

Sajtos László – Mitev Ariel

**SPSS KUTATÁSI
ÉS ADATELEMZÉSI
KÉZIKÖNYV**

Alinea Kiadó



2. fejezet

SPSS-ALAPOK

A fejezet céljai

A fejezet célja, hogy az olvasó:

- megismerje az SPSS program felépítését és működését,
- ismerje a programban a különböző nézeteket,
- ismerje az adatbeviteli módszereket.

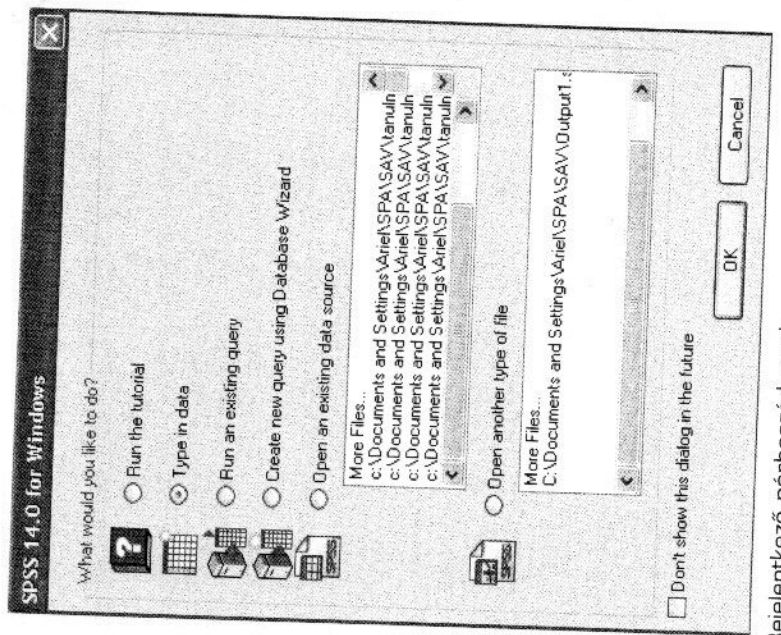
Az SPSS a Statistical Package for the Social Sciences (statisztikai programcsomag társadalomtudományok részére) rövidítése. Az e könyv megírásakor legújabbnak számító, SPSS 14.0 adatbázis-kezelő programmal nagyméretű, összetett adatbázist lehet feldolgozni gyorsan és hatékonyan. Ha ismerjük a program menürendszerét (File, Edit, View, Data menüpontokat), a megjelenő dialógusablakok megfelelő kitöltésével, kipipálásával – programozói jártasság nélkül – írjuk a parancsokat, de lehetőség van saját magunk által kifejlesztett, még nem létező utasítássorozat beírására is. Bármelyik utat választjuk is, a statisztikai eljárások alapos ismeretére és az eredmények pontos elemzésére feltétlenül szükség van, mert a program az értelmezést nem végzi el helyettünk. A fejezetben nem tárgyaljuk az ANALYZE (elemzés) menüt, mivel a különböző elemzési módszerekkel a könyv 3–9. fejezetében foglalkozunk. A többi menü bemutatásánál azt a logikát követjük, hogy a gyakrabban használt pontokat részletesebben, például alátámasztva fejtettük ki, míg a ritkábban használtakat csak röviden, a fejlesség igénye nélkül mutattuk be. Ennek egyrészt terjedelmi okai voltak, másrészt azonban az SPSS program kiváló HELP (segítség) menüvel rendelkezik, ahol az egyes kérdéseknek utána lehet nézni.

2.1. A program felépítése

Az SPSS felépítése és menürendszere sok tekintetben hasonlít a Microsoft Office programcsomag elemeinél megszokotthoz. Az ott rendszeresen használt műveletek közül számos alkalmazható (másolás, kivágás, beillesztés, törlés), azonban körültekintően kell eljárunk, mert néhány esetben még ezek az egyszerű műveletek is eltérően működnek. (Például az SPSS-ben a jól ismert visszavonás parancs csak az utolsó műveletet képes meg nem történné tenni, a beillesztéshez, így figyelmetlen használatával adatvesztést okozhatunk.)

2.2. A program indítása

A program indításakor megjelenik egy párbeszédablak, amely azt kérdezi, hogy mi legyen az első lépés:



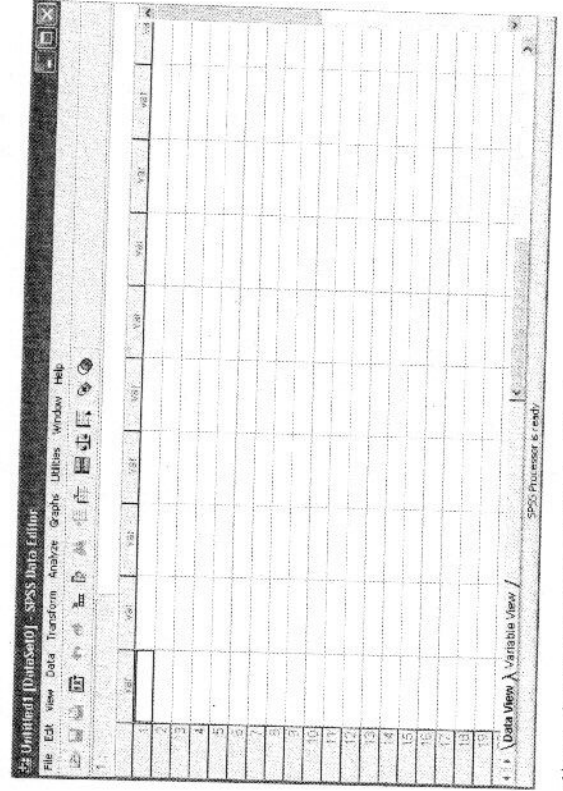
2.1. ábra. Bejelentkező párbeszédpanel

- Run the Tutorial (részletes leírást biztosító oktatóprogram futtatása),
- Type in data (adatok begépelése),
- Run an existing query (egy meglévő lekérdezés futtatása),
- Create new query using Database Wizard (adatok egy másik adatbázisból való beolvasása),
- Open an existing data source (egy korábbi SPSS-adatállomány betöltése),
- Open another type of file (más típusú fájl megnyitása).

Jelöljük meg a TYPE IN DATA-t vagy egyszerűen klikkeljünk a Cancel gombra, hogy egy üres adatbázist kapjunk, ahol rögzíteni tudjuk a változókat és az azokhoz tartozó adatokat a Data Editor ablakban.

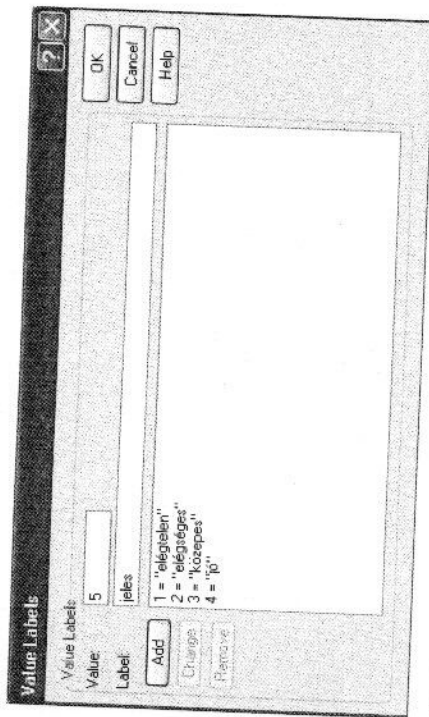
2.3. Data Editor (Adatszerkesztő)

Adatbázisunk feltöltött sorokból, úgynevezett rekordokból (vagy esetekből - case) és változókból, azaz oszlopokból áll. Az adatszerkesztő (Data Editor) két lapból áll: ezek DATA VIEW (2.2. ábra) és a VARIABLE VIEW (2.3. ábra), amelyeket az ablak bal alsó sarkában található fülére kattintva, illetve a CTRL+T billentyűkombinációval lehet váltogatni.



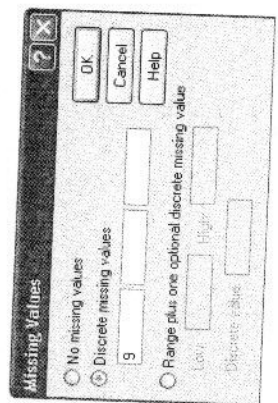
2.2. ábra. Adatbeviteli nézet (Data View)

A változó értékéhez (VALUE) jelentést kell hozzárendelni (VALUE LABEL), amelyet az „Add” gombbal lehet nyugtázni. Ha minden szükséges összetartozó párt meghatározunk, akkor az „OK”-ra kattintunk, s bezárjuk a panelt.



2.5. ábra. Értékcímkék megadása

Missing (hiányzó érték): Itt a nem kielégítő válaszokhoz rendelünk egy-az adat hiányzását mutató értéket. Például, ha egy kérdésnél a tanuló osztályzatát szeretnénk megtudni, akkor 1-5-ig jegyek valamelyiket várjuk eredményül. Ha a tanuló ezektől eltérő jegyet ad meg, például 6-ost, akkor ezt hiányzó értéknek minősíthetjük. Ha a hiányzó adatok kezelése nem oldjuk meg, hibák adódhatnak. (Például a program beleszámítja a 0-át is az átlagba, amely azonban valóban nem létező értéket, véleményt takar.) A különböző okból hiányzó adatokat meg kell jelölnünk ahhoz, hogy a statisztikai feldolgozás pontos és reális legyen. Olyan értéket kell kiválasztani, amely az amúgy érvényes adatok között nem fordulhat elő.



2.6. ábra. Missing values párbeszédablak. A 9-es mint hiányzóérték-jelölő meghatározása

A MISSING VALUES gomb megnyomása után megjelenik egy párbeszédablak (2.6. ábra). A kapott párbeszédablakban négy lehetőség közül választhatunk: *No missing values*: nem adunk meg hiányzó értéket. Ilyenkor a program azokra a helyekre, ahova nem írunk be értéket, automatikusan egy pontot tesz (System Missing). Ez az alapbeállítás, s ha nincs különösebb ok az eltérés-

Discrete missing values: itt néhány (legfeljebb három) egyedi hiányzóadat-kódot adhatunk meg (pl. hétfokozatú skála esetén ez lehet a 9-es). *Range plus one optional discrete missing value*: Itt megadhatunk egy számtartományt vagy az előző kettő kombinációját (egy tartományt és egy különálló értéket).

Columns (oszlopok): a DATA VIEW ablakban az oszlop szélességének mértéke, amelyet nem lehet kisebbre állítani, mint amilyen hosszú a változó neve.

Align: a DATA VIEW ablakban található adatbázisban lévő cellák tartalmának igazítási módja (balra, jobbra vagy középre zárt).

Measure (mérési skála): A kutatóknak célszerű megadni a skála típusát, mivel ez segít a további elemzésben (2.7. ábra). A mérési skálák közül itt három lehet beállítani: SCALE (metrikus, azaz intervallum- vagy arányskála), ORDINAL (sorrendi), NOMINAL (névleges). Az SPSS az intervallum- és az arányskálát nem különbözteti meg, hanem átfogó néven scale-lel jelöli. A skála típusa ezek után grafikusan jelenik meg az SPSS különféle párbeszédpaneleiben.




2.7. ábra. Mérési skálák grafikus megjelenítése

2.4. Adatbevitel

A változók meghatározása és paramétereik megadása után az esetek regisztrálása következik. Ennek két alapvető módja létezik: az elsődleges, illetve a másodlagos adatbevitel. Az elsődleges adatbevitel egyszerű begépeléssel történik a kódolási útmutató és a definiálandó változók ismerete mellett, míg a másodlagos rögzítésnél már létező adatbázisok (dBase, Excel vagy Access) importálásáról van szó.

2.4.1 Elsődleges adatbevitel

Az elsődleges adatbevitelnél a változók definiálása az első lépés (INSERT VARIABLE), majd a rekordok (adatsorok) begépelése következik (INSERT CASES). A Data menüpont alatt megtalálható az INSERT VARIABLE és az INSERT CASES, amelyeket a sorok vagy oszlopok elején történő jobb egérgéppel vagy az  ikonokkal szintén elő lehet hívni. Ha a VARIABLE VIEW ablakban alkalmazzuk az Insert VARIABLE menüpontot, akkor már csak a változó paramétereit kell beállítani, ha azonban a DATA VIEW ablakban tesszük mindezt, akkor az új változót a legutolsó oszlopba illeszti a program, s a változó paramétereinek definiálásához az oszlop tetején dupla kattintást kell alkalmazni.

Gondoljunk magunkat egy iskolaigazgató helyébe, akinek a következő feladatot kell megoldania: új tanulók jönnek a középiskolába, s szeretné az adataikat (kód, előző év végi informatika- és matematikajegyük, s szereténe az adataikat le az első nyelvvizsgájukat, milyen típusú osztályba jártak) gépre vinni. Először célszerű definiálni a változókat a VARIABLE VIEW ablakban, s csak ezután érdemes bevinni az adatokat a DATA VIEW ablakban.

Az SPSS megnyitása után váltunk VARIABLE VIEW nézetre, s definiáljuk a szükséges változókat (2.8. ábra).

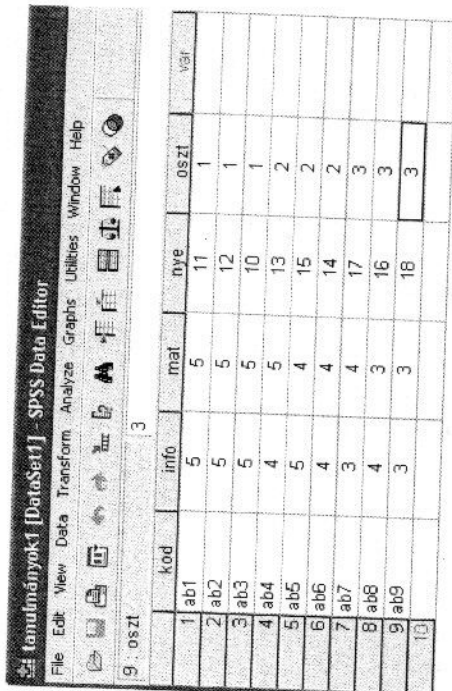
Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1 kod	String	4	0	halgatói kód	None	None	8	Left	Nominal
2 info	Numeric	2	0	informatika jegy	{1, elégtelen}	None	8	Center	Scale
3 mat	Numeric	2	0	matematika jegy	{1, elégtelen}	None	8	Center	Scale
4 nye	Numeric	2	0	hány éves korban tette le None	None	None	8	Center	Scale
5 oszt	Numeric	1	0	melyik osztályba jár	{1, A}	None	8	Center	Nominal

2.8. ábra. Változók definiálása a Variable View nézetben

A paraméterek meghatározása az egyes változónál úgy történik, hogy a változó sorában egyesével a cellákra állunk és

- a NAME és a LABEL esetében begépeljük a karaktereket,
- a TYPE, VALUES, MISSING esetében a cella jobb oldalán található gombra történő kattintással megjelenő panelt töltjük ki,
- a többi cella esetében a cellák jobb oldalán található görgetősávot használjuk.
- A „kod” szöveges változó (STRING), a többi numerikus.
- Nem szeretnénk tizedesjegyeket megjeleníteni (DECIMALS: 0).
- A VALUES parancsnál az informatika- és matematikajegyeknél az 1-es érték jelentse az elégtelent, az 5-ös a jelest, az osztály típusánál (oszt) az 1, 2, 3 értékekhez rendeljük hozzá az A, B, C karaktereket.
- Nem definiáljuk külön a hiányzó értékeket (MISSING: NONE).
- A „kod” és az „oszt” névleges skálák (NOMINAL), mivel az értékek csupán megkülönböztetésre vagy csoportosításra szolgálnak. A többi pedig metrikus (SCALE).

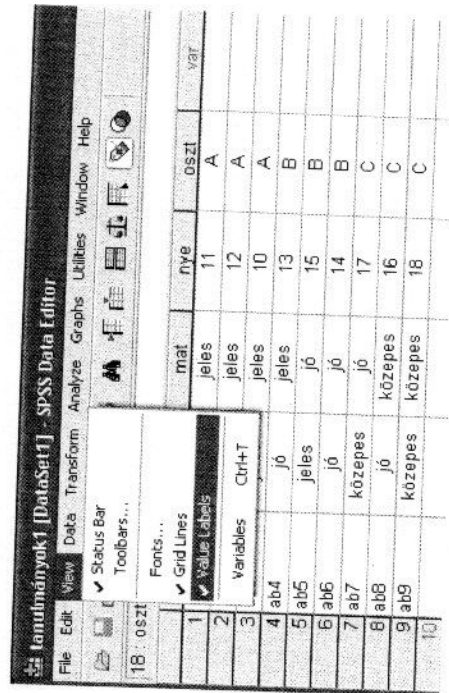
Ezután váltunk Data View nézetre, a 2.9. ábra alapján töltjük fel esetekkel a már definiált változókat.



	kod	info	mat	nye	oszt
1	ab1	5	5	11	1
2	ab2	5	5	12	1
3	ab3	5	5	10	1
4	ab4	4	5	13	2
5	ab5	5	4	15	2
6	ab6	4	4	14	2
7	ab7	3	4	17	3
8	ab8	4	3	16	3
9	ab9	3	3	18	3
10					

2.9. ábra. Esetek (case) megadása

Amennyiben a VIEW menüpontnál pipát látunk a VALUE LABELS opciónál (2.10. ábra) vagy rákattintunk a címkeré, a változók értékeihez hozzárendelt jelentést (az összetartozó érték-jelentés párokat a 2.3. ábrán láttuk) mutatja az SPSS, tehát a 2.9. és 2.10. ábra csak külső formában különböznek.

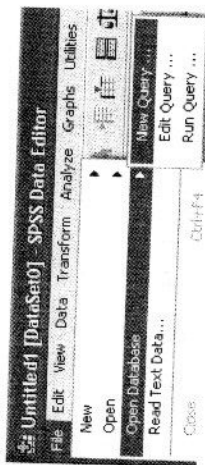


	mat	nye	oszt
1	jeles	11	A
2	jeles	12	A
3	jeles	10	A
4	jeles	13	B
5	jeles	15	B
6	jó	14	B
7	jó	17	C
8	közepes	jó	közepes
9	közepes	közepes	közepes
10			

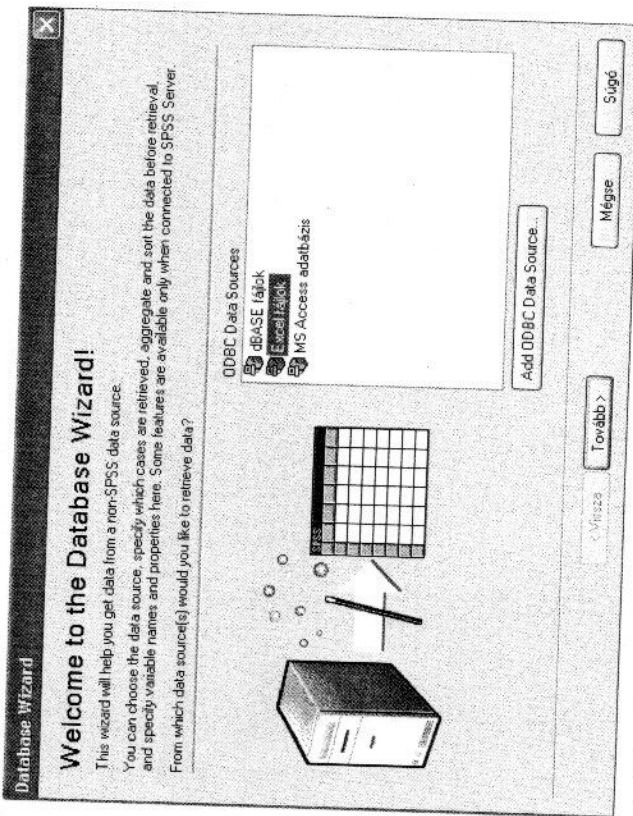
2.10. ábra. Esetek megadása

2.4.2 Másodlagos adatbevitel

Tegyük fel, hogy a mintapéldában szereplő iskolaigazgató az új tanulók adatait egy Excel fájlban megkapja. Ebben az esetben az igazgató élhet a másodlagos adatbevitellel. Első lépésben alkalmazzuk a FILE/OPEN DATABASE/NEW QUERY menüpontot (2.11. ábra), majd válasszuk az „Excel fájlok” lehetőséget (2.12. ábra), hiszen az importálandó adatforrás „xls” kiterjesztésű.

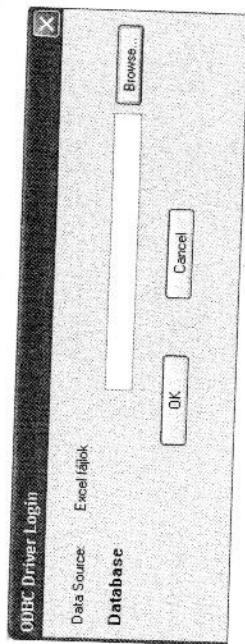


2.11. ábra. Másodlagos adatbevitel



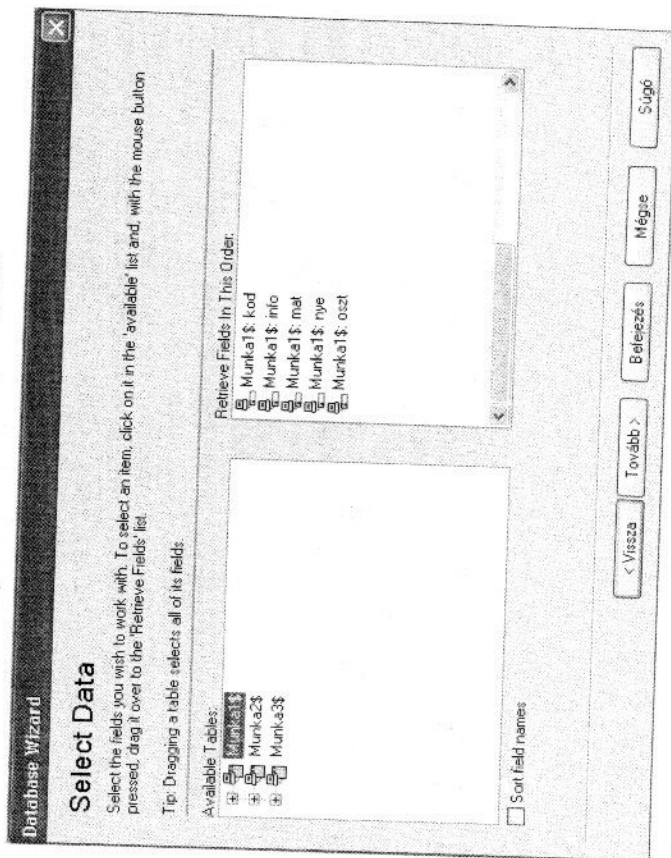
2.12. ábra. Az importálandó fájl típus meghatározása

A „TOVÁBB” gombra kattelve újabb ablakot kapunk, ahol az importálandó fájl elérési útját kell megadni (2.13. ábra). Miután a „BROWSE” gombbal kikerestük az elérési utat, nyomjuk meg az „OK”-t.



2.13. ábra. Az importálandó fájl elérési útjának megadása

Az Excel fájl két munkalapból áll, azonban nekünk csak az „új tanulók adatai” föltre van szükségünk, ezért az egérrel menjünk rá, s amikor a fehér kéz megjelenik, akkor a bal egérgomb folyamatos nyomva tartása mellett húzzuk át a jobb oldali üres ablakba az „új tanulók adatai”-t (2.14. ábra).



2.14. ábra. Az importálandó adatfájl adatainak megadása

Miután az SPSS rendelkezésére áll az importálandó fájl tartalma, felismeri a változókat és az eseteket. Ezután kattintsunk kétszer a további gombra (a köztes ablakokban a változók meghatározására és újrakódolására van lehetőség), és megkapjuk a végeredményt (2.15. ábra).

	kod	info	mat.	nyv.	oszt
1	ab1	5,00	5,00	11,00	1,00
2	ab2	5,00	5,00	12,00	1,00
3	ab3	5,00	5,00	10,00	1,00
4	ab4	4,00	5,00	13,00	2,00
5	ab5	5,00	4,00	15,00	2,00
6	ab6	4,00	4,00	14,00	2,00
7	ab7	3,00	4,00	17,00	3,00
8	ab8	4,00	3,00	16,00	3,00
9	ab9	3,00	3,00	16,00	3,00

2.15. ábra. Kész adatfájl

Tartalmát tekintve ugyanazt az adatbázist kaptuk, mint a 2.9. ábra esetében. A különbség a változók néhány paraméterének megadásában van, amit a 2.16. ábra jól szemléltet:

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1 kod	String	255	0		None	None	50	Left	Nominal
2 info	Numeric	8	2		None	None	8	Right	Scale
3 mat	Numeric	8	2		None	None	8	Right	Scale
4 nyv	Numeric	8	2		None	None	8	Right	Scale
5 oszt	Numeric	8	2		None	None	8	Right	Scale

2.16. ábra

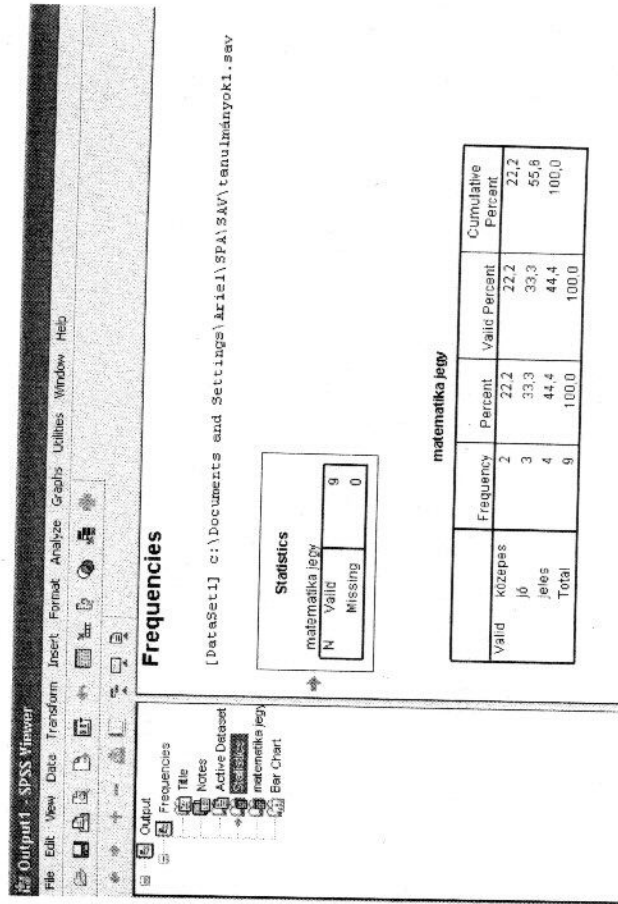
Vegyük észre, hogy az Excelből importált fájlnál több paraméter eltér:

- Ami eredetileg az Excelben oszlopcímként szerepelt (kód, info stb.), azok névvé (NAME) alakultak át.
- A szöveges (STRING) változót a program felismerte, s az alapbeállításként 50-es oszlopszélességet (COLUMNS), balra igazítást (ALIGN) és nominális skála megjelölést kapott.
- A címke (LABEL) üresen maradt, valamint a változó értékei (VALUES) szinten megjelölve.
- Minden numerikus változó SCALE (metrikus) minősítést kapott, amit szintén korrigálni kell.

Megjegyezzük, hogy ugyanez az Excel fájl megnyitható az OPEN FILE/DATA parancsoron keresztül is, ahol az xls kiterjesztésű fájlra rákattintva a program automatikusan beolvassa, és az excel fájl első sora alapján definiálja az SPSS a változókat (Read variable names from the first row of data). Ennek a megoldásnak előnye a gyorsaság, hátránya azonban az, hogy a változókon átalakítás előtt nem változtathatunk.

