

Az orvosi megismerés módszertana (és az orvosi kutatások kritikus értékelése)

Ferenci Tamás

2025. július 9.

Tartalomjegyzék

Előszó	3
1 Bevezető gondolatok, és pár történeti szempont	5
1.1 Az érvágás esete, avagy hogyan lehet valamit 2000 éven át rosszul csinálni úgy, hogy senkinek nem tűnik fel	6
1.2 Empirikus megismerés az orvostörténetben	9
1.3 Vita egy állásponttal	10
2 Az empirikus megismerés alapgondolata és a confounding	13
2.1 Az okozatiság nyomában	13
2.2 A confounding problémája	15
2.3 Zavaró változóktól a megzavart olvasóig	18
3 A confounding megoldásai – megfigyelés és kísérlet	22
3.1 Egy aranyérmes megoldás	22
3.2 Megfigyelés és kísérlet	25
3.3 A jó, a rossz, és a közepesnél némileg gyengébben jó	27
3.4 Jót tesz-e repülőgépből való kiesésnél, ha van nálunk ejtőernyő?	27
4 Megfigyeléses vizsgálatok a gyakorlatban	30
4.1 A rétegzés	30
4.2 A standardizálás	33
4.3 Kitérő arról, hogy miért jó, ha egy országban sokan halnak meg rákban	42
5 A véletlen szerepe az orvosi vizsgálatok kiértékelésében	45
5.1 A véletlen ingadozás átka	46
5.2 Isten létének bizonyítása a statisztika eszközeivel	47
5.3 A mágikus <i>p</i> -érték	50
5.4 Döntéshozatal véletlen jelenlétében	52
5.5 Az egyik legfontosabb félreértés	53
5.5.1 Terroristák a városban	53
5.5.2 Kis kitérő: na de mi köze ennek az orvosláshoz?	55
5.5.3 Nagy kitérő: Dr. GépBayes rendel	56
5.5.4 Na de mi köze ennek a véletlen ingadozás kezeléséhez?	59
5.6 A véletlenség megszelídítése	61
5.6.1 Megoldási lehetőségek nyomásban	62

5.6.2	Az orvosi kutatások edzeni mennek	65
5.6.3	Kutatások mintaméretének tervezése	67
6	Végpont és lemérés	71
6.1	Mikor adjunk gyógyszert?	71
6.2	Kísérletek résztvevői, avagy végre kiderül, hogy jót tesz-e repülőgépből kiesésnél, ha van nálunk ejtőernyő	72
6.3	Mutatók stabilitása	76
6.4	Döntéshozatal és az abszolút és relatív mutatók	79
6.5	Különböző kimenetek közös nevezőre hozása	81
6.6	Kitérő: patikamérlegen az emberélet	82
6.7	Végpontok megválasztásának problémái	83
6.8	Több végpont egyidejű vizsgálata	84
6.8.1	Vadászni mentünk... szignifikanciára	85
6.8.2	Egy nem csak biostatisztikai tanulságokkal bíró kitérő	88
6.8.3	A jóhiszemű kutatók nehézségei	89

Előszó

Manapság mind az érdeklődő laikusokra, mind az orvosokra, szakdolgozókra csak úgy zúdulnak a különféle egészségügyi, orvosi információk, melyeket a helyükön kell(ene) kezelni. Tényleg csökkenti-e a vérnyomást ez a vérnyomás-csökkentő gyógyszerjelölt? Okozhat-e agyvérzést mint mellékhatás? A vöröshús-fogyasztás valóban növeli a rákkockázatot? A császármetszéssel születés vezethet 1-es típusú cukorbetegséghez? Ahhoz, hogy értsük, hogy e kérdéseket hogyan lehet empirikusan megválaszolni, és ahhoz, hogy az ilyeneket megválaszoló kutatásokat értelmezni, értékelni és kritikusan értékelni tudjuk, az empirikus kutatások módszertani alapjait kell ismerni. Ez már ma is fontos, és napról-napra csak egyre fontosabb lesz, még pedig nem csak minden orvos és egészségügyis, de minden, a sajtóban olvasottakat valóban megérteni vágyó érdeklődő számára.

(Tapasztalataim szerint ez az orvosoknak sem könnyű: az „evidence-based medicine” varázsszó-ként terjedése ellenére számos orvos valójában nem rendelkezik kellő jártassággal e téren. Nem az ő hibájukból: az egyetemi biostatisztika oktatás inkább arra készít fel, hogy hogyan kell számokat gyártani, nem arra, hogy a mások által gyártott számokat hogyan kell értelmezni, értékelni.)

A fenti motivációból fakadóan szívügyem e témakör bemutatása. Nem, ez nem „biostatisztika”, legalábbis abban az értelemben, amit a legtöbb orvos ért ez alatt, nem képletek, levezetések és bizonyítások lesznek ebben a jegyzetben. Ha már mindenkor címkéznünk kell, akkor módszertan. Ahogy írtam is, ennek ismerete – szemben a biostatisztikával – véleményem szerint kivétel nélkül minden orvosnak, egészségügyi területen dolgozónak és minden érdeklődő laikusnak fontos, aki orvosi kutatások eredményeit olvassa és akarja megérteni; ehhez próbálok segítséget nyújtani e jegyzettel. A nyelvezete, felépítése olyan, hogy reményeim szerint minden különösebb előismeret, és orvosi tudás nélkül is végig követhető.

Különösen fontosnak tartom az orvosi kutatások kritikus értékelésének témakörét: mik a tipikus buktatók, csapdák, jóhiszemű és kevésbé jóhiszemű félreértési, félrevezetési lehetőségek. Sajnos e terület mágnesként vonzza mind félreértésekét, mind a szándékos félrevezetésekét, valószínűleg nem függetlenül attól, hogy sok részterülete van, amihez ideológiaiag nagyon motivált szereplők szólnak hozzá, illetve amelyek mögött komoly anyagi érdekek is megjelennek.

Remélem, hogy ez a jegyzet segítséget nyújt abban, hogy az olvasó mindezek ellenére mégis magabiztosabban tudjon navigálni az ilyen információk áradatában.

Az anyaggal kapcsolatban minden észrevételt, javaslatot, kritikát örömmel várok a tamas.fere_nci@medstat.hu email-címen!

Ez a jegyzet az Interpress Magazinban megjelent cikksorozatom kibővített, szerkesztett változata.

A dolgozathoz adott segítségükért szeretnék köszönetet mondani (alfabetikus sorrendben) Szilágyi Eszter Szabinának és Tóth Andrásnak.

1 Bevezető gondolatok, és pár történeti szempont

Nemrégiben egy cikket olvashattunk az egyik népszerű tudományos magazinban, mely egy érdekes orvosi eredménnyel indított: egy svéd egyetem kutatói félmillió gyermek környezetét vizsgálták meg, és azt találták, hogy ahol magasabb a légszennyezettség, ott több a mentálisan beteg gyermek. A légszennyezés tehát mentális betegséget okoz! Vagy mégsem...?

Az orvostudomány egy jelentős része ilyen, és ehhez hasonló kérdésekre igyekszik választ adni: okoz-e mentális betegséget a légszennyezés? A mobiltelefon-használat agydaganatot? A vöröshús-fogyasztás vastagbélrákot? A császármetszéssel születés megnöveli-e annak kockázatát, hogy a gyermeknek később 1-es típusú cukorbetegsége lesz? És ha az anya paracetamolt szed a terhesség alatt, attól lehet a gyermek autista? Itt van ez az új vérnyomás-csökkentő gyógyszerjelölt, vajon csökkenti-e tényleg a vérnyomást? És okozhat-e alvászavart mellékhatásértől?

E kérdésekre számos módszerrel kereshetjük a választ. Alapul vehetünk biológiai (élettani, körélettani) megfontolásokat, kereshetünk állatmodelleket, amik lehetővé teszik a jelenségek vizsgálatát, tekinthetünk analóg példákat más területről, gyárthatunk matematikai modellt, azonban a jelen cikksorozat tárgya egy más jellegű, ám egyre fontosabb módszer: az empirikus vizsgálat. Az empirikus annyit tesz: „tapasztalati”, úgyhogy rögtön pontosítanom kell: a legrosszabb orvosi megismerési módszerek is tapasztalatokon alapulnak, ezért talán jobb, ha úgy mondjuk: szisztematikus empirikus vizsgálat. Empirikus, mert az alapján próbáljuk megválaszolni a kérdést, hogy begyűjtünk tényadatokat gyermekek környezetének légszennyezettségéről és a tényleges megbetegedéseikről, és szisztematikus, mert ezt nem ötletszerűen, hanem valamilyen terv szerint tesszük. A kérdés tehát adott: miután megvannak ezek az adatok, hogyan következtethetünk azokból arra, hogy okoz-e mentális betegséget a légszennyezés? Látni fogjuk a következőkben, hogy az e kérdés megválaszolásához vezető úton nagyon sok csapda van elrejtve...

Számos esetben persze nem lehetséges, vagy felesleges az empirikus kipróbálás, hiszen az egyéb bizonyítékokat is elégségesnek gondoljuk. Urológusok emberemlékezet óta tanácsolják, hogy sok folyadékot kell inni a vesekő megelőzésére – de kipróbtalva-e bárki is empirikusan, hogy ez tényleg csökkenti-e a vesekő előfordulását?! Ugyan már! De minek is, a vesekőről tudjuk, hogy a túltelítődő vizelet talaján alakul ki (biológiai megfontolás), a sok folyadékivás pedig megakadályozza ezt, mivel csökkenti a vizelet sűrűségét (más tudományterületről vett ismeret) – tehát a sok folyadékivás nyilván csökkenti a vesekő kockázatát.

Az óvatosság persze mindenkor indokolt – sok történelmi példát lehetne hozni, amikor a „nyilvánvalóan jó” (és ezért empirikusan ki sem próbált) tanácsok bizonyultak nem is annyira jónak... Ennél ráadásul sokkal rosszabb dolgok is törtéhetnek akkor, ha a más jellegű alátámasztások sem túl acélosak.

1.1. Az érvágás esete, avagy hogyan lehet valamit 2000 éven át rosszul csinálni úgy, hogy senkinek nem tűnik fel

Erre – némi kírás ironikus módon – a legjobb példát pont az orvostörténelem egyik legáltalánosabban használt gyógyító beavatkozása jelenti: az érvágás. Az 1.1. ábrán láthatjuk az eljárás néhány ábrázolását, és ha jobban megnézzük, akkor nem sok kétségünk maradhat az előbbi áltítás igazságtartalma felől. Elképesztő az időbeli átfogás (láthatunk érvágást ókori görög vázán és fényképen is!), szinte teljes a térbeli elterjedés (használták keleten, használták a görög orvosok, használták az arab orvosok, használták a középkori és reneszánsz Európában), röviden szóval: legkevesebb kétezer éven keresztül gyakorlatilag minden kultúra orvoslása alkalmazta az érvágást a legkülönfélébb betegségek gyógyítására.

Miközben ma már elég világosan tudjuk, hogy a legtöbb betegség esetén minimum nem tesz jót, ha a beteget kivéreztetik... Akkor meg mégis, hogyan lehet, hogy ez vált az orvostörténelem legszélesebb körben alkalmazott gyógyító eljárásává? A válasz nagyon rövid: azért, mert senki nem próbálta ki, hogy használ-e. Senki. Nem lehet azt mondani, hogy nem „empirikus alapon” használták, ellenkezőleg, a legfontosabb indok az volt, hogy a „tapasztalataink szerint” működik, ám a „tapasztalataink szerint” magyará lefordítva azt jelenti, hogy a „benyomásaink szerint” – márpedig az ilyen benyomások rettenetesen hibásak lehetnek. Az érvágást a benyomások támasztották alá, valamint az támasztotta alá, hogy „a mestereinktől így tanultuk” (és emiatt így szoktuk meg), hogy a legnagyobb szaktekintélyek ajánlják... szisztematikus vizsgálat tárgyává azonban senki sem tette az érvágást.

Egészen 1828-ig. (1828-ig! – miközben már az ókorban is használták!) Ekkor egy Pierre-Charles-Alexandre Louis nevű francia orvos, az orvostörténelemben először, *kipróbálta*, hogy működik-e az érvágás (1.2. ábra).

Fogott 77 beteget, akit a La Charité kórházban kezeltek tüdőgyulladással (ebben a betegségen a kor orvosi gyakorlata szerint nem is volt kérdés, hogy érvágást kell alkalmazni, mindegyiknél meg is tették), ám Louis doktor kettéosztotta őket a szerint, hogy korán, 4 napon belül, vagy pedig későn, 4 napon túl hajtották-e végre az első érvágást. És láss csodát: a 41 beteg közül, akik pechükre a korai érvágásos csoportba kerültek, 18 halt meg (44%), a 36 beteg közül – akiknek volt egy kis idejük előbb spontán gyógyulni – pedig 9 (25%). A különbség majdnem kétszeres – és ebben a pillanatban, 2000 év után, az érvágás elkezdett kikerülni az orvosi gyakorlatból. (Nagyon jellemző, de külön tanulmány tárgya lehetne, hogy erre még ez után is majd’ száz évet kellett várni...) Ez a történet gyakran eszembe jut, amikor olyanokat hallok, hogy „ezt az eljárást már kétszáz éve használják, kiderült volna, ha nem is működik!”...



(a) Görög vázán, kb. Kr. 470-480.



(b) Fényképen, 1860.



(c) Iráni edényen, Kr. u. 1250 körül.



(d) Középkori kódexen, késői 13. század.

1.1. ábra. Az érvágás különböző ábrázolásai.



1.2. ábra. Pierre Charles Alexandre Louis (1787-1872).

Louis kutatásának számos limitációja volt, de a munkájában az a fantasztikus, hogy még ezekre is, legtöbb esetben mai szemmel nézve is teljesen helytállóan, rámutatott. Sok vita folyik arról mind a mai napig, hogy az orvosi döntéshozatalban mekkora szerepe lehet, legyen a szisztematikusan gyűjtött empirikus bizonyítékoknak, és mennyi az orvos nem-szisztematikus tapasztalatainak, benyomásainak. Anélkül, hogy ebben állást foglalnék, remélem a fenti példa legalábbis intő figyelmeztetés a benyomásokra, „így tanultam és így szoktam meg”, „híres orvosoknak is ez a véleménye” típusú bizonyítékokra történő alapozás kapcsán.

1.2. Empirikus megismerés az orvostörténetben

Az empirikus orvoslás története nagyon régi is, meg nagyon új is.

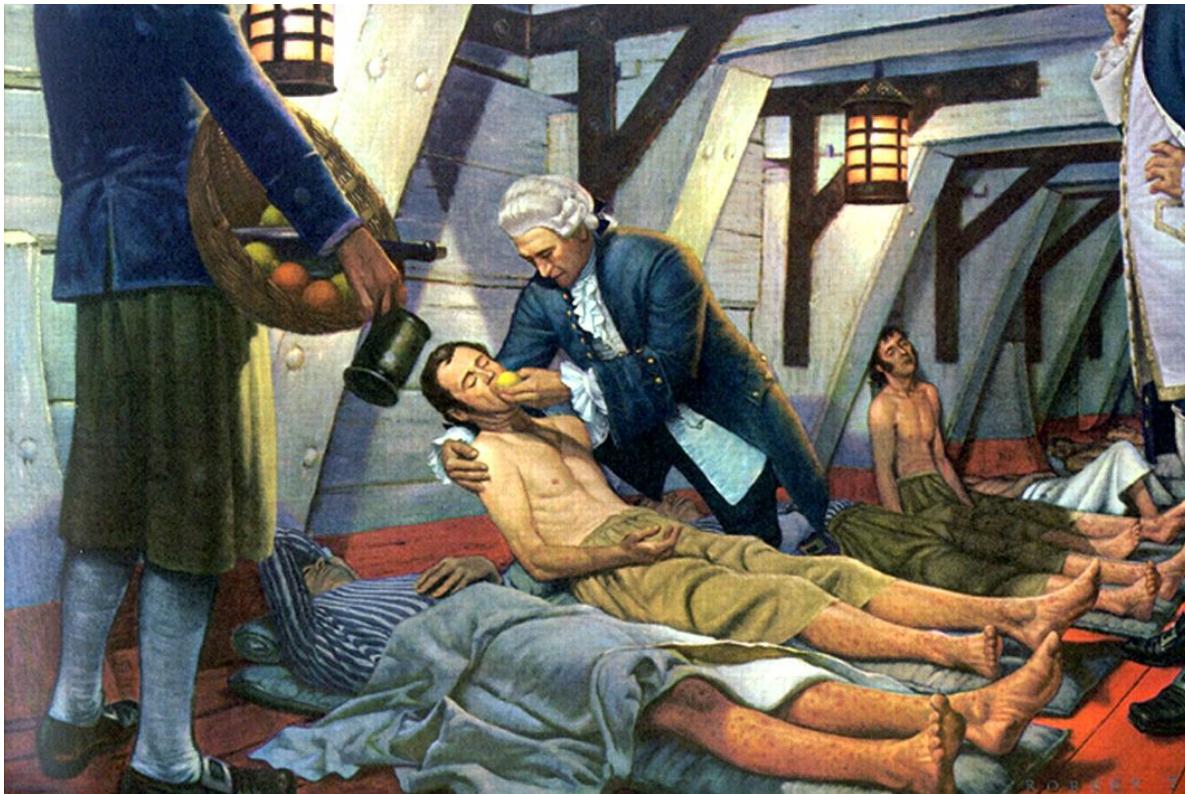
Az első említése valószínűleg sokak számára meglepő, úgyhogy hadd idézem szó szerint: „10 És mondá az udvarmesterek fejedelme Dánielnek: Félek én az én uramtól, a királytól, aki megrendelte a ti ételeteket és italotokat; minek lássa, hogy a ti orcátok hitványabb amaz ifjaknál, akik egykorúak veletek? és így bűnbe kevernétek az én fejemet a királynál. 11 És mondá Dániel a felügyelőnek, akire az udvarmesterek fejedelme bízta vala Dánielt, Ananiást, Misáelt és Azariást: 12 Tégy próbát, kérlek, a te szolgáiddal tíz napig, és adjanak nékünk zöldségféléket, hogy azt együnk, és vizet, hogy azt igyunk. 13 Azután mutassák meg néked a mi ábrázatunkat és amaz ifjak ábrázatát, akik a király ételével élnek, és aszerint cselekedjél majd a te szolgáiddal. 14 És engede nékik ebben a dologban, és próbát tőn velük tíz napig. 15 És tíz nap múlva szebbnek látszék az ő ábrázatuk, és testben kövreibungek valának mindenkorának az ifjaknál, akik a király ételével élnek vala.”

Ez az idézet Dániel könyvének 1. fejezetéből származik a Bibliából (Károli Gáspár fordítása); történészek a Kr. e. 2. századra teszik a szöveg keletkezését.

Noha vannak bizonyos problémái (nem világos, hogy Dániel beszerezte-e a Regionális Kutatásetikai Bizottság engedélyét, hogy az alanyok teljes írásos tájékozott beleegyezéssel vettek-e részt a kutatásban, hogy Dániel előzetesen regisztrálta-e a kutatást nemzetközi adatbázisban, az eredményközlés elégtelen, nem derül ki, hogy a hatás milyen statisztikai próbával került vizsgálatra stb.), mégis egy szempontból egészen fantasztikus, ami ide le van írva: felmerült a gondolat – még egyszer, Kr. e. 2. században vagyunk! –, hogy egy orvosi kérdést empirikus alapon döntsenek el. Nem benyomás alapján, mint az előbbi példában láttuk, nem mesterek hagyományai, szent iratok vagy vakszerencse alapján, hanem szisztematikus empirikus vizsgállattal.

Ez a gondolat, bár helyenként felbukkan az orvostörténetben, lényegében kétezer évre feledésbe merült és – egy-két említéstől eltekintve – csak a 18. századtól kezdve jött újra elő. A leghíresebb – bár egy-két részletében ma már vitatott – példa James Lind angol hajóorvos vizsgálata a skorbutról, 1747-ből. Lind a skorbut ellenészét keresve egy hajón a HMS Salisbury hajó 12 beteg tengerészét 6 kétfős csoportra osztotta, és – egyebekben ugyanolyan táplálkozás mellett – mindegyiknek olyan szert adott, amellyel kapcsolatban felmerült a lehetőség, hogy

gyógyítja a skorbutot. (Két tengerész szerencsétlenségére ezek között volt a hígított kénstav is...) Csakhogy az egyik csoport narancsot és citromot kapott – az ezekkel kezelt minden tengerész már 6 nap után sokkal jobban lett, az egyik szinte meg is gyógyult; semelyik másik csoportban sem tapasztalt ilyet Lind ([1.3. ábra](#)).



1.3. ábra. James Lind: A skorbut legyőzője. Robert A Thom festménye.

Ha jobban megnézzük, akkor világos, hogy ez szinte tökéletesen másolja Dániel könyvének leírását. Gondoljunk bele, majdnem pontosan 2000 év kellett ahhoz, hogy valaki először tényleg jól dokumentáltan végrehajtsa ezt egy konkrét orvosi probléma vizsgálatára!

Louis, Lind és társaik minden erőfeszítésével együtt is azonban e gondolatok igazán csak a 20. századra értek be, és kezdték egyre jobban formálni az orvosi megismerést.

1.3. Vita egy állásponttal

A bevezetést hadd zárjam egy kissé szubjektív gondolattal. Egy vitában mondta ezt egyszer nekem: „Én nem hiszek az ilyen statisztikai dolgoknak, én a saját tapasztalataimnak hiszek”. (A vitapartner az „ilyen statisztikai dolgok” alatt természetesen igazából nem statisztikai dolgokat értett, hanem a szisztematikus empirikus kutatásokat.) Nem amellett akarok érvelni,

hogy miért nem igaz ez a mondat – elvégre ez inkább egy vélemény –, hanem, hogy miért hamis már a kiindulópontja is.

Tény, hogy napjainkra az orvosi kutatások, különösen épp a szisztematikus empirikus vizsgálatok a legtöbb ember számára nehezen követhető, zavaros matematikai részletekkel teletűzdelt, átláthatatlan dolgokká váltak; ennyiben érhető a fenti álláspont, hogy az ember jobban bízik a „saját tapasztalatában”. Rettenetesen fontos azonban rögzíteni, hogy a valóságban ezek nem szembenálló dolgok, épp ellenkezőleg: a kutatásokban hajszálponosan ugyanúgy saját tapasztalatok vannak! minden kutatásban szereplő adat valakinek a „saját tapasztalata”! Csak épp egyrészt nagy mennyiségen összegyűjtve, másrészt szisztematikus összegyűjtve. A mennyiség azért fontos, mert egy sor kérdés pontos vizsgálatához olyan nagy mennyiségű adatra van szükség, amennyit többé már nem lehet egyénileg begyűjteni, illetve ránézésre áttekinteni. Hány emberről tudunk személyesen tapasztalatot szerezni, illetve azt észben tartani és értékelni? 100? 200? Ez megfelelő lehet akkor, ha annyira drámai a hatás, mind Lind kutatásában, ahol már 2 matróz elég volt a kimutatásához, de reménytelenül kevés akkor, ha valamilyen kevesebb embert érintő, vagy kevésbé drasztikus hatást eredményező tényezőt kell vizsgálnunk. Ekkor sok ember adatait kell begyűjtenünk, amelyek aggregálásához azonban már elkerülhetetlen valamilyen statisztikai módszer alkalmazása, hiszen többé már nem tekinthetők át szabad szemmel az adatok.

Még tanulságosabb azonban a második szempont: a szisztematikusság. Elvégre már az érvágás kapcsán is feltehetők volna a kérdést, hogy ugyan végül is miért nem vették észre, hogy a beavatkozás nagyon nem működik (sót)...? A kutya a szisztematikusságnál van elásva. Konkretizálva, két fontos problémához vezet a szisztematikusság hiánya, amelyek tipikusan megjelennek az ilyen „tapasztalataim szerint” jellegű kijelentésekben.

Az egyik: a kontrolláltság hiánya. Lehet, hogy az érvágás után nagy arányban haltak meg a betegek, de ezt nem volt mihez viszonyítani: érvágásban nem részesült kontrollcsoport híján ez akár még jó eredmény is lehetett volna! (Ha érvágás híján még többen haltak volna meg.) Kontrollálás hiányában egész egyszerűen nem is lehetett tudni, hogy ez az eredmény jó vagy rossz! Ez tükröződik vissza milliónyi hétköznapi kijelentésünkben, melyeket egészségügyi hatalások kapcsán teszünk. „Bevettem a gyógyszert, és két nap múlva el is múlt a megfázásom!” (A legtöbb ember számára ez azt jelenti, hogy hat a gyógyszer, miközben fogalmunk sem lehet, hogy mi történt volna gyógyszerbevétele nélkül! Ha három nap múlva gyógyultunk volna meg, akkor tényleg hatott a gyógyszer, ha két nap múlva, akkor nem ért semmit, ha egy nap múlva, akkor egyenesen lassította is a gyógyulást – miközben a fenti helyzet mindenki eshetőséget lehetővé teszi!) „10 ember lett cukorbeteg azok közül, akik ilyen gyógyszert kaptak!” (Ha gyógyszer híján közük csak 5 lett volna cukorbeteg, akkor tényleg a gyógyszer a ludas, ha 10, akkor semmi köze nem volt a megbetegedésükhez, ha 15, akkor még védett is ellene... Megint csak, nem arról van szó, hogy a 10-ne lenne igaz, hanem arról, hogy a 10-től még a három lehetőség mindegyike fennállhat!)

A másik problémát a nem szisztematikus módszerek kapcsán a különféle kognitív torzítások jelentik. Ilyenből milliónyi létezik; legjobban talán a megerősítési torzítás példáján magyarázható el, hogy miől van szó. Röviden: az emberek hajlamosak jobban emlékezni az olyan

példákra, amik megerősítik az előzetes elgondolásait, mint amik szembemennek velük. Ha tehát az ember „tapasztalataim szerint” alapon nyilatkozik, akkor – teljesen önkéntelenül, félreértés ne essék, tudattalan módon! – hajlamos arra, hogy azokra a példákra, amik egybevágnak az előzetes várakozásaikkal, emlékezzen, azokra viszont, amik cáfolnák azt, nem vagy kevésbé. Tehát amikor felidézzük, hogy „tapasztalataink szerint” hatott-e az érvágás (úgy, hogy ezt tanították a mestereink, mindenki más is ezt csinálja stb.), akkor kitűnően fogunk emlékezni azokra az esetekre, amikor a páciens hamar meggyógyult a beavatkozás után, és sokkal kevésbé azokra, amikor nem. Pontosan ugyanez játszódik le akkor, amikor nem szisztematikusan, hanem benyomásaink alapján gyűjtjük össze azokat az eseteket, amikor bevettük a gyógyszert és utána elmúlt a megfázásunk, feltéve, hogy erősen hiszünk benne, hogy hat a szer, vagy erősen hiszünk benne, hogy nem; amikor nem szisztematikusan gyűjtjük össze azokat az eseteket, amikor olyanról hallottunk, hogy valaki cukorbeteg lett a gyógyszert szedők közül, feltéve, hogy erősen hiszünk benne, hogy a betegséget a gyógyszer okozta, vagy erősen hiszünk benne, hogy nem. Újra fontos hangsúlyozni, hogy itt nem félrevezetésről vagy csalásról van szó, ez egy tisztán pszichológiai jelenség, ami elkerülhetetlenül jelentkezik akkor, ha nem szisztematikus módon dolgozunk (lévén, hogy a szisztematikus munka – bár egyébként nem teljesen érzéketlen az ilyen jellegű torzításokra! – legalábbis kevésbé érzékeny erre).

Félreértés ne essék, a szisztematikus empirikus vizsgálatoknak is vannak hátrányaik. Csak hogy a legkézenfekvőbbet említsem: a nagy, átfogó vizsgálatok gyakran igen költségesek, és előfordul, hogy e költségeket olyan szereplő – például gyógysergyár – állja, aki anyagilag érdekkelt abban, hogy egy adott eredmény jöjjön ki a kutatásból. Szó nincs arról, hogy erre ne lehetne rámutatni, épp ellenkezőleg, nagyon fontos, hogy e kérdésekről beszéljünk (én is sokszor meg fogom tenni a cikksorozatban), pont azért, hogy javítani lehessen az ebből fakadó torzításokat. De azt is világosan látni kell, hogy e rendszernek minden hibájával együtt is, az alternatívája az, amikor 2000 évig nem tünik fel, hogy az érvágás nem hatásos!

Louis egyik méltatója e szavakkal mutatta be művét 1836-ban az amerikai közönségnek: „[...] a tudományunkban [az orvoslásban] új korszak kezdődött azzal, hogy Louis bevezette a numerikus rendszernek nevezett eljárását. Végre megérkezett a valódi fény. Végre ráléptünk a biztos és nyílegyenes útra.” Mint az ki fog derülni, ez az út azonban végül is sajnos néha bizonytalannak, és szinte minden meglehetősen kacskaringósnak bizonyult...

2 Az empirikus megismerés alapgondolata és a confounding

Láttuk, hogy milyen veszélyeket rejthat magában, ha nem empirikusan, azaz tapasztalati úton vizsgálunk egy orvosi kérdést. Nem véletlen, hogy az utóbbi évtizedekben egyre jobban előtérbe kerültek a szisztematikus, empirikus vizsgálatok – csakhogy ezek értelmezése szintén számos csapdát tartogat! Úgyhogy térjünk vissza a kiindulóponthoz, az említett cikkhez: okoz-e mentális betegséget a légszennyezettség?

Első ránézésre könnyű dolgunk van: empirikusan dolgozunk ugyebár, ezért begyűjtünk tényadatokat gyermekek lakóhelyének légszennyezettségéről, és esetleges későbbi megbetegedéseikről. Szisztematikusan dolgozunk, ezért a mintavételt alkalmas terv szerint végezzük, például a néppességnélvántartó adataiból teljesen véletlenszerűen választunk ki kellően sok gyermeket. (Tehát véletlenül sem internetes kérdőívet küldünk ki, többek között a „Megbetegítették a gyermekemet a légszennyezettséggel!!!” Facebook-csoport tagjainak, megkérve, hogy idézzenek fel az összefüggést megerősítő illetve cáfoló példákat.)

A kapott eredmények, a példa kedvéért: a 100 ezer nem légszennyezett helyen élő gyermek közel 1310-nél lépett fel mentális betegség, a 100 ezer légszennyezett környezetben felnövő közül azonban 3750-ben. A különbség drámai. Márpedig szisztematikusan dolgoztunk, a lehető legjobban: véletlenszerűen választott adatokkal, empirikusan vizsgáltuk a kérdést, ráadásul igen nagy mintán, úgyhogy hátradőlhetünk és nagy nyugalommal mondhatjuk: a légszennyezettség mentális betegséget okoz!

Vagy mégsem?

2.1. Az okozatiság nyomában

A válaszhoz induljunk egy picit távolabbról. Érdemes felidézni, hogy milyen szerteágazóak azok a – fentihez hasonló – kérdések, melyekre választ szeretnénk adni az orvostudományban: Okoz-e agydaganatot a mobiltelefon használata? A vörös hús fogyasztása vastagbélrákot? A császármetszéssel születés megnöveli-e annak kockázatát, hogy a gyermeknek később 1-es típusú cukorbetegsége lesz? És lehet-e a gyermek autista attól, hogy az anya paracetamolt szed a terhesség alatt? Itt van ez az új vérnyomáscsökkentő gyógyszerjelölt, vajon tényleg csökkenti-e a vérnyomást? Okozhat-e alvászavart mellékhatásként?

Ebben a listában igyekeztem, teljesen szándékosan, a lehető legkülönbözőbb kérdéseket összegeűjteni, melyekben látszólag egyetlen közös pont sincs. Ezt azért tettem, hogy még meglepőbb legyen a következő kijelentésem: azt állítom ugyanis, hogy bármennyire is különbözőnek tűnnek, valójában kivétel nélkül az összes felsorolt kérdés mögött – és milliónyi egyéb orvosi, egészségügyi kérdés mögött – pontosan ugyanaz a séma van! Persze, a konkrét részletek nagyon eltérnek, de ha ezektől megtisztítjuk az egyes kérdéseket, akkor a mélyben minden esetben ugyanazt találjuk.

Az egyik komponens: minden kérdésben található valami, amit teljesen általános szóval *expozíciónak* fogunk hívni. Ez szó szerint „kitettséget” jelent, és tényleg azt is értjük alatta, hogy az alany ki volt téve valamilyen hatásnak. Ezt a legáltalánosabban értjük: a hatás lehet valami, amit szándékosan alkalmazunk az alanyon (pl. gyógyszert adunk neki), lehet valami, amit maga választ (pl. vörös húst fogyaszt), és olyasvalami is, aminek akaratán kívül van kitéve (pl. császármetszéssel született). Érdemes végignézni az összes előbbi példát, csakugyan minden egyikben azonosítható az expozíció, a mobiltelefon-használattól egészen a gyógyszerszedésig.

A másik komponens: minden kérdésben található valamiféle eredmény, kimenet – ezt fogjuk *végpontnak* hívni. Ez lesz az orvosilag lényeges, általunk vizsgált történés; ismét csak, érdemes egy pillanatra visszanézni az előbbi listára, és mindegyik elemnél megkeresni ezt, az agydaganattól az alvászavarig.

Ezzel megvan a séma két „oldala”, expozíció egyfelől, végpont másfelől, már csak egyetlen komponens hiányzik, amit vizsgálni akarunk minden ilyen és ezekhez hasonló kérdésben: az, hogy az expozíció és a végpont között van-e ok-okozati összefüggés. Szép szóval ezt hívhatjuk *kauzalitásnak*. Bármennyire is különbözőnek tűnnek a kérdések, e séma mindegyikre illeszkedik, minden egyikben azonosítható az expozíció, azonosítható a végpont, és mindenre igaz, hogy okozatiságot kutatunk.

De mit is jelent ez a fogalom? Kézenfekvő értelmezés, hogy megnézzük, hogy hány beteg volt a légszennyezéses csoportban, megnézzük, hogy hány lett *volna* köztük, ha minden más változatlan lett volna, *de* nem lett volna légszennyezés – és a kettő különbség a légszennyezés hatása. Sajnos ezt a másodikat soha nem ismerhetjük, hiszen ez egy valójában létre nem jött, képzeletbeli helyzet. Éppen ezért jobb híján a nem légszennyezéses csoporthoz viszonyítunk, tehát nem *ugyanazon* csoport tényleges és képzeletbeli helyzetét vizsgáljuk, hanem két *különböző* csoport tényleges helyzetét viszonyítjuk *egymáshoz*. Lényegében azt mondjuk, hogy a nem légszennyezett környezetben felnövő gyermekek adatai mutatják, hogy mi lett *volna*, ha a légszennyezett környezetben felnövőknél nem lett volna légszennyezés.

Csakhogy ez egy nagyon erős feltevés: kizárálag akkor igaz, ha a két csoport semmi másban nem tér el, egyedül a légszennyezettség tényében.

A téTELmondatot minden esetre megfogalmazhatjuk: Az expozíció akkor van okozati összefüggésben a végponttal, ha a csak az expozícióban eltérő csoportok eltérnek a végpontban, mégpedig olyan mértékben, ami már nem tudható be a véletlen ingadozásnak.

(Az utolsó tagmondat magyarázatára később még vissza fogok térni. Egyelőre gondoljunk arra, hogy ha nem 100 ezer gyermek vett volna részt a kutatásban csak 100, és mondjuk azt találjuk, hogy az egyik csoportban 2, a másikban 3 beteg gyermek van, akkor sokkal óvatosabban mondanánk, hogy különbséget találtunk – noha elvileg itt is másfélszeres az eltérés.)

2.2. A confounding problémája

Ebben a téTELmondatban tehát van egyetlen egy szó, ami iszonyatos bonyodalmakat okoz: az, hogy „csak”. Biztos, hogy az összehasonlított csoportjaink kizárálag csak az összehasonlítás tárgyában, tehát az expozícióban térnek el? Biztos, hogy a légszennyezett területen felnövő és az egészséges levegőben felnövő gyermekek között *csak és kizárálag* az a különbség, hogy milyen a levegőminőség a lakhelyükön...?

Dehogy! A nagyobb légszennyezettségű területek nagyon gyakran épp a városok peremkerületeit jelentik, az ipari negyedeket, az elavult fűtésű, leszakadt részeket. Itt azonban tendenciájában inkább rossz körülmények között élő, kevésbé tehetős családba született (egyszóval: rosszabb szocioökonómiai helyzetű) gyermekek fognak lakni. Igen ám, de a rosszabb szocioökonómiai helyzet egy sor betegség magasabb kockázatát hordozza – mi van, ha a mentális betegségek is ezek közé tartoznak? Mert a rosszabb szocioökonómiai szegmensben a várandósok kevésbé férnek hozzá a szülés előtti gondozáshoz, közülük több dohányzik vagy fogyaszt alkoholt a várandósság alatt. A gyermekkre sem csak a rosszabb levegő hat, hanem a rosszabb táplálkozás, hogy kevésbé vesznek részt szűréseken és így tovább, és így tovább.

InnenTől kezdve, ha találunk is különbséget a mentális megbetegedések előfordulásában a két csoport között, *nem tudhatjuk*, hogy az minek a következménye: az általunk vizsgált levegőminőségbeli eltérésnek, az ezzel – *óhatatlanul!* – *együttjáró* egyéb eltéréseknek, vagy esetleg ezek keverékének?! Ezt nem tudhatjuk – hiszen az összehasonlított csoportok nem *csak* az összehasonlítás szempontjában tértek el.

A problémát tehát az jelenti, hogy a két változó kapcsolatát *megzavarja* egy harmadik változó, mely egyszerre hat az expozícióra és a végpontra, ahogy az 2.1. ábra is mutatja. Ennek következtében elképzelhető, hogy a légszennyezettségnek valójában *nincs is a világon semmilyen* hatása a mentális betegségekre, amit látunk, az egy *látszólagos* hatás, annak következtében, hogy a légszennyezett területen felnövő gyermekek körében egész egyszerűen több a rossz szocioökonómiai helyzetű, ami pedig a *valódi* oka a több mentális betegségnak!

