Bevezetés a biostatisztikába

Ferenci Tamás

2025. december 4.

Tartalomjegyzék

Előszó **4**

1 A statisztika alapjai **5**
1.1 A statisztika alapfogalmai és ágai 5
1.2 Változók és mérési skálák 6
1.3 A biostatisztika kapcsolódó tudományai és elhatárolása 9
1.4 A biostatisztika számítástechnikai háttere 10
1.5 Futó példa 11

2 Deskriptív statisztika **14**
2.1 A deskriptív statisztikáról általában 14
2.2 A deskriptív statisztika módszereinek csoportosításáról 16
2.2.1 Grafikus és analitikus módszerek 16
2.2.2 Egy- és többváltozós módszerek 17
2.2.3 A vizsgált változó(k) mérési skálája 18
2.3 Minőségi változó egyváltozós elemzése 18
2.3.1 Analitikus eszközök 18
2.3.2 Grafikus eszközök 20
2.4 Mennyiségi változó egyváltozós elemzése 21
2.4.1 Analitikus eszközök 22
2.4.2 Grafikus eszközök 32
2.5 Minőségi változók kétváltozós elemzése 45
2.5.1 Analitikus eszközök 45
2.5.2 Grafikus eszközök 50
2.6 Mennyiségi változók kétváltozós elemzése 50
2.6.1 Analitikus eszközök 50
2.6.2 Grafikus eszközök 52
2.7 További többváltozós elemzések 55

3 Induktív statisztika **56**
3.1 A mintavételi helyzet és következményei 56
3.2 Becsléselmélet 59
3.3 Hipotézisvizsgálat 69

4 Haladó adatvizualizáció **79**
4.1 A base R grafika és korlátai 79

4.2 Haladó adatvizualizációs csomagok, a ggplot2 88

Előszó

Előszó.

A használt könyvtárak:

```
library(ggplot2)
```

1 A statisztika alapjai

A biostatisztika a *statisztika* egyik alkalmazott ága, mely az orvosbiológiai területen felmerülő, empirikus adatokkal leírt kérdések kvantitatív vizsgálatával foglalkozik.

Egy új vérnyomáscsökkentő gyógyszer-jelölt valóban csökkenti a vérnyomást? Nagyfeszültségű vezeték közelében tartózkodás növeli a rák-kockázatot? Van-e összefüggés egy gyermek táplálkozási energiabevitele és magasságának növekedése között? Ez csak pár példa olyan kérdésekre, melyek megválaszolásának egyik lehetősége az *empirikus adatok* alkalmazása: összegyűjtjük gyógyszert szedő és nem szedő emberek vérnyomását; emberek lakhelyének távolságát a nagyfeszültségű vezetékektől, és azt, hogy kialakult-e náluk rák; gyermekek táplálási adatait és magasságuk alakulását. Ezen adatok birtokában van remény a problémák vizsgálatára, orvosilag releváns kérdések megválaszolására. Továbbmenve az is látható, hogy végeredményben mind számszerű adatra vezet (vérnyomásalakulás, rákkockázat, magasságváltozás stb.), így adhatunk *kvantitatív* válaszokat is a kérdésekre (pontosan mekkora vérnyomás-csökkenést okoz várhatóan a gyógyszer-jelölt, ha egyáltalán okoz, hány százalékkal változtatja a rákkockázatot adott nagyfeszültségű vezeték, ha egyáltalán változtatja, nagyobb energiabevitel mennyiben módosítja a növekedés ütemét, ha egyáltalán módosítja stb.). Ehhez természetesen megfelelő elemzéseket kell végrehajtanunk, megfelelő modelleket kell alkotnunk. Ezzel foglalkozik a biostatisztika.

1.1. A statisztika alapfogalmai és ágai

Ahogy láttuk, a statisztika egyik fontos feladata lesz bizonyos kérdések szabatos megválaszolása empirikus adatok alapján. Ennek kapcsán be kell vezetnünk pár alapfogalmat, mely a statisztikusok beszédében lépten-nyomon előkerül.

Azt a halmazt, melyre a statisztikai eszközökkel megvizsgálandó kérdésünk vonatkozik (cél)populációnak, vagy *sokaságnak* szokás nevezni. A sokaság elemeit szokás *megfigyelési egységnek* is nevezni. Ha azt kérdezzük, hogy „Mennyi egy adott kurzus hallgatóinak átlagos testtömege?”, akkor a sokaság az adott kurzus hallgatóiból álló halmaz; a megfigyelési egységek az egyes hallgatók.

Azt a szempontot, amely szerint a sokaság elemeit vizsgálat alá vonjuk, *ismérvnek*, vagy más szóval *változónak* hívjuk. Az előbbi példa esetében a változó a testtömeg; más esetekben persze több változót is használunk. Azt a lépést, amikor adott változó értékét meghatározzák egy adott sokasági elemre, általában *megfigyelésnek* nevezik a statisztikában.

Nagyon sokszor nem tudunk a sokaság valamennyi egyedéről információt szerezni (azaz: nem tudjuk mindegyiket megfigyelni). Ilyenkor a sokaság azon részhalmazát, amelyet meg tudunk figyelni (tehát amelyről információnk van), *mintának* nevezzük, és ezt a helyzetet magát *mintavételi helyzetnek* hívjuk. Ennek egyrészt technikai okai lehetnek: sok esetben a sokaság valamennyi egységéről való adatgyűjtés (az ún. teljes körű megfigyelés) technikai okok miatt nehézkes vagy egyenesen lehetetlen (túl költséges, túl bonyolult a megszervezése, túl időigényes stb.). A biostatisztikában azonban ennél is fontosabb egy másik ok: az, hogy sok kérdés nem egy kézzelfogható, véges nagyságú sokaságra (mint egy adott kurzus hallgatói), hanem egy ún. fiktív sokaságra vonatkoznak. A kurzus hallgatóit fel lehet sorolni, felírhatjuk a neveiket egymás alá egy lapra. Egy ország lakosainál ugyan ez nehezebb a gyakorlatban, de elvileg minden további nélkül megtehető. De vessük ezt össze azzal a kérdéssel, hogy egy új vérnyomáscsökkentő gyógyszer-jelölt valóban csökkenti-e a vérnyomást – mi itt a sokaság? Itt valami alapvető különbség van: ennek a sokaságnak az elemeit nem tudjuk felírni egy lapra! Soha nem mondhatjuk azt, hogy tessék, itt a névsor, *konkrétan, név szerint őket* kell gyógyítania a gyógyszernek. E kérdés nem emberek egy konkrét, összeszedhető csoportjára vonatkozik, hanem egy képzeletbeli, megfoghatóan nem létező, absztrakt csoportra („aki megfelel a gyógyszer alkalmazási feltételeinek és nincs ellenjavallata"). Ez nem egy konkrét sokaság, hanem egy fiktív csoport; sokszor hasznos ha úgy gondolunk rá, mintha ebben végtelen sok elem lenne. Ebből az is következik, hogy akármennyi embert is vizsgálunk meg ebből a sokaságból, az szükségképp csak része lesz annak, azaz szükségképp csak mintát fog jelenteni a sokaságból. (Soha nem mondhatjuk, hogy mindenkin kipróbáltuk a gyógyszert, akin működnie kell.) Ilyenkor tehát *mindenképp* mintavételi helyzettel lesz dolgunk. Mivel ez a helyzet tipikus a biostatisztikában, így máris érthető, hogy miért mondtam, hogy a mintavételi helyzetnek – illetve kezelésének – kiemelt jelentősége van a biostatisztikában.

A statisztika azon ágát, mely sokaságról szerzett adatokkal foglalkozik, vagy mintabeliekkel de úgy, hogy elhanyagolja, hogy csak mintáról van szó (mintha a minta lenne a sokaság) *deskriptív (vagy leíró) statisztikának* nevezik; erről később bővebben lesz szó (2. fejezet). Ide tartoznak olyan kérdések, mint az információtömörítés, lényegkiemelés, adatvizualizáció. A statisztika azon ága, mely figyelembe veszi a mintavételi helyzetet, azaz mintabeli adatokkal foglalkozik, de úgy, hogy szem előtt tartja, hogy a kérdések valójában a sokaságra irányulnak, és – a minta alapján – arra próbál válaszolni, az *induktív (vagy következtető) statisztikának* névre hallgat, szintén részletesen lesz róla szó később (3. fejezet).

1.2. Változók és mérési skálák

Az előbbi pontban kissé nagyvonalúan csak annyit írtam, hogy a változó (vagy ismérv) az a szempont, ami alapján a megfigyelési egységeket vizsgálat alá vonjuk. (Természetesen több ilyen is szerepelhet egy vizsgálatban.) Ez meglehetősen kézenfekvő akkor, ha mondjuk az emberek testtömege a vizsgálati szempont – ekkor mondhatjuk egyszerűen, hogy lemérjük őket alkalmas módszerrel, és az e tulajdonságot leíró „testtömeg" változó legyen a lemért tömeg

mondjuk kilogrammban kifejezett értéke. Más esetekben azonban közel nem ilyen egyértelmű a változók megválasztásának a kérdése.

A statisztika alapvetően számszerű információk feldolgozásával foglalkozó tudomány, így ahhoz, hogy egy szempontot statisztikai úton tudjunk vizsgálni, előbb *számszerűen mérhetővé* kell tenni. Ez természetesen olyan információkkal is végrehajtható, melyek eredetileg nem számszerűek, ezt nevezzük *operacionalizálásnak*. Néha ez valóban szinte triviális feladat (a testtömeget mérjük az adott módon lemért és kilogrammban kifejezett testtömeggel), máskor viszont egyáltalán nem az. Gondoljunk arra, hogy hogyan lehet számszerűen mérhetővé tenni egy olyan jellemzőt, mint hogy milyen súlyos egy alany depressziója – szinte külön tudományág, hogy ehhez milyen kérdőívek, egyéb vizsgálatok kellenek, mellyel „lemérhető" ez. (Valójában a testtömeg mérése sem feltétlenül triviális. Mikor mérjük, reggel, délben, este? Ruhával, anélkül, mennyi ruhával? Milyen mérlegen?)

A változók kapcsán a másik probléma, hogy egy sor tulajdonság nem mérhető közvetlenül – akár technikai akadályok miatt, akár az operacionalizálás nehézségei miatt. Ez esetben gyakran kényszerülünk arra, hogy az eredetileg megcélzott változó helyett más, immár mérhető, és az eredetivel – lehetőleg minél szorosabb – kapcsolatban lévő változót vagy változókat mérjünk le. Az ilyen célból használt változót nevezzük *proxy változónak*. Például komoly gondban lennénk, ha az alany szocioökonómiai státuszát kéne lemérnünk egyetlen változóval – ezt ilyen formában aligha tehetjük meg, így a gyakorlatban proxykat próbálnánk hozzá keresni, például iskolai végzettséget mérnénk, jövedelmet, munkahelyi beosztást stb.

A következő kérdéskör, amiről a változók kapcsán beszélni kell, az a *mérési skála* fogalma. Mivel a statisztika végeredményben számszerű információkat dolgoz fel, így a változóinkat is tipikusan számokkal fogjuk leírni. Észre kell azonban venni, hogy vannak jellemzői a változóknak, amik *önmagukban* e számokból nem olvashatóak ki. Példának okáért tekintsük azt az adatot, hogy mi az alany szemszíne, és azt, hogy mennyi a CRP-je (ez egy laboreredmény). Tételezzük most fel, hogy a szemszínt úgy számszerűsítettük, hogy a barnához 1-et, a feketéhez 2-t, az egyébhez 3-at rendelünk; a CRP-nél pedig a koncentrációja számértékét adjuk meg, egész mg/l-ben. Mármost ekkor mindkét adat (a szemszín és a CRP) is lehet történetesen 1, 2 és 3 értékű – ám ettől még hatalmas különbség van köztük: a CRP-nél van értelme azt mondani, hogy 1,23 volt az alanyaink átlagos CRP-je, de annak nyilván nincs értelme, hogy 1,23 volt az átlagos szemszínük. E mögött az húzódik meg, hogy a CRP-k számértékeit van értelme összeadni egymással, a szemszínek számértékeit nem. Tehát: az, hogy milyen műveletek végezhetőek el az adott változóval, nem olvasható ki a változó által felvett értékekből. Ezeket a különbségeket a mérési skála fogalma ragadja meg, mely azt írja le, hogy hogyan viselkednek, viselkedhetnek az adataink. A leghíresebb Stanley Smith Stevens mérési skála modellje, mely négy lépcsőfokot különböztet meg. (Azért is beszélünk lépcsőfokokról, mert ez egy egymásra épülő, folyamatosan bővülő felosztás: a későbbi, magasabb skálák bírnak az összes többi korábbi, alacsonyabb skála tulajdonságaival, és még persze valamilyen többlettel is.) Stevens skálái a következőek:

1. *Névleges (nominális) skála* Ilyen skálán mért adatok esetén az adat számértékének valójában nincs semmi jelentősége, kizárólag az számít, hogy a számérték ugyanaz-e két

alanynál vagy sem: ha ugyanaz, akkor a két alany egyezik a változó szempontjából, ha nem, akkor nem – és kész, ennyit mondhatunk, semmi többet. Erre jó példa a beteg lakóhelye megye szerint; 1-től 20-ig kódolva. Ha az egyik betegnél ez 3, a másiknál 6, akkor kizárólag annyit mondhatunk, hogy különböző megyében laknak, semmi többet. Olyan kijelentéseknek, hogy a második „hárommal nagyobb megyében", „kétszer akkora megyében", vagy akár csak annak, hogy „nagyobb megyében lakik" nyilvánvalóan nincs értelmük. További tipikus példa nominális ismérvre a beteg neme, rassza, szemszíne stb.

2. *Sorrendi (ordinális) skála* Ilyen skála esetében már valamennyi jelentősége van a számértékeknek: számít ugyanis, hogy melyik *nagyobb* – ám ezen kívül semmi más. Ezzel tehát a lehetséges kimeneteket sorba rendeztük (innen a skála neve), ám egyebet nem mondhatunk. Tipikusan ide tartozik a különféle betegségek staging adata. Ha ez egyik beteg I., a másik II. stádiumban van, akkor mondhatjuk azt, hogy ez utóbbi állapota súlyosabb (ha ez nominális skálán mért ismérv lenne, akkor már ennyit sem mondhatnánk, csak annyit, hogy *nem ugyanaz* a súlyosság), ám olyan kijelentéseknek, hogy „eggyel súlyosabb", vagy „kétszer olyan súlyos" állapotban van, nincs értelme. Vegyük észre, hogy ez valóban tartalmazza a nominális skála jellemzőit (hiszen ha a kimenetek sorbarendezhetőek, akkor természetesen meg is különböztethetőek), azaz tényleg kibővítése annak.
3. *Valódi skálán mért ismérvek* Ide tartoznak azok az ismérvek, amelyek kimeneteivel már egyéb műveletek (nem csak az összehasonlítás és a sorbarendezés) is értelmezettek. Például ha egy beteg CRP-je 1 mg/l, egy másiké 2 mg/l, akkor mondhatjuk, hogy a kettő különbözik (nominális tulajdonság), mondhatjuk, hogy az utóbbi nagyobb (ordinális tulajdonság), *de* nyugodtan tehetünk olyan kijelentést is, hogy az utóbbi „eggyel nagyobb", vagy hogy „kétszer akkora" mint az előbbi! Ezek a skálán mért ismérvek, ide tartozik például a legtöbb laboreredmény. A statisztikai irodalomban ezen a kategórián belül két további csoportot szokás megkülönböztetni: a különbségi – vagy intervallum – skálán mért ismérveket, és az arányskálán mért ismérveket. Az eltérés a kettő között, hogy az előzőben csak az összeadás, míg az utóbbiban az összeadás *és* a szorzás is értelmezett. Például a CRP arányskálán mért, hiszen két érték vonatkozásában beszélhetünk arról, hogy az egyik mennyivel több, illetve hányszorosa a másiknak. A beteg testhőmérsékleténél, ha azt Celsius-fokban mérjük, már nem ez a helyzet! Annak van értelme, hogy az egyik beteg maghőmérséklete 5 fokkal több, de olyat nem mondhatunk, hogy 10%-kal magasabb[1].

Megjegyezzük, hogy az első két skálán mért változót nagyon gyakran *minőségi* (vagy kvalitatív) változónak nevezik közös néven, míg a valódi skálán mért változókat sokszor *mennyiségi* (vagy kvantitatív) változónak hívják.

Itt érdemes megemlíteni, hogy a változókat csoportosíthatjuk aszerint is, hogy hány lehetséges kimenetet vehetnek fel. Ha véges sokat vagy legfeljebb megszámlálhatóan végtelen sokat, akkor *diszkrét* változóról beszélünk, különben *folytonosról*. Folytonos változóra tipikus példa az olyan változó, melynek értékei a valós számok közül, vagy a valós számok valamilyen intervallumából

[1] Gondoljunk csak bele, ha ma 2 fok van kint egy téli napon, tegnap 1 fok volt, akkor aligha mondhatjuk, hogy ma kétszer olyan meleg van… Ez abból adódik, hogy a hőmérsékletnek nincsen rögzített nulla pontja – az teljesen esetleges, hogy a Celsius-skála hova rakta azt.

(pl. pozitív valós számok) kerülnek ki. Természetesen a gyakorlatban a korlátos mérési pontosság miatt elvileg minden változó diszkrét, de ha nagyon nagy a lehetséges kimenetek száma, és ezek egymáshoz sűrűn helyezkednek el, akkor általában nyugodtan alkalmazható a folytonos közelítés.

Nagyon sokszor a diszkrét változó fogalmat azonosítják a minőségi, a folytonosat pedig a mennyiségi változóval. Tisztán elméleti szempontból ez nem helyes (hiszen két különböző szempontról van szó), bár tény, hogy a legtöbb esetben valóban fennállnak ezek a megfeleltetések. Egy nevezetes kivétel ez alól a különféle darabszámokat, események számát stb. tartalmazó adatok, melyek a 0, 1, 2, 3 stb. értékeket vehetik fel (tehát diszkrétek), mégis skálán mértek, sőt, azon belül is arányskálán (tehát pont hogy a legmagasabb mérési skálán), hiszen általában van értelme nem csak különbségükről, de akár a hányadosukról is beszélni.

1.3. A biostatisztika kapcsolódó tudományai és elhatárolása

A biostatisztika az alkalmazott statisztika egyik ága, hasonlóan a pszichometriához, agrometriához stb. Látni kell, hogy a statisztika többé-kevésbé egységes tudomány, így végső soron hasonló módszereket alkalmaz az összes felsorolt ág, különbség inkább a részletekben (partikuláris problémákhoz testreszabott vagy kifejlesztett módszerek) és a az eljárások prezentációjában van.

Mint minden alkalmazott ágnak, a biostatisztikának is a statisztika, matematikai statisztika adja az alapját. Az itt bemutatott módszerek jó részéhez ugyan nincs szükség mélyebb matematikai statisztikai ismeretekre, de a manapság kifejlesztett új módszerek egyre komolyabb matematikai eszköztárat használnak.

A matematikai statisztika a matematika több ágára is épít, de ezek közül természetesen a valószínűségszámítás a kiemelkedően legfontosabb. (Ezt több más terület is kiegészíti természetesen, például a lineáris algebra.) Nem túlzás azt mondani, hogy a valószínűségszámítás a statisztika mögötti „alaptudomány”, melynek alapos ismerete elengedhetetlen a matematikai statisztika magas szintű műveléséhez. E jegyzetben azonban egyedül az induktív statisztikai rész fog komolyabb valószínűségszámítási alapismereteket feltételezni, a többi rész minden speciális matematikai ismeret nélkül is követhető lesz.

A valószínűségszámításon, matematikai statisztikán kívül természetesen orvosi ismeretekre is szükség van a biostatisztika műveléséhez. Ha nem is feltétlenül létkérdés, de a biostatisztikus munkáját megkönnyíti, ha legalább érti az orvosok szóhasználatát, valamint tisztában van az emberi test működésének élettani és a betegségek kórélettani alapjaival.

Ezt a szakaszt azzal zárom, hogy kísérletet teszünk a biostatisztika elhatárolására két olyan területtől, amellyel gyakran keveredik a fogalma. Az egyik a *bioinformatika*: ez a manapság rendkívül népszerű terület azonban inkább számítástechnikai, algoritmikus kérdésekkel foglalkozik (melyekkel nagy orvosbiológiai adatbázisokon is hatékonyan végezhetőek bizonyos műveletek, megválaszolhatóvá válnak bizonyos orvosilag releváns kérdések). A másik a *biomatematika*, ez

alatt azonban inkább olyan területet értünk, mely jellemzően nem statisztikai, hanem más matematikai (elsősorban analízisbeli) eszközöket, például differenciálegyenleteket használ, és a modellek adatokból történő becslése csak másodlagos kérdés.

1.4. A biostatisztika számítástechnikai háttere

Modern biostatisztika szinte elképzelhetetlen számítógépek, számítástechnikai támogatás nélkül. Ennek legalább három konkrét aspektusa van.

Először is, a leginkább „mechanikus" támogatás, amit a gépek adhatnak, hogy a szokásos számítási műveleteket (például egy átlag meghatározása vagy egy statisztikai próba kiszámítása) végrehajtják helyettünk. Bár sok statisztika kurzuson még ma is megtanítják a hallgatókat a kézi számításra (elsősorban azért, hogy jobban rögzüljenek a számítások részletei is), valójában már minden gyakorlati alkalmazásban számítógépek végzik a mechanikus kalkulációkat, érthető okokból kifolyólag.

A számítógépek ennél kicsit általánosabb módon is tudják támogatni a statisztikus munkáját. Azáltal, hogy segítik a nagy adatbázisok kezelését (szűrés, rendezés, keresés stb.), az adattranszformációkat (változók átkódolása, függvény szerint transzformálása stb.), lehetővé teszik, hogy könnyen kiszámoljunk mutatókat, vizualizáljunk adatokat és így tovább, a hatékonyabb, kreatívabb munkavégzést is segítik. (Részint azáltal, hogy csökkentik vagy szinte megszüntetik a rutinfeladatok időigényét, és így segítik, hogy a statisztikus a lényegre tudjon koncentrálni, részint azáltal, hogy számítógépek nélkül nem, vagy csak nagyon nehezen kivitelezhető segítségeket – pl. háromdimenziós ábrák – is tudnak adni a helyzet jobb megértéséhez.)

Végül pedig, vannak bizonyos módszerek, melyek nem csak nehézkesek lennének, de egyenesen elképzelhetetlenek számítástechnikai támogatás nélkül. Ezek az ún. *számításintenzív módszerek* (például az újramintavételezésen alapuló eljárások, a különféle algoritmikus modellek) mind rendkívüli számításigénnyel bírnak, így lényegében a számítógépekkel egyidősek, hiszen a nélkül kifejlesztésük, és különösen az érdemi használatuk nem volt elképzelhető.

Zárásként nagyon röviden megemlítem a talán legfontosabb programokat, melyeket a (bio)statisztikusok használnak:

- *SAS* A SAS egy komplex, nagyméretű és drága programcsomag. Legfőbb előnye, hogy jól standardizált, bejáratott, és a gyógyszeriparban – épp emiatt – előszeretettel alkalmazzák.
- *SPSS* Az SPSS egy általános célú statisztikai programcsomag (eredetileg szociológusoknak fejlesztették ki), funkcionalitása számos – egyenként megvásárolható – modullal állítható be a kívánt szintre. Grafikus kezelőfelülete rendkívül egyszerű és kényelmes (ráadásul nagyon sokan eleve ezt szokták meg), mellyel a beépített funkciók néhány kattintással végrehajthatóak. Cserében a bonyolultabb statisztikai problémák megoldása – noha van saját szkript-nyelve – nagyon nehézkes lehet. Összességében véve az alapvető dolgokat könnyű megcsinálni – a komplexebbeket viszont nagyon nehéz. Az SPSS-t bár sokan

használják, nincs mögötte széles, támogató nemzetközi közösség, mely érdemben bővítené a programcsomagot. Didaktikai hibái, gyatra adatvizualizációs lehetőségei, korlátozott bővíthetősége miatt nem ajánlható a használata biostatisztikai célokra.

- *R* Az R egy ingyenes és nyílt forráskódú programnyelv (http://www.rstudio.com/), egyben a talán legismertebb és legfontosabb biostatisztikai számítási környezet. Fő erejét az adja, hogy – az egyébként kezdők számára is nagyon támogató hozzáállású – virágzó nemzetközi felhasználói közösségnek köszönhetően hihetetlen mennyiségű kiegészítő érhető el hozzá a legkülönfélébb alkalmazásokhoz, AFT-modellektől a Zipf-eloszlásig, de ha valaki méhpopulációk ökológiájáról készítene statisztikát, még ahhoz is talál kész csomagot. (2025-ben több mint 22 ezer csomag érhető el, nem ritka, hogy napi 5-10 új jelenik meg!) Egy sor újonnan kifejlesztett statisztikai módszert elsőként R alatt implementálnak. E kiegészítő csomagokkal az R ereje hatalmas: rendkívül komplex feladatok is végrehajthatóak egysoros hívásokkal (néha szó szerint). Az R alapváltozatában még csak érdemi grafikus felület sincs hozzá és minden utasítást parancsként kell beírnunk; ezen segít az RStudio (szintén ingyenes és nyílt forráskódú) integrált fejlesztőkörnyezet (http://www.rstudio.com/) alkalmazása.
- *Stata* A Stata az R legfontosabb alternatívája, szintén széleskörű nemzetközi közösséggel és számos kiegészítővel, azonban az R-rel szemben nem ingyenes és nem nyílt forráskódú programról van szó. A Stata-nak van grafikus felülete, azonban a szkriptnyelve is elég erős, bár a közelében nincs az R elterjedtségének. Nem ingyenes és nem nyílt forráskódú jellege korlátozza a széleskörű használatot, illetve a nyelv néhány jellegzetessége is elég furcsa (például sokáig egyszerre csak egy adatbázist lehetett betölteni a memóriába és használni).

Jelen jegyzet mindenhol az R statisztikai programcsomagot fogja használni az elméleti mondanivaló illusztrálásához. Az R-be bevezetést nyújt a https://ferenci-tamas.github.io/r-nyelv/ címen elérhető elektronikus jegyzet.

1.5. Futó példa

A jegyzet hátralevő részében szereplő példák didaktikai okokból mind ugyanarra az adatbázisra vonatkoznak; ebben a szakaszban ezt mutatom be.

Az adatbázis egy klasszikus demonstrációs adatbázis, általánosan használt neve Low Infant Birth Weight (LOWBWT vagy BIRTHWT); a Baystate Medical Center (Springfield, Massachusetts, Egyesült Államok) kórházban végrehajtott kutatásból (1986) származik. A kutatás célja annak vizsgálata volt, hogy milyen tényezők befolyásolják, hogy egy világra jövő újszülött kis születési tömegű[2] lesz-e.

[2]Kis születési tömegről akkor beszélünk, ha az újszülött testtömege kisebb mint 2 500 gramm, akármennyi is a gesztációs kora.

Az adatbázis 189, Baystate Medical Center-ben lezajlott szülésről tartalmaz adatokat, egyrészt azt, hogy kis születési tömegű volt-e a világra jött újszülött, másrészt egy sor tényezőt, ami összefügghet a kis születési súllyal. A változók rövidítését, jelentését és mérési skáláját a 1.1. táblázat mutatja.

1.1. táblázat. A futó példa, a Low Infant Birth Weight adatbázisának változói főbb jellemzőikkel

Rövidítés	Tartalom	Mérési skála
low	Születési tömeg $< 2{,}5$ kg? [0:nem, 1:igen]	Nominális
age	Anya életkora [év]	Arányskála
lwt	Anya testtömege (UM) [font]	Arányskála
race	Rassz [1: kaukázusi, 2: afroamerikai, 3: egyéb]	Nominális
smoke	Anya dohányzik? [0:nem, 1:igen]	Nominális
ptl	Korábbi koraszülések száma [darab]	Arányskála
ht	Anyai hipertónia? [0:nem, 1:igen]	Nominális
ui	Irritábilis méh? [0:nem, 1:igen]	Nominális
ftv	Vizitek száma (1. trimeszter) [darab]	Arányskála
bwt	Születési tömeg [g]	Arányskála

Szemléltetésként az adatbázis első néhány megfigyelési egysége (az adatbázis megtalálható az R statisztikai környezet `MASS` nevű könyvtárában `birthwt` néven):

```
data(birthwt, package = "MASS")
head(birthwt, 10)
```

```
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
87   0  20 105    1     1   0  0  0   1 2557
88   0  21 108    1     1   0  0  1   2 2594
89   0  18 107    1     1   0  0  1   0 2600
91   0  21 124    3     0   0  0  0   0 2622
92   0  22 118    1     0   0  0  0   1 2637
93   0  17 103    3     0   0  0  0   1 2637
```

```
94    0   29 123     1       1    0   0   0    1 2663
95    0   26 113     1       1    0   0   0    0 2665
```

2 Deskriptív statisztika

Ebben az fejezetben a statisztika *deskriptív* (leíró) ágával fogunk foglalkozni. Már utaltam rá, hogy deskriptív statisztikáról akkor beszélünk, amikor kizárólag a mintában lévő információt igyekszünk valamilyen módon megragadni (és nem törődünk azzal, hogy a minta maga is csak a valóság egy „szelete", szebben megfogalmazva: figyelmen kívül hagyjuk a mintavételi helyzetet).

Először ezt a gondolatot fogjuk pontosítani, közelebbről körüljárni; majd pedig megismerkedünk a leíró statisztika legalapvetőbb módszereivel. Látni fogunk grafikus és analitikus módszereket, foglalkozunk egy- és (röviden) többváltozós helyzetekkel; az ismertetést pedig a vizsgált változók mérési skálája (1.2. alfejezet) szerint végezzük. (Azzal, hogy a nominális és az ordinális, illetve az intervallum- és arányskálán mért változókat nem választjuk szét, hanem minőségi és mennyiségi változókról fogunk beszélni.) Ezek után az olvasó számára ismerős lesz a mai orvostudományi cikkekben alkalmazott deskriptív eszköztár túlnyomó része; az elemi eszközöknek pedig szinte egésze.

Ebben a fejezetben a már említett módon a Low Infant Birth Weight adatbázist fogom futó példaként használni a módszertani mondanivaló illusztrálására. Az ábrák és a számítások R statisztikai környezet alatt készültek.

2.1. A deskriptív statisztikáról általában

Amint már többször említettük, a deskriptív statisztika definíciós jellemzője, hogy kizárólag a mintában lévő információval törődik, számára az az „univerzum", és teljes mértékben figyelmen kívül hagyja azt a kérdéskört, hogy a mintában lévő információ hogyan viszonyul a sokaságban lévő információhoz. Innen ered a módszer neve is: a deskripció leírást jelent, azaz a deskriptív módszerek *pusztán* a minta – valamilyen szempontból „jó" – leírását célozzák meg (nem pedig következtetést a sokaságra). Nem véletlen, hogy ebben a kontextusban nagyon sokszor minta helyett *adatbázist* mondunk (tükrözve, hogy itt igazából nincs is jelentősége annak, hogy az adataink csak – a szó statisztikai értelmében – egy mintát jelentenek).

A „jó leírás" alatt legtöbbször azt értjük, hogy a mintában lévő információt úgy próbáljuk *tömöríteni*, hogy közben – valamilyen elemzési célra tekintettel – *kiemeljük a lényeget.* Erre azért van szükség, mert a legtöbb esetben a mintában lévő információ (még ha csak néhány változóra, és néhány tucat megfigyelési egységre is gondolunk) feldolgozhatatlan „ránézésre". A számok tengeréből még a legalapvetőbb kérdésekre sem tudnánk válaszolni. Szükség van tehát olyan

módszerekre, melyek „emészthetővé teszik” ezt a számtengert: csökkentik a bonyolultságát, hogy tudjuk értelmezni azt, fel tudjuk használni kérdések megválaszolásához, illetve új megállapítások eléréséhez.

Nyilvánvaló, hogy a bonyolultság csökkentése csak úgy lehetséges, ha információt hagyunk el. Az egész művelet kritikus pontja épp ez: annak megválasztása, hogy mennyi információt hanyagoljunk el (és persze hogyan). A „hogyan” szerepe triviális: ha egy adott, mennyiségi változóra vonatkozó 100 elemű mintából elhagyjuk az első 99 elemet, akkor ugyan egyetlen számmá, azaz teljesen áttekinthetővé alakítjuk az információt – csak épp nyilván semmit nem érünk el vele. Ha viszont kiszámoljuk az átlagot, akkor ugyanúgy egyetlen számot kapunk, de immár úgy, hogy annak van értelme, azaz felhasználhatjuk kérdések megválaszolásához, illetve új megállapítások eléréséhez.

A meghatározó kulcskérdés az elhanyagolásban (az információtömörítésben) tehát a „mennyit”. Látható, hogy átváltás, trade-off áll fenn az *áttekinthetőség*, és a *reprodukciós hűség* között: minél többet hanyagolunk el, annál inkább segítjük az áttekinthetőséget, de annál többet vesztünk az eredeti információ hűséges reprodukciójából. A deskriptív statisztika igazi savaborsa (végeredményben a legtöbb módszer, így vagy úgy, de ebben foglal el egy álláspontot) a jó kompromisszum megkötése a kettő között. Példának okáért, adatbázisunkban a születési tömeg változó megfigyelései így néznek ki:

```
birthwt$bwt
```

```
  [1] 2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 2722 2733 2751 2750 2769
 [16] 2769 2778 2782 2807 2821 2835 2835 2836 2863 2877 2877 2906 2920 2920 2920
 [31] 2920 2948 2948 2977 2977 2977 2977 2922 3005 3033 3042 3062 3062 3062 3062
 [46] 3062 3080 3090 3090 3090 3100 3104 3132 3147 3175 3175 3203 3203 3203 3225
 [61] 3225 3232 3232 3234 3260 3274 3274 3303 3317 3317 3317 3321 3331 3374 3374
 [76] 3402 3416 3430 3444 3459 3460 3473 3544 3487 3544 3572 3572 3586 3600 3614
 [91] 3614 3629 3629 3637 3643 3651 3651 3651 3651 3699 3728 3756 3770 3770 3770
[106] 3790 3799 3827 3856 3860 3860 3884 3884 3912 3940 3941 3941 3969 3983 3997
[121] 3997 4054 4054 4111 4153 4167 4174 4238 4593 4990  709 1021 1135 1330 1474
[136] 1588 1588 1701 1729 1790 1818 1885 1893 1899 1928 1928 1928 1936 1970 2055
[151] 2055 2082 2084 2084 2100 2125 2126 2187 2187 2211 2225 2240 2240 2282 2296
[166] 2296 2301 2325 2353 2353 2367 2381 2381 2381 2410 2410 2410 2414 2424 2438
[181] 2442 2450 2466 2466 2466 2495 2495 2495 2495
```

Ezt a megadást nevezhetnénk az egyik végpontnak ebben a kompromisszumban: 100% reprodukciós hűség, de – szinte – 0% áttekinthetőség. Ez a legalapvetőbb kérdések megválaszolását, a legalapvetőbb észrevételek elérését is lehetetlenné teszi. (Képzeljük el, hogy mi van, ha tízezer alany lenne az adatbázisban!)

