

HUZSVAI LÁSZLÓ – VINCZE SZILVIA

SPSS-könyv

Seneca Books

2012

Minden jog fenntartva. Jelen könyvet vagy annak részleteit a Kiadó engedélye nélkül bármilyen formában vagy eszközzel reprodukálni, tárolni és közölni tilos.

Szerkesztette:

Dr. Huzsvai László

Írták:

© Dr. Huzsvai László

Dr. Vincze Szilvia

© Kiadó: SENECA BOOKS

ISBN:

978-963-08-5666-9

2012

TARTALOMJEGYZÉK

ELŐSZÓ	7
BEVEZETÉS	8
FILE MENÜ	9
READ TEXT DATA:	9
APPLY DATA DICTIONARY:	9
DISPLAY DATA INFO:	9
SZERKESZTÉS (EDIT) MENÜ	10
OPTIONS	10
NÉZET (VIEW) MENÜ	11
STATUS BAR	11
TOOLBARS	11
FONTS	11
GRID LINES	11
ADATOK (DATA) MENÜ	12
DEFINE VARIABLE	12
TEMPLATES	12
INSERT VARIABLE	12
INSERT CASE	12
GOTO CASE	12
SORT CASES	12
TRANPOSE	12
RESTRUCTURE	12
MERGE FILES	15
AGGREGATE DATA	16
ORTHOGONAL DESIGN	16
<i>Generate</i>	16
SPLIT FILE	17
SELECT CASES	17
<i>Nagy mennyiségű adat lekérdezése</i>	17
WEIGHT CASES	20
ÁTALAKÍTÁSOK (TRANSFORM) MENÜ	21
COMPUTE VARIABLE:	21
RANDOM NUMBER SEED:	21
RECODE:	21
CATEGORIZE VARIABLES:	22
RANK CASES:	22
AUTOMATIC RECODE:	22
RUN PENDING TRANSFORMS:	22
ELOSZLÁSOK	23
ANALÍZISEK	25
RIPORTOK	25
<i>OLAP Cubes</i>	25
<i>Case summaries</i>	29
<i>Report summaries in Rows</i>	30
<i>Report summaries in Columns</i>	30
LEÍRÓ STATISZTIKÁK (DESCRIPTIVE STATISTICS)	30

<i>Gyakoriságok (Frequencies...)</i>	30
<i>Descriptives</i>	32
<i>Explore</i>	32
<i>Keresztábrák (Crosstabs...)</i>	38
<i>Négy-mezős Chi2-próba függetlenség és homogenitás vizsgálatra</i>	39
CUSTOM TABLES.....	40
KÖZÉPÉRTÉK ÖSSZEHASONLÍTÁS (COMPARE MEANS)	40
<i>Középértékek (Means...)</i>	40
<i>Egy-mintás t-teszt (One Sample T Test...)</i>	41
<i>Egy-mintás z-próba</i>	42
<i>Két független minta középértékének összehasonlítása (Independent-Samples T Test...)</i>	43
<i>Két-mintás z-próba</i>	44
<i>Párosított t-próba (Paired-Samples T Test...)</i>	45
<i>Egy-énevezős variancia-analízis (One-Way ANOVA...)</i>	46
ÁLTALÁNOS LINEÁRIS MODELL (GENERAL LINEAR MODEL)	54
<i>Egy-változós variancia-analízis (Univariate...)</i>	56
<i>Többváltozós variancia-analízis (Multivariate...)</i>	57
KÍSÉRLETEK TERVEZÉSE ÉS ÉRTÉKELÉSE ÁLTALÁNOS LINEÁRIS MODELLEL	58
<i>ELMÉLETI ÁTTEKINTÉS</i>	58
EGY-TÉNYEZŐS VARIANCIA-ANALÍZIS AZ SPSS-BEN	63
A MODELL ÉRVÉNYESSÉGÉNEK VIZSGÁLATA	64
<i>Normalitás vizsgálat</i>	64
<i>Homogenitás vizsgálat</i>	66
<i>Kiugró értékek vizsgálata</i>	69
A VARIANCIA-ANALÍZIST KIEGÉSZÍTŐ KÖZÉPÉRTÉK ÖSSZEHASONLÍTÓ TESZTEK	72
<i>Kontrasztok</i>	72
<i>Szimultán vagy többszörös összehasonlító tesztek</i>	74
<i>Legkisebb szignifikáns differencia (LSD)</i>	76
<i>Newman-teszt</i>	76
<i>Bonferroni-teszt</i>	77
<i>Tukey-teszt, J.W. Tukey (1953)</i>	77
<i>H. Scheffé (1953) Scheffe-teszt</i>	78
<i>Dunnett-teszt</i>	78
<i>Student-Newman-Keuls próba</i>	81
<i>Duncan többszörös rang teszt (1955, 1965)</i>	81
ÁLTALÁNOS LINEÁRIS MODELLEK	86
<i>TOVÁBBI LEHETŐSÉGEK A GLM-BEN</i>	88
SZÁNTÓFÖLDI KÍSÉRLETEK TERVEZÉSE ÉS ÉRTÉKELÉSE	89
<i>KÍSÉRLETI ELRENDEZÉSEK</i>	91
EGY-TÉNYEZŐS KÍSÉRLETEK	99
<i>Teljesen véletlen elrendezés (CRD)</i>	99
<i>Véletlen blokk-elrendezés (RCBD)</i>	102
<i>Latin négyzet elrendezés</i>	104
<i>Latin téglalap elrendezés</i>	107
<i>Csoportosított elrendezés</i>	109
KÉT-TÉNYEZŐS KÍSÉRLETEK	113
<i>Véletlen blokk-elrendezés</i>	113
<i>Osztott parcellás (split-plot) elrendezés</i>	115
<i>Sávos elrendezés</i>	118
HÁROM- ÉS TÖBB-TÉNYEZŐS KÍSÉRLETEK	122
<i>Véletlen blokk-elrendezés</i>	122
<i>Kétszeresen osztott parcellás (split-split-plot) elrendezés</i>	124
<i>KOVARIÁNSOK ALKALMAZÁSA A LINEÁRIS MODELLBEN</i>	128
KORRELÁCIÓ- ÉS REGRESSZIÓSZÁMÍTÁS	133

KÉT-VÁLTOZÓS SZTOCHASZTIKUS KAPCSOLATOK.....	134
<i>ASSZOCIÁCIÓ.....</i>	<i>135</i>
<i>A - próba.....</i>	<i>136</i>
<i>Asszociáció és függetlenség -es táblában.....</i>	<i>136</i>
<i>A változók függetlenségének tesztelése.....</i>	<i>137</i>
<i>Az asszociáció mérése -es táblázat esetében.....</i>	<i>139</i>
<i>Asszociáció és függetlenség -s táblában.....</i>	<i>140</i>
<i>Az asszociáció mérése -s táblázat esetében.....</i>	<i>141</i>
<i>Nominális változókhoz tartozó asszociációs mutatók.....</i>	<i>141</i>
<i>Ordinális változókhoz tartozó asszociációs mutatók.....</i>	<i>141</i>
<i>Rangkorreláció.....</i>	<i>146</i>
<i>Vegyes kapcsolat.....</i>	<i>155</i>
<i>KÉT KVANTITATÍV VÁLTOZÓ KÖZÖTTI KAPCSOLAT ELEMZÉSE.....</i>	<i>155</i>
<i>Magas mérési szintű változók közötti kapcsolat vizsgálata.....</i>	<i>155</i>
<i>Pontdiagram.....</i>	<i>156</i>
<i>Lineáris korrelációs együttható.....</i>	<i>159</i>
<i>Korrelációs index.....</i>	<i>161</i>
<i>A lineáris korrelációs együttható meghatározása SPSS-ben.....</i>	<i>161</i>
<i>A regressziós egyenes.....</i>	<i>163</i>
<i>A legkisebb négyzetek módszere.....</i>	<i>164</i>
<i>A lineáris regressziószámítás menete.....</i>	<i>165</i>
<i>A lineáris függvény meghatározása.....</i>	<i>165</i>
<i>A korrelációs együttható és a determinációs együttható kiszámítása.....</i>	<i>167</i>
<i>A regresszió szignifikanciavizsgálata.....</i>	<i>169</i>
<i>A két változó összefüggésének szignifikanciavizsgálata.....</i>	<i>169</i>
<i>A regressziós egyenesből számított értékek hibája.....</i>	<i>170</i>
<i>A regressziós koeficiens statisztikai próbái.....</i>	<i>172</i>
<i>A regressziós koeficiens hibaszórása.....</i>	<i>172</i>
<i>A regressziós koeficiens konfidenciahatárai.....</i>	<i>172</i>
<i>A regressziós egyenlet konstans tagjának próbája.....</i>	<i>173</i>
<i>A korrelációs koeficiens statisztikai próbái.....</i>	<i>173</i>
<i>A LINEÁRIS REGRESSZIÓ ELVÉGZÉSE AZ SPSS-BEN.....</i>	<i>175</i>
TÖBBSZÖRÖS LINEÁRIS REGRESSZIÓSZÁMÍTÁS.....	184
<i>A STANDARD LINEÁRIS REGRESSZIÓS MODELL.....</i>	<i>184</i>
<i>Multikollinearitás.....</i>	<i>185</i>
<i>A multikollinearitás mérése.....</i>	<i>186</i>
<i>Autokorreláció.....</i>	<i>187</i>
<i>Az elsőrendű autokorreláció tesztelése.....</i>	<i>187</i>
<i>Heteroszkedaszticitás.....</i>	<i>189</i>
<i>A TÖBBSZÖRÖS LINEÁRIS REGRESSZIÓSZÁMÍTÁS LÉPÉSEI.....</i>	<i>189</i>
<i>A regressziós modell illeszkedésének vizsgálata.....</i>	<i>190</i>
<i>A paraméterek tesztelése.....</i>	<i>191</i>
<i>A becsült paraméterek jelentése.....</i>	<i>192</i>
<i>A reziduumok vizsgálata.....</i>	<i>192</i>
<i>KÉT FÜGGTLEN VÁLTOZÓS LINEÁRIS REGRESSZIÓELEMZÉS.....</i>	<i>193</i>
<i>A regresszió paramétereinek meghatározása kézi számítással.....</i>	<i>193</i>
<i>A regressziós paraméterek meghatározása az SPSS-vel.....</i>	<i>205</i>
<i>HÁROM FÜGGTLEN VÁLTOZÓS REGRESSZIÓANALÍZIS.....</i>	<i>208</i>
<i>NEMLINEÁRIS ÖSSZEFÜGGÉSEK VIZSGÁLATA.....</i>	<i>217</i>
<i>Lineárisra visszavezethető összefüggések vizsgálata.....</i>	<i>218</i>
<i>Logaritmikus regresszió.....</i>	<i>219</i>
<i>Exponenciális regresszió.....</i>	<i>226</i>
<i>Hatványkitevős regresszió.....</i>	<i>231</i>
<i>Parabolikus regresszió.....</i>	<i>237</i>
<i>Lineárisra nem visszavezethető összefüggések vizsgálata.....</i>	<i>241</i>
<i>Logisztikus függvény.....</i>	<i>241</i>
<i>A logisztikus függvény paramétereinek meghatározása.....</i>	<i>242</i>
ADATREDUKCIÓK.....	253

FŐKOMPONENS-ANALÍZIS.....	253
Korrelációs mátrix meghatározása.....	255
Az U sajátvektor mátrix és a sajátértékek (λ_j) meghatározása.....	256
Főkomponens együtthatók.....	256
Főkomponens változók.....	257
A főkomponens változók ábrázolása.....	258
A főkomponens súlyok meghatározása.....	259
Főkomponensek ábrázolása.....	262
A főkomponenssúlyok gyakorlati értelmezése.....	263
Főkomponens-analízis forgatóással.....	264
FAKTOR-ANALÍZIS.....	269
KATEGORIKUS FŐKOMPONENS-ANALÍZIS.....	269
NEM PARAMÉTERES PRÓBÁK.....	271
CHI-NÉGYZET TESZT.....	271
BINOMIÁLIS TESZT.....	272
RUNS TEST.....	273
EGYMINTÁS KOLMOGOROV-SMIRNOV TESZT (ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST).....	277
KÉT FÜGGETLEN MINTÁS TESZTEK (TWO INDEPENDENT SAMPLES TESTS).....	278
TÖBB FÜGGETLEN MINTÁS TESZT (K INDEPENDENT SAMPLES.....)	280
KÉT PÁRONKÉNT ÖSSZETARTÓZÓ MINTÁK TESZTJEI (2 RELATED SAMPLES.....)	280
K SZÁMÚ ÖSSZETARTÓZÓ MINTA TESZTJEI (K RELATED SAMPLES.....)	281
IDŐSOROK ANALÍZISE.....	283
TREND.....	284
RÖVID LEJÁRATÚ SZEZONÁLIS ÉS VÉLETLEN ÖSSZETEVŐK.....	284
A sorozat véletlenszerűségének vizsgálata.....	284
Periodogram-elemzés.....	286
Exponenciális simítás.....	286
A szezonális hatás felbontása.....	292
GRAFIKONOK.....	294
OSZLÓP DIAGRAMOK (BAR CHARTS).....	294
Egyszerű (Simple).....	294
Csoportosított (Clustered).....	295
Halmozott (Stacked).....	297
KÖRDIAGRAMOK (PIE CHARTS).....	298
KÉRDŐÍVEK TERVEZÉSE.....	301
KÉRDŐÍVEK KIÉRTÉKELÉSE.....	304
NOMINÁLIS TÍPUSÚ ADATOK KIÉRTÉKELÉSE.....	304
ORDINÁLIS TÍPUSÚ ADATOK KIÉRTÉKELÉSE.....	308
SKÁLA TÍPUSÚ ADATOK KIÉRTÉKELÉSE.....	310
TÖBBSZÖRÖS VÁLASZADÁSOK ELEMZÉSE 1.....	313
MAXIMUM K VÁLASZ ELEMZÉSE 2.....	319
GYAKORLÓ FELADATOK.....	320
FÜGGELÉK.....	322
AJÁNLOTT IRODALOM.....	331
GAUSS, CARL FRIEDRICH.....	333

ELŐSZÓ

A könyv megírásakor az egyik fontos célunk az volt, hogy a statisztikai és biometriai módszereket konkrét számítógépes környezetben mutassuk be, továbbá a módszerek elméleti elsajátításán túlmenően, azok számítógépen való helyes alkalmazását és a kapott eredmények tudományos igényű értelmezését is megismerje az olvasó.

Korábban ilyen átfogó mű Sváb János és Wellisch Péter munkássága nyomán jelent meg, melyben a szerzők a módszerek kézi számításait, valamint a publikációkban megjeleníthető eredmények, táblázatok tartalmát és formáját ismertették. Sok kutató a mai napig bibliaként használja.

Az azóta eltelt években sok új biometriai módszer került be a gyakorlatba és a számítógépes statisztikai programcsomagokba. Ez a tény teszi indokolttá, hogy egy olyan átfogó kiadványt jelentessünk meg, amelyben a Debreceni Egyetemen végzett több évtizedes kutatómunka eredményeit és tapasztalatait felhasználva mutatjuk be napjaink legkorszerűbb statisztikai és biometriai módszereit. Ez a könyv egy matematikus és egy mezőgazdász közös munkája során született meg.

A könyv egyik sajátossága, hogy a módszereket többnyire valós kísérleti adatokon keresztül szemlélteti. Néhány módszer ismertetésekor azonban a könnyebb érthetőség érdekében a példákban kitalált adatokat használunk fel; ilyenkor nem célunk a szakmailag helytálló következtetés levonása.

A könyv fejezetei azonos elvek alapján épülnek fel: először ismertetjük az elméletet, az alkalmazhatóság feltételeit, majd konkrét példán keresztül a számítógépes megoldást, végezetül megvizsgáljuk, hogy teljesültek-e az alkalmazhatóság feltételei.

Miért pont az SPSS?

Mert ezt a programot a felsőoktatási intézmények ingyen használhatják, és a statisztika minden területét felöleli.

Ajánljuk ezt a könyvet a felsőoktatási intézmények hallgatóinak, oktatóknak, kutatóknak, minden olyan embernek, akik munkájuk során a biometriához közel kerültek valamint nem utolsó sorban a mindenkori oktatási miniszternek.

A szerzők

Debrecen, 2012. március

BEVEZETÉS

Az SPSS hasonlóan több Windows programhoz többablakos technikával dolgozik. Külön ablakban kezelhetjük az adatbázist, mely leginkább egy

táblázatkezelő adatbázishoz hasonlít, külön ablakban jelenik meg az eredmény, és külön-külön ablakban szerkeszthetjük a syntaxokat és szkripteket.

A syntax az SPSS belső nyelve, melyben a párbeszédablakokban beállított utasításokat tárolhatjuk és futtathatjuk. Ezen belső nyelv segítségével olyan elemzéseket, ill. utasításokat is kiadhatunk, melyeket a párbeszédpanelből nem. Az SPSS kiterjesztett matematikai, ill. mátrix műveletei, melyekkel a legbonyolultabb számítási műveletek is elvégezhetők, csak ezen belső nyelven megírt nagyon egyszerű utasításokkal végezhetők el. A mátrix eljárás tartalmazza az elemi mátrix műveletektől (összeadás, kivonás, szorzás, osztás) kezdődően a determináns, inverzmátrix, sajátérték, sajátvektor, stb. meghatározását. Ezekre a korreláció- és regresszió számításban mutatunk be néhány példát. A ciklusutasítások, iterációs eljárások, automatikus adatbázis készítesek is csak a syntax editor ablakban futtathatók. Syntaxot a legegyszerűbb módon a párbeszédablakok Paste utasításával hozhatunk létre. Ilyenkor megnyílik a syntax editor ablak és megjelennek a parancssorok. Az ilyen módon el nem érhető parancsokat, a szintaktikai szabályoknak megfelelően, saját kezűleg kell beírni. A szintaktikai leírás megtalálható az SPSS Syntax Reference Guide-ban. A legfontosabb utasítások az aktuális fejezetekben kerülnek ismertetésre.

A szkriptek valójában sax basic nyelven írt függvények és eljárások sorozata. Ez a nyelv, néhány speciális szabálytól eltekintve, nagyon hasonlít a Visual Basicre. Aki már programozott Visual Basicben, a programhoz szállított példa szkriptek tanulmányozása után, könnyedén elkészítheti a saját szkriptjeit. A szkriptek segítségével az SPSS minden lehetőségét ki lehet aknázni. Az ún. autoszkriptek segítségével egy esemény bekövetkezésekor végrehajtódik egy utasítássorozat, amivel például egy kimutatástáblázat létrehozásakor automatikusan beállíthatjuk, hogy mely változók jelenjenek meg a kimutatás soraiban, oszlopaiban, legyen-e részösszegzés, és ezek milyen formátumot vegyenek fel. A szkriptek és syntaxok egymás között átjárhatók (szkriptből futtathatunk syntaxot és syntaxból szkriptet).

A könyvben leírtak az SPSS 9.0-tól kezdődően a későbbi verziókban is jól alkalmazhatók, mivel a program készítői csupán apróbb módosításokat eszközöltek, ami nem okozott lényegi változást. A kényelmi szolgáltatások beépítése a későbbi verziókban esetenként még egyszerűbbé teszi az egyes műveletek végrehajtását.

FILE MENÜ

Read Text Data:

text típusú adatok beolvasása, pl. automata meteorológia állomás adatait. *.dat kiterjesztéssel. *Fixed width*, a felső sor tartalmazza a változók neveit. A

változók régi neveinek újakat adhatunk. Mentsük el a fájl formátumát későbbi munkák számára *.tpf kiterjesztéssel.

Apply Data Dictionary:

Az SPSS-be már beolvasott adatok oszlop, címke, stb. kiegészítő adatait már meglévő adatbázisból is beolvashatjuk a fenti paranccsal, *.sav kiterjesztésű fájlt választva.

Display Data Info:

Lemezen tárolt adatbázis tulajdonságait, változóit, címkéit listázza ki.

Érdeemes néha *.por, portable formátumba menteni az adatokat, mert ezt még a DOS-os programok is el tudják olvasni, mivel majdnem szöveg fájlként menti. Excelből 4.0-s munkalapként kell menteni az adatokat.

SZERKESZTÉS (EDIT) MENÜ

Options...

Charts: A grafikonok formátumát, kinézetét lehet megadni. A mintát (template) előre szerkesztett formátumban, fájlban megőrizve is megadhatjuk. Figyeljünk arra, hogy a megadott könyvtárban ott legyen a *.sct kiterjesztésű fájl. Ha töröljük, a program indítása után hibajelzést kapunk. Betűtípusokat, színeket, vonalakat, mintázatot határozhatunk meg. A grafikon keretét, rácsozatát állíthatjuk be interaktív módon.

Alapbeállítások: Edit – Options – General, Output Labels, Data

NÉZET (VIEW) MENÜ

Status Bar

A táblázat alján található információs sávot jeleníthetjük meg vagy rejthetjük el.

Toolbars...

A menüsor alá különböző ikonokat rakhatunk ki, amelyek így gyors billentyűként szolgálnak. A leggyakrabban használt eljárásokat érdemes itt megjeleníteni. (Show Toolbars). A beállítás paranccsal (Customize...) elvégezhetjük a szükséges beállításokat. Az Edit Tool... billentyűvel még az ikonokat is átrajzolhatjuk kívánság szerint. Bal egér gombbal fogjuk meg az

ikonokat és vigyük a kívánt helyre. Az ikonok törlését is hasonló módon végezhetjük, egyszerűen vontassuk ki az ikon területről.

Fonts...

Meghatározhatjuk a betű típusát (Arial, Courier, stb., stílusát (normál, dőlt, félkövér, félkövér dőlt), méretét (8-72). Kiválaszthatjuk az alkalmazott írásrendszert (Közép-európai, Nyugati, Görög, stb.).

Grid Lines

Az adatbázis ablakban a rácsozatot tudjuk ki, illetve bekapcsolni.

ADATOK (DATA) MENÜ

Define Variable...

Az aktív adat editor ablakban a kiválasztott változó leíró fejléc adatait lehet megváltoztatni, vagy új adatbázis változóit lehet definiálni.

Templates...

Ha több változónak egyszerre akarjuk beállítani a tulajdonságait, akkor ezt a parancsot kell használni. Előzetesen az aktív editor ablakban a módosítandó változókat lenyomott egérbillentyűvel ki kell jelölni

Insert Variable

Új változó (oszlop) beszúrását végzi az aktív változó után.

Insert Case

Egy új eset (sor) beszúrását végzi az aktív eset után.

Goto Case...

Megkeresi az adott esetet. Ha nem az adat ablak az aktív, akkor ennek a parancsoknak hatására azzá válik. A kereső dobozt a kívánt eset megkeresése után a Close gomb megnyomásával lehet lezárni.

Sort Cases...

Az adatmátrix sorai csökkenő vagy növekvő sorrendbe rendezhetők. A parancsdobozban meghatározhatjuk, hogy melyik legyen az elsődleges, másodlagos, stb. kulcs.

Transpose...

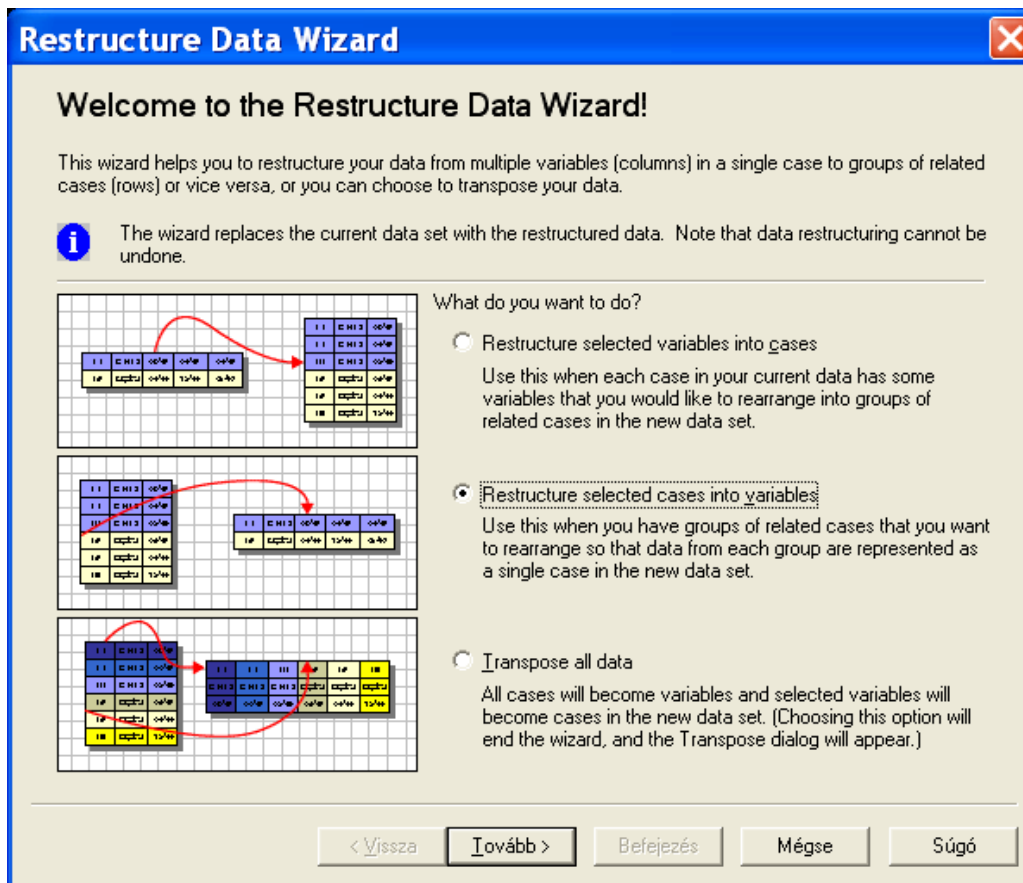
Az adatmátrix sorainak és oszlopainak felcserélése, ezzel az esetek és változók szerepei is felcserélődnek. A régi változók nevei a legelső új változó esetei lesznek, a többi új változó neve case_1, case_2, ... stb. lesznek.

Restructure...

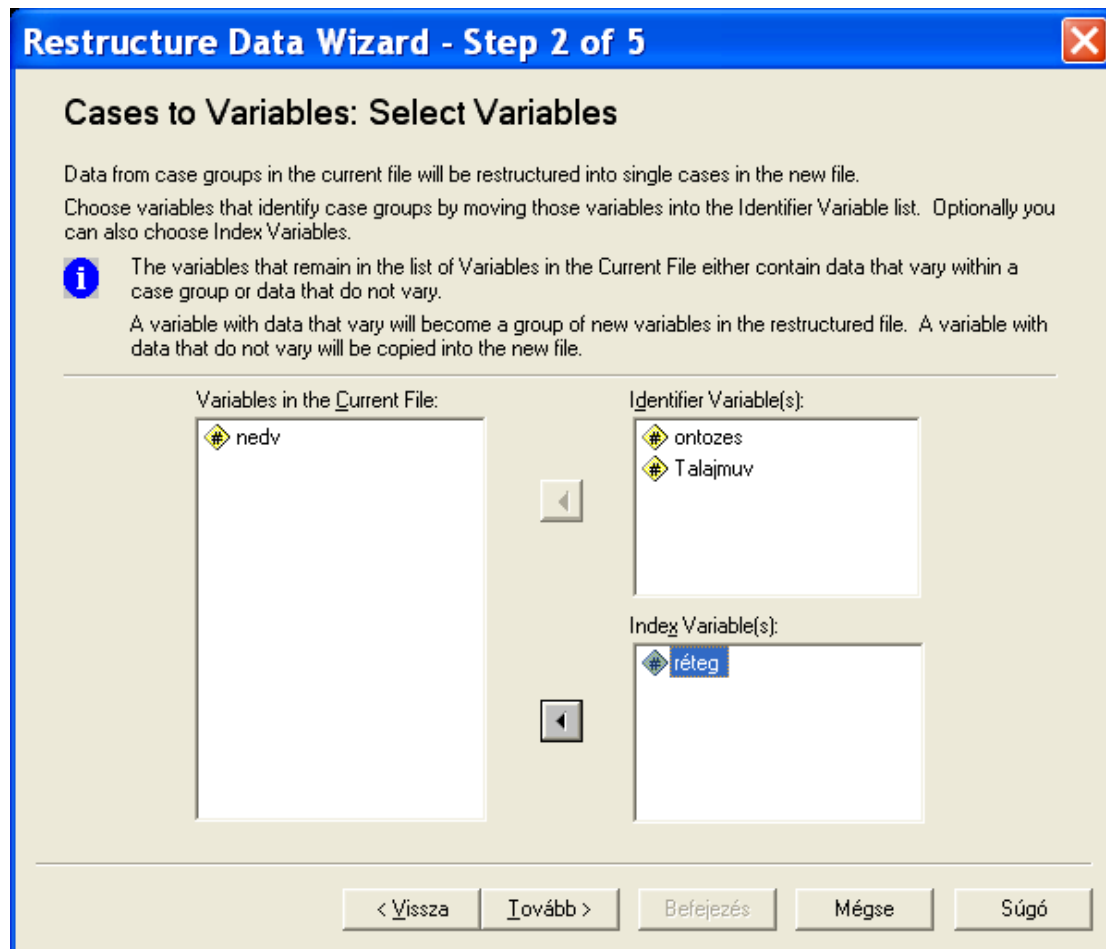
Itt az adatbázisok szerkezetét tudjuk megváltoztatni. Ezt átstrukturálásnak is nevezik. Vegyük az alábbi egyszerű adatbázist, és változtassuk meg a szerkezetét. A mért nedvességi értékek rétegenként kerüljenek új változóba.

Öntözés	Talajművelés	Réteg	Nedvesség
1,00	1,00	1,00	14,00
1,00	1,00	2,00	15,00
1,00	1,00	3,00	16,00
1,00	2,00	1,00	17,00
1,00	2,00	2,00	18,00
1,00	2,00	3,00	19,00
2,00	1,00	1,00	20,00
2,00	1,00	2,00	21,00
2,00	1,00	3,00	22,00
2,00	2,00	1,00	23,00
2,00	2,00	2,00	24,00
2,00	2,00	3,00	25,00

Data, Restructure... parancs után az alábbi párbeszédpanelt kapjuk. Itt kiválaszthatjuk, hogy a változókból csináljuk eseteket vagy fordítva, a kiválasztott esetekből legyenek új változók. A harmadik esetben az adatbázist transzponáljuk.



A Tovább billentyű után meg kell adni az új adatbázis szerkezetét. Szerintem, ez a párbeszédablakokban kissé nehézkes, sokkal egyszerűbb programból megadni. A baloldali ablakban láthatók a jelenlegi adatbázis változói (Variables in the Current File). Azonosító változóknak adjuk meg az öntözés és talajművelés változókat. Ezek külön sorokban fognak megjelenni az új adatbázisban. Index változónak jelöljük ki a réteg változót. Ez az adatbázis oszlopaiban fog megjelenni új változóként. Mivel három réteg van a nedvesség három új változóban fog megjelenni.



Öntözés	Talajművelés	Nedvesség_1	Nedvesség_2	Nedvesség_3
1,00	1,00	14,00	15,00	16,00
1,00	2,00	17,00	18,00	19,00
2,00	1,00	20,00	21,00	22,00
2,00	2,00	23,00	24,00	25,00

Merge Files

Fájlok bővítése, összekapcsolása. Új megfigyelésekkel (esetekkel) vagy új változókkal bővíthetjük az adatbázist. Az esetek bővítésével újabb megfigyeléseket csatolhatunk az adatainkhoz. Új változókkal történő bővítéskor több választási lehetőségünk is van, elő tudjuk állítani, pl. két fájl kombinációját egy kulcs változó felhasználásával. Legyen a *termes.sav* fájlnek három változója: *év*, *nPK*, *termes*. Összesen 84 megfigyelt terméseredményünk van, öt-öt 1990-től 2003-ig. Legyen a *csapadék.sav* fájlnek két változója: *év* és *csapadék*. Összesen 14 megfigyelésünk (rekord)

van, 1990-től 2003-ig. Ki szeretnénk bővíteni a *termés.sav* fájlunkat a csapadék értékekkel, hogy minden megfigyeléshez a megfelelő csapadéérték tartozzon. Nyissuk meg a *termés.sav* fájlt, és rendezzük növekvő sorrendbe az évek szerint. Válasszuk az Add Variables parancsot, a fájl megnyitás párbeszédpanelből válasszuk ki a csapadék.sav fájlt. Új párbeszédpanelt kapunk, amiben a két fájl információi láthatók. Válasszuk a Match cases on key variables in sorted files lehetőséget, és a rádiógombok közül External file is keyed table. A külső adatbázis lesz a kulcsmező tábla, ez tartalmazza a kulcsmezőt. A kulcsmező csak egyszer fordulhat elő a táblában. Az Excluded Variables: mezőben jelöljük ki az évváltozót, és húzzuk a Key Variables: mezőbe. Az OK gomb lenyomása után figyelmeztetést kapunk: ha nincsenek a fájlok a kulcsmező szerint sorba rendezve, rossz eredményt kapunk. Ez a lehetőség nagyon jól használható a logikailag összetartozó különböző táblák időszakos összekapcsolására, és elemzési feladatok elvégzésére. Ez nem más, mint az egy a többhez kapcsolat megteremtése egy relációs adatbázisban. Ennek két feltétele van, hogy mindkét fájlban legyen azonos kulcsmező, ami alapján össze lehet kapcsolni a két adatbázist, és mindkét fájl a kulcsmező szerint sorba legyen rendezve.

Aggregate Data

Break Variables: az a változó, ami szerint az összegzés ill. statisztika készüljön. Aggregate Variables: változó, amit összegezni szeretnénk. Create new data file: ezt választva egy új aggr.sav kiterjesztésű fájl készül az aggregált adatokkal.

Orthogonal Design

Generate...

Műtrágyázás	Öntözés	Status	Kártya
N 60	nem öntözött	Design	1
N 30	nem öntözött	Design	2
N 30	öntözött	Design	3
nem trágyázott	öntözött	Design	4
nem trágyázott	nem öntözött	Design	5
N 60	öntözött	Design	6

Több-tényezős kísérletek számára lineárisan független kezeléskombináció tervet készíthetünk véletlen szám generátor segítségével. A tényező nevének (Factor Name) és címkéjének (Factor Label) megadása után az Add billentyűvel felvesszük a tényezők ablakba. Az egérrel kiválasztva a tényezőt

definiálni kell a kezelésszintek számát (Define Values...), és el is lehet nevezni, pl. műtrágyából 1...3, nem trágyázott, 30 kg nitrogén, 60 kg nitrogén, stb.

Split File...

Lehetőségünk van az adatbázist felosztani és az elvégzett analíziseket így elvégezni. Három lehetőség közül választhatunk:

Minden esetet megvizsgálunk, nem képezünk csoportokat.

A csoportokat hasonlítjuk össze.

Az analízisek eredményét csoportonként jelenítjük meg.

Select Cases...

eseteket választhatunk ki az adatbázisból. Négy lehetőség közül választhatunk:

Minden eset részt vegyen az analízisben.

Ha valamilyen feltétel teljesül (if then)

Véletlen minta az esetekből

Kijelölhetjük az esetek bizonyos tartományát, az első és utolsó eset megjelölésével

Használhatunk szűrő változót

Mi legyen a ki nem választott esetek sorsa? Lehet szűrni és törölni őket az adatbázisból.

Nagy mennyiségű adat lekérdezése

Egy viszonylag nagy adatbázisból nagy mennyiségű adatot különbözőképpen kérdezhetünk le. Az egyik legegyszerűbb megoldás az adatok szűrése (select cases) parancs használata, azonban nagy mennyiségű adat, illetve több-szemponthoz lekérdezéskor nagyon sokat kell írni, és bonyolult logikai kifejezéseket kell megalkotni. Nagy a hibázási valószínűség. A másik nagyon hatékony megoldás, ha készítünk egy lekérdező adatbázist, és ehhez kapcsoljuk a nagy adatbázisból az adatokat az összekapcsol utasítással (merge files, add variables).

Pl.: a nagy adatbázis harminc év különböző kukorica hibridjeinek terméseit tartalmazza. Készítsük el az előre kiválasztott harminc hibrid egy-két vagy több éves terméseredményét. Az első lépés, alkossuk meg a lekérdező adatbázist. Rendezzük növekvő sorrendbe az adatokat a hibridek és év szempont alapján (Data, Sort Cases...).

A második lépésben kapcsoljuk hozzá a terméseredményeket a nagy adatbázisból.

SZÁSZ.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : HIBRIDEK 12

	HIBRIDEK	év	var	var	var	var
1	Borbála	2002				
2	Bourbon	1994				
3	Debreceni 377	1997				
4	Debreceni 377	1998				
5	Debreceni 377	1999				
6	Debreceni 377	2002				
7	Dekalb 386	1995				
8	Dekalb 391	2002				
9	Helga	1996				
10	Mv 277	2002				
11	Pactol	1994				
12	Pactol	1995				
13	BellaTC	1994				
14	BellaTC	1995				
15	BellaTC	1996				

Data View / Variable View /

SPSS Processor is ready

Add Variables: Read File

Hely: Biometria

- Legutóbbi dokumentumok
- Asztal
- Dokumentumok
- Sajátgép
- Hálózati helyek

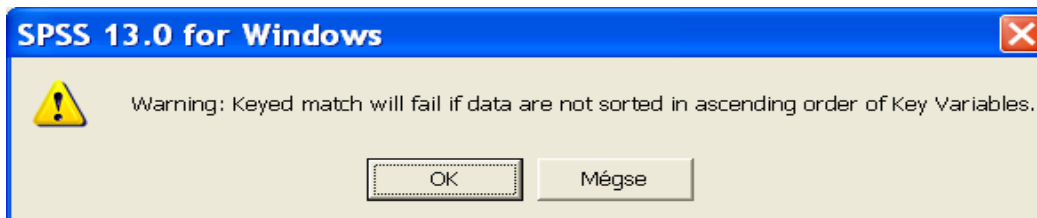
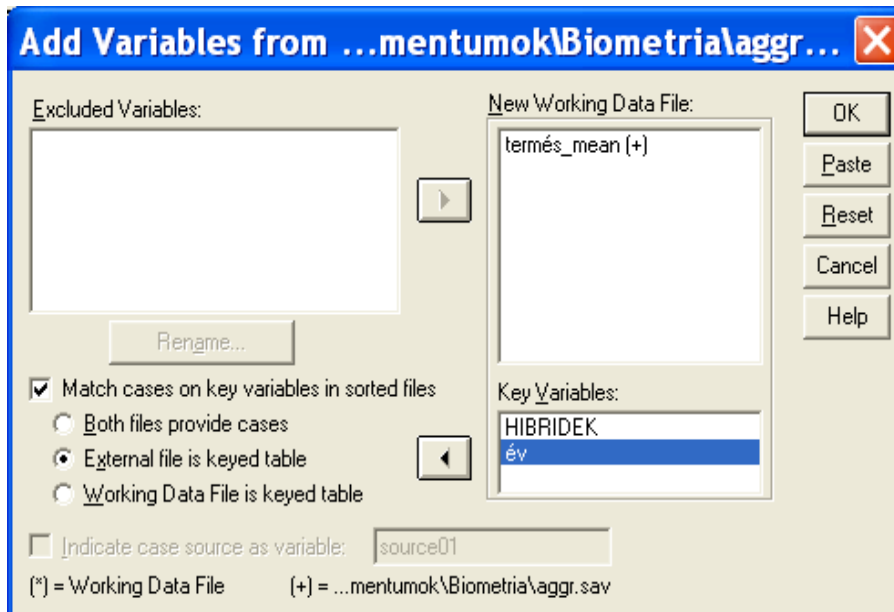
aggr
 KiteTalVizs
 KITEtermés
 Logisztikus_Függvény
 Sváb46o
 Sváb46Standard
 SZÁSZ
 Szerkezet
 TalajmuvTermések
 Termés1989

Fájlnév: aggr

Fájl típus: SPSS (*.sav)

Megnyitás

Mégse



SZÁSZ.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

3 :	HIBRIDEK	év	termés_mean	var	var	var
1	AW 043 (Perceval)	1994	8,74			
2	AW 043 (Perceval)	1995	8,41			
3	AW 043 (Perceval)	1996	8,89			
4	AW 143 (Durandal)	1994	9,14			
5	AW 143 (Durandal)	1995	7,32			
6	AW 143 (Durandal)	1996	10,18			
7	BellaTC	1994	8,39			
8	BellaTC	1995	5,81			
9	BellaTC	1996	10,30			
10	Borbála	2002	.			
11	Bourbon	1994	7,05			
12	Clarisia	1996	10,28			
13	Colomba	1994	9,19			
14	Colomba	1995	7,14			
15	Colomba	1996	10,49			

Data View Variable View / SPSS Processor is ready

Weight Cases...

Alul vagy túl reprezentált minták esetében lehet súlyzó tényezőt alkalmazni. Ha több ismérv alapján is alul vagy túl reprezentált a minta, akkor egyenként kell a súlyzó tényezőket kiszámítani, és az egyenkénti súlyzó tényezőket össze kell szorozni. (Ez a szociológiai és társadalomkutatásban elfogadott eljárás.) Pl.: 60 megfigyelésből 50 férfi és 10 nő. A férfiak túl reprezentáltak ebben a mintában ezért a két súlyzó tényező férfiak esetében 10/60, nők esetében 50/60.

ÁTALAKÍTÁSOK (TRANSFORM) MENÜ

Az adatmátrix elemeit lehet megváltoztatni, illetve új változókat lehet előállítani régi változók segítségével. Átkódolhatjuk a régi esetek értékeit akár új, akár a régi változókba. Az esetek rangszámait is kiszámíthatjuk.

Compute Variable:

Számított változó létrehozása. Meg kell adni a célváltozó nevét és a numerikus kifejezést. Lehetőség van arra is, hogy valamilyen logikai kifejezést is beállítsunk, és ilyenkor csak azoknál az eseteknél képződik a számított érték, amelyeknél a logikai érték igaz. A többi helyre *system missing value* kerül.

Gyakran előforduló feladat, hogy idősort kell előállítani, vagy meglévő idősort kell különböző szempontok szerint átalakítani. A talaj-növény-atmoszféra modellekben az időt az aktuális év január elsejétől eltelt napok számával jelölik (Julianus dátum). Havonkénti, negyedévenkénti összesítést ill. kimutatást így elég nehéz elvégezni. A program a különböző dátum függvényekkel lehetőséget biztosít az átalakításokra. Pl. DATE.YRDAY(év, az év napja) segítségével rendes dátumot lehet előállítani. A számított új változónak természetesen dátum típust kell megadni. A DATE.* függvényekkel számokból lehet különféle dátumot előállítani, az XDATE.* függvények pedig dátumból számokat, pl. napok száma, hónap száma, negyedévek száma, stb. Az így elkészített attribútumokkal különféle szempontok szerint csoportosíthatjuk az adatokat, készíthetünk statisztikákat, elemzéseket. (ld. *esztendő2002.sav*).

Véletlen számokat is elő tudunk állítani a beépített eloszlásfüggvények segítségével. Pl. RV.NORMAL(mean, stddev) normál eloszlás ismert középérték és szórás esetén.

Random Number Seed:

A számítógéppel generált u.n. pszeudó-véletlen számok előállításakor a kiindulási szám megadása. Csak sok számjegyű, páratlan szám adható meg.

Amennyiben sokszor generálunk véletlen számokat, időnként célszerű átállítani, nehogy ismétlődés lépjen fel a véletlen számok között.

Count:

Egy olyan új változó hozható létre, amelyben a változólistára felvitt változók együttes előfordulásait lehet regisztrálni.

Recode:

Előfordulhat, hogy ugyanazt a hibridet szintaktikailag kétféle módon rögzítettük, pl. Pelican és Pelikán. Az automatikus újrakódolás során két különböző szám fog hozzárendelődni a két megnevezéshez. Hogyan lehet ezt kijavítani?

Az újrakódolás során választhatjuk, hogy ugyanabba a változóba (Into Same Variables) vagy új változóba (Into Different Variables) kerüljenek az új értékek. Válasszuk, hogy ugyanabba a változóba kerüljenek az értékek. Fel kell sorolnunk a régi és új értékeket, és fel kell venni őket a listába, majd OK. Az újrakódolás megtörténik. Meg kell jegyezni, hogy a régi értékek, amelyek most már nem szerepelnek az adatbázisban, címkéi továbbra is megőrződnek.

A régi felesleges címkéket az Automatikus Újrakódolással (Automatic Recode) törölhetjük. Összefoglalásként: Automatic Recode → Recode Into Same Variable → Automatic Recode.

Categorize Variables:

Egy változó tartományát lehet felosztani kategóriákra, alapállapotban négy kategóriát ajánl fel a program, de lehet változtatni.

Rank Cases:

Egy változó értékeinek a nagyság szerinti sorrendben elfoglalt helyzetének megfelelő rangszámát generálja egy új változóba. Ha két egyforma érték áll a változóban, megfelezi a sorszámot, pl. 1,5 és 1,5.

Automatic Recode:

Változókat lehet automatikusan újrakódolni. A változó listából válasszuk ki az újrakódolandó változót, a New Name ablakba írjuk be az új változó nevét és nyomjuk meg a New Name gombot. OK után automatikusan újrakódolja a változót. Text típusú változó esetében, ha a változó különböző csoportokat jelöl nem érdemes a szöveget minden egyes rekordban tárolni, elég csak a kódokat. Ezzel az adatfájlt mérete jelentősen csökken. A kódok numerikus értékek lesznek. Az újrakódolt változóban a számokhoz címkék (labels) kapcsolódnak, melyek az eredeti text típusú változó tartalmát veszik fel.

Run Pending Transforms:

A felfüggesztett transzformációs parancsokat hajtja végre. Főként a syntax-ok futtatásakor használjuk, amelyeket a transzformációs opciókat használva a Preferences parancsdozobban felfüggesztettünk.

ELOSZLÁSOK

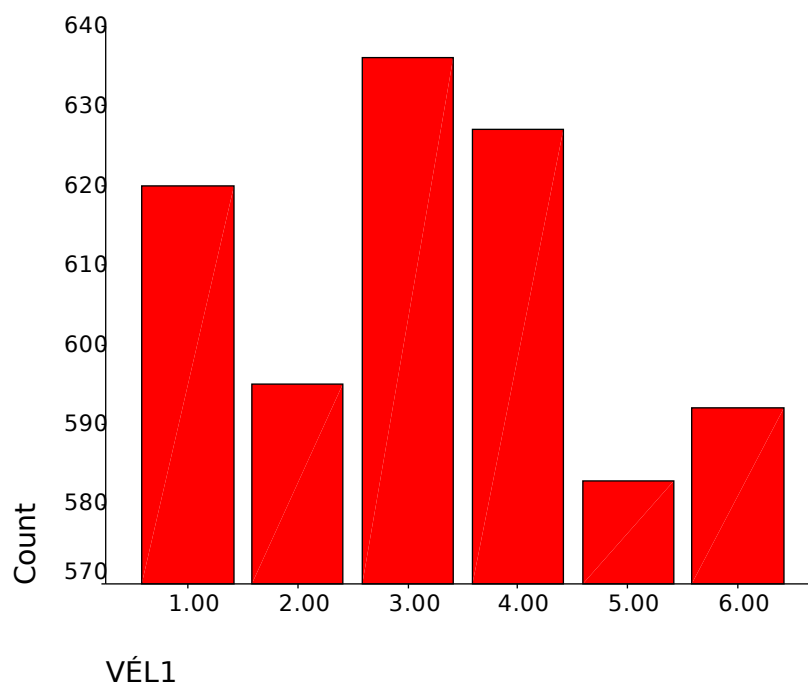
UNIFORM(max) = egyenletes eloszlású pszeudó véletlen számok előállítására a 0 és max tartományban.

RV.UNIFORM(min, max) = egyenletes eloszlású pszeudó véletlen számok előállítására min és max között.

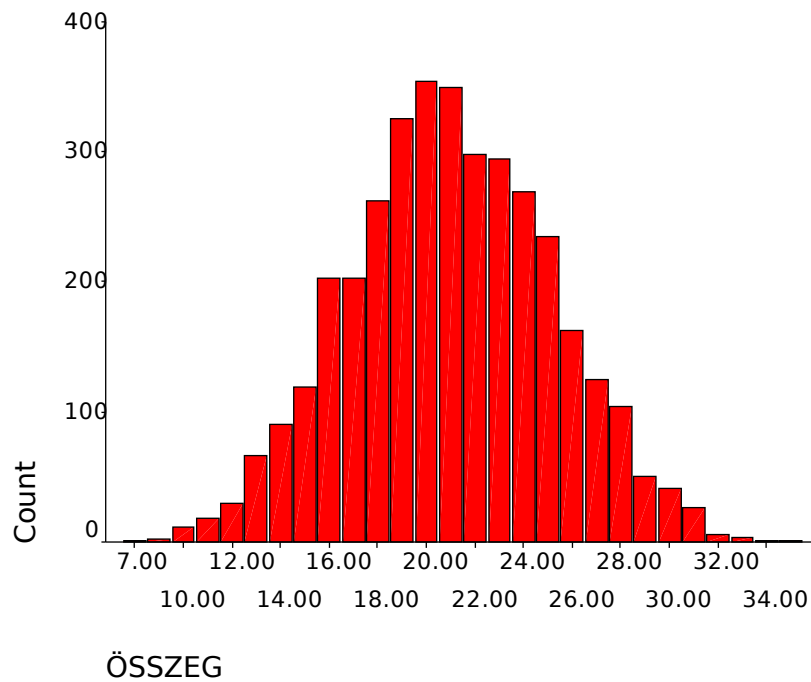
RND(numexpr) = egész rész függvény

Kockadobások szimulálása:

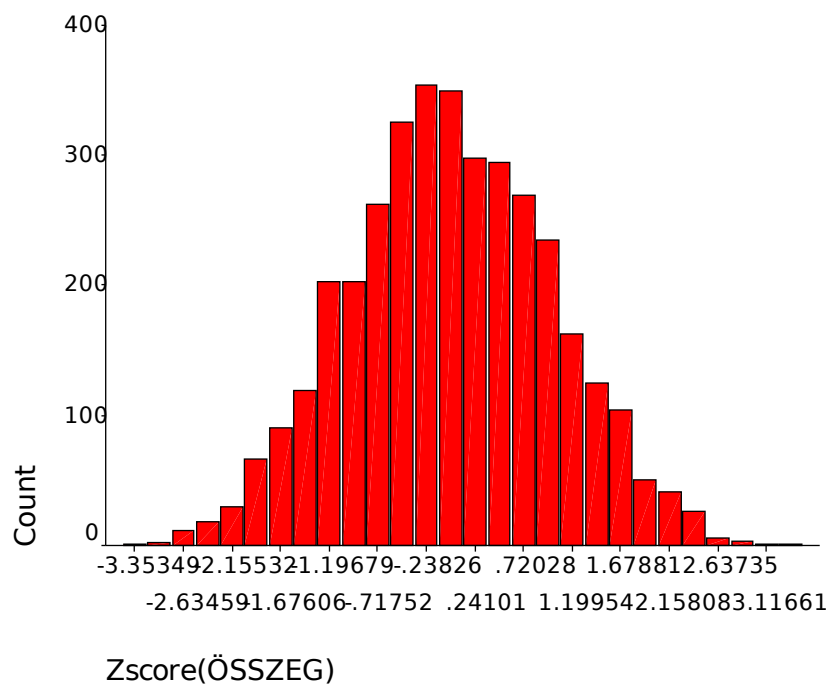
RND(UNIFORM(6)+0.5), egyenletes eloszlás 1-től 6-ig. ábrázolni a gyakoriságot oszlopdiagramon.



Hat új egyenletes eloszlású változó létrehozása, összeg kiszámítása. Ábrázoljuk az összeget!



Az adatok standardizálása, Analyze, Descriptive Statistics, Descriptives..., Save standardized values as variables. Ábrázolás.



ANALÍZEK

Riportok

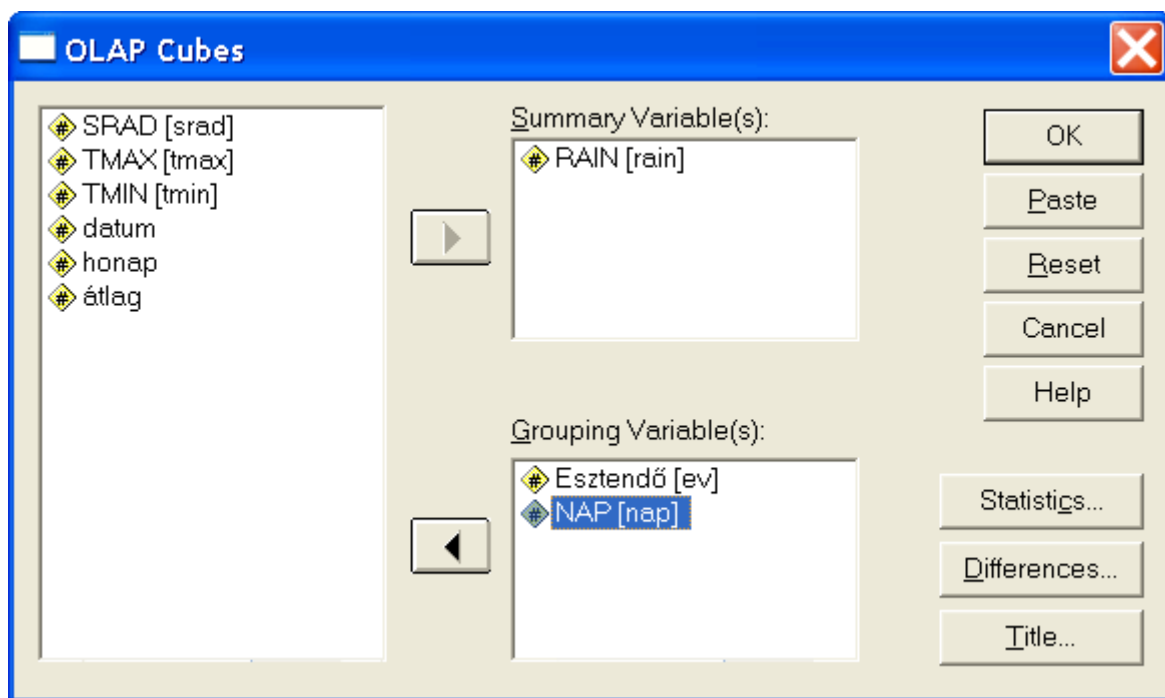
Adatbázisunkról különböző szempontok alapján készíthetünk kimutatásokat táblázatos formában.

OLAP Cubes...

Kimutatásokat, kimutató táblázatokat készíthetünk skála típusú adatokkal (Olap Cubes), Pivot tábla formátumban. OLAP (Online Analytical Processing). Réteg (layer), sor (row) és oszlop (column) változók szerint csoportosíthatjuk az adatainkat. Különböző statisztikákat jeleníthetünk meg, centrális mutatókat, szóródási és terjedelmi jellemzőket.

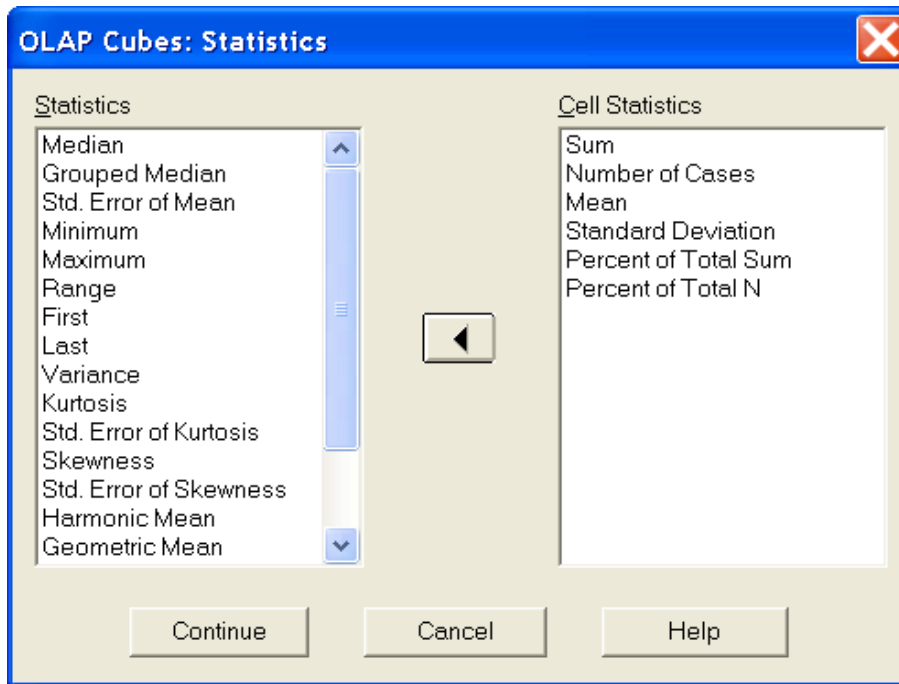
Analyze, Reports, OLAP Cubes...

Az elemezni kívánt skála típusú adatot vagy adatokat a



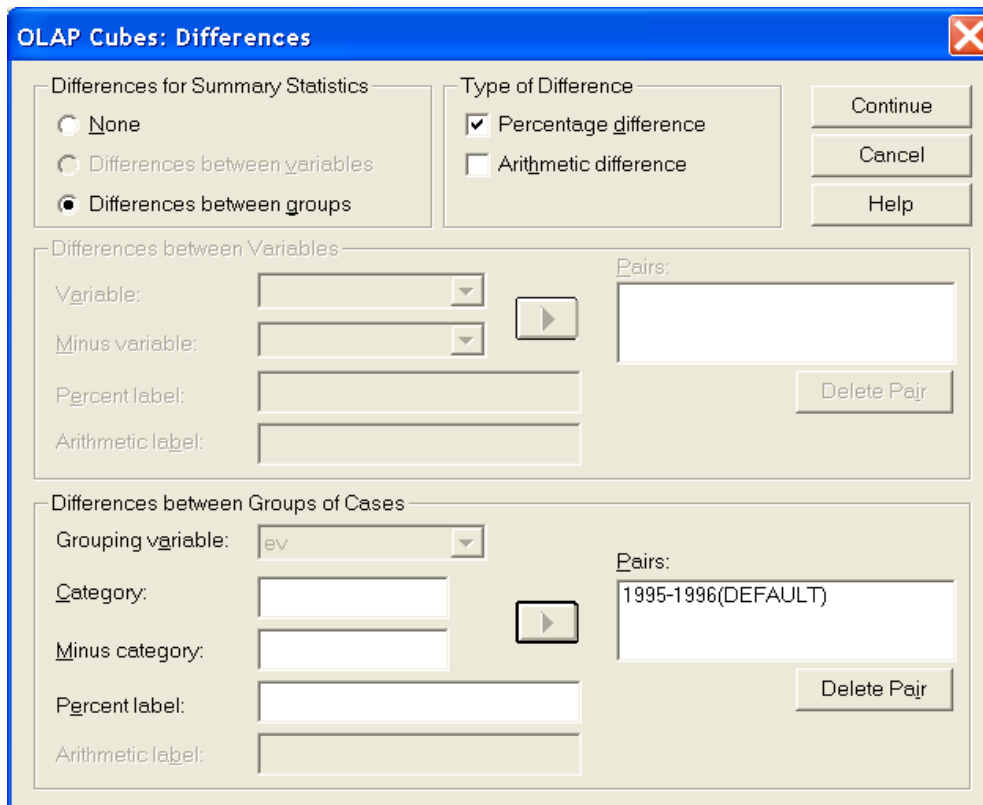
1. ábra: Kimutató varázsló párbeszédablaka

Summary Variable(s): ablakba tesszük. A csoportképző változókat a Grouping Variable(s) ablakba. A Statistics... gombra kattintva különböző statisztikai jellemzőket választhatunk.



2. ábra: A kimutatásban megjeleníthető statisztikák

Differences... gomb a változók, ill. csoportok közötti különbségeket jeleníti meg.



3. ábra: A kimutatásban megjeleníthető különbségek

Az OK gomb lenyomása után az Output ablakban megjelenik az eredmény összezárt formában, azaz minden csoportképző változó a rétegekben (layer) kerül.

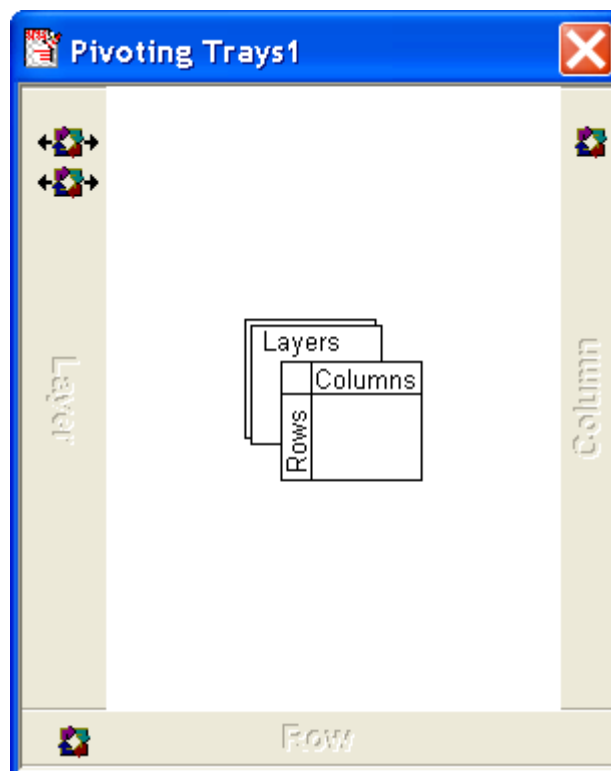
OLAP Cubes

Esztendő: Total

NAP: Total

Sum	
RAIN	5101.1

A kimutatást tetszés szerinti formába önthetjük, a rétegeket sorokba illetve oszlopokba húzhatjuk. Ehhez kattintsunk kettőt a táblázatban az egér balgombjával. A felső menüsoron megjelenik a Pivot parancs, melyben a Pivoting Trays parancs megnyitja a szerkesztési lehetőséget.



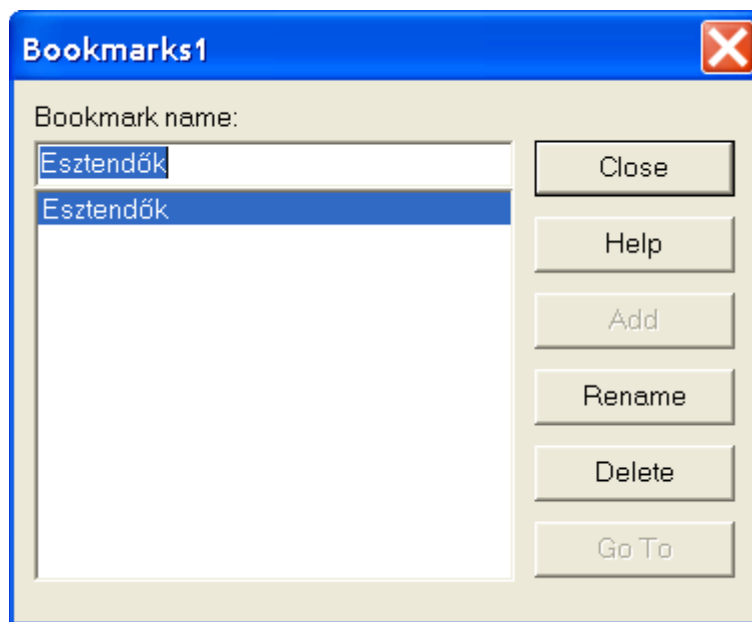
4. ábra: A kimutatás szerkezetének megváltoztatása

A baloldalon a réteg (layer), alul a sor (row) és jobboldalon az oszlop (column) található. A változókat az egérrel húzzuk a kívánt helyre, pl.

OLAP Cubes

NAP: Total		
Sum		
RAIN	1995	414.0
	1996	573.0
	1997	397.0
	1998	635.0
	1999	637.0
	2000	359.0
	2001	585.0
	2002	411.5
	2003	520.9
	2004	568.7
	Total	5101.1

A táblázat minden egyes elemét formázhatjuk, és elmenthetjük a kimutatás egyes változatait. Ehhez nyissuk meg a Bookmarks (könyvjelzők) parancsot.



5. ábra: A könyvjelzők megadása

Adjunk nevet az aktuális kimutatás változatnak, és az Add gombbal adjuk hozzá a könyvjelzőt.

A View menüparancsban válasszuk a Toolbars... lehetőséget, ekkor megjelennek a segédeszközök (tolltartó), melyek segítségével hasznos eszközök állnak rendelkezésünkre a kimutatások további elemzéséhez, formázásához.



6. ábra: Segédeszközök a kimutatások formázásához

Itt megtalálhatók a könyvjelzők is, amivel a kimutatások különböző változatai könnyen áttekinthetők.

Case summaries...

Nagyon hasonlít a pivot táblához, csak sokkal egyszerűbb formátumban jeleníti meg az adatokat. Jól használható a bevitt adatok ellenőrzésére, különböző csoportosítások szerinti adat-megjelenítéshez.

Case Summaries

Mean		
HONAP	TMAX	TMIN
1,000	-1,268	-7,496
2,000	-1,629	-12,093
3,000	7,677	-2,019
4,000	15,650	3,213
5,000	26,100	12,616
6,000	28,057	14,823
7,000	27,352	15,674
8,000	29,671	13,008
9,000	22,063	9,453
10,000	12,165	3,248
11,000	10,490	2,657
12,000	2,652	-3,039
Total	15,005	4,431

1. táblázat

Report summaries in Rows...

Report summaries in Columns...

A meteorológia adatbázisból minden kimutatás elvégezhető ezzel az eljárással. A **Data Columns** párbeszéd ablakban kell megadni az elemzendő változókat. Minden változóhoz különböző statisztikát rendelhetünk, sőt ugyanazt a változót többször is felvehetjük különböző számítási eljárásokkal. Pl. a hőmérsékletváltozóból az átlagot, minimumot, maximumot így egy táblázaton (kimutatáson) belül egyszerűen ki tudjuk számítani. A csoportképző változót a **Break Columns** ablakban kell megadni. Választhatunk növekvő, ill. csökkenő kiírás között. A kimutatás rtf formátumban készül. Nagyon jól használható az aggregált adatok megjelenítéséhez.

Leíró statisztikák (Descriptive Statistics)

Centrális mutatók: Átlag (várható érték), Medián (középső adat, gyakran helyettesíti a számtani közepet), Módusz (leggyakrabban előforduló elem)

Szóródási mutatók: Helyzeti és számított, Maximum (standardizált értéke), Minimum (standardizált értéke), Terjedelem (max.-min., range), Kiugró értékek, Kvartilisek (negyedelők), Interkvartilis $(Q3-Q1)/2$, Szórás (standard eltérés), Variancia (szórásnégyzet), Standard hibája az átlagnak, Standard hibája a mediánnak

Az eloszlás alakjának jellemzése: Ferdeség (skewness, jobbra-balra ferde eloszlások), Csúcsosság (kurtosis, 0 normális még -2 , $+2$ között), Boxplot ábrázolás

Trimmelt, csonkított, robusztus leíró statisztika, a kiugró értékek elhagyása.

Gyakoriságok (Frequencies...)

A megfigyelt változók relatív és kumulatív eloszlását tudjuk elemezni, ill. ábrázolni. Megjeleníthetjük a gyakorisági táblázatot (Display frequency tables). A százalékos értékeken belül (Percentile Values): a kvartiliseket, ahol az adatok 25, 50 és 75%-a található. Feloszthatjuk az adatokat egyenlő csoportokra (2-től 100-ig) (Cut points for x equal groups) valamint tetszőlegesen megadott százalékok alapján is megjeleníthetjük az adatok eloszlását. A centrális mutatók közül az átlagot (mean), mediánt, móduszt valamint a megfigyelések összegét (sum), az eloszlási mutatók közül a szórást (std. Deviation), a varianciát, a terjedelmet (range), a minimum és maximum értékeket valamint az átlag hibáját (S.E. mean) tudjuk kiszámítani.

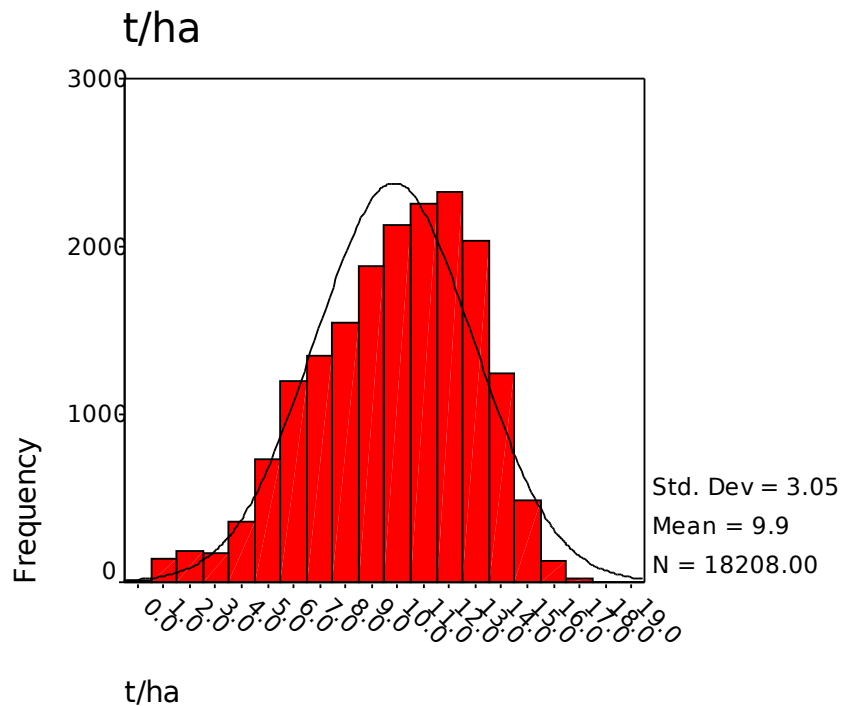
Statistics

t/ha		
N	Valid	18208
	Missing	0
Mean		9.86786
Std. Deviation		3.05116
Skewness		-.474
Std. Error of Skewness		.018
Kurtosis		-.207
Std. Error of Kurtosis		.036
Percentiles	5	4.59100
	25	7.75800
	50	10.22800
	75	12.22500
	95	14.20255

2. táblázat

Meghatározhatjuk az eloszlás jellemző paramétereit is. Az eloszlás szimmetriáját a ferdeségi mutatóval (skewness) jellemezhetjük. A normál eloszlás szimmetrikus és a ferdesége nulla. Pozitív ferdeségi érték mellett az eloszlásnak hosszú jobboldali része, farka van (right tail), ekkor balra ferdül, negatív érték esetében jobbra ferdül az eloszlás. Amennyiben a ferdeség értéke nagyobb, mint egy, az eloszlás nem normál. Az adatok középpont körüli csoportosulását a csúcsossági mutatóval (kurtosis) mérhetjük. Normál eloszlás esetén az értéke ennek is nulla. A csúcsosság pozitív értéke azt mutatja, hogy az adatok szélesebb csoportban helyezkednek el, az eloszlás két szélé hosszú. Negatív érték esetében kisebb csoportban helyezkednek el az adatok, az eloszlás két szélé rövidebb. A példa a kukorica termésének (t/ha) eloszlását mutatja be.

Ábrázolhatjuk az adatokat oszlop és kör diagramon, valamint hisztogram formájában is. A diagramokon ábrázolhatjuk a gyakoriságokat vagy a megfigyelések százalékos értékeit.

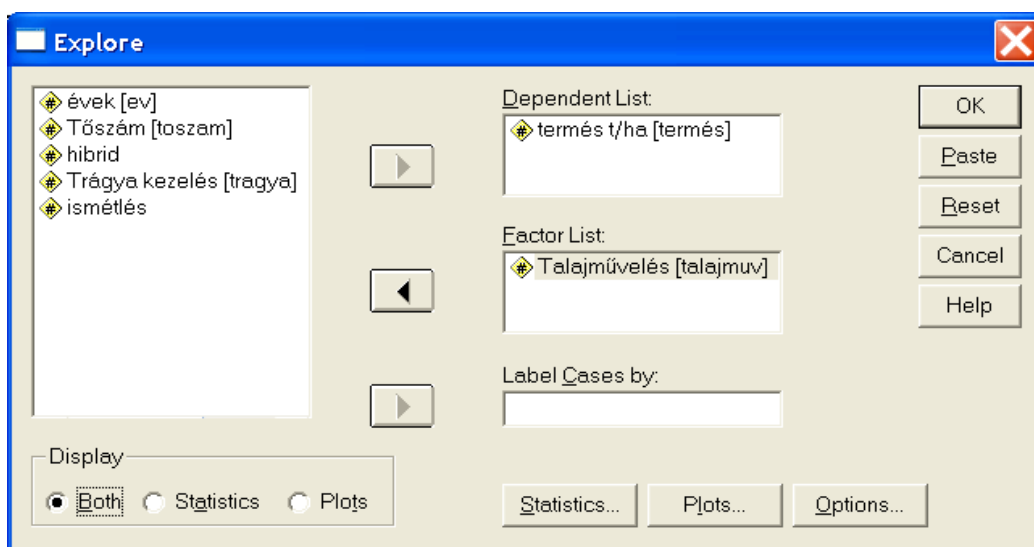


7. ábra

Descriptives...

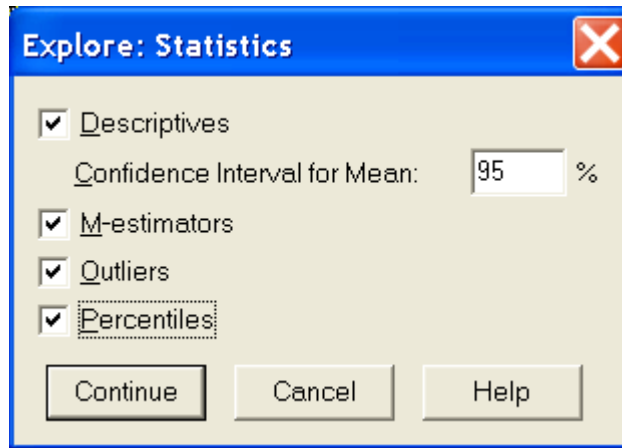
Explore...

Itt exploratív adatanalízist végezhetünk. Ez különösen fontos nagy adatbázisok esetében az adatok alapos megismerésére, felderítésére.



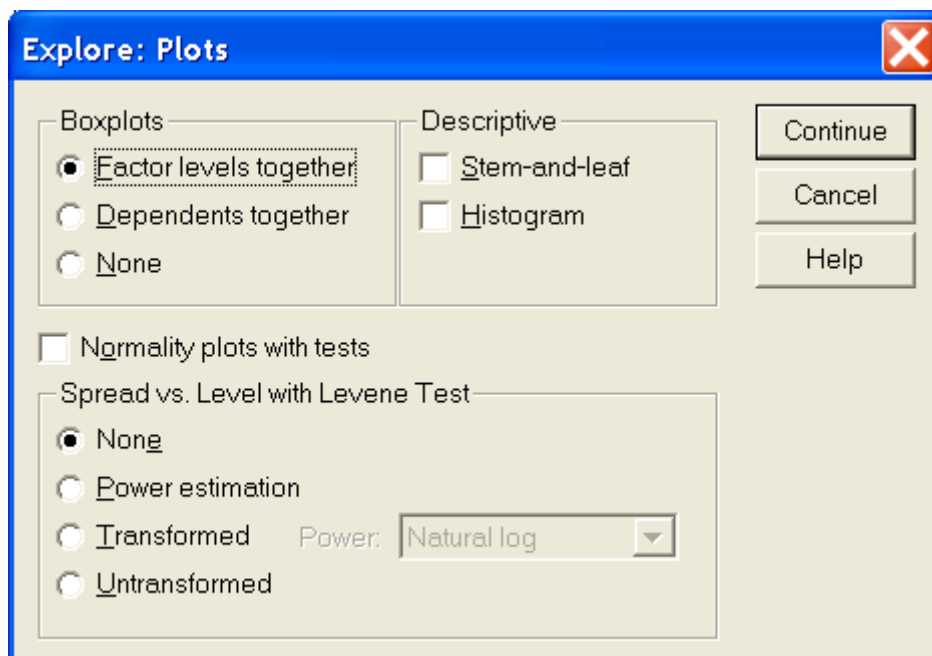
8. ábra

A Statistics... gombra kattintva különböző statisztikákat számíthatunk ki. Leíró statisztikák (Descriptives): átlag, medián, módusz, 5%-os csonkolt átlag, az átlag hibája, variancia, szórás, minimum, maximum, terjedelem, interkvartilisek, ferdeség, csúcsosság.



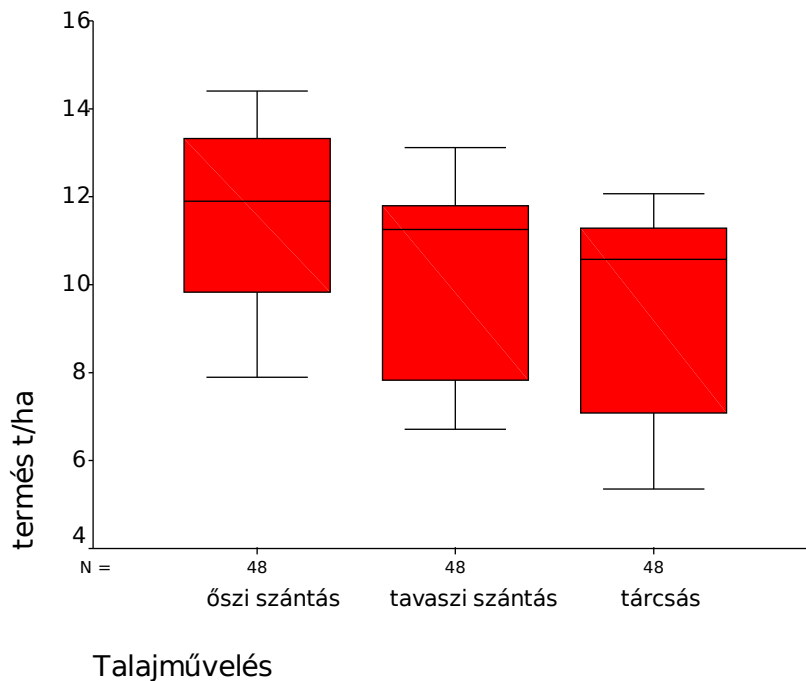
9. ábra

Robosztus centrális mutató meghatározása maximum-likelihood módszerrel (M-estimators). Négyféle módszerrel lehet meghatározni a centrális mutatót, mely torz eloszlás vagy extrém, kiugró értékek esetén jobb becslést ad, mint az átlag.



10. ábra

Az öt legnagyobb és legkisebb érték kijelzése (Outliers), ezeket az eredménylistában extrém értéként láthatjuk.



11. ábra

A megfigyelések százalékos eloszlását határozhatjuk meg, 5, 10, 25, 50, 75, 90, 95% (Percentiles).

Ábrák készítése, eloszlások tesztelése. Boxgrafikonok: a független változók függvényében készíthetünk kvartilis ábrát. A kiugróértékeket külön jelzi a program.

Az adatok eloszlásának leírása (Descriptive):

Stem-and-leaf grafikon: stem=szár, leaf=levél skála típusú adatok felbontása, hogy a fő értéket a szár, az utolsó jegyeket a leaf adja. Pl. 7.18 t/ha stem=7, leaf=1.

termés t/ha Stem-and-Leaf Plot for

TALAJMUV= őszi szántás

Frequency Stem & Leaf

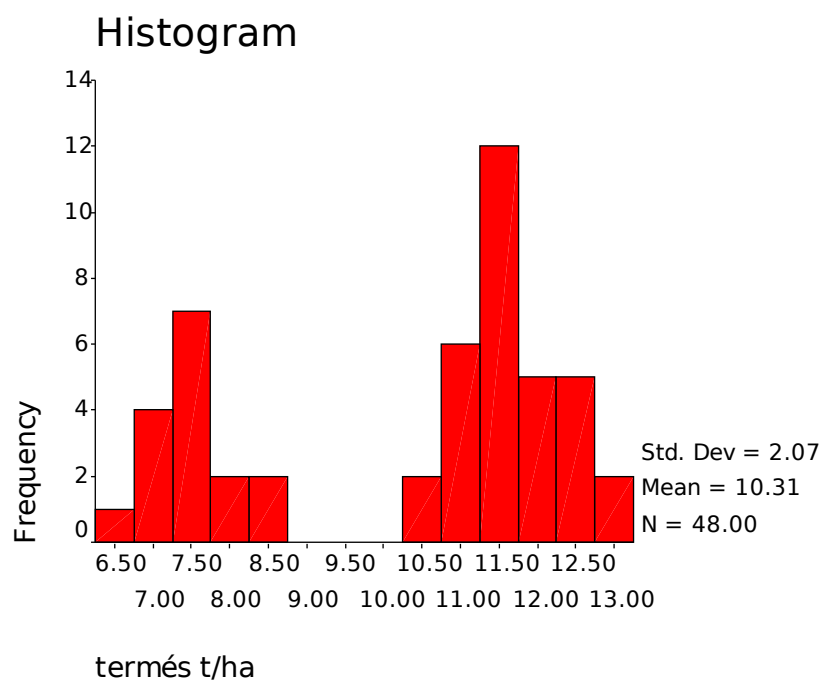
2.00 7 . 99

6.00 8 . 002458
 6.00 9 . 013699
 3.00 10 . 009
 5.00 11 . 02278
 8.00 12 . 00035679
 13.00 13 . 1223346666668
 3.00 14 . 233
 2.00 Extremes (>=113.5)

Stem width: 1.000

Each leaf: 1 case(s)

Hisztogram készítése (Histogram):



12. ábra

Normál eloszlás tesztelése Kolmogorov-Smirnov és Shapiro-Wilk próbával.

Tests of Normality

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
Talajművelés		Statistic	df	Sig.	Statistic	df	Sig.
termés t/ha	őszai szántás	.127	48	.050	.916	48	.002
	tavaszi szántás	.227	48	.000	.845	48	.000
	tárcsás	.263	48	.000	.817	48	.000

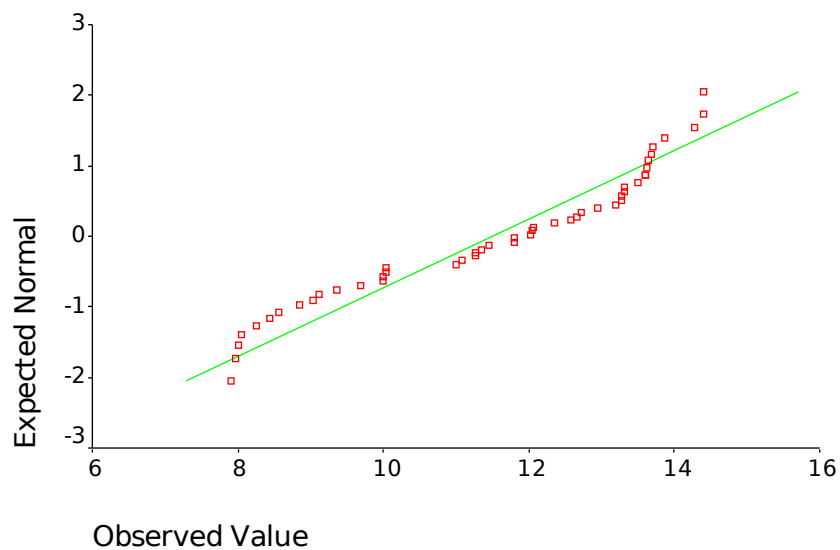
a. Lilliefors Significance Correction

Shapiro és Wilk's W-próba

Normális eloszlás tesztelésére szolgáló módszer, értéke maximum 1 lehet. Ennél jóval kisebb érték esetén nem normális az eloszlás. Szignifikancia vizsgálata megoldott, $\alpha = 0,05$. Akkor érdemes kiszámolni, ha a minta elemszáma nem haladja meg az 50-et.

Normal Q-Q Plot of termés t/ha

For TALAJMUV= őszai szántás



Detrended Normal Q-Q Plot of termés t_i

For TALAJMUV= őszi szántás



Keresztábrák (Crosstabs...)

A meteorológiai alapadatok ellenőrzését is el lehet végezni vele. Minden nap 24 darab nulla, negyed, fél és háromnegyed órás mérésnek kell lennie. Adjuk meg a napokat sorként, a negyedórákat oszlopként.

A hónap napja * Perc Crosstabulation

Count		Perc				Total
		0	15	30	45	
A	1	23	24	24	24	95
hónap	2	24	24	24	24	96
napja	3	24	24	24	24	96
	4	24	24	24	24	96
	5	24	24	24	24	96
	6	24	24	24	24	96
	7	24	24	24	24	96
	8	24	24	24	24	96
	9	24	24	24	24	96
	10	24	24	24	24	96
	11	24	24	24	24	96
	12	24	24	24	24	96
	13	24	24	24	24	96
	14	24	24	24	24	96
	15	24	24	24	24	96
	16	24	24	24	24	96
	17	24	24	24	24	96
	18	24	24	24	24	96
	19	24	24	24	24	96
	20	24	24	24	24	96
	21	24	24	24	24	96
	22	24	24	24	24	96
	23	24	24	24	24	96
	24	24	24	24	24	96
	25	24	24	24	24	96
	26	24	24	24	24	96
	27	24	24	24	24	96
	28	24	24	24	24	96
	29	24	24	24	24	96
	30	24	24	24	24	96
Total		719	720	720	720	2879

3. táblázat

Négy-mezős Chi²-próba függetlenség és homogenitás vizsgálatra

Osszunk fel egy véletlen minta alapján kiválasztott 100 személyt két alternatív ismérv szerint: nemek szerint és dohányzási szokás szerint.

	Nem dohányzó	Dohányzó	Σ
Nők	33	20	53
Férfiak	9	38	47
Σ	42	58	100

	-	+	Σ
-	a	b	a+b = n ₁
+	c	d	c+d = n ₂
Σ	a+c	b+d	a+b+c+d = n

Függetlenség esetén:

$$a/n_1 = c/n_2 = (a+c)/n \text{ vagy}$$

$$b/n_1 = d/n_2 = (b+d)/n \text{ stb}$$

$$Chi^2 = \frac{(n-1)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$Chi^2 = \frac{99(33 \cdot 38 - 20 \cdot 9)^2}{(33+20)(9+38)(33+9)(20+38)} = 18,819$$

$$DF = 1$$

Kritikus Chi²-értékek 5%-on: 3,841

Példa:

Kukorica fajták csövesedése:

FAJTA * CSÖVESD Crosstabulation

Count		CSÖVESD		Total
		Egy cső	Legalább két cső	
FAJTA	A fajta	73	23	96
	B fajta	48	8	56
Total		121	31	152

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	2.038 ^b	1	.153		
Continuity Correction ^a	1.486	1	.223		
Likelihood Ratio	2.123	1	.145		
Fisher's Exact Test				.210	.110
Linear-by-Linear Association	2.025	1	.155		
N of Valid Cases	152				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.42.

A Yates korrekcióval korigált különbség négyzetéből számított Chi-négyzetet a Continuity Correction mutatja. A két kukoricafajta a vizsgált tulajdonság szempontjából egyforma.

Custom Tables

Közéérték összehasonlítás (Compare Means)

A kezelésátlagok közötti különbségek megbízhatóságának igazolására többféle teszt ismeretes. Az összehasonlítás során, vagy két átlag különbségére vagyunk kíváncsiak, vagy a kezelésszintjeinket akarjuk összehasonlítani egymással, sorban tesztelve, hogy melyik kettő vagy több kezelés átlag tér el a többitől (szimultán vagy többszörös összehasonlítás). A kétféle eljárás kétféle összehasonlítási módszer csoportot takar. Az első módszer a páronkénti-tesztek csoportja a második a többszörös összehasonlító tesztek csoportja.

Közéértékek (Means...)

A függő változók (Dependent List) különböző statisztikai mutatóit lehet kiszámítani a független változók (Independent List) függvényében. Elkészíthetjük a variancia-táblázatot, tesztelhetjük az összefüggés linearitását és az összefüggés szorosságára az R és η paraméter nagyságából következtethetünk. Az R -érték, ill. R^2 a függő változó megfigyelt és becsült értékei közötti lineáris kapcsolat erősségét méri. Értéke 0,0 – 1,0 terjedhet. Kis érték esetében a függő és független változó között gyenge a kapcsolat vagy nem lineáris. Az η paraméter a korrelációs koefficienshez hasonlít, de itt a független változó nem folytonos, hanem kategória változó.

Report

termés t/ha

Talajművelés	Mean	N	Std. Deviation
őszi szántás	11.50673	48	2.06058
tavaszi szántás	10.30987	48	2.06889
tárcsás	9.56033	48	2.28744
Total	10.45898	144	2.27357

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
termés t/ha * Talajművelés	Between Groups	(Combined)	92.524	2	46.262	10.087	.000
		Linearity	90.923	1	90.923	19.825	.000
		Deviation from Linearity	1.601	1	1.601	.349	.556
	Within Groups		646.657	141	4.586		
	Total		739.181	143			

Measures of Association

	R	R Squared	Eta	Eta Squared
termés t/ha * Talajművelés	-.351	.123	.354	.125

4. táblázat

Egy-mintás t-teszt (One Sample T Test...)

Egy-mintás t-próba. Tesztelhetjük, hogy a valószínűségi változónk értéke megegyezik-e egy konkrét értékkel. Megválaszthatjuk a konfidencia intervallum nagyságát is.

Feltétel:

Normális eloszlású populáció, szigma ismeretlen és $n > 30$.

$$z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

DF = n-1

A minta elemszámának növekedésével a t – eloszlás egyre jobban közelíti a standard normális eloszlást.

Az X középértékű minta abban az esetben származhat a mű középértékű populációból ha t próbastatisztika abszolút értéke kisebb, mint az adott valószínűséghez tartozó kritikus t – érték.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
termés t/ha	144	10.45898	2.27357	.18946

One-Sample Test

	Test Value = 10					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
termés t/ha	2.423	143	.017	.45898	8.45E-02	.83349

5. táblázat

Egy-mintás z-próba

A minta középértékének összehasonlítása egy feltételezett középértékkel. Származhat-e az X középértékű minta egy μ_0 középértékű populációból? H_0 hipotézis:

$$H_0: \mu = \mu_0$$

Feltétel:

Normális eloszlású populáció, és ismert szórás,

Vagy tetszőleges eloszlású populáció, és $n > 30$.

A minta alapján számított X középérték standardizált érték felírható az alábbi formában:

$$z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Ahol:

- z a próbastatisztika minta alapján meghatározott értéke
- X a minta középértéke,
- μ a populáció feltételezett középértéke (adott középérték),
- σ a populáció (ismert) szórása,
- n a minta elemszáma.

A minta abban az esetben származhat az μ_0 középértékű populációból, ha a minta alapján meghatározott z próbastatisztika értéke kisebb az adott

valószínűségi szinthez tartozó kritikus z - értéknél. Egyoldali hipotézis esetén alfánál, kétoldali hipotézis esetén $\alpha/2$ -nél kell kikeresni.

$z < \text{kritikus } z$

Két független minta középértékének összehasonlítása (Independent-Samples T Test...)

Származhat-e a két független megfigyelés, minta azonos középértékű populációból?

Azonosnak tekinthető-e a két populáció középértéke, amelyekből a minták származnak? A két populáció, amelyekből a minták származnak, μ_1 , ill. μ_2 várható értékének becslésére a minták középértékei szolgálnak, $E(\bar{X}_1) = \mu_1$, ill. $E(\bar{X}_2) = \mu_2$.

$H_0: \mu_1 = \mu_2$

A középértékek összehasonlítására szolgáló statisztikai próbák – az egy-mintás próbákhoz hasonlóan – némileg eltérőek attól függően, hogy mekkora az egyes minták elemszáma, ill. hogy ismert-e az alappopulációk szórása.

Két független minta középértékének összehasonlítása. Feltétel:

Két független minta,

Normális eloszlású sokaságok,

A varianciák ismeretlenek, de azonosak

És $n < 30$ (n nem elég nagy a két-mintás z – próba alkalmazásához)

Ha a varianciák ismeretlenek, akkor azokat a mintákból számított szórásnégyzetekből becsülhetjük. A próbastatisztika értéke:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{(1/n_1) + (1/n_2)}}$$

$DF = n_1 + n_2 - 2$

A nevezőben az s_p a két minta összevont varianciájának (pooled variance) négyzetgyökét jelenti, melyet a két minta összevont szórásának nevezünk és az alábbi képlettel számítjuk ki:

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

A két populáció középértéke, amelyekből a minták származnak, abban az esetben tekinthetők azonosnak, ha:

$$|t| \leq t^*$$

A próba statisztika kritikus t – értékét kétoldali alternatív hipotézis esetén $\alpha/2$ -nél, egyoldali alternatív hipotézis esetén, α -nál kell a táblázatból

meghatározni. Ha a két populáció ismeretlen szórásnégyzete korábbi ismeretek, ill. a mintákból számított szórásnégyzetek alapján nem tekinthető azonosnak, akkor a t – próba helyett a Welch-próbát kell alkalmazni, mely igen hasonló a t-próbához, a különbség a szabadságfokok meghatározásában van.

A t-teszt alkalmazásakor előre tudni kell, hogy a két csoport szórása megegyezik-e, tehát tesztelni kell a csoportok szórását (Levene-póba). Amennyiben a szórások egyenlők, akkor a vizsgálatba vont összes csoportból kell a varianciát becsülni (pooled variancia). A próba valószínűségi változója t-eloszlású, így a középértékek különbségének szignifikanciája a t-érték táblázatból megállapítható.

Ha a két csoport szórása szignifikánsan különbözik, ilyenkor a két összehasonlítható csoport varianciáját súlyozni kell a variancia becsléséhez (separate variancia). A módosított variancia becslés az alábbi:

$$S_d = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

A próba valószínűségi változója ebben az esetben nem t-eloszlású, ezért nem a t-táblázatot, hanem a **Bonferroni-módosított** szignifikancia értékeket kell használni a középértékek különbözőségének elbírálásakor.

6. táblázat

Group Statistics

Trágya kezelés		N	Mean	Std. Deviation	Std. Error Mean
termés t/ha	nem trágyázott	48	7.66106	1.23444	.17818
	nitrogén 120	48	11.77213	1.08695	.15689

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
termés t/ha	Equal variances assumed	.472	.494	-17.317	94	.000	-4.11106	.23740	-4.58243	-3.63969
	Equal variances not assumed			-17.317	92.518	.000	-4.11106	.23740	-4.58253	-3.63959

Két-mintás z-próba

Feltétel:

Normális eloszlású független sokaságok, a variancia ismert,

Vagy tetszőleges eloszlású, mindkét mintában $n > 30$.

Az X_1 és X_2 középértékek különbsége akkor normális, ill. közelítőleg normális eloszlású, ha a sokaságok – amelyekből a minták származnak – normális eloszlásúak, illetve tetszőleges eloszlásúak, de a mintaelemek száma mindkét populációban nagyobb, mint 30.

A próbastatisztika:

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

A két populáció középértéke, amelyekből a minták származnak, abban az esetben tekinthetők azonosnak, ha:

$$|z| \leq z^*$$

A próba statisztika kritikus z-értékét kétoldali alternatív hipotézis esetén $\alpha/2$ -nél, egyoldali alternatív hipotézis esetén, α -nál kell a táblázatból meghatározni.

Párosított t-próba (Paired-Samples T Test...)

Párosított t-próba, két összefüggő minta középértékének összehasonlítására szolgál.

Ugyanazon egyeden két különböző időpontban mérünk egy tulajdonságot, vagy valamilyen csoportképző tulajdonság alapján párokat tudok képezni.

A két minta középértékének azonossága helyett a párosított minták d (előjeles) különbségének középértékére is megfogalmazhatjuk a H_0 hipotézist:

$$H_0: d_{\text{átlag}} = 0$$

Az előző eljárásokhoz hasonlóan itt is z- ill. t-próbát alkalmazhatunk attól függően, hogy ismert-e a d különbségek eloszlása és szórása, illetve mekkora a minta elemszáma?

Feltétel:

a d különbségek eloszlása normális, és σ_d ismeretlen (a mintából számított), és $n < 30$.

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \quad DF = n-1$$

A képletben s_d a párosított minták különbségének szórása, amelyet a minta alapján becsüljük.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Maximális hőmérséklet (C)	15.133	365	10.034	.525
	Minimális hőmérséklet (C)	5.710	365	7.868	.412

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Maximális hőmérséklet (C) & Minimális hőmérséklet (C)	365	.900	.000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Maximális hőmérséklet (C) - Minimális hőmérséklet (C)	9.424	4.521	.237	8.958	9.889	39.824	364	.000

7. táblázat

Átlag, az esetek száma, szórás, az átlag hibája. A két csoport közötti lineáris korrelációs együttható. Párosított t-próba eredmény táblázata: a két csoport különbségének átlaga, szórása, az átlag hibája, az átlag 95%-os konfidencia intervalluma, t-érték, szabadságfok, kétoldali szignifikancia szint.

Egy-tényezős variancia-analízis (One-Way ANOVA...)

Egy-tényezős variancia-analízis. Segítségével egy tényező hatását lehet vizsgálni a függő változó mennyiségi alakulására. A tényező, faktor valamilyen csoportképző ismérvvel rendelkezik, a függő változó pedig legtöbbször skála típusú adat. Egyszerre több függő változót is kijelölhetünk az analízis számára. A teszt során a nullhipotézis, hogy az átlagok egyenlők, nincs közöttük különbség. Ez a technika a két-mintás t-teszt általánosítása, kiterjesztése több mintára.

Közös szórásnégyzet (variancia) = Vizsgált tényezők + Hiba

A számítás során az SQ-t bontjuk fel.

$$x_{ij} = \mu + B_j + A_i + \varepsilon_{ij}$$

μ = a kísérlet főátlaga

$$B_i = (R_{\text{átlagok}} - \mu)$$

$$A_i = (\text{Kezelés}_{\text{átlagok}} - \mu)$$

$$\varepsilon_{ij} = \text{hiba}$$

A hiba normális eloszlású, független a blokk és kezelés hatástól. Mi van, ha nem teljesül? Lehet transzformálni az alapadatokat, logaritmikus vagy egyéb transzformációval. A blokk, kezelés és hiba hatások összege nulla.

Alkalmazási feltételek:

Független megfigyelések

Normális eloszlású sokaságok

Azonos szórások

Amennyiben az analízis az átlagok közötti egyenlőséget nem igazolja, szükséges az átlagok közötti különbségek kimutatása. A variancia-analízist kiegészítő középérték összehasonlító teszteknek kétféle típusa létezik:

előzetes, un. a priori kontrasztok és

az analízis után elvégezhető, un. post hoc analízisek

A kontrasztokat tehát a kísérleti adatok elemzése előtt kell előállítani, és így elvégezni az elemzést.

Az alábbi statisztikák készülnek: minden csoportról az esetek száma, átlag, szórás, az átlag hibája, minimum, maximum és az átlag 95%-os konfidencia intervalluma.

A csoportok varianciájának egyezőségét Levene's tesztel végezzük. Minden függő változóra elkészülnek a variancia-táblázatok. A post hoc range és többszörös középérték összehasonlító tesztek: Bonferroni, Sidak, Tukey's honestly szignifikáns differencia, Hochberg's GT2, Gabriel, Dunnett, Ryan-Einot-Gabriel-Welsch F test (R-E-G-W F), Ryan-Einot-Gabriel-Welsch range teszt (R-E-G-W Q), Tamhane's T2, Dunnett's T3, Games-Howell, Dunnett's C, Duncan's multiple range test, Student-Newman-Keuls (S-N-K), Tukey's b, Waller-Duncan, Scheffé, and least-significant difference.

