Pázmány Law Working Papers
2020/6

Szalai Ákos
Statisztika és a jog

PÁZMÁNY PÉTER KATOLIKUS EGYETEM
PÁZMÁNY PÉTER CATHOLIC UNIVERSITY BUDAPEST
HTTP://WWW.PLWP.JAK.PPKE.HU/

STATISZTIKA ÉS A JOG

Szalai Ákos[+]

A példa: egy munkaadót beperelnek a nála foglalkoztatott feketék azért, mert az átlagos bérük alacsonyabb, mint a fehéreké.

B példa: egy egyetemen a mesterszakos felvételi eredményekben jelentős eltérés mutatkozik a férfiak és a nők között. A férfiakat 49-át veszik fel, míg a nőknek csak 30%-át. A kérdés, hogy nemi diszkrimináció történik-e.

C példa: Egy perben azt kell megállapítani, hogy egy ingatlan értéke mekkora volt – mennyiért lehetett volna a piacon eladni. A probléma az, hogy a piacon eladott ingatlanok kisebb-nagyobb mértékben mindig különböznek az adott perben érintett ingatlantól.

Milyen olyan kérdések vannak a jogban, amelyekre adatokra hivatkozva válaszolunk? Hogyan olvashatók egyszerűen ezek az adatok – különösen, ha nagyon sok különböző van belőlük? Kiolvasható-e adatokból az, hogy az egyik jelenség okozza a másikat, vagy hogy egy személy jellemzője okozza-e azt, hogy hogyan viselkedik, milyen helyzetekbe kerül? Ha sok adatok ismerünk, de nem ismerjük mindet (például bizonyos emberek adatait ismerjük, de nem ismerjük mindenkiért), akkor következtethetünk-e azon adatokra, amelyeket nem ismerünk? Mikor? Hogyan? Ha egy jelenségre sok dolog hat, akkor szétválaszthatjuk-e az egyes hatások, megmondhatjuk-e, hogy ennyiben felelős ezért az egyik és mennyiben a másik hatás?

A statisztika tudomány alapvetően két kérdésre keresi a választ. Egyrészt arra, amit *adatsűrítés*nek nevezhetünk: ha van sok-sok adatunk, akkor ezt hogyan lehet viszonylag kevés mutatóval leírni. Vannak-e olyan jellemzői ezeknek az adatokat, amelyeket, ha megadunk, akkor többszáz, többezer adat helyett egy-két mutatóval helyettesíthetjük azokat. (Az adatsűrítés logikáját nem nehéz belátni: amikor azt kérdezzük, hogy milyen nehéz volt egy vizsga, akkor nem azt várjuk, hogy az adott vizsgán megszerzett rengeteg jegyet közöljék velünk – általában megelégszünk az átlaggal. Egy adattal. Tegyük hozzá – mint majd látjuk – a statisztikusok nem ezzel az egyetlen adattal dolgoznak; sőt azt is megkérdőjelezhetjük, hogy az átlag ebben az esetben jó mutató-e.)

A másik kérdés a következtetés. Ebben az esetben az a probléma, hogy tudjuk, hogy nem ismerünk minden adatot – csak azok egy részét. Mégis képesek vagyunk arra, hogy kijelentéseket tegyünk azokról is, akiket nem ismerünk. Ennek a módszereit írja le a statisztikai következtetéselmélet. Ezt is bemutatja, és egyben ennek a korlátaira is felhívja a figyelmet a fejezet.

Mindkét probléma esetén fel fog azonban bukkanni ugyanaz a „módszertani nyitottság". A statisztika sokféleképpen tud adatot sűríteni is, és a következtetései is sok elemtől függenek. És minden módszer más és más eredményt ad. A statisztikus számára a legjobb az (lenne), ha a megrendelő tisztában van (lenne) azzal, hogy pontosan mit is akar, és akkor ezen módszerek közül egyértelmű (lenne), hogy melyiket kell választania. De ez a ritkább eset. Általában a

[+] Pázmány Péter Katolikus Egyetem Jog és Államtudományi Kar egyetemi docens, Budapesti Corvinus Egyetem egyetemi docens. szalai.akos@jak.ppke.hu

megrendelő nem teszi meg ezeket a választásokat, és a statisztikusnak kell eldöntenie, hogy milyen eszközökkel nyúl az adott kérdéshez. És könnyen lehet, hogy ha más módszert választ, akkor ugyanarra a kérdésre gyökeresen eltérő választ ad. Például arra, hogy a mottóban szereplő A vagy B esetben történt-e diszkrimináció, illetve, hogy a C esetben mennyit ér az ingatlan.

Másik oldalon viszont a statisztikai vizsgálatok olvasói – így a jogászok is – általában elsikkadnak ezen módszertani döntések fölött. A fejezet célja annyi, hogy az olvasók, amikor egy statisztikai vizsgálat eredményeivel szembesülnek (nem az a cél, hogy ők tudjanak ilyeneket készíteni!), akkor értsék ezen döntések jelentőségét. Értsék, hogy az előttük levő érték nagyjából mennyiben lenne más, ha a statisztikus más módszertani döntéseket hoz. (Tegyük azonban hozzá: vannak „tilos" módszertani döntések is, amik a statisztika tudománya szerint bizonyosan rossz eredményeket adnak. Ilyennel, rendes statisztikai elemzésben nem találkozunk – de sajnos az olvasók találkozni fognak rossz elemzésekkel is. Ezek felismerésében is segíteni próbál ez a fejezet.)

14.1. Adatbázis

A statisztikai vizsgálatok adatokból indulnak ki, azokat elemzik. De mielőtt nekikezdünk az elemzésnek mindig először az adatbázis tartalmát kell pontosan megérteni, mert enélkül komoly tévedéseket követhetünk el. Az adatbázis alapesetben egy táblázat – sorokkal, oszlopokkal. A leginkább bevett felírás szerint az egyes sorokban vannak az egyes egyedek, azokról az oszlopokban szereplő információkat (változókat) gyűjtjük össze. Az adatbázis e két dimenzióját tekintjük át elsőként.

14.1.1. Egyedek

Azokat, akikről adataink vannak *egyedek*nek nevezzük. Az egyedek kapcsán három kérdést kell megválaszolni – attól függően, hogy mi a válasz, mást és mást kezdhetünk az adatbázissal. A kérdések: (i) mik az egyes egyedek, (ii) hány elem van, (iii) minden egyedet ismerünk, vagy csak egy részüket.

Az egyedek típusa alapján a legtöbbször ún. *keresztmetszeti elemzésről* van szó. Ebben az esetben az egyedek emberek, vállalatok, intézmények, állatok, stb. A lényeg hogy nem fontos a sorrendjük, nincsenek egymással határos, egymáshoz közelebb vagy egymásól távolabb álló elemek. A másik tipikus adatbázis az *idősor*. Ilyenkor ugyanazon emberek, vállalatok, intézmények, állatok, stb. adatait ismerjük, de különböző időpontokban. Itt a sorrend nyilvánvaló: a decemberi és a februári között szerepel a januári adat; a tavalyelőtti és az idei között a tavalyi. A harmadik adatbázistípus a *területi* adatsor. Ebben az esetben az egyes sorok egy-egy földrajz terület (város, megye, ország, stb.) adatait tartalmazzák.

> A földrajzi adatbázisok és az idősorok közös jellemzője, hogy az egymáshoz közeli egyedek adatai tipikusan befolyásolják egymást – például „húzzák egymást" lefelé, vagy felfelé. (Tegyük hozzá: olyan is van, amikor az egyik magas értékét „kiegyenlíti" a szomszéd alacsony értéke.) Az egyes egymáshoz közeli adatok nem függetlenek egymástól. A földrajzi adatok esetén a probléma bonyolultabb, mert a közelség és a távolság fogalma is bonyolultabb, mint idősornál. Utóbbi esetén minden egyednek két szomszédja van, egy kisebb és egy nagyobb. Az egyik szomszédtól a másikba csak a közöttük levő egyeden keresztül – ha tetszik: két lépésben – lehet eljutni. Ez a földrajzi adatok esetében nem igaz. Például Belgium határos Hollandiával és Franciaországgal is, és az is igaz, hogy egyikből a másikba csak két határt átlépve, Belgiumon keresztül lehet eljutni. Viszont,

> Belgium határos Németországgal is, de oda Franciaországból és Hollandiából is vezet közvetlen út is – hiszen ők is határosak vele.

Figyelni kell arra, hogy mit is írnak le az adatok, mik is az egyedek. „Kisebb egységre" általában nem vonhatunk le következtetéseket. Ez lenne az ún. *ökológiai tévedés*. Például, ha valaki azt mutatja meg, hogy azokban az országokban (államokban, megyékben), amelyekben magasabb valamilyen jellemző egy másik jellemző is magasabb, még nem bizonyítja, hogy e két jellemző az egyének szintjén is összefügg.

> Ilyenkor tipikusa az adott térség átlaga szerepel az adatbázisban. Ez az átlag azonban nagyon sokmindent elfed – például az adott országban élő egyes emberek közötti különbségeket. (Lásd erről a 2. alfejezetet.)
> Sir Robert Doll például 1955-ben publikálta ilyen tévedést (is) tartalmazó elemzését a dohányzás és a tüdőrák közötti kapcsolatról. (Doll [1955]) Ebben az egyik bizonyíték egy olyan ábra volt, amelyből egyértelműen látszott, hogy azokban az országokban, ahol magas az egy főre jutó cigarettafogyasztás magas a tüdőrák miatt bekövetkezett halálozás aránya is. De ebből (még) nem következik, hogy *azok, akik* többet dohányoznak nagyobb eséllyel betegszenek meg tüdőrákban. Csak az, hogy *azokban az országokban, ahol* sokat dohányoznak magas a tüdőrák miatti halálozás. Annak megmutatásához, hogy ez az összefüggés az emberek szintjén is igaz, egyes emberekről (dohányzási szokásairól, betegségeiről) kell adatokat gyűjteni. (Ez később meg is történt....)

Ebben a fejezetben alapvetően a keresztmetszeti vizsgálatok módszertanának alapjait fogjuk megismerni. Ez a legegyszerűbb. Amennyiben valamilyen technika nem használható a másik két adatbázis esetében, akkor azt jelezni fogjuk – viszont nem cél ezek elemzési módszertanát is megismertetni itt.

Az egyedek száma alapján meg szoktunk különböztetni *kis és nagy adatbázisokat*. Ez azért lesz fontos, mert egyes itt bemutatott módszerek csak nagy adatbázisok esetén használhatóak biztosan – kisebbek esetén más módszereket kell alkalmazni.

A teljesség kérdése pedig a statisztika legérdekesebb problémájához vezet el. Szinte soha nem ismerjük mindenki (minden ember, vállalat, intézmény, stb.) adatát. Csak egy részükét: csak egy *mintát* ismerünk – és nem a *teljes sokaság, populáció* adatait. A statisztika legizgalmasabb területe az ún. *következtetéselmélet*: ez arra keresi a választ, hogy az ismert minta alapján miként becsülhetjük meg, hogy a teljes sokaságban (a nem ismert egyedeket is ideértve) mekkora valaminek az értéke, igaz-e egy állítás, stb.. Az ismert mintából következtetünk a teljes sokaságra.

> Ki kell térni arra a mondatrészre, hogy „*szinte*" soha nem ismerjük. A statisztikát szabályozó jogszabályok ugyanis gyakran arra törekszenek, hogy teljeskörű adatbázisokat építsenek ki. Kötelezővé teszik az adatszolgáltatást, vagy éppen népszámlálásokat írnak ki. Ezzel kívánják elkerülni a „kimaradók" problémáját.[1]
> Ugyanakkor a statisztikusok ezeket a kötelező teljeskörű adatbázisokat is általában csak mintaként kezelik – nem a teljes populációként. Ennek oka az, hogy az igazán érdekes kérdések ritkán azok, amelyek azt kérdezik, hogy egy pillanatban a felvett adatok mit mutatnak. Inkább általános tendenciákra vagyunk kíváncsiak – amelyek az adott adatbázisba bekerülőkön túl is igaz lenne. Például, ha azt kérdezzük, hogy egy vizsga mennyire nehéz, akkor nem arra vagyunk kíváncsiak, hogy az egy, a két, vagy az öt évvel ezelőtti vizsgázók milyen eredményt értek el. Az általuk elért eredményből akarunk következtetni arra, hogy a következő vizsgaidőszaban (vagy

[1] Kimaradók természetesen mindig vannak – például, akik a kötelezettségüket nem teljesítik. Persze a számuk lényegesen kisebb, mint ha eleve nem is próbálkoznánk teljeskörű adatgyűjtéssel.

még később) mennyire lesz nehéz a vizsga. Ráadásul az előző évek eredményeinek alakulását sok más elem is befolyásolta, nem csak a vizsga nehézsége. (Például a hallgatók felkészültsége.) Ez az elmúlt évek vizsgáiból származó „minta" sajátossága – annak a véletlennek tudható be, hogy azokban az években éppen azok a hallgatók vizsgáztak, akik. Nem tudjuk, hogy ha mások vizsgáztak volna (mondjuk azok, akik idén fognak), ugyanolyan eredményt értek volna-e el. Vagyis az adatbázisból (amely valójában csak egy minta) akarunk következtetéseket levonni a teljes populáció más tagjaira, azokra, akik ezután fognak vizsgázni.

Vannak persze olyan elemek is, amikor a teljes populáció nyilvánvalóan megegyezik a mintával – például egy bíróság ügyben, amikor egyedi esetről kell véleményt mondani. De a statisztika ilyenkor is előszeretettel tekinti az adott ügyet „mintának": olyan hipotetikus „képzeletbeli ügyekkel" állítja azt szembe, amikor az adott bizonyítékok kerülnének elő, illetve amikor az adott üggyel megegyező helyzetű személyek kerülnének bíróság elé. (Németh [2020]) Ezzel arra hívjuk fel a figyelmet, hogy adott esetben is lehetnek véletlen zavaró tényezők. (Mint majd a 3. fejezetben látjuk, a statisztikai következtetéselmélet – különösen az ún. statisztikai bizonyítás – és a bírósági bizonyítás nagyon hasonló logikát követ.)

14.1. szövegdoboz: Az adatok forrásai és az ezzel összefüggő torzítások

Az adatbázis létrejötte kapcsán három (talán pontosabb, ha azt mondjuk: négy) fontosabb forrást kell megkülönböztetni: a kísérletet (ezen belül az ún. természetes kísérletet, megfigyelést), a kérdőíves vizsgálatot és a már publikált eredmények összefoglalását.

A *kísérlet* klasszikus – vélhetően mindenki előtt ismert – formája a gyógyszerkísérlet. Ennek példáján jól megérthetjük a kísérletek logikáját. Az egyedeket két csoportra osztjuk: az egyik lesz az ún. *kezelt*, a másik a *kontrollcsoport*. A gyógyszert, amelynek a hatására vagyunk kíváncsiak a kezelt csoport megkapja – a kontrollcsoport nem. *Kezelés*nek nevezzük azt, hogy az adott csoportot kitesszük annak a hatásnak, amelynek vizsgálata a célunk. A vizsgálat során általában kétszer végeznek mérést arról a jellemzőről, amelyre várakozásaink szerint a kezelés hat. Egyszer a kezelés előtt, egyszer akkor, amikor a kezelés hatása (várakozásaink szerint) már jelentkezik. Ha mind a kontrollcsoportnál, mind a kezelt csoportnál megtesszük ezt, akkor a *két mérés eltérésének különbsége* az, amit vizsgálni fogunk: mennyivel jobban változott a mért eredmény a kezelt csoportban, mint a kontrollcsoportban. (Ugyanis sok külső ok miatt a kontrollcsoportban is változhat az eredmény.[2]) Ugyanakkor a legtöbb kísérlet elemzésekor fontos, hogy a két csoport egyéb jellemzőit is „ellenőrzés alatt tartsuk" – azok hatását „kiszűrjük". Ezért a két csoport egyéb, a hatás szempontjából (várakozásaink szerint) fontos jellemzőit is mérni, rögzíteni szokták Az adatbázisnak ez is része.

A *természetes kísérlet* – amit nevezhetünk megfigyelésnek is – esetén a kezelést, a hatást nem a vizsgálat során kapják a kezelt csoport tagjai, hanem a kísérlettől függetlenül megkapták. Például valamilyen környezeti, vagy történeti hatásnak ki voltak téve. Ezek után egyszerűen megfigyeljük, hogy a két csoport jellemzői eltérnek-e. (Ezek a vizsgálatok általában csak egyszer mérnek: a kezelés, a „behatás" után.)

Mindkét kísérlet esetén elég fontos, hogy a kezelt és a kontrollcsoport tagjai között – lehetőleg – csak az legyen a különbség, hogy megkapták-e a kezelést. Más tekintetben legyenek – nagyjából – azonosak. Éppen ezért ezt vizsgálnunk kell. Mindkét kísérletnél vizsgálni lehet a

[2] De a placebo-hatás is megjelenik. Ennek lényege, hogy az adott egyed ne tudja, hogy melyik csoportba tartozik, és pusztán azért, mert azt hiszi, hogy ő is megkapja a kezelést, az ő esetében is megváltozhat az eredmény.

két csoport egyéb jellemzőit. (Ugyanakkor a klasszikus kísérletnél sokszor elég az is, ha egyszerűen véletlenszerűen osztjuk ketté a kísérletben résztvevőket.)

A kísérletek esetén tipikusan „objektív" adatokat rögzítünk. Az egyedek fizikai jellemzőit, vagy éppen adott helyzetben hozott döntéseiket. (Megvásárolnak-e valamit, elfogadnak-e egy ajánlatot, válaszolnak-e egy levélre, stb.) A kérdőíves vizsgálatok tipikusan hipotetikus vizsgálatok: úgy rögzítjük az egyes csoportok jellemzőit, hogy megkérdezzük őket valamilyen tulajdonságukról, vagy éppen arra kérjük őket, hogy képzeljék el hogyan döntenének egy elképzelt helyzetben. (Például kire szavaz*nának*, ha most vasárnap lenne a választás?) A kérdőíves vizsgálatokat épp ez a feltételesség, illetve az önbevallás különbözteti meg a kísérletektől. Emiatt az adatok „megbízhatatlanabbak": nem biztos, hogy a válaszadó igazat mond, és nem biztos, hogy valóban úgy döntene a valóságban is, ahogy egy ilyen hipotetikus kérdésre válaszol. Kérdőíves vizsgálatok esetén az irodalom (például: Jackson et al [2011] 472-474) többféle torzítást különböztet meg. Ilyen például:

- a kérdések félreértése: a kérdező és a kérdezett nem ugyanazt érti egy adott kérdésen. Gondoljunk például egy olyan kérdésre, amelyben a szürke, fekete jövedelmek elterjedtségére kíváncsi valaki: tapasztalt-e ilyet, fizetett-e, kapott-e ilyet a kérdezett. Ebben az esetben például komoly eltérést okozhat az, ha valaki a borravalót (pincérnek, taxisnak, stb.) ilyennek minősíti, vagy sem.

- a megfelelési kényszer: nyilvánvaló, hogy egy kérdőíves vizsgálatkor egy idegennek nem szívesen vallják be a kérdezettek, ha valami olyat tettek, olyan tulajdonságuk van, ami szerintük nem helyes, „megvetendő". Nyilvánvaló például, hogy viszonylag kevesen lennének hajlandóak bevallani, hogy egy „komolyabb" korrupciós cselekményben rész vettek. (De már más lesz a helyzet, ha valaki a hálapénzt, a borravalót is korrupciónak tekinti.) [3]

- a csomagolási hatás: a korlátozott racionalitás egyik fontos eleme a csomagolási hatás ebben az esetben is jelentkezhet. Ha ugyanazt a kérdést nem ugyanúgy (vagy ugyanúgy, de más kérdések után) tesszük fel, akkor emiatt megváltozhat a válasz.

- az elsietett (gyors) válaszok: különösen hosszabb kérdőívek kapcsán gyakori, hogy – egy idő után – a válaszadó a kérdéseket próbálja rövidre zárni, minél előbb szabadulni. A válaszok egyre kevésbé lesznek átgondoltak, megfontoltak. (Tegyük hozzá: ez lényegesen nagyobb problémát okoz, mint, ha egyszerűen megtagadnák a választ, mert a válaszmegtagadás, a hiányzó válasz látszik az adatbázisban. Az ilyen „összecsapott válaszok" viszont megjelennek benne – és nem lehet őket megkülönböztetni a „valós válaszoktól".)

A harmadik (vagy ha a klasszikus és a természetes kísérletet két külön forrásnak tekintjük, akkor a negyedik) forrás az ún. *összefoglaló elemzés*. Ilyenkor az adott kérdésben korábban mások által publikált eredményeket összesítjük. (Például az elmúlt időszakban a globális felmelegedés nagyságáról és annak hatásairól, vagy annak „okairól" publikált tanulmányokat, azok eredményeit vetjük össze.) A tudományos életben egyre inkább elvárás, hogy az ilyen elemzések adatbázisai (készüljenek azok akár kísérletek, akár kérdőívek alapján) elérhetőek legyenek más kutatók számára is. De ezek az adatbázisok további torzításokat tartalmazhatnak. Az egyik legfontosabb, hogy ha nem maga az elemző gyűjti az adatokat, akkor kevésbé fogja felismerni, ha a mérés nem pontosan azt méri, amit a kutató szerint kellene. Ha például tartunk attól, hogy egy kérdőíves vizsgálatkor nem valós válaszok születnek, akkor

[3] Ez a megfelelési kényszer különösen akkor zavaró, ha nem ilyen „nyilvánvaló". Például, ha a kérdezett olyan válaszokat igyekszik adni, amelyet szerinte a kérdező elvár. Ez sokszor nem egyértelmű a kérdező számára sem.

ezt a vizsgálatot végző tesztelni szokta – például mielőtt sok embert megkeresne a kérdőívvel, előbb egy kisebb csoporton teszteli azt. De, ha más adataiból dolgozunk, akkor nem biztos, hogy ezeket a torzításokat az új elemző felismeri. A másik roppant fontos torzítás az ún. *íróasztalfiók-hatás*. Tudjuk ugyanis, hogy a publikált eredmények (szinte mindig) azok, amelyek egy szakmai feltételezést igazolnak. Azok a vizsgálatok, amelyek nem igazolják a kutató feltevését, általában „elsüllyednek" egy-egy fiókban. Ezek a szakértői feltételezést alá nem támasztó vizsgálatok – tipikusan – nem kerülnek a közönség elé. (Kivétel ez alól az, ha a kutatói társadalom erősen megosztott – például mint a halálbüntetés hatásainak elemzésekor –, és mind a két oldal publikálja a saját feltevéseit igazoló eredményeket. Igaz egyik csoport sem teszi ezt meg azokkal az eredményeivel, vizsgálataival, amelyek a saját álláspontját gyengítenék, a másik oldal igazát támasztanák alá. Ezek ugyanúgy elsüllyednek egy-egy fiókban.)