Másik végpontnak vehetjük azt, amikor a fenti adatoknak csak az átlagát adjuk meg

```
mean(birthwt$bwt)
```

```
[1] 2944.587
```

Ez nagyon alacsony reprodukciós hűséget jelent (189 számból 1-et „gyártottunk”, szinte semmit nem tudunk reprodukálni az eredeti adatbázisból), viszont remek az áttekinthetősége, látható, hogy mi a „közepes” születési tömeg.

Az igazán érdekes az, hogy – természetszerűleg – a két végpont között számos egyéb kompromisszumot köthetünk. Megadhatjuk például (az utóbbi végponttól az előbbi felé haladva) az adatok átlagát *és* szórását: 2944.5873016 ± 729.2142952, az adatok átlagát, mediánját, szórását és interkvartilis terjedelmét[1]: 2944.6 (2977) ± 729.2 (1073), vagy épp az adatok átlagát, mediánját, szórását, interkvartilis terjedelmét, illetve minimumát és maximumát: 2944.6 (2977) ± 729.2 (1073) [709-4990].

Látszik, hogy minden ilyen megadás egyfajta kompromisszum: egyre több információt őrzünk meg (egyre kevesebb az adatvesztés, hűségesebb a reprodukció), viszont közben romlik a megadás áttekinthetősége.

Összefoglalva tehát megállapíthatjuk, hogy bár az információtömörítés ugyan szükségképp adatvesztést jelent, ez azonban nem feltétlenül baj, épp ellenkezőleg: ez teszi lehetővé, hogy a fontosat észrevegyük. A kulcs a kettő közötti egyensúlyozás.

2.2. A deskriptív statisztika módszereinek csoportosításáról

Azért, hogy az igen nagy számú leíró statisztikai módszert áttekinthetően tudjuk tárgyalni, érdemes megismerkedni pár szemponttal, melyek mentén e módszerek jellegzetes, és gyakorlati szempontból fontos csoportokba sorolhatóak. Ez egyrészt segít a módszerek áttekintését, megtanulását is, másrészt jól jön később, a gyakorlati munka során is: mindig érdemes végiggondolni, hogy egy adott helyzet melyik csoportba tartozik, és ez rögtön megadja, hogy mik a szóba jövő eszközök.

2.2.1. Grafikus és analitikus módszerek

A fent mutatott példák (átlagtól szóráson át a terjedelemig) mind ún. *analitikus* eszközök voltak, azaz a (számszerű) információból számszerű, csak épp tömörebb, lényeget kiemelő információt gyártottak. Az analitikus módszerek tipikus példái a mutatószámok, mint amilyen az átlag vagy a szórás, bár léteznek ennél komplexebb (nem egyetlen számból álló) eredményt

[1] Most még nem fontos, hogy ezek a mutatók pontosan mit jelentenek (a későbbiekből úgyis ki fog derülni), csak annyi számít, hogy a minta különböző leírói.

szolgáltató analitikus eszközök is – az azonban közös pont, hogy mindegyik számszerű kimenetet ad.

Ezzel állnak szemben a *grafikus* módszerek, melyek a bemenő (számszerű) információból valamilyen képi megjelenítést konstruálnak. Szokás ezért az ilyet *adatvizualizációnak* is nevezni, bár ezt a megnevezést gyakran csak a komplexebb módszerekre alkalmazzák.

A grafikus módszerek általában kevésbé tömörek és kevésbé objektivizálhatóak (ami gond lehet, ha például összehasonlításra van szükség), de cserébe nagyon sokszor jobban értelmezhető benyomást tudnak adni a vizsgált adatbázisról. Ennek hátterében az van, hogy az emberi agy különösen alkalmas struktúrák azonosítására, vizsgálatára grafikus információkban; így ha ügyesen tudjuk vizualizálni adatbázisunk tartalmát, azzal nagyban megkönnyíthetjük az elemzését. Nem véletlen, hogy John Wilder Tukey egyszer azt mondta: „There is no excuse for failing to plot and look!" („Nincs mentség arra, ha nem ábrázoljuk az adatokat és nézünk egyszerűen rá!").

2.2.2. Egy- és többváltozós módszerek

Szemben azzal, amit sokan elsőre gondolnának, hogy ti. az egyváltozós módszerekkel egyetlen változót vizsgálunk (míg a többváltozósakkal többet), valójában **egyváltozós** módszerekkel is vizsgálhatunk akárhány változót. A különbség tehát nem ez, hanem az, hogy az egyváltozós módszerekkel *egy időben* egyetlen változót vizsgálunk csak, míg a **többváltozós** módszerek egyidejűleg is több változót tekintenek. (Ha megadjuk, hogy pontosan hányat, akkor ezt az elnevezésben is szerepeltethetjük, pl. kétváltozós vizsgálat, háromváltozós vizsgálat stb.)

Hogy mit értünk az alatt, hogy „egy időben"? Képzeljünk el egy adatbázist, melyben emberek testmagasságát és testtömegét mértük le. Okkal várjuk azt, hogy a nagyobb testmagasság tendenciájában nagyobb testtömeggel jár együtt, tehát azoknak, akiknek nagyobb a testmagasságuk, várhatóan[2] nagyobb a testtömegük is. Igen ám, de ha *önmagában csak* a testmagasságot vizsgáljuk, vagy *csak* a testtömeget, akkor ezt soha nem vennénk észre! Vegyük észre, hogy bármilyen alapos elemzést is végeznénk (beleértve akár az összes megfigyelés tömörítés nélküli felsorolását), soha nem jövünk rá erre a kapcsolatra – hiszen a külön-külön végzett vizsgálatokban nem tudjuk összerendelni az ugyanazon emberhez tartozó testmagasságot és testtömeget (épp ez a definíciója a külön-külön végzésnek). Amit tehát elvesztünk, az a változók közötti *kapcsolatok* kérdésköre. Éppen ezért mondhatjuk azt, hogy egy többváltozós vizsgálat több, mint több egyváltozós vizsgálat – hiszen itt már megjelenik a változók közötti kapcsolatok kérdése is.

Végezetül megjegyezzük, hogy a többváltozós kategóriát néha szétbontják, arra tekintettel, hogy a többváltozós elemzés klasszikus arzenálja csak egy-két tucat változóig alkalmazható hatásosan (sőt, igazán hatásosan inkább csak 10-nél is kevesebb változóra). Az e fölötti tartományban néha megkülönböztetésül **sokváltozós** adatelemzésről beszélnek.

[2] E jelenséget később pontosabban is meg fogjuk ragadni, de most bőven elég lesz ez a kissé pontatlan megfogalmazás is.

2.2.3. A vizsgált változó(k) mérési skálája

A leíró statisztika módszerei jellegzetesen eltérnek aszerint is, hogy milyen mérési skálán mért változó elemzéséről van szó. Amint már említettem is, az ordinális és nominális változókat általában nem fogjuk megkülönböztetni, és egységesen minőségi változókról fogunk beszélni, hasonlóképp az intervallum- és arányskálán mért változók esetében is egységesen mennyiségi változókról lesz szó.

2.3. Minőségi változó egyváltozós elemzése

Minőségi változóra jó példa adatbázisunk rassz (`race`) változója, mely az alany rassz szerinti hovatartozását adja meg és ilyen módon nominális. A következőkben megvizsgáljuk, hogy egy ilyen változó leírására milyen analitikus (2.3.1. pont) és grafikus (2.3.2. pont) eszközeink vannak.

2.3.1. Analitikus eszközök

Ilyen változó elemzésének tipikus analitikus eszköze a **gyakoriségi sor**. A gyakoriségi sor a változó lehetséges kimeneteit (kategóriáit) tartalmazza, együtt azzal, hogy az adott kimenet hányszor fordult elő az adatbázisban. Az ilyen „darabszámot" a statisztikában általában is **gyakoriságnak** nevezik, és f-fel jelölik. (Illetve, ha utalni akarunk arra, hogy az i-edik kategória gyakoriságáról van szó, akkor f_i-vel.) Általában n-nel szokás jelölni a mintanagyságot, így $\sum_{i=1}^{n} f_i = n$.

Szokás még beszélni **relatív gyakoriságról** is, ami nem más, mint az előbbi (abszolút) gyakoriság osztva a mintanagysággal (azaz n-nel). A relatív gyakoriság tehát azt mutatja meg, hogy egy kategóriába a megfigyelési egységek mekkora hányada esik. Természetesen $\sum_{i=1}^{n} g_i = 1$.

Példának okáért, a rassz változó gyakoriségi sora:

```
birthwt$race <- factor(birthwt$race, levels = 1:3,
                       labels = c("Kaukázusi", "Afroamerikai", "Egyéb"))
table(birthwt$race)
```

```
   Kaukázusi Afroamerikai        Egyéb
          96           26           67
```

```
prop.table(table(birthwt$race))
```

```
   Kaukázusi Afroamerikai        Egyéb
   0.5079365    0.1375661    0.3544974
```

```
cbind(table(birthwt$race), prop.table(table(birthwt$race)))
```

```
             [,1]      [,2]
Kaukázusi      96 0.5079365
Afroamerikai   26 0.1375661
Egyéb          67 0.3544974
```

Megjegyezzük, hogy a teljes relatív gyakorisági sort a statisztikusok nagyon gyakran a változó **megoszlásának** hívják.

Vegyük észre, hogy ebben a speciális esetben az információtömörítés igazából semmilyen információveszteséggel nem járt: ez a három szám *pontosan ugyanúgy* hordoz *minden* információt erről a változóról mint az eredeti 189 szám! (Az adatbázis keresztmetszeti, az alanyok felsorolási sorrendjének nincsen semmilyen jelentősége.) Ez azonban egy speciális eset, ami kizárólag a változó minőségi mivoltának volt köszönhető.

A gyakorisági soron kívül egy mutatószámnak van még értelme ennél a mérési skálánál: a **módusznak**. A módusz (jele: Mo) nem más, mint a leggyakoribb[3] kimenet (tehát az a kimenet, melyhez tartozó gyakoriság a legnagyobb az adatbázisban). Nagyon formalizálva ezt írhatnánk:

$$\mathrm{Mo} = \arg\max_i f_i.$$

A példánkban tehát a rassz módusza a kaukázusi.

Érdemes megfigyelni, hogy itt viszont *már érvényesül* a kompromisszum a hűség és az áttekinthetőség között! Nyilván még áttekinthetőbb, ha a fenti 3 szám megadása helyett annyit mondunk, hogy „a módusz a kaukázusi”, de ebben már nagyon is lesz információveszteség: nem tudhatjuk, hogy a 189-ből 189 kaukázusi vagy 64 (vagy épp 96), és semmit nem tudunk a többi kategória gyakoriságáról.

[3]Érdemes az angol mode, vagy a német die Mode szavakra gondolni: a „legdivatosabb” érték.

Végezetül megjegyezzük, hogy az ordinalitás csak annyit módosít a fentieken, hogy a gyakorisági sorban a kategóriák felsorolási sorrendje kötött[4] lesz (nominális esetben, mint amilyen a mostani példánk is volt, érdektelen, hogy milyen sorrendben adjuk meg a kategóriákat, tetszőlegesen felcserélhettük volna a sorokat anélkül, hogy az változást okozott volna).

Ami a mutatószámokat illeti, ordinális esetben elvileg már definiálható lenne a medián fogalma is, de mivel használata itt nem tipikus, a bevezetését meghagyjuk későbbre.

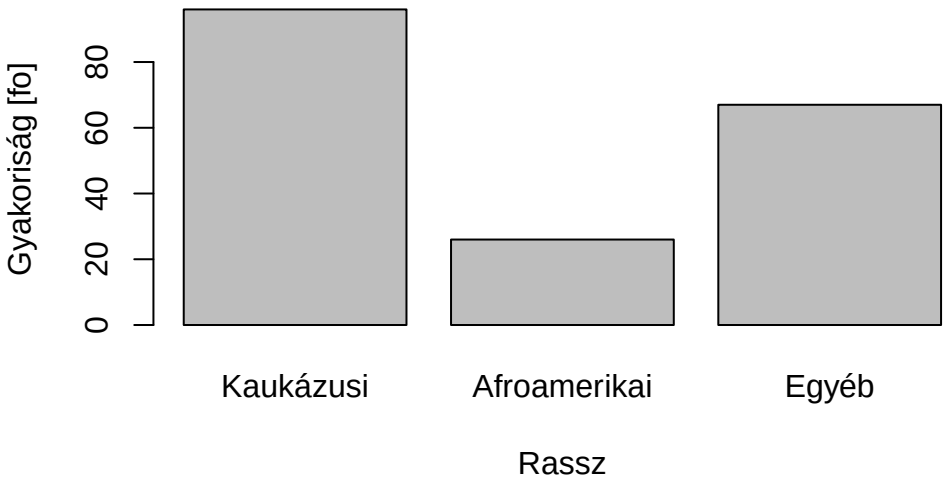
2.3.2. Grafikus eszközök

A minőségi változók grafikus elemzése lényegében a gyakorisági sor vizualizálását jelenti. Ennek két, gyakorlatban legtipikusabb eszköze az **oszlopdiagram** és a **kördiagram**. Az előbbi oszlopok magasságával, az utóbbi körcikkek területével szemlélteti a gyakoriságokat. (Bár ez utóbbi, jellegéből adódóan, igazából csak relatív gyakoriságokat tud szemléltetni. Oszlopdiagrammal gyakoriság és relatív gyakoriság is szemléltethető; sőt, a kettő lényegében ekvivalens, csak a függőleges tengely skálázása lesz más.)

Oszlopdiagramot használtunk a 2.1. ábrán.

```
barplot(table(birthwt$race), xlab = "Rassz",
        ylab = "Gyakoriság [fő]")
```

[4] Ebből a kötöttségből még egy dolog következik: lesz értelme beszélni arról is, hogy mennyi a gyakoriság egy adott kategóriá*ig*. (Nem csak adott kategóriá*ban*.) Ez nyilván értelmetlen fogalom mindaddig, amíg a kategóriák között nem értelmeztünk sorrendet. Éppen ezért ekkor bevezethető a **kumulált gyakoriság** fogalma (jele f'), mely adott kategóriára nem más, mint a gyakoriságok összege az adott a kategóriáig. (A szokásos definíció szerint: azt is beleértve.) Tehát formálisan: $f'_i = \sum_{j:C_j \leq C_i} f_j$. Hasonlóképp beszélhetünk *kumulált relatív gyakoriságról* (jele g'), mint a relatív gyakoriságok összege adott kategóriáig (azt is beleértve), tehát formálisan $g'_i = \sum_{j:C_j \leq C_i} g_j$. Nyilván $f'_{\max_j C_j} = n$ és $g'_{\max_j C_j} = 1$.



2.1. ábra. Példa egy minőségi változó ábrázolására oszlopdiagrammal.

Az oszlop- és kördiagramok használata kapcsán fontos, hogy tudományos munkákban általában az oszlopdiagram a preferált, pszichológiai vizsgálatok szerint ugyanis az emberi szem jobban tud lineáris mértékeket kezelni és értelmezni, mint területet. Az egyetlen megfontolás, ami mégis az oszlopdiagram ellen szólhat néha, hogy az oszlopok kirajzolási sorrendje már implikál egyfajta sorrendezést (a természetes balról-jobbra olvasás miatt), ami adott esetben nem következik az változó tartalmából.

Az ordinalitás e téren nem sok változást okoz: az oszlopok sorrendje kötött lesz, illetve ábrázolhatóvá válik a kumulált gyakoriság is (természetesen csak oszlopdiagrammal).

2.4. Mennyiségi változó egyváltozós elemzése

Mennyiségi változóra jó példa adatbázisunk születési tömeg (`bwt`) változója, mely az alany születési tömegét adja meg (és mint ilyen, arányskálán mért). Elsőként az ilyen változók analitikus (2.4.1. pont), majd pedig grafikus (2.4.2. pont) vizsgálati eszközeivel ismerkedünk meg.

2.4.1. Analitikus eszközök

Az analitikus eszközök közül először az osztályközös gyakorisági sort (2.4.1.1. pont), majd a különböző mutatószámokat (2.4.1.2. pont) tárgyaljuk meg.

2.4.1.1. Osztályközös gyakorisági sor

Gyakorisági sor mennyiségi változóra is készíthető, de csak módosításokkal. Annak ugyanis, hogy megszámoljuk, hogy az egyes előforduló kimenetekből mennyi van, nincs sok értelme (az itt tipikus folytonos változóknál könnyen lehet, hogy minden egyes előforduló kimenetből csak egyetlen egy lesz). A problémát a folytonosság jelenti, ami ellen úgy védekezhetünk, hogy nem adott értéket felvevő megfigyelési egységek számát adjuk meg, hanem *adott intervallumba esőek* számát. Így kapjuk az **osztályközös gyakorisági sort**. (Az elnevezés arra utal, hogy osztályközöket hozunk létre – így fogjuk hívni az előbb említett intervallumokat.) A gyakoriság, relatív gyakoriság, kumulált gyakoriság és kumulált relatív gyakoriság[5] értelmezése változatlan. A születési tömeg változó osztályközös gyakorisági sora (precízebben szólva: egy lehetséges osztályközös gyakorisági sora; hiszen ez már függeni fog az osztályközök megválasztásától is), a következő:

```
tab <- table(cut(birthwt$bwt, seq(500, 5000, 500)))
cbind(Ci0 = seq(500, 4500, 500), Ci1 = seq(1000, 5000, 500),
      fi = tab, gi = prop.table(tab))
```

```
                Ci0  Ci1 fi          gi
(500,1e+03]     500 1000  1 0.005291005
(1e+03,1.5e+03] 1000 1500  4 0.021164021
(1.5e+03,2e+03] 1500 2000 14 0.074074074
(2e+03,2.5e+03] 2000 2500 40 0.211640212
(2.5e+03,3e+03] 2500 3000 38 0.201058201
(3e+03,3.5e+03] 3000 3500 45 0.238095238
(3.5e+03,4e+03] 3500 4000 38 0.201058201
(4e+03,4.5e+03] 4000 4500  7 0.037037037
(4.5e+03,5e+03] 4500 5000  2 0.010582011
```

Itt C_{i0} és C_{i1} az i-edik osztályköz alsó és felső határát jelöli, rendre. (Az megállapodás kérdése, hogy a határon lévő megfigyelési egységeket, például egy pont 2000 grammos újszülöttet hová sorolunk, ennek természetesen csak a kerekítésből adódó diszkrétség miatt van egyáltalán jelentősége.)

[5] Emlékezzünk vissza, hogy a magasabb mérési skála minden alacsonyabb tulajdonságával bír, így természetesen az összes, alacsonyabb mérési skálán értelmezett módszer a magasabb mérési skálák esetében is alkalmazható.

Vegyük észre, hogy ez a megoldás lényegében azt jelenti, hogy a mennyiségi változónkat első lépésben „lefokozzuk” minőségi változóvá, és utána alkalmazzuk – mint teljesen közönséges minőségi változóra – a korábban megismert módszert!

Elöljáróban jegyezzük meg, hogy itt már a gyakorisági sor – szemben a minőségi esettel – igenis információvesztéssel jár: lehet 14 újszülött 1501 grammos, és lehet mind a 14 1999 grammos, mindkét esetben ugyanúgy a fenti osztályközös gyakorisági sort kapjuk. Az információvesztés mértékét az osztályközök hossza (a felosztás „finomsága”) fogja meghatározni.

A felosztás finomságára vonatkozó megjegyzés már utal arra, hogy mi az osztályközös gyakorisági sorok használatának legnagyobb kihívása: az osztályközök helyes megválasztása. A dolog azért nem könnyű – sőt, bizonyos szempontból lehetetlen – mert két, és egymásnak ellentmondó szempontnak kell megfelelni: van ok, ami miatt a minél szűkebb osztályközök a jók, és van, ami miatt a minél szélesebbek. E kérdéseket a 2.4.2.2. pontban fogjuk részletesebben megtárgyalni. Minden amit ott elmondok az osztályközök megválasztásáról, vonatkozik az osztályközös gyakorisági sorra is.

2.4.1.2. Mutatószámok

A mutatószámok a megfigyelések valamilyen jellemzőjét próbálják meg egy-egy számba tömörítve megragadni. A következőkben aszerint csoportosítva mutatom be őket, hogy mi ez a megragadott jellemző.

2.4.1.2.1. Középértékek (centrális tendencia)

A középértékek[6] egy nagyon érdekes állatfajt jelentenek: egyik oldalról ezek a leghétköznapibb mutatószámok, de valójában mégsincs igazán precíz definíciója annak, hogy mit értünk általában alattuk. Legtöbbször még a statisztika könyvek is inkább valamiféle verbális körülírással próbálkoznak, „mi körül csoportosulnak az értékek”, mi a „tipikus”, vagy „jellemző”, vagy „közepes” érték, de ez eléggé fából vaskarika, hiszen mégis mi a definíciója annak, hogy „jellemző” egy érték…? Ráadásul, mint az hamar ki fog derülni, ezek egy része még csak nem is igaz (pl. a megfigyeléseink fele 0, a másik fele 1000, akkor az 500-as átlag nemhogy nem tipikus vagy jellemző, de még csak elő sem fordul, sőt, még a környékén sincs megfigyelés). Az természetesen nagyon jó, ha az embernek van egy intuitív képe, amit a „közepes” magyar szó tényleg elég jól leír, de ezen túl nem hinném, hogy sokkal jobbat lehetne mondani, mint hogy a középérték az, amit a középérték-mutatók mérnek. És ezt egyáltalán nem viccből mondom, ennek a megfogalmazásnak fontos mondanivalója van: legyen az embernek intuitív képe, de ezen túl egész egyszerűen tudni kell, hogy pontosan mi a definíciója az adott mutatók, és az lesz a perdöntő.

Amikor azt mondtam, hogy leghétköznapibb, akkor nem csak azt értettem alatta, hogy közismert, hanem azt is, hogy bizonyos értelemben ez a legfontosabb mutatószám, jelesül, ha csak egyetlen

[6] Néha centrális tendenciáról szoktak beszélni, erős anglicizmussal.

számba kell sűrítenünk az egész eloszlást, akkor az tipikusan egy középérték. Ha több mutatót is közlünk egy eloszlásról, jellemzően akkor is a közép az, amit elsőként megadunk.

A legismertebb középérték a **(számtani) átlag**, jele $\overline{x}$. Definíciószerűen nem más, mint az a szám, amivel helyettesítve minden megfigyelési egység értékét, az ún. értékösszeg, tehát a változó megfigyeléseinek összege változatlan maradna, vagyis, amire igaz, hogy $\sum_{i=1}^{n} \overline{x} = n \cdot \overline{x} = \sum_{i=1}^{n} x_i$. Ebből már adódik, hogy az átlag:

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

Úgy is mondhatnánk: ha egyenletesen szétosztanánk az értékösszeget minden megfigyelés között, akkor ennyi jutna mindenkire.

Azonnal látható, hogy ennek akkor van értelme, ha a különböző megfigyelések számtani összege valamilyen értelmes tartalommal bír. Ez kézenfekvően megvalósul akkor, ha olyanokról beszélünk, mint például egy cég dolgozóinak átlagfizetése, vagy egy ország bányáinak átlagos széntermelése: az összeg itt az, hogy mennyi bért fizet ki a cég, vagy mennyi szenet termel az ország, ezek mind teljesen értelmes, tárgyterületi tartalommal, jelentéssel bíró számok. Érdekesebb a kérdés akkor, ha mondjuk egy osztály átlagos testtömegéről beszélünk, de kis ráolvasással ez is megindokolható (az összeg az, hogy mennyit mutatna a mérleg, ha együtt állna rá mindenki).

Amikor viszont biztosan nem alkalmas mutató az átlag, az az eset, ha az összegnek egyáltalán nincs értelme. Tipikus példa erre az, ha a változó valamilyen növekedési ütemet jelent időben: ha egy alany testtömege egy évben 1,2-szeresére nőtt, rákövetkező évben pedig 1,3-szeresére, akkor az össznövekedés nyilván nem a növekedések összege ($1{,}2 + 1{,}2 = 2{,}4$), hanem azok szorzata ($1{,}2 \cdot 1{,}2 = 1{,}44$) lesz. Ebben az esetben, tehát, ha nem az összeg, hanem a szorzat értelmes, a mértani átlag fogalmához jutunk el (az a szám, amivel a megfigyelések szorzata – nem összege – ugyanaz maradna).

A születési tömegek átlaga 2944.5873016 gramm. Meglehetősen erőltetett azt mondani (noha formailag természetesen helyes), hogy ez azt jelenti, hogy az adatbázisban szereplő újszülöttek össz-testtömege akkor maradna változatlan, ha mindegyikük 2944.5873016 gramm lenne; talán jobb, ha egyszerűen azt mondjuk, hogy ez egy középmutatója a csecsemők születési tömegei eloszlásának.

Az átlagnak két további nevezetes tulajdonsága említést érdemel. Az egyik, hogy a megfigyelések tőle vett eltéréseinek az összege zérus; ez könnyen belátható:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = n\overline{x} - n\overline{x} = 0.$$

A másik fontos tulajdonsága, hogy az összes szám közül ez az, amire igaz, hogy a megfigyelések tőle vett eltérés*négyzeteinek* az összege minimális, tehát a $\sum_{i=1}^{n} (x_i - c)^2$ kifejezés akkor

minimális[7], ha $c = \overline{x}$. Fontos megjegyezni: a négyzetre emelés eltünteti az előjelet, vagyis az eltérésnégyzet, szemben az előbbi előjeles eltéréssel, egyfajta távolság[8] – ilyen értelemben ez a megállapításunk azt mondja, hogy az átlag van a legközelebb a pontokhoz, ha ugyanannyira számít minden ponthoz a közelség! Ez egyfajta alátámasztását adja az átlag középérték jellegének. (Az összefüggés egyértelmű, ez a minimum-tulajdonság nem csak igaz az átlagra, de az átlag az egyetlen szám, amire igaz, tehát elvileg akár így is bevezethettük volna az átlagot, az értékösszeg egyenletes szétosztása helyett.)

Az átlag előnye, hogy mindenki számára közismert, kényelmesen kezelhető, bevett mutató. (Ez olyannyira erős tényező, hogy nagyon sok orvosi publikáció még akkor is erőlteti az átlag használatát, amikor az – a mindjárt részletezendő okokból – nem célszerű.)

Az átlag legnagyobb hátránya, hogy nem **robusztus**. A gyakorlatban ez két módon szokott megjelenni (matematikailag természetesen ugyanaz van a hátterében, ez ugyanazon jelenség két megjelenési formája, csak parktikus szempontokból érdemes különválasztani).

Az egyik probléma az **outlierek** ügye: így szokás hívni a csoportosulás alaptendenciájától jelentősen eltérő értéket vagy értékeket. Ez megint kettéoszlik jellegét tekintve: egyfelől előfordulhat, hogy az adatokat valamilyen eltérő mechanizmus generálta (például az alapvetően egészségesekből álló mintába bekerül néhány beteg is, akiknek a vizsgált laborváltozója lényegesen magasabb), másfelől idetartoznak az adatbeviteli hibák is. Akármelyikkel is állunk szemben, az átlag elveszíti szokásos tartalmát. Vegyük mondjuk az utóbbi példát: van 1000 újszülöttünk, és egyetlen egynél – de csak egynél – elrontjuk az adatbevitelt, mondjuk 3 kg helyett 3 tonnát írunk be a súlyaként. Ekkor (hiába is korrekt az adatok 99,9%-a!) az átlag teljesen értelmetlenné válik: az átlagos születési tömeg 6 kg lesz… (ha egyébként mindenki 3 kg körüli). Ezért nem robusztus az átlag: hiába volt az adatok abszolút túlnyomó többsége korrekt, minödssze egyetlen egy hiba elég volt ahhoz[9], hogy az átlag értelmetlenné váljon.

Az outlierek esetén két kérdés merül fel: a detekció, tehát annak azonosítása, hogy egy megfigyelés outlier (mi ennek a kritériuma?), valamint a kezelés. A fenti adathibás példa azt sugallja, hogy a „kezelés" az egyszerűen az ilyen megfigyelések törlése, és ha valóban elírás van a hátterében, úgy, hogy a valódi értéket már nem tudjuk kideríteni, akkor tényleg nem tehetünk sokkal jobbat. Ez azonban szükségessé teszi a megfelelő detekciót, így annak is lehet értelme, hogy e helyett inkább olyan statisztikai módszereket használjunk, amelyek robusztusak, azaz nem érzékenyek az outlier-ek jelenlétére (tehát például, mint láttuk, *nem* átlagot…), hiszen így nem kell – potenciálisan hibával terhelt módon – azonosítanunk, hogy egyáltalán mi az outlier, megspóroljuk ezt a definíciós problémát. Végezetül annak is lehet értelme, hogy a fentiekkel szemben ne – valamilyen módon – megszabadulni akarjunk az outlier-ektől, hanem ellenkezőleg, kimondottan megragadjuk ezeket, és egy finomított statisztikai megközelítés

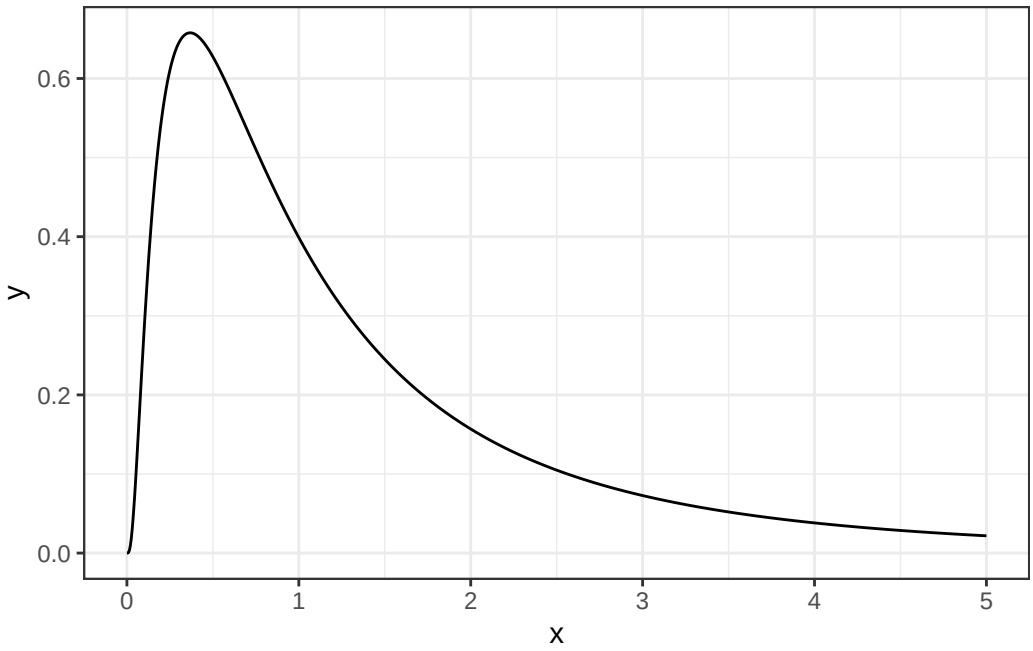
[7] A bizonyításhoz deriváljuk a kifejezést c szerint: $-2\sum_{i=1}^{n} x_i + 2nc$, ezt tegyük egyenlővé nullával és oldjuk meg c-re: $c = \frac{1}{n}\sum_{i=1}^{n} x_i$, ami valóban az átlag. A második derivált $2n$, ami pozitív, bizonyítván, hogy ez valóban szélsőértékhely és minimum.

[8] Ha valaki szeretné: L_2 távolság.

[9] Még szebben szólva: ha csak egyetlen egy megfigyelésre is $x_i \to \infty$, akkor $\overline{x} \to \infty$, *függetlenül* a megfigyelések számától.

segítségével explicite modellezzük őket is (például az egészséges-beteg laborváltozós esetben külön eloszlást illesszünk az egészségesekre és a betegekre).

A robusztusság azonban nem csak az outlierek kapcsán érdekes, sőt, a gyakorlatban nagyon fontos tud lenni akkor is, ha egy fia outlier nincs, semmilyen adatbeviteli hiba nem történt, semmilyen többitől eltérő csoportosulási tendencia nem lép fel. Mikor? Az úgynevezett ferde eloszlások esetén. A ferdeség azt jelenti, hogy az eloszlás aszimmetrikus: egyik irányban messzebre szóródik, mint a másikban. Erre orvosi példákat könnyű mondani, vegyünk egy olyan laborváltozót mint a CRP: ez egy gyulladásmarker, koncentrációja normálisan néhány mg/dl a vérben, csakhogy – és most jön a lényeg – lefelé nem tud szabadon szóródni, hiszen valaminek a koncentrációja, így negatív nem lehet. Felfelé azonban tud, hiszen ilyen felső korlát nincs, nyugodtan lehet 10, 50, 100, vagy akár annál is több az értéke. Lényegében arról van szó, hogy 0-nál egy „fal" van, ami megakadályozza a szóródást, de a kulcs, hogy ez csak egy egyik irányban történik meg – ez hozza létre a ferdeséget. Illusztratív példát mutat ilyen eloszlásra a 2.2. ábra.