2.5 Viewer: Eredmények

A Viewer ismertebb neve az Output (kimeneti, statisztikai) ablak, amely az elvégzett számítások eredményét (output) szemlélteti táblázatok, diagramok stb. formájában. (2.17. ábra). Az ablak bal oldalán fa szerkezetben mutatja a program egymás után az elkészített statisztikákat, s egy kis piros nyílal jelzi, hogy éppen melyiket nagyítja ki az ablak jobb oldalán. Az output ablakban előhívott grafikonokat, táblázatokat, diagramokat tovább lehet szerkeszteni, formázni, ha duplán klikkelünk azokra, majd alkalmazzuk az egér jobb gombjának megnyomása után a felkínált lehetőségeket.



2.17. ábra. Viewer ablak

2.6. A menüpontok leírása

A következő részben néhány menüponttal részletesebben foglalkozunk, hogy biztosabb háttérrel kezdhesük meg a konkrét statisztikai elemzések bemutatását, amelyeket az „ANALYZE” menüpontba gyűjtöttek össze.

2.6.1 File menü

Ezen menüben találjuk a fájlkezelő műveleteket.

New: új adatfájl (Data) vagy output fájl (Output) hozhatunk létre.

Open: már létező adatfájl (Data) vagy output állományt (Output) nyithatunk meg.

Open Database: új vagy létező SQL szervezésű adatfájl nyithatunk meg.

Read Text Data: szövegformátumú állomány megnyitása.

Save: az aktív állományt arra a helyre ment, ahonnan azt megnyitottuk. A fájl

előszöri mentésekor ebből a menüpontból a program automatikusan

továbblép a „SAVE AS” menüpontba.

Save As: az aktív állomány mentése megadott helyre, megadott néven és

megadott fájltypusban.

Save All Data: az összes nyitott állomány mentése.

Mark File Read Only: az adatfájl csak olvasható fájlként való megjelölése

(a továbbiakban a fájlban javítani nem lehet).

Rename Dataset: az adatbázis elnevezése, illetve annak megváltoztatása.

A fájlnev mellett az adatbázisnak adhatunk egy külön nevet, amely a fájl-

név után jelenik meg. Ez akkor előnyös, ha egy fájlból (azonos név alatt)

számos verzió létezik.

Display Data File Information: a .sav kiterjesztésű adatfájlokról ad össze-

sített információt egy output ablakban. Ha a Working file-ra megyünk,

akkor az aktuálisan betöltött adatbázis változóiról ad információkat

(2.18. ábra). A változó rövid neve (VARIABLE), pozíciója (POSITION), a címké-

je (LABEL), a mérési szint (MEASUREMENT LEVEL) mellett az oszlopszélesség-

ról és az igazítás módjáról is információkat kaphatunk. External File vá-

lasztásokor egy külső adatbázisról kaphatjuk meg ugyanezeket az infor-

mációkat.

Variable Information							
Variable Kód	Position	Label hailgatói kód	Measurement Level	Column Width	Alignment	Print Format	Write Format
info	1	Informatika	Nominal	8	Left	A4	A4
mat	2	Jegy	Scale	8	Center	F2	F2
nye	3	matematika	Scale	8	Center	F2	F2
	4	a jegy hány éves korban	Scale	8	Center	F2	F2
		tette le az első	Scale	8	Center	F2	F2
oszt	5	nyelvizsgát melyik osztályba jár.	Nominal	8	Center	F1	F1

Variables in the working file

2.18. ábra. Összefoglaló információk az adatbázisról

Variable Values

Value	Label
1	elégtelen
2	elégséges
3	közepes
4	jó
5	jeles
mat	elégtelen
	elégséges
	közepes
	jó
	jeles
oszt	A
	B
	C

2.19. ábra. Összefoglaló információk az adatbázisról

Cache Data: Ennek a funkciónak a futtatása (Cash Now) azt jelenti, hogy az adatokon addig senki nem tud változtatni, amíg be nem fejeztük a munkát rajtuk. Továbbá nagy adatbázisok esetén az adatoknak az adat-szerkesztőben való áttekintése gyorsabbá válik.

Print: az aktív ablak nyomtatási beállításait lehet megadni.

Print Preview: Nyomtatási kép. Nyomtatást megelőzően érdemes ellenőrizni a várható eredményt.

Swith Server: szervergépre történő csatlakozást tesz lehetővé.

Stop Processor: az SPSS számolási műveleteket végző egységének leállítását teszi lehetővé. Hibásan kiadott, nagy számolás- és így időigényű felada-

toknál lehet hasznos.

Recently Used Data: legutóbb használt .sav kiterjesztésű adatfájlok elérését teszi lehetővé.

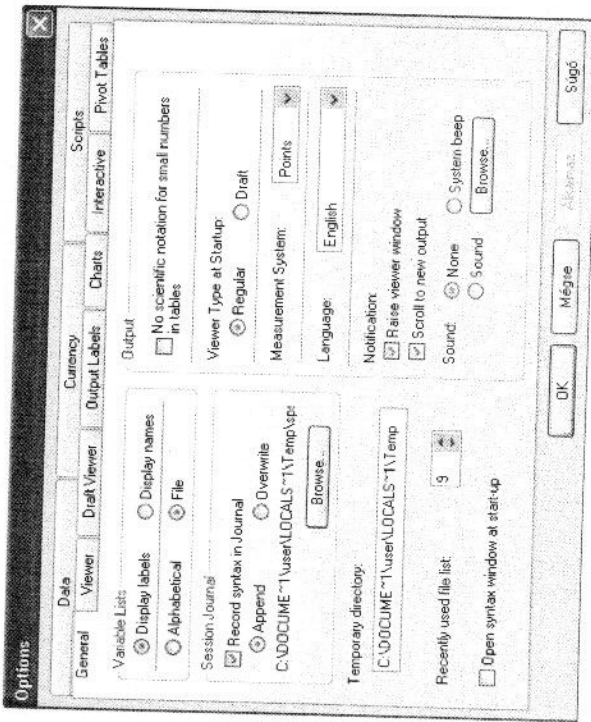
Recently Used Files: legutóbb használt nem .sav kiterjesztésű fájl elérését teszi lehetővé.

Exit: a program bezárására szolgál.

2.6.2. Edit menü

Ebben a menüpontban nagyrészt szerkesztési parancsokat találhatunk, adat-elemeket lehet másolni, kívágni, beilleszteni, törölni vagy megkeresni, illetve az adatelemeket, változókat és a korábban tárgyalt ablakokat lehet testreszabni.

- Undo:** segítségével az utójjára kiadott utasítást lehet visszavonni.
- Redo:** segítségével a visszavont utasítást lehet ismét érvényessé tenni.
- Cut:** az aktív ablakban valamely adat- vagy szövegrészt lehet kivágni, majd más helyre beilleszteni a „PASTE” paranccsal.
- Copy:** az aktív ablakban valamely adat- vagy szövegrészt lehet másolni, majd más helyre beilleszteni a „PASTE” paranccsal. A „CUT” és „COPY” parancsok alkalmazásakor az adat- vagy szövegrészek a Windows vágóasztalára kerülnek, így más alkalmazásokba is beilleszthetők.
- Paste:** beilleszti, azaz bemásolja a vágóasztalra helyezett adat- vagy szövegrészt.
- Paste Variables:** előzőleg kiválasztott változók bemásolása.
- Clear:** adat- vagy szövegrészek törlésére szolgál. Sorok vagy oszlopok törlésekor nem keletkeznek helyükön üres mezők.
- Insert Variable:** új változót (oszlopot) illeszt be attól az oszloptól balra, amelyen állunk. Ikonján oszlopok közti piros ík látható.
- Insert Cases:** új eset (sor) beillesztése azon sor fölé, ahol állunk. Sorok közti piros ík az ikonja.
- Find:** a DATA VIEW ablakban aktív, változókra lehet alkalmazni, esetekre nem (azaz függőlegesen keres, így ha a tanulmányok.sav fájl DATA VIEW nézetében a „mat” oszlopban állunk valamelyik cellán, akkor csak a 3, 4, 5 értékek valamelyikére lehet rákeresni, hiszen ebben az oszlopban más szám



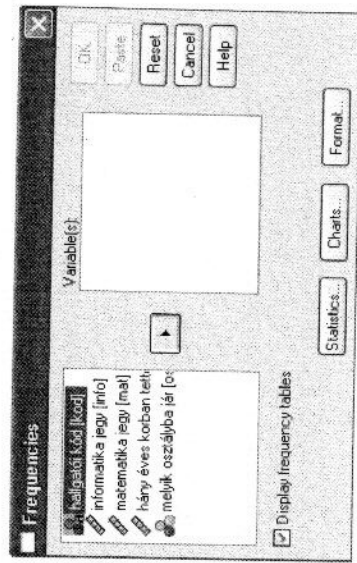
2.20. ábra. Edit./Options menüpont

nem fordul elő. Ha az „ab8” kódú diákot szeretném megtalálni, akkor a menüpont behívása előtt a „kod” oszlop valamelyik cellájára kell állni). A távcsovet ábrázoló ikonra kattintva is aktiválhatjuk a parancsot.

Go to Case: a megadott esethez (sorhoz) viszi a kurzort. Ikonján egy sor fölött álló piros nyíl látható.

Options: E menüpont alatt adatelemek, változók, ablakok beállítására és megadunk, a változók jelentése szerepel majd a felsorolásban, nevüket pedig zárójelben kapjuk csak meg. (A 2.21. ábrán például egy gyakoriságvizsgálat egyik paraméterezési ablakát látjuk.)

Ha később valamelyik statisztikai vizsgálatnál a bevinni kívánt változókat kell megadnunk, a változók jelentése szerepel majd a felsorolásban, nevüket pedig zárójelben kapjuk csak meg. (A 2.21. ábrán például egy gyakoriságvizsgálat egyik paraméterezési ablakát látjuk.)



2.21. ábra. A változók teljes nevének feltüntetése

Amennyiben a „DISPLAY NAMES” opciót jelöljük be, akkor a statisztikai elemzések során a változók bevitelkor a változók neve áll rendelkezésünkre, jelentésük nem (2.22. ábra). Ha nem tudjuk, hogy melyik változónév milyen jelentéssel bír adatmatrixunkban, érdemes a „DISPLAY LABELS”-t alkalmazni.

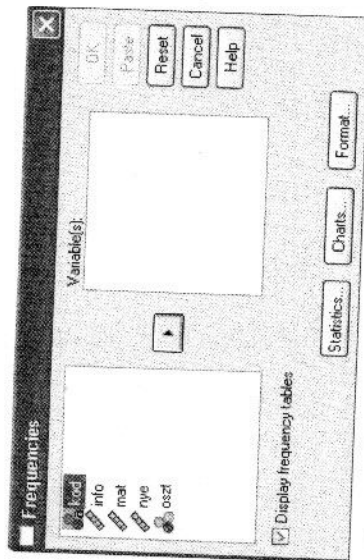
Viewer: az output ablakok beállítására szolgál (betűméret, betűtípus, szín).

Output Labels: segítségével azt lehet beállítani, hogy az output ablakban megjelenő táblázatokban, grafikonokban a változó neve (NAMES), jelentése (LABELS) vagy mindkettő (NAMES AND LABELS) szerepeljen.

Charts: az output ablakban megjelenő grafikonok beállításaira szolgál.

Interactive: az állomány mentési, nyomtatási jellemzőit lehet itt megadni.

Pivot Tables: az output ablakban megjelenő táblázatok formai beállításait teszi lehetővé.



2.22. ábra. A változók rövid nevének feltüntetése

Currency: pénznemek speciális formai beállítására szolgál. Itt lehet beállítani, hogy a tizedesek vesszővel vagy ponttal legyenek elválasztva, vagy melyik pénznem egyetemesen használja a jelet.

Data: adatok beállítási lehetőségeit lehet megtalálni. Például meg lehet határozni az új numerikus adatok formátumát (DISPLAY FORMAT FOR NEW NUMERIC VARIABLES), vagy a véletlenszám-generátor (RANDOM NUMBER GENERATOR) alapbeállítását.

2.6.3. View menü

A Nézet (VIEW) menü az aktív ablak szemmel látható tulajdonságainak beállítására ad lehetőséget. Igény szerint dönthetünk, milyen ikonok, feliratok jelenjenek meg.

Status Bar: az állapotosor beállítására szolgál. Megmutatja, hogy az SPSS matematikai műveleteket végző egysége (processor) dolgozik-e. Ezt az ablak alsó részének közepén lehet ellenőrizni, amennyiben az állapotosor aktív állapotban van.

Toolbars: a különböző ablakok eszköztárainak megjelenítésére szolgál. Beállítható, hogy milyen parancsok és ikonok jelenjenek meg az ablakok felső soraiban.

Fonts: az aktív ablakban alkalmazott karakterek betűtípusának, stílusának, méretének beállítására szolgál.

Grid Lines: aktív állapotban szemmel látható az ablak rácszata, ha kikapcsoljuk, akkor eltűnik.

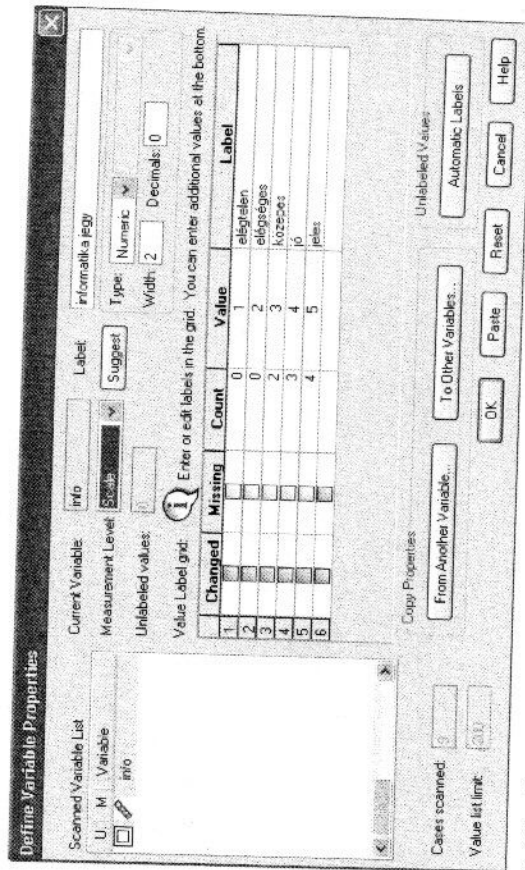
Value Labels: Értékcimkéek. Ha aktiváljuk, a program a változódefiniálási nézetben (VARIABLE VIEW) meghatározott változó jelentését mutatja a Data View ablakban, egyébként a változók értékeit.

Variabels/Data: a VARIABLE VIEW és a DATA VIEW ablakok közötti váltásra szolgál.

2.6.4. Data menü

Ez a menüpont elsősorban adatkezelésre szolgál.

Define variable properties: A változók tulajdonságainak meghatározására és megváltoztatására szolgál. Hasonló a változódefiniálási nézethez, amelytől például abban tér el, hogy itt ki lehet listázni, hogy az egyes értékekhez hány eset (COUNT) tartozik (2.23. ábra).



2.23. ábra. A változók tulajdonságainak definiálása

Copy data properties: az adattulajdonságok másolását teszi lehetővé. Vagy egy külső SPSS fájlban található adatok tulajdonságai másolhatók át az aktív adatkészletbe, vagy a jelenlegi adatkészlet tulajdonságai alapján definiálhatók további változók.

Define Dates: Datumformátumú változók meghatározására szolgál. A program külön változókat illeszt be az év, a hónap, a nap, valamint a másodperc pontossággal is meghatározható időpont számára.

Define Multiple Response Sets: többválaszos változók definiálása. Ez a funkció a DATA menüponton kívül az ANALYZE menüpontban is megtalálható Multiple response név alatt. A kettő közötti különbség abban áll, hogy a létrehozott változókat máshol tudjuk felhasználni. Az itt definiált változók az ANALYZE-TABLES menüponton belül a CUSTOM TABLES vagy

MULTIPLE RESPONSE SETS menüpontjában használhatók fel különböző táblázatok részeként, míg az ANALYZE – MULTIPLE RESPONSE – DEFINE SETS alatt létrehozott változó elemzésére külön menüpont van (FREQUENCIES, CROSSTABS), amely csak akkor aktiválódik, ha a változót már definiáltuk.

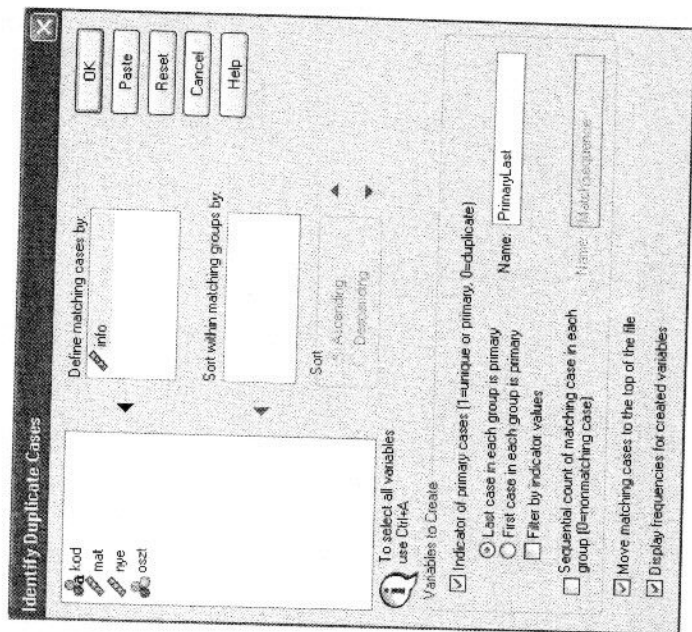
Ugyanakkor mindkét menüpont hasonló elven működik, azaz több változón belül számol össze egy vagy több értéket. Az elemezni kívánt változókat beviszük a „VARIABLES IN SET” ablakba, és ezután megadjuk, hogy egy értéket (dichotomies) vagy több értéket (categories) kívánunk összeszámolni, amely utóbbi esetében megadjuk a minimum- és a maximumértékeket (range). Az új, általunk definiált változónak nevet kell adnunk és az „ADD” gomb megnyomása után használni is tudjuk. Sajnos az adatelemzés/adat-előkészítés során nem mindig tudjuk előre, hogy melyik értékekre leszünk kíváncsiak az elemzés folyamán, ezért sok esetben az értékek átkódolására lesz szükség annak érdekében, hogy a minimum- és maximumértékek csak az általunk kívánatos értékeket tartalmazzzák.

A „dichotomies” opció olyan kérdések esetén segít, amikor ugyanazt kérdezzük sokszor és igen/nem vagy megjelölte/nem jelölte meg válaszok adhatók. Ha feltesszük azt a kérdést, hogy „Milyen napilapokat olvas ön?” és rendelkezünk a napilapok listájával, akkor az adatbázisban a napilapok lesznek a változónk és az oszlopokba olyan értékek kerülhetnek, hogy igen, olvassa (1) és nem, olvassa (0). Ebből természetesen általában csak az érdekel bennünket, ha valaki olvassa, tehát az összes napilap „1” értékeit számolja össze a program.

A „CATEGORIES” opció szintén akkor használatos, ha több hasonló kérdésünk van, azonban nemcsak egy, hanem több kategóriára vagyunk kíváncsiak. Például a „Nevezze meg az ön által leginkább olvasott 5 hetilapot”, kérdés esetén 5 változót hozunk létre az 5 hetilapnak megfelelően egységes kódolással. Ha például a listánk 30 hetilapot tartalmaz, akkor a program az 1 (első kód) és 30 (utolsó kód) közötti összes kódot megszámlolja és összesíti az 5 változóban. Ezekben az esetekben fontos, hogy a változók értékeit az adatbázisban pontosan definiáljuk.

Identify Duplicate Cases: A többször előforduló esetek azonosítását végzi el (2.24. ábra). Ezt tehetjük egyetlen változó alapján (pl. info), de elvégezhetjük akár az összes változó bevonásával is (ha kíváncsiak vagyunk, hogy ugyanazt az esetet nem rögzítettük-e tévedésből többször is).

Tegyük fel, hogy az igazgató arra kíváncsi, hogy az informatikai jegyeket tekintve mely hallgatók teljesítettek ugyanúgy. Ekkor csupán az „info” változót viszszük a fenti ablakba. A menüpont alkalmas arra, hogy az egyezőség jelzésére új változót hozzon létre (VARIABLES TO CREATE), ahol a program 1-gyel jelöli a figyelembe venni javasolt (UNIQUE OR PRIMARY), 0-val pedig a duplikált (DUPLICATE) értékeket. Az új változó neve PrimaryLast, mivel azt jelöltük be, hogy a hasonló elemekből álló csoport utolsó tagját tekintjük elsődlegesnek (LAST CASE IN EACH GROUP IS PRIMARY) (2.25. ábra)



2.24. ábra. A duplán előforduló esetek azonosítása

	kod	info	mat	nye	oszt	PrimaryLast
1	ab7	3	4	17	3	0
2	ab9	3	3	18	3	1
3	ab4	4	5	13	2	0
4	ab6	4	4	14	2	0
5	ab8	4	3	16	3	1
6	ab1	5	5	11	1	0
7	ab2	5	5	12	1	0
8	ab3	5	5	10	1	0
9	ab5	5	4	15	2	1

2.25. ábra. Egyedülálló és duplikált elemek azonosítása

A Data View-ban látható, a program az eseteket az info jegy alapján csoportosította. Emellett új változóval gazdagodtunk (PrimaryLast), ahol 1-et kapott a csoport utolsó eleme és 0-t a többi.

A 2.1. táblázatban azt láthatjuk, hogy három különféle érték van, amelyek hat-szor ismétlődnek.

Indicator of each last matching case as Primary

	Duplicate Case	Frequency	Percent	Valid Percent	Cumulative Percent
Valid		6	66,7	66,7	66,7
Primary Case		3	33,3	33,3	100,0
Total		9	100,0	100,0	

2. 1. táblázat. Az új változó (PrimaryLast) szerinti csoportosítás

Sort Cases: az esetek egy vagy több változó szerinti (csökkenő vagy növekvő) sorba rendezését lehet a menüpont alatt végrehajtani általunk megadott szempontok szerint.

Transpose: az adatbázis sorainak és oszlopainak felcserélésére van lehetőség, amely során azok szerepei is felcserélődnek. A régi változók nevei a legelső új változók eseteinek felelnek majd meg.

Restructure: a „TRANSPOSE” menüpont kiegészítése. Nem csak a teljes adatbázis, hanem az adatbázis néhány változójának vagy rekordjának transzponálását lehet végrehajtani általunk képzett csoportosítási alap szerint.

Merge Files: egy vagy több adatmátrix (állomány) eseteinek vagy változóinak összerűzését teszi lehetővé (példát lásd lentebb). A paranccsal jól kezelhetők azok az esetek vagy változók is, amelyek elternek (pl. a panel változása miatt) az összerűzendő adatbázisokban.

Aggregate: adattömörítésre, azaz az adatok aggregálására szolgál. A program az összevonást az általunk megadott funkció (például összeadás, átlag, legalacsonyabb vagy legmagasabb érték) alapján végzi. A program megkülönböztet csoportosító változót (break variable), például egy vállalat osztályai, illetve az összevonni kívánt változót (summaries of variables) például a dolgozók havi fizetése. Ebben az esetben a program egy új változóban tünteti fel az a dolgozók havi átlag fizetését, amely eltérő lesz az egyes osztályok esetében azé. Az új változót hozzáadhatjuk az eredeti fájlhoz vagy új fájlba rakhajtuk.

Orthogonal Design: ortogonális dizájn a conjoint elemzések esetében használatos (a módszerre a könyvben nem térünk ki részletesebben).

Copy Dataset: Az egész adatbázis másolására szolgál. Az SPSS új DATA EDITOR ablakot nyit ugyanazzal az adatbázissal. Használata akkor célszerű, ha valamilyen komolyabb változtatás előtt meg akarunk bizonyosodni arról, hogy az eredeti állományt semmiképpen sem írhatjuk tévedésből felül.

Split File: a parancs az adatbázist egy meghatározott változó szerint „gondolatban” részekre, csoportokra bontja, s további a statisztikai elemzéseket ezen elkülönített csoportokon végezzük. Ha az adatbázisunkra ezt a menüpontot aktiváljuk, akkor a „DATA EDITOR” ablak jobb alsó sarkán megjelenik egy felirat: „SPLIT FILE ON”. Kettévágott adatbázis az ikonja.

Select Cases: az adatbázisból eseteket lehet végleg vagy ideiglenesen kizárni. Az általunk megadott feltételeknek megfelelő rekordokkal dolgozunk tovább a statisztikai elemzések során. Ha ezt a menüpontot alkalmazzuk, a feltételeknek nem megfelelő rekordok sorszáma fekete vonallal van áthúzva, s a jobb alsó sarokban megjelenik egy felirat: „FILTER ON” (amennyiben csak ideiglenesen zártuk ki a rekordjainkat az adatbázisból). (A példát lásd lentebb.)

Weight Cases: A menüpont segítségével lehetőségünk van a reprezentatívást javítani az esetek súlyozásával azok mechanikus megsokszorozása helyett. A túlréprezentált eseteket kisebb, az alulréprezentált eseteket pedig nagyobb értékekkel súlyozza a program. A parancs ikonján egy mérleg látszik.

Fájlok egyesítése – Data/Merge Files

Korábbi példánknál maradva tételezzük fel, hogy újabb három tanuló érkezik az iskolába, akiknek az adatait egy SPSS fájl tartalmazza (2.26. ábra). Ezt az ábrát ön is elkészítheti, s lementheti „tanulmányok2” néven további felhasználáshoz. A változók paramétereit a korábban látott példával azonosak.

oszt	kod	info	mat	nye
1	ab10	2	3	16
2	ab11	3	2	17
3	ab12	2	2	18

2.26. ábra. Tanulmányok2 adatbázis

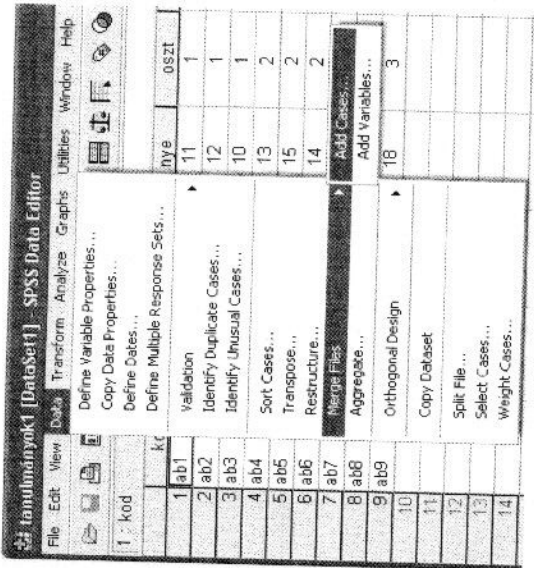
Feladatunk jelen esetben az, hogy az új tanulók adatait az előző adatbázishoz illesszük. Ezt a célt szolgálja a „DATA/MERGE FILES” menüpont, amelynek az alábbi két funkciója van:

Add Cases: az eredeti adatbázissal megegyeznek a változók, így csak eseteket illesztünk az eredeti adatbázishoz.

Add Variables: az új táblában az esetek azonosak, de más ismérveket tárolunk róluk. Így nem eseteket, hanem új változókat illesztjük be.

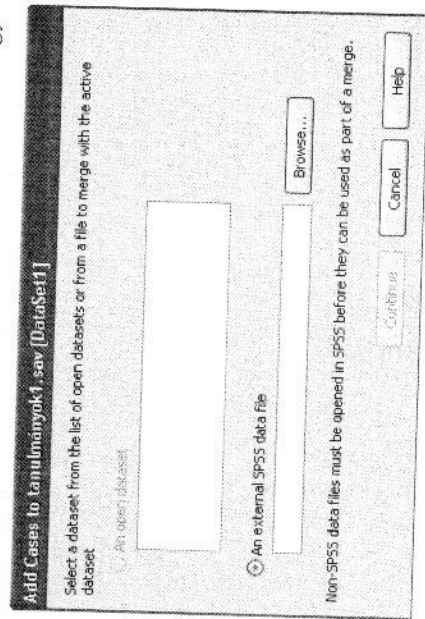
Példánk az „ADD CASES” (eset hozzáadása) használatát indokolja.

