Lázár Ede

KÖZGAZDASÁGI KUTATÁSMÓDSZERTAN

RISOPRINT
Cluj-Napoca • 2022

Lázár Ede

KÖZGAZDASÁGI KUTATÁSMÓDSZERTAN

A könyv a Pallas Athéné Domus Meriti Alapítvány támogatásával valósult meg.

SAPIENTIA – ERDELYI MAGYAR TUDOMANYEGYETEM

Lázár Ede

KÖZGAZDASÁGI KUTATÁSMÓDSZERTAN

egyetemi jegyzet

RISOPRINT KIADÓ

KOLOZSVÁR, 2022

© Editura RISOPRINT

Toate drepturile rezervate autorilor & Editurii Risoprint

*Editura **RISOPRINT** este recunoscută de **C.N.C.S.***
***(Consiliul Național al Cercetării Științifice)**.*
www.risoprint.ro ***www.cncs-uefiscdi.ro***

Opiniile exprimate în această carte aparțin autorilor și nu reprezintă punctul de vedere al Editurii Risoprint. Autorii își asumă întreaga responsabilitate pentru forma și conținutul cărții și se obligă să respecte toate legile privind drepturile de autor.

Toate drepturile rezervate. Tipărit în România. Nicio parte din această lucrare nu poate fi reprodusă sub nicio formă, prin niciun mijloc mecanic sau electronic, sau stocată într-o bază de date fără acordul prealabil, în scris, al autorilor.
All rights reserved. Printed in Romania. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a data base or retrieval system, without the prior written permission of the author.

Közgazdasági kutatásmódszertan : egyetemi jegyzet /
Lázár Ede. - Cluj-Napoca : Risoprint, 2022

ISBN 978-973-53-2923-5

Director editură: GHEORGHE POP

Olvasószerkesztő: **Benedek Enikő**
Lektorálta: **dr. Tóth József**
Tördelés: **Garda-Mátyás Zsolt**

TARTALOMJEGYZEK

ELŐSZÓ

„Akár egy halom hasított fa,
hever egymáson a világ,
szorítja, nyomja, összefogja
egyik dolog a másikát
s így mindenik determinált."

József Attila: *Eszmélet*

Reményeim szerint e könyv segíteni fogja az Olvasót a minket körülvevő gazdasági, társadalmi világ megértésében. Megtanulhatjuk a módját, módszertanát, hogyan szedjük szét és magyarázzuk meg az egymáson heverő gazdasági, üzleti világokat, és ez megadja a lehetőséget arra is, hogy beékeljük közéjük a számításaink szerinti kis világunkat. Ezeket a világokat ugyanis emberek alkotják a maguk motivációival, akaratával, hitével, álmaival vagy – ami mostanában hangsúlyosabb – a pánikra való hajlamával. Minden súlyos determinizmusuk mellett, ha képesek vagyunk megfigyelni, mérni és elemezni a gazdasági jelenségeket, akkor akaratunkat is érvényesíthetjük közöttük.

A lírainak szánt indítás ellenére a könyv célja nagyon gyakorlatias; bevezetni az olvasót a közgazdasági kutatás módszertanába, és megismertetni azokkal az elvekkel és gyakorlati technikákkal, amelyek alapján képes lesz felismerni, megmagyarázni és előrejelezni a gazdasági jelenségeket.

A könyv a Sapientia Erdélyi Magyar Tudományegyetem gazdasági szakos hallgatói számára készült jegyzet, a *Kutatásmódszertan* tárgy keretén belül kerül oktatásra. De az egyetemi hallgatókon, oktatókon túl az olvasói célcsoportba tartoznak azok a vállalati menedzserek, közgazdászok is, akik több, üzleti értéket jelentő információt szeretnének kinyerni a gyorsan növekvő vállalati adathalmazból vagy a piaci adatokból. Régebben a vállalati belső vagy külső adatok elemzése, a közgazdasági kutatás csak néhány szerencsés elemző drága és időigényes kiváltsága volt. Az információtechnológia fejlődésének, és ezzel párhuzamosan az adatelemzési módszertanok fejlődésének köszönhetően ma, a *Big Data* korszakában nemcsak lehetőség, hanem kihagyhatatlan a cégek számára, hogy a belső-külső adatokat megragadja, elemezze és a döntési folyamataiba integrálja.

Több mint száz piackutatási és más közgazdasági kutatási projekt tagjaként és vezetőjeként szerzett tapasztalat, illetve a közel húszéves oktatási tapasztalat alapján a gyakorlati alkalmazásra tettem a hangsúlyt. A könyv vezérfonala a közgazdasági kutatási folyamat, az olvasó – ha elsajátítja a leírtakat – képes lesz végigvinni egy kutatást a probléma felismerésétől a tanulmányírás befejezéséig.

A bemutatott módszerek, modellek elméleti, statisztikai alapjairól csak annyiban esik szó, amennyiben azt elengedhetetlenül szükségesnek tartottam az eredmények helyes értelmezéséhez. Magyarán szólva a „mire jó?" kérdésre keressük a választ, azután következik a „hogyan is működik?". Ha már kezdünk belejönni az „egymáson heverő világok" feltárásába, akkor ajánlott, hogy utánajárjunk a részletekben rejlő finomságoknak, utánaolvassunk a bemutatott módszerek irodalmának is.

Félreértjük az idézett költői hasonlat célját, ha azt hisszük, hogy a módszerek rutinos begyakorlásával „szellemi favágó munkába" foghatunk. A kutatási adatok elemzése, a folyamat egésze alapos ismeretek mellett kreativitást is igényel. Az elméleti részben sablonok és konzervek helyett támpontokat kívánok nyújtani a gazdasági jelenségekről való strukturált gondolkodáshoz, a gyakorlati részben pedig az egyik legelterjedtebb adatelemző szoftver, az SPSS használatával ismerkedünk.

Köszönettel tartozom dr. Tóth Józsefnek és dr. Farkas Csabának, akik hasznos észrevételeikkel, tanácsaikkal segítettek a könyv elkészítésében.

A költői hasonlattal élve lássunk neki a farakásnak, nem feledkezve meg arról, hogy egyetlen ismeret megszerzéséhez sem vezet királyi út.