Ha valaki nem hiszi el, hogy ilyen létezhet, akkor nézze meg képzeletbeli adatgyűjtésünk részletesebb eredményeit, melyet a következő táblázat mutat:

	Légszennyezett	Nem légszennyezett	Összesen
Rossz szocioökonómiai helyzet	6% (3300/55000)	6% (375/6250)	6% (3675/61250)

	Légszennyezett	Nem légszennyezett	Összesen
Jó szocioökonómiai helyzet	1% (450/45000)	1% (935/93750)	1% (1385/138750)
Összesen	3,8% (3750/100000)	1,3% (1310/100000)	2,5% (5060/200000)

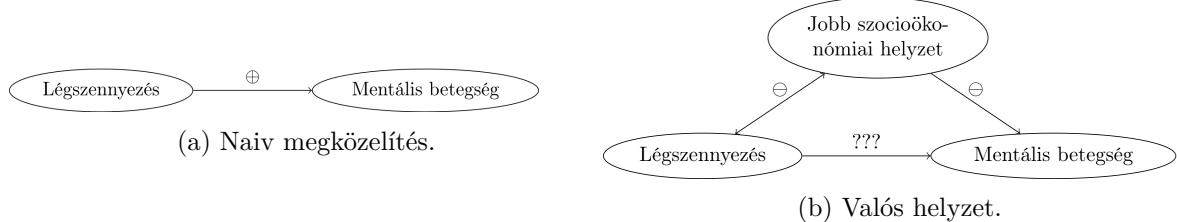
Ebben a táblázatban valami első ránézésre egészen paradox dolg látható. (Azért írtam oda nem csak a százalékokat, de a számokat is, mert néhányan azt szokták mondani, hogy ez matematikailag is lehetetlen. Erről szó nincs, ha valaki nem hiszi, adja össze és ossza el a feltüntetett számokat!) Mert mit látunk? Azt, hogy a rossz szocioökonómiai helyzetű gyermekek körében *nincs* hatása a légszennyezésnek (így is, úgy is 6% az előfordulása), a jó szocioökonómiai helyzetű gyermekek körében *szintén nincs* hatása (1% így is, úgy is) – összességében viszont *mégis van!* Hiszen az 1,3% megnőtt 3,8%-ra, ahogy azt a felvezetőben is írt számok mutatják. Ez meg hogy a csudában lehet? – kérdezhetné valaki. Se egyik csoportban nincs hatása, se a másikban, de összességében mégis van!?

Rakjuk most össze, hogy mi is történt itt. A problémát az okozta, hogy volt egy változónk, mely *egyszerre* tudott két dolgot: *egyrészt* összefüggött az expozícióval (nézzük meg, a jó szocioökonómiai helyzetűeknek csak harmada élt légszennyezett területen, a rosszaknak majdnem 90%-a), *másrészt* befolyásolta a végpontot a légszennyezettségtől *függetlenül, önmagában* is (az 1%-os előfordulást 6%-ra emelte). Az ilyen változókat szokás zavaró változónak, vagy – magyarul is gyakrabban használt angol kifejezéssel – confoundernek nevezni; a jelenségeknek magának pedig *confounding* a neve. (Ez egy nagyon találó angol kifejezés, amire sajnos nem honosodott meg hasonlóan frappáns magyar elnevezés. A „confounding” ugyanis szó szerint azt jelenti, hogy „egybemosódás”: a probléma valóban az, hogy az általunk vizsgált eltérés *egybe van mosódva* egy vagy több egyéb eltéréssel.)

Ez az oka annak, hogy a naiv módszer („több-e a mentálisan beteg a magasabb légszennyezettségű területeken?”) csábító mivolta ellenére is *teljesen fals!* Hiába is dolgoztunk empirikusan, és hiába is gyűjtöttünk szisztematikusan adatokat.

Fontos felhívni rá a figyelmet, hogy a „teljesen fals” természetesen nem azt jelenti, hogy az eredményünk akkor valójában azt jelenti, hogy nem okoz mentális betegséget a légszennyezettség – természetesen okozhat, csak ez *nem következik* abból, hogy több a mentálisan beteg a szennyezettebb levegőjű területeken! *Önmagában* ez az együttjárás *nagyon kevéssé bizonyítja* az okozati összefüggést. Azt mondhatjuk, hogy nagyobb légszennyezettség *együtt jár* a több mentális betegséggel, de hogy a nagyobb légszennyezettség *okoz-e* több mentális betegséget, az egy sokkal-sokkal fogósabb kérdés, aminek kapcsán, mint a fentiek is mutatják, roppant óvatosan kell eljárni.

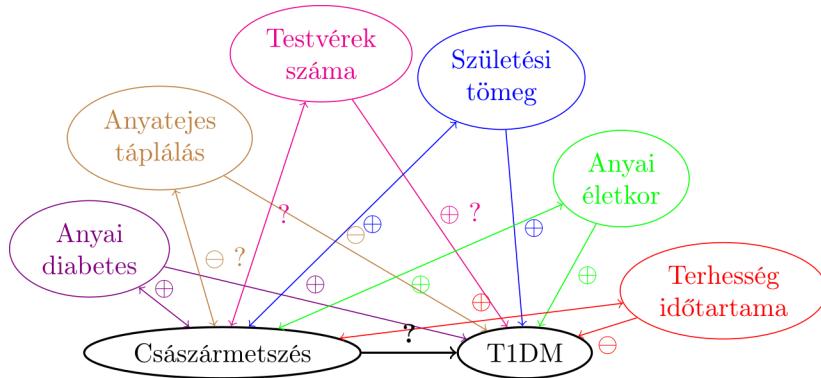
Ez az ilyen jellegű adatok értékelésének egyik legnagyobb problémája (mely a laikus sajtóban is lépten-nyomon visszaköszön). A valóságban ráadásul messze nem olyan egyszerű a helyzet,



2.1. ábra. A változók kapcsolatának megközelítései.

mint a fenti táblázatban, ahol van egy szem confounderünk. A valós helyzetek általában ennél sokkal-sokkal kuszábbak.

Ennek illusztrálására vegyük egy másik példát a cikk elejének listájáról: a császármetszéssel születés megnöveli-e az 1-es típusú cukorbetegség kockázatát? A császármetszéssel születők között több az 1-es típusú cukorbeteg, de – most már tudjuk – ez nem sokat jelent, hiszen mi van, ha vannak egyéb eltérések is a csoportok között? Ez csakúgyan így van: a 2.2. ábra mutatja, immár egy valós orvosi példán, a legfontosabb confoundereket ebben az esetben. Még itt se mondhatjuk persze, hogy ez az összes, de ez már valóságközelibb.



2.2. ábra. Okoz-e 1-es típusú cukorbetegséget (T1DM) a császármetszés? Potenciális confoundererek és hatásaik.

Példának okáért, az anyai diabetes és a császármetszés közötti nyílon pozitív jel van, mert a cukorbeteg anyáknak általában nagyobbak a magzataik, és ez a különféle téraránytalanságok miatt gyakrabban vezet császármetszéshez. Másrészt a cukorbetegségnek van egy erős genetikai komponense, így az anyai cukorbetegségből a gyermekéhez is pozitív nyíl vezet. És már ennyi is elég, hogy bajban legyünk: innentől kezdve, még ha azt is találjuk, hogy a császármetszéssel születettek körében több lesz később cukorbeteg (egyébként tényleg ez a helyzet), akkor sem tudhatjuk, hogy mi a valódi ok: csakúgyan a császármetszés, vagy egyszerűen csak az, hogy a császármetszéssel születőknek gyakrabban cukorbeteg az édesanya? És ez még csak az első confounder volt!

Érdemes megnézni a másodikat, az anyatejes táplálást is. Ez rámutat arra, hogy az expozició és a confounder között nem érdekes, hogy milyen az okozati kapcsolat iránya (az eddigi példákkal szemben itt most aligha arról van szó, hogy az anyatejes táplálás befolyásolja, hogy korábban császármetszés történt-e...), csak az fontos, hogy kapcsolat van. És csakugyan, a császármetszéssel szülő nők ritkábban táplálják anyatejjel a gyermeküket, ez így van a valóságban, és most az mindegy is, hogy ennek mi az oka. Másrészt az anyatejes táplálás – számos egyéb előnye mellett – csökkenti a cukorbetegség kockázatát is – és akkor e ponton megint meg vagyunk lőve... és még közel nem vagyunk a sor végén.

2.3. Zavaró változóktól a megzavart olvasóig

Nem véletlenül írtam korábban, hogy a naiv módszer nagyon „csábító” tud lenni. Képzeljük csak el, pláne némi marketinggel meghintve: látványos grafikon, rajta a nem légszennyezett területen felnövők körében a kockázat (kicsi oszlop), mellett a légszennyezett területek adata (oszlop kiüti az oldal tetejét), szomorú anyuka megrázó beszámolója mentálisan beteg gyermekéről, természetesen hangsúlyozva a levegőminőséget stb. stb. Vajon 100 emberből hánnyal hitetné el a légszennyezettség szerepét...? (És hányan mondanák azt, hogy „hohó, de hát itt óvatosnak kell lenni, mert a szocioökonómiai státuszon keresztül megvalósuló confounding van!”...?)

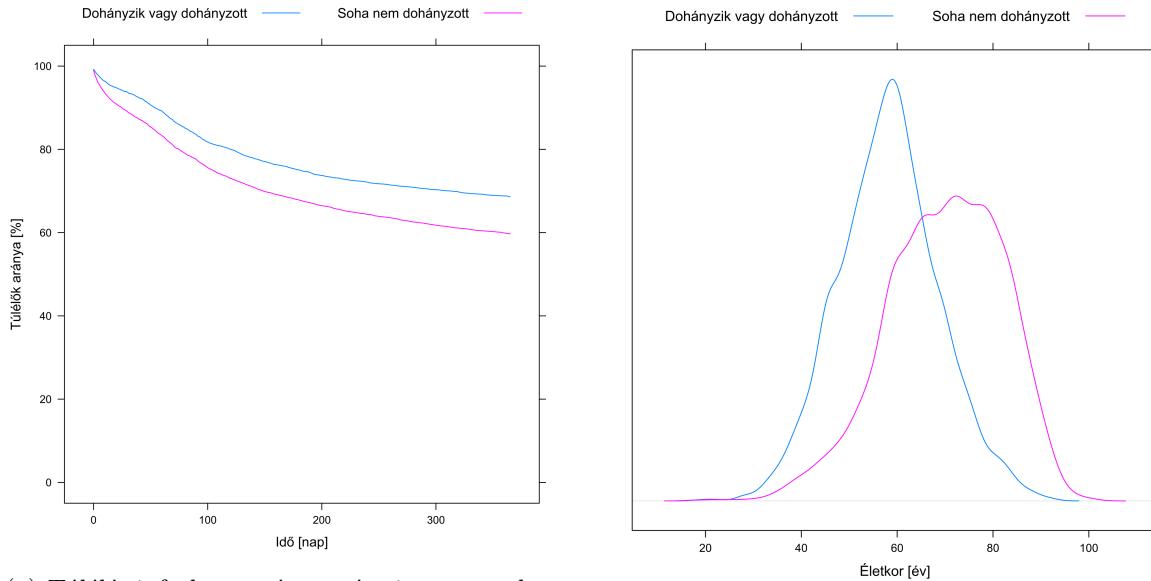
Ha jobban megnézzük, minden nap egészségi megállapításainkat lépten-nyomon átszövi ez a probléma. Nézzük meg kedvenc internetes portálunk egészségi rovatát is...

„A több zöldséget fogyasztók 10 évvel tovább élnek!” Biztos, hogy a több zöldséget fogyasztók csak a zöldségfogyasztás mértékében térnek el a kevesebb zöldséget fogyasztótól? Fogadjuk el, hogy igaz az állítás, és tényleg együtt jár a több zöldségfogyasztás a hosszabb élettartammal. Akkor végül is igaz ez a mondat, minek rajta kötözökön? – kérdezhetné valaki. Szó nincs erről, ez nem akadékoskodás, éppen ellenkezőleg, ez a legfontosabb kérdés. A mondat nyilván azt akarja sugallani, hogy együnk több zöldséget, hogy tovább éljünk. De ha valójában az előző nem okozati kapcsolat, csak együttjárás, akkor – mivel a valódi ok más volt – ezzel *nem megyünk semmire!* Márpedig számunkra ez a fontos: ha egy ilyen cikket olvasva életünket úgy változtatjuk meg, hogy növeljük a zöldségfogyasztásunkat, akkor *várhatjuk-e ettől*, hogy megnő az élettartamunk?

A sor ugyanerre a mintára sajnos igen hosszan folytatható. „Az indiai konyha rengeteg curry-t használ, és lám, velünk szemben ott szinte ismeretlenek a gyulladásos bélbetegségek!” (Biztos, hogy India és Magyarország között csak és kizárolag az az egyetlen különbség, hogy a főzéshez más mennyiségű curry-t használunk?) „Azokban az amerikai államokban, ahol többet alszanak, kevesebb a depressziós!” (Biztos, hogy ezen amerikai államok csak és kizárolag az alvással töltött órák számában térnek el a többitől?) A legszebb, amikor ugyanazt eljátszzuk oda és vissza is: „30 éve még nem használták ilyen széles körben a vérnyomás-csökkentőket, és jóval több is volt a magas vérnyomásos beteg!” (Biztos, hogy 2018 és 1988 között az egyetlen különbség a vérnyomás-csökkentők használatának a mértéke?) „30 éve még nem használták

ilyen széles körben a védőoltásokat, és jóval kevesebb is volt az autista!” (Biztos, hogy 2018 és 1988 között az egyetlen különbség a védőoltások használatának a mértéke?)

A jelenség egészen meglepő helyzetekben is felbukkanhat. A 2.3. ábra mutatja, hogy a magyar Nemzeti Szívinfarktus Regiszter 2013 és 2014 évi adatai – több mint 20 ezer infarktus! – szerint mik a túlélési kilátásai egy infarktuson átesett betegnek a szerint, hogy dohányzik-e vagy sem. Jól látható, hogy a dohányzók görbéje mindvégig a nemdohányzók felett halad, nem is kevessel: az infarktus utáni túlélésben *jót tesz* ha dohányzunk!



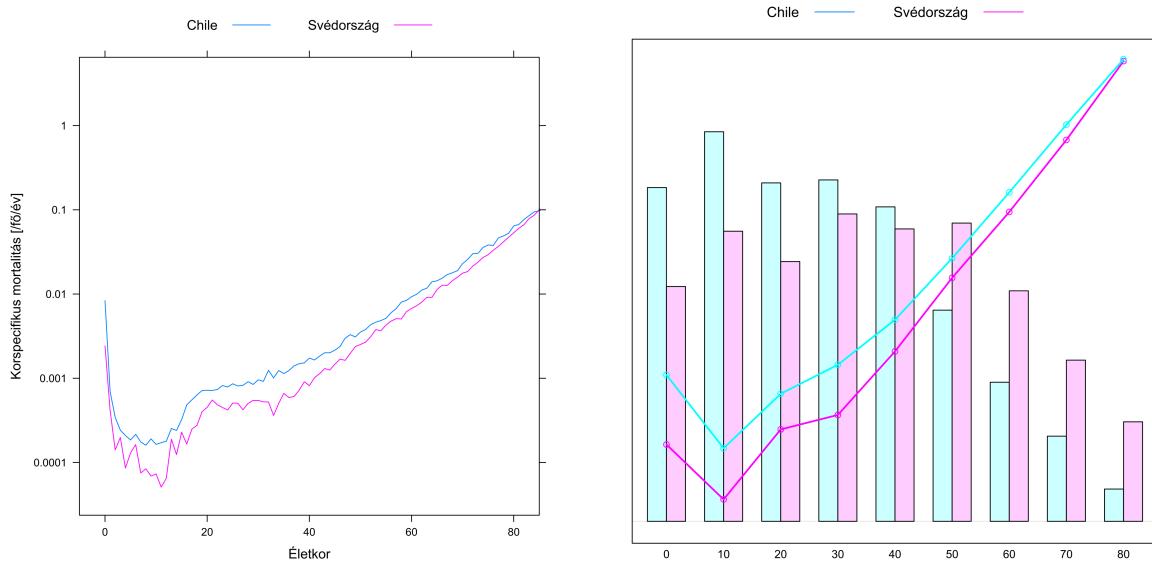
- (a) Túlélés infarktus után: a vízszintes tengelyen az infarktus óta eltelt idő, a függőleges tengelyen az (b) Az infarktust elszenvedett betegek életkorának adott időben még életben lévők aránya látható. eloszlása.

2.3. ábra. Infarktus túlélése és a dohányzás.

Vagy mégsem? Most már egész biztos mindenki rávágja: a dohányosok nem csak a dohányzás tényében térnek el a nem dohányosoktól! De még mennyeire, hogy így van, számos más dologban eltérnek; a számunkra most legfontosabb az életkor: az ábra jobb oldala mutatja – pontosan ugyanazon személyekre vonatkozóan! – az életkor eloszlását dohányzás szerint. Gyönyörűen látható, hogy a dohányosok, egész mellbevágó módon, átlagosan mintegy 15 évvel *korábban* kapnak infarktust! Márpedig az infarktus utáni túlélésben az életkor egy rendkívül fontos tényező, így ez a 15 év bőven kompenzálja a dohányzás – esetleges – negatív hatását. (Esetleges, hiszen innentől kezdve nem tudhatjuk biztosan mi a helyzet: lehet, hogy dohányzásnak tényleg negatív a hatása, ahogy várjuk, de elvileg ettől még az is lehet, hogy nincs hatása, vagy akár lehet továbbra is pozitív, csak nem annyira, mint az az ábrából következne.)

Vagy vegyük a következő esetet! 2005-ben Svédországból 9 millió 10 ezer 729 lakosból 91 ezer 710 halt meg, így ott a halálozási ráta 10,2/ezer fő/év. Ugyanebben az évben Chile-ben 15 millió 519 ezer 347 lakosból 86 ezer 100 halt meg, azaz ott a halálozási ráta 5,5/ezer fő/év.

Micsoda? Svédországban kétszer nagyobb a halandóság, mint Chilében?! A dolgot még furcsábbá teszi, ha megnézzük az ún. korspecifikus mortalitásokat (2.4. ábra), ami azt mutatja, hogy adott életkorban mekkora a halandóság.



- (a) Halandóság adott életkorban Chilében és Svédországban.
(b) Ugyanez 10 életévenként, a háttérben feltüntetve a két ország korfáját is.

2.4. ábra. Svédország és Chile halandósága.

Jól látszik az emberi halandóság nagyon jellegzetes alakja: csecsemőkorban magas, de nagyon gyorsan leeső mortalitás (gyermekbetegségek fázisa), majd egy minimum – már túlesett a gyermekbetegségeken, de még nem kezdődnek a felnőttkoriak – után folyamatos, de lassú növekedés („elhasználódásos” mortalitás fázisa). Ami azonban számunkra most sokkal izgalmasabb, az a megfigyelés, hogy a svéd görbe mindenkorban a chilei görbe *alatt* van! Azaz: minden életkorban jobbak a túlélési esélyek Svédországban!

Most akkor ez hogy is van? minden életkorban jobbak a túlélési esélyek Svédországban, de összességében sokkal rosszabbak?!

Első ránézésre teljesen paradoxnak tűnhet, mi azonban már nem lepődünk meg rajta – nyilván valamilyen confounding van a háttérben! Csakugyan: a két ország között egyéb eltérés is van; ami itt fontos lesz: a korfa. Amint azt az ábra jobb oldala mutatja is, a chilei lakosság sokkal-sokkal fiatalabb összetételeben, mint a svéd, így amikor a teljes halandóságot számoljuk, akkor ott az – előnyös – fiatalkori értékek jóval nagyobb súlyjal fognak latba esni, míg a svédeknek ezeknek kisebb lesz a súlya és a – sokkal rosszabb – időskori értékeknek lesz nagyobb. Így jöhét ki a fenti végeredmény (hiszen azért a chilei 20 éveskori halandóság még mindig jobb, mint a svéd 70 éves – ezen múlik a jelenség). Ez tehát, bármennyire is máshogyan hangzott első ránézésre, megint csak a confounding egy példája!

Most már itt az ideje, hogy átugorjunk a következő, nagyon kézenfekvő kérdésre: akkor mégis mit tegyük, hogy e jelenség fényében is megbízható következtetéseket tudjunk levonni?

3 A confounding megoldásai – megfigyelés és kísérlet

Láttuk, hogy miért jó ötlet szisztematikus empirikus módszerekkel megvizsgálni például azt a kérdést, hogy a légszennyezettség okoz-e mentális betegségeket: begyűjtünk tényadatokat a légszennyezettségről és a mentális megbetegedésekről, mégpedig szisztematikusan, majd ezeket elemezzük. Azt is láttuk, hogy mindenkorban óvatosan kell eljárunk: az, hogy a szennyezettebb területeken több a megbetegedés, még egyáltalán nem jelenti azt, hogy bizonyosan a szennyezés a ludas. De a legfontosabb kérdésre még mindig nem válaszoltunk: most akkor végül is mit tegyünk?

Idézzük fel az empirikus vizsgálatok téTELmondatát: egy vizsgált tényező akkor van okozati összefüggésben a végponttal, ha a kizárálag abban a tényezőben eltérő csoportok eltérnek a végpontban, mégpedig olyan mértékben, ami már nem tudható be a véletlen ingadozásnak. Azt is láttuk, hogy az egyik legalapvetőbb csapda a „csak” szóban van elrejtve: a naiv vizsgálati módszerek, például amikor összehasonlítjuk a szennyezettebb és a kevésbé szennyezett területeken a mentális betegségek előfordulását, nem garantálják, hogy az összehasonlított csoportok *csak* a csoportképzés szempontjában, jelen esetben a légszennyezettségben fognak eltérni. Mi van, ha a szennyezettebb területen élők kevésbé tehetősek, így kevésbé egészségesen táplálkoznak, kevésbé vesznek részt a szülés előtti gondozásban és így tovább. Innentől kezdve, ha találunk is különbséget a mentális betegségek előfordulásában, nem tudhatjuk, hogy az mitől van: a vizsgált különbség (légszennyezettség) miatt, a vizsgált különbséggel *automatikusan* együttjáró egyéb eltérések (táplálkozás, szülés előtti gondozás stb.) miatt, vagy ezek valamilyen keveréke miatt...? Ez volt a confounding problémája.

3.1. Egy aranyérmes megoldás

Mit tudunk tenni? *Törekedni* sokféleképp lehet arra, hogy a csoportok csak a vizsgált tényezőben térjenek el, de *biztosan* elérni csak egyféleképp. Tulajdonképpen az a meglepő, hogy a megoldás milyen későn merült fel. 1931-ben a michigan-i William H. Maybury Tüdőszanatórium orvosa, James Burns Amberson ([3.1. ábra](#)) ki akarta deríteni, mégpedig empirikusan, hogy egy sanocrysin nevű szervetlen aranyvegyület vajon gyógyítja-e a TBC-t (elég sok írás született ennek lehetőségéről akkoriban). Az – *ugybár!* – nem jó megoldás, hogy összehasonlítjuk a gyógyszert kapó és gyógyszert nem kapó betegek gyógyulását, hiszen mi van, ha ők másban is eltérnek a gyógyszerben részesülés tényén túl? Mi van, ha a gyógyszert inkább

kapták a fiatalok (vagy pont, hogy az idősek), inkább kapták a férfiak vagy a nők, inkább kapták a több vagy kevesebb társbetegséggel rendelkezők stb. Ez jelen esetben a legkevésbé sem elméleti spekuláció, nagyon is könnyen lehet, hogy egy új, még nem jól ismert kezelést inkább a jobb állapotú és így egyúttal legjobb gyógyhajlamú betegeknek írnak fel inkább az orvosok. Tehát a gyógyszert kapó és nem kapó csoportok ilyen összehasonlítása teljesen félrevezető lehet – belefutottunk a confounding problémájába.

Amberson és munkatársai egy huszárvágással megoldották a problémát: pénzfeldobással döntötték el, hogy ki kapjon sanocrysint! És ezt most nem irodalmi fordulatként mondjam, hanem a szó szoros értelmében: Amberson *konkrétan* feldobott egy pénzérmet és az alapján adott sanocrysint vagy egyszerű desztillált vizet a betegeknek, hogy fejet vagy írást kapott, ezt pontosan dokumentálta is a cikkében. Még arról is gondoskodott, hogy a két szer külsőleg ne legyen megkülönböztethető, és, hogy a dobás eredményéről ne tudjon a beteg, csak két orvos és a beadó nővér. (Cikksorozatunk későbbi részeiben erre majd azt fogjuk mondani: egyszeresen vak, placebo-kontrollált kutatást hajtott végre.)

És ennyi. Ezzel, a történelemben először, megoldódott a confounding problémája. Majd látni fogjuk, hogy az ismert confounderek kiszűrésére lesz módunk: ha eszünkbe jut, hogy a gyógyszert inkább fiatalabbak, vagy inkább férfiak kapják, és ezért feljegyezzük nem csak gyógyszerben részesülés tényét, hanem azt is, hogy az alany milyen idős és mi a neme, akkor ezeket – mint zavaró tényezőket – ki fogjuk tudni szűrni. De ennek minimális feltétele, hogy eszünkbe jusson, hogy mik a confounderek, és le is tudjuk őket mérni (egy olyannál, mint a „szocioökonómiai státusz” ez utóbbi sem nyilvánvaló). Amberson megoldásában, amit az orvosi irodalomban *randomizáció*nak szokás nevezni, az a zseniális, hogy *minden* confoundert kiszűr, azokat is, amiket nem tudunk feljegyezni, sőt, azokat is, amik eszünkbe sem jutnak! Tegyük fel például, hogy kiderül, hogy a kékszeműeknek az orvosok inkább adnak sanocrysint és a kék szem egyúttal növeli a TBC-ből való gyógyhajlamot. Ez csúnyán tönkretegné az összes vizsgálatot, hiszen ki gondolna arra, hogy a szemszínt is fel kell jegyezni, de vegyük észre, hogy – mert ez a lényeg – Amberson módszere *még ekkor is működik!* Hiszen a pénzfeldobás révén a kékszeműek arányában *sem* lesz szisztematikus különbség a két csoport között! Ugyanúgy, mint ahogy nem lesz szisztematikus különbség a nemi összetételben, az életkorban összetételben, és egyáltalán: *semmilyen* szempontban sem! Úgy is mondhatjuk, hogy a randomizáció kiszűri, ráadásul *automatikusan* kiszűri mind a végtelen számú potenciális confoundert – azokat is, amiket nem tudtunk feljegyezni, sőt, azokat is, amikről *eszünkbe sem jut*, hogy confounderek! Ez a randomizált kutatások hihetetlen nagy előnye.

(Ez a kiszűrés természetesen nem azt jelenti, hogy biztosan minden szempont tökéletesen kiegyensúlyozott lesz a csoportok között. A pénzfeldobás szeszélye folytán *előfordulhat*, hogy pusztá véletlenségből több kékszemű lesz az egyik csoportban, de be lehet látni, hogy mivel ez csak a véletlen szeszélye folytán állt elő, így nem befolyásolja a fenti állításokat.)



3.1. ábra. James Burns Amberson (1890-1979).

3.2. Megfigyelés és kísérlet

Amberson módszerének egy roppant fontos jellemzője van: befolyásolnunk kell hozzá, hogy ki kap gyógyszert (expozíciót). Azokat az orvosi vizsgálatokat, ahol a kutatók aktívan befolyásolják az expozíciót, *kísérletes vizsgálatnak*, azokat, ahol csak passzívan feljegyzik, hogy mi történt, de nem befolyásolják azt, *megfigyeléses vizsgálatnak* szokás nevezni.

Mint korábban is láttuk, a kísérletek története messzire nyúlik vissza. Kísérlet volt az is, amit James Lind végrehajtott a skorbut gyógyításának vizsgálatára – csak épp nem randomizált kísérlet. Az ilyenek problémája az, hogy minden ott van a lehetőség, hogy az orvos, akár teljesen tudattalanul is, de céllányaiban befolyásolja, hogy ki melyik csoportba kerül; például erősen hisz abban, hogy a citrusfélék jót tesznek, ezért, lehet, hogy egyáltalán nem tudatosan, de a legenyhébb eseteket rakja a citrusfélékkel kezelt csoportba (vagy pont fordítva). Ezt már a XIX. század végére felismerték, ezért akkorra divatba jöttek az úgynevezett „váltakozó besorolású” kutatások, ami azt jelentette, hogy minden második beteg kapta meg a vizsgált gyógyszert, minden második nem. Ez már egészen közel van a randomizált vizsgálatokhoz (az csak nem befolyásolja a gyógyulásomat, hogy páratlan sorszámú beteg voltam-e aznap a kórházban!), de valójában még itt is jelentkezhet az előbbi probléma: sokszor leírták például, hogy az orvosoknak meggeszett a szíve egy betegen, ezért igyekeztek úgy rendezni az ellátást, hogy a kezelt csoportba kerülhessen. Ez nyilvánvalóan elrontotta a dolgot, ha mondjuk a legrosszabb állapotú betegeknél került erre a leggyakrabban sor. Éppen ezért a váltakozó besorolás helyét a XX. század közepe felé átvette a randomizált besorolás, különösen, hogy a híres statisztikus Ronald Fisher ennek az elméletét is kidolgozta (egyébként már Amberson orvosi alkalmazása előtt).

Látható tehát, hogy a kísérletes vizsgálatok hihetetlenül nagy és roppant fontos előnye, hogy elvileg mentesek tudnak lenni a confoundingtól. (Gyakorlatilag persze nem feltétlenül: kísérletet is lehet rosszul csinálni – erről később még sok szó lesz.) A megfigyeléses vizsgálatoknál viszont, bármennyire is óvatosan járunk el, minden a fejünk fölött fog Damoklész kardjaként lebegni a confounding: biztos, hogy minden tényező, amiben az összehasonlított csoportok eltérnek – az összehasonlítás tárgyán kívül – eszünkbe jutott? Biztos, hogy mindegyiket le tudjuk mérni? Biztos, hogy mindegyiket jól ki tudjuk szűrni?

Mindezeket látva adja magát a kérdés: akkor miért nem csinálunk minden kísérletet?

Erre a kérdésre vannak nyilvánvaló és kevésbé nyilvánvaló válaszok. A legnyilvánvalóbb, hogy bizonyos helyzetekben egyszerűen lehetetlen: valószínűleg apróbb nehézségeink támadnának a kutatásetikai bizottság előtt egy olyan kutatási tervvel, amelyben szülőnőket randomizáltan akarunk „császármetszni” – függetlenül attól, hogy szükségük van-e rá – azért, hogy kiderítünk, hogy a császármetszés okoz-e cukorbetegséget (pedig, módszertani szempontból ez lenne a legjobb!). Hasonlóan nehéz embereket randomizáltan légszennyezett és kevésbé légszennyezett területen „lakatni”, csak hogy visszatérjünk az eredeti példánakra. Ilyen esetekben mindenki a maradnak a megfigyeléses vizsgálatok, azok minden bajával együtt is.

Az érdekes az, hogy néha akkor is csinálunk megfigyeléses vizsgálatot, ha lehetne kísérletet is (vagy akár ténylegesen végeztek is kísérlet). Ez is mutatja, hogy a kísérleteknek más hátrányaik is vannak, túl azon, hogy drágák, idő- és szervezésigényesek.

Az egyik probléma, hogy a kísérletekben, épp az említett szervezésigény miatt, korlátozott a bevonható betegek köre. A néhány ezer fős kísérlet a legtöbb területen már nagynak számít, a néhány tízezer fő pedig már nagyon nagynak, egy ennél is nagyobb kísérletet pedig csak extrém nehezen lehet megszervezni. (Ebből adódóan nagyon kevés ilyenre van példa. Az utóbbi idők legnagyobb orvosi kísérlete, melyben minden egyes alany egyénileg randomizálásra került, a CAPITA kutatás volt, melyben azt vizsgálták, hogy egy pneumococcus elleni oltás tényleg csökkenti-e a pneumococcus okozta tüdőgyulladások előfordulását 65 év felett. Elképesztő számú alanyt, 85 ezer főt vontak be, ehhez két év és 101 központ kellett, megszámlálhatatlan közreműködővel; sejthetőleg százmillió dolláros nagyságrendbe került ez az egyetlen kísérlet.) Hogy ez miért fontos? Azért, mert a nem elegendően nagy mintanagyság korlátozza, hogy milyen nagyságú hatást tudunk észrevenni, legyen szó akár kívánt hatásról, akár mellékhatásról, ha például egy gyógyszerről beszélünk. Ha kicsi a mintanagyság, akkor egy kis javulást, vagy egy ritkán jelentkező mellékhatást nincs sok esélyünk észrevenni. Pontosan az előbbi a magyrázat a CAPITA esetére is: a pneumococcus okozta tüdőgyulladás nem fordul elő sűrűn, így az oltás, legyen bármilyen hatásos is, „darabra” csak kevessel tudja csökkenteni a tüdőgyulladások számát. És csakugyan: még a 85 ezer alany is csak arra volt elég, hogy összesen kevesebb, mint 200 – a vizsgálat szempontjából fontos típusú – tüdőgyulladás előforduljon. De ugyanez a helyzet a mellékhatások terén is: ha egy mellékhatás csak minden 10 ezredik embert érinti, akkor minden matematikai indoklás nélkül is érezhető, hogy egy 5 ezer fős kutatásban esélyünk sem lesz észrevenni (pedig ez egyáltalán nem kis kísérlet!). Megfigyeléses vizsgálatokkal ezzel szemben összehasonlíthatatlanul könnyebben elérhető ilyen, vagy akár ennél is nagyobb mintanagyság. Gondoljunk arra, hogy a megfigyeléses vizsgálat sok esetben úgy néz ki, hogy adatbázisokból kérdezünk le alanyainkra vonatkozó információkat – itt a kutatás tehát nem azt jelenti, hogy *fizikailag* alanyakat kell kezelnünk, hanem azt, hogy a számítógép előtt ücsörögve lekérdezéseket kell írogatnunk. A kettő bonyolultságát egy napon nem lehet említeni...! Én magam is – harmincéves adjunktusként, 2 kutatótársammal – részt vettet olyan vizsgálatban, melyben néhány hónap alatt, és nulla finanszírozással, 400 ezer magyar beteg adatait dolgoztuk fel – a CAPITA esetében kutatók és segéderők ezreire és évekre volt szükség, meg mellesleg annyi pénzre, mint a Semmelweis Egyetem éves költségvetése, hogy 85 ezer alanyt össze tudjanak szedni...

A másik, előbbihez hasonló gyökerű probléma a kísérletekkel, hogy abban is korlátozottak, hogy mennyi ideig lehetséges az alanyak utánkövetése. A gyakorlatban néhány hónap vagy legfeljebb néhány év érhető el (de az alanyak kihullása a vizsgálatból – nem megy el a következő vizitre, mert elfelejti, elköltözik, elveszti az érdeklődését stb. – már ekkor is általában igen nagy probléma). Ennél hosszabb kísérlet lényegében kivitelezhetetlen, vagy csak a legelemeibb adatok (például: életben van-e egyáltalán még az alany) gyűjthetők be. Világos, hogy ez miért gond: amíg a kevés alany azt limitálja, hogy milyen nagyságú hatást tudunk észrevenni, addig a rövid utánkövetés azt korlátozza be, hogy mennyi idő alatt kialakuló hatást – legyen az akár kívánt hatás, akár mellékhatás – tudunk észrevenni. Szinte esélytelen, példának okáért,

kísérlettel eldöntení, hogy egy gyerekkor táplálkozási szokás vagy orvosi beavatkozás okozhat-e egy tipikusan időskorban, vagy akár felnőttkorban jelentkező betegséget. De itt is elmondható: megfigyeléses vizsgálatokkal nem feltétlenül reménytelen a helyzet, hiszen adatbázisokból sokszor akár több évtizedes átfogású adatok is könnyen kigyűjthetőek.

A harmadik lehetséges probléma a kísérletekkel, hogy a kísérletben részt vevő alanyok – még a legjóhiszeműbb tervezés esetén is – szükségképp egy elég speciális, „steril” populációt jelente-nek, már pusztán abból is adódóan, hogy hogyan verbuválják ezeket az alanyokat. Ez minden felveti azt a kérdést, hogy találunk bármit is a kísérlet alanyai körében, az vajon mennyire vonatkoztható az összes alanyra...? Megfigyeléses vizsgálatoknál ez a probléma sokkal kevésbé jelentkezik: gyakran akár az összes alany is bevonható a vizsgálatba, így aztán egész biztos nincs probléma az összes alanyra vonatkoztatással. E kérdésre később még visszatérünk.

3.3. A jó, a rossz, és a közepesnél némileg gyengébben jó

Összességében véve tehát a legfontosabb megállapítás, hogy nem lehet olyat mondani, hogy a kísérlet és a megfigyelés közül az egyik „jó”, a másik meg „rossz”. Mindkettőnek jellemző előnyei és hátrányai vannak, így az, hogy melyik a szerencsés választás, mindenkor kérdéstől függ: van ahol az egyik, van ahol a másik, a kérdés az, hogy az adott problémának mik a jellemzői. Az előbbi pontban mondottakat szem előtt tartva nagy vonalakban már mi is tudunk választani!

A „nincs jó meg rossz” a fentinél általánosabban is igaz. minden kutatásnak vannak hibaforrásai. Egy ilyet már láttunk is, a confoundingot, a későbbiekben még többet meg fogunk ismerni. Bizonyos kutatásokban több hibaforrás van, vagy komolyabb súlyúak vannak, másokban kevesebb. Van egy szó, amit nagyon szeretek erre: a *bizonyítóerő*. Kifejezi, hogy a tanulmányok – ilyen értelemben vett – értéke nem bináris, mint azt néhányan hajlamosak gondolni: nagyon ritkán van olyan, hogy egy kutatás „tökéletes” (és így ami abban olvasható, az úgy van és pont) vagy, hogy „teljesen hasznavezetetlen” (ezért bármi is olvasható benne, semmit nem jelent). A valóságban ez egy folytonos skála: arról, hogy a szennyezettebb területeken több mentálisan beteg gyermek él sem mondható, hogy semmit sem jelent (a confounding miatt) – csak épp borzasztóan alacsony a bizonyítóereje (arra nézve, hogy a légszennyezettség mentális betegséget okoz).

3.4. Jót tesz-e repülőgépből való kiesésnél, ha van nálunk ejtőernyő?

Valójában tehát nincs éles határvonal kísérletes és megfigyeléses bizonyíték között; minden kutatást a saját erényei és korlátai alapján kell értékelni. Ezt legékesebben az bizonyítja, hogy a különböző bizonyítékok „egy ligában játszanak”, már olyan értelemben, hogy lehet, hogy az általánosságban gyengébbnek tekintett bizonyítékok – például megfigyeléses vizsgálatok –

képesek lehetnek kiváltani a kísérletes bizonyítékokat. Kipróbálta-e bárki, hogy vakbélgyulladásban a vakbélműtét hatásos beavatkozás a semmittevéshez képest? Meglepődnék... Pedig borzasztó egyszerű volna! Csak fogni kellene 200 vakbélgyulladásos beteget, véletlenszerűen 100-at megműteni, és megvární, amíg 99 gyógyultan hazamegy (nem 100-at mondtam, mert legyen a műtétnek is valamicske kockázata), 100-zal nem csinálni semmit, és megvární, amíg 99 is az intenzív osztályra kerül perforált vakbéllel (nem 100-at mondtam, mert azért spontán is lehessen meggyógyulni), és voila, meg is van az igen magas bizonyítóerejű bizonyítékunk a vakbélműtét hatásosságára! Egész érthetetlen módon nem tudok róla, hogy ezt bárki megcsinálta volna... Vagy mondjuk kipróbálta-e bárki randomizált kísérletben, hogy ha nagy magasságban kiesünk egy repülőgépből, akkor jót tesz-e, ha van nálunk ejtőernyő?

