig

$$v_{\cdot j} = \sum_{i=1}^n v_{ij}$$

(az i -edik oszlopba, ill. a j -edik sorba eső mintapontok gyakoriságai), akkor a

$$p_i \approx \frac{v_i}{N}, \quad q_j \approx \frac{v_j}{N}$$

közelítés alapján a próbastatisztika:

$$\chi^2 = N \sum_{i=1}^n \sum_{j=1}^m \frac{(v_{ij} - \frac{v_{i.} \cdot v_{.j}}{N})^2}{\frac{v_{i.} \cdot v_{.j}}{N}}$$

amely ugyancsak χ^2 -eloszlású $(n-1)(m-1)$ szabadságfokkal. Ebben az esetben tehát a függetlenségvizsgálathoz az eloszlások ismeretére nincs szükség.

Amennyiben a függetlenségvizsgálat a H_0 hipotézis elutasításához vezet, akkor a két változó között korreláció- és regresszióelemzést kell végeznünk, amellyel a következőkben foglalkozunk.

3.6 A döntésfüggvények elméletének elemei

Amit láttunk, minden statisztikai eljárás valamilyen döntéssel végződik. Becslési problémánál a döntés abból áll, hogy elfogadunk valamilyen értéket a paraméter igazi értéke helyett. Valamely hipotézis elfogadása vagy elvetése tipikus döntés két alternatíva között. Egy konfidenciaintervallum meghatározása során a döntés arra vonatkozik, hogy az ipari paraméterérték adott intervallumba esik. Mindezek a döntések statisztikai mintán alapulnak, azért azokat statisztikai döntéseknek nevezzük. Mindegyik esetben a mintaelemekből alkotott függvény, tehát egy statisztika a döntés eszköze. Az alkalmas statisztika alapvető feladata. Például egy valószínűségi változó szórásnégyzetének becslésére az s_n^{*2} korrigált empirikus szórásnégyzetet javasoltuk, amely torzítatlan és erősen konzisztens becslése a szórásnégyzetnek, viszont $X_n^* - X_1^*$ mintaterjedelem segítségével sokkal egyszerűbben és minimális számolással tudjuk becsülni a szórásnégyzetet. Hipotézisvizsgálathoz kicsiny első- és másodfajú hibákat adó próbákat javasoltunk, amelyek viszont néha igen nagy megfigyelésszámot követelnek. Számos esetben a statisztikai eljárás megválasztásakor gazdasági követelményeket és figyelembe kell venni. A műszaki gyakorlatban a paraméterbecslési vagy hipotézisvizsgálati eljárás során hozott döntés egyben

meghatározhat bizonyos gyakorlati intézkedéseket, "döntéseket" amelyek helyesek vagy hibásak lehetnek a véletlen játéka következtében. A hibás döntés jelentős gazdasági kárt, veszteséget okozhat. A döntés alapjául valamely megfigyelt valószínűségi változó aktuális értéke (vagy értékeiből képezett statisztika) szolgál, ezért döntéseink bizonyos kockázattal járnak. Hogyan kell egy döntés kockázatát kiszámítani? Hogyan lehet olyan döntést választani, amely a lehető legkisebb kockázatú? Ezekre a kérdésekre a statisztikai döntésmélet segítségével válaszolhatunk.

3.6.1 A statisztikai döntési eljárás

A statisztikai döntés fogalmát és a döntési eljárást menetét egy példán keresztül mutatjuk be. Korábbi vizsgálataink alapján ismeretes, hogy valamely folyó árhullámainak tetőzési értéke exponenciális eloszlású:

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

ahol ν számunkra ismeretlen, tehát becsülendő paraméter, azaz ν numerikus értékére döntést kell hoznunk. A becslés eredménye valamely $\hat{\nu} = d$ valós szám. Amikor a becslés eredményeként a $\hat{\nu} = d$ ítéletet hozzuk, akkor statisztikai döntést végzünk. Példánkban d a pozitív félegyenes bármely pontja lehet. Az összes lehetséges d döntések halmaza tehát: $D = \{0 \leq d \leq \infty\}$. A D halmazt döntéstérnek nevezzük. Vizsgáljuk meg, hogy milyen alapon hozhatjuk a $\hat{\nu} = d$ döntést.

A ν paraméter értékét közvetlenül nem lehet megfigyelni, azonban többször is mérhető X valószínűségi változó (pl. a tetőzés) értéke. Legyen az X valószínűségi változóra vonatkozó statisztikai minta: $X = (X_1, X_2, \dots, X_n)$. X valószínűségi változó eloszlása a ν paramétertől, ezért X minden megfigyelt értéke információt hordoz ν -ra. Természetesen X vektortváltozó, tehát a minta általában még több információt hordoz ν -ra, mint egyetlen megfigyelés. Ezt az információt használjuk fel a ν -ra vonatkozó döntéshez, azaz X értékeitől tesszük függővé, hogy melyik $\hat{\nu} = d$ döntést hozzuk, azaz $d = \sigma(X)$. A $\sigma(X)$ szabályt - tehát azon előírást, hogy X észlelt értékéhez melyik d számot rendeljük - döntésfüggvénynek nevezzük. Minthogy példánkban ν az exponenciális eloszlás várható értékének recipokával egyenlő, ezért esetünkben ésszerű döntési szabály lehet pl.

$$\sigma_1(X) = \frac{1}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}};$$

$$\sigma_2(X) = \frac{1}{\sum_{i=1}^n p_i x_i}, \quad \sum_{i=1}^n p_i = 1;$$

$$\sigma_3(X) = \frac{2}{x_1^* + x_n^*}$$

A továbbiakban a jelölés egyszerűsége kedvéért a döntés alapjául szolgáló valószínűségi változókat X -szel fogjuk jelezni akkor is, ha X vektorváltozó. Amikor $X = (X_1, X_2, \dots, X_n)$ statisztikai minta, akkor a benne foglalt információt valamely $t(X) = t(X_1, \dots, X_n)$ statisztikába tömörítjük (lehetőleg elégséges statisztikát használva), amely ugyancsak valószínűségi változó, és ez szolgál a ψ paraméterre vonatkozó döntés alapjául.

A fenti példa alapján a statisztikai döntési eljárás kissé absztrakt módon - de igen általános érvényűen - a következőképpen fogalmazható: adott az M mintatér, megfigyeljük ennek egy $X \in M$ elemét, és ennek alapján választunk egy d döntést a lehetséges döntések D halmazából, amelyet a gyakorlati probléma határoz meg. A D halmazt döntéstérnek nevezzük. A döntés megválasztása bizonyos szabály alapján történik. Ezt a szabály döntésfüggvénynek nevezzük, és $\sigma(X)$ -szel jelöljük. A $\sigma(X)$ döntési szabály értelmében tehát azt a $d \in D$ döntést választjuk, amely az X megfigyelési eredményhez tartozik, azaz $d = \sigma(X)$.

3.6.2 Veszteségfüggvény és kockázatfüggvény

Különböző $\sigma(X)$ döntésfüggvények adott X esetén általában különböző d numerikus értéket szolgáltatnak, ezért felmerül a kérdés, hogy melyiket válasszuk közülük. A válaszadás érdekében meg kell vizsgálnunk döntésünk következményeit. Bizonyos esetekben megítélhető a rossz döntésből (azaz ha ψ nem egyenlő d -vel) származó kár vagy veszteség. Minél jobban eltér a döntés d numerikus értéke ψ igazi értékétől, annál nagyobb veszteséget okozhat a rossz döntés. A $\psi = d$ döntés numerikus értéke X ismeretében csak attól függ, hogy melyik $\sigma(X)$ döntésfüggvényt használjuk, így a döntés által okozott veszteség a $\sigma(X) \in [\psi, \sigma(X)]$ -szel, ha ψ a paraméter igazi értéke és a $\sigma(X)$ döntésfüggvényt használjuk (a továbbiakban legyen $X = x$).

Ilyen veszteség lehet pl:

$$L[\psi, \sigma(x)] = [\psi - \sigma(x)]^2.$$

vagy

$$L[\psi, \sigma(x)] = |\psi - \sigma(x)|.$$

Mivel X valószínűségi változó, ugyanazon döntésfüggvény a véletlen játéka következtében különböző d numerikus értékeket szolgáltat, tehát adott döntésfüggvény használata esetén különböző veszteségek állnak elő. Éppen ezért adott $\sigma(X)$ döntésfüggvény által előidézett átlagos veszteséget kell tekintenünk ahhoz, hogy meg tudjuk ítélni a döntésfüggvény hatékonyságát ψ igazi értékének megközelítése szempontjából.

Az adott $\sigma(X)$ döntésfüggvény által okozott átlagos veszteséget az

$$\begin{aligned} R_{\sigma(x)}(\psi) &= E_{\psi}[L(\psi, \sigma(x))] = \\ &= \int L[\psi, \sigma(x)] dF(x, \psi) \end{aligned}$$

függvény (mint ψ függvényét) a $\sigma(x)$ döntésfüggvény kockázatfüggvényének nevezzük. A kockázatfüggvény tehát a veszteségfüggvény feltételes várható értéke, azon feltétel mellett, hogy ψ a paraméter igazi értéke. $F(x, \psi)$ az X valószínűségi változó azon eloszlásfüggvénye, amelynél ψ értéket feltételezzük. A paraméterbecslés problémája tehát a statisztikai döntés egyik esete.

Vizsgáljuk meg most a hipotézisvizsgálat esetét mint döntési problémát! Legyen pl. valamely folyó évi maximális vízállása $N(\psi, \sigma_0)$ normális eloszlású. Ekkor a σ_0 ismert, és a ψ paraméterre a $H_0: \psi \leq \psi_0$ és $H_1: \psi > \psi_0$ hipotézisek között kívánunk dönteni. Jelölje H_0 elfogadását a d_1 döntés és H_1 elfogadását d_2 döntés. A D döntéstér tehát két pontból áll, azaz $D = \{d_1, d_2\}$. Az M mintatér [amennyiben a döntés alapjául az $X = (X_1, X_2, \dots, X_n)$ statisztikai minta szolgál] az R_n n -dimenziós euklideszi tér. A ψ paraméter lehetséges értékeinek halmaza Θ (ún. paramétertér): $\Theta = (-\infty, \infty)$. A lehetséges $\sigma(X)$ döntésfüggvények halmaza mindazon függvényekből áll, amelyek az M mintateret leképezik a $D = \{d_1, d_2\}$ halmazra, és amelyek tulajdonságai olyanok, hogy $P_{\psi}[d(x) = d_1]$ egyértelműen meghatározott minden $\psi \in \Theta$ -ra.

Legyen a veszteségfüggvény:

$$L(\psi, d_1) = \begin{cases} 1_1, & \text{ha } \psi > \psi_0 \\ 0, & \text{ha } \psi \leq \psi_0 \end{cases}$$

$$L(\psi, d_2) = \begin{cases} 0, & \text{ha } \psi > \psi_0 \\ 1_2, & \text{ha } \psi \leq \psi_0 \end{cases}$$

ahol 1_1 és 1_2 pozitív számok. (A veszteségfüggvényt rendszerint úgy választjuk, hogy helyes döntés esetén a veszteség zérus legyen.) A kockázatfüggvény ekkor:

$$R[\psi, \sigma(x)] = \begin{cases} 1_1 P_\psi [\sigma(x) = d_1], & \text{ha } \psi \geq \psi_0 \\ 1_2 P_\psi [\sigma(x) = d_2], & \text{ha } \psi \leq \psi_0 \end{cases}$$

A döntés során kétféle hibát követhetünk el. Amikor $\psi > \psi_0$ akkor $P_\psi [\sigma(x) = d_1]$ annak valószínűsége, hogy a d_1 döntést választva követünk el hibát, mivel a d_2 döntést kellene választanunk (feltéve, hogy ψ a paraméter igazi értéke). Hasonlóan $\psi \leq \psi_0$ esetén.

$$P_\psi [\sigma(x) = d_2] = 1 - P_\psi [\sigma(x) = d_1]$$

annak a valószínűsége, hogy a d_2 döntés választásával követünk el hibát, mivel a d_1 döntést kellene választanunk (feltéve, hogy ψ a paraméter igazi értéke).

A fenti példával érzékeltetni kívántuk, hogy a hipotézisvizsgálat problémája is tárgyalható a statisztikai döntéselmélet keretében.

3.6.3 A Bayes-féle döntési elv

Az előzőekben vázoltuk, hogy ha $F(x, \psi)$ eloszlásfüggvényű X valószínűségi változó eloszlásának ψ paraméterére kívánunk dönteni, és a $\sigma(x)$ döntésfüggvényt használjuk, akkor a kockázatfüggvény:

$$R_{\sigma(x)}(\psi) = E_\psi [L(\psi, \sigma(x))] = \int_M L[\psi, \sigma(x)] dF(x, \psi)$$

a ψ paraméter függvénye, ami azt jelenti, hogy ψ igazi értékétől függően egy adott $\sigma(x)$ döntésfüggvény kockázata igen különböző lehet.

Amikor két különböző döntésfüggvény, pl. $\sigma_0(x)$ és $\sigma_1(x)$ kockázatát kiszámítjuk, és azt találjuk, hogy egy adott ψ^* esetében

$$(d_1) \quad R \quad \sigma_0(x) (\psi^*) < R \quad \sigma_1(x) (\psi^*)$$

akkor úgy ítélünk, hogy ψ^* esetén $\sigma_0(x)$ jobb becslés, mint $\sigma_1(x)$. A $\sigma_0(x)$ döntésfüggvényt egyenletesen jobbnak nevezzük $\sigma_1(x)$ -nél, ha

$$(d_2) \quad R \quad \sigma_0(x) (\psi) \leq R \quad \sigma_1(x) (\psi)$$

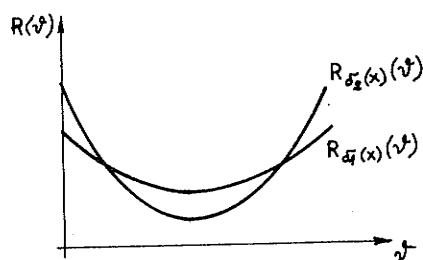
minden ψ -ra fennáll és legalább egy ψ^* paraméterértékre teljesül a (d_1) összefüggés.

Amikor $R \quad \sigma_0(x) (\psi) = R \quad \sigma_1(x) (\psi)$ minden ψ értékre, akkor a $\sigma_0(x)$ és $\sigma_1(x)$ döntésfüggvényeket ekvivalenseknek nevezzük.