2.2. ábra. Példa ferde eloszlásra

Mi fog történni, ha ilyen változónak számítjuk az átlagát? Az ábrán szereplő példában ez 1,6 körül lesz, ami csak azért „furcsa", mert első ránézésre meglepően magas: úgy érzi az ember, hogy valamiért egészen az eloszlás jobb széle felé van, egyáltalán nem a közepénél (ahol lennie „kellene", ha már egyszer középmutató). Ez bizonyos értelemben jogos, hogy számszerűsítsük a dolog: a megfigyelések nagyjából 69%-a lesz kisebb az 1,6-nél! Na de hogy lehet 1,6 az átlag, ha egyszer a megfigyelések több mint kétharmada kisebb nála?! – kérdezhetné valaki. A válasz

az, hogy nagyon könnyen: az eloszlás jobb szélén ugyan ritkán vannak megfigyelések, de azok értéke *lényegesen* nagyobb, mint a többi, ami fel fogja húzni az átlagot ugyanis – és most jön a lényeg! – a másik oldalról nem lesz, még kis számban sem, olyan érték, ami a túl oldalra kilógva tudná ezt ellensúlyozni! Ez egyáltalán nem outlier-probléma a fenti értelemben, mégis, az átlag használata megkérdőjeleződik[10].

Nagyon fontos hangsúlyozni, hogy ez a jelenség nem azt jelenti, hogy az átlag „elromlott" ezekben az esetekben. Az átlag *tényleg* annyi (ezen belül is különösen: *tényleg* 6 kg kell legyen mindenki születési tömege, hogy kiadja az összeget, amiben 3 tonna is van, *tényleg* 1,6 kell legyen mindenki CRP-je, hogy kiadja az összeget, amiben a kis számú nagy érték is benne van). Az átlag nem „rossz" ilyenkor, maximum nem felel meg annak, hogy mi a szubjektív képünk arról, hogy „közép"! De ez nem az átlag hibája; legfeljebb más mutatóra van szükségünk, ami jobban egybevág a szubjektív képünkkel[11]. (Itt is előjön, amit a felvezetőben mondtam: lehet verbális körülírásokkal élni, de egy ponton túl csak az lesz a perdöntő, hogy mi a definíció.)

Vannak bizonyos ad hoc javítások erre a robusztussági problémára, talán egyet érdemes itt megemlíteni, a **trimmelt (vagy nyesett) átlagot**: ezt úgy kapjuk, hogy elhagyjuk a legkisebb és legnagyobb adott számú elemet, és csak a maradékot átlagoljuk ki. Tipikusan az elhagyott megfigyelések száma alul és felül is a mintanagyság 2,5%-a; ebben az esetben 5%-os trimmelt átlagról beszélünk. (Bár elsőre ez szokatlan mutatónak tűnhet, és a tudományos irodalomban tényleg ritkábban is használják, de számos pontozásos sportágban épp ilyen elven alakítják ki a zsűri „átlagos" pontszámát.) A születési tömegek 5%-os trimmelt átlaga 2957.4152047 gramm, ami egyúttal azt is mutatja, lévén, hogy közel van a szokásos átlaghoz, hogy a születési tömegek aránylag szimmetrikus eloszlásúak, vélhetően komoly outlier nélkül.

Alapvetően más megközelítését jelenti a centrális tendencia megragadásának a **medián** használata, melynek jele Me_x. A medián nem más, mint a nagyság szerint sorbarendezett megfigyelések közül a középső. (Amennyiben a mintanagyság páros, úgy nyilván két „középső" is van, ez esetben megállapodás kérdése, hogy mit nevezünk mediánnak; vehetjük például a kettő átlagát.) Úgy is mondhatjuk, hogy a medián a felezőpont, az az érték, amiről elmondható, hogy alatta és felette is a mintaelemek fele található.

A medián szintén a centrális tendenciát jellemzi, csak épp kevésbé megszokott módon, mint az átlag – ez egyúttal használatának egyik fő gátja is: sok ember számára a medián tartalma (és egyáltalán, értelme) kevésbé ismert, így e mutató nem annyira jól kezelhető[12]. Előnye viszont a robusztusság, ilyen szempontból az átlaggal szemben a másik végpontot képviseli: míg az átlag extrém érzékeny volt, addig a medián extrém robusztus. Ez mind az outlier-es esetekben, mind a ferde eloszlásoknál érvényesül. A minta minden medián feletti értéke (az

[10] Bármely más ferde eloszlásnál ugyanaz a helyzet; a kérdés gyakran előjön például a keresetek kapcsán.

[11] Vegyük észre, hogy ha viszont áttérünk ilyenre, azzal az átlaghoz kapcsolódó értelmezést rontjuk el. Például az átlagos kereset úgy érezzük, hogy nem jó, mert túl nagy érték, ez esetben, ha valamilyen kisebbet mutató mérőszámot választunk, akkor elveszítjük az értékösszeg tartást: megváltozna a cég által kifizetett bértömeg, ha mindenki annyit keresne mint ez a kisebb középérték.

[12] Ami egyébként elég érdekes, már úgy értem, pusztán pszichológiailag is, hiszen valójában a medián még egyszerűbb is, mint az átlag: csak sorba kell rendezni hozzá, még csak összeadni és osztani sem kell.

egyszerűség kedvéért most gondoljunk páratlan mintanagyságra) tetszőlegesen megnövelhető (akár az összes egyszerre is), vagy a medián alatti értékek tetszőlegesen lecsökkenthetőek (akár az összes egyszerre is), vagy akár a kettő együtt is, a medián értéke *nem változik*! Az előző példánkban: hiába a 3 tonnás csecsemő, a medián marad értelmes, 3 kg körüli, a CRP esetében pedig 1 lesz a medián – nézzünk vissza az ábrára (2.2. ábra) ez valóban sokak érzete szerint közelebb van a „középhez", mint az 1,6-os átlag.

A medián hátránya (azon túl, hogy más az értelmezése, tartalma, de ez nem hátrány, csak egy jellemző), hogy a jó robusztusságért cserében kevesebb információt használ fel a mintából[13]; ezt épp a mintaértékek meglehetősen szabad „állítgathatósága" mutatja. Hogy ez miért baj, az precízen csak induktív statisztikai keretben lehet megérteni, az ottani tárgyalás után már érthető lesz, hogy mit jelent az, hogy a medián kevésbé hatásos becslő mint az átlag.

A tanulság összességében az, hogy ha előre tudható, hogy a háttéreloszlás szimmetrikus-közeli, akkor érdemes átlagot használni, ha nem, vagy outlier-ek jelenlétére is fel kell készülni, akkor jobb a medián ilyen szempontból. Az is gyakori megoldás – és ugyan az áttekinthetőségből feláldoz valamennyit, de cserében bombabiztos – ha egész egyszerűen közöljük mindkettőt.

Érdemes röviden megemlíteni, hogy a mediánnak is van távolság-optimum jellegű tulajdonsága, ráadásul egészen hasonló az átlagéhoz. Emlékeztetőül: az átlag az a szám, amire igaz, hogy a megfigyelések tőle vett távolságainak az összege minimális, ha a távolság alatt a különbség négyzetét értjük. Nos, a mediánra betű szerint ugyanez igaz, csak távolság alatt nem az eltérés négyzetét, hanem abszolútértékét[14] kell érteni! A $\sum_{i=1}^{n} |x_i - c|$ akkor minimális (és csak akkor), ha a c a medián[15]. Itt is elmondható ebből fakadóan, hogy ez újabb alátámasztását adja a medián középérték jellegének.

A születési tömegek mediánja 2977 gramm, azaz a 2977 gramm az a testtömeg, amiről elmondható, hogy az újszülöttek fele kisebb ennél, fele nagyobb. (Láthatóan közel van az átlaghoz, újból megerősítve, hogy ez valószínűleg egy szimmetrikus eloszlás, nagy outlier-ek nélkül.)

[13] Így már az is érthető, hogy a trimmelt átlag egyfajta kompromisszumnak tekinthető a kettő között, ti. a robusztusság és a mintaértékek mind teljesebb kihasználása között. Az is észrevehető, hogy bizonyos értelemben ez ráadásul általánosítja is a két mutatót: a 0%-os trimmelt átlag épp a „hagyományos" átlag, a 100%-os trimmelt átlag pedig épp a medián.

[14] Tehát ez az L_1 távolságmetrikát használja.

[15] A bizonyítás azért zűrösebb, mert az abszolútérték-függvényt nem olyan egyszerű deriválni, mint a négyzetet. Pontosabban szólva a derivált negatív számokra -1, pozitív számokra $+1$, 0-ban pedig nem értelmezett. A belső függvény deriváltja egy -1-es szorzót ad, tehát megfordítja az előbbit: negatív számokra $+1$, pozitív számokra -1; és ez van összeadva mindegyik mintaelemre. Mivel a negatív szám azt jelenti, hogy az adott mintaelem alatt vagyunk (a pozitív pedig azt, hogy felette), így az összeg egy adott ponton az lesz, hogy hánnyal több mintaelem van felettünk mint alattunk a kérdéses pontban, hiszen minden mintaelem, ami felettünk van – azaz amihez képest mi alatta vagyunk – $+1$-et ad az összeg és minden mintaelem, ami alattunk van – tehát ami felett vagyunk – pedig -1-et. Így rögtön látható, hogy a derivált ott lesz nulla, ahol pont ugyanannyi mintaelem van felettünk, mint alattunk! A teljesen precíz optimalizáláshoz még a mintaelemek pontjait meg kell nézni, hiszen ott a derivált nem volt értelmezett (ennek páratlan mintanagyságnál lesz jelentősége).

Ahogy a medián a minta „felezőpontja”, ugyanúgy definiálhatók általános osztópontok; ezeket **kvantiliseknek** nevezzük. A p-kvantilis ($0 < p < 1$) az az érték, amiről elmondható, hogy a megfigyelések p-ed része kisebb nála, $(1-p)$-ed része nagyobb nála. (Tehát a medián az 1/2-kvantilis.) Gyakorlati szempontból nagyobb jelentősége van még a negyedelőpontoknak, melyek neve **kvartilis**. Ilyenből tehát három van: a $p = 1/4, 2/4 = 1/2, 3/4$-kvantilis, ezek közül a középső persze ugyanaz mint a medián. A másik kettőt alsó és felső kvartilisnek szokták nevezni, és Q_1-gyel, illetve Q_3-mal jelölik. Tehát például Q_1 az a szám, amire igaz, hogy a minta egynegyede nála kisebb értékű, háromnegyede nála nagyobb. Ezek valójában már nem is a centrális tendenciát, hanem általában az eloszlás alakját mutatják, mégpedig robusztus módon (ugyanazon okból, mint amiért a medián is robusztus). Ritkábban, de szokták használni ugyanerre a célra a tizedelőpontokat, nevük decilis ($D_1, D_2, \ldots, D_9$) és a századolópontokat, nevük percentilis[16] ($P_1, P_2, \ldots, P_{99}$). A 90. percentilist például gyakran használják olyankor, ha az eloszlás széli viselkedésének jellemzésére van szükség – mi a legrosszabb eshetőség? – amit első ránézésre a maximum mutatna, csak a 90. percentilis sokkal robusztusabb: a maximum nagyon ingadozó, olyan értelemben, hogy egyetlen érték is odébbhúzza (az nagyon esetleges lehet, hogy pont a legnagyobb mennyi); a 90. percentilis viszont továbbra is az eloszlás szélét méri, de kevéssé ingadozó, sokkal robusztusabb módon[17].

A korábbi értelemben vett módusz használatának a folytonosság miatt általában nincs értelme mennyiségi változó esetén, hiszen még az is lehet, hogy minden konkrét értékből csak egyetlen egy fordul elő, ahogy arról már volt is szó. Folytonos változó esetén emiatt a módusz többé nem a leggyakoribb kimenet, hanem a sűrűségfüggvény maximumhelye[18] – csakhogy sűrűségfüggvényünk nincsen, azt legfeljebb közelíteni tudjuk a mintából, valamilyen simítóeljárással (részletesebb lásd a grafikus módszerek között, a 2.4.2. pontban). Ez lehet egy sima osztályközös gyakorisági sor / hisztogram (az is egyfajta simítás!), ez esetben modális osztályközről szokás beszélni, mint a legnagyobb gyakoriságú osztályköz / a hisztogram legmagasabb oszlopa, de használhatunk simításra magfüggvényes sűrűségbecslőt is (2.4.2.3. pont), ez esetben a maximum egyetlen pont lesz. Ez azonban még deskriptív statisztikai értelemben is csak egy közelítés. Az ilyen módon vett módusz használata ritka a mindennapi biostatisztikai gyakorlatban.

Ennek kapcsán még annyit megjegyzek, hogy átlagot, mediánt (és általában minden egyéb mutatószámot is) lehetséges osztályközös gyakorisági sorból, a nyers mintaelemek ismerete

[16] Különösen az angol irodalomban a percentilist nagyon gyakran szinte a kvantilis szinonimájaként használják. Ez csak szóhasználati könnyebbség, amennyiben ahelyett, hogy 0,123-kvantilis jobban hangzik azt mondani, hogy 12,3 percentilis.

[17] Azért vegyük észre, hogy igazából ez is egy kompromisszum: elvileg még jobb lenne a 99., a 99,9., a 99,99. stb. percentilis, csak ezeknél megint vissza fog jönni az a probléma, hogy csak nagy ingadozással lesznek meghatározhatóak, hacsak nincs hatalmas mintánk.

[18] Az érdekesség kedvéért megjegyzem, hogy valójában még a módusz is beleszuszakolható a „távolság-minimalizálási” keretbe: az átlagot akkor kaptuk, ha az L_2 távolságok összegét minimalizáltuk, a mediánt akkor, ha az L_1 távolságokét – nos, a móduszt akkor kapjuk, ha az L_0 távolságokét! Ez kicsit már feszegeti a szokásos kereteket, azt értjük alatta, hogy az L_p távolság akkor, ha $p \to 0$; belátható, hogy ez 1 értékű akkor, ha a két pont, aminek a távolságát nézzük, egybeesik, 0 különben. Az összeg tehát azt fogja jelenteni adott pontra, hogy hány megfigyelés van, ami *nem* az adott pontban található (függetlenül attól, hogy mennyire távol vagy közel), innen már nyilvánvaló, hogy ezt valóban az minimalizálja, ha a pontot oda tesszük, ahol a legtöbb megfigyelés van.

nélkül is számolni, persze ekkor már csak közelítő jelleggel.

2.4.1.2.2. Szóródás

Szóródásnak nevezzük azt, hogy a megfigyelések milyen szorosan csoportosulnak azon érték körül, ami körül csoportosulnak (lásd a centrális tendenciát!), más szóval mennyire ingadoznak a megfigyelések, mekkora változékonyság van bennük. A gyakorlatban ez a második legfontosabb kérdés: ha csak egy jellemzőt adhatunk meg, akkor az a centrális tendencia lesz, de ha kettőt, akkor megadjuk azt is, hogy mekkora a szóródás.

A minta szóródásának legegyszerűbb mérőszáma a legkisebb (Min) és a legnagyobb (Max) mintaelem értéke, a **mintaminimum** és **mintamaximum**, illetve kettejük különbsége, melyet **terjedelemnek** nevezünk és gyakran R-rel jelölünk: $R = \text{Max} - \text{Min}$. Ezek előnye, hogy teljesen egyértelmű a tartalmuk, hátrányuk, hogy rendkívül érzékenyek arra, hogy konkrétan milyen mintát vettünk a sokaságból, ezért, bár gyakran megadják információ gyanánt, a szóródás számszerű jellemzésére ritkán használják.

A születési tömegek mintaminimuma 709 gramm, mintamaximuma 4990 gramm, így e változó terjedelme 4281gramm.

Az egyik alapvető mutatója a szóródásnak a **szóránégyzet** (vagy **variancia**), jele általában s_x^2. A szórásnégyzet tulajdonképpen a legkézenfekvőbb jellemzője a szóródásnak, hiszen nem más, mint a átlagtól vett átlagos eltérés. Az egyetlen amire vigyázni kell, hogy az eltérés alatt mit értünk: ha egyszerűen a megfigyelés és az átlag különbségét vennénk, az nem lenne jó, mert a pozitív és a negatív eltérések csökkentenék (sőt, belátható, hogy kioltanák) egymás hatását. Ezért inkább négyzetre emeljük ezt a különbséget, hogy megszabaduljunk az előjeltől, hogy a $+1$ és a -1 eltérés hatása ugyanolyan legyen, és ezzel már jók vagyunk:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}.$$

A szórásnégyzet problémája, hogy a mértékegysége nem ugyanaz, mint az eredeti változóé (a négyzetremelés miatt). Ha szeretnénk a szóródást ugyanazon a skálán jellemezni, akkor egy gyökvonással visszatérhetünk; ezt a mutatót hívjuk **szórásnak**, jele s_x:

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}.$$

(A kettő neve nem keverendő: a „szóródás” a jellemző, a „szórás” egy lehetséges mutatószáma ennek a jellemzőnek.)

Deskriptív esetben néha inkább mintavarianciát, illetve mintaszórást mondanak (hogy a megfelelő sokasági jellemzőtől megkülönböztessék – sajnos nincs akkora szerencsénk, mint az

átlagnál, ahol az „átlag” és a „várható érték” révén két teljesen különböző szavunk van a mintabeli, tehát statisztikai és a sokasági, tehát valószínűségszámításos fogalomra).

A fent definiált mutatót szokás precízen *korrigálatlan* mintavarianciának, illetve mintaszórásnak nevezni, ezzel szemben a *korrigált* mutatóban nem n-nel, hanem $n-1$-gyel osztunk le. Például a korrigált mintaszórás, jele s_x^*:

$$s_x^* = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}.$$

A különbségük oka csak a következtető statisztikában válik világossá[19].

A születési tömegek korrigált mintaszórása 729.2142952 gramm, tehát az újszülöttek testtömegeinek átlaguk körül vett ingadozásának átlaga 729.2142952 gramm.

A szórás hátránya, hogy – az átlaghoz hasonlóan – nem robusztus[20] mutató. (Egyrészt azért nem, mert az átlagtól vett eltérések nézi, ami eleve nem robusztus középmutató, másrészt ezeket átlagolja, ami ugyebár nem robusztus, ráadásul a négyzetreemelés még ki is emeli a különbségeket.) Egyik gyakran alkalmazott robusztus alternatíva az **interkvartilis terjedelem** (jele IQR), ami a felső és az alsó kvartilis különbsége:

$$IQR = Q_3 - Q_1.$$

Mondhatjuk azt is, hogy az IQR az adatok középső 50%-ának szélessége.

Az interkvartilis terjedelem a robusztus kvartiliseken alapul, így robusztus mutató, és könnyen látható, hogy tartalmilag a szóródást jellemzi, hiszen minél jobban szóródott az eloszlás, annál távolabb lesz az alsó és a felső negyedelőpontja.

A születési tömegek interkvartilis terjedelme 1073 gramm, tehát az a tömeg, ami fölött az újszülöttek egynegyede (és alatta háromnegyede) van, 1073 grammal nagyobb annál a tömegnél, ami fölött az újszülöttek háromnegyede (és alatta egynegyede) van, tehát 1073 gramm szélességben szóródik a születési tömegek középső 50%-a.

Egy másik lehetőség a szórás „megjavítása”, olyan módon, hogy kiküszöböljük a nemrobusztusság fent említett forrásait: az eltéréseknek nem a négyzetét, hanem az abszolút értékét vesszük, az eltéréseket nem az átlagtól hanem a mediántól vesszük, végül pedig nem is átlagoljuk őket, hanem a mediánjukat képezzük. Az így kapott mutató neve **medián abszolút eltérés**, jele MAD, tehát

[19] A korrigálatlan mintavariancia, első ránézésre talán meglepő módon, *nem* torzítatlan becslője a sokasági varianciának, ezzel szemben a korrigált igen. A mintaszórásnál sajnos nem ilyen egyszerű a helyzet, ott a korrigált mutató is torzított (igaz, arra nem is létezik becslő, ami általában torzítatlan lenne). E kérdésekkel a következtető statisztikánál fogunk foglalkozni részletesen.

[20] Konkrétabban beszélve, például erre is igaz, hogy ha csak egyetlen egy megfigyelésre is $x_i \to \infty$, akkor $s_x \to \infty$, *függetlenül* a mintanagyságtól.

$$MAD = \mathrm{Me}\left(|x_i - \mathrm{Me}\left(x\right)|\right).$$

(A szakirodalom itt nem teljesen egyértelmű: néha MAD-nak nevezik azt a mutatót is, ahol csak az első javítást csinálják meg, tehát abszolútértéket vesznek, de azokat továbbra is csak átlagolják, és az eltéréseket is az átlagtól veszik.)

A születési tömegek medián abszolút eltérése 563 gramm, tehát az újszülöttek testtömegeinek mediánjuk körül vett (abszolút) ingadozásának mediánja 563 gramm.

2.4.1.2.3. Alakmutatók

A fenti két jellemzőn túlmenően néha egyéb, még inkább részletekbe menő jellemzőit is használják egy változó leírásának. Egy példát tulajdonképpen már láttunk is, a ferdeséget (szimmetriát), amire léteznek numerikus mutatók, melyek jellemzik az irányát és a mértékét. Vannak további mutatók, további statisztikai jellemzők (például csúcsosság) numerikus leírására, de ezeket, bár statisztika tankönyvek néha tartalmazzák, a mindennapi biostatisztikai gyakorlatban ritkán alkalmazzák, így most nem is részletezem bővebben.

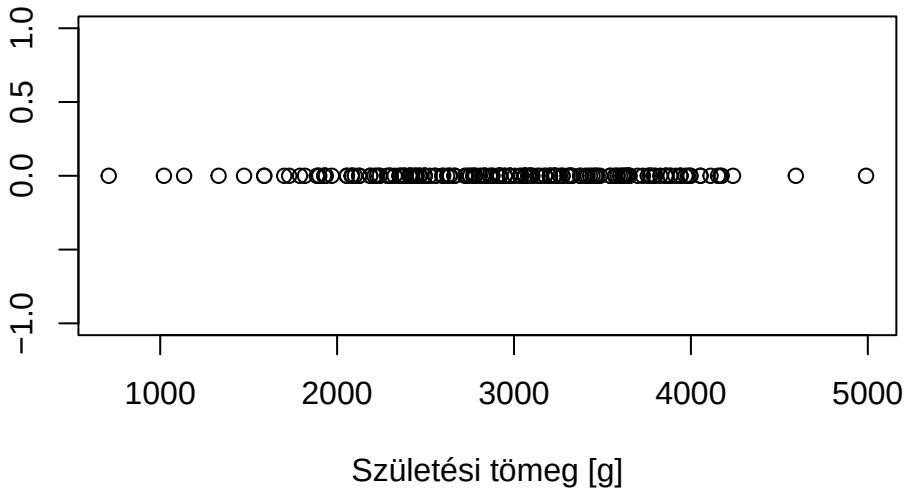
2.4.2. Grafikus eszközök

A grafikus eszközök közül először az egydimenziós szóródási diagramot (2.4.2.1. pont), aztán a hisztogramot (2.4.2.2. pont), utána a magfüggvényes sűrűségbecslőt (2.4.2.3. pont), majd végül a boxplotot (2.4.2.4. pont) tárgyaljuk meg.

2.4.2.1. Egydimenziós szóródási diagram

A történetet kezdjük ott, hogy van eszköz arra, hogy információveszteség nélkül vizualizáljunk mennyiségi változót is. Ez nem más, mint az egydimenziós szóródási diagram – az elnevezés oka később világosabb lesz – vagy angolul stripchart. Az ábrázolás nagyon egyszerű, fogunk egy számegyenest és minden megfigyelés értékéhez egy kis jelzést rakunk:

```
plot(birthwt$bwt, rep(0, nrow(birthwt)),
     xlab = "Születési tömeg [g]", ylab = "")
```

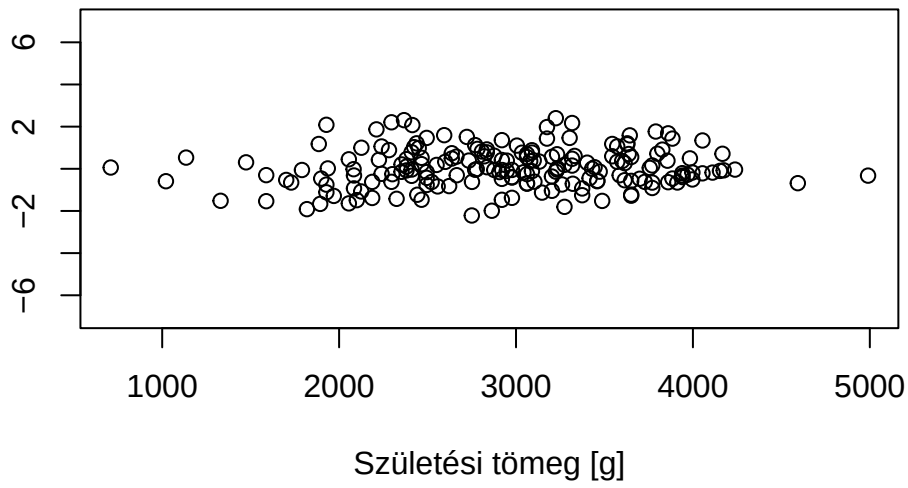


2.3. ábra. Példa egy mennyiségi változó ábrázolására egydimenziós szóródási diagrammal.

Ahogy ígértem, ezen valóban veszteség nélkül látszik minden információ, ami az eredeti változóban benne van, ami jól hangzik, de két problémája mégiscsak van ennek az ábrának. Az egyik, hogy a pontok – különösen az ábra közepe táján – elkezdenek egymás tetejére kerülni, így nagyon áttekinthetetlen lesz az ábra. Valójában az áttekinthetetlenségnél többről van szó: ha egyszer egy pont fekete lett a jelöléstől, akkor onnantól nem tudhatjuk, hogy ott 2, vagy 200 érték található egybeesve. Erről a problémakörről később, a szokásos szóródási diagramoknál (2.6.2. pont) jóval hosszabban fogunk beszélni.

Szerencsére most van egy nagyon egyszerű megoldási lehetőségünk. A probléma az, hogy valahogy „szét kellene szedni” az egybeeső pontokat, és erre most adja magát egy kézenfekvő lehetőség, van ugyanis egy dimenziónk, amit nem használtunk ki! Jelesül, a függőleges tengely: annak nincsen jelentősége (akárhol található függőlegesen egy pont, az ugyanazt jelenti, csak a vízszintes koordinátája számít), úgyhogy miért nem használjuk ki azt, hogy kicsit szétszedjük a pontokat? Ha már úgysincs jelentősége, akkor nyugodtan választhatunk egy véletlenszámot is függőleges koordinátaként:

```
plot(birthwt$bwt, rnorm(nrow(birthwt)),
     xlab = "Születési tömeg [g]", ylab = "", , ylim = c(-7, 7))
```



2.4. ábra. Példa egy mennyiségi változó ábrázolására jitter-elt egydimenziós szóródási diagrammal.

Ezt a módszert, tehát, amikor egy pici véletlen zajt keverünk a pontokra, hogy megszüntessük az egybeeséseéket, jitter-elésnek szokták nevezni. Később látni fogunk még rá a fentinél jóval nagyobb gyakorlati súlyú példát is.

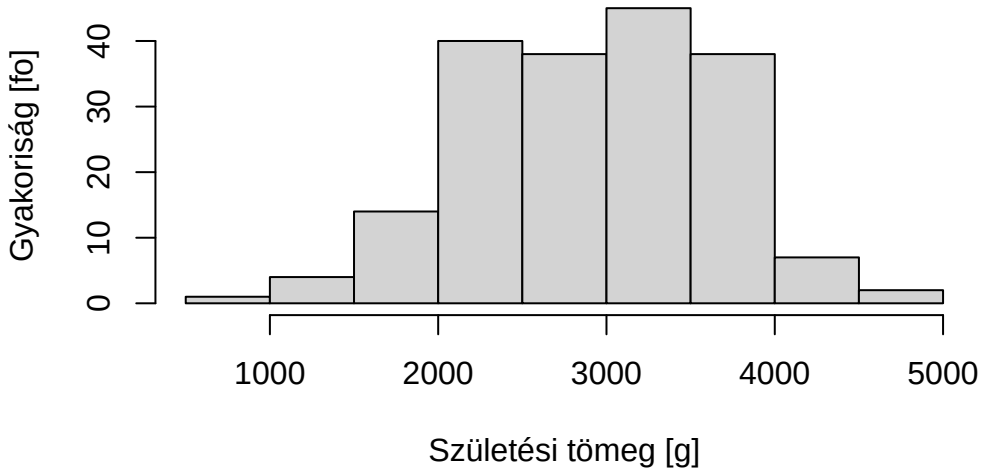
Ezzel már egy jó ábrát állítottunk elő, bár azt meg kell jegyezni, hogy ha még sokkal több pont van, akkor ez a jitter-elős módszer sem fog működni, valójában tehát az információvesztés nélküliség ez esetben inkább csak elméleti tulajdonság.

Van azonban egy másik probléma, ami még ezt a javított ábrát is sújtja, és emiatt a fontosabb – és általánosabb – gond a kettő közül. Jelesül: az ábra ugyan *érzékelteti*, hogy hol vannak sűrűbben a pontok, de nem ábrázolja ezt a sűrűséget közvetlenül. (A „közvetlenül” alatt az értem, hogy a sűrűség nem egy leolvasható érték, nincs egy skála meg egy jelölés, ami mutatja, hogy itt ennyi a sűrűség.) Ami baj: sokkal jobban tudunk értelmezni olyan ábrát, amin a vizsgált jellemző direkte leolvasható, és nem csak érzékelhető valamilyen áttételes vizuális módon. Érdekes módon ez annyira erősen így van, hogy ha egy ábra ezt megvalósítja, akkor azzal még akkor is jobban járunk, ha egyébként közben – a fentivel szemben – információvesztéssel jár. A következő pontok erre mutatnak példát.

2.4.2.2. Hisztogram

A hisztogram lényegében nem más, mint az osztályközös gyakorisági sor ábrázolása oszlopdiagramon, annyi specialitással, hogy az oszlopokat közvetlenül egymás mellé rajzoljuk, hely kihagyása nélkül (2.5. ábra).

```
hist(birthwt$bwt, xlab = "Születési tömeg [g]",
     ylab = "Gyakoriság [fő]", main = "")
```

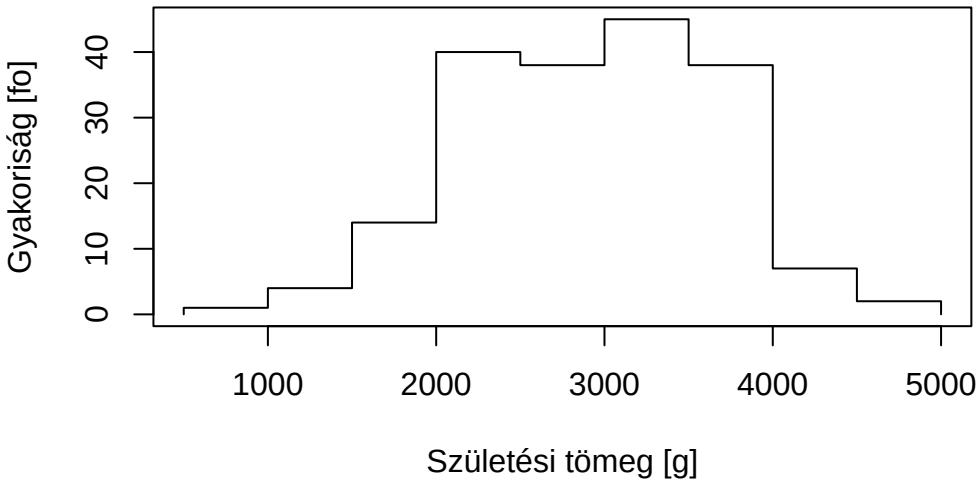


2.5. ábra. Példa egy mennyiségi változó ábrázolására hisztogrammal.

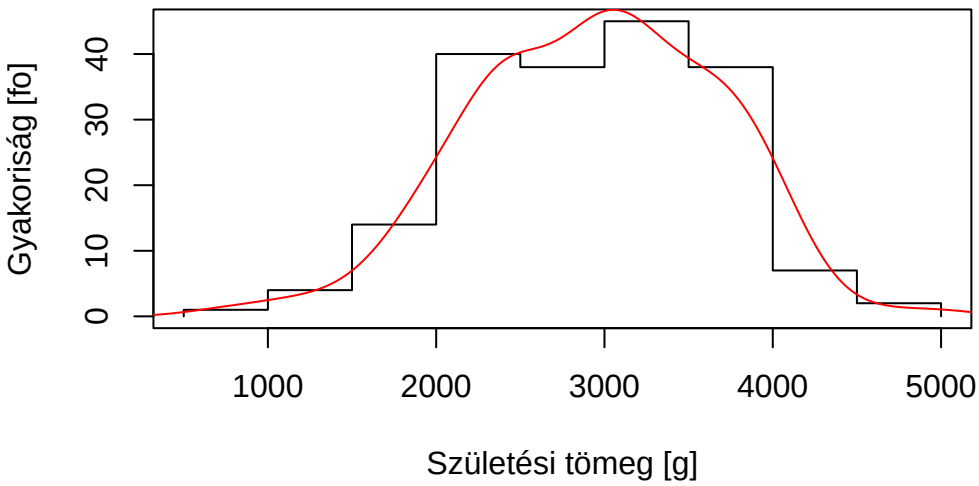
A hisztogram a háttérben lévő változó eloszlását (sűrűségfüggvényét) igyekszik közelíteni[21]. Ez azt is jelenti, hogy a fenti ábrán lévő függőleges vonalak igazából csak grafikai elemek, a hisztogram által reprezentált függvényben természetesen nem számítanak, az úgy néz ki, ahogy a 2.6. ábra mutatja.

A hisztogram tehát egy szakaszonként konstans, lépcsős függvényt ad – így írja le az eloszlást (ezzel igyekszik közelíteni a valódi sűrűségfüggvényt). A probléma csak az, hogy a valódi eloszlás szinte minden realisztikus esetben szép folytonos, simán változó, szakadások nélküli függvény – valahogy úgy, ahogy a 2.7. ábra mutatja.

[21] A „közelíteni” természetesen azt jelenti, hogy „becsülni”, de az már egy induktív statisztikai fogalom.