14.2. szövegdoboz: Mintavétel

Amikor eldöntjük, hogy ki kerülnek (kerülhetnek) a mintába, kézenfekvőnek tűnik, hogy *reprezentativitásra* törekedjünk. Vagyis arra, hogy a minta a lehető leginkább hasonlítson a teljes sokaságra. A teljes sokaságot felosztjuk különböző csoportokra, és azt vizsgáljuk, hogy az egyes soportokba tartozók aránya a mintében és a sokaságban közel ugyanolyan-e. (Tegyük azonban hozzá: a „hasonlóság" nem csak mérési kérdés, hanem – legalább ennyire – elméleti is. Ugyanis ennek kapcsán az első kérdés: milyen szempontból hasonlítson. A válaszadók testmagassága szerint, jövedelme szerint, végzettsége szerint, nyelvtudása szerint?)

A mintavétel kapcsán tipikusan véletlen és nem véletlen mintavételi technikákat különböztetünk meg.

A nem véletlen mintavételi technikák kapcsán általában vagy „garantálni akarjuk" a reprezentativitást, vagy egyszerűen a mintavétellel – egyébként – járó problémákat, költségeket akarjuk csökkenteni.

Ilyen mintavételi technika például az ún. *kvótás mintavétel*. Ekkor eleve előírjuk, hogy a mintába milyen csoportból hányan kerüljenek. A kvótás mintavétel esetén azt, hogy az egyes kvótaszámokat hogyan töltjük fel nem szabályozzuk. Például a kérdezőbiztosra van bízva, hogy hol talál megfelelő számú adott csoportba tartozó résztvevőt, válaszadót.

Míg a kvótás mintavételnél a reprezentativitás a fő szempont, addig például a *hólabda módszer* esetén elsősorban a gyorsaság, az egyszerűség a kérdés. Ilyenkor egyszerűen a vizsgálatba már bevont személytől kérünk ötleteket arra, hogy kiket keressünk még meg.

Véletlen mintavétel esetén a reprezentativitás kevésbé fontos: kis túlzással abból indulunk ki, hogy amennyiben a véletlen alakítja a mintát, akkor a „nagy számok törvénye szerint" nagyjából reprezentatív is lesz az.[4] A leggyakrabban alkalmazott (megcélzott) véletlen mintavételi eljárás az ún. *egyszerű véletlen mintavétel*, amikor a cél az, hogy a teljes sokaság minden egyede egyforma eséllyel kerüljön be a mintába. A legegyszerűbb eljárás, hogy egyszerűen sorsolják őket. (Ezen belül ismerünk visszatevéses és a visszatevés nélküli sorsolást is.) Ez a mintavételi eljárás ugyan roppant vonzó a statisztikusok, az elemzők számára – a véletlen mintáknak

[4] Bár ilyenkor is szokták utólag tesztelni, hogy a kialakuló minta reprezentatív-e – és például súlyozással korrigáljuk az eltéréseket: nagyobb súlyt adunk azok adatainak, akik nem kellően reprezentált csoportba tartoznak, stb.

nagyon sok „kedvező tulajdonsága” van –, azonban általában igen komoly problémát jelent a teljes populáció előzetes „listázása”.

Rétegzett mintavétel esetén a kvótás és a véletlen mintavételt ötvözzük: először meghatározzuk a minta „reprezentativitáshoz szükséges” összetételét. A teljes sokaságot felosztjuk ezekbe a kisebb csoportokba, majd az egyes csoportokon belül kisorsoljuk a mintába kerülőket. Ez a vizsgálat még bonyolultabb mint az egyszerű véletlen mintavétel: előzetesen a populáció minden tagjáról tudni kellene, hogy melyik csoportba tartozik.

A *csoportos* mintavétel nem az egyedeket választják ki, hanem azt a csoportot, amelynek az össze tagjának az adatait felveszik. A *többlépcsős* mintavétel esetén először egy csoportot választanak, majd azon belül az egyedeket véletlen mintavétel alapján választják ki.

És a mintavétel lezárásaként mindenképpen szólni kell az „önszelekció problémájáról”. Bármilyen módon választjuk is ki a mintát, mindenképpen számolni kell azzal, hogy a kiválasztottak nem akarnak részt venni az adatszolgáltatásban. Megtagadják az együttműködést. Nem nehéz persze „helyetteseket” találni (egyszerűen a kellő számú egyeden túl eleve „póttagokat” is ki kell jelölni). nagyobb gond az, hogy a minta reprezentativitását ez mindenképpen érinti. Azok, akik nem vesznek részt a vizsgálatban nem ugyanolyanok, mint azok, aki részt vesznek. Épp a „visszautasítás” mutatja, hogy eltérnek. (Az utóbbi években a választási előrejelzések esetén éppen ezért az egyik legfontosabb kérdés a „válaszmegtagadók”, a „rejtőzködők” tulajdonságainak azonosítása lett.)

14.1.2. Változók

Az adatbázis másik dimenzióját a *változók* adják. A változók kapcsán az alapvető kérdés az ún. *mérési szint*. Megkülönböztetünk kvalitatív (minőségi, nominális), ordinális és magas mérési szintű változókat. Ez határozza meg, hogy mit tehetünk az adatokkal. (Például, hogy értelmese átlagot számítani belőlük.) Azok a technikák, amit a „gyengébb változóknál” megfelelők, azok általában alkalmazhatók az erősebbeknél is; fordítva viszont nem.

Kvalitatív, vagy *minőségi*, vagy *nominális változók* esetén azok értékei között nincs jobb vagy rosszabb, több vagy kevesebb. Ezek olyan kérdések kapcsán állnak elő, amelyekre nem szám (vagy sorrend) a válasz. Egyszerűen azt írják le, hogy melyik egyed melyikhez hasonlít, és melyiktől különbözik.

> Nyilvánvalóan ilyen változó valaminek a színe. De ilyen például a földrajzi helyzet is. (Például egy adott bíróság székhelyét jelentő város, vagy megye megnevezése.)

A *dichotóm változók* a minőségi változók egy alesetét képezik: csak két minőségi kategória van. De ezek között itt sincs sorrend. Viszont mivel csak két kategória van, így ebben az esetben néhány technika alkalmazható, ami a kvalitatív változóknál nem.

> Például a színeket nem érdemes átlagolni. Viszont, ha egy adatbázisban a férfiakat 1-sel, a nőket pedig 0-val jelöljük, akkor ennek a „nem” változónak az átlaga értelmes: azt mutatja, hogy az adatbázisban szereplő egyedek mekkora része férfi. (Ha már nem csak férfi és nő nemeket ismerünk, vagyis többértékűvé válik a változó, akkor az átlagszámítás elveszíti értemét.)

A *kvantitatív* változók esetén egy szám a válasz. de nem mindegy, hogy milyen szám. Ezen belül két mérési szintet ismerünk: az ordinális, és az ún. magas szintű változókat.

Ordinális változók esetén a sorrend egyértelmű, de az értékek közötti pontos távolság nem számszerűsíthető. Például a bírósági szintek, vagy az iskolai végzettség (általános, középfok,

érettségi, alapdiploma, mesterképzés stb.) esetén van kisebb és nagyobb. Ezeket könnyű számmá kódolni: (például) az alacsonyabb kapjon kisebb számértéket (például 0-t vagy 1-et). De nem érdemes például átlagos iskolai végzettséget számolni ennek alapján.

A *magas mérési szintű* változók esetén már nem csak a sorrend, hanem két érték közötti különbség is egyértelmű. Ha az egyik ember ügyében tíz hónap alatt születik bírósági döntés, a másikében tizenegy, míg a harmadikében tizenkettő alatt – akkor ugyanúgy egy-egy hónap a különbség.

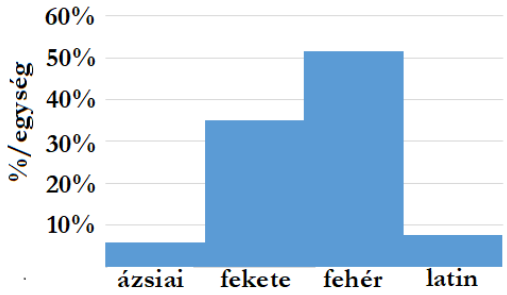
Időnként egy-egy változó többféleképpen is értelmezhető. Az például egyértelmű: ha egy írásbeli vizsgán elért pontszámokat (százalékokat) rögzítjük, az magas szintű változó. De az érdemjegy mérési szintje már erősen vitatható. Átlagot számítunk belőle mintha magas mérési szintű változó lenne. Holott az egyes jegyek közötti távolság nyilvánvalóan nem ugyanakkora

> A magas mérési szintű változók lehetnek diszkrétek és folytonosak. Diszkrét a változó, ha csak bizonyos értékeket vehet fel. Folytonos, ha adott (minimum és maximum) határok között bármilyen értéket.

14.2. Leíró statisztika: egyváltozós elemzés

A leíró statisztika az adatsürítés eszköze. Fő feladata az, hogy amikor rengeteg adattal rendelkezünk az egyedek valamilyen jellemzőjéről (változójáról), vagy éppen többféle változójáról, akkor ezt a rengeteg adatot egy-két ábrába, táblázatba, mutatóba sűrítsük. Elvileg választhatnánk azt a technikát is, hogy a teljes adatbázist, a sok-sok adatot odaadjuk. De a statisztikus arra törekszik, hogy a rengeteg adatot, eggyel-kettővel (egy-két mérőszámmal), illetve egy-két bevett ábrázolási formával helyettesítse.

Ebben az alfejezetben az ún. egyváltozós elemzéssel ismerkedünk meg, vagyis azzal, amikor egy változóról áll rendelkezésünkre rengeteg adat (szerencsés esetben annyi, ahány egyedünk van). Először egy vizuális megjelenítési formával, az eloszlás ábrázolásával ismerkedünk meg. Ennek célja, hogy felismerjünk abban valamilyen alakot felismerjünk. Ezt követően térünk rá a változók leírására használt két legfontosabb mutatótípus ismertetésére.



14.1. ábra: A foglalkoztatottak rassz szerinti megoszlása egy munkahelyen (Egy hipotetikus vállalat példáján)

14.2.1. Eloszlás, sűrűségfüggvény

Az egy változóval kapcsolatos adatsűrítés legfontosabb módja az, ha megadjuk, felrajzoljuk annak *sűrűségfüggvényét*. Ilyet láthatunk a 14.1 ábrán – kvalitatív változók esetére. A

koordinátarendszer két tengelyén egyrészt a különböző kategóriák, másrészt az egyes kategóriákba tartozó egyedek (elemek) aránya – statisztikai nyelven: gyakorisága – szerepel.

Az 14.1. ábrához hasonló ábra szerkesztése kapcsán az egyik legfontosabb probléma, hogy mik legyenek a kategóriák, és milyen sorrendben ábrázoljuk azokat.

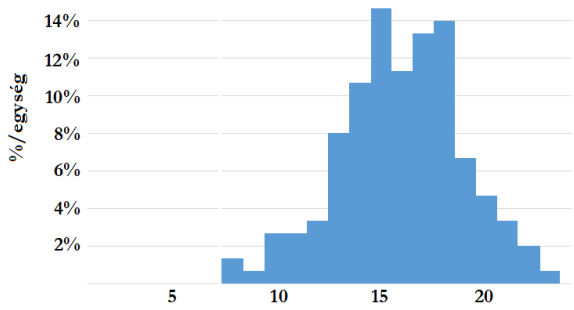
Kvalitatív változók esetén csak az utóbbi a kérdés: milyen sorrendben kerüljenek azok fel az ábrára. Az ábra alakja attól függően változik, hogy milyen ez a sorend.

Az *ordinális* változók esetén, amennyiben „megfelelő számú" értékünk van, akkor az ábrázolás egyszerű: azt vesszük fel, hogy az egyes katagóriákba az összes egyed hány százaléka tartozik. A sorrendet megadja az ordinális változó.

A legnagyobb problémát a *magas mérési szint* esetén tapasztaljuk. Itt mindenképpen dönteni kell arról, hogyan alakítjuk ki az egyes kategóriákat. (Ez ordinális mérési szintnél is probléma, ha nagyon sok érték jelentkezik abban.)

Ezen kategóriákat ebben az esetben *osztályközöknek* nevezzük: ez adja meg azt, hogy mettől meddig tart egy-egy kategória. (Például, ha egy csoport jövedelmeit írja le egy változó, akkor nem mindegy, hogy 1.000 forintos, 10.000 forintos, vagy 50.000 forintos osztályközöket használunk.) Az osztályközök kialakításakor az első kérdés, hogy hány kategória legyen. Ha nagyon kevés kategóriát veszünk fel, akkor az ábra semmit nem mutat. (Szélsőséges esetben képzeljük el, hogy csak egyetlen kategóriát veszünk fel, vagyis a kategória alsó határa a minimális, felső határa pedig a maximális érték: ha ezt tesszük, akkor az ábra azt mutatja majd, hogy az adatok 100%-a ebbe az osztályközbe tartozik.) Ha nagyon sok kategóriát különítünk el, akkor az ábra várhatóan „csipkézett" lesz: az egymást követő kategóriák közül várhatóan az egyik kicsit magasabb, a másik kicsit alacsonyabb lesz az őt megelőzőnél. (Ezt mutatja a 14.2. ábra.) Ezen az ábrán nehéz felismerni valamilyen alakot – holott az ábrázolás célja éppen ez.

> Persze időnként – ha ez a kérdés – éppen ez a „csipkézettség" az érdekes. Például a 14.2. ábrán, amely egy felvételi vizsgán elért eredmények megoszlását mutatja. Látjuk: a 16 ponthoz kisebb oszlop tartozik, mint a 15 és a 17 ponthoz. Amikor ilyet látunk elgondolkodhatunk, hogy miért esik vissza a 16 pontosok száma a 15 és a 17-hez képest. Eszünkbe juthat, hogy a 17 pontos eredmény nem volt-e valamiért kitüntetett – mondjuk, nem 17 pont kellett-e a sikeres vizsgához. Nem lehet-e, hogy a vizsgáztatók ezért „áttoltak" abba a kategóriába néhány vizsgázót a 16 pontosok közül.



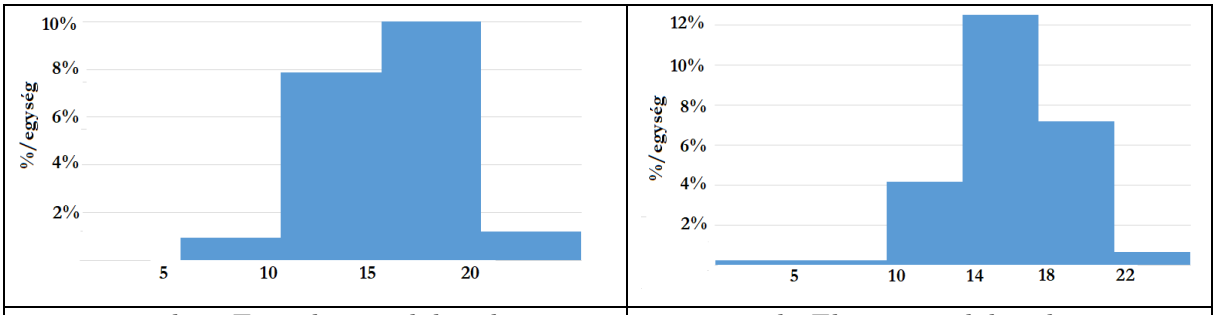
14.2. ábra: Egy felvételi vizsgán az elért pontszámok eloszlása
(A maximum 25 pont)

Tipikus megoldás az ún. *egyenlő osztályközös* ábrázolás: ilyenkor az egyes kategóriák ugyanolyan hosszúak. Ebben az esetben tarthatjuk magunkat a fent leíráshoz: az oszlopok magasságát az adja, hogy abba a kategóriába az összes egyed hány százaléka tartozik. De ez az egyenlő osztályköz praktikus okokból nem mindig (sőt esetben: nem) szerencsés.

> Például a jövedelmek esetén, ha alacsony jövedelmeknél érdemes is 30-50.000 forintonként kialakítani az osztályközöket, a magasabb jövedelmeknél tipikusan szélesebb kategóriákat alkalmazunk – sőt, a „legtetején" tipikusan azt mondjuk, hogy adott összegnél többet keresők mind egy „legmagasabb jövedelmi kategóriába" kerülnek.

Érdemes felfigyelni rá, hogy a függőleges tengelyen a mértékegység az egy egysége jutó százalék. Ennek akkor van igazán jelentősége, ha eltérő osztályközöket ábrázolunk. Ilyenkor fontos szem előtt tartani egy alapelvet: *nem az adott kategóriába tartozó oszlop magassága, hanem annak területe mutatja az abba a kategóriába tartozók arányt.* (Ezt láthatjuk a 14.3. ábrán is.) Ha egy kategória kétszer olyan széles, de ugyanannyi egyed tartozik abba, mint egy fele olyan széles kategóriába, akkor fele olyan magas oszlopot kap. A legjobb ha azt tartjuk szem előtt, hogy az egyes oszlopok magasságát úgy kapjuk, hogy az adott osztályközbe tartozó egyedek arányát elosztjuk az adott kategória szélességével (az intervallum hosszával). Az így szerkesztett ábrák az ún. *sűrűsödést* mutatják: az egy vízszintes egységre eső százalékarányt. Ez az ún. *sűrűségfüggvény.*

> A sűrűségfüggvény fontos tulajdonsága, hogy amennyiben minden egységnél megnézzük, hogy ahhoz milyen magas oszlop tartozik, és ezeket az értékeket összeadjuk, akkor 100%-ot kapunk.

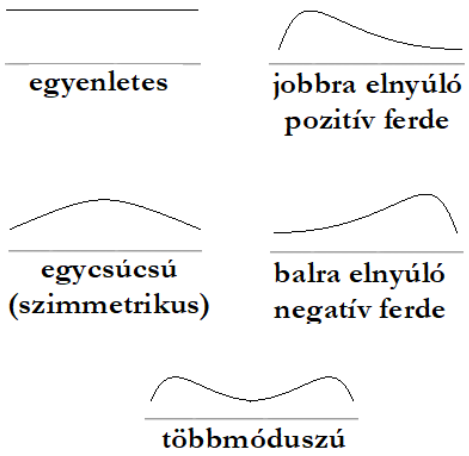


a. ábra: Egyenlő osztályközök | b. Eltérő osztályközök

14.3. ábra: A 14.2. ábrán szereplő pontszámok eloszlása különböző osztályközökkel ábrázolva
a. ábra: 5 egységenként; b ábra: 0-9 és onnan négy egységenként
(A b. ábrán a 0-9 osztályköznél az oszlop azért alacsony, mert a 8-9 pontos eredmények (lásd 14.2 ábra) a 0-9 osztályra széthúzva jelebik meg.)

Az ábrázolás célja az, hogy felismerjünk abban valamilyen alakot. A következő lépés ezért az, hogy az ábrát összevetjük bizonyos tipikus alakokkal, eloszlásfajtákkal. Amennyiben a kategóriák egyértelműen sorba rendezhetők (vagyis legalább ordinális változóval van dolgunk), akkor az eloszlás alakja kapcsán elsősorban a csúcspontok számát érdemes megvizsgálni. Háromféle eloszlást különböztethetünk meg: az egycsúcsú (vagy egymóduszú), a többcsúcsú (többmóduszú) és az egyenletes eloszlást. Ezekre láthatunk példát a 14.4. ábrán. (i) *Egyenletes eloszlás* esetén nincs olyan kategória, amely (jelentősen) kiemelkedne a többi közül, vagy jelentősen elmaradna a többitől. (ii) *Egycsúcsú, vagy egymóduszú* az *eloszlás*, ha ettől a csúcspontttól jobbra és balra távolodva folyamatosan csökken (vagy legalábbis jelentősen nem

nő) az egyes kategóriákba tartozó egyede száma, aránya. Azt a kategóriát, amelybe a legtöbb egyed tartozik nevezzük *módusz*nak. (iii) *Többcsúcsú (többmóduszú) eloszlás* esetén nő, és ezért több csúcspontot különíthetünk el.



14.4. ábra: Különböző tipikus eloszlások

Tegyük hozzá: az osztályközök száma erősen befolyásolja azt, hogy egy eloszlás egymóduszúnak, vagy többmóduszúnak látszik-e. Ha nagyon sok kategóriát veszünk fel, akkora „csipkézettség" pont azt jelent, hogy többcsúcsú lesz az ábra, mint a 14.2. ábra.

A következő kérdés a *ferdeség*, ami szintén az könnyedén leolvasható az ábráról. Amennyiben az eloszlás egycsúcsú, akkor megkülönböztetünk szimmetrikus, pozitív ferde (vagy más néven jobbra elnyúló) és negatív ferde (más néven: balra elnyúló eloszlást).

1. *Szimmetrikus* az eloszlás, ha a középponttól jobbra és balra ugyanannyival elmozdulva hasonló értékeket látunk.

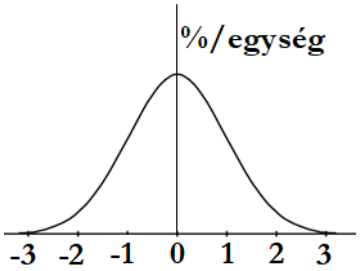
   Tegyük hozzá: a szimmetria pontos definíciója nem csak a 14.4. ábrán egycsúcsú (szimmetrikus) eloszlásnak nevezett alak esetén igaz, hanem az egyenletes, és az ábrán bemutatott többcsúcsú eloszlás esetén is. Ezeknél is igaz, hogy a középponttól (amely azonban ezek esetén nem „kiemelkedés", nem módusz) távolodva ugyanolyan függvényértékeket kapunk.

2. *Pozitív ferde*, vagy más néven *jobbra elnyúló eloszlás* esetén néhány kiugró magas érték okozza az eloszlás aszimmetriáját, vagyis a középpontól ugyanolyan távol jobbra nagyobb értékeket látunk (magasabbak az oszlopok, magasabb a függvény), mint balra.
3. *Negatív ferdeség*, vagy *balra elnyúló eloszlás* esetén fordítva: a középpontnál sokkal kisebb értékből van több.

   Érdemes felfigyelni rá, hogy attól függően, hogy ugyanazon adatsor (például a 14.2.-14.3 ábra) alakja eltérhet attól függően, hogy milyen osztályközöket használunk. A statisztikusok ezért inkább ferdeségi mérőszámokkal dolgoznak, amelyeket a nemsokára bemutatott két középérték (az átlag és a medián) közötti viszony alapján képeznek.

