

15.1 KÍSÉRLETEK TERVEZÉSE

15.1.1 Mintavétel

A statisztika egyik alapproblémája, hogy hogyan lehet egy sokaságra következtetéseket levonni anélkül, hogy a sokaság minden elemét megvizsgálánk. Tegyük fel például, hogy nagy mennyiségű tojást akarunk megvizsgálni. Ha csak nagyság szerinti eloszlásukra vagyunk kíváncsiak, azaz arra, hogy a tojások hány százaléka rendelkezik bizonyos nagysággal, akkor igen egyszerű módon, a tojások nagyság szerinti osztályozásával választ kaphatunk kérdésünkre. Ha azonban már a tojás sárgájának nagyságára akarunk felvilágosítást kapni (tojásexportnál például fontos kritérium, hogy a tojás sárgája elég nagy legyen), akkor nem járhatunk el úgy, hogy az összes tojást feltörjük és megmérjük a sárgáját, hanem csak azt tehetjük, hogy kiválasztunk a sokaságból néhány tojást (*mintát*), ezeket megvizsgáljuk, és ennek alapján próbálunk következtetéseket levonni az egész sokaságra vonatkozóan. Ebben a példában *kénytelenek vagyunk* az egész sokaság megvizsgálása helyett egy mintát megvizsgálni (nyilván nem törhetjük fel az összes exportálandó tojást) máskor, bár lehetséges lenne a sokaság összes elemének megvizsgálása, mégis csak egy mintát nézünk meg, mert az egész sokaság megvizsgálása túl költséges lenne, és úgy gondoljuk, hogy a minta alapján is elegendő felvilágosítást nyerhetünk a felvetett kérdésekre. Például, ha nagy mennyiségű naposcsibe súlyeloszlására vagyunk kíváncsiak, akkor bár lehetséges mindegyik csirke súlyát megmérni, mégis inkább egy minta megvizsgálása alapján döntünk, mert az összes csirke megmérése hatalmas munka lenne (sőt esetleg olyan hosszú időt venne igénybe a mérés, hogy az utoljára lemerít — immár mondjuk kétnapos — csibék súlya nagyobb lenne, már koruknál fogva is, mint az először lemerít csibéé, és így az idő múlása a kísérlet eredményeit oly mértékben zavarhatja, hogy az eredmény pontatlanabb, mintha mintavétel alapján kaptuk volna.

Az említett példák mutatják, hogy a mintavétel sok statisztikai vizsgálat első lépése. Bár első pillanatban úgy tűnik, hogy a mintavétel — egy sokaság néhány elemének véletlenszerű kiválasztása — igen egyszerű dolog, mégis szükséges, hogy néhány szabályára felhívjuk a figyelmet, annál is inkább, mert ha már maga a mintavétel nem volt megfelelő, akkor a későbbi vizsgálatok sem vezethetnek kielégítő eredményre.

A mintával kapcsolatban nyilván azt szeretnénk elérni, hogy a minta *reprezentatív* legyen, azaz ugyanolyan tulajdonságokkal rendelkezzen, mint az egész sokaság. Pontosabban a kérdéses adat (csirke súlya, tojás sárgájának nagysága) ugyanolyan eloszlással szerepeljen a mintában, mint a sokaságban. Nézzünk ezzel kapcsolatban néhány példát, amelyek mutatják a felvetődő nehézségeket. Például a naposcsibék esetében nagy mennyiségű csirke közül véletlenszerűen kiválasztani néhányat. Elképzelhető például, hogy az egészségesebb és ezzel együtt súlyosabb csibék mozgékonyabbak, mint a többiek. Bár azt hisszük, hogy véletlenszerűen választjuk a mintát, mégis általában a kevésbé mozgékony állatokat választjuk, és így a mintából megállapított átlagsúly kisebb, mint az egész sokaság átlagsúlya. Hasonló probléma vetődhet fel, ha egy erdőből kitermelhető fa mennyiségét akarjuk megbecsülni. Ekkor kiválaszthatunk mintaként az erdőből néhány fát, és ezeknek meghatározzuk a súlyát (vagy úgy, hogy a mintaként kijelölt fákat kivágjuk, vagy csak megbecsüljük a kijelölt élőfák súlyát). Természetesen a kijelölés módja itt is lényeges. Hiszen lehetséges, hogy az erdő valamely részén a talaj minősége jobb, és az ott élő fák nagyobbak,

így hibás következtetéseket vonhatunk le, ha a mintaelemeket túlnyomóan az erdő valamely speciális (például könnyen elérhető) részéből választjuk ki.

Gondot kell fordítanunk arra is, hogy a mintaelemeket egymástól függetlenül válasszuk ki. Például, ha nagy mennyiségű tojást kell megvizsgálnunk, nem járunk el helyesen, ha mintaelemként egymáshoz közel elhelyezett tojásokat vizsgálunk meg. Hiszen feltehető, hogy általában egy tyúk tojásai kerültek egymás mellé, és ebben az esetben a minta nem az egész sokaságra, hanem csak egy vagy néhány tojóra lesz jellemző.

Azt mondhatjuk, hogy a mintavétel biztosan helyes lesz, ha a sokaság elemeit megszámozzuk, a szereplő számokat jól összekeverve bedobjuk egy urnába, majd — mint a lottónál a nyertes számokat — kihúzzuk a mintába kerülő elemek számát. Bár ez az eljárás feltétlenül megfelelő eredményre vezet, mégis igen ritkán alkalmazható, mivel igen fáradságos munka lenne például, több ezer csirke mindegyikére egy számot tenni. Bárhogy is történik azonban a mintavétel, úgy kell eljárunk, hogy módszerünk minél jobban hasonlítson a sorsolással való kiválasztáshoz.

A kísérletet végző számára természetesen fontos tudni, hogy mekkora mintát kell vennie. Erre természetesen általános választ adni nem lehet, a minta nagysága függ a felvetett kérdéstől és attól, hogy mennyire pontos és megbízható választ akarunk arra kapni. Azonban a legpontosabban felvetett kérdés esetében sem biztos, hogy minden előzetes adat nélkül meg lehet mondani, hogy milyen nagyságú minta szükséges a kérdés megválaszolásához. A későbbiekben az egyes konkrét statisztikai módszerek ismertetésénél kitérünk majd arra is, hogy mit lehet mondani a szükséges mintaelemek számáról.

Itt csak azt jegyezzük meg meg, hogy a körülmények ismerete és a megfelelően végzett mintavétel a szükséges mintaelemszámot nagymértékben csökkentheti. Például, ha Magyarország tejtermelésének bizonyos adatairól akarunk képet kapni mintavétel útján, akkor eljárhatunk úgy is, hogy Magyarország tehénállományából, véletlenszerűen kiválasztunk teheneket, és megvizsgáljuk az ezek által adott tejet. Eljárhatunk azonban olyképpen is, hogy az országot felosztjuk tájegységekre úgy, hogy az egyes egységekről feltehető, ott a tehénállomány lényegében egységes, és a tejtermelést befolyásoló tényezők is változatlanok. A mintavételt az egyes egységekben külön-külön végezzük el, majd a kapott eredményeket összesítjük. Az ilyen eljárást *rétegezett mintavételnek* nevezzük. A rétegezett mintavételi eljárás segítségével — mint később látni fogjuk — sokkal kisebb elemszámú mintával dolgozhatunk, azonban ez az eljárás csak akkor alkalmazható, ha biztosak vagyunk abban, hogy a *rétegezés* helyes, azaz, ha az egyes egységek a kérdéses szempontból homogének.

15.1.2 Többfaktoros kísérletek

Igen gyakran találkozunk olyan kísérletekkel, amikor több különböző tényező (faktor) együttes hatását kívánjuk jellemezni. Például több különböző fajta műtrágya együttes hatására vagyunk kíváncsiak, és egyúttal azt is tudni akarjuk, hogy melyik fajtából mennyit célszerű adagolni. Az ilyen típusú kísérletek megtervezése különös gondot igényel, ha lehetőleg kevés kísérleti költséggel (pl. kis kísérleti földterülettel) akarunk az összes felvetődő kérdésre egyidejűleg választ kapni. A statisztikusok sok kísérleti tervet dolgoztak ki ilyen célokra. Itt példaként bemutatunk néhányat.

15.1.2.1 Latin négyzetek

Tegyük fel, hogy több (mondjuk n) különböző növényfajtát (pl. kukoricafajtát) akarunk összehasonlítani valamilyen szempontból (például terméshezam). A kérdés első pillanatban rendkívül egyszerűnek látszik, hiszen — mondhatjuk — nem kell mást tennünk, mint elvetjük a szóban forgó n kukoricafajtát és megnézzük, melyik adja a legtöbb termést. Azonban ezzel kapcsolatban két kérdés vetődik fel. Az egyik, hogy ha a rendelkezésünkre álló (pl. négyzet alakú) földterületet felosztjuk n részre párhuzamos vonalakkal (15.1. ábra), és az

1	2	...	n
---	---	-----	-----

15.1 ábra

egyes fajtákat az egyes sávokba vetjük, akkor a terméseredményt nemcsak a kísérletben szereplő növényfajták közötti, hanem a talajban levő inhomogenitások is nagymértékben befolyásolják. Előfordulhat például, hogy az 1. fajta adja a legjobb terméseredményt, bár az nem a legjobb, csak véletlenül a legjobb minőségű földbe került. (Meggjegyezzük, hogy a talaj inhomogenitása még viszonylag kis vetésterület esetén is jelentős lehet.)

A második kérdés a kísérlet eredményének kiértékelésével kapcsolatban vetődik fel. Ha az első növényfajtából hektáronként 2 q-val nagyobb termést kaptunk, mint a másodikból, és a talaj inhomogenitása nem befolyásolta az eredményt, akkor nyilván mondhatjuk, hogy az első fajta jobb, mint a második. Ha azonban a különbség hektáronként csak 5 kg, akkor nyilván senki sem meri azt állítani, hogy a különbség elég jelentős (*szignifikáns*) ahhoz, hogy az első fajtát jobbnak mondjuk, mint a másodikat, hiszen ilyen kis különbség — véletlentől függően — teljesen egyforma értékű fajták között is előfordulhat.

A talaj inhomogenitását már a kísérlet tervezésével ki kell küszöbölni. (Ezzel a kérdéssel foglalkozunk most.) Annak eldöntése, hogy mely különbség tekinthető szignifikánsnak, a kísérlet kiértékeléséhez tartozik. (Ezzel a 15.3 pontban foglalkozunk.)

Mint említettük, a 15.1 ábrán bemutatott vetésterv alkalmazása esetén a talaj inhomogenitása nagymértékben befolyásolhatja a kísérlet eredményét. Kiküszöbölhetjük ezt a tényezőt, ha a földterületünket felosztjuk n^2 kisebb négyzetre, sakktáblaszerűen, és vetéstervünket úgy készítjük el, hogy minden sorba és minden oszlopba mindegyik fajtából egy és csak egy jusson. $n = 3$, $n = 4$ és $n = 5$ esetén ezt például a 15.2 ábrán bemutatott módon tehetjük meg.

1	2	3
2	3	1
3	1	2

1	2	3	4
2	3	4	1
3	4	1	2
4	1	2	3

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

15.2 ábra

Nyilvánvaló, hogy egy ilyen vetésterv a talaj inhomogenitásának hatását szinte biztosan kiküszöböli.

Definíció. Az olyan elrendezést (amikor mindegyik sorba és mindegyik oszlopba a fajták mindegyikéből egy és csak egy kerül) *latin négyzetnek* nevezzük.

Tetszőleges n esetén latin négyzetet például úgy készíthetünk, hogy négyzetünk első sorába rendre beírjuk az 1, 2, ..., n számokat, a második sorba a 2, 3, ..., n , 1, a harmadik sorba a 3, 4, 5, ..., $(n-1)$, n , 1, 2, az utolsó sorba az n , 1, 2, ..., $(n-1)$ számok kerülnek. (Azaz az egyes sorokba az 1, 2, ..., n számok ciklikus permutációi kerülnek.)

15.1.2.2 Véletlen blokktervek

Mint említettük, a latin négyzet alkalmazásának nagy előnye, hogy a talaj inhomogenitásából származható hibákat csaknem teljesen kiküszöböli. Hátránya azonban, hogy túl bonyolult, pontosabban, hogy viszonylag kevés növényfajta kipróbálásához is túl sok blokkra van szükség (n növényfajta n^2 blokk szükséges). Ezt a hátrányt kívánjuk a véletlen blokk-elrendezés segítségével kiküszöbölni. Persze az egyszerűsítés csökkenti a kísérlet eredményének megbízhatóságát. (Sok esetben egy nagyfokú egyszerűsítés megéri a biztonság kismértvű csökkenését.)

Véletlen blokkelrendezést a következő módon készítünk. n növényfajta kipróbálása esetén osszuk fel kísérleti területünket k részre, blokkra. (k -t általában n -nél kisebbre választjuk, de pontos megválasztására nem tudunk szabályt mondani.) Ezután mindegyik blokkot osszuk fel n egyenlő részre. Az n növényfajta mindegyikét elvetjük mindegyik blokkba, de azt, hogy egy blokkon belül melyik fajta melyik helyre kerül, a véletlenre bizzuk (például sorsolással dönthetünk).

$n = 6$ esetén, k értékét 3-ra választva, például a következő véletlen blokkelrendezéshez juthatunk.

1. blokk	2	4	3	5	1	6
2. blokk	4	1	6	2	5	3
3. blokk	3	2	1	5	4	6

15.3 ábra

Itt az egyes blokkokon belüli elrendezést sorsolással határoztuk meg. Nyilvánvaló, hogy egy ilyen elrendezésnél nem lehetünk teljesen biztosak abban, hogy a talaj inhomogenitása nem befolyásolja-e kísérletünk eredményét, hiszen például előfordulhat, hogy valamelyik oszlopban valamelyik fajta túl sokszor előfordul, valamely másik fajta pedig egyáltalán nem fordul elő. Azonban, ha k értéke nem túl kicsi n -hez képest, akkor valószínűtlen, hogy az inhomogenitás komolyan zavarná kísérletünket.

Készíthetünk véletlen blokktervet úgy is, hogy a növényeknek az egyes blokkokon belüli elhelyezését nem bizzuk *teljesen* a véletlenre, hanem egy előre elkészített $n \times n$ -es latin négyzet k sorát választjuk ki véletlenszerűen, és az így kapott sorokat írjuk sorban a véletlen blokkterv egyes blokkjaiba.

15.1.2.3 Ortogonális latin négyzetek

A latin négyzetnél és a véletlen blokktervénél két faktort — a növényfajtákat és a talaj minőségét — vettük tekintetbe. (Ezek közül is az egyiknek — a talaj minőségének — a hatására nem voltunk kíváncsiak, hanem azt csak kiküszöbölni akartuk a kísérleti terv segítségével.) A következő kísérleti tervek azt a célt szolgálják, hogy egyidejűleg több faktor hatását is megfigyelhessük.

Tegyük fel például, hogy n kukoricafajtát és n kezelési módot, pl. trágyafajtát akarunk kipróbálni egyidejűleg, de azt is látni akarjuk, hogy az egyes kezelési módok milyen hatással vannak az egyes fajtákra. Azaz mindegyik kukoricafajtát mindegyik trágyafajttal kívánjuk kezelni. A vetésterv elkészítésénél a talaj inhomogenitásából származó hibákat is szeretnénk kiküszöbölni. Ezek a követelmények megvalósulnak, ha a kukoricafajtákat is egy $n \times n$ -es latin négyzetnek megfelelően vetjük, és a trágyafajtákat is egy $n \times n$ -es latin négyzet szerint helyezzük el, de a két latin négyzet „egymásra helyezését” vigyázunk arra, hogy mindegyik kukoricafajta mindegyik trágyafajttal együtt előforduljon. Két ilyen tulajdonságú latin négyzetet *ortogonálisnak* nevezünk.

Definíció. Két $n \times n$ -es latin négyzet ortogonális, ha a megfelelő elemeit egymásra helyezve minden lehetséges elempár előfordul. $n = 3$ esetén például a következő vetéstervet alkalmazhatjuk:

$k_1 t_1$	$k_2 t_2$	$k_3 t_3$
$k_2 t_3$	$k_3 t_1$	$k_1 t_2$
$k_3 t_2$	$k_1 t_3$	$k_2 t_1$

15.4 ábra

(Ez azt jelenti, hogy a kísérleti terület bal felső négyzetébe az 1. kukoricafajtát és az 1. trágyafajtát tesszük.)

Látható, hogy itt az egyes kukoricafajták (k_1, k_2, k_3) és az egyes trágyafajták (t_1, t_2, t_3) valóban latin négyzetet alkotnak, és mindegyik kukoricafajta mindegyik trágyafajttal együtt előfordul, azaz a k_1, k_2, k_3 elemeket és a t_1, t_2, t_3 elemeket tartalmazó latin négyzetek ortogonálisak.

A 4×4 -es ortogonális latin négyzetekre a következő példát adhatjuk:

1	2	3	4	1	3	2	4
2	1	4	3	4	2	3	1
4	3	2	1	2	4	1	3
3	4	1	2	3	1	4	2

Tekintsük ezen a két latin négyzeten kívül még az

1	3	4	2
4	2	1	3
3	1	2	4
2	4	3	1

4×4 -es latin négyzetet. Könnyű észrevenni, hogy ez a latin négyzet ortogonális a fenti két 4×4 -es latin négyzet mindegyikére. Így 3 darab 4×4 -es, páronként ortogonális latin négyzetet kapunk.

Ilyen elrendezést használhatunk abban az esetben, ha 3 különféle faktort kívánunk összehasonlítani. Például, ha 4 különféle kukoricafajtát akarunk összehasonlítani (első faktor) és két különféle trágyafajtát (második és harmadik faktor) használunk, de mindegyik trágyafajtából az optimális mennyiséget is meg akarjuk határozni. Legyen például a 4 kukoricafajta k_1, k_2, k_3, k_4 , az első trágyafajtából hektáronként próbaként t_1, t_2, t_3, t_4 mázsát helyezünk el, a második fajtból pedig s_1, s_2, s_3, s_4 mennyiségek alkalmazása esetén vagyunk kíváncsiak a hatásra. Ekkor a már felírt 3 páronként ortogonális latin négyzet alkalmazásával a következő kísérleti tervet készíthetjük:

$k_1 t_1 s_1$	$k_2 t_2 s_3$	$k_3 t_3 s_4$	$k_4 t_4 s_2$
$k_2 t_4 s_4$	$k_1 t_2 s_2$	$k_4 t_3 s_1$	$k_3 t_1 s_3$
$k_4 t_2 s_3$	$k_3 t_4 s_1$	$k_2 t_1 s_2$	$k_1 t_3 s_4$
$k_3 t_3 s_2$	$k_4 t_1 s_4$	$k_1 t_4 s_3$	$k_2 t_2 s_1$

15.5 ábra

(E terv szerint például a kísérleti terület bal felső négyzetébe az első kukoricafajtát vetjük, az első trágyafajtából t_1 , a második fajtból s_1 mennyiséget használunk.)

A latin négyzetek tárgyalásánál láttuk, hogy tetszőleges n esetén tudunk $n \times n$ -es latin négyzetet készíteni. Sokkal nehezebb kérdés annak eldöntése, hogy valamely n -re létezik-e két (vagy több) ortogonális $n \times n$ -es latin négyzet. Erre a kérdésre nem is kívánunk részletesebben kitérni, csak megemlítjük, azt, hogy egy n szám esetén hány páronként ortogonális $n \times n$ -es latin négyzet létezik, függ a szám számelméleti tulajdonságaitól. Például:

- $n = 3$ esetén 2 ortogonális latin négyzet van,
- $n = 4$ esetén 3 páronként ortogonális latin négyzet van,
- $n = 5$ esetén 4 páronként ortogonális latin négyzet van,
- $n = 6$ esetén nem létezik ortogonális latin négyzetpár,
- $n = 7$ esetén 6 páronként ortogonális latin négyzet van.

Az ortogonális latin négyzetek gyakorlati alkalmazásának egyik nehézsége az, hogy bizonyos n -ekre nem tudunk ortogonális latin négyzeteket konstruálni. Ezen úgy lehet segíteni, hogy egy nagyobb latin négyzetet tervezzünk, és mindegyik faktornak egy vagy több elemét ismételtén alkalmazzuk.

A gyakorlati alkalmazásnak másik nehézsége, hogy fel kellett tételeznünk, hogy mindegyik faktornak egyforma sok eleme van. (Például feltételeztük, hogy a kukoricafajták száma egyenlő a kipróbálásra kerülő trágyafajták számával. Ezen a nehézségen is segíthetünk oly módon, hogy azon faktornak, amelynek kevesebb eleme van, az egyik elemét többször alkalmazzuk.) Például, ha 3 kukoricafajtánk és csak 2 trágyafajtánk van, akkor is alkalmazhatjuk a 15.4 ábrán megadott vetéstervet úgy, hogy azokra a helyekre, ahová a terv szerint a t_3 trágyafajta kerül, például ismét a t_2 trágyafajtát tesszük.)

Bár a fenti két bekezdésben említett módszerek megoldják az ortogonális latin négyzetek alkalmazásával kapcsolatos problémákat, de csak azon az áron, hogy a kísérleti területet növeljük. Emiatt az ortogonális latin négyzetek alkalmazása csak akkor ajánlható, ha olyan kísérletekről van szó, amelyek nagy pontosságot igényelnek és fontosságuk miatt a kísérleti költségek magas volta nem tekinthető jelentős akadállynak.

15.1.2.4 Faktoriális kísérletek

Abban az esetben, amikor több faktort akarunk összehasonlítani, és mindegyik faktor több szinten fordul elő, általában úgynevezett faktoriális kísérleti terveket használunk. Tegyük fel például, hogy egy növényfajtát 3 különféle hatóanyagú trágyával kezelünk, és azt akarjuk kísérlettel eldönteni, hogy melyik fajtaból mennyit célszerű használni. Legyen a 3 trágyaféleség hatóanyaga: a kálium, a nitrogén és a foszfor. Káliumsóból a k_1, k_2 , pétisóból az n_1, n_2 és a szuperfoszfátból a p_1, p_2 mennyiségek hatására vagyunk kíváncsiak. (A k_1, k_2 , az n_1, n_2 , illetve a p_1, p_2 mennyiségek között 0 is lehet.) Ekkor a következő trágyázási lehetőségeket kell összehasonlítani:

$$k_1n_1p_1; k_1n_1p_2; k_1n_2p_1; k_2n_1p_1; k_2n_2p_1; k_2n_1p_2; k_1n_2p_2; k_2n_2p_2.$$

Például, ha $k_1 = n_1 = p_1 = 0$, akkor a felsorolt 8 lehetőség jelentése rendre: nem trágyázunk egyáltalán; csak foszforosat használunk; csak nitrogénest használunk; csak káliumosat használunk; káliumosat és nitrogénest használunk; káliumosat és foszforosat használunk; nitrogénest és foszforosat használunk; mind a három trágyafajtából használunk.

Ha ezt a 8 féle trágyázási módot egymás melletti területeken, pl. egy sorban próbáljuk ki, akkor ismételt figyelembe kell vennünk a talaj inhomogenitásából adódó hibákat. Ez kiküszöbölendő, használhatunk egy 8×8 -as latin négyzetet kísérleti tervként, amelynek természetesen minden sorában és minden oszlopában a felsorolt 8 trágyázási mód mindegyike egyszer és csak egyszer fordul elő. Gazdaságossági okokból azonban itt is többnyire eltekintünk a latin négyzet alkalmazásától, és véletlen blokktervet készítünk, amelynek egyes blokkjai 8–8 helyet tartalmaznak.

Általában, ha n különböző faktor szerepel kísérletünkben és az egyes faktorok $s_1, s_2, \dots, s_n$ szinten fordulnak elő, akkor azt mondjuk, hogy $s_1 \times s_2 \times \dots \times s_n$ típusú faktoriális kísérletet használunk. (Az említett példában $n = 3$ faktor fordul elő, mindegyik két szinten, azaz $s_1 = s_2 = s_3 = 2$ volt, tehát $2 \times 2 \times 2$ típusú faktoriális kísérlettel foglalkoztunk.) Könnyen belátható, hogy egy $s_1 \times s_2 \times \dots \times s_n$ típusú faktoriális kísérletben $s_1 \cdot s_2 \cdot \dots \cdot s_n$ lehetőséget kell összehasonlítani. Például egy $2 \times 2 \times 3$ típusú faktoriális kísérlet esetén 3 faktor szerepel, ezek közül az első két szinten, legyenek ezek a_1, a_2 , a második két szinten, b_1, b_2 , a harmadik 3 szinten c_1, c_2, c_3 . Ekkor a $2 \cdot 2 \cdot 3 = 12$ összehasonlításra váró eset a következő:

$$\begin{array}{cccc} a_1 & b_1 & c_1 & a_1 & b_2 & c_1 & a_2 & b_1 & c_1 & a_2 & b_2 & c_1 \\ a_1 & b_1 & c_2 & a_1 & b_2 & c_2 & a_2 & b_1 & c_2 & a_2 & b_2 & c_2 \\ a_1 & b_1 & c_3 & a_1 & b_2 & c_3 & a_2 & b_1 & c_3 & a_2 & b_2 & c_3 \end{array}$$

Az eddigiekben említett kísérleti tervekkel kapcsolatban mindig csak növénytermesztési példát hoztunk fel. Valóban ezeket a terveket főként növénytermesztési kísérletekkel kap-

csolatban szoktuk alkalmazni, de mindegyik szerepelt kísérleti tervnek több, más típusú alkalmazása is van. Most csak a faktoriális kísérleti tervek más alkalmazási területei közül említünk egyet.

Sértéstakarmányozási vizsgálatnál a következő 4 faktort akarjuk összehasonlítani.

1. Zöldtakarmány etetése (z), illetve nem (vagy minimális) juttatása ($\bar{z}$).
2. Korlátlan (t), illetve korlátozott ($\bar{t}$) mennyiségű takarmány adagolása.
3. Szabad (m), illetve korlátozott ($\bar{m}$) mozgási lehetőség.
4. Durva (f), illetve finomra őrölt ($\bar{f}$) takarmány.

Kísérletünk céljára $2 \times 2 \times 2 \times 2$ -es faktoriális kísérletet alkalmazhatunk, azaz a kísérleti állatokat 16 (lehetőleg egyenlő) csoportba osztjuk, és az egyes csoportokat rendre a következő módon kezeljük:

$$\begin{array}{cccc} 1. ztmf & 5. ztm\bar{f} & 9. z\bar{t}mf & 13. z\bar{t}m\bar{f} \\ 2. \bar{z}tmf & 6. \bar{z}tm\bar{f} & 10. z\bar{t}m\bar{f} & 14. \bar{z}t\bar{m}\bar{f} \\ 3. z\bar{t}mf & 7. z\bar{t}m\bar{f} & 11. z\bar{t}\bar{m}f & 15. z\bar{t}\bar{m}\bar{f} \\ 4. z\bar{t}m\bar{f} & 8. \bar{z}tm\bar{f} & 12. z\bar{t}\bar{m}\bar{f} & 16. \bar{z}t\bar{m}\bar{f} \end{array}$$

Tehát például az első csoportnak adunk zöldtakarmányt, egyéb takarmányt is korlátlan mértékben adagolunk, szabad mozgási lehetőséget biztosítunk, és finomra őrölt takarmányt juttatunk, míg a 7. csoportnak zöldtakarmányt nem adunk, de egyéb takarmányból korlátlan mennyiséget adagolunk, finomra őrölt takarmányt adunk, és mozgási lehetőséget csak korlátozottan biztosítunk.

15.2 VALÓSZÍNŰSÉGELOSZLÁSOK ISMERETLEN PARAMÉTEREKSEL

A valószínűség-számítási fejezetben általában abból indultunk ki, hogy adva van egy vagy több ismert eloszlással rendelkező valószínűségi változó. Majd olyan kérdéseket vetettünk fel, hogy mi mondható e valószínűségi változók várható értékéről, szórásáról, összegük eloszlásfüggvényéről stb. Természetesen a gyakorlatban ritkán fordul elő, hogy egy valószínűségi változónak az eloszlásfüggvényét teljesen ismerjük. Általában vagy az fordul elő, hogy semmit nem tudunk az eloszlásfüggvényről, vagy annak típusát ismerjük, csak paramétereit nem. Ilyen esetekben az ismeretlen eloszlásfüggvényt, vagy az ismert eloszlásfüggvény ismeretlen paramétereit statisztikai adatokból kell becsülnünk.

Láttuk — például a valószínűség-számítási fejezetben —, hogy az áruszállítványok egy raktárba, Poisson-eloszlás szerint érkeznek, azaz annak a valószínűsége, hogy T idő alatt k szállítvány érkezik:

$$\frac{(\lambda T)^k}{k!} e^{-\lambda T},$$

ahol λ az egységnyi idő alatt érkező szállítványok számának várható értéke. Nyilvánvaló, hogy λ értéke elméleti úton már nem határozható meg, hanem csak statisztikai adatokból becsülhető (azaz megnézzük, hogy átlagban időegységenként hány szállítvány érkezik).

Jelen fejezetben azzal a kérdéssel foglalkozunk, hogy egy ismert eloszlás ismeretlen paraméterére vonatkozóan milyen következtetések vonhatók le a statisztikai adatokból. A későbbiekben megvizsgáljuk azt a kérdést is, hogy a statisztikai adatokból milyen következtetést vonhatunk le egy eloszlásfüggvényre vonatkozóan, ha arról eleve semmit sem tudunk.

15.2.1 Ismeretlen valószínűség

15.2.1.1 A valószínűség becslése

Egy egyébként hatásos védőoltás után 10 000 beoltott állatból 30 elhullott. Azt kérdezzük, hogy e kísérlet alapján hogyan jellemezhetjük a gyógyszer veszélyességét, azaz mi mondható annak az eseménynek a valószínűségéről, hogy az oltás hatására egy állat elhullik.

Természetesnek látszik, hogy ezen ismeretlen valószínűséget $\frac{30}{10\,000} = 0,003$ -nek becsüljük. Hiszen a nagy számok törvénye (lásd a Valószínűségszámítás c. fejezetet) szerint nagy kísérletszám esetén (itt 10 000) a relatív gyakoriság (itt 0,003) nagy valószínűséggel közel lesz a valószínűséghez. Ennek ellenére persze előfordulhat, hogy a relatív gyakoriság pl. jóval kisebb a valószínűségnél, és így az ismeretlen valószínűség jóval nagyobb, mint a kísérlet alapján becsült érték.

Célunk az ismeretlen valószínűségnek egy (a statisztikai adatoktól függő) alsó és felső becslését megtalálni úgy, hogy azt mondhassuk, hogy a felső becslés szinte biztosan (pl. 0,95 valószínűséggel) nagyobb, mint az ismeretlen valószínűség, és az alsó becslés szinte biztosan kisebb annál.

Jelöljük az ismeretlen valószínűséget p -vel, és számítsuk ki először annak a valószínűségét, hogy az elhullott állatok számának $\frac{k}{10\,000}$ relatív gyakorisága (ahol k az elhullott állatok száma) $p - \varepsilon$ és $p + \varepsilon$ között van (ahol ε tetszőleges pozitív szám).