Szimultán vagy többszörös összehasonlítás (multiple comparison) a köztudatban a szórásanalízis kiegészítője, fejlődését főleg felhasználói igények indították útjára. Jelentősége azonban jóval nagyobb, különösen a nem paraméteres esetben, ahol szórásanalízisre, e normalitást feltételező eljárásra, nem kerülhet sor. Ha az egy-szemponthus szórásanalízis F-próbája szignifikáns, kíváncsiak vagyunk, mely populációk miatt nem homogén a minta. Eleinte csak páronként az összes lehetséges csoport párra két-mintás t-próbát hajtottak végre. Előfordulhat azonban, hogy adott α -szinten szignifikáns F-próba esetén egyik csoport pár sem mutat szignifikáns t-értéket

az adott α -szint mellett. A szimultán hipotézis vizsgálatok nemcsak az egy-szemponos szórásanalízisben hódítottak teret, hanem mindenütt, ahol egyidejű döntésre van szükség, pl. regresszió, kovariancia, több-szemponos szórásanalízis, stb.

Szimultán döntés, ha kettőnél több összehasonlítandó mintám van. Olyan állításokat fogalmazzunk meg, amelyek egyidejűleg érvényesek. Ezek lehetnek:

Egyidejűleg érvényes konfidencia intervallumok vagy

Szimultán végzett statisztikai próbák.

A többszörös statisztikai próbák zöme paraméteres, a normális eloszlásra épülő eljárás. Sorozatos statisztikai összehasonlítások végzésekor halmozódik a próbaként vállalt elsőfajú hiba (kockázat). A szimultán összehasonlítási módszerek fő célkitűzése ennek a halmozódásnak a csökkentése illetve megszüntetése. Ennek eredményeként az egyes összehasonlítások konzervatív irányba tolódnak el: a próbaként fenyegető elsőfajú hiba ténylegesen kisebb a vállalt (névleges) kockázatnál. Ez azonnal szembeötlő a többszörös összehasonlítások azon csoportjánál, amelyek az ún. Bonferroni-egyenlőtlenség alapján dolgoznak. Az első ilyen javaslat Fisher könyvében (1935) található. A lényege, hogy m összehasonlítás esetén, az egyes összehasonlításokat a névleges α szint helyett α/m valószínűségi szinten hajtják végre. A valószínűség szubadditív tulajdonsága miatt, ha az összehasonlításoként vállalt α_i kockázatok összege olyan nagy, mint a teljes sorozatra vállalt α valószínűségi szint, akkor annak valószínűsége, hogy m elvégzett összehasonlítás után valahol elkövetjük az elsőfajú hibát, legfeljebb α :

$$P(H) \leq \alpha = \sum_{i=1}^m \alpha_i$$

ahol: H esemény azt jelenti, hogy az állítások közt legalább egy hibás. Ha az egyes állítások (valószínűség-számítási értelemben) függetlenek lennének, akkor a fenti becslés helyett az

$$1 - P(H) = \prod_{i=1}^m (1 - \alpha_i)$$

egyenlőséget alkalmazhatnánk, ami azt mutatja, hogy az állítások között nincs hibás. Miller (1966) megmutatta, hogy a szimultán konfidencia-intervallumokra a fenti egyenlőség helyett mindig a \geq érvényes. A szimultán vizsgált minták között végezhető összehasonlítások nem függetlenek. Legyen valamennyi α_i valószínűsége egyforma: $\alpha_i = \alpha_m = \alpha/m$, akkor az összehasonlítások nem független természetét figyelembe véve, a szimultán próbák együttes kockázata:

$$P(H) \leq 1 - (1 - \alpha_m)^m$$

A levezetésből látszik, hogy az egyes szintek egyformaságának semmiféle szerepe nincs. Megtehetjük tehát, hogy a fontosabb összehasonlítások

számára magasabb szintet jelölünk ki, ezzel biztosítva számukra a nagyobb erőt.

Kontrasztok: a csoportok közötti eltérés négyzetösszeget (sums of squares) fel lehet bontani trend komponensekre, vagy előzetesen megadhatunk általunk definiált kontrasztokat is. A trendek között különböző hatvány függvényekkel leírható trend-összetevőket tesztelhetünk.

A kontrasztok az egyes csoportok várható értékeinek lineáris kombinációi. A súlyok segítségével meg lehet adni a csoportviszonyokat, akár több kontrasztot is egyidejűleg. Ilyen csoportviszonyok a mezőgazdaságban, pl. műtrágyadózis kísérletekben nagyon könnyen értelmezhetőek. A lineáris összehasonlító függvények elméletével több szerző is foglalkozott. Magyar nyelven ÉLTETŐ Ö.-ZIERMANN M. 1964 megjelent művében található meg. A módszer lényege, hogy egy olyan lineáris függvényt kell alkotni, mint pl.

$$\lambda_g = c_{g1}x_1 + c_{g2}x_2 + \dots + c_{gp}x_p.$$

és ha teljesül a

$$c_{g1} + c_{g2} + \dots + c_{gp} = 0$$

feltétel, akkor ez egy lineáris összehasonlító függvény. A fenti definícióból következően végtelen számú λ_g létezik.

A kontrasztokra vonatkozó nullhipotézis: $H_g: \lambda_g = 0$,

Az ellenhipotézis: $A_g: \lambda_g \neq 0$.

Ha pl. egy tényező hatását T1, T2, T3, T4 szinten vizsgálunk, akkor a (T1, T2) csoport egybevetését a (T3, T4) csoporttal a $\lambda_g = x_1 + x_2 - x_3 - x_4$ függvény segítségével végezhetjük el (itt $1+1-1-1=0$).

A fenti összehasonlítás a variancia-analízis által szolgáltatott pooled variancia felhasználásával történik, ezért követelmény, hogy a csoportok szórásai megegyezzenek, így gyakran a variancia-analízis kiegészítő részét képezi. A contrast fejezetben a hatótényezők sokféle csoportosítása útján kapott átlagok különbözőségét lehet vizsgálni, pl. műtrágyázás esetén, a feltételezésem az, hogy az őszi búza a legnagyobb termést a 120 kg nitrogén adag mellett éri el. Vizsgálhatom az ez alatti adagokat, mintát véve, vagy az e feletti adagokat, szintén mintát véve, véletlenszerűen, ha nem 120 kg-t alkalmazok, vajon milyen eredmény születne.

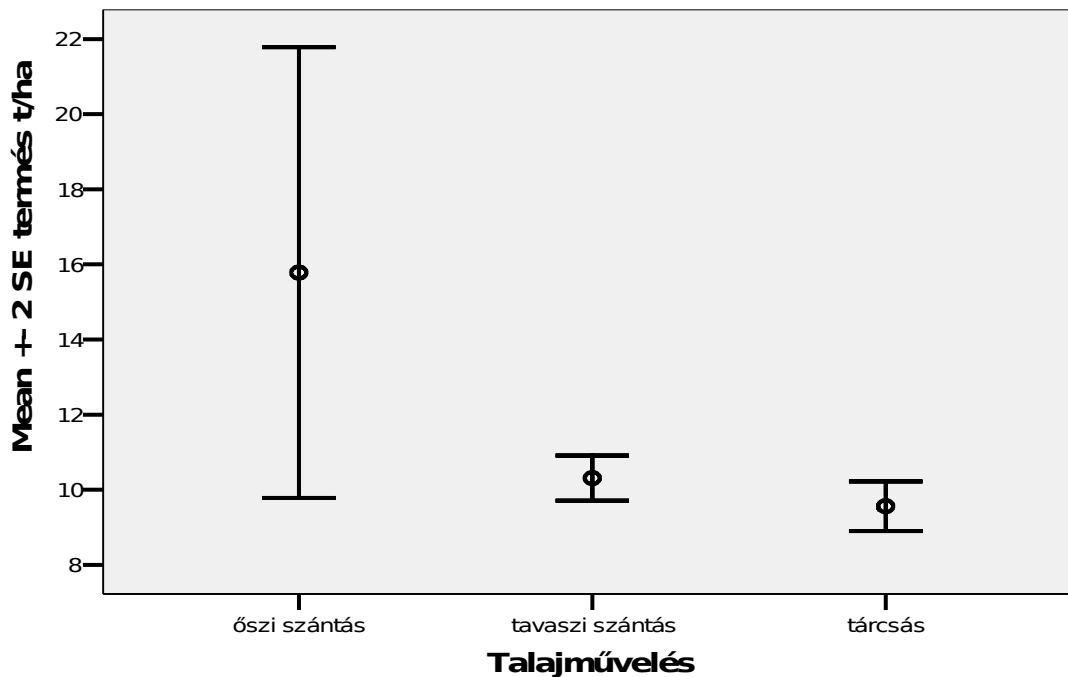
Az egy-szemponos szórásanalízis F-próbája akkor ad α -szinten szignifikáns eredményt, ha ezen a szinten létezik szignifikáns kontraszt a csoportok között.

Feladat:

Nyissuk meg a *Termés1989.sav* fájlt, és vizsgáljuk meg, hogy a talajművelésnek milyen hatása volt ebben az évben a kukorica termésére.

A legelső lépésben ábrázoljuk talajművelési változatonként a termések átlagának hibáját, pontosabban az átlagot \pm az átlag hibájának kétszeresét. Ebbe a tartományba fog legalább 95%-os valószínűséggel a valódi átlag esni.

Graph, Error Bar..., Simple, Summaries for groups of cases, Define, Variable: termés t/ha, Category Axis: Talajművelés, Bars Represent: Standard error of mean, Multiplier: 2.



13. ábra: Az átlag és az átlag hibájának kétszerese

Az őszi szántásos kezelés adataival valami probléma lehet, mert túlságosan nagy az átlag hibája, magába öleli a másik két kezelést is. Egyelőre hagyjuk így, és végezzük el a variancia-analízist.

Analyze, Compare Means, One-Way ANOVA, Dependent List: termés t/ha, Factor: Talajművelés, Options...: Homogeneity of variance test.

Test of Homogeneity of Variances			
termés t/ha			
Levene	df1	df2	Sig.
Statistic	2	141	.007

8. táblázat

A Levene-teszt azt mutatja, hogy a csoportokon belül a varianciák nem egyenlők. Ezek szerint valószínűleg az őszi szántásos parcellák terméseinek szórása szignifikánsan nagyobb, mint a másik kettőé. A variancia-analízis alkalmazásának egyik feltétele nem teljesül, ezért a lenti táblázat eredményét fenntartásokkal kell kezelni. Az elvégzett analízis a talajművelés szignifikáns hatását igazolja látszólag, sig. < 0,05. Ezt csak akkor fogadhatnánk el, ha a varianciák megegyeznének.

ANOVA

termés t/ha					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1106.779	2	553.390	3.758	.026
Within Groups	20763.765	141	147.261		
Total	21870.544	143			

9. táblázat

Végezzük el a középértékek többszörös összehasonlítását (Post Hoc analízis)! A többszörös összehasonlító teszteknek két nagy csoportja van: 1. a varianciáknak egyenlőknek kell lenni, 2. nem feltétel a varianciák egyenlősége. Válasszuk ki mindkét csoportból az elsőt!

One-Way ANOVA, Post Hoc..., Equal Variances Assumed: LSD, Equal Variances Not Assumed: Tamhane's T2.

Multiple Comparisons

Dependent Variable: termés t/ha

	(I) Talajművelés	(J) Talajművelés	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	őszi szántás	tavaszi szántás	5.470354*	2.477068	.029	.57336	10.36735
		tárcsás	6.219896*	2.477068	.013	1.32290	11.11689
	tavaszi szántás	őszi szántás	-5.470354*	2.477068	.029	-10.36735	-.57336
		tárcsás	.749542	2.477068	.763	-4.14745	5.64653
tárcsás	őszi szántás	-6.219896*	2.477068	.013	-11.11689	-1.32290	
	tavaszi szántás	-.749542	2.477068	.763	-5.64653	4.14745	
Tamhane	őszi szántás	tavaszi szántás	5.470354	3.015757	.211	-1.99077	12.93148
		tárcsás	6.219896	3.019043	.128	-1.24821	13.68800
	tavaszi szántás	őszi szántás	-5.470354	3.015757	.211	-12.93148	1.99077
		tárcsás	.749542	.445175	.260	-.33289	1.83197
	tárcsás	őszi szántás	-6.219896	3.019043	.128	-13.68800	1.24821
		tavaszi szántás	-.749542	.445175	.260	-1.83197	.33289

*. The mean difference is significant at the .05 level.

10. táblázat

Az LSD-teszt az őszi szántás és tavaszi szántás, valamint az őszi szántás és tárcsás talajművelés között 5%-os szignifikáns különbséget mutat. A Tamhane teszt egyik kezelés pár között sem mutat szignifikáns különbséget. Mivel a varianciák különbözősége miatt LSD tesztet nem csinálhatunk, a Tamhane-teszt eredményét kell elfogadni, és kideríteni, hogy miért nem tudjuk kimutatni a talajművelés okozta hatást.

Vizsgáljuk meg az őszi szántásos kezelések adatait!

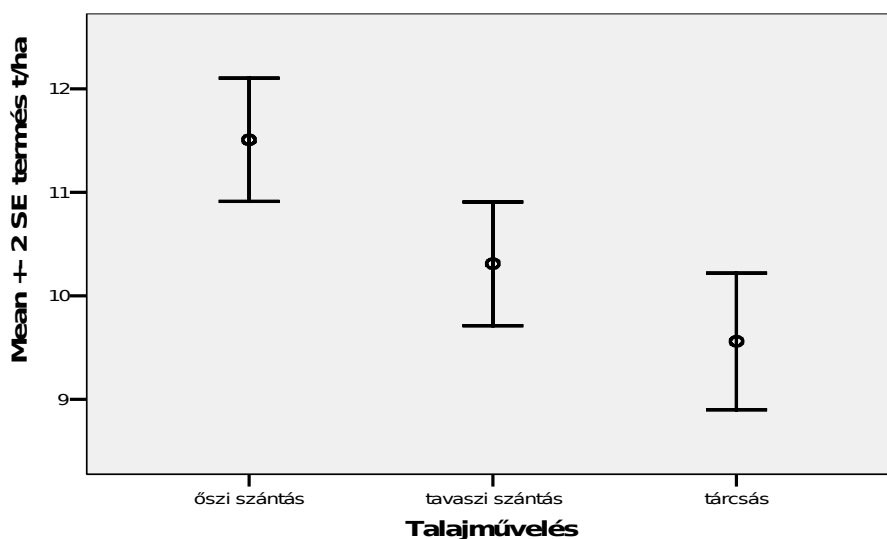
Analyze, Descriptive Statistics, Explore..., Dependent List: termés t/ha, Factor List: Talajművelés, Statistics..., Outliers.

Az eredménylistából csak a kiugró értékek táblázatát mutatjuk be.

Extreme Values					
		Case Number	Talajművelés	Value	
termés t/ha	Highest	1	9	őszi szántás	114.41
		2	10	őszi szántás	113.51
		3	32	őszi szántás	14.395
		4	35	őszi szántás	14.392
		5	36	őszi szántás	14.286
	Lowest	1	135	tárcsás	5.355
		2	134	tárcsás	5.421
		3	136	tárcsás	5.652
		4	122	tárcsás	5.697
		5	124	tárcsás	6.059

11. táblázat

Jól látható, hogy a 9. és 10. megfigyelés adatrögzítési hiba miatt egy nagyságrenddel nagyobb, mint a többi. Javítsuk ki a hibás adatokat és ismételjük meg az analízist a legelső lépéstől kezdődően!



14. ábra: Az átlag és az átlag hibájának kétszerese

Az átlagok ebben az esetben már jól elkülönülnek egymástól. Az átlagok hibáiból képzett intervallumok már kevésbé érnek egymásba.

Test of Homogeneity of Variances

termés t/ha				
Levene Statistic	df1	df2	Sig.	
1.096	2	141	.337	

12. táblázat

A variancia-analízis alkalmazási feltétele, a csoporton belüli varianciák egyezősége teljesül, tehát lehet variancia-analízist csinálni.

ANOVA

termés t/ha					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	92.524	2	46.262	10.087	.000
Within Groups	646.657	141	4.586		
Total	739.181	143			

13. táblázat

A variancia-analízis a talajművelés szignifikáns hatását mutatja. Az elvégzett többszörös középérték összehasonlító tesztek most már hasonló eredményt adnak. Mivel a varianciák megegyeznek, az LSD-teszt eredményét érdemes figyelembe venni, mert ennek a tesztnek ebben az esetben nagyobb a próba ereje. Ez azt jelenti, hogy a meglévő valódi különbséget nagyobb biztonsággal tudja kimutatni, mint a Tamhane teszt.

Multiple Comparisons

Dependent Variable: termés t/ha

	(I) Talajművelés	(J) Talajművelés	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	ősz szántás	tavaszi szántás	1.196854*	.437141	.007	.33266	2.06105
		tárcsás	1.946396*	.437141	.000	1.08220	2.81059
	tavaszi szántás	ősz szántás	-1.196854*	.437141	.007	-2.06105	-.33266
		tárcsás	.749542	.437141	.089	-.11466	1.61374
	tárcsás	ősz szántás	-1.946396*	.437141	.000	-2.81059	-1.08220
		tavaszi szántás	-.749542	.437141	.089	-1.61374	.11466
Tamhane	ősz szántás	tavaszi szántás	1.196854*	.421463	.017	.17227	2.22144
		tárcsás	1.946396*	.444371	.000	.86591	3.02689
	tavaszi szántás	ősz szántás	-1.196854*	.421463	.017	-2.22144	-.17227
		tárcsás	.749542	.445175	.260	-.33289	1.83197
	tárcsás	ősz szántás	-1.946396*	.444371	.000	-3.02689	-.86591
		tavaszi szántás	-.749542	.445175	.260	-1.83197	.33289

*. The mean difference is significant at the .05 level.

14. táblázat

Általános lineáris modell (General Linear Model)

Az általános lineáris modell a hagyományos variancia-analízis és a lineáris regresszió-analízis ötvözete. Egyetlen táblázatban jelenik meg a szórás elemzés és regresszió-analízis eredménye (15. táblázat). Napjainkban a variancia-analízisnek nagyon sokféle technikája létezik, amik lehetővé teszik a feladat sajátosságainak figyelembevételével a legalkalmasabb értékelési módszer kiválasztását. Az elemzés megbízhatósága a hiba (error) meghatározásának módjától függ, ami tulajdonképpen az eltérés négyzetösszeg (SQ) számítási technikájának függvénye. Az SPSS lehetővé teszi a kísérleti elrendezéshez hű, a felhasználó által megalkotott lineáris modell megbízható értékelését.

Tests of Between-Subjects Effects

Dependent Variable: X

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	119.248 ^a	3	39.749	4.706	.006
Intercept	20563.279	1	20563.279	2434.723	.000
FAJTA	119.248	3	39.749	4.706	.006
Error	439.184	52	8.446		
Total	21121.710	56			
Corrected Total	558.431	55			

a. R Squared = .214 (Adjusted R Squared = .168)

15. táblázat

A 15. táblázat 1-4 oszlopának értelmezése logikai sorrendben:

Total: az alapadatok négyzet összege (21 121), $\sum x^2$, szabadságfok (56)

Intercept: az alapadatok összegének négyzete osztva az adatok számával (20 563) $\frac{(\sum x)^2}{n}$, szabadságfok (1) valamint átlaga (20 563). Amennyiben az adatok egyáltalán nem szórnak (minden adat megegyezik), akkor a fenti két kifejezés értéke megegyezik. Az Intercept SS értéket Sváb könyveiben korrekciós tényezőként („C”) említi, mely nem más, mint a kísérlet főátlagának négyzetösszege, $\sum x^2$

Corrected Total: egyenlő Total – Intercept (558), vagyis $\sum x^2 - \frac{(\sum x)^2}{n}$, ez tulajdonképpen az alapadatok eltérésnégyzet-összege. Sváb könyveiben ez jelentette az „Összesen” sort. Szabadságfok (55).

Error: ebben a példában a négy FAJTA csoporton belüli eltérés négyzetösszege (439) $\sum_i \sum_j (x_{ij} - \bar{x}_i)^2$, szabadságfok (52), valamint ennek átlaga (8,446), ami gyakorlatilag a csoporton belüli varianciák átlaga. Sváb könyveiben a Hiba, a véletlen hatása, a meg nem magyarázott hatások. Minden FAJTA csoportban 14-14 megfigyelés van. Ebből az értékből gyököt vonva megkapjuk a csoporton belüli átlagos szórás nagyságát.

FAJTA: a kezelés okozta hatás, a négy fajta átlagának eltérése a főátlagtól.
 $r\sigma_{fajta}^2$

Corrected Model: a lineáris modellel becsült és a megfigyelt értékekre illesztett lineáris függvény jóságát mutatja. Eldönthető, hogy az alkalmazott modell megfelelő-e. $SS_R = \sum (\hat{Y}_i - \bar{Y})^2 = \frac{SP_{xy}^2}{SS_x}$

r-négyzet értéke Corrected Model SS/Corrected Total SS, (119/558).

Ha az általános lineáris modell alkalmazása során a becsült (predicted values) értékeket is elmentjük, elvégezhetjük a lineáris regresszió-analízist (16. táblázat). A regresszió eredménye megkönnyíti a GLM táblázatának újbóli értelmezését.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.462 ^a	.214	.199	2.8518

a. Predictors: (Constant), Predicted Value for X

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	119.248	1	119.248	14.662	.000 ^a
	Residual	439.184	54	8.133		
	Total	558.431	55			

a. Predictors: (Constant), Predicted Value for X

b. Dependent Variable: X

16. táblázat: A lineáris regresszió-analízis eredménye

A lineáris függvény illesztése során kapott eltérés négyzetösszegek teljesen megegyeznek a GLM-vel kapott értékekkel. A lineáris regresszió-analízis táblázatának (ANOVA) értelmezése:

Total: az alapadatok eltérés négyzetösszege, szabadságfoka. Ez megegyezik a GLM Corrected Total értékével. $SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$

Regression: a lineáris modellel becsült és a megfigyelt értékekre illesztett lineáris függvény jóságát mutatja. Eldönthető, hogy az alkalmazott modell megfelelő-e. $SS_R = \sum (\hat{Y}_i - \bar{Y})^2 = \frac{SP_{xy}^2}{SS_x}$

Residual: maradékok négyzetösszege, szabadságfok, négyzetösszeg átlagok. A lineáris egyenessel meg nem magyarázott hatás.

Az r-négyzet értéke 0,214. Ez a Regression SS/Total SS hányadosa (119/558).

Egy-változós variancia-analízis (Univariate...)

A variancia-analízis során négyféleképpen tudjuk kiszámítani az eltérés négyzetösszegeket (SS). Római számokkal jelölöm a négy típust (I-IV.). A programban kezdőértékként a III. jelenik meg, ezt használhatjuk az egy vagy több-tényezős, kiegyensúlyozott (balanced) vagy kiegyensúlyozatlan (unbalanced), teljes, azaz nincs hiányzó parcella adatú kísérletek kiértékelésekor (ez a leggyakoribb). Ez a módszer megegyezik a széles körben ismert Yates-féle módszerrel. A Yates módszer lényegében az átlagok súlyozott eltérésnégyzet technikáját használja a négyzetösszegek számításakor. Ez a módszer jól ismert a mezőgazdasági kutatásban, mivel Sváb könyveiben a variancia-analízis ismertetésekor ezt a technikát mutatja be.

Type I: ezt kell használni, ha a kezelésekben nem egyezik meg a megfigyelések száma, hiányzó parcellaadat van.

Többváltozós variancia-analízis, (Multivariate...)

Több kvantitatív tulajdonság együttes figyelembe vétele alapján kívánjuk kimutatni a kezelések hatása közötti különbségeket. Két kezelés közötti különbség szignifikanciájának vizsgálata, D^2 általánosított távolság tesztelése F-próbával. A DA a MANOVA határeset. Hotelling T^2 .

KÍSÉRLETEK TERVEZÉSE ÉS ÉRTÉKELÉSE ÁLTALÁNOS LINEÁRIS MODELLEL

Az alábbi fejezetekben a mezőgazdasági, földművelési, növénytermesztési, nemesítési, fajta összehasonlító, stb. kísérletek laboratóriumi és különböző szántóföldi kis-parcellás elrendezéseinek értékelését mutatom be a teljesség igénye nélkül. Az ismertetésre kerülő klasszikus elrendezések tanulmányozása és megértése segítséget nyújt a jövőbeli kísérletek megtervezéséhez és kiértékeléséhez. A fejezetekben az elrendezés rövid ismertetése után megadom a kísérlet vázrajzát, a matematikai modell leírását és a GLM-táblázat szerkezetét valamint a kiértékeléshez szükséges parancsokat, amit a parancsszerkesztő (syntax editor) ablakban lehet futtatni. Az elrendezéshez hű kiértékelés legfontosabb parancsa a DESIGN, ezért ezt a GLM-táblázat szerkezetében is megadom. Ezt követi a mintapélda GLM-táblázata, melyben a tényezők, négyzetösszegek, szabadságfokok, átlagos négyzetösszegek, F-próbák eredményei, valamint a szignifikancia szintek láthatók.

Elméleti áttekintés

A variancia-analízis modellben a függő változókat magyarázzuk független változó(k) segítségével. A magyarázat a függő változó teljes heterogenitásának¹ két részre bontását jelenti. A teljes heterogenitás egyik része az, amelynek „okai” a független változók, a másik heterogenitás-rész pedig az, amelynek „okait” az egyéb, általunk nem vizsgált tényezők tartalmazzák. Ez utóbbit sokszor a véletlen hatásaként is emlegetik. A heterogenitás mérésére többféle mérőszám szolgál:

(1) *range* (terjedelem); a legnagyobb és legkisebb érték közötti távolság

(2) átlagos eltérés; $\left(\delta = \frac{1}{N} \cdot \sum_{i=1}^N |x_i - \bar{x}| \right)$;

(3) *szórás*; $\left(\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \right)$;

¹ A változó heterogenitása azt jelenti, hogy az adott változó nem konstans.

(4) variancia- vagy szórásnégyzet; $\left(\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \right)$.

Ebből látszik, hogy a függő változónak magas (intervallum- vagy arányskála) mérési szintűnek kell lenni. Attól függően, hogy a független változók alacsony vagy magas mérési szintűek, eltérő magyarázó modelleket kell felépíteni. Ha ugyanis a független változóink nominális vagy ordinális mérési szintűek, akkor **variancia-analízissel** kereshetjük a magyarázatot a függő változó „viselkedésére”. Ha a független változók is magas mérési szintűek, akkor **regresszió-analízist** alkalmazhatunk. (Ha a függő változó alacsony mérési szintű, a magyarázatra szolgáló változók pedig magas mérési szintűek, akkor **diszkriminancia-analízist** használhatunk.)

A variancia-analízis során kettőnél több sokaság középértékeinek minta alapján történő összehasonlítása történik. Ezért nevezik a két-mintás t-próba általánosításának.

A variancia-analízis modellek olyan rugalmas statisztikai eszközök, amelyek alkalmasak valamely kvantitatív (numerikus vagy intervallum skálájú) változónak (függő változónak) egy vagy több nem feltétlenül kvantitatív változóval (független változók) való kapcsolata elemzésére. Arra vagyunk kíváncsiak, hogy van-e hatása a független változóknak a függő változóra, és a hatás különbözik-e vagy egyforma? A hatás, kapcsolat függvényyszerű leírása azonban nem célunk, még akkor sem, ha a független változók kvantitatívek. A regresszió-analízistől két szempont különbözteti meg a variancia-analízist:

A vizsgált független változók kvalitatívek is lehetnek (pl. a vizsgált személy neme, lakhelye stb.). Ebben az esetben ugyanis regresszió-analízis nem alkalmazhatunk.

Még ha a függő változók kvantitatívek is, nem cél a független változóval való kapcsolat természetének feltárása. A szórásanalízist tekinthetjük a regresszió-analízis vizsgálat megelőző vizsgálatának, ha ugyanis pozitív összefüggést kapunk a függő és független változó kapcsolatára, akkor van értelme vizsgálni az összefüggés jellegét.

Alap-fogalmak

Nézzük át azokat az alap-fogalmakat, amelyeket a variancia-analízis során használunk.

Faktor: Faktornak nevezzük a vizsgálatba bevont független változókat, pl. különböző kezeléseket, tényezőket.

Faktor szint: A faktor értékészletének az eleme, mely beállítása mellett vizsgálhatjuk meg a függő változónkat. A kezelések szintjei, pl. műtrágyaadagok.

Kvalitatív és kvantitatív faktorok: Ha a faktorszintek nem numerikusak vagy intervallum skálájúak, akkor kvalitatív, ellenkező esetben kvantitatív faktorokról beszélünk.

Kezelések (cellák): Egy-faktoros esetekben a kezelések megfelelnek a faktorok szintjeinek, több-faktoros esetben a figyelembe vett faktorok szintjeiből előálló kombinációk a kezelések. Pl. amikor a 2 faktor műtrágyaadagok és öntözési módok, akkor a kezelések a (műtrágyaadagok, öntözési módok) összes lehetséges kombinációjából áll.

Interakció: Két változó kapcsolatában akkor áll fenn interakció (kölsönhatás), ha x_1 változó hatása függ az x_2 változó szintjétől és fordítva.

Egy-szemponos variancia-analízis: Variancia-analízis, ahol csak egy faktor van.

Több-szemponos variancia-analízis: Variancia-analízis, ahol kettő vagy több faktor van.

Egy-változós variancia-analízis: ANOVA technika, amely egy függő változót használ.

Több-változós variancia-analízis: ANOVA technika, amely kettő vagy több függő változót használ.

A variancia-analízis alkalmazásának feltételei

A variancia-analízis adott n számú populáció középértékeinek minták alapján történő összehasonlítására szolgál (a két-mintás t-próba általánosításának tekinthető). A középértékre vonatkozó hipotézisek a következők:

H_0 : azoknak a populációknak a középértékei, amelyekből a minták származnak azonosak: $\mu_1 = \mu_2 = \dots = \mu_k$; H_A : legalább egy olyan középérték pár van, ahol a középértékek nem tekinthetők azonosnak: legalább egyszer $\mu_i \neq \mu_j$.

A variancia-analízis adatait a szokásos jelölésekkel 17. táblázat tartalmazza.

A statisztikai mintára alapozott variancia-analízis a következő lépésekben végezhető el:

A variancia-analízis modell felállítása.

A variancia-analízis kiszámítása, az F-próba.

A modell érvényességének ellenőrzése.

A középértékek többszörös összehasonlítása.

17. táblázat. A variancia-analízis adatai.

Sorszám	Populáció		Minta			
	várhatóérték	variancia	elemszám	mintaelemek	középértékek	variancia
1	μ_1	σ_1^2	r_1	$X_{11} X_{12} \dots X_{1r_1}$	\bar{X}_1	s_1^2
2	μ_2	σ_2^2	r_2	$X_{21} X_{22} \dots X_{2r_2}$	\bar{X}_2	s_2^2
·	·	·	·	·	·	·
n	μ_n	σ_n^2	r_n	$X_{n1} X_{n2} \dots X_{nr_n}$	\bar{X}_n	s_n^2

1. A variancia-analízis modell felállítása

A módszer alapgondolata szerint a modellben a mérési, megfigyelési értékeket összegként tekintjük. A k megfigyelés mindegyikére egy-egy modellegyenlet írható fel, amelynek alapján a mintaelemeken mért, ill. megfigyelt X_{ij} értékek felbonthatók a modell által meghatározott részekre és a hibára. A modell által meghatározott rész a szisztematikus hatásokat tartalmazza, a hibakomponens pedig a véletlen hatást jelenti.

A variancia-analízis legegyszerűbb modelljében a vizsgálatban szereplő n számú populációból egyszerűen véletlen mintát veszünk, majd a mintánkénti középértékeket hasonlítjuk össze, ezt nevezzük *egy-szemponthus variancia-analízisnek* (kísérlet esetén teljesen véletlen elrendezésnek). Az elrendezés modellegyenlete:

$$X_{ij} = \bar{\mu} + A_i + e_{ij}$$

ahol X_{ij} az i -edik minta j -edik eleme ($i=1, \dots, n$ $j=1, \dots, r_i$); $\bar{\mu}$ a kísérlet vagy minta főátlaga; A_i az i -edik mintához tartozó populáció hatása (növelheti vagy csökkentheti a főátlagot); e_{ij} véletlen hatás. Ebben a modellben a modell által meghatározott rész, csak az i -edik mintához tartozó populáció várható értékét tartalmazza, tehát szisztematikus különbséget csak a populációk várható értékei között feltételezhetünk. A véletlen okozta hatásokat a hibakomponens tartalmazza. Amennyiben teljesülnek a variancia-analízis alkalmazásának feltételei, akkor A_i összege nulla, és e_{ij} normális eloszlású nulla várhatóértékű sokaság, és független a blokk és kezeléshatástól.

A variancia-analízis alkalmazásának feltételei:

Az egyes kezelésekhez tartozó mintáknak függetleneknek kell lenniük. Ezt leginkább a kísérleti elrendezéssel, randomizálással biztosíthatjuk. A kísérleti elrendezésekről a vonatkozó fejezetben szólnunk.

A függő változó eloszlása normális legyen, pontosabban az e_{ij} maradéknak kell normális eloszlásúnak lennie. Attól, hogy egy normál eloszlású mintához egy konstans értéket hozzáadunk, vagy abból levonunk, az eloszlás és a minta szórása nem változik. A normalitás vizsgálatát korábban ismertetett módszerek valamelyikével ellenőrizhetjük. (Megjegyezzük, hogy a matematikai-statisztikai kézikönyvek az ANOVA-t robusztus eljárásnak tekintik, s azt állítják, hogy a függő változónak nem kell normális eloszlásúnak lennie). Ha matematikailag korrekt módon akarjuk az ANOVA-t használni, akkor a függő változót normális eloszlásúvá transzformálhatjuk.

A minták szórásnégyzetei egyezzenek meg $(\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2)$. (Az SPSS programnál ezt a homogenitást a Levene teszt alapján tesztelhetjük: ANALYZE/COMPARE MEANS/ONE-WAY ANOVA menüben az OPTIONS alatt jelölhetjük ki.)

Példa: egy-szemponos variancia-analízisre. Egy termesztő k kukoricafajta termesztése között választhat. Jelöljük a fajtákat A, B, C, D-vel. Döntsük el, hogy a 4 fajta termesztése esetén azonos terméseredményre számíthatunk-e.

18. táblázat. Kukoricatermés (t/ha)

Fajta	Termés (t/ha)		
A	9,3	7,2	8,2
B	5,4	7,1	5,9
C	4,5	2,9	5,0
D	3,5	0,9	2,5

A μ_i értékek a négy fajtapopuláció ismeretlen középértékeit jelentik, amiket az \bar{X}_i -vel tudjuk becsülni.

19. táblázat. Az alapadatok munkatáblázata

Fajta n_i	Termés (t/ha) X_{ij}			$\sum_i X_i$	\bar{X}_i
A	9,3	7,2	8,2	24,7	8,23
B	5,4	7,1	5,9	18,4	6,13
C	4,5	2,9	5,0	12,4	4,13
D	3,5	0,9	2,5	6,9	2,30
Összesen:				62,4	5,20

A közös $\bar{\mu}$ becslésére a kísérlet főátlaga szolgál.

EGY-TÉNYEZŐS VARIANCIA-ANALÍZIS AZ SPSS-BEN

Segítségével egy tényező hatását lehet vizsgálni a függő változó mennyiségi alakulására. A tényező, faktor valamilyen csoportképző ismérvvvel rendelkezik, a függő változó pedig legtöbbször skála típusú adat. Egyszerre több függő változót is kijelölhetünk az analízis számára.

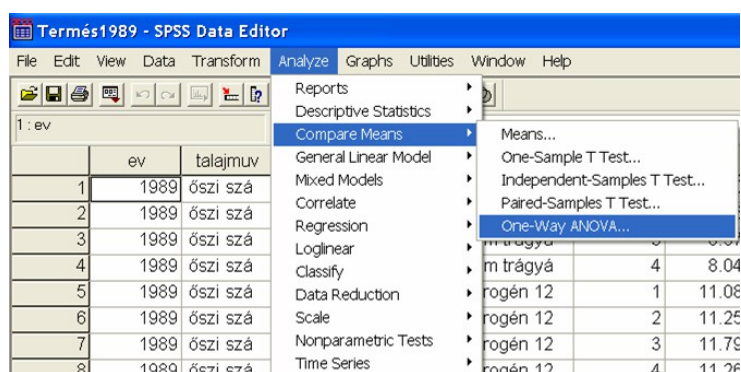
Amennyiben az analízis az átlagok közötti egyenlőséget nem igazolja, szükséges az átlagok közötti különbségek kimutatása. A variancia-analízist kiegészítő középérték összehasonlító teszteknek kétféle típusa létezik:

előzetes, ún. a priori kontrasztok és

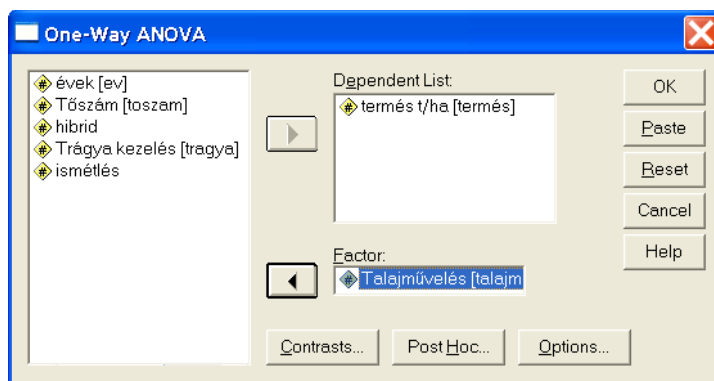
az analízis után elvégezhető, ún. post hoc analízisek

A kontrasztokat tehát a kísérleti adatok elemzése előtt kell előállítani, és így elvégezni az elemzést.

Az egy-szemponos szórásanalízis F-próbája akkor ad α -szinten szignifikáns eredményt, ha ezen a szinten létezik szignifikáns kontraszt a csoportok között.



15. ábra: Egy-tényezős variancia-analízis



16. ábra: A változók és a tényező megadása

változót helyezzük a DEPENDENT LIST ablakba, míg FACTOR-ént definiáljuk a talajművelést, hiszen a termésnek a talajművelési változatok közötti különbségét próbáljuk igazolni. Amennyiben a variancia-analízis a talajművelés szignifikáns hatását igazolja, kíváncsiak leszünk, hogy a három talajművelési változat közül vajon melyik között van lényeges (szignifikáns) különbség. A variancia-analízis után elvégzendő középérték összehasonlító tesztek helyes alkalmazásához azonban tudni kell, hogy a csoporton belüli

Példa: vizsgáljuk meg, hogy három talajművelési változatban hogyan alakul a kukorica termése. Az egy-tényezős variancia-analízis alkalmazásához kattintsunk az ANALYZE menüpont COMPARE MEANS almenüjében az ONE-WAY ANOVA parancsra (15. ábra).

A statisztikai számítás elvégzéséhez a vizsgált függő

variancia vajon megegyezik-e. Ezeket a tesztek a későbbi fejezetekben mutatjuk be.

20. táblázat: A variancia-analízis eredménye

ANOVA

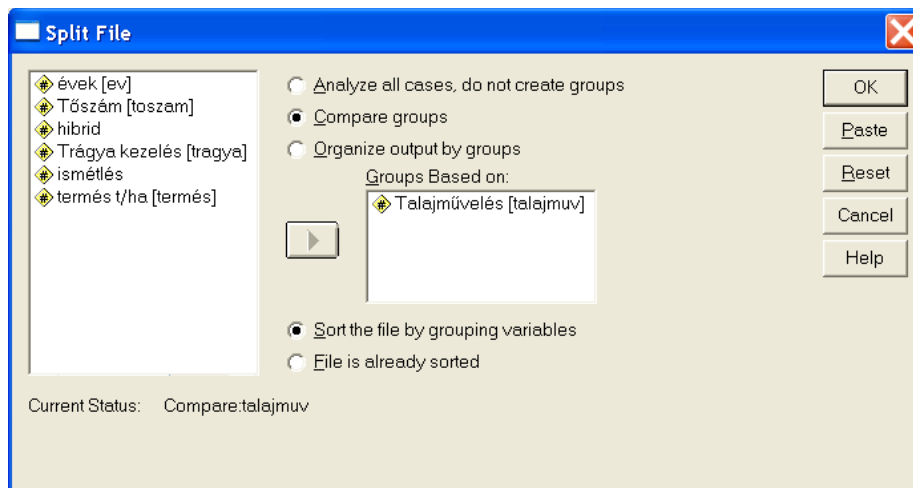
termés t/ha					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	92.524	2	46.262	10.087	.000
Within Groups	646.657	141	4.586		
Total	739.181	143			

A fenti táblázat a szórás-elemzés eredményét mutatja. 5%-os elsőfajú hibát választva, megállapítható, hogy a talajművelési változatokban a kukoricatermése szignifikánsan különbözik. Hangsúlyozzuk, hogy az F-próba eredménye csak akkor fogadható el, ha a vizsgált változó normál eloszlású és a csoportokon belüli varianciák megegyeznek.

A modell érvényességének vizsgálata

Normalitás vizsgálat

A variancia-analízis alkalmazhatóságának feltétele, hogy a függő változó normális eloszlású legyen, pontosabban a különböző kezelések mintáinak

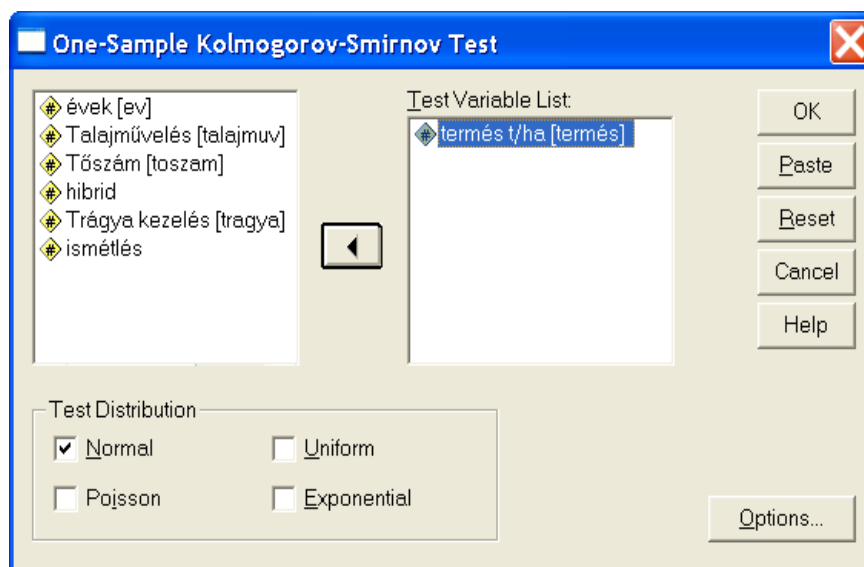


17. ábra. Az adatbázis megosztása

(lényegében azonban a hibának, vagy eltérésnek) kell normál eloszlásúnak lenni. A kezeléscsoportok elkülönített elemzéséhez meg kell osztani az adatbázist a DATA, majd a SPLIT FILE... kiválasztása után a megjelenő ablakban

válasszuk a COMPARE GROUPS rádiógombot, és a GROUPS BASED ON: ablakba helyezzük a „talajművelés” változót. Az OK gomb megnyomása után térjünk vissza az adatbázis ablakhoz.

Normalitás vizsgálatot az SPSS-ben többféleképpen is végezhetünk, pl. ANALYZE/NONPARAMETRIC TEST/1-SAMPLE K-S... a megjelenő párbeszédablakban (18. ábra) adjuk meg a vizsgálandó változót, és jelöljük be a normál eloszlást (alap esetben ez van megjelölve). A nullhipotézisünk ennek megfelelően az lesz, hogy a vizsgált változó eloszlása nem különbözik a normális eloszlástól. Válasszuk a szignifikancia szintet 5%-osra, és végezzük el az analízist az OK gomb megnyomásával. Az eredmény a 21. táblázatban látható.



18. ábra. Az egy-mintás Kolmogorov-Smirnov teszt

AZ ASYMP. SIG. (2-TAILED) sort tanulmányozva elmondható, hogy az őszi szántásos parcellák kukoricatermése normál eloszlású ($p > 0,05$), azonban a másik két talajművelési változat (tavaszi szántás, tárcsás) nem normál eloszlású, mert $p < 0,05$, vagyis elvetjük a nullhipotézist. A kapott eredmény alapján ebben az esetben nem szabadna variancia-analízissel értékelni a kísérletet. Vajon mi lehet ennek az oka? Sokszor a kiugró értékek, vagy adatrögzítési hiba okozza a hibát.

Homogenitás vizsgálat

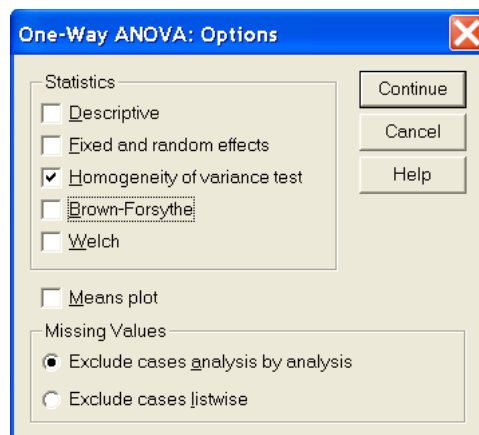
A varianciák homogenitásának ellenőrzésére az OPTIONS parancsgomb megnyomása után, a HOMOGENITY OF VARIANCE megjelölésével történik (19. ábra). A homogenitást Levene-teszttel állapíthatjuk meg. Visszatérve a variancia-analízis párbeszédablakhoz, és az OK gomb megnyomása után megkapjuk az eredményeket.

21. táblázat: A változó eloszlásának vizsgálata Kolmogorov-Smirnov próbával

One-Sample Kolmogorov-Smirnov Test			termés t/ha
Talajművelés őszai szántás	N		48
	Normal Parameters ^{a,b}	Mean	11.50673
		Std. Deviation	2.060577
	Most Extreme Differences	Absolute	.127
		Positive	.095
		Negative	-.127
	Kolmogorov-Smirnov Z		.882
	Asymp. Sig. (2-tailed)		.418
tavaszi szántás	N		48
	Normal Parameters ^{a,b}	Mean	10.30988
		Std. Deviation	2.068890
	Most Extreme Differences	Absolute	.227
		Positive	.148
		Negative	-.227
	Kolmogorov-Smirnov Z		1.574
	Asymp. Sig. (2-tailed)		.014
tárcsás	N		48
	Normal Parameters ^{a,b}	Mean	9.56033
		Std. Deviation	2.287441
	Most Extreme Differences	Absolute	.263
		Positive	.136
		Negative	-.263
	Kolmogorov-Smirnov Z		1.821
	Asymp. Sig. (2-tailed)		.003

a. Test distribution is Normal.

b. Calculated from data.



19. ábra. Homogenitás vizsgálat

Amennyiben a szignifikancia szintet előzetesen 5%-on rögzítettük, a talajművelés esetén megtartjuk a nullhipotézisünket ($p > 0,05$) (22. táblázat). Ez azt jelenti, hogy majd a kezelésátlagok összehasonlítása során nyugodtan

alkalmazhatjuk az egyenlő varianciákat feltételező tesztek. Abban az esetben, ha a Levene-teszt a varianciák különbözőségét igazolja (23. táblázat), nem használhatjuk a Fischer-féle tesztet. Ilyenkor robusztusabb próbát kell választani, pl. Brown-Forsythe vagy Welch próbát (WELCH, 1938).

22. táblázat: A talajművelési változatokon belüli varianciák egyenlőségének ellenőrzése

Test of Homogeneity of Variances

termés t/ha

Levene Statistic	df1	df2	Sig.
1.096	2	141	.337

23. táblázat: A talajművelési változatokon belüli varianciák egyenlőségének ellenőrzése

Test of Homogeneity of Variances

termés t/ha

Levene Statistic	df1	df2	Sig.
5,144	2	141	,007

A statisztika panelen különböző kiegészítő számításokat kérhetünk. Leíró statisztika (Descriptive): esetek száma, átlag, szórás, az átlag hibája, minimum, maximum, 95%-os konfidencia intervallum minden egyes csoportra. Fix és véletlen hatások (Fixed and random effects):

Brown-Forsythe próba

Ezt a próbát BROWN-FORSYTHE 1974-ben közölte először. A szórások különbözősége esetén meg kell vizsgálni, miért különbözik a szórás, milyen szakmai magyarázatot lehet rá adni. Ha a szórások különbözőségének semmilyen logikai vagy szakmai okát nem tudjuk megadni, nagy valószínűséggel a szórások véletlenül vagy valamilyen kísérleti hiba miatt különböznek.

A Welch és Brown-Forsythe-próba mezőgazdasági alkalmazásával még nem találkoztunk, ezért a több éves kutatómunka tapasztalatai alapján itt ragadjuk meg az alkalmat, hogy a használatukhoz néhány tanácsot adjunk. Ha a csoporton belüli szórás négyzetek (varianciák) nem egyformák nyugodtan használhatjuk a kezelésátlagok egyenlőségének tesztelésére bármelyiket a kettő közül. A legjobb, ha mindkettőt kipróbáljuk és összehasonlítjuk az eredményeket.

Válasszuk ki az OPTIONS párbeszédablakban (19. ábra) a Brown-Forsythe és Welch próbákat és futtassuk le a programot újból. A kapott eredményeket lentebb láthatjuk.

24. táblázat: A kezelés középértékek összehasonlítása robusztus tesztekkel

Robust Tests of Equality of Means

termés t/ha

	Statistic ^a	df1	df2	Sig.
Welch	3,238	2	83,571	,044
Brown-Forsythe	3,725	2	49,027	,031

a. Asymptotically F distributed.

Ebben az esetben a két teszt ugyanazt az eredményt adta, ha különbség lett volna a két eredmény között, tovább kell folytatni az értékelést. Ilyenkor szélsőséges esetben a Welch-próba szignifikáns különbséget mutathat a kezelés átlagok között, míg a Brown-Forsythe-próba nem. Mi lehet ennek az oka? Ez akkor következik be, ha a csoportok varianciája nagyon nagymértékben különbözik egymástól. Ilyenkor az elkülönített (separate) variancia tesztek a szabadságfok csökkentésével válaszolnak, és ezzel rontják a teszt eredményét. A varianciák nagyon nagy mértékű különbözőségét legtöbbször a csoportokon belüli kiugró értékek okozzák. A kiugró értékek zavaró hatását többféleképpen szűrhetjük ki. Az egyik hatásos eszköz a csonkított (trimmed) teszt, amikor minden egyes csoportból elhagyjuk a legnagyobb és legkisebb érték 15%-át. A csonkolás mértékét szakmai megfontolások miatt tetszőlegesen megváltoztathatjuk. A csonkolás után megismételt Brown-Forsythe próbában a szabadságfokok száma nőni fog és a teszt eredménye javul (25. táblázat). A fenti feltételek esetén a szórás hagyományos meghatározása helyett a *ROBUST SD* és *WINSORIZED SD* kiszámítása jobb becslést ad a csoporton belüli szórás nagyságára. Ezek a próbák kevésbé érzékenyek a kiugró értékekre. A különböző módon kiszámított szórások összehasonlítása közvetett módon, a csoporton belüli varianciák egyenlőségére vagy egyenlőtlenségére is rámutat. Szántóföldön tőszám kísérleteknél, ahol a varianciák egyezősége nem várható, a Welch vagy Brown-Forsythe- által kidolgozott variancia-analízist kell alkalmazni.

25. táblázat: A kezelés középértékek összehasonlítása robusztus tesztekkel a csonkolás után

Robust Tests of Equality of Means

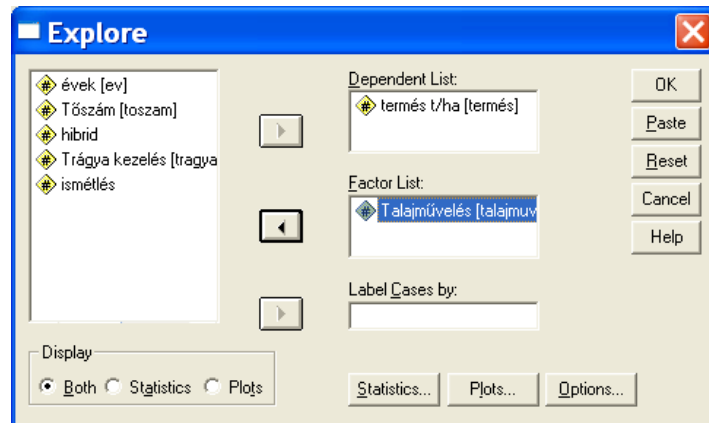
termés t/ha

	Statistic ^a	df1	df2	Sig.
Welch	9,905	2	93,797	,000
Brown-Forsythe	10,087	2	139,613	,000

a. Asymptotically F distributed.

Kiugró értékek vizsgálata

Az előző fejezetben láttuk, hogy a kiugró értékek milyen nagymértékben tudják megzavarni a varianciaanalízis eredményét. Ezért a statisztikai elemzések első és egyik legfontosabb lépése a kiugró értékek ellenőrzése. Az SPSS ennek ellenőrzésére is kínál lehetőséget.



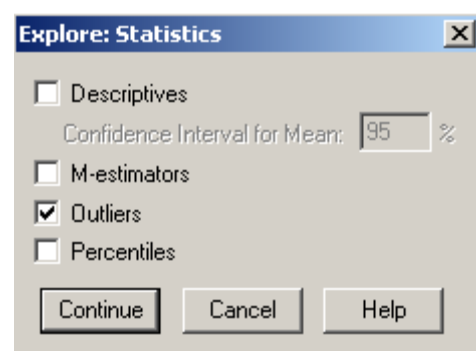
20. ábra. Az kiugró értékek vizsgálata

Az elemzés első lépéseként nézzük meg, hogy a kukoricatermés adataiban találunk-e kiugró értéket. Van-e vajon adatrögzítési, gépelési hiba?

A kiugró értékek ellenőrzése az ANALYZE / DESCRIPTIVE STATISTICS blokkjában az EXPLORE parancs szolgál.

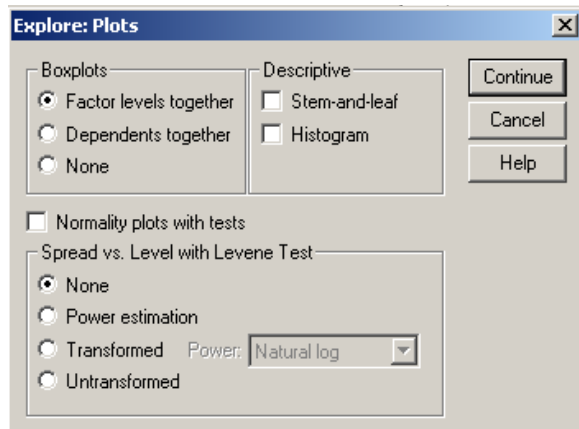
A DEPENDENT LIST mezőbe helyezzük a vizsgálni kívánt változót (változókat). Mivel mind a három talajművelésre ellenőrizni kívánjuk, hogy a kukoricatermés adatok tartalmazzanak-e kiugró értéket, a FACTOR LIST mezőbe helyezzük a „termés” változót. Ezzel érjük el, hogy a program a kiugró értékeket talajművelési változatonként külön-külön és nem összevont állományon ellenőrizze. A STATISTICS nyomógombra kattintva a megjelenő beszédpanelben válasszuk ki az OUTLIERS lehetőséget (21. ábra).

A beállítások elvégzése után futtassuk le a programot. A 26. táblázat talajművelési változatonként az öt legnagyobb és legkisebb értéket tartalmazza. A kiugró értékek ugyanis biztos, hogy itt keresendők, hiszen azok vagy sokkal nagyobbak, vagy sokkal kisebbek, mint a többi érték a mintában. A táblázatból jól látszik, hogy az őszi szántásos adatokban a 9. és 10. adat kiugró érték, adatrögzítési hiba miatt a tizedesvessző eggyel jobbra csúszott. A másik két talajművelésnél nem találunk kiugró eseteket. Hasonlóan végezhetjük el más változó esetében is a vizsgálatot.



21. ábra. Kiugró értékek megjelenítése

A kiugró értékek ellenőrzésének egy másik lehetséges módja az adatok grafikus megjelenítése. Az EXPLORE párbeszédpanel ablakából válasszuk a



22. ábra. A kiugró esetek grafikus ábrázolása

alatt és fölött jelennek meg a kiugró értékek (9. 10. adat őszi szántás) pontok formájában ábrázolva².

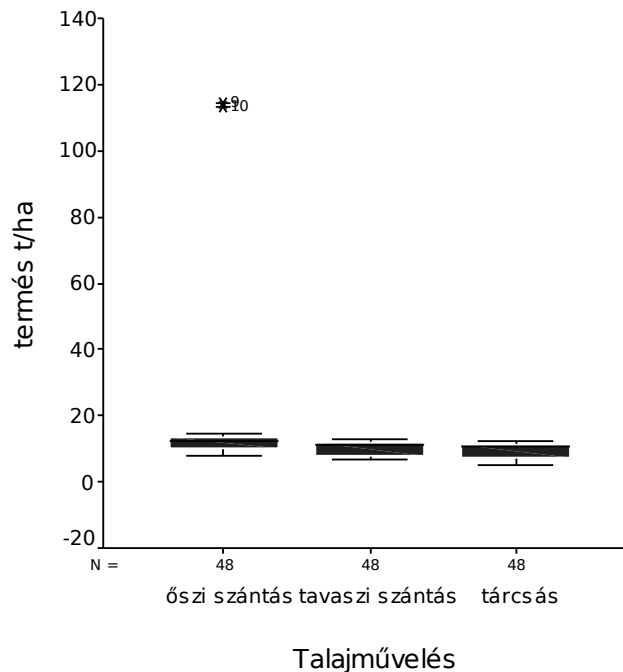
PLOTS lehetőséget (20. ábra). A BOXPLOTS panelrészben jelöljük meg a FACTOR LEVELS TOGETHER lehetőséget, majd a CONTINUES gombra kattintva menjünk vissza a főablakba, és ott az OK gomb megnyomásával hagyjuk jóvá a statisztika számítását. Ezt a vizsgálatot mindhárom talajművelésre futtatva a 23. ábrát kapjuk.

A megjelenő ábra legalsó és legfelső vonala a különböző talajművelésekben – a kiugró értékeket nem számítva – a mért legnagyobb és legkisebb értékeket jelölik. Az ábrán e két vonal

26. táblázat: A kiugró értéket ellenőrző táblázat

Extreme Values					
Talajművelés			Case Number	Value	
termés t/ha	őszi szántás	Highest	1	9	114,41
			2	10	113,51
			3	32	14,395
			4	35	14,392
			5	36	14,286
		Lowest	1	1	7,906
			2	2	7,957
			3	13	8,010
			4	4	8,043
			5	38	8,248
	tavaszi szántás	Highest	1	84	13,118
			2	83	12,859
			3	78	12,746
			4	82	12,672
			5	80	12,569
Lowest		1	88	6,715	
		2	87	6,865	
		3	85	6,929	
		4	49	7,078	
		5	62	7,099	
tárcsás	Highest	1	118	12,070	
		2	115	11,966	
		3	117	11,936	
		4	129	11,859	
		5	119	11,801	
	Lowest	1	135	5,355	
		2	134	5,421	
		3	136	5,652	
		4	122	5,697	
		5	124	6,059	

² A kiugró értékek ellenőrzésére az SPSS még számos lehetőséget kínál, ezek további bemutatására most nem kerül sor.



23. ábra. A kiugró értékek grafikus ellenőrzése (box-plot ábra)

A variancia-analízist kiegészítő középérték összehasonlító tesztek

Kontrasztok

A csoportok közötti eltérés négyzetösszeget (SUMS OF SQUARES) fel lehet bontani trend komponensekre, vagy előzetesen megadhatunk általunk definiált kontrasztokat is. A trendek között különböző hatványfüggvényekkel leírható trend-összetevőket tesztelhetünk.

A kontrasztok az egyes csoportok várható értékeinek lineáris kombinációi. A súlyok segítségével meg lehet adni a csoportviszonyokat, akár több kontrasztot is egyidejűleg. Ilyen csoportviszonyok a mezőgazdaságban, pl. műtrágyadózis kísérletekben nagyon könnyen értelmezhetőek. A lineáris összehasonlító függvények elméletével több szerző is foglalkozott. Magyar nyelven ÉLTETŐ Ö.-ZIERMANN M. 1964 megjelent művében található meg. A módszer lényege, hogy egy olyan lineáris függvényt kell alkotni, mint pl.

$$\lambda_g = c_{g1}x_1 + c_{g2}x_2 + \dots + c_{gp}x_p.$$

és ha teljesül a

$$c_{g1} + c_{g2} + \dots + c_{gp} = 0$$

feltétel, akkor ez egy lineáris összehasonlító függvény. A fenti definícióból következően végtelen számú λ_g létezik.

A kontrasztokra vonatkozó nullhipotézis: $H_g: \lambda_g = 0$, az ellenhipotézis: $A_g: \lambda_g \neq 0$.

Ha pl. egy tényező hatását T1, T2, T3, T4 szinten vizsgálunk, akkor a (T1, T2) csoport egybevetését a (T3, T4) csoporttal a $\lambda_g = x_1 + x_2 - x_3 - x_4$ függvény segítségével végezhetjük el (itt $1+1-1-1=0$).

A fenti összehasonlítás a variancia-analízis által szolgáltatott pooled variancia felhasználásával történik, ezért követelmény, hogy a csoportok szórásai megegyezzenek, így gyakran a variancia-analízis kiegészítő részét képezi. A contrast fejezetben a hatótényezők sokféle csoportosítása útján kapott átlagok különbözőségét lehet vizsgálni, pl. műtrágyázás esetén, a feltételezésem az, hogy az őszi búza a legnagyobb termést a 120 kg nitrogén adag mellett éri el. Vizsgálhatom az ez alatti adagokat, mintát véve, vagy az e feletti adagokat, szintén mintát véve, véletlenszerűen, ha nem 120 kg-t alkalmazok, vajon milyen eredmény születne.

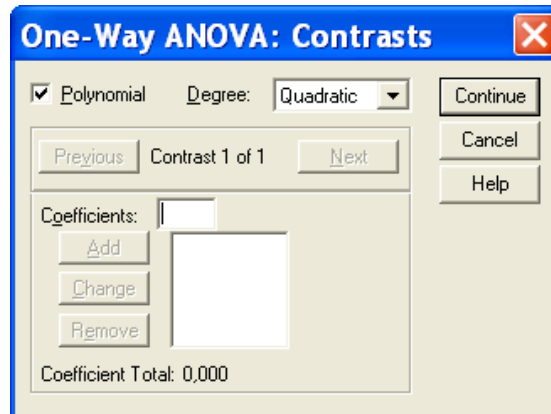
Korábban említettük, hogy a kontrasztok segítségével a csoportok közötti eltérés négyzetösszeget (Sums of Squares) fel lehet bontani trend komponensekre, ill. a trendek között különböző hatvány függvényekkel leírható trend-összetevőket tesztelhetünk. Ennek egyik gyakorlati alkalmazása műtrágyázás kísérletekben a hatásfüggvény lineáris, másodfokú, harmadfokú jellegének megállapítása.



24. ábra. Egy-tényezős variancia-analízis

Példa: kukorica műtrágyázási tartamkísérletben a nem trágyázott kezelésen felül öt különböző trágyaadagot alkalmaznak, ekvidisztáns távolságokra. (30, 60, 90, 120, 150 kg/ha N és PK). Függő változó a termés (t/ha), kezelés (factor) a műtrágyázás (24. ábra). A Contrasts... gombra kattintva állíthatjuk be a kontrasztokat. Két lehetőségünk is van. Polinomiális trendeket vizsgálhatunk, lineáris, kvadratikus, köbös, stb. tagokkal, vagy saját magunk adhatjuk meg a lineáris összehasonlító függvények együtthatóit. Az együtthatókra a korábban leírt szabályok az érvényesek. A műtrágyázás és termés közötti összefüggést gyakran másodfokú függvénnyel írják le. Erre jellemző, hogy egy határozott maximum pontja van, ami a legmagasabb termést jelenti. Az ehhez tartozó műtrágyaadag ismerete fontos a szakszerű tápanyag-visszapótlás elvégzéséhez. Válasszuk ki a legördülő listából a polinom fokát (Degree) másodfokúra (Quadratic). A Continue gomb

segítségével térjünk vissza az egyutas variancia-analízis ablakához, és az Ok gomb megnyomásával futtassuk az analízist.



25. ábra. Kontrasztok megadása

A korábban már megismert variancia-analízis eredménytáblázata további sorokkal bővült ki. Az első sor (Between) a műtrágyázás termésre gyakorolt hatását, ill. ennek F-próbáját mutatja. Amennyiben előzetesen a szignifikancia szintet 5%-osra választottuk, megállapíthatjuk, hogy a műtrágyázás szignifikánsan befolyásolja a kukoricatermést.

27. táblázat. Variancia-analízis eredménytáblázata, kiegészítve lineáris és másodfokú kontrasztokkal, trendkomponensekkel

TERMÉS			ANOVA				
			Sum of Squares	df	Mean Square	F	Sig.
Between	(Combined)		2025,772	5	405,154	77,247	,000
Groups	Linear Term	Contrast	1693,864	1	1693,864	322,955	,000
		Deviation	331,908	4	82,977	15,821	,000
	Quadratic	Contrast	296,019	1	296,019	56,439	,000
	Term	Deviation	35,889	3	11,963	2,281	,078
Within Groups			3493,101	666	5,245		
Total			5518,873	671			

A további sorokban a lineáris és négyzetes hatásgörbe tesztjei láthatók. A Linear Term a lineáris komponens hatását mutatja a termésre. A Contrast sora a lineáris tagot, a Deviation sora a maradék, nem lineáris, hanem egyéb, magasabb fokszámú polinomokkal jellemezhető részt mutatja. Lényegében az eltérés-négyzetösszegek (Sum of Squares) kerülnek felbontásra lineáris és egyéb összetevőkre. A lineáris tag szignifikáns $p < 0,05$. A maradék tag is szignifikáns, tehát érdemes megvizsgálni a másodfokú összetevőt is. A másodfokú összetevő sor (Quadratic Contrast) szintén a műtrágyázás szignifikáns hatását mutatja (Sig. 0,000), azonban a maradék tag már nem

$p > 0,05$ (Sig. 0,078). Ebben az esetben tehát a magasabb fokszámú polinomok bevonása a modellbe már nem indokolt.

Szimultán vagy többszörös összehasonlító tesztek

Szimultán vagy többszörös összehasonlítás (multiple comparison) a köztudatban a szórásanalízis kiegészítője, fejlődését főleg felhasználói igények indították útjára. Jelentősége azonban jóval nagyobb, különösen a nem paraméteres esetben, ahol szórásanalízisre, e normalitást feltételező eljárásra, nem kerülhet sor. Ha az egy-szemponos szórásanalízis F-próbája szignifikáns, kíváncsiak vagyunk, mely populációk miatt nem homogén a minta. Eleinte csak páronként az összes lehetséges csoport párra két-mintás t-próbát hajtottak végre. Előfordulhat azonban, hogy adott α -szinten szignifikáns F-próba esetén egyik csoport pár sem mutat szignifikáns t-értéket az adott α -szint mellett. A szimultán hipotézis vizsgálatok nemcsak az egy-szemponos szórásanalízisben hódítottak teret, hanem mindenütt, ahol egyidejű döntésre van szükség, pl. regresszió, kovariancia, több-szemponos szórásanalízis, stb.

Szimultán döntés, ha kettőnél több összehasonlítandó mintám van. Olyan állításokat fogalmazznak meg, amelyek egyidejűleg érvényesek. Ezek lehetnek:

Szimultán végzett statisztikai próbák vagy

Egyidejűleg érvényes konfidencia intervallumok

A többszörös statisztikai próbák zöme paraméteres, a normális eloszlásra épülő eljárás. Sorozatos statisztikai összehasonlítások végzésekor halmozódik a próbaként vállalt elsőfajú hiba (kockázat). A szimultán összehasonlítási módszerek fő célkitűzése ennek a halmozódásnak a csökkentése illetve megszüntetése. Ennek eredményeként az egyes összehasonlítások konzervatív irányba tolódnak el: a próbánként fenyegető elsőfajú hiba ténylegesen kisebb a vállalt (névleges) kockázatnál. Ez azonnal szembeötlik a többszörös összehasonlítások azon csoportjánál, amelyek az ún. Bonferroni-egyenlőtlenség alapján dolgoznak. Az első ilyen javaslat Fisher könyvében (1935) található. A lényege, hogy m összehasonlítás esetén, az egyes összehasonlításokat a névleges α szint helyett α/m valószínűségi szinten hajtják végre. A valószínűség szubadditív tulajdonsága miatt, ha az összehasonlításoként vállalt α_i kockázatok összege olyan nagy, mint a teljes sorozatra vállalt α valószínűségi szint, akkor annak valószínűsége, hogy m elvégzett összehasonlítás után valahol elkövetjük az elsőfajú hibát, legfeljebb α :

$$P(H) \leq \alpha = \sum_{i=1}^m \alpha_i$$

ahol: H esemény azt jelenti, hogy az állítások közt legalább egy hibás. Ha az egyes állítások (valószínűség-számítási értelemben) függetlenek lennének, akkor a fenti becslés helyett az

$$1 - P(H) = \prod_{i=1}^m (1 - \alpha_i)$$

egyenlőséget alkalmazhatnánk, ami azt mutatja, hogy az állítások között nincs hibás. Miller (1966) megmutatta, hogy a szimultán konfidencia-intervallumokra a fenti egyenlőség helyett mindig a \geq érvényes. A szimultán vizsgált minták között végezhető összehasonlítások nem függetlenek. Legyen valamennyi α_i valószínűsége egyforma: $\alpha_i = \alpha_m = \alpha/m$, akkor az összehasonlítások nem független természetét figyelembe véve, a szimultán próbák együttes kockázata:

$$P(H) \leq 1 - (1 - \alpha_m)^m$$

A levezetésből látszik, hogy az egyes szintek egyformaságának semmiféle szerepe nincs. Megtehetjük tehát, hogy a fontosabb összehasonlítások számára magasabb szintet jelölünk ki, ezzel biztosítva számukra a nagyobb erőt.

Legkisebb szignifikáns differencia (LSD)

R.A.Fisher 1935-ben úgy módosította az egyszerű t-próbát, amennyiben a szórásanalízis F-próbája szignifikáns, akkor alkalmazhatjuk a legkisebb szignifikáns különbség (LSD) próbát, amelyben a közös hiba négyzetösszeg osztva a szabadságfokával (error mean square) becsli a varianciát. A mezőgazdasági kutatásban, a kísérletek kiértékelésben, a legrégebben használt módszer a kezelésszintek különbségének vizsgálatára. A varianciaanalízis szolgáltatja Hiba MQ-ból kiszámolt SZDp% -ból ($SzD_{p\%} = t_{p\%} s_d$) levont következtetések azonban csak akkor érvényesek, ha az analízis előtt véletlenül választunk ki két kezelésátlagot, és ennek a különbségét teszteljük. Általában a legnagyobb és legkisebb értéket adó kezelések közötti különbségek akkor is nagyobbak, mint az SZDp%, ha a kezelések véletlen minták ugyanabból a sokaságból, tehát nincs közöttük különbség. Erre a következtetésre jutott Sváb, 1981 is és a fenti hátrányok kiküszöbölésére a Duncan-tesztet említi, de az értékelés körülményes voltára hivatkozva nem foglalkozik vele. Sajnos a mezőgazdasági kutatásban is sokszor tévesen alkalmazzák az SZDp%-t és gyakorlatilag sorba tesztelik a kezelésszinteket, és azt nézik melyik két kezelés közötti különbség nagyobb, mint az SZDp%. Az így kimutatott szignifikáns különbségek igen kétes értékűek, mivel az α -hiba valószínűsége (a kockázat) az összehasonlítások során halmozódik, mivel az elsőfajú hiba a páronkénti összehasonlításra rögzített. Ez a teszt nem alkalmaz semmiféle korrekciót.

Newman-teszt

D.Newman (1939) dolgozta ki az első, studentizált terjedelmeken alapuló többszörös összehasonlító tesztet. Erre az eloszlásra először ő állított fel táblázatokat, később Pearson és Hartley (1943) részletesebb táblázatot készített. Ha a próba érték szignifikáns, akkor elhagyják valamelyik szélső értéket, és a következő terjedelmet vizsgálják tovább. Newman a próbát Student (alias W.S. Gosset) (1927) cikke alapján dolgozta ki. Statisztikája:

$$q = \frac{\bar{x}_k - \bar{x}_1}{s}$$

k , v paraméterekkel, ahol k a normál eloszlású populációk száma és

$$s^2 = \sum_{i,j} (x_{ij} - \bar{x}_i)^2$$

négyzetösszeg, amelynek szabadságfoka $v=k(m-1)$, ahol m a minta elemszáma.

Bonferroni-teszt

Páronkénti átlagok különbségének vizsgálatára használható, a két csoport elemszáma lehet különböző is. Lényege, hogy az α -hibához tartozó t-értéket korrigálja a független összehasonlítások számának megfelelően, így az elsőfajú hiba az összes lehetséges összehasonlításra rögzített (experimentwise error). Amennyiben k a lehetséges páronkénti összehasonlítást jelenti, akkor egy összehasonlításban az elsőfajú hiba valószínűsége α/k .

$$L = t(\text{táblázatbeli}) \sqrt{S_p^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Tukey-teszt, J.W. Tukey (1953)

Studentizált terjedelem tesztjében a p -elemű részcsoportokat ugyanazzal a kritikus értékkel hasonlítja össze. Itt a teljes vizsgálat elsőfajú hibája rögzített, és az egyes összehasonlítások elsőfajú hibája n növekedésével csökken, s így a másodfajú nő. A Tukey teszt (1953) alapesetben egyforma minta nagyságú csoportok átlagainak különbségét tudja tesztelni, és a következő null-hipotézist vizsgálja:

$$H_0: \mu_1 = \dots = \mu_k.$$

Ezt felbontja a következő hipotézisek metszetére:

$$t_{ij} = (\bar{x}_i - \bar{x}_j) \frac{\sqrt{mv}}{\sqrt{2s^2}}$$

$$H_{ij} = \mu_i - \mu_j = 0,$$

Ellenhipotézis: $A_{ij}: \mu_i - \mu_j \neq 0$. Mivel a minták azonos elemszámúak: $n_i = m$, ezért $\nu = k(m-1)$. Tehát a páronkénti egyenlőségeket szimultán teszteli. Statisztikája:

A H_{ij} hipotézist elfogadja α -szinten, ha $t_{ij} < t_{\alpha/2}$, ahol

$$P(t_{ij} < t_{\alpha/2} \mid H) = 1 - \alpha$$

Annak a valószínűsége, hogy a számított érték kisebb a táblázati értéknél, ha a nullhipotézis igaz, tehát a teljes elsőfajú hiba α . A $\sqrt{2}t_{\alpha/2}$ értékre α , k , ν függvényében studentized range néven táblázatokat készítettek.

A fenti összefüggésből H.Scheffé (1959) készített λ_g -re konfidencia-intervallumot. A H_g hipotézist akkor fogadjuk el, ha a konfidencia-intervallum tartalmazza a 0-t. Ezt kiterjesztett Tukey-próbának vagy T-módszernek nevezik.

Tukey és tőle függetlenül C.Y.Kramer (1956) javaslata alapján kiterjesztették nem egyenlő elemszámra. Ez a módszer a **Tukey-Kramer** módosított teszt. A teszt a Newman-Keuls-teszttel kiszámított legnagyobb különbséggel egyenlő, ezért is hívják Tukey's Honestly (őszinte, becsület) Significant Difference-nek. Később általánosították tetszőleges kontrasztokra is.

$$L = q(\text{táblázatbeli}) \frac{s_p}{\sqrt{n}}$$

Dunett (1980a) cikkében számítógépes szimulációval több szerző hasonló eljárását hasonlította össze és ezek közül a Tukey-Kramer próbát találta a legjobbnak, azaz a különböző elsőfajú hibák mellett a konfidencia-intervallum hosszát a legrövidebbnek.

H. Scheffé (1953) Scheffe-teszt

A hagyományos tesztek közé tartozik. Ő már valóban a H_g hipotéziseket vizsgálta. Az egyszerű F-próba akkor utasítja el a H_0 -hipotézist, ha létezik egy $a < 0$ vektor, amelynél a konfidencia-intervallum nem tartalmazza a 0-t. Ha k darab összehasonlítandó csoportom van akkor $k(k-1)/2$ összehasonlítást kell végezni. A statisztikája:

$$L = \sqrt{s_p^2 (k-1) F_{(táblázatbeli)} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

A teszt tetszőleges elemszámok esetén érvényes, és a paraméterek valamennyi kontrasztjának egyidejű vizsgálatára alkalmas. A kontrasztok szimultán vizsgálata legtöbbször a szimultán konfidencia intervallumok felírásával történik, és nézzük, hogy azok tartalmazzák-e a nullát vagy nem. Mivel a kontrasztok száma végtelen, a Scheffé által kezdeményezett kiterjesztés igen lényeges általánosítást jelent. Ez a módszer a legáltalánosabb, egyedül ennek van meg az a tulajdonsága, hogy ekvivalens a szórásanalízissel. Az olyan vizsgálatokat azonban, amelyek megfelelnek a

Tukey vagy Dunnett-módszer eredeti kérdésfelvetésnek (egyenlő elemszámú csoportok közötti különbségek vizsgálata ill. ezek egy kontrollal való összehasonlítása a cél) érdemes ezekkel a módszerekkel elvégezni, erejük ilyenkor nagyobb a Scheffé-módszer erejénél. A Scheffé-módszer ereje a Bonferroni-egyenlőtlenség alapján kiterjesztett t-próbákénál is kisebb mindaddig, míg az elvégzett összehasonlítások m száma lényegesen meg nem haladja az elvégezhető összehasonlítások dimenzióját (Miller, 1966) k független csoport egyszempontos összehasonlításakor ez a dimenzió $(k-1)$.

A Scheffe-teszt gyakorlati alkalmazásához nyújt nagy segítséget O BRIEN R. R. 1983 megjelent műve és LOTHAR SACHS 1985.

Dunnett-teszt

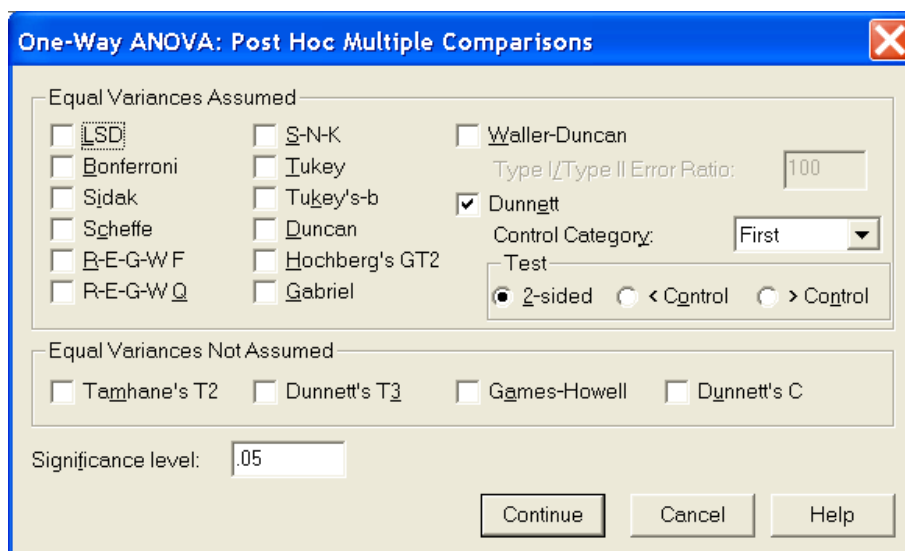
A Dunnett-teszt (1955) egy kijelölt csoportot (kontroll) hasonlít össze a többivel. Eredetileg egyenlő elemszámokra volt érvényes, de később elkészült az általánosítása egyenlőtlen elemszámokra is. Lényegét tekintve páronkénti összehasonlítást végez szimultán, de meg kell adni egy kezdő, kontroll csoportot, és ehhez hasonlítja a többi csoport átlagát. Statisztikája:

$$\bar{x}_i - \bar{x}_o \pm |d|s_p \sqrt{\frac{2}{n}}$$

\bar{x}_o = kontrol csoport átlaga

Statisztikája megegyezik Tukey statisztikájával, elfogadási tartománya viszont nem.

$$P(t_{ik} < t_{\alpha/2} | H) = 1 - \alpha$$



26. ábra. A Dunnett-teszt

Ehhez a statisztikához J. P. Shaffer készített konfidencia intervallumot, λ_g -re. Itt is a $H_0: \lambda_g = 0$ hipotézist elfogadják, ha az intervallum tartalmazza a 0-t. Ezt nevezik kiterjesztett Dunett-próbának.

Példa: 13 FAO 300-as éréscsoportba tartozó kukorica hibridet vizsgáltak azonos termesztési körülmények között. Az éréscsoport standardjának az Alpha hibridet választották, és ehhez hasonlították az összes többi. Az elsőfajú hiba halmozódásának elkerülésére a hibridek összehasonlítását Dunett-teszttel végezték.

A Dunett-tesztet az ANALYZE, COMPARE MEANS, ONE-WAY ANOVA, POST Hoc (26. ÁBRA) parancsok után érhetjük el. A teszt alkalmazása előtt ki kell választani a kontroll csoportot (Control Category). A párbeszéd ablakból csak az első vagy utolsó csoportot tudjuk kiválasztani a legördülő listából. Amennyiben más csoportot szeretnénk kontrollnak, ezt csak a Syntax Editor ablakban tudjuk megtenni. Továbbá meg kell adni, hogy az összehasonlítás egyoldalú vagy kétoldalú legyen. Alapbeállításaként kétoldalú összehasonlítás történik, kétoldalú szimmetrikus. Ebben az esetben nincs semmiféle előzetes információnk az összehasonlítandó párokról, bármelyik csoport lehet nagyobb, vagy kisebb, mint a kontroll. Egyoldalú próba esetében előzetesen már van információnk arról, hogy az összehasonlítandó csoport vagy csak nagyobb, vagy csak kisebb lehet, mint a kontroll csoport. Ez az információ sokszor valamilyen logikai feltételezésből ered. Az egyoldalú próba ereje nagyobb, mint a kétoldalú próbáé. Ez azt jelenti, hogy egy egyoldalú próbával ugyanolyan szignifikancia szint mellett már kisebb valódi különbség is kimutatható. Amennyiben nincs információnk a csoportok közötti relációról, mindig a kétoldalú próbát használjuk.

A CONTINUE gombra kattintással térjünk vissza az egy-tényezős variancia-analízis ablakhoz, és az OK gombbal futtassuk a programot. A Dunett-teszt eredményét,

nyét, az előre rögzített 5%-os szignifikancia szint mellett, a 28. táblázat mutatja.

Az első oszlopban az összehasonlítandó hibridek nevei, a másodikban a kontroll szerepel. A harmadik oszlopban a két hibrid termésének különbsége (t/ha) látható. A különbség melletti csillag 5%-os szinten szignifikáns különbséget jelez. A következő oszlopokban sorban a standard hiba, szignifikancia szint és a konfidencia intervallum alsó felső határa látható. A szignifikancia értéke a hibázás valószínűségét mutatja, ha elvetjük a nullhipotézist. Amennyiben a konfidencia intervallum magában foglalja a nullát, meg kell tartani a nullhipotézist.

28. táblázat. A Dunnett-teszt eredménye

Multiple Comparisons

Dependent Variable: TERMÉS
Dunnett t (2-sided)^a

(I) HIBRIDEK	(J) HIBRIDEK	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Debreceni 351	Alpha	-3.0217*	.87697	.013	-5.5696	-.4739
Debreceni 377	Alpha	-2.5845*	.87697	.045	-5.1323	-.0367
Ella (Sze 361)	Alpha	-.5730	.87697	.998	-3.1208	1.9748
Mv 370 Hunor	Alpha	-1.6240	.87697	.406	-4.1718	.9238
Norma	Alpha	-1.7915	.87697	.297	-4.3393	.7563
Occitán	Alpha	-.7653	.87697	.977	-3.3131	1.7826
DKC 3511	Alpha	.0128	.87697	1.000	-2.5351	2.5606
DKC 4626	Alpha	-.1147	.87697	1.000	-2.6626	2.4331
Goldacord	Alpha	-.6185	.87697	.996	-3.1663	1.9293
LG 3362	Alpha	.2450	.87697	1.000	-2.3028	2.7928
Szegedi 352	Alpha	-.8767	.87697	.945	-3.4246	1.6711
PR38A24	Alpha	-1.2873	.87697	.675	-3.8351	1.2606

*. The mean difference is significant at the .05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

A 28. táblázat alapján csak az első két hibrid termelt kevesebbet, mint a kontroll, a többi terméskülönbség statisztikailag nem igazolható.

Student-Newman-Keuls próba

M.Keuls (1952) Módosította a Newman próbát. A statisztikája megegyezik Newmanével, az elsőfajú hiba összehasonlításonként rögzített, ezért a teljes vizsgálat elsőfajú hibája n -nel együtt nő.

$$w_r = q_{\alpha, r, v} \sqrt{\frac{s_p}{n}}$$

A próba teszteli, hogy mely kezelés kombinációk tartoznak egy homogén csoportba. Kiszámítása bonyolultabb, ezért célszerű számítógéppel elvégezni. Az eredmény grafikusán ábrázolható és könnyen értelmezhető. Legtöbb számítógépes program először az átlagokat sorba rendezi, kicsitől a nagy felé és vízszintes vagy függőleges vonallal jelzi a homogén csoportokat, ahol nincs szignifikáns különbség a kezelés kombinációk között. Véleményem szerint a kezelés kombinációk sorba tesztelésére a mezőgazdaságban is az egyik legjobban használható próba.

Duncan többszörös rang teszt (1955, 1965)

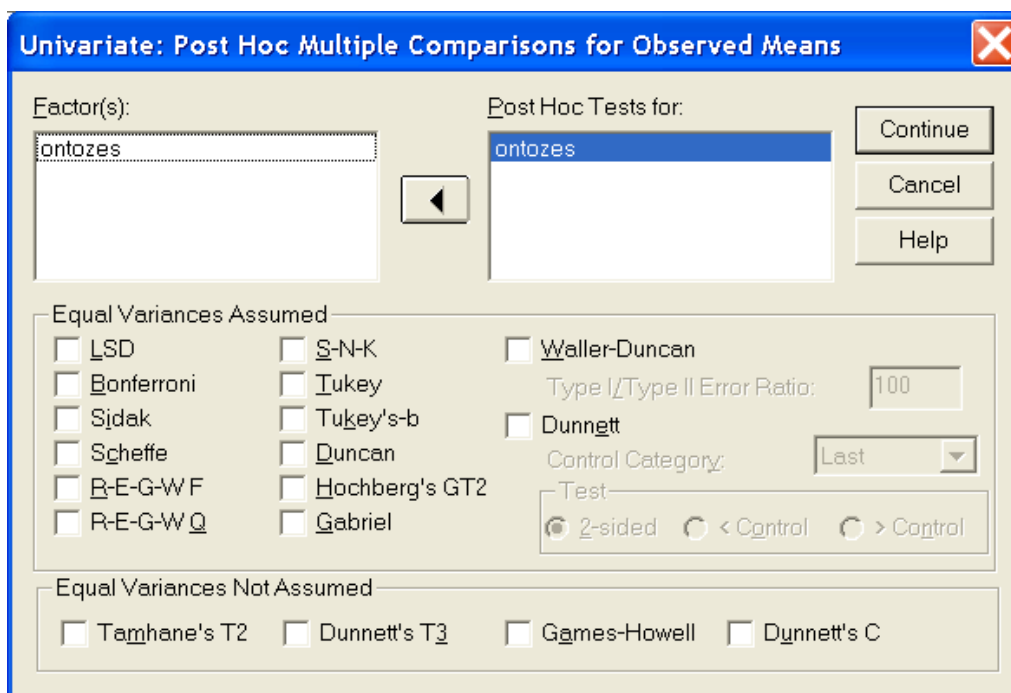
Itt is homogén csoportok képzése a cél. Napjainkban az egyik legjobbnak tartott többszörös összehasonlító teszt. Itt is a grafikus megjelenítés nagyban

segíti a kapott eredmények interpretációját. A mezőgazdasági kutatásban is potenciálisan nagy jelentőséggel bíró teszt.

Az SPSS-ben a variancia-analízis utáni középérték összehasonlító teszteket a POST Hoc... gombra kattintva érhetjük el (27. ábra).

Az újonnan megnyíló párbeszédablakban találjuk a teszteket. Az analízis utáni tesztek két nagy csoportját különíthetjük el: az egyikben a vizsgált csoportok varianciájának meg kell egyeznie (Equal Variances Assumed), míg a másik csoportban ennek a szigorú feltételnek nem kell teljesülnie (Equal Variances Not Assumed) (27. ábra). Ebben a fejezetben csak a páronkénti teszteket ismertetjük, a többszörös összehasonlító teszteket a következő fejezetben tárgyaljuk.

Példa: A homogén varianciákat feltételező tesztek közül válasszuk a leggyakrabban alkalmazottakat. Az LSD teszt (Least Significant Difference = legkisebb szignifikáns különbség) a legengedékenyebb a felsorolt tesztek közül, ami azt jelenti, hogy már nagyon kicsi középérték különbséget is szignifikánsnak mutat. A többi teszt ennél szigorúbb feltételeket támaszt, így sokszor előfordul, hogy az LSD-vel szignifikáns különbségek egy másik, szigorúbb teszt használatával már statisztikailag nem igazolhatók. Jelöljük meg a legengedékenyebb LSD és a legszigorúbb feltételeket támasztó Tukey tesztet. Válasszuk a szignifikancia szintet 10%-nak (Significance level = 0,10). Futtassuk le a programot.



27. ábra. Középérték összehasonlító tesztek

A 29. táblázat foglalja össze a vizsgált változó (Dependent Variable) és a talajművelések nevét. A kukorica termésének különbségét két-két talajművelést összehasonlítva a Mean Difference oszlop mutatja. Ezután következik a különbségek standard hibája (Std. Error), a számított p-érték (Sig.), illetve a 90%-os konfidencia intervallum alsó (Lower Bound) és felső (Upper Bound) határa.

Az értelmezést kezdjük az LSD-vel. Válasszunk ki egy sort, pl. őszi szántás tavaszi szántás. Itt a különbség plusz 1,197 t/ha, amit azt jelenti, hogy az őszi szántásban a kukorica ennyivel többet termelt. A szám mellett található csillag szignifikáns különbséget jelöl 10%-on. A Sig. oszlop az elsőfajú hiba valószínűségét mutatja, abban az esetben, ha elvetjük a nullhipotézist.

29. táblázat. Az LSD és Tukey teszt eredménye

Multiple Comparisons

Dependent Variable: termés t/ha

	(I) Talajművelés	(J) Talajművelés	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	őszi szántás	tavaszi szántás	1,19685*	,437141	,019	,29235	2,10136
		tárcsás	1,94640*	,437141	,000	1,04189	2,85090
	tavaszi szántás	őszi szántás	-1,19685*	,437141	,019	-2,10136	-,29235
		tárcsás	,74954	,437141	,203	-,15497	1,65405
	tárcsás	őszi szántás	-1,94640*	,437141	,000	-2,85090	-1,04189
		tavaszi szántás	-,74954	,437141	,203	-1,65405	,15497
LSD	őszi szántás	tavaszi szántás	1,19685*	,437141	,007	,47307	1,92064
		tárcsás	1,94640*	,437141	,000	1,22261	2,67018
	tavaszi szántás	őszi szántás	-1,19685*	,437141	,007	-1,92064	-,47307
		tárcsás	,74954*	,437141	,089	,02575	1,47333
	tárcsás	őszi szántás	-1,94640*	,437141	,000	-2,67018	-1,22261
		tavaszi szántás	-,74954*	,437141	,089	-1,47333	-,02575

*. The mean difference is significant at the .10 level.

A kockázat csak 0,7%, ami jóval kisebb, mint az előre megválasztott 10%, ezért nyugodtan, nagy biztonsággal elvethetjük a nullhipotézist. A konfidencia intervallum alsó és felső határa egyaránt pozitív előjelű, nem öleli körbe a nullát, ezért a két talajművelési változat között meglévő különbség valószínűleg tekinthető. Abban az esetben, ha a konfidencia intervallum magában foglalja a nullát, ld. Tukey teszt tavaszi szántás tárcsás sorát, akkor konzervatív irányba kell döntenet, meg kell tartani a nullhipotézist.

Az LSD módszer a legengedékenyebb az összes teszt közül, ezért ennél több szignifikáns különbséget nem lehet kimutatni a talajművelési változatok között. Vizsgáljuk meg a Tukey teszt eredményét, ami a legszigorúbb feltételeket támasztja az összehasonlítások során. Látjuk, hogy az LSD-vel szignifikáns tavaszi szántás és tárcsás talajművelés közötti különbség ezzel a teszttel statisztikailag már nem igazolható. Ezért is hívják „őszinte vagy becsületes” tesztnek, mert ha ezzel szignifikáns különbséget mutatunk ki, akkor az valódi különbség.

A Studentizált terjedelmet használó többszörös középérték összehasonlító tesztek (pl. Tukey) másik nagy előnye, hogy nem csak páronkénti összehasonlítás (29. táblázat) végezhető vele, hanem úgynevezett homogén csoportok is képezhetők a kezelések szintjeiből.

Két homogén alcsoportot kaptunk. Az elsőbe a tárcsás és tavaszi talajművelés parcelláinak terméseredményei nem különböznek szignifikánsan. Az elsőfajú hiba 20,3% százalék, ami sokkal nagyobb, mint a választott 10%, ezért homogénnek tekinthetők az ebben az alcsoportban található termésátlagok. A második alcsoportba egyedül az őszi szántás tartozik, ez szignifikánsan nagyobb termést eredményezett mind a tárcsás, mind a tavaszi talajműveléstől.

30. táblázat. Homogén alcsoportok képzése Tukey módszerével

termés t/ha

Talajművelés	N	Subset for alpha = .10	
		1	2
Tukey HSD ^a tárcsás	48	9,56033	
tavaszi szántás	48	10,30988	
őszi szántás	48		11,50673
Sig.		,203	1,000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 48,000.

Az One-Way ANOVA panel-ablakban megjelenő Options által felajánlott lehetőségeket tekintsük tovább (19. ábra). A Descriptive választásával egy összefoglaló táblázatot készíthetünk a vizsgált változóról. Ez a táblázat csoportonkénti bontásban tartalmazza a megfigyelések számát, átlagot, szórást, a középérték standard hibáját, a 95%-os konfidencia intervallum alsó és felső határát, valamint a minimum és maximum értékeket.

31. táblázat. A leíró statisztika eredménytáblázata

Descriptives

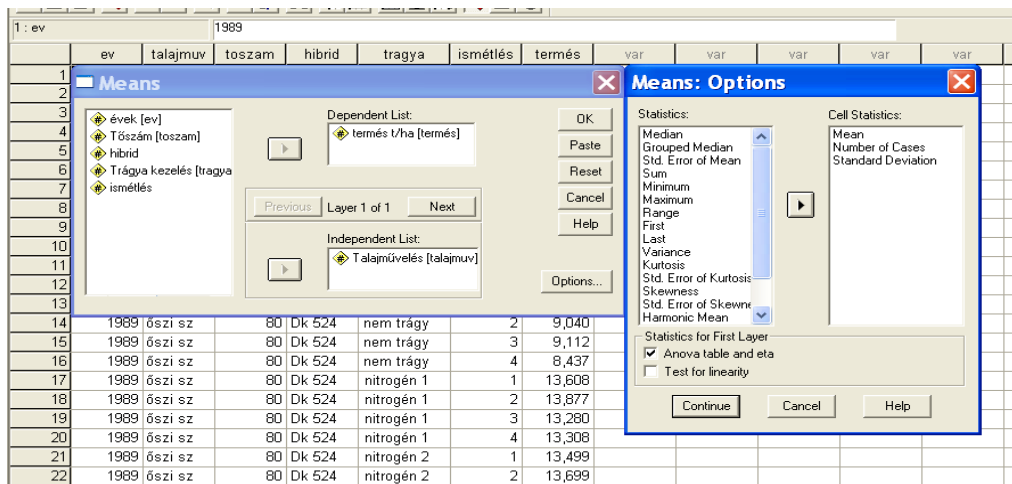
termés t/ha

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
őszi szántás	48	11,50673	2,060577	,297419	10,90840	12,10506	7,906	14,395
tavaszi szántás	48	10,30988	2,068890	,298619	9,70913	10,91062	6,715	13,118
tárcsás	48	9,56033	2,287440	,330164	8,89613	10,22454	5,355	12,070
Total	144	10,45898	2,273565	,189464	10,08447	10,83349	5,355	14,395

A Means Plot kiválasztásával a vizsgált változó csoportonkénti átlagértékei jelennek meg grafikusán. A Missing Values csoportban választható utasítások vagy az üres cellák (Exclude cases analysis by analysis), vagy pedig az üres

cellákat tartalmazó sorok (Exclude cases listwise) figyelmen kívül hagyását teszi lehetővé.

Már tudjuk, hogy a talajművelés szignifikánsan befolyásolja a kukorica termését. Vajon milyen mértékben magyarázható a talajműveléssel a termés varianciája? Ehhez nyissuk meg az Analyze menüpont Compare Means alpontjában a Means... párbeszéd ablakot (28. ábra).



28. ábra. Az ANALYZE/COMPARE MEANS/MEANS menü

A beállításokat a főablakban hasonlóan végezzük, mint korábban, majd az *OPTIONS* gombra kattintva tegyünk egy pipát az *ANOVA TABLE AND ETA* mellé. Futtassuk le a programot. Itt is megkapjuk a korábbi ANOVA táblát és a változók közötti összefüggés szorosságát mérő asszociációs táblázatot (32. táblázat).

32. táblázat. Az asszociáció mértéke a termés és a talajművelés között

Measures of Association		
	Eta	Eta Squared
termés t/ha * Talajművelés	,354	,125

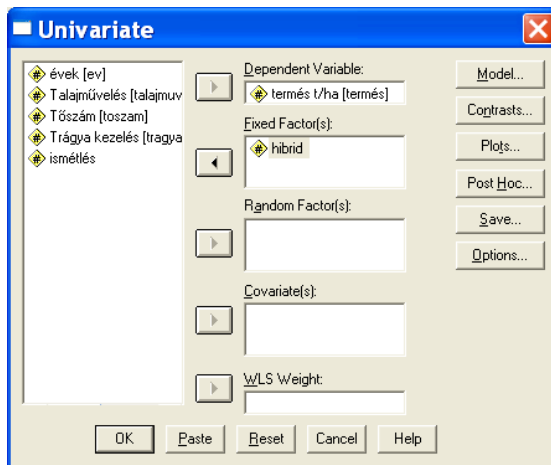
Az eta-négyzet alapján (Eta Squared) kijelenthetjük, hogy a talajművelés ebben az esztendőben 12,5%-ban befolyásolta a kukorica termésének változását. Ez elég kevésnek tűnik, de az adatok egy olyan tartamkísérletből származnak, ahol a talajművelés mellett tőszám, hibrid és műtrágyakezelések is szerepeltek. A műtrágyázás hatását a kukorica termésére a 33. táblázat mutatja.

33. táblázat. Az asszociáció mértéke a termés és a műtrágyázás között

Measures of Association		
	Eta	Eta Squared
termés t/ha * Trágya kezelés	,874	,763

ÁLTALÁNOS LINEÁRIS MODELLEK

Az SPSS programcsomagban az elrendezéshez hű egy-tényezős valamint több-tényezős variancia-analízist általános lineáris modellel helyettesítjük.



