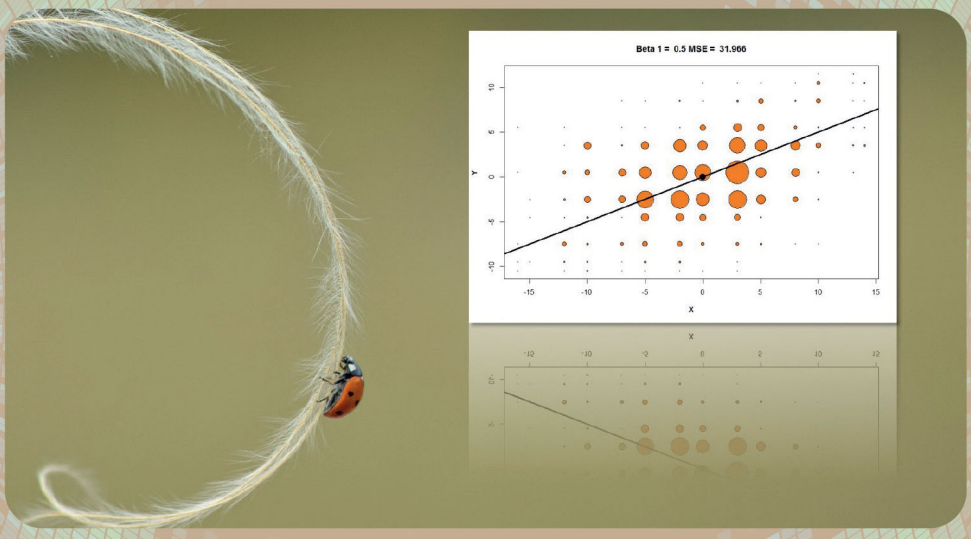
SAPIENTIA
ERDÉLYI MAGYAR
TUDOMÁNYEGYETEM
KOLOZSVÁRI KAR

jegyzet



ZSIGMOND ANDREA-REBEKA

ADATFELDOLGOZÁS ALAPJAI R-REL

ALKALMAZÁSOK A KÖRNYEZETTUDOMÁNYBAN

Scientia Kiadó | 2022

ZSIGMOND ANDREA-REBEKA

ADATFELDOLGOZÁS ALAPJAI R-REL. ALKALMAZÁSOK A KÖRNYEZETTUDOMÁNYBAN

SAPIENTIA ERDÉLYI MAGYAR TUDOMÁNYEGYETEM
KOLOZSVÁRI KAR

ADATFELDOLGOZÁS ALAPJAI R-REL. ALKALMAZÁSOK A KÖRNYEZETTUDOMÁNYBAN

ZSIGMOND ANDREA-REBEKA

Scientia Kiadó
Kolozsvár ▪ 2022

Felelős kiadó:
Sorbán Angella

Lektor:
Korponai János (Baja)

Borítóterv:
Tipotéka Kft.

Kiadói koordinátor:
Szabó Beáta

A szakmai felelősséget teljes mértékben a szerkesztők, illetve a szerzők vállalják.

Első magyar nyelvű kiadás: 2022

Descrierea CIP a Bibliotecii Naţionale a României
ZSIGMOND, ANDREEA-REBEKA
Adatfeldolgozás alapjai R-rel : alkalmazások a környezettudományban /
Zsigmond Andrea-Rebeka. - Cluj-Napoca : Scientia, 2022
Conţine bibliografie
ISBN 978-606-975-068-1

004

TARTALOM

CONTENTS

CUPRINS

ELŐSZÓ

Ez a könyv elsősorban környezettudomány szakos egyetemi hallgatók számára készült azzal a céllal, hogy megismerkedjenek az R-rel, amely egy ingyenes, nyílt forráskódú statisztikai programkörnyezet, és amelynek használata a tudományos világban mára már általánossá vált. A könyv nyolc fejezetre tagolódik, és az adatfeldolgozás alapfogalmainak tisztázása után útmutatót ad a tudományos kutatás megtervezésénél kulcsszerepet játszó szempontokra (mint például a minta elemszámának meghatározása, a mintavétel típusa stb.), részletezi az R nyelvezetét, kitér néhány egyszerű ábra szerkesztésére, megismerteti a környezettudományban használt leggyakoribb hipotézisvizsgálatokat. Ahogyan elvárjuk egy kézikönyvtől, nagy hangsúly esik az R parancsok részletes ismertetésére, amelyeken keresztül az olvasó jobban átlátja az R-rel való kommunikáció sokszínűségét.

Az R programnyelv elsajátítása hosszú időbe telik, és rendszeres gyakorlást igényel, viszont a siker biztosított. Nagy előnye ennek a programnak az, hogy minden lehetőséget megnyit az adatfeldolgozás előtt. Gyakorlatilag nincs olyan statisztikai művelet, olyan számítás, olyan próba, amelyet R-ben ne lehetne elvégezni. Az ingyenessége mellett ez a tulajdonsága az, ami miatt egyre szélesebb körben használják.

A tudományos kutatás olyan, mint egy izgalmas történet, aminek van bevezetője, bonyodalma, kifejtése, tetőpontja és megoldása. A témakör ismeretének alapján megfogalmazunk egy hipotézist, ami a történet bonyodalmát képezi, majd a kutatás kivitelezésén keresztül eljutunk a tetőpontig, ami az adatok értelmezéséből és a belőlük származó információ kinyeréséből áll. Jól megtervezett kutatásban – az R statisztikai eszköztár révén – könnyen értelmezhetőek az eredmények, és tudományosan helytálló következtetések vonhatók le.

A könyv átfogó irodalomra épül, számos magyar és angol, sőt, néhány román nyelvű tudományos mű is szerepel közöttük, amelyeket tanulmányozni lehet. Az angol szaknyelv és az R elsajátításához ajánlott a *The R Book* tanulmányozása (Crawley, 2013). A román nyelvű szakirodalom és az R-rel kapcsolatos további ismeretek elsajátításához három könyv is bátran ajánlható: *Analiza statistică folosind limbajul R* (Păun és Păun, 2009), illetve a *Noţiuni de statistică aplicată cu exemple în R* (Păun, 2016), míg magyarul a *Biostatisztika nem statisztikusoknak* (Reiczigel et al, 2007). Az R-ben használt írásjelekkel való összhang miatt a szövegtörzsben a számok tizedes vesszői kivételesen ponttal szerepelnek, illetve az R parancsokat leíró részekben az idézőjel megőrzi a program írásjelét. A könyv végén egy háromnyelvű szakszótár található, amely a magyar, román és angol szakkifejezéseket felelteti meg egymásnak.

A könyv fejezeteit a megadott sorrendben ajánlott végigvenni, mert logikai és ismereti szempontból így segíti a leghatékonyabban a fejlődést az adatfeldolgozás terén. Gyakorlott szakemberek számára a könyv kézikönyvként használható, ugyanis gyors segítséget tud nyújtani az adatfeldolgozásban R-ben.

1. BEVEZETŐ FOGALMAK

„A tudomány tényekből épül fel, ahogy egy ház
kövekből, de a tények puszta halmaza nem tudo-
mány, ahogy egy kőrakás sem tekinthető háznak."
(Henri Poincré)

A statisztikai elemzés egy hasznos eszköz arra, hogy az adatainkból információt nyerjünk, illetve tudományos alapokra helyezzük a válaszainkat, amelyeket környezeti problémákra szeretnénk adni. A statisztikának két fontos ága van: a leíró vagy deskriptív (*descriptive statistics*) és a következtető vagy induktív (*statistical inference*) statisztika. A leíró statisztika a vizsgált populációt középértékek és szóródási mutatók segítségével jellemezi, amire változatos grafikus módszerek léteznek. Információt szolgáltat a célcsoport (populáció, alapsokaság) eloszlásáról, be lehet azonosítani a kiugró pontokat, mintákat, egyedeket. A következtető statisztika lehetőséget ad arra, hogy populációkat hasonlítsunk össze, előre megfogalmazott hipotéziseket igazoljunk vagy cáfoljunk, trendeket, illetve jövőbeli események bekövetkezésének valószínűségét állapítsuk meg.

Egy tudományos kutatás megtervezéséhez alapos elméleti tudással kell rendelkeznünk, ugyanakkor aktív szemlélődőként tényeket kell gyűjtenünk a számunkra fontos célcsoportról. A tények alapján megfogalmazunk egy *hipotézist* a célcsoporttal kapcsolatban. Henri Poincré állításával összhangban a hipotézis egy olyan állítás, ami összefüggést feltételez a tények között, és ami a tudomány eszközeivel tesztelhető. A hipotézis alapvető tulajdonsága a cáfolhatóság.

Az 1.1. táblázatban megfogalmazott hipotézis tudományos, mert vannak tudományos eszközök arra, hogy az ellentétes állítást bizonyíthassuk: jelen esetben a hipotézis abban a pillanatban megdől, hogy egy test szabadesésben távolodik a földfelszíntől.

A hipotézisek összegzésével egy általános kijelentéshez jutunk, amit *elmélet*nek nevezünk. Az elmélet egy olyan fogalmi keret, ami megmagyaráz létező tényeket, és előrejelez újakat. A hipotézishez hasonlóan az elmélet is cáfolható kell legyen, azaz kell lennie olyan kísérletnek, ami be tudná bizonyítani az elmélet hamis voltát, ha valóban hamis.

Egy aktívan szemlélődő vegyész felfigyelhet arra, hogy a nátrium sói (például $NaCl$, $NaHCO_3$, $NaNO_3$, $NaIO_3$, $NaCN$ stb.) fehér színűek, és kijelentheti azt a hipotézist, hogy ha egy sóban nátriumion szerepel, akkor az a só fehér. Ez a kijelentés valóban egy hipotézis, hiszen a tudomány eszközeivel könnyen megcáfolható. A hipotézis hamisnak bizonyul, tudniillik létezik olyan vegyület, ami nátriumiont tartalmaz, és nem fehér, például a sárga színű nátrium-kromát (Na_2CrO_4).

1.1. táblázat. *Az aktív szemlélődéstől a hipotézis megfogalmazásáig*

Tények megfigyelése	Tények összegzése	Hipotézis
A falevél a földre hull.	A testek szabadesésben a földfelszínre kerülnek.	A testek halmazállapottól függetlenül, de tömegüktől függő sebességgel szabadesésben a földfelszínre kerülnek.
Az esőcsepp a földre hull.		
A tányér a kezemből a földre esik és eltörik.	Az esés halmazállapottól független.	
A toll a földre hull.	Az esés sebessége a tömegtől függ.	
A porszem a földre hull.		

Az 1.1. táblázatban említett példa általánosított formája a Newton-féle gravitációs elméletnek, amit a mai napig még nem sikerült megcáfolni. Innen következik az elméleteknek és a hipotéziseknek egy másik fontos tulajdonsága, a bizonyosság. Minden elmélet/hipotézis valamilyen mértékű bizonyossággal rendelkezik a biztosan hamis és a biztosan igaz között.

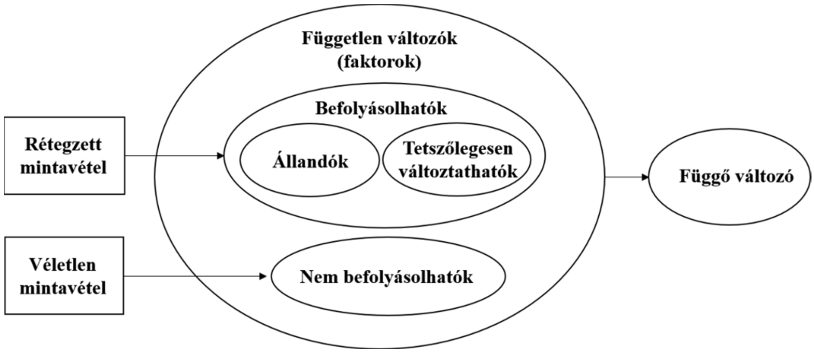
1.1. A tudományos kutatás eszközei

A tudományos kutatás eszközei a megfigyelés és a kísérlet. A megfigyelés során természetes kapcsolatokat állapíthatunk meg, de az egymással erősen korreláló változók között nem bizonyítható ok-okozati összefüggés. A kísérlet lehetővé teszi az ok-okozati kapcsolatok felderítését.

Kutatásunk során megfigyelhetjük egy populáció viselkedését a maga természetes környezetében. Ilyenkor adatokat gyűjtünk egy vagy több mérhető paraméterről, majd az adatsorokat egy középértékkel és egy szóródási mutatóval jellemezzük. Például több mezőgazdasági területen alkalmazva a városi szennyvíziszapot megfigyelték, hogy a kukorica termésében több fém mennyisége megnőtt, köztük említésre méltó a kadmium (Wang et al., 2017). A populációt minden esetben a kukorica egyedei képezték, a vizsgált paraméter pedig a kukorica termésének a kadmiumtartalma volt. A kutatások eredményeként megállapíthatjuk, hogy a szennyvíziszap feltehetően megnöveli a kukorica termésének a kadmiumtartalmát. Mivel a kutatást különböző mezőgazdasági területen végezték, így nem tudjuk megmondani, hogy a megemelkedett kadmiumtartalom nem más tényezőnek tulajdonítható-e, például a talaj tulajdonságainak, a kukoricafajtának vagy az öntözővíznek stb. Ha azt gyanítjuk, hogy a teszterületen a szennyvíziszap miatt növekedett meg a kukoricában a kadmiumtartalom, akkor kísérletet dolgozhatunk ki a hatás bizonyítására.

A tudományos kísérlet során egyetlen tényező hatását vizsgáljuk úgy, hogy az összes többi tényezőt állandó értéken tartjuk. A cél az, hogy megállapíthassuk, hogy vajon a tesztelendő paraméter hatással van-e a populáció valamely tulajdonságára. A vizsgált tényezőt faktornak nevezzük, a populáció által adott válasz pedig egy mérhető, statisztikailag kiértékelhető paraméter kell legyen, amit függő változónak nevezünk. Az előző példánál maradva, a kukoricát ellenőrzött körülmények között termesztjük. A kadmiumkoncentrációt befolyásoló tényezőket (kukoricafajta, talajtípus, öntözővíz, trágya/műtrágya) állandó értéken tartjuk az összes parcellán, majd a különböző parcellákon tetszőlegesen, de eltérő mértékben alkalmazzuk a városi szennyvíziszapot. Amennyiben a kukorica termésének a kadmiumtartalma és a szennyvíziszap mennyisége között pozitív kapcsolatot állapítunk meg, kijelenthetjük, hogy a szennyvíziszap növeli a kukorica termésének a kadmiumtartalmát.

A megfigyelés és a kísérlet között lényeges különbség van (1.1. ábra). A megfigyelés során nem tudunk minden tényezőt ellenőrzés alatt tartani, ezért nem tudjuk biztosítani a tesztelt paraméter (faktor) kizárólagos hatását a rendszer által szolgáltatott válaszra (függő változóra). Mint megfigyelők, nem hatunk semmilyen tényezőre (faktorra), ezeknek együttes hatását vizsgáljuk a populációra. A kísérlet azt feltételezi, hogy a faktorok hatását állandó értéken tartjuk, és kizárólag a tesztelt faktor értékét változtatjuk. Így a rendszer által adott választ – ideális esetben – kizárólag a tesztelt faktor váltja ki. A kísérleteket általában laboratóriumokban, üvegházakban vagy ugyanazon terület különböző parcelláin végzik, kijelölve egy kontroll- vagy referenciacsoportot/területet is.



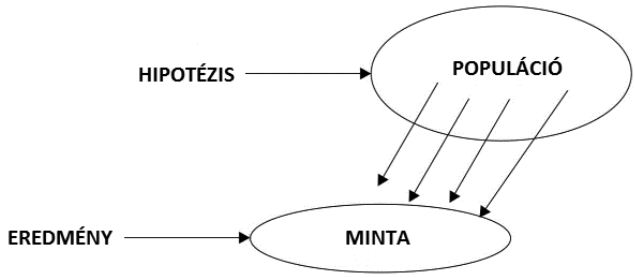
1.1. ábra. *A tudományos kísérlet és megfigyelés gondolati felépítése*

A kutató szempontjából a faktorok két csoportba sorolhatók: befolyásolhatók és nem befolyásolhatók. A befolyásolható faktorok közül általában egynek a hatását szeretnénk vizsgálni a célcsoporton, ezért a többit állandó értéken tartjuk. A nem befolyásolható faktorok hatását a célcsoport válaszára legfeljebb minimalizálni tudjuk, amit a véletlen mintavétellel kivitelezhetünk. Vannak olyan faktorok, amelyek

jól meghatározott értékeket vehetnek fel, vagy jól elhatárolható kategóriákat képezhetünk az értékeik alapján. Ilyen esetben a célcsoportot alcsoportokra oszthatjuk az illető faktor szerint, a módszert pedig rétegzett mintavételnek nevezzük. Ha a városokban szeretnénk felmérni a szálló por mennyiségét, akkor a városokból alcsoportokat képezhetünk a lakosság száma vagy a közúti forgalom nagysága alapján. Ha például a folyók nitráttartalmát szeretnénk felmérni, akkor a folyókat szakaszokra bonthatjuk (felső, középső, alsó szakasz). Egy ökológiai vizsgálatban például külön vizsgálhatjuk a nőstények és a hímek viselkedését adott környezeti tényezőre, a mért paraméter értékeit pedig külön-külön adjuk meg a rétegekre (alcsoportokra).

1.2. A véletlen mintavétel

A tudományos vizsgálat során egy hipotézist tesztelünk a populáción (1.2. ábra). Általában a kutatást nem tudjuk elvégezni az alapsokaság (populáció) minden egyes egyedén, mert sok esetben az alapsokaság végtelen számú elemből áll (például egy tó víztömege, talaj, levegő stb.). Ilyen esetben az alapsokaságból egy véges elemszámú mintát veszünk, és az összes mérést, kísérletet a mintán végezzük el. Nyilvánvalóan az eredményt is a mintára kapjuk, az eredményt pedig általánosítjuk az alapsokaságra. Ezt akkor és csakis akkor tehetjük meg, ha a minta reprezentatív, azaz képviseli az alapsokaságot. Abban az esetben, ha a célcsoport homogénnek tekinthető a vizsgált paraméter szempontjából, a minta reprezentativitását a véletlen mintavétellel biztosíthatjuk.

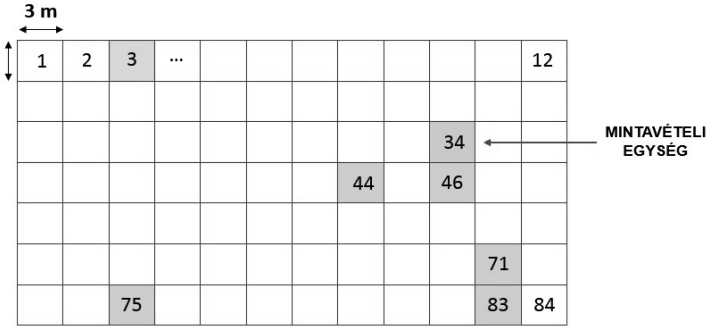


1.2. ábra. *A populáció és a minta kapcsolata*

A véletlen mintavétel lényege az, hogy a célcsoport minden egyedének/elemének azonos esélye van bekerülni a mintába, és ezek az egyedek függetlenek egymástól. Ha a célcsoport véges számú egyedből áll (N), akkor minden egyednek 1/N valószínűsége van bekerülni a mintába. Ha a célcsoport végtelen számú egyedből áll, az egységet a minta egy elemébe kerülő mennyiség szabja meg (például ha egy elem a mintában 1 liter vizet jelent, akkor a célcsoport az 1 literek összességét

jelenti; ha egy minta 2 m^3 levegőt jelent, akkor a célcsoport a 2 m^3 levegőtérfogatok összessége). A minta elemszáma (n) mindig véges. Ha az N értéke viszonylag kicsi (pl. pár száz), akkor az egyedek egymástól való függetlenségét úgy biztosítjuk, hogy a minta elemszáma nem haladja meg az N 10%-át.

A véletlen mintavétel kivitelezése egy véletlenszám generátor segítségével történik. A célcsoport egyedeit megszámozzuk 1 és N között, majd meghatározzuk a minta elemszámát. Az n értékeit véletlenszám-generátorral választjuk ki. Például az 1.3. ábrán egy mezőgazdasági területről szeretnénk 7 mintát begyűjteni. Megszerkesztünk egy négyzethálót, amely 84 kvadrátot tartalmaz (a kvadrátok élhosszúsága 3 m). Az $n < 0.1N$ feltétel teljesül (az N 10%-a 8.4). Megszámozzuk a kvadrátokat, majd véletlenszám-generátor segítségével – amit lekérhetünk az internetről (*https://www.random.org/integers/*), vagy futtathatunk hozzá egy telefonos alkalmazást (pl. Random Number Generator) – kiválasztjuk a 7 egységet: pl. 3, 34, 44, 46, 71, 75, 83. A mintavételi egységeket a földrajzi koordinátákkal is meg lehet jelölni. Ez a módszer a terepi munkában nagyon hasznos lehet.



1.3. ábra. *Véletlen mintavétel kivitelezése mintavételi háló segítségével*

A véletlen mintavétel annál torzítatlanabb becslést ad a populáció középértékére, minél homogénebb a populáció a mért paraméter szempontjából. Amint már szó esett róla, a homogenitást növelni lehet a rétegzett mintavétellel, a becslés torzítatlanságát pedig a minta elemszámának a növelésével.

1.3. A változók típusai

A tudományos kutatások a populáció által szolgáltatott paraméterek mérésén alapulnak, ezeket a paramétereket függő változóknak vagy egyszerűen változóknak nevezzük. A változók értékei különféle skálán értelmezettek, amelyek alapvetően meghatározzák az alkalmazható statisztikát is. Négy kategóriát különböztetünk meg:

nominális (nevesítő), ordinális (sorrendi) } minőségi változók,

intervallumskálájú, arányskálájú } mennyiségi változók.

Az alábbi táblázatban (1.2. táblázat) néhány székelyföldi ásványvízforrás fizikai és kémiai paraméterei közül figyelhetünk meg néhányat.

1.2. táblázat. *Néhány székelyföldi ásványvízforrás fizikai és kémiai adatai*

ID	Helység	EC1	EC2	pH	Ca	Fe	Kalciumos
AV2	Korond	2.47	közepes	6.09	433	7.77	1
AV8	Homoródfürdő	2.74	nagy	6.11	155	12.6	1
AV10	Szentegyháza	2.40	közepes	6.04	74	8.55	0
AV13	Csíksomlyó	2.85	nagy	6.30	76	9.30	0
AV17	Szeltersz	5.92	nagy	6.59	138	13.9	0
AV20	Kirulyfürdő	1.31	közepes	5.91	105	7.27	0
AV22	Tusnád	1.52	közepes	6.02	139	13.4	0
AV26	Bibarcfalva	2.50	közepes	6.25	239	5.03	1
AV28	Magyarhermány	0.92	kicsi	5.78	86	14.7	0
AV32	Zsögödfürdő	1.74	közepes	6.03	98	12.6	0

A táblázatban szándékosan nincsenek feltüntetve alapvető információk, mint például a mértékegység vagy az, hogy mit fejeznek ki az adatok (átlagot, mediánt, egyedi mérési eredményt). Ez a forma egyezik azzal, amit egy statisztikai programba olvashatunk be. Az adatok megértése végett elengedhetetlen egy magyarázó vagy leíró dokumentum (metaadat) társítása az adattáblához. Ez a dokumentum tartalmazza a kutatóintézet és a kutatók neveit, az adatgyűjtés időpontját, helyszínét, a változók megnevezését, mértékegységét, kimutatási határait. Ezenkívül bármilyen hasznos információt fel lehet tüntetni, ami könnyíti az adatok megértését.

Az 1.2. táblázat adataihoz a következő információk társíthatók:

– a vizsgálatot 2019–2021-ben végezték a Sapientia EMTE oktatói és hallgatói egy pályázat keretében, a négy évszaknak megfelelően;
– a kutatás célja az volt, hogy felmérjék a széles körben fogyasztott természetes ásványvízforrások fizikai és kémiai minőségét;
– az adatok négy mérési alkalom átlagértékeit adják meg;
– a változók mértékegységei:
  ▪ EC1 (elektromos vezetőképesség): mS/cm;

- EC2 (elektromos vezetőképesség): kicsi (EC1 ≤ 1.00), közepes (1.00 < EC1 ≤ 2.50), nagy (EC1 > 2.50);
- pH (kémhatás, ha pH < 7, a víz savas);
- Fe (vas): mg/l;
- Ca (kalcium): 1 (ha Ca ≥ 150 mg/l), 0 (ha Ca < 150 mg/l);

Egyéb információk: a Ca értékei azt jelölik, hogy a HG1020/2005 minisztériumi rendelet szerint az adott ásványvíz tekinthető-e kalciumos víznek (ebben az esetben az érték 1).

1.3.1. A nominális (nevesítő) változó

A nominális változó megnevez, kategorizál, csoportba sorol, kódol egy mintát vagy objektumot. Egy nominális változónak lehetnek számértékei is, de a számoknak nincs nagyság szerinti sorrendje. Nem lehet számolni velük. Például jelölhetnek gyártási sorozatszámokat, mintaszámokat. Az 1.2. táblázatban az ID, a Helység, Kalciumos változók tartoznak ebbe a kategóriába.

A nominális változókkal semmilyen matematikai művelet nem végezhető el, viszont meg lehet számolni, hogy egy adott név hányszor fordul elő, illetve hány százalékot tesz ki az összmintaszámra/csoporton belüli mintaszámra vonatkoztatva. Ilyen megközelítésben a nominális változók kategóriákat határozhatnak meg, feltéve, ha a készletükben levő nevek több mintánál vagy objektumnál előfordulnak. Ilyen kategorizáló változó a Ca. A két érték alapján, amit a változó felvehet (1 és 0), két csoportba sorolhatjuk a mintákat: kalciumos és kalciumban szegény vizek. A nyolc minta közül három kalciumos (37,5%) és öt kalciumban szegény (62,5%). Az olyan változókat, amelyek csak és kizárólag két értéket vehetnek fel, amelyek kölcsönösen kizárják egymást: 1 vagy 0, igen vagy nem, jelenlét vagy hiány, *bináris változó*knak nevezzük.

A nominális változók készletében szereplő kategóriákra abszolút vagy relatív gyakoriság adható meg, illetve meghatározható a leggyakoribb elem, a módusz. A statisztikai adatfeldolgozásban ezekkel a paraméterekkel jellemezhetjük őket.

1.3.2. Az ordinális (sorrendi) változó

Az ordinális változó megnevez, kategorizál, csoportba sorol, kódol egy mintát vagy objektumot, de értékeinek egyértelmű sorrendje van. Egy nominális változónak az értékei lehetnek számok is, és a számok sorba rendezhetők. Ilyen esetben a számok egy skálát képviselnek két végpont között. A végpontok ellentétes véleményt, minimális és maximális erősséget, mennyiséget, érzelmi hozzáállást stb. képviselnek. Az ordinális változónak névkészlete is lehet, a nevek meghatározott

sorrendbe tehetők (pl. kicsi, közepes, nagy; egyáltalán, inkább nem, közömbös, inkább igen, biztosan igen). Az 1.2. táblázatban ordinális változónak minősül az EC2, amelynek a névkészlete: kicsi, közepes, nagy. A nominális változóhoz hasonlóan az ordinális változóra meg lehet számolni, hogy adott név/szám hányszor fordul elő, illetve hány százalékot tesz ki az összmintaszámra/csoporton belüli mintaszámra vonatkoztatva. A statisztikában gyakorisággal és kumulatív gyakorisággal jellemezhetjük őket, a módusz mellett már a medián (helyzeti középérték) is meghatározható. Az EC2 változó „kicsi" értéke egyszer fordul elő a 10 mintában, tehát gyakorisága 1 (10%), a „közepes" hatszor, tehát gyakorisága 6 (60%), a „nagy" háromszor, tehát gyakorisága 3 (30%). A leggyakoribb érték (módusz) és a medián értéke egyaránt a „közepes". A kumulatív gyakoriságot úgy értelmezzük, hogy a sorba rendezett kategóriák gyakoriságait rendre hozzáadjuk az előző kategóriákhoz. Az EC2 változónál: a „kicsi" 10%, a „kicsi" és a „közepes" 70%, a „kicsi", a „közepes" és a „nagy" pedig együtt 100%.

1.3.3. Az intervallumskálájú változó

Az intervallumskálájú változónak numerikus értékskálája van, amin a nulla érték nem abszolút. Ez azt jelenti, hogy a nulla nem a paraméter vagy tulajdonság hiányát jelzi, hanem valós számértéknek minősül. Az intervallumskálájú változók általában származtatott skálával rendelkeznek, ilyen a hőmérséklet Celsius-skálája vagy a Fahrenheit-skála, a kémhatásnak a pH-skálája, a zajszintnek a dB- (decibel-) skálája. A hőmérséklet abszolút skálája a Kelvin-skála, az Anders Celsius által készített skála viszont egy relatív skála, amit Európában használnak. Az utóbbi skálán a nulla fok egy reális hőmérsékleti érték, ami a víz fagypontját jelöli 1 atm nyomáson. Hasonló módon a pH-skála a hidrogénionok koncentrációjából származtatott logaritmikus skála, amelyen a nulla érték (1 mol/l H^+ koncentráció) egy erősen savas oldat kémhatását jelöli. A származtatott skálákon értelmezhető az összeadás, kivonás és szorzás mint matematikai művelet, de nem értelmezhető az osztás. Az intervallumskálájú változók általában negatív értékeket is felvehetnek. Az 1.2. táblázatban intervallumskálájú változó a pH. Értéke 4.94 és 6.30 között változik.

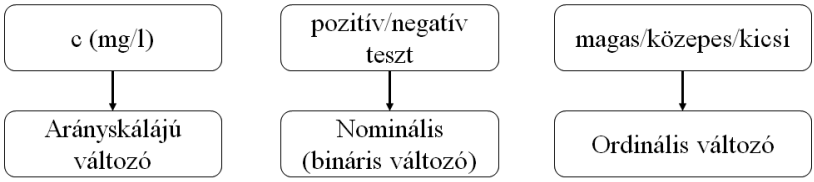
1.3.4. Az arányskálájú változó

Az arányskálájú változó numerikus skáláján a nulla az abszolút nulla. A változó ezt az értéket megközelítheti, de soha nem éri el. Ilyen változók a fizikai paraméterek többsége (hosszúság, térfogat, sebesség, fényintenzitás, nyomás, elektromos vezetőképesség, a hőmérséklet Kelvin-skálája stb.) és a kémiai paraméterek többsége (koncentráció, ionmozgékonyság, reakciósebesség), amelyeknek a mértékegysége azonos a nemzetközi mértékegységgel. Arányskálájú változókkal bármilyen matematikai

művelet elvégezhető. Az 1.2. táblázatban három arányskálájú változó szerepel: EC1, Ca, Fe. Azonos fizikai vagy kémiai paraméter változói (pl. különböző elemek koncentrációi) esetén minden elemre azonos mértékegységet célszerű használni, akkor is, ha nagyságrendbeli különbségek vannak. Az 1.2. táblázatban a Ca és a Fe mg/l-ben vannak megadva, annak ellenére, hogy a Ca a százas, a Fe a tizes skálán mozog.

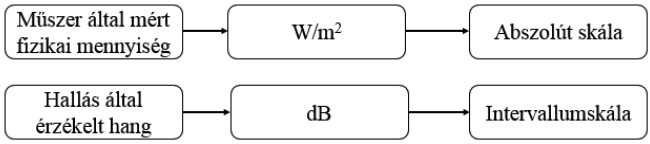
1.3.5. Sajátos esetek

Egy változónak nem feltétlenül adott a skálája (értékei különböző skálákon mozoghatnak). Észszerű szempontok szerint eldönthetjük, hogy melyik skálával dolgozunk. Például szükséges-e egy összetevő koncentrációját ismerni, vagy csak a jelenlétét/hiányát szeretnénk tudni, vagy arra vagyunk kíváncsiak, hogy az értékek hogyan viszonyulnak egy határértékhez. Például a vér magnéziumtartalma megadható háromféle módon (1.4. ábra). A legkimerítőbb információhordozó az arányskálájú változó, amiből könnyen származtatható egy bináris vagy egy ordinális változó, azonban számolnunk kell az információvesztéssel. Az ordinális minőségi skála akkor használható, ha az adott arányskálájú változó alapján csoportosítani szeretnénk a mintákat; a mennyiségi változóból így egy minőségi változót (faktort) hozunk létre. Visszafelé a konverzió nem működik, műtermékek, hamis eredmények fognak keletkezni.



1.4. ábra. *Egy véranalízis eredményének lehetséges közlési módjai*

A műszerek általában arányskálájú paraméterek nagyságát (intenzitását) mérik, viszont bizonyos esetben indokolt lehet az eredményeket intervallumskálán megadni (1.5. ábra). Például a zajt a műszer a fizikai inger szintjén méri (W/m^2), de a közérthetőség végett az értékeket átalakíthatjuk az emberi érzékelés szempontjából sokkal könnyebben értelmezhető (hallás által érzékelt különbségek) decibel értékeivé.



1.5. ábra. *A zaj értékei kétféle skálán adhatók meg*

Előfordulhat olyan sajátos eset, amikor a vizsgált populáció által adott választ képviselő változó nehezen mérhető, követhető vagy egyáltalán nem mérhető. Ilyen esetben kiválaszthatunk egy olyan változót, amiről tudjuk, hogy szorosan együtt változik a számunkra fontos változóval, és ami könnyen mérhető vagy követhető. Az ilyen segítő változókat helyettesítőknek (*proxy*knak) nevezi a szakirodalom. Például klímarekonstrukció esetén lehetetlen a levegő hőmérsékletének a változását évezredekre visszamenően megmérni. Tavak üledékében viszont meghatározhatók a régióban élt hideg/meleg, száraz/nedves éghajlatot kedvelő növények pollenjei, amiből következtetni lehet az éghajlatra. A vizekben oldott összsótartalom meghatározása hosszú és költséges munka lehet, mivel minden kation és anion mennyiségét meg kellene határozni. Megközelítő és gyors eredményt adhat a vizek elektromos vezetőképességének a megmérése a mintavétel helyszínén, amiből következtetni lehet a kívánt paraméterre.

2. AZ R STATISZTIKAI PROGRAM ALAPJAI

Az R egy szabadon hozzáférhető és használható statisztikai szoftvercsomag, tulajdonképpen egy programozási környezet, egy programnyelv. Nevét két újzélandi alkotójáról kapta, akiknek történetesen R-betűvel kezdődik a keresztnevük: Ross Ihaka és Robert Gentleman. Ők 1993-ban tették publikussá a szoftvert, amelyet 1995-ben átengedtek a GNU, General Public License ingyenes licencek sorozatába. Azóta a programot egy központi csapat (R Development Core Team) kezeli.

A hozzáférhetősége mellett a szoftver rendkívül széles körű lehetőséget nyújt a legkülönfélébb statisztikai eszköztár használatához a programkönyvtár által. Ez a könyvtár folyamatosan bővül új csomagok formájában, amelyek ingyen telepíthetők és futtathatók R-ben. A használati utasítások .html vagy .pdf formátumban tanulmányozhatók. Bárki létrehozhat új csomagokat, a szerzők nevei pedig megjelennek az adott csomag használati utasításában. A csomagok az Átfogó R Archívum Hálózaton (CRAN – Comprehensive R Archive Network) találhatók meg. A csomagokról szóló részletes tájékoztatók mellett számos internetes oldal nyújt segítséget a program használatával kapcsolatban, akár sajátos problémák megoldására is. Elterjedten használnak különféle társalgási oldalakat (bloggokat), ahol kérdéseket lehet feltenni. Néhány ajánlott oldal:

https://www.tutorialspoint.com/r/index.htm (az R saját kézikönyve);
https://support.rstudio.com/hc/en-us (az RStudio használati utasítása);
https://www3.nd.edu/~steve/computing_with_data_2014/ (Steven Buechler egyetemi kurzusa az R alapműveleteivel kapcsolatban);
https://rcompanion.org/handbook/ (Salvatore S. Mangiafico ingyenes internetes segédanyaga egyetemi hallgatók részére).

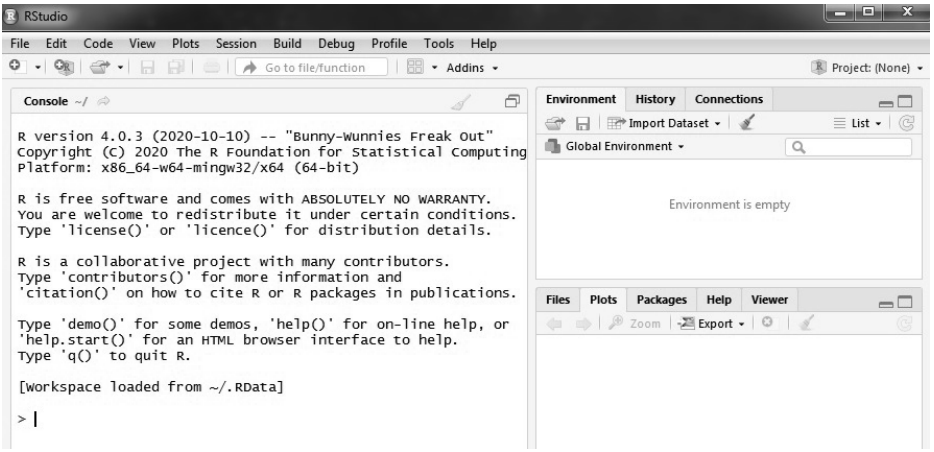
2009-ben létrehozták az *R Journal* (ISSN: 2073-4859) tudományos folyóiratot, amelynek cikkei szabadon elérhetők, és megjelentetés előtt független szakemberek általi bíráláton (peer review) esnek át. A folyóirat tematikájában szerepel az R-ben futtatható statisztikai adatelemzések széles körű alkalmazhatóságának bemutatása vagy új parancssorozatok (kódok) ismertetése az olvasó által reprodukálható formában. A folyóirat honlapja: *https://journal.r-project.org*.

Az R alapvetően parancssoros, de az operációs rendszereknek megfelelően biztosítanak egy alapfelületet (interface) is, amely gyakorlatilag a parancsokat futtató konzol (R Commander Widowson, R.app MacOSX-en, linuxon terminálban fut) és néhány kényelmi funkciót adó menüpont. Ezenkívül grafikus felhasználói felület (GUI: R Commander (Rcmdr), Deducer, jamovi, JASP, JGR, R-Instat, rattle, RKWard stb.), illetve fejlesztői környezet (IDE: RStudio) is rendelkezésünkre áll. Ezek között

vannak, amelyek az R környezeten belülről indíthatók (Rcmdr, Deducer, rattle, JGR), míg mások önállóan telepíthető környezetet biztosítnak (jamovi, JASP, RKWard, R-Instat). Az RStudio előnye a könnyű kezelhetőség, a szolgáltatások sokfélesége; például az, hogy beírás közben mutatja az alkalmazott parancs szintaxisát, jelzi az elírásokat, felajánlja a lehetőségeket, az ábrák könnyen méretezhetők, elmenthetők stb.

A felhasználó az R-rel egy programnyelv segítségével kommunikál, amelynek a megtanulása időigényes, de elsajátítható előzetes programozási ismeretek nélkül is. A párbeszéd az RStudióban a *Console* felületen zajlik parancsok formájában, (a konzol megegyezik az alap R konzollal). A parancsok együttesét, ami egy több lépésből álló műveletet hajt végre, kódnak (script) nevezzük. Maga az R program a *https://cran.r-project.org* internetes oldalról tölthető le, és futtatható Windows, Linux és MacOS rendszerben. Az RStudio a *https://www.rstudio.com/products/rstudio/* oldalról tölthető le, de Windows-rendszer alatt a használatához szükség van az R előzetes telepítéséhez.

A továbbiakban bemutatott parancsok, illetve kódok az RStudióban készültek (hasonló módon működnek minden R környezetben), amely az Rx64 4.0.5 verzió alatt fut.



2.1. ábra. *A felhasználófelület az RStudióban*

A felhasználófelületen (2.1. ábra) három ablakot látunk. Az első a *Console*, ahol a párbeszéd folyik az R-rel. Ez a futtató környezet, itt futnak a parancsok, ide küldjük a kódokat, és itt közli velünk az R a parancsok kimeneteit, a hibaüzeneteket, illetve egyéb hasznos információkat. A felső jobb ablaknak három füle van: *Environment*, *History* és *Connections*. Az Environment fülben a Global Environment folyamatosan megőrzi a Console-ban létrehozott adattárolókat (vektorokat, faktorokat, listákat, adattáblákat stb.), a History ennél többet tesz, minden műveletet elment 500 sorig. A Connections különféle oldalakkal köti össze a Console-t.

Bővebben a *https://db.rstudio.com/advanced/snippets/* oldalon lehet erről olvasni. A jobb oldali alsó ablakban lehetőség van a saját gépen levő mappákat behozni (Files), a Console-ban elkészített ábrákat megtekinteni és elmenteni (Plots), a letöltött csomagokat listázni (Packages), illetve a Console-ban feltett kérdésekre bővebb választ kapni (Help).

Az R egy saját munkamappával (*working directory*) van összeköttetésben, innen tölt be, illetve ide ment ki mindent a futtató környezetből. Ezt a mappát tetszőlegesen beállíthatjuk, és ez a beállítás mindenképpen ajánlott. Kétféleképpen járhatunk el: a mappát beállítjuk a menüben, vagy egy parancskódot használunk (2.1. táblázat). A **`getwd()`** utasítással megkapjuk az aktuális munkamappa helyét a gépen. A **`setwd()`** paranccsal megadjuk a munkakönyvtárunk helyét (a könyvtár teljes elérési útvonalát kell megadni).

2.1. táblázat. *A munkamappa (working directory) helyének megadása*

Szoftver	Beállítás	Kód
R	File → Change dir…	**`setwd()`**
RStudio	Session → Set Working Directory → Choose Directory	például: **`setwd("C:/Documents and Settings/Data")`**

A Console-ba beírt parancsok és kódok csak pillanatnyi, „élő beszélgetésnek” számítanak, ahhoz, hogy elmenthessük őket, meg kell nyitnunk egy munkafüzetet, ami kizárólag az R programban értelmezhető, kiterjesztése pedig „.R”. Ez a munkafüzet tetszőlegesen elnevezhető, véglegesítés után pedig el kell menteni (2.2. táblázat).

2.2. táblázat. *A munkafüzet (R Script) létrehozása és elmentése*

Szoftver	Beállítás
R	File → New script File → Save (amikor aktív a Script ablak)
RStudio	File → New File → R Script Save (ikon, a Script ablakon belül)

Jól bevált munkamenet R-ben, hogy az utasításokat a Scriptbe írjuk, majd átküldjük a Console-ba. Ezt az R Commander felületén a CTRL+R, az RStudio felületén pedig a CTRL+ENTER billentyűpárossal tehetjük meg.

2.1. Az R nyelvezetének alapjai

Az R megkülönbözteti a nagy és kis betűket. A parancsok írásánál erre különös figyelmet kell fordítani. Minden parancs egy kifejezésből áll, amit kerek zárójelekbe írt információk (adatok, illetve opcionális beállítások) követnek. A kapcsos zárójelet több parancs összefűzésekor, illetve programozáskor a ciklusok szervezésénél használjuk. A szögletes zárójelet az adatok (vektorok, mátrixok) elemeire történő hivatkozáskor, illetve a részhalmazok kijelölésére használjuk. Szavakat, kifejezéseket idézőjelek közé írunk, a számokat pedig egyszerűen vezetjük be. Továbbá betehetők olyan szöveges részek, amelyeket az R nem vesz figyelembe, ezeket a # előjellel látjuk el. Mindent, ami a # jel után következik egy sorban, az R nem hajtja végre. Ennek akkor van haszna, ha a Scriptben emlékeztetőket, magyarázatokat szeretnénk megőrizni egy parancs vagy egy kód mellett.

Az adattárolókban előfordulhat, hogy hiányzik egy vagy több adat. Ezeknek a helyén az R-ben az `NA` jelenik meg, ami a *not available* kifejezésből származik. Lehetetlen számok (pl. nullával való osztáskor) helyén a `NaN` szerepel, a *not a number* rövidítéseként.

2.3. táblázat. *Számtani és logikai műveletek R-ben*

Számtani műveletek				Logikai műveletek	
Jel	Művelet	Jel	Művelet	Jel	Művelet
`+`	összeadás	`round()`	kerekítés	`&`	és
`-`	kivonás	`sum()`	összeadja egy adatsor értékeit	`|`	vagy
`*`	szorzás	`mean()`	átlagolás	`!`	nem
`/`	osztás	`%/%`	maradék nélküli osztás	`==`	egyenlő
`^ vagy **`	hatványozás	`:`	sorozatot készít a két szám között	`!=`	nem egyenlő
`sqrt()`	gyökvonás	`min()`	minimum	`<`	kisebb
`log10()`	tízes alapú logaritmus	`max()`	maximum	`>`	nagyobb
`log()`	természetes alapú logaritmus	`floor()`	legközelebbi egész szám, nem nagyobb, mint x	`<=`	kisebb vagy egyenlő
`exp()`	exponenciális (e^x)	`celeing()`	legközelebbi egész szám, nem kisebb, mint x	`>=`	nagyobb vagy egyenlő
`abs()`	abszolút érték	`sin()` `cos()` `tan()`	trigonometria	`%in%`	egy elem benne van-e egy vektorban

Az R-ben érvényesek az egyszerű számtani és logikai műveletek, amelyeknek az írásjeleit a 2.3. táblázat tartalmazza. A számtani műveleteknél az R betartja a számtani műveletek sorrendjét: hatványozás, szorzás/osztás és összeadás/kivonás. A sorrend kerek zárójelekkel tetszőlegesen változtatható. A logikai műveletek logikai eredményt adnak, amelynek a lehetséges formái R-ben az igaz (TRUE) és a hamis (FALSE). Az R a logikai műveletek eredményeit csak ebben a formában értelmezi (nem helyettesíthetők például nullával és egyessel).

Az R-ben az adatok többféle módon tárolhatók: vektorban (vector), faktorban (factor), listában (list), mátrixban (matrix) vagy adattáblában (data.frame) (2.4. táblázat). Ezek a formák tetszőleges nevet kaphatnak, a megfeleltetés pedig az „=" vagy a „<–" írásjellel valósítható meg.

```
x = 5
y = 3
z = x + y
z
[1] 8

x > y
[1] TRUE

(x+z)**3-y
[1] 2194
```

Az elmentett x, y, z megjelenik a Global Environment ablakban (a munkafelület jobb felső részében). A bonyolultabb műveleteknél a műveleti sorrendet a kerek zárójelek használatával határozhatjuk meg.

```
((x+3*z)^0.5-(y-2*x))
[1] 12.4
```

A *vektor* egyetlen típusú változót tartalmazó adatsor. A típus lehet szám vagy karakter. A *faktor* hasonló a vektorhoz, azonban csak karaktereket tartalmazhat (a számokat is karaktereknek értelmezi), komponenseit rangsorolni lehet a `levels` paranccsal (lásd az 2.4. táblázatban a faktor2 példát). Ha a `levels` opciót nem használjuk, az R ábécésorrendbe teszi a szavakat, kifejezéseket (például ábrákon). Ha számok rangsorolásánál az `ordered` logikai opciót `TRUE`-ként adjuk meg, akkor a program a növekvő sorrend szerint rangsorolja őket (lásd az 2.4. táblázatban a faktor3 példát). A rangsor hasznos lehet a későbbi adatfeldolgozásnál (például a lineáris modellek alkalmazásánál). Több numerikus változó adatai *mátrix*ban vagy *adattábl*ában tárolhatók. A *lista* bármilyen adattípust tud tárolni.

2.4. táblázat. *Az adatok tárolása*

Objektum	Jellemzés	Példa
vektor	Egyetlen típusú változót tartalmaz. **c()**	**obj1 = c(5, 3, 7, 8, 2)** **obj2 = c(TRUE, TRUE, FALSE, TRUE, TRUE)** **obj3 = c("blue", "red", "green", "blue", "green")**
faktor	Csak karaktereket tartalmaz. **factor(c())** **factor(c(), levels = c())** **factor(c(), ordered = TRUE)**	**faktor1 = factor(c("M", "F", "F", "F", "M"))** **faktor2 = factor(c("Igen", "Igen", "Nem", "Igen", "Nem"), levels = c("Nem", "Igen"))** **faktor3 = factor(c(3,5,5,4,1,3,2), ordered = TRUE)**

mátrix	Csak egyféle típusú adatot tartalmazhat.
	matrix()
	A változók adatait egy vektorban adjuk meg, majd megadjuk, hogy hány sorba és oszlopba rendezze őket. A **byrow** a mátrix kitöltési sorrendjét adja meg, értéke alapból F (FALSE), vagyis oszloponként tölti fel az adatokat.
	A **dimnames = list()** paranccsal elnevezhetjük a sorokat és oszlopokat. Ezt megtehetjük a **colnames()** és **rownames()** parancsokkal is.
		m = matrix(c(3,2,6,4,87,34,23,94), nrow = 4, ncol = 2,byrow = F, dimnames = list(c("S1","S2","S3", "S4"), c("VAR1", "VAR2")))
		m
		`   VAR1 VAR2`
		`S1    3   87`
		`S2    2   34`
		`S3    6   23`
		`S4    4   94`
		colnames(m) = c("kontroll", "kezelt")
		rownames(m) = 1:4

lista	Többféle adattárolót tartalmazhat nem táblázatos formában, eltérő dimenziók is lehetségesek. **list()**	**lista = list(c(45,44,46), matrix(1:12, nrow = 4), list("kutya",FALSE, "tyúk",2,0,9), TRUE)** **paste(lista)**

adattábla	Vektorokat és faktorokat tartalmazhat táblázatos formában. Az adattárolók neveit (obj1-obj3) nem tesszük idézőjelbe.	**df = data.frame(obj1,obj2,obj3)**
		df
		`   obj1   obj2  obj3`
		`1     5   TRUE  blue`
		`2     3   TRUE  red`
		`3     7   FALSE green`
		`4     8   TRUE  blue`
		`5     2   TRUE  green`

Egy adattároló típusát a **class()** paranccsal kérhetjük ki (például **class(obj1)**).

```
class(obj1)
[1] "numeric"

class(lista)
[1] "list"
```

A számtani műveletek a számadatokat tartalmazó adattárolókra (vector, matrix, data.frame) is alkalmazhatók. Az R az adattároló összes elemére elvégzi a kijelölt műveletet. Az amerikai társadalom a testmagasságot lábban (ft, feet) méri. A méterbe való átszámítási képlet: feet/3.28008.

```
feet = c(4,4.5,5,5.5,6,6.5)
meter = feet/3.2808
meter
[1] 1.219215 1.371617 1.524019 1.676420 1.828822 1.981224

round(meter,2)
[1] 1.22 1.37 1.52 1.68 1.83 1.98

centimeter = feet*100/3.2808 #a cm-ben megadott magasság kö-
zelebb áll hozzánk

round(centimeter,1)
[1] 121.9 137.2 152.4 167.6 182.9 198.1
```

A listában szereplő adattárolóknak tetszőleges neveket lehet adni. Átnevezzük a 2.4. táblázatban feltüntetett lista adattárolóit:

```
names(lista) = c("Életkor", "Mátrix", "Állatok/db","Családos")
lista
$Életkor
[1] 45 44 46

$Mátrix
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
$`Állatok/db`
$`Állatok/db`[[1]]
[1] "kutya"
```

```
$`Állatok/db`[[2]]
[1] FALSE

$`Állatok/db`[[3]]
[1] "tyúk"

$`Állatok/db`[[4]]
[1] 2

$`Állatok/db`[[5]]
[1] 0

$`Állatok/db`[[6]]
[1] 9

$Családos
[1] TRUE
```

A vektorban tárolt számadatokra értelmezhetők a **sort()**, az **order()** és a **length()** parancsok is. Az első növekvő sorrendbe rendezi az értékeket, a második pedig megadja a növekvő sorba rendezett adatok helyét a vektorban. A csökkenő sorrendet úgy érjük el, hogy a vektor neve elé mínuszjelet teszünk. Legyen *cs* egy vektor, ami egy adott utcában a családtagok számát adja meg a házszámok növekvő sorrendje szerint. A harmadik parancs megadja az elemek számát egy vektorban.

```
cs = c(1,1,5,2,1,4,4,3,2)
sort(cs)
[1] 1 1 1 2 2 3 4 4 5

#Növekvő sorrend
order(cs)
[1] 1 2 5 4 9 8 6 7 3

#Fordított sorrend
order(-cs)
[1] 3 6 7 8 4 9 1 2 5
#vagy
order(cs, decreasing = T)

length(cs)
[1] 9
```

A **merge()** paranccsal több listát össze lehet olvasztani egy listába, illetve a listákból vektorok készíthetők az **unlist()** paranccsal. Ha a lista elemei nem csak számokat tartalmaznak, a belőle létrehozott vektor karakterek formájában tárolja az adatokat.

```
L1 = list(c("piros", "kék"))
L2 = list(1:3)
L12 = merge(L1, L2)
names(L12) = c("Szín","Csoport")
L12
     Szín  Csoport
1   piros        1
2     kék        1
3   piros        2
4     kék        2
5   piros        3
6     kék        3
```

Felbontjuk vektorokra a 2.4. táblázatban megadott listát:

```
data = unlist(lista)
#mivel van legalább egy karakter, az összes
#adat karakterként kerül a vektorba
class(data)
[1] "character"
paste(data)
[1] "45" "44" "46" "1" "2" "3" "4" "5" "6" "7"
[11] "8" "9" "10" "11" "12" "kutya" "FALSE" "tyúk" "2" "0"
[21] "9" "TRUE"
```

A vektorok összeolvasztása egyszerű művelet, feltétele, hogy a vektorok azonos típusú adatokat tartalmaznak (numerikus vagy nevesítő adatkat).

```
v1 = c(2,2,4,8,1)
v2 = c(5,5,5,7)
v3 = c(v1,v2)
v3
[1] 2 2 4 8 1 5 5 5 7
```

2.1.1. Sorok, oszlopok, adatok kiválasztása egy adattárolóból

Egy sor vagy oszlop kiválasztása egy mátrixból, listából, adattáblából a szögletes zárójel használatával történik: **[,].** A vessző előtt a sor (sorok) sorszámai, a vessző után az oszlop (oszlopok) sorszámai szerepelnek. Sorok vagy oszlopok eltávolíthatók a sorszámuk elé tett mínuszjellel.

```
m[,1]         #az m matrix első oszlopa
S1 S2 S3 S4 S5
 3  2  6  4  2
m[2,]         #az m matrix második sora
VAR1 VAR2
   2   34
```

Ha egy adattáblában szeretnénk oszlopokat kiválasztani, akkor a $ jelet használjuk. A jeltől balra írjuk az adattábla nevét, jobbra pedig az oszlop nevét. Az R-nek vannak beépített adattáblái, pl. az *mtcars*, ezen belül pedig van egy *hp* nevű változó. Ezt a következőképpen kérhetjük ki:

```
mtcars$hp
 [1] 110 110 93 110 175 105 245 62 95 123 123 180 180 180 205
215 230 66 52 65 97 150 150 245 175
[26] 66 91 113 264 175 335 109
```

Mivel egyetlen változó adatait tartalmazza azonos típusú értékekkel, ezért elmenthetjük vektorként, pl. *v1* néven, amely így numerikus vektor lesz.

```
v1 = mtcars$hp
class(v1)
[1] "numeric"
```

Egy adattábla egy oszlopára úgy is utalhatunk, hogy az adattábla neve után szögletes zárójelben megjelöljük a változó sorszámát. A *hp* az *mtcars* adattáblában a negyedik változó.

```
mtcars[4]
                     hp
Mazda RX4           110
Mazda RX4 Wag       110
Datsun 710           93
Hornet 4 Drive      110
Hornet Sportabout   175
...
```

2.1.2. Gyakran használt parancsok

A rep() parancs
Egy értéket ismétel megadott szorzóval:

```
rep("big",3)
[1] "big" "big" "big"

rep(1:4, 2)
[1] 1 2 3 4 1 2 3 4
```

A **c()** funkció használatával bonyolultabb műveleteket is végre lehet hajtani:

```
#háromszor ismételje az első három egész számot
c(rep(1,3),rep(2,3),rep(3,3))
[1] 1 1 1 2 2 2 3 3 3

#egy bináris csoportváltozó készítése (pl. urban/rural)
Site = c(rep("urban",3), rep("rural",3))
class(Site)
[1] "character"
Site
[1] "urban" "urban" "urban" "rural" "rural" "rural"
```

cbind() és rbind()
A változót utólag hozzá lehet adni egy adattáblához a **cbind()** paranccsal. A **cbind()** paranccsal oszlopot lehet hozzáragasztani egy olyan adattáblához vagy másik oszlpohoz, amelynek azonos a sorszáma. Hasonlóan, sorokat is lehet adattáblához ragasztani az **rbind()** paranccsal.

```
#A d adattáblához hozzáadjuk a Site változót
cbind(d,Site)
```

seq() parancs
A **seq()** paranccsal szakaszokat lehet létrehozni egy megadott logika szerint. Például megadjuk a szakasz kezdő- és végpontját, és az értékkészletből adott értékeket jelölünk.

```
#páros számok kiválasztása [0,10] között
seq(0,10, by = 2)
[1] 0 2 4 6 8 10
```

```
#10 egyenlő távolságra levő érték kiválasztása 10 és 1000
#között (az egyenlő távolság 110)
seq(10, 1000, length.out = 10)
[1] 10 120 230 340 450 560 670 780 890 1000
```

runif() parancs

A **runif()** parancs véletlenszerűen választ ki *n* számot egy egyenletes eloszlásból (minden tagnak egyforma esélye van bekerülni az új adatsorba). Alapból az intervallum [0,1]: **runif(n, min = 0, max = 1).**

```
#Tíz szám véletlenszerű kiválasztása 1 és 100 között.
round(runif(10,1,100),0)
[1] 55 98 87 83 16 64 20 80 66 91
```

Ha ismételjük a **runif()** parancsot, akkor mindig más sorozatot kapunk, lévén, hogy az adatsor véletlenszerűen áll össze. A **set.seed()** paranccsal rögzíthetjük a sorozatot. Így biztosítjuk, hogy bárhol bárki ugyanazt az adatsort generálja.

```
set.seed(456)
round(runif(10,1,100),0)
[1] 10 22 74 85 79 34 9 29 25 39
```

rnorm() parancs

Az **rnorm(n, mean = 0, sd = 1)** parancs *n* számot generál egy 0 középértékű és egységnyi szórású normális eloszásból. A normális eloszlás középértéke és szórása tetszőlegesen változtatható. A **set.seed()** ebben az esetben a létrehozott adatsort rögzíti.

```
set.seed(6284)
rnorm(10,3,0.6)
[1] 3.389476 3.815450 3.342745 4.096105 3.581106 3.192450
4.014564 2.823472 2.753615 2.950230
```

sample() parancs

A **sample(x, size, replace = FALSE)** parancs véletlenszerűen választ ki adott számú (*size*) elemet egy mintába a megadott *x* halmazból (sokaságból) visszatevős vagy nem visszatevős módszerrel. A visszatevős módszernél ugyanaz az elem többször bekerülhet a mintába, mialatt minden elem azonos valószínűséggel került a mintába, hiszen a halmaz elemszáma mindvégig azonos maradt.

```
sample(1:15, 7, replace = TRUE)
[1] 13 13 14 15 1 2 1
```

table() parancs
A **table()** parancs egy faktor szintjeinek a gyakoriságát adja meg.

```
szin = c("piros", "piros", "kek", "piros", "piros", "kek")
table(szin)
szin
 kek piros
 2 4
```

2.1.3. R csomagok letöltése, telepítése és aktiválása

Az R-nek egy saját alapcsomagja {base} van, ami a program telepítésekor elérhetővé válik, ezenkívül még néhány csomag települ alapból (datasets, utils, grDevices, graphics, stats, methods). Ezek az alapcsomagok számos művelet elvégzését teszik lehetővé, viszont eszköztáruk korlátolt. Az R rugalmasságából adódóan számos speciális statisztikai teszt vagy művelet elvégzésére csomagok készültek (például a {vegan} csomag elsősorban ökológiai elemzés elvégzését teszi lehetővé, a {cluster} csomag a klaszteranalízishez tartalmaz parancsokat, a {ggplot2} csomag pedig vonzó ábrák készítésére alkalmas). A telepített csomagokat az R egy könyvtárban őrzi, ezért aktiválnunk kell a kívánt csomagot. Ezt megtehetjük a **library()** paranccsal (2.5. táblázat).

2.5. táblázat. *Az R csomagok letöltése, telepítése és aktiválása*

Szoftver	Jellemzés	Aktiválás
R	Packages → Set CRAN mirror (földrajzilag minél közelebbit) Packages → Install package(s) (kiválasztjuk a kívánt csomagot)	**library(ggplot2)** Warning message: package 'ggplot2' was built under R version 4.0.5
RStudio	Tools → Install package(s) A megjelenő ablakba beírjuk a kívánt csomag nevét.	**library(ggplot2)** Loading required package: permute Loading required package: lattice This is vegan 2.5-7 Warning message: package 'ggplot2' was built under R version 4.0.5

Vannak csomagok, amelyek nem érhetőek el a CRAN csomagtárolójában. Egyes szerzők saját oldalukon vagy a GitHub-on teszik elérhetővé őket. Ezeknek a csomagoknak a könnyű letöltéséhez és telepítéséhez létrehozták az RTools esz-

köztárat, ami a következő oldalon érhető el: *https://cran.r-project.org/bin/windows/Rtools/*. Az R verziójának megfelelően telepíthető az RTools 4.0 vagy 4.2 verzió.

2.2. Az adattáblák kezelése R-ben

Az adattábla *n* sorból és *m* oszlopból áll, ahol az *n* a megfigyeléseket (mintákat, objektumokat) jelenti, az *m* pedig a változókat. Egy változó lehet numerikus (valós szám, egész szám vagy komplex szám), nominális (betűk, szavak) vagy logikai (igaz, hamis).

2.2.1. Adattábla készítése

Egy faluban meghatároztuk öt kút vizének a helyszíni paramétereit: a pH-t, a hőmérsékletet (°C) és a fajlagos vezetőképességet (EC – electric conductivity, μS/cm). A vizsgált kutak az utcán, illetve magánterületen (udvarokban) fordultak elő, ezért egy csoportváltozót is készítünk, amit faktorként mentünk el.

```
ID = c("K1", "K2", "K3", "K4", "K5")
pH = c(6.73, 6.45, 6.21, 7.03, 6.58)
Hom = c(12.4, 15.7, 12.8, 13.3, 14.2)
EC = c(345, 687, 397, 451, 430)
Csoport = factor(c("udvar", "utca", "udvar", "udvar", "utca"),
levels = c("udvar","utca"))
```

A változókból adattáblát készítünk, a sorokat pedig elnevezzük a minták neveivel (ID).

```
df = data.frame(pH, Hom, EC, Csoport)
row.names(df) = ID
#kikérjük az adattáblát
df
       pH   Hom    EC  Csoport
K1   6.73  12.4   345    udvar
K2   6.45  15.7   687     utca
K3   6.21  12.8   397    udvar
K4   7.03  13.3   451    udvar
K5   6.58  14.2   430     utca
```

2.2.2. Adattábla betöltése

1. Az R saját adattábláinak beolvasása

Az R alapcsomagjában számos adattábla található. Ezek kilistázhatók a **`library(help = "datasets")`** paranccsal. Egy adott adattábláról (pl. *volcano*) részletesebb információt találunk a Help fül alatt (jobb alsó ablak) a **`?volcano`** paranccsal. Az adattábla megnézhető a név egyszerű bevezetésével a Console-ba: **`volcano.`**

2. Adattábla beolvasása egy internetes oldalról

Különböző internetes oldalakon hasznos adattáblákat tettek közzé, amelyek elérhetők és letölthetők. Gazdag adattábla-gyűjtemény található a *https://archive.ics.uci.edu/ml/datasets.php* oldalon. Egy adattáblát az R a saját munkamappájába tölt le, tehát a letöltés előtt célszerű megadni ennek a mappának a helyét.

Az említett internetes oldalról a *flags* adattábla a következő módon tölthető le, és olvasható be az R-be: a listából kikeressük az adattáblát, majd behozzuk a Data Foldert. A *flag.names* tartalmazza az információkat a változókról, a *flag.data* pedig az adattáblát. Az utóbbit úgy töltjük le, hogy az egér jobb gombjával behozunk egy ablakot, és a *Save link content as…* parancsra kattintunk. A *flag.data* 194 országról és zászlóiról tartalmaz információkat. Az adattábla 1986-ból származik. Az adattábla *.data kiterjesztésű, ezért a beolvasása az R-be a következőképpen történik:

```
flags = read.table(file.choose(), sep = ",")
```

Megjelenik egy ablak a munkamappa tartalmával, ahonnan kiválaszthatjuk a kívánt dokumentumot.

3. Saját adattábla beolvasása

Saját adattábla beolvasására a legegyszerűbb mód az, ha az Excel-dokumentumot átalakítjuk .csv (comma separated value) formátumra: Save As → CSV (MS-DOS). A dokumentumot a munkamappába mentjük, majd a következő paranccsal olvassuk be az R-be:

```
df = read.csv(file="*.csv", header = T, colClasses = c(),
stringsAsFactors = T, row.names = m)
#vagy
df = read.csv2(file="*.csv", header = T, stringsAsFactors =
T, row.names = m)
```

A **`row.names`** opcionális, akkor használjuk, ha az objektumok elnevezésének/kódjának van oszlopa; az *m* annak az oszlopnak a sorszáma, amelyben az objektu-

mok nevei vannak. A **colClasses()** lehetőséget ad arra, hogy egyértelműsítsük a változók típusát. A **stringsAsFactors = T** lehetővé teszi a karaktervváltozók faktorként való értelmezését.

A **d = read.table(file.choose(), sep = ",")** paranccsal beolvasott adattáblánál az objektumok kódjának oszlopa az alábbi paranccsal állítható be:

```
rownames(d) = d[,m],
```

ahol m a kódokat tartalmazó változó neve.