Csíkszereda, 2022. augusztus Lázár Ede

1. A KUTATÁSI TÉMA MEGHATÁROZÁSA

Az elméleti részt két fejezetre tagoltam: az első a kutatási téma meghatározása, a „mit kutassunk?" kérdésre keresi a választ, míg a második fejezet a „hogyan kutassunk?" kérdés módszertani válaszait részletezi.

1.1 A tudományos megismerés

A tudományos kutatásmódszertannal való ismerkedésünket rövid tudományelméleti alapozással, az alapfogalmak meghatározásával kezdjük.

A **tudomány** tágabb értelemben a bennünket körülvevő világ megismerésének minden formáját jelenti, szűkebb (mai) értelemben csak **a tudományos módszertanon alapuló megismerési folyamatot tekintjük tudományosnak**. A tudományos megismerési folyamat bárki által megismételhető (reprodukálható), és azonos eredményre vezet. A szűkebb tudománydefiníció nem tartja tudományosnak a filozófiát, a teológiát és a művészeteket, de a szakmai tapasztalatok, technikai ismeretek megszerzésének folyamatát sem.

A tudományos megismerés folyamatának további bemutatásához szükséges definiálnunk néhány **alapfogalmat**:

- **Valóság**. Kétféle valóságot különböztetünk meg:
  – tapasztalati valóság – azok a dolgok, amelyeket saját közvetlen tapasztalatunkból ismerünk;
  – konszenzuális valóság – közvetlen tapasztalatunktól függetlenül azért fogadjuk el valóságnak, mert konszenzusos egyetértés van abban, hogy az.
- **Megfigyelés**: információk gyűjtése a valóságról. A tapasztalati valóságra vonatkozó megfigyelésünket **empirikus** (tapasztalati) megfigyelésnek nevezzük.
- **Mérés**: a megfigyelt jelenségek tulajdonságaihoz adott szabály alapján számot, adatot rendelünk.
- **Állítások**: a valóságról a megfigyelés során alkotott megállapítások, kijelentések.
- **Tény**: tudományos bizonyítási módszerekkel igazolt állítás.
- **Törvény**: a tények egy csoportjára vonatkozó, azok összefüggéseire is rámutató egyetemesen érvényes állítás, az elméletek alkotóeleme.

- **Elmélet**: a megfigyelt tények és törvények szisztematikus magyarázata. A makroelméletek a jelenség egészét átfogóan és empirikusan nehezen tesztelhető módon magyarázzák, míg a középszintű/metaelméletek a jelenség részleteit, okait, következményeit is bemutatják, empirikusan igazolt módon.
- **Paradigma**: a tudományos társadalom által legnagyobb konszenzussal elfogadott elmélet. Pl. a mikroökonómiában, marketingben a racionálisan döntő fogyasztó, a biológiában az evolúció elmélete stb.
- **Fogalom**: az elmélet legkisebb alkotóelemei.
- **Változó**: a fogalmak empirikusan mérhető, számszerűsíthető ismérvei.

Egy tudományosan elfogadott állításnak logikailag és empirikusan (tapasztalatilag) is alátámasztottnak kell lennie, elmondhatjuk, hogy a tudományos megismerés két pillére a **logika** és a **tapasztalati megfigyelés**.
A tudományos megismerésnek ezeken kívül még van néhány fontos jellemzője, amelyek azonban nem minden tudományterületen érvényesek:

- **Reprodukálhatóság**. Egy tudományos állítást akkor tekintünk ténynek, ha a tudományos bizonyítási, azaz kutatási folyamat bármennyiszer megismételhető, és az eredmények lényegében nem különböznek. Ez a feltétel érvényesül a legtöbb természettudományi területen (kivétel pl. a csillagászat), de általában nem érvényesül a társadalom- és humán tudományokban (pl. történelemtudomány, szociológia, közgazdaságtudomány). Reprodukálható kutatási folyamatok (kísérletek) a társadalomtudományokban gyakran előfordulnak a pszichológiában, de a közgazdaságtanon belül is fontos kutatási kérdések megválaszolásának egyedüli módja a piackutatási, gazdaságpszichológiai stb. kísérletek.
- **Kvantitatív leírás**. Általános jellemző, a tudományos megállapítások, összefüggések számszerűsítése. A jelenségek kvantitatív leírása és a statisztikai, matematikai módszereket használó hipotézist megfogalmazó, tesztelő és következtetést levonó folyamat jelenti a modern tudományos megismerést.
  Azonban vannak tudományterületek, illetve olyan jelenségek, amelyek kutatásánál nem, vagy csak részben használhatjuk a kvantitatív módszereket, itt előtérbe kerülnek a kvalitatív módszerek. Ilyenek a humán tudományok (pl. történelem), de mint látni fogjuk, a közgazdasági, üzleti kutatásokban is jól meghatározott helye és funkciója van a kvalitatív kutatásoknak.

- **Mérés, mérőeszközök és mértékegységek használata**. A kvantitatív leíráshoz szorosan kapcsolódik a megfigyelés során mérési skálák, mérőeszközök és szükségszerűen a mértékegységek használata.
- **Értéksemlegesség**. A tudományos megismerési, kutatási folyamat végeredményének tekinthető elméletek etikai értelemben értéksemlegesek, nincs jó vagy rossz tudomány.
- **Determinizmus, oksági kapcsolatok vizsgálata.** A közgazdaság-tudományi kutatásokban nagy jelentősége van az oksági kapcsolatok vizsgálatának, mivel a gazdasági kapcsolatok alapjában személyközi kapcsolatok, és az empirikusan gyakran megfigyelt determinizmus szemben áll az emberi természet szabad akarat iránti vágyával. Kérdés, hogy cselekvésünk (pl. közgazdasági preferenciarendezésünk, fogyasztói magatartásunk stb.) saját személyes akaratunk eredménye, vagy olyan erők és törvények kormányozzák, amit nem tudunk befolyásolni. E két szélsőség között a **sztochasztikus** – a hatások valószínűségét elemző – kapcsolatokat vizsgáljuk leginkább a közgazdasági kutatások során. Az oksági kapcsolat kritériumai:

  - az ok időben megelőzi az okozatot;
  - a két tényező között tapasztalati együttjárásnak kell lennie;
  - a két tényező közötti összefüggést ne lehessen valamely harmadik tényező hatásával megmagyarázni, amely mindkettőnek közös oka.