29. ábra. A GLM panelje

Az általános lineáris modell a hagyományos variancia-analízis és a lineáris regresszió-analízis ötvöze. Egyetlen táblázatban jelenik meg a szórás elemzés és regresszió-analízis eredménye (34. táblázat). Napjainkban a variancia-analízisnek nagyon sokféle technikája létezik, amik lehetővé teszik a feladat sajátosságainak figyelembevételével a legalkalmasabb értékelési módszer kiválasztását. Az elemzés megbízhatósága a hiba (error) meghatározásának módjától függ, ami

tulajdonképpen az eltérés négyzetösszeg (SQ) számítástechnikájának függvénye. Az SPSS lehetővé teszi a kísérleti elrendezéshez hű, a felhasználó által megalkotott lineáris modell megbízható értékelését.

34. táblázat. A GLM eredménytáblázata

Tests of Between-Subjects Effects					
Dependent Variable: X					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	119.248 ^a	3	39.749	4.706	.006
Intercept	20563.279	1	20563.279	2434.723	.000
FAJTA	119.248	3	39.749	4.706	.006
Error	439.184	52	8.446		
Total	21121.710	56			
Corrected Total	558.431	55			

a. R Squared = .214 (Adjusted R Squared = .168)

A megértés megkönnyítése érdekében az általános lineáris modellel kapott becsült (predicted values) értékeket mentjük el, és végezzük el a lineáris regresszió-analízist (35. táblázat). A regresszió eredménye megkönnyíti a GLM táblázatának értelmezését.

A függvényillesztés során kapott eltérés négyzetösszegek teljesen megegyeznek a GLM-vel kapott értékekkel. A GLM táblázatának értelmezése:

Corrected Model: a lineáris modellel becsült és a megfigyelt értékekre illesztett lineáris függvény jóságát mutatja. Eldönthető, hogy az alkalmazott modell megfelelő-e.

$$SS_R = \sum (\hat{Y}_i - \bar{Y})^2 = \frac{SP_{xy}^2}{SS_x}$$

Intercept: a kísérlet főátlaga

FAJTA: a kezelés okozta hatás.

Error: Sváb könyveiben a Hiba, a véletlen hatása, a meg nem magyarázott hatások.

Total: az adatok összes varianciáját mutatja.

Corrected Total: a lineáris regresszió-analízis összesen sora, a megfigyelt értékek eltérés négyzetösszege. Sváb könyveiben az Összesen sor.

$$SS_y = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

35. táblázat. A lineáris regresszió-analízis eredménye

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.462 ^a	.214	.199	2.8518

a. Predictors: (Constant), Predicted Value for X

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	119.248	1	119.248	14.662	.000 ^a
	Residual	439.184	54	8.133		
	Total	558.431	55			

a. Predictors: (Constant), Predicted Value for X

b. Dependent Variable: X

A variancia-analízis során négyféleképpen tudjuk kiszámítani az eltérés négyzetösszegeket (SS). Római számokkal jelölöm a négy típust (I-IV.). A programban kezdőértékként a III. jelenik meg, ezt használhatjuk az egy vagy több-tényezős, kiegyensúlyozott (balanced) vagy kiegyensúlyozatlan (unbalanced), teljes, azaz nincs hiányzó parcella adatú kísérletek kiértékelésekor (ez a leggyakoribb). Ez a módszer megegyezik a széles körben ismert Yates-féle módszerrel. A Yates módszer lényegében az átlagok súlyozott eltérésnégyzet technikáját használja a négyzetösszegek számításakor. Ez a módszer jól ismert a mezőgazdasági kutatásban, mivel Sváb könyveiben a variancia-analízis ismertetésekor ezt a technikát mutatja be.

Type I: ezt kell használni, ha a kezelésekben nem egyezik meg a megfigyelések száma, hiányzó parcellaadat van.

További lehetőségek a GLM-ben

Univariate options, Estimates of effect size. A hatás nagyságát tudjuk megbecsülni a parciális eta-négyzet meghatározásával. Ennek az értéke: $SSH/(SSH+SSE)$. Ahol SSH a független változó, vagy kölcsönhatás eltérés négyzetösszege, SSE a hiba, eltérés (error) négyzet összege. Ennek segítségével meghatározható a hatás nagysága, kiszűrhetők a legjelentősebb változók ill. kölcsönhatások.

SZÁNTÓFÖLDI KÍSÉRLETEK TERVEZÉSE ÉS ÉRTÉKELÉSE

Az alábbi fejezetekben a mezőgazdasági, földművelési, növénytermesztési, nemesítési, fajta összehasonlító, stb. kísérletek laboratóriumi és különböző szántóföldi kis-parcellás elrendezéseinek értékelését mutatjuk be a teljesség igénye nélkül. Az ismertetésre kerülő klasszikus elrendezések tanulmányozása és megértése segítséget nyújt a jövőbeli kísérletek megtervezéséhez és kiértékeléséhez.

A fejezetekben az elrendezés rövid ismertetése után megadjuk:

a kísérlet vázrajzát,

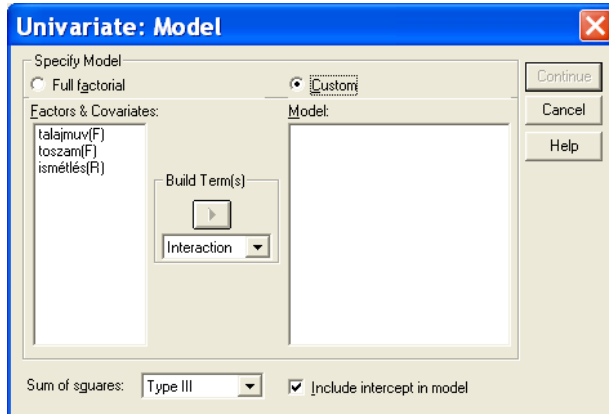
a matematikai modell leírását,

GLM-táblázat szerkezetét,

valamint a kiértékeléshez szükséges parancsokat, amit a parancsszerkesztő (SYNTAX EDITOR) ablakban lehet futtatni.

Az elrendezéshez hű kiértékelés legfontosabb parancsa a DESIGN, ezért ezt a GLM-táblázat szerkezetében is megadjuk. Ezt követi a mintapélda eredménytáblázata, melyben a tényezők, négyzetösszegek, szabadságfokok, átlagos négyzetösszegek, F-próbák eredményei valamint a szignifikancia

szintek láthatók. Az analízis után végezhető post hoc analízisekre nem térünk ki még egyszer, ezek teljesen megegyeznek a korábbi fejezetekben ismertetekkel. Ugyanezért nem ismételjük meg a GLM alkalmazási feltételeinek tételes vizsgálatát, mert ezek is megegyeznek a variancia-analízis alkalmazási feltételeivel. Néhány több-tényezős elrendezésben azonban

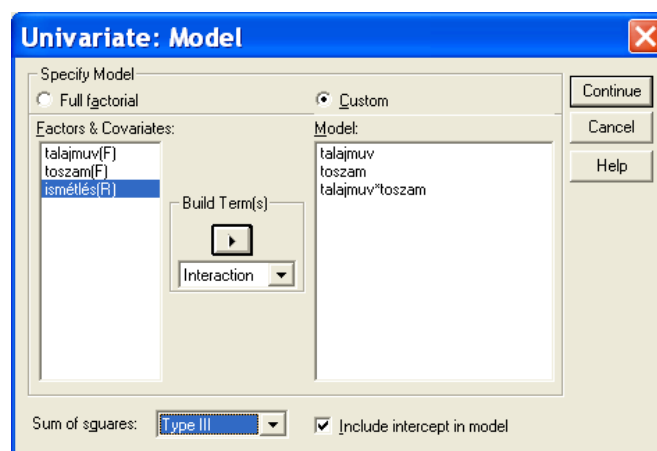


30. ábra. A GLM Modell ablaka

engedményeket tehetünk a szigorú alkalmazási feltételekből. Ilyenek például az osztott vagy kétszeresen osztott parcellás kísérletek.

A GLM nyitó párbeszédablakát mutatja a 29. ábra. Az egyik legfontosabb parancs a MODEL..., itt adhatjuk meg a kísérlet elrendezését. Alapbeállításként mindig teljes faktoriális kísérletként értékelődnek ki az adataink (Full factorial). Ilyenkor a főhatások mellett a tényezők kombinációjából

képezhető összes kölcsönhatás (interakció) is szerepel a lineáris modellben. Ezt úgy lehet megváltoztatni, hogy a CUSTOM rádiógombot jelöljük meg, amelynek hatására aktívvá válik a FACTORS & COVARIATES ablak. Ebben az ablakban kiválaszthatjuk a szerepeltetni kívánt változókat és ezek vizsgálni kívánt összefüggéseit. Jelöljük ki előbb a talajművelést, majd a tőszámváltozót, amiket a BUILD TERM(S) részben található nyilacska segítségével helyezhetünk a MODEL ablakba. A két változó kölcsönhatásának vizsgálatához egyidejűleg jelöljük ki mind a kettőt, és a legördülő listából válasszuk az INTERACTION lehetőséget, és a kis nyíl segítségével szintén helyezzük el a MODEL ablakban. A példában a talajművelés és a tőszám fix tényezőként, az ismétlés természetesen random tényezőként szerepel.



31. ábra. A GLM beállítása

A teljesen véletlen elrendezésű két-tényezős kísérlet lineáris modelljét mutatja az ábra. Alapbeállításként a lineáris modellben konstans is szerepel (INCLUDE INTERCEPT IN MODEL), ami legtöbb esetben a kísérlet főátlagának felel meg.

Az eltérés négyzetösszegek számítása a III. típus szerint fog történni. A CONTINUE gombra kattintva visszatérhetünk a főablakba, ahol futtathatjuk a programot (36. táblázat). Az elemzést a talajművelés*tőszám soron érdemes kezdeni. Látható, hogy e két tényező kölcsönhatása nem igazolható statisztikailag. A tőszám egymagában mint főhatás sem befolyásolta szignifikánsan a kukorica termését.

Egyedül a talajművelés befolyásolta jelentős mértékben a kukorica termését, amit a SIG. oszlopban található 0,000 érték mutat ($p < 0,05$).

A számítások automatikussá tehetők, ha a parancsszerkesztő (SYNTAX EDITOR) ablakban megadjuk a kiértékeléshez szükséges parancsokat. Ennek a legegyszerűbb módja, ha megnyomjuk a PASTE gombot a GLM ablakban (29. ábra). A gomb megnyomása után a parancsszerkesztő ablakba jutunk (32. ábra).

A beállításokat a szintaktikai szabályok betartása mellett szabadon megváltoztathatjuk. Végezhetünk fájl műveleteket, számításokat, ábrázolhatjuk az adatokat, különféle kimutatásokat készíthetünk, és a program olyan funkcióját is ki tudjuk használni így, amit a párbeszédpanelek segítségével nem tudunk beállítani.

36. táblázat. A több-tényezős GLM eredménytáblázata

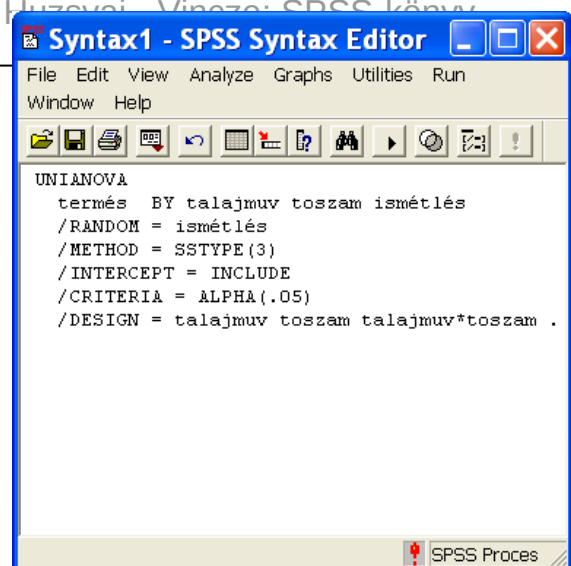
Tests of Between-Subjects Effects

Dependent Variable: termés t/ha

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	140,842 ^a	11	12,804	2,825	,002
Intercept	15752,195	1	15752,195	3475,103	,000
TALAJMUV	92,524	2	46,262	10,206	,000
TOSZAM	21,741	3	7,247	1,599	,193
TALAJMUV * TOSZAM	26,577	6	4,430	,977	,443
Error	598,339	132	4,533		
Total	16491,376	144			
Corrected Total	739,181	143			

a. R Squared = ,191 (Adjusted R Squared = ,123)

A RUN paranccsal azután futtathatjuk az általunk szerkesztett programot. Lehet soronként vagy egy kijelölt részfeladatonként futtatni a programot. Ez a legkényelmesebb módja az ismétlődő, nagy adatbázisokon végzett elemzések automatikussá tételéhez.



Az elrendezések ismertetésénél már nem térünk ki még egyszer az SPSS beállításaira, az itt leírtak minden egyes elrendezés esetén érvényesek.

32. ábra. A parancsszerkesztő

Kísérleti elrendezések

Leggyakrabban egy kísérlet célja az, hogy a különböző kezelések hatását összehasonlítsa, elemezze. A *kezelés* szót általánosságban kell érteni, nem szó szerint. Kezelésnek tekinthetők, pl. fajtakísérletekben a fajták, takarmányadag-kísérletekben a takarmányadagok. A kezeléshatásokat mérhetjük a termés hozamon, a növény magasságon, darabszámmal stb. A kezelés lehet egyetlen termés-kialakító tényező, pl. vetésidő, takarmányadag, fajta változásai változatai, vagy ezeknek a változatoknak különböző kombinációi. A kísérleteket eszerint két nagy csoportba osztjuk: egy-tényezős és több-tényezős csoportokba.

Egy-tényezős kísérletek esetében a kezelések egyetlen tényező változatai. Pl. pétisó-adag kísérlet esetében a pétisó adagja a vizsgált tényező és a változó adagok a kísérlet kezelései. Vetésidő kísérleteknél a vetés ideje a vizsgált tényező és a vetés változó időpontjai a kezelések; fajtakísérletekben a fajta a vizsgált tényező, és a különböző fajták jelentik a kezeléseket.

Több-tényezős kísérleteknél a kezelések több tényező változatainak kombinációi. Egyidejűleg kettő vagy több tényező változatait vizsgáljuk, és ezek kombinációit hasonlítjuk össze. Ha a vizsgált tényezők változatainak hatásai a kombinációkban nemcsak összegződnek hanem ezen túlmenően pozitív vagy depresszív összehatást is okoznak, akkor az a vizsgált tényezők kölcsönhatásban vannak egymással. Megkülönböztethetünk pozitív és negatív kölcsönhatásokat.

Míg egy-tényezős kísérletek esetében a kezelések száma a vizsgált tényező változatainak a számával egyezik meg, addig a több-tényezős kísérletekben többnyire a kezelések száma a tényezőkénti változatok összes lehetséges kombinációjának a száma. Pl. tekintsünk egy 6 fajtás búzakísérletben, ahol csak a fajta az egyetlen vizsgált tényező, ekkor a kezelések száma 6. Három-tényezős kísérletben 5 pétisó-adag, 3 műtrágyázási időpont és 4 fajta esetén a

három tényező változatainak a száma a kezelések száma: $5 \cdot 3 \cdot 4 = 60$. Jelöljük a kezelések számát V -vel, a tényezőket A, B, C, \dots -vel, ezek változatainak számát a, b, c, \dots -vel. Egy-tényezős kísérletekben a kezelések (kombinációk) száma V . Több-tényezős kísérleteknél $v = a \cdot b \cdot c \cdot \dots$.

Kvalitatív és kvantitatív tényezők, a kezelések megválasztása

A tényezők lehetnek kvalitatívek és kvantitatívek. A kvalitatív tényező változatai között minőségi különbség van, nem képeznek fokozatokat (pl. fajták, az öntözés módja, műtrágyakészítmények). A kvantitatív tényezők változatai fokozatokat jelentenek, amik folytonos és diszkrét értékeket vehetnek fel (pl. az öntözés mennyisége folytonos, az öntözés gyakorisága diszkrét tényező).

Attól függően, hogy a kísérletben kvalitatív vagy kvantitatív tényezőket vizsgálunk-e másként kell megfogalmazni a kérdést. Ennek megfelelően a kezelések megválasztása, az elrendezés és az értékelés is másképp alakul.

Kvalitatív tényező esetében a kérdés általában a vizsgált tényező meghatározott változataira vonatkozik. Ekkor a kezeléseket a vizsgált változatok képezik, és arra keressük a kérdést, hogy melyik két kezelés között van szignifikáns különbség és ez a különbség mekkora? Olyan kísérleti elrendezést kell megválasztani, amellyel a kezelések középértékei pontosabban hasonlíthatók össze.

Ha kvantitatív tényezőt vizsgálunk, akkor általában nem az a kérdés, hogy két meghatározott változata, fokozata között mekkora a különbség, hanem a hatásgörbe érdekel bennünket. Ezért úgy kell a kísérletet megtervezni, hogy a hatásgörbe minél több pontját meghatározzuk. Ha pl. összesen 24 parcellára van helyünk, akkor előnyösebb 8 kezelésfokozat egyenként 3 ismétlésben, mint 4 kezelésfokozat egyenként 6 ismétlésben. A könnyebb értékelés miatt a kezelésfokozatokat lehetőleg egyenlő „távolságokra” válasszuk meg és ne szabálytalan közökre. Pl. öntözési kísérletekben a vízmennyiség vizsgálata esetén 10, 30, 50, 70, 90 mm-es adagok képezzék a kvantitatív sor 5 szintjét³. A hatásgörbe meghatározáshoz legalább 4 kezelésfokozatnak kell lennie, mert ekkor tudjuk a lineáris és a négyzetes hatást elkülöníteni.

A hatótényezők (az általunk alkalmazott agrotechnikai beavatkozások) kvalitatív vagy kvantitatív jellegének elbírálása gyakran nagyon nehéz és nem egyértelmű.

Elővetemény, talajművelés, tőszám, fajta, öntözés és trágyázás hatását vizsgálva a termésátlag alakulására az alábbiakat kell figyelembe venni. Az

³ Ha az egyenletes távolság szakmailag nem helyes, akkor olyan fokozatokat érdemes megadni, aminél a fokozatok logaritmusai majd egységes fokozatokat képeznek (pl. 6, 12, 24, 48, 96).

előveteményt kvalitatív változóként érdemes figyelembe venni, mert olyan sokoldalú hatást fejt ki a talajra, hogy azt pontos számszerű paraméterekkel leírni nagyon nehézkes lenne. Kvantitatív változóként figyelembe véve meg kellene állapítani a különböző elővetemények talajra gyakorolt hatását, többek között, a teljesség igénye nélkül, hogyan hat a talaj vízgazdálkodására, mennyivel kevesebb vagy több vizet hagy maga után, mint az elővetemények átlaga. Nem is biztos, hogy az elővetemények átlagához kellene viszonyítani, és ha igen milyen növényeket, milyen súlyozással kellene bevonni az így kiszámítandó elővetemény átlagba. Vajon a hátrahagyott víz mennyisége vagy gradiense (mélységbeli, vertikális elhelyezkedése, rétegződése) számít? Valószínűleg mindkettő, de hogy milyen mélységben, milyen súllyal kell ezt figyelembe venni, függ attól a növénytől, aminek a termésátlag alakulását vizsgálom. A tápanyag-gazdálkodásra gyakorolt hatással a helyzet még bonyolultabb. Az egyes tápanyagok nem csak különböző mennyiségben és mélységben fordulnak elő a különböző elővetemények után, hanem különböző formákban is. A mikrobiológiai élet, biológiai aktivitás, gyomosság, növényegészségügyi kérdés számszerű megítélése a fentieknél is bonyolultabb. A felsorolt nehézségek miatt, egyelőre, célszerű az előveteményt kvalitatív tényezőként figyelembe venni.

Talajművelés, szintén nehéz megítélni a kvantitatív vagy kvalitatív jelleget. A kvantitatív jellegnél számszerűsíteni kellene a talajművelések közötti különbségeket. Ez lehetne a művelés mélysége, a lazultság állapotban bekövetkezett változás, a víz-levegő arány eltolódásának aránya stb. A változást nehéz számszerűsíteni, mert akadnak olyan változók is, amelyek térbeliek, pl. forgat vagy nem. Ezeket mátrixok segítségével vagy logikai változóként lehetne figyelembe venni. A talajművelés minőségének megítélése nehéz feladat közvetlenül a talaj-előkészítés után. Mi a jó talaj-előkészítés, ami a szemnek tetszetős, vagy ami után egyenletesen gyorsan kell a növényállomány, vagy ami után a legnagyobb termést kapjuk? Gyakran a fenti három meghatározás nem esik egybe és az egyéb körülmények hatása következtében a hatás nem egyértelmű. A talajművelést is véleményünk szerint helyesebb kvalitatív tényező gyanánt a vizsgálatba vonni.

A tőszámot mennyiségi tényezőként veszik figyelembe, ami véleményünk szerint helyes.

A fajta egyértelműen minőségi tényező. Ez az a "tényező" amit megfigyelünk, inkább megfigyelési egység (subject), mint kezelés. A fajta-összehasonlító kísérletek problematikája ezért egy kissé sajátos.

Az öntözést figyelembe szokták venni mind kvantitatív mind kvalitatív tényező gyanánt. A kvantitatív jellegnél a kiadagolt víz mennyiségét veszik figyelembe. Ez a vízmennyiség legtöbbször több öntözés összege, ezért nem egyértelmű a megítélés. Ugyanakkora vízmennyiség az öntözés időpontjától, a kiadagolt víznormától, intenzitástól stb. függően másképpen hat a termésátlag alakulására. Az öntözés hatása mindig összetett, nem csak a növény vízigény kielégítésén keresztül hat, hanem számos egyéb tényezőt is megváltoztat. Az öntözés lehűti a talajt, megváltoztatja a hőkapacitását, hőmérsékletvezető-

képességét ezeken keresztül az egész hőgazdálkodást. A talaj víz tartalmának változása a talaj levegő ellátottságát is megváltoztatja. A megváltozott hő-, víz-, levegőgazdálkodás megváltozott mikrobiológiai aktivitást von maga után. Megváltozik a tápanyagforgalom. Másképpen nő a növény, másképpen hat vissza a talajra, (árnyékolás, transzspiráció, stb.). A fenti problémákat mérlegelve érdemes az öntözést is kvalitatív tényezőként figyelembe venni a kísérletezés során.

A trágyázási kísérletek kiértékelésekor problémát szokott jelenteni, hogy az abszolút kontroll parcellát (ami nem kapott műtrágyát) a hatásgörbe kiszámításánál, amit legtöbbször másodfokú függvényvel közelítenek, figyelembe vegyék-e vagy sem. A kétféle elgondolás alapján számított egyenletek esetenként nagyon eltérhetnek egymástól. (SVÁB J., 1981) Vajon ekvidisztánsnak (egyenlő távolságúnak) vehető a nem műtrágyázott és az első műtrágya lépcső, a továbbiakban pedig a következő trágyalépcsők. Ha a fenti függvényt alkalmazzuk, a több éves tapasztalatok azt mutatják, hogy nem. A nem trágyázott és trágyázott kezelések minőségileg teljesen más kategóriába tartoznak, ezért trágyázott és nem trágyázott kezeléseket, mint kvalitatív tényezőket, érdemes elkülöníteni, és csak a trágyázott kezelésekből érdemes kiszámítani a hatásgörbét. Azonban a kutatási cél néha indokolhatja a kontroll parcella figyelembe vételét is.

A szervestrágyázási kísérletekkel viszonylag kevesebbet foglalkoznak, és itt is kvantitatív tényezőként veszik figyelembe a trágyázást. A szerves-trágyában lévő hatóanyag-tartalom alapján állítják a mennyiségi tényezők sorába. A kutatások kimutatták, hogy a szervestrágya jótékony hatása nem mindig a benne található makrotápanyag mennyiségtől függ, hanem az egyéb, a talajtulajdonságaira ill. a növény növekedésére kedvezően ható anyagok mennyiségétől. A szervestrágyázást is kvalitatív tényezőként vehetjük figyelembe a statisztikai elemzés során.

A fentiek ismeretében megállapítható, hogy a kísérletbe vont tényezők mindegyikét lehet kvalitatív tényezőnek tekinteni. A kvantitatív jelleg figyelembe vétele nagy körültekintést igényel és a szántóföldi kísérletezés terén szinte csak a műtrágyázás területén használható, bár itt is csak fenntartásokkal.

Az előbb ismertetett szempontok alapján az derül ki, hogy a mezőgazdasági szántóföldi kísérletek variancia-analízis útján történő értékeléséből a jelenségek kvalitatív leírására vállalkozhatunk csak. Egy tudományág az adott szakterületén mindig először a jelenség lefolyásának kvalitatív leírását adja meg. A felhalmozódott ismeretek és egy jó hipotézis eredményeképpen vállalkozhatnak a kvantitatív leírásra is és ez a szakember feladata. A matematikai leírás szolgáltatja mennyiségeket kísérleti úton ellenőrzik, és ha eltérés van korrigálják az egyenleteket. A mezőgazdasági kutatásban is először a jelenségek kvalitatív leírása a fő cél, amire a variancia-analízis az egyik hathatós eszköz lehet. Ha megvan a kvantitatív összefüggést leíró formula, amelynek egyes paraméterei szintén kísérleti úton lettek meghatározva, vállalkozhatunk a mezőgazdaságban is a számított és a

kísérletekben kapott, megfigyelt, értékek összehasonlítására. Az elméleti értékek és a megfigyelt értékek összehasonlításában szintén nagy szerepet kap a matematikai statisztika.

A lineáris modellek megalkotásakor el kell dönteni, hogy a hatásokat fix vagy random tényezőként vegyük figyelembe. A fix modellek főleg minősítő vizsgálatoknál használhatók, ahol adott feltételek mellett vizsgáljuk a hatótényezők viselkedését, és így az adott feltétel melletti dolgozás eredményeit kapjuk meg. A fix modellben legtöbbször kvalitatív tényezőket hasonlítunk össze. pl. az alábbi kérdésekre keresem a választ: Fajtáknál egyik fajta a másikonál, azonos termesztési technológia mellett, jobb-e? Előveteménynél, a szója jobb elővetemény, mint a kukorica? Öntözésnél, a háromszori 40 mm-es vízádagú öntözés jobb, mint az egyszeri 120 mm-es? Talajművelésnél, a 25 cm-es szántás jobb, mint a 15 cm-es tárcsázás? Műtrágyázásnál, az egyik műtrágyaféleség jobb, mint a másik, adott dózis mellett?

Random modellnél a tényezők hatásszintjei, amit a kísérletben alkalmazunk, az általunk vizsgált tényező reprezentatív mintája. Az ilyen modell általános érvényű összefüggések, törvényszerűségek felismerésének alapját jelentheti. Alkalmazása főleg több-szemponos szórás-elemzésnél a kevert modellek felépítésénél jelentős. A mezőgazdasági kutatásban való alkalmazás esetén felmerülhet a kérdés, milyen mintát vegyünk, ami hűen reprezentálja az általunk vizsgált tényezőt. A szakmai ismeretek birtokában erre szinte mindig megadható a válasz.

Ha az őszi búza trágya igényét akarjuk megállapítani, de nem érdekel bennünket a fajták közötti különbség, akkor az őszi búzát ilyenkor a köztermesztésben lévő fajtákkal jellemezhetjük. Milyen fajták vegyenek részt az analízisben? Célszerű a területi részesedés arányában kiválogatni a legjelentősebbeket. Abban az esetben, ha nincs *Fajta x Műtrágyázás* kölcsönhatás, amit előzetes vizsgálatok alapján meg lehet állapítani, elegendő lenne egyetlen fajta is.

Ha pl. kíváncsiak vagyunk, milyen változást okoz a talajművelés az őszi búza műtrágyázásában és nem célozom kiválasztani a legjobb talajművelést, csak jellemezni akarom a magyarországi őszi búza talajművelést. A random modellnél egyaránt vizsgálhatunk kvantitatív és kvalitatív tényezőket. Ha kvantitatív tényezőket vizsgálunk elsősorban a összefüggés milyensége (hatásgörbe) érdekel bennünket, és nem a konkrét dózisok közötti különbség. Ebben az esetben jó, ha ekvidisztánsan vagy logaritmikusan nőnek a kezelésfokokozatok. Ha nem valósítható meg, akkor sincs probléma, mert a legtöbb korszerű software -nél meg lehet adni a kezelésszintek egymástól való távolságát és az ortogonális polinomok segítségével így már a pontos hatás mutatható ki. Kvantitatív tényező vizsgálata esetén keverhetem a fix és random hatások elemzését.

A random vagy fix modell alkalmazása nem csak elméleti különbség, hanem a variancia-analízis számítása során, a variancia-komponensek különbözősége miatt, más számítási metódust is jelent.

A hatások felderítésére szolgáló modellek tehát legtöbbször lineáris matematikai modellek. Az alkalmazott matematikai modell nagyban meghatározza a kísérlet elrendezését is, egymástól elválaszthatatlanok. Fordítva is igaz, adott elrendezéshez csak meghatározott matematikai modell állítható fel.

Parcella, kísérleti egység

A *parcella*, *kísérleti egység*, *megfigyelési egység* kifejezéseket egymás szinonimájaként használjuk. A kísérletnek azt a legkisebb részét jelentik, amelyre a megfigyelésünk vonatkozik. Szántóföldi kísérletben a parcella szót alkalmazzuk és itt területet, az egész kísérleti tér legkisebb egységét értjük rajta.

Egy parcella csak egy kezelést reprezentálhat. Ezért kísérleti egység a parcella. Ha a kísérletben résztvevő parcellák eltérő kezelést kapnak, akkor más-más kezelést reprezentálnak, ha azonos kezelést kapnak, akkor ugyanannak a *kezelésnek az ismétlései*.

Az ismétlések jelentősége és száma

Különböző ellenőrizhetetlen hatások, az ún. kísérleti hibák parcellánként befolyásolhatják a kezeléshatásokat. Ha a kezeléseket több ismétlésben hasonlítjuk össze, feltételezhető, hogy a különböző ismétlésekben minden kezelést érnek pozitív és negatív hatású hibák. Az ismétlések számának növelésével egyre valószínűbb, hogy a pozitív és negatív hatású kísérleti hibák kiegyenlítődve jelentkeznek.

Az ismétléseknek kettős szerepet tulajdoníthatunk: (1) csökkenti a kísérleti hibák hatását, (2) lehetővé teszi a kísérleti hiba (ezen keresztül a középértékek közötti különbségek) becslését.

Az ismétlések számát azonban nem lehet minden határon túl növelni, hiszen a nagy ismétlésszám növeli a szükséges kísérleti egységek számát, ami által a kísérlet inhomogenitása is megnőhet.

Kísérletek elrendezési terve

Minél több a kísérleti egység, a parcella, annál kevésbé biztosítható minden parcellának azonos körülmény. Mivel a kísérleti hibaforrás egyik oka a

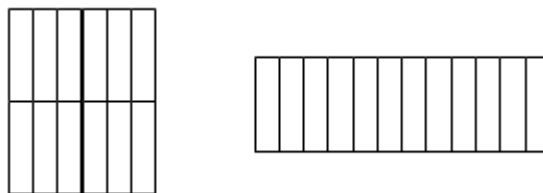
parcellák körülményeinek egyenlőtlensége, ezt csökkenteni kell. Ha magát az egyenlőtlenséget nem is tudjuk csökkenteni, akkor a parcellákból érdemes kisebb csoportokat képezni oly módon, hogy a csoporton belül a körülmények azonosak legyenek.

A parcellákból képzett csoportokat *blokkoknak* nevezzük. A blokk tehát valamilyen szempontból összetartozó parcellacsoportot jelent. Pl. szántóföldi kísérleteknél a szomszédos parcellák terület-blokkot, azonos időpontban végzett megfigyelések idő-blokkot, növényen az azonos rendű hajtások, azonos korú, azonos nemű állatok biológiai jellegű blokkot képeznek.

A legegyszerűbb és legáltalánosabb parcella-csoportosítás az, hogy az összes kezelés egy-egy parcellájából képzünk blokkot. Ekkor a kezelések teljes sorozata jelent egy blokkot. Az ilyen blokkot, amely az összes kezelés parcelláját tartalmazza *teljes blokk*nak nevezzük.

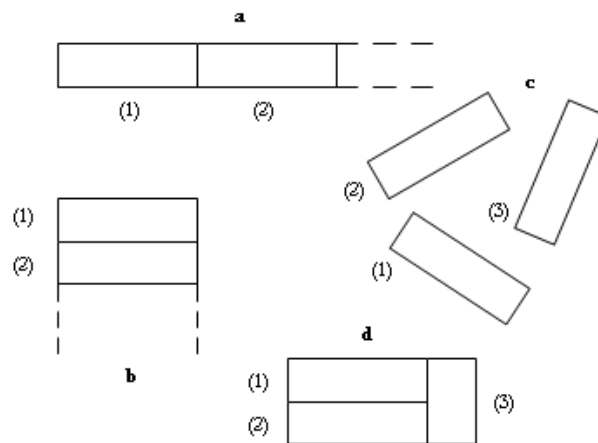
A kísérletek tervezésekor, amikor meghatározzuk a kezeléseket és az ismétlések számát, a kísérleti egységekből annyi teljes blokkot képezünk, ahány ismétlésünk van. Tíz kezeléssel négy ismétléses kísérlet összesen 40 kísérleti egységet 4 blokkba rendezzük, minden blokk 10 kísérleti egységet foglal magába, minden kezeléssel egy kísérleti egységet.

Egy 12 parcellás blokk számos kialakítási lehetőségei közül kettőt szemléltet az 33. ábra.



33. ábra. 12 parcellából képzett blokk (Forrás: Sváb János (1981): *Biometriai módszerek a kutatásban*, 93.o.)

A különböző ismétléseket jelentő blokkok lehetnek térben, vagy időben zárt egységben egymás mellett vagy szétszórva elhelyezve (34. ábra). A blokkok különböző alakúak is lehetnek, de legyenek azonos méretűek. A blokk mérete a parcellák (kísérleti egységek) számát jelenti.



34. ábra. A blokkok különböző elhelyezkedése. (Forrás: Sváb János (1981))

A kezelések elrendezése a parcellákon

A kísérleti terv elkészítésekor a kezelések elrendezése a parcellákon két fázisból áll:

A blokk-képzés után a megválasztott kísérleti elrendezésnek megfelelően elkészítjük az *elrendezés alaptervét*. Ez azt jelenti, hogy a kezelések sorszámát vagy egyéb jelét az elrendezés szerkezetének megfelelően beírjuk a parcellák helyére. Az alapterv mindig valamilyen szisztematikus elrendezés.

Az alapterv elkészítése után a kísérleti elrendezés szerkezetének keretén belül randomizáljuk, véletlenszerűen összekeverjük a kezeléseket. Így randomizálva kapjuk meg az *elrendezési tervet*. A randomizálással minden kezelésnek azonos esélyt adunk.

A 37. táblázat foglalja össze a továbbiakban bemutatásra kerülő kísérleti elrendezéseket.

37. táblázat. Kísérleti elrendezések összefoglaló táblázata.

Egy-tényezős kísérlek		Két-tényezős kísérletek
Teljesen elrendezés	véletlen	Véletlen blokkelrendezés
Véletlen elrendezés	blokk-	Osztott parcellás (split-plot) elrendezés
Latin négyzetes		Sávós elrendezés
		Az egyik tényező nincs

elrendezés	randomizálva
Latin téglá elrendezés	
Csoportosított elrendezés	
Három- és több-tényezős kísérletek	
Véletlen blokk elrendezés	
Osztott parcellás elrendezés	
Négy-tényezős kísérletek	

Egy-tényezős kísérletek

Teljesen véletlen elrendezés (CRD)

Példa: Négy kezelés hatását vizsgáljuk a tyúkok tojás-termelésére. Minden kezelésben 5 tojó van. Azonos idő alatt termelt tojások számán mérjük le a kezelések hatását. Az adatokat a 38. táblázat tartalmazza.

38. táblázat. 20 tyúk tojástermelése 4 kezeléssel 5 ismétléses teljesen véletlen elrendezésű kísérletben.

Kezelés	Adatok					Kezelés- összeg
1	94	86	52	83	60	375
2	11 4	81	97	10 1	12 8	521
3	90	88	78	10 2	45	403
4	70	58	90	54	65	337

Előfordul, hogy egyéb okok miatt az ismétlésekből nem lehetséges vagy nem célszerű a blokk-képzés, még akkor sem, ha azonos számú ismétlésünk van. Pl. állatokkal végzett kísérletekben kezelésként több állat lehet, és ezeket közösen tartjuk. Így állandóan keverednek, nem képezhetők fix blokkok.

A módszer általánosan alkalmazható azonos elemszámú minták illetve csoportok összehasonlítására is, ha (1) meghatározott szempontok szerint kiválasztott minták középértékeit hasonlítjuk össze, (2) utólag képezünk csoportokat, és ezek középértékeit hasonlítjuk össze. Jelöljük az alapadatokat x_1, x_2, \dots -vel, a kísérletek száma r , a kezelések száma v .

Az elrendezés matematikai modellje:

$$Y_{ij} = m + K_j + e_{ij}$$

ahol:

Y_{ij} = egy tyúk tojástermelése (db/tyúk)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

K_j = a kezelés hatása a tyúkok tojástermelésére

e_{ij} = a kísérlet hibája, a csoporton belüli szórás

39. táblázat. A GLM-táblázat szerkezete teljes véletlen elrendezésben

Tényező	SS	df	MS	F	Sig.	DESIGN
Korrigált modell						
Eltérés		1				
Kezelés (csop. között)		v-1				kezelés
Hiba (csoporton belül)		v(r-1)				
Összesen		rv				
Korrigált összesen		rv-1				

40. táblázat: A teljesen véletlen elrendezés SPSS parancsai

UNIANOVA
 tojás BY kezelés
 /METHOD = SSTYPE(3)
 /INTERCEPT = INCLUDE
 /CRITERIA = ALPHA(.05)
 /DESIGN = kezelés .

41. táblázat. A teljesen véletlen elrendezés eredménytáblázata,
 $v=4, r=5$

Tests of Between-Subjects Effects

Dependent Variable: Tojástermelés

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3784.000 ^a	3	1261.333	3.859	.030
Intercept	133824.8	1	133824.8	409.469	.000
KEZELÉS	3784.000	3	1261.333	3.859	.030
Error	5229.200	16	326.825		
Total	142838.0	20			
Corrected Total	9013.200	19			

a. R Squared = .420 (Adjusted R Squared = .311)

Véletlen blokk-elrendezés (RCBD)

Példa: A kukorica termését figyeljük 7 kezelés esetén. A kísérlet 5 ismétléses véletlen blokkelrendezésű. A parcellánkénti kezeléseket, és ismétléseket az alábbi táblázat mutatja:

Egyik legegyszerűbb és igen előnyös kísérleti elrendezés bármilyen témakörű kísérletben is az a fajta elrendezés, ahol a megfigyelési egységeket úgy csoportosítjuk, hogy egy csoportba minden kezelésből egy megfigyelési egység jusson.

Műtrágyázás							
1	2	3	4	5	6	7	(5)
2	7	5	6	7	3	1	(4)
5	6	2	3	1	7	4	(3)
3	1	4	5	7	6	2	(2)
4	3	1	7	2	5	6	(1)

ismétlés

35. ábra. Véletlen blokk elrendezés terve 7 kezelés (v) 5 ismétlésben (r)

Egy ilyen csoport képezi a blokkot, egyben egy ismétlés is. A blokkok száma így megegyezik az ismétlések számával. A blokkokon belül a kezelések randomizáljuk.

Az elrendezés előnye, hogy a kísérlet pontossága nem csökken, ha az ismétlések, azaz a blokkok, különböző körülmények között vannak. Az a fontos, hogy az egyes blokkon belül biztosítsuk az azonos feltételeket. A blokkok lehetnek egymástól távolabb is, ha ezt a terepakadályok szükségessé teszik, sőt lehetnek más körülmények között is.

A véletlenszerű blokkelrendezés hátránya, hogy minél nagyobb a blokk (vagyis minél több megfigyelési egységet tartalmaz), annál kevésbé biztosítható a megfigyelési egységek egyöntetősége és a kísérlet pontatlanabb lesz. Ha pl. négy ismétlést feltételezünk, 15-20 kezelésnél nagyobb véletlen blokkelrendezésű kísérletek nem ajánlottak.

Példa:

42. táblázat. *Parcella adatok kukorica kísérletben*

Kezelés	Ismétlés					
	(1)	(2)	(3)	(4)	(5)	(6)
1	18,9	17,6	16,4	16,4	14,4	14,8
2	16,4	16,7	14,7	14,4	12,6	13,8
3	10,4	13,5	13,9	8,7	11,5	12,3
4	17,4	17,7	15,7	17,5	16,8	18,3

Az elrendezés matematikai modellje:

$$Y_{ij} = m + R_i + M_j + e_{ij}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_i = blokk ill. ismétlés hatás a talaj heterogenitása, hogyan változik a talaj termékenysége fentről lefelé haladva

M_j = a műtrágyázás hatása a cirok termésére

e_{ij} = a kísérlet hibája

43. táblázat. A GLM-táblázat szerkezete véletlen blokk elrendezésben

Tényező	SS	df	MS	F	Sig.	DESIGN
Korrigált modell						
Eltérés		1				
Ismétlés		r-1				ismétlés
Kezelés (csop. között)		v-1				kezelés
Hiba		(r-1)(v-1)				
Összesen		rv				
Korrigált összesen		rv-1				

44. táblázat. A véletlen blokkelrendezés SPSS parancsai

```

UNIANOVA
  termés BY ismétlés kezelés
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = ismétlés kezelés.
    
```

45. táblázat. A véletlen blokkelrendezés eredménytáblázata, v=4, r=6

Tests of Between-Subjects Effects

Dependent Variable: cirok kg/parcella

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	124.943 ^a	8	15.618	7.416	.000
Intercept	5424.027	1	5424.027	2575.511	.000
ISMÉTLÉS	17.993	5	3.599	1.709	.193
KEZELÉS	106.950	3	35.650	16.928	.000
Error	31.590	15	2.106		
Total	5580.560	24			
Corrected Total	156.533	23			

a. R Squared = .798 (Adjusted R Squared = .691)

Latin négyzet elrendezés

Példa: Hat kezeléssel hat ismétléses kísérletben nitrogén műtrágyakezelések hatását vizsgálták és hasonlították össze az őszi búza szemtermésén.

1	2	3	4	5	6
6	1	2	3	4	5
5	6	1	2	3	4
4	5	6	1	2	3
3	4	5	6	1	2
2	3	4	5	6	1

36. ábra. Szisztematikus (diagonális elrendezés) 6x6 latin négyzet vázrajza

Egy-tényezős kísérletekben 4, 5, 6, 7 és 8 kezelés összehasonlítására alkalmas kísérleti elrendezés, feltételezve, hogy az ismétlések száma azonos a kezelések számával. Ebben az elrendezésben ugyanis a kezelések és az ismétlések számának meg kell egyezniük. Az elrendezés nagy előnye, ha ugyanabban a sorban vagy oszlopban több parcella is tönkremegy, akár egy sor vagy egy oszlop is kihagyható, és a kísérlet véletlen blokkelrendezésűnek tekinthető⁴.

A latin négyzet elrendezés legegyszerűbben a következőképpen szerkeszthető: a kezeléseket az első sorban 1-gyel kezdődő folyamatos számozással írjuk fel. A következő sorban ugyanebben a sorrendben, de egy parcellával jobbra eltolva kezdjük meg a felírást. Ezzel a módszerrel tehát minden egyes sorban eggyel jobbra tovább tolva, de ugyanabban a sorrendben írva töltjük ki a latin négyzetet. Ekkor kapjuk meg az ún. diagonális (átlós) latin négyzetet (36. ábra), amelyben minden sor és oszlop tartalmazza az összes kezelést. Aztán először a sorokat (46. táblázat), majd az oszlopokat (47. táblázat) véletlenszerűen felcseréljük. Az így szerkesztett latin négyzet már véletlen elrendezésű.

⁴ Ez az elrendezés abban különbözik a véletlen blokkelrendezéstől, hogy az összes kezelés egy-egy parcellájából két irányban képzünk blokkokat.

46. táblázat. Sorok felcserélése. 6x6 latin négyzet vázrajza

5	6	1	2	3	4
3	4	5	6	1	2
1	2	3	4	5	6
6	1	2	3	4	5
2	3	4	5	6	1
4	5	6	1	2	3

47. táblázat. Oszlopok felcserélése a sorok felcserélése táblázatból

1	3	6	2	5	4
5	1	4	6	3	2
3	5	2	4	1	6
2	4	1	3	6	5
4	6	3	5	2	1
6	2	5	1	4	3

Az elrendezés matematikai modellje:

$$Y_{ijk} = m + S_i + O_j + K_k + e_{ijk}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

S_i = blokk ill. ismétlés hatás a talaj heterogenitása, hogyan változik a talaj termékenysége fentről lefelé haladva

O_j = blokk ill. ismétlés hatás a talaj heterogenitása, hogyan változik a talaj termékenysége jobbról balra haladva

K_k = kezeléshatás

e_{ijk} = a kísérlet hibája

48. táblázat. A GLM-táblázat szerkezete véletlen blokk elrendezésben

Tényező	SS	df	MS	F	Sig.	DESIGN
Korrigált modell						
Eltérés		1				
Sor		r-1				sor
Oszlop		r-1				oszlop
Kezelés (csop. között)		v-1				kezelés
Hiba		(r-1)(v-2)				
Összesen		rv				
Korrigált összesen		rv-1				

49. táblázat. A latin négyzet elrendezés SPSS parancsai

```

UNIANOVA
  termés BY sor oszlop kezelés
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = sor oszlop kezelés .
    
```

50. táblázat. 5x5 latin négyzet eredménytáblázata

Tests of Between-Subjects Effects

Dependent Variable: TERMÉS

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4512.271 ^a	12	376.023	22.332	.000
Intercept	108860.4	1	108860.4	6465.188	.000
SOR	2326.386	4	581.597	34.541	.000
OSZLOP	901.374	4	225.344	13.383	.000
KEZELÉS	1284.510	4	321.128	19.072	.000
Error	202.055	12	16.838		
Total	113574.7	25			
Corrected Total	4714.326	24			

a. R Squared = .957 (Adjusted R Squared = .914)

Latin téglá elrendezés

A módszer a latin négyzet kiterjesztése 8, 9, 10, 12, 14, 15, 16, 18 kezelés összehasonlítására, feltételezve, hogy a kezelések száma kétszer, háromszor annyi, mint az ismétlések száma. Latin téglá elrendezésben ugyanis a kezelések száma az ismétlések számának egész számú többszöröse⁵.

A latin téglá elrendezés nagyon hasonlít a latin négyzet elrendezéshez: itt is sorokat és oszlopokat különböztetünk meg, a sorok és az oszlopok száma megegyezik egymással, illetve az ismétlések számával. Minden sorban és oszlopban az összes kezelés egy-egy parcellája szerepel. Mivel a kezelések száma az ismétlések számának két- vagy háromszorososa, bármely sor és oszlop kereszteződésében két vagy három kezelés parcellája van. Ez úgy lehet, hogy minden oszlop két vagy három részoszlopból áll⁶. Ettől eltekintve az elrendezés véletlenszerű.

Példa: 5 ismétléses 10 kezeléssel, 5x10-es latin téglá elrendezésű terv (37. ábra).

Sor	Oszlop									
	1		2		3		4		5	
1	8	5	9	7	2	4	10	3	1	6
2	9	7	8	1	3	10	4	6	5	2
3	6	3	5	10	1	7	9	2	8	4
4	4	10	6	2	8	9	1	5	7	3
5	1	2	3	4	5	6	7	8	9	10

37. ábra. 5x10-es latin téglá elrendezés, a 10 kezelés elhelyezése

Minden sor minden egymás melletti két részoszlop a kezelések egy teljes ismétlését tartalmazza. Két részoszlop együttesen képez egy oszlopot.

A latin téglá elrendezésben, a latin négyzet elrendezéssel teljesen egyezően, két alaptáblázatot készítünk. Az elsőben a termésadatokat az elrendezési terv alapján csoportosítjuk (51. táblázat).

⁵ Nem célszerű azonban, hogy a kezelések száma az ismétlések számának négyszerese, vagy annál is többszöröse legyen.

⁶ A latin négyzet módszernél minden kereszteződésben csak egy parcella van.

51. táblázat. Az adatok oszlopok és sorok szerinti elrendezése.

Sor	Oszlop									
	1		2		3		4		5	
1	0,9 6	1,1 7	1,1 2	1,0 3	1,3 8	1,5 0	2,2 2	2,0 4	2,2 4	1,2 9
2	1,7 0	1,9 0	1,4 8	1,4 1	1,9 7	2,3 5	2,4 0	2,0 5	1,9 5	1,7 0
3	2,1 2	1,7 3	1,9 7	1,6 8	1,8 6	1,6 9	1,8 0	1,7 5	1,9 2	1,5 3
4	1,9 4	2,4 2	1,7 5	1,9 3	1,7 9	1,6 4	1,9 3	1,5 9	1,8 2	1,4 5
5	1,0 0	1,6 9	2,0 6	1,5 9	1,9 3	1,7 0	1,9 2	1,7 6	1,7 6	1,1 1

A második alaptáblázatban a termésadatokat a kezelések és a sorok szerint csoportosítjuk (52. táblázat).

52. táblázat. Az adatok kezelések és sorok szerint rendezve.

Kezelés	Sor				
	1	2	3	4	5
1	2,24	1,41	1,86	1,93	1,00
2	1,38	1,70	1,75	1,93	1,69
3	2,04	1,97	1,73	1,45	2,06
4	1,50	2,40	1,53	1,94	1,59
5	1,17	1,95	1,97	1,59	1,93
6	1,29	2,05	2,12	1,75	1,70
7	1,03	1,90	1,69	1,82	1,92
8	0,96	1,48	1,92	1,79	1,76
9	1,12	1,70	1,80	1,64	1,76
10	2,22	2,35	1,68	2,42	1,11

Csoportosított elrendezés

Egy-tényezős kísérletek esetén, ha sok kezelést hasonlítunk össze gyakran olyan kezeléscsoportokat képezünk, amelyeken belül a kezelések összehasonlításának a pontosságára nagyobb súlyt helyezünk, mint a különböző csoportokban lévő kezelések összehasonlításának pontosságára. A kezelések közvetlen összehasonlításán túl vizsgálni akarjuk még a csoportátlagok közötti különbségeket is. Ilyen esetekben, ahelyett, hogy minden csoporttal külön elvégeznénk a kísérletet, a különböző csoportokat egy közös kísérletbe foglaljuk. A csoportonkénti kezelések száma különböző lehet.

Példa: Burgonya kísérletben 11 burgonyafajta három érési csoportba sorolható, egyenként 4 fajtaival. A kísérlet célja, hogy elsősorban összehasonlítsa az azonos érési csoportokon belüli fajták közötti terméskülönbséget. A különböző érési csoportokban lévő fajtákat ill. általában az érési csoportok átlagos termőképességének összehasonlítása csak másodlagos jelentőségű.

ismétlés												
1	III			I				II				
	10	9	11	1	3	4	2	8	6	5	7	
2	II				III				I			
	5	6	8	7	11	9	10	2	1	4	3	
3	I				II				III			
	4	1	3	2	6	7	5	8	10	11	9	
4	III				I				II			
	9	11	10	1	2	3	4	7	5	8	6	
5	II				III				I			
	6	8	7	5	10	9	11	2	3	4	1	

38. ábra. Csoportosított elrendezés terve, 11 kezeléssel, 3 csoportban, 5 ismétlésben

Az alaptáblázatban ismétlésenként és csoportonként, csoportokon belül kezelésenként rendezzük az adatokat.

53. táblázat. Az alaptáblázat

Érés csoportok	Kezelés	Ismétlés				
		(1)	(2)	(3)	(4)	(5)
I. középkorai	1	61	49	60	63	60
	2	42	33	66	64	51
	3	66	50	54	68	58
	4	54	50	53	60	49
II. közép	5	64	55	67	69	70
	6	47	39	41	37	52
	7	62	61	62	64	66
	8	72	56	62	60	71
III. Kései	9	61	62	74	60	80
	10	87	77	80	83	86
	11	82	73	85	83	74

Az elrendezés matematikai modellje:

$$Y_{ijk} = m + R_j + C_j + e_{ij} + K_k + e_{ijk}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_j = blokk ill. ismétlés hatás a talaj heterogenitása

C_j = az érécscsoportok termésre gyakorolt hatása

e_{ij} = az érécscsoportok közötti hiba

K_k = a fajták hatása a burgonya termésére

e_{ijk} = a kísérlet hibája

54. táblázat. A GLM-táblázat szerkezete csoportosított elrendezésben

Tényező	SS	df	M	F	Sig.	DESIGN
Eltérés		1	S			
Ismétlés		r-1				ismétlés
Csoportok között		cs-1				csoport
Hiba (cs)		(r-1)(cs-1)				ismétlés*csoport
Kezelés csp. belül		v-cs				kezelés
Hiba (v)		(r-1)(v-cs)				

55. táblázat. A csoportosított elrendezés SPSS parancsai

```

UNIANOVA
  termés BY ismétlés csoport kezelés
  /METHOD = SSTYPE(1)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /RANDOM = ismétlés
  /DESIGN = ismétlés csoport csoport*ismétlés kezelés .
    
```

Az eltérés négyzetösszeget az első típus szerint kell számítani.

56. táblázat. A csoportosított elrendezés eredménytáblázata, $cs=3$,
 $v=11$, $r=5$

Tests of Between-Subjects Effects

Dependent Variable: Termés kg/parcella

Source	Type I Sum of Squares	df	Mean Square	F	Sig.	
Intercept	Hypothesis	214531,364	1	214531,364	1096,328	,000
	Error	782,727	4	195,682(a)		
ismétlés	Hypothesis	782,727	4	195,682	3,708	,056
	Error	409,209	7,754	52,772(b)		
csoport	Hypothesis	4158,403	2	2079,202	39,761	,000
	Error	418,339	8	52,292(c)		
Hiba (cs)	Hypothesis	418,339	8	52,292	1,579	,170
	Error	1059,733	32	33,117(d)		
kezelés	Hypothesis	2520,433	8	315,054	9,513	,000
	Hiba (v)	1059,733	32	33,117(d)		

a MS(ismétlés)

b $1,025 \text{ MS(ismétlés * csoport) - } ,025 \text{ MS(Error)}$ c $\text{MS(ismétlés * csoport)}$ d MS(Error)

A csoportok közötti szignifikancia vizsgálatkor, ha a 'Hiba (cs) MS' kisebb, mint a 'Hiba (v) MS', akkor a csoportok közötti tényezőt az F-próbában a Hiba (v)-hez viszonyítjuk.

Két-tényezős kísérletek

Véletlen blokkelrendezés

A véletlen blokkelrendezés az egyik legegyszerűbb két-tényezős kísérleti elrendezés. Az egy-tényezős véletlen blokkelrendezéstől annyiban különbözik, hogy itt az egyes kezelések két tényező összes lehetséges kombinációi. Akkor alkalmazzuk, ha minden kombináció közötti különbséget azonos pontossággal akarunk elbírálni, és ha mindkét tényező változatai közötti különbségek elbírálására egyforma hangsúlyt fektetünk. Azonban ha az egyiket nagyobb pontossággal akarjuk elbírálni, akkor az osztott parcellás eljárást alkalmazzuk.

Példa: Három agrotechnikai eljárást hasonlítsunk össze, a tesztnövény burgonya legyen. Mivel feltételezhető, hogy a különféle burgonyafajták a vizsgált agrotechnikai eljárásokra különbözőképpen reagálnak, a kísérletet 2 burgonyafajtával állítjuk be (b_1, b_2) . 5 ismétléses véletlenszerű blokkelrendezésben. A kísérletben a következő kérdéseket lehet feltenni:

Melyik művelési móddal lehet a legnagyobb termést elérni a burgonyafajták átlagában?

Melyik burgonyafajta ad nagyobb termést a vizsgált művelési módok átlagában?

A művelési módok közti különbség változik-e burgonyafajták szerint, illetve

a burgonyafajták terméskülönbsége változik-e a művelési módok szerint?

57. táblázat. 5 ismétléses véletlen blokkelrendezésű, 2x3-as kísérlet termés adatai, a_1, a_2, a_3 művelési módok, b_1, b_2 .

ismétlés	kezelések					
1	$a_1 \cdot b_1$	$a_1 \cdot b_2$	$a_2 \cdot b_1$	$a_2 \cdot b_2$	$a_3 \cdot b_1$	$a_3 \cdot b_2$
2	$a_2 \cdot b_1$	$a_1 \cdot b_1$	$a_1 \cdot b_2$	$a_3 \cdot b_1$	$a_3 \cdot b_2$	$a_2 \cdot b_2$
3	$a_3 \cdot b_1$	$a_2 \cdot b_1$	$a_1 \cdot b_1$	$a_3 \cdot b_2$	$a_2 \cdot b_2$	$a_1 \cdot b_2$
4	$a_2 \cdot b_2$	$a_3 \cdot b_2$	$a_3 \cdot b_1$	$a_1 \cdot b_2$	$a_1 \cdot b_1$	$a_2 \cdot b_1$
5	$a_3 \cdot b_2$	$a_3 \cdot b_1$	$a_2 \cdot b_2$	$a_2 \cdot b_1$	$a_1 \cdot b_2$	$a_1 \cdot b_1$

Két-tényezős kísérleti elrendezést feltételezve az A tényező változatainak a számát jelöljük a -val, a B tényező változatainak a számát b -vel. A kezelések száma $a \cdot b$.

Az elrendezés matematikai modellje:

$$Y_{ijk} = m + R_i + A_j + B_k + AB_{jk} + e_{ijk}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_i = blokk ill. ismétlés hatás a talaj heterogenitását mutatja

A_j = az „A” tényező termésre gyakorolt hatása

B_k = a „B” tényező termésre gyakorolt hatása

AB_{jk} = a két tényező kölcsönhatása

e_{ijk} = a kísérlet hibája

58. táblázat. A GLM-táblázat szerkezete két-tényezős véletlen blokkelrendezésben

Tényező	SS	df	M	F	Si	DESIGN
			S		g.	
Korrigált modell		?				
Eltérés		1				
Ismétlés		r-1				ismétlés
A tényező		a-1				atényező
B tényező		b-1				btényező
AxB kölcsönhatás		(a-1)(b-1)				atényező*btényező
Hiba		(r-1)(ab-1)				
Összesen		rab				
Korrigált összesen		rab-1				

59. táblázat. A két-tényezős véletlen blokkelrendezés SPSS parancsai

UNIANOVA

termés BY ismétlés atényező btényező

/METHOD = SSTYPE(3)

/INTERCEPT = INCLUDE

/CRITERIA = ALPHA(.05)

/DESIGN = ismétlés atényező btényező atényező*btényező

60. táblázat. A két-tényezős véletlen blokkelrendezés eredménytáblázata, a=3, b=2, r=5

Tests of Between-Subjects Effects

Dependent Variable: Termés kg/parcella

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	62695 ^a	9	6966.130	18.867	.000
Intercept	1393207	1	1393207	3773.414	.000
ISMÉTLÉS	2720	4	679.917	1.842	.160
A TÉNYEZŐ	57965	2	28982.500	78.497	.000
B TÉNYEZŐ	3	1	2.700	.007	.933
A x B	2008	2	1003.900	2.719	.090
Error (Hiba)	7384	20	369.217		
Total	1463287	30			
Corrected Total	70080	29			

a. R Squared = .895 (Adjusted R Squared = .847)

Osztott parcellás (split-plot) elrendezés

Osztott parcellás elrendezés alkalmazásának a feltételei:

A kísérlet eredeti céljának megfelelően egy-tényezős (B). A kezelések közötti különbségeket azonban egy másik tényező (A) különböző változataival kombinálva akarjuk vizsgálni;

Az egyik vizsgált tényező (A) parcellánkénti változtatása technikai nehézségbe ütközik;

Mindkét vizsgált tényező változatai közötti különbségek, és ezek kölcsönhatása érdekel;

Az egyik vizsgált tényező (A) változatai közötti különbségek elbírálása nem elsődleges cél. A kísérlet kérdése elsősorban a másik tényező (B) változatainak értékelésére és az A×B kölcsönhatásra irányul.

Példa: 3 kukoricafajtát 4 időpontban vetve hasonlítunk össze 5 ismétléses kísérletben. Mérjük a vetéstől a hímvirágzásig eltelt napok számát. A kísérlet elsődleges célja, hogy a fajták vegetatív tenyészideje között milyen különbség van. Kérdés lehet még, hogy a vetéstől a hímvirágzásig eltelt idő hossza a fajták átlagában hogyan változik a vetésidőpont szerint? A vetésidő a kevésbé lényeges, *A* tényező, a fajta a fontosabb, *B* tényező.

ismétlés	kezelések					
1	$a_1 \cdot b_1$	$a_1 \cdot b_2$	$a_2 \cdot b_1$	$a_2 \cdot b_2$	$a_3 \cdot b_1$	$a_3 \cdot b_2$
2	$a_2 \cdot b_1$	$a_1 \cdot b_1$	$a_1 \cdot b_2$	$a_3 \cdot b_1$	$a_3 \cdot b_2$	$a_2 \cdot b_2$
3	$a_3 \cdot b_1$	$a_2 \cdot b_1$	$a_1 \cdot b_1$	$a_3 \cdot b_2$	$a_2 \cdot b_2$	$a_1 \cdot b_2$
4	$a_2 \cdot b_2$	$a_3 \cdot b_2$	$a_3 \cdot b_1$	$a_1 \cdot b_2$	$a_1 \cdot b_1$	$a_2 \cdot b_1$
5	$a_3 \cdot b_2$	$a_3 \cdot b_1$	$a_2 \cdot b_2$	$a_2 \cdot b_1$	$a_1 \cdot b_2$	$a_1 \cdot b_1$

39. ábra. 3x2 kéttényezős kísérlet elrendezési terve osztott parcellás elrendezésben, parcellánkénti adatokkal

A példában a vetésidő az **A** tényező négy változata az öt ismétléses kísérletben véletlen blokkelrendezésben van. A fajta a **B** tényező, ennek 3 változatát vizsgáljuk.

Az elrendezés matematikai modellje:

$$Y_{ijk} = m + R_j + A_j + e_{ij} + B_k + AB_{jk} + e_{ijk}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_j = blokk ill. ismétlés hatás a talaj heterogenitását mutatja

A_j = az „A” tényező termésre gyakorolt hatása

e_{ij} = az „A” tényező hibája

B_k = a „B” tényező termésre gyakorolt hatása

AB_{jk} = a két tényező kölcsönhatása

e_{ijk} = a „B” tényező hibája

61. táblázat. A GLM-táblázat szerkezete két-tényezős osztott parcellás elrendezésben

Tényező	SS	df	M	F	Sig	DESIGN
Eltérés		1	S		.	
Ismétlés		r-1				ismétlés
A tényező		a-1				atényező
Hiba (a)		(r-1)(a-1)				atényező*ismétl és
B tényező		b-1				btényező
AxB kölcsonhat ás		(a-1)(b-1)				atényező*btény ező
Hiba (b)		a(r-1)(b-1)				

62. táblázat. A két-tényezős osztott parcellás elrendezés SPSS parancsai

```

UNIANOVA
  napok BY ismétlés atényező btényező
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /RANDOM = ismétlés
  /DESIGN = ismétlés atényező atényező*ismétlés btényező
  atényező*btényező .
    
```

63. táblázat. A két-tényezős osztott parcellás elrendezés eredménytáblázata, a=4, b=3, r=5

Tests of Between-Subjects Effects

Dependent Variable: Napok száma vetéstől hímvirágzásig

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	675220,417	1	675220,417	58292,410	,000
Hypothesis Error	46,333	4	11,583(a)		
ismétlés Hypothesis	46,333	4	11,583	1,590	,240

	esis					
	Error	87,400	12	7,283(b)		
A tényező	Hypoth esis	1704,183	3	568,061	77,995	,000
	Error	87,400	12	7,283(b)		
Hiba (a)	Hypoth esis	87,400	12	7,283	,495	,902
	Error	470,667	32	14,708(c)		
B tényező	Hypoth esis	9168,233	2	4584,117	311,668	,000
	Error	470,667	32	14,708(c)		
A*B kölcsön hatás	Hypoth esis	19,767	6	3,294	,224	,966
	Hiba (b)	470,667	32	14,708(c)		

a MS(ismétlés)

b MS(ismétlés * atényező)

c MS(Error)

Az „A” tényező közötti szignifikancia vizsgálatkor az „A” tényező MS-t akkor kell osztani a Hiba (a) MS-vel, ha ez az érték nagyobb, mint a Hiba (b) MS. Egyéb esetben a Hiba (b)-hez kell viszonyítani az „A” tényező hatását.

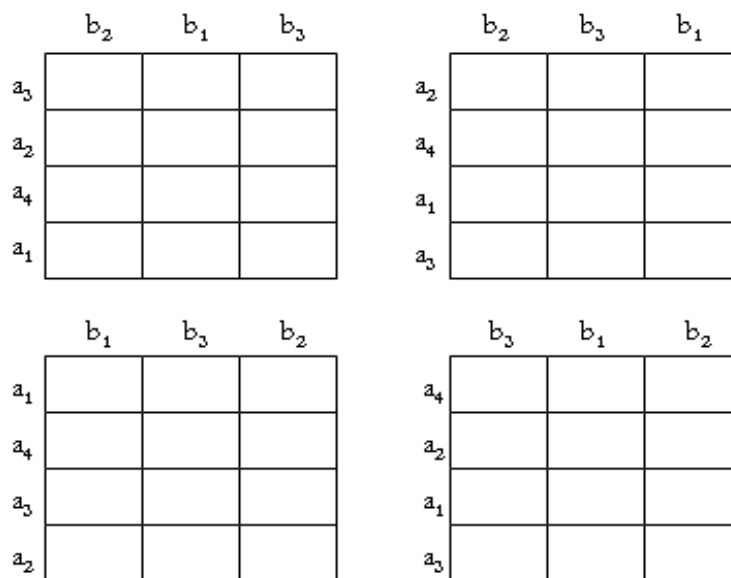
Sávós elrendezés

A két-tényezős kísérletek kevésbé javasolható elrendezése. Mégis gyakran ez az egyetlen megoldás, főként szántóföldi kísérletekben, ha a parcellaméret olyan kicsi, hogy azon technikai nehézségek miatt a tényezők egyikének vizsgálata sem kivitelezhető. Ilyenkor a pontosság rovására a belső ismétlések feláldozásával mindkét tényezőt főparcellákon helyezünk el. Előnye viszont, hogy a kölcsönhatást pontosabban lehet becsülni.

Példa: 4 vetésidőpont és 3 talaj-előkészítés hatását vizsgálják cukorrépán. Az összes kombináció száma 12. A kísérlet négy ismétléses, az összes

parcellaszám ezért 48. A rendelkezésre álló terület miatt egy parcella mérete csak 10 m² lehet. A talajművelés és vetés kombinációinak elhelyezése gyakorlatilag kivitelezhetetlen a gépek fordulása és helyigénye miatt. Mindkét kezeléshez nagyobb területre van szükség, mint 10 m². Így csak a sávos elrendezés nyújthat segítséget.

Az egész kísérleti teret annyi egyforma nagyságú részre, blokkra bontjuk, ahány ismétléses kísérletet tervezünk. Ezt követően ismétlésenként elhelyezzük az *A* tényező minden változatát, mintha nem is lenne *B* tényező, majd ezekre keresztbe helyezük el a *B* tényező minden változatát, mintha nem lenne *A* tényező. A két kezelés szintjeinek elhelyezését minden ismétlésben **újra randomizáljuk**.



40. ábra. Két-tényezős kísérlet sávos elrendezésben.

Az elrendezés matematikai modellje:

$$Y_{ijk} = m + R_j + A_j + e_{ij} + B_k + e_{ik} + AB_{jk} + e_{ijk}$$

ahol:

Y_{ijk} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_j = blokk ill. ismétlés hatás a talaj heterogenitását mutatja

A_j = az „A” tényező termésre gyakorolt hatása

e_{ij} = az „A” tényező hibája

B_k = a „B” tényező termésre gyakorolt hatása

e_{ik} = a „B” tényező hibája

AB_{jk} = a két tényező kölcsönhatása

e_{ijk} = a „B” tényező hibája

64. táblázat. A GLM-táblázat szerkezete két-tényezős sávos elrendezésben

Tényező	SS	df	M	F	Si	DESIGN
			S		g.	
Eltérés		1				
Ismétlés		r-1				ismétlés
A tényező		a-1				atényező
Hiba (a)		(r-1)(a-1)				atényező*ismétl és
B tényező		b-1				btényező
Hiba (b)		(r-1)(b-1)				btényező*ismétl és
AxB kölsönhatá s		(a-1)(b-1)				atényező*btény ező
Hiba (a x b)		(r-1)(a-1) (b-1)				

65. táblázat. A két-tényezős sávos elrendezés SPSS parancsai

```

UNIANOVA
  termés BY ismétlés atényező btényező
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /RANDOM = ismétlés
  /DESIGN = ismétlés atényező atényező*ismétlés btényező
  btényező*ismétlés atényező*btényező
  
```

66. táblázat. A kéttényezős sávos elrendezés eredménytáblázata, $a=4, b=3, r=4$

Dependent Variable: c.répa termés kg/10 m²

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	34133,333	1	34133,333	967,559	,000
Error	105,833	3	35,278(a)		
Ismétlés	105,833	3	35,278	2,988	,134
Error	59,652	5,053	11,806(b)		
A tényező	133,667	3	44,556	4,367	,037
Error	91,833	9	10,204(c)		
Hiba (a)	91,833	9	10,204	1,406	,257
Error	130,667	18	7,259(d)		
B tényező	1113,167	2	556,583	62,812	,000
Error	53,167	6	8,861(e)		
Hiba (b)	53,167	6	8,861	1,221	,341
Error	130,667	18	7,259(d)		
AxB kölcsönhatás	86,333	6	14,389	1,982	,122
Hiba (axb)	130,667	18	7,259(d)		

a MS(ismétlés)

b $MS(\text{ismétlés} * \text{atényező}) + MS(\text{ismétlés} * \text{btényező}) - MS(\text{Error})$

c $MS(\text{ismétlés} * \text{atényező})$

d $MS(\text{Error})$

e $MS(\text{ismétlés} * \text{btényező})$

Három- és több-tényezős kísérletek

Három vagy ennél több tényezős kísérleti elrendezések közül gyakorlatilag könnyen tervezhető és értékelhető a véletlen blokk, az osztott parcellás és a sávós elrendezés kombinációi.

Három-tényezős kísérletekben ebből a három alaptípusból a következő kombinációkat képezhetjük:

A három tényező változatainak minden kombinációját az ismétlésen belül véletlenszerűen rendezzük el (véletlen blokkelrendezés).

Az **A** és **B** tényezők változatainak kombinációit ismétlésen belül 1.) véletlenszerűen, 2.) osztott parcellásan, 3.) sávosan rendezzük el. Az a x b kombinációjú parcellákat osztjuk fel a **C** tényező szerint alparcellákra.

Az **A** tényező változatait ismétlésen belül véletlenszerűen rendezzük el. Az **A** tényező változatai tehát főparcellákat képeznek. Ezeket a főparcellákat osztjuk fel **B** és **C** tényezők változatainak kombinációira: 1.) véletlenszerű, 2.) osztott parcellás, 3.) sávós elrendezésben.

Véletlen blokkelrendezés

Három tényező vizsgálatakor ez az elrendezés főként laboratóriumi vagy tenyészedény kísérletekben előnyös, mivel minden kombináció azonos pontossággal hasonlítható össze. Szántóföldi kísérletekben ritkán alkalmazzák, mivel a sok kombinációhoz nagy blokkokat kell képezni, és egyes kezelések beállítása technikailag szinte lehetetlen, pl. talajművelés, öntözés, stb.

Az alábbi példa egy tőszám (2), hibrid (3) és műtrágyázási (3) kísérlet kiértékelését mutatja be.

Az elrendezés matematikai modellje:

$$Y_{ijkl} = m + R_j + A_j + B_k + C_l + AB_{jk} + AC_{jl} + BC_{kl} + ABC_{jkl} + e_{ijkl}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_j = blokk ill. ismétlés hatás a talaj heterogenitását mutatja

A_j = az „A” tényező termésre gyakorolt hatása

B_k = a „B” tényező termésre gyakorolt hatása

C_I = a „C” tényező termésre gyakorolt hatása
 AB_{jk} = a két tényező kölcsönhatása
 AC_{ji} = a két tényező kölcsönhatása
 BC_{ki} = a két tényező kölcsönhatása
 ABC_{jkl} = a három tényező kölcsönhatása
 e_{ijkl} = hiba

67. táblázat. A GLM-táblázat szerkezete háromtényezős véletlen blokkelrendezésben

Tényező	SS	df	M	F	Sig	DESIGN
			S		.	
Korrigált modell						
Eltérés		1				
Ismétlés		r-1				ismétlés
A tényező		a-1				toszam
B tényező		b-1				hibrid
C tényező		c-1				tragya
AxB kölcsönhatás		(a-1)(b-1)				hibrid*toszam
AxC kölcsönhatás		(a-1)(c-1)				toszam*tragya
BxC kölcsönhatás		(b-1)(c-1)				hibrid*tragya
AxBxC		(a-1)(b-1)(c-1)				hibrid*toszam*tragya
Hiba		(r-1)(abc-1)				
Összesen		rabc				
Korrigált összesen		rabc-1				

68. táblázat. A három-tényezős véletlen blokkelrendezés SPSS parancsai

```
UNIANOVA
  termes BY ismetles toszam hibrid tragya
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = ismetles toszam hibrid tragya hibrid*toszam
  toszam*tragya hibrid*tragya hibrid*toszam*tragya .
```

69. táblázat. A három-tényezős véletlen blokkelrendezés eredménytáblázata, a=2, b=3, c=3, r=4

Tests of Between-Subjects Effects

Dependent Variable: termés t/ha

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	204.634 ^a	20	10.232	14.208	.000
Intercept	6769.019	1	6769.019	9399.807	.000
ISMETLES	1.098	3	.366	.508	.678
TOSZAM	16.872	1	16.872	23.430	.000
HIBRID	23.914	2	11.957	16.604	.000
TRAGYA	147.237	2	73.618	102.230	.000
TOSZAM * HIBRID	3.873	2	1.936	2.689	.078
TOSZAM * TRAGYA	8.104	2	4.052	5.627	.006
HIBRID * TRAGYA	1.438	4	.360	.499	.736
TOSZAM * HIBRID * TRAGYA	2.098	4	.525	.728	.577
Error	36.726	51	.720		
Total	7010.379	72			
Corrected Total	241.360	71			

a. R Squared = .848 (Adjusted R Squared = .788)

Kétszeresen osztott parcellás (split-split-plot) elrendezés

Ez az elrendezés technikailag igen előnyös három-tényezős elrendezés, főként szántóföldi kísérletekben, mert a főparcellák, az elsőrendű és a másodrendű alparcellák eltérő méretei különböző tulajdonságú kezelések kombinációját teszik lehetővé. A kevésbé fontos tényezőt (kezelést) a főparcellán (A) helyezük el, a legfontosabbat a másodrendű alparcellákon (C). Az „A” tényező változatainak ismétlése megegyezik a valódi ismétlés számával (r). A „B” tényező ismétlés száma ra, amiből a a belső ismétlés. A „C” tényező ismétlése rab, amiből ab belső ismétlés. Amennyiben a

kölcsönhatások nem szignifikánsan, a belső ismétlések is valódi ismétlést jelentenek.

Egy debreceni kísérletben a főparcellán a tőszámot (A), az elsőrendű alparcellán a hibridet (B) és a másodrendű alparcellán a műtrágyakezeléseket (C) helyezték el négy ismétlésben (r).

		(1) ismétlés				(2) ismétlés			
Fő parcella		A1		A2		A2		A1	
Al parcella		B1	B2	B2	B1	B2	B1	B1	B2
	c1	c4	c3	c2	c2	c3	c4	c1	
	c2	c2	c4	c1	c4	c1	c2	c2	
	c3	c3	c1	c4	c3	c4	c3	c3	
	c4	c1	c2	c3	c3	c2	c1	c4	
Osztó területek					Osztó területek				

41. ábra. Három-tényezős kétszeresen osztott parcellás elrendezés terve

Az elrendezés matematikai modellje:

$$Y_{ijkl} = m + R_i + A_j + e_{ij} + B_k + AB_{jk} + e_{ijk} + C_l + AC_{jl} + BC_{kl} + ABC_{jkl} + e_{ijkl}$$

ahol:

Y_{ij} = egy parcella termése (kg/parcella)

m = a kísérlet becsült, számított átlaga, a kísérlet legjellemzőbb értéke

R_i = blokk ill. ismétlés hatás a talaj heterogenitását mutatja

A_j = az „A” tényező termésre gyakorolt hatása

e_{ij} = az „A” tényező hibája

B_k = a „B” tényező termésre gyakorolt hatása

AB_{jk} = a két tényező kölcsönhatása

e_{ijk} = a „B” tényező hibája

C_l = a „C” tényező termésre gyakorolt hatása

AC_{jl} = a két tényező kölcsönhatása

BC_{kl} = a két tényező kölcsönhatása

ABC_{jkl} = a három tényező kölcsönhatása

e_{ijkl} = hiba

70. táblázat. A GLM-táblázat szerkezete három-tényezős kétszeresen osztott parcellás elrendezésben

Tényező	SS	df	M	F	Si	DESIGN
			S		g.	
Eltérés		1				
Ismétlés		r-1				ismétlés
A tényező		a-1				toszam
Hiba (a)		(r-1)(a-1)				ismetlés*toszam
B tényező		b-1				hibrid
AxB kölcsonhatás		(a-1)(b-1)				hibrid*toszam
Hiba (b)		a(r-1)(b-1)				toszam(hibrid*ismetles)
C tényező		c-1				tragya
AxC kölcsonhatás		(a-1)(c-1)				toszam*tragya
BxC kölcsonhatás		(b-1)(c-1)				hibrid*tragya
AxBxC		(a-1)(b-1)(c-1)				hibrid*toszam*tragya
Hiba (c)		ab(r-1)(c-1)				

71. táblázat. Három-tényezős kétszeresen osztott parcellás elrendezés SPSS parancsai

```

UNIANOVA
  termes BY ismetles toszam hibrid tragya
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /RANDOM = ismetles
  /DESIGN = ismetles toszam ismetles*toszam hibrid
  hibrid*toszam toszam(hibrid*ismetles) tragya toszam*tragya
  hibrid*tragya hibrid*toszam*tragya .
    
```

72. táblázat. Három-tényezős kétszeresen osztott parcellás elrendezés eredménytáblázata, a=2, b=3, c=3, r=4

Dependent Variable: termés

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	6769,019	1	6769,019	18498,464	,000
Error	1,098	3	,366(a)		
Ismetles	1,098	3	,366	,336	,803
Error	3,266	3	1,089(b)		
Toszam	16,872	1	16,872	15,498	,029
Error	3,266	3	1,089(b)		
Hiba (a)	3,266	3	1,089	1,865	,189
Error	7,005	12	,584(c)		
Hibrid	23,914	2	11,957	20,482	,000
Error	7,005	12	,584(c)		
Toszam * Hibrid	3,873	2	1,936	3,317	,071
Error	7,005	12	,584(c)		
Hiba (b)	7,005	12	,584	,794	,653
Error	26,455	36	,735(d)		
Tragya	147,237	2	73,618	100,181	,000
Error	26,455	36	,735(d)		
Toszam * Tragya	8,104	2	4,052	5,514	,008
Error	26,455	36	,735(d)		
Hibrid * Tragya	1,438	4	,360	,489	,743
Error	26,455	36	,735(d)		
Toszam * Hibrid * Tragya	2,098	4	,525	,714	,588
Hiba (c)	26,455	36	,735(d)		

a MS(ismetles)

b MS(ismetles * toszam)

c MS(toszam(ismetles * hibrid))

d MS(Error)

Kovariánsok alkalmazása a lineáris modellben

A variancia-analízis során tekintettel kell lenni arra is, hogy a vizsgálni kívánt függőváltozót vagy változókat a számításba bevont, ill. kísérletbe állított tényezőknél túl egyéb más tényezők is befolyásolják. Keresni kell egy olyan változót, amelyet folyamatosan kontrolálunk, mérünk, és valószínűleg lineáris kapcsolatban van a függő változóval. Ezt a változót nevezzük kovariánsnak. A kovariáns(ok) bevonásakor az analízis során úgy hajtunk végre egyszerre variancia- és regresszió analízist, hogy a lineáris regresszióval korrigált, módosított függő változó varianciáját bontjuk fel kezeléshatásokra. A kovariánsnak folytonos, skála típusú adatnak kell lennie.

Alkalmazási feltétel:

A kovariáns lineáris kapcsolatban legyen a függőváltozóval

A kovariáns értéke nem függhet az alkalmazott kezelésektől, tényezőktől

Az első feltétel magától értetődő. Amennyiben a kovariáns és a függőváltozó között a kapcsolat nem lineáris, a regresszióval módosított adatok torz értéket fognak felvenni.

A második feltétel teljesülése gyakorlatilag azt eredményezi, hogy a kezelések minden egyes parcelláján, celláján, stb. a regressziós koefficiens értéke megegyezik, azaz egyetlen regressziós egyenessel írható le az összefüggés.

Egy két-tényezős lineáris modell kovariánssal kiegészített matematikai modellje:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + \beta x_{ijk} + \varepsilon_{ijk}$$

Ahol:

Y_{ijk} : a függőváltozó értéke

μ : a kísérlet főátlaga (fix hatás)

A_i : A tényező hatása

B_j : B tényező hatása

AB_{ij} : a két tényező kölcsönhatása

β : a függőváltozó és a kovariáns közötti lineáris regressziós együttható

x_{ijk} : a kovariáns értékei

ε_{ijk} : hiba, eltérés, a véletlen hatása

Egy modellbe több kovariáns is bevonható, ha teljesítik a fenti alkalmazási feltételeket.

Példa: Gyümölcsfákat kezeltek virágrügyet indukáló szerrel. Megmérték 50 kezeletlen és 50 kezelt vesszőn a virágrügyek számát. Az adatokat variancia-analízissel értékelték. A kezelés hatását az 73. táblázat mutatja.

73. táblázat. Virágrügyek száma a kezelés hatására

Estimates

Dependent Variable: virágrügyek száma (db)

permetezés	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
nem kezelt	1.900	.183	1.536	2.264
kezelt	1.400	.183	1.036	1.764

Jól látható, hogy a kezeletlen vesszőkön több virágrügy van, mint a kezeltlen. Vajon szignifikánsan több, vagy csak a véletlen ingadozásnak tudható be a különbség? Válasszuk a szignifikancia szintet 5%-ra, és végezzük el a variancia-analízist! Az eredményt az 74. táblázat mutatja.

74. táblázat. A permetezés hatása a virágrügyek számára

Tests of Between-Subjects Effects

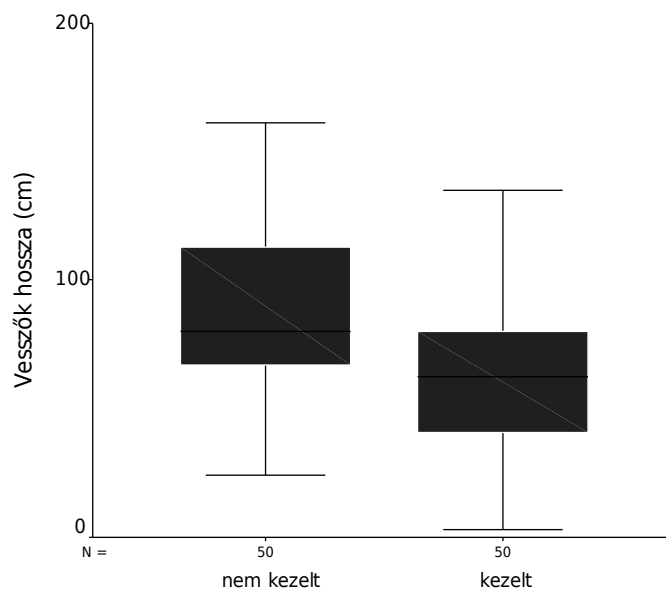
Dependent Variable: virágrügyek száma (db)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6.250 ^a	1	6.250	3.723	.057
Intercept	272.250	1	272.250	162.191	.000
KEZELÉS	6.250	1	6.250	3.723	.057
Error	164.500	98	1.679		
Total	443.000	100			
Corrected Total	170.750	99			

a. R Squared = .037 (Adjusted R Squared = .027)

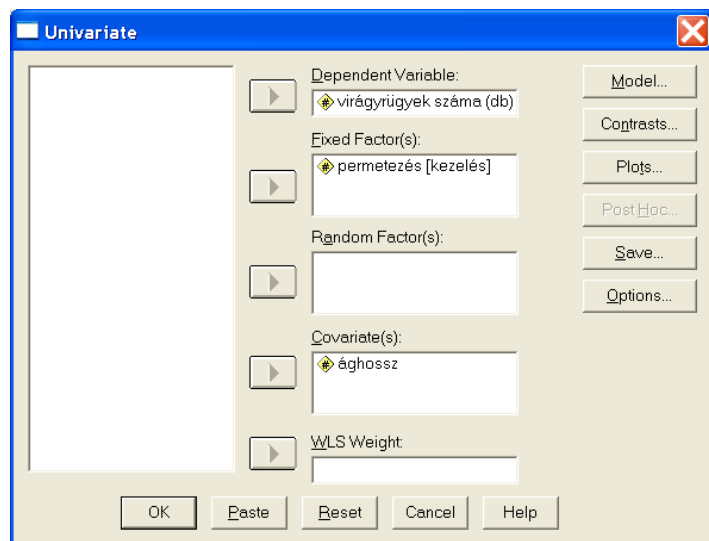
A „KEZELÉS” sort elemezve, elmondható, hogy 5%-os szignifikancia szint mellett nincs különbség a kezelt és kezeletlen vesszők virágszáma között. Amennyiben előzetesen az elsőfajú hiba valószínűségét 10%-ban állapítottuk volna meg, akkor ki kellene jelenteni, hogy a nem kezelt vesszőkön szignifikánsan több virágrügy található. Ez teljesen ellentmond a korábbi szakirodalmi megállapításoknak. Vajon mi lehet ennek az oka? Vizsgálódjunk tovább, nézzük meg az r-négyzet értékét! Ez nagyon alacsony 0,037, és a variancia-analízis táblázat „Corrected Model” sora sem igazolja a lineáris modell helyességét ($p > 0,05$). Valószínűleg egyéb tényező befolyásolja a virágrügyek számát, ami elfedi a kezelés hatását.

A virágrügyek a vesszők nóduszain fejlődnek. Két nódusz közötti távolság – ugyanazon faj esetében – eléggé konstans. Hosszú vesszőn potenciálisan több, rövidebb vesszőn potenciálisan kevesebb virágrügy indukálódhat. Vizsgáljuk meg a kezelt és kezeletlen vesszők hosszát (42. ábra).



42. ábra. Gyümölcsfa vesszőhossza a különböző kezelésekben

Ezek szerint a nem kezelt vesszők valamilyen szisztematikus hiba miatt eredetileg hosszabbak voltak, mint a kezelt vesszők. Mi okozhatja ezt? Valószínűleg két személy végezte a fák felvételezését, az egyik a nem kezelt



43. ábra. GLM modell kovariánssal kiegészítve

belátható, hogy a vesszők hossza nem függ a kezeléstől, permetezéstől.

vesszőket mérte és számolta meg rajta a virágrügyeket, a másik ugyanezt tette a kezelt vesszőkkel. Az első személynek azonban csak a hosszabb vesszők voltak a szimpatikusak, szisztematikusán válogatott a vesszők között, nem adta meg az esélyt, hogy bármilyen hosszúságú belekerüljön a mintába. Korábbi ismereteink alapján nyugodtan feltételezhetjük, hogy a vesszőhossz és virágrügyek száma lineáris kapcsolatban van egymással. Könnyen

A kovariancia analízis futtatásához térjünk vissza az ANALYZE/ GENERAL LINEAR MODEL UNIVARIATE parancsához, ahol a COVARIATE(S) ablakba helyezzük az ághossz változót (43. ábra). Az OK gombra kattintva futtassuk a programot.

75. táblázat. A GLM eredménytáblázata

Tests of Between-Subjects Effects

Dependent Variable: virágrügyek száma (db)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	151.532 ^a	2	75.766	382.412	.000
Intercept	22.848	1	22.848	115.322	.000
ÁGHOSSZ	145.282	1	145.282	733.279	.000
KEZELÉS	2.061	1	2.061	10.404	.002
Error	19.218	97	.198		
Total	443.000	100			
Corrected Total	170.750	99			

a. R Squared = .887 (Adjusted R Squared = .885)

A „Corrected Model” sor alapján megállapítható, hogy a lineáris modell helyes, az R-négyzet értéke magas, 0,887. Az „ághossz” változó, ami a vesszők hosszát jelenti, lineáris kapcsolatban van a virágrügyek számával. A kezelés szintén szignifikáns, a kezeletlen és kezelt csoportban a virágrügyek száma jelentősen eltér. Kovariáns alkalmazásakor korrigálni kell a mért csoport átlagokat, mintha minden csoportban a vesszők hossza megegyezett volna. Ezt egy átlagos vessző-hosszat figyelembe véve kell megtenni. (Gyakorlatilag ezzel „hendi kepeltük” a virágszámot.)

A 76,19cm-es vesszőhosszra korrigált virágszámokat az 76. táblázat mutatja. Itt már a szakirodalommal megegyező értékeket látunk, a kezelt vesszőkön 20%-kal több virág fejlődött.

76. táblázat. Korrigált virágrügy számok

Estimates

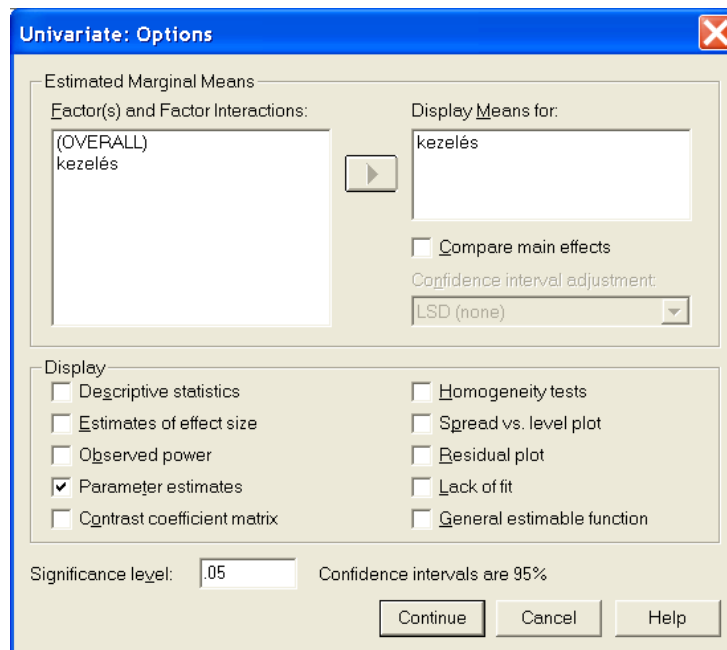
Dependent Variable: virágrügyek száma (db)

permetezés	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
nem kezelt	1.499 ^a	.065	1.370	1.627
kezelt	1.801 ^a	.065	1.673	1.930

a. Evaluated at covariates appeared in the model: ÁGHOSSZ = 76.19.

Mivel csak két csoportunk van nem kell további tesztek elvégzését a középértékek különbségének szignifikancia vizsgálatára.

Azonban kíváncsiak lehetünk a lineáris modell paramétereinek értékeire. Ehhez 43. ábra Options... parancsára kattintva ikszeljük be a Parameter estimates négyzetet (44. ábra). A paraméterek értékeit a 77. táblázat B oszlopából olvashatjuk le.



44. ábra. A paraméterek értékeinek meghatározása

77. táblázat. A lineáris modell paraméterei

Parameter Estimates

Dependent Variable: virágrügyek száma (db)

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-1.091	.111	-9.791	.000	-1.313	-.870
ÁGHOSSZ	3.80E-02	.001	27.079	.000	3.519E-02	4.075E-02
[KEZELÉS=0]	-.303	.094	-3.226	.002	-.489	-.116
[KEZELÉS=1]	0 ^a

a. This parameter is set to zero because it is redundant.

A fenti példa is jól igazolja, hogy a kísérlet kivitelezése, az adat felvételezés körülményei, a randomizáció hiánya döntő mértékben befolyásolhatja egy kísérletből levont következtetések helyességét.

KORRELÁCIÓ- ÉS REGRESSZIÓSZÁMÍTÁS

Egy statisztikai vizsgálatnak bármilyen legyen is a célja, megállapításokat, következtetéseket csak akkor tudunk levonni, ha a mért változók kölcsönhatásban vannak egymással és kapcsolataikat nemcsak a véletlen hatások befolyásolják, hanem szignifikáns összefüggések is kimutathatók.

A módszertani tanulmányok egyik leggyakrabban alkalmazott módszere a *korreláció- és regressziószámítás*, amelyek a változók közötti kapcsolatok elemzésének eszközei. A korrelációszámítás a magas mérési szintű változók kapcsolatainak vizsgálatával foglalkozik, elemzi a változók közötti kapcsolat meglétét, szorosságát és annak irányát. A regressziószámítás a változók közötti kapcsolat megléte esetén annak jellegét, minőségi jellemzőit vizsgálja.

KÉT-VÁLTOZÓS SZTOCHASZTIKUS KAPCSOLATOK

Egy gazdaságban termesztett haszonnövény pl. kukorica termését számos tényező befolyásolja: talajtípus, tőszám, a műtrágyázás jellege, kézi- és gépi munka ráfordítás, termőképesség stb. Ezek között a változók között különböző kapcsolatok alakulhatnak ki. Két változó között háromféle kapcsolat jöhet létre:

A két változó független egymástól, ha az egyik változó semmilyen információt nem szolgáltat a másik változóról.

Ha az egyik mutató hat a másik mutató alakulására, de a hatás véletlenszerű (következtetés szintű és csak közelítőleg becsülhető), akkor *sztochasztikus* a kapcsolat két mutató között.

Függvényszerű kapcsolatról akkor beszélünk, ha az egyik mutató változása egyértelműen befolyásolja a másik mutató megváltozását.

A mezőgazdasági termelési folyamatokban elsősorban sztochasztikus kapcsolattal találkozhatunk, ezeknek van kitüntetett szerepük. A sztochasztikus kapcsolatokat az alábbiak szerint csoportosíthatjuk.

Két minőségi ismérv közötti kapcsolatot az *asszociáció* fejez ki.

A *rangkorreláció* a sorba rendezett tényezők közötti kapcsolat elemzésének eszköze.

Vegyes kapcsolatról beszélünk, ha egy minőségi és egy mennyiségi változó közötti kapcsolatot elemzünk.

Két vagy több magas mérési szintű ismérv együttes vizsgálatakor (ha nem függetlenek egymástól) két kérdés merül fel: (1) milyen a két változó közötti kapcsolat erőssége és iránya; (2) hogyan lehet következtetni az egyik változó

értékeiből a másik változó értékeire. Az első kérdésre a *korrelációs számítás*, a másodikra a *regresszió-számítás* adja meg a választ.

A következőkben részletesen ismertetjük a fenti módszereket. A vizsgálat során el kell majd döntenünk, hogy melyik módszert kell alkalmazni; a döntést aszerint kell meghozni, hogy az adataink milyen mérési szintűek (nominális, ordinális, arány és intervallum).

Bármilyen vizsgálat megkezdése előtt azonban fontos annak átgondolása, hogy van-e valamilyen valóságos alapja a két változó közötti kapcsolatnak.

Asszociáció

Két minőségi mutató közötti kapcsolat szorosságát az asszociáció mutatószámaival mérjük. Ezek a mutatók a változók közötti kapcsolat szorosságát egy számban fejezik ki.

Mezőgazdasági vállalatoknál, vállalkozásoknál akarjuk azt megvizsgálni, hogy eltér-e a különböző végzettségű vezetők száma a veszteséges, a közepes nyereségű és a nagy nyereségű vállalkozásokban. Kérdőíves felmérésben 1500 vezetőt kérdeztek meg, az adatokat a 78. táblázat tartalmazza (*asszociacio.sav*).

78. táblázat. A vezetők megoszlása a különböző nyereségű mezőgazdasági vállalkozásokban

A vállalat	Alsó- fokú	Közép- végzettségű	Felső- vezetők	Összesen
Veszteséges	280	145	45	470
Közepes nyereségű	260	180	60	500
Nagy nyereségű	180	230	120	530
Összesen	720	555	225	1500

Forrás: VINCZE SZ. (2005): A korreláció- és regressziószámítás módszertani alapjai a területi statisztikai elemzésekben.

Amikor megkezdjük két minőségi ismérv kapcsolatának vizsgálatát és a kapcsolat erősségének meghatározását, elsőként az adatokat keresztáblába (kombinációs táblába) rendezzük. A keresztáblában az adatokat két (vagy

több) szempont / két (vagy több) változó szerint rendezve látjuk. A példánkban az egyik szempont (változó) a megkérdezett egyén iskolai végzettsége, ami három kategóriából áll: alsó-, közép- és felsőfokú végzettség. A másik szempont a vállalat minősítése aszerint, hogy veszteséges, közepes nyereségű vagy nagy nyereségű-e a vállalat.

Ha a kontingencia táblázatban a gyakoriságok elhelyezkedése valamilyen szabályosságot mutat, akkor érdemes konkrét mutatószámmal kimutatni a kapcsolat szorosságát.

Az asszociációnál alkalmazott mutatószámokat több megközelítés szerint kaphatjuk, ezek közül mi azzal foglalkozunk, amelyik függetlenséget tételez fel (χ^2 - próba).

A χ^2 - próba

A próba két változó közötti kapcsolat „valódiságának” az eldöntésére szolgál. Ez a módszer önmagában nem mutatja meg a kapcsolat erősségét, csak arra ad választ, hogy a változók között van-e ténylegesen kapcsolat egy bizonyos valószínűségi szint mellett.

Az egyik változónk legyen r osztályba sorolható, míg a másik változót c osztályba soroljuk. Jelöljük a keresztábla általános elemét x_{ij} -vel. A nullhipotézisünk (H_0) szerint a két vizsgált változó független egymástól.

A statisztikai próba célja az, hogy megállapítsuk, milyen mértékű eltérés tapasztalható a megfigyelt értékek és a nullhipotézisek alapján elméletileg várt értékek között. Az eltérés mértéke a változók egymásra hatásából adódik. Minél nagyobb ez az eltérés, annál nagyobb a valószínűsége, hogy a változók között tényleges kapcsolat van.

A próba:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*},$$

ahol f_{ij}^* az elvárt, elméleti gyakoriság (feltételezve a függetlenséget):

$$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{n}.$$

A χ^2 értékére érvényes a következő reláció: $0 \leq \chi^2 \leq N \cdot \min\{(r-1), (c-1)\}$, ahol a szorzandó a kapcsos zárójelben lévő számok kisebbike. A χ^2 értéke pontosan akkor nulla, ha a két ismérv függetlennek tekinthető, és akkor éri el a maximumát, ha a két ismérv között függvényeszerű kapcsolat van.

A továbbiakban különválasztjuk a 2×2 -es és az $r \times c$ -s ($r, c > 2$) táblákat, megnézzük, hogy hogyan vizsgáljuk a függetlenséget, és milyen asszociációs mérőszámokat lehet használni az egyes esetekben. Számos mutatót

dolgoztak ki az asszociáció mérésére, ezek közül a legáltalánosabban használtakat tekintjük át.