1. lépés: Használjuk a DATA/MERGE FILES/ADD CASES menüpontot (2.27. ábra).



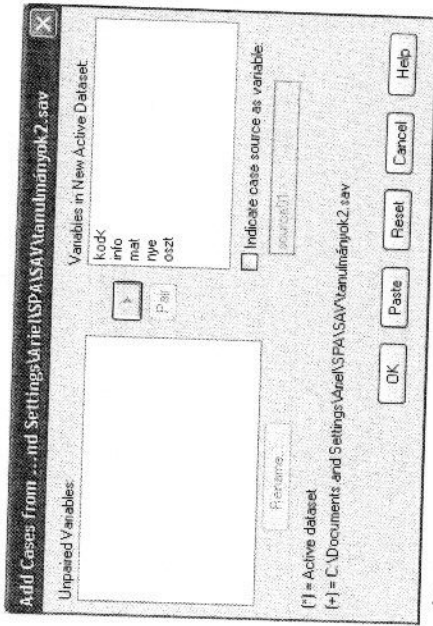
2.27. ábra. Esethozzáadás

2. lépés: Adjuk meg, melyik SPSS fájlban van a három új hallgató adata (2.28. ábra). Jelen esetben ezeket a „tanulmányok2.sav” állomány tartalmazza. Külső adatfájlt választva (An external SPSS data file) adjuk meg az elérési utat.
3. lépés: A „VARIABLES IN NEW ACTIVE DATASET” ablakba azok a változók kerülnek, amelyeket a program párosítani tudott (vagyis a két egyesíteni kívánt



2.28. ábra. Esethozzáadás egy másik SPSS fájlból

adatbázis egyaránt tartalmazza a változókat), s így benne lesznek az új, egyesített adatbázisban (2.29. ábra). Az „OK” gombra klikkelve megjelenik a kibővített adatmátrixunk (2.30. ábra). Az immáron 12 esetet tartalmazó adatbázist mentjük el „tanulmányok3” néven.



2.29. ábra. Új aktív adatbázis tartalma

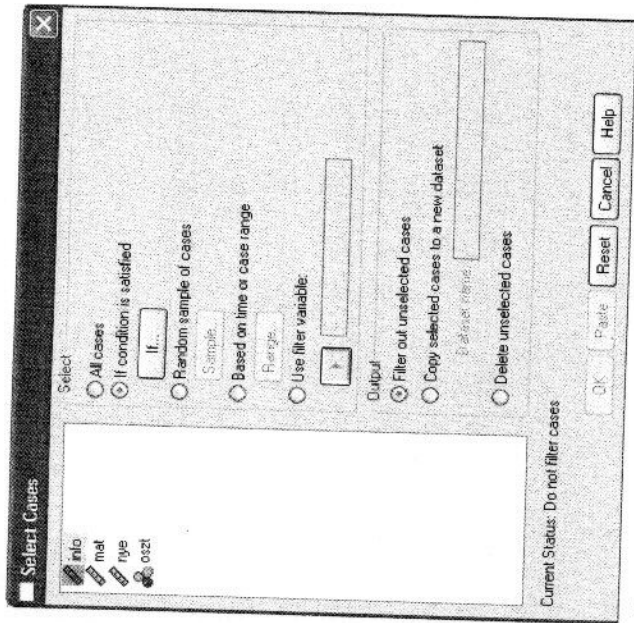
	kod	info	mat	nye	oszt
1	ab1	5	5	11	1
2	ab2	5	5	12	1
3	ab3	5	5	10	1
4	ab4	4	5	13	2
5	ab5	5	4	15	2
6	ab6	4	4	14	2
7	ab7	3	4	17	3
8	ab8	4	3	16	3
9	ab9	3	3	18	3
10	ab10	2	3	16	1
11	ab11	3	2	17	2
12	ab12	2	2	18	3

2.30. ábra. Az új adatbázis

Esetek kiválasztása/szűrése – DATA/SELECT CASES

Adatelemzés során előfordulhat, hogy nincs minden adatra szükségünk, csak bizonyos eseteket szeretnénk vizsgálni, ezért szűrni kell az eredeti adatbázis rekordjait. A menüpont segítségével mindez a következő lépésekben lehetséges:

1. lépés: Alkalmazzuk a „DATA/SELECT CASES” menüpontot. A 2.31. ábrán látott ablakban lehet beállítani a szűrőfeltételeket:



2.31. ábra. Esetek szelekciója/szűrése

All Cases: Minden esetet bevonunk az elemzésbe, vagyis nem szűrünk (alapbeállítás).

If condition is satisfied: Logikai feltétel alapján, egyszerűbb műveletek, relációs jelek, illetve függvények segítségével választunk ki az adatbázisból eseteket.

Random sample of cases: Véletlenszerűen választ ki eseteket kétféleképpen (2.32. ábra):

Approximately: Meg lehet adni, hogy az összes eset körülbelül hány százalékát vonja be az elemzésbe.

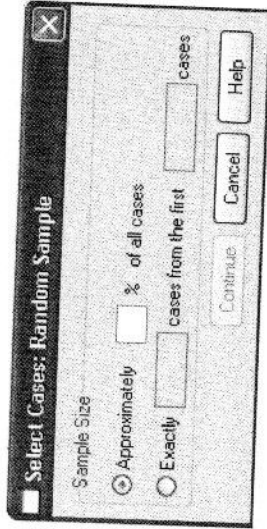
Exactly: Pontosán meg lehet adni, hogy hány esetet vonjon be az elemzésbe az első „x” darab esetből.

Based on time or case range: Szűrés időrend vagy sorrendiség szerint (például a 2. esettől a 10. esetig vonja be az elemzésbe a rekordokat).

Use filter variable: Szűrőfeltételként egy megadott változót használunk. Szűrés során a program szűrőváltozót kreál, amely természetesen újrafelhasználható.

Output: A szűrés eredményének sorsáról rendelkezhetünk itt.

Filter out unselected cases: A ki nem választott esetek az adatbázisban maradnak, de az elemzés során nem használjuk őket. Ezt a választásunkat a „DATA EDITOR” ablak jobb alsó sarkában megjelenő „FILTER ON” felirat jelzi.



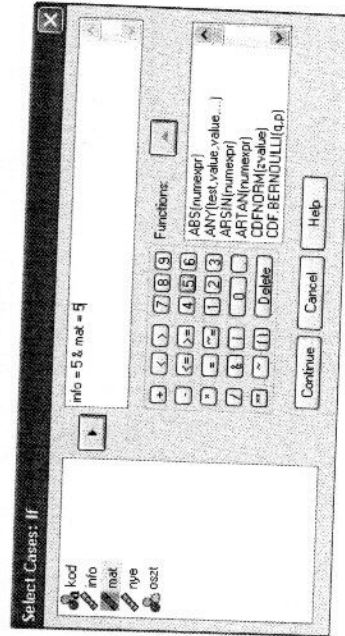
2.32. ábra. Esetek szelekciója/szűrése véletlen mintával

Copy selected cases to a new dataset: A kiválasztott eseteket egy új adatbázisba másolja.

Delete unselected cases: A ki nem választott eseteket töröljük az adatbázisból. Ennek használatát ajánljuk a legkevésbé.

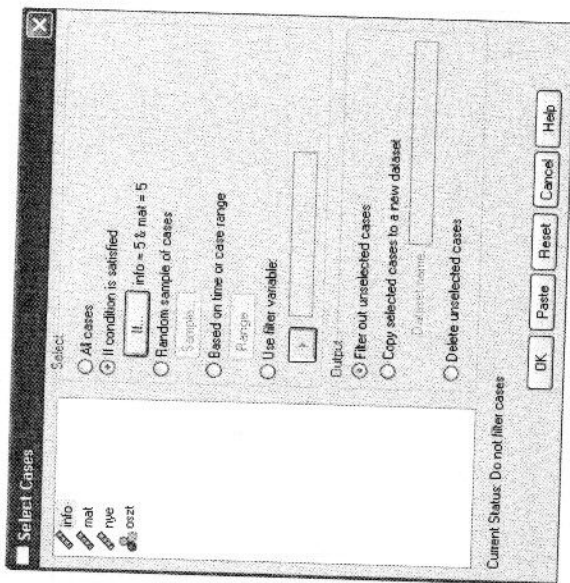
2. lépés: Alkalmazzuk az „If” lehetőséget szűrőfeltételnek. Konkrétan csak azokkal a hallgatókkal szeretnénk a továbbiakban foglalkozni, akik matematikából és informatikából is jelest kaptak, mivel közeledik az országos programozói verseny, és az iskola szűkös erőforrásai miatt csak a legjobbaknak tudja az előkészítő tanfolyamot biztosítani.

Klikkeljünk az „If” gombra és adjuk meg a feladatban megfogalmazott feltételt, akár gépeléssel, a szöközőkre figyelve, akár a gombok alkalmazásával (2.33. ábra), majd válasszuk a „CONTINUE” gombot.



2.33. ábra Feltételek megadása

3. lépés: A 2.34. ábrán látható a feltétel megadása, amelyet a gép elfogadott. Mivel a megadott feltételeknek meg nem felelő elemeket nem szeretnénk kitérölni az adatbázisból, csupán kiszűrni, ezért alkalmazzuk a Filter out unselected cases beállítást, majd kattintsunk az „OK”-ra és vizsgáljuk meg, hogy mi történt (2.35. ábra).



2.34. ábra. Feltételek rögzítése

	kod	info	mat	nye	oszt	filter_ \$
1	ab1	5	5	11	1	1
2	ab2	5	5	12	1	1
3	ab3	5	5	10	1	1
4	ab4	4	5	13	2	0
5	ab5	5	4	15	2	0
6	ab6	4	4	14	2	0
7	ab7	3	4	17	3	0
8	ab8	4	3	16	3	0
9	ab9	3	3	18	3	0
10	ab10	2	3	16	1	0
11	ab11	3	2	17	2	0
12	ab12	2	2	18	3	0

2.35. ábra. Az esetek állapota a feltételszabás után

Azokat az eseteket (9 darab), amelyek nem feleltek meg a feltételeknek, áthúzta a gép, a továbbiakban ezeket nem veszi figyelembe. Minden olyan helyzetben, amikor a „FILTER OUT...” lehetőséget választjuk, az utolsó oszlop után beszur a gép egy „filter_ \$” nevű változót, amely azt mutatja bináris kódolással, hogy a további elemzésbe bevontuk-e az adott eseteket (1), vagy sem (0).

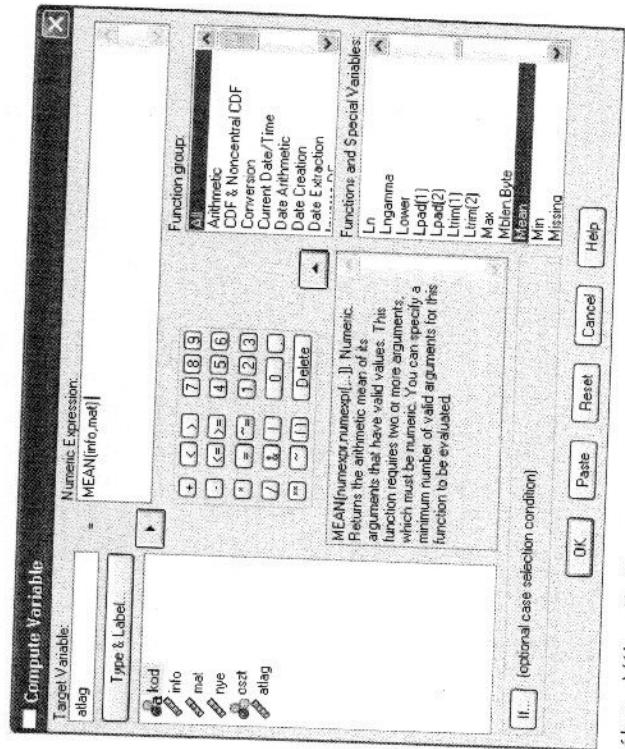
A „SELECT CASES” alkalmazásánál fontos odafigyelni arra, hogy mindaddig használatban maradjon a szűrő, amíg nem állítjuk vissza az „ALL CASES” opciót. Sok hiba adódott már abból, hogy azokra az elemzésekre is alkalmazza a program a szűrést, amelyekre eredetileg nem szerettük volna, mert elfelejtettük a kiválasztást kikapcsolni.

2.6.5. Transform menü (transzformálás)

Ez a menü szintén elsősorban adatkezelési lehetőségeket tartalmaz. Új változókat lehet előállítani régi változók segítségével, illetve az esetek ismérvétekeit lehet újrakódolni.

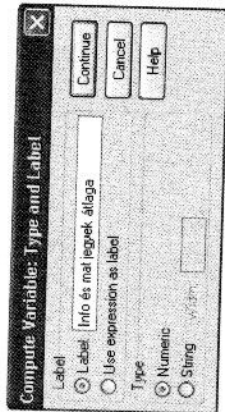
Compute: új változó számítáása

Ez a menüpont új változók létrehozását teszi lehetővé a régi változók felhasználásával. A régi változók felhasználása jelenthet köztük lévő függvényeszerű kapcsolatot, egyszerű vagy bonyolultabb logikai viszonyt. Nézzünk egy példát! Még szeretnénk határozni egy „atl” nevű változót, amely a tanulók informatika- és matematikajegyének az átlagát mutatja, s aki eléri a 4,5-ös átlagot, az év végén dícséretben részesül. Az év végi dícséret másik feltétele azonban az, hogy a tanuló junior nyelvizsgáját 14 éves kora előtt szerezte.



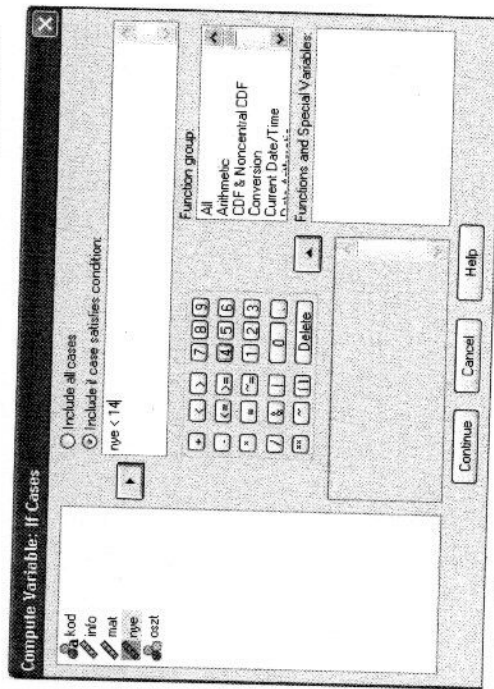
2.36. ábra. Változók képzése /számítása

1. lépés: Alkalmazzuk a „TRANSFORM/COMPUTE” menüpontot és adjuk meg a „NUMERIC EXPRESSION” panelben azt a képletet, amely segítségével a régi változókból az új kiszámolható, a „TARGET VARIABLE” dobozba pedig az új változó nevét írjuk. (2.36. ábra – a beírt képletben MEAN a számtani átlagot jelenti¹).
2. lépés: A „TYPE&LABEL” gombra klikkelve a változó bővebb jelentését adhatjuk meg a „Label”-ben. (A másik lehetőség, hogy címkéként magát a képletet használjuk. Szöveg típusú változóknál a végeredményül kapott változó hosszát is megadhatjuk.) Végül nyomjuk meg a „CONTINUE” gombot. (2.37. ábra)



2.37. ábra. Az új változó felcímkézése

3. lépés: A 2.36. ábrán látott ablak alján klikkeljünk az „IF” gombra, hogy a szűrőfeltételt megadhatssuk. Csak azokra az esetekre van szükségünk, ahol a nyelvvizsgaszerezés „életéve” kisebb mint 14 (2.38. ábra).



2.38. ábra. Az „IF” parancs használata

¹ Ehelyett használhatunk matematikai műveleteket is: (info + mat)/2.

4. lépés: A 2.38. ábrán a „CONTINUE” majd az „OK” gombra történő klikkelés után a végeredményt a 2.39. ábra mutatja.

kod	info	mal	nye	oszt	atlag
1 ab1	5	5	11	1	5,00
2 ab2	5	5	12	1	5,00
3 ab3	5	5	10	1	5,00
4 ab4	4	5	13	2	4,50
5 ab5	5	4	15	2	
6 ab6	4	4	14	2	
7 ab7	3	4	17	3	
8 ab8	4	3	16	3	
9 ab9	3	3	18	3	
10 ab10	2	3	16	1	
11 ab11	3	2	17	2	
12 ab12	2	2	18	3	

2.39. ábra. Az adatbázis az új adattal

Az első négy tanuló tette le junior nyelvvizsgáját 14 éves kora előtt, ők feleltek meg a szűrőfeltételnek, így csak az ő esetükben számolta ki a gép az informatika- és matematikajegyek átlagát, amely szerint mind a négyen kaphatnak dicséretet év végén².

RECODE: ÁTKÓDOLÁS

A menüpont segítségével már létező változóink kódolását módosíthatjuk gyorsan és egyszerűen. Két lehetőségünk van:

- INTO SAME VARIABLES (ugyanazon változókba): Ha az átkódolás után nincs szükség az eredeti változóra, akkor felülírjuk (lényegében eltüntetjük) a régi változót.
- INTO DIFFERENT VARIABLES (más változókba): az átkódolást a régi változó megtartásával végezzük.

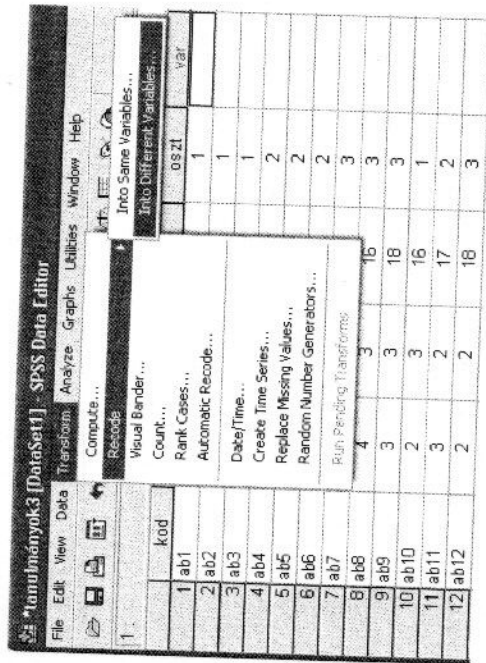
A két lehetőség paraméterezése között mindössze annyi a különbség, hogy az utóbbi esetben meg kell adni az új változó nevét és jelentését is. Az azonos néven való átkódolás ritkán praktikus, hiszen adatvesztéssel járhat. Célszerűbb új változót létrehozni, s ha valóban nem lesz szükség a régre, később azt bármikor törölhetjük.

Folytassuk előző példánkat egy átkódolási feladattal! Mivel a testnevelést órát bontott csoportban tartják, az igazgatónak be kell osztania a diákokat. Mivel

² Ezt a változót azonban a továbbiakban nem szeretnénk megőrizni adatbázisunkban, ezért töröljük onnan (ATLAG-ra rákattintva és DELETE gombot nyomva).

semmilyen megkülönböztetést nem szeretne, ezért a fennmaradó helyek alapján dönt, s ezért az első 8 diákot az első csoportba, míg a többieket a második³ időpontban induló kiscsoportba helyezi.

1. lépés: Alkalmazzuk a „TRANSFORM/RECODE/INTO DIFFERENT VARIABLES” menüpontot (2.40. ábra).



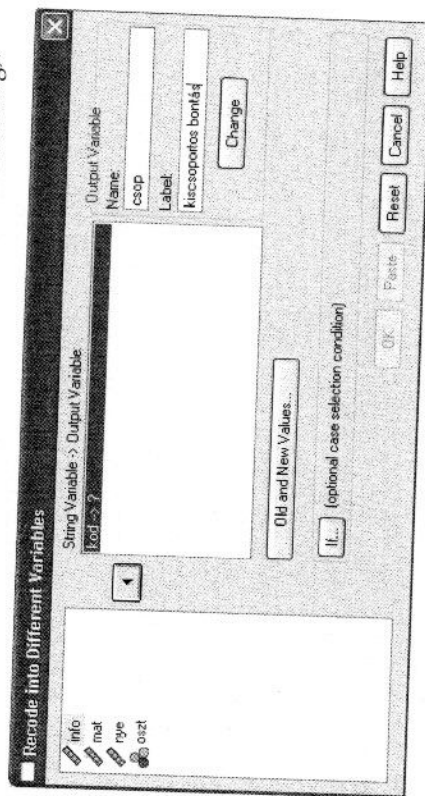
2.40. ábra. Kódolás új változóba

2. lépés: Mivel minden megkülönböztetés látszatát kerülni akarjuk, a hallgatói kódokból indulunk ki. Csakhogy így szöveges (STRING) változót kell numerikussá kódolni. Ez meg is jelenik a „STRING VARIABLE->OUTPUT VARIABLE” panelban. Az OUTPUT VARIABLE (kimeneti változó, vagyis a kódolás révén létrejövő új változó) nevét (NAME) és címkéjét (LABEL) is megadjuk. Nagyon fontos, hogy ezt a „CHANGE” gombbal mindenképpen aktiválni kell! Az újrakódolás előtt szűrőfeltételeket is megadhatunk az „IF” gombot megnyomva (2.39. ábra). Az „OLD AND NEW VALUES” gombra kattintva kell beállítani, hogy a régi változó mely értékeinek az új változó milyen értéke felel meg. (2.42. ábra)

3. lépés: Mivel szöveges változót alakítottunk át számokká, ezért bejelöljük a CONVERT NUMERIC STRINGS TO NUMBERS opciót. A második csoportba kerülő hallgatók kevesebben vannak, ezért kezdjük a kódolást velük. Az OLD VALUE/VALUE ablakba írunk ab9-et, a New VALUE/VALUE panelba

³ Az első és a második csoport nem minőségi megkülönböztetést jelent, csupán azt jelzi, hogy más időpontban van az óra (nominális változó).

pedig 2-t, s nyugtázzuk ezt a lépést az „ADD” gombbal, majd ismételjük meg ugyanezt ab10, ab11 és ab12 kóddal ellátott hallgatók esetében is. A fennmaradó értékek (ALL OTHER VALUES) 1-es értéket kapnak, mivel a többi diák az első csoportba kerül. Miután ezt is nyugtáztuk az „ADD” gombbal, látható, hogy a régi és új értékeket összehasonlító ablakban (OLD->NEW) hogyan történik az átkódolás. A 2.40. ábrán látott ablakban nem csak egy-egy konkrét értéket adhatunk meg a „VALUE” alatt. A többi lehetőséget tekintve („SYSTEM-MISSING, SYSTEM-OR USER-MISSING, RANGE...”) figyelembe vehetjük a hiányzó értékeket (ami adódhat véletlen vagy felhasználói hibából) vagy nem szöveges változók esetében adhatunk meg értékeket intervallumban (egy adott értéktől egy adott értékig, egy minimumértéktől kezdve egy a maximumértékkel bezárólag).



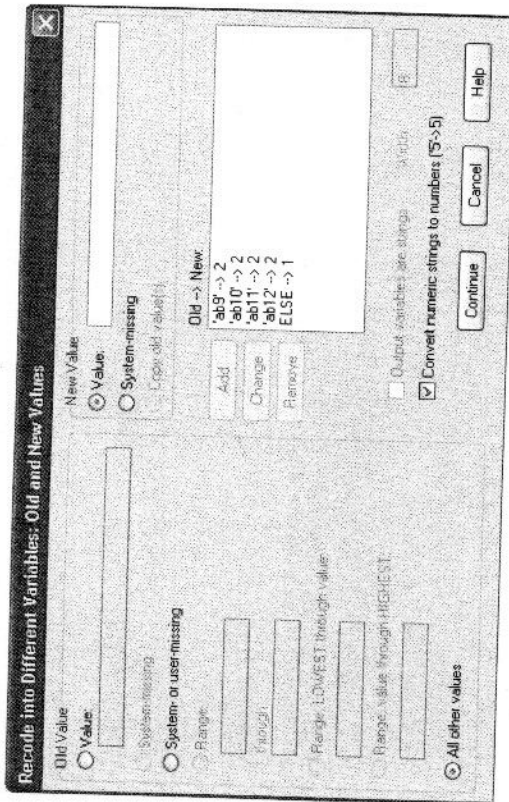
2.41. ábra. A változó bevitelre és megfeleltetésére

A „CONTINUE” gomb megnyomását követően visszaugrunk a kiinduló ablakhoz. A 2.42. ábra képernyőjén kattintunk a „CHANGE”-re, ha még ezt nem tettük volna meg, s az „Ok” alkalmazása után győződjünk meg az eredményről a 2.43. ábrán.

Már látjuk az új változót (csop), de még annak értékeihez jelentést is kell rendelnünk a VARIABLE VIEW-ban (1 – első, 2 – második). Ezek után mentjük el az adatbázist „tanulmányok4” címmel.

Nézzük meg az „EDIT/OPTIONS/DATA” parancssor segítségével, hogy a „TRANSFORMATION AND MERGE OPTIONS” a „CALCULATE VALUES IMMEDIATELY” (értékek azonnali kiszámolása) opcióra (alapértelmezés) vagy a „CALCULATE VALUES BEFORE USED” (értékek használat előtti kiszámolása) opcióra van-e állítva. Amennyiben a „CALCULATE VALUES IMMEDIATELY”-ra van állítva, az SPSS beolvassa az adatfájlt, és rögtön elvégzi a módosításokat. Amennyiben a „CALCULATE VALUES BEFORE USED” opció van megjelölve, az SPSS csak akkor végzi el a módosításokat, amikor az új

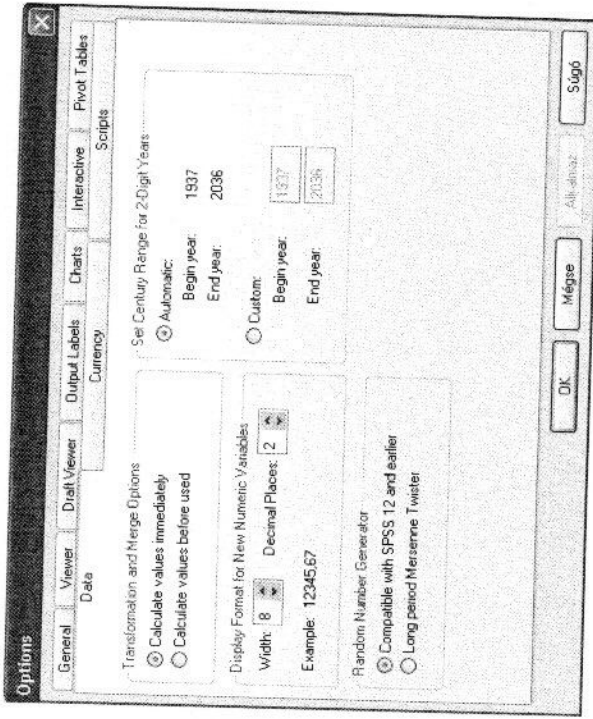
változóval valamilyen műveletet végzünk. Itt emellett azt is láthatjuk, hogy az új változó milyen alapbeállításokkal rendelkezik (DISPLAY FORMAT FOR NEW NUMERIC VARIABLES).