A tudományos megismerés egyik pillérét jelentő logikai következtetés két alapvető formáját még a görög filozófus, Arisztotelész fogalmazta meg:

- **Indukció** – egyedi, tapasztalati megfigyelésből kiindulva vonunk le következtetést és olyan általánosan érvényes elméletet fogalmazunk meg, ami magyarázza az egyedi megfigyelést.
- **Dedukció** – valamely általános törvényből indulunk ki és alkalmazzuk egy konkrét, egyedi esetre, ezáltal magyarázva azt.

A modern tudomány

Említettük már a **tudomány** fogalmát, a tudományos megismerés jellemzői után nézzük meg, hogy milyen főbb szakaszokból áll a megismerési folyamat. A modern tudomány a kutatás során a *pozitivistának* is nevezett módszertant követi. A modern, tudományos kutatásmódszertan folyamata:

- **Hipotézisek megfogalmazása**. A hipotézis egy olyan állítás, amit még bizonyítási eljárással nem fogadtunk vagy nem utasítottunk el. Az egzakt, matematikai módon megfogalmazott hipotéziseket modellnek nevezzük.
- **Tapasztalati megfigyelés**. A modern tudományszemlélet szerint a világ objektív módon létező és megismerhető, azaz minden visszavezethető a tapasztalati valóságra. A megfigyelés és kísérletezés a tudományos kutatás alapja, empirikus (tapasztalati) vizsgálatok. Lényeges, hogy a megfigyelést úgy tervezzük meg, hogy az eredmények alkalmasak legyenek a hipotézisek elfogadására vagy elutasítására.
- **Ellenőrzés**. A megfigyelés és/vagy a kísérlet adatainak elemzése, az eredmények és a hipotézisek összehasonlítása. Az ellenőrzés során általában előtérbe kerülnek a kvantitatív módszerek (matematika, statisztika).
- **Elméletalkotás**. Egy új elmélet megfogalmazását vagy egy meglévő elfogadását az ellenőrzés kiértékelése mellett más elméletekkel való kapcsolata, konzisztenciája is meghatározza.

A posztmodern tudomány

Az 1970-es évektől kezdődően egyre szélesebb körben kezdtek posztmodernnek nevezni egy, a modern tudományszemlélettől sok szempontból élesen különböző tudományfilozófiai irányzatot, a **posztmodernt**.
A posztmodern szerint az élet „szövegekből" áll, amelyeket állandóan és nagyon különböző módokon „olvasunk" (Drótos, 2000). A posztmodern tudományos módszertan jellemzői:

- A posztmodern a kutatási folyamat középpontjába **a jelenségek párhuzamosan érvényes olvasatainak megalkotását helyezi**.
- A posztmodernre az egyértelmű megtagadás helyett általában jellemzőbb **a különböző irányzatok tudatos keverése**, a klasszikusok újraértelmezése.

Valamennyi tudományterületre hatással volt és gyakran nemcsak egy újabb elméletet alkotott a meglévő paradigma mellé, hanem újszerű megközelítése akár új diszciplínákat is eredményezett. A közgazdaságtudományon belül a szervezés és vezetéstudományok nagyon sok diszciplínájában volt és van jelentős hatással.

1.2 A tudományos kutatás típusai és folyamata

Az adatgyűjtési technikák és az adatelemzési módszerek változatosságának köszönhetően sokféle kutatási módszer létezik. Röviden bemutatjuk négyféle tipológia szerint, mielőtt a kutatási folyamat tervezési szakaszában részletesen megismerjük.

I. A kutatási módszerek legalapvetőbb osztályozása a **kutatási adatok származása** alapján történik:

1. A szekunder kutatás olyan kutatás, amelynek adatait más, nem az adott kutatási probléma megoldása céljából gyűjtötték.
2. A primer kutatás adatait az adott kutatási probléma megoldására gyűjtik.

II. A primer kutatások a **kutatási adatok jellege** szerint két nagy csoportra bonthatók:

1. Kvantitatív – a kutatási adatokat számszerűsíti és általában statisztikai módszereket alkalmaz az elemzés során.
2. Kvalitatív – a kutatási probléma jobb megértését szolgáló módszer, amely kis mintán alapul, és az eredményei nem általánosíthatók a teljes alapsokaságra.

III. A **kutatás időbelisége** alapján is több típusú primer kutatást különböztetünk meg:

1. Keresztmetszeti kutatás, az információgyűjtés a sokaság elemeiből egyszeri alkalommal vett valamely mintán alapul. Megkülönböztetjük az egyszeri és a többszöri keresztmetszeti kutatást.
2. Longitudinális kutatás: rögzített, ugyanazon a mintán szabályos időközönként megismételt kutatást jelent.

IV. A **kutatási probléma definiáltsága** alapján (Malhotra, 2001):

1. Feltáró kutatás: a kutatási téma pontosabb definiálása, feltárása a cél.
2. Leíró kutatás: a következtető kutatás egy fajtája, amelynek a fő célkitűzése valamely gazdasági vagy társadalmi jellemzőknek vagy funkcióknak a leírása.
3. Ok-okozati: ok és hatás (ok-okozat) kapcsolatáról való bizonyosság megszerzésére használják.

A szakirodalom nagy része ezt a tipológiát tartja alapvetőnek, és ebből vezeti le a kvalitatív, kvantitatív megosztást is. Azonban a gyakorlati kutatások során ezek együttesen is megjelenhetnek, gyakoriak az olyan kutatások, amelyben mindhárom jelleg érvényesül, és szinte valamennyi jól elvégzett kutatásban találunk leíró és ok-okozati elemeket. Például egy primer kvantitatív kutatás során nagy valószínűséggel mindhárom elem megjelenik, az utóbbi kettő biztosan. Ezért nem annyira kutatástípusokról, mint **kutatási jellegről** beszélhetünk.