Asszociáció és függetlenség 2×2 -es táblában

Legyen két változónk, ezeket jelöljük A -val és B -vel, és mindkét változó legyen dichotóm (kétértékű). A két-két kategóriát jelöljük A_1 és A_2 -vel, illetve B_1 és B_2 -vel. A megkérdezett személyeket így négy típusra lehet bontani aszerint, hogy (A_1, B_1) , (A_1, B_2) , (A_2, B_1) és (A_2, B_2) kategóriák melyikébe esnek. Jelöljük f_{ij} -vel azoknak az eseteknek a számát, amelyek az (A_i, B_j) kategóriába esnek. Rendezzük a gyakoriságokat a 79. táblázatba.

79. táblázat. A 2×2 -es kontingencia táblázat

	B_1	B_2	Összesen
A_1	f_{11}	f_{12}	$f_{1.}$
A_2	f_{21}	f_{22}	$f_{2.}$
Összesen	$f_{.1}$	$f_{.2}$	N

Az utolsó oszlopban ($f_{i.}$) és az utolsó sorban ($f_{.j}$) szereplő gyakoriságokat peremgyakoriságoknak (feltétel nélküli eloszlásoknak) nevezzük, míg a többi gyakoriságot (f_{ij}) feltételes eloszlásoknak hívjuk. A táblázatban $f_{i.}$ az i -edik sor összegét (vagyis az A_i kategóriába eső válaszadók teljes számát), $f_{.j}$ a j -edik oszlop összegét (vagyis a B_j kategóriába eső válaszadók teljes számát) jelöli, míg N a teljes válaszadók számát jelenti.

A táblázat alapján az alábbi kérdésekre kereshetjük a választ:

Az A és B változók kapcsolódnak egymáshoz, vagy függetlenek?

Ha nem függetlenek, hogyan kapcsolódnak egymáshoz?

A változók függetlenségének tesztelése

Ha az A és B változók függetlenek egymástól, akkor ez azt jelenti, hogy az A kategóriájának ismerete semmiféle információt nem ad a B kategóriájára

nézve. Matematikailag ha A és B függetlenek, akkor $\frac{f_{11}}{f_{.1}} = \frac{f_{12}}{f_{.2}}$ és $\frac{f_{11}}{f_{1.}} = \frac{f_{21}}{f_{2.}}$ függetlenség tesztelésekor a két-változós valószínűségeloszlást kell tekinteni. Tekintsük a 80. táblázatot, ami az elméleti valószínűségeloszlást tartalmazza.

80. táblázat. 2×2 -es táblázat elméleti valószínűség eloszlása

	B_1	B_2	Összesen
A_1	p_{11}	p_{12}	$p_{1.}$
A_2	p_{21}	p_{22}	$p_{2.}$
Összesen	$p_{.1}$	$p_{.2}$	1

A táblázatban szereplő p_{ij} elméleti valószínűség annak a valószínűségét adja meg, hogy véletlenszerűen kiválasztva egy megfigyelési egységet, az éppen az (i, j) cellához tartozik-e. A peremeloszlásokat a következőképpen lehet felírni: $p_{i.} = \sum_{j=1}^2 p_{ij}$, $p_{.j} = \sum_{i=1}^2 p_{ij}$ és $\sum_i \sum_j p_{ij} = 1$. (A teljes valószínűség 1-gyel egyenlő, hiszen egy válaszoló a négy cella valamelyikébe mindenképpen beletartozik.)

Ha az A és B változók függetlenek, akkor a B_1 kategóriába tartozók azon aránya, akik az A_1 kategóriába tartoznak, meg kell hogy egyezzen a B_2 kategóriába tartozók azon arányával, akik az A_1 kategóriába tartoznak, vagyis:

$$\frac{p_{11}}{p_{.1}} = \frac{p_{12}}{p_{.2}} = \frac{p_{1.}}{1},$$

azaz $p_{11} = p_{1.} \cdot p_{.1}$.

Ha a B_1 kategóriát tekintjük feltételnek, akkor az A kategóriára akkor nincs hatással ez a feltétel, ha:

$$\frac{p_{11}}{p_{1.}} = \frac{p_{21}}{p_{2.}} = \frac{p_{.1}}{1},$$

vagyis $p_{11} = p_{1.} \cdot p_{.1}$. Általánosságban azt mondhatjuk, hogy az A és B függetlenek egymástól, ha $p_{ij} = p_{i.} \cdot p_{.j}$, $(i, j = 1, 2)$.

A tényleges valószínűségeket rendszerint nem ismerjük, azok becslését azonban a gyakoriságok alapján megkaphatjuk: $\hat{p}_{ij} = \frac{f_{ij}}{N}$. Függetlenséget feltételezve:

$$\frac{f_{ij}}{N} = \frac{f_{i.} \cdot f_{.j}}{N^2}.$$

A várható gyakoriság az A és B változók függetlenségét feltételezve:

$$f_{ij}^* = N \cdot \hat{p}_{ij} = \frac{f_{i.} \cdot f_{.j}}{N}.$$

A tényleges és a várható gyakoriság alapján χ^2 -függvény tapasztalati értékét kiszámítjuk és összehasonlítjuk az elméleti értékkel, amit α szignifikancia-szint és $(i-1) \cdot (j-1)$ szabadsági fok mellett keresünk meg. Ha az empirikus χ^2 érték nagyobb, mint az elméleti érték, az adott valószínűségi szinten elvetjük a függetlenségre vonatkozó nullhipotézist. A χ^2 kiszámítása 2×2 -es táblázat esetén:

$$\chi^2 = N \cdot \frac{(f_{11} \cdot f_{22} - f_{12} \cdot f_{21})^2}{f_{i.} \cdot f_{.2} \cdot f_{.1} \cdot f_{.2}}.$$

Ha a függetlenséget elvetjük, akkor a kapcsolat erősségét is kiszámíthatjuk. Az asszociáció mérésére több mutatót is kidolgoztak, ezek közül a legáltalánosabban használtakat tekintjük át.

Az asszociáció mérése 2×2 -es táblázat esetében

Yule-féle asszociációs együttható: $Q = \frac{f_{11} \cdot f_{22} - f_{12} \cdot f_{21}}{f_{11} \cdot f_{22} + f_{12} \cdot f_{21}}$. Az együttható a $[-1,1]$ intervallumban vehet fel értéket. Ha N értéke viszonylag nagy, akkor Q normális eloszlású.

Goodman és Kruskal-féle τ mérték: A Goodman és a Kruskal-féle τ mérték 2×2 -es tábla esetében megegyezik a χ^2 statisztikával, ha a χ^2 értékét elosztjuk N -nel: $\tau = \frac{(f_{11} \cdot f_{22} - f_{12} \cdot f_{21})^2}{f_{i.} \cdot f_{.2} \cdot f_{.1} \cdot f_{.2}}$.

Közvetlenül χ^2 -en alapuló mértékek: A függetlenség tesztelésére alkalmas χ^2 érték az asszociáció mérésére is alkalmas, ha különböző transzformációt hajtunk végre rajta. A transzformáció végrehajtására azért van szükség, mert a χ^2 értéke a $[0, \infty]$ intervallumba esik. A következő két mutató alkalmas az asszociáció jellemzésére:

$$\Phi = \sqrt{\frac{\chi^2}{N}}, \text{ ahol } \Phi^2 \text{ értéke megegyezik a } \tau \text{ értékkel;}$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}.$$

A C -t kontingencia-együtthatónak nevezzük. Ennek a mutatónak az a hátránya, hogy maximuma nem 1, hanem $\frac{1}{\sqrt{2}}$.

Asszociáció és függetlenség $r \times c$ -s táblában

Tekintsük általánosságban az r sorból és c oszlopból álló kétdimenziós kontingencia táblázatot (81. táblázat).

A 2×2 -es tábla elemzésénél láttuk, hogy függetlenség esetén annak valószínűsége, hogy a mintából egy esetet véletlenül kiválasztva az a táblázat i -edik sor j -edik oszlopába, vagyis az (i, j) cellába esik: $p_{ij} = p_{i.} \cdot p_{.j}$, ($i = 1, \dots, r$; $j = 1, \dots, c$). Itt sem ismerjük a p_{ij} elméleti valószínűséget, de meg lehet becsülni a mintából:

$$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{N}.$$

A függetlenség tesztelése a χ^2 próbával:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*},$$

ahol f_{ij}^* az elvárt, elméleti gyakoriság (feltételezve a függetlenséget):

$$f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{n}.$$

81. táblázat. Kontingencia táblázat

		Az Y ismerv szerinti osztályok					Összesen	
		C_1^Y	C_2^Y	...	C_j^Y	...		C_c^Y
Az X ismerv szerinti osztályok	C_1^X	f_{11}	f_{12}	...	f_{1j}	...	f_{1c}	$f_{1.}$
	C_2^X	f_{21}	f_{22}	...	f_{2j}	...	f_{2c}	$f_{2.}$
	·							
	C_i^X	f_{i1}	f_{i2}	...	f_{ij}	...	f_{ic}	$f_{i.}$
	·							
	C_r^X	f_{r1}	f_{r2}	...	f_{rj}	...	f_{rc}	$f_{r.}$
Összesen		$f_{.1}$	$f_{.2}$...	$f_{.j}$...	$f_{.c}$	N

Az asszociáció mérése $r \times c$ -s táblázat esetében

Az asszociáció mérésénél a változók két mérési típusát különböztetjük meg: a nominális változókat és az ordinális változókat. Nézzük meg, hogy a két változó esetén milyen asszociációs mérőszámokat használhatunk.

Nominális változókhoz tartozó asszociációs mutatók

Közvetlenül a χ^2 -en alapuló mértékek: A 2×2 -es táblánál alkalmazott mértékek ugyanúgy alkalmazhatók az $r \times c$ -s táblák esetében is.

Cramer-féle asszociációs együttható: $V = \sqrt{\frac{\chi^2}{N \cdot (r-1)}}$ ha $r \leq c$; illetve $V = \sqrt{\frac{\chi^2}{N \cdot (c-1)}}$ ha $r > c$. A Cramer-féle asszociációs együttható értéke 0, akkor a két mutató független, míg az 1-hez közeli érték nagyon erős kapcsolatra utal.

A Csuprov-féle asszociációs együttható: $T = \sqrt{\frac{\chi^2}{N \cdot \sqrt{(r-1) \cdot (c-1)}}}$. A T értéke szintén 0 és 1 között mozog.

Goodman és Kruskal-féle τ mérték: A τ mérték értéke a $[0,1]$ intervallumba

esik és kiszámításának képlete:
$$\tau = \frac{N \cdot \sum_{i=1}^r \sum_{j=1}^c \left(\frac{f_{ij}^2}{f_i} \right) - \sum_{j=1}^c f_{.j}^2}{N^2 - \sum_{j=1}^c f_{.j}^2}.$$

Ordinális változókhoz tartozó asszociációs mutatók

A következőkben olyan táblákkal foglalkozunk, amelyben az A és B változók kategóriái rendezettek, vagyis, ha mondjuk valaki az A változó első kategóriájába kerül magasabbra rangsorolt, mint aki a második kategóriába került.

82. táblázat. Az ordinális változók esetén bevezetett négy mennyiség

Az új jelölés:	A megfigyelési egységek azon párojainak a teljes száma, amelyekre:
S	vagy $i > i'$ és $j > j'$ vagy $i < i'$ és $j < j'$
D	vagy $i > i'$ és $j < j'$ vagy $i < i'$ és $j > j'$
T_a	$i = i'$
T_b	$j = j'$

Vezessünk be négy új jelölést. Tekintsük a megfigyelt személyek egy általános párosítását. Az egyik személy tartozzon az (i, j) cellához, vagyis az A változó i -edik kategóriájához és a B változó j -edik kategóriájához. A másik személy kerüljön az (i', j') cellába. Az asszociáció ordinális mértéke a következő négy mennyiségnek a függvénye (82. táblázat):

Ha az A és B változók között erős az asszociáció értéke akkor az S értéke nagy, és D értéke kicsi lesz. Ez azt jelenti, hogy az asszociációt az S és a D különbségével, ennek a különbségnek a standardizálásával kell mérni.

Goodmann és Kruskal-féle γ : $\gamma = \frac{S-D}{S+D}$. A γ mértéknek a valószínűségi értelmezése: annak a valószínűségéből, hogy a mintából véletlenszerűen kiválasztott két megfigyelés hasonlóan rendezett vonjuk ki annak a valószínűségét, ha nem hasonlóan rendezett, eltekintve azoktól a pároktól, amelyek valamelyik változó azonos kategóriájába esnek. A γ a $[-1,1]$ intervallumban veheti fel az értékét. Ha az A és B változók függetlenek, akkor γ átlagosan nulla.

Kendall-féle τ^* : Kendall-mértéke figyelembe veszi az azonos kategóriákba való esést is: $\tau = \frac{2 \cdot (S-D)}{\sqrt{(S+D+T_a) \cdot (S+D+T_b)}}$.

Somer-féle d^{} :** Somer javasolta, hogy a vizsgálat során vegyék figyelembe azt is, hogy B függ A -tól, vagy fordítva. Ha a B a függő változó a d mértékét Somer a következő képlettel definiálta: $d_{ba} = \frac{S-D}{S+D+T_b}$. Hasonló értelmezést adhatunk ennek a Somer-féle értéknek, mint a γ -nak, azzal a különbséggel, hogy most azt feltételezzük, hogy az A változó szerint nincsenek kategóriaegyezések, vagyis $i = i'$.

Térjünk vissza a kiinduló feladatunkhoz, amelyben azt akarjuk megvizsgálni, hogy a különböző végzettségű vezetők és a mezőgazdasági vállalkozások jövedelmezősége között van-e összefüggés. Első lépésként – a számítások egyszerűsége miatt – megmutatjuk, hogy hogyan lehet kiszámítani a fenti

képletekben is szereplő, és a statisztikai vizsgálatokban gyakran alkalmazott χ^2 értékét.

Ehhez készítsünk el egy olyan táblázatot (83. táblázat), ami a tapasztalati gyakoriságokat tartalmazza a $f_{ij}^* = \frac{f_{i.} \cdot f_{.j}}{n}$ képlet alapján. A táblázatba az adatok

számítása a következők szerint történik: $f_{11}^* = \frac{720 \cdot 470}{1500} = \frac{338400}{1500} = 225,6$;

$f_{12}^* = \frac{555 \cdot 470}{1500} = \frac{260850}{1500} = 173,9$, stb.

83. táblázat. A tapasztalati gyakoriságok

A vállalat	Alsó- fokú végzettségű vezetők	Közép- vezetők	Felső- vezetők	Összesen
veszteséges	225,6	173,9	70,5	470
közepes nyereségű	240	185	75	500
nagy nyereségű	254,4	196,1	79,5	530
Összesen	720	555	225	1500

A Cramer-féle asszociációs együttható kiszámításához a 84. táblázat ad segítséget.

84. táblázat. Munkatábla a Cramer-féle együttható kiszámításához

A vállalat		f_{ij}	f_{ij}^*	$f_{ij} - f_{ij}^*$	$(f_{ij} - f_{ij}^*)^2$	$\frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$
veszteséges	alsófokú	280	225,6	54,4	2959,36	13,11773
	középfokú	145	173,9	-28,9	835,21	4,802818
	felsőfokú	45	70,5	-25,5	650,25	9,223404
közepes nyer.	alsófokú	260	240	20	400	1,666667
	középfokú	180	185	-5	25	0,135135

	felsőfokú	60	75	-15	225	3
nagy nyer.	alsófokú	180	254,4	-74,4	5535,36	21,75849
	középfokú	230	196,1	33,9	1149,21	5,860326
	felsőfokú	120	79,5	40,5	1640,25	20,63208
Összesen		1500	1500	0	-	80,19665

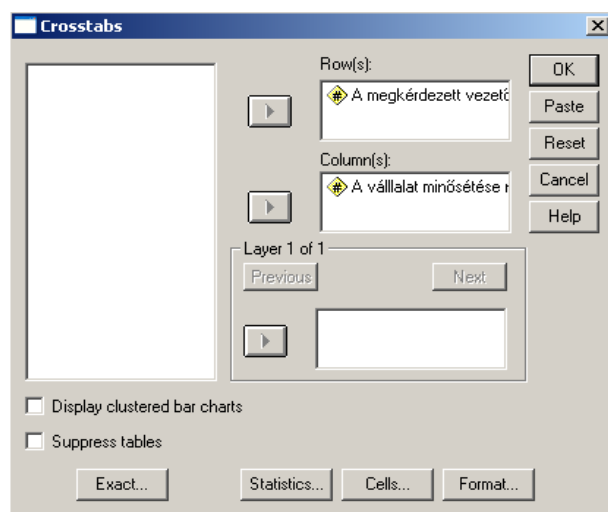
A táblázat alapján a χ^2 értéke 80,197. A kapott χ^2 érték segítségével már a korábban ismertetett asszociációs mérőszámok meghatározhatók.

Mivel a példában $r = c$, így ha a Cramer-féle együtthatót számítjuk ki, akkor a $v = \sqrt{\frac{\chi^2}{N \cdot (r-1)}}$ képletbe helyettesítünk és ezzel: $v = \sqrt{\frac{80,197}{1500 \cdot 2}} = 0,163$. Ha a Csuprov-féle asszociációs együtthatót határozzuk meg, akkor:

$$T = \sqrt{\frac{\chi^2}{N \cdot \sqrt{(r-1) \cdot (c-1)}}} = \sqrt{\frac{80,197}{1500 \cdot \sqrt{3-1}}} \cong 0,19$$

A V és a T értéke alapján a vezetői szint és a veszteség közötti kapcsolat nem tűnik jelentősnek.

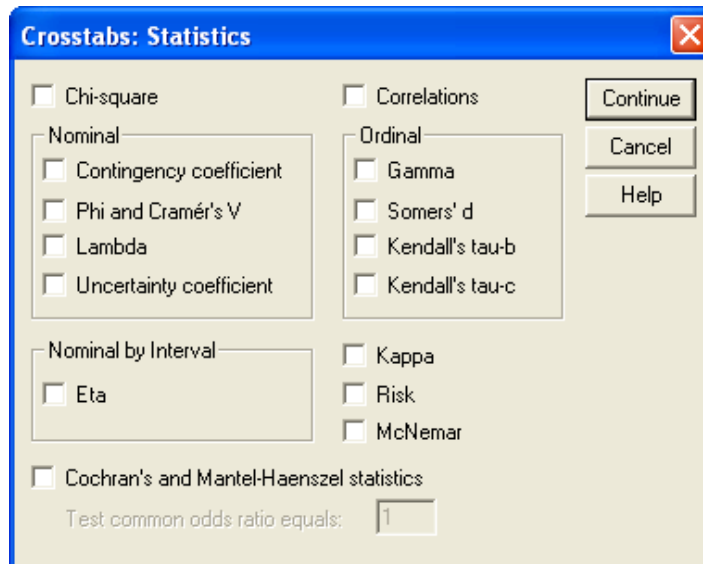
Az SPSS-ben az asszociációs vizsgálatot a következők szerint végezhetjük el. Először elkészítetjük a keresztábrát, amit az ANALYZE menüpont DESCRIPTIVE STATISTICS menüjének CROSSTABS... parancsán belül tehetünk meg.



Ahogy azt megszoktuk a bal oldali ablakból a megfelelő változókat tegyük a Row(s) és a COLUMN(s) ablakokba. A DISPLAY CLUSTERED BAR CHARTS mellé tegyük pipát és egyelőre semmilyen más beállítást ne hajtsunk végre, csak egyszerűen futtassuk le a programot.

45. ábra. A keresztábra elkészítéséhez tartozó panel

A továbbiakban nézzük meg a parancsablak egyéb beállítási lehetőségeit. Ha a STATISTICS gombra kattintunk, megjelenik a 47. ábrán látható panel. Itt állathatjuk be azt, hogy a program írja ki a Khi-négyzet statisztika értékét (CHI-SQUARE).



47. ábra. A STATISTICS parancsgomb beállításai

Az elméleti összefoglaláshoz hasonlóan láthatjuk felsorolva nominális és ordinális változók esetén a különböző asszociációs mérőszámokat. Megjelölve a khi-négyzet statisztikát és a PHI AND CRAMÉR'S V asszociációs mérőszámot, futtassuk le a programot.

Rangkorreláció

Két változó közötti összefüggés vizsgálatának egyik egyszerű és gyors módszere a rangkorreláció. Ilyen esetben első lépésként a változók megfigyelt értékeit rangsoroljuk és az egyes megfigyeléseknek a rangsoruknak megfelelő rangszámot adunk 1-től n -ig, ahol n a megfigyelési egységek száma. Azt vizsgáljuk, hogy a változók rangszámai az azonos megfigyelési egységeken mennyire egyeznek meg.

Az ordinális mérési szintű változók közötti kapcsolat jellemzésére használhatjuk a Spearman-féle rangkorrelációs együtthatót (ρ), a Kendall-féle rang- vagy konkordancia mutatót (w), ezek a legismertebb rangkorrelációs együtthatók.

Ha a két ordinális skálán mért változók 1 és n közötti rangjait (sorszámait) R_{x_i} -vel és R_{y_i} -vel, akkor a két változó közötti kapcsolat szorosságának mérésére bevezetett Spearman-féle rangkorrelációs együtthatót az alábbi képlettel határozhatjuk meg:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n (R_{x_i} - R_{y_i})^2}{n \cdot (n^2 - 1)}$$

A Spearman-féle rangkorrelációs együttható értéke -1 és 1 közé esik. Ha az érték 1-hez közeli, akkor a két sorrend azonosnak tekinthető, a -1-hez közeli érték a két sorrend fordítottságára utal. A 0 közeli eredmény azt mutatja, hogy a két sorrend között nincs kapcsolat.

86. táblázat. Az almafajták sorrendje az íz és szín szerint

Alma sorszáma	Íz szerinti sorrend	Szín szerinti sorrend
1	6	6
2	2	3
3	3	1
4	5	7
5	1	2
6	4	4
7	8	8
8	7	5

8 almafajta íz és szín közötti összefüggését keressük (*rangkorrelacio.sav*). A 8 almát bármilyen sorrendben 1-től 8-ig sorszámozzuk, majd íz és szín szerint rangsoroljuk őket (86. táblázat).

A legrosszabb ízű alma az 1-es, a legjobb a 8-as rangszámot kapja, míg szín szerint a legvilágosabbnak az 1-es, a legsötétebbnek a 8-as értéket adjuk. Előfordulhat, hogy két vagy több megfigyelés között nem tudunk különbséget tenni, ilyenkor ezeknek azonos rangszámot adunk. Az azonos rangszámú megfigyelések ún. kötést képeznek.

Vegyük a megfelelő rangszámok különbségének négyzetét (87. táblázat).

Behelyettesítve a Spearman-féle rangkorrelációs együtthatót megadó képletbe (a megfigyelt esetek száma $n=8$).

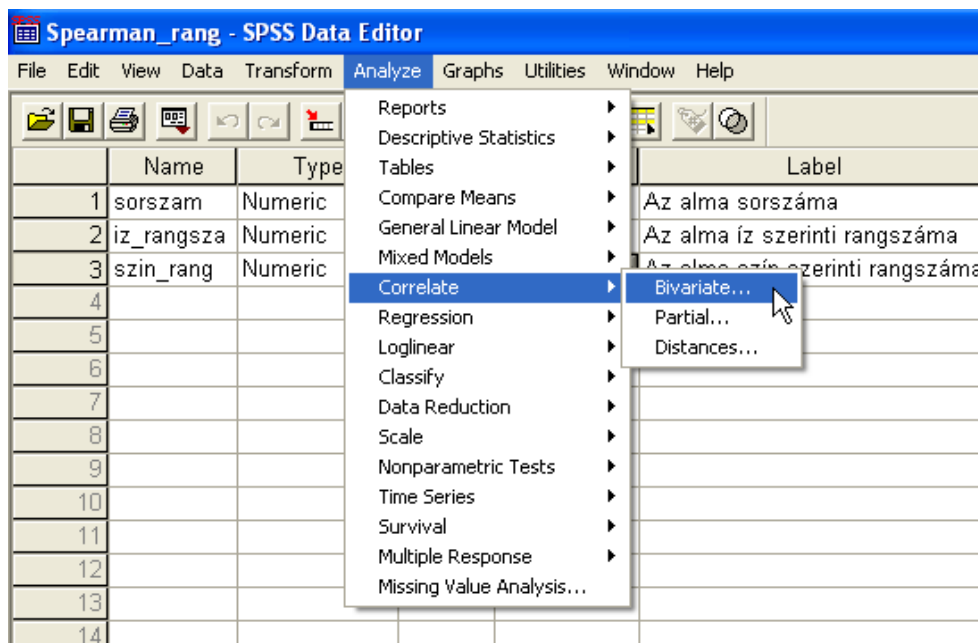
$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n (R_{x_i} - R_{y_i})^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot (0+1+4+4+1+0+0+4)}{8 \cdot (8^2 - 1)} \cong 0,833$$

87. táblázat. Munkatábla a rangkorreláció számításához

Alma sorszáma	R_{x_i}	R_{y_i}	$R_{x_i} - R_{y_i}$	$(R_{x_i} - R_{y_i})^2$
1	6	6	0	0
2	2	3	1	1
3	3	1	2	4
4	5	7	-2	4
5	1	2	-1	1
6	4	4	0	0
7	8	8	0	0
8	7	5	2	4

A rangkorrelációs koefficiens statisztikai próbájához alkalmazhatjuk az ρ -táblázatot (. melléklet) $df = n - 2$ szabadsági fokkal. Példánkban a számított ρ nagyobb, mint $df = 6$ esetén az 5%-os szinten megadott táblázati ρ érték $0,7067$, ami azt jelenti, hogy az almák színe és íze közötti kapcsolat szignifikáns.

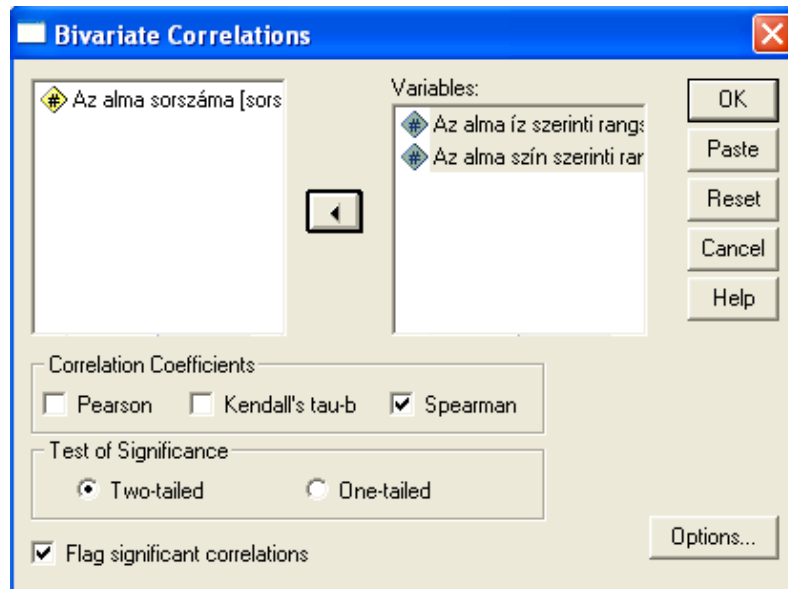
Az SPSS statisztikai programcsomagban végezzük el ugyanezt a számítást. Nyissuk meg az adatokat tartalmazó „Spearman_rang.sav” fájlt, majd kattintsunk az ANALYZE menüpont CORRELATE almenüjében a BIVARIATE... parancsra (48. ábra).



48. ábra. A Spearman-féle rangkorreláció parancssora az SPSS-ben

A megjelölt panelban (49. ábra) a bal oldali ablakrészben vannak a változók megadva, amelyek közül ki kell választanunk azokat a változókat, amelyek

között a Spearman-féle rangkorrelációt ki akarjuk számítani. Jelöljük ki ezeket a változókat, majd helyezük a VARIABLES ablakba. A CORRELATION COEFFICIENTS részben meg kell adni, hogy milyen korrelációt kívánunk számítani, itt a SPEARMAN felirat mellé tegyük pipát.



49. ábra. A rangkorreláció elvégzése az SPSS-ben

Miután a megfelelő beállításokat elvégeztük, futtassuk le a programot, majd elemezzük a kapott eredményt (88. táblázat).

88. táblázat. Az SPSS által végzett rangkorreláció-számítás eredménye

Correlations			Az alma íz szerinti rangszáma	Az alma szín szerinti rangszáma
Spearman's rho	Az alma íz szerinti rangszáma	Correlation Coefficient	1,000	,833*
		Sig. (2-tailed)	.	,010
		N	8	8
	Az alma szín szerinti rangszáma	Correlation Coefficient	,833*	1,000
		Sig. (2-tailed)	,010	.
		N	8	8

*.Correlation is significant at the 0.05 level (2-tailed).

Az eredményül kapott táblázatban a vizsgált változók közötti kapcsolat szorosságáról (Correlation Coefficient), a korreláció szignifikanciaszintjéről (Sig. 2-tailed) és a változónként rendelkezésre álló elemszámról (N)

tájékozódhatunk. Először a szignifikancia értéket nézzük meg, ami a hipotézisvizsgálat eredménye. Nullhipotézisünk alapján a két változó között nincs kapcsolat. Mivel a szignifikancia sorában $p < 0,05$, így elvetjük a nullhipotézist, azaz az alma íze és színe között van kapcsolat. Mivel a kapcsolat szignifikáns, megnézzük a Spearman-féle rangkorrelációs együttható értékét, amit a CORRELATION COEFFICIENT sorban találunk. Az itt szereplő 0,833 érték megegyezik a kézi számítás során kapott értékkel. Mivel a korreláció értéke pozitív, ez azt jelenti, hogy nagyobb „íz-rangszámhoz” nagyobb „szín-rangszámok” tartoznak.

Korábban utaltunk arra, hogy előfordul olyan eset is, amikor két vagy több megfigyelt eset között nem tudunk különbséget tenni, vagyis rangsorolásuk nem egyértelmű. Az ilyen egyedeknek adjunk azonos rangszámot, s ahogy azt korábban jeleztük, ezek az egyedek ún. kötésben állnak egymással. Jelöljük a kötés elemeinek a számát t -vel. A példánkat módosítsuk olyan formában, hogy ízben a 3. és 6. sorszámú almát ne tudjuk megkülönböztetni, így mindkettő a 3,5-es rangszámot fogja kapni. Színben a 2., 5. és 6. illetve 4. és 7. sorszámú almákat ne tudjuk megkülönböztetni. Mivel az első három a 2., 3. és 4. szín-ranghelyeken vannak, így átlagosan a 3-as számot kapják, az utóbbi kettő pedig a 7. és 8. szín-ranghelyeket megosztva átlagosan a 7,5 rangszámot kapja.

89. táblázat. Az almafajták sorrendje íz és szín szerint

Alma sorszáma	Íz szerinti sorrend	Szín
1	6	6
2	2	3
3	3,5	1
4	5	7,5
5	1	3
6	3,5	3
7	8	7,5
8	7	5

Az ízben egyetlen kötés van $t = 2$ elemmel, míg a színben két kötés van $t = 3$ és $t = 2$ elemmel. A kötések a ρ rangszám kiszámításakor figyelembe kell venni úgy, hogy a kötésekől korrekciós tényezőt kell kiszámítani.

Jelöljük T_A -val az A tulajdonság, T_B -vel a B tulajdonság korrekciós tényezőjét. Ezzel a rangkorrelációs képlet az alábbiak szerint módosul:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n [(R_{x_i} - R_{y_i})^2 + T_A + T_B]}{n \cdot (n^2 - 1)},$$

ahol $T_A = \frac{\sum t_A \cdot (t_A^2 - 1)}{12}$ és $T_B = \frac{\sum t_B \cdot (t_B^2 - 1)}{12}$. A \sum jel az azonos tulajdonságon belüli különböző kötésekre vonatkozik.

Példánkban az **A** tulajdonságban (íz) egy kötés van, így $t_A = 2$ elemmel, így:

$$T_A = \frac{2 \cdot (2^2 - 1)}{12} = 0,5.$$

A **B** tulajdonságban két kötés van $t = 3$ és $t = 2$ elemmel:

$$T_B = \frac{3 \cdot (3^2 - 1) + 2 \cdot (2^2 - 1)}{12} = 2,5.$$

A kapott értékeket helyettesítsük be a $\rho = 1 - \frac{6 \cdot \sum_{i=1}^n [(R_{x_i} - R_{y_i})^2 + T_A + T_B]}{n \cdot (n^2 - 1)}$ képletbe:

$$\rho = 1 - \frac{6 \cdot (22 + 0,5 + 2,5)}{8 \cdot (8^2 - 1)} = 0,702.$$

Mivel a számított ρ értéke kisebb, mint a $df = 8 - 2 = 6$ szabadsági foknál és 5%-os szignifikancia-szintnél megadott elméleti ρ érték ($\rho = 0,7067$), így csak 10%-os szignifikancia-szint mellett kapunk szignifikáns összefüggést.

Ha az SPSS-el az eddig szokásos módon végeznénk el a vizsgálatot, más eredményt kapnánk, ugyanis az SPSS-be beépített ρ nem számol a kötésekkel. Futtassuk le erre az adatbázisra is a vizsgálatot (*Spearman_kotes.sav*) és az eredményül kapott táblázatunkat (90. táblázat) vessük össze a kézi számítás eredményével.

90. táblázat. A Spearman-féle korreláció értéke

Correlations			Az alma íz szerinti sorszáma	Az alma szín szerinti sorszáma
Spearman's rho	Az alma íz szerinti sorszáma	Correlation Coefficient	1,000	,729*
		Sig. (2-tailed)	.	,040
		N	8	8
	Az alma szín szerinti sorszáma	Correlation Coefficient	,729*	1,000
		Sig. (2-tailed)	,040	.
		N	8	8

*. Correlation is significant at the 0.05 level (2-tailed).

Előfordul, hogy nem két rangsort, hanem többet kell összehasonlítani. Ilyen típusú feladatoknál a *Kendall-féle konkordancia*, vagy *egyetértési mutató-t* használjuk, melyet a

$$W = \frac{12 \cdot \sum_{i=1}^n (R_i - \bar{R})^2}{m^3 \cdot (n^3 - n)}$$

képlet alapján kapunk meg. A képletben lévő m a különböző sorrendek száma, n az elemek száma, R_i az i -edik elem rangszám-összege és \bar{R} az

$$\bar{R} = \frac{m \cdot (n+1)}{2}$$

átlagos oszlopösszeg, vagyis $\frac{m \cdot (n+1)}{2}$. Az egyetértési mutató értéke 0 és 1 közé esik. Azt mondjuk, ha ez az érték 0,6 fölötti, akkor a felállított sorrendek azonosnak tekinthetők.

91. táblázat. Az almák íz, szín és eladás szerinti sorrendje

Az alma sorszáma	Íz	Szín	Eladási ár	R_i
1	6	6	7	19
2	2	3	3	8
3	3	1	2	6
4	5	7	5	17
5	1	2	1	4
6	4	4	4	12
7	8	8	8	24
8	7	5	6	18

A táblázat utolsó oszlopa az egyes változók rangszám-összegével van kiegészítve. Határozzuk meg az átlagos oszlopösszeg értékét:

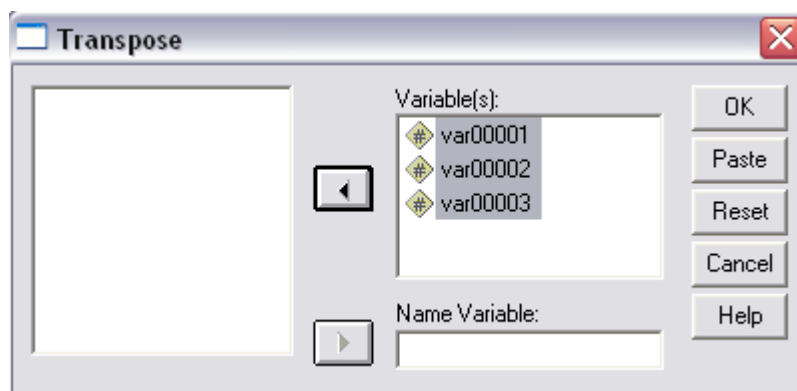
$$\bar{R} = \frac{m \cdot (n+1)}{2} = \frac{3 \cdot (8+1)}{2} = 13,5, \text{ mivel } m = 3 \text{ és } n = 8.$$

A Kendall-féle mutató értéke:

$$W = \frac{12 \cdot [(5,5)^2 + (-5,5)^2 + (-7,5)^2 + (3,5)^2 + (-9,5)^2 + (-1,5)^2 + (10,5)^2 + (4,5)^2]}{3^2 \cdot (8^3 - 8)} \cong 0,931.$$

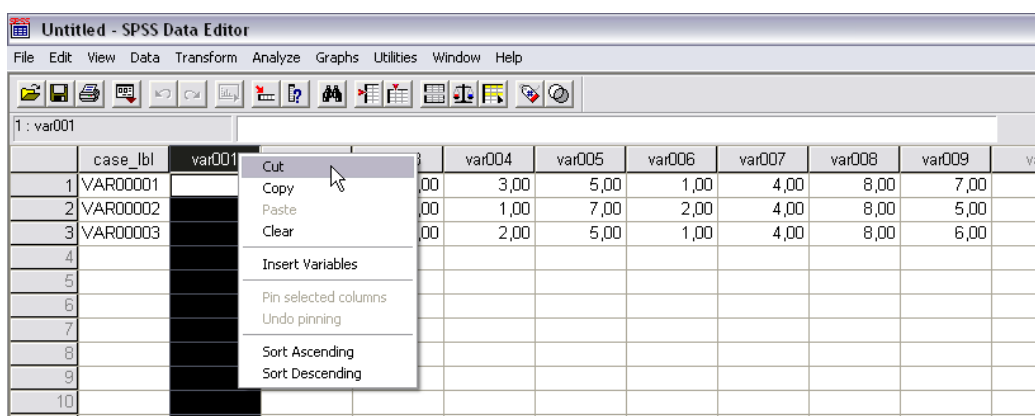
Mivel 0,6-nál nagyobb értéket kaptunk, így elmondhatjuk, hogy az almák sorrendje a három változó tekintetében hasonlóan tekinthető (közelítőleg 93%-ban tekinthetők a sorrendek azonosnak).

Nyissuk meg a KendallW.xls fájlt, ami Excel táblázatban tartalmazza az adatainkat. Ebben a táblázatban három oszlopban jelenítjük meg az „íz”, „szín” és „eladási ár” változókat. Jelöljük ki a táblázatot, majd másoljuk át a *KendallW_rang.sav* név alatt megnyitott SPSS fájl DATA VIEW adattáblájába. A másolás után az első sor üresen maradt, nem jelent meg adat, ezzel most ne foglalkozzunk. A részletes leírást azért mutatjuk be, mert ennél a mutató kiszámításánál az adatmátrixunkat transzponálni kell, hiszen nem az „íz”, „szín” és „eladási ár” változókat akarjuk összehasonlítani, hanem az almák sorrendjére vagyunk kíváncsiak a három változó tekintetében.



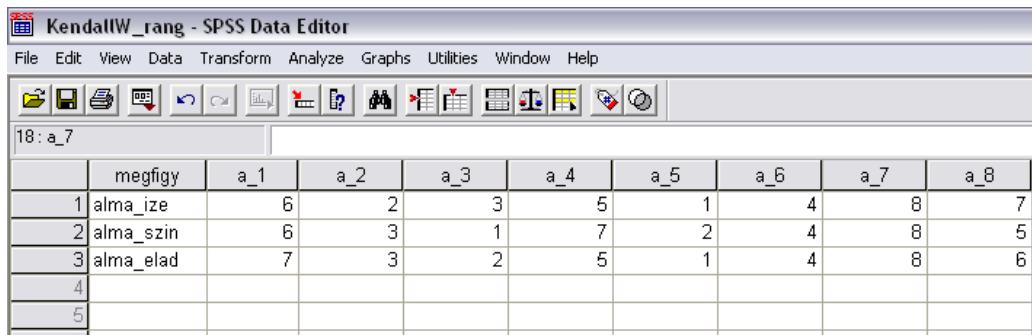
50. ábra. A transzponálás művelet elvégzése az SPSS-ben

A transzponálás műveletét a DATA menü TRANSPOSE... parancsa alatt végezzük el. A megjelent panel (50. ábra) bal oldali ablakából a VARIABLE(S) ablakba helyezzük át a még varR000001, var00002 és var00003 változókat, majd kattintsunk az OK gombra.



51. ábra. A transzponálás után a DATA VIEW

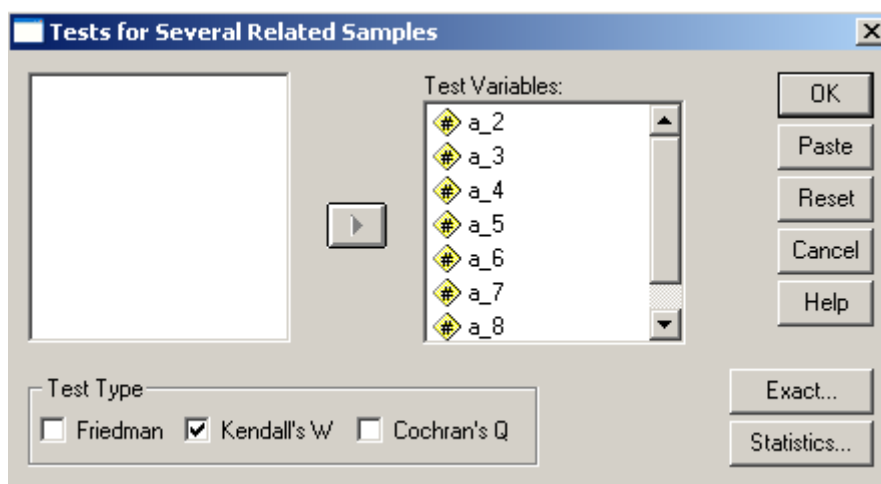
A tranzponálás elvégzése után a DATA VIEW ablak a 51. ábrán látható módon fog kinézni. Jelöljük ki a var001 oszlopot és töröljük. Legyen VAR00001: alma_ize, a VAR00002: alma_szin és VAR00003: alma_elad, ezeket az átnevezéseket egyszerűen az adott cellára lépve és beírva módosíthatjuk. Ezután felcímkézhetjük a változókat. A Case_Ibl változónak név helyett adjuk a „megfigy” nevet, majd az egyes almafajtákat rendre jelöljük a_1, a_2,...,a_8 jelölésekkel. Ezzel elkészült az az adatfájl, amin most már elvégezhetjük a rangkorrelációs vizsgálatunkat (52. ábra).



	megfigy	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
1	alma_ize	6	2	3	5	1	4	8	7
2	alma_szin	6	3	1	7	2	4	8	5
3	alma_elad	7	3	2	5	1	4	8	6
4									
5									

52. ábra. A KendallW_rang.sav fájl a DATA VIEW ablaka

Kattintsunk az ANALYZE menü NONPARAMETRIC TEST almenüjének K RELATED SAMPLES... parancsára (53. ábra). Jelöljük ki az almákat, ezek sorrendjét akarjuk ugyanis összehasonlítani a változók tekintetében és tegyük át a TEST VARIABLES listába ezeket.



53. ábra. Több rangsor összehasonlítása a

Kendall-féle konkordancia mutató segítségével

Az alkalmazott teszt típusa (TEST TYPE) mezőben a KENDALL'S W tesztet jelöljük meg. A beállítások után futtassuk le a programot, majd elemezzük az eredményül kapott 92. táblázatot.

92. táblázat. A Kendall-teszt eredménye

Test Statistics	
N	3
Kendall's W	,931
Chi-Square	19,556
df	7
Asymp. Sig.	,007

a. Kendall's Coefficient of Concordance

A Kendall-féle egyetértési mutató értékét a második sorban olvassuk le, ami látható, hogy nagyobb 0,6-nél (és megegyezik a kézi számítás eredményével), vagyis az almák sorrendje azonosnak tekinthető a vizsgált változók tekintetében, és a kapcsolat szignifikáns $p < 0,05$.

Vegyes kapcsolat

Vegyes kapcsolatról akkor beszélünk, ha mennyiségi és minőségi változók közötti kapcsolatot vizsgálunk (pl. a talajművelés és a termésátlag közötti kapcsolat). A vegyes kapcsolatok vizsgálatára a varianciaanalízist használjuk, amivel már a korábbi fejezetek egyikében részletesen megismerkedtünk.

Két kvantitatív változó közötti kapcsolat elemzése

A kvantitatív változók közötti kapcsolatok jellemzésére – ahogy arra már korábban is utaltunk – a korreláció- és regressziószámítást alkalmazzuk. Amikor magas mérési szintű változók közötti kapcsolatokat elemezzük, több kérdésre keressük a választ: (1) Van-e kapcsolat a változók között? (2) Milyen szoros ez a kapcsolat? (3) Hogyan tudunk következtetni az egyik változó megváltozásából a másik változó megváltozására?

Magas mérési szintű változók közötti kapcsolat vizsgálata

Egy gazdaságban 15 földterületen 15 búzakaralász hosszát és kalásonkénti szemszámát jegyezték fel (93. táblázat). Jelentse az x_i az i -edik kalász

hosszát, míg y_i az i -edik kalász esetén a szemszám mennyiségét (db). Számítsuk ki, hogy milyen erős és milyen irányú a kapcsolat a két változó között. Az adatokat a *lienaris.sav* fájl tartalmazza.

93. táblázat. Búzakaralász hossza és a kalásonkénti szemszám

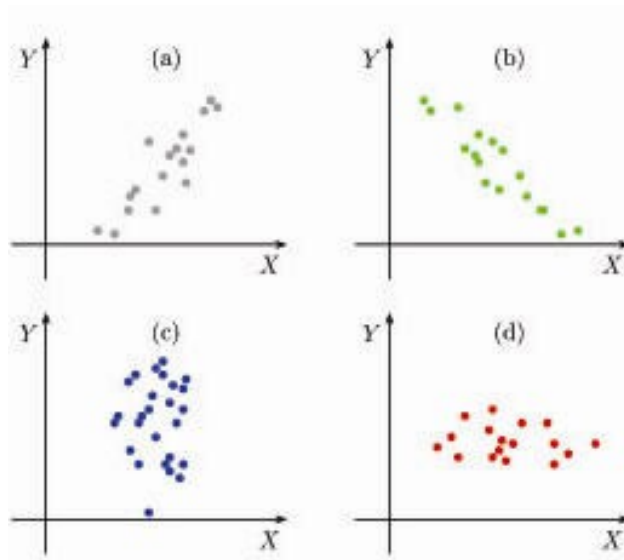
Földterület sorszáma	A kalász hossza (cm) – x_i	Kalásonkénti szemszám (db) – y_i
1	7,1	26
2	7,3	24
3	7,4	25
4	7,6	27
5	7,7	22
6	8,1	30
7	8,2	32
8	8,2	31
9	8,3	33
10	8,5	29
11	9,3	35
12	9,4	37
13	9,5	38
14	9,7	40
15	10,5	41

A kalásonkénti szemszám függ-e a kalász hosszától, ha igen milyen erősségű és milyen irányú a kapcsolat? Állapítsuk meg továbbá, hogy a független változó alakulásából a függő változó alakulására tudunk-e következtetést adni.

Pontdiagram

Amikor mennyiségi mutatók kapcsolatát vizsgáljuk, a mutatószámok meghatározása előtt érdemes ún. pontdiagram-t készíteni. Ekkor az együttesen előforduló (x_i, y_i) mutatókat ábrázoljuk, és az empirikus adatok alapján a pontok elrendeződéséből próbálunk következtetni a kapcsolatra. A pontdiagram segítségével azonban csak szemléletes képet kapunk a kapcsolat erősségéről és irányáról. A korreláció pozitív irányú, ha a pontok elhelyezkedése megegyezik a 54. ábra (a) részén látható pontfelhővel; negatív irányú a kapcsolat a (b) esetben. Minél „vékonyabb” a pontfelhő, annál

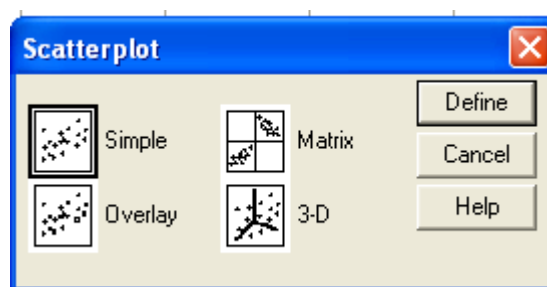
erősebb a kapcsolat, függetlenül annak irányától. A (c) és (d) esetben a változók között nincs kapcsolat.



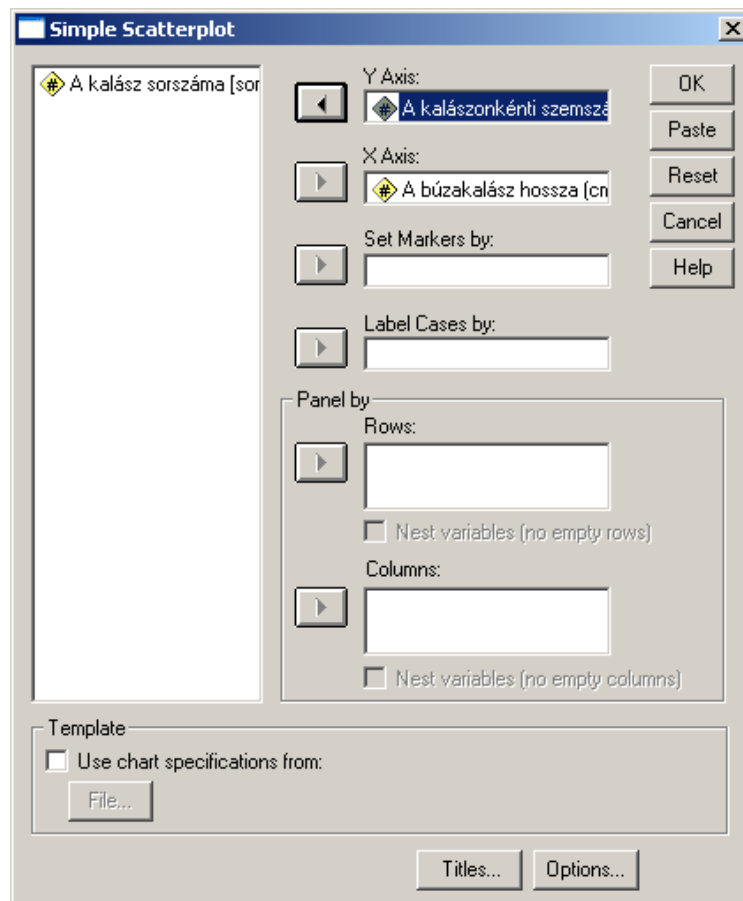
54. ábra. A lineáris korreláció: (a) pozitív, (b) negatív korreláció, (c) és (d) x és y korrelálatlanok

Forrás: ZAR, J. H. (1996)

Az SPSS-ben a pontdiagram készítéshez kattintsunk a **GRAPHS** menü **SCATTER...** parancsára. A megjelent panelben (55. ábra) hagyjuk megjelölve a **SIMPLE** parancsgombot, majd a **DEFINE** gombra kattintsunk.

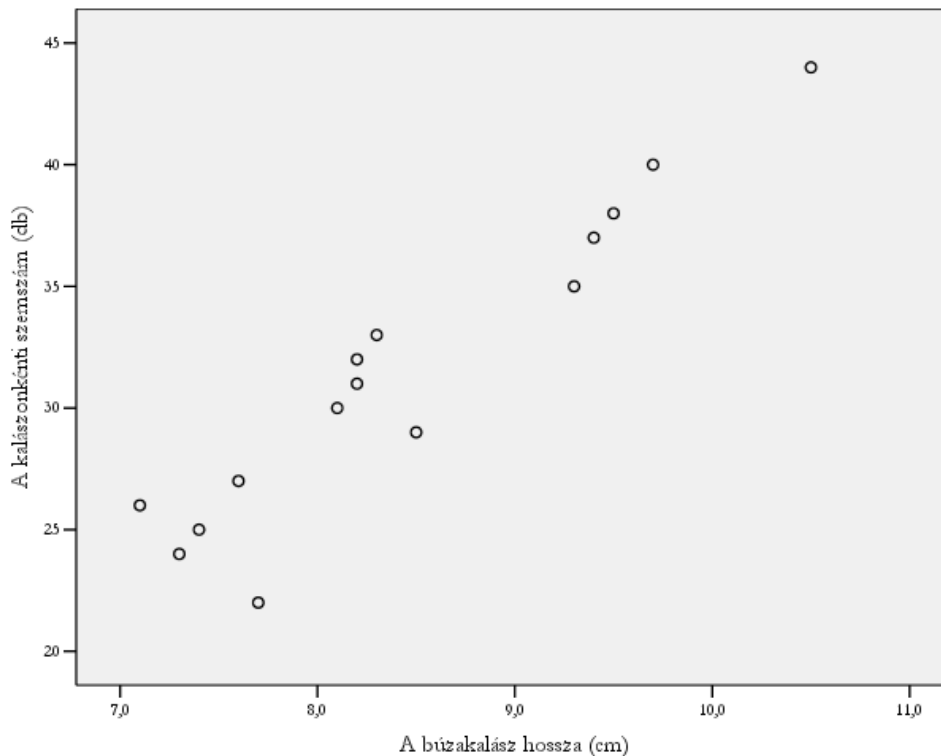


55. ábra. A **SCATTERPLOT** ablak



56. ábra. A SIMPLE SCATTERPLOT... ablak beállításai

A megjelent ablakban (56. ábra) végezzük el a következő beállításokat. A bal oldali ablakból válasszuk ki a független változót és a függő változót, majd tegyük át az „A búzakaralász hossza” változót az „x-tengely” ablakba, a „Kalászonkénti szemszám” függő változót pedig az „y-tengely” ablakba. Ezek után kattintsunk az OK gombra, amelynek eredményeképpen az OUTPUT ablakban megjelenik a pontdiagram.



57. ábra. A karalás hossza és a karalásonkénti szemszám közötti pontdiagram

A 57. ábra a változók közötti kapcsolatot szemlélteti. Kisebb ellentmondások ellenére úgy tűnik, hogy az empirikus megfigyelési pontokra képzeletbeli egyenes illeszthető, amely balról jobbra határozott emelkedő irányt mutat. (Megjegyezzük, hogy az esetek száma igen kevés, az empirikus elemzéseknél azonban nem szabad kevés számú megfigyelés alapján statisztikai összefüggéseket keresni.) A kapott ábra azonban csak vizuálisan mutatja meg a változók közötti kapcsolat jellegét és irányát, számszerű eredményeket itt nem tudunk leolvasni. Hogyan tudjuk azt eldönteni, hogy milyen erős a kapcsolat? Határozzuk meg továbbá azt a függvényt, amely legjobban illeszkedik a ponthalmazra!

Először tekintsük át, hogy a kapcsolat erősségét és irányát milyen mutatószámokkal jellemezhetjük!

Lineáris korrelációs együttható

Amennyiben két változó között lineáris kapcsolat áll fenn, vagyis a pontdiagramon látható pontok közelítően egy képzeletbeli egyenes körül csoportosulnak, akkor a kapcsolat erősségének és irányának számszerűsítésére a *Pearson-féle korrelációs együtthatót* használjuk:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \cdot \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}}$$

A kifejezésben szereplő (x_i, y_i) az X -re és Y -ra vonatkozó n elemű minta ($i = 1, 2, \dots, n$), továbbá $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i$ és $\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i$.

A két változó kapcsolatának mérésére alkalmazott Pearson-féle korrelációs együttható számítására vonatkozó módszer bár a legáltalánosabban használt módszer, alkalmazásának feltételei azonban szigorúak:

Mindkét változó intervallumszintű;

Mindkét változó normális eloszlású;

Feltételezhető, hogy a két változó között lineáris kapcsolat van.

A gyakorlatban ezek a feltételek csak ritkán teljesülnek maradéktalanul (a mutató legkevésbé a normalitásra érzékeny).

Ha tudjuk, hogy a két változó közötti kapcsolat szignifikáns, akkor a gyakorlatban az r értéke alapján a következőket mondhatjuk (94. táblázat):

94. táblázat. A korreláció értékei alapján a változók közötti lineáris kapcsolat jellege

Nincs kapcsolat a két változó között	$-0,25 < r < 0,25$
Gyenge sztochasztikus kapcsolat	$-0,5 < r \leq -0,25$ vagy $0,25 \leq r < 0,5$
Közepes sztochasztikus kapcsolat	$-0,75 < r \leq -0,5$ vagy $0,5 \leq r < 0,75$
Erős sztochasztikus kapcsolat	$-1 < r \leq -0,75$ vagy $0,75 \leq r < 1$
A kapcsolat függvényyszerű	$r = -1$ vagy $r = 1$

A különböző kutatásoknál jelentős szerepet kap a lineáris korrelációs együttható négyzete (r^2), amit *determinációs együtthatónak* nevezünk. A determinációs együttható értéke megmutatja, hogy az X értékei hány százalékban magyarázzák az Y értékeinek az alakulását. A determinációs együttható értékére: $r^2 \in [0,1]$.

Mivel az r és az r^2 szimmetrikus kapcsolatot kifejező mutatószámok, így az X -nek Y -nal való korrelációja megegyezik az Y -nak az X -szel való

korrelációjával, azaz nincs jelentősége annak, hogy melyiket tekintjük függő illetve független változónak. A Pearson-féle korrelációs együttható a lineáris kapcsolatok erősségét méri, így nem alkalmazható nem lineáris kapcsolatok esetében. Az $r=0$ csupán azt jelenti, hogy nincs lineáris kapcsolat az X és Y változó között és nem azt, hogy nincs közöttük kapcsolat. Ha ugyanis a két változó között nem lineáris a kapcsolat, akkor azt az r értéke nem mutatja meg.

Korrelációs index

Az elemzéseknél gyakori, hogy a változók közötti kapcsolat nem lineáris. Ha a két vizsgált változó közötti kapcsolat nem lineáris, akkor a változók közötti kapcsolat erősségének megadására nem a korrelációs együtthatót, hanem a korrelációs indexet szoktuk használni (I):

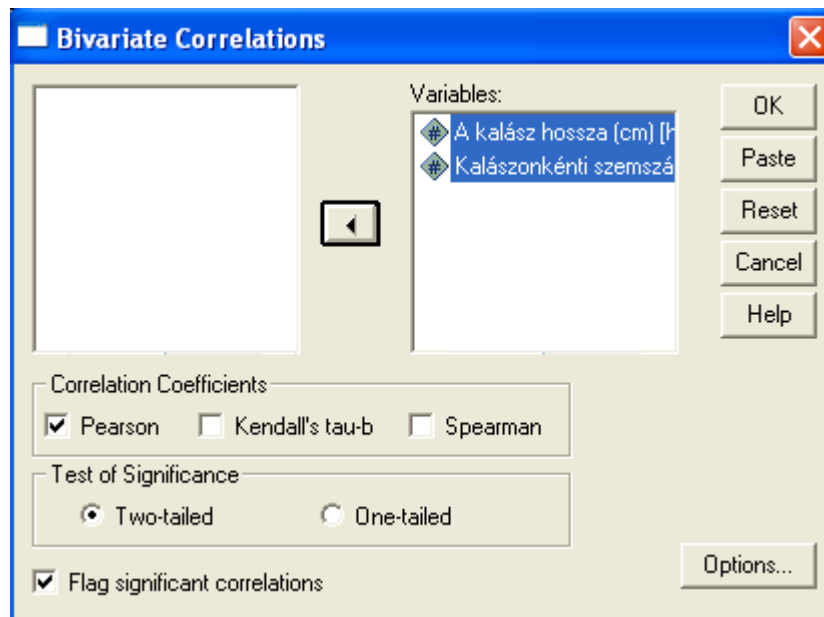
$$I = \sqrt{1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

ahol e_i a tapasztalati (mért) y_i és a becsült függvény által számolt \hat{y}_i értékek közötti eltérést jelenti.

A korrelációs index előjelét nem tudjuk értelmezni, csak az abszolút nagyságát, amelyre: $I \in [0,1]$.

A lineáris korrelációs együttható meghatározása SPSS-ben

Két változó közötti kapcsolat erősségének és irányának számszerű megadására készítsük el az SPSS-ben a korrelációs mátrixot. Az SPSS-ben a Pearson-féle korrelációs együttható kiszámítását az ANALYZE / CORRELATE / BIVARIATE... menüpontban végezhetjük el – éppen úgy, ahogy azt tettük a Spearman-féle korrelációs együttható kiszámításánál is. A megjelenő panelben (58. ábra) a bal oldali ablakból a két változót helyezzük a jobb oldali ablakba, majd a Pearson-korreláció beállítása mellett kérjük le az eredményt az Ok gombra kattintva.



58. ábra. A Pearson-féle korreláció vizsgálatnak parancssora

A korrelációs számítás eredményeképpen a 95. táblázatot kapjuk.

95. táblázat Az SPSS outputja A Pearson-féle korreláció elvégzésekor

Correlations			
		A kalász hossza (cm)	Kalászonkénti szemszám (db)
A kalász hossza (cm)	Pearson Correlation	1	,938*
	Sig. (2-tailed)		,000
	N	15	15
Kalászonkénti szemszám (db)	Pearson Correlation	,938*	1
	Sig. (2-tailed)	,000	
	N	15	15

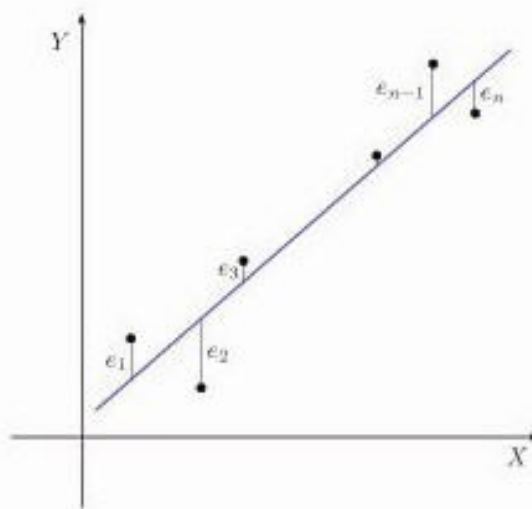
** .Correlation is significant at the 0.01 level (2-tailed).

A korrelációs együttható szignifikancia-vizsgálata jelenti a kapott táblázat elemzésének első lépését. Azt kell eldönteni, hogy a kapott r érték valódi, szignifikáns kapcsolatot jelent-e a két változó között, vagy csak a véletlen hatások eredőjeként keletkezett. A statisztikai próba nullhipotézise szerint a két változó között nincs kapcsolat. Mivel $p < 0,05$, ezért elvetjük a nullhipotézist, azaz a két változó között van kapcsolat. Ha tudjuk, hogy a két változó közötti kapcsolat nem a véletlennek köszönhető, megnézhetjük a kapcsolat szorosságát. A korrelációs együttható értéke 0,938, ami igen erős

sztochasztikus kapcsolatot jelent. A változók közötti kapcsolatot lineáris függvénnyel tudjuk legjobban közelíteni.

A regressziós egyenes

Miután tudjuk, hogy a két változó közötti kapcsolat lineáris függvénnyel modellezhető, adódik a kérdés, hogy hogyan kapjuk meg azt az egyenest, ami a pontokra legjobban illeszkedik. A pontok és az egyenes távolságának meghatározása az adott pontból az illesztett egyenesig (regressziós egyenesig) húzott függőleges (Y tengellyel párhuzamos) távolság alapján történik.



59. ábra. A regressziós egyenes illesztése az egyenes és pont távolságának mérése alapján

A regressziós egyenes illesztése úgy történik, hogy a távolságokat összegezzük, majd ezt az összeget minimalizáljuk. Ugyanis a legjobban illeszkedő regressziós egyenes az az egyenes, ahol a távolságok összege minimális. Erre a gyakorlatban a *legkisebb négyzetek módszerét* szoktuk használni, ami a nevében is mutatja, hogy az eljárás a távolságok négyzetösszegeit minimalizálja. (A legkisebb négyzetek elvének kidolgozása Gauss (1777-1855) német matematikushoz köthető.)

Tekintsünk egy n elemből álló mintát. Két-változós modell esetében a regressziós egyenes általános képlete $y_i = \beta_1 \cdot x_i + \beta_0$ alakban írható fel, ahol β_0 és β_1 a regressziós paraméterek. Ez a modell determinisztikus kapcsolatot ír le, amelyben az x teljesen meghatározza y -t. A regressziós egyenes illesztése az empirikus vizsgálatokban soha sem tökéletes, az általa meghatározott értékek eltérnek a tényleges értékektől. Az eltérés a *hibatag*-ban nyilvánul meg, amit jelöljünk ε_i -vel.

Ezek alapján a következő összefüggést írhatjuk fel: $y_i = \beta_1 \cdot x_i + \beta_0 + \varepsilon_i$. A β_0 és β_1 tényleges paraméterek regressziós egyenes alapján becsült értékeit jelöljük $\hat{\beta}_0$ és $\hat{\beta}_1$ -el, míg a hibatagok becsült értékeit (az ún. *reziduum*-okat) e_i -vel. Ezek alapján:

$$y_i = \hat{\beta}_1 \cdot x_i + \hat{\beta}_0 + e_i \quad (i = 1, 2, \dots, n)$$

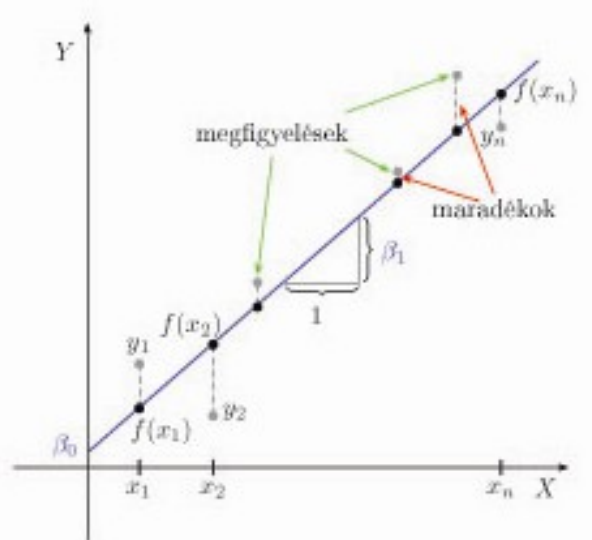
illetve

$$\hat{y}_i = \hat{\beta}_1 \cdot x_i + \hat{\beta}_0,$$

ahol $\hat{y}_i = y_i - e_i$. Az e_i maradékok fontos szerepet játszanak a modellezésben, ugyanis megmutatják, hogy a modell mennyire közelíti a valóságot, hiszen e_i kis értékei jó, nagy értékei pedig gyenge illeszkedést jeleznek.

A legkisebb négyzetek módszere

Tekintsünk a síkban n számú pontot (60. ábra): (x_i, y_i) , ahol $i = 1, 2, \dots, n$; $n \in \mathbb{N}$, feltéve, hogy $x_i \neq x_j$ ha $i \neq j$. A minta alapján becsült regressziós függvény $\hat{y}_i = \hat{\beta}_1 \cdot x_i + \hat{\beta}_0$ ($i = 1, 2, \dots, n$).



60. ábra. Legkisebb négyzetek módszere

Keressük a $\hat{\beta}_0$ és $\hat{\beta}_1$ becsült paraméterek értékeit, amely mellett a megfigyelésekből származó $(f(x_i) = y_i)$ és a regressziós függvény alapján becsült értékek (\hat{y}_i) különbségének négyzetösszege minimális:

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i - y_i)^2 \rightarrow \min.$$

A feladat tehát azon $\hat{\beta}_0$ és $\hat{\beta}_1$ (becsült) értékek meghatározása, amelyekre a $Q(\hat{\beta}_0, \hat{\beta}_1)$ két-változós függvény minimális értéket vesz fel. (Egy adott függvény szélsőértéke létezésének szükséges feltétele, hogy az első deriváltja nullával legyen egyenlő. Mivel két-változós függvényről van szó, így el kell készíteni a két változó szerinti elsőrendű parciális deriváltakat, és ezeket kell nullával egyenlővé tenni.) Elvégezve a deriválást $\hat{\beta}_0$ és $\hat{\beta}_1$ függvényében a következő egyenletrendszert kapjuk:

$$Q'_{\hat{\beta}_0}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n 2 \cdot (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i - y_i) = 0$$

$$Q'_{\hat{\beta}_1}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n 2x_i \cdot (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i - y_i) = 0.$$

Az egyenletek kettővel egyszerűsíthetők. A zárójelek felbontása és egyenletrendezés után jutunk az ún. *normálegyenletek*-hez:

$$\left(\sum_{i=1}^n x_i \right) \cdot \hat{\beta}_1 + n \cdot \hat{\beta}_0 = \sum_{i=1}^n y_i$$

$$\left(\sum_{i=1}^n x_i^2 \right) \cdot \hat{\beta}_1 + \left(\sum_{i=1}^n x_i \right) \cdot \hat{\beta}_0 = \sum_{i=1}^n x_i \cdot y_i.$$

A számítógépes programok a lineáris regressziós függvények paramétereinek becslésére a fenti egyenletrendszer megoldásával kész eljárást adnak meg.

A továbbiakban a lineáris regressziós feladaton keresztül kézi számítással bemutatjuk a vizsgálat menetét, majd az SPSS-ben történő elemzésre térünk át. A nemlineáris regressziószámítás esetében csak az SPSS segítségével történő elemzésre térünk ki.

A lineáris regressziószámítás menete

A lineáris függvény meghatározása

Számítsuk ki a lineáris regressziófüggvény paramétereit! Ehhez meg kell oldanunk a legkisebb négyzetek módszerénél levezetett normálegyenletekből álló lineáris egyenletrendszert:

$$\left(\sum_{i=1}^n x_i \right) \cdot \hat{\beta}_1 + n \cdot \hat{\beta}_0 = \sum_{i=1}^n y_i$$

$$\left(\sum_{i=1}^n x_i^2 \right) \cdot \hat{\beta}_1 + \left(\sum_{i=1}^n x_i \right) \cdot \hat{\beta}_0 = \sum_{i=1}^n x_i \cdot y_i$$

Az egyenletrendszer felírásához szükséges adatokat összefoglaltuk a 96. táblázatban.

Mivel $n=15$, $\sum_{i=1}^{15} x_i = 126,8$, $\sum_{i=1}^{15} y_i = 470$, $\sum_{i=1}^{15} x_i^2 = 1086,18$ és $\sum_{i=1}^{15} x_i \cdot y_i = 4052,2$, így a megoldandó egyenletrendszer:

$$126,8 \cdot \hat{\beta}_1 + 15 \cdot \hat{\beta}_0 = 470$$

$$1086,18 \cdot \hat{\beta}_1 + 126,8 \cdot \hat{\beta}_0 = 4052,2$$

96. táblázat. A regressziós paraméterek kiszámítása

	x_i	y_i	x_i^2	$x_i \cdot y_i$
1	7,1	26	50,41	184,6
2	7,3	24	53,29	175,2
3	7,4	25	54,76	185
4	7,6	27	57,76	205,2
5	7,7	22	59,29	169,4
6	8,1	30	65,61	243
7	8,2	32	67,24	262,4
8	8,2	31	67,24	254,2
9	8,3	33	68,89	273,9
10	8,5	29	72,25	246,5
11	9,3	35	86,49	325,5
12	9,4	37	88,36	347,8
13	9,5	38	90,25	361
14	9,7	40	94,09	388
15			110,2	
	10,5	41	5	430,5
Σ	126,8	470	1086,18	4052,2

Az egyenletrendszer megoldás: $\hat{\beta}_0 = -15,45$, $\hat{\beta}_1 = 5,53$. A kapott paraméterekkel az illesztett egyenes egyenlete:

$$\hat{y} = -15,54 + 5,53 \cdot x$$

Értelmezzük a kapott eredményeket! A regressziós paraméterek értelmezésekor elsősorban statisztikai, szakmai értelmezést kell adnunk, mert a matematikai értelmezés nem elegendő. A $\hat{\beta}_1$ regressziós együttható megmutatja, hogy az x magyarázó változó egységnyi növekedése az eredményváltozó mekkora változásával jár együtt. Tehát az x változó értékét 1 egységgel növelve az y változó értéke átlagosan $\hat{\beta}_1$ értékkel növekszik, vagy csökken. A regressziós együttható pozitív vagy negatív előjele a kapcsolat irányát fejezi ki. A $\hat{\beta}_0$ paraméter az $x=0$ esetre ad elméleti értéket. Természetesen ez csak abban az esetben értelmezhető, ha a 0 érték beletartozik az x -ek közé.

Visszatérve most a példára: A kapott $\hat{\beta}_1$ regressziós együttható értéke alapján azt mondhatjuk, hogy egy cm-rel nagyobb kalászhossz esetében átlagosan 5-6 szemmel több a kalásonkénti szemszám. A $\hat{\beta}_0$ paraméter értelmezésének példánkban nincs értelme.

A korrelációs együttható és a determinációs együttható kiszámítása

Számítsuk ki a Pearson-féle korrelációs együtthatót, amit a

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \cdot \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]}}$$

képlettel definiáltunk.

97. táblázat. A Pearson-féle korreláció munkatáblázata

(x_i)	(y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
7,1	26	1,35	5,33	7,22	1,83	28,44
7,3	24	1,15	7,33	8,46	1,33	53,78
7,4	25	1,05	6,33	6,67	1,11	40,11
7,6	27	0,85	4,33	3,70	0,73	18,78
7,7	22	0,75	9,33	7,03	0,57	87,11
8,1	30	0,35	1,33	0,47	0,12	1,78

8,2	32	-	0,67	-0,17	0,06	0,44
8,2	31	-	0,33	0,08	0,06	0,11
8,3	33	-	1,67	-0,26	0,02	2,78
8,5	29	0,05	2,33	-0,11	0,00	5,44
9,3	35	0,85	3,67	3,10	0,72	13,44
9,4	37	0,95	5,67	5,36	0,90	32,11
9,5	38	1,05	6,67	6,98	1,10	44,44
9,7	40	1,25	8,67	10,80	1,55	75,11
10,5	41	2,05	9,67	19,78	4,19	93,44

Vezessük be a következő jelöléseket: az SP jelentse az X és Y változók összes eltérésszorzatát; SQ_x az X változó összes eltérésnégyzetét és SQ_y pedig az Y változó összes eltérésnégyzetét. A korrelációs együttható kiszámításához készítettük el a 97. táblázatot.

A táblázat alapján tekintsük a következő összegzések eredményeit:

$$\frac{\sum_{i=1}^{15} x_i}{15} = \bar{x} = 8,45, \quad \frac{\sum_{i=1}^{15} y_i}{15} = \bar{y} = 31,33, \quad \sum_{i=1}^{15} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = 79,13, \quad \text{azaz } SP = 79,13;$$

$$\sum_{i=1}^{15} (x_i - \bar{x})^2 = 14,3, \quad \text{azaz } SQ_x = 14,3$$

$$\sum_{i=1}^{15} (y_i - \bar{y})^2 = 497,33, \quad \text{azaz } SQ_y = 497,33.$$

A kapott részeredményekből:

$$r = \frac{79,13}{\sqrt{14,3 \cdot 497,33}} \cong 0,938.$$

Az r értéke alapján a 98. táblázattal összhangban azt mondhatjuk, hogy a két vizsgált változó között erős, sztochasztikus kapcsolat van.

A korrelációs koefficiens négyzetét kiszámítva kapjuk meg a determinációs együtthatót: $r^2 \cong 0,8798$. A determinációs együttható értelmezése szerint a vizsgálati mintában a búza kalásonkénti szemszámának változatosságát közel 88%-ban tulajdoníthatjuk a kalász hosszának, és csak 12%-ban befolyásolják ezt egyéb tényezők.

A regresszió szignifikanciavizsgálata

A két változó összefüggésének szignifikanciavizsgálata

A két változó összefüggésének szignifikanciavizsgálatát a varianciaanalízisnél tárgyalt F -próbával végezzük el. A varianciaanalízis táblázat szerkezetével már a korábbi fejezetekben megismerkedtünk (98. táblázat).

98. táblázat. A varianciaanalízis táblázat

A szóródás oka	Az eltérések négyzetösszege	Szabadsági fok	Szórásnégyzetek becslése	F
Regresszió	$SP^2/SQ_x = SSR$	1	$MSR = SSR/1$	$\frac{MSR}{MSE}$
Hiba	$SQ_y - SP^2/SQ_x = SSE$	$n - 2$	$MSE = SSE/(n - 2)$	
Összesen	$SQ_y = SST$	$n - 1$		

Készítsük el a varianciaanalízis táblázatunkat a fenti sémának megfelelően úgy, hogy a megfelelő értékeket behelyettesítjük (99. táblázat).

99. táblázat. A varianciaanalízis táblázat

A szóródás oka	Az eltérések négyzetösszege	Szabadsági fok	Szórásnégyzetek becslése	F
Regresszió	$79,13^2 / 14,3 = 437,8$	1	437,87	$\frac{437,87}{4,57} = 95,81$
Hiba	$497,33 - 437,87 = 59,46$	13	4,57	
Összesen	497,33	14		

A 99. táblázat alapján a számított F érték: 95,81. Az „ F -próba kritikus értékei” táblázatból keressük ki a tapasztalati F értéket: $F_{0,1\%} = 17,81$. Mivel a számított érték nagyobb, mint a tapasztalati érték, így azt mondhatjuk, hogy a két változó között az adott szignifikancia szinten szignifikáns összefüggés van.

A regressziós egyenesből számított \hat{y}_i értékek hibája

Két esetet kell vizsgálnunk:

Az első esetben az a kérdés, hogy a független változó valamely meghatározott x_i értékéhez tartozó átlagos \hat{y} értéknek a becslés során milyen a hibája. A konfidenciahatárok számítását a regressziós egyenes egyenletéből az x_i ponthoz számított \hat{y}_i értéktől, vagyis az ábrázolt egyenestől függőleges \pm

irányban: $h_{\hat{y}_i} = \pm t_{p\%} \cdot \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SQ_x} \right)}$, $df = n - 2$ esetén megadott t értékkel

($t_{5\%} = 2,16$). Feladatunk ebben az esetben az, hogy pl. a kiválasztott $x_i = 7,5$ cm-es kalászok esetén meghatározzuk, hogy mekkora lesz az átlagos kalásonkénti szemszám becslésnek a hibája? Válasszuk a $p = 5\%$ -os szignifikancia szintet. A kapott regressziós egyenletbe ($\hat{y} = -15,54 + 5,53 \cdot x$) ha behelyettesítjük az x helyére a 7,5 értéket, akkor $\hat{y}_{7,5} = 25,935$ átlagos szemszámot kapunk. A becslés konfidenciahatára:

$h_{\hat{y}_{7,5}} = \pm 2,16 \cdot \sqrt{4,57 \cdot \left(\frac{1}{15} + \frac{(7,5 - 8,45)^2}{14,3} \right)} = \pm 1,66$. Ez azt jelenti, hogy a becslésünk

két konfidenciahatára 7,5 cm hosszú kalász esetén $25,935 \pm 2,74$, vagyis $[23,195; 28,675]$ szem/kalász.

A másik esetben arra vagyunk kíváncsiak, hogy egy egyedre vonatkozó becslésünk milyen hibával terhelt. A konfidenciahatárok képlete:

$h_{\hat{y}_i} = \pm t_{p\%} \cdot \sqrt{MSE \cdot \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SQ_x} \right)}$, $df = n - 2$ esetén megadott t értékkel. $x_i = 7,5$

cm-es kalászt kiválasztva a regressziós egyenes egyenletéből 25,935 db a becsült szemek száma. Ezt a becslést

$h_{\hat{y}_i} = \pm 2,16_{p\%} \cdot \sqrt{4,57 \cdot \left(1 + \frac{1}{15} + \frac{(7,5 - 8,45)^2}{14,3} \right)} = \pm 4,91$ hibahatár terheli. A becslésünk

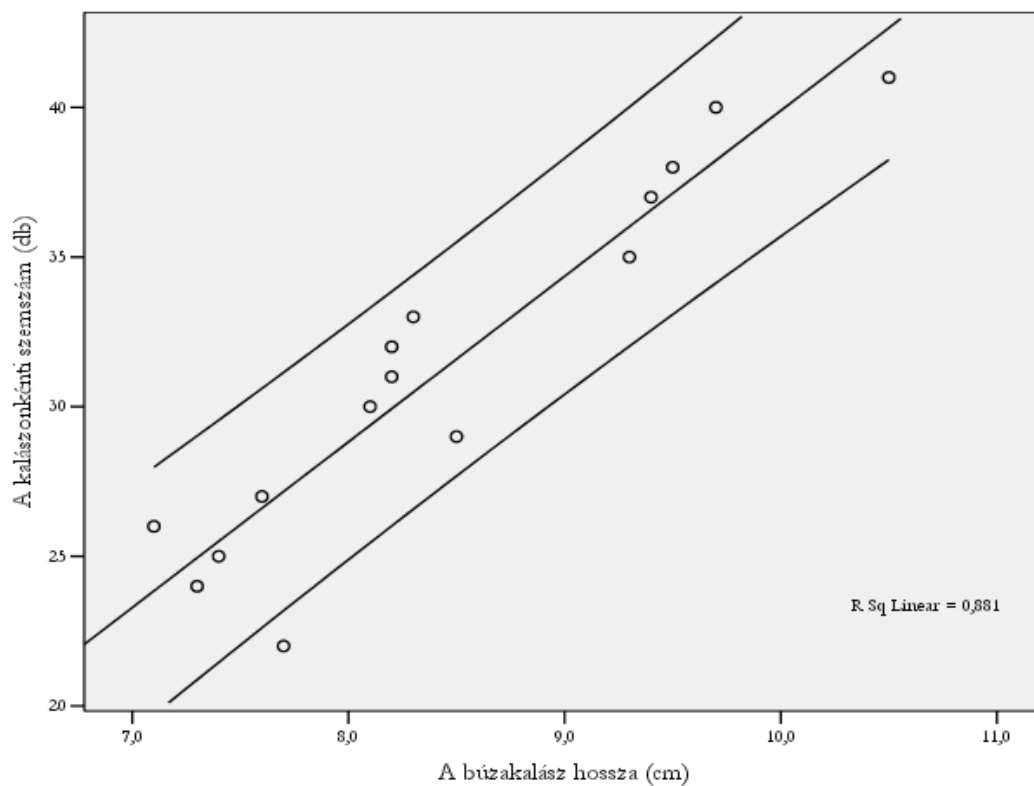
két konfidenciahatára véletlenül kiválasztott 7,5 cm-es kalászra vonatkozóan 21,025 és 30,845 szem.

Számítsuk ki az előbbi konfidenciahatárokat az összes x_i értékekre (100. táblázat). A konfidenciasáv kifejezést általában az első esethez tartozó képlettel határozzuk meg és nem az egyes egyedekre megadott képlet alapján.

Látható, hogy a konfidenciasáv az \bar{x} körül lesz a legkisebb, azonban minél jobban távolodunk az \bar{x} -től, egyre nagyobb konfidencia intervallumokat kapunk (61. ábra). A 61. ábra az 5%-os szignifikancia szinten mutatja a konfidenciasávot.

100. táblázat. A konfidenciasávok $p=5\%$ -s szignifikanciaszinten

x_i	$x_i - \bar{x}$	\hat{y}_i	$h\hat{y}_i$	Alsó	Felső
				Konfidenciahatár	
7,1	-1,35	23,84	$\pm 5,05$	18,80	28,89
7,3	-1,15	24,95	$\pm 4,97$	19,98	29,92
7,4	-1,05	25,50	$\pm 4,94$	20,56	30,44
7,6	-0,85	26,61	$\pm 4,88$	21,73	31,49
7,7	-0,75	27,16	$\pm 4,86$	22,31	32,02
8,1	-0,35	29,38	$\pm 4,79$	24,59	34,17
8,2	-0,25	29,93	$\pm 4,78$	25,15	34,71
8,2	-0,25	29,93	$\pm 4,78$	25,15	34,71
8,3	-0,15	30,48	$\pm 4,77$	25,71	35,26
8,5	0,05	31,59	$\pm 4,77$	26,82	36,36
9,3	0,85	36,02	$\pm 4,88$	31,14	40,90
9,4	0,95	36,57	$\pm 4,91$	31,66	41,48
9,5	1,05	37,13	$\pm 4,94$	32,19	42,06
9,7	1,25	38,23	$\pm 5,01$	33,23	43,24
10,5	2,05	42,66	$\pm 5,39$	37,28	48,05



61. ábra. A lineáris regressziós egyenlet ábrája a konfidenciasávval

A regressziós koeficiens statisztikai próbái

A regressziós koeficiens hibája és statisztikai próbái analógiát mutatnak a középérték hibájával és statisztikai próbáival.

A regressziós koeficiens hibaszórása

A regressziós koeficiens hibaszórását az $\sigma_{\beta_1} = \sqrt{\frac{MSE}{SQ_x}}$ képlet alapján számítjuk

ki, ami a példánkban: $\sigma_{\beta_1} = \sqrt{\frac{4,57}{143}} = 0,565$.

A regressziós koeficiens konfidenciahatárai

A regressziós koeficiens hibahatárait h_1 és h_2 -vel jelölve: $\beta_1 \pm (t_{p\%} \cdot \sigma_{\beta_1})$, $df = n - 2$ szabadságfokú t értékkel. Példánkban 5%-os szignifikancia szint mellett a konfidenciahatárok: $5,53 \pm (2,16 \cdot 0,565)$, azaz 4,31 és 6,75.

A számított (β_1) és a hipotetikus (β) regressziós koeficiens közötti különbség szignifikanciapróbája:

$$t = \frac{\beta_1 - \beta}{s_{\beta_1}}$$

$df = n - 2$ szabadsági fokú t érték alapján történik.

Ellenőrizzük, hogy az 5,53-os regressziós koeficiens eltér-e a sok-éves tapasztalat alapján meghatározott $\beta = 4,57$ -s regressziós koeficiensről. Behelyettesítve:

$$t = \frac{5,53 - 4,57}{0,565} = 1,699$$

Ez a számított t érték kisebb, mint $df = 15 - 2 = 13$ szabadsági fokra a $p = 5\%$ szinten megadott táblázati t érték (2,16). Ez azt jelenti, hogy a mintánk alapján kapott regressziós koeficiens érték nem tér el bizonyíthatóan a sok-éves adatok alapján számított regressziós koeficiens értéktől.

A regressziós egyenlet konstans tagjának próbája

A regressziós egyenes konstans tagja (β_0) az $x=0$ helyen adja meg az Y értékét. Általában azonban az $x=0$ az analízisek túlnyomó többségében a megfigyelési tartományon kívül esik, ezért annak értékéből semmiféle következtetést nem tudunk megállapítani, vagyis hibája nem adható meg.

A korrelációs koefficiens statisztikai próbái

Vizsgáljuk meg, hogy a korrelációs koefficiens értéke szignifikánsan eltér-e nullától? A korrelációs koefficiens statisztikai próbáit ismertetjük a következőkben, ennek vizsgálatára ugyanis több lehetőségünk van.

A varianciaanalízis táblázatából közvetlenül kiszámíthatjuk a determinációs koefficienset, hiszen $r^2 = \frac{SSR}{SST} = \frac{437,87}{497,33} = 0,88$. Ennek négyzetgyöke a korrelációs koefficiens: $r = \sqrt{r^2} = 0,938$. Ez alapján azt mondhatjuk, hogy a regresszió statisztikai próbája megegyezik a korrelációs koefficiens próbájával.

„A korrelációs koefficiens kritikus értékei” táblázatból keressük ki megfelelő szabadságfokhoz tartozó kritikus r értéket. A szabadságfok $df = n - 2$, ahol az n az adat-párok számát jelenti. A példában a szabadságfokok száma 13, az ehhez tartozó kritikus r érték $p = 0,1\%$ -s szignifikancia szinten: $r_{krit} = 0,760$. Mivel a számított r érték ($r = 0,938$) nagyobb, mint a kritikus érték. Ez azt jelenti, hogy $p = 0,1\%$ -os szignifikancia szinten bizonyítottnak tekinthetjük, hogy az r értéke szignifikánsan eltér nullától, azaz az X és Y változók egymással összefüggésben vannak.

Előfordulhat, hogy nem áll rendelkezésünkre táblázat, ebben az esetben a $t = \sqrt{\frac{r^2 \cdot (n-2)}{1-r^2}}$ képletet alkalmazzuk, és $df = n - 2$ szabadságfok mellett végzünk

t -próbát. A példabeli adatokat használva: $t = \sqrt{\frac{0,88 \cdot 13}{1-0,88}} = 9,76$. Ez az érték nagyobb, mint $df = 13$ -hoz tartozó $p = 0,1\%$ esetén leolvasott táblázatbeli t érték ($t = 4,22$), azaz az összefüggés ezen a szignifikancia szinten igazolt.

A továbbiakban határozzuk meg a korrelációs koefficiens hibáját, a konfidenciaintervallumot. Ennek a vizsgálatához első lépésként „A korrelációs koefficiens transzformált Z -értékei” táblázatból keressük ki a korrelációs koefficiens értékét (amit kiszámítottunk) és a bal oldali Z oszlopban, valamint a felső Z sorban leolvassuk a megfelelő Z értéket. A példabeli r érték alapján $z = 1,72$. Ezzel az r értéket Z értékévé transzformáltuk, és látható, hogy a Z értéke felveszi az r érték előjelét. Ezután határozzuk meg a Z

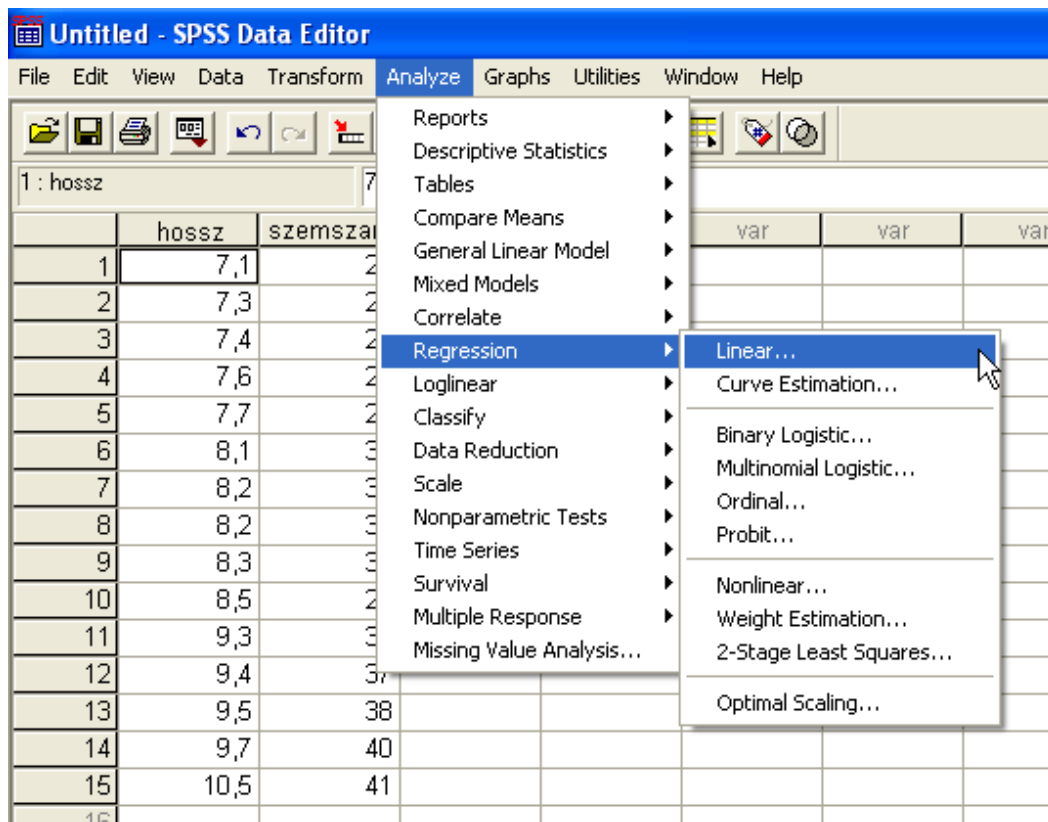
szórását ehhez a $\sigma_z = \sqrt{\frac{1}{n-3}}$ képletet használjuk: $\sigma_z = 0,288$. A képletből Z két konfidenciahatára: $h_1 = z - (t_{p\%} \cdot \sigma_z)$ és $h_2 = z + (t_{p\%} \cdot \sigma_z)$ alapján határozható meg (a t érték mindig $df = \infty$ -re megadott táblázati t érték). Behelyettesítve az adatokat. $h_1 = 1,72 - 1,96 \cdot 0,288 = 1,15$ és $h_2 = 1,72 + 1,96 \cdot 0,288 = 2,28$. Ez a két érték 1,72-es Z -érték konfidenciahatárai $p = 5\%$ -s szignifikanciaszinten. A táblázatból h_1 -et és h_2 -t Z -ről r -re transzformálva kapjuk meg $p\%$ szinten a korrelációs koefficiens két konfidenciahatárát: 0,8178 és 0,9793, ezek a számított r érték konfidencia határai lesznek.

Ha a korrelációs koefficiens pozitív és szignifikáns, akkor a két konfidenciahatár is pozitív; ha a korrelációs koefficiens negatív és szignifikáns, akkor a két konfidenciahatár is negatív. Ha a korrelációs koefficiens nem szignifikáns, akkor a két konfidenciahatár ellenkező előjelű és a 0-t is közrefogja.

A lineáris regresszió elvégzése az SPSS-ben

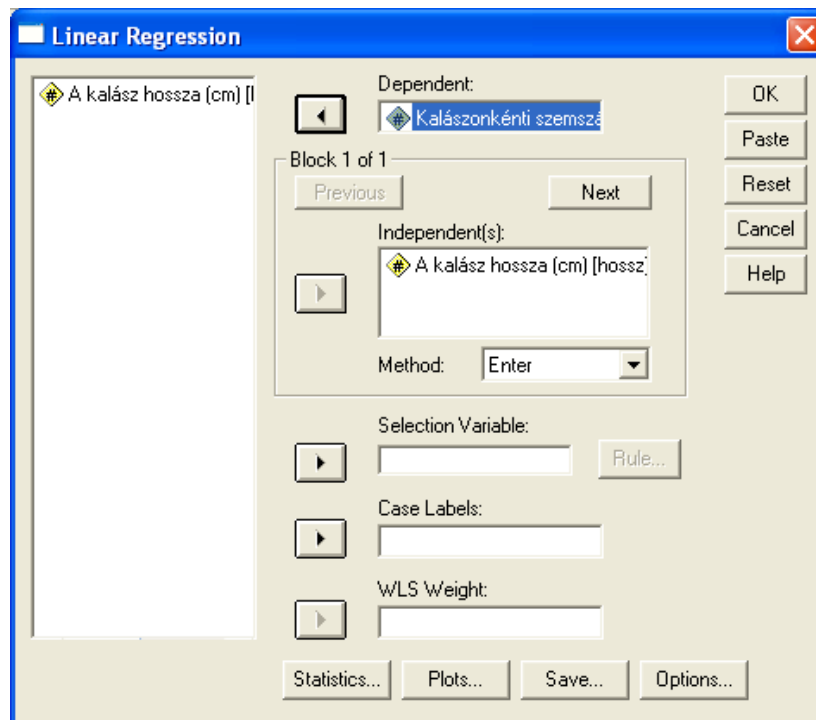
Korábban a pontdiagram alapján azt már láttuk, hogy a tapasztalati értékekre leginkább egyenes illeszthető, így lineáris regressziót kell végeznünk.

Ha a számításokat az SPSS programban végezzük, akkor a regressziószámítás elvégzéséhez kattintsunk az ANALYZE menüpont REGRESSION almenüjében a LINEAR... parancsra (62. ábra).



62. ábra. A lineáris regresszióanalízis parancssora

A bal oldali ablakból (63. ábra) az INDEPENDENT(S) ablakba a független változót („A kalász hossza”), míg a DEPENDENT ablakba a függő változót („Kalásonkénti szemszám”) tesszük át a nyílacska segítségével. A többi beállítással egyelőre ne foglalkozzunk, majd kattintsunk az OK gombra. A beállítások elvégzése után futtassuk le a programot és elemezzük az OUTPUT ablakban megjelent táblázatokat.



63. ábra. A lineáris regresszió beállításai

Több táblázat jelenik meg a program eredményeképpen, az első (101. táblázat) táblázat számunkra ebben a feladatban nem informatív, így ennek elemzésével most nem foglalkozunk (később – a több-változós lineáris regressziónál – térünk erre vissza). A lineáris regresszió eredményeként megjelent következő táblázat (102. táblázat) második oszlopában (R) a korrelációs együttható értékét látjuk.

101. táblázat. Az összesítő táblázat

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,938 ^a	,881	,871	2,137

a. Predictors: (Constant), A kalász hossza (cm)

A harmadik oszlop (R SQUARE) a korrelációs együttható négyzetét tartalmazza. A determinációs együttható (R^2) megadja, hogy a tényezőváltozó az eredményváltozó variációját hány százalékban magyarázza. (Ha $R^2 = 1$, akkor a pont-párok tökéletesen illeszkednek a regressziós egyenesre, ha $R^2 = 0$, akkor a változók között nincs lineáris kapcsolat.) Példánkban $R^2 = 0,881$, vagyis a modell 88,1%-ban tudja magyarázni az Y értékek eltérés

négyzetösszegét, ez az érték azonban torzított becslés. A valóságos, az alapsokaságbeli megmagyarázott hányad torzítatlan becslését az ADJUSTED R SQUARE oszlopban olvassuk le $\left(R_A^2 = R^2 - \frac{1-R^2}{n-2} \right)$. Ez az ún. módosított megbízhatósági együttható (R_A^2) megkísérli kiküszöbölni a mintavételezéskor elkövetett esetleges hibát a két változó elméleti lineáris regressziója erősségének megítélésekor. A STD. ERROR OF THE ESTIMATE oszlopban található érték is bizonyos értelemben szintén a regressziós egyenes illeszkedését jelzi, hiszen ez az érték a reziduálisok szórását jelenti. Minél nagyobb ez az érték, annál inkább számíthatunk olyan kalászonkénti szemszám adatokra, amelyek messze esnek a regressziós egyenes által becsült értéktől.

Azt, hogy sikerült-e a regressziós egyenes segítségével akkora részt „megragadni” a függő változó varianciájából, hogy a független változó hatását szignifikánsnak tekinthessük, varianciaanalízissel teszteljük. Az erre vonatkozó összefoglaló táblázat a következő „elemzésre váró” táblázat (103. táblázat). A táblázat szerkezetét, felépítését a kézi számításoknál részletesen bemutattuk (de a varianciaanalízis fejezetnek köszönhetően már egyébként is ismerős lehet). Ha összevetjük az ott kapott táblázattal (99. táblázat) azt látjuk, hogy egy oszloppal (SIG.) több van az SPSS által készített táblázatban. Gyakorlatilag ez az oszlop az, ami számunkra az elemzés során az eredmény leolvasásához szükséges. A regressziós modell helyességére vonatkozó próba nullhipotézis azt mondja, hogy az Y értékek véletlenszerűen szóródnak, vagyis nem a modellel magyarázható a változásuk. Mivel a szignifikancia érték kisebb 0,05-nél, így elvetjük a nullhipotézist, tehát a lineáris modellünk helyes.

102. táblázat. A regressziós modell helyességét tesztelő varianciaanalízis táblázat

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	437,990	1	437,990	95,947	,000
	Residual	59,344	13	4,565		
	Total	497,333	14			

a. Predictors: (Constant), A kalász hossza (cm)

b. Dependent Variable: Kalászonkénti szemszám (db)

A regressziós együtthatókat és ezek statisztikai próbáit tartalmazza a 103. táblázat.