2.42. ábra. Az eredeti és az újrakódolt változók értékeinek megadása

	kod	info	mat	nye	oszt	csop
1	ab1	5	5	11	1	1,00
2	ab2	5	5	12	1	1,00
3	ab3	5	5	10	1	1,00
4	ab4	4	5	13	2	1,00
5	ab5	5	4	15	2	1,00
6	ab6	4	4	14	2	1,00
7	ab7	3	4	17	3	1,00
8	ab8	4	3	16	3	1,00
9	ab9	3	3	18	3	2,00
10	ab10	2	3	16	1	2,00
11	ab11	3	2	17	2	2,00
12	ab12	2	2	18	3	2,00

2.43. ábra. Az újrakódolt változó (csop)

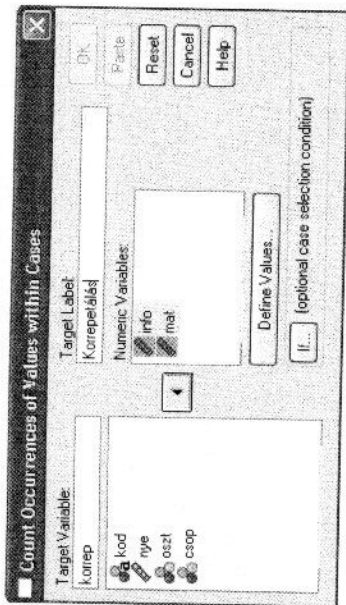


2.44. ábra. Options menüpont

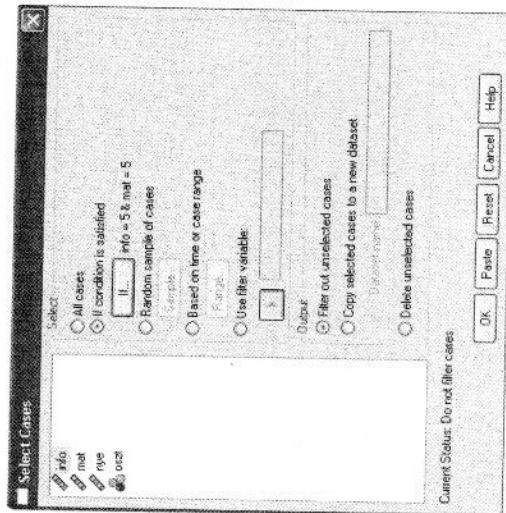
COUNT: ELŐFORDULÁSOK MEGSZÁMLÁLÁSA

A Count menüpont segítségével új változó hozható létre, amelyben az általunk megadott régi változók együttes előfordulásait lehet megadni. Tegyük fel, hogy az iskola vezetősége korrepetálni szeretné azokat a diákokat, akik informatikából vagy matematikából legfeljebb 3-as osztályzatot szereztek.

1. lépés: Alkalmazzuk a „TRANSFORM/COUNT” menüpontot.
2. lépés: Töltsük ki a paneleket a 2.43. ábrán látható módon, ahol a „TARGET VARIABLE” az új változó neve, a „TARGET LABEL” az új változó bővebb jelölése, a „NUMERIC VARIABLES” pedig azokat a régi változókat jelöli, amelyek a csoportosítás alapját képezik, majd klikkeljünk a „DEFINE VALUES”-ra.
3. lépés: A 2.46. ábrán adhatjuk meg a csoportképzésbe előzetesen bevont változók értékeit, „VALUES TO COUNT”. A feladat többféleképpen megoldható: a) az 1-es, 2-es és 3-as értékeket beírjuk egyesével a „VALUE” ablakba, majd az „ADD” gombra klikkelünk; b) a RANGE (terjedelem) ablakba beírjuk, hogy 1 THROUGH 3 (1-től 3-ig); c) felső korlátot adunk meg (RANGE, LOWEST THROUGH VALUE: 3 – Tartomány, Legalacsonyabbtól 3-ig).



2.45. ábra. A célváltozó (target variable) definiálása



2.46. ábra. A változó értékeinek meghatározása

Miután megadtuk az értékeket, a „CONTINUE”-t választjuk, amivel eljutunk a 2.47. ábrához, ahol az „If” gomb esetleges választásával további szűrőfeltételeket lehet megadni, az előző példában már látott módon.

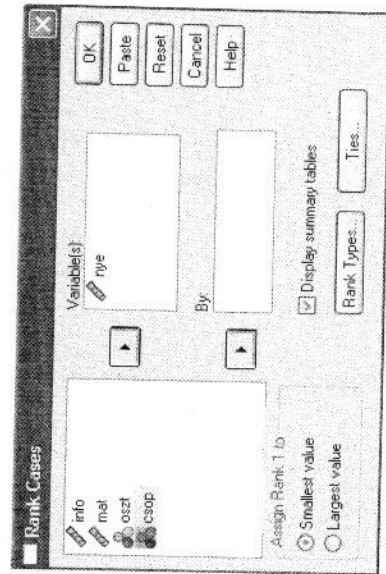
Az utolsó oszlopban a számítógép létrehozta az általunk elnevezett új változót (korrep), melynek 0, 1, 2 értéket adott. A „0” azt jelenti, hogy a tanulónak nem volt 4-esnél rosszabb jegye. Az „1”-es érték azt jelenti, hogy pontosan egy ilyen jegye volt, a „2”-es pedig azt, hogy mindkét jegye rosszabb 4-esnél. Összességében tehát 6 hallgató szorul korrepetálásra, s közülük negyven két tárgyból is⁴.

⁴ Az új változót nem kívánjuk megtartani, ezért töröljük az adatbázisból.

kod	info	mat	nye	oszt	csop	korrep
1 ab1	5	5	11	1	1	,00
2 ab2	5	5	12	1	1	,00
3 ab3	5	5	10	1	1	,00
4 ab4	4	5	13	2	1	,00
5 ab5	5	4	15	2	1	,00
6 ab6	4	4	14	2	1	,00
7 ab7	3	4	17	3	1	,00
8 ab8	4	3	16	3	1	,00
9 ab9	3	3	18	3	2	,00
10 ab10	2	3	16	1	2	,00
11 ab11	3	2	17	2	2	,00
12 ab12	2	2	18	3	2	,00

2.47. ábra. Az adatbázis az új változóval

RANK CASES (esetek rangsorolása): az eseteket rangsorolja az alapul vett változók értékei alapján. A változót, mely szerint a rangsorolást akarjuk végezni, a VARIABLES ablakba vesszük (például a „nye”-t), míg ha csoportokon belül akarunk rangsorolni, akkor a párbeszédablak jobb alsó „By...” dobozát is használjuk. Az „ASSIGN RANK 1 TO” utasításnál azt lehet beállítani, hogy az egyes helyezés a legkisebb (SMALLEST VALUE) vagy a legnagyobb értékhez (LARGEST VALUE) tartozzon-e. Ezenkívül a program lehetőséget nyújt arra, hogy meghatározzunk specifikus rangsorolási eljárásokat (RANK TYPES), illetve, hogy módosítsuk az azonos értékű változókra jutó rangszámokat (TIES). A rangsorolás elvégzése után új változót kapunk (Rnye).



2.48. ábra. Esetek rangsorolásának beállításai

AUTOMATIC RECODE: AUTOMATIKUS ÁTKÓDOLÁS

Ezt az eljárást akkor alkalmazzuk, ha az eredeti változóértékek valamilyen szempontból nem alkalmasak a feldolgozásra és azokat át kell kódolnunk. Először átvisszük az átkódolandó változókat. Ezeket újra elnevezzük a „NEW NAME” segítségével. Majd azt is meghatározhatjuk, hogy a program az újrakódolást a legnagyobb (RECODE STARTING FROM HIGHEST VALUE) vagy a legkisebb értéktől kezdje.

DATA/TIME: Időpontok és dátumok varázslója, ezekkel kapcsolatos beállítások és transzformációk hozhatók létre.

CREATE TIME SERIES (IDŐSOROK LÉTREHOZÁSA):

Idősorok alatt időben egymás után következő értékek sorát értjük. A „CREATE TIME SERIES” parancs segítségével idősor jellegű változókból más típusú idősort hozhatunk létre. Minden adat más időpontban történő megfigyelést jelent, amelyeknél az időintervallumok azonos nagyságúak (például egy részvények tőzsdéi záróára napokon keresztül). A megnyíló párbeszédablakban az eddigiekhez hasonlóan ki kell választani a feldolgozandó változókat. Ezután a „FUNCTION” alatti cellában meg kell határozni a függvény típusát (pl. szezonális ingadozás, simítás, mozgóátlagolás⁵), amellyel az átalakítást végre kívánjuk hajtani. Miután választottunk a listából, nyomjuk meg a „CHANGE” feliratú gombot. (Idősorok elemzésével részletesebben nem foglalkozunk a könyvben).

REPLACE MISSING VALUES (HIÁNYZÓ ÉRTÉKEK PÓTLÁSA):

Ezt az eljárást abban az esetben alkalmazzuk, ha minden esetben szükségünk van érvényes ismérvertékre, s nem használhatunk hiányzóérték-kódot. A párbeszédablakban a szokásos módon meg kell adni a változókat, amelyekkel dolgozni akarunk. A „METHOD” alatt választhatjuk ki a lehetséges értékeket (teljes átlag, szomszédos pontok átlaga, szomszédos pontok mediánja, lineáris interpoláció, az adott pontra vonatkozó lineáris trend), amelyekkel helyettesíthjük a hiányzó adatokat.

RANDOM NUMBER GENERATORS: (VÉLETLENSZÁM-GENERÁTOROK):

Két véletlenszám-generátort tartalmaz a program, az elavultabb, de a 12-es verzióval kompatibilis megoldást (SPSS 12 COMPATIBLE), illetve az újabb és megbízhatóbb MERSENNE TWISTER-t. Az ACTIVE GENERATOR INITIALIZATION menüpontnál megadhatunk fix indulóértéket (FIXED VALUE) is.

VISUAL BANDER: VÁLTOZÓK KATEGORIZÁLÁSA

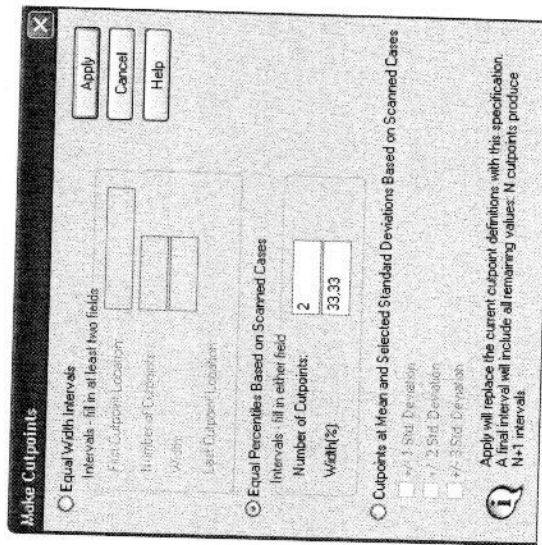
Sokszor, ha numerikus változót (például jövedelmet) elemzünk, az eredeti (folytonos) változó mellett szükségünk van annak kategorizált változatára is. Ennek

5 Idősorok elemzéséről lásd részletesebben Hunyadi–Mundruczó–Vita (2001).

létrehozásában segít a VISUAL BANDER parancs, amely lehetővé teszi, hogy a folytonos változót gyorsan és egyszerűen kategóriákra bontsuk. Ehhez egy hisztogramot rajzol, ahol megadhatók a kategóriák alsó és felső értékei.

Alakítsunk ki például a „nye” változó alapján három kategóriát! Alakítsuk a „TRANSFORM/VISUAL BANDER” menüpontot. És válasszuk ki a „nye” változót a VARIABLES ablakból, ahol csak azok a változók vannak feltüntetve, amelyek ordinálisak vagy metrikusak (hiszen a nominális változóknál nincs értelme a kategorizálásnak). Miután átvittük a kategorizálandó változókat (VARIABLES TO BAND), nyomjuk meg a folytatás (CONTINUE) gombot.

Ezután a SCANNED VARIABLE LIST ablakban kicikkeljük a „nye” változóra, s megjelenik a változót ábrázoló hisztogram, amely segíthet annak eldöntésében, hány kategóriát hozunk létre. A hisztogram ugyanis megmutatja az eloszlás képét, vagyis azt, mely érték(ek) körül sűrűsödnek az adatok. Ezután meghatározzuk az osztópontokat (MAKE CUTPOINTS) (2.49. ábra).



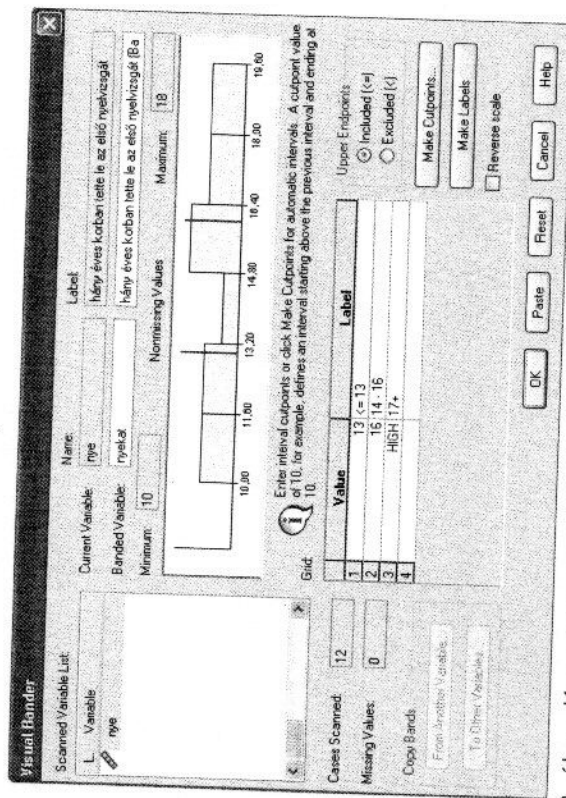
2.49. ábra. Az osztópontok meghatározása

Az osztópontok meghatározása háromféleképpen történhet:

- Azonos szélességű intervallumok (EQUAL WIDTH INTERVALS), ahol meghatjuk az első osztópontot, az osztópontok számát, valamint a szélességet is (hisztogram segít).
- Egyenlő percentílesek alapján (EQUAL PERCENTILES). Itt nem feltétlenül lesznek azonos szélességűek az intervallumok, azonban ugyanannyi esetet tartalmaznak. Válasszuk ezt a megoldást, és hozzunk létre három kategóriát. Ebben az esetben 2 osztópontot (NUMBER OF CUTPOINTS) van szükségünk.

- Átlag és szórás alapján (CUTPOINTS AT MEAN AND SELECTED STANDARD DEVIATIONS). Itt meghatározhatjuk, hogy az átlagon kívül akörül még milyen szórásértékeknél (1, 2 vagy 3) legyenek az osztópontok.

Az APPLY gomb megnyomása után visszatérünk a kiinduló ábrához, ahol láthatjuk, hogy az értékeknél (VALUE) megjelent a két osztópont értéke, amelyek a hisztogramon pedig kék vonal jelöl.



2.50. ábra. Kategóriaképzés

A MAKE LABELS parancsal a program automatikusan hozzárendeli az értékekhez (VALUE) a címkét (LABEL). Ezek után nevezzük el a kategorizált változót (BANDED VARIABLE) ebben az esetben „nyekat”-nak, s kattintsunk az OK gombra.

A 2.51. ábrán látható, hogy az utolsó oszlopban a számítógép létrehozott egy új változót („nyekat” néven) három kategóriával. A címkék pedig automatikusan megjelennek a VARIABLE VIEW-ban is.

2.6.6. Analyze menü (elemzés)

Az SPSS legfontosabb, legösszetettebb menüje, amely a tényleges statisztikai elemzéseket tartalmazza. Az „ANALYZE” menüpontot részletesen későbbi fejezetekben tárgyaljuk.

	kod	info	mat	nye	oszt	csop	nyekat
1	ab1	5	5	11	1	1	1
2	ab2	5	5	12	1	1	1
3	ab3	5	5	10	1	1	1
4	ab4	4	5	13	2	1	1
5	ab5	5	4	15	2	1	2
6	ab6	4	4	14	2	1	2
7	ab7	3	4	17	3	1	3
8	ab8	4	3	16	3	1	2
9	ab9	3	3	18	3	2	3
10	ab10	2	3	16	1	2	2
11	ab11	3	2	17	2	2	3
12	ab12	2	2	18	3	2	3

2.51. ábra. Az adatbázis az új változóval

2.6.7. Graphs menü (ábrázolás)

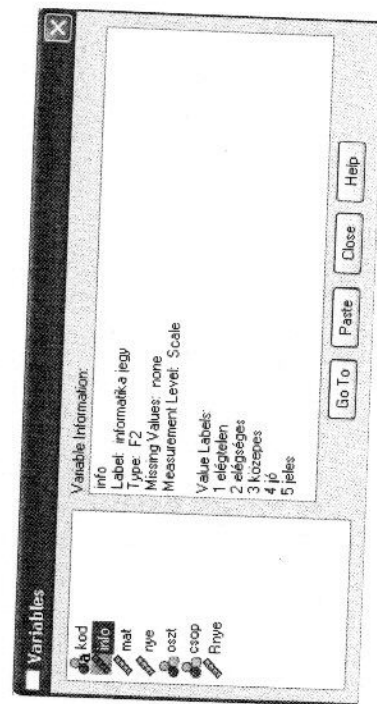
A statisztikai elemzésekből nyert adatok gyors, szemléletes megjelenítését segítik a menüpont alatt található különböző grafikonok, ábrák, diagramok.

INTERACTIVE: a fentiekben említett grafikai lehetőségek finombeállításait teszi lehetővé.

2.6.8. Utilities menü (támogató eszközök)

A menüből néhány hasznos kiegészítő lehetőség érhető el.

VARIABLES: a változók paramétereinek leírását lehet megtekinteni egyesével egy output ablakban.



2.52. ábra. Változó információ

BAR: az oszlopdiagramok egy kategorizált változóra építve a sokaság szerkezetét szemléltetik.



SCATTER: értékpárok vagy érték-hármasok pontfelhőszerű megjelenítésére szolgál, elsősorban mennyiségi ismérvek közötti kapcsolatot szemléltetésére alkalmas.



LINE: a vonaldiagram az adatpontokat folytonos vonallal köti össze, idősorokat lehet vele jól ábrázolni.

HISZTOGRAM: egy változó eloszlását szemlélteti, metrikus skálák esetén használható.



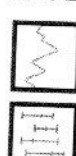
AREA: a sávidiagram egy vagy több görbe által lefedett tartomány megjelenítését szolgálja.

P-P: egy változó eloszlásfüggvényét rajzolja ki a normális eloszlásfüggvényvel összevetve.



PIE: a kördiagrammal elsősorban egy sokaság szerkezetét lehet szemléltetni.

Q-Q: egy változó kvantilisai és a normális eloszlás elméleti kvantilisai között lehet párhuzamot vonni.



HIGH-LOW: összetartozó értékpárok vagy értékhármasok grafikonja.

SEQUENCE: Esetek sorban történő ábrázolása, főként idősorok esetében használatos.



PARETO: egy oszlopdiagram és a kumulált összeget mutató vonaldiagram kombinációja.

AUTOCORRELATIONS: Segítségével idősoroknál vizsgálható az autokorreláció, vagyis a hibataragok korrelációja.



CONTROL: folyamatirányítási grafikon, amellyel ellenőrizhető, hogy a tesztelt változó a határok között mozog-e.

CROSS-CORRELATIONS: Két idősor közti korreláció megjelenítésére szolgál.



BOXPLOT: A „dobozábráról” leolvasható a medián, a kvantilisok, a minimum és maximum-érték, illetve a kiugró elemek is.

SPECTRAL: Idősorok elemzése (spektrumbecslés) esetén használatos komplex vonaldiagram.



ERROR BAR: Az átlagbecslés konfidenciaintervallumát, standard hibéjét vagy szórását mutatja változónként.

POPULATION PYRAMID: Népeségpiramis. Két, egymással szembe fordított hisztogram (metrikus változónál) vagy oszlopdiagram (kategorizált változónál).



3-D BAR: Háromdimenziós oszlopdiagram, látványosabb megjelenítést tesz lehetővé.

MIXED: Kombinált változat (pl. oszlop- és vonaldiagram együtt).

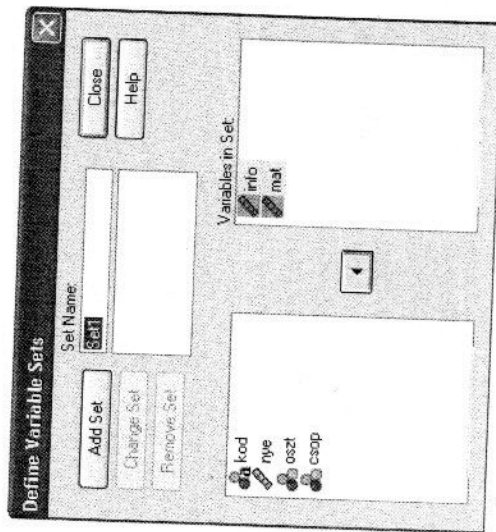
2.53. ábra. Különböző grafikai megjelenítési lehetőségek

OMS CONTROL PANEL: A kimeneti menedzsmentrendszer (Output Management System vagy OMS) lehetővé teszi, hogy a kiválasztott kategóriákat automatikusan a különféle kimenetifájl-típusokba (sav, xml, html vagy text formátumba) írja.

OMS IDENTIFIERS: Segítségével OMS parancsok írhatók.

DATA FILE COMMENTS: Megjegyzések írhatók az adatfájlhoz.

DEFINE SETS: sok változónál statisztikai elemzésünket leszűkíthetjük a változók egy részhalmozára úgy, hogy a részhalmozba felvevő változókat megjelöljük. Például az info és a mat változókból új csoportot hozunk létre (VARIABLES IN SET) és ellátjuk a Set1 névvel.



2.54. ábra. Változócsoporthoz definiálása

USE SETS: ebben a menüpontban szűkíthető le az elemzés a változók adott részhalmozára.

MENU EDITOR: a menüsor szerkesztését, testreszabását (például menüpontok beszűrésát és eltávolítását) teszi lehetővé.

2.6.9. Window menü (ablakkezelés)

A felhasznált ablakok méreteit lehet beállítani.

MINIMIZE ALL WINDOWS: az összes ablakot a tálcára helyezi, kicsire zárva.

SPLIT: Képernyő felosztása, a kisablakok külön nézhetők és mozgathatók.

Ezekon kívül még lehetőség van az aktív ablakok közötti váltásra.

2.6.10. Help menü (segítség)

Általános információk lekérdezésére szolgál, egyfajta szöveges segítségnyújtás az SPSS program használatához.

TOPICS: témák alapján kereshetők az információk.

TUTORIAL: oktatóprogram feladatokkal.

CASE STUDIES: Esettanulmányok, amelyek a tanulást segítik.

STATISTICS COACH: eredmények megjelenítéséhez nyújt segítséget.

COMMAND SYNTAX REFERENCE: az SPSS által használt parancsok leírása.

ALGORITHMS: eljárások, képletek tárháza.

SPSS HOME PAGE: <http://www.spss.com/> A program honlapját hívja be.

ABOUT: rövid leírás az SPSS-ről, főbb adatok.

LICENSE AUTHORIZATION WIZARD: licenc engedélyeztetése.

REGISTER PRODUCT: termék regisztrálása nyomtatott formában, levélben.

CHECK FOR UPDATES: Programfrissítések keresése az interneten keresztül.

Kódolási esettanulmány

Egy kérdőív kódolása nem mindig könnyű feladat. A továbbiakban olyan típuskérdéseket mutatunk be, amely a kevésbé tapasztalt kutatókat gondolkodóba ejtheti. A következő fiktív kérdőív egy olyan kutatás része, amely a csokoládéfogyasztást hivatott feltérképezni a fiatalok körében. Azt javasoljuk, hogy aki gyakorolni szeretne, először – a kódolási tippek elolvasása nélkül – próbálja egyedül definiálni a változókat a VARIABLE VIEW-ben.

KÉRDŐÍVRÉSZLET

1. Milyen szó jut eszedbe a csokoládéről?

Kódolási tipp: Célszerű rövid névnek vagy a változóra utaló rövidítést, vagy pedig kérdés alapján történő számozást (pl. NAME: k1) használni. Hosszabb kérdőívnél mindenképpen az utóbbi javasolt. A címke (LABEL) meghatározásához bemásolhatjuk az egész kérdést.

A problémát ennél a kérdésnél az jelenti, hogy legrosszabb esetben annyi-féle alternatívát rögzíthetünk (VALUES), ahányféleképpen válaszoltak a kérdésre (vagyis ez nem túl kódolóbarát kérdés). A válaszokat rögzíthetjük szöveges (string) változóként (például: „barna”, „ét”), vagy azonnal bekódolhatjuk, minden egyes alternatívát külön kóddal illetve (1=barna, 2=ét). Az első eset előnye, hogy a szöveges válaszokat kilistázhadjuk és utólag cso-

portokat képezve bekódolhatjuk. Bármelyik alternatívát is választjuk, biztos, hogy nominális (MEASURE: NOMINAL) változót kapunk.

2. Mi alapján választasz csokoládét?

Rangsorold 1–5-ig az alábbi szempontokat! (1 – legfontosabb szempont, 5 – legkevésbé fontos szempont).

- Márka
- Ár
- Minőség
- Csomagolás
- Reklám

Kódolási tipp: Itt egy kérdésen belül nem egy, hanem több (öt) változóval találkozunk, amelyeket mind külön be kell kódolni. A márka a k2.1 nevet kapja (NAME), a címke (LABEL) megadására pedig több lehetőség is adódik: a) kiírjuk a teljes kérdést választal (Mi alapján választasz csokoládét? – Márka), b) rövid címkét választunk (pl. Márka). A kutató ezek közül saját preferenciája alapján dönt. Az ár a k2.2, a minőség a k2.3, a csomagolás a k2.4, a reklám pedig a k2.5 nevet kapja, ezzel jelezzük, hogy a változók a 2-es kérdéshez tartoznak. Mindegyik változó rangsorszáma (MEASURE: ORDINAL). Értékeket (VALUES) itt nem szükséges definiálni.

3. Mennyire értesz egyet az alábbi állításokkal a csokoládékkal kapcsolatosan? (5 – teljes mértékben egyetérték, 4 – inkább egyetérték, 3 – egyet is értek, meg nem is, 2 – inkább nem értek egyet, 1 – egyáltalán nem értek egyet)

A csokoládé örömet okoz	1	2	3	4	5
A csokoládé egyszerű ajándék	1	2	3	4	5
Akkor eszem csokit, ha bánatos vagyok	1	2	3	4	5

Kódolási tipp: Többféle, ötfokozatú Likert skálát látunk, ahol az állítások külön változónak számítanak (NAME: k3.1, k3.2, k3.3). Itt definiálhatjuk az egyes értékeket (VALUES: 5 – teljes mértékben egyetérték, 4 – inkább egyetérték, 3 – egyet is értek, meg nem is, 2 – inkább nem értek egyet, 1 – egyáltalán nem értek egyet), amelyeket aztán a COPY/PASTE parancsokkal átmosolhatunk a k3.1 változóból a k3.2 és a k3.3 VALUES dobozába. Ez egy intervallumskála (azaz metrikus, MEASURE: SCALE).

4. Milyen márkákat szoktal fogyasztani? (Több válasz is bejelölhető)

- A márka
- B márka
- C márka
- D márka

Kódolási tipp: Mivel több válasz is megjelölhető, ezért minden egyes alternatíva külön változóként szerepel (k4.1-től k4.4-ig). Az értékek (VALUES) megadása annyit jelent, hogy az adott márkát bejelölte-e a válaszadó vagy sem (0=nem, 1=igen). Ennek alapján nominális (MEASURE: NOMINAL) változót kapunk.

5. Jellemezd „A” márkát az alábbiak jellemzők alapján! Minél közelebb teszed az X-et egy tulajdonság mellé, annál jellemzőbb rá az a tulajdonság!

barátságos	-----	barátságtalan
szomorú	-----	vidám
keserű	-----	édes

Kódolási tipp:

Szemantikus differenciáliskála, ahol két ellentétes tartalmú jellemző olvasható. Itt is mindegyik sor külön változónak feleltethető meg (k5.1, k5.2, k5.3), a címkek (LABEL) azonban különbözők lehetnek (pl. „barátságos – barátságtalan” vagy „barátságosság mértéke”). Véggponi értékek (VALUES) megadhatók (pl. 1=barátságos, 7=barátságtalan), de nem feltétlenül szükséges. Intervallumskála (MEASURE: SCALE).

6. Legmagasabb iskolai végzettséged:

- Általános iskola
- Középiskola
- Egyetem / főiskola

Kódolási tipp:

Egyetlen változó (k6), amelynek több értéke (VALUE) lehet (1=általános iskola, 2=középiskola, 3=egyetem/főiskola). Rangsorskála (MEASURE: ORDINAL).