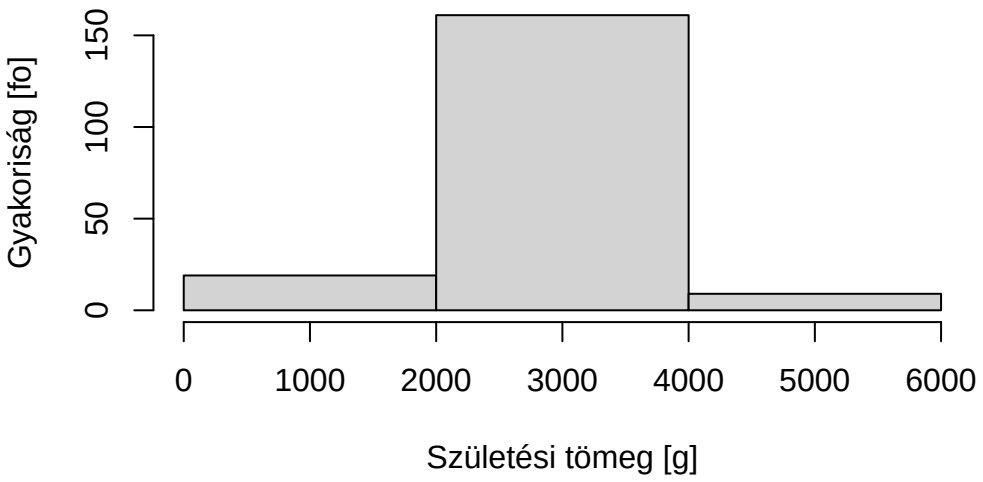
2.6. ábra. A hisztogram által reprezentált függvény.



2.7. ábra. Hisztogram, hátterében a valódi eloszlás

Nagyon fontos hangsúlyozni, hogy ez csak egy *illusztráció*: a valóságban mi magunk sem tudhatjuk, hogy mi az eloszlás a háttérben! Pont ez a lényeg, ami miatt szükség van egyáltalán a hisztogramra. Az ábra tehát pusztán szemléltetni kívánja, ahogy a hisztogram egy lépcsős függvénnyel igyekszik leírni egy – általunk sem ismert, de mindenesetre nem lépcsős – függvényt.

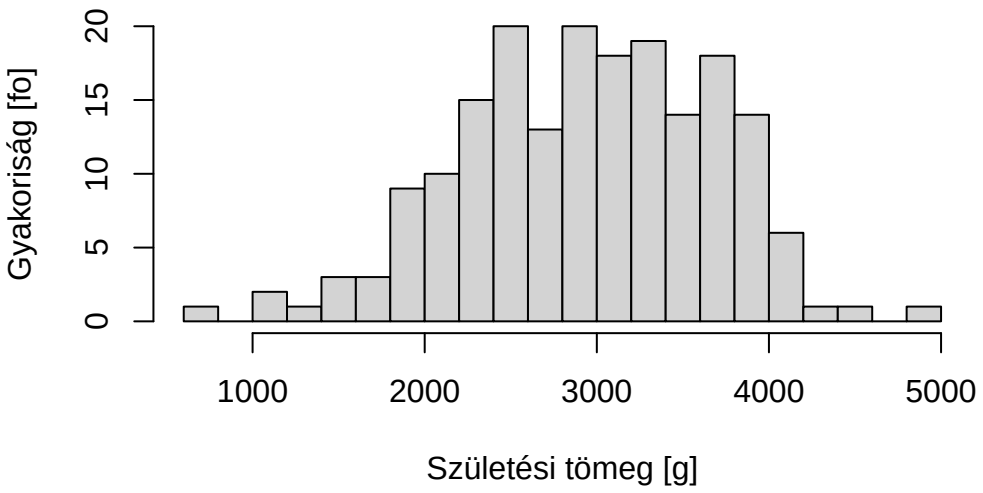
Mi következik mindebből? Először is az, hogy ez a leírás annál pontosabb tud lenni, minél szűkebbek az osztályközök – annál finomabb a lépcsős függvény, annál jobban tud követni egy (bármilyen alakú) valódi függvényt! Ez tehát a szűk osztályközök mellett szól, avagy, fordítva megfogalmazva, a nagyon széles osztályközök azért lesznek rosszak, mert lehetetlen lesz vele követni a valódi függvényt, mert össze fog mosni különböző dolgokat. (Később úgy fogjuk mondani: torzított lesz. Érdemes összevetni ezt azzal, amit az osztályközös gyakorisági soroknál mondtunk: ilyenkor nagy lesz az információvesztés.) Ezt mutatja a 2.8. ábra, ahol nagyon széles osztályközöket választottam a születési tömegek ábrázolására.



2.8. ábra. Születési tömegek hisztogramja, széles osztályközökkel

Akkor tehát vegyünk fel nagyon szűk osztályközöket? Ez a fenti szempontból jó választásnak tűnik (jól tudjuk követni az igazi függvényt, nem mosunk össze különböző dolgokat, kicsi a torzítás), de okoz egy másik problémát. Ez jól látszik a 2.9. ábrán, ahol szűk osztályközökkel ábrázoltam a születési tömegeket.

Mit látunk az ábrán? Nézzük meg a hisztogramot hogyan alakul, ahogy haladunk balról jobbra, hogyan változik az értéke ahogy egyesével ugrálunk az oszlopokon jobbra: felmegy, felmegy, felmegy, lemegy, felmegy, lemegy, felmegy, lemegy, felmegy, lemegy, felmegy… Ennek aligha



2.9. ábra. Születési tömegek hisztogramja, szűk osztályközökkel

van biológiai realitása, hogy a *valódi* születési tömeg eloszlása *tényleg* így változik! Akkor mi történik? A magyarázat az, hogy ha szűk osztályközöket választunk, akkor az egy osztályközbe jutó megfigyelések száma (ami alapján ugye meghatározzuk az oszlop magasságát!) nagyon kicsi lesz. Hogy ez miért probléma, azt igazán majd csak a következő statisztikából fogjuk megérteni, most annyit mondok, hogy nagyon ingadozó lesz az oszlop magassága a véletlen szeszélyétől függően is (gondoljunk arra, hogy ha egy osztályközbe 2 pont esik, akkor mindössze egyetlen érték kicsit odébbmozdítása is lehet, hogy megfelezi, vagy épp másfélszeresére növeli az oszlop magasságát!). Voltaképp az történik, hogy van a valódi függvény, a hisztogram ingadozik körülötte – és minél szűkebbek az osztályközök, annál nagyobb ez az ingadozás. Ezt szép szóval úgy mondják, hogy a variancia a probléma.

Összefoglalva a helyzetet: a széles és a szűk osztályközök mellett is szólnak érvek. Lehet a kettő között egy optimumot találni, de az csak optimum (legkisebb hibájú helyzet), nem ideális (nulla hibájú helyzet). Ez a konkrét eset egy példa arra a statisztikában több helyen is előforduló jelenségre, amit torzítás-variancia dilemmának hívnak; lesz még róla szó.

(Zárójelben: érdemes észrevenni, hogy az egyetlen megoldás, ami nem kompromisszumos, tehát nem valaminek a feláldozása árán javítja az egyik szempontot – ugye az osztályközök szűkítése vagy bővítése ilyen! – az a mintanagyság növelése. Ha tudjuk növelni a mintanagyságot, akkor ugyanolyan széles osztályközökbe is több pont jut, tehát a torzítás növelése *nélkül* csökkentjük a varianciát, avagy ugyanannyi osztályközbe jutó pontot megtarthatunk szűkebb osztályközökkel is, tehát a variancia növelése *nélkül* csökkenthetjük a torzítást. Vagy valahol a kettő között, ha

kisebb mértékben is, de egyszerre javíthatjuk mindkettőt.)

A probléma tehát kettős: egyrészt még ha meg is találjuk az optimumot, az sem lesz tökéletes, másrészt meg kell találni valahogy ezt az optimumot. A gyakorlatban ez utóbbi a probléma (pláne, hogy az első elkerülhetetlen, így azzal nincs mit tennünk). Hogyan találjuk meg tehát az optimális osztályköz-szélességet? Ez a legnagyobb kihívás a hisztogram használatakor. A kérdés egyáltalán nem mellékes: mint a fenti ábrák is mutatják, az, hogy a gyakorisági sor milyen képest sugall számunkra a vizsgált változóról sajnos nagyban függhet az osztályközök megválasztásától, különösen kis mintanagyságnál.

A gyakorlatban két megközelítése van a problémának. Az egyik lehetőség, ha *szakmai alapon* választjuk meg az osztályközöket: úgy rakjuk le az osztópontokat, hogy az osztályközök valamilyen értelmes, tárgyterületi tartalommal bíró kategóriákat jelöljenek ki. Kis költői szabadsággal élve mondhatjuk, hogy már a 2.5. ábra hisztogramján is ez történt – az osztópontok szép kerek, 500-al osztható számokra kerültek. A valódibb példa természetesen az, ha az osztályközök valamilyen szakmai tartalommal bírnak. Példának okáért, a szülészetben általánosan bevett módon használják azt a terminológiát[22], hogy kis születési súlyról beszélnek 1500 és 2499 gramm között, igen kis születési súlyról 1000 és 1499 gramm között és igen-igen kis (extrém kis) születési súlyról 1000 gramm alatt – megtehetjük, hogy az osztópontokat is így rakjuk le, így az osztályközök az általánosan használt szülészeti kategóriáknak fognak megfelelni.

A másik lehetőség, hogy *statisztikai alapon* választjuk meg az osztályközöket: nem is törődünk a változó tárgyterületi tartalmával, egyszerűen azt nézzük, hogy pusztán a megfigyelések statisztikai jellemzőit figyelembe véve mi az a választás, ami optimális a torzítás-variancia dilemma tükrében, tehát az össz-hibát minimalizálja. Esetleg reménykedhetünk abban, hogy erre van egyetlen, egyértelmű válasz, tehát, hogy megcsináljuk az optimalizálást, majd a végén kijön, hogy mi „az” optimum, de sajnos ez nem így van. A probléma az, hogy bár lehet ilyen levezetéseket csinálni, de az eredmény függeni fog attól is, hogy pontosan milyen feltételezésekkel élünk, így valójában a válasz nem egyértelmű. Példának okáért, az egyik ilyen ismert analitikus szabály a Sturges-szabály, ami azt javasolja, hogy $\lceil \log_2 n + 1 \rceil$ darab azonos szélességű osztályközt vegyünk fel a mintaminimum és -maximum között.

A fentiekben elmondottak természetesen nem csak a hisztogramra érvényesek, hanem az osztályközös gyakorisági sor kontrukciójára is (2.4.1.1. pont).

Végezetül beszéljünk kicsit arról, hogy hogyan lehet konkrétan R-ben megadni az osztályközöket. Erre 4 lehetőségünk is van:

1. Explicite megadjuk az osztályközök határait: `hist(birthwt$bwt, breaks = c(500, 1500, 2000, 2500, 2750, 3000, 3250, 3500, 5000))`. Az R ezt feltételezi akkor, ha a `breaks` argumentum értéke egy vektor.
2. Megadjuk az osztályközök számát: `hist(birthwt$bwt, breaks = 10)`. Az R ezt feltételezi akkor, ha a `breaks` argumentum értéke egy szám (persze, mint tudjuk az R-ben ez azt jelenti, hogy vektor, de 1 hosszúságú).

[22] https://real.mtak.hu/86899/1/650.2018.31199.pdf

3. Megadjuk a szabály nevét, amivel kérjük az osztályközök számának kiszámolását: `hist(birthwt$bwt, breaks = "Sturges")`. Az R ezt feltételezi akkor, ha a `breaks` argumentum értéke egy sztring.
4. Saját függvényt adunk meg, mely az osztályközök számát adja vissza (bemenetként a megfigyeléseket kapja meg). Lényegében implementálhatunk egy saját szabályt a beépítettek mellett. Az R ezt feltételezi akkor, ha a `breaks` argumentum értéke egy függvény.

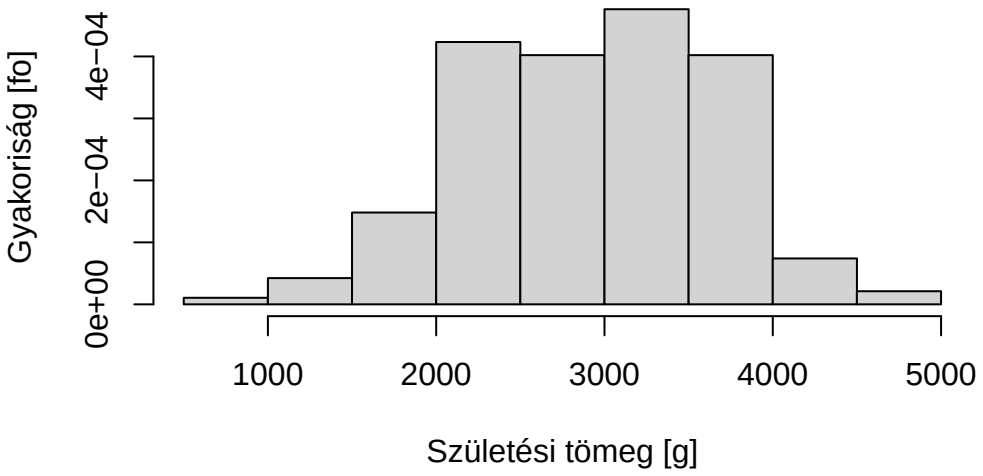
Megjegyzendő, hogy az utolsó 3 esetnél a beállítást az R csak ajánlásnak tekinti: a konkrét osztópontokat ilyenkor kicsit odébbmozgathatja, ugyanis ilyenkor arra is törekszik, hogy szép kerek számokra rakjak az osztásokat (a `pretty` függvény használatával).

Most, hogy kiveséztük az oszlopok szélességeit, beszéljünk picit a magasságukról is!

A hisztogramokra adott definíció alapján („az osztályközös gyakorisági sor oszlopdiagrammal ábrázolva, csak oszlopok közötti rések nélkül") alapján ez egyszerűnek tűnik: az oszlopok magassága f_i, tehát az adott intervallumba eső megfigyelések száma. Ez lehet teljesen működőképes, hogy mást ne mondjak, az összes eddigi hisztogram így készült, egy baja azonban van: a görbe alatt terület (emlékezzünk a 2.6. ábrára!) nem 1 lesz. Miért érdekes ez? Az esetek túlnyomó részében semmiért, egy orvosi közleményben tökéletesen értelmesek és értelmezhetőek a fenti hisztogramok (sőt, valószínűleg így értelmezhetőek a legjobban). Ha azonban valamiért komolyan kell vennünk azt a kitételt, hogy a hisztogram a sűrűségfüggvényt igyekszik közelíteni, akkor baj van: a sűrűségfüggvény görbe alatti területe viszont biztosan 1, tehát, ha a hisztogramé nem, akkor itt valami biztosan nem stimmel[23]. A problémán azonban könnyen segíthetünk. Mennyi a hisztogram görbe alatti területe, ha nem 1? Nagyon egyszerű: egy téglalap területe $h \cdot f_i$, ha h az osztályközök szélessége, így a teljes görbe alatti terület $\sum_i h \cdot f_i = h \cdot \sum_i f_i = h \cdot n$, ahol a szummázás az egyes oszlopokra (osztályközökre) fut. A megoldás tehát kézenfekvő: normalizáljunk ezzel! Ha minden oszlop magasságát leosztjuk $h \cdot n$-nel, akkor máris 1 lesz a görbe alatti terület. Így készült hisztogramot mutat a 2.10. ábra.

```
hist(birthwt$bwt, freq = FALSE, xlab = "Születési tömeg [g]",
     ylab = "Gyakoriság [fő]", main = "")
```

[23] Most már elárulhatom, hogy a 2.7. ábránál csaltam: direkt felnagyítottam a sűrűségfüggvényt, hogy ez a probléma ne jelentkezzen; tehát azon az ábrán igazából nem is valódi sűrűségfüggvény van.



2.10. ábra. Példa egy mennyiségi változó ábrázolására hisztogrammal.

Amit látunk, az persze előre is megjósolható lett volna: mivel minden oszlopmagasságot pontosan ugyanúgy $h \cdot n$-nel osztottunk le, így végeredményben az ábra változatlan néz ki (nyugodtan mondhattuk volna azt is, hogy csak a függőleges tengelyt skálázzuk át). Ezért is mondhatjuk, hogy a dolognak nincs nagy jelentősége.

Egy kivétel azonban van: ha az osztályközök nem ugyanolyan szélesek. Ekkor ugyanis nem ugyanazzal a számmal kell leosztani az oszlopok magasságát; belátható, hogy ilyenkor az oszlopok

$$\frac{f_i}{n \cdot h_i}$$

magasságúak kellenek legyenek (ellenőrizzük le, $\sum_i h_i \cdot \frac{f_i}{n \cdot h_i} = 1$ valóban). Ez alapvetően változtatja meg a helyzetet: ilyenkor a magasságok *nem* lehetnek a gyakoriságok, hiszen ez így már nem csak a függőleges tengely átskálázása, az egyes oszlopok egymáshoz való viszonya is megváltozik. Ilyenkor tehát csak a fenti, sűrűség-jellegű magasságok használatának van értelme. (Az R beépített működése pontosan ezt tükrözi: ha nem adjuk meg kézzel, hogy mit szeretnénk, akkor megnézi az osztályközöket, ha azonos szélességűek, akkor gyakoriságokat használ magasságként, ha nem, akkor a fenti sűrűséget. Ez utóbbi esetben kézzel ugyan visszaállíthatjuk gyakoriságok használatára, de ilyet ne tegyünk, ez hibás lesz – az R megengedi, de figyelmeztetést ad.) A különböző szélességű osztályközök használatának lehet értelme, elég

kézenfekvő például az ötlet, hogy ahol ritkábban vannak a pontok, ott szélesebb osztályközt vegyünk fel (mondván, hogy bár így torzítottabb lehet, de kevés ponttal úgysem tudunk sokkal jobbat csinálni), ahol meg sűrűn, ott bátran felvehetünk szűkebb intervallumokat is, de a gyakorlatban az ilyen megoldásokat ritkán használják, és a hisztogramokat általában azonos szélességű osztályközökkel készítik.

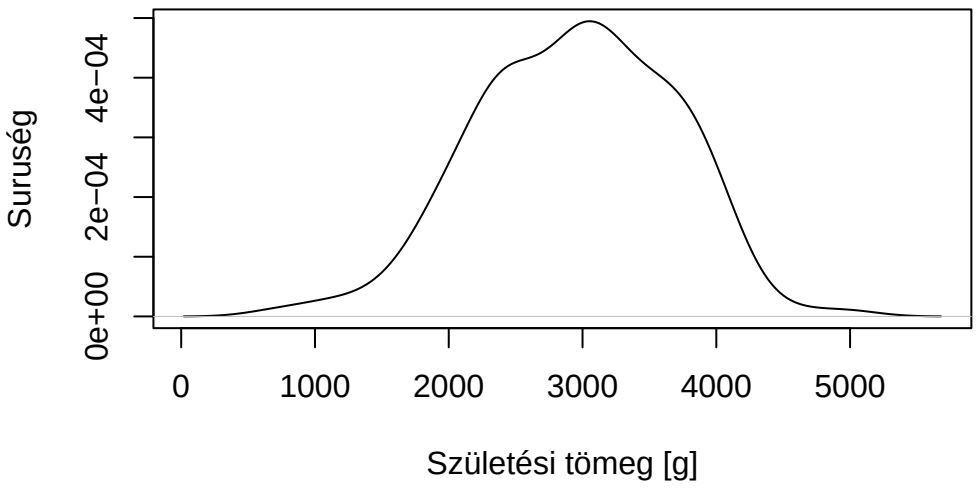
Hogyan értékelhetjük összességében a hisztogramot, mint adatvizualizációs eszközt? Kezdjük a legfontosabb előnnyel: a hisztogram könnyen értelmezhető, és jól ismert. Nagyon szemléletes, jól mutatja az egész eloszlás alakját, annak minden részletével (különösen, ha a mintanagyság nem nagyon kicsi).

Ezek mellett azonban a hisztogramnak több komoly problémája is van. Az egyik hátránya, hogy a lépcsős függvény jellegű közelítés ugyan matematikáját, konstrukcióját – pláne kézi konstrukcióját – tekintve kényelmes, de nem túl természetes: a valódi függvények, amiket közelíteni igyekszünk, nem így néznek ki, így ez a fajta közelítés zavaró lehet. A másik probléma, hogy sok helyet foglal, nem túl kompakt – ha többet kell egymás mellé rajzolnunk, akkor az hamar nehezen értelmezhető lesz. Márpedig – és itt jön a másik probléma – hisztogramokat muszáj egymás mellé plottolni: hisztogramokat nem igazán lehet – például különböző színnel megkülönböztetve – egymás*ra* plottolni, mert az ábra szinte azonnal teljesen áttekinthetetlen lesz.

2.4.2.3. Magfüggvényes sűrűségbecslő

A **magfüggvényes sűrűségbecslők** más kiindulóponttal épülnek fel, de céljuk hasonló a hisztograméhoz: a sűrűségfüggvény közelítése. A legfontosabb különbség, hogy a hisztogramoknak a megértését, konstrukcióját (pláne kézi konstrukcióját) nagyban segíti a lépcsős függvény jellegük, de a dolognak az az ára, hogy a kapott közelítés igazából elég természetellenes lesz – a valóságban aligha van bármilyen orvosbiológiai változó, ami lépcsőkben ugrál. A magfüggvényes sűrűségbecslők matematikai felépítése ugyan bonyolultabb, többé nem lehetséges kézi rajzolásuk, viszont cserében elérnek valami nagyon fontosat: a közelítés szép sima, szakadás nélkül, folytonos függvénnyel oldják meg (2.11. ábra).

```
plot(density(birthwt$bwt), xlab = "Születési tömeg [g]",
     ylab = "Sűrűség", main = "")
```



2.11. ábra. Példa egy mennyiségi változó ábrázolására magfüggvényes sűrűségbecslővel.

A további részletekkel itt nem foglalkozunk, de annyit megjegyzek, hogy ezek is igénylik egy, a hisztogramok osztályköz-szélességével analóg paraméter hangolását (sávszélesség), sőt, itt még egy paramétert, a magfüggvényeket is meg kell választani. Szerencsére ezekre elég jól bevált megoldások érhetőek el.

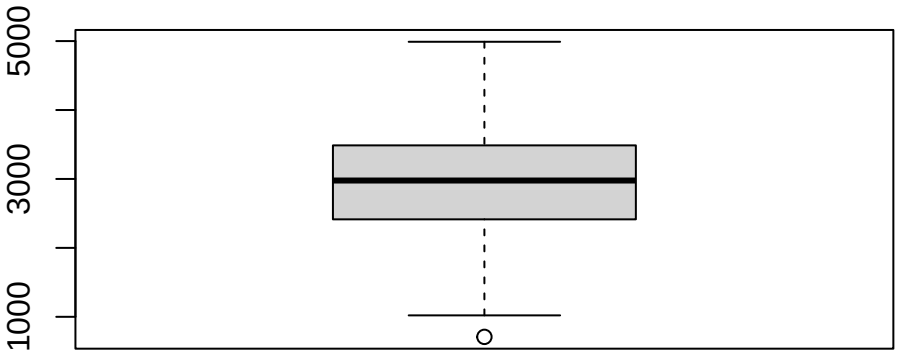
A magfüggvényes sűrűségbecslők legfontosabb előnye, hogy sokkal természetesebben néznek ki, mint a hisztogramok. Az igazság az, hogy emiatt indokolt lenne sokkal gyakrabban használni őket a hisztogramok hátrányára; ennek valószínűleg a bonyolultabb matematikájuk szab gátat. Az azonban megjegyzendő, hogy abban a hátrányban, hogy van paraméterük, amit a felhasználónak kell behangolnia – ebből fakadóan lehet, hogy hibásan állítja be – a magfüggvényes sűrűségbecslők is osztoznak a hisztogrammal; itt is igaz, hogy ettől akár nagymértékben is függhet sajnos a végeredmény. Végezetül még egy előnyét megemlítem a magfüggvényes sűrűsébecslőknek a hisztogramokkal szemben: épp a simaság miatt sokkal inkább egymásra lehet belőle többet is (például különböző színnel) plottolni.

2.4.2.4. Boxplot

Végül egy egész más elven felépülő, de szellemes, és a gyakorlatban is nagyon hasznos vizualizációs módszerrel ismerkedünk meg, a (Tukey-féle) *boxplottal* (vagy ritkán használt magyar nevén: dobozábrával).

A boxplot nem más, mint egy számegyenes fölé rajzolt téglalap, mely egy adott változót reprezentál úgy, hogy a téglalap alsó széle az alsó kvartilisnél (Q_1-nél), a felső széle pedig a felső kvartilisnél (Q_3-nál) van. A téglalapon belül egy vastagabb függőleges vonal található a mediánnál (2.12. ábra).

```
boxplot(birthwt$bwt)
```



2.12. ábra. Példa egy mennyiségi változó ábrázolására boxplottal.

A boxplotból két „antenna” nyúlik ki felfelé és lefelé. A boxplot alapváltozatában ezek a mintaminimumig és mintamaximumig nyúlnak ki, de a némileg haladóbb megvalósításban (amit a fenti ábra is mutat) az alsó antenna nem a minimumig terjed, hanem a legkisebb elemig, ami nem kisebb, mint $\mathrm{Q1} - \alpha \cdot IQR$; hasonlóképp a felső antenna nem a maximumig terjed, hanem a legnagyobb elemig, ami nem nagyobb mint $\mathrm{Q3} + \alpha \cdot IQR$. α egy előre megadott konstans, tipikusan $\alpha = 1{,}5$. Azokat az elemeket melyek ezen kívül helyezkednek el, külön szimbólum, például kis karika jelöli. E mögött az a megfontolás, hogy így a boxplot egyszerű outlier-szűrést is lehetővé tesz: azok az elemek minősülnek outliernek, melyek az antennákon kívül helyezkednek el.

A boxplot jóval nagyobb információtömörítést hajt végre mint akár a hisztogram, akár a magfüggvényes becslő – ez alapvető hátránya, bár ennek ellenére gyakorlott szem számára így is meglehetősen jó információt hordoz az eloszlás alakjáról. Azonban ugyanez előnye is, hiszen kompakt, ami különösen jól jön akkor, ha például több boxplotot kell ábrázolni – elég

sok egymás mellé rajzolható úgy, hogy összehasonlíthatóak és még áttekinthetőek maradnak. További nagy előnye, hogy – szemben mind a hisztogrammal, mind a magfüggvényes becslővel – semmilyen paraméter hangolását nem igényli, kinézete teljesen egyértelműen meghatározott az adatok által, semmilyen felhasználó által beállítandó (és így potenciálisan hibalehetőséget adó) paramétere nincsen.

2.5. Minőségi változók kétváltozós elemzése

Minőségi változók kapcsolatát **asszociációnak** szokás nevezni a statisztikában. Erre jó példa adatbázisunk rassz (`race`) és irritábilis méh (`ui`) változói, mely az alany rassz szerinti hovatartozását és az irritábilis méh szindróma fennállását adja meg.

2.5.1. Analitikus eszközök

Ahogy már megbeszéltük, a kétváltozós vizsgálatok sava-borsa az lesz, hogy a változók *kapcsolatáról* is képesek leszünk nyilatkozni. Ahhoz, hogy precízen definiáljuk, hogy mit értünk kapcsolat alatt, elsőként bemutatjuk az **kontingenciatáblát** (vagy kombinációs táblát vagy kereszttáblát), mely egyúttal az egyik legfontosabb analitikus eszköz is lesz két minőségi változó kapcsolatának vizsgálatában. Ezt követően nagyon röviden beszélünk a kapcsolat jellemzésére használható mutatószámokról is.

2.5.1.1. Kontingenciatábla

A **kontingenciatábla** egy olyan táblázat, melynek soraiban és oszlopaiban a két változó lehetséges kimenetelei vannak, az egyes cellákban pedig azon megfigyelési egységek darabszáma (gyakorisága), melyek a cella sora és oszlopa szerinti kimenetűek a sorhoz illetve az oszlophoz rendelt változó szerint. Például, a rassz és az irritábilis méh kontingenciatáblája így néz ki:

```
table(birthwt$race, birthwt$ui)
```

```
              0  1
  Kaukázusi    83 13
  Afroamerikai 23  3
  Egyéb        55 12
```

Tehát például 83 olyan megfigyelési egység van az adatbázisban, ahol az anya rassza kaukázusi *és* nincs irritábilis méh szindrómája 12 egyéb rasszú, és irritábilis méh szindrómában szenvedő alany van, és így tovább.

A kontingenciatábla szigorúan véve csak a 3×2 darab gyakoriságot jelenti; de néha összegző sorokat vagy oszlopokat írunk mellé:

```
tab <- table(birthwt$race, birthwt$ui)
rbind(cbind(tab, margin.table(tab, 1)),
      cbind(t(margin.table(tab, 2)), margin.table(tab)))
```

```
               0  1
Kaukázusi     83 13  96
Afroamerikai  23  3  26
Egyéb         55 12  67
             161 28 189
```

Ezek neve: **perem- vagy vetületi gyakoriság**. (Mindkét elnevezés logikus: perem, hiszen a kontingenciatábla peremére kell ezeket ráírni, és vetületi, hiszen úgy kaphatjuk, hogy a kontingenciatáblát levetítjük vízszintesen vagy függőlegesen „levetítjük", vetítés alatt most azt értve, hogy az egymásra „vetülő" elemeket összeadjuk.) A 189 természetesen a mintanagyság.

A fenti gyakoriságokon túl természetesen relatív gyakoriságokról is beszélhetünk. A relatív gyakoriság definícióját közvetlenül alkalmazva kapjuk azt a lehetőséget, hogy mindegyik cellát leosztjuk a mintanagysággal, például a bal felső $83/189 = 43{,}9\%$ lesz. Ez az irritábilis méh szindrómában szenvedő kaukázusiak aránya a teljes mintán belül. A relatív gyakoriságokkal kitöltött kontingenciatábla peremei a **relatív peremgyakoriságok** (vagy relatív vetületi gyakoriságok). Szokás ezt peremmegoszlásnak vagy vetületi megoszlásnak is nevezni.

Kontingenciatábla esetén azonban van egy másik – logikus – mód arra, hogy relatív gyakoriságot értelmezzünk: a 43,9% megadja, hogy az összes alany mekkora hányada kaukázusi *és* irritábilis méh szindrómában nem szenvedő, de minket érdekelhet az is, hogy az (összes helyett) csak az irritábilis méh szindrómában nem szenvedők mekkora hányada kaukázusi. Azaz: a 83-at nem a 189-cel, hanem a 161-gyel osztjuk le: $83/161 = 51{,}6\%$. Ezt nevezzük **feltételes relatív gyakoriságnak**. Azért feltételes, mert ez egy relatív gyakoriság *azon feltétel mellett*, hogy valaki nem szenved irritábilis méh szindrómában. Más szóval: ha *feltesszük*, hogy az alanyaink nem szenvednek irritábilis méh szindrómában akkor közöttük 51,6% a kaukázusiak aránya. Ez természetesen kiszámolható a rassz változó másik két kimenetére is; az így kapott 51,6% – 14,3% – 34,2% egy teljes (csak épp feltételes) relatív gyakorisági sor, összege természetesen 100%. Szokás ezt a sorváltozó (esetünkben a rassz) feltételes megoszlásának is nevezni, az oszlopváltozó (esetünkben az irritábilis méh) *adott értéke* (esetünkben: »igen«) mint feltétel mellett. Természetesen ugyanezek kiszámolhatóak a jobb oldali oszlopra is, ez magyarul azt jelenti, hogy az irritábilis méh »nem« kimenetére feltételezünk. Az eljárás ugyanez, azzal a különbséggel, hogy a jobb oldali számokat nyilván 28-cal kell leosztani. A feltételes relatív gyakoriság tehát nem más, mint a gyakoriság adott peremgyakorisággal osztva.

Természetesen nem csak az oszlopváltozóra feltételezhetünk! Pontosan ugyanígy van értelme beszélni az oszlopváltozó feltételes eloszlásáról a sorváltozó adott értéke, mint feltétel mellett.

Például kijelenthetjük, hogy annak feltételes relatív gyakorisága, hogy egy alany nem szenved irritábilis méh szindrómában 83/96 = 86,5% *azon feltétel mellett*, hogy kaukázusi a rassza. Hasonlóan továbbmenve azt is mondhatjuk, hogy az irritábilis méh fennállásának feltételes megoszlása azon feltétel mellett, hogy az alany kaukázusi, 86,5% – 13,5%.

Összefoglalva, egy cellához négyféle számot is rendelhetünk, a bal felső példáján: 83 (gyakoriság), 43,9% (relatív gyakoriság), 51,6% (feltételes relatív gyakoriság azon feltétel mellett, hogy nem áll fenn irritábilis méh szindróma) és 86,5% (feltételes relatív gyakoriság azon feltétel mellett, hogy a rassz kaukázusi). Mindezeket szemléltetik a következő táblázatok.

Relatív gyakoriságok (peremeken a vetületi megoszlásokkal):

```
tab <- prop.table(table(birthwt$race, birthwt$ui))
rbind(cbind(tab, margin.table(tab, 1)),
      cbind(t(margin.table(tab, 2)), margin.table(tab)))
```

```
                    0         1
Kaukázusi    0.4391534 0.06878307 0.5079365
Afroamerikai 0.1216931 0.01587302 0.1375661
Egyéb        0.2910053 0.06349206 0.3544974
             0.8518519 0.14814815 1.0000000
```

Irritábilis méh feltételes relatív gyakoriságai a rassz különböző kimenetei, mint feltétel esetén:

```
tab <- prop.table(table(birthwt$race, birthwt$ui), 1)
rbind(cbind(tab, margin.table(tab, 1)))
```

```
                    0         1
Kaukázusi    0.8645833 0.1354167 1
Afroamerikai 0.8846154 0.1153846 1
Egyéb        0.8208955 0.1791045 1
```

Rassz feltételes relatív gyakoriságai az irritábilis méh különböző kimenetei, mint feltétel esetén:

```
tab <- prop.table(table(birthwt$race, birthwt$ui), 2)
rbind(tab, t(margin.table(tab, 2)))
```

```
                    0         1
Kaukázusi    0.5155280 0.4642857
Afroamerikai 0.1428571 0.1071429
Egyéb        0.3416149 0.4285714
             1.0000000 1.0000000
```

Az, hogy a fentiek közül melyiket használjuk, az elemzési céltól függ. Statisztikai értelemben felcserélhető az, hogy „a kaukázusiak mekkora hányada szenved irritábilis méh szindrómában?" és az, hogy „az irritábilis méh szindrómában szenvedők mekkora hányada kaukázusi?", de tartalmilag nem: feltételezni olyan információra van értelme, amit ismerünk. (Ez a feltételes valószínűség fogalmának a lényege: ismert információ beépítése egy valószínűségbe.) Képzeljük magunkat egy orvos helyébe, aki ül a rendelőjében, vele szemben a páciens. Ha azt vesszük, hogy „az irritábilis méh szindrómában szenvedők mekkora hányada kaukázusi?" akkor azt mondjuk, hogy tudjuk, hogy az alany beteg-e, és kérdezzük, hogy ennek figyelembevételével mekkora a valószínűsége, hogy kaukázusi… Tehát hiába is egyenértékű statisztikailag, tartalmilag a „kaukázusiak mekkora hányada szenved irritábilis méh szindrómában?" kérdés – és az azt megválaszoló feltételes eloszlás – lesz a releváns: látván, hogy a betegek kaukázusi, kiderül, hogy *így* mennyi annak a valószínűsége, hogy szenved ebben a betegségben.