103. táblázat. Az regressziós együtthatók és statisztikai próbái

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-15,454	4,808		,007
	A kalász hossza (cm)	5,535	,565	,938	,000

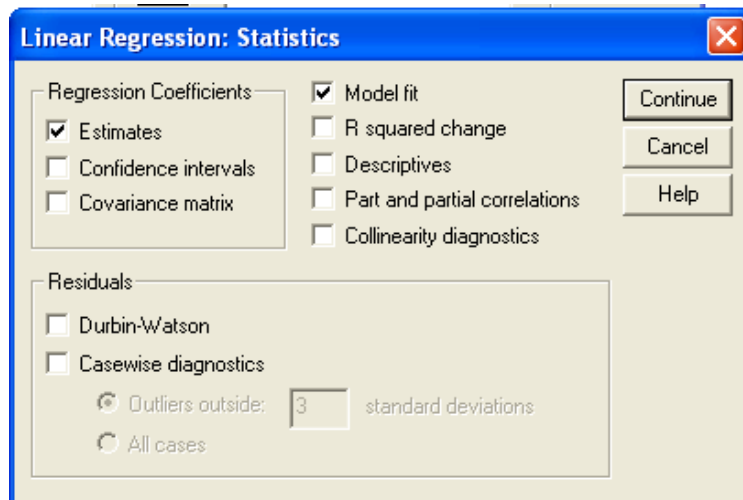
a. Dependent Variable: Kalászonkénti szemszám (db)

A paraméterek becslésére kapott értékeket a táblázat B oszlopában olvassuk le. A konstans értéke: $\beta_0 = -15,454$ és $\beta_1 = 5,535$, így a becült regressziós egyenes:

$$\hat{y} = -15,454 + 5,535 \cdot x$$

A táblázat STD. ERROR feliratú oszlopában az együtthatók becslési hibáját láthatjuk. A BETA oszlop a standardizált együtthatókat adja meg, jelentését akkor fogjuk megérteni, amikor kettő vagy több független változót építünk be a regressziós modellbe. A t oszlop a számított t értékeket tartalmazza. A program mindkét együtthatóra t-próbát végez – korábban részletesen bemutattuk a β_1 együtthatóra végzett próbát –, amely nullhipotézise szerint az együtthatók értékei 0-val egyenlők (azaz nincs szerepük a modellben, rossz a modell). Az utolsó oszlopban a paraméterek tesztelésének az eredménye jelenik meg. Ha az itt szereplő érték 0,05 alatt van, akkor 95%-os megbízhatósági szinten mondhatjuk, hogy a kapott paraméterértékek becslése megbízható, a modellben való szereplésük igazolt.

Nézzük meg, hogy milyen további statisztikákat kérhetünk még a lineáris regresszió elvégzésekor. Ha visszatérünk a lineáris regresszió paneljához a panel alsó részében négy parancsgombot látunk, kattintsunk ezek közül először a STATISTICS... gombra (64. ábra).

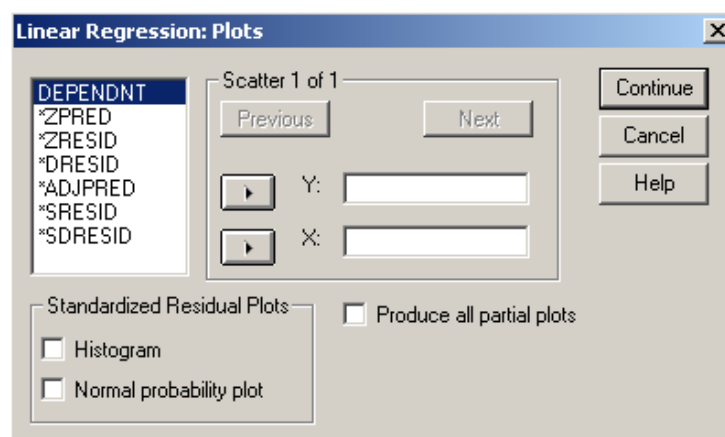


A megjelenő panelban több beállítást végezhetünk el. A regressziós koefficiensek ablakrészben (REGRESSION COEFFICIENTS) az ESTIMATES parancs megjelölésével a regressziós együtthatók becült értékét kapjuk – ez az alapbeállítás.

64. ábra. A LINEAR REGRESSION menü STATISTICS... parancsának beállításai

A CONFIDENCE INTERVALS megjelölésével az együtthatókra vonatkozóan a konfidencia intervallumokat kérhetjük, míg a kovariancia mátrix kiíratására is lehetőségünk van (COVARIANCE MATRIX). A MODEL FIT a modell helyességére vonatkozó jellemzőket számítja ki (R , R^2 , ANOVA) – alapbeállítás ez is. Az R SQUARED CHANGE bejelölése esetén többváltozós regresszió esetén kapjuk meg az R^2 értékét. A RESIDUALS panel-részben a hibatagok statisztikáit kapjuk meg.

A PLOTS... gombra kattintva az X és Y megadásával különböző rajzokat készíthetünk.: pl. DEPENDNT – függő változók, PRED végűek a becült értékek, RESID – a hibatagok. A HISTOGRAM a hibatagok eloszlását vizsgálja, míg a NORMAL PROBABILITY PLOT a hibatagok normális eloszlását mutatja meg (65. ábra).



65. ábra. A Linear Regression menü Plots alpontja

A SAVE gombra kattintva megjelenő ablakban (66. ábra) az eltérések (RESIDUALS) és a modellnek az esetekre való érzékenységének széles körű

elemzésére van lehetőségünk. Az alábbi változókat menthetjük el a regressziós vizsgálat eredményeként.

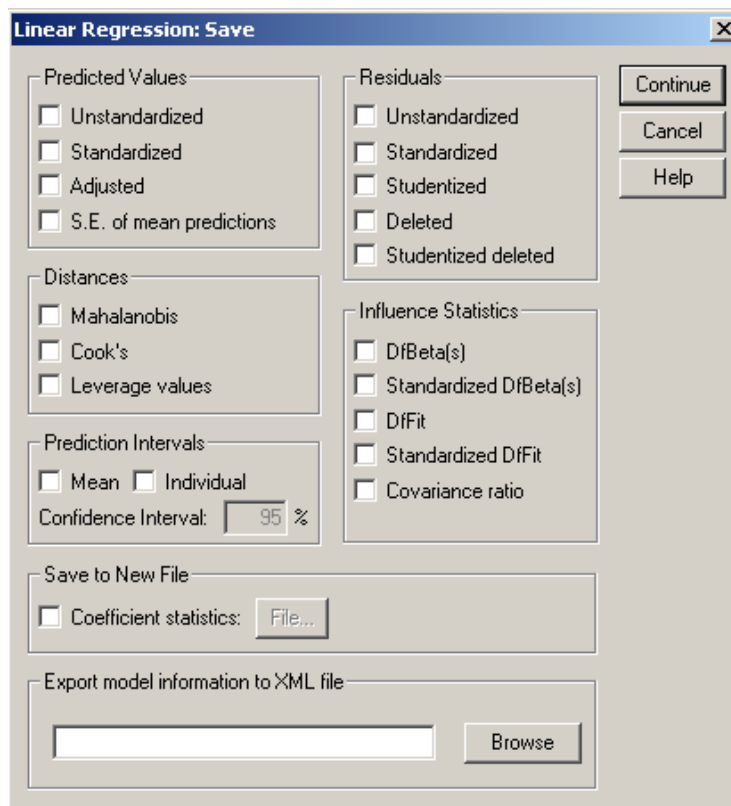
A PREDICTED VALUES részben a becsült értékeket adja meg a program.

UNSTANDARDIZED: a regressziós összefüggésnek az egyes esetekhez kiszámított értéke, amivel a célváltozót közelítjük.

STANDARDIZED: az előző értékek standardizált változata (\hat{x}_{ZPRED}).

ADJUSTED: módosított előrejelző érték, amit úgy kapunk, hogy az i -edik eset becslésénél a regressziót azon $n-1$ esetre számoljuk, ahol az i -edik eset nem szerepel ($\hat{x}_{ADJPRED}$).

S.E. OF MEAN PREDICTIONS: minden esethez számolt várható becslési pontosság.



66. ábra. A Linear Regression menü Save alpontja

A RESIDUALS a hibatagokat menti el:

UNSTANDARDIZED: a tényleges és az előre jelzett értékek különbsége;

STANDARDIZED: az előző hibatag standardizáltja (\hat{x}_{ZRESID}).

STUDENTIZED: az esetektől függően súlyozza az értékeket, az átlagostól jobban eltérő független eseteknél kisebb súllyal, az átlagoshoz közeli helyeken nagyobb súllyal veszi figyelembe a hibatagokat. (\hat{x}_{SRESID})

DELETED: abban az esetben tartalmazza az eltérést, amikor a regressziós sík éppen a vizsgált pont kihagyásával készült (x_{DRESID}).

STUDENTIZED DELETED: az előző eltéréseket súlyozza át attól függően, hogy a bemenő adatok milyen messze esnek az átlagos esettől ($x_{SDRESID}$).

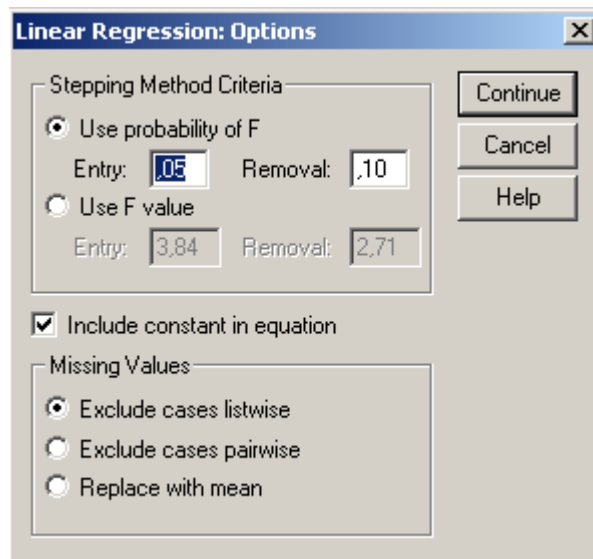
Az SPSS alapértelmezésben nyilvántartja a szélsőséges eseteket, amelyekről listát kérhetünk. Ezt a DISTANCES panelrészben a tehetjük meg.

MAHALANOBIS (távolság): $D_i = \left(\frac{x_i - \bar{x}}{\sigma_x} \right)^2$. Ez megadja minden input-eset vektornak az átlagvektortól vett távolságát, ami igen érzékeny az átlagostól jelentősen eltérő szélsőséges esetek detektálására.

COOK's (távolság): $C_i = \frac{\sum_{j=1}^n (\hat{y}_j^{(i)} - \hat{y}_j)^2}{2 \cdot \sigma^2}$. Ez a távolság az előre jelzett értékekben

fellépő azon négyzetes eltéréseket méri esetenként, ami akkor keletkezne, amikor az adott esetet kihagynánk a regresszióból. (Azok az esetek a legtipikusabbak, amelyeknél ez a távolság nagy.)

LEVERAGE VALUES: az esetek fontosságát méri a regressziós összefüggésben.

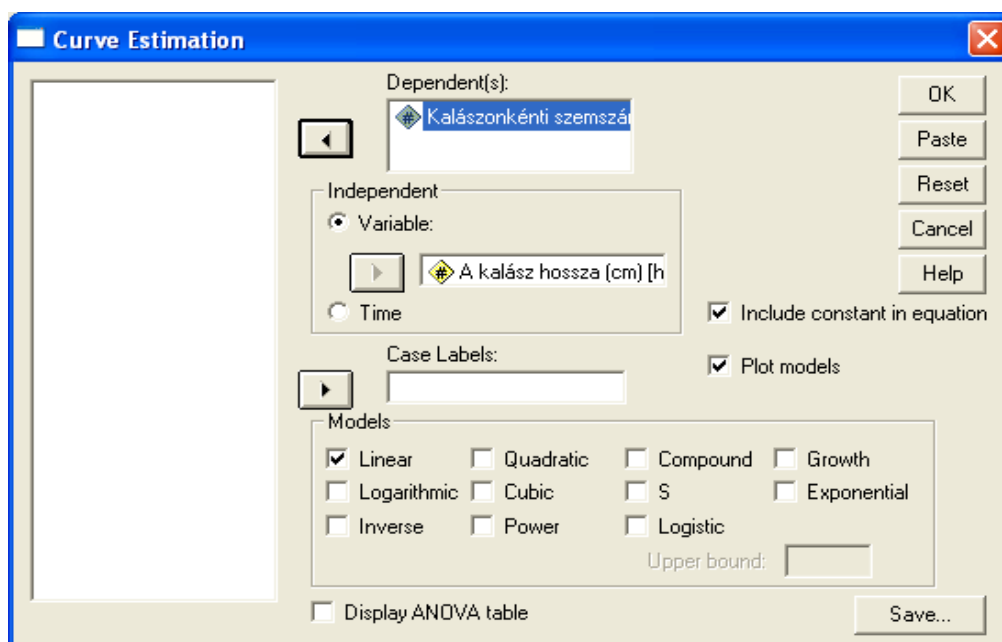


67. ábra. A Linear Regression menü Options alpontja

A PREDICTION INTERVALS részben a konfidencia intervallumok határai jeleníthetők meg tetszőlegesen beállítható szignifikancia szinten – a részletek ismertetésére az elméleti részben már sor került.

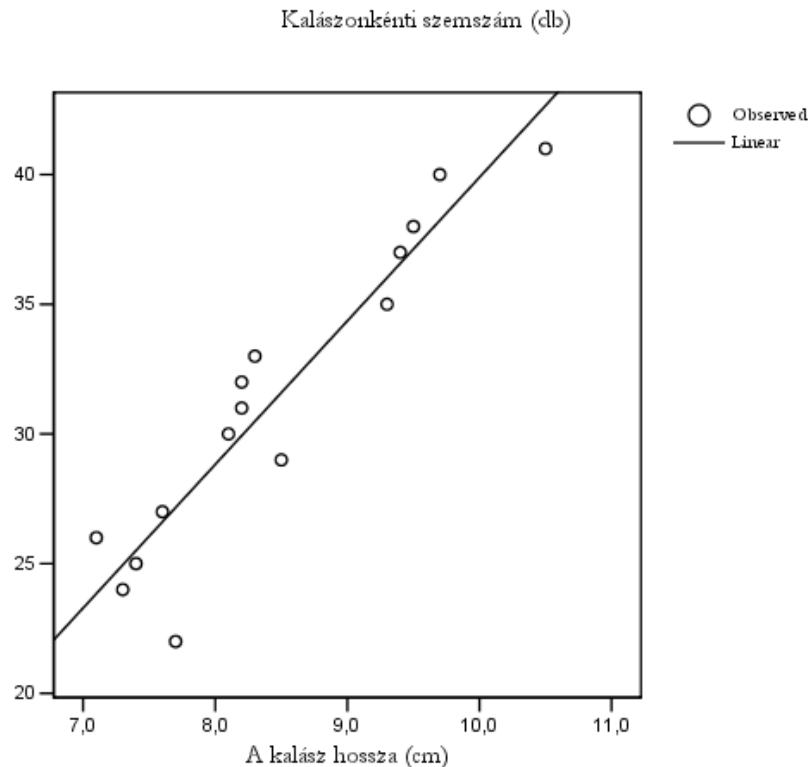
AZ **OPTIONS** beállításainál (67. ábra) a **STEPPING METHOD CRITERIA** a **STEPWISE** módszer feltételeinek beállításra alkalmas (ennek több-változós regressziónál van értelme), az **INCLUDE CONSTANT IN EQUATION** a konstans tag megadására ad lehetőséget (hogyan legyen-e a modellben konstans tag), míg a **MISSING VALUES** a hiányzó tagok kezelésre szolgál.

Ha a kapott pontokra (számított értékekre) egyenest is illeszteni szeretnénk, ezt ebben a menüpontban nem tudjuk megtenni. Kattintsunk az **ANALYZE / REGRESSION / CURVE ESTIMATION...** parancsra (68. ábra), ahol a fent bemutatott számítások mellett ábrát is készíthetünk.



68. ábra. *Analyze/Regression/ Curve Estimation...* menüpont

A bal oldali ablakból válasszuk ki a független változót („A kalász hossza”), amit helyezünk a nyilacska segítségével a **VARIABLE** mezőbe, majd a függő változót („Kalászonkénti szemszám”) a **DEPENDENT(S)** mezőbe tesszük. Az SPSS alapbeállításaként a **MODELS** részben hagyjuk meg a **LINEAR** megjelölést, majd futtassuk le a programot. A program futtatásának eredményeként többek között azokat a táblázatokat is megkapjuk, amelyek a feladat kapcsán már elemzésre kerültek, számunkra azonban most az egyenes illesztése (69. ábra) a lényeges, amit az előző menüpontban a program nem végzett el.



69. ábra. Regressziós egyenes illesztésének eredménye

TÖBBSZÖRÖS LINEÁRIS REGRESSZIÓSZÁMÍTÁS

Egy jelenség vizsgálata során általában az adott jelenséget több tényező befolyásolja, vagyis többnyire nem elegendő a két-változós modell elemzése. Szükség van további olyan magyarázó változók vizsgálatára, amik a jelenség egzaktabb leírását teszik lehetővé. Azokat a kapcsolatokat, amelyeknél az egyik tényezőre több másik tényező is hatással van többszörös kapcsolatoknak nevezzük, a kapcsolatok mennyiségi jellemzőinek, illetve szorosságának vizsgálatát pedig *többszörös korreláció- és regressziószámításnak* hívjuk.

A két-változós regressziós modell problémáit tárgyalva bemutattuk a regressziós modell alapjait, most ezt kiterjesztjük arra az esetre, amikor több tényező befolyásolja egy jelenség alakulását, s mindezt oly módon tesszük meg, hogy felhasználjuk mindazokat a módszereket és elveket, amiket két-változós esetben megismertünk. A két-változós esethez hasonlóan itt is megkülönböztetünk lineáris és nemlineáris típusú összefüggéseket.

A fejezet következő részében a többszörös modellek elemzésének lehetőségeit ismertetjük, ám csak az egyszerűbb lineáris modellt tárgyaljuk. Ugyanakkor megjegyezzük, hogy a kutatások során gyakran nemlineáris regressziót kell alkalmazni.

A standard lineáris regressziós modell

Ha n elemű mintát veszünk, akkor a többszörös lineáris összefüggések általános matematikai egyenlete:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_m \cdot x_{im} + \varepsilon_i,$$

$$(i = 1, 2, \dots, n, m + 1 < n < N),$$

ahol $\beta_1, \beta_2, \dots, \beta_m$ a függő változóra ható tényezőket jelenti; β_0 a függvény konstans tagja, az ε_i pedig a regressziós egyenes hibatagja.

Tekintsük a regressziós modell mátrixalgebrai jelölését:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \cdot & & & & \\ \cdot & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \beta_n \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix},$$

ahol m a magyarázó változók száma és \mathbf{X} első oszlopa mindig egy összegzővektor.

A következőkben az alábbi feltételezésekkel fogunk élni:

A modellben szereplő x_{im} tényezők a független változók, amelyek a feltevésünk szerint lineárisan befolyásolják az \mathbf{Y} függő változó alakulását (azaz $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{y} = \mathbf{X} \cdot \hat{\boldsymbol{\beta}} + \mathbf{e}$ illetve $\hat{\mathbf{y}} = \mathbf{X} \cdot \hat{\boldsymbol{\beta}}$).

Az \mathbf{X} mátrix olyan mennyiségi változókat tartalmaz, amelyek nem valószínűségi változók, értékük nem függ a véletlentől.

Az \mathbf{X} hatása az \mathbf{Y} -ra nem determinisztikus, hanem sztochasztikus, amit kifejez az összefüggésben szereplő $\boldsymbol{\varepsilon}$, amely a független változókon túl hat az \mathbf{Y} -ra, vagyis az \mathbf{Y} változó értéke függ a véletlentől, azaz valószínűségi változó.

A hibatagok nulla várható értékű, konstans varianciájú, korrelálatlan valószínűségi változók, amelyek normális eloszlásúak.

Az ismertett feltételeknek eleget tevő modelleket *standard lineáris regressziós modellnek* hívjuk. A feltételek azonban többnyire nem teljesülnek, az okok közül a 3 legfontosabbat emeljük ki:

Multikollinearitás: a magyarázó változók nem lineárisan függetlenek;

Autokorreláció: a hibatagok lineárisan nem függetlenek;

Heteroszkedaszticitás: a hibatagok szórásnégyzete nem állandó.

A könyv keretein belül nem foglalkozunk azzal, hogy mi lenne annak a következménye, ha a standard lineáris regressziós modell ellentmond valamelyik feltétel teljesülésének. Megmaradunk az alapfokú tárgyalás mellett, és abból indulunk ki, hogy a feltételek ellenőrzései megerősítik a feltételek teljesülését.

Multikollinearitás

A standard lineáris regressziós modell feltételezi, hogy a magyarázó változók egymástól lineárisan függetlenek. Ha azonban valamelyik magyarázó változó kifejezhető a többi tényező lineáris kombinációjaként (azaz függvényyszerű kapcsolatban áll a többi magyarázó változóval) akkor multikollinearitásról beszélünk.

A multikollinearitás kiküszöbölése viszonylag egyszerűen megoldható (lenne), hiszen a lineáris függőség megszüntethető azzal, hogy a vizsgálatba bevont változók közül kizárjuk a lineáris függőségben lévőket. Annak eldöntése azonban, hogy melyik a lineáris függőségben lévő változó, nem könnyű. Mivel a magyarázó változók közötti összefüggések sztochasztikus jellegűek, a jelenség felismerése és a tényezők hatásainak szétválasztása külön számításokat, elemzési módszereket igényel.

Ha a magyarázó változók lineárisan nem függetlenek, akkor az alábbi következményekkel kell számolni:

A becslés és az előrejelzés torzított marad;

A regressziós együtthatók standard hibái nőnek;

A becsléseink bizonytalanokká válnak;

Az egyes magyarázó változók hatásainak elkülönítése nem lehetséges.

A magyarázó változók lineáris függetlenségének tesztelését a többszörös lineáris regresszió elvégzése előtt meg kell vizsgálni. A multikollinearitás meghatározására a multikollinearitás mérőszámai szolgálnak.

A multikollinearitás mérése

Alapelvként abból indulunk ki, hogy a magyarázó változók determinációs együtthatóinak összege, ha megegyezik a többszörös determinációs együttható értékével, akkor nem áll fenn a magyarázó változók között a multikollinearitás, ellenkező esetben igen, mégpedig a különbség nagyságával arányosan.

Ha egy új magyarázó változót vonunk be az elemzésbe, akkor a többszörös determinációs együttható értéke vagy növekszik, vagy nem változik. Ezért a multikollinearitást kiszámíthatjuk, ha minden magyarázó változóra meghatározzuk, hogy a modellbe utolsó változóként bevonva mennyivel növeli a determinációs együttható értékét. Ha a hatásoknak az összege egyenlő a többszörös determinációs együtthatóval, akkor azt mondjuk, hogy a magyarázó változók lineárisan függetlenek – az alapelvvel egybehangzóan. Ellenkező esetben az eredményváltozó négyzetének van olyan része, ami együttesen magyaráz több változót. A multikollinearitás nagyságát pedig ezzel az együttesen magyarázott résszel mérhetjük:

$$M = r_{y.x_1, x_2, \dots, x_m}^2 - \sum_{i=1}^m \left(r_{y.x_1, x_2, \dots, x_m}^2 - r_{y.x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m}^2 \right).$$

Az M értéke alapján azt mondhatjuk, hogy minél kisebb az eltérés közte és a többszörös determinációs együttható között, annál jelentősebb a multikollinearitás, nullához közeli értéke a multikollinearitás hiányát mutatja.

Autokorreláció

Autokorrelációról akkor beszélünk, ha a hibatagok lineárisan nem függetlenek. Az autokorreláció különböző rendű lehet, attól függően, hogy a hibatag i -edik értéke melyik értékkel van kapcsolatban. Ha a hibatag i -edik értéke közvetlenül az előtte lévő értékkel áll korrelációs kapcsolatban, akkor *elsőrendű autokorrelációról* beszélünk. Az elsőrendű autokorreláció modellje:

$$\varepsilon_i = \rho \cdot \varepsilon_{i-1} + \lambda_i,$$

ahol ρ az autokorrelációs együttható.

Az elsőrendű autokorreláció tesztelése

Az elsőrendű lineáris autokorreláció tesztelésére a *Durbin-Watson-féle próbát* alkalmazzuk, a próba a regressziós reziduumokra épít és próbafüggvényét azokból állítja elő:

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2},$$

ahol e_i a legkisebb négyzetek módszerével kapott reziduumokat jelenti (ezt a hibatagok becslésének tekintjük). A d statisztika értékei 0 és 4 közé esnek.

A ρ lineáris autokorrelációs együttható értékét a

$$\hat{\rho} = \frac{\sum_{i=2}^n e_i \cdot e_{i-1}}{\sqrt{\sum_{i=2}^n e_i^2} \cdot \sqrt{\sum_{i=2}^n e_{i-1}^2}}$$

képlet alapján becsüljük.

Mivel $\sum_{i=1}^n e_i^2 \approx \sum_{i=2}^n e_i^2 \approx \sum_{i=2}^n e_{i-1}^2$, így a Durbin-Watson-féle próbafüggvény $d \approx 2 \cdot (1 - \hat{\rho})$ alakra hozható.

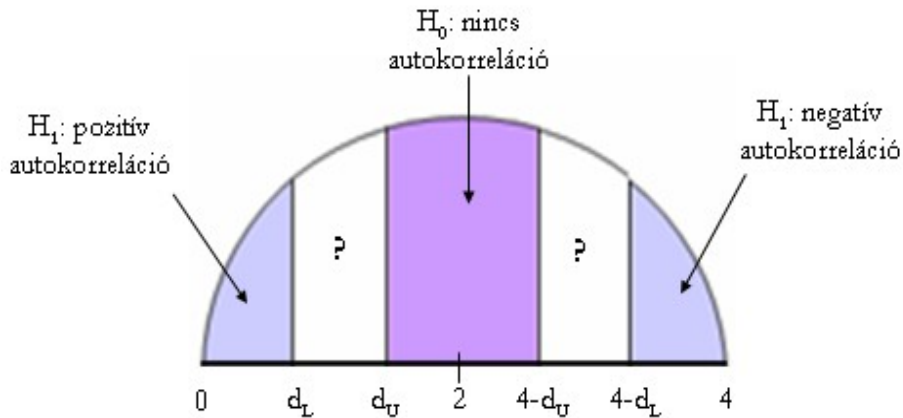
A próba nullhipotézise az elsőrendű autokorreláció hiányát fogalmazza meg, azaz e szerint $\rho=0$. Ha a próbafüggvény értéke 2-nél nagyobb, akkor az alternatív hipotézisünk a negatív autokorreláció ($H_1 : \rho < 0$), amennyiben 2-nél kisebb, akkor a pozitív autokorreláció ($H_1 : \rho > 0$).

Az elsőrendű lineáris autokorreláció tesztelésekor a 104. táblázat relációi alapján döntünk. A kritikus értékek (d_L és d_U) meghatározása a megfigyelések számának és a magyarázó változók számának függvényében a „Durbin-Watson-féle próba kritikus értékei” táblázatból kereshetők ki.

104. táblázat. A Durbin-Watson-féle próba döntési táblája

Alternatív hipotézis	$H_0 : \rho = 0$		
	Elfogadjuk	Elvetjük	Nincs döntés
$\rho > 0$	$d > d_U$	$d < d_L$	$d_L \leq d \leq d_U$
$\rho < 0$	$d < 4 - d_U$	$d > 4 - d_L$	$4 - d_L \leq d \leq 4 - d_U$

A döntésszabály szemléltetésére tekintsük a 70. ábrat.

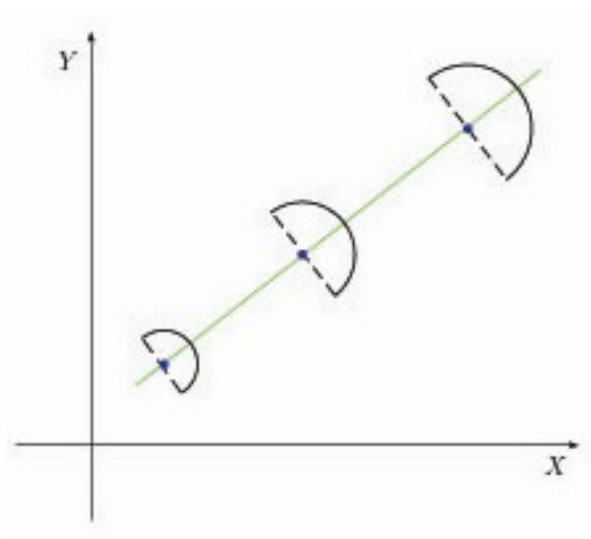


70. ábra. A Durbin-Watson teszt döntési szabálya

Amennyiben a teszt alapján nem tudunk döntést hozni, vagyis a próbafüggvény értéke a semleges zónák valamelyikébe esik, akkor több lehetőséggel is élhetünk:

- A modell paramétereinek a becslését újra el kell végezni, de nagyobb minta alapján.
- Meg kell változtatni a szignifikancia-szintet úgy, hogy döntési helyzetbe kerüljünk.
- Más próbafüggvényt kell alkalmazni.

Heteroszkedaszticitás



71. ábra. A heteroszkedaszticitás interpretációja

A keresztmetszeti vizsgálatoknál gyakori probléma, hogy a hibatagok varianciái nem állandóak (71. ábra), pedig standard lineáris regressziós modell esetében ez követelmény.

Azt, hogy a varianciák hibatagjainak az állandósága nem áll fenn okozhatja az, hogy a hibatag nagysága függ valamelyik változótól.

A heteroszkedaszticitás tesztelésénél ellenőrizni kell, hogy milyen szoros a kapcsolat az egyes változók és a hibatagok abszolút értékei között. A használandó próbafüggvény:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} .$$

Ki kell számítani külön az egyes magyarázó változóknak, illetve a becsült eredményváltozóknak a reziduumok abszolút értékeivel való szorosságát jellemző lineáris korrelációs együtthatót, amelyek közül a legnagyobb abszolút értékű kerül tesztelésre. Ha a nullhipotézist elvetjük, akkor a modell heteroszkedasztikusnak tekinthető.

A többszörös lineáris regressziószámítás lépései

A többszörös regresszióelemzés folyamata hasonlít a két-változós regresszióelemzés folyamatához.

A regressziós modell illeszkedésének vizsgálata

A regressziós modell illeszkedésének vizsgálatához definiáljuk az alábbi eltérés-négyzetösszegeket:

$$\sum_{i=1}^n (y_i - \bar{y})^2 := SST, \quad \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 := SSR \quad \text{és a} \quad SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 .$$

Ha a modell tartalmaz konstans tagot (vagyis $\beta_0 \neq 0$), akkor: $SST = SSR + SSE$. A

lineáris determinációs együttható, ami megadható az $r^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

képlettel (is), felírható a következő alakban:

$$r^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

Egy modell illeszkedésének mértékét az határozza meg, hogy a teljes eltérés-négyzetösszegének mekkora részét teszi ki a regresszió által magyarázott és a hibataggal kapcsolatos négyzetösszeg.

A modell illeszkedésének jóságát a varianciaanalízis segítségével teszteljük (globális F -próba). Ez egy olyan hipotézisvizsgálat, amelynél a nullhipotézis: $\beta_1 = \beta_2 = \dots = \beta_m = 0$, azaz a β_j meredekségek mind egyenlők zérussal (csak a konstans tag értéke különbözik szignifikánsan nullától). Az alternatív hipotézis szerint: $\beta_j \neq 0$, valamelyik j -re, $j = 1, 2, \dots, m$. A nullhipotézis elfogadása azt jelenti, hogy az adott változókkal felírt regressziós modell nem alkalmas az \mathcal{Y} becslésére.

Az F próba:

$$F = \frac{\frac{SSR}{m}}{\frac{SSE}{n-m-1}} = \frac{MSR}{MSE}.$$

A varianciaanalízis táblázatból olvasható le a modell illeszkedésének helyessége, ebben a táblázatban a tapasztalati F -értékek vannak összevetve a megfelelő elméleti értékekkel. A varianciaanalízis egyoldalú próba, ami azt jelenti, hogyha a tapasztalati F érték kisebb az elméleti értéknél, akkor a nullhipotézist elfogadjuk (az adott szignifikancia szint mellett), vagyis ebben az esetben a vizsgált modell nem alkalmas a megfigyelt jelenség elemzésére. A nullhipotézis elvetése azonban nem jelenti automatikusan a modell illeszkedésének jóságát.

Az ANOVA táblázat felépítését a 105. táblázatban mutatjuk be. A regressziós modellben a teljes eltérés-négyzetösszeg két részre bontható: (1) regressziós hatásra és (2) hibahatásra. Azzal, hogy az együtthatók legkisebb négyzetes becslése során az SSE-t minimalizáljuk, az SSR-t maximalizáljuk. Az átlagértékek aránya – az F -hányados – „nagy” lesz, ha van lineáris összefüggés a függő és független változók között.

105. táblázat. Az ANOVA táblázat

A szóródás oka	Az eltérések négyzetösszege	Szabadsági fok	Szórásnégyzetek becslése	F
Regresszió	SSR	m	MSR	$\frac{MSR}{MSE}$
Hiba	SSE	$n - m - 1$	MSE	
Összesen	SST	$n - 1$		

A paraméterek tesztelése

Fentebb az egész modell illeszkedését vizsgáltuk, most egyetlen magyarázó változó fontosságát, magyarázó erejét teszteljük. Gyakorlatilag ez azt jelenti, hogy minden becsült paraméterértékre végzünk egy hipotézisvizsgálatot, amelynek a nullhipotézise szerint: $H_0: \beta_j = 0, j = 1, 2, \dots, m$; míg a kétoldali alternatív hipotézis: $H_1: \beta_j \neq 0, j = 1, 2, \dots, m$.

A tesztelésre az alábbi próbafüggvényt használjuk:

$$F = \frac{\hat{\beta}_j^2}{\text{Var}(\hat{\beta}_j)},$$

ahol $\text{Var}(\hat{\beta}_j)$ a $\text{Var}(\hat{\beta}) = \frac{\mathbf{e}^T \cdot \mathbf{e}}{n - m - 1} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} = s_e^2 \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1}$ variancia-kovariancia mátrix főátlójában lévő j -edik elem. (Az s_e^2 az ún. reziduális szórásnégyzet, ami torzítatlan becslése a σ^2 -nek.) Ez a statisztika $f_1 = 1, f_2 = n - m - 1$ szabadsági fokú F -eloszlást követ.

Ha t -próbát végzünk, akkor a próbafüggvény alakja:

$$t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}},$$

ahol $s_{\hat{\beta}_j}$ a fentebb definiált variancia négyzetgyöke. Ha az empirikus t -érték abszolút értéke kisebb, mint az elméleti t -érték, akkor a nullhipotézist elfogadjuk, ami azt jelenti, hogy a vizsgált magyarázó változó nem befolyásolja az eredményváltozót. Ebben az esetben nem érdemes szerepeltetni a modellben a magyarázó változót.

A becsült paraméterek jelentése

Miután elvégeztük a modellt, a paraméterek vizsgálatát – és az megfelelő volt –, értelmezni kell a kapott $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m$ becsült regressziós paramétereket. A $\hat{\beta}_j$ ($j=1,2,\dots,m$) azt mutatja meg, hogy az x_j magyarázó változó egységnyi növekedése az eredményváltozó mekkora változásával (átlagos) jár együtt, ha a többi magyarázó változó értéke nem változik.

A reziduumok vizsgálata

A reziduumok pontdiagramjainál a reziduumokat az \hat{y}_i becsült értékekkel, vagy magyarázó változókkal (vagy az idővel) együtt szoktuk ábrázolni, ezek a pontdiagramok jelzik a feltételek teljesülését és a regressziós modell illeszkedését.

A reziduumokra vonatkozó feltételek közül először a normalitást vizsgáljuk. A hibatenyező normális eloszlásának ellenőrzésére több módszert ismerünk. A grafikus teszteket elsősorban vizuális eszköznek tekintjük az egyes hipotézisek vizsgálatára, a több létező grafikus teszt közül megemlítjük a hisztogramot, és az illeszkedésre szolgáló ún. P-P diagramot, amelyek a leggyakrabban alkalmazott grafikus eszközök. További bizonyítékokat kaphatunk az eloszlás jellegéről, ha megvizsgáljuk, hogy a reziduumok hány százaléka esik a $\pm 1SE$, vagy $\pm 2SE$ intervallumba. A százalékok összehasonlíthatók azzal, ami a normális eloszlás mellett várható (68% ill. 95%). Az egymintás Kolmogorov-Smirnov próbával azonban pontosabb értékelést kaphatunk.

A hibatenyező konstans varianciájára vonatkozó feltevés tesztelhető, ha a reziduumokat a függő változó becsült \hat{y}_i értékeivel együtt ábrázoljuk. Ha ugyanis a ponthalmazban szereplő pontok elrendeződése nem véletlenszerű, akkor a hibatenyező varianciája nem konstans.

Két független változós lineáris regresszióelemzés

A regresszió paramétereinek meghatározása kézi számítással

A következő példában azt fogjuk megvizsgálni, hogy egy mennyiségi változó hogyan függ másik két mennyiségi változótól. A probléma matematikai egyenlete:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2,$$

ahol \hat{y} a függő változó becsült értéke, x_1, x_2 a független változók, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ az egyenlet becsült paraméterei.

106. táblázat. Két független változós többszörös regresszióanalízis adatmátrixa

$P_2O_5 (x_1)$	$H_{\%} (x_2)$	AK (y)
5,4	2,9	23,0
4,0	2,9	26,9
7,0	1,9	19,0
7,8	4,4	19,4
8,0	2,5	21,0
10,3	3,1	31,0
16,1	3,6	31,8
13,1	2,5	28,0
5,0	2,5	15,0
9,6	2,3	28,0
5,0	2,5	14,0
12,4	3,6	31,0
10,2	2,1	28,0
20,7	2,5	35,2
15,0	2,5	28,0
10,0	2,5	22,0
2,6	2,5	20,8
6,3	2,9	14,3

Forrás: SVÁB JÁNOS (1981), 332.o.

Az alábbi példában azt szeretnénk meghatározni, hogy egy gazdaság napraforgó táblái esetében (106. táblázat) a táblák aranykorona értéke (AK) hogyan függ a talaj foszfor tartalmától (P_2O_5), humuszszázalékától ($H_{\%}$). A példa 1976-os adatokat tartalmaz.

Mielőtt elvégezzük a modell paramétereinek a becslését, nézzük meg, hogy teljesül-e a standard lineáris regressziós modell feltételrendszere.

Elsőként a magyarázó változók lineáris függetlenségét teszteljük. Számítsuk ki a korrelációs mátrixot (ezt az SPSS-el végezzük el), amit a 107. táblázat tartalmaz.

107. táblázat. Az SPSS által készített korrelációs mátrix

Correlations

		foszfor_x1	humusz_x2	aranykorona ertek_y
foszfor_x1	Pearson Correlation	1	,091	,764**
	Sig. (2-tailed)		,718	,000
	N	18	18	18
humusz_x2	Pearson Correlation	,091	1	,122
	Sig. (2-tailed)	,718		,629
	N	18	18	18
aranykoronaertek_y	Pearson Correlation	,764**	,122	1
	Sig. (2-tailed)	,000	,629	
	N	18	18	18

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

A fenti táblázatból felírva a korrelációs mátrixot:

$$\mathbf{R} = \begin{pmatrix} 1 & 0,091 & 0,764 \\ 0,091 & 1 & 0,122 \\ 0,764 & 0,122 & 1 \end{pmatrix}.$$

A szimmetria miatt a mátrixnak csak az alsó háromszögét tekintjük. Az egyes értelmezések a két változós korrelációnál tanultak alapján egyszerű: például a 0,764 azt mutatja, hogy a talaj foszfor tartalma pozitív és közepesnél erősebb kapcsolatban van a talaj aranykorona értékével.

Mivel a mátrixban a két független változó közötti korrelációs érték (0,091) nullához közeli, feltételezhető, hogy a magyarázó változók egymástól függetlenek. Ellenőrizzük a multikollinearitást, amihez helyettesítsünk be a

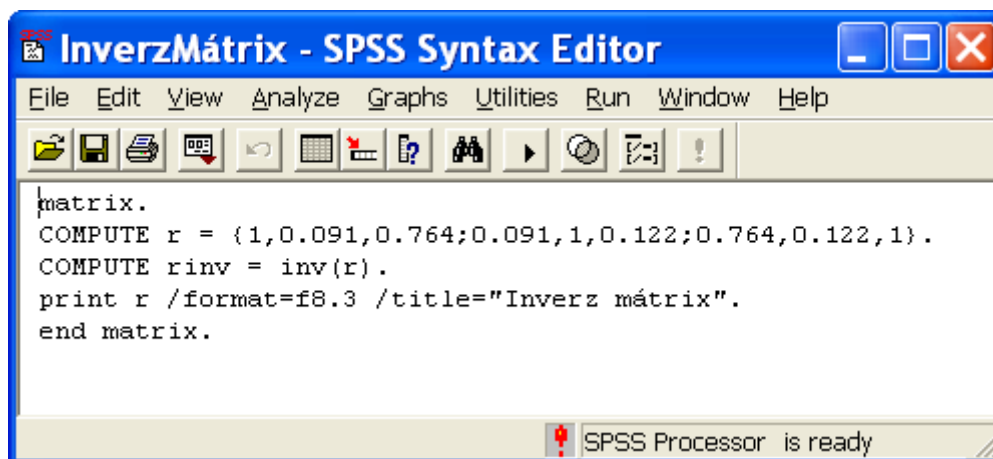
$$M = r^2_{y.x_1, x_2, \dots, x_m} - \left[\sum_{i=1}^m r^2_{y.x_1, x_2, \dots, x_m} - r^2_{y.x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m} \right]$$

képletbe. A megfelelő páronkénti korrelációs együtthatók: $r_{yx_1} = 0,764$; $r_{yx_2} = 0,122$. Mivel három dimenziós a modell, ezért meg kell adni a többszörös determinációs együtthatót, amit az $r^2_{y.x_1, x_2, \dots, x_m} = 1 - \frac{1}{\mathbf{R}_{yy}^{-1}}$ képlet szerint fogunk kiszámolni az \mathbf{R}^{-1} (inverz) mátrix segítségével. (Az inverz mátrix olyan mátrix, amelyre teljesül a következő összefüggés: $\mathbf{R} \cdot \mathbf{R}^{-1} = \mathbf{R}^{-1} \cdot \mathbf{R} = \mathbf{E}$).

A korrelációs mátrix inverze:

$$\mathbf{R}^{-1} = \begin{pmatrix} 2,402 & 0,005 & -1,836 \\ 0,005 & 1,015 & -0,128 \\ -1,836 & -0,128 & 2,418 \end{pmatrix}.$$

Az inverz meghatározását az SPSS mátrix utasításaival végeztük a MATRIX – END MATRIX eljárás segítségével. A program Syntax Editor ablakát nyissuk meg, és írjuk be az eredeti korrelációs mátrixot (72. ábra). A mátrix sorelemeit vesszővel, az oszlopait pontosvesszővel kell elválasztani. A mátrixot kapcsos zárójelek között kell definiálni, ezt a Compute paranccsal tehetjük meg. Szintén ezzel a paranccsal számítottuk ki az inverz mátrixot is. A beépített függvények közül válasszuk az inv(mátrix)-t, és a mátrix helyére írjuk be az eredeti korrelációs mátrixot, esetünkben r-t. A print-tel kezdődő sor csak az inverz mátrix kiíratásának formáját szabályozza. A cím Inverz mátrix, és minden szám nyolc karakter hosszúságban, három tizedes pontossággal fog megjelenni. A további számításokat is az SPSS-vel végeztük el, ahol az inverzen kívül a transzponálás és mátrixszorzás függvényeit használtuk fel.



72. ábra. Az SPSS utasítástervező ablaka

A többszörös determinációs együttható értéke:

$$r_{y.x_1, x_2, \dots, x_m}^2 = 1 - \frac{1}{\mathbf{R}_{yy}^{-1}} = 1 - \frac{1}{2,418} = 0,586.$$

Ez azt jelenti, hogy az eredményváltozó szórásnégyzetének 58,6%-át tudjuk megmagyarázni az x_1 és x_2 változókkal.

Most már a megfelelő adatokat helyettesítsük be a multikollinearitás képletébe:

$$M = r_{y.x_1, x_2, \dots, x_m}^2 - \left[\sum_{i=1}^m r_{y.x_1, x_2, \dots, x_m}^2 - r_{y.x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m}^2 \right]$$

$$= 0,586 - \left[(0,586 - (0,764)^2) + (0,586 - (0,122)^2) \right] \cong 0,013.$$

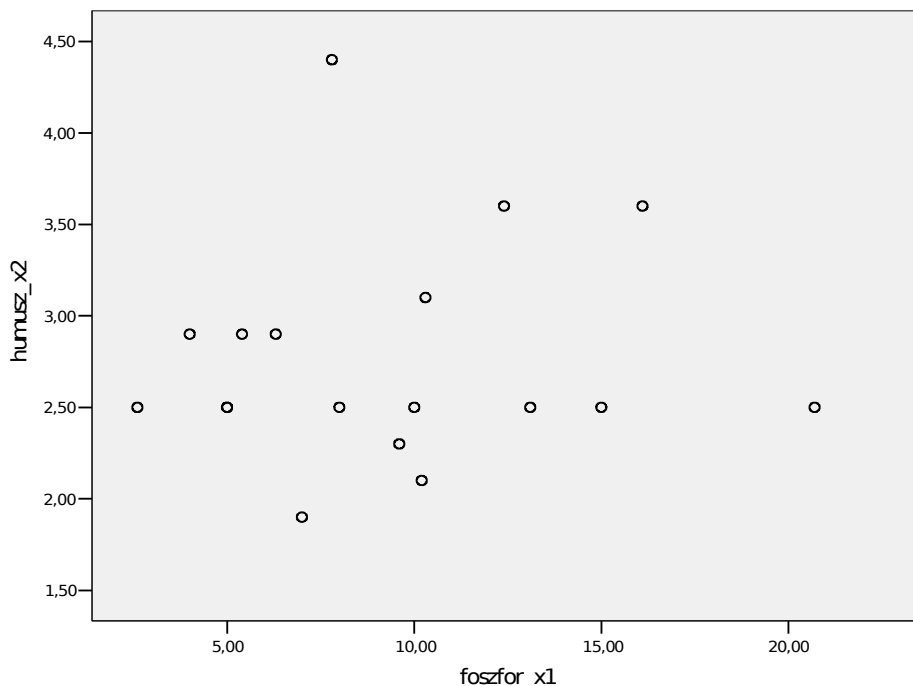
Az M értéke alapján azt mondhatjuk, hogy nullához közeli értéke a multikollinearitás hiányát mutatja.

A két magyarázó változó kapcsolatának szorosságát a $t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$ próbafüggvénnyel teszteljük:

$$t = \frac{0,091 \cdot \sqrt{16}}{\sqrt{1-0,0083}} \cong 0,365.$$

Kétoldali próba esetén ($\alpha = 0,05$ és $df = 16$) az elméleti t-érték 2,11 (Student-féle t-eloszlású változó eloszlásának kvantilisértékei táblázat kétoldali próbákhoz). Az empirikus $t = 0,365$ kisebb ennél az értéknél, ezért a nullhipotézist 5%-os szignifikanciaszinten megtartjuk, ami a magyarázó változók lineáris függetlenségét támasztja alá.

Grafikusan is ábrázolhatjuk a két magyarázó változót (73. ábra). A kapott pontok elhelyezkedése alapján azt mondhatjuk, hogy a pontok elrendeződése véletlenszerű. A grafikus megjelenítés alapján is ugyanarra a következtetésre jutottuk a magyarázó változók esetében, mint azt a számolásokkal is kaptuk, vagyis nincs multikollinearitás.



73. ábra. A magyarázó változók pontdiagramja

A multikollinearitás tesztelése után az autokorrelációra vonatkozó nullhipotézist vizsgáljuk meg, amelyhez a reziduumokra van szükség. Kiindulásként felírtuk a több-változós lineáris regressziós egyenletet a következő alakban:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \dots + \hat{\beta}_m \cdot x_{im} + e_i .$$

A fenti kifejezés felírható egyszerűbben mátrixalgebrai jelöléssel:

$$\mathbf{y} = \mathbf{X} \cdot \hat{\boldsymbol{\beta}} + \mathbf{e} .$$

Helyettesítsük be az adatokat a mátrixegyenletbe:

$$\begin{bmatrix} 23 \\ 26,9 \\ \cdot \\ \cdot \\ \cdot \\ 14,3 \end{bmatrix} = \begin{bmatrix} 1 & 5,4 & 2,9 \\ 1 & 4,0 & 2,9 \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ 1 & 6,3 & 2,9 \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ e_2 \end{bmatrix}$$

Az ismeretlen $\hat{\boldsymbol{\beta}}$ oszlopvektorának a meghatározásához használjuk fel a $\hat{\boldsymbol{\beta}} = (\mathbf{x}^T \cdot \mathbf{x})^{-1} \cdot \mathbf{x}^T \cdot \mathbf{y}$ képletet, ahol \mathbf{x}^T az \mathbf{X} mátrix transzponáltját jelenti.

Először meghatározzuk az $\mathbf{x}^T \cdot \mathbf{x}$ kifejezés értékét:

$$\begin{aligned} \mathbf{x}^T \cdot \mathbf{x} &= \begin{bmatrix} 1 & 1 & \cdot & \cdot & \cdot & 1 \\ 5,4 & 4,0 & & & & 6,3 \\ 2,9 & 2,9 & & & & 2,9 \end{bmatrix} \cdot \begin{bmatrix} 1 & 5,4 & 2,9 \\ 1 & 4,0 & 2,9 \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ 1 & 6,3 & 2,9 \end{bmatrix} = \\ &= \begin{bmatrix} 18 & 168,5 & 49,7 \\ 168,5 & 1955,81 & 469,68 \\ 49,7 & 469,68 & 143,43 \end{bmatrix} . \end{aligned}$$

Vegyük az $\mathbf{x}^T \cdot \mathbf{x}$ mátrixszorzat inverzét:

$$(\mathbf{x}^T \cdot \mathbf{x})^{-1} = \begin{bmatrix} 1,43 & -0,02 & -0,43 \\ -0,02 & 0,003 & -0,002 \\ -0,43 & -0,002 & 0,163 \end{bmatrix} .$$

Képezzük az $\mathbf{x}^T \cdot \mathbf{y}$ szorzatot:

$$\mathbf{X}^T \cdot \mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & \dots & 1 \\ 5,4 & 4,0 & & & 6,3 \\ 2,9 & 2,9 & & & 2,9 \end{bmatrix} \cdot \begin{bmatrix} 23 \\ 26,9 \\ \dots \\ 14,3 \end{bmatrix} = \begin{bmatrix} 436,4 \\ 4478,81 \\ 1213,02 \end{bmatrix}$$

Most már könnyen megkapjuk $\hat{\beta}$ -t, ha elvégezzük a $(\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$ mátrixszorzást:

$$\hat{\beta} = \begin{bmatrix} 1,43 & -0,02 & -0,43 \\ -0,02 & 0,003 & 0,002 \\ -0,43 & -0,002 & 0,163 \end{bmatrix} \cdot \begin{bmatrix} 436,4 \\ 4478,81 \\ 1213,02 \end{bmatrix} = \begin{bmatrix} 13,0165 \\ 1,0335 \\ 0,5627 \end{bmatrix}$$

A fenti mátrixműveletek eredményeit az SPSS segítségével gyorsan megkaphatjuk. Nyissuk meg a *Syntax Editor*t és írjuk be az alábbi utasításokat, majd kattintsunk a *Run* gombra:

```
MATRIX.
COMPUTE x = {1,5.4,2.9;...;1,6.3,2.9}.
COMPUTE y = {23;26.9;...;14.3}.
COMPUTE BETA = INV(T(x)*x)*T(x)*y.
PRINT BETA /FORMAT=F8.4 /TITLE="Együtthatók".
END MATRIX.
```

A becsült $\hat{\beta}$ paraméterek oszlopvektora segítségével a táblák aranykorona értéke (az $\hat{y} = \mathbf{x} \cdot \hat{\beta}$ képletbe helyettesítve).

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 & 5,4 & 2,9 \\ 1 & 4,0 & 2,9 \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ 1 & 6,3 & 2,9 \end{bmatrix} \cdot \begin{bmatrix} 13,0165 \\ 1,0335 \\ 0,5627 \end{bmatrix} = \begin{bmatrix} 20,229 \\ 18,782 \\ \dots \\ \dots \\ 21,159 \end{bmatrix}$$

Az autokorreláció teszteléséhez használjuk az alábbi munkatáblázatot (108. táblázat).

108. táblázat. A regressziós függvény becsült értékei és a hibatagok

	y_i	\hat{y}_i	e_i	e_i^2	e_{i-1}	$(e_i - e_{i-1})^2$	$ e_i $
1	23	20,2 3	2,77	7,68	-	7,6729	2,77
2	26,9	18,7 8	8,12	65,90	2,77	28,622 5	8,12
3	19	21,3 2	-2,32	5,38	8,12	108,99 36	2,32
4	19,4	23,5 5	-4,15	17,25	-2,32	3,3489	4,15
5	21	22,6 9	-1,69	2,86	-4,15	6,0516	1,69
6	31	25,4 1	5,59	31,30	-1,69	52,998 4	5,59
7	31,8	31,6 8	0,12	0,01	5,59	29,920 9	0,12
8	28	27,9 6	0,04	0,00	0,12	0,0064	0,04
9	15	19,5 9	-4,59	21,07	0,04	21,436 9	4,59
10	28	24,2 3	3,77	14,20	-4,59	69,889 6	3,77
11	14	19,5 9	-5,59	31,25	3,77	87,609 6	5,59
12	31	27,8 6	3,14	9,88	-5,59	76,212 9	3,14
13	28	24,7 4	3,26	10,63	3,14	0,0144	3,26
14	35,2	35,8 2	-0,62	0,38	3,26	15,054 4	0,62
15	28	29,9 3	-1,93	3,71	-0,62	1,7161	1,93
16	22	24,7 6	-2,76	7,61	-1,93	0,6889	2,76
17	20,8	17,1 1	3,69	13,61	-2,76	41,602 5	3,69
18	14,3	21,1 6	-6,86	47,05	3,69	111,30 25	6,86
Σ	436, 4	436, 4	0	289,7 7	6,85	663,14 3	--

A 108. táblázat adatait felhasználva helyettesítsünk be a $d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$

képletbe, így a próbafüggvény értéke: $d = 2,288$. Az autokorreláció becslése: $\hat{\rho} = 1 - \frac{d}{2} = 1 - \frac{2,288}{2} = -0,144$. A kapott eredmény alapján alternatív hipotézisünk a negatív autokorreláció (mivel a próbafüggvény értéke kisebb 2-nél). A „Durbin-Watson-féle próba kritikus értékei” táblázat alapján $\alpha = 0,05$ szignifikancia szint mellett: $d_U = 1,545$. Mivel $d = 2,288 < 4 - d_U = 2,457$, ezért a Durbin-Watson-féle próba nullhipotézisét elfogadjuk, azaz a hibatagok lineárisan nem autokorrelálnak.

A feltételek ellenőrzése között még a heteroszkedaszticitást is tesztelnünk kell, ami a reziduumok abszolút értékei és a változók értékei közötti lineáris korreláció kiszámításának segítségével történik (a reziduumok abszolút értékeit már a . táblázatban meghatároztuk). A korrelációs mátrix meghatározását az SPSS-ben végezzük el, amelynek eredménye a 109. táblázat.

109. táblázat. A korrelációs mátrix a reziduumok abszolútértékeivel kiegészítve

Correlations					
		foszfor_x1	humusz_x2	aranykoronaeretek_y	a reziduumok abszolútértéke
foszfor_x1	Pearson Correlation	1	,091	,764**	-,677**
	Sig. (2-tailed)		,718	,000	,002
	N	18	18	18	18
humusz_x2	Pearson Correlation	,091	1	,122	,112
	Sig. (2-tailed)	,718		,629	,658
	N	18	18	18	18
aranykoronaeretek_y	Pearson Correlation	,764**	,122	1	-,439
	Sig. (2-tailed)	,000	,629		,068
	N	18	18	18	18
a reziduumok abszolútértéke	Pearson Correlation	-,677**	,112	-,439	1
	Sig. (2-tailed)	,002	,658	,068	
	N	18	18	18	18

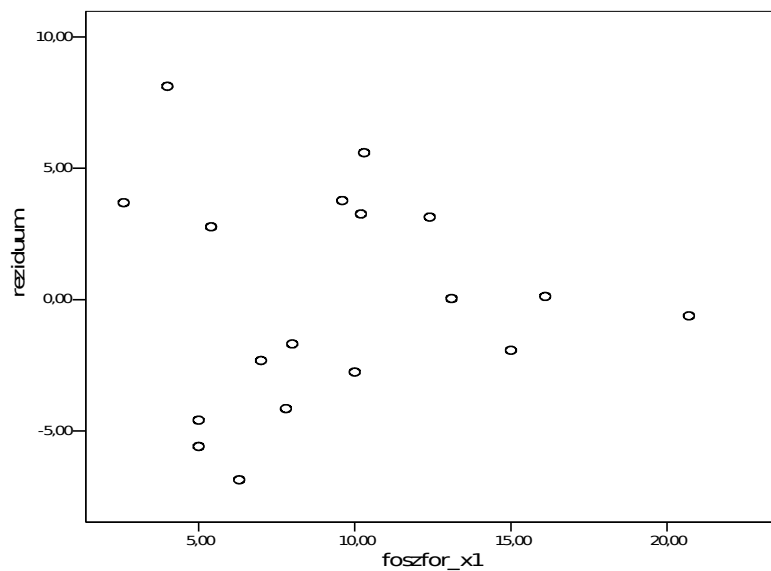
**Correlation is significant at the 0.01 level (2-tailed).

A táblázat alapján: $r_{e|\hat{y}} = -0,439$; $r_{e|x_1} = -0,677$; $r_{e|x_2} = -0,112$. A legnagyobb abszolút értékű az $r_{e|x_1}$, az kell ellenőrizni, hogy ez szignifikánsan különbözik-e nullától. A teszteléshez a t -próbafüggvényt használjuk:

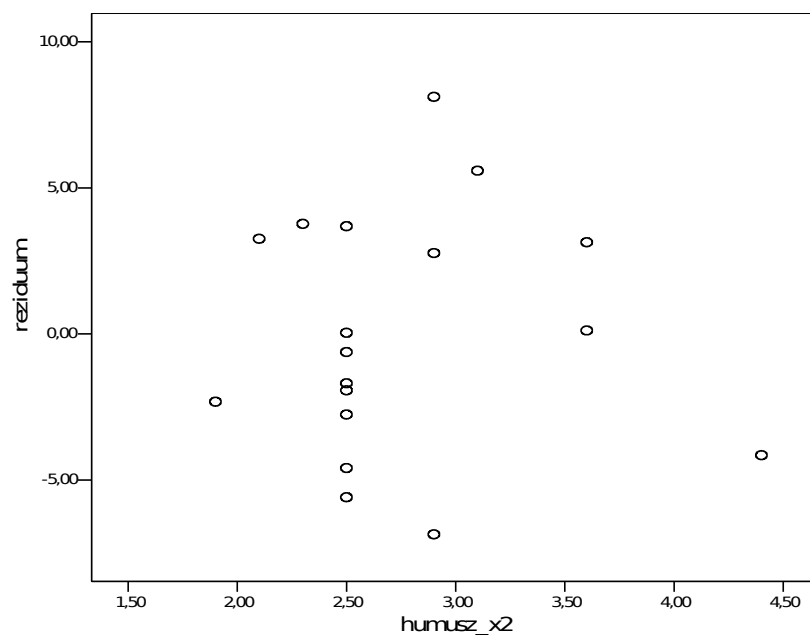
$$t = \frac{0,677 \cdot \sqrt{16}}{\sqrt{1 - 0,4583}} \cong 3,679 .$$

Az elméleti t -érték 2,12 ($\alpha=0,05$ és $df=16$), az empirikus t -érték ($t=3,679$) nagyobb ennél az értéknél, ezért a nullhipotézist 5%-os szignifikanciaszinten elvetjük.

Grafikusan is ellenőrizhetjük a heteroszkedaszticitást, ha ábrázoljuk az egyes változók és a reziduumok közötti kapcsolatokat (74. ábra, 75. ábra).



74. ábra. A talaj foszfor tartalma és a reziduum pontdiagramja



75. ábra. A talaj humusz tartalma (%) és a reziduum pontdiagramja

A lineáris regressziós függvény a kapott eredmény alapján:

$$\hat{y} = 13,0165 + 1,0335 \cdot x_1 + 0,5627 \cdot x_2.$$

A többszörös determinációs együttható $r^2_{y,x_1,x_2} = 0,586$ értéke alapján nem tudjuk objektívan megítélni, hogy megfelelő-e a modell illeszkedése. Azonban ellenőrizzük ezt a feltételezésünket a globális F -próba segítségével.

Az F -próba nullhipotézise szerint $\hat{\beta}_1 = \hat{\beta}_2 = 0$, míg az alternatív hipotézis szerint $\hat{\beta}_j \neq 0$ valamelyik j -re ($j = 1, 2$). A próbafüggvény:

$$F = \frac{\frac{SSR}{m}}{\frac{SSE}{n-m-1}} = \frac{MSR}{MSE},$$

amelyhez készítsük el az ANOVA táblázatunkat (110. táblázat).

110. táblázat. Az ANOVA táblázat

A szóródás oka	Az eltérések négyzetösszege	Szabadsági fok	Szórásnégyzet becslése	F
Regresszió	(SSR=) 411,15	(m=) 2	(MSR=) 205,575	10,64
Hiba	(SSE=) 289,77	(n-m-1=) 15	(MSE=) 19,318	
Összesen	700,92	17	--	

Az SSR eltérés négyzetösszeg kiszámítása: $SSR = 1,033393621 + 0,5638071$,

ahol $393621 = 5,4 \cdot 23 + 4 \cdot 269 + \dots + 6,3 \cdot 143 - \frac{1685 \cdot 4364}{18}$ és

$8,0711 = 2,9 \cdot 23 + 2,9 \cdot 269 + \dots + 2,9 \cdot 143 - \frac{497 \cdot 4364}{18}$.

Az 5%-os szignifikancia szint mellett az elméleti F érték: $F_{(2,15)} = 3,68$. Mivel a próbafüggvény értéke ennél nagyobb, így a nullhipotézist elvetjük, vagyis a modell illeszkedése megfelelő.

Most már csak a regressziós paraméterek tesztelése van hátra, amihez a paraméterek standard hibáit kell meghatározni. Ez a

$Var(\hat{\beta}) = \frac{\mathbf{e}^T \cdot \mathbf{e}}{n-m-1} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} = s_e^2 \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1}$ képlet segítségével történik.

A számítás alapján:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{289,77}{15} \cdot \begin{bmatrix} 1,43 & -0,02 & -0,43 \\ -0,02 & 0,003 & -0,002 \\ -0,43 & -0,002 & 0,163 \end{bmatrix} = \begin{bmatrix} 27,63 & -0,38 & -8,33 \\ -0,38 & 0,05 & -0,04 \\ -8,33 & -0,04 & 3,14 \end{bmatrix}$$

A standard hibákat a főátlóban lévő elemek négyzetgyökei adják: $s_{\hat{\beta}_0} = 5,256$, $s_{\hat{\beta}_1} = 0,227$ és $s_{\hat{\beta}_2} = 1,772$.

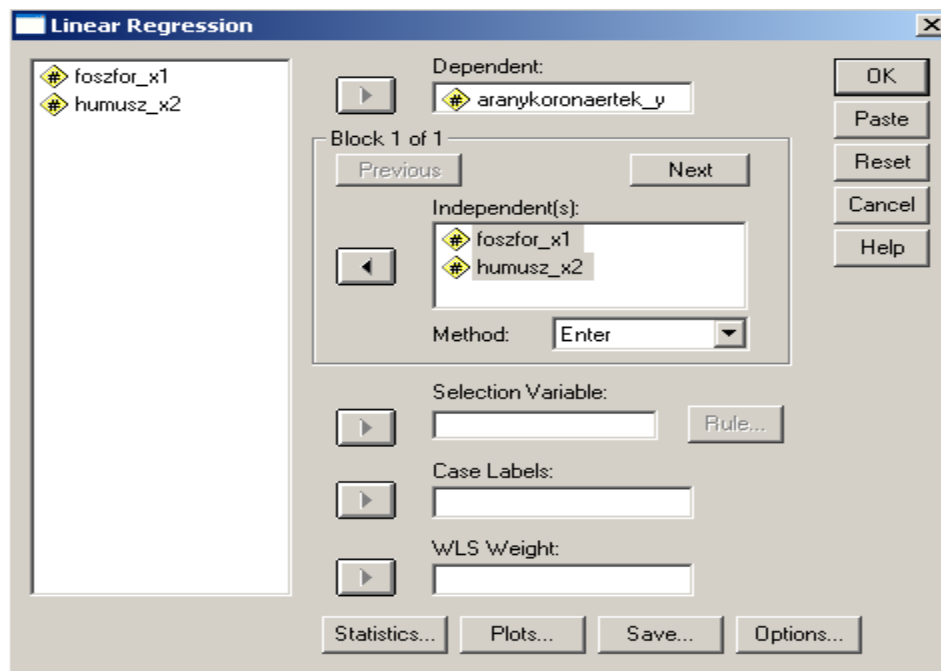
A parciális F -teszt próbafüggvényének az értékei $t = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$ alapján:

$$t_{\hat{\beta}_0} = \frac{13,0165}{5,256} \cong 2,476; \quad t_{\hat{\beta}_1} = \frac{1,0335}{0,227} \cong 4,553 \quad \text{és} \quad t_{\hat{\beta}_2} = \frac{0,5627}{1,772} \cong 0,317.$$

Kétoldali t -próba esetén ($\alpha = 0,05$ és $df = 15$) az elméleti t -érték: 2,1315. Mivel $|t_{\hat{\beta}_1}| = 4,553 > 2,1315$, ez azt jelenti, hogy az x_1 változó szignifikánsan befolyásolja a függő változó alakulását.

A regressziós paraméterek meghatározása az SPSS-vel

A kézi számítás után – ami igen hosszadalmas –, nézzük meg az SPSS-ben, hogyan lehet több-változós lineáris regressziót kiszámítani. A fenti példán keresztül csak bemutatjuk a több-változós lineáris regressziós beállításokat és összevetjük a kapott eredményeket a kézi számítás eredményeivel, majd egy példán keresztül részletesebben ismertetésre kerül a több-változós regressziós elemzés menete.



76. ábra. A több-változós lineáris regresszió elvégzésének panelja az SPSS-ben

Töltsük be a példához tartozó adattáblázatot, amit a „*Tobbszoros_linreg1.sav*” fájl tartalmaz. Kattintsunk ANALYZE menü REGRESSION almanüjének LINEAR... parancsára. A megjelenő panelban (76. ábra) végezzük el az alábbi beállításokat.

A bal oldali ablakban jelöljük ki a független változókat (foszfor_x1 és humusz_x2) majd helyezzük ezeket az INDEPENDENT(S) ablakba; a függő változót (aranykoronaertek_y) pedig a DEPENDENT ablakba tegyük. A METHOD ablakban az ún. ENTER nevű módszer van megadva alapállapotban, ez azt jelenti, hogy a vizsgálat minden független változót bevon az elemzésbe (a későbbiekben ennek a részletes elemzésére visszatérünk). Minden egyéb beállítást hagyjunk változatlanul. Kattintsunk az OK gombra, amelynek az eredményeképpen az Output ablakban kapott táblázatokat kell elemezni. Elsőként megjelenik az a táblázat, ami az alkalmazott modellt tartalmazza (111. táblázat), jelen esetben ez az ENTER módszer volt.

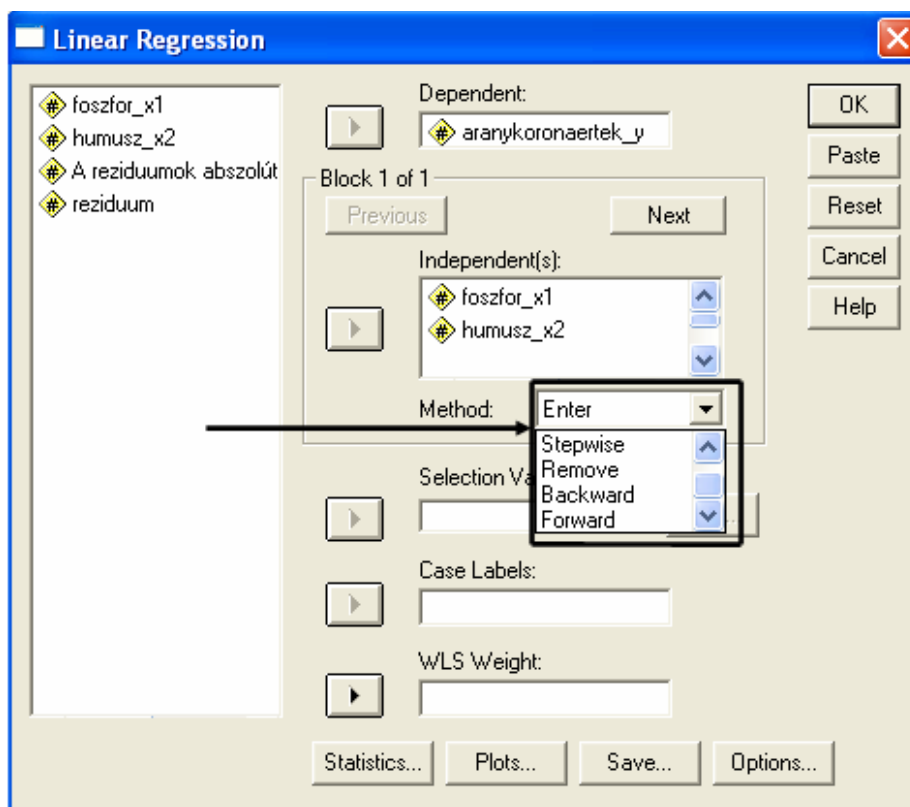
111. táblázat. Az Output első táblázata, ami a kiválasztott módszert takarja, a magyarázó változókat megjelenítve

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	humusz_x2 foszfor_x1	.	Enter

a. All requested variables entered.
 b. Dependent Variable: aranykoronaertek_y

Nézzük meg azt, hogy az SPSS milyen módszereket tud használni a változók kiválasztására. Ha visszamegyünk a főablakba, akkor a METHOD ablakra kattintva megjeleni még az ENTER módszer mellett másik négy eljárás is, ezek a FORWARD, a BACKWARD, a STEPWISE és a REMOVE (77. ÁBRA).



77. ábra. A változók kiválasztásnak módszerei

A FORWARD módszer minden lépésben azt a magyarázót vonja be a vizsgálatba, amelyik parciális F tesztjéhez a legkisebb P (vagyis hibázási) valószínűség tartozik. A bevonás folyamata addig tart, amíg a P az előre rögzített maximum érték (P_{IN}) alatt marad, vagy minden változót bevon.

A **BACKWARD** elimináció az induló lépésben mindegyik változót tartalmazza, és lépésenként mindig azt az egyet hagyja ki, amelyiknek a legkisebb a parciális korrelációja. Ekkor a parciális F teszthez a legnagyobb P valószínűség (a legnagyobb elsőfajú hiba) tartozik. Akkor áll le a módszer, ha a P kisebb, mint a küszöbérték (P_{OUT}), vagy már nincs változó a modellben.

A **STEPWISE** módszer a **FORWARD** szelekciótól annyiban tér el, hogy minden lépésben ellenőrzi a modellbe korábban bevont változók P valószínűségét, és ha a P értéke nagyobb, mint a küszöbérték, akkor a változót kihagyja a modellből. (Szokásos beállítás: $P_{IN}=0,05$; $P_{OUT}=0,1$.) Nem kerülünk végtelen ciklusba, ha $P_{IN} \leq P_{OUT}$.

A **REMOVE** eljárás a független változók közül eltávolítja azokat, amelyeknél az együtttható nem szignifikáns. A végső kifejezésben csak a maradék független változók szerepelnek.

AZ **OUTPUT** ablakban megjelenő következő táblázat (112. táblázat) a többszörös korrelációt, a determinációs együttthatót, a korrigált r^2 értékét, a regressziós modell standard hibáját tartalmazza (ezeket összevetve a kézi számolás eredményével, látható, hogy ugyanazok az értékek adódtak).

112. táblázat. Az **ENTER** módszer összefoglaló táblázata

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,766	,587	,532	4,39523

a.Predictors: (Constant), humusz_x2, foszfor_x1

A harmadik táblázat a modell tesztelésére szolgáló ANOVA táblázat (113. táblázat), amiből látszik, hogy a regressziós modell jól magyarázza az Y értékek szóródását ($p < 0,05$), vagyis a modell alkalmas az Y értékek becslésére.

113. táblázat. Az ANOVA tábla

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	411,334	2	205,667	10,646	,006
	Residual	289,771	15	19,318		
	Total	701,104	17			

a.Predictors: (Constant), humusz_x2, foszfor_x1

b. Dependent Variable: aranykoronaertek_y

A táblázat utolsó oszlopa szerint elvetjük a nullhipotézist, ami azt jelenti, hogy a modell alkalmas a függő változó magyarázatára. Abból azonban, hogy elvetjük a nullhipotézist még nem következtethetünk arra, hogy a függő változónak jó becslését tudjuk megadni, mert előfordulhat, hogy a modellben vannak olyan változók, amik nem szignifikáns paraméterűek. Erről a 114. táblázat ad tájékoztatást.

114. táblázat. Az együtthatók táblázata

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13,017	5,256		2,476	,026
	foszfor_x1	1,033	,227	,759	4,555	,000
	humusz_x2	,563	1,772	,053	,317	,755

a. Dependent Variable: aranykoronaertek_y

A t -próbához tartozó szignifikancia értékek alapján a humusz magyarázó változó szerepeltetése nem helyes a modellben ($p > 0,05$), azaz a humusz és az aranykorona érték között nincs lineáris kapcsolat.

Három független változós regresszióanalízis

A három független változós regresszióanalízis esetén: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \hat{\beta}_3 \cdot x_3$, ahol \hat{y} a függő változó becsült értéke, x_1, x_2, x_3 a független változók, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ az egyenlet becsült paraméterei.

Vizsgáljuk meg, hogy a micélium tömege hogyan függ a talaj N, P és K tartalmától, melyik tápanyag növelése mekkora hatással van a micélium súlyának alakulására (115. táblázat)

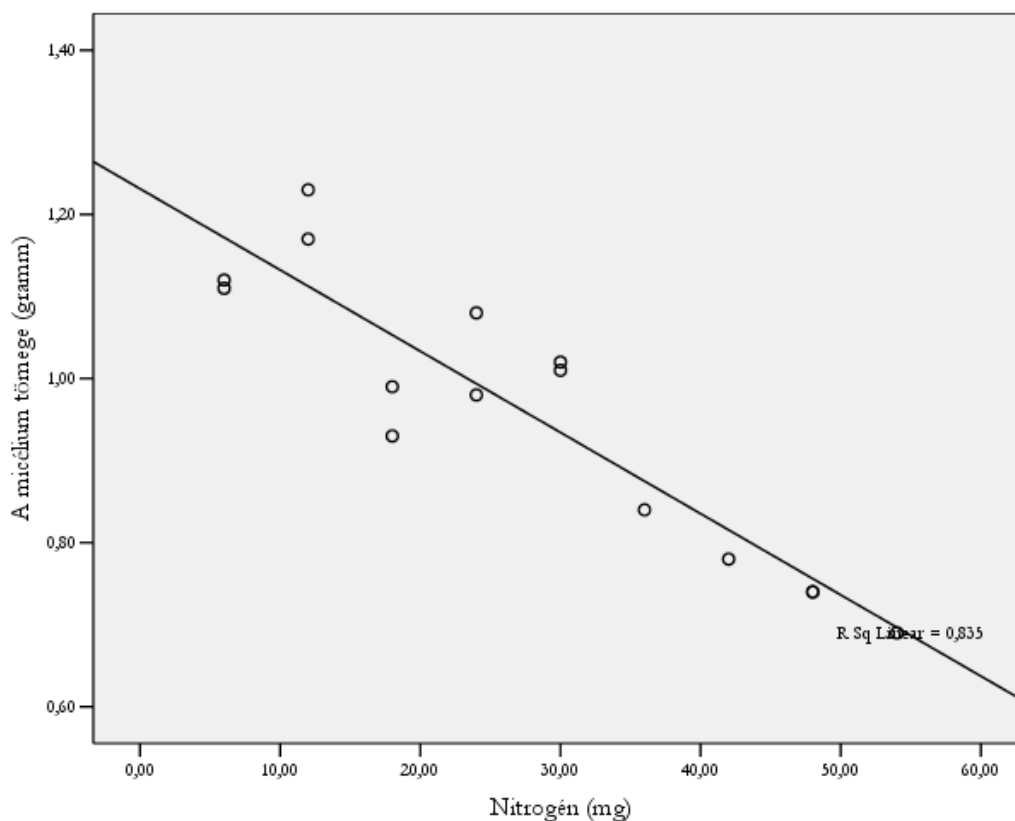
115. táblázat. A N, P és K különböző kombinációinak hatása az *Aspergillus niger* micéliumának tömegére

$N(x_1)$ mg	$P(x_2)$ mg	$K(x_3)$ mg	y
12	72	4	1,23
12	48	4	1,17
6	72	32	1,12
6	24	16	1,11
24	8	20	1,08
30	32	12	1,02
30	32	20	1,01

18	16	12	0,99
24	16	24	0,98
18	40	28	0,93
36	24	28	0,84
42	8	8	0,78
48	56	36	0,74
48	40	32	0,74
54	56	36	0,69

Forrás: SVÁB JÁNOS (1981), 317.o.

Az SPSS alkalmazása előtt nézzük meg grafikus módszerrel (2-3 dimenziós ábrákkal), hogy közelítően teljesülnek-e a lineáris regressziószámítás előfeltételei, használható-e a modell (*tobbszoros_linreg2.sav*). Mivel a grafikus ábra magasabb dimenzióban nem készíthető el, ez a lépés nem helyettesítheti a modell jóságát vizsgáló tesztet, de arra alkalmas, hogy a teljesen hasznavehetetlen számításokat megelőzzük.



78. ábra. A micélium tömege a talaj nitrogén tartalma (mg) függvényében

Az Y és az x_1, x_2, \dots változók pontdiagramját vizsgálva leolvashatók a következők:

Lineáris-e a kapcsolat, jogos-e a lineáris modell alkalmazása, vagy más függvénytypust célszerű választani?

Az x növekedésével az y adatok szórása változatlan marad-e, vagyis a hibatag konstans szórása feltételezhető-e?

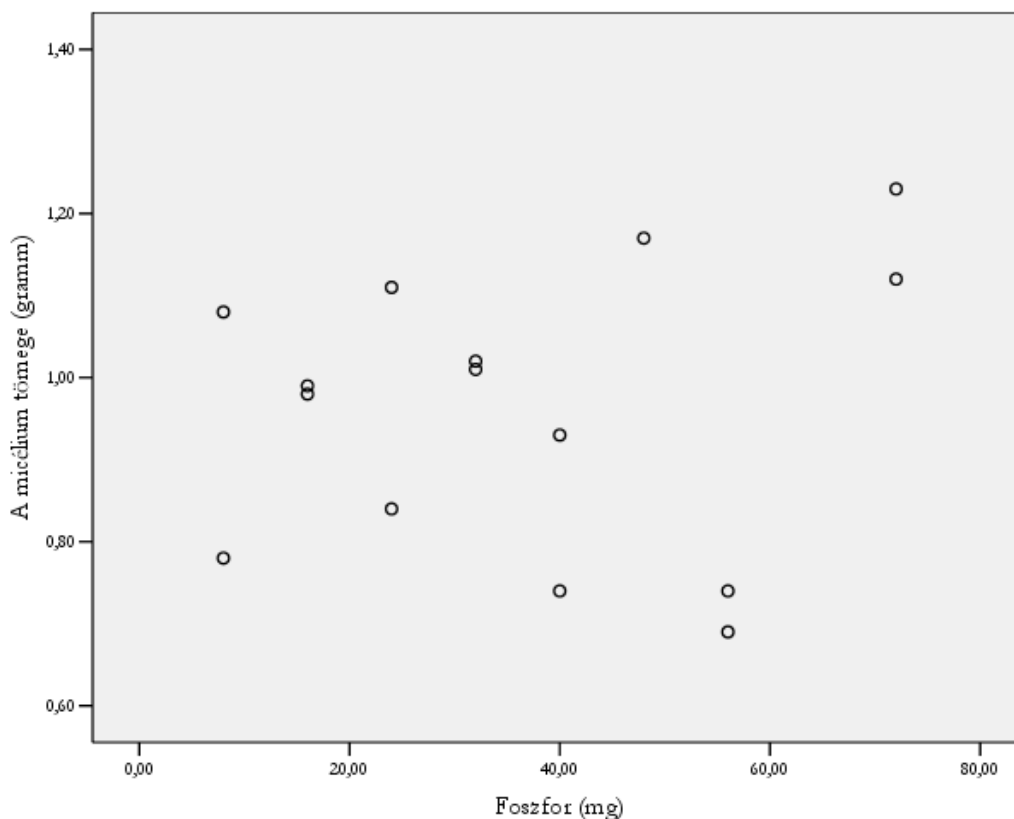
Homogén-e a minta, vagy alminták láthatók, amelyekben más-más tendencia érvényesül a változók között?

Vannak-e kiugró pontok és milyen az elhelyezkedésük?

Ábrázoljuk pontdiagramon (GRAPHS / SCATTER) minden egyes független változó és a függő változó kapcsolatát külön-külön, ezt a 78. ábra, 79. ábra, 80. ábra mutatja.

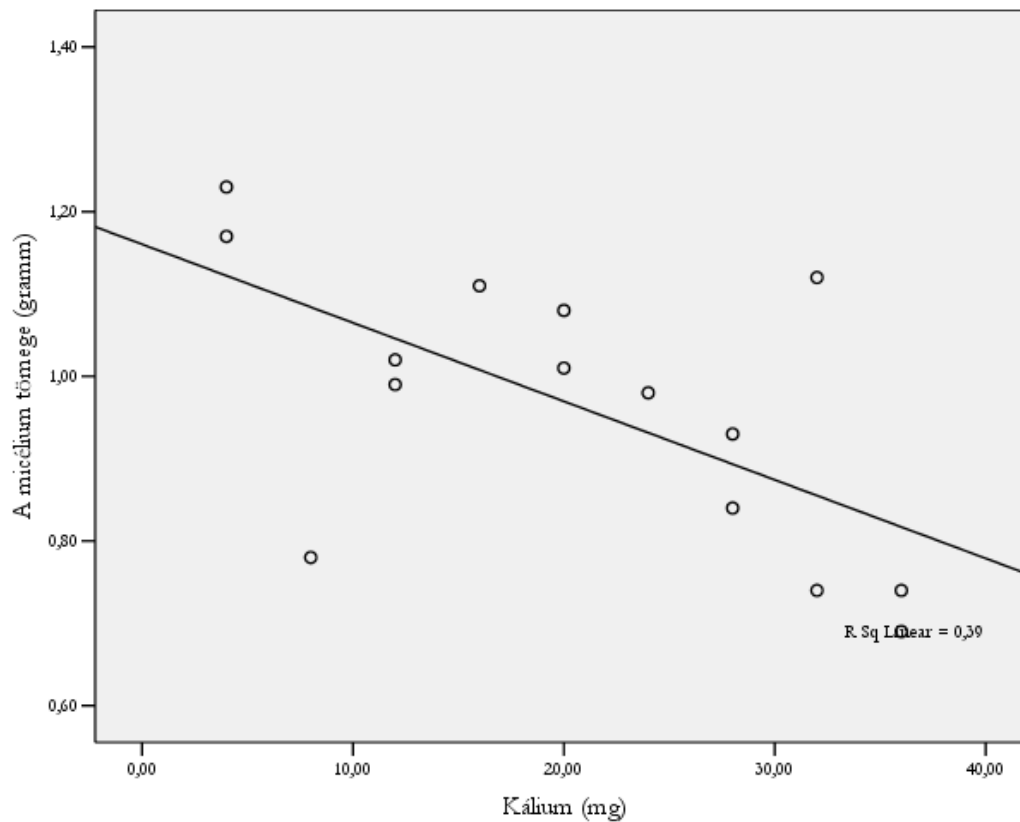
A talaj nitrogén tartalma és a micélium tömege közötti kapcsolat lineárisnak tekinthető, a közöttük lévő kapcsolat ellentétes irányú, vagyis növekvő nitrogéntartalomhoz csökkenő micéliumtömeg tartozik.

Ha a talaj foszfor tartalmának a függvényében nézzük a micélium tömegét (79. ábra), akkor a ponthalmaz elhelyezkedése alapján nem lehet tendenciózus megállapítást levonni. Függvényszerű kapcsolatot nem lehet leolvasni, még akkor sem, ha esetleg almintára bontanánk a mintát (hangsúlyozzuk az alacsony mintaszám erre egyébként nem ad lehetőséget).



79. ábra. A micélium tömege a talaj foszfor tartalma (mg) függvényében

A talaj kálium tartalmának a függvényében megvizsgálva a micélium tömegét, a pontdiagrammot a 80. ábra mutatja.



80. ábra. A micélium tömege a talaj kálium tartalma (mg) függvényében

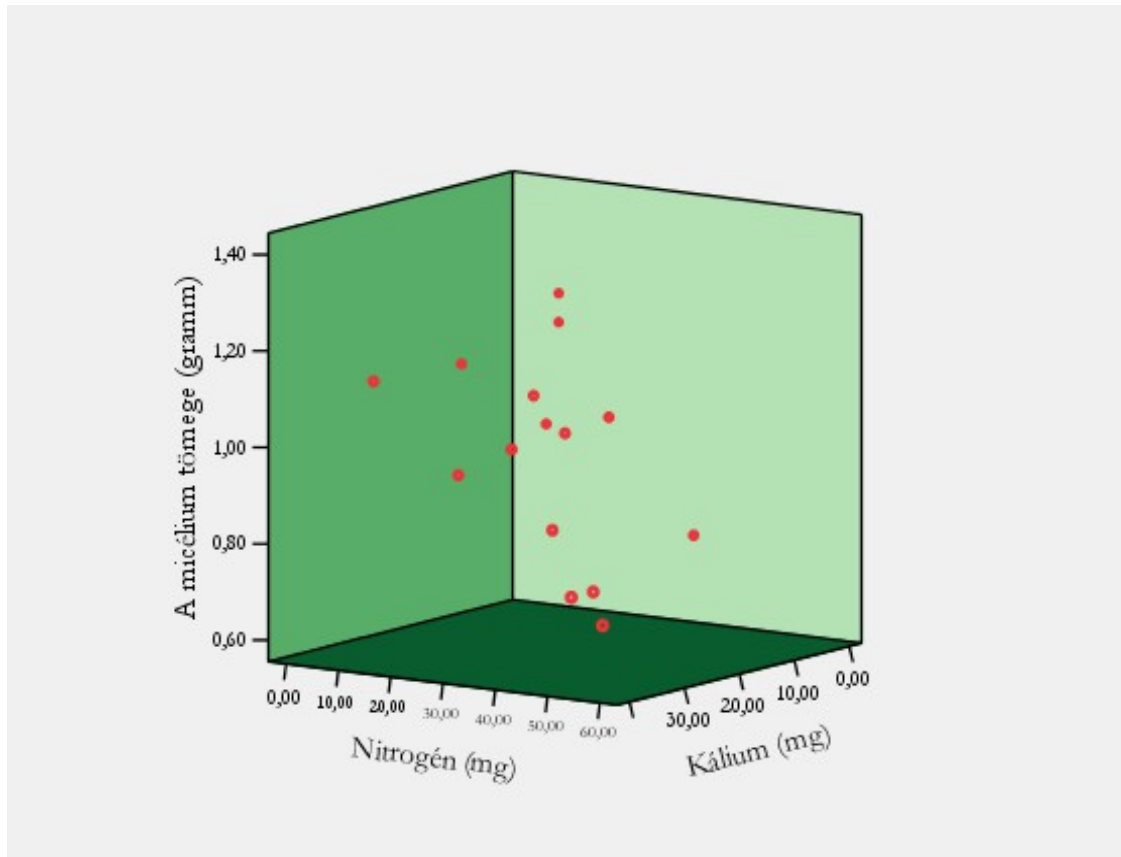
A pontok itt is szóródnak, ám első közelítésben megpróbáltunk a pontokra egyenest illeszteni. A minta jó közelítéssel homogénnek tekinthető, van néhány kiugró érték, amely elhagyásával a lineáris regressziós egyenes illeszkedését javítani lehetne, ám az alacsony mintaszám miatt ettől eltekintünk.

A kapott ábrák alapján úgy tűnik, hogy a három független változó közül a talaj foszfor tartalma az a változó ami nem illeszkedik a lineáris modellbe. Három dimenziós ábrán jelenítsük meg a másik két független változót (a talaj nitrogén és kálium tartalmát) és a micélium tömegének az alakulását (81. ábra).

Bár a három dimenziós ábrák elemzése nem könnyű, ám a pontok elhelyezkedése alapján durva közelítésben mondhatjuk, hogy lineáris összefüggés látható a vizsgált változók között.

Az ábrák elkészítése után végezzük el a regressziószámítást. A többszörös regressziószámítás elvégzéséhez kattintsunk az ANLYZE menüpont REGRESSION almenüjének LINEAR... parancsára. A megjelenő panelban (82. ábra) végezzük el a következő beállításokat: a függő változó ablakba (DEPENDENT)

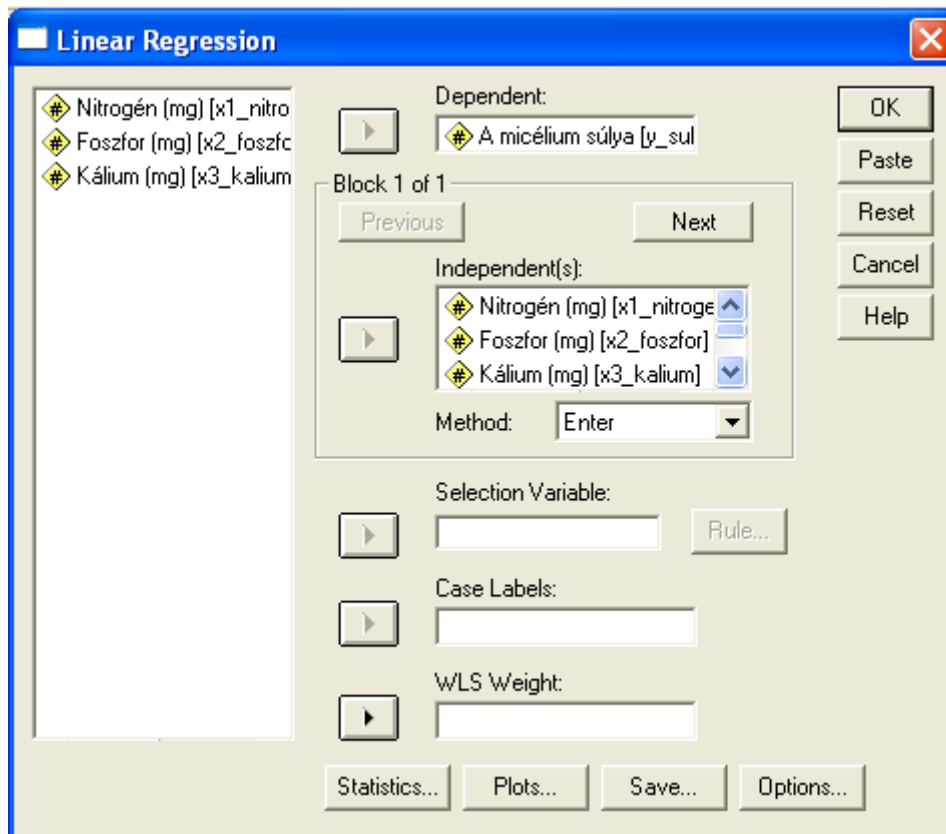
helyezzük a micélium tömege változót, míg a független változók közé helyezzük be a talaj nitrogén tartalma (x_1), foszfor tartalma (x_2) és a kálium tartalma (x_3) változókat.



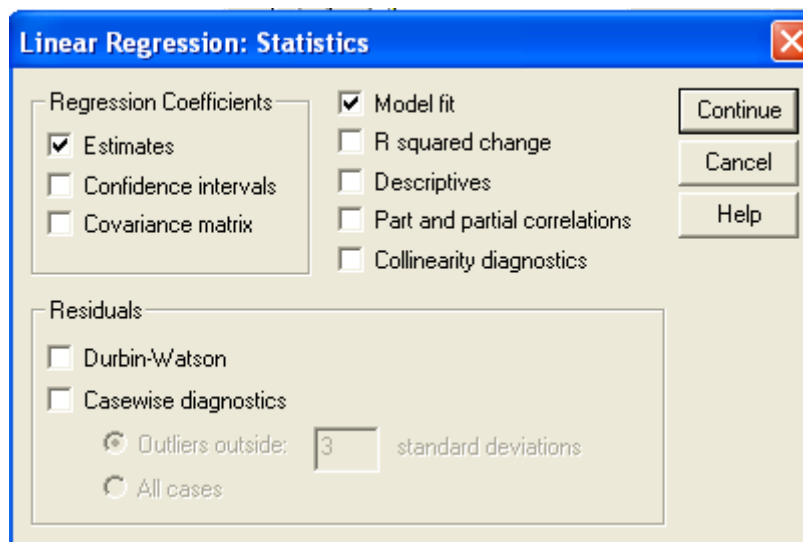
81. ábra. A változók három dimenziós ábrája

Azt, hogy a független változókat hogyan válassza be a program, a METHOD ablakban állíthatjuk be. Megint az ENTER módszert jelöltük meg, ahogy azt az előbbi feladatban is tettük.

Kattintsunk a STATISTICS... parancsgombra (83. ábra). A REGRESSION COEFFICIENTS részben az ESTIMATES parancs megjelölésével azt érjük el, hogy a program a regressziós paramétereket írja ki. De ebben az ablakban van arra lehetőségünk, hogy konfidencia intervallumot (CONFIDENCE INTERVALS) és kovariancia mátrixot (COVARIANCE MATRIX) is lekérjünk.



82. ábra. A többszörös lineáris regresszió beállításai



83. ábra. A STATISTICS... parancsgomb beállításai

A modell illeszkedését (MODEL FIT), az r^2 változását (R SQUARED CHANGE), a leíró statisztikákat (átlag, szórás, megfigyelések száma) (DESCRIPTIVES), a parciális korrelációt (PART AND PARTIAL CORRELATIONS) és multikollinearitási méreteket

(COLLINEARITY DIAGNOSTICS) a jobb oldali panelrészben történő megjelölésekkel kérhetünk. Ezek közül mi most a kérjük az r^2 változását. A reziduális részben Durbin-Watson tesztet és esetenkénti diagnosztikát kérhetünk. Ha a vizsgálati minta száma nagy, érdemes kiírni a kiugró értékeket, amelyek az átlagtól 2-3 szórásnyi távolságra vannak, ezek ugyanis nagymértékben torzíthatják a kapott eredményeket. A 83. ábra csak a program alapbeállításait mutatja, mi most azonban jelöljük meg minden lehetőséget, majd futtassuk le a programot.

Az elsőként kapott táblázatban (116. táblázat) a leíró statisztika eredményeit látjuk, a változók átlagát, szórását és a megfigyelt esetek számát közli a program.

116. táblázat. A változók átlaga és szórása

Descriptive Statistics			
	Mean	Std. Deviation	N
A micélium tömege (gramm)	0,9620	,16992	15
Nitrogén (mg)	27,2000	15,68985	15
Foszfor (mg)	36,2667	21,13719	15
Kálium (mg)	20,8000	11,13040	15

A leíró statisztikákat tartalmazó táblázatból a micélium tömegének átlaga 0,962 gramm, a talaj nitrogén tartalmának átlaga 27,2 mg, a foszfortartalom átlaga 36,27 mg és a káliumtartalom átlaga 20,8 mg. A szórások alapján a foszfortartalom esetében legnagyobb a szórás, ez összhangban van a két dimenziós ábrán kapott képpel. A mintaszám minden változó esetén 15.

A korrelációs mátrixban (117. táblázat) a függő és a magyarázó változók páronkénti korrelációi, a szignifikancia-szintek és a minta mérete szerepel.

A szignifikancia értéke alapján a micélium tömege (y) a talaj nitrogén tartalmával (x_1) és kálium tartalmával (x_3) van szignifikáns kapcsolatban. A Pearson-féle korreláció értéke azt mutatja, hogy a nitrogéntartalom ($r = -0,914$) erős sztochasztikus kapcsolatban van a micélium tömegével, de ez a kapcsolat ellentétes irányú; míg a káliumtartalomnál közepes erősségű a sztochasztikus kapcsolat, és ez a változó is negatív hatással van a micélium tömegére, vagyis növekvő káliumtartalom esetén csökkenő micéliumtömeget kapunk.

117. táblázat. Korrelációs mátrix

Correlations					
		A micélium tömege (gramm)	Nitrogén (mg)	Foszfor (mg)	Kálium (mg)
Pearson Correlation	A micélium tömege (gramm)	1,000	-,914	,141	-,625
	Nitrogén (mg)	-,914	1,000	-,096	,485
	Foszfor (mg)	,141	-,096	1,000	,237
	Kálium (mg)	-,625	,485	,237	1,000
Sig. (1-tailed)	A micélium tömege (gramm)	.	,000	,309	,006
	Nitrogén (mg)	,000	.	,367	,033
	Foszfor (mg)	,309	,367	.	,197
	Kálium (mg)	,006	,033	,197	.
N	A micélium tömege (gramm)	15	15	15	15
	Nitrogén (mg)	15	15	15	15
	Foszfor (mg)	15	15	15	15
	Kálium (mg)	15	15	15	15

A 118. táblázatban a többszörös korreláció és a determinációs együttható, a korigált r^2 , a regressziós modell standard hibája szerepel. Az utolsó oszlopban a Durbin-Watson teszt eredményét látjuk.

118. táblázat. A Stepwise módszer összefoglaló táblázata

Model Summary										
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,946	,895	,866	,06215	,895	31,220	3	11	,000	2,092

a. Predictors: (Constant), Kálium (mg), Foszfor (mg), Nitrogén (mg)

b. Dependent Variable: A micélium tömege (gramm)

A gyakorlatban a táblázatból számunkra a második és a harmadik oszlop az érdekes. A táblázat második oszlopában a többszörös korrelációs koefficiens értéke található ($r = 0,946$), ami a függő változó és a független változók közötti lineáris összefüggés szorosságát fejezi ki. A harmadik oszlopban a többszörös determinációs koefficiens értéke olvasható le ($r^2 = 0,895$), ez az érték azt mutatja meg, hogy az Y függő változó szóródásából mennyi tulajdonítható a független változók hatásának. A micélium tömege szóródásának 89,5%-a a kapott eredmény alapján a N, P és K tápanyag változásával magyarázható.

A program elkészíti a regresszióanalízis varianciaanalízis táblázatát is (119. táblázat), ami a modell tesztelésére szolgál. Ebből a táblázatból olvashatjuk le, hogy a modell mennyire jól magyarázza meg az Y értékek szóródását. A

táblázat utolsó oszlopából láthatjuk ($p < 0,05$), hogy a nullhipotézisünket el kell vetni, ami azt jelenti, hogy a modell jó.

119. táblázat. Az Anova táblázat

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,362	3	,121	31,220	,000 ^b
	Residual	,042	11	,004		
	Total	,404	14			

a. Predictors: (Constant), Kálium (mg), Foszfor (mg), Nitrogén (mg)

b. Dependent Variable: A micélium tömege (gramm)

A 120. táblázatban kapjuk meg a többszörös lineáris regressziós modell felírásához szükséges paramétereket, valamint, hogy a változók egyenként szignifikánsan befolyásolják-e az Y változót.