A sok eloszlás között kitüntetett szerepet játszik az ún. *normális eloszlás*, a *normálgörbe*. (De ez csak elnevezés – ez nem jelenti azt, hogy a többi eloszlás „hibás".) Ilyet látunk a 14.5. ábrán. Ez egy szimmetrikus eloszlás; a sokak által jól ismert „haranggörbe" – pontosabba annak egy konkrét változata.

Haraggörbe sokféle van. A normális eloszlás pontos definícióját annak képlete adja meg. Ez nagyon csúnya: $y = \frac{100\%}{2\pi} e^{-x^2/2}$ (Itt e az ún. Euler-féle szám, a természetes alapú logaritmus alapja, ami 2,719). Egészen pontosan ez az ún. standard normális eloszlás, amikor a átlag éppen nulla, vagyis az x tengelyen az átlagtól való (pozitív vagy negatív) eltérés szerepel. (Egészen pontosan az, hogy a 3. pontban bemutatásra kerülő szórásban mérve mekkora az eltérés.)



14.5. ábra: Normális eloszlás

A normál eloszlás a továbbiakban fontos szerepet játszik majd. Ez a függvény igazából az, ami leírja azt, amit a köznyelv a „nagy számok törvényeként" ismer. A nagy számok törvénye ugyanis nem azt mondja ki, hogy ha egy véletlen által befolyásolt (kockázatos) esemény sokszor megtörténik, akkor annak bekövetkezési gyakorisága az ún. várható értékhez (az átlaghoz) tart. (Vagyis, például, ha többször dobunk egy normális pénzérmével, akkor a fejek aránya közelíteni fog az 50%-hoz.) A nagy számok törvénye azt mondja ki, hogy ha sokszor megismételjük a véletlen által befolyásolt „kísérletet", akkor annak valós értéke úgy fog ingadozni a várható érték körül, ahogyan ez a függvény leírja. Minél alacsonyabb a függvény, annál kisebb az esélye (de soha nem kizárt), hogy a valós érték épp annyival térjen el a várható értéktől.

Ezek közül a statisztika tudománya számára legfontosabb összefüggés az, hogy ha véletlenszerűen (visszatevéssel) kerülnek egyedek egy kellően nagy mintába, akkor a minta átlaga úgy fog viszonyulni a valósághoz (a teljes sokaság átlagához) ahogyan ez a függvény leírja. Lehet, hogy elmarad tőle, lehet, hogy nagyobb lesz annál. De a függvényről leolvasható, hogy mekkora eséllyel, milyen gyakran lesz az átlagtól való eltérés éppen akkora.

Érdemes már ezen a ponton kiemelni a *sűrűségfüggvények* egyik fontos tulajdonságát. Ha tetszőlegesen kijelölünk két értéket a vízszintes tengelyen, akkor megbecsülhetjük, hogy az összes egyed hány százaléka esik e két érték közé. Ehhez nem kell mást tennünk, mint megnézni, hogy e két érték között az oszlopok területe mekkora.

> Például a normálgörbén tudható, hogy -1 és 1 között 68%; -2 és 2 között 95%; -3 és 3 között 99,7% van. Itt azonban nem oszlopokkal dolgozunk, hanem egyszerűen a függvény alatti területet számoljuk ki.

14.2.2. Középértékek

Az eloszlások jellemzésére nem csak ábrákat használunk, hanem számokat is. Ezek közül a leginkább „kézenfekvő" a középérték különösen az átlag. De ilyen „tipikus" értékként szoktunk beszélni a móduszról és a mediánról is.

Az *átlag* képlete, számítása (remélhetően) nem okoz gondot: $\bar{x} = \frac{\sum x_i}{n}$. Összeadjuk a változók értékeit (x_i) és elosztjuk azok számával (n). Érdemes azonban egy-két megjegyzést tenni. Egyrészt mivel a változókat össze kell adni, így csak magas mérési szintű változók esetén alkalmazható.

> Annak az állításnak nyilvánvalóan van értelme, hogy ha valaki az egyik hónapban 300 ezer forintot keres a másikban pedig 200 ezret, akkor a két hónapban összesen ötszázat. Ebből tudjuk, hogy átlagosan 250 ezret keresett. De annak van értelme, hogy ha valaki az egyik tárgyból kettest kap a másikból pedig hármast, akkor kettőből együtt ötöst? Márpedig, ha ennek nincs, akkor erősen megkérdőjelezhető, hogy annak van-e értelme annak, hogy azt mondjuk, hogy az átlaga 2,5. (Más lenne a helyzet, ha azt mondanánk, hogy az egyik vizsgán 65%-ra a másikon 55%-ra teljesített, és így a kettő átlagában 60%-ot ért el. Mert a százalékban megadott teljesítmény már egyértelműen magas mérési szintű változó!)

Másrészt, az átlag nem középen áll. Bár kézenfekvőnek tűnik, hogy az egyedek fele az átlag alatt, a fele meg fölötte van, de ez nincs így. (Pontosabban: csak szimmetrikus eloszlás esetén van így.) Az átlag ún. *számított középérték* – nem helyzeti. Amennyiben az eloszlás ferde, akkor ez a ferdeség húzza az átlagot is. Egy pozitív ferde, jobbra elnyúló eloszlás esetén a kiugró magas értékek miatt magasabb lesz az átlag. (Egy balra elnyúló, negatív ferde eloszlásnál pedig alacsonyabb lesz az átlag.)

> Az átlag megértéséhez érdemes talán Freedman és szerzőtársai példáját segítségül hívni! Ha előző pontban látott sűrűségfüggvény oszlopait úgy képzeljük el, mint egy deszkán elhelyezkedő különböző nagyságú dobozokat (amelyek súlya arányos a nagyságukkal), akkor az átlag mutatja azt a pontot, ahol a deszkát alá kell támasztani ahhoz, hogy az ne billenjen el se jobbra se balra. „A mérleghintán egy kicsi gyerek a középponttól távolabb ül, hogy egyensúlyt tartson a középponthoz közelebb ülő nagyobb gyerekkel." Vagyis az „egyensúlyi pont", az átlag a nagyobbhoz közelebb lesz. (Freedman et al [2005] 83-84.)

Az az érték, amelytől jobbra és balra épp ugyanannyi egyed van a *medián*. A medián ún. *helyzeti középérték*. Számítása – mivel az adatokon belül vannak kisebbek és nagyobbak – csak akkor lehetséges, ha minimum ordinális változókkal dolgozunk.

A *móduszról* az eloszlás kapcsán előbb volt szó: ez a változó leggyakoribb értéke. Ugyanakkor ezt igazán akkor érdemes csak használni, ha a kategóriák „egyértelműek". Ha a látott osztályköz-problémák fellépnek, akkor a módusz nagyon érzékeny lesz arra, hogy milyen értékeket sorolunk egy-egy kategóriába.

A különböző mérési szintű változóknál tehát más és más középértéket érdemes használni. Ezt mutatja a 14.1. táblázat.

Mérési szint	Használható középérték	
Kvalitatív	módusz	
Orinális	módusz/medián	
Magas mérési szintű	átlag/medián	ha szimmetrikus: átlag = medián ha pozitív ferde (jobbra elnyúló): átlag > medián ha negatív ferde (jobbra elnyúló): átlag < medián

14.1. táblázat: Használható középértékek változók mérési szintje szerint

14.2.3. Sokszínűség, szóródás

A középérték csak azt mutatja, hogy mi lenne a „tipikus". De ez a tipikus nagyon sok módon előállhat. Az átlag mögött nagyon sokszínű lehet a valóság.

> Például lehet valakinek úgy hármas az átlaga (már persze, ha ennek van bármi jelentése...), hogy
> a. minden tárgyból közepes
> b. a tárgyak feléből jeles, a feléből megbukik
> c. a tárgyak ötödéből bukott, ötödéből elégséges, ötödéből közepes, ötödéből jó és ötödéből jelese volt.

A – statisztika nyelvén a – *szóródás* mérésekor ezeket az eltéréseket, ezt a sokszínűséget próbáljuk egy-két viszonylag bevett mutató segítségével számszerűsíteni. Itt is igaz, hogy a különböző mérési szintű változók esetében másként és másként lehet ezt megtenni. Használhatunk távolság-mutatókat, vagy a – statisztikusok által leginkább kedvelt, és legtöbbször alkalmazott – szórást, illetve varianciát.

Az első eszköz, amelyet meg kell ismernünk a *kvantilis*ekre bontás. Ebben az esetben arról van szó, hogy az egyedeket egyforma létszámú, egyedszámú csoportokra bontjuk aszerint, hogy az adott változó értéke kinél kisebb, kinél nagyobb. Például, ha negyedeket, ún. kvartiliseket, keresünk, akkor megkeressük azt a három értéket, amelynél épp az adatok 25%-a, 50%-a és 75%-a kisebb. Ezzel négy csoport áll elő: az egyedek legkisebb értékkel rendelkező, a második, a harmadik és legnagyobb értékkel rendelkező negyede. De bonthatjuk akármennyi egyforma nagyságú csoportra. Vannak kitüntetett, névvel ellátott kvantlisek. Ezek

- a *kvartilisek*, ami – mint láttuk – négy egyenlő csoportra osztja az egyedeket hoz létre.
- a *kvintilisek*, ezek öt csoportot hoznak létre (ezért azokat az értékeket keressük, amelyeknél az változók 20%-a, 40%-a, 60%-a és 80%-a kisebb)
- a *decilisek*, amelyek tíz egyenlő csoportra bontják a teljes csoportot,
- a *percentilisek*, amelyek 100 egyenlő csoportra. Vagyis a 32. percentilis értéknél épp a változók 32%-a kisebb (és 68%-a nagyobb). A medián pedig egyszerűen az 50. percentilis-érték.

Ezeket az értékeket – sok más mellett – használhatjuk például ún. *terjedelemmutatók*, vagy más néven *távolságmutatók* készítésére. Ez az egyik leggyakrabban használt szóródási, sokszínűségi mérőszám. Azt vizsgálják, hogy milyen távol van egymástól valamely csoportok legkisebb és legnagyobb értéke.

> Mit mond el az a sokaságról, ha a legnagyobb adat (de csak az) kétszeresre nő? Ez egy ilyen, a minimumot a maximummal összevető mutatót kétszeresére növelne. Még akkor is, ha a többi adat változatlan.

Tipikusan terjedelmi mérőszám az ún. *interkvartilis terjedelem*. Ez a kvartiliseknél látott négyrészre-osztással dolgozik, de csak a középső két negyedre figyel, ennek alsó és felső határát veti össze. A statisztikában igen ritka, hogy egyszerűen a legkisebb és legnagyobb értéket hasonlítsuk össze. Ennek oka, hogy ezek sokszor ún. *kilógó adatok*: nagyon távol esnek a többitől, kifejezetten nem-tipikusak. Az interkvartilis terjedelem épp ezeket a kilógó adatokat vágja le.

14.3. szövegdoboz: A jövede*lemegyenlőtlenség mérése – sokszínűség a jövedelmekben*

A jövedelmi egyenlőtlenség (az ún. jövedelemeloszlás) vizsgálatakor általában nem kvartiliekkel (és interkvartilis terjedelemmel), hanem a decilisekkel dolgozunk: az legalsó decilis felső határa (a 10%-os pont) és a legfelső declis alsó határa (a 90%-os pont) egymáshoz viszonyított arányát adjuk meg.

De még gyakoribb, hogy inkább az ún. Lorenz-görbét rajzoljuk fel, és a Gini-mutatót adjuk meg. A Lorenz-görbét a 14.sz.1. ábrán látjuk. A görbe egyes pontjai azt mutatják, hogy a jövedelemeloszlás alján található meghatározott nagyságú (10%, 16%, 45%, 73, %, stb.) csoportok az összjövedelem hány százalékával rendelkeznek. A görbe úgy készül, hogy

(i) a jövedelemnagyság szerint növekvő sorrendbe rendezzük az egyedeket, majd
(ii) mindenkinél azt az értéket vesszük fel, amennyivel (amilyen aránnyal) ő és a nála szegényebbek az összjövedelemből rendelkeznek.

A görbe alakja – épp mert a legszegényebbek vannak balra, és jobbra haladva egyre tehetősebbek következnek – homorú, mint az ábrán is látszik.

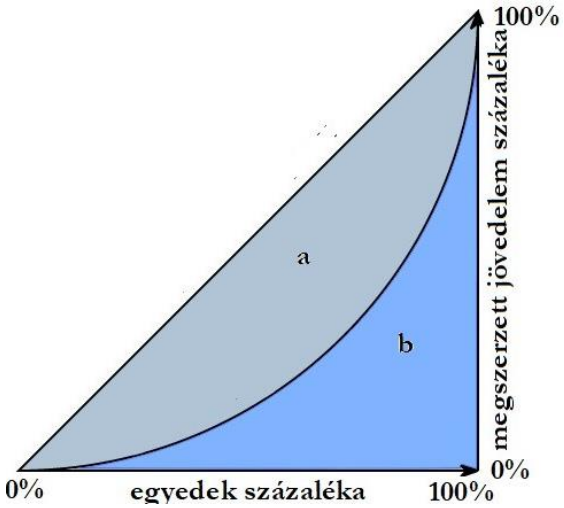
A Gini-mutató, ebből a Lorenz-görbéből készül. Ez két részre osztja a 45°-os egyenes által kirajzolt háromszöget. Van a görbe fölötti terület (ezt jelöli az ábrán a), és a görbe alatti (ez a b). Ha az a terület nagyságát elosztjuk a teljes háromszög területével, vagyis az $(a+b)$-vel, akkor megkapjuk a Gini-mutatót. Ez egy 0 és 1 közötti érték lesz. 0 akkor lenne, ha a Lorenz görbe épp a 45°-os egyenes lenne – vagyis, ha mindenkinek ugyanakkora a jövedelme. (Vagyis az alsó x% épp az összjövedelem ugyanekkora, x%-ával rendelkezne. Mert mindenki ugyanannyival rendelkezik.) 1 pedig akkor lenne a Gini, ha a görbe alatt nem lenne semmi. Ez akkor állna elő, ha egyetlen ember kivételével senkinek nem lenne jövedelme – egyetlen ember kapja az összest. Vagyis

– ha a Gini 0, akkor az a tökéletes egyenlőtlenséget,

– ha pedig 1, akkor az a legnagyobb egyenlőtlenséget jelzi.

Ha két csoport (vagy két ország) Gini-ét összevetjük, akkor a nagyobb érték nagyobb egyenlőtlenséget jelez.

A jövedelemeloszlás vizsgálatakor a Lorenz-görbe és a Gini-mutató kedveltebb, mint távolságmutatók (például a p90/p10). Mindenekelőtt azért, mert a távolságmutatók érzéketlenek arra, hogy mit látunk a jövedelemeloszlás közepén. Csak azt elemzik, hogy adott nagyságú csoport (p90/p10 esetén az emberek 80%-a) milyen szélső értékek között mozog. A Gini viszont, mivel a görbe fölötti terület a kiindulópontja, arra is érzékeny, hogy a középosztályoknál mennyire magasan, vagy alacsonyan, mennyire állandó meredekséggel halad a görbe.



14.sz.1. ábra: Lorenz-görbe

Magas mérési szintű változók esetén a leggyakrabban alkalmazott sokszínűségi, szóródási mutató a *szórás*. Ez röviden: ez az átlagtól vett átlagos négyzetes eltérés gyöke. Képlettel:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}},$$

Amennyiben egy változó csak két értéket (egy kisebbet és egy nagyobbat) vehet fel akkor a szórás képlete még egyszerűbb: $\sigma = (X - Y) \times \sqrt{P_x \times P_Y}$,
ahol X: a nagyobb érték, Y: a kisebb érték, P_X: a nagyobb érték aránya (gyakorisága) az egyedek között, P_Y: a kisebb érték aránya (gyakorisága).

A szórás ugyanabban a mértékegységben lesz, mint a változó, könnyű azzal, vagy az átlaggal összevetni. Könnyű például megadni egy olyan sávot, amely az átlagtól egy vagy két szórásnyira van. (Egyszerűen egyszer hozzáadjuk, és egyszer kivonjuk az átlagból a szórást, vagy annak kétszeresét.) Ezeknek a sávoknak a kijelölése nem érdektelen! Az ún. *Csejbisev-egyenlőtlenség* szerint például biztos, hogy az egyedek minimum 75%-a esetén a változó értéke az átlagtól két szórásnyira húzott sávban van. (És minimum 8/9-e maximum három szórásnyira.)

A Csejbisev-egyenlőtlenség képlete: $P(|X - \bar{X}| < k \times \sigma) > 1 - \frac{1}{k^2}$. (ahol k>1) Ez a legáltalánosabb egyenlet, itt semmit nem kell tudni az eloszlásról (a 14.4 ábra alapján bármilyen lehet). Ha tudjuk, hogy az eloszlás szimmetrikus, akkor az „erősebb" Gauss-féle egyenletet használhatjuk. E szerint: $P(|X - \bar{X}| < k \times \sigma) > 1 - \frac{4}{9k^2}$

Ugyanakkor a szórás az átlaggal dolgozik, amely az eloszlás ferdeségére, a nagyon magas, vagy nagyon alacsony adatokra érzékeny. Ezért (erősen) ferde eloszlásoknál sok statisztikus nem is javasolja annak használatát. (Simon [2020])

A *variancia* a szórás négyzete, vagyis az átlagtól vett négyzetes eltérés. Ez ugyanakkor – mivel négyzetre emelt értékekkel dolgozik, a gyökvonás „korrekciója" nélkül – lényegesen nagyobb értékeket mutat majd, mint a változók. Ha a szórás 30, akkor a variancia 900. A szórás közvetlenül összevethető az átlaggal. A variancia annál lényegesen nagyobb – gyakorlatilag más mértékegységben van..

14.3. Leíró statisztika: változók közötti összefüggések

Adatsűrítésre nem csak akkor van szükség, ha egy változót akarunk jellemezni, hanem akkor is, ha az a kérdés, hogy két változó között milyen a kapcsolat. Például, általános feltételezés (és általában igazolható is), hogy aki magasabb végzettséget szerez, az magasabb jövedelemre tesz szert. Ezt az összefüggést is megmutathatjuk úgy, hogy sok-sok egyed képzettségi és jövedelmi adatait mind odaadjuk az érdeklődőnek, de – ha értjük a most következő statisztikai fogalmakat – lényegesen egyszerűbb, ha egy-két ennek az összefüggésnek az erősségét mérő mutatószámot adunk meg.

Ebben a fejezetben csak a legegyszerűbbeket az ún. Cramer-féle asszociációs mutatót és a korrelációt mutatjuk be, illetve két olyan eszközt, amely az összefüggés vizuális megjelenítését teszi egyszerűbbé. (A harmadik fontos, a magas mérési szintű változóknál leggyakrabban alkalmazott mutatót a regressziót az ötödik fejezet tárgyalja majd.) De először a változók közötti kapcsolat egy-két fontos definícióját kell megismerni.

14.3.1. Változók közötti összefüggések formái

Első lépésként a változók közötti kapcsolat lehetséges „erősségét" leíró fogalmakat kell számba venni. Két változó között

- lehet *függvényszerű*, vagy *determinisztikus kapcsolat*, amikor az egyik változó *egyértelműen meghatározza* a másik értékét.
- lehet *sztochasztikus kapcsolat*, amikor az egyik nem határozza ugyan meg a másik értékét (sokféle lehet az), de az egyik változó valamilyen értéke *valószínűbbé teszi* a másik valamilyen értékét.

  Például az előbb említett képzettség és jövedelem közötti kapcsolat ilyen: a magasabb képzettségből nem következik, hogy ki mennyit fog keresni, de a magasabb képzettséget szerzők valószínűleg többet.

- lehet, hogy semmilyen kapcsolat nincs – ekkor beszélünk a változók *függetlenségéről*.

Az összefüggésmutatók általában arra keresik a választ, hogy a háromféle kapcsolat közül melyik áll fenn – illetve, sztochasztikus kapcsolat esetén azt, hogy az összefüggés mennyire erős, mennyire van közel a determinisztikus viszonyhoz.

14.3.2. A változók közötti kapcsolatok ábrázolása: kereszttáblák és pontdiagramok

A változók közötti kapcsolat érzékeltetésére a legegyszerűbb módszer a *kereszttábla*. Ilyet mutat a 14.2 táblázat. Ez a (már kvalitatív szintű változók esetén is alkalmazható) eszköz egyszerűen azt mutatja, hogy ha az egyedeket a két változó kategóriái szerint csoportokra bontjuk, akkor az egyes „kombinált" csoportokban hány egyed, vagy az összes egyed mekkora része lesz. Érdemes már itt kiemelni, hogy a kereszttáblák mindig tartamaznak egy „összesen" sort és oszlopot is. Ezek azt mutatják, hogy összesen hány egyed tartozik az egyes sorokban, illetve oszlopokban szereplő kategóriákba.

Amennyiben két (minimum) ordinális szintű változó közötti kapcsolatot akarunk ábrázolni, akkor kézenfekvő választás a 14.6. ábrán látható *pontdiagram* is. Ebben az egyes pontok az egyes egyedeket jelzik – helyzetük pedig azt, hogy esetükben a két változó értéke mekkora. Az ábrára (az összes pontra) ránézve először azt keressük, hogy kirajzolódik-e valamilyen alakzat. Például emelkedő, vagy csökkenő sávban helyezkednek-e el a pontok; valamilyen U-alakot vesznek-e fel; hasonlítanak-e valamilyen klasszikus függvényre (egy egyenesre, egy

négyzetfüggvényre, egy gyökfüggvényre, egy logaritmus-függvényre, egy hiperbolára, stb.). Ezeken az ábrákon a kapcsolat erősségét az mutatja, hogy ez az alakzat mennyire egyértelmű – az egyes pontok mennyire „rendeződnek" valamilyen ilyen alakzatba. És fordítva: mennyi „képzelőerő kell", hogy belelássuk az alakzatot a ponthalmazba.