Amennyiben egy adott döntési problémánál létezik egy olyan $\sigma_0(x)$ döntésfüggvény, amelyre a (d_1) és (d_2) relációk teljesülnek bármilyen más $\sigma_1(x)$ döntésfüggvényre, akkor $\sigma_0(x)$ az egyenletesen legjobb döntésfüggvény.

Amikor egy adott döntési problémánál azt találjuk, hogy $\sigma_0(x)$ egyenletesen jobb döntésfüggvény, mint $\sigma_1(x)$, akkor a $\sigma_1(x)$ döntésfüggvény alkalmazása nem jöhet szóba, tehát ekkor $\sigma_1(x)$ megengedhetetlen döntésfüggvény.

Amennyiben viszont az adott döntési problémára nem létezik egyenletesen legjobb döntésfüggvény, akkor nem tehetünk jobbat, minthogy egy megengedhető döntésfüggvényt használunk. Amikor mind $\sigma_1(x)$ mind $\sigma_2(x)$ megengedhető döntésfüggvények, akkor ez azt jelenti, hogy $\sigma_1(x)$ bizonyos ψ értékekre jobb, mint $\sigma_2(x)$ (azaz kisebb kockázatú), más ψ értékekre viszont $\sigma_2(x)$ a jobb. Melyiket válasszuk tehát a két döntésfüggvény (57. ábra) közül?



57. ábra

Amennyiben ismernénk ψ igazi értékét, akkor könnyű lenne választani $\sigma_1(x)$ és $\sigma_2(x)$ között, de akkor nem is kellene döntenünk. Amikor legalább azt ismernénk, hogy ψ értéke milyen valószínűséggel esik adott intervallumba, akkor is könnyebb lenne a döntés.

Tekintsük a ψ paramétert valószínűségi változónak, és tegyük fel, hogy ψ a priori eloszlását ismerjük (pl. korábbi tapasztalatok alapján).

Legyen a ψ valószínűségi változó eloszlásfüggvénye a Θ paraméterterén $\tau(\psi)$. Adott $\tau(\psi)$ a priori eloszlás mellett keressük most olyan $\sigma(x)$ döntésfüggvényt, amelyre az

$$r(\tau, \sigma) = \int_{\Theta} R_{\sigma(x)}(\psi) d\tau(\psi)$$

ún. Bayes-féle kockázat minimális lesz, ahol az integrandusban szereplő kockázatfüggvény:

$$(r_1) R_{\sigma(x)}(\psi) = \int_M L[\psi, \sigma(x)] dF(x|\psi).$$

Azt a $\sigma(x) = d$ döntést, amelyre a (r_1) kockázat minimális, a $\tau(\psi)$ a priori eloszlásához tartozó Bayes-döntésnek nevezzük. A Bayes-döntés tehát minimális Bayes-kockázatú döntés.

Az (r_1) Bayes-kockázat kiszámításához vegyük figyelembe a teljes valószínűség tételének következő folytonos analogonjait:

$$F(x) = \int_{\Theta} F(x|\sigma) d\tau(\psi),$$

valamint

$$\tau(\psi) = \int_M \tau(\psi|x) dF(x).$$

Ezek alapján:

$$\begin{aligned} r(\tau, \nu) &= \int_{\Theta} \left\{ \int_{\mathcal{M}} L(\nu, \sigma(x)) dF(x | \nu) \right\} d\tau(\nu) = \\ &= \int_{\mathcal{M}} \left\{ \int_{\Theta} L(\nu, \sigma(x)) d\tau(\nu | x) \right\} dF(x) \end{aligned}$$

a minimalizálandó Bayes-kockázat (feltételezzük, hogy a fenti kettős integrálban az integrálás sorrendje felcserélhető).

A fenti Stieltjes-integrálos jelölés helyett szokásosabb a Riemann-integrálos jelölés a sűrűségfüggvények segítségével. Legyen

$$\begin{aligned} \frac{d\tau(\nu)}{d\nu} &= g(\nu), \quad \frac{d\tau(\nu | x)}{d\nu} = g(\nu | x), \\ \frac{dF(x)}{dx} &= f(x), \end{aligned}$$

ekkor

$$r(\tau, \sigma) = \int_{\mathcal{M}} \left\{ \int_{\Theta} L(\nu, \sigma(x)) g(\nu | x) d\nu \right\} f(x) dx.$$

A kettős integrállal kifejezett $r(\tau, \nu)$ mennyiség minimalizálásához tehát az

$$\int_{\Theta} L(\nu, \sigma(x)) g(\nu | x) d\nu = E_{\nu} [L(\nu, d) | X = x]$$

belső integrál értékét kell minimalizálni.

A $g(\nu | x)$ feltételes sűrűségfüggvényt a ν valószínűségi változó a posteriori sűrűségfüggvényének nevezzük. A Bayes-döntés tehát az a $\sigma(x)$ döntésfüggvény, amely minimalizálja az a posteriori (feltételes) várható veszteségeket. Tekintsük most azt az esetet, amikor

$$L(\nu, \sigma(x)) = (\nu - d)^2; \quad d = \sigma(x),$$

ha $X = x$.

Ekkor A Bayes-döntés az a $\sigma(x) = d$ döntés, amelyre

$$E_{\nu} [L(\nu, d) | X = x] = \int_{\Theta} (\nu - d)^2 g(\nu | x) d\nu = \min$$

Ekkor

$$\frac{\partial E_{\psi} [L(\psi, d) | X = x]}{\partial d} = -2 \int_{\theta} (\psi - d) g(\psi | x) d\psi = 0$$

Innen:

$$d = \frac{\int_{\theta} \psi g(\psi | x) d\psi}{\int_{\theta} g(\psi | x) d\psi} = \int_{\theta} \psi g(\psi | x) d\psi$$

Tehát $\sigma(x) = d = E(\psi | X = x)$ a ψ paraméter feltételes várható értéke az adott a posteriori sűrűséggel számítva.

Amennyiben paraméterbecslés esetén a veszteségfüggvény

$$L[\psi, \sigma(x)] = |\psi - \sigma(x)|,$$

akkor a Bayes-döntés a ψ paraméter a posteriori eloszlásának mediánja. Ekkor ugyanis az

$$E_{\psi} [L(\psi, d) | X = x] = \int_{\theta} |\psi - d| g(\psi | x) d\psi$$

kifejezést kell minimalizálni, ahol $d = \sigma(x)$.

Tegyük fel, hogy $\theta = (-\infty, \infty)$, ekkor:

$$(r_2) \int_{-\infty}^{\infty} |\psi - d| g(\psi | x) dx = \int_{-\infty}^d (d - \psi) g(\psi | x) d\psi + \int_d^{\infty} (\psi - d) g(\psi | x) d\psi$$

$$= d \int_{-\infty}^d g(\psi | x) d\psi - \int_{-\infty}^d \psi g(\psi | x) d\psi + \int_d^{\infty} \psi g(\psi | x) d\psi - d \int_d^{\infty} g(\psi | x) d\psi =$$

$$+ \int_0^{\infty} \psi g(\psi | x) d\psi - d \int_d^{\infty} g(\psi | x) d\psi = \min$$

Vezessük be az $\int_{-\infty}^{\infty} \psi g(\psi | x) d\psi = m$ jelölést ψ a posteriori várható értékére, továbbá vegyük figyelembe, hogy

$\int_{-\infty}^{\infty} g(\psi | x) d\psi = 1$, ekkor a () összefüggés így írható:

$$(r_3) \int_{-\infty}^{\infty} |\psi - d| g(\psi | x) d\psi = d \int_{-\infty}^d g(\psi | x) d\psi - \int_{-\infty}^d \psi g(\psi | x) d\psi +$$

$$+ \int_d^{\infty} \psi g(\psi | x) d\psi - d \int_d^{\infty} g(\psi | x) d\psi = d [1 - \int_{-\infty}^d \psi g(\psi | x) d\psi] = \min$$

Az (r_3) összefüggés d szerinti minimuma csak ott lehet, ahol a d szerinti derivált zérus. Képezzük ezért a kifejezés d szerinti deriváltja:

$$\begin{aligned}
 & - \int_{-\infty}^d g(\psi | x) d\psi + d g(d|x) - d g(d|x) - dg(d|x) = 1 + \\
 & + \int_{-\infty}^d g(\psi | x) d\psi + dg(d|x) = 0
 \end{aligned}$$

Innen:

$$\int_{-\infty}^d g(\psi | x) d\psi = \frac{1}{2}$$

Tehát d a ψ paraméter a posteriori eloszlásának mediánja.

Megjegyezzük, hogy a Bayes-féle kockázat kiszámítása és a Bayes-döntés megtalálása a ψ -ra vonatkozó a priori eloszlás ismeretében akkor is lehetséges, ha ψ nem X valószínűségi változó eloszlásának paramétere. Amennyiben X és ψ két olyan valószínűségi változó, amelyek között függőség, sztochasztikus kapcsolat van és ismerjük egymásra vonatkozó feltételes eloszlásaikat, valamint az egyik változó eloszlását, akkor a fentebb ismertetett eljárás alkalmazható.

3.7 Korreláció- és regresszióelmélet

3.7.1 A valószínűségi változók közötti sztochasztikus kapcsolatokról

A műszaki gyakorlatban és kutatásban a korreláció és regresszióelmélet igen fontos segédeszköz különböző mennyiségek közötti összefüggések vizsgálatára.

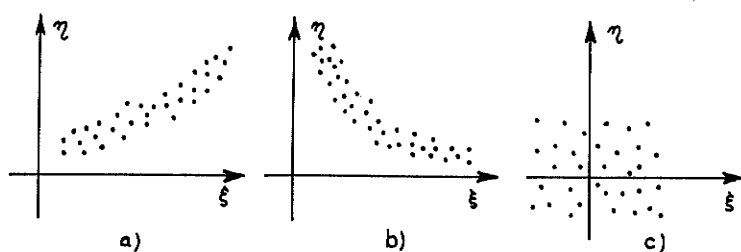
A korrelációelmélet tárgya valószínűségi változók közötti kapcsolat szorosságának vizsgálata, a kapcsolat szorosságának mérése alkalmas mérőszámok konstruálása, azok statisztikai mintából történő becslése, és a becslés tulajdonságainak vizsgálata.

A valószínűségi változók közötti kapcsolat mérőszámai közül megismerkedünk az ún. korrelációs együtthatóval és annak néhány tulajdonságával. Más mérőszámot is ismertetünk a kapcsolat jellemzésére, továbbá foglalkozunk az ismertetett mérőszámok statisztikai becslésével.

A regresszióelmélet tárgya valószínűségi változók közötti kapcsolat függvényalakjának vizsgálata. Az ide tartozó gyakorlati problémák különböző jellegűek lehetnek. A leggyakrabban felmerülő feladatokat általában a következő kategóriák valamelyikébe sorolhatjuk:

1. Két vagy több valószínűségi változó közötti sztochasztikus kapcsolat vizsgálata. Ilyen feladatot kell megoldanunk, amikor meg akarjuk határozni, milyen kapcsolat van adott méretű acélrúd maximális teherbírása, valamint a rúd kezdeti görbesége, a rúdban lévő feszültségek stb. között. Vizsgálhatjuk adott betonkockán áthaladó ultrahang terjedési sebessége és a kockák törési szilárdsága közötti kapcsolatot (roncsolásmentes szilárdságvizsgálat céljából). Kereshetjük valamely folyó vízgyűjtőterületén adott időintervallumban lehullott csapadék mennyisége és a folyó vízállása vagy vízhozama közötti kapcsolatot stb. E feladatok közös jellemzője, hogy a bennük szereplő változók mindegyike valószínűségi változó, s a függetleneknek tekintett változók értékei általában nem határozzák megegyértelműen a függő változó értékeit (nem tudjuk azt kiszámítani a független változók aktuális értékeiből). Ha két valószínűségi változó X és Y közötti kapcsolatot vizsgáljuk, akkor - a mért összehasonlító $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$ értékpárokat síkbeli pontthalmazként ábrázolva - a "pontfelhő" alakja utalhat a változók közötti közös - pl. monoton növekvő vagy csökkenő - tendenciára, amint azt az 58.a.) és b.) ábra mutatja. Ilyenkor a pontfelhő alakjától függően igyekszünk valamilyen jól kezelhető függvény (pl. egyenes vagy alacsony fokszámú polinom) segítségével kifejezni a közös tendenciát. Előfordulhat, hogy a pontfelhő alakja azt sugallja, hogy semmiféle kapcsolat nincs a két valószínűségi változó között (lásd: 50.c.) ábrát), ilyenkor célszerű függetlenségvizsgálatot végezni.

2. Egy valószínűségi változónak egy vagy több funkcionális változótól való függésének vizsgálata. Ez az eset fordul elő pl. ha vizsgáljuk egy jármű féktávolságának a sebességtől való függését, amelynél a sebességet "beállítjuk", a féktávolság adott útviszonyok mellett is valószínűségi változó, amelynek értékét nyilván nem tudjuk előírni. Hasonló a helyzet, ha textiltípust a fonalszakadások számát vizsgáljuk a motorfordulatszám függvényében. Ilyenkor tehát a független változó (vagy változók) nem valószínűségi változók, hanem beállított értékek (amelyek meghatározása kísérlettervezési feladat); a függő változó valószínűségi változó, amelynek eloszlásáról a független változók rögzített értéke esetén beszélhetünk.



58. ábra

3. Ismert, egzakt fizikai törvényszerűséget kifejező

$$y = f(x_1, x_2, \dots, x_s; a_1, a_2, \dots, a_r)$$

egyenletben az ismeretlen paraméterek meghatározása. Az ebbe a kategóriába tartozó feladatoknál sem az $x_1, \dots, x_s$ független változók, sem az y függvényérték nem valószínűségi változók, s a paraméterek ismeretében a független változók adott értékeihez tartozó y érték kiszámítható. A paraméterek azonban rendszerint esetről esetre meghatározandók a független változók és a függő változó összetartozó mért értékei alapján, tehát mérési eredményekből. A paraméterek meghatározását éppen az ismert f függvénykapcsolatra támaszkodás teszi lehetővé. Külön probléma annak eldöntése, hogy a független változók mely értékeit írjuk elő mérési helynek, valamint hány mérést kell végezni, hogy adott számú ismeretlen paramétert meghatározhassunk. E problémakörrel nem foglalkozunk könyvünkben.

A felsorolt problémátípusok közös vonása, hogy mennyiségek közötti összefüggéseket vizsgálunk kvalitatív és kvantitatív szempontból. E vizsgálat fő módszere a "legkisebb négyzetek elvének" alkalmazása, amelyet a pontban ismertetünk.