Ez a valószínűség nyilván (lásd ismét a Valószínűségszámítás c. fejezetet):

$$\begin{aligned} P\left(p - \varepsilon \leq \frac{k}{10\,000} \leq p + \varepsilon\right) &= P(10\,000p - 10\,000\varepsilon \leq k \leq \\ &\leq 10\,000p + 10\,000\varepsilon) = \sum_{k=10\,000p-10\,000\varepsilon}^{10\,000p+10\,000\varepsilon} \binom{10\,000}{k} p^k (1-p)^{10\,000-k} \approx \\ &\approx \Phi\left(\frac{10\,000\varepsilon}{\sqrt{10\,000p(1-p)}}\right) - \Phi\left(\frac{-10\,000\varepsilon}{\sqrt{10\,000p(1-p)}}\right) = \\ &= 2\Phi\left(\frac{100\varepsilon}{\sqrt{p(1-p)}}\right) - 1. \end{aligned}$$

Következő lépésként válasszuk meg ε -t úgy, hogy ezen valószínűség 0,95 legyen, azaz:

$$\begin{aligned} 2\Phi\left(\frac{100\varepsilon}{\sqrt{p(1-p)}}\right) - 1 &= 0,95, \\ \Phi\left(\frac{100\varepsilon}{\sqrt{p(1-p)}}\right) &= 0,975 \end{aligned}$$

legyen. A Φ függvényre vonatkozó táblázat szerint:

$$\Phi(1,97) = 0,975,$$

$$\text{így } \frac{100\varepsilon}{\sqrt{p(1-p)}} = 1,97,$$

$$\text{és } \varepsilon = 0,0197 \sqrt{p(1-p)}.$$

Eredményünk szerint a relatív gyakoriság 0,95 valószínűséggel $p - 0,0197 \sqrt{p(1-p)}$ és $p + 0,0197 \sqrt{p(1-p)}$ között lesz. Ez más szóval azt jelenti, hogy ha a relatív gyakoriság kísérletünkben $\gamma = \frac{k}{10\,000}$, akkor 0,95 valószínűséggel $\gamma + 0,0197 \sqrt{p(1-p)}$ nagyobb az ismeretlen p valószínűségnél, míg $\gamma - 0,0197 \sqrt{p(1-p)}$ kisebb annál, vagyis

$$P(\gamma - 0,0197 \sqrt{p(1-p)} \leq p \leq \gamma + 0,0197 \sqrt{p(1-p)}) = 0,95.$$

Vizsgáljuk meg, hogy milyen p számok tesznek eleget az itt szereplő kikötésnek, azaz oldjuk meg a

$$\gamma - 0,0197 \sqrt{p(1-p)} \leq p \leq \gamma + 0,0197 \sqrt{p(1-p)}$$

egyenlőtlenséget.

Kapjuk, hogy

$$10^4(p - \gamma)^2 \leq (1,97)^2 p(1-p),$$

$$\text{azaz } [10^4 + (1,97)^2] p^2 - [2\gamma 10^4 + (1,97)^2] p + 10^4 \gamma^2 \leq 0.$$

Ezen másodfokú egyenlőtlenség megoldásaként kapjuk, hogy

$$\begin{aligned} \frac{2\gamma 10^4 + (1,97)^2 - 1,97 \sqrt{4 \cdot 10^4 \gamma + (1,97)^2 - 4 \cdot 10^4 \gamma^2}}{2(10^4 + (1,97)^2)} &\leq p \leq \\ &\leq \frac{2\gamma 10^4 + (1,97)^2 + 1,97 \sqrt{4 \cdot 10^4 \gamma + (1,97)^2 - 4 \cdot 10^4 \gamma^2}}{2(10^4 + (1,97)^2)}. \end{aligned}$$

Eredményünket összefoglalva azt mondhatjuk, hogy ha az oltás hatására 10 000 állatból 30 hullik el, akkor annak a valószínűsége, hogy az oltás megöli a beoltott állatot, szinte biztosan (0,95 valószínűséggel) nem nagyobb 0,0042-nél, és nem kisebb 0,0022-nél.

Ez azt jelenti, hogy ha egy gyógyszergyár gyógyszerének kipróbálásánál az említett kísérleti eredményt kapta, akkor állíthatja, hogy annak a valószínűsége, hogy a gyógyszer öl, nem nagyobb mint 0,0042. Természetesen, ha ez a valószínűség valóban 0,0042, akkor előfordulhat, hogy a beoltott állatoknak több mint 0,42%-a hullik el az oltás hatására. Könnyen kiszámítható, az előbbi módszer segítségével, hogy ha a szer 0,0042 valószínűséggel öl, akkor legfeljebb hány állat hullhat el (pl. 0,95 valószínűséggel) a szer alkalmazása folytán. Más esetekben is teljesen hasonló módon becsülhetjük a valószínűséget a relatív gyakoriság segítségével. Általános formulák kidolgozása helyett inkább még egy konkrét példán mutatjuk be a módszert.

Tegyük fel, hogy nagy mennyiségű (N darab) tojás közül vizsgálat céljára kiválasztunk 200 darabot, és úgy találjuk, hogy e 200 elemű mintában 10 darab nem felel meg a minőségi előírásoknak. Feladatunk megbecsülni azt, hogy az összes tojás között hány a nem megfelelő, más szóval megbecsülni annak a valószínűségét, hogy egy véletlenszerűen kiválasztott tojás nem megfelelő. (Ha ez a valószínűség p , és az összes tojások száma N , akkor a nem megfelelő tojások száma Np). Ezen ismeretlen p valószínűség becslésére előbbi módszerünket alkalmazhatjuk. Nevezetesen először számítsuk ki annak a valószínűségét, hogy a mintában levő selejtesek száma $200p - l$ és $200p + l$ között van. A mintában levő selejtesek számát k -val jelölve kapjuk, hogy ez a valószínűség:

$$P(200p - l \leq k \leq 200p + l) = \sum_{k=200p-l}^{200p+l} \binom{200}{k} p^k (1-p)^{200-k} \approx \\ \approx \Phi\left(\frac{l}{\sqrt{200p(1-p)}}\right) - \Phi\left(\frac{-l}{\sqrt{200p(1-p)}}\right) = 2\Phi\left(\frac{l}{\sqrt{200p(1-p)}}\right) - 1.$$

Megjegyzés. A binomiális eloszlás e helyütti alkalmazása csak abban az esetben jogosult, ha a 200-as mintát visszatevéssel választjuk ki az N tojás közül, azaz, ha a minta elemeit egyenként, véletlenszerűen húzzuk ki, és a kihúzott mintaelemet a vizsgálat után visszatesszük a többi tojás közé. Gyakorlatilag ritkán járunk így el, de abban az esetben, ha N jóval nagyobb, mint 200, akkor majdnem mindegy, hogy a mintavétel visszatevéssel vagy visszatevés nélkül történik.

Következő lépésként válasszuk meg az l számot úgy, hogy

$$P(200p - l \leq k \leq 200p + l) \approx 2\Phi\left(\frac{l}{\sqrt{200p(1-p)}}\right) - 1 = 0,9$$

legyen,

$$\text{azaz} \quad \Phi\left(\frac{l}{\sqrt{200p(1-p)}}\right) = 0,95$$

legyen. A Φ függvényre vonatkozó táblázatból kapjuk, hogy

$$\Phi(1,65) = 0,95,$$

$$\text{így} \quad \frac{l}{\sqrt{200p(1-p)}} = 1,65$$

$$\text{és} \quad l = 23,3 \sqrt{p(1-p)}.$$

Ez azt jelenti, hogy

$$P = (200p - 23,3 \sqrt{p(1-p)} \leq k \leq 200p + 23,3 \sqrt{p(1-p)}) = \\ = P\left(\frac{k - 23,3 \sqrt{p(1-p)}}{200} \leq p \leq \frac{k + 23,3 \sqrt{p(1-p)}}{200}\right) = 0,9.$$

$$\text{A} \quad \frac{k - 23,3 \sqrt{p(1-p)}}{200} \leq p \leq \frac{k + 23,3 \sqrt{p(1-p)}}{200}$$

egyenlőtlenséget megoldva kapjuk, hogy

$$(200p - k)^2 \leq (23,3)^2 p(1-p), \\ \text{azaz} \quad (4 \cdot 10^4 + (23,3)^2) p^2 - (400k + (23,3)^2) p + k^2 \leq 0, \\ \text{amiből} \quad \frac{400k + (23,3)^2 - 23,3 \sqrt{800k + (23,3)^2 - 4k^2}}{2(4 \cdot 10^4 + (23,3)^2)} \leq p \leq \\ \leq \frac{400k + (23,3)^2 + 23,3 \sqrt{800k + (23,3)^2 - 4k^2}}{2(4 \cdot 10^4 + (23,3)^2)}.$$

Vagyis, ha a 200 elemű mintában 10 darab hibás tojást találtunk, akkor $k = 10$ esetén

$$\frac{400k + (23,3)^2 - (23,3) \sqrt{800k + (23,3)^2 - 4k^2}}{2(4 \cdot 10^4 + (23,3)^2)} = 0,031 \\ \text{és} \quad \frac{400k + (23,3)^2 + (23,3) \sqrt{800k + (23,3)^2 - 4k^2}}{2(4 \cdot 10^4 + (23,3)^2)} = 0,083$$

szinte biztosan (0,9 valószínűséggel) $N \cdot 0,083$ és $N \cdot 0,031$ között lesz a hibások száma (az N tojás között).

E két példában láttuk, hogy statisztikai adatok segítségével egy ismeretlen valószínűségre alsó és felső becslés adható. Természetesen sohasem állítható, hogy a felső becslés biztosan nagyobb és az alsó becslés biztosan kisebb az ismeretlen valószínűségnél. Állíthatjuk azonban, hogy ez nagy valószínűséggel igaz, és előírhatjuk, hogy milyen biztonságot kívánunk meg. Első példánkban azt kívántuk, hogy 0,95 valószínűséggel legyen becslésünk helyes, második példánkban 0,9 volt az előírt biztonsági szint. Az, hogy milyen biztonsági szintet írunk elő, a körülményektől függ, nevezetesen attól, hogy mennyire veszélyes az, ha becslésünk nem helyes. Tudnunk kell azonban, hogy a biztonsági szint túl magasan történő megállapítása azzal a hátránnyal jár, hogy az alsó és felső becslés különbsége túl nagy lesz, és így igen kevés (bár igen nagy valószínűséggel helyes) információt nyerünk a valószínűségről.

Gyakran vetődik fel a kérdés úgy, hogy milyen nagyságú mintát kell vennünk ahhoz, hogy az ismeretlen valószínűségről adott biztonsággal, adott pontossággal becslést nyerjünk. Most oldjunk meg egy ilyen típusú feladatot. Térjünk vissza például első feladatunkhoz. Kérdésünk, hogy hány állatot kell beoltani ahhoz, hogy az elhullás valószínűségét 0,9 biztonsággal és 0,001 pontossággal becsülhessük. Más szóval azt kérdezzük, hogy hogyan kell megválasztani a beoltandó állatok számát, n -t, ha azt akarjuk elérni, hogy

$$P\left(\frac{k}{n} - 0,001 \leq p \leq \frac{k}{n} + 0,001\right) \geq 0,9$$

legyen, ahol k jelenti a beoltott állatok közül az elhullottak számát. Nyilván

$$P\left(\frac{k}{n} - 0,001 \leq p \leq \frac{k}{n} + 0,001\right) = P\left(p - 0,001 \leq \frac{k}{n} \leq p + 0,001\right) = \\ = P(np - n \cdot 0,001 \leq k \leq np + n \cdot 0,001) = \sum_{k=np-n \cdot 0,001}^{np+n \cdot 0,001} \binom{n}{k} p^k (1-p)^{n-k} \approx \\ \approx \Phi\left(\frac{0,001 \sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-0,001 \sqrt{n}}{\sqrt{p(1-p)}}\right) = 2\Phi\left(\frac{0,001 \sqrt{n}}{\sqrt{p(1-p)}}\right) - 1.$$

Most kísérjük meg n -t úgy választani, hogy

$$2\Phi\left(\frac{0,001\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 \cong 0,9,$$

azaz
$$\Phi\left(\frac{0,001\sqrt{n}}{\sqrt{p(1-p)}}\right) \cong 0,95$$

legyen. Mivel $\Phi(1,65) = 0,95$, ez azt jelenti, hogy

$$n \cong p(1-p) \left(\frac{1,65}{0,001}\right)^2.$$

A $p(1-p)$ szorzat értéke ismeretlen (hiszen éppen p értékét akarjuk becsülni), de könnyen belátható, hogy $p(1-p) \leq \frac{1}{4}$, akármilyen p értéke. Így a kívánt pontosságú becsléshez

$$n \cong \frac{1}{4} \left(\frac{1,65}{0,001}\right)^2 = 680\,625$$

kísérletre van szükség. Vagyis azt a teljesen képtelen eredményt kaptuk, hogy több mint fél-millió állatot kell beoltanunk, ha a kívánt pontosságú becslést akarjuk kapni. Ennek az az oka, hogy azt mondtuk: a p számról semmit sem tudunk. Ha például eleve tudjuk, hogy $p \cong 0,005$, akkor $p(1-p) \cong 0,005$, és így elegendő, ha

$$n \cong 13\,612.$$

Megjegyzés. Ha a p szám $\frac{1}{2}$ közelében lenne, akkor *ténylegesen* félmilliónál több kísérletre volna szükség, hogy a kívánt pontosságú becslést kapjuk. Persze ilyen p -k esetében általában nem kell 0,001 pontosságot megkívánnunk.

Ez a példa is mutatja, hogy a statisztika módszerei csak akkor hatásosak, ha megfelelő statisztikai adatok is állnak rendelkezésünkre. Ha a körülményekről semmilyen ismeretünk sincs, akkor a statisztikai módszerektől sem lehet semmilyen eredményt várni (vagy, mint ezt példánk mutatja, a kapható eredmény szinte semmit sem mond).

Itt kívánjuk megjegyezni, hogy abban az esetben, ha az ismeretlen valószínűség maga is a véletlentől függ, azaz valószínűségi változó, amelynek eloszlásáról bizonyos ismeretekkel rendelkezünk, akkor a fentiekől eltérő módszereket, az úgynevezett *Bayes*-módszert célszerű használni. Ez a helyzet például akkor, ha rendszeresen kapunk tojásslálmányokat, de a selejtesek száma szállítmányonként véletlenszerűen változik, és így a tapasztalatból tudjuk, hogy ez a szám általában milyen értékkel egyenlő. Természetesen ekkor is probléma, hogy az éppen érkező szállítmány milyen mennyiségű hibás tojást tartalmaz, de ekkor a szállítmányból vett mintából nyert információk kívül még a régebbi tapasztalatokból származó információkat is fel tudjuk használni.

15.2.1.2 Valószínűségekre vonatkozó hipotézisek

Időnként nem az a célunk, hogy a statisztikai adatokból egy ismeretlen valószínűségét becsüljük, hanem csak azt akarjuk eldönteni, hogy egy, a valószínűségekre vonatkozó hipotézis összeegyeztethető-e a statisztikai adatokkal, azaz, hogy a hipotézis elhíhető-e a statisztikai adatok ismeretében.

Például minőségellenőrzési vizsgálatoknál ritkán akarunk becslést adni a tételben lévő selejtszázalékra (azaz annak az eseménynek a valószínűségére, hogy a tétel egy véletlenszerűen kiválasztott eleme selejtes). Általában megelégszünk azzal, ha el tudjuk dönteni, hogy a selejtszázalék egy bizonyos megengedett szintet (például 5%) meghalad-e vagy sem. Más szóval a statisztikai vizsgálat célja nem a selejtszázalék becslése, hanem csak annak eldöntése, hogy az a hipotézis, miszerint a selejtszázalék kisebb 5%-nál, elhíhető-e a vett minta alapján.

Genetikai vizsgálatokban is általában az a helyzet, hogy van egy hipotézisünk arra vonatkozólag, hogy mi a valószínűsége annak, hogy egy keresztezésnél a születendő utód valamilyen tulajdonsággal rendelkezik. Ismeretes, hogy annak a valószínűsége, hogy az utódok egy adott tulajdonsággal rendelkeznek, nem lehet akármilyen p szám, ez a valószínűség csak bizonyos speciális értékekkel lehet egyenlő, például $\frac{1}{2}$, $\frac{1}{4}$, $\frac{5}{9}$... stb.

Példaként induljunk ki Mendel egyik kísérletéből. Mint tudjuk, Mendel többek között gömbölyű és szögletes magvú borsófajtákat keresztezett egymással. Itt a gömbölyű fajta a domináns tulajdonságú és a szögletes a recesszív tulajdonságú. Így az utódok mind gömbölyű magvúak lesznek. Az első hibridnemzedék (gömbölyű magvú) egyedeit egymás között keresztezve a második hibridnemzedék egyedei — a génelmélet szerint — $\frac{3}{4}$ valószínűséggel lesznek gömbölyű magvúak, és $\frac{1}{4}$ valószínűséggel szögletesek, azaz az utódoknak kb. $\frac{3}{4}$ -e lesz gömbölyű, és $\frac{1}{4}$ -e szögletes magvú.

Mendel eredeti kísérletében azt találta, hogy a második hibridnemzedék 7324 egyede közül 5474 gömbölyű és 1850 szögletes magvú, vagyis az utódok $\frac{5474}{7324} = 0,748$ -ed része gömbölyű magvú. Kérdés, vajon az az elméleti hipotézisünk, hogy a második hibridnemzedék egy egyede $\frac{3}{4}$ valószínűséggel gömbölyű magvú, összeegyeztethető-e a kapott kísérleti eredménnyel.

Ha egy hipotézis (például az utódok $\frac{3}{4}$ valószínűséggel gömbölyű magvúak) igaz vagy hamis voltáról akarunk dönteni, kétfajta hibától kell tartanunk.

Definíció. Az egyik hibalehetőség az, hogy a hipotézis helyes, és azt mégis elvetjük. Ezt a hibát elsőfajú hibának nevezzük. A második hibalehetőség az, hogy a hipotézis helytelen, és azt mégis elfogadjuk. Ezt a hibát másodfajú hibának nevezzük.

Természetesen nem várhatjuk, hogy olyan módszert találunk, amelynek alkalmazása esetén biztosak lehetünk abban, hogy nem követjük el az említett hibákat. Legfeljebb azt remélhetjük, hogy módszerünk segítségével a hibák elkövetésének valószínűsége kicsi.

Tegyük fel, hogy genetikai hipotézisünk helyes vagy hamis voltáról úgy akarunk dönteni, hogy az elsőfajú hiba elkövetésének valószínűsége ne haladja meg a 0,05-öt. (Arra, hogy a másodfajú hibáról ekkor mi mondható, a későbbiekben visszatérünk.)

Vizsgáljuk meg először, hogy mi mondható a relatív gyakoriságról, ha hipotézisünk helyes. Pontosabban, nézzük meg, hogy mely k számokra teljesül a

$$P\left\{\left|\frac{k}{7324} - \frac{3}{4}\right| \leq \varepsilon\right\} \leq 0,05$$

egyenlőtlenség, ahol k jelenti a második hibridnemzedék 7324 egyede között levő gömbölyű magvú egyedek számát.

A Moivre – Laplace-tétel alkalmazásával kapjuk, hogy

$$\begin{aligned} P\left\{\left|\frac{k}{7324} - \frac{3}{4}\right| \leq \varepsilon\right\} &= 1 - \sum_{k=\frac{3}{4} \cdot 7324 - \varepsilon \cdot 7324}^{\frac{3}{4} \cdot 7324 + \varepsilon \cdot 7324} \binom{7324}{k} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{7324-k} \approx \\ &\approx 1 - \left[\Phi\left(\frac{\varepsilon \cdot 7324}{\sqrt{7324 \cdot \frac{3}{4} \cdot \frac{1}{4}}}\right) - \Phi\left(\frac{-\varepsilon \cdot 7324}{\sqrt{7324 \cdot \frac{3}{4} \cdot \frac{1}{4}}}\right) \right] = \\ &= 2 - 2\Phi\left(\frac{4\varepsilon \sqrt{7324}}{\sqrt{3}}\right) \leq 0,05. \end{aligned}$$

$$\text{Így} \quad \Phi\left(\frac{4\varepsilon \sqrt{7324}}{\sqrt{3}}\right) \geq 0,975,$$

és ebből, a Φ függvény táblázatát használva, adódik, hogy

$$\begin{aligned} \frac{4\varepsilon \sqrt{7324}}{\sqrt{3}} &\geq 1,96, \\ \varepsilon &\geq \frac{\sqrt{3}}{4 \sqrt{7324}} 1,96 = 0,01. \end{aligned}$$

Vagyis azt kaptuk, hogy ha valóban $\frac{3}{4}$ valószínűséggel lesz a második hibridnemzedék egy egyede gömbölyű magvú, és ha k jelenti a második nemzedék 7324 egyede közül a gömbölyű magvúak számát, akkor

$$P\left\{\frac{3}{4} - 0,01 \leq \frac{k}{7324} \leq \frac{3}{4} + 0,01\right\} \geq 0,95.$$

Ezért, ha kísérletünk eredményeként k -ra olyan értéket kapunk, amelyre $\frac{k}{7324}$ nincs 0,74 és 0,76 között, akkor hipotézisünket elvetjük, hiszen ha hipotézisünk igaz lenne, akkor igen valószínűtlen volna, hogy $\frac{k}{7324}$ nincs az említett két határ között.

Természetesen előfordulhat, hogy annak ellenére, hogy hipotézisünk igaz, a $\frac{k}{7324}$ relatív gyakoriság nem esik a két határ közé. Ennek valószínűsége azonban 0,05. Más szóval 0,05 annak a valószínűsége, hogy a hipotézist elvetjük, bár az helyes, azaz az elsőfajú hiba elkövetésének valószínűsége 0,05.

Mivel Mendel kísérletében $k = 5474$ adódott, és így $\frac{k}{7324} = 0,748$ a kapott két határ között van, a hipotézis elfogadható.

Természetesen előfordulhat, hogy genetikai hipotézisünk helytelen, és csupán véletlenül adódott olyan k érték, amely a hipotézissel összeegyeztethető, vagyis elképzelhető, hogy másodfajú hibát követünk el. Persze, ha annak az igazi valószínűsége, hogy a második hibrid-

nemzedék egy egyede gömbölyű magvú, nem $\frac{3}{4}$, de közel van ehhez a számhoz, akkor a másodfajú hiba elkövetése igen könnyen lehetséges. Ha azonban az igazi valószínűség távol van $\frac{3}{4}$ -től, akkor a másodfajú hiba elkövetésének valószínűsége igen kicsi. Tegyük fel, hogy

a szóban forgó valószínűség igazi értéke p , és számítsuk ki, hogy e feltétel esetén mi a valószínűsége, hogy az eredeti hipotézist elfogadjuk. Mint láttuk, az eredeti hipotézist akkor fogadjuk el, ha

$$0,74 \leq \frac{k}{7324} \leq 0,76,$$

tehát azt kell meghatároznunk, hogy mi a valószínűsége ezen egyenlőtlenség teljesülésének, ha a kérdéses valószínűség igazi értéke p . Kapjuk, hogy

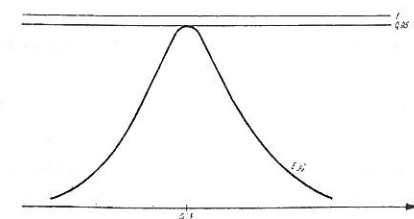
$$\begin{aligned} P\left\{0,74 \leq \frac{k}{7324} \leq 0,76\right\} &= P\{5419 \leq k \leq 5566\} = \\ &= \sum_{k=5419}^{5566} \binom{7324}{k} p^k (1-p)^{7324-k} \approx \Phi\left(\frac{5566-7324p}{\sqrt{7324p(1-p)}}\right) - \\ &- \Phi\left(\frac{5419-7324p}{\sqrt{7324p(1-p)}}\right), \end{aligned}$$

vagyis annak a valószínűsége, hogy hipotézisünket elfogadjuk, bár a valószínűség valódi értéke p :

$$E(p) = \Phi\left(\frac{5566-7324p}{\sqrt{7324p(1-p)}}\right) - \Phi\left(\frac{5419-7324p}{\sqrt{7324p(1-p)}}\right).$$

Az $E(p)$ függvény grafikonja világosan mutatja azt az egyébként is nyilvánvaló tény, hogy ha p értéke $\frac{3}{4}$ közelében van, akkor könnyen dönthetünk úgy, hogy a $p = \frac{3}{4}$ hipotézis helyes.

Gyakran találkozunk olyan valószínűségekre vonatkozó hipotézisekkel, amelyek nem egy esemény valószínűségére, hanem két különböző esemény valószínűségei közötti viszonyra vonatkoznak. Például lehet az a hipotézisünk, hogy két esemény ismeretlen valószínűségei egyenlőek. Képzeljük el, hogy két, lényegében egyező körülmények között levő állatállományt ugyanaz a betegség támad meg, de a két állatállománynál a betegség legyőzésére két különböző kezelési módot használnak. Tegyük fel, hogy az egyik



15.6 ábra

állomány 1000, a másik 2000 egyedből áll, és a betegség miatt az 1000 állatból 112, a 2000 állatból (viszonylag kevesebb) 220 állat hullik el. Kérdés, mondhatjuk-e, hogy a második kezelési mód eredményesebb, vagy azt kell mondanunk, hogy ilyen kis különbség esetén még teljesen elképzelhető, hogy a két kezelés egyformán hatásos, és az ilyen kis különbség véletlen tényezők folytán is felléphetett. Más szóval, ha p_1 jelenti annak a valószínűségét, hogy egy az első módon kezelt állat elhullik a betegség hatására, és p_2 annak a valószínűsége, hogy a második módon kezelt állat hullik el, akkor a $p_1 = p_2$ hipotézis helyes vagy helytelen voltáról kell döntenünk. E hipotézis helyes, illetve hamis voltának eldöntése valamivel bonyolultabb módszereket igényel, így ennek a kérdésnek a tárgyalására a későbbiekben térünk vissza. (Lásd t -próba.)

15.2.1.3 Megjegyzések az ismeretlen valószínűségek vizsgálatával kapcsolatban

Mind a valószínűség becslésénél, mind a valószínűségre vonatkozó hipotézisek ellenőrzésénél abból indultunk ki, hogy az ismeretlen valószínűségű eseményre vonatkozóan nagy számú kísérletet végzünk. Bár általában külön nem hangsúlyoztuk, mindig feltételeztük, hogy a kísérletek eredményei egymástól függetlenek, ugyanis csak ebben az esetben állíthatjuk, hogy az illető esemény gyakorisága binomiális eloszlású.

Ismeretett példáinkban a binomiális eloszlás segítségével kapott valószínűségeket a Moivre – Laplace-formula segítségével becsültük. Mint tudjuk, a Moivre – Laplace-formula alkalmazhatóságának bizonyos feltételei vannak. Példáinkban ellenőrizhetjük, hogy a formula alkalmazása indokolt volt, illetve megbecsülhetjük a formula alkalmazása során elkövetett hibát. A valószínűségszámítási fejezetben láttuk, hogy a binomiális eloszlás bizonyos esetekben nem a normális eloszlással, hanem a Poisson-eloszlással közelíthető. (Nevezetesen akkor, amikor a p valószínűség a kísérletek számához, n -hez képest kicsi. Például, ha $np \leq 5$.) Ilyen esetben természetesen módszerünket ennek megfelelően kell megváltoztatnunk.

Megjegyzés. Bizonyos esetekben a relatív gyakoriság eloszlásának vizsgálata helyett a relatív gyakoriság valamely függvényének (például $\arcsin \frac{k}{n}$ -nek) eloszlásfüggvényét vizsgáljuk. Kimutatható, hogy ez a módszer időnként a becslés pontosságát fokozni képes. (A módszer ismertetésére itt nem térünk ki.)

15.2.2 Ismeretlen várható érték és szórás

15.2.2.1 A várható érték és szórás becslése

Jelentsé ξ egy bizonyos korú csirkeállomány egy véletlenszerűen kiválasztott egyedének súlyát. (ξ értéke a véletlentől, nevezetesen a csirke kiválasztásától függ, így ξ valószínűségi változó.) Célunk a ξ valószínűségi változó $M(\xi)$ várható értékének és $D(\xi)$ szórásának becslése egy n elemű mintából. Azaz véletlenszerűen kiválasztunk n csirkét az állományból, ezeknek lemérjük a súlyát, jelentsé $\xi_1, \xi_2, \dots, \xi_n$ a mérés eredményeit, és kérdezzük, hogy

a $\xi_1, \xi_2, \dots, \xi_n$ értékek ismeretében mi mondható $M(\xi)$ -ről (a csirkeállomány átlagos súlyáról) és $D(\xi)$ -ről.

Mindenekelőtt jegyezzük meg, hogy a $\xi_1, \xi_2, \dots, \xi_n$ mennyiségek is tekinthetők valószínűségi változónak. (ξ_1 jelenti az elsőként kiválasztott csirke súlyát, és a kiválasztás előtt ez a súly még függ a véletlenszerű választástól.) Természetesen e változók eloszlásfüggvényei (és így szórásai továbbá várható értékei is) egyenlők. Abban az esetben, ha a mintavételnél a mintaelemeket egymástól függetlenül (visszatevéssel) választjuk, akkor a $\xi_1, \xi_2, \dots, \xi_n$ változók függetlenek. Így a nagy számok törvényeinek alkalmazásával adódik, hogy

$$(1) \quad \bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} \rightarrow M(\xi) \quad (n \rightarrow \infty),$$

$$(2) \quad \frac{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}{n} \rightarrow M(\xi^2) \quad (n \rightarrow \infty).$$

(Itt a $\rightarrow$ jel pontos jelentésére nem térünk ki. Számunkra jelenleg csak annyi a lényeges, hogy ha n nagy, akkor

$$\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}$$

nagy valószínűséggel közel van $M(\xi)$ -hez, és hasonlóan

$$\frac{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}{n}$$

nagy valószínűséggel közel van $M(\xi^2)$ -hez.)

(1)-ből és (2)-ből következik, hogy az ismeretlen $M(\xi)$ várható értéket lehet a

$$\bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}$$

értékkel, az ismeretlen

$$D(\xi) = \sqrt{M(\xi^2) - M^2(\xi)}$$

$$\text{szórását a } \sigma_n = \sqrt{\frac{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}{n} - \left(\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}\right)^2}$$

mennyiséggel becsülni.

A szórás becslésére általában a kissé különböző

$$s = \sqrt{\frac{n}{n-1}} \sigma_n$$

mennyiséget használják*.

* Belátható, hogy az s^2 valószínűségi változó várható értéke az ismeretlen σ^2 szórásnégyzet. Ezt úgy fejezzük ki, hogy s^2 torzítatlan becslése σ^2 -nek.

Definíció. Az s számot *tapasztalati szórásnak* nevezzük.