Különböző kód használatával információt kaphatunk az adattábláról. A kódokat a 2.6. táblázat összegezi. Az adattáblát a közérthetőség kedvéért *df*-fel jelöltük (**df = airquality**), és az R alapcsomagjában tárolt `airquality` adatait tartalmazza (2.6. táblázat). Az adatok New York levegőminőségére vonatkoznak 1973 nyári időszakára (május–szeptember). Összesen 153 nap adatairól van szó. A változók: ózonkoncentráció (ppb), napsugárzás (kal/cm^2), szélsebesség (mérföld/óra), hőmérséklet (F), hónap és nap.

2.6. táblázat. *Az* airquality *adattábláról kikérhető információk*

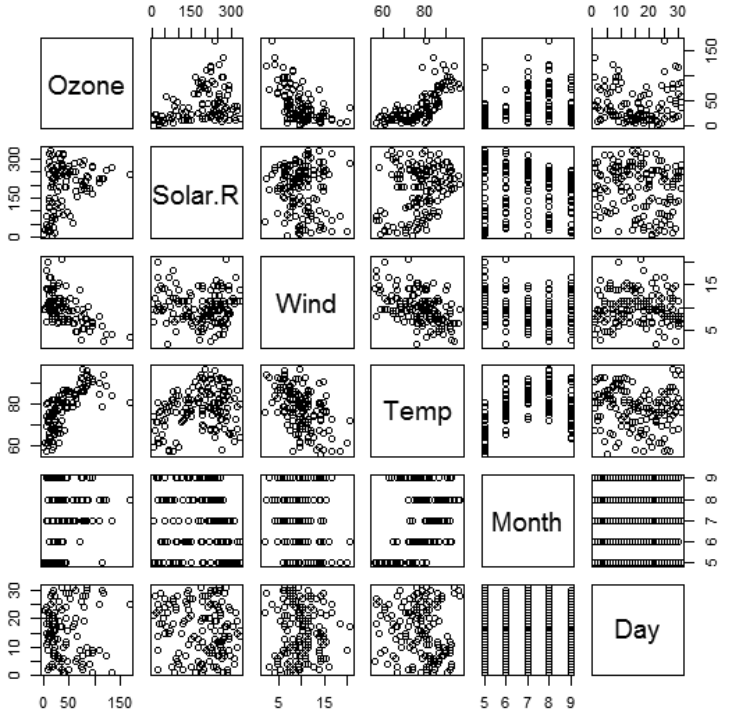
Információ	Kód	Példa
Leírás	**?airquality**	`New York Air Quality Measurements Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973. (…)`
Típusa	**class(df)**	`[1] "data.frame"`
Sorok, oszlopok száma)	**dim(df)**	`[1] 153 6`
Oszlopok nevei	**names(df)**	`[1] "Ozone" "Solar.R" "Wind" "Temp" "Month" "Day"`

Az adattábla első hat, ill. első 10 sora	**head(df)** **head(df, 10)**	`  Ozone Solar.R Wind Temp Month`
		`1 41 190 7.4 67 5`
		`2 36 118 8.0 72 5`
		`3 12 149 12.6 74 5`
		`4 18 313 11.5 62 5`
		`5 NA NA 14.3 56 5`
		`6 28 NA 14.9 66 5`

Az adattábla utolsó hat sora	**tail(df)**	`  Ozone Solar.R Wind Temp Month`
		`148 14 20 16.6 63 9`
		`149 30 193 6.9 70 9`
		`150 NA 145 13.2 77 9`
		`151 14 191 14.3 75 9`
		`152 18 131 8.0 76 9`
		`153 20 223 11.5 68 9`

Információ	Kód	Példa
A változók típusa	**str(df)**	'data.frame': 153 obs. of 6 variables: $Ozone: int 41 36 12 18 NA 28 23… $Solar.R: int 190 118 149 313 NA… $Wind: num 7.4 8 12.6 11.5 14.3… $Temp: int 67 72 74 62 56 66… $Month: int 5 5 5 5 5 5 5 5 5 5… $Day: int 1 2 3 4 5 6 7 8 9 10…
A változók statisztikai jellemzői	**summary(df)**	Ozone Solar.R Min. : 1.00 Min. : 7.0 1st Qu. : 18.00 1st Qu. : 115.8 Median : 31.50 Median : 205.0 Mean : 42.13 Mean : 185.9 3rd Qu. : 63.25 3rd Qu. : 258.8 Max. : 168.00 Max. : 334.0 NA's : 37 NA's : 7

A változók páronkénti kapcsolatát a **pairs(df)** paranccsal nézhetjük meg (2.2. ábra).



2.2. ábra. *Az* airquality *adattábla változóinak páronkénti kapcsolata*

2.2.3. Adattáblán belüli műveletek

Egy adattáblából aladattáblák készíthetők egy vagy több logikai szempont alapján. A `[]` zárójelek használatával egy logikai művelet végezhető el az adattáblán, például az `airquality` adattáblából azokat a sorokat szeretnénk megtartani, amelyre az ózonkoncentráció 70 ppb fölött volt, az al-adattáblát pedig elmentjük `df70` név alatt.

```
df = airquality
df70 = df[df$Ozone > 70,]
```

Ugyanez a művelet elvégezhető a **`subset()`** paranccsal is. Ennek a parancsnak az az előnye is megvan, hogy egyszerre több kritériumot is kezelni tud. Egy olyan aladattáblát készíthetünk, amelyben csak azok a sorok szerepelnek, amelyekre az ózonkoncentráció > 70, a szélsebesség pedig kisebb vagy egyenlő 10. Az új adattáblának a `df2` nevet adjuk.

```
df2 = subset(df, Ozone > 70 & Wind <= 10)
head(df2)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
30	115	223	5.7	79	5	30
62	135	269	4.1	84	7	1
68	77	276	5.1	88	7	7
69	97	267	6.3	92	7	8
70	97	272	5.7	92	7	9
71	85	175	7.4	89	7	10

```
dim(df2)
[1] 23 6
```

A **`dim(df2)`** megmutatja, hogy összesen 23 olyan nap volt május és szeptember között, amikor az ózonkoncentráció 70 ppb fölött volt, a szélsebesség pedig 10 mérföld/órával volt egyenlő vagy annál kisebb.

A **`table()`** parancs egy faktor szintjeinek megfelelően megadja a logikai műveletnek megfelelő sorok számát. Az előző esetnél maradva tudjuk, hogy az említett két feltétel összesen 23 napon fordult elő egyszerre. A **`table()`** paranccsal megtudhatjuk a 23 nap havi lebontását.

```
table(df2$Ozone > 70, df2$Month)
      5 7 8 9
 TRUE 1 9 9 4
```

Az eredeti df adattáblára a két feltételnek megfelelő napok havi bontását az alábbi paranccsal kérhetjük ki:

```
table(df$Ozone > 70 & df$Wind <= 10, df$Month)
        5  6  7  8  9
 FALSE 28 17 20 20 26
 TRUE   1  0  9  9  4
```

Ebben az esetben megjelent még egy sor, ami megmutatja, hogy hány sor nem felelt meg az előírt feltételeknek.

Egy adattábla soraira és oszlopaira elvégezhető pár egyszerű művelet a **rowSums()**, **rowMeans()**, **colSums()** és **colMeans()** parancsokkal. Ezek a parancsok összeadják, ill. átlagolják a sorok és oszlopok értékeit. Az **apply()** parancs lehetővé teszi egy adattáblában egy művelet elvégzését több sorra vagy oszlopra. A **tapply()** parancs különösen hasznos lehet olyan adattáblánál, amelyben faktor (csoportosító változó) is van. A csoportosító változó minden szintjére elvégez egy művelet, az eredményeket pedig táblázat formájában adja meg, innen a *t* betű a parancsban. Az említett parancsok alkalmazásának részletei a 2.7. táblázatban láthatók.

2.7. táblázat. *Néhány hasznos parancs részletei adattáblákra*

Parancs	Értelmezés	Részletek
rowSums(x, na.rm = FALSE) rowMeans(x, na.rm = FALSE) colSums(x, na.rm = FALSE) colMeans(x, na.rm = FALSE) apply(x, MARGIN, FUN, na.rm = F) tapply(x, factor, FUN, na.rm = F)		
x		egy adattábla olyan sora(i) vagy oszlopa(i), amely(ek) számadatokat tartalmaznak
na.rm	TRUE/FALSE	hagyja ki az R a számításból a hiányzó adatokat?
MARGIN	MARGIN = 1 MARGIN = 2 MARGIN = c(1,2)	MARGIN – 1 vagy 2 (1 sorokra, 2 oszlopokra)
factor		egy csoportosító változó
FUN		FUN – elvégzendő művelet

```
#Hiányzó adatoknál az R nem átlagol
colMeans(df[,1:3], na.rm = F)
```

```
    Ozone  Solar.R     Wind
       NA       NA 9.957516

#Eltávolítjuk a hiányzó adatokat
colMeans(df[,1:3], na.rm = T)
     Ozone    Solar.R       Wind
 42.129310 185.931507   9.957516

#Átlagok az airquality első három változójára
apply(df[,1:3],2,mean, na.rm = T)
     Ozone    Solar.R       Wind
 42.129310 185.931507   9.957516

#Szórások ózonra hónapok szerint
tapply(df$Ozone, df$Month, sd, na.rm = T)
       5        6        7        8        9
22.22445 18.20790 31.63584 39.68121 24.14182

#Shapiro-Wilk-próba elvégzése ózónra havi bontásban
tapply(df$Ozone, df$Month, shapiro.test)
$`5`
Shapiro-Wilk normality test
data: X[[i]]
W = 0.71401, p-value = 8.294e-06

$`6`
Shapiro-Wilk normality test
data: X[[i]]
W = 0.84328, p-value = 0.0628

$`7`
Shapiro-Wilk normality test
data: X[[i]]
W = 0.97967, p-value = 0.8669

$`8`
Shapiro-Wilk normality test
data: X[[i]]
W = 0.93279, p-value = 0.09032

$`9`
Shapiro-Wilk normality test
```

```
data: X[[i]]
W = 0.78373, p-value = 4.325e-05
```

A **tapply()** paranccsal akár statisztikai próba is elvégezhető, ilyen a Shapiro–Wilk-próba egy eloszlás tesztelésére (lásd a 4.2. alfejezetet). Ez a próba egy művelettel elvégezhető egy csoportváltozó minden szintjén bármely numerikus folytonos változóra.

2.3. Az adattábla előkészítése a statisztikai elemzéshez

A legtöbb esetben az adattábla nem felel meg a további statisztikai műveletek elvégzéséhez. Ezért előbb formázni kell. Ezek a formázások magukba foglalják a hiányzó adatok kezelését, az oszlopok formázását, a szavak átírását, a mértékegységek átalakítását stb. Példaként az R saját adatbázisában lévő *airquality* adattáblával foglalkozunk, amit elneveztünk *df*-nek (**df = airquality**).

2.3.1. A hiányzó adatok kezelése

A leíró statisztikai adatok kimutatásából (**summary(df)**) kiderült, hogy az ózonkoncentrációnál 37 napra nincs adat, a napsugárzás adatai úgyszintén hét esetben hiányosak. Az R a hiányzó adatokat egy karakterpárral értelmezi és jelöli (NA – not available). Megtörténhet, hogy az adattáblában bizonyos formában már jelölték a hiányzó adatokat (például *nd*-vel, ami a *not detected* rövidítése). Az R az NA-tól eltérő rövidítéseket nominális elemekként értelmezi, ezért ezeket NA-val kell helyettesíteni a következő kóddal: **df[df == „nd”] = NA.**

A legtöbb statisztikai művelet nem végezhető el olyan változóra, amely hiányzó adatokat tartalmaz, ezért ezeket valamilyen formában kezelni kell. Erre több lehetőség van:

1) Ha az objektumok számához viszonyítva a hiányzó adatokat tartalmazó objektumok száma elhanyagolható (általában < 5%), akkor ezeket a sorokat eltávolíthatjuk.
   a) A sorok kihagyhatók a **na.omit()** parancs használatával. Az új, *df1* adattáblában már hiányzanak egyes sorok (pl. 5, 6).

```
df1 = na.omit(df)
head(df1)
    Ozone  Solar.R     Wind   Temp  Month    Day
1      41      190      7.4     67      5      1
2      36      118      8.0     72      5      2
```

```
3      12      149     12.6     74       5       3
4      18      313     11.5     62       5       4
7      23      299      8.6     65       5       7
8      19       99     13.8     59       5       8
```

b) Egyes parancsok egy logikai vektor formájában tüntetik fel a hiányzó adatok helyét. A **complete.cases()** „TRUE” választ ad a létező adatokra, és „FALSE” választ a hiányzó adatokra, az **is.na()** pedig fordítva működik, „TRUE” választ ad a hiányzó adatokra. A hiányzó adatokat tartalmazó sorok eltávolítása a következő módon is megvalósítható:

```
complete.cases(df)
df2 = df[complete.cases(df),]
```

Ha azokat az objektumokat szeretnénk látni, amelyekre nem teljesek az adatok, a következő parancsot alkalmazzuk (a felkiáltójel tagadást jelent):

```
df[!complete.cases(df),]
```

2) Vannak esetek, amikor az objektumok helyett célszerű lenne pár változónak az eltávolítása, amennyiben ezek értékei többnyire a módszer kimutatási határa alatt vannak, vagy több alkalommal nem voltak mérhetők/megfigyelhetők. Az NA értékek száma változónként megfigyelhető a **summary(df)** kimutatásnál, ugyanakkor egy R kóddal is azonosíthatjuk őket:

```
apply(apply(df, 2, is.na), 2, sum)
Ozone Solar.R Wind Temp Month Day
   37       7    0    0     0   0
```

A két változó nem tartalmaz annyi hiányzó adatot, hogy indokolt lenne az eltávolításuk.

3) Ha túl sok objektum tartalmaz hiányzó adatokat, akkor az NA helyettesíthető egy számmal. Kevésbé elfogadott megoldás az, ha az NA-t 0-val helyettesítjük, bizonyos esetekben viszont indokolt lehet. Ha a 0 nem elfogadott, akkor esetenként a változó átlagértékével vagy mediánjával helyettesíthetjük, ill. ha az adat azért hiányzik, mert az adott objektumra a változó értéke annyira kicsi volt, hogy az alkalmazott módszerrel vagy

használt mérőműszerrel nem volt mérhető, akkor a módszer kimutatási határának[1] törtrészével helyettesítjük.

a) *A hiányzó adatok helyettesítése 0-val*

```
df[is.na(df)] = 0
```

b) *A hiányzó adatok helyettesítése a változó értékeinek átlagával vagy mediánjával*
Abban az esetben, ha nincsenek csoportváltozók az adattáblában, azaz az összes objektum egy populációból származik, az R kód egyszerű:

```
df$Ozone[is.na(df$Ozone)] = mean(df$Ozone, na.rm = T)
df$Solar.R[is.na(df$Solar.R)] = mean(df$Solar.R, na.rm
= T)
```

Abban az esetben, ha az objektumok különböző populációkból származnak, amit egy csoportváltozó jelöl, az R kód bonyolultabb. Az átlagok, amivel a hiányzó adatokat helyettesítjük, csoportátlagok kell legyenek. Ez a helyzet áll fenn az *airquality* adattáblanál is, ahol az `NA`-kat az adott hónap átlagértékeivel célszerű helyettesíteni.
Az R kód megírható az alapcsomag segítségével, és annyi lépést tartalmaz, ahány hónap hiányos adatsorral rendelkezik az adott változóra. Az ózonkoncentráció esetében mind az öt hónapra, a napsugárzás esetében pedig az 5. és 8. hónapra végezzük el a műveletet:

```
df$Ozone[df$Month==5 & is.na(df$Ozone)] =
mean(df$Ozone[df$Month==5], na.rm=TRUE)
...
df$Ozone[df$Month==9 & is.na(df$Ozone)] =
mean(d$Ozone[df$Month==9], na.rm=TRUE)
df$Solar.R[df$Month==5 & is.na(df$Solar.R)]=
mean(df$Solar.R[df$Month==5], na.rm=TRUE)
df$Solar.R[df$Month==8 & is.na(df$Solar.R)] =
mean(df$Solar.R[df$Month==8], na.rm=TRUE)
head(df, 10)
      Ozone  Solar.R  Wind  Temp  Month   Day
1  41.00000 190.0000   7.4    67      5     1
2  36.00000 118.0000   8.0    72      5     2
```

1 MKH (módszer kimutatási határ): az alkalmazott analitikai módszer által mérhető legkisebb koncentrációérték.

```
 3  12.00000 149.0000  12.6    74       5     3
 4  18.00000 313.0000  11.5    62       5     4
 5  23.61538 181.2963  14.3    56       5     5
 6  28.00000 181.2963  14.9    66       5     6
 7  23.00000 299.0000   8.6    65       5     7
 8  19.00000  99.0000  13.8    59       5     8
 9   8.00000  19.0000  20.1    61       5     9
10  23.61538 194.0000   8.6    69       5    10
```

Az aktualizált adattáblában az *Ozone* változónál az 5. és 10. objektumnál, a *Solar.R* változónál pedig az 5. és 6. objektumnál megjelentek a májusi hónap átlagértékei.

A {dplyr} csomag egy elegánsabb kód használatára ad lehetőséget:

```
library(dplyr)
df = df %>% group_by(Month) %>%
 mutate(Ozone = ifelse(is.na(Ozone), mean(Ozone, na.rm
= T), Ozone))
df
# A tibble: 153 x 6
# Groups: Month [5]
      Ozone  Solar.R  Wind  Temp  Month    Day
      <dbl>    <dbl> <dbl> <int>  <int>  <int>
 1       41      190   7.4    67      5      1
 2       36      118     8    72      5      2
 3       12      149  12.6    74      5      3
 4       18      313  11.5    62      5      4
 5     23.6     181.  14.3    56      5      5
 6       28     181.  14.9    66      5      6
 7       23      299   8.6    65      5      7
 8       19       99  13.8    59      5      8
 9        8       19  20.1    61      5      9
10     23.6      194   8.6    69      5     10

# … with 143 more rows
```

Lényegesen rövidebb idő alatt ugyanarra az eredményre jutottunk.

c) *A hiányzó adatok helyettesítése a módszer kimutatási határának felével*

Minden változónak más-más MKH értéke van. Ebben az esetben a helyettesítéseket változónként kell elvégezni. A *df* adattáblában két változóra (ózonkoncentrációra és napsugárzásra) nincsenek teljes adatok.

Az egyszerűség kedvéért legyen az ózonmeghatározás módszerének a kimutatási határa 0.6 ppb, a napsugárzás méréséé pedig 2 kal/cm^2.

```
df$Ozone[is.na(df$Ozone)] = 0.3
df$Solar.R[is.na(df$Solar.R)] = 1
head(df)
```

```
      Ozone  Solar.R  Wind  Temp  Month     Day
1      41.0      190   7.4    67      5       1
2      36.0      118   8.0    72      5       2
3      12.0      149  12.6    74      5       3
4      18.0      313  11.5    62      5       4
5       0.3        1  14.3    56      5       5
6      28.0        1  14.9    66      5       6
```

Minden sor megmaradt, az NA-k helyében pedig a kimutatási határok félértékei szerepelnek. A mérési adatok leírását figyelembe véve egyértelművé válik, hogy nem mérési problémáról van szó, hanem az adatok egyszerűen hiányoznak. Ezért az *airquality* adattáblánál a legmegfelelőbb eljárás a havonkénti átlagok használata.

2.3.2. Az adattábla formázása

Megváltoztathatjuk az adattábla változóinak a neveit:

```
colnames(df) = c("Ozon", "Sugarzas","Szelseb", "Hom", "Honap",
"Nap")
head(df)
```

```
           Ozone  Solar.R  Wind  Temp  Month     Day
1  41.00000 190.0000   7.4    67      5       1
2  36.00000 118.0000   8.0    72      5       2
3  12.00000 149.0000  12.6    74      5       3
4  18.00000 313.0000  11.5    62      5       4
5  23.61538 181.2963  14.3    56      5       5
6  28.00000 181.2963  14.9    66      5       6
```

A numerikus (arányskálájú) változók értékei kerekíthetők:

```
df = round(df, 1)
head(df)
```

```
           Ozon Sugarzas  Szelseb   Hom  Honap     Nap
1      41.0    190.0      7.4    67      5       1
2      36.0    118.0      8.0    72      5       2
3      12.0    149.0     12.6    74      5       3
```

```
4      18.0    313.0     11.5    62       5       4
5      23.6    181.3     14.3    56       5       5
6      28.0    181.3     14.9    66       5       6
```

A változók értékeit célszerű egységesíteni. Például a koncentrációértékeket egységesen mg/l, μg/l vagy mol/l formában tüntessük fel, illetve amennyiben lehetséges, a nemzetközi mértékrendszerhez igazodjunk. Az *airquality* adattáblában a hőmérsékletet Fahrenheitban (°F), a napsugárzást pedig kal/cm^2-ben adták meg. Ezeket alakítsuk át Celsius-fokokba, illetve kJ/m^2-be. Az átalakítások a következő összefüggésen alapulnak: 1 °C = (°F−32)/1.8, illetve 1 kal/cm^2 = 41868 J/m^2. A napsugárzás értékeit a könnyebb értelmezés céljából MJ/m^2-re számoljuk át.

A hőmérséklet átalakítása:

```
df$Hom = (df$Hom-32)/1.8
```

A napsugárzás átalakítása:

```
df$Sugarzas = df$Sugarzas*41868/10^6
df = round(df, 1)
head(df)
```

```
        Ozon Sugarzas  Szelseb   Hom  Honap     Nap
 1      41.0      8.0      7.4  19.4      5       1
 2      36.0      4.9      8.0  22.2      5       2
 3      12.0      6.2     12.6  23.3      5       3
 4      18.0     13.1     11.5  16.7      5       4
 5      23.6      7.6     14.3  13.3      5       5
 6      28.0      7.6     14.9  18.9      5       6
```

2.4. Az adatok megismerése grafikus eszközökkel

Az adattábla rövid áttekintése és formázása után, de a statisztikai feldolgozás előtt célszerű megismerni a változók tulajdonságait (eloszlását, szóródását, kiugró értékek jelenlétét), két változó egymáshoz való viszonyulását (együtt változnak-e vagy nem), illetve több változó egymáshoz való viszonyulását (mintázatok felderítését). A személyes kép kialakításának a legjobb módszerei a grafikus eszközök. A leíró statisztikában használt ábráknak tehát nem az a célja, hogy nagyközönség elé kerüljenek. Ha szeretnénk őket bemutatni, illetve leközölni, akkor általában utólagos formázást igényelnek. Ezek az ábrák az adatok több szempontból való megismerését teszik lehetővé, rövid idő alatt elkészíthetők. A színeknek és a méreteknek elsősorban információközlő szerepe van, nem esztétikai jellegűek.

A leíró statisztikai adatelemzésben a következő ábrákat használjuk:
1. hisztogram (információt szolgáltat egy változó adatainak eloszlásáról);
2. dobozdiagram (információt szolgáltat egy változó adatainak eloszlásáról, azonosítható rajta a változó középértéke, illetve a kiugró értékek);
3. sűrűségdiagram (egy változó eloszlásának nem parametrikus becslését végzi el);
4. szórásdiagram (két változó egymáshoz való viszonyulásáról nyújt információt);
5. oszlopdiagram (a középértékek változása követhető egy csoportváltozó szerint, a szóródás is feltüntethető rajta).

2.4.1. A hisztogram vagy gyakorisági görbe

A hisztogram a vizsgált változó eloszlását teszi láthatóvá, megmutatja a változó egyenlő hosszúságú szakaszokra osztott értékskáláján (osztályok) az egyes szakaszokba bekerült minták számát. Az értékközök az ábrán egyenlő szélességű oszlopok formájában jelennek meg, az oszlopok magasságai pedig az adott értékközben található minták számát vagy gyakoriságát (százalékos arányát) mutatják. A hisztogramot a móduszok száma (egy-, két-, több móduszú vagy egyenletes), illetve a felfelé és lefelé ívelő szakaszok egymáshoz viszonyított formája (jobb oldali vagy bal oldali ferdeség) szerint jellemezzük.

A hisztogram az R-ben a **`hist()`** paranccsal végezhető el.

```
hist(df$Ozon)
rug(df$Ozon)
#Az ábra alatt vonalakként tünteti fel a mintákat.
```

Egy másik módja a hisztogram készítésének a **`with(df, hist())`** kód.

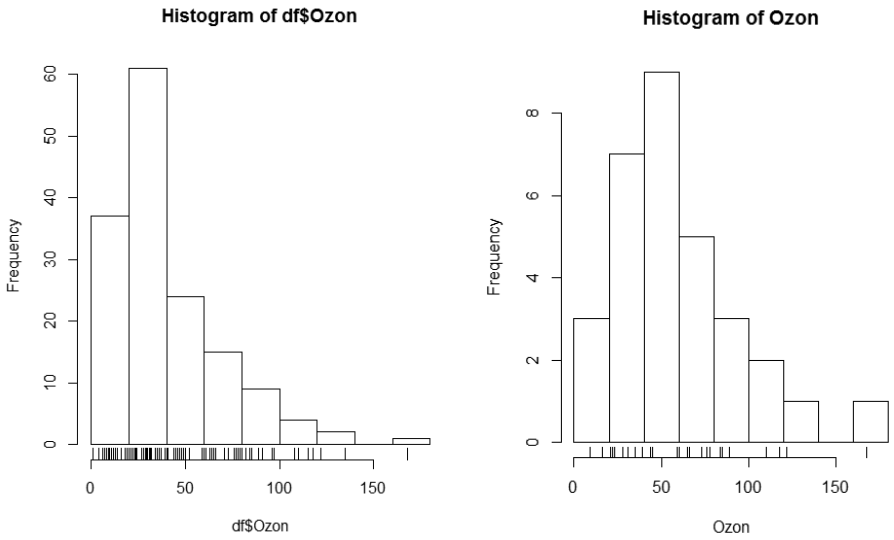
```
with(df, hist(Ozon))
```

A **`with(data, expr)`** parancs segítségével megadjuk azt a környezetet (data), ahol az adott parancsot (expr) alkalmazni szeretnénk, így nem kell az abszolút hivatkozást megadni (`df$Ozon`).

Az augusztusi hónapra kapott ózonkoncentrációk hisztogramjának kódja:

```
with(subset(df, Honap == 8), hist(Ozon))
with(subset(df, Honap == 8), rug(Ozon))
```

Mind a két esetben a hisztogram egymóduszú, és jobb oldali ferdeséget mutat (2.3. ábra), azaz a kisebb koncentrációk a gyakoribbak, a vizsgált időszakban a levegő ózonkoncentrációja ritkán mutatott nagy értéket.

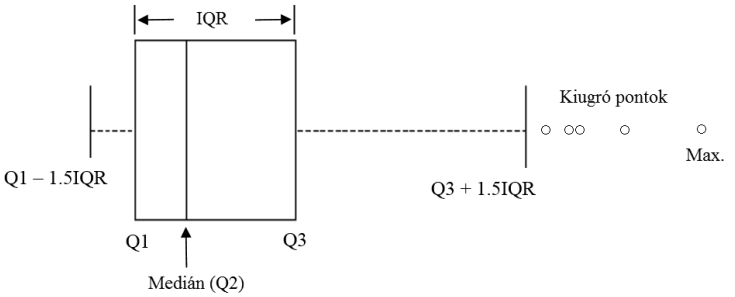


2.3. ábra. *Az ózonkoncentráció eloszlása hisztogramon: a teljes mérési időtartamra (balra) és az augusztusi hónapra (jobbra). Az adatok az* airquality *adattáblából származnak.*

2.4.2. A dobozdiagram

A hisztogramhoz hasonlóan a dobozdiagram egy változó eloszlásáról nyújt információt, ugyanakkor jelzi a középértéket, valamint a kiugró értékeket is (2.4. ábra). Tetszés szerint a minimum és a maximum értékeket is fel lehet tüntetni rajta. A növekvő sorrendbe rendezett adatok közül megjelöljük azt, ami a sornak az első negyedét választja el a második negyedtől, ez az első kvartilis (Q1); megjelöljük azt, ami a sort két egyenlő részre osztja (Q2), és azt, ami a harmadik negyedet választja el a negyedik negyedtől (Q3). A Q2 az adatsor mediánjával egyenlő. Az adatok középső 50%-át a Q1 és Q3 közötti szakasz tartalmazza, ez a diagram doboz része, és interkvartilis tartománynak (IQR) nevezzük. A dobozon kívül eső alsó és felső szakasz határvonala tetszés szerint jelölheti a minimum és a maximum értékeket vagy a Q1 – 1.5IQR és a Q3 + 1.5IQR értékeket. Ekkor a range argumentumot 0-ra kell állítani, alapesetben a range=1.5. Az utóbbi esetben létezhetnek értékek, amelyek az IQR 1.5-szörös tartományán kívül esnek. Ezek a kiugró értékek, a IQR tartomány háromszorosát meghaladó értékek pedig az extrém értékek. A (Q1 – 1.5IQR, Q3 +

1.5IQR) intervallum azt a tartományt jelöli, amelyben nagy valószínűsége van annak, hogy megtaláljuk mindazokat a mintákat, amelyek egy populációból származnak. A diagramon a dobozok függőlegesen vagy vízszintesen helyezkedhetnek el.

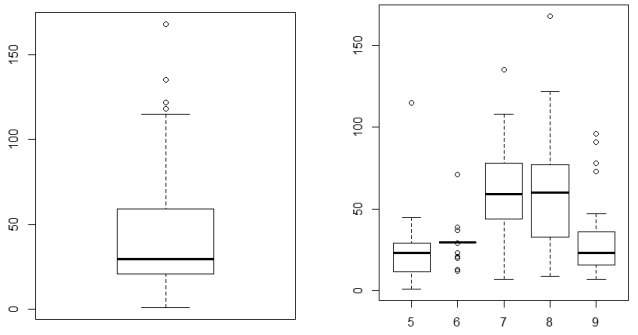


2.4. ábra. *A dobozdiagram felépítése*

A dobozdiagramot R-ben a **`boxplot()`** paranccsal készítjük el.

```
boxplot(df$Ozon) #A teljes időszakra.
boxplot(df$Ozon~df$Honap) #Felbontva hónapokra.
#A with() kóddal:
with(df, boxplot(Ozon))
with(df, boxplot(Ozon~Honap))
```

A 2.5. ábrán megfigyelhetők a dobozdiagramok a teljes időtartamra, illetve hónapokra lebontva. Az R alapbeállítása szerint a szakaszok határait a Q1 – 1.5IQR és a Q3 + 1.5IQR értékek adják meg, így a kiugró értékek is láthatók, pontok formájában.



2.5. ábra. *Az ózonkoncentráció eloszlása dobozdiagramon: a teljes mérési időtartamra (balra) és a hónapok lebontva (jobbra). Az adatok az* airquality *adattáblából származnak.*

2.4.3. A sűrűségdiagram

A sűrűségdiagram egy változó eloszlásának nem parametrikus becslését végzi el. Nagyban hasonlít a hisztogramhoz, azzal a különbséggel, hogy az oszlopközök lényegesen kisebbek, az oszlopok csúcsait összekötő görbe pedig egy algoritmust követ, általában a Gauss-féle függvényt. Az y-tengely skálája dimenziómentes, és az $f(y)$ értékek eleget tesznek annak a feltételnek, hogy a görbe alatti terület eggyel legyen egyenlő. A sűrűségfüggvényt az R-ben a **density()** és a **plot()** parancsok felhasználásával készíthetjük el.

```
d = density(df$Ozon)
plot(d)
```

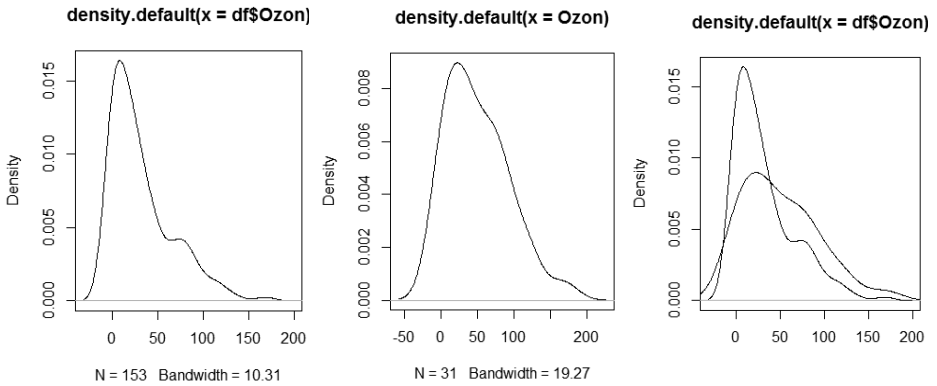
Az augusztusi hónap ózonkoncentrációnak a sűrűségdiagramja:

```
ds = with(subset(df, Honap == 8), density(Ozon))
plot(ds)
```

A 2.6. ábrán láthatók a sűrűségdiagramok. Mivel a két ábra x-tengelyének a skálája megegyezik, egy ábrán is megjeleníthetők. A **lines()** paranccsal vonal szerkeszthető. Megadható egy vonal *x* és *y* értékeinek a vektora, ebben az esetben úgy a d, mint a ds adattábla a sűrűségdiagram x és y koordinátáit tartalmazza. Az ábrát kétféle módon vehetjük fel:

```
d = density(df$Ozon)
ds = with(subset(df, Honap == 8), density(Ozon))
ds
Call:
density.default(x = Ozon)
Data: Ozon (31 obs.);  Bandwidth ‚bw' = 19.27
 x y
 Min. :-57.50 Min. :7.503e-06
 1st Qu.: 13.33 1st Qu.:6.358e-04
 Median : 84.15 Median :2.495e-03
 Mean : 84.15 Mean :3.526e-03
 3rd Qu.:154.97 3rd Qu.:6.706e-03
 Max. :225.80 Max. :8.986e-03

#Elso lehetőség
plot(d, xlab = "") #nem ir nevet az x-tengelynek
lines(ds)
#Második lehetőség
```

```
plot(d, type = "n", xlab = "") #csak a tengelyek
lines(d)
lines(ds)
```



2.6. ábra. *A sűrűségdiagramok az ózonkoncentrációra: a teljes mérési időtartamra (balra) és az augusztusi hónapra (középen), valamint a két grafikon egymásra helyezése (jobbra). Az adatok az* airquality *adattáblából származnak.*

2.4.4. A szórásdiagram

A szórásdiagram két változó egymáshoz való viszonyulását szemlélteti. A szórásdiagram láthatóvá teszi a két változó kapcsolatát, ami lehet erős vagy gyenge korreláció, vagy egyéb függvénnyel megközelíthető kapcsolat. Az R-ben a szórásdiagram a **`plot()`** paranccsal ábrázolható.

```
plot(Ozon~Hom, df)
#vagy
with(df, plot(Ozon~Hom))

#szórásdiagram az augusztusi hónapra
with(subset(df, Honap == 8), plot(Ozon~Hom))
```

A 2.7. ábrán látható, hogy az öt hónapon keresztül regisztrált értékek alapján a hőmérséklet növekedésével a levegő ózonkoncentrációja is enyhén növekedett. Néhány kiugróan magas ózonkoncentráció-értéket is láthatunk a 25–35 °C-os hőmérsékleti tartományban. Ezek ellenére az ábra azt sugallja, hogy New Yorkban, az adott időszakban, a hőmérséklet növekedése hozzájárulhatott az ózonkoncentráció növekedéséhez. A hőmérséklet és az ózonkoncentráció közti pozitív kapcsolat egyértelműen kimutatható az augusztusi hónapban is. Az ózon-

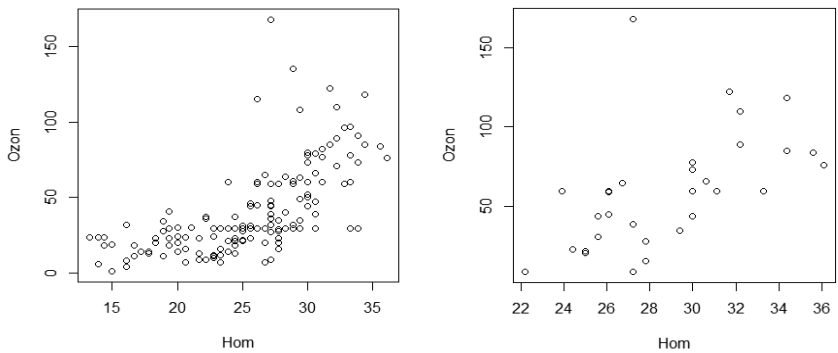
képződés kémiai hátterét ismerve ez az eredmény várható volt, mivel az ózon, eltérően sok más légszennyező anyagtól, elsősorban reakciók során képződik, amelyeket a magasabb hőmérséklet és az erős napsugárzás is felgyorsít. Érdemes megvizsgálni az ózonkoncentráció és a napsugárzás erőssége közti kapcsolatot is egy szórásdiagramon.

```
with(df, plot(Ozon~Sugarzas))
with(subset(df, Honap == 8), plot(Ozon~Sugarzas))
```

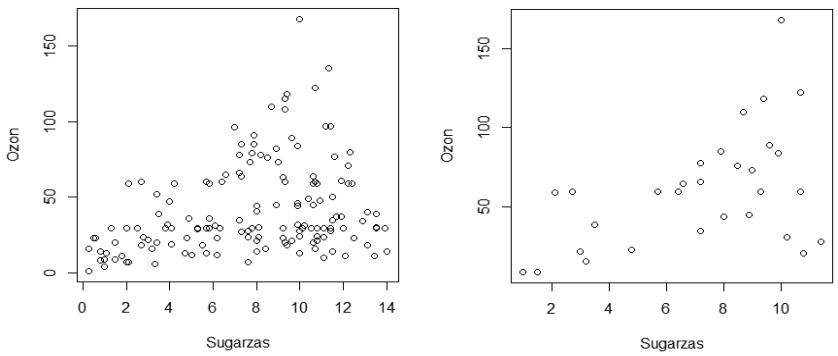
A 2.8. ábrán az ózonkoncentráció kapcsolata a napsugárzással nem tűnik annyira egyértelműnek, mint a hőmérséklettel, noha a kapcsolat mindenképp pozitív. Ennek az is lehet az oka, hogy a napsugárzás nem változott annyira erőteljesen, mint a hőmérséklet. Ezt megtudhatjuk a hőmérséklet és a napsugárzás hisztogramjáról (2.9. ábra).

```
with(df, hist(Hom))
with(df, hist(Sugarzas))
```

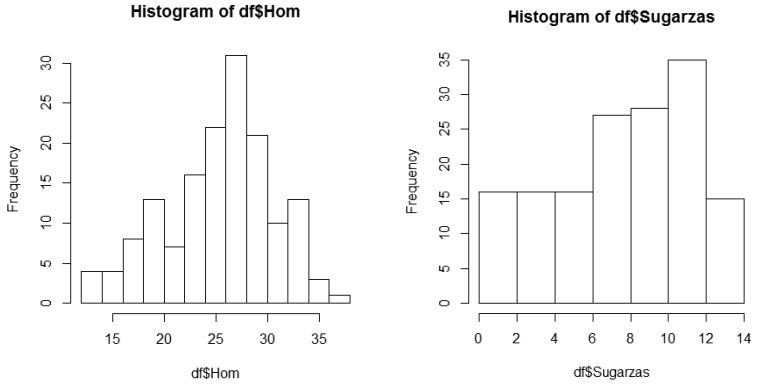
A hisztogramok szerint a mérési időszakban a levegő hőmérséklete nagyobb mértékben változott, mint a napsugárzás, ami magyarázhatja a napsugárzás gyengébb kapcsolatát az ózonkoncentrációval, a levegő hőmérsékletéhez képest.



2.7. ábra. *Az ózonkoncentráció és a levegő hőmérsékletének kapcsolata a teljes mérési időszakra (balra) és az augusztusi hónapra (jobbra)*



2.8. ábra. *Az ózonkoncentráció és a napsugárzás kapcsolata a teljes mérési időszakra (balra) és az augusztusi hónapra (jobbra)*



2.9. ábra. *A levegőhőmérséklet és a napsugárzás hisztogramja*

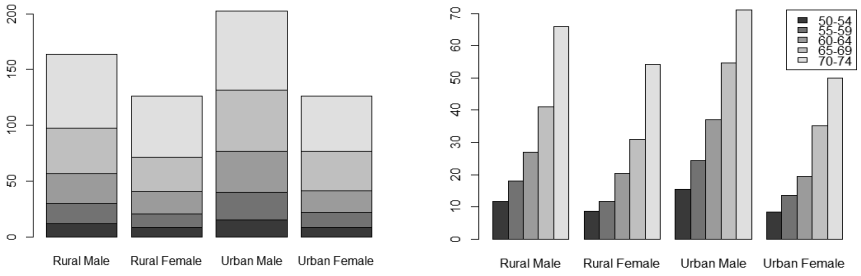
2.4.5. Az oszlopdiagram

Az oszlopdiagram diszkrét, numerikus változókra alkalmazható, és információt szolgáltat a változó abszolútértékéről vagy középértékéről egy ordinális vagy egy nominális változó szerint. Az R-ben a mátrix típusú adattábla egy lépésben ábrázolható. Az R könyvtárában megtalálható VADeaths adatmátrix öt sorból és négy oszlopból áll, és Virginia (Amerikai Egyesült Államok tagállama) halálozási arányát (1000 emberre eső halottak számát) tartalmazza nemek és települések szerint, 1940-re. Oszlopdiagramot a **`barplot()`** paranccsal készíthetünk. Magyarázat nélkül az oszlopok különböző színének nincs jelentése, ezért célszerű egy

magyarázatot feltüntetni. Ezt a **legend()** paranccsal tehetjük meg. A mátrix sorainak a nevei a korosztályokat jelölik, tehát ezek szolgáltatják a magyarázatot.

```
dm = VADeaths
class(dm)
barplot(dm) #A különböző korosztályokra vonatkozó
            #értékeket egymás fölé rendezi.
barplot(dm, beside = T) #A különböző korosztályokra
                        #vonatkozó értékeket
                        #egymás mellé rendezi.
legend("topright", legend = rownames(dm),
 fill = gray.colors(5))
```

A 2.10. ábrán látjuk a halálozási arányszámokat a korosztályok, nemek és települések szerint. Az ábra segít a különböző csoportok egymással való összehasonlításában.



2.10. ábra. *Virginia állam halálozási arányszámai (a* VADeaths *adatmátrix alapján)*

Ha az adatok `data.frame` típusúak (pl. df), akkor egy csoportváltozó (pl. F) szintjeinek az abszolút gyakoriságát a következő paranccsal ábrázolhatjuk: **barplot(table(df$F))**, a szintek relatív gyakoriságát pedig a következő paranccsal, ahol a **length()** a Faktorban szereplő adatok számát adja meg: **barplot(table(df$F)/length(df$F))**. Az *mtcars* adattábla *cyl* változója az amerikai márkájú autók hengereinek a számát tartalmazza, értékei: 4, 6, 8. Az alábbi kódot futtatva egy összetett ábrán láthatók a három típusú autó abszolút és relatív gyakoriságai.

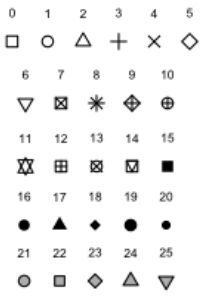
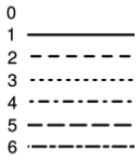
```
#Készítünk egy összetett ábrát
par(mfrow = c(1,2)) #két ábra egy sorba, két oszlopba
#Abszolút gyakoriságok
barplot(table(mtcars$cyl), xlab = "hengerek száma",
 ylab = "Abszolút gyakoriság")
```

```
#Relatív gyakoriság
barplot(table(mtcars$cyl)/length(mtcars$cyl),
 xlab = "hengerek száma",
 ylab = "Relatív gyakoriság"))
```

2.4.6. Diagramok formázási lehetőségei

Az ábrák számos paramétere tetszőlegesen beállítható. Ezek a beállítások működnek a **plot(),boxplot(),hist(),barplot()** parancsokon belül. A **?plot.default** paranccsal a Help menüben megjelenik számos részlet az ábrákkal kapcsolatban. A legfontosabb beállítások funkcióit a 2.8. táblázat tartalmazza.

2.8. táblázat. *A diagramok fontosabb formázási lehetőségei*

Formázási művelet	Funkció R-ben	Példa
Tengelyek elnevezése	xlab, ylab	xlab = "Ozone (ppb)"
Ábra elnevezése	main	main = "Airquality"
Vastagított betű: tengely neve, tengely skálája, ábra neve	font.lab, font. axis, font.main Lehetséges értékek: 1,2,3,4	font.lab = 2
Betűméret, számméret	cex.lab, cex.axis, cex.main alapérték: 1	cex.lab = 1.2 cex.axis = 0.9
Ábra színe Ábrák színe faktor szerint	col = " " col = faktor	col = "red" col = df$Month
Pontok típusa (pch)		pch = 16 pch = c(1,2,8,11)
Vonal típusa (lty)		lty = 5 lty = c(1,2,5)
Vonal vastagsága	lwd Alapérték: 1	lwd = 2

2.4.7. A kiugró értékek kezelése

Egy kiugró érték olyan érték az adatsorban, ami nem várt mértékben távolabb helyezkedik el a többi értéktől. Egy ilyen érték problémát okozhat azáltal, hogy befolyásolhatja a statisztikai próbákat, illetve a teljes kutatás végkövetkeztetését. Ugyanakkor érdekes adat is lehet, feltárhat „nem normális" (nem jellemző) eseteket a vizsgált populációban, azaz olyan eseteket, amelyek ritkán fordulnak elő. Amikor előfordul kiugró érték az adatsorban, döntést kell hozni a megtartásáról vagy az elvetéséről. A döntés meghozatalában ajánlott végigmenni a 2.11. ábrán feltüntetett logikai érvelésen.

1) A kiugró érték egy hiba következménye

Kiugró érték megjelenhet egy hiba elkövetésének következményeként. Ez a hiba lehet egy terepi elírás, egy mérési hiba, lehet a mintába került szennyezőanyag következménye vagy bármilyenféle hiba a mintatárolás, -előkészítés, -elemzés során. Ezek a hibák általában könnyen beazonosíthatók, különösen akkor, ha az adott érték nagyságrendekkel eltér a többitől. Ebben az esetben eltávolítható az adott objektum. Amennyiben nem lehet megtudni a hiba okát, akkor fel kell tennünk a következő kérdést:

2) A kiugró érték befolyásolja-e az eredményeket?

Ha a kiugró érték valós (a vizsgált populációból származik), akkor fel kell mérni azt, hogy lényegesen befolyásolja-e a kutatási eredményeket (például a függő változó átlagát, vagy egyéb mutatóját), vagy elméletileg nem magyarázható eredményhez vezetett.

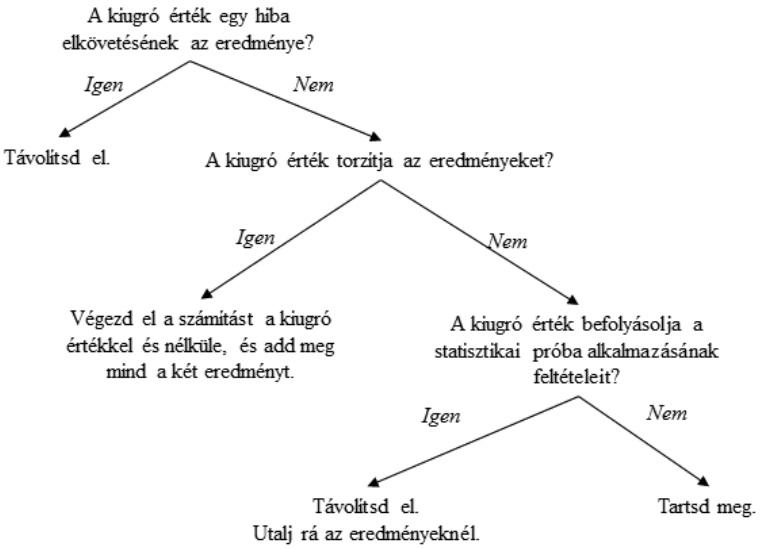
Ilyen esetben jó megoldás az, hogy feltüntetjük az eredményt kiugró értékkel és nélküle kiszámolva. Például a vizek vezetőképessége egyenes arányban van a bennük oldott sók mennyiségével. Egy kutatás során a 2.9. táblázatban látható eredményeket kapták kútvizekre.

2.9. táblázat. *Kútvizek kémiai adatai*

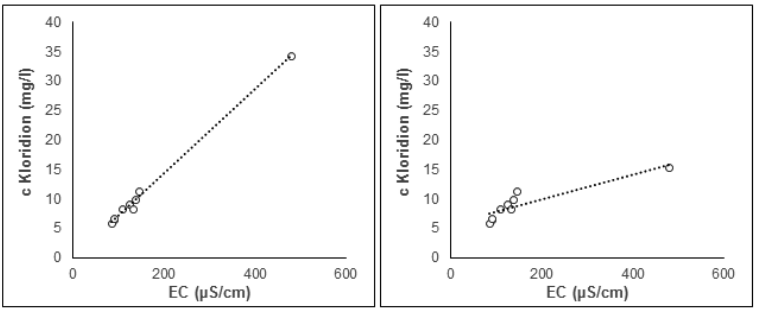
EC (μS/cm)	145	90	126	132	110	85	91	138	480
Cl^- (mg/l)	11.3	6.3	9.0	8.2	8.2	5.7	6.5	9.9	34.3

A 2.12. ábra bal oldalán látható a két kémiai paraméter közti összefüggés. A kiugró érték nem módosítja a többi pont által meghatározott lineáris összefüggést. Ezért a kiugró értéket meg lehet tartani. Érdemes utánaérdeklődni, hogy az utolsó mintában miért olyan magasak az értékek (pl. szennyezés eredménye lehet). A 2.12. ábra jobb oldalán egy olyan kiugró érték látható, ami jelentősen torzítja a többi pont által meghatározott lineáris összefüggést. Ebben az esetben ezt a pontot el lehet távolítani úgy, hogy az eredményeknél külön kitérünk rá. Statisztikai próbáknál, amelyek szignifikánsan

különböző eredményt adnak a kiugró érték jelenlétében és hiányában, tanácsos mind a két eredményt megadni, és meg kell indokolni, hogy melyiket tartjuk meg.



2.11. ábra. *A kiugró érték kezelésével kapcsolatos gondolatmenet*



2.12. ábra. *A kútvizek fajlagos vezetőképessége és kloridiontartalma közti kapcsolat*

3) *A kiugró érték befolyásolja-e a statisztikai próbák alkalmazhatóságát?*

Ha a kiugró érték nem befolyásolja az eredményeket, de megakadályozza egy parametrikus próba használatát, akkor két dolgot tehetünk:

– elvetjük, és megindokoljuk az eredményeknél az elvetés okát;

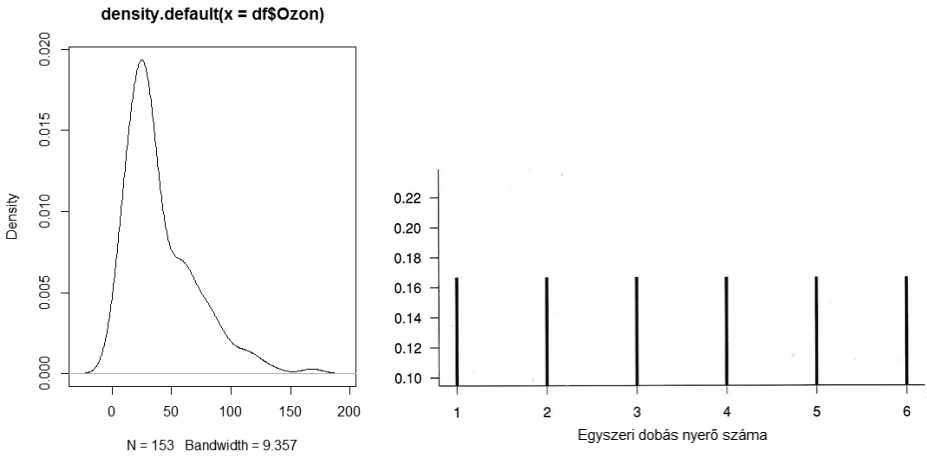
– átalakítjuk az adatsort: gyököt vonunk, vagy logaritmáljuk az adatokat.

Az utóbbi két módszer elvégzésével az adatsor normális eloszlást mutathat, így lehetővé válik a kiválasztott statisztikai próba elvégzése.

3. A VÁLTOZÓK ELOSZLÁSÁNAK VIZSGÁLATA

Az R software saját könyvtárában levő *airquality* adattábla ózon változójának koncentrációeloszlása látható a 3.1. ábrán. Az ábra egy sűrűségdiagramot mutat, amelynek vízszintes tengelyén a koncentrációskála, a függőleges tengelyen pedig a különböző koncentrációértékeknek az előfordulási valószínűsége látható. Ismerve az ózonkoncentráció sűrűségdiagramját, ezt a változót *valószínűségi változó*nak is tekinthetjük. A valószínűségi változó lényeges tulajdonsága az, hogy értékei egy számhalmaz elemei úgy, hogy a halmaz minden eleméhez és részhalmazához tartozik egy valószínűség. Ha az ózonkoncentráció értékkészletéből véletlenszerűen választunk ki egy mérési adatot, nagy esélyünk van arra, hogy ez az érték 30 ppb legyen, és sokkal kisebb esélyünk lenne arra, hogy ez az érték 150 ppb legyen.

Ha a változót sokszor megfigyeljük vagy megmérjük, akkor a megfigyelések vagy mérési eredmények ott helyezkednek el sűrűbben (nagyobb valószínűséggel rendelkeznek), ahol a sűrűségfüggvény értéke nagyobb. A valószínűségi változó lehet diszkrét, ha az értékkészlete véges vagy megszámlálhatóan végtelen; és folytonos, ha az értékkészlete végtelen, és az adott tartományban bármely valós számértéket tartalmazhat. Diszkrét valószínűségi változó pl. a dobókocka nyerő száma. A dobókocka számainak értékkészlete (Ω) hat egész számból áll: $\Omega = \{1, 2, 3, 4, 5, 6\}$. A sűrűségfüggvény ebben az esetben hat vonalból áll, és mindegyik számnak ugyanannyi a valószínűsége, hogy nyerő legyen: $1/6 = 0.167$ (3.1. ábra).



3.1. ábra. *Az ózonkoncentráció (balra) és a dobókocka nyerőszámainak (jobbra) sűrűségfüggvényei*

Ha ismerjük egy változó lehetséges értékeinek valószínűségi eloszlását, akkor matematikai függvénnyel, az ún. eloszlásfüggvénnyel jellemezhetjük. Ez a függvény nagy előrelépést jelent az adatfeldolgozásban, mivel az adott változó viselkedése matematikailag leírható, ennek ismeretében pedig egyszerű algoritmusú statisztikai tesztek alkalmazhatók, amelyeknek az eredménye könnyen kiértékelhető. A matematikai függvény állandói a változóra jellemző paraméterek (például az adatsor átlagértéke, standard deviációja, varianciája stb.).

3.1. Folytonos változók eloszlásának vizsgálata

3.1.1. Az egyenletes eloszlás

Az egyenletes eloszlás minden értékében maximummal rendelkezik. Ha egy [a,b] intervallumra értelmezzük, akkor az $x < a$ és $x > b$ (intervallumon kívüli) értékek gyakorisága nulla. Mivel minden x értékre, ami az [a,b] intervallumhoz tartozik, ugyanaz a gyakoriság jellemző, az eloszlás szimmetrikus az (a+b)/2 értékben emelt egyenesre. A 3.1. ábrán a dobókocka nyerőszámainak sűrűségfüggvénye diszkrét egyenletes eloszlás.

3.1.2. A normális (Gauss-féle) eloszlásfüggvény

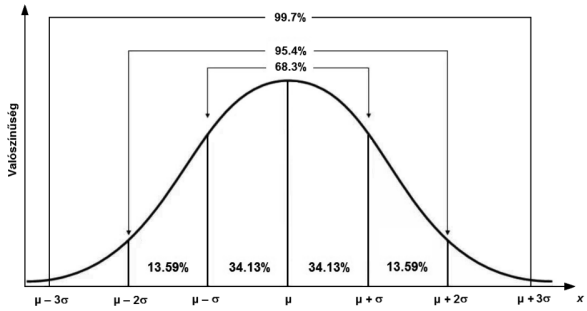
A természetben egy populációnak adott mérhető vagy valamilyen szinten kiértékelhető tulajdonsága általában normális eloszlást követ, vagy nagyon közel áll hozzá. Ilyen tulajdonságok például egy növényfaj, állatfaj testrészeinek a méretei, a felnőttek magassága, intelligenciaszintje, sportteljesítménye, talajvíz, ivóvíz, ásványvíz, esővíz kémiai összetevőinek a koncentrációja, a levegő hőmérséklete, nedvességtartalma egy adott időszakban stb.

A normális eloszlásfüggvény folytonos mennyiségi változók eloszlásának a jellemzésére alkalmas. Két állandója van: a populáció középértéke (μ) és standard deviációja (σ), jelölése pedig N(μ,σ). A középérték torzítatlan mutatója az átlagérték, ami ebben az esetben a mediánnal és a módusszal is egybeesik. A függvény matematikai kifejezése:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}.$$

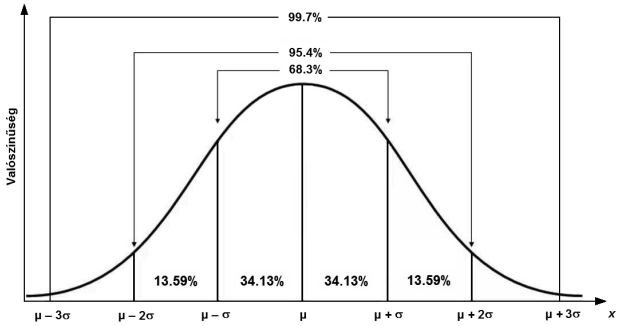
A függvény egy harang alakú görbe (3.2. ábra), a harang csúcsához (a maximális valószínűségi értékhez) az adatsor átlagértéke tartozik, a görbe alatti terület nagysága pedig 1. A függvény szimmetrikus az átlagértékben emelt merőlegesre.

A lefutó ágak aszimptotikusan közelítenek a vízszintes tengelyhez. Az áthajlási pontokhoz a $\mu-\sigma$ és $\mu+\sigma$ értékek tartoznak. A $\mu \pm \sigma$, $\mu \pm 2\sigma$ *és* $\mu \pm 3\sigma$ intervallumba tartozik az adatok 68.3%, 95.4% és 99.7%-a. A $\mu \pm 3\sigma$ intervallum gyakorlatilag magában foglalja a teljes adatsort. A kiugró értékek azonosítása ezen a kritériumon alapul, azaz mindazok az értékek, amelyek kívül esnek ezen a tartományon, nagy eséllyel nem tartoznak hozzá az adatsorhoz. A kritériumot 3σ (három szigma) kritériumnak is nevezik. Az adatsor várható minimum (MIN) és maximum (MAX) értékei tehát a $\mu \pm 3\sigma$ értékekkel tehetők egyenlővé. Ennek a két értéknek az ismerete lehetővé teszi a normális eloszlás két állandójának a kiszámolását, hiszen a két szám különbsége 6σ-val egyenlő. Innen kiszámolható a szórás ((MAX – MIN)/6)), illetve a középérték (MIN + 3σ vagy MAX–3σ) képletekkel.



3.2. ábra. *A normális eloszlásfüggvény (μ és σ a populáció középértéke és szórása)*

A 3.3. ábrán látható három normális eloszlásfüggvény, amelyeknek különböző a középértéke és a szórása. Mivel a görbe alatti terület nagysága változatlanul 1, következik, hogy minél kisebb a szórás, annál csúcsosabb a függvény.



3.3. ábra. *Három szilvafajta tömegének normális eloszlása*

Az alábbi kód egy olyan ábrát szerkeszt, amelyen három normális eloszlásfüggvény látható, nulla középértékkel és változó (1, 0.5 és 2) szórással.

```
curve(dnorm(x, 0,1), -5, 5, ylim=c(0,0.8),
 ylab="probability")
par(new=T) #Új görbét tesz ugyanarra az ábrára.
curve(dnorm(x, 0,0.5), -5, 5, ylim=c(0,0.8), lty=2,
 ylab="", xaxt="n", yaxt="n")
par(new=T)
curve(dnorm(x, 0,2), -5, 5, ylim=c(0,0.8), lty=3,
 ylab="",xaxt="n", yaxt="n")
legend(2,0.76, lty=c(1,2,3), bty = "n",
 legend = c(expression(sigma~"= 1"),
 expression(sigma~"= 0.5"),
 expression(sigma~"= 2")))
```

Ha ismerjük a populáció valamely normális eloszlású paraméterének az eloszlásfüggvényét, kiszámolható bármely érték előfordulási valószínűsége a populáció egyedeiben. Például ha tudjuk, hogy a ringló szilvafajta egyedeinek az átlagtömege 12 ± 4 g, és a tömegeloszlás normális, akkor a függvény egyenletéből kiszámolható valamely tömeghez tartozó valószínűség. Ez a számítás bonyolult és időigényes, sőt alapos matematikai tudást igényel, ezért a statisztikai adatfeldolgozásban egy egyszerűsített módszert használnak. Meghatároztak egy sablon- vagy mintafüggvényt, amelynek a független változója a *z*, és a neve a *standardizált normál eloszlásfüggvény*, *f*(*z*). A 3.4. ábra jobb oldalán látható ez a függvény, amelynek középértéke 0, szórása 1, a függvény egyenlete pedig:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}.$$

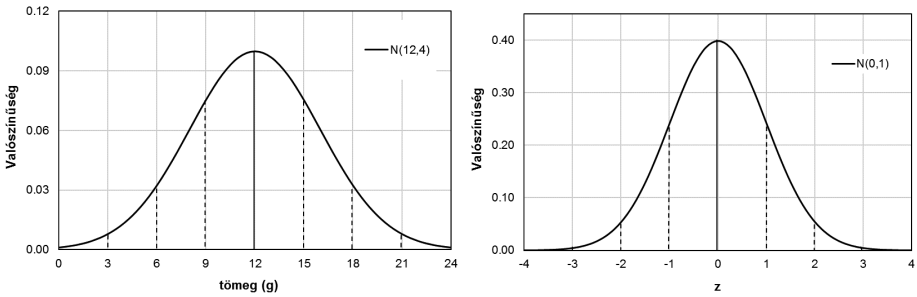
Bármely normális eloszlásfüggvény *x* paramétere átalakítható *z* paraméterré az ún. *z-transzformációval*.

$$z = \frac{x - \mu}{\sigma}$$

A statisztikusok által használt z-táblázat tartalmazza a (−∞, z] intervallum által meghatározott görbe alatti terület nagyságát, ami annak a valószínűségével egyenlő, hogy a z kisebb vagy egyenlő legyen az adott értékkel.

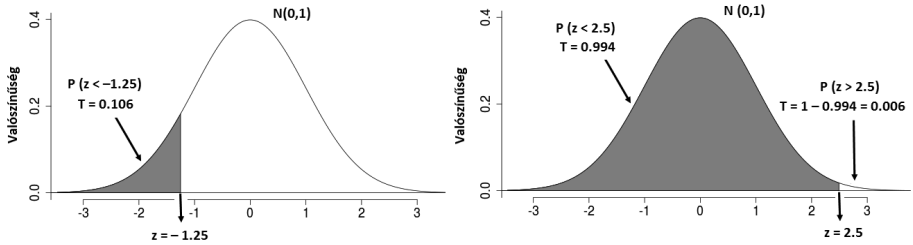
A standardizált normális eloszlásfüggvény segítségével választ adhatunk arra a kérdésre, hogy adott szilvafajta populációjából véletlenszerűen kiválasztva egy szilvát, mekkora a valószínűsége annak, hogy a tömege (m) kisebb legyen 7 g-nál, illetve nagyobb legyen 22 g-nál, ha a tömeg eloszlása N(12,4). Mind a két tömeg-

értéket átalakítjuk *z*-értékké a *z*-transzformáció segítségével. A 7 g-nak megfelelő *z*-érték (7−12)/4 = −1.25, a 22 g-nak pedig (22−12)/4 = +2.5.



3.4. ábra. *Az N(12,4) és az N(0,1) normális eloszlásfüggvények*

A valószínűség a (−∞, −1.25], illetve a [+2.5, +∞) intervallumok által határolt görbe alatti területek nagysága a standardizált normál eloszlásfüggvényen (3.5. ábra). A táblázat szerint a (−∞, −1.25] által határolt terület nagysága: 0.106, a (−∞, +2.5] által határolt területé pedig 0.994. A keresett valószínűségek: P(m < 7) = 0.106 (10.6%), P(m > 22) = 0.006 (0.6%).



3.5. ábra. *A z-táblázat által megadott valószínűségi értékeknek megfelelő területek a standardizált normális eloszlásfüggvényen (P – valószínűség; T – terület)*

A *https://gallery.shinyapps.io/dist_calc/* oldal lehetővé teszi a sajátos esetek szimulálását és a valószínűségi értékek kiszámítását.

3.1.3. A normális eloszlásfüggvény használata R-ben

R-ben az alapcsomag használatával előállíthatunk egy normális eloszlású adatsort tetszőleges átlaggal, szórással és egyedszámmal. Erre az általános parancs a következő: **`rnorm(n, mean, sd)`.**

Állítsunk elő egy *data_norm* vektort, ami 100 adatot tartalmaz egy normális eloszlásból, amelynek átlaga 32, szórása pedig 7.

```
data_norm = rnorm(100, 32, 7)
round(data_norm, 2) #Két tizedesre kerekített értékek.
[1] 26.47 36.54 28.52 24.38 33.40 32.45 33.77 43.44 24.31 ...

class(data_norm)
[1] "numeric"
```

A *data_norm* egy vektor, ami valós számokat tartalmaz. Tartalma annyiszor változik, ahányszor futtatjuk a **rnorm(100, 32, 7)** parancsot, mivel a megadott kritériumok alapján számtalan adatsor generálható. Ha szeretnénk állandósítani a *data_norm* tartalmát (hogy bármikor, bárhol ugyanazt a vektort állítsuk elő), akkor ezt megtehetjük a **set.seed()** paranccsal, amivel beállítható a kikért adatsor sorszáma.