A tudományos kutatás során megfigyelni kívánt valóság sokszínűsége miatt nehéz a kutatási típusok, módszerek átfedés nélküli, következetes számbavétele. Az alábbi ábrán az előbbi három osztályozási szempont, dimenzió alapján soroljuk be a leggyakrabban használt kutatási módszereket.[1]

1. táblázat. Kutatástípusok osztályozása

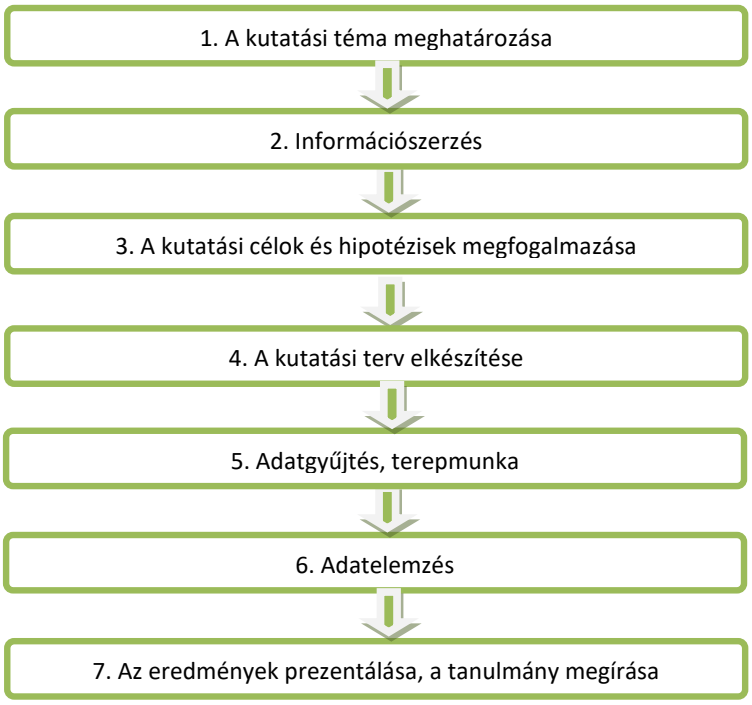
	Kvalitatív	**Kvantitatív**
Szekunder adat	-	- keresztmetszeti: gyakori - longitudinális: gyakori
Primer adat	- keresztmetszeti: nagyon gyakori - longitudinális: nagyon ritka	- keresztmetszeti: nagyon gyakori - longitudinális: ritka

Forrás: saját szerkesztés

A kutatási folyamat (1. ábra) alkotja a könyv gerincét, a gyakorlatnak megfelelően lineáris szerkezetben fogjuk megismerni az egymást követő szakaszokat. Ez a linearitás a folyamat elején visszacsatolásokat tartalmaz, az első négy szakasz kölcsönhatásban van egymással, iteratív eljárással csiszoljuk kutatási tervünket a végleges formára. Nagyon gyakori, hogy a rendelkezésünkre álló (ingyen beszerezhető) szekunder adatok struktúrája a kutatási céljaink vagy akár a témánk újragondolására kényszerít.
A primer és a szekunder kutatás, illetve a kvantitatív és kvalitatív kutatás folyamata a negyedik szakaszban, a kutatási terv részletes kidolgozásánál kezd eltérni egymástól. Ennek ellenére valamennyi típusra érvényes a fenti folyamatábra, a tartalmi eltéréseket a későbbiekben részletezzük.

[1] Léteznek szekunder adatforrásokra támaszkodó kvalitatív kutatások is például a dokumentumelemzés különböző módszerei, de ezeket a jegyzetben nem részletezzük.

1. ábra. A kutatás folyamata



Forrás: saját szerkesztés

A kutatási folyamat részletes ismertetése előtt nézzük meg, hogy a kutatás során milyen **nehézségekkel**, problémákkal kell szembenéznünk:

- A kutatást emberekkel végezzük, ezért:
  - viselkedésüket sok tényező befolyásolja, amelyeket nem lehet mind figyelembe venni;
  - gyakran nehéz vagy nagyon időigényes homogén – fontos ismérvek mentén hasonló – mintát venni;
  - adatvédelmi törvények szabályozzák az adatgyűjtést.
- Általában nem reprodukálható (szemben a természettudományos kutatással).
- Nem mérhető minden ismérv (lehetnek olyan fontos ismérvek, amelyek nem számszerűsíthetők, kvantifikálhatók).
- Felmerülnek etikai kérdések:
  - a részvétel önkéntességével kapcsolatban;
  - a résztvevők jogait és biztonságát garantálni kell;
  - senkit sem érinthet hátrányosan a kutatásban való részvétel;
  - az eredmények nem hamisíthatóak, nem tulajdoníthatóak el (plágium);

  - az eredmények tudatosan nem értelmezhetőek félre.
- Időbeliség – a kutatás folyamán a vizsgált személyek/megfigyelési egységek változhatnak.

1.3 A kutatási téma, a kutatási probléma meghatározása

A kutatási folyamat első és legalapvetőbb szakasza a kutatási téma kiválasztása.

Mindenekelőtt a tudományos kutatás **céljával** kell tisztában lennünk, ami lehet:

- egy tudományterület, diszciplína fejlesztése (pl. a világ százötven országára kiterjedő GLOBE-kutatás jelentős eredményeket hozott a szervezeti kultúra területén);
- valamely gyakorlati vagy elméleti probléma megoldása (pl. egy vállalatvezető választ keres arra a kérdésre, hogy miért csökken az értékesítés, vagy például azt szeretnénk tudni, hogy a romániai vállalatvezetők vezetési stílusa miben tér el mondjuk a csehországiaktól);
- a tudományos munkára való alkalmasság bizonyítása értekezéssel (pl. államvizsga-dolgozat, disszertáció).

Ezután a kutatási **témát** két irányból is meghatározhatjuk: vagy egy tágabb témakör szakirodalmából, gyakorlati munkáiból keressük és jelöljük ki a problémát, vagy a problémát az (üzleti) élet veti fel, és ahhoz keressük a releváns válaszokat. Egy diszciplína fejlesztését célzó tudományos kutatás során főképp az előbbi, míg egy gyakorlati üzleti kutatás során inkább az utóbbi irányból indulunk ki. Államvizsga-dolgozat szerzője számára lehetséges mindkét irányból elindulni.

A témaválasztás során a következő **elvárásokra** figyeljünk:

- a kutatási téma legyen aktuális és releváns;
- legyen behatárolható, hogy milyen tudományterület(ek)hez tartozik a témánk;
- időbeli határai legyenek egyértelműek;
- világos legyen, hogy kikre vonatkozik a kutatás;
- valódi kutatási célokat és hipotéziseket tartalmazzon.

