

## 1.6. Adatminták és rögzítésük a ROPstat programcsomagban

Az empirikus vizsgálatokban a statisztikai elemzéseket nem közvetlenül a populációból véletlenszerűen kiválasztott mintán, hanem az ahhoz tartozó **adatmintán** végezzük. A kiválasztott minta ugyanis az empirikus vizsgálat megfigyelési egységeinek (pl. személyeknek vagy más eseteknek) az együttese, az adatminta pedig a megfigyelési egységek leírására felhasznált változók értékeinek együttese. Az adatmintában tehát egy személyhez mindig egy egész adatlista tartozik, mely annyi adatból áll, amennyi változó jellemző egy személyt. Ha egy vizsgálatban például minden személynek csak a nemét, az életkorát és az iskolázottságát jegyezzük fel (ez utóbbit mondjuk a végzett osztályok számával mérve), akkor itt az adatminta minden személyéhez három adat tartozik. Tekintve, hogy statisztikai elemzéseket csakis adatmintákon lehet végezni, könyvünk további részében – ha ez nem okoz majd félreértést – *adatminta* helyett az egyszerűbb *minta* szóhasználatot részesítjük előnyben.

A számítógépes statisztikai adatfeldolgozás első lépése mindig a számítógépes adat-rögzítés. Ez sokféle szövegszerkesztővel, a ROPstat programcsomagon belül például az indítás után megjelenő adattábla kellő méretűre hozásával (a szükséges számú sor és oszlop beillesztésével), majd kitöltésével végezhető el. A számítógépes adatfelvitel illusztrálására az alábbiakban bemutatunk egy konkrét példát.

Egy 117 sportolóval végzett vizsgálatban (Vincze Pál vizsgálata) többek között a következő változókra vonatkozó adatok álltak rendelkezésre: üzőtt sportág (*Sportág*, értékek: 1 = tollaslabda, 2 = cselgáncs, 3 = súlyemelés), edzés után mért pulzus (*Pulzus*), edzés után mért szisztolés vérnyomás (*Szisztol*), edzés után mért diasztolés vérnyomás (*Diasztol*), valamint a Vargha-féle Szondi-skálák (vö. Vargha, 1994) közül a Dominancia-preferencia (*Dom\_pref*), az Aktivitás-preferencia (*Akt\_pref*), a Hosztilitás-preferencia (*Hos\_pref*), a Feminitás-preferencia (*Fem\_pref*) és az Instabilitás skála (*Instab*). E személyek közül kiválasztottuk a nőket, akik 25-en voltak, s adataikat – azonosító kódszámukkal együtt – külön táblázatba foglaltuk (lásd 1.3. táblázat). Ez egy ún. **eset-változó adattáblázat**, melynek soraiiban az egyes esetek (személyek), oszlopaiban pedig az egyes változók (*Kódszám*, *Pulzus*, *Szisztol* stb.) helyezkednek el. Ez az alaptípusa a nyersadatok táblázatának. Bármilyen empirikus vizsgálatot is végzünk, az adatokat első lépésben mindig ilyen formájú táblázatban kell elhelyezni. Ebben a vizsgálatban egyes személyeknek hiányoznak egyes adatai (pl. azért, mert nagy ritkán elromlott egy műszer). Ezeket a táblázatban üres rovat jelzi. Itt jegyezzük meg, hogy a ROPstatban nemcsak numerikus, hanem szöveges adatok (pl. nevek) is felvihetők, ha az illető változót szöveges típusúnak definiáljuk. A program azonban csak numerikus változók feldolgozását teszi lehetővé, ezért a ROPstatban az elemzésbe bevonni tervezett változók értékeit numerikus kóddal kell jelölni (például a *Személy neve* változó esetében alkalmazható a szokásos 1 = férfi, 2 = nő jelölés, a *Sportág* változó esetében pedig lásd a fenti kódolást).

Az eset-változó adattáblázat téglalap alakú, s két fő jellemzője a sorok (személyek, esetek) száma, mely a mi esetünkben 25, valamint az oszlopok (változók) száma, mely esetünkben 10.

Az adatfelvitel eredményeként olyan adatfájlhoz jutunk, amelyen a későbbiekben a ROPstat segítségével statisztikai elemzéseket tudunk végezni. Ez a ROPstatra specifikus, ún. **MSW-fájl**. Az MSW-fájlokat csak a ROPstat tudja közvetlenül elolvasni, de a ROPstat ezeket igen egyszerűen át tudja konvertálni olyan adatfájlokká, amelyek több más statisztikai programban (például Excel, SPSS stb.) is beolvashatók.

**1.3. táblázat.** Egy adattáblázat (a táblázatban az üres rovat az adat hiányzó voltára utal)

| Kód-szám | Sportág | Pulzus | Szisztol | Diasztol | Dom_<br>pref | Akt_<br>pref | Hos_<br>pref | Fem_<br>pref | Instab |
|----------|---------|--------|----------|----------|--------------|--------------|--------------|--------------|--------|
| 5        | 1       |        | 120      | 73       | 57           | 47           | 40           | 47           | 51     |
| 6        | 1       | 77     | 127      | 73       |              |              |              |              |        |
| 7        | 1       | 76     | 120      | 73       | 62           | 57           | 45           | 43           | 46     |
| 8        | 1       | 83     | 115      | 60       | 29           | 46           | 58           | 59           | 61     |
| 9        | 1       | 132    | 136      | 93       | 41           | 51           | 61           | 36           | 57     |
| 10       | 1       | 64     | 122      | 74       |              |              |              |              |        |
| 17       | 1       | 85     | 141      | 91       | 35           | 50           | 47           | 53           | 50     |
| 25       | 1       | 116    | 126      | 88       | 70           | 59           | 39           | 38           | 44     |
| 26       | 1       | 71     | 132      | 71       | 50           | 50           | 64           | 43           | 56     |
| 30       | 1       | 112    | 129      | 82       |              |              |              |              |        |
| 31       | 1       | 89     | 123      | 78       | 55           | 47           | 63           | 50           | 56     |
| 32       | 1       | 141    | 148      | 83       | 56           | 51           | 57           | 50           | 53     |
| 33       | 1       | 124    | 144      | 78       |              |              |              |              |        |
| 37       | 1       | 109    | 140      | 83       | 62           | 54           | 39           | 51           | 52     |
| 51       | 1       | 163    | 133      | 88       | 43           | 34           | 56           | 57           | 43     |
| 52       | 1       | 117    | 124      | 83       | 61           | 50           | 46           | 50           | 34     |
| 59       | 1       | 110    | 129      | 63       | 56           | 53           | 55           | 57           | 47     |
| 60       | 2       | 119    | 164      | 82       | 61           | 52           | 37           | 38           | 52     |
| 61       | 1       | 122    | 130      | 78       | 41           | 45           | 73           | 71           | 46     |
| 82       | 2       | 97     | 140      | 82       | 47           | 49           | 46           | 29           | 41     |
| 88       | 2       | 56     | 109      | 60       | 59           | 68           | 54           | 42           | 47     |
| 89       | 2       | 105    | 107      | 65       | 36           | 53           | 61           | 39           | 57     |
| 105      | 2       | 83     | 118      | 74       |              |              |              |              |        |
| 114      | 3       | 68     | 125      | 87       | 47           | 50           | 53           | 46           | 49     |
| 116      | 3       | 79     | 135      | 89       | 61           | 65           | 44           | 36           | 53     |

Az. 1.3. táblázat adatait például a következőképpen írhatjuk egy MSW-fájlba.

1. Indítsuk el a ROPstat programcsomagot! Az indítás után megjelenik a ROPstat adattáblázata, melynek kezdetben egyetlen sora (indexe: 1) és egyetlen oszlopa (neve: var\_1) van. A táblázat sorait a képernyő felső részén látható „Esetek” gomb (parancsikon) „Új esetek beillesztése” menüpontja, oszlopait pedig a „Változók” parancsikon „Új változók beillesztése” menüpontja segítségével bővíthetjük.

2. Az adatfelvitel két fő lépésből áll. Először célszerű megadni a változók főbb adatait (rövid és hosszú nevét, a lehetséges legkisebb és legnagyobb értéket, kódolt változók esetében a kódok neveit stb.). A változók jellemzőinek megadásához ablakot kell váltanunk, rákattintva a képernyő bal alsó sarkában látható „Változók deklarációi” parancsikonra.

3. A változódeklarációs ablakban minden változóhoz egy sor tartozik. A változó sorában látható, illetve adható meg a változó rövid és hosszú neve, típusa (numerikus vagy szöveges), numerikus változó esetén a változó alsó és felső korlátja, kódolt értékek esetén a kódok és azok nevei stb. A feladatunk abból áll, hogy az 1.3. táblázat összesen 10 változójának egymás után meg kell nyitni ebben az ablakban egy sort és beírni a fenti jellemzőiket.

A rövid változónév maximum 8 betűből, illetve más jelből (pl. számból) állhat, a hosszú név esetében ilyen megkötés nincs. Az alsó és felső limitek használata az adatbevitel során csökkenti a hibás adatok felkerülésének esélyét.

Kezdetben csak egy változósort látunk a képernyőn. Ennek indexe (sorszáma 1), rövid neve var\_1, hosszú neve var\_1\_long, típusa numerikus, alsó és felső limitje nincs (ezt num=0\_0 jelzi). Helyezzük el itt *Kódszám* változónk jellemzőit! Ehhez csupán be kell írni a megfelelő rovatokba a megfelelő adatokat. Például a Rövid név Var\_1-et tartalmazó rovatára kétszer rákattintva megjelenik egy kis szerkesztő ablak, ahova beírhatjuk a "Kódszám" nevet. Az 1.3. táblázatból kiolvashatjuk, hogy a *Kódszám* legkisebb és legnagyobb értéke 5, illetve 116. Legyenek tehát ezek a *Kódszám* esetében a beírandó alsó és felső limitek, amelyeket a Típus rovat jobb széle mellett látható kis négyzet alakú ikonra rákattintva megjelenő menüablakban állíthatunk be. Ezeket rögzítve a program nem engedi majd meg, hogy e limiteken kívül eső értékek kerüljenek fel a *Kódszám* adatoszlopába. Ez persze nem jelent végleges kötöttséget, mert ha később nagyobb kódú személy adatait is szeretnénk felvinni, a felső limit könnyűszerrel módosítható majd.

A második változó (*Pulzus*) információinak beviteléhez először is illesszünk be újabb változót a „Változók” parancsikonra rákattintva, s adjuk meg ennek jellemzőit is ugyanúgy, mint a *Kódszám* esetében! E változó javasolt alsó és felső korlátja az 1.3. táblázat alapján mondjuk 50 és 170 lehet. Ezt a műveletsort folytatva írhatjuk be a többi 8 változó fő jellemzőit is.

4. Az adatok beírásához vissza kell mennünk az adatok táblázatához, rákattintva a képernyő bal alsó sarkában látható „Adattáblázat” parancsikonra. Itt most már csak ki kell töltenünk a táblázat rovatait. Alternatív megoldásként az adatok felírhatók személyenként oszlopos táblázat formájában is, mely az „Oszlopinput” parancsikonra kattintva jelenik meg.

5. A kitöltés után az adatfájl elmentése a „Fájl” menüparancs „Mentés másként” menüpontjával történhet. Mentsük el az adatállományt mondjuk SPORTN.MSW néven, mely fájl immár az 1.3. táblázat összes adatát minden változójellel együtt tartalmazza.

### 2.4.1. Diszkrét változók empirikus eloszlása

Diszkrét esetben a tapasztalati eloszlást úgy készítjük el, hogy számba vesszük: az X változó különböző értékei a mintában rendre hányszor fordulnak elő. Ezeket a **gyakorisági értékeket** (röviden: **gyakoriságokat**) táblázatba foglaljuk, s így jutunk **az X-minta gyakorisági eloszlásához**. Példaként az 1.3. táblázat *Sportág* változójának gyakorisági eloszlását a 2.6. táblázat mutatja be. Az egyszerű gyakoriságoknál sokszor informatívabbak a **relatív gyakoriságok**, amelyeket a fenti mintára vonatkozóan szintén a 2.6. táblázat mutat be százalékos formában.

**2.6. táblázat.** Az 1.3. táblázat *Sportág* változójának gyakorisági eloszlása és relatív gyakoriságai

| Változó:              | Sportág            |                  |                   |          |
|-----------------------|--------------------|------------------|-------------------|----------|
| Értékek:              | 1 =<br>Tollaslabda | 2 =<br>Cselgáncs | 3 =<br>Súlyemelés | Összesen |
| Gyakoriságok:         | 18                 | 5                | 2                 | 25       |
| Relatív gyakoriságok: | 72 %               | 20 %             | 8 %               | 100 %    |

Bár egy 25 fős mintában nem nehéz kézi számolással sem elkészíteni egy gyakorisági táblázatot, gyakorlás céljából nézzük meg, hogy a ROPstat hogyan birkózik meg ezzel a feladattal. Indítsuk el a ROPstatot, majd nyissuk meg az 1.3. táblázat adatait tartalmazó SPORTN.MSW fájlt.

A *Pulzus* változó gyakorisági eloszlásának elkészítése céljából a „Statisztikai elemzések” parancsikorra kattintva az 'Alapok' menüpontot, majd ezen belül a 'Gyakoriság, hisztogram' menüpontot kell kiválasztanunk. Ekkor a megjelenő elemzés-specifikáló ablakban tegyük a *Pulzus* változót (ez a változók sorában a 3.) a kijelölt változók ablakába, majd az ettől jobbra látható ablakban az egér jobb oldali gombjának segítségével állítsuk be a változó típusát diszkrétre, hogy pontos képünk legyen az egyes értékek előfordulási gyakoriságáról! Ha ezután a „Futtat” parancsikkal elindítjuk az elemzést, a 2.7. táblázatban látható gyakorisági eloszláshoz jutunk.

A 2.7. táblázat 'Érték' oszlopában a program a változó adott mintában talált különböző értékeit sorolja fel, mellette megadja az érték mintabeli gyakoriságát (Gyak) és százalékos relatív gyakoriságát (%), majd végül százalékos kumulatív gyakoriságát (Kum%). Ez utóbbi adat minden érték esetében arról tájékoztat, hogy az illető érték, valamint a nála kisebb értékek összesen a minta hány százalékát képezik. A 2.7. táblázatból például kiolvashatjuk, hogy a 83-as pulzusérték két személy esetében fordult elő. Ez a 24 fős minta (mint az 1.3. táblázatból látható, az eredeti 25-ből 1 személynek hiányzik a pulzus adata) 8.3%-a, s ekkora vagy ennél kisebb érték jellemzi a minta 37.5%-át.

## 2.7. táblázat Az 1.3. táblázat *Pulzus* változójának ROPstat segítségével elkészített gyakorisági eloszlása (változótípus: diszkrét)

| Gyakoriság, hisztogram             |      |     |       |
|------------------------------------|------|-----|-------|
| A beolvasott összes eset száma: 25 |      |     |       |
| VÁLTOZÓ: Pulzus                    |      |     |       |
| Érték                              | Gyak | %   | Kum%  |
| 56                                 | 1    | 4.2 | 4.2   |
| 64                                 | 1    | 4.2 | 8.3   |
| 68                                 | 1    | 4.2 | 12.5  |
| 71                                 | 1    | 4.2 | 16.7  |
| 76                                 | 1    | 4.2 | 20.8  |
| 77                                 | 1    | 4.2 | 25.0  |
| 79                                 | 1    | 4.2 | 29.2  |
| 83                                 | 2    | 8.3 | 37.5  |
| 85                                 | 1    | 4.2 | 41.7  |
| 89                                 | 1    | 4.2 | 45.8  |
| 97                                 | 1    | 4.2 | 50.0  |
| 105                                | 1    | 4.2 | 54.2  |
| 109                                | 1    | 4.2 | 58.3  |
| 110                                | 1    | 4.2 | 62.5  |
| 112                                | 1    | 4.2 | 66.7  |
| 116                                | 1    | 4.2 | 70.8  |
| 117                                | 1    | 4.2 | 75.0  |
| 119                                | 1    | 4.2 | 79.2  |
| 122                                | 1    | 4.2 | 83.3  |
| 124                                | 1    | 4.2 | 87.5  |
| 132                                | 1    | 4.2 | 91.7  |
| 141                                | 1    | 4.2 | 95.8  |
| 163                                | 1    | 4.2 | 100.0 |
| Össz.:                             | 24   |     |       |

Bár a *Pulzus* változó formailag diszkrét (értékét egész értékre kerekítve szokás megadni, mint a testmagasságot), ebben a mintában majdnem mindenki más pulzusértéket produkált (a kivétel éppen a fent említett 83-as érték). Emiatt az értékenkénti felsorolásból nem igazán

látható, hogy ezek az edzés után mért pulzusértékek hol tömörülnek leginkább. E célból célszerű övezeteket kialakítani. Ezeket a ROPstat automatikusan elkészíti, ha a változó típusát folytonosra állítjuk be, vagyis ha alapértelmezés szerint futtatjuk a fenti programot. Ezt is elvégezve a 2.8. táblázat eredménylistájához jutunk.

**2.8. táblázat.** Az 1.3. táblázat *Pulzus* változójának ROPstat segítségével elkészített gyakorisági eloszlása (változótípus: folytonos)

| Gyakoriság, hisztogram             |      |      |       |       |
|------------------------------------|------|------|-------|-------|
| A beolvasott összes eset száma: 25 |      |      |       |       |
| VÁLTOZÓ: Pulzus                    |      |      |       |       |
| Oszt.köz.                          | Gyak | %    | Kum%  |       |
| 63.6                               | 4    | 16.7 | 16.7  | ===   |
| 78.9                               | 6    | 25.0 | 41.7  | ===== |
| 94.2                               | 2    | 8.3  | 50.0  | =     |
| 109.5                              | 6    | 25.0 | 75.0  | ===== |
| 124.8                              | 4    | 16.7 | 91.7  | ===   |
| 140.1                              | 1    | 4.2  | 95.8  |       |
| 155.4                              | 1    | 4.2  | 100.0 |       |
| Össz.:                             | 24   |      |       |       |

A 2.8. táblázat gyakorisági eloszlása csak abban tér el a 2.7. táblázatban láthatótól, hogy az egyes sorok információi nem a különböző mintabeli értékekre, hanem a program által automatikusan létrehozott azonos szélességű övezetekre (osztályokra) vonatkoznak. Az eredménylista ezeknek az övezeteknek csak a közepeit tünteti fel (Oszt.köz.). De mivel a szomszédos közepek közti eltérések megegyeznek az osztályok közös szélességével (ami esetünkben  $155.4 - 140.1 = 15.3$ ), az övezethatárok is egyszerűen kiszámíthatók (osztályközép  $\pm$  osztályszélesség fele). A táblázatból egyébként szakmailag azt olvashatjuk ki, hogy ezeknél a női sportolóknál az edzés után mért pulzusértékeknek két sűrűsödési pontja van: az egyik a 80-as, a másik a 110-es pulzusérték környékén. Megjegyezzük, hogy a ROPstatban mód van arra, hogy szabályozzuk a program által automatikusan létrehozott övezetek számát. A 2.8. táblázat egy alapértelmezés szerinti osztályozást mutat be, amely a minimális számú övezetet hozza létre. Ezt akár meg is hatszorozhatjuk (az osztályszám-szorító értékének megfelelő beállításával).

Egy másik lehetőség az övezetek szabályozására, ha mi magunk állítjuk be őket egyenként. Erre szolgál a CsopDef típus. E célból a változóhoz csoportokat kell definiálni övezetek definiálásával (pl. a változódeklarációs ablakban, vagy úgy, hogy a ki nem választott változók listájában az egérrel a változóra lépve megnyomjuk az egér jobb oldali gombját), de ennek technikai részletekre itt most nem térünk ki. Ha hasonló módon a *Sportág* változó csoportjait is kijelöljük (ez esetben persze nem övezetekkel, hanem kódszámokkal), akkor a két változónál a CsopDef típus beállítása után a 2.9. táblázatban látható eredményt kapjuk.

A 2.10. táblázat egy többesreprezentatív KSH-minta adatain alapuló gyermek-egészségügyi longitudinális vizsgálat néhány diszkrét változójának ROPstat segítségével elkészített gyakorisági eloszlását tartalmazza. Itt érdemes figyelni néhány érdekes eredményre. A nemmel kapcsolatban meglepőnek tűnhet, hogy a fiúk aránya mintegy 4.2%-kal nagyobb, mint a lányoké. Az újszülöttek között a fiúknak ez a dominanciája arra utal, hogy a természet genetikailag igyekszik kompenzálni a férfi nem nagyobb sérülékenységét és magasabb halálráta-rátáját. Érdekes továbbá, hogy a gyerekek között milyen nagy a barna hajszín előfordulása (világos és sötét árnyalatával együtt a gyerekek kb. 80%-át ez jellemzi). Végül a szoptatással kapcsolatban az tűnik elgondolkodtató adatnak, hogy a csecsemők több mint a felét (pontosan 54.5%-át) csak legfeljebb 2 hónapos korukig szoptatják, s mintegy 21%-uk anyatej

nélkül nő fel. Meg kell jegyeznünk, hogy a táblázatban 6 és 12 hónap között azért nem szerepelnek páratlan értékek, mert a csecsemők megfigyelését 6 és 12 hónapos koruk között csak páros hónapokban végezték, s csupán azt rögzítették, hogy a csecsemő napi táplálékának nagy részét szopással szerzi-e meg, vagy sem.

**2.9. táblázat.** Az 1.3. táblázat *Sportág* és *Pulzus* változójának ROPstat segítségével elkészített csoportkijelöléses gyakorisági eloszlása

| VÁLTOZÓ: Sportág            |            |               |          |           |       |
|-----------------------------|------------|---------------|----------|-----------|-------|
| Index                       | Csoportnév | Csoport gyak. | Teljes % | Relatív % | Kum%  |
| 1                           | Tollas     | 18            | 72,0     | 72,0      | 72,0  |
| 2                           | Cselg      | 5             | 20,0     | 20,0      | 92,0  |
| 3                           | Súlyem     | 2             | 8,0      | 8,0       | 100,0 |
| Összesen:                   |            | 25            | 100,0    | 100       |       |
| Megadott csoportokon kívül: |            | 0             | 0,0      |           |       |
| Mindösszesen:               |            |               | 25       | 100       |       |
| VÁLTOZÓ: Pulzus             |            |               |          |           |       |
| Index                       | Csoportnév | Csoport gyak. | Teljes % | Relatív % | Kum%  |
| 1                           | 40-60      | 1             | 4,0      | 4,2       | 4,2   |
| 2                           | 61-80      | 6             | 24,0     | 25,0      | 29,2  |
| 3                           | 81-100     | 5             | 20,0     | 20,8      | 50,0  |
| 4                           | 101-120    | 7             | 28,0     | 29,2      | 79,2  |
| 5                           | 121-140    | 3             | 12,0     | 12,5      | 91,7  |
| 6                           | 141-160    | 1             | 4,0      | 4,2       | 95,8  |
| 7                           | 161-180    | 1             | 4,0      | 4,2       | 100,0 |
| Összesen:                   |            | 24            | 96,0     | 100       |       |
| Megadott csoportokon kívül: |            | 1             | 4,0      |           |       |
| Mindösszesen:               |            |               | 25       | 100       |       |

**2.10. táblázat.** Egy gyermekegészségügyi longitudinális vizsgálat néhány diszkrét változójának ROPstat segítségével készített gyakorisági eloszlása (Nem = Az újszülött neme, Hajlév = A gyerek hajának színe 1 éves korában, SzopHón = Hány hónapos koráig szopott a gyerek)

| VÁLTOZÓ: GyNeme             |            |               |          |           |       |
|-----------------------------|------------|---------------|----------|-----------|-------|
| Index                       | Csoportnév | Csoport gyak. | Teljes % | Relatív % | Kum%  |
| 1                           | Fiú        | 3503          | 52,1     | 52,1      | 52,1  |
| 2                           | Lány       | 3217          | 47,9     | 47,9      | 100,0 |
| Összesen:                   |            | 6720          | 100,0    | 100       |       |
| Megadott csoportokon kívül: |            | 0             | 0,0      |           |       |
| Mindösszesen:               |            | 6720          | 100      |           |       |
| VÁLTOZÓ: Hajlév             |            |               |          |           |       |
| Index                       | Csoportnév | Csoport gyak. | Teljes % | Relatív % | Kum%  |
| 1                           | fekete     | 558           | 8,3      | 9,0       | 9,0   |
| 2                           | sötétbarna | 1547          | 23,0     | 25,0      | 34,0  |
| 3                           | barna      | 1888          | 28,1     | 30,5      | 64,5  |
| 4                           | vilbarna   | 1507          | 22,4     | 24,3      | 88,8  |
| 5                           | szőke      | 592           | 8,8      | 9,6       | 98,4  |
| 6                           | lenszőke   | 57            | 0,8      | 0,9       | 99,3  |
| 7                           | vörös      | 44            | 0,7      | 0,7       | 100,0 |
| Összesen:                   |            | 6193          | 92,2     | 100       |       |
| Megadott csoportokon kívül: |            | 527           | 7,8      |           |       |
| Mindösszesen:               |            | 6720          | 100      |           |       |
| VÁLTOZÓ: SzopHón            |            |               |          |           |       |

| Érték  | Gyak | %    | Kum%  |       |
|--------|------|------|-------|-------|
| 0      | 1297 | 20,8 | 20,8  | ===== |
| 1      | 1159 | 18,6 | 39,4  | ===== |
| 2      | 943  | 15,1 | 54,5  | ===== |
| 3      | 761  | 12,2 | 66,7  | ===== |
| 4      | 586  | 9,4  | 76,1  | ===== |
| 5      | 464  | 7,4  | 83,6  | ===== |
| 6      | 514  | 8,2  | 91,8  | ===== |
| 8      | 272  | 4,4  | 96,2  | ===== |
| 10     | 133  | 2,1  | 98,3  | ===== |
| 12     | 104  | 1,7  | 100,0 | ===== |
| Össz.: | 6233 |      |       |       |

A *SzopHón* változó gyakoriságai mellett látható vonalas grafikon a 2.10. táblázatban a gyakorisági adatok egyfajta ábrázolása. Itt a vonalak hossza a gyakoriságokkal arányos és szemléletesen tükrözi a különböző értékek előfordulási arányai közti különbségeket. Az ilyen jellegű grafikonokat a statisztikában **hisztogramnak** nevezik. A CsopDef típusúra beállított változók esetében ez azért nem szerepel, mert az egyes csoportokra vonatkozó információk mellett már nincs elég hely ezek felrajzolására. Hisztogram készítéséhez tehát a változó típusát folytonosra vagy diszkrétre kell beállítani.

## 2.4.2. Folytonos változók empirikus eloszlása

Folytonos változók esetén a változó értékskáláját övezetekre kell osztani. Ezt megtehetjük mi magunk is, mint a *Pulzus* változó 2.9. táblázatban bemutatott eloszlása esetében, vagy rábízhathatjuk magunkat a programra, mint a fentebbi 2.8. táblázatban bemutatott esetben. Példaként előbb az 1.3. táblázat edzés után mért szisztolés vérnyomás változójának (*Szisztol*) empirikus

**2.11. táblázat.** Az 1.3. táblázat *Szisztol* vérnyomás változójának empirikus eloszlása

| VÁLTOZÓ: Szisztol (Szisztolés vérnyomás) |      |      |       |       |
|------------------------------------------|------|------|-------|-------|
| Oszt.köz.                                | Gyak | %    | Kum%  |       |
| 111,1                                    | 3    | 12,0 | 12,0  | ==    |
| 119,2                                    | 5    | 20,0 | 32,0  | ===== |
| 127,4                                    | 7    | 28,0 | 60,0  | ===== |
| 135,5                                    | 4    | 16,0 | 76,0  | ===== |
| 143,6                                    | 4    | 16,0 | 92,0  | ===== |
| 151,8                                    | 1    | 4,0  | 96,0  | ===== |
| 159,9                                    | 1    | 4,0  | 100,0 | ===== |
| Össz.:                                   | 25   |      |       |       |

eloszlását készítsük el! Az ehhez vezető lépések ugyanazok, mint a 2.8. táblázat esetén részletezettek. Sikeres végrehajtás után a 2.11. táblázatban látható gyakorisági eloszláshoz jutunk. Ez esetben a program a kis mintanagyság miatt alakított ki olyan kevés övezetet. A fentebb már említett longitudinális vizsgálat két folytonos változójának (születéskori testhossz és testsúly) empirikus eloszlása esetében már jóval részletesebb gyakorisági táblázatok és tetszetősebb hisztogramok tárulnak a szemünk elé (lásd 2.12. és 2.13. táblázat).

Persze egy grafikus program, mint amilyen a Microsoft Excel is, a ROPstaténál szebb hisztogramot tud készíteni (lásd 2.3. ábra). Ha pedig a 2.3. ábra osztályközépeinél erre a hisztogramra egy jól illeszkedő görbét helyezünk, akkor a kapott grafikon, az ún. **empirikus eloszlásgörbe** már igen jól megközelítheti a sűrűségfüggvény grafikonját (lásd 2.4. ábra). Ezeken az ábrákon a legkisebb és a legnagyobb érték (azaz 1104 és 4866) a ROPstat által készített gyakorisági eloszlás legalsó és legfelső övezetének egészen kerekített osztályközépe (vö. a 2.13. táblázat Tsúly0 változóra vonatkozó hisztogramjával).

**2.12. táblázat.** Egy gyermekegészségügyi longitudinális vizsgálat születési testhossz és testsúly folytonos változójának ROPstat segítségével elkészített gyakorisági eloszlása

| VÁLTOZÓ: Thossz0 (születési testhossz) |      |      |       |       |
|----------------------------------------|------|------|-------|-------|
| Oszt.köz.                              | Gyak | %    | Kum%  |       |
| 35,96                                  | 3    | 0,0  | 0,0   |       |
| 37,69                                  | 3    | 0,0  | 0,1   |       |
| 39,42                                  | 8    | 0,1  | 0,2   |       |
| 41,14                                  | 18   | 0,3  | 0,5   |       |
| 42,87                                  | 44   | 0,7  | 1,2   |       |
| 44,60                                  | 113  | 1,8  | 3,1   | ==    |
| 46,32                                  | 401  | 6,5  | 9,6   | ===== |
| 48,05                                  | 1047 | 17,1 | 26,7  | ===== |
| 49,78                                  | 1907 | 31,1 | 57,8  | ===== |
| 51,50                                  | 1558 | 25,4 | 83,2  | ===== |
| 53,23                                  | 714  | 11,6 | 94,8  | ===== |
| 54,96                                  | 226  | 3,7  | 98,5  | ===== |
| 56,68                                  | 66   | 1,1  | 99,6  | =     |
| 58,41                                  | 22   | 0,4  | 99,9  |       |
| 60,14                                  | 5    | 0,1  | 100,0 |       |
| Össz.:                                 | 6135 |      |       |       |

| VÁLTOZÓ: Tests0 (születési testsúly) |      |      |       |       |
|--------------------------------------|------|------|-------|-------|
| Oszt.köz.                            | Gyak | %    | Kum%  |       |
| 1104,3                               | 8    | 0,1  | 0,1   |       |
| 1373,0                               | 7    | 0,1  | 0,2   |       |
| 1641,7                               | 24   | 0,4  | 0,6   |       |
| 1910,3                               | 63   | 1,0  | 1,7   | =     |
| 2179,0                               | 144  | 2,3  | 4,0   | ===== |
| 2447,7                               | 354  | 5,8  | 9,8   | ===== |
| 2716,3                               | 834  | 13,6 | 23,3  | ===== |
| 2985,0                               | 1202 | 19,6 | 42,9  | ===== |
| 3253,7                               | 1354 | 22,0 | 64,9  | ===== |
| 3522,3                               | 1141 | 18,6 | 83,5  | ===== |
| 3791,0                               | 608  | 9,9  | 93,4  | ===== |
| 4059,7                               | 259  | 4,2  | 97,6  | ===== |
| 4328,3                               | 112  | 1,8  | 99,4  | ===== |
| 4597,0                               | 31   | 0,5  | 100,0 |       |
| 4865,7                               | 3    | 0,0  | 100,0 |       |
| Össz.:                               | 6144 |      |       |       |

### 3.2.4. A minta módusza

A **minta módusza** (szokásos jelöléssel  $M_o$ ) a mintában legtöbbször, legnagyobb gyakorisággal előforduló adatérték. Csak nagyobb – legalább 50–100 elemű – minták esetén szokták értelmezni. Például a 2.6. táblázat adatmintája esetében a minta módusza az 1 = Tollaslabda érték, a 2.10. táblázatban bemutatott minták esetében az 1 = Fiú ( $X = Nem$ ), az 1 = RH-Poz ( $X = Vércso$ ), a 3 = barna ( $X = Hajl$ ), illetve a 0 ( $X = SzopH$ ).