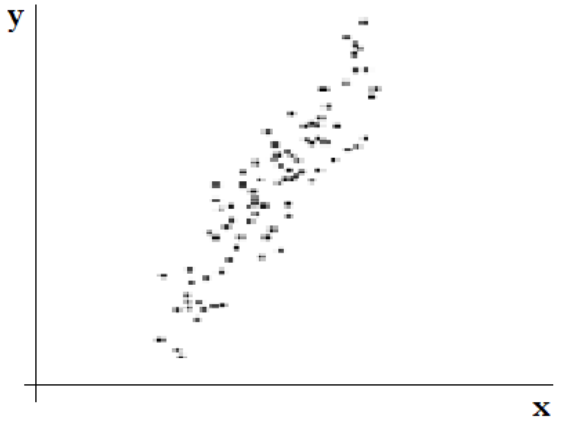
	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	17	20	30	20	13	100
Nő	8	30	45	55	12	150
Összesen	25	50	75	75	25	250

a. Kereszttábla elemszámmal

	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	4%	8%	12%	12%	4%	40%
Nő	6%	12%	18%	18%	6%	60%
Összesen	10%	20%	30%	30%	10%	100%

b. Kereszttábla gyakorisággal

14.2. táblázat: Egy vizsga eredményei jegyek és nemek szerinti bontásban



14.6. ábra: Pontdiagram: egyedek két változó (X és Y) szerint

	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	25	50	0	0	25	100
Nő	0	0	75	75	0	150
Összesen	25	50	75	75	25	250

a. Kereszttábla elemszámmal

	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	10%	20%	0%	0%	10%	40%
Nő	0%	0%	30%	30%	0%	60%
Összesen	10%	20%	30%	30%	10%	100%

b. Kereszttábla gyakorisággal

14.3. táblázat: Egy lehetséges determinisztikus viszony két változó között (A 14.2. táblázatban leírt összefüggés esetén – egy lehetséges determinisztikus viszony)

14.3.3. Az összefüggés erősségének mérése

Az összefüggéseket nem csak vizualizálni tudjuk, hanem azok erősségét statisztikai mutatókkal is megadhatjuk. A két legegyszerűbbet mutatjuk most be: a kereszttábla esetén alkalmazott Cramer-féle mutatót és a pontdiagramok esetén alkalmazható a regresszió is.

A kereszttáblák esetén először is érdemes végiggondolni, hogy hogyan néznének azok ki determinisztikus kapcsolat esetén, és akkor, ha a változók között nem lenne semmiféle kapcsolat.

Determinisztikus kapcsolatot látunk a 14.3. táblázatban. Azért determinisztikus a viszony, mert minden oszlopban egyetlen olyan mező van, ahol nullától eltérő érték szerepel. Vagyis, ha tudjuk valamelyik egyedről, hogy a melyik oszlopban van (az oszlopokban ábrázolt kategóriák közül melyikbe tartozik), akkor ebből egyértelműen következik, hogy melyik sorba kerül. Az oszlopban szereplő változó a *magyarázó (független) változó*; a sorokban szereplő változó a *függő változó.*

> Vegyük azonban észre, hogy ez fordítva nem igaz: van olyan sor, ahol két nullánál magasabb szám szerepel.

A kapcsolat hiányára a 14.4. táblázat mutat példát. Ennek áttekintése, megértése kissé bonyolultabb – de elengedhetetlen ahhoz, hogy a kapcsolat erősségének mérőszámait értelmezni tudjuk. A legegyszerűbben akkor láthatjuk ezt be, ha minimális matematikát bevetünk. A 14.4.b. táblázatban nem az szerepel, hogy melyik kategóriába hány egyed tartozik, hanem azt, hogy az összes egyed mekkora része. Viszont az is igaz, hogy az egyes kombinált kategóriákban szereplő értékek pontosan megegyeznek az „összesen" sorokban és oszlopokban szereplő értékek szorzataival. Ez a matematikai összefüggés a függetlenség definíciója.

	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	10	20	30	30	10	100
Nő	15	30	45	45	15	150
Összesen	25	50	75	75	25	250

a. Kereszttábla elemszámmal

	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	4%	8%	12%	12%	4%	40%
Nő	6%	12%	18%	18%	6%	60%
Összesen	10%	20%	30%	30%	10%	100%

b. Kereszttábla gyakorisággal

	Elégtelen	Elégséges	Közepes	Jó	Jeles	Összesen
Férfi	10%	20%	30%	30%	10%	100%
Nő	10%	20%	30%	30%	10%	100%
Összesen	10%	20%	30%	30%	10%	100%

c. Kereszttábla feltételes eloszlásokkal

14.4. táblázat: Két független változó közötti viszony
(A 14.2. táblázatban leírt összefüggés esetén)

Ugyanezt a kapcsolatot írják le a klasszikus statisztika tankönyvek úgy, hogy bevezetnek két új kategóriát: a feltételes és a feltétel nélküli megoszlást. Ez ugyan két új (talán bonyolult) kategória, de a függetlenség fogalmát (talán) könnyebben értelmezhetővé teszik.

- A *feltétel nélküli megoszlás* nem más, mint, hogy az összes egyed hogyan oszlik meg a sorokban, illetve az oszlopokban szereplő kategóriák között. Vagyis a feltétel nélküli eloszlás szerepel a 14.1.b táblázat utolsó sorában és oszlopában. (A 14.3.b. táblázatban ez egyes cellákban ezen feltétel nélküli valószínűségek szorzata szerepel.)
- A *feltételes valószínűséget* pedig úgy kapjuk, ha megnézzük, hogy egyes sorokon belül hogyan oszlanak meg az egyes egyedek. Vagyis azt nézzük, hogy az első, a második, a harmadik, stb. sorban szereplő egyedek mekkora része melyik oszlopba tartozik. Ezekben az esetekben az egyes sorok utolsó oszlopa 100% lesz – hiszen az ebben a sorban szereplő egyedeket osztottuk csak fel az oszlopok között.

A 14.3.c táblázat ezt a felosztást mutatja – ugyanarra az esetre, amire az 14.3.a táblázat az egyedek számát mutatta. Látszik, hogy az egyes sorokban (beleértve az utolsó „összesen" sort is) ugyanazok a számok szerepelnek. A statisztika úgy fogalmaz: „a feltételes megoszlások megegyeznek egymással". Statisztikus számára ezt jelenti az, hogy az oszlopokban jelzett kategóriák függetlenek a sorban szereplő kategóriáktól.[5]

A Cramer-féle asszociációs mutató, mint a kapcsolat erősségének mutatószáma egyszerűen azt méri, hogy egy kereszttábla mennyiben tér el attól, mint amilyen akkor lenne, ha a változók függetlenek lennének. Ez a mutató egy nulla és egy közötti számot ad. Nulla akkor, ha a változók függetlenek, és egy akkor, ha azok függvényszerű, determinisztikus kapcsolatban vannak. Minél nagyobb az érték, annál erősebb a kapcsolat.

A mutató értékét ugyan a statisztikusok is egyszerűen számítógépes programokkal számoltatják ki, de érdemes a logikát áttekinteni. A számítás két lépésből áll. Először kiszámoljuk a χ^2-nek nevezett mutatót. Majd ebből az asszociációs mutatót.

A Cramer-féle asszociációs mutató képlete: $\sqrt{\frac{\chi^2}{min(sorok\ száma,\ \ oszlopok\ száma)-1}}$,

ahol $\chi^2 = \sum\sum \frac{\left(n_{ij}-n_{ij}^*\right)^2}{n_{ij}^*}$

Vagyis χ^2 kiszámítása három lépésből áll.

- Először minden cellában megnézzük, hogy az egyedek száma (n_{ij}) mennyivel tér el attól, ami függetlenség esetén ott lenne (n_{ij}^*).
- Majd ezen eltéréseknek a négyzetét vesszük, és elosztjuk azzal az elemszámmal, ami függetlenség esetén szerepelne ott. Így minden cellának lesz egy „eltérésértéke".
- A χ^2 egyszerűen ezen értékek összege.

A Cramer-féle mutató pedig ebből számítható, úgy, hogy...

- ...az így kapott értéket elosztjuk az ún. *szabadságfokkal*, ami nem más, mint az oszlopok vagy a sorok száma közül a kisebbik mínusz egy. (Vagyis, ha öt sorkategóriánk és nyolc oszlopkategóriánk lenne, akkor néggyel – mivel az öt a kisebb, és ennél eggyel kevesebbel kell osztani.)
- ...majd az így kapott értéknek a négyzetgyökét vesszük – így „küszöbölve ki" azt, hogy a χ^2 számításakor négyzetre emeltünk.

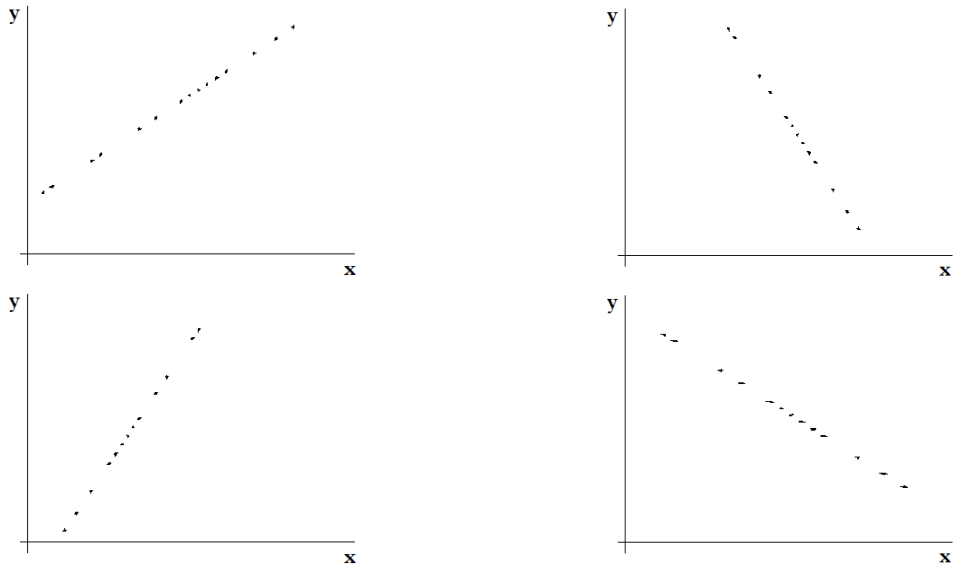
Érdemes kiemelni, hogy sztochasztikus kapcsolat esetén az, hogy melyik változó függő és melyik független nem állapítható meg úgy, mint függvényszerű kapcsolatnál. (Ott, ugye az lehetett a szabály, hogy magyarázó változó az, ami determinál, aminek minden értékénél csak egyetlen

[5] A vizsgálat elvégezhető fordítva is, vagyis, amikor az egyes oszlopokban szereplő csoportokban szereplő egyedekről írjuk fel, hogy mekkora részük melyik sorba kerül. Függetlenség esetén ebben az esetben is igaz, hogy az egyes (feltételes) megoszlások megegyeznek.

nullától eltérő értéket találunk.) Itt mind a sor, mind az oszlop játszhatja mindkét szerepet: statisztikai értelemben magyarázhatjuk a sorral az oszlopot, de az oszloppal is a sort

A magas mérési szintű változók közötti kapcsolat (vagyis pontdiagramok) esetében – tipikus esetben – az ún. *Pearsons-féle lineáris korrelációs együtthatót*, vagy *Pearsons-féle r*-t számítjuk. Ez egy -1 és 1 közötti mutató, amely azt mutatja, hogy a pontok mennyire illeszkednek egy egyenesre. Ha az értéke

- 1, akkor egy növekvő egyenesen van az összes pont (lásd 14.7.a ábra),
- -1, akkor egy csökkenő egyenesen (lásd 14.7.b. ábra).
- 0 akkor nincs lineáris kapcsolat a két változó között.

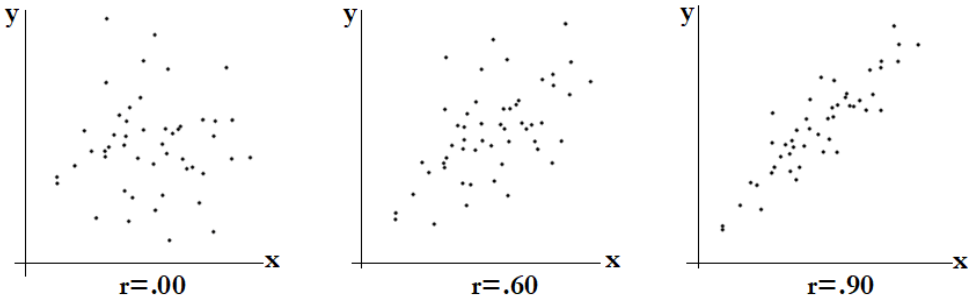


a. r = 1 b. r = -1

14.7. ábra: Két magas mérési szintű változó közötti determinisztikus kapcsolat pontdiagramon, korrelációs együtthatókkal

Érdemes kiemelni (látszik a 14.7. ábrán is), hogy a korrelációs együttható nagysága arra nem érzékeny. hogy milyen meredek egyenesre illeszkednek a pontok. Ha bármilyen (akármilyen meredek) egyenesre tökéletesen illeszkednek, akkor az értéke 1, vagy -1 lesz. (Az ötödik fejezetben bemutatott regressziószámítás foglalkozik majd az egyenes meredekségével, azzal, hogy e tekintetben milyen a két változó közötti összefüggés.)

Minél távolabb van *r* értéke 0-tól (minél közelebb 1-hez vagy -1-hez), annál erősebb a viszony. Hogy ez az erősség mit jelent, azt a 14.8. ábrán láthatjuk.



14.8. ábra: Különböző korrelációs együtthatók és a pontdiagram alakja

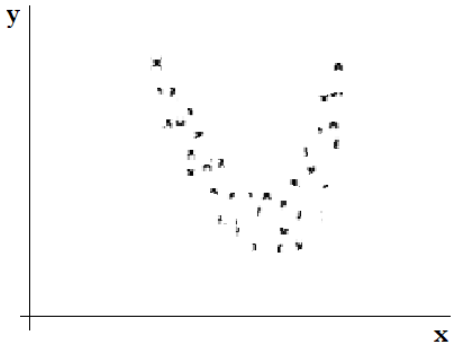
A korrelációs együttható, r kiszámítása, képlete:

$$r = \frac{\sum\left[\frac{(x_i - \bar{x})}{\sigma_x} \times \frac{(y - \bar{y})}{\sigma_y}\right]}{n}$$

Vagyis minden egyed minden változójánál (x_i és y_i) megnézzük, hogy hány szórásnyira (σ_x és σ_y) is van az az adott változó átlagától ($\bar{x}$ és $\bar{y}$), és minden egyednél összeszorozzuk a két, így kapott értéket. A korrelációs együttható ezen szorzatok átlaga.[6]
Ebből is következik, hogy a korrelációs együttható „szimmetrikus", vagyis az x és y közötti korreláció megegyezik az y és x közötti korrelációval.

Ki kell emelni, hogy ez a korrelációs mutató csak lineáris kapcsolatot mutat ki. Ha nem lineáris az összefüggés, akkor nem tér el az értéke nullától.

Tegyük fel például, hogy a két változó közötti viszony nem lineáris – mondjuk egy U alakot követ. (Mint a 14.9. ábrán.) A magyarázó változó növekedésével a függő változó előbb csökken, majd nő. (Ilyen például az életkor és a templomba járás közötti viszony.) Ez esetben a lineáris korrelációs együttható alacsony lesz. Holott van kapcsolat – csak nem egy egyenesre, hanem más függvényre illeszkednek a pontok.



14.9. ábra: U alakú összefüggés és a pontdiagram alakja

[6] A korreláció legtöbb könyvben szereplő képlete nem ez, hanem az, amikor *n-1*-gyel osztunk. A gyakorlatban ugyanis szinte mindig ezt használjuk. Ez azért van, mert – szinte – soha nem arra vagyunk kíváncsiak, amit most – didaktikai okból – tárgyalunk: nem egy adatbázison belül keresünk összefüggést, hanem a rendelkezésre álló adatokból (a mintából) próbálunk következtetni a teljes sokaságban e két változó között meglevő kapcsolatra. És ezt már az n-1-gyel osztó képlettel kell becsülni.

Nem lineáris kapcsolat esetén tehát a Pearsons-féle r „fals alacsony" értéket mutat. De kaphatunk „fals magas" értéket is. Elsősorban akkor, ha kilógó pontokat tartalmaz az ábra. Ezek „el tudják húzni" az összefüggést a saját irányukba. Minél távolabb vannak a többi ponttól, annál inkább.

14.3.4. „Statisztikai okság"

Az előbb úgy fogalmaztunk, hogy magyarázó (független) és függő változók vannak. Nem úgy, ahogyan „kézenfekvő lenne": nem okról és okozatról beszéltünk. Ennek „okaira" érdemes egy – az oksággal sokat foglalkozó, arra sokszor hivatkozó – jogászok számára készülő könyvben külön is kitérni. Két kérdést fogunk tárgyalni. Egyrészt azt, hogy a statisztikusok mikor hajlandóak valamit oknak tekinteni. (Látjuk majd: az eddig látott összefüggés még nem biztos, hogy oksági viszony szerintük.) Másrészt azt, hogy ez a statisztikusok körében bevett okság-fogalom mennyiben egyezik meg a jogtudományban (és a tudományfilozófiában) használatos okfogalommal – és miben tér el tőle.

Kezdjük azzal, hogy a statisztikusok akkor beszélnek okságról, ha három feltétel fennáll. (Babbie [2009] 87.)

1. a két változó között empirikus összefüggés mutatható ki (például az egyik nagyobb értéke mellett tipikusan nagyobb a másik értéke, vagy az egyik jelenlétében nő a másik megjelenési valószínűsége, stb.);
2. az egyik (az ok) időben megelőzi a másikat (az okozatot);
3. kizárható, hogy ezt az összefüggést valamilyen harmadik – mindkét változóra egyformán ható – tényezőre (az ún. közös okra) lehessen visszavezetni.

Az első feltétellel foglalkoztunk az előző pontban.

Az időbeliség, a megelőzés kapcsán a legfontosabb probléma az, hogy sokszor nem egyértelmű, hogy melyik dolog volt előbb. A közmondásos tyúk-tojás probléma a társadalmi viszonyok között is gyakran felbukkan.

> Tegyük fel például, hogy azt tapasztaljuk, hogy egy városon belül az adott városrészben lakó romák aránya és az adott városrész ingatlanárai között erős összefüggést tudunk kimutatni. De itt nem egyértelmű, hogy melyik volt előbb. Ugyanis, ha csökken az ingatlanár, akkor nőni szokott a romák (illetve általában a szegények) aránya – a szegényebb csoportok költöznek oda. De, ha nő a romák (illetve általában a szegények) részaránya egy adott városrészben, akkor csökken az ingatlanár is.

A harmadik pontot szokás az ún. *összemosó tényező*, vagy *közös ok* problémájának is nevezni. Ez azt jelenti, hogy a két dolog nincs egymással érdemi kapcsolatban – csak valamiféle „egybeeséssel" van dolgunk. Ezzel szemben mind a kettő összefügg egy harmadikkal: az „okozza" mind a kettőt.

> A probléma jól érthető Jon Elster – Alexis de Tocqueville-től kölcsönzött – példáján. (Elster [1995] 11–12) A kérdés: igaz-e, hogy a szerelemre alapuló (és nem a szülők által elrendezett) házasságok hosszabb távon boldogtalanabbak? Egyelőre tegyük fel, hogy az összefüggést ki is lehet mutatni: a szerelmi házasságok rövidebb ideig tartanak, az abban élő felek boldogtalanabbak, stb. Ez azonban mégsem jelenti azt, hogy az ilyen házasságok boldogtalanságát az okozza, hogy (korábban) szerelemből kötötték azokat. Mindenekelőtt azért nem, mert meghúzódhat a háttérben egy összemosó tényező. Például ha az adott kultúrában az elrendezett házasság a bevett, akkor a

szerelmi házasság egyfajta lázadást jelent a bevett normák ellen. Lázadók kötnek szerelmi házasságot – „öntörvényű emberek". És két öntörvényű ember nehezebben alkalmazkodik egymáshoz is. (További példákat láthatunk az összemosó tényezőre a 14.4. szövegdobozban.)

Az egyik leggyakoribb összemosó tényező az idő. Vannak olyan jelenségek, amelyeknek ún. *trendjük* van: idővel valami tipikusan nő, vagy csökken. Például az elmúlt időszakban ilyen – a legtöbbek által elfogadott – trend az átlaghőmérséklet emelkedése. De ugyanígy kimutatható bizonyos valuták leértékelődése is. Ha idősort vizsgálunk, akkor azt fogjuk találni, hogy e két jelenség (a hőmérséklet és az árfolyam alakulása) között összefüggés van. Holott csak „összemossa őket" az idő. (Éppen ezért jeleztük az 1. alfejezetben, hogy az idősorok elemzésekor nagyon kell figyelni. Többek között az összefüggések vizsgálatakor, a következtetések levonásakor is.)

Ez a harmadik feltétel úgy szólt, hogy „kizárható" ilyen harmadik (közös) ok léte. Ugyanakkor az összemosó tényező problémáját a statisztika egymaga soha nem tudja teljesen kizárni. A statisztika csak arra alkalmas, hogy ha valaki vélelmez egy ilyen közös okot, akkor teszteli annak hatását. (Ennek a tesztnek a legfontosabb módszerével, a regressziós modellel az ötödik alfejezetben találkozunk majd.)

A 14.4. táblázat a mottó B példája mögött meghúzódó adatokat bontja ki. Azt látjuk, hogy miközben a nők felvételi eredményei rosszabbak (30% szemben a 49%-kal), aközben egyetlen szak sincs, ahol jelentősen rosszabb eredményeket érnének el. Sőt, a szakok között csak egy jelentős eltérést láthatunk – és azt is a nők javára: az A szak esetén őket 82%-os arányban veszik fel, míg a férfiakat csak 62%-os arányban.

A táblázatot áttekintve előttünk áll az összemosó tényező: az ún. *összetételhatás*. A nők „összetétele" eltér a férfiakétól. Nem csak a nemük jelenti a különbséget, hanem az is, hogy az egyes szakokra más arányban jelentkeztek: a nők lényegesen nagyobb arányban jelentkeztek azokra a szakokra ahova nehezebb volt bekerülni.

	Férfiak		Nők	
Szak	Jelentkezők száma	Felvettek %-a	Jelentkezők száma	Felvettek %-a
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7
Összesen	2691	49	1835	30

14.4. táblázat: A férfiak és a nők felvételi eredményei szakonként
(Forrás: Freedman et al [2005] 36)

A legtöbb statisztikus oksággal kapcsolatos álláspontja ezért az, hogy az okság fogalma a statisztikán kívüli fogalom. Az okság az összefüggést magyarázó megalapozott (megalapozottnak tűnő) szakmai hipotézis – a statisztikai csak ennek az okság által vélelmezett összefüggésnek a fennálltát, erősségét tudja tesztelni.