Bocsánat, ez utóbbi kérdésre lehet pontos választ adni: Smith és szerzőtársa 2003-as cikkükben – a neves orvosi folyóirat, a British Medical Journal karácsonyi különszámban jelent meg – nagyon alapos irodalomkutatást végeztek a témaban. Pontosan definiálták az expozíciót (ejtőernyővel rendelkezés szabadesés esetén) és a végpontot (halál, vagy komoly trauma – a traumatológiában általánosan használatos ISS séreléssúlyossági pontszám 15-nél nagyobb – fellépése a földbecsapódáskor), rendkívül átfogó, több adatbázisra kiterjedő, pontosan dokumentált irodalomkeresést végeztek, majd arra a megdöbbentő eredményre jutottak, hogy elköpesztő módon egyetlen egy vizsgálat sem volt, melyben embereket repülőgépből dobáltak volna ki, randomizáltan ellátva őket ejtőernyővel és vizsgálva a végpontot! Azaz, mondják a szerzők – nyilván a kísérletek mindenekfelellett mivoltát hirdetőön gúnyolódva – igazából nem tudhatjuk, hogy jót tesz-e, ha van nálunk ejtőernyő, ha kiesünk egy repülőgépből...

A másik dolog, amit mindig észben kell tartani: ha el kell döntenünk egy kérdést, akkor – természetesen – az összes rendelkezésre álló bizonyítékot fel kell használnunk. A második kifejezés, amit nagyon szeretek: *a „bizonyítékok összessége” szemlélet*. Nem lehet kiragadni egy konkrét kutatást, különösen, ha rengeteg készült a számunkra érdekes kérdés vizsgálatára. Márpedig egy sor ilyen téma kör van; ezekben az esetekben az, hogy *egy konkrét* kutatás mit talált, nem sokat jelent. Szoktam mondani, hogy számos kérdés esetében, ha kapok öt percet és egy számítógépet internetkapcsolattal, akkor *legalább egy* kutatást minden állításra és az ellenkezőjére is találok... El kell tehát felejteni az olyan szalagcímeket, hogy „A legújabb kutatás bizonyította, hogy” – nem az az érdekes, hogy a legújabb mit bizonyított, hanem az, hogy összességében mit bizonyítanak a kutatások! Hasonlóan félrevezetések alapjai lehetnek az olyan mondatok – noha elsőre nagyon tudományosnak látszódnak! – miszerint „ez tehát ilyen hatást okoz [Doe, 2016]” (különösen laikusok megtévesztésére alkalmas ez, akik hajlamosak azt gondolni, hogy mivel ez egy ilyen komolyan kinéző, tudományos hivatkozással ellátott állítás, akkor így kell legyen – ha egyszer itt az alátámasztó kutatás...) Valójában azonban ez nem sokat jelent, még ha Doe tényleg ezt is találta, azonban 20 másik kutatás meg az ellenkezőjét.

Láttuk tehát, hogy ha a kérdésünk vizsgálatára tudunk kísérletet végezni (azaz tudjuk aktívan befolyásolni az expozíciót), akkor jó helyzetben vagyunk, mert tudunk randomizálni, és innentől elég egyenes az út: a csoportok közti különbség ez esetben tényleg biztosan a vizsgált expozícióknak – és ezen felül legfeljebb a véletlen ingadozásnak, amit később meg fogunk beszélni – tudható be. Néha azonban ez nem célszerű, vagy lehetetlen – mint épp a légszennyezéses

példánkban is. Erről eddig annyit mondtunk, hogy ez esetben, ha eszünkbe jut, hogy mik a confounderek, és le tudjuk őket mérni, akkor valamilyen módon ki lehet szűrni a hatásukat – hiába megfigyelésesek az adataink. De mégis hogyan? Mi ez a „valamilyen mód”...?

4 Megfigyeléses vizsgálatok a gyakorlatban

Fontos előrebocsátani, hogy ez egy olyan komplex téma, így a mostani cél a legkevésbé sem a módszerek teljeskörű bemutatása. A cél ezzel szemben annak szemléltetése, hogy egyáltalán *van* ilyen módszer – mert ez sem nyilvánvaló! Mégis, hogyan tudjuk megmondani, hogy mi a légszennyezettség valódi hatása a mentális betegségekre, ha egyszer a légszennyezettsébb területen élők nem csak a légszennyezettségnek kitettségen térnek el a többiektől?

Nagyon fontos megismételni, hogy e módszerek, és egyáltalán, bármilyen szóba jövő módszer *kizárálag* azon zavaró változók hatását tudja kiszűrni, amikről egyáltalán eszünkbe jut, hogy confounderek, és amiknek az értékét le tudjuk mérni (azaz meg tudjuk határozni, fel tudjuk jegyezni mindegyik alanyra). Ezeket sem biztos, hogy tökéletesen szűrik, de az biztos, hogy amiről nincs információink, pláne amiről eszünkbe sem jutott, hogy confounder, azt *nem* tudjuk kiszűrni – szemben a randomizálással! Hát persze: ez volt a kísérlet hatalmas előnye.

Lássunk tehát néhány ilyen módszert! Az áttekintés olyannyira nem lesz teljeskörű, hogy még a legnépszerűbb módszer, az ún. többváltozós regressziós modellezés sem fog szerepelni benne – a cél ugyanis nem a részletek megismertetése, hanem az, hogy érezhető legyen a „dolog íze”, hogy hogyan lehet egyáltalán ezt a problémát leküzdjeni.

4.1. A rétegzés

Az érdekes az, hogy az egyik legkézenfekvőbb módszert tulajdonképpen már láttuk! Idézzük fel a táblázatot, melyben a képzeletbeli légszennyezettség/mentális betegség kutatásunk eredményeit közöltük a cikksorozat korábbi részében:

	Légszennyezett	Nem légszennyezett	Összesen
Rossz szocioökonómiai helyzet	6% (3300/55000)	6% (375/6250)	6% (3675/61250)
Jó szocioökonómiai helyzet	1% (450/45000)	1% (935/93750)	1% (1385/138750)
Összesen	3,8% (3750/100000)	1,3% (1310/100000)	2,5% (5060/200000)

Látszik, hogy az alsó Összesen sor – a naiv vizsgálat – hibás eredményt szolgáltat: úgy tűnik belőle, hogy a légszennyezettség növeli a mentális betegségek előfordulását: 3,8% az 1,3%-kal szemben. Igen ám, de ha *megbontjuk* a számításokat a confounder szerint, azaz a zavaró

változó minden lehetséges értékére *külön-külön* is elvégezzük az összes számítást, akkor helyes eredményt kapunk: 6% a 6%-kal szemben és 1% az 1%-kal szemben. Így ugyan már nem egy eredményünk van, hanem kettő, de látható belőlük, hogy a légszennyezettségnek nincs valódi hatása.

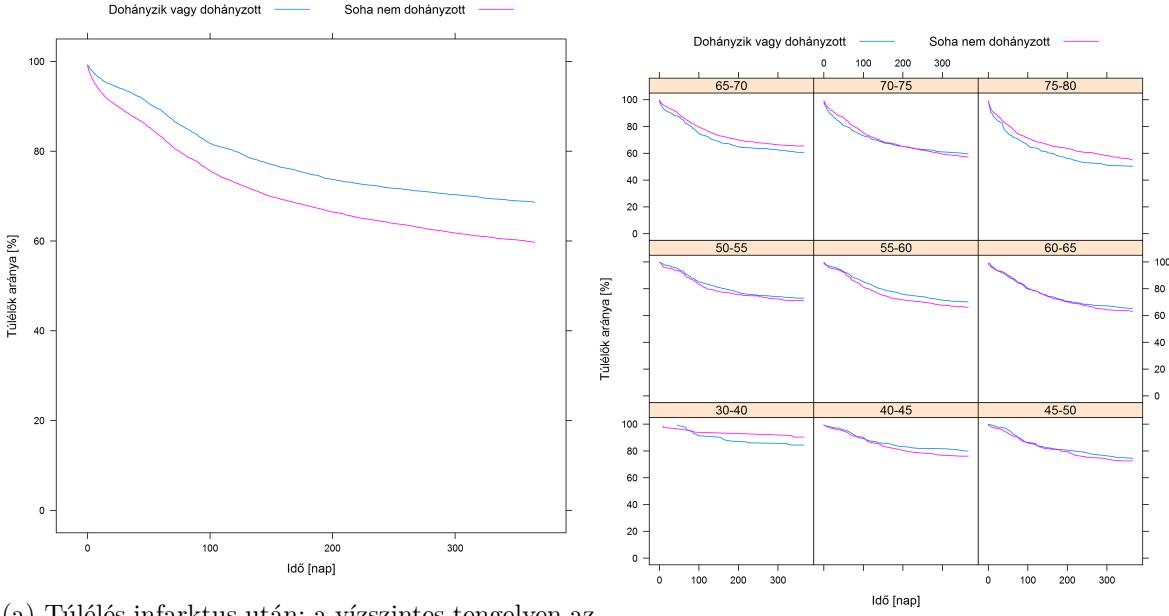
Az orvosi irodalomban ezt a módszert szokás rétegzésnek hívni. (A szakcsargonban ugyanis az ilyen megbontással kapott csoportokat rétegnek hívják: az emberek szocioökonómiai helyzet szerint a „jó” és a „rossz” réteg valamelyikébe tartoznak a mostani példánkban.)

A módszer tehát pofonegyszerű: bontsuk a confounder szerint rétegekre a mintánkat, és egyszerűen annyit tegyük meg, hogy az elemzést nem összességében, hanem rétegenként külön-külön hajtjuk végre.

Sajnos a gyakorlatban ezt sokszor könnyebb mondani, mint megcsinálni. A módszer hatalmas előnye, hogy semmilyen feltételezéssel nem él arra vonatkozólag, hogy az egyes változók milyen kapcsolatban vannak egymással, tehát akármilyenben is vannak, mindenképp működik, de ennek ára van. Az egyik probléma, hogy a valóságban nem egy confounder van: nézzük meg a cukorbetegség és a császármetszéssel születés példáját, nem kevesebb, mint 6 confoundert soroltunk fel. A helyzet valójában ennél is rosszabb: a valóságban ugyanis általában nem tudhatjuk, hogy mik a confounderek! Orvosi megfontolások, korábbi irodalmi adatok stb. alapján tippelnünk kell, azaz valójában inkább *potenciális* confounderekről beszélhetünk, melyek hatását mind szűrni kell, hiszen nem tudhatjuk, hogy közülük mely(ek) az igazi zavaró változó(k). Emiatt vannak olyan vizsgálatok is, ahol akár 10, 20 vagy annál is több confoundert akarunk kiszűrni.

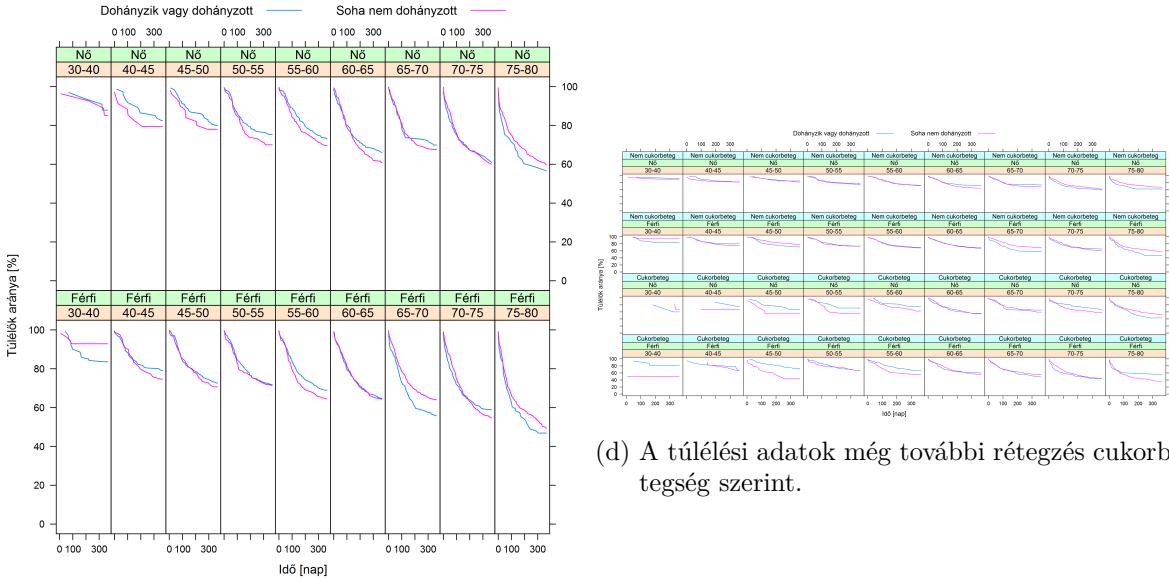
Hogy ez miért probléma? Vegyük a legegyszerűbb esetet, amikor mindegyik zavaró változó két réteget határoz meg. A gond az, hogy ha *egyszerre* kettő szerint rétegzünk, akkor már 4 rétegünk lesz (például férfi – cukorbeteg, férfi – nem cukorbeteg, nő – cukorbeteg, nő – nem cukorbeteg), ha 3 szerint, akkor 8, és így tovább, a rétegek száma mértani haladvány szerint nő, képzelhetjük, hogy mennyi lesz, ha 20 zavaró változó szerint szűrünk... És ez még a legegyszerűbb eset, amikor mindegyiknek két kimenete van, még nagyobb szorzótényezők jönnek be, ha valamelyik kettőnél több értéket vehet fel. Külön probléma a folytonos változók kérdése, tehát azok, amelyeknek nem csak néhány kategória lehet a kimenete, hanem egy szám, mint például az életkor. Itt muszáj csoportosítanunk (pl. 0-10 év, 10-20 év stb.), csak hogy ezt meg lehetetlen jól megcsinálni: ha túl tág intervallumokat veszünk fel, akkor egybemosunk akár nagyon különböző dolgokat (az előbbi csoportosítás esetében egy 1 napos újszülött és egy 9 éves kisiskolás ugyanolyan életkorú!), ha viszont erre tekintettel szűk intervallumokat veszünk fel, akkor rengeteg rétegünk lesz. Mindezeket együtt szemlélteti az [4.1. ábra](#) annak példáján, hogy a dohányzás hogyan hat a szívinfarktus utáni túlélésre.

A probléma igazából kettős. Egyrészt rétegzés esetén nem egyetlen eredményt kapunk, hanem sokat, adott esetben nagyon sokat – ez minden mást félretéve sem túl praktikus, ahogy az előbbi ábra is mutatja, amin jól érzékelhető, hogy a rétegek számának növekedésével egyre áttekinthetetlenebbé válik a helyzet. De a helyzet ennél rosszabb, ez ugyanis nem csak kényelmi problémákat okoz: ha nagyon sok rétegünk van, akkor az egyes rétegekbe egyre kevesebb



(a) Túlélés infarktus után; a vízszintes tengelyen az infarktus óta eltelt idő, a függőleges tengelyen az adott időben még életben lévők aránya látható.

(b) A túlési adatok rétegzése életkor szerint.



(c) A túlélési adatok további rétegzése nem szerint.

(d) A túlélési adatok még további rétegzés cukorbetegség szerint.

4.1. ábra. Az infarktus utáni túlélés, különféleképp rétegzett vizsgálatokban.

alany fog jutni, hiszen ők is szétoszlanak a rétegek között. Nagyon könnyen lehet, hogy még egy egyébként sok alanyt bevonó vizsgálatban is kevés 60 és 70 év közötti, cukorbeteg, nem magasvérnyomásos, korábbi infarktuson átesett, de stroke-on át nem esett, dohányzó nő lesz... Ami azért gond, mert ekkor a rétegenként külön-külön végzett vizsgálatok nagyon kevés alany adatait fogják felhasználni, ami miatt az eredményük bizonytalan lesz! Nagyon sok réteg esetén az egyes rétegekbe egyre kevesebb beteg jut, így azokból csak egyre bizonytalanabban lehet becsülni a túlélési görbéket – ez a magyarázat arra, hogy a jobb alsó ábrán miért néznek ki egyes görbék olyan furcsán: van köztük, amit minden össze 5-10, sőt, 2 vagy 3 betegből kellett megbecsülni (noha az egész adatbázis több ezer beteget tartalmaz). Jól látszik, hogy az életkor milyen problémás önmagában is: 5 évet egybemos ez a felbontás – de még így is 9 kategóriát ad.

4.2. A standardizálás

E problémák közül az első kezelésére van egy egyszerű megoldás, ami ugyan nem tekinthető korszerűnek, de ennek ellenére, jórészt történeti okokból, számos területen használják, így érdemes megismerkedni vele – ez a standardizálás.

Az alapötlet bemutatásához vegyük elő a svéd/chilei halandóság példáját. Emlékeztetőül: Svédországban 2005-ben 10,2 halálozás történt ezer lakosra vonatkoztatva, Chilében 5,5/ezer fő volt a halandóság. Mint láttuk, a meglepő számok magyarázata a confounding: a zavaró tényező az életkor, ugyanis külön-külön vizsgálva minden életkorban kisebb a svéd halandóság (ahogy azt vártuk is!), csak épp nagyon más a két ország korfája: a chilei lakosság összetelében sokkal-sokkal több a fiatal. Így amikor a halandóságot számoljuk, akkor Chilében a fiatalkori, sokkal jobb értékek (mert azért a chilei fiatalok halandósága még mindig jobb, mint a svéd időseké – ezen múlik a dolog) nagyobb súllyal esnek latba, ezért lesz ott alacsonyabb a halálozási ráta. Az előző részben ezt grafikusan láttuk, most nézzük számszerűen (a jobb szélső oszloppal egyelőre ne törődjünk):

Halandóság adott Korcsoporthozban [%/ezer fő/év]		Létszám meg- oszlás (korfa) [%]			
	Svédország	Chile	Svédország	Chile	Segi-Doll standard
0-14	0,28	0,73		17,6	24,2
15-39	0,49	0,90		32,3	39,9
40-44	1,11	1,86		6,8	7,8
45-49	1,83	2,80		6,5	6,6
50-54	3,14	4,27		6,5	5,3

Halandóság adott Korcsoportban [/ezer fő/év]	Létszám meg- oszlás (korfa) [%]			
55-59 5,04	6,86	7,1	4,4	4,0
60-64 8,20	11,33	6,0	3,5	4,0
65-69 13,49	16,86	4,5	2,9	3,0
70-74 21,63	28,92	3,9	2,2	2,0
75- 82,33	76,15	8,8	3,5	2,0

Ha valaki szeretné, ellenőrizheti is a számításokat: $0,176 \cdot 0,28 + 0,323 \cdot 0,49 + \dots + 0,088 \cdot 82,33 = 10,2$, és hasonlóan Chile esetében.

Erre mondhatjuk, hogy ez egy szokásos demográfiai táblázat, de ha ilyen szemmel megnézzük, akkor ez igazából nem más, mint egy confounding megoldása rétegzéssel! A zavaró változó az életkor volt, e szerint rétegeztünk, és ez csakugyan meg is oldotta a problémát: jól látszik, hogy Svédország gyakorlatilag minden korcsoporthoz jobb halandósági mutatókkal rendelkezik.

És gyönyörűen látszik a rétegzés utolsóként említett problémája is: nem egyetlen eredményt kapunk. Milyen kényelmes volt (helytelenül) azt mondani, hogy „Svédország: 10,2, ezzel szemben Chile 5,5”, most ezzel azt tudjuk csak (helyesen) szembeállítani, hogy „0-14 év között Svédország 0,28, ezzel szemben Chile 0,73, 15-39 év között Svédország 0,49, ezzel szemben Chile 0,90, 40-44 év között...” – nem valami praktikus! Mennyivel jobb lenne, ha tudnánk egyetlen, de mégis korrekt számot mondani!

Szerencsére erre a problémára van megoldás. A trükk az lesz, hogy első lépéskor rétegzünk, azaz szétbontunk, csakhogy utána csinálunk egy második lépést, amelynek a keretében újra összetesszük a rétegeket, hogy egyetlen számhoz jussunk eredményként – csakhogy ezt az összerekést okosan hajtjuk végre! Mi volt a probléma, ami a rossz eredményre vezetett? Az, hogy az egyes korosztályokra jellemző halandóságokat különböző korfákkal súlyozva adtuk össze: Chileben a fiatalkori, Svédországban az idősebb életkorbeli halandóságok kaptak nagyobb súlyt. A megoldás kínálkozik: számolunk ugyanúgy, mint eddig, de használjuk *ugyanazt* a korfát mindenkorral! Hiszen a „rossz” számok is felfoghatóak úgy, mint amik ilyen lebontás-összerakás módszerrel jönnek ki, jelesül, ha két különböző korfát használunk; legyen akkor az említett okos összerakás az, ha ugyanazt a korfát használjuk. Például a svéd korfát használva immár *mindkét* országban, a Svédországra kapott halandóság 10,2 (értelemszerűen), a Chile-re kapott halandóság $0,176 \cdot 0,73 + 0,323 \cdot 0,90 + \dots + 0,088 \cdot 76,15 = 10,8$. Hoppá! Máris helyreállít a világ rendje, a chilei érték rosszabb. Persze használhatjuk a chilei korfát is, ekkor Svédországra kapunk 4,9-et – megint csak rendben vagyunk, hiszen Chilére pedig értelemszerűen 5,5 marad az eredmény! Természetesen használhatnánk mindenkorral Magyarország korfáját, Trinidad és Tobago korfáját, Középfölde korfáját, a lényeg, hogy ugyanazt használjuk mindenkorral! Más kérdés, hogy a gyakorlatban ez így nem túl célszerű, hiszen ha

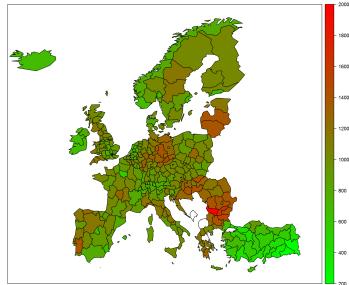
Ha ezzel számolunk, akkor Svédországra $0,31 \cdot 0,28 + 0,37 \cdot 0,49 + \dots + 0,02 \cdot 82,33 = 3,6$ -ot, Chile-re pedig $0,31 \cdot 0,73 + 0,37 \cdot 0,90 + \dots + 0,02 \cdot 76,15 = 4,4$ -et kapunk. Íme, a rossz módszernek megfelelő egyszerűség (két, minden további nélkül összehető szám), de korrekt eredménnyel! Ezt az eljárást szokás direkt standardizálásnak nevezni, az így kapott mutatót pedig standardizált mutatónak, jelen esetben tehát standardizált halálozási rátának. (Ezt fontos hozzátenni, hiszen ezek nem igazi halálozási ráták, lévén, hogy az adott korcsoport halandóságát nem az országban neki megfelelő aránnyal szoroztuk, tehát lényegében egy fiktív helyzetet számoltunk ki. A számértéknek így nincs önálló értelme, csak összehasonlításban értelmezhető.)

Ha megegyezünk, hogy minden ország például a Segi-Doll korfához standardizálva közli a halálozási rátáját, akkor az így kapott eredmények már az – adott esetben rettenetesen különböző – életkori összetételek ellenére is összehetők lesznek! Természetesen az sem kötelező, hogy az összes halálozást számoljuk, nyugodtan nézhetjük csak egy konkrét betegségből előforduló halálozásokat, sőt, vehetjük a megbetegedések fellépését is. A probléma ekkor is ugyanúgy jelentkezne, ha a betegség előfordulása összefügg az életkorral (ami meglehetősen tipikus) és az országok korfái lényegesen eltérnek. Az egyszerűen kiszámolt megbetegedési vagy halálozási ráták – szép néven egyébként ezeket nyers mutatóknak szokták nevezni – nagyon félrevezetőek lehetnek ez esetben, de a standardizált mutatók használata megoldja ezt a problémát.

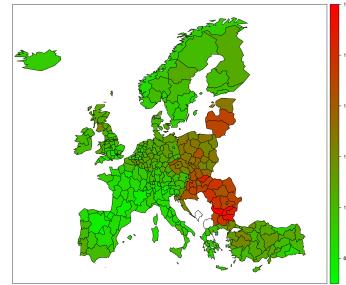
Legalábbis az életkorra vonatkozóan! Mert ne felejtsük el, hogy ezzel csak az életkori eltérések hatását szűrtük ki, semmi mászt. Természetesen, ha nem egy, hanem két szempont szerint tudunk rétegezni, mondjuk életkor és nem szerint is le tudjuk bontani a halálozásokat, a standardizálós trükk ekkor is bevethető, hogy mégis visszajussunk egy – immár nem a és életkori eltérésekre *egyaránt* korrigált! – számhoz, ehhez persze ekkor már a standardnak is életkor és nem szerint bontottnak kell lennie. Ha még több szempont van, akkor a helyzet pontosan ugyanúgy bonyolódik, mint a rétegzésnél láttuk, hiszen a standardizálás első lépése is – természetesen – egy rétegzés.

Példának okáért, az 4.2. ábra mutatja az európai országok halálozási rátáit nyersen és standardizálva. Látszik az egészen drámai hatás: a nyers rátákból azt gondolhatnánk, hogy Törökországban a legjobb a halandósági helyzet, a többi ország között pedig nincsenek túl nagy különbségek. A standardizáltból viszont kiderül, hogy valójában Törökország helyzete egyáltalán nem kiugróan jó (nem halnak meg sokan, de a fiatalabb lakosság *ellenére* sem kifejezetten alacsony a halálozás), a többi ország között pedig nagyon is vannak különbségek (ugyan hasonló arányban halnak meg, csakhogy közben eltérő a korfa!).

Természetesen a probléma nem csak különböző országok összehetésekkel merülhet fel, hanem magasabb szinten (például különböző kontinensek összehetése) vagy alacsonyabb szinten (egy



(a) Nyers rátá.



(b) Standardizált rátá.

4.2. ábra. Halálozási ráták (minden lényeges halálok ból) az Eurostat által gyűjtött országokban NUTS2 szinten. Az adminisztratív határokra vonatkozó térkép forrása: © Eurographics.

ország különböző megyéinek összevetése) is, sőt, akár akkor is, ha ugyanazt a területi egységet vetjük össze, de eltérő időpontokban. A lényeg, hogy ha eltérnek a korfák, akkor bajban leszünk, feltéve, hogy a kor hat a megbetegedési vagy halálozási kockázatra, ezért a standardizálás használata nagyon elterjedt az epidemiológiában. Példának okáért, a 4.3. ábra mutatja a magyarországi vastagbélrákos esetek előfordulásának nyers incidenciáját a Nemzeti Rákregiszter adatai alapján. (Ezekkel az ábrákkal az Olvasó maga is kísérletezhet, például kipróbálhatja más időszakokra vagy más tumor-típusokra a <https://research.physcon.uni-obuda.hu/> címen elérhető Rákregiszter vizualizátor alkalmazás segítségével.)

Jól látható, hogy a nyers ráták szerint meglehetősen drámai romlás mutatkozik 2000 és 2018 között: az 50 körüli érték 60 főlő nőtt, ez több mint 20%-os emelkedés!

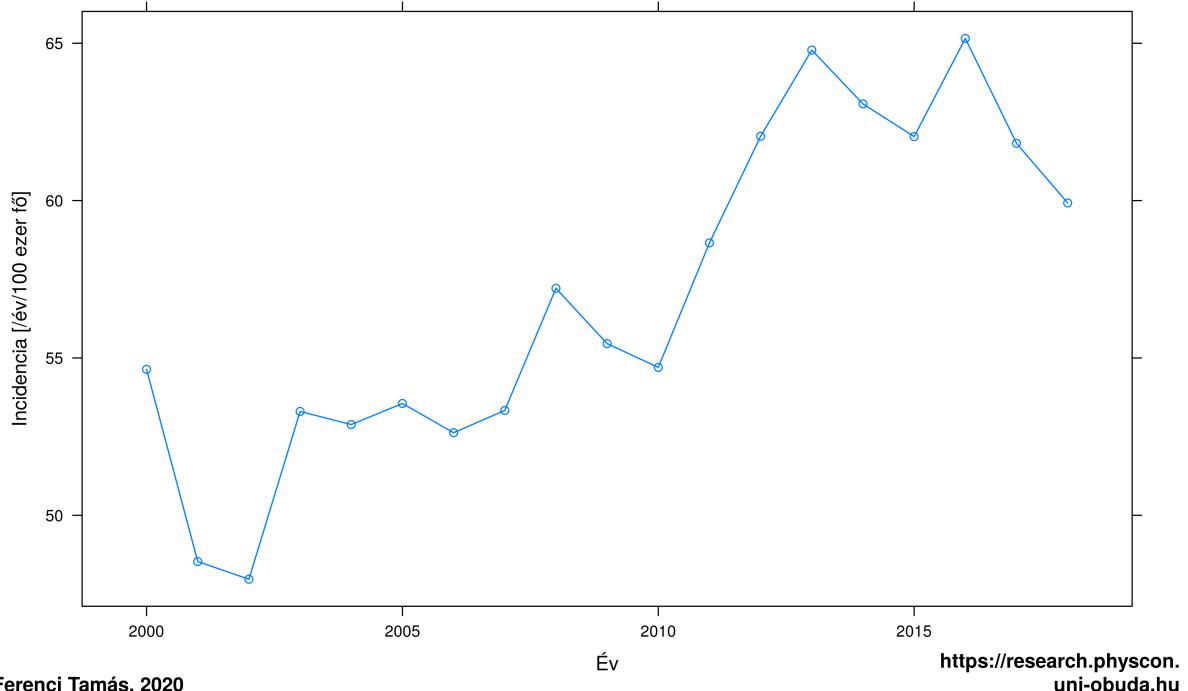
Igen ám, de nézzük meg mit látunk, ha korcsoportonként külön-külön nézzük az előfordulást (4.4. ábra)! Rögtön feltűnik, hogy egyetlen korcsoportban sem volt érdemi romlás.

Ez meg mégis hogyan lehet?! A fentiek alapján már biztos mindenki rávágja a választ: öregedett a népesség! Ezt a Rákregisztertől függetlenül tudjuk (az expozíció, tehát, hogy melyik évben vagyunk, összefügg a zavaró változóval, az életkorral, 4.5. ábra) és a növekvő életkor emeli a vastagbélrák kockázatát (a zavaró változó hat a végpontra, 4.6. ábra). Ne feledjük, a problémához mindenkit tényező egyidejű jelenléte kellett!

Kész is tehát a confounding: valójában a vastagbélrákosok számának növekedése, legalábbis nagy részben, egyszerűen annak tudható be, hogy idősebb lett a népesség! Az 4.7. ábra ezt mutat meg, immár számszerű pontossággal: a standardizált megbetegedési rátákból látható, hogy szó nincs ennyire drámai – vagy egyáltalán, bármilyen érzékelhető – növekedésről.

Bár a standardizálás leggyakrabban a fenti kontextusban, különböző országok megbetegedési és halálozási mutatóinak összevetése kapcsán jön elő, valójában egy teljesen általános megoldási lehetősége a confoundingnak. Térjünk vissza az eredeti példánakra! A probléma, ilyen szemmel ránézve, hogy bár minden csoportban ugyanúgy 6 és 1% a megbetegedési arány minden-

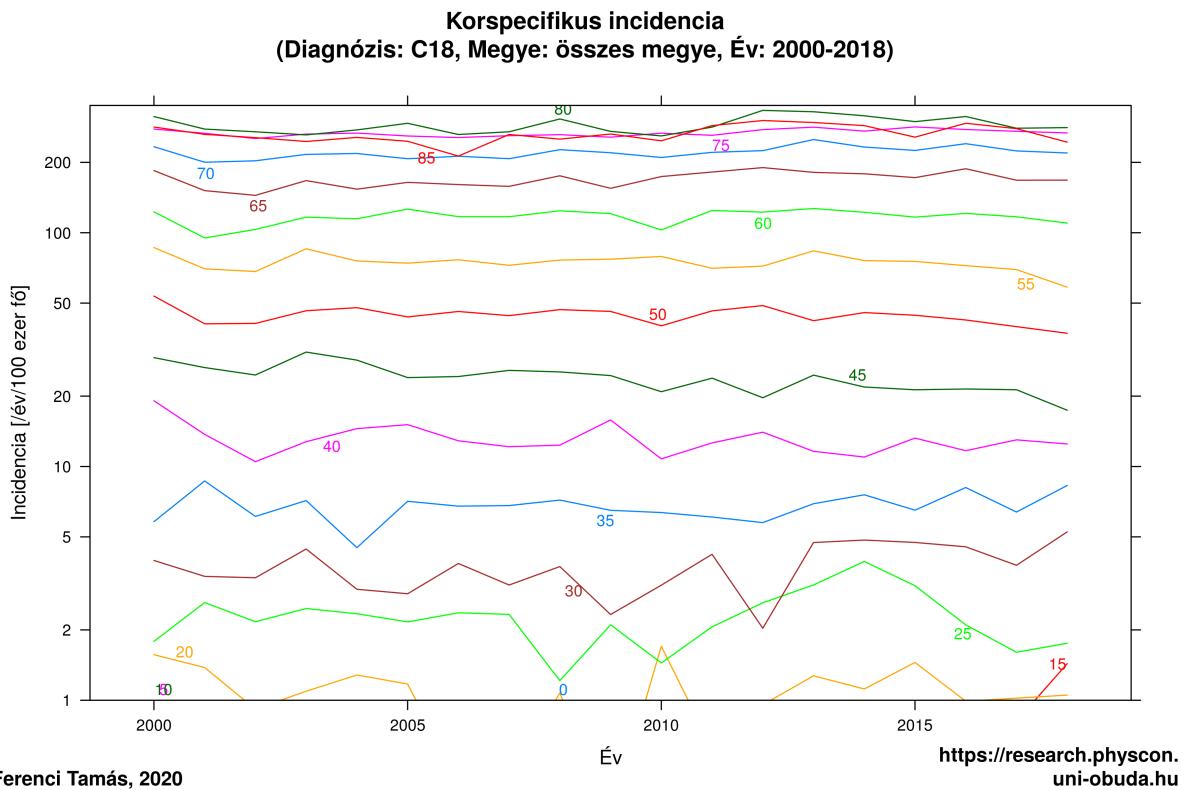
Nyers incidencia alakulása időben
(Diagnózis: C18, Megye: összes megye, Év: 2000-2018)



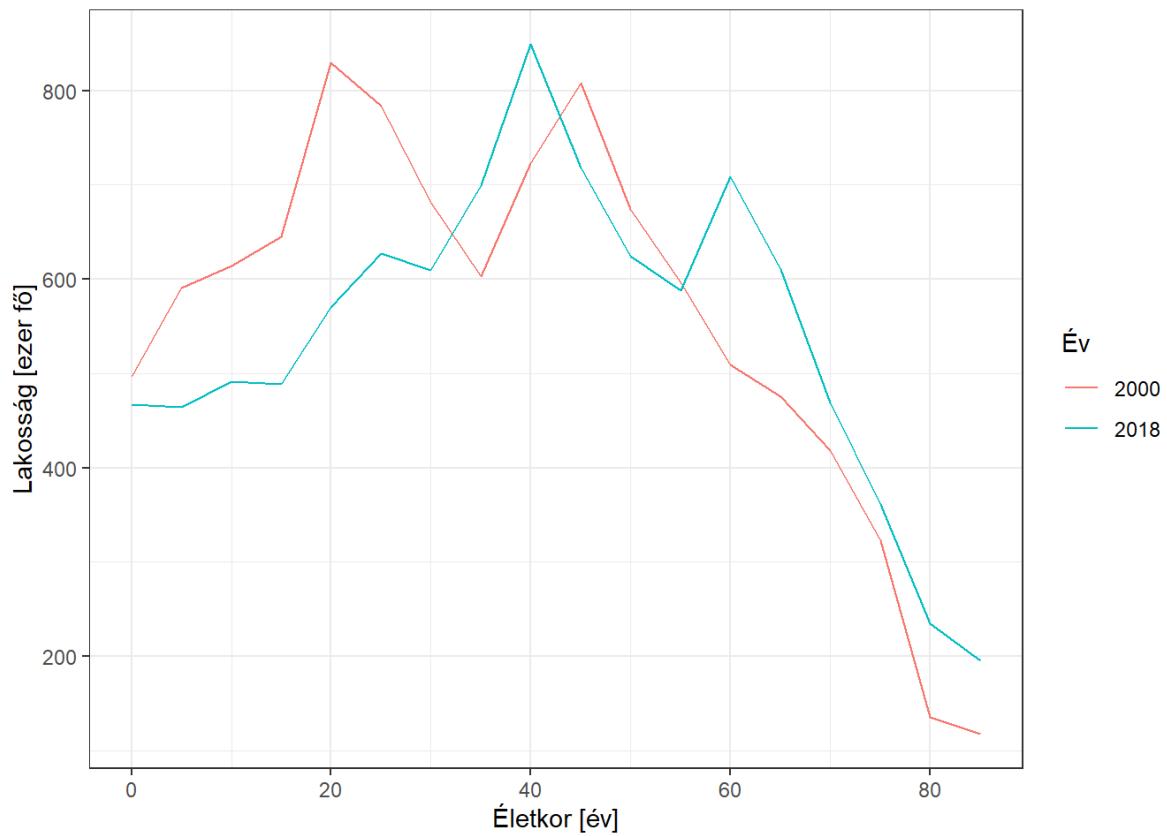
Ferenci Tamás, 2020

<https://research.physcon.uni-obuda.hu>

4.3. ábra. A vastagbélrák nyers megbetegedési rátáinak (eset/100 ezer fő/év) alakulása Magyarországon 2000 és 2018 között.