Továbbhaladva, tökéletesen látható, hogy miért mondtuk, hogy a többváltozós elemzés az egyváltozós elemzések kiterjesztése: a fenti kétdimenziós kontingenciatáblában *minden* információ benne van, amit a két változót külön-külön elemezve látnánk: egyszerűen levetítjük a kontingenciatáblát egy dimenzió mentén és kapott vetületi gyakoriságok nem mások lesznek, mint a megfelelő változó gyakorisági sora! Az tehát egyértelmű, hogy ez tartalmazza mindazt az információt, amit a két változó külön-külön végzett vizsgálata – csakhogy én azt állítottam, hogy többet is. Ez vezet el minket a változók kapcsolatának kérdéséhez.

Minőségi változók esetében (kontingenciatáblán) akkor mondjuk, hogy két változó kapcsolatban van egymással, ha a sorváltozó feltételes megoszlásai *ugyanazok*, az oszlopváltozó *bármely* értékére is feltételezünk. Vagy, ami ezzel egyenértékű[24]: az oszlopváltozó feltételes megoszlásai *ugyanazok*, a sorváltozó *bármely* értékére is feltételezünk.

Ez a definíció jogos: általánosságban véve is, az, hogy két változó között nincs kapcsolat, azt jelenti statisztikai nyelven, hogy az egyikre vonatkozó információból nem nyerünk információt a másikra vonatkozóan. Így már érthető ez a kontingenciatáblákra alkalmazott definíció: ha nincs kapcsolat, akkor hiába mondjak meg valaki, hogy mi – például – a sorváltozó értéke, ebből semmit nem tudunk meg az oszlopváltozó feltételes megoszlásáról (hiszen az minden sorban ugyanaz!). Ha van kapcsolat, akkor nyerünk plusz-információt (hiszen más lesz a feltételes megoszlása).

Látható, hogy ebben az esetben csak nagyon gyenge kapcsolatról beszélhetünk: a sorváltozó feltételes megoszlása mindkét oszlopban (precízen: az oszlopváltozó mindkét kimenetére feltételezve) nagyjából ugyanaz (kb. 50% – kb. 10% – kb. 40%), és az oszlopváltozó feltételes megoszlása is nagyjából ugyanaz mindhárom sorban (kb. 85% – kb. 15%). Ahogy már volt róla szó, az előbbi mondat bármelyik feléből automatikusan következik a másik fele. Itt tehát szemléletesen is látható a kapcsolat hiányának tartalma: a rasszra vonatkozó információ nem adott szinte semmilyen információt a betegség fennállásáról, abban a precíz értelemben, hogy *hiába is* mondja meg valaki, hogy az alany rassza kaukázusi, afroamerikai vagy egyéb, szinte

[24]Ez bizonyítást igényelne, de belátható, hogy az egyikből következik a másik.

ugyanúgy csak azt tudjuk mondani, hogy „akkor 85% – 15% a megoszlás az irritábilis méh fennállása szerint”.

Képzeljünk el ezzel szemben – másik végletként – egy olyan esetet, melyben a 189 alany közül 100 kaukázusi irritábilis méh szindróma nélkül, és 89 egyéb rasszú irritábilis méh szindrómával! Ebben az esetben az egyik változóra vonatkozó információ nem egyszerűen „elárul valamit” a másik változóról, hanem egyenesen determinálja azt: ha valaki elárulja, hogy egy alany kaukázusi rasszú, akkor *biztosan tudjuk*, hogy nem szenved irritábilis bél szindrómában, ha pedig azt mondja, hogy egyéb rasszú, akkor *biztosan tudjuk*, hogy szenved ebben. (Itt rögtön jól látszik, hogy a dolog fordítva is működik: ha tudjuk, hogy egy alany nem szenved irritábilis méh szindrómában, akkor azonnal tudjuk, hogy kaukázusi, ha pedig szenved ebben, akkor biztos, hogy egyéb rasszú.) Ez a kapcsolat erősségének másik végpontja.

Zárásként megjegyezzük, hogy a statisztikában valójában nem így szokták bevezetni a kapcsolat fogalmát, hanem úgy, mint azt az esetet, amikor a két változó nem független egymástól; függetlenség alatt pedig azt értik, hogy az együttes megoszlás a vetületi megoszlások szorzataként áll elő. Érdemes végiggondolni, hogy ez valóban egybeesik a hétköznapi „függetlenség” fogalommal. Szintén érdemes végiggondolni, hogy ebből valóban következik a fenti definíció, de ezzel részletesebben nem foglalkozunk most.

2.5.1.2. Mutatószámok

A kapcsolat *erősségének* kvalitatív fogalmát fent megadtuk; erre több mutatót is definiáltak, melyekkel az erősség számszerűen is lemérhető. Amennyiben a változók nominálisak, úgy pusztán erre van lehetőség.

Ha azonban a változók ordinálisak, úgy értelmet nyert a kapcsolat *irányának* fogalma is. Ordinális változók esetén ugyanis a sorok és oszlopok sorrendje nem tetszőleges, van értelme mindkét változó szerint „nagyobb” és „kisebb” kimenetről beszélni. Innentől kezdve tehát nem csak azt mondhatjuk, hogy van kapcsolat, ha más oszlopban más a feltételes megoszlás, hanem értelmet nyer az a kijelentés is, hogy nagyobb oszlopban a feltételes megoszlás úgy más, hogy inkább nagyobb sorbeli érték szerepelnek, vagy épp úgy, hogy inkább kisebbek. (Itt is egyenértékű, ha ugyanezt a sorok és oszlopok fordított szerepével mondjuk el.) Ezt ragadja meg a kapcsolat irányának fogalma: ha van kapcsolat (nem 0 az erőssége), akkor az pozitív, amennyiben az oszlopváltozó szerinti nagyobb érték tendenciájában a sorváltozó szerinti nagyobb értékkel jár együtt (és fordítva), negatív, ha az oszlopváltozó szerinti nagyobb érték tendenciájában a sorváltozó szerint kisebb értékkel jár együtt (és fordítva). Ordinális változónál erről is lehet nyilatkozni mutatókkal.

A konkrét mutatószámokkal most nem foglalkozunk (többek között azért sem, mert meglehetősen sok van belőlük, attól függően, hogy pontosan hogyan viselkednek az egyes változók).

2.5.2. Grafikus eszközök

Kontingenciatáblát vizualizálni ún. mozaikábrával és asszociációs ábrával lehet, ezek azonban nem túl látványos, és emiatt nem is túl gyakran használt módszerek, így most mi sem részletezem ezeket.

Ami bevettebb, az a vetületi megoszlások (vagy nevezetes feltételes megoszlások) ábrázolása egyszerűen oszlopdiagramon (vagy kördiagramon), ez azonban ugyanaz a feladat, amit már minőségi változók egyváltozós elemzésénél megbeszéltünk.

2.6. Mennyiségi változók kétváltozós elemzése

Mennyiségi változók kapcsolatát **korrelációnak** szokás nevezni a statisztikában. Erre jó példa adatbázisunk anyai testtömeg (`lwt`) és újszülött születési tömege (`bwt`) változói, melyek az anya illetve az újszülött testtömegét tartalmazzák.

2.6.1. Analitikus eszközök

A kapcsolat fogalmát mennyiségi változókra is ugyanazon gondolatot követve értelmezzük, mint amit minőségi változóknál már láttunk. Azt mondjuk, hogy két változó kapcsolatban van egymással, ha az egyik változó átlag feletti értékei tendenciájában a másik változó átlag feletti értékeivel járnak együtt (és ekkor persze fordítva is: az egyik változó átlag alatti értékei tendenciájában a másik változó átlag alatti értékeivel járnak együtt). Azaz: ha egy megfigyelési egység értéke az egyik változó szerint átlag feletti, akkor várhatóan a másik változó szerint is átlag feletti[25] lesz. Pontosabban szólva ez a *pozitív* kapcsolat definíciója, a negatív esetén az egyik változó átlag feletti értékei tendenciájában a másik átlag alatti értékeivel járnak együtt, és fordítva. Itt természetesen *sztochasztikus* kapcsolatról beszélünk, ezért a „tendenciájában" kifejezés: nem arról van szó, hogy ha a megfigyelési egység egyik változója átlag feletti, akkor *biztos*, hogy a másik is, de az esetek *többségében* érvényesül ez a tendencia.

Érdemes megfigyelni, hogy itt mindenképp van értelme az iránynak (összhangban azzal, hogy a mennyiségi változók bírnak az ordinális tulajdonságaival is).

Két mennyiségi változó fent definiált kapcsolatát klasszikusan a **kovarianciával** szokás lemérni, jele $\operatorname{cov}(x, y)$. Ennek definíciója:

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} \left[(x_i - \overline{x}) (y_i - \overline{y}) \right]}{n}.$$

[25] Az átlag itt természetesen minden esetben a szóban forgó változó átlagát jelenti. A használatára azért van szükség (és azért nem mondhatjuk egyszerűen azt, hogy „a változó nagy értékei"), mert hozzáadva valamilyen nagy konstanst a változóhoz, annak összes értéke nagy lesz, tehát mindenképp valamilyen viszonyításra van szükség.

A számítás logikája vegytisztán tükrözi a definíciót: az $(x_i - \overline{x})$ tükrözi az egyik, az $(y_i - \overline{y})$ a másik változó szerint azt, hogy az adott megfigyelési egység átlag alatti vagy átlag feletti. Vegyük észre, hogy a kettő szorzata pedig *pontosan akkor* lesz pozitív, ha vagy mindkét változó szerint átlag feletti a megfigyelési egység, vagy mindkét változó szerint átlag alatti – azaz ha az adott megfigyelési egység a pozitív kapcsolatot erősíti meg! Ha a szorzat negatív, akkor az adott megfigyelési egység a negatív kapcsolatot erősíti.

Ennél azonban több is igaz: a szorzatnak nem csak az előjele stimmel, de a nagysága is, az ugyanis kifejezi, hogy mennyire erősít meg bennünket az adott megfigyelési egység a kapcsolat fennállásában. Ha a megfigyelési egység egyik (pláne ha mindkét) változó szerint közel van az átlaghoz, akkor az csak gyenge „bizonyíték" a kapcsolat mellett (kis módosulással lehet, hogy az ellenkező irányú kapcsolatot erősítené), viszont ha mindkét változó szerint távol van az átlagtól, az erős érv a kapcsolat mellett.

A szummázás ezeket a hatásokat fogja összeadni megfigyelési egységről megfigyelési egységre, így előjele a kapcsolat irányát mutatja, abszolút értéke pedig annak erősségét. (Az n-nel való leosztás nyilván szükséges, különben a kétszer megismételt adatbázison kétszer akkora lenne a kovariancia, holott az információ ugyanaz; tehát ezeket a szorzatokat átlagolni kell.)

Hogy mi a kovariancia problémája, az azonnal kiderül, ha közöljük az anyai és az újszülött testtömeg közti kovarianciát: 4141.6518913. Ami kétségtelenül kiolvasható ebből, hogy az anyai és az újszülött testtömeg között van kapcsolat, mégpedig pozitív irányú (nagyobb anyai tömeg – nagyobb újszülött tömeg, és fordítva), hiszen az előjel pozitív. Amiről viszont lényegében semmit nem tudunk meg, az az erősség! Annál is inkább, mert a kovariancia mértékegység-függő: más értéket kapunk, ha az újszülött testtömegét nem grammban, hanem kilogrammban rögzítjük. Tekintetbe véve, hogy az információ ettől még ugyanaz marad, ez nyilván nem szerencsés. A probléma lényegében az, hogy honnan tudhatnánk, hogy a 4141.6518913 sok vagy kevés…? Ebben segít minket az a matematikai észrevétel, hogy mindenképp fennáll az $-s_x s_y \leq \operatorname{cov}(x, y) \leq s_x s_y$ összefüggés, tehát a kovariancia abszolút értéke nem lehet nagyobb mint a két változó szórásának szorzata. Így máris van mihez viszonyítani a kovariancia nagyságát! Ez elvezet minket a következő mutatóhoz, a neve **korreláció**, jelben $\operatorname{corr}(x, y)$:

$$\operatorname{corr}(x, y) = \frac{\operatorname{cov}(x, y)}{s_x s_y}.$$

Ez az előjel értelmezésén semmit nem változtat, hiszen a kovariancia előjelét meghagyja (a nevezőben szórások szerepelnek, így mindkettő szükségképp pozitív), viszont az abszolút értéket értelmezhetővé teszi, hiszen a korrelációra már az teljesül, hogy $-1 \leq \operatorname{corr}(x, y) \leq 1$. A korreláció tehát minél közelebb van ± 1-hez, annál erősebb a két változó közötti kapcsolat.

Például, az anyai testtömeg és az újszülött születési tömege közti korrelációs együttható értéke 0.1857333. Ez alapján nem csak azt tudjuk mondani, hogy van kapcsolat és az pozitív irányú (a 0.1857333 előjele pozitív), de most már azt is, hogy ez a kapcsolat igen gyenge (ha elhelyezzük a 0.1857333-ot a 0–1 között).

A dologban azonban van egy csavar, ami a definícióból egyáltalán nem látható, de bebizonyítható, hogy igaz: az így definiált korreláció nem általában mér bármilyen kapcsolatot a változók között, hanem csak egyféle kapcsolatot mér, azt, hogy van-e *lineáris* kapcsolat a változók között (szokás emiatt lineáris korrelációs együtthatónak is nevezni). Ha a korreláció abszolút érték 1, az épp azt jelenti, hogy $y = ax + b$ függvényszerű kapcsolat van a két változó között. Fordítva, ha a korreláció 0, az nem azt jelenti, hogy nincs kapcsolat, hanem azt, hogy nincs *lineáris* kapcsolat! Másféle kapcsolat lehet, akár még függvényszerű is, úgy, hogy közben ez a korreláció nulla. Általában is, a korreláció „erősségét" úgy kell érteni, hogy mennyire szorosan valósul meg az *egyenesre* illeszkedés.

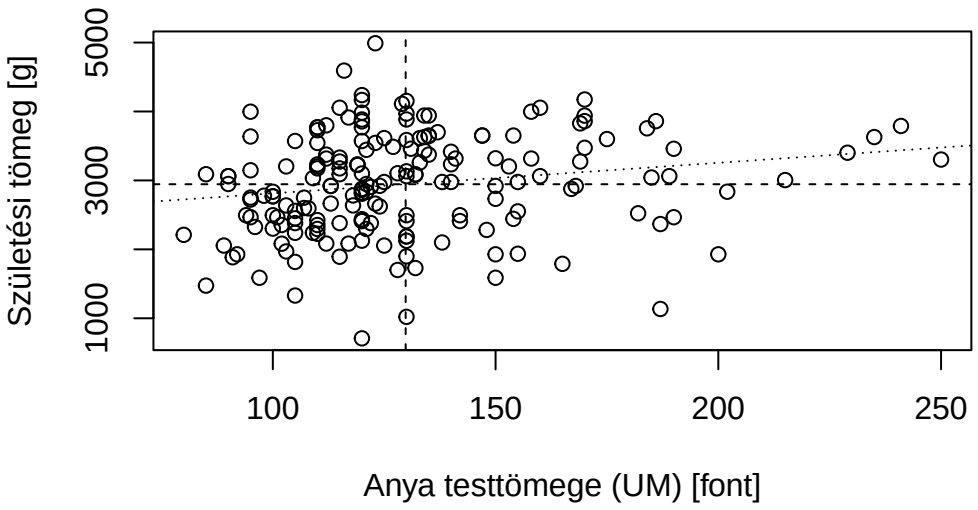
Erre tekintettel szokás más korrelációs együtthatókat is definiálni, ezek közül megemlítjük a Spearman-ρ és a Kendall-τ mutatókat, ezek ún. rangkorrelációs mutatók, amik nem konkrétan lineáris, hanem általános *monoton* kapcsolat erősségét mérik. Nem foglalkozunk vele részletesen, de megemlítem, hogy itt is igaz, hogy a kapcsolat erőssége azzal van összefüggésben, hogy az egyik változó ismerete mennyi információt árul el a másik változóról (természetesen sztochasztikus értelemben).

Végül egy figyelmeztetés. Mint általában, természetesen itt is elmondható, hogy a mutatószám használata nagyon nagy információtömörítést jelent. Éppen ezért ne támaszkodjunk önmagában egy korrelációs együtthatóra (és különösen ne önmagában egy lineáris korrelációs együtthatóra) két változó kapcsolatának megítéléséhez, hiszen ez elfedi az esetleges nemlineáris kapcsolatokat, az outlier-eket stb. Erre a következő pontban látni is fogunk egy nevezetes példát.

2.6.2. Grafikus eszközök

Két mennyiségi változó kapcsolatának legfontosabb ábrázolási eszköze az **szóródási diagram**. A szóródási diagramot úgy kapjuk, hogy minden megfigyelési egységnek egy pontot feleltetünk meg a síkban úgy, hogy a pont egyik koordinátája a megfigyelési egység egyik, a másik koordinátája a másik változó szerinti értéke. Az anyai és újszülött testtömeg szóródási diagramját a 2.13. ábra mutatja.

```
plot(bwt ~ lwt, data = birthwt,
     xlab = "Anya testtömege (UM) [font]",
     ylab = "Születési tömeg [g]" )
abline( h = mean(birthwt$bwt), v = mean(birthwt$lwt),
        lty = "dashed")
abline(lm(bwt ~ lwt, data = birthwt), lty = "dotted")
```



2.13. ábra. Két mennyiségi változó kapcsolatának ábrázolása szóródási diagrammal.

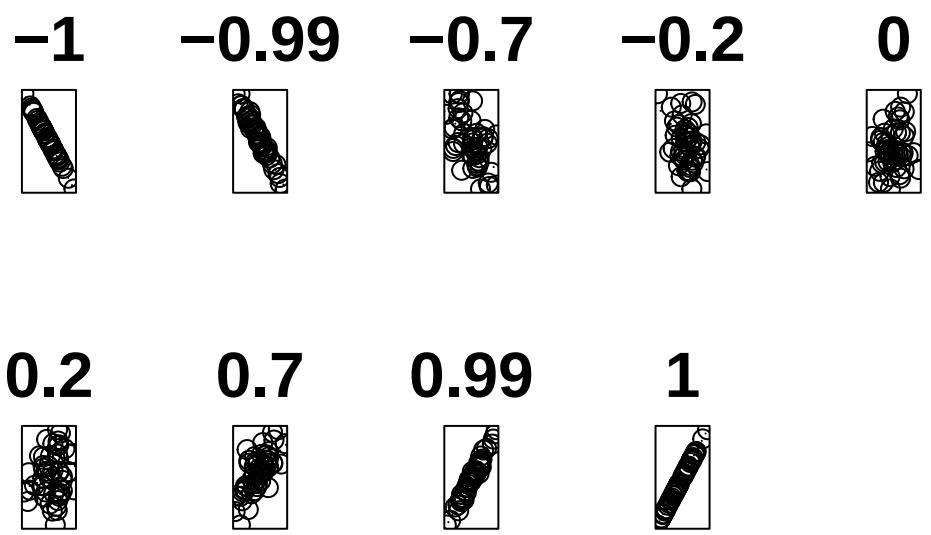
Az ábrán bejelöltem (szaggatott vonallal, a két tengellyel párhuzamosan) a két változó átlagát is.

Jól látható, immár grafikusan is, hogy mit értünk a két változó közötti (pozitív) kapcsolat fogalmán: a pontok tendenciájukban a szaggatott vonalak által kijelölt koordináta-rendszer jobb felső és bal alsó kvadránsában találhatóak (átlag feletti – átlag feletti és átlag alatti – átlag alatti zónák). Természetesen látszik az is, hogy a kapcsolat sztochasztikus, azaz van pont a több kvadránsban is (itt aztán pláne, hiszen a kapcsolat nem is túl erős). Ahogy láttuk, a kovarianciában / korrelációban persze nem csak a pontok darabszáma számít, hanem a konkrét helyzetük is.

Ráerősítve az előbb mondottakra, az ábrán behúztam a pontokra legjobban illeszkedő egyenest is. Ne feledjük, hogy a szokásos korrelációs együttható esetén, a kapcsolat „erőssége” egyúttal azt is jelenti, hogy a pontok mennyire szorosan illeszkednek a rájuk legjobban illeszkedő egyenesre (látható, hogy itt nem túl szorosan).

Mindezeket szemlélteti a 2.14. ábra is, mely különböző korrelációs együtthatójú kapcsolatokat (különböző előjelekkel és abszolút értékekkel, azaz különböző irányú és erősségű kapcsolatokat) mutat be példákkal.

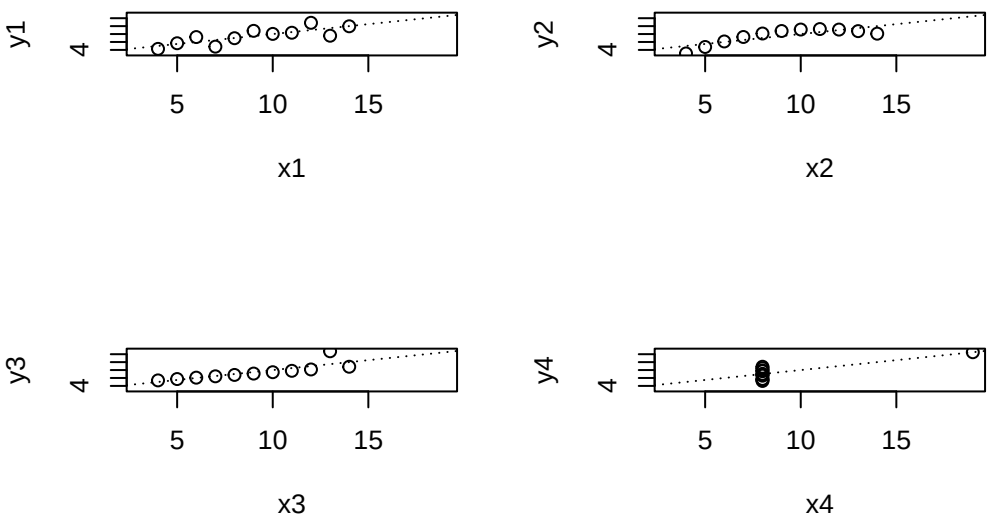
A grafikus ábrázolás előnye, hogy (szemben a korrelációs együtthatóval) nem okoz gondot semmilyen outlier, nemlineáris kapcsolat stb. – ezek mind láthatóak lesznek az ábrán. (Itt is



2.14. ábra. Különféle korrelációs együtthatók szemléltetése.

hangsúlyosan él tehát Tukey már említett tanácsa!) Erre mutat példát a nevezetes Anscombe-kvartett (2.15. ábra). Az ábrák négy kétváltozós adatsor szóródási diagramját mutatják. Mindegyiknek *hajszálpontosan ugyanaz* a korrelációs együtthatója (sőt, az átlaguk és a szórásuk is – így ugyanaz a rájuk legjobban illeszkedő egyenes is), mégis, a valós helyzet drámaian más. Outlierek, nemlineáris kapcsolatok vannak jelen; ez azonban csak ábrázolás után derül ki, a korrelációs együttható használata mindezt teljesen elfedné!

Zárásként megjegyzem, hogy ebben a grafikus ábrázolásban nincsen semmilyen információtömörítés. Az is igaz, hogy a kétváltozós elemzés tartalmaz minden információt, amit a két egyváltozós elemzés: a pontokat levetítve valamelyik tengelyre, visszakapjuk az adott tengely változójának adatait; azokat csoportosítva (a tengelyt osztályközökre bontva) rögtön készíthető például hisztogram. Szemléletesen látszik azonban az is, hogy *pusztán* a hisztogramokból (tehát az egyváltozós adatokból) *lehetetlen* lenne nyilatkozni a két változó közti kapcsolatról. (Képzeljünk egy egy olyan esetet, melyben a változók között erős kapcsolat van, de úgy, hogy mindkét változó önmagában szimmetrikus. Ekkor nyugodtan tükrözhetnénk a szóródási diagramot bármelyik átlagot jelentő szaggatott vonalra, az egyváltozós adatok ugyanazok maradnának, noha kétváltozósan pont hogy megfordult a kapcsolat iránya.) Ezért több a kétváltozós elemzés mint két egyváltozós elemzés.



2.15. ábra. Az Anscombe-kvartett.

2.7. További többváltozós elemzések

A kétváltozós esetek tárgyalásából a fentiekben kimaradt az az eset, amikor egy minőségi és egy mennyiségi változó kapcsolatát kell vizsgálni. Ezt *vegyes kapcsolatnak* szokás nevezni; részletesebben most nem foglalkozunk vele.

A másik kérdés, ami felmerül, hogy mi a helyzet kettőnél több változó esetén. Ha nem lényegesen több változóról van szó, akkor a fenti módszerek – több-kevesebb módosítással – de kiterjeszthetőek. Például a szóródási diagram elvileg három változós esetre változatlanul kiterjeszthető (bár a gyakorlatban már ezt sem nagyon szokták használni, hiszen egy három dimenziós pontfelhő csak számítógépen tekinthető meg érdemben, és ott se túl áttekinthető emberi szemnek). Négy és annál több dimenziónál már trükkre van szükség; a tipikus megoldás, hogy minden lehetséges koordináta-párra levetítik a sokdimenziós pontfelhőt, és az így kapott kétdimenziós szóródási diagramokat mutatják meg (mátrix szóródási diagram). Egy-két tucat változó felett azonban már ez sem igazán tekinthető át, illetőleg már nem nevezhető érdemben kettőnél több dimenziós elemzésnek. Hasonló a helyzet a korrelációs együtthatóval, illetve a kontingenciatáblával és elemzési eszközeivel.

3 Induktív statisztika

Ebben az alfejezetben röviden, az alapkoncepciókra fókuszálva bemutatom a statisztika induktív (következtető) ágát. Már volt róla szó, hogy az induktív statisztika jellemzője, hogy *tekintettel van* a mintavételi helyzetre, azaz arra, hogy mi csak egy részét ismerjük azon sokaságnak, melyre a kérdésünk irányult: épp azzal foglalkozik, hogy hogyan lehet pusztán a mintában lévő információ alapján mégis a sokaságról nyilatkozni. Innen a módszer neve: indukció, annyi mint következtetés, tudniillik következtetés a mintából a sokaságra.

Elsőként röviden megismételjük, és pár fontos részlettel kibővítjük a **mintavételi helyzettel** kapcsolatos ismereteinket (3.1. pont); ezt követően nagyon tömören, az alapelvekre szorítkozva bemutatom az induktív statisztika két nagy területét: a becsléselméletet és a hipotézisvizsgálatot. A **becsléselmélet** (3.2. pont) azzal foglalkozik, hogy egy sokaságot jellemző paramétert, például a sokaság várható érték pusztán a minta alapján „megtippeljünk” (valamilyen szempontok szerint a lehető legjobban). A **hipotézisvizsgálat** (3.3. pont) ennek bizonyos értelemben az ikertestvére: célja, hogy a sokaság valamely jellemzőjére tett állítás – például a sokaság várható értéke egy adott szám – helyességét „megtippeljük” pusztán a minta alapján.

3.1. A mintavételi helyzet és következményei

Ahogy már volt róla szó, mintavételi helyzetről akkor beszélünk, ha a **sokaságnak** (definíció szerint: a halmazra, amire a kutatási kérdésünk vonatkozik), csak egy részét tudjuk megfigyelni. Ezt a megfigyelt részt nevezzük **mintának**. Szintén volt róla szó, hogy a mintavételi helyzet jelentősége a biostatisztikában hatalmas: nem csak azért, mert egy sor esetben még ha a sokaság elvileg teljeskörűen megfigyelhető is lenne, erre gyakorlati okok (költség, időigény stb.) miatt nincs mód, hanem azért is, mert biostatisztikában tipikusak az olyan kérdések, melyek fiktív, végtelen sokaságra vonatkoznak (például: „Egy új vérnyomáscsökkentő gyógyszer-jelölt valóban csökkenti a vérnyomást?”). Ilyen esetekben bármennyi megfigyelést is végzünk, az szükségképp minta lesz, azaz szükségképp mintavételi helyzettel lesz dolgunk.

Adódik tehát a feladat: annak ellenére nyilatkozzunk a sokaságról, hogy mi csak egy részét ismerjük. Nagyon sokan ezen a ponton valószínűleg azt gondolnák, hogy ez lehetetlen feladat – valóban, ha mondjuk 1000 elemből csak 100-at ismerünk, akkor *elvileg* bármennyi lehet a sokaság (azaz mind az 1000 elem) átlaga, akármik is voltak a minta elemei… akkor meg mégis hogyan tudnánk ezt megmondani?! Sehogy, nyilván. Sőt, voltaképp ha 999-et ismerünk, és

csak 1-et nem, már *akkor is* ugyanez a helyzet. Tehát, ha mintavételes helyzet van, akkor nem tudjuk megmondani, hogy mi a sokaság átlaga.

Szerencsére ennél azért kicsivel jobb a helyzet. Az igaz, hogy *biztosan* nem tudjuk megmondani a sokaság átlagát, ez tény – de az nem igaz, hogy *semmit* nem tudunk róla mondani! Valamit fogunk tudni mondani, ha ugyanis megfelelően történt a mintavétel, akkor már a minta is elárult *valamit* a sokaságról. Mi az, hogy „valamit"? Ha a mintavétel teljesen véletlenszerű volt, akkor az épp azt jelenti, hogy a kivett elemek eloszlása ugyanolyan, mint a ki nem vett elemeké! Ez a tartalma annak a kifejezésnek, hogy „valamit" tudni: nem tudjuk biztosan az értékeiket, de az sem igaz, hogy nem tudunk semmit, mert tudjuk az eloszlásukat. Ez, tehát az eloszlás ismerete jelenti a középutat a között, hogy nem tudunk róla semmit, meg hogy tudjuk mennyi az értéke. Sztochasztikus ismeretekkel rendelkezünk.

Ennek van egy nagyon fontos következménye: tudunk észszerű becslést tenni. Kritikusan fontos, hogy ebben a mondatban mit értünk az alatt, hogy „észszerű", úgyhogy ezt azonnal pontosítom: azt, hogy bizonytalansággal terhelt, *de úgy*, hogy a bizonytalanság mértéke maga is jellemezhető! Itt álljunk meg egy kicsit és emésszük meg jobban ezt a mondatot, mert az egész következtető statisztikának ez a lényege. Illusztráló példaként képzeljük el, hogy a barátunk gondol egy valós számot, és a feladat az, hogy eltaláljuk mit gondolt. Három eset lehetséges:

- Megmondja, hogy mi a gondolt szám. Ez a „biztosan tudjuk" esete; statisztika alkalmazására nincs szükség.
- Nem mond semmit a világon a gondolt számról. Ez a „nem tudunk róla semmit" esete; statisztika alkalmazására megint csak nincs szükség (vagy inkább úgy mondom: nincs rá lehetőség). Éppenséggel tippelhetünk valamit – mondjuk 42 vagy $-\log \pi$ – de ezek mögé nem rakható semmilyen észszerű okfejtés, hogy miért azt tippeltük, semelyik tipp sem lesz „jobb", mint bármelyik másik, *és* semelyik tippnél nem fogunk tudni semmit mondani a tipp várható hibájáról.
- Elárulja, hogy a gondolt szám standard normális eloszlású. Ez a következtető statisztika esete! A fent emlegetett középút a „biztosan tudjuk" és a „nem tudunk róla semmit" között, amikor sztochasztikus információnk van. Vegyük észre, hogy ekkor *két* nagyon fontos dolog is lehetővé válik. *Az egyik*, hogy immár lehet észszerűen tippelni: minden matematikai levezetés nélkül is érezhető, hogy a 0 jobb tipp mint a 100 (noha elvileg kaphatunk 100 körüli értéket is standard normális eloszlásból, tehát az se lenne kizárt). Mi az, hogy „jobb"? Ahhoz, hogy ilyet mondjunk, kell a jóságnak valamilyen mérőszáma, itt lehet a definíció az, hogy ami a hibát minimalizálja – amely definíció azért fog működni, mert így már lesz információnk a hibáról! Megfelelő matematikai apparátussal lehet precízen definiálni, hogy mekkora a várható hiba ha 0-t tippelünk, és mekkora, ha 100-at. És ezzel elértünk a *másik* fontos dologhoz: ahhoz, hogy ez esetben, miután adtunk egy tippet, annak a várható hibázását magát is meg tudjuk[1] mondani! Mindkét dolog nagyon

[1] A gondolatmenet egyszerűsítése érdekében egy dolgot szőnyeg alá söpörtem: hogy a mintába be nem került elemeknek valójában eloszlását sem fogjuk *biztosan* ismerni, hiszen arra csak a – véges sok – mintába bekerült elem alapján tudunk következtetni, ami persze szintén hibával lesz terhelt. Vagyis az eloszlásban is lesz bizonytalanság (ha valaki szeretné: a sztochasztikus információt is csak sztochasztikusan fogjuk ismerni…),

fontos, úgyhogy megismétlem: immár tudunk racionálisan tippelni *és* a tippünk várható hibázást tudjuk jellemezni.