14.4. szövegdoboz: Példák összemosó tényezőkre

Tanulságos példákat hoznak az összemosó tényezőre Freedman és szerzőtársai. (Freedman et al [2005]. 179)

Az egyik példájukban a rákot okozó tényezőkkel kapcsolatos vizsgálatok problémáira hívják fel a figyelmet. Megfigyelhető például, hogy azokban az országokban, ahol sok zsírt fogyasztanak, magas bizonyos rákos megbetegedések aránya. Ebből az adatból azonban (még) két ok miatt sem vonhatjuk le a következtetést, hogy a zsírfogyasztás a rák esélyének növekedését okozza. Az ökológiai tévedést már tárgyaltuk: országos szintű adatokból nem következtethetünk egyéni, egyes emberek szintjén igaz összefüggésekre. (Tegyük hozzá: napjainkban a rák esélyét növelő tényezőket vizsgáló kutatások nem országok, hanem egyének adataival dolgoznak.) A másik a mostani témánk: az összemosó tényezők hatása. Tudjuk például, hogy azon országokban, ahol sok zsírt fogyasztanak, például a cukorfogyasztás is magas. Illetve az is igaz, hogy azokban az országokban, ahol több zsírt fogyasztanak, a jövedelem is magasabb. Vagyis a rák esélye és a zsírfogyasztás közötti feltárt összefüggés mögött meghúzódhat más táplálkozásban jelentkező, vagy egyéb (a magas jövedelemmel összefüggő) életmódbeli hatás is.

Másik példájuk különösen azért tanulságos, mert egy gyakran alkalmazott magyarázó változó a képzettséget mérő változó tulajdonságára hívja fel a figyelmet. A példájuk az, hogy a nagy gazdasági válság idején (1929-33) azt figyelték meg, hogy az iskolázottabbak tipikusan rövidebb ideig maradnak munka nélkül. Ebből első ránézésre levonható a következtetés: az alacsony iskolázottság miatt nő a munkanélküliség hossza – az alacsony iskolázottság okozza azt. Azonban nem feledkezhetünk el arról, hogy abban az időben is igaz volt (ahogy azóta is igaz), hogy a fiatalabb generációk iskolázottabbak, mint a korábbiak – folyamatosan nő az iskoláztatás. Vagyis könnyen lehet, hogy nem a magasabb iskolázottság miatt (annak okán) csökkent a munkanélküliség hossza, hanem csak annyi történt, hogy a fiatalabbak könnyebben találtak munkát, mint az idősebbek.

A problémát az okozza, hogy az iskoláztatás és az életkor, vagyis az, hogy ki milyen generációhoz tartozik, között erős az összefüggés. (Ugyanígy az iskolázottság és a jövedelem között is – ahogyan erre más vizsgálatoknál hívták fel a figyelmet. Az iskoláztatásnak tulajdonított hatás lehet, hogy csak azt jelenti, hogy a magasabb jövedelemmel rendelkezők inkább tesznek valamit.)

14.3.5. Az okság egyéb fogalmai

A természet- és társadalomtudományban többféle, egymással vitatkozó oksági modellt ismerünk. Ezeket három nagyobb csoportba sorolhatjuk: 1. a regularitásra épülő, 2. a tényellentétes és 3. a valószínűségi oksági tesztekre. Lássuk ezeket – és a közöttük levő eltéréseket.

A *regularitásra épülő okfogalom* lényege, hogy az egyik dolgot (az okot) a másik (az okozat) *szokta* követni.

> Ezen okfogalom klasszikus megfogalmazását David Hume adta: az okot olyan dologként határozhatjuk meg, amelyet egy másik követ, „éspedig olyképpen, hogy az elsőhöz hasonló összes dolgot a másodikhoz hasonló dolgok követik" (1751; Hume, 1973, 117). Ez az állandó együttjárásra („összes dolgot") épülő okfogalom a későbbiek során finomodott: kialakult az ún. regularitásteszt mai formája. Eszerint az okot az okozat nem mindig követi, hanem csak követni szokta.

Ez a teszt erősen támaszkodik a hasonlóság fogalmára. A modell igazából azt keresi, hogy az esetéhez *hasonló körülmények* az esetben fellépő következményhez *hasonló eseményekhez* szoktak-e vezetni. Ezzel azonban egy újabb kérdés bukkan fel: választ kell adni arra, hogy mi a hasonló, mi használható analógiaként.

> A United Novelty v Daniels (43 So.2d 395, 1949) esetben – többek között – ez volt a kérdés. Az ügyben azt kellett eldönteni, hogy okozója-e a munkavállaló balesetének az a munkaadó, aki azzal a feladattal küldte őt egy olyan szobába, ahol tűz nyílt lánggal égett, hogy benzinnel tisztítson meg valamit. A szobában – mint kiderült – volt egy patkány is. A benzin ráfröccsent a patkányra, aki ettől megijedt, elszaladt. A tűz közelébe érve pedig a rajta levő benzin berobbant. Mi az ehhez hasonló helyzet? Az, ha valaki benzint használ egy szobában, ahol nyílt láng van – függetlenül attól, hogy mekkora a szoba? Az hasonló helyzet, ha valaki egy ilyen szobában pontosan olyan messze a tűztől használja a benzint – függetlenül attól hogy van-e patkány? Csak az a hasonló helyzet, ha egy olyan szobában történik mindez, ahol valamilyen kisállat van – nem feltétlenül patkány? Csak akkor hasonló a helyzet, ha az állat patkány? Csak miután erre e kérdésre választ adtunk, tehetjük fel a regularitás-teszt alapkérdését: szokásos következmény a robbanás ebben a helyzetben.

A *tényellentétes okság, az ún. condotio sine qua non feltétel* leggyakoribb megfogalmazása: az egyik dolog akkor oka a másiknak, ha annak hiányában az utóbbi nem jelentkezne. (Vagy megfordítva: ha az okozat megjelenik, akkor nem képzelhető el, hogy az ok nem volt jelen korábban.) Ezt a definíciót általában az ún. *kemény szükségszerűség tesztjének* nevezik. (Honoré [1995] 363) De a gyakorlatban általában nem ezt alkalmazzuk (bár erre szoktunk hivatkozni), hanem a puhább teszteket. Ezek már a *ceteris paribus feltevésből* (az adott feltételek változatlanságából) indulnak ki, és azt kérdezik, hogy ha minden más változatlan lenne, akkor adott dolog (az ok) megjelenéséből következik-e az adott esemény (az okozat). A puhább tesztek nem várják el, hogy mindig (minden körülmények között) igazolható legyen az ok és az okozat közötti kapcsolat – csak adott körülmények között..

> A kemény szükségesség tesztjének megfogalmazása szintén David Hume-hoz köthető: „ha [az ok] nem lett volna, [az okozat] sose létezhetett volna" (Hume [1751/1973] 117.) Hume ezt a regularitás-teszt kiegészítéseként fogalmazza meg, de a modern irodalomban már inkább alternatív okságfogalomként kezeli (Huoranszki [2001] 110).

A puhább tesztek közül érdemes kiemelni az ún. INUS-feltételt [(insufficient but necessary part of a condition which itself unnecessary but sufficient for the result]. (Mackie [1965, 1974]). Ez abból indul ki, hogy egy adott eseményt (az okozatot) többféle körülmény is előidézhette *volna*. A konkrét feltételrendszerből viszont azokat a nem redundáns elemeket tekintjük oknak, amelyek nélkülözhetetlenek az adott esemény előállásához.

> Például egy betörés nagyon sok módon megeshetne. Ahogy a valóságban megesett az annak egy elégséges feltételrendszerét adja. (Elégséges feltétel, mert, mivel bekövetkezett, ezért a betörés is bekövetkezett.) Ezen körülmények között azonban van olyan, ami elhagyható: ha nem az történt volna, akkor is van betörés. (Például ilyen az, hogy esett az eső.) Mások viszont ezen belül szükséges feltételek voltak: ha nem az történik, nincs betörés. (Például: ha az üveg, amit kivágtak a betörők, erősebb, „vághatatlan" üveg lett volna.)

A tényellentétes okság kapcsán a két fő (leggyakrabban tárgyalt) probléma annak *hipotetikussága* és a túlderminált ság.

A teszt a valós, ismert folyamatot egy hipotetikussal veti össze – azzal, hogy mi lett volna az ok hiányában. Tudni kell, hogy mi lépne az ok helyébe a hipotetikus helyzetben. És azt is hogy

az esemény bekövetkezne-e (az okozat megjelenne-e) ebben az alternatív helyzetben. És tipikusan mind a kettő meglehetősen bizonytalan.

> A probléma jól megérhető, ha a halálbüntetés hatásáról szóló vitára gondolunk. Ahhoz, hogy kimondhassuk, hogy a halálbüntetés okozza-e a bűncselekmények (bizonyos bűncselekmények) számának csökkenését nem elég azt tudni, hogy halálbüntetés esetén hány bűncselekményt követnek el. Azt is tudni kellene, hogy...
>
> 1. ...mi lenne az adott bűncselekmény büntetése halálbüntetés hiányában?
> 2. ...mekkora lenne ezen bűncselekmény száma ezen alternatív büntetés mellett?

Az ún. *túldetermináltság* akkor jelentkezik, ha adott eseményt nagyon sok dolog idézhette volna elő. Ekkor egyik sem megy át a tényellentétes bizonyítás resztjén: egyik sem lesz ok.

> Például a Caesart érő több késszúrás közül nehéz akár egyetlent is okként azonosítani (Huoranszki [2001] 115).

A harmadik oksági modell a *valószínűségi okság*. Ennek megértéséhez érdemes egy példával folytatni. A dohányzás és a dohányzás miatti megbetegedések közötti kapcsolat sem a tényellentétes elemzés, sem a regularitás-teszt alapján nem igazolható. A tényellentétes teszt alapján nem állítható az okság: nem tudható, hogy adott esetben (adott ember esetén) mi történt volna dohányzás hiányában. A regularitás tesztje alapján pedig azért nem, mert nem biztos, hogy a dohányzás után a betegég is fellép. Sőt, még az sem biztos, hogy az esetek többségében fellép. (Például, szerencsére a dohányzás következtében is csak az esetek kisebb részében alakul ki tüdőrák.) Amikor a dohányzást mint okot azonosítjuk, akkor általában csak azt állítjuk, hogy a dohányzás növeli adott betegség esélyét.

Pontosabban a valószínűségi okságot is (legalább) két formában fogalmazhatjuk meg. Az egyik (a gyengébb változata) szerint az ok megjelenése emeli az okozat bekövetkezési valószínűségét. (Hitchcock [2018]) A másik (kicsit erősebb) változata szerint: az ok jelenlétében az okozat bekövetkezési valószínűsége *jelentősen, nem elhanyagolható mértékben* nő. Érdemes ezt a tesztet összeveti a regularitás-teszttel! Utóbbi azt tekinti okozatnak, ami valószínűleg követi az okot – függetlenül attól, hogy az okozat hiányában mekkora lenne annak az esélye. A valószínűségi teszt pedig azt, aminek a megjelenési esélye emelkedik – függetlenül attól, hogy az ok jelenlétében mekkora ennek az esélye.

> Másik példán: a valószínűségi teszt alapján például a gyorshajtás az oka a balesetnek, ha növeli annak esélyét – a regularitásteszt szerint nem, ha ez a megnövelt baleseti esély is kisebb, mint amit rendszeres együttjárásnak tekinthetünk.

A valószínűségi okság kapcsán a probléma az, hogy pusztán a valószínűségemelkedés alapján nem tudjuk megmondani, hogy mi az ok és mi az okozat. Ugyanis (matematikai okokból): amennyiben E megjelenési esélye nagyobb C jelenlétében, mint annak hiányában, akkor C megjelenési esélye is nagyobb E jelenlétében, mint annak hiányában.

14.5. szövegdoboz: A jog okfogalmai

A jogtudományban is gyakran megjelenik az okozás fogalma. (Beszélünk például károkozóról, halált okozó testi sértésről stb.) A joggyakorlatban, a bíróságok tipikusan kétféle kérdést tesznek fel (nevezhetjük ezt kétlépcsős tesztnek). Megkülönböztetünk

– ún. *okazonosító* elméleteket, amely az okok szélesebb körét, az ún. *természetes okokat* jelölik ki, és

– ún. *okszűrő* elméleteket, amelyek e természetes okok közül választják ki az ún. *jogi (jogilag releváns) okokat*.

A mi mostani kérdésünk – vagyis a statisztika, a más tudományágak okfogalma kapcsán – a fontosabb az első, az okazonosítás. (Ez keresi a természetes okot.) A legismertebb okazonosító elmélet (sőt a büntetőjogi tankönyvekben megjelenő egyetlen oksági modell) a *conditio sine qua non* elve, amely az okozat szükséges feltételeit tekinti oknak.

Az egyik legfontosabb vita a jog okfogalma kapcsán: a *conditio sine qua non* feltétel hiányában (vagy annak bizonyíthatatlansága esetén) más elvek segítségével is találhat-e a jog okokat. Például, segítségül hívható-e a valószínűségi vagy a regularitás teszt? Mint látjuk a dohányzással kapcsolatos esetekben nem lennénk képesek pusztán ennek alapján okokat azonosítani – de sok orvosi műhiba esetén sem. [7]

Érdemes a három okfogalom lezárásaként egy közös problémájukra felhívni a figyelmet. Mindegyik esetén felmerül a kérdés, hogy annak alapján valóban csak az okokat találjuk-e meg. Az irodalomban általában különbséget teszünk *okok és feltételek* között. Például azon lehet vitatkozni, hogy az áldozat figyelmetlenség (vagy éppen kihívó viselkedése) szükséges, vagy elégséges *feltétele* volt-e a vele szembeni bűncselekménynek – de bármiként is vélekedünk erről, az áldozatot *okozónak* soha nem tartjuk. Az okok „különleges feltételek". A fenti tesztek viszont – sokak szerint – csak a feltételek megtalálására alkalmasak, önmagukban ezek segítségével nem tudjuk kiválasztani az okokat. Ez a probléma (egyes feltételek „kiemelése") a jog oksági modelljének is kulcskérdése. (Ez a 14.5. szövegdoboz tárgya.)

14.4. Statisztikai következtetések

A statisztika tudományának másik – talán még az adatsűrítésnél is érdekesebb – kérdése a következtetés: egy mintából vonunk le következtetéseket egy egész populációra. Egy mintából következtetünk a populáció azon tagjainak tulajdonságaira, akiket nem ismerünk, akikről nincs is adatunk. Ez, pontosabban ennek két fő alkérdése, a becslés és a hipotézistesztelés (más néven: bizonyítás) lesz ennek az alfejezetnek a tárgya.

Mielőtt azonban hozzáfognánk, be kell vezetnünk két fogalmat. A vizsgálat során statisztikákból következtetünk paraméterekre. *Statisztikák*nak nevezzük azokat a mutatókat, amelyeket a mintából megismerhetünk. Statisztika például a minta átlaga, szórása, valamilyen jellemző előfordulási gyakorisága, az egyes változók közötti összefüggés a mintában. A *paraméterek* ugyanezen mutatók – de a teljes sokaságban. Vagyis például a teljes sokaság (nem megismerhető) átlaga, szórása, valamilyen jellemző gyakorisága, vagy két változó közötti összefüggése a teljes sokaságban.

14.4.1. Becslés

Az alapvető problémát az jelenti, hogy az elemző csak a mintát, illetve az azt leíró statisztikákat ismeri – a paramétereket csak szeretné. Azokat becsülni próbálja. És a statisztikai

[7] Azonban ennek kacsán is több kérdés felmerül. Például: lehet-e ok (akár részben) nem emberi magatartás – figyelembe veszi-e ezt a jog? Milyen emberi magatartás lehet ok – például bármilyen emberi magatartás, vagy csak jogellenes, felróható magatartás? Mi lehet okozat – lehet-e, hogy az okozat nem valaminek a megjelenése, hanem csak az esélyének az emelkedése? (Blutman [2011] 314; Boronkay [2007] 191; Dósa [2010] 113; Fuglinszky [2015] 244., Menyhárd [2015] 299; Szalai [2017] 22–28, 39-41.)

következtetéselmélet (és a mögötte meghúzódó valószínűségszámítási modell) bizonyítja, hogy ezt meg is teheti – bizonyos minták esetén.

Egészen pontosan: amikor a statisztikus becslést ad egy paraméterre (vagyis például arra, hogy a teljes sokaságban hány százalék valamilyen jellemzővel bíró egyedek aránya, vagy mekkora a teljes sokaság átlagértéke), akkor két dolgot mond meg. Két becslést ad. Létezik pontbecslés és intervallumbecslés. A pontbecslés a paraméter értékének legvalószínűbb értékét adja meg; az intervallumbecslés pedig egy olyan tól-ig-határt ad, amely „meghatározott megbízhatósággal" tartalmazza a paramétert.

> A logika legegyszerűbben talán akkor érthető, ha a választási közvéleménykutatások példáját vesszük. Ezek meg szokták adni, hogy
> 1. mekkora a különböző pártokra szavazók várható aránya (az eddigi nyelvhasználattal: mekkora azon egyedek aránya a populációban, akiknek az a tulajdonságuk, hogy az adott pártra szavaznak)
> 2. mekkora a „hibahatár". (Például ± x százalékpont.)
>
> Az előbbi a pont-, az utóbbi az intervallumbecslés.

A *pontbecslés* nem túl bonyolult: torzítatlan egyszerű véletlen minta esetén a mintában látott érték (a statisztika) várhatóan megegyezik a populáció adott értékével (a paraméterrel).

De csak „*várhatóan*" egyezik meg. Egészen pontosan, ha sok véletlen mintát veszünk, akkor az azokban megjelenő statisztika várható értéke (átlaga) egyezik meg a paraméterrel. Adott minta szinte biztos, hogy nem ezt az értéket adja. Épp a minta kiválasztásakor szerepet kapó véletlen miatt nem várhatjuk, hogy a mintabeli statisztika „telibe trafálja" a valós paramétert. Előfordulhat ugyanis, hogy véletlenül „túl sok" kis vagy nagy érték kerül a kiválasztott mintába.

Érdemes abból az egyszerű összefüggésből kiindulni, hogy a mintában „talált" jellemző, statisztika alapvetően három tényező összegétől függ:

(i) a paramtér értékétől (vagyis a populáció „valós" jellemzőjétől)
(ii) a minta torzítása miatti eltéréstől, és
(iii) a véletlen hibától.

Induljunk hátulról! A véletlen hiba ezen várható nagyságát nevezik *standard hibá*nak. Ennek nagysága becsülhető statisztikai módszerekkel. Ez lesz az intervallumbecslés lényege.

Ez függ a minta nagyságától: csökken annak növelésével – de nem arányosan. Ahhoz, hogy a hibát felére szorítsuk le, négyszer akkora minta kell – ha harmadára akarjuk leszorítani, akkor kilencszer akkora.

> Egészen pontosan, ha k-szorosára növeljük a mintát, akkor a pontosság $\sqrt{k}$–szorosára nő (a standard hiba $\sqrt{k}$-adrészére csökken). Ebből következik, hogy ha négyszeresére növeljük a minta nagyságát, akkor felére csökken a százalékarány véletlen hibájának valószínű nagysága.

A pontosság és a mintanagyság tehát összefügg. De a becslés pontossága, a standard hiba nagysága attól nem függ, hogy a teljes sokaságnak mekkora részét tartalmazza a minta. Nem igaz, hogy ugyanahhoz a pontossághoz nagyobb populáció esetén nagyobb minta kell: ugyanakkora mintából ugyanolyan pontos becslést lehet adni akkor is, ha tizedakkora, és akkor is, ha tízszer akkora a populáció.

> 1.000 ember megkérdezésével Budapesten ugyanolyan pontos előrejelzést lehet adni az önkormányzati választás végeredményére, mint 1.000 fős országos mintából az országgyűlési választásokra. Ennek megértéséhez érdemes ismét Freedman és szerzőtársai egyik példáját segítségül hívni. „Képzeljük el, hogy vegyelemzéshez egy csepp mintát veszünk egy folyadékból.

Ha a folyadék jól el van keveredve, akkor a csepp kémiai összetétele tükrözi az egész üveg összetételét, és igazán nem számít, hogy egy kis üvegcséből vagy egy nagy kancsóból vettük a mintát. A vegyész mit sem törődik azzal, hogy a csepp az oldatnak 1%-a vagy 0,01%-a." (Freedman et al [2005] 414.)

Míg a véletlen hiba becsülhető statisztikai módszerekkel – és a mintanagyság növelésével csökkenthető is – a *minta torzítása* statisztikai eszközökkel nem értékelhető. A minta kapcsán ezért a fő kérdés a kiválasztás módja. Ha ugyanis a kiválasztás egyszerű véletlennel (vagy ahhoz közel álló módon) történt, akkor a következőkben bemutatott módszerek alkalmazhatók.

Pontosabban: a most bemutatott becslések teljesen pontosak akkor, ha visszatevéses húzásokkal választjuk ki a mintát. De jó közelítésnek tekinthető visszatevés nélküli húzások esetén is.

Igen kicsi tehát az esélye, hogy egy torzítatlan mintából kiszámolt, abban megfigyelt érték (statisztika) megegyezzen a valós (a populációban meglevő) paraméterrel. Éppen ezért ad a statisztikai intervallumbecslést is: meghatározza az ún. *konfidencia (megbízhatósági) intervallumot*. Egészen pontosan két adatot közöl:

- egy felső és alsó értéket és
- egy *megbízhatósági (konfidencia) szintet*.

Ha például azt mondjuk, hogy a 95%-os megbízhatósági szinttel számolt konfidencia intervallum alsó határa 900 a felső pedig 1005, akkor 95%-ig biztosak lehetünk abban, hogy ezen két határ között lesz a populáció értéke, a paraméter.

Pontosabban ez csak az első – szemléletes, de meglehetősen pontatlan – közelítése a konfidencia-intervallum jelentésének. Ugyanis a paraméter a valóságban adott: a véletlen nem annak nagyságát befolyásolja, hanem azt, hogy mi milyen mintát kaptunk – és ezért abban mekkora az adott érték. Az igazság az, hogy semmiféle módon nem tudjuk azt megmondani, hogy mekkora az esélye annak, hogy a valós érték a két megadott határ közé esik. Amit konfidencia-intervallum valójában leír az az, hogy ha száz vizsgálatot végeznénk el (mindegyiknél véletlen mintát választva a sokaságból), akkor az a módszer, amivel a konfidencia-intervallumot most számítjuk, a 100 vizsgálatból 95-ször tartalmazná a becsülni próbált paramétert.

A konfidenciaintervallum számítási mechanizmusa egyszerű. A mintában mért értékből (a statisztikából) kivonjuk, illetve ahhoz hozzáadjuk a standard hiba megfelelő számú többszörösét. Ehhez „már csak" azt kell tudni, hogy (i) mekkora a standard hiba, és (ii) mekkora a „megfelelő számú többszörös".