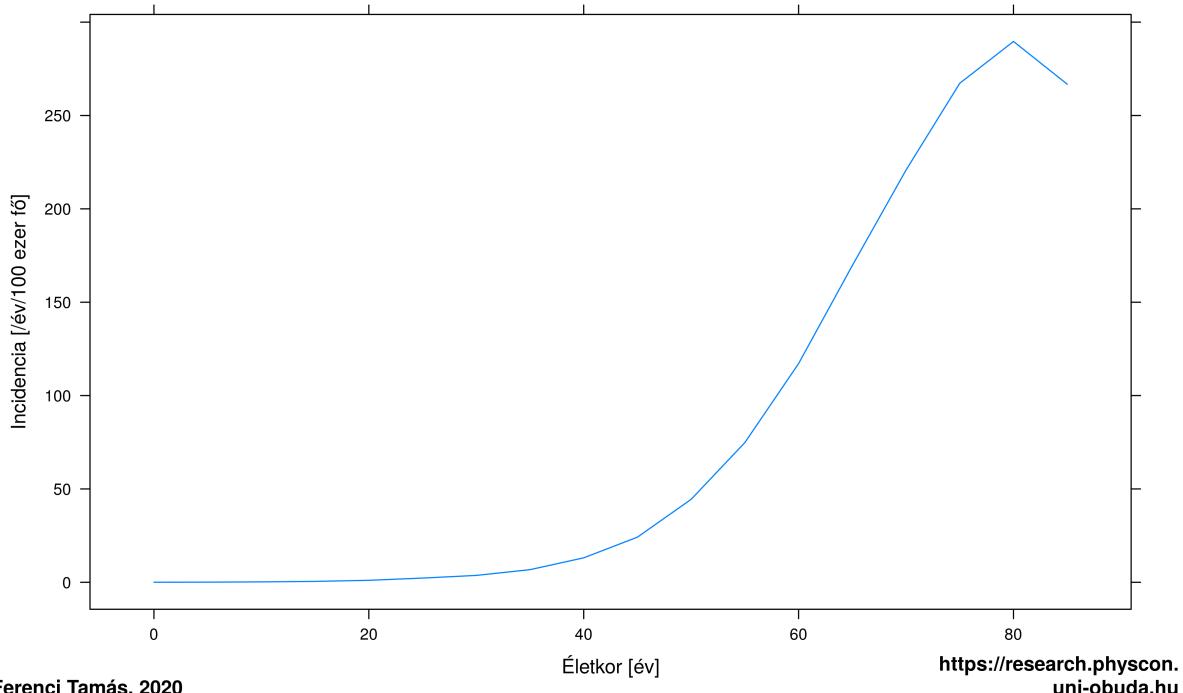


4.4. ábra. Az egyes életkorok nyers vastagbélrák megbetegedési rátáinak alakulása időben (a függőleges tengely logaritmikus beosztású, hogy a nagyon különböző értékek is jól láthatóak legyenek egy ábrán).



4.5. ábra. A magyar társadalom öregedése (életkori megoszlás 2000-ben és 2018-ban).

Korspecifikus incidencia
(Diagnózis: C18, Megye: összes megye, Év: 2000-2018)

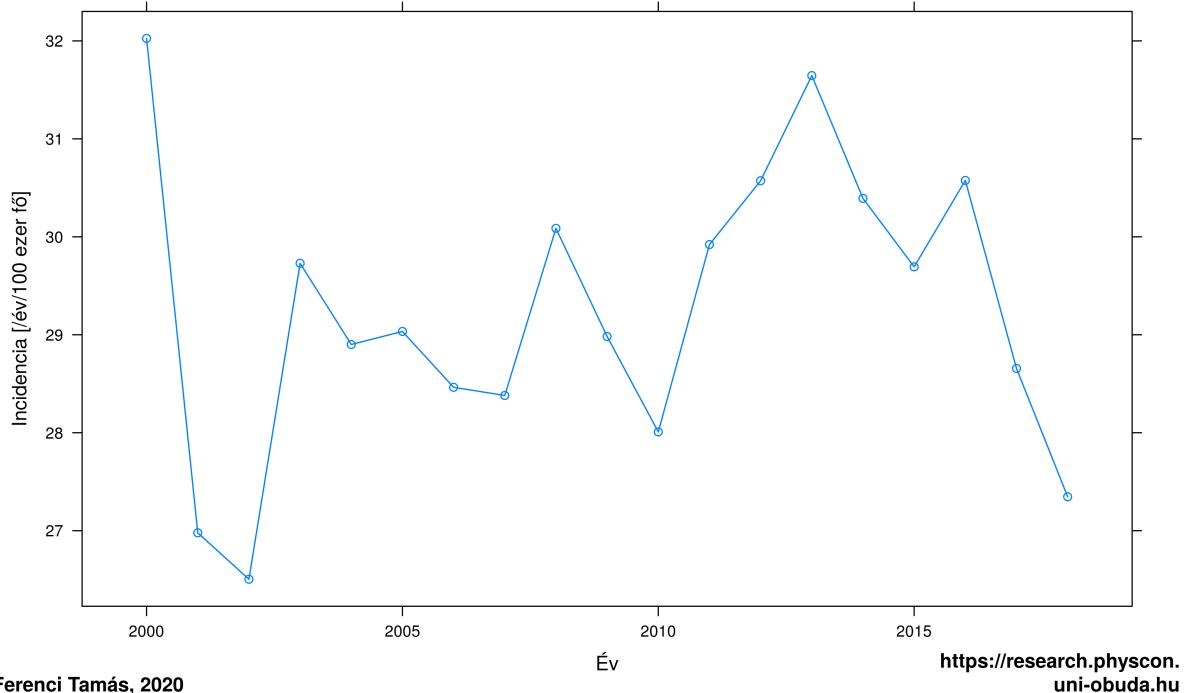


Ferenci Tamás, 2020

[https://research.physcon.
uni-obuda.hu](https://research.physcon.uni-obuda.hu)

4.6. ábra. A vastagbélrák megbetegedési rátájának függése az életkortól.

Standardizált incidencia alakulása időben
(Diagnózis: C18, Megye: összes megye, Év: 2000-2018, Standard: Segi-Doll, 1960)



4.7. ábra. A vastagbélrák standardizált megbetegedési rátájának alakulása 2000 és 2018 között.

rétegekben, a 6 és 1%-ot nagyon eltérő súlyokkal kombináljuk össze: a légszennyezett területen élő csoport esetében a kevésbé tehetősek 6%-os értéke 55% súlyt kap, a tehetősebbek 1%-os száma pedig 45%-ot (ellenőrizzük le: $0,55 \cdot 6 + 0,45 \cdot 1 = 3,8\%$), addig a nem légszennyezett területen felnövő csoport esetében a 6% minden össze 6,25% súlyt kap (itt kevésbé tehetős lakosság él – ugye épp ez volt az egyik oka a confoundingnak), az 1% viszont 93,75%-ot ($0,0625 \cdot 6 + 0,9375 \cdot 1 = 1,3\%$). A probléma tehát itt is az, hogy nagyon más „korfával” (szocioökonómiai helyzet-fával) kombináljuk össze az egyes rétegek értékeit. Hogy egy, de immár korrekt számba tükrítsük a probléma megoldását, egészítük ki a rétegzést standardizálással! Ahelyett, hogy különböző súlyokat használunk, válasszunk egy közöset, legyen ez a Nagy és Alapos Standard IPM módra (röviden: NASI). A NASI-ban a rossz szocioökonómiai csoport aránya legyen 30%, a jóé 70%. Ha ezek után kiszámoljuk a mentális betegségek előfordulását minden két csoportban, de immár egységesen a NASI-t használva, akkor a légszennyezett területen felnövő csoportra $0,3 \cdot 6 + 0,7 \cdot 1 = 2,5\%$ -t, a nem légszennyezettre $0,3 \cdot 6 + 0,7 \cdot 1 = 2,5\%$ -ot kapunk. Ezek lesznek a standardizált mentális megbetegedési ráták a két területen – ezekből már jól látszik, hogy *valójában* nincs különbség! A 3,8%-os és 1,3%-os (nyers) ráták közti különbség teljes egészében a szocioökonómiai helyzeten keresztül megvalósuló confoundingból jött, a légszennyezésnek nincs valóságos hatása a gyermekek mentális megbetegedéseire.

Összefoglalva: a standardizálás legalább azt az egy problémáját megoldja a rétegzésnek, hogy áttekinthetetlen, mert nem egy, hanem sok eredményt ad – azon az áron, hogy adott esetben az így kapott eredmény a standard megválasztásától is függeni fog.

4.3. Kitérő arról, hogy miért jó, ha egy országban sokan halnak meg rákban

Ezen a ponton érdemes kitérni pár kérdésre, melyek ugyan nem kapcsolódnak a szorosan vett témánkhöz, de nagyon logikusan következnek a fentiekből, és meglehetősen tanulságosak – nem is feltétlenül (tisztán) orvosilag.

Az első, hogy a betegségek nyers előfordulási rátája, különösen, ha hosszú időtávot nézünk, vagy nagyon különböző országokat hasonlítunk össze, általában nem túl értelmes mutató. Ezt a fentiekben láttuk, ám érdemes jobban is belegedolni, hogy a véiggondolatlan használata milyen – akár első ránézésre teljesen paradox – helyzetekhez vezethet. Az egyik legjobb példa erre: miért kell néha örülni annak, ha egy országban sokan halnak meg rákban?

Az előbbieket átgondolva adódik a válasz: minden olyan betegség esetén, melynek a kockázata nő az életkor növekedtével (ez számos ráktípusra igaz), a növekvő előfordulásnak nem csak az lehet az oka, hogy romlik a népegészségügyi helyzet, hanem az is, ha tovább élnek a lakosok! Szavaziföldön például szinte ismeretlen a rákos halálozás, de ennek annyira valószínűleg nem örülnek, ugyanis a fő oka nem a remek szváziföldi onkológiai helyzet, hanem egész egyszerűen az a tény, hogy az AIDS viszi el a lakosok majdnem kétharmadát, többségüköt borzasztó fiatalon. (A szülőnők majdnem fele HIV-fertőzött, a válság legsúlyosabb pontján, a 2000-es évek

elején a születéskor várható élettartam 45 év alá esett, majdnem 10%-os csecsemőhalandóság-gal.) Azaz: egyszerűen nem élnek addig az emberek, hogy elkezdjenek rákosak lenni! Hiszen az, vagy legalábbis számos fajtája, idősebb korban jelent igazán kockázatot. A szváziföldi népegészségügy roppant boldog lesz, ha majd azt látja, hogy egyre többen halnak meg rákban – ez ugyanis azt jelenti, hogy sikerült megfélezni a HIV/AIDS-járványt.

A másik doleg, aminek pontosan emiatt nincs túl sok értelme (önmagában) az a haláloki összetétel. Ez alatt azt értik, hogy az elhunytak mekkora hányada halt meg rákban, szívinfarktusban stb., a közbeszédben gyakran hangzanak el olyan kijelentések, hogy „Úristen, a helyzet drámai, a magyar lakosok 25%-a rákban hal meg!”.

Na és? Én személy szerint nagyon örülnék neki, ha a magyar lakosság 100%-a rákban halna meg – feltéve, hogy ezt 130 éves korukban teszik!

A probléma tehát ma Magyarországon nem az, hogy *túl sokan* halnak meg rákban, hanem az, hogy *túl korán* halnak meg rákban!

Elege, a „túl sok” jelzőnek, és ebből adódóan az ilyen politikusi „az a célunk, hogy kevesebben haljanak meg rákban Magyarországon!” jelszavaknak túl sok értelmük nincs: magától érte-tődően, ha kevesebben halnak meg rákban, akkor *automatikusan*, szükségképp *többen* halnak meg infarktusban, agyér-betegségen és egyéb halálokban... (Mindaddig legalábbis, amíg nem sikerül megoldani azt a jelenleg kissé bonyolultnak tűnő orvosi problémát, hogy *valami-ben* mindenki meghal előbb-utóbb.) Ha valaki azt mondja, hogy „az a célunk, hogy minél kevesebben haljanak meg rákban Magyarországon!” az *tökéletesen egyenértékű* azzal, mintha azt mondaná, hogy „az a célunk, hogy minél többen haljanak meg nem rákban Magyarországon!”... A valódi cél tehát nem az, hogy *kevesebben* haljanak meg, hanem az, hogy *később* haljanak meg az emberek (kivéve persze, ha valakinek speciális preferenciája van, ami miatt egy bizonyos halálokban jobban vagy kevésbé jobban szeretne meghalni, mint a többiben).

Ez elvezet minket egy mutatóhoz, ami a „sokan halnak meg rákban” szintű közbeszédben kevés helyet kap, ami azért különösen sajnálatos, mert valójában egyáltalán nem bonyolult módon ragadja meg a fenti problémát. Ez nem más mint az elvesztett potenciális életévek (általánosan használt angol rövidítéssel: YLL) fogalma. A YLL számítása tulajdonképpen pofonegyszerű: kijelölünk egy életkort, ez tipikusan 70 év, és minden egyes halálozásnál megnézzük, hogy az elhunyt mennyivel korábban halt meg – ez lesz az elvesztett potenciális életévek száma az Ő esetében. Egy csecsemőhalálozásnál ez tehát 70 év, ha valaki 60 évesen hal meg, akkor 10, ha 70 évesen, akkor 0, ha 71 évesen (vagy bármikor később), akkor ugyanúgy 0. Ha ezt összeadjuk az összes adott országban adott évben elhunytra, akkor megkapjuk, hogy ott mennyi volt a teljes potenciális életév-veszteség. Ezt tipikusan 100 ezer lakosra leosztva szokás közelíni, Magyarországon 2016-ban az OECD adatai szerint 4806 év/100 ezer fő volt. Ez első ránézésre meglehetősen abszurd mutató („akkor az nem is számít, ha 70 éves korom után halok meg!” – kérdezhetné valaki), ám második ránézésre már látható, hogy valami nagyon fontosat mutat meg. Megtehetjük ugyanis, hogy az elvesztett potenciális életéveket is kiszámoljuk halálokkonként, és ez olyat fog felfedni, amit az egyszerű haláloki összetétel elfedett:

azt, hogy melyik halálok öli *korán* a lakosokat! Magyarországon például a KSH utolsó, 2007-es elemzése szerint a daganatos megbetegedések felelnek a nagy betegségcsoportok közül a legnagyobb potenciális életév-veszteségért, miközben a halálok között csak másodikak a szívérrendszeri halálozások mögött. Még érdekesebb (de ha végiggondoljuk, teljesen logikus!), hogy a szándékos önarálom a halálok között 2%-ot sem ér el, de potenciális életév-veszteségen még az egyébként halálozások 10%-áért felelős agyérbetegségeket is megelőzi! Hogy ez miért fontos? Azért, mert segít végiggondolni, hogy az egészségügy – minden szűkös – erőforrásait hová érdemes csoportosítani: egyáltalán nem biztos, hogy arra racionális, amiben sokan halnak meg!

A kérdésnek persze a valóságban számos, nem könnyen megválaszolható további problémája is van (megelőzhetőség figyelembevétele, költségek figyelembevétele, nem halálos, de életminőséget rontó állapotok figyelembevétele, a különböző életkorú halálozások súlyozása stb.), de itt most a cél az alapgondolás bemutatása volt.

Láttuk tehát, hogy kísérletes adatokból hogyan tudunk okozati viszonyokra következtetni (randomizálással), és most már azt is láttuk, hogy hogyan tudjuk ugyanezt megtenni megfigyeléses adatokkal is. A problémáknak azonban még nem értünk a végére! Van egy kérdés, amiről még egyáltalán nem beszélünk: hogyan vesszük figyelembe a véletlen ingadozás hatását?

5 A véletlen szerepe az orvosi vizsgálatok kiértékelésében

Láttuk, hogy ha kíváncsiak vagyunk arra, hogy egy tényező bír-e valamilyen egészségügyi hatással, akkor végezhetünk kísérletet és megfigyelést, és alaposan körbejártuk azt, hogy ez utóbbi esetében mire kell figyelni. Van azonban egy tényező, ami mind a megfigyeléseket, mind a kísérleteket érinti, de eddig még egyáltalán nem ejtettünk róla szót: a véletlen szerepe.

Emlékezzünk vissza a téTELmondatunkra: Az expozíció akkor van okozati összefüggésben a végponttal, ha a csak az expozícióban eltérő csoportok eltérnek a végpontban, mégpedig olyan mértékben, ami már nem tudható be a véletlen ingadozásnak. Ennek az első felét alaposan kivesztük (különös tekintettel a „csak” szóra!), de a második felével még adósak vagyunk. Egyáltalán, mit értünk e véletlen ingadozás alatt?

A véletlenség problémája abból fakad, hogy a biológiában a dolgok általában nem determinisztikusak. Mi az, hogy „nem determinisztikus”? Az, hogy a gyógyszerrel kezeltek nem fognak 100%-ban meggyógyulni, és hasonlóan, kezelés nélkül sem fog 100% belehalni a betegségebe. Ha ez így lenne, akkor semmiféle problémánk nem lenne a véletlenséggel: ha tudunk kísérletet csinálni, akkor összesen 2 beteg bevonásával biztos választ kaphatnánk a gyógyszer működőképességéről. Az egyik – véletlenszerűen kiválasztott – beteget kezeljük a gyógyszerrel, a másikat nem: ha a kezelt beteg meggyógyul és a nem kezelt meghal, akkor a gyógyszer *biztosan* hatott, minden más esetben a gyógyszer *biztosan* hatástalan (vagy kifejezetten káros). Abban a pillanatban azonban, hogy ez nem így van, gondban leszünk a 2 szem betegünkkel.

Mondjuk, hogy a kísérletünkben mindenki meghal, kijelenthetjük-e ekkor, hogy a gyógyszer hatástalan? Nem! Tegyük fel, hogy kezelés nélkül 99% hal meg, kezeléssel 1%, azaz igen csak hatásos a szer. Ekkor 99% valószínűsége van annak, hogy a nem kezelt beteg meghal *mivel* nem kezeltük, és 1% valószínűsége, hogy a kezelt beteg meghal *annak ellenére*, hogy kezeltük. Hogy kézzelfoghatóvá tegyük ezeket a valószínűségeket, gondolatkísérletként képzeljük el, hogy 10 000-szer elvégezzük ezt a kísérletet. Ekkor várhatóan $10000 \cdot 99\% = 9900$ esetben meghal a nem kezelt beteg. Mivel a két beteg sorsa értelemszerűen teljesen független egymástól, ezért ezen a 9 900 eseten belül *is* 1% lesz azon esetek aránya, amikor a kezelt beteg meghal. Ez $9\ 900 \cdot 1\% = 99$ eset – ennyiszer fordul elő, hogy a nem kezelt és a kezelt beteg egyaránt meghal (miközben a kezelés roppant hatásos volt!). Mivel erre 10 000 esetből került sor, így $99/10\ 000 = 0,99\%$ ennek a valószínűsége. Összefoglalva: hiába volt *majdnem* biztos, hogy kezelés nélkül meghal a beteg, és *majdnem* biztos, hogy kezeléssel nem hal meg, már ez a „majdnem” is elég volt ahhoz, hogy többé ne tudjunk biztos állítást tenni. Mert nem tudunk:

biztosan nem mondhatjuk minden beteg halálát látva, hogy a gyógyszer nem hat! Hiszen láttuk, ha hat, *akkor is* előfordulhat – még ha csak szűk 1% valószínűséggel is –, hogy *mégis* meghal mindenki. A két halálozás tehát nem jelenti *biztosan* azt, hogy nem hatott a gyógyszer. Ha ráadásul nem is ennyire hatásos a kezelés, akkor ez a valószínűség jóval nagyobb is lehet: ha 10% és 90% a két halálozási arány, akkor már 9% valószínűsége van annak, hogy minden beteg meghal (ellenőrizzük a számot!), noha a kezelés továbbra is hatásos.

A dolog fordítva is igaz: ha azt látjuk, hogy a kezelt túlél, a nem kezelt meghalt, mondhatjuk-e, hogy a kezelés biztosan hat? Nem! Ha kezeléssel és kezelés nélküli is 50% a halálozási kockázat (tehát a kezelésnek a világban semmi hatása nincs), akkor is lehet, hogy túlél a kezelt beteg, és meghal a nem kezelt. Az előzőek ismeretében ennek a valószínűségét is könnyedén ki tudjuk számolni: 10 000 esetből várhatóan 5 000-szer fog túlélni a kezelt beteg, ezen belül 2 500-szor fog meghalni a nem kezelt – összességében tehát 25% a valószínűsége, hogy bár a gyógyszer egyáltalán nem hat, mi mégis azt tapasztaljuk, hogy a kezelt alany túlél, a nem kezelt pedig meghalt. Ugyanaz történik mint az előbbi esetben: megint csak nem tehetünk biztos kijelentést! Hiába halt meg a nem kezelt beteg és élt túl a kezelt, ennek hátterében lehet az is, hogy nincs hatása a kezelésnek, tehát megint csak, *biztosan* nem mondhatjuk azt, hogy hatott a gyógyszer.

5.1. A véletlen ingadozás átka

A probléma egy szóval: a véletlen ingadozás. Az a megfogalmazás, hogy „a beteg adott körülmények között 99% valószínűséggel” hal meg, véletlen ingadozást jelent: 100 beteget kitéve ezeknek a körülményeknek a halálozások száma 0 és 100 között bárhol lehet (persze nem ugyanolyan valószínűséggel, például, ahogy érezhető is, a legvalószínűbb a 99 halálozás lesz – de sem a 0 halálozásnak, sem a 100-nak nem nulla a valószínűsége!). Pontosan ugyanúgy, mint ahogy egy szabályos pénzérmet feldobva sem biztos, hogy *pont* 50 fejet és 50 írást kapunk.

Amikor az előbb úgy fogalmaztunk, hogy „hat a gyógyszer *mégis* meghalt minden beteg”, vagy „nem hat a gyógyszer *mégis* túlél a kezelt és meghalt a kezelés nélküli”, akkor bátran hozzátehetünk volna, hogy „...pusztán a véletlen ingadozás szeszélye folytán”. Mert itt csakugyan erről van szó: nincsen semmi ámítás és csalás, kísérletet végeztünk (tehát nem volt confounding), tételezzük azt is fel, hogy a kísérlet minden szempontból tökéletesen volt kivitelezve, tehát semmilyen más hibaforrás sincs – így az iménti kellemetlen eredményeink egyetlen és kizárálagos oka a véletlen ingadozásból fakadó, kiküszöbölhetetlen pechünk. Természetesen a mostani megállapítások ugyanúgy érvényesek megfigyeléses vizsgálatokra is!

Adja magát a kérdés: most mégis mit tehetünk? Sőt, igazából egy ennél erősebb állítás is adja magát: a kissé defetista szemléletűek esetleg ezen a ponton azt mondhatják, hogy akkor kár is bármilyen orvosi vizsgálatot végezni. Soha nem lehet eldönteni, hogy hat-e a gyógyszer, hiszen *bármi* jön is ki eredményként, attól még egyaránt lehet az is, hogy *igazából* hat a gyógyszer, meg az is, hogy *igazából* nem.

A valóságban ennél sokkal jobb a helyzet: igen, *biztos* állítást tényleg nem tehetünk, de *valószínűségi* állítást igen. Bizonytalan állításokat tudunk tenni, és, ami még fontosabb, úgy tudunk tenni, hogy közben a bizonytalanság mértékét *magát* is tudjuk jellemezni. Hibázhattunk – akár akkor, amikor azt mondjuk, hogy hat a gyógyszer, akár akkor, amikor azt mondjuk, hogy nem – de az állításainkat terhelő hibát ismerjük. Az erre szolgáló statisztikai apparátust szokás következtető statisztikának nevezni. De mégis mi ez az apparátus, hogyan tudjuk ezt a helyzetet kezelni?

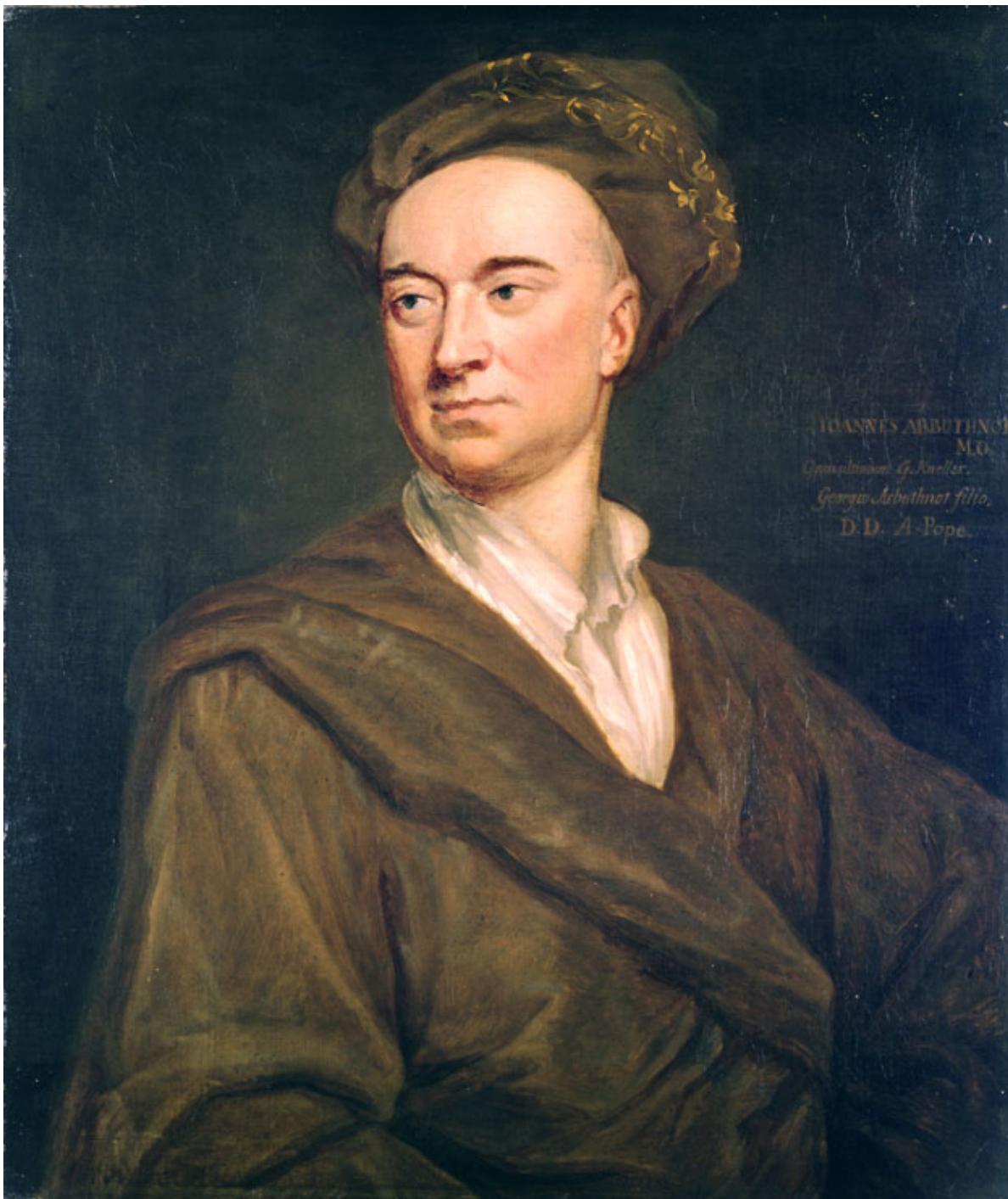
5.2. Isten létenek bizonyítása a statisztika eszközeivel

A meglehetősen kalandos sorsú angol polihisztor, John Arbuthnot ([5.1.](#) ábra), aki – amellett, hogy orvosként praktizált – kora jelentős irodalmi levelezője és politikai szatíraírója volt, félhobbi matematikus, fordító és mellesleg az Angliát megszemélyesítő John Bull figurájának kitalálója, 1711-ben közölt egy esszét „Érv az isteni gondviselés léte mellett a mindenbeli keresztelesekben megfigyelhető állandó szabályosság alapján” címmel.

Arbuthnot az 1629 és 1710 közötti londoni keresztelek adatait dolgozta fel, egész pontosan azt nézte, hogy hány fiút és hány lányt kereszteltek, ez elég jó közelítéssel mutatja, hogy hány fiú és lány született; az adatait a [5.2.](#) ábra mutatja. (Később még lesz róla szó, hogy még a teljesen banális tévesztések, például a szedési hibák szerepét sem szabad alábecsülni, erre ez a táblázat is egy korai példát szolgáltat: nézzük meg közelebbről az 1674-es és az 1704-es évet! A vonaldiagramos ábrázolás egyébként nem Arbuthnot-tól származik, annál is, mert ezt a vizualizációs módot ugyanis csak a 18. század végén öltötte ki William Playfair – szintén skót – matematikus, aki emellett elsőként használt oszlopdiagramot, vonaldiagramot és kördiagramot is adatok vizualizálására.)

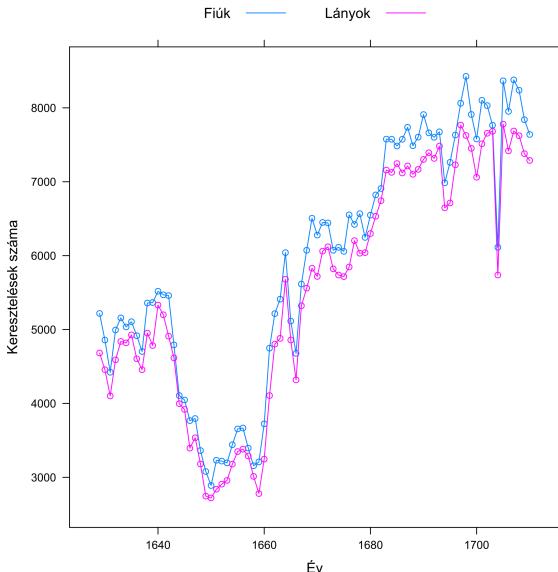
Arbuthnotnak az szúrt szemet, hogy az újszülöttek között mintha több lenne a fiú. Ez meglepő, hiszen első ránézésre 50-50%-os arányt várt volna az ember. Jó, meglepőnek meglepő, mondhatjuk, de hogy jön ide Isten léte?! Úgy, hogy az már Arbuthnot korában is ismert volt, hogy a fiúk halandósága – különösen csecsemőkorban – némi magasabb. Azaz, folytatta okfejtését Arbuthnot, ha az isteni gondviselés nem szólna közbe, és tériénél az 50-50%-os arányt, akkor a végeredmény az lenne, hogy nem jut minden nőnek férfi. Szerencsére azonban annyival több fiú születik, hogy az ellensúlyozza a nagyobb halandóságukat, így mire házasodásra kerül sor, mindenki talál párt. Ugyan mi más lenne ez, ha nem az isteni gondviselés?

Egy bökkenő azonban van; és ezt már Arbuthnot is felismerte. *Biztos*, hogy több fiú születik, mint lány? A naiv válasz, hogy persze, nézzünk rá az ábrára. Azonban az előzőekben mondottak ismeretében valószínűleg sokan rájönnek, hogy ez egyáltalán nem ilyen egyszerű. Mi van akkor, ha *valóságban* 50-50% a két valószínűség, és pusztán a véletlen ingadozás miatt kaptunk ettől eltérő számokat...? Az előzőek fényében világos, hogy ezt nem zárhatjuk ki, elvégre ez olyan, mint a pénzfeldobás: fej a fiútöbbség, írás a lánytöbbség, és minden év egy dobás. Ha a pénzérme szabályos, tehát nincs különbség a születéskori arányban, *akkor is* lehet, hogy – pusztán a véletlen ingadozás szeszélye folytán – csupa fejet dobunk. Ez az, amit Arbuthnot



5.1. ábra. John Arbuthnot (1667-1735).

Christened.					
Anno.	Males.	Females.	Anno.	Males.	Females.
1667	5616	5322	1689	7604	7167
68	673	5560	90	7909	7302
69	6506	5829	91	7662	7392
70	6278	5719	92	7602	7316
71	6449	6061	93	7676	7483
72	6443	6120	94	6985	6647
73	6073	5822	95	7263	6713
74	6113	5738	96	7632	7229
75	6058	5717	97	8002	7767
76	6552	5847	98	8426	7626
77	6423	6203	99	7911	7452
78	6568	6033	1700	7578	7061
79	6247	6041	1701	8102	7514
80	6548	6299	1702	8031	7656
81	6822	6533	1703	7765	7683
82	6909	6744	1704	6113	5738
83	7577	7158	1705	8366	7779
84	7575	7127	1706	7952	7417
85	7484	7246	1707	8379	7687
86	7575	7119	1708	8239	7623
87	7737	7214	1709	7840	7380
88	7487	7101	1710	7640	7288



(a) Facsimile oldal Arbuthnot eredeti, 1711-es publikációjából, mely a táblázatának a végét mutatja; minden két hasában az évszám melletti bal oldali oszlop a fiúk, a jobb oldali a lányok száma.

(b) Arbuthnot adatai grafikusan ábrázolva.

5.2. ábra. Arbuthnot születésekre vonatkozó adatai.

felismert, és ez az, ami összeköti ezt a kérdést a gyógyszer hatásosságával. Az, hogy a 82 év mindegyikében fej jött ki, nem jelenti tehát *biztosan* azt, hogy a pénzérme nem szabályos – azaz az isteni gondviselés létét – hiszen szabályos pénzérmével is dobhatunk sokszor fejet egymás után, akár 82-szer is.

És itt jön a kulcsgondolat, amiben Arbuthnot továbbment: számszerűvé tette ezt a bizonytalanságot. Igen, mondta, dobhatunk fejet 82-szer egymás után szabályos érmével is, de ennek *nagyon pici a valószínűsége*. Számoljuk is ki konkrétan: ha szabályos a pénzérme, akkor minden dobásnál $\frac{1}{2}$ a fej valószínűsége. Mennyi annak a valószínűsége, hogy kétszer egymás után fejet dobunk? Alkalmazzuk az előző logikát: az esetek felében, azaz 4 esetből 2-szer dobunk várhatóan fejet először, ezen két eseten belül, mivel a dobás az előzőtől független, ismét csak az esetek felében, azaz egyszer dobunk fejet – összességében tehát 4-ből 1-szer, 25% valószínűséggel fordul elő a dupla fej. Azt látjuk, hogy ha először is $\frac{1}{2}$ a valószínűség és másodszor is, akkor annak a valószínűsége, hogy *mindkétszer* bekövetkezik az esemény, feltéve, hogy egymástól függetlenek, egyszerűen a két valószínűség szorzata: $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Ez a szabály általában is igaz: annak a valószínűsége, hogy háromszor egymás után fejet dobunk egy szabályos pénzérmével, $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^3} = \frac{1}{8}$ és így tovább. És akkor jöjjön az izgalmas rész: mennyi annak a valószínűsége, hogy 82-ből 82-szer fejet dobunk, pusztán a véletlen ingadozás peche folytán, ha valójában szabályos a pénzérme: $\frac{1}{2^{82}} = \frac{1}{4835703278458516699884444} = 0,000000000000000000000000002068\%$. Ilyen számot felfogni is nehéz, úgyhogy a viszonyítás kedvéért: ennél 57-szer valószínűbb, hogy megnyerjük a lottót... háromszor egymás után.

Tehát: bár azt nem mondhatjuk, hogy biztosan egyenlőtlen a fiú-lány arány, de ha egyenlő lenne, akkor egy elképesztően ritka esemény történt volna, hogy mi mégis 82-ből 82-ször fiútöbbséget kaptunk (pusztán a véletlen ingadozás szeszélye folytán). Azt nem mondhatjuk, hogy „kizárt”, hogy ez legyen a helyzet, de azt mondhatjuk, hogy „nem hisszük el”, hogy ez történt. És igen, ezzel hibázhatunk, sőt, az is látható, hogy mennyit: ha így járunk el, akkor minden 4835703278458516699884444. esetben rossz döntést fogunk hozni. (Filozófiai gondolkodnivaló: azonban egy konkrét esetben nem tudhatjuk, hogy az véletlenül nem az ennyiedik-e...!) Ha azonban ezt nem vállaljuk, akkor soha nem fogunk tudni döntést hozni, tehát ésszerűnek tűnik azt mondani, hogy ilyen eredmény „már nem tudható be” a mintavételei ingadozásnak. A kérdés egyedül az, hogy hol húzzuk meg a határt: akkor is elvetjük, hogy a kocka szabályos, ha 2-ből 2-szer dobunk fejet? (Ez 25% valószínűséggel fordulhat elő a szabályosság esetén.) Vagy csak ha 3-ból 3-szor? (12,5%) Vagy csak 10-ből 10-szer? (0,1%) ...vagy csak 82-ből 82-ször esetén? Vagy ez se elég biztos, legyen inkább 100-ból 100? Ez persze már nem tisztán statisztikai kérdés, hiszen azon is műlik, hogy mi a hozzáállásunk a hibázáshoz, a statisztika első szerepe, hogy egyáltalán megmutatta, hogyan tudjuk ezt számszerűen tenni, hogy mit jelent a „betudható-e a véletlen ingadozásnak” objektíven, számokra lefordítva.

(Arbuthnot megfigyelése egyébként mai szemmel nézve is helytálló. A legfrissebb adatok szerint fogantatáskor valóban igen pontosan 50-50% a fiú-lány arány, az Arbuthnot által is leírt eltérés valódi oka az, hogy a lánymagzatok méhen belüli vesztesége némileg magasabb, ezért van végeredményben születéskor fiútöbbség; Magyarországon jelenleg 100 lányra kb. 106 fiú jut születéskor. Ez a világon nem teljesen egységes, ennek sajnos sejthetőleg nem csak biológiai okai vannak, hanem a szelektív abortusz is belejátszik, például Kínában ugyanez az arány 100:115 feletti, de vannak tartományok, ahol elképesztő módon a 100:130-at is meghaladja ez az arány.)

5.3. A mágikus p -érték

Egyetlen kérdést kell már csak tisztáznunk. Mi van akkor, ha Arbuthnot azt tapasztalta volna, hogy a 82 évből 81 évben van fiútöbbség? A logikánk az volt, hogy feltéve, hogy nincs nem különbség, mekkora valószínűséggel jön ki az, ami ténylegesen ki is jött. Ez jelen esetben azt jelenti, hogy ha szabályos a pénzérme, akkor mekkora annak a valószínűsége, hogy a 82-ből 81-szer fejet dobunk. Ennek kiszámolása nem túl nehéz: annak a valószínűsége, hogy elsőre írást dobunk és utána csupa fejet, pontosan ugyanannyi, mint hogy csupa fejet dobunk, $\frac{1}{2^{82}}$. Igen ám, de akkor is megvalósul a 82-ből 81, ha másodiknak dobunk írást (még plusz $\frac{1}{2^{82}}$ valószínűség), ha harmadikra, ha negyedikre, és így tovább, egész addig, hogy 82. dobásra jön ki az írás. Egy szó mint száz, annak a valószínűsége, hogy 81 fejet dobunk és 1 írást, $82 \cdot \frac{1}{2^{82}}$.

Az igazán fontos dolog azonban nem ez – hanem az, hogy ez a szám félrevezető. A mi kérdésünk szempontjából ugyanis ennek az eredménynek a fontos momentuma nem az, hogy *pont* 81 fej (azaz fiútöbbség) volt, hanem az, hogy *legalább* 81 fej volt – hiszen ha 82 lett volna, az csak még jobban ellentmondott volna az egyenlőségnek! A jó mérőszám tehát nem az lesz, hogy

feltéve az egyenlőséget mekkora valószínűséggel jön ki az, ami ténylegesen ki is jött, hanem az, hogy mekkora valószínűséggel jön ki az, vagy az egyenlőségnek még jobban ellentmondó eredmény. Ez a jelen esetben tehát $82 \cdot \frac{1}{2^{82}} + \frac{1}{2^{82}}$, ez fogja jól megmutatni, hogy az adataink mennyire kompatibilisek azzal a feltevéssel, hogy 50-50% a fiúk és lányok aránya.