A regressziós koefficiensek értékeit a táblázat második oszlopából olvashatjuk le, ez alapján a micélium tömege és a talaj nitrogén-, kálium- és foszfor tartalma közötti összefüggés $y = 1,238 - 0,008x_1 + 0,001x_2 - 0,004x_3$ formában írható fel, ahol Y a micélium tömege, x_1, x_2, x_3 pedig a talaj nitrogén-, kálium- és foszfor tartalma. A standardizált koefficiens oszlopban lévő „Beta” értékről már korábban szoltunk, ám igazi jelentését most érthetjük meg. Többszörös lineáris regressziónál minél közelebb van a „Beta” értéke az 1-hez, annál inkább

Azt, hogy az egyes regressziós koefficiensek valóban befolyásolják-e az Y változót t -próbával döntjük el a $b_i = 0$ ($i = 1, 2, 3$) nullhipotézissel szemben. A t -próba eredményét a t oszlopban láthatjuk, míg a szignifikancia oszlopában olvassuk azt le, hogy melyik regressziós együttható hatása szignifikáns.

120. táblázat. A regressziós együtthatók

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
	Nitrogén (mg)	,008	,001	-,761	-6,600	,000	-,011	-,005	-,914	-,894	-,645	,718	1,39
	Foszfor (mg)	,001	,001	,136	1,309	,217	-,001	,003	,141	,367	,128	,886	1,13
	Kálium (mg)	-,004	,002	-,288	-2,434	,033	-,008	,000	-,625	-,592	-,238	,684	1,46

a. Dependent Variable: A micélium tömege (gramm)

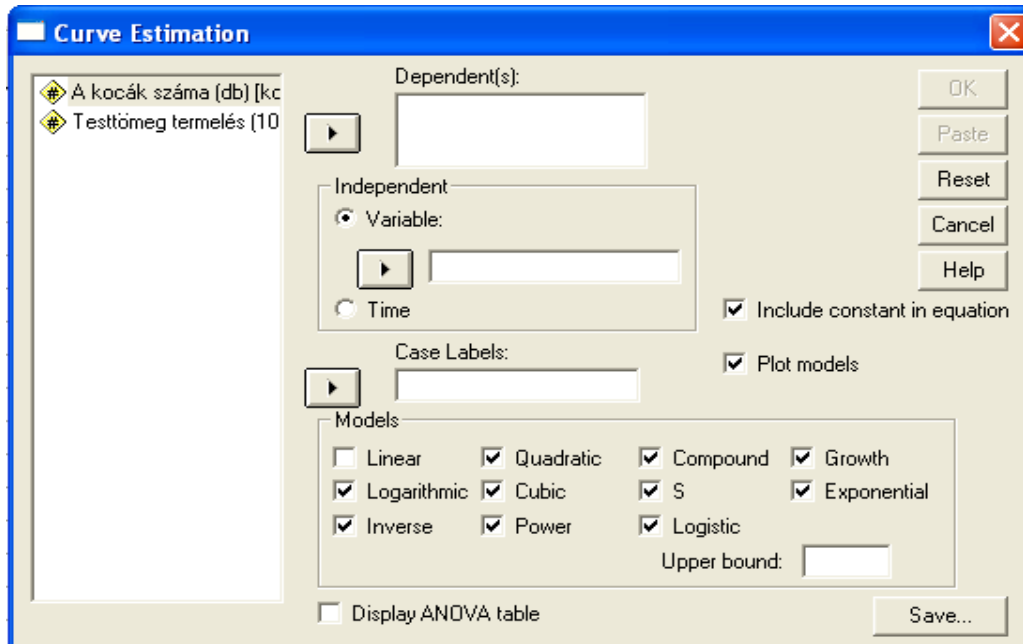
A nitrogén növelésének depresszív hatása szignifikáns, a foszfor változásának hatása nem bizonyítható, míg a kálium hatása $p=5\%$ -s szinten bizonyítható. A N és a K hatása negatív.

A parciális regressziós koefficiensek szignifikanciájának az alakulását befolyásolja a független változók egymás közötti korrelációja. Ha a független változók között erős a korreláció, akkor az értelmezésnél a kapott eredményekkel óvatosan kell bánni.

Nemlineáris összefüggések vizsgálata

Az előző példákban a változók közötti kapcsolat elemzésekor (mind az egyszeres, mind a többszörös esetben) lineáris regressziót alkalmaztunk. A biometria témakörébe tartozó jelenségek között azonban gyakrabban fordul elő az, hogy a függő változó a független változó 1 egységnyi változására nem állandó változással reagál a különböző x pontokban. A statisztikai gyakorlat éppen ezért gyakran nemlineáris függvények illesztését igényli.

A nemlineáris függvényeket statisztikai szempontból két csoportra osztjuk: lineárisra visszavezethető és lineárisra nem transzformálható modellekre.



84. ábra. *Nemlineáris, de linearizálható függvények az SPSS-ben*

Ezek az illesztések az SPSS-ben az alábbi parancssorral indíthatók el: ANALYZE / REGRESSION / CURVE ESTIMATION... (84. ábra).

Lineárisra visszavezethető összefüggések vizsgálata

Ha a lineáris regresszió feltételei nem teljesülnek, vagy rossz illesztést kapunk, akkor meg kell próbálkozni a függő és a független változók transzformációjával. A transzformált adatokon már lineáris regressziós elemzést hajtunk végre, de ez az eredeti adatoknál már nem lineáris összefüggést fog magyarázni. A továbbiakban ismertetünk néhány lehetőséget a nemlineáris kapcsolatnak a lineáris regresszió segítségével való megadására.

A 84. ábrán a MODELS részben pipával jelöltük a nemlineáris, de linearizálható függvényeket. Ezek megnevezését és képletét a 121. táblázatban foglaltuk egybe.

121. táblázat. A legfontosabb nemlineáris (de linearizálható) regressziós függvények

SPSS elnevezés	Típus	Egyenlet
Logarithmic	Logaritmikus	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \ln x$
Inverse	Inverz	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1/x$
Quadratic	Parabolikus	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \hat{\beta}_2 \cdot x^2$
Cubic	Harmadfokú	$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \hat{\beta}_2 \cdot x^2 + \hat{\beta}_3 \cdot x^3$
Power	Hatványkitevős	$\hat{y} = \hat{\beta}_0 \cdot x^{\hat{\beta}_1}$ vagy $\ln \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \ln x$
Compound	Vegyes	$\hat{y} = \hat{\beta}_0 \cdot \hat{\beta}_1^x$ vagy $\ln \hat{y} = \ln \hat{\beta}_0 + (\ln \hat{\beta}_1 \cdot x)$
S	Szigmoid	$\hat{y} = e^{\hat{\beta}_0 + \hat{\beta}_1/x}$ vagy $\ln \hat{y} = \hat{\beta}_0 + \hat{\beta}_1/x$
Logistic	Logisztikus	$\hat{y} = \frac{1}{(1/u) + \hat{\beta}_0 \cdot \hat{\beta}_1^x}$ vagy $\ln \left(\frac{1}{\hat{y}} - \frac{1}{u} \right) = \ln \hat{\beta}_0 + (\ln \hat{\beta}_1 \cdot x)$
Growth	Növekedési	$\hat{y} = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x}$ vagy $\ln \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$
Exponential	Exponenciális	$\hat{y} = \hat{\beta}_0 \cdot \hat{\beta}_1^x$ vagy $\hat{y} = \hat{\beta}_0 \cdot e^{\hat{\beta}_1 \cdot x}$ vagy $\ln \hat{y} = \ln \hat{\beta}_0 + \hat{\beta}_1 \cdot x$

A továbbiakban az alábbi függvényekkel foglalkozunk részletesen egy-egy példán keresztül:

Logaritmusfüggvény esetén az Y változó az X logaritmusával van lineáris összefüggésben, azaz X szorzatos változására Y additívan reagál.

Exponenciális összefüggés esetén a logaritmusfüggvénnyel ellentétben az Y logaritmusa az X -szel van lineáris összefüggésben. Az exponenciális összefüggésben Y növekedésének a sebessége arányos v már elért értékével.

Hatványfüggvény esetén Y logaritmusa az X logaritmusával van lineáris összefüggésben.

A *parabolikus függvény* és a *harmadfokú függvény* a polinomiális függvénycsaládba tartozik, ez a függvénycsalád gyakorlatilag bármilyen összefüggés leírására alkalmas, de az összefüggés törvényszerűségét legtöbbször nem jellemezik.

Logisztikus függvény esetén a függő változó értékei először lassan, majd egyre gyorsabban növekednek, majd ismét lassulnak egy felső határ felé közelítve.

Logaritmikus regresszió

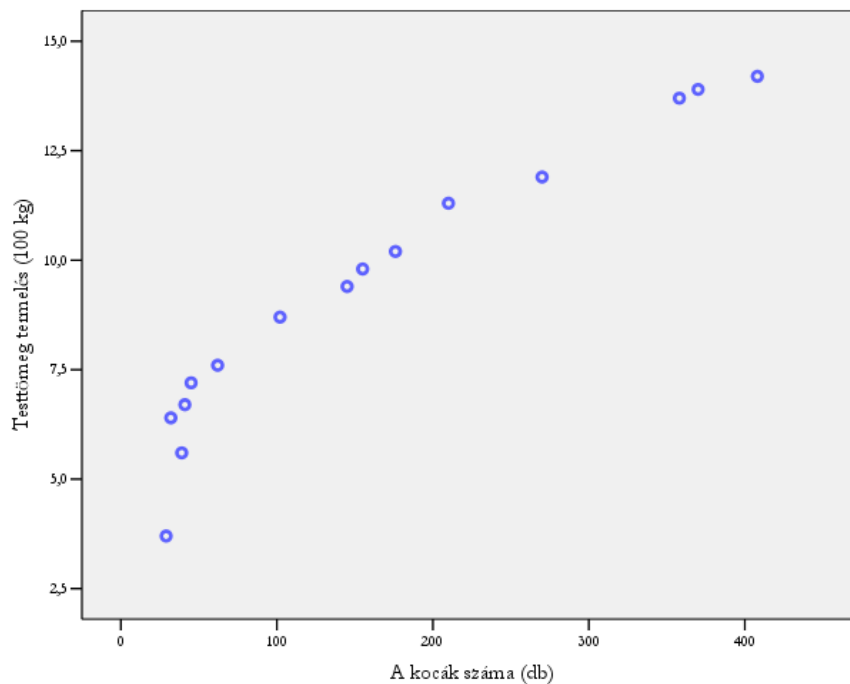
122. táblázat. A logaritmikus regresszióhoz tartozó adattáblázat

A kocák száma (db)	Testtömeg termelés (100 kg)	A kocák száma (db)	Testtömeg termelés (100 kg)
29	3,7	155	9,8
32	6,4	176	10,2
39	5,6	210	11,3
41	6,7	270	11,9
45	7,2	358	13,7
62	7,6	370	13,9
102	8,7	408	14,2
145	9,4	--	--

Forrás: MANCZEL (1983): Statisztikai módszerek alkalmazása a mezőgazdaságban

Vizsgáljuk meg a 122. táblázathoz tartozó adatok alapján, hogy egy sertéstelepen fokozatosan növelve a kocalétszámot, az állatok testtömegtermelése hogyan változik. A kocák számát és a testtömegtermelés közötti adatokat tartalmazó SPSS fájlt „*logaritmus_regresszio.sav*” név alatt mentettük el.

A két változó közötti kapcsolat jellegének szemléltetéséhez ábrázoljuk a pontpárokat az SPSS-ben. A pontdiagramot a tanult módon készítjük el (GRAPHS / SCATTER...). A két változó közötti kapcsolatot a 85. ábra szemlélteti



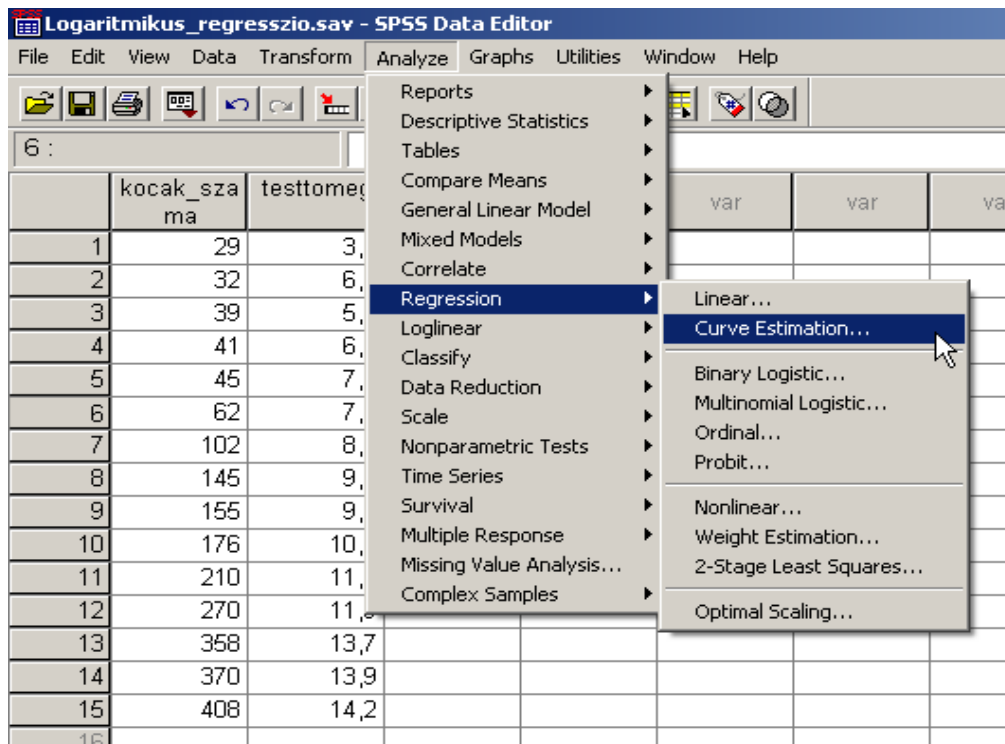
85. ábra. A két változó közötti pontdiagram

Az ábráról látszik, hogy ebben az esetben a ponthalmazra nem az egyenes illesztése tűnik megfelelőnek, hanem a logaritmussfüggvény, amelynek az egyenlete $\hat{y} = \beta_1 \cdot \ln x + \hat{\beta}_0$ alakú.

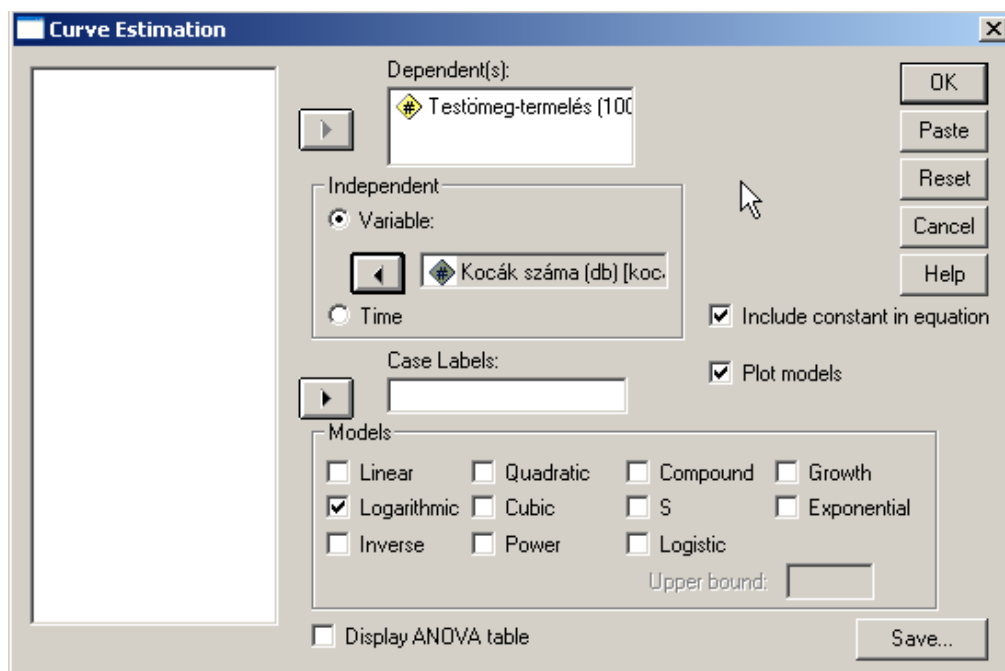
A továbbiakban nem ismertetjük a regressziós paraméterek manuális kiszámításának módszerét és menetét, hanem a könyv céljának alárendelten az SPSS-ben mutatjuk meg a számításokhoz szükséges beállításokat.

Kattintsunk az ANALYZE menüpont REGRESSION almenüjének CURVE ESTIMATION... parancsára (86. ábra).

A megjelent panelben a bal oldali ablakból (87. ábra) válasszuk ki a független változót (*kocák száma*) és a nyilacska segítségével tegyük át a VARIABLE ablakba, majd a függő változót (*testtömeg termelés*) a DEPENDENT(S) ablakba.



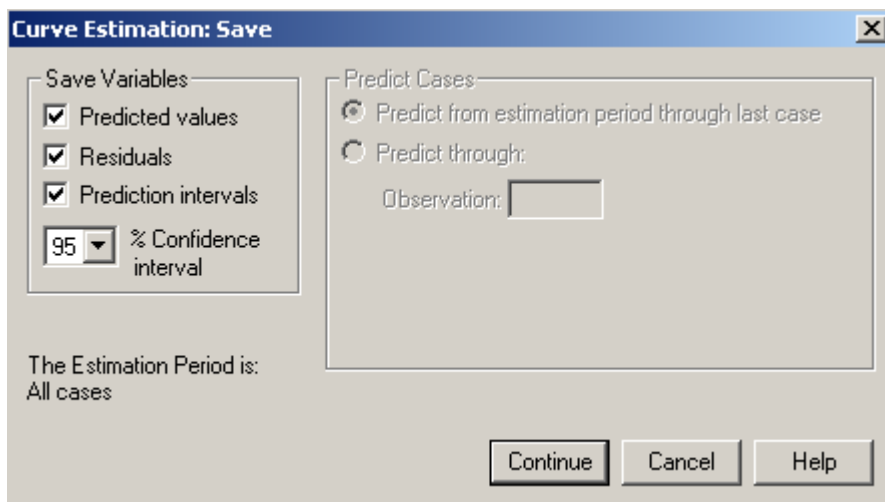
86. ábra. A logaritmus regressziós függvény illesztésének parancssora



87. ábra. A logaritmus regressziós függvény illesztése SPSS-ben

A MODELS részben van arra lehetőség, hogy a különböző regressziós függvények közül válasszunk (egyszerre több regressziós függvényt is kijelölhetünk). Jelöljük meg a LOGARITHMIC függvényt.

Más beállításokat is végezhetünk. Ha azt akarjuk, hogy a program a konstans tagra is adjon becslést, akkor az INCLUDE CONSTANT IN EQUATION mellett hagyjuk meg az alapbeállításban megjelenő pipát. A PLOT MODELS megjelölésével (ami szintén alapbeállítás) a program grafikusán jeleníti meg a megfigyelési pontokra illesztett, általunk kiválasztott regressziós függvényt. A DISPLAY ANOVA TABLE megjelölésével variancia táblát készítetünk a programmal.



88. ábra. A SAVE parancs beállításai

A SAVE parancsra kattintva (88. ábra) a következő beállításokra van lehetőségünk:

PREDICTED VALUES: Megjelölve a regressziós függvény által becsült \hat{y} értékeket írja ki a program az adatmátrixba új változóként FIT_1 név alatt.

RESIDUALS: Ha bejelöljük, akkor a maradékok egy külön változóban jelennek meg az adatmátrixban ERR_1 név alatt.

123. táblázat. A logaritmikus regressziós függvény összefoglaló táblázata

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,976	,953	,949	,722

The independent variable is A kocák száma (db).

PREDICTION INTERVALS: Kipiálva, a megadott szignifikancia szinten (alapbeállításban 95%), akkor két újabb változóban (LCL_2 és UCL_2) változónév alatt a konfidencia intervallum határait adja meg a program.

A bemutatott beállítások mindegyikét megjelölve futtassuk a programot, majd elemezzük az Output ablakban megjelent táblázatokat és a kapott ábrát.

A 123. táblázat első oszlopa tartalmazza a lineáris korrelációs együttható értékét ($r = 0,976$). A második oszlopban a determinációs együttható értékét ($r^2 = 0,95$) látjuk, ami szerint a modell 95%-ban tudja magyarázni az Y értékek eltérés négyzetösszegét. Ez jó eredménynek tekinthető, hiszen csak 5% a hibából adódó eltérés. A harmadik oszlop a korrigált r^2 adja, míg az utolsó oszlopban a regressziós modell standard hibája szerepel.

A 124. táblázat az ANOVA táblázat, amely tartalmazza többek között az eltérés- és átlagos négyzetösszegeket, az F -próba értékét (ezek korábban ismertetésre kerültek). Az utolsó oszlop jelenti számunkra a legfontosabb információt, innen olvasható le, hogy a kiválasztott modell helyes-e. A nullhipotézis szerint az \hat{y} értékek véletlenszerűen szóródnak. Mivel a szignifikancia érték kisebb 0,05-nél, így elvetjük a nullhipotézist, tehát a logaritmikus modellünk helyes.

124. táblázat. A logaritmikus regressziós függvény illesztéséhez tartozó ANOVA táblázat

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	137,415	1	137,415	263,379	,000
Residual	6,783	13	,522		
Total	144,197	14			

The independent variable is A kocák száma (db).

A 124. táblázatból a regressziós paramétereket (B) és azok tesztelését kapjuk meg (Sig). A „kocák száma” sorban a $\hat{\beta}_1$ paraméter becsült értékét látjuk, míg a konstans (*Constans*) sorban a β_0 együttható értéke olvasható le.

125. táblázat. A paramétereket megadó táblázat

Coefficients					
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
A kocák száma	3,289	,203	,976	16,229	,000
(Constant)	-6,151	,973		-6,319	,000

A paraméterek alapján a regressziós egyenlet:

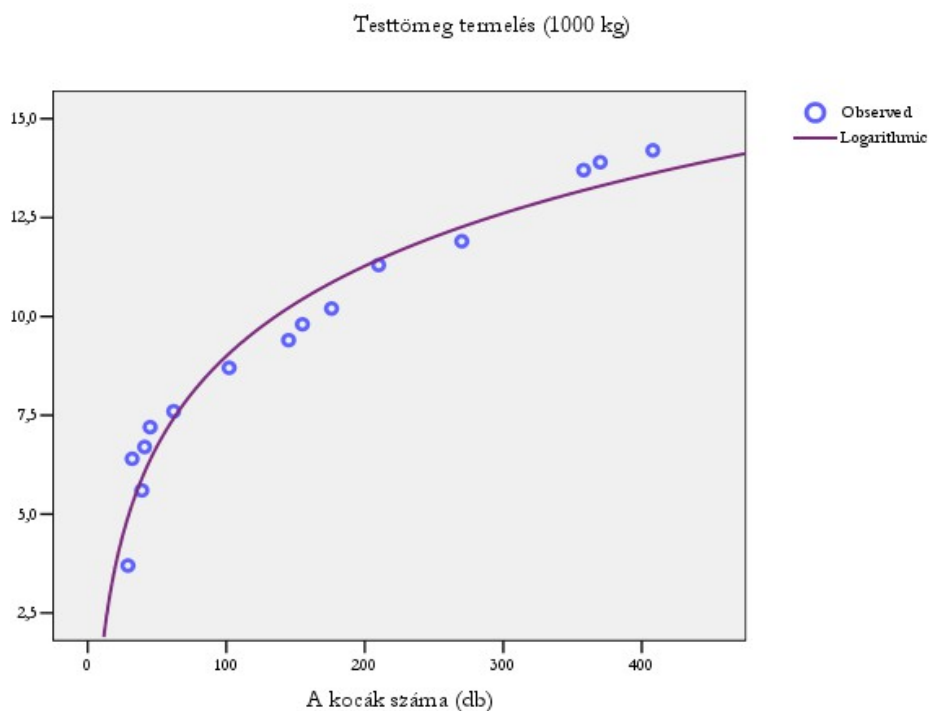
$$\hat{y} = 3,289 \cdot \ln x - 6,115 .$$

A kapott regressziós görbe egyenlete alapján megbecsülhetjük például azt, hogy 250 darabos kocalétszámhoz mekkora testtömegtermelés társul. A függvénybe helyettesítve az $x = 250$ -et megkaphatjuk a testtömegtermelést:

$$\hat{y} = 3,289 \cdot \ln 250 - 6,115 , \text{ ahonnan } \hat{y} = 12,45 .$$

Megállapíthatjuk, hogy a kocalétszám 250 darabra növelésével a testtömegtermelés 12,45 · 100 kg lesz.

A kiszámított értékeket, azaz a program által a ponthalmazra illesztett logaritmikus függvényt a 89. ábra mutatja.



89. ábra. Az empirikus adatokra illesztett logaritmikus függvény

A nemlineáris, de linearizálható kapcsolatok esetében a korrelációs index szolgál mérőszámmal a két változó közötti kapcsolat jellemzésére, amit az

$$I = \sqrt{1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

képlettel értelmeztünk ($e_i = y_i - \hat{y}_i$). Először meghatározzuk a korrelációs indexet, a transzformált változók közötti lineáris korrelációs együtthatót, majd értelmezzük a kapott eredményeket.

A számításhoz szükséges részeredményeket a 126. táblázat tartalmazza.

126. táblázat. A korrelációs index kiszámításához szükséges munkatábla

	x_i	y_i	\hat{y}_i	e_i	$(e_i)^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	29	3,7	4,96	-1,26	1,588	-5,65	31,923
2	32	6,4	5,28	1,12	1,246	-2,95	8,703
3	39	5,6	5,93	-0,33	0,112	-3,75	14,063
4	41	6,7	6,10	0,60	0,361	-2,65	7,023
5	45	7,2	6,41	0,79	0,632	-2,15	4,623
6	62	7,6	7,46	0,14	0,020	-1,75	3,063
7	102	8,7	9,10	-0,40	0,157	-0,65	0,423
8	145	9,4	10,25	-0,85	0,728	0,05	0,003
9	155	9,8	10,47	-0,67	0,453	0,45	0,203
10	176	10,2	10,89	-0,69	0,477	0,85	0,722
11	210	11,3	11,47	-0,17	0,029	1,95	3,803
12	270	11,9	12,30	-0,40	0,159	2,55	6,503
13	358	13,7	13,23	0,47	0,225	4,35	18,923
14	370	13,9	13,33	0,57	0,320	4,55	20,703
15	408	14,2	13,66	0,54	0,296	4,85	23,523
Σ	--	140,3	140,8 4	--	6,802	--	144,19 8
átla g	--	9,35	--	--	--	--	--

A 126. táblázat adatait helyettesítsük be a korrelációs index képletébe:

$$I = \sqrt{1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} = \sqrt{1 - \frac{6,802}{144,198}} \cong 0,9761.$$

Az eredmény nagyon erős nemlineáris korrelációs kapcsolatra utal. Ez azt jelenti, hogy a kocák testtömeg-termelése és a létszám között szoros kapcsolat áll fenn.

Exponenciális regresszió

Két kvantitatív változó közötti kapcsolat exponenciális függvénnyel adható meg, ha a független x változó egységnyi növekedése hatására a függő y változó konstans értékkel szorozódik, vagyis konstans százalékos értékkel változik.

Exponenciális függvénnyel írható le pl. a természet számos törvényszerűsége, a biológiában általában a fejlődés kezdeti szakasza jellemezhető exponenciális függvénnyel. Ekkor a független változó általában az idő, a vizsgált élőlény életkora, a fejlődés egymást követő szakaszai stb., míg a függő változó a mért tulajdonság.

Az exponenciális regresszió bemutatásához vizsgáljuk meg, hogy a talaj különböző mélysége (cm) és a búza gyökértömege (g/m^2) között milyen jellegű kapcsolat van. Az adatokat a 127. táblázat tartalmazza.

127. táblázat. A feladathoz tartozó adattáblázat

Mélység (cm)	Gyökértömeg (g/m^2)
0–10	26,87
10,1–20	15,66
20,1–30	6,18
30,1–40	2,9
40,1–50	1,5
50,1–60	0,65
60,1–70	0,28
70,1–80	0,13
80,1–90	0,06
90,1–100	0,03

Forrás: SVÁB JÁNOS (1981), 381. o.

A talaj mélységéhez tartozó adatok intervallumban és cm mértékegységben, míg a gyökérsúlyhoz tartozó adatok g/m^2 -ben vannak megadva. A

talajmélységhez tartozó adatokat alakítsuk át konkrét értékekre úgy, hogy a 0–10 cm-es intervallumhoz rendeljük az 1 dm, a 10,1–20 cm-es intervallumhoz a 2 dm stb. adatokat. A gyökértömeg adatokat váltsuk át mg/m²-re (128. táblázat).

128. táblázat. A feladathoz tartozó adatok a transzformáció után

Mélység (dm) x	Gyökértömeg (mg/m ²) y	$\ln y$
1	26870	10,20
2	15660	9,66
3	6180	8,73
4	2900	7,97
5	1500	7,31
6	650	6,48
7	280	5,63
8	130	4,87
9	60	4,09
10	30	3,40

Ezeket az adatokat vigyük be az SPSS táblába és mentjük el „*Exponencialis_regresszio.sav*” név alatt. Elsőként ábrázoljuk pontdiagramon a pont-párokat (90. ábra).

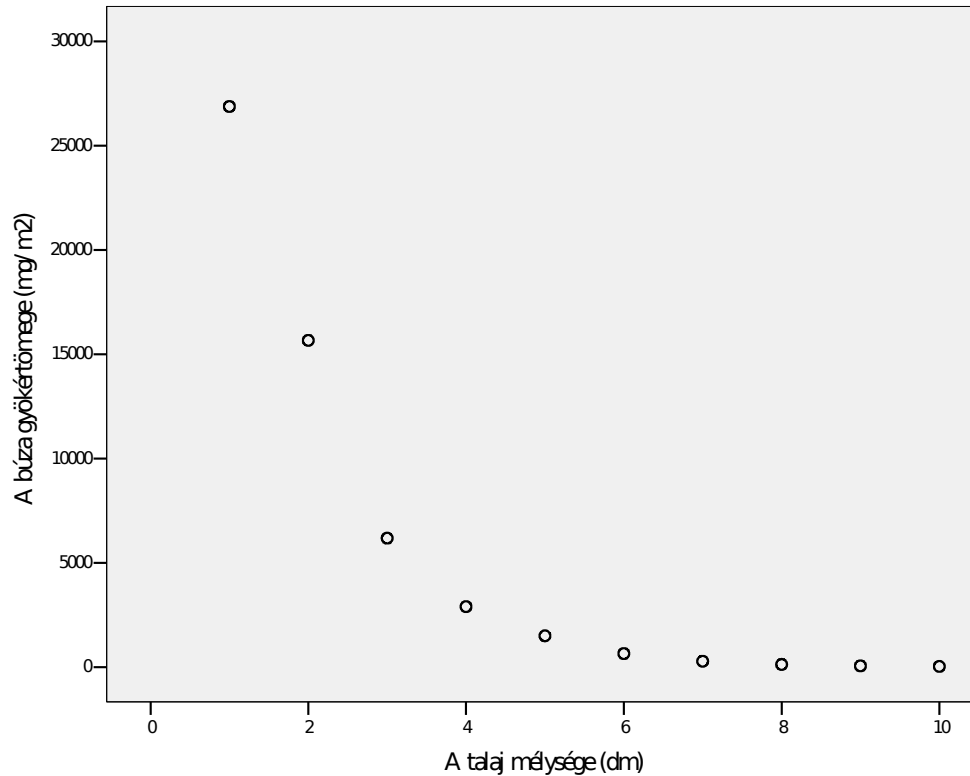
A kapott pontok elhelyezkedése alapján látható, hogy a ponthalmazra leginkább illeszkedő függvény ebben az esetben az exponenciális függvény. Az exponenciális regressziós függvény alakja: $\hat{y} = \hat{\beta}_0 \cdot \hat{\beta}_1^x$. A függvény logaritmikus transzformáció segítségével a következő lineáris összefüggésé alakítható:

$$\log \hat{y} = \log \hat{\beta}_0 + x \cdot \log \hat{\beta}_1 .$$

(A transzformációhoz tetszőleges alapú logaritmust használhatunk.) Az exponenciális regressziós függvény paramétereit kézi számításokkal úgy határozhatjuk meg, hogy alkalmazzuk a lineáris regressziónál tanultakat a transzformált változókra, majd a kapott eredményeket visszatranszformáljuk.

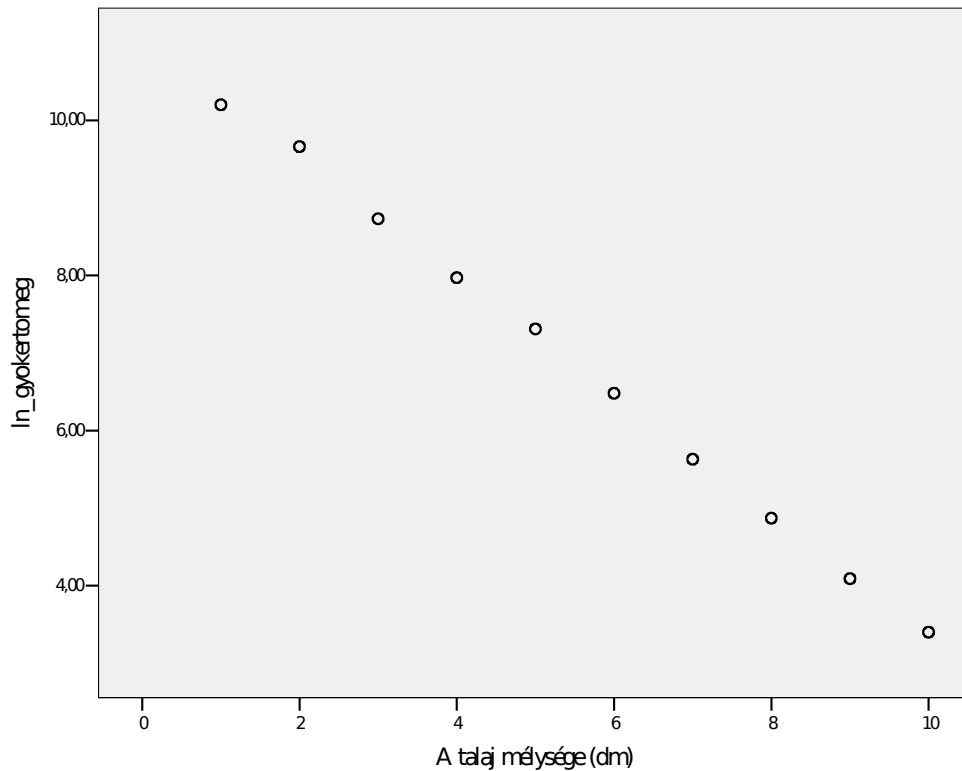
Könnyen ellenőrizhetjük, hogy a változók közötti kapcsolat valóban az exponenciális függvénnyel közelíthető-e. Ehhez vegyük az y értékeknek pl. a természetes alapú logaritmusát ($\ln y$) és az x , valamint $\ln y$ adat-párokból készítsünk grafikont. Amennyiben a vízszintes tengely beosztása az x változó természetes léptéke, a függőleges beosztása pedig a logaritmus és a kapott pont-párok egy képzeletbeli egyenes mentén helyezkednek el, akkor az

összefüggés exponenciális (91. ábra). Ha az adat-párok elhelyezkedésében „hajlás” található, akkor az összefüggés nem exponenciális, más regressziós függvényt kell keresni.

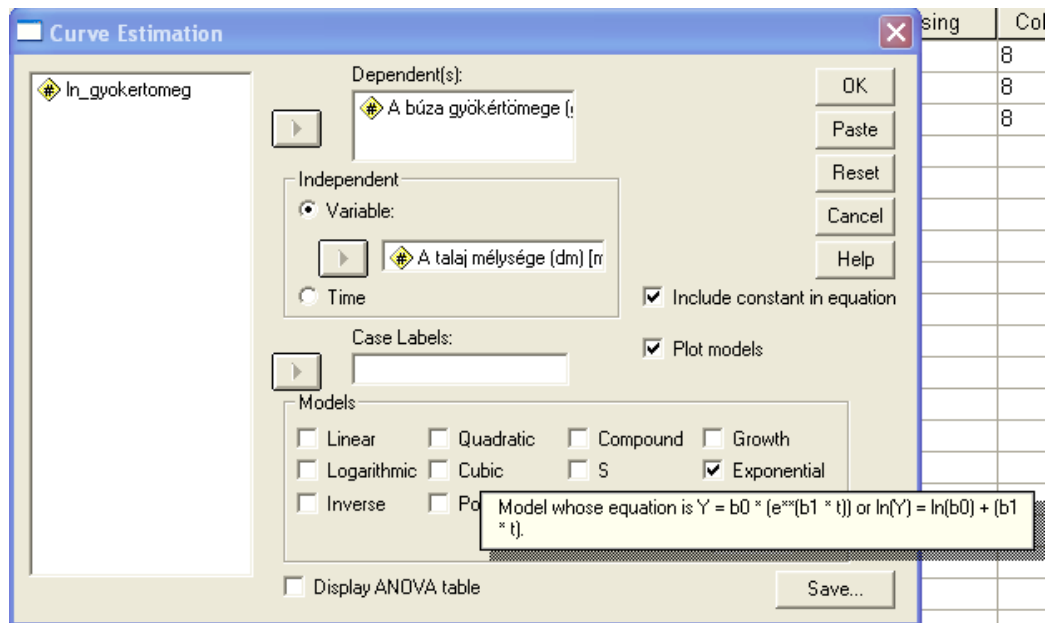


90. ábra. A változók közötti pontdiagram

Az SPSS-ben az exponenciális illesztést ugyanúgy végezzük, mint azt tettük pl. a logaritmikus regressziófüggvény alkalmazásánál. Kattintsuk végig az alábbi parancssort: `ANALYZE/REGRESSION/CURVE ESTIMATION....` A megjelenő panelban (92. ábra) válasszuk ki a független változót („A talaj mélysége”) amit a `VARIABLE` mezőbe helyezünk, a függő változót („A búza gyökértömege”) a `DEPENDENT(S)` ablakba tegyük. Most válasszuk ki az `EXPONENTIAL` függvényt, majd futtassuk le a programot.



91. ábra. A változók közötti pontdiagram



92. ábra. Az exponenciális regresszió beállításai

Az exponenciális regressziós függvény illesztése után elemezzük a 129. táblázatot.

129. táblázat. Az exponenciális regressziós függvény illesztéséhez tartozó táblázat

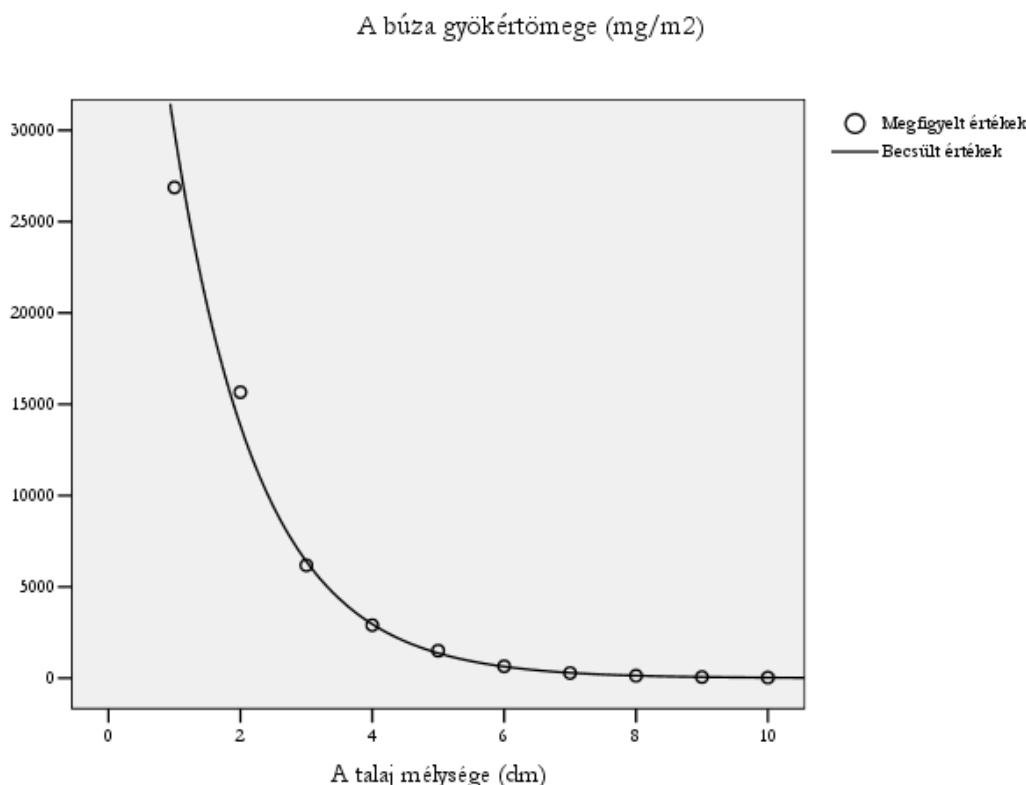
Model Summary and Parameter Estimates

Dependent Variable: A búza gyökértömege (g/m²)

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Exponential	,999	8872,148	1	8	,000	64706,304	-,771

The independent variable is A talaj mélysége (dm).

Az első oszlopban lévő determinációs érték ($r^2 = 0,999$) alapján azt mondhatjuk, hogy, a modell majdnem 100%-ban tudja magyarázni az Y értékek eltérés négyzetösszegét. A szignifikancia oszlopában a regressziós modell helyességét tesztelve kapjuk, hogy az exponenciális modell megfelelően írja le a vizsgált jelenséget ($p < 0,05$). Az utolsó két oszlopból tudjuk megadni a keresett paraméterértékeket.



93. ábra. Az exponenciális regresszió függvény

A konstans (Constant) oszlopban a $\hat{\beta}_0$ együttható értékét látjuk, míg a b_1 oszlophoz tartozó értékből a $\hat{\beta}_1$ paramétert a $\hat{\beta}_1 = e^{b_1}$ helyettesítéssel kapjuk, ugyanis az SPSS az exponenciális függvényt $y = b_0 \cdot e^{b_1 \cdot t}$ alakban illeszti. Így a

becsült regressziós paraméterek $\hat{\beta}_0 = 64706,304$ és $\hat{\beta}_1 = 0,463$, ezzel az illetet regressziós függvény: $\hat{y} = 64706,304 \cdot 0,463^x$.

A kapott regressziós egyenlet alapján azt mondjuk, hogy a gyökérsúly 10 cm-enként a megelőző 10 cm gyökérsúlyának 0,463-szorosa, vagyis kevesebb, mint a fele.

Az exponenciális regressziós függvényt a 93. ábra mutatja.

Hatványkitevős regresszió

Hatványkitevős regressziót két kvantitatív változó között akkor alkalmazunk, ha a független x változó szorzatos (százalékos) növekedésével a függő y változó is szorzatosan (százalékosan) változik. A hatványfüggvény szerinti regressziós kapcsolatot könnyen felismerhetjük, ugyanis akkor találkozunk vele, amikor mindkét változó pl. időegységben, évenként stb. exponenciálisan változik.

Hatványfüggvény-kapcsolat szokott lenni pl. különböző testrészek fejlődése között, a kórokozók terjedése és a szimptomák, valamint a betegségek szimptomái és az okozott kár között stb.

A hatványkitevős regressziófüggvény alakja:

$$\hat{y} = \hat{\beta}_0 \cdot x^{\hat{\beta}_1}.$$

Ezt a függvényt elsősorban akkor használjuk, ha az x és y változók logaritmusai között van lineáris összefüggés. A $\hat{y} = \hat{\beta}_0 \cdot x^{\hat{\beta}_1}$ összefüggést logaritmikus transzformáció segítségével visszavezethetjük lineáris alakúra. Ha mindkét oldalnak vesszük a logaritmusát, akkor az alábbi összefüggéshez jutunk:

$$\log \hat{y} = \log \hat{\beta}_0 + \hat{\beta}_1 \cdot \log x.$$

Ha bevezetjük a következő jelöléseket: $\log \hat{y} = \hat{y}^*$, $\log \hat{\beta}_0 = \hat{\beta}_0^*$ és $\log x = x^*$, akkor a regressziós függvény az alábbi alakban írható fel:

$$\hat{y}^* = \hat{\beta}_0^* + \hat{\beta}_1 \cdot x^*.$$

A transzformált modell megoldása után a $\hat{\beta}_0$ értéket kell a $\log \hat{\beta}_0$ megfelelő alapú hatványozásával kiszámítani, ugyanis a $\hat{\beta}_1$ -et közvetlenül megkapjuk.

Mivel a $\hat{y}^* = \hat{\beta}_0^* + \hat{\beta}_1 \cdot x^*$ egyenlet „hasonlatos” a lineáris regressziónál kapott $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ egyenlethez, így a $\hat{\beta}_0^*$ és $\hat{\beta}_1$ becslések ha a számításokat kézzel végeznénk, az alábbi normál-egyenletekből nyernénk

$$\hat{\beta}_0^* \cdot n + \hat{\beta}_1 \cdot \sum_{i=1}^n x_i^* = \sum_{i=1}^n \hat{y}_i^*$$

$$\hat{\beta}_0^* \cdot \sum_{i=1}^n x_i^* + \hat{\beta}_1 \cdot \sum_{i=1}^n (x_i^*)^2 = \sum_{i=1}^n x_i^* \cdot \hat{y}_i^* .$$

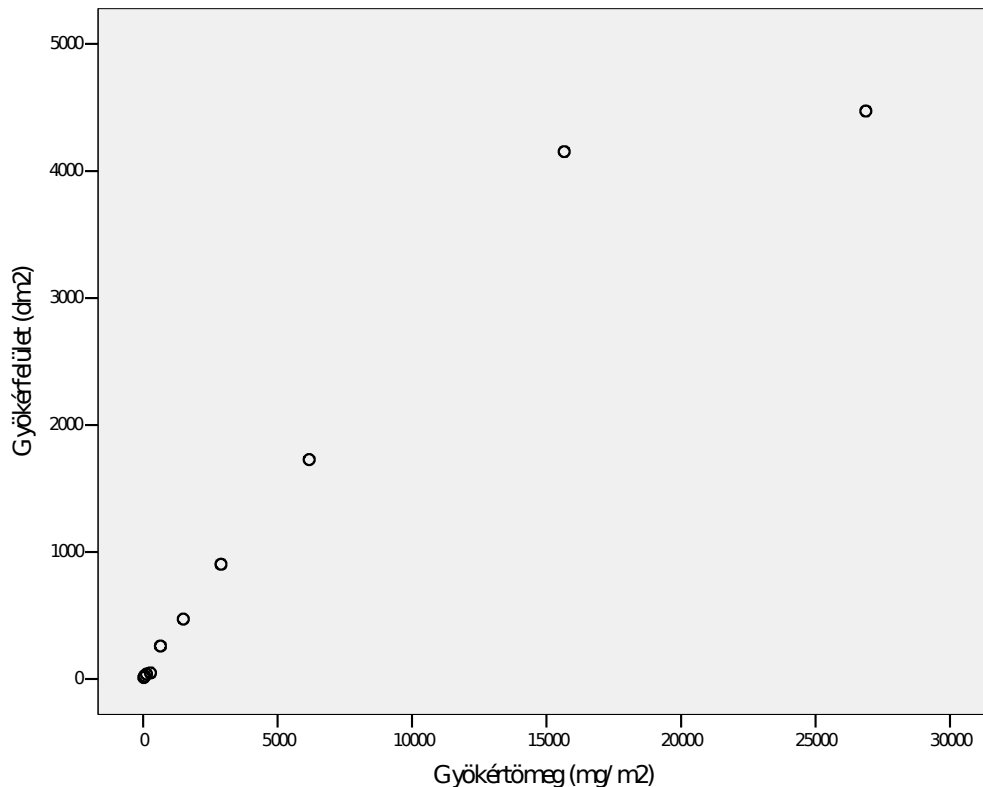
A hatványkitevős regresszió vizsgálatához bővítsük ki az előző feladatunkat azzal, hogy megadjuk a gyökértömeghez tartozó gyökérfelület-adatokat. Nézzük meg, hogy milyen összefüggésben van egymással a gyökértömeg és a gyökérfelület a talaj különböző szintjeiben. A gyökértömeg (mg/m^2) változó a független változó (x), a gyökérfelület (dm^2) pedig a függő (y) változó (130. táblázat).

130. táblázat. Alaptáblázat a hatványfüggvény szerinti összefüggés-vizsgálathoz

Mélység (cm)	Gyökértömeg (mg/m^2) x	Gyökérfelület (dm^2) y
0–10	26870	4472
10,1–20	15660	4152
20,1–30	6180	1728
30,1–40	2900	904
40,1–50	1500	472
50,1–60	650	260
60,1–70	280	48
70,1–80	130	39
80,1–90	60	24
90,1–100	30	12

Forrás: SVÁB JÁNOS (1981), 386. o.

Az adatokat tartalmazó „Hatvany_regresszio.sav” fájlt nyissuk meg és ábrázoljuk az adat-párok összefüggését (49. ábra).

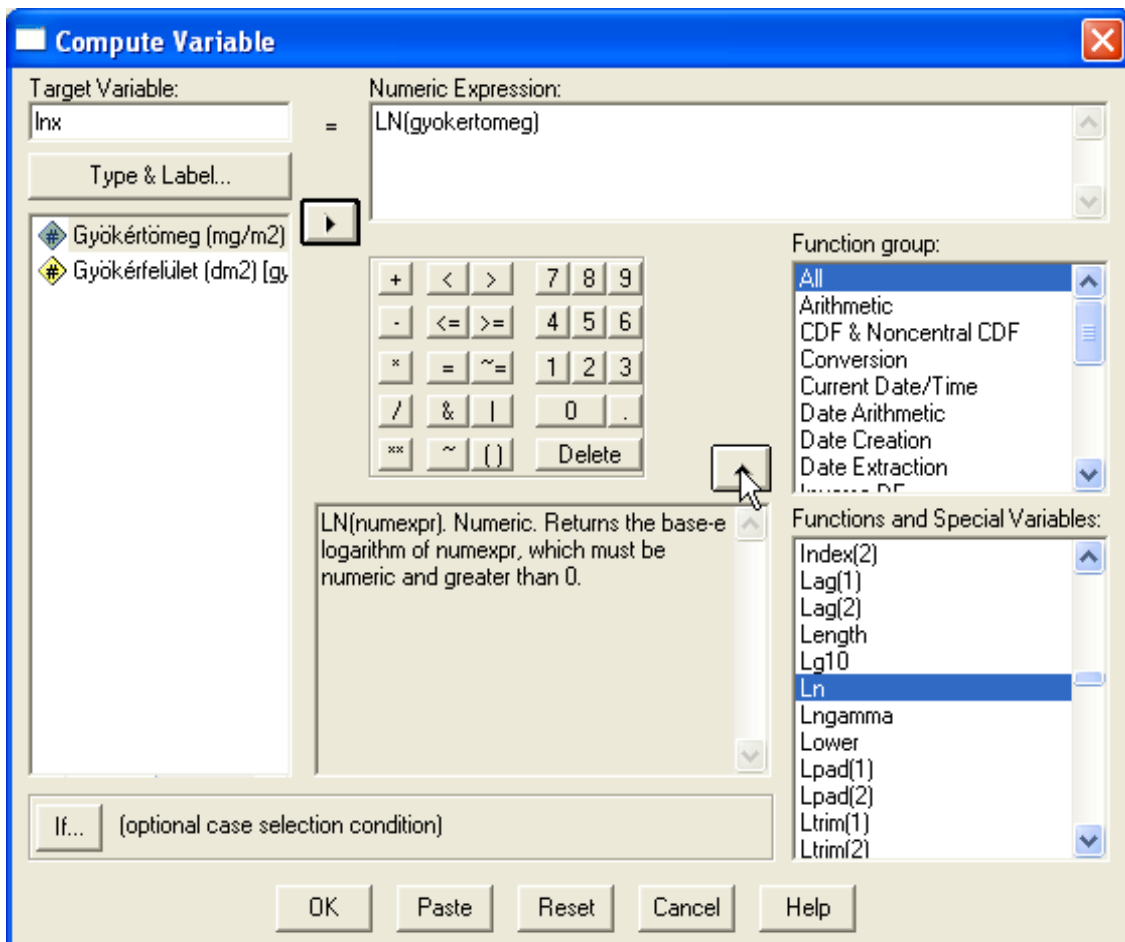


94. ábra. Az adat-párok pontdiagramja

Megvizsgálva a 49. ábrát látható, hogy a kisebb értékek nagyon összetömörülnek, és csak nehezen vehető ki az adat-párok elhelyezkedése. Ilyenkor célszerű az x és y értékek logaritmussá alakított értékeiből készíteni ábrát ($\ln x$ -ből és $\ln y$ -ből), amelynek egyenest kell adnia. Ha ugyanis az adat-párok nem egy képzeletbeli egyenes mentén helyezkednek el, hanem görbe vonalat mutatnak, akkor az összefüggés törvényszerűsége nem követi a hatványfüggvényt, és más függvényt kell választani.

EI kell készíteni a változók természetes alapú logaritmusait, amihez kattintsunk a TRANSFORM menü COMPUTE... parancsára.

A megjelent panelban (95. ábra) a TARGET VARIABLE mezőbe írjuk az új változó nevét, ami először legyen „lnx”. Ezt követően a FUNCTION GROUP ablakban válasszuk ki az ALL funkciót, ami azokat a függvényeket és speciális változókat jelenít meg, amelyek be vannak építve az SPSS-be. Ezek közül válasszuk ki a természetes alapú logaritmus függvényt (LN), majd a nyilacska segítségével helyezzük ezt a függvényt a NUMERIC EXPRESSION ablakba.



95. ábra. A Transform menü Compute... parancsa, ahol új változókat definiálunk

Ekkor az LN szimbólum után megjelenik egy zárójel, ahová helyezzük a bal oldalon lévő változókat tartalmazó ablakból a „gyökértomeg” változót. A beállítások után az OK gombbal hagyjuk jóvá az új változó definiálását, aminek következtében az adatmátrixban megjelenik az új változónk az értékeivel együtt. Ugyanezt a műveletet végezzük el az $\ln y$ függvény elkészítésére is, majd nézzük meg a 96. ábra adatmátrixának alakulását.

A kapott két új változóval készítsünk pontdiagramot, melynek eredményét az 97. ábra mutatja.

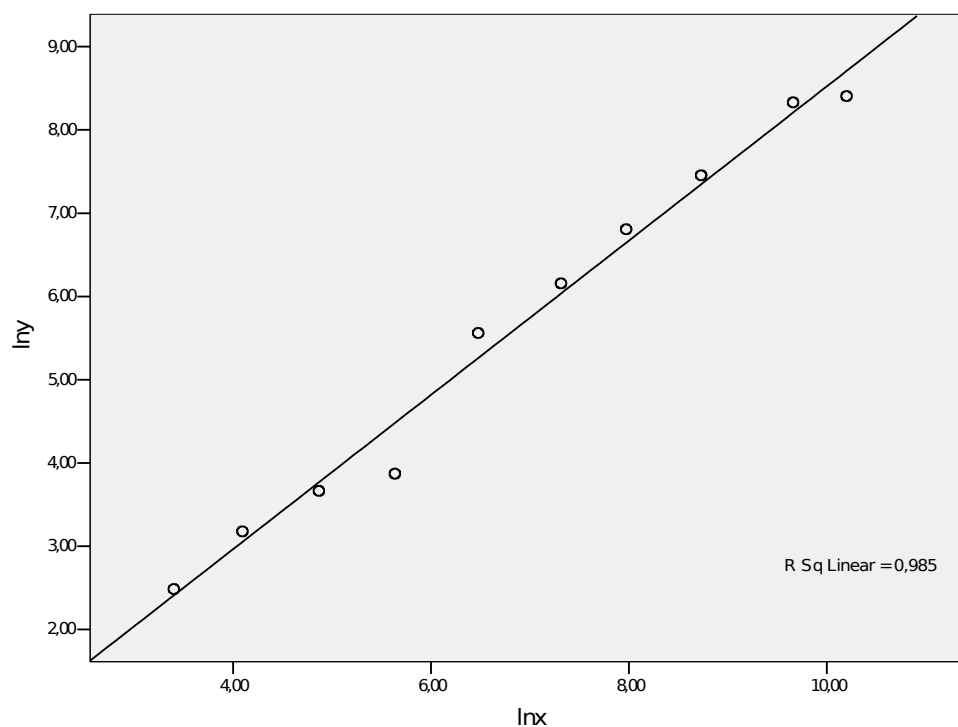
hatvany_regresszio.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : gyokertomeg 26870

	gyokertomeg	gyokerfelulet	lnx	lny	var
1	26870	4472	10,20	8,41	
2	15660	4152	9,66	8,33	
3	6180	1728	8,73	7,45	
4	2900	904	7,97	6,81	
5	1500	472	7,31	6,16	
6	650	260	6,48	5,56	
7	280	48	5,63	3,87	
8	130	39	4,87	3,66	
9	60	24	4,09	3,18	
10	30	12	3,40	2,48	

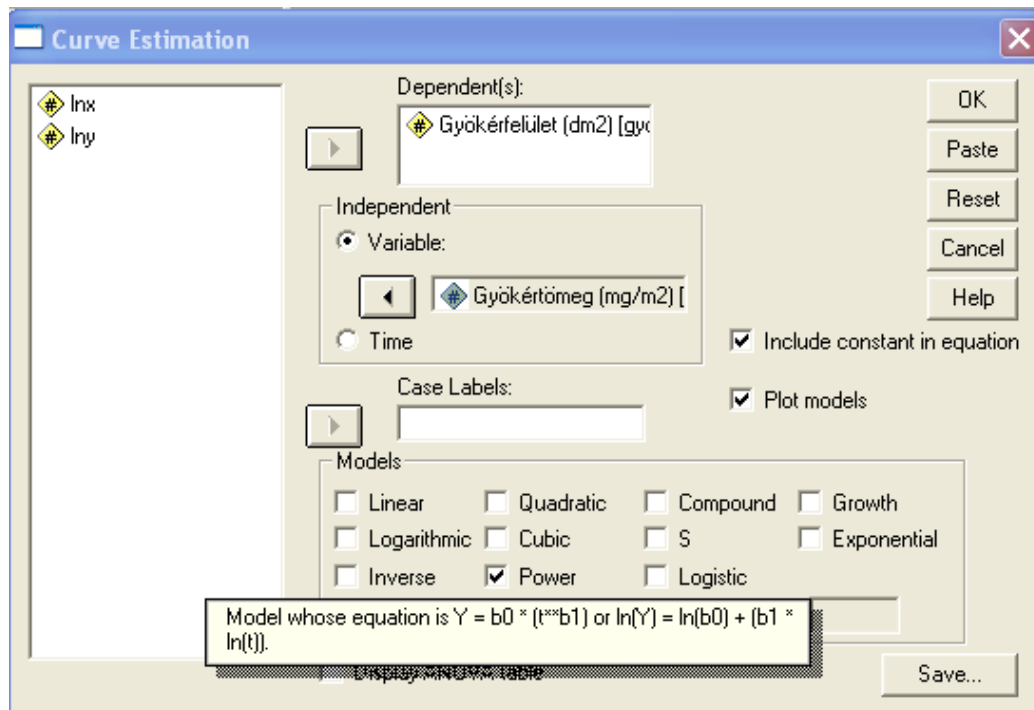
96. ábra. Az új változók definiálása után az SPSS Data View ablaka



97. ábra. A változók (gyökértömeg és felület) természetes alapú logaritmusai közötti összefüggése.

Az ábra alapján elmondhatjuk – a fentiekkel összhangban –, hogy a hatványkitevős regresszió függvény illesztése megfelelőnek tűnik.

Végezzük el a regressziós függvény illesztését, amihez kattintsunk az ANALYZE menüpont REGRESSION almenüjének CURVE ESTIMATION... parancsára. A megjelent panelban (98. ábra) a VARIABLE mezőbe helyezük a független változót, ami a „gyöktömeg”, míg a DEPENDENT(S) ablakba a „gyökérfelület” változót helyezük. A MODELS részben most a POWER függvényt választjuk, ez a hatványkitevős regresszió-függvény.



98. ábra. A hatványkitevős regresszió-függvény illesztésnek beállítása

A beállítások és a számítások elvégzése után nézzük meg a kapott eredményt a 131. táblázat alapján. Ebből a táblázatból olvashatjuk azt ki, hogy megfelelő-e a modellünk, valamint hogyan alakulnak a paraméterértékek.

131. táblázat. A hatványkitevős regresszió paraméterei

Model Summary and Parameter Estimates

Dependent Variable: Gyökérfelület (dm2)

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Power	,985	517,285	1	8	,000	,476	,927

The independent variable is Gyökértömeg (mg/m2).

A paraméterek becslése az utolsó két oszlopból olvasható le: $\hat{\beta}_0 = 0,476$ és $\hat{\beta}_1 = 0,927$, így a keresett hatványkitevős regresszió függvény alakja:

$$\hat{y} = 0,476 \cdot x^{0,927}$$

A determinációs együttható értéke 0,985, ez alapján azt mondhatjuk, hogy a modell közelítőleg 99%-ban tudja magyarázni az Y értékek eltérés négyzetösszegét. A szignifikancia oszlopában a regressziós modell helyességét tesztelve azt kapjuk ($p < 0,05$), hogy a hatványkitevős modell megfelelően írja le a vizsgált jelenséget. Elmondhatjuk még, mivel $\hat{\beta}_1$ értéke kisebb mint 1, hogy a gyökérfelület növekedése lassúbb, mint a gyökértömeg növekedése.

Parabolikus regresszió

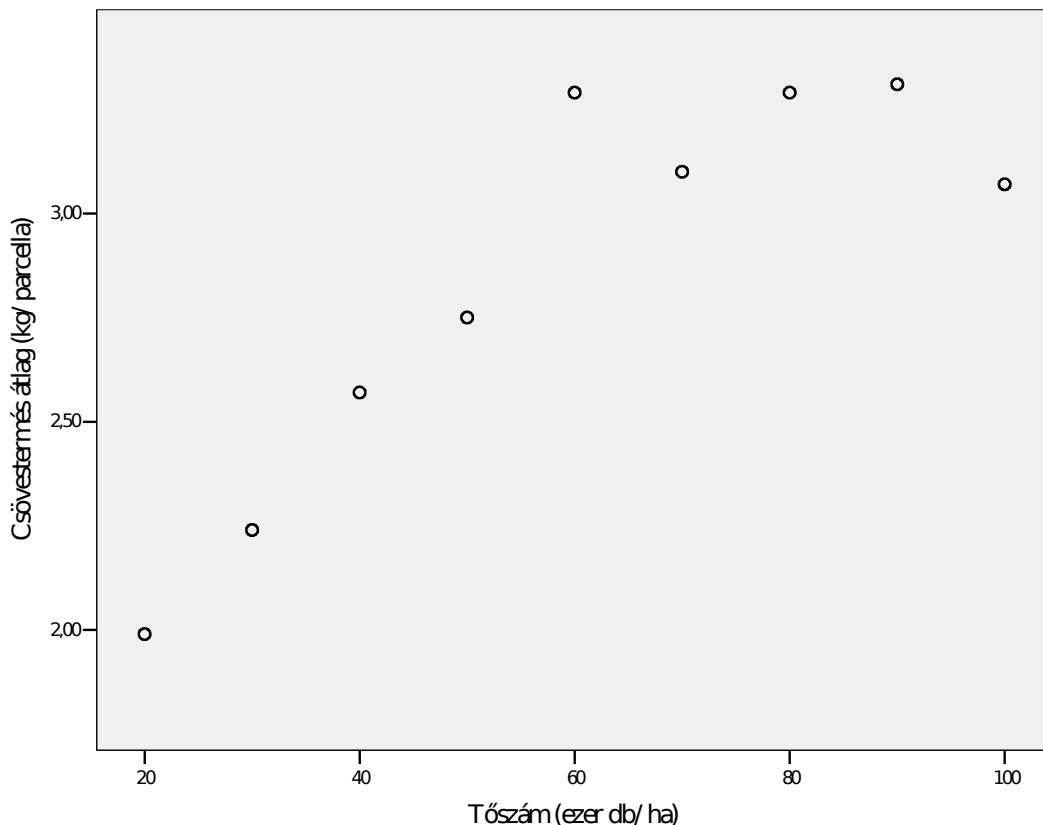
A parabolikus regresszió vizsgálatára használjuk a 132. táblázat adatait. Vizsgáljuk meg a kukorica tőszámnövekedésének a hatását a csöves termés mennyiségére. A gyakorlat azt mutatja, ha egy adott területen növeljük a tőszámot, a termésmennyiség egy bizonyos pontig növekszik, azután csökken. A kérdés azonban az, hogy meddig lehet sűríteni a kukoricát terméscsökkenés nélkül.

132. táblázat. A feladathoz tartozó adattáblázat

Tőszám (ezer db/ha)	Csőves termésátlag (kg/parcella) Sze SC 352,FAO 340
20	1,99
30	2,24
40	2,57
50	2,75
60	3,29
70	3,10
80	3,29
90	3,31
100	3,07

Forrás: HUZSVAI L. (2003)

Az SPSS-ben is készítsük el az adatfájlt, ennek neve legyen „Parabolikus_regresszio”. Első lépésként ábrázoljuk a pont-párokat, ezt a 99. ábra tartalmazza.



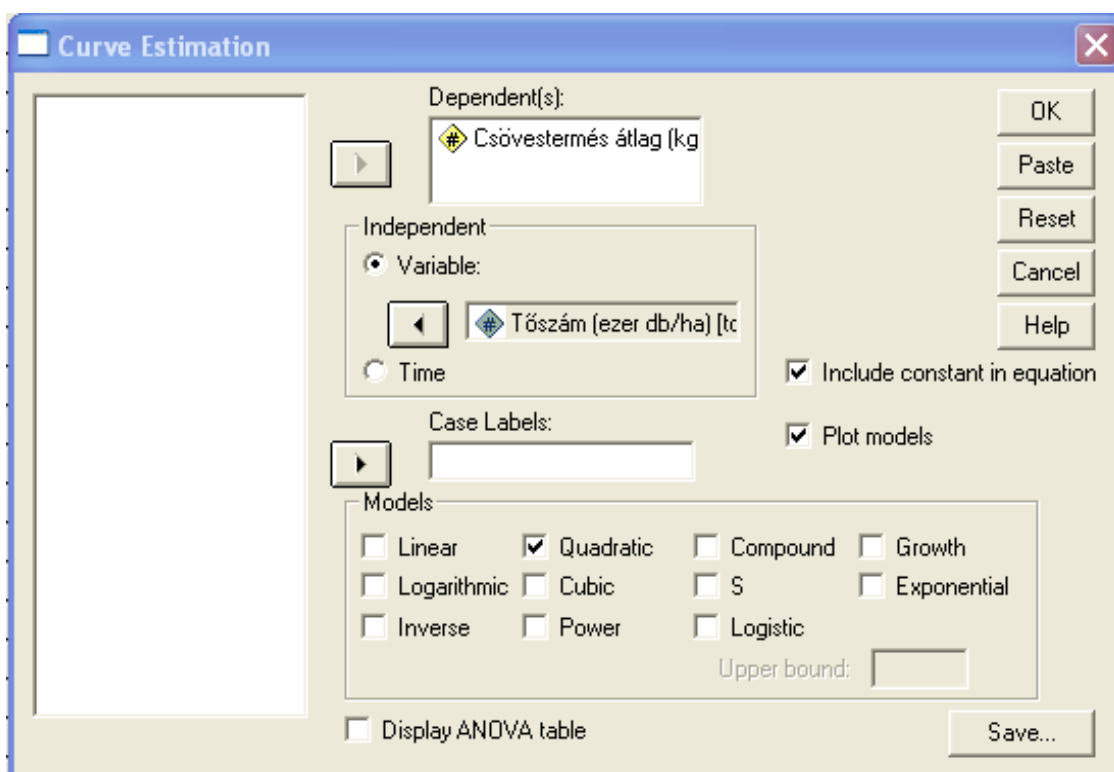
99. ábra. A változók közötti pontdiagram

A pontfelhő elhelyezkedése alapján leginkább a másodfokú függvény illeszkedik a ponthalmazra. A parabolikus regressziós függvény alakja:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \hat{\beta}_2 \cdot x^2.$$

A regressziós függvény illesztéséhez kattintsunk az **ANALYZE** menüpont **REGRESSION** almenüjének **CURVE ESTIMATION...** parancsára, ahol a megjelent panelban a *100. ábra* szerint végezzük el a beállításokat.

A modellhez tartozó determinációs együttható értéke 0,974, ami azt jelenti, hogy közel 97%-ban tudja magyarázni a modell az Y értékek eltérés négyzetösszegét. A szignifikancia oszlopában a regressziós modell helyességét tesztelve kapjuk, hogy a másodfokú modell megfelelően írja le a vizsgált jelenséget.



100. ábra. Parabolikus regressziós függvény illesztése

Az utolsó három oszlopból tudjuk megadni a keresett paramétereket (133. táblázat).

133. táblázat. A parabolikus regresszió paraméterei

Model Summary and Parameter Estimates

Dependent Variable: Csövestermés átlag (kg/parcella)

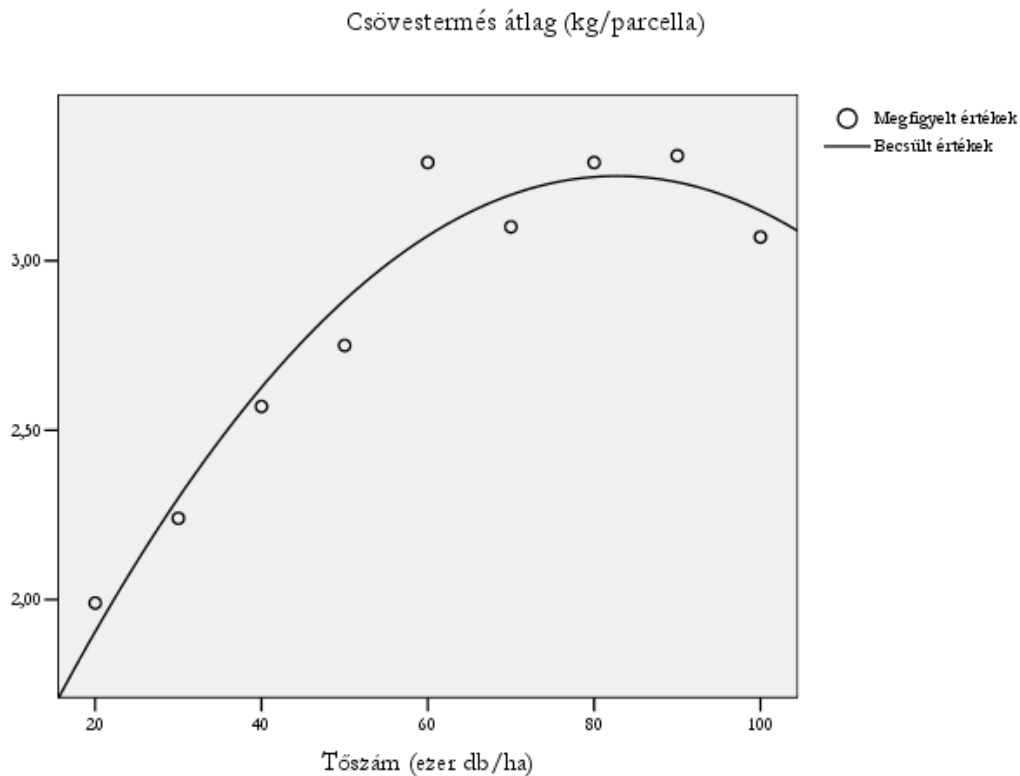
Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	,947	53,340	2	6	,000	,910	,057	-,00034

The independent variable is Tőszám (ezer db/ha).

A számított paraméterek alapján a parabolikus regressziófüggvény:

$$\hat{y} = 0,91 + 0,057 \cdot x - 0,00034 \cdot x^2$$

alakban írható fel, míg az illesztés eredményeképpen az 101. ábra mutatja a parabolikus regressziós függvényt. A függvény alakjából azt a következtetést vonhatjuk le, hogy a hektáronkénti tőszám növelése csak egy bizonyos pontig jár együtt a hektáronkénti csöves termés mennyiségének növekedésével. Amennyiben meghatározzuk a függvény maximum pontját, megállapíthatjuk, hogy mennyi az a hektáronkénti tőszámérték, ami még a termést növeli.



101. ábra. Az empirikus adatokra illesztett parabolikus regressziófüggvény

A függvény szélsőérték helyének meghatározása differenciál-számítás segítségével történik. Egy függvénynek ott lehet szélsőérték helye, ahol az első deriváltja nulla. Ismerjük az illesztett függvényt: $\hat{y} = 0,91 + 0,057 \cdot x - 0,00034 \cdot x^2$, ezt kell deriválnunk. A derivált-függvény: $\hat{y}' = 0,057 - 0,00068 \cdot x$. Egyenlővé téve a kifejezést nullával, majd megoldva az egyenletet az $x = 83,82$ értéket kapjuk. Ebben a pontban a függvénynek akkor van biztosan szélsőérték helye, ha a második derivált értéke nem nulla. Mivel a második derivált ezen a helyen kisebb, mint nulla, így a másodfokú függvénynek ezen a helyen maximuma van.

Határozzuk meg a 83,82 ezer/ha tőszámhoz tartozó csöves termés mennyiségét (kg/parcella). A tőszám értéket behelyettesítve a regressziós függvény képletébe megkapjuk azt a termésmennyiséget, ami az adott tőszámhoz társul: $0,91 + 0,057 \cdot 83,82 - 0,00034 \cdot 83,82^2 = 3,299$.

Ez azt jelenti, hogy az elérhető legmagasabb termés 3,299 kg/parcella csöves termés.

Lineárisra nem visszavezethető összefüggések vizsgálata

Logisztikus függvény

A biológia egyik legáltalánosabb és legfontosabb törvényszerűségét fejezi ki a logisztikus függvény, amely jellemzője, hogy a függő változó eleinte lassan, majd mind gyorsabban növekszik, aztán a növekedése lelassul és egy felső határ, a maximum felé közeledik.

134. táblázat. A növény tömegének gyarapodása

Sorszám	Eltelt napok száma (x)	A növény tömege (gramm) (y)
1	1	0,19
2	7	0,96
3	14	3,01
4	21	6,59
5	28	12,25
6	35	19,73
7	42	30,30
8	49	43,14
9	56	57,06
10	63	73,44
11	70	89,99
12	77	104,97
13	84	118,24
14	91	129,55
15	98	141,46
16	105	155,40
17	112	166,84
18	119	175,32
19	126	181,74
20	133	186,06
21	140	187,76

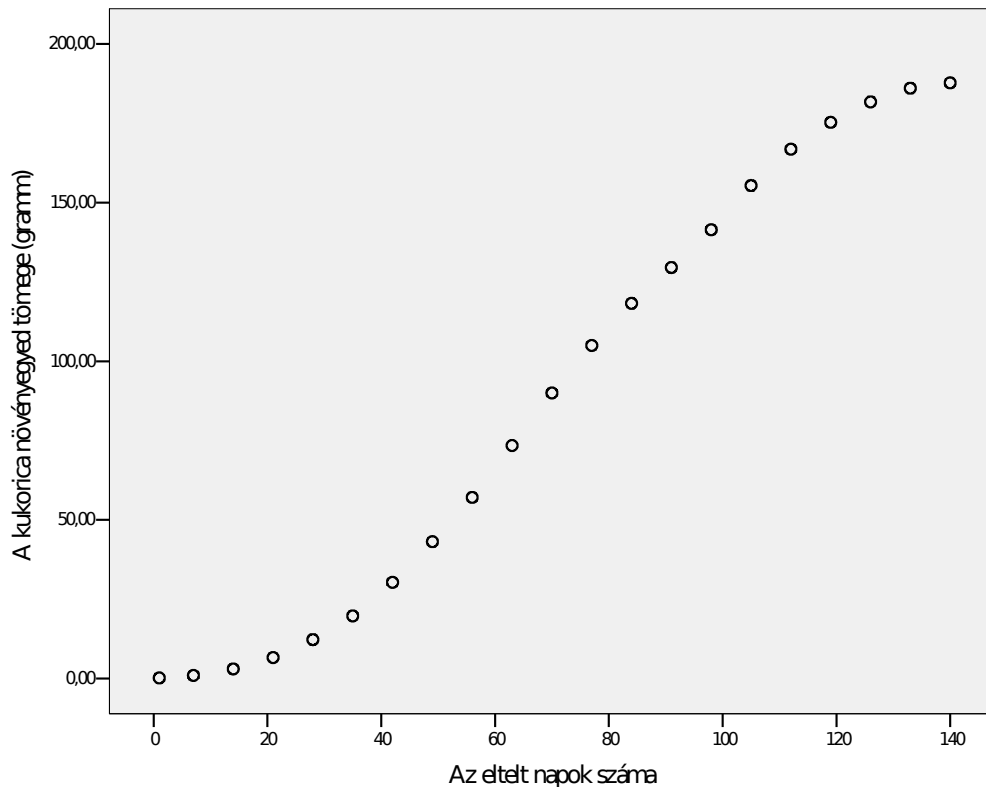
Forrás: Huzsvai L. kézirat

A logisztikus függvény megadásához az $\hat{y} = \frac{\hat{y}_{\max}}{1 + m}$ ($m = e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x}$, \hat{y}_{\max} : telítődési szint, a függvény felső aszimptotája) képlettel definiált becslőfüggvényt fogjuk használni.

Kukorica növények tömegének növekedését vizsgálták. A kelés első napjától kezdve hetente mérték a növényeket, az adatokat grammban jegyezték fel.

Milyen függvénnyel írható le a kukorica növekedése? Az adatokat az 134. táblázatban közöljük.

Készítsünk az adatokból SPSS fájlt és mentjük el „Logisztikus_regresszio.sav” név alatt. Kezdjük a vizsgálatot a változók közötti pontdiagram elkészítésével (102. ábra).



102. ábra. Kukorica növényegyed növekedése

Az adatok ábrázolása alapján a logisztikus függvénykapcsolat látszik a legmegfelelőbbnek a napok száma és a növény tömege közötti kapcsolat leírására, ugyanis a megfigyelt értékek jellemzően először lassan, majd egyre gyorsabban növekednek, azután ismét lassulnak, majd egy felső határ felé közelítenek.

A logisztikus függvény paramétereinek meghatározása

A megfigyelt ponthalmazra illeszkedő logisztikus függvényt ebben az esetben a korábban leírtakkal összhangban az:

$$\hat{y} = \frac{\hat{y}_{\max}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot x}}$$

képlet alapján keressük. (Ahogy nő az x értéke, a függvény értéke is állandóan nő, az x -nek végtelen nagy értékére a függvény határértéke \hat{y}_{\max} lesz.)

Feladatunk az lesz, hogy meghatározzuk az \hat{y}_{\max} , $\hat{\beta}_0$ és $\hat{\beta}_1$ paraméterek értékeit. Mivel a logisztikus trend paramétereinek a meghatározása a legkisebb négyzetek módszere szerint igen bonyolult, ezért egy kevésbé egzakt módszert mutatunk be, az ún. „három kiválasztott pont módszerét”. Ennek a módszernek a lényege, hogy az említett három szakaszra jellemző helyen kiválasztunk három pontot, amelyek egymástól közelítőleg azonos távolságra vannak (ezeket a pontokat a 134. táblázatban kiemelten jelöltük). A három pontot jelöljük $x_0, x_0 + m, x_0 + 2m$ szimbólumokkal, ahol m a kiválasztott pontok egymástól való távolságát jelenti és $x_0 = 0$. A pontok kiválasztás után meg kell határozni a kiválasztott pontok környezetéhez tartozó átlagos adatokat $(\bar{y}_{x_0}, \bar{y}_{x_0+m}, \bar{y}_{x_0+2m})$. Ezek után van aztán arra lehetőség, hogy a függvény paramétereit meghatározzuk, amelyhez a következő összefüggéseket használjuk fel:

$$\hat{y}_{\max} = \frac{2 \cdot \bar{y}_{x_0} \cdot \bar{y}_{x_0+m} \cdot \bar{y}_{x_0+2m} - \bar{y}_{x_0+2m}^2 \cdot (\bar{y}_{x_0} + \bar{y}_{x_0+m})}{\bar{y}_{x_0} \cdot \bar{y}_{x_0+2m} - \bar{y}_{x_0+m}^2}$$

$$\hat{\beta}_0 = \ln \left(\frac{\hat{y}_{\max} - \bar{y}_{x_0}}{\bar{y}_{x_0}} \right)$$

$$\hat{\beta}_1 = \frac{1}{m} \cdot \ln \left(\frac{\hat{y}_{x_0} \cdot (\hat{y}_{\max} - \bar{y}_{x_0+m})}{\bar{y}_{x_0+m} \cdot (\hat{y}_{\max} - \bar{y}_{x_0})} \right).$$

Első lépésként adjuk meg az önkényesen kiválasztott három pontot $(x_0, x_0 + m, x_0 + 2m)$, amelyek egymástól nagyjából azonos távolságra legyenek és úgy válasszuk meg a pontokat, hogy x_0 a kisebb, $x_0 + m$ a középső és $x_0 + 2m$ a legnagyobb értékekhez tartozzanak.. A paraméterek meghatározásához szükséges számításokat a 135. táblázatban foglaltuk össze.

135. táblázat. A logisztikus trendfüggvény illesztéséhez szükséges részeredmények

A kiválasztott pontok sorszáma	A pontok új jelölése	Átlagok ($\bar{y}_{x_0}, \bar{y}_{x_0+m}, \bar{y}_{x_0+2m}$)
2	$x_0 = 0$	$\bar{y}_0 = \frac{\frac{0,19}{2} + 0,96 + \frac{3,01}{2}}{2} \cong 1,28$
11	$x_0 + m = 9$	$\bar{y}_9 = \frac{\frac{73,44}{2} + 89,99 + \frac{104,97}{2}}{2} \cong 89,6$
20	$x_0 + 2m = 18$	$\bar{y}_{18} = \frac{\frac{181,74}{2} + 186,06 + \frac{187,76}{2}}{2} \cong 185,4$

A táblázat adatait helyettesítsük be a paramétereket megadó képletekbe:

$$\hat{y}_{\max} = \frac{2 \cdot 1,28 \cdot 89,6 \cdot 185,405 - (89,6)^2 \cdot (1,28 + 185,405)}{1,28 \cdot 185,405 - (89,6)^2} ;$$

$$\cong 189,91$$

$$\hat{\beta}_0 = \ln\left(\frac{186,91 - 1,28}{1,28}\right) \cong 4,977 ;$$

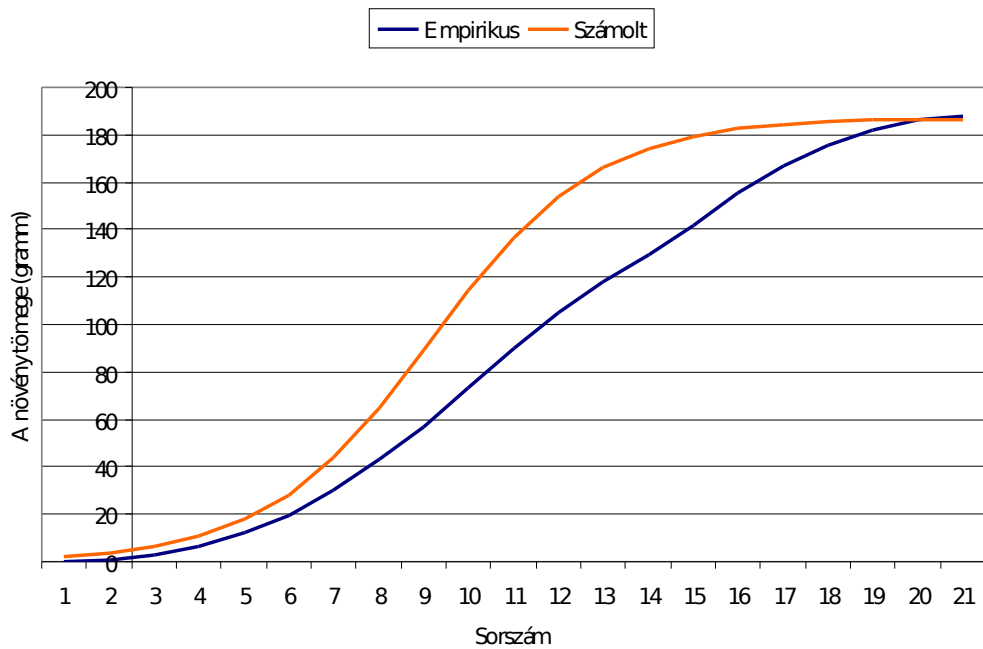
$$\hat{\beta}_1 = \frac{1}{9} \cdot \ln\left(\frac{1,28 \cdot (186,91 - 89,6)}{89,6 \cdot (186,91 - 1,28)}\right) \cong -0,543 .$$

A logisztikus trendfüggvény paraméterei: $\hat{y}_{\max} = 186,91$; $\hat{\beta}_1 = 4,977$ és $\hat{\beta}_0 = -0,543$. A paraméterekkel a logisztikus trendfüggvény:

$$\hat{y} = \frac{186,91}{1 + e^{4,977 - 0,543 \cdot x}} ,$$

ahol x a sorszám, és a kiindulópont az $x = 2$.

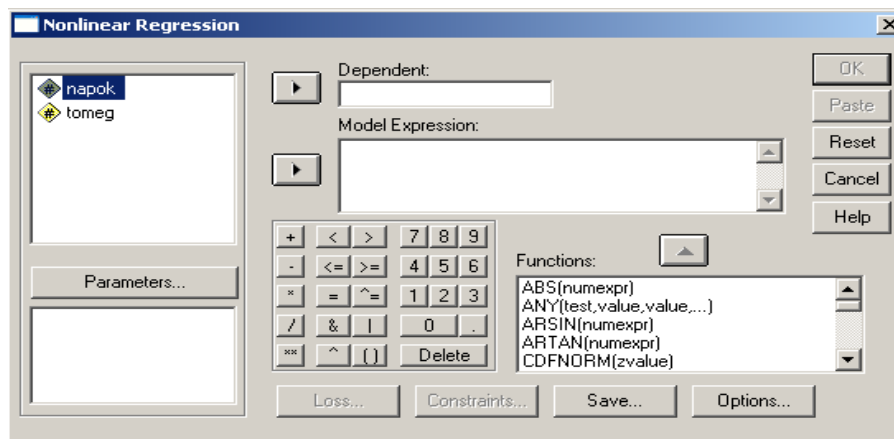
Az empirikus és a fentebb számolt függvény alapján kapott adatokat az 103. ábra mutatja.



103. ábra. Kukorica növényegyed növekedése az empirikus és a számolt adatokkal

A logisztikus regressziós-függvény meghatározásának ez a módszere önkényes elemeket tartalmazott, így joggal merül fel a kérdés, hogy mennyire megbízható ez a módszer.

Az SPSS-ben a nemlineáris regresszió elvégzéséhez kattintsunk az ANALYZE, REGRESSION, NONLINEAR...menüre. A megjelent panel (104. ábra) beállításait az alábbiak szerint végezzük.



104. ábra. Az Analyze / Regression / Nonlinear... ablak beállításai

A DEPENDENT mezőbe írjuk a függő változót (104. ábra). A MODEL EXPRESSION panelrészben a függő változó becslésére alkalmas függvényt kell megadni, aminek legalább egy független változót kell tartalmaznia.

A **PARAMETERS** ablakban azokat a paramétereket adjuk meg, amiket felhasználunk a modellben. A regressziós függvényekben szereplő paraméterek kezdeti értékeit nekünk kell megbecsülni és megadni, a program csak ezután, közelítő eljárást használva határozza meg a paraméterek legjobb értékét, úgy, hogy a hiba eltérés négyzetösszegét minimalizálja (az előző részben bemutatott kézi számítás eredményei segítséget adhatnak a paraméterek becsült értékeinek megadásához).

136. táblázat. A nemlineáris panel parancsgombjai és azok funkciója

Parancsgomb	Funkció
Loss...	PARAMETER CONSTRAINTS: Lineáris kifejezésekkel egy vagy több paraméter értékének korlátozó feltételeket adhatunk meg.
Constraints ...	LOSS FUNCTION: A regressziós egyenlet meghatározásának módját határozhatjuk itt meg. Alapesetben a maradékok eltérés négyzetösszegének minimalizálásával folyik a regressziós egyenlet meghatározása. Lehetőség van általunk készített és definiált módszerrel is meghatározni a regressziós függvényt (USER-DEFINED LOSS FUNCTION). Ha pl. a keresett függvény képe $f = p_1 + p_2 \cdot x$, akkor a legkisebb négyzetes eltérést így adhatjuk meg $x_{\text{loss}} = [y - (p_1 + p_2 \cdot x)]^2$, az illesztés során ezen értékek összegének a minimalizálása folyik. Az x_{loss} -ba csak paraméterek is tartalmazó függvényt érdemes megadni, hisz a számítások során a p értékek változnak, és csak ezek tudnak konvergálni a megadott feltételek szerint. Az SPSS-ben a változók között megtalálhatók a becsült és maradék értékek, PRED_ és RESID_ jelöléssel.
Save...	Ezzel a paranccsal elmenthetjük a becsült, maradék, derivált és ha volt az x_{loss} függvény értékeit. Ezek az értékek új változóként az adatbázisban megjelennek, és további elemzést végezhetünk rajtuk.
Options...	Az OPTIONS az iteráció módszerét és feltételeit állítja be. Ebben a menüpontban nem kapunk automatikusan rajtot az illesztet görbéről, azért a becsült értékek elmentése fontos. Az ábrázolást a GRAPHS/SCATTER/OVERLAY menüben végezhetjük el. A pontokat célszerű a SPLINE módszerrel összekötni.

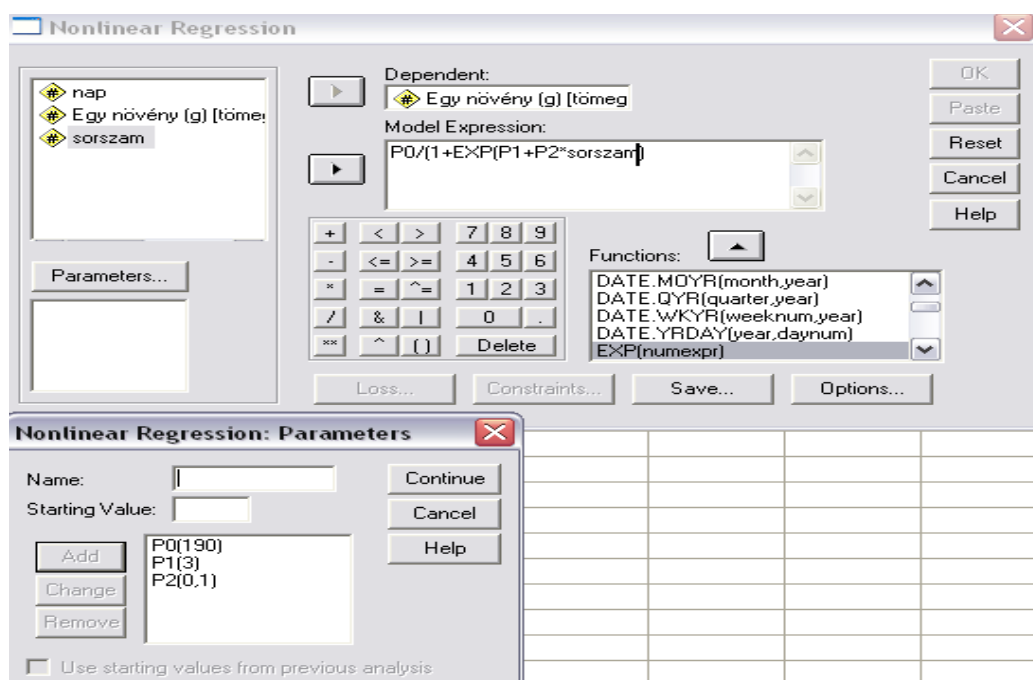
Az iterációs eljárás alkalmazása miatt az eredmény kismértékben függ a paraméterek kezdeti értékétől is.

Ha a megadott lépésszám után nem kapunk jó eredményt, vagyis az R^2 értéke nagyon kicsi, akkor érdemes az odáig kiszámított paramétereket megadni kezdő paraméterként és újratekdeni a számítást.

A 104. ábrán látható panelben alul még négy parancsgombot találunk, amelynek funkciót a következő (136. táblázatban) foglaltuk össze.

Most vegyük sorba a kísérlethez tartozó beállításokat. A DEPENDENT mezőbe a „tömeg” változót helyezzük A MODEL EXPRESSION ablakban beírjuk a logisztikus függvény képletét, amelyhez segítségül használjuk a beépített függvényeket tartalmazó FUNCTIONS parancsot. A PARAMETERS ablakba magunk adjuk meg a paraméterek értékeit, amivel a program a számításokat kezdi (105. ábra). Fontos a becsült értékek elmentése az illesztett görbe kirajzolásához, valamint mentsük el a hibtagokat is azok további vizsgálata céljából. Ezeket a SAVE... parancsgombra kattintva megjelenő panelben tehetjük meg.

A beállítások elvégzése után elemezzük az Outputban megjelent táblázatokat. Az iteráció (137. táblázat) a paraméterek általunk megadott kezdeti értékeiből indul ki, és akkor áll le, amikor a hiba eltérés négyzetösszege már csak 10^{-8} -on nagyságrendű értékkel csökkent.



105. ábra. Az ANALYSE/REGRESSION/NONLINEAR menüpont beállítása

137. táblázat. Az iteráció eredménye

Iteration History

Iteration Number ^a	Residual Sum of Squares	Parameter		
		P0	P1	P2
1.0	246851,26	190,000	3,000	,100
1.1	267535,32	2127,405	222,580	-3,304
1.2	267535,32	20784,732	104,123	-2,508
1.3	267535,32	20576,229	47,467	-2,126
1.4	1931701,85	706,070	20,449	-1,942
1.5	167850,14	477,234	1,196	-,273
2.0	167850,14	477,234	1,196	-,273
2.1	4425,993	297,872	1,060	-,221
3.0	4425,993	297,872	1,060	-,221
3.1	50134,657	100,866	,082	-,317
3.2	2912,611	267,339	,916	-,221
4.0	2912,611	267,339	,916	-,221
4.1	1897,044	219,410	,618	-,256
5.0	1897,044	219,410	,618	-,256
5.1	751,158	195,241	,343	-,314
6.0	751,158	195,241	,343	-,314
6.1	295,599	191,442	,233	-,356
7.0	295,599	191,442	,233	-,356
7.1	285,111	192,831	,236	-,355
8.0	285,111	192,831	,236	-,355
8.1	285,110	192,843	,236	-,355
9.0	285,110	192,843	,236	-,355
9.1	285,110	192,844	,236	-,355

Derivatives are calculated numerically.

- a. Major iteration number is displayed to the left of the decimal iteration number is to the right of the decimal.
- b. Run stopped after 23 model evaluations and 9 derivative evaluations because the relative reduction between successive residual squares is at most SSCON = 1,00E-008.

A P0, P1 és P2 oszlopok legutolsó sorából tudjuk leolvasni a paraméterek értékeit, ami azt jelenti, hogy az SPSS által számolt P0 érték 192,844, a P1=0,236 és P3=-0,355.

A 138. táblázat megadja a kapott paraméterek értékeit, hibáit és közli a konfidencia intervallum alsó és felső határát 95%-os megbízhatósági szinten.

138. táblázat. A paraméterek és azok standard hibái

Parameter Estimates				
Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
P0	192,844	3,229	186,061	199,627
P1	,236	,051	,129	,343
P2	-,355	,015	-,386	-,325

Ez a programrész nem számít t-próbát a paraméterekre vonatkozóan, de a konfidencia intervallum alapján lehet dönteni (138. táblázat). Mivel egyik intervallum sem tartalmazza a 0-t, így 95%-os megbízhatósági szinten egyik paraméter értéke sem lehet 0-val egyenlő.

Az ANOVA táblát a 139. táblázat tartalmazza.

139. táblázat. Az ANOVA tábla és a determinációs együttható

ANOVA			
Source	Sum of Squares	df	Mean Squares
Regression	267250,212	3	89083,404
Residual	285,110	18	15,839
Uncorrected Total	267535,322	21	
Corrected Total	98513,840	20	

Dependent variable: Egy növény (g)

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = ,997.

Nem számol F -próbát a függvény, de a kapott két szórásnégyzet hányadosából ki tudjuk az F értékét számítani:

$$F = \frac{MQ_{modell}}{MQ_{hiba}}$$

A modell miatt szórásnégyzet jóval nagyobb, mint a hiba miatti, ezért az F érték nagy lesz, vagyis miszerint az \hat{y} értékek szóródása véletlenszerű, biztosan elutasítjuk.

Az R^2 érték alapján azt mondhatjuk, hogy modellünk 99,7%-ban tudja magyarázni az \hat{y} értékek szóródását, azaz a modell nagyon jó.