- Egyszerű véletlen minta esetén a minta szórását használhatjuk a standard hiba becslésére. Nagy minta esetén ez jó becslést ad.

Az átlag és az arány (a két talán legtöbbet becsült mutató) standard hibája becsülhető egy egyszerű képlettel:

$$\text{standard hiba} = \text{a minta szórása} \times \sqrt{n}/\text{n},$$

ahol n a minta elemszáma.

A standard hiba nagyságát megfigyelni nem, csak becsülni lehet. Például akkor, ha néhányszor megismételnénk a vizsgálatot – új és új mintákon, mindig ugyanúgy. Ekkor az egymást követő vizsgálatok persze nem ugyanazt az eredményt adnák – épp a véletlen, a véletlen hiba miatt. De ha megnéznénk a vizsgálatok eredményeit, akkor azok szórása becslést ad arra, hogy körülbelül mekkora lesz a mérési hiba egyetlen mérésben.

- A „megfelelő szám" (a legtöbbször) a normálgörbéről olvasható le. Tudjuk például, hogy normálgörbe esetében az adatok 95%-a -2 és +2 között van. Vagyis a 95%-os

megbízhatósághoz a standard hiba kétszeresét kell a megfigyelt értékhez hozzáadni, illetve abból kivonni. És azt is tudjuk (a statisztikus, a valószínűségszámítással foglalkozó matematikus fejből is), hogy az adatok 99,7%-a pedig -3 és +3 között van. Vagyis a 99,7%-os megbízhatósághoz a standard hiba háromszorosát kell a megfigyelt értékhez hozzáadni, illetve abból kivonni. De bármilyen megbízhatósági szinthez megadható az a negatív és a pozitív érték, amelyek között az adatok éppen adott aránya szerepel.

A konfidencia-intervallum képlete talán még egyszerűbb is, mint az eddig leírás. Ha a teljes sokaság átlagát akarjuk becsülni, akkor a képlet:

$$\mathrm{x} \pm \Delta, \text{ ahol } \Delta = z_p \times \sigma/\sqrt{n}$$

Itt σ a szórás, míg z_p a „megfelelő szám". Ez a z_p a p megbízhatósági szintből kiszámolt szorzó, ami a normálgörbe egyenletéből jön. (De soha nem számoljuk ki, hanem vagy táblázatokból keressük ki, vagy számítógépes programok számolják. Megadjuk a p-t, vagyis az elvárt megbízhatóságot, és ehhez kapjuk az adott z-értéket.)

A képlet igazán azért érdekes most a számunkra, mert megjelenik benne az „$/\sqrt{n}$" tényező. Látszik, hogy ez a konfidencia-intervallum szélességét, vagyis Δ-t befolyásolja. Az csökken az elemszám (n) emelésével – de csak az elemszám négyzetgyökével arányosan.

14.4.2. Hipotézistesztelés, bizonyítás

Talán a hipotézistesztelés kapcsán érthető meg a leginkább a statisztika alapvető funkciója. A statisztika ugyan adatokat elemez, de ezt nem öncélúan teszi. A cél az, hogy egy statisztikán kívüli – szerencsés esetben a más tudományok által jól megalapozott, plauzibilis – állítást teszteljen az elé kerülő adatokon. Tipikusan egy mintán.

A becslés arra keresi a választ, hogy „mennyi?": valamilyen sokasági értéket (paramétert) mennyire becslünk a minta alapján? De feltehető egy kérdés úgy is, hogy a minta alapján mit gondolunk: igaz-e egy állítás a sokaság kapcsán. Ezt az állítást, hipotézist akarjuk tesztelni – más szóval bizonyítani (vagy cáfolni). Azonban mind a „hipotézis", mind a „tesztelés, bizonyítás" fogalmát értenünk kell. Különbséget kell tennünk szakmai, null- és ellenhipotézis között. A bizonyítás kapcsán pedig (az egyébként a jogi bizonyításnál is használt) indrekt bizonyítás fogalmát (és jelentését), valamint az bizonyítékok „erejét" (ami a jogban az „ítéleti bizonyosság", a statisztikában a „szignifikancia" fogalmában ölt testet) kell megérteni.

A tesztelendő állítást fogjuk most „*szakmai hipotézisnek*" nevezni. Első lépésként ezt kell lefordítani az adatok nyelvére. Ez a legtöbbször kézenfekvő: A mottóban szereplő A példánál például az, hogy a feketék átlagkeresete *kisebb*, a B példánál pedig az, hogy a nők felvételi aránya *alacsonyabb*, mint a másik csoportban.

Tegyük fel, hogy az adataink (egy mintából) származó adataink ezt is mutatják. (A feketék jövedelme *alacsonyabb*; a nők felvételi aránya *alacsonyabb*, stb.) A probléma az, hogy ez még nem igazolja a szakmai hipotézist. Egy ilyen (bizonyítéknak tűnő) eredmény ugyanis három ok miatt is előállhat. Lehet, hogy az adott szakmai hipotézis, magyarázat tényleg igaz. De lehet, hogy az eredményt azért kaptuk, mert a minta torz; vagy egyéb érvényességi hiba jelentkezett. És az is lehet, hogy – bár a minta nem torz – az eredmény a mintavételi ingadozásból fakadó véletlen miatt állt elő. Amiatt, mert véletlenül éppen azok kerültek a mintába, akik. A statisztikai hipotézistesztelés ez utóbbi lehetőséget igyekszik kizárni. (A minta torzításáról láttuk, hogy a statisztika nem tudja tesztelni.)

A hipotézistesztelés az *indirekt bizonyításon* nyugszik. Vagyis feltesszük, hogy a (szakmai) hipotézis nem igaz. Első lépésként, meghatározzuk, hogy milyen lenne akkor a mutató az adott mintában. (A válasz megint nem bonyolult: a feketék ugyanannyit keresnének, mint a fehérek; a nők felvételi aránya ugyanannyi, mint a férfiaké.) A mintából kapott eredményt pedig összevetjük ezzel a feltételezett eredménnyel. Eltér tőle... Persze... De kérdés, hogy okozhatja-e ezt egyszerűen a mintavételben rejlő véletlen. Igen... Persze... De minél nagyobb az eltérés, annál kisebb az esélye ennek. Ennek az esélynek a megadása a hipotézistesztelés lényege. Ezt az esélyt nevezi a statisztika *szignifikanciának* – ha ennek értéke alacsony, akkor mondjuk, hogy a magyarázatot, a szakmai hipotézist alátámasztó eredmény szignifikáns.

> A köznyelvben a szignifikáns hatás általában jelentős hatást, jelentős eltérést jelent. A statisztikában kicsit mást értünk alatta. De a jelentős hatások valóban általában statisztikai értelemben is szignifikánsak.

A hipotézistesztelés három lépésből áll:

1. Meg kell fogalmazni (le kell fordítani a statisztika, az adatok nyelvére) a nullhipotézist.
2. Ki kell választanunk egy „próbastatisztikát", „tesztstatisztikát" – ez méri, mennyire térnek el az eredmények a nullhipotézis alapján várttól.
3. A tesztstatisztika alapján ki kell számítanunk az ún. empirikus szignifikanciaszintet, az ún. p-értéket. Ez méri annak *esélyét*, hogy az eredményt csak a mintavételben rejlő véletlen ingadozás okozza.

(ad 1) A *nullhipotézis* fogalmát az imént vezettük be. Ez arra feltevésre alapul, hogy a mintában – a szakmai hipotézist alátámasztó – eredményt pusztán a véletlen okozza. Ahhoz, hogy a szakmai hipotézist bizonyítani tudjuk „ki kell zárni" ennek esélyét. Ezt a nullhipotézist kell cáfolni.

> Tegyük hozzá: a büntetőeljárásban ismert bizonyítás is ilyen indirekt bizonyítás. Ott a szakmai hipotézis az, hogy a vádlott bűnös. A nullhipotézis (a vélelem) azonban az, hogy ártatlan. Ezt próbálja cáfolni, „megdönteni" az ügyész.

A statisztikai bizonyításhoz, a hipotézistesztéléshez, a nullhipotézissel szemben meg kell fogalmazni egy *ellenhipotézist* is. A statisztikai elemzésben ez lesz a szakmai hipotézis. Az ellenhipotézis tipikusan három formát ölthet: a paraméter

(i) kisebb, mint amit a nullhipotézis állít.
(ii) nagyobb, mint amit a nullhipotézis állít.
(iii) nem egyenlő azzal, amit a nullhipotézis állít (vagyis kisebb és nagyobb is lehet).

> Látszólag az utóbbinak nincs értelme. Ha azt állítjuk, hogy a feketék kevesebbet keresnek, vagy a nőket kisebb arányban veszik fel, akkor ez az állítás pontosabb, mint az, hogy nem annyit keresnek, vagy, hogy nem olyan arányban veszik fel őket. Van azonban olyan eset, amikor az állítás csak az, hogy a két dolog eltér. Ilyen állítást fogalmazhatunk meg például a 14.2. táblázat kapcsán: a nők és a férfiak *nem ugyanúgy* teljesítenek a vizsgán.

(ad 2) A statisztikus számára a központi kérdés a megfelelő *próba-* vagy *tesztstatisztika* kiválasztása. Ezzel mérjük, hogy a kapott eredmények mennyivel térnek el a nullhipotézis alapján várható értéktől. Egész pontosan: hány standard hibányira vannak attól. Alapvető logikája:

$$tesztstatisztika\ értéke = \frac{megfigyelt\ érték - nullhipotézisben\ megfogalmazott\ érték}{standard\ hiba}$$

Általában elmondható, hogy minél inkább eltér ez a tesztstatisztika nullától (akár negatív, akár pozitív irányba), annál kisebb az esélye annak, hogy a nullhipotézis igaz legyen, vagyis, hogy a megfigyelt eltérést pusztán a véletlen okozza.

Azt, hogy a tesztstatisztikát hogyan konvertáljuk p-értékre (statisztikai nyelven: milyen próbát alkalmazunk) alapvetően két szempontra figyelemmel döntjük el:

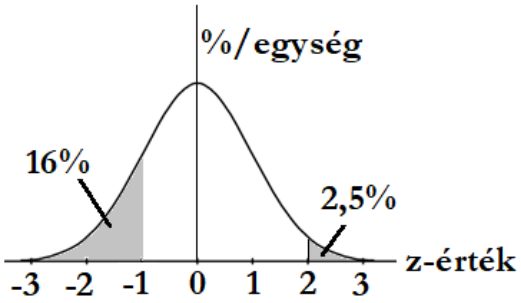
(i) mi a kérdés (mi a szakmai hipotézis) és
(ii) milyen a minta.

Például, ha (i) a szakmai kérdés és a nullhipotézis az átlagra, vagy valamilyen tulajdonágú egyedek gyakoriságára vonatkozik, és (ii) a minta nagy, akkor ún. *z-próbát* végezhetünk. Vagyis ekkor a tesztstatisztika az ún. *z*-értéket adja. (A módszer mindjárt következik.) Kisebb mintán a tesztstatisztika érték az ún. t-érték lesz.

A z-értékkel a becsléskor már találkoztunk. Gyakorlatilag ugyanazt a képletet írjuk fel fordítva: $z = \frac{x-\mu}{\sigma/\sqrt{n}}$. Az ún. t-próba képlete megegyezik az előzővel: $t = \frac{x-\mu}{\sigma/\sqrt{n}}$. (Az eltérés csak annyi, hogy az így kapott értéket más módon fogjuk „valószínűséggé konvertálni".)

A tesztstatisztika (a z-érték, vagy a t-érték) önmagában semmit nem jelent, mértékegység nélküli mutató. De az értéke összevethető egy-egy (a statisztikusok és a valószínűségszámításban jártasemberek számára) jól ismert eloszlással. A z-érték például a normálgörbével. Megvizsgálható, hogy a normálgörbén mekkora annak az esélye, hogy az érték annyival (vagy annál nagyobb mértékben), tér el nullától, mint amit a tesztstatisztika mutat. Ezt az értéket nevezzük empirikus szignifikanciának, vagy p-értéknek. (Az ábrázolást látjuk a 14.10. ábrán.) A t-értékből ugyanígy nyerjük az empirikus szignifikanciát – csak nem a normális eloszlást vizsgáljuk, hanem az ún. Student-féle eloszlást.

> Az előbb az intervallumbecslésnél ezt az eljárást írtuk körül úgy, hogy, ha megadjuk a p-t, akkor a számítógép megadja a z-értéket. Itt az történik, hogy ha például a z-érték -2, akkor megnézzük, hogy az esetek hány %-a van a 0-tól ilyen távol (vagy távolabb.) A becslésnél fordítva dolgoztunk: megmondtuk, hogy mekkora lehet a hiba, és megnéztük, hogy milyen z-érték az, amelynél az eseteknek éppen ekkora része van távolabb.[8]



14.10. ábra: A normális eloszlás (normálgörbe) és a p-érték leolvasása

[8] Fontos különbség azonban, hogy a becslésnél általában „kétoldalúan" gondolkodunk: a túl nagy és a túl kicsi is hiba. A hipotézistesztelésnél ez attól függ, hogy milyen az ellenhipotézis. Ha annak állítása az, hogy „nem egyenlő", akkor itt is összeadódik az adott z-érték negatív pontja alatti és pozitív pontja fölötti esetek aránya (gyakorisága). Ha egyoldalú az ellenhipotézis, vagyis egyértelműen azt állítjuk, hogy „kisebb" vagy „nagyobb", akkor csak az adott oldali értéktől távolabb levő esetek arányát keressük.

Az így kapott *p-érték*, vagyis az *empirikus szignifikancia* fogalmának pontos értelmezése: ha a nullhipotézis igaz lenne, és a vizsgálatot nagyon sokszor elvégeznénk (véletlenül kiválasztott mintákon), akkor ilyen arányban mutatna az eredmény a most kapott (vagy annál nagyobb) eltérést. A kis p-értéket a nullhipotézis ellen szóló bizonyítékként szoktuk értelmezni. Ha a nullhipotézis igaz lenne, akkor – pusztán a véletlen miatt – nagyon kevés esetben kapnánk ilyen „szélsőséges" eredményt. Vagyis, ha a p-érték alacsony, akkor feltehetjük, hogy a véletlenen kívül valami egyébnek is hatnia kellett.

> A statisztika könyvek mindig hangsúlyozzák, hogy a p-érték nem annak valószínűségét adja meg, hogy a nullhipotézis igaz. „Csak" azt mondja meg, hogy milyen valószínűséggel kapunk ennyire erős, vagy ennél erősebb bizonyítékot a nullhipotézis ellen – ha a nullhipotézis igaz.

De ki kell emelni, hogy ennek az ellenkezőjét nem szoktuk mondani: ha a p-érték viszonylag magas az nem jelenti azt, hogy a nullhipotézis igaz. Csak azt, hogy nem találtunk (kellően) erős bizonyítékot ellene.

> A logika jól érhető a bírósági bizonyítás esetén. A büntetőjogban az, hogy valakit nem találunk bűnösnek, mert nincsenek kellően erős, „megdönthetetlen" bizonyítékok az „ártatlanság ellen", nem jelenti azt, hogy ártatlan. A jog nyelvén csak annyit mondanánk: nem bizonyított a bűnösség. A bűnösség a szakmai hipotézis – az ártatlanság a nullhipotézis (a vélelem).

Ezt az állítást jobban megérthetjük, ha bevezetjük a hibás döntések osztályozását. Érdemes az alábbi két kérdést szétválasztani. (A logika követhető a 14.5 táblázatban is.) Egyrészt egy hipotézis (így a nullhipotézis is) vagy igaz vagy hamis/téves. Másrészt a hipotézist vagy elfogadjuk vagy elvetjük. Ennek alapján négy helyzet állhat elő: (i) elfogadjuk a valóban igaz hipotézist; (ii) elvetjük a hipotézist, pedig az helyes; (iii) elfogadjuk a hipotézist, pedig az hamis; (iv) elvetjük a téves hipotézist. A négy lehetőség közül kettő helyes döntés (elfogadjuk az igazat, elvetjük a hamist) és kettő hibás. E két hibát azonban ne keverjük össze! A statisztika és a logika éles különbséget tesz köztük. *Elsőfajú hibának* nevezzük azt, ha az igaz nullhipotézist elvetjük. *Másodfajú hibának* pedig azt, ha a hamis nullhipotézist elfogadjuk.

Hipotézis	Igaz	Hamis
Elfogadása	OK	Másodfajú hiba
Elutasítása	Elsőfajú hiba	OK

14.5. táblázat: A döntési hibák

A büntetőperek példáján, ahol a nullhipotézis, ugye, az, hogy a vádlott ártatlan,

- elsőfajú hiba az, ha az ártatlant elítéljük,
- másodfajú hiba az, ha a bűnöst felmentjük.

A szignifikancia-vizsgálat az elsőfajú hibára tekint, ennek a valószínűségét próbálja számszerűsíteni: mekkora az esélye annak, hogy *ha a (null)hipotézis igaz*, akkor elvetjük azt. Az empirikus szignifikancia, a p-érték épp ezt mondja meg. Ugyanakkor azt is tudni kell, hogy a másodfajú hiba valószínűségét (pontosabban: annak valószínűségét, hogy bár hamis a nullhipotézis, elfogadjuk azt) számszerűsíteni nem tudjuk. Csak annyit mondhatunk: ha

emeljük a minta nagyságát, akkor ezzel *általában* csökkenthető a másodfajú hiba valószínűsége. (A másodfajú hiba el nem követésének valószínűségét értjük „a *teszt erejé*n".)

> Az első és a másodfajú hiba alakulása kapcsán is érdemes szem előtt tartani a büntetőperek példáját: ha csökkentjük az esőfajú hiba valószínűségét, akkor általában emeljük a másodfajúét. Ha keményebb bizonyítékokat követelünk ahhoz, hogy elítéljünk egy vádlottat (elvessük az ártatlanságát), akkor ezzel növeljük annak az esélyét, hogy bűnösöket fogunk felmenteni.

Végezetül le kell szögezni: ha egy „eredmény szignifikáns" az csak annyit jelent, hogy kicsi az esélye, hogy a hatás, a nullhipotézistől mért eltérés pusztán a véletlen műve legyen. De ez semmit nem mond arról, hogy az adott hatás mekkora – szakmailag is releváns-e. Éppen ezért komoly viták folynak a statisztikában arról, hogy a szignifikancia-teszt, a szignifikancia kiszámítása önmagában értelmes-e. A – ma már talán – többségi álláspont szerint, a szignifikanciát csak azzal együtt érdemes elemezni, hogy mekkora is maga a hatás.

> A statisztikai szignifikancia és a szakmai relevancia közötti eltérésre jelentkezik akkor, ha egy munkahelyen azt találjuk, hogy a nők és a férfiak közötti bérkülönbség szignifikáns lesz, de elenyésző – mondjuk 1-2 %-os eltérés van köztük.

14.6. szövegdoboz: Az érvénytelenség problémája

Érvénytelen egy vizsgálat, ha az eredményei nem arra adnak választ, amire kellene – vagy amit mi kiolvasni vélünk belőlük. Megkülönböztetjük a külső és a belső érvénytelenséget.

A *belső érvénytelenség* tipikusan akkor jelentkezik, ha összemosó tényező zavarja össze a következtetést. Vagyis az azonosított hatás nem csak a magyarázó változónak tudható be – hanem valamilyen más az „okkal" együtt jelentkező egyéb hatásnak. (Tegyük fel, hogy azt vizsgáljuk: egy büntetési tétel megváltozást követő időszakban változott-e egy adott bűncselekmény száma. Csakhogy nem feledkezhetünk el arról, hogy egyik évről a másikra nemcsak a büntetés mértéke módosul, hanem nagyon sok egyéb, az adott bűncselekményre ható feltétel is.) Éppen ezért törekszünk *ceteris paribus* vizsgálatra: megpróbáljuk az összes egyéb hatást kiszűrni.

A *külső* érvényesség vagy *érvénytelenség* kapcsán pedig az a kérdés, hogy az adott elemzés eredménye kiterjeszthető-e más körülmények közé. Erre láttunk példát fent a szerelmi házasságok kapcsán. Tegyük fel, hogy az adataink meggyőzőnek tűnnek (a hipotézistesztelés szerint szignifikáns is a szerelmi házasság hatása a házasságok tartósságára, boldogságára), egy olyan kultúrában, ahol az elrendezett házasság a bevett. Kérdéses azonban, hogy ugyanez az összefüggés „átvihető-e" egy olyan társadalomra is, ahol nem az a „normális".

Ugyanilyen problémát okoz az 1. alfejezetben bemutatott ökológiai tévedés is, vagyis amikor az eltérő tulajdonságú csoportok értékeiből akarunk következtetni arra, hogy az eltérő tulajdonságú emberek hogyan reagálnak.

Ugyanakkor azt is érdemes kiemelni, hogy a szinfikancia hiánya sem jelenti azt, hogy a szakmai összefüggés nem létezik. Ez a tévedés elsősorban akkor jelentkezik, ha nem megfelelő módon „fordítjuk le" a szakmai hipotézist a statisztika nyelvére – és ezért a nullhipotézis sem lesz megfelelő. Láttunk ezt a korreláció példáján: ha lineáris kapcsolatot keresünk, de nem ilyen a valós viszony (mint például a 14.1. táblázat esetén), akkor a valóságban nem igaz nullhipotézist fogjuk fenntartani. (Az ilyen típusú problémák is megjelennek az ún. érvényességi problémák között, amelyeket a 14.6. szövegdoboz tárgyalja.)

Tegyük fel, hogy az életkor és a templomba járás közötti kapcsolatot keressük. És a nullhipotézist úgy írjuk fel, hogy a templomba járók átlagos életkora nem tér el a többiekétől. Ebben az esetben nem látunk majd szignifikáns eltérést ettől. Ugyan az életkor szerint nyilvánvalóan eltérő a templomba járás, de U-függvény szerint. Az átlag azonban nem biztos, hogy eltér a templomba járók és a templomba nem járók csoportjában..

14.5. Regressziószámítás – többváltozós elemzés

A korreláció definíciójakor azt mondtuk, hogy az azt méri, hogy a pontdiagram pontjai mennyire illeszkednek egy egyenesre. Ugyanakkor azt is kiemeltük, hogy mindegy, hogy milyen ez az egyenes. (Láttuk a 14.7. ábrán, hogy akármilyen meredek is egy emelkedő egyenes, ugyanakkora a korrelációs együttható.) A regressziót lényegesen többször alkalmazza a statisztika: ez ugyanis az egyenes alakját is megadja – amellett, hogy azt az összefüggés-erősséget is számszerűsít, mint a korreláció.