Ezt szokás az orvosi statisztikában *p*-értéknek nevezni, ez az orvosi vizsgálatok kiértékelésében már-már mágikus jelentőséggel bír (később majd látni fogjuk, hogy néha sajnos kissé túlságosan is mágikussal). Azt, hogy egy eredmény mennyire hihető, hogy pusztán a mintavételi ingadozás miatt jött ki, napjaink orvosi vizsgálatainak abszolút túlnyomó része a *p*-érték közlésével oldja meg: minél alacsonyabb ez, annál kevésbé hihető, hogy az eredmény pusztán a véletlen ingadozás műve, és annál biztosabbak lehetünk benne, hogy valódi hatás van a hátterében. Ez kissé olyan, mint a középiskolai matematikaórán az indirekt bizonyítás: feltessük, hogy igaz egy állítás (50-50% a nemi arány), ebből olyasvalamire jutunk, ami ha nem is lehetetlen, de nagyon valószínűtlen (valószínűtlenebb, mint háromszor egymás után megnyerni a lottót), úgyhogy ebből arra a következtetésre jutunk, hogy akkor az állítás minden bizonnal mégsem volt igaz.

Ezt a kiinduló állítást az orvosi statisztikában nullhipotézisnek szokták nevezni, ahogy a fenti példák is mutatják, ez általában azt fogalmazza meg, hogy „nincs eltérés”, „nincs hatás” (ugyanaz a nemi arány, a valóságban nem hat a gyógyszer stb.). A *p*-érték pedig azt fogja megmutatni, hogy az adatok – a véletlen ingadozás figyelembevételével! – mennyi bizonyítékot szolgáltatnak arra, hogy a nullhipotézis nem áll fenn, mennyire mondanak ellent a nullhipotézisnek. Megjegyzendő, hogy a gyógyszeres példa annyiban bonyolultabb, hogy ott nem egy eredményt ($82/82$) hasonlítunk egy általunk megadott értékhez (50-50%), hanem két eredményt (kezelt beteg meghalt-e – nem kezelt beteg meghalt-e) egymáshoz, de ez pusztán technikai kérdés. Ahhoz, hogy „kezelt beteg túlélt, nem kezelt meghalt” szokásos módszerekkel kiszámolhatjuk a *p*-értéket (ugyanúgy, mint minden más esethez), ami jellemzni fogja, hogy ez mennyire tudható be a véletlen ingadozásnak, ha a gyógyszer igazából nem hat. Amennyiben ez nagyon pici, úgy arra következtetünk, hogy a gyógyszer minden bizonnal hat.

Van ebben az egész logikában egy rendkívül fontos dolog, amire feltétlenül érdemes külön is felhívni a figyelmet. Jelesül: vegyük észre, hogy nem a természetes kérdésre adunk választ, hanem annak *fordítottjára!* Nem azt mondjuk meg, hogy ha 82-ből 82 fiútöbbséget tapasztaltuk, akkor mennyire valószínű, hogy a valóságban egyenlő az arány, hanem azt, hogy ha egyenlő *lenne* az arány, akkor mekkora valószínűséggel kapnánk 82-ből 82-t. Nem azt mondjuk meg, hogy ha a kezelt beteg túlélt, a nem kezelt viszont meghalt, akkor mekkora valószínűséggel hatástanak a gyógyszer, hanem azt, hogy ha hatástanak *lenne* a gyógyszer, akkor mekkora valószínűséggel kapjuk azt, hogy a kezelt beteg túlélt, a nem kezelt viszont meghalt. Egyfajta fordított logikát alkalmaz, ami – mint láttuk – egyáltalán nem ésszerűtlen (ha nagyon-nagyon pici a valószínűsége, hogy a gyógyszer hatástanásága esetén azt kapjuk, amit ténylegesen kapunk is, akkor elég kézenfekvő azt mondani, hogy „minden bizonnyal” hat a gyógyszer), de ettől még fordított a logika. És ennek bizony lesznek következményei: nem véletlenül van a „minden bizonnyal” idézőjelben – de erről majd kicsit később.

5.4. Döntéshozatal véletlen jelenlétében

Pár apróságot még helyre kell rakkunk a bemutatott megközelítés kapcsán. Először is, meg kell válaszolnunk azt a kérdést, hogy hogyan hozzunk döntést! Mert végül csak azt kell kimon-dani, hogy hat a gyógyszer, vagy nem hat a gyógyszer. Ennek nagyon sok statisztikai, sőt, filozófiai kérdése (és problémája) van, e helyütt elég, ha annyit mondunk, hogy napjainkban az orvosi vizsgálatok abszolút többsége esetén azt az elvet követik, hogy a hatást akkor tekin-tik valósnak, akkor mondják, hogy nem igaz, hogy a gyógyszer hatástalan és a vizsgálatban kapott hatás pusztán a véletlen ingadozás műve, ha a *p*-érték 5%-nál kisebb. Más szóval: úgy választjuk meg, hogy mikor mondjuk, hogy a gyógyszer hatásos, hogy 5% valószínűsége legyen annak, hogy egy hatástalan gyógyszer esetén pusztán a véletlen ingadozás miatt előálló eredmény miatt azt mondjuk, hogy a gyógyszer hatásos. Ezt az általában 5%-ra rakott küszö-böt *szignifikanciaszintnek* hívjuk. Azt, hogy a gyógyszer hatása – adott szignifikanciaszinten – már nem tudható be a véletlen ingadozásnak, gyakran úgy is mondják: szignifikáns hatást találtunk.

Hibázhatunk ezzel a kijelentéssel? Hogyne, sőt, az egészben az a jó, hogy az előbbi mondatból ennek a pontos valószínűsége is látható! Ha olyan hatásosságoknál mondjuk ki, hogy a gyógy-szer hat, ami a gyógyszer hatástanansága esetén minden össze 5%-os valószínűséggel fordulhatna elő, akkor természetesen 5% annak a valószínűsége, hogy egy hatástalan gyógyszert tévesen hatásosnak minősítünk. Ezt szokás *elsőfajú hibának* nevezni. A szignifikanciaszint tehát meg fog egyezni ennek a hibázásnak a valószínűségével.

Ezt látva egy kérdés azonnal adja magát: ha ez így van, akkor miért 5%-os szignifikanciaszintet használunk? Miért nem 1%-ot? Miért nem 0,1%-ot? Mielőtt erre válaszolunk, gondoljuk végig, hogy ez mit jelent: ha lejjebb visszük a szignifikanciaszintet, akkor csak a kutatásban nagyon-nagyon hatásosnak bizonyult gyógyszerekre mondhatjuk azt, hogy „na, ez tényleg hat” (és nem csak a mintavételi ingadozás miatt tűnik úgy, miközben valójában hatástalan).

Ez tényleg lecsökkenti a hatástalan gyógyszerek – téves – törzskönyvezésének a valószínűségét, ez tény, csakhogy a probléma, hogy az elsőfajú hibával szemben áll egy másik lehetséges tévedés: az, hogy hatásos gyógyszert tévesen hatástanannak minősítünk; ezt szokás *másodfajú hibának* nevezni. (A kutatás *erejének* hívjuk annak a valószínűségét, hogy ezt nem követjük el: az erő annak a valószínűsége, hogy a hatásos gyógyszert *tényleg* hatásosnak minősítjük. A legtöbb kutatásban igyekeznek ezt 80-90% körülire belőni, később látni fogjuk, hogy hogyan.) Ha nagyon lecsökkentjük a szignifikanciaszintet, akkor a hatástalan gyógyszereket csakugyan valószínűtlen, hogy törzskönyvezzük, de közben nagyon megnöveljük annak a valószínűségét, hogy a valóban hatásosakat sem fogjuk! Azaz, lecsökkentjük az erőt. A szignifikanciaszintet tehát túlságosan leszorítani sem érdemes, kompromisszumot kell kötni a két szempont, a kétféle hibázás között. Az 5% a gyógyszerészeti erős óvatosságát fejezi ki: sokkal inkább óvakodunk attól, hogy egy hatástalan – pláne káros – szert törzskönyvezzünk, minthogy egy hatásosat visszatartsunk.

5.5. Az egyik legfontosabb félreértés

Láttuk tehát, hogy a véletlen kezelésére szolgáló apparátusunk egyfajta fordított logikát alkalmaz: nem arra ad választ, hogy a kutatási eredményünk fényében mennyire valószínű, hogy igazából hatástan a gyógyszer, hanem arra, hogy ha igazából hatástan lenne, akkor mekkora valószínűséggel kaphatnánk olyasféle eredményt, mint amit ténylegesen kaptunk. Ez a fordítottság azonban komoly félreértések forrása lehet – és ez a gondolkodási hiba messze nem csak itt jelentkezik.

A véletlen ingadozás miatt soha nem tudunk biztos döntést hozni: ha a kutatásban hatásosnak is bizonyul például egy gyógyszer, minden fennáll a lehetősége, hogy igazából nem hat, csak pechünk volt a véletlen ingadozás miatt (úgy, ahogy egy szabályos pénzérmével is előfordulhat, hogy 10-ből 10-szer fejet dobunk – legfeljebb kicsi a valószínűsége). Azért, hogy egyáltalán tudjunk dönteni, valahol határt kell húzni: ha ez a valószínűség nagyon kicsi, akkor (elfogadva, hogy ezzel hibázhatunk!) azt mondjuk, hogy „minden bizonnyal” hat a gyógyszer – noha az említett kis valószínűség épp ahoz tartozik, hogy igazából nem hat. Mi mégis azt mondjuk, hogy hat: ez a kis valószínűség lesz tehát annak a valószínűsége, hogy a hatástan gyógyszert hatásosnak mondjuk. Ott hagytuk legutóbb abba, hogy a napjainkban általánosan használt mérték szerint ezt 5%-ra állítjuk: úgy állítjuk be a szigorúságunkat, hogy a hatástan gyógyszer esetén 5% valószínűsséggel mondjuk azt, hogy hatásos. (Ezt lehetne ugyan lejebb vinni, de akkor megnőne annak a valószínűsége, hogy hatásos gyógyszereket sem törzskönyvezünk.)

Ezen rövid ismétlés után lássuk most, hogy mi az emlegetett leggyakoribb gondolkodási hiba ennek kapcsán!

5.5.1. Terroristák a városban

Kezdjünk egy – látszólag – teljesen ide nem vágó találós kérdéssel. (Valódi kérdés: a kedves olvasó is bátran tippeljen! Ami az igazán érdekes, hogy akkor mi jön ki, ha nem „tudományosan” levezeti az ember, hanem gondolkodás nélkül, zsigerből válaszol.)

Képzeljünk el egy várost, ahol 1 millióan laknak, köztük 100 keresett bűnöző (drámaibb változatban: terroristák). A városban fölszerelnek egy körözési adatbázison alapuló, automatikus arcfelismerővel ellátott kamerarendszert, hogy megkeresse a terroristákat. Ez az arcfelismerő rendszer nagyon jól működik: amennyiben tényleg egy terrorista kerül be a képbe, akkor 99% valószínűsséggel azt fogja mondani, hogy az illető terrorista, ha pedig egy ártatlan sétál a képbe, akkor 99% valószínűsséggel azt mondja, hogy az illető ártatlan. A kérdés a következő: besétál a kamera képébe egy ember és megjelenik alatta a felirat, hogy terrorista. Mekkora a valószínűsége, hogy a 99%-os pontossággal működő kamerarendszerünk jól tippelt, azaz, hogy az illető tényleg terrorista?

És akkor most, hölgyeim és uraim, kérem, tegyék meg tétjeiket...!

Az emberek többsége 90-95%, vagy annál is nagyobb számokat tippel; a leggyakoribb a 99% – hát meg is mondta, hogy ilyen pontos a kamera, mi itt a kérdés? Ez a válasz azonban teljesen rossz. A valóságban annak a valószínűsége, hogy az illető tényleg terrorista, ha egyszer a gép kiírta róla, hogy terrorista, valójában... kevesebb, mint 1%!

Amit sokan elfelejtenek az az, hogy az 1 millió lakosból mindössze 100 terrorista van. Kicsi a terroristák „alapgyakorisága”, azaz annak a valószínűsége, hogy a kamera előtt álló ember pont terrorista legyen, még mielőtt egyáltalán megnéztük volna, hogy a gép mit írt ki. Szakkifejezéssel élve: a prior valószínűsége az adott személy terroristá voltának igen kicsi, mindössze $100/1$ millió. (A prior latinul annyit tesz: előzetes; ez olyan értelemben „előzetes valószínűség”, hogy az információk begyűjtése, jelen esetben a gép által kiírt azonosítás megtékinése előtti valószínűsége annak, hogy a személy terroristá.)

A probléma, hogy amikor a gép azt jelzi, hogy terroristát lát, az az esetek túlnyomó többségében nem abból fog származni, hogy egy valódi terroristá sétált be és helyesen azonosította a szoftver, hanem abból, hogy egy ártatlan sétált be a képbe és tévedésből minősítette terroristának. Igaz, hogy ez utóbbinak mindössze 1 százalék a valószínűsége, de annyival sokkal, sokkal, sokkal több ártatlan fog átsétálni a kamera előtt, hogy ennek még az 1%-a is jóval több embert fog jelenteni, mint a mindössze 100 terroristá 99%-a.

A dolgot számszerűsíthetjük is. Képzeljük el, hogy minden egyes lakost átküldünk a kamera előtt: a 999 900 ártatlan 1%-a 9999 (fals) terroristá jelzést ad, a 100 terroristá 99%-a 99 (helyes) jelzést, azaz a terroristának minősített emberek mindössze a $\frac{99}{99+9999} = 0,98\%$ -a lesz ténylegesen terroristá – ez a válasz a kérdésre! E számítás logikáját hívják a matematikában Bayes-tételnek.

Mi a 99%-tól nagyon eltérő válasz oka? Az, hogy megfordítottuk a kérdést! Nem azt kérdeztük, hogy feltéve, hogy valaki terroristá, mekkora valószínűsséggel minősíti annak (*ez* a 99%), hanem azt, hogy feltéve, hogy terroristának minősített valakit, mekkora valószínűsséggel tényleg az. Ami viszont már nagyon nem 99%! A Bayes-tétel tehát azt teszi lehetővé, hogy megfordítsuk az ilyen feltételes valószínűségeket – csakhogy ehhez a prior valószínűségre is szükség van.

Érdemes megjegyezni, hogy a 0,98%-ot szokás a terroristá mivolt pozsterior („utólagos”) valószínűségének nevezni: a $100/1$ millió volt a valószínűség az arcfelismerő rendszer információjának megismerése előtt, a 0,98% pedig az után. Ha így nézzük, akkor a Bayes-tétel alkalmazása lehetővé tette, hogy *beépítsünk egy információt* a valószínűségbé: meg tudtuk pontosan határozni, hogy az az információ, hogy a gép kiírta, hogy terroristá, hogyan módosítja a terroristá mivolt valószínűségét. Természetesen a dolog nem muszáj, hogy itt véget érjen: elképzelhető, hogy a képfelismerő után használunk mondjuk egy bombakereső szkennert is, ennek a nézőpontjából a 0,98% lesz a prior valószínűség! Majd a bombakereső eredményének a fényében fog ez nőni vagy csökkeni, és így tovább. A bayes-i eljárás tehát lehetővé teszi, hogy a valószínűséget folyamatosan frissítsük a rendelkezésre álló információk alapján.

5.5.2. Kis kitérő: na de mi köze ennek az orvosláshoz?

Aki 99%-ot vagy hasonló értéket tippelt, az tehát lényegében figyelmen kívül hagyta a prior valószínűséget, azaz az alapgyakoriságot. Ez egy általános gondolkodási hiba, amire számtalan a fentihez hasonló példát lehetne még hozni. Mielőtt rátérünk a mostani témánkra, a véletlen ingadozás kezelésére szolgáló apparátusra, említsünk meg még egyet e példák közül, ugyanis orvosi is, izgalmas is – ez nem más, mint a diagnosztika!

Ha valaki ilyen szemmel néz rá az előbbi példára, akkor valószínűleg nagyon hamar meglátja a kapcsolatokat. A lakosok az emberek, a terroristák a beteg emberek, a kamera a diagnosztikai módszerünk, és végül a terroristák, azaz a betegek alapgyakorisága a populációban – a 100/1 millió – pedig nem más, mint a betegség elterjedtsége, az orvosok úgy szokták híjni, hogy prevalenciája. Az, hogy a kamera mekkora valószínűsséggel írja ki egy terroristára, hogy terrorista, nem más, mint hogy egy beteg embernél a diagnosztikai tesztünk mekkora valószínűsséggel mutatja ki a betegséget, az orvosok ezt szokták a teszt érzékenységének vagy szenzitivitásának nevezni. A másik 99%, hogy a kamera egy ártatlannál mekkora valószínűsséggel írja ki, hogy ártatlan (egy egészséges embert mekkora valószínűsséggel minősít egészségesnek a teszt) pedig a fajlagosság vagy specifikitás.

Mit mond tehát az előbbi példa nekünk, ha ilyen szemmel nézünk rá? Azt, hogy ha egy betegség ritka, akkor még egy egészen kitűnő – 99%-os szenzitivitású és 99%-os specifikású – teszt alkalmazva is igaz, hogy a pozitív lelet is csak azt jelenti, hogy kevesebb, mint 1% valószínűsséggel vagyunk tényleg betegek!

Ennek messzemenő következményei vannak például népegészségügyi szűrőprogramok tervezésekor: ha egy ilyen ritka betegséget igyekszünk kiszűrni, akkor még egy egyébként kiváló tesztet használva is igaz lesz, hogy a betegnek minősített emberek több mint 99%-a valójában *nem* beteg! (Ezt azért kell mindenkor figyelembe venni a szűrések tervezésekor, mert a betegnek minősítés általában további, néha veszélyesebb vagy költségesebb vizsgálatokat von maga után, lelkileg megterheli az alanyt stb.)

Az érdekes az, hogy bár a pszichológiai vizsgálatok szerint az emberek hétköznapi gondolkodása általánosságban véve nem bayes-i, tehát nem vesszük figyelembe a prior valószínűségeket (ezért is esnek sokan bele a terroristás kérdés csapdájába), az orvosi diagnosztikai logika – szerencsére! – sokkal inkább bayesiánus. Bár a legtöbb agydaganat jár fejfájással, a fejfájós betegnél nem ez az első, amire az orvos gondol, hanem egy sor betegség megelőzi, olyanok is, amik a ráknál ritkábban okoznak fejfájást – egész egyszerűen azért, mert annyira ritka betegség az agytumor. Ez a logika ép a prior valószínűség figyelembevétele! („A gyakori betegségek gyakoriak, a ritka betegségek ritkák” – szokták mondani az orvosok; ez pont a bayes-i gondolkodásmód megjelenése.) Pontosan ugyanazon infarktusra gyanús EKG birtokában egy 20 éves, makkekészséges lány esetén lehet, hogy egy tucat dologra előbb gondol az orvos, mint infarktusra, míg egy 75 éves, elhízott, cukorbeteg férfi esetén szinte biztosnak veszi a szív-rohamot. Nem azért, mintha nem ugyanúgy értelmezné az EKG-t: a pozitív EKG minden esetben megnöveli a szívinfarktus valószínűségét, de az első esetben annyira csekély a prior

valószínűsége, hogy még a megnövelés *után* is meglehetősen alacsony lesz (hiszen az EKG-nak sem 100%-os a specificitása és a szennitivitása). Egyébként olyan skálát is lehet találni, amin még az is igaz, hogy minden esetben *pontosan ugyanannyira* növeli meg a hiedelmünket az infarktusban a pozitív EKG.

Természletesen itt is működik az új információk beépítése – sorozatos vizsgálatokkal folyamatosan finomítható a valószínűség, például ha a lány laborja is pozitív lesz, akkor az már az ō esetében is első helyre katapultálhatja az infarktus diagnózisát a potenciális diagnózisok listájában. De nem csak ilyen tesztekre lehet gondolni: az is egyfajta diagnosztikai módszer, hogy az orvos megvizsgálja a beteget, vagy akár csak kérdez tőle valamit – például kiderül, hogy a lány családjában számos korai szíveredetű halálozás fordult elő. Az orvos ilyen lépések nyomán állítja elő a diagnózist, még ha ez implicate és nem is számszerűsítve zajlik az agyában (noha sokszor az is rendkívül hasznos lenne!).

5.5.3. Nagy kitérő: Dr. GépBayes rendel

Ezen a ponton adja magát az ötlet: miért nem automatizáljuk az egészet? Hiszen ez teljes egészében megvalósítható gépi úton, nem is kell ide orvos! Egy hatalmas adatbázisban eltároljuk egyrészt, hogy az egyes betegségek mekkora valószínűséggel okoznak egy adott tünetet vagy teszteredményt (ez alapvetően orvosi, biológiai kérdés, tehát elég stabil), másrészt, hogy mennyi az egyes betegségek gyakorisága, ez lesz a prior valószínűség (ez persze időben és populáció szerint is változhat), és ennyi. Innentől a számítógép elvégzi a beszorzást, kiírja az eredményt, kiválasztja legjobbként a legvalószínűbb diagnózist, és kész is vagyunk!

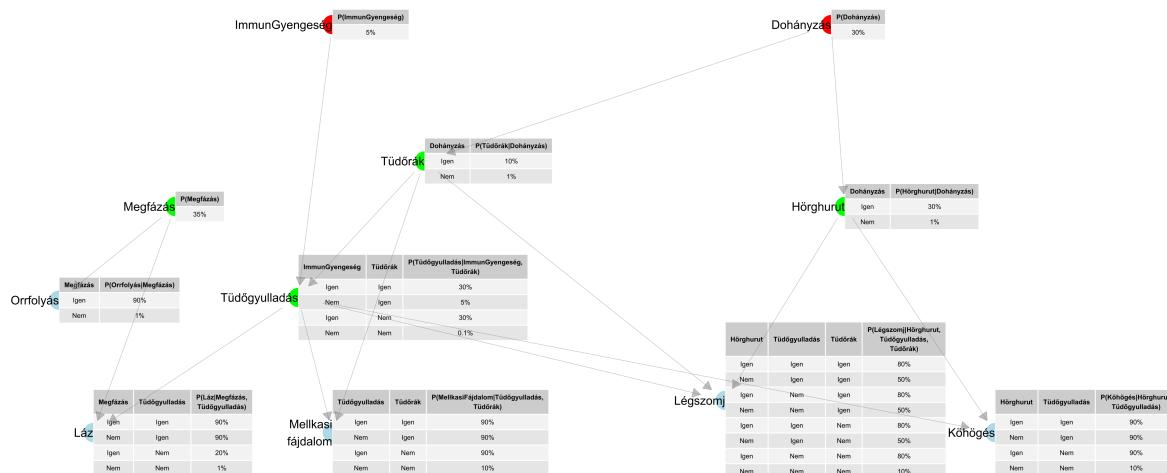
Az ötlet nem is annyira elborult, mint amennyire először hangzik; nagyjából a '70-es évek óta kísérleteznek is ilyen rendszerekkel. A legfontosabb probléma, hogy valójában nem egyetlen tünetet kell figyelembe vennünk. Ha például nem csak a beteg feje fájhat, hanem a hasa is, akkor már nem egyetlen valószínűséget kell letárolnunk, hanem 3-at. Ha 3 tünetünk van, akkor már 7 valószínűséget kell tárolni, és így tovább, miközben a valóságban nyilván több száz, vagy akár több ezer tünet lehetséges, pláne, hogy ugyebár ide tartoznak a kórelőzményi, vizsgálati adatok, teszteredmények is. A kezelendő valószínűségek száma már 100 tünetnél is ezer kvadrilliárd (leírva harmincegy számjegy...), és akkor még azt feltételeztük, hogy minden tünetünk bináris.

Hogyan tehetjük ezt a helyzetet kezelhetővé? Leegyszerűsíti a helyzetet, ha a tünetek függetlenek. Volt már róla szó, hogy függetlenségnél az események együttes bekövetkezésének a valószínűsége egyszerűen a külön-külön vett valószínűségeik szorzata, így ha a betegek tizedének fáj a feje, és tizedének a hasa, akkor (ha ezek függetlenek!) századunknak fog egyszerre fájni mindenkető – továbbra is elég tehát egy-egy valószínűséget, a két 10%-ot tárolni, az 1%-ot nem kell külön, mert kiadódik ezekből. Így 100 tünetnél is elég lesz 100 valószínűséget tárolni, függetlenség esetén ebből már minden kombináció valószínűsége kiszámolható. Valójában ennél egy kicsit kevesebb is elég: mivel úgy kérdezzük, hogy feltéve, hogy adott betegségen szenved, mi a tünetek előfordulásának a valószínűsége, így elég, ha a betegséget *feltéve* függetlenek.

Például a mellkasi fájdalom és a szívrohamra jellemző EKG-eltérés nyilván nem függetlenek, de *feltéve* a szívinfarktus tényét már azok: ha adottnak vesszük a tényt, hogy az alanyunknak van-e infarktusa, akkor már egyiknek sincs hatása a másikra, hiszen mindenkiüknek a következménye (az infarktusnak). Csak ez okozza az összefüggésüket, de közvetlen hatása egyiknek sincs a másikra: nem a rossz EKG-től fog megfájdulni a beteg mellkasa, és nem is a mellkasában érzett fájdalom rontja el az EKG-t.

Elképzelhető persze olyan helyzet is, amikor nem egy, hanem két dolgot – két betegséget vagy kockázati tényezőt – feltéve függetlenek a tünetek. Ez már kezd kicsit átláthatatlanná válni, hacsak...

...hacsak nem jövünk rá, hogy mennyivel szemléletesebb az egész, ha ábrát készítünk belőle! Rajzolunk pontokat, ezek jelöljék a betegségeket, tüneteket, kockázati tényezőket (az egyszerűség kedvéért most legyen minden bináris: fennáll vagy nem áll fenn), és köztünk húzzunk nyilakat, melyek azt mutatják, hogy mi hat mire közvetlenül. Lesznek pontok, amikbe nem fut nyíl – ezek nem függnek semmitől, egyszerűen azt kell odaírnunk, hogy mekkora valószínűséggel állnak fent. Amelyik pontba nyíl vagy nyilak futnak, ott a fennállás valószínűsége azoktól a pontktól függ, ahonnan a nyilak jönnek (nevezük ezeket szülőknek), így ott egy kis táblácskát kell odaírnunk, mely a szülő csomópontok összes lehetséges kombinációjára megadja, hogy a pont mekkora valószínűséggel áll fenn, ha a szülők adott értékeik. Ezt szokás Bayes-hálónak nevezni, a 5.3. ábra egy példát mutat ilyenre. (Az ábrán a színezésnek nincs számítási jelentősége, csak a pontokat csoportosítja: halványkékkel a tünetek, zöldek a betegségek, pirosak a kockázati tényezők.)



5.3. ábra. Egyszerű légúti diagnosztikai rendszer Bayes-hálója (Wiegerinck, Burgers és Kappen példája).

A Bayes-hálók sava-borsát az adja, hogy a nyilakat úgy húzzuk be, hogy a függetlenségi viszonyokat írják le. Még pontosabban: annak kell teljesülnie, hogy egy csomópont csak a szüleitől

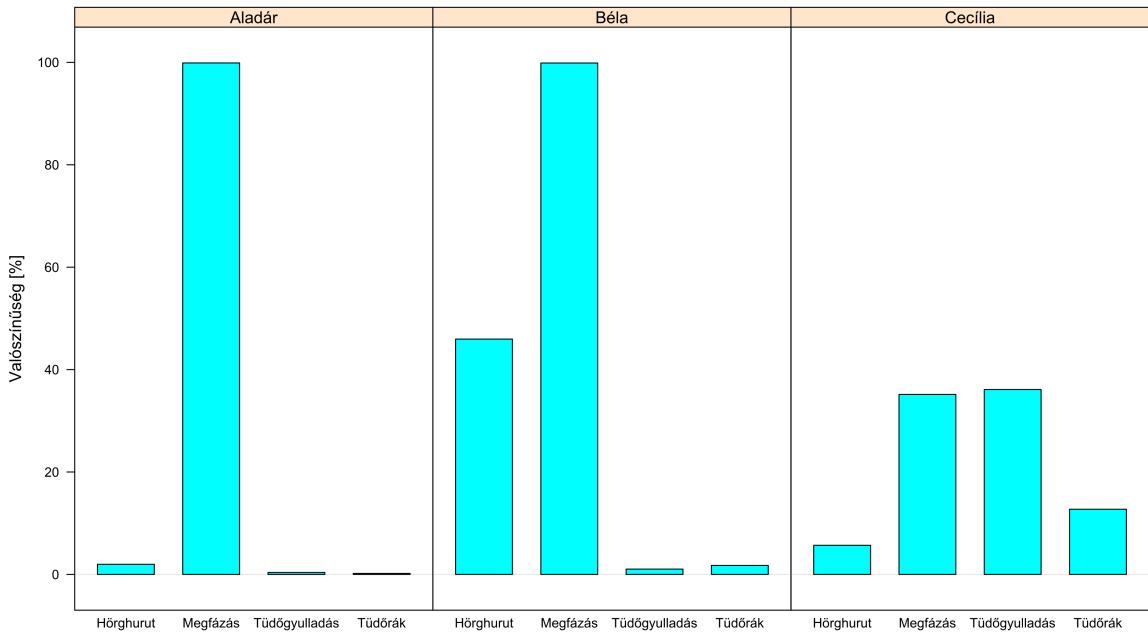
és azoktól a pontoktól függhet, amelyek belőle kiindulva elérhetők a nyilakon – de a többiből nem. Ebben az a fantasztikus, hogy ha ezt megvalósítjuk, akkor a pontok mellett látható néhány valószínűségből *minden* valószínűség kiszámítható! Az ábrán látható esetben például a teljeskörű leírás több mint 2000 valószínűség megadását igényelné, a Bayes-hálóban viszont minden össze 33 van. És mégis, ebből a 33-ból mind a 2000 kiszámolható – ehhez kellettek a függetlenségek, amelyeknek a Bayes-háló nagyon jól használható reprezentációját adja.

Na de mire jó ez az egész? A kezdeti orvosi szakértői rendszerek *diagnosztikusak* voltak: olyan jellegű szabályokat tartalmaztak, melyek a megfigyelésekkel vezettek az okok felé, „ha köhög a beteg, akkor ilyen valószínűséggel van hörghurutja”. Érdekes módon hiába tűnik kézenfekvőnek, hamar kiderült, hogy ez tévűt. A jó szakértői rendszerben *ok-okozati* szabályokat kell rögzíteni, olyanokat, melyek az okkból vezetnek a megfigyelések felé: „ha hörghurutja van a betegnek, akkor ilyen valószínűséggel köhög”. Ezeket szokták *modell-alapú* szakértői rendszereknek nevezni – a Bayes-háló pont ennek a filozófiának felel meg. A diagnosztikus szabályok kevésbé stabilak (ha kitör egy járvány, az egész szabályrendszer átalakul), diagnosztikus szabályból általában több kell, és gyakran kevésbé egyszerűen határozhatóak meg a hozzájuk tartozó valószínűségek, ráadásul nagy rendszerekben néha egészen természetellenes függőségekhez kell valószínűséget rendelni.

Igen ám, de az ok-okozati szabályok „iránya” viszont fordított, hogyan használjuk ezt diagnosztikára? A rövid válasz az, hogy minden további nélkül! A fordítottság semmilyen problémát nem jelent: ha egyszer megvan a Bayes-hálónk, akkor abból ugyebár minden valószínűséget ki tudunk számolni – ebben pedig az is benne van, hogy a betegségek valószínűségeit is meg tudjuk határozni a tünetek alapján! A Bayes-hálónak tehát mindegy, hogy milyen „irányban” következtetünk rajta. A fontos az, hogy a *valóságot írjuk le* (márpedig ezt az ok-okozati szabályok jelentik!), ha ez megvan, akkor onnan már minden irányban tudunk következtetni. A Bayes-hálókra elérhetők hatékony algoritmusok, amik ezt a következtetést „végigfuttatják” a hálón és meghatározzák a szükséges valószínűségeket.

A 5.4. ábrán erre látunk példákat: megadtuk a hálónak, hogy mit tudunk a betegről (azaz bizonyos csomópontok nem valószínűségek, hanem beállítottuk a konkrét értékét), ezután megkértük, hogy ezt futtassa végig a hálón a többi ponthoz beírt feltételes valószínűségek segítségével, majd lekérdeztük a végeredményt a minket érdeklő csomópontokon.

Aladár lázról, köhögésről és orrfolyásról panaszkodva érkezik a rendelőnkbe, egyébként egészséges nemdohányzó, mellkasi fájdalma, légszemje nincs – szinte biztos a megfázás. A köhögés ugyan nem illeszkedik a képhez, de egy ilyen egészséges alanynál nagyon valószínűtlen bármi más. Béla annyiban tér el, hogy dohányzik: ez azonnal játékba hozza – nézzük a Bayes-hálót! – a hörghurutot is. (Ne lepődjünk meg, hogy a valószínűségek összege nem 100 százalék, hiszen nincs olyan kikötés, hogy a beteg a 4 lehetséges körből pontosan 1-ben szenved.) Cecília állapota hasonlít Aladárhoz, viszont neki nem folyik az orra; sajnos őt nem személyesen vizsgáltuk meg, a munkatársunk pedig elfejtette megkérdezni, hogy van-e légszemje vagy fáj-e a mellkasá. Ez példa arra, hogy a Bayes-háló a hiányzó teszteredmények helyzetét is gond nélkül tudja kezelni: ezeket a csomópontokat meghagyjuk valószínűségeinek! Így sokkal bizonytalanabb az ábra (nézzük a hálót: a megfázás majdnem minden orrfolyással járna), de mivel a többi ok



5.4. ábra. Három beteg diagnosztikája az előbbi Bayes-hálóval..

nagyon ritka egy ilyen egészséges embernél, így még mindig nagy a valószínűsége. Gyönyörűen látszik a bayes-i logika: a láz és a köhögés sokkal jobban megfelelne a tüdőgyulladásnak, de annak a valószínűsége így is csak döntetlenig kúszik fel, mivel kockázati tényezők nélkül kicsi a tüdőgyulladás prior valószínűsége.

A valós orvosi szakértői rendszereknek persze akár több száz, sőt, több ezer csomópontjuk is lehet.

5.5.4. Na de mi köze ennek a véletlen ingadozás kezeléséhez?

E hatalmas kitérő után térjünk vissza konkrét témánkra, az egész történetnek ugyanis van egy közvetlenebb köze is a mi mostani kérdésünkhez. Valószínűleg sokan látják már, miről van szó: a mintázat teljesen ugyanaz. A véletlen ingadozás hatásának fenti módon történő kezelése is egy természetes kérdéshez képest fordított kérdésre ad választ, ahogy azt már a módszer bevezetésekor is megállapítottuk. Hiszen a természetes kérdés az, hogy „feltéve, hogy ezt meg ezt kaptuk a kutatásban, mekkora a valószínűsége annak, hogy valójában nem hat a gyógyszer”, mi viszont arra adunk választ az apparátusunk segítségével, hogy „feltéve, hogy valójában nem hat a gyógyszer, mekkora a valószínűsége annak, hogy ezt meg ezt kapjuk a kutatásunkban”. Sokan – ugyanazt a hibát elkövetve – azt gondolják, hogy ha a gyógyszer

hatása 5%-on szignifikáns, az azt jelenti, hogy 5% a hibavalószínűség, tehát, hogy valójában nem hat a gyógyszer.

Hogy a doleg végére járunk, játsszuk végig itt is ugyanazt a számítást! A szignifikanciaszint legyen 5%, az erő 80%; ezek a legszokásosabb értékek. Ha valaki követi a terroristás analógiánkat, akkor hamar rá fog jönni, hogy még egy furcsa dologra szükségünk van: arra, hogy mennyi a „hatástanlanság prevalenciája”, azaz milyen gyakori, mennyire valószínű az ilyen vizsgálatoknál, hogy a gyógyszer hatástanlan. Ez első ránézésre elég bizarr (épp azért végezzük a kutatást, hogy kiderítsük, hogy hat-e, honnan tudnám a vizsgálat előtt, hogy mennyire valószínű, hogy hat-e?!), de ezzel most ne törődjünk, fogadjuk el, hogy ismert a gyógyszer hatástanlanságának prior valószínűsége. A prior ezúttal is előzetes, olyan értelemben, hogy még a kutatás megkezdése előtt, a kutatás kimenete, mint információ begyűjtése előtt mennyire valószínű, hogy hatástanlan a gyógyszer. Ha valaki valami konkréthoz szeretné kötni, akkor gondolhat arra, hogy egy számos már sikeres gyógyszerrel rendelkező gyógyszercsalád egy minimálisan módsított új tagjánál ez a valószínűség nagy, annak viszont, hogy a hiperpulzatív mágneses térrrel kvantumtranszformált rezgőkristály hat, a prior valószínűsége csekély. A mostani példánkban legyen a hatástanlanság prior valószínűsége 90%. Lefuttatjuk a kísérletet, 5%-on szignifikánsnak bizonyul a beavatkozás (tehát ha nem hat *na*, legfeljebb 5% valószínűséggel jött volna ki olyan eredménynél, aminél azt mondjuk, tévesen, hogy hat). Akkor tehát 5% a valószínűsége annak, hogy valójában nem hat? Számolunk!

Képzeljük el, hogy 1000 párhuzamos világegyetemben állnak neki a kutatók teszteli az új szert. Ezen világokból várhatóan 100-szor lesz hatásos a szer, 900-szor nem – itt jelenik meg a prior valószínűség. Az előbbi esetekben, tehát amikor tényleg hat a szer, ezt 80% valószínűsséggel tudjuk kimutatni (erő), azaz 80 alkalommal minősítjük – helyesen – hatásosnak a készítményt. Az utóbbi esetekben, tehát amikor valójában nem hat a szer, ezt 5% valószínűséggel tévesztjük el (szignifikanciaszint), azaz 45-ször minősítjük – helytelenül – hatásosnak a szert. Összességeiben véve $80+45=125$ esetben lesz „hatásos” a minősítés. Mi persze nem tudhatjuk, hogy a 80-ba vagy a 45-be tartozunk, így azt mondhatjuk, hogy a „hatásos” címke esetén $45/125=36\%$ a valószínűsége annak, hogy valójában *nem* hat a gyógyszer! Ami ugye *nagyon nem* 5%! A kézenfekvő és közvetlenül releváns kérdésre („mekkora valószínűséggel hatástanlan a gyógyszer?”) nem a szignifikanciaszint ad választ, hanem a fenti – bayes-i – számítás.