Természetesen itt is fontos volna tudnunk, hogy mi mondható e becslések pontosságáról, azaz felvetődik az a kérdés, hogy található-e olyan intervallum, amely legalábbis nagy valószínűséggel tartalmazza a keresett várható értéket és szórást. Erre a kérdésre sajnos nem adható pozitív válasz, hacsak nincs valami további információ a ξ valószínűségi változó eloszlásfüggvényéről. Azonban nagyon sok gyakorlati esetben (így például a súlyeloszlásnál is) feltehető, hogy ξ normális (vagy közel normális) eloszlású, és ekkor már kérdéseink megoldhatóak.*

Térjünk vissza a csirkeállomány átlagsúlyának vizsgálatára, de most már a további feltételezéssel, hogy a súlyeloszlás normális eloszlás. További egyszerűsítésként tegyük fel egyelőre, hogy ezen eloszlás szórását is ismerjük. Legyen ez σ , és célunk csak a várható érték becslése a statisztikai adatokból, azaz olyan intervallum megadása, amely nagy valószínűséggel tartalmazza az ismeretlen várható értéket.

Az ismeretlen várható értékét m -mel jelölve, és figyelembe véve, hogy $\bar{\xi}$ egy m várható értékű és $\frac{\sigma}{\sqrt{n}}$ szórású normális eloszlású valószínűségi változó, kapjuk, hogy

$$P\left\{\left|\frac{\bar{\xi}-m}{\frac{\sigma}{\sqrt{n}}}\right| \leq \varepsilon\right\} = P\left\{\bar{\xi}-\varepsilon\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{\xi}+\varepsilon\frac{\sigma}{\sqrt{n}}\right\} = 2\Phi(\varepsilon)-1.$$

Az ε értékét például 1,96-nak választva kapjuk, hogy

$$P\left\{\bar{\xi}-\varepsilon\frac{\sigma}{\sqrt{n}} \leq m \leq \bar{\xi}+\varepsilon\frac{\sigma}{\sqrt{n}}\right\} = 2\Phi(\varepsilon)-1 = 0,95,$$

vagyis a $\left(\bar{\xi}-\varepsilon\frac{\sigma}{\sqrt{n}}; \bar{\xi}+\varepsilon\frac{\sigma}{\sqrt{n}}\right)$ intervallum szinte biztosan (0,95 valószínűséggel) tartalmazza az ismeretlen m várható értéket. Jegyezzük meg, hogy ezen intervallum tényleges meghatározásához szükség van $\bar{\xi}$ értékére, vagyis a mintaátlagra, n -re, vagyis a minta elemszámára és σ -ra, ami a csirkék súlyának, ξ -nek a szórása. Ezen adatok közül egyedül σ ismerete okoz nehézséget. Általában σ -t is a statisztikai adatokból kell becsülnünk, de előfordulhat, hogy értékét eleve ismerjük, pl. ha egy gazdaság rendszeresen szállít csirkéket, akkor könnyen feltehető, hogy az egyes szállítmányok szórásai lényegében megegyeznek, és csak a szállítmány egyedeinek átlagsúlya, ami szállítmányonként változik. Ekkor σ értékét az előzőleg megvizsgált szállítmányok adataiból ismerhetjük, és célunk csak a konkrét szállítmány átlagsúlyának meghatározása a minta alapján, felhasználva a szórásnak az előzőekből ismert értékét.

Ezután rátérünk arra a kérdésre, hogy mi mondható az ismeretlen várható értékről, ha a szórást nem ismerjük eleve, de feltételezzük, hogy a változó normális eloszlású. Mint láttuk,

* Megjegyezzük, hogy ha ξ eloszlásfüggvényének típusát nem ismerjük, de tudjuk, hogy ξ lehetséges értékei bizonyos ismert korlátok között vannak, akkor is lehetséges a statisztikai adatok felhasználásával olyan intervallum megadása, amely nagy valószínűséggel tartalmazza az ismeretlen várható értéket, illetve szórást.

az ismert szórás esetében módszerünk azon alapult, hogy a $\frac{\bar{\xi}-m}{\frac{\sigma}{\sqrt{n}}}$ valószínűségi változó

0 várható értékű, 1 szórású, normális eloszlású. Természetesnek látszik ismeretlen szórás esetében e valószínűségi változó helyett a

$$t_n = \frac{\bar{\xi}-m}{\frac{s}{\sqrt{n}}}$$

valószínűségi változót vizsgálni, ahol

$$s = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (\xi_k - \bar{\xi})^2},$$

azaz a σ szórást annak s becslésével helyettesíteni. Ahhoz, hogy m -re becslést kaphassunk, tudnunk kellene, hogy t_n milyen valószínűséggel vesz fel nagy értéket, azaz a $P(|t_n| > z)$ valószínűségeket kellene ismernünk. Ugyanis, ha tudunk egy olyan z értéket találni, amelyre

$$0,95 = P(|t_n| \leq z) = P\left(\bar{\xi}-z\frac{s}{\sqrt{n}} \leq m \leq \bar{\xi}+z\frac{s}{\sqrt{n}}\right),$$

akkor azt mondhatjuk, hogy $\bar{\xi}-z\frac{s}{\sqrt{n}}$ az alsó és $\bar{\xi}+z\frac{s}{\sqrt{n}}$ a felső becslése a várható értéknek.

A nehézséget az okozza, hogy a t_n valószínűségi változó eloszlása nem normális, hanem jóval bonyolultabb. Bebizonyítható, hogy t_n sűrűségfüggvénye

$$t_n(x) = \frac{1}{\sqrt{\pi f}} \frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)} \left(1 + \frac{x^2}{f}\right)^{-\frac{f+1}{2}},$$

ahol $f = n-1$, és $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ ($\alpha > 0$).

Gyakorlati használat céljaira a $T_n(z) = \int_{-z}^z t_n(x) dx = P(|t_n| \leq z)$ függvény értékeinek táblázatait adják meg.

Most vizsgáljunk meg egy konkrét példát.

Egy tejtermeléssel foglalkozó gazdaságban 20 alkalommal megvizsgáljuk a tejet zsírtartalom*

* A zsírtartalom fontos jellemzője a tej minőségének. Az átvétőhelyek általában e szerint fizetnek. A zsírtartalom rendszeres mérése nagy mennyiségű tej esetén azonban komoly feladatot jelent. Ezért fontos kérdés, hogy csak időnkénti ellenőrzőméréseket alkalmazva, milyen pontossággal becsülhető a zsírtartalom.

szerint. Tegyük fel, hogy eredményünk a következő:

nap	zsírszázalék	nap	zsírszázalék
1.	3,0	11.	3,2
2.	3,1	12.	3,4
3.	3,0	13.	3,6
4.	3,2	14.	3,5
5.	3,3	15.	3,7
6.	3,2	16.	3,2
7.	3,4	17.	3,3
8.	3,3	18.	3,4
9.	3,5	19.	3,3
10.	3,3	20.	3,5

Becsüljük meg a kapott adatok alapján a zsírszázalék várható értékét. Jegyezzük meg, hogy az egyes tehenek, és így az egész állomány tejhozama, de a termelt tej zsírszázaléka is változó, például éves periodicitást mutat. Feltételezhető azonban, hogy míg a zsírszázaléknak (és a tej mennyiségnek) a várható értéke változik, a szórása változatlan, és így azt előző mérésekből már ismerjük. Most a szórás ismerete nélkül kíséreljük meg a várható értéket becsülni.

A T -táblázat alkalmazásával* kapjuk, hogy

$$T_{20}(2,093) = 0,95.$$

$$\text{Így } P\left(\left|\frac{\bar{\xi} - m}{\frac{s}{\sqrt{20}}}\right| \leq 2,093\right) = P\left(\bar{\xi} - 2,093 \frac{s}{\sqrt{20}} \leq m \leq \bar{\xi} + 2,093 \frac{s}{\sqrt{20}}\right) = 0,95.$$

Ennek megfelelően azt mondhatjuk, hogy $\bar{\xi} - 2,093 \frac{s}{\sqrt{20}}$ szinte biztosan (0,95 valószínűséggel) kisebb az ismeretlen várható értéknél, és $\bar{\xi} + 2,093 \frac{s}{\sqrt{20}}$ szinte biztosan nagyobb annál.

Esetünkben adataink átlaga

$$\bar{\xi} = 3,32,$$

tapasztalati szórása:

$$s = 0,185.$$

Így m -re a 3,23 alsó és a 3,41 felső becslést kaptuk.

Eredményünk gyakorlati alkalmazása szempontjából az a kérdés is lényeges, hogy a kapott tej mennyisége és zsírszázaléka között van-e valami kapcsolat. Elképzelhető ugyanis, hogy olyan napokon, amikor a termelt tej mennyisége nagy, akkor általában a zsírszázalék viszonylag kisebb. Ha ez valóban így van, akkor csupán az átlagos zsírszázalékból és a napi tej mennyiségéből nem lehet következtetéseket levonni a napi tej zsírmennyiségére. Ennek a kérdésnek a vizsgálatára a későbbiekben még majd visszatérünk.

* Természetesen feltételeztük, hogy a zsírmennyiség normális eloszlású, ami általában teljesül.

Végül röviden ismertetjük a szórás becslésének módszerét normális eloszlású valószínűségi változók esetében.

Mint említettük, az

$$s = \sqrt{\frac{n}{n-1} \sigma_n^*}; \quad \left(\sigma_n^* = \sqrt{\frac{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}{n} - \left(\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}\right)^2} \right)$$

mennyiséget használhatjuk a szórás becslésére. Utolsó példánkban (a tej zsírszázalékával kapcsolatban) ki is számítottuk, hogy a tapasztalati szórás értéke, $s = 0,185$.

Most azt kívánjuk vizsgálni, hogy mi mondható e becslés pontosságáról, azaz azt, hogy találhatunk-e olyan alsó, illetve felső becslést, amely nagy valószínűséggel jó. Ehhez az s valószínűségi változó sűrűségfüggvényére lenne szükségünk. Ehelyett a $\chi_n^2 = \frac{s^2(n-1)}{\sigma^2}$ valószínűségi változó sűrűségfüggvényét adjuk meg, ahol σ jelenti a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók (ismeretlen) szórását. Bebizonyítható, hogy a χ_n^2 valószínűségi változó sűrűségfüggvénye

$$\chi_n^2(t) = \frac{1}{2^{\frac{f}{2}}} \frac{1}{\Gamma\left(\frac{f}{2}\right)} t^{f-2} e^{-\frac{t^2}{2}},$$

ahol $f = n-1$.

Gyakorlati szempontból ismét kizárólag a χ_n^2 valószínűségi változó eloszlásfüggvényének, azaz a

$$\int_0^z \chi_n^2(t) dt$$

függvénynek a táblázatát használják.

A táblázatból például azt kapjuk, hogy

$$P(\chi_{20}^2 \leq 30,144) = P\left(\frac{s^2 \cdot 19}{\sigma^2} \leq 30,144\right) = 0,95,$$

$$\text{és } P(\chi_{20}^2 \leq 10,117) = P\left(\frac{s^2 \cdot 19}{\sigma^2} \leq 10,117\right) = 0,95,$$

vagyis σ^2 -nek (a szórásnégyzetnek) $\frac{s^2 \cdot 19}{30,144}$ szinte biztosan (0,95 valószínűséggel) az alsó becslése és $\frac{s^2 \cdot 19}{10,117}$ szinte biztosan a felső becslése. Esetünkben

$$\frac{s^2 \cdot 19}{10,117} = 0,064, \quad \frac{s^2 \cdot 19}{30,144} = 0,021.$$

15.2.2.2 Az átlag és a tapasztalati szórás kiszámításának módszerei

Bár az átlag és a tapasztalati szórás meghatározása semmi elvi nehézséget nem okoz, mégis, ha sok adatról van szó, akkor konkrét kiszámításuk jelentős munkát igényel. Az átlag kiszámítása egyszerűbb, ha a szereplő adatok általában kicsinyek, illetve viszonylag

15.2.2.3 A várható értékre és a szórásra vonatkozó hipotézisek

Az ismeretlen valószínűségek vizsgálatánál láttuk, hogy sok esetben nem tartjuk fontosnak becslés megadását, megelégszünk annak eldöntésével, hogy az ismeretlen érték eleget tesz-e bizonyos kikötéseknek. Hasonlót mondhatunk abban az esetben, ha egy valószínűségi változó várható értéke vagy szórása az ismeretlen. A zsírszázalékra vonatkozó példánkkal kapcsolatban is előfordulhat, hogy nem akarunk a várható értékre becslést adni, hanem csak azt akarjuk eldönteni, hogy az átlagos zsírszázalék elér-e bizonyos adott értéket. Annak a kérdésnek eldöntése, hogy a kapott adatokkal összeegyeztethető-e például az a hipotézis, hogy a zsírszázalék elér-e a 4%-ot, nagyon hasonló az előző pontban ismertetett feladat megoldásához. Ezért most egy kissé eltérő típusú feladatot oldunk meg.

Célunk két gazdaság tehénállományának összehasonlítása tejhozamuk zsírtartalma szempontjából. Tegyük fel, hogy az első gazdaság 250, a második pedig 200 tehénnel rendelkezik, és egy alkalommal lemérve a termelt tejben levő zsírmennyiséget, a következő adatokat kapjuk:

I. gazdaság		II. gazdaság	
zsír kp	tehenek száma	zsír kp	tehenek száma
0,28—0,30	6	0,28—0,30	10
0,30—0,32	9	0,30—0,32	21
0,32—0,34	21	0,32—0,34	28
0,34—0,36	39	0,34—0,36	35
0,36—0,38	51	0,36—0,38	35
0,38—0,40	49	0,38—0,40	30
0,40—0,42	39	0,40—0,42	20
0,42—0,44	20	0,42—0,44	15
0,44—0,46	11	0,44—0,46	5
0,46—0,48	5	0,46—0,48	1

A kérdést nehezíti, hogy adataink nem pontosak, azaz nem tudjuk az egyes tehenek által termelt tej pontos zsírtartalmát, hanem csak 0,02 pontossággal megadott adatokkal rendelkezünk. Az egyszerűség kedvéért úgy számolunk, mintha a zsírmennyiségek mind a megadott intervallumok középpontjaiba esnének*, azaz táblázatunkat a következővel helyettesítjük:

I. gazdaság		II. gazdaság	
zsír kp	tehenek száma	zsír kp	tehenek száma
0,29	6	0,29	10
0,31	9	0,31	21
0,33	21	0,33	28
0,35	39	0,35	35
0,37	51	0,37	35
0,39	49	0,39	30
0,41	39	0,41	20
0,43	20	0,43	15
0,45	11	0,45	5
0,47	5	0,47	1

* Ezzel természetesen hibát követünk el. A hiba nagyságának becslésére és kiküszöbölésének lehetőségére a sűrűségfüggvény becslésével kapcsolatban visszatérünk.

E táblázatokból azt kapjuk, hogy az első gazdaságban a zsír súlyátalaga [kp]

$$\bar{\xi}_1 = 0,380,$$

a második gazdaságban pedig

$$\bar{\xi}_2 = 0,365.$$

Kérdésünk, mondhatjuk-e ennek alapján, hogy az első gazdaság tehénállománya jobb azaz az egy tehén által termelt tejzsír mennyisége nagyobb, mint a második gazdaságban, vagy azt kell mondanunk, hogy ilyen kis különbség alapján még ezt nem állíthatjuk. Más szóval, az $m_1 = m_2$ hipotézist akarjuk ellenőrizni, ahol m_1 , ill. m_2 az első, ill. második gazdaság egy tehene által adott zsírmennyiség várható értékét jelenti. Tegyük fel, hogy úgy akarunk dönteni, hogy az elsőfajú hiba (vagyis az igaz hipotézis elvetésének) valószínűsége 0,05 legyen.

Abban az esetben, ha a szórásokat az előző adatokból ismerjük (és ezek σ_1 és σ_2), akkor hipotézisünk helyes vagy hamis volta felől igen egyszerűen dönthetünk. Ugyanis, mint ez könnyen látható*, ha hipotézisünk helyes, akkor

$$u = \frac{\bar{\xi}_1 - \bar{\xi}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

0 várható értékű és 1 szórású normális eloszlású valószínűségi változó. Így a normális eloszlás táblázatát felhasználva, σ_1 és σ_2 értékeinek ismeretében könnyen eldönthetjük, hogy hipotézisünket elfogadhatjuk-e.

Abban az esetben, ha a szórásokat eleve nem ismerjük, akkor ezeket is a kapott adatokból kell becsülnünk, azaz ebben az esetben a

$$t = \frac{\bar{\xi}_1 - \bar{\xi}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (n_1 = 250, n_2 = 200)$$

valószínűségi változót kell vizsgálnunk, ahol

$$s_1 = \sqrt{\frac{n_1}{n_1 - 1} \left[\frac{1}{n_1} \sum_{k=1}^{n_1} \xi_{1k}^2 - \left(\frac{1}{n_1} \sum_{k=1}^{n_1} \xi_{1k} \right)^2 \right]},$$

$$s_2 = \sqrt{\frac{n_2}{n_2 - 1} \left[\frac{1}{n_2} \sum_{k=1}^{n_2} \xi_{2k}^2 - \left(\frac{1}{n_2} \sum_{k=1}^{n_2} \xi_{2k} \right)^2 \right]},$$

és $\xi_{11}, \xi_{12}, \dots, \xi_{1n_1}$ az első, $\xi_{21}, \xi_{22}, \dots, \xi_{2n_2}$ a második gazdaság tehenei által adott zsírmennyiségek.

Ha a két minta szórása megegyezik, akkor a t valószínűségi változó eloszlása ismét az előző pontban megismert t_f eloszlás, $f = n_1 + n_2 - 2$ paraméterrel. A t táblázat alkalmazásával azt kapjuk, hogy (0,05 szinten) hipotézisünket el kell vetnünk, ha t abszolút értéke-

* Természetesen a zsírmennyiségről most is feltételezzük, hogy normális eloszlású.

ként 1,96-nál nagyobb szám adódik, míg ha $|t|$ értéke ennél kisebb, a hipotézis elfogadható. Esetünkben

$$\begin{aligned}\bar{\xi}_1 &= 0,380, & \bar{\xi}_2 &= 0,365, \\ s_1^2 &= 0,002, \\ s_2^2 &= 0,0015,\end{aligned}$$

és így

$$t = 2,25.$$

Tehát $m_1 = m_2$ hipotézisünket 0,05 szinten nem fogadhatjuk el, vagyis a kapott különbség alapján már állíthatjuk, hogy az első gazdaság tehénállománya jobb. A szórásra vonatkozó hipotézisek vizsgálatára itt sem térünk ki, csak azt jegyezzük meg, hogy a χ^2 -eloszlás alkalmazásával hasonlóan történhet, mint a szórás becslése.

15.2.3 A becsléelmélet általános kérdései

15.2.3.1 A becsléelméletben szereplő fogalmak

Legyenek $\xi_1, \xi_2, \dots, \xi_n$ független egyforma eloszlású valószínűségi változók. Jelentsé $F(x)$ a $\xi_1, \xi_2, \dots, \xi_n$ változók közös eloszlásfüggvényét. Kérdésünk, hogy $\xi_1, \xi_2, \dots, \xi_n$ értékeinek ismeretében mi mondható az $F(x)$ eloszlásfüggvényről, illetve annak paramétereiről. Más szóval azt kérdezzük, hogy a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók milyen $\eta_n = t_n(\xi_1, \xi_2, \dots, \xi_n)$ függvénye alkalmas az eloszlás valamely paraméterének becslésére. A $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók egy $\eta_n = t_n(\xi_1, \xi_2, \dots, \xi_n)$ függvényét *statisztikának* (a változók egy statisztikájának) nevezzük. Általában — mint azt példánkban láttuk — $\xi_1, \xi_2, \dots, \xi_n$ kísérletek, a mérések eredményei, amelyek felhasználásával az $F(x)$ eloszlás várható értéke, szórása vagy más paramétere becsülhető. A várható értéket például a

$$\eta_n = t_n(\xi_1, \xi_2, \dots, \xi_n) = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}$$

statisztikával becsültük.

Most vizsgáljuk meg, mikor mondhatjuk, hogy egy $t_n(\xi_1, \xi_2, \dots, \xi_n)$ statisztika az $F(x)$ eloszlásfüggvény valamely paraméterének jó becslése. Az első követelmény, amit támasztani szoktunk az, hogy a statisztika várható értéke meg egyezzen a becsülendő a paraméterrel, azaz hogy

$$M(t_n(\xi_1, \xi_2, \dots, \xi_n)) = a$$

legyen.

Definíció. Abban az esetben, ha ez a kikötés teljesül, az mondjuk, hogy a t_n statisztika az a paraméternek *torzítatlan becslése*. Mint láttuk, az

$$\eta_n = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}$$

átlag torzítatlan becslése a várható értéknek, hasonlóan a

$$s_n^2 = \frac{\xi_1^2 + \xi_2^2 + \dots + \xi_n^2}{n-1} - \left(\frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} \right)^2 \frac{n}{n-1}$$

statisztika a szórásnégyzetnek torzítatlan becslése. Jegyezzük meg, hogy a

$$\sigma_n = \sqrt{\sigma_n^2}$$

vagy az

$$s_n = \sqrt{\frac{n}{n-1} \sigma_n^2}$$

statisztika általában nem torzítatlan becslése a szórásnak. Ennek ellenére e becsléseket is gyakran alkalmazzuk.

A második követelmény, amit becslésekkel szemben támasztani szoktunk az, hogy ha a kísérletek száma, n nagy, akkor a t_n statisztika valóban közel legyen az a paraméterhez. Természetesen különböző módon definiálhatjuk azt, hogy mit értünk azon, hogy egy t_n statisztika (valószínűségi változó) közel van a -hoz. Egy lehetséges kikötés az, hogy azt tesszük fel, hogy $|t_n - a|$ nagy valószínűséggel kicsi, ha n elég nagy, azaz, hogy tetszőleges $\varepsilon > 0$ -hoz és $\delta > 0$ -hoz található olyan $n_0 = n_0(\varepsilon, \delta)$, hogy

$$P(|t_n - a| \leq \varepsilon) \geq 1 - \delta,$$

ha

$$n \geq n_0.$$

Definíció. Abban az esetben, ha ez a feltétel teljesül, akkor azt mondjuk, hogy a t_n statisztika konzisztens becslése az a paraméternek.

Például a nagy számok törvényei szerint az átlag és a σ_n^2 statisztika konzisztens becslései a várható értéknek és a szórásnégyzetnek (ha ezek léteznek). Mint említettük, σ_n és s_n általában nem torzítatlan becslései a szórásnak, de könnyen belátható, hogy konzisztens becslések. Ez a körülmény indokolja, hogy a σ_n , illetve az s_n statisztikák használhatók a szórás becsléseként. Annak a követelménynek, hogy t_n nagy n esetén legyen közel a -hoz, egy másik lehetséges megfogalmazása, hogy $t_n - a$ négyzetének várható értéke legyen kicsi. Azaz, hogy

$$M[(t_n - a)^2] \rightarrow 0 \quad (n \rightarrow \infty).$$

Definíció. Ha ez a követelmény teljesül, akkor azt mondjuk, hogy a t_n statisztika erősen konzisztens becslése a -nak.

Ugyancsak a nagy számok törvényeiből következik, hogy az átlag erősen konzisztens becslése a várható értéknek, ha a szórás létezik, és a σ_n^2 statisztika erősen konzisztens becslése a szórásnégyzetnek, ha a negyedik momentum is létezik.

Egy ismeretlen paramétert általában különböző statisztikákkal becsülhetünk, sőt általában több torzítatlan erősen konzisztens becslést találhatunk. Fontos kérdés, hogy a lehetséges statisztikák közül melyik a legjobb. Különböző szempontokat vehetünk figyelembe annak eldöntésekor, hogy két becslés közül melyik a jobb, de általában két torzítatlan becslés közül azt tekintjük a jobbiknak, amelyiknek a szórása kisebb. Vagyis, ha $t_n(\xi_1, \xi_2, \dots, \xi_n)$

valamint $u_n(\xi_1, \xi_2, \dots, \xi_n)$ két torzítatlan becslése egy paraméternek,

$$\text{és} \quad M[(t_n - a)^2] = D^2(t_n) < D^2(u_n) = M[(u_n - a)^2],$$

akkor azt mondjuk, hogy t_n jobb becslés, mint u_n .

Bebizonyítható, hogy igen általános feltételek mellett megadható a torzítatlan becslések szórásának egy alsó korlátja.

Tétel. Nevezetesen, ha $f(x, a)$ egy sűrűségfüggvény és $t_n(\xi_1, \xi_2, \dots, \xi_n)$ egy torzítatlan becslés* az a paraméternek,

$$\text{akkor} \quad D^2(t_n) \geq \frac{1}{nM \left[\left(\frac{d \ln f(\xi, a)}{da} \right)^2 \right]}.$$

Ezen egyenlőtlenséget *Cramer—Rao-egyenlőtlenségnek* nevezzük. Azt mondhatjuk tehát, hogy az a legjobb torzítatlan becslés, amelynek szórásnégyzete megegyezik a Cramer—Rao-egyenlőtlenség jobb oldalán álló mennyiséggel. Az olyan torzítatlan becslést, amelyre ez teljesül, *efficiens becslésnek* nevezzük.

Definíció. Egy *torzítatlan becslés hatásfokán* az illető becslés szórásnégyzetének és ezen minimális szórásnégyzetnek a hányadosát értjük, tehát egy $t_n(\xi_1, \xi_2, \dots, \xi_n)$ torzítatlan becslés hatásfoka a

$$\frac{D^2(t_n)}{nM \left[\left(\frac{d \ln f(\xi, a)}{da} \right)^2 \right]}$$

hányados.

Statisztikákkal kapcsolatban más követelményeket is szoktak támasztani, de ezeket itt nem tárgyaljuk.

Most áttérünk annak a kérdésnek a vizsgálatára, hogy mi mondható egy becslés pontosságáról. Ahhoz, hogy erre a kérdésre valami választ adhassunk, ismernünk kell a szóban forgó eloszlásfüggvény alakját, vagy legalább tudnunk kell valamit az eloszlásfüggvény típusáról. Így kérdésünket a következő módon fogalmazhatjuk.

Legyen $F(x; a_1, a_2, \dots, a_k)$ egy eloszlásfüggvény, ahol $a_1, a_2, \dots, a_k$ ismeretlen paraméterek. Továbbá legyenek $\xi_1, \xi_2, \dots, \xi_n$ független valószínűségi változók közös $F(x; a_1, a_2, \dots, a_k)$ eloszlásfüggvénnyel. Tegyük fel, hogy az ismeretlen paraméterek egyikét, például a_1 -et akarjuk becsülni a $\xi_1, \xi_2, \dots, \xi_n$ statisztikai adatok segítségével. Pontosabban, keresünk egy olyan intervallumot (alsó és felső becslést), amely nagy valószínűséggel (általában előírt valószínűséggel) tartalmazza az ismeretlen a_1 paramétert. Más szóval, keresünk egy $u_n(\xi_1, \xi_2, \dots, \xi_n)$ és egy $v_n(\xi_1, \xi_2, \dots, \xi_n)$ statisztikát úgy, hogy teljesüljön a

$$P[u_n(\xi_1, \xi_2, \dots, \xi_n) \leq a_1 \leq v_n(\xi_1, \xi_2, \dots, \xi_n)] \geq p$$

egyenlőtlenség, ahol p adott valószínűség. (Például 0,9, 0,95, 0,99 stb.)

* Itt $\xi_1, \xi_2, \dots, \xi_n$ független $f(x, a)$ sűrűségfüggvényű valószínűségi változók.

Definíció. Az $[u_n, v_n]$ intervallumot *konfidenciaintervallumnak* (az a paraméter konfidencia-intervallumának) nevezzük. A konfidenciaintervallum nyilván függ a p számtól is. Ha p nagy, akkor a konfidenciaintervallum hossza is nagy. p értékének megválasztása a konkrét körülményektől függ, vagyis attól, hogy mennyire veszélyes, ha ítéletünk hibás, azaz a konfidenciaintervallum nem tartalmazza az ismeretlen paramétert.

15.2.3.2 Maximum likelihood módszer

Az előzőekben láttuk, hogy a várható érték és a szórás becslésére igen természetes statisztikákat találhatunk, nevezetesen az átlagot és a tapasztalati szórást. Gyakran előfordul, hogy az eloszlás- vagy sűrűségfüggvénynek egy más paraméterét kell becsülnünk statisztikai adatokból. Más szóval problémánk a következő: legyen $f(x, a)$ egy sűrűségfüggvény és legyenek $\xi_1, \xi_2, \dots, \xi_n$ független $f(x, a)$ sűrűségfüggvényű valószínűségi változók. Kérdés, hogy hogyan található egy, az a paramétert becslő $t_n(\xi_1, \xi_2, \dots, \xi_n)$ statisztika.

Erre a kérdésre ad egy választ a *maximum likelihood módszer*. Számítsuk ki először annak valószínűségét, hogy első kísérletünk eredménye x_1 és $x_1 + \Delta x$ közé esik, adódik, hogy

$$P(x_1 \leq \xi_1 < x_1 + \Delta x) = \int_{x_1}^{x_1 + \Delta x} f(t, a) dt \approx f(x_1, a) \Delta x,$$

ha Δx elég kicsi.

Hasonló módon, annak a valószínűsége, hogy második kísérletünk eredménye x_2 és $x_2 + \Delta x$ közé essék, kis Δx esetén $f(x_2, a) \Delta x$ -szel becsülhető. Mivel a kísérleteket egymástól függetlenül végezzük,

$$P(x_i \leq \xi_i < x_i + \Delta x; i = 1, 2, \dots, n) \approx f(x_1, a) f(x_2, a) \dots f(x_n, a) (\Delta x)^n.$$

Ezután a következőképpen okoskodhatunk. Általában kísérletünk eredményeként nem adódnak olyan értékek, amelyek igen valószínűtlenek, vagyis kísérletünk eredményei általában olyan $x_1, x_2, \dots, x_n$ számok lesznek, amelyekre az

$$f(x_1, a) f(x_2, a) \dots f(x_n, a) (\Delta x)^n$$

valószínűség, illetve az

$$f(x_1, a) f(x_2, a) \dots f(x_n, a)$$

függvény nagy. Ennek az elvnek megfelelően a maximum likelihood módszer (a maximális valószínűség módszere) abban áll, hogy az a paramétert úgy választjuk meg, hogy az

$$f(x_1, a) f(x_2, a) \dots f(x_n, a)$$

mennyiség a lehető legnagyobb legyen. Az a megválasztása természetesen függ a kísérletek eredményétől, azaz az $x_1, x_2, \dots, x_n$ számoktól, más szóval becsléseként a $\xi_1 = x_1, \xi_2 = x_2, \dots, \xi_n = x_n$ valószínűségi változóknak egy statisztikája adódik.

Az $f(x_1, a) f(x_2, a) \dots f(x_n, a)$

függvény maximumának megkeresése helyett általában egyszerűbb

$$L(a) = \ln (f(x_1, a) f(x_2, a) \dots f(x_n, a))$$

maximumát megkeresni. (E két függvény maximumhelye nyilván egybeesik.) Ugyanis e függvény maximumának megtalálása a

$$\frac{dL(a)}{da} = \frac{d}{da} \sum_{i=1}^n \ln f(x_i, a) = \sum_{i=1}^n \frac{d}{da} \ln f(x_i, a) = 0$$

egyenlet megoldására vezet. (Természetesen be kell látni, hogy a differenciálhányados 0-helyén az $L(a)$ függvénynek valóban maximuma van, de erre a kérdésre itt nem térünk ki.)

Megjegyzés. A maximum likelihood módszer igen általános feltételek mellett több szempontból a legjobb becslést szolgáltatja.