Ha  $X$  folytonos változó, akkor mintabeli móduszának megállapításához olyan gyakorisági eloszlást kell készíteni, ahol a lehetséges értékek tartományát övezetekre (osztályokra) osztjuk fel, s az adatokat ezek szerint csoportosítjuk (lásd például a 2.12. táblázatban bemutatott két gyakorisági eloszlást). Ilyen esetben a minta móduszának a legnagyobb gyakoriságú osztály osztályközepét tekintjük. Például a 2.12. táblázatban látható ROPstat eredménylistán a *Thossz0* változó (születéskori testhossz cm-ben) módusza 49.78 (cm), a *Tsúly0* változó (születéskori testsúly grammban) módusza pedig 3253.7 g  $\approx$  3.25 kg. Ilyenformán folytonos esetben a módusz erősen függ az osztályok helyének kijelölésétől.

Ezen esetlegesség csökkentése érdekében ha a minta nagysága lehetővé teszi, válasszuk az osztályok számát kellően nagyra! Például a *Thossz0* változó esetében megduplázva az osztályok számát (ehhez a ROPstatban a 'Gyakoriság, hisztogram' rutinban az osztályszám szorzó-



ját 2-re kell beállítani), a 3.5. táblázatban feltüntetett részletesebb gyakorisági eloszláshoz jutunk, ahonnan kiolvasható, hogy a minta módusza az  $M_o = 50.21$  (cm) érték. Ez feltehetőleg jobban megközelíti az elméleti értéket, mint a 2.12. táblázatból megállapítható 49.78, bár a kapott két érték szemre nem tér el túlságosan egymástól. Érdekes azonban felfigyelni arra, hogy ez az 50.21 módusz érték pontosan megegyezik a 3.2.2. alpontban közelítő számítással meghatározott ( $M = 50.21$ ), illetve valódi mintabeli ( $M = 50.20$ ) medián értékkel. Tekintve, hogy esetünkben mind a mintaátlag ( $\bar{x} = 50.20$ ), mind a trimmelt mintaátlag szinte ugyanekkora ( $\bar{x}_t = 50.23$ ), a kapott közös 50.2 cm a vizsgált minta megbízható, jó középpértékének tekinthető.

A *Thossz0* változó négy különböző középpértéke az eloszlás szép, meglehetősen szimmetrikus alakja miatt egyezik meg ilyen nagy mértékben (lásd 3.5. táblázat). Ha az eloszlás erősen ferde, akkor ilyenre nem számíthatunk. Például a 2.10. táblázat *SzopHón* változója esetében a módusz a 0, a medián a 2 érték (lásd 2.10. táblázat), a táblázatbeli adatokból kiszámított átlag, illetve 10%-os trimmelt átlag pedig 2.86, illetve 2.48 (vö. 3.2.3. alpont). Ebben az esetben nem igazán tudjuk egyetlen számmal jellemezni a 80-as évek magyar viszonyai között a szoptatási szokást. A legtipikusabb érték (a módusz) a 0, vagyis szopási hónapszámban a 0 a leggyakoribb érték. Viszont az eloszlás közepén (a mediánnál) a 2 hónap található. Ha pedig leátlagoljuk az adatokat, a viszonylag ritkább, de mégis csak előforduló magas szopási hónapértékek 3 körüli (pontosan 2.86) átlagszintet eredményeznek. Az, hogy adott esetben melyik középpértéket használjuk, a megfogalmazott szakmai kérdéstől függ.

**3.5. táblázat.** A 2.12. táblázat *Thossz0* (születés kori testhossz cm-ben) változójának részletesebb (kétszer annyi osztályt tartalmazó) gyakorisági eloszlása

| VÁLTOZÓ: Thossz0 (születési testhossz) |      |      |       |
|----------------------------------------|------|------|-------|
| Oszt.köz.                              | Gyak | %    | Kum%  |
| 35,53                                  | 1    | 0,0  | 0,0   |
| 36,40                                  | 2    | 0,0  | 0,0   |
| 37,26                                  | 2    | 0,0  | 0,1   |
| 38,12                                  | 1    | 0,0  | 0,1   |
| 38,98                                  | 5    | 0,1  | 0,2   |
| 39,85                                  | 3    | 0,0  | 0,2   |
| 40,71                                  | 9    | 0,1  | 0,4   |
| 41,58                                  | 9    | 0,1  | 0,5   |
| 42,44                                  | 15   | 0,2  | 0,8   |
| 43,30                                  | 29   | 0,5  | 1,2   |
| 44,16                                  | 45   | 0,7  | 2,0   |
| 45,03                                  | 68   | 1,1  | 3,1   |
| 45,89                                  | 152  | 2,5  | 5,6   |
| 46,76                                  | 249  | 4,1  | 9,6   |
| 47,62                                  | 472  | 7,7  | 17,3  |
| 48,48                                  | 575  | 9,4  | 26,7  |
| 49,34                                  | 836  | 13,6 | 40,3  |
| 50,21                                  | 1071 | 17,5 | 57,8  |
| 51,07                                  | 951  | 15,5 | 73,3  |
| 51,94                                  | 607  | 9,9  | 83,2  |
| 52,80                                  | 474  | 7,7  | 90,9  |
| 53,66                                  | 240  | 3,9  | 94,8  |
| 54,52                                  | 128  | 2,1  | 96,9  |
| 55,39                                  | 98   | 1,6  | 98,5  |
| 56,25                                  | 36   | 0,6  | 99,1  |
| 57,12                                  | 30   | 0,5  | 99,6  |
| 57,98                                  | 15   | 0,2  | 99,8  |
| 58,84                                  | 7    | 0,1  | 99,9  |
| 59,70                                  | 3    | 0,0  | 100,0 |
| 60,57                                  | 2    | 0,0  | 100,0 |
| Össz.:                                 | 6135 |      |       |

### 3.5. A minta statisztikai jellemzőinek kiszámítása a ROPstat segítségével

A ROPstatban két rutin segítségével számíthatjuk ki legkönnyebben egy adatminta statisztikai jellemzőit. Mindkettő az 'Alapok (egymintás eljárások)' programcsoportban található. Az 'Alapstatisztikák' rutin segítségével változónként alapstatisztikák (átlag, szórás, variációs együttható, legkisebb és legnagyobb érték stb.) számíthatók ki, míg a 'Részletesebb statisztikák' rutinban az előbbieken túl meghatározható többek közt a medián, a trimmelt átlag, a winszorizált szórás, az IT és az IF is. Az alábbiakban ezeket kissé részletesebben is ismertetjük.

1. Először számítsuk ki az 'Alapstatisztikák' rutinnal a korábban már bemutatott SPORTN.MSW adatállomány (lásd 1.6. alfejezet) változóinak alapstatisztikáit! Ez a következő lépésekben végezhető el, miután az 1.6. alfejezet instrukciói alapján az 1.3. táblázat adatait beírtuk egy SPORTN.MSW nevű MSW-fájlba.

Indítsuk el a ROPstatot, majd a „Megnyit” parancsikon segítségével nyissuk meg a SPORTN.MSW fájlt! A „Statisztikai elemzések” parancsikonra kattintva az 'Alapok' menüpontot, majd ezen belül az 'Alapstatisztikák' menüpontot kell kiválasztanunk. Ekkor a megjelenő elemzés-specifikáló ablakban tegyük az összes változót a kijelölt változók ablakába! Ha ezután a „Futtat” parancsikkal elindítjuk az elemzést, a 3.7. táblázatban látható eredménylistához jutunk.

A 3.7. táblázatból kiolvashatjuk például, hogy a *Pulzus* változó átlaga 99.92, szórása 26.84, variációs együtthatója 27%, legkisebb értéke 56, legnagyobb értéke 163. Mivel ugyanezen változó esetében az előforduló legkisebb standard érték:  $z_{\min} = -1.64$  (ez az  $x = 56$  értékhez tartozó  $z$  érték) és a legnagyobb standard érték:  $z_{\max} = 2.35$  (az  $x = 163$  értékhez tartozó  $z$  érték), amelyek abszolút értéke egyaránt kisebb 3-nál, ebben a mintában nem találunk a mintából kilógó, extrém értékeket. A táblázatból láthatjuk, hogy ugyanez egyébként minden változóra igaz. Ha viszont ugyanebben a mintában egy 77-es pulzusértéket tízszeresére, 770-re cserélnénk fel, akkor a  $z_{\max}$  érték menten felugrana 4.61-re.

### 3.7. táblázat. A SPORTN.MSW fájl változóinak ROPstat segítségével elkészített alapstatisztikái

| Alapstatisztikák                                                  |          |        |        |        |          |       |       |       |       |
|-------------------------------------------------------------------|----------|--------|--------|--------|----------|-------|-------|-------|-------|
| Jelölés:                                                          |          |        |        |        |          |       |       |       |       |
| - Var. eh. = Variációs együttható = Relatív szórás = szórás/átlag |          |        |        |        |          |       |       |       |       |
| - X_min = Talált legkisebb érték                                  |          |        |        |        |          |       |       |       |       |
| - X_max = Talált legnagyobb érték                                 |          |        |        |        |          |       |       |       |       |
| - z_min = Standardizált legkisebb érték = (X_min - átlag)/szórás  |          |        |        |        |          |       |       |       |       |
| - z_max = Standardizált legnagyobb érték = (X_max - átlag)/szórás |          |        |        |        |          |       |       |       |       |
| A beolvasott összes eset száma: 25                                |          |        |        |        |          |       |       |       |       |
| Index                                                             | Változó  | Esetek | Átlag  | Szórás | Var. eh. | X_min | X_max | z_min | z_max |
| 1.                                                                | Kódszám  | 25     | 46,12  | 35,41  | 0,768    | 5     | 116   | -1,16 | 1,97  |
| 2.                                                                | Sportág  | 25     | 1,360  | 0,638  | 0,469    | 1     | 3     | -0,56 | 2,57  |
| 3.                                                                | Pulzus   | 24     | 99,92  | 26,84  | 0,269    | 56    | 163   | -1,64 | 2,35  |
| 4.                                                                | Szisztol | 25     | 129,48 | 12,59  | 0,0972   | 107   | 164   | -1,79 | 2,74  |
| 5.                                                                | Diasztol | 25     | 78,04  | 9,392  | 0,120    | 60    | 93    | -1,92 | 1,59  |
| 6.                                                                | Dom_pref | 20     | 51,45  | 11,07  | 0,215    | 29    | 70    | -2,03 | 1,68  |
| 7.                                                                | Akt_pref | 20     | 51,55  | 7,207  | 0,140    | 34    | 68    | -2,44 | 2,28  |
| 8.                                                                | Hos_pref | 20     | 51,90  | 9,936  | 0,191    | 37    | 73    | -1,50 | 2,12  |
| 9.                                                                | Fem_pref | 20     | 46,75  | 9,862  | 0,211    | 29    | 71    | -1,80 | 2,46  |
| 10.                                                               | Instab   | 20     | 49,75  | 6,447  | 0,130    | 34    | 61    | -2,44 | 1,74  |

2. Második feladatként készítsük el a korábban már többször hivatkozott longitudinális antropometriai vizsgálat egy demo adatállománya (ANTR100.MSW) változóinak részletesebb

statisztikai jellemzését, és pedig nemek szerinti bontásban! Az ANTR100.MSW fájl 100 személyét (47 fiú és 53 lány) véletlenszerűen választottuk ki a teljes 6720 személyt tartalmazó teljes adatállományból. Az állományon lévő változók a következők: *Tsúly0* (születéskori testsúly), *Tsúly10* (testsúly 10 éves korban), *Thossz0* (születéskori testhossz), *Tmag10* (testmagasság 10 éves korban), *AnyaSúly* (anya testsúlya), *ApaSúly* (apa testsúlya), *AnyaMag* (anya testmagassága), *ApaMag* (apa testmagassága). A kívánt számításokat a következőképpen végeztethetjük el a ROPstat programcsomaggal.

Indítsuk el a ROPstatot, majd a „Megnyit” parancsikon segítségével töltsük be az ANTR100.MSW fájlt! A „Statisztikai elemzések” parancsikonra kattintva az 'Alapok' menüpontot, majd ezen belül most a 'Részletesebb statisztikák' menüpontot válasszuk ki! Ekkor a megjelenő elemzés-specifikáló ablakban tegyük a Nem kivételével az összes változót a kijelölt változók ablakába, majd a Nem változót a „Feltételes csoportosító változó” ablakába! Ha ezután a „Futtat” parancsikonnal elindítjuk az elemzést, a 3.8. táblázatban látható eredménylistát kapjuk.

**3.8. táblázat.** Az ANTR100.MSW fájl változóinak ROPstat segítségével elkészített részletesebb statisztikái, ha a trimmelési százalék 0

Részletes leíró statisztikák

Feltételes csoportosító változó: Nem

Feltételes csoportok gyakoriságai

| Index | Csoportnév | Gyakoriság |
|-------|------------|------------|
| 1     | fiú        | 47         |
| 2     | lány       | 53         |

CSOPORTINDEX: 1. Csoportnév: fiú

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 47

| Index | Változó  | Esetek | Medián | Átlag  | St.hiba | 95%-os konf. int. | Ferdeség | Csúcsosság |
|-------|----------|--------|--------|--------|---------|-------------------|----------|------------|
| 2.    | Tsúly0   | 47     | 3,240  | 3,220  | 0,0548  | 3,109 3,330       | -0,11    | -0,82      |
| 3.    | Tsúly10  | 47     | 32     | 33,57  | 0,973   | 31,61 35,53       | 1,33***  | 1,88**     |
| 4.    | Thossz0  | 47     | 50     | 50,51  | 0,298   | 49,91 51,11       | -0,26    | -0,35      |
| 5.    | Tmag10   | 47     | 139    | 138,19 | 0,845   | 136,49 139,89     | -0,35    | 0,22       |
| 6.    | AnyaSúly | 47     | 65     | 66,26  | 2,092   | 62,04 70,47       | 1,20***  | 2,32**     |
| 7.    | ApaSúly  | 47     | 80     | 81,47  | 2,192   | 77,05 85,88       | 0,95**   | 0,79       |
| 8.    | AnyaMag  | 47     | 160    | 161,40 | 0,942   | 159,51 163,30     | 0,28     | -0,21      |
| 9.    | ApaMag   | 47     | 172    | 172,77 | 0,903   | 170,95 174,58     | 0,16     | -0,79      |

A normalitás tesztelése a Ferdeség és a Csúcsosság segítségével: \*: p < 0,05 \*\*: p < 0,01 \*\*\*: p < 0,001

CSOPORTINDEX: 2. Csoportnév: lány

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 53

| Index | Változó  | Esetek | Medián | Átlag  | St.hiba | 95%-os konf. int. | Ferdeség | Csúcsosság |
|-------|----------|--------|--------|--------|---------|-------------------|----------|------------|
| 2.    | Tsúly0   | 53     | 3,130  | 3,169  | 0,0694  | 3,030 3,309       | -0,14    | -0,74      |
| 3.    | Tsúly10  | 53     | 34     | 33,89  | 0,983   | 31,91 35,86       | 1,36***  | 2,90***    |
| 4.    | Thossz0  | 53     | 50     | 49,60  | 0,274   | 49,05 50,15       | -0,54    | -0,00      |
| 5.    | Tmag10   | 53     | 140    | 139,74 | 0,826   | 138,08 141,39     | 0,27     | -0,47      |
| 6.    | AnyaSúly | 53     | 63     | 66,75  | 1,900   | 62,94 70,57       | 0,52     | -0,70      |
| 7.    | ApaSúly  | 53     | 79     | 78,72  | 1,760   | 75,18 82,25       | 0,67*    | 0,47       |
| 8.    | AnyaMag  | 53     | 162    | 161,75 | 0,861   | 160,03 163,48     | 0,26     | -0,37      |
| 9.    | ApaMag   | 53     | 175    | 174,49 | 0,926   | 172,63 176,35     | -0,35    | -0,34      |

A normalitás tesztelése a Ferdeség és a Csúcsosság segítségével: \*: p < 0,05 \*\*: p < 0,01 \*\*\*: p < 0,001

Ebben a táblázatban a 3.7. táblázathoz képest többlet a medián, valamint a ferdeségi és a csúcsossági együttható értéke, amelyeket a program változónként kiszámít. Az eredménylis-

tán szintén megtalálható 95%-os konfidenciahatárok azt jelzik, hogy mekkora lehet, milyen övezetben található 95%-os valószínűséggel az elméleti átlag. Ez azonban egy későbbi (5.) fejezet témája lesz, ezért ezzel most ne foglalkozzunk, mint ahogy a St.hiba (standard hiba) oszlopával, a "normalitásvizsgálat szignifikanciájával", valamint a ferdeségi és a csúcossági együttható mellett időnként feltűnő jelzésekkel sem. Ezekre majd akkor térünk vissza, ha a szükséges elméleti háttér már felvázoltuk (lásd 5. és 6. fejezet).

3. Harmadik problémaként végezzük el a fenti adatállomány *Tsúly10* (tízéves kori testsúly) változójának a 3.8. táblázatban láthatónál részletesebb elemzését! Ehhez ugyanúgy kell eljárni, mint az előző példában, csak az elemzés végrehajtása előtt a trimmelési százalékot be kell állítani egy 0-tól különböző értékre, mondjuk 10%-ra. Az eredmény a 3.9. táblázatban látható. Ez az eredménylista még több ismeretlen fogalmat tartalmaz, mint az előző. Ezekkel most ne foglalkozzunk! Ami az eddigiek alapján érthető: a 'Pontbecslés' oszlopban található értékek a 'Paraméter' oszlopban felsorolt eloszlásjellemzők mintabeli (tapasztalati) értékei.

Például a *Tsúly10* változóra vonatkozó eredmények szerint ebben a mintában a fiúk 10 éves kori testsúlyának átlaga 33.57 kg, szórása 6,67 kg, variációs együtthatója pedig:

$$VE = s / \bar{x} = 6,67/33.57 = 0,20.$$

A trimmelt átlag (32.87 kg) 0,7 kg-mal kisebb a rendes átlagnál (33.57 kg), a winsorizált szórás (4.98 kg) pedig több mint 1,5 kg-mal a rendes szórásnál (6,67 kg). Ezek az értékek arra utalnak, hogy a 10 éves kori testsúly fölfelé szélsőségesebben leng ki, mint lefelé, hiszen a trimmelés eredményeképpen a trimmelt átlag érzékelhetően lefelé mozdult el.

A medián (32 kg) érezhetően közelebb van a trimmelt átlaghoz, mint a sima átlaghoz. Ebből arra következtethetünk, hogy a trimmelés után az eloszlás valamelyest szimmetrikusabbá vált. Minthogy a medián a (30; 37) interkvartilis tartománynak érezhetően nem a közepén helyezkedik el, ez arra utal, hogy az eloszlás a szélek levágása után sem lesz szimmetrikus.

A ferdeségi és a csúcossági együttható 0-tól jelentősen különböző értéke (1,33 és 1,88) a fiúk 10 éves kori testsúlyának határozott aszimmetrikus és nehéz szélű eloszlásáról árulkodik.

A lányok esetében analóg következtetések vonhatók le.

### 3.9. táblázat. Az ANTR100.MSW fájl változóinak ROPstat segítségével elkészített részletesebb statisztikái 10%-os trimmelés beállítása esetén

|                                      |                  |                  |                              |
|--------------------------------------|------------------|------------------|------------------------------|
| Részletes leíró statisztikák         |                  |                  |                              |
| Feltételes csoportosító változó: Nem |                  |                  |                              |
| Feltételes csoportok gyakoriságai    |                  |                  |                              |
| Index                                | Csoportnév       | Gyakoriság       |                              |
| 1                                    | male             | 47               |                              |
| 2                                    | female           | 53               |                              |
| CSOPORTINDEX: 1. Csoportnév: male    |                  |                  |                              |
| A beolvasott összes eset száma: 100  |                  |                  |                              |
| Az aktuális csoport elemszáma: 47    |                  |                  |                              |
| VÁLTOZÓ: SÜLY10                      |                  |                  |                              |
| Érvényes értékek száma: 47           |                  |                  |                              |
| Paraméter                            | Pont-<br>becslés | Standard<br>hiba | Konfidencia<br>határok (95%) |
| Átlag:                               | 33,57            | 0,97             | (31,61; 35,53)               |
| Szórás:                              | 6,67             |                  |                              |
| Variációs eh.:                       | 0.20             |                  |                              |

Trimmelés = 10%  
 Trimmelt átlag: 32,87 0,91 (31,65; 34,10)  
 Winsorizált  
   szórás: 4,98  
 Medián: 32  
 (Min; Max): (24; 56)  
 (Kvart1; Kvart3): (30; 37)  
 (Kvart3-Kvart1)/2: 3,50  
  
 A normalitás tesztelése a Ferdeség és a Csúcsosság együtthatója segítségével:  
 Ferdeség: 1,33 (p = 0,000)\*\*\*  
 Csúcsosság: 1,88 (p = 0,008)\*\*

CSOPORTINDEX: 2. Csoportnév: lány  
 A beolvasott összes eset száma: 100  
 Az aktuális csoport elemszáma: 53

VÁLTOZÓ: Tsúly10 (testsúly 10 éves korban)  
 Érvényes értékek száma: 53  

| Paraméter              | Pont-<br>becslés | Standard<br>hiba | Konfidencia<br>határok (95%) |
|------------------------|------------------|------------------|------------------------------|
| Átlag:                 | 33,89            | 0,98             | (31,91; 35,86)               |
| Szórás:                | 7,15             |                  |                              |
| Variációs eh.:         | 0,21             |                  |                              |
| Trimmelés = 10%        |                  |                  |                              |
| Trimmelt átlag:        | 33,23            | 0,90             | (32,02; 34,45)               |
| Winsorizált<br>szórás: | 5,26             |                  |                              |
| Medián:                | 34               |                  |                              |
| (Min; Max):            | (24; 62)         |                  |                              |
| (Kvart1; Kvart3):      | (29; 37)         |                  |                              |
| (Kvart3-Kvart1)/2:     | 4,00             |                  |                              |

A normalitás tesztelése a Ferdeség és a Csúcsosság együtthatója segítségével:  
 Ferdeség: 1,36 (p = 0,000)\*\*\*  
 Csúcsosság: 2,90 (p = 0,000)\*\*\*

### 4.3. Néhány speciális valószínűségeloszlás és generálásuk a ROPstat segítségével

A 4.2.3. alpontban bemutatott normális eloszlás igen fontos szerepet játszik mind a matematikai statisztikában, mind a különböző empirikus tudományokban. Az előbbi oka, hogy számos kvantitatív változókkal kapcsolatos statisztikai eljárás érvényességét csak a változók normalitása esetén lehet garantálni. A normális eloszlásnak a pszichológiában és más tudományokban betöltött fontos szerepét pedig az indokolja, hogy a kvantitatív szakmai változók közül sokan jól közelíthetők a normális eloszlással (vö. a 4.2.3. alpont 9. pontjában mondotakkal). Ugyanakkor azt is tudomásul kell venni, hogy az empirikus kutatások igen sok változója nem normális eloszlású. Micceri (1989) egy hosszabb tanulmányában több száz pszichológiai kutatás változóinak eloszlását elemezve arra a következtetésre jutott, hogy a nem normális eloszlások sokkal gyakrabban fordulnak elő a pszichológiában, mint azt régebben gondolták.

Nem kell messzire menni, hogy az ember nem normális eloszlású változóba botoljon. Például egy longitudinális fejlődéslélektani vizsgálatban csaknem 15 ezer gyermek sorsát követték nyomon a születéstől 11 éves korukig. A vizsgálat eredményeit ismertető kiadvány arról árulkodik, hogy a nem normális eloszlások megdöbbentő gyakorisággal fordulnak elő az empirikus kutatásokban. Ízelítőül a 4.7. ábra mutat be néhány extrém mértékben nem normális eloszlású változót ebből a vizsgálatból.

A normalitás szempontjából megvizsgáltuk a Rorschach-teszt 235 leggyakoribb jegye és mutatója alapján definiált változó eloszlását is 359 mentálisan egészséges személy mintája alapján. Ezen Rorschach-változók mindegyikének megvizsgáltuk a normalitását a ROPstat normalitásvizsgáló rutinja segítségével. Eredményül azt kaptuk, hogy a 235 változóból 187, vagyis a változók 80%-a bizonyíthatóan nem normális eloszlású volt. Példaként ezek közül az egy Rorschach-válaszra eső átlagos idő (RI) eloszlását a 4.8. ábra szemlélteti.

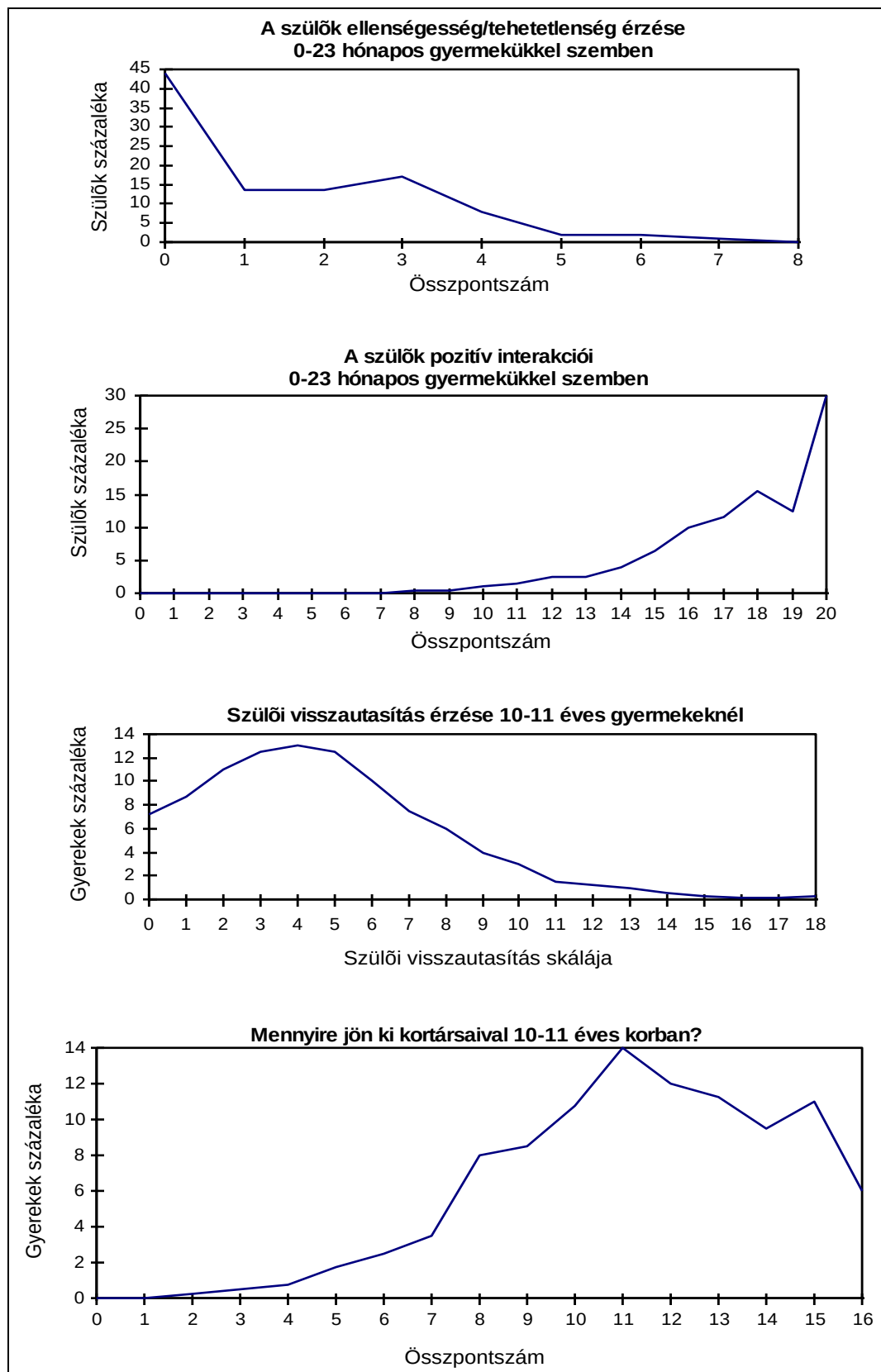
A fenti példák bemutatásával nem az volt a célunk, hogy a normális eloszlás jelentőségét vitassuk. Mindössze arra akartuk felhívni a figyelmet, hogy az empirikus változók között szép számmal fordulnak elő nem normális eloszlásúak is. Így egy új kísérleti vagy vizsgálati változó kapcsán, amíg nincs elegendő információnk az eloszlás alakjáról, nagy merészség azt normális eloszlásúnak feltételezni.

A normalitás megsértése gyakran abból fakad, hogy nem a legjobb mutatót használjuk a vizsgált jelenség mérésére. Lehetséges, hogy alkalmas nemlineáris standardizálás vagy skálát-transzformáció a 4.7. ábra változóiból is a normálist megközelítő eloszlású változókat hozna létre. A normalitás sok esetben azért sérül, mert az értékskálának van egy természetes alsó vagy felső korlátja. Például pozitív mennyiségek esetén a 0 érték mindig áthághatatlan alsó korlát, százalékban kifejezett gyakoriságok esetén pedig a 100 egy ilyen természetes felső korlát. Ha az egyének nagy része az alsó vagy a felső korláthoz közel eső értékkel jellemezhető, akkor ott az adatok feltorlódnak – mint lezárt sorompó előtt a gépkocsik –, s nem tudnak a közép körül mindkét irányban szépen, normálisan szétszóródni (vö. 4.7. és 4.8. ábra).

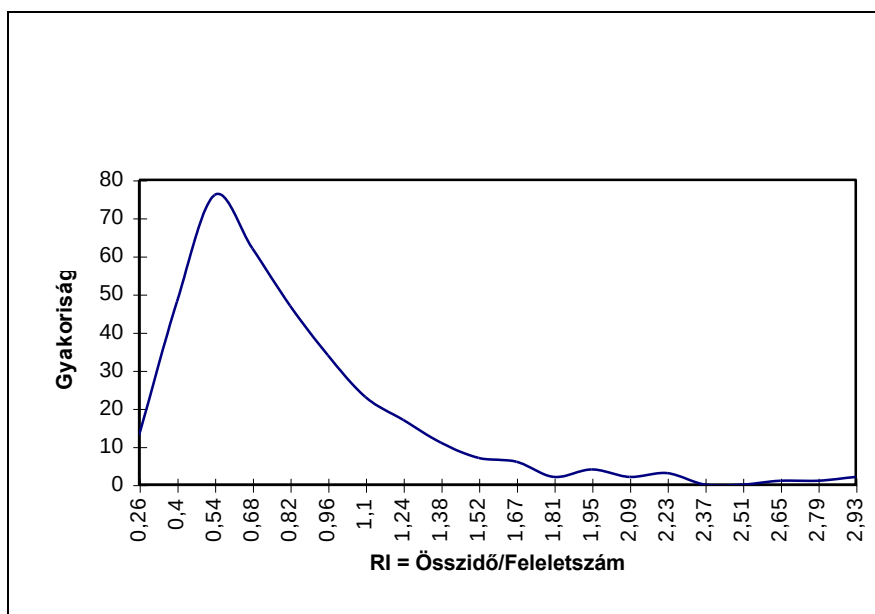
Ebben az alfejezetben néhány olyan nem normális valószínűségeloszlást ismertetünk, amelyek a különféle társadalomtudományi és biológiai empirikus kutatásokban, illetve a matematikai statisztikai módszerek kapcsán előfordulnak. Ezek többsége a ROPstatban speciális beépített rutinok segítségével generálható is, így alakjuk a ROPstat segítségével is szemléltethető. Emiatt alfejezetünk első alpontjában a változógenerálás technikáját mutatjuk be a normális eloszlás példáján (lásd 4.3.1. alpont).