A fenti, meglehetősen filozofikus felvezetés után nézzük meg kicsit közelebbről ezt a hibázást. Vegyük a véges sokaság esetét; ezt általában is ajánlom, ha valaki most kezd ismerkedni a témával, egész egyszerűen azért, mert jól elképzelhető. Szó szerint is: képzeljünk magunk elé egy urnát, mint a lottóhúzásnál, benne golyókkal. A golyókban számok vannak, úgy, hogy azokat nem látjuk – tényleg mint a lottóhúzásnál. Ez a sokaság! Ezután megkeverjük jól az urnát, és kiveszünk néhány golyót – ez a mintavétel. A kihúzott golyókat kinyitjuk, és megnézzük mik benne a számok – ez a minta. A valós biostatisztikai helyzetek általában nem ilyenek (már volt róla szó, hogy általában fiktív, végtelen sokaságunk van), de szerintem mégis fontos és hasznos, pláne elsőként, ez a példa, mert segít, hogy ezt tényleg fizikai valójában magunk elé tudjuk képzelni.

Legyen mondjuk 100 golyónk az urnában, amiből 10-et húzunk ki – ekkor a 100 száz szám (ismeretlen) átlaga a keresett sokasági jellemző, és a 10 ismert szám a mintánk.

Bármit is számolunk ki a mintából, hogy becsüljük a 100 száz szám átlagát, az eredmény két dologtól fog függeni:

1. a 100 golyó átlagától, tehát az ismeretlen sokasági jellemzőtől, és
2. attól, hogy pont melyik mintát húztuk ki, tehát éppen melyik 10 szám került (a véletlen szeszélye folytán) a kezünkbe.

Ez utóbbi a fontos most számunkra. A jobb megértés végett képzeljük el azt, hogy nem csak egyszer húzunk: kihúzzuk a számokat, felírjuk, csakhogy utána visszadobjuk őket, átkeverjük az urnát és újra húzunk. Aztán újra, aztán újra, aztán újra, ahányszor csak szeretnék. Ez most itt egy kulcsfontosságú ötlet, ami még sokszor elő fog jönni később: a valóságban csak egyetlen húzásunk és egyetlen mintánk van, de hogy ennek a viselkedését és tulajdonságait jobban megértsük, *képzeletben* eljátsszuk, hogy mi történik, ha újra meg újra mintát vennénk! A valóságban ilyen nem lesz, hiszen csak egyetlen mintánk van, de ezek az ismeretek segíteni fognak abban, hogy jobban értsük, hogy annál az egy mintánál mire számíthatunk.

Képzeljük magunk elé ezt a helyzetet! Mit fogunk látni? Az első fontos megállapítás, hogy *hiába* is van lerögzítve a sokaság (vannak mindig ugyanazok a golyók az urnában), *hiába* is teljesen véletlen a mintavétel (mindig jól átkeverjük az urnát), *mégis* minden húzásnál más számokat kapunk. Néha kicsit kisebbeket, néha kicsit nagyobbakat. Miközben a valódi átlag mindig ugyanaz, fix, egy adott, rögzített érték (csak mi nem tudjuk, hogy mennyi). A konklúzió ebből nagyon egyszerű: lehetetlen, hogy a feladatot hiba nélkül oldjuk meg. Akármit is számolunk ki a mintából, az nem tudja mindig a jó értéket adni, lehetetlen, hogy akkor is a jó értéket szolgáltassa, ha véletlenül a 10 legkisebb számot húzzuk ki, meg akkor is, ha a 10 legnagyobbat.

---

ebből fakadóan a becslésünk hibája *maga is* egy becsült érték lesz. De igaz lesz erre is minden pozitív jellemző, amit a becslésekről mondtunk.

Bármilyen módszert is találunk ki arra, hogy a mintából hogyan következtessünk a sokaságra, teljesen biztos, hogy annak a végeredménye *mintáról-mintára változni* fog, azaz függeni fog attól, hogy konkrétan „hogyan nyúltunk bele a sokaságba", konkrétan milyen mintát vettünk. Ezt a jelenséget hívjuk **mintavételi ingadozásnak**. A szerencse épp az lesz, amit fent megtárgyaltunk, hogy ez a mintavételi ingadozás követni fog bizonyos (valószínűségi) törvényszerűségeket, így bár a fenti miatt elkerülhetetlenül hibázhatunk a következtetésnél, de ennek a hibázásnak a természetéről fogunk tudni nyilatkozni.

Amit nagyon fontos megérteni, hogy az előbb említett „hibázás" alatt nem arra kell gondolni, hogy valamilyen értelemben rosszul vesszük a mintát: a *legtökéletesebben véletlen* mintavétel mellett is előfordulhat, hogy a 10 legkisebb vagy 10 legnagyobb értéket vesszük. A *legtökéletesebb véletlen* mintavétel mellett is néha kisebb számokat húzunk, néha nagyobbakat. Ez az, amit az előbb úgy hívtam, hogy a „véletlen szeszélye"; egy elkerülhetetlen jelenség.

A mintavételi ingadozásból fakadó hibát hívjuk **mintavételi hibának**. A fenti okfejtés azt mondja, hogy a mintavételi hiba elkerülhetetlen ugyan, de a mértékéről fogunk tudni nyilatkozni.

Egy orvosi vizsgálatnak azonban nem csak olyan hibája lehet, ami az – elkerülhetetlen – mintavételi ingadozásból adódik. Minden más hibaforrást összefoglaló néven **nem-mintavételi hibának** nevezünk. Talán a legfontosabb: megfigyeléses vizsgálatoknál a confounding. De ide tartozik az is, ha nem véletlenszerűen választjuk a mintát, sok egyéb hiba mellett.

3.2. Becsléselmélet

A becsléselmélet az induktív statisztika egyik fő ága, feladata valamilyen sokasági jellemző értékének minta alapján történő számszerű meghatározása. A „becslés" szó használata azért indokolt, mert az előbb kifejtettekből adódik, hogy mintavételi helyzetben valójában nem tudjuk „meghatározni" a jellemző értékét (olyan értelemben, hogy „megmondjuk biztosan"), csak valószínűségi jellegű, potenciálisan hibával terhelt kijelentések tételére van mód, azzal, hogy a hibázás mértékét magát is fogjuk tudni jellemezni. Ettől válik a becsléselmélet tudománnyá.

A becsülni kívánt sokasági jellemzőt általánosságban θ-val jelöljük, ha nem kívánjuk konkrétan megmondani, hogy micsoda, mert általában beszélünk becsléselméletről. Ez lehet a sokaság átlaga, szórása, valamilyen tulajdonsággal rendelkező elemeinek az aránya stb. Két pontosítást, illetve kommentárt a fenti definíció általánosságáról tennem kell:

1. A becslési feladat egyik megközelítése az, amikor a sokaság eloszlására feltételezünk valamit, és ezáltal a feladat ezen eloszlás – egy vagy több – paramétérének megbecslésére redukálódik. Például feltételezzük, hogy a sokaság eloszlása normális, és így a feladat két szám (a várható érték és a szórás) megbecslése, hiszen ha azokat ismerjük, akkor már tudunk mindent az eloszlásról. Ezt hívjuk **paraméteres becslésnek**. Elképzelhető olyan feladat is azonban, ahol nem ez a helyzet, nem arról van szó, hogy a sokaság

eloszláscsaládját feltételezzük, és a becslés paraméterre irányul; kérdezhetjük például azt, hogy mi maga az eloszlás, mondjuk „becsüljük meg a sűrűségfüggvényét" formában. Ilyenkor beszélünk **nemparaméteres becslésről**.

2. A θ jelölés arra utal, hogy egyetlen számot kell becsülnünk, ez az egydimenziós paraméterbecslés. Ez a feladat például ha feltételezzük, hogy a sokasági eloszlás Bernoulli és a cél a p paraméter becslése, de az is idetartozik, ha normális háttéreloszlást feltételezünk, de úgy, hogy a szórás adott, ismert érték. Az is egydimenziós becslés, ha külön-külön megbecsüljük a várható értéket és a szórást a normális eloszlásos példa esetén, azonban annak is van értelme, hogy ezeket egyszerre, egyidejűleg becsüljük. Ez ugyanaz a feladat, csak a becsülendő objektum nem egy szám (θ), hanem egy vektor ($\underline{\theta}$). Ezt hívjük többdimenzós becslésnek.

A továbbiakban elsőként az egydimenziós paraméteres becslésekkel fogunk foglalkozni.

Maradjunk most annál a példánál, ha a feladat a sokaság átlagának/várható értékének[2] megbecslése. Ekkor egy teljesen természetes gondolat, hogy ezt a jellemzőt a minta átlagával igyekezzünk megbecsülni.

Ez a naiv „tipp" is mutatja már, hogy mit értünk precízen becslés alatt: egy olyan függvényt (neve **becslőfüggvényt** vagy egyszerűen **becslő**), melynek bemenetül a minta elemeit kell megadni, eredményként pedig kidobja a becslést az ismeretlen sokasági jellemzőre. Egy θ sokasági jellemző becslőfüggvénye tehát egy

$$\hat{\theta}\left(x_1, x_2, \dots, x_n\right)$$

függvény. (A becsült értéket a statisztikában általában is kalappal jelöljük.) A fenti jelölés a korrekt, hiszen a becslőfüggvény a mintaelemek függvénye, de néha ezt $\hat{\theta}_n$ formában rövidítjük, ha pusztán annyit kívánunk jelezni, hogy hány elemű mintából becslünk, vagy $\hat{\theta}$ formában, ha még ezt sem.

Az előbbi példánk azt jelenti, hogy ha a becsülni kívánt jellemző a sokasági várható érték ($\theta = \mu$), akkor reményeink szerint arra jó becslő lesz az

$$\hat{\theta}\left(x_1, x_2, \dots, x_n\right) = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

függvény. Ahogy már korábban is volt róla szó, ennek értéke két dologtól fog függeni: a μ értékétől (a valódi sokasági jellemzőtől), és attól, hogy konkrétan milyen mintát vettünk –

[2] Átlagról általában akkor beszélünk, amikor a sokaság véges, ilyenkor tipikusan úgy képzeljük („elemeivel adott sokaság"), hogy a sokaságot véges sok érték felsorolásával megadhatjuk; várható értéket általában akkor mondjuk, ha a sokaság fiktív, végtelen, ilyen tipikusan úgy gondoljuk („eloszlásával adott sokaság"), hogy azt a háttéreloszlást ismerjük, melyet a sokaság minden egyes eleme követ, legegyszerűbb esetben független és azonos eloszlású módon, és ezt az eloszlást adjuk meg. Az első esetekben tehát felsorolunk elemeket, a másodikban megadunk egy valószínűségszámítási eloszlást.

a véletlen szeszélyétől; ezt hívtuk mintavételi ingadozásnak. Eddig úgy fogalmaztunk, hogy a mintaelemek ingadoznak, de mivel a becslőfüggvényt a mintaelemekből számoljuk ki, így természetesen a becslőfüggvénnyel becsült érték is mintáról-mintára ingadozni fog. Gondoljunk a képzeletbeli újra-mintavételezésre: ha kihúzzuk a 10 számot, felírjuk *az átlagukat*, visszadobjuk, megkeverjük, újra húzunk, újra átlagolunk, újra visszadobunk, újra keverünk, újra húzunk, újra átlagolunk stb., akkor minden egyes alkalommal más és más átlagot fogunk kapni, *hiába* ugyanaz a sokaság és *hiába* volt teljesen véletlenszerű minden mintavétel. (Ez természetesen nem csak az átlagra igaz: bármi mást számolnánk ki a mintaelemekből, az is ugyanúgy fog viselkedni.) Ez a mintavételi ingadozás, immár a becslőfüggvényre vonatkozóan.

Ha mármost a fenti eljárást sokszor megcsináljuk, és a végén megnézzük a felírt – ingadozó – mintaátlagokat, akkor azoknak lesz egy eloszlása[3]. Ezt hívjuk **mintavételi eloszlásnak**.

Érdemes ezt egy szimulációval is megnézni! Vegyünk egy adott, rögzített a sokaságból, melyet itt eloszlásával adtunk meg ($\mathcal{N}(30, 70)$), teljesen véletlenszerű módon 10 darab 30 elemű mintát, majd mindegyiknek számoljuk ki az átlagát:

```
replicate(10, mean(rnorm(30, 70, 10)))
```

```
 [1] 70.82458 71.32775 71.10278 71.13333 66.69972 72.37003 70.66829 67.62002
 [9] 70.24407 71.36787
```

Látszik, hogy – noha a sokaság állandó, és így a várható értéke, tehát a becsülendő jellemző is állandó, fixen 10 – a minták átlaga, tehát a sokaság várható értékének mintából *becsült* értéke ingadozik.

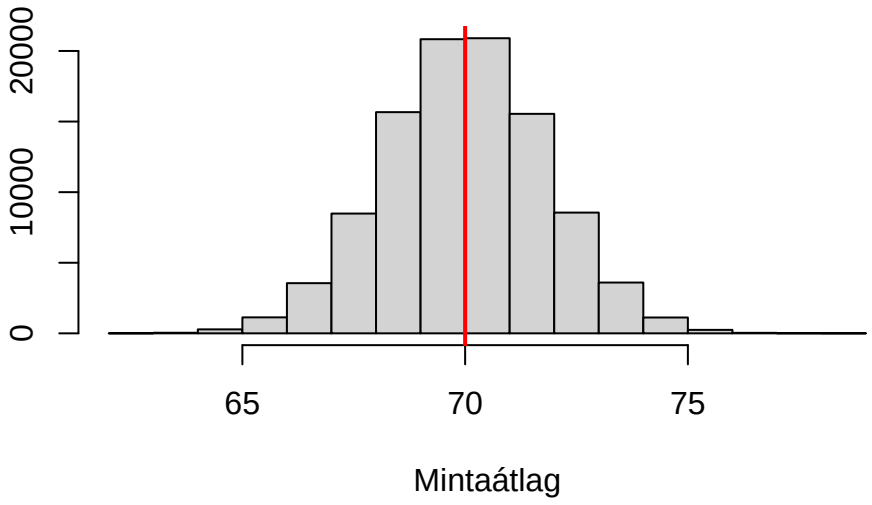
Elég sok ilyen szimulációt végezve, ez az ingadozás jól feltérképezhető, ez lesz az emlegetett mintavételi eloszlás (3.1. ábra).

```
res <- replicate(100000, mean(rnorm(30, 70, 10)))
mean(res)
```

```
[1] 69.99897
```

```
hist(res, main = "", ylab = "", xlab = "Mintaátlag")
abline(v = 70, col = "red", lwd = 2)
```

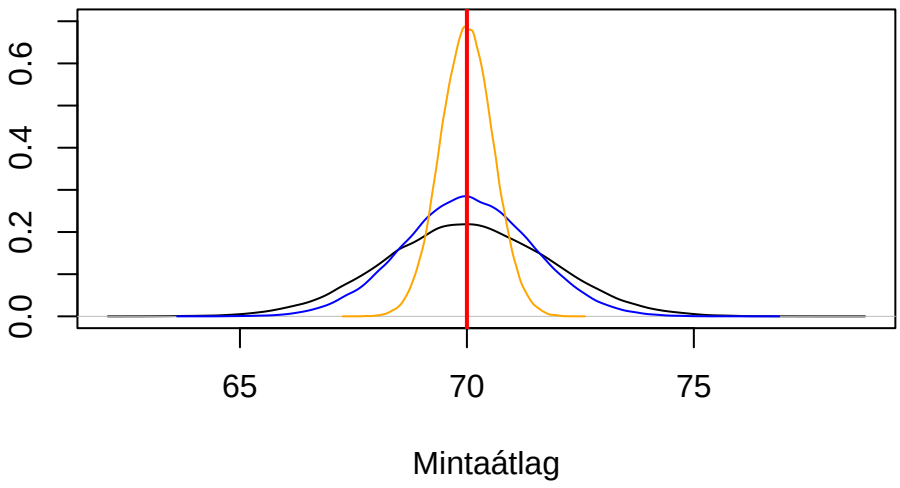
[3] A valóságban ez nem igazán definíció volt, hanem inkább egy illusztráció; erre illene egy valószínűségszámítási definíciót is adni. Ezt később meg fogom tenni.



3.1. ábra. A mintaátlag mintavételi eloszlásának meghatározása szimulációval, normális háttéreloszlás mellett.

Ilyen módon további fontos kérdések is vizsgálhatóak, például megnézhetjük, hogy a becsült érték ingadozása hogyan függ a mintanagyságtól (3.2. ábra).

```
res <- replicate(100000, mean(rnorm(30, 70, 10)))
plot(density(res), ylim = c(0, 0.7), main = "", ylab = "",
     xlab = "Mintaátlag")
res50 <- replicate(100000, mean(rnorm(50, 70, 10)))
lines(density(res50), col = "blue")
res300 <- replicate(100000, mean(rnorm(300, 70, 10)))
lines(density(res300), col = "orange")
abline(v = 70, col = "red", lwd = 2)
```



3.2. ábra. A mintavételi eloszlás függése a mintanagyságtól.

Az ilyen vizsgálatok (szokták ezt Monte Carlo szimulációnak is nevezni) könnyen kivitelezhetőek, és megfelelő számítási kapacitás mellett bonyolult problémák kezelésére is alkalmas. Hátránya viszont, hogy nem kapunk analitikus eredményt (tehát a mintavételi eloszlást nem kapjuk meg matematikai képlettel felírt függvényként); ebben az egyszerű példában ez sem jelent problémát, nemsokára vissza is fogunk rá térni.

Az előbb lényegében hasraütésszerűen választottuk a mintaátlagot mint a sokasági várható érték becslőfüggvényét. (Maximum annyit mondtam, hogy „kézenfekvő”.) De miért pont a mintaátlagot használjuk becslőfüggvényként? Elvégre nagyon sok más választással is élhetnénk…!

A helyzet nagyon egyszerű: a legjobb becslőfüggvényt kell választani. Igen ám, de mit értünk az alatt, hogy egy becslőfüggvény „jó”…? A gyakorlatban a következő tulajdonságok különösen fontosak:

1. Elfogadjuk, hogy a becslőfüggvény által szolgáltatott becslés mintáról-mintára ingadozik (mást nem is tehetünk…), de legalább az teljesüljön, hogy a becslés *a jó érték körül* ingadozik. Ez volt a konyhanyelvi megfogalmazás, a precíz az, hogy *átlagosan* jó legyen a becsült érték: egy becslőfüggvényt **torzítatlannak** mondunk, ha a mintavételi eloszlásának a várható értéke a valódi (sokasági) jellemző. Formálisan: a $\hat{\theta}$ becslőfüggvény torzítatlan, ha $\mathbb{E}\hat{\theta} = \theta$ minden θ-ra. (A „minden θ-ra” kitétel fontos: a $\hat{\theta} = 123$ várható értéke prímán a valódi érték, ha véletlenül $\theta = 123$, de azonnal elromlik, ha θ bármi más. Látható tehát, hogy a definíciónak csak akkor van értelme, ha megköveteljük, hogy működjön *bármi*

legyen is a θ értéke.) Egy becslőfüggvény torzításának, jele $Bs\left(\hat{\theta}\right)$ nevezzük a mintavételi eloszlásának várható értéke és a valódi érték közötti különbséget: $Bs\left(\hat{\theta}\right) = \mathbb{E}\hat{\theta} - \theta$. (A torzítatlan becslő tehát az, aminek a torzítása nulla.) A fenti szimulációk azt sugallják, hogy az előbbi példában a mintaátlag torzítatlan becslője a sokasági várható értéknek (ezt persze még bizonyítani kellene matematikailag is, hiszen egy ilyen szimuláció soha nem tud szó szerint bizonyítani).

2. Egy becslő **aszimptotikusan torzítatlan**, ha a torzítása 0-ba tart, midőn a mintanagyság tart a végtelenbe: $\lim_{n\to\infty} Bs\left(\hat{\theta}_n\right) = 0$. (Itt kitettem egy n alsó indexet a becslőhöz, hogy jelezzem: az értéke függ a mintanagyságtól.) Következésképp minden torzítatlan becslő egyúttal aszimptotikusan is torzítatlan, de lehet olyan becslő – és van is – ami véges mintán torzított ugyan, de a torzítása nullába tart; ez lesz aszimptotikusan torzítatlan.
3. A torzítatlan becslők a jó érték körül ingadoznak (most konyhanyelvileg szólva), ilyenkor a következő logikus elvárásunk, hogy ez az ingadozás minél kisebb legyen. E tulajdonság neve: **hatásosság**. A hatásosságot a mintavételi eloszlás szórásával (röviden mintavételi szórással) mérhetjük: két torzítatlan becslőfüggvény közül azt mondjuk hatásosabbnak, amelyiknek kisebb a mintavételi szórása. Ha egyszerűen hatásosnak nevezünk egy becslőfüggvényt, az alatt pedig azt értjük, hogy torzítatlan, és a torzítatlan becslők körében minimális szórású[4]. A mintavételi szórásra mintából adott becslést standard hibának szokás nevezni.
4. Végezetül, kissé leegyszerűsítve fogalmazva, egy becslőfüggvényt **konzisztensnek** mondunk, ha a mintanagyság növekedtével nullába tart a mintavételi szórása (a mintavételi eloszlás összemegy egy tüskévé) *és* az a tüske jó helyen van. Tehát: a becslőfüggvény (legalább aszimptotikusan) torzítatlan és nullába tart a mintavételi szórása; ami lényegében azt jelenti, hogy a becslőfüggvény tart[5] a valódi értékhez. Ne keverjük ezt össze a torzítatlansággal, vagy aszimptotikus torzítatlansággal: az csak annyit követel meg, hogy a *várható értéke* tartson a valódi értékhez; a konzisztencia azt mondja, hogy az eloszlás *egésze* tartson hozzá (tehát kiköti a nullába tartó szórást is). Érdemes végiggondolni: milyen lehet az a becslő, ami torzítatlan, de mégsem konzisztens…?

A torzítatlanság és a hatásosság *véges mintás* tulajdonságok: csak akkor állnak fenn, ha *minden* n-re fennállnak. (Hiszen a definíció nem mondott semmit n-ről, tehát csak akkor működik, ha minden n-re igaz.) Az aszimptotikus torzítatlanság és a konzisztencia *aszimptotikus* tulajdonságok, hiszen $\lim_{n\to\infty}$-re követelnek meg valamit. A statisztikai zsargon az előbbit

[4] Gondoljuk végig, hogy miért kötelező kikötni, hogy csak torzítatlan becslők körében gondolkozunk! Miért válna értelmetlenné a mintavételi szórás minimalizálása, ha nem kötnénk ki, hogy a becslőfüggvény torzítatlan legyen…?

[5] Itt egy elég nagyon egyszerűsítettem, de a konzisztencia valódi definíciója ez: hogy a becslőfüggvény, mint valószínűségi változó, konvergál a valódi értékhez. Azért egyszerűsítettem, hogy ne kelljen behozni a valószínűségi változók konvergenciájának kérdéskörét; de ha valaki ebben jártas, akkor elég annyit tudnia, hogy egy becslőfüggvényt adott értelemben – erősen, gyengén, négyzetes középben – konzisztens, ha abban az értelemben tart a valódi értékhez, midőn $n \to \infty$. Az én definícióm igazából a négyzetes középben – és így gyengén – konzisztens becslő definíciója, azt persze külön tételként kellene bizonyítani, hogy ha egy becslő aszimptotikusan torzítatlan és nullába tart a mintavételi szórása, akkor valóban négyzetes középben konzisztens.

néha kismintás, az utóbbit néha nagymintás tulajdonságnak nevezi. (Érthető, hogy honnan jönnek az elnevezések, de azért különösen az előbbi nem tökéletesen szerencsés: a kismintás természetesen nem azt jelenti, hogy kis mintára működik, hanem azt, hogy kis mintára *is* működik.)

Azzal a kérdéssel, hogy hogyan lehet egy becslőfüggvényt „kitalálni" (tehát, ha megadnak egy paramétert, akkor mutatni egy rá vonatkozó, és persze lehetőleg minél jobb statisztikai tulajdonságokkal bíró becslőfüggvényt) nem foglalkozunk részletesebben, csak megemlítem, hogy erre vonatkozóan jól bejáratott módszerek, ún. becslési elvek léteznek. (A legnevezetesebb közülük a maximum likelihood-elv, továbbá a plug-in becslés, a legkisebb négyzetek elve, a momentumok módszere és a Bayes-becslés.)

Nézzünk minderre egy példát! Tekintsünk egy (eloszlásával adott) sokaságot, mely $X \sim \mathcal{N}(\mu, \sigma_0^2)$ eloszlást követ. (Tehát tetszőleges számú mintát vehetünk belőle; minden egyes ilyen mintaelem egy ilyen eloszlásból származó, egymástól független szám lesz.) A 0 index a σ_0^2-ben arra utal, hogy ez nem becsülendő jellemző, az értékét ismertnek vesszük. Azt állítom (és ezt hamarosan szabatosabban is be fogjuk bizonyítani), hogy ekkor a belőle vett n elemű minták átlaga, azaz a μ sokasági várható érték (mint sokasági jellemző) fenti becslőfüggvénye $\overline{x} \sim \mathcal{N}(\mu, \sigma_0^2/n)$ eloszlást fog követni. (Tehát most feltételeztük, hogy azt *a priori* tudjuk, hogy normális eloszlású a sokaság, sőt, mivel σ_0-t is ismertnek vettük, azaz csak a μ a kérdés.) Jegyezzük meg, hogy a sokasági jellemző, amit becsülni szeretnénk, itt a μ maga; az tehát nem követ semmilyen eloszlást, egy – konstans – szám! Csak mi nem ismerjük. A következőkben ezt az állítást fogjuk matematikai úton, valószínűségszámítási eszközökkel bebizonyítani, mégpedig a legegyszerűbb esetre, a fent vázolt független és azonos eloszlású mintavételre.

Legyen az n elemű mintánk $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma_0^2)$ függetlenül (mivel a mintavétel azonos eloszlású is, így mindegyik ugyanolyan eloszlást követ, ezért volt azt elég egyszer leírni). Figyeljük meg, hogy itt nagy betűket írtam: ezek nem konkrét (realizálódott) értékek, hanem maguk is valószínűségi változók. (Most ugyanis statisztikai analízisét adjuk a helyzetnek: úgy képzeljük, hogy még nem vettünk mintát, hanem épp ellenkezőleg, azt vizsgáljuk, hogy „mi minden történhet" amikor majd mintát veszünk.) Ezzel a becslőfüggvényünk:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Valószínűségszámításból tudjuk, hogy

1. Normális eloszlású valószínűségi változók összege normális (szépen megfogalmazva: a normális eloszláscsalád zárt a konvolúcióra).
2. A várható érték lineáris, így egy összeg várható értéke a várható értékek összege.
3. Ha ráadásul korrelálatlan (de csak ez esetben!), akkor a szórásnégyzetek – nem a szórások! – is összeadódnak.

Ebből a háromból már következik, hogy

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma_0^2\right).$$

Szintén valószínűségszámításból tudjuk, hogy $\mathbb{E}\left(aX\right) = a \cdot \mathbb{E}X$ és $\mathbb{D}^2\left(aX\right) = a^2 \cdot \mathbb{D}^2 X$, ezekből pedig már következik, hogy

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \mathcal{N}\left(\mu, \sigma_0^2/n\right),$$

ahogy azt eredetileg állítottam is.

Ezzel igazoltuk, hogy ilyen körülmények mellett a mintaátlag torzítatlan becslője a sokasági átlagnak, sőt, kiszámoltuk a mintavételi szórását is. (Be lehetne látni kicsit komolyabb matematikai statisztikai eszközökkel, hogy ez ráadásul e körülmények között hatásos becslő is, tehát ennél kisebb mintavételi szórás el sem érhető a torzítatlan becslő segítségével.)

Ez tehát azt jelenti, hogy a 2944.5873016 gramm nem csak a születési tömegek átlaga (amint a deskriptív statisztikai résznél szerepel), hanem egyúttal a „vizsgálat beválogatási feltételeinek megfelelő újszülöttek" (fiktív, végtelen!) sokaságának várható értékének becslője is! Nem csak azt mondhatjuk, hogy 2944.5873016 gramm a mintaátlag (biztosan), hanem azt is, hogy ez a legjobb tippünk arra, hogy mennyi a sokaság várható értéke. Vegyük észre, hogy minket valójában ez utóbbi érdekel! Tehát bár a számérték itt pont ugyanaz lett (ez nincs mindig így!), az igazán érdekes eredmény az utóbbi megfogalmazás (hiszen minket nem *konkrétan* ez a 189 újszülött érdekel, hanem *általában* az ilyen újszülöttek jellemzőinek viselkedése).

Mind ez idáig azonban csak olyan becslőfüggvényekről beszéltünk, melyek egyetlen értéket, „a" legjobb becslést adják vissza eredményként. Az ilyen becslést hívjuk **pontbecslésnek**. (Hiszen az eredménye egyetlen pont a számegyenesen.) Ez azonban még nem teljesíti a kitűzött céljainkat: eredetileg azt mondtam, sőt, ez volt a pláne, hogy a becslés bizonytalanságát magát is tudjuk jellemezni – csakhogy a pontbecslést ezt a célt nem valósítja meg!

A megoldás nyomai azért már az eddigiekben is felbukkantak: a fenti levezetésből kiderült, hogy az átlag ingadozásának a mértéke σ_0^2/n. Ez már feltétlenül információ, csak az is fontos lenne, hogy ezt jól megragadható formában adjuk meg.

Azt a becslési módszert, ami ezt a célt igyekszik megvalósítani, megragadva és megjelenítve a becslésben lévő bizonytalanságot is, **intervallumbecslésnek** nevezzük. Az intervallumbecslés központi eszköze az **konfidenciaintervallum** (CI): ez egy olyan intervallum, melyre igaz, hogy a hogy ha sokszor megismételnék a mintavételt, és mindegyik mintából megszerkesztenénk a CI-t, akkor ezen CI-k várhatóan adott, nagy hányada (például 95%-a) tartalmazná az igazi (sokasági) értéket. Ez esetben ezt az intervallumot 95% megbízhatóság melletti konfidenciaintervallumnak nevezzük. A 95%, mint paraméter neve **megbízhatósági szint**, általában $1-\alpha$-nak nevezzük (tehát $\alpha = 0{,}05$ mellett beszélünk 95%-os megbízhatóságról). Első ránézésre kicsit furcsa lehet

ez a jelölés, de majd a hipotézisvizsgálatnál is látni fogjuk, hogy α-val valamilyen hibázás jellegű mennyiséget szeretnénk jelölni, nem jóságot.

Az induktív statisztikában tehát elfogadjuk (kénytelenek vagyunk elfogadni), hogy a becslésünk eredménye mintáról mintára változik, és így nem tudhatjuk biztosan, hogy *adott mintából* számolt becslés hogyan viszonyul a valódi (sokasági) értékhez – a konfidenciaintervallum azonban épp azt próbálja megragadni, hogy – adott minta alapján! – mire tippelhetünk, „vélhetően" hol lehet a valódi sokasági érték (adott, nagy megbízhatósággal). Ez természetesen már nem egyetlen szám, hanem egy tól-ig intervallum lesz a jellemzőre vonatkozóan. Hogy mit jelent a „vélhetően" és a „megbízhatóság", az pontosításra szorul, erre tárgyalásunk legvégén fogok visszatérni.

Adott megbízhatósági szint mellett minél szűkebb a CI, annál kisebb a bizonytalanság a becslésünkben. Természetesen adott becslés mellett a CI szélességét a megbízhatósági szint fogja meghatározni: kis megbízhatóság mellett szűk intervallumot is mondhatunk, de ha nagy megbízhatóságra van szükségünk, akkor csak széles limiteket tudunk szabni. Itt tehát kompromisszumot kell kötnünk: az se jó, ha nagy biztonsággal tudjuk, hogy nem igazán tudjuk, hogy hol van az igazi érték, és az se, ha nagyon kis biztonsággal tudjuk, hogy igen pontosan hol van… A 95% egy tipikus, gyakorlatban igen sokszor használt kompromisszum ez ügyben.

Nézzünk erre is egy számszerű példát! Folytatva előző példánkat, tudjuk, hogy $\overline{X} \sim \mathcal{N}\left(\mu, \sigma_0^2/n\right)$. Ebből következik, hogy

$$\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}\left(0,1\right),$$

azaz

$$\mathbb{P}\left(-z < \frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} < z\right) = \Phi\left(z\right) - \Phi\left(-z\right) = \Phi\left(z\right) - \left[1-\Phi\left(z\right)\right] = 2\Phi\left(z\right) - 1.$$

Ha ezt a valószínűséget $(1-\alpha)$-nak választjuk (a megbízhatósági szint fenti értelme miatt), akkor kapjuk, hogy $\Phi\left(z\right) = 1-\frac{\alpha}{2}$ azaz $z = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$. Erre a mennyiségre bevezetve a $z_{1-\frac{\alpha}{2}}$ jelölést, rögtön látható, hogy a $\left[\mu - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}, \mu + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right]$ tartományba $1-\alpha$ valószínűséggel esik $\overline{X}$. Ezt nevezhetnénk „deduktív statisztikának", hiszen itt a sokaságot tekintettük ismertnek, és ez alapján következtettünk a minta viselkedésére.

Átrendezve „kapjuk" a minket érdeklő az induktív statisztikát:

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Rightarrow \mathbb{P}\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right) = 1-\alpha.$$

Ekkor a konfidenciaintervallum immár egy konkrét mintára a fenti alapján:

$$\left[\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right].$$

Tipikusan $\alpha = 0{,}05$, amint mondtuk, ekkor $1-\alpha = 95\%$-os konfidenciaintervallumról beszélünk.