```
set.seed(2345) #Mindig ugyanazt az adatsort tárolja.
data_norm = rnorm(100, 32, 7)
round(data_norm, 2)
[1] 23.66 35.84 31.56 33.86 30.36 25.02 22.21 ...
```

A **pnorm(x, mean, sd)** parancs megadja az N(átlag, sd) normál eloszlásfüggvény adott x értékéhez tartozó valószínűséget. A görbe alatti területet a $(-\infty, x]$ tartományra vonatkozik. Mekkora a valószínűsége annak, hogy az N(12,4) tömegeloszlású szilvafajtából véletlenül kiválasztott egyed tömege kisebb legyen 7 g-nál, ill. nagyobb legyen 22 g-nál?

```
pnorm(7, 12, 4)
[1] 0.10565
1-pnorm(22,12,4)
[1] 0.006210
```

Értelmezés szerint a standardizált normál eloszlásfüggvényt a **pnorm(z, 0, 1)** paranccsal használhatjuk. Ennek létezik egy egyszerűsített formája: **pnorm(z)**. Az előző példa két tömegértékének megfelelő z-érték a -1.25 és a $+2.5$.

```
pnorm(-1.25)
[1] 0.10565
1-pnorm(2.5)
[1] 0.006210
```

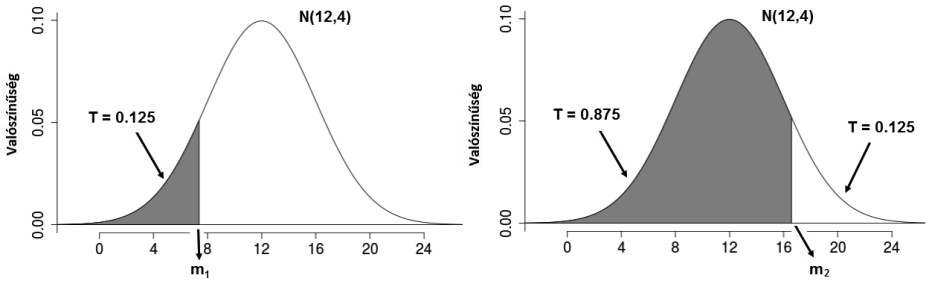
A kérdés fordítva is megközelíthető, azaz azt szeretnénk megtudni, hogy melyik *x*-értéknél nagyobb vagy kisebb értékek véletlenszerű megjelenésének a valószínűsége egy adott szám. Ennek a kérdésnek a megválaszolására a **qnorm(p, mean, sd)** parancsot használjuk. Például milyen tömegnél kisebb szilvák teszik ki a populáció 25%-át; mekkora tömegű szilvák teszik ki a populáció alsó és felső 5%-át; milyen tömegértékek között helyezkedik el a szilvák 75%-a?

```
qnorm(0.25, 12, 4)
[1] 9.3020
```

A második kérdésnél azokat a tömegértékeket keressük, amelyekre a görbe alatti terület 0.05, illetve 0.95 (1–0.05).

```
qnorm(0.05,12,4)
[1] 5.4206
qnorm(0.95,12,4)
[1] 18.5794
```

A harmadik kérdésnél a görbe alatti középső részt úgy kell behatárolnunk, hogy a területe 0.75 legyen. Mivel a görbe szimmetrikus a középértékben emelt függőlegesre, ezért ez a félegyenes két egyforma részre osztja a görbe alatti részt. A középérték két oldalán 0.75/2 = 0.375 területű részt kell elhatárolnunk. A kieső területek nagysága 1–0.75 = 0.25, ami egyforma két részt jelent, azaz 0.125 területű részt a két extrém szakaszban. Tehát keressük azokat a tömegértékeket, amelyek határolják az alsó és a felső szakaszban a 0.125 nagyságú területet: $T_1 = 0.125$, $T_2 = 0.875$ (3.6. ábra).



3.6. ábra. *Az N(12,4) normális eloszlású görbe alatti 0.125 és 0.875 nagyságú területeket meghatározó m_1 és m_2 pontok*

```
qnorm(0.125,12,4)
[1] 7.3986
qnorm(0.875,12,4)
[1] 16.6014
```

A három kérdésre adott válaszok:
1) a populáció 25%-át a 9.3 g-nál kisebb tömegű szilvák teszik ki;
2) az 5.4 g-nál kisebb és a 18.6 g-nál nagyobb tömegű szilvák teszik ki a populáció alsó és felső 5%-át;
3) a szilvák 75%-a 7.4 g és 16.6 g közötti tömeggel rendelkeznek.

3.1.4. A Student-féle t-eloszlásfüggvény

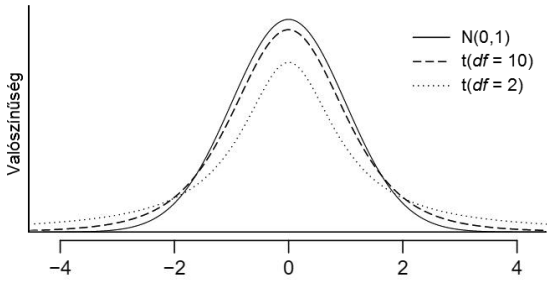
A bevezető részben tárgyaltunk arról, hogy a kísérleti kutatás során mintát veszünk abból a populációból, amellyel kapcsolatban megfogalmaztunk egy hipotézist, a mintára kapott eredmények alapján pedig megtartjuk vagy elvetjük azt. Tételezzük fel, hogy n elemszámú mintát veszünk egy normális eloszlású populációból, amelynek középértéke μ, szórása pedig σ. Ebben az esetben a mintaátlagból ($\bar{x}$), az n elemszámból és a populáció paramétereiből képezett z változó normális eloszlást követ, aminek középértéke 0 és szórása 1.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Általában a populáció szórása nem ismert, ezért azt a minta szórásával: s tudjuk megbecsülni. Lényeges különbség a σ és az s között, hogy amíg a σ egy szám, addig az s egy valószínűségi mutató (értéke mintánként változik), aminek a lehetséges értékeit egy valószínűségi függvény írja le. A módosult képlet által kifejezett paraméter nem követi a normális eloszlást, hanem az ún. t-eloszlást.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Az $s/\sqrt{n}$ a standard hibát (standard error, SE) fejezi ki, ami a mintaátlag szórásának felel meg.



3.7. ábra. *A* z- *és a* t-*eloszlás egymáshoz való viszonyulása (a* df *szabadságfokokat jelent, és függ a minta elemszámától)*

A *t* és a *z* valószínűségi változókat kifejező képletek nagyon hasonlítanak egymásra, ezért a *t*-eloszlás is hasonlít a normális eloszlásra, viszont a harang alakú görbe lapultabb, a lefutó ágak szélei alatti területek pedig nagyobbak (3.7. ábra). A *t*-eloszlás függ a mintaszámtól, ezért minden *n* értékre más *t*-eloszlás érvényes, amelynek szabadságfoka ($n-1$)-gyel egyenlő.

3.1.5. A Student-féle t-eloszlásfüggvény alkalmazása R-ben

R-ben az alapcsomag használatával előállíthatunk adott *df* szabadságfokkal rendelkező *t*-eloszlású adatsort *n* elemmel. Erre az általános parancs a következő: **`rt(n, df)`**. A testmagasság példájánál maradva, generálhatunk egy *tdistr* vektort, ami 5000 mintavétel átlagait tartalmazza, ha az összes minta elemszáma $n = 50$ ($df = 49$), a populáció szórását pedig nem ismerjük.

```
tdistr = rt(5000, 49)
d = density(tdistr) #elkészítjük a sűrűségdiagramot
plot(d) #ábrázoljuk
```

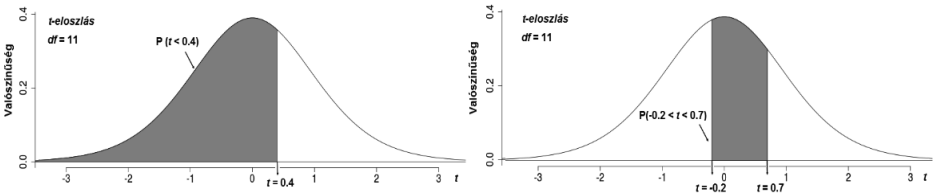
A **`pt(q, df)`** parancs megadja a *df* szabadságfokú *t*-eloszlásfüggvény adott *t* értékéhez tartozó valószínűséget. A görbe alatti terület a $(-\infty, t]$ tartományra vonatkozik. A művelet egyezik a **`pnorm()`** művelettel, azzal a különbséggel, hogy az R a *t*-eloszlásfüggvénnyel számol. Mekkora a valószínűsége annak, hogy egy *t*-érték 0.4 alatt, illetve $(-0.2, 0.7)$ közé essen, ha a $df = 11$ (3.8. ábra)?

```
pt(0.4,11)
[1] 0.652
```

A valószínűsége annak, hogy $t < 0.4$-nél 65.2%. A $t_{t=0.7} - T_{t=-0.2}$ különbséggel fejezhetjük ki.

```
T0.7 = pt(0.7,11)
T0.7
[1] 0.751

T0.2 = pt(-0.2,11)
T0.2
[1] 0.423
T = T0.7-T0.2 #a keresett terület
T
[1] 0.328
```

Annak a valószínűsége, hogy t értéke $(-0.2, 0.7)$ intervallumba essen, 32.8%.



3.8. ábra. *A* t-*eloszlás (*df = *11) adott* t-*értékéhez, illetve* t-*értékei közé eső görbe alatti területek*

A **qt(p, df)** paranccsal azt a kritikus t_k-értéket kérhetjük ki, amelyre a $P(t < t_k) = p$. Például melyik az a t_k-érték, amelyre a $t < t_k$ valószínűsége $p = 25\%$, ha $n = 5$, illetve $t > t_k$ valószínűsége 33%, ha $n = 25$?

```
qt(0.25,4)
[1] -0.74 #t_k = -0.741
```

A második esetben a t_k által határolt jobb oldali területről van szó, amelynek nagysága T = 0.33. A **qt()** parancs a $(-\infty, t]$ tartományra vonatkozó területtel számol, amely ebben az esetben 1–T = 0.67, a *df* pedig 24.

```
qt(0.67,24)
[1] 0.45 #t_k = 0.45
```

A konfidenciaintervallum kiszámítására nincs parancs az alapcsomagban, de vannak különféle csomagok, amelyekben van lehetőség a konfidenciaintervallum kiszámítására, pl. {Rmisc} csomag **CI()** parancsa. A lépéseket elvégezhetjük lépésről lépésre, vagy írhatunk egy függvényt, ami bármely esetben megadja a kívánt paramétereket. Adjuk meg a férfiak testmagasságának középértékbecslését konfidenciaintervallummal, amit $n = 50$ elemszámú mintára kaptunk, ha az átlag 172.0 cm, a szórás pedig 4.9 cm.

```
KI = qt(0.975,49)*4.9/sqrt(50)
KI
[1] 1.39
also = 172-KI
also
[1] 170.6
felso = 172+KI
felso
[1] 173.4
```

Ezek a lépések akár a Microsoft Excellel is elvégezhetők. Az R-ben rejlő lehetőségek egyike, hogy ezeket a lépéseket lerövidíthetjük és általánosíthatjuk egy saját függvénnyel. Legyen a függvényünk neve *konf_int*, ami négy paraméterrel (atlag, szoras, msz, alfa) számol, ahol *msz* a minta elemszámát jelöli, az *alfa* pedig a szignifikanciaszintet. Az alfa értékét alapból beállíthatjuk 0.05-re, de a parancs alkalmazásánál tetszőlegesen változtathatjuk.

```
konf_int = function(atlag, szoras, msz, alfa = 0.05) {
 tort=round(qt(1-(alfa/2),msz-1)*szoras/sqrt(msz),1)
 KI = tort*2
 also = atlag-tort
 felso = atlag+tort
 return(data.frame(KI,also,atlag,felso,alfa))
}
```

Miután a függvényt lefuttattuk, az R parancsként megjegyzi, és amíg nem zárjuk le a programot, bármikor futtathatjuk. A függvényt elmenthetjük scriptként (pl. konf_int.R), amit később bármikor használhatunk. A **konf_int(atlag, szoras, msz, alfa)** formában futtatott parancs megadja a KI-t, a KI alsó és felső határát, az átlagot, valamint az alfa értékét.

```
#ha alfa = 0.05, nem kell beírni a parancsba
konf_int(172,4.9,50)
   KI  also atlag felso alfa
1 2.8 170.6   172 173.4 0.05

konf_int(172,4.9,50,0.01) #alfa = 0.01
   KI  also atlag felso alfa
1 3.8 170.1   172 173.9 0.01

#három alfa érték
konf_int(172,4.9,50, c(0.01,0.05, 0.1))
   KI  also atlag felso alfa
1 3.8 170.1   172 173.9 0.01
2 2.8 170.6   172 173.4 0.05
3 2.4 170.8   172 173.2 0.10
```

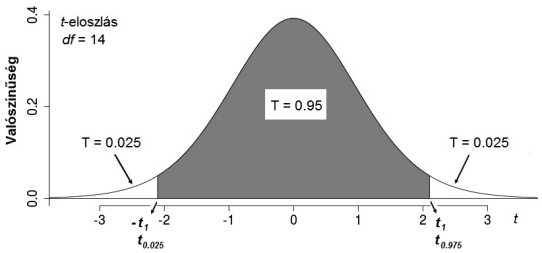
3.1.6. A populáció középértékének és szórásának becslése

A populáció középértékét és szórását a minta átlagával és szórásával becsülhetjük meg. Szem előtt kell tartanunk, hogy mind a két paraméter valószínűségi változó. Ez azt jelenti, hogy többszörösen megismételt mintavétel átlagainak eloszlása valószínűségi függvénnyel írható le. Az átlagok átlaga ($\bar{X}$) várhatóan nagyon közel áll a populáció középértékéhez. Ha ismert a populáció szórása, akkor a mintaátlagok normális eloszlásúak lesznek, ha a szórás nem ismert, akkor *t*-eloszlást követnek. Figyelembe véve a mintavétel és mintaelemzés sokszor időigényes és költséges voltát, a többszörös mintavétel módszere, habár jó megközelítést biztosít a populáció középértékéhez, a gyakorlatban mégsem járható út. Elfogadott módszer az egyszeri mintavétel, amelynek segítségével a populáció középértékét a mintaátlaggal, valamint a mintaszórás és -elemszám alapján meghatározott megbízhatósági (konfidencia) intervallummal fejezzük ki. Minél kisebb a konfidenciaintervallum (KI), annál biztosabbak vagyunk abban, hogy a mintaátlag közel áll a populáció középértékéhez. A populáció középértékét az alábbi összefüggésből fejezzük ki:

$$\mu = \bar{x} \pm t_{\alpha=0.05} \cdot \frac{s}{\sqrt{n}},$$

ahol a $\pm t_{\alpha=0.05} \cdot s/\sqrt{n}$ kifejezés megadja a konfidenciaintervallum határait, az átlagérték pedig az intervallum közepén helyezkedik el. Az egyenlet alapján megértjük azt, hogy a t-eloszlás tulajdonképpen az n elemszámú számtalan minta átlagainak az eloszlását mutatja, amelyekhez az $s/\sqrt{n}$ szórás tartozik, és amelyiknek a középértéke az általunk vett minta átlaga. Az $s/\sqrt{n}$ szórást az átlag standard hibának nevezzük, és SE formában jelöljük (SE – standard error), ez megmutatja, hogy a minta átlaga mennyire tér el a populáció átlagától, a várható értéktől (μ).

A fenti egyenlet rámutat egy nagyon fontos alapigazságra a tudományos kutatások korlátoltságával kapcsolatban: a populáció várható értékét nem tudjuk egyértelműen meghatározni (például a mintaátlaggal), csak becsülni tudjuk pl. a mintaátlaggal, és csupán egy intervallumot adhatunk meg, ami a *t* értéke alapján növelhető vagy csökkenthető, és ami adott bizonyossággal tartalmazza a μ-t. A bizonyosság mértékét általában 95%-ra teszik. Megfordítva, a bizonytalanság 5%, amit szignifikanciaszintnek nevezünk ($\alpha = 5\%$ vagy $\alpha = 0.05$). Ennek az értéknek statisztikai alapja van, optimálisnak bizonyult kétféle hiba kiegyensúlyozásában. A 95%-os bizonyosságot biztosító *t*-értéket hasonló módon állapítjuk meg, mint a normális eloszlásnál a *z*-értéket. A *t*-eloszlásfüggvényen keressük a középső 0.95 nagyságú területet határoló t_1 és t_2 értéket (3.9. ábra). Mivel a függvény szimmetrikus harang alakú, a $t_2 = -t_1$. A *t* értéket kikereshetjük táblázatból, vagy használhatjuk az R-t (lásd a 2.5.4. alfejezetet). A *t* értékeket általában úgy jelölik, hogy indexben feltüntetik a $(-\infty, t]$ intervallum által meghatározott görbe alatti terület nagyságát. Mivel az *n* értéke szerint változik a *t*-eloszlásfüggvény, a szabadságfokok számát is feltüntetik. Így a $-t_1$ megszokott jelölése $t_{0.025,14}$, a t_1-é pedig $t_{0.975,14}$.



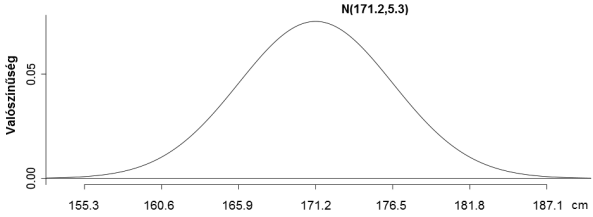
3.9. ábra. *A* t*-eloszlásfüggvény, ha a minta elemszáma 15 (*df *= 14). A konfidenciaintervallum meghatározásához keressük a* t_1 *és* $-t_1$ *értékeket, amelyek meghatározzák a T = 0.95 területet a* t *= 0 érték körüli részben.*

Példa: férfiak átlagos testmagasságának becslése

Világviszonylatban (az országos és néhány régió hivatalos adatai szerint, N = 200) a férfiak átlagmagassága 1996-ban 171.2 cm volt, 5.3 szórással (*https://ourworldindata.org/human-height*). A populáció eloszlása megközelítően az N(171.2, 5.3) normális eloszlással jellemezhető (3.10. ábra). Tételezzünk fel két esetet:

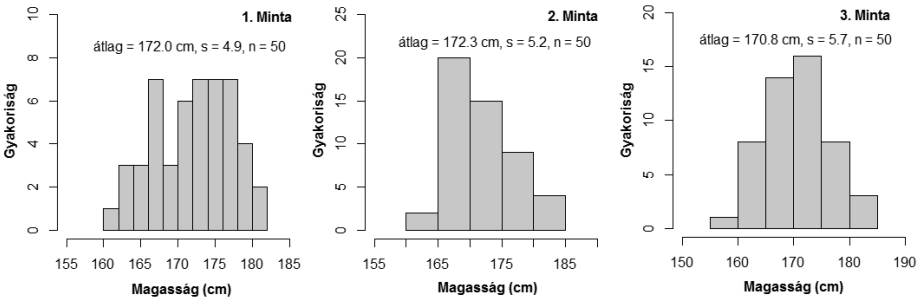
a) nem ismerjük a μ-t, de a σ ismert;

b) nem ismerjük a μ-t és a σ-t sem.



3.10. ábra. *A férfiak testmagasságának megközelítő normális eloszlása:* $\mu = 171.2$ *cm,* $\sigma = 5.3$ *cm*
(1996-os adatok, forrás: https://ourworldindata.org/human-height*)*

A mintavételt m-szer ismételjük meg, így m mintaátlagunk és m mintaszórásunk lesz. Minden esetben $n = 50$ férfit véletlenszerűen választunk ki. A 3.11. ábrán látható az első három mintavétel eredménye: az átlagok és a szórások eltérőek, de közel állnak a valós értékekhez. Ugyanakkor az egyes mintákon belüli magasságértékek eloszlása többé-kevésbé normálisnak tűnik.



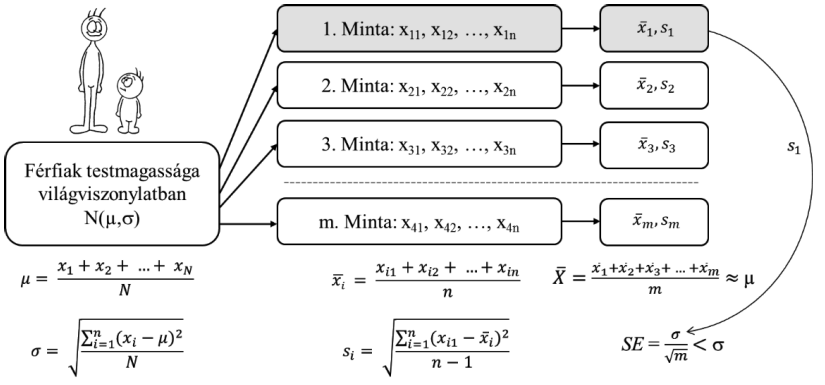
3.11. ábra. *Három mintavétel eredménye. Minden esetben az elemszám 50.*

a) *Nem ismerjük a μ-t, de a σ ismert*

Az m minta átlagainak az eloszlása a N(μ, $\frac{\sigma}{\sqrt{m}}$) normális eloszlást követi, ahol az átlagok átlaga $\bar{X} \approx \mu$ (3.11. ábra). A $\sigma/\sqrt{m}$ a mintavételek standard hibája, ami kisebb a populáció szórásánál, ha $m > 1$. Tehát a mintaátlagok eloszlása csúcsosabb a populáció eloszlásához viszonyítva. A mintaátlagok kisebb szórása érthető, hiszen a populációhoz tartozó extrém magas és alacsony értékek az átlagok eloszlásában hiányoznak. Az átlagok eloszlása annál csúcsosabb, minél nagyobb a minta elemszáma. A testmagasságok eloszlásának szórása 5.3 cm, $m = 10$ mintavétel esetében ez az érték 1.7 cm-re csökken, $m = 100$ esetén 0.53 cm-re.

b) *Nem ismerjük a μ-t és σ-t*

Az m minta átlagainak az eloszlása a t(μ, $\frac{S}{\sqrt{m}}$) t-eloszlást követi, ahol az átlagok átlaga $\bar{X} \approx \mu$ (3.12. ábra). A σ ismeretének hiányában a szórást a mintaátlagok S szórásával becsüljük, a mintaátlagok szórása pedig $\frac{S}{\sqrt{m}}$-mel fejezhető ki, ami a mintavételek standard hibája.



3.12. ábra. *A populáció, az egyedi minták és a mintaátlagok statisztikai mutatói*

A gyakorlatban egyetlen mintát veszünk (pl. az 1. minta), amelynek eredménye: $\bar{x} = 172.0$ cm, $s = 4.9$ cm, $n = 50$. Ebben az esetben $\mu \neq \bar{x}$, hanem $\mu = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$, ha σ ismert, és $\mu = \bar{x} \pm t_{0.05} \cdot \frac{s}{\sqrt{n}}$ ($df = 49$), ha σ nem ismert. A populáció középértékét mind a két esetben a konfidenciaintervallum segítségével becsüljük meg, 5%-os bizonytalanság mellett ($\alpha = 0.05$).

a) *Konfidenciaintervallum megadása, ha σ ismert*

Az átlagok eloszlása normális, tehát a *z*-eloszlásfüggvényen keressük azt a $\pm z$ értéket, amely a középső $T = 0.95$ területet határolja. Ez a *z*-érték ± 1.96. Ha 99%-ban szeretnénk bizonyosak lenni, hogy a KI tartalmazza μ-t ($\alpha = 0.01$), akkor a *z*-érték nyilvánvalóan nagyobb (2.576), így a KI is szélesebb lesz.

$$\mu = 172.0 \pm 1.96 \cdot \frac{5.3}{\sqrt{50}} \text{ cm } (\alpha = 0.05)$$
$$\text{KI} = (170.5, 173.5), \alpha = 0.05$$
$$\text{KI} = (169.4, 174.6), \alpha = 0.01$$

A konfidenciaintervallumnak akkor van értelme, ha feltüntetjük a bizonytalansági (szignifikancia) szintet. Ezzel azt fejeztük ki, hogy 95%-ban bizonyosak vagyunk abban, hogy $\mu \in (170.5, 173.5)$. Mivel ebben a sajátos esetben ismerjük a μ-t (171.2), biztosan tudjuk, hogy a KI valóban tartalmazza μ-t.

b) *Konfidenciaintervallum megadása, ha nem ismerjük σ-t*

Az átlagok eloszlása a $df = 49$ szabadságfokkal jellemzett *t*-eloszlást követi, ezért keressük azt a $\pm t$ értéket, amely a középső $T = 0.95$ területet határolja. Ez a *t*-érték ± 2.01. A $T = 0.99$ területet határoló *t*-értékek ± 2.68.

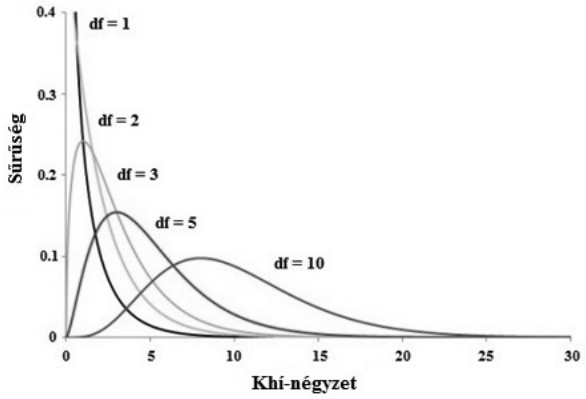
$$\mu = 172.0 \pm 2.01 \cdot \frac{4.9}{\sqrt{50}} \text{ cm } (\alpha = 0.05)$$
$$\text{KI} = (170.6, 173.4), \alpha = 0.05$$
$$\text{KI} = (170.1, 173.9), \alpha = 0.01$$

Tehát 95%-ban, illetve 99%-ban bizonyosak vagyunk abban, hogy a (170.6, 173.4) intervallum, illetve a (170.1, 173.9) intervallum tartalmazza a μ-t. Ezt úgy is értelmezhetjük, hogy ha 100-szor végeznénk el a mintavételt, várhatóan 95 esetben a konfidenciaintervallum tartalmazni fogja a μ-t.

3.1.7. A χ^2 (khí-négyzet) eloszlás

A χ^2-eloszlás a görög χ betűről kapta. Folytonos valószínűségi eloszlás, ami szorosan kapcsolódik a normális eloszláshoz. Ha a *z* változó normális eloszlást követ, akkor a z^2 χ^2-eloszlással jellemezhető, amelynek egy szabadságfoka van (3.13. ábra). Ha $z_1, z_2, \ldots, z_k$ változók normális eloszlást követnek, akkor a $z_1^2 + z_2^2 + \ldots + z_k^2$ változó

χ^2-eloszlást követ k szabadságfokkal. Mivel négyzetek összegéből áll, ezért értékei a $[0, \infty)$ intervallumban helyezkednek el. Ha a $k = 1$, a függvény aszimptotikusan közelít a nullához.



3.13. ábra. *A χ^2-eloszlás különböző szabadságfokokkal*

A χ^2-eloszlás paraméterei: $\mu = k$, $\sigma^2 = 2k$, ahol k a szabadságfokok számát jelenti. A görbe a maximumát a $\mu-2$ értékben éri el, ha df > 3. Például a $k = 10$ függvény maximuma 8-ban van, középértéke 10, szórása pedig 4.5. Ahogy nő a k értéke, úgy egyre nő a szórás, és egyre lapultabb a függvény. A szabadságfokok számának növekedésével az eloszlás formája megváltozik, a középérték nagyobb értékek felé mozdul el, a jobb oldali ferdeség pedig csökken, az eloszlás egyre inkább hasonlít a normális eloszláshoz.

A χ^2-eloszlást hipotézisek tesztelésére használjuk, például tesztelhetjük vele, hogy egy adatsor egy adott parametrikus eloszlásból (pl. binomiális, normális, Poisson stb.) származik-e. Ezt a próbát az illesztés jóságának (goodness of fit) nevezzük. Ezt az alkalmazást részletesen kifejtjük a 7.5.1. alfejezetben.

3.2. Diszkrét változók eloszlásának vizsgálata

3.2.1. A binomiális eloszlás

A binomiális eloszlás nominális változók viselkedését írja le abban az esetben, ha a változónak csak két, egymást kölcsönösen kizáró értéke van. Ezeket bináris változóknak nevezzük. A változót gyakran eseménynek tekintik, amelynek két, egymást kizáró értéke a siker és a kudarc (jelenlét és hiány, igen és nem). A siker

azt jelenti, hogy egy vizsgált objektumon/egyeden a várt esemény bekövetkezett, a kudarc pedig azt, hogy nem. Az objektumok/egyedek száma a próbálkozások számát jelenti. A binomiális eloszlás általánosítja egy mintán megfigyelt sikerek gyakoriságát tetszőleges egyedszámra (próbálkozásra), ugyanakkor megadható k tetszőleges siker bekövetkezésének valószínűsége is egy adott n értékre. Jelölése: B(n, p), ahol n a próbálkozások száma, p pedig a sikerek bekövetkezésének valószínűsége. Matematikai kifejezése megadja egy X változóra a k-szor bekövetkezett siker valószínűségét:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k},$$
$$k \in \{0, 1, 2, 3, \ldots, n\},$$
$$0 < p < 1,$$

ahol k a sikerek száma n próbálkozásból, p pedig a siker bekövetkezésének valószínűsége. Az $\binom{n}{k}$ a kombinatorikában használt kifejezés, jelentése:

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

ahol $n!$ (n-faktoriális) = $1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$. A tört megmutatja, hogy hányféleképpen lehet az n értékkészletéből kiválasztani k darabot, ha nem számít a kiválasztás sorrendje, és mindegyik értéket csak egyszer választhatjuk.

Egy pénzérme feldobásából két esemény következhet be: fej vagy írás. A két esemény kölcsönösen kizárja egymást. Ha a fej eseményt tekintjük sikeres dobásnak, akkor az érme egyszeri feldobásával a fej esemény valószínűsége 50% (p = 0.5). Ha a próbálkozások száma 30, akkor a sikeres esemény bekövetkezésének valószínűségét a B(30, 0.5) eloszlásfüggvény írja le:

$$P(X = k) = \binom{30}{k} \cdot 0.5^k \cdot 0.5^{30-k}.$$

Az egyenlet alapján kiszámolható annak a valószínűsége, hogy 30 próbálkozásból 5, 12, 20, 27 alkalommal következzen be a siker.

$k = 5$ esetén $P(5) = \binom{30}{5} \cdot 0.5^5 \cdot 0.5^{25}$, innen $P(5) = \frac{26 \cdot 27 \cdot 28 \cdot 29 \cdot 30}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} \cdot 0.5^{30} = 0.00013$.

A fenti példához hasonlóan a következő valószínűségeket kapjuk a k (5, 12, 20, 27) értékekre: 0.013%, 8.1%, 2.8%, 0.0004%. Beláthatjuk, hogy a legvalószínűbb az, hogy a 30 dobásból 15 lesz fej, hiszen a fej bekövetkezésének valószínűsége 0.5 (3.14. ábra). A felsorolt értékek közül a 12 sikeres dobás áll legközelebb a 15-höz, ezért ennek a legnagyobb a valószínűsége (8.1%). Ha azt szeretnénk megtudni, hogy mekkora a valószínűsége annak, hogy legfeljebb 12-szer dobjunk fejet, akkor:

$$P(k \leq 12) = P(0) + P(1) + P(2) + \ldots + P(12).$$

Ilyen esetben ki kell számolni az összes esemény valószínűségét, ami eleget tesz a feltételnek.

$$P(0) = \binom{30}{0} \cdot 0.5^0 \cdot 0.5^{30} = 9 \cdot 10^{-10}$$

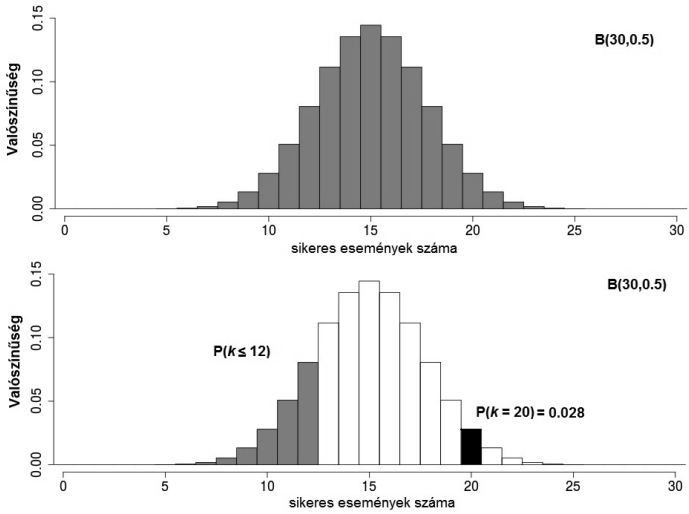
$$P(1) = \binom{30}{1} \cdot 0.5^1 \cdot 0.5^{29} = 2.8 \cdot 10^{-8}$$

$$\dots$$

$$P(12) = \binom{30}{12} \cdot 0.5^{12} \cdot 0.5^{18} = 8.1 \cdot 10^{-2}$$

$$\mathrm{P}(k \leq 12) = 0.181$$

Annak a valószínűsége, hogy legfeljebb 12-szer dobjunk fejet a 30 próbálkozásból, 18.1%. A feladat hosszadalmas, a számítást sokkal egyszerűbb elvégezni R-ben.

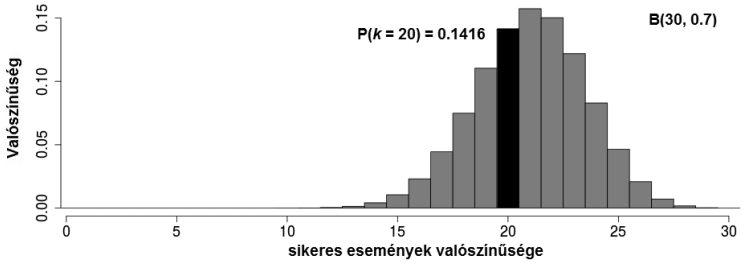


3.14. ábra. *A B(30,0.5) binomiális eloszlás (n = 30, p = 0.5)*

Egy hamis érmével növeljük az esélyét annak, hogy a vártnál többször dobjunk fejet. Ha az érmét úgy hamisítjuk, hogy a fejnek a valószínűsége 0.7 legyen, akkor annak az esélye, hogy 20-szor dobjunk fejet a 30-ból, 2.8%-ról 14.2%-ra nő.

$$P(12) = \binom{30}{12} \cdot 0.7^{12} \cdot 0.3^{18} = 0.1416$$

Ebben az esetben a B(30,0.7) eloszlással jellemezhetjük a sikerek valószínűségét (3.15. ábra).



3.15. ábra. *A B(30,0.7) binomiális eloszlás (n = 30, p = 0.7)*

3.2.2. A binomiális eloszlás alkalmazása R-ben

Egy vagy több esemény bekövetkezésének valószínűsége a **dbinom(k, n, p)** paranccsal kérhető ki. Mekkora a valószínűsége annak, hogy 30 próbálkozásból 5-ször dobjunk fejet egy érmével? A fej esemény valószínűsége 0.5.

```
p5 = dbinom(5, 30, 0.5)
p5
[1] 0.0001327191
```

Mekkora a valószínűsége annak, hogy 30 próbálkozásból 5-ször, 12-szer, 20-szor, 27-szer dobjunk fejet egy érmével?

```
p = dbinom(c(5,12,20,27),30,0.5)
[1] 1.327191e-04 8.055309e-02 2.798160e-02 3.781170e-06
```

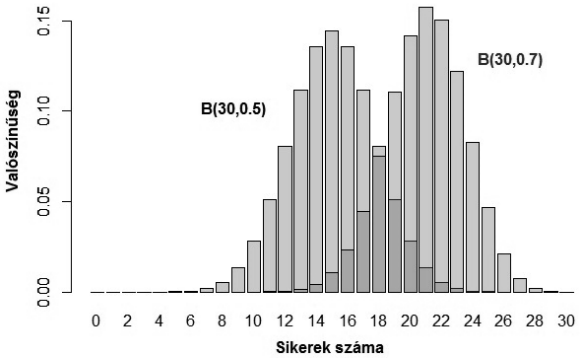
Mekkora a valószínűsége annak, hogy 30 próbálkozásból legfeljebb 12-szer dobjunk fejet egy érmével?

```
#a {0,1,..,12} sikerek valószínűségeinek összege
sum(dbinom(0:12,30,0.5))
[1] 0.1807973
```

Az előbbi parancs egyszerűbben is felírható a **pbinom(k, n, p)** formában, amely a {0,1,..,12} sikereses események kumulatív valószínűségét számolja ki.

```
pbinom(12,30,0.5)
[1] 0.1807973
```

A binomiális eloszlás ábrázolása R-ben oszlopdiagramként valósítható meg, a **barplot()** paranccsal. Egy ábrán több eloszlás is feltüntethető, például a B(30,0.5) és a B(30,0.7) eloszlások, amelyek megfelelnek az igazi és a hamis érmékkel való sikeres dobások eloszlásának (3.16. ábra).



3.16. ábra. *A B(30,0.5) és a B(30,0.7) binomiális eloszlások ábrázolása R-ben*

A 3.16. ábra R-kódja:

```
barplot((dbinom(0:30,30,0.5)), xlab ="Sikerek száma",
 ylab = "Valószínuség", font.lab = 2,
 names = 0:30, ylim = c(0,0.16),
 col = rgb(100,100,100, alpha = 70, max = 255))
barplot((dbinom(0:30,30,0.7)), xlab ="Sikerek száma",
 ylab = "Valószínuség", font.lab = 2,
 ylim = c(0,0.16),
 col = rgb(30,130,250, alpha = 80, max = 255),
 add = T)
```

3.2.3. A binomiális eloszlás megközelítése a normális eloszlással

A 3.13. és 3.14. ábrákon látható binomiális eloszlások harang alakúak és szimmetrikusak a maximum értéken átmenő függőlegesre, ezért a binomiális eloszlás megközelíthető a normális eloszlásfüggvénnyel. Ennek a jelentősége abban mutatkozik meg, hogy a diszkrét eloszlás jó megközelítéssel helyettesíthető egy folytonos eloszlással, amire természetesen alkalmazhatók a folytonos eloszlás törvényei.

A B(n,k) eloszlást egy olyan normális eloszlásfüggvénnyel írhatjuk le, amelynek középértéke

$$\mu = p \cdot n,$$

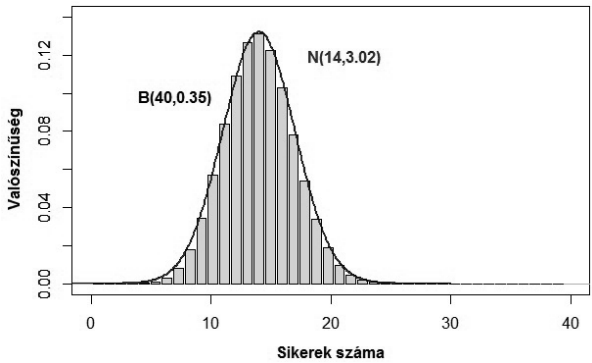
szórása pedig

$$\sigma = \sqrt{p \cdot (1-p) \cdot n}.$$

A binomiális eloszlás normális eloszlással való megközelítésének feltételei vannak, amelyek elsősorban azon alapulnak, hogy az eloszlásnak szimmetrikusnak kell lennie. Több javaslat született ezzel kapcsolatban, amit Emura és Liao (2017) összesített. Noha engedékenyebb feltételek mellett szokták alkalmazni a közelítést, a legmegbízhatóbb feltételek a következők: $n \cdot p \geq 10$ és $p \geq 0.1$, vagy $n \cdot p \geq 15$ (Emura és Lin, 2015).

Például a B(30,0.5) eloszlás az N(15,2.74) eloszlással közelíthető meg, mivel $30 \cdot 0.5 = 15 > 10$, a B(10,0.7) eloszlás viszont nem közelíthető meg a normális eloszlással, mivel $10 \cdot 0.7 = 7 < 10$. A művelet végrehajtásához növelni kell az n értékét, például 15-re ($15 \cdot 0.7 = 10.5$), ahol már a B(15,0.7) eloszlás megközelíthető az N(10.5,1.77) normális eloszlással.

A B(40,0.35) binomiális eloszlás megközelíthető az N(14,3.02) normális eloszlással, mivel $p \cdot n = 0.35 \cdot 40 = 14 > 10$, $(1-p) \cdot n = 0.65 \cdot 40 = 26 > 10$. A számított középérték 14, a számított szórás pedig 3.02. A két eloszlás a 3.17. ábrán látható.



3.17. ábra. *A B(40,0.35) eloszlás megközelítése az N(14,3.02) eloszlással*

A két eloszlás R-kódja a következő:

```
#generálunk 10e+07 adatot az N(14,3.02) eloszlásból
dn = density(rnorm(10000000,14,3.02))
plot(dn, lwd = 2, col = "blue", ylim = c(0,0.14),
   xlim = c(0,40), xlab = "Sikerek száma",
   ylab = "Valószínuség", font.lab = 2, main = "")
barplot(dbinom(0:40,40,0.35), xlab ="Sikerek száma",
 ylim = c(0,0.14), ylab = "Valószínuség",
 font.lab = 2, xlim = c(0,40), width = 0.8,
```

```
    col = rgb(100,100,100, alpha = 60, max = 255),
    add = T)
#ha add = T, akkor a második ábrát ráteszi az elsőre.
```

3.2.4. A Poisson-eloszlás

A Poisson-eloszlás μ átlagértékkel rendelkező diszkrét változó valószínűségi eloszlását írja le. Hasonlóan a binomiális eloszláshoz, a változó itt is eseményként fogható fel, de ennek a bekövetkezése egy próbálkozás alatt többször megtörténhet. Például egy faj egyedeinek a száma egy mintavételi egységben, beteg egyedek száma adott megfigyelési helyen, áthaladó járművek száma egy adott megfigyelési ponton adott időegység alatt stb. A Poisson-eloszlás a következő matematikai függvénnyel írható le:

$$f(x) = \frac{\mu^x \cdot e^{-\mu}}{x!},$$

ahol $x \geq 0$ és egész szám, μ az x változó átlagértéke. Az x valószínűségi változó akkor jellemezhető a Poisson-eloszlással, ha az események függetlenek egymástól (egy esemény semmilyen információt nem ad a többi eseményről), és véletlenszerűen következnek be (az esemény bekövetkezésének a valószínűsége időben állandó), és az átlaguk és a varianciájuk közel egyenlő ($\overline{x} \approx \sigma^2$). Kis átlagértékekre az eloszlás ferde, nagy középértékekre viszont szimmetrikus. A középérték lehet tizedes szám, viszont az x csak pozitív egész szám. Két Poisson-eloszlású valószínűségi változó (x_1 és x_2) összege is Poisson-eloszlást követ, amelynek átlagértéke $\mu_1 + \mu_2$.

Németországi kutatók a lombkorona épsége alapján 16 x 16 km-es egységű négyzethálót elemeztek ki, és megállapították, hogy az elpusztult lucfenyők átlagos gyakorisága 8.4‰, az erdei fenyőké pedig 3.8‰ (Brandl et al., 2020). Ha a kihalt fák előfordulása a négyzetekben Poisson-eloszlást követ, akkor mekkora a valószínűségé annak, hogy egy adott területen a lucfenyő és az erdei fenyő kihalási gyakorisága 2‰, illetve 10‰ legyen? Mennyi a valószínűsége annak, hogy egy adott területen legalább, illetve legfeljebb 5‰ legyen a lucfenyő kihalási gyakorisága?

A lucfenyő kihalási rátájára a következő Poisson-eloszlás írható fel:

$$f(x) = \frac{8.4^x \cdot e^{-8.4}}{x!}.$$

A 3.18. ábrán látható a $\mu = 8.4$ átlagértékű Poisson-eloszlás sűrűségfüggvénye és a kumulatív valószínűségek függvénye. Az eloszlás enyhe jobb oldali ferdeséget mutat. Mivel az eloszlás diszkrét, ezért a legcélszerűbb az oszlopdiagram használata. A mellékelt táblázatban az egyes események valószínűség- és kumulatív valószínűségértékei láthatók. A táblázatból kivehető, hogy az átlagérték előtti és utáni értékek összesített valószínűsége ~0.5 (azaz ~50%). A függvény az

átlagértéktől távolodva rohamosan közelíti a nullát, a lucfenyő esetében az f(22) gyakorlatilag nullának tekinthető.

Ha $x = 2‰$, a valószínűség:

$$f(2) = \frac{8.4^2 \cdot e^{-8.4}}{2!} = 0.008,$$

ha $x = 10‰$, akkor a valószínűség:

$$f(10) = \frac{8.4^{10} \cdot e^{-8.4}}{10!} = 0.108.$$

Annak a valószínűsége, hogy $x \leq 5‰$:

$$f(x \leq 5) = f(1) + f(2) + f(3) + f(4) + f(5),$$
$$f(x \leq 5) = 0.1573.$$

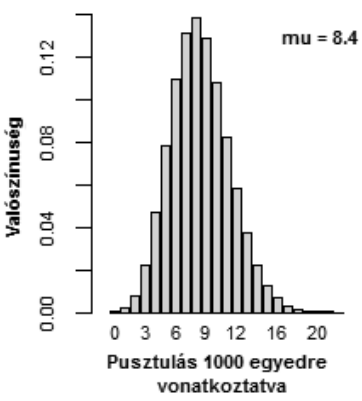
Annak a valószínűsége, hogy $x \geq 5‰$:

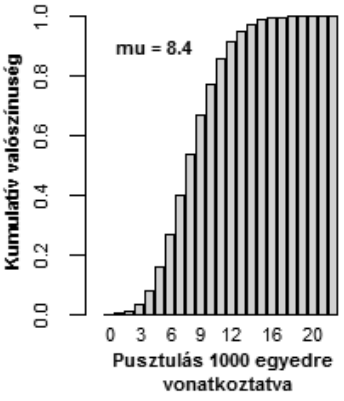
$$(x \geq 5) = 1 - [f(0) + f(1) + f(2) + f(3) + f(4)]$$
$$f(x \geq 5) = 1 - 0.0789 = 0.9211.$$

Az erdei fenyő kihalási rátájának a Poisson-eloszlási egyenlete:

$$f(x) = \frac{3.8^x \cdot e^{-3.8}}{x!}.$$

Annak a valószínűsége, hogy 1000 egyedből két kihalt fát találjunk, 0.1615, annak pedig, hogy 10 kihalt fát találjunk, 0.0039. Ezek az értékek lényegesen eltérnek a lucfenyőre kapott hasonló arányokra. Mind a két esetben az átlagértékhez közeli esetszámok valószínűségei a nagyobbak.





μ = 8.4		
x	p	∑p
0	0.0002	0.0002
1	0.0019	0.0021
2	0.0079	0.01
3	0.0222	0.0322
4	0.0467	0.0789
5	0.0784	0.1573
6	0.1097	0.267
7	0.1316	0.3986
8	0.1382	0.5368
9	0.1290	0.6658
10	0.1084	0.7742
11	0.0828	0.857
12	0.0579	0.9149
13	0.0374	0.9523
14	0.0225	0.9748
15	0.0126	0.9874
16	0.0066	0.994
17	0.0033	0.9973
18	0.0015	0.9988
19	0.0007	0.9995
20	0.0003	0.9998

3.18. ábra. *A $\mu = 8.4$ átlagértékű Poisson-eloszlás sűrűségfüggvénye és kumulatív valószínűségek függvénye*

3.2.5. A Poisson-eloszlás alkalmazása R-ben

Egy esemény bekövetkezésének valószínűsége a **dpois(x, lambda)** paranccsal kérhető ki, ahol *x* az egyedek száma egy területen, egy időintervallumban stb., a *lambda* pedig az eloszlás átlagértéke. Mekkora a valószínűsége annak, hogy 2‰, 10‰ legyen a lucfenyő kihalási gyakorisága, ha az átlagérték 8.4‰?

```
p2 = dpois(2, 8.4)
p2
[1] 0.007933319
P10 = dpois(10, 8.4)
P10
[1] 0.1083818
```

Mekkora a valószínűsége annak, hogy a lucfenyő kihalási gyakorisága legfeljebb 5‰ legyen?

```
p = sum(dpois(0:5,8.4))
p
[1] 0.1572768
```

A kumulatív gyakoriságokat a *ppois(x, lambda)* paranccsal kérhetjük ki.

```
ppois(5, 8.4)
[1] 0.1572768
#A valószínűsége annak, hogy a lucfenyo kihalási
#gyakorisága 5‰ legyen 0.157 avagy 15.7%.
```

Mekkora a valószínűsége annak, hogy a lucfenyő kihalási gyakorisága legkevesebb 5‰ legyen?

```
1-ppois(4,8.4)
[1] 0.9210917
#A valószínűsége annak, hogy a lucfenyo kihalási
#gyakorisága legkevesebb 5‰ legyen 0.921 avagy 92.1%.
```

Egy adott valószínűséghez rendelhető esemény meghatározása a **qpois(p, lambda)** paranccsal végezhető el. Mivel a *q* csak egész számokat vehet fel, az R a megadott *p* valószínűségnek leginkább megfelelő *q* értéket adja meg. A 0.02 valószínűség a 3 egyednek megfelelő valószínűségi értékhez (0.0222) van a legközelebb, tehát az eredmény 3.

```
qpois(0.02, 8.4)
[1] 3
```

A 3.18. ábra szerkesztése R-ben:

```
#A mu = 8.4 Poisson-eloszlás sűrűségfüggvénye
par(mfrow = c(1,2))
barplot(dpois(0:22, lambda = 8.4),
 xlab ="Pusztulás 1000 egyedre \n vonatkoztatva",
 cex.lab = 0.7, cex.axis = 0.7,
 ylim=c(0,0.14), ylab="Valószínuség", names=0:22,
 font.lab=2, xlim=c(0,40), width=0.8,
 col = rgb(100,100,100, alpha = 60, max = 255))
text(20, 0.13, "mu = 8.4", font = 2, cex = 0.7)
#A mu = 8.4 Poisson-eloszlás kumulatív valószínűségeinek
függvénye
barplot(ppois(0:22, lambda = 8.4),
 xlab ="Pusztulás 1000 egyedre \n vonatkoztatva",
 cex.lab = 0.7, cex.axis = 0.7,
 ylim = c(0,1), ylab = "Kumulatív valószínuség",
 names=0:22, font.lab=2, xlim=c(0,40), width=0.8,
 col = rgb(100,100,100, alpha = 60, max = 255))
text(5, 0.9, "mu = 8.4", font = 2, cex = 0.7)
```

3.2.6. A negatív binomiális eloszlás

A negatív binomiális eloszlás egy diszkrét eloszlás, amelyet Pascal-eloszlásnak is nevezünk. A valószínűségi változó binomiális, amely két egymást kizáró értéket vehet fel, és bármely próbálkozás sikerének a valószínűsége azonos (p). A *negatív* kifejezés onnan ered, hogy a binomiális eloszlással ellentétben ennek az eloszlásnak a valószínűségi változója a próbálkozások száma (n), míg a sikerek száma (r) állandó érték.

A negatív binomiális eloszlás olyan valószínűségi változót jellemez, amelyre az utolsó próbálkozásnál elérjük a kitűzött célt, az r-edik sikert. Tehát a próbálkozások x össz-száma r siker és k kudarc összege.

A függvény a következő egyenlettel fejezhető ki:

$$f(x,r,p) = \left(\frac{(x-1)!}{(r-1)!\cdot(x-r)!}\right)\cdot p^r \cdot (1-p)^{x-r},$$

ahol: x – a próbálkozások száma ($x = k + r$);
r – a sikeres események előre meghatározott száma;
p – a sikeres esemény bekövetkezésének a valószínűsége egyszeri próbálkozáskor;
$1-p$ – a kudarc valószínűsége egyszeri próbálkozáskor.

Az egyenlet felírható úgy is, hogy a valószínűségi változó a kudarcok száma (k):

$$f(k,r,p) = \left(\frac{(k+r-1)!}{(r-1)!\cdot k!}\right)\cdot p^r \cdot (1-p)^k.$$

A negatív binomiális eloszlásnak két állandója van: az r és a p, jelölése pedig NB(k, r, p). Az eloszlás középértéke és varianciája:

$$\mu = \frac{r(1-p)}{p} \text{ és } \sigma^2 = \frac{r(1-p)}{p^2}.$$

Például, ha tudjuk, hogy egy régióban 0.4 a valószínűsége annak, hogy egy vidéki településen gólyafészket találjunk, akkor mekkora a valószínűsége annak, hogy az első 10 települést bejárva elérjünk a második, illetve a hetedig településig, ahol van gólyafészek?

Az első esetben: $p = 0.4$, $x = 10$, $r = 2$, tehát $k = 8$.

$$f(8,2,0.4) = \left(\frac{(10-1)!}{(2-1)!\cdot(10-2)!}\right)\cdot 0.4^2 \cdot 0.6^8 = 0.024$$

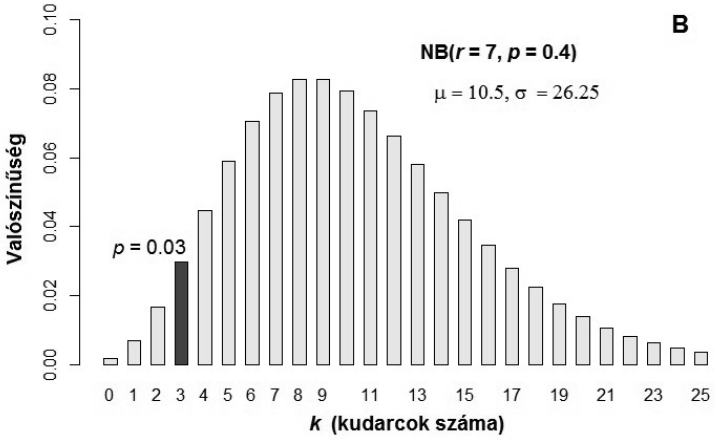
Ennek az eloszlásnak a középértéke és varianciája: $\mu = \frac{2\cdot 0.6}{0.4} = 3$ és $\sigma^2 = \frac{2\cdot 0.6}{0.16} = 7.5$.

A második esetben: $p = 0.4$, $x = 10$, $r = 7$, tehát $k = 3$.

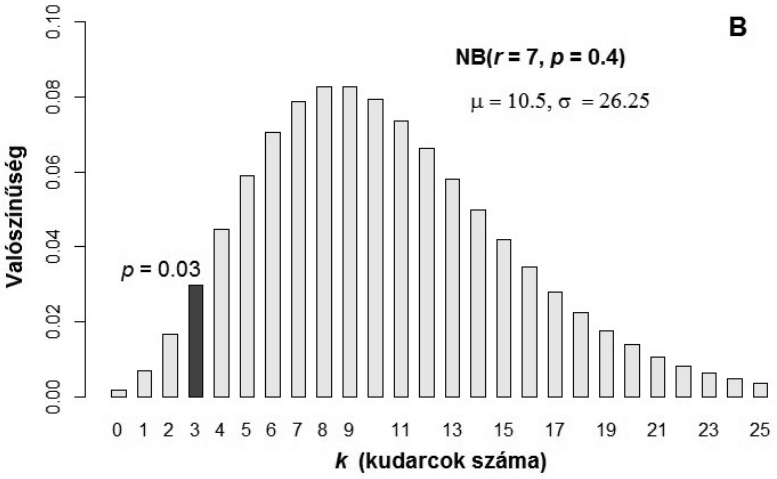
$$f(3,7,0.4) = \left(\frac{(10-1)!}{(7-1)!\cdot 3!}\right)\cdot 0.4^7 \cdot 0.6^3 = 0.030$$

Ennek az eloszlásnak a középértéke és varianciája:

$$\mu = \frac{7\cdot 0.6}{0.4} = 10.5 \text{ és } \sigma^2 = \frac{7\cdot 0.6}{0.16} = 26.25.$$



3.19. A ábra. *A negatív binomiális eloszlás r = 2 és r = 7 sikerre, ha a siker valószínűsége 0.4*



3.19. B ábra. *A negatív binomiális eloszlás* r = *2 és* r = *7 sikerre, ha a siker valószínűsége 0.4*

A két eredményből látható, hogy megközelítően azonos valószínűséggel találunk kettő, illetve hét települést gólyafészekkel, ha a tizedik településnél járunk utoljára sikerrel. Mivel egy sikeres eredmény valószínűsége 0.4 (azaz 10-ből 4), a legvalószínűbb az, ha

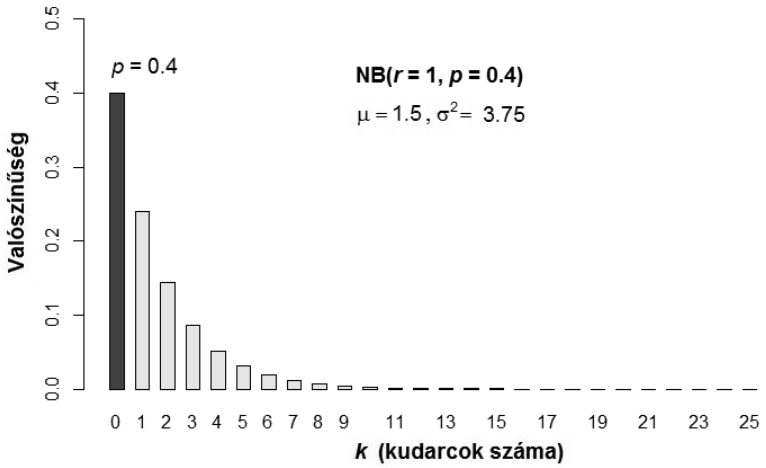
– egy kudarc után találjuk meg a második települést gólyafészekkel ($k = 1$, 3.19.A. ábra);
– nyolc-kilenc kudarc után találjuk meg a hetedig települést gólyafészekkel ($k = \{8,9\}$, 3.19.B. ábra).

Mekkora a valószínűsége annak, hogy az első településen van gólyafészek? Ebben az esetben $r = 1$, a kudarcok száma pedig $k = 0$.

$$f(0{,}1{,}0.4) = \left(\frac{(1-1)!}{(1-1)! \cdot 0!}\right) \cdot 0.4^1 \cdot 0.6^0 = 0.4$$

Az $r = 1$, azaz az első sikerig tartó próbálkozások valószínűségi eloszlása a 3.20. ábrán látható.

Ha az r értéke 1, akkor a negatív binomiális eloszlás azonos a geometriai eloszlással. Ez annak az esetnek felel meg, amikor k kudarc után sikerrel járunk. Például hányszor kell megismételni egy laboratóriumi kísérletet, amíg a várt hatás jelentkezik, vagy például hány használat után romlik el egy gép?



3.20. ábra. *A negatív binomiális eloszlás* r = *1 sikerre, ha a siker valószínűsége 0.4*

3.2.7. A negatív binomiális eloszlás alkalmazása R-ben

A negatív binomiális eloszlással kapcsolatos műveleteket R-ben a **`dnbinom(k, r, p, mu)`** paranccsal végezhetjük el, ahol *k* a kudarcok száma, *r* a sikerek száma, *p* egy sikeres esemény bekövetkezésének a valószínűsége, *mu* pedig az eloszlás várható értéke (átlaga). A binomiális eloszlás vagy a *p* vagy a *mu* segítségével fejezhető ki, így a kettő közül csak az egyiket kell megadni. Pár példa, ami az előző alfejezetben meg volt említve:

```
#k = 8, r = 2, p =0.4
dnbinom(8,2,0.4)
[1] 0.02418647

#k = 3, r = 7, p = 0.4
dnbinom(3,7,0.4)
[1] 0.02972713

#több k értékhez tartozó valószínűség kikérése
dnbinom(c(3,5,9),7,0.4)
[1] 0.02972713 0.05885972 0.08263904

#kumulatív valószínűségek kikérése
#például k < 4
```

```
pnbinom(3,7,0.4)
[1] 0.00163840 0.00851968 0.02503475 0.05476188

#A p helyett a mu ismerete
#dnbinom(k, r, mu)
dnbinom(3, 2, mu = 10.5)
[1] 0.06069289
```

A 3.18.A. ábra készítése:

```
barplot(dnbinom(0:25, 2, p = 0.4), names = 0:25,
        ylim = c(0,0.28), xlab = "k (kudarcok száma)",
 col = rgb(0,100,200, alpha = 30, max = 255),
        ylab = "Valószínuség",
        space = 0.8, font.lab = 2)
```

3.2.8. A diszkrét eloszlások jelentősége az ökológiai kutatásokban

Ökológiai kutatásokban a legtöbb esetben az a cél, hogy adott területeken felmérjék bizonyos egyedek előfordulási gyakoriságát. Ennek a gyakorlati kivitelezése legtöbbször abban áll, hogy a területen mintavételi kvadrátokat jelölnek ki (pl. 1 x 1 m-eseket, vagy 10 x 10 km-eseket) vagy csapdákat helyeznek ki, és a mintavételi kvadrátban/csapdában megszámolják az illető faj vagy fajon belül bizonyos állapotban levő egyedek (fiatalok/felnőttek, egészségesek/betegek) számát. A felmérés eredménye egy egész számokból álló adatsor, amelyben nulla is szerepelhet. Ilyen adatsorok általában a Poisson-eloszlással jellemezhetők, viszont az adatok szóródása rendszerint jóval nagyobb, mint amit ez az eloszlás előír. A túlszórással rendelkező Poisson-eloszlású adatsorokat a negatív binomiális eloszlás sokkal jobban le tudja írni. A két eloszlás szoros kapcsolatban van egymással. Tulajdonképpen a Poisson-eloszlás a negatív binomiális eloszlásnak egy szélsőséges esete (Zuur et al., 2009).

Abban az esetben, ha az eloszlás átlagértékét és varianciáját ismerjük (amelyek az empirikusan nyert adatsorból kiszámolhatók), a binomiális eloszlás r paramétere kiszámolható a variancia alábbi kifejezéséből:

$$\sigma^2 = \mu + \frac{\mu^2}{r}.$$

Innen az r megbecsülhető, ha a μ-t az adatsor átlagával, a varianciát pedig az adatsor varianciájával közelítjük meg:

$$r = \frac{\text{átlag}^2}{var - \text{átlag}}.$$

A számított r értékből megállapíthatjuk, hogy melyik függvény jellemzi jobban a vizsgált faj eloszlását a területen. Ha a variancia nagyobb, mint az átlag, akkor túlszórással állunk szemben, ezért az r az adataink szóródását mutatja. Ha az r értéke nagy, a négyzetes tag értékéhez képest, akkor a μ^2/r közelít a nullához, ekkor a variancia megegyezik a várható értékkel, és ez esetben a Poisson-eloszláshoz jutunk. Ha az r kicsi, akkor túlszórás jellemzi az adatokat.

Ha r = 1, a variancia $\mu+\mu^2$, és az eloszlást geometriai eloszlásnak nevezzük. Kis átlagértéknél és nagy szóródásnál (kicsi r értéknél) a nulla értéknek a legnagyobb a valószínűsége.

Az *r* kiszámítására egy R függvényt lehet írni, amelynek legyen a neve *emp_r*. Ez a függvény oszloponként kiszámolja az adattábla változóinak átlagát és varianciáját, és ezek alapján az *r* értékét:

```
emp_r = function (x, MARGIN=2){
 atlag = apply(x, MARGIN=MARGIN, mean)
 variancia = apply(x, MARGIN=MARGIN, var)
 r = apply(x, MARGIN=MARGIN,
 function(x) mean(x)**2 / (var(x) - mean(x)))
 tmp = data.frame(atlag, variancia, r)
 tmp
}
```

A μ, σ^2 és *r* ismeretében a binomiális eloszlás *p* paramétere is megállapítható:

$$p = \frac{\mu}{\sigma^2}.$$

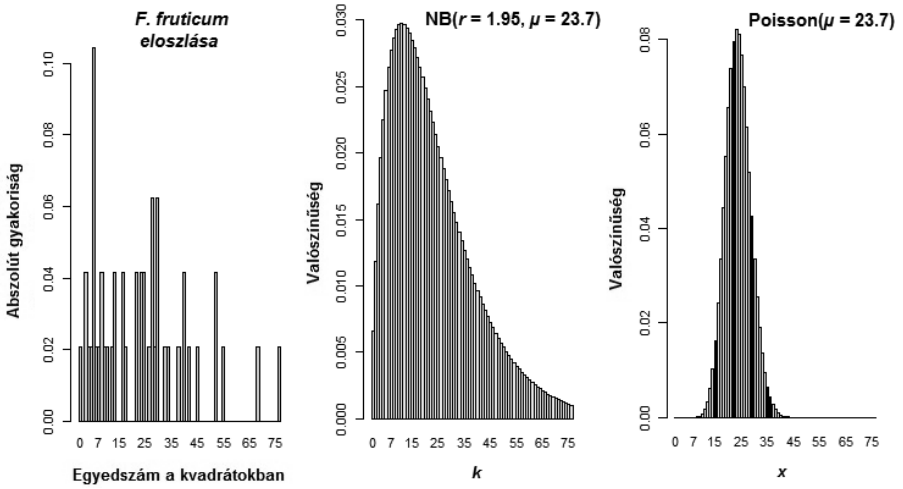
Vizsgáljuk meg a Gheoca et al. (2021) által gyűjtött pár csigafaj eloszlását Délkelet-Erdélyben a Nagy-Küküllő és az Olt mentén. A csigákat 48 kvadrátból gyűjtötték be, összesen 71 fajt azonosítottak. Két gyakori, egy ritka és egy nagyon ritka faj adatsorát vizsgáljuk meg. Gyakori fajnak számított az éticsiga (*Helix pomatia*) és a berki párduccsiga (*Fruticicola fruticum*), ritka fajnak az ugarcsiga (*Helix lutescens*), nagyon ritka fajnak minősült a harántfogú törpecsiga (*Vertigo angustior*). A négy faj adatait a 48 kvadrátban a H.pom, F.frut, H.lut és V.ang vektorok tartalmazzák.