7. Kérlek sorold fel azokat a médiumokat, ahonnan a kerületben történt eseményekről tájékozódsz:

	Spontán említés
Helyi tv	
Helyi rádió	
Helyi lap	
Országos napilap	

Kódolási tipp: Ennek a kérdésnek a kódolása kétféleképp történhet, vagy a sorok, vagy pedig az oszlop lesz a változó. Ha a sorokból, azaz az egyes médiatípusokból indulunk ki, négy változót kapunk a négy médiatípusnak

megfelelően, és két érték (VALUES) lesz minden változóban, az 1=említetté, illetve 2=nem említetté.

Ha az oszlop alapján alakítjuk ki a változóinkat, akkor ez azt jelenti, hogy az említések kerülnek rögzítésre, amelyekben belül négy érték (VALUES) lesz a négy médiatípusnak megfelelően (1=helyi tv, 2=helyi rádió, 3=helyi lap, 4=országos napilap). Ilyenkor, ha a megfigyeléseink szerint a válaszadók maximálisan 3 említést tettek, akkor 3 változó elegendő a rögzítésre⁶.

Miután definiáltuk a változókat, következhet az adatok bevitelére a DATA VIEW-ba. Az adatok rögzítése után meg kell vizsgálni azok konzisztenciáját. Ezt a folyamatot nevezik adattisztításnak. Nem egy esetben a megkérdezettek figyelmetlenül töltik ki a kérdőíveket, és így abban ellentmondásokat lehet felfedezni. A nem konzevens adatok megtalálására a leggyorsabb lehetőség az „ANALYZE” menüponton belül a „FREQUENCES” alkalmazása (lásd 3. fejezet).

⁶ Ennél is elegánsabb megoldás, ha a kódolást a kettős számmendszert követve végezzük (1=helyi tv, 2=helyi rádió, 4=helyi lap, 8=országos napilap). Ekkor adatvesztés nélkül tárolhatjuk egyetlen változóban az információt, ha annak értékeként azt a számot rögzítjük, amelyet a bejelölt médiumok kódjának összegeként kapunk. Ez a technika különösen több változó összehasonlásakor nyújt hasznos segítséget. (A szerkesztő.)

3. fejezet

EGYVÁLTOZÓS ELEMZÉSEK

A 3. fejezettől kezdve a könyv egyes fejezeteiben különböző elemzési módszereket mutatunk be az egyszerűbbtől a bonyolultabb felé haladva. Az elemzési módszerek közül az első az egyváltozós elemzési módszer, amelyet a két-, illetve, többváltozós elemzési módszerek követnek.

A fejezet célja

A fejezet célja, hogy az olvasó a fejezet elolvasása után

- ismerje az egyváltozós elemzések technikáját, lényegét és elvégzésének feltételeit,
- átfogó képet tudjon alkotni egy adott mintáról néhány alapvető statisztikai mutató alapján
- tudja előállítani és elemezni a különböző statisztikákat és mutatókat,
- képes legyen következtetéseket levonni az elvégzett elemzések alapján.

3.1. Gyakorlati alkalmazhatóság

Az egyváltozós elemzések célja az adatstruktúrába való elsődleges betekintés az adatbázisban lévő változók egyenkénti elemzése által. A változók egyenkénti elemzése önmagában fontos lehet, például ha kíváncsiak vagyunk arra, hogy melyik válasz a leggyakoribb egy kérdésre, ugyanakkor az egyváltozós elemzési módszerek hozzásegítenek a többváltozós elemzési módszerekhez szükséges feltételek vizsgálatához is. Számos esetben találkozunk azzal a problémával, hogy az egyváltozós elemzések jelentősége alábecsült, illetve nem kerülnek elvégzésre, különös-képp, ha többváltozós módszerek kerülnek alkalmazásra. Tisztában kell lennünk

azonban azzal, hogy a többváltozós módszerek – a nevüknek megfelelően – számos változó közötti összefüggést mutatnak, amely összefüggésre minden változónak hatása van. Ennek megfelelően fontos az, hogy minden változót jól ismerjünk, mielőtt többváltozós elemzésnek vetjük alá őket, s a kutató legyen tisztában az adatok jellegével, struktúrájával. Más megközelítésben: ha megkívánunk egy piros almát a fán, rögtön leszakíthatjuk, s beleharaphatunk, de szerencsésebb először megmosni a permetezettől, esetleg szemrevételezni, hogy nincs-e szezonális albról az almán belül. S ha ezekre az egyszerű kérdésekre megadtuk a választ, valamint rendelkezünk a szükséges információkkal, akkor továbbléphetünk, s nyugodtan elfogaszthatjuk a gyümölcsöt.

Néhány egyszerű példán keresztül nézzük meg az egyváltozós elemzések gyakorlati alkalmazhatóságát.

- Egy országos elosztási hálózattal rendelkező, üzemanyagot árusító vállalat ki szeretné választani a leggyengébben és a legjobban működő telephelyét. Emellett az adott évi teljesítményt alapul véve előrejelzést kíván adni a következő évi benzinteljesítmény várható nagyságáról a különböző telephelyek mindegyikére.
- Egy üzletlajdonos ismerni szeretné az üzletében található minőségileg kifogástalan (azaz leértékelés nélkül eladható) ruhadarabok számát, azok átlagos, minimum (legolcsóbb árú cikk), illetve maximum (legdrágább árú cikk) árát, illetve hogy a ruhák többsége milyen árkategóriába esik. Ennek megfelelően az üzletlajdonos meg tudja határozni a potenciális fogyasztók körét, illetve hogy a bolt arculata, illetve pozicionálása összhangban van-e a domináns árkategóriával. Az ilyen és az ehhez hasonló kérdésekre az egyváltozós elemzések segítségével felelhetünk.

Az előbbi példákban egyszerű kérdésekre, vezetői döntési problémákra keresünk a választ egyszerű statisztikai mutatók segítségével:

- Hány érvényes elem áll rendelkezésre (különböző telephelyek összessége, ruhadarabok száma)?
- Melyik a legkisebb elem (leggyengébben működő telephely, legolcsóbb ruhadarab)?
- Melyik a legnagyobb elem (legjobban működő telephely, legdrágább ruhadarab)?
- Mennyi az átlagos érték (heti várható benzinteljesítmény érték, átlagos ruhaár)?
- Mekkora az elemeknek az átlagtol való átlagos eltérése (szórása), például mekkorák a benzinteljesítmény térfelnyagság eltérései az átlagos értéktől?
- Van-e „gyanús” elem (kiugró érték) a mintában? Ha például az egyik benzinkút egymillió liter üzemanyagot adna el egy nap alatt vagy az egyik ruhadarab ötszáz ezer forintba kerülne, akkor ezek az értékek a többi „átlagos” benzinkút és ruha értékétől jelentősen eltérnek. Ezekben az esetekben a kutató feladata, hogy ezeket az értékeket azonosítsa és megvizsgálja a torzítások elkerülése érdekében.

3.2. Általános elméleti áttekintés

Az egyváltozós elemzés minden adatbázis-elemzés első lépései között szerepel. Ennek során a változókat egyenként, egymástól függetlenül elemezzük. Az egyváltozós elemzések során a leggyakrabban alkalmazott mutatókat négy csoportba sorolhatjuk: helyzet-, szóródási, alak- és egyéb mutatószámok, amelyeket a 3.1. táblázat tartalmaz.

Helyzetmutató számok	Szóródási mutatószámok	Alakmutató számok	Egyéb mutatószámok
Átlag (Mean) Medián (Median) Módusz (Mode)	Terjedelem (Range) Szórás (Standard deviation) Varianscia (Variance)	Csúcsosság (Kurtosis) Ferdesség (Skewness)	Összeg (Sum) Elemek száma (Number of cases) Minimum (Minimum) Maximum (Maximum)

3.1. táblázat. A leggyakrabban alkalmazott statisztikai mutatószámok (zárójelben az SPSS által is használt angol megfelelő)

• Helyzetmutatók:

- **ÁTLAG (MEAN):** az elemek számtani átlaga. Az átlag alkalmazása intervallum és arányszála esetén a legmegfelelőbb, ugyanakkor nem alkalmazható például sorrendi, illetve nominális (névleges) skálák esetén. Az átlag hátránya az, hogy minden érték ugyanolyan hatással van rá, azaz ha van néhány kiugróan alacsony/magas értékünk, akkor ezek jelentős mértékben befolyásolják az átlagot. Ugyanez nem áll fenn a medián és a módusz esetében, azaz a kiugróan magas vagy alacsony értékek nincsenek hatással ezekre a mutatókra. Ez azt jelenti, hogy egy olyan arányszálan mért változó esetén, amelyben vannak kiugró értékek, érdemesebb a mediánt használni, mintsem az átlagot. Erre példa az, hogy egy adott város vagy régió stb. ingatlanértékesítéseinek átlagos értékét a kiugróan magas vagy éppen alacsony eladási árak miatt – a mediánnal szokták kifejezni és nem az átlaggal.

- **MEDIÁN (MEDIAN):** olyan középérték, amelynél az elemek fele nagyobb, a fele pedig kisebb az esetek sorrendbe állítása esetén. A medián alkalmazása sorrendi skála esetén a legmegfelelőbb, és értéke megegyezik a középső értékkel, amennyiben páratlan, illetve a középső két érték átlagával, amennyiben páros számú esetünk van.

- **MÓDUSZ (MODE):** a leggyakrabban előforduló elem (diszkrét ismérv esetén), illetve a gyakorisági görbe maximumhelye (folytonos ismérv esetén). Akkor jó mutatószám, ha a változó kategorizált. Névleges skálánál csak ez használható, ugyanakkor intervallum és arányszála esetében is alkalmazható.

A fentiek tükrében tehát minden skálatípusnak megvan a legmegfelelőbb mutatója, és az alkalmazható mutatók köre bővíti a magasabb rendű skálák esetében. Ez

azt jelenti, hogy míg a nominális skálák esetében kizárólag a módusz alkalmazható, addig az arányskálák esetében egyaránt alkalmazható a módusz, a medián és az átlag is.

• **Szóródási mutatószámok:**

- **TERJEDELEM (RANGE):** a legnagyobb és legkisebb elem közötti különbség. Metrikus skála esetén alkalmazzuk.
- **SZORÁS (STANDARD DEVIATION):** megmutatja, hogy az elemek az átlagos értéktől, mennyire térnek el átlagosan¹. Metrikus skála esetén alkalmazott.
- **VARIANCIA (VARIANCE):** a szórás négyzete, vagyis az átlagtól való négyzetes eltérés. Metrikus skála esetén használjuk.

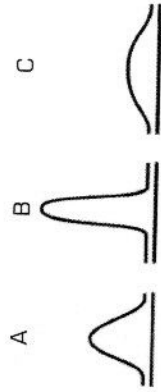
• **Alakmutató számok:**

- **CSÚCROSSÁG (KURTOSIS):** az eloszlás alakját vertikálisan leíró mutatószám. A normális eloszlás csúcossága nulla. Pozitív érték esetén az eloszlás ehhez képest csúcsosabb, negatívnál laposabb (lásd 3.1. ábra). A felvehető értékeknek nincs felső vagy alsó határa.

A) normál eloszlás

B) csúcsos, azaz pozitív értékkel rendelkezik

C) lapos, azaz negatív értékkel rendelkezik



3.1. ábra. Normál, csúcsos és lapos eloszlások

- **FERDESÉG (SKEWNESS):** az eloszlás horizontális alakját leíró mutatószám. Pozitív érték esetén az eloszlás jobbra ferde (jobbra dől), míg negatív érték esetén balra ferde (balra dől) (lásd 3.2. ábra). A felvehető értékeknek nincs felső vagy alsó határa.

A) pozitív, vagy jobbra ferde eloszlás

B) negatív, vagy balra ferde eloszlás

Amint azt említettük, a csúcosság és a ferdeség esetében a felvehető értékeknek nincs felső vagy alsó határa, azonban a kutatónak el kell döntenie, hogy mikor tér el egy eloszlás a normál eloszlástól, amely a normalitás témaköréhez vezet el bennünket. A normalitást, illetve a ferdeség és a csúcosság megállapítására a fejezet

¹ Mivel az SPSS minden adatbázist mintának tekint, ezért automatikusan korrigált szórást számol.



3.2. ábra. Jobbra és balra ferde eloszlás

és különösképp az esettanulmány során kitérünk, azonban jelenleg a következő szabályokat fogalmazzuk meg. Egyrészt a normál eloszlás esetében az átlag, a medián és a módusz ugyanoda esik, azaz egyenlők. Mindazonáltal sok eloszlás egy-szere ferde és csúcsos, ezért az átlag, medián, a módusz egymáshoz való viszonya csálóka lehet. Másrészt a normáloszlás-görbe és a hisztogram sokat segít a normalitás eldöntésében, ugyanakkor a megítésés helyessége jelentős mértékben függ a kutató tapasztaltságától. Harmadrészt, amennyiben a csúcosság és/vagy a ferdeség mutatók értéke meghaladja a +1 vagy -1 értéket, akkor ez azt jelenti, hogy az adott eloszlás különbözik a normál eloszlástól. Negyedrész, a ferdeség és a csúcosság értéke, illetve azok standard hibájának a hányadosa nem haladhatja meg a $\pm 2,58$ -at, illetve szigorúbb esetben az $\pm 1,96$ -ot. Ez utóbbi szabály részletes kifejtését lásd az esettanulmánynál.

• **Egyéb mutatók:**

- **ÖSSZEG (SUM):** az elemek összege.
- **ESETSZÁM (NUMBER OF CASES):** a megfigyelt elemek száma.
- **MINIMUM:** a legkisebb elem.
- **MAXIMUM:** a legnagyobb elem. A minimum és a maximum megjelenítésekor kiszűrhetjük a rögzítési hibák egy részét (például egy elvileg 1-7 tartományú skálán 9-es a maximumértéket találunk).

Összefoglalva megállapíthatjuk, hogy az egyváltozós elemzések bármilyen adatbázisra s minden változóra alkalmazhatók, habár ismernünk kell a változók jellegét annak érdekében, hogy a megfelelő mutatószámot alkalmazzuk. Az egyváltozós elemzések használata a többváltozós elemzések szempontjából kifejezetten fontos, ugyanis a változók közötti bonyolult kapcsolatok elfedhetik a feltételek megsértésének „jeleit”, s így eredményeinkből helytelen következtetéseket vonhatunk le. Tehát az egyváltozós elemzések a többváltozós elemzések elvégzéséhez szükséges feltételeket is feltárja, ugyanis a program akkor is becsli a modellt és elvégzi a számításokat, ha alkalmazásának feltételei nem teljesülnek, ezért azok kielégítéséről a kutatónak kell megbizonyosodnia. Az egy- és többváltozós elemzések szempontjából legfontosabb előfeltételek 1. a normalitás, 2. a kiugró (OUTLIER) és hiányzó (MISSING) értékek 3. a szóráshomogenitás és 4. a linearitás vizsgálata. Ezen feltételek közül az esettanulmányban az első kettővel foglalkozunk részletesebben, az utóbbi kettő a későbbi fejezetekben kerül bemutatásra.

3.3. Egváltozós elemzés az SPSS-ben

Az SPSS-ben számos módon végezhetünk egyváltozós elemzést, azonban a fejezet során nem törekszünk a teljességre sem a módszerek, sem pedig az SPSS parancsokat illetően, hanem a célunk az, hogy egy lehetséges és viszonylag egyszerű, de átfogó egyváltozós adatelemzési stratégiát mutassunk be. A fejezet során az előző fejezetben elkészített tanulmányok4.sav fájlját használjuk, azonban ezúttal arra vagyunk kíváncsiak, hogy a tanulói teljesítményt a szülők mennyi zsebpénzzel honorálják a családi kassza nagyságától függetlenül. Ennek megfelelően az adatbázis a következő kérdésre adott válaszokkal egészült ki:

Mennyi zsebpénzt kapsz a szüleidtől havonta?

A 12 válaszadó diák az alábbi válaszokat adta a havi zsebpénzét illetően: 10 000, 15 000, 20 000, 25 000, 30 000, 35 000, 5000, 5000, 20 000, 20 000, 40 000, 20 000, amely adatokat a zseb változó tartalmazza. Ezt az új adatbázist „tanulmányok5” néven mentjük el (lásd 3.3. ábra).

kod	info	mal	nye	oszt	csop	zseb	
1	ab1	5	11	1	1	10000	
2	ab2	5	12	1	1	15000	
3	ab3	5	10	1	1	20000	
4	ab4	4	13	2	1	25000	
5	ab5	5	4	15	2	1	30000
6	ab6	4	4	14	2	1	35000
7	ab7	3	4	17	3	1	5000
8	ab8	4	3	16	3	1	5000
9	ab9	3	3	18	3	2	20000
10	ab10	2	3	16	1	2	20000
11	ab11	3	2	17	2	2	40000
12	ab12	2	2	18	3	2	20000

3.3. ábra. Tanulmányok5.sav

A statisztikai elemzés során a következő kérdésekre keressük a választ:

- Hány diák adott választ a kérdésre?
- Mekkora a tanulók között a minimális zsebpénz?
- Mekkora a tanulók között a maximális zsebpénz?
- Mekkora a tanulók átlagos zsebpénze?
- A szülők összesen mennyi zsebpénzt adtak a tanulóknak?
- Mekkora a szórás, azaz a tanulók zsebpénze átlagosan mennyivel tér el az átlagos zsebpénztől?

Az egyváltozós elemzés mutatószámainak nagy része az „ANALYZE” menü „REPORTS” és a „DESCRIPTIVES STATISTICS” almenüpontjaiban található. A két menüpont közötti különbséget az elemzés célja határozza meg, ha gyors vizsgálat ké-

szül, akkor a „REPORTS”, ha leíró statisztika prezentációs elemekkel (grafikonokkal, ábrákkal), akkor pedig a „DESCRIPTIVES STATISTICS” menüpont az ajánlott, azonban ezek a végeredmény szempontjából természetesen nem különböznek.

Az „ANALYZE” menüponton belül a következő egyváltozós elemzésekkel foglalkozunk:

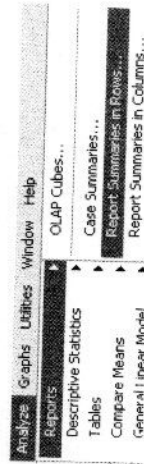
- ANALYZE
- REPORTS (jelentések)
- REPORT SUMMARIES IN ROWS (összesítő jelentés sorokban)
- REPORT SUMMARIES IN COLUMNS (összesítő jelentés oszlopokban)
- DESCRIPTIVES STATISTICS (leíró statisztika)
- FREQUENCIES (gyakoriság)
- DESCRIPTIVES (leíró elemzések)

3.3.1. Összefoglaló jelentés sorokban (Reports/Report Summaries in Rows)

Ebben az egyváltozós elemzésben egy vagy több változóra alkalmazhatjuk a helyzet, szóródási, alak és egyéb mutatószámokat úgy, hogy mindegyik változóra – amennyiben több változót jelöltünk meg – külön kiszámolja azokat.

A „REPORT SUMMARIES IN ROWS” egyváltozós elemzést az SPSS-ben a következő parancssorral érhetjük el:

ANALYZE/REPORTS/REPORT SUMMARIES IN ROWS (3.4. ábra)



3.4. ábra. Összesítő jelentés sorokban (Report Summaries in Rows)

A „REPORT SUMMARIES IN ROWS” párbeszédpanelben a „zseb” változót a fekete nyíl (4) segítségével vigyük át a „DATA COLUMN” dobozba (3.5. ábra), majd kattintsunk a „SUMMARY” gombra, ahol a szükséges statisztikai mutatószámokat találhatjuk (lásd 3.6. ábra).

Néhány opció és jelentése

Format: ez az opció lehetőséget ad arra, hogy megváltoztassuk a változó nevét úgy, ahogy azt az outputban szeretnénk látni, illetve további módosításokat tehetünk például a változó karakterszámára vonatkozóan.
OPTIONS: a hiányzó értékre tehetünk megkötéseket.
LAYOUT: a készülő jelentés oldalaira tehetünk beállításokat (pl.: oldalszám stb.).

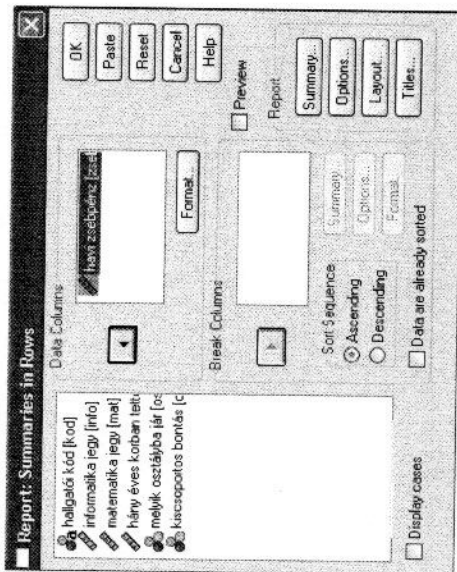
TITLES: a készülő jelentés címére tehetünk beállításokat.

DISPLAY CASES: a tényleges adatokat is mutatja és nem csak a kért mutatószámokat. Komolyabb méretű adatbázis folytonos változóinál használatos az őrítési listaméret miatt nem célszerű.

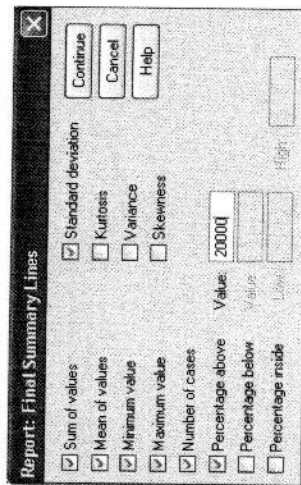
ASCENDING: az adatokat növekvő sorrendbe rakja.

DESCENDING: az adatokat csökkenő sorrendbe rakja.

DATA ARE ALREADY SORTED: jelöljük be, ha az adataink már előre csoportosítva vannak, ha nincsenek, akkor ezt a kockát hagyjuk üresen.



3.5. ábra. Összesítő jelentés sorokban – változókiválasztás



3.6. ábra. Mutatószámok kiválasztása

A „REPORT: FINAL SUMMARY LINES” párbeszédpanelben válasszuk ki a szükséges mutatószámokat (3.6. ábra) és kattintsunk ez után a „CONTINUE” majd az „OK” gombra. A kapott eredményt a 3.2. táblázat mutatja.

Összesen 12 tanuló töltötte ki a kérdőívet (N). A szülők által adott zsebpénz havonta összesen 245 000 forint (GRAND TOTAL SUM). Ha a tanulók között ezt egyenlően osztották volna szét, akkor mindenki 20 417 forintot kapott volna havonta (MEAN). A legkisebb összeg 5000 forint (MINIMUM), a legnagyobb pedig 40 000 forint (MAXIMUM) volt. A 20 417 forintos átlagos értéktől a zsebpénzek átlagosan 10 967 forinttal térnek el (STDDEV). A minimum, a maximum, az átlag és a szórás értékek alapján megállapíthatjuk, hogy az értékek jelentős része az átlagtól egy szórótávolságra szóródik (20 417+10 967), azaz 9450 és 31 384 között és minden érték két szórás távolságon belül helyezkedik el normálosztást feltételezve². A 3.4. ábra alapján könnyen kiszámolható, hogy a 12 érték közül 8 1 szórás távolságon belül van. Továbbá, 20 000 forintnál nagyobb összegű zsebpénzt a válaszadók harmada kapott.

Havi zsebpénz	
Grand Total Sum	245 000
Mean	20 417
Minimum	5000
Maximum	40 000
N	12
StdDev	10 967
> 20 000	33,3%

3.2. táblázat. Eredmények (Output)

3.3.2. Összefoglaló jelentés oszlopokban (Reports/Report Summaries in Columns)

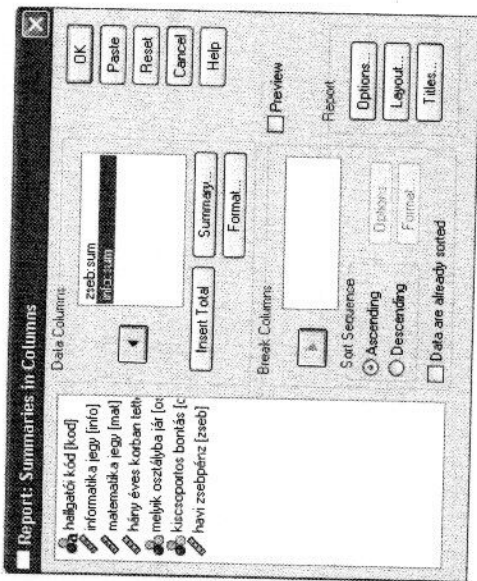
Ebben az egyváltozós elemzésben egy vagy több változóra csak egyetlen mutatószámot kérhetünk. Az előnye a „REPORT SUMMARIES IN ROWS” elemzéshez képest az, hogy a változók különböző mutatószámait egyszerre láthatjuk, s nem kell minden mutatószámot minden egyes változóra alkalmazni. Nézzünk egy példát az információkajegyek és a szülők által kiadott összes zsebpénz egyidejű vizsgálatára.

A „REPORT SUMMARIES IN COLUMNS” egyváltozós elemzést az SPSS-ben a következő paranccssorral érhetjük el:

ANALYZE/REPORTS/REPORT SUMMARIES IN COLUMNS

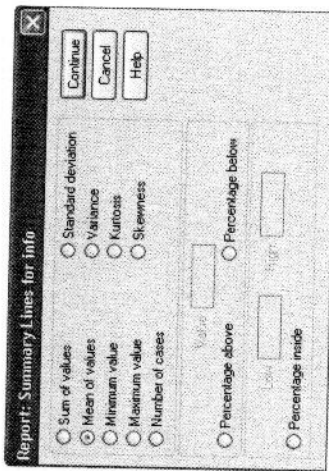
A „REPORT: SUMMARIES IN COLUMNS” párbeszédpanelben a „zseb” és az „info” változót a fekete nyíl 4 segítségével vigyük át a „DATA COLUMN” dobozba (3.7. ábra), álljunk ebben a dobozban a „ZSEB:SUM” feliratra. Látható, hogy az alapbeállítás az, hogy mindkét értéknek az átlagát számolja ki a program (ezt jelenti a :sum).

² A normálosztási görbe azt mutatja, hogy az értékek 68%-a az átlagtól 1,95%-a 2., és 99,5%-a 3 szórás távolságon belül helyezkedik el.



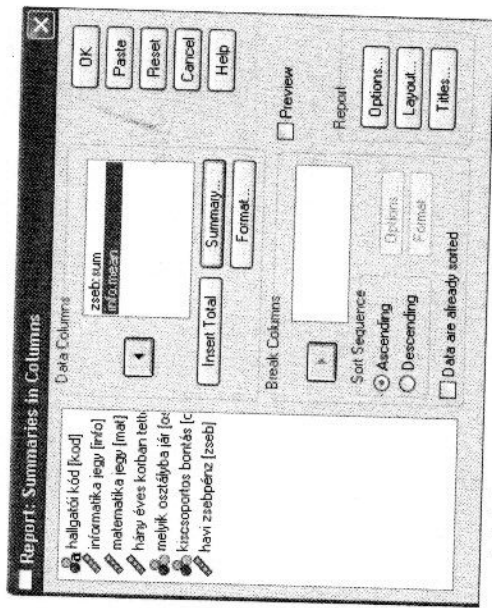
3.7. ábra. Összesítő jelentés oszlopokban – változókiválasztás

Ha azonban ezen változtatni szeretnénk, akkor kattintsunk a „SUMMARY” gombra, ahol meghatározhatjuk a kiszámolandó mutatószámokat. Mivel az informatikai jegyeknél az átlaga vagyunk kíváncsiak, ezért ezt változtatlanul hagyjuk. A zsebpénznél azonban az összegre van szükségünk, ezért válasszuk ki az INFO:SUM-ot és kattintsunk a SUMMARY menüre (3.8. ábra).



3.8. ábra. Mutatószámok kiválasztása

A „REPORT: SUMMARY LINES FOR INFO” párbeszédpanelben válasszuk ki a megfelelő mutatószámot (Mean of values), ezután kattintsunk a „CONTINUE” gombra. Látható, hogy a kiinduló ablakban az SPSS megváltoztatta a paramétereket (INFO:MEAN).