A módszer megvilágítása céljából most két példát vizsgálunk meg. Legyen

$$f(x, m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2}},$$

vagyis $f(x, m)$ egy $\sigma = 1$ szórású és ismeretlen m várható értékű normális eloszlású valószínűségi változó sűrűségfüggvénye. Az m paramétert a maximum likelihood módszer segítségével oly módon becsülhetjük, hogy maximalizáljuk az

$$L(m) = \sum_{i=1}^n \ln \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-m)^2}{2}}$$

függvényt, azaz megoldjuk a

$$\frac{d}{dm} L(m) = - \sum_{i=1}^n \frac{d}{dm} (x_i - m)^2 = 2 \sum_{i=1}^n (x_i - m) = 2 \left[\sum_{i=1}^n x_i - nm \right] = 0$$

egyenletet. Ennek megoldásaként azt kapjuk, hogy

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

Így az m paraméter maximum likelihood becsléseként az eddig is használt $\frac{\sum_{i=1}^n x_i}{n}$ átlag adódik.

Második példaként tekintsünk egy $[a; b]$ intervallumban egyenletes eloszlású $f(x; a, b)$ sűrűségfüggvényt, azaz legyen

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & \text{ha } a \leq x \leq b, \\ 0 & \text{különben,} \end{cases}$$

ahol a és b ismeretlen paraméterek. Ekkor

$$(x_1; a, b) f(x_2; a, b) \dots f(x_n; a, b) = \begin{cases} \frac{1}{(b-a)^n}, & \text{ha } a \leq x_i \leq b \quad (i = 1, 2, \dots, n) \\ 0 & \text{különben.} \end{cases}$$

Így $f(x_1; a, b) f(x_2; a, b) \dots f(x_n; a, b)$

legnagyobb értékét akkor veszi fel, ha a -t és b -t úgy választjuk meg, hogy $\frac{1}{(b-a)^n}$ olyan nagy legyen, mint csak lehetséges, de az összes statisztikai adat (az $x_1, x_2, \dots, x_n$ számok) a és b közé essék. Ez más szóval azt jelenti, hogy a -t a statisztikai adatok legkisebbjének, b -t ezen adatok legnagyobbikának kell választanunk. Vagyis az a paraméter maximum likelihood becslése az

$$s_n = \min \{x_1, x_2, \dots, x_n\}$$

statisztika és b_n maximuma likelihood becslése a

$$t_n = \max \{x_1, x_2, \dots, x_n\}$$

statisztika.

Ebből az eredményből az egyenletes eloszlású valószínűségi változó várható értékének becslésére is érdekes következtetést lehet levonni, ugyanis mint tudjuk, az $[a; b]$ intervallumban egyenletes eloszlású változó várható értéke $\frac{a+b}{2}$. Így a várható érték becslésére az $\frac{s_n + t_n}{2}$ statisztikát használhatjuk, ami általában különbözik a számtani középtől. Megmutatható, hogy egyenletes eloszlás esetén ez a statisztika több szempontból jobb becslése a várható értéknek, mint az átlag (például a szórása kisebb).

15.2.4 A hipotézisvizsgálat általános kérdései

15.2.4.1 A hipotézisvizsgálatban szereplő általános fogalmak

Mint láttuk, a becslésméletben egy (vagy több) ismeretlen mennyiség (szám) értékét (nagyságát) akartuk meghatározni statisztikai adatok alapján. A hipotézisvizsgálatban olyan kérdésekkel foglalkozunk, amelyekre egyszerű igen vagy nem választ adhatunk (válaszunk ismét statisztikai adatok vizsgálatán alapszik). Ilyen kérdések például:

- Egy eloszlásfüggvény valamely paramétere (pl. várható értéke, szórása stb.) nagyobb-e (kisebb-e) valamely számnál (illetve egyenlő-e azzal)?
- Egy eloszlásfüggvény két (vagy több) paramétere eleget tesz-e valamely kikötésnek. Például, igaz-e hogy, egy eloszlás relatív szórása, σ/m , kisebb valamely adott számnál?
- Igaz-e, hogy egy valószínűségi változó valamely adott (pl. normális) eloszlással rendelkezik?
- Két (vagy több) valószínűségi változó független-e?
- Igaz-e, hogy két (vagy több) valószínűségi változó eloszlásfüggvényének valamely paramétere (pl. várható értéke) megegyezik?
- Két valószínűségi változó korrelációs együtthatója nagyobb-e valamely adott számnál?
- Két valószínűségi változó eloszlásfüggvénye megegyezik-e?

A példaként említett kérdéseket két csoportba oszthatjuk aszerint, hogy egy vagy több eloszlásra vonatkoznak-e. Az $a)$, $b)$, $c)$ kérdések egy, a $d)$, $e)$, $f)$, $g)$ kérdések több eloszlásra vonatkoznak. Az említett két típusú problémáknak megfelelően megkülönböztetünk egy- és két- (illetve több-) mintás próbákat.

Az egymintás próbákban döntésünk egy $\xi_1, \xi_2, \dots, \xi_n$ mintán alapul, ahol $\xi_1, \xi_2, \dots, \xi_n$ független, egyforma eloszlású valószínűségi változók. A kétmintás próbákban döntésünk egy $\xi_1, \xi_2, \dots, \xi_n$ és egy $\eta_1, \eta_2, \dots, \eta_m$ minta összehasonlításán alapul. (Itt $\xi_1, \xi_2, \dots, \xi_n$ független, egyforma eloszlású és hasonlóan $\eta_1, \eta_2, \dots, \eta_m$ is független, egyforma eloszlású valószínűségi változók, de általában a ξ -k az η -któl nem függetlenek, és eloszlásuk különböző.) Az egymintás esetben a $\xi_1, \xi_2, \dots, \xi_n$ mintát úgy lehet tekinteni, mint az n -dimenziós tér egy véletlenszerűen kiválasztott pontját, hasonlóan a kétmintás esetben a $\xi_1, \xi_2, \dots, \xi_n; \eta_1, \eta_2, \dots, \eta_m$ mintákat az $(n+m)$ -dimenziós tér egy pontjaként foghatjuk fel. A hipotézisvizsgálat egy problémájának megoldása azt jelenti, hogy megmondjuk, mely minták (azaz az n , illetve $(n+m)$ -dimenziós tér mely pontjai) esetében fogadjuk el és melyek esetében vetjük el hipotézisünket.

Definíció. Az n - (illetve $(n+m)$ -) dimenziós tér azon pontjainak halmazát, amelyek esetén a hipotézist elfogadjuk, *elfogadási tartománynak*, ezen halmaz komplementerét *elutasítási* vagy *kritikus tartománynak* nevezzük.*

A 15.2.2.3 pontban foglalkoztunk két gazdaság tehénállományának összehasonlításával. Azt találtuk, hogy az a hipotézis, amely szerint a két gazdaság tehénállománya a vizsgált szempontból megegyező, elfogadható, ha

$$|t| = \left| \frac{\bar{\xi}_1 - \bar{\xi}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right| \leq 1,96.$$

Itt t értéke a $\xi_{11}, \xi_{12}, \dots, \xi_{1n_1}; \xi_{21}, \xi_{22}, \dots, \xi_{2n_2}$ kísérleti eredményektől függ, azaz a t statisztika az (n_1+n_2) dimenziós téren értelmezett függvény. Eredményünk szerint az elfogadási tartomány az (n_1+n_2) dimenziós tér azon pontjainak halmaza, amelyekben a t függvény abszolút értéke nem nagyobb 1,96-nál. A kritikus tartomány pedig azon pontok halmaza, amelyekben $|t| > 1,96$.

Mint láttuk, bárhogya is választjuk meg a kritikus tartományt, elképzelhető, hogy nem döntünk helyesen, azaz elképzelhető, hogy egy valóban helyes hipotézisről úgy döntünk, hogy az nem helyes (*első fajú hiba*), illetve hogy egy valóban helytelen hipotézist elfogadjuk (*másodfajú hiba*). Általában előre kikötjük, hogy az első fajú hiba elkövetésének mennyi lehet a valószínűsége, és ennek megfelelően választjuk meg a kritikus tartományt. Így, ha feladatunk az, hogy valamely hipotézis helyes vagy hamis volta felől p szinten döntsünk, akkor ez azt jelenti, hogy a kritikus tartományt úgy kell megválasztanunk, hogy ha a hipotézis igaz, akkor annak a valószínűsége, hogy a minta a kritikus tartományba esik, p legyen. A p értéket általában a 0,10, 0,05, 0,03, 0,02, 0,01 számok valamelyikének választjuk.

Megjegyezzük, hogy a p érték megadása nem definiálja egyértelműen a kritikus tartományt. Ezért különböző módszereket, úgynevezett *próbákat* dolgozhatunk ki egyes hipotézis helyes vagy hamis voltának eldöntésére, azaz a kritikus tartomány megválasztására.

* Feltételezzük, hogy bármilyen eredményt is kapunk a mintavétel során, azaz az n - (illetve $(n+m)$ -) dimenziós tér bármely pontjához jutunk, a hipotézisről vagy elfogadólag vagy elutasítólag döntünk. A statisztika más ágaiban, például az úgynevezett szekvenciális analízisben, bizonyos kísérleti eredmények esetén nem döntünk sem elfogadólag, sem elutasítólag, hanem úgy is határozhatunk, hogy csak a kísérletek folytatása esetén lehet megbízható döntést hozni.

A p szint, a kritikus tartomány megválasztása után foglalkozhatunk azzal a kérdéssel, hogy mi mondható a másodfajú hiba valószínűségéről. A másodfajú hiba elkövetésének valószínűsége általában attól függ, hogy a hipotézis „mennyire nem igaz”. A Mendel-kísérlettel kapcsolatban már láttuk, hogy a másodfajú hiba elkövetésének valószínűsége attól függ, hogy mi a szóban forgó valószínűség igazi értéke, illetve, hogy az igazi valószínűség milyen távol van a hipotetikus valószínűségtől. Általában, ha eredeti hipotézisünket H_0 -val jelöljük (a Mendel-kísérlet esetében H_0 jelenti a $p = \frac{1}{4}$ hipotézist, a két gazdaság tehénállományával kapcsolatos problémánál H_0 jelenti, hogy a két várható érték megegyezik: $m_1 = m_2$),

és H_1 valamilyen más lehetséges ellenhipotézis, (a Mendel-kísérletnél H_1 lehet például a $p = 0,28$ hipotézis, a tehénállományokkal kapcsolatban H_1 lehet az $m_1 - m_2 = 0,1$ hipotézis), akkor a másodfajú hiba elkövetésének valószínűsége függ a H_1 ellenhipotézistől.

Definíció. Azt a függvényt, amely megadja a másodfajú hiba elkövetésének valószínűségét valamely H_1 ellenhipotézis esetén, *erőfüggvénynek* nevezzük.

Akárcsak a becslésméletben, itt is felvetődik a kérdés, hogyan lehet két próbát (két döntési módszert) összehasonlítani. Mivel az elsőfajú hiba valószínűsége előre adott, két próba közül azt célszerű jobbnak tekinteni, amely esetén a másodfajú hiba valószínűsége kisebb. Azonban — mint említettük — a másodfajú hiba valószínűsége csak egy függvénnyel (az erőfüggvénnyel) jellemezhető. Abban az esetben, ha egy próbának az erőfüggvénye minden H_1 ellenhipotézis esetén kisebb, mint valamely más próba erőfüggvénye, akkor az első próbára azt mondjuk, hogy *egyenletesen jobb*, mint a második próba. Általában azonban két próba nem hasonlítható ilyen egyszerűen össze, ugyanis bizonyos ellenhipotézisek esetén az egyik, más ellenhipotézisek esetén a másik erőfüggvénye kisebb. Ilyen esetekben csak olyan kijelentést tehetünk, hogy valamely próba bizonyos ellenhipotézisekre nézve jobb, mint egy másik próba.

Említettük, hogy az első fajú hiba valószínűségét (a döntési szintet) általában a 0,1, 0,05, 0,03, 0,02, 0,01 számok valamelyikének választjuk. Azt, hogy a döntési szintet valamely konkrét esetben hogyan célszerű megválasztani, elég nehéz megmondani. Erről a kérdéstről itt csak annyit jegyzünk meg, hogy a szint megválasztása általában gazdasági jellegű megfontolásoktól függ, nevezetesen attól, hogy milyen anyagi kárt jelent az első-, illetve a másodfajú hiba elkövetése. Matematikai szempontból a *döntésfüggvények* elmélete foglalkozik ilyen kérdésekkel.

A következőkben néhány gyakrabban használt próbát ismertetünk.

15.2.4.2 u -próba

Az u -próbát általában ismert szórású, normális eloszlású valószínűségi változó(k) várható értékére (értékeire) vonatkozó hipotézisekkel kapcsolatban használjuk.

Legyen
$$F(x; m, \sigma) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

egy ismert σ szórású és ismeretlen m várható értékű normális eloszlású eloszlásfüggvény. Továbbá legyenek $\xi_1, \xi_2, \dots, \xi_n$ független $F(x; m, \sigma)$ eloszlású valószínűségi változók, azaz $\xi_1, \xi_2, \dots, \xi_n$ a szóban forgó eloszlásból vett minta.

Feladatunk eldönteni annak a H_0 hipotézisnek helyes vagy helytelen voltát, hogy az ismeretlen várható érték megegyezik-e valamely adott m_0 számmal. Ezt általában úgy fejezzük ki, hogy hipotézisünk

$$H_0: m = m_0.$$

Abban az esetben, ha hipotézisünk helyes, akkor az

$$u = \frac{\bar{\xi} - m_0}{\frac{\sigma}{\sqrt{n}}} \quad \left(\bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} \right)$$

statisztika 0 várható értékű, 1 szórású, normális eloszlású valószínűségi változó. Így $m = m_0$ hipotézisünket p szinten elfogadhatjuk, ha

$$(3) \quad \left| \frac{\bar{\xi} - m_0}{\frac{\sigma}{\sqrt{n}}} \right| \leq \Phi^{-1} \left(1 - \frac{p}{2} \right).$$

Ugyanis, ha hipotézisünk helyes, akkor

$$P \left(\left| \frac{\bar{\xi} - m_0}{\frac{\sigma}{\sqrt{n}}} \right| \leq \Phi^{-1} \left(1 - \frac{p}{2} \right) \right) = 2\Phi \left(\Phi^{-1} \left(1 - \frac{p}{2} \right) \right) - 1 = 1 - p,$$

tehát valóban p a valószínűsége annak, hogy a (3) egyenlőtlenség nem teljesül, vagyis, hogy a hipotézist elvetjük. Azt mondhatjuk tehát, hogy a próba kritikus tartománya az n -dimenziós tér azon $\xi_1, \xi_2, \dots, \xi_n$ pontjainak összessége, amelyekre

$$\left| \frac{\bar{\xi} - m_0}{\frac{\sigma}{\sqrt{n}}} \right| > \Phi^{-1} \left(1 - \frac{p}{2} \right).$$

Hasonló módon dönthetünk például a következő hipotézisek helyes vagy helytelen voltáról.

a) Egy ismert szórású, normális eloszlású változó várható értéke nagyobb (kisebb) egy adott számnál.

$$H_0: m > m_0 \quad (H_0: m < m_0).$$

b) Két független, ismert szórású, normális eloszlású változó várható értéke egyenlő egymással

$$H_0: m_1 = m_2.$$

c) Két független, ismert szórású, normális eloszlású változó közül az első várható értéke nagyobb (kisebb), mint a másodiké.

$$H_0: m_1 > m_2 \quad (H_0: m_1 < m_2).$$

d) Két független, ismert szórású, normális eloszlású változó várható értékének különbsége adott számmal egyenlő (adott számnál kisebb, adott számnál nagyobb):

$$H_0: m_1 - m_2 = m_0 \quad (H_0: m_1 - m_2 < m_0; m_1 - m_2 > m_0).$$

Az a) pontban említett hipotézist elfogadhatjuk, ha

$$\frac{\bar{\xi} - m_0}{\frac{\sigma}{\sqrt{n}}} > \Phi^{-1}(p)$$

(ahol n a minta elemszáma, $\bar{\xi}$ a mintaátlag, σ az ismert szórás, p az előírt szint). Ugyanis ha a hipotézis helyes, akkor

$$P \left(\frac{\bar{\xi} - m_0}{\frac{\sigma}{\sqrt{n}}} > \Phi^{-1}(p) \right) = 1 - \Phi(\Phi^{-1}(p)) = 1 - p.$$

A b) pontban említett hipotézist elfogadhatjuk, ha

$$\left| \frac{\frac{\bar{\xi}_1}{n_1} - \frac{\bar{\xi}_2}{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| < \Phi^{-1} \left(1 - \frac{p}{2} \right),$$

ahol n_1 (illetve n_2) az első (illetve második) minta elemszáma, $\bar{\xi}_1$ (illetve $\bar{\xi}_2$) az első (illetve a második) minta átlaga, σ_1 (illetve σ_2) az első (illetve második) eloszlás szórása, p az előírt szint, ugyanis, ha a hipotézis helyes, akkor

$$P \left(\left| \frac{\frac{\bar{\xi}_1}{n_1} - \frac{\bar{\xi}_2}{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| < \Phi^{-1} \left(1 - \frac{p}{2} \right) \right) = 2\Phi \left(\Phi^{-1} \left(1 - \frac{p}{2} \right) \right) - 1 = 1 - p.$$

A c) és d) pontban említett hipotézisek helyes vagy hamis voltáról teljesen hasonló módon dönthetünk. Így e próbák részletes kidolgozását az olvasóra bizzuk.

Nézzük meg például a b) esetben, hogy mi a próba erőfüggvénye, vagyis mi a hipotézis elfogadásának valószínűsége, ha a hipotézis nem teljesül: $m_1 \neq m_2$, hanem például $m_1 = m_2 + m_0$. Mivel hipotézisünket akkor fogadjuk el, ha

$$\left| \frac{\frac{\bar{\xi}_1}{n_1} - \frac{\bar{\xi}_2}{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| < \Phi^{-1} \left(1 - \frac{p}{2} \right),$$

tehát azt kell megvizsgálunk, mi a valószínűsége ezen egyenlőtlenség teljesülésének, ha $m_1 = m_2 + m_0$. Kapjuk, hogy

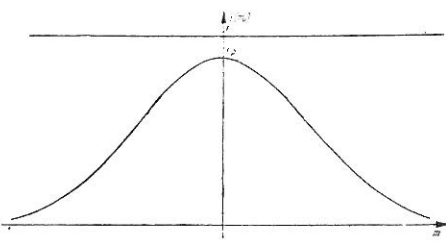
$$P\left(\left|\frac{\frac{\bar{x}_1}{n_1} - \frac{\bar{x}_2}{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| < \Phi^{-1}\left(1 - \frac{p}{2}\right)\right) = \Phi\left(\Phi^{-1}\left(1 - \frac{p}{2}\right) - m_0\right) - \Phi\left(-\Phi^{-1}\left(1 - \frac{p}{2}\right) - m_0\right).$$

Tehát a próba erőfüggvénye

$$E(m_0) = \Phi\left(\Phi^{-1}\left(1 - \frac{p}{2}\right) - m_0\right) - \Phi\left(-\Phi^{-1}\left(1 - \frac{p}{2}\right) - m_0\right).$$

A 15.7 ábra szemlélteti az erőfüggvényt.

Megjegyzés. Az u -próba alkalmazása esetén mindig feltételeztük, hogy a vizsgált eloszlás normális, de valójában csak azt használtuk fel, hogy a minta átlaga normális eloszlású. Ez utóbbi kikötés viszont — a centrális határeloszlás-tétel szerint — igen általános feltételek mellett közelítőleg teljesül, ha a mintaelemszám elég nagy. Így az u -próbát gyakran alkalmazhatjuk akkor is, ha az alapeloszlás nem normális. Ezt tettük például a Mendel-kísérlet vizsgálatával kapcsolatban is.



15.7 ábra

15.2.4.3 A t -próba

A t -próbát általában ismeretlen szórású, normális eloszlású változó(k) várható értékére (értékeire) vonatkozó hipotézisekkel kapcsolatban használjuk. Tehát ugyanazon hipotézisekről dönthetünk, mint az u -próba alkalmazásával, de most nem kell feltételeznünk, hogy a szórás ismert.

$$\text{Legyen} \quad F(x; m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt$$

egy ismeretlen, m várható értékű és σ szórású normális eloszlású eloszlásfüggvény. Továbbá legyenek $\xi_1, \xi_2, \dots, \xi_n$ független, $F(x; m, \sigma)$ eloszlású valószínűségi változók. Feladatunk eldönteni annak a H_0 hipotézisnek a helyes vagy helytelen voltát, hogy az ismeretlen várható érték megegyezik-e valamelyik adott m_0 számmal: $H_0: m = m_0$.

Abban az esetben, ha hipotézisünk helyes, akkor a

$$t = \frac{\bar{x} - m_0}{\frac{\sigma_n}{\sqrt{n}}} \quad \left(\bar{x} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n}; \quad \sigma_n = \sqrt{\frac{\sum_{i=1}^n (\xi_i - \bar{x})^2}{n-1}} \right)$$

statisztika eloszlása az úgynevezett $f = n-1$ paraméterű, (*szabadságfokú*) t_f eloszlás. Így $m = m_0$ hipotézisünket p szinten elfogadjuk, ha

$$\left| \frac{\bar{x} - m_0}{\frac{\sigma_n}{\sqrt{n}}} \right| \leq t_f^{-1}\left(1 - \frac{p}{2}\right).$$

Ugyanis, ha hipotézisünk helyes, akkor

$$P\left(\left| \frac{\bar{x} - m_0}{\frac{\sigma_n}{\sqrt{n}}} \right| \leq t_f^{-1}\left(1 - \frac{p}{2}\right)\right) = t_f\left(t_f^{-1}\left(1 - \frac{p}{2}\right)\right) - t_f\left(-t_f^{-1}\left(1 - \frac{p}{2}\right)\right).$$

A normális eloszláshoz hasonlóan a t_f eloszlás is szimmetrikus, így

$$t_f(-\varepsilon) = 1 - t_f(\varepsilon),$$

$$\text{tehát} \quad P\left(\left| \frac{\bar{x} - m_0}{\frac{\sigma_n}{\sqrt{n}}} \right| \leq t_f^{-1}\left(1 - \frac{p}{2}\right)\right) = 2t_f\left(t_f^{-1}\left(1 - \frac{p}{2}\right)\right) - 1 = 1 - p.$$

A 15.2.3 pontban már részletesen kidolgoztunk egy példát a t -próba alkalmazására. Így itt már csak azt említjük meg, hogy — akárcsak az u -próba esetén — itt is gyakran alkalmazható a próba abban az esetben is, ha az alapeloszlás nem normális. Annak eldöntése azonban, hogy a próba valóban mikor alkalmazható, mindig a konkrét eset részletes vizsgálatát igényli.

15.2.4.4 A χ^2 -próba

A χ^2 -próbának sok különböző alkalmazása közül először a normális eloszlás szórására vonatkozó hipotézisek vizsgálatát említjük.

$$\text{Legyen} \quad F(x; 0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{t^2}{2\sigma^2}} dt$$

egy 0 várható értékű, ismeretlen σ szórású normális eloszlás. Továbbá legyenek $\xi_1, \xi_2, \dots, \xi_n$ független, $F(x; 0, \sigma)$ eloszlású valószínűségi változók.

Feladatunk eldönteni annak a H_0 hipotézisnek a helyes vagy hamis voltát, hogy az ismeretlen szórás egyenlő-e valamely adott σ_0 számmal: $H_0: \sigma = \sigma_0$.

Abban az esetben, ha hipotézisünk helyes, akkor a

$$\chi_f^2 = \frac{\sum_{i=1}^n \xi_i^2}{\sigma_0^2}$$

valószínűségi változó $f = n$ szabadságfokú, χ^2 eloszlású változó, azaz sűrűségfüggvénye

$$\chi_f^2(t) = \frac{1}{2^{f/2} \Gamma\left(\frac{f}{2}\right)} t^{f/2-2} e^{-\frac{t}{2}},$$

$$\text{ahol } \Gamma\left(\frac{f}{2}\right) = \begin{cases} \left(\frac{f}{2}-1\right)\left(\frac{f}{2}-2\right)\dots 3,2,1, & \text{ha } f \text{ páros és } f > 2, \\ \left(\frac{f}{2}-1\right)\left(\frac{f}{2}-2\right)\dots \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi}, & \text{ha } f \text{ páratlan és } f > 2, \\ 1, & \text{ha } f = 2, \\ \sqrt{\pi}, & \text{ha } f = 1. \end{cases}$$

A χ^2 valószínűségi változó várható értéke

$$M(\chi^2) = \frac{\sum_{i=1}^n M(\xi_i^2)}{\sigma_0^2} = n.$$

Így a $\sigma = \sigma_0$ hipotézist elfogadhatjuk, ha χ^2 az n várható érték közelében van.

A kritikus tartomány meghatározásához találnunk kell egy $(a; b)$ intervallumot úgy, hogy

$$P(a < \chi^2 < b) = 1-p$$

legyen, ahol p az elfogadási szint. Az u - és t -próba esetén az elfogadási intervallumot úgy választottuk meg, hogy az a várható értékre (ami ott 0 volt) szimmetrikus legyen. Valójában ez ott is önkényes választás volt, de mivel az u - és a t -eloszlás szimmetrikus, ez a választás számítástechnikai szempontból igen kényelmes. A χ^2 eloszlás esetében a keresett $(a; b)$ intervallumot például a következő módon választhatjuk meg. Legyen a az a szám, amelyre

$$P(a < \chi^2) = 1 - \frac{p}{2},$$

b pedig az, amelyre

$$P(\chi^2 < b) = 1 - \frac{p}{2}.$$

Ekkor

$$P(a < \chi^2 < b) = 1-p.$$

Tehát $\sigma = \sigma_0$ hipotézisünket elfogadhatjuk, ha χ^2 értéke a most definiált a és b számok közé esik.

Mint említettük, az a és b értékek megválasztása bizonyos mértékig önkényes, pontosabban a és b egyikének megválasztása már egyértelműen meghatározza a másikat. Így az elfogadási intervallumot néha (főként, ha f értéke nem túl nagy) úgy választjuk meg, hogy $a = 0$ legyen, és ekkor b -t úgy kell megválasztanunk, hogy

$$P(\chi^2 < b) = 1-p$$

legyen.

Abban az esetben, ha a várható érték is ismeretlen, akkor a $H_0: \sigma = \sigma_0$ hipotézisről hasonló módon dönthetünk, felhasználva azt a tényt, hogy a

$$\chi_1^2 = \frac{\sum_{i=1}^n (\xi_i - \bar{\xi})^2}{\sigma_0^2} \quad \left(\bar{\xi} = \frac{\sum_{i=1}^n \xi_i}{n} \right)$$

statisztika $f = n-1$ szabadságfokú, χ^2 eloszlású valószínűségi változó, ha a hipotézis igaz. Így a H_0 hipotézis helyes vagy hamis volta felől a χ^2 táblázat felhasználásával tetszőleges szinten dönthetünk, abban az esetben is, ha a várható érték ismeretlen.

Bár nem bizonyítjuk be, hogy a χ_1^2 valószínűségi változó valóban χ^2 eloszlású, néhány megjegyzést kívánunk tenni ezzel kapcsolatban. Könnyen belátható, hogy ha $M(\xi_i) = m$, akkor

$$\sum_{i=1}^n (\xi_i - m)^2 = \sum_{i=1}^n (\xi_i - \bar{\xi})^2 + n(\bar{\xi} - m)^2.$$

Az előzőekből már tudjuk, hogy a formulánk bal oldalán álló valószínűségi változó n szabadságfokú, χ^2 eloszlású változó, és mivel $\bar{\xi}$ is normális eloszlású, a jobb oldalon álló második tag 1 szabadságfokú χ^2 eloszlású. Tehát χ_1^2 egy n és egy 1 szabadságfokú χ^2 eloszlású változó különbsége. Így nem meglepő az, hogy χ_1^2 egy $n-1$ szabadságfokú χ^2 eloszlású valószínűségi változó. (Természetesen ez a megjegyzés nem tekinthető bizonyításnak.)

Talán furcsának tűnhet, hogy a χ^2 eloszlás paraméterét szabadságfoknak nevezzük. Ezért megpróbáljuk megvilágítani ennek az elnevezésnek az okát. Eredetileg az n szabadságfokú χ^2 eloszlású valószínűségi változót mint n független 0,1 paraméterű normális eloszlású változó négyzetösszegét definiáltuk. Azaz a „véletlennek” n -szeres teljes „szabadsága van befolyásolni” χ^2 értékét. A χ_1^2 valószínűségi változó definíciójában szintén n tag szerepel, de ezek nem függetlenek, sőt, köztük lineáris összefüggés áll fenn, nevezetesen

$$\sum_{i=1}^n (\xi_i - \bar{\xi}) = 0.$$

Így, ha az n tag közül $n-1$ értéke már ismert, az n -edik kiszámítható. Tehát a „véletlennek” csak $n-1$ „lehetősége van χ_1^2 értékét befolyásolni”. Általában n változó szabadságfokát úgy kapjuk meg, hogy az n számból levonjuk a változók között levő lineáris kapcsolatok számát. A χ^2 -próba lényegesen más típusú alkalmazása a függetlenségvizsgálat. Térjünk vissza a 15.2.2.1 pontban vizsgált problémánkhoz. Említettük, hogy a tej átlagos zsírszázalékából és a beadott tejmenyiségből csak akkor tudunk következtetést levonni a tejben levő zsírmennyiségre, ha a tej mennyisége és a zsírszázalék függetlenek. Ugyanis elképzelhető, hogy olyan napokon, amikor a kifejt (beadott) tej mennyisége nagy, akkor zsírszázaléka kicsi és megfordítva, amikor a tej mennyisége kicsi, akkor zsírszázaléka nagy. Abban az esetben, ha ez valóban így van (és a beadott zsírmennyiségre a tej mennyiségéből, valamint az átlag zsírszázalékból akarunk következtetéseket levonni), akkor olyan esetekben, amikor a beadott tej mennyisége nagy, a zsírmennyiséget felül, más esetekben alul becsüljük. Vizsgáljunk meg egy konkrét példát.