3.7.2 Korreláció- és regresszióelmélet két valószínűségi változó esetén

Két valószínűségi változó - X és Y - sztochasztikus kapcsolatát illetőleg legteljesebb információt a két változó együttes eloszlásából nyerünk. Az alábbiakban tehát a kétdimenziós folytonos eloszlást vizsgáljuk. Megmutatjuk, hogy a folytonos eloszlású X és Y valószínűségi változók $h(x, y)$ együttes sűrűségfüggvénye segítségével néhány mennyiséget kell kiszámítani ahhoz, hogy a két változó kapcsolatát elméletileg jellemezzhessük.

A korábbiakban láttuk (lásd:2.1.2 pont), hogy az X valószínűségi változó $f(x)$ sűrűségfüggvényét, ill az Y változó $g(y)$ sűrűségfüggvényét a $h(x,y)$ együttes sűrűségfüggvényből megkaphatjuk a következő módon:

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

Ugyancsak kiszámíthatók $h(x,y)$ segítségével az alábbi momentumok:

$$\begin{aligned} M(X) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x h(x,y) dx dy = \int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} h(x,y) dy \right) dx = \\ &= \int_{-\infty}^{\infty} x f(x) dx = \mu_1 , \end{aligned}$$

$$\begin{aligned} M(Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y h(x,y) dx dy = \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} h(x,y) dx \right) dy = \\ &= \int_{-\infty}^{\infty} y g(y) dy = \mu_2 , \end{aligned}$$

$$\begin{aligned} M(X^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 h(x,y) dx dy = \int_{-\infty}^{\infty} x^2 \left(\int_{-\infty}^{\infty} h(x,y) dy \right) dx = \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx = M(X^2) = \sigma_1^2 + \mu_1^2 , \end{aligned}$$

$$\begin{aligned} M(Y^2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y^2 h(x,y) dx dy = \int_{-\infty}^{\infty} y^2 \left(\int_{-\infty}^{\infty} h(x,y) dx \right) dy = \\ &= \int_{-\infty}^{\infty} y^2 g(y) dy = M(Y^2) = \sigma_2^2 + \mu_2^2 , \end{aligned}$$

$$M(X \cdot Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy h(x,y) dx dy = \mu_{12} .$$

Definíció.

Az X és Y valószínűségi változók közötti korrelációs együttható

$$\begin{aligned}\rho &= \frac{M[X - M(X)] [Y - M(Y)]}{D(X) D(Y)} = \frac{M(X \cdot Y) - M(X)M(Y)}{D(X) D(Y)} = \\ &= \frac{\mu_{12} - \mu_1 \mu_2}{\sigma_1 \sigma_2}.\end{aligned}$$

A számlálóban lévő mennyiséget az X és Y közötti kovarianciának nevezzük, a nevezőben pedig a szórások szorzata van.

A korrelációs együtthatóra vonatkozólag kimutatható, hogy ha $\rho = \pm 1$, akkor $Y = aX + b$, vagyis lineáris kapcsolat áll fenn a két változó között, azaz X értéke egyértelműen meghatározza y értékét; ez esetben ha $a > 0$, akkor $r = .1$, ha $a < 0$, akkor $r = -1$.

Definíció. Amennyiben $\rho = 0$, akkor az X és Y valószínűségi változókat korrelátlanoknak nevezzük.

Ha két valószínűségi változó független, akkor korrelátlatlan is, hiszen független valószínűségi változók szorzatának várható értéke a várható értékek szorzatával egyenlő, tehát függetlenség esetén a korrelációs együttható számlálója zérus.

Abból a tényből viszont, hogy $\rho = 0$, általában még nem következik a két változó függetlensége. Látni fogjuk, hogy a gyakorlati szempontból fontos kétdimenziós normális eloszlás esetén a korrelátlanság maga után vonja a két változó függetlenségét. Más eloszlások esetében előfordulhat, hogy $\rho = 0$, a két változó között mégis szoros kapcsolat (sőt az is lehet, hogy függvénykapcsolat) áll fenn. Ha értéke közel van 1-hez vagy (-1)-hez, akkor mindenesetre azt mondhatjuk, hogy a vizsgált valószínűségi változók között szoros lineáris sztochasztikus kapcsolat van.

Ahhoz, hogy az X és Y valószínűségi változók kapcsolatot megbízhatóan jellemezhessük, és a kapcsolat függvényalakjáról is képet kapjunk, a $h(x,y)$ együttes sűrűségfüggvény segítségével további mennyiséget is ki kell számítanunk.

Definíció.

A $h(x,y)$ kétdimenziós sűrűségfüggvény ismeretében az Y valószínűségi változó feltételes várható értéke az $X = x$ feltétel mellett:

$$M(Y|X = x) = \frac{\int_{-\infty}^{\infty} y h(x,y) dy}{\int_{-\infty}^{\infty} h(x,y) dy} = \mu_2(x).$$

Az $M(Y|X = x)$ feltételes várható értéket mint x függvényét az Y változó X -re vonatkozó regressziós görbéjének nevezzük. Az $M(Y|X = x) = \mu_2(x)$ regressziós görbe az X és Y valószínűségi változók közötti kapcsolat függvényalakját, a két változó közötti közös tendenciát fejezi ki; mivel azt tükrözi vissza, hogyan függ az Y változó várható értéke X megfigyelt értékétől. Hasonlóan értelmezhető az $M(X|Y = y) = \mu_1(y)$ feltételes várható érték mint y függvénye. A gyakorlati probléma dönti el, hogy melyiket vizsgáljuk a két függvény közül.

A regressziós görbének fontos optimumtulajdonsága van, amelyet a következő tétel fejez ki:

Tétel.

Ha $\varphi = \varphi(X)$ tetszőleges függvénye az X változónak, akkor tetszőleges $X = x$ értékre:

$$M[Y - \mu_2(x)]^2 \leq M[Y - \varphi(x)]^2,$$

azaz minden görbe közül a regressziós görbe az, amelytől Y értékeinek négyzetes átlageltérése a legkisebb.

Bizonyítás.

A szórásnégyzetre vonatkozó Steiner-formula szerint, ha $M(Y) = \mu_2$, akkor tetszőleges a számra

$$M(Y - \mu_2)^2 \leq M(Y - a)^2,$$

és egyenlőség csakis az $a = \mu_2$ esetben áll fenn.

Legyen $\varphi = \varphi(X)$ tetszőleges függvénye X -nek, akkor rögzített $X = x$ értékre $\mu_2(x) = M(Y|X = x)$ az Y feltételes várható értéke, $\varphi(x) = a$ valamely szám, tehát minden x -re:

$$M[Y - \mu_2(x)]^2 \leq M[Y - \varphi(x)]^2.$$

Az $M(Y|X = x)$ feltételes várható érték valószínűségi változónak tekinthető, hiszen értéke attól függ, hogy X melyik lehetséges x értékét vette fel, amennyiben Y nem független X -től. Ez utóbbi esetben ugyanis $M(Y|X = x) = M(Y) = \text{const.}$ Ilyenformán, ha az $M(Y|X = x) = \mu_2(x)$ függvény nem az x tengellyel párhuzamos egyenes, akkor máris az X és Y változók közötti függésre következtethetünk. E függés szorosságát a következő módon is mérhetjük.

Számítsuk ki az $M(Y|X = x)$ feltételes várható értékének mint valószínűségi változónak a várható értékét és a szórásnégyzetét! Minthogy

$$M(Y|X = x) = \frac{\int_{-\infty}^{\infty} y h(x, y) dy}{\int_{-\infty}^{\infty} h(x, y) dy} = \int_{-\infty}^{\infty} \frac{h(x, y)}{f(x)} dy ,$$

ezért

$$\begin{aligned} M M(Y|X = x) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \frac{h(x, y)}{f(x)} dy \right) f(x) dx = \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} h(x, y) dx \right) dy = \int_{-\infty}^{\infty} y g(y) dy = M(Y) = \mu_2 . \end{aligned}$$

Ez utóbbi összefüggés alapján a következő tételt mondhatjuk ki.

Tétel.

Az Y változó $M(Y|X = x)$ feltételes várható értékének várható értéke az Y változó (feltétel nélküli) várható értékével egyenlő. Az $M(Y|X = x)$ valószínűségi változó szórásnégyzete:

$$D^2[M(Y|X)] = \int_{-\infty}^{\infty} [\mu_2(x) - \mu_2]^2 f(x) dx .$$

Az Y valószínűségi változó feltételes várható értékének szórásnégyzete és Y szórásnégyzete segítségével az X és Y változók kapcsolata szorosságának mérésére igen jó mérőszámot konstruált K. Pearson.

Definíció.

Ha ez Y változó X -re vonatkozó feltételes várható értékének szórásnégyzetét elosztjuk Y szórásnégyzetével, vagyis képezzük a

$$\theta_{yx}^2 = \frac{D^2[M(Y|X)]}{D^2(Y)} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} [\mu_2(x) - \mu_2]^2 f(x) dx$$

menyiséget akkor olyan számot kapunk, amelyre teljesül a

$$0 \leq \theta_{xy}^2 \leq 1$$

egyenőtlenség. Ennek a mennyiségnek a pozitív négyzetgyökét a $\theta_{Y,X}$ számot az X és Y változók közötti korrelációhányadosnak nevezzük. A korrelációhányados fontos tulajdonságát fejezi ki az alábbi tétel, amelyet bizonyítás nélkül közlünk.

Tétel.

A $\theta_{Y,X}$ korrelációhányados értéke akkor és csak akkor 0, ha X és Y függetlenek, továbbá $\theta_{Y,X}$ értéke akkor és csak akkor 1, ha X és Y között függvénykapcsolat van.

A korrelációhányados tehát két valószínűségi változó kapcsolata szorosságának igen jó mérőszáma, előnyösebb tulajdonságai vannak, mint a korrelációs együtthatónak, amelynek bevezetése ugyancsak K. Pearson nevéhez fűződik.

3.7.3 A regressziós egyenes meghatározása a legkisebb négyzetek módszerével

A legkisebb négyzetek módszere a valószínűségszámításnak és a matematikai statisztikának jelentős technikai segédeszköze. Használatának jogosultságát az eredmények matematikai statisztikai vizsgálata támasztja alá. A módszer alkalmazását először két valószínűségi változó közötti kapcsolatot kifejező közelítő függvény keresésére vonatkozólag mutatjuk be. Ilyen közelítő függvény keresését indokolhatja, hogy az egyik változó értékei könnyebben mérhetők, mint a másiké, vagy pl. egyik változó aktuális értéke korábban megfigyelhető, mint a másik változóé.

Tekintsük a folytonos eloszlású X és Y valószínűségi változókat, amelyek összetartozó mért értékeit tartalmazó

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

kétdimenziós mintával rendelkezünk. A mintát síkbeli pontthalmazként ábrázolva, a pontok elhelyezkedése sokszor arra utal, hogy megelégedhetünk a két változó közötti közös tendenciának valamely $Y = aX + b$ egyenessel való kifejezésére. Ilyen pontelhelyezkedést láthattunk a 50.a.) ábrán. Amikor az Y változó értékét a X változó $aX + b$ lineáris függvényével közelítünk, akkor feladatunk az a és b állandók meghatározása. E közelítés esetében tényleges értéke és a lineáris közelítés közötti eltését az

$$Y - (aX + b)$$

különbség adja.

Definíció.

Lineáris közelítés esetén a legkisebb négyzetek módszere szerint a és b értéket úgy kell meghatároznunk, hogy az eltérés négyzetének várható értéke a lehető legkisebb legyen, azaz fennálljon:

$$M[Y - (aX + b)]^2 = \text{minimum.}$$

Határozzuk meg mely a, b értékre áll fenn ez a minimum. A fenti összefüggésben szereplő kifejezés a és b függvénye

$$f(a,b) = M[Y - (aX + b)]^2.$$

Az $f(a,b)$ kétváltozós függvény szélsőértékét csak ott veheti fel, ahol mindkét változó szerint parciális deriváltja zérussal egyenlő:

$$\frac{\partial f(a,b)}{\partial a} = 0, \quad \frac{\partial f(a,b)}{\partial b} = 0.$$

Ha az X és Y változók $h(x,y)$ együttes sűrűségfüggvényét ismerjük, akkor az $f(a,b)$ függvény a következő:

$$f(a,b) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - ax - b)^2 h(x,y) dx dy.$$

Innen

$$\begin{aligned} \frac{\partial f(a,b)}{\partial a} &= -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(y - ax - b) h(x,y) dx dy = \\ &= -2 \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyh(x,y) dx dy - a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 h(x,y) dx dy - \right. \\ &\quad \left. - b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xh(x,y) dx dy \right] = 0, \end{aligned}$$

$$\begin{aligned}
 \frac{\partial f(a,b)}{\partial b} &= -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - ax - b) h(x,y) dx dy = \\
 &= -2 \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y h(x,y) dx dy - a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x h(x,y) dx dy - \right. \\
 &\quad \left. - b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) dx dy \right] = 0.
 \end{aligned}$$

Figyelembe véve a 260. oldalon szereplő formulákat, továbbá (-2)vel osztva, a következő lineáris egyenletrendszert kapjuk:

$$\mu_{12} - a(\sigma_1^2 + \mu^2) - b\mu_1 = 0.$$

$$\mu_2 - a\mu_1 - b = 0$$

Innen $b = \mu_2 - a\mu_1$, tehát

$$\mu_{12} - a(\sigma_1^2 + \mu_1^2) - \mu_1 \mu_2 + a\mu_1^2 = 0,$$

azaz

$$a = \frac{\mu_{12} - \mu_1 \mu_2}{\sigma_1^2} = \frac{\sigma_2}{\sigma_1} \frac{\mu_{12} - \mu_1 \mu_2}{\sigma_1 \sigma_2}.$$

Ha most figyelembe vesszük a ρ korrelációs együtthatóra vonatkozó összefüggést, akkor

$$a = \rho \frac{\sigma_2}{\sigma_1}, \quad b = \mu_2 - \rho \frac{\sigma_2}{\sigma_1} \mu_1.$$

Könnyű belátni, hogy a kapott (a,b) értékpárra $f(a,b)$ valóban minimumát veszi fel. A keresett egyenes egyenlete:

$$\begin{aligned}
 y &= \rho \frac{\sigma_2}{\sigma_1} x + \mu_2 - \rho \frac{\sigma_2}{\sigma_1} \mu_1 = \\
 &= \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1) + \mu_2.
 \end{aligned}$$

A regressziós egyenes iránytangensét, az $a = \rho \frac{\sigma_2}{\sigma_1}$ mennyiséget az Y változó X-re vonatkozó regressziós együtthatójának nevezzük. A közelítő egyenes egyenlete a következő alakban írható:

$$\frac{y - \mu_2}{\sigma_2} = \rho \frac{x - \mu_1}{\sigma_1}.$$

Ebből az alakból látható, hogy $x = \mu_1$ esetén $y = \mu_2$ adódik, vagyis a regressziós egyenes átmegy a (μ_1, μ_2) ponton, az eloszlás súlypontján. Ha $\rho = 0$, vagyis X és Y korrelátlanak, akkor $y = \mu_2$ az x tengellyel párhuzamos egyenes.