3.9. ábra. Összesítő jelentés oszlopokban – megváltozott paraméter

Havi zsebpénz/Sum	informatikajegy/Mean
Grand Total	245 000
	4

3.3. táblázat. Eredmények (Output)

A szűlők által összesen kiadott pénzeszeg tehát – ahogy az előbb is láttuk – 245 000 forint, míg az informatikajegyek átlaga 4.

3.3.3. Leíró statisztikák és gyakoriság (Descriptive Statistics/Frequencies)

A „DESCRIPTIVES STATISTICS” menüponton belül a DESCRIPTIVES elemzést akkor érdemes használni, ha:

- csak összesítő statisztikára van szükségünk, mert ebben az esetben nem készül gyakorisági tábla (ez lényegében egyezik a „REPORT SUMMARIES IN ROWS” egyváltozós elemzéssel, de itt az output esztétikusabb, mivel táblázat formájában jelenik meg az eredmény).
- a változók intervallum vagy arányskálán mértek.

A „DESCRIPTIVES STATISTICS” menüponton belül a FREQUENCIES elemzést akkor érdemes használni, ha:

- nem csak összesítő táblára van szükségünk, hanem gyakorisági táblára, s az eredmények grafikus ábrázolására is.
- a változók nominális vagy ordinális skálán mértek (a „FREQUENCIES”-ben az intervallum, illetve arányskálán mért változókat is vizsgálhatjuk).

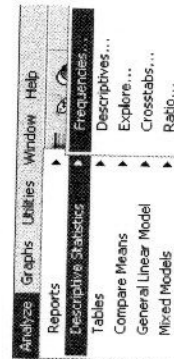
Mindkét elemzésnél egyszerre több változót is kiválaszthatunk. A könnyebb áttekinthetőség érdekében érdemes egyszerre csak azonos mérési skálájú változókat vizsgálni. A 3.4. táblázatban (Kovács, 2004 alapján) összefoglaljuk, melyik SPSS utasításban található meg a leíró statisztika eszközei a mérési skálák szerinti bontásban. Alapvető szabály, hogy a magasabb szintű mérési skálákon az alacsonyabb rendű skálákhoz rendelt eljárások mindig alkalmazhatók. A 3.4. táblázatban a D jelöli a „DESCRIPTIVES” és F a „FREQUENCIES” parancsot.

Mutatók	Nominális	Ordinális	Intervallum/arány
Helyzetmutatók	Módusz F	Medián, (Módusz) F	Átlag, (Medián, Módusz) F, D
Egyéb mutatók		Min., Max. F, D	Minimum, Maximum F, D
Szóródási mutatók	Gyakoriság, relatív gyakoriság F	Terjedelem F, D	Szórás, variancia F, D
Alakmutatók			Ferdesség, csúcsosság F, D
Grafikus megjelenítés	Gyakoriságra oszlop- és kördiagram F		Hisztogram F

3.4. táblázat. A skálátípusok és az alkalmazott mutatók összefüggése, valamint az SPSS-ben használt parancsok

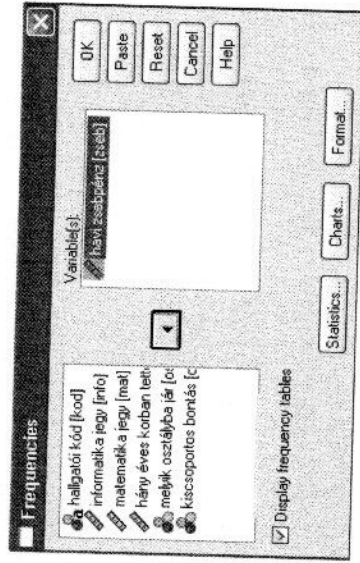
3.3.3.1. Gyakoriság (Descriptive Statistics/Frequencies)

A „FREQUENCIES” egyváltozós elemzést az SPSS-ben a következő úton érhetjük el: ANALYZE/DESCRIPTIVE STATISTICS/FREQUENCIES (3.10. ábra)



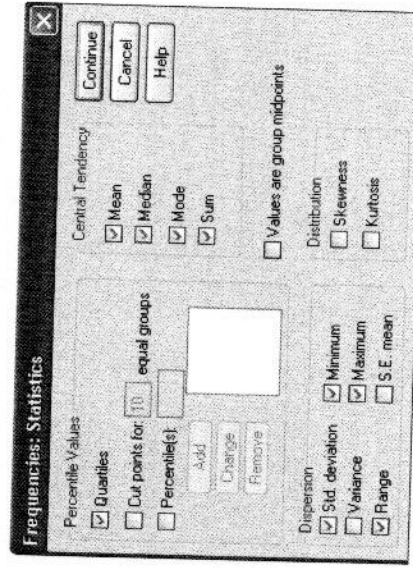
3.10. ábra. Gyakoriság (Frequency)

A „FREQUENCIES” párbeszédpanelben a „zseb” változót a fekete nyíl (4) segítségével vigyük át a „VARIABLE(S)” dobozba (3.11. ábra), majd kattintsunk a „STATISTICS” gombra.



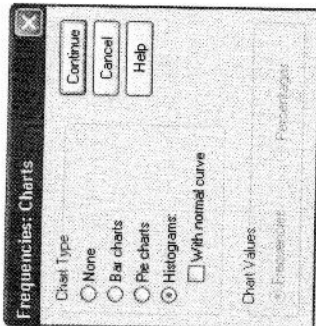
3.11. ábra. Változókiválasztás

A „FREQUENCIES: STATISTICS” párbeszédpanelben válasszuk ki a megfelelő mutatókat (3.12. ábra), majd kattintsunk a „CONTINUE”-ra.



3.12. ábra. A mutatók kiválasztása

A CHART opció kiválasztásával meghatározható, hogy milyen grafikai megjelenítést alkalmazunk (3.13. ábra). Jelen esetben kiválasztjuk a hisztogramot, majd kattintsunk a „CONTINUE”-ra, s azután az „OK” gombra.



3.13. ábra. Ábratípus kiválasztása

A mutatószámokat a 3.5. táblázatban láthatjuk, amelyek közül csak azokat tárgyaljuk, amelyek a 3.2 táblázatban nem szerepeltek.

- a táblázat az érvényes eseteken kívül mutatja a hiányzó esetek (MISSING) számát is, amely ebben az elemzésben nincs.
 - a medián (MEDIAN) 20 000 forint, azaz az esetek felében ennél az összegnél kevesebb pénzt adnak a szülők, míg az esetek másik felében többet.
 - a módusz (MODE) 20 000 forint, a leggyakrabban előforduló érték, amelyet leginkább nominális skálák esetén elemzünk, azonban értelmezhető ebben az esetben is, azaz arányskála esetén. Az átlag, a medián és a módusz közeli értékei alapján azt állapíthatjuk meg, hogy az eloszlás nem pontosan normál eloszlás, de nagyon jól közelíti azt.
 - a terjedelem (RANGE) 35 000 forint, azaz a legmagasabb és a legalacsonyabb összegű zsebpénzek közötti különbség.
 - az első kvartilis (PERCENTILES 25) 11 250 forint, azaz a válaszadók 25 százaléka kevesebb, mint 11 250 forintot kapott a szüleitől.
 - a második kvartilis (PERCENTILES 50 vagy medián) 20 000 forint, azaz az esetek felében ennél az összegnél kevesebb pénzt adnak a szülők, míg az esetek másik felében többet.
 - a harmadik kvartilis (PERCENTILES 75) 28 750 forint, azaz a válaszadók 75 százaléka kevesebb, mint 28 750 forintot kapott a szüleitől.
- A 3.6. táblázat a gyakorisági táblát mutatja, amelynek értelmezése a következő:
- az első oszlopban a válaszadók által megjelölt zsebpénzértékek láthatók (a folytonos skála ellenére szerencsére kevésféle értéket azonosítottunk).
 - a második oszlop (FREQUENCY) a különböző összegű zsebpénzek gyakoriságát mutatja (abszolút gyakoriság).
 - a harmadik oszlop (PERCENT) a százalékos megoszlást mutatja az összes megkérdezett, a negyedik oszlop (VALID PERCENT) pedig az érvényes válaszadók százalékában (relatív gyakoriság).
 - az utolsó oszlop (CUMULATIVE PERCENT) pedig a kumulált gyakoriságot mutatja,

például a 25 000 forinthez tartozó 75 százalékos kumulált gyakoriság azt jelenti, hogy a válaszadók 75 százaléka maximum 25 000 forintot kap.

Statistics

	Valid	Missing
N	12	0
Mean	20416,67	
Median	20000,00	
Mode	20000	
Std. Deviation	10966,547	
Range	35000	
Minimum	5000	
Maximum	40000	
Sum	245000	
Percentiles	25	
	50	
	75	

3.5. táblázat. Eredmények (Output)

havi zsebpénz

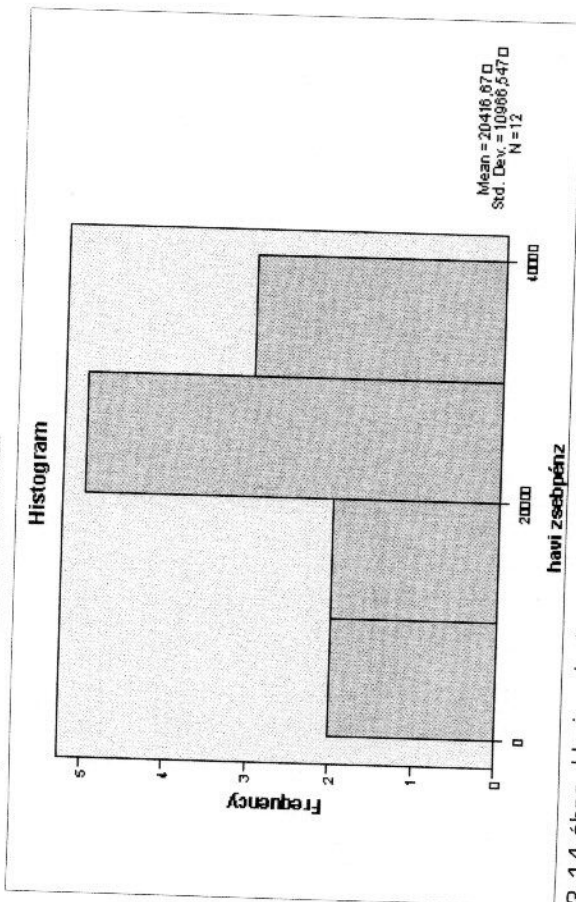
Valid	Frequency	Percent	Valid Percent	Cumulative Percent
5000	2	16,7	16,7	16,7
10000	1	8,3	8,3	25,0
15000	1	8,3	8,3	33,3
20000	4	33,3	33,3	66,7
25000	1	8,3	8,3	75,0
30000	1	8,3	8,3	83,3
35000	1	8,3	8,3	91,7
40000	1	8,3	8,3	100,0
Total	12	100,0	100,0	

3.6. táblázat. Havi zsebpénz gyakorisági táblázat

A fentiekben leírtakat a következő hisztogram szemlélteti (3.14. ábra).

A hisztogram az egyik legszemléletesebb és ennek megfelelően a leggyakrabban alkalmazott módja egy eloszlás ábrázolásának, amely egy változó értékeinek előfordulási gyakoriságát mutatja. A gyakoriság grafikus ábrázolásával a válaszok eloszlásának alakját vizsgálhatjuk meg. A hisztogram alkalmas mind folyamatos,

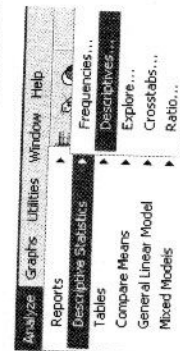
mind díszkrét változók ábrázolására. A fő különbség a hisztogram és az oszlopdiagram között abban rejlik, hogy a hisztogram oszlopai – az oszlopdiagramtól eltérően – mindig összeérnek, ezáltal is kifejezve az értékek folytonosságát. A hisztogram alapján az előbbi megállapításunk, mely szerint a változó normáloszlást követ, nem látszik teljesülni, hanem ez egy balra ferde (jobbra dől), azaz negatív ferde és enyhén csúcsos eloszlásnak tűnik.



3. 14 ábra. Havi zsebpénz-gyakorisági táblázat

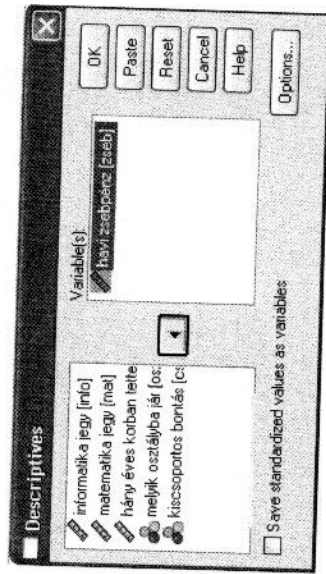
3.3.3.2. Leíró statisztikák (Descriptive Statistics/Descriptives)

A „DESCRIPTIVES” egyváltozós elemzést az SPSS-ben a következő módon érhetjük el: ANALYZE/DESCRIPTIVE STATISTICS/DESCRIPTIVES (3.15. ábra).



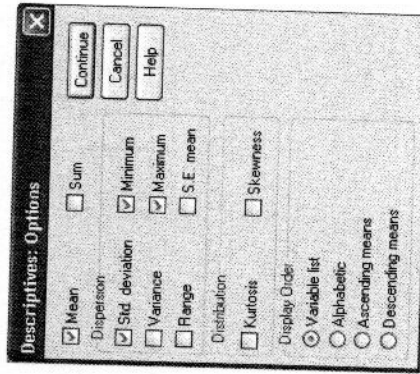
3. 15. ábra. Leíró statisztikák

A „DESCRIPTIVES” párbeszédpanelben a „zseb” változót a fekete nyíl (4) segítségével vigyük át a „VARIABLE(S)” dobozba (3.16. ábra), majd kattintsunk az „OPTIONS” gombra.



3. 16. ábra. Változó kiválasztása

A „DESCRIPTIVES: OPTIONS” párbeszédpanelben válasszuk ki a megfelelő mutatókat (3.17. ábra), azaz az átlagot, a szórást, a minimumot és a maximumot. A párbeszédpanel alján a „display order” alatt látható részben azt állíthatjuk be, hogy a változókat – amennyiben több van – milyen sorrendben szeretnénk látni. A következő 4 opció közül lehet választani, 1. olyan sorrendben, ahogy az adatbázisban szerepel (variable list), 2. ábécésorrendben (Alphabetic), 3. az átlagok növekvő (Ascending means), illetve 4. csökkenő sorrendjében (Descending means). Ezután kattintsunk a „CONTINUE”, s azután az „OK” gombra.



3. 17. ábra. A mutatók kiválasztása

A mutatószámok értelmezése és a végeredmény természetesen megegyezik az előző példában leírtakkal.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
havi zsebpénz	12	5000	40000	20416,67	10966,547
Valid N (listwise)	12				

3.7. táblázat. Eredmények [Output]

Az adatok megoszlásának grafikus ábrázolására gyakran alkalmaznak a pontfelhő diagramot (Scatterplot), amely alkalmas a kiugró és extrém értékek ábrázolására is. Próbáljuk ezt ki most mi is!

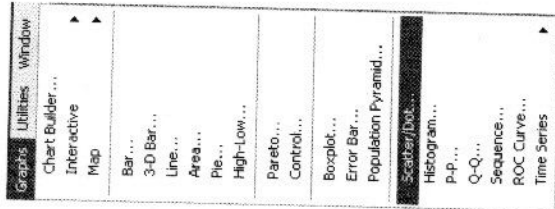
A „tanulmányok5.sav” fájlt egészítsük ki a 3.18. ábrán látható módon, a 11. sor-számú hallgató esetében a zsebpénzt módosítsuk 80 000 forintra (a „VARIABLE VIEW” nézetben a „VALUES” tulajdonságnál a 8-as értékhez előzetesen rendelt 40 000 forintot meg kell változtatni 80 000 forintra), majd mentjük a fájlt „tanulmányok6.sav” néven.

kod	info	mat	nye	oszt	csoport	zseb
1 ab1	5	5	11	1	1	10000
2 ab2	5	5	12	1	1	15000
3 ab3	5	5	10	1	1	20000
4 ab4	4	5	13	2	1	25000
5 ab5	5	4	15	2	1	30000
6 ab6	4	4	14	2	1	35000
7 ab7	3	4	17	3	1	5000
8 ab8	4	3	16	3	1	5000
9 ab9	3	3	18	3	2	20000
10 ab10	2	3	16	1	2	20000
11 ab11	3	2	17	2	2	80000
12 ab12	2	2	18	3	2	20000

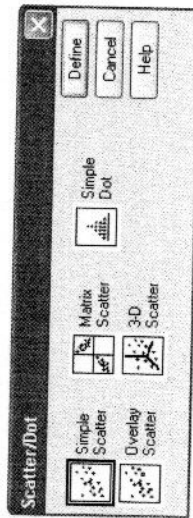
3.18. ábra. A módosított adatfájl

Most ábrázoljuk a zsebpénzt a sorszám függvényében. Hívjuk be a „GRAPH” menüpontból a „SCATTER/DOT”-ot (pontdiagramot, 3.19. ábra).

Válasszuk a „SIMPLE” lehetőséget (példánknl ezt a legegyszerűbb használni) és nyomjuk meg az „DEFINE”-t (3.20. ábra).

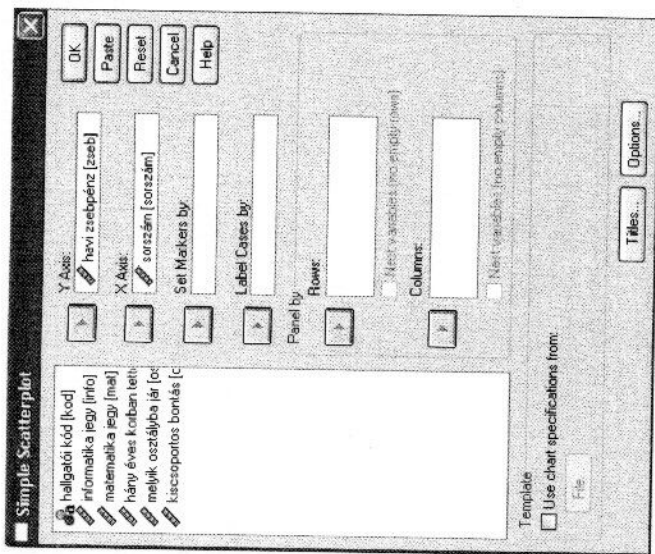


3.19. ábra. Pontfelhődiagram (Scatter graph) kiválasztása



3.20. ábra. Pontfelhődiagram (Scatter graph) típusának kiválasztása

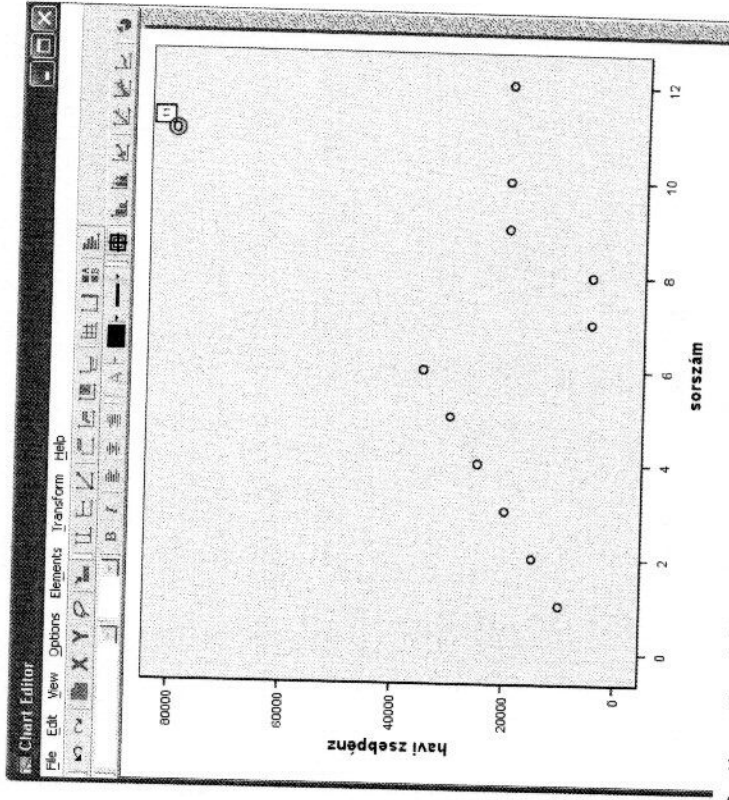
Általános szabályként elmondható, hogy az X (vízszintes) tengelyen a független, míg az Y (függőleges) tengelyen a függő változó szerepel. Ebben az esetben nem lehet egyértelműen eldönteni, hogy melyik a független és melyik a függő változó, ezért azt a logikát érdemes követni, hogy az elemzés szempontjából elsődleges változó (zsebpénz) lesz a függő változó (Y tengely), és ami szerint szeretnénk ezt vizsgálni (egykes tanulók), az lesz a független változó (X tengely). Ennek megfelelően az Y tengelyre a „zseb”, az X tengelyre pedig a „sorszám” változót vigyük be, majd nyomjuk meg az „Ok” gombot (3.21. ábra).



3.21. ábra. A változók kiválasztása a tengelyek szerint

Az eredményeket a 3.22. ábra mutatja.

Az ábra elemzése során észrevehetjük, hogy egy kiugró érték van, amelyik nagyon magas a többihez képest, ez pedig a 11. hallgató 80 000 forintos zsebpénze. A CHART EDITOR-ban a célkereszt ikon segítségével kiválaszthatjuk az elemeket, s ebben az esetben a program kiírja a sorszámot.



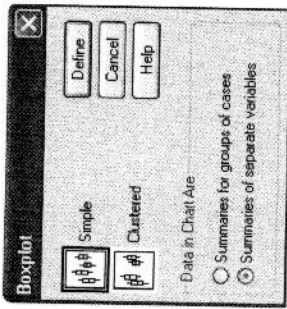
3.22. ábra. Pontfelhődiagram a tanulók sorszama és a havi zsebpénz szerint

Dobozdiagram (Boxplot)

A pontfelhődiagramnál professzionálisabb, ha a dobozdiagram (boxplot) segítségével határozzuk meg a kiugró elemeket. Amennyiben ez valóban kiugró elem, akkor ajánlatos kizárni a további elemzésekből – a már bemutatott „DATA/SELECT CASES” menüpont segítségével –, mert torzítaná az eredményeinket.

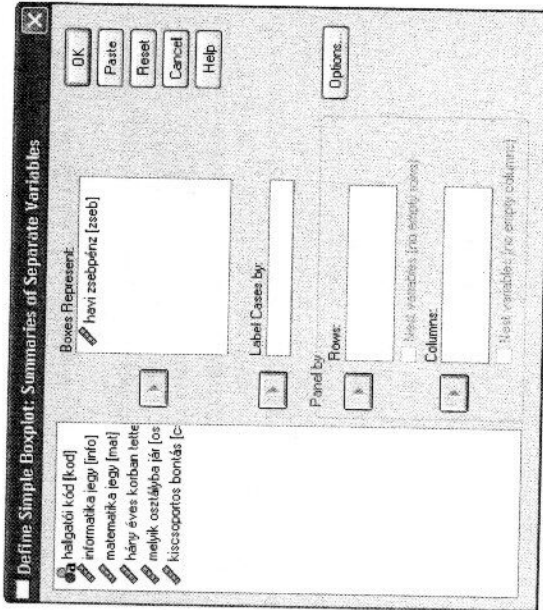
A „GRAPH/BOXPLOT” menüponttal olyan ábrát készíthetünk, amelyről egyértelműen kiderül, hogy van-e kiugró elemünk. A „GRAPH/BOXPLOT” menüponton belül két opció van, a chart típusa, amely lehet „SIMPLE” és „CLUSTERED”, illetve az adatok típusa (Data in Chart are), amely lehet eset (Summaries for groups of cases) vagy pedig változó (summaries of separate variables – (3.23 ábra), amelyek összegében negyféle dobozra-lehetőséget hoznak létre. A „SIMPLE” chart típusban a változók szerinti elemzés a legegyszerűbb, mivel ebben az esetben egy vagy több változó eloszlása önmagában kerül bemutatásra, amelyet a 3.25 ábrán látni is fogunk. Az esetek szerinti ábrázolás azt jelenti, hogy az adott változót (pl. jövedelem) egy másik változó kategóriái (pl. nem, férfi és nő) szerint ábrázoljuk. A változók szerinti elemzésnél a vizsgált változó az X tengelyen van, míg az esetek szerinti opcióban, a kategóriák vannak az X tengelyen és a vizsgált változó az Y-on.

A „CLUSTERED” opcióban a változók szerinti elemzés azt jelenti, hogy legalább két változót ábrázolunk egyszerre (pl. jövedelem és kiadás) egy másik változó (pl. nem) kategóriái szerint. A „CLUSTERED” opcióban az esetek szerinti elemzés a legösszetettebb, ahol habár egy változó eloszlását ábrázoljuk, azonban két változó szerinti bontásban.



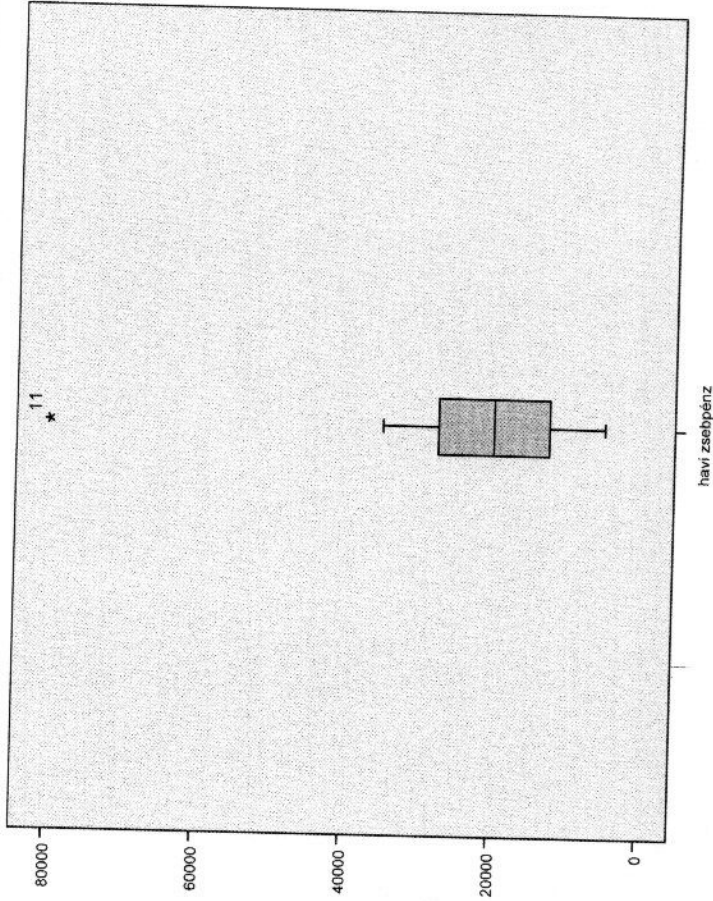
3.23. ábra. Boxplot diagram típusának kiválasztása

A megjelenítendő (BOXES REPRESENT) ablakba tegyük be a nyíl segítségével a „zseb” változót, és kattintsunk az Ok gombra (3.24. ábra).



3.24. ábra. Változó kiválasztása

A Boxplot diagramot a 3.25. ábra mutatja, amely megegyezik a pontfelhoábra tartalmával, különösképp a kiugró értékre vonatkozóan.