A nem normális eloszlások közül elsőként az elméleti jelentőségű **egyenletes eloszlást** ismertetjük (lásd 4.3.2. alpont), majd két olyan diszkrét eloszlástípust (binomiális és Poisson), amelyek jellegzetes természeti, illetve társadalmi jelenségek leírására alkalmasak. Közelítőleg **Poisson-eloszlású** például egy adott időtartam alatt elbomló radioaktív részecskék száma vagy a Mammut üzletközpontba három óra alatt belépő személyek száma (lásd 4.3.3. alpont), **binomiális eloszlású** pedig mondjuk 100 véletlenszerűen kiválasztott ember közül azoknak a száma, akik a kormányt (vagy egy pártot, egy politikust vagy valamilyen termékfajtát) pozitívan minősítik (lásd 4.3.4. alpont).

Érdekes, hogy a normális eloszláshoz ezen eloszlástípusok határeseteként is eljuthatunk, tehát a könyvünkben ismertetett út (vö. 4.2.3. alpont) csak egy a lehetséges utak közül, amelyek a normális eloszláshoz elvezet. A normalitás sérülése néha abból fakad, hogy egy normális eloszlású változó markánsan eltérő átlagú és szórású egy populáció két alpopulációjában, s e változónak a teljes populációbeli eloszlását vizsgáljuk. Ez a **kevert normális eloszlás** (lásd 4.3.5. alpont).



**4.7. ábra.** Különböző nem normális eloszlások egy longitudinális vizsgálatból



**4.8. ábra.** A Rorschach-teszt RI változójának empirikus eloszlása (n = 359)

Az alfejezetben ismertetett eloszlástípusok egy másik csoportját azok az ún. **statisztikai eloszlások** képezik, amelyek a statisztikai eljárások végrehajtásában játszanak alapvető, de technikai jellegű szerepet. Ezek lényegében mind a normális eloszlásból származtathatók. A sort a **khi-négyzet-eloszlás** vezeti, melyhez független  $N(0, 1)$  eloszlású változók négyzetösszegeként jutunk (lásd 4.3.6. alpont). A normális és a khi-négyzet-eloszlás segítségével definiálható a matematikai statisztikában szintén nélkülözhetetlen **t-eloszlás** (lásd 4.3.7. alpont) és **F-eloszlás** (lásd 4.3.6. alpont).

Mindezekkel nem merítjük ki a természetben és a matematikai statisztikában előforduló összes nem normális eloszlást. Külön alpontot érdemelne például a **lognormális eloszlás**, mely többnyire olyan helyzetekben jön létre, amikor egy eredendően normális eloszlás valamilyen természetes alsó vagy felső korlát miatt nem tud valamilyen irányban kiteljesedni (jó példa erre a jövedelem vagy a testsúly eloszlása; vö. 5.7. táblázat). Hasonlóképpen az **exponenciális eloszlás**, amellyel jól közelíthető egy véletlen telefonbeszélgetés időtartamának, illetve egy radioaktív atom vagy technikai berendezés élettartamának az eloszlása. Ezekkel sajnos könyvünk terjedelmi korlátai miatt nem tudunk részletesebben foglalkozni. Az érdeklődő olvasónak ezekkel és más fontos valószínűségeloszlásokkal kapcsolatban Feller (1978), Rényi (1968) és Vincze (1968) művét ajánljuk a figyelmébe.

#### 4.3.1. Normális eloszlású változók generálása a ROPstat segítségével

Az eljárás illusztrálására egy  $N(50, 10)$  eloszlású véletlen változót generálunk 1000 esettel. Ez egy T-skálájú változót eredményez, amely a kérdőíves személyiségvizsgálatok standard skálátípusa. A T-skála 50-es átlaga és 10-es szórása biztosítja, hogy a T-értékek gyakorlatilag mindig a (0–100) övezeten belül (a legtöbb esetben 30 és 70 között) maradnak. Az  $N(50, 10)$  eloszlású véletlen változó generálását és hisztogramjának elkészítését a következő lépésekben tehetjük meg.

1. Indítsuk el a ROPstatot, de ne egy már korábban létrehozott fájlt hívjunk be, hanem kattintunk az „Új” parancsikonra. Ezzel megjelenik egy egyetlen cellából álló adattáblázat.



2. Ekkor a korábban már ismertetett módon (vö. 1.6. alfejezet) illesszünk be 999 új esetet („Esetek” parancsikon, majd ’Esetek beillesztése’ menüpont). Ezzel mindjárt 1000 sorosra bővült adattáblázatunk.
3. Ezután kattintsunk a „Transzformációk” parancsikonra és válasszuk ki itt az ’Új random változó’ menüpontot. Ennek hatására megjelenik a statisztikai eloszlások alábbi listája.
  - Egyváltozós normális
  - Kétváltozós normális
  - Egyenletes (diszkrét)
  - Egyenletes (folytonos)
  - Binomiális
  - Poisson
  - Lambda
  - Student t
  - Fisher F
  - Khi-négyzet
  - r és Z(r)
  - Kever normális
4. Válasszuk ki ebből a listából az elsőt, ami után a program rákérdez a generálandó változó eloszlásának paramétereire. Az itt megjelenő ’Átlag’ ablakba írjunk be 50-et, a ’Szórás’ ablakba pedig 10-et.
5. Ezután a „Futtat” parancsikonra kattintva nevet adhatunk az új véletlen változónak (alapértelmezés: RanNorm1).
6. A névadás után a „Folytatás” parancsikonra kattintva a program létrehozza az új véletlen változót annyi esetre, ahány sora van az adattáblázatunknak.
7. Ha akarjuk, a leírt módon akárhány új random változót generálhatunk, most azonban hagyjuk el a generálás ablakát (pl. ablak jobb felső sarkába kattintva).
8. A hisztogramot a 2.4.1. és 2.4.2. alpontban leírt módon készíthetjük el. Ez minden bizonynyal a 4.4. táblázatban láthatóhoz fog hasonlítani.

**4.4. táblázat.** Egy ROPstat segítségével generált normális eloszlású véletlen változó hisztogramja

| Gyakoriság, hisztogram                                     |      |      |       |       |
|------------------------------------------------------------|------|------|-------|-------|
| A beolvasott összes eset száma: 1000                       |      |      |       |       |
| VÁLTOZÓ: RanNorm1 (Random normális, átlag: 50; szórás: 10) |      |      |       |       |
| Oszt.köz.                                                  | Gyak | %    | Kum%  |       |
| 19,26                                                      | 2    | 0,2  | 0,2   |       |
| 23,58                                                      | 4    | 0,4  | 0,6   |       |
| 27,91                                                      | 13   | 1,3  | 1,9   | ==    |
| 32,23                                                      | 32   | 3,2  | 5,1   | ===== |
| 36,56                                                      | 64   | 6,4  | 11,5  | ===== |
| 40,89                                                      | 123  | 12,3 | 23,8  | ===== |
| 45,21                                                      | 149  | 14,9 | 38,7  | ===== |
| 49,54                                                      | 180  | 18,0 | 56,7  | ===== |
| 53,86                                                      | 168  | 16,8 | 73,5  | ===== |
| 58,19                                                      | 133  | 13,3 | 86,8  | ===== |
| 62,51                                                      | 60   | 6,0  | 92,8  | ===== |
| 66,84                                                      | 53   | 5,3  | 98,1  | ===== |
| 71,16                                                      | 13   | 1,3  | 99,4  | ==    |
| 75,49                                                      | 3    | 0,3  | 99,7  |       |
| 79,81                                                      | 3    | 0,3  | 100,0 |       |
| Össz.:                                                     | 1000 |      |       |       |

### 4.3.2. Az egyenletes eloszlás

Ha egy  $X$  változó értékei mind ugyanakkora valószínűséggel fordulnak elő, akkor  $X$ -et **egyenletes eloszlású** változónak nevezzük.

Ha  $X$  diszkrét és lehetséges különböző értékeit  $x_1, x_2, \dots, x_k$  jelöli, akkor tehát az eloszlás egyenletes volta azt jelenti, hogy

$$P(x_1) = P(x_2) = \dots = P(x_k) = 1/k. \quad (4.9)$$

Egyenletes például a 3.9. ábra jobb felső sarkában látható diszkrét eloszlás. Ha  $X$  egy szabályos dobókocka dobását jelző szám (a szokásos 1, 2, 3, 4, 5, 6 értékekkel), akkor  $X$  egyenletes eloszlású diszkrét változó.

Ha  $X$  folytonos és lehetséges értékei a számegyenes valamely  $[a, b]$  szakaszát töltik ki, akkor az eloszlás egyenletessége azt jelenti, hogy tetszőleges, az  $[a, b]$  szakasz részét képező  $[x, y]$  szakasz esetén az  $x$  és az  $y$  érték közé esés valószínűsége e szakasz hosszával, vagyis  $(y-x)$ -szel arányos:

$$P(x < X < y) = (y - x)/(b - a) \quad (4.10)$$

$X$ -et ilyenkor az  **$[a, b]$  intervallumon egyenletes eloszlású** változónak nevezzük. Egyenletes például a 3.7. ábra bal felső sarkában látható folytonos eloszlás. Folytonos egyenletes eloszlású véletlen változóhoz jutunk például, ha egy hagyományos óra egyenletes sebességgel mozgó másodpercmutatóját egy véletlenszerű időpontban megállítjuk, s megnézzük, mennyi idő telt el azóta, hogy ez a mutató a 12 óránál állt.

A ROPstatban egyenletes eloszlású *diszkrét* változót ugyanúgy lehet generálni, mint normális eloszlásút (lásd alfejezetünk elejét), csak a „Transzformációk” menüparancsra kattintva a megjelenő eloszlástípusok közül az 'Egyenletes (diszkrét)' menüsort kell kiválasztani, majd megadni a változó lehetséges legkisebb és legnagyobb értékét (például az 1-et és az 5-öt).

Egyenletes eloszlású *folytonos* véletlen változót a ROPstatban ugyanúgy képezhetünk, mint diszkrét egyenletest, csak a megjelenő eloszlástípusok közül a 'Egyenletes (folytonos)' menüsort kell kiválasztani, majd megadni a változó lehetséges legkisebb és legnagyobb értékét.

Meg kell jegyeznünk, hogy az egyenletes eloszlás lényegében csak elméleti jelentőségű. Sőt, hozzá kell tenni, hogy amit az ember naivan egyenletesnek gondolna, azt a statisztikában a normális eloszlás testesíti meg.

### 4.3.3. A Poisson-eloszlás

Ha egy  $X$  diszkrét változó lehetséges értékei a

$$0, 1, 2, 3, \dots, k, k+1, \dots$$

egész számok, amelyek  $P(k)$  előfordulási valószínűségei a

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (k = 0, 1, 2, \dots) \quad (4.11)$$

képlet által határozhatók meg ( $e$  képletben  $k! = 1 \cdot 2 \cdot \dots \cdot (k-1) \cdot k$ , illetve  $0! = 1$ ;  $e \approx 2.7183$  a természetes logaritmus alapszáma,  $\lambda$  pedig egy tetszőleges, de rögzített pozitív szám), akkor  $X$ -et  **$\lambda$  paraméterű Poisson-eloszlású** változónak nevezzük.

Poisson-eloszlású változóval a természetben akkor találkozunk, ha egy gyakorlatilag végtelennek tekinthető (pl. idői vagy téri) tartományban véletlenszerűen, de állandó sűrűséggel eloszló pontok (részecskék, elemek, időpontok stb.) közül kell összeszámlálni azokat, amelyek egy előre kijelölt véges résztartományba esnek. Megközelítőleg Poisson-eloszlást követ:

- valamely radioaktív anyagban adott idő alatt elbomló atomok száma;
- egy telefonközpontba adott időszakban, adott napszakban, meghatározott időtartam alatt beérkező hívások száma;
- adott személy 1 ml vérében található fehérvérsejtek vagy vörösvértestek száma;
- valamely üzletbe adott időszakban, adott napszakban, meghatározott időtartam alatt érkező vásárlók száma.

A Poisson-eloszlás néhány fontosabb jellemzőjét az alábbiakban foglaljuk össze.

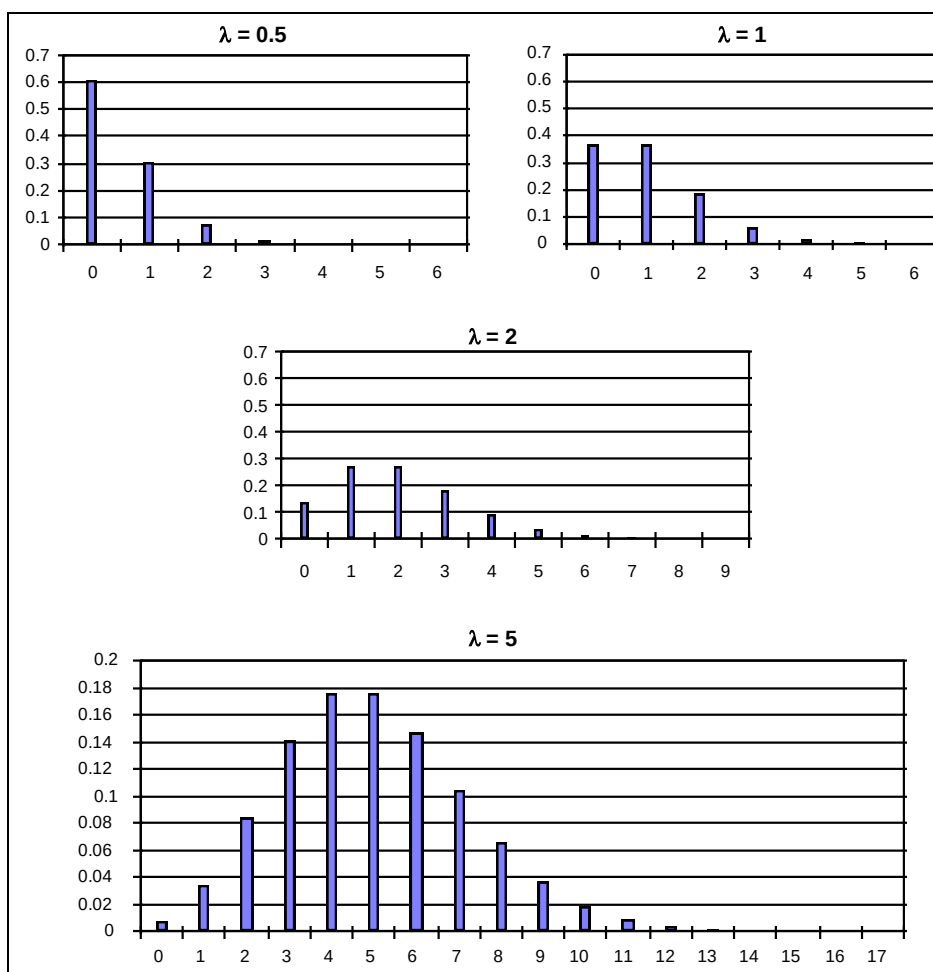
1. Ha az  $X$  változó  $\lambda$  paraméterű Poisson-eloszlást követ, akkor elméleti átlaga és varianciája egyaránt  $\lambda$ -val egyenlő:

$$E(X) = \lambda \quad \text{és} \quad \text{Var}(X) = \lambda. \quad (4.12)$$

2. Ha  $X_1, X_2, \dots, X_n$  páronként független Poisson-eloszlású változók, rendre  $\lambda_1, \lambda_2, \dots, \lambda_n$  paraméterrel, akkor összegük is Poisson-eloszlású változó lesz, éspedig

$$\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_n$$

paraméterrel.



**4.9. ábra.** A Poisson-eloszlás a  $\lambda$  paraméter néhány értékére

3. Ha a  $\lambda$  paraméter elég nagy (mondjuk  $\lambda > 30$ ), akkor a Poisson-eloszlás jól közelíthető egy olyan normális eloszlással, amelynek elméleti átlaga és varianciája egyaránt  $\lambda$ .

Illusztrációképpen a Poisson-eloszlást néhány  $\lambda$ -ra a 4.9. ábrán szemléltetjük. Poisson-eloszlású véletlen változót a ROPstatban a „Transzformációk” parancsikon segítségével képezhetünk.

#### 4.3.4. A binomiális eloszlás

Ha egy  $X$  diszkrét változó lehetséges értékei a

$$0, 1, 2, 3, \dots, n-1, n$$

egész számok, amelyek  $P(k)$  előfordulási valószínűségei a

$$P(k) = \frac{n!}{k! \cdot (n-k)!} p^k (1-p)^{n-k} \quad (k = 0, 1, 2, \dots, n) \quad (4.13)$$

képlet által határozhatók meg (e képletben  $n!$ ,  $k!$ , illetve  $(n-k)!$  rendre az első  $n$ ,  $k$ , illetve  $n-k$  természetes szám szorzata,  $0! = 1$  figyelembevételével;  $n$  tetszőleges pozitív egész szám,  $p$  pedig tetszőleges 0 és 1 közti pozitív szám), akkor  $X$ -et **(n, p) paraméterű binomiális eloszlású** változónak nevezzük.

Binomiális eloszlású például rulett játékban a fekete szín bekövetkezéseinek a száma 25 pörgetésből, egy pénzdarab 10-szeri (egymástól független) feldobása során a fejek száma, kockadobásnál 100 dobásból a hatosok száma, a nők száma egy 20 fős mintában, ha a mintát véletlenül választottuk ki egy gyakorlatilag végtelen nagy populációból stb.

Binomiális eloszlású változóhoz általában úgy juthatunk, ha valamely  $A$  eseményre vonatkozóan egymás után, pontosan ugyanolyan körülmények között, több független kísérletet hajtunk végre, majd megszámláljuk, hogy az  $A$  esemény hányszor következett be. Ez esetben az  $n$  paraméter az  $A$ -ra vonatkozóan elvégzett független kísérletek számát,  $p$  pedig a  $P(A)$  valószínűséget, vagyis egyetlen kísérlet során  $A$  bekövetkezésének esélyét adja meg.

Konkrét példaként tekintsük a következő  $X$  változót! Dobjunk fel egy szabályos dobókockát egymás után, egymástól függetlenül háromszor, majd számoljuk meg, hogy hány hatost dobtunk! A hatos dobások számát jelöljük  $X$ -szel és határozzuk meg az  $X$  változó eloszlását! A kísérlet feltételei alapján megállapíthatjuk, hogy az  $X$  változó binomiális eloszlású,  $n = 3$  (háromszor dobtunk!) és  $p = 1/6$  (a hatos dobás valószínűsége szabályos kocka esetén) paraméterrel, így ezt az eloszlást a (4.13) képlet segítségével határozhatjuk meg a következőképpen (3 tizedes jegyre kerekítve):

$$P(0) = \frac{3!}{0! \cdot 3!} \left(\frac{1}{6}\right)^0 \cdot \left(\frac{5}{6}\right)^3 = 1 \cdot 1 \cdot \frac{5^3}{6^3} = 0.579$$

$$P(1) = \frac{3!}{1! \cdot 2!} \left(\frac{1}{6}\right)^1 \cdot \left(\frac{5}{6}\right)^2 = \frac{1 \cdot 2 \cdot 3}{1 \cdot (1 \cdot 2)} \cdot \frac{1}{6} \cdot \frac{5^2}{6^2} = \frac{3 \cdot 5^2}{6^3} = 0.347$$

$$P(2) = \frac{3!}{2! \cdot 1!} \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^1 = \frac{1 \cdot 2 \cdot 3}{(1 \cdot 2) \cdot 1} \cdot \frac{1}{6^2} \cdot \frac{5}{6} = \frac{3 \cdot 5}{6^3} = 0.069$$

$$P(3) = \frac{3!}{3! \cdot 0!} \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^0 = 1 \cdot \frac{1}{6^3} \cdot 1 = 0.005.$$

Ha ezt a négy valószínűséget összeadjuk, 1-et kapunk eredményül, ami azt mutatja, hogy feltehetőleg jól számoltunk. Érdekes megfigyelni, hogy határozottan nagyobb 0.5-nél annak a valószínűsége, hogy 3 dobásból egy hatost sem dobunk.

A binomiális eloszlás néhány fontosabb jellemzőjét az alábbiakban foglaljuk össze.

1. Ha  $X(n, p)$  paraméterű binomiális eloszlású véletlen változó, akkor elméleti átlaga  $n \cdot p$ , varianciája pedig  $n \cdot p \cdot (1-p)$ :

$$E(X) = n \cdot p \quad \text{és} \quad \text{Var}(X) = n \cdot p \cdot (1 - p). \quad (4.14)$$

2. Ha  $X_m$  és  $X_n$  egymástól független  $(m, p)$ , illetve  $(n, p)$  paraméterű binomiális eloszlású változók, akkor összegük, az  $X = X_m + X_n$  változó is binomiális eloszlású változó lesz, és pedig  $(N, p)$  paraméterrel, ahol  $N = m+n$ .

3. Bármely  $p$  esetén, ha  $n$  elég nagy (mondjuk  $n > 9/[p \cdot (1-p)]$ ), akkor az  $(n, p)$  paraméterű binomiális eloszlás igen jól közelíthető egy olyan normális eloszlással, amelynek elméleti átlaga  $n \cdot p$ , varianciája pedig  $n \cdot p \cdot (1-p)$  (szemléltetésképpen lásd a 4.5. táblázat 2. eloszlását).

Binomiális eloszlású véletlen változót a ROPstatban ugyanúgy képezhetünk, mint diszkrét egyenletes vagy Poisson eloszlásút (vö. 4.3.2. alpont). A binomiális eloszlás alakjának szemléltetésére a ROPstat segítségével generáltunk két véletlen binomiális eloszlású változót ( $n = 4$  és  $p = 0.1$ , illetve  $n = 10$  és  $p = 0.4$  paraméterrel) 10 000 véletlen megfigyeléssel, majd elkészítettük hisztogramjukat (lásd 4.5. táblázat).

**4.5. táblázat.** Két binomiális eloszlású véletlen változó ROPstat segítségével készített hisztogramja

| VÁLTOZÓ: RanBin1 (Random binomiális: n = 4; p = 0,1)  |       |      |       |       |
|-------------------------------------------------------|-------|------|-------|-------|
| Érték                                                 | Gyak  | %    | Kum%  |       |
| 0                                                     | 6478  | 64,8 | 64,8  | ===== |
| 1                                                     | 2997  | 30,0 | 94,8  | ===== |
| 2                                                     | 488   | 4,9  | 99,6  | ==    |
| 3                                                     | 37    | 0,4  | 100,0 |       |
| Össz.:                                                | 10000 |      |       |       |
| VÁLTOZÓ: RanBin2 (Random binomiális: n = 10; p = 0,5) |       |      |       |       |
| Érték                                                 | Gyak  | %    | Kum%  |       |
| 0                                                     | 7     | 0,1  | 0,1   |       |
| 1                                                     | 102   | 1,0  | 1,1   | =     |
| 2                                                     | 471   | 4,7  | 5,8   | ===== |
| 3                                                     | 1166  | 11,7 | 17,5  | ===== |
| 4                                                     | 2015  | 20,1 | 37,6  | ===== |
| 5                                                     | 2442  | 24,4 | 62,0  | ===== |
| 6                                                     | 2119  | 21,2 | 83,2  | ===== |
| 7                                                     | 1109  | 11,1 | 94,3  | ===== |
| 8                                                     | 452   | 4,5  | 98,8  | ===== |
| 9                                                     | 105   | 1,1  | 99,9  | =     |
| 10                                                    | 12    | 0,1  | 100,0 |       |
| Össz.:                                                | 10000 |      |       |       |

### 4.3.5. A kevert normális eloszlás

Gyakori eset, hogy egy populáció több markánsan különböző alpopulációra bontható (például férfiakra és nőkre, alsó-, közép- és felsőfokú végzettségűekre, többségi és kisebbségi anyanyelvűekre stb.). Mi történik akkor, ha ezt a heterogenitást egy  $X$  változó vizsgálatakor nem vesszük figyelembe, hanem a populációt egységes összességnek tekintjük?

Ilyen esetben az  $X$  változó eloszlása még akkor sem lesz feltétlenül normális, ha  $X$  minden alpopulációban normális eloszlást követ. Jól látható ez például a 4.3. ábra bal felső grafikonján. Ez egy olyan  $X$  változó eloszlását ábrázolja, amely  $N(0, 1)$  eloszlású a vizsgálati populáció 70%-ában, és  $N(4, 1)$  eloszlású a populáció 30%-ában, vagyis az utóbbiban az  $X$  változó elméleti átlaga 4 egységgel magasabb. Egy ilyen több normális eloszlásból összetevődő eloszlást **kevert normális eloszlásnak** nevezünk.

**4.6. táblázat.** Két kevert normális eloszlású véletlen változó ROPstat segítségével készített hisztogramja. Az MixNor1 változó egy  $N(0, 1)$  és egy  $N(0, 3)$  eloszlás keveréke 95%-5% arányban, MixNor2 pedig egy  $N(0, 1)$  és egy  $N(3, 3)$  eloszlás keveréke, szintén 95%-5% arányban

| VÁLTOZÓ: MixNor1 (Random kevert normális) |      |      |       |       |
|-------------------------------------------|------|------|-------|-------|
| Oszt.köz.                                 | Gyak | %    | Kum%  |       |
| -6,18                                     | 2    | 0,2  | 0,2   |       |
| -5,27                                     | 2    | 0,2  | 0,4   |       |
| -4,35                                     | 1    | 0,1  | 0,5   |       |
| -3,43                                     | 3    | 0,3  | 0,8   |       |
| -2,52                                     | 19   | 1,9  | 2,7   | ==    |
| -1,60                                     | 116  | 11,6 | 14,3  | ===== |
| -0,69                                     | 255  | 25,5 | 39,8  | ===== |
| 0,23                                      | 334  | 33,4 | 73,2  | ===== |
| 1,15                                      | 220  | 22,0 | 95,2  | ===== |
| 2,06                                      | 41   | 4,1  | 99,3  | ===== |
| 2,98                                      | 5    | 0,5  | 99,8  |       |
| 3,89                                      | 1    | 0,1  | 99,9  |       |
| 4,81                                      | 0    | 0,0  | 99,9  |       |
| 5,73                                      | 0    | 0,0  | 99,9  |       |
| 6,64                                      | 1    | 0,1  | 100,0 |       |
| Össz.:                                    | 1000 |      |       |       |
| VÁLTOZÓ: MixNor2 (Random kevert normális) |      |      |       |       |
| Oszt.köz.                                 | Gyak | %    | Kum%  |       |
| -2,42                                     | 20   | 2,0  | 2,0   | ==    |
| -1,60                                     | 92   | 9,2  | 11,2  | ===== |
| -0,79                                     | 224  | 22,4 | 33,6  | ===== |
| 0,03                                      | 297  | 29,7 | 63,3  | ===== |
| 0,85                                      | 222  | 22,2 | 85,5  | ===== |
| 1,66                                      | 91   | 9,1  | 94,6  | ===== |
| 2,48                                      | 33   | 3,3  | 97,9  | ===== |
| 3,30                                      | 7    | 0,7  | 98,6  |       |
| 4,11                                      | 1    | 0,1  | 98,7  |       |
| 4,93                                      | 2    | 0,2  | 98,9  |       |
| 5,75                                      | 3    | 0,3  | 99,2  |       |
| 6,56                                      | 2    | 0,2  | 99,4  |       |
| 7,38                                      | 1    | 0,1  | 99,5  |       |
| 8,20                                      | 1    | 0,1  | 99,6  |       |
| 9,01                                      | 4    | 0,4  | 100,0 |       |
| Össz.:                                    | 1000 |      |       |       |

Ha egy kevert normális eloszlás két összetevője csak csekély mértékben különbözik egymástól (vagyis ha nagyon hasonló az elméleti átlaguk és szórásuk), akkor a közös eloszlás nem különbözik jelentősen a normálistól. Néha azonban egy kevert normális eloszlás hisztogramja még számottevően eltérő összetevők esetén sem jelzi egyértelműen, hogy a populáció

valójában heterogén együttes. Illusztrációképpen nézzük meg a 4.6. táblázat két hisztogramját, amelyet két ROPstattal generált kevert normális eloszlású változóról készítettünk! Az első változó (MixNor1) egy  $N(0, 1)$  és egy  $N(0, 3)$  eloszlás 95%–5%-os arányú keveréke. Ez olyan helyzetnek felelhet meg, amikor egy változó mérésekor az esetek 5%-ában véletlenszerűen az alapszóródáshoz viszonyítva háromszoros ingadozású értékek fordulnak elő az eloszlás mindkét oldalán szimmetrikusan. A második változó (MixNor2) egy  $N(0, 1)$  és egy  $N(3, 3)$  eloszlás ugyancsak 95%-5%-os arányú keveréke. Ez pedig egy olyan helyzetnek felelhet meg, amikor egy változó mérésekor az esetek 5%-ában véletlenszerűen olyan műtermék adatok jelennek meg, amelyek szintje 3 szórással magasabb, mint az eredeti változóé, s ingadozása is háromszoros. A 4.6. táblázat hisztogramja alapján hajlamosak lennénk azt mondani, hogy az ábrázolt változók közel normális eloszlásúak. Ha azonban kiszámítjuk csúcsossági együtthatójukat (a ROPstatban az 'Alapok' programcsoport 'Részletesebb statisztikák' rutinjával megtehető), akkor a kapott  $g_2(\text{MixNor1}) = 4,38$  és  $g_2(\text{MixNor2}) = 10,64$  csúcsossági értékek jól jelzik a két eloszlásnál az extrém értékek nem elhanyagolható arányú jelenlétét.

#### 4.3.6. A khi-négyzet-eloszlás

Egy ELTÉ-n oktató tanár azt állítja, hogy több évtizedes személyes tapasztalata alapján az ELTÉ-re járó egyetemi hallgatók intelligenciája nem nagyobb a magyar átlagnál, s még az IQ-szerinti variabilitás sem tér el a teljes magyar populációtól. Gondolatban végezzünk el egy ellenőrző vizsgálatot! Tételezzük fel, hogy az ELTE-hallgatók közül egy véletlen 25 fős mintát kiválasztva azt kapjuk, hogy a minta átlaga 110.5 IQ-pont és szórása 12 IQ-pont! Ha a tanár úrnak igaza van, akkor milyen valószínűséggel lesz egy 25-elemű véletlen mintában ilyen vagy ennél szélsőségesebb az átlag és a szórás?