A fenti egyenlet azt is mutatja, hogy ha az X és Y valószínűségi változók

$$x^* = \frac{x - \mu_1}{\sigma_1}, \quad y^* = \frac{y - \mu_2}{\sigma_2}$$

standardizált változóit képezzük, akkor y^* -nak az x^* -ra vonatkozó regressziós egyenese az origén átmenő egyenes amelynek iránytangense éppen a korrelációs együttható.

Hasonló módon határozható meg a legkisebb négyzetek módszerével az X valószínűségi változó Y-ra vonatkozó regressziós egyenese, amelynek egyenlete:

$$\begin{aligned} x &= \rho \frac{\sigma_1}{\sigma_2} y + \mu_1 - \mu_2 \rho \frac{\sigma_1}{\sigma_2} = \\ &= \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2) + \mu_1 \end{aligned}$$

Az X változó Y-ra vonatkozó regressziós együtthatója tehát:

$$a' = \rho \frac{\sigma_1}{\sigma_2}$$

A két regressziós együttható szorzata:

$$aa' = \rho \frac{\sigma_2}{\sigma_1} \cdot \rho \frac{\sigma_1}{\sigma_2} = \rho^2.$$

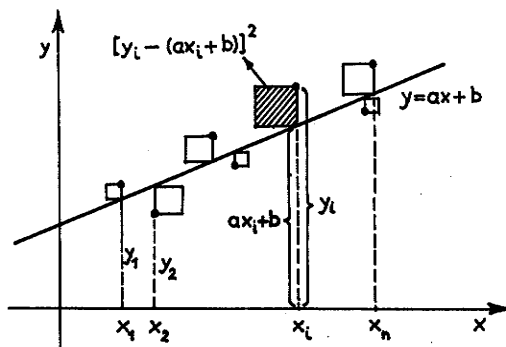
Láthatjuk, hogy a legkisebb négyzetek módszere az X és Y valószínűségi változók közötti lineáris kapcsolatának közelítésére két egyenest szolgáltat.

Amennyiben X értéke pontosabban mérhető (vagy beállítható) mint Y értéke, akkor X -et tekintjük független változónak és Y -nak X -re vonatkozó regressziós egyenesét határozzuk meg.

3.7.4 A regressziós egyenes paramétereinek becslése.

A korrelációs együtttható becslése

A gyakorlatban az $(X; Y)$ valószínűségi változó pár együttes eloszlását rendszerint nem ismerjük, így a regressziós egyeneseket az előző pontban ismertetett módszerrel nem tudjuk meghatározni. Ilyenkor a $(X; Y)$ párra vonatkozó $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$ kétdimenziós mintából indulunk ki, amelyet rendszerint egy (x, y) derékszögű koordináta-rendszerben ábrázolunk. Ha a pontelhelyezkedés lineáris közelítés célszerűségét sugallja, akkor keressük azt az $y = ax + b$ egyenest, amely a legkisebb négyzetek elve értelmében legjobban illeszkedik a pontfelhőhöz (lásd: 59. ábra).



59. ábra

Az együttthatókat a

$$\varphi(a, b) = \sum_{i=1}^n [y_i - (ax_i + b)]^2 = \text{minimum}$$

feltétel alapján becsüljük. A $\varphi(a, b)$ kétváltozós függvény szélsőértékhelyének megkeresése céljából az előző pontban ismertetett eljárással teljesen analóg módon eljárva a parciális deriváltakat kell zérussal egyenlővé tenni, majd a

$$\frac{\partial \varphi(a, b)}{\partial a} = 0, \quad \frac{\partial \varphi(a, b)}{\partial b} = 0$$

egyenleteket kell megoldani.

$$\frac{\partial \varphi(a,b)}{\partial a} = -2 \sum_{i=1}^n (y_i - ax_i - b) x_i = 0 ,$$

$$\frac{\partial \varphi(a,b)}{\partial b} = -2 \sum_{i=1}^n (y_i - ax_i - b) = 0 .$$

Ezek alapján a következő lineáris egyenletrendszert nyerjük:

$$a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i ,$$

$$a \sum_{i=1}^n x_i + bn = \sum_{i=1}^n y_i .$$

a megoldásrendszer:

$$\hat{a} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

$$\hat{b} = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

Ezzel gyakorlatilag megoldottuk a feladatunkat. Figyelembe kell venni, hogy az a , ill. b mennyiségek valószínűségi változók mivel az (x_i, y_i) mintaelemek függvényei $(i=1, 2, \dots, n)$.

Igy a , ill b az

$$a = \rho \frac{\sigma_2}{\sigma_1} , \text{ ill. } b = \mu_2 - \frac{\sigma_2}{\sigma_1} \mu_1 \text{ elméleti para-}$$

méterek becslései. Ha a képletében a számlálót és nevezőt n^2 -tel osztjuk, akkor

$$\hat{a} = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \bar{y}}{\frac{\sum_{i=1}^n x_i^2}{n} - (\bar{x})^2} = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \bar{y}}{s_1^2} =$$

$$= \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \bar{y}}{s_1 s_2} \cdot \frac{s_2}{s_1}$$

ahol s_1^2 az X változó empirikus szórásnégyzete, s_2^2 pedig Y empirikus szórásnégyzete. Bevezetve az

$$r = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \bar{y}}{s_1 s_2}$$

jelölést kapjuk:

$$\hat{a} = r \frac{s_2}{s_1}.$$

Összehasonlítva az $a = \rho \frac{\sigma_2}{\sigma_1}$ elméleti regressziós együtt-
hatót $\hat{a} = r \frac{s_2}{s_1}$ becslésével, indokolt az r mennyiséget a
 ρ korrelációs együtttható becslésének tekinteni. Az X és Y
valószínűségi változók közötti

$$\rho = \frac{M(X \cdot Y) - M(X) M(Y)}{D(X) D(Y)}$$

korrelációs együttható becslése az

$$r = \frac{\frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x} \bar{y}}{s_1 s_2}$$

empirikus korrelációs együttható. Az r empirikus (tapasztalati) korrelációs együttható maga is valószínűségi változó, amelynek várható értéke és szórásnégyzetére vonatkozólag eléggé enyhe feltételek mellett a következő aszimptotikus formulák érvényesek:

$$M(r) = \rho + O\left(\frac{1}{n}\right); D^2(r) = \frac{(1 - \rho^2)^2}{n} + O\left(\frac{1}{n^{3/2}}\right),$$

ahol $O\left(\frac{1}{n}\right)$, ill. $O\left(\frac{1}{n^{3/2}}\right)$ olyan mennyiségeket jelölnek, amelyek $n \rightarrow \infty$ esetén legalább olyan gyorsan tartanak 0-hoz mint $\frac{1}{n}$, ill. $\frac{1}{n^{3/2}}$. Ez azt jelenti, hogy r a ρ -nak

asszimptotikus torzítatlan és konzisztens becslése. A bevezetett jelölésekkel némi számolással belátható, hogy a regressziós egyenes b elméleti paraméterének b becslése az alábbi alakban írható:

$$\hat{b} = \bar{y} - r \frac{s_2}{s_1} \bar{x}$$

vagyis a keresett egyenes egyenlete:

$$y = r \frac{s_2}{s_1} x + \bar{y} - r \frac{s_2}{s_1} \bar{x} = r \frac{s_2}{s_1} (x - \bar{x}) + \bar{y}.$$

Az összefüggés átrendezhető

$$\frac{y - \bar{y}}{s_2} = r \frac{x - \bar{x}}{s_1}$$

alakra, amely összhangban van az elméleti

$$\frac{y - \mu_2}{\sigma_2} = \rho \frac{x - \mu_1}{\sigma_1} \text{ formulával, csak itt a megfele}$$

lő empirikus mennyiségek (becslések) szerepelnek.

3.7.5 Regressziós parabolák

Tekintsük ismét a folytonos eloszlású (X, Y) valószínűségi változó párt, amelyről feltételezzük, hogy együttes $h(x, y)$ sűrűségfüggvényét ismerjük.

Ha az (X, Y) párra vonatkozó statisztikai minta síkbeli pontthalmazként ábrázolva szemlátomást nem teszi lehetővé az Y változónak az X lineáris függvényével való közelítését, akkor (lehetőleg alacsony fokszámú) polinom segítségével közelítünk.

Tegyük fel, hogy Y értékeit az X változó k -adfokú polinomja segítségével kívánjuk közelíteni:

$$Y \approx a_0 + a_1 X + a_2 X^2 + \dots + a_k X^k$$

akkor a legkisebb négyzetek elvét alkalmazva az $a_0, a_1, \dots, a_k$ együtthatókat úgy kell meghatároznunk, hogy

$$M[Y - (a_0 + a_1 X + \dots + a_k X^k)]^2 = \text{minimum}$$

legyen. Ez a várható értéke az $a_0, a_1, \dots, a_k$ ismeretlenek függvénye:

$$\begin{aligned} \varphi(a_0, a_1, \dots, a_k) = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - a_0 - a_1 x - a_2 x^2 - \dots \\ & \dots - a_k x^k)^2 h(x, y) dx dy. \end{aligned}$$

Mivel a φ függvénynek szélsőértéke csak ott lehet, ahol az a_i változók szerinti parciális deriváltak zérussal egyenlők, a keresett együtthatókra az alábbi egyenletrendszer adódik:

$$\begin{aligned} - \frac{1}{2} \frac{\partial \varphi}{\partial a_i} = & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - a_0 - a_1 x - a_2 x^2 - \dots - \\ & - a_k x^k) x^i h(x, y) dx dy = 0 \quad (i=0, 1, 2, \dots, k). \end{aligned}$$

Innen:

$$\begin{aligned} a_0 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i h(x, y) dx dy + a_1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^{i+1} h(x, y) dx dy + \dots \\ \dots + a_k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^{i+k} h(x, y) dx dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y x^i h(x, y) dx dy \\ (i=0, 1, 2, \dots, k). \end{aligned}$$

Bevezetve a momentumokra szokásos

$$\mu_{0,0} = 1, \quad \mu_{i,0} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i h(x,y) dx dy,$$

$$\mu_{i,1} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i y h(x,y) dx dy$$

jelölést, egyenletrendszerünk az alábbi alakú lesz:

$$\mu_{i,0} a_0 + \mu_{i+1,0} a_1 + \mu_{i+2,0} a_2 + \dots + \mu_{i+k,0} a_k = \mu_{i,1}$$

(i=0,1,2,...,k).

Mivel (k+1) lineáris egyenletünk és (k+1) ismeretlen együttható van, az egyenletrendszer megoldása megadja a keresett polinom együtthatóit. Be lehet bizonyítani, hogy a nyert megoldásrendszer ψ -nek minimumhelye.

A gyakorlatban az X és Y változók együttes eloszlását általában nem ismerjük. Ha az (X, Y) változópár összetartozó értékeire mérések útján az $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ statisztikai mintát nyerjük, akkor a legkisebb négyzetek módszerével a közelítő polinom együtthatóinak becslését a

$$\psi(a_0, a_1, \dots, a_k) = \sum_{i=1}^k [y_i - (a_0 + a_1 x_i - a_2 x_i^2 - \dots - a_k x_i^k)]^2 = \text{minimum}$$

feltétel mellett keressük. A parciális deriváltakra most a

$$\frac{\partial \psi}{\partial a_j} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i - \dots - a_k x_i^k) x_i^j = 0$$

$$(j=0,1,2,\dots,k)$$

feltételnek kell teljesülnie.

!Innen a megoldandó egyenletrendszer:

$$a_0 \sum_{i=1}^n x_i^j + a_1 \sum_{i=1}^n x_i^{j+1} + a_2 \sum_{i=1}^n x_i^{j+2} + \dots$$

$$\dots + a_k \sum_{i=1}^n x_i^{j+k} = \sum_{i=1}^n y_i x_i^j \quad (j=0,1,2,\dots,k).$$

Az egyenlet mindkét oldalát n -nel végigosztva és a tapasztalati momentumokra szokásos

$$m_{0,0} = 1, \quad m_{j,0} = \frac{\sum_{i=1}^n x_i^j}{n}, \quad m_{j,1} = \frac{\sum_{i=1}^n y_i x_i^j}{n}$$

jelölés bevezetésével megoldandó egyenletrendszer:

$$m_{j,0} a_0 + m_{j+1,0} a_1 + m_{j+2,0} a_2 + \dots + m_{j+k,0} a_k = m_{j,1}$$

($j=0,1,2,\dots,k$)

amely analóg az elméleti momentumokat tartalmazó egyenletrendszerrel, csak benne az elméleti momentumok helyett a tapasztalati (empirikus) momentumok szerepelnek. A fenti egyenletrendszert k kicsiny értékeire ($k=2,3$) a Cramér-számba segítségével megoldhatjuk. Ha magasabb fokszámú polinommal közelítünk, akkor általában számítógép alkalmazása szükséges. A kapott $a_0, a_1, \dots, a_k$ megoldásrendszer az elméleti egyenletrendszer megoldás rendszerének statisztikai becslése.

3.7.6 Korreláció és regresszió kétdimenziós normális eloszlás esetében

A műszaki gyakorlatban betöltött fontos szerepe miatt külön is vizsgáljuk a korrelációs és regressziós viszonyok szempontjából a kétdimenziós normális eloszlás esetét, amikor az X és Y valószínűségi változók együttes sűrűségfüggvénye a

$$h(x,y) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho \frac{(x-m_1)(x-m_2)}{\sigma_1 \sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2} \right]}$$

ahol $m_1 = M(X)$, $m_2 = M(Y)$, $\sigma_1^2 = D(X)$,

$\sigma_2^2 = D(Y)$, $\rho = \rho(X, Y)$ pedig a korrelációs együttható.

Ha a $\rho = 0$, azaz X és Y korrelálatlanok, akkor

$$h(x, y) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x-m_1)^2}{\sigma_1^2}} \cdot \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(y-m_2)^2}{\sigma_2^2}} = f(x)g(y)$$

Minthogy ez esetben a X és Y együttes sűrűségfüggvénye az egyes változók sűrűségfüggvényeinek szorzata, következik, hogy X és Y függetlenek, tehát igaz az alábbi tétel.

Tétel.