Egy gazdaságban egy tehén által adott tejet 50 fejés alkalmával megvizsgáltuk. A következő eredményeket kaptuk:

liter \ zsír %	3,5	3,6	3,7	3,8	Össz.
10,0	1	3	3	5	12
10,5	3	3	4	3	13
11,0	3	4	4	3	14
11,5	4	3	3	1	11
Összesen	11	13	14	12	50

Tehát az 50 alkalomból 4-szer kaptunk 11 liter 3,6% zsírtartalmú tejet, 14 alkalommal kaptunk 11 liter tejet, és 13 alkalommal fejtünk 3,6% zsírtartalmú tejet. Így annak a relatív gyakorisága, hogy egy fejésnél 11 liter 3,6% zsírtartalmú tejet kapunk, $\frac{4}{50} = 0,08$, annak az eseménynek a relatív gyakorisága, hogy 11 liter tejet kapunk $\frac{14}{50} = 0,28$, és annak az eseménynek a relatív gyakorisága, hogy 3,6% zsírtartalmú tejet kapunk, $\frac{13}{50} = 0,26$. Abban az esetben, ha a zsírszázalék és tej mennyisége függetlenek, akkor az elsőként említett (szorzat) esemény relatív gyakorisága közel egyenlő kell, hogy legyen a második két esemény relatív gyakoriságának szorzatával. Esetünkben a $0,08$ és a $\frac{14}{50} \cdot \frac{13}{50} = 0,0728$ számok különbségének kicsi vagy nagy volta alapján dönthetünk arról, hogy azt a hipotézisünket, hogy a zsírszázalék és a tejmennyiség függetlenek, elfogadjuk-e. Természetesen, ha a zsírszázalék és a tejmennyiség valóban függetlenek, akkor a többi relatív gyakoriságokra is hasonló kikötéseknek kell teljesülni. Nevezetesen vizsgálnunk kell az

$$\begin{aligned} \frac{1}{50} - \frac{11}{50} \frac{12}{50} &= -0,0328, & \frac{3}{50} - \frac{14}{50} \frac{12}{50} &= -0,072, \\ \frac{3}{50} - \frac{11}{50} \frac{13}{50} &= 0,0028, & \frac{4}{50} - \frac{14}{50} \frac{13}{50} &= +0,072, \\ \frac{3}{50} - \frac{11}{50} \frac{14}{50} &= -0,0016, & \frac{4}{50} - \frac{14}{50} \frac{14}{50} &= 0,0016, \\ \frac{4}{50} - \frac{11}{50} \frac{11}{50} &= +0,0316, & \frac{3}{50} - \frac{14}{50} \frac{11}{50} &= -0,0016, \end{aligned}$$

$$\begin{aligned} \frac{3}{50} - \frac{13}{50} \frac{12}{50} &= -0,024, & \frac{5}{50} - \frac{12}{50} \frac{12}{50} &= 0,0424 \\ \frac{3}{50} - \frac{13}{50} \frac{13}{50} &= -0,0076, & \frac{3}{50} - \frac{12}{50} \frac{13}{50} &= -0,0024 \\ \frac{4}{50} - \frac{13}{50} \frac{14}{50} &= 0,0082, & \frac{3}{50} - \frac{12}{50} \frac{14}{50} &= -0,0072 \\ \frac{3}{50} - \frac{13}{50} \frac{11}{50} &= 0,0028, & \frac{1}{50} - \frac{12}{50} \frac{11}{50} &= -0,0328 \end{aligned}$$

különbségeket. Ezek nagyságának együttes áttekintése céljából általában a

$$\begin{aligned} &(-0,0328)^2 + (-0,0024)^2 + (-0,0072)^2 + (0,0424)^2 + (0,0028)^2 + (-0,0076)^2 + \\ &+ (0,0082)^2 + (-0,0024)^2 + (-0,0016)^2 + (0,0072)^2 + (0,0016)^2 + (0,0072)^2 + \\ &+ (0,0316)^2 + (0,0028)^2 + (-0,0016)^2 + (-0,0328)^2 = 0,0052 \end{aligned}$$

összeget vagy ennek valamely normáltját szoktuk vizsgálni. Abban az esetben, ha ez az összeg nagy, a függetlenségi hipotézist elvetjük, ha kicsi, megtartjuk. Természetesen annak a kérdésnek az eldöntésére, hogy *mennyire* kell a fenti összegnek nagy, illetve kicsinek lennie ahhoz, hogy a függetlenségi hipotézist elvethessük, illetve elfogadhassuk, statisztikai próbát kell alkalmaznunk. Ahhoz, hogy megtaláljuk a megfelelő statisztikai próbát, általánosan fogalmazzuk meg kérdésünket. Egyidejűleg két, különböző valószínűségi változó (pl. tejmennyiség, zsírszázalék) értékét határozzuk meg kísérleti úton. Legyen e két valószínűségi változó ξ és η ; ξ lehetséges értékei legyenek $x_1, x_2, \dots, x_r$; η lehetséges értékei legyenek $y_1, y_2, \dots, y_s$; a kísérletek száma legyen n . Jelölje v_{ij} azon kísérletek számát, amelyeknek eredményeül a $\xi = x_i, \eta = y_j$ párt kaptuk. Így $\frac{v_{ij}}{n}$ a $\xi = x_i, \eta = y_j$ esemény relatív gyakorisága. Bevezetve a

$$\sum_{i=1}^r v_{ij} = v_{.j}, \quad \sum_{j=1}^s v_{ij} = v_{i.}$$

jelölést*, azt mondhatjuk, hogy $\frac{v_{i.}}{n}$ a $\xi = x_i$ és $\frac{v_{.j}}{n}$ az $\eta = y_j$ esemény relatív gyakorisága. A függetlenségi hipotézis eldöntése céljából tekintsük a

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(v_{ij} - \frac{v_{i.} v_{.j}}{n}\right)^2}{\frac{v_{i.} v_{.j}}{n}}$$

* A statisztikában általában szokásos, hogy ha egy többindexű számsorozat tagjait néhány index szerint összegezzük, akkor az összeget úgy jelöljük, hogy az összegzett indexek helyére pontokat írunk.

statisztikát. Bebizonyítható, hogy a χ^2 valószínűségi változó közelítőleg $f = (r-1)(s-1)$ paraméterű, χ^2 eloszlású. Így a χ^2 -táblázat alkalmazásával adott szinten dönthetünk a hipotézis elfogadásáról vagy elvetéséről.

Példánkra visszatérve, mivel táblázatunkban azt találtuk, hogy $\chi^2_5(2,088) = 0,99$ és $\chi^2_5(21,67) = 0,01$, tehát függetlenségi hipotézisünket 0,02 szinten elfogadhatjuk, ha az adatokból számított χ^2 érték 2,088 és 21,67 között van. Esetünkben χ^2 értékére 4,5 adódik, tehát függetlenségi hipotézisünket elfogadhatjuk.

Megjegyzés. A kérdés általános megfogalmazásánál diszkrét valószínűségi változók szerepeltek, míg a konkrét példában szereplő valószínűségi változók folytonosak, és csak azáltal váltak diszkrétte, hogy az adatok csak bizonyos pontossáig voltak megadva. Mivel minden mérési adat csak bizonyos pontossáig (pl. 2 vagy 3 jegy pontossáig) van megadva, a statisztikai adatok általában úgy kezelhetők, mintha egy diszkrét eloszlásból vett adatok lennének.

A χ^2 -próba számos alkalmazása közül következőként a homogenitás-vizsgálattal foglalkozunk. Induljunk ki ismét egy konkrét példából. Tegyük fel, hogy két különböző gépcsoport hatását vizsgáljuk, és azon hipotézisről akarunk dönteni, hogy a két gépcsoport egyforma hatású-e vagy sem. Természetesen ahhoz, hogy e hipotézisről dönthessünk, szükséges, hogy a két gépcsoportot ugyanolyan körülmények között működtessük.

Például ha két különböző gépcsoporttal szántunk, akkor ahhoz, hogy a két gépcsoport teljesítményét és munkájának minőségét összehasonlíthassuk, szükséges, hogy a két gépcsoportot azonos (vagy közel azonos) minőségű talajon dolgoztassuk.

A gépek munkájának minőségét ellenőrizhetjük például oly módon, hogy különböző helyeken megmérjük a szántás mélységét és ezek átlagát hasonlítjuk össze. Azonban ez az adat igen keveset mond, ugyanis itt nemcsak az átlagmélység, hanem a szántás egyenletessége is fontos. Így akkor járunk el helyesen, ha a szántásmélység eloszlásfüggvényét vizsgáljuk. Azaz ξ legyen az első, η a második gépcsoport által végzett szántás mélysége a szántott tábla egy véletlenszerűen kiválasztott pontjában. Feladatunk annak a hipotézisnek a vizsgálata, hogy ξ és η eloszlásfüggvénye megegyezik-e. (Homogenitás-vizsgálat.)

Tegyük fel, hogy mindkét gépcsoport szántásmélységét ellenőriztük 200–200 helyen, és a következő adatokat kaptuk:

Szántásmélység	20 cm	21 cm	22 cm	23 cm	24 cm	25 cm	összesen
1. gépcsoport	8	32	56	64	28	12	200
2. gépcsoport	24	36	40	44	32	24	200
Összesen	32	68	96	108	60	36	400

Nyilvánvaló, hogy ha a két eloszlás megegyezik, akkor a

$$24-8=16, \quad 36-32=4, \quad 40-56=-16, \quad 44-64=-20, \quad 32-28=4, \\ 24-12=12$$

különbségeknek kicsiknek kell lenniük a kísérletek számához, 400-hoz képest.

Az alkalmazható statisztikai próba megtalálása céljából ismét általánosan fogalmazzuk meg kérdésünket. Legyen ξ és η két valószínűségi változó (pl. a két gépcsoport által végzett szántás mélysége), mindkettő lehetséges értékei legyenek az $a_1, a_2, \dots, a_r$ számok*. Tegyük fel, hogy ξ értékét n , η értékét m különböző alkalommal mértük**, és az n kísérletből az $a_1, a_2, \dots, a_r$ értékek rendre $v_1, v_2, \dots, v_r$ -szer ($v_1+v_2+\dots+v_r=n$), az m kísérletből ugyanezek az értékek rendre $\mu_1, \mu_2, \dots, \mu_r$ -szer ($\mu_1+\mu_2+\dots+\mu_r=m$) fordulnak elő. Így kísérletünk eredményét a

	a_1	a_2	$\dots$	a_r	összesen
ξ	v_1	v_2	$\dots$	v_r	n
η	μ_1	μ_2	$\dots$	μ_r	m
összesen	$\delta_1 = v_1 + \mu_1$	$\delta_2 = v_2 + \mu_2$	$\dots$	$\delta_r = v_r + \mu_r$	$n+m$

táblázat szemlélteti. Abban az esetben, ha a két eloszlás megegyezik (és n és m elég nagy) akkor a

$$\frac{v_1}{n} - \frac{\mu_1}{m}, \quad \frac{v_2}{n} - \frac{\mu_2}{m}, \quad \dots, \quad \frac{v_r}{n} - \frac{\mu_r}{m}$$

és a

$$\frac{v_1}{n} - \frac{\delta_1}{n+m}, \quad \frac{v_2}{n} - \frac{\delta_2}{n+m}, \quad \dots, \quad \frac{v_r}{n} - \frac{\delta_r}{n+m};$$

$$\frac{\mu_1}{m} - \frac{\delta_1}{n+m}, \quad \frac{\mu_2}{m} - \frac{\delta_2}{n+m}, \quad \dots, \quad \frac{\mu_r}{m} - \frac{\delta_r}{n+m}$$

különbségeknek kicsinyeknek kell lenniük. Belátható, hogy a

$$\chi^2 = \sum_{i=1}^r \frac{\left(v_i - \frac{\delta_i n}{n+m}\right)^2}{\frac{\delta_i n}{n+m}} + \sum_{i=1}^r \frac{\left(\mu_i - \frac{\delta_i m}{n+m}\right)^2}{\frac{\delta_i m}{n+m}}$$

valószínűségi változó $f = r-1$ paraméterű, χ^2 eloszlású. Így a χ^2 valószínűségi változó értékének ismeretében és a χ^2 -táblázat felhasználásával tetszőleges szinten dönthetünk homogenitási hipotézisünk helyes vagy hamis voltáról.

Példánkban $f = r-1 = 5$ és χ^2 értékére adatainkból 18,88 adódik. Táblázatunk szerint $\chi^2_5(0,554) = 0,99$ és $\chi^2_5(15,09) = 0,01$, tehát hipotézisünk 0,02 szinten elvethető.

* Ha a lehetséges értékek különbözők lennének, akkor a két eloszlás is különböző lenne.

** Példánkban $n = m = 200$. Általában a kiértékelés megkönnyítése céljából igyekszünk n és m értékét egyenlőnek választani. Néha azonban ezt nem tudjuk megvalósítani. Például állatkísérleteknél, még ha a két csoportba sorolt állatok száma kezdetben egyenlő is, akkor is előfordulhat, hogy a két csoport valamelyikéből egy (vagy több) állat elhullik, és ekkor a kiértékelhető adatok száma különböző lesz.

Megjegyzés. A táblázatból elég világosan látszik, hogy az első gépcsoport munkája sokkal egyenletesebb, a 22–23 cm szántásmélységtől ritkábban mutat komoly eltérést. Statisztikai próbánk eredménye folytán most már azt is mondhatjuk, hogy a táblázat által mutatott eltérés elég jelentős (szignifikáns) ahhoz, hogy állíthassuk, hogy a két eloszlás valóban különböző.

Hasonló próba dolgozható ki arra az esetre is, amikor nem két, hanem több mintát kell összehasonlítani, de erre itt nem térünk ki.

A χ^2 -próba utolsó alkalmazásaként az úgynevezett illeszkedésvizsgálatot említjük meg. Illeszkedésvizsgálaton olyan hipotézisek ellenőrzését értjük, amelyek azt állítják, hogy valamely valószínűségi változó adott eloszlással rendelkezik.

A raktározási probléma vizsgálatánál például láttuk, hogy elméleti megfontolások alapján arra a következtetésre lehet jutni, hogy egy raktárba a szállítmányok Poisson-eloszlásnak megfelelően érkeznek. Természetesen ahhoz, hogy biztosak lehessünk abban, hogy a szállítmányok valamely konkrét esetben valóban Poisson-eloszlás szerint érkeznek, vagy ellenőriznünk kell, hogy az elméleti megfontolásokban alkalmazott feltevések valóban teljesülnek-e a vizsgált esetben, vagy mintavétel segítségével és statisztikai próba alkalmazásával kell eldöntenünk, hogy a szállítmányok beérkezése valóban tekinthető-e Poisson-eloszlásúnak. Az elméleti megfontolás feltételeinek gyakorlati ellenőrzése igen nehéznek látszik, így inkább a második módszerhez szoktunk folyamodni.

Tegyük fel például, hogy 50 napon keresztül feljegyeztük a vizsgált raktárba érkező szállítmányok számát, és a következő adatokat kaptuk:

Szállítmányok száma	0	1	2	3	4	5	6
Napok száma	7	10	15	8	5	3	2

vagyis az 50 megfigyelt nap közül 7 olyan nap volt, amikor egyáltalán nem érkezett szállítmány, 10 olyan nap volt, amikor 1, 15 olyan nap, amikor 2, ..., és 2 olyan nap, amikor 6 szállítmány érkezett. Így átlagban naponként

$$\frac{0 \cdot 7 + 1 \cdot 10 + 2 \cdot 15 + 3 \cdot 8 + 4 \cdot 5 + 5 \cdot 3 + 6 \cdot 2}{50} = 2,22$$

szállítmány érkezik. A $\lambda = 2,22$ paraméterű Poisson-eloszlás táblázata szerint annak a valószínűsége, hogy 0, 1, 2, ..., 6 szállítmány érkezzék, rendre

$$e^{-2,22} = 0,11, \quad 2,22e^{-2,22} = 0,24, \quad \frac{2,22^2}{2} e^{-2,22} = 0,26,$$

$$\frac{(2,22)^3}{3!} e^{-2,22} = 0,19, \quad \frac{(2,22)^4}{4!} e^{-2,22} = 0,11,$$

$$\frac{(2,22)^5}{5!} e^{-2,22} = 0,05, \quad \frac{(2,22)^6}{6!} e^{-2,22} = 0,02.$$

Így, ha a szállítmányok valóban Poisson-eloszlás szerint érkeznek, akkor a

$$\frac{7}{50} - 0,11, \quad \frac{10}{50} - 0,24, \quad \frac{15}{50} - 0,26, \quad \frac{8}{50} - 0,19, \quad \frac{5}{50} - 0,11, \\ \frac{3}{50} - 0,05, \quad \frac{2}{50} - 0,02$$

különbségeknek kell kicsinyeknek lenniük.

Általában, ha azt a hipotézist akarjuk ellenőrizni, hogy egy ξ valószínűségi változó valóban Poisson-eloszlású-e, és n kísérlet során azt tapasztaltuk, hogy ξ a 0, 1, 2, ..., r értékeket rendre $v_0, v_1, v_2, \dots, v_r$ -szer ($v_0 + v_1 + v_2 + \dots + v_r = n$) veszi fel, akkor a

$$\chi^2 = \sum_{i=1}^r \frac{(v_i - np_i)^2}{np_i} \quad \left(p_i = \frac{\lambda^i}{i!} e^{-\lambda}, \lambda = \frac{v_1 + 2v_2 + \dots + rv_r}{n} \right)$$

valószínűségi változó közelítőleg $f = r$ szabadságfokú χ^2 eloszlású.

Mivel adataink alapján esetünkben χ^2 értéke 2,43-nak adódik, a szabadságfok $f = 6$, és χ^2 -táblázatunk szerint $\chi^2_6(0,872) = 0,99$ és $\chi^2_6(16,81) = 0,01$, így hipotézisünk 0,02 szinten elfogadható.

Megjegyezzük, hogy a χ^2 -próba akkor is alkalmazható, ha azt a hipotézist akarjuk ellenőrizni, hogy valószínűségi változónk eloszlása megegyezik valamely — Poisson-tól különböző — diszkrét eloszlással.

A függetlenségvizsgálattal kapcsolatban megjegyeztük, hogy bár a próbát elvileg csak diszkrét eloszlások esetén alkalmazhatjuk, nem diszkrét eloszlású valószínűségi változók esetére is jó közelítéssel alkalmazható. Hasonló elmondható a homogenitás- és az illeszkedés-vizsgálatról is, de e problémák vizsgálatára a folytonos eloszlású valószínűségi változók esetében más próbák is rendelkezésre állnak (lásd a 15.4 pontot).

15.2.4.5 v^2 -próba

Említettük, hogy normális eloszlású valószínűségi változók szórására vonatkozó hipotézisek helyes vagy hamis voltáról általában a χ^2 -próba alkalmazásával dönthetünk. Nyilvánvaló módon alkalmazható ez a próba például annak a hipotézisnek a vizsgálatára, hogy két független normális eloszlású valószínűségi változó szórása megegyezik-e.

E fejezetben megismerkedünk egy más módszerrel is, amely e hipotézis vizsgálatára használható. A szóban forgó próba számos alkalmazásával a következő fejezetben (Szórásanalízis) fogunk megismerkedni.

Legyen ξ és η két normális eloszlású valószínűségi változó (pl. két különböző gépcsoport által végzett szántás mélysége). Végezzünk ξ -re vonatkozólag n , η -ra vonatkozólag m kísérletet. A ξ -re vonatkozó kísérletek eredményei legyenek $\xi_1, \xi_2, \dots, \xi_n$, az η -ra vonatkozó kísérletek eredményei legyenek $\eta_1, \eta_2, \dots, \eta_m$. Feltesszük, hogy a ξ -re és az η -ra vonatkozó kísérletek függetlenek, azaz a $\xi_1, \xi_2, \dots, \xi_n; \eta_1, \eta_2, \dots, \eta_m$ valószínűségi változók függetlenek. Ekkor ξ tapasztalati szórásnégyzete

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2 \quad \left(\bar{\xi} = \frac{\xi_1 + \xi_2 + \dots + \xi_n}{n} \right),$$

és η tapasztalati szórásnégyzete

$$s_2^2 = \frac{1}{m-1} \sum_{i=1}^m (\eta_i - \bar{\eta})^2 \quad \left(\bar{\eta} = \frac{\eta_1 + \eta_2 + \dots + \eta_m}{m} \right).$$

Belátható, hogy ha ξ és η szórásnégyzete megegyezik, akkor a

$$v^2 = \frac{s_1^2}{s_2^2}$$

valószínűségi változó sűrűségfüggvénye:

$$f(v) = \frac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} f_1^{\frac{f_1}{2}} f_2^{\frac{f_2}{2}} \frac{(v^2)^{\frac{f_1}{2}-1}}{(f_2+f_1v^2)^{\frac{f_1+f_2}{2}}}$$

($0 < v < \infty$; $f_1 = n-1$, $f_2 = m-1$)

úgynevezett $f_1 = n-1$, $f_2 = m-1$ paraméterű v^2 -eloszlás.

Az
$$F(x) = \int_0^x f(t) dt$$

eloszlásfüggvény értékeit különböző, x , f_1 , f_2 értékekre táblázatokba szokták foglalni.

15.3 SZÓRÁSANALÍZIS

A hipotézisvizsgálattal kapcsolatban említettük, hogy foglalkozhatunk olyan hipotézisek vizsgálatával is, amelyek több eloszlásra vonatkoznak. Valójában mégis csak egy vagy legfeljebb két eloszlásra vonatkozó hipotéziseket vizsgáltunk. Ennek oka az, hogy bár a hipotézisvizsgálattal kapcsolatban ismertett próbák elvileg alkalmasak több eloszlásra vonatkozó hipotézisek vizsgálatára, ilyen típusú kérdések megválaszolására alkalmasabb módszerek is állnak rendelkezésünkre. Nevezetesen az úgynevezett szórásanalízis módszerei. A kísérlettervezéssel kapcsolatban már találkoztunk olyan problémákkal, amelyek több eloszlásra vonatkoznak. Például a latin négyzetekkel kapcsolatban azzal foglalkoztunk, hogy n növényfajtát hogy lehet egyetlen kísérletben összehasonlítani. Arra a kérdésre, hogy hogyan lehet ilyen célra alkalmas kísérleti elrendezést találni, választottunk az 15.1 pontban. Most rá fogunk térni az ilyen típusú kísérletek eredményeinek kiértékelésével foglalkozó módszerekre.

15.3.1 n normális eloszlású változó várható értékének összehasonlítása

Tegyük fel, hogy egy gépalakatrész élettartamát vizsgáljuk különböző körülmények között. Például kérdezhetjük, hogy különböző olajfajták alkalmazása hogyan befolyásolja az illető alkatrész élettartamát, vagy kérdezhetjük, hogy különböző terhelések hogyan befolyásolják az élettartamot stb.

Mivel a kísérlet eredményeit véletlen tények is befolyásolják, megbízható képet csak akkor kapunk, ha a vizsgált hatások (olajozás, terhelés) mindegyike mellett több kísérletet végzünk (többször mérjük az élettartamot). Tegyük fel, hogy az első hatás eredményét k_1 , a második hatás eredményét k_2 , ..., az n -edik hatás eredményét k_n alkalommal figyeltük meg.* Jelentse

$$\xi_{11}, \xi_{12}, \dots, \xi_{1k_1}$$

az első hatásnak kitett k_1 gépalakatrész élettartamát, hasonlóan

$$\xi_{21}, \xi_{22}, \dots, \xi_{2k_2}$$

$$\xi_{31}, \xi_{32}, \dots, \xi_{3k_3},$$

$$\dots$$

$$\dots$$

$$\dots$$

$$\xi_{n1}, \xi_{n2}, \dots, \xi_{nk_n}$$

a második, harmadik, ..., n -edik hatásnak kitett alkatrész élettartamát.

Abban az esetben, ha a hatások között élettartam szempontjából nincs különbség, akkor

$$\bar{\xi}_1 = \frac{\xi_{11} + \xi_{12} + \dots + \xi_{1k_1}}{k_1}$$

$$\bar{\xi}_2 = \frac{\xi_{21} + \xi_{22} + \dots + \xi_{2k_2}}{k_2}$$

$$\dots$$

$$\dots$$

$$\dots$$

$$\bar{\xi}_n = \frac{\xi_{n1} + \xi_{n2} + \dots + \xi_{nk_n}}{k_n}$$

átlagoknak körülbelül egyenlőknek kell lenniük.

Feltehetjük, hogy a különböző hatások csak az élettartam várható értékét változtatják meg de az élettartam szórása változatlan lesz. Ugyanis a szórás az határozza meg, hogy az illető alkatrészek között azok gyártásánál milyen különbségek lépnek fel. Természetesen feltesz-szük, hogy az élettartamok normális eloszlásúak.

Így kérdésünk a következőképpen fogalmazható. Legyenek

$$\xi_{11}, \xi_{12}, \dots, \xi_{1k_1}$$

független, (m_1, σ) paraméterű, normális eloszlású valószínűségi változók, hasonlóan

$$\xi_{21}, \xi_{22}, \dots, \xi_{2k_2}$$

$$\xi_{31}, \xi_{32}, \dots, \xi_{3k_3}$$

$$\dots$$

$$\dots$$

$$\dots$$

$$\xi_{n1}, \xi_{n2}, \dots, \xi_{nk_n}$$

* Természetesen mind a $k_1 + k_2 + \dots + k_n$ alkalommal ugyanazt az alkatrészt vizsgáljuk (pl. kardán-tengely), csak különböző gépekben (pl. különböző teherautókban).

független (m_2, σ) , (m_3, σ) , ..., (m_n, σ) paraméterű, normális eloszlású változók. Feltesszük hogy nemcsak az egyes mintákban belül függetlenek a mintaelemek, hanem a különböző minták egymás között is függetlenek, azaz a $\xi_{11}, \xi_{12}, \dots, \xi_{1k_1}; \xi_{21}, \xi_{22}, \dots, \xi_{2k_2}, \dots; \xi_{n1}, \xi_{n2}, \dots, \xi_{nk_n}$ valószínűségi változók teljesen függetlenek.

Célunk a $\xi_{11}, \xi_{12}, \dots, \xi_{1k_1}; \xi_{21}, \xi_{22}, \dots, \xi_{2k_2}, \dots; \xi_{n1}, \xi_{n2}, \dots, \xi_{nk_n}$ minták alapján a

$$H_0: m_1 = m_2 = \dots = m_k$$

hipotézis helyes vagy hamis voltának eldöntése.

Bevezetve az

$$\bar{m} = \frac{m_1 + m_2 + \dots + m_n}{n};$$

$$\mu_1 = m_1 - \bar{m}$$

$$\mu_2 = m_2 - \bar{m}$$

$$\dots \dots \dots$$

$$\dots \dots \dots$$

$$\dots \dots \dots$$

$$\mu_n = m_n - \bar{m}$$

jelöléseket, hipotézisünket így is fogalmazhatjuk:

$$\mu_1 = \mu_2 = \dots = \mu_n = 0.$$

A szórásanalízis módszerének lényege, hogy egy ilyen típusú hipotézisről úgy döntünk, hogy a σ szórás különböző módokon becsüljük meg, és a különböző becsléseket hasonlítjuk össze. Vizsgáljuk először a

$$\chi_i^2 = \sum_{j=1}^{k_i} (\xi_{ij} - m_i)^2 \quad (i = 1, 2, \dots, n)$$

valószínűségi változókat. Mint tudjuk, a χ_i^2 valószínűségi változó k_i szabadságfokú, χ^2 eloszlású és χ_i^2 a σ^2 szórásnak egy becslése. Könnyen belátható, hogy

$$(4) \quad \chi_i^2 = \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2 + k_i (\bar{\xi}_i - m_i)^2 \quad \left(\bar{\xi}_i = \frac{\xi_{i1} + \xi_{i2} + \dots + \xi_{ik_i}}{k_i} \right),$$

ahol a jobb oldalon álló első tag egy $k_i - 1$ szabadságfokú — a második tag egy 1 szabadságfokú χ^2 eloszlású valószínűségi változó. Belátható, hogy ezen két valószínűségi változó független.

Az első tagok összegezéséből adódó

$$s_1^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2}{\sum_{i=1}^n (k_i - 1)}$$

valószínűségi változó egy, a szórásnak a várható értékektől független becslése.

A (4) jobb oldalán álló második tagok összegezéséből nyert valószínűségi változót további két tag összegére bonthatjuk. Mivel

$$\sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi}) = 0 \quad \left(\bar{\xi} = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i} \xi_{ij}}{\sum_{i=1}^n k_i} \right)$$

$$\sum_{i=1}^n k_i \mu_i = \sum_{i=1}^n k_i (m_i - \bar{m}) = 0$$

$$\begin{aligned} \text{és} \quad (\bar{\xi}_i - m_i)^2 &= [(\bar{\xi}_i - \bar{\xi}) - (m_i - \bar{m}) + (\bar{\xi} - \bar{m})]^2 = [(\bar{\xi}_i - \bar{\xi} - \mu_i) + (\bar{\xi} - \bar{m})]^2 = \\ &= (\bar{\xi}_i - \bar{\xi} - \mu_i)^2 + (\bar{\xi} - \bar{m})^2 + 2(\bar{\xi} - \bar{m})(\bar{\xi}_i - \bar{\xi} - \mu_i) \end{aligned}$$

adódik, hogy

$$(5) \quad \sum_{i=1}^n k_i (\bar{\xi}_i - m_i)^2 = \sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi} - \mu_i)^2 + \left(\sum_{i=1}^n k_i \right) (\bar{\xi} - \bar{m})^2,$$

ahol $\sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi} - \mu_i)^2$ egy $n - 1$ szabadságfokú és $(\bar{\xi} - \bar{m})^2$ egy 1 szabadságfokú χ^2 eloszlású valószínűségi változó.

Abban az esetben, ha a $\mu_1 = \mu_2 = \dots = \mu_n = 0$ hipotézis teljesül, akkor a

$$\sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi})^2$$

valószínűségi változó $n - 1$ szabadságfokú, χ^2 eloszlású, és így

$$v^2 = \frac{s_2^2}{s_1^2} \quad \left(s_2^2 = \frac{1}{n-1} \sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi})^2; s_1^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2}{\sum_{i=1}^n (k_i - 1)} \right)$$

egy $f_1 = n - 1$, $f_2 = \sum_{i=1}^n (k_i - 1)$ paraméterű, v^2 eloszlású valószínűségi változó. Tehát a $\mu_1 = \mu_2 = \dots = \mu_n = 0$ hipotézisünket elfogadhatjuk, ha a mintából számított v^2 érték 1-hez közel van. Pontosabban, v^2 értékének ismeretében, a v^2 -táblázat felhasználásával tet-szőleges szinten dönthetünk hipotézisünk helyes vagy hamis voltáról.

(4 és 5) egybevetéséből azt kapjuk, hogy

$$(6) \quad \sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - m_i)^2 = \sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2 + \sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi} - \mu_i)^2 + \left(\sum_{i=1}^n k_i \right) (\bar{\xi} - \bar{m})^2,$$

illetve, ha a $\mu_1 = \mu_2 = \dots = \mu_n = 0$ hipotézis helyes, akkor

$$(7) \quad \sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - m_i)^2 = \sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2 + \sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi})^2 + \left(\sum_{i=1}^n k_i \right) (\bar{\xi} - \bar{m})^2.$$

A (6), illetve (7) formulánk jobb oldalán szereplő három tagot a következőképpen interpretálhatjuk. Az első tag az egyes mintákban belüli szorzásokból adódik, így például

$$\sum_{j=1}^{k_1} (\xi_{1j} - \bar{\xi}_1)^2$$

jelenti az első hatásnak kitett alkatrészek élettartamának szórását. A második tag a különböző hatások közti szórást jellemzi, tehát a második tag akkor nagy, ha a különböző hatások között lényeges különbség van. A harmadik tagot *maradék*nak nevezzük, ez a teljes minta-átlagnak a szórása. Így azt mondhatjuk, hogy a teljes mintának (a $k_1 + k_2 + \dots + k_n$ mintaelemnek) a szórása három összetevőre bontható: (i) az egyes mintákon belüli szórások összege, (ii) az egyes minták közötti szórások és (iii) a maradék.

Bebizonyítható, hogy e három tag teljesen független és mindegyike χ^2 eloszlású.