Először az átlagot vegyük szemügyre! Tudjuk, hogy a mintaátlag véletlen mintavétel esetén véletlen változó, amely  $X$  normalitása, illetve a mintaelemszám elegendően nagy volta esetén normális eloszlású,  $E(\bar{x}) = E(X)$  elméleti átlaggal és  $D(\bar{x}) = D(X)/\sqrt{n}$  elméleti szórással (vö. 4.2.2. alpont). Esetünkben maga az IQ is közel normális eloszlású, s mivel  $n = 25$  itt elegendően nagyra tekinthető (vö. 4.3. és 4.4. ábra), az IQ-átlag normális eloszlásában nem szükséges kételkednünk. Mivel tehát az IQ közel  $N(100, 15)$  eloszlású, az  $\bar{x}$  mintaátlagra a fentiek szerint igaz az, hogy

$$E(\bar{x}) = 100 \text{ és } D(\bar{x}) = 15/\sqrt{25} = 15/5 = 3.$$

Emiatt az  $\bar{x}$  mintaátlag most  $N(100, 3)$  eloszlású változó. Ebben az eloszlásban a kapott 110.5 IQ-átlagnak megfelelő standard érték:

$$z = (110.5 - 100)/3 = 10.5/3 = 3.5.$$

A standard normális eloszlásban 3.5-nél kisebb érték az esetek 99.98%-ában fordul elő (a Melléklet 1. táblázatában a  $z = 3.50$  mellett  $p = 0.9998$  található). Következésképpen annak a valószínűsége, hogy egy standard normális eloszlású változó 3.5-nél nagyobb értékű legyen, csak 0.02%, ami egyben annak a valószínűsége is, hogy egy  $N(100, 3)$  eloszlású változó a saját átlagánál legalább 3.5 szórással, vagyis esetünkben legalább 10.5-tel nagyobb értékű legyen. Ha tehát a tanár úrnak igaza van, akkor egy 25 fős véletlen minta IQ-átlaga 0.0002 = 0.02% valószínűséggel lesz 110.5 vagy annál nagyobb. Ez olyan kis valószínűségű esemény, ami átlagosan csak minden húszezredik véletlen mintavétel esetén következik be. Lehet, hogy

a hallgatók IQ-szintje tekintetében a tanár úr tévedett? Az eredmények ismeretében szinte bizonyos.

Mit mondhatunk a szórással kapcsolatos feltételezésről? Milyen valószínűségű egy 12 vagy ennél nagyobb mintaszórásnak a fellépte, ha az IQ elméleti szórása 15 és a minta-elemszám:  $n = 25$ ? Ehhez jó lenne tudni, hogy a mintaszórás, mint véletlen változó milyen eloszlású. Ha a mintaszórás eloszlását a mintaszórás képletében szereplő négyzetgyökök miatt (vö. (3.17) formula) nehéz is matematikailag pontosan meghatározni, négyzete, a mintavariancia esetében már biztatóbb a helyzet. A mintavariancia definiáló képlete:

$$\text{Var}_x = \frac{\sum_{i=1}^n (X_i - \bar{x})^2}{f}$$

(vö. a (3.14) és a (3.15) formulával). Itt  $X_i$  az  $i$ -edik mintaelem (most azért írtuk nagy betűvel, hogy véletlen változó voltát hangsúlyozzuk),  $\bar{x}$  a mintaátlag és  $f = n-1$  a mintavariancia szabadságfoka. Ezt a képletet kissé átrendezve és mindkét oldalt az  $X$  változó  $\sigma^2$  elméleti varianciájával leosztva azt kapjuk, hogy

$$\frac{f \cdot \text{Var}_x}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{x})^2}{\sigma^2}.$$

Matematikailag igazolható, hogy ha  $X$  normális eloszlású, akkor a fenti összefüggés jobb oldala mindig átalakítható úgy, hogy  $f$  darab független standard normális eloszlású  $Z_i$  véletlen változó négyzetösszege legyen. Ez azt jelenti, hogy a  $\text{Var}_x$  mintavarianciához mindig található  $f$  darab olyan független standard normális eloszlású  $Z_i$  változó, amelyekre fennáll a következő összefüggés:

$$\frac{f \cdot \text{Var}_x}{\sigma^2} = \sum_{i=1}^f (Z_i)^2. \quad (4.15)$$

Az ilyen független  $N(0, 1)$  eloszlású változók négyzetösszegeként előálló változók eloszlását a matematikai statisztikában **khi-négyzet-eloszlásnak** (alternatív felírással  **$\chi^2$ -eloszlásnak**) nevezik, az összegzett tagok számát ( $f$ ) pedig a  $\chi^2$ -eloszlás **szabadságfokának**.

A fenti gondolatmenetnek tehát az a lényege, hogy az  $(n-1)/\sigma^2$  számmal megszorozott mintavariancia  $f = n-1$  szabadságfokú  $\chi^2$ -eloszlást követ.

A (4.15) összefüggés megvilágítja a tapasztalati variancia  $f$  szabadságfokának jelentését is: ennyi független négyzetes eltérés komponensre bontható a  $Q_x = f \cdot \text{Var}_x$  négyzetes összeg. Így  $\text{Var}_x = Q_x/f$  alapján a tapasztalati variancia a standardizált (a  $\sigma^2$  elméleti varianciával leosztott) négyzetes eltérések átlaga.

A (4.15) formulából kiolvashatjuk, hogy egyetlen  $N(0, 1)$  eloszlású változó négyzete  $f = 1$  szabadságfokú  $\chi^2$ -eloszlást követ. Két független  $N(0, 1)$  eloszlású  $Z_1$  és  $Z_2$  változó  $Z = (Z_1)^2 + (Z_2)^2$  összege  $f = 2$  szabadságfokú  $\chi^2$ -eloszlást követ stb. Egy  $f$  szabadságfokú  $\chi^2$ -eloszlást röviden  $\chi^2(f)$ -eloszlásnak is szoktak nevezni.

Konkrét példánkhoz visszatérve a fentiek fényében azt mondhatjuk, hogy ha a tanár úrnak igaza van abban, hogy a ELTE-hallgatók populációjában az  $X = \text{IQ}$  változó  $N(100, 15)$  eloszlású, akkor a

$$Z = \frac{f \cdot \text{Var}_x}{\sigma^2} = \frac{(25-1) \cdot \text{Var}_x}{15^2} = \frac{24}{225} \cdot \text{Var}_x$$



változó  $f = 24$  szabadságfokú  $\chi^2$ -eloszlást követ. A  $\chi^2$ -eloszlás különböző szabadságfokokhoz tartozó egyes kvantilisei megtalálhatók a Melléklet 3. táblázatában. Jelen esetben a  $Z$  változó értéke:

$$Z = \frac{24}{225} \cdot 12^2 = 15.36.$$

A Melléklet 3. táblázatában az  $f = 24$  szabadságfok sorában azt találjuk, hogy

$$Q_{0.10} = 15.6587 \approx 15.66.$$

Ez azt jelenti, hogy

$$P(Z < 15.66) = 0.10,$$

vagyis ha egy változó 24 szabadságfokú  $\chi^2$ -eloszlást követ, akkor értékei az esetek kb. 10%-ában lesznek csak 15.66-nál kisebbek, s még ennél is ritkábban lesznek kisebbek 15.36-nál. Ha tehát igaz az a feltételezés, hogy az IQ szórása 15, akkor egy 25-fős véletlen minta szórása 10%-nál kisebb valószínűséggel lesz 12 vagy ennél kisebb, ami persze nem túl kis esély.

A  $\chi^2$ -eloszlás alapvető fontosságú a matematikai statisztikában, főleg a diszkrét változókkal kapcsolatos elemzésekben (lásd 17. fejezet).

A  $\chi^2$ -eloszlással kapcsolatban érdemes megemlíteni néhány további összefüggést.

1. Ha egy  $X$  változó  $f$  szabadságfokú  $\chi^2$ -eloszlást követ, akkor elméleti átlaga és varianciája a következőképpen határozható meg:

$$E(X) = f \quad \text{és} \quad \text{Var}(X) = 2f. \quad (4.16)$$

2. Ha  $X_f$  és  $X_g$  egymástól független  $f$ , illetve  $g$  szabadságfokú  $\chi^2$ -eloszlású változók, akkor  $X = X_f + X_g$  összegük is  $\chi^2$ -eloszlású változó lesz, és pedig  $N = f + g$  szabadságfokkal.

3. Az  $f$  szabadságfok növekedtével a  $\chi^2$ -eloszlás egyre jobban hasonlít a normálisra (vö. 4.7. táblázat). Ha  $f > 60$ , akkor a közelítés már igen jónak mondható.

Illusztrációképpen a 4.7. táblázatban mutatjuk be néhány  $\chi^2$ -eloszlású változó empirikus eloszlását. Ezeket a ROPstat segítségével generáltuk, 10 000 véletlen megfigyelésre vonatkozóan.

#### 4.3.7. A $t$ -eloszlás

A matematikai statisztika lényege, hogy egy véletlen minta alapján következtetéseket vonunk le a populációra, illetve a vizsgált véletlen változó eloszlására vonatkozóan. Ezért olyan fontos, hogy a minta statisztikai jellemzői milyen statisztikai eloszlásokat követnek. A mintaátlag eloszlásának a vizsgálata vezetett a statisztika talán legfontosabb eloszlás-típusához, a normális eloszláshoz (vö. 4.2.2. alpont). Egy szórással, illetve varianciával kapcsolatos kérdés

**4.7. táblázat.** Egy 2, egy 5 és egy 20 szabadságfokú khi-négyzet-eloszlású véletlen változó ROPstat segítségével készített hisztogramja

| VÁLTOZÓ: RanKhi1 (Random khi-négyzet, szab.fok = 2) |      |      |      |
|-----------------------------------------------------|------|------|------|
| Oszt.köz.                                           | Gyak | %    | Kum% |
| 0,61                                                | 4598 | 46,0 | 46,0 |
| 1,84                                                | 2496 | 25,0 | 70,9 |
| 3,06                                                | 1273 | 12,7 | 83,7 |
| 4,29                                                | 745  | 7,5  | 91,1 |
| 5,51                                                | 424  | 4,2  | 95,4 |

|        |       |     |       |   |
|--------|-------|-----|-------|---|
| 6,74   | 200   | 2,0 | 97,4  | = |
| 7,96   | 131   | 1,3 | 98,7  |   |
| 9,19   | 52    | 0,5 | 99,2  |   |
| 10,41  | 35    | 0,3 | 99,5  |   |
| 11,64  | 23    | 0,2 | 99,8  |   |
| 12,86  | 13    | 0,1 | 99,9  |   |
| 14,09  | 5     | 0,1 | 99,9  |   |
| 15,31  | 3     | 0,0 | 100,0 |   |
| 16,54  | 1     | 0,0 | 100,0 |   |
| 17,76  | 1     | 0,0 | 100,0 |   |
| Össz.: | 10000 |     |       |   |

VÁLTOZÓ: RankHi2 (Random khi-négyzet, szab.fok = 5)

| Oszt.köz. | Gyak  | %    | Kum%  |       |
|-----------|-------|------|-------|-------|
| 0,96      | 1406  | 14,1 | 14,1  | ===== |
| 2,83      | 2715  | 27,1 | 41,2  | ===== |
| 4,70      | 2418  | 24,2 | 65,4  | ===== |
| 6,56      | 1632  | 16,3 | 81,7  | ===== |
| 8,43      | 884   | 8,8  | 90,6  | ===== |
| 10,30     | 490   | 4,9  | 95,5  | ===== |
| 12,17     | 207   | 2,1  | 97,5  | ==    |
| 14,04     | 131   | 1,3  | 98,8  | =     |
| 15,91     | 70    | 0,7  | 99,5  |       |
| 17,77     | 26    | 0,3  | 99,8  |       |
| 19,64     | 14    | 0,1  | 99,9  |       |
| 21,51     | 6     | 0,1  | 100,0 |       |
| 23,38     | 0     | 0,0  | 100,0 |       |
| 25,25     | 0     | 0,0  | 100,0 |       |
| 27,12     | 1     | 0,0  | 100,0 |       |
| Össz.:    | 10000 |      |       |       |

VÁLTOZÓ: RankHi3 (Random khi-négyzet, szab.fok = 20)

| Oszt.köz. | Gyak  | %    | Kum%  |       |
|-----------|-------|------|-------|-------|
| 5,04      | 21    | 0,2  | 0,2   |       |
| 8,26      | 300   | 3,0  | 3,2   | ===== |
| 11,48     | 1030  | 10,3 | 13,5  | ===== |
| 14,70     | 1790  | 17,9 | 31,4  | ===== |
| 17,92     | 2031  | 20,3 | 51,7  | ===== |
| 21,14     | 1900  | 19,0 | 70,7  | ===== |
| 24,36     | 1293  | 12,9 | 83,7  | ===== |
| 27,58     | 823   | 8,2  | 91,9  | ===== |
| 30,80     | 433   | 4,3  | 96,2  | ===== |
| 34,02     | 196   | 2,0  | 98,2  | ==    |
| 37,24     | 110   | 1,1  | 99,3  | =     |
| 40,46     | 40    | 0,4  | 99,7  |       |
| 43,68     | 22    | 0,2  | 99,9  |       |
| 46,90     | 9     | 0,1  | 100,0 |       |
| 50,12     | 2     | 0,0  | 100,0 |       |
| Össz.:    | 10000 |      |       |       |

kapcsán jutottunk a  $\chi^2$ -eloszláshoz is. Ebben az alponban olyan újabb eloszlástípust ismertettünk, amely szintén összefüggésbe hozható a mintabeli jellemzőkkel.

Ha van egy standard normális eloszlású  $Y$ , és egy ettől független,  $f$  szabadságfokú  $\chi^2$ -eloszlású  $Z_f$  véletlen változónk, akkor a

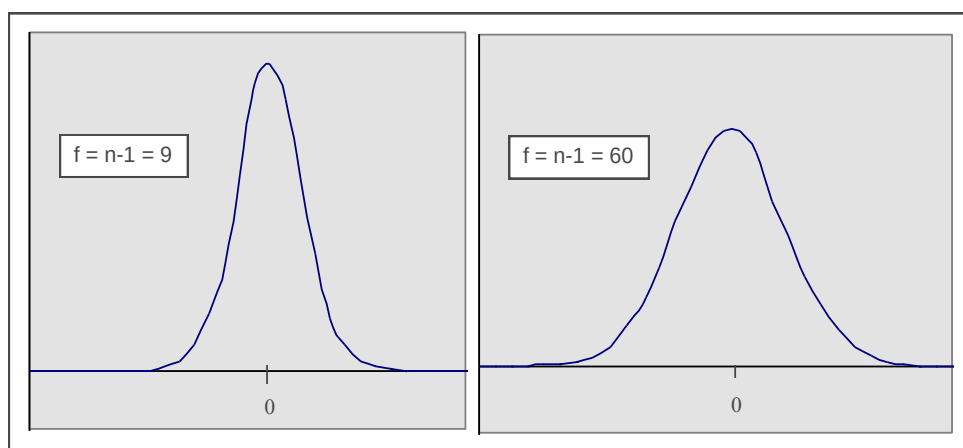
$$t = \frac{Y}{\sqrt{\frac{1}{f} Z_f}} \quad (4.17)$$

képlettel definiált véletlen változó eloszlását  $f$  szabadságfokú **t-eloszlásnak** vagy **Student eloszlásnak** nevezzük (lásd 4.10. ábra). Egy  $f$  szabadságfokú t-eloszlást röviden  $t(f)$ -eloszlásnak is szoktak nevezni. A t-eloszlás igen fontos szerephez jut majd az elméleti átlaggal kapcsolatos becslésekben és hipotézisvizsgálatokban (lásd 5.3.1. alpont, illetve 7.1, 8.1 és 9.2. alfejezet).

A t-eloszlás szimmetrikus, mint a normális eloszlás, de csúcsossága nagyobb, mint a normálisé. Kis szabadságfok esetén a csúcsossági mutató igen magas is lehet (például  $f = 5$  esetén  $\gamma_2 = 6$ ), a szabadságfok növekedésével azonban a t-eloszlás görbéje egyre jobban megközelíti a standard normális eloszlását. Ez a közelítés  $f > 60$  esetén már igen jónak mondható. A t-eloszlás elméleti átlaga csak 1-nél nagyobb szabadságfokra létezik, s ekkor szabadságfokától függetlenül 0. Varianciája pedig  $f > 2$  esetén létezik, s ekkor

$$\text{Var}(t_f) = f/(f - 2). \quad (4.18)$$

A t-eloszlás egyes kiválasztott kvantiliseit a Melléklet 4. táblázata tartalmazza. Itt az utolsó, végtelen szabadságfokhoz tartozó sor az  $N(0, 1)$  eloszlás kvantiliseit tartalmazza.



**4.10. ábra.** A t-eloszlás sűrűségfüggvénye két különböző szabadságfok esetén

Ha egy  $N(\mu, \sigma)$  eloszlású  $X$  változó értékeiből kiválasztunk véletlenszerűen  $n$  darabot, akkor ennek az  $n$ -elemű véletlen mintának az átlaga  $N(\mu, \sigma/\sqrt{n})$  eloszlású lesz (vö. 4.2.2. alpont), így ennek standardizáltja, az

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

változó szükségképpen  $N(0, 1)$  eloszlású. Az előző alpont fejtegetései alapján a fenti véletlen  $X$ -minta  $\text{Var}_x$  varianciájára  $X$  normalitása miatt igaz az, hogy a

$$Z = \frac{f \cdot \text{Var}_x}{\sigma^2}$$

változó  $f$  szabadságfokú  $\chi^2$ -eloszlást követ.

A matematikai statisztika módszereivel igazolható, hogy ha  $X$  normális eloszlású, akkor egy véletlen  $X$ -minta átlaga és szórása, mint véletlen változók, statisztikailag függetlenek egymástól. Emiatt ebben az esetben a

$$t = \frac{Y}{\sqrt{\frac{1}{f}} Z} = \frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{1}{f} \cdot \frac{f \cdot \text{Var}_x}{\sigma^2}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{\text{Var}_x}{n}}} = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}} \quad (4.19)$$

változó  $f = n-1$  szabadságfokú t-eloszlást követ.

Ahhoz, hogy a t-statisztika t-eloszlást kövessen, az  $X$  változó normalitása fontos, különösen kis minták esetén.

Tekintve, hogy az  $\bar{x}$  mintaátlag elméleti varianciája  $\sigma^2/n$ , a (4.19) formula nevezőjében lévő ezzel analóg

$$\text{Var}_{\bar{x}} = \text{Var}_x/n = s^2/n \quad (4.20)$$

statisztikai jellemzőt a **mintaátlag tapasztalati varianciájának**, ennek négyzetgyökét, az

$$s_{\bar{x}} = s/\sqrt{n} \quad (4.21)$$

mennyiséget pedig a **mintaátlag tapasztalati szórásának** vagy **standard hibájának** nevezzük. A standard hiba az elméleti átlag becslésekor jut majd fontos szerephez (vö. 5. fejezet).

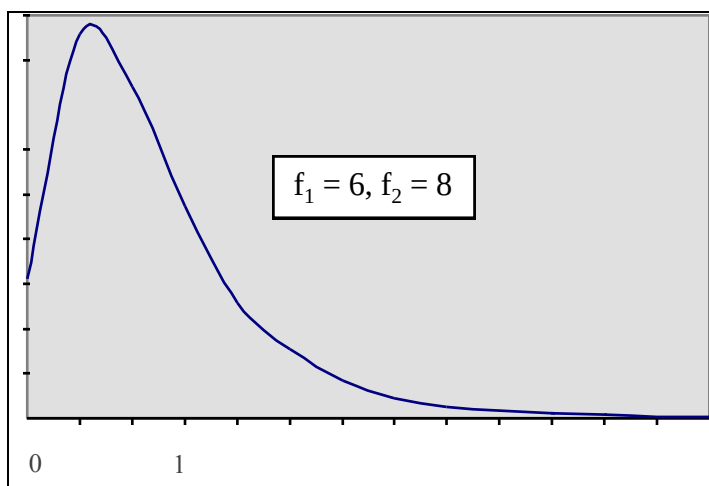
#### 4.3.8. Az F-eloszlás

Ha  $Z_f$   $f$  szabadságfokú  $\chi^2$ -elozzlású,  $Z_g$  pedig  $g$  szabadságfokú  $\chi^2$ -elozzlású változó és teljesül az, hogy  $Z_f$  és  $Z_g$  egymástól függetlenek, akkor az

$$F = \frac{\frac{1}{f} Z_f}{\frac{1}{g} Z_g} \quad (4.22)$$

változó eloszlását **(f, g) szabadságfokú F-eloszlásnak**, vagy röviden **F(f, g)-eloszlásnak** nevezzük.

Egy F-eloszlású változó csak pozitív értékeket vehet fel. Az eloszlás balra többnyire betömörül a 0 köré, jobbra pedig hosszan elnyúlik. Illusztrációképpen a 4.11. ábrán egy  $F(6, 8)$ -eloszlás sűrűségfüggvényét láthatjuk. A ROPstatban F-eloszlású véletlen változók is generálhatók különböző  $(f, g)$  szabadságfokpárookra ugyanolyan módon, mint a korábban ismertetett eloszlások.



**4.11. ábra.** Az  $f_1 = 6$ ,  $f_2 = 8$  szabadságfokú F-eloszlás sűrűségfüggvénye

Ha egy normális eloszlásból egymás után, egymástól függetlenül kiválasztunk egy  $n_1$ - és egy  $n_2$ -elemű véletlen mintát, akkor a nyert adatok variáciáját mindkét mintában kiszámítva hányadosuk, az

$$F = \frac{\text{Var}_1}{\text{Var}_2} \quad (4.23)$$

menyiség F-eloszlást követ  $f_1 = n_1 - 1$ ,  $f_2 = n_2 - 1$  szabadságfokokkal. Röviden: két független normális eloszlásból származó mintavariancia hányadosa mindig F-eloszlást követ, ha elméleti variáciájuk megegyezik (ez most azért teljesül, mert a két mintát ugyanabból a populációból választottuk ki). Ennek bizonyítása nagyon egyszerű, a (4.15) és a (4.22) formulából közvetlenül adódik. A normalitás itt ugyanolyan fontos feltétel, mint a  $\chi^2$ -eloszlással kapcsolatos (4.15) formula esetében. Az F-eloszlás igen fontos szerephez jut majd az elméleti átlaggal és szórással kapcsolatos hipotézisvizsgálatokban (lásd 9.7. alfejezet, illetve 13., 14. és 16. fejezet).

A (4.17) és a (4.22) formula összevetésével megállapíthatjuk azt az érdekes összefüggést, hogy egy  $t(f)$ -eloszlású változó négyzete mindig F-eloszlású, éspedig  $f_1 = 1$ ,  $f_2 = f$  szabadságfokokkal:

$$[t(f)]^2 = F(1, f). \quad (4.24)$$

### Normalitásvizsgálat a ferdeségi és a csúcsossági együttható segítségével

A vizsgált  $X$  változó eloszlásának nem normális volta számos statisztikai eljárás esetén kétséggé teheti az eljárás érvényességét. A normalitástól való eltérés hatása némely esetben (például az átlagra vonatkozó intervallumbecslés esetén) kisebb mértékű, más esetben (például a szórásra és a variációra vonatkozó intervallumbecslés során) igen erős. Jó lenne ezért, ha lenne valamilyen statisztikai támpontunk annak megítélésére, hogy mikor kételkedhetünk egy  $X$  kvantitatív változó normalitásában.

Megjegyezzük, hogy ha bizonyos statisztikai elemzés arra az eredményre vezet, hogy a normalitást nincs okunk elutasítani, akkor ez nem jelenti azt, hogy a vizsgált változó igazoltan normális eloszlású. Nem túlságosan kicsi elemszámok esetén azonban már ez is elegendő ah-

hoz, hogy a normalitásra építő statisztikai módszerek (becslések és hipotézisvizsgálatok) ne torzítsanak túl nagy mértékben a normalitás esetleges megsértése miatt.

Szélsőséges esetben az eloszlás nem normális volta a változó hisztogramjáról egy pillanat alatt kétség nélkül megállapítható (lásd például a 2.11. táblázatban a 'SzopHón' változó hisztogramját). Vannak azonban olyan esetek is, amikor a hisztogram önmagában nem alkalmas erre a feladatra (lásd például a 4.6. táblázat két hisztogramját).

Ebben az alponthoz a ferdeségi és a csúcsossági együttható felhasználásával fogunk a vizsgált változó normalitásáról véleményt formálni. Gondolatmenetünk az alábbi összefüggéseken alapul.

1. Normális eloszlású változók ferdeségi és csúcsossági együtthatója nulla:

$$\gamma_1 = 0 \quad \text{és} \quad \gamma_2 = 0$$

(vö. 3.4.2 és 3.4.3. alpont).

2. Normális eloszlású változók esetén a  $g_1$  tapasztalati ferdeségi együttható standard hibája:

$$\sigma_1 = D(g_1) = \sqrt{\frac{6}{n}}, \quad (5.26)$$

a  $g_2$  tapasztalati csúcsossági együttható standard hibája pedig:

$$\sigma_2 = D(g_2) = \sqrt{\frac{24}{n}} \quad (5.27)$$

(lásd Cramér, 1946, 375. o.; idézi Dixon, 1990, 137. o.).

3. Az (5.26) és az (5.27) formula érvényessége igen erősen függ az  $X$  változó normalitásától.

A fenti megállapítások alapján a következőképpen lehet egy eloszlás nem normális voltára az intervallumbecslés technikájának alkalmazásával következtetni.

Először is rögzítsünk egy megbízhatósági szintet ugyanúgy, mint az intervallumbecslések során szoktuk, legyen ez mondjuk  $1-\alpha$ . Ezzel választjuk meg a normalitást esetleg elutasító következtetésünk megbízhatóságának a mértékét. Például  $\alpha = 5\%$  a normalitás elutasítása esetén 95%-os megbízhatóságú döntést garantál.

Ha  $X$  normális eloszlású és az  $n$  mintaelemszám elég nagy (mondjuk  $n > 30$ ), akkor  $g_1$  és  $g_2$  közel normális eloszlású, továbbá standardizáltjaik, a

$$Z_1 = \frac{g_1}{\sigma_1}, \quad Z_2 = \frac{g_2}{\sigma_2}$$

változók pedig közel  $N(0, 1)$  eloszlásúak. Emiatt a  $g_1/\sigma_1$ ,  $g_2/\sigma_2$  értékeknek  $1-\alpha$  valószínűséggel az  $N(0, 1)$  eloszlás középső  $1-\alpha$  valószínűségű tartományában, vagyis  $-u_{\alpha/2}$  és  $u_{\alpha/2}$  között kell elhelyezkedniük:

$$P(-u_{\alpha/2} \leq \frac{g_1}{\sigma_1} \leq u_{\alpha/2}) = 1 - \alpha$$

és

$$P(-u_{\alpha/2} \leq \frac{g_2}{\sigma_2} \leq u_{\alpha/2}) = 1 - \alpha$$

(vö. 5.3. ábra). Ez  $\sigma_1$  és  $\sigma_2$  értékének behelyettesítésével (vö. (5.26) és (5.27)) és egyszerű átrendezéssel a következőt jelenti  $g_1$ -re és  $g_2$ -re:

$$P(-u_{\alpha/2} \cdot \sqrt{\frac{6}{n}} \leq g_1 \leq u_{\alpha/2} \cdot \sqrt{\frac{6}{n}}) = 1 - \alpha \quad (5.28)$$

és

$$P(-u_{\alpha/2} \cdot \sqrt{\frac{24}{n}} \leq g_2 \leq u_{\alpha/2} \cdot \sqrt{\frac{24}{n}}) = 1 - \alpha. \quad (5.29)$$

Ha  $X$  normális eloszlású, akkor  $g_1$ -re és  $g_2$ -re teljesülnie kell az (5.28)-ban és (5.29)-ben felírt egyenlőtlenségnek, éspedig  $1-\alpha$  megbízhatósággal. Ha bármelyikük is kívül esik a rá vonatkozó kijelölt határokon, az vagy a véletlen műve, aminek a valószínűsége  $\alpha$  (például  $1-\alpha = 0.95$  esetén  $\alpha = 0.05$ , azaz 5%), vagy pedig annak a következménye, hogy  $X$  nem normális eloszlású. Ha (5.28), illetve (5.29) nem teljesülése esetén mindig úgy következtetünk, hogy  $X$  *nem* normális eloszlású, akkor csupán  $\alpha$  valószínűségű (például 5%) lesz a tévedésünk esélye, így ekkor  $1-\alpha$  valószínűséggel helyesen döntünk. Ez a valószínűség itt megint úgy érthető, hogy ha a leírt eljárás alapján mondjuk 95%-os megbízhatósági szinten 100 független vizsgálatban azt a döntést hozzuk, hogy  $X$  nem normális eloszlású, akkor e 100 döntés közül várhatóan legalább 95-ször igazunk lesz, vagyis a hibás elutasító döntések aránya várhatóan nem lesz nagyobb, mint 5%.

Más normalitásvizsgálatokkal kapcsolatban lásd a 17.1.3. alpontot.

#### 5.4. példa

A 3.8. táblázatban egy  $n = 100$  fős minta néhány antropometriai változójának alapstatisztikái láthatók nemek szerinti bontásban. E táblázatból többek között kiolvashatjuk, hogy a 10 éves kori testsúly ( $Tsúly_{10}$ ) ferdeségi és csúcsossági együtthatója a fiúk  $n = 47$  fős mintájában:

$$g_1 = 1,33 \text{ és } g_2 = 1,88,$$

a lányok  $n = 53$  fős mintájában pedig:

$$g_1 = 1,36 \text{ és } g_2 = 2,90.$$

Ezen adatok birtokában vizsgáljuk meg 99%-os megbízhatósággal, hogy normális eloszlásúnak tekinthető-e a 10 éves kori testsúly a fiúk és a lányok populációjában!

#### Megoldás

A fiúk esetében

$$\sigma_1 = \sqrt{\frac{6}{47}} = 0,128 \text{ és } \sigma_2 = \sqrt{\frac{24}{47}} = 0,511.$$

Mivel  $1-\alpha = 0.99$ , ezért  $\alpha = 0.01$  és

$$u_{\alpha/2} = u_{0.005} = Q_{0.995} = 2.576$$

(vö. Melléklet, 2. táblázat). Ekkor (5.28) és (5.29) most így néz ki:

$$P(-2,576 \cdot 0,128 \leq g_1 \leq 2,576 \cdot 0,128) = 0.99$$

és

$$P(-2,576 \cdot 0,511 \leq g_2 \leq 2,576 \cdot 0,511) = 0.99,$$

azaz (elvégezve a kijelölt műveleteket):

$$P(-0,33 \leq g_1 \leq 0,33) = 0.99$$

és

$$P(-1,32 \leq g_2 \leq 1,32) = 0.99.$$

Minthogy a  $g_1 = 1,33$  és a  $g_2 = 1,88$  értékre ezek a normalitás esetén 99%-os valószínűséggel fennálló összefüggések nem teljesülnek, ez ellentmond a normalitás feltételezésének. Így 99%-os megbízhatósággal állíthatjuk, hogy a 10 éves fiúk testmagassága nem normális eloszlású. A  $g_1$  mutató nullánál érezhetően magasabb értéke az eloszlás jobbra elnyúló aszimmetrikus voltát támasztja alá,  $g_2$  magas értéke pedig azt jelzi, hogy az eloszlás a normálisnál csúcsosabb, vagyis a széli extrém értékek ennél az eloszlásnál nagyobb arányban fordulnak elő, mint a normálisnál.

A lányok esetében

$$\sigma_1 = \sqrt{\frac{6}{53}} = 0,113 \quad \text{és} \quad \sigma_2 = \sqrt{\frac{24}{53}} = 0,453.$$

Ugyanúgy, mint a fiúk mintájában,  $\alpha = 0.01$  és  $u_{\alpha/2} = 2,576$ . Ekkor (5.28) és (5.29) most így néz ki:

$$P(-0,29 \leq g_1 \leq 0,29) = 0.99$$

és

$$P(-1,17 \leq g_2 \leq 1,17) = 0.99.$$

Mivel a lányok mintájában  $g_1 = 1,36$  és  $g_2 = 2,90$ , a 10 éves lányok esetében is kijelenthetjük 99%-os megbízhatósággal, hogy testmagasságuk nem normális eloszlású.

Bár a  $g_1$  és a  $g_2$  együttható a normalitástól való eltérés más vonásaira érzékeny, mindkét eljárás külön-külön is elégséges a normalitás elvetésére. Együttes végrehajtásuk, illetve figyelembevételük nem szükséges. Persze ha mindkettő csilingel, az nagyobb súlyt ad a normalitás elvetésének, ha pedig egyik sem, akkor ez még erősebben alátámasztja azt, hogy a vizsgált változóval végzett normalitást feltételező eljárások érvényessége nem csorbulhat számottevően.