Az illesztett függvény a becsült paraméterek alapján:

$$\hat{y} = \frac{192,844}{1 + e^{4,14 - 0,355 \cdot x}}$$

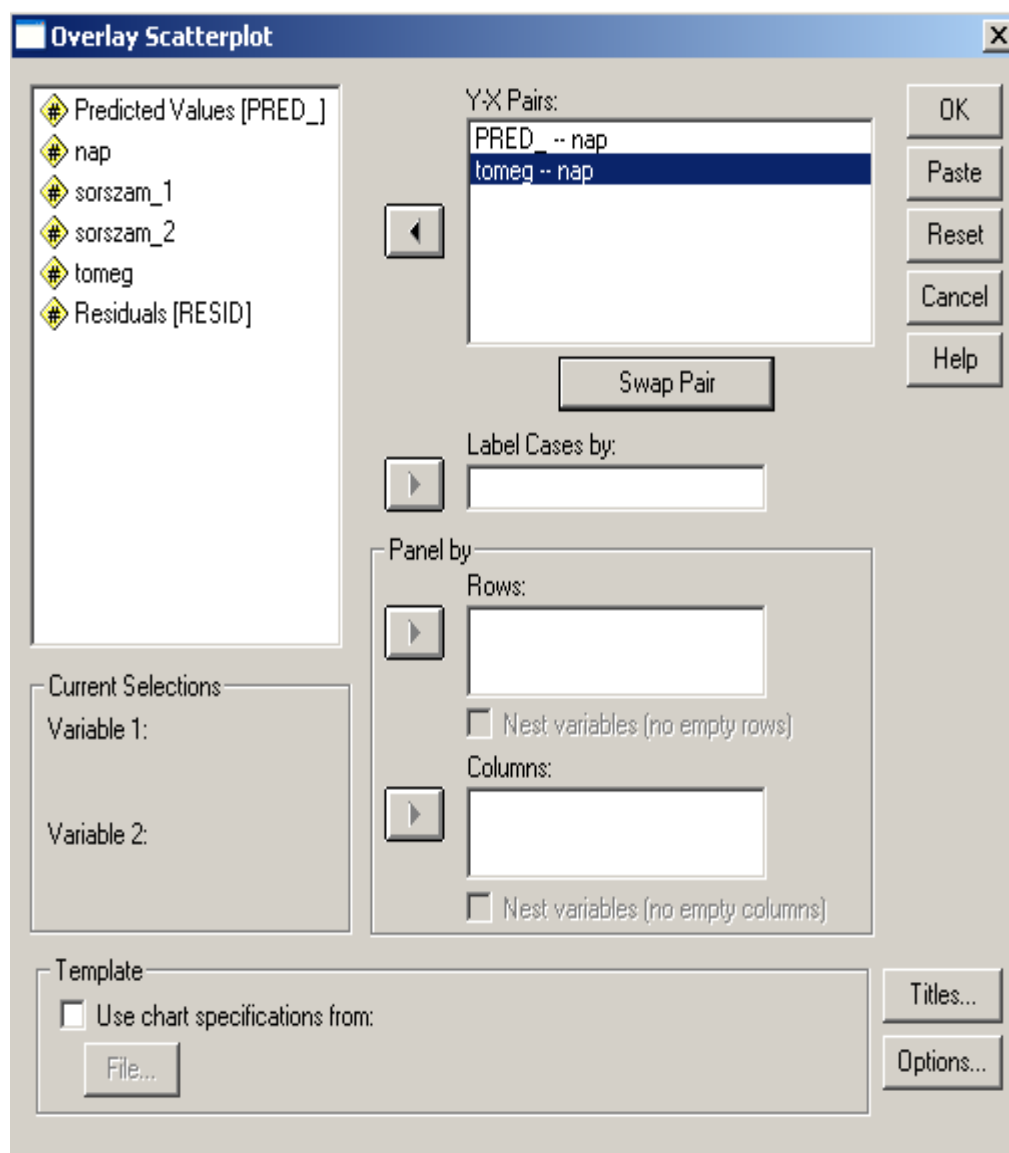
ahol x a sorszám.

	nap	sorszam_1	sorszam_2	tomeg	PRED	RESID	var
1	1	1,00	-10,00	,19	4,26	-4,07	
2	7	2,00	-9,00	,96	6,03	-5,06	
3	14	3,00	-8,00	3,01	8,48	-5,48	
4	21	4,00	-7,00	6,59	11,88	-5,28	
5	28	5,00	-6,00	12,25	16,51	-4,26	
6	35	6,00	-5,00	19,73	22,73	-3,00	
7	42	7,00	-4,00	30,30	30,87	-,58	
8	49	8,00	-3,00	43,14	41,23	1,92	
9	56	9,00	-2,00	57,06	53,90	3,16	
10	63	10,00	-1,00	73,44	68,70	4,74	
11	70	11,00	,00	89,99	85,08	4,91	
12	77	12,00	1,00	104,97	102,15	2,82	
13	84	13,00	2,00	118,24	118,87	-,63	
14	91	14,00	3,00	129,55	134,27	-4,72	
15	98	15,00	4,00	141,46	147,69	-6,23	
16	105	16,00	5,00	155,40	158,81	-3,41	
17	112	17,00	6,00	166,84	167,66	-,82	
18	119	18,00	7,00	175,32	174,47	,84	
19	126	19,00	8,00	181,74	179,59	2,15	
20	133	20,00	9,00	186,06	183,36	2,71	
21	140	21,00	10,00	187,76	186,09	1,67	
??							

106. ábra. A nemlineáris regresszió végrehajtása után a bővült adattáblázat

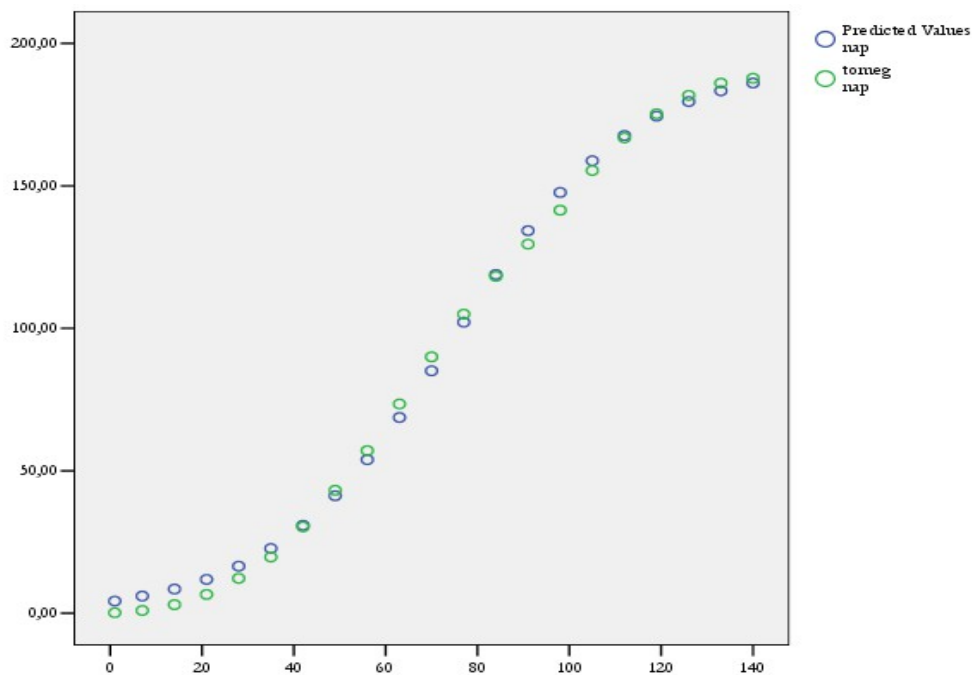
Az illesztett görbe kirajzolása a **GRAPHS / SCATTER / OVERLAY** menüben történik. Erre akkor van lehetőségünk, ha a nemlineáris regresszió **Save...** beállításánál megjelöltük a **PREDICTED VALUES** és a **RESIDUALS** parancsokat. Ennek hatására ugyanis az **SPSS DATA VIEW** ablakában 2 új változó jelenik meg **PRED_** és **RESID_** változónevekkel (106. ábra).

A logisztikus regressziófüggvény kirajolásához *tomeg-napok* és *pred_-napok* változó-párokat vigyük be (107. ábra).



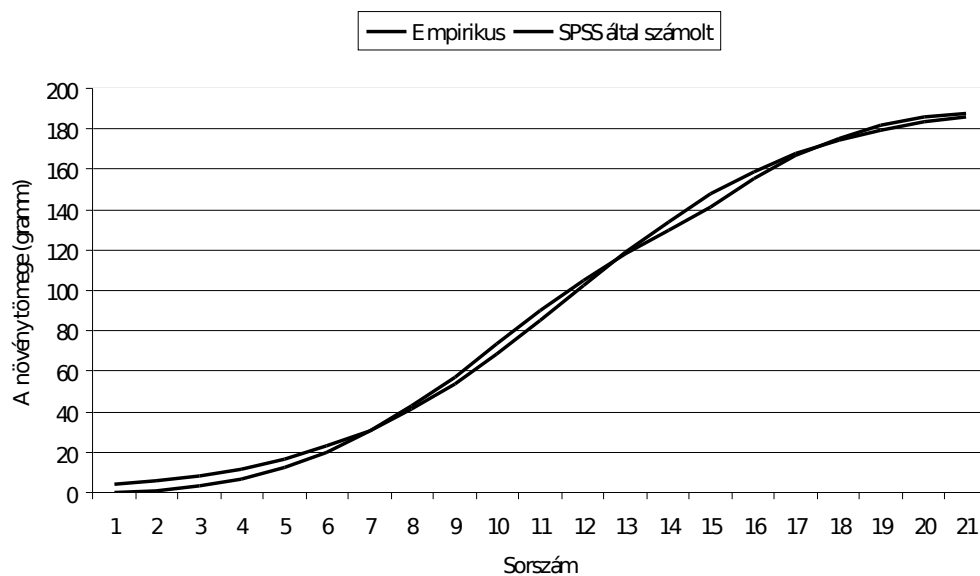
107. ábra. Az *GRAPHS/ SCATTER / OVERLAY* menüpont beállításai

A beállítások elvégzése után futtassuk le a programot, majd megkapjuk azt az ábrát, ami közösen szemlélteti az eredeti Y és a számolt \hat{Y} értékeket (108. ábra).



108. ábra. Az empirikus y és a számolt \hat{y} értékek közös pontdiagramja

A program alapbeállításban nem köti össze a pontokat, ha ez is cél, akkor azt a grafikon szerkesztőben állíthatjuk be (109. ábra).



109. ábra. Kukorica növényegyed növekedése az empirikus és az SPSS által számolt adatokkal

ADATREDUKCIÓK

Főkomponens-analízis

Sajátérték számításon alapuló valódi több-változós eljárás. Az x változó $s_i^2 = 1$ varianciáját bonjuk fel. Az eredetileg megfigyelt változókat korrelációjuk alapján kevesebb számú főkomponens változóvá vonjuk össze. Gyakran már 2-3 főkomponens változóval kielégítő pontossággal helyettesíthetjük a „ p ” számú megfigyelt változót. Minden megfigyelési egység annyi főkomponens értéket kap, ahány főkomponens-változót kiszámítunk.

A főkomponens-analízis (principal component analysis) a több-változós módszerek közül a legfontosabb. Gyakorlati alkalmazásuk a bonyolult és számításigényes sajátérték számítás miatt csak számítógépen valósítható meg. A módszer előnyei:

A változók számának csökkentése, a jelentéktelen változók kiszűrése.

A vizsgált változók csoportosítása az egymás közötti korrelációjuk alapján. Megállapíthatjuk, hogy hány ilyen csoport van, és csoporton belül a változók kapcsolata milyen, pozitív vagy negatív.

Közös háttérváltozó ill. faktor felismerése, mely valamely változócsoporttal szoros összefüggésben van. (pl. levegő, talajhőmérséklet közötti kapcsolat, melynek közös háttérváltozója a napenergia)

A változók térbeli elhelyezkedését, csoportosulását lehet ábrázolni. A főkomponensek lesznek a koordinátarendszer tengelyei.

A főkomponens változók kiszámításával osztályozni tudjuk a megfigyelési egységeket több tulajdonság, ill. változó együttes figyelembevételével. Minden megfigyelés annyi főkomponensértéket kap, ahány főkomponens változót kiszámítunk. A főkomponens változók fogják képezni a két-, esetleg három dimenziós ábrák tengelyeit.

A főkomponens változók és egy adott függőváltozó között két-változós vagy többszörös regresszióanalízist végezhetünk, ezt nevezik főkomponens regressziónak.

140. táblázat. Alapadatok

Fajta	farinográf érték	sikér terülés	sikér mennyiség	fehérje %
Mironovszkaja 808.	81.8	3.0	34.3	14.8
Fertődi 293.	75.9	6.4	39.3	16.1
Bezostája	79.9	2.6	32.6	14.2
Martonvásári 1.	68.6	3.7	31.7	14.5
Martonvásári 2.	77.4	3.2	33.0	14.5

Martonvásári 16.	68.7	6.0	37.1	14.8
Martonvásári 24.	73.6	3.2	31.7	13.4
Jubilejnaja	73.3	2.1	31.4	14.5
Avróra	66.8	5.1	34.1	14.5
GK-Fertődi 2.	58.3	6.5	33.4	15.0
Kavkáz	61.2	5.1	33.3	14.5
Rannaja	59.6	2.9	30.4	15.1
Kiszombori	52.6	7.9	35.8	14.6
Burgas	44.2	10.8	36.1	14.0
Összesen:	941.9	68.5	474.2	204.5

SPSS Analyze, Descriptive Statistics, Descriptives...

Options... Mean, Std. Deviation

Save standardized values as variables

141. táblázat. Átlagok és szórások

Descriptive Statistics			
	N	Mean	Std. Deviation
Farinograf érték	14	67.279	10.9255
Sikér terület	14	4.893	2.4474
Sikér mennyisége	14	33.871	2.4703
Fehérje %	14	14.607	.6044
Valid N (listwise)	14		

Standardizálás után az alábbi értékeket kapjuk:

142. táblázat. Standardizált adatok, Z mátrix

Fajta	farinograf érték	sikér terület	sikér mennyiség	fehérje %
Mironovszkaja 808.	1.33	-.77	.17	.32
Fertődi 293.	.79	.62	2.20	2.47
Bezostája	1.16	-.94	-.51	-.67
Martonvásári 1.	.12	-.49	-.88	-.18

Martonvásári 2.	.93	-.69	-.35	-.18
Martonvásári 16.	.13	.45	1.31	.32
Martonvásári 24.	.58	-.69	-.88	-2.00
Jubilejnaja	.55	-1.14	-1.00	-.18
Avróra	-.04	.08	.09	-.18
GK-Fertődi 2.	-.82	.66	-.19	.65
Kavkáz	-.56	.08	-.23	-.18
Rannaja	-.70	-.81	-1.41	.82
Kiszombori	-1.34	1.23	.78	-.01
Burgas	-2.11	2.41	.90	-1.00
Összesen:	0	0	0	0

A standardizált értékek tulajdonságai: összegük, ill. az átlaguk egyenlő nullával, a szórásuk egy. A standardizálással egy nulla várhatóértékű, egy szórású sokaságot állítottunk elő.

SPSS Analyze, Data Reduction, Factor...
Descriptives, Correlation Matrix

Korrelációs mátrix meghatározása

143. táblázat. Korrelációs mátrix, **R** mátrix

	Farinograf érték	Sikér terület	Sikér mennyisége	Fehérje %
Correlation Farinograf érték	1.000	-.774	-.126	.103
Sikér terület	-.774	1.000	.681	.087
Sikér mennyisége	-.126	.681	1.000	.480
Fehérje %	.103	.087	.480	1.000

Az **U** sajátvektor mátrix és a sajátértékek (λ_j) meghatározása

144. táblázat. Sajátvektor mátrix és sajátértékek, **U** mátrix és λ

Változó	U ₁	U ₂	U ₃	U ₄
Farinograf érték	-.4787	.5312	.5045	.4838
Sikérterület	.6560	-.2008	.1514	.7116
Sikér mennyiség	.5361	.4144	.5454	-.4933
Fehérje %	.2303	.7111	-.6520	.1270
sajátértékek (λ_j)	2.1524	1.3316	0.4989	0.0170

A sajátvektorok sor és oszlop irányban normáltak, azaz a négyzetösszegük egy sor-, ill. oszlopvektoron belül 1.

A sajátvektorok további tulajdonsága, hogy sorpáronkénti és oszloppáronkénti szorzatösszegük nulla, azaz a sorok és oszlopok páronként ortogonálisak (függetlenek egymástól). Az **U** mátrix ortonormált.

Ha a sajátértékeket összeadjuk, megkapjuk a változók számát, a mátrix rangját.

Főkomponens koeficiensek

A főkomponens koeficienseket (Component Score Coefficient) a sajátvektor mátrixból állítjuk elő súlyozással, tehát a sajátvektorokat osztjuk a hozzá tartozó sajátértékek gyökével.

$$wu_{ij} = \frac{u_{ij}}{\sqrt{\lambda_j}}$$

145. táblázat. Súlyozott főkomponens-koeficiensek **WU**

Component Score Coefficient Matrix

	Component			
	1	2	3	4
Farinograf érték	-,326	,460	,714	3,706
Sikér terület	,447	-,174	,214	5,451
Sikér mennyisége	,365	,359	,772	-3,778
Fehérje %	,157	,616	-,923	,973

Extraction Method: Principal Component Analysis.
Component Scores.

Főkomponens változók

Főkomponens-változók kiszámítása: **Z** mátrix * Súlyozott főkomponens-koefficiensek.

146. táblázat. Főkomponens-változók C mátrix

Fajta	C ₁	C ₂	C ₃	C ₄
Mironovszkaja 808.	-,66600	1,00537	,62309	,36481
Fertődi 293.	1,20873	2,56748	,11257	,38011
Bezostája	-1,08968	,09483	,84878	,46418
Martonvásári 1.	-,60646	-,28447	-,53313	,94043
Martonvásári 2.	-,76830	,31087	,40475	,82342
Martonvásári 16.	,68755	,64722	,90447	-1,67999
Martonvásári 24.	-1,13286	-1,15980	1,42983	-,24724
Jubilejnaja	-1,08353	-,01630	-,45970	-,56999
Avróra	,05812	-,11090	,22191	-,22301
GK-Fertődi 2.	,59406	-,16055	-1,19359	1,88714
Kavkáz	,10701	-,46317	-,39425	-,89894
Rannaja	-,52031	-,18406	-2,51418	-,94032
Kiszombori	1,27122	-,55915	-,08266	-1,24292
Burgas	1,94044	-1,68739	,63211	,94232
Összesen:	0	0	0	0

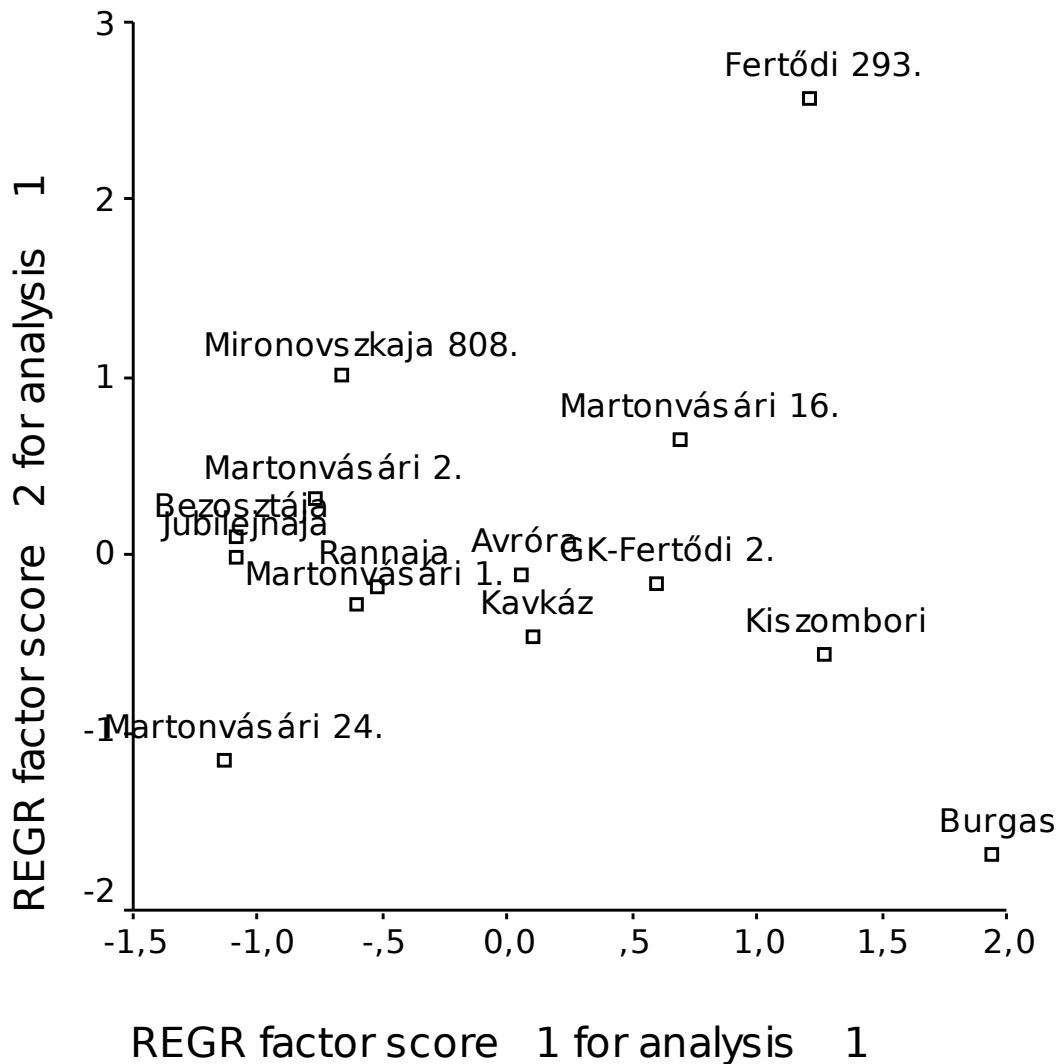
A főkomponens-változók középértéke nulla, szórásnégyzetük egyenlő eggyel. Tehát tulajdonságban hasonlítanak a Z-mátrixhoz, azonban van egy nagyon jelentős eltérés. A főkomponens-változók egymástól függetlenek, azaz az egymás közötti korrelációjuk nulla. (A kovariancia-mátrixa is ugyanígy néz ki.) A standardizált változók és a főkomponens-változók szórásnégyzeteinek összege, valamint a sajátértékek összege azonos.

147. táblázat. Főkomponens-változók korrelációs mátrixa

	C ₁	C ₂	C ₃	C ₄
C ₁	1	0	0	0
C ₂	0	1	0	0
C ₃	0	0	1	0
C ₄	0	0	0	1

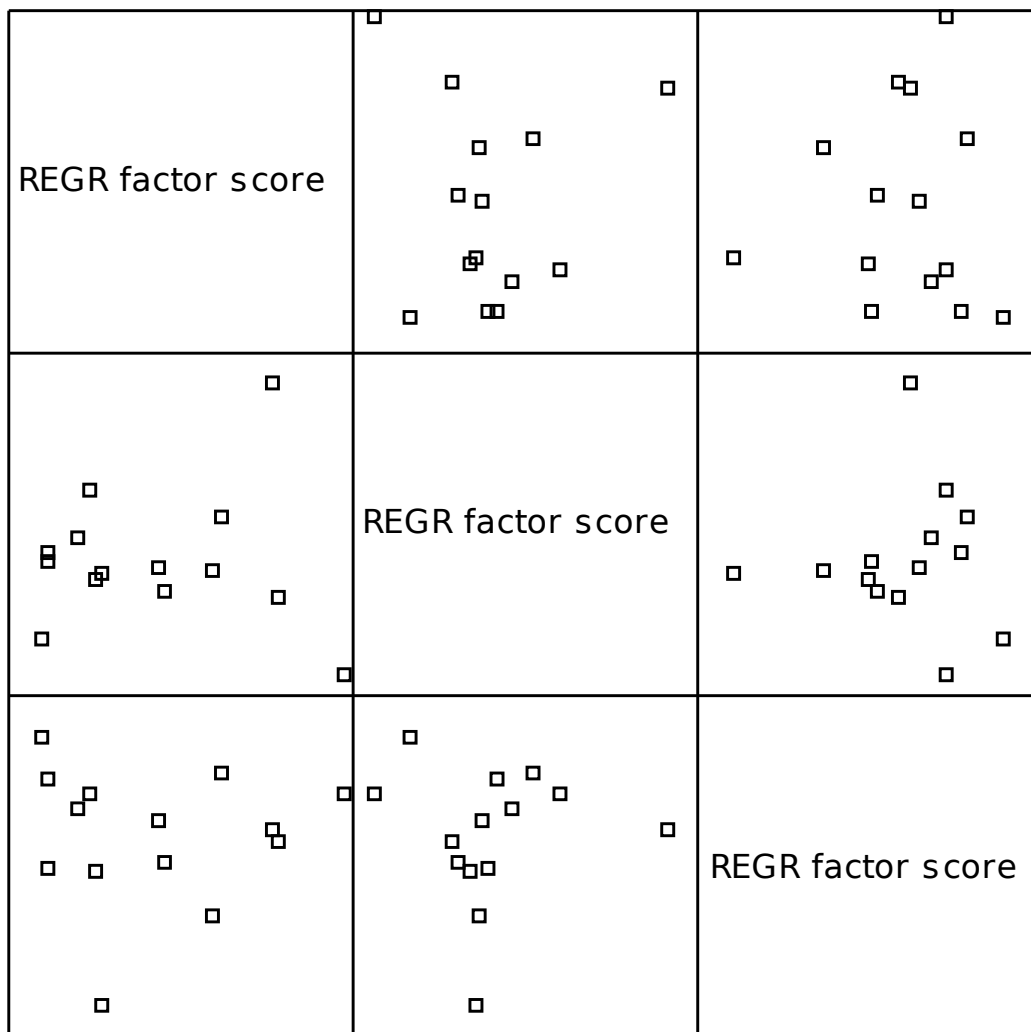
A főkomponens változók ábrázolása

A vízszintes tengely a C_1 , a függőleges a C_2 változó. A különböző őszi búzafajták főkomponens-változó értékeit az alábbi ábra mutatja.



110. ábra. A főkomponens-változók ábrázolása

Három főkomponens-változó két dimenziós ábrázolásához válasszuk a Scatterplot Matrix menüpontot, és adjuk meg az első három főkomponens-változót.



111. ábra. Három főkomponens-változó ábrázolása

Az átló elemei a főkomponens-változók. Az első oszlopban az első változó az x-tengely, a másodikban a második, és így tovább. Az y-tengelyt a sorok mutatják.

A főkomponens súlyok meghatározása

A sajátvektorok elemeit megszorozzuk a hozzá tartozó sajátérték négyzetgyökével, vagyis a szórással.

$$a_{ij} = u_{ij} \sqrt{\lambda_j}$$

148. táblázat. Főkomponenssúly mátrix, **A**-mátrix

Component Matrix^a

	Component			
	1	2	3	4
Farinograf érték	-,702	,613	,356	6,317E-02
Sikér terület	,962	-,232	,107	9,291E-02
Sikér mennyisége	,787	,478	,385	-6,44E-02
Fehérje %	,338	,821	-,461	1,658E-02

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

A főkomponens-súly mátrix tulajdonságai:

Számszerű értéke csak -1 és +1 között lehet. Az oszloponkénti négyzetösszeg egyenlő a hozzá tartozó sajátértékkel.

A soronkénti négyzetösszeg egyenlő eggyel.

Tehát oszlop irányban a főkomponensek, sor irányban a megfigyelt változók varianciáját bontottuk fel.

A súlyok négyzeteinek főösszege egyenlő a mátrix rangjával, az egész rendszer összvarianciájával.

Kommunalitás, h^2 . Ha sor irányban balról jobbra haladva összegezzük a főkomponens-súly négyzeteit, megkapjuk a kumulált értéküket, és ezeket nevezzük kommunalitásnak.

149. táblázat. Kommunálisok

Communalities

	Initial	Extraction
Farinograf érték	1,000	1,000
Sikér terület	1,000	1,000
Sikér mennyisége	1,000	1,000
Fehérje %	1,000	1,000

Extraction Method: Principal Component Analysis.

Bármely két oszlop szorzata nulla. A főkomponenssúly vektorok ortogonálisak (függetlenek).

Bármely két sor szorzata a két változó két-változós korrelációs koefficiensét adja. Ha megszorozzuk az **A**-mátrixot a transzponáltjával, visszakapjuk az **R**-mátrixot, azaz az eredeti változók korrelációs koefficiensét.

Factor Analysis, Descriptives..., Correlation Matrix, Reproduced

150. táblázat. Korrelációs mátrix reprodukálása a főkomponenssúlyokból, maradékok

Reproduced Correlations

		Farinograf érték	Sikér terület	Sikér mennyisége	Fehérje %
Reproduced Correlation	Farinograf érték	1,000 ^a	-,774	-,126	,103
	Sikér terület	-,774	1,000 ^b	,681	8,740E-02
	Sikér mennyisége	-,126	,681	1,000 ^b	,480
	Fehérje %	,103	8,740E-02	,480	1,000 ^b
Residual ^a	Farinograf érték		,000	1,665E-16	,000
	Sikér terület	,000		-4,441E-16	-2,78E-17
	Sikér mennyisége	1,665E-16	-4,441E-16		1,110E-16
	Fehérje %	,000	-2,776E-17	1,110E-16	

Extraction Method: Principal Component Analysis.

- a. Residuals are computed between observed and reproduced correlations. There are 0 (,0%) nonredundant residuals with absolute values greater than 0.05.
- b. Reproduced communalities

A főkomponens-analízissel a varianciákat átrendeztük. A standardizált változóknál minden változó azonos jelentőséggel szerepel a variancia szempontjából. A főkomponens-analízisben az eredeti változók összefüggése miatt az első főkomponens varianciája magába foglalja az összes változó varianciájának legnagyobb közös részét, második főkomponens a maradék varianciák legnagyobb közös részét és így tovább, míg az utolsó főkomponensekre alig marad varianciarész. Ezért ezeket jelentéktelennek tekinthetjük, és elhanyagolhatjuk. Az átrendezett varianciákban figyelembe vettük az X változó összes varianciáját és egymás közötti korrelációját. A főkomponensek egymással már nem korrelálnak.

A λ sajátértékeket főkomponensenként kumulálva mutatja a 151. táblázat. Leolvasható, hogy a különböző főkomponensek hány százalékát értelmezik az összes varianciának.

151. táblázat. Az összes variancia felbontása

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,152	53,810	53,810	2,152	53,810	53,810
2	1,332	33,290	87,100	1,332	33,290	87,100
3	,499	12,473	99,574	,499	12,473	99,574
4	1,704E-02	,426	100,000	1,704E-02	,426	100,000

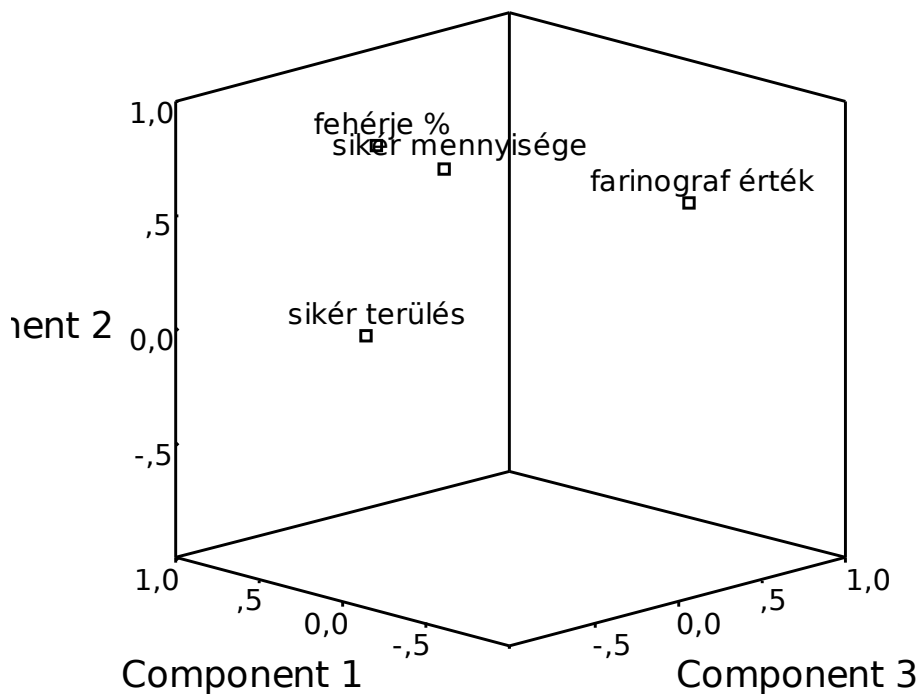
Extraction Method: Principal Component Analysis.

Főkomponensek ábrázolása

A főkomponensek ábrázolása a főkomponenssúlyok alapján történik, ezért az **A** mátrixot főkomponensmintázatnak (pattern) is nevezik. Két legfeljebb három dimenziós ábrát készíthetünk.

Factor Analysis, Rotation, Display, Loading plot(s)

Component Plot



112. ábra. A változók három dimenziós konfigurációja

A főkomponenssúlyok gyakorlati értelmezése

A főkomponenssúlyok a megfigyelt változók és a főkomponens-változók közötti korrelációs koefficiensek, melyet a 152. táblázat mutat.

152. táblázat. A korrelációs koefficiensek, ill. főkomponenssúlyok

		Correlations				
		Farinograf érték	REGR factor score 1 for analysis 1	REGR factor score 2 for analysis 1	REGR factor score 3 for analysis 1	REGR factor score 4 for analysis 1
Farinograf érték	Pearson Correlation	1	-,702**	,613*	,356	,063
	Sig. (2-tailed)	,	,005	,020	,211	,830
	N	14	14	14	14	14
REGR factor score 1 for analysis 1	Pearson Correlation	-,702**	1	,000	,000	,000
	Sig. (2-tailed)	,005	,	1,000	1,000	1,000
	N	14	14	14	14	14
REGR factor score 2 for analysis 1	Pearson Correlation	,613*	,000	1	,000	,000
	Sig. (2-tailed)	,020	1,000	,	1,000	1,000
	N	14	14	14	14	14
REGR factor score 3 for analysis 1	Pearson Correlation	,356	,000	,000	1	,000
	Sig. (2-tailed)	,211	1,000	1,000	,	1,000
	N	14	14	14	14	14
REGR factor score 4 for analysis 1	Pearson Correlation	,063	,000	,000	,000	1
	Sig. (2-tailed)	,830	1,000	1,000	1,000	,
	N	14	14	14	14	14

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

* . Correlation is significant at the 0.05 level (2-tailed).

A főkomponensekkel háttérváltozókat (okváltozókat) akarunk felderíteni. A főkomponenssúlyok azt fejezik ki, hogy milyen jelentősége és súlya van valamely főkomponensnek (háttérváltozónak) a megfigyelt változók varianciájában.

A megfigyelt változók közötti korrelációs koefficiensek felbontása. Az **A**-mátrix bármely két sorának skaláris szorzata megadja a két változó közötti korrelációs koefficienst. Két változó skaláris szorzata akkor lehet nagy, ha a két változó nagy főkomponenssúlyai ugyanazokban a főkomponensekben vannak és a különböző főkomponensekben a szorzatuk azonos előjelű. A korrelációs koefficienst így egymástól független tényezőkre bontottuk fel.

A főkomponenssúlyok csoportosulása. Ha kettőnél több nagy főkomponens ugyanabba a főkomponensben van, akkor a változók egymással páronként, ezáltal közösen, csoportosan korrelálnak. Közös háttérváltozót kereshetünk.

A változók ábrázolásakor legjobban a kör, ill. gömb kerületén elhelyezkedő változók korrelálnak a legszorosabban. Az egymással negatívan korreláló

változókat az origóra középpontosan tükrözni lehet, hogy könnyebben felismerjük az összefüggést.

Mit jelent a nagy vagy kis főkomponenssúly? Sváb szerint, ha a változók között nincs korreláció, úgy $p=2$ esetén $\frac{1}{\sqrt{2}}$, kb. 0,7 körüliek az a_{ij} értékek, $p=10$ változó esetén $\frac{1}{\sqrt{10}}$, kb. 0,3 körüli értékeket kapunk véletlenszerű elosztásban, minthogy a négyzetek összege mindenképpen csak 1 lehet.

Hány főkomponens jelentős? A λ legalább egy, vagyis eléri az átlagot. Ezt alkalmazzák a statisztikai programcsomagokban is. Más ajánlás szerint az összes variancia legalább 80%-át magyarázzák a főkomponensek, azaz a kumulált λ százalék legalább 80% legyen. Sváb szerint ez, ha túl sok változó van, magas követelmény. Egyesek a faktoranalízisben képletet is megadnak, hogy legfeljebb hány faktort érdemes meghatározni.

$$q \approx \frac{(2p+1) - \sqrt{8p+1}}{2}$$

Főkomponens-analízis forgatással

A faktoranalízist jóval korábban fejlesztették ki, mint a főkomponens-analízist, így a forgatást is eredetileg a faktoranalízisre dolgozták ki. Mind a faktoranalízisben, mind a főkomponens-analízisben ugyanazokat a forgatási módszereket használjuk. A faktoranalízis kidolgozásakor az volt az elképzelés, hogy p számú X változó kevesebb $q < p$ számú háttérváltozóval, faktorról értelmezhető, mert ugyanaz a faktor több X változót értelmez.

Az X változó faktorsúlyai azonban többnyire megoszlanak kettő vagy annál is több faktorra, annak ellenére, hogy az ábrázolás szerint a változók csoportokba tömörülnek. A csoportok ugyanis nem egyetlen tengely mentén, hanem a hipergömb több tengelye által közrefogott valamely szektorában fekszenek. Ilyenkor a tengelyek elforgatásával meg tudjuk tenni, hogy a tengelyek áthaladjanak a csoportokon. Az azonos csoportokba tartozó X változók faktorsúlyai a közös faktorban -1-hez és +1-hez lesznek közel, a többi faktorban viszont nullához közelítenek. Így az eredeti p számú változót $q < p$ számú faktorról tudjuk leírni, amelynek nagy része közös faktor, és sokszor szakmailag értelmezhető háttérváltozót ismerhetünk fel benne. Ezt az eljárást nevezik *faktorextrahálásnak*.

Amennyiben a forgatással nem sikerül az X változókat kevesebb számú faktorról előállítani, ez azt jelenti, hogy egyáltalán nincs vagy csak kevés a változócsoporthoz, és az egyes csoportok is csak kevés változóból tevődnek össze.

A forgatás lehet derékszögű és ferdeszögű. A forgatás szöge mindig tengelypáronként értendő. A derékszögű forgatáskor az új koordináta rendszer is derékszögű marad, ezért a faktorok függetlensége továbbra is megmarad, a forgatás tehát ortogonális.

Ferdeszögű forgatás esetén a faktorok nem lesznek függetlenek, a közöttük fennálló korreláció mértéke:

$$r_{I, II} = \cos(90^\circ + \alpha_I - \alpha_{II})$$

ahol: α_I : az I., vízszintes tengely elfordításának szöge

α_{II} : a II., függőleges tengely elfordításának szöge

Ez a megoldás az „elsődleges faktor” (primary factor) szerinti ferdeszögű forgatás. Ennek továbbfejlesztése a ferdeszögű „vetületi vektorok” (reference vector) szerinti forgatás.

Derékszögű forgatás Varimax módszerrel

Többféle derékszögű forgatás létezik. A legelterjedtebb eljárás H.F. Kaiser módszere a Variamax rotáció. Ez elégtí ki legjobban a Thurstone-féle egyszerű struktúra követelményeit. A Varimax kritérium: a főkomponenssúly négyzetek oszloponkénti varianciáinak összege (V) maximum legyen.

$$V = \sum_{j=1}^q s_{a_j}^2 \quad \text{maximum}$$

153. táblázat. Főkomponenssúly mátrix Varimax rotáció után, A_0 mátrix

Rotated Component Matrix^a

	Component			
	1	2	3	4
Farinograf érték	,997	-4,30E-02	5,821E-02	3,083E-02
Sikér terület	-,753	,639	-6,40E-03	,155
Sikér mennyisége	-,102	,951	,291	-1,02E-02
Fehérje %	5,514E-02	,212	,976	-5,15E-04

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Ez akkor a legnagyobb, ha a főkomponens súlyok $|1|$ és 0-hoz közeli értékek. A forgatás során megváltoznak az a_{ij} értékek. Az új A_0 mátrix is ortogonális, soronként a a_{ij}^2 összege továbbra is 1, azonban az oszloponkénti összegük módosul, többé már nem azonosak λ_j -vel. Az ortogonális forgatás a

főkomponensekben azonban csak átrendezi a varianciák változónkénti megoszlását, de az összes varianciát nem módosítja.

A fenti Varimax kritériumban minden változó azonos súllyal, jelentőséggel vesz részt. Ezért a számítások során a tényleges kritérium: az a_{ij}^2 értékeket az X_i változók h_i^2 kommunalitásával súlyozzák.

$$V = p \sum_{j=1}^q \sum_{i=1}^p \left(\frac{a_{ij}^2}{h_i^2} \right)^2 - \sum_{j=1}^q \left(\sum_{i=1}^p \frac{a_{ij}^2}{h_i^2} \right)^2$$

154. táblázat. Transzformáló mátrix, **T** mátrix

Component Transformation Matrix

Component	1	2	3	4
1	-,691	,681	,238	,055
2	,587	,341	,734	-,017
3	,421	,645	-,636	,048
4	,028	-,063	,030	,997

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

155. táblázat. Főkomponens-koefficiensek forgatás után

Component Score Coefficient Matrix

	Component			
	1	2	3	4
Farinograf érték	,900	,162	-,084	3,704
Sikér terület	-,168	,039	,004	5,474
Sikér mennyisége	,177	1,108	-,252	-3,717
Fehérje %	-,108	-,340	1,105	,924

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Component Scores.

156. táblázat. Főkomponens-változók forgatás után, C mátrix

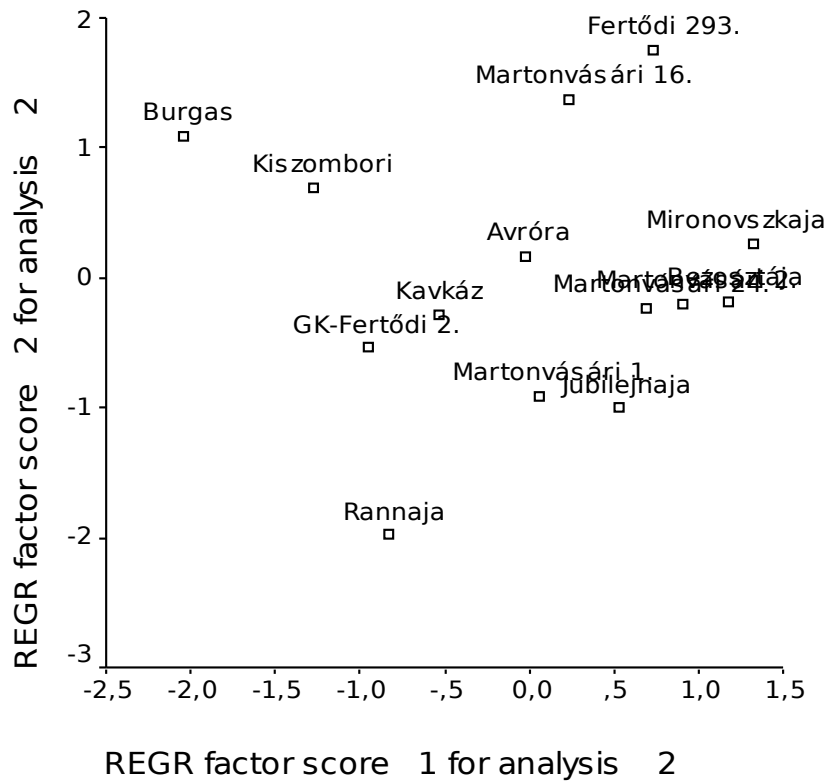
Fajta	C ₁	C ₂	C ₃	C ₄
Mironovszkaja 808.	1,32	,27	,19	,34
Fertődi 293.	,73	1,75	2,11	,41
Beosztája	1,18	-,19	-,72	,44
Martonvásári 1.	,05	-,91	,01	,88
Martonvásári 2.	,91	-,21	-,19	,79
Martonvásári 16.	,24	1,38	,01	-1,60
Martonvásári 24.	,70	-,23	-2,04	-,22
Jubilejnaja	,53	-1,00	,01	-,65
Avróra	-,02	,16	-,22	-,21
GK-Fertődi 2.	-,95	-,54	,84	1,86
Kavkáz	-,54	-,28	-,09	-,90
Rannaja	-,83	-1,98	1,31	-1,08
Kiszombori	-1,28	,70	-,09	-1,16
Burgas	-2,04	1,09	-1,15	1,11
Összesen:	0	0	0	0

157. táblázat. Az összes variancia felbontása

Total Variance Explained

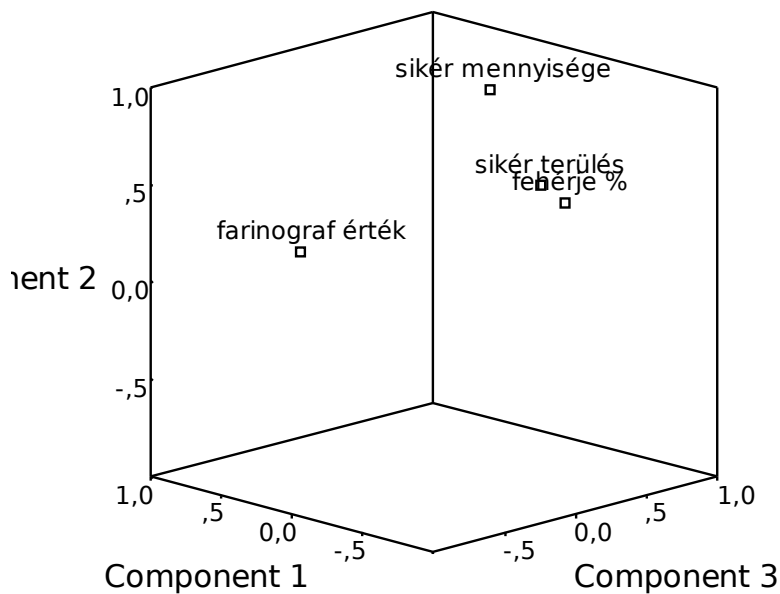
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,152	53,810	53,810	2,152	53,810	53,810	1,575	39,364	39,364
2	1,332	33,290	87,100	1,332	33,290	87,100	1,360	33,998	73,362
3	,499	12,473	99,574	,499	12,473	99,574	1,040	26,011	99,373
4	1,704E-02	,426	100,000	1,704E-02	,426	100,000	2,509E-02	,627	100,000

Extraction Method: Principal Component Analysis.



113. ábra. A főkomponensváltozók ábrázolása

Component Plot in Rotated Space



114. ábra. A változók három dimenziós konfigurációja

Faktor-analízis

X változó h^2 részének (közös faktorok) felbontása történik. A faktorok lehetnek:

közös faktor

általános faktor

csoport faktor

egyedi faktor

hiba faktor

a hiba faktor származhat a lineáris korrelációs közelítésből is, ill. egyéb zavaró hatásokból. Ebben az eljárásban alapesetben a faktorok nem korrelálnak egymással. Csak a közös faktorokat számítjuk ki. A korrelációs mátrix főátlójába a kommunalításokat helyettesítjük be.

Kategorikus főkomponens-analízis

A kategorikus főkomponens-analízis (CATPCA) az egyszerű PCA általánosítása kevert mérési szintű változók összefüggésrendszerének elemzésére. Ez nagyban hasonlít a többszörös korrespondencia-analízishez. CATPCA segítségével például meghatározhatjuk az autómárkák és az ár, tömeg, üzemanyag-fogyasztás, egyéb közötti kapcsolatokat. Vagy osztályokba sorolhatjuk a különböző típusú autókat több jellemvonás egyidejű figyelembevételével.

A főkomponens-analízis célja a változók eredeti számának csökkentése kisebb egymással nem korreláló komponensekre, amik hordozzák az eredeti adatok információinak jelentős részét. A standard főkomponens-analízisben feltételezzük, hogy a változók között lineáris kapcsolat van. A kategorikus főkomponens-analízisben a változók közötti nem lineáris kapcsolatot modellezzük.

Hogyan lehet használni? Analyze, Data Reduction, Optimal Scaling..., Selected Analysis. Itt választhatjuk ki, hogy milyen kategorikus analízist szeretnénk végezni. Három közül választhatunk:

Többszörös korrespondencia-analízis

Kategorikus főkomponens-analízis

Nemlineáris kanonikus korreláció

A megfelelő analízis kiválasztása az Optimal Scaling Level és a Number of Sets of Variables rádiógombok kombinációjával történik.

Optimal Scaling Level:

Minden változó többszörös nominális változó

Néhány változó nem többszörös nominális változó (egy vagy több változó skála típusú a többi többszörös nominális. Vagy lehetnek még egyszerű nominális ordinális és diszkrét értékek is.

Number of Sets of Variables: meg kell adni hogy a változók csoportjából hányat akarunk összehasonlítani más változó csoportokkal.

Egy csoport

Több csoport

	Minden	változó	Néhány nem	
	többszörös	nominális		
Egy csoport	Többszörös		Kategorikus	
	korrespondencia		főkomponens-analízis	
	analízis			
Több csoport	Nemlineáris	kanonikus	Nemlineáris	kanonikus
	korreláció		korreláció	

NEM PARAMÉTERES PRÓBÁK

Chi-négyzet teszt

A Chi-négyzet teszt a változókat kategóriákba rendezi, és utána számítja ki a statisztikát. A teszt során a megfigyelt és feltételezett relatív gyakoriságokat hasonlítja össze. Lehetőségünk van több csoport eloszlásának homogenitását tesztelni vagy egy megadott relatív gyakorisággal való egyezés tesztelésére.

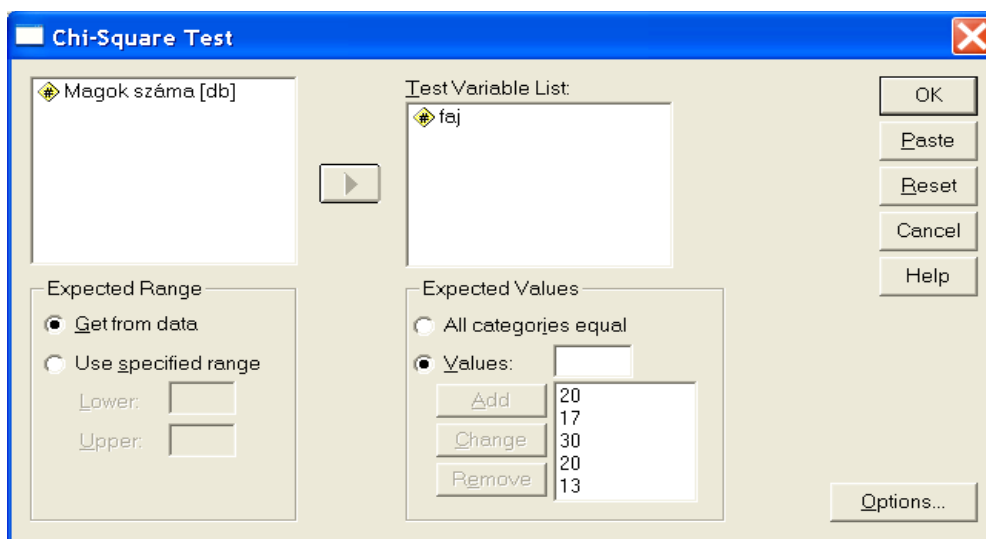
Feladat:

Származhat-e egy gyepmag keverék egy 20, 17, 30, 20, 13%-os összetételű keverékből? A mintavételezés során az alábbi eredményt kaptuk:

Faj	magok száma (db)
Réti perje	236
Angolperje	241
Réti komócsin	443
Réti csenkesz	252
Fehérhere	155
Összesen:	1 327

Az SPSS-ben a fenti adatbázissal csak akkor lehet gyakoriságokat számítani, ha a **Faj** változót súlyozzuk a **magok száma** változóval. Date, Weight Cases...

Analyze, Nonparametric Tests, Chi-Square.



FAJ			
	Observed N	Expected N	Residual
1.00	236	265.4	-29.4
2.00	241	225.6	15.4
3.00	443	398.1	44.9
4.00	252	265.4	-13.4
5.00	155	172.5	-17.5
Total	1327		

Megfigyelt gyakoriság, várható gyakoriság és a kettő különbsége.

Test Statistics	
	FAJ
Chi-Square ^a	11.827
df	4
Asymp. Sig.	.019

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 172.5.

A vetőmagkeverék aránya nem felel meg az előírásnak. Mi lehet ennek az oka?

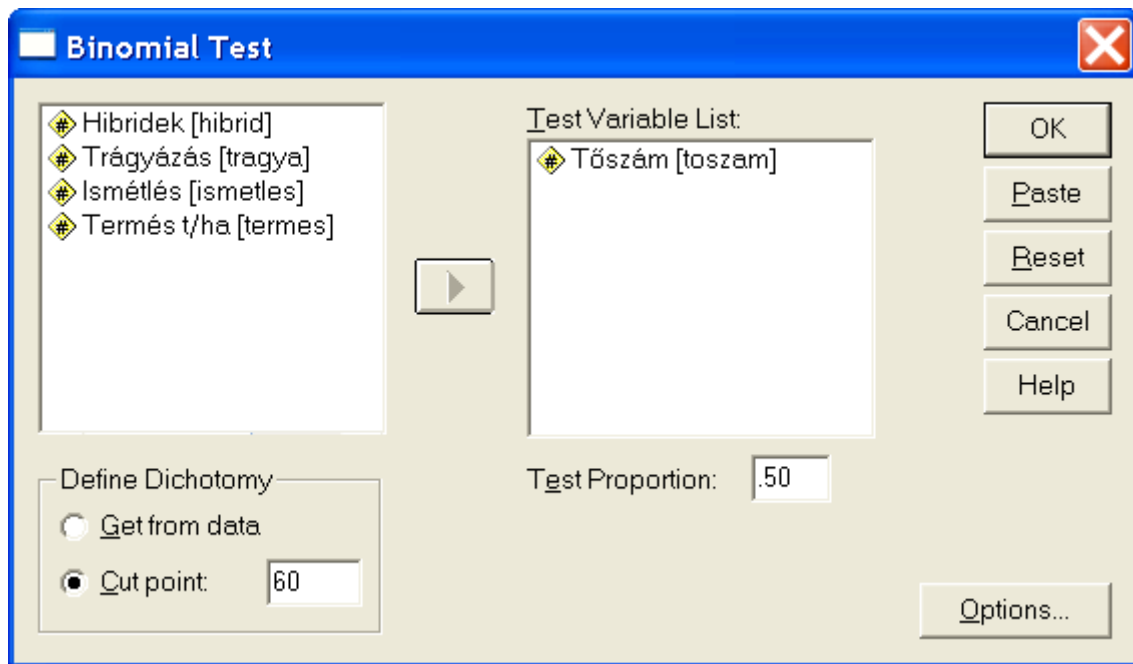
Binomiális teszt

Ezzel a teszttel két csoport (kategória) megfigyelt relatív gyakoriságát lehet összehasonlítani a binomiális eloszlás alapján. A valószínűségi változó kezdeti értéke mindkét csoportban 0,5. A valószínűség megváltoztatásakor az első csoport előfordulását tesztelhetjük. A második csoport előfordulásának valószínűsége 1 mínusz az első csoportra megadott valószínűség.

Feladat:

Megegyezik a 60 ezer alatti tőszám parcelláinak száma az e feletti parcellák számával?

Tőszám, Cut point 60. Test Proportion 0.5, OK.



Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (2-tailed)
Tőszám	Group 1	<= 60	36	.50	.50	1.000 ^a
	Group 2	> 60	36	.50		
Total			72	1.00		

a. Based on Z Approximation.

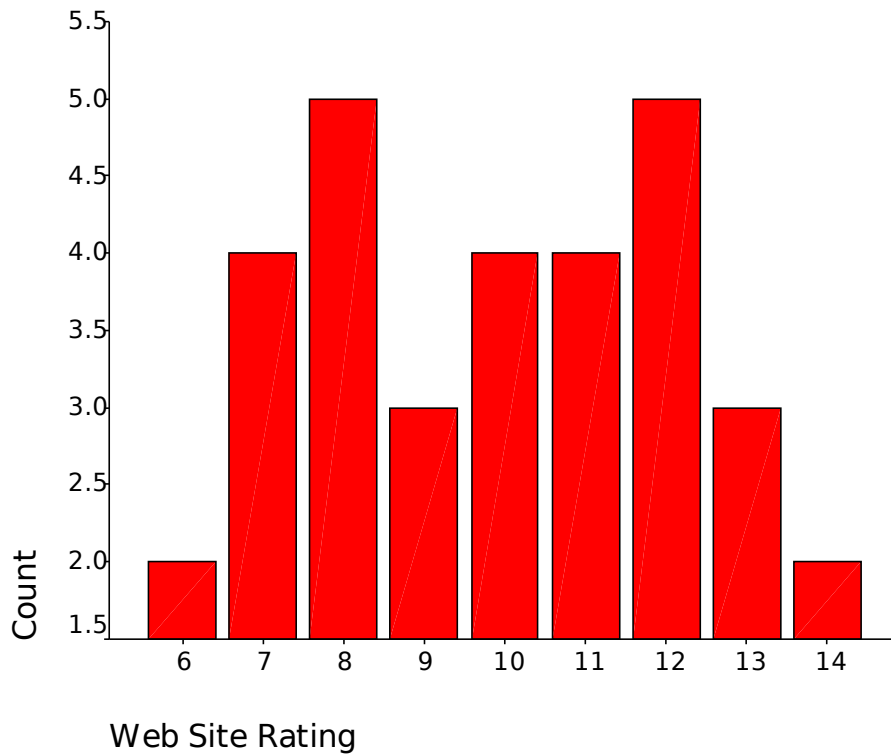
A két csoport relatív gyakorisága tökéletesen megegyezik.

Runs Test

Sok statisztikai teszt alkalmazásának feltétele, hogy a mintában a megfigyelések függetlenek legyenek. Ezt csak akkor tudjuk leellenőrizni, ha ismert a mintaelemek kihúzásának időpontja vagy sorrendje, főként idősoros elemzésnél hasznos. Ezzel a teszttel leellenőrizhetjük, hogy a mintánk véletlen mintának tekinthető-e. Legelőször választani kell egy jellemző értéket – ami legtöbbször valamilyen centrális mutató – és ehhez hasonlítjuk a megfigyelt értékeket. Az eljárás során a változó minden egyes értékét osztályozzuk, hogy a töréspont alatt vagy felett helyezkedik el. Ez után megállapítjuk, hogy van-e valamilyen szabályosság a sorozatban, hányszor ismétlődik egymásután ugyanabba az osztályba tartozó elem, azaz egy sorozat. Egy sorozatnak (run) legalább egy tagja van.

A teszt eredménye attól is függ, hogy mit választunk ki töréspontnak (medián, módusz, átlag, stb.).

Az eloszlás egy bimodális sokaságot mutat melynek két módusza van. Az SPSS a módusz meghatározásakor a nagyobbikat adja meg.



Descriptive Statistics

		Web Site Rating
N		32
Mean		9.94
Std. Deviation		2.368
Minimum		6
Maximum		14
Percentiles	25th	8.00
	50th (Median)	10.00
	75th	12.00

Runs Test	
	Web Site Rating
Test Value ^a	10.00
Cases < Test Value	14
Cases >= Test Value	18
Total Cases	32
Number of Runs	10
Z	-2.283
Asymp. Sig. (2-tailed)	.022

a. Median

Ha a minta teljesen véletlen lenne, akkor a sorozatok száma 17 körüli lenne. Mivel a megfigyelt sorozatok száma csak 10, ezért a Z-statisztika értéke negatív. Túl alacsony a szignifikancia értéke, ezért nem tekinthető véletlennek a minta.

Ratin	cut	point
g	10	
8	1	
7	1	
8	1	
6	1	
10	2	
8	1	
6	1	
7	1	
8	1	
9	1	
7	1	
10	2	
7	1	
8	1	
12	2	
10	2	
12	2	

9	1
11	2
12	2
10	2
13	2
13	2
12	2
11	2
14	2
9	1
14	2
11	2
12	2
11	2
13	2

Runs Test 2

	Web Site Rating
Test Value ^a	12 ^b
Cases < Test Value	22
Cases >= Test Value	10
Total Cases	32
Number of Runs	16
Z	.315
Asymp. Sig. (2-tailed)	.752

a. Mode

b. There are multiple modes. The mode with the largest data value is used.

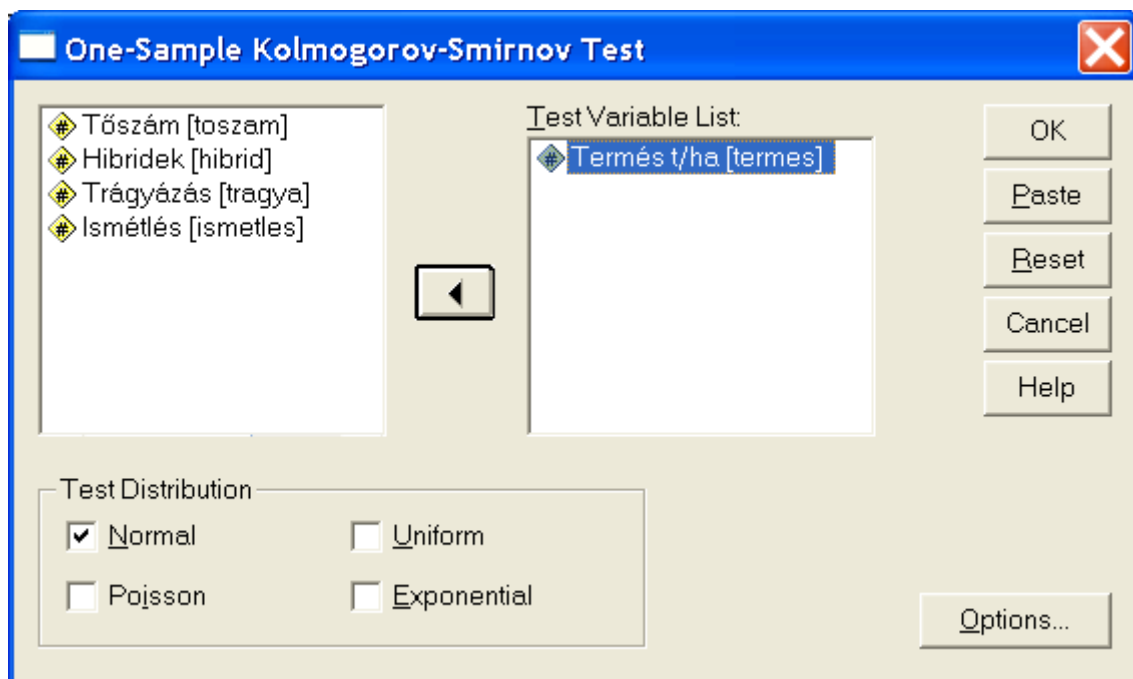
A módusz alatt több, mint kétszer annyi elem fordul elő, mint felette. Ennek az az oka, hogy 12 felett az adatoknak csak a 25%-a helyezkedik el. Mivel a teszt a töréspont alatti ill. feletti elemeket különíti el, a várható sorozatok száma a törésponttól függ. Ebben az esetben a sorozatok várható száma 15 körüli. A számított érték nagyon közel van hozzá, ezért véletlennek tekinthető a minta, amit a szignifikancia értéke is megerősít.

	Web Site Rating
Test Value ^a	8
Total Cases	32
Number of Runs	11
Z	.000
Asymp. Sig. (2-tailed)	1.000

a. User-specified.

A bimodális eloszlás első móduszát választottuk töréspontnak. A sorozatok várható száma ekkor 11. A számított érték pontosan megegyezik a várható értékkel, ezért a minta véletlennek tekinthető.

Egymintás Kolmogorov-Smirnov teszt (One-Sample Kolmogorov-Smirnov Test)



Milyen eloszlásba tartozik a minta? Normál, Poisson, egyenletes (Uniform) és exponenciális (Exponential) eloszlás tesztelése. A megfigyelt adatok kumulatív eloszlás függvényét (cumulative distribution function, CDF) hasonlítja össze a teoretikus eloszlás kumulatív függvényével. A Kolmogorov-Smirnov Z-érték a megfigyelt és teoretikus kumulált eloszlás függvények közötti legnagyobb abszolút különbségből számítják. Ezt az értéket szorozzák a megfigyelések

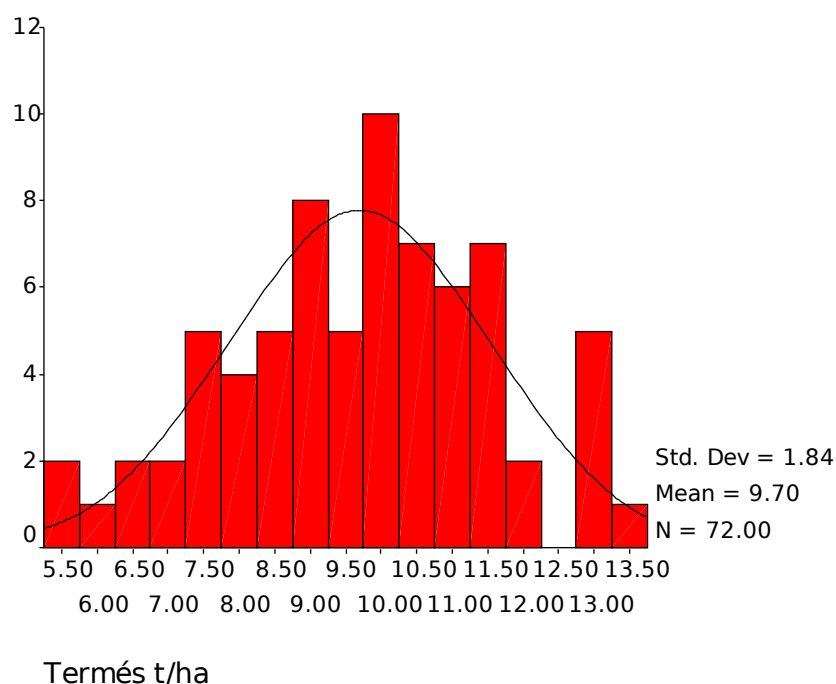
négyzetgyökével. Sok paraméteres teszt megköveteli, hogy a változó normális eloszlású legyen.

One-Sample Kolmogorov-Smirnov Test		
		Termés t/ha
N		72
Normal Parameters ^{a,b}	Mean	9.69609
	Std. Deviation	1.843756
Most Extreme Differences	Absolute	.075
	Positive	.047
	Negative	-.075
Kolmogorov-Smirnov Z		.635
Asymp. Sig. (2-tailed)		.814

a. Test distribution is Normal.

b. Calculated from data.

A nullhipotézis: a mért változó normál eloszlású. A hipotetikus és mért eloszlás nem különbözik egymástól. A nullhipotézist megtartjuk, mivel nagyon kicsi az eltérés a kettő között, és a szignifikancia szint is magas.



Két független mintás tesztek (Two Independent Samples Tests)

Mann-Whitney U-próba

A Mann-Whitney U és a Wilcoxon W statisztika.

Két független minta medián egyezésének igazolására való eljárás (két-mintás t-teszt). A nullhipotézis, hogy a két sokaság ugyanabba az eloszlásba tartozik. Ordinális típusú adatoknál használható, vagy skála típusú adatoknál, ahol nem feltétel a normál eloszlás. Csak az egyezésre ad elfogadható, megbízható eredményt. Ha ettől eltérő eredményt kapunk, nem tudhatjuk biztosan, hogy mi a valóság.

Alkalmazási feltétel:

Hasonló alakú eloszlások (tesztelhető a két-mintás Kolmogorov-Smirnov próbával)

Független minták

Null hipotézis: $M(x) = M(y)$. A hipotézisvizsgálat céljára konstruált valószínűségi változó: n_1+n_2 elemű mintából egyetlen rangsor felállítása, „x” mintára vonatkozó rangszámok összege: R1 vagy W-érték.

$$m = \frac{n_1(n_1 + n_2 + 1)}{2}$$

$$\sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

A próba változójának eloszlása, ha n_1 és n_2 elég nagy, megközelítően $N(m, \sigma)$.

Kolmogorov-Smirnov Z-próba

Két eloszlás összehasonlítására szolgáló eljárás. A nullhipotézis, hogy a két sokaság ugyanabba az eloszlásba tartozik. A Kolmogorov-Smirnov Z-értéket a két csoport kumulált eloszlás függvényei közötti legnagyobb abszolút különbségből számítják. A változóknak ezért illik folyamatos eloszlásúnak lenni. A két csoportban a megfigyelések számának nem kell megegyeznie. Nagyon rugalmas a teszt, nem kell az eloszlásoknak hasonló alakúnak lennie, hisz az eljárás ezt is teszteli.

Alkalmazási feltétel:

Csak folytonos eloszlások hasonlíthatók össze.

Független minták

A próba érzékeny a helyzeti különbségekre és az eloszlások alakjára. A helyzeti különbség azt jelenti, hogy a két eloszlás hol helyezkedik el a skálán. A Kolmogorov-Smirnov teszt akkor is különbözőnek mutatja a két eloszlást, ha az alakjuk (shape) megegyezik, de egymástól távol helyezkednek el. Ezek szerint két eloszlás akkor különbözik, ha vagy az alakjuk, vagy az elhelyezkedésük különbözik, vagy mindkettő. Amennyiben a két eloszlás helyzeti különbsége nem érdekel bennünket, toljuk el a skálát az origóra, aminek a legegyszerűbb módja az adatok standardizálása (ettől az eloszlások alakja semmit sem változik). A standardizálással skála-eltolást és skála transzformációt is végrehajtottunk egyszerre.

Alternatívaként használhatjuk a Crosstabs eljárásokat is kettő vagy több ordinális vagy nominális változó közötti különbség kimutatására.

Amennyiben a t-teszt alkalmazásának feltételei teljesülnek, akkor azt kell használni.

Több független mintás teszt (K Independent Samples...)

Kruskal-Wallis H próba

Rendezett mintán alapuló, több mintás hipotézis vizsgálat, amelynek null hipotézise: minden minta azonos eloszlású sokaságból származik. A próba segítségével „h” darab „nh” elemszámú mintát vizsgálhatunk. Ezt ismételt Wilcoxon-próbákkal is elvégezhetnénk, de ebben az esetben az ismétlések megnövelik az elsőfajú hibát (analóg a középértékek többszörös összehasonlításának, szimultán próbák problematikájával).

Két páronként összetartozó minták tesztjei (2 Related Samples...)

Wilcoxon teszt (Wilcoxon signed-rank test)

Két eloszlás egyezésének vizsgálatára alkalmas. Sokszor használják két várható érték egyezésének vizsgálatára is. A két minta elemei páronként összefüggnek. n_1+n_2 elemű mintából egyetlen rangsort képeznek. Konstruált valószínűségi változó „u”. A nullhipotézis: a páronkénti különbségek a nulla körül szimmetrikusan helyezkednek el.

Előjel próba (Sign)

Összetartozó elem párok vizsgálata. Hipotézis, hogy $x_1 \dots x_n$ minta elemei nagyobb (vagy kisebb) értéket vesznek fel, mint $y_1 \dots y_n$ elemei, ahol az azonos indexű minta elemek között valamilyen logikai kapcsolat van (pl. ugyanazon jelenség két különböző időpontban vagy helyen mért értékei).

Első lépésben meghatározzuk az x_i-y_i különbségek előjelét, utána megszámloljuk, hogy hány darab „-” és „+” előjelű különbség adódott. Az előjel próba, ellentétben a rendezett mintás próbákkal szemben, kisebb elemszámokra erősebb. Így kétszeresen nem indokolt nagy elemszámok esetén az előjel próba használata: Nagyobb minták esetén relatíve gyengébb a próba ereje. Elvész az előjel próba jelentős előnye, a gyors alkalmazhatóság.

McNemar teszt

Két-értékű, bináris vagy dichotóm változók összehasonlítására szolgáló módszer. Tipikusan megismételt mérések esetében használható, amikor ugyanazon egyedeket figyeljük meg: bizonyos esemény bekövetkezése (pl. kezelés) megváltoztatja-e az egyedek állapotát (az esemény előtti és utáni állapot összevetése). Nullhipotézis: a kezelés utáni állapot egyenlő a kezdeti állapottal.

Ez a teszt főként nominális vagy ordinális változók tesztelésére alkalmas.

K számú összetartozó minta tesztjei (k Related Samples...)

Friedman teszt

Több eloszlás homogenitás vizsgálatára alkalmas, összetartozó több változó esetén. Paraméteres megfelelője a kéttényezős variancia-analízis. Feltételezzük, ha az eloszlás megegyezik a várható érték is megegyezik nagy valószínűséggel. Fordítva ez nem igaz. Null hipotézis: a k darab összetartozó változó ugyanabba a sokaságba tartozik.

$$F(x) = G(x) = \dots = K(x)$$

Alkalmazási feltétel: több rendezett minta azonos elemszámokkal, g és h elég nagy, ahol g a minta elemszáma a szempont egy szintjére (blokk), 'h' a szempontonkénti vagy szintenkénti minták száma (kezelés).

A próba változójának eloszlása *Chi-négyzet*, szabadságfoka $k-1$.

Blok	1	r_{11}	...	r_{1k}	A minta elemeinek sorrendje az első szempont szerint
	⋮			⋮	
	g	r_{g1}	...	r_{gk}	A minta elemeinek sorrendje az utolsó szempont szerint
		R_1		R_k	A változók átlagos rangszámai.

Megjegyzés: a Friedman-teszt és a Kendall-féle konkordancia együtttható ugyanannak a problémának a tesztelésére használható. Szignifikancia-szintjeik megegyeznek, mindkettő két-tényezős problémát tárgyal.

Kendall konkordancia együtthatója W

Kettőnél több „bíró” rangsora áll rendelkezésre. Van-e különbség a bírók együttesét tekintve a közöttük lévő egyetértésnek, vagy van-e szignifikáns mértéke? Milyen az egyetértés (konkordancia) a rangsorok együttesében.

(egyáltalán nem egyezik a bírálók véleménye) $0 \leq W \leq 1$ (tökéletesen egyezik a bírálók véleménye). A próba változójának eloszlása Chi-négyzet, szabadságfoka $m-1$.

Pl.: több oktató a hallgatókat rangsorolja tudás szerint. Minden oktató sorba rendezi a hallgatót 1-től m -ig, m a hallgatók száma. Az oktatók száma legyen n . Vajon megegyeznek az oktatók véleményei, van közöttük egyetértés?

Hallgatók	Hallgató1	Hallgató2	Hallgató3	Hallgató4	Hallgató5	Hallgató6	Hallgató7	Hallgató8
Kovács	6	2	3	5	1	4	8	7
Kiss	6	3	1	7	2	4	8	5
Szabó	7	3	2	5	1	4	8	6

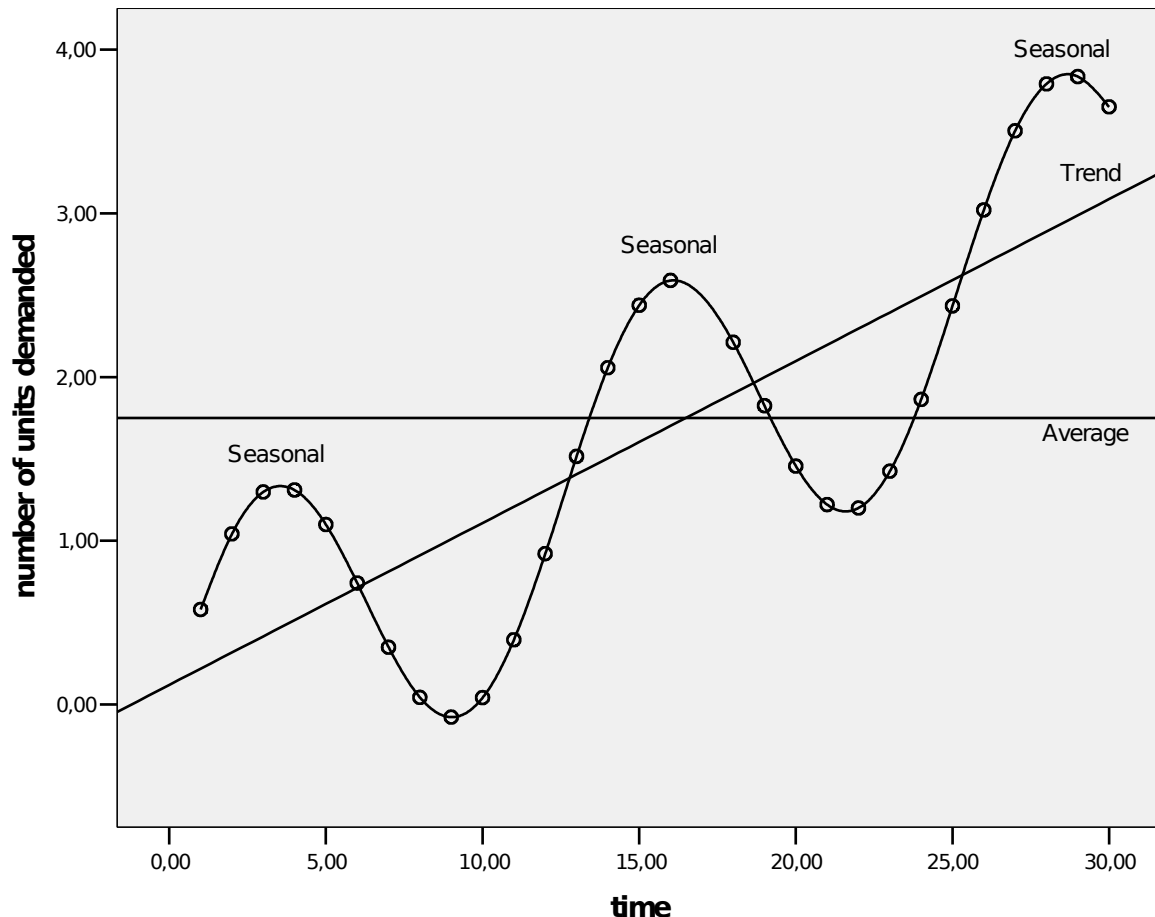
Test Statistics

N	3
Kendall's W ^a	.931
Chi-Square	19.556
df	7
Asymp. Sig.	.007

a. Kendall's Coefficient of Concordance

Idősorok ANALÍZISE

Az idősornak a különböző időpontokban végzett megfigyeléseket nevezzük. Az adatok sorrendje nagyon fontos, a különböző rendezéseknek itt nincs értelme. A megfigyeléseket egyenlő időközökben végezzük. $t=1, 2, 3, \dots, n$.



A mért értékeket $u_1, u_2, u_3 \dots u_t$ -vel jelöljük.

$t=0$ időpontból kiindulva nemcsak előre, hanem hátra is haladhatunk, ekkor az indexeket -1, -2 stb. jelöljük.

Az idősorok elmélete más típusú adatokra is alkalmazható, pl. földszív mentén különböző pontokban mért nitrogéntartalom, amelyben az időbeli változás helyébe térbeli változás lép fel. Műtrágyadózisok is felfoghatók idősornak. A módszer felhasználható olyan esetben, ahol egy valószínűségi változó egy „ t ” változótól függ, ahol „ t ” egyaránt vonatkozhat időre vagy lineáris térre.

Az „ u ” változó lehet diszkrét, pl. emberek száma, és lehet folytonos változó, pl. hőmérséklet, légnyomás, stb.

Az idősornak négy összetevője lehet:

Trend, hosszú időszakon keresztül érvényesülő változás

Szezonális ingadozás, rövid ideig tartó szisztematikus ingadozás

Periodikus ingadozás, mely hosszabb időtávon jelentkezik

Véletlen ingadozás

Az idősorok analízise során ezt a négy összetevőt kell elkülöníteni, ami sokszor elég nehéz feladat.

Trend

Regressziós technikával valamilyen alkalmas függvény illesztése az adatokra. Lehet lineáris ill. nem lineáris, pl. polinomok illesztése. A magasabb fokú polinomok illesztése azonban sokszor hátrányos, mivel sok számítást igényel, és újabb tagok csatolása esetében az illesztést előlről kell kezdeni.

Mozgóátlagolás. A leggyakoribb a 3, 5, 7, 9, 15 és 21 pontos mozgóátlagolás. Ezzel a módszerrel a szezonális hatások kiküszöbölhetők.

Rövid lejáratú szezonális és véletlen összetevők

Feltételezzük, hogy az idősor trend mentes, vagy a trendet már korábban kiszűrtük (detrendelés). Sorozatunk ekkor többé-kevésbé szabálytalanul ingadozik valamilyen középponti érték körül.

A sorozat véletlenszerűségének vizsgálata

Vizsgáljuk meg, hogy milyen sorozatot várhatunk abban az esetben, ha az ingadozás teljesen véletlenszerű, azaz ha az egymást követő tagok függetlenek, és a sorozat egy ismeretlen sokaságból származó minta véletlen elrendezéseként fogható fel. Az ettől az állapottól való eltérést különböző mérőszámokkal mérhetjük, pl.:

Csúcsponatok és mélyponatok előfordulása a sorozatban

A szomszédos tagok közötti korreláció

Csúcsponatok és mélyponatok előfordulása a sorozatban

$u_{t-1} < u_t > u_{t+1}$ csúcsponat vagy $u_{t-1} > u_t < u_{t+1}$ mélyponat. Mindkét esetben u_t fordulópont. Két fordulópont közötti intervallumot fázisnak nevezünk. Egy oszcilláló idősor véletlenszerűsége a fordulópontok számának meghatározásával jól vizsgálható, ez n tagú véletlen sorozatban $\frac{2}{3}(n-2)$ várható értékű és $(16n-29)/90$ szórású.

A szomszédos tagok közötti korreláció, sorozatkorreláció

Egy sorozat szomszédos tagjai korrelációs együtthatóját elsőrendű autókorrelációs együtthatónak nevezzük. A k távolságra lévő tagok korrelációs együtthatójának elnevezése k -ad rendű autókorrelációs együttható.

$$\rho_k = \frac{\text{COV}(u_t, u_{t+k})}{D(u_t)D(u_{t+k})}$$

Hosszú sorozatban $D^2(u_t)$ és $D^2(u_{t+1})$ gyakorlatilag azonosak, és emiatt a fenti képlet a következő módon adható meg:

$$\rho_k = \frac{\text{COV}(u_t, u_{t+k})}{D^2(u_t)}$$

Rövid megfigyelési sorozatok esetében $D^2(u_t)$ becslésének jobb az egész sorozat (n tagból számított) szórásnégyzetét tekinteni, bár a kovarianciát csak $(n-k)$ tagból határozzuk meg. Hasonlóképpen jobb u_t és u_{t+k} szorzatösszegének meghatározásánál az u eltéréseket a teljes sorozat számtani közepétől mérni.

Amennyiben a sorozat tagjait az összes tag számtani közepétől mérjük, akkor:

$$r_k = \frac{n}{n-k} \frac{\sum_{t=1}^{n-k} u_t u_{t+k}}{\sum_{t=1}^n u_t^2}$$

Amennyiben a sorozat véletlen jellegű, akkor ρ_k elméleti értéke minden k -ra nulla. Ennélfogva a sorozatkorrelációs együtthatók nullától való eltérését felhasználhatjuk a sorozat véletlenszerűségének vizsgálatára. Véletlen sorozatban nagy n -re ρ_k szórásnégyzete közelítően:

$$D^2(r_k) \approx \frac{1}{n-k}$$

A ρ_k autókorrelációs együtthatót k függvényében ábrázoló görbét korrelogrammnak nevezzük. Ennek segítségével megkülönböztethetők a harmonikus sorozatok és az autoregresszív sorozatok.

Periodogram-elemzés

Számos oszcilláló fizikai jelenség bizonyos számú „tisztá” harmonikus hullámra bontható fel, amely mindegyike egy-egy szinusz vagy koszinusz függvénnyel írható le. Egy tiszta oszcillátor időbeli mozgása az $A \sin\left(\alpha + \frac{2\pi}{\lambda} t\right)$ függvénnyel fejezhető ki, ahol λ a hullámhossz és A az amplitúdó. Az oszcillációs jelenség pedig gyakran állítható elő ilyen tagok összegeként:

$$u_t = A_1 \sin\left(\alpha_1 + \frac{2\pi}{\lambda_1} t\right) + A_2 \sin\left(\alpha_2 + \frac{2\pi}{\lambda_2} t\right) + \dots +$$

Idősor periodicitásának keresése harmonikus analízis segítségével

Nincs zavar. Egy rádiókészülék behangolásával hasonlítható össze. Ismert hullámhosszú sorozatokat korrelálunk az adott sorozatokkal, ha összhangba jutnak, akkor intenzív korrelációt kapunk. Hibákkal terhelt, régi módszerek tartják napjainkban.

Autoregresszív sorozatok

Autoregresszív sorozatnak olyan sorozatot nevezünk, amely minden pontban az előző pontban felvett értékek, plusz egy zavar függvénye. Amennyiben a függvény lineáris, akkor lineáris autoregresszív függvényről beszélünk. A módszer figyelembe veszi, hogy zavar előfordulása esetén ez a rendszer változóba beleolvad. Nem szabályos ingadozáshoz hasonlít, melyet néha meglöknek. A kilengések közötti idő nem állandó, valamint a kilengés sem mindig azonos mozgású. Nagyon hasonló ahhoz, ahogyan sok oszcilláló idősor viselkedik. Ebből kifolyólag az autoregresszív sorozatnak nincs szigorú értelemben vett periódusa. A csúcspontok közötti átlagos távolság teljesen különbözhet a korrelogram periódusától. Tegyük fel, hogy egy rendszer mozgását két tényező határoz meg:

Belső tulajdonságainak összessége, pl. rugalmasság, kényszer, ezek a külső hatás nélküli mozgást határozzák meg

Külső lökések sorozata

Az autoregresszív sorozatokban a két legfontosabb eset:

$$u_{t+1} = \mu u_t + \varepsilon_{t+1} \quad (1.)$$

$$u_{t+2} + \alpha u_{t+1} + \beta u_t = \varepsilon_{t+2} \quad (2.)$$

Exponenciális simítás

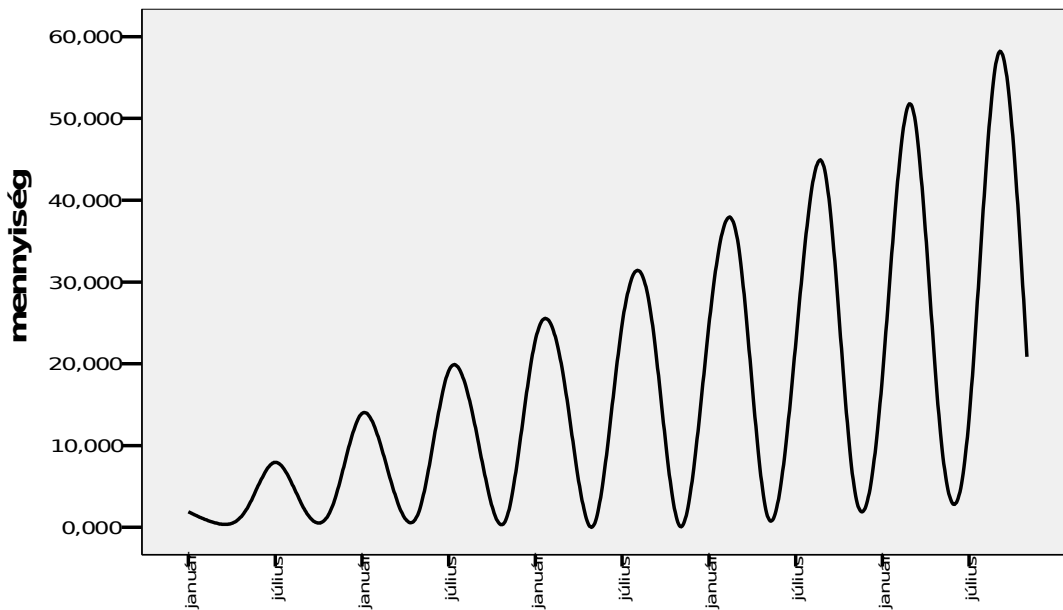
Négyféle modellt állíthatunk fel a trend és szezonaritás figyelembe vételének kombinációjával.

Egyszerű (Simple) modell: nincs trend és nincs szezonális hatás, ill. változás.

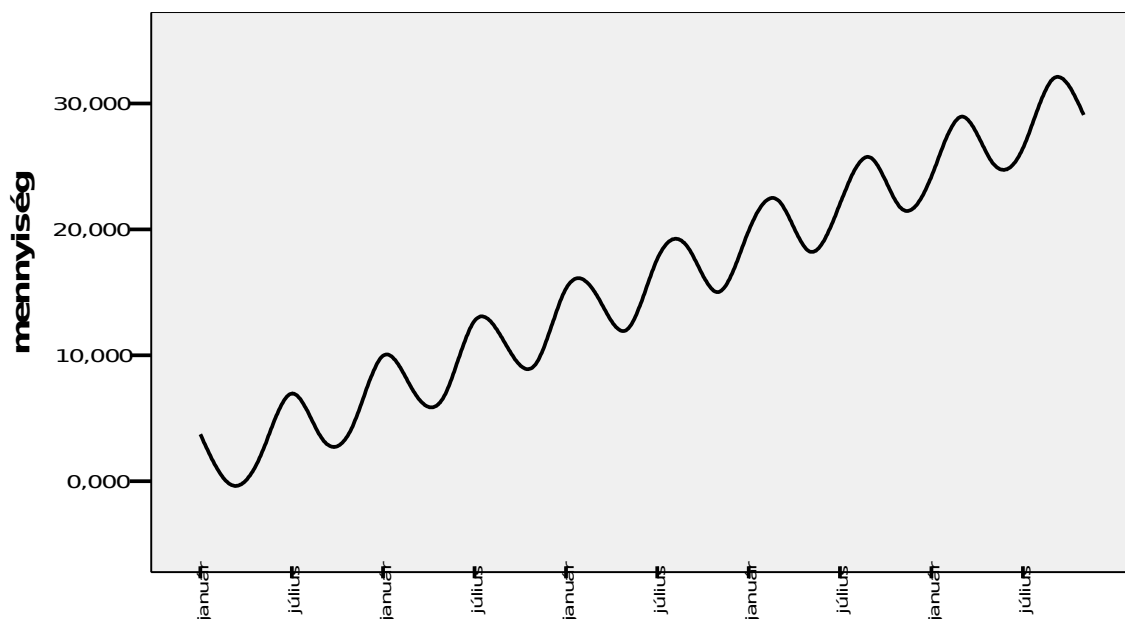
Holt modell: lineáris trend szezonális hatás nélkül.

Winters modell: lineáris trend és multiplikatív szezonális hatás. Az ingadozás nagysága nő vagy csökken a sorozat értékétől függően

Felhasználó által definiált modellek: a felhasználó állíthatja be a trend és szezonális hatásokat.



Multiplikatív szezonális dinamika, növekvő ingadozás



Additív szezonális dinamika

A fenti modellek négy paramétert használnak

Alfa (általános paraméter), minden modell használja, értéke 0,00-1,00. Ha az alfa 1, kizárólag a legfrissebb megfigyeléseket használjuk, ha alfa 0, akkor a régebbi megfigyelések is befolyásolják az aktuális érték alakulását.

Gamma Akkor használjuk, ha feltételezzük, hogy az idősnak van trendje. Értéke 0,00-1,00. A gammát csak lineáris vagy exponenciális trendnél, vagy csillapított trendnél, ahol nincs szezonális komponens, használjuk. Egyszerű modell esetében nincs értelme.

Delta. Ez a paraméter a szezonalitást írja le. Értéke 0,00-1,00, egyhez közeli értéke magasabb súlyt jelent. Csak szezonális hatást tartalmazó modellben kerül meghatározásra, nem használjuk egyszerű, ill. Holt modell esetében.

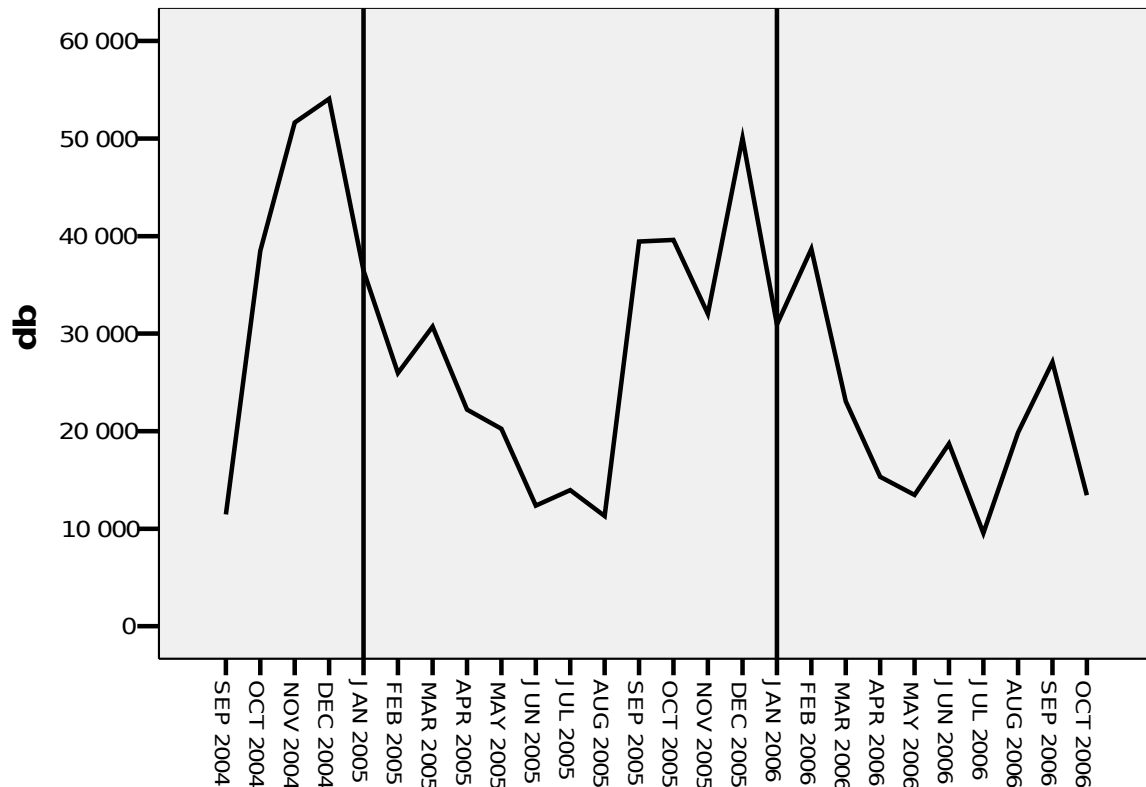
Phi. Az exponenciális simítás ezen paramétere kontrollálja, hogy a trend „damped”, csillapított vagy milyen gyorsan csökken a nagysága az idő függvényében. Értéke 0,00-1,00 (de soha sem éri el az egyet), egyhez közeli értéke nagyobb fokú csillapítást jelez. A phi csak csillapított trendet tartalmazó modellben használható, nincs értelme a szimpla, a Holt ill. a Winters modellben.

Az exponenciális simítás legelső lépése az ábrázolás, mivel a adatok időbeli alakulása segít a megfelelő modell kiválasztásában.

Van-e a sorozatnak egyáltalán trendje? Milyen a trend: változatlan vagy változik az idő függvényében?

Látható-e az adatokon szezonális ingadozás? A szezonális ingadozások idővel nőnek, vagy változatlanok.

Válasszuk a grafikonok közül a Szekvenciális grafikonokat.

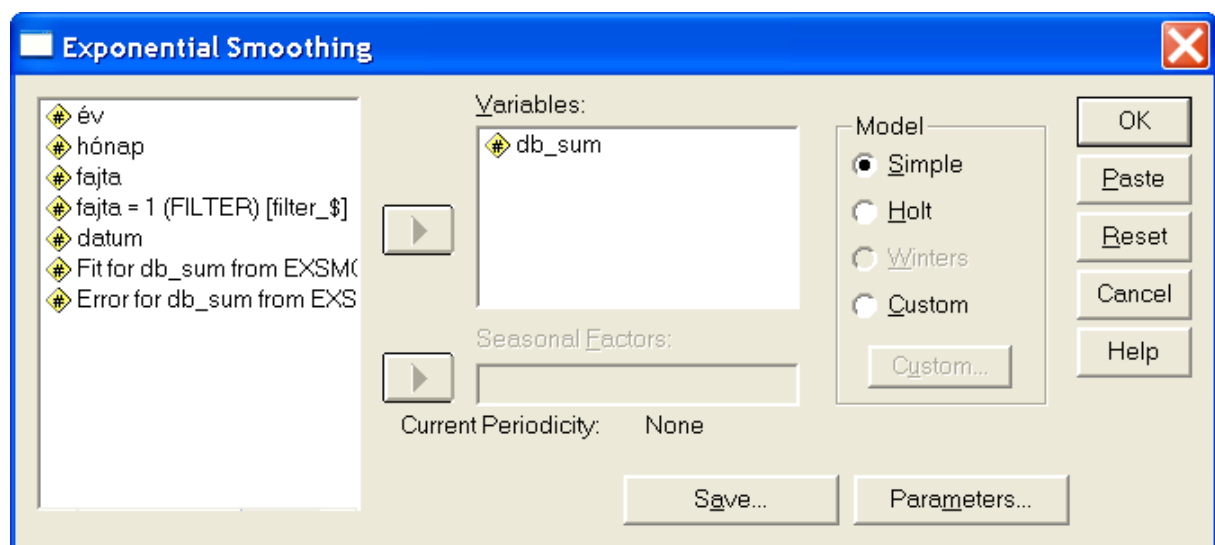


Filteres teafigyaszítás

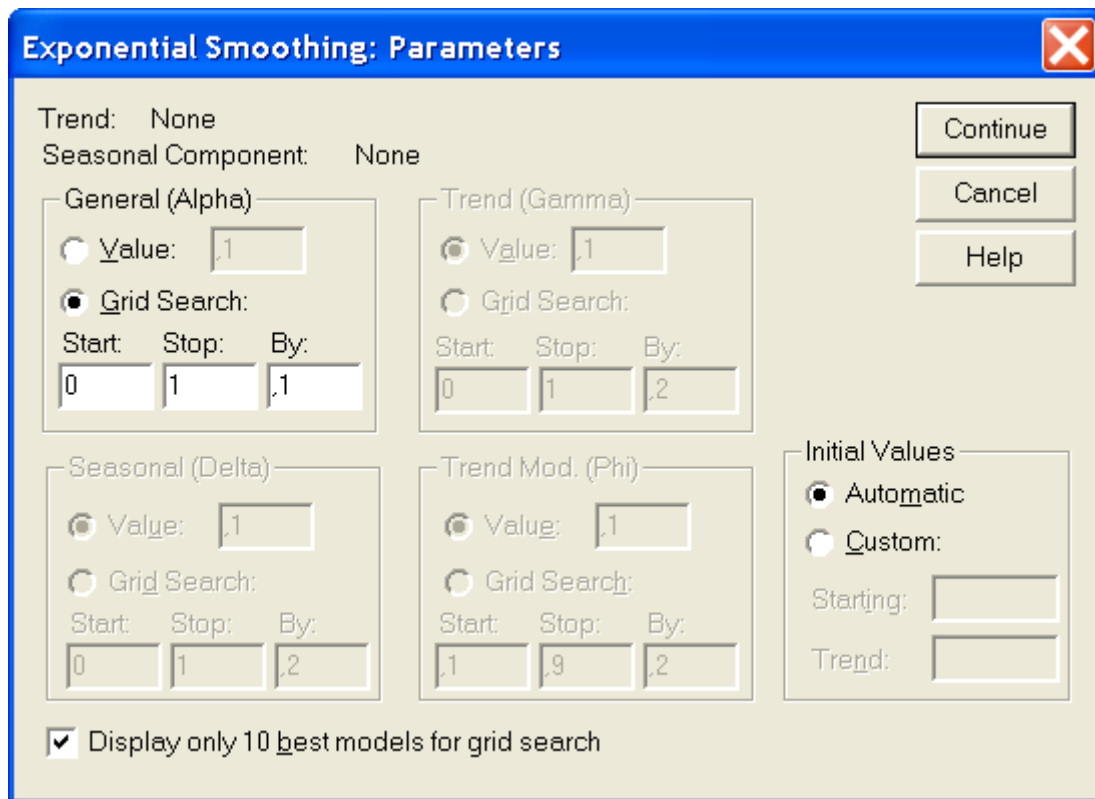
Trend nehezen ismerhető fel, vagy nincs, vagy enyhén csökkenő lineáris trendet feltételezhetünk. (nincs elég adat, hogy biztonságosan megítéljük konstans-e a trend).

Szezonális dinamika figyelhető meg: a hidegebb hónapokban több, a nyári időszakban kevesebb teát fogyasztanak az emberek.

A fentiek ismeretének ellenére legelőször a legegyszerűbb modellt állítsuk fel, ahol nincs trend és nincs szezonális hatás.