A minta szórásának említett felbontását a következő táblázattal szemléltethetjük:

Szórások		Szabadsági fok
Mintákon belüli szórások	$\sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2$	$\sum_{i=1}^n (k_i - 1)$
Minták közötti szórás	$\sum_{i=1}^n k_i (\bar{\xi}_i - \bar{\xi})^2$	$n - 1$
Maradék	$\left(\sum_{i=1}^n k_i \right) (\bar{\xi} - \bar{m})^2$	1
Összesen	$\sum_{i=1}^n \sum_{j=1}^{k_i} (\xi_{ij} - m_i)^2$	$\sum_{i=1}^n k_i$

A szórásanalízis módszere — mint látni fogjuk — általában azon alapul, hogy a teljes minta tapasztalati szórását felbontjuk különböző összetevőkre, és ezek egymáshoz való viszonyát vizsgáljuk.

Most nézzünk egy konkrét példát.

Tegyük fel, hogy négy különböző hatás esetén vizsgáljuk valamely gépalkatrész élettartamát, és mindegyik hatásra vonatkozóan öt kísérletet végzünk. Kísérletünk eredményei legyenek:

	Élettartam napokban				
1. hatás	190	190	210	190	200
2. hatás	180	200	190	210	190
3. hatás	190	200	210	200	200
4. hatás	180	200	180	180	170

Számításaink eredményét célszerű a kísérletünk eredményeit is feltüntető táblázaton megadni. Így táblázatunkat a következő módon egészíthetjük ki:

	Élettartam napokban	Átlagok $\bar{\xi}_i$	$\sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2$	Szabadsági fok
1. hatás	190 190 210 190 200	196	320	4
2. hatás	180 200 190 210 190	194	520	4
3. hatás	190 200 210 200 200	200	200	4
4. hatás	180 200 180 180 170	182	480	4
Összeg		772	1520	16

Számításainkat folytatva a következőket kapjuk:

$$\bar{\xi} = \frac{\bar{\xi}_1 + \bar{\xi}_2 + \bar{\xi}_3 + \bar{\xi}_4}{4} = \frac{772}{4} = 193,$$

$$\sum_{i=1}^4 k_i (\bar{\xi}_i - \bar{\xi})^2 = 5 \sum_{i=1}^4 (\bar{\xi}_i - \bar{\xi})^2 = 900.$$

$$\text{Ebből } s_2^2 = \frac{1}{3} 5 \sum_{i=1}^4 (\bar{\xi}_i - \bar{\xi})^2 = 300$$

$$\text{és } s_1^2 = \frac{1}{16} \sum_{i=1}^4 \sum_{j=1}^{k_i} (\xi_{ij} - \bar{\xi}_i)^2 = \frac{1520}{16},$$

$$\text{tehát } v^2 = \frac{s_2^2}{s_1^2} = \frac{300 \cdot 16}{1520} = 3,15.$$

Az $f_1 = 3$, $f_2 = 16$ paraméterű v^2 -táblázatban azt találjuk, hogy

$$v^2(3,24) = 0,95,$$

így hipotézisünket 0,05 szinten elfogadhatjuk, vagyis azt mondhatjuk, hogy a különböző hatások az illető gépkatrész élettartamának várható értékét nem befolyásolják. Ha eredményül azt kapjuk, hogy $\mu_1 = \mu_2 = \dots = \mu_n = 0$, hipotézisünket el kell vetnünk, akkor általában még további vizsgálatra is szükség van. Ugyanis abból, hogy a várható értékek nem egyenlők, még nem feltétlenül világos, hogy ezt a különbséget mi okozza; nem tudjuk például, hogy a legnagyobb átlag (a legnagyobb $\bar{x}_i$) nagyobb-e szignifikánsan a többinél, vagy pedig a legkisebb szignifikánsan kisebb (esetleg mindkettő). Ha például célunk annak bizonyítása, hogy valamely hatás esetén a várható érték szignifikánsan nagyobb a többi várható értéknél, és már beláttuk, hogy a várható értékek különbözőek, akkor például a legjobbnak feltételezett mintát elhagyva, az eredeti módszer ismételt alkalmazásával vizsgálhatjuk azt a hipotézist, hogy a többi minta várható értéke egyenlő. Ha e hipotézis elfogadható, akkor már állíthatjuk, hogy a legnagyobb átlagú hatás várható értéke valóban nagyobb a többi hatás várható értékénél. E megjegyzésünkkel — anélkül, hogy a részletekbe mennénk — fel akartuk hívni a figyelmet arra, hogy ha a várható értékek között szignifikáns eltérés mutatkozik, akkor még részletes vizsgálatokra lehet szükség ahhoz, hogy a várható értékek nagyságrendjéről végleges következtetéseket vonhassunk le.

15.3.2 n normális eloszlású változó várható értékének összehasonlítása különböző hatások figyelembevétele esetén

Az előző pontban már említettük, hogy valamely gépkatrész élettartamát számos tényező befolyásolja. Az előző pontban feltételeztük, hogy módunkban van valamely hatásfajta lényegében homogén körülmények között figyelni. Ez sok esetben igen nehezen megvalósítható követelmény.

A nehézséget szemlélteti a következő példa. Célunk eldönteni, hogy különböző oktánszámú benzinek esetén hogy működik valamely autótípus (például mi a fogyasztása, végsebessége stb.). Természetesen a fogyasztás nemcsak a benzin oktánszámától, hanem a terheléstől, a sebességtől és más tényezőktől is függ. Viszonylag könnyű kérdéseinkre (hogy függ a fogyasztás az oktánszámtól) választ adni, ha lehetőségünk van arra, hogy autónkat (autóinkat) egyforma körülmények között működtessük, és csak a benzin oktánszámát változtassuk. Ilyen ideális kísérleti körülmények biztosítása általában csak úgy lehetséges, ha autóinkat a kísérlet közben ténylegesen nem használjuk. Ez egyrészt a kísérlet költségét emeli erősen, másrészt esetleg a kísérlet megbízhatóságát is csökkenti az a körülmény, hogy a valóságos körülményektől lényegesen eltérő helyzetben próbáltuk ki gépkocsinkat.

Az egyszerűség kedvéért tételezzük fel, hogy a sebességtől eltekintve homogén körülményeket tudunk biztosítani; azaz csak két tényezővel kell számolnunk, az oktánszámmal és a sebességgel (és pl. a terhelések a kísérlet folyamán lényegében változatlanok).

Legyenek a kipróbálásra kerülő oktánszámok a_1, a_2, a_3 , és tegyük fel, hogy autóink b_1, b_2, b_3, b_4 km/h sebességgel szoktak közlekedni. Az a_1, a_2, a_3 oktánszámú benzinek mindegyikét 30 esetben próbáljuk ki, de az egyes esetekben a sebességek véletlentől függőek. Kísérletünk eredményeit a következő táblázatok adják:

Oktánszám: a_1

Sebesség	Fogyasztás: liter/100 km								Átlag
b_1	22,5	22,0	22,5	22,5	22,5	23,0	22	—	22,43
b_2	22,0	22,0	22,0	21,5	22,0	22,5	21	21,0	21,75
b_3	22,0	22,0	21,5	21,0	21,0	21,5	22	22,5	21,69
b_4	23,0	22,5	22,5	23,0	22,5	23,5	23	—	22,86

Oktánszám: a_2

Sebesség	Fogyasztás: liter/100 km								Átlag
b_1	21,0	20,5	21,0	21	21,0	20,5	20,0	—	20,71
b_2	21,0	20,5	20,0	20	20,5	20,0	19,5	19	20,11
b_3	20,5	20,0	20,5	20	21,0	20,0	20,5	20	20,31
b_4	21,0	21,5	20,5	21	21,5	20,0	—	—	20,92

Oktánszám: a_3

Sebesség	Fogyasztás: liter/100 km								Átlag
b_1	21,0	21,0	20	20,5	21,0	20,0	20,5	21,0	20,63
b_2	20,5	21,0	20	20,0	19,5	20,0	19,0	20,5	20,06
b_3	20,5	20,0	20	19,5	20,0	20,5	21,0	—	20,21
b_4	21,0	20,5	21	21,5	20,0	21,0	—	—	20,83

Kísérletünk célja annak eldöntése, hogy az oktánszám és a sebesség hatással van-e a fogyasztásra.*

Problémánk általánosan a következő módon fogalmazható. Kétfajta hatás eredményét kívánjuk kipróbálni. Az első hatást (pl. oktánszám) n , a másodikat (pl. sebesség) m szinten vizsgáljuk. Ha az első hatást az i -edik, a második hatást a j -edik szinten vizsgáljuk, akkor a kísérlet eredménye egy μ_{ij} várható értékű és σ szórású** normális eloszlású valószínűségi változó.

Tegyük fel, hogy az i, j szintpár ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$) alkalmazása esetén k_{ij} kísérletet végeztünk. E kísérletek eredményei legyenek

$$\xi_{ij1}, \xi_{ij2}, \dots, \xi_{ijk_{ij}}.$$

Célunk e kísérleti eredmények felhasználásával a

$$H_0: \mu_{11} = \mu_{12} = \dots = \mu_{1m} = \mu_{21} = \mu_{22} = \dots = \mu_{2m} = \dots = \mu_{n1} = \mu_{n2} = \dots = \mu_{nm}$$

hipotézis helyes vagy hamis voltáról dönteni.

Ismételten a szórásanalízis módszerét használjuk, azaz a minta teljes tapasztalati szórását, az

$$s^2 = \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} (\xi_{ijv} - \mu_{ij})^2,$$

mennyiséget bontjuk fel egy összegre, és ezen összeg tagjait hasonlítjuk össze.

Vezessük be a következő jelöléseket:

$$\begin{aligned} \sum_{j=1}^m k_{ij} &= N_i, \quad \sum_{i=1}^n N_i = N, \\ \frac{1}{k_{ij}} \sum_{v=1}^{k_{ij}} \xi_{ijv} &= \bar{\xi}_{ij}, \quad \frac{1}{N_i} \sum_{j=1}^m k_{ij} \bar{\xi}_{ij} = \frac{1}{N_i} \sum_{j=1}^m \sum_{v=1}^{k_{ij}} \xi_{ijv} = \bar{\xi}_i, \\ \frac{1}{N} \sum_{i=1}^n N_i \bar{\xi}_i &= \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} \xi_{ijv} = \bar{\xi}, \\ \frac{1}{m} \sum_{j=1}^m \mu_{ij} &= \bar{\mu}_i, \quad \frac{1}{n} \sum_{i=1}^n \bar{\mu}_i = \bar{\mu}, \\ v_{ij} &= \mu_{ij} - \bar{\mu}_i, \quad v_i = \bar{\mu}_i - \bar{\mu}. \end{aligned}$$

Tehát N_i jelenti az első hatás i -edik szintjére vonatkozó kísérletek számát, N jelenti az összes kísérletek számát, $\bar{\xi}_{ij}$ az i, j szintpárra végzett kísérletek eredményeinek átlaga, $\bar{\xi}_i$ az első hatás i -edik szintjére vonatkozó kísérletek átlaga, $\bar{\xi}$ az összes kísérletek átlaga. Ha az első hatás i -edik szintjén a második hatások nem befolyásolják a várható értékeket, azaz, ha

$$\mu_{i1} = \mu_{i2} = \dots = \mu_{im},$$

$$\text{akkor} \quad v_{i1} = v_{i2} = \dots = v_{im} = 0,$$

ismeretes, hogy az oktánszám növelése bizonyos határ után nem csökkenti a fogyasztást. Például lehet kísérletünk célja ennek a határnak a megtalálása.

** Most is feltesszük, hogy a különböző hatások a szórást nem befolyásolják.

és, ha az első hatás n szintje között nincs különbség, akkor

$$v_1 = v_2 = \dots = v_n = 0.$$

Számítsuk ki a szóban forgó mennyiségeket az oktánszámra és sebességre vonatkozó példánkban:

$$N_1 = N_2 = N_3 = 30 \quad N = 90$$

$$\bar{\xi}_{11} = \frac{22,5 + 22 + 22,5 + 22,5 + 22,5 + 23 + 22}{7} = 22,43,$$

$$\bar{\xi}_{12} = 21,75, \quad \bar{\xi}_{13} = 21,69, \quad \bar{\xi}_{14} = 22,86$$

$$\bar{\xi}_{21} = 20,71, \quad \bar{\xi}_{22} = 20,11, \quad \bar{\xi}_{23} = 20,31, \quad \bar{\xi}_{24} = 20,92$$

$$\bar{\xi}_{31} = 20,62, \quad \bar{\xi}_{32} = 20,05, \quad \bar{\xi}_{33} = 20,21, \quad \bar{\xi}_{34} = 20,83.$$

$$\bar{\xi}_1 = \frac{7 \cdot 22,43 + 8 \cdot 21,75 + 8 \cdot 21,69 + 7 \cdot 22,86}{30} = \frac{664,5}{30} = 22,15$$

$$\bar{\xi}_2 = 20,47 \quad \bar{\xi}_3 = 20,24$$

$$\bar{\xi} = 21,01.$$

Jelöléseinket és a

$$\sum_{v=1}^{k_{ij}} (\xi_{ijv} - \bar{\xi}_{ij}) = 0, \quad \sum_{j=1}^m k_{ij} (\bar{\xi}_{ij} - \bar{\xi}) = 0, \quad \sum_{j=1}^m k_{ij} v_{ij} = 0,$$

$$\sum_{i=1}^n N_i (\bar{\xi}_i - \bar{\xi}) = 0, \quad \sum_{i=1}^n N_i v_i = 0$$

nyilvánvaló formulákat felhasználva, a szórásfelbontást a következő módon végezhetjük el:

$$\begin{aligned} s^2 &= \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} (\xi_{ijv} - \mu_{ij})^2 = \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} (\xi_{ijv} - \bar{\xi}_{ij} + \bar{\xi}_{ij} - \mu_{ij})^2 = \\ &= \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} (\xi_{ijv} - \bar{\xi}_{ij})^2 + \sum_{i=1}^n \sum_{j=1}^m k_{ij} (\bar{\xi}_{ij} - \mu_{ij})^2 \end{aligned}$$

és

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^m k_{ij} (\bar{\xi}_{ij} - \mu_{ij})^2 &= \sum_{i=1}^n \sum_{j=1}^m k_{ij} (\bar{\xi}_{ij} - \bar{\xi}_i + \bar{\xi}_i - \bar{\mu}_i + \bar{\mu}_i - \mu_{ij})^2 = \\ &= \sum_{i=1}^n \sum_{j=1}^m k_{ij} (\bar{\xi}_{ij} - \bar{\xi}_i - v_{ij})^2 + \sum_{i=1}^n N_i (\bar{\xi}_i - \bar{\mu}_i)^2, \end{aligned}$$

továbbá

$$\sum_{i=1}^n N_i (\bar{\xi}_i - \bar{\mu}_i)^2 = \sum_{i=1}^n N_i (\bar{\xi}_i - \bar{\xi} + \bar{\xi} - \bar{\mu}_i + \bar{\mu}_i - \bar{\mu})^2 = \sum_{i=1}^n N_i (\bar{\xi}_i - \bar{\xi} - v_i)^2 + N(\bar{\xi} - \bar{\mu})^2.$$

Eredményeinket összevetve, az s^2 tapasztalati szórást a következő módon bonthatjuk fel:

$$\begin{aligned} s^2 &= \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} (\xi_{ijv} - \mu_{ij})^2 = \sum_{i=1}^n \sum_{j=1}^m \sum_{v=1}^{k_{ij}} (\xi_{ijv} - \bar{\xi}_{ij})^2 + \\ &+ \sum_{i=1}^n \sum_{j=1}^m k_{ij} (\bar{\xi}_{ij} - \bar{\xi}_i - v_{ij})^2 + \sum_{i=1}^n N_i (\bar{\xi}_i - \bar{\xi} - v_i)^2 + N(\bar{\xi} - \bar{\mu})^2. \end{aligned}$$

Belátható, hogy a jobb oldalon szereplő valószínűségi változók függetlenek, és rendre $N - mn, mn - n, n - 1, 1$ szabadságfokú χ^2 eloszlással rendelkeznek.

Az első tag az osztályokon belüli szórásokból tevődik össze,

$$\text{így} \quad \sum_{k_{ij}}^r (\xi_{ijv} - \bar{\xi})^2$$

jelenti az első hatás i -edik és a második hatás j -edik szintje mellett végzett kísérletek tapasztalati szórását.

A második tag jelenti az első hatás adott szintje mellett, a második hatás különböző szintjei közötti szórások összegét. A harmadik tag az első hatás egyes szintjei közötti szórásokat adja. Míg a negyedik tag a maradék.

E szórásfelbontást táblázatban a következőképpen foglalhatjuk össze:

Szórások	Osztályokon belüli szórások	A második hatás szintjei közötti szórások	Az első hatás szintjei közötti szórás	Maradék	Összesen
Szabadsági fok	$N - mn$	$mn - n$	$n - 1$	1	N
	$s_1^2 = \sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n (\xi_{ijv} - \bar{\xi}_i)^2$	$s_2^2 = \sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n k_{ij} (\bar{\xi}_{ij} - \bar{\xi}_j - p_{ij})^2$	$s_3^2 = \sum_{i=1}^r N_i (\bar{\xi}_i - \bar{\xi} - p_i)^2$	$s_4^2 = N(\bar{\xi} - \mu)^2$	$s^2 = \sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n (\xi_{ijv} - \mu_{ij})^2$

Ha az első hatást fixen tartva a második hatás szintjei között nincs különbség, azaz, ha $v_{ij} = 0$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$), akkor az

$$\frac{s_2^2}{s^2} = \frac{\sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n k_{ij} (\bar{\xi}_{ij} - \bar{\xi}_i)^2}{\sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n (\xi_{ijv} - \bar{\xi}_{ij})^2}$$

valószínűségi változó $f_1 = mn - n, f_2 = N - mn$ paraméterű, v^2 eloszlású. Így a v^2 táblázat felhasználásával és s_2^2/s_1^2 értékeknek ismeretében dönthetünk a

$$H_0: v_{ij} = 0 \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$$

hipotézis helyes vagy hamis voltáról.

Ha az első hatás egyes szintjei között sincs különbség, azaz ha

$$v_1 = v_2 = \dots = v_n = 0,$$

akkor az

$$\frac{s_3^2}{s_1^2} = \frac{\sum_{i=1}^r N_i (\bar{\xi}_i - \bar{\xi})^2}{\sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n (\xi_{ijv} - \bar{\xi}_{ij})^2}$$

valószínűségi változó $f_1 = n - 1, f_2 = N - mn$ paraméterű, v^2 eloszlású, így a

$$H_0: v_1 = v_2 = \dots = v_n = 0$$

hipotézis helyes vagy hamis voltáról is tudunk dönteni kísérleti eredményeink birtokában. Térjünk vissza példánkra. Mindenekelőtt számítsuk ki a szórás felbontásban szereplő mennyiségeket. Kapjuk:

$$s_1^2 = \sum_{i=1}^r \sum_{j=1}^m \sum_{k_{ij}}^n (\xi_{ijv} - \bar{\xi}_{ij})^2 = 19,78$$

$$s_2^2 = 12,51$$

$$s_3^2 = 65,40.$$

Így esetünkben — a $v_{ij} = 0$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$) hipotézist feltételezve — a szórásfelbontási táblázat a következőképpen alakul:

Szórások	Osztályokon belüli szórások	A második hatás szintjei közötti szórások	Az első hatás szintjei közötti szórás	Szabadsági fok
	$s_1^2 = 19,78$	$s_2^2 = 12,51$	$s_3^2 = 65,40$	2
	79	9		

Tehát a $v_{ij} = 0$ hipotézis vizsgálatához az

$$\frac{s_2^2}{s_1^2} = \frac{12,51}{19,78} = 0,63$$

mennyiséget kell tekintenünk. Az $f_1 = 9, f_2 = \infty$ paraméterű táblázatban azt találjuk, hogy ez az érték még nem tér el szignifikánsan az 1-től. Így adataink alapján nem állíthatjuk, hogy a sebesség a fogyasztást befolyásolja.

A $v_i = 0$ hipotézis vizsgálatához az

$$\frac{s_3^2}{s_1^2} = \frac{65,40}{19,78} = 3,32$$

menyiséget kell tekintenünk. Az $f_1 = 2, f_2 = \infty$ paraméterű táblázatban azt találjuk, hogy

$$v^2(0,95) = 3,00.$$

Igy 0,95 szinten nem fogadhatjuk el azt a hipotézist, hogy $v_i = 0$, tehát azt mondhatjuk, hogy az oktatászám valóban befolyásolja a fogyasztást. Természetesen további vizsgálatot igényelne annak eldöntése, hogy mely oktatászámok hatása között van szignifikáns különbség, de részletes statisztikai analízis nélkül is világos, hogy az a_1 oktatászámú benzín az, ami mellett a fogyasztás szignifikánsan nagyobb. Aki részleteiben is végigkövette számításaink menetét, láthatja, hogy bár csak két hatást, és azokat is csak 3, illetve 4 szinten vizsgáltuk, a kísérlet kitértékése mégis farradtságos számításai nagyobb — kísérletek kitértékéséhez feltétlenül szükség van számológépekre. Azonban még nagy teljesítményű számológépek esetében is fontos, hogy a számítás menetét jól tervezzük meg. Az, hogy a számítást hogy célszerű végrehajtani, erősen függ a rendelkezésünkre álló számológéptől is. Ezért erre a kérdésre itt nem térünk ki, csak utalunk a 15.2.1.2 pontban mondottakra.

15.3.3 Többfaktoros kísérletek kitértékése

Már a kísérletek megtervezésénél fontos, hogy gondosan járjunk el, és be is mutattunk néhány példát arra vonatkozóan, hogy bizonyos esetekben miképpen célszerű megtervezni kísérletünket. Most azt vizsgáljuk, hogy a 15.1 pontban bemutatott kísérlet tervek alkalmas-e azokra, hogy kapott statisztikai adatokat miként lehet a szórásanalízis módszerével kitértékálni.

15.3.3.1 Latin négyzetes kísérletek kitértékése

A 15.1.2.1 pontban a következő példát említettük a latin négyzetekkel kapcsolatban. n kukoricafajtát akarunk összehasonlítani terméshozam szempontjából, de a kísérletünket úgy akarjuk megtervezni, hogy a talaj inhomogenitása minél kevésbé zavarja a kísérlet eredményét. Ekkor $n = 4$ esetben (a kukoricafajtákat F_1, F_2, F_3, F_4 -gyel jelölve) például a következő vetéstervet készíthetjük:

F_3	F_4	F_1	F_2
F_1	F_3	F_4	F_2
F_2	F_1	F_3	F_4

Azaz a vetésterületet 4 sorra és 4 oszlopra bontjuk, és például a második sor 3. oszlopába az F_1 kukoricafajtát vetjük. Jeleltse ξ_{ij} ($i = 1, 2, 3, 4; j = 1, 2, 3, 4$) az i -edik sor j -edik oszlopában a terméseredményt. Természetesen most is feltehetjük, hogy ξ_{ij} -k független, egyforma szórású, normális eloszlású valószínűségi változók.

Vezessük be a következő jelöléseket:

$$\bar{\xi}_{F_1} = \frac{\xi_{11} + \xi_{21} + \xi_{31} + \xi_{41}}{4}; \quad \bar{\xi}_{F_2} = \frac{\xi_{12} + \xi_{22} + \xi_{32} + \xi_{42}}{4};$$

$$\bar{\xi}_{F_3} = \frac{\xi_{13} + \xi_{23} + \xi_{33} + \xi_{43}}{4}; \quad \bar{\xi}_{F_4} = \frac{\xi_{14} + \xi_{24} + \xi_{34} + \xi_{44}}{4};$$

$$\bar{\xi}_i = \frac{\xi_{i1} + \xi_{i2} + \xi_{i3} + \xi_{i4}}{4} \quad (i = 1, 2, 3, 4)$$

$$\bar{\xi}_j = \frac{\xi_{1j} + \xi_{2j} + \xi_{3j} + \xi_{4j}}{4} \quad (j = 1, 2, 3, 4)$$

$$\bar{\xi} = \frac{\xi_{11} + \xi_{12} + \xi_{13} + \xi_{14} + \xi_{21} + \xi_{22} + \xi_{23} + \xi_{24} + \xi_{31} + \xi_{32} + \xi_{33} + \xi_{34} + \xi_{41} + \xi_{42} + \xi_{43} + \xi_{44}}{16}$$

Célunk a $H_0: m_{F_1} = m_{F_2} = m_{F_3} = m_{F_4}$ azaz $\bar{\xi}_{F_1} = \bar{\xi}_{F_2} = \bar{\xi}_{F_3} = \bar{\xi}_{F_4}$ rendre az F_1, F_2, F_3, F_4 kukoricafajták átlagtermését jelenti. $\bar{\xi}_j$ a j -edik oszlop átlagtermését jelenti.

hipotézis helyes vagy hamis voltának ellenőrzése. E hipotézist ismételen a szórásanalízis módszerének segítségével ellenőrizhetjük, azaz a tapasztalati szórást különböző szempontoknak megfelelően bontjuk fel, és ezen szórásokat hasonlítjuk össze. Most nem fogjuk elvégezni a szórásfelbontás egészét — bár ez teljesen az előzőekhez hasonlóan megtehető —, csak végérményként megadjuk a szórásfelbontó táblázatot.

Szórások	Oszlopok közötti szórás	Sorok közötti szórás	Fajták közötti szórás	Maradék	Összesen
$n-1$	$s_o^2 = n \sum_{j=1}^4 (\bar{\xi}_j - \bar{\xi})^2$	$s_s^2 = n \sum_{i=1}^4 (\bar{\xi}_i - \bar{\xi})^2$	$s_f^2 = n \sum_{j=1}^4 (\bar{\xi}_{F_j} - \bar{\xi})^2$	$s_m^2 = n \sum_{j=1}^4 \sum_{i=1}^4 (\xi_{ij} - \bar{\xi}_i - \bar{\xi}_j + \bar{\xi})^2 + 2\bar{\xi}^2$	$s^2 = n \sum_{j=1}^4 \sum_{i=1}^4 (\xi_{ij} - \bar{\xi})^2$
Szabadsági fok	$n-1$	$n-1$	$n-1$	$(n-1)(n-2)$	n^2-1

Esetünkben $n = 4$.

Megjegyzés. s_m^2 kifejezésében a ξ_{F_k} úgy értendő, hogy k értékét a latin négyzet i -edik sorának j -edik oszlopában szereplő fajta indexével megegyezően kell megválasztani.

Ha $H_0: m_{F_1} = m_{F_2} = m_{F_3} = m_{F_4}$

hipotézisünk helyes, akkor a

$$v^2 = \frac{s_f^2}{s^2}$$

valószínűségi változó $f_1 = n-1$; $f_2 = n^2-1$ paraméterű, v^2 eloszlású. Így adataink alapján dönthetünk hipotézisünk helyes vagy hamis voltáról.

Bizonyos esetekben nemcsak azt akarjuk vizsgálni, hogy a fajták között van-e különbség, hanem azt is tudni akarjuk, hogy az oszlopok, illetve sorok között van-e különbség. Más szóval a

$$H_0^{(1)}: M(\xi_{.1}) = M(\xi_{.2}) = M(\xi_{.3}) = M(\xi_{.4}),$$

illetve a

$$H_0^{(2)}: M(\xi_{1.}) = M(\xi_{2.}) = M(\xi_{3.}) = M(\xi_{4.})$$

hipotézis helyes vagy hamis voltáról akarunk dönteni.

A $H_0^{(1)}$ hipotézis teljesülése esetén a

$$v^2(1) = \frac{s_0^2}{s^2}$$

valószínűségi változó $f_1 = n-1$; $f_2 = n^2-1$ paraméterű v^2 eloszlású, míg a $H_0^{(2)}$ hipotézis teljesülése esetén a

$$v^2(2) = \frac{s_s^2}{s^2}$$

valószínűségi változó $f_1 = n-1$; $f_2 = n^2-1$ paraméterű, v^2 eloszlású. Így a v^2 -próbával $H_0^{(1)}$ és $H_0^{(2)}$ hipotézisünkről is dönthetünk.

Említett példánkban a $H_0^{(1)}$ és $H_0^{(2)}$ hipotézisek teljesülése azt jelenti, hogy a kísérleti talaj homogén. Más esetekben a sor-, illetve oszlophatás jelentősége egész más. Nézzük a következő példát.

Tegyük fel, hogy n egyforma típusú gépen n különböző cikket gyártunk. Tudni akarjuk, hogy melyik cikket legkifizetődőbb gyártani, azaz egy időegység alatt melyik gyártása esetén nyerjük a legnagyobb hasznot. A kísérlet elvégzésére több lehetőség kínálkozik. Lehetséges például egy munkanapon keresztül mindegyik géppel egy más terméket gyártani, ekkor azonban féltő, hogy a kísérlet eredménye nemcsak a termékek közti, hanem a gépek (gépeken dolgozó munkások) közötti különbségektől is erősen függ. Másik lehetőség, hogy a munkanapot felosztjuk n részre, és az egyes részekben mindegyik géppel ugyanazt a terméket gyártjuk. Ez a módszer természetesen kiküszöböli a gépek közötti különbségekből eredő hibát, de hibája, hogy nem számol azzal, hogy a munkanap végére — pl. a munkások fáradtsága miatt — csökken a termelékenység.

Így leghelyesebben akkor járunk el, ha latin négyzetszerű elrendezést készítünk. Például $n = 4$ esetben és az egyes termékeket t_1, t_2, t_3, t_4 -gyel jelölve a következő kísérleti tervet készíthetjük:

	1. gép	2. gép	3. gép	4. gép
1. szakasz	t_1	t_2	t_3	t_4
2. szakasz	t_2	t_3	t_4	t_1
3. szakasz	t_3	t_4	t_1	t_2
4. szakasz	t_4	t_1	t_2	t_3

A kísérlet kiértékelése pontosan ugyanúgy történhet, mint azt az előzőekben, a kukorica-fajtákkal kapcsolatos kísérleteknél láttuk, azonban most a $H_0^{(1)}$ hipotézis vizsgálatával arra a kérdésre kaphatunk választ, hogy a munkások fáradtsága befolyásolja-e számottevően a termelést. Míg a $H_0^{(2)}$ hipotézis vizsgálata a termelésben részvevő gépek (illetve munkások) közti különbségeket világítja meg.

Természetesen egy ilyen latin négyzet elrendezésű kísérlet megszervezése (anyagellátás stb.) igen nehéz és költséges, de képes a felvetett kérdésekre viszonylag rövid idő alatt választ adni.

15.3.3.2 Faktoriális kísérletek

Mint említettük, a faktoriális kísérletekben (15.1.2.4 pont) egyszerre több hatást vizsgálunk, amelyek mindegyike több szinten fordul elő. Így például $2 \times 2 \times 2$ -es faktoriális kísérlet az, amelyben 3 különböző faktor szerepel, mindegyike 2 szinten, és a szóba kerülő 8 lehetőség mindegyikét vizsgálni akarjuk. Ebben az értelemben a 15.3.2 pontban tárgyalt autókísérlet egy 3×4 -es faktoriális kísérletnek tekinthető.

Most példaként vizsgáljuk meg egy $2 \times 2 \times 2$ -es faktoriális kísérlet kiértékelésének módját, azaz adjuk meg a szórásfelbontó táblázatot.