A regresszió tehát „többet tud". Rádásul több további olyan problémát is kezelni lehet általa, amit korrelációval nem – vagy csak nehezen. Láttuk például, hogy a korrelációs együtthatót mindig befolyásolhatja, összezavarhatja valamilyen összemosó tényező. A regresszióval képesek lehetünk kiszűrni ezt a hatást: ha konkrét zavaró tényezővel kapcsolatos szakmai hipotézis fogalmazódik meg, akkor annak hatást tesztelni lehet. Azt is láttuk, hogy a korreláció csak akkor alkalmazható, ha mindkét változó magas mérési szintű. A regresszió azonban – némi manipuláció után – alkalmas arra is, hogy minőségi változókat építsünk be az elemzésbe.

A regressziós vizsgálat módszerét és tulajdonságait tekinti át ez az alfejezet. Először azt az esetet mutatjuk be, amire a korrelációt is felírtuk: két változó közötti összefüggést keressük. Itt látszik majd, hogy a regresszió mennyivel tud „többet", mint a korreláció. A második pont a statisztikai következtetések fejezetben megismert problémát mutatja be: amennyiben csak egy mintát ismerünk, akkor a regresszió segítségével képesek lehetünk becslést adni a teljes sokaságban meglevő összefüggésekre is. A harmadik pont mutatja be az ún. többváltozós regressziós becslést, amely már az összemosó tényezők kiszűrésére és a nem magas szintű változók bevonására is alkalmas.

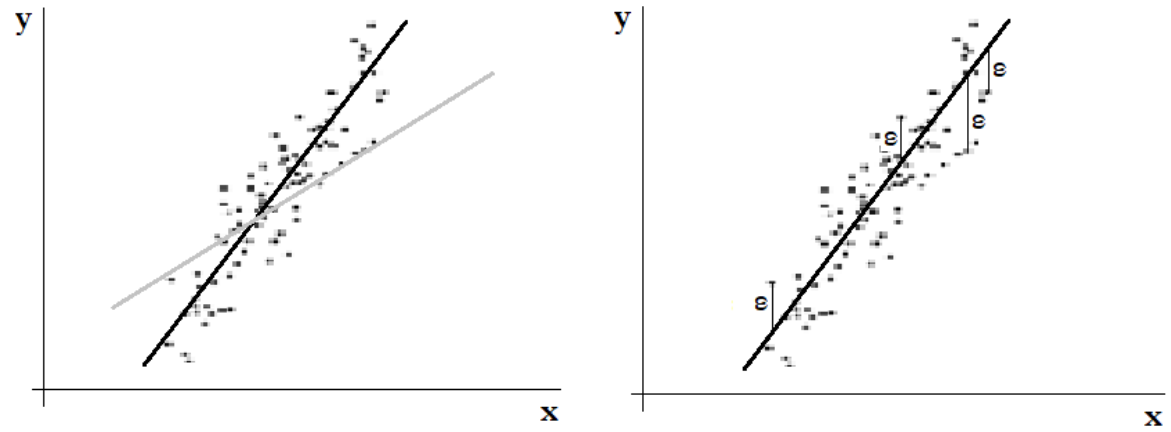
14.5.1. Kétváltozós regresszió

A korreláció azt mutatja, hogy mennyire illeszkednek egy pontdiagram pontjai egy egyenesre. Ez az egyenes az ún. *regressziós egyenes*. A regressziószámítás első lépése ennek az egyenesnek a leírása. Második lépése pedig annak számszerűsítése, hogy mennyire illeszkednek a pontok erre az egyenesre – vagyis ez ugyanazt adja meg, amit a korreláció.

A *regressziós egyenes* az az egyenes, amely a legközelebb halad a pontokhoz. Ez azt jelenti, hogy minden pont esetén megvizsgáljuk azt, hogy az egyenes milyen messze halad el a pontoktól. (Lásd a 14.2.b ábrán látható távolságokat!) Ezeket nevezzük a regresszió hibatagjának, vagy reziduumnak. Jele: ε. Képlete: $\varepsilon_i = Y_i - \widehat{Y}_i$, ahol Y_i az adott pont Y értéke, $\widehat{Y}_i$ az egyenesen az ugyanazon X-hez tartozó Y érték.

A 14.11.a ábrán látható két egyenes közül a vastagabb nyilvánvalóan közelebb halad a pontokhoz, jobban leírja a pontdiagramot, mint a vékonyabb. A grafikus megjelenítés ebben az esetben egyértelmű, más esetekben már nehezebb megmondani, hogy két egyenes közül melyik „halad közelebb" a pontokhoz. Ebben az esetben a statisztikusok (legtöbbször) az ún. *legkisebb négyzetek módszer*ét hívják segítségül: az az egyenes halad legközelebb a pontokhoz, amelytől azok függőleges

távolságainak négyzetösszege a legkisebb. (A leírás bonyolultnak hangzik, de még számítógép nélkül is viszonylag könnyen felírható az egyenlet. Persze a számítógépes programok jelentősen leegyszerűsítik az életünket.)



a. b.

14.11. ábra: A regressziós egyenes elhelyezkedése

Egy egyenest – így a regressziós egyenest is – a meredekségével és a „magasságával" (tengelymetszetével) írhatunk le. Ez a két paraméter egy képletben így néz ki:

$$\hat{Y}_i = B_0 + B_1 X_i$$

Ezen egyenlet segítségével adja meg a regressziószámítás Y „legjobb becslését". Ezzel előrejelezhetjük Y értékét X ismeretében. A képlet azt mutatja, hogy amikor az X változó éppen konkrét X_i értéket vesz fel, akkor a regressziós függvény éppen az $\hat{Y}_i$ értéknél halad.

A két paraméter közül...

- ...B_1 mutatja a meredekséget, vagyis azt, hogy amennyiben X értéke egy egységgel nő, akkor $\hat{Y}$ mennyivel nő. Ezt nevezik *regressziós együtthatónak*.
- ...B_0 mutatja a tengelymetszetet, vagyis azt, hogy amennyiben X éppen 0, akkor milyen magas értéket vesz fel $\hat{Y}$. Ezt gyakran nevezik *konstansnak* is.

Például, ha azt gondoljuk, hogy az adott munkahelyen töltött idővel nő a jövedelem, akkor a regressziós egyenletben X az adott személy által az adott helyen töltött évek száma és Y az adott személy jövedelme. A két paraméter közül (i) B_1 azt mutatja, hogy aki egy évvel régebben dolgozik ott, az átlagosan mennyivel keres többet; (ii) B_0 pedig azt, hogy a most belépők (akiknek az adott helyen töltött idejük 0) átlagosan mekkora jövedelmet érnek el.

A statisztikusok ebben az esetben is a függő és a magyarázó változó – és nem az ok és az okozat – fogalmait használják. A függő változó, vagyis Y az, amelynek az értékét előre szeretnénk jelezni. A magyarázó változó, X, aminek alapján az előrejelzést elkészítjük.

Vegyük észre: a regressziós egyenesen levő pontok, vagyis a képletből kapott $\hat{Y}$ értékek szinte soha nem esnek egybe a függő változó valós értékekeivel, a pontdiagram pontjaival. A regressziós egyenes ugyan a pontokra legjobban illeszkedő egyenes, de ez még nem jelenti azt, hogy „közel" lesz a pontokhoz, „kicsi" lesz az összesített hiba. Ennek a közelségnek, ennek az

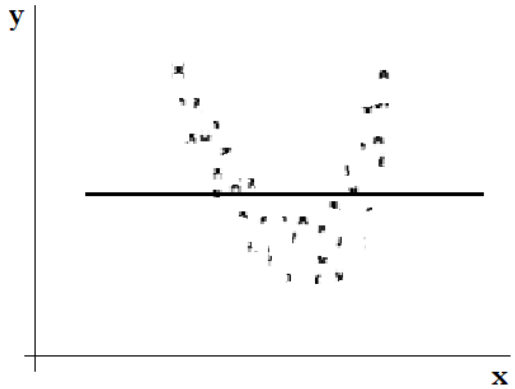
összesített hibának a megállapítása a regressziós elemzés második lépése. Tipikus módszere az ún. *determinációs együttható*, az R^2 becslése.

Az R^2-et tipikusan így szoktuk interpretálni: ez azt mutatja, hogy a függő változó, vagyis Y értékéből (példánkban a jövedelem nagyságából) a regresszió mennyit magyaráz. (Ha R^2= 0.6, akkor a bevett formula szerint a regresszió Y értékét 60%-ban magyarázza.) De érdemes R^2 jelentését kicsit pontosabban is megérteni. A determinációs együttható valójában azt teszteli, hogy amennyiben Y-t értékét előre akarjuk jelezni, akkor a regresszió mennyivel ad pontosabb előrejelzést, mint amit az átlag alapján kapnánk. Ha nem lenne regresszió, akkor Y legjobb becslését Y átlaga adná. A determinációs együttható azt kérdezi, hogy (i) amennyiben az adott egyed kapcsán a magyarázó változó értékét tudjuk és (ii) ismerjük a regresszió képletét, akkor ez a becslésünk mennyivel lesz ennél pontosabb.

R^2 valójában a – 14.3.2. pontban megismert – korrelációs együtthatónak felel meg. Egészen pontosan az R^2 értéke X és Y közötti korrelációs együttható négyzete lesz. (Ez az összefüggés azért is fontos, mert ez a korábban látott korrelációs együttható egy fontos tulajdonságát is megvilágítja: a korreláció azt mutatja, hogy egy ilyen – ott be nem mutatott – regressziós becslés mennyivel pontosabb, mint, ha csak Y átlagértékét ismernénk.)

Mielőtt továbblépünk, ki kell térni a regressziós vizsgálat korlátaira.

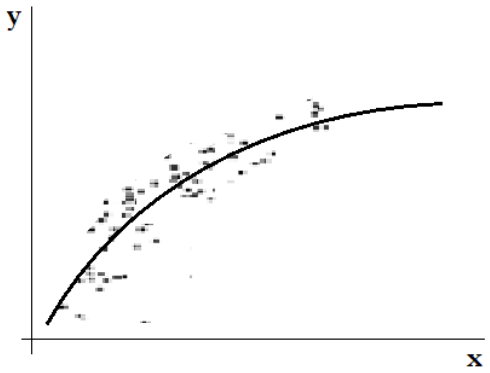
Az egyik legfontosabb probléma – mint a korreláció kapcsán is láttuk –, hogy az összefüggés lineáris kapcsolatot tételez. A regressziós egyenes mindig egyenes: a regresszió a pontokhoz legközelebb fekvő egyenest keresi. Akkor is egyenest keres, ha a pontdiagram pontjai nem egy egyenesre illeszkednek. (Lásd a 14.12. ábrán.) De a regressziószámításnál ez az akadály könnyen leküzdhetető. Ha felismerjük, hogy a pontdiagram nem egy egyenesre, hanem valami másra „hasonlít", valamilyen más alakzatra illeszkedik inkább, akkor ezt – némi „manipulációval" – beépíthetjük a modellbe. Például, ha úgy tűnik, hogy a ponthalmaz jobban illeszkedik egy másik (a matematikában jól ismert) függvényalakra, akkor az egyenletben a magyarázó változó nem X lesz, hanem X-nek az adott függvénye. Ebben az esetben ún. *nem-lineáris regressziót* végzünk. (De az egyenlet ugyanaz, mint fent, csak az X nem önmagában, hanem egy ilyen átalakítás után szerepel benne.)



14.12. ábra: Nem lineáris összefüggés és a regressziós egyenes

Sokszor előfordul – különösen a gazdasági, társadalmi folyamatok elemzésekor –, hogy a ponthalmaz alakja nem lineáris, vagyis nem egy egyenesre illeszkedik, hanem inkább egy

logaritmusfüggvényre emlékeztet. (Lásd a 14.13. ábrán.) Ilyenkor a látott regressziós egyenlet helyett egyszerűen a $\widehat{Y}_i = B_0 + B_1 lnX_i$-t írjuk fel.
Persze az interpretációkor vigyázni kell! Itt B_1 nem azt írja e, hogy ha X eggyel nő, akkor Y mennyivel nő, hanem azt, hogy ha X logaritmusa 1-gyel nő, akkor Y mennyivel nő. Matematikában jártasak emlékezhetnek: X logaritmusa akkor nő 1-gyel, ha X értéke körülbelül 2,71-szeresére nő. Az értelmezés tehát ez: ha X értéke 2,71-szeresére nő, akkor Y értéke várhatóan B_1-gyal nő.



14.13. ábra: Logaritmikus összefüggés

A másik fontos probléma az X-ek egymástól való *függetlensége*. A regresszió akkor ad pontos képet, ha az adatbázisban egymást követő X értékek között nincs összefüggés. Ez így bonyolultan hangzik (az is...), de a mi számunkra most elég, ha visszautalunk az 1. alfejezetben látott példára. Az idősorok és a területi adatok esetén az egymáshoz közeli adatok nem függetlenek egymástól. Ezért ezek esetében az itt bemutatott elemzés téves eredményeket ad.

A statisztikusok azonban nem esnek kétségbe, ha a függetlenség nem teljesül. Regressziószámítás esetén ugyanis ezt a problémát is kezelni lehet az adatok „manipulálásával". Például általában elég (vagyis a regressziószámítás már jó eredményt ad), ha nem a magyarázó és a függő változó konkrét értékeit írják be a képletbe, hanem az éves változását. Például nem azt tesszük be a képletbe, nem azt elemezzük, hogy 2020-ban mennyi a GDP, hanem azt, hogy 2020-ban mennyivel változott 2019-hez képest.

A harmadik fontos probléma az ún. *heteorszkedaszticitás*. Ez a hibára vonatkozó feltétel. Az a kérdés, hogy a reziduumok miként alakulnak. Heteroszkedaszticitásról akkor beszélünk, ha ezek a reziduumok X növekedésével tipikusan nőnek, vagy csökkennek. Ha a hibatagok nagyságában ilyen tipikus „trendet" látunk, akkor a regresszió vélhetően nem jó becslést ad.

És végül ne feledkezzünk el a kilógó adatokról sem: azok a pontok, amelyek a többitől nagyon eltérő helyen, nagyon távol helyezkednek el, általában a maguk irányába húzzák el a regressziós egyenest és a regressziós egyenletet. (Magyarán, ha csak a többi adat alapján számítanánk ki a regressziós egyenest, akkor nagyon eltérő egyenletet kapnánk.)

14.5.2 Statisztikai következtetés regresszióra

A regresszió eddigi leírásában csak az adatbázisban levő adatok érdekeltek bennünket. De általában csak a teljes sokaság egy mintáját ismerjük – a pontdiagram is csak a mintát írja le. A teljes sokaságban meglevő összefüggést csak becsülhetjük. Becsülhetjük a „valós" regressziós egyenes két paraméterét is, a tengelymetszetet (B_0-t) és a meredekséget (B_1-et).

Gyakran keveredést okoz, hogy a regressziószámításkor a becslés szót két dologra is használjuk. Gyakran úgy fogalmazunk, hogy a regressziós függvény „becslést ad" Y értékére. Ez az $\widehat{Y}_i$. Abban az értelemben becslés ez, hogy a egyenes pontjai és a valós pontok mindig eltérnek egymástól. Mindig lesz hiba, ε.

De maga ez az egyenes (egészen pontosan annak a két paramétere, a B_0, és a B_1) is csak egy becslése annak az egyenesnek, amit akkor kapnánk, ha a teljes sokaságot, minden egyed adatát ismernénk. Ezt a becslést nevezzük innentől statisztikai becslésnek. (És a keveredés elkerülése érdekében mondtuk eddig, hogy $\widehat{Y}_i$ előrejelzést ad és nem becslést Y értékére.)

Ha egy mintából következtetünk egy sokaságra, akkor – mint láttuk – kétféle statisztikai becslést kell adnunk: pontbecslést és intervallumbecslést. A pontbecslés ebben az esetben sem bonyolult: amit a minta alapján kiszámolunk, azt tekinthetjük a „valóság legjobb becslésének". Ha a minta alapján a B_0 és a B_1 érték jött ki, akkor nincs okunk feltételezni, hogy a teljes sokaságot leíró regressziós egyenes együtthatóit valamilyen más érték jobban leírná.

Ahogy általában a becslésnél mindig, ebben az esetben is az *intervallumbecslés* az érdekesebb. Az intervallumbecslés logikája itt sem tér el attól, amit fent láttunk: adott konfidencia (megbízhatósági) érték mellett megadjuk azt a felső és egy alsó határt, amik között a regressziós együttható a valóságban lehet. (Elvileg megadhatjuk ezt mind a tengelymetszetre, mind a meredekségre – de a gyakorlatban általában csak a meredekség érdekes.) Hasonló módon: adott megbízhatósági (konfidencia) érték mellett megadjuk azt a felső és alsó határt, amik között a regresszió alapján kapott értékbecslés lehet.

Regressziószámításkor intervallumbecslést azonban ritkán közölnek a statisztikusok. Inkább a statisztikai következtetéselmélet másik pontjára a *hipotézistesztelésre* koncentrálnak. Emlékezhetünk: ennek lényege, hogy empirikus szignifikancia-szinteket, p-értékeket adunk meg. Ezek azt írják le, hogy a szakmai hipotézisünkkel ellentétesen megfogalmazott nullhipotézis igaz lehet-e, ha a mintában adott értékeket látjuk. Pontosabban: mekkora az esélye, hogy a mintából épp az adott pontbecslés adódik, ha a nullhipotézis igaz. A regressziószámítás kapcsán két nullhipotézist tesztelünk.

1. Az egyik nullhipotézis a regressziós együtthatóra vonatkozik. E szerint a kapott regressziós együtthatót csak véletlenül találtuk – és a teljes sokaságban a regressziós együttható nulla. Ha egy együttható „szignifikáns", akkor nagyon kicsi a p-érték, nagyon kicsi az esélye annak, hogy ilyen nagyságú együtthatót találjunk egy mintában, ha a teljes sokaságban X és Y közötti nincs összefüggés (vagyis a valóságban $B_1=0$).

   A mért érték és a nullhipotézis szerinti érték eltérését ebben az esetben t-értékkel fejeztünk ki. Ez a fent látott tesztstatisztika. Sok esetben a statisztikusok ezt a t-értéket is közlik. (Esetleg a p-értékkel együtt.)

2. A másik nullhipotézis szerint a teljes sokaságban a regressziós egyenlet semmit nem magyaráz Y értékéből; a mintában csak véletlenül olyan magas a determinációs együttható. A kérdés az, hogy a teljes sokaságban a regresszió hozzátesz-e bármit az átlagon alapuló előrejelzéshez. Ezt a „hozzáadott értéket", ugye, az R^2 értéke becsli. A nullhipotézis ezért az, hogy a determinációs együttható a teljes sokaságban 0. A regresszió hatása szignifikáns, ha R^2 olyan nagy, hogy kicsi az esélye, hogy ha a valóságban az egyenletnek nincs magyarázó ereje, akkor ezt mutassa a minta.

   Egészen pontosan egy ún. F-értéket számolunk – ez a tesztstatisztika. (Ez hasonlít a fent látott z- és t-értékekhez). Majd ehhez az F-értékhez keressük ki az empirikus szignifikanciát, p-értéket..

14.5.3. Többváltozós regressziós

Érdemes azzal folytatni, hogy a regressziószámítást tipikusan két célra alkalmazzuk. Mindkettőt az ún. többváltozós regresszió segítségével lehet leginkább elérni.

Az egyik cél, hogy előrejelezzük a segítségével Y értékét. Általában nem gondolhatjuk, hogy Y értéke egyetlen dologtól függene. Nyilvánvaló például, hogy a mottó C példájában szereplő ingatlan értéke sok elemtől függ. A többváltozós elemzés erre képes: sokféle hatást egymás mellett is be tud építeni az előrejelzésbe.

A másik cél – amire a mottó A és a B példája szolgál –, hogy megmutassuk egy magyarázó változó hatását a függő változóra. (Valóban számít-e a jövedelemnél az, hogy valaki fekete, vagy a felvételinél az, hogy nő.) Emlegettük többször, hogy ennek az összefüggésnek a vizsgálatát mindig összezavarhatja egy összemosó tényező. Elképzelhető, hogy csak azért találunk viszonylag szignifikáns együtthatót – és kétváltozós regresszió esetén magas determinációs együtthatót –, mert valamilyen mögöttes hatás hat X-re és Y-ra is. Miközben köztük nincs is valódi kapcsolat. A többváltozós regresszió egyik előnye, hogy annak révén tesztelhetjük az ilyen egyéb tényezők hatását. Egészen pontosan: úgy vizsgálhatjuk X hatását Y-ra, hogy „kiszűrjük" ezen „háttérváltozók" hatását.

Érdemes tehát áttekinteni a többváltozós regresszió logikáját. A többváltozós regresszió esetén nem egytelen magyarázó változó kerül a képletbe, hanem több. (Az eddig elemzett kétváltozós regressziónál csak két változóval dolgozunk, egy magyarázó és egy függő változóval.) Ennek módszere, értelmezése azonban, szerencsére, nem sokban tér ez el a kétváltozós regresszióétól.

Kiindulásként képzeljük el a ponthalmazt, ha nem egyetlen magyarázó változónk van, hanem kettő, X_1 és X_2! (Mondjuk az ingatlan kora és az, hogy milyen messze van a legközelebbi metrómegállótól.) Természetesen továbbra is egy függő változónk lesz, Y. (Most: az ingatlan négyzetméter-ára.) Ebben az esetben a ponthalmazt nem rajzolhatjuk le egy papírra, hanem egy háromdimenziós térben kell elképzelni. A térben egy-egy pontot három koordinátával tudunk leírni: az egyik lesz az egyik magyarázó változó, a másik a másik, míg a harmadik (a pont „magassága") pedig a függő változó. A regressziószámítás ebben az esetben ugyanazt teszi mint a kétdimenziós esetben: megadja annak előrejelzésnek a képletét, amelytől a valós Y_i értékek a lehető lekevésbé térnek el.[9] Az egyenlet most

$$\hat{Y}_i = B_0 + B_1 X_{1i} + B_2 X_{2i}$$

A hiba továbbra is: $\varepsilon_i = Y_i - \hat{Y}_i$

Ez a képet nagyon hasonlít a fentihez. Csak a regressziós együtthatók (most kettő van: B_1 és B_2) jelentését kell pontosan érteni. Ezek most *parciális regressziós együtthatók.* Azért parciálisak, mert azt becslik, hogy amennyiben a képletben szereplő többi magyarázó változó értéke nem változik, csak az adott változó értéke nő egy egységgel, akkor hogyan reagál erre a függő változó – mennyivel nő várhatóan Y értéke. A parciális regressziós együttható épp a sokszor emlegetett *ceteris paribus* hatást becsli: semmi más nem változik, csak az adott magyarázó változó. (Vagyis, ha X_1 az ingatlan életkora és X_2 a metróállomástól vett távolság, akkor B_1 azt becsli, hogy két a metróállomástól ugyanolyan távol levő ingatlan közül az egy évvel idősebb átlagosan mennyivel ér többet, vagy kevesebbet.)