```
H.pom = c(22,14,12,5,37,20,15,26,20,17,12,25,6,0,1,
          40,14,45,48,3, 1,0,5,0,0,0,0,7,5,20,16,45,
          1,12,16,18,25,7,0,0,5,1,3,0,0,4,11,6)
H.lut = c(2,0,0,2,0,3,0,0,2,0,0,0,0,0,1,3,0,2,0,1,2,
          0,1,0,3,7,12, 10,0,4,0,0,2,6,0,7,6,3,2,7,0,
          7,0,5,3,8,3,3)
```

```
F.frut = c(18,17,40,52,76,30,6,22,30,25,8,13,27,6,38,52,
           10,6,25,28,3,6,9,14,3,14,42,9,22,45,24,0,33,34,
           17,30,24,11,28,29,4,6,5,7,68,28,40,55)
V.ang = c(0,3,0,0,0,1,0,42,0,3,3,0,0,0,0,0,
          7,0,0,3,2,0,0,0,0,0,0,1,1,0,0,0,
          0,4,0,0,0,0,0,0,0,0,4,0,0,3,0,0)
#A fajok adatait egyesítjük a csiga adattáblába.
csiga = data.frame(H.pom, F.frut, H.lut, V.ang)

#Az emp_r segítségével kiszámoljuk az r értékeit,
#és az értékeket két tizedesre kerekítjük.
round(emp_r(csiga),2)
```

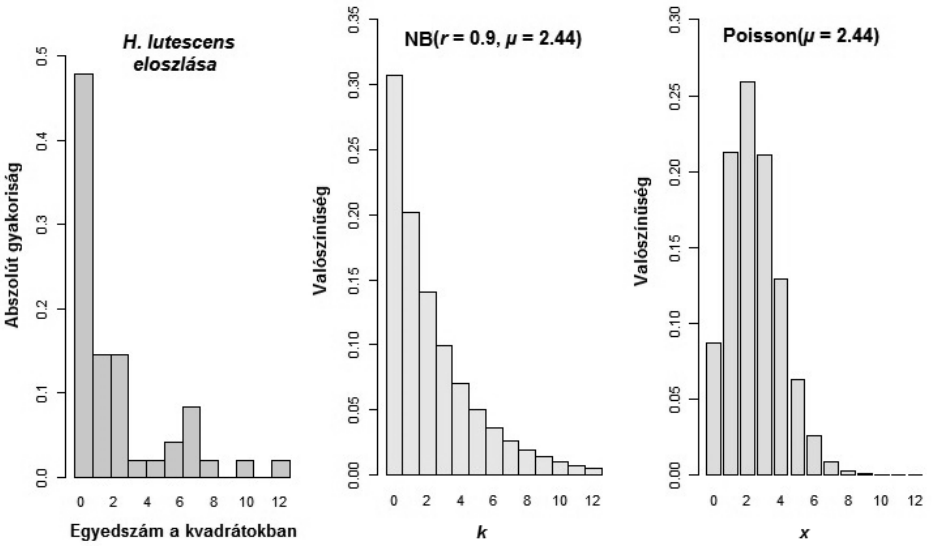
	atlag	variancia	r
H.pom	12.29	176.25	0.92
F.frut	23.73	312.20	1.95
H.lut	2.44	9.06	0.90
V.ang	1.60	37.73	0.07

A nagy átlagos egyedszámú berki párduccsiga eloszlása a 3.21. ábrán látható. Az *r* értéke 1.95, ami nagyobb 1-nél, ezért tekinthetjük úgy, hogy a Poisson-eloszlás jól leírja a faj előfordulását a vizsgált területen.



3.21. ábra. *A berki párduccsiga (*F. fruticum*) eloszlása a mintaterületen és az eloszlás megközelítése diszkrét eloszlásokkal*

Az éti- és az ugarcsiga esetében az $r \sim 1$, így az adatok szóródása geometrikus eloszlást feltételez. Ebben az esetben a negatív binomiális eloszlás jobban jellemzi a csigák eloszlását, mint a Poisson-eloszlás (3.22. ábra).

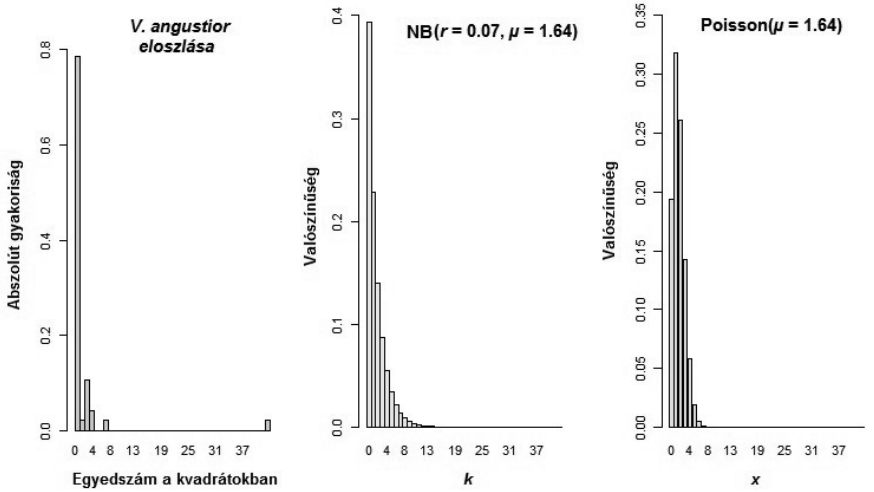


3.22. ábra. *Az ugarcsiga (*H. lutescens*) eloszlása a mintaterületen és az eloszlás megközelítése diszkrét eloszlásokkal*

A 3.22. ábra R kódjai:

```
#Ábra a Helix lutescens fajra
par(mfrow=c(1,3))
hist(H.lut, breaks=12, ylim = c(0,0.5), freq=F, main="")
barplot(dnbinom(0:12,0.9,mu=2.44), names=0:12,
 ylim=c(0,0.35),
 col=rgb(0,100,200, alpha=30, max=255))
barplot(dpois(0:12, lambda=2.44), ylim = c(0,0.3),
 names=0:12,
 col=rgb(100,150,150, alpha=60, max=255))
```

A harántfogú törpecsiga ritka fajnak számít a vizsgált területen, az ilyen esetekre pedig a nagyon kis r értékek jellemzőek ($r = 0.07$), ugyanakkor az adatoknak nagyon nagy a varianciája. Ennek a csigafajnak az átlagos egyedszáma is kicsi (1.64), ami az eloszlást erősen csúcsossá teszi nullában. Ebben az esetben is az eloszlást jobban leírja a negatív binomiális eloszlás (3.23. ábra).



3.23. ábra. *A harántfogú törpecsiga (*V. angustior*) eloszlása a mintaterületen és az eloszlás megközelítése diszkrét eloszlásokkal*

3.3. A kiugró értékek jelentősége

A kiugró értékek érdekesek lehetnek, különösen akkor, ha a mintavétellel és a méréssel kapcsolatos hibák kizárhatók. A nem várt magas, illetve nem várt alacsony érték valamilyen faktornak, külső tényezőnek a hatása lehet, ami csak az adott egyedre vagy objektumra hatott. Általában ezeket a tényezőket utólagos kutatással fel lehet deríteni. Például az *airquality* R adattáblában a teljes vizsgált időszakra (öt hónap) négy kiugró pontot kaptunk az ózonkoncentrációra (4.3. ábra). Egy egyszerű kód segítségével kikérhetjük a magas ózonkoncentrációjú napoknak a teljes adatsorát.

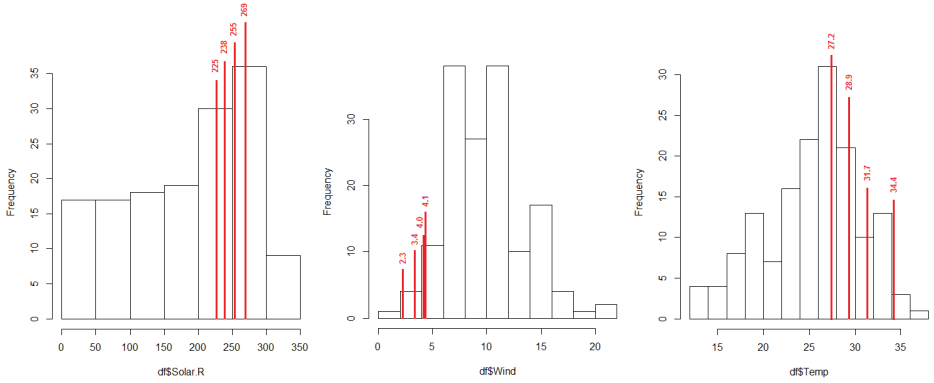
```
df = airquality
df$Temp = (df$Temp-32)/1.8
df1 = subset(df, Ozone > 115)
df1
```

	Ozone	Solar.R	Wind	Temp	Month	Day
62	135	269	4.1	28.9	7	1
99	122	255	4.0	31.7	8	7
117	168	238	3.4	27.2	8	25
121	118	225	2.3	34.4	8	29

Ezekre a mérésekre júliusban és augusztusban került sor, a hőmérséklet és a napsugárzás átlagon felüli, a szélsebesség pedig jóval átlagon aluli volt (3.24. ábra). Az ózon a városi levegőben a nitrogén-oxidok és az oxigén reakciójából képződik erős napsugárzás hatására. A légköri adatokból az derült ki, hogy ezeken a napokon elsősorban a szélsebesség volt különösen alacsony értéken. Szélcsendben a szennyezők feldúsulnak a városi levegőben, a viszonylag magas napsugárzás és hőmérséklet pedig hozzájárult az ózonképződéshez. Noha erre nincs adat, valószínű, hogy a közúti forgalom is fokozottabb volt ezeken a napokon, ami növelhette a levegő nitrogén-oxid tartalmát.

Ez a példa alátámasztja azt, hogy ha a kiugró értékek háttértényezők hatásának az eredményei, akkor értelmezésük gyakorlati magyarázatot szolgáltathat az elméleti tudásnak, fényt deríthet a lokális hatásokra, esetleg szennyezőforrások beazonosítására vagy egyéb extrém körülmények környezetre gyakorolt hatásának a kimutatására.

Azok az adatok, amelyek a normális tartományba esnek, azt jelzik, hogy az illető objektumokra vagy mintákra azonos környezeti hatások hatnak, egyformán vannak kitéve természetes és esetleges stresszhatásoknak.



3.24. ábra. *A kiugróan magas ózonkoncentrációk mellett mért napsugárzás (lang), szélsebesség (mph) és hőmérséklet (°C) az* airquality *adattáblában*

3.4. Léptékváltás

Sokváltozós adatelemzéskor fennállhat az a probléma, hogy a változók különböző léptékű skálákon mozognak. Például elemanalízis során a makroelemek koncentrációja több 100 mg/l lehet, másoké (pl. a nyomelemeké) pár μg/l. Ebből természetszerűen adódik, hogy a makroelemek varianciái nagyságrenddel nagyob-

bak, mint a nyomelemeké. Ahhoz, hogy a varianciákat össze tudjuk hasonlítani, az összes elem szórását azonos skálára kell hozni. Ezt szolgálja a léptékváltás (transzformáció). A léptékváltást többféleképpen végezhetjük el, a fő szempont a változó típusa.

3.4.1. Léptékváltás arányskálájú változóknál

A leggyakrabban használt művelet a standardizálás, amelynek során egységnyi szórásra és nulla középre alakítjuk a változók adatsorait. A skála vagy zsugorodik, vagy kitágul oly módon, hogy az új adatsor szórása 1 lesz, számtani középértéke pedig 0. Az új adatsor dimenziómentessé válik.

$$x'_{ij} = \frac{\sum_{i=1}^{N} x_{ij} - \bar{x}_j}{s_i},$$

ahol x'_{ij} a j változó i objektumára kapott standardizált értéke, x_{ij} a j-dik változó i objektumára kapott eredeti koncentrációérték, $\bar{x}_j$ a változó értékeinek számtani középarányosa, az s_j pedig a változó szórása. Az $x_j \in [1,N]$, ahol N az objektumok száma. Például adott a {154, 120, 147, 161, 139} vektor, az elemeinek a számtani középértéke 144.2, szórása pedig 15.8. Az adatokat a fenti egyenlet alapján standardizáljuk:

$$x'_1 = \frac{(154 - 144.2)}{15.8} = 0.62,\ x'_2 = \frac{(120 - 144.2)}{15.8} = -1.53 \text{ stb.}$$

A standardizált vektor {0.62, −1.53, 0.18, 1.06, −0.33} elemeinek az átlaga 0, a szórása pedig 1. Az R-ben a művelet a **`scale()`** paranccsal végezhető el.

```
v = c(154, 120, 147, 161, 139)
v_sd = scale(v)
v_sd = round(vs_d, 2)
#Két tizedesre kerekítjük az értékeket.
v_sd
 [,1]
[1,] 0.62
[2,] -1.53
[3,] 0.18
[4,] 1.06
[5,] -0.33
attr(,"scaled:center")
[1] 144.2
attr(,"scaled:scale")
[1] 15.8019
```

Az R-ben egy adattábla tetszőleges változói is standardizálhatók, például az *airquality* (*df*) adattábla első négy változója.

```
df = airquality
df_stand_m = scale(df[,1:4]) #Mátrixban menti el.
df_stand = data.frame(df_stand_m, df [,5:6])
head(df_stand)
#Visszakaptuk a df-et, négy standardizált változóval.
```

```
         Ozone      Solar.R       Wind       Temp Month Day
1  0.005025743   0.05075471 -0.7259482 -1.1497140     5   1
2 -0.163920873  -0.76756846 -0.5556388 -0.6214670     5   2
3 -0.974864627  -0.41523487  0.7500660 -0.4101682     5   3
4 -0.772128689   1.44872346  0.4378323 -1.6779609     5   4
5 -0.582388644  -0.04816810  1.2326091 -2.3118573     5   5
6 -0.434235458  -0.04816810  1.4029185 -1.2553634     5   6
```

Arányskálájú változóknál megtörténhet az, hogy az adatsor eloszlása ferde a kiugró, illetve extrém értékek miatt. A ferdeség kiküszöbölhető az adatok logaritmálásával. Általában a természetes alapú vagy a tizes alapú logaritmálást szokták használni. Ennek az átalakításnak az az előnye, hogy az adatok utólag is értelmezhetők.

3.4.2. Léptékváltás intervallumskálájú változóknál

Az intervallumskálájú változóknál a terjedelem-léptékváltást szokták alkalmazni. A skála ebben az esetben is zsugorodik vagy kitágul, az értékek pedig 0 és 1 között helyezkednek el. A változó új értékei dimenziómentesek.

$$x_{ij}^{'}=\frac{x_{ij}-\min(\mathrm{x_{ij}})}{\max(\mathrm{x_{ij}})-\min(\mathrm{x_{ij}})}$$

ahol $x_{ij}^{'}$ a *j* változó *i* objektumára kapott standardizált értéke, x_{ij} a *j*-dik változó *i* objektumának eredeti koncentrációértéke. Az előző példánál maradva, azonosítjuk a legkisebb (120) és a legnagyobb (161) elemet, majd kiszámoljuk a terjedelmet (161−120 = 41).

A vektor új elemei {0.83, 0.00, 0.66, 1.00, 0.46} a 0 és 1 között helyezkednek el. Az R alapcsomagjában nincs kód ennek a műveletnek az elvégzéséhez, viszont szerkeszthetünk egy függvényt, és elnevezhetjük **terjedelem()**-nek, illetve használhatjuk a {vegan} csomag **decostand()** parancsát a range módszerrel **(decostand(x, "range")**.

```
terjedelem = function(x) {
 round((x-min(x))/(max(x)-min(x)),2)
 }
```

```
#A min(), max(), mean() parancsokban használható a
#na.rm = T utasítás az NA-k eltávolítására.
v = c(5,12,5,1,6,9)
v_terj = terjedelem(v)
v_terj
[1] 0.36 1.00 0.36 0.00 0.45 0.73
```

A függvény egy adattábla oszlopaira is alkalmazható, például az *airquality* (*df*) első négy változójára, miután eltávolítottuk a hiányzó adatokat tartalmazó sorokat.

```
df = airquality
df1 = na.omit(df)
#Terjedelem-léptékváltás a df1 első négy oszlopára
df1_terj = terjedelem(df1[,1:4])
head(df1_terj)
    Ozone Solar.R Wind Temp
1    0.12    0.57 0.02 0.20
2    0.11    0.35 0.02 0.21
3    0.03    0.44 0.03 0.22
4    0.05    0.94 0.03 0.18
7    0.07    0.89 0.02 0.19
8    0.05    0.29 0.04 0.17

#A négy oszlophoz hozzáragasztjuk a df1
#5. és 6. oszlopát
df_terj = data.frame(df1_terj, df1[,5:6])
head(df_terj)
  Ozone Solar.R Wind Temp Month Day
1  0.12    0.57 0.02 0.20     5   1
2  0.11    0.35 0.02 0.21     5   2
3  0.03    0.44 0.03 0.22     5   3
4  0.05    0.94 0.03 0.18     5   4
7  0.07    0.89 0.02 0.19     5   7
8  0.05    0.29 0.04 0.17     5   8
```

3.4.3. Léptékváltás diszkrét változóknál

A diszkrét változók léptékváltása elsősorban a többváltozós adatelemzési módszerek (főkomponens-analízis, redundanciaanalízis, korreszpondanciaanalízis, klaszteranalízis) előfeltétele. Ilyen esetekben az általánosan elfogadott módszerek a Chord-, ill. a Hellinger-transzformáció (Legendre és Gallagher, 2001).

Ezeknek a módszereknek az előnye az, hogy megőrzik a metrikus távolságokat, és egyenértékűvé tehetők az arányskálájú változókra alkalmazott euklideszi távolságokkal. Az ökológiában egy objektumban (pl. kvadrátban, csapdában) talált fajok abundanciaadataira használhatók.

A Chord-transzformáció lényege az, hogy egy objektumra az összes faj abundanciáját úgy alakítja át, hogy az értékek összege 1 legyen. Hasonlóan, a Hellinger-transzformáció egy objektumra megadja egy faj relatív abudanciájának a négyzetgyökét. Elsősorban akkor ajánlott a használata, ha az objektumok méretei eltérőek.

Egy *objektum x faj* adattáblánál, amelynek mérete *n x m*, a Chord-transzformáció az *i* objektumra a következő képlettel értelmezhető:

$$y_{ij}^{chord} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{m} y_{ij}^2}}.$$

A Hellinger-transzformáció az *i* objektumra:

$$y_{ij}^{H} = \sqrt{\frac{y_{ij}}{\sum_{j=1}^{m} y_{ij}}}.$$

Az említett transzformációk R-ben a {vegan} csomagban található **`decostand()`** paranccsal végezhetők el (3.1. táblázat).

3.1. táblázat. *Abundanciaadatok transzformációja R-ben*

Parancs	Részletek	Magyarázat
decostand(x, method) {vegan}		
x	adattábla	adattáblában levő fajok oszlopai
method	"normalize", "hellinger", "total"	Chord, Hellinger, relatív gyakoriság

4. HIPOTÉZISVIZSGÁLATOK

A tudományos kutatás során általában a populáció egyik tulajdonságára megfogalmazott hipotézist teszteljük az általunk vett minta alapján. Ezzel foglalkozik a következtetési statisztika. A *Bevezető*ben tisztáztuk a hipotézis fogalmát mint egy olyan állítást, ami összefüggést feltételez a tények között, és ami a tudomány eszközeivel tesztelhető. A statisztikai értelemben vett hipotézisvizsgálatnak kissé szokatlan alapelve van: két egymással ellentétes állítást vizsgálunk, és azt fogjuk igaznak tekinteni a populációra nézve, amelyik a minta alapján hihetőbbnek tűnik a másiknál. Ennél a pontnál világossá válik a tudományos kutatásnak egy lényeges korlátja: nem tudunk teljes bizonyossággal kijelenteni vagy cáfolni valamit a populációval kapcsolatban. A kijelentésünk tehát lehet igaz, de lehet hamis. *Azt viszont meg tudjuk mondani, hogy az állításunk igaz, illetve hamis voltának mekkora a valószínűsége.*

Az állításunkat nullhipotézisnek nevezzük, és H_0-val jelöljük. A statisztikai tesztek általában úgy működnek, hogy nullhipotézisük egyenlőséget feltételez. Ha egy átlagot egy számhoz viszonyítunk, vagy két átlagot egymáshoz, akkor abban vagyunk a legbiztosabbak, hogy a köztük lévő különbség akkor nulla, ha a kettő egyenlő egymással. Tehát a nullhipotézis a következő formában írható fel:

H_0: $\bar{x} = a$ *vagy* $\bar{x} - a = 0$ (ahol a egy valós vagy egy egész szám)
vagy
H_0: $\bar{x} = \mu$ *vagy* $\bar{x} - \mu = 0$.

Például a Kolmogorov–Szmirnov-próba azt teszteli, hogy egy valószínűségi változónak az eloszlása megegyezik-e az általunk feltételezett eloszlással. Ennek a próbának a nullhipotézise azt jelenti ki, hogy a minta eloszlása egyezik az elméleti eloszlással.

A nullhipotézis ismeretében megfogalmazzuk az alternatív hipotézist, amelyet H_A vagy H_1 formában jelölünk. A tagadást többféle módon fogalmazhatjuk meg. Ha a H_0 egyenlőséget feltételez, akkor ezt három módon tagadhatjuk: nem egyenlő, kisebb vagy nagyobb.

H_A: $\bar{x} \neq a$ vagy $\bar{x} < a$ vagy $\bar{x} > a$ (ahol a egy valós vagy egy egész szám)
vagy
H_A: $\bar{x} \neq \mu$ vagy $\bar{x} < \mu$ vagy $\bar{x} > \mu$

Ha a tagadást a *nem egyenlő* feltételhez kötjük, akkor kétoldali tesztet végzünk, ugyanis az eltérés lehet pozitív vagy negatív (az átlag nagyobb vagy kisebb az a-hoz vagy a μ-höz viszonyítva). Ha a *kisebb* vagy a *nagyobb* feltételt választjuk alternatívaként a H_0 hipotézisre, akkor egyoldali tesztet végzünk.

A statisztikai próbát előre megszabott szignifikanciaszint (α) mellett végezzük. Az R által adott eredmény a következőket tartalmazza: a próbában tesztelt paraméter (próbastatisztika) értékét, szabadságfokainak a számát és a paraméterhez tartozó valószínűségi értéket (p-értéket). A döntést a p-érték alapján hozzuk meg, amit az α értékhez viszonyítunk. Ha $p \leq \alpha$, akkor elvetjük a H_0 hipotézist, ha $p > \alpha$, akkor megtartjuk a H_0 hipotézist. Az α szint azt jelenti, hogy ha H_0 igaz, akkor száz mintavételből 100α alkalommal kapnánk olyan eredményt, ami alapján elvetjük a H_0-t. A tudományos világban az $\alpha = 0.05$ értéket fogadják el mint kritikus szintet. Ez azt jelenti, hogy ha H_0 igaz, akkor nagy valószínűséggel 100 esetből 95-ször hoznánk helyes döntést, és csupán 5 alkalommal döntenénk úgy, hogy elvetjük a H_0-t. Tehát ha a statisztikai próba alapján ($p > \alpha$) úgy döntünk, hogy megtartjuk a H_0 hipotézist, akkor 5% a valószínűsége annak, hogy tévedtünk.

Indokolt esetben, ha a p-érték szignifikáns eltérést jelzett, akkor a hatásnagyságot (ES – effect size) is megadjuk. A hatásnagyság kiegészíti a p-értéket azzal, hogy megmutatja az eltérés nagyságát. Ennek azért van jelentősége, mert nagy mintaelemszámnál a p-érték a kis hatásnagyság mellett is szignifikáns eltérést mutathat. Ennek az esetnek egy klasszikus példája az aszpirin hatékonyságának vizsgálata a szívinfarktus megelőzésében (Bartolucci et al., 2011). Öt éven át vizsgálták 22 000 alanyon az aszpirin hatását, és megállapítottak egy szignifikáns eltérést a teszt- és a kontrollcsoport között ($p < 0.00001$). Erre a tanulmányra alapozva széles körben kezdték alkalmazni az aszpirint szív- és érrendszeri betegeknél a szívinfarktus megelőzésére. A szerzők nem tértek ki a hatásnagyságra, ami utólag kiszámolva 0.77% kockázati különbséget adott a két csoport között, ami egy roppant kis érték. Azóta több kutatás ennél kisebb hatásnagyságot kapott eredménynek, ezért idővel elvetették ezt a megelőzési módszert (Sullivan és Feinn, 2012).

Hatásnagyság

A hatásnagyság értelmezése a próbák függvényében változik, így a megfelelő helyen részletesebben kitérünk rá. A leggyakoribb esetekben a Jacob Cohen és Larry V. Hedges által javasolt képleteket használjuk (Cohen, 1988; Hedges, 1981). Cohen egy értékskálát javasolt a hatásnagyságnak (4.1. táblázat).

4.1. táblázat. *A hatásnagyság kiértékelése Jacob Cohen szerint*

Hatásnagyság	Relatív nagyság	Sűrűségfüggvény alatti terület	Nem átfedett rész (%)
0	Jelentéktelen	50	0
0.2	Kicsi	58	15
0.5	Közepes	69	33

Hatásnagyság	Relatív nagyság	Sűrűségfüggvény alatti terület	Nem átfedett rész (%)
0.8	Nagy	79	47
1.0		84	55
1.5		93	71
2.0	Óriási	97	81

Nulla hatásnagyságnál a második minta átlaga egybeesik az első minta átlagával, vagyis az első minta sűrűségfüggvénye alatti területet 50−50%-ra osztja. A két minta eloszlásai teljesen fedik egymást. Ha a hatásnagyság 0.8, akkor a második minta átlaga az első minta sűrűségfüggvénye alatti területet 79%−21%-ra osztja, a két eloszlás 47%-ban nem fed át egymással.

A Cohen-féle *d*-t akkor használjuk, ha $n_1 + n_2 > 20$, amelynek képlete:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}}.$$

A Hedges-féle g értékkel (ha $n_1 + n_2 \leq 20$) egy korrekciót végzünk d-re, amely egy szorzótényező, és értéke annál jelentéktelenebb, minél nagyobb az $n_1 + n_2$ (ha $n_1 + n_2 \to \infty$, akkor $g \to 1$):

$$g = d \cdot \left(1 - \frac{3}{4 \cdot (n_1 + n_2) - 9}\right).$$

Abban az esetben, ha a két minta szórása szignifikánsan eltér egymástól, vagy a kontroll és a teszt csoportra kapott szórások nagyon eltérőek, akkor a Glass-féle deltát használjuk. Glass azt ajánlotta, hogy a párosított minták esetén a kontroll szórását használjuk:

$$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{s_{kontroll}}.$$

Természettudományos kutatásokban általában független mintákat hasonlítunk össze, ezért Klein és Dabney (2013) azt javasolta, hogy ilyen esetben két hatásnagyságot számoljunk, külön-külön a két szórásra.

R-ben több csomag lehetőséget ad a hatásnagyság kiszámítására (effsize, effectsize, lsr), ezek közül a legátfogóbb az {effectsize} csomag. Ebben a csomagban a 4.2 táblázatban feltüntetett parancsok találhatók.

4.2. táblázat. *A mintaelemszám meghatározása a statisztikai próbák mutatói alapján*

Parancs	Csoportszám	Részletek
{effectsize} csomag		
cohens_d, hedges_d	egy mintára vagy egymástól függő vagy független két mintára (átlagok)	x (num), y (nominal), pooled_sd = T, mu = 0, paired = F, alternative = c("two.sided", "less", "greater"), ci = 0.95
glass_delta	egymástól függő vagy független két csoportra (nagy különbségek a szórások között)	x (num), y (nominal), mu = 0, paired = F, alternative = c("two.sided", "less", "greater"), ci = 0.95
cohens_f*	több csoport (átlagok)	ANOVA modell, partial = F, ci = 0.95, altenative = "two.sided"
eta_squared*	több csoport (átlagok)	ANOVA modell, partial = F, ci = 0.95, altenative = "two.sided"

* Az *f* és az *eta-négyzet* mutatók az ANOVA próbánál kerülnek bemutatásra.

A döntéshozás hibái

A 4.3. táblázat tartalmazza a döntéshozás hibáinak a valószínűségeit. Ha elvetjük a H_0-t, noha az igaz, akkor elsőfajú hibát követünk el, aminek értéke α. Az α értékét tetszőlegesen határozzuk meg, így az elsőfajú hiba elkövetésének valószínűsége ismert. Ha megtartjuk a H_0 hipotézist, noha az hamis állítás, akkor másodfajú hibát követünk el, aminek valószínűsége β. Ha egy tényező hatását vizsgáljuk egy populációra, akkor β hibát abban az esetben követünk el, ha a tényező valóban hat a populációra, de mi azt a döntést hozzuk, hogy nem hat. Ilyenkor a kutatásnak nem volt elég statisztikai ereje ahhoz, hogy kimutassa a hatást a populációra. Természettudományos kutatásokban általában β-t nem ismerjük, viszont ha α-t rögzítjük, a β nagysága két tényezőtől fog függeni: a minta elemszámától és a nullhipotézistől való eltérés mértékétől. Ha előzetesen eldöntjük, hogy mekkora legyen a legkisebb eltérés a nullhipotézistől, amit szakmailag már valós eltérésnek tartunk, akkor kiszámolható a mintának egy olyan elemszáma, amely biztosítja, hogy az α és β ne haladja meg a kívánt hibakorlátot. A problémát másként is felvethetjük: ha rögzítjük α-t, a minta elemszámát és a legkisebb lényeges eltérést a nullhipotézistől, akkor kiszámolhatjuk β-t, illetve az $1-\beta$-t (a próba erejét).

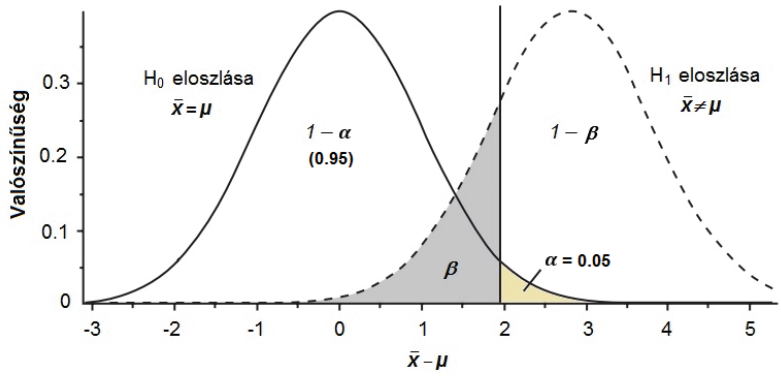
Például a kórházi laboratóriumi tesztek esetében lehet tudni a β-t, mivel előzetes vizsgálatokkal megállapítható az, hogy a teszt milyen hatékonysággal mutatja ki a betegség jelenlétét (H_0 azt feltételezi, hogy az ember egészséges, tehát azt az esetet vizsgálják, amikor H_0 hamis). Ha a teszt ereje ($1-\beta$) nagyobb 80%-nál, (a betegek legalább 80%-át betegnek nyilvánítja) akkor általában hatékonynak minősítik a tesztet (Hintze, 2008; Serdar et al., 2021). Minél nagyobb a statisztikai próba ereje, annál kisebb az esélye annak, hogy másodfajú hibát követünk el.

4.3. táblázat. *A minta alapján hozott döntés lehetséges hibái*

		A valóság	
Döntés a minta alapján		H_0 igaz	H_0 hamis
	H_0-t elfogadjuk	Helyes döntés	Másodfajú hiba
		$1-\alpha$	β
	H_0-t elvetjük	Elsőfajú hiba	Helyes döntés
		α	$1-\beta$

A 4.1. ábrán látható a nullhipotézis eloszlása nulla középértékkel, amelyre részben rátevődik az alternatív hipotézis görbéje (aminek a pontos helyét nem tudjuk). A nullhipotézis eloszlása megmutatja az összes lehetséges eredményt, amit akkor kapnánk, ha a H_0 igaz lenne. Bármely $\bar{x}-\mu$ értékre, amely a görbéhez tartozik, azt a helyes döntést hozzuk meg, hogy megtartjuk a H_0 hipotézist. Az alternatív hipotézis eloszlása megmutatja az összes lehetséges eredményt, amit akkor kapnánk, ha a H_0 hamis lenne. Bármely $\bar{x}-\mu$ értékre, amely a görbéhez tartozik, a helyes következtetés az, hogy elvetjük a H_0 hipotézist. Az α és a β függnek egymástól, csak az egyik rovására növelhetjük meg a másikat (a 4.1. ábrán látható függőlegest elmozdítva az egyik nő, a másik csökken). Értéket adva az α-nak, indirekt módon hatunk β-ra is.

A két görbének van egy metszete, ahol adott $\bar{x}-\mu$ érték mind a két görbéhez tartozik. Ezekre az $\bar{x}-\mu$ értékekre nem tudjuk megmondani, melyik döntés helyes. Itt kap értelmet az α szignifikanciaszint meghatározása. Alacsony értéket adva α-nak, lecsökkentjük az elsőfajú hiba elkövetését, de megnöveljük a másodfajú hiba lehetőségét. Ha megnöveljük egy próba erejét, megnő az α is, ekkor lecsökkentjük a másodfajú hiba elkövetésének valószínűségét, de megnöveljük az elsőfajú hiba elkövetésének valószínűségét.



4.1. ábra. *A H_0 és H_1 hipotézisek eloszlása és egymáshoz való viszonyulása*

Statisztikai szempontból az elsőfajú hiba elkövetése súlyosabb, mint a másodfajú hiba elkövetése. A természettudomány területén, a témának megfelelően bármelyik hiba elsőbbséget kaphat. Az elsőfajú hiba elkövetése azt jelenti, hogy tévesen elutasítjuk a nullhipotézist. A nullhipotézis elvetése akkor, amikor a valóságban az igaz, olyan következményekkel járhat, mint például természetvédelmi, szennyezésmentesítési, egyéb beavatkozási tevékenységek kezdeményezése, amelyek teljes mértékben szükségtelenek. A másodfajú hiba elkövetésekor fenntartjuk azt az állításunkat, hogy a vizsgált tényezőnek nincs semmilyen hatása a vizsgált területre, közösségre, pedig a valóságban van. Ebben az esetben szükség lenne a beavatkozásra, de erre nem kerül sor.

Beláthatjuk, hogy a próba szignifikanciaszintje és az ereje, a hatásnagyság és a minta elemszáma összefüggő tényezők. Ezek közül bármelyik kiszámolható a többinek az ismeretében.

Minták elemszámának a meghatározása

Kutatástervezésnél kulcsfontosságú a minta elemszámának a meghatározása. Ilyenkor előre leszögezzük a hatásnagyságot és a próba erejét. R-ben létezik egy {pwr} csomag, amit kizárólag ezeknek a tényezőknek a kiszámítására dolgoztak ki, különféle statisztikai próbák esetén (4.4. táblázat). Ezeket a próbákat a következő alfejezetekben részletesen is tárgyaljuk.

4.4. táblázat. *A mintaelemszám meghatározása a statisztikai próbák mutatói alapján*

Parancs	Csoportszám	Részletek
{pwr} csomag		
pwr.t.test	egymástól függő vagy független két csoportra, ill. egy csoportra (átlagok)	d* = Cohen-féle d, n = elemszám, sig.level = 0.05, power = 1–β, type = c("two.sample", "one. sample", "paired"), alternative = c("two.sided", "less", "greater")
pwr.t2n.test	egymástól független két csoportra (átlagok)	d = Cohen-féle d, n1, n2 = elemszámok, sig.level = 0.05, power = 1–β, alternative = c("two.sided", "less", "greater")
pwr.anova.test	több csoport (átlagok)	f* = Cohen-féle f, k = csoportok száma, n = elemszám csoportonként, sig.level = 0.05, power = 1–β
pwr.2p.test	két csoport, arányokra (azonos n)	h* = ES, n = elemszám, sig.level = 0.05, power = 1–β, alternative = c("two.sided", "less", "greater")

*d, f, h – hatásnagyság mérői

Például ha két, ill. három területen végzünk felmérést, szeretnénk egy változó átlagértékeinek az egyenlőségét tesztelni úgy, hogy $\alpha = 0.05$ és $1-\beta = 0.8$ mellett biztonságosan kimutassunk egy közepes hatásnagyságot (legyen ez az érték 0.5). Ebben az esetben a minták szükséges elemszáma:

```
library(pwr)
#Két független minta esetén
pwr.t.test(d = 0.5, power = 0.8)
 Two-sample t test power calculation
              n = 63.766
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
NOTE: n is number in *each* group

#Három egymástól független minta esetén
pwr.anova.test(k =3, f = 0.5, power = 0.8)
Balanced one-way analysis of variance power calculation
              k = 3
              n = 13.895
              f = 0.5
      sig.level = 0.05
          power = 0.8
NOTE: n is number in each group
```

Ilyen feltételek mellett kettő, illetve három csoport esetén legalább 64, illetve 14 elemet kell begyűjteni területenként, a véletlen mintavétel módszerével.

A hipotézisvizsgálat lépései

A könnyű követhetőség érdekében a hipotézisvizsgálat elvégzésekor ajánlott betartani a következő lépéseket:

– a hipotézis megfogalmazása;
– a szignifikanciaszint meghatározása (általában $\alpha = 0.05$);
– a teszt típusának eldöntése: egyoldali, kétoldali;
– a próbafüggvény kiválasztása;
– a teszt alkalmazhatósági feltételeinek ellenőrzése;
– a hipotézis tesztelése;
– döntéshozás.

4.1. Parametrikus és nem parametrikus statisztikai próbák

A statisztikai próbákkal populációkat hasonlíthatunk össze, változók közti összefüggést mutathatunk ki, vagy regresszióanalízist végezhetünk. Az utóbbival kimutatható, hogy egy vagy több független változó hatást gyakorol-e egy függő változóra. A statisztikai próbák kétfélék lehetnek: parametrikusak és nem parametrikusak. A parametrikus tesztek normális eloszlást feltételeznek a mintára vagy mintákra, a nem parametrikus tesztek alkalmazása viszont nem kötött a normális eloszláshoz. A parametrikus teszteket hatékonyabbaknak tartják egy létező hatás kimutatására, mint a nem parametrikus teszteket.

A parametrikus tesztek használata a következő feltételekhez kötött:

– a minta elemszáma elég nagy ahhoz, hogy reprezentatív legyen a populációra nézve, amiből származik;
– a populáció, amiből a minta származik, normális eloszlású legyen;
– csoportok összehasonlításakor a csoportok varianciái megközelítően azonosak legyenek.

A populációk összehasonlítására többféle próbát lehet használni, a leggyakrabban használtak listája a 4.5. táblázatban látható. A *t*-tesztek a Student-féle *t*-paraméterrel dolgoznak, innen kapták a nevüket. A kétmintás *t*-teszt és a Welch-teszt abban különbözik egymástól, hogy az első egyenlő varianciákat feltételez, a második viszont nem. Az ANOVA (ANalysis Of VAariances) varianciaanalízist jelent. A próba a csoportok varianciáját viszonyítja az összevont adatsorok varianciájához, amely arány *F*-eloszlást követ, így a számított paraméter az *F*, Ronald Fisher neve után, aki az ANOVA próbát kidolgozta.

Az ANOVA egyenlő varianciákat feltételez a csoportokra, a Welch–ANOVA a nem egyenlő varianciákra alkalmazható. A Mann–Whitney–Wilcoxon-teszt és a Kruskal–Wallis-teszt nem parametrikus tesztek, amelyek mediánokat hasonlítanak össze, és nem írnak elő semmilyen előfeltételt.

4.5. táblázat. *Statisztikai próbák populációk összehasonlítására egy vagy több csoport alapján*

Próba	Típus	Összehasonlítás	Csoportok száma
Egymintás *t*-teszt	parametrikus	átlag egy értékkel	1
Kétmintás *t*-teszt	parametrikus	két átlag	2
Welch-teszt (*d*-teszt)	parametrikus	két átlag	2
Mann–Whitney–Wilcoxon-teszt (*U*-teszt)	nem parametrikus	medián egy értékkel két medián	1 2

Egyszempontú ANOVA	parametrikus	átlagok	3+
Welch–ANOVA	parametrikus	átlagok	3+
Faktoriális ANOVA	parametrikus	átlagok	2 vagy több faktor
Kruskal–Wallis	nem parametrikus	mediánok	3+
Jonckheere–Terpstra-próba	nem parametrikus	mediánok	3+
Friedman-próba	nem parametrikus	mediánok	3+
Welch–Johansen	nem parametrikus	mediánok	2 vagy több faktor

A kettőnél több csoportot összehasonlító teszteknél egy szignifikáns különbséget adó eredmény azt jelenti, hogy legalább két csoport átlaga vagy mediánja eltér egymástól, ezért megköveteli egy ún. *post-hoc* teszt elvégzését, ami arra hivatott, hogy megmutassa a páronkénti szignifikáns különbségeket a csoportátlagokban vagy -mediánokban. A *post-hoc* latin eredetű kifejezés azt jelenti, hogy *ez után*. A post-hoc teszt elvégzése elengedhetetlen az előbb említett esetben, és nem helyettesíthető a páronkénti *t*-teszt elvégzésével. Ennek oka az, hogy a post-hoc tesztek ellenőrzés alatt tartják az α hibaértéket a végrehajtott műveletek számától függően úgy, hogy a teljes műveletsorra vonatkozó α értéke 0.05 (vagy az általunk megadott érték) legyen. Ezzel szemben a páronként elvégzett *t*-teszt mindegyikének az α értéke 0.05, így megnövekszik az összhiba, és fennáll a veszélye annak, hogy elsőfajú hibát követünk el.

A változók közti összefüggések kimutatására a 4.6. táblázatban felsorolt próbák állnak rendelkezésünkre. A Pearson- és a Spearman-próba két változó közötti összefüggést tesztel. A Pearson-próba folytonos, normális eloszlású változókra alkalmazható, a Spearman-próba viszont egy nem parametrikus teszt, ami nem támaszt feltételeket a változók eloszlásával kapcsolatban.

4.6. táblázat. *Statisztikai próbák a változók közti összefüggések kimutatására*

Próba	Típus	Változók típusa
Pearson	parametrikus	arányskálájú és intervallumskálájú
Spearman	nem parametrikus	arányskálájú, intervallumskálájú és ordinális
Khí-négyzet	nem parametrikus	ordinális és nominális
Kendall-féle tau	nem parametrikus	arányskálájú, intervallumskálájú és ordinális
Somers-féle delta	nem parametrikus	ordinális

A Pearson-próba lineáris összefüggést feltételez a két folytonos változó között. A szignifikanciaszint megállapításánál egy normális eloszlássorozatot képez, és megvizsgálja, hogy az eloszlások közti eltérések összege nagyobb, mint amit a véletlennek tulajdonítanánk. A Spearman-próba rangokat társít az értékekhez,

és vizsgálja, hogy létezik-e egy monoton növekedés vagy csökkenés, majd a szignifikanciát egy parametrikus teszttel állapítja meg. A Pearson-korrelációvizsgálat eredménye kevésbé van alávetve az első- és másodfajú hibának, mint a Spearman-próba. A Kendall-féle tau a Pearson-féle korreláció nem parametrikus verziójaként használható, de ordinális változókra is megbízhatóan működik. A Somers-féle delta egymással összefüggő ordinális változókra használható.

A χ^2-próba a függetlenség tesztelésére két ordinális vagy két nominális változó közötti összefüggést tesztel. A próba szignifikáns eredménye azt jelenti, hogy a két változó független egymástól. A χ^2-paraméter értékét a két változó értékeiből számolja ki, majd a χ^2-eloszlás és a szabadságfokok alapján megállapítható a *p*-érték.

Az ok-okozati kapcsolat vizsgálatára regresszióanalízis végezhető. A lényege az, hogy egy vagy több egymástól független magyarázó változó alapján modellt lehet készíteni egy függő változó értékeinek jósolására. A regresszióanalízis típusainak tárgyalására a 9. fejezetben kerül sor.

4.2. Adatsor normális eloszlásának megállapítása

A parametrikus próbák alkalmazásának feltétele az, hogy a minta normális eloszlású kell hogy legyen. Egy adatsor normális eloszlását vizuálisan lehet ellenőrizni, ill. statisztikus próbákkal tesztelhető. A vizuális ellenőrzés háromféle módon történhet: grafikusan (hisztogramon és egy Q-Q ábra kiértékelésével), a normalitás mutatóinak kiértékelésével (ferdeség és csúcsosság) és statisztikai próbák segítségével.

Grafikus vizsgálat

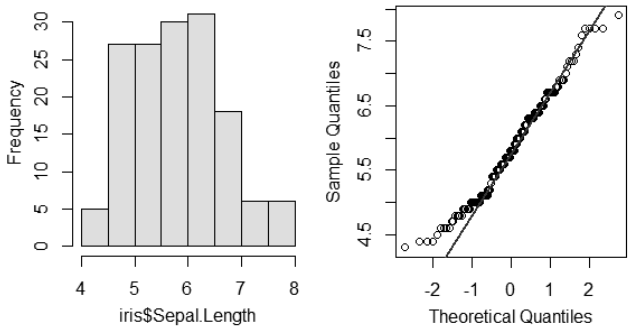
A hisztogram haranghoz közelítő alakja jelzi a normális eloszlást, a Q-Q ábrának a lényege az, hogy a vízszintes tengelyen a normális eloszlásnak megfelelő kvantiliseket tartalmazza, a függőleges tengelyen pedig a minta értékeit. Normális eloszlású adatsor esetén a pontok egy egyenes mentén helyezkednek el. Minél több pont tér el az egyenestől, annál kevésbé felel meg a minta eloszlása a normális eloszlásnak.

Az R adatbázisában levő *iris* adattáblában szerepelő három nősziromfaj csészelevélhosszának az eloszlását láthatjuk a 4.2. ábrán.

A 4.2. ábra R-kódja:

```
par(mfrow=c(1,2))
hist(iris$Sepal.Length, col = "grey90", main = "")
qqnorm(iris$Sepal.Length, pch = 21, main = "")
qqline(iris$Sepal.Length, col = "blue", lwd = 2)
```

A vizuális vizsgálat szerint a csészelevelek hossza megfelel a normális eloszlásnak.



4.2. ábra. *A normális eloszlás vizuális ellenőrzése*

Ferdeség és csúcsosság

A ferdeség (skewness) és a csúcsosság (kurtosis) két olyan mutató, amelyek az eloszlás alakjáról nyújtanak információt. A ferdeség az eloszlás aszimmetriáját méri. Ha értéke negatív, akkor az eloszlás bal oldali ferdeséget mutat, ha értéke pozitív, akkor jobb oldali ferdeség áll fenn. A csúcsosság azt méri, hogy az eloszlás lapultabb vagy csúcsosabb-e, mint a normális eloszlás. Ha a csúcsosság 3-nál kisebb szám, akkor az eloszlás lapultabb, mint a normális eloszlás, azaz kevesebb kiugró és extrém értéket jelez, mint a normális eloszlás. Ellenkező esetben az eloszlás csúcsos, és sokkal több kiugró és extrém értéket jelez, mint a normális eloszlás. Amennyiben a ferdeség értéke nulla és a csúcsosság értéke 3, az eloszlás egybeesik a normális eloszlással.

Normalitást tesztelő próba

A normális eloszlást tesztelő próbák nullhipotézise azt állítja, hogy a tesztelt adatsor eloszlása egyezik a normális eloszlással. A leggyakrabban használt próbák a Shapiro–Wilk-próba és a Kolmogorov–Smirnov-próba. A Shapiro–Wilk-próba a minta értékei és a normális eloszlás által elvárt értékek közötti korrelációt teszteli, a Kolmogorov–Smirnov-próba pedig az elméleti kumulatív eloszlásfüggvényt hasonlítja össze a minta empirikus kumulatív eloszlásfüggvényével. A Shapiro–Wilk-próbát főleg kis mintaelemszám mellett használják ($n < 50$), a Kolmogorov–Smirnov-próbát pedig a nagy mintaelemszámok esetében ($n > 50$). Az utóbbi próba érzékeny az extrém értékekre, viszont a Lilliefors-korrekció beépítésével ezt az érzékenységet korrigálták (Gashemi és Zahediasl, 2012).

A Jarque–Bera-próba azt a hipotézist teszteli, hogy a minta eloszlásának a csúcsossága és ferdesége megegyezik-e a normális eloszlásra jellemző értékekkel. A próba egy χ^2-eloszlást követő próbastatisztkát számol ki, aminek két szabadságfoka van:

$$JB = \frac{n}{6} \cdot \frac{S^2 + (K-3)^2}{4},$$

ahol S a ferdeség, a K pedig a csúcsosság értéke (Bowman és Shenton, 1975).

A különböző próbák megtalálhatók az R alapcsomagjában, illetve a {moments} csomagban (4.7. táblázat).

4.7. táblázat. *A normális eloszlás tesztelésére alkalmas R-parancsok*

Parancs	Részletek	Magyarázat
skewness(x) {moments} kurtosis(x) {moments} jarque.test(x) {moments} shapiro.test(x) ks.test(scale(x), "pnorm", alternative)		
x	adatsor	df$Ca, df[,1]
scale(x)	x adatsor standardizált változata	$\mu = 0$, sd = 1
alternative	"two-sided", "less", "greater"	alapból "two-sided"

```
#A csészelevelek eloszlásának tesztelése
library(moments)
skewness(iris$Sepal.Length)
[1] 0.3117531
kurtosis(iris$Sepal.Length)
[1] 2.426432

#Jarque-Bera-próba
jarque.test(iris$Sepal.Length)
Jarque-Bera Normality Test
data: iris$Sepal.Length
JB = 4.4859, p-value = 0.1061
alternative hypothesis: greater

#Shapiro-Wilk-próba
shapiro.test(iris$Sepal.Length)
Shapiro-Wilk normality test
data: iris$Sepal.Length
W = 0.97609, p-value = 0.01018

#Kolmogorov-Smirnov-próba
ks.test(scale(iris$Sepal.Length), "pnorm")
One-sample Kolmogorov-Smirnov test
data: scale(iris$Sepal.Length)
D = 0.088654, p-value = 0.1891
alternative hypothesis: two-sided
```

A ferdeség (0.312) és a csúcsosság (2.43) kismértékben eltérnek a 0 és 3 értékektől. Ebből az derül ki, hogy az eloszlás enyhe jobb oldali ferdeséget mutat, és kissé lapultabb, mint a normális eloszlás. A normalitást tesztelő próbák szerint a Shapiro–Wilk-próba kivételével az eloszlás nem tér el szignifikánsan a normális eloszlástól.

Általánosan elfogadott eljárás, hogy a minta eloszlását vizuálisan teszteljük. Tsagris és Pandis (2021) szerint a normalitást tesztelő próbák kis elemszámnál engedékenyebbek, mivel kicsi az erejük, viszont nagy elemszámnál a normálistól kismértékű eltérések esetén is szignifikáns eltérést adnak, ráadásul a különféle próbák eltérő eredményeket adhatnak.

A centrális határeloszlás tétele

A centrális határeloszlás tétele (CLT – Central Limit Theorem) megkönnyíti a statisztikai adatfeldolgozást, ha a mintavételt úgy tervezzük meg, hogy nagy mintaelemszámmal dolgozzunk. A tétel kimondja, hogy a nagy elemszámú minták átlagai normális eloszlást követnek, függetlenül a populáció eloszlásától. Ilyen esetben a parametrikus próbák megbízhatóan használhatók akkor is, ha a minta eloszlása nem normális.

5. HIPOTÉZISVIZSGÁLATOK FOLYTONOS VÁLTOZÓKRA

5.1. Az egymintás *t*-próba

Az egymintás *t*-próba arra alkalmas, hogy egy normális eloszlású minta átlagértékét összehasonlítsuk egy számmal, a populáció várható értékével (μ). A szignifikanciát a *t*-paraméter segítségével állapítjuk meg, amelynek Student-féle *t*-eloszlása van. A próba alkalmazásának feltételei:

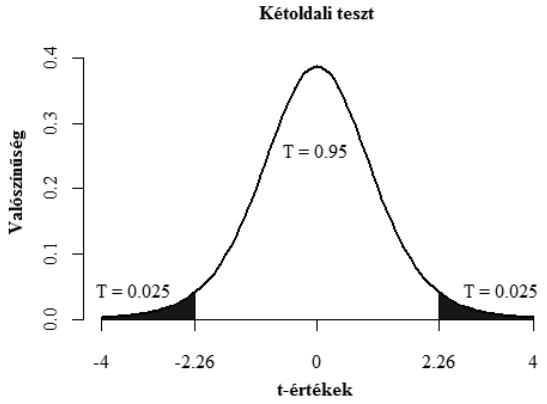
– egyszerű véletlen mintavétel;
– normális eloszlású folytonos változó;
– a populáció varianciája nem ismert, a minta varianciájával becsüljük meg;
– a minta elemszámát a hatásnagyság is meghatározza, ha az utóbbit előre leszögezzük.

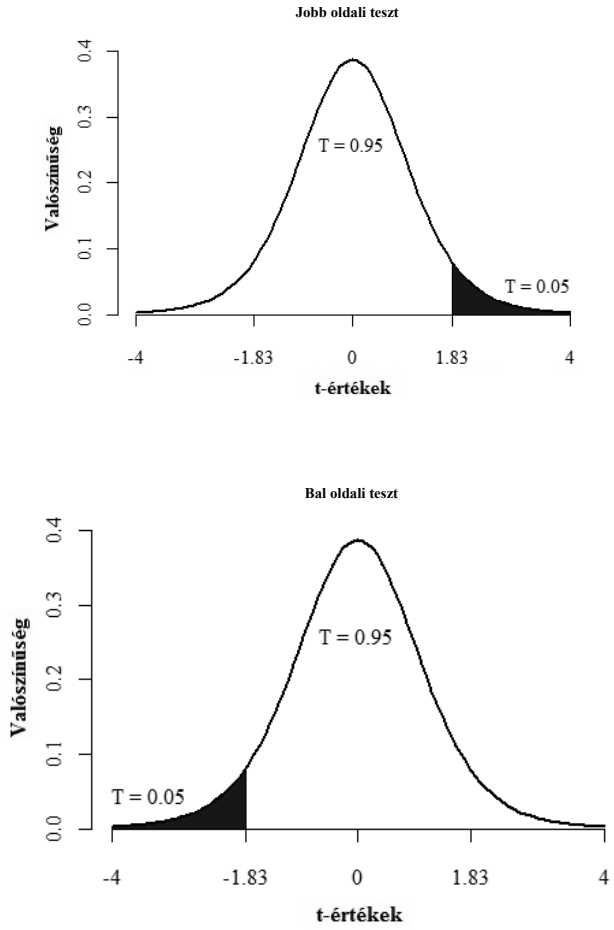
A *t*-próba nullhipotézise azt jelenti ki, hogy a mintaátlag egyenlő egy adott számmal (a populáció várható értékével (μ)). Az ellenhipotézis lehetséges esetei: a mintaátlag nem egyenlő / kisebb / nagyobb az illető számnál.

A próbastatisztika (*t*-érték) a következő kifejezéssel értelmezhető:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}.$$

A μ a populáció középértéke (várható érték), az s a mintából számolt szórás. Az α ismeretében azt vizsgáljuk meg, hogy a számított *t*-érték az elfogadási vagy a kritikus tartományba esik a Student-féle eloszlásfüggvényen *df* szabadságfokkal. Három esettel állunk szemben, annak függvényében, hogy bal oldali, jobb oldali vagy kétoldali tesztet végzünk (5.1. ábra).





5.1. ábra. *A három eset grafikus szemléltetése az alternatív hipotézis megfogalmazása szerint a* t*-eloszlásfüggvényen (*df *= 9), ha* $\alpha = 0.05$

A kétoldali teszt ábrájának R-kódja:

```
#A t-eloszlásfüggvény (df = 9), [-4,4]
par(family="serif") #Mindent TNR betűtípussal ír az ábrára
curve(dt(x, df = 9), #Generálunk egy t-eloszlást
 xlim = c(-4, 4), ylim = c(0,0.4),
 main = "Kétoldali teszt", cex.main = 1,
 yaxs = "i", font.lab = 2, cex.lab = 1,
 xlab = "t-értékek",
```

```
 ylab = "Valószínűség",
 lwd = 2,
 axes = "F") #Nem tünteti fel a tengelyeket.
#A tengelyek önkényes formázása
axis(1,
 at = c(-4, -2.26, 0, 2.26, 4),
 padj = 0.5,
 labels = c(-4, -2.26, 0, 2.26, 4))
axis(2)
#Elutasítási tartományok kijelölése poligonnal
polygon(x = c(2.26, seq(2.26, 4, 0.01), 4),
 y = c(0, dt(seq(2.26, 4, 0.01), df = 9), 0),
 col = "darkblue")
polygon(x = c(-4, seq(-4,-2.26, 0.01), -2.26),
 y = c(0, dt(seq(-4, -2.26, 0.01), df = 9), 0),
 col = "darkblue")
text(-0.02,0.26, "T = 0.95")
text(-3.4,0.045, "T = 0.025")
text(3.4,0.045, "T = 0.025")
```

Ha a *t*-érték a kritikus tartományba esik (kék terület), elvetjük a H_0 hipotézist α szignifikanciaszint mellett. A kétoldali próbánál nagyobb kritikus *t*-értéknél vetjük el a H_0 hipotézist a jobb oldali próbához viszonyítva, és kisebb *t*-érték mellett a bal oldali próbához viszonyítva. Változatlan α-érték mellett az egyoldali próbáknál engedékenyebbek vagyunk a H_0 hipotézis elvetésénél, mint a kétoldali próbánál, mivel nagyobb valószínűséggel követhetünk el elsőfajú hibát. Ezért egyoldali próbáknál az α-értéket felére lehet csökkenteni, azaz 0.025-re tenni. A *t*-próba R kódjának részletei az 5.1. táblázatban láthatók.

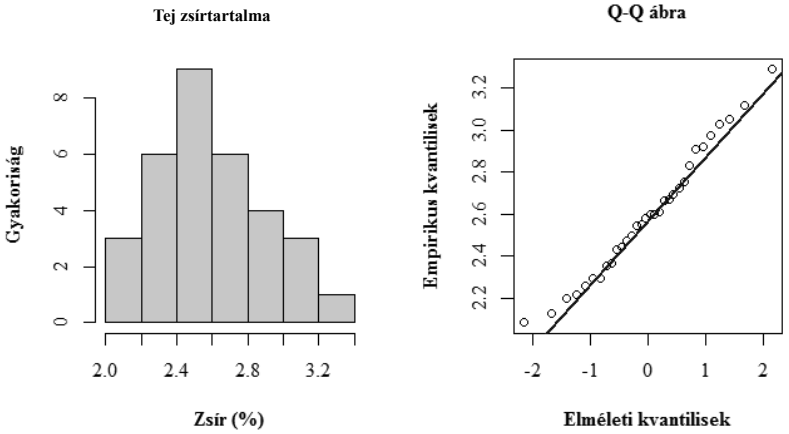
5.1. táblázat. *Az egymintás* t*-próba elvégzése R-ben*

Parancs	Részletek	Magyarázat
t.test(adatsor, mu, alternative, conf.level)		
adatsor	pl. a Ca változó a df adattáblából	df$Ca, df[,1]
mu	pl. mu = 59	a várt érték (μ)
alternative	"two.sided", "less", "greater"	a kétoldali teszt alapból működik
conf.level	0.95, 0.90 stb.	Az 1–α-érték, alapból 0.95

Egy tejgyár mintát vesz a legyártott sorozatból, a minta elemszáma pedig 32. Ellenőrzik a tej zsírtartalmát, a mintaátlag 2.81%, a szórás pedig 0.39%. Eltér-e szignifikánsan a minta átlagértéke a várt 3%-tól?

H_0: $\bar{x} = 3\%$
H_1: $\bar{x} \neq 3\%$
Az $\alpha = 0.05$, a próba kétoldali.

Az egymintás *t*-teszt feltétele, hogy a minta normális eloszlást kövessen. Az 5.2. ábrán látható a minta eloszlása és a Q-Q ábra, ami megközelíti a normális eloszlást. R-ben elvégezhetjük a Shapiro–Wilk-tesztet a normalitás ellenőrzésére.



5.2. ábra. *A minta eloszlásának vizsgálata*

Az 5.2. ábra kódja R-ben:

```
set.seed(1542)  #A generált normális adatsor rögzítése.
tej = rnorm(32, 2.77, 0.39)
#Összetett ábra szerkesztése (két ábra egymás mellé)
par(mfrow = c(1,2), family = "serif") #TNR betűtípus
hist(tej, main = "Tej zsírtartalma", xlab = "Zsír (%)",
 ylab = "Gyakoriság", font.lab = 2, cex.main = 1)
qqnorm(tej, xlab = "Elméleti kvantilisek",
 ylab = "Empirikus kvantilisek",
 main = "Q-Q ábra", font.lab = 2, cex.main = 1)
qqline(tej, col = "darkblue", lwd = 2)
```

Kiszámoljuk a *t*-értéket:

$$t = \frac{2.81 - 3}{\frac{0.39}{\sqrt{32}}} = -2.76$$

A kritikus *t*-érték ($\alpha = 0.05$, $df = 31$): −2.04. Mivel a számított *t*-érték kisebb a kritikus *t*-értéknél, *t* a kritikus tartományba esik, ezért elvetjük a H_0 hipotézist.

Tehát a minta alapján megállapítható, hogy a 2.81%-os átlagérték nem a véletlen ingadozás eredménye, hanem a tej minősége gyengült.

A teszt elvégzése R-ben: legyen az adattábla d.

```
head(d) #az első hat sor
 Zsir
1 2.34
2 2.63
3 3.65
4 2.74
5 2.73
6 2.74

mean(d) #átlag
[1] 2.775
sd(d) #szórás
[1] 0.413

#Normális eloszlás tesztelése
shapiro.test(d$Zsir)
Shapiro-Wilk normality test
data: d$Zsir
W = 0.976, p-value = 0.6675
#A teszt p-értéke > 0.05,
#tehát elfogadjuk a normális eloszlást.

#Az egymintás t-próba
t.test(d$Zsir, mu = 3)
data: tej$Zsir
t = -3.0774, df = 31, p-value = 0.004342
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.626402 2.924223
sample estimates:
mean of x
 2.775313
```

A *t*-próba szerint a *p*-érték = 0.008 < 0.05. Ez alapján elvetjük a H_0 hipotézist, tehát a minta azt jelzi, hogy a tej minősége nem felel meg az elvárásnak. Mivel kétoldali tesztet végeztünk, nem volt szükség az *alternative* utasítás megváltoztatására. Az R az eredményben megadja a *t*-értéket, a konfidenciaintervallumot a

változóra, és jelzi a teszt típusát azáltal, hogy kiírja az alternatív hipotézist. Ebben az esetben a *nem egyenlő* szerepel (tehát a teszt kétoldali).

A szignifikáns próba megköveteli a hatásnagyság megállapítását. Az egymintás *t*-próbára alkalmas a Cohen-féle *d* kiszámítása. Cohen a következő kiértékelést javasolta *d*-nél a kicsi, közepes és nagy hatásnagyságra: 0.2–0.3, 0.5 és 0.8. A *d* tulajdonképpen azt mutatja meg, hogy az átlag és a várt érték különbsége hányad része a szórásnak.

$$d = \frac{\bar{x} - \mu}{s}$$

A tejmintákra a *d* értéke 0.54, tehát közepes méretű eltérésről beszélünk, így helyes döntésnek látszik a H_0 hipotézis elvetése. R-ben a Cohen-féle *d*-t az {effectsize} csomag segítségével számolhatjuk ki.