A **szerző szempontjából** jó, ha a következő kérdéseket megvizsgáljuk:

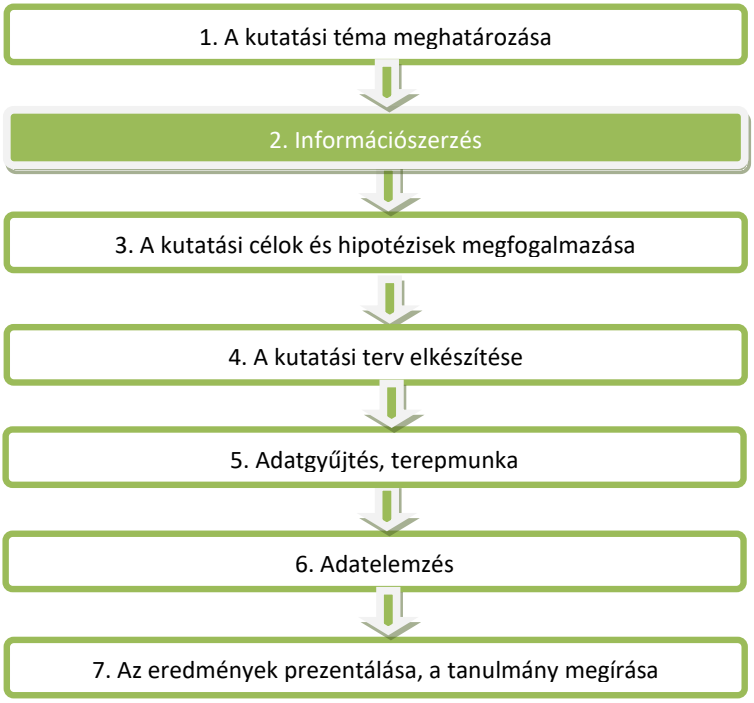
- érdekes, motiváló-e számunkra;
- ismerjüke-e vagy el tudjuk-e sajátítani a téma kidolgozásához szükséges módszertant;
- hozzáférünk-e a szükséges szakirodalomhoz és adatforrásokhoz;
- belefér-e a rendelkezésünkre álló időkeretbe?

Legtöbbször nem ismerjük előzetesen annyira a kiszemelt kutatási témát, hogy megnyugtató válaszokat tudjunk adni a fenti kérdésekre, ezért már itt jelentős a **témavezető** szerepe.
Gyakorlati jó tanács, hogy készítsünk egy feltételes vázlatot az elképzelésünkről, a kutatási kérdésekről, amit a kutatási folyamat következő szakaszában, az előzetes tájékozódás során viszonyítási pontként használunk, de szükség esetén módosítunk.

Szakdolgozatpélda: a kutatás folyamatát egy államvizsga-dolgozat elkészítésének folyamatával példázom. Egy végzős diák úgy érzi, hogy „valamilyen banki témában" szeretne dolgozatot írni. Mivel tanulta és már használja az internetes banki szolgáltatásokat, ezért a kutatási témát a lakossági internetes banki szolgáltatásokra szűkíti. A célcsoport tehát banki ügyfelek lakossági szegmense.

1.4 Információszerzés

A kutatási téma második szakasza olyannyira összetartozik az elsővel, hogy a gyakorlatban nincs jó témaválasztás a téma szakirodalmának alaposabb ismerete nélkül! Kiemelten érvényes megállapítás ez kezdő szerzők, egyetemi hallgatók esetében; annál jobb, minél előbb derül ki a választott téma és a szerző közötti összeférhetetlenség, és választhatunk más témát. Tapasztalatom szerint egyetemi hallgatók számára a szakdolgozat-készítés folyamatában ez a legidőigényesebb szakasz, ezért egy tudatos és célirányos hozzáállással heteket, hónapokat nyerhetünk.



Ebben a kutatási szakaszban bármilyen ötlet vagy tanács jól jöhet elméleti vagy gyakorlati szakemberektől, de legfontosabb a releváns szakirodalom behatárolása és megismerése. Természetesen a kutatási téma meghatározása már a szakirodalom valamilyen szintű ismeretét feltételezte, de most a kutatási témánk alapvető **fogalmi keretével, elméleti modelljeivel és legfontosabb kutatási eredményeivel** ismerkedünk meg.

A szakirodalom megismerése

„Nincs új a nap alatt!", legalábbis előzménye van mindennek. Ne csökkentse az új eredmények iránti elkötelezettségünket a tudat, hogy minden, eszünkbe jutó témának könyvtárnyi szakirodalma van. Még az új jelenségek (pl. a COVID-világjárvány hatása a gazdaságra) vizsgálatánál is figyelembe kell vennünk hasonló témák elemzésének módszertanát és eredményeit.

A szakirodalom elektronikus vagy fizikai megjelenésétől függetlenül a szakirodalmi forrásokat két nagy csoportra osztjuk:

- **Elsődleges forrásoknak** nevezzük a különböző kutatások eredményeiről közvetlenül beszámoló kutatási beszámolókat, monográfiákat[2], disszertációkat, szakcikkeket.
- **Másodlagos források**. Az elsődleges források alapján készített összefoglalók, elemzések, amelyek áttekinthetőbb képet adnak az adott kutatási területről. Például lexikonok, enciklopédiák, tankönyvek, jegyzetek, tanulmánykötetek, szakirodalmi áttekintést nyújtó cikkek.

A témában újdonsült kutatónak ajánlott a másodlagos források feldolgozásával kezdeni, a téma fogalmi keretét, főbb elméleti modelljeit megismerni.
A kutatás folyamatán belül talán a szakirodalom megismerésének módja változott a legradikálisabban az elmúlt húsz évben. A könyvtári katalógusrendszerekben való keresgélés nagyrészt a múlté, az újabb dokumentumok (cikkek, könyvek stb.) elektronikus kiadása ma már alapvető elvárás, és folyamatosan digitalizálják a korábbi tartalmakat is. Már sokszor megjósolták a Gutenberg-galaxis végét, de a nyomtatott könyv és a digitális világhoz alkalmazkodó könyvtárak tartják pozícióikat. A másodlagos forrásokat hatékonyabban megtaláljuk egy jól felszerelt egyetemi könyvtárban, és onnan érdemes az elsődleges, online forrásokat is keresni, mivel hozzáférést biztosíthat fizetős tartalmakhoz is.

A szakirodalom internetes keresése

Könyvek online keresése esetében szinte biztosra mehetünk, ha csak néhány információra van szükségünk a bibliográfiához vagy néhány oldalt, kivonatot keresünk egy pontos idézet miatt, de nehezebb és költségesebb dolgunk van, ha a teljes könyvet szeretnénk.