Gyakorlásképpen készítse el az olvasó a *Tsúly10* változó hisztogramját mind a két nemre vonatkozóan. E hisztogramok alakja összhangban van a példában kapott eredményekkel. A 10 éves kori testsúly – főleg a fiúk esetében – érzékelhetően aszimmetrikus eloszlású, a magas testsúlyértékek felé hosszán elnyúló széllel.

A jelen alponban ismertetett normalitásvizsgálat érdekessége, hogy a mintabeli ferdeségi és csúcsossági együttható értéke alapján vonunk le következtetést, de nem a megfelelő elméleti paraméterre, hanem a vizsgált változó normalitására. A leírt módszer akár konfidencia-intervallum szerkesztésére is alkalmas lenne, csak éppen az a bökkenő, hogy amikor a konfidencia-intervallum érvényes lenne, akkor az egész teljesen felesleges. A konfidencia-intervallum érvényességének ugyanis az a feltétele, hogy  $X$  normális eloszlású legyen. De ha tudjuk, hogy ez a feltétel teljesül, akkor a normális eloszlás tulajdonságai miatt szükségképpen  $\gamma_1 = 0$  és  $\gamma_2 = 0$ , tehát nem kell őket a mintából megbecsülni. Ha viszont a normalitásról nem tudunk semmit, vagy történetesen azt tudjuk, hogy  $X$  nem normális eloszlású, akkor érvénytelen  $\gamma_1$ -re és  $\gamma_2$ -re a standard hibájuk segítségével megszerkesztett konfidencia-intervallum. Emiatt olyan helyzet is előállhat, hogy  $\gamma_1$  és  $\gamma_2$  közül az egyik 0, tehát  $X$  tekintetben olyan, mint egy normális eloszlású változó (például  $\gamma_1 = 0$  esetén szimmetrikus), a másik viszont jelentősen különbözik a 0-tól. Ennek következtében a normalitást feltételező standard hibák eltérhetnek – s többnyire el is térnek – a valódi értéktől. Ha emiatt a becsült standard hibák kisebbek, mint a valódi,



akkor ez könnyen vezethet olyan eredményre, hogy a 0 elméleti érték ellenére a mintabeli mutató eltérést jelez a 0-tól. Következésképpen a ferdeségi és csúcsossági mutató standard hibához viszonyított mintabeli nagyságából csak a normalitástól való eltérésre következtethetünk, elméleti értékeik 0-tól való különbözőségére nem, vagy legalább is nem a fent ismertetett módon.

E jelenség illusztrálására öt diszkrét egyenletes eloszlású változót (lehetséges értékeik a 0 és 100 közötti egész számok) és ugyancsak öt  $f = 5$  szabadságfokú  $t$ -eloszlású változót generáltunk a ROPstat segítségével. Ezután az 'Alapok' programcsoport 'Részletesebb statisztikák' rutinjával statisztikai elemzést végeztünk velük, melynek során kiszámítottuk ferdeségi és csúcsossági együtthatójukat és ezek segítségével normalitásvizsgálatot is végrehajtottunk. Az eredmények az 5.7. táblázatban láthatók. Itt a ferdeségi és a csúcsossági együtthatók melletti jelzésből olvasható ki, hogy a normalitás milyen megbízhatósággal vethető el. Egy plusz jel (+) esetén 90, egy csillag (\*) esetén 95, két csillag (\*\*) esetén 99%-os biztonsággal, három csillag (\*\*\*) esetén pedig 99,9%-os biztonsággal jelenthetjük ki, hogy az adott sorban található változó nem normális eloszlású. A 'Normalitásvizsgálat szignifikanciája' sor a táblázat alatt ezt adja tudtul, csak a döntési hiba oldaláról nézve (a 90, 95, 99 és 99,9%-os megbízhatóságnak rendre 10, 5, 1, illetve 0,1%-os hiba felel meg). A 'szignifikancia' kifejezés pontos jelentését majd a statisztikai hipotézisvizsgálat módszerének ismertetésekor adjuk meg (lásd 6.1. alfejezet).

Mind az egyenletes, mind a  $t$ -eloszlás szimmetrikus, ezért elméleti ferdeségi együtthatójuk 0. A csúcsosság tekintetében az egyenletes eloszlás laposabb, mint a normális, ezért csúcsossági együtthatója 0-nál kisebb (és pedig a jelen esetben  $-1,2$ ), a  $t$ -eloszlás viszont csúcsosabb, mint a normális, szélei hangsúlyosabbak, ezért csúcsossági együtthatója 0-nál határozottan nagyobb ( $f = 5$  esetén a  $t$ -eloszlás csúcsossága:  $\gamma_2 = 6$ ). Azt, hogy ezek a változók nem normális eloszlásúak, a program normalitásvizsgálata mind a tíz véletlen változó esetén helyesen jelzi. Ugyanakkor azt is érdemes észrevenni, hogy az öt szimmetrikus  $t(5)$  eloszlású változó közül négy esetében a ferdeség mutatója is határozottan jelzi a normalitástól való eltérést, viszont ezek nem tekinthetők olyan információknak, amelyek a  $t$ -eloszlás szimmetrikusságát cáfolnák. A táblázatból inkább azt a következtetést vonhatjuk le, hogy a  $g_1$  ferdeségi együttható standard hibája az egyenletes eloszlás esetén kisebb, a  $t(5)$  eloszlás esetén pedig nagyobb, mint a normálisé.

#### 5.7. táblázat. Öt diszkrét egyenletes és öt $t(5)$ -eloszlású véletlen változó normalitásvizsgálata egy $n = 10\,000$ elemű véletlen minta alapján

| Részletes leíró statisztikák                                                                                   |          |        |         |         |         |                   |        |          |            |
|----------------------------------------------------------------------------------------------------------------|----------|--------|---------|---------|---------|-------------------|--------|----------|------------|
| A beolvasott összes eset száma: 10000                                                                          |          |        |         |         |         |                   |        |          |            |
| Index                                                                                                          | Változó  | Esetek | Medián  | Átlag   | St.hiba | 90%-os konf. int. |        | Ferdeség | Csúcsosság |
| 2.                                                                                                             | E-1      | 10000  | 50      | 50,10   | 0,292   | 49,62             | 50,59  | -0,01    | -1,22***   |
| 3.                                                                                                             | E-2      | 10000  | 50      | 49,91   | 0,292   | 49,43             | 50,39  | 0,01     | -1,19***   |
| 4.                                                                                                             | E-3      | 10000  | 50      | 49,90   | 0,292   | 49,42             | 50,38  | 0,01     | -1,20***   |
| 5.                                                                                                             | E-4      | 10000  | 50      | 49,70   | 0,292   | 49,22             | 50,18  | 0,02     | -1,21***   |
| 6.                                                                                                             | E-5      | 10000  | 50      | 50,04   | 0,290   | 49,57             | 50,52  | -0,01    | -1,19***   |
| 7.                                                                                                             | $t(5)_1$ | 10000  | -0,0028 | 0,0094  | 0,0129  | -0,0117           | 0,0306 | 0,11***  | 4,60***    |
| 8.                                                                                                             | $t(5)_2$ | 10000  | 0,0136  | 0,0161  | 0,0129  | -0,0051           | 0,0373 | 0,13***  | 3,83***    |
| 9.                                                                                                             | $t(5)_3$ | 10000  | 0,0312  | 0,0213  | 0,0128  | 0,0002            | 0,0424 | -0,11*** | 3,78***    |
| 10.                                                                                                            | $t(5)_4$ | 10000  | 0,0066  | -0,0065 | 0,0131  | -0,0279           | 0,0150 | -0,18*** | 3,46***    |
| 11.                                                                                                            | $t(5)_5$ | 10000  | 0,0245  | 0,0194  | 0,0132  | -0,0023           | 0,0412 | -0,03    | 3,92***    |
| A normalitás tesztelése a Ferdeség és a Csúcsosság segítségével: *: $p < 0,05$ **: $p < 0,01$ ***: $p < 0,001$ |          |        |         |         |         |                   |        |          |            |

#### 5.4. Becslések a ROPstat segítségével

1. Egy kvantitatív változó két legfontosabb statisztikai jellemzője elméleti átlaga és szórása. Ha van egy véletlen mintánk a változó eloszlásából, akkor ezt a két elméleti paramétert a megfelelő mintabeli statisztikai jellemzővel, vagyis a mintaátlaggal és a mintaszórással becsülhetjük (vö. 5.2. alfejezet). A ROPstatban ezeket legegyszerűbben az 'Alapok' programcsoport 'Alapstatisztikák' rutinjában lehet kiszámítani, mely az átlag és a szórás mellett a minta variációs együtthatóját, valamint legkisebb és legnagyobb értékét is kiszámítja (lásd 3.7. táblázat).

2. Részletesebb statisztikai jellemzés végezhető el ugyanezen programcsoport 'Részletesebb statisztikák' rutinja segítségével. Itt alapértelmezésben a program kiszámítja a mediánt, az átlagot, az átlag standard hibáját, valamint a ferdeségi és a csúcsossági együtthatót, melyek a megfelelő elméleti értékek torzítatlan vagy aszimptotikusan torzítatlan pontbecslései (vö. 5.2. alfejezet). Ezenfelül a program konfidencia-intervallumot is szerkeszt az elméleti átlagra. Ennek megbízhatósági szintje alapértelmezésben 95%, de ez egyszerűen átállítható 90 vagy 99%-ra is. Példaként nézzük meg a 3.8. táblázatban látható ROPstat eredménylistát, mely 100 gyermek és szülei néhány antropometriai változójának részletes statisztikai jellemzését mutatja be nemenkénti bontásban! Figyeljük meg, hogy a program változónként normalitásvizsgálatot is végez. Például a 'T súly10' (10 éves kori testsúly) változó ferdeségi együtthatója melletti három csillag (\*\*\*) alapján 99,9%-os biztonsággal kijelenthetjük, hogy ez a változó sem a fiúk, sem a lányok esetében nem normális eloszlású. Hasonló megbízhatósággal vethetjük el a normalitást a 'Anya Súly' (anya testsúlya) változó esetében, de csak a fiúk mintájában. Egyébként, ha egy változóval kapcsolatban csak a normalitás fennállása érdekel bennünket, akkor az 'Alapok' programcsoport 'Normalitásvizsgálat' rutinját is alkalmazhatjuk.

3. Ugyanebben a 'Részletesebb statisztikák' rutinban még a 2. pontban említetttnél is részletesebb eredménylistát kapunk, ha a trimmelési százalékot 0-ról valamilyen pozitív értékre állítjuk be. Illusztrációképpen tekintsük meg a 3.9. táblázat eredménylistáját, melyen a fenti 'T súly10' változó részletes statisztikai jellemzése látható nemenkénti bontásban. Itt a fenti 2. pontban említettekén kívül megtalálhatjuk a szórást, a variációs együtthatót, a trimmelt átlagot, a winsorizált szórást, a minta minimumát és maximumát, alsó és felső kvartilisét ( $K_1$ -et és  $K_3$ -at), az interkvartilis félterjedelmet, továbbá több mutató standard hibáját és intervallumbecslését. Ismét hangsúlyozzuk, hogy a ferdeségi és a csúcsossági együttható standard hibája csak normális eloszlás esetén érvényes. Mivel az empirikus változók eloszlása igen gyakran nem normális (vö. Micceri, 1989), ezekről csak nagy óvatossággal szabad bármi következtetést levonni. Ugyanezen okból a program nem szerkeszti meg a szórásra vonatkozó konfidencia-intervallumot, mert az nem normális eloszlású  $X$  változók esetén nagyon megbízhatatlan (vö. 5.3.4. alpont).

## 6. fejezet: A statisztikai hipotézisvizsgálat

A különböző statisztikai programok (köztük a ROPstat is) nemcsak azt tüntetik fel az egyes statisztikai próbák eredménylistáján, hogy a próba milyen szinten szignifikáns vagy nem szignifikáns, hanem megadják ennek az informatív  $p$  valószínűségnek is a becsült értékét. Röviden: a próbastatisztikához tartozó  $p$ -érték az I. fajta hiba valószínűsége a  $H_0$  esetleges elutasítása esetén, s egy olyan megállapítás, hogy  $p < 0,05$  szignifikáns, éppen azt jelenti, hogy a  $H_0$ -t elutasító döntés esetén a hibás döntés valószínűsége nem haladja meg az 5%-ot.

### 7.6. A középértékekre vonatkozó hipotézisek vizsgálata a ROPstatban

A jelen fejezetben ismertetett statisztikai próbák mind elvégezhetők a ROPstat 'Középértékekre vonatkozó próbák' rutinja segítségével. A lényeges lépések a következők:

1. ROPstat indítása
2. Adatállomány behívása vagy létrehozása
3. A „Statisztikai elemzések” menüben az 'Alapok (egymintás eljárások)', majd ezen belül a 'Középértékekre vonatkozó hipotézisek vizsgálata' menüpont kiválasztása
4. Elemzendő változó (változók) kijelölése
5. Hipotetikus érték (értékek) megadása
6. Trimmelési százalék megadása (opcionális, az alapértelmezés 0)
7. Az elemzés elindítása a „Futtat” parancsikorra kattintással.

Például a 7.1. példa súlycsökkenésre vonatkozó adatainak a ROPstat segítségével elvégzett elemzésével, 10%-os trimmelés beállításával a 7.5. táblázatban látható eredményekhez jutotunk.

**7.5. táblázat.** A 7.1. példa súlycsökkenésre vonatkozó adatainak az elemzése a ROPstat 'Középértékekre vonatkozó próbák' rutinja segítségével

Változók középértékére vonatkozó hipotézisek vizsgálata

A beolvasott összes eset száma: 12

Jelölés: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$  \*\*\*:  $p < 0,001$

VÁLTOZÓ: Súlycsök (Súlycsökkenés kg-ban)

Érvényes értékek száma: 12

Átlag: 14

Szórás: 8,421

Trimmelt átlag: 13,50

Winszorizált szórás: 3,204

Medián: 13

Minimum: -2

Maximum: 35

Hipotetikus érték: 10

A normalitás tesztelése a Ferdeség és a Csúcsosság segítségével:

Ferdeség: 0,94 ( $p = 0,185$ )

Csúcsosság ( $g_4 = a_4 - 3$ ): 1,33 ( $p = 0,348$ )

A  $H_0$ : Az elméleti átlag = 10 nullhipotézis vizsgálata:

- Egymintás t-próba:  $t(11) = 1,646$  ( $p = 0,1281$ )

- Johnson-próba:  $J(11) = 1,935$  ( $p = 0,0791$ )+

- Gayen-próba szignifikanciája:  $p = 0,1377$

A  $H_0$ : Az elméleti trimmelt átlag = 10 nullhipotézis vizsgálata:

- Trimmelt t-próba (trimmelés = 8%):  $Tt(9) = 2,649$  ( $p = 0,0265$ )\*

A  $H_0$ : Az elméleti medián = 10 nullhipotézis vizsgálata:

- Wilcoxon-próba:  $R^- = 9,0$ ,  $R^+ = 46,0$  ( $p = 0,0586$ )+

A  $H_0$ :  $P(X < 10) = P(X > 10)$  nullhipotézis vizsgálata:

- Előjelpróba:  $\#(X < 10) = 1$  (8,3%),  $\#(X > 10) = 9$  (75,0%) | ( $p = 0,0215$ )\*

A kapott eredményekhez a következő megjegyzéseket fűzzük.

1. Az extrém értékek (-2 és 35) jelenléte a mintában (lásd 7.1. példa) megakadályozza, hogy az egymintás t-próba és két robusztus változata (Johnson- és Gayen-próba) 5%-os szin-

ten szignifikáns legyen. Ezt abból láthatjuk, hogy kiszámított próbastatisztikáik mellett nem található '\*' jel. A  $J = 1,935$  melletti 0,10-nél kisebb  $p$ -érték ( $p = 0,0791$ ), illetve az ezt kihangsúlyozó '+' jel adja tudtul, hogy a Johnson-próba 10%-os, tehát tendenciaszinten szignifikáns.

2. A trimmelt  $t$ -próba a szélsőséges értékek hatását kiküszöbölve 5%-os szinten jelzi, hogy a súlycsökkenés várható nagysága nagyobb 10 kg-nál ( $p = 0,0385 < 0,05$ ).

3. Az átlagra és a trimmelt átlagra vonatkozó különböző statisztikai próbák próbastatisztikája ( $t$ ,  $J$  és  $T_1$ ) után zárójelben mindig a próbastatisztika szabadságfoka látható.

4. Az előjelpróba nem veszi figyelembe adatonként a hipotetikus értéktől való eltérés nagyságát, csak az irányát. Az 5%-os szinten szignifikáns eredmény annak köszönhető, hogy a minta 12 személyéből 9 esetben volt 10 kg-nál nagyobb a súlycsökkenés, 1 esetben kisebb, s ilyen mértékű különbség  $n_+$  és  $n_-$  között a

$$H_0: P(X > 10) = P(X < 10)$$

hipotézis igaz volta esetén 5%-nál kisebb ( $p = 0,0215$ ) eséllyel fordul elő.

5. A Gayen-próba olyan speciális hipotézisvizsgálati módszer, amely nem próbastatisztika segítségével dönt a nullhipotézis elutasíthatóságáról. Ehelyett közvetlenül azt számítja ki, hogy a nullhipotézis igaz volta esetén milyen valószínűséggel lesz a mintaátlag és a hipotetikus érték eltérése a kapott mintában tapasztalttal egyenlő vagy annál szélsőségesebb, vagyis közvetlenül a szignifikancia mértékét ( $p$  érték) adja meg. Az adott példában a Gayen-próba szignifikanciájának mértéke:  $p = 0,377$ . Mivel ez nagyobb, mint 0,10, a Gayen-próba még 10%-os szinten sem szignifikáns (ugyanúgy, mint az egymintás  $t$ -próba).

6. A trimmelt egymintás  $t$ -próba sorában a ROPstat eredménylistája feltünteti a trimmelés valódi mértékét (8%). A mintából ténylegesen elhagyott adatok aránya alulról és felülről egyaránt 8% annak ellenére, hogy mi 15%-ot adtunk meg. Ennek az az oka, hogy  $n = 12$  esetén 1 adat éppen 8%-nak felel meg, 2 adat pedig 16%-nak. Ha tehát maximum 15%-ot vághatunk le a minta alsó és felső széléről, akkor 1-nél több adatot egyik oldalról sem hagyhatunk el (a levágásnak mindkét oldalon ugyanolyan mértékűnek kell lennie).

## 8.6. Két változó ugyanakkorosságának vizsgálata a ROPstat segítségével

A jelen fejezetben ismertetett statisztikai próbák mind elvégezhetők a ROPstatban az 'Összetartozó minták (változók) egyszempontos összehasonlítása' rutin segítségével. A lényeges lépések a következők:

1. ROPstat indítása
2. Adatállomány behívása vagy létrehozása
3. A „Statisztikai elemzések” menüben a 'Csoportok és változók összehasonlítása', majd ezen belül az 'Összetartozó minták (változók) egyszempontos összehasonlítása' menüpont kiválasztása
4. Elemzendő változók (egy-egy  $X$  és  $Y$  változó) kijelölése
5. Trimmelési százalék megadása (opcionális, az alapértelmezés 0)
6. Az elemzés elindítása a „Futtat” parancsikorra kattintással.

### 8.5. példa

A 8.2. példa ingerexpozíciós adatainak ROPstat segítségével végrehajtott elemzésével a 8.5. táblázatban bemutatott eredményekhez jutunk.

A 8.5. táblázat ROPstat eredménylistájához a következő megjegyzéseket fűzzük.

1. A  $V = Y - X$  különbségváltozó mintabeli ferdeségi együtthatója ( $g_1 = 1,463$ ) alapján  $V$  normalitása ebben az esetben megkérdőjelezhető. A ferdeség magas volta extrém értékek jelenlétét valószínűsíti. Ilyen esetben a  $t$ -próba ereje gyakran lecsökken. Feltehetőleg ez okozza a  $t$ -próba és robusztus alternatíváinak a szignifikanciát meg sem közelítő eredményét.

2. Az átlagra és trimmelt átlagra vonatkozó különböző statisztikai próbák próbastatisztikája ( $t$ ,  $J$  és  $T_t$ ) után zárójelben mindig annak szabadságfoka látható.

3. A trimmelt  $t$ -próba és a Wilcoxon-próba már tendencia szinten jelez, de utóbbit a  $V$  változó erős ferdesége miatt nem értelmezzük. Az előjelpróba 5%-os szinten szignifikáns eredményét úgy értelmezhetjük, hogy bár a két inger felismeréséhez szükséges ingerexpozíciók átlagos száma nem különbözik jelentősen egymástól (ez feltehetően a nagy egyéni variabilitás eredménye), de a 2. ingert a személyek közül szignifikánsan többen észlelik hamarabb, mint az 1. ingert. Az előbbi a  $P(Y < X)$  valószínűség becslése alapján kb. 78%-os, az utóbbi pedig a  $P(Y > X)$  valószínűség becslése alapján kb. 17%-os gyakorisággal következik be.

4. A 2. inger felismerési fölényét az 1. ingerrel szemben az  $A(Y, X)$  valószínűségi fölény mutatója jelzi, melynek pontbecslése esetünkben:

$$\hat{A}(Y, X) = 0,194.$$

Ez az érték azért kisebb 0.5-nél, mert a felismerés könnyűsége és a felismeréshez szükséges expozíciók száma ellentétes pólusú skála.  $X$  és  $Y$  szerepének felcserélésével egyébként a következő pontbecslést kapjuk a valószínűségi fölény mutatójára (vö. 8.14), mely irányában is helyesen jelzi a 2. ingerábra felismerési könnyűségét az 1. ingerábrával szemben:

$$\hat{A}(X, Y) = 1 - 0.194 = 0.806.$$

5. Az előjelpróba sorában  $\#(X < Y)$  és  $\#(X > Y)$  az  $n_-$  és az  $n_+$  mennyiséget jelöli. Ezek összege adja ki  $n$ -et, a csökkentett elemszámot, mely a felette lévő sorban található Wilcoxon-próba szignifikanciájának kiértékeléséhez is szükséges.

**8.5. táblázat.** A 8.2. példa ingerexpozíciós adatainak az elemzése a ROPstat 'Összetartozó minták (változók) egyszempontos összehasonlítása' rutinja segítségével

Összetartozó minták egyszempontos összehasonlítása

A beolvasott összes eset száma: 18

Érvényes esetek száma: 18

Jelölés: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$  \*\*\*:  $p < 0,001$

Elméleti átlagok egyenlőségének tesztelése:

- Egymintás  $t$ -próba:  $t(17) = -1,477$  ( $p = 0,1580$ )
- Johnson-próba:  $J(17) = -1,169$  ( $p = 0,2586$ )
- Gayen-próba szignifikanciája:  $p = 0,1718$

Elméleti trimmelt átlagok egyenlőségének tesztelése (trimmelés: 18%):

- Összetartozó mintás trimmelt  $t$ -próba:  $Tt(11) = -1,945$  ( $p = 0,1677$ )

A "Nincs  $X$ -ről  $Y$ -ra változás" hipotézis vizsgálata:

- Wilcoxon-próba:  $R+ = 37,0$ ,  $R- = 116,0$  ( $p = 0,0638$ )+

"A két változó sztochasztikusan ugyanakkora" hipotézis vizsgálata:

- Előjelpróba:  $\#(X < Y) = 3$ ,  $\#(X > Y) = 14$  | ( $p = 0,0127$ )\*

| Index  | Változó | Átlag  | Szórás | Trimmett<br>átlag | Winsor.<br>szórás | Rang-<br>átlag |
|--------|---------|--------|--------|-------------------|-------------------|----------------|
| X:     | Inger1  | 69,28  | 34,43  | 67,83             | 22,09             | 1,81           |
| Y:     | Inger2  | 52,89  | 28,96  | 48,17             | 17,63             | 1,19           |
| Y - X: |         | -16,39 | 47,08  | -19,67            | 28,18             |                |

Ferdeség- és csúcsosságérték és normalitásvizsgálat

Az Y - X változó mintabeli ferdesége = 1,463 ( $p = 0,011$ )\*

Az Y - X változó mintabeli csúcsossága ( $g4 = a4 - 3$ ) = 0,936 ( $p = 0,417$ )

95%-os konfidencia-intervallum az Y-X különbség elméleti átlagára  $C(0,95) = (-39,80, 7,02)$

Pontbecslés a valószínűségi fölény  $A(Y, X)$  mutatójára:

$A(Y, X)^{\wedge} = 0,194$  ( $(X < Y)\% = 16,7$  ( $Y = X)\% = 5,6$  ( $X > Y)\% = 77,8$ )

95%-os konfidencia-intervallum  $A(Y, X)$ -re:  $C(0,95) = (0,020, 0,369)$

## 8.6. példa

Másik példaként egy 1996-ban lefolytatott vajdasági szociolingvisztikai vizsgálat (Göncz, 1999, 77. o.) adatainak ROPstat segítségével elvégzett elemzését mutatjuk be. Ebben a vizsgálatban 144 véletlenszerűen kiválasztott, Vajdaságban élő magyar nemzetiségű felnőtt személyt többek között arról is megkérdeztek, hogy mennyire kötődik a mai Jugoszláviához, illetve Magyarországhoz. A kötődés mértékét egy ötfokú skálán lehetett megadni, melyen az 1-es érték azt jelentette, hogy egyáltalán nem kötődik, az 5-ös pedig azt, hogy igen erősen kötődik. Ezekre a kérdésekre a 144 közül 99 személy adott érvényes választ. Az elemzések eredményeit a 8.6. táblázat tartalmazza.

**8.6. táblázat.** A mai Jugoszláviához (MaiJugó) és Magyarországhoz (MagyarO.) való kötődés összehasonlítása ROPstat segítségével egy 1996-ban lefolytatott vajdasági szociolingvisztikai vizsgálat adatai alapján. A kötődést ötfokú skálán mérték (1 = egyáltalán nem kötődik, 5 = nagyon erősen kötődik)

Összetartozó minták egyszempontos összehasonlítása

A beolvasott összes eset száma: 144

Érvényes esetek száma: 99

Jelölés: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$  \*\*\*:  $p < 0,001$

Elméleti átlagok egyenlőségének tesztelése:

- Egymintás t-próba:  $t(98) = 3,104$  ( $p = 0,0025$ )\*\*

- Johnson-próba:  $J(98) = 3,114$  ( $p = 0,0024$ )\*\*

- Gayen-próba szignifikanciája:  $p = 0,0025$ \*\*

Elméleti trimmett átlagok egyenlőségének tesztelése (valódi trimmelés: 14%):

- Összetartozó mintás trimmett t-próba:  $Tt(70) = 3,370$  ( $p = 0,0028$ )\*\*

A "Nincs X-ről Y-ra változás" hipotézis vizsgálata:

- Wilcoxon-próba:  $R+ = 1456,5$ ,  $R- = 559,5$ ,  $z = 3,125$  ( $p = 0,0018$ )\*\*

"A két változó sztochasztikusan ugyanakkora" hipotézis vizsgálata:

- Előjelpróba:  $\#(X < Y) = 41$ ,  $\#(X > Y) = 22$ ,  $z = 2,268$  ( $p = 0,0233$ )\*

| Index  | Változó  | Átlag | Szórás | Trimmett<br>átlag | Winsor.<br>szórás | Rang-<br>átlag |
|--------|----------|-------|--------|-------------------|-------------------|----------------|
| X:     | MaiJugó  | 2,222 | 1,217  | 2,028             | 0,880             | 1,40           |
| Y:     | MagyarO. | 2,707 | 1,247  | 2,662             | 1,104             | 1,60           |
| Y - X: |          | 0,485 | 1,554  | 0,634             | 1,340             |                |

Ferdeség- és csúcsosságérték és normalitásvizsgálat

Az Y - X változó mintabeli ferdesége = 0,029 ( $p = 0,907$ )

Az Y - X változó mintabeli csúcsossága ( $g4 = a4 - 3$ ) = -0,077 ( $p = 0,876$ )

95%-os konfidencia-intervallum az  $Y-X$  különbség elméleti átlagára  $C(0,95) = (0,174; 0,795)$

Pontbecslés a valószínűségi fölény  $A(Y, X)$  mutatójára:

$$\hat{A}(Y, X) = 0,596 \quad ((X < Y)\% = 41,4 \quad (Y = X)\% = 36,4 \quad (X > Y)\% = 22,2)$$

95%-os konfidencia-intervallum  $A(Y, X)$ -re:  $C(0,95) = (0,520; 0,672)$

A 8.6. táblázat ROPstat eredménylistájához a következő megjegyzéseket fűzzük.

1. Minden próba egybehangzóan azt támasztja alá, hogy a vajdasági magyarok általában jobban kötődnek Magyarországhoz, mint a mai Jugoszláviához. A  $V = Y - X$  különbségváltozó mintabeli ferdeségi együtthatója (0.029) igen közel van a 0-hoz, így  $V$  eloszlásának szimmetrikus volta ellen nem szól különösebb érv. Emiatt a Wilcoxon-próba érvényességét nem kell megkérdőjeleznünk (a folytonosság sérülése legfeljebb a próba erejét csökkenti valamelyest).

2. A szimmetriából az is következik, hogy az elméleti átlag, az elméleti trimmelt átlag és az elméleti medián megegyezik, s ennek alapján az sem meglepő, hogy a rájuk vonatkozó próbák egyöntetűen ugyanazon az 1%-os szinten szignifikánsak. Ezek alapján 99%-os biztonsággal kijelenthetjük, hogy a Magyarországhoz való kötődés szintje magasabb, mint a mai Jugoszláviához való kötődése.

3. A kötődési szintkülönbség valószínű mértékét az  $Y - X$  különbségre vonatkozó konfidencia-intervallum jelzi, mely 95%-os megbízhatósággal 0.17 és 0.80 közötti érték.

4. A Wilcoxon- és az előjelpróba sorában található  $z$ -érték azt mutatja, hogy a nagy elemszám miatt e próbák próbastatisztikája a normális eloszlással való közelítés alapján lett kiértékelve, így ezek szignifikanciájának megállapításához a standard normális eloszlás táblázata használható.

5. Az előjelpróba ekkora elemszám ( $n = 99$ ) melletti csupán 5%-os szinten szignifikáns eredménye abból fakad, hogy az egyértelmű kötődési szintkülönbség ellenére nem elhanyagolható azoknak a személyeknek a száma, akik nem Magyarországhoz, hanem a jelenlegi Jugoszláviához kötődnek jobban. A táblázatból kiolvasható, hogy az adott két kérdésre érvényes választ adó 99 személy 41.4%-a Magyarországhoz, 22.2%-a pedig a mai Jugoszláviához kötődik jobban (a maradék 36.4% egyforma mértékben kötődik a két országhoz). A mai Jugoszláviával szemben inkább Magyarországhoz való kötődés valószínűségi fölényének pontbecslése:

$$\hat{A}(Y, X) = 0.596$$

és intervallumbecslése:

$$C_{0.95}(A_{YX}) \approx (0.520; 0.672).$$

Ez utóbbi becslés, mely elég mérsékelten emelkedik a sztochasztikus egyenlőség 0.50-es szintje fölé, jól mutatja, hogy a szintátlagok erősen szignifikáns különbségét nem szabad úgy értelmeznünk, hogy az adott irányú és mértékű különbség jellemzi az egész populációt. A  $t$ -próba és alternatívái szignifikanciájának erőssége ( $\alpha = 1\%$ ) csupán arra vonatkozik, hogy igen biztosak lehetünk abban, hogy az elméleti átlagok egyenlőségére vonatkozó nullhipotézis nem igaz, vagyis a jelen esetben abban, hogy a Magyarországhoz és a mai Jugoszláviához való kötődés szintje nem ugyanakkora. Valamivel erősebb a kötődés Magyarországhoz, de ez nem mindenkire igaz. Annnyira, hogy az adott mintában a Magyarországhoz jobban kötődő személyek aránya (41.4%) még az 50%-ot sem éri el.