```
library(effectsize)
cohens_d(tej$Zsir, mu = 3)
Cohen's d | 95% CI
-------------------------
-0.54 | [-0.93, -0.17]
```

5.2. Nem parametrikus próbák a várt érték és a medián összehasonlítására

5.2.1. Előjelpróba

Az előjelpróba azt teszteli, hogy a minta mediánja egyenlő-e egy feltételezett mediánnal. Ezt úgy valósítja meg, hogy meghatározza azoknak a mintaelemeknek a számát, amelyek értékei nagyobbak a feltételezett mediánnál (N^+), majd a kritikus tartományt egy olyan binomiális eloszláson határozza meg, amelyre *n* a minta elemszámának az a része, ami nem egyenlő a feltételezett mediánnal, a siker valószínűsége pedig 0.5.

A 2000-es évek elején Kelet-Európa országaiban felmérték a költő gólyapárok populációját (BirdLife International, 2015). Romániában a becsült szám 2.3/100 km^2. Különbözik-e szignifikánsan a romániai populációsűrűség a kelet-európai országok középértékétől? Az országok adatai az 5.2. táblázatban láthatók. Az adatok nem normális eloszlásúak, Lengyelország, Bosznia-Hercegovina és Albánia extrém kiugró értékekkel szerepelnek. Ilyen esetben egy nem parametrikus próbát kell alkalmazni.

5.2. táblázat. *Kelet-Európa országaiban költő gólyapárok populációsűrűsége*

Ország	Bulgária	Moldova	Szerbia	Szlovénia	Horvátorsz.
Költőpárok /100 km^2	4.6	1.5	1.5	1.1	2.1
Előjel	+	–	–	–	–
Ország	Bosznia-HG	Albánia	Lengyelorsz.	Ukrajna	Görögorsz.
Költőpárok /100 km^2	0.1	0.03	16.9	4.9	1.5
Előjel	–	–	+	+	–

Az előjelpróba nullhipotézise: a minta mediánja = elméleti medián = 2.3. Előjeleket társítunk az országokhoz, alapul véve az elméleti mediánt (2.3). Az N^+ értéke 3, az N^- értéke pedig 7. A 10 országból 10-nél az érték nem volt egyenlő a mediánnal, tehát a binomiális eloszlás n-értéke 10, $p = 0.5$. Az előjelpróbát R-ben a **`binom.test()`** paranccsal végezhetük el (5.3. táblázat).

5.3. táblázat. *A binomiális teszt értelmezése R-ben*

Parancs	Részletek	Magyarázat
binom.test(x, n, p = 0.5, alternative, conf.level)		
x	N^+ vagy c(N^+, N^-)	a sikerek, vagy sikerek és kudarcok száma
n	max. minta elemszáma	a mediánnal nem egyenlő értékek összege
alternative	"two.sided", "less", "greater"	alapból kétoldali próba
conf.level	0.95, 0.90, 0.99 stb.	az (1-α)-érték, alapból 0.95

```
    binom.test(3,10)
    Exact binomial test
    data: 3 and 10
    number of successes = 3, number of trials = 10, p-value = 0.3438
    alternative hypothesis: true probability of success is not
equal to 0.5
    95 percent confidence interval:
     0.06673951 0.65245285
    sample estimates:
    probability of success: 0.3
```

A binomiális teszt alapján, p = 0.3438 > 0.05, megtartjuk a H0 hipotézist, tehát a 2.3/100 km^2-es érték nem tér el a kelet-európai mediánértéktől.

5.2.2. Wilcoxon-féle előjeles rangpróba

A próbát Frank Wilcoxon amerikai vegyész dolgozta ki 1945-ben. Az előjelpróbához hasonlóan a Wilcoxon-próba azt teszteli, hogy a minta mediánja egyenlő-e

egy feltételezett mediánnal. A próbastatisztika (V) értékét egy kritikus értékhez hasonlítjuk, adott szignifikanciaszint mellett. A W két érték (V^+ és V^-) közül a kisebb. A két statisztikai mutatót úgy kapjuk meg, hogy kiszámoljuk az adatok eltérését a feltételezett mediántól, a különbségek abszolút értékeit rangsoroljuk, majd összeadjuk a mediántól kisebb, illetve nagyobb eltérések rangjait. A két érték közül a kisebbet választjuk próbastatisztikának. Ha a $V > 20$, akkor a kritikus értéket egy normál eloszlás alapján állapítjuk meg, amelynek a középértéke és a szórása:

$$\mu = \frac{n(n+1)}{4} \text{ és } \sigma = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

Ha valamely érték egyenlő a feltételezett mediánnal, akkor azt kihagyjuk a sorból, így csökken az elemszám. Egyenlő eltérést adó értékeket kapcsolt értékeknek (*ties*) nevezzük, és mindegyikük a rájuk eső rangok átlagát kapja (pl. ha a 7. és 8. rangra azonos eltérés jut, akkor mindkét eltérés 7.5-ös rangot kap). Kapcsolt értékekre módosul a normális eloszlás szórása:

$$\sigma = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}},$$

ahol t a kapcsolt értékeknek a rangjai.

Annak ellenére, hogy a Wilcoxon-próba nem parametrikus, alkalmazása bizonyos feltételekhez kötött:

– az adatok eltérései a feltételezett mediántól folytonos skálán mozognak;
– a minta eloszlása szimmetrikus;
– a megfigyelések/mérések egymástól függetlenek;
– az adatsor legalább ordinális skálájú.

A Wilkoxon-próbához az 5.2. táblázat adatai további adatokkal bővülnek (5.4. táblázat). A medián értéke 2.3 (a Romániára közölt adat). A táblázatból kitűnik, hogy a Moldovára, Szerbiára és Görögországra kapott eltérések azonosak, tehát ezek kapcsolt értékek. Mind a három ország rangja 3, a 2, 3, 4 helyett.

5.4. táblázat. *Kelet-Európa országaiban költő gólyapárok populációsűrűsége*

Ország	Bulgária	Moldova	Szerbia	Szlovénia	Horvátorsz.
Költőpárok /100 km^2	4.6	1.5	1.5	1.1	2.1
Eltérés a mediántól	2.3	−0.8	−0.8	−1.2	−0.2
Rang (abszolút értékekre)	8	3	3	5	1
Előjel	+	–	–	–	–
Ország	Bosznia-HG	Albánia	Lengyelorsz.	Ukrajna	Görögorsz.
Költőpárok /100 km^2	0.1	0.03	16.9	4.9	1.5

Ország	Bulgária	Moldova	Szerbia	Szlovénia	Horvátorsz.
Eltérés a mediántól	−2.2	−2.27	14.6	2.6	−0.8
Rang	6	7	10	9	3
Előjel	–	–	+	+	–

A pozitív előjelű adatokhoz tartozó rangok összege: $V^+ = 27$, a negatív előjelű adatokhoz tartozóké pedig: $V^- = 28$. A V-érték ebben az esetben 27.

Ha nincsenek kapcsolt értékek, R-ben a **`wilcox.test()`** parancsot használjuk. Kapcsolt értékek esetében a próba hibaüzenetet ír ki. Ilyenkor az {exactRankTest} csomagban levő **`wilcox.exact()`** próbát használjuk, mint az előbb említett példában is (5.5. táblázat).

5.5. táblázat. *A Wilcoxon-próba értelmezése R-ben*

Parancs	Részletek	Magyarázat
wilcox.test(x, mu, alternative, conf.level) wilcox.exact(x, mu, alternative, conf.level) {exactRankTest}		
x	adatsor	egy oszlop egy adattáblából vagy egy vektor
mu	feltételezett medián	
alternative	"two.sided", "less", "greater"	alapból kétoldali próba
conf.level	0.95, 0.90, 0.99 stb.	az (*1-α*)-érték, alapból 0.95

A fészkelő gólyapárokra teszteljük, hogy a romániai átlag (2.3/100 km^2) eltér-e szignifikánsan a többi kelet-európai ország középértékétől. Ezt a Wilcoxon-próbával végezzük el ($\alpha = 0.05$).

```
golya = c(4.6,1.5,1.5,1.5,1.1,2.1,0.1,0.03,16.9,4.9)
wilcox.test(golya, 2.3)
Wilcoxon signed rank test with continuity correction
data: golya
V = 27, p-value = 1
alternative hypothesis: true location is not equal to 2.3
Warning message:
In wilcox.test.default(golya, mu = 2.3) :
 cannot compute exact p-value with ties
```

A teszt nem adott szignifikáns eltérést ($p = 1$), viszont hibaüzenetet írt ki, miszerint az adatsorban kapcsolt értékek vannak. Így a helyzetnek megfelelő próbát végezzük el.

```
wilcox.exact(golya, 2.3)
Exact Wilcoxon signed rank test
```

```
data: golya
V = 27, p-value = 1
alternative hypothesis: true mu is not equal to 2.3
```

Az eredmény ebben az esetben is ugyanaz. Nincs szignifikáns különbség a romániai érték és a kelet-európai medián között.

5.3. A kétmintás *t*-próba

A kétmintás *t*-próbával két minta átlagát hasonlítjuk össze. A nullhipotézis azt állítja, hogy a minták egy populációból származnak, az átlagok között nincs eltérés, az ellenhipotézis pedig három lehetőséget ad: nem egyenlő / kisebb / nagyobb. A két minta lehet független egymástól (két területről származók, vagy két módszerrel kapott adatsorok, vagy egy kontroll- és egy tesztcsoportra kapott adatsorok), és lehetnek egymástól függők (önkontrollos csoport, előtte-utána típusú adatsorok, időben megismételt kísérletek, mintavételek). A minták típusa szerint változik a t-próba is, ugyanis a szórásnál figyelembe kell venni azt a tényt, hogy ugyanarról a területről származó minták vagy ugyanazokkal az alanyokkal/műszerrel elvégzett kísérletek eredményei kevésbé szóródnak, mint az egymástól független adatsorok értékei.

5.3.1. Kétmintás t-próba egyenlő varianciák esetében

A próba bizonyos feltételekhez kötött, amelyeknek teljesülniük kell ahhoz, hogy megbízható eredményt kapjunk. Ezek a feltételek a következők:

– egyszerű véletlen mintavétel;
– minták függetlensége;
– normális eloszlású folytonos változó mindegyik mintára;
– a populáció varianciája nem ismert, a minták varianciáival becsüljük meg;
– a varianciák egyenlősége;
– ha a minta elemszáma nagyobb 30-nál, akkor a két adatsor normális eloszlása nem feltétele a *t*-próbának (a centrális határeloszlás tétel miatt).

A nullhipotézis kijelenti, hogy a két minta átlaga egyenlő (egy populációból származnak). Az ellenhipotézis lehetséges esetei: a mintaátlagok nem egyenlők, az egyik kisebb vagy nagyobb a másiknál.

A próbastatisztika (*t*-érték) a következő kifejezéssel értelmezhető:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_d^2}{n_1} + \frac{s_d^2}{n_2}}},$$

ahol s_d^2 a közös variancia becslése a minták varianciáiból:

$$s_d^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}.$$

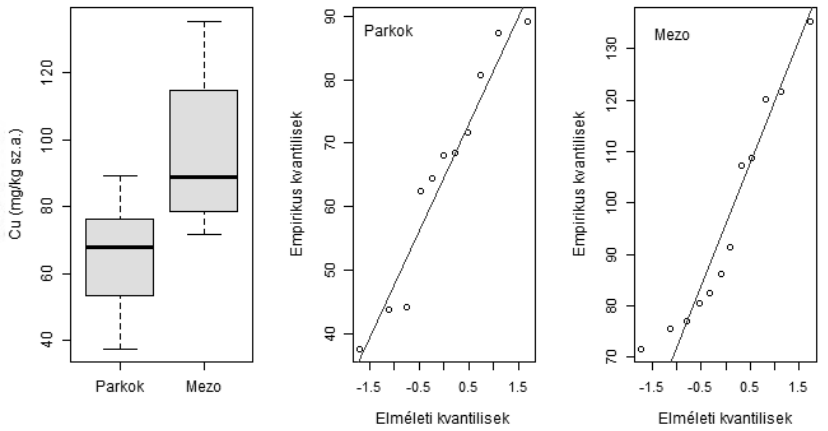
Az α ismeretében azt vizsgáljuk meg, hogy a számított *t*-érték az elfogadási vagy a kritikus tartományba esik a Student-féle eloszlásfüggvényen, amelynek df $= n_1 + n_2 - 2$ szabadságfoka van. A számolás időigényes, ezért a kétmintás *t*-próbát az R-ben célszerű elvégezni, ahol az eredmény az a *p*-érték, ami a számított *t*-értéknek felel meg.

Kolozsvár területén és a környékén vizsgáltuk a mezei csiperke (*Agaricus campestris*) elemösszetételét (Zsigmond et al., 2018). Egyik fontos elem a réz, amit a csiperkék általában hatékonyan felhalmoznak a termőtestben, ugyanakkor a városokban is feldúsul a szálló porban, ami a talajra ülepedve elérhetővé válik a gombák számára. Az adatok az 5.6. táblázatban találhatók.

5.6. táblázat. *A csiperkegombák kalapjának Cu-tartalma (mg/kg) száraz anyagra*

Parkok	44.2	87.4	37.6	68.5	64.4	62.4	43.9	89.2	71.7	80.7	68.0	–
Mező	121.7	108.9	77.0	107.4	91.3	82.6	135.4	120.2	71.7	75.5	86.3	80.5

Van-e szignifikáns különbség a mezei csiperke kalapjának a réztartalmában az élőhely függvényében? Az 5.3. ábrán láthatók a két területről származó csiperke réztartalmának a dobozdiagramjai, valamint a Q-Q ábrák is. A dobozdiagram szerint a mezőről származó csiperkék kalapjában több réz található, mint a parkokból származókban. A kérdés megválaszolására *t*-próbát végzünk.



5.3. ábra. *Kolozsvár parkjaiból és a környező mezőkről származó mezei csiperke kalapjának réztartalma dobozdiagramon és Q-Q ábrákon*

```
#Csiperkekalapok parkokból és mezőről
d = read.csv("Agaricus_caps.csv", header = T, row.names = 1)
#Három ábra szerkesztése
par(mfrow = c(1,3))
#Dobozdiagram
boxplot(d$Cu~d$Source, col = "grey90",
 xlab = "", ylab = "Cu (mg/kg sz.a.)",
 names = c("Parkok","Mező"), cex.lab = 1.1,
 cex.axis = 1.1)
#Q-Q ábrák
qqnorm(d$Cu[d$Source == "city park"],
 main = "", xlab = "Elméleti kvantilisek",
 ylab = "Empirikus kvantilisek", cex.lab = 1.1)
qqline(d$Cu[d$Source == "city park"], lwd = 1.5, col = "blue")
text(-1.1,88,"Parkok", cex = 1.1)

qqnorm(d$Cu[d$Source == "meadow"],
 main = "", xlab = "Elméleti kvantilisek",
 ylab = "Empirikus kvantilisek", cex.lab = 1.1)
qqline(d$Cu[d$Source == "meadow"], lwd = 1.5, col = "blue")
text(-1.1,133,"Mező", cex = 1.1)
```

A minták egymástól függetlenek, a minták elemszámai 11 (parkok) és 12 (mező). A Q-Q ábrák szerint a pontok viszonylag jól illeszkednek az egyeneshez. A normalitás tesztelésére Shapiro–Wilk-próbát használunk. A varianciák egyenlőségének a tesztelésre az *F*-próba használható. A független mintákra alkalmas kétmintás *t*-próba R kódjának részletei a 5.7. táblázatban láthatók.

5.7. táblázat. *Az egymintás* t*-próba elvégzése R-ben*

Parancs	Részletek	Magyarázat
t.test(adatsor~csoport, var.equal, alternative, conf.level)		
adatsor~csoport	pl. d$Cu~d$Csoport	függő~független változók; a függő és független változók mindkét csoport adatait tartalmazzák
var.equal	pl. var.equal = TRUE	TRUE opciót választjuk egyenlő varianciákra
alternative	"two.sided", "less", "greater"	a kétoldali teszt alapból működik
conf.level	0.95, 0.90, 0.99 stb.	az (1-α)-érték, alapból 0.95

A *t*-próbát a következő R-kóddal végezhetjük el:

```
head(d)
 Cu Source
K1 44.24 city park
K2 87.35 city park
K3 37.64 city park
K4 68.48 city park
K5 64.41 city park
K6 62.40 city park
dim(d)
[1] 23 2

#Normális eloszlás tesztelése
shapiro.test(d$Cu[d$Source == "city park"])
Shapiro-Wilk normality test
data: d$Cu[d$Source == "city park"]
W = 0.93211, p-value = 0.4326
shapiro.test(d$Cu[d$Source == "meadow"])
Shapiro-Wilk normality test
data: d$Cu[d$Source == "meadow"]
W = 0.90675, p-value = 0.1938
#A két p-érték > 0.05. Elfogadjuk H0-t.
#A két csoport normális eloszlású.

#A varianciák egyenlőségének tesztelése F-teszttel.
var.test(d$Cu~d$Source)
F test to compare two variances
data: d$Cu by d$Source
F = 0.66705, num df = 10, denom df = 11, p-value = 0.5314
alternative hypothesis: true ratio of variances is not equal
to 1
95 percent confidence interval:
 0.1891988 2.4446909
sample estimates:
ratio of variances
 0.6670527
#Az F-teszt nem adott szignifikáns eredményt (p > 0.05).
#Elfogadjuk H0-t. A varianciák közt nincs különbség.

#Elvégezzük a t-próbát
t.test(d$Cu~d$Source, var.equal = T)
Two Sample t-test
```

```
    data: d$Cu by d$Source
    t = -3.8331, df = 21, p-value = 0.0009677
    alternative hypothesis: true difference in means is not equal
to 0
    95 percent confidence interval:
     -48.24376 -14.30700
    sample estimates:
    mean in group city park mean in group meadow
     65.26545 96.54083
```

A *t*-próba szignifikáns eredményt adott ($t = -3.83$, $df = 21$, $p = 0.00097$), ez azt jelenti, hogy elvetjük a H_0 hipotézist, tehát szignifikáns különbség van a két élőhelyről származó csiperke kalapjának réztartalma között.

A Cohen-féle *d* a csiperkeadatokra:

$$d = \frac{96.54 - 65.27}{\sqrt{\frac{10 \cdot 302.9 + 11 \cdot 454.1}{21}}} = 1.6.$$

A Cohen-féle *d*-érték erős hatást mutat, így nagy valószínűséggel jól döntöttünk a H_0 hipotézis elvetésekor. A próba elvégezhető R-ben az {effectsize} csomag segítségével, amelyben a **cohens_d()** parancsot használjuk, ha $n > 20$, és **hedges_g()**, ha $n \leq 20$. A minták elemszáma 23, tehát a két próba nagyjából egyforma eredményt ad.

```
library(effectsize)
cohens_d(d$Cu, d$Source)
Cohen's d | 95% CI
--------------------------
-1.60 | [-2.53, -0.64]

hedges_g(d$Cu, d$Source)
Hedges' g | 95% CI
-------------------------
-1.54 | [-2.44, -0.61]
```

5.3.2. Kétmintás t-próbák nem egyenlő varianciák esetében

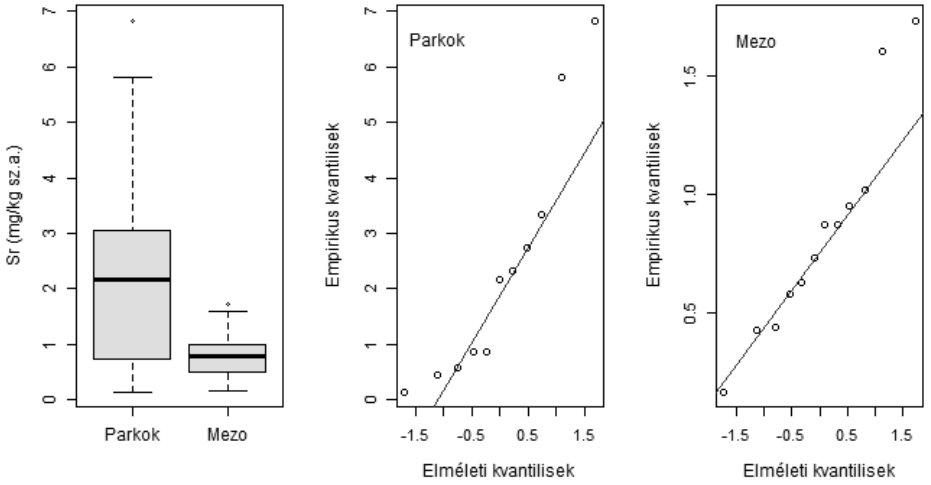
Welch-féle U-próba

Olyan adatsoroknál, amelyek függetlenek egymástól, normális eloszlásúak, de a varianciák nem egyenlők, a Welch-féle *U*-próbát végezzük el (Welch, 1938). A próbastatisztika (*t*-érték) a következő kifejezéssel értelmezhető:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

A szabadságfokok számának kiszámítása bonyolult, nem mindig egész szám, ilyenkor a legközelebb álló egész számra kerekítjük.

Az 5.3.1. alfejezetben említett tanulmánynak egy másik adatsorát vizsgáljuk. Keressük a választ arra a kérdésre, hogy van-e szignifikáns különbség a mezei csiperke kalapjának stronciumtartalma között élőhelytől függően. Az 5.4. ábrán láthatók a két területről származó csiperkék stronciumtartalmának a dobozdiagramjai és a Q-Q ábrák. A dobozdiagram szerint a parkokból származó csiperkék kalapjában több stroncium található, mint a mezőről származókban. A dobozdiagramon látható kiugró értékek a Q-Q ábrán messze esnek az egyenestől, viszont a pontok többsége az egyeneshez közel helyezkedik el.



5.4. ábra. *Kolozsvár parkjaiból és egy közeli mezőről származó mezei csiperke kalapjának stronciumtartalma dobozdiagramon és Q-Q ábrákon*

A feltételeket (a minták normális eloszlását és a varianciák egyenlőségét) R-ben ellenőrizzük.

```
#Normális eloszlás tesztelése
tapply(d$Sr,d$Source,shapiro.test)
$`city park`
Shapiro-Wilk normality test
data: X[[i]]
```

```
    W = 0.86292, p-value = 0.06282
    $meadow
    Shapiro-Wilk normality test
    data: X[[i]]
    W = 0.92568, p-value = 0.3366

    #Varianciák egyenlőségének tesztelése
    var.test(d$Sr~d$Source)
    F test to compare two variances
    data: d$Sr by d$Source
    F = 23.325, num df = 10, denom df = 11, p-value = 1.104e-05
    alternative hypothesis: true ratio of variances is not equal
to 1
    95 percent confidence interval:
     6.615888 85.485760
    sample estimates:
    ratio of variances
     23.32545
```

A próbákból az következik, hogy a két minta normális eloszlású ($W = 0.863$, $p = 0.063$, és $W = 0.926$, $p = 0.337$), a varianciák viszont nem egyenlők ($F = 23.32$, $df_1 = 10$, $df_2 = 11$, $p < 0.001$). Ilyen esetben a Welch-féle *U*-próbát ajánlott használni.

```
    #A Welch-féle t-teszt
    t.test(d$Sr~d$Source, var.equal = F)
    Welch Two Sample t-test
    data: d$Sr by d$Source
    t = 2.2599, df = 10.786, p-value = 0.04555
    alternative hypothesis: true difference in means is not equal
to 0
    95 percent confidence interval:
     0.03648396 3.04260695
    sample estimates:
    mean in group city park mean in group meadow
     2.374545 0.835000
```

A Welch-féle *U*-próba ($t = 2.26$, $df = 11$, p-value $= 0.0456$) szignifikáns különbséget adott a két területről származó csiperke átlag stronciumtartalmára, noha a *p*-érték közel van az α szignifikanciaszinthez.

Mivel a varianciák nem egyenlők, ezért a hatásnagyságot a Glass-féle deltával adjuk meg külön-külön mind a két szórásra.

$$\Delta_1 = \frac{\bar{x}_1 - \bar{x}_2}{s_1} = \frac{2.375 - 0.835}{0.459} = 3.35,$$
$$\Delta_2 = \frac{\bar{x}_1 - \bar{x}_2}{s_2} = \frac{2.375 - 0.835}{2.216} = 0.70.$$

A két érték között nagy különbség van, közepes, ill. nagyon nagy hatást jeleznek. R-ben az {effectsize} csomagban levő **glass_delta()** paranccsal számolhatjuk ki az értéket. A konfidenciaintervallum nagyon tág, tehát az érték nem megbízható. Ennek oka a viszonylag kis mintaelemszám.

```
library(effectsize)
glass_delta(d$Sr~d$Source)
Glass' delta | 95% CI
--------------------------
3.35 | [0.07, 6.52]
```

Érdekes eredményhez jutottunk: egy kismértékben szignifikáns Welch-próbához ($p \sim \alpha$) jelentős hatásnagyság (~ 3.35) társult. A 4.6. ábrán látható, hogy mindkét csoport eloszlásán kiugró értékek azonosíthatók, amelyek olyan átlagokhoz vezetnek, amelyek nagymértékben torzítják a valós középértéket. Ezért érdemes a Welch-próba helyett a Yuen–Welch-próbát elvégezni és az annak megfelelő hatásnagyságot kiszámolni.

Yuen–Welch-próba

1974-ben Karen K. Yuen közölt egy módszert kiugró értékeket tartalmazó adatsorok átlagának összehasonlítására (Yuen, 1974). Az átlag erősen érzékeny a kiugró értékekre, így a *t*-próba olyan átlagokat hasonlít össze, amelyek nem jellemzik megfelelően az eloszlások középértékeit. Kiugró értékek mellett az eloszlás elnyúlik ezeknek az értékeknek az irányába, ezért Yuen a trimmelt átlag használatát javasolta. Ennek értelmében újraértékelte a *t*-eloszlás szabadságfokainak a számát. A trimmelt átlagnál megengedett a 20%-os adateltávolítás az adatsor mindkét végéről, viszont túl sok adat eltávolítása bizonytalanná teszi a próba eredményét. Ha a trimmelést nem végezzük el, akkor a Yuen-próba megegyezik a Welch-próbával.

Az előző példában láttuk, hogy a stroncium dobozdiagramjain mind a két csoportban voltak kiugró értékek, annak ellenére, hogy az eloszlások normálisnak tekinthetők a Shapiro–Wilk-próba szerint. Ezért a Yuen–Welch-próba alkalmasabb az átlagok egyenlőségének tesztelésére. R-ben a Yuen–Welch-próbát a WRS2 csomag **yuen()** parancsával lehet elvégezni.

```
#Yuen-Welch-próba a csiperke Sr-tartalmára
library(WRS2)
```

```
yuen(d$Sr~d$Source, tr = 0.2) #Trimmelés 20%
Test statistic: 1.8144 (df = 6.39), p-value = 0.1166
Trimmed mean difference: 1.08161
95 percent confidence interval:
-0.356 2.5192
Explanatory measure of effect size: 0.65
```

A Yuen–Welch-próba nem adott szignifikáns eredményt a csiperke átlag stronciumtartalmára ($t = 1.814$, $df = 6$, $p = 0.117$). Tekintettel a kiugró értékek jelenlétére, a Yuen–Welch-próba eredménye megbízhatóbb a Welch-próbánál. A **yuen()** parancs megadja a Cohen-féle hatásnagyság korrigált értékét is, ami ebben az esetben 0.65. A hatásnagyság kiszámolásához trimmelt átlagot és winsorizált varianciákat (s_w^2) használunk (Algina et al., 2005). A winsorizálás úgy kezeli a kiugró értékeket, hogy az adatsor két végén azonos százalékban kijelöl egy-egy szakaszt, és egy szakaszhoz tartozó értékeket a szakasz közvetlen szomszédságában levő értékkel helyettesíti.

$$s_w^2 = \frac{(n_1 - 1)s_{w1}^2 + (n_2 - 1)s_{w2}^2}{n_1 + n_2 - 2}$$

A winsorizált varianciákat korrigálni kell ahhoz, hogy reálisan becsüljék a populáció varianciáját. Ha a trimmelés 20%, akkor ez a szorzótényező 0.412 a varianciára és 0.642 a szórásra. Így a Cohen-féle d a következő formában írható fel a trimmelt átlaggal és winszorizált szórással:

$$d_t = 0.642\frac{\bar{x}_{t1} - \bar{x}_{t2}}{s_w}$$

A hatásnagyságot a **yuen.effect.ci()** paranccsal is elvégezhetjük, a parancs előnye, hogy a d_t-re a konfidenciaintervallumot is megadja. Ha egyenlők a varianciák, akkor az **akp.effect()** parancsot használjuk.

```
library(WRS2)
yuen.effect.ci(d$Sr~d$Source,tr=0.2)
#nem egyenlő varianciákra
$effsize
[1] 0.6483503
$alpha
[1] 0.05
$CI
[1] 0.000 0.960
```

A hatásnagyság tehát 0.65, viszont a konfidenciaintervalluma nagyon tág (0, 0.96), a 0.05 szignifikanciaszint mellett belefér a jelentéktelen és a nagy hatás-

nagyság is. Ez az eredmény azt sugallja, hogy a mintaszám alacsony ahhoz, hogy a próba jól működjön. Ha figyelembe vesszük azt, hogy a Yuen–Welch-próba eltávolítja az adatoknak egy jelentős részét, akkor ebben az esetben valóban nagyon leesik a tényleges elemszám, és a próba megbízhatatlanná válik.

5.3.3. Páros t-próba

Egymástól függő adatsoroknak azokat tekintjük, amelyeknek az értékeit ugyanaz a mintaterület, ugyanaz a műszer vagy ugyanazok az egyedek szolgáltatják, és ún. önkontrollos kísérletet végzünk. Az ilyen jellegű kutatások leggyakoribbak az orvostudomány, pszichológia, ill. a gyógyszerkísérletek területén, de előfordulnak a természettudományokban is. Természetükből adódóan az egymástól függő adatsoroknak kisebb a szórása, mint az egymástól független adatsoroké.

A páros t-próba

Egymástól függő adatsoroknál a *t*-próba a különbségek eloszlását vizsgálja, így a nullhipotézis azt jelenti ki, hogy az átlagok közti különbség nulla, a varianciák egyenlőségének a tesztelése pedig értelmét veszíti. Bizonyos feltételeknek teljesülniük kell ahhoz, hogy megbízható eredményt kapjunk. Ezek a feltételek a következők:

– egyszerű véletlen mintavétel;
– egymástól függő minták;
– a két sorozat különbsége normális eloszlású;
– a populáció varianciája nem ismert, a minták varianciáival becsüljük meg;
– ha a minta elemszáma 30-nál nagyobb, akkor a különbség normális eloszlása nem szükséges feltétel (a centrális határeloszlás tétele szerint).

A nullhipotézis kijelenti, hogy a két minta átlagának különbsége nulla (egy populációból származnak). Az ellenhipotézis lehetséges esetei: a mintaátlagok nem egyenlők, az egyik kisebb vagy nagyobb a másiknál.

A próbastatisztika (*t*-érték) a következő kifejezéssel értelmezhető a *t*-próbánál:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s_d}{\sqrt{n}}},$$

ahol $n = n_1 = n_2$, a df $n-1$, s_d pedig az 5.3.1. alfejezetben értelmezett szórás. Az α ismeretében azt vizsgáljuk meg, hogy a számított *t*-érték az elfogadási vagy a kritikus tartományba esik a Student-féle eloszlásfüggvényen, amelynek $df = n-1$ szabadságfoka van. A számolás időigényes, ezért az R-ben végezzük el a próbát (`t.test(x,y,paired=T)`), ahol az eredmény az a *p*-érték, ami a számított *t*-értéknek felel meg.

5.4. Nem parametrikus próbák két medián összehasonlítására

5.4.1. Mann–Whitney–Wilcoxon kétmintás próba

A független adatsorok esetén a próba azt vizsgálja, hogy azonos-e annak az esélye, hogy az egyik, illetve a másik minta értékei lesznek nagyobbak. Ha X és Y mintát tekintjük, akkor a nullhipotézis kimondja, hogy $P(X < Y) = P(X > Y)$, ami egyenértékű azzal, hogy $P(X > Y) = 0.5$.

Az egymintás Wilcoxon-próbának feltétele az adatsor szimmetrikus eloszlása. Ez a feltétel a két minta különbségére kell teljesüljön. Matematikailag bizonyított, hogy két egyforma vagy két szimmetrikus eloszlású adatsor különbségének az eloszlása szimmetrikus. Tehát ebben az esetben elégséges, ha a két minta eloszlása egyforma, vagy mind a kettő szimmetrikus. További alkalmazhatósági feltétel a minták egymástól való függetlensége.

A próbastatisztika (U) azoknak a pároknak a száma, amelyekre $x_i > y_i$. Ha a két mintában előfordul ugyanaz az érték, akkor az adott érték 0.5-ös rangot kap. Ha $n, m \geq 8$, és nincsenek kapcsolt értékek, akkor a próbastatisztika eloszlása megfelel az alábbi normális eloszlásnak:

$$\mu = \frac{n \cdot m}{2} \text{ és } \sigma = \sqrt{\frac{nm(n+m+1)}{12}},$$

ahol n és m a két minta elemszámai.

Az 5.8. táblázat tartalmazza az egészséges fák számát két gyümölcsösben, fajok szerint. Igaz-e, hogy az A gyümölcsösben az egészséges fák mediánja nagyobb a B gyümölcsös mediánjánál? Növekvő sorrendbe rendezve az adatokat: $A = \{3,4,5,5,9\}$, $B = \{3,3,4,6\}$. A minták elemszáma: 5 és 4. A normális eloszlás paraméterei: $\mu = 4x5/2 = 10$, $\sigma = \sqrt{\frac{20 \cdot 10}{12}} = 4$. *Az U* számításánál sorra vesszük az A értékeit, és megállapítjuk, hány értéknél nagyobb, mint B-ben. $U = 1 + 2.5 + 3 + 3 + 4 = 13.5$. Kiszámoljuk az U-ra a z-értéket: $z = (U-\mu)/\sigma = (13.5-10)/4 = 0.875$. A számított z-érték kisebb, mint az $\alpha = 0.025$ ($p = 0.975$) értéknek megfelelő z-érték $= 1.96$, tehát megtartjuk a H_0 hipotézist, miszerint a két gyümölcsösben egyenlő számban fordulnak elő az egészséges fák.

5.8. táblázat. *Az egészséges gyümölcsfák száma két gyümölcsösben*

Gyümölcsös	Körtefa	Szilvafa	Meggyfa	Almafa	Cseresznyefa
A	5	3	5	9	4
B	3	3	4	6	–

A kétmintás Mann–Whitney–Wilcoxon-próbát az R-ben a **wilcox.test()**, illetve a **wilcox.exact()** paranccsal végezhetjük el (lásd az 5.5. táblázatot).

```
A = c(3,4,5,5,9)
B = c(3,3,4,6)
wilcox.test(A,B)
Wilcoxon rank sum test with continuity correction

data: A and B
W = 13.5, p-value = 0.4509
alternative hypothesis: true location shift is not equal to 0
Warning message:
In wilcox.test.default(A,B):
cannot compute exact p-value with ties
```

Jelen vannak a kapcsolt adatok, ezért az egzakt próbát végezzük el.

```
library(exactRankTests)
wilcox.exact(A,B)
Exact Wilcoxon rank sum test
data: A and B
W = 13.5, p-value = 0.4444
alternative hypothesis: true mu is not equal to 0
```

Az egzakt Mann–Whitney–Wilcoxon-próba nem adott szignifikáns eredményt, ezért azt mondhatjuk, hogy a két gyümölcsösben levő egészséges fák száma között nincs szignifikáns különbség.

Függő adatsoroknál a próba a különbségekkel dolgozik, a *V* próbastatisztika pedig a negatív vagy pozitív értékekhez tartozó rangok összege, attól függően, hogy a kettő közül melyik összeg a kisebb. Tekintsük a gyümölcsfák egészségi állapotát kizárólag az *A* gyümölcsösben. Tegyük fel, hogy a gazda kezeli a fákat, aminek a hatására a következő évben megnőtt az egészséges fák száma. A kérdés az, hogy a számbeli növekedés a kezelésnek tulajdonítható-e, vagy csupán a véletlen eredménye. A rangokat hasonló módon adjuk a különbségekhez, mint az egymintás Mann–Whitney–Wilcoxon-próbánál. Az 5.9. táblázat alapján a $V^- = 3.5$, a $V^+ = 11.5$, tehát a próbastatisztika értéke 3.5.

A csatolt értékek miatt az egzakt **wilcox.exact()** próbát végezzük el az {exactRankTest} csomagból (5.5. táblázat).

```
A1 = c(5,3,5,9,4)
A2 = c(9,2,8,7,6)
library(exactRankTests)
wilcox.exact(C,D, paired = T, alternative = "less")
```

```
#C kisebb mint D?
Exact Wilcoxon signed rank test
data: A1 and A2
V = 3.5, p-value = 0.1875
alternative hypothesis: true mu is less than 0
```

5.9. táblázat. *Az egészséges gyümölcsfák száma két gyümölcsösben*

Gyümölcsös	Körtefa	Szilvafa	Meggyfa	Almafa	Cseresznyefa
A (1. év)	5	3	5	9	4
A (2. év)	9	2	8	7	6
2. év – 1. év	4	−1	3	−2	2
Rangok	5	1	4	2.5	2.5
Előjelek	+	-	+	-	+

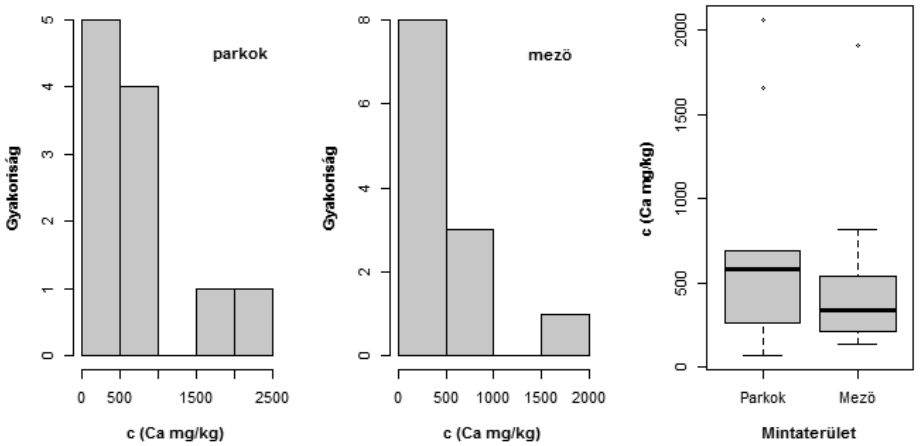
A próba eredménye szerint ($p = 0.1875 > 0.05$) megtartjuk a H_0 hipotézist, az egészséges fák számának növekedése véletlenszerűnek tekinthető.

A csiperkegombák kalciumtartalma a két csoportban nem normális eloszlású, viszont egyformán jobb oldali ferdeséget mutatnak (5.5. ábra), ezért a Mann–Whitney–Wilcoxon-próbát használjuk a mediánok összehasonlítására. A próba nem adott szignifikáns eredményt ($W = 79$, $p = 0.449$).

```
wilcox.test(d$Ca~d$Source)
Wilcoxon rank sum exact test

data: d$Ca by d$Source
W = 79, p-value = 0.4491
alternative hypothesis: true location shift is not equal to 0
```

Az 5.5. ábrát az alábbi kóddal lehet elkészíteni R-ben:

```
par(mfrow=c(1,3), font.lab = 2)
hist(d$Ca[d$Source == "city park"], xlab = "c (Ca mg/kg)",
 ylab = "Gyakoriság", main ="", font.lab = 2)
text(2100,4.5, "parkok")
hist(d$Ca[d$Source == "meadow"], xlab = "c (Ca mg/kg)",
 ylab = "Gyakoriság", main = "")
text(1600,7.2, "mezö")
boxplot(d$Ca~d$Source,
    xlab = "Mintaterület",
    names = c("Parkok", "Mezö"),
   ylab = "c (Ca mg/kg)")
```



5.5. ábra. *A csiperke kalciumtartalmának eloszlása az élőhely szerint*

A mediánokra a hatásnagyságot a *WRS* csomagban található **med.effect()** paranccsal hajthatjuk végre. A két csoportot külön-külön kell bevezetni, ezért készítünk két vektort. Ha a WRS csomagot nem lehet telepíteni az *RStudió*ból, akkor az alábbi módon kell megpróbálni:

```
#A WRS csomag telepítése
#Kiegészítő csomagok telepítése
install.packages(c("MASS", "akima", "robustbase"))
install.packages("WRS",repos="http://R-Forge.R-project.org")
#Az élőhelyek szerinti vektorok
city = subset(d, Source == "city park")
meadow = subset(d, Source == "meadow")
#A hatásnagyság számolása mediánokra
library(WRS)
med.effect(city,meadow)
[1] 0.4251077
```

Annak ellenére, hogy a hatásnagyság nem bizonyult jelentéktelennek, a Mann–Whitney–Wilcoxon-próba *p*-értéke elég nagy ahhoz, hogy elfogadjuk azt, hogy nincs különbség a két területről származó csiperke kalciumtartalmában.

5.4.2. Kétmintás előjelpróba függő változókra

A próba algoritmusa egyezik az egymintás előjelpróbáéval azzal a különbséggel, hogy ebben az esetben a két helyzetre kapott értékek páronkénti különbségeire végezzük el. Az 5.9. táblázatban három pozitív és két negatív különbséget kaptunk, tehát $N^+ = 3$, $N^- = 2$. Összesen $n = 5$ elemszámú mintapárunk van, egy sikeres esemény (pl. pozitív különbség) valószínűsége pedig 0.5. Elvégezzük az előjelpróbát a **`binom.test()`** paranccsal, amelynek részletes leírása a 5.3. táblázatban látható. Általában a kisebb N értéknek megfelelő eseményt tekintjük sikernek.

```
binom.test(2,5)
Exact binomial test
data: 2 and 5
number of successes = 2, number of trials=5, p-value=1
alternative hypothesis:
true probability of success is not equal to 0.5
95 percent confidence interval:
0.05274 0.85337
sample estimates:
probability of success
 0.4
```

A két év alapján a próba nem adott szignifikáns eltérést a fák egészségi állapotával kapcsolatban.

5.4.3. Kétmintás permutációs próba függő és független adatsorokra

A permutációs próba alapelve az, hogy ha két adatsor azonos populációból származik, akkor az értékek véletlenszerűen kerültek a két mintába. Ez alapján a próba nem igényel sem mintaátlagokat, sem varianciákat, és az sem feltétel, hogy a mintavétel véletlenszerű legyen (Bonini et al., 2014). A próba a két mintát egyként kezeli, és az adatokból véletlenszerűen hoz létre két mintát az n és m eredeti mintaelemszámmal, majd kiszámolja a T próbastatisztikát, ami a két átlag különbsége. Ezt a műveletet sokszor megismételi (általában 5000, 10000 alkalommal), és a T értékek eloszlásától elvárja, hogy ne tartalmazzon sem túl kis, sem túl nagy értékeket. Egy permutáció alapján kapott T érték:

$$T^* = \bar{X}_1^* - \bar{X}_2^*.$$

A permutációs próba az R-ben a **`perm.test()`** paranccsal végezhető el. A parancs alkalmas egy minta, illetve az egymástól függő adatsorok permutációs próbájára is (5.10. táblázat).

5.10. táblázat. *A permutációs próba értelmezése R-ben*

Parancs	Részletek	Magyarázat
perm.test(x, y, mu = 0, paired = F, alternative, conf.level) {exactRankTests}		
x, y	adatsorok	két oszlop egy adattáblából vagy két vektor
mu	feltételezett középérték	egymintás próbánál
paired	függő vagy független adatsorok	alapból független adatsorokra működik (= F)
alternative	"two.sided", "less", "greater"	alapból kétoldali próba
conf.level	0.95, 0.90, 0.99 stb.	az (*1-α*)-érték, alapból 0.95

Elvégezzük a kétmintás permutációs próbát az 5.8. táblázatban feltüntetett két adatsorra.

```
library(exactRankTests)
perm.test(A,B)
2-sample Permutation Test
data: A and B
T = 26, p-value = 0.4762
alternative hypothesis: true mu is not equal to 0
```

A Mann–Whitney–Wilcoxon-próbához hasonlóan a permutációs próba sem adott szignifikáns eltérést a két minta középértékére. Hasonló a helyzet a csiperkegombák medián kalciumtartalmával is. A kétmintás permutációs próba *p*-értéke (0.436) hasonló a Mann–Whitney–Wilcoxon-próba *p*-értékéhez (0.449).

```
perm.test(d$Ca~d$Source)
2-sample Permutation Test
(scores mapped into 1:(m+n) using rounded scores)

data: d$Ca by d$Source
T = 78, p-value = 0.436
alternative hypothesis: true mu is not equal to 0
```

5.5. Alkalmazási szempontok a statisztikai próbákra

A kétmintás próbák alkalmazhatóságánál három szempontot kell figyelembe venni: egyenlők-e a varianciák a két adatsorban; az eloszlások normálisak-e vagy ferdék a kiugró értékek miatt; illetve a minták elemszámai egyenlők-e. A helyzetet

bonyolítja az *F*-próba, mivel a nem szignifikáns eredmény csak azt jelenti, hogy nincs elég bizonyítékunk arra, hogy elvessük a nullhipotézist. Tehát a nullhipotézis megtartásával annyit jelenthetünk ki, hogy a varianciák nem különböznek szignifikánsan, de azt nem állíthatjuk, hogy egyenlők. A *t*-próba bizonyos mértékben robusztus a varianciák különbözőségére, de csak abban az esetben, ha a minták elemszáma közt nincs nagy eltérés. Ezt még súlyosbítja az *F*-próbának a kismértékű megbízhatósága kis mintaelemszámoknál.

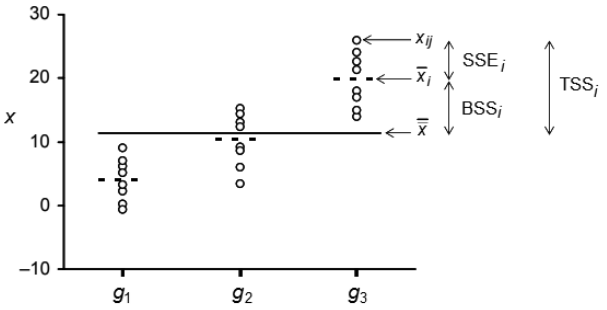
Általánosan elfogadott, hogy nagy mintaelemszámoknál ($n > 30$) a *t*-próba hatékonyan működik abban az esetben is, ha az eloszlások nem követik szigorúan a normális eloszlást, a varianciák kismértékben eltérnek egymástól (arányuk nem nagyobb 2-nél), a minták elemszámai pedig közeliek (Fagerland, 2012). Környezettudományi kutatásokban a minta elemszáma a legtöbb esetben kisebb 30-nál, a *t*-próba megbízhatósága pedig erősen függ a fent említett alkalmazhatósági feltételektől. Kis elemszámoknál és normális eloszlású mintáknál a legkisebb kockázattal járó helytelen döntést a Welch-féle *U*-próba biztosítja (Fagerland és Sandvik, 2009; Ruxton, 2006).

Kiugró értékeket tartalmazó normális eloszlású mintákra a Yuen–Welch-próbát érdemes használni. Egyoldali próbánál, különösen, ha a minták elemszámai különböznek, célszerűbb a Yuen–Welch-próba bootstrap-*t* változatát használni, ami szimulációt futtat a meglévő adatokkal úgy, hogy a két adatsort úgy csúsztatja el, hogy a trimmelt átlagok egybeessenek. Ezt követően ismételten mintákat vesz az összeolvasztott adatokból, amelyek alapján megbecsüli a *t*-eloszlás alsó és felső kritikus értékét, α szignifikanciaszint mellett (Westfall és Young, 1993). Ez a próba az R-ben a WRS2 csomag **`yuenbt(formula, data, tr = 0.2, nboot = 599, side = TRUE)`** parancsával végezhető el. A parancs alapból 599 ismételt mintavételt végez el.

Az egymintás Wilcoxon-féle előjeles rangpróba sokkal erősebb, mint az előjelpróba, viszont csak akkor megbízható, ha az adatsor eloszlása szimmetrikus. Ezt a feltételt a hisztogramon ellenőrizhetjük. Amennyiben az eloszlás ferde, javasolt az előjelpróba használata.

5.6. Varianciaanalízis. Kettőnél több mintaátlag egyenlőségének tesztelése. Parametrikus próbák

A varianciaanalízis egy vagy több faktor hatását vizsgálja egy folytonos függő változóra azáltal, hogy a függő változó szóródásának mekkora részét magyarázza a faktor (faktorok). A faktor által meghatározott csoportok közti eltéréseket maga a faktor okozhatja, de olyan tényezők is okozhatják, amelyekről nem tudunk semmit. Az utóbbiaknak tulajdonítható szóródás a zajt vagy a hibatagot eredményezi (5.6. ábra).



5.6. ábra. *A faktor által megmagyarázott (BSS_i) szóródás és a hibatag (SSE_i) viszonya egymáshoz egy csoproton belül. A TSS_i a teljes szóródás egy csoportra vonatkoztatva.*

Egy faktor által meghatározott több minta középértékének összehasonlítására az egyszempontos ANOVA-t használjuk. A név az ANalysis Of VAriances kifejezésből származik, ami varianciaanalízist jelent. Nevét onnan kapta, hogy a mintaátlagok varianciáját viszonyítjuk az összevont adatsorok varianciájához. Első megközelítésből egyértelműnek látszik a minták páronkénti összehasonlítása *t*-próbával, viszont ez az út azért nem járható, mert a számos *t*-próba külön-külön α szinten tartott elsőfajú hibája összeadódik, és így megnő annak az esélye, hogy az egyetlen kérdésre (miszerint van-e különbség a csoportátlagok között) adott válasz elsőfajú hibája túlságosan nagy lesz.

Az ANOVA parametrikus próba, ami megköveteli az adatok normális eloszlását minden mintára. Amennyiben ez a feltétel teljesül, a következő változatai közül választhatunk:

– egyszempontos ANOVA (egyetlen szempont szerint csoportosítjuk a teljes adatsort);
– kétszempontos vagy faktoriális ANOVA (két szempont szerint csoportosítjuk a teljes adatsort).

Az egyszempontos ANOVA hasonlít a *t*-próbára, például több változata van:

– ANOVA, ha a varianciák egyenlősége teljesül;
– Welch–ANOVA, ha a varianciák nem egyenlők;
– ismételt méréses ANOVA, ha a minták nem függetlenek egymástól.

5.6.1. Egyszempontos ANOVA

Az egyszempontos ANOVA azt vizsgálja, hogy kettőnél több, egymástól független minta átlagai egyenlők-e egymással. Ez a próba nullhipotézise. Az ellenhipotézis azt állítja, hogy legalább két minta átlaga különbözik egymástól, azonban ezek a minták nem azonosíthatók be. A próba akkor alkalmazható, ha teljesülnek az alábbi feltételek:

– a mintákat az egyszerű véletlen mintavétellel képezték;
– minden minta normális eloszlású;
– a varianciák egyenlők;
– a minták egymástól függetlenek.

A próba két variancia (mintaátlagok varianciája és csoportokon belüli variancia) arányának a vizsgálatán alapuló *F*-próba, amelynek két szabadságfoka van: $df_1 = g - 1$, $df_2 = N - g$, ahol g a csoportok száma, N az adatok össz-száma.

$$F = \frac{\text{csoprtok közti variancia}}{\text{csoportokon belüli variancia}} = \frac{s_g^2}{s_N^2},$$

ahol:

$$s_g^2 = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_g\left(\bar{x}_g - \bar{x}\right)^2}{g - 1},$$

$$s_N^2 = \frac{\sum(\bar{x}_1 - \bar{x})^2 + \sum(\bar{x}_2 - \bar{x})^2 + \cdots + \sum\left(\bar{x}_g - \bar{x}\right)^2}{N - g},$$

az $\bar{x}$ az átlagok átlaga, az $\bar{x}_1, \bar{x}_2, \bar{x}_g$ pedig a minták átlagai. Az F-értéket az F-eloszlás alapján értékeljük ki. Az R az ANOVA eredményét az 5.11. táblázat formájában adja meg, az angol jelölésekkel. A BSS, WSS és TSS négyzetösszegek (between, within és total sum of squares) az s_g^2 és s_N^2 varianciák képletében szereplő számlálókat jelentik. Az átlagos csoportok közötti variancia (MBSS – mean between groups sum of squares) az átlagok különbözőségét jellemzi, ezért ezt hatásvarianciának nevezzük. Az átlagos csoportokon belüli varianciát (MWSS – mean within groups sum of squares) hibavarianciának nevezzük, mert a csoportokon belüli szóródást a véletlen ingadozások adják (amennyiben a mintákra a véletlen mintavétel módszerét alkalmaztuk). R-ben az egyszempontos ANOVA-t az alapcsomagban levő *aov(függő változó~faktor)* paranccsal végezhetjük el.

Amennyiben az ANOVA a szignifikanciaszintnél kisebb *p*-értéket ad, elvetjük a H_0 hipotézist. Ilyen esetben egy post-hoc próbát használunk a mintaátlagok páronkénti összehasonlítására úgy, hogy a műveletek össz-szignifikanciaszintje α legyen.

5.11. táblázat. *Az egyszempontos ANOVA eredményének megadása*

A szóródás forrása	Szabadságfokok	Négyzetösszegek	Variancia	*F*-test
Csoportok között	$g-1$	*BSS*	$M_{BSS}\,(s_g^2)$	M_{BSS}/M_{WSS}
Csoportokon belül	N−g	WSS (SSE[#])	$M_{WSS}\,(s_N^2)$	
Teljes variancia[*]	$N-1$	*TSS*	$M_{TSS}(s^2)$	

[#] *SSE – sum of square error*

[*] *TSS = BSS + WSS*

A feltételek tesztelése

Az egyszempontos ANOVA feltételeinek tesztelését különböző próbákkal végezzük el. A csoportok normális eloszlását a Shapiro–Wilk- vagy Kolgomorov–Smirnov-próbával végezhetjük el, vagy a grafikus ellenőrzésre is hagyatkozhatunk (hisztogram vagy Q-Q ábra).

A varianciák egyenlőségét többféle próbával ellenőrizhetjük, a próba kiválasztását a minták eloszlása dönti el: a Bartlett-próba normális eloszlásoknál ad megbízható eredményt, a Levene-próba kevésbé érzékeny a normális eloszlástól való eltérésekre, a Brown–Forsythe-próba pedig a legalkalmasabb erősen ferde eloszlásokra. A szórások inhomogenitását, illetve az eloszlások ferdeségét ki lehet küszöbölni az adatok átalakításával (pl. logaritmálás). Másik lehetőség az ANOVA helyettesítése a Welch–ANOVA vagy a Brown–Forsythe-féle *F*-próbába (Allen, 2017).

Az ANOVA alkalmazhatóságának egy másik feltétele az, hogy a reziduálisok (az egyedi eltérések a csoportátlagtól) normális eloszlást kövessenek. Ezt ellenőrizhetjük egy Q-Q ábrán vagy hisztogramon, illetve elvégezhetünk egy próbát a normális eloszlásra. R-ben a reziduálisok megtalálhatók vektorként az *aov* típusú objektumban *residuals* név alatt.

Post-hoc tesztek

Az ANOVA-próbára a leggyakrabban használt post-hoc tesztek a Tukey, Duncan, Scheffe, Student-Newman-Keuls (SNK), illetve a legkisebb szignifikáns különbség (LSD – least significant difference) (Allen, 2017). Ezek közül a helyzetnek megfelelően a legalkalmasabbat válasszuk ki. A Scheffe-próba a legszigorúbb az elsőfajú hiba elkövetésével szemben. Ez egyben azt is jelenti, hogy a legnagyobb a veszélye annak, hogy egy másodfajú hibát követünk el, nem mutatjuk ki a hatást, amikor az a valóságban fennáll. A Tukey, Duncan, SNK, LSD sokkal engedékenyebbek az elsőfajú hiba elkövetésével szemben, ezért ezeket a teszteredményeket óvatosan kell kezelni. A Scheffe-teszt az egyedüli, amely megengedi a különböző elemszámot a mintákban, a Duncan-teszt pedig akkor is elvégezhető, ha az ANOVA-teszt nem adott szignifikáns eredményt. A Scheffe-próba további különlegessége, hogy kevésbé érzékeny arra, ha nem teljesül a normális eloszlás vagy a varianciák egyenlősége, és más algoritmusra épül, mint a többi próba: nem páronkénti összehasonlításokat végez, hanem az átlagok közti lineáris kontrasztokat értékeli ki (Gad, 2010). R-ben a post-hoc próbákat többféle csomag segítségével lehet elvégezni, az 5.12. táblázatban a {DescTools} csomag parancsa látható.

Hatásnagyság az ANOVA-próbára

Az egyszempontos ANOVA hatásnagyságát (η^2, eta-négyzet) a csoportok közötti variancia és a teljes variancia aránya fejezi ki, ami megmutatja, hogy a csoportok közötti szóródás hányad része a teljes szóródásnak:

$$\eta^2 = \frac{s_g^2}{s^2}.$$

Az η^2 mértékét Cohen-féle kritikus értékek négyzetével értelmezhetjük, azaz a 0.10, 0.25, 0.40 megfelelnek a kis, közepes és nagy hatásnak (Allen, 2017). A Cohen-féle d kiterjeszthető kettőnél több csoportra is, ebben az esetben a Cohen-féle *f* mutatót számoljuk ki:

$$f = \frac{s_m}{s},$$

ahol s_m a csoportátlagok szórása, s pedig a teljes adatsor szórása. Az s_m kifejezése függ attól, hogy a csoport elemszámai egyenlők-e vagy nem.

$$s_m = \sqrt{\frac{(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2 + \cdots + (\bar{x}_g - \bar{x})^2}{g}},$$

ha a csoportok elemszámai azonosak, és

$$s_m = \sqrt{\frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_g(\bar{x}_g - \bar{x})^2}{N}},$$

ha a csoportok elemszámai nem azonosak. Mivel a $d = 2f$, ezért a 0.1, 0.25, ill. 0.5 *f*-értékek megfelelnek a kis, közepes és nagy hatásnak (Salkind, 2010).

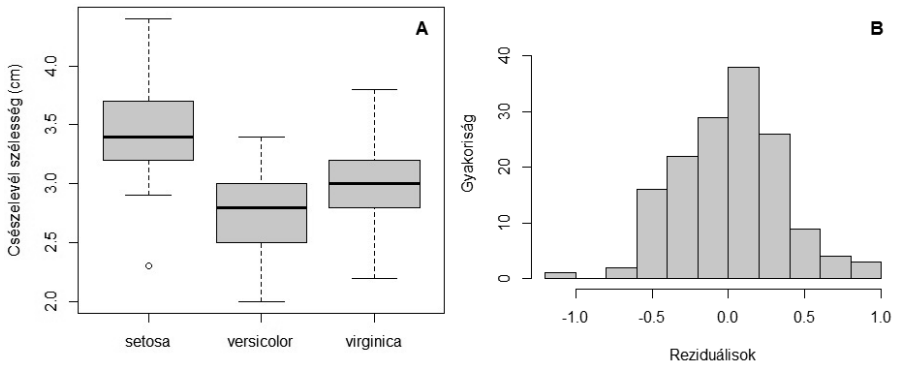
5.12. táblázat. *Az egyszempontos ANOVA-próba R-ben*

Parancs	Részletek	Magyarázat
model = aov(x~g, data) PostHocTest(model, method, conf.level = 0.95, ordered) {DescTools} eta_squared(model, partial = F) {effectsize}		
x	Numeric	Függő, numerikus változó
g	Factor	Független változó
model	egy aov típusú objektum	Az ANOVA eredményét örző objektum.
method	"hsd", "bonferroni", "lsd", "scheffe", "newmankeuls", "duncan"	Tukey HSD, Bonferroni/Dunn, LSD, Scheffe, SNK, Duncan
ordered	FALSE / TRUE	TRUE = legkisebb átlagtól halad a legnagyobb felé
partial	TRUE/FALSE	Egyszempontú ANOVA-nál a parciális és a teljes eta-négyzet ugyanaz az érték. Alapból a TRUE van beállítva.

Példa: Három nősziromfaj morfológiai adatai

Az R alapcsomagjában levő *iris* adattábla három nősziromfaj csésze- és sziromlevelének a hosszát és szélességét tartalmazza, fajonként 50-50 egyedre. Megvizsgáljuk a csészelevelek fajonkénti átlagszélességének egyenlőségét (5.7. ábra,

A). Az ANOVA alkalmazásának feltételei a normális eloszlás és a varianciák egyenlősége. Mivel minden csoportban 50 egyedet vizsgáltak, a Shapiro–Wilk-próbával ellenőrizzük az eloszlást, és a Levene-próbával a varianciák egyenlőségét.



5.7. ábra. *A három nősziromfaj csészelevél-szélességének dobozdiagramja (A) és az ANOVA-próba reziduálisainak eloszlása (B)*

Hipotézisek:

H_0: a három nősziromfaj csészelevél-átlagszélességei egyenlők. ($p > 0.05$)

H_1: a csészelevelek átlagszélességei legalább két fajnál különböznek. ($p \leq 0.05$)

Az adattábla három csoportja szerinti dobozdiagram megközelítően szimmetrikus eloszlásokat mutat, az *Iris setosa* fajnál enyhe jobb oldali ferdeség áll fenn.

```
#Normális eloszlások tesztelése Shapiro–Wilk-próbával
tapply(iris$Sepal.Width, iris$Species, shapiro.test)
$setosa
Shapiro-Wilk normality test
data: X[[i]]
W = 0.97172, p-value = 0.2715
$versicolor
Shapiro-Wilk normality test
data: X[[i]]
W = 0.97413, p-value = 0.338
$virginica
Shapiro-Wilk normality test
data: X[[i]]
W = 0.96739, p-value = 0.1809

#Varianciák egyenlőségének tesztelése Levene-próbával
library(DescTools)
```

```
LeveneTest(iris$Sepal.Width~iris$Species)
Levene's Test for Homogeneity of Variance
(center = median)
 Df F value Pr(>F)
group 2 0.5902 0.5555
 147
#A p-érték 0.555 > 0.05, a varianciák egyenlők.

#Az ANOVA-próba
modell = aov(iris$Sepal.Width~iris$Species)
summary(modell)
 Df Sum Sq Mean Sq F value Pr(>F)
iris$Species 2 11.35 5.672 49.16 <2e-16 ***
Residuals 147 16.96 0.115
#A p-értél << 0.05, legalább két átlag nem egyenlő.

#Post-hoc teszt
posth = PostHocTest(modell,
method = "hsd", ordered = T)
print(posth, digits = 2)
 Posthoc multiple comparisons of means : Tukey HSD
 95% family-wise confidence level
 factor levels have been ordered
$`iris$Species`
 diff lwr.ci upr.ci pval
virginica-versicolor 0.20 0.043 0.36 0.0088 **
setosa-versicolor 0.66 0.497 0.82 3.1e-14 ***
setosa-virginica 0.45 0.293 0.61 1.4e-09 ***
#A p-értékek (utolsó oszlop) << 0.05

#Reziduálisok eloszlásának tesztelése
hist(modell$residuals) #6.7.B ábra
shapiro.test(modell$residuals)
Shapiro-Wilk normality test
data: iris_a$residuals
W = 0.98948, p-value = 0.323
#A p-érték > 0.05, a reziduálisok normális eloszlásúak.
```

Az ANOVA-próba szignifikáns eredményt adott ($F = 49.16$, $df_1 = 2$, $df_2 = 147$, $p < 0.001$). A Tukey HSD post-hoc próba szerint mind a három faj páronként különböző átlagszélességű csészelevéllel rendelkezik. A hisztogram és a Shapiro–Wilk-próba szerint a reziduálisok eloszlása normális (4.9. ábra, B).

A hatásnagyság megadható az *f* mutatóval. R-ben az {effectsize} csomagban megtalálhatók a megfelelő parancsok. A számítások alapján azt mondhatjuk, hogy a hatás nagy, ami alátámasztja az ANOVA szignifikáns eredményét.