Legyen a kísérletben szereplő 3 faktor A, B, C . Tegyük fel, hogy az A faktor A_1, A_2 szinten, a B faktor B_1, B_2 szinten, míg a C faktor C_1, C_2 szinten fordul elő. Kísérletünk eredményeit a következő táblázatban foglalhatjuk össze:

	A_1		A_2	
	B_1	B_2	B_1	B_2
C_1	ξ_{111}	ξ_{121}	ξ_{211}	ξ_{221}
C_2	ξ_{112}	ξ_{122}	ξ_{212}	ξ_{222}

Itt például $\bar{\xi}_{122}$ jelenti azon kísérletek átlagát, amelyeket az A hatás A_1 , a B hatás B_2 és a C hatás C_2 szintje mellett vizsgáltunk.

Ha például azt akarjuk tudni, hogy az A hatás két szintje között van-e különbség, akkor nyilván a

$$\bar{\xi}_{1..} - \bar{\xi}_{2..} \quad \left(\bar{\xi}_{1..} = \frac{1}{4} \sum_{j=1}^2 \sum_{k=1}^2 \bar{\xi}_{1jk}; \quad \bar{\xi}_{2..} = \frac{1}{4} \sum_{j=1}^2 \sum_{k=1}^2 \bar{\xi}_{2jk} \right)$$

valószínűségi változót kell vizsgálnunk.

Hasonlóan, például $\bar{\xi}_{11.} - \bar{\xi}_{21.}$, illetve $\bar{\xi}_{12.} - \bar{\xi}_{22.}$ jellemzi az A hatás 2 szintje közötti különbséget, ha a B hatást a B_1 , illetve a B_2 szinten alkalmazzuk. Így

$$\frac{1}{2} (\bar{\xi}_{11.} - \bar{\xi}_{21.} - \bar{\xi}_{12.} + \bar{\xi}_{22.})$$

jellemzi A és B kölcsönhatását.

Feltesszük, hogy mind a nyolc esetre vonatkozóan n kísérletet végzünk. Jelentsé ξ_{ijkv} ($v = 1, 2, \dots, n; i = 1, 2; j = 1, 2; k = 1, 2$) az $A_i B_j C_k$ feltételek mellett végzett v -edik kísérlet eredményét. Ekkor nyilván

$$\bar{\xi}_{ijk} = \frac{1}{n} \sum_{v=1}^n \xi_{ijkv}.$$

A szórásfelbontó táblázatot ismét bizonyítás nélkül közöljük.

Szórások		Szabadsági fok
A_1 és A_2 közötti szórás	$s_5^2 = 2n(\bar{\xi}_{2..} - \bar{\xi}_{1..})^2$	1
B_1 és B_2 közötti szórás	$s_7^2 = 2n(\bar{\xi}_{.2.} - \bar{\xi}_{.1.})^2$	1
C_1 és C_2 közötti szórás	$s_8^2 = 2n(\bar{\xi}_{..2} - \bar{\xi}_{..1})^2$	1
A és B kölcsönhatása	$s_9^2 = \frac{n}{2} (\bar{\xi}_{22.} - \bar{\xi}_{12.} + \bar{\xi}_{11.} - \bar{\xi}_{21.})^2$	1
A és C kölcsönhatása	$s_4^2 = \frac{n}{2} (\bar{\xi}_{2.2} - \bar{\xi}_{1.2} + \bar{\xi}_{1.1} - \bar{\xi}_{2.1})^2$	1
B és C kölcsönhatása	$s_3^2 = \frac{n}{2} (\bar{\xi}_{.22} - \bar{\xi}_{.12} + \bar{\xi}_{.11} - \bar{\xi}_{.21})^2$	1
A, B, C kölcsönhatása	$s_2^2 = \frac{n}{8} (\bar{\xi}_{222} - \bar{\xi}_{122} - \bar{\xi}_{212} + \bar{\xi}_{112} - \bar{\xi}_{221} + \bar{\xi}_{121} + \bar{\xi}_{211} - \bar{\xi}_{111})^2$	1

Szórások		Szabadsági fok
Osztályon belüli szórás	$s_1^2 = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 \sum_{v=1}^n (\xi_{ijkv} - \bar{\xi}_{ijk})^2$	$8(n-1)$
Összesen	$s^2 = \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 \sum_{v=1}^n (\xi_{ijkv} - \bar{\xi}_{...})^2$	$8n-1$

E szórásfelbontó táblázat alkalmazásával, ha például azt a hipotézist akarjuk vizsgálni, hogy az A hatás két szintje között nincs-e szignifikáns eltérés, azaz a

$$H_0: M(\bar{\xi}_{1..}) = M(\bar{\xi}_{2..})$$

hipotézis helyes vagy helytelen voltáról akarunk dönteni, akkor a

$$v^2 = \frac{s_8^2}{s^2}$$

valószínűségi változót kell vizsgálnunk. Nevezetesen, ha a H_0 hipotézis helyes, akkor a v^2 valószínűségi változó $f_1 = 1$, $f_2 = 8n-1$ paraméterű, v^2 eloszlású, és így a v^2 -próba segítségével dönthetünk hipotézisünk helyes vagy hamis voltáról.

15.4 ELOSZLÁS- ÉS SŰRŰSÉGFÜGGVÉNYEK MEGHATÁROZÁSA. (NEMPARAMÉTERES STATISZTIKÁK, RENDEZETT MINTÁK)

A 15.1, 15.2 és 15.3 pontban mindig feltételeztük, hogy a szereplő valószínűségi változó (változók) eloszlását ismerjük, és a mintából csak a változó (változók) paramétereire (általában várható érték és szórás) akarunk következtetéseket levonni. Általában abból indultunk ki, hogy a szereplő valószínűségi változók normális eloszlásúak, ugyanis az elméleti megfontolások és a tapasztalati adatok is arra mutatnak, hogy a gyakorlatban előforduló folytonos eloszlású valószínűségi változók túlnyomóan normális eloszlásúak.

Ebben a fejezetben abból indulunk ki, hogy a változók eloszlásáról nem tudunk semmit, és célunk a kapott minta alapján becslést adni a változók eloszlás- és sűrűségfüggvényére, illetve a minta alapján dönteni arról, hogy a változók eloszlására vonatkozó hipotézisek helyesek-e. Az ilyen típusú vizsgálatokat nemparaméteres próbáknak nevezzük, szemben azokkal a vizsgálatokkal, amelyek kizárólag az eloszlás paramétereire vonatkoznak. E vizsgálatoknál általában az ún. rendezett minták elméletét használjuk, (az elnevezés magyarázatára később visszatérünk).

15.4.1 A sűrűségfüggvény becslése

15.4.1.1 Hisztogram

A 15.2.4.4 pontban foglalkoztunk két gépcsoport által végzett szántás mélységének összehasonlításával. Azt mondtuk, hogy mérjük le a szántás mélységét mindkét gépcsoport esetében, véletlenszerűen kiválasztott pontokban. Számítsuk ki a két átlagmélységet. Ha ezek mindkettője megfelel az előírtnak, akkor számítsuk ki a két tapasztalati szórást és ezeket hasonlítsuk össze. Azt a gépcsoportot tekintjük a jobbnak, amelynél a szántás-mélység tapasztalati szórása kisebb, ugyanis ez nyilván egyenletesebben dolgozik.

Tegyük fel például, hogy mindkét gépcsoport szántásmélységét 10 helyen ellenőriztük és a következő adatokat kaptuk:

	Szántásmélység
1. gépcsoport	20, 20, 20, 18, 20, 20, 22, 20, 20, 20
2. gépcsoport	20, 19, 20, 21, 19, 21, 19, 19, 21, 21

Ezen adatok alapján az átlagmélységre és a szórásra a következőket kapjuk:

	Átlag	Szórás
1. gépcsoport	$\bar{x}_1 = 20$	$s_1 = \frac{2\sqrt{2}}{\sqrt{9}}$
2. gépcsoport	$\bar{x}_2 = 20$	$s_2 = \frac{2\sqrt{2}}{\sqrt{9}}$

azaz a két gépcsoport által végzett szántás mélységének átlaga és tapasztalati szórása megegyezik. Mégis nehéz lenne azt állítani, hogy a két gépcsoport egyformán dolgozik, hiszen az első gépcsoport általában 20 cm mélyen szánt, és csak ritkán tér el ettől lényegesen, míg a második gépcsoport általában a 20 cm átlagmélységtől eltérő szántásmélységekkel dolgozik.*

Két ilyen típusú eloszlás közötti különbséget egy közgazdasági jellegű példán még jobban meg tudjuk világítani. Tegyük fel, hogy két szövetkezetben vizsgáljuk a tagok egy ágazat-

* Természetesen a gyakorlatban ilyen adatok nem fordulnak elő. Különösen a második gépcsoportra vonatkozó adatok teljesen irreálisak. Ugyanis a szántásmélység eloszlása általában normális (vagy közel normális), és a második gépre vonatkozó adatok alapján a szántásmélység feltehetően különbözik a normális eloszlástól.

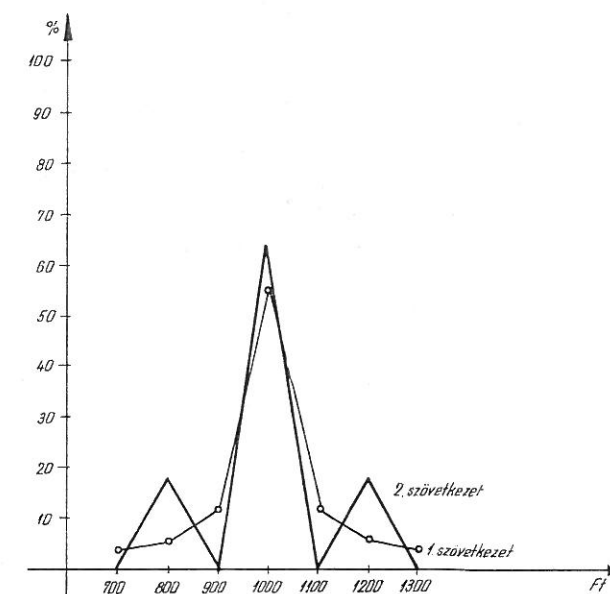
ból szerzett jövedelemeloszlását. Az adatokat a következő táblázat szemlélteti:

	Havi jövedelem						
	700	800	900	1000	1100	1200	1300
1. szövetkezet	4	6	12	56	12	6	4
2. szövetkezet	0	18	0	64	0	18	0

Ezen adatok alapján az átlagjövedelem és a szórás

	Átlag	Szórás
1. szövetkezet	$\bar{x}_1 = 1000$	$s_1 = \frac{400}{\sqrt{11}}$
2. szövetkezet	$\bar{x}_2 = 1000$	$s_2 = \frac{400}{\sqrt{11}}$

Tehát, ha csak az átlagot és a szórást vizsgáljuk, akkor azt mondhatjuk, hogy a két szövetkezet a tagok jövedelme szempontjából egyforma. Azonban a jövedelemeloszlás táblázatát nézve, valószínűleg mindenki azt fogja mondani, hogy az első szövetkezet sokkal egészségesebb, mint a második. A másodikban ugyanis a tagok egy nagy csoportja (18 fő) csak kis mértékben veszi ki részét a munkából (vagy kiveszi ugyan, de rosszul jövedelmező munkát kap), míg egy másik nagy csoport tagjainak keresete nagy.



15.8 ábra

E két példa azt mutatja, hogy bár az átlag és a szórás sokat elmond egy statisztikai sokaság struktúrájáról, de nem feltétlenül ad kielégítő képet. (Persze, ha eleve tudjuk, hogy az eloszlás normális, akkor a várható érték és a szórás teljesen meghatározza az eloszlást.) Próbáljuk most a két szövetkezet jövedelemeloszlását grafikonon ábrázolni. A következő két grafikont kapjuk (15.8 ábra).

E két grafikont hisztogramnak nevezzük. A hisztogram célja, hogy szemléletes képet nyerjünk arról, hogy az adatok hány százaléka esik bizonyos adott határok közé. Például az 1. szövetkezetre vonatkozó adatoknak a $4\% + 6\% = 10\%$ -a esik 700 és 800 közé, más szóval a hisztogram alatti terület 700 és 800 közé eső része jellemzi, hogy az adatok hány százaléka esik a két határ közé.

Hisztogramot adatainkból a következőképpen készíthetünk. Jelentsé $\xi_1, \xi_2, \dots, \xi_n$ adatfelvételünk eredményét, tegyük fel, hogy ezen adatok mindegyike valamely $[a; b]$ intervallumba esik, azaz a -t és b -t úgy választjuk meg, hogy

$$a \leq \min(\xi_1, \xi_2, \dots, \xi_n),$$

$$\text{és} \quad b \geq \max(\xi_1, \xi_2, \dots, \xi_n)$$

legyen. Majd osszuk fel az $[a; b]$ intervallumot részekre. Legyen például egy felosztás

$$a = x_0 < x_1 < x_2 < \dots < x_{k-1} < x_k = b.$$

Arról, hogy hogyan célszerű a -t és b -t pontosan megválasztani, valamint, hogy hogyan célszerű a felosztást elkészíteni, a későbbiekben lesz szó. Egyelőre csak annyit jegyünk meg, hogy számítástechnikai okokból a -t általában $\min(\xi_1, \xi_2, \dots, \xi_n)$ -nél nem sokkal kisebb kerek számnak, b -t $\max(\xi_1, \xi_2, \dots, \xi_n)$ -nél nem sokkal nagyobb kerek számnak választjuk. k -t (azaz az intervallumok számát) körülbelül $\sqrt{n}$ -nek választjuk, és az $x_1, x_2, \dots, x_{k-1}$ osztópontokat úgy választjuk meg, hogy az $[x_i; x_{i+1}]$ ($i = 0, 1, 2, \dots, k-1$) intervallumok mindegyikébe körülbelül $\sqrt{n}$ adat essék.

Ha most az $[a; b]$ intervallumon egy $h(t)$ hisztogramot úgy akarunk definiálni, hogy $\int_{x_i}^{x_j} h(t) dt$ ($0 \leq i < j \leq k$) megegyezzen az $[x_i; x_j]$ intervallumba eső adatok relatív gyakoriságával, akkor a következő módon járhatunk el. Számoljuk meg az $[x_{i-1}; x_i]$ ($i = 1, 2, \dots, k-1$) intervallumba eső adatok számát, legyen ez a szám z_i és legyen z_k az $[x_{k-1}; x_k]$ intervallumba eső adatok száma. Most $h(t)$ -t a következőképpen definiálhatjuk:

$$h(t) = \frac{z_i}{(x_i - x_{i-1})n}, \quad \text{ha} \quad x_{i-1} \leq t < x_i.$$

Látható, hogy

$$\int_{x_i}^{x_j} h(t) dt = \frac{\sum_{l=i+1}^j z_l}{n},$$

$$\text{és} \quad \int_a^b h(t) dt = 1.$$

Most vizsgáljunk egy számpéldát. Feladatunk a

11,34, 41,4, 45,4, 40,7, 20,2, 23, 31, 37,5, 41, 36,9, 26, 27,5, 28,2, 35,5, 37, 35,2, 30,2, 31,5, 22,2, 33,9, 35,5, 33,5, 33, 34,2, 12,5, 30,8, 33,7, 20, 34,9, 35,9, 23,5, 27,2, 30,5, 22,2, 36,5, 38,5, 28,15, 18,29, 40, 41,8, 24, 22,5, 32,7, 44, 46,5, 34, 28,5, 34,7, 57, 54, 36, 32, 32,5, 35,5, 35,6, 10,2, 28,8, 50,2, 39, 34,4, 33,7, 59,8, 26,5, 46,2, 42, 36,2, 36,4, 29,5, 41,6, 47,6, 39,2, 39,4, 40,5, 33,2, 38, 49,5, 42,5, 58,3, 36,7, 39,5, 35,7, 46, 49, 45, 35, 43,5, 58, 21, 25, 30,3, 43,2, 52,2, 10,5, 24,5, 19, 49

adatok hisztogramját elkészíteni. Legyen

$$a = 10$$

$$b = 60$$

$$k = 10$$

és az osztópontok legyenek:

$$x_1 = 22; x_2 = 27; x_3 = 30,5; x_4 = 33,5; x_5 = 35; x_6 = 36,5; x_7 = 39,5; x_8 = 43; x_9 = 48;$$

$$\text{ekkor} \quad z_1 = z_2 = \dots = z_{10} = 10,$$

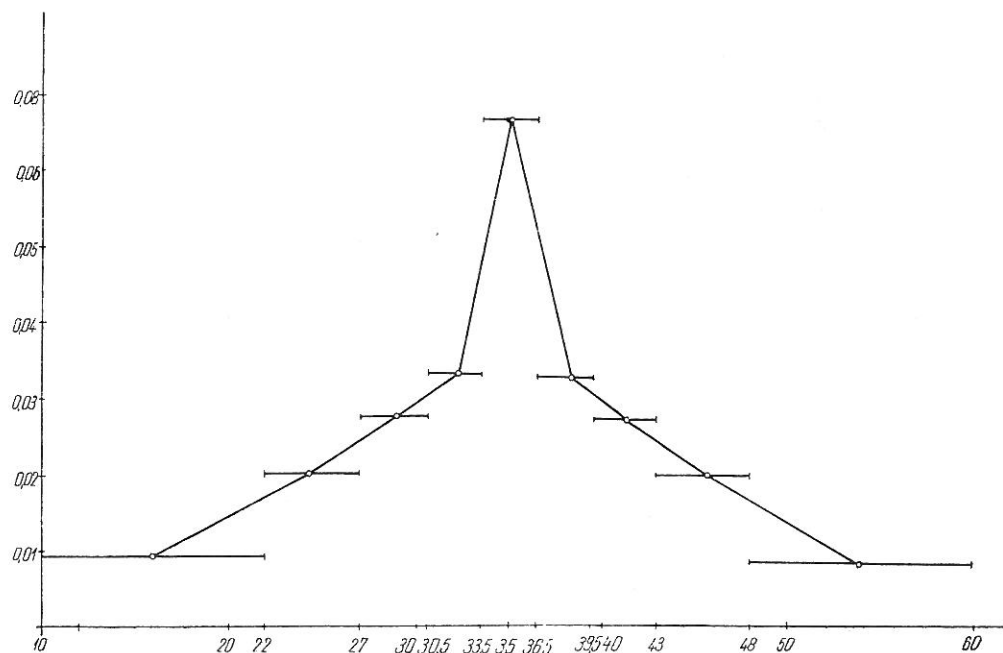
$$\text{így} \quad h(t) = \begin{cases} 0,00833, & \text{ha } 10 \leq t < 22 \\ 0,02, & \text{ha } 22 \leq t < 27 \\ 0,0285, & \text{ha } 27 \leq t < 30,5 \\ 0,0333, & \text{ha } 30,5 \leq t < 33,5 \\ 0,0666, & \text{ha } 33,5 \leq t < 35 \\ 0,0666, & \text{ha } 35 \leq t < 36,5 \\ 0,0333, & \text{ha } 36,5 \leq t < 39,5 \\ 0,0285, & \text{ha } 39,5 \leq t < 43 \\ 0,02, & \text{ha } 43 \leq t < 48 \\ 0,00833, & \text{ha } 48 \leq t < 60 \end{cases}$$

azaz hisztogramunk a 15.9 ábra szerint alakul.

Megjegyzés. A hisztogram elkészítésében mindig van bizonyos fokú önkény, azaz ha két különböző statisztikust kérünk meg, hogy készítsen ugyanazokból az adatokból hisztogramot, akkor általában két különböző grafikont kapunk, de ha az adatok száma nagy, a két görbe között csak minimális eltérés lesz.

A definíció szerint hisztogram egy lépcsős függvény, időnként azonban jól közelíthető valamely folytonos függvénnyel. Példánkban is szép görbét kapunk, ha a hisztogram egyenes szakaszainak középpontjait összekötjük egy folytonos vonallal (15.9 ábra).

A hisztogram fogalmával kapcsolatban abból indultunk ki, hogy az átlag és a tapasztalati szórás nem feltétlenül ad kielégítő képet a vizsgált statisztikai adatokról. Úgy gondoltuk,



15.9 ábra

hogy a hisztogram segítségével egy részletesebb áttekintést nyerhetünk. Valóban, ha a hisztogramra ránézünk, képet szerzünk arról, hogy mely intervallumokba esik sok adat (itt a hisztogram magas) és melyekbe esik kevés adat (itt a hisztogram alacsony). Pontosabban — mint láttuk — a hisztogram integrálja valamely intervallumon megegyezik az illető intervallumba eső adatok relatív gyakoriságával.

Annak jellemzésére, hogy a hisztogram valóban szinte a lehető legteljesebb jellemzését adja a statisztikai adatoknak, most példaként megmutatjuk, hogy a hisztogram segítségével miképpen lehet a várható értéket és a tapasztalati szórást meghatározni, vagy legalább megbecsülni.

Legyen a $h(t)$ hisztogram az $[a; b]$ intervallumon a következő módon definiálva:

$$h(t) = y_j, \quad \text{ha} \quad x_{j-1} \leq t < x_j, \quad (j = 1, 2, \dots, k),$$

ahol

$$a = x_0 < x_1 < x_2 < \dots < x_k = b$$

az $[a; b]$ intervallumnak egy felosztása. Ha a $h(t)$ hisztogramot n adatból szerkesztettük, akkor az $[x_{i-1}; x_i]$ ($i = 1, 2, \dots, k$) intervallumba eső adatok száma $y_i(x_i - x_{i-1})n$. Tételezzük fel, hogy az $[x_{i-1}; x_i]$ intervallumba eső összes adatok x_{i-1} -gyel egyenlők. E feltetelezés teljesülése esetén az adatok átlaga:

$$\sum_{i=1}^k x_{i-1} y_i (x_i - x_{i-1})$$

lenne. Ez az érték nyilván nem nagyobb, mint az adatok igazi átlaga. Az átlagra felső becslést teljesen hasonlóan nyerhetünk. Így azt mondhatjuk, hogy ha a $\xi_1, \xi_2, \dots, \xi_n$ adatokból a $h(t)$ hisztogram nyerhető, akkor ezen adatok $\bar{\xi}$ átlagára kizárólag a hisztogram felhasználásával a

$$\bar{\xi}_a = \sum_{i=1}^k x_{i-1} y_i (x_i - x_{i-1}) \leq \bar{\xi} \leq \sum_{i=1}^k x_i y_i (x_i - x_{i-1}) = \bar{\xi}_f$$

becslés nyerhető.

Utoljára vizsgált példánkban a szereplő 100 adat átlaga 34,799. Az adatokból nyert hisztogramot (és nem magukat az adatokat) felhasználva az átlagra a

$$\bar{\xi}_a = 32,5$$

alsó, és a

$$\bar{\xi}_f = 37,5$$

felső becslés nyerhető.

Megjegyzés. A hisztogramból nem olvasható le, hogy azt hány adatból nyertük, de az átlag alsó, illetve felső becslésének kiszámításához erre nincs is szükség.

Az átlag becslésére használhatjuk például az alsó és felső becslések számtani közepét, a

$$\bar{\xi}_k = \frac{\bar{\xi}_a + \bar{\xi}_f}{2} = \sum_{i=1}^k \frac{x_{i-1} + x_i}{2} y_i (x_i - x_{i-1})$$

mennyiséget.

A szórás becslését hasonlóan végezhetjük el. Tegyük fel, hogy a $h(t)$ hisztogramot a $\xi_1, \xi_2, \dots, \xi_n$ adatokból kaptuk, akkor

$$\sum_{i=1}^k x_{i-1}^2 y_i (x_i - x_{i-1}) \leq \frac{1}{n} \sum_{i=1}^n \xi_i^2 \leq \sum_{i=1}^k x_i^2 y_i (x_i - x_{i-1})$$

$$\text{és} \quad \left(\sum_{i=1}^k x_{i-1} y_i (x_i - x_{i-1}) \right)^2 \leq \left(\frac{1}{n} \sum_{i=1}^n \xi_i^2 \right)^2 \leq \left(\sum_{i=1}^k x_i y_i (x_i - x_{i-1}) \right)^2,$$

$$\text{így} \quad s_a^2 = \sum_{i=1}^k x_{i-1}^2 y_i (x_i - x_{i-1}) - \left(\sum_{i=1}^k x_{i-1} y_i (x_i - x_{i-1}) \right)^2 \leq s^2 = \\ = \frac{1}{n} \sum_{i=1}^n \xi_i^2 - \left(\frac{1}{n} \sum_{i=1}^n \xi_i \right)^2 \leq \sum_{i=1}^k x_i^2 y_i (x_i - x_{i-1}) - \left(\sum_{i=1}^k x_i y_i (x_i - x_{i-1}) \right)^2.$$

Az s_a^2 és s_f^2 becslések kizárólag a hisztogram felhasználásával, az eredeti statisztikai adatok ismerete nélkül is megkaphatók. Azonban az

$$\frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$$

mennyiség alsó és felső becsléséhez már a hisztogramon kívül legalább az adatok számára (azaz n -re) is szükség van.

Az átlag becsléséhez hasonlóan itt is természetesnek látszik a

$$\sum_{i=1}^k \left(\frac{x_{i-1} + x_i}{2} \right)^2 y_i (x_i - x_{i-1}) - \left(\sum_{i=1}^k \frac{x_{i-1} + x_i}{2} y_i (x_i - x_{i-1}) \right)^2$$

menntiséget, mint a szórás becslését tekinteni. (Ez a mennyiség nyilván az alsó és felső becslés, s_a^2 és s_b^2 között van.)

Belátható azonban, hogy ha az adatokat egy folytonos eloszlásból nyertük, akkor azáltal, hogy az adatokat csoportokba vontuk össze, a szórást megnöveltük. E hiba kiküszöbölésére egyenletes felosztás esetén használjuk az úgynevezett *Sheppard-féle* korrekciót. Eszerint a szórást az

$$s^2 - \frac{1}{12} h^2$$

menntiséggel becsülhetjük, ahol h a felosztásban szereplő intervallumok hossza. Részletesebb magyarázat nélkül megemlítjük, hogy a $\frac{h^2}{12}$ mennyiség azzal magyarázható, hogy az nem egyéb, mint egy h hosszúságú intervallumon értelmezett egyenletes eloszlású valószínűségi változó szórása.

15.4.1.2 A hisztogram mint a sűrűségfüggvény becslése

Legyenek $\xi_1, \xi_2, \dots, \xi_n$ független, közös $f(t)$ sűrűségfüggvényű valószínűségi változók. Természetesnek látszik az a gondolat, hogy az $f(t)$ sűrűségfüggvényt a $\xi_1, \xi_2, \dots, \xi_n$ adatok hisztogramjával becsüljük (ha az $f(t)$ sűrűségfüggvény ismeretlen, és célunk azt a statisztikai adatokból becsülni).

Pontosabban problémánkat a következőképpen fogalmazhatjuk meg. Legyenek $\xi_1, \xi_2, \dots$ független, egyforma eloszlású valószínűségi változók, és jelents $f(t)$ közös sűrűségfüggvényüket. Készítsük el a $\xi_1, \xi_2, \dots, \xi_n$ statisztikai adatok $h_n(t)$ hisztogramját. Kérdezzük, hogy állítható-e, hogy $h_n(t)$ (valamilyen értelemben) konvergál $f(t)$ -hez, ha $n \rightarrow \infty$.

Mindenekelőtt gondoljunk vissza a hisztogram definíciójára. Láttuk, hogy a hisztogram elkészítésére (pontosabban az $[a; b]$ intervallumnak és a felosztásnak a megválasztására) több lehetőség kínálkozik. Most azért, hogy pontosan megfogalmazhassuk tételeinket, definiáljuk egyértelműen n adat hisztogramjának fogalmát, de hangsúlyozni kívánjuk, hogy nem feltétlenül szükséges hisztogram készítésénél az itt közölt definíciót használni. A hisztogram elkészítésénél, első lépésként $\xi_1, \xi_2, \dots, \xi_n$ adatainkat rendezzük nagyság szerint sorba. Jelents $\xi_1^* < \xi_2^* < \dots < \xi_n^*$ a már nagyság szerint rendezett mintát, azaz ξ_1^* a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemeknek legkisebbike, ξ_2^* a nagyság szerinti második, és így tovább. A $\xi_1^*, \xi_2^*, \dots, \xi_n^*$ valószínűségi változókat *rendezett mintának*, (a $\xi_1, \xi_2, \dots, \xi_n$ adatokból nyert rendezett mintának) nevezzük.

Ezután a $h(t)$ hisztogramot a következőképpen definiálhatjuk:

$$h_n(t) = \frac{1}{\xi_{(k+1)}^* - \xi_{ka(n)}^*} \frac{a(n)}{n}, \quad \text{ha} \quad \xi_{ka(n)}^* \leq t < \xi_{(k+1)}^* a(n)$$

$$\left(k = 1, 2, \dots, \frac{n}{a(n)} \right),$$

ahol $a(n) = \lfloor \sqrt{n} \rfloor$. Természetesen az $a(n)$ függvényt másképpen ($\lfloor \sqrt{n} \rfloor$ -től különbözően) is definiálhatnánk, most azonban az így definiált hisztogramsorozat tulajdonságait kívánjuk vizsgálni.

Megjegyzés. $h_n(t)$ értéke (fix n és t mellett) a véletlentől függ, azaz valószínűségi változó, így csak olyan eredményeket remélhetünk, hogy $h_n(t)$ nagy valószínűséggel (esetleg 1 valószínűséggel) tart $f(t)$ -hez, de nem remélhetjük, hogy $h_n(t)$ az analízisben megszokott konvergenciafogalmaknak megfelelően konvergál $f(t)$ -hez.

Könnyű belátni, hogy tetszőleges $-\infty < \alpha < \beta < +\infty$ esetén, $\int_{\alpha}^{\beta} h_n(t) dt$ (nagy n -re) közel van $\int_{\alpha}^{\beta} f(t) dt$ -hez, pontosabban

$$P \left(\left| \int_{\alpha}^{\beta} h_n(t) dt - \int_{\alpha}^{\beta} f(t) dt \right| \leq \varepsilon \right) \rightarrow 1 \quad (n \rightarrow \infty)$$

tetszőleges $\varepsilon > 0$ esetén. Ugyanis $\int_{\alpha}^{\beta} f(t) dt$ jelenti annak a valószínűségét, hogy a kísérlet eredménye α és β közé esik, míg $\int_{\alpha}^{\beta} h_n(t) dt$ közelítőleg megegyezik az $[\alpha; \beta]$ intervallumba eső kísérletek számának relatív gyakoriságával. Pontosabban

$$(8) \quad \left| \int_{\alpha}^{\beta} h_n(t) dt - \frac{k}{n} \right| \leq 2 \frac{a(n)}{n} = 2 \frac{\lfloor \sqrt{n} \rfloor}{n},$$

ahol k jelenti az n kísérlet közül az α és β közé esők számát. (A (8) egyenlőtlenségünk bizonyítását az olvasóra bizzuk.) A nagy számok törvényei alapján azonban tudjuk, hogy a relatív gyakoriság nagy valószínűséggel közel van a valószínűséghez, ami szemlélteti (8) formulánk helyességét.

Természetesen (8)-ból még nem következik, hogy fix t mellett $h_n(t)$ konvergál $f(t)$ -hez, akárcsak nagy valószínűséggel. Belátható azonban, hogy ha $f(t)$ elég sima (pl. egy zárt korlátos intervallumon kívül eltűnik, és azon belül folytonos), akkor $h_n(t)$ 1 valószínűséggel egyenletesen konvergál $f(t)$ -hez.