Adódik mindezek után a kérdés: ha ez így van, és a bayes-i módszer ad választ a természetes – és nekünk fontosabb – kérdésre, akkor miért nem minden ezt használjuk? Miért használjuk egyáltalán, pláne miért meghatározó a „fordított logikán” alapuló elv? Azon túl, hogy a dologan vannak bizonyos történelmi okai (például a bayes-i eljárások általában számításigényesek, ami egészen a legutóbbi évtizedekig komoly problémát okozott), a talán legfontosabb ok a prior valószínűségek szükségessége. Nagyon sokan ódzkodnak attól, hogy illet kelljen megadniuk, mert úgy érzik, hogy szubjektív, hogy mi a gyógyszer hatásosságának prior valószínűsége, úgy érzik, hogy ez egy indokolhatatlan paraméter, ami bármilyen értékre beállítható és ezzel igazából akármi kihozható a vizsgálatból. Valójában a legtöbb ezzel foglalkozó kutató egyetért abban, hogy a ma használt eljárásban nem kevesebb a szubjektív döntési lehetőség, legfeljebb azok kevésbé vannak szem előtt, kevésbé explicitek, ez azonban nem feltétlenül előny (sőt). A

dolog mögött inkább a tehetetlenség a fontos faktor – így tanultuk, mindenki más is így csinálja, minden korábbi elemzés így készült stb. – ami pláne igaz egy olyan konzervatív területen, mint a gyógyszertudomány.

De mégis hogyan válasszuk meg a prior valószínűséget? Ez számos matematikai és filozófiai kérdést felvet; itt most talán csak egy elvre érdemes felhívni a figyelmet, ugyanis általánosabb tanulságokkal is bír, ez a Cromwell-elv. A Cromwell-elv azt mondja ki, hogy egy prior valószínűséget soha ne állítsunk 0%-nak (vagy 100%-nak) – azért, mert ebben az esetben semmilyen bizonyíték nem tudja megváltoztatni az álláspontunkat! Ahogy mondani szokták: „ahhoz se nulla prior valószínűséget rendeljünk, hogy sajtból van a Hold, különben egy hadseregnyi, sajttal visszatérő ūrhajós sem győz meg minket erről”. Tehát: ahhoz se nulla prior valószínűséget rendeljünk, hogy a hiperpulzatív mágneses térrel kvantumtranszformált rezgőkristály hat, bármennyire is úgy gondoljuk, hogy ez lehetetlen. Rendeljünk hozzá 0,1% vagy 0,001% vagy épp 0,00000001% prior valószínséget, de ne 0-t – különben akárhány beteg meggyógyulása sem fog tudni meggyőzni minket róla, hogy hat (és így elvileg is elzárjuk magunkat egy új, fantasztikus orvostudományi felfedezéstől). Vegyük észre, hogy ez nem megy szembe a józan ésszel, ellenkezőleg, teljesen logikus a következménye: ha 0,00000001%-ra tesszük ezt a prior valószínűséget, az magyarul azt fogja jelenteni, hogy rettenetesen erős empirikus bizonyítékot várunk el a módszertől (messze többet, mint egy 5%-on szignifikáns vizsgálat), hogy elhiggyük, hogy tényleg működik – de ha ezt tudja produkálni, akkor elhísszük.

5.6. A véletlenség megszelídítése

A korábbiakban körbejártuk a véletlen ingadozás problémáját. Láttunk, hogy ez minden empirikus orvosi kutatás elkerülhetetlen velejárója, ami miatt soha nem tudunk biztosat mondani. De másrészről azt is láttuk, hogy milyen eszközökkel kezelhető, láttuk, hogy ha megszüntetni nem is tudjuk, hogyan tudjuk mérni az ebből fakadó hibát. De a talán legfontosabb kérdés még nyitva maradt: mi nem csak jellemezni akarjuk – hogyan tudjuk csökkenteni ezt a hibát? Hogyan vehetjük figyelembe, hogy az orvosi kutatásokat racionálisan tervezzük?

Az előbbiekből látott véletlenség problémája úgy is megfogalmazható, hogy a vizsgálatunkban kapott eredmény *ingadozni* fog a valódi érték körül. Ha a pénzérmén szabályos, akkor 82 dobásból *elvileg* 41 fej felel meg a valódi aránynak, de ettől még dobhatunk 42 fejet, 43 fejet vagy akár 50 fejet is – pusztán a véletlen ingadozás miatt, mindenféle csalás és ámítás nélkül, miközben a pénzérme szabályos volt. Sőt, amint azt megtárgyalunk, akár 82 fejet is dobhatunk, továbbra is teljesen szabályos érmével, csak ennek hihetetlenül kicsi a valószínűsége.

Fogjuk ezt fel a következőképp: adnak nekünk egy pénzérmét, és ki kell derítenünk, hogy mekkora valószínűsséggel dobunk vele fejet, e célból feldobjuk 100-szor és megszámoljuk, hogy hány fej jött ki. Az előzőeket kicsit általánosítva így is megfogalmazhatjuk az ekkor jelentkező problémát: a *valódi* érték, a fejek igazi valószínűsége egy adott, rögzített szám (csak mi nem tudjuk, hogy mennyi!), a dobálgatás-sorozatban kapott arány viszont *ingadozni* fog, olyan

értelemben, hogy ha újra elvégeznénk a dobálgatást, akkor más eredményt kapnánk, ha harmadszor is, akkor megint mást, és így tovább. A kutatás olyan, mint egy homályos szemüveg: amit rajta keresztül látunk, az nem biztosan a valódi érték, hanem egy elmosódott kép, mert a kapott eredmény ingadozni fog a valódi körül. A véletlen szeszélyén múlik, hogy pont mennyi fejet kapunk, erről mi csak annyit mondhatunk, hogy a valószínűségeiket kiszámíthatjuk; ezt mutatja a 5.5. ábra. Az itt megjelenített példában a pénzérme szabályos, tehát 50%-os valószínűséggel dob fejet – a kutatásban (dobálás-sorozatban) kapott fejek aránya viszont ingadozni fog e körül; az ábra pontosan mutatja hogyan, egy 100 feldobásból álló sorozatra.

Nekünk persze a probléma fordított irányban jelentkezik: mondjuk 60%-os arányt kaptunk, mondhatjuk-e ekkor, hogy szabálytalan a pénzérme? Nem biztos, mert az 50%-os valódi érték mellett is beingadozhat a kutatásban kapott érték 60%-ba. Az orvosi kutatásokkal való kapcsolat világos: gyógyszer nélkül 50% gyógyul meg, gyógyszerrel 60%, hat-e a gyógyszer? Nem biztos, mert ha a gyógyszer mellett *is* 50% a gyógyulási arány, akkor is előfordulhat, hogy – pusztán a véletlen ingadozás miatt – a mi mintánkban 60% gyógyult meg. Gyógyszer nélkül 50% gyógyul meg, gyógyszerrel is 50%, hatástan a gyógyszer? Nem biztos, mert ha a gyógyszer mellett 60% a gyógyulási arány, akkor is előfordulhat, hogy – pusztán a véletlen ingadozás miatt – a mi mintánkban csak 50% gyógyult meg. Az előző részekben pont az e jelenség kezelésére szolgáló apparátust ismertük meg, mellyel – noha biztos döntést nem hozhatunk, ez ellen nincs mit tenni – a bizonytalanság mértéke, a hibázás jellemzhető.

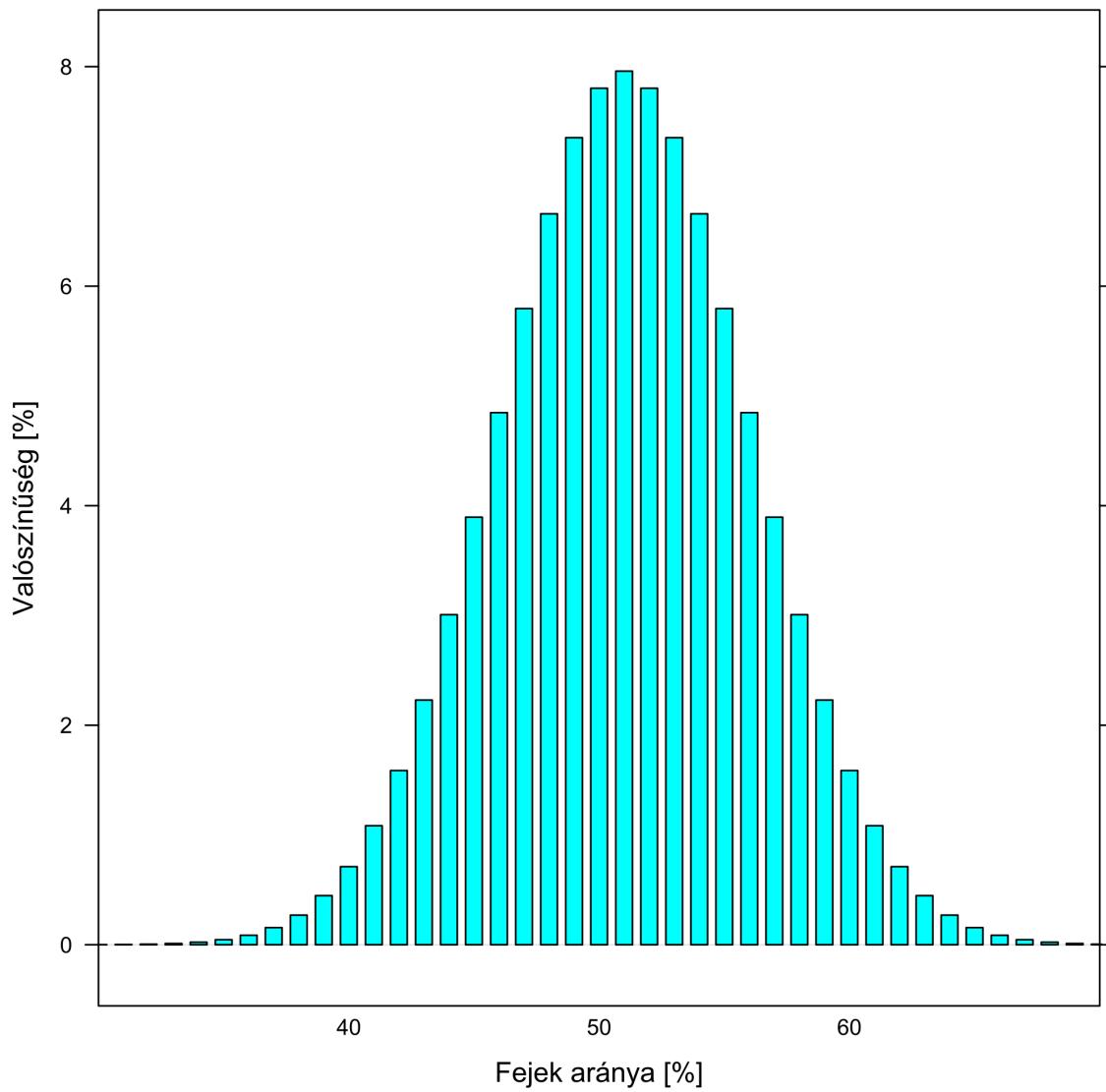
5.6.1. Megoldási lehetőségek nyomásban

Ennyi ismétlés után tegyük fel az egyik legfontosabb kérdést: mit tehetünk ez ellen? Mert az egy dolog, félreértés ne essék, nagyon fontos dolog, hogy jellemzni tudjuk a hibázást, de hogyan lehet csökkenteni? Ha egyszer a kutatásban mindenkiépp homályosan látunk, legalább a homályosság mértéke csökkenthető valahogyan? Tisztítható-e a szemüveg...?

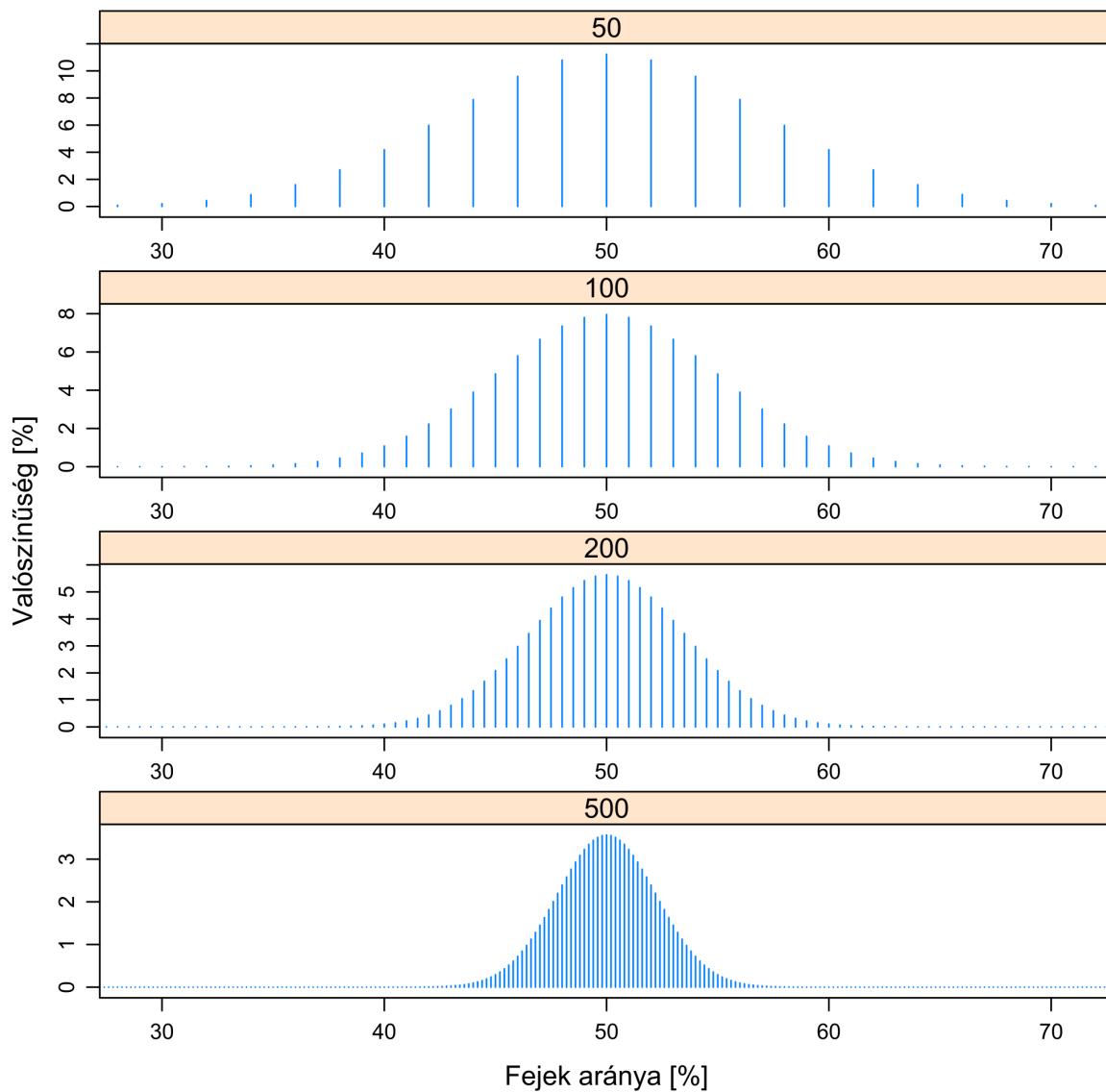
E kérdésre a válasz pozitív. Egyetlen dolgot tehetünk e bizonytalanság csökkentésére: meg-növelhetjük a mintanagyságot! (A kutatásban részt vevő alanyok számát, a pénzfeldobások számát stb.) Ekkor ugyanis pontosabban fogunk látni: minél nagyobb a mintanagyság, annál kisebb lesz az ingadozás; ezt illusztrálja a 5.6. ábra. Az előző ábra folytatásaként azt látjuk, hogy különböző mintanagyságok mellett milyen lesz a kutatásban kapott érték eloszlása; a valódi érték továbbra is 50%. Mindig ingadozunk e körül, de minél nagyobb a mintanagyság, annál kevésbé. (A tüskék egyre sűrűbben vannak, hiszen 50-es mintanagyság mellett minden egyes fej 2%-kal változtatja az arányt, 100-nál már csak 1-gyel és így tovább.)

Jól látszik, hogy bár az ingadozás nem szüntethető meg (összhangban azzal, hogy biztos döntést soha nem tudunk hozni!), de a mértéke csökkenthető a mintanagyság növelésével. Mindig homályosan látunk, de minél nagyobb mintánk van, annál kevésbé homályosan: minél nagyobb mintánk van, annál inkább a valódi érték körül fog tömörülni a kutatásban kapott eredmény.

Mi történik? Ahogy a köznyelv mondani szokta: „kiátlagolónak az ingadozások”, és most csakugyan erről van szó. Minél nagyobb a minta, annál kisebb lesz a szóródás, mert az egyik



5.5. ábra. A kutatás mint homályos szemüveg.



5.6. ábra. A kutatás mint homályos szemüveg homályosságának a függése a mintanagyságtól.

irányú kilengéssel szemben egyre valószínűbb, hogy lesz másik irányú is, ami ellensúlyozza a hatását.

A dolog kicsit arra hasonlít, mint a jel-zaj viszony a mérnöki szóhasználatban. Ezt eredetileg olyan helyzetekre értették, mint amikor zenét (jel) kell kihallanunk egy sercegő (zaj) rádióból. Annál több esélyünk van erre, minél hangosabb a zene (nagyobb a jel), illetve minél halkabb a sercegés (kisebb a zaj). A helyzet itt is ugyanaz! A jel a gyógyszer hatása, hogy mennyivel távolodott el a kezelt csoport halálozása a kontrollcsoportétől, a zaj a korlátos mintanagyságból fakadó ingadozás. Mikor tudjuk észrevenni – azaz igazolni, hogy nem csak a véletlen ingadozásnak tudható be – a gyógyszer hatását? Akkor, ha nagy a jel, illetve ha kicsi a zaj. A jelre, hogy milyen hangosan szól a zene, nincs ráhatásunk, ez a gyógyszer tulajdonsága. Amit tehetünk, azért, hogy mégis meghalljuk: kellően le kell halkítanunk a zajt. Ez az ugyanis, amit tudunk befolyásolni, mégpedig a mintaméret megnövelésével. Ha hangos a zene (nagyon hatásos a gyógyszer), akkor maradhat akár sok zaj is, de ha halk a zene (csak kicsit javít a gyógyszer), akkor nagyon le kell halkítanunk a zajt, azaz nagy mintára van szükség, hogy meghalljuk a zenét.

Természetesen minden ugyanúgy érvényes a káros hatásokra is: ha csak kicsi a különbség, akkor nagy minta kell, hogy észrevegeyük (azaz igazolni tudjuk, hogy nem pusztán a véletlen műve), ha nagy a hatás, akkor kis minta is elég.

5.6.2. Az orvosi kutatások edzeni mennek

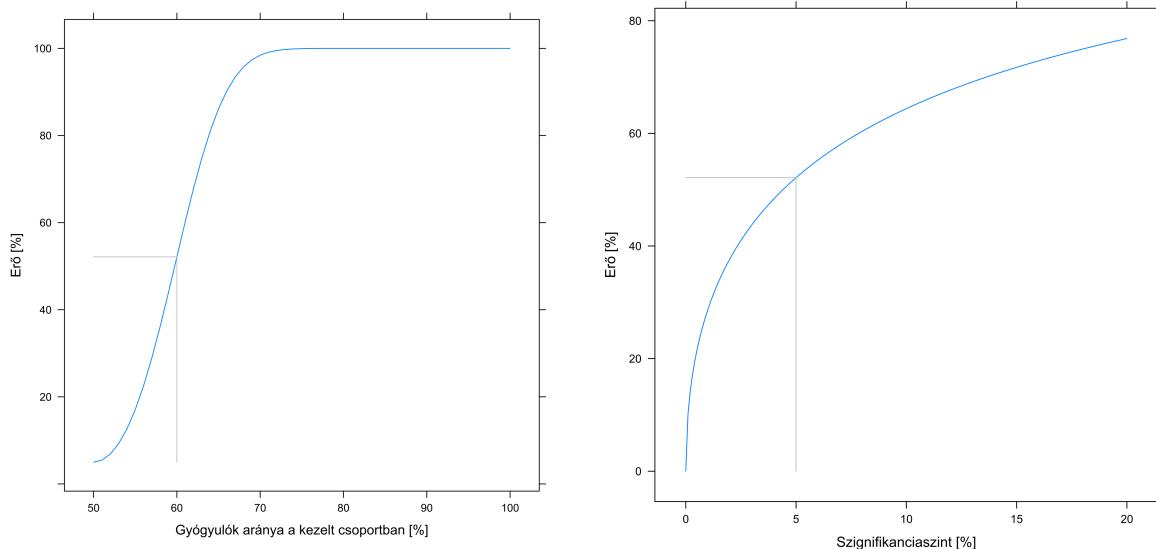
Eddig egyetlen paraméterét láttuk a véletlen jelenléte melletti döntéshozatalnak: a szignifikanciaszintet. A szignifikanciaszint megadja, hogy ha valójában nincs hatás, akkor mi mekkora valószínűsséggel mondjuk – tévedésből – mégis azt, hogy van. Ez fogja tehát meghatározni a szigorúságunkat: hogy milyen nagy hatásnál mondjuk azt, hogy „na, erről már nem hisszük, hogy a véletlen ingadozás miatt jött ki”. Ha 50% gyógyul gyógyszer nélkül, akkor milyen aránynál mondhatjuk, hogy a gyógyszer hatásos és az eredmény nem csak a véletlen műve? Ha 60%-ra emeli az arányt? Vagy elég az 55? Vagy 70 kell? Ezt szabjuk meg a szignifikanciaszint megválasztásával. Minél kisebb az értéke, azaz minél magasabbra rakjuk a limitet, annál valószínűlenebb, hogy egy hatástalan gyógyszert a véletlen ingadozás miatt hatásosnak minősítünk, de annál valószínűbb, hogy a hatásosakra is azt mondjuk, hogy hatástalan. Hiszen lehet, hogy az 50-et megemeli tényleg, de csak 60-ra, amire mi még azt mondjuk, hogy „á, ez lehetett a véletlen ingadozás miatt is”.

Az előzőekkel egybevetve látható, hogy mi a másik fogalom, amire szükségünk van a szignifikanciaszint mellett: az, hogy ha valójában hat a gyógyszer, akkor mi azt mekkora valószínűsséggel vesszük észre. Ezt szokták egy kutatás *erejének* nevezni; ennek a kiszámítására a megfelelő képletek rendelkezésre állnak. Ez természetesen mindig egy adott méretű hatásra vonatkozik, például úgy fogalmazhatunk: *ha a gyógyszer az 50%-os gyógyulási arányt 60%-ra emeli, akkor a kutatásunk erre vonatkozó ereje (100 beteg és 5%-os szignifikanciaszint mellett) 52,1%*. Azaz,

ha a gyógyszer *tényleg* 50-ről 60%-ra emeli a túlélést, és nagyon sok ilyen kutatást végrehajtanánk, akkor ezek kb. felében igazolódna, hogy a gyógyszer hat, a felében sajnos nem: hiába hat, a véletlen ingadozás miatt olyan eredmény jönne ki, amire azt mondánánk, hogy ez betudható lehet a véletlen ingadozásnak. Természetesen, ha a hatásnagyság nagyobb, mondjuk a gyógyszer 65%-ra emeli a túlélést, akkor a kutatás is erősebb lesz: ugyanezen paraméterek mellett már 86,1% az erőnk, azaz a gyógyszer hatásosságát már igen jó eséllyel ki fogjuk tudni mutatni.

Az erő természetesen a választott szignifikanciaszinttől is függ. Ha ezt 5%-ről 1%-ra csökkentjük, akkor ez utóbbi, 65%-ra növelő esetben is 68,1%-ra esik vissza az erő. Érthető: ahogy volt is róla szó, ha szigorúbbak vagyunk, akkor ugyan ritkábban veszünk észre nem valódi hatást, de ennek az ára, hogy a valódi hatások észlelése is nehezebb lesz.

Mindezeket összefoglalóan mutatja a 5.7. ábra. (Az ábrán a szürke vonalak az összetartozó – szövegben is tárgyalt – esetet mutatják. A kontrollcsoportban a gyógyulási arány 50%; az egyszerűség kedvéért feltételeztük, hogy ez fix érték, nem a kutatásból mértük ki.) Jól látszik, hogy minél hatásosabb a gyógyszer, annál valószínűbb, hogy ezt kimutatjuk, illetve, hogy minél szigorúbbak vagyunk, annál valószínűbb, hogy nem vesszük észre, hogy a gyógyszer hat.



- (a) Az erő függése attól, hogy a gyógyszer mennyire javítja a gyógyulási arányt, 100 beteg és 5%-os szignifikanciaszint mellett.
(b) Az erő függése attól, hogy milyen szignifikanciaszintet választottunk, 100 beteg és az 50%-os túlélést 60-ra emelő gyógyszer mellett

5.7. ábra. A kutatás erejének függése különböző tényezőktől.

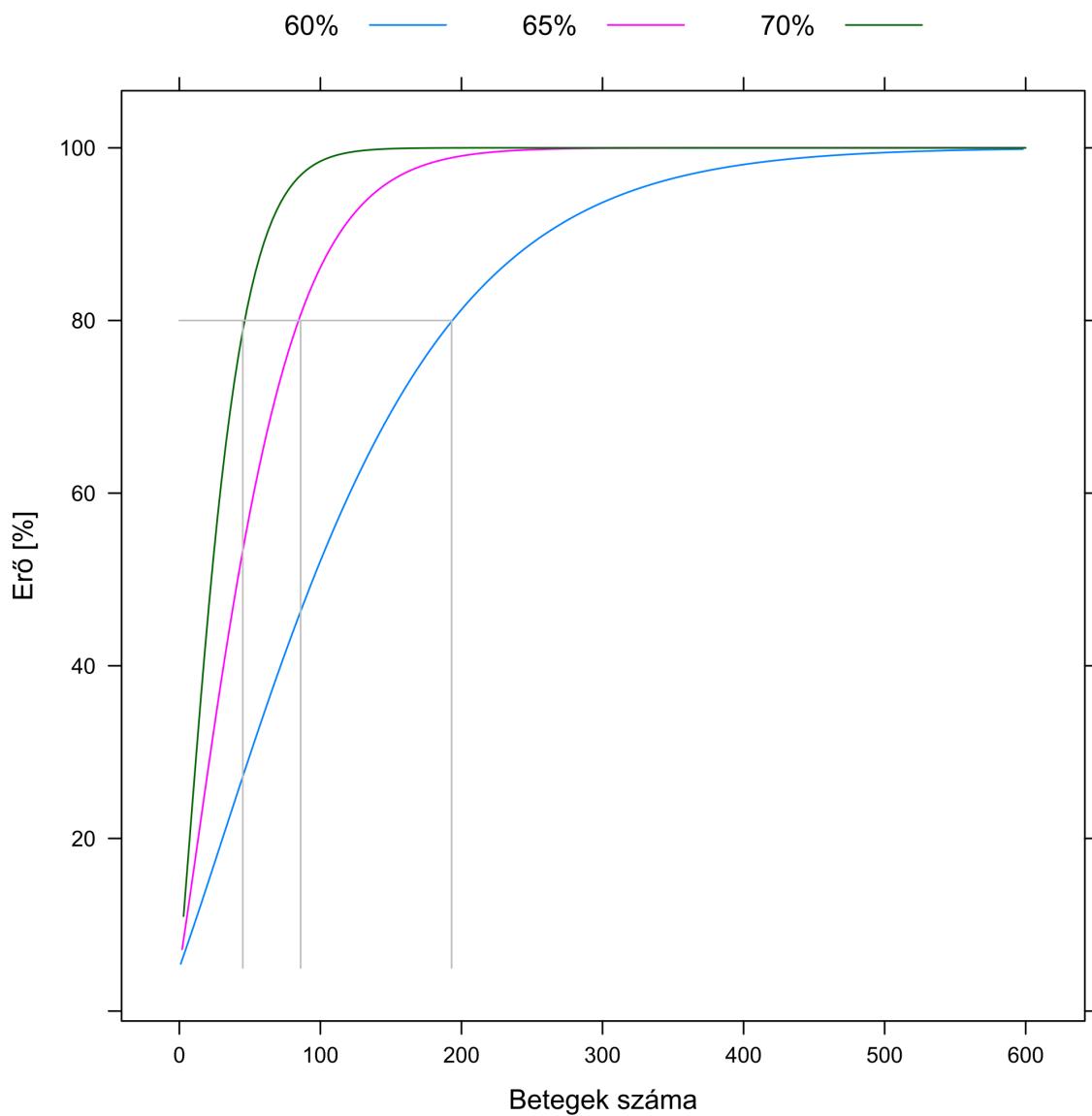
5.6.3. Kutatások mintaméretének tervezése

Kicsit elkanyarodtunk a jel-zaj viszonyánál tárgyalt fő kérdésüktől, a mintanagyság megválasztásától – de csak kicsit. Ugyanis a mintanagyság is egy ugyanolyan tényező az erő meghatározásában, mint a szignifikanciaszint vagy a hatásnagyság! A már említett képletek segítségével nyugodtan kísérletezhetünk azzal is, hogy különféle mintanagyságok mellett nézzük meg az erőt – a korább elmondottak fényében azt várjuk, hogy az egyre nagyobb mintanagyság, minden másról változatlanul tartva, egyre nagyobb erőt jelent. És csakugyan, ezt mutatja a 5.8. ábra. Az ábrán a kontrollcsoportban a gyógyulási arány 50%; a különböző színek azt mutatják, hogy ezt a gyógyszer mennyire növeli meg, a szignifikanciaszint 5%. Jól látszik, hogy minél hatásosabb a gyógyszer, annál nagyobb az erő (ezt már korábban is láttuk), de az is, hogy adott gyógyszerhatás mellett is növelhető az erő – a kutatásba bevont betegek számának emelésével. A szürke vonalak azt mutatják, hogy hány beteg kell a 80%-os erő eléréséhez.

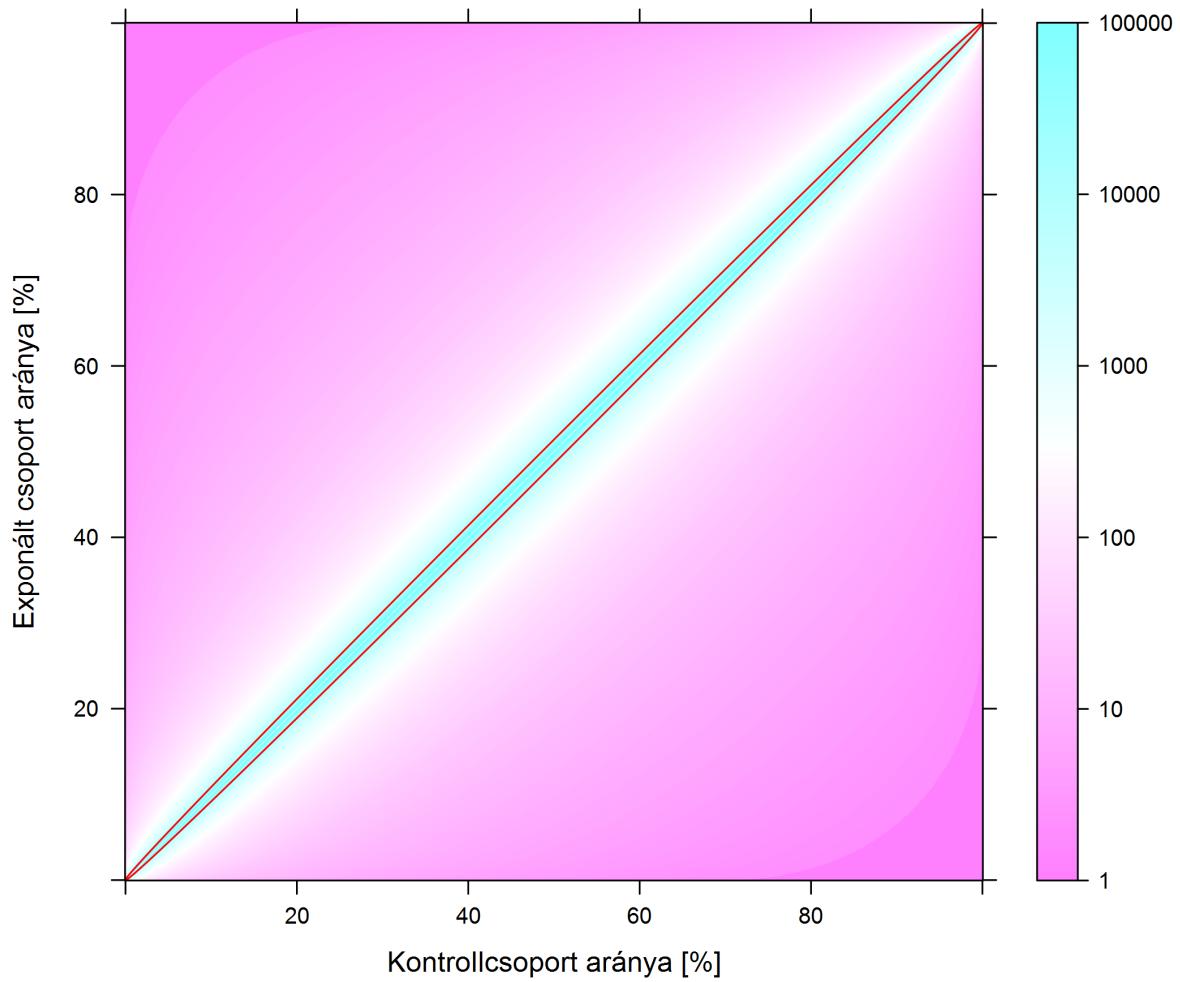
Egy a gyakorlati munka szempontjából roppant fontos összefüggéshez jutunk, ha az ábrára „fordítva” nézünk rá: nem azt nézzük, hogy adott betegszám mellett mekkora az erő, hanem azt, hogy adott erő eléréséhez mekkora betegszám kell! (Ezt mutatják a szürke vonalak.) Ez ugyanis lehetővé tesz valami nagyon fontos dolgot: annak racionális meghatározását, hogy egy vizsgálatba hány beteget kell bevonni! Megmondjuk, hogy milyen nagyságú hatást mekkora valószínűséggel akarunk kimutatni (azaz mekkora legyen az erő), és a képletből kipotyog, hogy ehhez hány betegre van szükség! Természetesen csak a feltételezett hatásra vonatkozóan, ha a valóságban kisebb a hatás, akkor gyengébb lesz a kísérletünk, de ha nagyobb, akkor erősebb.

(Kitérő megjegyzés: első ránézésre esetleg meglepő lehet, hogy a kiszámításhoz szükség van a gyógyszer hatására. Hiszen épp azért csináljuk a kutatást, hogy megmondjuk mekkora a hatás, akkor hogyhogy ezt előre meg kell adnunk?! – mondhatja valaki. Amellett, hogy a fentiekből látszik, hogy erre szükség van, hiszen a hatás nagyságától alapvetően függ az erő, második ránézésre talán logikus is: ez olyan, mint az, hogy egy csillag fényességének megfigyeléséhez mekkora távcsövet vegyük. Muszáj előzetesen feltételeznünk valamit a csillag fényességéről, mert ha fényes, akkor elég kis távcső is, de ha nagyon halvány, akkor nagy távcsővel kell készülnünk. Pontosan ugyanez a helyzet itt is, a csillag fényessége a gyógyszer hatása, a távcső nagysága pedig az, hogy hány beteget vonunk be a kísérletbe.)

Lássuk tehát az összefüggést, ezt mutatja a 5.9. ábra. (A két tengelyen a kezelt- és a kontrollcsoportbeli arány van, a színezés pedig a betegszámot adja meg. A színezés logaritmikus, tehát egy árnyalattal sötétebb szín egy nagyságrenddel nagyobb beteglétszámot jelent. A piros vonal 10 ezer fő, az ezen belüli esetekben kell 10 ezer főnél is nagyobb beteglétszám a 80%-os erővel történő kimutatáshoz.) Az ábra univerzális: akkor is használható, ha az arány valami jó dolog, amit emelni akarunk (pl. gyógyulók aránya) vagy valamilyen káros dologról van szó, például az expozíció egy káros behatás, esetleg gyógyszer, de a mellékhatását vizsgáljuk; ezekben az esetekben az exponált (kezelt) csoport aránya a kontrollnál nagyobb. De akkor is jó, ha az arány valami rossz dolog, például megbetegedés, amit elkerülni akarunk, ekkor az exponált (kezelt) csoportban az arány reményeink szerint alacsonyabb.



5.8. ábra. Az erő függése a kutatásba bevont betegek számától.



5.9. ábra. A 80%-os erő eléréséhez szükséges betegszám (5%-os szignifikanciaszinten).

Gyönyörűen látszik (és az előzőeket végiggondolva remélhetőleg logikus is!), hogy mi a könnyű és mi a nehéz. Egy gyógyszer hatásának a példáján: legkönnyebb akkor a helyzetünk, ha kezelés nélkül szinte senki nem gyógyul meg, kezeléssel pedig szinte mindenki (bal felső sarok környéke). Hatalmas a hatásnagyság, akár 10 beteg is elég a kimutatásához. Ha a gyógyszer hatása továbbra is kitűnő, de kezelés nélkül is sokan meggyógyulnak (haladunk a bal felsőből a jobb felső sarokba), akkor nehezedik a helyzet, 100 betegre is szükség lehet. Nem arról van szó, hogy a gyógyszer rosszabb, a probléma oka, hogy a különbség kisebb – márpedig a hatás kimutatásához a különbség az érdekes! A legnehezebb a helyzet akkor, ha ráadásul a gyógyszer sem átütően hatásos, azaz csak kicsit emeli a gyógyulók arányát, ez esetben hirtelen eldurvul a helyzet, és több ezer, sőt tízezer betegre, vagy annál is többre lehet szükség. Ezt még jobban kiemeli a piros vonal, mely a tízezres beteglétszámot jelenti: egy kísérlet esetében ez nagyjából a gyakorlati elvégezhetőség határa – az ezen belül eső hatások tehát kísérletben nagyon nehezen kimutathatóak. Ha a mellékhatásokra gondolunk, akkor ugyanez a tanulság azt jelenti: egy mellékhatás kimutatása akkor könnyű, ha a gyógyszer nagyon megnöveli az arányát.

Egy dologra fontos ezek kapcsán felhívni a figyelmet, mert az ábrán talán nem egyértelmű első ránézésre: a kimutathatóság a két arány különbségén múlik – nem a hányadosán. Ha az egyik arány 10% ponttal nagyobb, mint a másik, akkor (többé-kevésbé) állandó beteglétszám kell a kimutatásához, de ha a hányadosuk adott, például a gyógyszer 10%-kal növeli meg egy nemkívánatos esemény előfordulási valószínűségét, akkor a kimutathatóság attól is függ, hogy honnan indulunk, azaz mi a gyógyszer nélküli arány. Hiszen 50%-ról indulva a 10% növekedés az 5% pont, de 0,5%-ról indulva már csak 0,05% pont. Azaz: még ha a gyógyszer állandó arányban növeli is a mellékhatás kockázatát, kísérletben nagyon nehéz kimutatni, ha egy eleve is ritkán jelentkező betegségről van szó.