3.25. ábra. A havi zsebpénz változó Boxplot diagramja

A 3.25 ábrán szürke téglalapok mutatják az alsó és felső kvartilis közötti távolságot, és a doboz közepén lévő vízszintes fekete vonal pedig a mediánt. A dobozról felfelé és lefelé irányuló vonalak hossza az interkvartilis terjedelem (ami a felső és alsó kvartilis különbsége) másfélszerese. A függőleges vonalak két végén lévő vízszintes vonalak jelzik azt a terjedelmet, amelyen belül minden megfigyelésnek lennie kellene, a normál eloszlás. Ha az adott megfigyelés a doboz szélétől számított 1,5-3 interkvartilis terjedelemben van, akkor kiugró értéknek (jele: O) tekintjük, míg ha a megfigyelés három interkvartilis terjedelmen kívül van, akkor az extrém érték (jele: *). Ebben az esetben a 11. sorszámu elem extrém érték. A fenti elemzések alapján megállapítható, hogy a zsebpénzváltozó az extrém érték kizárása esetén normál eloszlású.

Esettanulmány

Az egyváltozós elemzések mélyebb tárgyalásához nagyobb adatbázisra van szükség. Vegyük most ismét elő az 1. fejezet végén látott gyógyszerárak példát. A 3.8. táblázat a gyógyszerárak forgalomszámaira vonatkozó leíró statisztikákat tartalmazza. Annak ellenére, hogy kényes kérdéstről van szó, a 172 kitöltőből 161 válaszolt rá. A forgalom átlagértéke 9,35 millió forint, a medián pedig 8 millió forint (vagyis az átlag nagyobb, mint a medián), ami arra utal, hogy az eloszlás bal oldali aszimmetrikus (jobbba hosszabban elnyúló), s ezt támasztja alá a ferdeségi mutató (SKEWNESS) is, amelynek értéke 1,496 (nagyobb, mint 1). Mint irányelv megállapítható, hogy ha a ferdeség értéke több mint kétszer nagyobb, mint a standard hibája, akkor a szimmetrikusság nem feltételezett.

Statistics

Havi átlagos forgalom2 (ezer Ft)		
N	Valid	Missing
Mean	9349,8820	161
Median	8000,0000	11
Mode	8000,00	
Std. Deviation	6321,429	
Variance	4E+007	
Skewness	1,496	
Std. Error of Skewness	,191	
Kurtosis	3,967	
Std. Error of Kurtosis	,380	
Range	41994,00	
Minimum	6,00	
Maximum	42000,00	
Sum	1505331	

3.8. táblázat. A gyógyszerárak forgalmi adatainak leíró statisztikái

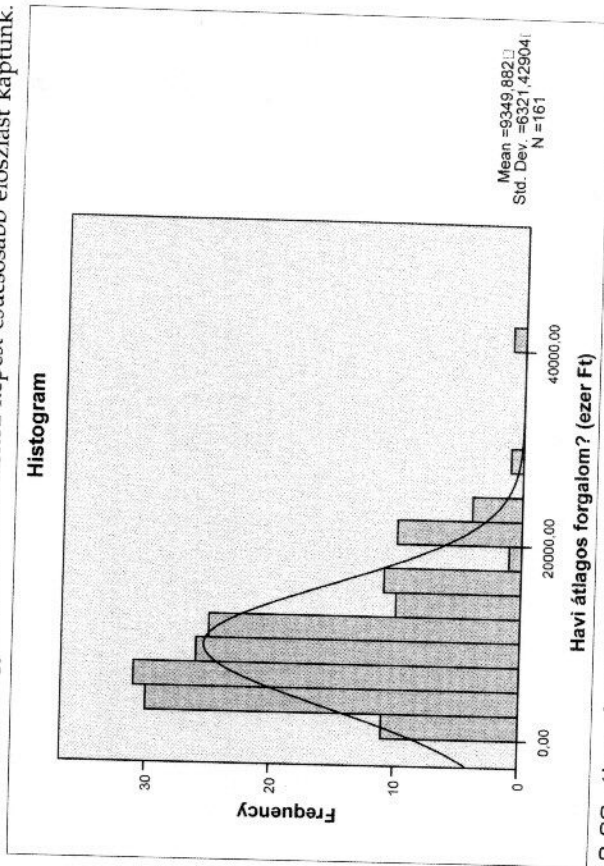
Az adatbázisban találhatóunk meglehetősen alacsony és magas értékeket. Például a havi minimum 6 ezer forintos forgalomérték meglehetősen alacsonynak tűnik, amellyel kapcsolatban eredetileg úgy véltük, hogy ez adatbeviteli hiba, azonban az ellenőrzés után ez nem volt igazolható, hanem úgy tűnik, hogy ez válaszadói hiba lehet (a kérdőív kitöltője nem olvasta el alaposan a kérdést, és ezer forint helyett millióban válaszolt). Hasonlóan matematikusnak tűnik a maximumérték, azaz a havi 42 millió forint, bár ez elvileg lehetséges.

A továbbiakban tekintjük át a fejezet elején említett feltételeket, amelyek közül a normalitással, a kiugró és a hiányzó értékek elemzésével és kezelésével foglalkozunk.

Normalitásvizsgálat

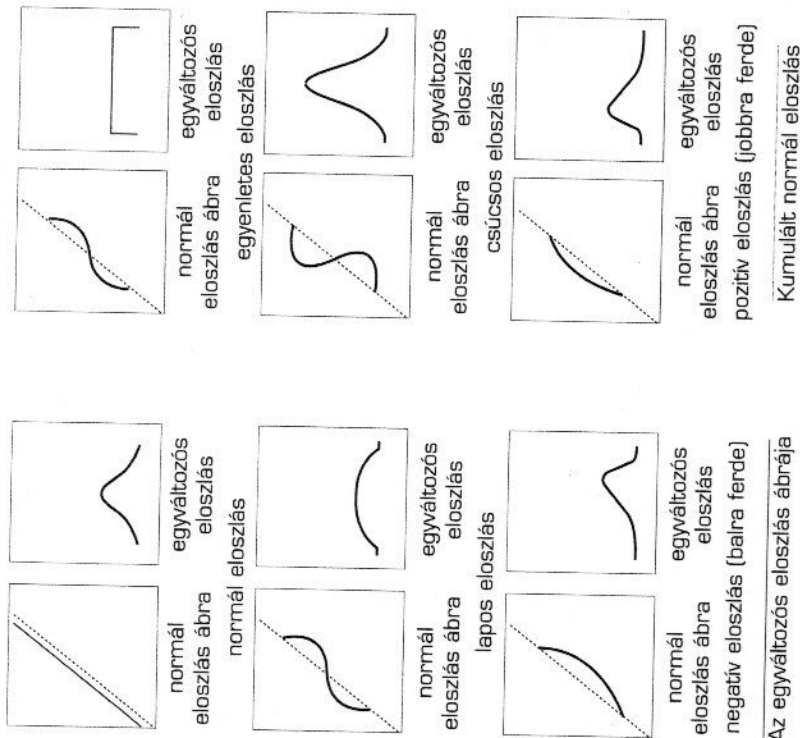
Számos elemzés legalapvetőbb feltétele a normalitás, amely egy metrikus változó eloszlásának alakjára, illetve annak a normális eloszlással való egyezőségére utal. Ha a normális eloszlástól való eltérés nagy, akkor nagyon sok – többek között az F és a t – statisztikai próba nem alkalmazható, így eredményeinket nem tudjuk értelmezni. Ahogy az egyváltozós elemzéseknel van egyváltozós normalitás, a többváltozós elemzéseknel többváltozós normalitásról beszélünk. Ha egy változóra teljesül a többváltozós normalitás, akkor az egyváltozós normalitás is teljesül, azonban sajnos ennek fordítottja nem igaz. Tehát az egyváltozós normalitás vizsgálata szükséges, de nem elégséges a többváltozós normalitás garantálásához.

Az előbbieken már bemutatott hisztogram (3.26. ábra), illetve annak normális eloszlással való összevetése segít a normalitás kérdésének eldöntésében. (A 3.14. ábrán bemutatott hisztogram vizuális képe az alacsony esetszám miatt is tűnhet torzítottnak, így az ilyen esetekben a hisztogram normális eloszlással való összevetése nem célszerű.) Jelen példánkban az esetszám alapján teljesen indokolt az adatok eloszlását hisztogramon ábrázolni, ami megerősíti a KURTOSIS pozitív értékét, azaz, hogy a normális eloszláshoz képest csúcsosabb eloszlást kaptunk.



3.26. ábra. A gyógyszerárak forgalmának hisztogramja

A hisztogramhoz hasonlóan két másik diagramtípus is alkalmas a normalitás vizsgálatára. Az egyik az úgynevezett stem-and-leaf (szár-levél) diagram, a másik pedig a normális eloszlás ábra (normal probability plot). Mindkét az ANALYSE



3.30. ábra. Eloszlástípusok

A normalitás vizsgálatára alkalmas tesztek közül a két leggyakrabban alkalmazott próba a Kolmogorov-Smirnov- és a Shapiro-Wilk-próba, amelyeket az EXPLORE menüpont alapbeállításaként kiszámol. A próbák szignifikanciaszintje kis (kevesebb mint 30 megfigyelésből álló) mintáknál nem használható, míg nagy minták (1000 fő felett) esetében igen érzékeny. A próba nullhipotézise az, hogy a változó nem normál eloszlású és a szignifikanciaszint alapján (3.9. táblázat) az adatok eloszlása szignifikánsan különbözik a normáloszlástól, tehát ez azt jelenti, hogy a változó nem normál eloszlású.

A normalitás teszteléskor mind a grafikai, mind pedig a statisztikai vizsgálatok elengedhetetlenek. A normalitás feltételének megsértése esetén adattranszformáció végrehajtása szükséges, amely segít normál eloszlásúvá alakítani az adatokat.

Az adattranszformáció a változók módosítását jelenti. Ennek elsődleges célja, hogy az elemzés feltételeinek sérülését korrigáljuk, illetve a változók közötti

Tests of Normality

Havi átlagos forgalom? (ezer Ft)	Kolmogorov-Smirnov ^a		Shapiro-Wilk	
	Statistic	df	Statistic	df
	,125	161	,898	161
			Sig.	Sig.
			,000	,000

a. Lilliefors Significance Correction

3.9 táblázat. Normalitás próbák

kapcsolatot javítsuk. Az adattranszformáció két csoportját különböztetjük meg: egyrészt az elméleti, másrészt a tapasztalati transzformációt. Az elméleti adattranszformáció alapja az adatok jellege, míg a tapasztalati megoldás az adatok elemzésének következménye. Tisztában kell lennünk azszal, hogy a változótranszformációra sajnos nincs egyetlen járható út, hanem ez számos próbálkozás alapul (trial-and-error), még a tapasztalt kutatók esetében is. A kutatónak több transzformációt kell alkalmaznia ahhoz, hogy a legmegfelelőbbet megtalálja, és habár az itt leírt szabályok segítenek, semmiképp nem mondják meg a jó megoldást. Továbbá, amint azt az előbb említettük, az adattranszformáció nemcsak a változó eloszlását, de a jelentését is módosítja, ezért az adattranszformáció alkalmazása sok esetben igen problematikus.

Az adattranszformáció alapvető szabályai a következők: 1. a változó átlag-szórás aránya legalább 4 kell, hogy legyen, hogy a folyamat eredményre vezetessen; 2. ha több változóról van szó, akkor a kisebb átlag-szórás arányút alakítsuk át. 3. a transzformációt – a heteroszkedaszticitás³ kivételével – a független változóra kell alkalmazni; 4. a homoszkedaszticitás⁴ és a linearitás⁵ nem teljesülése esetén mind a független, mind pedig a függő változót át kell alakítani; 5. a transzformáció megváltoztatja a változó jelentését is, például a logaritmus-transzformáció esetén a változó jelentése arányváltozás.

A nem normális eloszlás két gyakori jellemzője a lapos, illetve a ferde eloszlás. Lapos eloszlásnál az inverz (például $1/X$), a negatív ferdeségnél a négyzetgyök, míg a pozitív ferdeségnél a logaritmus-transzformáció a legcélszerűbb módszer. Ennek megfelelően esetünkben valószínűleg először a logaritmus-transzformációt kell alkalmazni. Mindazonáltal ki kell hangsúlyozni, hogy a nem normál eloszlás valamelyik másik feltétel megsértésének eredménye, ezért azokat kell először vagy a normalitással egyidejűleg megvizsgálni.

³ Lásd később a homoszkedaszticitásról.

⁴ A szóráshomogenitás azt jelenti, hogy a függő változónak azonos szórással kell rendelkeznie a független változó különböző szintjei mellett. Ha ez nem teljesül, akkor ezt heteroszkedaszticitásnak nevezzük. Ezzel bővebben többek között a variancia és a diszkriminációelemzésnél foglalkozunk.

⁵ Ennek kifejezésére szolgál, hogy egy adott, például regressziós modell olyan értékeket becsül, amelyek egy egyenesre esnek. Egy adott példán ez azt jelenti, hogy a jövedelem 1000 Ft-os növekedése minden esetben az élelmiszerre való kiadás 500 Ft-os növekedését vonja maga után. A linearitással részletesen többek között a diszkriminációelemzésnél találkozunk.

A kiugró értékek vizsgálata

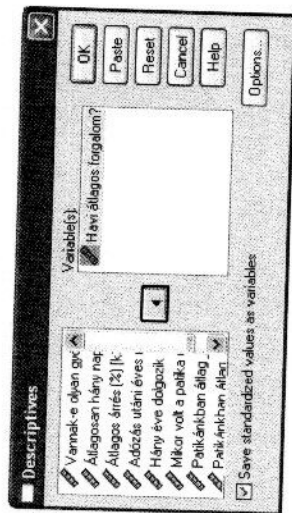
A kiugró és extrém érték olyan megfigyelés, amely egyértelműen elkülönül a többi megfigyeléstől. A kiugró értékeknek négy típusa van: 1. adatbeviteli vagy kódolási hiba, amelyeket az adattisztítás során kell felismerni; 2. különleges, egyedi esemény, amelyről a kutatónak kell eldöntenie, hogy megtartja vagy kizárja; 3. különleges esemény „látható” indok nélkül, azaz olyan értékek, amelyekre a kutatónak nincs magyarázata. Ezeket az eseteket rendszerint kizárjuk. 4. kombinációs érték, azaz a megfigyelés minden változó szerint az „elfogadható” tartományba esik, azonban a változók kombinációjában kiugró értéké válik. A kiugró értékek vizsgálatakor elsődlegesen azt a kérdést kell feltennünk, hogy mi legyen a kiugró értékek kiválasztására vonatkozó küszöbérték, amelyeket az előzőkben részletesen bemutatunk a boxplot diagram kapcsán. Mindazonáltal a küszöbérték valószínűleg elemzésről elemzésre valamelyest változni fog a kutató döntésének függvényében.

A kiugró értékek feltárása után felmerül a kérdés, hogy mi történjen ezekkel az értékekkel. A problémára a szakirodalom számos megoldást kínál, azok kizárása és megtartása mellett egyaránt érvelve. Ezek az értékek egyrészt többletinformációt hordoznak magukban, megmutatva a sokaság egy elkülönülő vetületét, másrészt viszont olyan, a sokaságra nem jellemző értékekről van szó, amelyek súlyosan torzítják a kapott eredményeinket. Ezen értékek megtartása akkor indokolt, ha a megfigyelések valós szegmensét képezhetik a sokaságnak, és így növelik a minta általánosíthatóságát. Ha ezeket a megfigyeléseket kizárjuk, a kutató valószínűleg javítja az elemzések statisztikáit, azonban csökkenti a minta általánosíthatóságát. Mindazonáltal ha egy megfigyelés több változó szempontjából is kiugró értéket mutat a fentebb említett módszerek alapján, akkor érdemes lehet két- és többváltozós elemzéseket végezni. A kiugró értékek vizsgálatának első, leggyakrabban alkalmazott módszere a standardizált értékek képzése (3.31 ábra), amely egy-olyan változó létrehozását jelenti, ahol minden egyes megfigyelésből kivonjuk az átlagot és a maradékot elosztjuk a szórás és a mintanagyság hányadosával, amelyet standard hibának is nevezünk. Ennek az új standardizált változónak az a jellegzetessége, hogy az átlagértéke 0 és szórása 1, amely a változók közötti összehasonlíthatóságot leegyszerűsíti. A standardizálás kifejezetten alkalmas különböző skálák egy alpra hozására.

Egy változó standardizálását a DESCRIPTIVES menüpontban végezhetjük el a „SAVE STANDARDIZED VALUES AS VARIABLES” parancs kiválasztásával.

A változó (forgalomzseb) és a standardizált változó (SAVE STANDARDIZED VALUES AS VARIABLES) parancs alapján az SPSS egy új változót hoz létre ZK37 néven. Az erre elvégzett gyakorlati elemzés táblázatát a következőkben láthatjuk (3.10. táblázat).

A 3.10. táblázatban belül az első két oszlop fontos számunkra. Az első oszlopban a standardizált érték, míg a másodikban ennek gyakorisága szerepel. A standardizált változó eloszlása egy olyan standard normal eloszlás, amely esetében bizonyos szabályok teljesülnek, mint például, hogy az esetek 68 százaléka egy,



3.31. ábra. Descriptives menüpont – standardizált értékek létrehozása

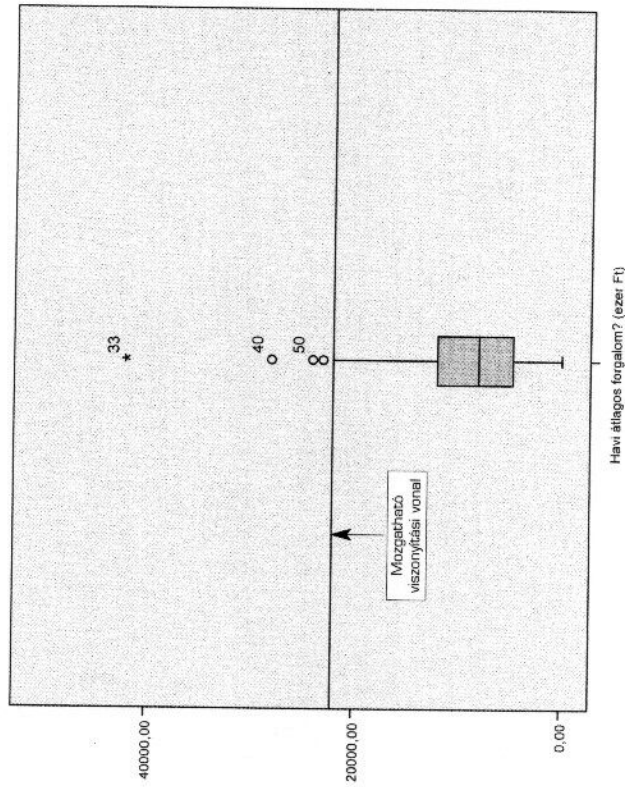
Zsomb: Havi állásig: forgalom? (lezer FI)

Value	1	2	3	4	5	6	Cumulative
-1,42813	1	0	0	0	0	0	0
-1,47512	1	0	0	0	0	0	1,2
-1,46879	1	0	0	0	0	0	1,9
-1,43182	1	0	0	0	0	0	2,6
-1,40780	1	0	0	0	0	0	3,3
-1,32086	1	0	0	0	0	0	4,0
-1,27343	1	0	0	0	0	0	4,7
-1,24179	1	0	0	0	0	0	5,4
-1,14633	1	0	0	0	0	0	6,1
-1,15090	1	0	0	0	0	0	6,8
-1,08778	1	0	0	0	0	0	7,5
-1,00450	3	1,7	1,9	0	0	0	8,1
-97286	3	1,7	1,9	0	0	0	8,9
-95704	1	0	0	0	0	0	9,6
-92540	1	0	0	0	0	0	10,3
-90999	1	0	0	0	0	0	11,0
-88213	1	0	0	0	0	0	11,7
-86031	0	4,7	5,0	18,3	0	0	14,3
-84985	1	0	0	0	0	0	15,0
-78300	1	0	0	0	0	0	15,7
-78139	2	2,9	3,1	23,6	0	0	18,9
-73557	1	0	0	0	0	0	19,6
-68817	6	2,9	3,1	25,5	0	0	22,8
-66902	2	1,2	1,2	1,2	1,2	1,2	24,8
-60156	2	1,2	1,2	1,2	1,2	1,2	26,8
-55962	8	4,7	5,0	36,0	0	0	31,1
-50096	1	0	0	0	0	0	31,8
-48247	1	0	0	0	0	0	32,5
-46037	2	1,2	1,2	1,2	1,2	1,2	33,8
-37173	0	0	0	0	0	0	34,5
-29584	2	1,2	1,2	1,2	1,2	1,2	35,8
-21354	0	7,8	8,1	4,0	0	0	44,7
-13444	3	1,7	1,9	5,0	0	0	49,0
-05535	7	4,1	4,3	40,2	0	0	53,0
.02775	1	0	0	0	0	0	53,7
.10284	6	4,7	5,0	65,8	0	0	60,9
.91104	6	3,5	3,7	89,6	0	0	67,1
.20470	1	0	0	0	0	0	68,4
.41923	0	5,8	6,2	76,4	0	0	74,2
.48833	3	1,7	1,9	77,8	0	0	76,4
.57742	5	3,1	3,1	80,7	0	0	81,9
.73561	3	1,7	1,9	82,9	0	0	83,9
.83340	6	2,9	3,1	85,7	0	0	87,1
.98872	1	0	0	0	0	0	88,4
1,05200	4	2,3	2,5	88,4	0	0	90,1
1,21019	1	0	0	0	0	0	90,1
1,38538	1	0	0	0	0	0	91,1
1,58476	6	3,5	3,7	93,3	0	0	93,3
2,16924	4	2,3	2,5	96,3	0	0	96,3
2,21753	3	1,7	1,9	96,9	0	0	98,9
2,86030	1	0	0	0	0	0	98,9
3,16436	1	0	0	0	0	0	99,4
Total	161	93,5	100,0	0	0	0	100,0
Missing	11	6,4	0	0	0	0	0
Total	172	100,0	0	0	0	0	0

3.10. táblázat. A forgalom standardizált változata

95 százaléka kettő, 99 százaléka pedig három szóráségsnyvi távolságon belül helyezkedik el. Ez alapján tehát lényegében minden értéknek az első oszlopban -3 és +3 között kell lennie, amely teljes mértékben teljesül, ugyanakkor ez az érték a mintanagysággal együtt változik. A standardizált normál eloszlás elmélete alapján tehát ha a mintanagyság 80 vagy annál kisebb, akkor a 2,5-nél nagyobb standardizált értékkel rendelkező megfigyelések kiugró értékek minősülnek, míg az ennél nagyobb minták esetében a 3-nál nagyobb értékek számítanak kiugró értékeknek. Ennek megfelelően a táblázat egy ilyen kiugró értéket tartalmaz (5,16).

A kiugró értékek elemzésének másik módszere a boxplot diagram (3.32. ábra), amely nemcsak a kiugró értékeket mutatja meg (például a 40-es vagy az 50-es számú megfigyelés⁶), hanem az extrém esetet is (33-as megfigyelés). A kiugró és a hiányzó értékek megállapításához segítséget nyújthat a CHART EDITOR-ban lévő vonatkoztatási egyenes (ADD A REFERENCE LINE TO THE Y AXIS).



3.32. ábra. Boxplot diagram

⁶ Az 50-es számú elem körül még tömörül több más elem is, de az SPSS alapbeállításaként nem írja ki mindégüknek a számát, mivel az olvashatatlanná tenné a számokat. Ez azt jelenti, hogy több, az 50-es esethez hasonló érték van, amelyet az ábra nem mutat, főleg amiatt, mert ez egy számkavalkádót okozna. Ha ezeket az eseteket törölni akarjuk, akkor érdemes azt a technikát alkalmazni, hogy az eset törlése után újra kémi a ábrát, és akkor a következő kiugró eset is látható lesz az ábrán.

Ez az egyenes a 20 000-es értéknél jelenik meg, azonban ez függőlegesen mozgatható, ha rákattintunk. Miután az egyenest a boxplot felső záróvonaláig toltuk, klikkeljünk rá kettőt ismételtlen a vonalra, ahol megjelenik az Y AXIS POSITION négyzetben az az érték, ahol a vonatkoztatási egyenes metszi az Y tengelyt. Nekünk pontosan erre az értékre van szükségünk (esetünkben 22 040 körül van). A kiugró adatok kiszűrésére használjuk a DATA/SELECT CASES/IF CONDITION IS SATISFIED/IF/ K37<22 040 parancssort, amelynek eredményeként a program a 22 040-nél nagyobb értékek mellett a hiányzó értékeket is áthúzza, vagyis figyelmen kívül hagyja az elemzés során (egészen addig, amíg vissza nem kapcsoljuk az összes elem opciót).

A hiányzó értékek vizsgálata

Hiányzó értékek nélkül ritkán létezik kutatás. Hiányzó adat alatt egyrészt a válaszadón kívüli okokat (például adatgyűjtés, illetve adatbevitel), másrészt pedig a válaszadón által kiváltott okokat (például válaszmegtagadás) értjük, azonban még előbbieket határolja ismert, az utóbbiaké csak ritkán. Ugyanakkor a hiányzó adatok kezelése rendkívül fontos, ugyanis ez meghatározza az adatok általánosíthatóságát. A kutatónak azonosítani kell azt a folyamatot, amely a hiányzó értékekhez vezet, ugyanis csak ezáltal lehet kiválasztani a kezelésükre legmegfelelőbb módszert. A kutatónak a hiányzó értékek elemzésekor fel kell ismernie, hogy a hiányzó értékek véletlenszerűen fordulnak-e elő vagy valamilyen szabály szerint, illetve mennyire gyakoriak. A hiányzó értékekkel leginkább akkor célszerű foglalkozni, ha ezek számottevő mennyiségben és valamilyen szabály (pattern) szerint fordulnak elő, ugyanis valószínű, hogy az adatok szerkezetét a hiányzó értékek befolyásolják. Ha a hiányzó értékekkel rendelkező megfigyelésekre nem alkalmazunk semmilyen megoldást, akkor ezeket az eseteket az SPSS alap esetben nem veszi figyelembe, azonban ez a mintanagyság radikális csökkenését is magával vonhatja. Ennek eldöntésére, hogy milyen megoldást alkalmazhatunk a hiányzó adatokra, a kutatónak meg kell állapítania a hiányzó adatok véletlenségének szintjét. Ehhez ugyanazon változó hiányzó adatokkal és anélkül való változatát hasonlítják össze. Ha a két változat között szignifikáns különbség van, akkor a hiányzó adatok nem véletlenszerűek, vagy csak bizonyos szinten véletlenszerűek. Ennek jelölése MAR (Missing at Random). Ez azt jelenti, hogy Y hiányzó értékei X-től és nem Y-tól függenek. A másik – optimális – eset, ha a hiányzó értékek teljesen véletlenszerűek, amelynek jelölése MCAR (Missing Completely at Random), amely esetben a változók egyes – hiányzó értékkel és anélkül – változatai között nincs szignifikáns különbség. A két típus közötti különbség abban rejlik, hogy ha bármilyen szabályszerűséget megfigyelhetünk a hiányzó értékek eloszlásában, akkor a hiányzó értékek nem véletlenszerűek.

Mit jelent ez a gyakorlatban? Például tegyük fel, hogy a háztartás jövedelme változóinkban számos hiányzó érték van, ugyanis a válaszadók nem voltak hajlandók ezt megadni. A hiányzó értékek vizsgálatát azzal kezdjük, hogy a háztartás jövedelme változót két csoportra bontjuk, az egyik az összes (azaz az

értékes és a hiányzó) értéket, a másik pedig az érvényes megfigyeléseket tartalmazza, és ezekben a csoportokban vizsgáljuk a nemek arányát. Amennyiben azt találjuk, hogy a hiányzó értékek véletlenszerűek mind a férfiak, mind a nők esetében, azonban sokkal magasabb a férfiaknál, akkor valószínűsíthető, hogy a hiányzó értékek nem véletlenszerűek. Ennek következtében még ha a hiányzó értékek véletlenszerűek is, a nem hatással lesz a háztartási jövedelem értékeinek az eloszlására, s ez kihat a minta általánosíthatóságára is. Teljesen véletlenszerű hiányzó értékekről akkor beszélhetünk, ha a háztartási jövedelem adatai véletlenszerűen hiányoznak egyenlő arányban a férfiaknál és a nőknél.