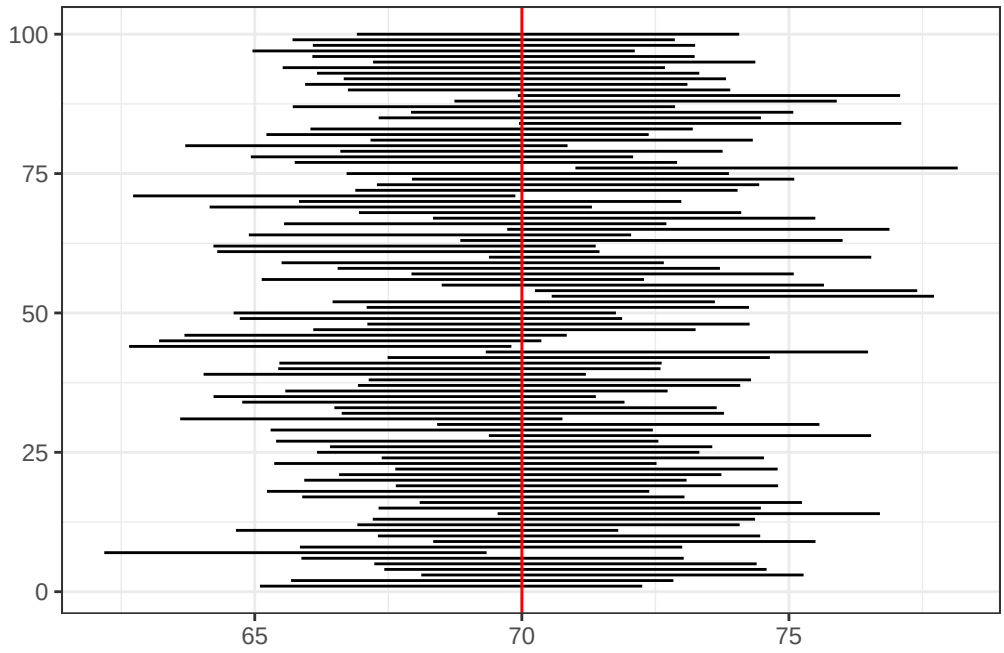
Nagyon fontos megfigyelni, hogy csak mintavétel *előtt* vannak valószínűségi változók („nagy betűk"), *utána* már nem („kis betűk") – ezért használtuk a megbízhatóság szót a valószínűség helyett. Mintavétel *után* ugyanis már nem tehetünk olyan kijelentést, hogy a megkonstruált CI 95%-os „valószínűséggel" tartalmazza a valódi, sokasági paramétert, hiszen ha már egy realizálódott minta van a kezünkben, akkor elvileg akárhol lehet a valódi érték, erről semmi közelebbit nem tudunk mondani. Valószínűséget csak a (szükségképp képzeletbeli) „ismételt mintavételi" értelemben tudunk behozni a feladatba, ezért használjuk megkülönböztetésül a megbízhatóság szót. Így kell érteni, hogy a konfidenciaintervallum jellemzi, hogy „hol lehet" a valódi (sokasági) paraméter.

A születési tömegek 95%-os konfidenciaintervalluma [2840,0–3049,2] gramm. (Megjegyezzük, hogy ez a fentitől kissé eltérő módszerrel készült, ami tekintettel van arra is, hogy itt most – szemben a fenti példával – nem ismerjük *a priori* a sokaság szórását.) Ez azt jelenti, hogy a *legjobb* tippünk a születési tömeg sokasági várható értékére a 2944.5873016 gramm, de azt is tudjuk ezen felül mondani, hogy bár ez csak bizonytalan tipp (hiszen a becsült érték mintáról-mintára ingadozik), de 95%-os *megbízhatósággal* azért kijelenthető, hogy nem kisebb a keresett, ismeretlen sokasági várható érték mint 2840,0 gramm és nem nagyobb mint 3049,2 gramm. (Amit úgy értünk, hogy azt becsüljük, hogy ha a sokaságból 100 mintát vennénk, és mindegyikből ugyanígy megkonstruálnánk a konfidenciaintervallumokat, akkor várhatóan 95 esetben tartalmazná a CI a valódi, sokasági értéket.) Érdemes megfigyelni, hogy a konfidenciaintervallum két végpontja szimmetrikus a pontbecslésre; ez a várható érték becslésére jellemző, de más paramétereknél nem feltétlenül van így.

Itt is hasznos mindezeket egy szimulációval szemléltetni (3.3. ábra).

```
SimData <- data.frame(
  idx = 1:100,
  CI = t(replicate(
    100, TeachingDemos::z.test(rnorm(30, 70, 10),
                               stdev = 10)$conf.int)))

ggplot(SimData, aes(xmin = CI.1, xmax = CI.2, y = idx)) +
  geom_linerange() +
  geom_vline(xintercept = 70, color = "red") + labs(y = "")
```



3.3. ábra. Konfidenciaintervallumok szemléltetése szimulációval.

3.3. Hipotézisvizsgálat

Az induktív statisztika másik nagy ága a hipotézisvizsgálat. A hipotézisvizsgálat nagyon sok szempontból a becsléselmélet, ezen belül is az intervallumbecslés elméletének ikertestvére (ami ekvivalens, csak átfogalmazottan felírt egyenletekre vezet), mégis, saját szóhasználata, fogalomköre, és hatalmas gyakorlati jelentősége indokolja, hogy külön tárgyaljuk.

Amíg a becsléselmélettől azt vártuk, hogy nyilatkozzon egy számunkra ismeretlen jellemzőről, addig a hipotézisvizsgálat esetében van előzetes elképzelésünk a jellemző értékéről (például, hogy egy adott számmal egyenlő) – csak épp nem tudjuk, hogy ez igaz-e. Ha az előzetes feltevésünk mintára vonatkozna, akkor nem is volna semmi probléma: kiszámítjuk a jellemzőt a mintából, és megnézzük, hogy teljesült-e a feltevésünk. Mivel azonban a feltevés a sokaságra vonatkozik, így megint csak visszatérünk oda, hogy erről biztos döntést hozni lehetetlen minta alapján – de valószínűségit lehet. Nem tudjuk megmondani, hogy a sokaság átlagos testtömege 70 kg-e, ha a mintabeli átlag 65 kg… de meg fogjuk tudni mondani (egyéb mintaadatok felhasználásával), hogy *mennyire hihető*, hogy 70 kg a sokasági átlag. Erre szolgál a hipotézisvizsgálat. Már most fontos megjegyezni, hog a hipotézisvizsgálat logikája bizonyos szempontból fordított: az előbbi kérdés ellentétére keresi a választ, arra, hogy ha 70 kg *lenne* a sokasági átlag, akkor mennyire lenne valószínű, hogy ettől olyannyira eltérő eredményt kapunk, mint a 65 (vagy

annál is kisebb). Ha nagyon, akkor azt mondjuk, hogy „minden bizonnyal” nem 70 kg volt az átlag.

A problémát az adja, hogy – maradva a fenti példánál – nem tudhatjuk, hogy mi okozta ezt az 5 kg különbséget. Valójában tényleg 70 kg a sokaság átlaga, csak a mintavételi ingadozás játéka miatt pont olyan mintát fogtunk ki, amiben picit kisebb volt az átlag, vagy ez az 5 kg különbség olyan nagy, ami túlmutat a mintavételi ingadozáson, és azt kell feltételeznünk, hogy a hátterében sokasági hatás (is) van (tehát, hogy a sokasági átlag kisebb mint 70 kg)…?

Amint a fentiekből is kiderült, a hipotézisvizsgálat mindig a sokaságra megfogalmazott állításból indul ki. Valójában nem is egy, hanem rögtön két állítást használ a hipotézisvizsgálat; nevük nullhipotézis (H_0) és ellenhipotézis (H_1) melyek jellemzően egymás komplementerei. (Azaz egymást kizárják, de a kettőből valamelyik biztosan fennáll.) A fenti példát így írhatnánk:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

úgy, hogy μ_0=70 kg.

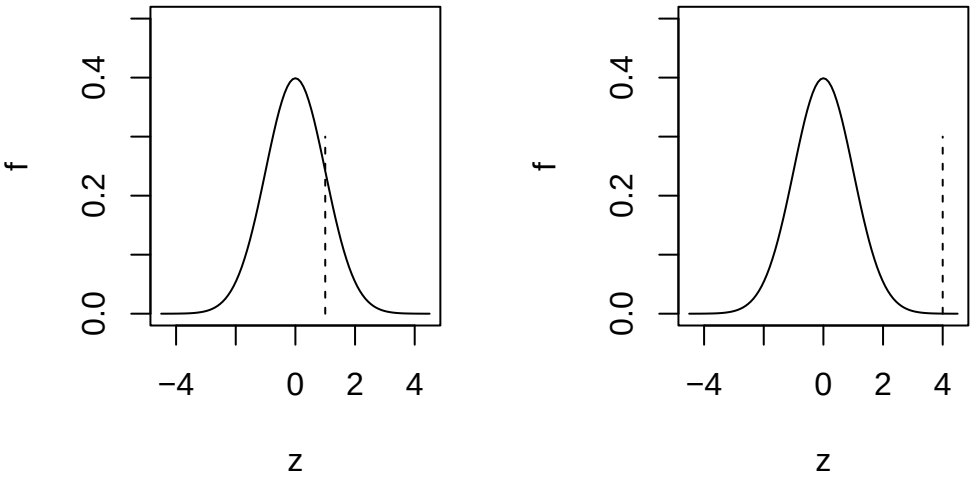
Amit fontos észben tartani, hogy hipotézisvizsgálatnál az erős döntés mindig az elutasítás tud lenni, ezért a legtöbb próba úgy van megszerkesztve, hogy a szakmailag „izgalmas” állítás, a tudományos nóvum (hatásos a gyógyszer, van eltérés a laboreredményben stb.) az ellenhipotézisbe kerüljön. Pontosan emiatt az elutasítás esetén nagyon gyakran – szinonimaként – azt mondjuk, hogy a „próba szignifikáns”.

A hipotézisvizsgálat központi eszköze a **próbafüggvény** (vagy más szóval **tesztstatisztika**). Az egész eszközt együtt **tesztnek** vagy **próbának** nevezzük. A próbafüggvény a mintaelemek függvénye, ilyen módon a próbafüggvénynek is eloszlása lesz. És itt jön a kulcs: a próbafüggvényt úgy választjuk meg, hogy H_0 fennállása esetén valamilyen *pontosan ismert* eloszlást kövessen; ezt szokás *nulleloszlásnak* is nevezni. Természetesen a próbafüggvény *konkrét értéke* függeni fog a mintaelemektől, de az *eloszlása* nem függhet ettől (sem más, ismeretlen paramétertől, ha volna ilyen).

Hogy megértsük, hogy ez miért lesz alkalmas a hipotézispárról történő (valószínűségi) döntéshozatalra, nézzünk egy konkrét példát. Folytatva az előző példát, tegyük fel, hogy sokaságunk eloszlása normális, ismert szórással. Amint már megbeszéltük, ekkor $\overline{X} =\sim \mathcal{N}\left(\mu, \sigma_0^2/n\right)$. Ez tehát a mintaelemek függvénye, és elvileg próbafüggvénynek is nevezhető, mert ha érvényesítjük rajta H_0-t (azaz H_0-t igaznak fogadjuk el), akkor azt kapjuk, hogy $\overline{X} =\sim \mathcal{N}\left(\mu_0, \sigma_0^2/n\right)$, ami valóban már nem függ ismeretlen paramétertől. Ezzel, és a technikailag szintén megfelelő $\overline{X} - \mu_0 \sim \mathcal{N}\left(0, \sigma_0^2/n\right)$-nel is az a gyakorlati baj azonban, hogy nagyon nehézkes lenne a használatuk, hiszen bár a nulleloszlás ismert, de minden μ_0-ra, σ_0-ra és n-re más és más – azaz ezektől függően minden egyes hipotézisvizsgálathoz elő kéne keresni az adott eloszlást.

A $\overline{X} - \mu_0$ azonban már mutatja az utat: próbálkozzunk a $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ próbafüggvénnyel (jele általában Z)! Ez már minden szempontból tökéletes lesz, hiszen nulleloszlása $\mathcal{N}(0,1)$, azaz minden paramétertől függetlenül ugyanaz; egyetlen eloszlással elvégezhető az összes ilyen típusú hipotézisvizsgálat e körülmények között.

Foglaljuk össze hol tartunk! Konstruáltunk egy olyan függvényét a mintaelemeknek, melynek ismerjük az eloszlását *ha* fennáll a nullhipotézis. Ki tudjuk azt is számolni, hogy mennyi ennek a próbafüggvénynek az értéke a konkrét (realizálódott) mintánkból; ezt szokás empirikus értéknek (z_{emp}) is nevezni. Innentől úgy okoskodhatunk: biztos döntést lehetetlen hozni (ez az előbbi példán nagyon jól látszik: a $\mathcal{N}(0,1)$ nulleloszlás tartója az egész számegyenes, tehát még ha fenn is áll a nullhipotézis, elvileg *akármilyen* szám realizálódhat belőle, az elvileg bármilyen szám lehet a mintából kiszámított próbafüggvény értéke, azaz z_{emp}), de mégis, mennyire hihető, hogy a szaggatott vonallal jelölt érték a folytonosan behúzott eloszlásból realizálódott a következő esetekben (3.4. ábra).



3.4. ábra. A hipotézisvizsgálat alapgondolatának szemléltetése.

Bár *elvileg* mindkettő előfordulhat, de a bal oldalit *hajlamosak vagyunk* elhinni, a jobb oldalinál viszont épp ellenkezőleg, *hajlunk arra*, hogy azt gondoljuk, hogy az empirikus érték valójában más eloszlásból realizálódott. Noha elvileg a bal oldali is jöhet más eloszlásból, és a jobb oldali is ebből – ezért a bizonytalan megfogalmazások, mutatva, hogy ezek csak valószínűségi állítások.

Precízebben megfogalmazva: az kicsi valószínűségű esemény ($\mathcal{N}(0,1)$ eloszlás esetén), hogy

± 3-on kívül számot kapjunk. Ha *mégis* ilyen érték jön ki, akkor joggal kérdőjelezzük meg, hogy a próbafüggvény ilyen eloszlást követett – márpedig, ha fennáll a nullhipotézis, akkor ilyen eloszlást *kellett* követnie, így más szóval mi most arra következtettünk, hogy nem áll fenn a nullhipotézis!

Ez persze bizonytalan döntés, és itt jól látszik ennek az oka: nagyon is kijöhet ± 3-on kívül szám *még akkor is*, ha fennáll a nullhipotézis, sőt, ennek a valószínűsége akár számszerűen is meghatározható ($\Phi(-3) + [1 - \Phi(3)]$ ami kb. 0,27%). Ha a ± 3-on kívüli tartományra mondjuk az, hogy ide eső empirikus tesztstatisztika esetén „már nem hisszük el", hogy fennállt a nullhipotézis, akkor pontosan (és előre megmondható, általunk tudott módon) 0,27% valószínűséggel fogunk hibás döntést hozni: ekkora a valószínűsége ugyanis, hogy fennálló H_0 esetén is ilyen extrém tesztstatisztika jöjjön ki.

Ha ez számunkra túl nagy, akkor megtehetjük, hogy mondjuk csak a ± 4-en kívüli értékeket tekintjük „gyanúsnak" – csakhogy ekkor a valódi különbségek felderítését is megnehezítjük.

Az tehát egy kompromisszum eredménye, hogy „hol húzzuk meg a határt". A gyakorlatban ezt úgy hajtjuk végre, hogy az eloszlás legextrémebb, tehát a nullhipotézis fennállása esetén várt értéktől legtávolabb eső részein (a mostani példánkban: mindkét szélén szimmetrikusan) kijelölünk egy olyan tartományt, melynek egy adott, kicsi érték (jele α) a valószínűsége[6]. Más szóval azt mondjuk, hogy ebbe az intervallumba elvileg ugyan eshet egy realizálódott érték akkor is, ha a nulleloszlás fennáll, de ennek olyan kicsi a valószínűsége, hogy ezt már nem tartjuk hihetőnek (hivatkozva arra, hogy ez a tartomány fekszik a legtávolabb nullhipotézis fennállása esetén várt értéktől). Tökéletesen látszik azonban, hogy csak bizonytalan döntést tudunk hozni: ez a kijelentésünk *automatikusan* az esetleges hibázás elfogadását jelenti – nagyon is tudjuk, hogy ebbe a tartományba eshet a realizálódott érték a nulleloszlás fennállása esetén is, mi *mégis* azt mondjuk, hogy ekkor már nem hisszük el a nullhipotézist. Mivel a normális eloszlás tartója az egész számegyenes, így egyértelmű, hogy ennél jobbat nem tudunk tenni, valahol korlátot kell húznunk.

Ilyen módon kijelöltük, hogy milyen empirikus tesztstatisztika-értékek esetén fogadjuk el a nullhipotézist (**elfogadási tartomány**), és milyenek esetén nem (**elutasítási (vagy kritikus) tartomány**). Látható, hogy a tartományok helyét az α valószínűség szabja meg, ennek a valószínűségnek a neve: **szignifikanciaszint**.

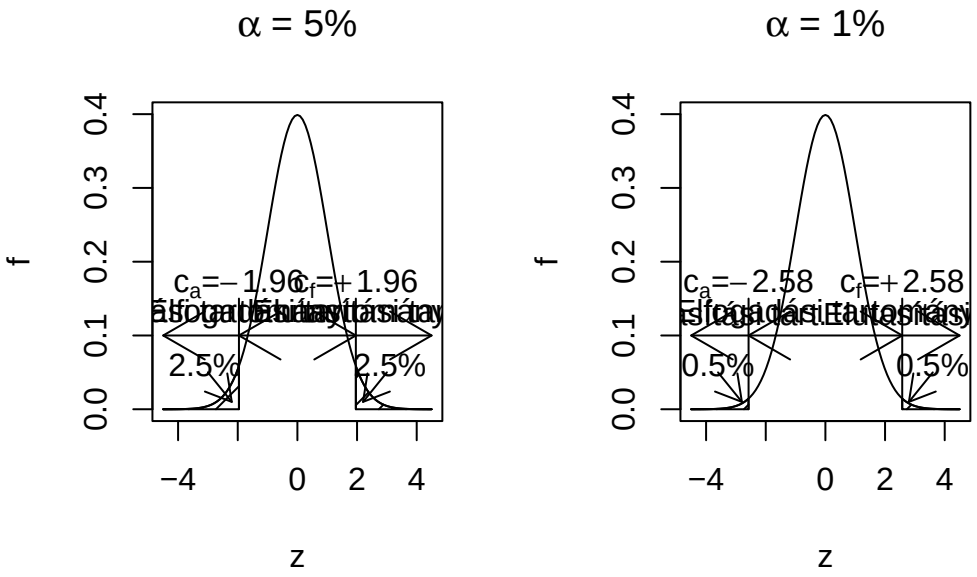
Ebben a feladatban a túl magas és a túl alacsony tesztstatisztika érték is ugyanúgy az elvetés irányába mutat[7], így az elfogadási tartományt valóban a nullára szimmetrikusan jelöljük ki. Ha például azt mondjuk, hogy a szignifikanciaszint 5%, azaz a legextrémebb 5%-nyi területen utasítsunk el, akkor azt úgy tehetjük meg, hogy a nulleloszlás alsó és a felső szélén is 2,5-2,5%-nyi valószínűséget vágunk le. Ezeket a „szétvágási pontokat", melyek az elutasítási és az elfogadási

[6] Egy tartomány valószínűsége alatt most azt értjük, hogy adott eloszlás mellett mekkora annak a valószínűsége, hogy az eloszlásból realizálódott érték a tartományba esik, azaz mennyi a sűrűségfüggvény integrálja a tartomány felett.

[7] Ez nem szükségszerű, a hipotézispár függvényében léteznek ún. egyoldali próbák is, de ezzel most nem foglalkozunk.

tartományokat határolják, **kritikus értékeknek** szokás nevezni. Mivel a nulleloszlás ismert, így ezek könnyen számszerűsíthetőek is mint a 0,025-ös és a 0,975-ös kvantilisei az eloszlásnak; például $\alpha = 5\%$-ra a két kritikus érték a $c_a = -1{,}96$ alsó kritikus érték és a $c_f = +1{,}96$ felső kritikus érték.

Mindezeket összefoglalóan szemlélteti a 3.5. ábra, $\alpha = 5$ és $\alpha = 1\%$-os szignifikanciaszintekre.



3.5. ábra. A hipotézisvizsgálat döntésének szemléltetése két szignifikanciaszint mellett.

Amint arra már utaltunk is, α beállításával a hipotézisvizsgálatban elkövethető kétféle hiba között egyensúlyozunk. Az egyik tévedési lehetőség, hogy fennáll a nullhipotézis, mi mégis elvetünk (ennek neve **elsőfajú hiba**; a valószínűsége felett nagyon is erős kontrollunk van, hiszen az épp α); a másik hibázási lehetőség, hogy elvethetnénk a nullhipotézis, mi mégis elfogadunk (ennek neve **másodfajú hiba**, a valószínűségét β-val szokás jelölni; β értékét nem tudjuk jól kézben tartani, hiszen attól is függ, hogy konkrétan milyen ellenhipotézis áll fenn, amit általában mi sem tudhatunk). Ha α-t növeljük („beljebb húzzuk" a kritikus értékeket, növeljük az elutasítási, csökkentjük az elfogadási tartomány méretét), akkor megemeljük a téves elutasítás, és lecsökkentjük a téves elfogadás valószínűségét, ha α-t csökkentjük („kijjebb toljuk" a kritikus értékeket, növeljük az elfogadási, csökkentjük az elutasítási tartomány méretét), akkor megemeljük a téves elfogadás, és lecsökkentjük a téves elutasítás valószínűségét. Az $\alpha = 5\%$ egy tipikus kompromisszum a kétféle hibázás között. Kiegészítésként megjegyzem, hogy $(1 - \beta)$-t a próba **erejének** szokás nevezni (hiszen azt mutatja meg, hogy ha a valóságban nem áll fenn a nullhipotézis, akkor azt mekkora valószínűséggel fogjuk detektálni).

A fentiekből is érezhető, hogy egy próba eredményének olyan formában történő megadása,

hogy „5%-on szignifikáns” nem a legszerencsésebb, hiszen rögtön adódik a kérdés: vajon 1%-on is szignifikáns lett volna? És 0,1%-on? Nem mindegy, hiszen egy olyan eredmény, mely 5%-on szignifikáns, de 4%-on nem, sokkal nagyobb bizonytalanságú, mint egy olyan, ami 0,1%-on is szignifikáns. Megoldás lehetne a tesztstatisztika konkrét értékének megadása, ez azonban gyakorlati szempontból nehézkes, hiszen így minden esetben meg kéne nézni, hogy mi a nulleloszlás (hiszen a tesztstatisztika empirikus értékét muszáj ahhoz viszonyítani). Éppen ezért a mai gyakorlatban inkább azt adják meg, hogy *melyik lenne* az a szignifikanciaszint, ami mellett a tesztstatisztika empirikus értéke épp az elutasítás és az elfogadás határa kerülne. Ennek neve: ***p*-érték** (vagy empirikus szignifikanciaszint). Például, gondoljuk azt, hogy próbánk 5%-on elutasít. Ekkor elkezdjük az α-t csökkenteni (ezzel kijjebb húzzuk a kritikus értékeket, bővítjük az elfogadási, szűkítjük az elutasítási tartomány). Elérjük a 4%-ot, az empirikus tesztstatisztikánk még mindig az elutasítási tartományban van, tovább csökkentjük az α-t, és így tovább… míg nem egyszer csak azt vesszük észre, hogy mondjuk 2,31%-on még elutasít a teszt, de 2,29%-on már nem. Ekkor azt mondjuk, hogy a teszt p-értéke 2,3%.

A p-érték tehát nem más, mint a szignifikanciaszint akkor, ha a megfelelő (alsó vagy felső) kritikus értéket a tesztstatisztika empirikus értékének helyére helyezzük át. Ebből az is következik, hogy a p-érték számszerűen a nulleloszlás integrálja az empirikus tesztstatisztikától extrémebb irányba (illetve ennek kétszerese), ugyanúgy, ahogy az α is – definíció szerint – a nulleloszlás integrálja a kritikus értékektől extrémebb irányokba (és itt, ahogy megbeszéltük, a kritikus érték szerepét az empirikus tesztstatisztika játssza). Ennek meghatározása tehát manapság már számítástechnikai szempontból is problémamentes.

Világos, hogy p-érték az elvetésben való bizonyosságunkat fejezi ki. Ez az eredményközlés azért rendkívül praktikus, mert – szemben az előzőekkel – az olvasó „elvégezheti magának” a hipotézisvizsgálatot, és *bármilyen szignifikanciaszinten* döntést hozhat. A p-értéknél magasabb szignifikanciaszinteken elutasítás lesz a döntés (ekkor bővebb az elutasítási tartomány, bele fog esni az empirikus tesztstatisztika), a p-értéknél alacsonyabb szinteken pedig elfogadás (az elutasítási tartomány szűkebb, az empirikus tesztstatisztika az elfogadási tartományba fog esni).

Végezetül egy fontos gyakorlati kérdésre hívom még fel a figyelmet. Amint már beszéltük, az α azt mutatja meg, hogy egy adott próba mekkora valószínűséggel ad téves jelzést. (Emlékezzünk rá, hogy általában mi az elutasítást keressük!) Igen ám, de ha mi két próbát végzünk egymástól függetlenül *úgy*, hogy akkor is találatot deklarálunk, ha *legalább* az egyik teszt szignifikáns lett, akkor valójában már *nem* α valószínűséggel kapunk jelzést akkor is, ha nincs hatás (egyik esetben sem), hanem $1-(1-\alpha)^2$ valószínűséggel! (Hiszen a hibás jelzés annak a komplementere, hogy mindkét teszt jó döntés ad, mivel pedig függetlenek, ezek valószínűsége összeszorzódik.) Ez pedig nagyon nem mindegy, a tipikus $\alpha = 5\%$-ra ez a valószínűség már 9,75%! Tehát valójában majdnem a nominális szignifikanciaszint kétszerese lesz annak a valószínűsége, hogy kapunk elutasítást – miközben a valóságban nincs is hatás egyik esetben sem! Ezt a jelenséget szokás α-inflációnak nevezni. (A kétféle α-t pedig néha megkülönböztetésül comparisonwise (α_C) α-nak illetve familywise (α_F) α-nak nevezik. Az előbbi annak a valószínűsége, hogy egy

teszt hibás jelzést ad (ez az eddig tárgyalt α), az utóbbi annak a valószínűsége, hogy tesztek egy családjából *legalább egy* lesz, ami hibás jelzést ad.) Az összefüggés a kettő között tehát:

$$\alpha_F = 1 - (1 - \alpha_C)^k ,$$

ahol k az elvégzett próbák száma.

Azt a helyzetet, amikor egymással párhuzamosan több, egymástól független hipotézisvizsgálatot futtatunk (és vagylagosan keresünk szignifikáns eredményt), **többszörös összehasonlítások helyzetének** szokás nevezni.

A dolog azt sugallja számunkra, hogy ha sok tesztet végzünk párhuzamosan, akkor valamit tenni kell az ellen, hogy ne találjuk túl könnyen fals elutasításokat. A legegyszerűbb megoldás, ha a tesztenkénti (comparisonwise) szignifikanciaszintet lecsökkentjük. Például, az ún. Bonferroni-egyenlőtlenség szerint $1 - (1 - \alpha)^k \leq \alpha \cdot k$, ezért durva becsléssel úgy korrigálhatjuk a szignifikanciaszintet, hogy elosztjuk a célszintet az elvégzett hipotézisvizsgálatok számával. Ez garantálja, hogy a k teszt elvégzését *együttesen tekintve* sem lehet a kitűzött szignifikanciaszint feletti az elsőfajú hibák aránya.

A módszer hátránya, hogy túl drasztikus: annyira megnehezíti a nullhipotézis elvetését, hogy a valós különbségek is „el fognak veszni". Vannak módszerek, melyek ezt enyhítik (pl. Holm–Bonferroni-korrekció), illetve melyek teljesen más elven próbálják elérni az α-infláció enyhítését (pl. FDR).

Itt hívom fel a figyelmet az ún. **szignifikanciavadászat** jelenségére. Ez lényegében nem más, mint a többszörös összehasonlítások helyzetének rosszindulatú kiaknázása inkorrekt következtetésre. A szignifikanciavadászat jelenségét inkább egy példával illusztráljuk: tegyük fel, hogy bizonyítani akarjuk, hogy a hétfőn és kedden született emberek laboreredményei között szignifikáns eltérés van. Bár ez ránézésre látható módon abszurdum, a fentiek kihasználásával tulajdonképpen nem is nehéz bizonyítani: manapság már a rutinszerűen vizsgált laborparaméterek száma is eléri a 20-30-at, így nincs más dolgunk, mint mindegyiket összehasonlítani! Természetesen valós különbség sehol nem lesz, de mivel 5% valószínűséggel mindegyik adhat téves jelzést, így 30 között már az lenne a meglepő, ha nem kapnánk egyetlen elutasítást sem. Ha a vizsgálatot – korrekt módon – úgy publikáljuk le, hogy összehasonlítottunk 30 laborváltozót 5%-on, és közülük 1 esetben, az XYZ-nél szignifikáns különbséget találtunk, akkor mindenki rögtön tudni fogja, hogy mi történt (azaz, hogy nem jelenthetjük ki, hogy találtunk bármit is). Igen ám, de ha inkorrekt módon játszunk, akkor azt tesszük, hogy a cikket úgy írjuk meg, hogy mi *előre* tudtuk, hogy XYZ-ben lesz különbség (mert van egy ragyogó kórélettani modellünk, mely az XYZ termelését a születés napjával hozza összefüggésbe), és ezért *célirányosan* XYZ-t leteszteltük, és lám: valóban szignifikáns különbséget is kaptunk…! Ezzel szemben nehéz védekezni, hiszen magából az eredményközlésből nem lehet rájönni, hogy mi történt (de természetesen a vizsgálat reprodukciója azonnal lebuktatja a csalást).

Zárásként részletesebb indoklás nélkül felhívom három összefüggésre a figyelmet.

1. Nagyon fontos gyakorlati probléma, hogy adott feladat vizsgálatára konkrétan melyik próbát használjuk. Ez közel sem triviális kérdéskör, ugyanis a feladat önmagában még nem determinálja a próbát: sok feladat van, amire akár tucatnyi különböző próba is elérhető; ezek tipikusan az előfeltevéseikben különböznek. (Azaz, hogy milyen *a priori* megkötésekkel élnek a sokaságra vonatkozóan.) Ennek kapcsán arra fontos felhívni a figyelmet, hogy egyrészt ha egy próba előfeltevései nem teljesülnek, de mi mégis alkalmazzuk, akkor nem garantált, hogy valid végeredményt kapunk (abban az értelemben, hogy az elsőfajú hibák várható aránya nem fog egyezni a szignifikanciaszinttel), másrészt viszont a több előfeltevésre építő próbáknak általában kisebb az erejük. A tanulság, hogy mindig annyi előfeltevésre építő próbát használjunk, amennyit tudunk, se többet se kevesebbet: amely előfeltevésekről tudjuk, hogy teljesülnek (a priori!) azokat építsük be a próbaválasztásba… de többet ne.
2. Rögtön itt érdemes megjegyezni, hogy – bár egyes statisztikai programcsomagok notóriusan az ellenkezőjét sugallják – elvileg nem illik az alapján dönteni, hogy milyen próbát használunk, hogy az előfeltevéseit *ugyanazon* mintán *egy másik próbával* leellenőrizzük. Ezért hangsúlyoztuk az előbbi pontban, hogy a feltevésekről *a priori* kell döntenünk (korábbi eredmény, másik mintán végzett teszt stb. alapján).
3. Végül felhívjuk a figyelmet, hogy egy próba erejét önmagában növeli a nagyobb mintanagyság. Pontosan ezért a klasszikus mondás szerint: „kis hatás kimutatásához nagy minta kell, nagy hatáshoz elég a kisebb minta is!”.

Zárásként nézzünk meg egy konkrét példát a hipotézisvizsgálat alkalmazására is: vizsgáljuk meg azt a kérdést, hogy a dohányzó anyák újszülötteinek születési tömege eltér-e a nemdohányzó anyák újszülötteitől!

Az első kérdés, hogy mit értünk az alatt, hogy „eltér”. Ezt többféleképp is lehetne operacionalizálni, most maradjunk annál a – kézenfekvő, és klinikailag is releváns – megközelítésnél, hogy a várható születési tömegük kisebb-e. (Tehát a kérdést a várható értékek egyezésére hegyezzük ki, nem az érdekel minket, hogy például a szórása a születési tömegeknek eltér-e a két csoportban.)

Az adatbázisban 115 nemdohányzó és 74 dohányzó anyától származó újszülött van. Gyorsan kiszámolhatjuk, hogy az előbbi csoportban az újszülöttek átlagos születési tömege 3055,7 gramm, míg az utóbbiban 2771,9 gramm. Mondhatjuk akkor, hogy a dohányzó anyák újszülöttjei kisebb tömegűek? Természetesen nem! Ez ugyanis csak annyit mondott, hogy a *mintában* kisebb a tömegük, de minket természetesen nem a konkrét minta érdekel, hanem a sokaság! Kijelenthetjük ez alapján, hogy a sokaságban is kisebb a dohányzó anyák újszülöttjeinek a várható születési tömege? Nem, a helyzet nem ilyen egyszerű: elképzelhető, hogy mindkét csoportnak ugyanannyi (a sokaságban!) a várható születési tömege, csak épp pont olyan mintát vettünk, amiben a dohányzó anyáknál ez kisebb. (Hiszen ez tökéletesen véletlen mintavétel esetén is előfordulhat – mintavételi ingadozás, ugyebár!) Sőt, akár az is lehet, hogy épp a dohányzó anyák újszülöttei nagyobb születési súlyúak várhatóan, csak a mintavétel ördöge az ő csoportjukból pont kicsi, a nemdohányzó csoportból meg nagyobb újszülötteket dobott ki.

A kérdésről tehát *biztosat* nem lehet mondani – de statisztikai próbával *valószínűségi kijelentést* tehetünk. Elsőként döntenünk kell arról, hogy milyen próbát alkalmazzunk. Ennek a részletei számunkra most nem fontosak, a lényeg csak a végeredmény: a körülmények (két független csoport, aránylag nagy mintanagyság mindkét csoportban, *a priori* nem ismert sokasági szórás) a választásunk az ún. Welch-próbára esik. Ennek nullhipotézise, hogy a két csoport várható értéke között nincs különbség, ellenhipotézise, hogy van, a két várható érték nem egyezik.

Végezzük el a próbát:

```
t.test(bwt ~ smoke, data = birthwt)
```

```
    Welch Two Sample t-test

data:  bwt by smoke
t = 2.7299, df = 170.1, p-value = 0.007003
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
  78.57486 488.97860
sample estimates:
mean in group 0 mean in group 1
       3055.696        2771.919
```

A p-érték: $p = 0{,}007$, ez minden szokásos szignifikanciaszintnél kisebb (még az 1%-ot sem éri el), így kijelenthetjük: a várható értékek egyezésére vonatkozó nullhipotézis minden szokásos szignifikanciaszinten elvethető, azaz minden szokásos szignifikanciaszinten kijelenthető, hogy a két csoport (sokaságbeli!) várható értéke között különbség van. Azaz: a mintában tapasztalt különbség olyan nagy (a minta egyéb jellemzőit is figyelembe véve), hogy az már túlmutat a mintavételi ingadozás hatásán, nem hihető, hogy betudható pusztán a mintavételi ingadozás hatásának. Azt kell feltételeznünk, hogy mögötte sokasági hatás, vagyis sokaságban is eltérő várható érték van.