Kétdimenziós normális eloszlás esetében a két valószínűségi változó korrelálatlanságából a változók függetlensége következik. Számítsuk ki az Y változó X -re vonatkozó regressziós görbáját, az $M(Y|X = x) = \mu_2(x)$ függvényt! Mint ismeretes

$$M(Y|X = x) = \int_{-\infty}^{\infty} y \frac{h(x, y)}{f(x)} dy = \int_{-\infty}^{\infty} yg(y|x) dy ,$$

ahol a $\frac{h(x, y)}{f(x)} = g(y|x)$ függvény az Y valószínűségi változó feltételes sűrűségfüggvénye az $X = x$ feltétel mellett. Ha a $h(x, y)$ együttes sűrűségfüggvényét eloszjuk az X változó $f(x)$ sűrűségfüggvényével, akkor

$$g(y|x) = \frac{h(x, y)}{f(x)} = \frac{1}{\sigma_2 \sqrt{2\pi(1-\rho^2)}} e^{-\frac{1}{2\sigma_2^2(1-\rho^2)} \left[y - m_2 - \rho \frac{\sigma_2}{\sigma_1} (x - m_1) \right]^2}$$

Ebből a formából közvetlenül kiolvasható, hogy az valószínűségi változó feltételes várható értéke:

$$M(Y|X = x) = \mu_2(x) = \rho \frac{\sigma_2}{\sigma_1} (x - m_1) + m_2 .$$

Ez utóbbi formula egy egyenes egyenlete. Ennek alapján a következő tételt mondhatjuk ki.

Tétel.

Ha az X és Y valószínűségi változók együttes eloszlása kétdimenziós normális eloszlás, akkor az Y változó X -re vonatkozó regressziós görbéje az

$$y = \rho \frac{\sigma_2}{\sigma_1} (x - m_1) + m_2 \text{ egyenes.}$$

Látható, hogy $x = m_1$ esetén $y = m_2$, vagyis a regressziós egyenes átmegy az $(m_1; m_2)$ ponton, az eloszlás súlypontján. A regressziós egyenes egyenlete

$$\frac{y - m_2}{\sigma_2} = \rho \frac{x - m_1}{\sigma_1}$$

alakban is írható, ami mutatja, hogy az $X^* = \frac{x - m_1}{\sigma_1}$ és $Y^* = \frac{y - m_2}{\sigma_2}$ standardizált változók esetén Y^* -nak az X^*

változóra vonatkozó regressziója az origón átmenő egyenes, amelynek iránytangense a ρ korrelációs együttható.

Analóg módon számítható az X valószínűségi változó Y -ra vonatkozó regressziós görbéje. A számítás az

$$x = \rho \frac{\sigma_1}{\sigma_2} (y - m_2) + m_1$$

egyenest szolgáltatja.

A gyakorlatban általában mintából kell becsülnünk a regressziós egyenes paramétereit.

Mint hogy normális eloszlás esetében a várható érték legjobb becslése a mintaközép, a szórásnégyzetek torzítatlan becslései a korrigált empirikus szórásnégyzetek, ezért az alábbi becsléseket alkalmazzuk:

$$m_1 \approx \bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad m_2 \approx \bar{y} = \frac{\sum_{i=1}^n y_i}{n};$$

$$s_1^{*2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$S_2^{x2} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

ρ korrelációs együttható becslésére az

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

empirikus korrelációs együtthatót használjuk, amelynek kétváltozós normális eloszlás esetében a sűrűségfüggvénye is ismert és ne elemű minta esetén a következő:

$$f_n(z, \rho) = \frac{n-2}{\pi} (1-\rho^2)^{\frac{n-1}{2}} (1-z^2)^{\frac{n-1}{2}}$$

$$\int_0^1 \frac{x^{n-2}}{(1-\rho xz)^{n-1}} \frac{dx}{\sqrt{1-x^2}} \quad (-1 \leq z \leq 1),$$

ahol ρ az elméleti korrelációs együttható.

Az r statisztikából R. A. Fisher a következő statisztikát képezte:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

Megmutatható, hogy ha n nagy, akkor Z elég általános feltételek mellett normális eloszlású

$$M(Z) \approx \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}$$

várható értékkel és

$$D^2(Z) \approx \frac{1}{n-3}$$

szórásnégyzettel.

Ezen formulák segítségével kétdimenziós normális eloszlás esetében vizsgálhatjuk az X és Y valószínűségi változók függetlenségét. Ekkor ugyanis a

$$H_0 : P(X < x, Y < y) = P(X < x)P(Y < y)$$

hipotézis ekvivalens a $H'_0 : \rho = 0$ hipotézissel, mivel ez esetben a korrelátlansága maga után vonja a függetlenséget. H'_0 fennállása esetén pedig

$$P\left(-\frac{2}{\sqrt{n-3}} \leq Z < -\frac{2}{\sqrt{n-3}}\right) \approx 0,95, \text{ ha } n \text{ elég nagy } (n > 30).$$

3.7.7 Korrelációs és regressziós több változó esetén

A műszaki gyakorlatban sokszor tisztáznunk kell, hogy bizonyos mennyiség(ek) hogyan függ(nek) más mennyiségektől, ezek közül melyektől függ(nek) 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, így együttesük - egy n dimenziós vektor - ugyan-csak 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 fokig önkényes, mindenestre igyekszünk a vizsgált jelenség alakulásában lényeges szerepet játszó valamennyi komponenst figyelembe venni).

Jelöljük a szóban forgó valószínűségi változók együttesét az $(X_1, X_2, \dots, X_n)$ véletlen vektorral, s ennek n dimenziós sűrűségfüggvénye legyen $f(x_1, x_2, \dots, x_n)$. Vizsgálhatjuk, hogy az X_1 változó értékeit az $X_2, X_3, \dots, X_n$ változók mely függvényének értékei közelítik legjobban. A legkisebb négyzetek módszere most is a feltételes várható értékhez vezet, amely most $n-1$ változós függvény:

$$\begin{aligned} M(X_1 | X_2 = x_2, X_3 = x_3, \dots, X_n = x_n) &= g(x_2, x_3, \dots, x_n) = \\ &= \frac{\int_{-\infty}^{\infty} x_1 f(x_1, x_2, \dots, x_n) dx_1}{\int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1} \end{aligned}$$

A $g(x_2, \dots, x_n)$ függvényt X_1 -nek, az $(X_2, \dots, X_n)$ -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 becslést X_1 megfigyelendő értékre vonatkozólag. A regressziós felület viszonylag egyszerű n dimenziós normális eloszlás esetében, amint az alábbi tételből látható, amelyet bizonyítás nélkül közlünk.

Tétel.

Abban az esetben, ha az $(X_1, X_2, \dots, X_n)$ változók együttes eloszlása n -dimenziós normális eloszlás, akkor a regressziós felület sík, vagyis

$$x_1 = M(X_1 | X_2 = x_2, \dots, X_n = x_n) = a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n + a_1$$

Az X_1 változó helyett természetesen az X_1 változók bármelyikét kifejezhetjük a többi változó függvényeként, esetleg nem is mindegyik változót vesszük az $X_1, \dots, X_n$ változók közül.

Annak kifejezésre juttatása, hogy melyik változót emeljük ki, és mely többi változó függvényeként közelítjük, meglehetősen bonyolult jelölésmóddal lehetséges, amelyet nem kívánunk bevezetni, inkább megállapodunk, hogy mindig az aktuális kiválasztott változót jelöljük X_1 -gyel.

Hasonlóan a két valószínűségi változót közötti regressziós görbe egyenessel való közelítéséhez, több valószínűségi változó esetében is gyakran megelégszünk egy kiválasztott X_1 változónak lineáris függvénnyel való közelítésével akkor is, ha az együttes eloszlás nem n -dimenziós normális eloszlás. Meghatározzuk azt a regressziós "síkot" (valójában $n - 1$ dimenziós ún. hipersíkot), amelytől X_1 értékei a legkisebb négyzetek módszere alapján a legkevésbé térnek el. Egyszerűség kedvéért feltesszük, hogy

$$M(X_i) = 0 \quad (i=1, 2, \dots, n).$$

Ezt mindig elérhetjük azáltal, hogy mindegyik valószínűségi változóból levonjuk a várható értékét.

Keressük tehát azt a

$$X_1 = a_{12} X_2 + a_{13} X_3 + \dots + a_{1n} X_n$$

$(n - 1)$ változós lineáris függvényt, amelyre

$$M(X_1 - a_{12} X_2 - a_{13} X_3 - \dots - a_{1n} X_n)^2 = \text{minimum}.$$

Feltételezve, hogy a szóban forgó változók folytonos eloszlásúak, a fenti összefüggést integrál alakban is írhatjuk:

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (x_1 - a_{12}x_2 - a_{1n}x_n)^2 f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots \dots dx_n = \text{minimum.}$$

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

$$b_{ii} = M(x_i^2) = D^2(x_i) \quad (i=1, 2, \dots, n).$$

Differenciálással az ismeretlen a_{ij} együtthatókra a következő lineáris egyenletrendszert kapjuk:

$$b_{i2}a_{12} + b_{i3}a_{13} + \dots + b_{in}a_{1n} = b_{i1} \quad (i=2, 2, \dots, n).$$

Feltéve, hogy

$$\underline{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{12} & b_{22} & \dots & b_{2n} \\ \cdot & & \cdot & \\ \cdot & & & \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix}$$

kovarianciamátrix nem szinguláris (azaz determinánsa nem zérus), megoldásként - ha $B_{11} \neq 0$ - az

$$a_{1k} = \frac{B_{1k}}{B_{11}} \quad (k=2, \dots, n),$$

kifejezést nyerjük, ahol B_{1k} , ill. B_{11} a b_{1k} , ill. b_{11} elemekhez tartozó al-determinánsokat jelölik.

Az X_i és X_j valószínűségi változók korrelációs együtthatója

$$\rho_{ij} = \frac{M(X_i \cdot X_j)}{D(X_i)D(X_j)} = \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} \quad (i, j=1, 2, \dots, n).$$

Képezzük most az

$$\begin{aligned} Y_1 &= X_1 - a_{12} X_2 - a_{13} X_3 - \dots - a_{1n} X_n = \\ &= \frac{1}{B_{11}} \sum_{j=1}^n B_{1j} X_j \end{aligned}$$

ún. maradékot, amelyet tehát úgy nyerünk, hogy X_1 értékből levonjuk a lineáris közelítés értékét. Könnyen beláthatjuk, hogy Y_1 korrelálatlan az $X_2, X_3, \dots, X_n$ változókkal, viszont X_1 -gyel pozitívan korrelált. Ugyanis:

$$\begin{aligned} M(Y_1 X_i) &= M(Y_1 X_i) - \sum_{i=1}^n a_{1j} M(X_j X_i) = \\ &= \frac{1}{B_{11}} \sum_{j=1}^n B_{1j} b_{ij} = \begin{cases} \frac{|B|}{B_{11}}, & \text{ha } i = 1 \\ 0 & \text{ha } i \neq 1. \end{cases} \end{aligned}$$

(Itt $|B|$ a B kovarianciamátrix determinánsát jelöli). Mint-hogy az $X_1, X_2, \dots, X_n$ változók mindegyikének várható értéke zérus, a várható érték tulajdonságai alapján következik, hogy $M(Y_1) = 0$.

Megjegyezzük, hogy abból a tényből, hogy a B kovarianciamátrix szimmetrikus, következik, hogy $|B|$ és B_{11} nem negatív értékek, s abban az esetben, ha az $X_1, X_2, \dots, X_n$ változók között nem áll fenn lineáris függés, a fenti determinánsok értéke pozitív. Az Y_1 és X_1 változók közötti korrelációs együttható:

$$\rho(Y_1 X_1) = \frac{M(X_1 \cdot Y_1)}{D(X_1)D(Y_1)} = \sqrt{\frac{|B|}{b_{11} B_{11}}}$$

Ha az $(X_1, X_2, \dots, X_n)$ véletlen vektoreloszlása n dimenziós normális eloszlás, akkor Y_1 független az $X_2, \dots, X_n$ változótól, X_1 -gyel pedig pozitívan korrelált. Ebben az esetben azt mondhatjuk, hogy Y_1 -ben az $X_2, X_3, \dots, X_n$ változók hatását teljesen kiküszöböltük.

3.7.8 Parciális korreláció

Tekintsük ismét az $X_1, X_2, \dots, X_n$ valószínűségi változókat.

Definíció.

$$-1 \leq \rho_{ik} = \frac{M[X_1 - M(X_1)] [X_k - M(X_k)]}{D(X_1)D(X_k)} \leq 1$$

korrelációs együtthatót totális korrelációs együtthatónak nevezzük.

Ha ρ_{ik} abszolút értéke közel van 1-hez, akkor azt mondjuk, hogy X_1 és X_k között szoros korreláció áll fenn. Szoros korreláció mutatkozhat annak következtében is, hogy mind X_1 , mind X_k értékeinek alakulását a többi valószínűségi változó értékeitől való függés irányítja, egymással pedig nincsenek is okozati összefüggésben. Ilyen esetben, ha a többi változó hatásait kiküszöbölnénk, nem is találnánk közöttük szoros korrelációt. A többi változó hatásának kiküszöbölése céljából úgy járunk el, hogy tekintsük most az X_1 és X_2 változókat - mind X_1 -et, mind X_2 -t kifejezzük az $X_3, \dots, X_n$ változókra vonatkozó legjobb közelítő regressziós felületekkel:

$$X_1 \approx \sum_{j=3}^n a_{1j} X_j ; X_2 \approx \sum_{j=3}^n a_{2j} X_j$$

Ezután képezzük az

$$Y_1 = X_1 - \sum_{j=3}^n a_{1j} X_j \text{ és } Y_2 = X_2 - \sum_{j=3}^n a_{2j} X_j$$

maradékokat, amelyekről tudjuk, hogy az $X_3, \dots, X_n$ változókkal korrelátlanok, együttes normális eloszlás esetén pedig függetlenek is tőlük, viszont az Y_1 pozitívan korrelált X_1 -gyel, Y_2 pedig pozitívan korrelált X_2 -vel. Az Y_1 és Y_2 maradékok közötti korrelációs együtthatót az X_1 és X_2 változók közötti parciális együtthatónak nevezzük. Úgy tekintjük, hogy a parciális korrelációs együttható X_1 és X_2 kapcsolatát tükrözi, megtisztítva ezt a kapcsolatot az $X_3, \dots, X_n$ valószínűségi változók hatásától.