A hiányzó értékek kezelésére különböző megközelítéseket alkalmaznak; egyrészt 1. a fent említett eljárást, amely szerint egy változóból két csoportot képzünk, az egyik tartalmazza a megfigyelt és a hiányzó adatot, a másik csak a megfigyelt adatot. Amennyiben a változó metrikus, t-próbát alkalmazhatunk. Másrészt 2. ha a változókat dichotomizáljuk, azaz a megfigyelt értéket 1-gyel, a hiányzó értéket 0-val jelöljük, és korrelációelemzést végzünk. A nem szignifikáns, illetve alacsony korreláció azt jelenti, hogy a hiányzó adatok teljesen véletlenszerűek, míg a szignifikáns magas ennek ellentettjére utal. Harmadrészt 3. létezik véletlenszerűségi teszt a változók hiányzó értékeinek a vizsgálatára, amely ha nem talál szignifikáns különbséget, akkor az adatok teljesen véletlenszerűek (MCAR). A hiányzó adatok kezelésére a következő eljárások alkalmazhatók:

1. „**Teljes eset**” eljárás, azaz csak az összes adattal rendelkező esetet vesszük figyelembe, amely szinte minden statisztikai elemzésben alapbeállítás. Ez az eljárás MCAR hiányzó értékek esetében alkalmazható, és akkor előnyös, ha a hiányzó adatok száma a mintaelemszámhoz képest kicsi. Észre kell vennünk azonban, hogy az eljárás csökkentheti az elemzés általánosíthatóságát.
2. **A hiányzó esetek vagy változók törlése** – A módszer alkalmazása akkor célszerű, ha a hiányzó értékek nem véletlen eloszlásúak. Az eljárás során a kutatónak először fel kell mérnie a hiányzó esetek mértékét, illetve azok elhelyezkedését az esetekben és a változókban. Általában igaz az a megfigyelés, hogy a hiányzó esetek többsége a változók/ esetek kis csoportjában található, ezért ezek kizárásával jelentősen csökkenthető a számuk. Mindazonáltal, nincs elfogadott irányelv, hogy mennyi legyen a hiányzó értékek minimális/maximális száma. Továbbá a hiányzó értékek kizárása a függő változóban minden esetben indokolt, ugyanis ezek szükségtelenül növelhetik az elemzés magyarázó erejét⁷. A kutatónak tehát tisztában kell lennie, hogy a hiányzó érték „forrásának” (esetkizárás) megszüntetése vagy a változónak az elemzésből való kihagyása az előnyösebb.
3. **Imputációs eljárások** – Ennek a módszernek a lényege, hogy a hiányzó értékeket az adatbázis változóinak/eseteinek vagy más mintáknak az érvényes értékei alapján becsüljük. A módszer a minta érvényes értékeinek ismert

⁷ Ez főként akkor fordul elő, ha a kutató előzetes vizsgálatok nélkül pótolja a hiányzó értékeket, és ezt használja fel a további vizsgálatokhoz.

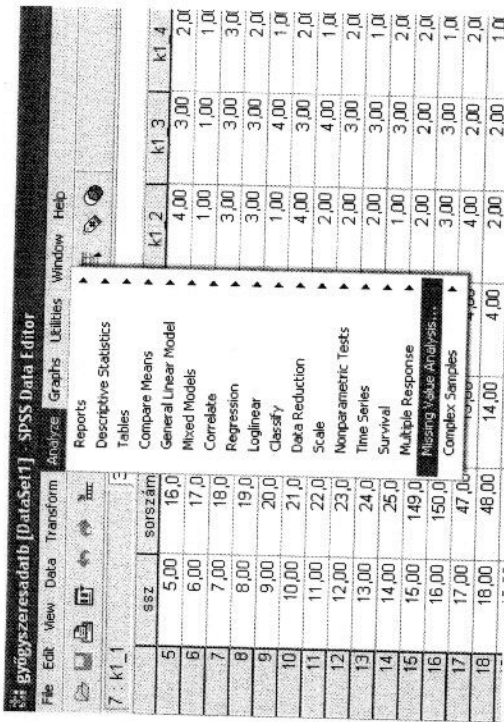
összefüggéseire támaszkodik, amelynek alkalmazása veszélyes lehet abból a szempontból, hogy elhitheti a kutatóval, hogy teljes adatbázissal dolgozik. Az eljárás kizárólag metrikus skálákra (például attitűd vagy jövedelem) alkalmazható és a következő fajtái vannak:

- a) Az „**Összes adat**” módszer valójában nem helyettesíti az adatokat, hanem csak azok eloszlási statisztikáit (például: átlag, szórás), illetve összefüggéseinek mutatóit (például: korrelációs együttható) az érvényes esetek alapján. Ez a módszer csak MCAR hiányzó értékeknél alkalmazható.
- b) **A hiányzó adatok pótlása**, amely valójában már pótolja a hiányzó értékeket, amelynek főbb formái a következők:
 1. **Az esetpótlás** azt jelenti, hogy például egy olyan háztartás adatait tudjuk ilyen módszerrel pótolni, illetve helyettesíteni egy másik háztartás adataival, amelyik vagy nem elérhető, vagy pedig túl sok hiányzó értékkel rendelkezik. Ilyen esetben az újonnan választott háztartás jellemzőinek azonosnak vagy hasonlóknak kell lennie a helyettesítettel.
 2. **Az átlagpótlás** esetén az adott változó hiányzó értékeit az érvényes esetek átlagával helyettesítjük. A módszer hátránya az adatok torzítottasága, ugyanis a variancia és a korreláció alulbecsült.
 3. **„Külő pótlás” módszere** (cold deck imputation) – A módszer hasonlós az az előző eljáráshoz, azonban a kutató külső forrásból vagy előzetes kutatásból származó konstans értékkel helyettesíti a hiányzó értékeket. A kutatónak meg kell bizonyosodnia arról, hogy a külső forrásból származó adatok sokkal érvényesebbek, mint egy belsőleg generált adat.
 4. **Regressziós pótlás** – A módszer a regresszióelemzést alkalmazza a hiányzó értékek pótlására. A módszer hátrányai a következők. 1. Az adatbázisban már meglévő összefüggéseket erősíti meg, ami csökkentheti az adatok általánosíthatóságát. A módszer emellett feltételezi, hogy az alkalmazott (hiányzó értékkel rendelkező) változó erősen korrelál más, a modellben szereplő változókkal. 2. Az eloszlás varianciája alulbecsült. 3. Előfordul, hogy a pótlások nem érvényes értéktartományok közé esnek, például adódhat 11-es érték a 10-es skálán.

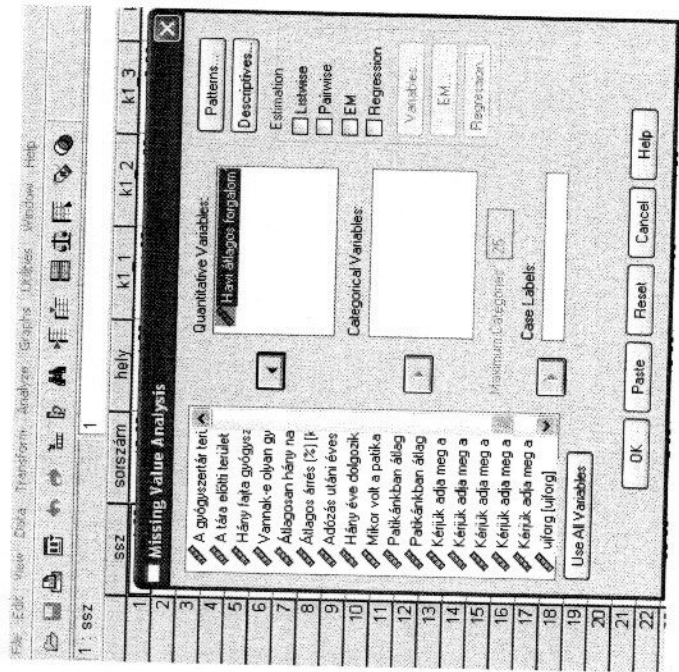
4. **Modelllezési eljárások.** Ezen módszerek alkalmazása akkor javasolt, ha a hiányzó értékek nem teljesen véletlenszerűek. Ennek egyik típusa az SPSS-ben használt EM eljárás, amely kétlépcsős iteratív módszer, és a hiányzó értékeket, statisztikákat becsli.

Ezek alapján próbáljuk meg pótolni adatbázisunkban a havi átlagos forgalom hiányzó adatait. Az SPSS-ben a MISSING VALUE ANALYSIS végzi a hiányzó értékek elemzését, amelyet a következő úton érhetünk el: ANALYSE/MISSING VALUE ANALYSIS (3.33. ábra).

A hiányzó eseteket tartalmazó változók kiválasztására, azaz annak eldöntésére, hogy mi az a határ amely fölött a hiányzó értékekkel foglalkozni kell, nincs egyértelmű szabály, hanem ez a kutató felelőssége.



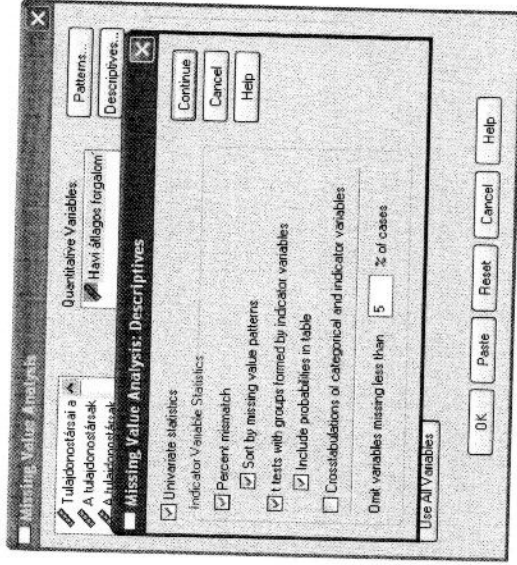
3.33. ábra. Hiányzó értékek elemzése menüpont elérése



3.34. ábra. Hiányzó értékek párbeszédpanel

A 3.34. ábra a változó bevitelét mutatja, amely esetünkben a havi átlagos forgalom (k37), azaz egy metrikus (intervallum- vagy arányskálán mért) változó (QUANTITATIVE VARIABLES). Amennyiben az elemezni kívánt változó nem metrikus, akkor az ez alatti dobozba kell bevinni az adott változó (CATEGORICAL VARIABLES).

Az elemzésen belüli beállításokat tekintve a DESCRIPTIVES menüpontban az első opció (UNIVARIATE STATISTICS) alapbeállítás, és az egyváltozós leíró statisztikákat mutatja. A második opció (PERCENT MISMATCH), a hiányzó értékek százalékos arányát számítja ki külön-külön az egyes változókra, illetve a változók interakciójára. A harmadik opció pedig az elemzésben résztvevő változók közötti t-próbákat mutatja (3.35. ábra).



3.35. ábra. Descriptive menüpont

A kapott 3.11. összefoglaló táblázat (UNIVARIATE STATISTICS) bemutatja az érvényes esetek számát, az átlagot, a szórást, illetve hiányzó esetek számát (11) és arányát (6,4%). Továbbá látható, hogy a változónak hat extrém értéke van magas kategóriában. Extrém érték alatt – mint ahogy a táblázat alatti magyarázatban is látható – az alsó kvartilisból kivont, illetve a felső kvartilishez hozzáadott interkvartilis terjedelem másfélszeresénél nagyobb értéket értjük.

Univariate Statistics

k37	N	Mean	Std. Deviation	Missing ^a	
				Count	Percent
	161	9349,8820	6321,42904	11	6,4
				Low	0
				High	6

^a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

3.11. táblázat. Egyváltozós statisztika

A DESCRIPTIVES menüpont második opciója alapján (PERCENT MISMATCH) láthatók (3.12. táblázat) az egyes változókban, illetve a változók interakciójában lévő hiányzó értékek aránya. Az újonnan bevitt két változó (k34 – a gyógyszer-fajták átlagos száma a készletben, k39 – adózás utáni éves nyereség) jobban megvilágítja a táblázat lényegét.

A 3.12. táblázat átlójában lévő értékek a hiányzó elemek arányát mutatják, amely rendre 6,40 százalék, 21,51 százalék és 29,07 százalék a k37-es, k34 és k39-es változóban, ahol az utóbbi aránya a legmagasabb. (Ez a szám könnyen ellenőrizhető, ugyanis 50 hiányzó értékünk és összesen 172 esetünk van.) Az átlón kívüli értékek az egyes változók közötti interakciójában megmutatkozó hiányzó értékek arányát mutatják, amelyek közül a k34-es és a k39-es változók interakciójában a legmagasabb, 27,33 százalék, azaz a két változó együttes elemzése esetén a teljes mintán belül az adatok több mint 27%-a hiányzik, míg a k34-es és k37-esé a legalacsonyabb (16,28 százalék).

Percent Mismatch of Indicator Variables.^{a, b}

	k37	k34	k39
k37	6,40		
k34	16,28	21,51	
k39	22,67	27,33	29,07

The diagonal elements are the percentages missing, and the off-diagonal elements are the mismatch percentages of indicator variables.

- a. Variables are sorted on missing patterns.
b. Indicator variables with less than 5% missing values are not displayed.

3. 12. táblázat. A hiányzó értékek százalékos aránya

A 3.13. táblázat a hiányzó értékek véletlenszerűségének egyfajta értékelését adja. A táblázat minden cellája hét értéket tartalmaz. Ezek (fentről lefelé): 1. a t-érték, az érvényes (A csoport) és a hiányzó értékeket (B csoport) tartalmazó sorváltozó oszlop változó szerinti összehasonlítását mutatja; 2. szabadságfok (df); 3. a t-érték szignifikanciaszintje; 4. az érvényes esetek száma (A csoport); 5. hiányzó esetek száma (B csoport); 6. az érvényes esetek átlaga; 7. hiányzó esetek átlaga.

A jobb felső cellában a k34-es változó A (érvényes érték) és B (hiányzó érték) csoportjának k39-es változó szerinti összehasonlítása alapján kapott t-érték 1,4, amelynek szabadságfoka 51,7 és szignifikanciaszintje 0,157. Az A és a B csoport mintanagyságai 107 és 17, és a csoportok átlagai 5554,6542 és 4152,3824. A táblázatban számos mutatószám segít annak eldöntésében, hogy a hiányzó értékek véletlenszerűek vagy sem. 1. Amennyiben a táblázat t-értékei szignifikánsak, akkor valószínűleg a hiányzó értékek nem véletlenszerűek. A táblázatban a t-ér-

tékek közül csak egy szignifikáns, ugyanakkor számos hiányzik. 2. Egy másik irányelv szerint amennyiben a t-értékek nagyobbak 2-nél, akkor a hiányzó értékek nem véletlenszerű eloszlásúak, és minél nagyobb mértékben térnek el a 2-től, annál kevésbé véletlenszerű a hiányzó értékek eloszlása. Ez szintén egy esetben fordul elő a táblázatban (2.7). 3. Egy harmadik megközelítés szerint amennyiben a szabadságfok jóval kisebb, mint a mintanagyság, akkor feltételezhető, hogy a két csoport varianciája különbözik egymástól, azaz a hiányzó értékek véletlenszerűek a mintában. Ez alapján a hiányzó értékek véletlenszerűnek tűnnek, azonban néhány celláról nincs információnk a hiányzó adatok alacsony száma miatt. Ezt mutatja a táblázat alatti megjegyzés (b), mely szerint azok a változók nincsenek feltüntetve a táblázatban, amelyeknél a hiányzó értékek aránya kisebb, mint 5 százalék.

Separate Variance t Tests^a

t	Variables Tested		
	k34	k37	k39
df		3	1,4
P(2-tail)		41,0	51,7
# Present		785	157
# Missing	137	136	107
Mean(Present)	0	27	17
Mean(Missing)	2712,0949	9366,8382	5554,6542
		9034,8519	4152,3824
t			
P(2-tail)			
# Present	136	163	124
# Missing	1	0	0
Mean(Present)	2718,8015	9311,8466	5362,4073
Mean(Missing)	1800,0000		
t	,1	2,7	
df	45,8	94,6	
P(2-tail)	,918	,008	
# Present	107	124	124
# Missing	30	39	0
Mean(Present)	2719,9720	9914,4355	5362,4073
Mean(Missing)	2684,0000	7395,9231	

For each quantitative variable, pairs of groups are formed by indicator variables (present, missing).

a. Indicator variables with less than 5% missing are not displayed.

3. 13. táblázat. A hiányzó értékek véletlenszerűségének értékelése – érvényes és hiányzó értékek közötti megfigyelések összehasonlítása

A hiányzó értékek véletlenszerű eloszlásának vizsgálatára a korrelációelemzés⁸ is alkalmas, amely azonban az adatok átkódolását igényli. Az átkódolás során az érvényes adatok 1-est kapnak, míg a hiányzóok értéke 0. Az átalakítás után az eredmények a következőképp alakulnak (3.14. táblázat).

Correlations

k34uj	Pearson Correlation	k34uj	1	k37uj	,442**	k39uj	,291**
	Sig. (2-tailed)		,000		,000		,000
	N	174	174	174	174	174	174
k37uj	Pearson Correlation	,442**	1	k39uj	,409**		
	Sig. (2-tailed)	,000	,000		,000		
	N	174	174	174	174		
k39uj	Pearson Correlation	,291**	,409**	1			
	Sig. (2-tailed)	,000	,000	,000			
	N	174	174	174			

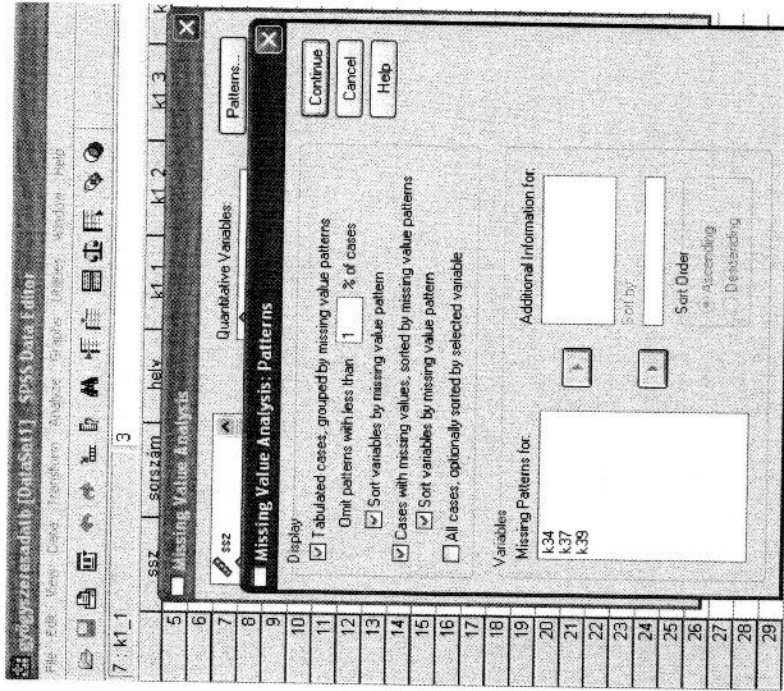
** Correlation is significant at the 0.01 level (2-tailed).

3.14. táblázat A hiányzó értékek véletlenszerűségének értékelése – dichotomizált változók korrelációja alapján

A 3.14. táblázat azt mutatja, hogy az egyes változók közötti korreláció közepes nagyságú és szignifikáns, azaz a változók hiányzó értéket jelentős mértékben összefüggnek egymással. Ennek megfelelően már egyáltalán nem alátámasztott, hogy a hiányzó értékek véletlenszerűek.

A MISSING VALUE ANALYSIS-en belül a „PATTERNS” menüpont (3.36. ábra) a hiányzó (és kiugró) értékek részletes „elhelyezkedését”, illetve a hiányzó értékek összesített statisztikáit mutatja. A 3.15. táblázatban az első oszlop az egyes eseteket, a második és a harmadik oszlop a hiányzó értékek számát és ezek arányát mutatja (a három változón belül). A többi oszlopban az S jelöli a hiányzó értékeket, a + pedig az extrém értéket (98, 105, 112 számú esetek). Ez a táblázat is azt a célt szolgálja, hogy van-e valamilyen rendszeresség (pattern – minta) a hiányzó adatok elhelyezkedésében. A 3.16. összesítő táblázat alapján az eddigi-ekhez hasonló adatok olvashatók le: például a legtöbb hiányzó érték a k39-es változóban van (29+10+10 = 49), s látjuk a változók interakciójában hiányzó értékek számát is.

⁸ A korrelációelemzés, két metrikus változó összefüggésének erősségét vizsgálja, amellyel a 6. fejezetben foglalkozunk részletesen.



3.36. ábra. Patterns menüpont

Missing Patterns (cases with missing values)

Case	Missing and Extreme Values Pattern ^a
1	33,3 S
15	1 33,3 S
28	1 33,3 S
38	1 33,3 S
45	1 33,3 S
49	1 33,3 S
51	1 33,3 S
52	1 33,3 S
58	1 33,3 S
62	1 33,3 S
77	1 33,3 S
85	1 33,3 S
92	1 33,3 S
96	1 33,3 S
97	1 33,3 S
101	1 33,3 S
102	1 33,3 S
105	1 33,3 S
113	1 33,3 S
123	1 33,3 S
129	1 33,3 S
132	1 33,3 S
133	1 33,3 S
134	1 33,3 S
142	1 33,3 S
153	1 33,3 S
161	1 33,3 S
170	1 33,3 S
172	1 33,3 S
94	2 66,7 S
44	2 66,7 S
20	2 66,7 S
26	2 66,7 S
53	2 66,7 S
56	2 66,7 S
124	2 66,7 S
126	2 66,7 S
30	2 66,7 S
137	2 66,7 S
24	1 33,3 S
91	1 33,3 S
111	1 33,3 S
112	1 33,3 S
39	1 33,3 S
115	1 33,3 S
122	1 33,3 S
87	1 33,3 S
127	1 33,3 S
7	1 33,3 S
130	1 33,3 S
88	1 33,3 S
86	1 33,3 S
138	1 33,3 S
146	1 33,3 S
155	1 33,3 S
157	1 33,3 S
88	3 100,0 S S S
4	3 100,0 S S S
89	3 100,0 S S S
16	3 100,0 S S S
71	3 100,0 S S S
84	3 100,0 S S S
108	3 100,0 S S S
100	3 100,0 S S S
31	3 100,0 S S S
78	3 100,0 S S S
81	2 66,7 S S

^a indicates an extreme low value, while + indicates an extreme high value. The format used is (01: 1 5(N)02: 03 + 1 5(N)01).
^b Cases and variables are sorted on missing patterns.

3.15. táblázat.
 A hiányzó és extrém értékek részletes bemutatása

Tabulated Patterns

Number of Cases	Missing Patterns ^a
105	
29	X
10	X X
17	X X
10	X X X

Patterns with less than 1% cases (2 or fewer) are not displayed.

- a. Variables are sorted on missing patterns.
- b. Number of complete cases if variables missing in that pattern (marked with X) are not used.

3.16. táblázat. A hiányzó értékek összesített bemutatása

A 3.13. táblázat azt valószínűsítette, hogy a hiányzó értékek véletlenszerű eloszlásúak, míg a 3.14. táblázat a magas korrelációk alapján ennek ellentettjét jelezte. A 3.15. és a 3.16. táblázatban számos olyan eset volt, ahol a hiányzó értékek két vagy mindhárom változóban együtt jelentkeztek (10 hiányzó érték a k34 és k39 változó esetében, és 10 mindhárom változó esetében), feltételezve a szabályszerűséget. Összességében kijelenthető, hogy a hiányzó értékek között van szabályszerűség, azaz nem véletlenszerűek.

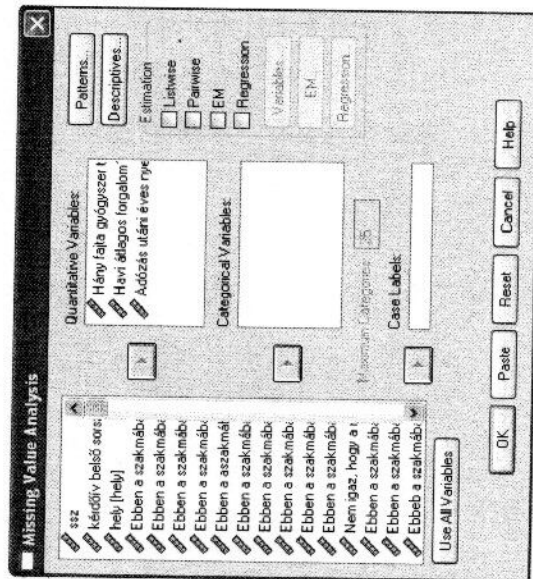
A hiányzó esetek kezelésére számos módszert tudunk alkalmazni. A „teljes eset” eljárás alapján 107 esetet tudnánk elemezni, azonban ez csak MCAR feltétellel mellett alkalmazható. Az esetkizárás alapján a fentiekben említett 20 esetből legalább 10 esetet kellene kizárni, mégpedig azt a 10-et, amelyik mindhárom változó esetében hiányzik. Míg a változó kizárásakor a k39-es változót zárnánk ki, amely természetesen függ a változó helyettesíthetőségétől, azaz van-e olyan változó, amellyel ez szorosan összefügg. A tíz összefüggő eset kizárása esetén látható, hogy a korreláció már nem szignifikáns a változók között (3.17. táblázat).

Correlations

	k34d	k37d	k39d
k34d			
Pearson Correlation	1	-.035	,131
Sig. (2-tailed)		,658	,095
N	164	164	164
k37d			
Pearson Correlation	-.035	1	,138
Sig. (2-tailed)	,658		,078
N	164	164	164
k39d			
Pearson Correlation	,131	,138	1
Sig. (2-tailed)	,095	,078	
N	164	164	164

3.17. táblázat. Korrelációs mátrix

A változók becslésére az SPSS négy eljárást kínál: az esetenkénti, a páros, az EM és a regressziós becslést, amelyek közül az EM eljárás a legelterjedtebb.



3.37. ábra. Becslési eljárások

Az EM becslés alapján kapott főbb táblázatok bemutatják a becslés utáni átlag-, illetve korrelációs értékeket, illetve az ehhez tartozó Little-féle MCAR próbákat, amelyek az adatok véletlenszerűségét igazolják a próba nem szignifikáns volta alapján.

EM Means^a

	k37	k39
k34		
Mean	2729,1379	9343,1222
N	4981	1776

a. Little's MCAR test: Chi-Square = 7,234, DF = 6, Sig. = ,300

3.18. táblázat. EM eljárással becsült átlagértékek

EM Correlations^a

	k37	k39
k34		
Mean	1	1
Std. Deviation	,309	,659
N	1	1

a. Little's MCAR test: Chi-Square = 7,234, DF = 6, Sig. = ,300

3.19. táblázat. EM eljárással becsült korrelációs értékek

Az esettanulmányban bemutattuk az egyváltozós elemzések főbb típusait, illetve az egyváltozós és számos többváltozós elemzésekhez szükséges feltételeket, úgymint normalitást, linearitást, szóráshomogenitást, kiugró és hiányzó értékek vizsgálatát. Ezeknek a mutatóknak, jelenségeknek a vizsgálatát rendkívül fontos, és habár sok esetben az elemzés szempontjából elenyészők, azonban az elemzés eredményét s ezáltal az azokból levont következtetések helyességét jelentős mértékben befolyásolják.

Összefoglalás

Ebben a fejezetben megismerkedtünk az egyváltozós elemzések technikájával. A példákon keresztül bemutattuk, melyek a kulcsfontosságú változók és az elemzések elvégzésének milyen feltételei vannak. Az olvasó ezen ismeretek elsajátítása után néhány alapvető mutatót keresztül átfogó képet tud adni egy adatbázisról, értelmezni tudja az egyváltozós elemzések során használt statisztikákat, s képes levonni a szükséges következtetéseket. Áttekintettük az egyváltozós és többváltozós elemzésekhez szükséges feltételeket: a normalitást, kiugró és hiányzó értékek vizsgálatát. Ezek elemzése rendkívül fontos, s bár sok esetben az elemzés szempontjából elenyésző a jelentőségük, a teljes elemzés eredményét s ezáltal az azokból levont következtetések helyességét jelentős mértékben befolyásolják.