[2] Önálló dolgozat, szakmai írásmű, amely egy tudományos kérdést minden szempontból kimerítően, egységbe foglalva tárgyal.

Előbbi célból használhatjuk a crossref.org, a Google Books, az Amazon honlapjait, az utóbbi kettő egyes könyvekre ingyenesen hozzáférhető teljes szöveget kínál, és még több megvásárolhatót. Magyar nyelvű könyvek adatait az Országos Széchenyi Könyvtár oldalain (www.oszk.hu) és a fenntartásában levő Magyar Országos Közös Katalógus (www.mokka.hu) oldalain kereshetjük. Szintén az OSZK fenntartásában működik a kezdetek óta hasonló megjelenésű honlappal a Magyar Elektronikus Könyvtár (www.mek.oszk.hu), ahol több mint húszezer mű teljes szövegét tölthetjük le nagy szépirodalmi és nagyon szerény közgazdasági kínálatból.

Lexikonokban és **enciklopédiákban** is érdemes rákeresnünk a választott téma alapvető fogalmaira, elsősorban az elérhető fizikai könyvtárak kínálatából kiindulva az ingyenesen elérhető online tartalmakig.
Magyar kutatók munkáit, teljes életművét a **Magyar Tudományos Művek Tára** (www.mtmt.hu) tartalmazza, és fő erőssége a pontossága, mivel a felsőoktatásban, akadémiai szférában oktató és kutató szerzők maguk töltik fel műveiket. „Teljes dokumentum" jelzéssel, vagyis letölthető teljes szöveggel itt is csak elvétve találkozunk. Megállapíthatjuk, hogy teljes könyvet ma még mindig a fizikai könyvtárakból vagy vásárlással (e-könyv vagy nyomtatott formában) tudunk megszerezni.
Az MTMT-hez hasonló funkciót tölt be oktató-kutatók számára a **Goggle Tudós** (Scholar), de globális méretekben. A professzionálisabb keresőmotornak köszönhetően nagyobb valószínűséggel találunk a szerző által feltöltött teljes dokumentumot.

Szakcikkek teljes körű hozzáférésének legjobb útja az **Elektronikus Információszolgáltatás (EISZ)** Nemzeti Program által nyújtott hozzáférés több nagy nemzetközi kiadó több ezer folyóiratához. Ennek keresője a COMPASS, útmutató a Magyarországon elérhető elektronikus tudományos tartalmakhoz – amint önmagát definiálja. Angolul értők számára az EISZ-en keresztül kitárul a szakcikkek tudományos világa, annak mértékében, hogy milyen előfizetéssel rendelkezik az intézménye. Egyetemi hallgatók általában az intézményi internethálózaton keresztül automatikusan tudnak csatlakozni erre a felületre.

Új fogalom a **repozitórium**, amely az egyetemek, intézmények saját elektronikus, tudományos dokumentumai hozzáférését biztosító tárhely. Az intézményi repozitóriumtól megkülönböztetjük a **tematikust**, amely az intézményi kereteken túllépve tematikailag gyűjti össze, archiválja és bocsátja a kutatók rendelkezésére a

publikációkat. A közgazdaságtan területén ilyen – a csatlakozott szerzők feltöltött dokumentumait és kiterjedt bibliográfiát tartalmazó – repozitórium a RePEc (Research Papers in Economics).

Mivel nincs fejlődés kutatás nélkül és nincs kutatás, illetve kutatási eredmény publikálása a szakirodalom ismerete nélkül, ezért az internet térhódításával párhuzamosan a szakcikkeket megjelentető folyóiratok kiadása hatalmas forgalmú üzleti szolgáltatássá nőtte ki magát. A hozzáférési díjak jelentősen emelkedtek, és annak ellenére, hogy az egyetemek megpróbálják ezt finanszírozni, gyakran kerülhetünk olyan helyzetbe, hogy nem tudjuk letölteni a kutatásunk szempontjából fontos szakcikket. Vannak olyan oldalak, ahol a **dokumentum elektronikus azonosítója** (DOI száma) alapján nagy valószínűséggel hozzáférünk a keresett cikkhez. Ilyen például a SCI-HUB, amely több mint 88 millió letölthető dokumentumot tartalmaz, jelszava, hogy „bontsunk le minden akadályt a tudomány útjából!", és amelyet ezúton nem szeretnénk reklámozni.

A szakirodalom feldolgozásának különböző technikái vannak. A hagyományos technikák – mint például a jegyzetelés, cédulázás, kulcsszavak definiálása és keresése – számítógépesített alkalmazása felgyorsítja a szakirodalom feltárásának folyamatát.

Szakdolgozatpélda. A szakdolgozatát tervező hallgatónk eldöntötte, hogy kutatási témája az online, lakossági banki szolgáltatások vizsgálata lesz. Kezdeti lelkesedése kissé megtört, amikor az egyetemi könyvtárban közölték vele, hogy ilyen címmel vagy tárgyszóval rendelkező, magyar nyelvű könyvük nincs. Nem lett boldogabb akkor sem, amikor *ráguglizott* a témára, és a több mint egymillió találat bősége hozta zavarba. Ezek között volt néhány biztató cikk, illetve egyetemikurzus-tananyag PowerPoint-os formátumban, amelyek segítettek az alapvető fogalmakat tisztázni. Témavezetője javaslatára célzottan keresett és talált hasonló témájú mesteri és doktori disszertációkat, így már lehetséges kutatási kérdésekkel és módszertannal is megismerkedett. Legjobban a dolgozatok bibliográfiáinak örült, ami után célzottabban kereshette a szakirodalmat, és egy kis támpontot kapott a különböző szakirodalmi tételek fontosságának megítéléséhez is.