És most emlékezzünk vissza, mit mondtunk annak idején! Hogy a kísérletek egyik korlátja, hogy „...korlátozott a bevonható betegek köre ... a nem elegendően nagy mintanagyság korlátozza, hogy milyen nagyságú hatást tudunk észrevenni, legyen szó akár kívánt hatásról, akár mellékhatásról, ha például egy gyógyszerről beszélünk ... ha kicsi a mintanagyság, akkor egy kis javulást, vagy egy ritkán jelentkező mellékhatást nincs sok esélyünk észrevenni”. Pontosan ezt látjuk most!

6 Végpont és lemérése

Megbeszéltük, hogy a végpont a számunkra érdekes kimenet, például, hogy a beteg infarktust kap. De hogyan lehet ennek a kockázatát számszerűsíteni, és hogyan lehet gyógyszer ezt csökkentő (vagy ártalmatlan expozíció ezt növelő) hatását egyetlen számban kifejezni?

A hatás számszerű „lemérése” ugyanis ezt jelenti: hogyan tudjuk egy számba sűríteni azt, hogy az adott expozíció hogyan változtatja a kockázatot? Az egyszerűség kedvéért szorítkozzunk most egyetlen esetre, mely a gyakorlatban is az egyik legfontosabb és egyúttal az általános jelenségek is jól illusztrálhatóak rajta: arra, ha a végpont két lehetséges kimenetet vehet fel, úgy is szokták mondani, bináris vagy dichotóm. Ilyen az, hogy az alany meghalt-e adott időn belül, rákos lett-e, mentális beteg lett-e, és így tovább. Hogy egyéb problémákkal ne bonyolítsuk a kérdést, mondjuk, hogy két összehasonlított csoport van (exponált és nem exponált: gyógyszert kapó a gyógyszert nem kapóval szemben, légszennyezetterméken elő a nem légszennyezetterméken élővel szemben és így tovább), valamint, hogy nincs confounding, egy tökéletesen kivitelezett kísérletet végeztünk.

Az első megállapítás, hogy ebben az esetben a kockázat egy arány: a csoportból az alanyok mekkora hányada érte el a végpontot, például hány százalék kapott adott időn belül infarktust. A „hatás lemérése” tehát azt jelenti, hogy két arányt kell összevetnünk: ha 100 exponáltból 1 érte el a végpontot, akkor 1% az egyik arány, ha a nem exponált 200 főből 4, akkor 2% a másik. Na de mit értünk összevetés alatt? Ennek megválaszolása mondja meg, hogy hogyan mérjük le a hatást. Az érdekes az, hogy még ebben az elképzelhető legegyszerűbb helyzetben (egyetlen bináris végpont, két csoport) is nagyon nem nyilvánvaló problémákra vezet ez a kérdés...!

6.1. Mikor adjunk gyógyszert?

A konkrétság kedvéért mondjuk, hogy egy koleszterinszintet csökkentő gyógyszert vizsgálunk, a végpont, hogy az alany infarktust kap-e adott időn belül. Azt tapasztaljuk, hogy a gyógyszerrel kezelt csoportban egy év alatt 1% kap infarktust, a gyógyszert nem kapó kontrollcsoportban 2%. Akkor most mit mondunk?

- A. A fantasztikus gyógyszerünk 50%-kal csökkenti az infarktus-rizikót!
- B. A fantasztikus gyógyszerünk nélkül 100%-kal nagyobb az infarktus-rizikó!

C. Ezzel a gyógyszerrel 100 embert kezelve 98-at feleslegesen kezelünk (hiszen gyógyszer nélkül sem kapna infarktust), 1-et hiába kezelünk (hiszen a gyógyszerrel együtt is infarktust kap), és 1 az, akinél elérünk valamit. Miközben mind a 100-at kitesszük a mellékhatások kockázatának és mind a 100-zal kifizettetjük a gyógyszert.

Ugye mennyire máshogy hangzik? Pedig csak osztani kell tudni, hogy lássuk: ez a három igazából *ugyanaz!*

A fenti példák rámutatnak a hatás lemérésnek két alapvető lehetőségrére: arra, ha a két arányt elosztjuk egymással (ezt szokás relatív rizikónak nevezni), és arra, ha kivonjuk őket egymásból (abszolút rizikó-különbség). Ha osztunk, akkor kapjuk az első két megfogalmazást (pl. $1\% / 2\% = 0,5$, amit úgy is kifejezhetünk, hogy -50%), ha kivonunk, akkor a harmadikat ($2\% - 1\% = 0,02 - 0,01 = 0,01 = 1\%$ pont, ami épp az 1 a 100-ból).

Az első tanulság a fentiekből, hogy a racionális gyógyszerelés alapját az abszolút mutató jelzi. Szemben azzal, amit esetleg elsőre gondolhatna az ember, hogy ti. a gyógyszer adásának mérlegelésekor a gyógyszer mellékhatásait kell a megelőzni kívánt végpont szövődményeivel összevetni, a fenti mutatja, hogy valójában a gyógyszer mellékhatásait *be kell szorozni 100-zal*, és úgy hasonlítni az infarktushoz! Hiszen 100 embert kell kitenni ezeknek ahoz, hogy egyetlen infarktust megelőzzünk. Egy következmény azonnal látható: ritka betegségek megelőzésére csak nagyon biztonságos gyógyszerek használhatóak. (A nem megelőző, hanem gyógyító jellegű készítmények pedig végképp előnyben vannak ilyen szempontból, hiszen ott nem kell arra tekintettel lenni, hogy kezelés nélkül sem biztos, hogy baja lesz az alanynak.) A klinikai döntéshozatal szempontjából tehát az abszolút hatás a mérvadó. De akkor miért használjuk egyáltalán a relatív mutatókat?

6.2. Kísérletek résztvevői, avagy végre kiderül, hogy jót tesz-e repülőgépből kiesésnél, ha van nálunk ejtőernyő

Kezdjük a kérdést egy picit messzebbőről! Korábban már volt róla szó, hogy a klinikai kísérleteknek, azon hatalmas előny mellett, hogy teljes mértékben védettek tudnak lenni a confo- undinggal szemben, három, helyzettől függően kisebb vagy nagyobb hátrányuk van. Az egyik, hogy korlátozott az elérhető mintanagyság – emiatt kicsi hatásokat (kis mértékű, vagy keveseket érintő hatásokat), legyen szó akár pozitív hatásról, akár mellékhatásról, nem tudunk észrevenni. A második, az előzőhez nagyon hasonló limitáció, hogy korlátozott az utánkövetési idő, ezért a lassan fellépő hatásokat nem tudjuk észrevenni.

A harmadikról azonban eddig még nem beszélünk részletesen: arról, hogy a klinikai kísérletekben részt vevő betegek jellemzői szinte mindenkor előternekk, néha nem is kicsit, a betegek összességének jellemzőitől. Nevesítve: a klinikai kísérletekben részt vevő betegek legtöbbször fiatalabbak, mint általában véve a betegek, több közöttük a férfi, ritkábbak a társbetegségek. De miért van ez így?

A jelenségek vannak jóhiszemű és kevésbé jóhiszemű magyarázatai. Az első szűrő kapásból az, hogy bár a történelemben ez sajnos nem volt mindig így, de manapság már a klinikai kísérletekben kizárolag önkéntesek vesznek részt. Ez azt jelenti, hogy még ha a felkért betegek tökéletesen meg is felelnek összetételben az összes betegnek, azok akik *vállalják* a részvételt már rögtön nem fognak, hiszen az önkéntesség önmagában jelent eltérő jellemzőket (például a férfiak inkább hajlandóak kipróbalni ilyen kockázatos dolgokat). Ezzel pedig lehetetlen bármit is kezdeni, hiszen az az önkéntesség megsértését jelentené. A valóságban ráadásul a felkért betegek köre már önmagában is eltérő lesz: az ilyen klinikai kísérleteket tipikusan nagy, magas szinten lévő, városi centrumok végzik; az önmagában sokszor szűrés, például szocioökonómiai státusz szerint, hogy egyáltalán ki az, akit ilyen centrum lát el. A végeredmény az, hogy simán előfordulhat, például kardiologiában, hogy egy kísérletben a – „standard ellátást” kapó – kontrollcsoport halálozása fele az országos adatnak! Mindezek tetejébe jönnek a bevonási és kizárási kritériumok, ezek határozzák meg, hogy milyen betegek vehetnek részt a kísérletben, és kik azok, például életkor, társbetegségek, körrelőzményi adatok vagy épp súlyosság szerint, akik nem. Itt megjelennek a kevésbé jóhiszemű szempontok is: a szponzornak (a szakzsargonban így szokták hívni a kísérlet finanszírozóját) gyakran érdeke, hogy olyan betegek kerüljenek be, akiknek a legjobb a gyógyhajlama, hogy a vizsgált szer a legjobb színben tűnhessen fel. Erre pedig a bevonási és kizárási kritériumok meghatározásán keresztül lehet ráhatásuk.

Csak egyetlen példa mindezek eredőjének illusztrálására: Travers és munkatársai egy emlékezetes 2007-es cikkükben véletlenszerűen kiválasztott, „való életbeli” asztmás betegeknél néztek meg, hogy milyen gyógyszereket szednek a betegségükre, és hogy azokat a gyógyszereket milyen bevonási és kizárási kritériumú kísérletek alapján törzskönyveztek. Ezután összevetették a betegek adatait e kritériumokkal, és megnézték, hogy mekkora hányaduk vehetett volna részt az egyes kísérletekben; íme a százalékok: 5, 7, 6, 6, 0, 4, 2, 1, 7, 8, 7, 36, 2, 1, 1, 2 és 3. Látható, hogy *egyetlen* kísérlet volt, ahol legalább a kétszámjegyű százalékot sikerült elérni, egyébként az 1-3% a tipikus, de a legjobb, hogy olyan kísérlet is volt, aminél konkrétan nem találtak elő beteget, aki jogosult lett volna részt venni benne... Amiben ugye az a vicces, hogy utána a betegeket kezelik azokkal a gyógyszerekkel, amiket ilyen kísérletek alapján törzskönyveztek!

Ezek az arányok egyáltalán nem kiugróak, számos területen tapasztalható, hogy éves nagyságrendben kell várni, hogy elég beteg összegyűljen, mivel 80, 90 vagy annál is nagyobb százalékok nem jogosult a részvételre. Ezt a jelenséget hívják *szelekciós torzításnak*; szokás beszálni a kísérlet *kiülső validitásáról* vagy *általánosíthatóságáról* is, hiszen a probléma az, hogy bármit is találunk a kísérletbe bevont alanyok csoportján *belül*, az vajon mennyire vonatkoztatható az összes betegre *általában*. Így már talán még jobban érthető, hogy miért mondta annak idején, hogy ez a megfigyeléses vizsgálatok egyik előnye: ott sokkal kevésbé kell aggódnunk azon, hogy az eredmények mennyire vonatkoztathatóak az összes betegre, hiszen nem ritka, hogy akár az összes beteg is bevonható a vizsgálatba.

Ha már a korábban mondottaknál tartunk: talán emlékszik a nyájas Olvasó, hogy a tudomány jelen állása szerint nem tudhatjuk „biztosan”, hogy jót tesz-e, ha van nálunk ejtőernyő amennyiben kiesünk egy repülőgépből, legalábbis ha a biztosan alatt azt értjük, hogy „kísérlettel megvizsgálva”. Hivatkoztam is Smith és szerzőtársa cikkére, mely az igen tekintélyes

British Medical Journal 2003-as karácsonyi számában jelent meg, és amelyikben rendkívül alapos kutatással feltárták, hogy egyetlen egy ilyen kísérletet sem végeztek! (Nyilván gúnyolódva azokon, akik a kísérleteket mindenhatónak állítják be.) Nos, jelenthetem, hogy az orvostudomány fejlődése megállíthatatlan, ugyanis a helyzet azóta megváltozott! A British Medical Journal-ben a napokban, egész pontosan december 13-án, természetesen csak teljes véletlenségből megint a karácsonyi különszámban megjelent a történelem első kísérletes vizsgálata, amely az ejtőernyő hatásosságát vizsgálta! A kísérlet alanyai tökéletesen randomizáltan kaptak vagy ejtőernyőt, vagy ejtőernyőt nem tartalmazó hártsákat, ezt követően kiugrottak a repülőgépből, majd a kutatók rögzítették a földbe csapódáskor fellépő halálozások, illetve súlyos sérülések (a traumatológiában általánosan használatos ISS sérüléssúlyossági pontszám 15-nél nagyobb) fellépését. A kísérlet igen gondos tervezésű volt, az előző részben látott módon határozták meg a mintanagyságot, rögzítettek számos fontos betegjellemzőt, még arra is ügyeltek, hogy a felhasznált ejtőernyők, sőt, hártsákok típusát és gyártóját dokumentálják, beszerezték az etikai engedélyt a kísérletre stb. Egyszóval egy minden elvárásnak megfelelő kutatásról van szó. Hadd fassak előre: a vizsgálat szerint az ejtőernyő nem csökkentette a földet éréskor bekövetkező halálozások és súlyos sérülések számát!

Hogy ezt miért pont most, a klinikai kísérletek résztvevőinek speciális összetételéről szóló résznél mondjam? Ehhez érdemes közelebbről megnézni a kísérlet pontos lefolytatását!

A részvételre felkért alanyok két csoportból kerültek ki: egy részüket sugárhajtású repülőgépen repülés közben interjúvolták meg, hogy vállalják-e a kiugrást randomizáltan ejtőernyővel vagy hártsákkal, más részüknek repülőgép-múzeumban, egy földön álló kisrepülőben tették fel ugyanezt a kérdést (**6.1.** ábra; fontos megjegyezni, hogy a képen látható alany nem halt meg, illetve nem szenvedett súlyos traumás sérülést a földbe csapódáskor.). A fent említett eredményhez talán azt a mellékes információt érdemes hozzátenni, hogy az előbbi csoportból 0% vállalta a részvételt, míg az utóbbiak közül 100% (erre nincs ráhatásunk, önkéntesség, ugyebár!), így apróbb eltérések keletkeztek a klinikai kísérletbe bekerülő és be nem kerülő alanyok között: az utóbbiak esetében a repülőgép átlagos sebessége 800 km/h volt, az előbbieknél 0 km/h, az utóbbiaknál az átlagos ugrási magasság 9146 méter, az előbbieknél 60 centiméter...

Az említett eredmény úgy jött ki, hogy mind az ejtőernyős csoportban, mind a kontrollcsoportban 0 halálozás, illetve súlyos sérülés fordult elő. Tehát: nincs különbség...

Nagyon fontos újra hangsúlyozni, hogy a kísérlet szervezői mindenféle kockázatnak kitett alanyt *igyekeztek* verbuválni, arról már nem lehetnek a szerzők, hogy történetesen a részvételt vállalók köre „némileg” speciálisra sikeredett – és pont arra akarják felhívni a figyelmet, hogy ez egy valódi klinikai kísérletben is előfordulhat. Zárásként adjuk vissza a szót a szerzőknek: „a magas kockázatnak kitett alanyok részvételének a hiánya elképzelhető, hogy befolyásolta a vizsgálat végeredményét”.



6.1. ábra. Reprezentatív példa a kísérletben résztvételt vállaló alanyra (kontroll – hátizsák – csoport).

6.3. Mutatók stabilitása

Egy doleg azonban még mindig lög a levegőben: mi köze ennek a relatív és abszolút mutatókhoz? Bármilyen furcsa is lehet elsőre, de nagyon sok: ha relatív mutatót használunk, az az egész fenti problémát sok esetben meg tudja oldani, vagy legalábbis jelentősen enyhíti!

A klinikai kísérletekben fiatalabbak az alanyok, kevésbé súlyos az állapotuk, kevesebb társbetegségiük van? Igen. Emiatt jobb a gyógyhajlamuk, mint a betegeknek általában? Igen. Csakhogy mi, ha relatív mutatót használunk, akkor nem is ezt nézzük, hanem azt, hogy *egymáshoz képest* hogyan viselkednek a kezelt- és kontrollcsoportok! Igen, fiatalabbak, de a kezelt és a kontrollcsoport *egyaránt* fiatalabb, márpédig őket *egymáshoz* hasonlítjuk! Lehet, hogy a klinikai kísérletben 10% a halálozás a kontrollcsoportban, míg a valóságban 20, de ha ez 8-ra megy le a kezelés hatására, akkor reménykedhetünk benne, hogy a 20 meg 16-ra fog. Igen, a klinikai kísérlet betegeinek összetétele eltérő volt, ezért a halálozási arányok is mások voltak, de a *relatív viszonyok* állandóak! Ha ez igaz, tehát a relatív mutató stabil, akkor onnantól nem is annyira számít, hogy a klinikai kísérlet betegei tényleg speciális populációt jelentenek-e, hiszen mi úgysem az abszolút számokat fogjuk felhasználni, hanem a relatív viszonyokat – ami viszont a nem speciális populációra is érvényes.

De tényleg stabilak a relatív mutatók? A tapasztalatok szerint igen! Fontos előrebocsátani, hogy ez nem valamiféle matematikai törvényszerűség, és nincs is rá garancia, hogy minden teljesüljön (épp emiatt még ennek fényében is igenis hasznos, ha a klinikai kísérlet alanyai nem nagyon speciálisak!), de *nagy általánosságban* véve a relatív mutatók meglepően stabilak. Erre mutat példát a 6.1. táblázat egy koleszterinszintet csökkentő gyógyszercsalád néhány kísérletének példáján keresztül. (A táblázat a szív-érrendszeri eredetű halálozások arányát mutatja a kísérlet utánkövetése alatt. Az utolsó oszlop az utánkövetési idő hossza, átlag vagy medián, függően attól, hogy a tanulmány mit közölt.)

6.1. táblázat. Különböző kísérletek, melyben a sztatinnak nevezett koleszterinszint-csökkentő készítmények hatását vizsgálták.

Kísérlet neve	Kontrollcsoport rizikója	Relatív rizikó	Abszolút rizikó-különbség	Utánkövetés hossza [év]
JUPITER	0,48%	0,81 (-19%)	0,09 %pont	1,9
AFCAPS/TexAS	0,78%	0,68 (-32%)	0,24 %pont	5,2
ASCOT-LLA	1,60%	0,90 (-10%)	0,16 %pont	3,3
WOSCOPS	2,22%	0,68 (-32%)	0,70 %pont	4,9
CARE	6,26%	0,86 (-14%)	0,87 %pont	5,0
HPS	9,13%	0,83 (-17%)	1,52 %pont	5,0
4S	9,31%	0,66 (-34%)	3,19 %pont	5,4
LIPID	9,62%	0,76 (-24%)	2,28 %pont	6,1
PROSPER	10,06%	0,86 (-14%)	1,38 %pont	3,2

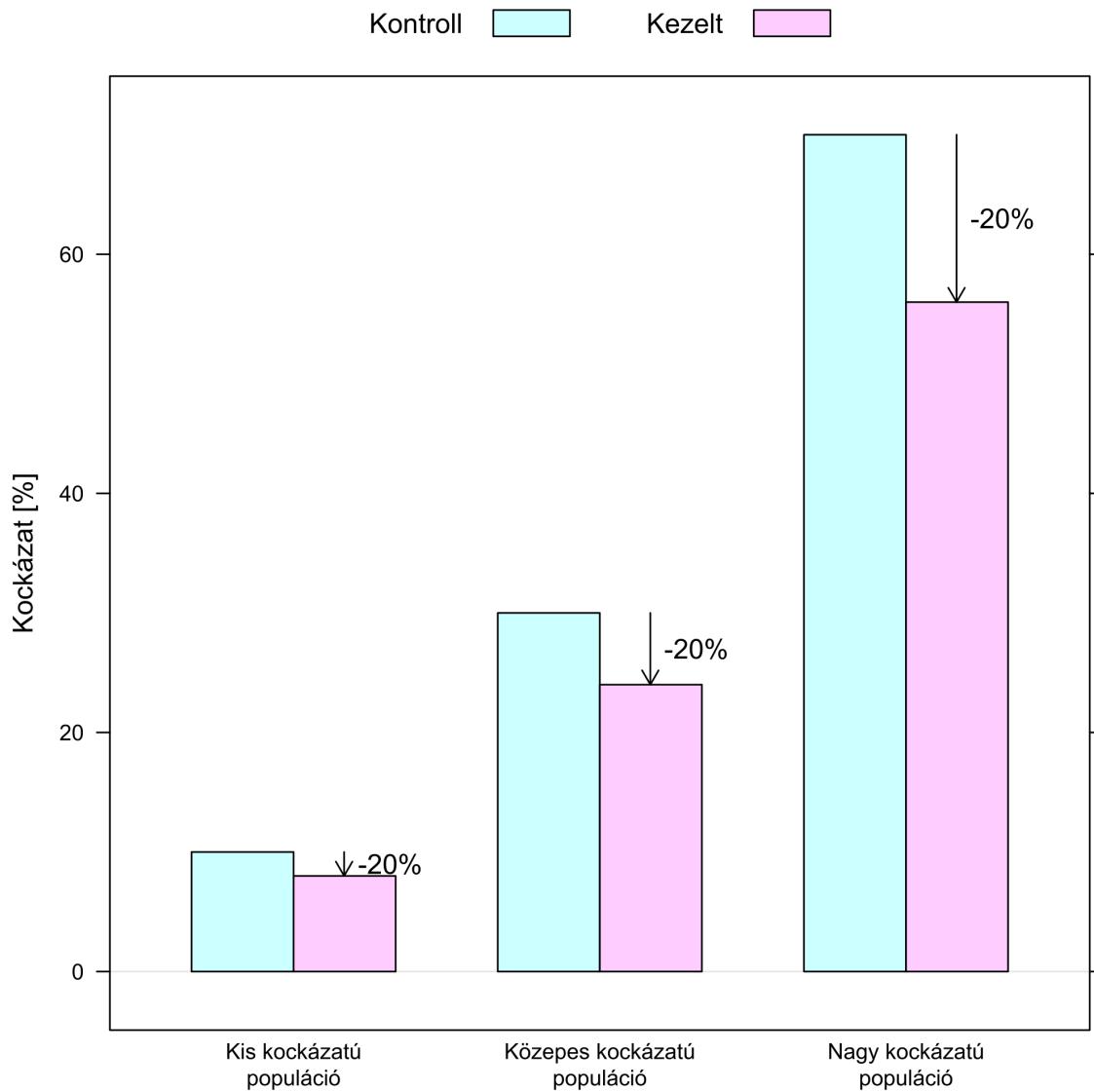
Gyönyörűen látható, hogy a gyógyszereket nagyon-nagyon különböző populációkban próbálták ki: volt, ahol csak 0,48% halt meg szív-érrendszeri okból a kontrollcsoportban, de volt, ahol több mint húszszor ennyi. (Ez a kezelés nélküli rizikó, ez jellemzi tehát, hogy milyen alanyok körében végezték a kísérletet.) Az abszolút rizikó-különbség ennek megfelelő drámai eltéréseket mutat, a legkisebb és a legnagyobb között több mint harmincötösök a különbség. Igen ám, de – és most jön a lényeg – minden közben a relatív hatás bámulatosan állandó, 10 és 30% közötti kockázatcsökkenés látható *függetlenül* attól, hogy milyen kockázatú populációban végezték a vizsgálatot! Az elsőként felsorolt kísérletet olyan populációban végezték, hogy kezelés nélkül fél százalék halt meg, az utolsót olyanban, ahol több mint 10, de a gyógyszer *relatív* hatása alig tér el!

Mindezeket úgy is elmondhatjuk: a jelek szerint a gyógyszerre saját magára jellemző tulajdon-ság a relatív hatás, az az, ami állandó. Az abszolút hatás egy származtatott mutató, egy eredő: a (gyógyszerre jellemző) relatív hatás, és az (adott populációra jellemző) kockázat szorzata. A relatív hatás állandó, az abszolút hatás attól függ, hogy mennyi a kezelés nélküli kockázat: ahol nagy (pl. idős, sok társbetegséggel rendelkező beteg), ott az abszolút csökkenés is nagy lesz, ahol kicsi, ott kicsi ([6.2. ábra](#)).

Ennek két nagyon fontos következménye van. Az egyik, hogy egy kísérlet eredményének megadásakor igenis jogos a relatív mutató használata, hiszen egy gyógyszerkísérletben értelemszerűen azt kell kimérni, ami magára a gyógyszerre jellemző. A másik, hogy ez persze nem változtat azon, hogy a klinikai döntéshozatal szempontjából az abszolút különbség a mérvadó. Sőt, ez rögtön érthetővé teszi, hogy miért van az, hogy egy fiatal, terhelő kórelőzmény nélküli, egyébként egészséges betegnek lehet, hogy nem ilyen gyógyszert fog felírni az orvos, míg egy idős, korábban szívinfarktuson átesett, cukorbeteg páciensnek igen – nem azért, mert azt gondolná, hogy az előbbi esetben nem hat a gyógyszer, az utóbbi esetben viszont igen. Könnyen lehet, hogy *pontosan ugyanúgy* hat a gyógyszer, azaz pontosan ugyanúgy 20% kockázatot csökkent, csakhogy ez a -20% az előbbi esetben, mivel alacsonyról indulunk, nagyon kis abszolút kockázatcsökkenés (és így a gyógyszer mellékhatásai nagyobb súlyval esnek latba), míg az utóbbi esetben fordított a kockázat/haszon mérleg.

A kettőt összerakva láthatjuk, mi a helyes eljárás: a kísérletben azt kell kimérni, ami stabil és ami a gyógyszerre jellemző, aztán ezt az információt a konkrét klinikai alkalmazásban *kontextusba kell helyezni*. Azaz: a – kísérletből ismert, gyógyszerre jellemző – relatív mutatót az adott konkrét beteg jellemzői, például társbetegségei vagy életkora alapján át kell számolni abszolút mutatóra... és ez alapján dönten!

Most, hogy minden értünk, némileg a legelső példa is újraértekelhető. Abból ugyanis nem került ki túl jól a gyógyszer (100 beteget kellett kezelni egy infarktus elkerüléséhez), de nézzük csak meg a számokat: 2% infarktusrizikó még kezelés nélkül is? Miközben Magyarországon 15 ezer ember kap infarktust – minden egyes évben! Ez meg hogyan lehet? Az 1 és 2% persze nyilván kerek szám volt a példa kedvéért, de ha megnézzük a táblázatot, nagyságrendileg nem tévesek, tényleg van számos nemzetközi kutatás ilyen számokkal (pedig máshol sem sokkal kisebb az infarktus-rizikó). A választ akkor kapjuk meg, ha ránézünk a táblázat jobb szélső oszlopára is: e kutatások utánkövetési ideje mindössze néhány év volt! Az infarktus-rizikó



6.2. ábra. Ugyanolyan (20%-os) relatív csökkenés abszolút hatása attól függ, hogy honnan indulunk.

azonban nemhogy több év, hanem inkább évtizedek alatt épül fel, e gyógyszereket is ilyen távon szedik igazából a betegek. E kutatásokat tehát nem lehet a valós helyzetre közvetlenül rávettíteni... illetve nem lehetne, ha nem lenne a relatív rizikó! Ugyanis a kis kockázatú populáció nem csak azt jelentheti, hogy fiatal meg nem cukorbeteg, hanem azt is, hogy kevés ideig utánkövetett, a nagy kockázatú meg nem csak az idős és cukorbeteg lehet, hanem a realisztikus ideig utánkövetett. Az előző megállapításunk tehát azt mondja ez esetben, hogy a kísérletből ne az abszolút kockázatot olvassuk ki, hiszen az rövid utánkövetésre vonatkozik, nem a valóságra. Olvassuk ezzel szemben ki a relatív kockázatot, hiszen az stabil (reményeink szerint a különböző utánkövetési időkre nézve is!), és azt használjuk: számítsuk át, hogy mi történne hosszabb, azaz a valóságnak megfelelő utánkövetési idő alatt. A példa 1 éves utánkövetést írt; kiszámítható, hogy 5 év alatt a kontrollcsoport kockázata már 9,6%, nem 2%. Ha a -50% állandó marad, akkor az abszolút különbség már 4,8 %pont, nem 1 %pont – 21 embert kell kezelni egy infarktus megelőzéséhez. 10 év alatt a kockázat 18,3% (látszik, hogy szépen közeledünk a valós kockázatokhoz!), a gyógyszer abszolút hatása 9,1 %pont csökkenés – 11 embert kell kezelni egy infarktus megelőzéséhez, azaz már csak 11-gyel kell szorozni a gyógyszer mellékhatásait a kockázat/haszon mérlegelésnél. Már más hogy hangzik!

6.4. Döntéshozatal és az abszolút és relatív mutatók

Vajon hogyan használhatók a különböző mutatók arra, hogy jobban megértsük, mi a helyes döntés egy beteg ellátása során? A statisztika természetesen nem tudja átvenni az orvosi döntéshozatal szerepét, de hatalmas segítséget adhat annak megalapozottabbá tételeben, azáltal, hogy a választási lehetőségeket jobban megértjük.

Érdemes a kiindulópontunkat felidézni: egy gyógyszer az infarktus kockázatát 2%-ról 1%-ra csökkenti. Ha azt mondjuk, hogy „50%-kal csökkenti az infarktusrizikót”, akkor relatív mutatóval írtuk le a hasznosságát, ha azt mondjuk „1%ponttal csökkenti az infarktusrizikót”, akkor abszolút mutatót használtunk. Megbeszéltük, hogy az előbbi mutatót érdemes klinikai kísérletben kimérni, mert stabilabb: azt reméljük, hogy az 50%-os csökkenés az, ami magára a gyógyszerre jellemző, és így állandó marad akkor is, ha a kezelés nélküli kockázat nem 2%. Ez tehát azt jelenti, hogy a relatív mutatók használata lehetővé teszi a klinikai kísérlet eredményeinek az általánosítását: ha a rendelónkben ülő beteg kockázata nem 2% hanem 10%, akkor is tudjuk, hogy mire számíthatunk – arra, hogy ezt 5%-ra csökkenti a gyógyszer.

Az abszolút mutatóról azt mondtauk, hogy ez fontos a klinikai döntéshozatalhoz, például annak eldöntéséhez, hogy adjunk-e ilyen gyógyszert a betegnek. Ennek logikáját érdemes részletesen is felidézni: a fenti számok azt jelentik, hogy ezzel a gyógyszerrel 100 embert kezelve 98-at feleslegesen kezelünk (hiszen gyógyszer nélkül sem kapna infarktust), 1-et hiába kezelünk (hiszen a gyógyszerrel együtt is infarktust kap), és 1 az, akinél elérünk valamit – miközben minden a 100-at kitesszük a mellékhatások kockázatának. Ez az „1 a 100-ból” épp az abszolút mutató 1%pontja. Az, hogy 100 beteget kell kezelnünk 1 végpont megelőzéséhez, azért nagyon fontos, mert így már látható, hogy mit kell mérlegelnünk, ha racionálisan akarunk dönteni

a gyógyszeradásról: azt, hogy mi a rosszabb, az infarktus szövődményei, vagy a gyógyszer mellékhatásai *beszorozva 100-zal!* Érhető tehát, hogy miért mondhatjuk, hogy ez mutatja a gyógyszer klinikai előnyét, miért ez a releváns a döntéshozatalhoz: ha a betegünk kockázata kezelés nélkül 10%, akkor az abszolút előny 5% pont, tehát nála már csak 20-szal kell szorozni a gyógyszer mellékhatásait a kockázat/haszon mérlegelés során.

Azonban nem ez az egyetlen terület, ahol az abszolút mutatók jól jönnek.

Elsőként kezdjünk egy kis ismétléssel; hogy látványosabb legyen a doleg, nézzünk egy konkrét, való életbeli példát! Lipson és szerzőtársai egy 2018 májusában megjelent cikkükben a krónikus obstruktív tüdőbetegség (továbbiakban röviden: COPD) kezelési lehetőségeit vizsgálták. A COPD, amely elsősorban a dohányosok betegsége, a légutak beszűkülezésével és a tüdő szövetének pusztulásával járó visszafordíthatatlan, folyamatosan súlyosbodó gyulladásos folyamat. Ebben a betegségen az egyik doleg, amitől félünk, az a nagyon hirtelen kezdődő, átmeneti állapotrosszabbodás. Ez sajnos időről-időre előfordul a COPD-s betegeknél, és egyszer – különösen a rosszabb állapotú alanyoknál – akár nagyon súlyos lefolyású is lehet, másrészről még ha nem is történik nagy baj, a beteg későbbi kilátásait akkor is rontja minden egyes ilyen epizód. Éppen ezért nem meglepő, hogy a gyógyszerek kezelések egyik fontos célja is ezen állapotromlások megelőzése. Az említett cikk két gyógyszert hasonlít össze: az egyik két szokványos hatóanyagot tartalmaz, a másik kiegészíti ezt egy harmadikkal, egy szteroiddal.

A kutatás egyik végpontja tehát az állapotromlások előfordulási gyakorisága volt; a hagyományos kezelést kapóknál 1,21 ilyen történt átlagosan egy évben, a tripla kombináció esetében ez lement 0,91-re.

Egyfelől tehát mondhatjuk, hogy az új gyógyszer kb. 25%-kal csökkenti az ilyen állapotromlások előfordulási gyakoriságát (relatív mutató), másrészről mondhatjuk, hogy évi 0,3 állapotromlást előz meg (abszolút mutató).

Első lépés: Nekünk a relatív mutató a lényeges eredmény a kísérletből. Ha a rendelőnkben épp velünk szemben ülő beteg esetében neme, életkora, társbetegségei stb. alapján mondjuk 2 állapotromlás várható évente a hagyományos kezeléssel, akkor arra számíthatunk, hogy ezt a gyógyszer átlagosan 1,5-re fogja csökkenteni (elfogadva, hogy a relatív mutató állandó). Hiába volt tehát más a konkrét betegünk kockázata, a relatív mutató használata lehetővé tette, hogy rá nézve is következtetést tudjunk levonni.

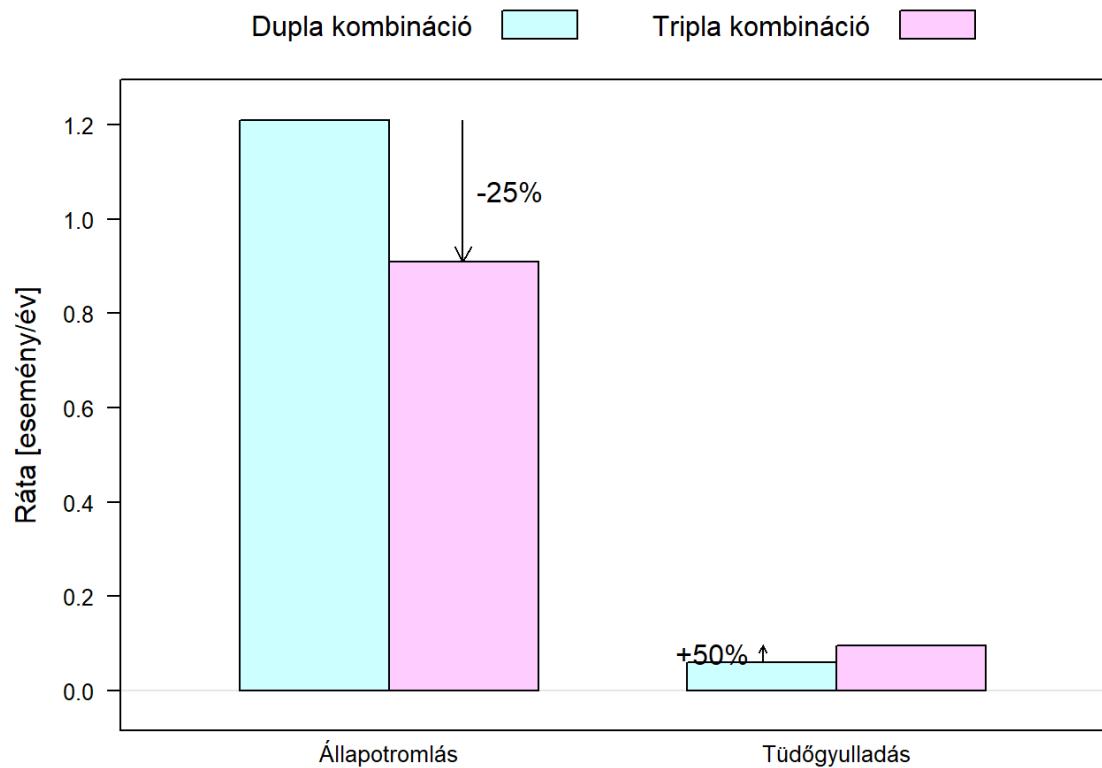
Második lépés: A gyógyszer abszolút előnye tehát ennél a betegnél 0,5 állapotromlás megelőzése évente. Itt is célszerűbb áttérni a fordított mutatóra, és azt mondani, hogy 2 évnnyi kezeléssel előünk meg egy állapotromlást.

A kérdés tehát így már egyértelmű: megéri-e két éven át kezeln a beteget (két éven át tartó kezelés mellékhatásainak kitenni, két évnyi gyógyszert kifizettetni) azért, hogy egy állapotromlást megelőzzünk? Ez a kérdés természetesen nem dönthető el statisztikai úton – azon múlik, hogy egyszerűt milyen súlyú az állapotromlás, másrészről milyen súlyúak a mellékhatások – de a statisztika segít abban, hogy világosan megfogalmazzuk, lássuk, és így jól meg tudjuk érteni, hogy egyáltalán mi a kérdés, milyen alternatívák között kell dönten.

6.5. Különböző kimenetek közös nevezőre hozása

De mi az a másik terület, ahol jól jönnek az abszolút mutatók? Ez azonnal világos lesz, ha nem csak a főhatással, hanem a mellékhatásokkal is elkezdünk számolni. A tripla kombináció egyik problémája, hogy a szteroid-tartalma miatt megnöveli a tüdőgyulladás esélyét: a kísérlet eredményei szerint a dupla kombináció esetében évi átlag 0,061 ilyen fordult elő, de az új szernél már 0,096.

Mondhatjuk, hogy kb. 50%-kal megnöveli a tüdőgyulladás kockázatát, ami nagyon hasznos mutató, ha más betegre vagy betegcsoportra akarjuk vetíteni ezt, tehát a kutatásból tényleg ezt kell kiolvasni, ahogy láttuk is, de nem sokat segít az összehasonlításban! Vegyük ugyanis észre, hogy a „25%-kal csökkenti az állapotromlás kockázatát” és az „50%-kal növeli a tüdőgyulladás kockázatát” egymással *totálisan összevethetetlen* kijelentések! Miért? Azért, mert nagyon más a kiindulási alap! Állapotromlásból évi 1 fordul elő, tüdőgyulladásból tizedannyi (6.3. ábra).