A parciális korrelációs együtthatóra bevezetjük a következő jelölést:

$$\rho_{3,4,\dots,n}^{(1,2)} = \frac{M(Y_1 \cdot Y_2)}{D(Y_1) D(Y_2)} = - \frac{B_{12}}{\sqrt{B_{11}B_{22}}}$$

ahol B_{ik} a $\underline{B}$ korrelációs mátrix b_{ik} eleméhez tartozó algebrai aldetermináns jelöli.

A korreláció- és regresszióelmélettel kapcsolatos mondanivalónk befejezéséül még egy megjegyzést teszünk. Fejtetéseink 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 sok esetben kielégítő a lineáris közelítés.

FÜGGELÉK

Táblázatok

Binomiális eloszlás

I. táblázat

10 n	10 k	p											
		0,01	0,02	0,04	0,05	0,08	0,1	0,2	0,3	0,4	0,5		
5	0	0,95099	0,90392	0,81537	0,77378	0,65908	0,59049	0,32768	0,16807	0,07776	0,03125	5	5
	1	0,04803	0,09224	0,16987	0,20363	0,28656	0,32805	0,40960	0,36015	0,25920	0,15625	4	4
	2	0,00097	0,00376	0,01416	0,02143	0,04984	0,07290	0,20480	0,30870	0,34560	0,31250	3	3
	3	0,00001	0,00008	0,00059	0,00113	0,00433	0,00810	0,05120	0,13230	0,23040	0,31250	2	2
	4			0,00001	0,00003	0,00019	0,00045	0,00640	0,02835	0,07680	0,15625	1	1
10	5						0,00001	0,00032	0,00243	0,01024	0,03125	0	0
	0	0,90438	0,81707	0,66483	0,59874	0,43439	0,34868	0,16737	0,02825	0,00605	0,00908	10	10
	1	0,09135	0,16675	0,27701	0,31512	0,37773	0,38742	0,26844	0,12106	0,04031	0,00977	9	9
	2	0,00415	0,01531	0,05194	0,07463	0,14781	0,19371	0,30199	0,23347	0,12093	0,04395	8	8
	3	0,00011	0,00083	0,00577	0,01048	0,03427	0,05740	0,20133	0,26683	0,21499	0,11719	7	7
	4		0,00003	0,00042	0,00096	0,00522	0,01116	0,08808	0,20012	0,25082	0,20508	6	6
	5			0,00002	0,00006	0,00054	0,00149	0,02642	0,10292	0,20066	0,24609	5	5
	6					0,00004	0,00014	0,00551	0,03676	0,11148	0,20508	4	4
	7						0,00001	0,00079	0,00900	0,04247	0,11719	3	3
	8							0,00007	0,00145	0,01062	0,04395	2	2
15	9								0,00014	0,00157	0,00977	1	1
	10								0,00001	0,00010	0,00098	0	0
	0	0,86006	0,73857	0,54209	0,46329	0,28630	0,20589	0,03518	0,00475	0,00047	0,00003	15	15
	1	0,13031	0,22609	0,33880	0,36576	0,37343	0,34315	0,13194	0,03052	0,00470	0,00046	14	14
	2	0,00921	0,03230	0,09882	0,13475	0,22731	0,26690	0,23090	0,09156	0,02194	0,00320	13	13
	3	0,00040	0,00286	0,01784	0,03073	0,08565	0,12851	0,25014	0,17004	0,06339	0,01389	12	12
	4	0,00001	0,00017	0,00223	0,00485	0,02234	0,04284	0,18760	0,21862	0,12678	0,04166	11	11
	5		0,00001	0,00020	0,00056	0,00427	0,01047	0,10318	0,20613	0,18594	0,09164	10	10
	6			0,00001	0,00005	0,00062	0,00194	0,04299	0,14724	0,20660	0,15274	9	9
	7					0,00007	0,00028	0,01382	0,08113	0,17708	0,19638	8	8
	8					0,00001	0,00003	0,00345	0,03477	0,11806	0,19638	7	7
	9							0,00067	0,01159	0,01621	0,15274	6	6
	10							0,00010	0,00298	0,02449	0,09164	5	5
	11							0,00001	0,00058	0,00742	0,04166	4	4
	12								0,00008	0,00165	0,01389	3	3
	13								0,00001	0,00025	0,00320	2	2
	14									0,00002	0,00046	1	1
	15										0,00003	0	0
20	0	0,81791	0,66761	0,44200	0,35849	0,18869	0,12158	0,01153	0,00080	0,00004	0,00000	20	20
	1	0,16523	0,27249	0,36834	0,37735	0,32816	0,27017	0,05765	0,00684	0,00049	0,00002	19	19
	2	0,01586	0,05283	0,14580	0,18868	0,27109	0,28518	0,13691	0,02785	0,00309	0,00018	18	18
	3	0,00096	0,00647	0,03645	0,05958	0,14144	0,19012	0,20536	0,07160	0,01235	0,00109	17	17
	4	0,00004	0,00056	0,00645	0,01333	0,05227	0,08978	0,21820	0,13042	0,03492	0,00462	16	16
	5		0,00004	0,00086	0,00224	0,01454	0,03192	0,17456	0,17886	0,07465	0,01479	15	15
	6			0,00009	0,00030	0,00316	0,00887	0,10910	0,19164	0,12441	0,03696	14	14
	7			0,00001	0,00003	0,00055	0,00197	0,05455	0,16426	0,16588	0,07393	13	13
	8					0,00008	0,00036	0,02216	0,11440	0,17971	0,12013	12	12
	9					0,00001	0,00005	0,00739	0,06537	0,15974	0,16018	11	11
	10						0,00001	0,00203	0,03082	0,11714	0,17620	10	10
	11							0,00046	0,01201	0,07099	0,16018	9	9
	12							0,00009	0,00386	0,03550	0,12013	8	8
	13							0,00001	0,00102	0,01456	0,07393	7	7
	14								0,00022	0,00485	0,03696	6	6
	15								0,00004	0,00129	0,01479	5	5
	16								0,00001	0,00027	0,00462	4	4
	17									0,00004	0,00109	3	3
	18										0,00018	2	2
	19										0,00002	1	1
	20											0	0
		0,99	0,98	0,96	0,95	0,92	0,9	0,8	0,7	0,6	0,5	k	n
p													

I. táblázat folytatása

n	k	P											
		0,01	0,02	0,04	0,05	0,08	0,1	0,2	0,3	0,4	0,5		
25	0	0,77782	0,60346	0,36040	0,27739	0,12436	0,07179	0,00378	0,00013			25	25
	1	0,19642	0,30789	0,37541	0,36499	0,27036	0,19942	0,02361	0,00144	0,00005		24	
	2	0,02381	0,07540	0,18771	0,23052	0,28211	0,26589	0,07084	0,00739	0,00038	0,00001	23	
	3	0,00184	0,01180	0,05996	0,09302	0,18807	0,22650	0,13577	0,02428	0,00194	0,00007	22	
	4	0,00010	0,00132	0,01374	0,02693	0,08995	0,13842	0,18668	0,05723	0,00710	0,00038	21	
	5	0,00000	0,00011	0,00240	0,00595	0,03285	0,06459	0,19602	0,10302	0,01989	0,00158	20	
	6		0,00001	0,00033	0,00104	0,00952	0,02392	0,16335	0,14717	0,04420	0,00528	19	
	7			0,00004	0,00015	0,00225	0,00722	0,11084	0,17119	0,07999	0,01433	18	
	8				0,00002	0,00044	0,00180	0,06235	0,16508	0,11998	0,03223	17	
	9					0,00007	0,00038	0,02944	0,13364	0,15109	0,06089	16	
	10					0,00001	0,00097	0,01178	0,09184	0,16116	0,09742	15	
	11						0,00001	0,00001	0,05355	0,14651	0,13284	14	
	12							0,00117	0,02678	0,11395	0,15498	13	
	13							0,00029	0,01148	0,07597	0,15498	12	
	14							0,00006	0,00422	0,04341	0,13284	11	
	15							0,00001	0,00132	0,02122	0,09742	10	
	16								0,00035	0,00884	0,06089	9	
	17								0,00008	0,00312	0,03223	8	
	18								0,00002	0,00092	0,01433	7	
	19								0,00021	0,00528		6	
	20									0,00005	0,00158	5	
	21									0,00001	0,00038	4	
	22										0,00007	3	
	23										0,00001	2	
30	0	0,73970	0,54548	0,29386	0,21464	0,08197	0,04239	0,00124	0,00002			30	30
	1	0,22415	0,33397	0,36732	0,33890	0,21382	0,14130	0,00928	0,00029			29	
	2	0,03283	0,09883	0,22192	0,25864	0,26961	0,22766	0,03366	0,00180	0,00004		28	
	3	0,00310	0,01182	0,08630	0,12705	0,21881	0,23609	0,07853	0,00720	0,00027		27	
	4	0,00021	0,00259	0,02427	0,04514	0,12843	0,17707	0,13252	0,02084	0,00120	0,00003	26	
	5	0,00001	0,00028	0,00526	0,01235	0,05807	0,10230	0,17228	0,04644	0,00415	0,00013	25	
	6		0,00002	0,00091	0,00271	0,02104	0,04736	0,17946	0,08293	0,01152	0,00055	24	
	7			0,00013	0,00049	0,00627	0,01804	0,15382	0,12185	0,02634	0,00190	23	
	8			0,00002	0,00007	0,00157	0,00576	0,11056	0,15014	0,05049	0,00545	22	
	9				0,00001	0,00033	0,00152	0,96756	0,15729	0,08228	0,01332	21	
	10					0,00006	0,00037	0,03547	0,14156	0,11519	0,02798	20	
	11					0,00001	0,00007	0,01612	0,11031	0,13962	0,05088	19	
	12						0,00001	0,00638	0,07485	0,14738	0,08055	18	
	13							0,00221	0,04442	0,13604	0,11154	17	
	14							0,00067	0,02312	0,11013	0,13544	16	
	15							0,00018	0,01057	0,07831	0,14446	15	
	16							0,00004	0,00425	0,04895	0,13544	14	
	17							0,00001	0,00150	0,02687	0,11154	13	
	18								0,00046	0,01294	0,08055	12	
	19								0,00013	0,00545	0,05088	11	
	20								0,00003	0,00200	0,02798	10	
	21								0,00001	0,00063	0,01332	9	
	22									0,00017	0,00545	8	
	23									0,00004	0,00190	7	
	24									0,00001	0,00055	6	
	25										0,00013	5	
	26										0,00003	4	
		0,99	0,98	0,96	0,95	0,92	0,9	0,8	0,7	0,6	0,5	k	n
P													

Poisson-eloszlás

II. táblázat

$k \backslash \lambda$	0,1	0,2	0,3	0,4	0,5
0	0,90484	0,81873	0,74082	0,67032	0,60653
1	0,09048	0,16375	0,22225	0,26813	0,30327
2	0,00452	0,01637	0,03334	0,05362	0,07581
3	0,00015	0,00109	0,00333	0,00715	0,01263
4		0,00005	0,00025	0,00071	0,00158
5			0,00001	0,00005	0,00016
6					0,00001
$k \backslash \lambda$	0,6	0,7	0,8	0,9	
0	0,54881	0,49659	0,44933	0,40657	
1	0,32929	0,34761	0,35946	0,36591	
2	0,09878	0,12166	0,14379	0,16466	
3	0,01976	0,02838	0,03834	0,04939	
4	0,00296	0,00496	0,00766	0,01111	
5	0,00035	0,00069	0,00123	0,00200	
6	0,00003	0,00008	0,00016	0,00030	
7			0,00001	0,00003	
$k \backslash \lambda$	1	2	3	4	5
0	0,36788	0,13534	0,04978	0,01831	0,00673
1	0,36788	0,27067	0,14936	0,07326	0,03369
2	0,18394	0,27067	0,22404	0,14653	0,08422
3	0,06131	0,18045	0,22404	0,19537	0,14037
4	0,01532	0,09022	0,16803	0,19537	0,17547
5	0,00306	0,03609	0,10082	0,15629	0,17547
6	0,00051	0,01203	0,05040	0,10420	0,14622
7	0,00007	0,00343	0,02160	0,05954	0,10444
8		0,00085	0,00810	0,02977	0,06527
9		0,00019	0,00270	0,01322	0,03626
10		0,00003	0,00081	0,00529	0,01813
11			0,00022	0,00192	0,00824
12			0,00005	0,00064	0,00343
13			0,00001	0,00019	0,00132
14				0,00005	0,00047
15				0,00001	0,00015
16					0,00004
17					0,00001

II. táblázat folytatása

$k \backslash \lambda$	6	7	8	9	10
0	0,00247	0,00091	0,00033	0,00012	0,00004
1	0,01487	0,00638	0,00268	0,00111	0,00045
2	0,04461	0,02234	0,01073	0,00499	0,00227
3	0,08923	0,05212	0,02862	0,01499	0,00756
4	0,13385	0,09122	0,05725	0,03373	0,01891
5	0,16062	0,12772	0,09160	0,06072	0,03783
6	0,16062	0,14900	0,12214	0,09109	0,06305
7	0,13768	0,14900	0,13959	0,11712	0,09007
8	0,10326	0,13038	0,13959	0,13176	0,11260
9	0,06883	0,10140	0,12408	0,13176	0,12511
10	0,04130	0,07098	0,09926	0,11858	0,15511
11	0,02252	0,04517	0,07219	0,09702	0,11374
12	0,01126	0,02635	0,04812	0,07276	0,09478
13	0,00519	0,01418	0,02961	0,05037	0,07290
14	0,00222	0,00709	0,01692	0,03238	0,05207
15	0,00089	0,00331	0,00902	0,01943	0,03471
16	0,00033	0,00144	0,00451	0,01093	0,02169
17	0,00011	0,00059	0,00121	0,00578	0,01276
18	0,00003	0,00023	0,00094	0,00289	0,00709
19	0,00001	0,00008	0,00039	0,00137	0,00373
20		0,00003	0,00015	0,00061	0,00183
21			0,00006	0,00026	0,00088
22			0,00002	0,00010	0,00040
23				0,00004	0,00017
24				0,00001	0,00007
25					0,00002
26					0,00001

A nem teljes gammafüggvény

III. táblázat

n	x=0,001	x=0,002	x=0,003	x=0,004
1	0,000999	0,001988	0,002995	0,003992
2	000001	000002	000005	000008
n	x=0,005	x=0,006	x=0,007	x=0,008
1	0,004987	0,005982	0,006976	0,007968
2	000013	000018	000024	000032
n	x=0,009	x=0,01	x=0,02	x=0,03
1	0,008960	0,009950	0,019801	0,029555
2	000040	000050	000197	000441
3			000002	000004
n	x=0,04	x=0,05	x=0,06	x=0,07
1	0,039211	0,048771	0,058236	0,067606
2	000779	001209	001730	002339
3	000010	000020	000034	000054
4				000001
n	x=0,08	x=0,09	x=0,10	x=0,11
1	0,0976884	0,086069	0,095163	0,104166
2	003034	003815	004679	005624
3	000080	000114	000155	000204
4	000002	000002	000003	000006
n	x=0,12	x=0,13	x=0,14	x=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	x=0,16	x=0,17	x=0,18	x=0,19
1	0,147856	0,156335	0,164730	0,173041
2	011513	012912	014381	015915
3	000606	000721	000850	000992
4	000024	000031	000038	000047
5	000001	000001	000001	000001