Tétel. Legyenek $\xi_1, \xi_2, \dots$ független, $f(t)$ sűrűségfüggvényű valószínűségi változók. Tegyük fel, hogy $f(t)$ egy $[a; b]$ intervallumon kívül eltűnik, és az $[a; b]$ intervallumon pedig folytonos. Jelents $h_n(t)$ a $\xi_1, \xi_2, \dots, \xi_n$ adatok hisztogramját. Ekkor

$$P \left(\sup_t |h_n(t) - f(t)| \rightarrow 0 \right) = 1.$$

15.4.2 Az eloszlásfüggvény meghatározása statisztikai adatokból

15.4.2.1 Az eloszlásfüggvény becslése

Legyen $F(x)$ egy ismeretlen eloszlásfüggvény, és legyenek $\xi_1, \xi_2, \dots, \xi_n$ független $F(x)$ eloszlásfüggvényű valószínűségi változók. Mint tudjuk $F(x_0)$ annak a valószínűsége, hogy egy $F(x)$ eloszlásfüggvényű valószínűségi változó értéke (vagy a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek

valamelyike) x_0 -nál kisebb legyen. Így természetesnek látszik $F(x_0)$ -t a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek közül az x_0 -nál kisebbek számának relatív gyakoriságával becsülni. Azaz jelentse k a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek közül az x_0 -nál kisebbek számát, és definiáljuk az $F_n(x)$ tapasztalati eloszlásfüggvény értékét az x_0 pontban az

$$F_n(x_0) = \frac{k}{n}$$

képlettel.

A nagy számok törvénye értelmében az $F_n(x_0)$ relatív gyakoriság nagy n esetén közel van az $F(x_0)$ valószínűséghez, pontosabban

$$(9) \quad P(|F_n(x_0) - F(x_0)| \leq \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty),$$

ahol ε tetszőleges pozitív szám. Mivel (9) formulánk minden x_0 -ra teljesül, az $F_n(x)$ tapasztalati eloszlásfüggvényt joggal tekinthetjük az $F(x)$ eloszlásfüggvény becslésének. Érdekes megemlíteni, hogy az $F_n(x)$ tapasztalati eloszlásfüggvényt definiálhatjuk a rendezett minták fogalmának felhasználásával is. Nevezetesen legyen $\xi_1^* < \xi_2^* < \dots < \xi_n^*$ a $\xi_1, \xi_2, \dots, \xi_n$ mintából nyert rendezett minta, akkor

$$F_n(x) = \frac{k}{n}, \quad \text{ha} \quad \xi_k^* \leq x < \xi_{k+1}^*.$$

Könnyen belátható, hogy a tapasztalati eloszlásfüggvény ezen és az előzőleg adott definíciója ekvivalens.

A statisztikusok sokat foglalkoztak azzal, hogy hogyan jellemezhető a tapasztalati eloszlásfüggvénynek az elméleti eloszlásfüggvényhez való konvergenciája. Az ezzel kapcsolatos alapvető tétel — amelyet a *statisztika alaptételének* vagy *Glivenko–Cantelli-tételnek* is is szoktak nevezni — azt állítja, hogy a tapasztalati eloszlásfüggvény 1 valószínűséggel egyenletesen konvergál az elméleti eloszlásfüggvényhez.

Tétel. $P\left(\sup_{-\infty < x < +\infty} |F_n(x) - F(x)| \rightarrow 0\right) = 1.$

Sokan vizsgálták azt a kérdést is, hogy milyen sebességgel konvergál az $F_n(x)$ tapasztalati eloszlásfüggvény az $F(x)$ elméleti eloszlásfüggvényhez. Bár az ilyen típusú eredmények is nagy gyakorlati jelentőségűek, itt nem tudunk rájuk kitérni.

15.4.2.2 Az eloszlásfüggvényre vonatkozó hipotézisek vizsgálata

Legtöbb gyakorlati esetben, ha nem is tudjuk biztosan, hogy vizsgált valószínűségi változónk milyen eloszlású, de van valami elképzelésünk legalább az eloszlás típusáról. Mint már többször említettük, nagyon sok gyakorlati esetben indokolt az a hipotézis, hogy kísérletünk eredménye normális eloszlású valószínűségi változó. Most azt a kérdést vizsgáljuk, hogy ezt vagy más, az eloszlásra vonatkozó hipotézist, hogy ellenőrizhetjük a statisztikai adatok alapján.

Ezzel a kérdéssel már foglalkoztunk a χ^2 -próbával kapcsolatban. Az illeszkedésvizsgálat célja ott is annak eldöntése volt, hogy statisztikai adataink alapján elhihető-e az eloszlás-

függvényre vonatkozó hipotézisünk. Bár a χ^2 -próba tanulmányozásánál megemlítettük, hogy az illeszkedésvizsgálat nem diszkrét eloszlásokra is alkalmazható, mégis a folytonos eloszlásokra vonatkozó hipotézisek vizsgálatára jobb módszerek is rendelkezésünkre állnak. Most térünk rá a folytonos eloszlásokra vonatkozó hipotézisek vizsgálatára. Tehát feladatunk azon hipotézisek helyes vagy hamis voltáról dönteni, hogy kísérletünk eredményének eloszlásfüggvénye valamely adott folytonos eloszlásfüggvénnyel megegyezik-e. Kérdésünk pontosabban így fogalmazható.

Legyenek $\xi_1, \xi_2, \dots, \xi_n$ független egyforma eloszlású valószínűségi változók. Jelentse $G(x)$ a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók közös eloszlásfüggvényét. Feladatunk a

$$H_0: G(x) = F(x)$$

hipotézis helyes vagy hamis voltát eldönteni, ahol $F(x)$ adott folytonos eloszlásfüggvény. Az ötletet egy, a kérdést eldöntő statisztikai próba kidolgozásához a statisztika alaptétele adja meg. Eszerint, ha hipotézisünk helyes, akkor a $\xi_1, \xi_2, \dots, \xi_n$ minta $F_n(x)$ tapasztalati eloszlásfüggvénye nagy valószínűséggel egyenletesen közel van a hipotetikus eloszlásfüggvényhez. Vagyis a következő módon járhatunk el. Először elkészítjük a $\xi_1, \xi_2, \dots, \xi_n$ mintaelemek $F_n(x)$ tapasztalati eloszlásfüggvényét, majd meghatározzuk e tapasztalati eloszlásfüggvény és a hipotetikus eloszlásfüggvény közötti legnagyobb eltérést, azaz a

$$t_n = \sup_{-\infty < x < +\infty} |F_n(x) - F(x)|$$

értéket. A H_0 hipotézis elfogadható, ha ez az érték kicsi, és elvetendő, ha nagy.

Természetesen tudnunk kell, hogy mennyire szabad nagyra lennie a t_n mennyiségnek ahhoz, hogy még a hipotézis elfogadható legyen. A hipotézisvizsgálat kifejezéseivel élve, meg kell határoznunk az elfogadási és a kritikus tartományt. Ehhez azonban azt kell tudnunk, hogy a t_n statisztika milyen valószínűséggel vesz fel valamely adott értéknél nagyobb, más szóval a t_n statisztika (valószínűségi változó) eloszlásfüggvényét kellene tudnunk.

Bebizonyítható, hogy t_n úgy tart 0-hoz, mint $\frac{1}{\sqrt{n}}$, és hogy $\sqrt{nt_n}$ eloszlásfüggvénye nagy n -re közelítőleg

$$K(y) = \begin{cases} \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 y^2}, & \text{ha } y > 0, \\ 0 & \text{különben.} \end{cases}$$

Tétel.

$$(10) \quad \lim_{n \rightarrow \infty} P(\sqrt{nt_n} < y) = K(y).$$

A (10) formula KOLMOGOROV-tól származik, ezért az e formulán alapuló próbát *Kolmogorov-próbának* szoktuk nevezni.

A $K(y)$ függvény értékeit táblázatban adják meg. Táblázatunkban például azt találjuk, hogy

$$K(1,36) = 0,95,$$

így H_0 hipotézisünket 0,05 szinten elfogadhatjuk, ha $\sqrt{nt_n} < 1,36$.

Nézzünk egy konkrét példát! Tegyük fel, hogy szegecsket gyártunk és 25 véletlenszerűen kiválasztott szegecs átmérőjét lemérve a

10,05	10,28	10,60	10,90	11,30
10,10	10,30	10,62	10,92	11,32
9,76	9,50	9,00	9,32	8,65
10,12	10,70	11,40	8,80	9,50
9,30	9,22	8,68	9,72	9,56

adatokat kapjuk. Ellenőrizni akarjuk, hogy

$$H_0 : F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt; \quad m = 10, \quad \sigma = 1$$

hipotézisünk teljesül-e, azaz hipotézisünk az, hogy a szegecs átmérője $m = 10$ várható értékű és $\sigma = 1$ szórással normális eloszlású valószínűségi változó.*

Ábrázoljuk a 15.10 ábra szerint egy grafikonon az adatok $F_n(x)$ ($n = 25$) tapasztalati eloszlásfüggvényét és a hipotetikus

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-10)^2}{2}} dt$$

elméleti eloszlásfüggvényt. A grafikon alapján azt találjuk, hogy $F_n(x)$ és $F(x)$ között a legnagyobb eltérés az $x = 8,65$ pontban van és ott

$$F_n(x) - F(x) = 0,08.$$

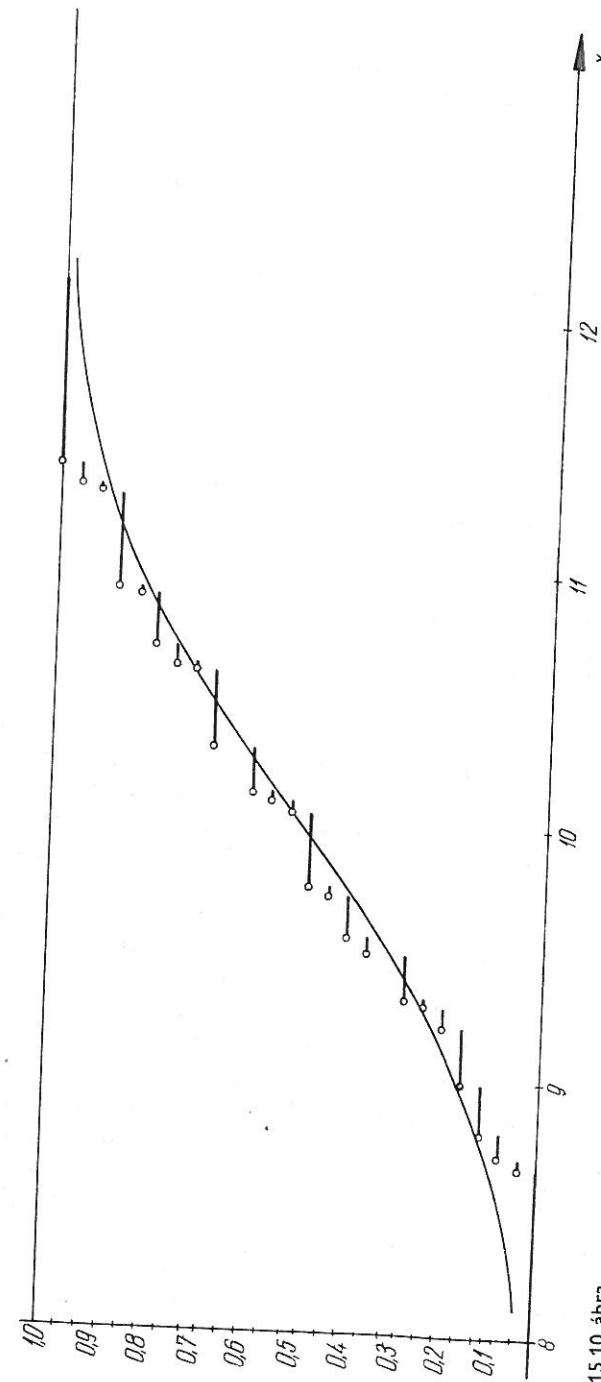
Így hipotézisünket 0,05 szinten elfogadhatjuk.

Hangsúlyozni kívánjuk, hogy más — a normálistól különböző, de folytonos hipotetikus eloszlásfüggvények esetén a Kolmogorov-próba teljesen hasonló módon alkalmazható.

15.4.2.3 Grafikus eljárások

Abban az esetben, ha csak egy durva, de gyors választ akarunk kapni arra a kérdésre, hogy statisztikai adataink származhatnak-e valamely adott eloszlásból, akkor a Kolmogorov-próba helyett célszerű grafikus eljárásokkal próbálkozni. Itt csak a normális eloszlás ellenőrzésére (normalitásvizsgálatokra) szolgáló *Gauss-papírok* használatával fogunk foglalkozni.

* Gyakorlatban legtöbbször már tudjuk, hogy az átmérő normális eloszlású, és csak a várható értékre, továbbá a szórással vonatkozó hipotéziseket ellenőrizzük, például t -, illetve χ^2 -próbával. Más esetekben azonban éppen azt akarjuk tudni, hogy az eloszlás valóban normális-e. Például ha gyártás közben változtak a körülmények, akkor az eloszlás általában különbözni fog a normálistól, vizsgálatunk célja lehet éppen ennek a körülménynek a kiderítése. Azt mondhatjuk, hogy legtöbb gyakorlati esetben, ha egy folytonos eloszlás nem normális, akkor szükséges, hogy felderítsük az okát, hogy miért nem normális, és ez az ok értékes információkat nyújthat az eloszlásról.



15.10 ábra

Legyen ξ egy m várható értékű és σ szórású normális eloszlású valószínűségi változó, azaz ξ eloszlásfüggvénye

$$F(x; m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-m)^2}{2\sigma^2}} dt.$$

Könnyen belátható, hogy az $F(\xi; m, \sigma)$ valószínűségi változó a $[0; 1]$ intervallumban egyenletes eloszlású, ugyanis

$$\begin{aligned} P(F(\xi; m, \sigma) < y) &= P(\xi < F^{-1}(y; m, \sigma)) = F(F^{-1}(y; m, \sigma)) = \\ &= \begin{cases} 0, & \text{ha } y < 0, \\ y, & \text{ha } 0 \leq y \leq 1, \\ 1, & \text{ha } y > 1. \end{cases} \end{aligned}$$

Ennek megfelelően, ha $\xi_1, \xi_2, \dots, \xi_n$ független m várható értékű és σ szórású normális eloszlású valószínűségi változók, akkor az $F(\xi_1; m, \sigma), F(\xi_2; m, \sigma) \dots F(\xi_n; m, \sigma)$ valószínűségi változók tapasztalati eloszlásfüggvénye közelítőleg a $[0; 1]$ intervallumban egyenletes eloszlás eloszlásfüggvényével, vagyis az

$$E(x) = \begin{cases} 0, & \text{ha } x < 0, \\ x, & \text{ha } 0 \leq x \leq 1, \\ 1, & \text{ha } x > 1 \end{cases}$$

függvénnyel egyenlő. Így, ha nem az eredeti adatoknak, hanem az $F(x; m, \sigma)$ függvény segítségével transzformált adatoknak, vagyis az $F(\xi_1; m, \sigma), F(\xi_2; m, \sigma) \dots$ értékeknek készítjük el a tapasztalati eloszlásfüggvényét, akkor az ábrára *ránézve* is megmondhatjuk, hogy a közel van-e az $E(x)$ függvényhez vagy sem. Ezen eljárásnak hátránya, hogy a grafikon elkészítése előtt ki kell számítani az adatok transzformáltjait, vagyis az $F(\xi_1; m, \sigma), F(\xi_2; m, \sigma), \dots, F(\xi_n; m, \sigma)$ mennyiségeket.

E nehézségen úgy lehet segíteni, hogy ha eleve olyan koordináta-rendszert használunk, amelynek ordinátatengelye nem egyenletesen, hanem az $F(x; m, \sigma)$ függvénynek megfelelően van beosztva. Ugyanis, ha ezen grafikonon ábrázoljuk az eredeti $\xi_1, \xi_2, \dots, \xi_n$ adatok tapasztalati eloszlásfüggvényét, akkor az (a hipotézis teljesülése esetén) egyenes lesz.

Ennek az elvnek megfelelően készítik a Gauss-papírt. Pontosabban a Gauss-papíron a koordináta-rendszer y -tengelyét az $F(x; 0, 1) = \Phi(x)$ függvénynek megfelelően osztják be. Az így készült Gauss-papíron egy $m=0, \sigma=1$ paraméter normális eloszlás tapasztalati eloszlásfüggvénye az $E(x)$ függvény lesz, míg egy tetszőleges m, σ paraméterű normális eloszlásból származó minta tapasztalati eloszlásfüggvénye az $y = \sigma x + m$ egyenes lesz. Így statisztikai adatainkról akkor mondhatjuk hogy azok egy normális eloszlásból származnak, ha tapasztalati eloszlásfüggvényük a Gauss-papíron közelítőleg egyenes.

Hangsúlyozni kívánjuk, hogy a Gauss-papír (és a többi hasonló papírok) csak első tájékozódásra szolgálnak, de nem pótolják a pontos statisztikai elemzést.

15.5 VALÓSZÍNŰSÉGI VÁLTOZÓK KÖZÖTTI KAPCSOLATOK JELLEMZÉSE

A χ^2 -próbával kapcsolatban foglalkoztunk függetlenségvizsgálattal, azaz kidoigoztunk egy próbát, amely megmondja, hogy az a hipotézis, hogy két valószínűségi változó független, összeegyeztethető-e a két valószínűségi változóra vonatkozó kísérleti adatokkal. Jelen fejezetben azt vizsgáljuk, hogy két (vagy több) nem független valószínűségi változó közötti kapcsolat (a változók közötti függés típusa és erőssége) hogyan jellemezhető.

15.5.1 Korrelációs együttható

Mint tudjuk két valószínűségi változó közötti kapcsolat jellemzésére általában a korrelációs együtthatót használjuk. Emlékeztetül jegyezzük meg, hogy két valószínűségi változó ξ és η korrelációs együtthatója az

$$R(\xi, \eta) = \frac{M((\xi - M(\xi))(\eta - M(\eta)))}{D(\xi)D(\eta)} = \frac{M(\xi\eta) - M(\xi)M(\eta)}{D(\xi)D(\eta)}$$

mennyiség, amely a következő tulajdonságokkal rendelkezik:

- a) $|R(\xi, \eta)| \leq 1,$
- b) $R(\xi, \eta) = 0,$ ha ξ és η függetlenek.

A korrelációs együttható szemléletes jelentését a következő módon világíthatjuk meg. Tegyük fel, a két valószínűségi változó közötti kapcsolat olyan, hogy abból, hogy az egyik változó értéke nagy (saját várható értékéhez viszonyítva), következik, hogy a másik változó értéke is általában nagy, és fordítva, ha az egyik változó értéke kicsi, akkor általában a másik értéke is kicsi. Ez más szóval azt jelenti, hogy a $\xi - M(\xi)$ és az $\eta - M(\eta)$ valószínűségi változók előjele általában megegyezik, azaz szorzatuk (illetve annak várható értéke) pozitív. Így tehát azt mondhatjuk, hogy ha két valószínűségi változó, ξ és η közötti korreláció pozitív, $R(\xi, \eta) > 0$, akkor a közöttük levő kapcsolat olyan jellegű, hogy az az esemény, hogy az egyik változó értéke nagy, általában maga után vonja azt, hogy a másik változó is nagy.

Fordítva, negatív korreláció esetén az egyik változó nagy voltából a másik változó kicsiny voltára lehet következtetni. A korrelációs együttható e szemléletes tulajdonságát később a lineáris regresszió segítségével pontosabban is meg tudjuk világítani.

Ezután rátérünk annak vizsgálatára, hogy statisztikai adatokból miképpen lehet két valószínűségi változó korrelációs együtthatóját becsülni.

Tegyük fel, hogy két valószínűségi változó, ξ és η értékét egyidejűleg, n alkalommal megmérjük. Például egy erdőből kivágott n darab fának megmérjük a magasságát ξ , és súlyát η ; vagy egy állatállományban meghatározzuk n állatnak életkorát ξ , és súlyát η ; vagy egy tehénállományban vizsgáljuk n tehénnek a tejhozama ξ , és súlya η közötti összefüggést. Kísérletünk eredménye n számpár lesz. Legyenek ezek $(x_1; y_1), (x_2; y_2), \dots, (x_n; y_n)$. Ekkor $R(\xi, \eta)$ definíciójában szereplő $M((\xi - M(\xi))(\eta - M(\eta)))$, $D(\xi)$, $D(\eta)$ mennyiségek rendre a következő

értékekkel becsülhetők:

$$m_{\xi\eta} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \frac{1}{n(n-1)} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$

$$\sigma_{\xi} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{1}{n-1} \cdot \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2}$$

$$\sigma_{\eta} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n y_i^2 - \frac{1}{n-1} \cdot \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2}$$

Azt a tényt, hogy a σ_{ξ} és a σ_{η} mennyiségek valóban becslései a megfelelő szórásoknak az előzőekben láttuk. Az, hogy $m_{\xi\eta}$ torzítatlan konzisztens becslése $M((\xi - M(\xi)) \cdot (\eta - M(\eta)))$ -nak a nagy számok törvényeiből közvetlenül következik. E becsléseket felhasználva arra jutunk, hogy $R(\xi, \eta)$ -t a

$$(11) \quad \rho = \frac{m_{\xi\eta}}{\sigma_{\xi}\sigma_{\eta}} = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) \left(\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right)}}$$

mennyiséggel becsülhetjük. Jegyezzük meg, hogy az előzőekből azonnal következik, hogy ρ a korrelációs együttható konzisztens becslése, de általában nem torzítatlan becslés. (Már σ_{ξ} sem volt torzítatlan becslés.)

A ρ mennyiség kiszámításánál számítástechnikai okokból legcélszerűbb a (11) formulát használni. Ennek alkalmazásához az adatokból meg kell határoznunk a

$$\sum_{i=1}^n x_i, \quad \sum_{i=1}^n y_i, \quad \sum_{i=1}^n x_i y_i, \quad \sum_{i=1}^n x_i^2, \quad \sum_{i=1}^n y_i^2$$

mennyiségeket, amelyeknek birtokában a (11) formula felhasználásával, már igen könnyen kaphatjuk ρ értékét.

Megjegyzés. A szórások becslésénél, azaz a σ_{ξ} , σ_{η} mennyiségek meghatározásánál használhatjuk a Sheppard-féle korrekciót.

Nézzünk egy konkrét példát a korrelációs együttható becslésére.

Tegyük fel, hogy teheneknél az anyaállat és az elsőszülött leány tejhozama közötti összefüggést akarjuk vizsgálni, és a kapcsolat jellegét a korrelációs együttható értékével akarjuk kifejezni. (Természetesen ahhoz, hogy a helyzetről reális képet kapjunk szükséges, hogy a kísérleti állatokat egyforma korban és körülmények között vizsgáljuk. Így kísérletünk eredménye akkor lesz csak teljesen megbízható, ha a kísérletben szereplő leányállatok mindegyikénél ugyanaz az apaállat.)

Tételezzük fel, hogy kísérletünk folyamán 50 pár tehen tejhozamát vizsgáltuk meg. A kapott eredményeket a következő táblázatban foglalhatjuk össze:

Anyanya	18 l	19 l	20 l	21 l	22 l	23 l	24 l	25 l	26 l
Leány									
18 l									
19 l	1	2	1	1					
20 l		1	2		1	1			
21 l	1		1	2					
22 l				5	1		2		
23 l					10	3	2	1	
24 l					4	3			1
25 l									1
26 l						1		1	
27 l								1	

Számítsuk ki először a korrelációs együttható becslésének, ρ -nak meghatározásához szükséges mennyiségeket. Adódik, hogy

$$n = 50, \quad \sum_{i=1}^{50} x_i = 1100, \quad \sum_{i=1}^{50} y_i = 1119, \quad \sum_{i=1}^{50} x_i y_i = 24\,744,$$

$$\sum_{i=1}^{50} x_i^2 = 24\,366, \quad \sum_{i=1}^{50} y_i^2 = 25\,219,$$

ahol x_i az i -edik anya, y_i az i -edik leányállat által adott tej mennyisége. Ebből kapjuk, hogy

$$\rho = \frac{\sum_{i=1}^{50} x_i y_i - \frac{1}{50} \sum_{i=1}^{50} x_i \sum_{i=1}^{50} y_i}{\sqrt{\left(\sum_{i=1}^{50} x_i^2 - \frac{1}{50} \left(\sum_{i=1}^{50} x_i \right)^2 \right) \left(\sum_{i=1}^{50} y_i^2 - \frac{1}{50} \left(\sum_{i=1}^{50} y_i \right)^2 \right)}} =$$

$$= \frac{126}{\sqrt{166 \cdot 175,8}} = 0,72.$$

Megjegyzés. Hogy ha az x_i , y_i mennyiségek mindegyikéből például 20-at levonunk, azzal ρ értékét nem változtatjuk, de a számítási munkát megkönnyítjük.

A korrelációs együttható értékének pontos ismerete számos genetikai vizsgálatban játszik jelentős szerepet. A genetikában általában elméleti úton tudunk következtetni a korrelációs együttható értékére, és a statisztikai vizsgálat feladata az elméleti úton kapott hipotézis megerősítése vagy elvetése.

A korrelációs együttható értékére vonatkozó hipotézisek vizsgálata általában azon alapul, hogy nagy n esetén ρ normális eloszlású változó R várható értékkel és $\frac{1-R^2}{\sqrt{n}}$ szórással,

ahol R a valódi korrelációs együttható. Így a korrelációs együtthatóra vonatkozó hipotéziseket a normális eloszlás táblázatában ellenőrizhetjük, pl. azt mondhatjuk, hogy a

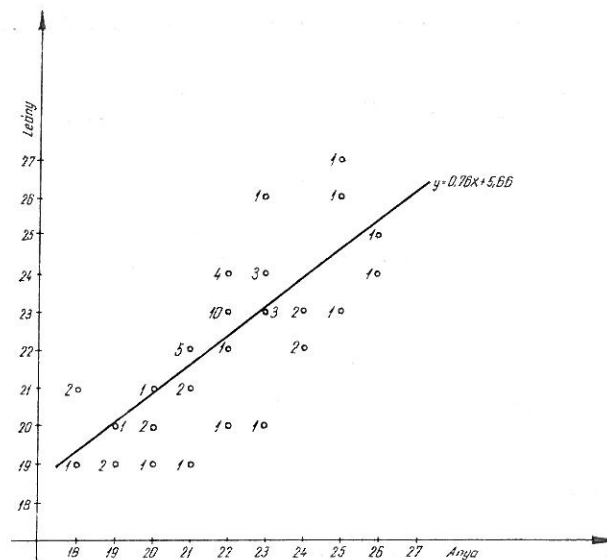
$$H_0: R = \frac{1}{2}$$

hipotézis elfogadható 0,95 szinten, ha q értéke $\frac{1}{2} + \frac{3}{4\sqrt{n}}$ és $\frac{1}{2} - \frac{3}{4\sqrt{n}}$ közé esik.

Megjegyzés. q eloszlásának a normális eloszlással való közelítése igen rossz közelítés. Csak nagy n -nek ($n \geq 500$) esetén alkalmazható. Kisebb n -ek esetén más közelítést kell alkalmaznunk, de erre itt nem térünk ki.

15.5.2 A regresszióanalízis

Az előző pontban kiszámítottuk az anya- és a leányállat tejhozama közötti korrelációs együttható értékét. Azt kaptuk, hogy ez az érték 0,72. Ebből (az esetleges genetikai következtetésektől eltekintve) azt a következtetést vonhatjuk le, hogy az anya- és a leányállat



15.11 ábra

tejhozama között elég szoros kapcsolat van, nagy tejhozamú anyaállatok leányának tejhozama is általában nagy, a tejhozam jól öröklődő tulajdonság. A kapcsolatot jobb megvilágítása céljából próbáljuk meg az 1209. oldalon szereplő táblázatunk adatait egy koordináta-rendszerben ábrázolni.

A 15.11 ábrán a pontok mellett levő számok azt jelentik, hogy az illető adat hányszor fordult elő a mérések során.

Az ábrára nézve úgy látszik, hogy a pontok általában az $y = x + 1$ egyenes körül helyezkednek el. Ha e szemléletes kép helyesnek bizonyul, azt mondhatjuk, hogy egy x liter tejhozamú tehén leányának tejhozama általában $x + 1$ (feltéve, hogy az apaállat ugyanaz). Most elsősorban azt vizsgáljuk, hogy a 15.11 ábrán levő pontokat valójában mely egyenessel lehet legjobban közelíteni. Azaz van-e olyan $y = ax + b$ egyenes, amely az $y = x + 1$ egyenesnél közelebb van az ábrán levő pontokhoz. Később térünk rá arra a kérdésre, hogy a legjobban közelítő egyenes mennyire és milyen értelemben tekinthető megbízhatónak.

Természetesen több lehetőség van arra, hogy a sík pontjainak egy halmaza (adott esetben az $(x_1, y_1), (x_2, y_2), \dots, (x_{50}, y_{50})$ pontok halmaza) és a sík egy egyenese közötti távolságot definiáljuk, illetve hogy megmondjuk, melyik egyenes van legközelebb valamely ponthalmazhoz.

Mi abból indulunk ki, hogy egy egyenest akkor tekintünk egy ponthalmazhoz közelinek, ha az a halmaz minden pontjához közel van. Pontosabban: legyen $y = ax + b$ egy egyenes, és legyen $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ a sík pontjainak egy halmaza. Az (x_i, y_i) pontnak az $y = ax + b$ egyenestől való távolságán az $(y_i - ax_i - b)^2$ mennyiséget értjük, és a ponthalmaznak az egyenestől való távolságán pedig a

$$(12) \quad \sum_{i=1}^n (y_i - ax_i - b)^2$$

összeget.

A vizsgált példánkban szereplő pontok és a szemléletesen jó közelítést nyújtó $y = x + 1$ egyenes közötti távolság például 106.

Abból a kérdésből indultunk ki, hogy az $y = x + 1$ egyenes valóban a legjobb közelítése-e az adatoknak, vagy van-e olyan $y = ax + b$ egyenes, amely közelebb van a példában szereplő $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ pontokhoz, azaz, amelyre

$$\sum_{i=1}^{50} (y_i - ax_i - b)^2 < 106.$$

Méginkább azt szeretnénk tudni, hogy hogyan lehet megtalálni azt az $y = ax + b$ egyenest (illetve az a, b paramétereket), amely legközelebb van az adott pontokhoz, illetve amelyre az

$$F(a, b) = \sum_{i=1}^{50} (y_i - ax_i - b)^2$$

összeg a lehető legkisebb.

(Még egyszer hangsúlyozzuk, az hogy a $\sum_{i=1}^n (y_i - ax_i - b)^2$ összeget kívánjuk minimalizálni elég önkényes választás, bár bizonyos fokig elméletileg is indokolható lenne. Valójában mondhatnánk, hogy a legjobban közelítő egyenes az, amelyre a

$$\sum_{i=1}^{50} |y_i - ax_i - b|$$

$$\text{vagy a} \quad \sum_{i=1}^{50} (y_i - ax_i - b)^4$$