## 9.8. Két független minta összehasonlítása paraméteres eljárásokkal a ROPstatban

A jelen fejezetben ismertetett statisztikai próbák mind elvégezhetők a ROPstatban a 'Független minták (csoportok) egyszempontos összehasonlítása' rutin segítségével. A lényeges lépések a következők:

1. ROPstat indítása
2. Adatállomány behívása vagy létrehozása
3. A „Statisztikai elemzések” menüben a 'Csoportok és változók összehasonlítása', majd ezen belül a 'Független minták (csoportok) egyszempontos összehasonlítása' menüpont kiválasztása
4. Csoportosító változó kijelölése
5. Az elemzendő változó (változók) kijelölése
6. Trimmelési százalék megadása (opcionális, az alapértelmezés 0)
7. Az elemzés elindítása a „Futtat” parancsikonra kattintással.

Például a 9.4. példa adatainak ROPstat segítségével elvégzett elemzésével a 9.7. táblázatban látható eredményekhez jutunk. A táblázatban 'ElítSzáma' jelöli az 'Elítélések száma', 'BcsTípus' pedig a 'Bűncselekmény típusa' változót. Ez utóbbi szerint csoportosítjuk a személyeket.

A 9.7. táblázat eredménylistájához a következő megjegyzéseket fűzzük.

1. A különböző statisztikai próbák próbastatisztikája (F, t, d stb.) után zárójelben mindig annak szabadságfoka (szabadságfokai) látható (láthatók).

2. Az 'Elítélések száma' változó varianciája a Levene- és az O'Brien-próba alapján egyöntetűen szignifikánsan különbözik a két bűnözői csoportban. A mintaszórások alapján azt láthatjuk, hogy a jogerős elítélések száma tekintetében az erőszakos bűnözők populációja homogénebb. A szórások közti különbség azonban nem olyan mértékű, ami számottevően rontaná a kétmintás t-próba érvényességét, tekintetbe véve, hogy a két mintaszórás hányadosa:  $2,094/1,502 = 1,39$  és a két minta elemszáma (47 és 53) is közel ugyanakkora (a nagyobbik csak kb. 13%-kal nagyobb a kisebbiknél).

**9.7. táblázat.** A 9.4. példa adatainak az elemzése a ROPstat 'Független minták (csoportok) egyszempontos összehasonlítása' rutinja segítségével

|                                                                             |          |        |       |        |      |      |          |            |
|-----------------------------------------------------------------------------|----------|--------|-------|--------|------|------|----------|------------|
| Független minták egyszempontos összehasonlítása                             |          |        |       |        |      |      |          |            |
| A beolvasott összes eset száma: 100                                         |          |        |       |        |      |      |          |            |
| Csoportosító változó: BcsTípus                                              |          |        |       |        |      |      |          |            |
| Jelölés: +: $p < 0,10$ *: $p < 0,05$ **: $p < 0,01$ ***: $p < 0,001$        |          |        |       |        |      |      |          |            |
| FÜGGŐ VÁLTOZÓ: ElítSzáma                                                    |          |        |       |        |      |      |          |            |
| Csoportonkénti alapstatisztikák                                             |          |        |       |        |      |      |          |            |
| Index                                                                       | BcsTípus | Esetek | Átlag | Szórás | Min. | Max. | Ferdeség | Csúcsosság |
| 1                                                                           | Erőszak  | 47     | 4,489 | 1,502  | 2    | 9    | 0,665+   | 0,753      |
| 2                                                                           | NemErő   | 53     | 4,189 | 2,094  | 2    | 9    | 0,823*   | -0,243     |
| A Ferdeség vagy a Csúcsosság szignifikanciája a normalitás sérülését jelzi. |          |        |       |        |      |      |          |            |
| Elméleti szórások egyenlőségének tesztelése                                 |          |        |       |        |      |      |          |            |
| - O'Brien-próba (Welch-féle): $F(1,0; 89,4) = 4,950$ ( $p = 0,0286$ )*      |          |        |       |        |      |      |          |            |
| - Levene-próba (Welch-féle): $F(1; 96,2) = 6,946$ ( $p = 0,0098$ )**        |          |        |       |        |      |      |          |            |



#### Elméleti átlagok egyenlőségének tesztelése

Hagyományos eljárás, amely feltételezi a szóráshomogenitást:

- Kétmintás t-próba :  $t(98) = 0,816$  ( $p = 0,4168$ )

Robusztus eljárás, amelynél nem szükséges a szóráshomogenitás:

- Welch-féle d-próba:  $d(94,0) = 0,832$  ( $p = 0,4078$ )

95%-os konfidencia-intervallum a két elméleti átlag m1-m2 különbségére

- a kétmintás t-próba alapján:  $C(0,95) = (-0,432, 1,033)$

- a Welch-féle d-próba alapján:  $C(0,95) = (-0,418; 1,020)$

A trimmelt átlagra vonatkozó elemzések (max. trimmelés = 10%)

| Index | BcsTípus | n  | n_trim | Trim. átlag | Winsor. szórás | Trim. min. | Trim. max. |
|-------|----------|----|--------|-------------|----------------|------------|------------|
| 1     | Erőszak  | 47 | 39     | 4,410       | 1,137          | 3          | 6          |
| 2     | NemErő   | 53 | 43     | 3,930       | 1,797          | 2          | 7          |

Elméleti trimmelt átlagok egyenlőségének tesztelése

- Yuen-próba:  $Y(71,4) = 1,316$  ( $p = 0,1925$ )

95%-os konfidencia-intervallum az elméleti trimmelt átlagok

Tm1-Tm2 különbségére:  $C(0,95) = (-0,248; 1,208)$

3. A két minta nagyság szintje sem az átlagok, sem a trimmelt átlagok tekintetében nem különbözik szignifikánsan egymástól, ugyanúgy, mint a 9.4. példában, ahol a medián-próba segítségével hasonlítottuk össze őket egymással.

4.

Szakmailag tehát nem tudunk igazolt különbséget megállapítani az erőszakos és a nem erőszakos bűnözők populációja között az '*Elítélések száma*' változó *nagysága* tekintetében, mindamellett e változó szerint az erőszakos bűnözők populációja valamelyest homogénebbnek tűnik.

A medián-próbát a ROPstatban ugyanezen rutin segítségével lehet végrehajtani, de ehhez mindenekelőtt ki kell jelölni a közös mediánnál kisebb és nagyobb értékek övezetét (a „Változók deklarációi táblában” vagy a szóban forgó statisztikai menüablakban a változó kiválasztása előtt az egér jobb gombjával a változó nevére kattintva). Maga a medián egyébként a ROPstatban a 'Részletesebb statisztikák' vagy a 'Gyakoriság, hisztogram' rutin segítségével határozható meg. Esetünkben az '*Elítélések száma*' változó esetében a kijelölendő két övezet '2–3', illetve '5–9'. Ezután az elemzendő változó típusát (a 'Skálátípus' ablakban egérrel a változóval azonos sorban) át kell állítani intervallumról nominális skálájúra, s indíthatjuk is az elemzést. Ezzel a beállítással a 9.8. táblázatban látható eredménylistához jutunk. Itt a medián-próba eredménye a khi-négyzet-próba sorából olvasható ki, amely természetesen megegyezik a 9.4. példa megoldásában kapottal. Az eredménylista néhány további információt is tartalmaz (kapcsolati mutatók, illetve a khi-négyzet-próba egy alternatívája, a Fisher-egzakt-próba), amelyekkel a 17. fejezetben foglalkozunk majd részletesebben.

#### 9.8. táblázat. A medián-próba végrehajtása ROPstatban a 9.4. példa adatai esetében

Független minták egyszempontos összehasonlítása

A beolvasott összes eset száma: 100

Csoportosító változó: BcsTípus

Jelölés: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$  \*\*\*:  $p < 0,001$

FÜGGŐ VÁLTOZÓ: ElítSzá

KÉTSZEMPONTOS GYAKORISÁGI TÁBLÁZAT - Változók: BcsTípus és ElítSzá

|          |                                            |     |          |
|----------|--------------------------------------------|-----|----------|
|          | Az ElítSzám változó csoportjai/kategóriái: |     |          |
| BcsTípus | 2-3                                        | 5-9 | Összesen |
| Erőszak  | 13                                         | 22  | 35       |
| NemErő   | 26                                         | 21  | 47       |
| Összesen | 39                                         | 43  | 82       |

Sorösszegek szerinti százalékok táblázata

|          |                                            |      |          |
|----------|--------------------------------------------|------|----------|
|          | Az ElítSzám változó csoportjai/kategóriái: |      |          |
| BcsTípus | 2-3                                        | 5-9  | Összesen |
| Erőszak  | 37,1                                       | 62,9 | 100,0    |
| NemErő   | 55,3                                       | 44,7 | 100,0    |
| Összesen | 47,6                                       | 52,4 | 100,0    |

KAPCSOLATI MUTATÓK  
fi kontingencia-együttható ( = Cramér-féle V) = -0,180  
Yule-féle Q asszociációs együttható: Q = -0,354  
Esélyhányados (odds ratio): alfa = 0,477

FÜGGETLENSÉG TESZTELÉSE  
Khi-négyzet-próba (f = 1): Khi2 = 2,657 (p = 0,1031)  
Fisher-egzaktp-próba:  
- egyoldalú: p = 0,0795+  
- kétoldalú: p = 0,1217

## 10.5. Két független minta összehasonlítása rangsorolós eljárásokkal a ROPstatban

A jelen fejezetben ismertetett statisztikai próbák mind elvégezhetők a ROPstatban a 'Független minták (csoportok) egyszempontos összehasonlítása' rutin segítségével. A lényeges lépések ugyanazok, mint az átlagok összehasonlítása esetén (lásd 9.8. alfejezet), egyetlen különbséggel. Ez esetben az elemzésre kijelölt változók skálátípusát ordinálisra kell beállítani (a „skálátípus” ablakban az egér jobb billentyűje segítségével). Például a 10.2. példa adatai esetében, ahol az ausztriai magyar kisebbség 18–32 és 33–53 éves korosztályát hasonlítottuk össze a magyar nemzethez való kötődés szempontjából, a 10.6. táblázatban látható eredménylistát kaptuk.

**10.6. táblázat.** A sztochasztikus egyenlőség vizsgálata ROPstat segítségével a 10.2. példa adatai esetében

|                                                                                        |         |        |                |                 |         |         |
|----------------------------------------------------------------------------------------|---------|--------|----------------|-----------------|---------|---------|
| Független minták egyszempontos összehasonlítása                                        |         |        |                |                 |         |         |
| A beolvasott összes eset száma: 35                                                     |         |        |                |                 |         |         |
| Csoportosító változó: Csoport                                                          |         |        |                |                 |         |         |
| Jelölés: +: p < 0,10 *: p < 0,05 **: p < 0,01 ***: p < 0,001                           |         |        |                |                 |         |         |
| FÜGGŐ VÁLTOZÓ: MNemzet                                                                 |         |        |                |                 |         |         |
| Csoportonkénti alapstatisztikák                                                        |         |        |                |                 |         |         |
| Index                                                                                  | Csoport | Esetek | Rang-<br>átlag | Rang-<br>szórás | Minimum | Maximum |
| 1.                                                                                     | 18-32   | 15     | 13,77          | 9,194           | 1       | 5       |
| 2.                                                                                     | 33-53   | 20     | 21,18          | 7,884           | 3       | 5       |
| Elméleti rangszórások egyenlőségének tesztelése                                        |         |        |                |                 |         |         |
| - Levene-próba (Welch-féle): F(1; 29,8) = 0,826 (p = 0,3709)                           |         |        |                |                 |         |         |
| Sztochasztikus egyenlőség tesztelése                                                   |         |        |                |                 |         |         |
| Hagyományos eljárás, amely feltételezi a szóráshomogenitást:                           |         |        |                |                 |         |         |
| - Mann-Whitney-próba (normális közelítés): Z = -2,375 (p = 0,018)*                     |         |        |                |                 |         |         |
| Szóráshomogenitást nem igénylő robusztus közelítő eljárások:                           |         |        |                |                 |         |         |
| - Fligner-Policello-próba Welch-féle szabadságfokkal: FPW(23,3) = -2,501 (p = 0,0198)* |         |        |                |                 |         |         |

- Brunner-Munzel-próba:  $BM(28,5) = -2,549$  ( $p = 0,0164$ )\*

Pont- és intervallumbecslés a valószínűségi fölény A mutatójára:

$A^{\wedge} = 0,288$  [  $P(Csop1 > Csop2) = 0,123$ ,  $P(Csop1 < Csop2) = 0,547$  ]

$C(0,95) = (0,118; 0,458)$

A 10.6. táblázat eredménylistájához a következő megjegyzéseket fűzzük.

1. Az elméleti rangszórások próbái (Levene-próba), továbbá az egymáshoz nagyon hasonló szintű két rangszórás (9,194 és 7,884) alapján nincs jele annak, hogy a szóráshomogenitás feltevése sérülne.

2. Ekkor a sztochasztikus egyenlőség vizsgálatára a robusztus eljárások (az FPW- és a Brunner–Munzel-próba) mellett a hagyományos Mann–Whitney-próba szintén elfogadható érvényességgel alkalmazható. Ez utóbbi próba sorában most nem rangszámösszegeket, hanem egy Z-értéket találunk ( $Z = 2,375$ ). Ez amiatt van, mert ekkora elemszámok (15 és 20) esetén a ROPstat a Mann–Whitney-próbát a normális eloszlással való közelítéssel értékeli ki (lásd a 10.1. alfejezet végén a Mann–Whitney-próbához fűzött 2. megjegyzést). Mindegyik próba egyöntetűen szignifikáns, ami jelzi a fiatalabb korosztály némileg gyengébb kötődését a magyar nemzethez.

3. A valószínűségi fölény mutatójának pontbecslése:  $\hat{A} = 0.288$ , továbbá a Brunner–Munzel-próba alapján készített 95%-os konfidencia-intervallum:  $C_{0,95} = (0,118; 0,458)$ . Ez azt jelenti, hogy az  $A_{12}$  mutató elméleti értéke (a fiatalok sztochasztikus dominanciája az idősebbekkel szemben) nagy valószínűséggel kb. 0,20 és 0,45 közé esik, az  $A_{21} = 1 - A_{12}$  mutatóé (az idősebbek sztochasztikus dominanciája a fiatalokkal szemben) pedig 0,55 és 0,80 közé, mely a 10.2. táblázat szerint legalább közepes mértékű attitűdbeli eltérést valószínűsít a két korosztály között.

#### 11.4.4. A korrelációs együttható intervallumbecslése

A Z-transzformáció erejének illusztrálása céljából a ROPstat 'Új random változó' rutinja lehetővé teszi, hogy olyan véletlen változót generáljunk, amelynek az eloszlása megegyezik a tapasztalati korrelációs együtthatóéval, vele együtt létrehozva annak Z-transzformáltját is.

A generálás ugyanúgy végezhető el, mint a 4. fejezetben ismertetett eloszlásoké (normális, binomiális, khi-négyzet stb.), csak most az eloszlás típusaként az 'r és Z(r)' menüsort kell kiválasztani, majd megadni a tapasztalati korrelációs együttható eloszlását meghatározó két fő paramétert, az n mintaelemszámot és a  $\rho$  elméleti korrelációt. A 11.6. táblázatban egy ilyen módon létrehozott változópár hisztogramját mutatjuk be. A generálás során 10 000 véletlen esetet hoztunk létre, vagyis egy 10 000-elemű véletlen kétdimenziós (X, Y)-mintát. A korrelációs együttható paramétereiként  $n = 10$ -et és  $\rho = 0.8$ -at adtunk meg. A két változó hisztogramját a ROPstat 'Gyakoriság, hisztogram' rutinjával készítettük el. A 11.6. táblázatban látható eredménylista meggyőzően igazolja, hogy a Z-transzformáció még az erősen ferde r-eloszlást is képes tetszetős normális alakúra transzformálni.

**11.6. táblázat.** Egy  $\rho = 0,8$  elméleti átlagú  $r$  tapasztalati korrelációs együtthatónak, valamint Z-transzformáltjának ROPstat segítségével elkészített hisztogramja

|                                                        |       |      |       |       |
|--------------------------------------------------------|-------|------|-------|-------|
| Gyakoriság, hisztogram                                 |       |      |       |       |
| A beolvasott összes eset száma: 10000                  |       |      |       |       |
| VÁLTOZÓ: RCorr1 (Random korreláció, n = 10, rho = 0,8) |       |      |       |       |
| Oszt.köz.                                              | Gyak  | %    | Kum%  |       |
| -0,343                                                 | 2     | 0,0  | 0,0   |       |
| -0,251                                                 | 1     | 0,0  | 0,0   |       |
| -0,159                                                 | 4     | 0,0  | 0,1   |       |
| -0,067                                                 | 8     | 0,1  | 0,2   |       |
| 0,024                                                  | 15    | 0,2  | 0,3   |       |
| 0,116                                                  | 25    | 0,3  | 0,6   |       |
| 0,208                                                  | 35    | 0,4  | 0,9   |       |
| 0,300                                                  | 72    | 0,7  | 1,6   |       |
| 0,392                                                  | 151   | 1,5  | 3,1   | =     |
| 0,484                                                  | 256   | 2,6  | 5,7   | ==    |
| 0,576                                                  | 582   | 5,8  | 11,5  | ===== |
| 0,668                                                  | 1192  | 11,9 | 23,4  | ===== |
| 0,760                                                  | 2265  | 22,7 | 46,1  | ===== |
| 0,852                                                  | 3457  | 34,6 | 80,7  | ===== |
| 0,944                                                  | 1935  | 19,4 | 100,0 | ===== |
| Össz.:                                                 | 10000 |      |       |       |
| VÁLTOZÓ: RanZ(r)                                       |       |      |       |       |
| Oszt.köz.                                              | Gyak  | %    | Kum%  |       |
| -0,31                                                  | 3     | 0,0  | 0,0   |       |
| -0,11                                                  | 14    | 0,1  | 0,2   |       |
| 0,10                                                   | 46    | 0,5  | 0,6   |       |
| 0,30                                                   | 141   | 1,4  | 2,0   | ==    |
| 0,51                                                   | 434   | 4,3  | 6,4   | ===== |
| 0,71                                                   | 1048  | 10,5 | 16,9  | ===== |
| 0,91                                                   | 1853  | 18,5 | 35,4  | ===== |
| 1,12                                                   | 2249  | 22,5 | 57,9  | ===== |
| 1,32                                                   | 2020  | 20,2 | 78,1  | ===== |
| 1,53                                                   | 1247  | 12,5 | 90,6  | ===== |
| 1,73                                                   | 600   | 6,0  | 96,6  | ===== |
| 1,93                                                   | 256   | 2,6  | 99,1  | ===== |
| 2,14                                                   | 68    | 0,7  | 99,8  |       |
| 2,34                                                   | 18    | 0,2  | 100,0 |       |
| 2,55                                                   | 3     | 0,0  | 100,0 |       |
| Össz.:                                                 | 10000 |      |       |       |

## 11.8. Lineáris korreláció és regresszió a ROPstatban

### 11.8.1. A ROPstat 'Korreláció, egyszerű regresszió rutinja

A fejezetünkben leírt korrelációs és regressziós eljárások nagy többségét a ROPstat 'Változók kapcsolatának vizsgálata' programcsoportjának 'Korreláció, egyszerű regresszió' rutinja segítségével lehet végrehajtani. Ennek szemléltetésére most nem valódi kutatási adatokat használunk fel, hanem a ROPstat által generált véletlen adatokat, hogy a rutin segítségével egyben a kétdimenziós normális eloszlást is bemutathassuk. E célból először is létrehoztunk egy kétdimenziós normális eloszlású (X, Y) változópárt 100 esettel a ROPstat 'Új random változó' rutinjában. A generálás ugyanúgy végezhető el, mint a 11.4.4. alpontban ismertetett (r, Z(r)) változópár esetében, csak most a 'Kétdimenziós normális' eloszlástípust kell kiválasztani. Ekkor beírható a két új változó neve (pl. RNormX és RNormY), továbbá a két változó elméleti átlaga és szórása, illetve a köztük lévő  $\rho$  elméleti korreláció. Ezzel a generálással 100 esetet hoz-

tunk létre. Az átlagok és szórások esetében adjunk meg  $\mu_X = 50$  és  $\sigma_X = 10$ , illetve  $\mu_Y = 50$  és  $\sigma_Y = 10$  értékeket, s az elméleti korrelációra a  $\rho = -0,70$  értéket írjuk be.

Miután létrehoztuk ennek a kétdimenziós normális eloszlású (X, Y) változópárnak egy n = 100 elemű véletlen mintáját, a ROPstat statisztikai menüjében kiválasztjuk a 'Változók kapcsolatának vizsgálata' programcsoportot, majd ezen belül a 'Korreláció, egyszerű regresszió' menüpontot. E rutin menüablakában tegyük egérrel az RNormX változót az „X változók”, a RNormY változót pedig „Y változók” ablakába. Eredményül a 11.7. táblázatban látható eredménylistát kaptuk, amelyhez a következő megjegyzéseket fűzzük.

1. Először is megjegyezzük, hogy az Y változó X-re vonatkozó *elméleti* lineáris regressziós egyenesének  $\beta_1$  együtthatójára  $\sigma_X \beta_1 = \rho \sigma_Y$ , amiből egyszerű átrendezéssel és  $\sigma_1 = \sigma_2 = 10$ , valamint  $\rho = -0.70$  figyelembevételével adódik, hogy

$$\beta_1 = \rho \cdot \sigma_1 / \sigma_2 = \rho = -0.70.$$

A  $\beta_0$  együtthatót a (11.20) formulába behelyettesítve kapjuk:

$$\beta_0 = \mu_Y - \beta_1 \mu_X = 50 - (-0.70) \cdot 50 = 50 + 0.70 \cdot 50 = 50 + 35 = 85.$$

Így tehát az Y változó X-re vonatkozó elméleti lineáris regressziós függvénye:

$$\hat{Y} = 85 - 0.70 \cdot X.$$

Mivel esetünkben  $\mu_1 = \mu_2 = 50$  és  $\sigma_1 = \sigma_2 = 10$ , ugyanígy néz ki az X változó Y-ra vonatkozó elméleti lineáris regressziós függvénye is:

$$\hat{X} = 85 - 0.70 \cdot Y.$$

**11.7. táblázat.** Egy normális eloszlású (X, Y) változópár ( $\mu_X = 50$ ,  $\mu_Y = 50$ ,  $\sigma_X = 10$ ,  $\sigma_Y = 10$ ,  $\rho = -0,70$ ) kapcsolatának részletes vizsgálata a ROPstat 'Korreláció, regresszió, kétváltozós diagram' rutinja segítségével egy 100-elemű véletlen (X, Y)-minta alapján

|                                                                              |           |              |                        |       |       |                      |
|------------------------------------------------------------------------------|-----------|--------------|------------------------|-------|-------|----------------------|
| Korreláció, lineáris regresszió                                              |           |              |                        |       |       |                      |
| A beolvasott összes eset száma: 100                                          |           |              |                        |       |       |                      |
| Jelölés:                                                                     |           |              |                        |       |       |                      |
| r: Pearson-féle lineáris korrelációs együttható                              |           |              |                        |       |       |                      |
| Rpb: Wilcoxon-féle robusztus korrelációs együttható                          |           |              |                        |       |       |                      |
| Szignifikancia: +: p < 0,10 *: p < 0,05 **: p < 0,01 ***: p < 0,001;         |           |              |                        |       |       |                      |
| X = RNormX, Y = RNormY                                                       |           |              |                        |       |       |                      |
| A mind X, mind Y esetében érvényes értékkel rendelkező esetek száma: N = 100 |           |              |                        |       |       |                      |
|                                                                              | Változó   | Átlag        | Var.                   | Min.  | Max.  | Regressziós egyenlet |
| X:                                                                           | RNormX    | 50,91        | 106,37                 | 30,48 | 86,57 | x = 84,91 - 0,690y   |
| Y:                                                                           | RNormY    | 49,26        | 102,46                 | 23,70 | 79,55 | y = 83,11 - 0,665x   |
| X: Reziduális variancia (Hibavariancia): MSerr(X) = 57,58                    |           |              |                        |       |       |                      |
| Y: Reziduális variancia (Hibavariancia): MSerr(Y) = 55,46                    |           |              |                        |       |       |                      |
| Korrelációk és 90%-os intervallumbecslés az elméleti értékre:                |           |              |                        |       |       |                      |
| r =                                                                          | -0,677*** | (p = 0,0000) | C90 = (-0,757; -0,578) |       |       |                      |
| Rpb =                                                                        | -0,692*** | (p = 0,0000) | C90 = (-0,769; -0,596) |       |       |                      |

A közös elméleti reziduális hiba a (11.42) formula segítségével adódik:

$$\text{Var}(\varepsilon) = (1 - \rho^2) \cdot \text{Var}(Y) = (1 - (-0.70)^2) \cdot 10^2 = (1 - 0.49) \cdot 100 = 0.51 \cdot 100 = 51.$$

2. A jelen esetben a tapasztalati korrelációs együttható értéke,  $r = -0,677$  viszonylag közel van az elméleti  $-0,70$ -hez. Az elméleti regressziós függvény mintabeli becslése:

$$\hat{x} = 84,91 - 0,690 \cdot y,$$

illetve

$$\hat{y} = 83,11 - 0,665 \cdot x,$$

amelyek szintén közel esnek a fentebb kiszámított elméleti egyenes paramétereire (ahol is  $\beta_0 = 85$  és  $\beta_1 = -0,70$ ). A kapott mintabeli reziduális hibák: 50,91 és 49,26 csekély mértékben kisebbek, mint az elméleti 51.

3. Az adott 100-elemű minta alapján az elméleti  $\rho$  értékre vonatkozó 95%-os konfidencia-intervallum:

$$C_{0.95} = (r_1; r_2) = (-0,757; -0,578).$$

Ez tartalmazza az elméleti  $-0,70$ -es értéket, de e minta alapján az elméleti érték lehetne akár  $-0,75$  vagy  $-0,60$  is. Ha bármilyen hipotézisünk van a  $\rho$  elméleti értékre, az megvizsgálható e konfidencia-intervallum segítségével. Például a  $H_0: \rho_{XY} = -0,50$  hipotézis 5%-os szignifikanciaszinten elutasítható, mert a  $-0,50$ -es érték kívül esik a fenti 95%-os szintű konfidencia-intervallumon. Mivel  $r_2$  kisebb, mint  $-0,50$ ,  $H_0$ -lal szemben ez esetben a  $\rho < -0,50$  ellenhipotézist fogadjuk el.

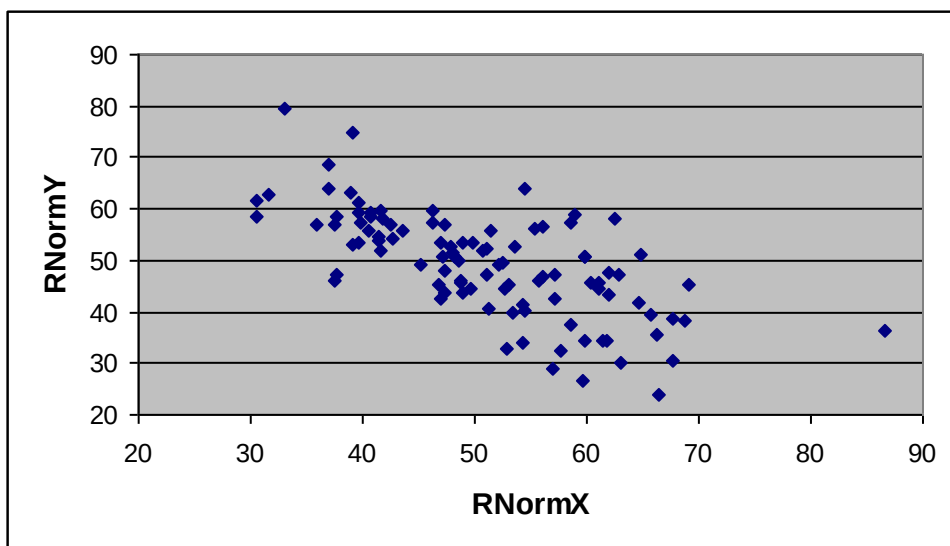
4. A Wilcoxon-féle robusztus tapasztalati korrelációs együttható értéke:

$$r_{pb} = -0,692.$$

Mivel a két változó együttes eloszlása normális, ez a korreláció csak kis mértékben különbözik a hagyományos Pearson-féle  $r$  együtthatótól.

5. Ha a két változó együttes eloszlását szeretnénk pontdiagramon ábrázolni, ezt a legegyszerűbben a következőképpen tehetjük meg.

- (a) A ROPstatban menjünk vissza az adatok táblázatához, kattintsunk a képernyő bal felső sarkában látható „Kiválaszt” parancsikorra, majd jelöljük ki és másoljuk be a memóriába az RNormX és RNormY változóoszlopot!
- (b) Indítsuk el az Excelt és illesszük be a táblázatba az RNormX és az RNormY adatoszlopot!
- (c) Végül a kijelölést megőrizve, az Excel ábrakészítő ikonjára rákattintva válasszuk ki az XY pontdiagram ábratípust, melynek eredményeképpen a 11.9. ábrán látható grafikonhoz jutunk, ha a megfelelő rovatokba beírjuk a két változó nevét (RNormX és RNormY), továbbá az X- és az Y-tengely léptékét (10) és határait (20 és 90).



**11.9. ábra.** A 11.7. táblázat RNormX és RNormY normális együttes eloszlású változójának Excel segítségével készített kétváltozós pontdiagramja

### 11.8.2. Korrelációs és parciális korrelációs mátrix készítése a ROPstat segítségével

A 11.8.1. alponthan ismertetett 'Korreláció, egyszerű regresszió' rutint akkor érdemes használni, ha kevés számú változópár kapcsolatáról szeretnénk részletes információt kapni. Ha több változónk van és megelégszünk a köztük lévő korrelációk kiszámításával és szignifikanciájuk vizsgálatával, akkor ugyanazon programcsoport 'Korrelációs és parciális korrelációs mátrix készítése' rutint érdemes használnunk. Az elemzésbe bevonandó változók kijelölése ugyanolyan egyszerű, mint az előző esetben. Megjegyezzük, hogy a korrelációk kiszámításakor ez a program csak azokat az eseteket (személyeket) veszi figyelembe, amelyeknek az elemzésre kijelölt *minden* változó szerint érvényes értékük van. Ha egy konkrét változópár esetében a köztük lévő korreláció kiszámításához minden olyan esetet figyelembe akarunk venni, amelyek mindkét változó esetében érvényes értékkel rendelkeznek, akkor az előző alponthan ismertetett rutint használjuk.

Példaként elkészítettük az ANTR100.MST fájl (lásd 3.5. alfejezet) gyerekre vonatkozó antropometriai változóira a páronkénti korrelációs együtthatók táblázatát, melyet **korrelációs mátrixnak** hívunk. Az elemzéseket a fiúkra és a lányokra külön végeztük el. Ehhez a 'Nem' változót egérrel betettük a feltételes csoportosító változó ablakába. A két csoport korrelációinak összehasonlításához bejelöltük a „Feltételes csoportok páronkénti összehasonlítása” opciót. Az elemzés eredményét a 11.8. táblázatban foglaltuk össze.

A 11.8. táblázat listájához a következő megjegyzéseket fűzzük.