A változók és az értékbecslés kapcsán ebben az esetben is ugyanazokat a vizsgálatokat végezzük el, amiket az előbb láttunk.

---

[9] Amelytől a hibák négyzetösszege minimális.

- Elemezhetjük adott változó hatásának szignifikanciáját: vajon csak véletlenül találtunk-e olyan regressziós együtthatót (B-értéket) a mintában. (Vagyis számíthatjuk a *t*-értéket és az empirikus szignifikanciát mutató *p*-értéket.)
- Elemezhetjük azt, hogy a modell teljes magyarázó ereje hogyan alakul – mekkora az R^2. És azt is, hogy vajon ezt csak véletlenül találtuk-e.

Az *F*-érték képlet:

$$F = \frac{\sum(\hat{Y}_i - \bar{Y})^2 / m}{\sum \varepsilon_i^2 / (n - m - 1)}$$

ahol m a regresszióba bevont magyarázó változók száma; n a minta elemszáma. Látszik, hogy F értéke annál nagyobb

- minél inkább eltérnek a regresszió által előrejelzett Y értékek (vagyis $\hat{Y}_i$) Y átlagától (vagyis $\bar{Y}$-től).
- minél kisebb a regressziós előrejelzés mellett megmaradó hibatagok négyzetössszege

A többváltozós regresszió logikájának áttekintése után rátérhetünk a két előbb említett célra. Kezdjük az előrejelzéssel! Ez, ugye, a probléma a mottó C példájában. Ha egy regresszió nagyobb R^2 értéket (és ezzel együtt nagyobb *F*-értéket ad), akkor annak segítségével kisebb hibák mellett jelezhető előre Y (példánkban: az adott ingatlan értéke). Ennek kapcsán egy fontos matematikai összefüggést mindig szem előtt kell tartani: ha újabb és újabb változókat vonunk be egy regresszióba, akkor R^2 mindig nőni fog.[10]

Amennyiben a cél nem az előrejelzés, hanem annak becslése, hogy valamilyen változónak mekkora a magyarázó ereje (mint az A és a B példákban), akkor a többváltozós regresszió (legalább) kétféle módon segíthet ebben.

1. Bármilyen (majdnem bármilyen) olyan összemosó változó kiszűrhető általa, amelynek megjelenésétől félünk. Ezen egyéb változókat ilyen vizsgálat esetén gyakran *kontrollváltozók* nevezzük – ezek hatását szűrjük ki.

   Ha az az állítás, hogy a feketék azért keresnek kevesebbet, mint a fehérek, mert alacsonyabb az iskolai végzettségük, akkor ezt beépíthetjük a regresszióba. Lásd a 14.6. táblázatot! Ebben a faji hovatartozás mellett három másik magyarázó változó szerepel: a képzettség (a mesterszintű diploma léte), az adott munkahelyen eltöltött idő hossza és az alkalmazott beosztása. Vagyis a fajhoz tartozó regressziós együttható azt mondja meg, hogy ha két olyan ember közül, akik hasonló végzettségűek, hasonló ideje dolgoznak a cégnél, és hasonló beosztásban vannak, egy fehér átlagosan mennyivel keres többet, mint egy fekete. A *t*-érték (és a táblázatban nem szereplő – nagyon alacsony – empirikus szignifikancia-érték) azt is megmondja, hogy ez az eltérés „mennyire erős". Pontosabban: mekkora az esélye annak, hogy ha a fehérek a valóságban nem keresnek többet, akkor egy mintában ekkorra eltérést találunk.

   Érdemes kiemelni, hogy több változó regresszióba építése úgy is hathat, hogy „felszínre hozzák" az összefüggést. Ilyen az, amikor a kétváltozós elemzés kapcsán nem látjuk valamely magyarázó változó hatását, de amint többváltozóssá tesszük az elemzést megjelenik az.

[10] Más kérdés, hogy megéri-e az adatbázis ehhez szükséges növelése – az újabb változó számbavétele minden egyed esetén többletköltséggel jár. Ezt a költséget szembe kell állítani azzal az R^2 növekményével, és az empirikus szignifikancia csökkenéssel, amit ennek révén elérhetünk.

Gondoljunk az iskolai végzettségre, amely ugye a fiatalabbak körében tipikusan magasabb. Tegyük fel azt a kérdést, hogy a képzettség hogyan hat adott munkahelyen a fizetésekre. Könnyen lehet, hogy azt fogjuk találni, hogy csak elenyésző a hatása. De annak, hogy ilyen alacsony az együttható (és esetleg a szignifikancia-szint sem kielégítő) könnyen lehet az oka az, hogy a fiatalabbak jövedelme tipikusan alacsonyabb, mint az idősebbeké. Ha többváltozós elemzést készítünk, akkor megkérdezhetjük azt is, hogy az ugyanolyan korú emberek között a képzettebbek átlagosan többet keresnek-e. Vélhetően itt a képzettség hatása már erősebb lesz.

A *t-érték* és az *empirikus szignifikancia-szint* elemzésének logikája ugyanaz a többváltozós és a kétváltozó regresszió esetén. (Csak arra kell figyelni, hogy itt a hatás „parciális", abból már kiszűrtük más változók hatását.)

2. A többváltozós regresszió kínál egy másik módszert is a hatás erősségének mérésére. Megvizsgálhatjuk, hogy az adott változó mennyivel emeli a regresszió magyarázó erejét, vagyis az R^2-et. Ehhez nem kell mást tennünk, mint kétszer elvégezni a regressziót: egyszer úgy, hogy csak a kontrollváltozók szerepelnek benne, egyszer pedig úgy, hogy a bennünket érdeklő változó is. A két regresszió R^2-ének eltérése azt megmutatja, hogy mennyivel javítja az adott változó „beemelése" az értékbecslés pontosságát.

A 14.6. táblázatban például az látszik, hogy az a modell (az i modell), amelyben a faji hovatartozás is szerepel 10 százalékponttal többet magyaráz a jövedelmek sokszínűségéből.

Magyarázó változó	(i)			(ii)		
	B	Standard hiba	t	B	Standard hiba	t
mester diploma	898,55	140,36	6,40	1091,23	142,71	7,65
alkalmazás hossza	59,06	8,47	6,97	62,61	7,56	8,28
vezetői pozíció	5221,19	232,28	22,48	7001,92	202,34	34,60
ügynöki pozíció	2404,44	170,58	14,10	2741,71	192,61	14,23
associate pozíció	918,82	174,42	5,27	1081,23	246,21	4,39
fehér	394,80	137,67	2,87			
konstans	9291,51			10002,13		
R2	,76			,66		

14.6. táblázat: Jövedelem-regresszió (hipotetikus) egy vállalatnál
Forrás: Jackson et al [2011] 518

Akik eddig nagyon figyeltek, azok számára hibásnak tűnhet az érvelés. A 14.3. alfejezetben azt mondtuk, hogy a korreláció akkor alkalmazható, ha mindkét változó magas mérési szintű. És eddig a regressziót is mindig a korreláció képéből vezettük le: a ponthalmaz és az ahhoz a lehető legközelebb elmenő regressziós egyenes volt a kiindulás. A mottó A és B példájában szereplő két magyarázó változó, vagyis a nem és a rassz azonban kvalitatív, minőségi, nominális változó. Mégis beemelhetjük őket egy többváltozós regresszióba, tesztelhetjük a hatásukat. Ennek kulcsa az ún. dummy-változók módszere.

Dummy-változók használata esetén a nominális változókat olyan módon kódolják le, hogy azokból kétértékű (0-1 értéket felvevő) változók lesznek.

- Dichotóm változók esetén a helyzet egyszerű: az egyik csoport 1-es értéket kap, a másik 0-t. Innentől az ehhez a változóhoz tartozó B-érték azt mutatja, hogy az 1-essel kódolt csoportnál Y értéke átlagosan mennyivel magasabb – ceteris paribus. (Vagyis, ha az 1-

es a nő, a 0 a férfi, akkor a B-érték azt jelzi, hogy két, a többi változó szerint hasonló egyed közül átlagosan mennyivel magasabb, vagy alacsonyabb egy nő átlagos Y-értéke.)

- Többértékű változók esetén kicsit bonyolultabb a helyzet. Ilyenkor több dummy-változót képezünk. Egész pontosan eggyel kevesebbet, mint ahány kategóriánk van. Egy kivétellel minden minőségi kategória kap egy „saját dummyt": az abba a kategóriába tartozók értéke 1 lesz a többié 0. A kihagyott kategóriába tartozók az összes dummy-ban nullás értéket kapnak. A regresszió minden egyes dummy-hoz meg fog adni egy B-értéket (és kiszámolhatjuk annak *t*- és *p*-értékét). Az értelmezéskor azonban figyelni kell: itt az egyes B-értékek azt mutatják, hogy az adott csoportba tartozók, ceteris paribus, átlagosan mennyivel magasabb vagy alacsonyabb Y értéket érnek el, *mint a kihagyott csoport.*

A 14.6. táblázat kapcsán úgy fogalmaztunk, hogy abban a faji hovatartozáson túl három szempontot kontrolláltunk, mégis hat változó lett. Ez azért van, mert a beosztás jellemzésre három dummyt hoztunk létre – mivel az adott elemzés négy beosztást különített el. A vezetői pozíciót, az ügynöki szintet, az „associate" beosztást, és a legalacsonyabb pozíciókat. A dummykat úgy képeztük, hogy a legalacsonyabb pozíció lett a kihagyott változó, ezért a 14.6. táblázatban szereplő egyes együtthatók úgy olvasandók, hogy az – egyéb tekintetben ugyanolyan jellemzőkkel bíró – vezetők ennyivel keresnek többet, mint a legalacsonyabb pozícióba tartozók. Stb.) A dummyk felírását mutatja a 14.7. táblázat.

	Dummy változók és értékeik		
Beosztás	vezető	ügynök	associate
vezetői pozíció	1	0	0
ügynöki pozíció	0	1	0
associate pozíció	0	0	1
legalacsonyabb	0	0	0

14.7. táblázat: Dummy változók négy foglalkoztatási kategória esetén

A többváltozós regresszió révén ugyan sok, más eszközöknél megjelenő problémát megoldhatunk, de ennek kapcsán is megmarad jónéhány. Mindenekelőtt, itt is figyelni kell arra, hogy (i) az egyes magyarázó változók egymást követő értékei függetlenek legyenek egymástól, és arra, hogy (ii) ne jelentkezzen heteroszkedaszticitás. Ezen kívül többváltozós elemzéskor roppant fontos a multikollinearitás és az endogenitás tesztelése. Ezek ugyanis szintén téves következtetésekhez vezethetik az elemzőt.

Multikollinearitásról akkor beszélünk, ha a különböző változók között erős a statisztikai összefüggés – például magas köztük a korreláció. Ebben az esetben ugyanis az ilyen magyarázó változók regressziós együtthatói megbízhatatlanok lesznek. (Ugyanakkor, ha az elemzés célja az előrejelzés, akkor a multikollinearitás nem jelent komoly problémát.)

Egyszerű hipotetikus példán talán könnyen megérhető a probléma. Tegyük fel, hogy a lábméret valaminek jó magyarázó változója. A regressziós egyenletben azonban szerepeltetjük a jobbláb és a balláb nagyságát is – két külön változóként. (E kettő között, ugye, erős a korreláció.) Ebben az esetben a regressziószámítás komoly probléma előtt „áll": a jobb- vagy a balláb mérete hat-e –

másként: a lábméret hatását melyikre „számolja el". Végső soron véletlenszerűen fogja „megosztani" a hatást a jobb és a bal között. És lehet, hogy mindkettőnek alacsony lesz a *t*-értéke is.

Az *endogenitás problémáját* tipikusan a kölcsönös okozás problémájaként írjuk le. Ha egy többváltozós regresszió erős hatást mutat, akkor nem lehetünk biztosak abban, hogy valóban a magyarázó változó magyarázza-e a függőt. Elképzelhető ugyanis, hogy a hatás fordított.

> Azt láttuk például a 14.6. táblázatban, hogy akik régebb óta dolgoznak az adott vállalatnál, azoknak tipikusan magasabb a jövedelme. (Még akkor is, ha kiszűrjük, kontrolláljuk az egyetemi végzettség, a pozíció, vagy a faji hovatartozás, stb. hatását.) De jelenti ez azt, hogy a vállalat azért fizet többet azoknak, akik régebben dolgoznak ott, mert jobban ismerik a helyi viszonyokat, hűségesebbek, stb.? Lehet. De nem feledkezhetünk el a „fordított hatásról" sem. Az is lehet, hogy azok dolgoznak régebb óta (azok maradnak tovább) adott helyen, akik többet keresnek. Azok, akik úgy érzik, hogy őket kevésbé becsülik meg, akik kevesebbet keresnek, vélhetően korábban elhagyják a vállalatot, korábban új munkahely után néznek. Vagyis elképzelhet az is, hogy a régi munkakapcsolat „okozza" a magas jövedelmet, de az is, hogy a magas jövedelem „okozza" a hosszabb munkaviszonyt.

14.7. szövegdoboz: A regresszós tévkövetkeztetés – téves regresszió

Tegyük fel, hogy egy oktatási program kapcsán azt mérjük, hogy a gyerekek teszteredményei javulnak-e a program végére. Elvégzünk egy tesztet a program elején, és egyet a program végén. Azt az eredményt kapjuk, hogy azok, akik a program elején az átlagnál jobbak voltak, a végén közelebb kerülnek az átlaghoz (csökken az előnyük). És fordítva: azok, akik az átlagnál rosszabbak voltak, szintén közelebb kerültek az átlaghoz (csökken a hátrányuk). Jelenti ez azt, hogy a program csökkentette az eltéréseket? Vagy azt, hogy a program visszafogta a jobbakat? Vagy azt, hogy a program alapvetően a rosszabb tanulóknak segít? Ha ezeket a következtetéseket megtesszük, akkor elfeledkezünk egy fontos – és sajnos tesztelhetetlen – összemosó tényezőről. A szerencséről.

Azok között ugyanis, akik a program elején jobban teljesítettek ott vannak azok is, akiket a szerencse segített. (Éppen olyan kérdéseket kaptak, amelyek nekik jobban feküdtek, amelyeket ők könnyebben átláttak, megértettek.) És azok között, akik rosszabbul teljesítettek, ott voltak a balszerencsések. Amennyiben a szerencse fordul, akkor pusztán ez is a kiegyenlítődés irányába hat.

Ennek a – legtöbbször kiszűrhetetlen – hatásnak időnként külön nevet is adunk. Freedman és szerzőtársai ezt nevezik „regressziós effektusnak", illetve „regressziós tévkövetkeztetésnek". (Freedman et al [2005] 200)

És végül: soha nem zárhatjuk ki, hogy az összefüggés összemosó változók következménye. A többváltozós regresszió csak arra képes, hogy azokat a változókat szűrje ki, amiket beteszünk a képletbe. Ami nem jut eszünkbe, ami kimarad a képletből, az még összezavarhatja a képet. (Egy ilyen tipikus – gyakran elfeledett – hatást mutat be Freedman et al. nyomán a 14.7. szövegdoboz.)

Röviden: fenntarthatjuk amit eddig mondtunk, vagyis a legjobb, ha statisztikai elemzése kapcsán csak összefüggésekről, (kölcsönös) hatásokról beszélünk, és kerüljük az okság kifejezést!

14.5. Összefoglalás

A statisztika tudomány az adatok segítéségével próbál szakmai (vagyis nem statisztikai) állításokat igazolni, vagy cáfolni. A statisztika önmagában semmit nem tud bizonyítani – külső inputokra van szüksége. Meg kell mondjuk, hogy mit tartunk valószínűnek, mi a sejtésünk – a statisztika innentől lép a képbe. Ezt a sejtést fordítja le a maga nyelvére. Ennek alapján tesz javaslatot (ha még az adatgyűjtés előtt vagyunk) arra, hogy egy adatbázis mit tartalmazzon, illetve ennek alapján elemzi az adatbázist.

Láttuk a fejezetben a statisztika legfontosabb eszközeit: adathalmazokat egyszerűsít le egy-két változóra (adatsűrítést végez), egyes jelenségek (a statisztika, az adatbázisok nyelvén: változók) közötti összefüggéseket vizsgál, mintákból következtet hiányzó adatokra (ún. teljes sokaságokra, populációkra). A fejezet az ún. lineáris regressziós modellig jutott el. Ez az eszköz (különösen annak többváltozós formája) az, amely a legtöbb ilyen sejtés, szakmai hipotézis tesztelésére alkalmas. A lineáris regresszió révén adhatunk előrejelzést (becslést) egy változó alakulásáról – így egy ingatlan áráról, értékéről is. (Mint a mottó C példájában.) De a lineáris regresszió segíthet abban is, hogy egy tényező (például a nem vagy a faji hovatartozás) hatását is elemezzük – úgy, hogy az esetleg a hátérben meghúzódó összemosó változók már nem zavarják össze a képet. (Mint a mottó A és B példájában.)

A fejezet – többek között azért, mert alapvetően jogászok számára készül – a szokásosnál kicsit hosszabban tért ki az okság fogalmára. Láttuk, hogy a legjobb, ha a statisztikai adatok alapján inkább kerüljük annak használatát, és csak összefüggésekről, hatásokról beszélünk. (Bár néhány statisztikus is használja az okság fogalmat.)

Fogalmak

adatsűrítés
átlag
Cramer-féle asszociációs mutató
decilisek
determinációs együttható
dichotóm változók
dummy változók
egyed
egyenletes eloszlás
egymóduszú eloszlás
ellenhipotézis
előrejelzés
elsőfajú hiba
empirikus szignifikancia, p-érték
endogenitás
feltételek (vs. okok)
ferdeség
F-érték
függetlenség
függvényszerű kapcsolat
heteoszkedasztiicitás
hiba, reziduális, reziduum
hipotetikusság-probléma
idősoros vizsgálat
indirekt bizonyítás
interkvartilis terjedelem
intervallumbecslés
keresztmetszeti vizsgálat
kereszttábla
kétváltozós regresszió
kilógó pontok
konfidencia (megbízhatóság)

korrelációs együttható (Pearsons féle lineáris), r

következtetés

kvantilisek

kvantitatív, minőségi, nominális változó

kvartilisek

kvintilisek

magas mérési szintű változók

másodfajú hiba

medián

minta

minta torzítása

módusz

multikollineritás

negatív ferde, balra elnyúló eloszlás

nem-lineáris regresszió

normál eloszlás, normálgörbe, normálfüggvény

nullhipotézis

okság

ordinális változók

osztályköz

ökológiai tévedés

összemosó tényező, közös ok

parciális regressziós együttható

percentilisek

pontbecslés

pontdiagram

pozitív ferde, jobbra elnyúló eloszlás

próbastatisztika, tesztstatisztika

regressziós egyenes

regressziós egyenlet

regressziós együttható

regularitásra épülő okság

statisztikák vs. paraméterek

sűrűségfüggvény

szakmai hipotézis

számított középérték, helyzeti középérték

szignifikancia

szimmetria

szórás

szóródás, sokszínűség

sztochasztikus kapcsolat

távolság-, terjedelmi mutatók

teljes sokaság, populáció

tényellentétes okság, *conditio sine qua non* feltétel

t-érték

területi adatsor

többmóduszú eloszlás

többváltozós regresszió

trend

túldetermináltság

valószínűségi okság

variancia

z-érték

Irodalom

Babbie, Earl [2009]: A társadalomtudományi kutatás gyakorlata. *Budapest, Balassi*

Blutman László [2011]: Okozatosság, oksági mércék és a magyar bírói gyakorlat. *Jogtudományi Közlöny* 309-320.

Boronkay Miklós [2007]: A deliktuális felelősség határai. *Iustum Aequum Salutare* III 2007/4.

Doll, Richard [1955], Etiology of Lung Cancer. *Advances in Cancer Research* 3 1-50.

Dósa Ágnes [2010]: *Az orvos kártérítési felelőssége.* (2. kiadás) Budapest,HVG-ORAC

Elster, Jon [1995]: *A társadalom fogaskerekei.* Budapest, Osiris

Freedman, David – Robert Pisani – Roger Purves [2005]: *Statisztka.* Typotex, Budapest

Fuglinszky Ádám [2015]: *Kártérítési jog.* Budapest, HVGOrac

Hitchcock, Christipher [2018]: Probabilstic Causation. in: Edward N. Zalta (szerk.): *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition) https://plato.stanford.edu/archives/fall2018/entries/causation.probabilistic

Honoré, Tony [1995]: Necessary and Sufficient Conditions in Tort Law. in David G. Owen (szerk.): *Philosophical Foundations of Tort Law.* Oxford, Clarendon Press 363–385.

Hume, David [1751/1973]: *Tanulmány az emberi értelemről.* Budapest, Magyar Helikon – Európa

Huoranszki Ferenc [2001]. *Modern metafizika.* Budapest, Osiris

Jackson, Howell – Louis Kaplow – Steven Shavell – W. Kip Viscusi – David Cope [2011]: *Analytical Methods for Lawyers.* 2nd edition. New York: Thompson Reuters/Foundation Press, Chapter 8-9

Mackie, John L. [1965]: Causes and Conditions. *American Philosophical Quarterly* 2 245-264

Mackie, John L. [1974]: *The Cement of Universe. A Study of Causation.* Oxford, Oxford University Press.

Menyhárd, Attila [2015]: Basic Questions of Tort Law from a Hungarian Perspective. In Helmut Koziol (szerk.): *Basic Questions of Tort Law from a Comparative Perspective.* Wien, Jan Sramek Verlag

Németh Renáta [2020]: A statisztikai megközelítés. in: Jakab András – Sebők Miklós (szerk): *Empirikus jogi tanulmányok paradigmái és módszertana. Gyakorlati bevezetés jogászoknak.* megjelenés alatt

Simon Dávid [2020]: Leíró statisztikai alapok. in: Jakab András – Sebők Miklós (szerk): *Empirikus jogi tanulmányok paradigmái és módszertana. Gyakorlati bevezetés jogászoknak.* megjelenés alatt

Szalai Ákos [2017]: Okozatosság a kártérítési jogban – joggazdaságtani megfontolások. *Polgári Jog.* 2017/1