III. táblázat folytatása

n	x=0,20	x=0,22	x=0,24	x=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	x=0,28	x=0,30	x=0,40	x=0,50
1	0,244216	0,259182	0,320680	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
n	x=1,0	x=,15	x=2,0	x=2,5
1	0,63212	0,77687	0,86466	0,91792
2	26424	44218	59399	71270
3	08030	19115	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	x=3,0	x=3,5	x=4,0	x=4,5
1	0,95021	0,96980	0,98168	0,98889
2	80085	86411	90842	93890
3	57681	67915	76190	82642
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

III. táblázat folytatása

n	x=3,0	x=3,5	x=4,0	x=4,5
13	00002	00008	00027	00081
14		00001	00008	00025
15			00002	00007
16			00001	00002
17				00001
n	x=5,0	x=5,5	x=6,0	x=6,5
1	0,99326	0,99591	0,99752	0,99850
2	95957	97345	98265	98872
3	87535	91162	93804	95696
4	73497	79830	84880	88816
5	55951	64248	71494	77633
6	38404	47108	55433	63096
7	23782	31396	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

III. táblázat folytatása

n	x=7,0	x=7,5	x=8,0	x=8,5	x=9,0	x=9,5	x=10,0
1	0,99909	0,99945	0,99966	0,99980	0,99988	0,99993	0,99996
2	99271	99530	99698	99807	99877	99921	99950
3	97036	97975	98625	99072	99377	99584	99724
4	91823	94085	95762	96989	97877	98514	98966
5	82701	86794	90037	92564	94504	95974	97075
6	69929	75856	80876	85040	88431	91147	93291
7	55029	62184	68663	74382	79322	83505	86986
8	40129	47536	54704	61440	67610	73134	77978
9	27091	33803	40745	47689	54435	60818	66719
10	16950	22359	28338	34703	41259	47817	54207
11	09852	13776	18412	23664	29401	35467	41696
12	05335	07924	11192	15134	19699	24801	30322
13	02700	04267	06380	09092	12423	16338	20844
14	01281	02157	03418	05141	07385	10186	13554
15	00572	01026	01726	02743	04147	05999	08346
16	00241	00461	00823	01383	02204	03347	04874
17	00096	00196	00372	00661	01111	01773	02704
18	00036	00079	00160	00300	00533	00893	01428
19	00013	00031	00065	00130	00243	00428	00719
20	00005	00011	00025	00054	00106	00196	00345
21		00004	00010	00020	00044	00086	00159
22		00001	00003	00008	00018	00036	00070
23			00001	00003	00006	00015	00030
24				00001	00003	00006	00012
25						00002	00004
26						00001	00001

Normális eloszlás

IV. táblázat

x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$
0,00	0,5000	0,35	0,6368	0,70	0,7580	1,05	0,8531
01	5040	36	6406	71	7611	06	8554
02	5080	37	6443	72	7642	07	8577
03	5120	38	6480	73	7673	08	8599
04	5160	39	6517	74	7703	09	8621
05	5199	0,40	0,6554	75	7734	1,10	0,8643
06	5239	41	6591	76	7764	11	8665
07	5279	42	6628	77	7794	12	8686
08	5319	43	6664	78	7823	13	8708
09	5359	44	6700	79	7853	14	8729
0,10	0,5398	45	6736	0,80	7881	15	8749
11	5438	46	6772	81	7910	16	8770
12	5478	47	6808	82	7939	17	8790
13	5517	48	6844	83	7967	18	8810
14	5557	49	6879	84	7995	19	8830
15	5596	0,50	0,6915	85	8023	1,20	0,8849
16	5636	51	6950	86	8051	21	8869
17	5675	52	6985	87	8078	22	8888
18	5714	53	7019	88	8106	23	8907
19	5753	54	7054	89	8133	24	8925
0,20	0,5793	55	7088	0,90	0,8159	25	8944
21	5832	56	7123	91	8186	26	8962
22	5871	57	7157	92	8212	27	8980
23	5910	58	7190	93	8238	28	8997
24	5948	59	7224	94	8264	29	9015
25	5987	0,60	7257	95	8289	1,30	0,9032
26	6026	61	7291	96	8315	31	9049
27	6064	62	7324	97	8340	32	9066
28	6103	63	7357	98	8365	33	9082
29	6141	64	7389	99	8389	34	9099
0,30	0,6179	65	7422	1,00	0,8413	35	9115
31	6217	66	7454	01	8438	36	9131
32	6255	67	7486	02	8461	37	9147
33	6293	68	7517	03	8485	38	9162
34	6331	69	7549	04	8508	39	9177
1,40	0,9192	1,76	0,9608	2,24	0,9875	2,96	0,9985
41	9207	77	9616	26	9881	98	9986
42	9222	78	9625	28	9887	3,00	0,99865
43	9236	79	9633	30	9893	10	99903
44	9251	1,80	0,9641	32	9898	20	99931
45	9265	81	9649	34	9904	30	99951
46	9279	82	9656	36	9909	40	99966
47	9292	83	9664	38	9913	50	99976

IV. táblázat folytatása

x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$	x	$\phi(x)$
48	9306	84	9671	2,40	0,9918	60	99984
49	9319	85	9678	42	9922	70	99989
1,50	0,9332	86	9686	44	9927	80	99993
51	9345	87	9693	46	9931	90	99995
52	9357	88	9699	48	9934	4,00	0,999968
53	9370	89	9706	50	9938	50	99997
54	9382	1,90	0,9713	52	9941	5,00	99997
55	9394	91	9719	54	9945		
56	9406	92	9726	56	9948		
57	9418	93	9732	58	9951		
58	9429	94	9738	2,60	0,9953		
59	9441	95	9744	62	9956		
1,60	0,9452	96	9750	64	9959		
61	9463	97	9756	66	9961		
62	9474	98	9761	68	9963		
63	9484	99	9767	70	9965		
64	9495	2,00	0,9772	72	9967		
65	9505	02	9783	74	9969		
66	9515	04	9793	76	9971		
67	9525	06	9803	78	9973		
68	9535	08	9812	2,80	0,9974		
69	9545	10	9821	82	9976		
1,70	0,9554	12	9830	84	9977		
71	9564	14	9838	86	9979		
72	9572	16	9846	88	9980		
73	9582	18	9854	90	9981		
74	9591	2,20	0,9861	92	9982		
75	9599	22	9868	94	9984		

χ^2 -eloszlás

V. táblázat

n \	0,99	0,98	0,95	0,90	0,80	0,70	0,50
1	0,000	0,000	0,003	0,016	0,064	0,148	0,455
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366
4	0,297	0,429	0,711	1,964	1,649	2,195	3,357
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340
13	4,107	4,765	5,892	7,042	8,634	9,926	12,340
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338
17	6,408	7,255	8,672	10,085	12,002	13,531	16,338
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337
26	12,198	13,409	15,379	17,292	19,820	21,792	25,336
27	12,879	14,125	16,151	18,114	20,703	22,719	26,336
28	13,565	14,847	16,928	18,939	21,588	23,647	27,336
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336
n \	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	3,665	4,642	6,251	7,815	9,837	11,345	16,268
4	4,878	5,989	7,779	9,488	11,668	13,277	18,465
5	6,064	7,289	9,236	11,070	13,388	15,086	20,517

V. táblázat folytatása

n \ P	0,30	0,20	0,10	0,05	0,02	0,01	0,001
6	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	31,391	34,027	37,916	41,337	45,419	48,278	56,793
29	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	33,530	36,250	40,256	43,773	47,962	50,892	59,703

Student- vagy t-eloszlás (t-próbához)

VI. táblázat

$\frac{P}{n}$	0,90	0,80	0,70	0,60	0,50	0,40	0,30
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076
15	0,128	0,258	0,393	0,536	0,691	0,866	1,974
16	0,128	0,258	0,392	0,535	0,690	0,865	1,971
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058
27	0,127	0,256	0,389	0,531	0,684	0,856	1,057
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041
	0,126	0,253	0,385	0,524	0,674	0,842	1,036

VI. táblázat folytatása

$\begin{matrix} p \\ n \end{matrix}$	0,20	0,10	0,05	0,02	0,01	0,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,553	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,859
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,578
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,291	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

**Az F-próba kritikus értékei 99%-os egyoldali, ill.
98%-os kétoldali szintre**

VII. táblázat

A névező szabadsági foka f_2	A számoló szabadságfoka f_1																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	Az első sorban ($f_2=1$) számokat 10-zel meg kell szorozni.																	
1	405	500	540	563	576	586	593	598	602	606	608	611	613	614	616	617	618	619
2	98,5	99,0	99,2	99,2	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4	99,4
3	34,1	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1	27,0	26,9	26,9	26,8	26,8	26,8
4	21,2	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,5	14,4	14,4	14,3	14,2	14,2	14,2	14,1	14,1
5	16,3	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	9,96	9,89	9,82	9,77	9,72	9,68	9,64	9,61
6	13,7	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,52	7,48	7,45
7	12,2	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,36	6,31	6,27	6,24	6,21
8	11,3	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,61	5,56	5,52	5,48	5,44	5,41
9	10,6	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,00	4,96	4,92	4,89	4,86
10	10,0	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,65	4,60	4,56	4,52	4,49	4,46
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,21	4,18	4,15
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,97	3,94	3,91
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,91	3,86	3,82	3,78	3,75	3,72
14	8,86	6,51	5,56	5,04	4,70	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,62	3,59	3,56
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,49	3,45	3,42
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,62	3,55	3,50	3,45	3,41	3,37	3,34	3,31
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,27	3,24	3,21
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,43	3,37	3,32	3,27	3,23	3,19	3,16	3,13
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,12	3,08	3,05
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,18	3,13	3,09	3,05	3,02	2,99
21	8,04	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,99	2,96	2,93
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,94	2,91	2,88
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,89	2,86	2,83
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,85	2,82	2,79
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22	3,13	3,06	2,99	2,94	2,89	2,85	2,81	2,78	2,75
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,90	2,86	2,82	2,78	2,74	2,72
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,99	2,93	2,87	2,82	2,78	2,75	2,71	2,68
28	7,64	5,45	4,57	4,07	3,75	3,54	3,36	3,23	3,12	3,03	2,96	2,90	2,84	2,79	2,75	2,72	2,68	2,65
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,93	2,87	2,81	2,77	2,73	2,69	2,66	2,63
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,79	2,74	2,70	2,66	2,63	2,60
32	7,50	5,34	4,46	3,97	3,65	3,43	3,26	3,13	3,02	2,93	2,86	2,80	2,74	2,70	2,66	2,62	2,58	2,55
34	7,44	5,29	4,42	3,93	3,61	3,39	3,22	3,09	2,98	2,89	2,82	2,76	2,70	2,66	2,62	2,58	2,55	2,51
36	7,40	5,25	4,38	3,89	3,57	3,35	3,18	3,05	2,95	2,86	2,79	2,72	2,67	2,62	2,58	2,54	2,51	2,48
38	7,35	5,21	4,34	3,86	3,54	3,32	3,15	3,02	2,92	2,83	2,75	2,69	2,64	2,59	2,55	2,51	2,48	2,45
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,48	2,45	2,42
42	7,28	5,15	4,29	3,80	3,49	3,27	3,10	2,97	2,86	2,78	2,70	2,64	2,59	2,54	2,50	2,46	2,43	2,40
44	7,25	5,12	4,26	3,78	3,47	3,24	3,08	2,95	2,84	2,75	2,68	2,62	2,56	2,52	2,47	2,44	2,40	2,37
46	7,22	5,10	4,24	3,76	3,44	3,22	3,06	2,93	2,82	2,73	2,66	2,60	2,54	2,50	2,45	2,42	2,38	2,35
48	7,19	5,08	4,22	3,74	3,43	3,20	3,04	2,91	2,80	2,72	2,64	2,58	2,53	2,48	2,44	2,40	2,37	2,33
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,79	2,70	2,63	2,56	2,51	2,46	2,42	2,38	2,35	2,32
55	7,12	5,01	4,16	3,68	3,37	3,15	2,98	2,85	2,75	2,66	2,59	2,53	2,47	2,42	2,38	2,34	2,31	2,28
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,44	2,39	2,35	2,31	2,28	2,25
65	7,04	4,95	4,10	3,62	3,31	3,09	2,93	2,80	2,69	2,61	2,53	2,47	2,42	2,37	2,33	2,29	2,26	2,23
70	7,01	4,92	4,08	3,60	3,29	3,07	2,91	2,78	2,67	2,59	2,51	2,45	2,40	2,35	2,31	2,27	2,23	2,20
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55	2,48	2,42	2,36	2,31	2,27	2,23	2,20	2,17
90	6,93	4,85	4,01	3,54	3,23	3,01	2,84	2,72	2,61	2,52	2,45	2,39	2,33	2,29	2,24	2,21	2,17	2,14
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,43	2,37	2,31	2,26	2,22	2,19	2,15	2,12
125	6,84	4,78	3,94	3,47	3,17	2,95	2,79	2,66	2,55	2,47	2,39	2,33	2,28	2,23	2,19	2,15	2,11	2,08
150	6,81	4,75	3,92	3,45	3,14	2,92	2,76	2,63	2,53	2,44	2,37	2,31	2,25	2,20	2,16	2,12	2,09	2,06
200	6,76	4,71	3,88	3,41	3,11	2,89	2,73	2,60	2,50	2,41	2,34	2,27	2,22	2,17	2,13	2,09	2,06	2,02
300	6,72	4,68	3,85	3,38	3,08	2,86	2,70	2,57	2,47	2,38	2,31	2,24	2,19	2,14	2,10	2,06	2,03	1,99
500	6,69	4,65	3,82	3,36	3,05	2,84	2,68	2,55	2,44	2,35	2,28	2,22	2,17	2,12	2,07	2,04	2,00	1,97
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,66	2,53	2,43	2,34	2,27	2,20	2,15	2,10	2,06	2,02	1,99	1,95
	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,13	2,08	2,04	2,00	1,97	1,93