```
library(effectsize)
eta_squared(modell, partial = F)
#Effect Size for ANOVA (Type I)

 Parameter | Eta2 | 95% CI
---------------------------------
iris$Species | 0.40 | [0.30, 1.00]

cohens_f(modell, partial = F)
#Effect Size for ANOVA (Type I)

Parameter | Cohen's f | 95% CI
-----------------------------------------
iris$Species | 0.82 | [0.66, Inf]
```

5.6.2. Egyszempontos ANOVA nem egyenlő varianciákra

Az egymástól független minták átlagainak egyenlőségére, ha a varianciák nem egyenlők a csoportváltozó szintjeire, robusztus varianciaelemzést használhatjuk. A leggyakrabban használt próbák: a Welch–ANOVA, a Brown–Forsythe-próba és a James-próba (Brown és Forsythe, 1974; James, 1951; Welch, 1951).

A Welch–ANOVA a Welch-féle *U*-próba általánosítása. Akkor működik jól, ha az eloszlások normálisak, a mintaelemszámok nagyok és egyenlőek. Az elsőfajú hiba felügyelése viszont kevésbé hatékony, ha az elemszámok kicsik és eltérőek, amihez bizonyos fokú ferdeség is társul a minták eloszlásában. A próba megbízhatóságát tovább gyengíti a csoportok nagy száma (Lix és Keselman, 1995; Wilcox, 1988).

A Brown–Forsythe-próba az egyszempontos ANOVA robusztus változata, amely jól ellenőrzi az elsőfajú hibát akkor is, ha a csoportokon belüli eloszlások különböző mértékben ferdék. Ellenben kevésbé jól teljesít, ha a csoportok varianciái erősen eltérnek egymástól, és az elemszám alacsony (Lix et al., 1996). Normális eloszlású mintáknál a Welch–ANOVA megbízhatóbban működik.

A James-próba egy Q próbastatisztika kiértékelésén alapul, amely χ^2-eloszlást követ. A próba sokkal hatékonyabban működik a másik kettőnél, ha az eloszlások szimmetrikusak, ellenben a kis elemszám és ferde eloszlás esetén nem tanácsos a használata (Lix et al., 1996; Wilcox, 1988).

Post-hoc próba és hatásnagyság

A nem egyenlő varianciákkal rendelkező minták páronkénti összehasonlítására a Games–Howell-próba alkalmas. A próba a szabadságfokok Welch-féle korrekciója és a Tukey-próba Student-féle változatára épül. Az átlagok közti eltérések kiértékelésekor korrigálja az ismételten elvégzett műveletek *p*-értékét, ezért ezt nem kell külön elvégezni. Az elsőfajú hibát akkor is jól kontrollálja, ha a minták elemszámai különböznek, ellenben kis mintaelemszámoknál engedékenyebbé válik (a javaslat az, hogy a minták elemszámai ne legyenek hatnál kisebbek). A hatásnagyságot az egyszempontú ANOVA-nál tárgyalt paraméterekkel lehet megadni.

A próbákat R-ben a {onewaytests} csomag segítségével, a post-hoc próbát pedig az {rstatix} csomagban levő **`games_howell_test()`** paranccsal végezhetjük el (5.13. táblázat). A Welch–ANOVA próba lehetővé teszi az adatsorok trimmelését a *rate* opció által, amelynek értéke [0, 0.49] között változhat.

5.13. táblázat. *A robusztus egyszempontos ANOVA-próba R-ben*

Parancs	**Részletek**	**Magyarázat**
model = welch.test(x~g, data, rate, alpha) {onewaytests} model = bf.test(x~g, data, alpha) {onewaytests} model = james.test(x~g, data, alpha) {onewaytests} games_howell_test(x~g, data, conf.level = 0.95) {rstatix} eta_squared(model, partial = F) {effectsize}		
x	Numeric	függő, numerikus változó
g	Factor	független változó
data	Dataframe	a változókat tartalmazó adattábla
rate	trimming	Alapból az értéke 0. Változtatható 0.49-ig. Például a 0.1 érték 10% trimmelést jelent.
alpha	signifikanciaszint	Alapból az értéke 0.05.
partial	TRUE/FALSE	Egyszempontú ANOVA-nál a parciális és a teljes eta-négyzet ugyanaz az érték. Alapból a TRUE van beállítva.

5.6.3. Ismételt méréses ANOVA

A próba olyan mintákra alkalmazható, amelyek nem függetlenek egymástól. A leggyakoribb esetek azok, amikor időben ugyanazon a mintán többször elvégzünk adott méréseket, megfigyeléseket, vagy ugyanazt a mintát különböző feltételek mellett vizsgáljuk. Ezért ezt a próbát csoporton belüli ANOVA-nak (within-subjects ANOVA) is nevezik. A független változó lehet nominális vagy ordinális (csoportok), a függő változó pedig folytonos numerikus (arány- vagy intervallumskálájú). Mivel ugyanazt a mintát használjuk, a kutatás során a minta

elemszáma nem változik. A próba hipotézisei megegyeznek az egyszempontos ANOVA hipotéziseivel:

H_0: a mintára kapott átlagok egyenlők ($p > 0.05$);

H_1: legalább két átlag különbözik egymástól ($p \leq 0.05$).

A próba akkor alkalmazható, ha teljesülnek az alábbi feltételek:

– a mintában levő objektumok véletlen mintavétellel lettek kiválasztva;

– a vizsgált változó minden csoportban normális eloszlású;

– szfericitás fennállása.

Szfericitás

Szfericitás alatt varianciák különbségének az egyenlőségét értjük. Két-két csoportra kiszámoljuk az egyedekre kapott értékek különbségeit, majd a különbségek varianciáját. Ha három csoportot (1, 2, 3) vizsgáltunk, akkor három varianciát hasonlítunk össze, amelyeket az 1-2, 1-3, 2-3 adatsorokra kaptunk. Erre alkalmas próba a Mauchly-féle szfericitási vizsgálat. A Mauchly-próba nullhipotézise a különbségek varianciáinak egyenlősége. A próbastatisztika a Mauchly W-értéke, amely χ^2-eloszlást követ g–1 szabadságfokkal, ahol g a csoportok száma. A próba nem végezhető el két csoport esetén (egyetlen varianciaérték van). A Mauchly-próba ereje gyengül kis elemszámok mellett, illetve nagyon megnő nagy elemszámú mintáknál. Ennek ellenére a szfericitás tesztelésére ez a jelenleg bevált próba.

Amennyiben a próba szignifikáns eredményt ad, korrekciót kell végezni a szabadságfokok számára. A korrekció eredményeként a kapott *p*-érték nagyobb lehet. A korrekciót többféle módon lehet elvégezni, a leggyakoribbak:

– Greenhouse–Geisser-módszer;

– Huynh–Feldt- (kevésbé konzervatív) módszer.

R-ben a Mauchly-próba be van építve az **`anova_test()`** parancsba, ami az {rstatix} csomagban található meg. Szignifikáns eredménynél a parancs az ANOVA-próba szabadságfokait helyben módosítja a Huynh–Feldt- és a Greenhouse–Geisser-módszer szerint.

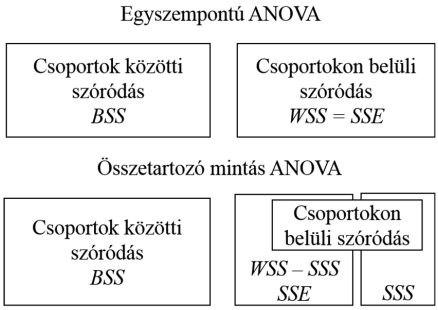
F-próba

Az egyszempontos ANOVA hibatagjának a független csoportokon belüli eltérések négyzetösszege (*WSS*) felelt meg. Az ismételt méréseknél a méréseket ugyanazon az objektumon végezzük, így a kapott szórás kisebb, mint amit különböző objektumok esetében kapunk (5.8. ábra). A hibanégyzetek összege (*SSE*) ebben az esetben a következőképpen módosul:

$$SSE = WSS - SSS,$$

ahol az *SSS* az alanyok által meghatározott variabilitás (Sum of Squares of the Subjects). Mivel a hibanégyzetek összege kisebb, mint a független minták esetében,

ezért az *F*-érték nagyobb lesz, így a próba ereje nagyobb, mint az egyszempontos ANOVA esetében.



5.8. ábra. *A független és az egymástól függő minták által meghatározott hibanégyzetek összege (*SSE*).*

Az *SSS* az alábbi képlettel számolható ki:

$$SSS = g \cdot \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2,$$

ahol $\bar{x}_i$ az alanyok (subject) értékeinek átlaga az *i*-dik csoportban, $\bar{x}$ a nagy átlag (csoportátlagok átlaga), g pedig a csoportok száma, k a csoportok száma.

Az *F*-értéket az M_{BSS} és M_{SSE} arányából kapjuk meg, a két szabadságfok pedig: $df_1 = g - 1$, $df_2 = (N - 1)(g - 1)$, ahol *g* a csoportok száma, *N* az adatok össz-száma.

$$F = \frac{M_{BSS}}{M_{SSE}},$$

ahol M_{BSS} azonos az egyszempontú ANOVA-nál használt mutatóval, az M_{SSE} pedig:

$$M_{SSE} = \frac{SSS}{(N-1)(g-1)}.$$

R-ben a próbát végre lehet hajtani az alapcsomagban levő **`aov()`** parancscsal, beillesztve egy hibatagot is az összefüggésbe: **`aov(függő.v~csoport.v + Error(egyedek vektora))`**. Ennek a parancsnak feltétele, hogy az egyedek vektorában egy adott egyed kódja minden csoportban azonos legyen. A post-hoc teszt a **`pairwise.t.test()`** paranccsal futtatható. Az {rstatix} csomagban található parancsok több információt szolgáltatnak a puszta *F*-próbánál. Például az **`anova_test()`** parancsba be van építve a varianciák homogenitásának a tesztelése Mauchly-módszerrel, és a szabadságfokok korrigálása is, ha nem teljesül a feltétel, valamint az eta-négyzet hatásnagyságot is megadja. A post-hoc teszt a **`pairwise_t_test()`** paranccsal végezhető el (5.14. táblázat).

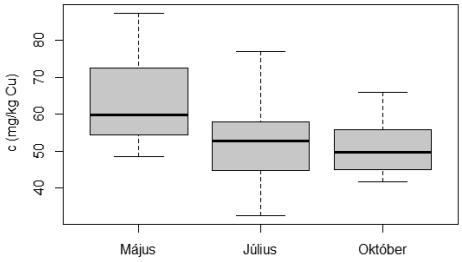
5.14. táblázat. *Az ismételt méréses ANOVA elvégzése R-ben*

Parancs	Részletek	Magyarázat
modell = anova_test(data, dv, wid, within) {rstatix} get_anova_table(modell) {rstatix} pairwise_t_test(x~g, paired = TRUE, p.adjust.method) {rstatix}		
data		adattábla
dv	dependent variable	függő változó
wid	within identities	objektumok kódjait tartalmazó változó (a kódok minden csoportban ugyanazok); át kell alakítani faktor változóra
within	groups	csoportváltozó (faktorrá kell alakítani)
x		függő változó
g	groups	csoportváltozó
p.adjust.method	*p*-érték korrekciója	"holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none"

Példa

Kolozsvár közelében (18 km-re északra), egy legelőnek használt területen vizsgáltuk a mezei csiperke (*Agaricus campestris*) réztartalmának a változását három termési időszakban, 2020-ban. Minden alkalommal 15 mintát gyűjtöttünk be, egy-egy mintában 5-7 egyeddel. Az 5.9. ábrán látható a rézkoncentrációk dobozdiagramja a három időszakra. Az ábra alapján úgy tűnik, hogy az első termési időszakban volt a legmagasabb a rézmennyiség a gombában. Az adattábla tartalmazza a mintakódokat (ID), amelyek ismétlődnek a három időszakban, a rézkoncentrációt (Cu) szárazanyagra mg/kg-ban kifejezve, és az időszakot (május, július és október).

```
head(g)
  ID Cu Season
1 DK1 77.1 Május
2 DK2 71.7 Május
3 DK3 87.5 Május
4 DK4 58.1 Május
5 DK5 52.7 Május
6 DK6 70.0 Május
```



5.9. ábra. *Mezei csiperke réztartalma az idő függvényében*

Tesztelni szeretnénk a rézkoncentráció változását a három egymást követő időszakban. Mivel a minták nem függetlenek egymástól, hiszen három alkalommal ugyanarról a helyről gyűjtöttük be őket, ismételt méréses ANOVA próbával végezzük a tesztelést.

A feltételek tesztelése:

```
#Normális eloszlás tesztelése
library(rstatix)
library(tidyverse)
g %>% group_by(Season) %>% shapiro_test(Cu)
#A tibble: 3 x 4
 Season variable statistic p
 <chr> <chr> <dbl> <dbl>
1 First Cu 0.922 0.205
2 Second Cu 0.950 0.529
3 Third Cu 0.946 0.471
#Mind a három minta normális eloszlású.

#ANOVA és Mauchly-próba
anova_test(data = g, dv=Cu, wid=ID, within = Season)
ANOVA Table (type III tests)
$ANOVA
 Effect DFn DFd F p p<.05 ges
1 Season 2 28 6.347 0.005 * 0.225

$`Mauchly's Test for Sphericity`
 Effect W p
1 Season 0.993 0.957
#A Mauchly-próba eredménye nem szignifikáns.
```

Az ANOVA-próba szignifikáns eredményt adott ($F = 6.35$, $df = 2$, $p = 0.005$), ezért post-hoc tesztet végzünk, Bonferroni-féle p-korrekcióval. Az eta-négyzet hatásnagyságértéke 0.225, ami jelentős hatást mutat.

```
#Post-hoc próba
comp_Cu = g %>%
 pairwise_t_test(Cu~Season, paired = TRUE,
 p.adjust.method = "bonferroni")
comp_Cu
# A tibble: 3 x 10
.y. group1 group2 n1 n2 statistic df p p.adj p.adj.sig
 <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <dbl> <chr>
1 Cu First Second 15 15 2.80 14 0.014 0.043 *
2 Cu First Third 15 15 3.20 14 0.006 0.019 *
3 Cu Second Third 15 15 0.485 14 0.636 1 ns
```

A post-hoc próba szerint beigazolódott a megfigyelésünk, miszerint a mezei csiperkében az első termés idején van a legnagyobb rézkoncentráció, ami szignifikánsan eltér a második ($t = 2.80$, $df = 14$, $p = 0.043$) és a harmadik ($t = 3.20$, $df = 14$, $p = 0.019$) terméstől. A második és a harmadik termés során a gomba rézkoncentrációja között nincs szignifikáns különbség ($t = 0.485$, $df = 14$, $p = 1$).

5.7. Nem parametrikus próbák kettőnél több csoport mediánjának egyenlőségére

5.7.1. A Kruskal–Wallis-próba

A próba William H. Kruskal és W. Allen Wallis statisztikusokról kapta a nevét, akik 1952-ben közölték az általuk kidolgozott alternatív megoldást az egyszempontos ANOVA próbához, ha annak nem teljesülnek az alkalmazhatósági feltételei (Kruskal és Wallis, 1952). Ahhoz, hogy a mediánokra tudjunk következtetéseket levonni, teljesülnie kell egy feltételnek: a minták eloszlása azonos kell hogy legyen; eltolással egymással fedésbe kell hogy kerüljenek. Ellenkező esetben csak az eloszlások egyformaságát vagy különbözőségét állapíthatjuk meg, több független minta egyforma méretéről kapunk információt (Vargha, 2015). A változó legalább ordinális skálán értelmezett kell hogy legyen.

A próba alapgondolata az, hogy az összes értéket rangsoroljuk, és elvárjuk, hogy a rangok eloszlása egyenletes legyen a mintákon belül, ha az összes minta egy populációból származik.

Hipotézisek:

H_0: a minták azonos populációból származnak (ill. a minták mediánjai egyenlők).

H_1: a minták nem azonos populációból származnak (ill. legalább két minta mediánja eltér egymástól).

A H próbastatisztika a következőképpen számolható ki, ha nincsenek kapcsolt rangok:

$$H = \frac{12}{N(N+1)} \cdot \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(N+1),$$

ahol N az össz-elemszám, g a minták (csoportok) száma, n_i az elemek száma csoportonként, R_i a rangok összege csoportoként. A H olyan χ^2-eloszlást követ, amelynek szabadságfoka $df = g - 1$. Ha kapcsolt rangok vannak jelen, akkor a H-t korrigálni kell úgy, hogy elosztjuk az alábbi kifejezéssel:

$$1 - \frac{\sum_i^g (t_i - 1) \cdot t_i \cdot (t_i + 1)}{N^3 - N},$$

ahol t (*ties*) egy adott csoportban levő kapcsolt rangok száma.

Post-hoc próba

A Kruskal–Wallis-próbára kapott szignifikáns eredmény szükségessé teszi a csoportok páronkénti összehasonlítását. Erre alkalmas a Dunn-próba. A *p*-értékek korrigálását több módszerrel végezhetjük el. A leggyakrabban használt módszerek a Bonferroni-, Holm-, Hommel-, Rom-, Hochberg-, Benjamini–Hochberg-módszer. A Bonferroni *p*-korrekció lényege, hogy egy összehasonlítás szignifikanciaszintje α/k, ahol k az összehasonlítások száma. Például, négy csoportnál összesen hat összehasonlítást végzünk, tehát a k értéke 6, minden próba α' értéke pedig $0.05/6 = 0.008$. A többi módszer ennek az alapmódszernek a továbbfejlesztése. Bonferroni módszerét Holm fejlesztette tovább egy szekvenciális algoritmust bevezetve. Ez az eljárás az elsőfajú hibát ugyanolyan hatékonyan korrigálja, mint a Bonferroni-módszer, viszont a próba ereje nagyobb (Eichstaedt et al., 2013). Hochberg módszere ugyancsak egy szakaszos módszer, ún. csökkenő eljárás, amelynek a részletesebb kifejtése betekintést enged a hasonló algoritmusok megértésébe. Hochberg csökkenő sorrendbe rendezte a k darab összehasonlítás *p*-értékeit: $p_1 > \ldots > p_k$. Első lépésben a legnagyobb *p*-értéket (p_1) az α szignifikanciaszinttel veti össze, és ha $p_1 < \alpha$, akkor az összes *p*-értéket szignifikánsnak tekinti. Ha $p_1 > \alpha$, akkor a következő *p*-értéket ($k = 2$) vizsgálja; ha $p_2 \leq \alpha/2$, akkor az összes $k > 2$ értéket szignifikánsnak tekinti. Ellenkező esetben a p_3 értéket viszonyítja $\alpha/3$-hoz stb. Általánosan felírva az algoritmus akkor áll le, ha:

$$p_k \leq \frac{\alpha}{k}.$$

Ezeknek a módszereknek a gyenge pontja viszonylag sok csoport összehasonlításánál jelentkezik, ugyanis ilyen esetben csökken a próba ereje. Sok pár összehasonlításánál annyira kis értékre csökkennek az egyedi α-értékek, hogy csak szélsőséges esetben vetjük el a nullhipotézist, így megnő a másodfajú hiba esélye (nem mutatjuk ki a hatást ott, ahol az a valóságban létezik). A Benjamini–Hochberg-módszer ellenőrzés alatt tartja a másodfajú hibát, ezért ezt elsősorban sok csoport összehasonlításakor érdemes használni (Benjamini és Hochberg, 1995). További részletek, ill. algoritmusok leírása Keselman et al. (2011) publikációjában tanulmányozható.

Hatásnagyság

Kruskal–Wallis-próba esetén a hatásnagyságot az eta-négyzettel jellemezhetjük, ami a H próbastatisztikából számolható ki az alábbi képlet segítségével:

$$\eta^2 = \frac{H - g + 1}{N - g},$$

ahol g a csoportok száma, N pedig az összes elem.

Példa

Az R alapcsomagjában levő *mtcars* adattábla 15 autójára az 5.15. táblázatban láthatjuk a lóerő értékeket az összsebességi fokozatok függvényében. Teszteljük a három csoport mediánjának egyenlőségét. A rangok csoportokon belüli összege alapján kiszámoljuk a *H* próbastatisztikát.

$$H = \frac{12}{15(15+1)} \cdot \frac{2450.25 + 289 + 2862.25}{5} - 3(15+1) = 8.0125$$

A korrekciós tag értéke:

$$1 - \frac{2 \cdot 3 \cdot 4 + 1 \cdot 2 \cdot 3 + 0 \cdot 1 \cdot 2}{15^3 - 15} = 0.991,$$

innen pedig a *H* értéke

$$H = \frac{8.0125}{0.991} = 8.085.$$

5.15. táblázat. *Az* mtcars *adattábla 15 autómárkájára megadott lóerő értékek az össz-sebeségfokozatok függvényében*

Mutatók	Sebességi fokok					
	3		4		5	
	Megfi-gyelés	*Rang*	Megfi-gyelés	*Rang*	Megfi-gyelés	*Rang*
	97	*6*	66	*3.5*	91	*5*
	150	*9.5*	52	*1*	113	*8*
	150	*9.5*	65	*2*	264	*14*
	245	*13*	66	*3.5*	175	*11.5*
	175	*11.5*	109	*7*	335	*15*
Medián	150		66		264	
R		*49.5*		*17*		*53.5*
R^2		*2450.25*		*289*		*2862.25*

A $df = 2$ szabadságfokkal rendelkező χ^2-eloszlás kritikus értéke 5.991 ($\alpha = 0.05$). A *H* értéke ennél nagyobb, tehát elvetjük a nullhipotézist, és fenntartjuk, hogy legalább két medián különbözik egymástól.

R-ben a **`kruskal.test()`** vagy **`kruskalTest()`** paranccsal végezhetjük el a próbát, a post-hoc tesztre **`kwAllPairsDunnTest()`**, az eta-négyzetet pedig a **`kruskal_effsize()`** paranccsal számolhatjuk ki. A parancsok részletes leírása az 5.16. táblázatban látható.

5.16. táblázat. *A Kruskal–Wallis- és a post-hoc próba mediánokra*

Parancs	Részletek	Magyarázat
kruskal.test(x, g, data) kruskalTest(x, g, data, dist) {PMCMRplus} kwAllPairsDunnTest(x, g, p.adjust.method) {PMCMRplus} kruskal_effsize(x∼g, data) {rstatix}		
x		függő változó
g		csoportváltozó
data		adattábla
dist	"Chisquare", "KruskalWallis", "FDist"	a próbastatisztika eloszlása
p.ajust.method	"holm", "bonferroni", "hommel", "hochberg", "BH"	a *p*-érték korrigálása BH: Benjamini–Hochberg

```
#K-W próba az mtcars adattábla utolsó 15 objektumára
library(PMCMRplus)
d = mtcars[18-32,]
kruskalTest(hp~gear, data = d, dist = "Chisquare")
Kruskal-Wallis test
data: hp by gear
chi-squared = 8.0582, df = 2, p-value = 0.01779
#Mivel a p < 0.05, elvetjük a H0 hipotézist.
#Post-hoc próba
kwAllPairsDunnTest(d$hp~d$gear, p.adjust.method = "holm")
Pairwise comparisons using Dunn's all-pairs test
data: d$hp by d$gear
 3 4
4 0.042 -
5 0.777 0.029
P value adjustment method: holm
alternative hypothesis: two.sided
Warning message:
In kwAllPairsDunnTest.default(c(66,52,65,97,150,150,245,175:
Ties are present. z-quantiles were corrected for ties.
#Hatásnagyság (eta-négyzet)
kruskal_effsize(hp~gear, data = d)
#A tibble: 1 x 5
 .y. n effsize method magnitude
* <chr> <int> <dbl> <chr> <ord>
1 hp 31 0.427 eta2[H] large
```

A post-hoc próbát a Holm-módszerrel végeztük el. Az R ad egy üzenetet, amelyben jelzi, hogy az adatsorban kapcsolt rangok vannak, és hogy ezeket sikerült-e korrigálni. Ebben az esetben sikerült a korrekció. A hatásnagyság nagy (eta-négyzet = 0.427), tehát a Kruskal–Wallis-próba szignifikáns eredménye elfogadható.

5.7.2. Jonckheere–Terpstra-próba

Abban az esetben, ha a minták egy szempont szerint növekvő sorrendbe rendezhetők (például vizsgáljuk egy kultúrnövény biomasszáját olyan területeken, amelyekre egyre nagyobb mennyiségben adagolunk műtrágyát), akkor a Jonckheere–Terpstra-próba megbízhatóbb eredményt ad, mint a Kruskal–Wallis-próba (Jonckheere, 1954; Terpstra, 1952). A Jonckheere–Terpstra-próba nullhipotézise a mediánok egyenlősége, az alternatív hipotézis pedig azt állítja, hogy legalább két csoport mediánjában eltérés tapasztalható. R-ben a {PMCMRplus} csomagban megtalálható a **`jonckheereTest()`** parancs erre a próbára (5.17. táblázat).

5.17. táblázat. *A Jonckheere–Terpstra-próba rendezett mintákra*

Parancs	Részletek	Magyarázat
jonckheereTest(x, g, alternative) {PMCMRplus}		
x		függő változó
g		csoportváltozó
alterantive	"two.sided", "greater", "less"	A kétoldali teszt egyenértékű a korreláció teszttel Kendall tau-ra. greater: növekvő sorrend less: csökkenő sorrend
continuity	FALSE/TRUE	FALSE (kapcsolt rangok jelenlétében) TRUE (nincsenek kapcsolt rangok)

Legyen egy laboratóriumi kísérlet, amely során magok csíráztatásának hatékonyságát tesztelik különböző rézkoncentrációjú (0, 5, 10, 25, 50 μg/l) oldatokban, 6-6 Petri-csészében. Az adatokat bevezetjük R-be:

```
conc = c(rep(0,6), rep(5,6),rep(10,6),rep(25,6))
siker = c(30,27,25,23,28,23,24,27,25,24,27,28,22,
 25,21,23,25,20,18,22,16,20,19,18)
library(PMCMRplus)
```

```
    jonckheereTest(siker, conc, alternative = "less")
    Jonckheere-Terpstra test
    data: siker and conc
    z = -3.9314, p-value = 4.223e-05
    alternative hypothesis: less
    sample estimates:
    JT 32
    Warning message:
    In jonckheereTest.default(siker, conc, alternative = "less":
Ties are present. Jonckheere z was corrected for ties.
```

A próba szignifikáns eredményt adott ($z = -3.93$ p < 0.00005), tehát állíthatjuk, hogy a rézkoncentráció növekedésével csökkent a kicsírázott magok száma. A réz 5–50 μg/l koncentrációban gátolja a magok csírázását. A mintában kapcsolt rangok vannak, amelyeket a próba kezelni tud.

5.7.3. Friedman-próba több összetartozó minta összehasonlítására

A Friedman-próba több összetartozó minta összehasonlítására alkalmas nem parametrikus próba, amely az ismételt méréses ANOVA alternatívája, amikor nem teljesül a minták normális eloszlásának feltétele. A próba nullhipotézise azt állítja, hogy a g_1, g_2, …, g_k mintáknak azonos eloszlása van. A próba a Kruskal–Wallis-próbához hasonlóan rangsorolja az N elemet, minden csoportra kiszámolja a rangok összegét (R_1, R_2, …, R_k), majd kiszámolja a G próbastatisztikát. Ha az $N > 5$, akkor a G megközelítően azt a χ^2-eloszlást követi, amelynek a szabadságfoka $df = g - 1$.

A G próbastatisztikát a következő képlettel értelmezzük:

$$G = \left(\frac{12}{N \cdot k(k+1)} \sum_{i=1}^{g} R_i^2 \right) - 3N(k+1),$$

ahol N az összmintaelemszám, k a minták száma, g csoportok száma, R a rangösszegek mintánként.

Post-hoc próbák és a hatásnagyság

A szignifikáns eredmény után post-hoc teszttel azonosíthatjuk az egymástól eltérő mintákat. A Neményi-próba rangösszegek közti különbségek kimutatására alkalmas, ami a Tukey-próba nem parametrikus változata, ill. a Connover-próba is használható. Ezek a próbák az elsőfajú hibát tartják ellenőrzés alatt. A {PMCMRplus} csomagban mind a két próba megtalálható. A Friedman-próba hatásnagyságát az {rstatix} csomag **`friedman_effsize()`** parancsával vizsgálhatjuk. A Friedman-

próbát elvégezhetjük az {rstatix}, valamint a {PMCMRplus} csomag használatával, a **`friedman_test()`**, ill. a **`friedmanTest()`** paranccsal (5.18. táblázat).

5.18. táblázat. *A Friedman-próba összetartozó mintákra*

Parancs	Részletek	Magyarázat
friedmanTest (x, g, ID, data) {PMCMRplus} friedman_test(x∼g | ID, data) {rstatix} friedman_effsize(x∼g | ID, data) {rstatix} kwAllPairsNemenyiTest(x, g, data) {PMCMRplus} kwAllPairsConoverTest(x, g, data) {PMCMRplus}		
x	Numeric	függő változó
g	Numeric	csoportváltozó
ID	Factor	a minták kódjai
data	Data frame	adattábla a változókkal

Példa

Módosítjuk az előző pontban tárgyalt példát úgy, hogy az egyetlen magféleség helyett hat különféle magtípus (M1, …, M6) csírázását vizsgáljuk négy különböző rézkoncentrációjú (0, 5, 10, 25, 50 μg/l) tápoldatban. A rézkoncentrációk és a kicsírázott magok száma mellett bevezetjük a magféleségek kódját is, majd a három változót egy adattáblába rendezzük:

```
ID = c(rep(c("M1","M2","M3","M4","M5","M6"),4))
ID = as.factor(ID)
conc = c(rep(0,6), rep(5,6),rep(10,6),rep(25,6))
siker = c(30,27,25,23,28,23,24,27,25,24,27,28,22,
 25,21,23,25,20,18,22,16,20,19,18)
d = data.frame(ID, conc, siker)
head(d)
  conc siker ID
1    1    30 M1
2    1    27 M2
3    1    25 M3
4    1    23 M4
5    1    28 M5
6    1    23 M6

#Friedman-próba
library(PMCMRplus)
friedmanTest(siker, conc, ID, data = d)
Friedman rank sum test
```

```
data: y, groups and blocks
Friedman chi-squared = 16.158, df = 3, p-value = 0.001053

#Post-hoc: Conover-próba
kwAllPairsConoverTest(siker, conc, data = d)
Pairwise comparisons using Conover's all-pairs test
data: siker and conc
   1       5       10
5  0.99989 -       -
10 0.06677 0.05846 -
25 0.00016 0.00014 0.05846

P value adjustment method: single-step
Warning message:
In kwAllPairsConoverTest.default(siker, conc, data = d) :
Ties are present. Quantiles were corrected for ties.

#Hatásnagyság
library(rstatix)
friedman_effsize(siker~conc|ID, data = d)
# A tibble: 1 x 5
 .y. n effsize method magnitude
* <chr> <int> <dbl> <chr> <ord>
1 siker 6 0.898 Kendall W large
```

A Friedman-próba szignifikáns eredményt adott ($G = 16.16$, $df = 3$, $p = 0.001$). Ezzel összhangban a hatásnagyság is jelentős (0.898), tehát elmondhatjuk, hogy a rézkoncentráció hatással van a magok csírázására. A páronkénti összehasonlítás az 1 és 25 μg/l, illetve az 5 és 25 μg/l oldattal kezelt magok csírázási képessége között mutatott szignifikáns eltérést, ami azt mutatja, hogy a tömény (10 μg/l-nél nagyobb) rézkoncentráció esetén jelentősen csökken a magok csírázási képessége.

5.8. Faktoriális (többszempontos) ANOVA

Egy környezeti változóra gyakorolt hatások nagyon bonyolultak, összetettek lehetnek. Amennyiben egy témát a maga valóságában szeretnénk vizsgálni, akkor lehetséges, hogy több faktor hatását is figyelembe kell vennünk. A faktoriális ANOVA képes egyszerre több tényező hatását kiértékelni, amit két lépésben tesz meg: egyrészt külön-külön vizsgálja a faktorok hatását a csoportátlagokra az egyszempontú ANOVA szerint, másrészt pedig vizsgálja, hogy az egyik faktor hatását

befolyásolja-e egy másik faktor. Ezt interakcióvizsgálatnak nevezzük. A faktoriális ANOVA esetén tehát két kérdésre keressük a választ:

– van-e szignifikáns interakció a faktorok között;
– megváltoztatja-e a függő változó átlagértékét egy változás a független változóban?

Ennek alapján megkülönböztetünk főhatásokat (egy-egy faktornak a vizsgált függő változóra gyakorolt önálló hatása), és interakciós hatást (a faktorok együttes hatása a függő változóra). Innen következik, hogy a faktoriális ANOVA esetében több null-, illetve alternatív hipotézist fogalmazhatunk meg (5.19. táblázat).

5.19. táblázat. *A faktoriális ANOVA hipotézisei*

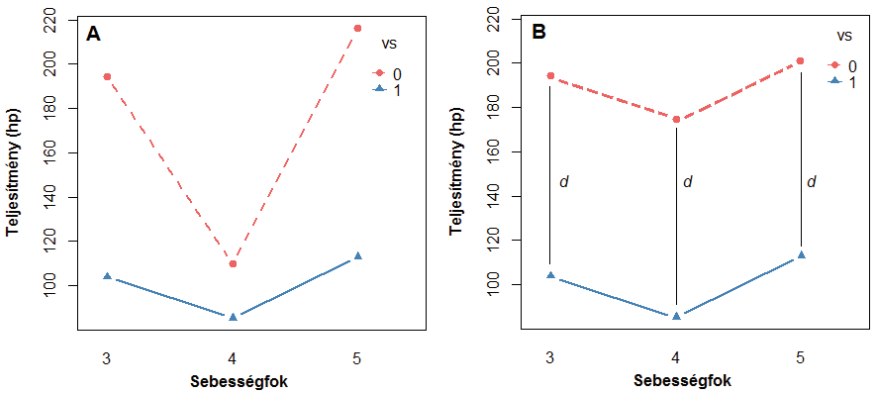
Nullhipotézisek	Ellenhipotézisek
Nincs eltérés az F_1 faktor szintjeire a függő változó átlagaiban.	Legalább két szinten eltérés van az átlagokban.
Nincs eltérés az F_2 faktor szintjeire a függő változó átlagaiban.	Legalább két szinten eltérés van az átlagokban.
Az egyik fakor hatása a függő változó átlagára nem függ a másik faktortól.	Interakció van a két faktor között (egymást befolyásolják).

Például az mtcars adattáblában a különféle autók teljesítményét két szempontból vizsgáljuk: a sebességi fokok száma (3, 4, 5) és a motor alakja (V-alakú: 0, egyenes: 1) szerint (5.20. táblázat). A két faktornak összesen hat szintje van, így ezt a modellt 2x3-as mintázatnak nevezzük (5.10.A. ábra).

5.20. táblázat. *Az mtcars autóinak átlagteljesítménye (lóerő)*

Motor alakja	Sebességfok			Átlag
	3	4	5	
0	194.2	110.0	216.2	189.7
1	104.0	85.4	113.0	91.4
Átlag	176.1	89.5	195.6	146.7

```
#Az 5.10.A. ábra kódja
?mtcars
d = mtcars
attach(d)
interaction.plot(gear, vs, hp, fun = mean, type = "b",
 col = c("indianred2", "steelblue"), pch = c(19,17),
 font.lab = 2, lwd = 2, xlab = "Sebességfok",
 ylab = "Teljesítmény (hp)")
#type = "b" (jelzi és összeköti az átlagokat)
```



5.10. ábra. *Egy 2x3-as mintázatú faktoriális ANOVA interakciós ábrája: A. valós helyzet; B. elméleti ábra (ha nem lenne interakció a két faktor között)*

```
#Átlagok kikérése
round(tapply(d[,4],d[,c(8,10)],mean),1)
 gear
vs 3 4 5
 0 194.2 110.0 216.2
 1 104.0 85.4 113.0
round(tapply(d[,4],d[,8],mean),1)
 0 1
189.7 91.4
round(tapply(d[,4],d[,10],mean),1)
 3 4 5
176.1 89.5 195.6
mean(hp)
[1] 146.7
```

Amennyiben nincs interakció a két faktor között, azt várjuk el, hogy a motor két alakjának esetében az egyes sebességfokozatokra a teljesítménykülönbség állandó legyen (5.10.B. ábra, d = állandó). Az 5.10.A. ábra alapján feltételezzük, hogy a két faktor hat egymásra, viszont ennek a hatásnak a valódiságát a faktoriális ANOVA próbával ellenőrizni kell.

Alkalmazhatósági feltételek:

A faktoriális ANOVA alkalmazhatósági feltételei egyeznek az egyszempontos ANOVA feltételeivel. Ezenkívül feltétel, hogy a faktorok egymástól függetlenek

legyenek. Ezt nehezen lehetne tesztelni, így a kutatástervezésnél úgy kell kiválasztani a faktorokat, hogy elméletileg egymástól függetlenek legyenek.

– Az adatsorok normális eloszlásúak.
– Homogén szórások minden csoportra.
– A csoportok egymástól függetlenek (faktoron belül és faktorok között is).

Post-hoc próba és hatásnagyság

A post-hoc próbák egyeznek az egyszempontos ANOVA próbáival (lásd az 5.6.1. alfejezetet), azzal a kiegészítéssel, hogy az interakciókra is elvégezhetők. A hatásnagyságot a parciális η^2-tel fejezhetjük ki, ami ebben az esetben az adott faktor szórásnégyzetének aránya a faktor szórásnégyzetének és a hiba szórásnégyzetének összegével:

$$\eta_p^2 = \frac{Q_1}{Q_1 + Q} \text{ az } F_1 \text{ faktorra,}$$

ahol Q_1 az F_1 szórásnégyzeteinek összege, a Q a hibatag szórásnégyzeteinek összege.

Egyszempontos ANOVA-nál a parciális és a teljes η^2 megegyeznek. A parciális η^2 kiszámolható az ANOVA *F*- és *df*-értékének a felhasználásával is:

$$\eta_p^2 = \frac{F \cdot df_1}{F \cdot df_1 + df_{hiba}}.$$

A parancsok R-ben az 5.21. táblázatban láthatók.

5.21. táblázat. *A többszempontos ANOVA elvégzése R-ben*

Parancs	Részletek	Magyarázat
model = aov(x~F1+F2+F1*F2, data) summary(model) PostHocTest(model, method, conf.level = 0.95, ordered = F) {DescTools} eta_squared(model, partial = T, ci = 0.95, alternative = "two.sided") {effectsize} cohens_f(model, partial = T, ci = 0.95, alternative = "two.sided") {effectsize}		
x	Numeric	függő változó
F1, F2	Faktorok	csoportváltozók
data	Data frame	adattábla a változókkal
method	"hsd", "bonferroni", "lsd", "scheffe", "newmankeuls", "duncan"	a post-hoc próba típusa
conf.level, ci	0.95 ($\alpha = 0.05$)	konfidenciaintervallum
partial	TRUE/FALSE	a parciális érték megadása (alapbeállítás)
alternative	"two.sided", "greater", "less"	*ci* tartomány megadása

A kétszempontos ANOVA próbát a **`model = aov()`** paranccsal végezhetjük el R-ben. Az eredmény egy táblázat formájában tekinthető meg a **`summary(model)`** paranccsal, és ami egyezik az 5.22. táblázattal. A szignifikáns eredményekre post-hoc próbát végezhetünk, valamint a hatásnagyságot is kiszámolhatjuk az egyszempontos ANOVA-ra érvényes próbákkal.

5.22. táblázat. *A kétszempontos ANOVA eredményének megadása*

Faktorok	Szabadságfokok	Négyzetösszegek	Variancia	*F*-test
F_1	$df_1 = G_1-1$	Q_1	$MS_1 = Q_1/df_1$	MS_1/MS
F_2	$df_2 = G_2-1$	Q_2	$MS_2 = Q_2/df_2$	MS_2/MS
$F_1 \times F_2^*$	$df_{12} = (G_1-1)\cdot(G_2-1)$	Q_{12}	$MS_{12} = Q_{12}/df_{12}$	MS_{12}/MS
Hibatag	$df = N-G_1\cdot G_2$	Q	$MS = Q/df$	

* *interakció*

G_1, G_2 – *a két faktor szerinti csoportok*

A példánknál a következő hipotéziseket teszteljük:

H_0: Az átlagteljesítmény nem függ a sebességi fokok számától.
Az átlagteljesítmény nem függ a motor alakjától.
A sebességi fokok száma és a motor alakja nem hatnak egymásra a teljesítmény tekintetében.

H_1: Az átlagteljesítmény függ a sebességi fokok számától.
Az átlagteljesítmény függ a motor alakjától.
A sebességi fokok száma és a motor alakja hatnak egymásra a teljesítmény tekintetében.

A faktoriális ANOVA lépésről lépésre R-ben:

```
#Kétszempontos ANOVA a teljesítményre (hp)
attach(mtcars)
model=aov(hp~gear+vs+gear*vs)
summary(model)
 Df Sum Sq Mean Sq F value Pr(>F)
gear 2 64213 32106 15.911 3.07e-05***
vs 1 23753 23753 11.771 0.00202**
gear:vs 2 5296 2648 1.312 0.28647
Residuals 26 52465 2018
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#A két faktor szignifikáns eredményt adott,
#az interakció nem szignifikáns.
```

```
#Post-hoc próba (Scheffe)
library(DescTools)
PostHocTest(model, method = "scheffe")
 Posthoc multiple comparisons of means: Scheffe Test
 95% family-wise confidence level

$gear
            diff     lwr.ci    upr.ci       pval
4-3   -86.63333 -149.20225 -24.06442     0.0025**
5-3    19.46667  -63.95855 102.89189     0.9810
5-4   106.10000   20.10725 192.09275     0.0086**

$vs
            diff     lwr.ci    upr.ci      pval
1-0    -43.1746  -100.7435  14.39427    0.2384

$`gear:vs`
                 diff      lwr.ci     upr.ci        pval
4:0-3:0    -84.16667  -207.55423   39.22090      0.33508
5:0-3:0     22.08333   -71.18890  115.35557      0.97979
3:1-3:0    -90.16667  -194.44819   14.11486      0.12284
4:1-3:0   -108.76667  -177.93921  -39.59413     0.00054***
5:1-3:0    -81.16667  -249.31558   86.98224     0.69830
5:0-4:0    106.25000   -33.65835  246.15835     0.22662
3:1-4:0     -6.00000  -153.47635  141.47635     1.00000
4:1-4:0    -24.60000  -149.73783  100.53783     0.99126
5:1-4:0      3.00000  -194.86029  200.86029     1.00000
3:1-5:0   -112.25000  -235.63757   11.13757     0.09219.
4:1-5:0   -130.85000  -226.42560  -35.27440     0.00290**
5:1-5:0   -103.25000  -283.87090   77.37090     0.53023
4:1-3:1    -18.60000  -124.94671   87.74671     0.99491
5:1-3:1      9.00000  -177.54447  195.54447     0.99999
5:1-4:1     27.60000  -141.83743  197.03743     0.99636
#p<0.05 csillaggal jelöltek.

#Hatásnagyság
library(effectsize)
eta_squared(model, partial = T, alternative = "two.sided")
#Effect Size for ANOVA (Type I)
Parameter | Eta2 (partial) |       95% CI
-----------------------------------------
gear      |           0.55 | [0.26, 0.71]
vs        |           0.31 | [0.06, 0.54]
```

```
gear:vs   |           0.09 | [0.00, 0.31]
#A hatásnagyság jelentős a gear és vs változó esetén.
```

A kétszempontos ANOVA szignifikáns eredményt adott az F_1 és F_2 faktor hatására, az interakció ellenben nem szignifikáns. Noha az 5.10. ábrán úgy tűnt, hogy van eltérés a két csoporthoz (*vs*) tartozó átlagok (*gear*) között, a valóságban ezek az eltérések nem szignifikánsak. Ezt támasztják alá a parciális eta-négyzet értékek is. A post-hoc próba rávilágított a páronkénti szignifikáns eltérésekre.

5.8.1. Robusztus alternatívák a faktoriális ANOVA-ra

A faktoriális ANOVA eléggé robusztus a normalitással és a varianciák egyenlőségével szemben támasztott feltételek bizonyos szintű megsértésére. Ha a függő változó eloszlása a faktorok szerint egyirányú ferdeséget mutat, és a minták elemszámai nem térnek el jelentősen egymástól, akkor a próbát a varianciák egyenlőségének a megsértése mellett is lehet használni. Ellenkező esetben egy robusztus változatot kell használni. Erre több lehetőség van, ezzel kapcsolatban részletesebb információk olvashatók Erce–Hurn és Mirosevich (2008), valamint Keselman et al. (2003) munkáiban. Az egyik legelfogadottabb módszer a Welch–James-próba, amely a Welch-próbát alkalmazza a faktorok szerinti átlagok összehasonlítására, és a Johansen által leírt módszer az interakciók vizsgálatára (Johansen, 1980).

A {welchADF} csomag lehetővé teszi ennek a próbának az elvégzését. A szerző a csomag részletes jellemzéséről cikket jelentett meg az *R Journal*ben (Villacorta, 2017). A csomagban szerepel egy adattábla, amely egy 2005-ben megjelent pszichológiai kísérlet eredményeit reprodukálja, sarkítva a csoportokon belüli elemszámokat (a szintekhez tartozó elemszámok 45-50 között változnak) a Welch–James-próba használatához. A kutatás célja a sztereotípiák érvényességének vizsgálata volt, miszerint a nők gyengébben teljesítenek számtanpéldákban, mint a férfiak (Wicherts et al., 2005). Az adattábla két faktort (condition, sex) és egy numerikus (y) változót tartalmaz. A *condition* faktornak három szintje van: kontroll, lenullázott és sztereotípiás csoport. A lenullázott csoportban a tesztlapon közölték, hogy ebben a feladattípusban a nők bizonyítottan egyformán teljesítenek a férfiakkal, a sztereotípiás csoportnál pedig azt, hogy a férfiak bizonyítottan jobban teljesítenek. Az *y* változó a teszten elért pontszámokat tartalmazza.

```
library(welchADF)
d = womenStereotypeData
head(d)
```

```
          condition    sex     y
      1       control   male   10
      2       control   male   10
      3       control   male   16
      4       control   male   20
      5       control   male   35
      6       control   male   13
```

A Welch–James-próbát R-ben a **`welchADF.test()`** paranccsal végezhetjük el (5.23. táblázat).

5.23. táblázat. *A Welch–James-próba R-ben*

Parancs	Részletek	Magyarázat
model = welchADF(data, response, between.s, constrast, trimming, per, effect.size, bootstrap) summary(model) {welchADF}		
data	Dataframe	adattábla a változókkal
response	Numeric	függő változó
between.s	Factors	faktorok
contrast	"omnibus", "all.pairwise"	a post-hoc próba típusa
trimming	FALSE/TRUE	szélsőértékek eltávolítása
per	0.2	Ha trimmelést végzünk, beállíthatjuk a trimmelés mértékét. Alapból 20%.
effect.size	FALSE/TRUE	hatásnagyság megadása

```
#Welch–James-próba
model = welchADF.test(d, response = "y",
 between.s = c("condition", "sex"))
summary(model)
Call:
          WJ statistic  NumeratorDF DenominatorDF   Pr(>WJ)
   condition     2.151            2         154.7   0.11986
         sex     2.933            1         216.4   0.08824.
condition:sex     2.521            2         154.7   0.08368.
---
Signif. codes: 0 ,***’ 0.001 ,**’ 0.01 ,*’ 0.05 ,.’ 0.1
```

Az eredmény azt mutatja, hogy sem a befolyásolásnak (kondicionálásnak), sem a kitöltő nemi hovatartozásának nincs szignifikáns hatása a teszteredményekre. Az interakció sem szignifikáns. Ennek ellenére, ha a szélsőértékeket eltávolítjuk trimmeléssel (20%), eltérő eredményhez jutunk:

```
    #Welch-James-próba trimmelt adatokra
    model = welchADF.test(d, response = "y",
     between.s = c("condition", "sex"),
     trimming = TRUE)
    summary(model)
    Call:
           WJ statistic NumeratorDF DenominatorDF   Pr(>WJ)
    condition     5.205           2         93.38  0.007189**
          sex     5.754           1        130.06  0.017875*
condition:sex     3.130           2         93.38  0.048347*

---
    Signif. codes: 0 ,***' 0.001 ,**' 0.01 ,*' 0.05 ,.' 0.1

    #Post-hoc próba az interakciókra (bootstrap)
    posthoc = welchADF.test(y ~ condition*sex, data = d,
     contrast = "all.pairwise",
     effect = c("condition","sex"),
     trimming = TRUE,
                                bootstrap = TRUE)
    summary(posthoc)
     significant?
  control:stereotype    x  female:male      no
 nullified:stereotype   x  female:male      no
    control:nullified   x  female:male     yes
    Bootstrap critical value: 5.211

    #Post-hoc próba a befolyásolásra (Hochberg)
    ph_condition = welchADF.test(y~condition, data = d,
     contrast = "all.pairwise",
     effect = c("condition"),
     trimming = TRUE,
     bootstrap = F)
    summary(ph_condition)
                     WJ statistic   adj.pval
      control:stereotype    5.1031    0.05168.
     nullified:stereotype    8.3165    0.01437*
        control:nullified    0.1856    0.66743
    ---
    Signif.codes (Hochberg p-values: 0'***' 0.001'**' 0.01'*'
```

Itt már az látszik, hogy mind a befolyásolásnak, mind a nemekhez való tartozásnak hatása van a matematikai teszteredményekre. A post-hoc próbát elvégeztük az interakciókra, majd a *condition* faktor szintjeire is. Az elsőt a bootstrap ismételt véletlenszerű mintavétellel végeztük, a második esetben pedig a Hochberg-módszert használtuk. Az első esetben csupán egy érték volt nagyobb a kritikus értéknél (5.766 > 5.211), amikor is a kontroll és a sztereotípiát kioltott csoportok esetében a dolgozatok eredménye szignifikánsan függött attól, hogy melyik nemhez tartozik a kitöltő. A második esetben, amikor az egyszempontos ANOVA post-hoc próbája szerint értelmeztük az eredményt, a legnagyobb különbség a pontszámokban a sztereotípiát kioltott, ill. a sztereotípiákkal rendelkező csoportok között volt tapasztalható.

```
    #Hatásnagyság az interakciókra
    effect = update(model, effect.size = TRUE)
    effect
    Effect size standardized via square root of the average of
cell variances:
    control:stereotype x female:male    -0.072170
    nullified:stereotype x female:male   -0.582649
    control:nullified x female:male       0.523038
```

A három interakció közül kettőnél találtunk nagy hatásnagyságot. Ezek közül az utolsó az, amelyre a post-hoc próba is szignifikáns eredményt adott.

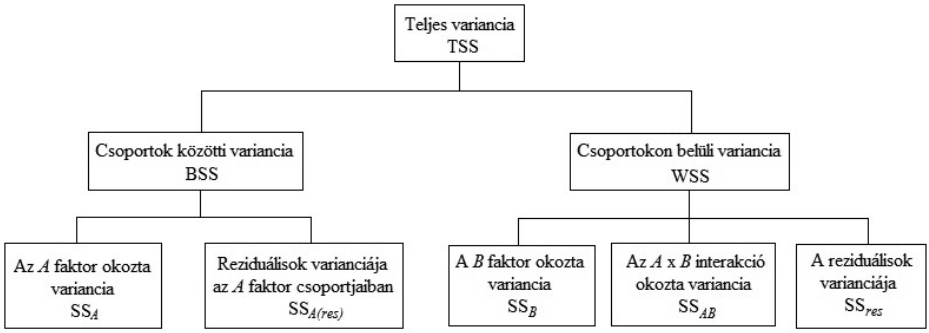
5.8.2. Beágyazott varianciaanalízis

Léteznek olyan helyzetek, amikor egy folytonos, normális eloszlású változóra két olyan független változó hat, amelyek közül az egyiknek egymástól független szintjei vannak, a másiknak pedig egymástól függő szintjei. Például több éven át, különböző hatásoknak kitett területeken kijelölt kvadrátokban vizsgálják a biodiverzitás változását. Ebben az esetben a területek egymástól függetlenek, a különböző évek területre szabott adatai egymástól függenek. Tehát a területeket összesítő *A* változó (*I* szinttel) független mintákat tartalmaz, az időbeli *B* változó (*J* szinttel) pedig egymástól függő mintákat.

A próba hipotézisei egyeznek a többszempontos független mintás ANOVA hipotéziseivel: külön teszteljük a két faktor hatását (ezek a főhatások) és az interakciót. Mivel a két faktor lényegesen eltér egymástól, az *A* a csoportok közti varianciát (BSS) határozza meg, a *B* a csoportokon belüli varianciát (WSS), amelyek további résztagokból épülnek fel (5.11. ábra).

Az ismételt méréses ANOVA a csoportok közötti varianciát vizsgálja, lehetővé téve az alanyok/egyedek varianciájának eltávolítását azzal a céllal, hogy a

faktor által okozott varianciára összpontosítson az alanyok között. Jelen esetben, ahol méréseket ismételünk a *B* faktor szerint, de nem *A* szerint, a csoportok közötti variancia (BSS) két tagra bomlik: SS_A és $SS_{A(res)}$. Az SS_A a különböző szintek (*I*) által okozott szóródás mértéke (pl. különböző mértékű beavatkozások, kezelések, hatások miatt). Az $SS_{A(res)}$ az *A* faktorhoz tartozó egyedek/alanyok egyéni variabilitásának összege. A H_0(A) hipotézis azt teszteli, hogy az átlagos SS_A (MS_A) szignifikánsan nagyobb-e az átlagos $SS_{A(res)}$ ($MS_{A(res)}$) értékénél, azaz az *A* faktor által okozott szóródás a csoportok között nagyobb-e az alanyok/egyedek egyéni szóródásánál.



5.11. ábra. *A variancia összetétele a vegyes kétszempontú ANOVA esetén*

Minden alany/egyed részesült a *B* faktor minden szintjéből, ezért a H_0(B) azt teszteli, hogy a különböző szintek átlagértékeinek varianciája (SS_B) eltér-e a tesztelt alanyoknak/egyedeknek a varianciájától (SS_{res}). Mivel ez egy vegyes kétszempontos ANOVA, megvan annak is a valószínűsége, hogy a két faktor befolyásolja egymást. Ez abban nyilvánul meg, hogy az *A* faktor által meghatározott csoportokban a *B* faktor hatása különböző mértékben érvényesül. Ezt teszteli a H_0(A*x*B) hipotézis, ami akkor ad szignifikáns eredményt az interakcióra, ha az interakcióból adódó átlagos SS_{AB} (MS_{AB}) nagyobb, mint a reziduálisok átlagos négyzetösszege (MS_{res}). Az 5.24. táblázat összegzi ezeket a varianciákat és az *F*-értékeket (próbastatisztikákat).

Alkalmazhatósági feltételek

A beágyazott ANOVA alkalmazásához több feltételnek egyszerre kell teljesülnie:

– a függő változó folytonos skálájú kell legyen;
– minden csoport normális eloszlású kell legyen (például egy 2x3 szintes tesztben mind a 6 csoportot ellenőrizni kell);
– egyetlen csoportban sem szabad legyen kiugró érték;
– az *A* faktor által meghatározott csoportok varianciái egyenlők kell legyenek (Levene-próba);

– a *B* faktor által meghatározott csoportok páronkénti különbségeinek a varianciái egyenlők kell legyenek (Mauchly-próba).

5.24. táblázat. *A vegyes kétszempontos ANOVA eredményének megadása*

Faktorok	Szabadságfokok	Négyzetösszegek	Variancia	*F*-test
		Csoportok között		
A-faktor	$df_A = I-1$	Q_A	$MS_A = Q_A/df_A$	$MS_A/MS_{A(res)}$
A-hibatag	$df_{A(res)} = (n-I)$	$Q_{A(res)}$	$MS_{A(res)} = Q_{A(res)}/df_{A(res)}$	
		Csoportokon belül		
B-faktor	$df_B = J-1$	Q_B	$MS_B = Q_B/df_B$	MS_B/MS_e
B-hibatag	$df_{res} = I\cdot(n-I)\cdot(J-1)$	Q_{res}	$MS_{res} = Q_{res}/df_{res}$	
AxB interakció	$df_{AB} = (I-1)\cdot(J-1)$	Q_{AB}	$MS_{AB} = Q_{AB}/df_{AB}$	MS_{AB}/MS_{res}

A-faktor (független mintás faktor), B-faktor (ismételt mérések faktora)

I, J – az A és B faktorok szerinti csoportok
n – egy csoportban levő egyedek száma

Post-hoc próba

A post-hoc próbákat arra a faktorra, illetve arra az interakcióra kell elvégezni, amelyre szignifikáns eredményt kaptunk. Az *A* faktor esetén az egyszempontú ANOVA próbánál tárgyalt próbákat lehet használni, a *B* faktorra pedig az ismételt méréses ANOVA próbánál tárgyaltakat.

A vegyes kétszempontos ANOVA R-parancsai az 5.25. táblázatban láthatók.

5.25. táblázat. *A beágyazott ANOVA elvégzése R-ben*

Parancs	Részletek	Magyarázat
modell = anova_test(data, dv, wid, within, between) {rstatix} get_anova_table(modell) {rstatix} pairwise_t_test(x~g, paired = TRUE/FALSE, p.adjust.method) {rstatix}		
data		adattábla
dv	dependent variable	függő változó
wid	within identities	objektumok kódjait tartalmazó változó (a kódok minden csoportban ugyanazok); át kell alakítani faktor változóra
within	groups	*B* faktor (faktorrá kell alakítani)
between	groups	A faktor (faktorrá kell alakítani)
p.adjust.method	*p*-érték korrekciója	"holm", "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr", "none"

Az R adatbázisában van egy *ToothGrowth* nevű adattábla, ami Crampton (1947) adatait tartalmazza. Crampton a fogak növekedéséért felelős odontoblaszt sejtek hosszát vizsgálta tengerimalacoknál a C-vitamin-adagolás függvényében (*dose*: 0.5, 1, ill. 2 mg/nap). A C-vitamint kétféleképpen adagolta: narancslé és aszkorbinsav formájában (*supp*: OJ és VC). Összesen hat csoportot képezett, mindegyikben 10-10 egyeddel dolgozott ($N = 60$). A *supp* faktor csoportjai egymástól függetlenek (*A* faktor), a *dose* faktor csoportjai egymástól függnek (*B* faktor). A két faktor hatását egy 2x3 típusú beágyazott ANOVA próbával tesztelhetjük.

```
    #Adattábla beolvasása és rendezése
    d = ToothGrowth
    d$id = rep(1:10, 6) #Hozzáadjuk az egyedeket
    #Faktorrá alakítjuk a faktorváltozókat
    d$dose = as.factor(d$dose)
    d$id = as.factor(d$id)
    head(d)
       len supp  dose  id
     1  4.2   VC   0.5   1
     2 11.5   VC   0.5   2
     3  7.3   VC   0.5   3
     4  5.8   VC   0.5   4
     5  6.4   VC   0.5   5
     6 10.0   VC   0.5   6
#Feltételek tesztelése
    library(tidyverse)
    #az adattábla egyszerűbb kezeléséért használjuk
    library(rstatix)

    #Kiugró értékek jelenlétének tesztelése
    d %>%
     group_by(supp, dose) %>%
     identify_outliers(len)
    #A tibble: 2 x 6
       supp  dose    len     id  is.outlier  is.extreme
      <fct> <fct>  <dbl>  <fct>       <lgl>       <lgl>
     1    OJ     2   30.9      6        TRUE       FALSE
     2    VC     1   22.5      5        TRUE       FALSE
#Nincsenek extrém kiugró értékek
```

```
#Normális eloszlás a 6 csoportban
d %>%
 group_by(supp, dose) %>%
 shapiro_test(len)
#A tibble: 6 x 5
     supp   dose variable   statistic       p
    <fct>  <fct>    <chr>       <dbl>   <dbl>
 1     OJ    0.5      len       0.893   0.182
 2     OJ      1      len       0.927   0.415
 3     OJ      2      len       0.963   0.815
 4     VC    0.5      len       0.890   0.170
 5     VC      1      len       0.908   0.270
 6     VC      2      len       0.973   0.919
#Mind a hat csoport eloszlása normális.

#Varianciák egyenlősége A-faktor szerint
d %>%
 group_by(dose) %>%
 levene_test(len~supp)
# A tibble: 3 x 5
     dose    df1   df2 statistic       p
    <fct> <int> <int>     <dbl>   <dbl>
1    0.5      1    18      3.38  0.0825
2      1      1    18      1.67   0.212
3      2      1    18      2.51   0.130

#A varianciák a B faktor mindhárom
#szempontja szerint egyenlők.

#A B faktorra a Mauchly-próbát az ANOVA parancs
#tartalmazza.

#ANOVA-próba
model = anova_test(data = d, dv = len, wid = id,
 within = dose, between = supp)
get_anova_table(model)
ANOVA Table (type II tests)
       Effect   DFn   DFd      F        p  p<.05    ges
 1      supp     1    18 30.215 3.21e-05      *  0.224
 2      dose     2    36 74.055 1.75e-13      *  0.773
 3 supp:dose     2    36  3.306 4.80e-02      *  0.132
```

```
   #Szfericitási próba a B faktorra
   model$`Mauchly's Test for Sphericity`
          Effect       W       p
   1        dose   0.989   0.908
   2   supp:dose   0.989   0.908
#A p > 0.05, a feltétel teljesül.
```

Az ANOVA-próba eredménye szerint mind a három esetben szignifikáns eredményt kaptunk. Az interakcióra kapott eredmény határeset, ugyanis a *p*-érték 0.048. Az eta-négyzet mind a három esetben közepes, ill. nagy hatásnagyságot jelez (*ges* értékek az ANOVA Table-ben).

Post-hoc próbák

1) Az *A* csoportjaira a *B* három szintjén

```
cA = d %>% group_by(dose) %>%
 pairwise_t_test(len~supp,
 p.adjust.method = "bonferroni")
cA
#A tibble: 3 x 10
dose    .y.   group1 group2     n1     n2      p p.signif
   * <fct>   <chr>   <chr>  <chr> <int> <int>     <dbl>  <chr>
   1   0.5     len      OJ     VC     10    10    0.0053     **
   2     1     len      OJ     VC     10    10 0.000781    ***
   3     2     len      OJ     VC     10    10     0.964     ns
```

2) A *B* faktor csoportjaira

```
cB = d %>%
 group_by(supp) %>%
 pairwise_t_test(len~dose, paired = TRUE,
 p.adjust.method = "bonferroni") %>%
 select(-df, -statistic, -p) #részletek kihagyása
cB
#A tibble: 6 x 8
  supp   .y.   group1 group2     n1     n2      p.adj p.adj.sig
 <fct> <chr>   <chr>  <chr>  <int>  <int>      <dbl>     <chr>
1    OJ   len     0.5      1     10     10      0.007        **
2    OJ   len     0.5      2     10     10   0.000112       ***
3    OJ   len     1        2     10     10      0.251        ns
4    VC   len     0.5      1     10     10   0.000516       ***
5    VC   len     0.5      2     10     10  0.0000128      ****
6    VC   len     1        2     10     10      0.001        **
```

3) Az *A* faktorra a *B* faktor három szintjén (interakció)

```
effect1 = d %>%
 group_by(dose) %>%
 anova_test(dv = len, wid = id, between = supp) %>%
 get_anova_table() %>%
 adjust_pvalue(method = "bonferroni")
effect1
#A tibble: 3 x 9
 dose Effect   DFn   DFd     F       p `p<.05`    ges   p.adj
<fct> <chr> <dbl> <dbl> <dbl>   <dbl>   <chr>  <dbl>   <dbl>
1   0.5  supp     1    18  10.0   0.005     "*"  0.358   0.015
2     1  supp     1    18  16.3 0.00078     "*"  0.475 0.00234
3     2  supp     1    18 0.002   0.964      "" 0.0001       1
```

4) Post-hoc a *B* faktorra az *A* faktor két szintjén (interakció)

```
effect2 = d %>%
 group_by(supp) %>%
 anova_test(dv = len, wid = id, within = dose) %>%
 get_anova_table() %>%
 adjust_pvalue(method = "bonferroni")
effect2
#A tibble: 2 x 9
 supp Effect   DFn   DFd     F          p `p<.05`   ges    p.adj
<fct> <chr> <dbl> <dbl> <dbl>      <dbl>   <chr> <dbl>    <dbl>
1    OJ  dose     2    18  23.9   0.000008       *   0.7  0.00002
2    VC  dose     2    18  57.8 0.00000001       *  0.83 0.000003
```

A post-hoc próbák értelmezése:

1) A C-vitamin adagolása narancslé vagy tabletta formájában (*A* faktor) hatással volt a sejtek hosszára, ha a mennyiség 0.5 vagy 1 mg volt naponta. A 2 mg-os adagoknál nem számított a vitamin beviteli módja.

2) A beviteli módtól függetlenül, a napi adag befolyásolta a sejtek hosszát, egy kivétellel (1 vagy 2 mg C-vitamin adagolásánál narancslé formájában).

3) Ha 0.5, ill. 1 mg C-vitamin volt a napi adag, akkor számított, hogy narancslé vagy tabletta formájában adagolták.

4) A napi adag mindkét beviteli formánál szignifikáns hatást gyakorolt a tesztparaméterre.

6. HIPOTÉZISVIZSGÁLAT DISZKRÉT VÁLTOZÓKRA

Diszkrét változókkal általában ökológiai felméréseknél találkozunk, amikor adott faj egyedeit kell felmérni térben vagy időben, vagy egyéb környezeti mutatókat szintekre bontunk (pl. a terület kihasználási módjai, trágyázás mértéke, nedvességtartalom, árnyékos/napsütötte területek stb.). Ha diszkrét változót folytonos változóval szeretnénk összehasonlítani, akkor a folytonos változót kategóriákra lehet bontani és így diszkrét változóvá alakítani.

6.1. Mintabeli arány összehasonlítása a várt értékkel

Binomiális eloszlást követő mintánál a statisztikai próba egy mintabeli arányt (az összes vizsgált egyedből/objektumból hány teljesítette az elvárást) hasonlít össze a várt (feltételezett) értékkel. A próba nullhipotézise azt állítja, hogy a mintabeli arány egyenlő a várt aránnyal. Az ellenhipotézis három módon fogalmazható meg: nem egyenlő, kisebb vagy nagyobb. A tesztelésre alkalmas az egzakt binomiális próba, amely viszont akkor ad egzakt eredményt, ha egy végtelen populációból vagy egy véges populációból visszatevéses módszerrel mintázunk véletlen mintavétellel egymástól független egyedeket. Véges populáció esetén a próba akkor ad egzakt eredményt, ha a minta kicsi a populációhoz viszonyítva.

R-ben a próbát a **`binom.test()`** paranccsal végezhetjük el (6.1. táblázat). Egy egyszerűbb megoldás a z-próba, amit olyan feltételek mellett (pl. $n \cdot p \geq 10$ és $p \geq 0.1$ vagy $n \cdot p \geq 15$) lehet alkalmazni, amelyek lehetővé teszik a binomiális eloszlás megközelítését a normális eloszlással (lásd a 3.2.3. alfejezetet). A próbastatisztika a z^2, ami χ^2-eloszlást követ $df = 1$ szabadságfokkal (Reiczigel et al., 2007).

$$z^2 = \frac{n \cdot (\bar{p} - p_0)^2}{p_0 \cdot (1 - p_0)},$$

ahol n a próbálkozások (egyedek) száma, $\bar{p}$ a mintára kapott gyakoriság, p_0 pedig a várt (elméleti) gyakoriság. A z-próbát a **`prop.test()`** paranccsal hajthatjuk végre.

A Poisson-eloszlást követő adatokra a Poisson-próba végezhető el a **`poisson.test()`** paranccsal. Az egymintás próba tulajdonképpen egy binomiális próba nagy mintaelemszám esetén. A Poisson-eloszlás *adott eseményből hány következik be* típusú adatok eloszlását írja le, ahol az események összszámát a tér- vagy időegységek összessége jelenti (lásd a 3.2.4. alfejezetet). Például fél óra megfigyelési idő alatt hány egerészölyv jelenik meg a látómezőben. Ezért az R-parancs x, T, r paramétereivel a sikerek számát, az időt és az előfordulási arányt adhatjuk meg.

6.1. táblázat. *A mintabeli arány összehasonlítása a várt értékkel*

Parancs	Részletek	Magyarázat
binom.test(x, n, p, alternative, conf.level) prop.test(x, n, p, alternative, conf.level, correct) poisson.test(x, T, r, alternative, conf.level) ES.h(p1, p0) {pwr}		
x	Integer	sikerek száma, vagy egy vektor, ami a sikerek és a kudarcok számát tartalmazza
n	Integer	próbálkozások száma
p, p0	[0,1]	feltételezett arány
T	time	az idő, ami alatt bekövetkezett az x esemény
r	ratio	Értéke alapból 1. A várt x/T arány.
alternative	"two.sided", "less", "greater"	a próba típusa
conf.level	0.95	alapból $1-\alpha = 0.95$
correct	TRUE/FALSE	Yates-féle korrekciót végez. Ezt ki kell zárni (FALSE).

Hatásnagyság

A hatásnagyság meghatározására nem alkalmas a Cohen-féle *d*, mivel az folytonos változókra működik. Diszkrét változókra Cohen-féle *h* használható, aminek képlete:

$h = 2 \cdot (arcsin\sqrt{p_1} - arcsin\sqrt{p_0})$, *mintabeli arány,* p_0 pedig a várt arány. A hatásnagyság R-ben az `ES.h()` paranccsal végezhető el a {pwr} csomagból (6.1. táblázat).

Példa binominális eloszlást követő változóra

A Nemzeti Statisztikai Intézet adatai szerint Románia összlakossága 2020-ban 22 142 153 volt, amelynek férfi : nő aránya 48.8 : 51.2 (6.2. táblázat). Egy felméréshez 6520 személyt választottak ki véletlenszerűen, amelyből 3243 férfi és 3277 nő volt. Eltér-e szignifikánsan a férfiak aránya a hivatalos 48.8%-tól? Ebben az esetben úgy a binomiális próba, mint a *z*-próba is használható, mivel egy véges populációból kis mintát vettünk, ugyanakkor az $n \cdot p = 3260 > 10$.

6.2. táblázat. *Románia népességének megoszlása nemek szerint (INS, 2021)*

Románia	Összlakosság	Férfi (egyed)	Nő (egyed)
2020	22 142 153	10 813 692	11 328 461
%	100	48.8	51.2

A próba nullhipotézise azt állítja, hogy a férfiak aránya nem tér el a $p = 0.488$ értéktől. A próbát elvégezzük a binomiális eloszlás és a normális eloszlás alapján is.

```
binom.test(x = 3243, n = 6520, p = 0.488)
Exact binomial test
data: 3243 and 6520
number of successes = 3243, number of trials = 6520,
p-value = 0.1307
alternative hypothesis: true probability of success is not
equal to 0.488
95 percent confidence interval:
 0.4851830 0.5096046
sample estimates:
probability of success
 0.4973926

prop.test(x = 3243, n = 6520, p = 0.488, correct = F)
1-sample proportions test
 without continuity correction
data: 3243 out of 6520, null probability 0.488
X-squared = 2.3021, df = 1, p-value = 0.1292
alternative hypothesis: true p is not equal to 0.488
95 percent confidence interval:
 0.4852614 0.5095270
sample estimates:
 p
0.4973926
```

A két próba hasonló eredményt adott. Mind a két esetben a mintában tapasztalt arány (0.497) nem tér el szignifikánsan a Nemzeti Statisztikai Intézet által közzétett értéktől.

Példa Poisson-eloszlást követő változóra

A Brandl et al. (2020) által közölt cikkben a fafajok pusztulási gyakoriságával kapcsolatban azt találták, hogy 65416 lucfenyőből 5098 elpusztult. Ez 78 fenyőt jelent 1000-ből, az arány tehát 0.078. Teszteljük azt a hipotézist, hogy a kapott arány eltér-e szignifikánsan a 0.08 értéktől.