Mindezt röviden úgy is megfogalmazhatjuk, hogy a különbség szignifikáns, még más szóval, hogy a két csoport között lényeges különbség van. (Ebben a kontextusban a „lényeges” statisztikai értelemben szignifikánsat jelent.) Ez a jó pont arra, hogy felhívjuk a figyelmet a különbségre a – most definiált – *statisztikai szignifikancia* és a – köznapi értelmű – *klinikai szignifikancia* között. E kettőt mindig szigorúan különböztessük meg egymástól! A köznapi szóhasználatban a „lényeges különbség” alatt ugyanis azt értjük, hogy a tárgyterületi (esetünkben: orvosi) skálán mi bír jelentőséggel. 1 grammal nagyobb születési tömegnek semmi (klinikai) jelentősége (nem gondol az orvos más klinikai helyzetre, nem rendel más vizsgálat, más kezelést stb.), 500 grammnak nagyon is lehet. A statisztikai szignifikancia viszont *teljesen mást* mér: azt, hogy mennyire hihető, hogy a különbség betudható a mintavételi ingadozásnak! Adott esetben lehet 500 gramm különbség is – statisztikailag – inszignifikáns (ha nagy a szórás, vagy kicsi a

mintanagyság), és lehet 1 gramm különbség is – statisztikailag – szignifikáns (ha kicsi a szórás, vagy nagy a mintanagyság).

4 Haladó adatvizualizáció

A deskriptív statisztika kapcsán (2. fejezet) lépten-nyomon láttunk adatvizualizációkat. Még csoportosítási szempont is volt, úgy hívtam ott az ilyeneket, hogy a deskriptív statisztika grafikus eszközei. Annyiban azonban korlátozott volt az adatvizualizáció ottani tárgyalása, hogy kizárólag az R beépített[1] lehetőségeit használtuk; ezt úgy is szokták hívni, hogy base R grafika.

Ebben a fejezetben elsőként (4.1. alfejezet) jobban kontextusba helyezzük a base R grafikát: mi az egyáltalán, milyen előnyei és milyen korlátai vannak? Erre már csak azért is szükség van, mert korábban, a deskriptív statisztikáról szóló fejezetben ezeket az eszközöket magától értetődőnek vettem, anélkül, hogy jobban vizsgáltuk volna az előnyeit és hátrányait. A korlátok miatt több, tudományos adatvizulizációra kifejezetten alkalmas haladó adatvizualizációs csomag is kialakult az évtizedek alatt – mi is megismerkedünk ezt követően az egyik ilyen csomaggal, a `ggplot2`-vel (4.2. alfejezet). Elsőként megismerjük az általános működését, majd megnézzük a korábban már látott, base R grafikás – elemi – vizualizációk `ggplot2` alatti megvalósítását. Ha ez megvan, akkor továbbléphetünk a fejezet fő céljára: meg fogunk ismerkedni a többváltozós vizualizációkkal, különösen hangsúllyal kitérve a vizualizáció stratégiai kérdéseire, tehát immár elszakadva a technikai megvalósítástól, arra fogunk fókuszálni, hogy adott tudományos kérdéshez mi lesz a legcélszerűbb ábra, hogyan tudjuk megtervezni úgy a vizualizációt, hogy legjobban segítse a kérdés megválaszolását.

4.1. A base R grafika és korlátai

A deskriptív statisztikánál (2. fejezet) mindenhol base R grafikát használtam az adatvizualizációk elkészítésére, így az ottani kódok és ábrák áttekintése már jó képet ad e vizualizációs rendszer főbb jellemzőiről. A base R grafikában minden ábratípushoz egy saját függvény tartozik, és az ábra testreszabását a függvény argumentumainak beállításával tudjuk végezni. (Az argumentumok meglehetősen konzisztensek base R grafikán belül, tehát ha valahol megtanultuk, hogy a `main` argumentummal lehet címet adni az ábrának, akkor jó eséllyel számíthatunk arra, hogy ez minden base R grafikás ábránál így lesz.) Base R grafikával jellemzően elemi ábrákat lehet legyártani, azaz olyanokat, amik egyszerűbb esetekre önmagukban is megoldást jelentenek, a bonyolultabbakat pedig nagyon gyakran ezekből, mint építőkövekből kell összerakni (vagy több

[1] Beépített, azaz külön csomag betöltése nélkül is elérhető. Valójában ezek a függvények is valamilyen csomagban vannak – jellemzően a `graphics`-ban – csak ezek a csomagok automatikusan betöltődnek az R indulásakor.

kombinálásával, vagy egynek valamilyen variálásával). Mindazonáltal ehhez az építkezéshez az base R grafika már jellemzően kevés segítséget ad.

A base R grafika megtanulása, korlátaival együtt is, hasznos. Egyrészt, mert egyszerű, és abban a fejezetben a hangsúly sokszor az elméleti kérdéseken volt, így jól jött, hogy az R-es kivitelezés nem vonta el a figyelmet, gyorsan, könnyen megírható, magától értetődő kódokat lehetett használni. Ez nem csak didaktikai kérdés: a valós napi gyakorlatban is sokszor előfordul, hogy valami egyszerű ábrára kell villámgyorsan rápillantani, amit csak „belső használatra" gyártana le az ember; ilyenkor sokszor az egyébként haladó grafikát használók is inkább csak gyorsan a base R grafikához nyúlnak. Úgyhogy hasznos, ha az ember ezt is ismeri (túl azon, hogy az alapokat „illendő" ismerni). Mindemellett, a base R grafikának előnyei is vannak, például jól támogatja azt, ha egy meglevő ábrára akarunk rárajzolni valamit.

Ezzel együtt is, a base R grafikának komoly limitáció vannak. Az egyik az esztétika: sokan azt mondják, hogy a base R grafikás ábrák nem néznek ki túl jól. Ez persze elég szubjektív (mi az, hogy „jól néz ki" egy ábra?), de ami objektív, hogy base R grafikában az elemi ábrák ugyan könnyen megvannak, de ha komplexebb ábrákat akarunk összerakni, akkor nagyon hamar jönnek a gondok. Egy ponton túl ezeket egész egyszerűen nem is lehet megvalósítani base R grafikában, de ha meg is lehet, az is sokszor macerás, időigényes, sok kódolást igényel. De, ami sokkal fontosabb, hogy ez nem csak idő és energia kérdése: a még nagyobb baj, hogy a sok macera mind-mind hibalehetőség. Nézzünk is meg pár példát erre!

Alanyunk a megszokott adatbázis lesz:

```
data(birthwt, package = "MASS")
birthwt$race <- factor(birthwt$race, levels = 1:3,
                       labels = c("Kaukázusi", "Afro-amerikai", "Egyéb"))
```

A feladat pedig legyen a következő: vizualizáljuk a születési tömegek eloszlását rassz szerint! Azt is mondhattam volna: vizualizáljuk a születési tömegek és a rassz kapcsolatát; de az előbbi megfogalmazás már a megoldást is sugallja: kell egy eszköz mennyiségi változó vizualizálására (amit használnánk ha csak simán a születési tömegek eloszlását akarnánk ábrázolni, mondjuk egy hisztogram vagy magfüggvényes becslő), majd ezt kell alkalmazni, csak most nem egyszer, hanem többször – minden rasszra külön-külön. Ezeket az ábrákat utána persze megfelelően ki is kell rajzolni. (Ha valaki szeretné a dolgot a deskriptív statisztikáról szóló fejezet tipológiájába beilleszteni: ez lényegében egy kétváltozós kapcsolat vizualizálása volt, ahol az egyik változó minőségi, a másik mennyiségi!)

Válasszuk mondjuk először eszközként a hisztogramot! Ekkor tehát több hisztogramot kell legyártanunk, majd valahogy ábrázolnunk. Az utóbbi kapcsán belefutunk a hisztogramok egyik problémájába: nem lehet őket egymás*ra* plottolni, például különböző színekkel megkülönböztetve (pedig az lenne az ideális – erről később még sok szó lesz), ezért csak egymás mellé lehet őket plottolni. Ami nem annyira jó, illetve azt a kérdést is felveti, hogy egymás alá vagy egymás mellé plottoljuk? Erről később még fogunk beszélni, most maradjunk abban, hogy egymás

alá plottolunk. Base R grafikában ezt a feladat csak úgy tudjuk megoldani, hogy a plottolási felületet kézzel szétosztjuk több, kisebb részfelületre – erre szolgál az `mfrow` nevű opció. Ha ezt beállítjuk, akkor az első kiplottolt ábra a bal felső kis részfelületre kerül, és minden további ábra, ha kiadunk valamilyen plottolási utasítást, a következőre (ha elérünk az utolsóig, akkor kezdődik a dolog elölről a bal felső résznél, felülírva az ábrákat).

Nekünk tehát most három sorra (és egy oszlopra, hiszen oszlopokat nem akarunk) kell osztanunk a felületet. Illetve bocsánat, általában véve nem tudhatjuk, hogy mennyire lesz szükségünk, úgyhogy először ezt ki kell derítenünk:

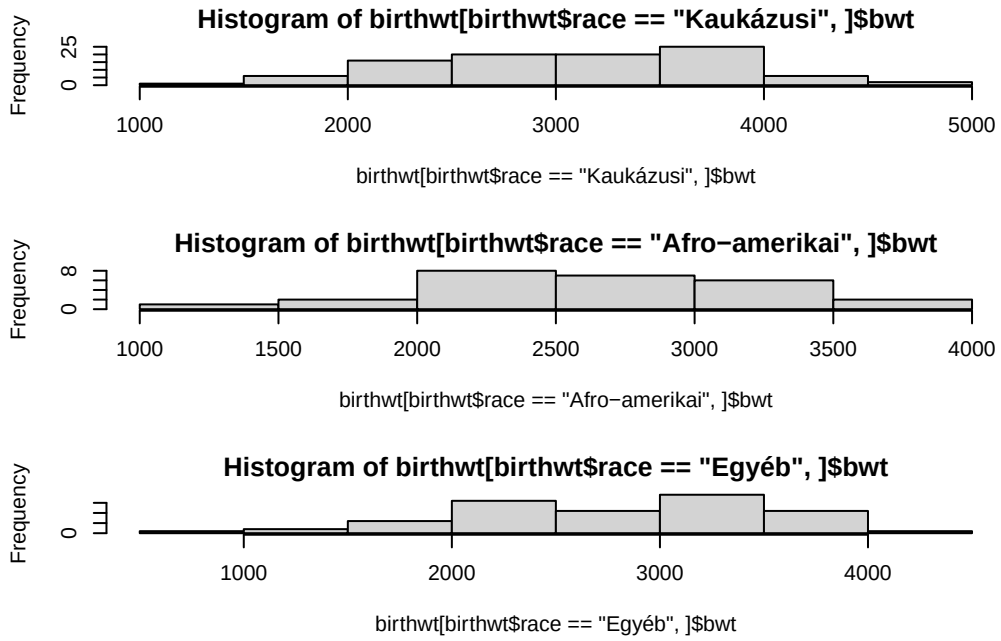
```
length(unique(birthwt$race))
```

```
[1] 3
```

Ha ez megvan, akkor a `par(mfrow = c(3, 1))` paranccsal végrehathatjuk a szétosztást (a `par` parancs szolgál a grafikus paraméterek beállítására). Ezt követően kirajzolhatjuk[2] a hisztogramokat:

```
par(mfrow = c(3, 1))
par(mar = c(5.1, 4.1, 2.1, 2.1))
hist(birthwt[birthwt$race == "Kaukázusi",]$bwt)
hist(birthwt[birthwt$race == "Afro-amerikai",]$bwt)
hist(birthwt[birthwt$race == "Egyéb",]$bwt)
```

[2] A `mar` paraméter beállítására tisztán technikai okokból van szükség. Ez az ábra körüli margók szélességét adja meg base R grafikában; kicsit le kell csökkenteni, hogy kiférjen az oldalon.



Figyeljük meg, hogy a dolog két szempontból is macerás: egyrészt nekünk, kézzel kell megoldani az adatbázis rászűrését a megfelelő rasszra, ráadásul még azt is nyomon kell követnünk[3], hogy milyen rasszok vannak!

Egy apró technikai megjegyzés: ha már nincs szükségünk a felület szétosztására, akkor adjuk ki a `dev.off()` parancsot. Ez ugyanis nem csak törli a plottolási felületet, de egyúttal reset-eli a grafikus paramétereket, így az `mfrow`-t is.

A lényeg mindenesetre, hogy az ábra elkészült, és tudjuk használni!

…vagy mégsem? Túl azon, hogy nagyon rossz a helykihasználás a feleslegesen sok felirat miatt, még egy, elég nagy probléma van: nem ugyanaz a vízszintes tengelyek skálázása! Ami teljesen érthető is: az egyes `hist` hívások a többitől teljesen függetlenül futnak, így mindegyik a *saját* (egy rasszra leszűrt) adatbázisára fogja beállítani a vízszintes tengely határait! Ami igencsak nagy baj, hiszen mi az egész ábrát arra akarjuk használni, hogy a balra-jobbra eltolódásokat keressük, ehhez képest itt *lehetetlen* hogy legyen ilyen: még ha valamelyik rassz 1000 grammal nagyobb vagy kisebb is átlagosan, a fenti ábrán *akkor is* mindenki szép középen lesz… hiszen az `hist` így állítja be. És erre nem válasz az, hogy de ott van a vízszintes tengely beosztása, és azon ez látszik – az egész adatvizualizáció lényege, hogy *segítsük* az olvasót, az a lényeg, hogy minél egyértelműbben, minél automatikusabban, a legkevesebb „kognitív munkával” látszódjon az eredmény az ábrán. Ha számokat kell kiolvasni, több tengelyről, összehasonlítani, majd ez alapján elképzelni, hogy mi hol van, az már régen rossz. (Ezt az „elképzelést” kell nekünk, egy

[3]Ez utóbbin lehetne segíteni egy `for`-ciklussal – itt most joggal használnánk `for`-ciklust, a `hist` egy mellékhatásos függvény – ami ezt a problémát megoldaná, de behozná a `for`-ciklust, mint plusz programozási eszközt; úgyhogy most maradtam az egyszerű, kézi megvalósításnál.

jó ábrával megvalósítani, az olvasó helyett, levéve róla ezt a terhet! Ekkor lesz az ábra igazán jó!)

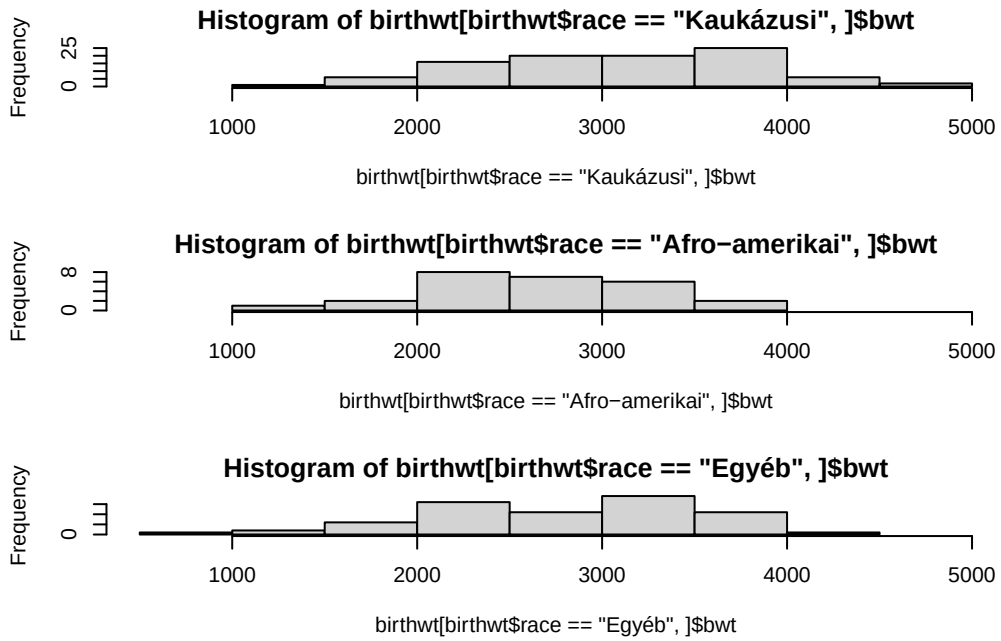
Mi a megoldás? Természetesen az, hogy minden hisztogramot ugyanazzal a vízszintes tengelybeosztással kell kiplottolni. Igen ám, csakhogy ezt mi sem tudhatjuk, hogy mi! Azt kell megnézni, hogy mi a `bwt` teljes terjedelme, hogy mindegyik ábrán kiférjenek az adatok, és erre kell állítani az összes hisztogramot. De ezt mi sem tudhatjuk, hogy mennyi, ezért először kérdezzük le kézzel:

```
range(birthwt$bwt)
```

```
[1]  709 4990
```

Ezt némileg kerekítve most már elkészíthetjük a jó ábrát, természetesen minden egyes hisztogramnál kézzel beállítva ezeket a határokat:

```
par(mfrow = c(3, 1))
par(mar = c(5.1, 4.1, 2.1, 2.1))
hist(birthwt[birthwt$race == "Kaukázusi",]$bwt, xlim = c(500, 5000))
hist(birthwt[birthwt$race == "Afro-amerikai",]$bwt, xlim = c(500, 5000))
hist(birthwt[birthwt$race == "Egyéb",]$bwt, xlim = c(500, 5000))
```



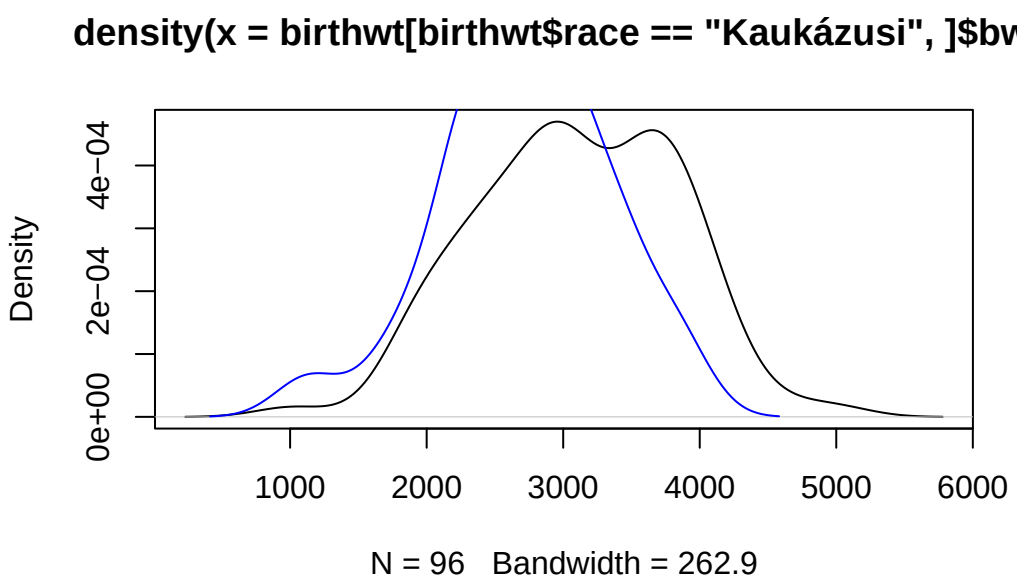
Sikerült létrehozni az ábrát? Igen. Egyszerű volt a dolog? Nem. Meg tudtuk oldani, de oda kellett figyelni, *rengeteg* dolgot kézzel kellett állítani, mindegyik ilyen kivétel nélkül hibalehetőség, a maceráról nem beszélve. És még így is egy olyan ábra jött létre ami… khm, komoly vizuális kihívásokkal terhelt.

Talán ennél is nagyobb problémákra mutat rá a következő példa. A feladat legyen ugyanez, annyi módosítással, hogy most nem hisztogramot, hanem magfüggvényes sűrűségbecslőt akarunk alkalmazni az egyes eloszlások vizualizálására.

A magfüggvényes sűrűségbecslő előnye, hogy ilyenből több is ráplottolható *ugyanarra* az ábrára (nyilván valamilyen módon, például színnel megkülönböztetve ezeket), ami azért előnyös, mert jobban összehasonlíthatóak azok a dolgok, amik ugyanazon az ábrán vannak.

Álljunk neki a feladatnak! Az egyetlen dolog, amire figyelni kell, hogy az *első* KDE-t ábrázolhatjuk `plot`-tal, de a *másodikat* már nem, mert a `plot` mindig elsőként letörli a plottolási felületet. De semmi gond nincs: a `density` objektumot átadhatjuk egy `lines`-nak is, ami pont ezt a problémát oldja meg, mert ugyanazt csinálja mint a `plot`, de nem üríti a felületet, hanem a meglevőre *rá*plottol – pont amire szükségünk van! Nézzük is; természetesen a színt nekünk kell, kézzel beállítanunk:

```
plot(density(birthwt[birthwt$race == "Kaukázusi",]$bwt))
lines(density(birthwt[birthwt$race == "Afro-amerikai",]$bwt), col = "blue")
```

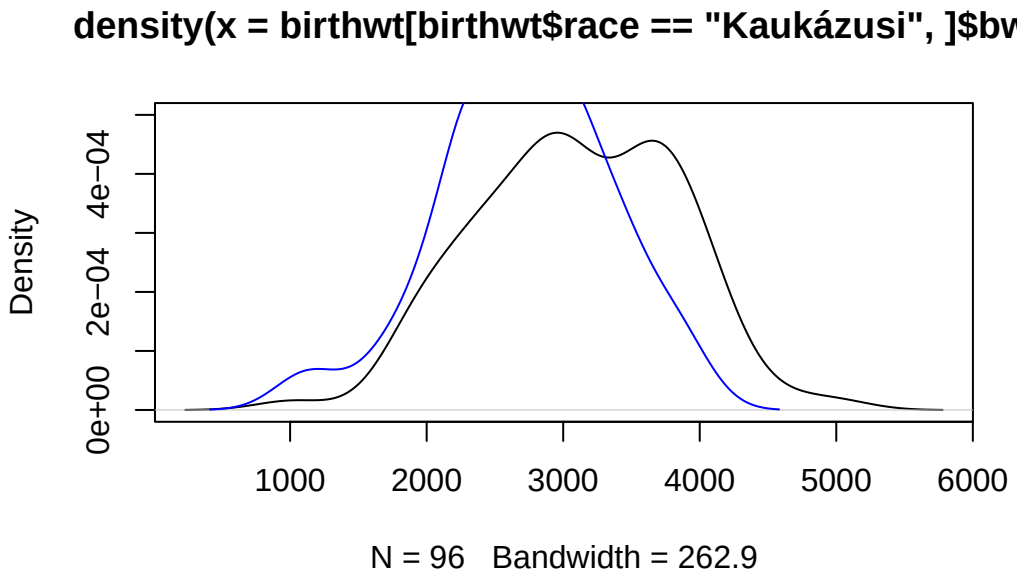


Baj van: a második görbe kifut az ábra tetején. De ha végiggondoljuk, hogy mi történt, akkor nagyon hamar rájövünk, hogy mi a probléma forrása: az ábra koordinátatengelyeit az első

hívás állította be – természetesen az ott használt, kaukázusira leszűrt adatbázis alapján! A `lines` *rá*plottol, vagyis természetesen nem tudja, visszamenőleg, megváltoztatni a függőleges tengely határait…!

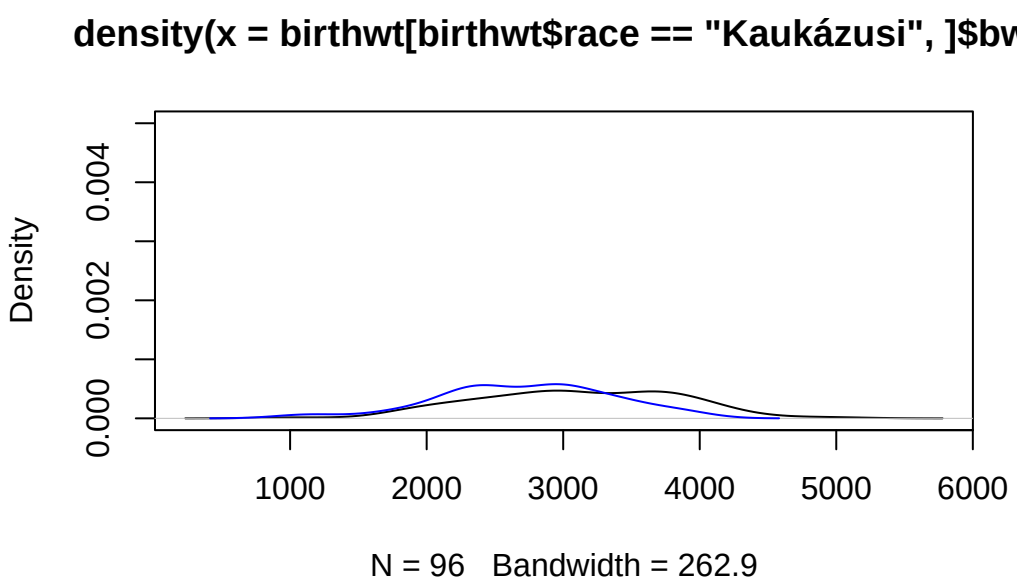
Próbáljunk a dolgon javítani. A megoldás nem bonyolult: egyszerűen meg kell növelni a függőleges tengely tartományát. (Természetesen kézzel beállítva!) Itt azonban van egy plusz-probléma: nem nyilvánvaló, hogy mire kell állítanunk. Az előző példánál is volt hasonló probléma, csak ott legalább ez a gond nem volt, mert a `range` használatával egyszerűen lekértük, hogy mik a határok. De a függőleges tengelyre, tehát, hogy a sűrűség meddig fut fel, nincsen `range`, ezt nem tudjuk sehogy lekérdezni, vagy megtudni…! Mondhatnánk, hogy akkor először plottoljuk a második rasszt (hiszen az úgy beállított tengelyekre az első már biztosan ráfér), de ez sem oldja meg a problémát, mert mi van, ha a harmadik még nagyobb? Voltaképp arról van szó, hogy a kísérletezést sehogy nem tudjuk megspórolni, akkor kísérletezzünk (próbálgatásos alapon – hiszen jobb eszközünk nincs!) inkább a függőleges tengely skálázásával:

```
plot(density(birthwt[birthwt$race == "Kaukázusi",]$bwt), ylim = c(0, 5e-04))
lines(density(birthwt[birthwt$race == "Afro-amerikai",]$bwt), col = "blue")
```



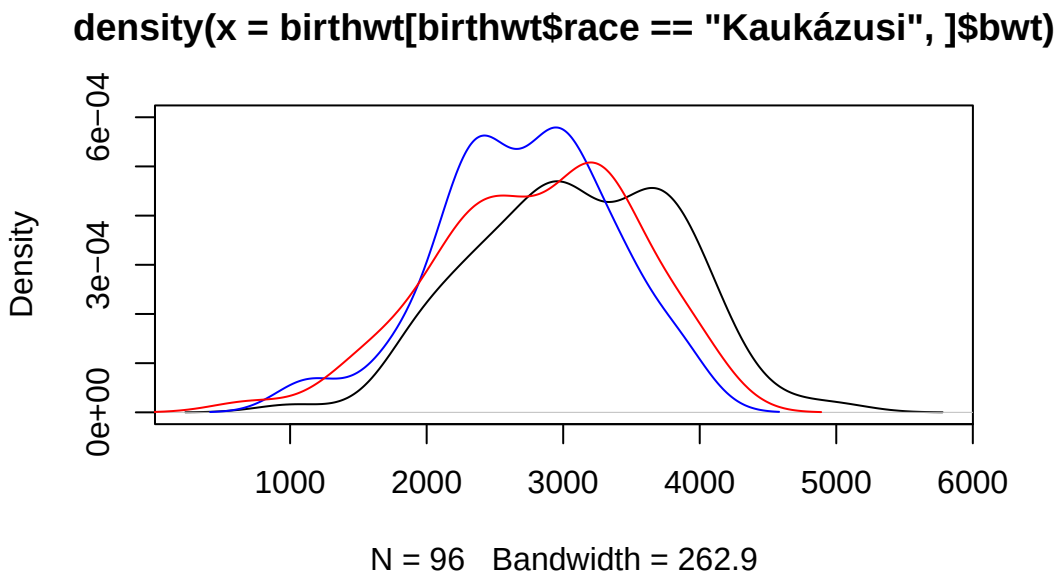
Sajnos nem jó a dolog, ez a felső határ nem elég. Próbáljuk újra:

```
plot(density(birthwt[birthwt$race == "Kaukázusi",]$bwt), ylim = c(0, 5e-03))
lines(density(birthwt[birthwt$race == "Afro-amerikai",]$bwt), col = "blue")
```



Na, ezzel meg túllőttünk a célon – belefér minden ugyan, de nagyon kicsire vannak összenyomva, nem használjuk ki a területet, ami szintén nem jó. Próbáljuk még egyszer:

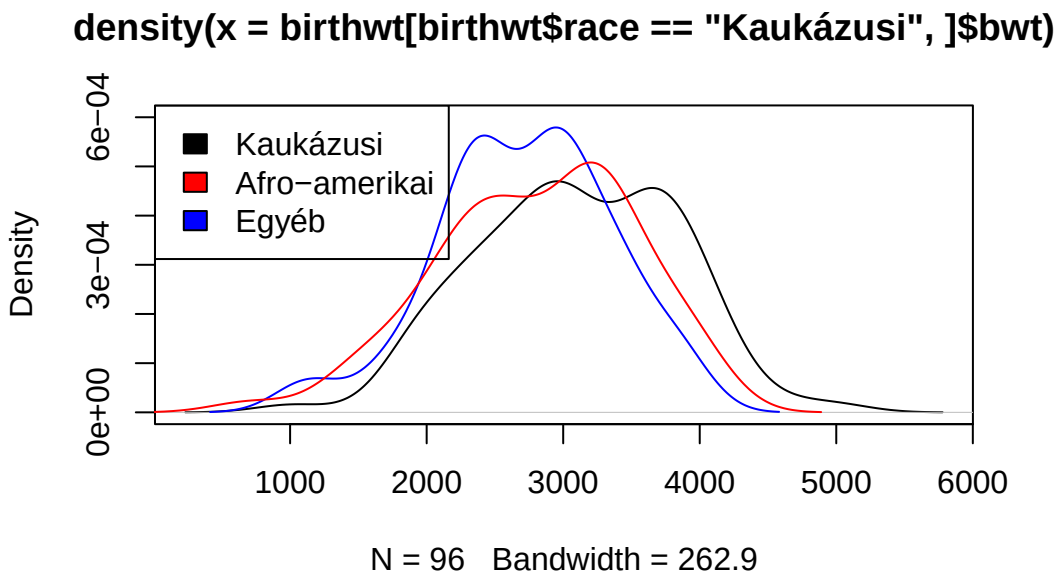
```
plot(density(birthwt[birthwt$race == "Kaukázusi",]$bwt), ylim = c(0, 6e-04))
lines(density(birthwt[birthwt$race == "Afro-amerikai",]$bwt), col = "blue")
lines(density(birthwt[birthwt$race == "Egyéb",]$bwt), col = "red")
```



Most már stimmel!

Illetve… egy apró, de azért meglehetősen kézenfekvő probléma még mindig van: melyik szín mit jelent?? Na igen, ugyanis nincsen jelmagyarázat, e nélkül nem sokra megyünk az ábrával… Szerencsére a base R grafikának van jelmagyarázatot készítő függvénye, az a neve, hogy `legend`, át kell neki adni a jelmagyarázat pozícióját, valamint a feltüntett szövegeket és színeket:

```
plot(density(birthwt[birthwt$race == "Kaukázusi",]$bwt), ylim = c(0, 6e-04))
lines(density(birthwt[birthwt$race == "Afro-amerikai",]$bwt), col = "blue")
lines(density(birthwt[birthwt$race == "Egyéb",]$bwt), col = "red")
legend("topleft", c("Kaukázusi", "Afro-amerikai", "Egyéb"),
       fill = c("black", "red", "blue"))
```



Most már minden tökéletes! Ugye?

…vagy mégsem?

Aki szerint minden tökéletes, nézze meg még egyszer, jobban a színeket…

A harci helyzet ugyanis az, hogy *elrontottam* a jelmagyarázatot. Igen, elrontottam, ugyanis *semmi* nem kapcsolja össze, hogy a jelmagyarázatban mi jelenik meg, és hogy az ábrán mi van…! Következésképp *simán* meg lehet tenni, hogy a jelmagyarázatban *mást* tüntetünk fel, mint ahogy az ábra készült! E felett semmiféle kontroll, ellenőrzés, figyelmeztetés nincs; egyetlen pillatnyi figyelmetlenség a mi részünkről, és teljesen rossz ábra készül! (Ha ráadásul az R kódot nem adjuk meg, csak az ábrát, ez soha ki sem derül…!)

Remélem ez a példa végképp demonstrálta, hogy az ilyen ábrák összerakása base R grafikával, még ha lehetséges is, nagyon macerás, és, ami még nagyobb baj, hatalmas hibalehetőségeket rejt magában. Szoktam mondani, hogy ha egy ábrára `legend` kell vagy több részábrából áll, az jó jel arra nézve, hogy elgondolkozzunk, hogy inkább ne base R grafikában valósítsuk meg.

De akkor miben?

4.2. Haladó adatvizualizációs csomagok, a ggplot2

Az elmúlt évtizedek alatt nagyon komoly munka folyt olyan haladó adatvizualizációs csomagok kialakítására, melyek lehetővé teszik a komplex, adott esetben nagyon komplex tudományos

adatvizualiciók relatíve egyszerű, hibaálló, jól átlátható kóddal történő elkészítését R alatt. Azért fogalmaztam úgy, hogy „relatíve”, mert az ilyen megoldások a legegyszerűbb feladatokra tipikusan bonyolultabbak, mint a base R grafika – de cserében a bonyolultabbakra jobban működnek, vagy egyáltalán, működnek.

A mai napra két csomag kristályosodott ki, amelyeket a fenti célra széleskörűen használnak.