VII. táblázat folytatása

A számított szabadságfokok: f ₁																	
9	20	22	24	26	28	30	35	40	45	50	60	80	100	200	500	∞	
Az előző sorban(f ₂ =1) számokat 10-zel meg kell szorozni																	
620	621	622	623	624	625	626	628	629	630	630	631	633	633	635	636	637	1
95,4	95,4	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	95,5	2
26,7	26,7	26,6	26,6	26,6	26,5	26,5	26,5	26,4	26,4	26,4	26,4	26,3	26,3	26,2	26,2	26,1	3
14,0	14,0	14,0	13,9	13,9	13,9	13,9	13,8	13,7	13,7	13,7	13,7	13,6	13,6	13,5	13,5	13,5	4
9,56	9,55	9,51	9,47	9,43	9,40	9,38	9,33	9,29	9,29	9,24	9,20	9,16	9,13	9,08	9,04	9,02	5
7,42	7,40	7,35	7,31	7,28	7,25	7,23	7,18	7,14	7,11	7,09	7,06	7,01	6,99	6,93	6,90	6,88	6
6,16	6,16	6,11	6,07	6,04	6,02	5,99	5,94	5,91	5,88	5,86	5,82	5,78	5,75	5,70	5,67	5,65	7
5,38	5,36	5,32	5,28	5,25	5,22	5,20	5,15	5,12	5,09	5,07	5,03	4,99	4,96	4,91	4,88	4,86	8
4,63	4,61	4,57	4,53	4,50	4,47	4,45	4,40	4,37	4,34	4,32	4,28	4,24	4,21	4,16	4,13	4,11	9
4,43	4,41	4,36	4,33	4,30	4,27	4,25	4,20	4,17	4,14	4,12	4,08	4,04	4,01	3,96	3,93	3,91	10
4,12	4,10	4,06	4,02	3,99	3,96	3,94	3,89	3,86	3,83	3,81	3,78	3,73	3,71	3,66	3,62	3,60	11
3,86	3,86	3,82	3,78	3,75	3,72	3,70	3,65	3,62	3,59	3,57	3,54	3,49	3,47	3,41	3,38	3,36	12
3,69	3,66	3,62	3,58	3,56	3,53	3,51	3,46	3,43	3,40	3,38	3,34	3,30	3,27	3,22	3,19	3,17	13
3,53	3,51	3,46	3,43	3,40	3,37	3,35	3,30	3,27	3,24	3,22	3,18	3,14	3,11	3,06	3,03	3,00	14
3,40	3,37	3,33	3,29	3,26	3,24	3,21	3,17	3,13	3,10	3,08	3,05	3,00	2,98	2,92	2,89	2,87	15
3,28	3,26	3,22	3,18	3,15	3,12	3,10	3,05	3,02	2,99	2,97	2,93	2,89	2,86	2,81	2,78	2,75	16
3,16	3,16	3,12	3,08	3,05	3,03	3,00	2,96	2,92	2,89	2,87	2,83	2,79	2,76	2,71	2,68	2,65	17
3,10	3,08	3,03	3,00	2,97	2,94	2,92	2,87	2,84	2,81	2,78	2,75	2,70	2,68	2,62	2,59	2,57	18
3,03	3,00	2,96	2,92	2,89	2,87	2,84	2,80	2,76	2,73	2,71	2,67	2,63	2,60	2,55	2,51	2,49	19
2,96	2,94	2,90	2,86	2,83	2,80	2,78	2,73	2,69	2,67	2,64	2,61	2,56	2,54	2,48	2,44	2,42	20
2,90	2,88	2,84	2,80	2,77	2,74	2,72	2,67	2,64	2,61	2,58	2,55	2,50	2,48	2,42	2,38	2,36	21
2,83	2,83	2,78	2,75	2,72	2,69	2,67	2,62	2,58	2,55	2,53	2,50	2,45	2,42	2,36	2,33	2,31	22
2,80	2,78	2,74	2,70	2,67	2,64	2,62	2,57	2,54	2,51	2,48	2,45	2,40	2,37	2,32	2,28	2,26	23
2,76	2,74	2,70	2,66	2,63	2,60	2,58	2,53	2,49	2,46	2,44	2,40	2,36	2,33	2,27	2,24	2,21	24
2,72	2,70	2,66	2,62	2,59	2,56	2,54	2,49	2,45	2,42	2,40	2,36	2,32	2,29	2,23	2,19	2,17	25
2,69	2,66	2,62	2,58	2,55	2,53	2,50	2,45	2,42	2,39	2,36	2,33	2,28	2,25	2,19	2,16	2,13	26
2,66	2,63	2,59	2,55	2,52	2,49	2,47	2,42	2,38	2,35	2,33	2,29	2,25	2,22	2,16	2,12	2,10	27
2,63	2,60	2,56	2,52	2,49	2,46	2,44	2,39	2,35	2,32	2,30	2,26	2,22	2,19	2,13	2,09	2,06	28
2,60	2,57	2,53	2,49	2,46	2,44	2,41	2,36	2,33	2,30	2,27	2,23	2,19	2,16	2,10	2,06	2,03	29
2,57	2,55	2,51	2,47	2,44	2,41	2,39	2,34	2,30	2,27	2,25	2,21	2,16	2,13	2,07	2,03	2,01	30
2,53	2,50	2,46	2,42	2,39	2,36	2,34	2,29	2,25	2,22	2,20	2,16	2,11	2,08	2,02	2,00	1,96	32
2,49	2,46	2,42	2,38	2,35	2,32	2,30	2,25	2,21	2,18	2,16	2,12	2,07	2,04	1,98	1,94	1,91	34
2,45	2,43	2,38	2,35	2,32	2,29	2,26	2,21	2,17	2,14	2,12	2,08	2,03	2,00	1,94	1,90	1,87	36
2,42	2,40	2,35	2,32	2,28	2,26	2,23	2,18	2,14	2,11	2,09	2,05	2,00	1,97	1,90	1,86	1,80	38
2,39	2,37	2,33	2,29	2,26	2,23	2,20	2,15	2,11	2,08	2,06	2,02	1,97	1,94	1,87	1,83	1,84	40
2,37	2,34	2,30	2,26	2,23	2,20	2,16	2,13	2,09	2,06	2,03	1,99	1,94	1,91	1,85	1,80	1,78	42
2,35	2,32	2,28	2,24	2,21	2,18	2,15	2,10	2,06	2,03	2,01	1,97	1,92	1,89	1,82	1,78	1,75	44
2,33	2,30	2,26	2,22	2,19	2,16	2,13	2,08	2,04	2,01	1,99	1,95	1,90	1,86	1,80	1,75	1,73	46
2,31	2,28	2,24	2,20	2,17	2,14	2,12	2,06	2,02	1,99	1,97	1,93	1,88	1,84	1,78	1,73	1,70	48
2,29	2,27	2,22	2,18	2,15	2,12	2,10	2,05	2,01	1,97	1,95	1,91	1,86	1,82	1,76	1,71	1,68	50
2,25	2,23	2,18	2,15	2,11	2,08	2,06	2,01	1,97	1,93	1,91	1,87	1,81	1,78	1,71	1,67	1,64	55
2,22	2,20	2,10	2,12	2,08	2,05	2,03	1,99	1,94	1,90	1,88	1,84	1,78	1,75	1,68	1,63	1,60	60
2,20	2,17	2,09	2,06	2,03	2,00	1,98	1,93	1,88	1,85	1,81	1,75	1,72	1,65	1,60	1,57	1,54	65
2,18	2,15	2,11	2,07	2,03	2,01	1,98	1,93	1,89	1,85	1,83	1,78	1,73	1,70	1,62	1,57	1,54	70
2,14	2,12	2,07	2,03	2,00	1,97	1,94	1,89	1,85	1,81	1,79	1,75	1,69	1,66	1,58	1,53	1,49	80
2,11	2,09	2,04	2,00	1,97	1,94	1,92	1,86	1,82	1,79	1,76	1,72	1,66	1,62	1,54	1,49	1,46	90
2,09	2,07	2,02	1,98	1,94	1,92	1,89	1,84	1,80	1,76	1,73	1,69	1,63	1,60	1,52	1,47	1,43	100
2,05	2,03	1,98	1,94	1,91	1,88	1,85	1,80	1,76	1,72	1,69	1,65	1,59	1,55	1,47	1,41	1,37	125
2,03	2,00	1,96	1,92	1,88	1,85	1,83	1,77	1,73	1,69	1,66	1,62	1,56	1,52	1,43	1,38	1,33	150
2,00	1,97	1,93	1,89	1,85	1,82	1,79	1,74	1,69	1,66	1,63	1,59	1,52	1,48	1,39	1,33	1,28	200
1,97	1,94	1,89	1,85	1,82	1,79	1,76	1,71	1,66	1,62	1,59	1,55	1,48	1,44	1,35	1,28	1,22	300
1,94	1,92	1,87	1,83	1,79	1,76	1,74	1,68	1,63	1,60	1,56	1,52	1,45	1,41	1,31	1,23	1,16	500
1,92	1,89	1,85	1,81	1,77	1,74	1,72	1,66	1,61	1,57	1,54	1,50	1,43	1,38	1,28	1,19	1,11	1000
1,90	1,88	1,83	1,78	1,75	1,72	1,70	1,64	1,59	1,55	1,52	1,47	1,40	1,36	1,25	1,15	1,00	

17. fejelet

T₂ teljes létszámra igazolható

A Kolmogorov-féle $K(z)$ függvény

VIII. táblázat

z	K(z)	z	K(z)	z	K(z)
0,28	0,000001	0,71	0,305471	1,14	0,851394
0,29	0,000004	0,72	0,322265	1,15	0,858038
0,30	0,000009	0,73	0,339113	1,16	0,864442
0,31	0,000021	0,74	0,355981	1,17	0,870612
0,32	0,000046	0,75	0,372833	1,18	0,876548
0,33	0,000091	0,76	0,389640	1,19	0,882258
0,34	0,000171	0,77	0,406372	1,20	0,887750
0,35	0,000303	0,78	0,423002	1,21	0,893030
0,36	0,000511	0,79	0,439505	1,22	0,898104
0,37	0,000826	0,80	0,455857	1,23	0,902972
0,38	0,001285	0,81	0,472041	1,24	0,907648
0,39	0,001929	0,82	0,488030	1,25	0,912132
0,40	0,002808	0,83	0,503808	1,26	0,916432
0,41	0,003972	0,84	0,519366	1,27	0,920556
0,42	0,005476	0,85	0,534682	1,28	0,924505
0,43	0,007377	0,86	0,549744	1,29	0,928288
0,44	0,009730	0,87	0,564546	1,30	0,931908
0,45	0,012590	0,88	0,579070	1,31	0,935370
0,46	0,016005	0,89	0,593316	1,32	0,938682
0,47	0,020022	0,90	0,607270	1,33	0,941848
0,48	0,024683	0,91	0,620928	1,34	0,944872
0,49	0,030017	0,92	0,634286	1,35	0,947756
0,50	0,36055	0,93	0,647338	1,36	0,950512
0,51	0,042814	0,94	0,660082	1,37	0,953142
0,52	0,50306	0,95	0,672516	1,38	0,955650
0,53	0,058534	0,96	0,684636	1,39	0,958040
0,54	0,067497	0,97	0,696444	1,40	0,960318
0,55	0,077183	0,98	0,707940	1,41	0,962486
0,56	0,087577	0,99	0,719126	1,42	0,964552
0,57	0,098656	1,00	0,730000	1,43	0,966516
0,58	0,110395	1,01	0,740566	1,44	0,968382
0,59	0,122760	1,02	0,750825	1,45	0,970158
0,60	0,135718	1,03	0,760780	1,46	0,971846
0,61	0,149223	1,04	0,770434	1,47	0,973448
0,62	0,163225	1,05	0,779794	1,48	0,974970
0,63	0,177753	1,06	0,788860	1,49	0,976412
0,64	0,192677	1,07	0,797636	1,50	0,977782
0,65	0,207987	1,08	0,806128	1,51	0,979080
0,66	0,223637	1,09	0,814342	1,52	0,980310
0,67	0,239582	1,10	0,822282	1,53	0,981476
0,68	0,255780	1,11	0,829950	1,54	0,982578
0,69	0,272189	1,12	0,837356	1,55	0,983622
0,70	0,288765	1,13	0,844502	1,56	0,984610

VIII. táblázat folytatása

z	K(z)	z	K(z)	z	K(z)
1,57	0,985544	1,93	0,998837	2,29	0,999944
1,58	0,986426	1,94	0,998924	2,30	0,999949
1,59	0,987260	1,95	0,999004	2,31	0,999954
1,60	0,988043	1,96	0,999079	2,32	0,999958
1,61	0,988791	1,97	0,999149	2,33	0,999962
1,62	0,989492	1,98	0,999213	2,34	0,999965
1,63	0,990154	1,99	0,999273	2,35	0,999968
1,64	0,990777	2,00	0,999329	2,36	0,999970
1,65	0,991364	2,01	0,999380	2,37	0,9999
1,66	0,991917	2,02	0,999428	2,38	0,999976
1,67	0,992438	2,03	0,999474	2,39	0,999978
1,68	0,992928	2,04	0,999516	2,40	0,999980
1,69	0,993389	2,05	0,999552	2,41	0,999982
1,70	0,993828	2,06	0,999588	2,42	0,999984
1,71	0,994230	2,07	0,999620	2,43	0,999986
1,72	0,994612	2,08	0,999650	2,44	0,999987
1,73	0,994972	2,09	0,999680	2,45	0,999988
1,74	0,995309	2,10	0,999705	2,46	0,999989
1,75	0,995625	2,11	0,999723	2,47	0,999990
1,76	0,995922	2,12	0,999750	2,48	0,999991
1,77	0,996200	2,13	0,999770	2,49	0,999992
1,78	0,996460	2,14	0,999790	2,50	0,999925
1,79	0,996704	2,15	0,999806	2,55	0,999956
1,80	0,996932	2,16	0,999822	2,60	0,999974
1,81	0,997146	2,17	0,999838	2,65	0,999984
1,82	0,997346	2,18	0,999852	2,70	0,999990
1,83	0,997533	2,19	0,999864	2,75	0,999994
1,84	0,997707	2,20	0,999874	2,80	0,999997
1,85	0,997870	2,21	0,999886	2,85	0,999982
1,86	0,998023	2,22	0,999896	2,90	0,999990
1,87	0,998145	2,23	0,999904	2,95	0,999994
1,88	0,998297	2,24	0,999912	3,00	0,999997
1,89	0,998421	2,25	0,999920		
1,90	0,998536	2,26	0,999926		
1,91	0,998644	2,27	0,999934		
1,92	0,998744	2,28	0,999940		