1. *Fiúk* A korrelációs mátrix sorait és oszlopait az elemzésbe bevont változók fejlecezik. Két változó korrelációját a megfelelő sor és oszlop kereszteződésénél lévő érték adja meg. Például a Tsúly0 (születési testsúly) és Thossz0 (születési testhossz) közötti korreláció 0.794, igen tetemes. A 0,655 melletti két csillag (\*\*) adja tudtul, hogy ez az érték 1%-os szinten szignifikáns (azaz szignifikánsan különbözik a 0-tól). A korrelációk közös szabadságfoka ( $f = 45$ ) a táblázat fölött található  $f$  érték. Itt jegyezzük meg, hogy a ROPstatban külön kérésre a

korrelációk szignifikanciájának pontos p-értéke is kilistázható, amiből az 1%-nál erősebb szignifikanciák is azonosíthatók.

A ROPstatban az „X változók” ablakba betett változók a létrehozott korrelációs mátrix sorait, az „Y változók” ablakba betettek pedig a mátrix oszlopait fejlecezik. Ilyen esetben a program csak az X-ek és az Y-ok között számít páronként korrelációt (X-eken vagy Y-okon belül nem). Ha a kijelölésnél csak egy ablakba teszünk változót, mint a 11.8. táblázat esetében is, akkor a korrelációs mátrix sorait és oszlopait ugyanazok a változók képezik, így ekkor a kapott korrelációs mátrix szimmetrikus. Például a 2. sor 3. oszlopában található korreláció (0,021) ugyanakkora, mint a 3. sor 2. oszlopában helyet foglaló érték. Ez esetben a főátlóban – az (1, 1), (2, 2), (3, 3) stb. indexű helyeken – a változók saját magukkal való korrelációi láthatók, amelyek szükségképpen 1-gyel egyenlők.

A korrelációkat szakmailag elemezve elmondhatjuk, hogy a legszorosabb lineáris típusú kapcsolat az egy időben felvett (születési súly és testhossz, illetve 10 éves kori testsúly és testmagasság) között van. Lényegesen szorosabb összefüggés van például Tsúly0 és Thossz0 között (0,655), mint Tsúly0 és Tsúly10 (–0,083) vagy Thossz0 és Tmag10 (0,256) között.

A korrelációs mátrix alatt a **teljes függetlenség** Bartlett-féle próbája (vö. Appelbaum és McCall, 1983) azt teszteli, hogy van-e egyáltalán valamilyen elméleti lineáris típusú kapcsolat az elemzésbe bevont változók együttesében. Ez azt a hatást próbálja meg kiküszöbölni, hogy igen sok változó esetén egy-egy szignifikáns korreláció a véletlen eredménye is lehet. Ha tehát ez a Bartlett féle  $\chi^2$  próbastatisztika nem szignifikáns (a mi esetünkben nem kell aggódni, mert  $p = 0,0000$ , tehát  $\chi^2$  még 0,1%-os szinten szignifikáns) és a korrelációs mátrix terjedelmes, akkor némi fenntartással kell fogadni az esetleg fellépő kevés számú korrelációt. 5%-os szinten szignifikáns eredmények közül ugyanis várhatóan minden huszadik a véletlennek köszönhető.

2. *Lányok* A fiúk csoportjában megfigyelt jelenségek a lányok korrelációs mátrixára is jellemzők.

3. Az, hogy a vizsgált négy változó kapcsolatrendszer (lineáris korrelációinak mintázata) a két csoport esetében ugyanolyan jellegű, a két korrelációs mátrix statisztikai összehasonlításából is kiderül. Az összehasonlítás a Z-transzformáció segítségével történt, minden változó-párra külön-külön. A táblázatbeli értékek a 11.4.6. alpontban ismertetett  $Z^*$  próbastatisztika értékét adják meg. Például a 3. sor (Thossz0) és az 1. oszlop (Tsúly0) kereszteződésénél található –1,698 érték azt jelzi, hogy a

$$H_0: \rho(\text{Thossz0}, \text{Tsúly0} \mid \text{Fiú}) = \rho(\text{Thossz0}, \text{Tsúly0} \mid \text{Lány})$$

nullhipotézis vizsgálatában definiált  $Z^*$  próbastatisztika értéke:  $Z^* = -1,698$  (vö. 11.4.6. alpont). Az, hogy ezen érték mellett egy + jelet látunk, úgy értelmezendő, hogy jelen esetben egy tendenciát figyelhetünk meg a fiúk és lányok közötti korreláció különbségére a jelzett két változó szerint (a két tapasztalati korreláció: 0,655 és 0,812 közötti különbség:  $p < 0,10$  szintű tendenciát jelez). Mintha a lányok esetében a születési súly és a születési testhossz között valamivel szorosabb lenne a kapcsolat, mint a fiúk esetében.



## 11.8. táblázat. A ROPstat 'Korrelációs mátrix készítése' rutinjának egy eredménylistája

Korrelációs mátrix készítése

Feltételes csoportosító változó: Nem

Feltételes csoportok gyakoriságai

| Index | Csoportnév | Gyakoriság |
|-------|------------|------------|
| 1     | fiú        | 47         |
| 2     | lány       | 53         |

CSOPORTINDEX: 1. CSOPORTNÉV: fiú

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 47

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 47

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE KORRELÁCIÓS MÁTRIX

(f = 45; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | -0,083  | 0,655** | 0,210   |
| Tsúly10 | -0,083  | 1       | 0,021   | 0,547** |
| Thossz0 | 0,655** | 0,021   | 1       | 0,256+  |
| Tmag10  | 0,210   | 0,547** | 0,256+  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata: Khi2(6) = 45,895 (p = 0,0000)

CSOPORTINDEX: 2. CSOPORTNÉV: lány

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 53

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 53

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE KORRELÁCIÓS MÁTRIX

(f = 51; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | 0,132   | 0,812** | 0,068   |
| Tsúly10 | 0,132   | 1       | 0,227   | 0,639** |
| Thossz0 | 0,812** | 0,227   | 1       | 0,289*  |
| Tmag10  | 0,068   | 0,639** | 0,289*  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata: Khi2(6) = 89,894 (p = 0,0000)

A FELTÉTELES CSOPORTOK KORRELÁCIÓS MÁTRIXAINAK PÁRONKÉNTI ÖSSZEHASONLÍTÁSA

Az összehasonlítandó két csoport:

| Index | Csoportnév | Elemszám |
|-------|------------|----------|
| 1     | fiú        | 47       |
| 2     | lány       | 53       |

A Pearson-féle korrelációk standardizált különbségeinek táblázata (Csop1-Csop2)

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10 |
|---------|---------|---------|---------|--------|
| Tsúly0  | .       | -1,043  | -1,698+ | 0,699  |
| Tsúly10 | -1,043  | .       | -1,017  | -0,691 |
| Thossz0 | -1,698+ | -1,017  | .       | -0,174 |
| Tmag10  |         |         |         | 0,699  |
| 0,174   |         |         |         | .      |

-0,691

-

4. A vizsgált antropometriai változók (testsúly és testhossz különböző életkorokban) ki-sebbs-nagyobb mértékű genetikai hatás alatt állnak. Elképzelhető, hogy ez a közös biológiai gyökér a köztük lévő kapcsolatra is kihat. Vajon milyen lenne közöttük a korreláció, ha sikerülne ezt a genetikai hatást kiszűrni?

A bemutatott változók mellett a szóban forgó ANTR100.MST adatfájlban rendelkezésre álltak a szülők súly- és magasságadatai is. Az említett genetikai hatás nyilván összefüggésben

van az ő adataikkal. Ezért e hatás kiküszöbölését olyan parciális korrelációk kiszámításával próbáltuk elérni, amelyben a kiszűrendő változók – összesen négy – az anya és az apa test- súlya, illetve testmagassága. Az elemzés eredményeit a 11.9. táblázat tartalmazza.

A 11.9. táblázat eredménylistájából azt olvashatjuk ki, hogy a szülők antropometriai változóinak kiszűrése nem változtatott lényegesen a korrelációs összképen. A rövid szakmai konklúzió a következő.

- Születéskor, illetve 10 éves korban a testsúly és a testhossz/testmagasság igen erős, esetenként 0,6-0,8 körüli korrelációban van egymással.
- A testmagasság születéskori és 10 éves kori értéke lineárisan összefügg egymással, de a kapcsolat szorossága elég gyenge, 0,25 és 0,30 körüli.
- A testsúly születéskori és 10 éves kori értéke között nincs lineáris típusú kapcsolat.
- Különböző korokban mért különböző antropometriai mértékek általában nincsenek közvetlen lineáris kapcsolatban egymással.
- A fiúk és a lányok mintájában kiszámított korrelációk mintázata nem tér el számottevő mértékben.

## 11.9. táblázat. Parciális korrelációs mátrix készítése a ROPstattal

Korrelációs mátrix készítése

Feltételes csoportosító változó: Nem

Feltételes csoportok gyakoriságai

| Index | Csoportnév | Gyakoriság |
|-------|------------|------------|
| 1     | fiú        | 47         |
| 2     | lány       | 53         |

CSOPORTINDEX: 1. CSOPORTNÉV: fiú

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 47

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 47

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE PARCIÁLIS KORRELÁCIÓS MÁTRIX

(f = 41; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

Kiszűrve az alábbi változók lineáris hatása:

(6) Anyasúly (7) Apasúly (8) Anyamag (9) Apamag

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | -0,122  | 0,662** | 0,176   |
| Tsúly10 | -0,122  | 1       | 0,034   | 0,601** |
| Thossz0 | 0,662** | 0,034   | 1       | 0,270+  |
| Tmag10  | 0,176   | 0,601** | 0,270+  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata: Khi2(6) = 47,298 (p = 0,0000)

CSOPORTINDEX: 2. CSOPORTNÉV: lány

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 53

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 53

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE PARCIÁLIS KORRELÁCIÓS MÁTRIX

(f = 47; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

Kiszűrve az alábbi változók lineáris hatása:

(6) Anyasúly (7) Apasúly (8) Anyamag (9) Apamag

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | -0,022  | 0,790** | 0,012   |
| Tsúly10 | -0,022  | 1       | 0,045   | 0,388** |
| Thossz0 | 0,790** | 0,045   | 1       | 0,246+  |
| Tmag10  | 0,012   | 0,388** | 0,246+  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata: Khi2(6) = 59,979 (p = 0,0000)

A FELTÉTELES CSOPORTOK KORRELÁCIÓS MÁTRIXAINAK PÁRONKÉNTI ÖSSZEHAJONLÍTÁSA

Az összehasonlítandó két csoport:

| Index | Csoportnév | Elemszám |
|-------|------------|----------|
| 1     | fiú        | 47       |
| 2     | lány       | 53       |

Kovariáns változók száma: 4

A Pearson-féle parciális korrelációk standardizált különbségeinek táblázata (Csop1-Csop2)

| Változó | Tsúly0 | Tsúly10 | Thossz0 | Tmag10 |
|---------|--------|---------|---------|--------|
| Tsúly0  | .      | -0,466  | -1,280  | 0,770  |
| Tsúly10 | -0,466 | .       | -0,054  | 1,318  |
| Thossz0 | -1,280 | -0,054  | .       | 0,120  |
| Tmag10  |        |         |         | 0,770  |
| 0,120   |        |         |         | .      |

1,318

## 11.8. Lineáris korreláció és regresszió a ROPstatban

### 11.8.1. A ROPstat 'Korreláció, egyszerű regresszió rutinja

A fejezetünkben leírt korrelációs és regressziós eljárások nagy többségét a ROPstat 'Változók kapcsolatának vizsgálata' programcsoportjának 'Korreláció, egyszerű regresszió' rutinja segítségével lehet végrehajtani. Ennek szemléltetésére most nem valódi kutatási adatokat használunk fel, hanem a ROPstat által generált véletlen adatokat, hogy a rutin segítségével egyben a kétdimenziós normális eloszlást is bemutathassuk. E célból először is létrehoztunk egy kétdimenziós normális eloszlású (X, Y) változópárt 100 esettel a ROPstat 'Új random változó' rutinjában. A generálás ugyanúgy végezhető el, mint a 11.4.4. alpontban ismertetett (r, Z(r)) változópár esetében, csak most a 'Kétdimenziós normális' eloszlástípust kell kiválasztani. Ekkor beírható a két új változó neve (pl. RNormX és RNormY), továbbá a két változó elméleti átlaga és szórása, illetve a köztük lévő  $\rho$  elméleti korreláció. Ezzel a generálással 100 esetet hoztunk létre. Az átlagok és szórások esetében adjunk meg  $\mu_X = 50$  és  $\sigma_X = 10$ , illetve  $\mu_Y = 50$  és  $\sigma_Y = 10$  értékeket, s az elméleti korrelációra a  $\rho = -0,70$  értéket írjuk be.

Miután létrehoztuk ennek a kétdimenziós normális eloszlású (X, Y) változópárnak egy  $n = 100$  elemű véletlen mintáját, a ROPstat statisztikai menüjében kiválasztjuk a 'Változók kapcsolatának vizsgálata' programcsoportot, majd ezen belül a 'Korreláció, egyszerű regresszió' menüpontot. E rutin menüablakában tegyük egérrel az RNormX változót az „X változók”, a RNormY változót pedig „Y változók” ablakába. Eredményül a 11.7. táblázatban látható eredménylistát kaptuk, amelyhez a következő megjegyzéseket fűzzük.

**11.7. táblázat.** Egy normális eloszlású (X, Y) változópár ( $\mu_X = 50$ ,  $\mu_Y = 50$ ,  $\sigma_X = 10$ ,  $\sigma_Y = 10$ ,  $\rho = -0,70$ ) kapcsolatának részletes vizsgálata a ROPstat 'Korreláció, regresszió, kétváltozós diagram' rutinja segítségével egy 100-elemű véletlen (X, Y)-minta alapján

|                                                                               |           |              |                        |       |       |                      |
|-------------------------------------------------------------------------------|-----------|--------------|------------------------|-------|-------|----------------------|
| Korreláció, lineáris regresszió                                               |           |              |                        |       |       |                      |
| A beolvasott összes eset száma: 100                                           |           |              |                        |       |       |                      |
| Jelölés:                                                                      |           |              |                        |       |       |                      |
| r: Pearson-féle lineáris korrelációs együttható                               |           |              |                        |       |       |                      |
| Rpb: Wilcox-féle robusztus korrelációs együttható                             |           |              |                        |       |       |                      |
| Szignifikancia: +: $p < 0,10$ *: $p < 0,05$ **: $p < 0,01$ ***: $p < 0,001$ ; |           |              |                        |       |       |                      |
| X = RNormX, Y = RNormY                                                        |           |              |                        |       |       |                      |
| A mind X, mind Y esetében érvényes értékkel rendelkező esetek száma: N = 100  |           |              |                        |       |       |                      |
|                                                                               | Változó   | Átlag        | Var.                   | Min.  | Max.  | Regressziós egyenlet |
| X:                                                                            | RNormX    | 50,91        | 106,37                 | 30,48 | 86,57 | $x = 84,91 - 0,690y$ |
| Y:                                                                            | RNormY    | 49,26        | 102,46                 | 23,70 | 79,55 | $y = 83,11 - 0,665x$ |
| X: Reziduális variancia (Hibavariancia): MSerr(X) = 57,58                     |           |              |                        |       |       |                      |
| Y: Reziduális variancia (Hibavariancia): MSerr(Y) = 55,46                     |           |              |                        |       |       |                      |
| Korrelációk és 90%-os intervallumbecslés az elméleti értékre:                 |           |              |                        |       |       |                      |
| r =                                                                           | -0,677*** | (p = 0,0000) | C90 = (-0,757; -0,578) |       |       |                      |
| Rpb =                                                                         | -0,692*** | (p = 0,0000) | C90 = (-0,769; -0,596) |       |       |                      |

1. Először is megjegyezzük, hogy az Y változó X-re vonatkozó *elméleti* lineáris regressziós egyenesének  $\beta_1$  együtthatójára  $\sigma_X \beta_1 = \rho \sigma_Y$ , amiből egyszerű átrendezéssel és  $\sigma_1 = \sigma_2 = 10$ , valamint  $\rho = -0.70$  figyelembevételével adódik, hogy

$$\beta_1 = \rho \cdot \sigma_1 / \sigma_2 = \rho = -0.70.$$

A  $\beta_0$  együtthatót a (11.20) formulába behelyettesítve kapjuk:

$$\beta_0 = \mu_Y - \beta_1 \mu_X = 50 - (-0.70) \cdot 50 = 50 + 0.70 \cdot 50 = 50 + 35 = 85.$$

Így tehát az Y változó X-re vonatkozó elméleti lineáris regressziós függvénye:

$$\hat{Y} = 85 - 0.70 \cdot X.$$

Mivel esetünkben  $\mu_1 = \mu_2 = 50$  és  $\sigma_1 = \sigma_2 = 10$ , ugyanígy néz ki az X változó Y-ra vonatkozó elméleti lineáris regressziós függvénye is:

$$\hat{X} = 85 - 0.70 \cdot Y.$$

A közös elméleti reziduális hiba a (11.42) formula segítségével adódik:

$$\text{Var}(\varepsilon) = (1 - \rho^2) \cdot \text{Var}(Y) = (1 - (-0.70)^2) \cdot 10^2 = (1 - 0.49) \cdot 100 = 0.51 \cdot 100 = 51.$$

2. A jelen esetben a tapasztalati korrelációs együttható értéke,  $r = -0,677$  viszonylag közel van az elméleti  $-0,70$ -hez. Az elméleti regressziós függvény mintabeli becslése:

$$\hat{x} = 84,91 - 0,690 \cdot y,$$

illetve

$$\hat{y} = 83,11 - 0,665 \cdot x,$$

amelyek szintén közel esnek a fentebb kiszámított elméleti egyenes paramétereire (ahol is  $\beta_0 = 85$  és  $\beta_1 = -0,70$ ). A kapott mintabeli reziduális hibák: 50,91 és 49,26 csekély mértékben kisebbek, mint az elméleti 51.

3. Az adott 100-elemű minta alapján az elméleti  $\rho$  értékre vonatkozó 95%-os konfidencia-intervallum:

$$C_{0,95} = (r_1; r_2) = (-0,757; -0,578).$$

Ez tartalmazza az elméleti  $-0,70$ -es értéket, de e minta alapján az elméleti érték lehetne akár  $-0,75$  vagy  $-0,60$  is. Ha bármilyen hipotézisünk van a  $\rho$  elméleti értékre, az megvizsgálható e konfidencia-intervallum segítségével. Például a  $H_0: \rho_{XY} = -0,50$  hipotézis 5%-os szignifikanciaszinten elutasítható, mert a  $-0,50$ -es érték kívül esik a fenti 95%-os szintű konfidencia-intervallumon. Mivel  $r_2$  kisebb, mint  $-0,50$ ,  $H_0$ -lal szemben ez esetben a  $\rho < -0,50$  ellenhipotézist fogadjuk el.

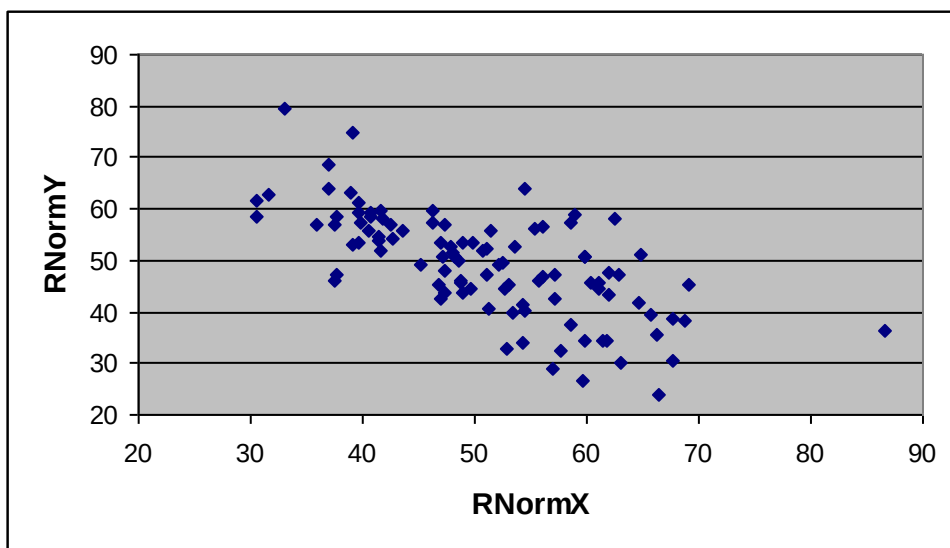
4. A Wilcoxon-féle robusztus tapasztalati korrelációs együttható értéke:

$$r_{pb} = -0,692.$$

Mivel a két változó együttes eloszlása normális, ez a korreláció csak kis mértékben különbözik a hagyományos Pearson-féle  $r$  együtthatótól.

5. Ha a két változó együttes eloszlását szeretnénk pontdiagramon ábrázolni, ezt a legegyszerűbben a következőképpen tehetjük meg.

- (a) A ROPstatban menjünk vissza az adatok táblázatához, kattintsunk a képernyő bal felső sarkában látható „Kiválaszt” parancsikorra, majd jelöljük ki és másoljuk be a memóriába az RNormX és RNormY változóoszlopot!
- (b) Indítsuk el az Excelt és illesszük be a táblázatba az RNormX és az RNormY adatoszlopot!
- (c) Végül a kijelölést megőrizve, az Excel ábrakészítő ikonjára rákattintva válasszuk ki az XY pontdiagram ábratípust, melynek eredményeképpen a 11.9. ábrán látható grafikonhoz jutunk, ha a megfelelő rovatokba beírjuk a két változó nevét (RNormX és RNormY), továbbá az X- és az Y-tengely léptékét (10) és határait (20 és 90).



**11.9. ábra.** A 11.7. táblázat RNormX és RNormY normális együttes eloszlású változójának Excel segítségével készített kétváltozós pontdiagramja

### 11.8.2. Korrelációs és parciális korrelációs mátrix készítése a ROPstat segítségével

A 11.8.1. alponthban ismertetett 'Korreláció, egyszerű regresszió' rutint akkor érdemes használni, ha kevés számú változópár kapcsolatáról szeretnénk részletes információt kapni. Ha több változónk van és megelégszünk a köztük lévő korrelációk kiszámításával és szignifikanciájuk vizsgálatával, akkor ugyanazon programcsoport 'Korrelációs és parciális korrelációs mátrix készítése' rutint érdemes használnunk. Az elemzésbe bevonandó változók kijelölése ugyanolyan egyszerű, mint az előző esetben. Megjegyezzük, hogy a korrelációk kiszámításakor ez a program csak azokat az eseteket (személyeket) veszi figyelembe, amelyeknek az elemzésre kijelölt *minden* változó szerint érvényes értékük van. Ha egy konkrét változópár esetében a köztük lévő korreláció kiszámításához minden olyan esetet figyelembe akarunk venni, amelyek mindkét változó esetében érvényes értékkel rendelkeznek, akkor az előző alponthban ismertetett rutint használjuk.

Példaként elkészítettük az ANTR100.MST fájl (lásd 3.5. alfejezet) gyerekre vonatkozó antropometriai változókra a páronkénti korrelációs együtthatók táblázatát, melyet **korrelációs mátrixnak** hívunk. Az elemzéseket a fiúkra és a lányokra külön végeztük el. Ehhez a 'Nem' változót egérrel betettük a feltételes csoportosító változó ablakába. A két csoport korrelációinak összehasonlításához bejelöltük a „Feltételes csoportok páronkénti összehasonlítása” opciót. Az elemzés eredményét a 11.8. táblázatban foglaltuk össze.

A 11.8. táblázat listájához a következő megjegyzéseket fűzzük.

1. *Fiúk* A korrelációs mátrix sorait és oszlopait az elemzésbe bevont változók fejlecezik. Két változó korrelációját a megfelelő sor és oszlop kereszteződésénél lévő érték adja meg. Például a Tsúly0 (születési testsúly) és Thossz0 (születési testhossz) közötti korreláció 0.794, igen tetemes. A 0,655 melletti két csillag (\*\*) adja tudtul, hogy ez az érték 1%-os szinten szignifikáns (azaz szignifikánsan különbözik a 0-tól). A korrelációk közös szabadságfoka ( $f =$

45) a táblázat fölött található  $f$  érték. Itt jegyezzük meg, hogy a ROPstatban külön kérésre a korrelációk szignifikanciájának pontos  $p$ -értéke is kilistázható, amiből az 1%-nál erősebb szignifikanciák is azonosíthatók.

A ROPstatban az „X változók” ablakba betett változók a létrehozott korrelációs mátrix sorait, az „Y változók” ablakba betettek pedig a mátrix oszlopait fejlécezik. Ilyen esetben a program csak az X-ek és az Y-ok között számít páronként korrelációt (X-eken vagy Y-okon belül nem). Ha a kijelölésnél csak egy ablakba teszünk változót, mint a 11.8. táblázat esetében is, akkor a korrelációs mátrix sorait és oszlopait ugyanazok a változók képezik, így ekkor a kapott korrelációs mátrix szimmetrikus. Például a 2. sor 3. oszlopában található korreláció (0,021) ugyanakkora, mint a 3. sor 2. oszlopában helyet foglaló érték. Ez esetben a főátlóban – az (1, 1), (2, 2), (3, 3) stb. indexű helyeken – a változók saját magukkal való korrelációi láthatók, amelyek szükségképpen 1-gyel egyenlők.

A korrelációkat szakmailag elemezve elmondhatjuk, hogy a legszorosabb lineáris típusú kapcsolat az egy időben felvett (születési súly és testhossz, illetve 10 éves kori testsúly és testmagasság) között van. Lényegesen szorosabb összefüggés van például Tsúly0 és Thossz0 között (0,655), mint Tsúly0 és Tsúly10 (–0,083) vagy Thossz0 és Tmag10 (0,256) között.

A korrelációs mátrix alatt a **teljes függetlenség** Bartlett-féle próbája (vö. Appelbaum és McCall, 1983) azt teszteli, hogy van-e egyáltalán valamilyen elméleti lineáris típusú kapcsolat az elemzésbe bevont változók együttesében. Ez azt a hatást próbálja meg kiküszöbölni, hogy igen sok változó esetén egy-egy szignifikáns korreláció a véletlen eredménye is lehet. Ha tehát ez a Bartlett féle  $\chi^2$  próbastatisztika nem szignifikáns (a mi esetünkben nem kell aggódni, mert  $p = 0,0000$ , tehát  $\chi^2$  még 0,1%-os szinten szignifikáns) és a korrelációs mátrix terjedelmes, akkor némi fenntartással kell fogadni az esetleg fellépő kevés számú korrelációt. 5%-os szinten szignifikáns eredmények közül ugyanis várhatóan minden huszadik a véletlennek köszönhető.

2. *Lányok* A fiúk csoportjában megfigyelt jelenségek a lányok korrelációs mátrixára is jellemzők.

3. Az, hogy a vizsgált négy változó kapcsolatrendszer (lineáris korrelációinak mintázata) a két csoport esetében ugyanolyan jellegű, a két korrelációs mátrix statisztikai összehasonlításából is kiderül. Az összehasonlítás a Z-transzformáció segítségével történt, minden változó-párra külön-külön. A táblázatbeli értékek a 11.4.6. alpontban ismertetett  $Z^*$  próbastatisztika értékét adják meg. Például a 3. sor (Thossz0) és az 1. oszlop (Tsúly0) kereszteződésénél található –1,698 érték azt jelzi, hogy a

$$H_0: \rho(\text{Thossz0}, \text{Tsúly0} \mid \text{Fiú}) = \rho(\text{Thossz0}, \text{Tsúly0} \mid \text{Lány})$$

nullhipotézis vizsgálatában definiált  $Z^*$  próbastatisztika értéke:  $Z^* = -1,698$  (vö. 11.4.6. alpont). Az, hogy ezen érték mellett egy + jelet látunk, úgy értelmezendő, hogy jelen esetben egy tendenciát figyelhetünk meg a fiúk és lányok közötti korreláció különbségére a jelzett két változó szerint (a két tapasztalati korreláció: 0,655 és 0,812 közötti különbség:  $p < 0,10$  szintű tendenciát jelez). Mintha a lányok esetében a születési súly és a születési testhossz között valamivel szorosabb lenne a kapcsolat, mint a fiúk esetében.

**11.8. táblázat.** A ROPstat 'Korrelációs mátrix készítése' rutinjának egy eredménylistája

#### Korrelációs mátrix készítése

Feltételes csoportosító változó: Nem

Feltételes csoportok gyakoriságai

| Index | Csoportnév | Gyakoriság |
|-------|------------|------------|
| 1     | fiú        | 47         |
| 2     | lány       | 53         |

CSOPORTINDEX: 1. CSOPORTNÉV: fiú

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 47

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 47

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE KORRELÁCIÓS MÁTRIX

(f = 45; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | -0,083  | 0,655** | 0,210   |
| Tsúly10 | -0,083  | 1       | 0,021   | 0,547** |
| Thossz0 | 0,655** | 0,021   | 1       | 0,256+  |
| Tmag10  | 0,210   | 0,547** | 0,256+  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata:  $\chi^2(6) = 45,895$  (p = 0,0000)

CSOPORTINDEX: 2. CSOPORTNÉV: lány

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 53

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 53

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE KORRELÁCIÓS MÁTRIX

(f = 51; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | 0,132   | 0,812** | 0,068   |
| Tsúly10 | 0,132   | 1       | 0,227   | 0,639** |
| Thossz0 | 0,812** | 0,227   | 1       | 0,289*  |
| Tmag10  | 0,068   | 0,639** | 0,289*  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata:  $\chi^2(6) = 89,894$  (p = 0,0000)

A FELTÉTELES CSOPORTOK KORRELÁCIÓS MÁTRIXAINAK PÁRONKÉNTI ÖSSZEHASONLÍTÁSA

Az összehasonlítandó két csoport:

| Index | Csoportnév | Elemszám |
|-------|------------|----------|
| 1     | fiú        | 47       |
| 2     | lány       | 53       |

A Pearson-féle korrelációk standardizált különbségeinek táblázata (Csop1-Csop2)

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10 |
|---------|---------|---------|---------|--------|
| Tsúly0  | .       | -1,043  | -1,698+ | 0,699  |
| Tsúly10 | -1,043  | .       | -1,017  | -0,691 |
| Thossz0 | -1,698+ | -1,017  | .       | -0,174 |
| Tmag10  | 0,699   | -0,691  | -0,174  | .      |

4. A vizsgált antropometriai változók (testsúly és testhossz különböző életkorokban) kisebb-nagyobb mértékű genetikai hatás alatt állnak. Elképzelhető, hogy ez a közös biológiai gyökér a köztük lévő kapcsolatra is kihat. Vajon milyen lenne közöttük a korreláció, ha sikerülne ezt a genetikai hatást kiszűrni?

A bemutatott változók mellett a szóban forgó ANTR100.MST adatfájlban rendelkezésre álltak a szülők súly- és magasságadatai is. Az említett genetikai hatás nyilván összefüggésben van az ő adataikkal. Ezért e hatás kiküszöbölését olyan parciális korrelációk kiszámításával



próbáltuk elérni, amelyben a kiszűrendő változók – összesen négy – az anya és az apa testsúlya, illetve testmagassága. Az elemzés eredményeit a 11.9. táblázat tartalmazza.

A 11.9. táblázat eredménylistájából azt olvashatjuk ki, hogy a szülők antropometriai változóinak kiszűrése nem változtatott lényegesen a korrelációs összképen. A rövid szakmai konklúzió a következő.

- Születéskor, illetve 10 éves korban a testsúly és a testhossz/testmagasság igen erős, esetenként 0,6-0,8 körüli korrelációban van egymással.
- A testmagasság születéskori és 10 éves kori értéke lineárisan összefügg egymással, de a kapcsolat szorossága elég gyenge, 0,25 és 0,30 körüli.
- A testsúly születéskori és 10 éves kori értéke között nincs lineáris típusú kapcsolat.
- Különböző korokban mért különböző antropometriai mértékek általában nincsenek közvetlen lineáris kapcsolatban egymással.
- A fiúk és a lányok mintájában kiszámított korrelációk mintázata nem tér el számottevő mértékben.