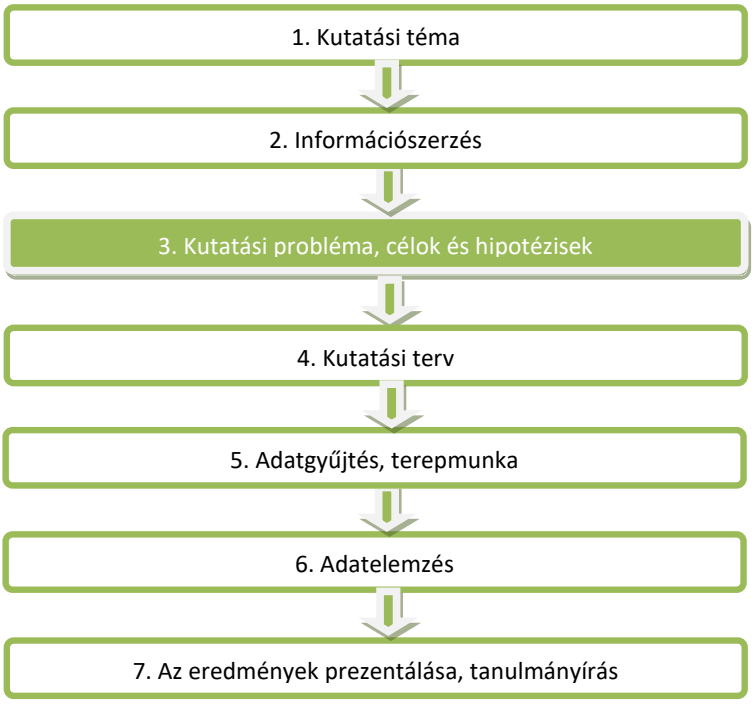
Szakértői megkérdezés – az adott kutatási téma szakértőivel folytatott interjúk, amelyek segíthetnek a kutatási téma pontosabb behatárolásában. Tőlük általában a kvalitatív kutatási módszereknél részletezett egyéni mélyinterjúhoz hasonló módszertannal, formális kérdőív nélküli személyes interjúval nyerhetünk információkat. Olyankor érdemes a szakértői interjút lefolytatnunk, amikor már

képbe kerültünk a téma alapvető fogalmaival és fontosabb elméleteivel, illetve többször konzultáltunk a témavezetőnkkel is.

Szakdolgozatpélda: az internetes banki szolgáltatások államvizsgatémát választó hallgatónk az egy-két szakkönyvrészletből és több jó szakfolyóiratból álló szakirodalom megismerése után interjút kér egy banki szakemberrel és/vagy a témában jártas kutatóval.

1.5 A kutatási probléma, célok és hipotézisek megfogalmazása

Ebben a szakaszban arra keressük a választ, hogy milyen tudományos állítást akarunk megfogalmazni, vagy milyen üzleti problémára keressük a megoldást. A **kutatási probléma** a kutatási téma számunkra releváns részének a behatárolását jelenti.



A kutatási probléma szakszerű meghatározása lehetetlen a szakirodalmi tájékozódás során megismert fogalmak nélkül. **Konceptualizálásnak** nevezzük a kutatásban alkalmazott fogalmi keret meghatározását, a kutatás szempontjából fontos fogalmak egyértelmű definiálását.

Erre egyrészt nyilvánvalóan azért van szükség, hogy kapcsolatot teremtsünk a szakirodalommal, másrészt ugyanannak a fogalomnak eltérő értelmezése egymással nem összehasonlítható mérési skálák alkalmazásához vezet. Empirikus kutatás során ugyanis a vizsgált fogalmakhoz mérési lehetőségeket kell rendelnünk. A konceptualizálás azt szolgálja, hogy **valóban azt vizsgáljuk és mérjük, amire szükségünk van**. Például ha egy ország gazdaságának *teljesítményéről* van szó, akkor szinte mindenki a GDP-re gondol, már csak a mérési skálát kell pontosítanunk, hogy az összes GDP-t abszolút értékben nemzeti devizában vagy euróban fejezzük ki, egy főre jutó GDP-re gondolunk, vagy épp annak éves szintű változására százalékban kifejezve. Jóval nehezebb a dolgunk, ha összetettebb fogalmat, például egy ország gazdaságának *versenyképességét* szeretnénk vizsgálni. Olyannyira összetett fogalom, hogy eltérő elméleti modellek végeredményei azok az összetett mutatók, indexek, amelyet különböző nemzetközi rangsorokban használnak.

A kutatási téma szempontjából fontos fogalmak szétbontását dimenzióira, kutatási változókká és ezekehez mérési skálák rendelését **operacionalizálásnak** nevezzük. A **változó** tehát egy fogalom ismérve, jellemzője, amihez egyértelmű mérési skála rendelhető.

A kutatási probléma meghatározásában segít, ha megválaszoljuk a következő, egymással összefüggő kérdéseket:

- Mit tudunk már? A szakirodalmi tájékozódás alapján megfogalmazunk néhány vizsgált kutatási problémát.
- Mit kell tudnunk? Megvizsgáljuk – lehetőleg a témavezetővel közösen –, hogy a szakirodalmi kutatási problémák közül melyek adaptálhatók a kutatási környezetünkre, lehetőségeinkre.
- Miért kell tudnunk? Az előzőekben behatárolt lehetséges kutatási problémákat most a relevanciájuk szerint rangsoroljuk és választunk közülük. Azt vizsgáljuk, hogy miért fontos és miért érdekes számunkra az adott kutatási probléma.
- Mit fogunk tenni, hogy megtudjuk? Itt már a továbblépés, a kutatási terv irányába kell gondolkodnunk, ki kell tűznünk a kutatási célokat.

Egy szerényebb kutatási probléma önmagában egyetlen kutatási célt is jelenthet, de általában többre bonthatjuk. A **kutatási célok**[3] megfogalmazása a kutatási

[3] A szakirodalomban kutatási kérdésként is szerepel, amennyiben nem állításként, hanem kérdőmondatként fogalmazzuk meg a kutatási céljainkat.

probléma több, egyértelműen és tömören megfogamazott állításra való szétbontását jelenti. A kutatási cél legyen:

- Specifikus. Nem általánosságban, hanem konkrétan fogalmaz.
- Mérhető. A kutatási cél fogalmaihoz mérhető változók rendelhetők.
- Reális. Kutatási lehetőségeinknek, erőforrásainknak megfelelő cél.
- Határidőhöz kötött. Meg kell becsülnünk – a tapasztalt témavezetőnkkel egyetemben – a tervezett kutatási cél időigényét.