```
poisson.test(5098, 65416, 0.08)
Exact Poisson test
data: 5098 time base: 65416
```

```
number of events = 5098, time base = 65416,
p-value = 0.06201
alternative hypothesis:
true event rate is not equal to 0.08
95 percent confidence interval:
 0.07580726 0.08010121
sample estimates:
event rate
 0.077932
```

Az egzakt Poisson-próba nem adott szignifikáns eltérést ($p = 0.062$) a várt érték és a mintaátlag között.

6.2. Két minta függetlenségének vizsgálata

Egy diszkrét valószínűségi változó két mintabeli arányát hasonlítjuk össze, ha a minták egymástól függetlenek. A kétmintás próbák kiválasztásának szempontjai a változó típusa, illetve a minták elemszámai, amint azt az előző alfejezetben tárgyaltuk. A használható próbák a 6.4. táblázatban találhatók. A hatásnagyság meghatározására a Cohen-féle h használható, aminek képlete:

$h=2\cdot(arcsin\sqrt{p_1}-arcsin\sqrt{p_0})$, p_2 a két mintabeli arány. R-ben az **`ES.h()`** paranccsal hajtható végre (6.4. táblázat).

A kategóriákat tartalmazó változók függetlenségét a χ^2-próbával, illetve a Fisher-féle egzakt próbával teszteljük (lásd a 7.5.1. alfejezetet).

A binomiális próba megközelíthető a normális eloszlással, ha teljesülnek az $n\cdot p \geq 10$ és $p \geq 0.1$ vagy $n\cdot p \geq 15$ feltételek. A z-értéket a következő képlettel értelmezzük (Reiczigel et al., 2007):

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{p_p(1-p_p)}{n_1} + \frac{p_p(1-p_p)}{n_2}}}, p_p = \frac{f_1 + f_2}{n_1 + n_2},$$

ahol n_1, n_2 az elemszámok, f_1, f_2 a sikerek száma a két mintában. A normális megközelítésen alapuló próbát a **`prop.test()`** paranccsal végezhetjük el (6.1. táblázat).

A Poisson-próba elvégezhető a **`poisson.test()`** és a **`rateratio.test()`** paranccsal, utóbbi a {rateratio.test} csomagban található (6.4. táblázat).

Két egymástól függő minta arányának egyenlőségét a McNemar-próbával tesztelhetjük. Párosított minta alatt olyan helyzetet értünk, amikor kontroll/teszt párokkal dolgoznak, és egy tünet/elváltozás/betegség megjelenését vizsgálják. A vizsgált paraméter dichotóm kell legyen (a lehetséges két szintje egymást ki kell zárja). A kontroll/teszt párban ugyanazt azt egyedet vizsgáljuk két különböző

időpillanatban: az első a kontroll állapot, a második pedig a hatásnak való kitettség utáni állapot. Az eredményeket a 6.3. táblázat szerint adják meg. A McNemar-próba χ^2-eloszlás alapján ($df = 1$) értékeli ki a próbastatisztikát.

6.3. táblázat. *A kutatás eredménye párosított mintákra a McNemar-próbához*

		A		Összesen
		+	–	
B	+	a	b	
	–	c	d	
Összesen				

A próbastatisztika:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}.$$

A b és c értékek a két változó közötti eltéréseket jelölik (diszkordáns értékek). A nevezőben szereplő korrekciós tag (−1) a folytonossági korrekciót képezi. Abban az esetben, ha a $b + c < 10$, akkor a McNemar-próba egzakt változatát javasolják használni (Rufibach, 2011).

A hatásnagyságot a Cohen-féle g értékkel határozhatjuk meg. Ezt a következő képlettel számolhatjuk ki:

$$g = \frac{b}{b + c} - 0.5.$$

A Cohen-féle g értéke 0.05, 0.15, ill. 0.25 alatt nagyon kicsi, kicsi, közepes jelentőségű, 0.25 fölött pedig nagy hatásnagyságot jelent. Az R-ben a McNemar-próbát a **`mcnemar.test()`** és az **`exact2x2()`** paranccsal végezhetjük el. A Cohen-féle g-t a **`cohenG()`** parancs segítségével számoljuk ki. Az adatokat a próbákba mátrix formájában kell bevezetni (6.4. táblázat).

6.4. táblázat. *Két minta arányainak az összehasonlítása R-ben*

Parancs	Részletek	Magyarázat
rateratio.test(c(s1,s2), c(T1, T2), r, alternative, conf.level) {rateratio.test} mcnemar.test(matrix, correct) exact2x2(matrix, alternative, conf.level, paired) {exact2x2} ES.h(p1, p2) {pwr} cohenG(matrix) {rcompanion}		
s_1, s_2	Integer	A sikerek száma a két mintában.
T_1, T_2	Integer	Próbálkozások száma.
p_1, p_2	Numeric	A sikerek aránya a két mintában.

Parancs	Részletek	Magyarázat
matrix	matrix	Mátrixba kell rendezni a kereszttábla adatait.
r	ratio	Értéke alapból 1.
alternative	"two.sided", "less", "greater"	A próba típusa.
conf.level	0.95	Alapból $1-\alpha = 0.95$.
paired	FALSE/TRUE	TRUE (McNemar exact test)
correct	TRUE/FALSE	Yates-féle korrekciót végez, ha 2x2-es a táblázat.

Példa binomiális eloszlást követő adatokra

Eidenbenz et al. (2021) vizsgálták a lavina túlélés esélyeit 1997−2018 között Svájcban, abban az esetben, ha a hó alá temetődés időtartama minimum 60 perc és maximum 24 óráig tartott. Az adatok alapján keressük a választ arra a kérdésre, hogy nagyobb eséllyel élik-e túl a férfiak a balesetet a nőknél. Összesen 138 esetet vizsgáltak, amiből 113 férfi volt. Összesen 25-en maradtak életben, ebből 19 férfi és 6 nő. A 6.5. táblázatban látható a kontingenciatáblázat (kereszttábla).

6.5. táblázat. *A lavinatúlélés esélyei Svácban (1997–2018) (Eidenbenz, 2021)*

Lavina	Túlélő	Halott	Összesen
Nő	6	19	25
Férfi	19	94	113
Összesen	25	113	138

A nők túlélési aránya tehát 6/25, a férfiaké pedig 19/113. Mivel csupán két, egymást kizáró eseményről beszélünk (életben marad, meghal), a *2x2*-es táblázatra a Fisher-féle egzakt próbát végezzük el, vagy az arányok tesztelését a **`prop.test()`** paranccsal.

```
#Elkészítjük a mátrixot
a = matrix(c(6,19,19,94), nrow = 2)
a
     [,1] [,2]
[1,]    6   19
[2,]   19   94

#Fisher-próba
fisher.test(a, alternative = "two.sided")
Fisher's Exact Test for Count Data
data: a
p-value = 0.3981
```

```
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.4487829 4.7874246
sample estimates:
odds ratio
 1.556819

#Megközelítés a normális eloszlással
prop.test(a, correct = F)
2-sample test for equality of proportions
 without continuity correction
data: a
X-squared = 0.71258, df = 1, p-value = 0.3986
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1092001  0.2529169
sample estimates:
   prop 1    prop 2
0.2400000 0.1681416
```

A két eredmény hasonló (p = 0.398, ill. p = 0.399), egyik szerint sincs szignifikáns különbség a nők (24%) és a férfiak (17%) túlélési arányai között.

```
#Hatásnagyság
library(pwr)
ES.h(0.24,0.17)
[1] 0.1739678
```

A hatásnagyság < 0.2, így kicsinek minősül.

Példa Poisson-eloszlást követő adatokra

A Brandl et al. (2020) által közölt cikk alapján vizsgáljuk a bükk és a tölgy túlélési esélyei közti különbséget: 23 648 bükkből 1004 pusztult el, 14 412 tölgyből pedig 699. A két arány közti különbséget a **`poisson.test()`**-tel és a **`rateratio.test()`**-tel vizsgálhatjuk.

```
poisson.test(c(1004,699),c(23648,14412))
Comparison of Poisson rates
data: c(1004, 699) time base: c(23648, 14412)
count1 = 1004, expected count1 = 1058.1,
p-value = 0.007504
alternative hypothesis: true rate ratio is not equal to 1
95 percent confidence interval:
```

```
 0.7939999 0.9654731
sample estimates:
rate ratio
 0.8753593

library(rateratio.test)
rateratio.test(c(1004,699),c(23648,14412))
Exact Rate Ratio Test, assuming Poisson counts
data: c(1004, 699) with time of c(23648, 14412),
null rate ratio 1
p-value = 0.007634
alternative hypothesis: true rate ratio is not equal to 1
95 percent confidence interval:
 0.7939999 0.9654731
sample estimates:
Rate Ratio     Rate 1     Rate 2
0.87535935 0.04245602 0.04850125

#Hatásnagyság
library(pwr)
ES.h(0.794,0.965)
[1] -0.5658288
```

A próbák szignifikáns eredményt adtak ($p = 0.008$), tehát azt mondhatjuk, hogy nem egyforma a bükk és a tölgy pusztulási aránya. A hatásnagyság -0.56, ami közepes hatást jelez.

Példa párosított mintákra

Legyen egy példa a 6.6. táblázatban kapott eredményekkel (A: teszt, B: betegség). Összesen 92 emberen teszteltek egy vérből kimutatható markert egy adott betegségre a betegség alatt és a felgyógyulás után két héttel. A kérdés, hogy a teszt alkalmas-e a betegség akut fázisának kimutatására. A nullhipotézis: a marker egyformán kimutatható a betegség akut fázisában és közvetlenül a felgyógyulás után is.

6.6. táblázat. *A teszt és a betegség közötti kapcsolat*

		A (teszt)		Összesen
	+		–	
B (betegség)	+	16	7	23
	–	3	66	69
Összesen		19	73	92

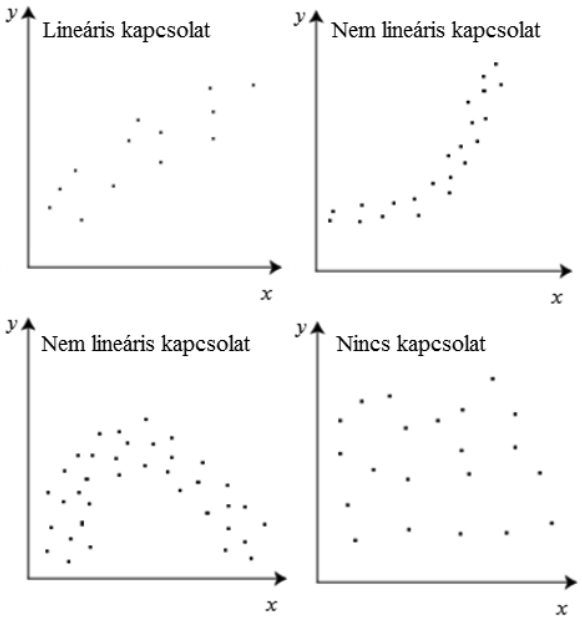
A 6.6. táblázat szerint hét esetben a teszt nem mutatta ki a marker jelenlétét beteg embereknél, és három esetben mutatta ki a markert a betegségből felgyógyult embereknél. Mivel a két diszkordáns érték összege 10, ezért érdemes az egzakt McNemar-próbával tesztelni a hipotézist.

```
#A mátrix elkészítése
b = matrix(c(16,3,7,66), nrow = 2)
b
     [,1] [,2]
[1,] 16    7
[2,]  3   66

library(exact2x2)
exact2x2(b)
exact2x2(b, paired = T)
Exact McNemar test
(with central confidence intervals)
data: b
b = 7, c = 3, p-value = 0.3438
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.5326778 13.9836279
sample estimates:
odds ratio
 2.333333

#Hatásnagyság
library(rcompanion)
cohenG(b)
$Global.statistics
  Dimensions    OR    P   g
1      2 x 2  2.33  0.7 0.2

#Hatásnagyság kiértékelése
library(effectsize)
interpret_cohens_g(0.2)
[1] "medium"
(Rules: cohen1988)
```

Az egzakt McNemar-próba nem adott szignifikáns eredményt ($p = 0.344$), noha a hatásnagyság közepesnek mutatkozik ($g = 0.2$). Így azt állíthatjuk, hogy vagy a teszt nem alkalmas a betegség akut fázisának kimutatására, vagy pedig növelni kell a tesztalanyok számát.

7. KÉT VÁLTOZÓ KÖZÖTTI KAPCSOLAT VIZSGÁLATA

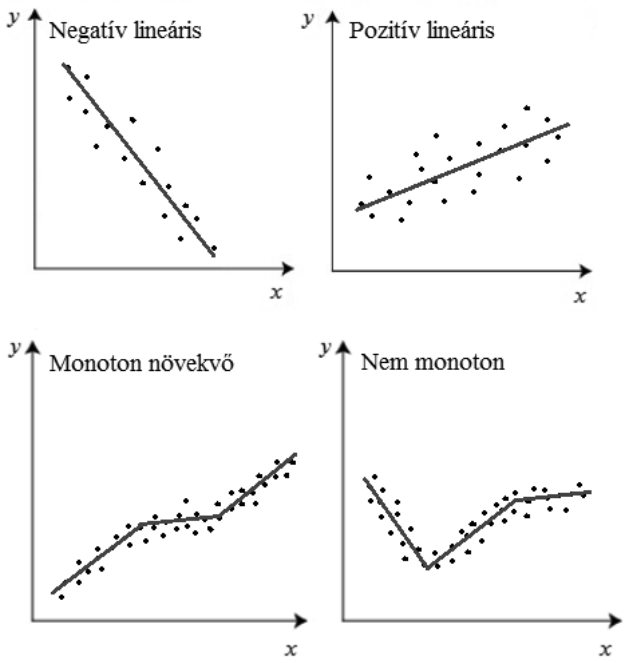
Korrelációvizsgálattal két változó közötti összefüggés létezését és irányát vizsgáljuk. A változók értékeiből számpárokat képezünk. A 7.1. ábrán két folytonos változó közötti lehetséges kapcsolatok láthatók: lineáris kapcsolat (a kapcsolat egy egyenessel jellemezhető), nem lineáris kapcsolat (többféle görbével közelíthető meg, amit például vagy négyzetes tagú, vagy más exponenciális függvény határoz meg) és az az eset, amikor nincs kapcsolat a két változó között (*x* és *y* egymástól független).



7.1 ábra. *Két folytonos x és y változó közötti lineáris, nem lineráis kapcsolatok, illetve a kapcsolat hiánya*

Két változó között akkor beszélünk lineáris kapcsolatról, ha a pontokhoz egyenes illeszthető. A lineáris korreláció lehet pozitív (ha *x* értéke nő, *y* értéke is nő), és lehet negatív (ha *x* értéke nő, *y* értéke csökken) (7.2. ábra). Előfordulhat, hogy van kapcsolat a két változó között, de a pontok egyetlen egyenes helyett több lineáris szakasszal közelíthetők meg. Ha ezekben a szakaszokban végig pozitív

kapcsolat van a két változó között, akkor monoton növekedésről beszélünk. Ha a kapcsolat szakaszonént hol pozitív, hol negatív, akkor nem beszélünk monoton kapcsolatról (7.2. ábra).



7.2. ábra. *Lineáris és monoton kapcsolatok két folytonos változó között*

Korrelációvizsgálattal ok-okozati viszony vizsgálható. Ebben az értelemben megkülönböztetünk egy független (magyarázó) változót (x) és egy függő változót (y). Az x meghatározza az y értékét.

Folytonos változók közötti összefüggés jellemzésére a Pearson-, Spearman- és Kendall-féle korrelációs tényezők alkalmasak, ordinális változókra a Kendall-féle tan, a Somers-féle delta vagy a Cramer-féle V használható.

7.1. Pearson-féle korrelációs együttható

Folytonos változók esetén egy lineáris kapcsolat erősségét a Pearson-féle korrelációs együtthatóval (R_{xy} vagy r_{xy}) szokták mérni. r_{xy} értéke $[-1,1]$ közötti értéket vehet fel. A 7.1. táblázat tartalmazza az R kiértékelésének kritériumait.

7.1. táblázat. *A Pearson-féle korrelációs együttható értkéinek jelentősége*

Az összefüggés mértéke	R_{xy}^{*}	
	pozitív	negatív
kicsi	0.1 − 0.3	−0.1 − −0.3
közepes	0.3 − 0.5	−0.3 − −0.5
nagy	0.5 − 1.0	−0.5 − −1.0

*Az értékek csak tájékoztató jellegűek. Vannak esetek, amikor újra kell értékelni őket.

Abban az esetben, ha az r_{xy} értéke 1, x és y között függvényszerű kapcsolat van, amit az egyenes egyenletével jelemezhetünk: $y = ax + b$, ahol a az egyenes iránytangense (meredeksége), b pedig a szabadtag (az y tengellyel való metszéspont). Az r_{xy} előjele megegyezik a előjelével.

A Pearson-féle együttható nem függ a két változó értékkészletének sem a mértékegységétől, sem a nagyságrendjétől, és a változók sorrendjétől sem. Az r_{xy} értelmezése:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

ahol n az x és y objektumainak a száma.

Mivel x és y valószínűségi változók, az átlagok is azok, így előfordulhat, hogy annak ellenére, hogy a két változó független egymástól, mégis az r_{xy} értéke nem 0. Ezért a két változó függetlenségének tesztelését hipotézisvizsgálattal végezzük el. A próbastatisztika a $df = n-2$ szabadságfokú Student-féle t-eloszlás t paramétere, amit a következő képlettel számolunk ki:

$$t = r_{xy} \cdot \sqrt{\frac{n-2}{1-R_{xy}^2}}.$$

A próba nullhipotézise: az x és y egymástól független ($r_{xy} = 0$), ellenhipotézise pedig, hogy x és y nem független egymástól ($r_{xy} \neq 0$). Az eredményt a következő formában ajánlott közölni: $r(df)$ = érték, p-érték. R-ben az alapcsomagban levő `cor.test()` paranccsal végezhetjük el (7.2 táblázat).

7.2. táblázat. *A korrelációs próba R-ben*

Parancs	**Részletek**	**Magyarázat**
cor(x, y, method) *#az együtthatót számolja ki* cor.test(x, y, method, conf.level, continuity) #elvégzi a korrelációs próbát		
x, y	Numeric	két folytonos változó
method	"pearson", "kendall", "spearman"	korrelációs együttható típusa alapból a Pearson-félét számolja
conf.level	0.95	szignifikanciaszint (alapból $\alpha = 0.05$)
continuity	FALSE/TRUE	Spearman- és Kendall-együttható esetén, ha vannak csatolt rangok, értéke FALSE

Az *iris* adattáblában a csésze- és sziromlevelek hossza közötti kapcsolat vizsgálata:

```
d = iris
cor.test(d$Sepal.Length, d$Petal.Length,
 method = "pearson")
Pearson's product-moment correlation
data: d$Sepal.Length and d$Petal.Length
t = 21.646, df = 148, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8270363 0.9055080
sample estimates:
 cor
0.8717538
```

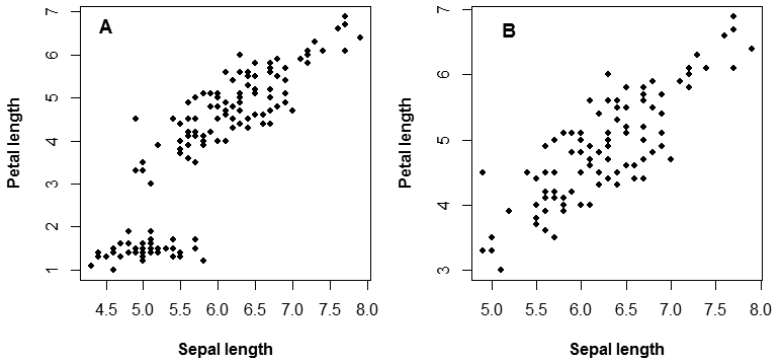
A korrelációs próba eredménye szignifikáns ($t = 21.6$, $df = 148$, $p < 0.001$), így a két változó között kapcsolat van, az r_{xy} értéke 0.872, ami erős kapcsolatora utal.

7.1.1. A Pearson-féle korrelációs együttható alkalmazhatósági feltételei

A Pearson-féle korrelációs együttható a lineáris kapcsolat erősségét méri, tehát első feltétel az, hogy a két változó közötti kapcsolat lineáris legyen.

1) Lineáris kapcsolat vizuális ellenőrzése

A kapcsolat lineáris voltát egy *xy*-ábra segítségével vizuálisan ellenőrizni lehet (7.3. ábra). Amennyiben láthatóan a kapcsolat nem lineáris, értelmetlen ezt az együtthatót kiszámolni.



7.3. ába. *A nősziromfajok csésze- és sziromlevél hosszai közötti kapcsolat: A – három faj; B – két faj (forrás: iris adattábla R-ben)*

A 7.3.A. ábrán látható, hogy a bal alsó részen elkülönül egy pontcsoport. Ez az *Iris setosa* faj adatait mutatja. Mivel lineáris összefüggést keresünk, ezért ezt a fajt eltávolítjuk az adattáblából (B. ábra), marad a másik két faj, amelynél láthatóan lineáris az összefüggés. A 7.3. ábra R-kódja:

```
d = iris
#Eltávolítjuk a "setosa" adatokat.
d1 = subset(d, Species != "setosa")
#Képezünk egy összetett ábrát, és beállítjuk a margókat.
par(mfrow=c(1,2), mar=c(4.2,4.2,1,1))
#Elkészítjük a szórásdiagramokat.
plot(d$Sepal.Length, d$Petal.Length,
 xlab = "Sepal length",
 ylab = "Petal length", font.lab = 2,
 pch = 16, cex = 0.9)
text(4.5,6.7,"A", font = 2, cex = 1.2)
plot(d1$Sepal.Length, d1$Petal.Length,
 xlab = "Sepal length",
 ylab = "Petal length", font.lab = 2,
 pch = 16, cex = 0.9)
text(5.2,6.7,"B", font = 2, cex = 1.2)
```

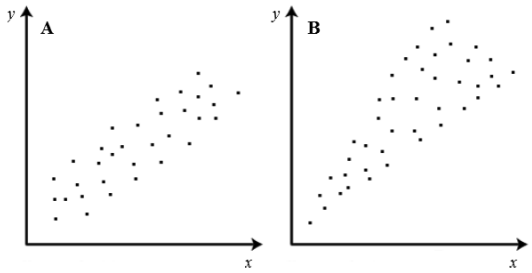
2) Normális eloszlás

A Pearson-féle korrelációs együttható érvényességének feltétele a kétváltozós normális eloszlás. Ezt nehéz ellenőrizni, ezért elfogadható a két változó normális eloszlása, külön-külön. A megszokott módon lehet ezt ellenőrizni egy hisztogramon vagy QQ-ábrán, illetve tesztelni Shapiro–Wilk- vagy Kolmogorov–Smirnov-próbával (lásd a 4.2. alfejezetet).

3) Homoszkedaszcitiás

A homoszkedaszticitás azt jelenti, hogy a variancia a két változó szórásdiagramján azonos szóródást mutat a diagram bármely szakaszában (az illesztett egyenes körül a pontok szóródása állandó) (7.4. ábra). A két nősziromfaj szórásdiagramjának (7.3.B. ábra) a középső szakaszán nagyobb a szóródás, mint a széleken.

R-ben a nem állandó variancia tesztelését az `ncvTest()` segítségével lehet elvégezni a {car} csomag segítségével, vagy a Breusch–Pagan-próbával a {lmtest} csomagból. A nullhipotézis azt állítja, hogy a variancia állandó (a homoszkedaszticitás fennáll) (7.3. táblázat).



7.4. ábra. *Szóródás a szórásdiagramon: A – homoszkedasztikus (egyenletes); B – heteroszkedasztikus (a szóródás változó)*

7.3. táblázat. *A homoszkedaszticitás és többváltozós kiugró értékek tesztelése R-ben*

Parancs	Részletek	Magyarázat
ncvTest(lm(x~y)) {car} bptest(lm(x~y)) {lmtest} outlierTest(lm(x~y))		
x, y	Numeric	két folytonos változó

A két változó homoszkedaszticitásának tesztelése R-ben:

```
library(car)
ncvTest(lm(d1$Sepal.Length~d1$Petal.Length))
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.04710648, Df = 1, p = 0.82818
```

Az NCV-próba alapján a szóródás állandónak tekinthető ($\chi^2 = 0.047$, $df = 1$, $p = 0.828$).

4) Kiugró értékek

A Pearson-féle korrelációs együttható értékét jelentősen befolyásolja az egy- és kétváltozós kiugró érték(ek) jelenléte. Az egyváltozós kiugró érték alatt azt értjük, hogy egy adatsorban létezik olyan érték, ami nem tartozik bele a $\pm 3s$ tartományba, ahol s a minta szórása). Nem normális eloszlások esetén pedig nem tartozik bele a ± 1.5IQR tartományba, ahol az IQR az interkvartilis tartományt jelenti. A kiugró értékeket könnyen lehet azonosítani a dobozdiagramon.

A kétváltozós (ill. sokváltozós) kiugró érték azt jelenti, hogy két (több) változó szempontjából is kiugró értéknek minősül.

Ezt a feltételt nem lehet dobozdiagramon tesztelni, noha az egyedi dobozdiagramokon azonosított kiugró értékek lehetnek többváltozós kiugró értékek is. R-ben a {car} csomagban levő Bonferroni-féle próbával ellenőrizni lehet a jelenlétüket (7.3. táblázat).

```
#Kétváltozós kiugró értékek tesztelése
library(car)
outlierTest(lm(d1$Sepal.Length~d1$Petal.Length))
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
 rstudent unadjusted p-value Bonferroni p
57 -3.069037 0.0027847 0.27847
```

A Bonferroni-féle próba szerint nincsenek kétváltozós kiugró értékek ($t = -3.07$, *p-korr* = 0.278), tehát ki lehet számolni a Pearson-féle korrelációs együtthatót.

```
#A Pearson-féle korrelációs együttható
cor(d1$Sepal.Length,d1$Petal.Length)
[1] 0.8284787
#A korrelációs próba
cor.test(d1$Sepal.Length,d1$Petal.Length)
Pearson's product-moment correlation
data: d1$Sepal.Length and d1$Petal.Length
t = 14.645, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7549049 0.8814586
sample estimates:
 cor
0.8284787
```

A kapcsolat a két nősziromfaj csésze- és sziromlevél hossza között nagy ($r = 0.828$) és szignifikáns ($t = 14.645$, $df = 98$, $p < 0.001$). Rövid formában: a korreláció $r(98) = 0.28$, $p < 0.001$.

7.2. Spearman-féle rangkorreláció

A Spearman-féle rangkorrelációs együttható ordinális változók közti kapcsolat megállapítására alkalmas mérőszám, ugyanakkor egy nem parametrikus változata a Pearson-féle korrelációs együtthatónak. Az utóbbival ellentétben a

Spearman-féle együttható két változó közti monoton kapcsolat szorosságát és irányát vizsgálja, és nem egy lineáris kapcsolatot értékel ki, így kevésbé korlátozott, mint a Pearson-féle együttható. Alkalmazásának feltétele, hogy a rangok közti távolságok egyenlők legyenek, túl sok kapcsolt rang és kis mintaelemszám esetén nem javasolt a használata. Az ilyen helyzetben a kapcsolat vizsgálatára a Kendall-féle *tau*-t érdemes használni.

Az együtthatót r_s vagy ρ (rho) görög betűvel jelöljük, értékei pedig −1 és +1 között változhat. A +1 és −1 erős monoton növekvő, ill. csökkenő kapcsolatot mutatnak, a 0 érték pedig a két változó monoton kapcsolatának hiányát jelenti.

A statisztikai próba nullhipotézise azt állítja, hogy a két változó között nincs monoton kapcsolat. R-ben a próbát a `cor.test()` paranccsal végezhetjük el (8.2. táblázat).

A Spearman-féle együttható kiszámításához mind a két változó összes értékét külön-külön rangsoroljuk, a legnagyobb érték kapja az 1-es rangot, a legkisebb pedig a legnagyobb rangot. A rangpárokból különbségeket számolunk (d_i), majd ennek segítségével kiszámoljuk az r_s értéket. Ha nincsenek csatolt rangok, akkor a számítás viszonylag egyszerű:

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i}{n(n^2 - 1)},$$

ahol n a két változó elemszáma. Ha csatolt rangok vannak jelen, akkor a számítás a Pearson-féle r értelmezésével analóg, azzal a különbséggel, hogy a változókra kapott értékek helyett a rangokat használjuk:

$$r_s = \frac{\frac{1}{n}\sum_{i=1}^{n}(r_{xi} - \bar{r}_x) \cdot (r_{yi} - \bar{r}_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{xi} - \bar{r}_x)^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_{yi} - \bar{r}_y)^2}},$$

ahol r_{xi} és r_{yi} az x és y változók rangjai, $\bar{r}_x$ és $\bar{r}_y$ az x és y változók rangjainak az átlagértékei.

A próbastatisztikát (t-érték) a következő képlettel számoljuk ki:

$$t = \frac{r_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}}.$$

A Spearman-féle rangkorrelációs együtthatót a következő formában közöljük: $r_s(df)$ = érték, p = érték, ahol a $df = n - 2$.

Példa

A 7.4. táblázat tartalmazza az *airquality* adattábla augusztusi adatait az ózonkoncentrációra és a napsugárzásra, valamint a rangokat. Az adattábla nem tartalmazza azokat a napokat, amelyekre nem volt mérési adat. Mivel az ózonkoncent-

rációnál vannak csatolt rangok, ezért a bonyolultabb összefüggést használjuk a Spearman-féle rangkorreláció kiszámításához.

7.4. táblázat. *Az* airquality *adattábla adatai az ózonkoncentrációra és a napsugárzásra, valamint az értékekhez tartozó rangok*

Nap	Ózon (ppb)	Rang (ózon)	Napsugárzás (lang)	Rang (napsugárzás)
1	39	15	83	18
2	9	22.5	24	23
3	16	21	77	19
7	122	2	255	3
8	89	5	229	7
9	110	4	207	11
12	44	13.5	192	13
13	28	17	273	1
14	65	10	157	16
16	22	19	71	20
17	59	11	51	21
18	23	18	115	17
19	31	16	244	4
20	44	13.5	190	14
21	21	20	259	2
22	9	22.5	36	22
24	45	12	212	10
25	168	1	238	5
26	73	9	215	9
28	76	8	203	12
29	118	3	225	8
30	84	7	237	6
31	85	6	188	15
		12		12

$$r_s = \frac{20.7}{6.63 \cdot 6.63} = 0.47$$

A próbastatisztika *t*-értéke:

$$t = \frac{0.47}{\sqrt{\frac{1 - 0.47^2}{23 - 2}}} = 2.44.$$

A kritikus *t*-érték 2.08 ($df = 21$ és $\alpha = 0.05$). Mivel a $t > t_{krit}$, ezért a két változó között szignifikáns kapcsolatot, korrelációt feltételezhetünk. R-ben a Spearman-féle rangkorreláció az alábbi kóddal számolható ki, ill. tesztelhető a szignifikáns kapcsolat érvényessége is.

```
d = airquality
#Kiválasztjuk az augusztusi hónap adatait
a = subset(d[,c(1,2,5,6)], Month == 8)
head(a)
   Ozone Solar.R Month Day
93    39      83     8   1
94     9      24     8   2
95    16      77     8   3
96    78      NA     8   4
97    35      NA     8   5
98    66      NA     8   6

#Eltávolítjuk az NA-kat.
a = na.omit(a)
head(a)
    Ozone Solar.R Month Day
93     39      83     8   1
94      9      24     8   2
95     16      77     8   3
99    122     255     8   7
100    89     229     8   8
101   110     207     8   9

#Kiszámoljuk a Spearman-féle rho-t.
cor(a$Ozone, a$Solar.R, method = "spearman")
[1] 0.4695997

#A szignifikáns korreláció tesztelése.
cor.test(a$Ozone,a$Solar.R, method = "spearman")
Spearman's rank correlation rho
data: a$Ozone and a$Solar.R
S = 1073.5, p-value = 0.02377
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
 rho
0.4695997
```

Az ózonkoncentráció és a napsugárzás között szignifikáns pozitív korreláció van, $r_s(21) = 0.47$, $p = 0.024$. Ez azt jelenti, hogy ha erősebb a napsugárzás, akkor több ózon képződik New York légkörében.

7.3. Kendall-féle tau

A Kendall-féle tau (τ) monoton kapcsolatot vizsgál két ordinális vagy numerikus skálán mozgó változó között (Kendall, 1938). Nem feltétel a rangok közti azonos távolság, ezért akkor is lehet használni, amikor a Spearman-féle együttható nem megbízható. A Pearson-féle együttható nem parametrikus változataként is használható, ha nem teljesülnek annak feltételei. Jól működik kis mintaelemszám és kapcsolt rangok esetén is.

Értékei -1 és $+1$ között változnak. Ideális monoton növekvő kapcsolat esetén értéke $+1$, ideális monoton csökkenő kapcsolat esetén pedig -1. A nulla érték a kapcsolat hiányát jelenti.

A Spearman-féle módszerhez hasonlóan az értékekhez rangokat társítunk, általában a legkisebb értékhez társítjuk az 1-es rangot. A két változó közül az egyiket „lehorgonyozzuk”, rangjait etalonnak használjuk, a másik változó rangjait pedig ehhez viszonyítjuk (referencia változó). Ideális monoton növekvő kapcsolat esetén az első változó növekvő rangjaihoz automatikusan a második változó növekvő rangjai társulnak. Ideális monoton csökkenő kapcsolat esetén a második változó rangjai csökkenő sorrendben társulnak az elsőhöz. E két extrém helyzettől való eltérés a monotonitástól való eltérést jelenti, illetve a kapcsolat gyengülését a két változó között.

A Kendall-féle tau kiszámítását az airquality adattábla két változójára számoljuk ki (7.5. táblázat).

7.5. táblázat. *Az* airquality *adattábla két változójára megállapított rangok (az első változó rangjai növekvő sorrendbe rendezve)*

Nap	Rang ($_X$)	Rang (Y)
25	1	5
7	2	3
29	3	8

Nap	Rang ($_X$)	Rang (Y)
9	4	11
8	5	7
31	6	15
30	7	6
28	8	12
26	9	9
14	10	16
17	11	21
24	12	10
12	13.5	13
20	13.5	14
1	15	18
19	16	4
13	17	1
18	18	17
16	19	20
21	20	2
3	21	19
2	22.5	23
22	22.5	22

A Kendall-féle tau kiszámítása hosszadalmas feladat, ezért a számítási módszert az első öt adat alapján mutatjuk be. Az első változó rangjai szerinti rendezés után a második változó rangjait páronként összehasonlítjuk, és megszámoljuk a pozitív előjelű különbségeket (P). A maradék párok összegének (negatív előjelű különbségek) jele az I.

$R_{y1} - R_{y2} \rightarrow +$ $R_{y2} - R_{y3} \rightarrow -$ $R_{y3} - R_{y4} \rightarrow -$ $R_{y4} - R_{y5} \rightarrow +$
$R_{y1} - R_{y3} \rightarrow -$ $R_{y2} - R_{y4} \rightarrow -$ $R_{y3} - R_{y5} \rightarrow +$
$R_{y1} - R_{y4} \rightarrow -$ $R_{y2} - R_{y5} \rightarrow -$
$R_{y1} - R_{y5} \rightarrow -$

A különbségek alapján $P = 3$, $I = 7$.

A Kendall-féle tau értelmezése, ha nincsenek kapcsolt rangok az adatsorokban (τ_a):

$$\tau_a = \frac{P - I}{n \cdot \frac{n-1}{2}}$$

$$\tau_a = \frac{3 - 7}{5 \cdot \frac{5-1}{2}} = -0.4$$

Az öt adat alapján a Kendall-féle tau értéke -0.4.

Abban az esetben, ha vannak kapcsolt rangok valamely adatsorban, a Kendall-féle tau b (τ_b) jellemzi a kapcsolat erősségét:

$$\tau_b = \frac{P - I}{\sqrt{\left(\frac{n \cdot (n-1)}{2} - T\right) \cdot \left(\frac{n \cdot (n-1)}{2} - U\right)}},$$

ahol n a minták elemszáma, a P a pozitív különbségek összege, az I a negatív különbségek összege, a T és az U az X és az Y változók kapcsolt rangjait veszik figyelembe. A T és az U értelmezése:

$$T = \frac{\sum_{i=1}^{b} t_i(t_i - 1)}{2}, U = \frac{\sum_{i=1}^{c} u_i(u_i - 1)}{2},$$

ahol b és c a különböző kapcsolt rangok száma az X és Y változóban, t_i és u_i pedig adott kapcsolt rang összes előfordulása X, ill. Y változóban.

Például a 7.5. táblázatban az X változóra két különböző kapcsolt rang van ($b = 2$), azon belül pedig mind a kettőből kettő-kettő fordul elő. Az Y változóban nincsenek kapcsolt rangok ($c = 0$). Ebben az esetben az $U = 0$, a T pedig:

$$T = \frac{2 \cdot (2-1) + 2 \cdot (2-1)}{2} = 2.$$

Minél több kapcsolt rang fordul elő az adatsorokban, a τ_a annál kisebb lesz a τ_b-hez viszonyítva.

A Kendall-féle tau könnyen számolható R-ben (lásd a 7.2. táblázatot). Az *airquality* adattáblában az ózonkoncentrációra és a napsugárzásra az augusztusi hónapra a Kendall-féle tau b-t kapjuk meg a kapcsolt rangok miatt. Az *a* adattáblához a 7.2. alfejezetben leírt módon jutottunk el.

```
#Kiszámoljuk a Kendall-féle taut.
cor(a$Ozone, a$Solar.R, method = "kendall")
[1] 0.4695997

#A szignifikáns korreláció tesztelése.
cor.test(a$Ozone,a$Solar.R, method = "kendall")
Kendall's rank correlation tau
```

```
data: a$Ozone and a$Solar.R
z = 2.6165, p-value = 0.008885
alternative hypothesis: true tau is not equal to 0
sample estimates:
 tau
0.3928602
```

Az ózonkoncentráció és a napsugárzás között szignifikáns pozitív korreláció van, $\tau_b = 0.39$, $p = 0.009$. Az érték közeli a Spearman-féle rho értékhez (0.47).

7.4. Somers-féle delta

A Somers-féle delta két ordinális változó közötti asszociációt jellemzi. Noha a Kendall-féle tau is alkalmas erre a feladatra, a Somers-féle deltát akkor használjuk, amikor a két változó nem független egymástól (Somers, 1962). Ebből adódik, hogy a két változó nem egyenértékű, megkülönböztetünk egy független (X) és egy függő változót (Y), így a Somers-féle D egy aszimmetrikus mérőszám. Azt fejezi ki, hogy mennyivel nagyobb a konkordancia (egyirányú párok) aránya a diszkordanciánál (fordított párok), ha az X értékei nem egyeznek meg. Kapcsolt pároknak azokat tekinti, amelyeknél a két változó értéke azonos.

$$D = \frac{P - I}{P + I + T_x},$$

ahol P az összes lehetséges objektumpárra (egyedpárra) a két változóban azonos irányba mutató rangok száma, az I az ellentétes irányba mutató rangok száma, T_x a kapcsolt rangok összege az X változóban.

A manuális számítás bonyult, ezért a Somers-féle deltát az R-ben számoljuk ki. Az alábbiakban részszámításokat mutatunk be. A 7.6. táblázatban egy részeredményt ragadtunk ki egy terepi felmérésből, ahol egy növény fénykedvelő tulajdonságát vizsgáltuk.

7.6. táblázat. *Terepi felmérés (egy ordinális és egy diszkrét változó alapján)*

Kvadrát	Fényviszonyok (X)	Egyedszám (Y)
K1	Erős (3)	4
K2	Gyenge (1)	3
K3	Közepes (2)	2
K4	Gyenge (1)	2

A K1-K2 pár esetében a fényviszonyok és az egyedszám tekintetében nagyobb értéket találunk a K1 kvadrátban, tehát a rangok azonos irányba mutatnak (konkordancia). A K2-K3 párnál a fényviszonyok jobbak a K3-nál, de az egyedszám kisebb (diszkordancia). A K2-K4 párnál a fényviszonyok egyenlők, tehát ez az eset a T_x kapcsolt rangok közé sorolódik. A K3-K4 párnál a függő változó értékei azonosak, ezért ez a pár kiesik a számításokból. A négy kvadrátra kapott adatok alapján: $P = 3$, $I = 1$, $T_x = 1$, a delta pedig:

$$D = \frac{3 - 1}{3 + 1 + 1} = 0.4.$$

R-ben a Somers-féle delta a {DescTools} csomagban levő **`SomersDelta()`** paranccsal számolható ki, ahol a konfidenciaintervallumot is kikérhetjük adott α mellett, ami alapján eldönthetjük, hogy a delta értéke szignifikáns asszociációt mutat-e (7.7. táblázat). Amennyiben a konfidenciaintervallum tartalmazza a nullát, nincs szignifikáns kapcsolat a két változó között.

7.7. táblázat. *A Somers-féle delta kiszámítása R-ben*

Parancs	Részletek	Magyarázat
SomersDelta(x, y, direction, conf.level) {DescTools}		
x	Numeric	Független változó
y	Numeric	Függő változó
direction	"row", "column"	Alapból a sorokat veszi pároknak.
conf.level		Alapból nem adja meg. 0.95 ($\alpha = 0.05$)

```
f = c(3,1,2,1)
e = c(4,3,2,2)
SomersDelta(f,e, conf.level = 0.95)
     somers     lwr.ci     upr.ci
  0.4000000 -0.5075289  1.0000000
```

Annak ellenére, hogy a Somers-féle delta értéke viszonylag nagy (0.4), ez az érték mégsem mutat szignifikáns kapcsolatot a két változó között, mivel a konfidenciaintervallum a nulla értéket is tartalmazza. A minta elemszáma túl kicsi, ha valós eredményt akarunk, akkor több kvadrátot kell felmérni.

A {vegan} csomagban a *dune.env* és a *dune* adattáblák környezeti, ill. egyedszámváltozókat tartalmaznak 20 kvadrátból. Tesztelhetjük a közönséges cickafark (*Achillea millefolium*) nedvességkedvelő tulajdonságát a *Moisture* ordinális változó és az *Achimill* egyedszámváltozó alapján. A Moisture változónak öt szintje van, az 5-ös érték a legnedvesebb talajt jelöli.

```
library(vegan)
data("dune.env")
data("dune")
library(DescTools)
SomersDelta(dune.env$Moisture, dune$Achimill,
 conf.level = 0.95)
 somers lwr.ci upr.ci
-0.6037736 -0.9774908 -0.2300564
```

A Somers-féle delta (-0.604) és a konfidenciaintervallum (-0.977, -0.230) alapján azt mondhatjuk, hogy szignifikáns negatív kapcsolat áll fenn a két változó között, ami azt mutatja, hogy a közönséges cickafark nem kedveli a nedves talajt.

7.5. Két diszkrét változó függetlenségének vizsgálata

Diszkrét változók esetén, amelyek lehetnek nominálisak vagy ordinálisak, kontingenciatáblázatba rendezzük az adatokat. A táblázat segít átlátni a változók különböző szintjéhez tartozó adatokat, illetve vizsgálható a szintek egymáshoz való viszonya is. Legyen egy példa, ami egy kontroll- és egy tesztcsoportra tartalmaz dichotóm nominális adatokat: kezelés (igen, nem), javulás (igen, nem). Az adatok a 7.8. táblázatban láthatók.

7.8. táblázat. *Két diszkrét változó adatai kontingenciatáblázat*

igen		Javulás		*Összeg*
		igen	nem	
Kezelés	igen	26	29	*55*
	nem	35	15	*50*
Összeg		*61*	*44*	*105*

A két változó közötti kapcsolat létezésének megállapításához a két változó közötti függetlenség állapotából indulunk ki. Ez a feltétel akkor teljesül, ha a változók szintjei által meghatározott csoportokban az eloszlás megegyezik az összevont szintek eloszlásával. A *Kezelés* változó eloszlása a szintjeinek megfelelően: 55 igen, 50 nem. Ha a változó nem függ a *Javulás* változótól, akkor az utóbbi szintjein is azonos eloszlást várunk. Ez könnyen kiszámolható a két szélső oszlop segítségével (7.9. táblázat). A várt értékek és a tényleges értékek között eltérés van. Ez az eltérés lehet a véletlen eredménye vagy annak a kö-

vetkezménye, hogy a két változó nem független egymástól. Az eltérés okának megállapítására χ^2-próbát végezhetünk el.

7.9. táblázat. *Egymástól független változó várt értékei az összes szinten*

igen		Javulás: nem		*Összeg*
Kezelés	igen	$\frac{61 \cdot 55}{105} = 31.95$	$\frac{44 \cdot 55}{105} = 23.05$	*55*
	nem	$\frac{61 \cdot 50}{105} = 29.05$	$\frac{44 \cdot 50}{105} = 20.95$	50
Összeg		61	44	105

A várt értékek számításának általános képlete:

$$\text{várt érték} = \frac{\text{sor összeg} \cdot \text{csoport összeg}}{\text{teljes összeg}}.$$

7.5.1. A χ^2- és Fisher-próba a függetlenség tesztelésére

A χ^2-próba nullhipotézise azt állítja, hogy a két változó független egymástól. A χ^2, amiről a próba a nevét kapta, az alábbi képlettel számolható ki:

$$\chi^2 = \sum_{i=1}^{s} \sum_{i=1}^{o} \frac{\left(d_{ij} - d_{ij}^{exp}\right)^2}{d_{ij}^{exp}},$$

ahol d_{ij} a kereszttábla (kontingenciatáblázat) *i* sorában és *j* oszlopában levő érték, d_{ij}^{exp} pedig a d_{ij} helyett várt érték. Az előbb említett példa esetében a χ^2 értéke:

$$\chi^2 = \frac{(26 - 31.95)^2}{31.95} + \frac{(29 - 23.05)^2}{23.05} + \frac{(35 - 29.05)^2}{29.05} + \frac{(15 - 20.95)^2}{20.95}.$$

Az így kapott $\chi^2 = 5.56$. A próbastatisztikát azzal a χ^2-eloszlással értékeljük ki, amelynek szabadságfoka: df = (sorok – 1)·(oszlopok – 1). A *2x2* táblázat esetén $df = 1$. Ha a szignifikanciaszint 0.05, akkor a kritikus χ^2-érték 3.841. A próbastatisztika nagyobb a kritikus értéknél, tehát a próba szignifikáns eredményt adott, a két változó nem független egymástól. R-ben a χ^2-próbát a `chisq.test()` paranccsal lehet elvégezni (7.10. táblázat).

7.10. táblázat. *A χ^2- és a Fisher-próba R-ben*

Parancs	Részletek	Magyarázat
proba = chisq.test(x, y, correct) fisher.test(x, y, conf.level)		
x	Numeric/Nominal	diszkrét változó
y	Numeric/Nominal	diszkrét változó
correct	FALSE/TRUE	2x2 táblázatnál a Yates-féle korrekció elvégzése
proba$expected	mátrix	megadja a várt értékeket (ellenőrizni lehet a kis értékeket)
conf.level	0.95	értéke változtatható

```
#X²-próba a 7.9. táblázatra
d = read.csv("https://goo.gl/j6lRXD")
d = d[,c(2,3)]
table(d)
 improvement
treatment improved not-improved
 not-treated 26 29
 treated 35 15
chisq.test(d$treatment,d$improvement, correct =F)
   Pearson's Chi-squared test
data: d$treatment and d$improvement
X-squared = 5.5569, df = 1, p-value = 0.01841

#A várt értékek táblázata
proba = chisq.test(d$treatment,d$improvement, correct =F)
proba$expected
 d$improvement
d$treatment improved not-improved
 not-treated 31.95238 23.04762
 treated 29.04762 20.95238
```

A χ^2-próba szignifikáns eredményt adott ($\chi^2 = 5.56$, $df = 1$, $p = 0.018$). A χ^2-értéke annál nagyobb, minél erősebb a kapcsolat a két változó között.

χ^2-próba az illesztés jóságára

A χ^2-próba lehetővé teszi, hogy egy változó értékeit egy elméleti eloszláshoz hasonlítsuk. Ebben az esetben a változó minden szintjére kiszámolható az elméleti eloszlás által meghatározott érték. A χ^2 értékét tehát a megfigyelt és az elméleti eloszlás szerint elvárt értékek közti különbségekből lehet kiszámolni. Ezt a módszert az illesztés jóságának (*goodness of fit*) nevezzük.

Alkalmazási feltételek

1) 2x2 típusú táblázatok

A χ^2-próba nem megbízható abban az esetben, ha a kereszttáblában kis számok vannak. Ha az összelemszám kisebb 20-nál, vagy 20 és 40 között van, de a legkisebb várt érték 5-nél kisebb, a Fisher-féle exakt próbát kell használnunk (7.10. táblázat).

Ha az összelemszám 100-nál kisebb, vagy bármely érték 10-nél kisebb, akkor a Yates-féle korrekciót lehet alkalmazni. Ez a próba be van építve az R alapcsomagjában levő **`chisq.test()`** parancsba.

2) Bármely más esetben

A χ^2-próba nem alkalmazható, ha a várt értékek ~20%-a 5-nél kisebb, vagy bármely várt érték kisebb 1-nél. Ilyen esetben a kutatást újra kell gondolni, növelve a minta elemszámát.

7.5.2. Az asszociáció mérőszámai

A φ együttható

A változók szintjeinek növekedésével a χ^2 értéke nő, ezért ilyen esetben korrekciót kell végezni. Ha az egyik változónak két szintje van, akkor a legmegfelelőbb korrekció a φ (fí).

$$\varphi = \sqrt{\frac{\chi^2}{n}},$$

ahol n a minta elemszáma.

A kapcsolat a két változó között nagyon erős, ha a φ értéke > 0.7, erős, ha $0.4 < \varphi$ értéke < 0.7, közepes, ha $0.3 < \varphi$ értéke < 0.4, kicsi, ha $0.2 < \varphi$ értéke < 0.3, és elhanyagolható, ha φ értéke < 0.2.

A φ értéke nem esik minden esetben $[-1, +1]$ közé, van, amikor $\varphi > 1$. A φ maximális értékét a következő képlettel számolhatjuk ki:

$$\varphi_{max} = \sqrt{\min(\text{sorok száma}, \text{oszlopok száma}) - 1}.$$

A képlet alapján belátható, hogy a φ akkor lesz kisebb 1-nél, ha vagy a sorok, vagy az oszlopok száma 2. Tehát a φ együtthatót akkor használhatjuk, ha legalább a sorok vagy az oszlopok száma 2, vagy mindkettőé 2.

A Pearson-féle kontingencia együttható

A kontingencia együttható (C) értéke [0,1] között változik, azonban nem mindig az 1-es érték mutatja a tökéletes kapcsolatot. Az együttható az alábbi képlettel értelmezhető:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}},$$

ahol n a minta elemszáma.

Ha a C nullához közeli, akkor a kapcsolat a két változó között elhanyagolható, ha 1-hez közelít, akkor a változók között erős kapcsolat van. A tökéletes kapcsolat akkor jellemezhető az 1-es értékkel, ha *5x5*-ös vagy ennél nagyobb táblázattal dolgozunk. Ezért kisebb szintek esetén a kontingencia együtthatót nem is érdemes használni. A C_{max} kiszámolható a következő képlettel:

$$C_{max} = \sqrt{\frac{\min(\text{sorok száma}, \text{oszlopok száma}) - 1}{\min(\text{sorok száma}, \text{oszlopok száma})}}$$

annak érdekében, hogy a C_{max} minden esetben 1 legyen, Sakoda (1977) korrigálta a C értéket:

$$C_{korr} = \sqrt{\frac{\chi^2}{\chi^2 + n}} \cdot \sqrt{\frac{\min(\text{sorok száma}, \text{oszlopok száma})}{\min(\text{sorok száma}, \text{oszlopok száma}) - 1}}.$$

A Cramer-féle V együttható

A Cramer-féle V együttható a leggyakrabban használt mutatója két diszkrét változó kapcsolatának. Értéke bármely kereszttábla esetén [0,1] között változik. Tulajdonképpen a φ együttható kibővítése:

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min(s, o) - 1)}} = \varphi \cdot \sqrt{\frac{1}{\min(s, o) - 1}},$$

ahol a s a sorok számát, az o az oszlopok számát jelöli. A 7.11. táblázat tartalmazza az együttható kiértékelési kritériumait.

7.11. táblázat. *A Cramer-féle* V *együttható kiértékelése*

V-érték	Asszociáció mértéke
0–0.1	nincs asszociáció
0.1–0.3	gyenge
0.3–0.5	közepes
0.5–1.0	erős

A 7.12. táblázat tartalmazza az asszociációs együtthatók és a Cramer-féle *V* konfidenciaintervallumának számolását R-ben. A {confintr} csomag lehetővé teszi a konfidenciaintervallum meghatározását bootstrap módszerrel.

7.12. táblázat. *A φ, a* C *és a Cramer-féle* V *együtthatók R-ben*

Parancs	Részletek	Magyarázat
Phi(x, y) {DescTools} ContCoef(x, y, correct) {DescTools} CramerV(x, y, conf.level) {DescTools} ci_cramersv(d, probs, type, boot_type) {confintr}		
x	Numeric/Nominal	diszkrét változó
y	Numeric/Nominal	diszkrét változó
conf.level	NA	Értékét meg kell adni a CramerV-próbánál. Ha nem tartalmazza a 0-t, akkor az asszociáció szignifikáns.
correct	FALSE/TRUE	Alapból nem korrigál. Ha TRUE, akkor a Sakoda-féle korrekciót alkalmazza *C*-re.
d	data.frame chisq.test result	Adattábla a két változóval vagy egy chisq.test eredmény.
probs	c(0.025, 0.975)	kétoldali teszt ($\alpha = 0.05$)
type	c("chi-squared", "bootstrap")	A konf. int. meghatározási módszere. Javasolt a "bootstrap".
boot_type	c("bca", "perc", "norm", "basic")	Javasolt a "bca".

Példa

Vizsgáljuk meg egy ordinális (*Manure*) és egy nominális (*Management*) változó kapcsolatát a *dune.env* adattáblában, ami a {vegan} csomagban található. Az adattábla 20 kvadrátban vizsgálja a terület felhasználását a földművelés és területmenedzsment szempontjából. A *Manure (M1)* változó a trágyázás mértékét mutatja 0 és 4 közötti szintekkel, a *Management (M2)* nominális változónak négy szintje van: BF (Biological farming), HF (Hobby farming), NM (Nature Conservation Management) és SF (Standard Farming). Az adatokat egy kereszttáblába rendezzük (7.13. táblázat).

7.13. táblázat. *Két diszkrét változó adatai kereszttáblában*

		M2				*Összeg*
		BF	HF	SF	NM	
M1	0	0	0	0	6	*6*
	1	2	1	0	0	*3*
	2	1	2	1	0	*4*
	3	0	2	2	0	*4*
	4	0	0	3	0	*3*
Összeg		*3*	*5*	*6*	*6*	20

Kérdés: van-e kapcsolat a két változó között? Mivel a kereszttábla 4x5-ös típusú, és sok benne a nulla, a Cramer-féle V a legmegfelelőbb mutató.

```
#Beolvassuk és rendezzük az adattáblát
library(vegan)
data(dune.env)
d = dune.env
d = d[,c("Management", "Manure")]
table(d)
 Management
 Manure   BF   HF   NM   SF
      0    0    0    6    0
      1    2    1    0    0
      2    1    2    0    1
      3    0    2    0    2
      4    0    0    0    3

#Cramer-féle V
library(DescTools)
CramerV(d$Manure,d$Management)
[1] 0.7533874

#Konfidenciaintervallum meghatározása
ci_cramersv(d, type = "bootstrap")
   Two-sided 95% bootstrap confidence interval for the population
Cramer's V based on 9999 bootstrap replications and the bca method
   Sample estimate: 0.7533874
   Confidence interval:
    2.5% 97.5%
   0.6411795 0.8076975
```

Az eredmények alapján azt mondhatjuk, hogy a két változó között erős pozitív asszociáció van (Cramer-féle $V = 0.753$), és az asszociáció szignifikáns (a konfidenciaintervallum 0.64 és 0.81 közé esik, ha $\alpha = 0.05$). Tehát a terület felhasználási módja és a trágyázás mértéke függnek egymástól.

8. REGRESSZIÓANALÍZIS

A regresszióanalízis a legszélesebb körben alkalmazott statisztikai módszer, amellyel ok-okozati összefüggések modellezhetők. Ez az elemzés lehetővé teszi annak a tesztelését, hogy egy függő változó értékeit milyen mértékben lehet megbecsülni egy vagy több független változó értékeiből.

Maga a fogalom Francis Galton angol polihisztor nevéhez kötődik, aki először használta a regressziót, amikor azt vizsgálta, hogy milyen mértékben határozza meg a szülők átlagmagássa az utódaik magasságát. Klasszikussá vált tanulmányában: *Regression towards mediocrity in hereditary stature* (Az örökölt testmagasság regressziója a középszerűség felé) azt a törvényszerűséget fedezte fel, hogy a szülők átlagmagasságából előre megmondható az utód felnőttkori testmagassága, és ez a magasság az átlagos érték felé mozdul el (Galton, 1886). Ennek a megállapításnak megfelelően a magas, ill. alacsony szülők gyermekei is magasak, ill. alacsonyak lesznek, de magasságuk közelebb lesz az átlagmagassághoz, mint a szülőké. Galton úttörő munkája után számos kutatás bizonyította, hogy a testmagasság az egyik legerősebb örökletes tulajdonság. A nem- és életkor szerinti változatosság figyelembevételével a testmagasság 80%-ban örökletes (Visscher et al., 2006).

A regresszióanalízisnek több változata van, amit a független változók típusa határoz meg. Ezek a változatok a 8.1. táblázatban figyelhetők meg.

8.1. táblázat. *A regresszióanalízis főbb típusai*

Típus	Adatok átalakítása	A függő változó eloszlása	A magyarázó változó eloszlása
lineáris	–	normális	normális
varianciaelemzés	–	normális	kategorikus
kovarianciaelemzés	–	normális	normális és kategorikus
logisztikus	logit	binomiális	normális és kategorikus
Poisson	logaritmus	Poisson	normális és kategorikus
negatív-binomiális	logaritmus	negatív binomiális	normális és kategorikus

Egyes típusoknál szükségessé válik az adatok átalakítása. A legalkalmasabb átalakítás a logaritmizálás, ugyanis ez a fajta átalakítás könnyen értelmezhető utólag. Az átalakított függő változó eloszlásának sajátosságai jobban megfelelnek a lineáris regresszió követelményeinek, a modellalkotásnál pedig használható a lineáris regressziónál bevált legkisebb négyzetek módszere (Moksony, 2006).

8.1. A lineáris regresszió

A lineáris regresszió olyan parametrikus eljárás, amely a függő és független változó között lineáris kapcsolat létezését modellezi, és megadja az *Y* becsült értékét az *X*-ből. A függő változót *Y*-nal, a független *X*-szel szokták jelölni. Az *X* alapján becsült *Y* értéke az alábbi lineáris összefüggés segítségével számolható ki:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon,$$

ahol $\hat{\beta}_0$ és $\hat{\beta}_1$ becsült értékek. Az *XY* koordináta-rendszerben a $\hat{\beta}_0$ az egyenes metszéspontja az Y tengellyel, a $\hat{\beta}_1$ pedig az egyenes iránytangense (meredeksége). Mivel az *Y* valószínűségi változó, a $\hat{\beta}_0$ és $\hat{\beta}_1$ is azok, ezért ezeknek is standard hibája van, az egyenletbe bekerült értékek pedig a becsült átlagértékek ($\hat{\beta}_0$ és $\hat{\beta}_1$). Az ϵ hibatag is valószínűségi változó, a véletlen hibák összességét jelenti, ezért eloszlása az $N(0,\sigma^2)$ standardizált normál eloszlásfüggvényt követi. Értelmezése:

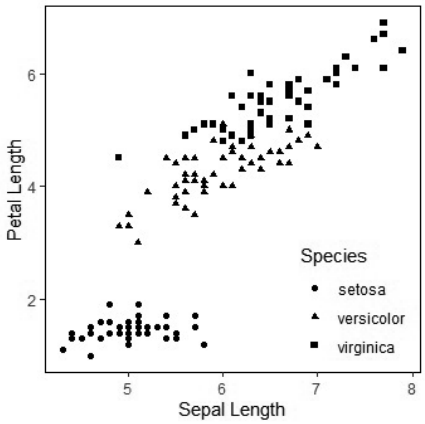
$$\epsilon = \sum_{i=1}^{n} e_i^2,$$

ahol e_i az y_i értékhez kapcsolodó hiba, σ^2 pedig a hibák varianciája.

A lineáris regresszióanalízis tesztelését egyszempontos ANOVA-val végezzük, amely a következő hipotézisekre épül:

H_0: Az *X* értékeihez *Y* értékei véletlenszerűen társulnak, azaz $\hat{\beta}_1$ nem különbözik nullától (az egyenes párhuzamos az *X* tengellyel).

H_1: Az *X* értékeiből megbecsülhetők *Y* értékei, azaz $\hat{\beta}_1$ különbözik nullától (az egyenes nem párhuzamos az *X* tengellyel).



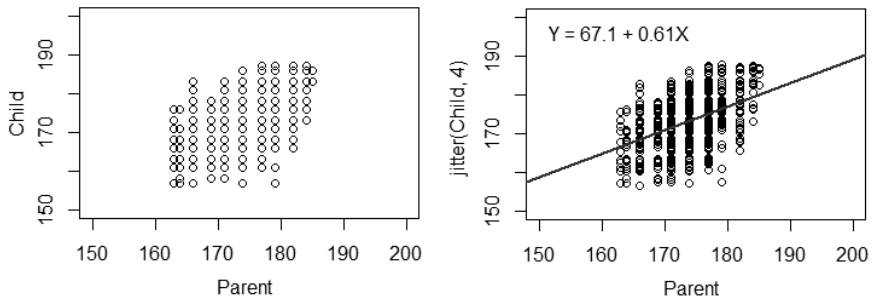
8.1. ábra. *Az* Iris setosa *faj adatai különálló csoportot képeznek*

Amennyiben az X és Y nem normális eloszlásúak, és a mintaszám kicsi ($n < 30$), az adatokat transzformálni kell (ferde eloszlásnál a logaritmálást tanácsos alkalmazni). Az X változó kiugró értékei jelentősen torzíthatják a modellt. Szintén torz, nem valós eredményekhez vezet, ha az adatok heteroszkedasztikusak, vagy egymástól távol eső csoportosulásokat mutatnak. Példa erre az *iris* adattábla, amelyben az *I. setosa* faj egyes értékei eltérnek a másik két fajétól (8.1. ábra). Ilyen esetben a modell hatékonysága romlik. Tanácsos a faj adatait eltávolítani az adattáblából.

Példa: Galton adatai ma nyilvánosak, a {UsingR} csomagból betölthetők R-be, jelen helyzetben példaként használjuk a következőkben. Az adattábla 928 szülő–utód pár adatait tartalmazza. Az utódok magassága (child) az Y változó szerepét veszi fel, a szülők átlagmagassága (parent) pedig az X változóét. A magasságok inch-ben vannak megadva, ezért a jobb követhetőség miatt átalakítjuk cm-be (1 inch = 2.54 cm).