6.3. ábra. A kétféle végponton mért hatás nem összehomogén, ha a hatást relatív mutatóval mérjük, mert teljesen más a kiindulópontjuk.

Itt jön az abszolút mutatók másik előnye: megteremtik az összehasonlíthatóságot! A példa

kedvéért mondjuk, hogy a betegünk tüdőgyulladás-kockázata a régi kezeléssel 0,1 eset évente, akkor – elfogadva a relatív mutató stabilitását ebben is – az új kezelés mellett 0,15-re számíthatunk. A tripla kombináció tehát évi 0,05 tüdőgyulladást okozott, vagy – megint csak megfordítva a jobb érthetőség kedvéért – 20 évnyi kezeléssel okozunk mi, a gyógyszeradással egy többlet-tüdőgyulladást.

És akkor azonnal látható, hogy miért beszélhetünk az összehasonlíthatóság megteremtéséről: ha 2 év kezeléssel előzünk meg egy állapotromlást, de 20 év kezeléssel okozunk egy tüdőgyulladást, akkor egész egyszerűen azt mondhatjuk, hogy 10 állapotromlást előzünk meg 1 tüdőgyulladás okozása árán! (Az egyszerűség kedvéért vegyük úgy, hogy a gyógyszernek nincs más előnye mint az állapotromlás megelőzése és nincs más hátránya mint a tüdőgyulladás okozása.)

Megint csak: az természetesen nem statisztikai kérdés, hogy ez megéri-e, az azon múlik, hogy a tüdőgyulladás milyen súlyú az állapotromláshoz képest, de így legalább már látható, értelmesen, hogy egyáltalán mit kell összevetni – a relatív mutatókból ez nem derült ki!

6.6. Kitérő: patikamérlegen az emberélet

De tényleg nem statisztikai kérdés az előbbi? Talán nem tisztán az, de igenis lehet még szerepe a statisztikának, hogy segítsük az orvosi döntéshozatalt! Mondjuk, hogy egyetlen dologtól félünk minden az állapotromlás, minden a tüdőgyulladás kapcsán, ez pedig az, hogy a beteg belehal. (Sajnos csakugyan mindenkihez bele lehet halni, pláne egy rosszabb állapotú COPD-s beteg esetén.) Ekkor igenis továbbvihető a doleg számszerűsítése: nézzük meg, hogy mennyire a halálozási kockázat az állapotromlás és mennyire a tüdőgyulladás esetén! A befejezés innentől már egyértelmű: ha az utóbbi kockázat kevesebb, mint tízszer akkora, akkor megéri az új gyógyszer, ha nem, akkor a régivel jár jobban a beteg.

Első hallásra kicsit ijesztő lehet, hogy ilyen szikár számokra fordítjuk le emberek életét, de valójában a statisztikai adatokkal explicite nem támogatott orvosi döntéshozatal is *hajszálpon-tosan ugyanezeket* a megfontolásokat alkalmazza, legfeljebb implicite – akkor viszont már jobb, ha explicitté tesszük!

A probléma sokkal inkább az, hogy a döntési helyzet nem ennyire „egydimenziós”. Nem csak arról van szó, hogy a gyógyszernek nem az állapotromlás megelőzése az egyetlen előnye és a tüdőgyulladás okozása az egyetlen hátránya, hanem arról is, hogy a kimeneteket nem lehet egyszerűen arra szűkíteni, hogy a beteg meghalt-e vagy sem. Ezzel ugyanis azt mondjuk, hogy mindenki, aki nem halt meg az pontosan ugyanolyan állapotban van: nincs különbség között, hogy makkegészséges vagy mondjuk ágyhoz kötött, önellátásra képtelen beteg.

A doleg kézenfekvő továbbfejlesztési lehetősége tehát az, hogy valamilyen formában ezt az ún. egészségi állappittal összefüggő életminőséget is figyelembe vesszük. Mondjuk ha valaki 10 évet nyer a gyógyszer hatására, de ágyhoz kötött betegként, az kevesebbet ér, mintha 10 tökéletes egészségen töltött évet nyerne. Ezek a megközelítések matematikai szemmel ugyan

tetszetősek lehetnek, de etikailag nagyon problémásak. A legnagyobb gond, hogy valamilyen formában le kellene mérni azt, hogy a különböző életminőségek „mennyt érnek” a tökéletes egészséghez képest. Hány százalék életminőség romlás a tökéletes egészséghez képest az, ha vak vagyok? Ha süket? Ha ágyhoz között? Ha amputált? Ezekre a kérdésekre borzasztó nehéz válaszolni, és az is nagy kérdés, hogy van-e egyáltalán értelmük ilyen formában.

(Mindazonáltal kutatók intenzíven foglalkoznak ezekkel a problémákkal. Az egyik lehetőség, hogy megkérdeznek betegeket, hogy hány évnyi tökéletes egészségen töltött időre cserélnék le 10, adott állapotban töltött évüket. Ha 10-et mondanak, akkor tökéletes az életminőségiuk, ha 9-et, akkor 10% a betegségük életminőséget rontó hatása, ha 8-at akkor 20% és így tovább.)

6.7. Végpontok megválasztásának problémái

E kitérő után térjünk vissza a végpontokhoz, mert egy nagyon fontos kérdést még nem érintettünk. Eddig a lemérésről beszélünk, de van egy alapvetőbb kérdés is: a végpont megválasztása! Ez első hallásra elég felesleges kérdésnek hangzik (mit kell ezen megválasztani? – az a végpont, ami érdekel minket!), de valójában ez sem ilyen egyszerű kérdés.

A legfontosabb probléma az, hogy sok esetben, ha szó szerint vesszük azt, hogy „ami érdekel minket” akkor olyan kísérlethez jutunk, amit lehetetlen, vagy nagyon nehéz végrehajtani. Vegyünk egy példát! Miért adunk vérnyomás-csökkentő gyógyszert a magas vérnyomású betegeknek? Azért, hogy csökkentsük a stroke-kockázatukat, csökkentsük az infarktus-kockázatukat és így tovább. Akkor tehát mi legyen a végpont egy vérnyomás-csökkentő gyógyszerjelölt vizsgálatában? Természetesen a stroke-rizikó, az infarktusrizikó, és így tovább.

Ez elmondva nagyon logikus, de valójában van egy hatalmas problémája, amiről már esett is sok szó: az, hogy rendkívül nehéz lesz kimérni kísérletben a hatást! Az infarktus, pláne a stroke nem fordul elő túl sűrűn, ezért vagy hatalmas mintanagyságra lenne szükség, vagy nagyon hosszú után követésre (vagy leginkább mindenkor...), és láttuk, hogy pont ez a kettő problémás egy kísérletben.

A megoldás tehát az lesz, hogy azt mérjük, hogy a vérnyomás-csökkentő csökkenti-e a vérnyomást. (Hiszen az nagyon rövid idő után kimérhető, és mindenki kimérhető.) Ez így már-már módosan egyértelműnek is tűnhet, de vegyük észre, hogy valójában van mögötte egy rendkívül fontos háttérfeltételezés: az, hogy a normalizált vérnyomás *tényleg* kisebb szív-érrendszeri kockázattal jár együtt! Ez viszont már egyáltalán nem nyilvánvaló, és ezen belül is különösen fontos az, hogy *önmagában* az a tény, hogy a normális vérnyomású embereknek kisebb a szív-érrendszeri rizikója még *nem* jelenti azt, hogy a vérnyomás mérése jó megoldás! Ha ez így lenne, akkor egy fogfehérítő készítmény könnyen hatásosnak bizonyulhatna a tüdőrák megelőzésére (gondoljuk végig!).

Az ilyen mutatókat, melyek kevesebb betegen, illetve hamarabb is kimérhetőek, ám közben jól mutatják, hogy milyen lenne a hatás a – csak sok betegen, illetve lassan kimérhető – valódi végponton, szokás helyettesítő végpontnak nevezni. Hiszen a betegnek nem lesz attól

jobb, hogy egy műszer által kiírt szám valamilyen értéket mutat (a magas vérnyomás, kevés kivételtől eltekintve, önmagában semmilyen panaszt nem okoz), neki attól lesz jobb, ha nem kap infarktust. Fontos újra hangsúlyozni, hogy mindenekkel csak akkor jogos, ha tényleg igaz, hogy a helyettesítő végpont a fenti értelemben jó, tehát, hogy a csökkenése *tényleg* maga után vonja – ha kevesebb beteget érintően és később is – az igazi végpont javulását.

Összefoglalva, a helyettesítő végpontok szerepe hatalmas, hiszen lehetővé teszik, hogy a gyógyszerek hatását könnyebben megítéljük, vagy egyáltalán, képesek legyünk megítélni. Azt is látni kell azonban, hogy a helyettesítő végpontok használata „veszélyes üzem”, épp az előbb tárgyalt feltétel miatt. Az ideális eset az, ha a betegség valódi végpontra gyakorolt hatása teljes egészében átmegy a helyettesítő végponton, és a vizsgált beavatkozás a betegség és a helyettesítő végpont között hat. Sajnos több történeti példát lehetne hozni arra, hogy nagyon logikusnak tűnő helyettesítő végpontok bizonyultak katasztrofálisan rossznak.

Az egyik leghíresebb eset a CAST kísérlet, melyben azt vizsgálták a '80-as évek végén, hogy bizonyos kamrai szívritmuszavarokat megelőzni hivatott gyógyszerek (flekainid, enkainid és moracizin) hogyan hatnak a halálozásra infarktuson átesett betegek esetében. Az korábbról is ismert volt, hogy e gyógyszerek csakugyan csökkentik a kérdéses szívritmuszavarok előfordulását – márpedig az jól ismert, hogy ezek nagyban megnövelik a hirtelen szívhallál esélyét. A szívritmuszavar halálhoz vezethet, a gyógyszer megelőzi a szívritmuszavart, tehát a gyógyszer csökkenti a halálozási kockázatot. Tiszta sor, nem? Sajnos nem! A gyógyszereket törzskönyvezésük után évente 200 ezer beteg kezdte Amerikában szedni, mire az említett CAST kísérletben végre kiderült, hogy bár a ritmuszavarok előfordulását csakugyan csökkentik, de van velük egy apróbb probléma: a halálozást viszont *megnövelik!* A valószínűsíthető magyarázat, hogy az alapbetegség más úton is hat a hirtelen szívhallálra, nem csak a ritmuszavaron keresztül, és erre az útvonalra a gyógyszernek kimondottan *rossz* hatása volt. Kevesebb ritmuszavar, több halál – pedig ugye mennyire logikus volt az ellenkezője?

A helyettesítő végpontokkal tehát mindig óvatosnak kell lenni, és ellenőrizni kell, hogy a fenti feltételt teljesítik-e (szép szóval úgy szokták mondani: jól validált helyettesítő végpont-e).

6.8. Több végpont egyidejű vizsgálata

Mostanra alaposan kiveséztük a végpontok kérdését – azonban csak egyetlen végpontról beszélve! A valódi klinikai vizsgálatokban általában nem csak egyetlen kimenetel érdekel minket, az azonban további problémákat vet fel, ha több végpontot egyszerre kell figyelnünk.

Az egyetlen végpont összehasonlítása kapcsán megtárgyal számos kérdésből egy dolgot érdekes most felidézni: azt, hogy a szokásos (bár számos szempontból kritizálható – néhányról már e hasábokon is volt szó) eljárás szerint úgy határozzuk meg, hogy mikor mondjuk, hogy egy hatás nem pusztán a véletlen műve, hogy igaz legyen, hogy ha valójában nincs hatás, akkor 5%-os valószínűsséggel mondjuk mégis azt, tévesen, hogy van. (Miért 5% és nem kevesebb, ha egyszer mi mondjuk meg és ez egy tévedés valószínűsége? Lecsökkenthetnénk, de akkor

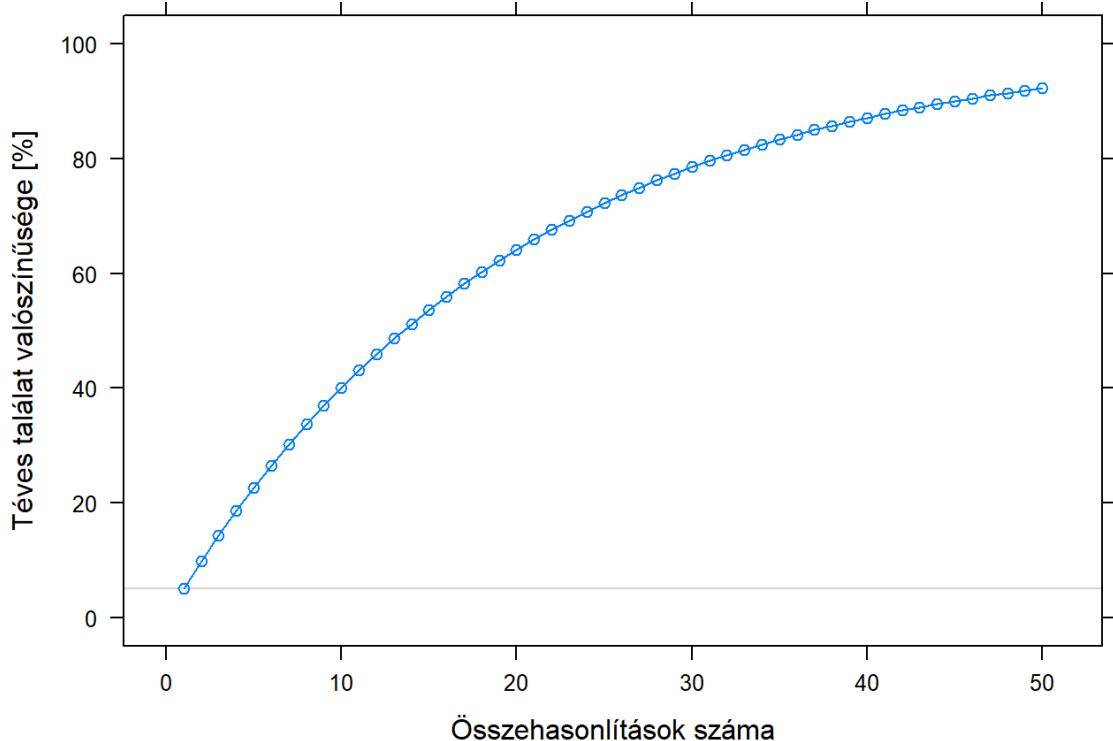
megnőne annak a valószínűsége, hogy a valódi hatásokra is azt mondjuk, hogy csak a véletlen miatt vannak.) A szokásos szóhasználat szerint ilyenkor fogalmazunk úgy, hogy „szignifikáns” a hatás, kifejezve azt, hogy ilyenkor már nem hisszük – de nem kizárt! – hogy a hatás a véletlen miatt van. Hibázhatunk, de erre van szükség ahhoz, hogy egyáltalán bármit tudjunk mondani.

6.8.1. Vadászni mentünk... szignifikanciára

Ez az 5% ilyen értelmű hibavalószínűség remekül működik akkor, ha egyetlen végpontunk van. Bonyolódik azonban a helyzet akkor, ha több végpontot hasonlítunk és azokat „vagylagosan” kezeljük, tehát akkor kiáltunk fel, hogy találtunk valamit, ha *bármelyik* végpontban van eltérés. Mondjuk, hogy a gyógyszer hatását nem csak azzal mérjük le, hogy csökkenti-e a szíviroham kockázatát, hanem azzal is, hogy csökkenti-e az agyvérzését. *Külön-külön* 5% a valószínűsége, hogy alaptalanul kiáltunk fel (tehát mondjuk, hogy az adott végponton hat a gyógyszer, miközben a valóságban nem is), de együtt már közel sem! Hiszen ezen hozzáállás mellett akkor is hibásan kiáltunk fel, ha *vagy* az egyik, *vagy* a másik végponton hibásan kiáltunk fel; ennek valószínűsége természetesen összekombinálódik a külön-külön vett tévedések valószínűségéből. Mintha lenne két húszoldalú, szabályos dobókockánk, és azt kérdeznénk, hogy mekkora a valószínűsége, hogy a kettőből *valamelyik* 1-est dob – ami értelemszerűen nagyobb, mint, hogy bármelyikkkel külön 1-est dobunk (ami 5% ebben a példában). Nem a két 5% összege, ahogy azt sokan gondolnák elsőre, a matematika ennél egy nagyon kicsit bonyolultabb, de most ez nem különösebben érdekes, a végeredmény: 9,75%.

Az orvosi példánkra visszatérve mindez azt jelenti, hogy bár mi azt hirdetjük magunkról, hogy 5%-on vizsgálódtunk, azaz a téves találat (nem hat a gyógyszer de mi mégis azt mondjuk, hogy igen) valószínűsége 5%, a valóságban e vagylagos hozzáállás mellett az ilyen tévedés valószínűsége ennek majdnem kétszerese lesz! Ahogy nő a végpontok száma a helyzet csak egyre romlik, ezt mutatja az 6.4. ábra (a halvány szürke vonal az 5%-ot mutatja, mely az egyes összehasonlítások külön-külön vett ilyen hibavalószínűsége). Érdemes úgy is ránézni a kérdésre, hogy az 5% azt jelenti, hogy ha sehol nincsen semmilyen hatás, akkor is várhatóan minden 20. esetben mégis találunk. Avagy: aki keres, az talál – csak itt ez nem feltétlenül jó hír...

A problémakör e ponton kettéágazik. Az egyik irány a nyílt rosszhiszeműség: addig növelni az összehasonlítások számát, amíg valahol csak találunk valamit. Például be akarjuk bizonyítani, hogy az emberek vére szisztematikusan eltér aszerint, hogy hosszú-e a vezetéknévük. Ez első ránézésre elég megmosolyogtatónak hangzik, pedig valójában pofonegyszerűen bizonyítható! Semmi más dolgunk nincs mint fogni 100-100 rövid és hosszú vezetéknévű embert és egy teljesen rutin laborvizsgálatnak alávetni őket. Manapság ezek is 20, 30, vagy akár ennél is több paramétert mérnek le, ha ezeket mind összehasonlítjuk egyesével, akkor elég nyugodtan hátradőlhetünk, hogy legalább egy különbséget találni fogunk (lásd 1. ábra!).



6.4. ábra. Adott számú végpont vagylagos összehasonlítása esetén annak a valószínűsége, hogy legalább egy tévesen hatást mutat, ha valójában semelyik végponton nincsen hatás (és a különböző végpontok függetlenek egymástól).

Innen től két lehetőségünk van. Az egyik, hogy leírjuk a cikkben, hogy 30 összehasonlítást végeztünk, ebből 29 esetben nem találtunk szignifikáns különbséget, 1 esetben igen. Ebből minden olvasó tudni fogja, hogy mit találtunk: semmit, hiszen az az 1 különbség tökéletesen megfelel annak, amennyit akkor várunk, a véletlen ingadozásnak köszönhetően, ha semmiben nincs különbség. Igen ám, csakhogy. Van egy másik lehetőség is: „megfeledkezünk” róla, hogy mi valójában 30 dolgot hasonlítottunk össze, kitalálunk utólag egy filozófiát ahhoz az 1-hez, és úgy írjuk le, mintha eleve is, célirányosan azt néztünk volna meg. Mondjuk egy gyulladásmarkerben találunk eltérést, akkor úgy írjuk meg a cikket, hogy van egy fantasztikus körélettani teoriánk, miszerint a hosszú vezetéknélküli embereket frusztrálja, hogy olyan lassú amíg aláírják a papírokat, és ez a frusztráció egy szervezettszintű gyulladást hoz létre bennük – éppen ezért célirányosan megmértük egy gyulladásos paraméterüket, és láss csodát, tényleg eltér. (Tételezzük fel, hogy nincs confounding, tehát nem lehet, hogy valami összefügg a vezetéknévvel, ami hat a gyulladásra is.) Az igazán nagy baj ebben az, hogy az ilyen típusú csalást magából a cikkből nem lehet lebuktatni! Hiszen ha tényleg célirányosan csak a gyulladásmarkert vizsgáltuk volna meg, akkor teljesen rendben is lenne ez az eredmény. Ami természetesen nem azt jelenti, hogy biztosan igaz lenne, de tényleg 5% lenne ennek az eredménynek a valószínűsége akkor, ha valójában nem lenne különbség. (Ami természetesen, emlékezzünk vissza a korábbi írásokra, nem ugyanaz, mint hogy 5% valószínűséggel nincs különbség!) Persze, ha a megdöbbentő eredményen fellekesülve egy másik kutatócsoport megpróbálja reprodukálni a vizsgálatunkat, akkor jó eséllyel lebukunk; egész pontosan csak 5% a valószínűsége, természetesen, hogy – véletlenből kifolyólag – ők is ugyanezt találják.

Számos csalás vagy félreértés megy ugyanerre a kaptafára; ezt szokás, meglehetősen találó kifejezéssel, *szignifikanciaavadászatnak* nevezni az irodalomban. Nézzünk egy másik példát!

Ez a jelenség természetesen nem csak több végpont esetén fordulhat elő – a kritérium az, hogy több összehasonlítást hajtsunk végre (nem véletlen az @ref(fig:alfainflacio). ábrán a tengely felirata!), ezért szokás egyébként ezt többszörös összehasonlítások problémájának nevezni. Ez lehet több végpont összehasonlítása, de lehet több kezelés összehasonlítása ugyanazon a végponton, vagy egy kezelés összehasonlítása egy végponton, de több csoportban. Ez utóbbihoz tekintsük a következő kijelentést: „a gyógyszer 5%-os szignifikanciaszinten szignifikáns hatás-sal bír a 30-40 év közötti cukorbeteg férfiak körében”. Rendben van ez így? Az őszinte válasz erre a kérdésre, hogy nem tudjuk! Ha csakugyan célirányosan a 30-40 év közötti cukorbeteg férfiakat vonták be a vizsgálatba, akkor minden rendben, legalábbis ebből a szempontból. (Ami persze nem azt jelenti, hogy biztosan igaz az állítás, és még csak azt sem, hogy 5% a tévedés, azaz a gyógyszer hatástalanságának a valószínűsége, ezt nem lehet elégszer hangsúlyozni.) Azonban enyhén szólva is erős lehet a gyanunk, hogy erről szó sincs, hanem egyszerűen nem hatott a gyógyszer, ezért elkezdtek próbálkozni. Esetleg csak a férfiakban? Csak a nőkben? Csak a cukorbetegekben? Ha ezt a fenti módon, kombinatorikusan tesszük, akkor nagyon hamar rengeteg összehasonlításunk lesz (10 korcsoporttal, 2 nemmel és 2 cukorbetegség szerinti állapottal számolva 40!), ami pontosan ugyanahhoz a problémához vezet mint a vezetéknéves példa.

6.8.2. Egy nem csak biostatisztikai tanulságokkal bíró kitérő

A szignifikanciaavadászatra sok példát lehetne hozni; az alábbi azért érdekes, mert a szignifikanciaavadászaton (sőt, általában a statisztikai kérdésekben) messze túlmutató tanulságokkal is bír. A történet eredetileg nem a szignifikancia-vadászatra van kihegyezve, hanem az újságírók alaposságát akarta megvizsgálni, de a sztori magvában, mint majd látni fogjuk, a szignifikancia-vadászat van.

2015-ben egy Johannes Bohannon nevű szerző és munkatársai, a német Táplálkozási és Egészségügyi Intézet kutatói, közöltek egy tanulmányt az International Archives of Medicine nevű orvosi lapban, ami nagyon leegyszerűítve arról szólt, hogy a csokoládéevés segít a fogyásban. Egy rendes klinikai kísérletről volt szó, empirikusan vizsgálták meg a kérdést: az alanyokat véletlenszerűen több csoportra osztották, és azt találták, hogy a csokoládét evő csoport testtömege szignifikánsan csökkent a kontrollokéhoz viszonyítva. Gyönyörű grafikonok, táblázatok, *p*-értékek, részletes diszkusszió az eredményekhez, számos irodalmi hivatkozással, ahogyan kell.

A tanulmánnyal azonban van pár apróbb probléma. Az egyik, hogy „Táplálkozási és Egészségügyi Intézet” nem létezik, a másik, hogy „Johannes Bohannon” nevű kutató nem létezik (más intézetben sem, ez ugyanis álnév, a szerző igazi neve John Bohannan, és valójában angol tudományos újságíró, nem német orvos), a harmadik, hogy a folyóirat egy jól ismert kamufelirat. (Kitérő a kitérőben: a tudományos folyóiratok klasszikus modelljével szemben, melynél az olvasó fizet a lap vagy cikk elolvasásáért, egy új modell is kialakult, melyben a cikkek ingyen elolvashatóak, cserében viszont a szerzőknek kell fizetniük a megjelenésért. Ez sok szempontból pozitív és rendkívül szimpatikus kezdeményezés, azonban sajnos létrejött egy erre rátelepedő csalási iparág is. Ebben a „folyóiratok” valójában semmilyen bírálatot nem alkalmaznak, akármit közölnek – néha *szó szerint* akármit, véletlenszerűen egymás után rakott szavakból álló cikket is... – majd begyűjtik a pénzt a szerzőktől a közlésért. Bárki, aki valaha írt igazi tudományos cikket, tudja miről van szó, ugyanis a postaládájába naponta tízesével érkezik a levélszemét az ilyen „folyóratoktól”... Az International Archives of Medicine egy ilyen lap; ez amúgy nagyon gyorsan ki is deríthető róla.)

Összefoglalva, egy nem létező intézet nem létező kutatója közzétesz egy cikket egy kamufeliratban, amely cikk ráadásul módszertanilag is botránnyos (erről majd pici később); és a kérdés: vajon mi fog történni? Hányan veszik ezt észre? A válasz: újságok, tévék, internethelyszínek, beleértve egészségügyi rovatokat, sőt, magukat a táplálkozástudományban jártasnak mondó szerzők, százával vették át a hírt az egész világon, egyetlen újságíró nem akadt, aki utánanezett volna, hogy mit is közöl. Amiben az az igazán kétségejtő, hogy nem kéthates kutatóinkra lett volna szükség a dolog felderítéséhez: Bohannonék minden lehetségeset megtettek, hogy a lehető legkönnyebben lebukjanak. Ha beütíti valaki az intézet nevét a Google-be, kiderült volna, hogy sehol nincs nyoma sem, a weboldalát akkor hozták létre, ha beütik a kutató nevét bármelyik tudományos adatbázisba, kiderült volna, hogy nem is létezik. 30 másodperc munka, ha lassan végzik, 2 perc. És akkor arról még nem beszélünk, hogy esetleg, netalántán, mielőtt egy ilyen cikk konklúzióját változtatás nélkül közlik, talán meg lehetne kérdezni egy hozzáértőt, hogy maga a cikk amúgy rendben van-e módszertani szempontból. De nem, senki nem akadt,

aki ezt megtette volna, sőt, a bulvárlapok még fokozták is a hírt („miért kell [!] minden nap csokoládét enned”)... A helyzet odáig fajult, hogy végül Bohannon-nak saját magát kellett lebuktatnia egy cikkben, hogy megállítsák a dolgot. Jusson ez eszünkbe, ha kedvenc lapunkban legközelebb az épp aktuális új tudományos világszenzációról olvasunk...

Na de mi köze ennek az egésznek a szignifikancia-vadászathoz? A helyzet az, hogy a szerző kamú, az intézet kamú, a folyóirat kamú, de egy doleg valódi volt: a kutatás! Bármilyen meglepő, az nem volt kamú, Bohannonék *tényleg* verbuváltak kísérleti alanyokat, tényleg véletlenszerűen csoportba osztották őket, tényleg etettek velük csokit (vagy sem) és tényleg lemérték, hogy mi történik velük. És a testtömegük tényleg csökkent!

Ez tehát igaz, csak épp – és most jön a lényeg – valójában *18 különböző* dolgot mértek le, egyáltalán nem csak a testtömeget! Az tehát, hogy találtak 1 különbséget, *tökéletesen megfelel* annak, hogy semmiben nincsen semmilyen különbség. A nem statisztikai tanulságra visszatérve, az igazán szomorú az, hogy – szemben a vezetékneves példával – ezt meg sem próbálták eltitkolni, épp ellenkezőleg, teljesen világosan leírták, hogy 18 összehasonlítást végeztek. Tehát még a módszertani hibát sem lett volna nehéz megtalálni (a többihez hasonlóan, mert nem ez volt az egyetlen!) – ennek ellenére vették át a hírt a különféle lapok, internetes oldalak, rendkívül lehangoló képet festve arról, hogy mennyire működik valójában a kritikus szemlélet. Sajnos ez nem csak a laikus sajtóra igaz...

(Záró megjegyzésként fontos ugyanazt a figyelmeztetést hozzátenni, mint a confounding esetében: ez nem azt jelenti, hogy akkor cáfoltuk, hogy a csokoládé-evés fogyasztana – éppenséggel fogyaszthat is, csak épp ez a tanulmány ezt rettenetesen kevéssé bizonyítja.)

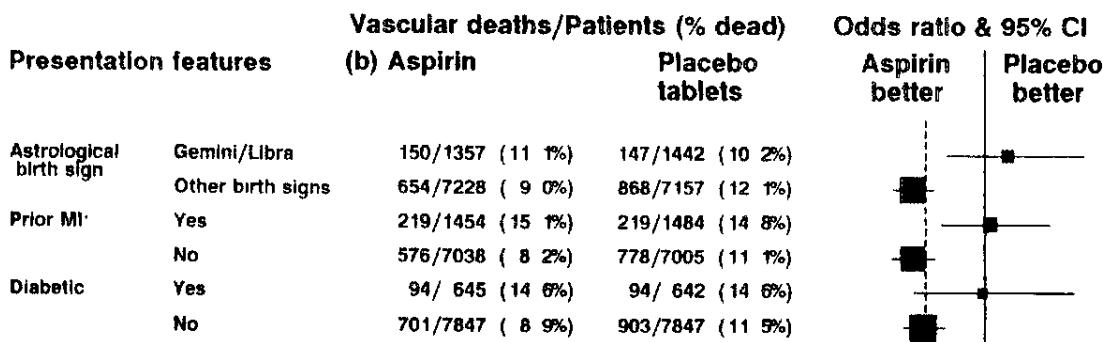
6.8.3. A jóhiszemű kutatók nehézségei

Bizonyos értelemben a fenti még a jobbak eset. Ez ugyanis direkt csalás, joggal mondhatjuk, hogy ilyet nem lehet véletlenül csinálni, és így a jóhiszemű kutatóknak nem kell aggódniuk. A probléma, mint sok más esetben, akkor jelentkezik, amikor az ilyen többszörös összehasonlításoknak teljesen legitim okai vannak. A vezetékneves példa lehet, hogy abszurd, de mi van, ha egy betegségnél mi sem tudjuk, hogy melyik laborváltozóban okoz eltérést, és ezt akarjuk kideríteni? Mi van, ha egy környezeti hatás esetében több száz betegségről szeretnénk kideríteni, hogy valamelyiknek növeli-e a kockázatát?

A legjobb példa mindenre talán az alcsoport-elemzések intézménye. Alcsoport-elemzés alatt azt értik az orvosi irodalomban, amikor egy vizsgálat végén az egész elemzést megismétlik úgy, hogy az adatokat leszűkítik különböző csoportokra. Hogyan hat a gyógyszer csak a férfiakban és csak a nőkben? Cukorbetegekben és nem cukorbetegekben? Idősekben és fiatalokban?

Itt is az a nehéz helyzet, amikor nem rosszhiszeműen járunk el, mint az írás elején felhozott hasonló példában, hanem *tényleg* kíváncsiak vagyunk erre, mert elvileg *tényleg* lehet, hogy a gyógyszer bizonyos csoportokban nem hat (vagy csak bizonyos csoportokban hat). Ezért az alcsoport-elemzéseknek van létfogosultságuk, de hogy legalábbis óvatosan kell velük bánni, arra

az irodalom egyik legszórakoztatóbb példáját egy 1988-as kutatás hozta. Ezt ma már minden bizonnyal nem így neveznék el – annak idején ugyanis az ISIS-2 nevet kapta... Ez a kísérlet azt vizsgálta, hogy 3 kezelési stratégia (egy sztreptokináznak nevezett vérrögoldó adása, aspirin adása, vagy mindenki adása egyszerre) hogyan hat infarktusban a halálozás megelőzésére. A 6.5. ábra mutatja az eredményeket. A jobb oldalon lévő folytonos függőleges vonal a „hatástaranság vonala”: amelyik csoport vízszintes vonala metszi, ott az adott csoportban nem hat a gyógyszer. Az alsó kettő meglehetősen szokásos alcsoport-analízis: a cukorbetegekben nem hat a gyógyszer (a nem cukorbetegekben igen), az infarktuson átesettekben nem hat a gyógyszer (a korábban infarktust el nem szenvedőknél igen). Ha csak az ábra jobb oldalát nézzük, akkor a legfelső pár is nagyon hasonló ezekhez, de nézzük csak meg jobban a feliratot a bal oldalon! A legfelső alcsoport-analízis tárgya, hogy a betegnek mi a csillagjegye – az eredmények szerint Ikek és Mérleg jegyűekben nem hat a gyógyszer, a többiekben igen...



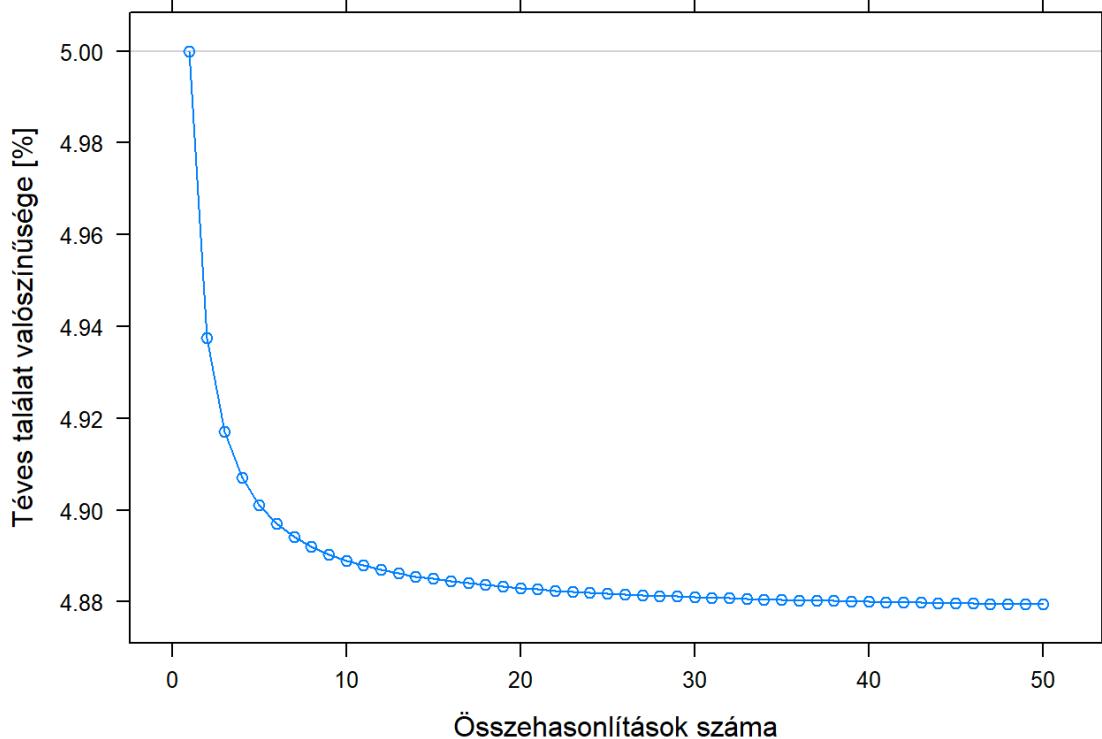
6.5. ábra. Kissé meglepő eredmények az aspirin hatására vonatkozóan: facsimile oldal az ISIS-2 vizsgálat közleményéből.

A megállapítás, amelyet ez a példa nagyon szellemesen szemléltet, hogy a cukorbetegség és a csillagjegy között az *egyetlen* különbség, hogy az előbbihez jobb filozófiát tudnak kitalálni az orvosok, hogy ott miért nem hat a gyógyszer, az utóbbihoz meg talán picit kevésbé... De az eredmények statisztikai bizonyítóereje *pontosan ugyanaz* minden esetben!

Mi akkor a megoldás mindenekkel problémára? A kérdés összetett, itt talán csak elég két pontot kiemelni. Az egyik, hogy bizonyos esetekben nem akarjuk megoldani a problémát. Az alcsoport-elemzésekben sokszor ez a helyzet: a cikkek nem próbálják meg statisztikai úton helyezni a dolgot, cserében viszont ekkor az alcsoport-elemzések eredményét nem lehet bizonyítékként kezelni (legfeljebb felvetésként: ha valahol gyanús eltérést látunk, akkor azt rendes vizsgálatban célrányosan meg kell nézni, de az alcsoport-analízis önmagában nem bizonyít, csak felveti a gyanút).

Máskor szeretnénk korrigálni a többszörös összehasonlítások helyzetét; a jó hír, hogy erre vannak módszerek. Az alkalmazásuk kapcsán sok vita van, de az fontos, hogy egyáltalán lehet a problémán statisztikai úton segíteni. Vegyük például a következő ötletet. Mi a probléma alapja? Az, hogy hiába rakjuk 5%-ra a szintet összehasonlításonként, összességében ez meg fog nőni, ahogy azt az 6.4. ábra is mutatja, jóval 5% fölé. Akkor mit csinálunk? Vegyük le

a szintet összehasonlításunként, hogy összességében kapjunk 5%-ot! Bebizonyítható matematikai úton példának okáért, hogy ha az összehasonlításunkénti szintet úgy állítjuk be, hogy az 5%-ot elosztjuk az összehasonlítások számával (tehát például 10 összehasonlítás esetén 0,5%-ra rakjuk), akkor az összességében vett hibavalószínűség – annak a valószínűsége, hogy téves módon hatást mutatunk ki *bárhol*, miközben igazából sehol nincs hatás – nem lehet több mint 5%. A módszer neve Bonferroni-korrekción, az eredményét a 6.6. ábra szemlélteti. Itt az alaphelyzet ugyanaz, mint az előbbi ábra esetében, de az összehasonlításunkénti szignifikanciaszintet lecsökkentettük, annyiadrészre ahány összehasonlítást végeztünk. Jól látszik az eredmény: a téves találat valószínűsége igen pontosan 5% marad (illetve soha nem megy fölé), akárhány összehasonlítást is végzünk.



6.6. ábra. A Bonferroni korrekció hatása.

A Bonferroni-korrekciónak nagyon sok baja van (például rendkívül szigorú, nagyon megnehezíti a valódi hatások észrevételét is – azaz nagyon lecsökkenti az erőt), de azt jól szemlélteti, hogy a probléma kezelhető statisztikai úton.