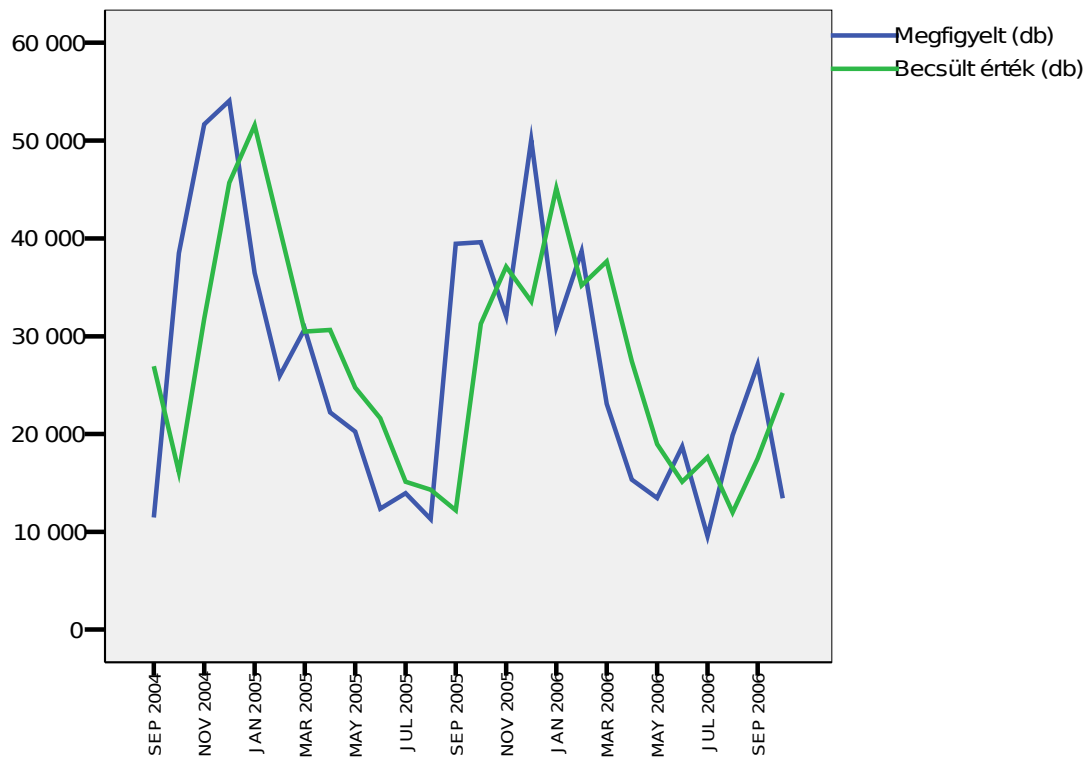
Az alfa paraméter meghatározását bízzuk a programra (válasszuk a Grid Search lehetőséget). Írassuk ki a legjobb 10 modell paramétereit.



A legjobb 10 modell alfa értéke az alábbi volt. A legpontosabb értéket $\alpha=0,7$ értéknél kaptuk, ekkor volt az eltérés négyzetösszeg a legkisebb.

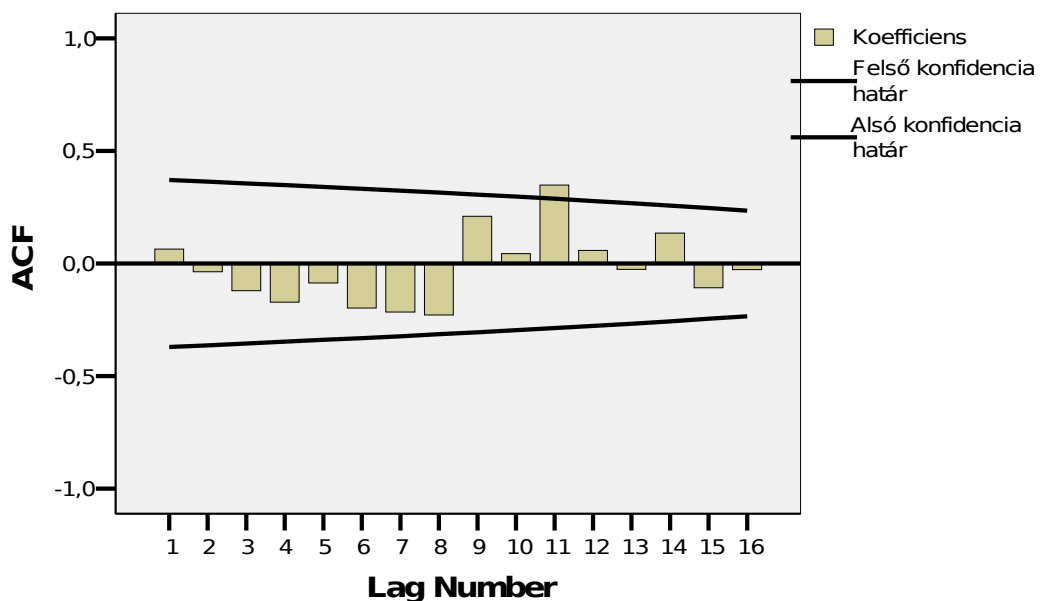
Smallest Sums of Squared Errors

Series	Model rank	Alpha (Level)	Sums of Squared Errors
db_sum	1	,70000	3907879011,54776
	2	,80000	3910177609,03472
	3	,60000	3965966440,05588
	4	,90000	3974185801,05280
	5	,50000	4081471836,26282
	6	1,00000	4103872124,85208
	7	,40000	4242310694,02872
	8	,30000	4418217451,93448
	9	,00000	4485032083,84616
	10	,20000	4560252585,14987



Az egyszerű modellel kapott becült és tényleges forgalmi adatok

Az exponenciális simítás szépen leírja az adatokat, azonban van egy időbeli elcsúszás, ami több ezer darabos alá és felül becslést jelent. Ezért érdemes a maradékok autókorrelációs grafikonját elkészíteni, és megvizsgálni, hogy van-e valamilyen jól felismerhető szezonális hatás.



Az ábrán jól látható, hogy van egy 11 hónapos szignifikáns hatás, ami erős szezonális hatást jelent.

A becsült adatok rossz illeszkedése, és a 11 hónapos autókorrelációs együttható miatt az egyszerű exponenciális modell nem alkalmas a teakereslet pontos előrejelzésére.

A szezonális hatás felbontása

A szezonális analízis négy új változót hoz létre, amelyek az adatbázisban megjelennek és az alábbi kezdőbetűkkel azonosíthatók:

SAF. Szezonális faktorok, a szezonális változásokat mutatják. A multiplikatív modellben az 1 érték a szezonális ingadozás hiányát mutatja. Az additív modellben ugyanezt a 0 érték jelenti. A szezonális faktorokat használhatjuk inputként az exponenciális simítás modelljeiben.

SAS. A szezonális hatástól megtisztított eredeti idősor. Ezzel a sorozattal trend-analízist, vagy más független szezonális összetevő kimutatását végezhetjük el. Trend meghatározása regresszió-analízis segítségével függő változóként lehet megadni. Autoregresszió számítása.

STC. Rövidebb trend-ciklus összetevők. Trend meghatározása regresszió-analízis segítségével függő változóként lehet megadni. Autoregresszió számítása.

ERR. Maradék tagok.

A szezonális hatások leválasztásával az egyszerű szezonális hatásokat távolíthatjuk el a ciklikus idősorokból.

GRAFIKONOK

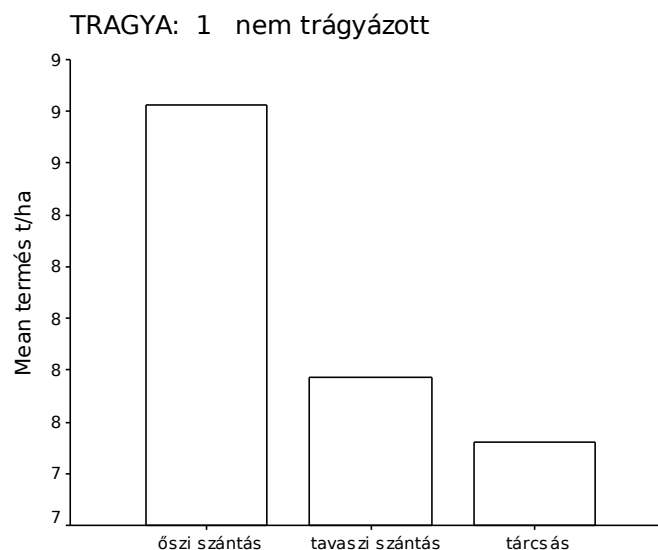
Grafikon készítésekor nemcsak az alapadatokat ábrázolhatjuk, hanem a csoportképző változó szerint összesített, számított értékeket is. Pl. napi hőmérsékleti átlagok, minimumok, maximumok ábrázolása.

Oszlop diagramok (Bar Charts)

Egyszerű (Simple)

Csoportosított megfigyelések ábrázolása (Summaries for groups of cases):

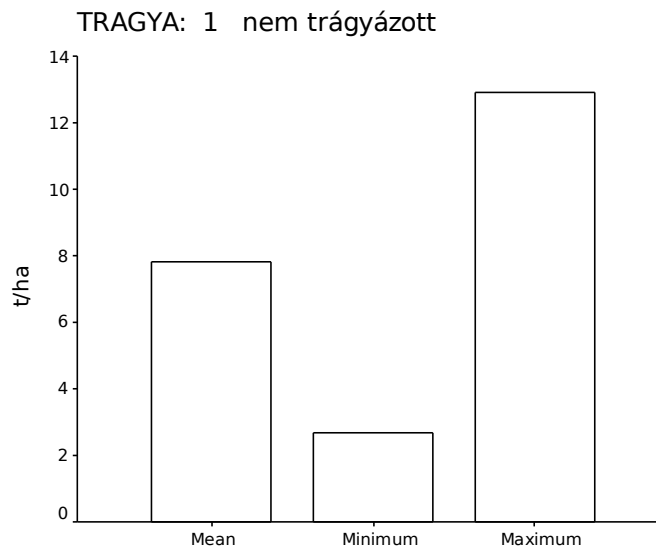
A kategória tengelyen (Category Axis) a csoportképző változó szerepel, pl. a kezelés (öntözés, hibrid, trágyázás, stb.). Az oszlopok mutathatják a kezelésszintek megfigyeléseinek, eseteinek számát, kumulált értékeit és ezek százalékos részesedéseit. Egy függő változót kijelölve különböző statisztikai mutatókat ábrázolhatunk a csoportképző változó függvényében.



115. ábra: A kukorica termése (t/ha) különböző talajművelésekben

Különböző változók ábrázolása egy diagramon (Summaries of separate variables):

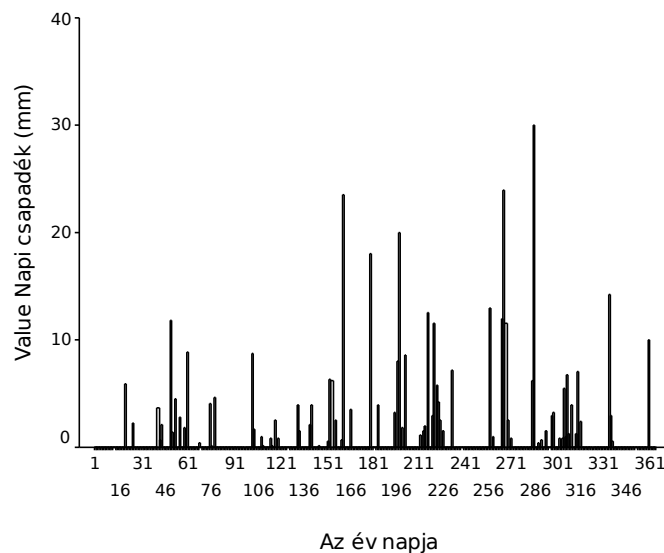
Csoportképző változó nélkül több változót, vagy ugyanannak a változónak a különböző statisztikai mutatóit ábrázolhatjuk a grafikonon.



116. ábra: A termés (t/ha) különböző statisztikai mutatói

A megfigyelt értékek ábrázolása (Values of individual cases):

A változó minden egyes értékét ábrázolhatjuk. A megfigyelések száma nem lehet több, mint 3000. A megfigyeléseknek, eseteknek magyarázatokat, címkéket is adhatunk.

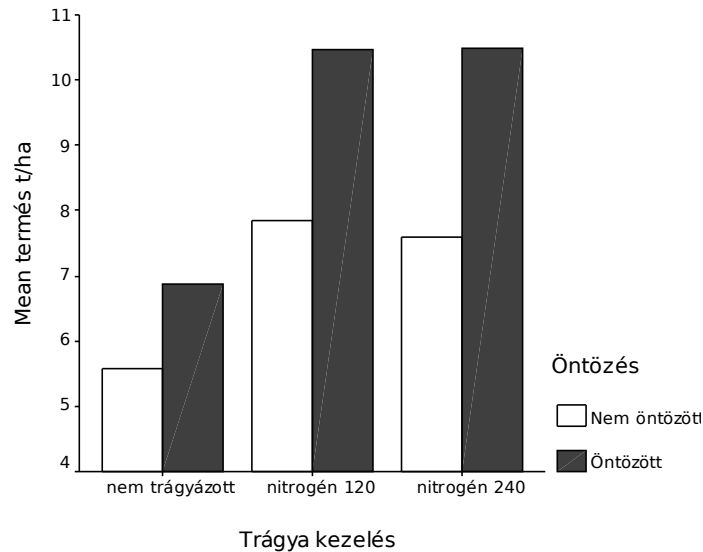


117. ábra: 2002. év napi csapadékadatai (mm)

Csoportosított (Clustered)

Csoportosított megfigyelések ábrázolása (Summaries for groups of cases):

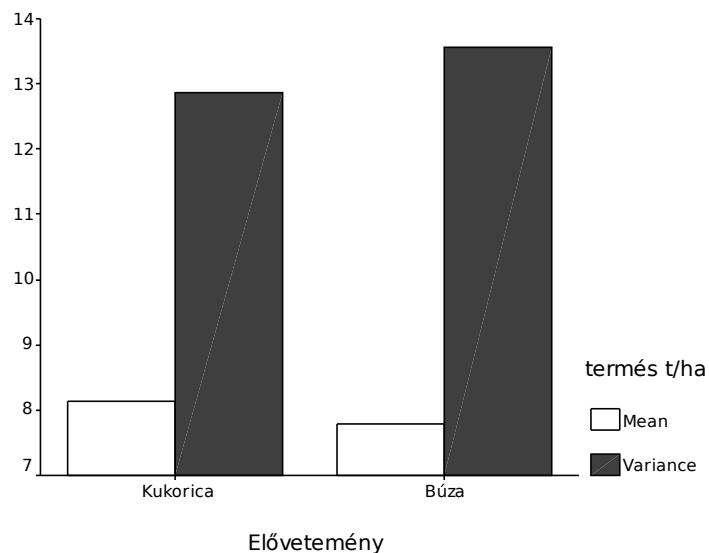
Egy változó különböző statisztikai jellemzőinek ábrázolása két ismérv alapján. A kategória tengelyen a trágyázás, klaszterként az öntözés szerepel.



118. ábra: A trágyázás hatása a kukorica termésére nem öntözött és öntözött kezelésekben

Különböző változók ábrázolása egy diagramon (Summaries of separate variables):

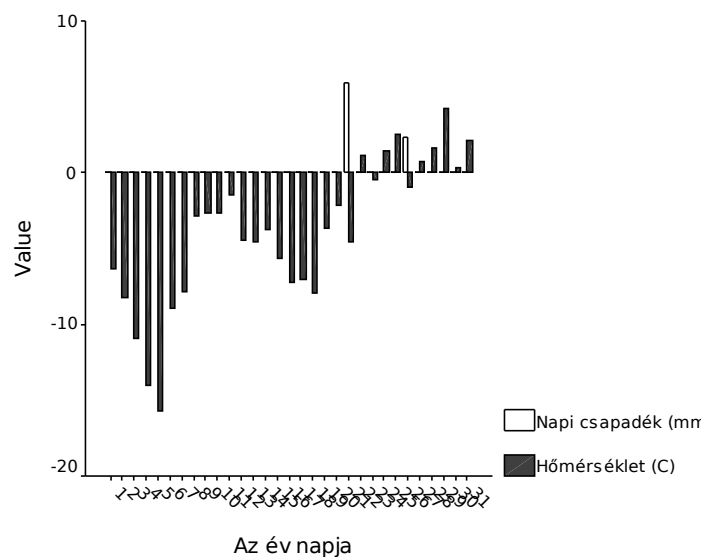
Több változót, vagy ugyanannak a változónak a különböző statisztikai mutatóit ábrázolhatjuk a grafikonon. A csoportképző változó a kategória tengelyen jelenik meg.



119. ábra: Az elővetemény hatása a kukorica termésére és varianciájára

A megfigyelt értékek ábrázolása (Values of individual cases):

Több változó minden egyes értékét ábrázolhatjuk. A megfigyelések száma nem lehet több, mint 3000. A megfigyeléseknek, eseteknek magyarázatokat, címkéket is adhatunk.



120. ábra: 2002. év január havi napi hőmérséklet és csapadékadatai

Halmozott (Stacked)

Egyetlen változó számított értékeinek ábrázolása vonaldiagram segítségével: (Graphs, Line Charts Simple)

... Summaries for groups of cases, Define. A változó kiválasztása után megadható az összesítés módja (statisztikája). A kategória tengelyen a csoportképző változót kell megadni.

... Summaries of separate variables

... Values of individual cases: a változó minden egyes előfordulását ábrázolja, nem számít statisztikát.

Többszörös ábrázolás: (Graphs, Line Charts, Multiple)

... Summaries for groups of cases. Két csoportképző ismérv szerint ábrázolhatjuk a kiválasztott változókat, pl. kukorica terméseket a termőhely és idő függvényében. Az x-tengely (Category Axis) lehet az idő, pl. év, a vonalak (Define Lines by) pedig a termőhelyenkénti termések.

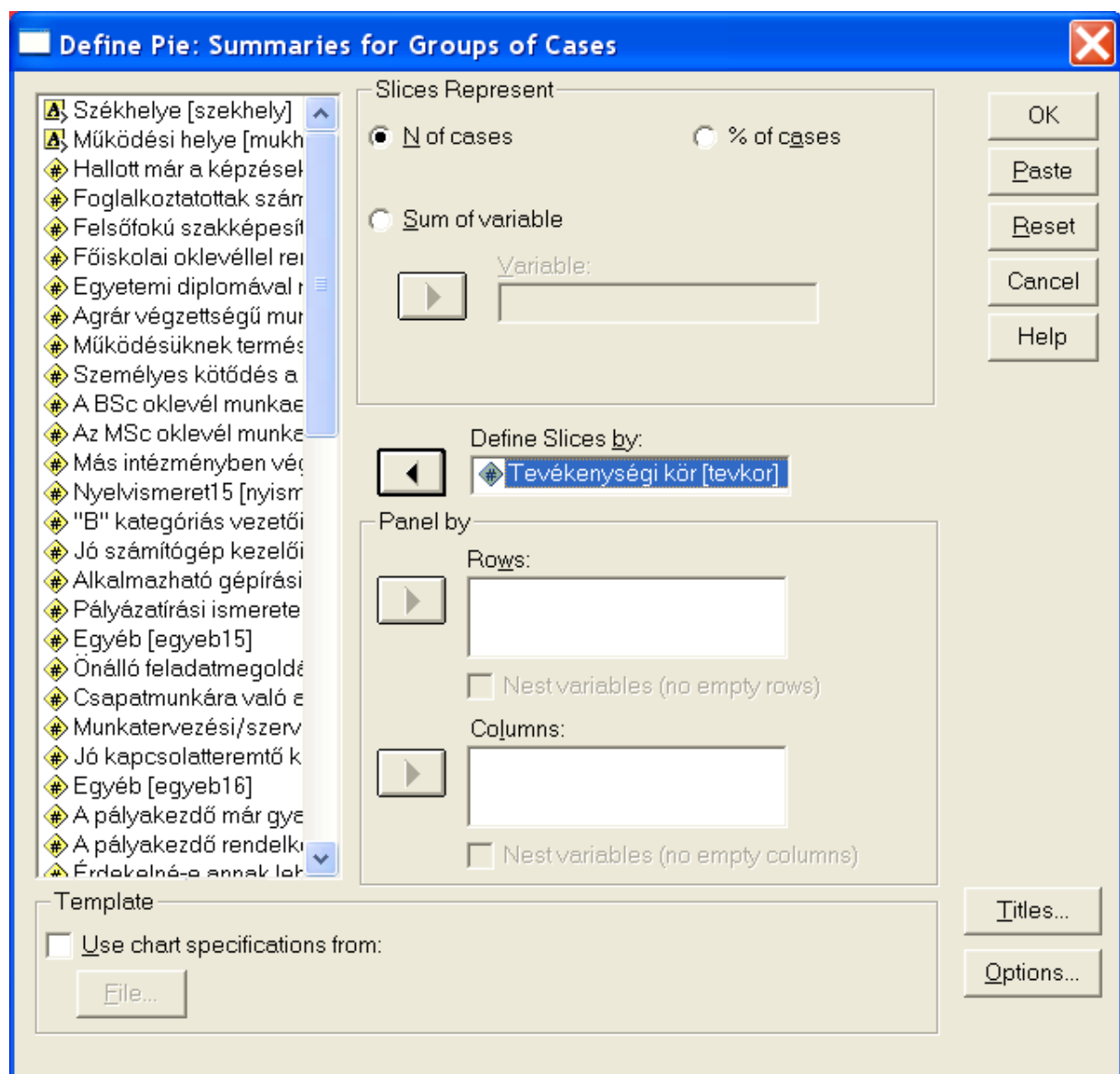
... Summaries of separate variables, Define. A Lines Represent ablakban az ábrázolandó változókat lehet megadni. A változókhöz különböző összesítési eljárások választhatók. Jelöljük ki egérrel a változót (kék szín) és Change Summary gombbal adjuk meg a számítási eljárást (átlag, medián, módusz, esetek száma, összeg, szórás, variancia, minimum, maximum, kumulatív összeg). Lehetőség van különböző százalékokban is megjeleníteni a változó értékeit.

... Values of individual cases:

Minden grafikonnak azonos formátumot biztosíthatunk, ha a mintát (template) alkalmazunk. A minta *.sct fájlban található. Figyelem: bonyolult elérési útvonal esetén nem mindig találja meg a fájlt. Érdekes az SPSS alkönyvtárban tárolni ezeket a fájlokat.

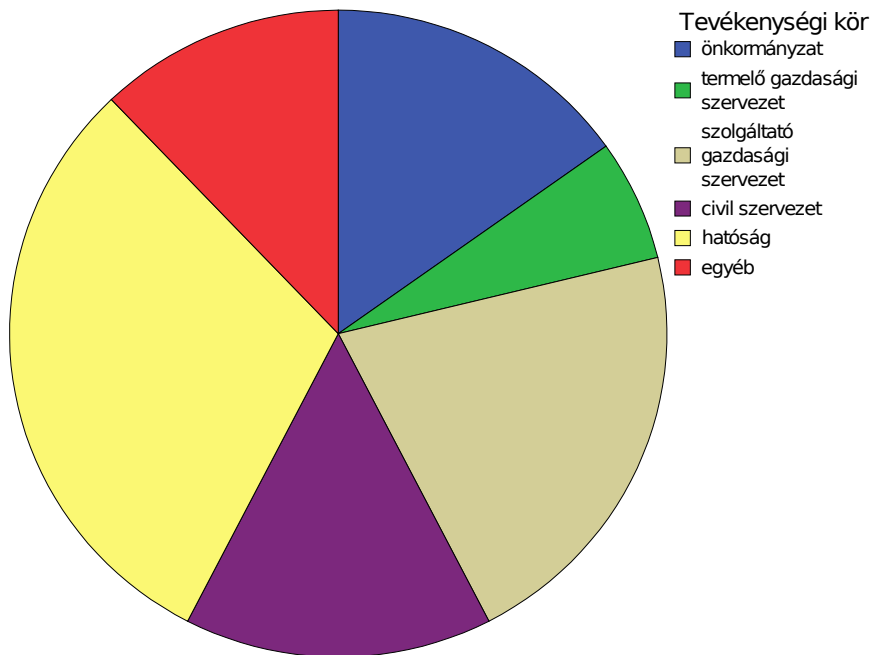
Kördiagramok (Pie Charts)

Kördiagramot főként nominális változók gyakoriságának, vagy egy változó összetételének bemutatására használunk.

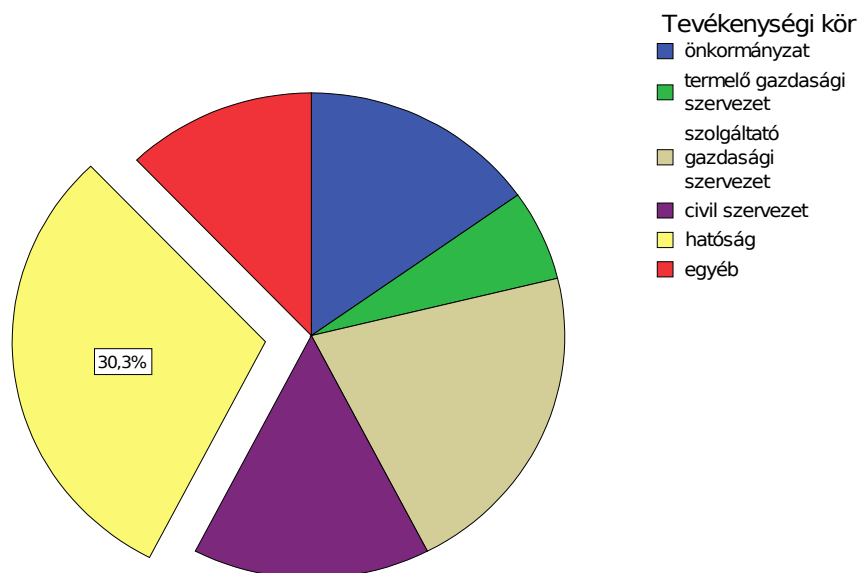


121. ábra: A kördiagram beállítása

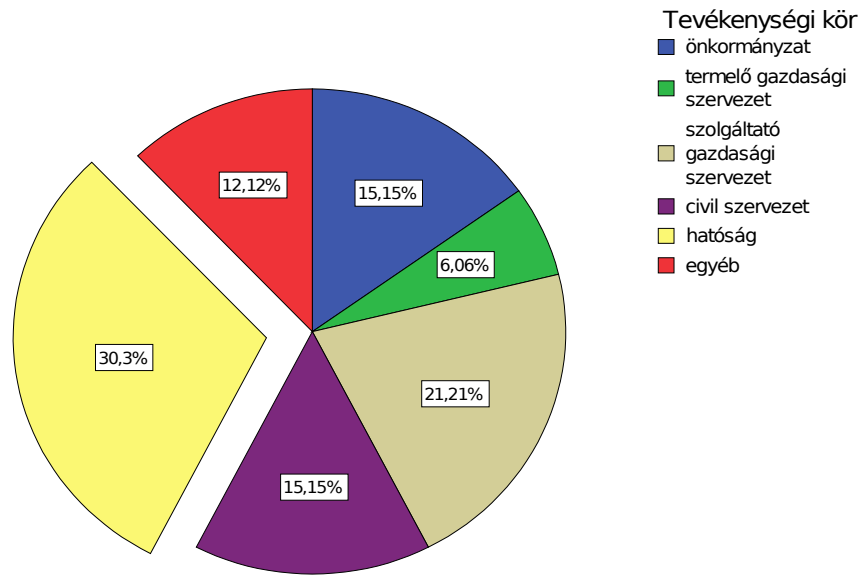
Készíthetünk egyszerű, úgynevezett robbantott ill. különböző információkkal kiegészített kördiagramokat.



122. ábra: Egyszerű kördiagram



123. ábra: Robbantott kördiagram, a leggyakoribb érték jelölésével



124. ábra: Robbantott kördiagram, a százalékok feltüntetésével

KÉRDŐÍVEK TERVEZÉSE

A tesztkészítéssel, tervezéssel és kiértékeléssel foglalkozó tudomány a tesztelmélet. Ebben a fejezetben nem akarok a teljességre törekedni, mert akkor több száz oldalt kellene igénybe venni. Csak az SPSS programhoz szükséges alapvető ismereteket tárgyalom, és megpróbálok némi gyakorlati útmutatót adni a helyes kérdőív kiértékeléshez.

A szakemberek nagyon sokféle kérdést különböztetnek meg, pl.:

- Igen/Nem kérdés
- Nyitott kérdés (1 soros válasz)
- Nyitott kérdés (több soros válasz)
- Nyitott kérdés (számjegyes válasz)
- Többszörös nyitott kérdés (számjegyes válasz).
- Egyszerű választás (egy válaszlehetőség)
- Többszörös választás (több válaszlehetőség)
- Mátrix-kérdés (soronként egy válaszlehetőség)
- Mátrix-kérdés (soronként több válaszlehetőség)
- Értékelő kérdés
- Többszörös értékelő kérdés
- Osztályozó kérdés 1-től 10-ig
- Többszörös osztályozó kérdés 1-től 10-ig A kérdőív létrehozása
- Időpontra vonatkozó kérdés
- Érték-relevanciakérdés

És még biztosan lehetne kitalálni még egy párat. A könnyű eligazodás érdekében le fogom egyszerűsíteni a kérdések csoportosítását, főként az adatbázis tulajdonságaik alapján, mivel a különböző típusú kérdéseket különbözőképpen kell beépíteni az adatbázisba. Vannak olyan kérdések, amikre csak egyetlen választ lehet adni, pl. a lenti kérdés (melyik korosztályba tartozik), és vannak többszörös válaszadásúak is. A válaszadó csak egyetlen korcsoportba tartozhat. Az ilyen típusú válaszokat rádiógombokkal szokták jelezni, ezzel is sugalmazva, hogy csak egyetlen választ vár a kérdést feltevő személy. Az adatbázisban ezt egyetlen nominális vagy ordinális típusú változóban tárolhatjuk. Érdeemes számokkal kódolni az egyes korosztályokat, és címkéket használni a megnevezésükhöz.

Melyik korosztályhoz tartozik?

- 0-20 év
- 21-40 év
- 41-60 év
- 61-80 év
- 80 felett

Az alábbi kérdésre (hogyan van megelégedve a munkahelyével) is csak egyetlen válasz adható. Ez egy minősítő, eldöntendő kérdés. Az adatbázisban ezt is egyetlen változóban tároljuk, ordinális adatként. Értéke lehet szöveg vagy szám. Célszerű számokat megadni, és címkéket használni, mivel így sokkal kisebb méretű adatbázist kapunk.

„Hogyan van megelégedve a munkahelyével?”

- Nagyon
- Közepesen
- Kevésbé

Az előző két kérdéstípusra adott válaszokat tehát egyetlen változóban kell tárolni. Az olyan kérdéseket, ahol több válasz is lehetséges, kicsit bonyolultabb az adatbázisba elhelyezni. Ilyen típusú kérdés az alábbi:

„Van-e a lakásban?”

- Vezetékes víz
- Központi fűtés
- Telefon
- Színes televízió
- Számítógép

A válaszoló akár mindet megjelölheti vagy egyiket sem. Ilyenkor a lehetséges válaszok mindegyikére egy-egy dichotóm (két-értékű, 0=nincs, 1=van) változót képezünk. Ezek a változók egy csoportot alkotnak, érdemes a változók nevével is jelezni a csoportba tartozást. Pl. ha a fenti kérdés a nyolcadik, akkor a válaszokat K8_1, K8_2, K8_3 stb. jelölhetjük. A többszörös válaszadások elemzése bonyolultabb, mint az egyszerű választó kérdések kiértékelése.

Milyen sorrendben tartalmazzák a kérdéseket a kérdőívek?

Ez elég szubjektív, függ a kérdőív típusától. Általában a közvélemény kutatásban használt kérdőívekben az elején egyszerű, figyelemfelkeltő kérdések vannak. A bizalmas jellegű, személyes kérdések a kérdőív végére

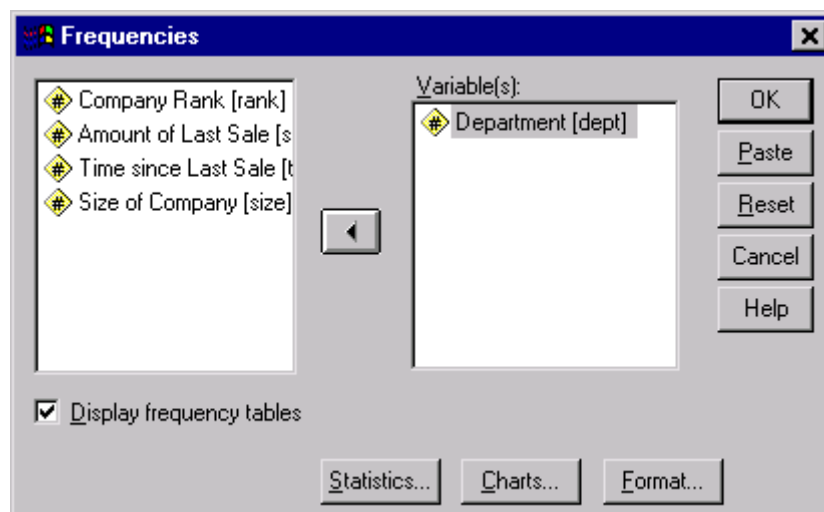
kerülnek. Bizalmatlanságot és ellenkezést válthat ki az olyan kérdőív, aminek az első kérdése azt firtatja, hogy hány évesek vagyunk és mennyit keresünk, kövérek vagyunk-e vagy soványak. Egy jó kérdőív betartja az íratlan udvarlási szabályokat. Előbb az érdeklődést, szimpátiát, bizalmat kell megszerezni, és csak utána jöhetnek az esetleges bizalmas kérdések.

KÉRDŐÍVEK KIÉRTÉKELÉSE

Először az egyetlen választ adó kérdések értékelését mutatom be. Ezek a válaszok lehetnek nominális, ordinális és skála típusú adatok.

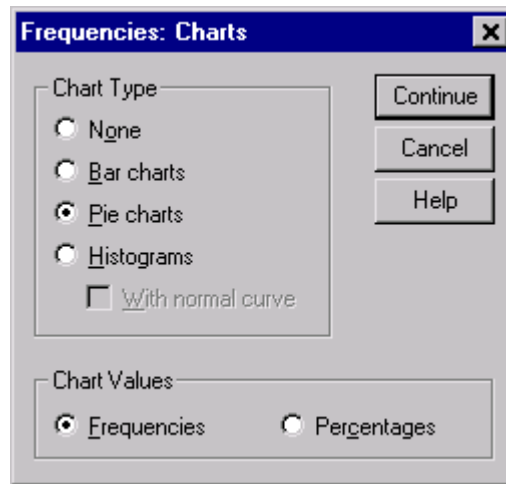
Nominális típusú adatok kiértékelése

Az elemzést az SPSS-hez mellékelt *contact.sav* adatbázison mutatom be. Nyissuk meg az adatbázist és válasszuk Analyze, Descriptive Statistics, Frequencies menü pontot. Ezután a Department változót tegyük a változók ablakba.



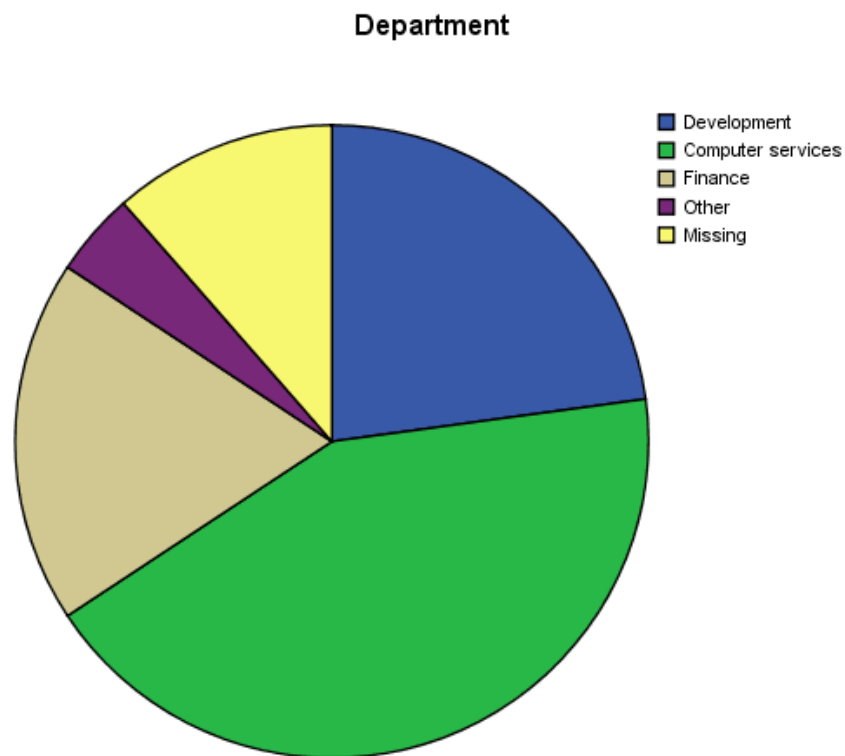
125. ábra: Gyakoriságok párbeszédpanel

Kattintsunk a Charts... gombra (grafikonok), és a grafikon típusának (Chart Type) adjuk meg a kördiagramot (Pie Charts). A diagram ábrázolhatja az adatok gyakoriságát (az előfordulás számát) vagy százalékát. Ezt a Chart Values területen tudjuk beállítani.



126. ábra: Grafikonok kiválasztása és beállítása

A folytatáshoz kattintsunk a Continue gombra.



127. ábra: Kördiagram, a kategóriák feltüntetésével

A kördiagram szemléletesen ábrázolja a különböző kategóriák relatív gyakoriságát a megfigyelések egészéhez viszonyítva. A gyakorisági táblázat pontosan megmutatja az egyes kategóriák gyakoriságát (Frequency), százalékban kifejezve (Percent), az érvényes megfigyelések százalékában kifejezve (Valid Percent) és a kumulatív eloszlást százalékban (Cumulative

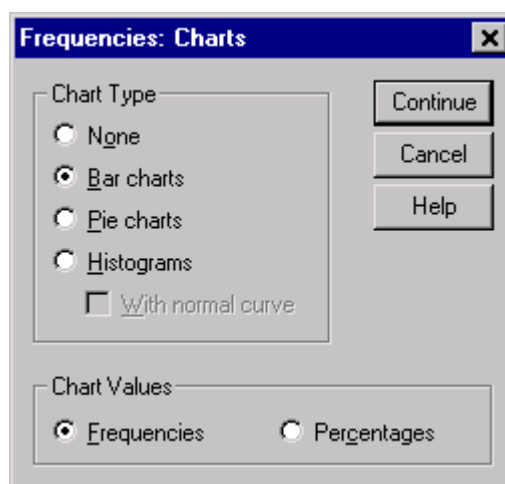
Percent). Az első oszlopban a Valid az érvényes megfigyeléseket, a Missing a hiányzó értékeket jelöli. Az adatbázisban összesen 70 megfigyelés szerepel. Ebben 62 érvényes és 8 hiányzó adat van. Általános érvényű következtetés levonásához mindig az érvényes adatokat kell figyelembe venni. Nominális típusú adatoknál a kumulatív eloszlás nem ad többlet információt, gyakorlatilag nem is lehet használni semmire. Ez inkább az ordinális típusú adatok esetén hasznos.

		Department			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Development	16	22.9	25.8	25.8
	Computer services	30	42.9	48.4	74.2
	Finance	13	18.6	21.0	95.2
	Other	3	4.3	4.8	100.0
	Total	62	88.6	100.0	
Missing	Don't know	8	11.4		
Total		70	100.0		

A leggyakoribb adat a Computer Services, az érvényes adatok százalékában 48,4%.

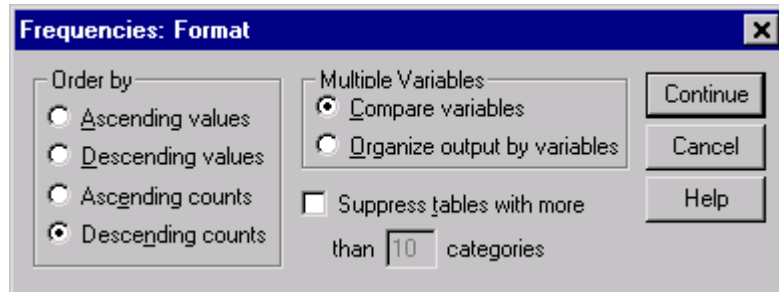
A kördiagram helyett oszlopdiagramot választva, amiben a kategóriákat csökkenő gyakorisággal ábrázoljuk, gyorsan megállapíthatjuk a sokaság móduszát (leggyakrabban előforduló kategória), illetve a relatív gyakoriságot szemléletesen ábrázolhatjuk.

Válasszuk újból a grafikonok menü pontot, és most az oszlop diagramot aktivizáljuk.



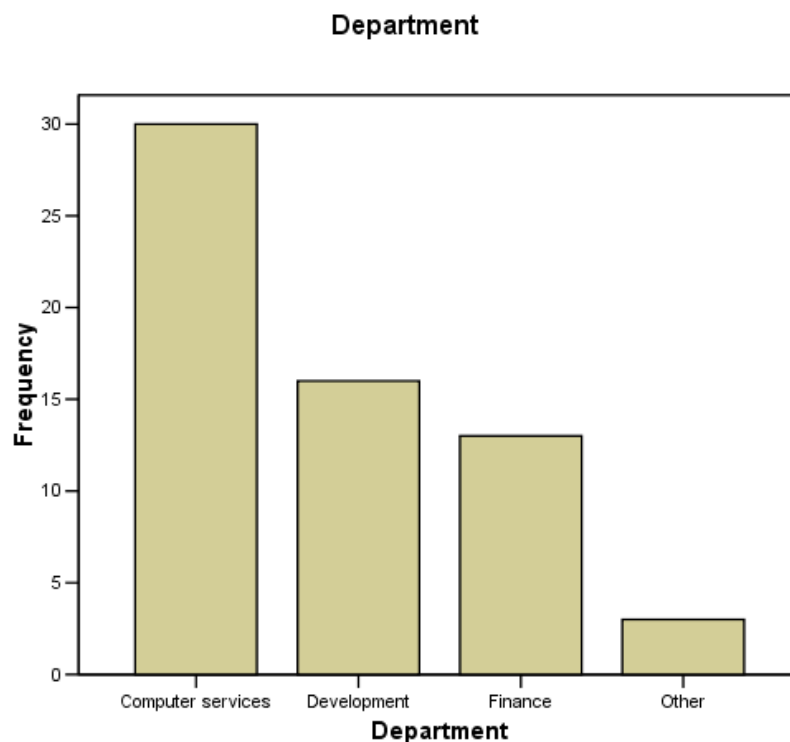
128. ábra: Oszlopdiagram kiválasztása

A folytatás után kattintsunk a Format gombra, ahol a frekvenciák táblázatos megjelenítésének és ábrázolásának módját lehet beállítani. A gyakoriságok megjelenítését végezzük a kategóriagyakoriságok csökkenő sorrendjében (Descending Counts). A Continue után megkapjuk az oszlopdiagramot.



129. ábra: A gyakoriságok megjelenítésének beállítása

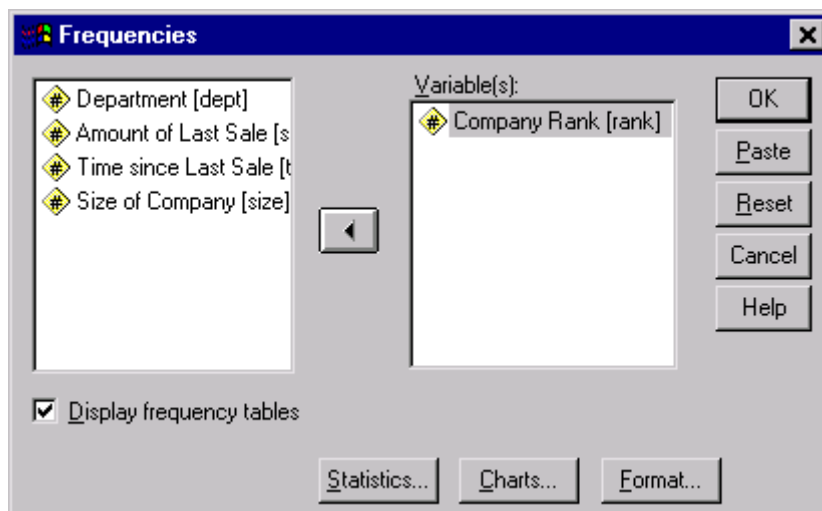
A gyakoriságok az előfordulások nagyságának csökkenő sorrendjében jelennek meg.



130. ábra: Oszlopdiagram, csökkenő gyakorisági sorrend

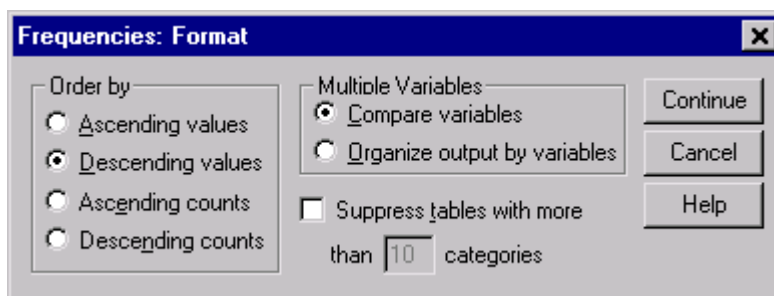
Ordinális típusú adatok kiértékelése

Olyan adatok kiértékelése, amik valamilyen szempont alapján sorba rendezhetők, valamilyen rangsor felállítható közöttük. Ugyanazt az eljárást fogjuk használni, mint az előbb, csak a beállítások lesznek mások. Az adatbázis alapján elemezzük a cég rangsor változót. Tegyük be a Company Rank változót a vizsgálati ablakba. Ez egy ordinális típusú változó, amit a változó definiálásakor nekünk kellett beállítani az SPSS adatbázis ablakában.



131. ábra: Gyakoriságok elemzése

Készítsünk oszlopdiagramot.. A Format... beállításai az alábbiak lesznek: csökkenő rendezés a változó értékei szerint. Tehát nem az előfordulás gyakorisága szerint, hanem a rangsorban elfoglalt értéke alapján fognak megjelenni a frekvenciák. A rangsorban legértékesebb kategória gyakorisága fog az első helyen (balra) megjelenni, és utána a többi.



132. ábra: Gyakoriságok megjelenítése csökkenő értékek szerint

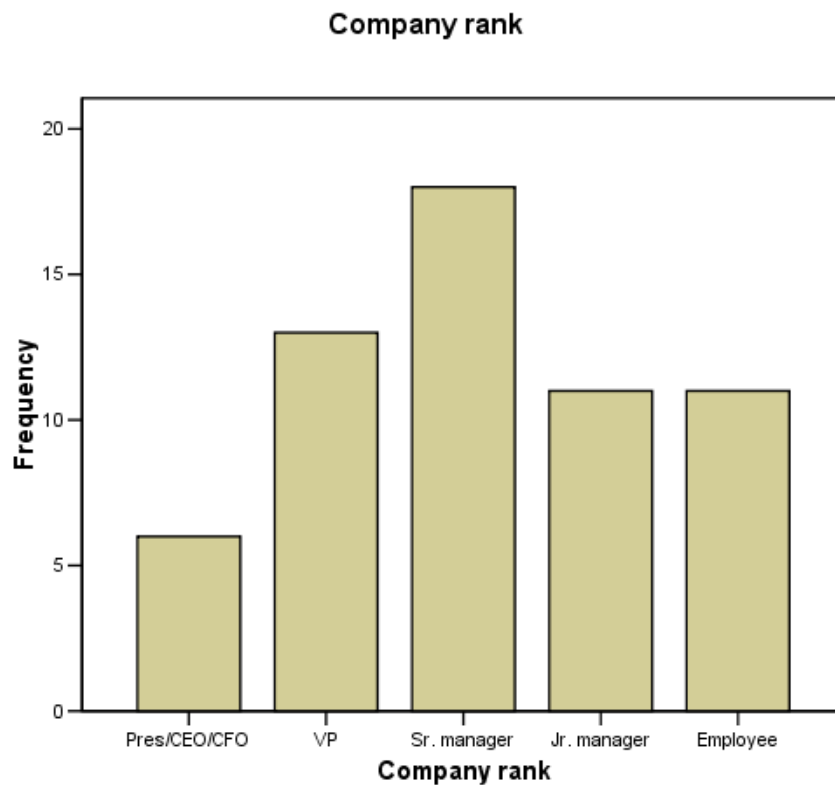
Ez alapján fog elkészülni a frekvenciatáblázat és az oszlopdiagram. A gyakorisági táblázat fentről lefelé csökkenő rangszámok alapján mutatja a

frekvenciákat. Képzelnék el, hogy iskolai osztályzatokat értékeltünk. Ekkor a jeles áll az első helyen és az elégtelen az utolsón. Jeles – csak az érvényes megfigyeléseket figyelembe véve – 10,2%, jó 22, közepes 30,5%, elégséges 18,6% és végül elégtelen szintén 18,6%. A kumulatív eloszlásból (Cumulative Percent) egyéb értékes megállapítások is tehetők. Pl. a legalább közepes eredményt elért hallgatók aránya 62,7%. Sikeres vizsgák aránya 81,4% és így tovább.

Company rank

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Pres/CEO/CFO	6	8.6	10.2	10.2
	VP	13	18.6	22.0	32.2
	Sr. manager	18	25.7	30.5	62.7
	Jr. manager	11	15.7	18.6	81.4
	Employee	11	15.7	18.6	100.0
	Total	59	84.3	100.0	
Missing	Don't know	11	15.7		
Total		70	100.0		

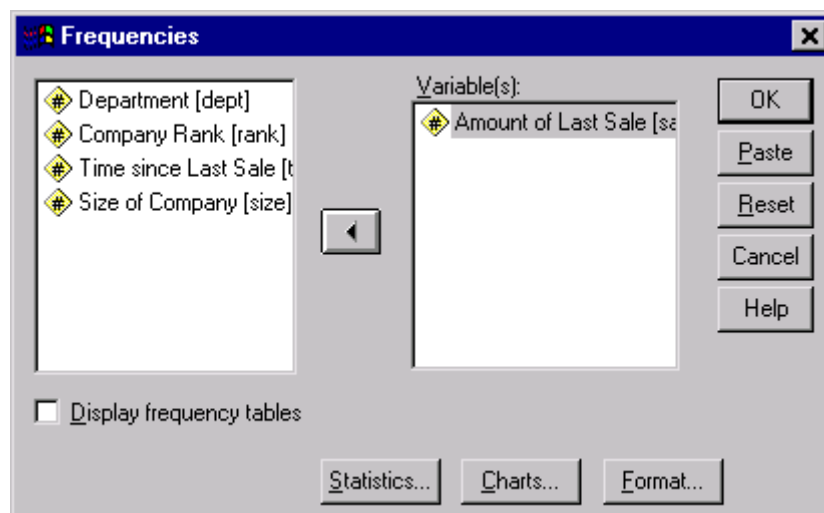
Az oszlop diagram a gyakoriságokat mutatja az „osztályzatok” csökkenő sorrendjében. Leggyakoribb osztályzat a közepes, jelesből van a legkevesebb, stb.



133. ábra: Oszlopdiagram, gyakoriságok csökkenő érték kategória szerint rendezve

Skála típusú adatok kiértékelése

Itt is ugyanazt a programot fogjuk használni, mint az előbb. A skála típusú adat valamilyen fizikai mennyiséget jelöl, legtöbbször mértékegységgel is rendelkezik. Az előbbi adatbázis segítségével a legutóbbi értékesítési eredményeket fogjuk elemezni. Kattintsunk a Reset gombra, hogy alaphelyzetbe hozzuk a párbeszédablakot. Ezután válasszuk ki a Amount of Last Sale változót és tegyük be a változók ablakba. Ne készítsünk frekvencia táblázatot, mivel skála típusú adatnál nagyon sok „kategória” van, ezért a Display Frequency Tables jelölőnégyzetet töröljük.



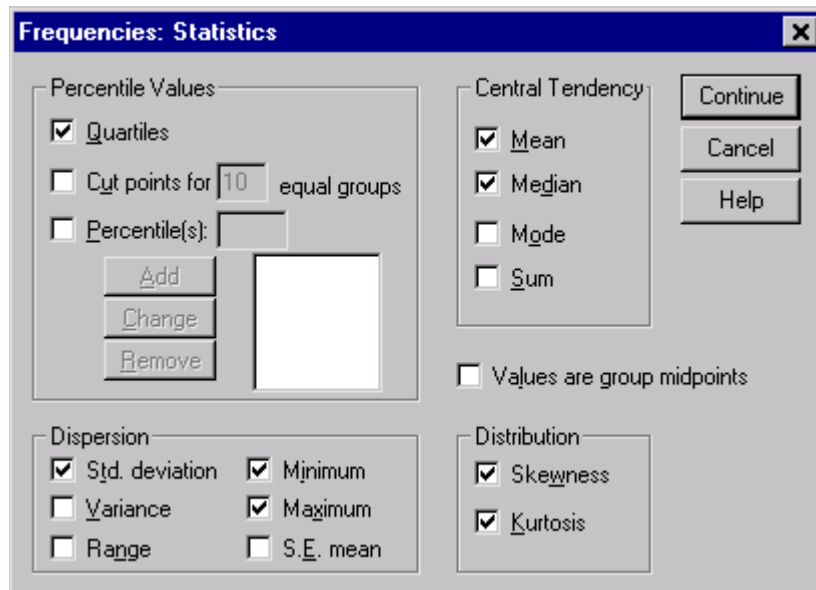
134. ábra: Skála típusú adat gyakoriságának elemzése

Valószínűleg kapunk egy hibaüzenetet, ami arra figyelmeztet, hogy minden outputot letiltottunk. Ez nem baj, majd a későbbiekben beállítjuk amire szükségünk van. Kiklikeljük az OK-ra.



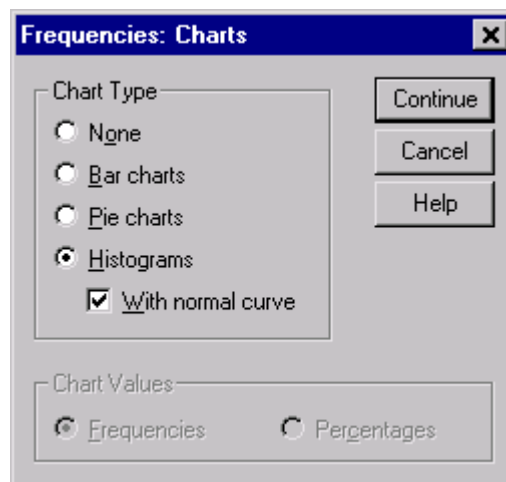
135. ábra: Az SPSS figyelmeztető üzenete

Kattintsunk a Statistics gombra a Frequencies párbeszédablakban. Válasszuk ki a kvartiliseket, a szórást, a minimumot és maximumot, az átlagot, a mediánt, a ferdeséget és végül az eloszlás csúcsosságát.



136. ábra: A különböző statisztikák beállítása

Klikk Continue. Válasszuk ki a grafikonok menüből a hisztogramot és jelöljük be a normál eloszlás jelölőnégyzetét. Continue.



137. ábra: Hisztogram beállítása, a normál eloszlás görbéjének kiválasztása

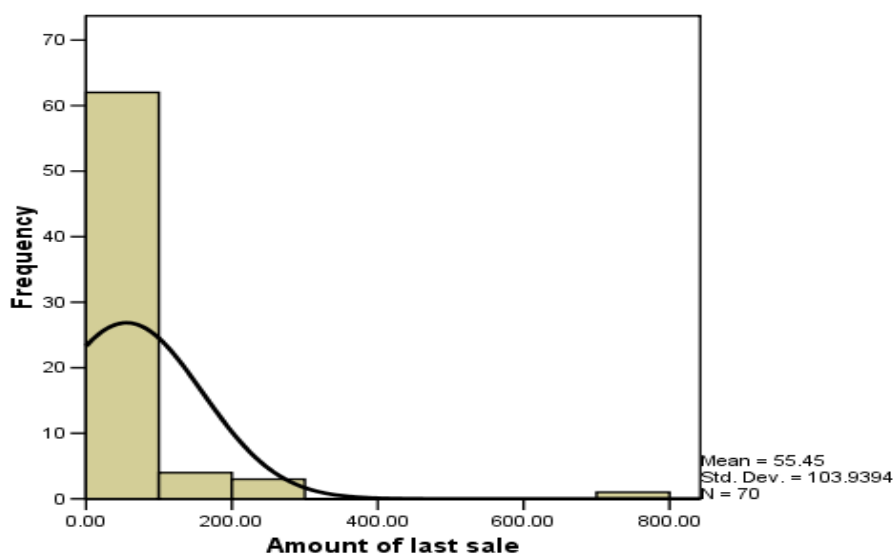
Az elkészült statisztikák az eladások jellemző értékeit mutatják. A megfigyelések száma 70, nincs hiányzó adat. Az átlag 55,45, a medián 24, a szórás 103,94. Az eloszlás ferdesége 5,33, csúcossága 34,29. A legkisebb értéke 6, a legnagyobb 776,5. Az első kvartilis értéke 12, a másodiké 24, a harmadiké 52,88. Az eladások eloszlásának közepét a medián, illetve a második kvartilis mutatja, értéke 24. Az eloszlás középpontja körül az adatok

fele 12 és 52,88 közé esik. Ezt az első és harmadik kvartilis mutatja, aminek a különbségét interkvartilisnek neveznek.

Amount of last sale		
N	Valid	70
	Missing	0
Mean		55.4500
Median		24.0000
Std. Deviation		103.93940
Skewness		5.325
Kurtosis		34.292
Minimum		6.00
Maximum		776.50
Percentiles	25	12.0000
	50	24.0000
	75	52.8750

Az eladások két extrém értéke a minimum és a maximum. Az átlag és a medián nagyon különbözik, ami azt sugalmazza, hogy az eloszlás erősen aszimmetrikus. Ezt erősíti meg a ferdeségi mutató nagy pozitív értéke (5,33), ami azt mutatja, hogy az eladások eloszlásának hosszú jobboldali farka van, azaz balra ferde az eloszlás. Kis gyakorisággal nagyon nagy eladások is előfordulnak, szinte nincs felső határa az eladás nagyságának. Az alsó határa azonban csak nulla lehet, negatív eladás nem létezik. Az eladások ezen tulajdonsága okozza a balra ferde eloszlást, vagyis a pozitív ferdeséget. A nagy pozitív ferdeség jól látható a medián és átlag elhelyezkedésén is, az átlag jobbra található a mediántól. A szórás nagyon nagy 103,94. A nagy pozitív csúcsosság (34,29), a normál eloszlásnál csúcsosabb eloszlást jelez.

Histogram



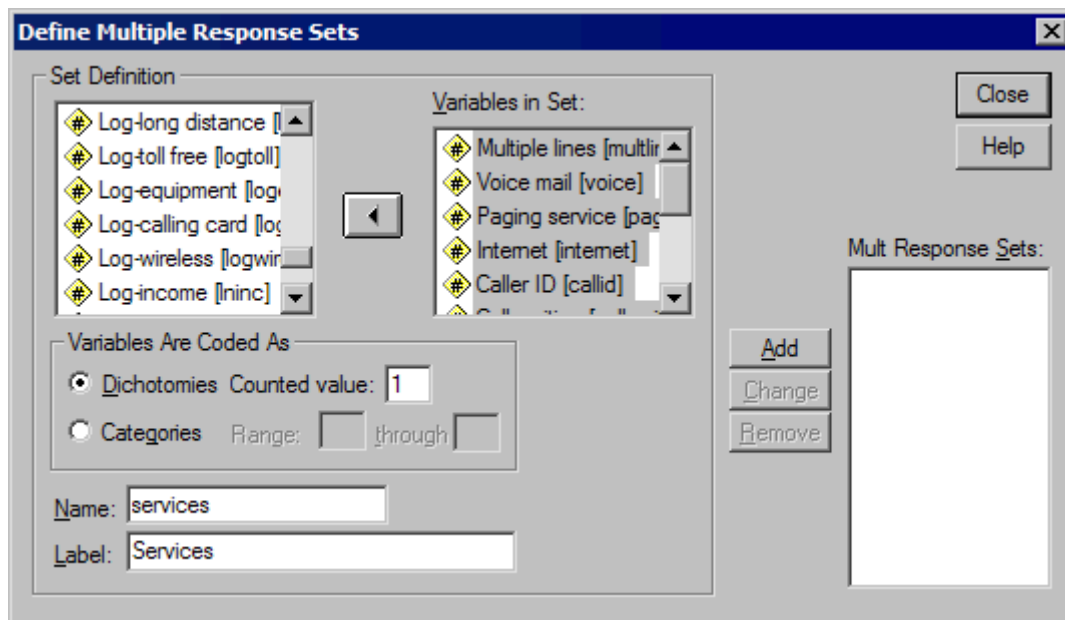
138. ábra: Az eladások hisztogramja

A hisztogram vizuálisan mutatja az eladások eloszlását. A normál eloszlás vonala segít a tényleges eloszlás tulajdonságainak megítéléséhez.

Többszörös válaszadások elemzése 1.

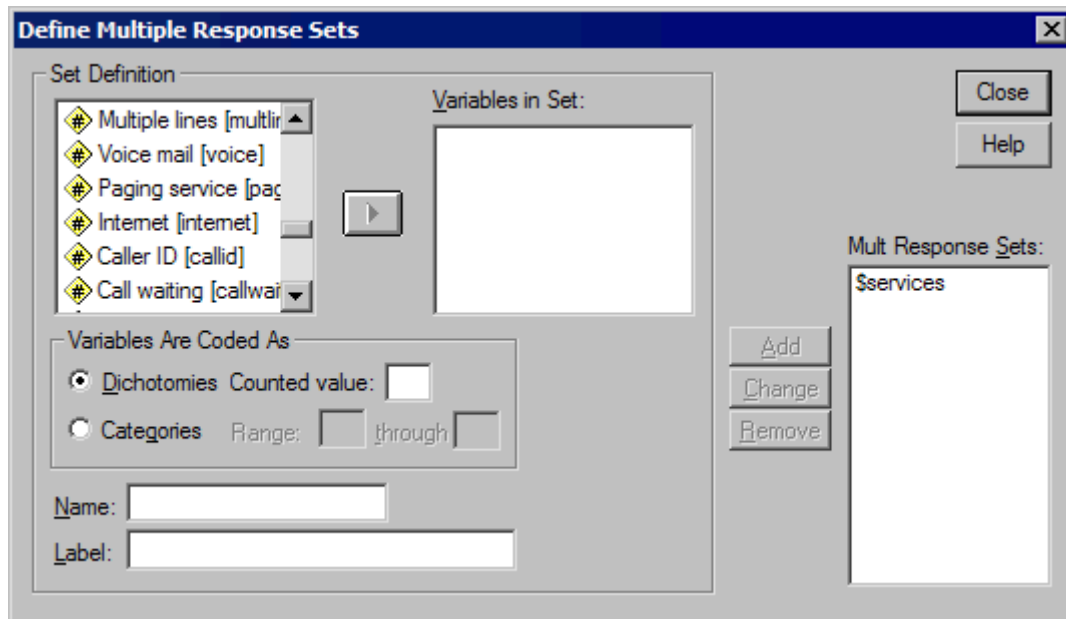
A többszörös válaszadások elemzése gyakorisági-, és kereszttáblázatok létrehozását jelenti az általunk előre definiált többszörös válaszadások csoportjai, szettjei alapján. A többszörös válaszadás szettje logikailag összetartozó változók együtteséből áll. Ezek a változók legtöbbször dichotóm, két-értékű vagy kategória változók. A többszörös dichotóm szett a gyakorlatban sokszor igen/nem (1/0 vagy true/false) típusú válaszok csoportját jelenti. Pl. milyen eszközökkel rendelkezik a válaszadó az alább felsoroltak közül? Többszörös kategória szettet akkor készítünk, amikor maximalizáljuk a válaszok számát. Ebben az esetben a megkérdezettek maximális válaszainak száma jelentősen kevesebb, mint a lehetséges válaszok száma.

A többszörös válaszadás szettjének létrehozásához válasszuk az Analyze, Multiple Response, Define Sets... parancsot. A példaadatbázist az SPSS-hez kapjuk. Ebben egy telekommunikációs kérdőív adatai szerepelnek, hogy milyen szolgáltatásokat vesznek igénybe a megkérdezettek. Az adatbázis változói közül válasszuk ki a logikailag összetartozókat, és tegyük bele Variables in Set ablakba. Összesen maximum 20 ilyen szettet tudunk megadni. Mindegyiknek egyedi névvel kell rendelkeznie.



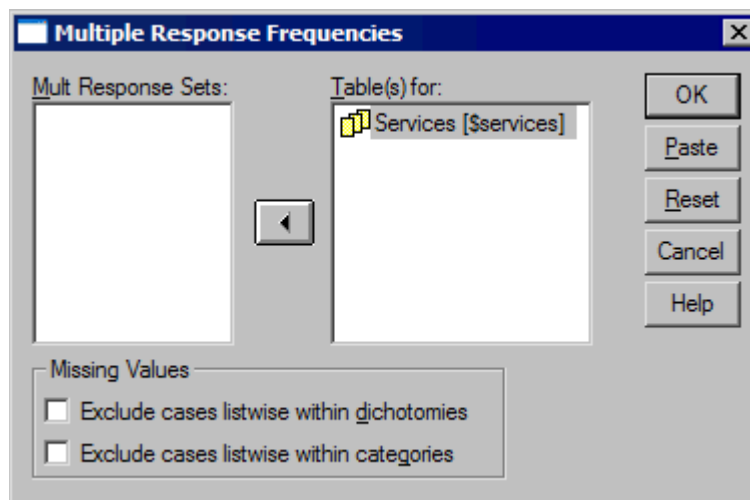
139. ábra: Többszörös válaszadások szettjének megadása

A változók most két-értékűek és az összeszámlálандó értéket az 1 jelenti. A szett neve 'servises' és a címkéje 'Services'. Klickr Add gomb és Close.



140. ábra: Többszörös válaszadások neve és címkéje

A gyakorisági táblázat elkészítéséhez válasszuk az *Analyze, Multiple Response, Frequencies...* parancsot. A *Services* változót tegyük a *Table(s) for:* ablakba.



141. ábra: Gyakoriság elemzése többszörös választású változóval

Klikk *OK*. Az adatbázisban összesen 1 000 megfigyelés szerepel 111 hiányzó adattal. A hiányzó adat ebben a példában azokat a személyeket jelenti, akik egyetlen szolgáltatásra sem fizetnek elő. Érvényes válasznak tekintjük azt, ha legalább egy szolgáltatást igényben vesz az illető.

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
\$services ^a	889	88.9%	111	11.1%	1000	100.0%

a. Dichotomy group tabulated at value 1.

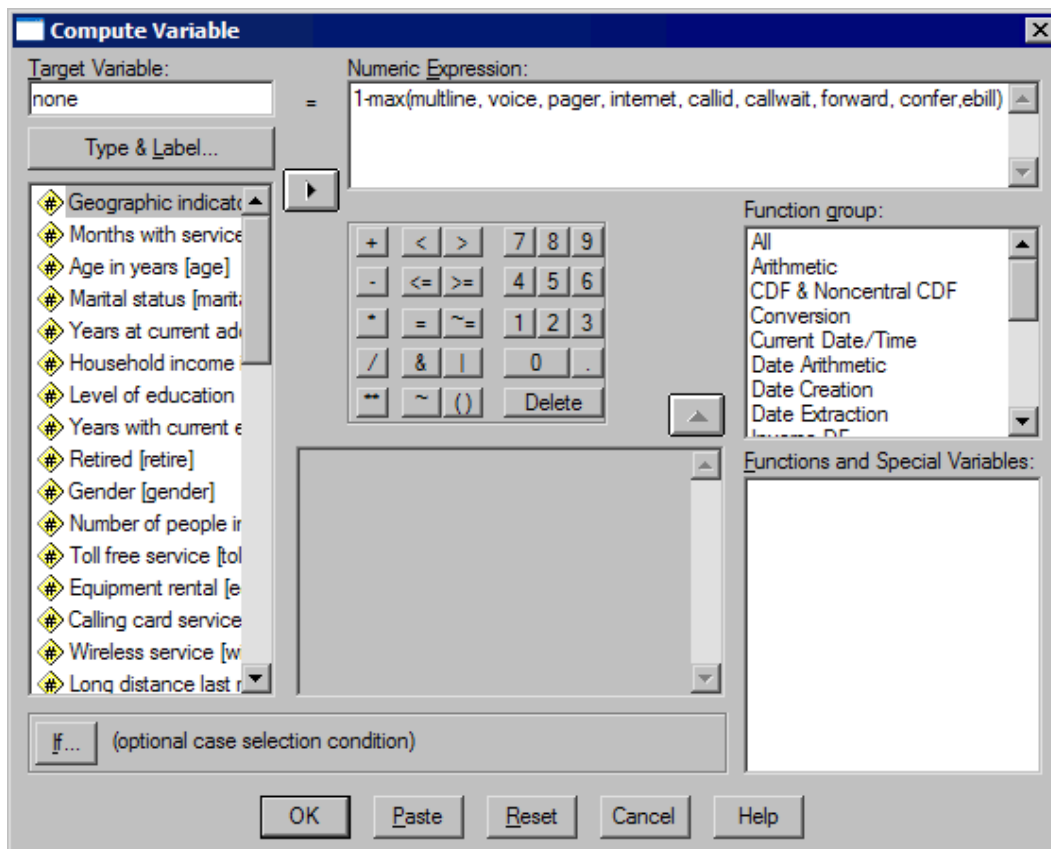
A gyakorisági táblázat a megkérdezettek válaszait tartalmazza. Jól látható, hogy milyen szolgáltatást hányan vesznek igénybe. Pl. az 1 000 megkérdezett közül 368 használja az Internetet. Természetesen az összes válaszok száma meghaladhatja az 1 000, mivel egy válaszadó több szolgáltatást is használhat egyszerre. Ez a többszörös válaszadás lényege. Az egyszerű gyakorisági értékek mellett egyéb fontos információ is kiolvasható a táblázatból. Az N jelenti, hogy hányan használják vagy fizetnek elő a szolgáltatásra, a Total N az összes előfizetői szerződés számát mutatja. A százalék (Percent) oszlop az összes válaszok százalékában adja meg az igénybe vett szolgáltatások nagyságát. Ezt többféleképpen is lehet értelmezni. Pl. a válaszadók mindennapi tevékenységük során ilyen arányban használják a különböző szolgáltatásokat vagy a különböző szolgáltatások piaci részesedése. Ilyen kimutatást egy egyszerű gyakorisági táblázattal nem lehet készíteni csak többszörös gyakorisági táblázattal.

		Responses		Percent of Cases
		N	Percent	
Services ^a	Multiple lines	475	12.7%	53.4%
	Voice mail	304	8.1%	34.2%
	Paging service	261	7.0%	29.4%
	Internet	368	9.8%	41.4%
	Caller ID	481	12.9%	54.1%
	Call waiting	485	13.0%	54.6%
	Call forwarding	493	13.2%	55.5%
	3-way calling	502	13.4%	56.5%
	Electronic billing	371	9.9%	41.7%
Total		3740	100.0%	420.7%

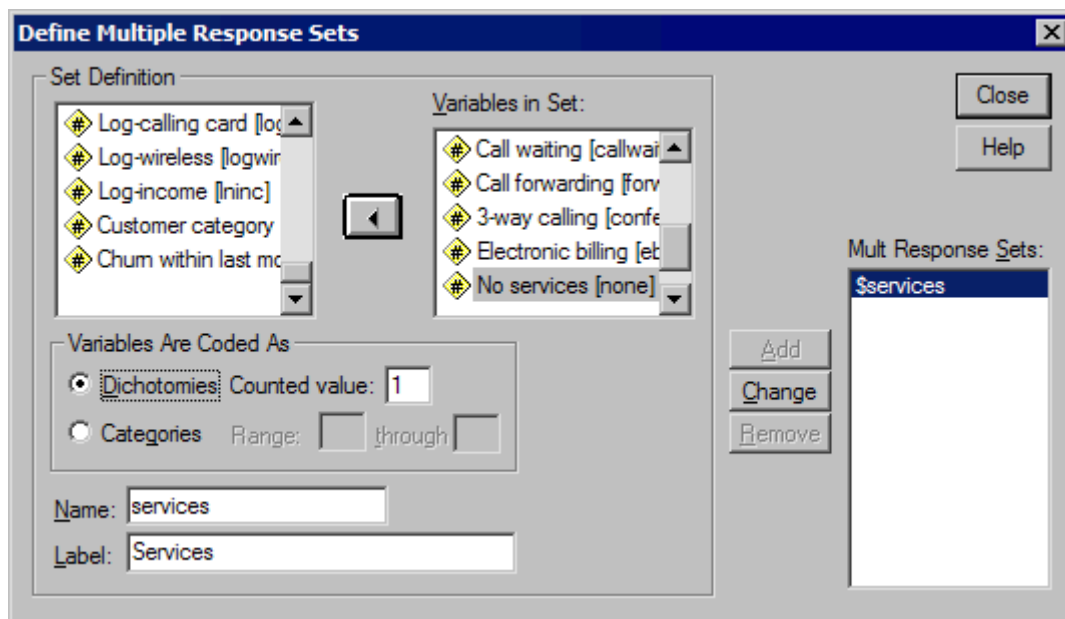
a. Dichotomy group tabulated at value 1.

A Percen of Cases oszlopban az érvényes válaszok százalékában látjuk az adott szolgáltatás igénybevételi arányát. A 889 válaszoló 41,4%-a használja az Internetet. Ezek a számok azonban nem mutatják azokat a felhasználókat, akik egyetlen szolgáltatást sem választanak (ezek a hiányzó adatok). Amennyiben ezekre is kíváncsiak vagyunk elő kell állítani egy új változót, amit a Transform, Compute paranccsal végezhetünk el (142. ábra). Legyen a változó neve „none”. A függvény: 1-max(a szolgáltatások listája vesszővel elválasztva). Ennek az értéke 1, ha egyetlen szolgáltatást sem vesz igénybe az illető. A változó címkéje legyen „No services”. Klickelj a Continue gombra.

Ezek után adjuk hozzá az új változót a már meglévő többszörös dichotóm ssethez (143. ábra).

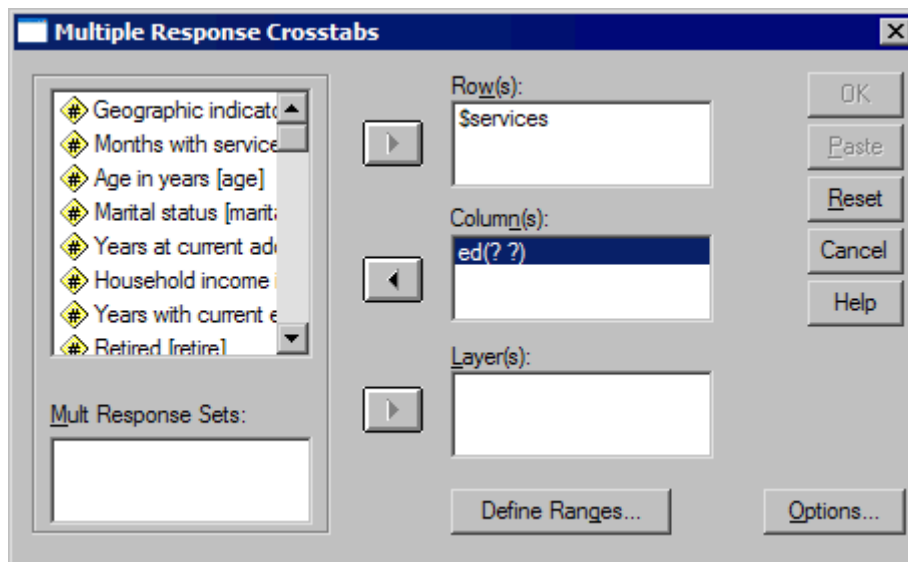


142. ábra: A szolgáltatást igénybe nem vevők számítása



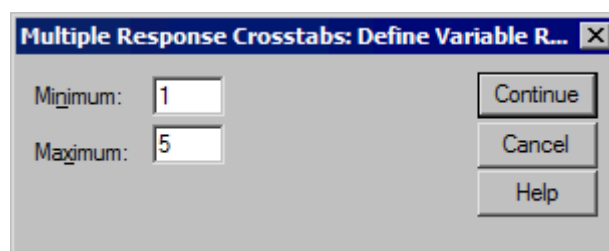
143. ábra: A többszörös választás bővítése

Van-e összefüggés az iskolai végzettség és a szolgáltatások kedveltsége között? Ehhez készítsünk keresztábrát az igénybe vett telekommunikációs szolgáltatások gyakorisága és az azt igénybe vevő ügyfelek iskolai végzettség változók felhasználásával. Analyze, Multiple Response, Crosstabs...



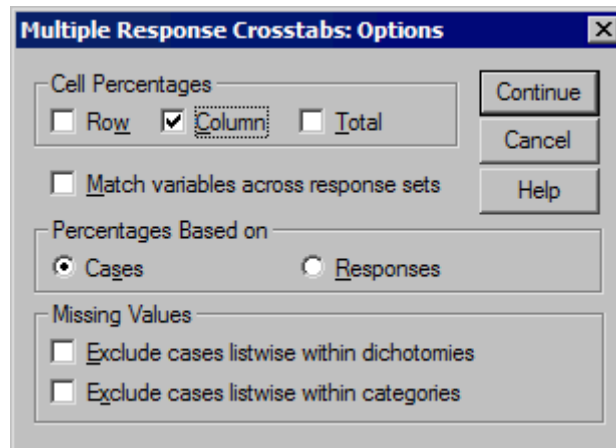
144. ábra: Keresztábrát készítése többszörös változóval

A sorba (Row) tegyük bele a 'services' változót, az oszlopba (Column) az 'ed' (level of education) változót. Az iskolai végzettség változónak meg kell adni az elemzésbe vont tartományát (Define Ranges). Jelen esetben 1-5 kategóriák vesznek részt az elemzésben.



145. ábra: Iskolai végzettség kategóriájának megadása

Ezek után állítsuk be, hogy a cellákban a százalékok oszloponként számíthatódnak, vagyis iskolai végzettség szerint (146. ábra).



146. ábra: Cellák százalékszámítása oszlop szerint

A számítások elvégzése után az alábbi eredménytáblázatot kapjuk.

			Level of education					Total
			Did not complete hig	High school degree	Some college	College degree	Post-graduate d	
Services	Multiple lines	Count	57	121	107	139	51	475
		% within ed	27.9%	42.2%	51.2%	59.4%	77.3%	
	Voice mail	Count	21	70	68	104	41	304
		% within ed	10.3%	24.4%	32.5%	44.4%	62.1%	
	Paging service	Count	14	61	53	101	32	261
		% within ed	6.9%	21.3%	25.4%	43.2%	48.5%	
	Internet	Count	14	71	86	145	52	368
		% within ed	6.9%	24.7%	41.1%	62.0%	78.8%	
	Caller ID	Count	87	142	103	121	28	481
		% within ed	42.6%	49.5%	49.3%	51.7%	42.4%	
	Call waiting	Count	92	145	101	116	31	485
		% within ed	45.1%	50.5%	48.3%	49.6%	47.0%	
	Call forwarding	Count	94	141	106	119	33	493
		% within ed	46.1%	49.1%	50.7%	50.9%	50.0%	
	3-way calling	Count	98	145	106	120	33	502
		% within ed	48.0%	50.5%	50.7%	51.3%	50.0%	
	Electronic billing	Count	14	84	88	141	44	371
		% within ed	6.9%	29.3%	42.1%	60.3%	66.7%	
	No services	Count	48	36	17	7	3	111
		% within ed	23.5%	12.5%	8.1%	3.0%	4.5%	
Total		Count	204	287	209	234	66	1000

Percentages and totals are based on respondents.

a. Dichotomy group tabulated at value 1.

A sorokban a szolgáltatások, az oszlopokban az iskolai végzettségek láthatók. A legutolsó sor (Total Count) mutatja a megkérdezettek iskolai végzettségének megoszlását (204, 287...66). A táblázat azt sugalmazza, hogy összefüggés van az iskolai végzettség és a használt szolgáltatások között. Pl. a 'Caller ID'-től '3-way calling' szolgáltatások igénybevétele közel azonos a különböző iskolai végzettségű kategóriákban. A többi szolgáltatások használatának gyakorisága azonban nő az iskolai végzettséggel. Minél magasabb

végzettséggel rendelkezik a válaszadó, annál több szolgáltatást használ. Az összefüggés pozitív. A szolgáltatást nem használó kategóriában ('No services') az iskolai végzettség növekedésével csökkennek a százalékos értékek. A két változó közötti összefüggés negatív.

Maximum k válasz elemzése 2.

Az SPSS programban egyéb lehetőség is rendelkezésre áll a többszörös választásos kérdések kiértékelésére. Nagyon hasonló az előbb leírtakhoz az Analyze, Tables, Multiple Response Sets... menüpont. Ebben is dichotóm és kategória változókból készíthetünk összeállításokat, szettekét. A kategória változók kódolásának tökéletesen meg kell egyezniük. A kategória változókból készített szett előállítását nem írtam le részletesen korábban, most pótolom. Az alábbi kérdésre várjuk a válaszokat:

Jelölje be a legszimpatikusabb nemzeteket! Maximum hármat választhat.

magyar

olasz

angol

román

.

.

.

egyéb

A lehetséges válaszok száma 41, de ebből csak maximum hármat választhat a megkérdezett. Az adatbázisban ilyenkor három kategória változót kell definiálni. Az első, második és harmadik választásra. Mindhárom változóban 41 kategória van (magyartól-egyéb nációig).

GYAKORLÓ FELADATOK

Meglévő adatbázis tulajdonságainak megtekintése, (Display Data Info)

Olvassa be a *meteorológia.txt* fájlt. Mentse el a fájl formátumát *.tpf kiterjesztéssel! Text típusú adatok beolvasása.

Egészítse ki a változókat további információkkal meglévő adatbázisból! (Apply Data Dictionary)

Rendezze növekvő, majd csökkenő sorrendbe a *termés1989.sav* fájlt!

Bővítse ki a *termés1989.sav* fájlt a *termés1995.sav* fájl eseteivel! (Merge Files, Add Cases...)

Kapcsolja össze a *termés.sav* fájlt a *csapadék.sav* fájjal! (Merge Files, Add Variables, Match cases on key variables in sorted files, External file is keyed table, Excluded Variables: év változó → Key Variables. OK) Őrizze meg az utasításokat!

Alakítsa át az *esztendő2002.sav* fájl „jdate” változóját hónapokra! (Transform, Compute, DATE.YRDAY(2002,jdate) dátum előállítás, XDATE.MONTH(datum) a hónap számainak képzése)

Agregálja az *esztendő2002.sav* fájlt változóit értelemszerűen, a hónapok alapján! (Data, Aggregate..., grad.sum; tmax.mean; tmin.mean; csapad.sum)

A *termés.sav* fájlban kódolja át az aktuális műtrágyaadagokat egy másik változóba (1→0, 60-180→90, 240-300→150). 1990 és 1991 esztendőben! (Transform, Recode, Into Different Variables)

Ossza fel a *termés.sav* fájl „termés” változóját négy egyforma számú kategóriára! (Transform, Categorize Variables)

Számítsa ki a *termés.sav* fájl „termés” változójának rangszámait és ábrázolja vonal diagrammal! (Transform, Rank Cases), (Graphs, Line Charts, Category Axis: rank of termés, Variable: mean of termés)

Próbálja ki az automatikus újrakódolást saját készítésű text típusú adatbázison!

Számítsa ki *termés.sav* fájl „termés” változójának legfontosabb statisztikai mutatóit a különböző műtrágya kezeléseknél! (Analyze, Reports, Case summaries, Variables: termés, Grouping Variables: műtrágyázás)

Számítsa ki az *esztendő2002.sav* fájl minden változójának jellemző éves értékét! (Analyze, Reports, Reports summaries in Columns...)

Állapítsa meg, hogy a *Termés1989.sav* fájlban hány parcellát figyeltünk meg a talajművelés x tőszám kombinációban! (Analyze, Descriptive Statistics, Crosstabs...)

Ábrázolja oszlopdiagrammal a *Termés1989.sav* fájlból a termés nagyságát a műtrágyázás függvényében! (Graphs, Bar..., Simple, Summaries for groups of cases)

Ábrázolja oszlopdiagrammal az *esztendő2002.sav* fájlból a globálsugárzás, minimális, maximális és átlagos értékét! (Graphs, Bar..., Simple, Summaries of separate variables)

Ábrázolja az *esztendő2002.sav* fájlból a globálsugárzás éves menetét! (Graphs, Line..., Simple, Values of individual cases, Line Represents: globálsugárzás, Category Labels Variable: az év napja)

A *termés1989.sav* fájlból számítsa ki a „termés” változó különböző statisztikai mutatóit a talajművelés függvényében! Állapítsa meg az összefüggés szorosságát és lineáris jellegét! Mennyiben határozza meg a talajművelés a kukorica termését ebben az esztendőben?

Állapítsa meg, hogy a kukorica termése 10t/ha volt-e az 1989-as esztendőben!

Van-e különbség a nem trágyázott és a 120 kg/ha nitrogénnel műtrágyázott kukorica termése között 1995-ben? (kétmintás t-próba)

FÜGGELÉK

Az *esztendő2002.sav* fájl szerkezete:

Name		Position
JDATE	Az év napja	1
	Measurement Level: Scale	
	Column Width: Unknown Alignment: Right	
	Print Format: F3	
	Write Format: F3	
GRAD	Globálsugárzás (MJ/m2 nap)	2
	Measurement Level: Scale	
	Column Width: Unknown Alignment: Right	
	Print Format: F4.1	
	Write Format: F4.1	
TMAX	Maximális hőmérséklet (C)	3
	Measurement Level: Scale	
	Column Width: Unknown Alignment: Right	
	Print Format: F4.1	
	Write Format: F4.1	
TMIN	Minimális hőmérséklet (C)	4
	Measurement Level: Scale	
	Column Width: Unknown Alignment: Right	
	Print Format: F4.1	
	Write Format: F4.1	
CSAPADÉK	Napi csapadék (mm)	5
	Measurement Level: Scale	
	Column Width: Unknown Alignment: Right	

Print Format: F4.1

Write Format: F4.1

A *csapadék.sav* fájl szerkezete:

Name	Position
<p>ÉV Esztendő</p> <p>Measurement Level: Scale</p> <p>Column Width: Unknown Alignment: Right</p> <p>Print Format: F4</p> <p>Write Format: F4</p>	1
<p>CSAPADÉK Éves csapadék (mm)</p> <p>Measurement Level: Scale</p> <p>Column Width: Unknown Alignment: Right</p> <p>Print Format: F8.2</p> <p>Write Format: F8.2</p>	2

A *termés.sav* fájl szerkezete:

Name	Position
<p>ÉV</p> <p>Measurement Level: Scale</p> <p>Column Width: 4 Alignment: Right</p> <p>Print Format: F4</p> <p>Write Format: F4</p>	1
<p>NPK műtrágyázás</p> <p>Measurement Level: Nominal</p> <p>Column Width: 7 Alignment: Right</p>	2

Print Format: F3

Write Format: F3

Value Label

1	nem trágyázott
30	N30, P23, K27
60	N60, P45, K53
90	N90, P68, K80
120	N120, P90, K106
150	N150, P113, K133
180	N180, P135, K159
240	N240, P180, K212
300	N300, P225, K265

TERMÉS Termés (t/ha) 3

Measurement Level: Scale

Column Width: Unknown Alignment: Right

Print Format: F8.2

Write Format: F8.2

A *termés1989.sav* fájl szerkezete:

Name Position

EV évek 1

Measurement Level: Scale

Column Width: 5 Alignment: Right

Print Format: F4

Write Format: F4

TALAJMUV Talajművelés 2

Measurement Level: Nominal

Column Width: 5 Alignment: Right

Print Format: F1

Write Format: F1

Value Label

- 1 őszi szántás
- 2 tavaszi szántás
- 3 tárcsás

TOSZAM Tőszám 3

Measurement Level: Scale

Column Width: 4 Alignment: Right

Print Format: F2

Write Format: F2

HIBRID 4

Measurement Level: Nominal

Column Width: 6 Alignment: Right

Print Format: F2

Write Format: F2

Value Label

- 1 De 351
- 2 De 377
- 3 De 382
- 4 Dk 366
- 5 Dk 373
- 6 Dk 391
- 7 Dk 471
- 8 Dk 477
- 9 DK 493

- 10 Dk 524
- 11 Dk 527
- 12 Kanada
- 13 Katinka
- 14 LG 2298
- 15 Lipesa
- 16 Maya
- 17 Mv 444
- 18 Mv 484
- 19 MV 487
- 20 Occitán
- 21 Pannónia
- 22 Pelikán
- 23 Sprinter
- 24 Stira
- 25 Szegedi 348
- 26 Veronika (Sze 427)
- 27 Volga SC
- 28 Szegedi 463
- 29 Dk 440
- 30 Ella
- 31 Hunor

TRAGYA Trágya kezelés

5

Measurement Level: Ordinal

Column Width: 9 Alignment: Right

Print Format: F1

Write Format: F1

Value Label

1 nem trágyázott

2 nitrogén 120

3 nitrogén 240

ISMÉTLÉS 6

Measurement Level: Nominal
 Column Width: 5 Alignment: Right
 Print Format: F1
 Write Format: F1

TERMÉS termés t/ha 7

Measurement Level: Scale
 Column Width: 6 Alignment: Right
 Print Format: F6.3
 Write Format: F6.3
 Missing Values: -999.0

A termés1995.sav fájl szerkezete:

Name	Position
EV évek	1
Measurement Level: Scale Column Width: 5 Alignment: Right Print Format: F4 Write Format: F4	
ONTOZES Öntözés	2
Measurement Level: Ordinal Column Width: 9 Alignment: Right Print Format: F1 Write Format: F1	
Value Label	

1	Nem öntözött	
2	Öntözött	
ELOVET	Elővetemény	3
	Measurement Level: Nominal	
	Column Width: 7 Alignment: Right	
	Print Format: F1	
	Write Format: F1	
	Value Label	
	1 Kukorica	
	2 Búza	
TALAJMUV	Talajművelés	4
	Measurement Level: Nominal	
	Column Width: 5 Alignment: Right	
	Print Format: F1	
	Write Format: F1	
	Value Label	
	1 őszi szántás	
	2 tavaszi szántás	
	3 tárcsás	
TOSZAM	Tőszám	5
	Measurement Level: Scale	
	Column Width: 4 Alignment: Right	
	Print Format: F2	
	Write Format: F2	
HIBRID		6

Measurement Level: Nominal

Column Width: 6 Alignment: Right

Print Format: F2

Write Format: F2

Value Label

- 1 De 351
- 2 De 377
- 3 De 382
- 4 Dk 366
- 5 Dk 373
- 6 Dk 391
- 7 Dk 471
- 8 Dk 477
- 9 DK 493
- 10 Dk 524
- 11 Dk 527
- 12 Kanada
- 13 Katinka
- 14 LG 2298
- 15 Lipesa
- 16 Maya
- 17 Mv 444
- 18 Mv 484
- 19 MV 487
- 20 Occitán
- 21 Pannónia
- 22 Pelikán
- 23 Sprinter
- 24 Stira
- 25 Szegedi 348
- 26 Veronika (Sze 427)

- 27 Volga SC
- 28 Szegedi 463
- 29 Dk 440
- 30 Ella
- 31 Hunor

TRAGYA Trágya kezelés 7

Measurement Level: Ordinal

Column Width: 9 Alignment: Right

Print Format: F1

Write Format: F1

Value Label

- 1 nem trágyázott
- 2 nitrogén 120
- 3 nitrogén 240

ISMÉTLÉS 8

Measurement Level: Nominal

Column Width: 5 Alignment: Right

Print Format: F1

Write Format: F1

TERMÉS termés t/ha 9

Measurement Level: Scale

Column Width: 6 Alignment: Right

Print Format: F6.3

Write Format: F6.3

Missing Values: -999.0

AJÁNLOTT IRODALOM

SPSS:

- Falus István – Ollé János: Statisztikai módszerek pedagógusok számára, OKKER Kiadó, 2000. (Excel és SPSS alkalmazásokkal)
- Huzsvai L. (2004): Biometriai módszerek az SPSS-ben. <http://www.agr.unideb.hu/~huzsvai>
- Katona Tamás - Lengyel Imre (szerk.): Statisztikai ismerettár - fogalmak, képletek, módszerek Excel és SPSS alkalmazásokkal. JATEPress, Szeged, 1999. 121 oldal, (közgazdász, jogász, kísérletes és társadalomtudomány)
- Ketskemény L. – Izsó L.: Bevezetés az SPSS programrendszerbe. Módszertani útmutató és feladatgyűjtemény statisztikai elemzésekhez. ELTE Eötvös Kiadó, Budapest, 2005.
- Ketskemény L. – Izsó L.: Az SPSS for Windows programrendszer alapjai, Felhasználói útmutató és oktatási segédlet. Budapest, 1996.
- Moksony Ferenc: Gondolatok és adatok: Társadalomtudományi elméletek empirikus ellenőrzése. Budapest, Osiris Kiadó, 1999.
- Székelyi Mária - Barna Ildikó: Túlélőkészlet az SPSS-hez. TYPOTEX, 2002, ISBN 963 9326 429

Statisztika:

- Anscombe, F.J. (1973). Graphs in statistical analysis. *American Statistician*, 27, 17-21.
- Baráth Cs.-né. - Ittész A. - Ugrósd Gy.: 1996. Biometria: módszertan és a MINITAB programcsomag alkalmazása. Mezőgazda Kiadó, Budapest
- G.U. Yule – M.G. Kendall: Bevezetés a statisztika elméletébe. Közgazdasági és Jogi könyvkiadó, Budapest. 1964.
- Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting*, 4, 1-28.
- Harnos ZS. szerk.: 1993. Biometriai módszerek és alkalmazásaik MINITAB programcsomaggal. AKAPRINT, Budapest
- Lothar Sachs.: Statisztikai módszerek. Mezőgazdasági Kiadó, Budapest, 1985.
- Makridakis, S. G., S. C. Wheelwright, and R. J. Hyndman. 1997. *Forecasting: Methods and Applications*. New York: John Wiley & Sons.
- Mérő, L.: 1986. A többdimenziós skálázás alapelvei. *Pszichológia*, (6), 3, 399-433.
- Móri F.T. – Székely J.G.: Többváltozós statisztikai analízis. Műszaki Könyvkiadó, Budapest, 1986. (ISBN 963 10 6684 3)
- Sváb J.: Biometriai módszerek a kutatásban. Mezőgazdasági Kiadó, Budapest, 1973. (második, átdolgozott, bővített kiadás)

Sváb J.: Többváltozós módszerek a biometriában. Mezőgazdasági Kiadó, Budapest, 1979. (ISBN 963 230 011 4)

Szűcs I. (szerk.)(2002): Alkalmazott statisztika. Tankönyv, Agroinform K. Budapest

GAUSS, CARL FRIEDRICH

(1777. 04. 30. - 1855. 02. 23.)



Német matematikus, csillagász és fizikus. Őt tartják minden idők egyik legnagyobb matematikusának. Így is nevezik: "A matematikusok fejedelme." Euler mellett ő a matematika legsokoldalúbb tudósa.

Braunschweigben született, édesapja nyergesmester volt. Már 6 éves korában kitűnt matematikai tehetségével. Tanítója egyszer azt a feladatot adta a kis tanulóknak, hogy adják össze a számokat 1-től 40-ig, mivel a tanító úr addig egy másik évfolyammal akart foglalkozni, és így akarta addig a kicsiket lefoglalni. De a kis Gauss hamarosan jelentkezett a jó eredménnyel. Csodálkozó tanítójának el is magyarázta, hogyan csinálta.

Párba állította a számokat $40 + 1 = 39 + 2 = 38 + 3$ stb. Ezek a párok mindig 41-t adnak összegül, és mivel 20 ilyen pár van, az eredmény 820. Ez a gondolkodás megegyezik a számtani sorozat összegének meghatározásánál alkalmazottal. Tanítója felismerve a kisfiú rendkívüli képességeit, jelentette az esetet előjáróinak. Így jutott el a híre braunschweig-i herceghez, aki felkarolta a kis Gauss-t. Gimnáziumba került, majd a göttingeni egyetemre.

Pályája töretlenül ívelt felfelé. Ismerte és barátjának nevezte Bolyai Farkast, ennek ellenére fiát Bolyai Jánost nem támogatta, és ezzel igen nagy csalódást okozott mindkettőjüknek. Sajnos Gauss mások elismerésével is fukarkodott. Így például Abel tehetséges norvég matematikussal kapcsolatban is. Lobacsevszkij orosz matematikust ugyan beajánlotta a Göttingeni Tudományos Társaságba, de a nem euklideszi geometria megalkotásának területén végzett munkásságának közvetlen elismerésétől tartózkodott, akárcsak Bolyai János esetében.

Gauss csillagászként is számottevőt alkotott. 1801-ben egy új és egyszerűbb módszert dolgozott ki a bolygó pályájának kiszámítására. 1820 körül geodéziával (földmérés) kezdett foglalkozni. Fizikai munkássága is említésre méltó. Göttingenben egy szobor ábrázolja őt és Wilhelm Webert a távíró 1833-ban történő feltalálása közben. Ő alkotja meg az első abszolút fizikai mértékegységrendszert. Még számológép fejlesztéssel is foglalkozott, Leibniz gépét tökéletesítette. Ez a gép az ő idejében népszerű volt egész Németországban.

Gauss békés, hosszú és elismert életet élt. Igazi zsenialitást még így is nehéz teljes egészében felmérni, mert nagyon sok felfedezését, elgondolását, így a nem euklideszi geometria felfedezése irányába tett gondolatait sem publikálta.

Utolsó kívánsága az volt, hogy egyik korai és számára legkedvesebb felfedezésének, a 17 oldalú szabályos sokszög szerkesztésének emlékére sírkövére egy szabályos 17-szöget véssenek. Ezt ugyan nem teljesítették, de szülővárosában a tiszteletére emelt szobor talapzatán látható a szabályos 17 oldalú sokszög.

Matematikai munkásságáról:

Egyik legkedvesebb matematikai szakterülete a számelmélet volt. Tőle származik az a mondás, hogy: "A matematika a tudományok királynője, és a matematika királynője a számelmélet." 1791-ben, 14 éves korában becslést adott a prímszámok eloszlására, miszerint ezres számkörben a prímszámok száma fordítottan arányos a számok természetes alapú logaritmusával. Ezt ugyan később többen is pontosították, de ez semmit nem von le a fiatal Gauss érdemeiből.

Ő volt az, aki felfedezte, hogy kapcsolat van a prímszámok és a szabályos sokszögek szerkeszthetősége között. Egy "n" oldal számú szabályos sokszög csak akkor szerkeszthető euklideszi szerkesztéssel ha "n" prímtényező felbontásában csak a 2 szerepel tetszőlegesen nem negatív egész kitevőjű hatványon és az ún. Fermat-féle prímelek (3,5,17,65537) első kitevőjű hatványon.

Azaz $n = 2^k \cdot p_1 \cdot p_2 \cdot \dots \cdot p_k$, ahol p_1, p_2, p_k különböző Fermat-féle prímelek. Tehát szerkeszthető a 3, 4, 5, 6, 8, 10, 12, 14, 15, 17, ..., 257 és 65537 oldalú szabályos sokszög, de nem szerkeszthető például a 7, 9, ill. 11 oldalú. A 17 oldalú szabályos sokszög szerkesztésének a módját ő meg is oldotta.

Gauss foglalkozott a szakaszos tizedes törtekkel, és tisztázta, mikor kapunk tiszta vagy vegyes szakaszos tizedes törtet, és mekkora lehet a szakasz hosszúsága. 1799-ben a doktori értekezésében az "algebra alaptételét" igazolta, amely szerint minden algebrai egyenletnek van gyöke. Ezek a gyökök nem okvetlenül valósak, hanem lehetnek komplex számok is, és nem biztos, hogy ezek a gyökök mind különböznek egymástól. A gyökök száma (beleértve az azonosakat is) az egyenlet fokszámával egyenlő. 1827-ben jelent meg „A görbe felületekre vonatkozó általános vizsgálatok” című műve, amelynek eredményei geodéziai munkásságára vezethetők vissza. Gauss ötlete, hogy a komplex számokat a sík pontjaiként ábrázolhatjuk. 1837-ben megjelent értekezése a komplex számok algebráját és aritmetikáját tartalmazza. A nem euklideszi geometria megalkotásának területén végzett kutatásairól csak levelezéseiből tudunk, és feltételezhető, hogy ezen a területen is messzire jutott.