### 11.9. táblázat. Parciális korrelációs mátrix készítése a ROPstattal

Korrelációs mátrix készítése

Feltételes csoportosító változó: Nem

Feltételes csoportok gyakoriságai

| Index | Csoportnév | Gyakoriság |
|-------|------------|------------|
| 1     | fiú        | 47         |
| 2     | lány       | 53         |

CSOPORTINDEX: 1. CSOPORTNÉV: fiú

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 47

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 47

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE PARCIÁLIS KORRELÁCIÓS MÁTRIX

(f = 41; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

Kiszűrve az alábbi változók lineáris hatása:

(6) Anyasúly (7) Apasúly (8) Anyamag (9) Apamag

| Változó | Tsúly0  | Tsúly10 | Thossz0 | Tmag10  |
|---------|---------|---------|---------|---------|
| Tsúly0  | 1       | -0,122  | 0,662** | 0,176   |
| Tsúly10 | -0,122  | 1       | 0,034   | 0,601** |
| Thossz0 | 0,662** | 0,034   | 1       | 0,270+  |
| Tmag10  | 0,176   | 0,601** | 0,270+  | 1       |

A változók teljes függetlenségének Bartlett-féle vizsgálata: Khi2(6) = 47,298 (p = 0,0000)

CSOPORTINDEX: 2. CSOPORTNÉV: lány

A beolvasott összes eset száma: 100

Az aktuális csoport elemszáma: 53

Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 53

A jelen elemzésben csak ezek az esetek vesznek részt.

PEARSON-FÉLE PARCIÁLIS KORRELÁCIÓS MÁTRIX

(f = 47; +: p < 0,10 \*: p < 0,05 \*\*: p < 0,01)

Kiszűrve az alábbi változók lineáris hatása:

(6) Anyasúly (7) Apasúly (8) Anyamag (9) Apamag

| Változó | Tsúly0 | Tsúly10 | Thossz0 | Tmag10  |
|---------|--------|---------|---------|---------|
| Tsúly0  | 1      | -0,022  | 0,790** | 0,012   |
| Tsúly10 | -0,022 | 1       | 0,045   | 0,388** |

|         |         |         |        |        |
|---------|---------|---------|--------|--------|
| Thossz0 | 0,790** | 0,045   | 1      | 0,246+ |
| Tmag10  | 0,012   | 0,388** | 0,246+ | 1      |

A változók teljes függetlenségének Bartlett-féle vizsgálata:  $\text{Khi}^2(6) = 59,979$  ( $p = 0,0000$ )

A FELTÉTELES CSOPORTOK KORRELÁCIÓS MÁTRIXAINAK PÁRONKÉNTI ÖSSZEHASONLÍTÁSA

Az összehasonlítandó két csoport:

| Index | Csoportnév | Elemzés |
|-------|------------|---------|
| 1     | fiú        | 47      |
| 2     | lány       | 53      |

Kovariáns változók száma: 4

A Pearson-féle parciális korrelációk standardizált különbségeinek táblázata (Csop1-Csop2)

| Változó | Tsúly0 | Tsúly10 | Thossz0 | Tmag10 |
|---------|--------|---------|---------|--------|
| Tsúly0  | .      | -0,466  | -1,280  | 0,770  |
| Tsúly10 | -0,466 | .       | -0,054  | 1,318  |
| Thossz0 | -1,280 | -0,054  | .       | 0,120  |
| Tmag10  | 0,770  | 1,318   | 0,120   | .      |

## 12.4. A monotonitás vizsgálata a ROPstatban

Rangkorrelációk, illetve monotonitási együtthatók a ROPstatban a 'Változók kapcsolatának vizsgálata' programcsoport alábbi három rutinja segítségével számíthatók.

1. A 'Korrelációs és parciális korrelációs mátrix készítése' rutinban nemcsak Pearson-féle, hanem rangkorrelációs mátrixok is készíthetők a Spearman-féle  $r_s$  és a Kendall-féle tau együtthatókra. Az ehhez szükséges lépések ugyanazok, mint a Pearson-féle korrelációs együttható esetében (lásd 11.8.2. alpont), csak a korrelációs együttható típusát Pearson helyett Spearman-re vagy Kendall-ra kell átállítani. Parciális rangkorrelációkhoz a kiszűrendő változókat a kovariáns változók ablakába kell betenni. Emlékeztetünk arra, hogy ez a rutin csak azokat az eseteket (személyeket) veszi figyelembe, amelyeknek az elemzésre kijelölt minden változó szerint érvényes értékük van.

Ezen elemzések szemléltetésére bemutatjuk az ANTR100.MST fájl (vö. 3.5. alfejezet) gyermekkel és szülővel kapcsolatos változóinak Spearman-féle (lásd 12.3. táblázat) és Kendall-féle (lásd 12.4. táblázat) rangkorrelációs mátrixát.

### 12.3. táblázat. Spearman-féle rangkorrelációs mátrix készítése ROPstatban

|                                                                                |          |         |         |         |
|--------------------------------------------------------------------------------|----------|---------|---------|---------|
| Rangkorrelációs mátrix készítése                                               |          |         |         |         |
| A beolvasott összes eset száma: 100                                            |          |         |         |         |
| Az összes kiválasztott változóra érvényes adattal rendelkező esetek száma: 100 |          |         |         |         |
| A jelen elemzésben csak ezek az esetek vesznek részt.                          |          |         |         |         |
| SPEARMAN-FÉLE RANGKORRELÁCIÓS MÁTRIX                                           |          |         |         |         |
| (Jelölés: +: $p < 0,10$ *: $p < 0,05$ **: $p < 0,01$ )                         |          |         |         |         |
| Változó                                                                        | Anyasúly | Apasúly | Anyamag | Apamag  |
| Tsúly0                                                                         | 0,273**  | 0,104   | 0,129   | 0,028   |
| Tsúly10                                                                        | 0,289**  | 0,377** | 0,126   | 0,193+  |
| Thossz0                                                                        | 0,180+   | 0,190+  | 0,158   | -0,038  |
| Tmag10                                                                         | 0,090    | 0,511** | 0,434** | 0,487** |

### 12.4. táblázat. Kendall-féle tau-b rangkorrelációs mátrix készítése ROPstatban

|                                  |
|----------------------------------|
| Rangkorrelációs mátrix készítése |
|----------------------------------|

A beolvasott összes eset száma: 100

#### KENDALL-FÉLE RANGKORRELÁCIÓS MÁTRIX

(Jelölés: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$ )

| Változó | Anyasúly | Apasúly | Anyamag | Apamag  |
|---------|----------|---------|---------|---------|
| Tsúly0  | 0,187**  | 0,076   | 0,081   | 0,019   |
| Tsúly10 | 0,195**  | 0,261** | 0,083   | 0,127+  |
| Thossz0 | 0,123+   | 0,125+  | 0,110   | -0,029  |
| Tmag10  | 0,057    | 0,368** | 0,290** | 0,341** |

A 12.3. és a 12.4. táblázatból azt olvashatjuk ki, hogy bár a két rangkorrelációs együttható értékszintje jól érzékelhetően különbözik (a Spearman-féle értékek itt minden esetben nagyobb abszolút értékűek), a szignifikanciák mintázata csaknem teljesen ugyanúgy néz ki.

2. A Kendall-féle tau és tau-b elméletileg jól értelmezhető monotonitási együttható. Ezért előfordulhat, hogy szükség van az elméleti értékek intervallumbecslésére is. Ezt a ROPstat ugyanazon rutinja ('Korreláció, egyszerű regresszió') segítségével lehet elkészíteni, mint a Pearson-féle lineáris korrelációs együttható esetében (vö. 11.8.2. alpont). Ehhez csupán egy dolgot kell még megtenni: a rutin menüablakában külön kérni kell a monotonitási mutatók kiszámítását, mert alapértelmezésben a program nem számítja ki a Kendall-féle együtthatókat. Ezt az elemzést a 12.3. példában említett vizsgálat *h4-preferencia* és *iskolázottság* változójával végeztük el, azzal a csekély eltéréssel, hogy az iskolai végzettséget most egy finomabb skála, a *végzett osztályok száma* segítségével mértük. Az eredmények a 12.5. táblázatban láthatók.

**12.5. táblázat.** A Kendall-féle tau és tau-b monotonitási együtthatók kiszámítása és az elméleti érték intervallumbecslése a ROPstat 'Korreláció, egyszerű regresszió' rutinjában

#### Korreláció, lineáris regresszió

A beolvasott összes eset száma: 277

Jelölés:

r: Pearson-féle lineáris korrelációs együttható

Rpb: Wilcoxon-féle robusztus korrelációs együttható

Tau-b: Kendall-féle tau-b monotonitási együttható

Tau: Kendall-féle tau monotonitási együttható

p\_pos: Pozitív együttjárás (konkordancia) százaléka

p\_neg: Negatív együttjárás (diszkordancia) százaléka

Szignifikancia: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$  \*\*\*:  $p < 0,001$ ;

X = h4, Y = Isk

A mind X, mind Y esetében érvényes értékkel rendelkező esetek száma: N = 277

|    | Változó | Átlag | Var.  | Min. | Max. | Regressziós egyenlet |
|----|---------|-------|-------|------|------|----------------------|
| X: | h4      | 1,834 | 1,610 | 0    | 4    | $x = 3,589 - 0,150y$ |
| Y: | Isk     | 11,74 | 13,01 | 3    | 18   | $y = 13,95 - 1,209x$ |

X: Reziduális variancia (Hibavariancia): MSerr(X) = 1,319

Y: Reziduális variancia (Hibavariancia): MSerr(Y) = 10,66

Korrelációk és 95%-os intervallumbecslés az elméleti értékre:

|         |           |              |                        |               |               |
|---------|-----------|--------------|------------------------|---------------|---------------|
| r =     | -0,425*** | (p = 0,0000) | C95 = (-0,517; -0,324) |               |               |
| Rpb =   | -0,423*** | (p = 0,0000) | C95 = (-0,514; -0,321) |               |               |
| tau-b = | -0,344*** | (p = 0,0000) | C95 = (-0,437; -0,252) |               |               |
| tau =   | -0,280*** | (p = 0,0000) | C95 = (-0,359; -0,201) | p_poz = 19,2% | p_neg = 47,2% |

A Kendall-féle monotonitási együtthatók megint kisebbek, mint a többi (Pearson- és Wilcoxon-féle) korreláció, de ennek nem kell különösebb jelentőséget tulajdonítani. Egyszerűen arról van szó, hogy a Kendall-féle együtthatók más skálán mérnek, mint *r* és a vele rokon korrelációk. Nagyságszint tekintetében mindig csak egyazon típusú kapcsolati mutatók vethetők össze egymással. A szignifikancia mértéke tekintetében azonban itt sincs különbség: mind a

negatív lineáris, mind a monoton fogyó kapcsolat 1%-os szinten szignifikáns, tehát 99%-os megbízhatósággal állíthatjuk, hogy az iskolai végzettség negatív kapcsolatban van a h4 Szondi-kép preferálásával (vö. 12.3. példa). Figyeljük meg, hogy a program kiszámítja a konkordanciák és diszkordanciák mintabeli arányát is (19,2, illetve 47,2%), ami jól mutatja ebben az esetben a negatív kapcsolat dominanciáját.

3. Ha X és Y értékskálája csak néhány különböző értékből áll, vagy övezetek segítségével kevés számú kategóriára redukáljuk a vizsgált változók értékskáláját, akkor a köztük lévő monoton összefüggés elemzéséhez célszerű a 'Változók kapcsolatának vizsgálata' programcsoport 'Diszkrét változók kapcsolatvizsgálata' rutinját használni. Itt ugyanúgy lehet az elemzendő változókat kijelölni, mint a fenti 1. és 2. pontban említett rutinok esetében. A monotonitási együtthatók kiszámítását itt is külön kérni kell. Szemléltetésképpen ismét a 12.3. példa *h4-preferencia* és *iskolázottság* változójának kapcsolatát elemezzük. Az eredményeket a 12.6. táblázat tartalmazza.

## 12.6. táblázat. Monotonitási együtthatók kiszámítása a ROPstat 'Diszkrét változók kapcsolatvizsgálata' rutinjában

|                                                                                              |                       |    |     |    |    |          |
|----------------------------------------------------------------------------------------------|-----------------------|----|-----|----|----|----------|
| Diszkrét változók kapcsolatvizsgálata                                                        |                       |    |     |    |    |          |
| Jelölés: +: $p < 0,10$ *: $p < 0,05$ **: $p < 0,01$ ***: $p < 0,001$                         |                       |    |     |    |    |          |
| A beolvasott összes eset száma: 277                                                          |                       |    |     |    |    |          |
| KÉTSZEMPONTOS GYAKORISÁGI TÁBLÁZAT                                                           |                       |    |     |    |    |          |
| Sorváltozó: Isk (Grades completed)                                                           |                       |    |     |    |    |          |
| Oszlopváltozó: h4                                                                            |                       |    |     |    |    |          |
|                                                                                              | A h4 változó értékei: |    |     |    |    |          |
| Isk                                                                                          | 0                     | 1  | 2   | 3  | 4  | Összesen |
| alsófok                                                                                      | 9                     | 7  | 48  | 19 | 31 | 114      |
| középfok                                                                                     | 20                    | 26 | 34  | 7  | 5  | 92       |
| felsőfok                                                                                     | 25                    | 15 | 22  | 6  | 3  | 71       |
| Összesen                                                                                     | 54                    | 48 | 104 | 32 | 39 | 277      |
| KAPCSOLATI MUTATÓK                                                                           |                       |    |     |    |    |          |
| Cramér-féle kontingencia-együttható, $V = 0,334$                                             |                       |    |     |    |    |          |
| Monotonitási együtthatók                                                                     |                       |    |     |    |    |          |
| - Kendall-féle gamma monotonitási együttható: $\gamma = -0,507$                              |                       |    |     |    |    |          |
| - Kendall-féle tau-b monotonitási együttható: $\tau_b = -0,367$                              |                       |    |     |    |    |          |
| - Oszlopváltozó függése a sorváltozótól: Somers-féle $D(\text{oszlop} \text{sor}) = -0,395$  |                       |    |     |    |    |          |
| - Sorváltozó függése az oszlopváltozótól: Somers-féle $D(\text{sor} \text{oszlop}) = -0,341$ |                       |    |     |    |    |          |
| FÜGGETLENSÉG TESZTELÉSE                                                                      |                       |    |     |    |    |          |
| Khi-négyzet-próba ( $f = 8$ ): $\text{Khi}^2 = 61,842$ ( $p = 0,0000$ )***                   |                       |    |     |    |    |          |
| MONOTONITÁS TESZTELÉSE (Kendall-féle gamma alapján): $z = -7,996$ ( $p = 0,0000$ )***        |                       |    |     |    |    |          |

Ezekhez nem kell különösebb kommentárt fűznünk. Összevetve őket a 12.3. példában kiszámított értékekkel, jól láthatók az egyezések. Ez a rutin nemcsak ordinális, hanem tetszőleges nominális változók kapcsolatának vizsgálatára is alkalmas. A Cramér-féle kontingencia-együttható ilyen esetben is alkalmazható általános kapcsolati mutató, két tetszőleges diszkrét változó függetlensége pedig szokásosan a khi-négyzet-próbával tesztelhető. Az eredménylista az ezzel kapcsolatos számítások eredményeit is tartalmazza, részletesebben azonban majd csak a 17. fejezetben foglalkozunk a nominális változókat is felölelő diszkrét változók kapcsolatával (lásd 17.2. alfejezet).

### 13.5. Több független minta összehasonlítása paraméteres eljárásokkal a ROPstatban

Kettőnél több független minta pontosan ugyanúgy hasonlítható össze a ROPstat segítségével, mint két független minta, vagyis a 'Csoportok és változók összehasonlítása' programcsoport 'Független minták (csoportok) egyszempontos összehasonlítása' rutinja segítségével. A futtatáshoz szükséges lépések ugyanazok, mint a 9.8. alfejezetben, ezért ezeket most nem részletezzük. A program onnan veszi észre, hogy kettő vagy pedig kettőnél több csoport van, hogy azonosítja a csoportosító változóval kapcsolatban definiált csoportok számát, s ennek megfelelően választja meg a szükséges statisztikai módszereket is. Például intervallum-skálájú függő változó esetén, ha a program két csoportot talál, akkor kétmintás t-próbát és robusztus változatait, több csoport esetén pedig a VA-t és annak robusztus változatait hajtja végre. Kontrasztokat és kontrollcsoporttal való összehasonlítást a ROPstat jelenlegi változata nem tud végrehajtani, ezzel kapcsolatban a könyvünk első kiadásában ismertetett MiniStat programcsomagot ajánljuk az olvasó figyelmébe (vö. Vargha, 2000, A. melléklet, 2.1.4. és 6.3. alpont).

Szemléltetésképpen a 13.1. ábra B) diagramján látható RI = 'Átlagos Ro-válaszidő' változó szerint összehasonlítottuk az *Egészséges*, a *Személyiségzavar* és a *Holocaust* mintát. 10%-os trimmelést beállítva, trimmeléses elemzést is előírtunk a program számára. A ROPstat eredménylistáját a 13.10. táblázat tartalmazza.

A 13.10. táblázat eredménylistájához a következő megjegyzéseket fűzzük.

1. A program először az elméleti szóródási mutatókat hasonlítja össze az O'Brien- és a Levene-próbával. Egyik sem jelez tendenciát a különbségre, így nem cáfolják a VA alkalmazásának jogosságát. Persze ilyen kis minták esetén igen nagyok kellene lennie a szórások közti különbségnek ahhoz, hogy a szóráshomogenitás hipotézise elutasítható legyen. A VA érvényességét erősíti azonban a mintaelemszámok egyenlősége, mely növeli a VA robusztuságát mind a normalitás, mind a szóráshomogenitás feltételének megsértésével szemben.

2. A hagyományos VA F-értéke:  $F = 2,316$ . Mivel az F-érték mellett feltüntetett p-érték ( $p = 0,1132$ ) nagyobb 0,10-nél, a VA még 10%-os szinten sem szignifikáns. A VA összefoglaló táblázata helytakarékoság miatt nem szerepel a ROPstat eredménylistáján. Ugyanakkor ennek legfontosabb elemeit a ROPstat mind közli. A VA F-mennyisége után zárójelben található a próbastatisztika két szabadságfoka (esetünkben  $f_1 = 2$  és  $f_2 = 36$ ), melyek egyben a hatásvariancia és a hibavariancia szabadságfokai is.

Ez utóbbi mennyiségek a VA F-értéke alatti sorban foglalnak helyet ( $Vark = 0.5232$  és  $Varb = 0.2259$ ). Az ez alatti sorban pedig a nemlineáris korrelációs együttható látható ( $e = 0.338$ ), mely a csoportosító változó gyenge hatását mutatja a függő változóra.

3. A robusztus VA-k közül mind a Welch-próba, mind a James-próba 10%-os szinten jelez, s ennek alapján tendenciát állapíthatunk meg a csoportok különbözőségére. Ez a mintaszórások kb. 1:2 arányú különbségének köszönhető, amelynek a torzító hatását a robusztus eljárások kiszűrik.

4. Mivel a hagyományos VA nem szignifikáns, nem lepődhetünk meg azon, hogy az átlagok Tukey–Kramer-féle páronkénti összehasonlításai sem jeleznek különbséget az átlagok között.

#### 13.10. táblázat. Több független minta összehasonlítása ROPstat segítségével

|                                                 |
|-------------------------------------------------|
| Független minták egyszempontos összehasonlítása |
|-------------------------------------------------|

A beolvasott összes eset száma: 39

Csoportosító változó: Csoport

Jelölés: +:  $p < 0,10$  \*:  $p < 0,05$  \*\*:  $p < 0,01$  \*\*\*:  $p < 0,001$

FÜGGŐ VÁLTOZÓ: RI

Csoportonkénti alapstatisztikák

| Index | Csoport   | Esetek | Átlag | Szórás | Min.  | Max.  | Ferdeség | Csúcsosság |
|-------|-----------|--------|-------|--------|-------|-------|----------|------------|
| 1     | Egészség  | 13     | 0,632 | 0,281  | 0,190 | 1,360 | 1,207+   | 3,408*     |
| 2     | SzemZav   | 13     | 1,018 | 0,547  | 0,440 | 2,590 | 2,160**  | 5,828***   |
| 3     | Holocaust | 13     | 0,917 | 0,548  | 0,310 | 2,270 | 1,396*   | 2,076      |

A Ferdeség és a Csúcsosság szignifikanciája a normalitás sérülését jelzi.

Elméleti szórások egyenlőségének tesztelése

- O'Brien-próba:  $F(2; 36,0) = 0,696$  ( $p = 0,5050$ )
- Levene-próba:  $F(2; 36,0) = 1,481$  ( $p = 0,2409$ )

Elméleti átlagok egyenlőségének tesztelése

Hagyományos eljárás, amely feltételezi a szóráshomogenitást:

- Varianciaanalízis:  $F(2; 36) = 2,316$  ( $p = 0,1132$ )
- Hatásvariancia = 0,5232, Hibavariancia = 0,2259
- Korrelációs hányados (nemlineáris korrel. együtttható):  $e = 0,338$

Robusztus eljárások, amelyeknél nem szükséges a szóráshomogenitás:

- Robusztus Welch-féle varianciaanalízis:  $W(2; 21,5) = 3,229$  ( $p = 0,0594$ )+
- James-próba:  $U = 6,657$  ( $p < 0,10$ )+
- Brown-Forsythe-próba:  $BF(2; 30) = 2,316$  ( $p = 0,1161$ )

Átlagok Tukey-Kramer-féle páronkénti összehasonlítása ( $k = 3$ ,  $df = 36$ ):

T12= 2,93 T13= 2,16 T23= 0,77

Átlagok Games-Howell-féle páronkénti összehasonlítása

(elméleti szórások különbözhetnek, zárójelben a szabadságfokok):

T12(3; 18)= 3,21+ T13(3; 18)= 2,36 T23(3; 24)= 0,67

A trimmelt átlagra vonatkozó elemzések (max. trimmelés = 10%)

| Index | Csoport   | n  | n_trim | Trim. átlag | Winsor. szórás | Trim. min. | Trim. max. |
|-------|-----------|----|--------|-------------|----------------|------------|------------|
| 1     | Egészség  | 13 | 11     | 0,605       | 0,158          | 0,390      | 0,810      |
| 2     | SzemZav   | 13 | 11     | 0,928       | 0,302          | 0,550      | 1,440      |
| 3     | Holocaust | 13 | 11     | 0,849       | 0,424          | 0,420      | 1,640      |

Elméleti trimmelt átlagok egyenlőségének tesztelése

- $Ft(2; 17,2) = 4,625$  ( $p = 0,0247$ )\*

Trimmelt átlagok Games-Howell-féle páronkénti összehasonlítása

(elméleti szórások különbözhetnek, zárójelben a szabadságfokok):

T12(3; 15)= 4,44+ T13(3; 13)= 2,53 T23(3; 18)= 0,71

5. A trimmelés során a szélsőséges RI-értékek kimaradtak az elemzésből. Ennek hatására a szórás lecsökkent (lásd a táblázatban a winsorizált mintaszórásokat), így az átlagok közti különbségek felértékelődtek. Az eredmény: a trimmelt átlagok VA-ja már 5%-os szinten szignifikáns ( $Ft(2; 17,2) = 4,625$ ,  $p = 0,0247$ ).

6. A trimmelt átlagok páronkénti összehasonlításai azt mutatják, hogy a három csoport közül a legnagyobb eltérés az Egészséges és a Személyiségzavar RI-szintje között van – az utóbbi csoport tagjainak egy Rorschach-válasz megfogalmazásához szignifikánsan több időre van szükségük, mint az előbbieknél.

## 14.4. Több összetartozó minta átlagának összehasonlítása a ROPstatban

Kettőnél több összetartozó minta pontosan ugyanúgy hasonlítható össze a ROPstat segítségével, mint két összetartozó minta, vagyis a 'Csoportok és változók összehasonlítása' program-csoport 'Összetartozó minták (változók) egyszempontos összehasonlítása' rutinja segítségével. A futtatáshoz szükséges lépések ugyanazok, mint a 8.6. alfejezetben, ezért ezeket itt most nem részletezzük. A program a kijelölt függő változók számának megfelelően választja meg a szükséges statisztikai módszereket. Például intervallum-skálájú függő változó esetén, ha a program két változót talál, akkor összetartozó mintás t-próbát és annak robusztus paraméteres, valamint nemparaméteres változatait, több változó esetén pedig egyszempontos összetartozó mintás VA-t és annak robusztus paraméteres, illetve rangsorolós változatait hajtja végre. Ez utóbbi módszereket a 15. fejezetben ismertetjük majd.

**14.5. táblázat.** A saját falujához vagy városához (Saját), Vajdasághoz (Vajdaság), Szerbiához (Szerbia), a régi Jugoszláviához (RégiJugó), a mai Jugoszláviához (MaiJugó) és a mai Magyarországhoz (MagyarO.) való kötődés összehasonlítása ROPstat segítségével egy 1996-ban lefolytatott vajdasági szociolingvisztikai vizsgálat adatai alapján. A kötődést egy ötfokú skálán mérték (1 = egyáltalán nem kötődik, 5 = nagyon erősen kötődik)

|                                                                                        |          |       |        |                |                |
|----------------------------------------------------------------------------------------|----------|-------|--------|----------------|----------------|
| Összetartozó minták egyszempontos összehasonlítása                                     |          |       |        |                |                |
| A beolvasott összes eset száma: 144                                                    |          |       |        |                |                |
| Érvényes esetek száma: 97                                                              |          |       |        |                |                |
| Jelölés: +: $p < 0,10$ *: $p < 0,05$ **: $p < 0,01$ ***: $p < 0,001$                   |          |       |        |                |                |
| Elméleti átlagok egyenlőségének tesztelése:                                            |          |       |        |                |                |
| - Varianciaanalízis (VA): $F(5,480) = 75,733$ ( $p = 0,0000$ )***                      |          |       |        |                |                |
| (Hatásvariancia = 85,4203, Hibavariancia = 1,1279)                                     |          |       |        |                |                |
| - Robusztus VA szabadságfok-korrekcióval                                               |          |       |        |                |                |
| Geisser-Greenhouse (epszilon = 0,802): $F(4,0; 385,0) = 75,733$ ( $p = 0,0000$ )***    |          |       |        |                |                |
| Huynh-Feldt (epszilon = 0,841): $F(4,2; 403,9) = 75,733$ ( $p = 0,0000$ )***           |          |       |        |                |                |
| Elméleti trimmelt átlagok egyenlőségének tesztelése (valódi trimmelés: 9%):            |          |       |        |                |                |
| - Robusztus trimmelt VA szabadságfok-korrekcióval                                      |          |       |        |                |                |
| Geisser-Greenhouse (epszilon = 0,767): $F(3,8; 299,3) = 92,998$ ( $p = 0,0000$ )***    |          |       |        |                |                |
| Huynh-Feldt (epszilon = 0,803): $F(4,0; 313,3) = 92,998$ ( $p = 0,0000$ )***           |          |       |        |                |                |
| Index                                                                                  | Változó  | Átlag | Szórás | Trimmelt átlag | Winsor. szórás |
| 1.                                                                                     | Saját    | 4,258 | 1,034  | 4,456          | 0,768          |
| 2.                                                                                     | Vajdaság | 4,124 | 1,043  | 4,278          | 0,823          |
| 3.                                                                                     | Szerbia  | 2,247 | 1,191  | 2,127          | 1,086          |
| 4.                                                                                     | RégiJugó | 3,794 | 1,290  | 3,975          | 1,108          |
| 5.                                                                                     | MaiJugó  | 2,227 | 1,221  | 2,076          | 1,074          |
| 6.                                                                                     | MagyarO. | 2,722 | 1,248  | 2,658          | 1,102          |
| Átlagok Tukey-féle páronkénti összehasonlítása ( $k = 6$ , $df = 480$ ):               |          |       |        |                |                |
| T12= 1,24 T13= 18,64** T14= 4,30* T15= 18,83** T16= 14,24** T23= 17,40** T24= 3,06     |          |       |        |                |                |
| T25= 17,59** T26= 13,00** T34= 14,34** T35= 0,19 T36= 4,40* T45= 14,53** T46= 9,94**   |          |       |        |                |                |
| T56= 4,59*                                                                             |          |       |        |                |                |
| Trimmelt átlagok Tukey-féle páronkénti összehasonlítása ( $k = 6$ , $df = 313$ ):      |          |       |        |                |                |
| T12= 1,72 T13= 22,67** T14= 4,68* T15= 23,16** T16= 17,50** T23= 20,95** T24= 2,96     |          |       |        |                |                |
| T25= 21,44** T26= 15,77** T34= 17,99** T35= 0,49 T36= 5,17** T45= 18,48** T46= 12,81** |          |       |        |                |                |
| T56= 5,67**                                                                            |          |       |        |                |                |

Illusztrációként a 8.6. alfejezetben már említett, 1996-ban lefolytatott vajdasági szociolingvisztikai vizsgálat (Göncz, 1999, 77. o.) adatainak ROPstat segítségével elvégzett elemzését mutatjuk be. Ebben a vizsgálatban 144 véletlenszerűen kiválasztott, Vajdaságban élő magyar nemzetiségű felnőtt személyt többek között arról is megkérdeztek, hogy mennyire kötődik saját falujához vagy városához, Vajdasághoz, Szerbiához, a régi Jugoszláviához, a mai Jugoszláviához, illetve Magyarországhoz. A kötődés mértékét ötfokú skálán lehetett megadni,

melyen az 1-es érték azt jelentette, hogy egyáltalán nem kötődik, az 5-ös pedig azt, hogy igen erősen kötődik. Ezekre a kérdésekre a 144 közül 97 személy adott érvényes választ. Az elemzések eredményeit – a nemparaméteres eljárások eredményeit most mellőzve – a 14.5. táblázat tartalmazza. Ezekhez a következő megjegyzéseket fűzzük.

1. A hagyományos VA F-értéke:  $F = 75,733$ . Mivel ezen eredmény mellett 0,001-nél kisebb p-érték és három csillag található, a VA 0,1%-os szinten szignifikáns. A VA összefoglaló táblázata helytakarékoság miatt itt sem szerepel az eredménylistáján, de ennek legfontosabb elemeit a ROPstat közli. A VA F-mennyisége után zárójelben található a próbastatisztika két szabadságfoka (esetünkben  $f_1 = 5$  és  $f_2 = 480$ ), melyek egyben a hatásvariancia és a hibavariancia szabadságfokai is. Ez utóbbi mennyiségek a VA F-értéke alatti sorban foglalnak helyet ( $Var_k = 85,4203$  és  $Var_e = 1,1279$ ).

2. Minden próba (hagyományos VA, Geisser-Greenhouse és Huynh-Feldt-féle szabadságfok-korrektúrák VA, trimmelt VA) egybehangzóan azt támasztja alá, hogy a vajdasági magyarok kötődése a felsorolt régiókkal, illetve országokkal szemben eltérő szintű. A részletes eredmények a mintaátlagok páronkénti összehasonlításaiból következtethetők ki.

3. Mind az eredeti, mind a trimmelt átlagok összehasonlítása alapján egyöntetűen azt láthatjuk, hogy a vajdasági magyarok leginkább saját településükhöz, illetve a Vajdasághoz kötődnek, mert ezek átlagai szignifikánsan nagyobbak, mint a többihez tartozó átlag. Például T13, T14, T15 és T16 szignifikanciája a saját (1. számú változó) fölényét jelzi a Szerbia (3. változó), RégiJugó (4. változó), MaiJugó (5. változó) és MagyarO. (6. változó) változóval szemben. Egy kivétel van csak: a Vajdaság és a RégiJugó kötődési átlaga közti különbség nem szignifikáns ( $T24 = 3,06$ ,  $p > 0,10$ ).

4. A RégiJugó (4. változó) átlaga szignifikánsan nagyobb, mint bármely jelenlegi országé (Szerbia, MaiJugó, MagyarO.).

5. A Magyarországhoz való kötődés csak a Szerbiához és a mai Jugoszláviához való kötődésnél erősebb.