```
install.packages("UsingR")
library(UsingR)
d = galton
dim(d)
[1] 928 2
head(d)
    child  parent
 1   61.7    70.5
 2   61.7    68.5
 3   61.7    65.5
 4   61.7    64.5
 5   61.7    64.0
 6   62.2    67.5

#Az értékeket átalakítjuk cm-be.
d = round(d*2.54, 0)
names(d) = c("Child","Parent")
head(d)
    Child  Parent
 1    157     179
 2    157     174
 3    157     166
 4    157     164
 5    157     163
 6    158     171
```

A 8.2. ábra bal oldali ábráján látható a kapcsolat a két változó között. Az ábrán sok pont egymásra tevődik, ezért a pontok sűrűségének érzékeltetésére kismértékben eltávolíthatók egymástól (jobb oldali ábra). Az utóbbin megfigyelhető, hogy a legtöbb esetben a szülők és az utódok magassága a középértékek köré tömörülnek.



8.2. ábra. *A szülők átlagmagassága és az utódok magassága közti összefüggés Francis Galton adatai alapján (Galton, 1886)*

Ritkán fordultak elő olyan esetek, amikor a szülők és utódaik nagyon alacsonyak vagy nagyon magasak voltak. Az ábrán az is látható, hogy adott szülők átlagmagasságához képest szélesebb tartományban változott az utódok testmagassága. Kérdés, hogy ennek ellenére létezik-e törvényszerűség a két magasság között, vagy teljesen véletlen az utódok magassága. Erre a kérdésre ad választ a lineáris regresszióanalízis.

A 8.2. ábra kódja:

```
par(mfrow=c(1,2))
plot(Child ~ Parent, galton,
 xlim = c(150,200), ylim = c(150,200))
plot(jitter(Child,4)~Parent,galton,
 xlim = c(150,200), ylim = c(150,200))
model = lm(Child~Parent, galton)
abline(model,lwd=2, col= "blue")
legend(150,199, "Y = 67.1 + 0.61X", bty = "n")
summary(model)
Call:
lm(formula = Child ~ Parent, data = d)
Residuals:
    Min      1Q  Median      3Q     Max
-19.336  -3.454  -0.167   4.546  14.597
```

```
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  67.10332     6.86354    9.777     <2e-16   ***
     Parent   0.61024     0.03955   15.431     <2e-16   ***
---
Signif.codes:0‘***’ 0.001‘**’ 0.01‘*’ 0.05‘.’
Residual standard error: 5.636 on 926 degrees of freedom
Multiple R-squared: 0.2046, Adjusted R-squared: 0.2037
F-statistic: 238.1 on 1 and 926 DF, p-value: < 2.2e-16
```

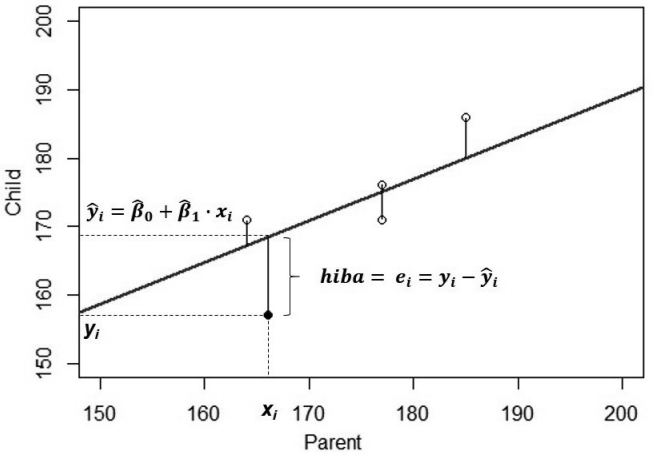
A 8.2. ábrán látható a lineáris regresszió által kapott egyenlet a becsült $\hat{\beta}_0$ és $\hat{\beta}_1$ értékekkel. Az X együtthatója egy pozitív szám, ami pozitív összefüggést mutat X és Y között. A modell szerint a szülők átlagmagasságának egységnyi (1 cm) növekedésénél az utód magassága csupán 0.61 cm-rel növekszik. Ez a növekedés független a szülők átlagmagasságától.

A hipotézisvizsgálatot a $\hat{\beta}_1$ együtthatóra végezzük el, amivel a $\hat{\beta}_1$-t becsüljük. A modell akkor lesz szignifikáns, ha a $\hat{\beta}_1$ értéke szignifikánsan eltér nullától. A $\hat{\beta}_1$ t-eloszlást követ. A $\hat{\beta}_1$ varianciája ($\sigma^2_{\hat{\beta}_1}$) tulajdonképpen azt fejezi ki, hogy hogyan aránylik az e hibatag varianciája az X varianciájához. Minél kisebb a hibatag varianciája, annál szorosabban illeszkednek a pontok az egyeneshez, így annál kisebb lesz a $\sigma^2_{\hat{\beta}_1}$. Minél nagyobb az X varianciája, annál kisebb a $\sigma^2_{\hat{\beta}_1}$, tehát minél jobban szórodnak az X értékei, annál jobb lesz az illesztés, és annál nagyobb lesz a modell magyarázóereje.

Az ANOVA elemzést a **`summary(model)`** paranccsal kapjuk meg. A parancs kimenete a két együtthatóra (*Coefficients*), az *Intercept* ($\hat{\beta}_0$) és a Parent ($\hat{\beta}_1$) értékelését mutatja. Az *Estimate* adja meg a becsült értékeket, emellé társul a standard hiba ($\frac{s}{\sqrt{n}}$), az a t-érték, amelyre $\hat{\beta}_0$ és $\hat{\beta}_1$ várt értéke nulla, és ehhez a t-értékhez tartozó p-érték. Mind a két együtthatóra a $p < 0.001$, tehát a $\hat{\beta}_0$ és $\hat{\beta}_1$ szignifikánsan különböznek a nullától. A modell kiértékeléséhez további paramétereket is figyelembe kell venni, mint például a reziduumok eloszlását, standard hibáját, illetve az R^2-értéket. Ezek értelmezéséhez szükség van a lineáris modell illesztési algoritmusának megértésére.

8.1.1. Egyenes illesztése a pontokhoz

A lineáris regresszió alapja az egyenes illesztésének módja a pontokhoz, ami a *legkisebb négyzetek módszerével* történik. A pontok kisebb-nagyobb szóródása miatt bármilyen egyenest illesztünk rájuk, mindig lesznek olyan pontok, amelyek nem lesznek rajta az egyenesen. A modell csupán azokat a pontokat tudja helyesen megjósolni, amelyek rajta vannak az egyenesen. Minél távolabb esik egy pont az egyenestől, annál nagyobb a modell hibája (8.3 ábra).



8.3. ábra. *A lineáris regressziós modell hibájának értelmezése egy pont esetében*

Az (x_i, y_i) pont becslésének a hibája e_i, a 8.2. ábrán kitüntetett pontté ez negatív érték. A különböző előjelek hatásának a kiköszöbölésére a hibákat négyzetre emelik. Az egyenes illesztése úgy történik, hogy a $\hat{\beta}_1$ értéke eleget kell hogy tegyen annak a feltételnek, hogy a függőleges hibanégyzetek összege minimális legyen. Matematikailag ez a következő kifejezéssel írható le:

$$\min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right).$$

A módszer tehát innen kapta a nevét. A két együttható értéke a következő értelmezést kapja:

$$\hat{\beta}_1 = cor(Y, X) \cdot \frac{sd(Y)}{sd(X)},$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}.$$

A $\hat{\beta}_0$ értelmezéséből következik, hogy az $(\bar{X}, \bar{Y})$ pont rajta van az egyenesen, az egyenes iránytangense $(\hat{\beta}_1)$ pedig függ a két változó korrelációs együtthatójától és a minták szórásaitól. Ezekből következik pár érdekesség:

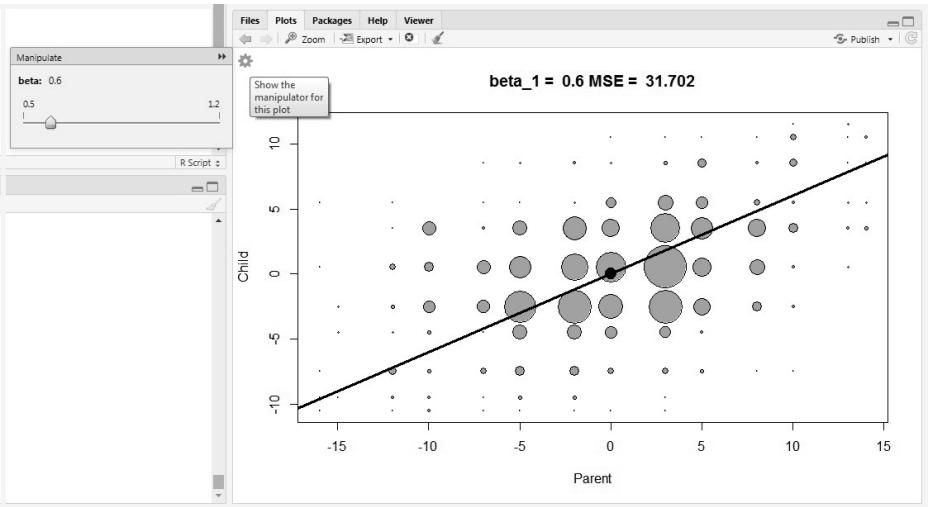
– ha X és Y mértékegysége megegyezik, akkor $\hat{\beta}_1$ dimenziómentes szám (egy szórzótényező), a $\hat{\beta}_0$ viszont az $\bar{Y}$ mértékegységével rendelkezik;
– ha a (0,0) pontra centráljuk a két adatsort (kivonjuk az átlagokat az értékekből), akkor a $\hat{\beta}_0 = 0$ (kiesik a modellből), az iránytangens viszont változatlan marad;

– ha z-transzformációt végzünk a két adatsorral, akkor a $\hat{\beta}_0 = 0$, az iránytangens pedig a korrelációs együtthatóval lesz egyenlő, függetlenül attól, hogy melyik változót választjuk függetlennek.

Belátható, hogy minél inkább elforgatjuk az egyenest, a szélsőértékek túlnyomó része nagyon távol fog esni az egyenestől, így a hibanégyzetek összege növekedni fog. Az egyenes forgatása és a négyzetösszeg értékének változása tanulmányozható az alábbi kóddal elkészített 8.4. ábrán, amelynek eredeti formáját Caffo (2015) készítette el Galton adataira. A forgatás kezeléséhez az összes egyenesnek a (0,0) ponton kell átmennie ($\hat{\beta}_0 = 0$), tehát a két változót az origóba kell centrálni.

```
install.packages("manipulate")
library(manipulate)
myPlot = function(beta){
 y = d$Child - mean(d$Child)
 x = d$Parent - mean(d$Parent)
 freqData = as.data.frame(table(x, y))
 names(freqData) = c("Child", "Parent", "freq")
 plot(
 as.numeric(as.vector(freqData$Parent)),
 as.numeric(as.vector(freqData$Child)),
 pch = 21, col = "black", bg = "grey70",
 cex = 0.15*freqData$freq,
 xlab = "Parent",
 ylab = "Child"
 )
 abline(0, beta, lwd = 3)
 points(0, 0, cex = 2, pch = 19)
 mse = mean((y - beta*x)^2)
 title(paste("beta_1 = ", beta, "MSE = ", round(mse, 3)))
}
#Megkeressük azt a β_1 értéket, amelyre a szórásnégyzetek
#átlaga(MSE)minimális.
manipulate(myPlot(beta, beta = slider(0.5,1.2,step=0.02))
```

A szimulálás során a legkisebb MSE értéket, a modell által becsült becsült $\hat{\beta}_1$ értékre adja, ami 0.61. A modell által becsült *y*-értékeket a **predict(model)** vagy a **model$fitted.values** paranccsal kérhetjük ki, és elmenthetjük egy vektor formájában. A **model = lm(y~x)**, a lineáris regresszió modellje.



8.4. ábra. *Az iránytangens és a szórásnégyzetek átlagának változása Galton adataira (magasságok cm-ben)*

8.1.2. A hibatag vizsgálata

A modell hibatagját (ϵ) a reziduumok összessége képezi. A reziduumok, a pontok eltérései a regresszós egyenestől, ahogy a 8.3. ábra is mutatja. Ezeket az eltéréseket az ANOVA nem értékeli ki, de tanulmányozható az eloszlásuk. Ezzel ellentétben a modell együtthatóinak hibái valódi hibák, amelyek a valószínűségi változók véletlen hibái. A reziduumokat a következőképpen kapjuk meg:

$$e_i = y_i - \hat{y}_i,$$

kifejtve:

$$e_i = y_i - \hat{\beta}_0 \ - \hat{\beta}_1 \cdot x_i.$$

A modell a reziduumok négyzetösszegét minimalizálja, így a reziduumok a modell hibáját is képezik, és egyben a modell sikerességének is a mutatói. R-ben a **`resid(model)`** vagy **`model$residuals`** parancsokkal kérhetjük ki őket, ha a **`model = lm(y~x)`**. A reziduumok összege, illetve a reziduumok és a független változó szorzatának összege nulla vagy nullához közeli szám.

```
e = resid(model)
sum(e)
[1] -5.327516e-12
sum(e*d$Parent)
[1] -8.728023e-10
```

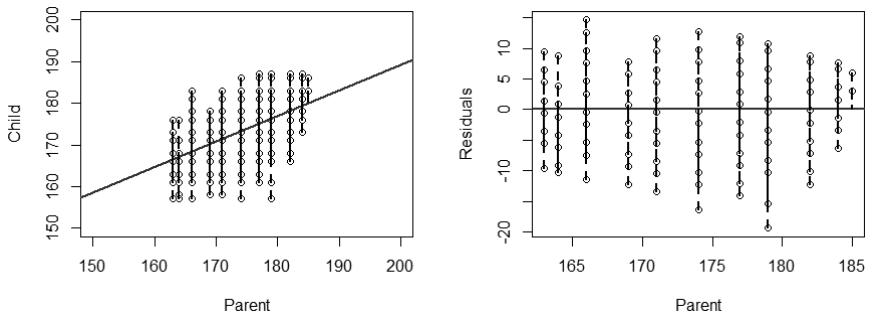
A reziduumok szabadságfoka a mintaegyedszámnál annyival kevesebb, mint ahány együttható van a modellben. Az egyszerű lineáris regressziónál tehát a $df = n - 2$, mivel a modell két együtthatót tartalmaz. A reziduumok szóródását a reziduális varianciával ($\hat{\sigma}^2$), ill. a reziduális szórással (a variancia négyzetgyökével) értelmezzük:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

A reziduális szórást az *R*-modell is megadja *Residual standard error* név alatt. A Galton-adatokra ez az érték 5.636. A két mutató az alábbi parancsokkal számolható ki R-ben:

```
#Reziduális variancia Galton adataira
sum(resid(model)^2)/(nrow(d)-2)
#Reziduális szórás Galton adataira
sqrt(sum(resid(model)^2)/(nrow(d)-2))
```

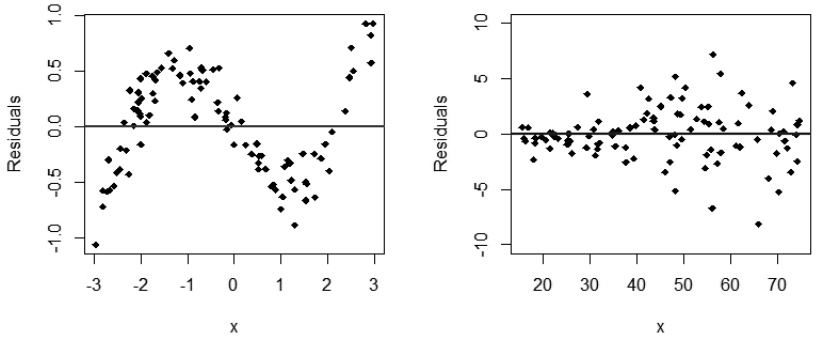
A modell sikerességét a reziduumok grafikus elemzésével, ill. egyes próbák elvégzésével is ellenőrizhetjük. Sok esetben az *XY* koordináta-rendszerben nem követhető tisztán a reziduumok mintázata (8.5. ábra, balra), ezért a leghasznosabb a reziduumok ábrázolása az *X* változó függvényében (8.5. ábra, jobbra). A reziduumok eloszlása egyenletes kellene legyen az *X* -tengely mentén, az értékeknek pedig nulla körül kell mozogniuk. Ideális esetben a reziduumok eloszlása a normális eloszlást követi. A modell akkor nem megbízható, ha a reziduumok eloszlásában valamilyen mintázat rajzolódik ki, mint a 8.6. ábrán. Az első ábrán látható egy mintázat, ami alapján nem minden *X* értékre teljesül az, hogy a reziduumok a nulla körül szóródnak. A második ábrán *X* értékeinek növekedésével a reziduumok szóródása egyre nagyobb, az adatokra heteroszkedasztikusok (vagyis a hibatag varianciája nem egyforma). A normális eloszlást általában vizuálisan teszteljük, a homoszkedaszcitiást pedig az NCV-próbával.



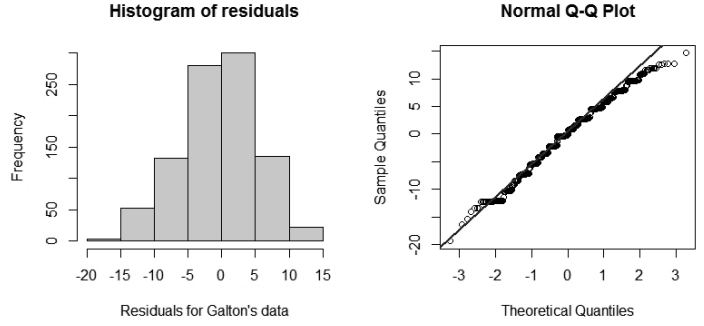
8.5. ábra. *A reziduumok grafikus vizsgálata.*
Balra: az *XY* koordináta-rendszerben.
Jobbra: A reziduumok ábrázolása az *X* változó függvényében.

A Galton adataira kapott lineáris modell reziduumai megközelítően a nulla körül szóródnak, és a szóródás nem mutat egy rendezett mintázatot (8.5. ábra). A normális eloszlást egy hisztogramon, ill. egy Q-Q ábrán ellenőrizhetjük (8.7. ábra). A homoszkedaszticitásra elvégezhetjük az NCV-próbát (lásd a 7.3. táblázatot). Az NCV-próba alapján a variancia állandónak tekinthető a Galton-adatokra ($\chi^2 = 0.499$, df = 1, $p = 0.480$). Ezzel szemben a 9.6. ábra jobb oldali adataira a próba szignifikáns eredményt ad ($\chi^2 = 16.46$, $df = 1$, $p < 0.001$), tehát az adott kapcsolat leírására a lineáris regressziós modell nem megfelelő.

```
library(car)
ncvTest(lm(d$Child~d$Parent))
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.4994732, Df = 1, p = 0.47973
ncvTest(lm(y~x))
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 16.45993, Df = 1, p = 4.9689e-05
```



8.6. ábra. *Mintázatok a reziduumok eloszlásában*



8.7. ábra. *A Galton-adatok lineáris regresszió által kapott reziduumainak eloszlása*

8.1.3. A lineáris regresszió magyarázóereje

A modell magyarázóerejét a következő mutatókkal jellemezhetjük: a determinációs együttható (R^2), a korrigált R^2, az átlagos négyzetösszeg (*MSE*) és ennek a négyzetgyöke (*RMSE*), ill. a reziduumok varianciája.

A determinációs együttható (R^2) megadja a függő változó (Y) teljes varianciájának azt a hányadát, amelyet a lineáris regresszió magyaráz. A lineáris regresszió azt magyarázza, hogy a becsült $\hat{Y}$ értékek mennyire térnek el az átlagértéktől ($\bar{Y}$). Mivel a szabadságfokok megegyeznek, a képletben a négyzetösszegek szerepelnek:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

A determinációs együttható kifejezhető az R-ben használt jelölésekkel is, a következő módon:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS},$$

ahol az *RSS* a regressziós (magyarázó) négyzetösszeg, a *TSS* a teljes négyzetösszeg, az *ESS* pedig az eltérés-négyzetösszeg.

Az R^2 értéke [0,1] között változik, és a két változó közötti Pearson-féle korrelációs együttható négyzetével egyenlő. Minél nagyobb az értéke, annál jobban magyarázza a modell az Y változó értékeit. Az R^2 értéke függ a minta elemszámától (adatok hozzáadásával az értéke csökken). Ebből adódik, hogy a modell hatékonyságát nem lehet megítélni pusztán az R^2 értékével, a reziduumok eloszlását is mindenképp ki kell ehhez értékelni.

A korrigált R^2 értéke ugyancsak [0,1] közötti érték, viszont figyelmbe veszi a független változók számát, ugyanis ennek a számnak a növelésével az R^2 értéke növekszik. Azt a látszatot kelti, hogy a modell megbízhatósága javul. Ezt küszöböli ki a korrigált R^2. Egyetlen független változó esetén az R^2 és a korrigált R^2 érték egymáshoz közel állnak.

Az *átlagos négyzetösszeg – MSE* (mean squared error) – a modell által becsült Y értékek és a valós Y értékek közti különbségek négyzetösszegeinek az átlagát fejezi ki.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Ha az illesztett egyenes az Y összes értékeit tartalmazza, akkor az $MSE = 0$. Ilyen eset a valóságban nem szokott előfordulni. Ha vannak pontok, amelyek nincsenek rajta az egyenesen, az *MSE* értéke egy pozitív szám lesz. Minél nagyobb ez az érték, annál rosszabb a modell magyarázóképessége. Az *MSE* értékét nehezebb értelmezni, mint a determinációs együtthatóét, mivel abszolút értéke van, míg a determinációs együttható egy viszonyszám. Általában a modell nagyon gyenge teljesítménye mutatható ki vele.

R-ben az átlagos négyzetösszeg többféleképpen számolható ki. Legyen a lineáris modell a `model = lm(x~y, data = d)`.

```
#A hiba átlagos négyzetösszege (MSE)
MSE = mean(model$residuals^2)
#vagy
#Kiszámoljuk a reziduálisokat
MSE = mean((d$y - predict(model))^2)
```

A *gyök átlagos négyzetösszeg – RMSE (root mean squared error*) – előnye az átlagos négyzetösszeggel szemben az, hogy a reziduumok átlagos szóródását adja meg az Y változó értékskálájának megfelelően. A mutató egy standardizált mérőszám, az értéke erősen függ az Y értékeitől.

Ha a lineáris modell: `model = lm(x~y, data = d)`, és az MSE értékét a fenti parancsok egyikével kaptuk meg, akkor R-ben az RMSE a következő parancscsal számolható ki:

```
#A gyök átlagos négyzetösszeg kiszámítása:
RMSE = sqrt(MSE)
```

8.2. Általánosított lineáris modellek (GLM)

Az általánosított lineáris modelleket Nelder és Wedderburn dolgozta ki 1972-ben (Nelder és Wedderburn, 1972). Az ő meghatározásuk szerint „a modell a legnagyobb valószínűség módszerét használja arra, hogy megbecsülje azokat a paramétereket, amelyekre a megfigyelt értékek eloszlása az exponenciális függvénycsaládhoz tartozó függvénnyel jellemezhető, a szisztematikus hatások pedig egy megfelelő transzformációval linearizálhatók”. Az általánosított lineáris modellek (Generalized Linear Models – GLM) egy modellcsaládot alkotnak, amelyek közös jellemzője az, hogy a függő változó (Y) várható értékét egy kapcsolati függvény segítségével hozza összefüggésbe a lineáris modellel. Az értelmezés szerint három részből tevődnek össze: szisztematikus és véletlen komponensből, valamint egy kapcsolati függvényből:

– szisztematikus komponens (az Y várható értéke): $E(Y) = \mu$;
– véletlen komponens (lineáris prediktor, amit X határoz meg): η;
– kapcsolati függvény: $g(\mu) = \eta$,

ahol μ az Y várható értéke, η az Y értékből logaritmikusan transzformált érték, a g pedig az a függvény, ami megteremti a kapcsolatot a μ és az η között.

Nedler és Wedderburn (1972) zsenalitása a g függvény bevezetésében rejlik, ami az Y diszkrét értékeiből folytonos skálájú változót (η) hoz létre, aminek az értékei ($-\infty$, $+\infty$) között változnak.

A modell koefficienseinek ($\hat{\beta}_0$ és $\hat{\beta}_1$) a meghatározása a legnagyobb valószínűség (maximum likelihood) módszerrel történik egy iteratív algoritmus segítségével, ami több próbálkozás árán megkeresi a legjobb megoldást. A következő egyenletre keresi a megoldást:

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\Phi \cdot var(Y_i)} \cdot W_i = 0,$$

ahol $var(Y_i)$ az eloszlás varianciája, a W_i egy származtatott érték, a Φ pedig egy korrekciós tényező. A valószínűségi változó típusa szerint a varianciát háromféleképpen értelmezhetjük:

– Y normális eloszlású: var(Y) = σ^2;
– Y binomális eloszlású: var(Y) = $\mu \cdot (1-\mu)$;
– Y Poisson-eloszlású: var(Y) = μ.

A modellek R-ben az alapcsomagban megtalálható *glm()* paranccsal készíthetők el, kivételt képez a negatív binomiális eloszlásra épülő modell, amit a {MASS} csomagban található `glm.nb()` paranccsal készíthetünk el (8.2. táblázat).

8.2. táblázat. *Általánosított lineáris modellek R-ben*

Parancs	Részletek	Magyarázat
model = glm(y~x, data, family) model.bn = glm.nb(y~x, data) {MASS} summary(model)		
x	Numeric/Nominal	független (prediktor) változó
y	Numeric/Binomial/Poisson	függő változó
data		adattábla
family	"gaussian", "binomial", "quasibinomial", "poisson", "quasipoisson"	normális, binomiális vagy Poisson-eloszlású *Y*-ra

Ha a Φ értéke 1, akkor a binomiális vagy a Poisson-eloszlással dolgozik a modell. Abban az esetben, ha az X szélső tartományaiban túl sok 0 van (vagy 1-es a binomiális eloszlásnál), akkor Φ értéke eltérő, így a modellt a kvázi-binomiális, ill. kvázi-Poisson-eloszlásra kell építeni.

8.2.1. A logisztikus regresszió

A logisztikus regressziót abban az esetben alkalmazzuk, amikor a függő változó bináris, eloszlása pedig a binomiális eloszlással írható le. Ilyen esetben a modell a következő komponenseket használja:

– szisztematikus komponens: $E(Y) = \mu$;
– lineáris prediktor: $\eta = \beta_0 + \beta_1 \cdot X$;
– kapcsolati függvény: $g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right)$.

A $\mu/(1\text{-}\mu)$ arányt esélyértéknek (O – odds) nevezzük, a μ pedig annak a valószínűsége, hogy egy esemény (siker) bekövetkezik. Tehát jelölhetjük p-vel. Az esélyérték kifejezi, hogy hányszor akkora a valószínűsége annak, hogy a siker bekövetkezik, mint annak, hogy nem. Ha például a p értéke 0.5, akkor az O értéke $0.5/0.5 = 1$, ha a $p = 0.75$, akkor az O értéke $0.75/0.25 = 3$. Az esélyérték természetes alapú logaritmusa logit néven ismert. A két paraméter könnyen átalakítható egymásba:

$$O = \frac{p}{1-p} \text{ és } p = \frac{O}{1+O}.$$

A p, O és η értékeinek kapcsolatát a 8.3. táblázat tartalmazza.

8.3. táblázat. *A siker valószínűségének (*p*), az esélyértéknek (*O*) és a lineáris prediktornak (*η*) a kapcsolata*

Mutatók	Értékek						
p	0	0.01	0.10	0.5	0.9	0.99	1
O	0	0.01	0.11	1	9	99	$+\infty$
η	$-\infty$	−4.60	−2.20	0	2.20	4.60	$+\infty$

Mivel $\hat{\beta}_0$ és $\hat{\beta}_1$ az η értékeit határozzák meg X alapján, ők maguk nem értelmezhetők valószínűségi értékekként (p). Át kell őket alakítani a $g(\mu)$ függvény inverzével:

$$p_i = \frac{e^{\eta i}}{1 + e^{\eta i}} = \frac{\exp(\beta_0 + \beta_1 \cdot x_i)}{1 + \exp(\beta_0 + \beta_1 \cdot x_i)}.$$

A modell alapján $\hat{\beta}_0$ visszaalakítva kifejezi annak a valószínűségét, hogy Y bekövetkezzen, ha X értéke nulla. A $\hat{\beta}_1$ visszaalakítva kifejezi Y-nak az esélyértékét, ha X egy egységgel növekszik. A modell szerint:

$$p_{x=0} = \frac{e^{\beta 0}}{1 + e^{\beta 0}}, p_{\Delta x=1} = e^{\beta 1}.$$

A logisztikus modell egy S-alakú görbe, ami 0 és 1 között változik a modellezett Y tengelyén. Minél inkább közelít $\hat{\beta}_1$ a nullához, annál inkább kisimul a görbe (8.8. ábra).

Példa

A *titanic.csv* adattábla 887 utas pár adatát tartalmazza a Titanic 2229 utasából. Ezek közt szerepel az életkor és az, hogy az illető túlélte vagy nem a katasztrófát. Az életkor, mint magyarázó változó, a túlélés mint függő változó szerepel a modellben.

Az adattábla az alábbi linken érhető el:
https://gist.github.com/sachinsdate/8b212e2c589a70910dfd04fae7d0f788

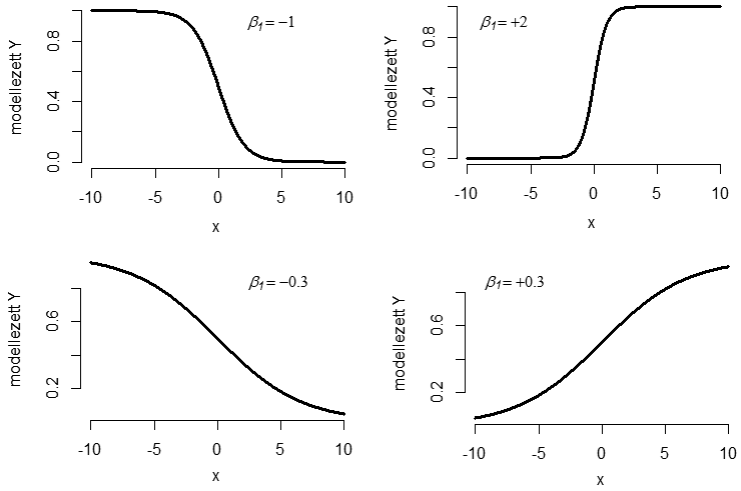
```
df = read.csv(file = "titanic.csv", header = T)
dim(df)
[1] 887 8
model = glm(Survived~Age, family = "binomial", data = df)
summary(model)
Call:
glm(formula = Survived ~ Age, family = "binomial", data = df)
Deviance Residuals:
 Min 1Q Median 3Q Max
-1.0864 -1.0017 -0.9439 1.3562 1.5806
Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.209189 0.159494 -1.312 0.1897
Age -0.008774 0.004947 -1.774 0.0761 .
---
(Dispersion parameter for binomial family taken to be 1)
 Null deviance: 1182.8 on 886 degrees of freedom
Residual deviance: 1179.6 on 885 degrees of freedom
AIC: 1183.6
Number of Fisher Scoring iterations: 4
```

A modell szerint sem a β_0 ($p = 0.1897$), sem a β_1 ($p = 0.076$) nem tér el szignifikánsan a nullától, tehát a modell nem tudja magyarázni a túlélést az életkor alapján. Ezzel ellentétben, ha az utasok besorolását választjuk magyarázó változónak (Pclass), ami három osztályt tartalmaz, a modell szignifikáns lesz.

```
model = glm(Survived~Pclass, family = "binomial", data = df)
summary(model)
Call:
glm(formula=Survived~Pclass, family="binomial", data=df)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4382  -0.7602  -0.7602   0.9374   1.6629
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  1.43907    0.20742   6.938 3.98e-12 ***
Pclass      -0.84423    0.08719  -9.683  < 2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 1182.8  on 886  degrees of freedom
Residual deviance: 1082.1  on 885  degrees of freedom
AIC: 1086.1
Number of Fisher Scoring iterations: 4
```

A modell szerint az, hogy egy utas milyen osztályon utazott a Titanicon, befolyásolta a túlélési esélyeit. A $p = 0.430$, vagyis a túlélési esély átlagosan 0.43-szor kisebb osztályról osztályra. Átlagosan 57%-kal (100%–43%) csökkent egy utas túlélési esélye, ha egy szinttel olcsóbb osztályon utazott.



8.8. ábra. Y *a valóságban csak 0 és 1 értéket vehet fel, a modellezett értékek nem lépik túl a [0,1] intervallumot*

A 8.8. ábra egyik grafikonjának a kódja:

```
library(manipulate)
x = seq(-10, 10, length = 1000)
manipulate(
 plot(x, exp(beta0 + beta1*x)/(1 + exp(beta0 + beta1*x)),
 type = "l", lwd = 3, frame = FALSE,
    ylab = "modellezett Y"),
 beta1 = slider(-2, 2, step = 0.1, initial = 2),
 beta0 = slider(-2, 2, step = 0.1, initial = 0)
)
```

8.2.2. A Poisson-regresszió

A Poisson-regresszió olyan diszkrét változóra alkalmazható, amely csak pozitív értékeket vehet fel, és varianciájára érvényes az, hogy egyenlő a középértékkel, $\text{var}(Y) = \mu$. A modell a következő komponenseket használja:

– szisztematikus komponens: $E(Y) = \mu$;

– lineáris prediktor: $\eta = \beta_0 + \beta_1 \cdot X$;

– kapcsolati függvény: $g(\mu) = \eta = \log \mu$.

Poisson-eloszlást olyan Y változókra alkalmazhatunk, amelyek kifejezik, hogy hány esemény következik be egy vizsgált területen vagy időpillanatban, hány esetből hány felel meg az előfeltételeknek, továbbá alkalmazható binomiális változóra is, amikor a siker valószínűsége kicsi, és az elemszám nagy, valamint kontingenciatáblázatokra is. Az utóbbinál a független változó egy kategorikus változó. Például egy adott területen felmérik egy faj egyedszámát napos, borús és esős időben egyaránt. A független változó folytonos skálán is mozoghat, például abban az esetben, ha a faj egyedszámát a levegő hőmérsékletének függvényében mérik fel.

A Poisson-regresszió alkalmazási feltételei:

1) A függő változó nullát vagy diszkrét pozitív értékeket vehet fel, és Poisson-eloszlást követ. Ez a feltétel sok esetben nem teljesül, mivel a variancia általában nagyobb az átlagnál.

2) A megfigyelések (Y értékei) egymástól függetlenek (egy esemény bekövetkezése nem befolyásolja egy másik esemény bekövetkezésének a valószínűségét).

A logisztikus regresszióhoz hasonlóan a Poisson-regresszió lehetőséget ad a $\hat{\beta}_1$ alapján arra, hogy felmérjük azt, hogy a folytonos prediktor változó egységnyi változására hány százalékban módosul a függő változó értéke. Kategorikus változó esetén a modell azt mutatja meg, hogy két kategória Y értékeinek eltérése mekkora. A $\hat{\beta}_1$ azt fejezi ki, mennyivel nő vagy csökken a függő változó átlagos előfordulási gyakoriságának a logaritmusa, ha a magyarázó változó értéke 1 egységgel növekszik. A $\hat{\beta}_1$ értéknek a reális skálán való értéke $e^{\hat{\beta}_1}$, a függő változó százalékos változását pedig az alábbi képlettel számolhatjuk ki:

$$\Delta Y(\%) = 100 \cdot \left(e^{\beta 1} - 1\right).$$

Például, ha a $\hat{\beta}_1$ értéke 0.45, akkor $e^{\hat{\beta}_1}$ értéke 1.57, tehát a magyarázó változó egységnyi növekedésére az Y változó értéke 57%-kal növekszik. Ha a $\hat{\beta}_1$ értéke −0.08, akkor $e^{\hat{\beta}_1}$ értéke 0.92, tehát az X változó egységnyi növekedésére az Y változó értéke 8%-kal csökken.

A modell hatékonysága a reziduális deviancia alapján értékelhető ki, ami megadja a különbséget a jelenlegi modell devianciája és az ideális modell maximális devianciája között. Az ideális modell minden megfigyelt értéket tökéletesen megjósol. A modell hatékonyságát χ^2-próbával tesztelhetjük az illesztés jóságára.

Ha a reziduális deviancia elég alacsony, akkor a χ^2-próba eredménye nem lesz szignifikáns, azaz a modell jól illeszkedik a megfigyelt értékekhez.

A Poisson-eloszlás alkalmazása korlátozott. Ha a variancia jóval nagyobb az átlagértéknél, akkor túlszóródásról beszélünk. Ilyen esetben a Poisson-eloszláson alapuló modell a koefficiensek konfidenciaintervallumait szűkebbre veszi, hiszen kisebb szórást feltételez, így szignifikáns eredményt adhat olyan esetre, ami a valóságban nem az. Erre az esetre vezették be a kvázi-Poisson-eloszlást, ami részben figyelembe veszi a nagyobb varianciát, és a *z*-eloszlás helyett a *t*-eloszlással értékeli ki a konfidenciaintervallumokat. A modell együtthatóinak az értéke nem változik, csupán szigorúbban ítéli meg a szignifikanciájukat.

Első példa: a *cyclones* adattábla a {GLMsData} csomagban az Ausztrália környékén észlelt trópusi ciklonok számát tartalmazza 1969 és 2005 között (N = 37). Az éves átlag 12, a variancia pedig 13.7, ami nagyjából megfelel a Poisson-eloszlásnak. A ciklonok száma tehát diszkrét változó, és ugyanúgy az év is. Ilyen esetben a Poisson-regresszió alkalmazható.

```
library(GLMsData)
data(cyclones)
d = cyclones
dim(d)
[1] 37 8
model_p = glm(Total~Year, family = "poisson", data = d)
summary(model_p)
Call:
glm(formula = Total~Year, family = "poisson", data = d)

Deviance Residuals:
 Min 1Q Median 3Q Max
-2.2639 -0.6484 0.1130 0.5929 1.9500

Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) 26.978363 8.871739 3.041 0.00236 **
Year -0.012331 0.004468 -2.760 0.00578 **
---
(Dispersion parameter for poisson family taken to be 1)

 Null deviance: 42.278 on 36 degrees of freedom
Residual deviance: 34.621 on 35 degrees of freedom
AIC: 197.25
Number of Fisher Scoring iterations: 4
```

A modell a 8.9. ábra bal oldalán látható. A tengelymetszet (Intercept) ~ 27, ami azt jelenti, hogy a nulladik évben elméletileg 27 ciklon volt Ausztrália közelében. Ez az érték szignifikánsan különbözik nullától (p = 0.002). Ennek az értéknek nincs jelentősége, mivel időben óriási extrapolációt jelent. A β_1 értéke –0.0123, aminek a valós skálán $e^{-0.0123}$ érték felel meg, azaz 0.988. Innen:

$$\Delta Y(\%) = 100 \cdot (0.988 - 1) = -1.2\%,$$

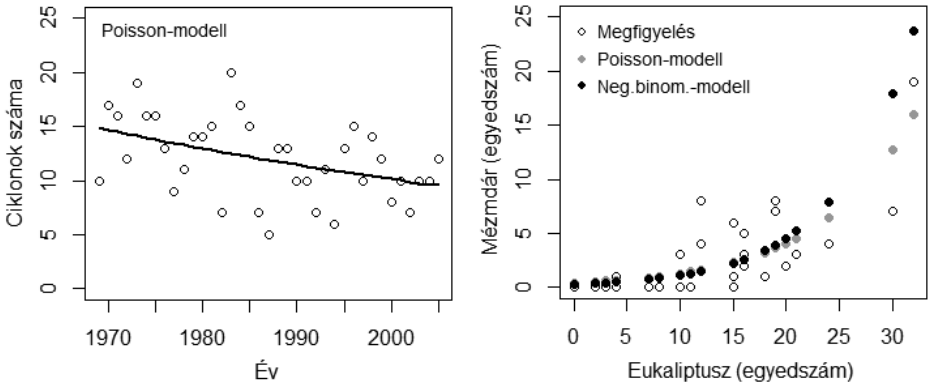
tehát a modell szerint 1969 és 2005 között évente 1.2%-kal csökkent a ciklonok gyakorisága. A β_1 értéke szignifikánsan meghatározza a modellt (p = 0.006), a konfidenciaintervallum pedig kiszámolható R-ben:

```
#beta1 értéke a valós skálán
exp(model_p$coeff[2])
 Year
0.9877445
#beta1 konfidenciaintervalluma a valós skálán
exp(confint(model_p))
 2.5 % 97.5 %
(Intercept) 1.507529e+04 1.941574e+19
Year 9.791077e-01 9.964146e-01
```

A β_1 értéke tehát 0.988 (0.979 – 0.996, $\alpha = 0.05$). A modell érvényességének és megbízhatóságának kiértékelését két χ^2-próbával végezhetjük el: az egyik az ANOVA alapján megállapítja, hogy a modell eltér-e szignifikásan a véletlenszerű eloszlástól, a másik pedig a Poisson-regresszió megbízhatóságát ellenőrzi úgy, hogy a modell által megjósolt *Y*-értékeket összeveti a megfigyelt *Y*-értékekkel.

```
#A modell érvényességének tesztelése
anova(model_p, test = "Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: Total
Terms added sequentially (first to last)
 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL 36 42.278
Year 1 7.6567 35 34.621 0.005656 **
---
Signif.codes:0‘***’ 0.001‘**’ 0.01‘*’ 0.05‘.’ 0.1‘ ’

#A Poisson-regresszió megbízhatósága
pchisq(34.621,35, lower.tail = FALSE)
[1] 0.4862805
```

Az ANOVA alapján a modell érvényes ($\chi^2 = 34.6$, $df = 35$, $p = 0.006$), és megbízhatóan jelzi a ciklonok várható számát (a χ^2-próba nem adott szignifikáns eltérést a megfigyelt és a modellezett értékek között).



8.9. ábra. *A* cyclones *és az* nminer *adattáblák alapján elvégzett általánosított lineáris modellek viszonya a megfigyelt értékekhez*

A 8.9. ábra R-kódja

```
par(mfrow=c(1,2)) #összetett ábra szerkesztése
#a
plot(d$Year,model_p$fitted.values,pch=19,col="black",
 xlab="Év",ylab="Ciklonok száma",
 ylim = c(0,25), type = "l", lwd = 2)
points(d$Total~d$Year)
text(1976,24, "Poisson-modell", cex = 0.9)
#b
plot(dt$Eucs,model_pois$fitted.values,pch=19,
 col="darkgrey",xlab="Eukaliptusz (egyedszám)",
 ylab="Mézmadár (egyedszám)", ylim = c(0,25))
points(dt$Minerab~dt$Eucs)
points(dt$Eucs,model_nb$fitted.values,pch=19,col="black")
legend("topleft", pch = c(1,16,16), cex = 0.9, bty = "n",
 col = c("black", "darkgrey","black"),
 legend = c("Megfigyelés","Poisson-modell",
                    "Neg.binom.-modell"))
```

Második példa: az *nminer* adattábla {GLMsData} csomagban a feketehomlokú mézmadár (*Manorina melanocephala*) egyedszámát tartalmazza Délkelet-Ausztrália erdős térségéből (Buloke Woodlands). A felmérést kéthektáros transzektek mentén végezték, transzektenként 3 x 20 perces megfigyelés során. Összesen 31

transzektet mértek fel; a mézmadár átlagos egyedszáma 2.7, varianciája pedig 16.4 volt. Ebből a két adatból látszik, hogy túlszórás áll fenn, így a Poisson-eloszlásra épülő modell nem megbízható. A legvalószínűbb oka a túlszórásnak az, hogy a tesztterületeken a különféle tényezők eltérően hatnak a madár egyedszámára. Ezért különböző szempontok szerint alcsoportokat kellene létrehozni egy rétegzett mintavétel keretében. Jelen esetben két lehetőség van: a modellt vagy a kvázi-Poisson-eloszlásra ($\Phi \neq 1$), vagy a negatív binomiális eloszlásra építeni. A negatív binomiális eloszlásnak nincs olyan feltétele, hogy az átlagérték egyenlő legyen a varianciával, ezért a ráépülő modell is sikeresebb lehet.

```
library(GLMsData)
data(nminer)
dt = nminer
dim(dt)
[1] 31 8

#Kvázi-Poisson-eloszláson alapuló modell
model_qpois = glm(Minerab~Eucs, family = "quasipoisson", data
= nminer)
summary(model_qpois)
Call:
glm(formula = Minerab ~ Eucs, family = "quasipoisson", data
= nminer)
Deviance Residuals:
 Min 1Q Median 3Q Max
-2.1454 -1.2530 -0.9673 0.5634 3.5603
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.87621 0.43149 -2.031 0.0515 .
Eucs 0.11398 0.01897 6.009 1.55e-06 ***
---
(Dispersion parameter for quasipoisson family taken to be
2.328067)
 Null deviance: 150.545 on 30 degrees of freedom
Residual deviance: 63.318 on 29 degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 5

#A modell érvényessége
anova(model_qpois, test = "Chisq")
Analysis of Deviance Table
Model: quasipoisson, link: log
```

```
Response: Minerab
Terms added sequentially (first to last)
 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL 30 150.545
Eucs 1 87.227 29 63.318 9.294e-10***

#A modellezett értékek illeszkedése a megfigyelt értékekre
pchisq(63.318,29, lower.tail = FALSE)
[1] 0.0002346215

#Negatív binomiális eloszláson alapuló modell
library(MASS)
model_nb = glm.nb(Minerab~Eucs, data = dt)
summary(model_nb)
Call:
glm.nb(formula = Minerab ~ Eucs, data = dt, init.theta =
1.718868357, link = log)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6981  -0.9661  -0.7248   0.3468   2.3120
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.25960    0.44978  -2.800   0.0051**
Eucs         0.13826    0.02477   5.582 2.37e-08***
---
(Dispersion parameter for Negative Binomial(1.7189) family
taken to be 1)
    Null deviance: 66.706  on 30  degrees of freedom
Residual deviance: 31.353  on 29  degrees of freedom
AIC: 113
Number of Fisher Scoring iterations: 1

#A modell érvényessége
anova(model_nb, test = "Chisq")
Analysis of Deviance Table
Model: Negative Binomial(1.7189), link: log
Response: Minerab
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    30     66.706
Eucs  1   35.353        29     31.353 2.75e-09***
```

```
#A modellezett értékek illeszkedése a megfigyelt értékekre
pchisq(31.353,29, lower.tail = FALSE)
[1] 0.3489757
```

A két modell által megjósolt *Y*-értékek a 8.9. ábra jobb oldalán láthatók. Noha az ábrán a két modell között nincs jelentős eltérés, a statisztikai mutatókban különböznek, különösképpen a reziduális devianciák terén. A kvázi-Poisson-eloszlásnál ez az érték 63.3 ($df = 29$), a negatív binomiális eloszlásnál pedig 31.4 ($df = 29$). Kizárólag az utóbbinál érvényesül az, hogy a reziduális deviancia megközelíti a szabadságfokok számát. Amíg mind a két modell ANOVA-tesztje szignifikáns, addig a modellezett és a megfigyelt értékek illeszkedése csak a negatív binomiális eloszlás esetén tekinthető jónak. A két eloszlásra épülő modell adatai a 8.4. táblázatban figyelhetők meg.

8.4. táblázat. *A két modell mutatóinak az összehasonlítása az* nminer *adattábla adataira (magyarázó változó: eukaliptuszok száma, függő változó: mézmadár abundanciája)*

Mutató	Kvázi-Poisson	Negatív binomiális
$\hat{\beta}_1$	0.114	0.138
$e^{\hat{\beta}_1}$	1.121 (1.091−1.220)	1.148 (1.080−1.164)
átlagos $\triangle$Y (%)	12.1	14.8
reziduális deviancia	63.318	31.353
Teta	2.328	1.719
szabadságfokok	29	29
ANOVA-próba	$p < 0.001$	$p < 0.001$
χ^2-próba	$p = 0.0002$	$p = 0.349$
megbízható	nem	igen

A negatív binomiális eloszlásra alapuló modellben mind a tengelymetszet, mind a meredekség szignifikánsan meghatározza a modellt: a β_0 és a β_1 koefficienseknek a nullától való eltérése egyaránt szigifikáns ($p = 0.005$, ill. $p < 0.001$). A tengelymetszet negatív értékét nem lehet átültetni a valóságba, mivel az azt fejezi ki, hogy ha a területen nem lennének eukaliptuszfák, akkor a mézmadár egyedszáma -1 lenne. Az érték egyrészt közel áll a nullához, másrészt pedig a modell szerint a mézmadár nem lesz megfigyelhető még akkor sem, ha már csak pár eukaliptuszfa van a területen.

Ha egy esemény bekövetkezése növeli egy másik esemény bekövetkezésnek a valószínűségét, akkor nagyobb lesz a magas és az alacsony értékek előfordulási gyakorisága, ennek eredményeként pedig emelkedik a változó szórása. Az eloszlást ebben az esetben túlszorás jellemzi, ezért nem felel meg a Poisson-eloszlásnak. Ilyen esetben a negatív binomiális eloszlásra alapuló modell sikeresebb.

HÁROMNYELVŰ SZAKKIFEJEZÉSEK

Magyar	Román	Angol
adatsor	şir de date	data series
alternatív hipotézis	ipoteza alternativă	alternative hypothesis
arány	proporţie	ratio
arányskála	scală de raport (proporţii)	ratio scale
állandó	constantă	constant
általánosított lineáris modellek	modele liniare generalizate	generalized linear models
asszociáció	asociere	association
átlag	media	mean
átlagos négyzetes eltérés	eroare pătrată medie	mean squared error (MSE)
bal oldali próba	test unilateral stâng	left-tailed test
beágyazott ANOVA	ANOVA mixtă	mixed ANOVA
becslés	estimare	estimation
bináris változó	variabilă binară	binary variable
centrális határeloszlás tétel	teorema tendinţei centrale	central limit theorem
csoportok közötti négyzetösszeg	suma pătratelor dintre grupe	between-group sum of squares (BSS)
csoportokon belüli négyzetösszeg	suma pătratelor din interiorul grupelor	within-group sum of squares (WSS)
csúcsosság	indice de aplatizare	kurtosis
determinációs együttható	coeficientul de determinare	coefficient of determination
diszkrét eloszlás	distribuţie discretă	discrete distribution
diszkrét változó	variabilă discretă	discrete variable
egyenletes eloszlás	distribuţie uniformă	uniform distribution
egymintás t-próba	testul t pentru un singur eşantion	one-sample t-test
egyoldalú próba	test unilateral	one-tailed test
egyszempontos ANOVA	ANOVA unifactorială	one-way ANOVA
együttható	coeficient	coefficient
elemszám	dimensiune / volum	sample size
elméleti eloszlás	distribuţie teoretică	theoretical distribution
előjelpróba	testul semnului	sign-test

Magyar	Román	Angol
előtanulmány	studiu pilot	pilot study
elsőfajú hiba	eroare de tip I	type I error
esettanulmány	studiu de caz	case study
esélyhányados	riscul relativ	odds ratio
faktor (független változó)	factor (variabilă independentă)	factor (indipendent variable)
ferdeség	indice de asimetrie	skewness
fő hatás	efect principal	main effect
független változó	variabilă independentă	independent variable
függő változó	variabilă dependentă	dependent variable
függetlenség tesztelése	test de independenţă	test of independence
gyakoriság	frecvenţă	frequency
gyök átlagos négyzetes eltérés	eroarea rădăcinii medie pătrată	root mean square error (RMSE)
hatásnagyság	mărimea efectului	effect size
heteroszkedaszticitás	heteroscedascititate	heteroscedasticity
hibatag	eroare de estimare	error term
hipotézis	ipoteză	hypothesis
hisztogram	histogramă	histogram
illesztés jósága	(test de) concordanţă	goodness of fit
interakció	interacţiune	interaction
interkvartilis tartomány	interval intercuartilic	interquartile range
intervallumskála	scală de interval	interval scale
ismételt méréses ANOVA	ANOVA cu măsurători repetate	repeated measures ANOVA
jobb oldali próba	test unilateral dreapta	right-tailed test
kapcsolt értékek	valori legate	ties
kereszttábla / kontingenciatáblázat	tabel de contingenţă	contingency table
kétmintás t-próba	testul t pentru eşantioane independente	two sample t-test
kétoldalú próba	test bilateral	two-tailed test
kétszempontos ANOVA	analiza dispersională bifactorială	two-way ANOVA
kiugró érték	valoare aberantă	outlier
kísérleti kutatás	studiu experimental	experimental study

Magyar	Román	Angol
konfidenciaintervallum	interval de încredere	confidence interval
kontrollcsoport	grup martor (de control)	control group
korreláció	corelaţie	correlation
korrelációs együttható	coeficient de corelaţie	correlation coefficient
középtendencia	tendinţă centrală	central tendency
következtető statisztika	statistică inductivă	inferential statistics
kvantilis	quantilă	quantile
kvartilis	quartilă	quartile
legkisebb négyzetek módszere	regula celor mai mici pătrate	least squares method
leíró statisztika	statistică descriptivă	descriptive statistics
lineáris regresszió	regresie liniară	linear regression
magyarázó változó	variabilă explicativă	explanatory variable
másodfajú hiba	eroare de tip II	type II error
medián	mediană	median
megfigyeléses kutatás	studiu observaţional	observational study
mennyiségi változó	variabilă cantitativă	quantitative variable
minta	eşantion	sample
mintavételi egység	unitatea experimentală	sampling unit
mintavételi stratégia	program de colectare a probelor (eşantionare)	sampling design
minőségi változó	variabilă calitativă	qualitative variable
módusz	mod	mode
nemparametrikus próba	test neparametric	nonparametric test
nominális skála	scală nominală (de clasificare)	nominal scale
normál QQ-ábra	diagrama QQ-normal	QQ-normal plot
normális eloszlás	distribuţia / repartiţia normală	normal distribution
nullhipotézis	ipoteza nulă	null hypothesis
ordinális skála	scală ordinală (de rang)	ordinal scale
oszlopdiagram	diagrama cu bare	barplot
páros t-próba	testul t pentru eşantioane perechi	paired t-test
p-érték (valószínűségi érték)	valoare de probabilitate	probability value

Magyar	Román	Angol
populáció	populaţie	population
post-hoc analízis (páronkénti összehasonlítás)	analiza post-hoc (comparaţii multiple)	post-hoc analysis (multiple comparisons)
próbastatisztika	statistica testului	test statistics
regresszió a középértékre	regresie către medie	regression to the mean
reprezentatív	reprezentativ	representative
reziduális	reziduu	rezidual
rétegzett mintavétel	eşantionare stratificată	stratified sampling
standard hiba	eroare standard	standard error (SE)
standardizálás	standardizare	standardization
szabadságfok	grad de libertate	degree of freedom
szignifikanciaszint	prag de semnificaţie	significance level
szórás (standard deviáció)	abatere standard	standard deviation
szórásdiagram	diagrama de dispersie	scatterplot
szóródás	împrăştiere	spread
terjedelem	amplitudine (domeniu)	range
teszt (kísérleti) csoport	grup experimental	experimental group
teszt ereje	puterea testului	power of a test
trimmelt átlag	media redusă	trimmed mean
túlszórás	supradisperise	overdispersion
valószínűségi változó	variabilă aleatoare	random variable
variancia	dispersie (varianţă)	variance
varianciaanalízis	analiza dispersională	analysis of variances
változó	variabilă	variable
várt érték	valoare aşteptată	expected value
véletlen mintavétel	eşantionare aleatorie	random sampling

BIBLIOGRÁFIA

Algina, J., Keselman, H. J., Penfield, R. D. (2005). An Alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317–328.

Allen, M. (2017). *The SAGE Encyclopedia of Communication Research Methods*. SAGE Publications, Los Angeles, USA. p. 1295.

Bartolucci, A. A., Tendera, M., Howard, G. (2011). Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *American Journal of Cardiology*, 107, 1796–801.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B*, 57, 289–300.

BirdLife International. (2015). *European Red List of Birds*. Office for Official Publications of the European Communities, Luxembourg.

Bonini, S., Corain, L., Marozzi, M., Salmaso, L. (2014). *Nonparametric Hypothesis Testing. Rank and Permutation Methods with Applications in R*. John Wiley & Sons Ltd, Chichester, UK.

Bowman, K. O., Shenton, B. R. (1975). Omnibus test contours for departures from normality based on b_2. *Biometrika*, 64, 243–250.

Brandl, S., Paul, C., Knoke, T., Falk, W. (2020). The influence of climate and management on survival probability for Germany's most important tree species. *Forest Ecology and Management*, 458, 117652.

Brown, M. B., Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 129–132.

Caffo, B. (2015). *Regression Models for Data Science in R*. Lean Publishing.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press. New York, USA.

Crampton, E. W. (1947). The growth of the odontoblast of the incisor teeth as a criterion of vitamin C intake of the Guinea pig. *The Journal of Nutrition*, 33, 491–504.

Crawley, M. J. (2013). *The R Book*. Second ed. Wiley, London, UK.

Eichstaedt, K. E., Kovatch, K., Maroof, D. A. (2013). A less conservative method to adjust for familywise error rate in neuropsychological research: The Holm's sequential Bonferroni procedure. *NeuroRehabilitation*, 32, 693–696.

Eidenbenz, D., Techel, F., Kottmann, A., Rousson, V., Carron, P. N., Albrecht, R., Pasquier, M. (2021). Survival probability in avalanche victims with long burial *Resuscitation*, 166, 93–100.

Emura, T., Liao, Y. T. (2017). Critical review and comparison of continuity correction methods: the normal approximation to the binomial distribution, *Communications in Statistics – Simulation and Computation*, 47, 2266–2285.

Emura, T., Lin, Y. S. (2015). A comparison of normal approximation rules for attribute control charts. *Quality and Reliability Engineering International*, 31, 411–418.

Erce-Hurn, D. M., Mirosevich, V. M. (2008). Modern Robust Statistical Methods: An Easy Way to Maximize the Accuracy and Power of Your Research. *American Psychologist*, 63, 591–601.

Fagerland, M. W., Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials*, 30, 490–496.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies – A paradox of statistical practice? *BMC Medical Research Methodology*, 12, 78.

Gad, S. C. (2010). Statistical Methods in Toxicology. In: McQueen, C. A. (ed.) *Comprehensive Toxicology*. Elsevier, p. 183–197.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

Gashemi, A., Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, 10, 486–489.

Gheoca, V., Benedek, A. M., Schneider, E. (2021). Exploring land snails' response to habitat characteristics and their potential as bioindicators of riparian forest quality. *Ecological Indicators*, 132, 108289.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.

Hintze, J. L. (2008). Power analysis and sample size system (PASS) for Windows User's Guide. NCSS, Kaysville, USA.

INS, Institutul Naţional de Statistică (2021). *Romania in Figures*. Bucureşti.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38, 324−329.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67, 85–92.

Jonckheere, A. R. (1954). A distribution-free k-sample test again ordered alternatives. *Biometrika*, 41, 133−145.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30, 81–93.

Keselman, H. J., Wilcox, R. R., Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophyisiology*, 40, 586–596.

Keselman, H. J., Miller, C. W., Holland, B. (2011). Many tests of significance: New methods for controlling type I errors. *Psychological Methods*, 16, 420–431

Klein, G., Dabney, A. (2013). The Cartoon Introduction to Statistics. New York: Hill and Wang.

Kruskal, W. H., Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583–621.

Legendre, P., Gallagher, E. D. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271–280.

Lix, L. M., Keselman, H. J. (1995). Approximate Degrees of Freedom Tests: A Unified Perspective on Testing for Mean Equality. *Psychological Bulletin*, 117, 547–560.

Lix, L. M., Keselman, J. C., Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the oneway analysis of variance F test. *Review of Educational Research*, 66, 579–619.

Moksony, F. (2006). A Poisson-regresszió alkalmazása a szociológiai és demográfiai kutatásban. *Demográfia*, 4, 366–382.

Nedler, J. A., Wedderburn, R. W. M. (1972). Generalized linear modells. *Journal of the Royal Statistical Society. Series A*, 135, 370–384.

Păun, A., Păun, M. (2009). *Analiza statistică folosind limbajul R*. Ed. Matrix Rom, Bucureşti.

Păun, M. (2016). *Noţiuni de statistică aplicată cu exemple în R*. Ed. Matrix Rom, Bucureşti.

Reiczigel, J., Harnos, A., Solymosi, N. (2007). *Biostatisztika nem statisztikusoknak*. Pars. Kft., Nagykovácsi.

Rufibach, K. (2011). Assessment of paired binary data. *Skeletal Radiology*, 40, 1–4.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, 17, 688–90.

Sakoda, J. M. (1977). Measures of association for multivariate contingency tables. *Proceedings of the Social Statistics Section of the American Statistical Association (Part III)*, 777–780.

Salkind, N. J. (2010). *Encyclopedia of Research Design*. SAGE Publications, London, UK.

Serdar, C. C., Cihan, M., Yücel, D., Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, 31, 010502.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27, 799–811.

Sullivan, G. M., Feinn, R. (2012). Using effect size – or Why the p value is not enough. *Journal of Graduate Medical Education*, 4, 279–282.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14, 327−333.

Tsagris, M., Pandis, N. (2021). Normality test: Is it really necessary? *American Journal of Orthodontics and Dentofacial Orthopedics*, 159, 548–549.

Vargha, A. (2015). *Matematikai statisztika pszichológiai, nyelvészeti és biológiai alkalmazásokkal*. Második kiadás, Pólya Kiadó, Budapest.

Villacorta, P. J. (2017). The welchADF package for robust hypothesis testing in unbalanced multivariate mixed models with heteroscedastic and non-normal data. *R Journal*, 9, 309–328.

Visscher, P. M., Medland, S. E., Ferreira, M. A., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., Martin, N. G. (2006). Assumption-free estimation of herita-

bility from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, 2, e41.

Wang, S., Wu, W., Liu, F., Liao, R., Hu, Y. (2017). Accumulation of heavy metals in soil-crop systems: a review for wheat and corn. *Environmental Science and Pollution Research*, 24, 15209–15225.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350–362.

Welch, B. L. (1951). On the comparison of several means: An alternative approach. *Biometrika*, 38, 330–336.

Westfall, P. H., Young, S. S. (1993). *Resampling based multiple testing*. Wiley, New York, USA.

Wicherts, J. M, Dolan, C. V., Hessen, D. J. (2005). Stereotype Threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, 89, 696–716.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology*, 41, 109–117.

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika*, 61, 165–170.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York, USA, 199.

Zsigmond, A. R., Varga, K., Kántor, I., Urák, I., May, Z. Héberger, K. (2018). Elemental composition of wild growing Agaricus campestris mushroom in urban and peri-urban regions of Transylvania (Romania). *Journal of Food Composition and Analysis*, 72, 15–21.

REZUMAT

Procesarea datelor cu R. Aplicaţii în ştiinţa mediului

Cartea este destinată studenţilor care au ca specializare ştiinţa mediului. Este un ghid practic, care îi ajută să se familiarizeze cu programul R. Acesta este un software gratuit larg utilizat de comunitatea ştiinţifică. Cauza principală pentru care a devenit atât de popular este paleta largă de operaţiuni statistice pe care o oferă utilizatorilor. Prin pachetele care se găsesc pe reţeaua de internet şi se pot instala gratuit, R oferă posibilităţi nelimitate în procesarea statistică a datelor.

Cartea conţine opt capitole şi începe cu informaţii despre bazele statisticii şi despre proiectarea unei cercetări ştiinţifice în domeniul ştiinţei mediului. Capitolul al doilea prezintă limbajul R şi principalele comenzi de tratare a datelor şi editare a diagramelor simple precum histograma, diagrama prin puncte, diagrama cu bare. Capitolele de la trei la şapte introduc informaţii despre distribuţia variabilelor continue şi discrete şi tratează detaliat ipotezele statistice. Această secţiune reprezintă centrul cărţii, şi oferă cititorului o vaste cunoştinţe în procesarea statistică a datelor de mediu într-o manieră ştiinţifică. Capitolul al optulea este destinat analizei de regresie. Sunt prezentate bazele regresiei liniare şi a modelelor liniare generalizate. Aceste metode sunt limitate doar la o singură variabilă independentă.

Cartea se bazează pe o bibliografie amplă şi conţine referiri la articole ştiinţifice şi cărţi în trei limbi (maghiară, română şi engleză). Pentru studenţii care vor să aprofundeze limbajul statistic în engleză, este recomandată cartea *The R Book* (Crawley, 2013), iar pentru limba română sunt recomandate cărţile: *Analiza statistică folosind limbajul R* (Păun Păun, 2009) şi *Noţiuni de statistică aplicată cu exemple în R* (Păun, 2016).

Pentru începători este recomandat să studieze cartea în mod cronologic deoarece informaţiile sunt construite logic. Pentru cei care sunt familiarizaţi cu procesarea statistică a datelor, această carte poate fi utilizat ca un îndreptar pentru a găsi soluţiile cele mai eficiente în R pentru problema dată.

ABSTRACT

Data Processing with R. Applications in the Environmental Sciences

This book is designed for students involved with environmental science studies. It is a useful guide for them to become acquainted with R, which is a free software for statistical analyses and data processing. Today, R is accepted and widely used in the scientific community. The main reason of its popularity is the boundless possibilities R gives for users through the packages developed by statisticians all over the world.

The book contains eight chapters, at first providing information about the basics of statistics and experimental design in the environmental sciences. The following chapters present the language of R and the main commands for handling a data frame and constructing simple plots such as histograms, box plots, or scatter plots. Through chapters three to seven, the distributions of continuous and discrete data and the most relevant hypothesis testing procedures are detailed. This part is the core of the book, which provides an expansively treated knowledge for the reader to feel themselves confident in processing experimental data with R in a scientific manner. The eighth chapter is dedicated to regression analysis. This chapter presents the linear regression and the generalized linear models confined to one independent variable.

The book relies on a comprehensive bibliography containing Hungarian, English, and Romanian scientific articles and books. For students who want to deepen their English statistical language, *The R Book* (Crawley, 2013) is highly recommended. There are also some books written in Romanian language that are also recommended: *Analiza statistică folosind limbajul R* (Păun Păun 2009) and *Noţiuni de statistică aplicată cu exemple în R* (Păun 2016).

For learners, it is recommended to pass through the chapters chronologically because the knowledge the book gives is logically constructed. For those who are familiarized with the statistical tools, the book can be used as a handbook for finding a useful solution in R for the given problem.

A SZERZŐRŐL

Zsigmond Andrea-Rebeka Kolozsváron született 1978-ban, tanulmányait az Apáczai Csere János Líceumban végezte 1997-ben matematika–fizika szakon. Egyetemi tanulmányait a Babeş–Bolyai Tudományegyetem Kémia és Vegyészmérnöki Karán folytatta 1998–2002 között, majd ugyanott mesterizett és doktorált is. Doktori dolgozatát az atomspektroszkópiai módszerek fejlesztésének témakörében írta. 2004-től a Sapientia Erdélyi Magyar Tudományegyetem oktatója, ahol jelenleg egyetemi adjunktusként dolgozik a Kolozsvári Kar Környezettudományi Tanszékén. Kutatási területe a környezeti minták elemanalízise spektroszkópiai módszerekkel, emellett pedig 2015 óta továbbképzéseken vett részt a statisztikai adatfeldolgozás területén, különös tekintettel az R számítógépes program kezelésére. Tudományos publikációi elsősorban a város hatásának vizsgálatára összpontosulnak különféle növények és gombák elemösszetételére, illetve a székelyföldi természetes ásványvízforrások kémiai elemzésére.

Scientia Kiadó
400112 Kolozsvár (Cluj-Napoca)
Mátyás király (Matei Corvin) u. 4. sz.
Tel./fax: +40-364-401454
E-mail: scientia@kpi.sapientia.ro
www.scientiakiado.ro

Korrektúra:
Szabó Beáta

Műszaki szerkesztés:
Metaforma Kft.

Tipográfia:
Könczey Elemér

Nyomdai munkálatok:
F&F INTERNATIONAL Kft.
Felelős vezető: Ambrus Enikő igazgató

A könyv segítséget nyújt az adatfeldolgozásban kevésbé jártas hallgatóknak abban, hogy megértsék a legalapvetőbb statisztikai próbák működését, jól megtervezett környezettudományos kutatásokat végezhessenek, és értelmezni tudják a nemzetközi tudományos közlemények eredményeit. Ugyanakkor hasznos útmutató az R statisztikai szoftvercsomaghoz, amely szabadon hozzáférhető, és amelyet egyre szélesebb körben használnak a tudományos világban, hiszen tág lehetőségeket kínál a legkülönfélébb adatelemzésekhez.

A könyv nyolc fejezetre tagolódik, és az adatfeldolgozás alapfogalmainak tisztázása után útmutatót ad a tudományos kutatás megtervezésénél kulcsszerepet játszó szempontokhoz, részletezi az R nyelvezetét, kitér néhány egyszerű ábra szerkesztésére, és megismerteti a környezettudományban használt leggyakoribb hipotézisvizsgálatokat.

ISBN 978-606-975-068-1