Továbbhaladva az operacionalizálás útján, egy kutatási célt egy vagy több kutatási hipotézisnek feleltetünk meg. Ezzel elértünk az operacionalizálási folyamat utolsó szakaszához, ez lesz elméletünk (szerényebben fogalmazva kutatási eredményünk) „atomi" része. A **kutatási hipotézis** olyan állítás, amelyben a kutatási témára, annak változóira, vagy ezek kapcsolatára vonatkozó, önálló feltételezéseinket fejezzük ki. Olyan állítás, amely korábbi kutatási eredményekre támaszkodik, megteremti az empirikus kutatásunk és az elmélet közötti kapcsolatot.
A jó kutatási hipotézis az alábbi **jellemzőkkel** rendelkezik:

- egyértelmű kijelentő mondatban van megfogalmazva, amelyben azonosíthatók a kutatási változók;
- logikailag magyarázatot ad a kutatási problémára;
- ok-okozati kapcsolat vizsgálata esetén egyértelmű kapcsolatot jelöl;
- lehetőleg új ismeretet eredményez vagy megkérdőjelezi a korábbiakat;
- a kutatás tervezése során fogalmazzuk meg, hogy meghatározza az adatgyűjtési és elemzési terveinket;
- minden esetben ellenőrizzük, kvantitatív kutatásnál statisztikai hipotézisvizsgálattal.

Eljutottunk odáig a kutatás folyamatában, ameddig „papír-ceruzával" és a szakirodalommal körülvéve magunkat el lehet jutni.
A kvalitatív és kvantitatív kutatások vízválasztója a kutatási hipotézisek megfogalmazása és ellenőrzése, tesztelése. A kvantitatív kutatások során a kutatási hipotéziseket a statisztikai hipotézisvizsgálat módszereivel teszteljük, míg a kvalitatív kutatási adatok elemzésénél a kutató szubjektív elemzőképessége a döntő.

1.6 Statisztikai hipotézisvizsgálat

A kutatási hipotézist mérhetővé kell tennünk, ezért megfogalmazzuk a **statisztikai hipotézist**, ami a megfigyelt sokaság valamely ismérvére vonatkozó matematikai formába öntött állítás. Kvantitatív kutatások során valamennyi kutatási célt és az abból következő kutatási hipotézist végső soron statisztikai hipotézisvizsgálattal ellenőrizzük. A hipotézisvizsgálat statisztikai módszer, amely segítségével eldöntjük, hogy hipotézisünket elfogadjuk vagy elutasítjuk.
A **hipotézisvizsgálat folyamata** (Tóthné Lőkös 2008 alapján[4]):

1. a statisztikai hipotézis megfogalmazása a kutatási hipotézis alapján, H_0 és H_1 felállítása;
2. a szignifikanciaszint (α) kiválasztása;
3. a próbafüggvény megválasztása és aktuális értékének kiszámítása;
4. kritikus érték kikeresése a megfelelő táblázatból;
5. döntés a hipotézis (H_0) elfogadásáról vagy elvetéséről;
6. szakmai következtetés levonása a hipotézisnek megfelelően.

A statisztikai hipotézisvizsgálatot tehát valamennyi hipotézisünkre el kell végeznünk. Ezért annak ellenére, hogy a mai adatelemző, statisztikai programcsomagok a fenti folyamat nagyját automatikusan elvégzik, a következőkben részletesen ismertetjük a hipotézisvizsgálat folyamatát. Egy tapasztalt kutató a hipotézisvizsgálat folyamatából többnyire csak az első (a statisztikai hipotézis megfogalmazása) és az utolsó (a szakmai következtetés levonása) szakasznál „dolgozik", a többi a számítógép feladata, de természetesen jó, ha ismeri a többi részfolyamatot is.

1. A statisztikai hipotézis megfogalmazása

A statisztikai hipotézist a kutatási hipotézisből vezetjük le úgy, hogy a verbális állítást matematikai, logikai formába öntjük. A statisztikai hipotézis a sokasági ismérvek (változók) eloszlásának a paramétereire (átlag, szórás vagy az eloszlás típusa) vonatkozik. A kutatási hipotézis elfogadását úgy szigorítjuk, hogy a hipotézis tagadását jelentő állítást tekintjük kiindulásként érvényesnek, ettől pedig csak akkor állunk el, ha ez a hipotézisvizsgálat alapján indokolt (Hajdu, 2003). E nyakatekertnek tűnő megközelítés mögött matematikai-statisztikai indokok állnak.

[4] Az említett szakirodalomban a hipotézisvizsgálat első lépése a szakmai probléma felvetése, a kutatási hipotézis megfogalmazása, mi ezt a korábbiakban külön tárgyaltuk.

A kutatási hipotézis érvénytelen voltát jelentő állítást nevezzük alap- vagy **nullhipotézisnek (H_0),** az alternatíváját képező kutatási hipotézist pedig **alternatív hipotézisnek (H_1).** A nullhipotézis (H_0) mindig az ismérvek egyenlőségét fogalmazza meg, az alternatív hipotézis (H_1) pedig értelemszerűen ennek alternatíváját.

A hipotézisvizsgálatot egy, illetve két változó esetén végezhetjük el, a null-, illetve az alternatív hipotézis megfogalmazását a már elkezdett szakdolgozatpéldánkon keresztül mutatjuk be.

Szakdolgozatpélda: megállapítottuk, hogy kutatási célunk a csíkszeredai X bankfiók lakossági ügyfélkörén belül az online banki szolgáltatások iránti potenciális kereslet vizsgálata, illetve a szolgáltatást már igénybe vevők szegmensének a jellemzése. Ezek alapján az egyik kutatási hipotézisünk az, hogy a X bankfiók lakossági ügyfélkörén belül az online banking szolgáltatást igénybe vevők aránya eltér az országos – banktól független – átlagtól.

I. Egy változó esetén – egy mintabeli változót össze akarunk hasonlítani egy külső, nem mintabeli értékkel.

Alaphipotézis (H_0). Pl. a bankfiók lakossági ügyfélkörén belül az „online banking" szolgáltatást igénybe vevők aránya megegyezik az országos – banktól független – aránnyal.

Alternatív hipotézis (H_1):

Kétoldali alternatív hipotézis:		Pl. a két arány nem egyenlő
Egyoldali alternatív hipotézis:	- bal oldali:	Pl. a helyi arány nagyobb az országosnál
	- jobb oldali:	Pl. a helyi arány kisebb az országosnál

II. Két változó esetén – két mintabeli változó statisztikáit (átlagát, szórását vagy eloszlását) hasonlítjuk össze egymással.

Alaphipotézis (H_0): Pl. a bankfiók lakossági ügyfélkörén belül az "online banking" szolgáltatást igénybe vevő férfiak és nők aránya azonos.

Alternatív hipotézis (H_1):

Kétoldali alternatív hipotézis:		Pl. a két arány nem egyenlő
Egyoldali alternatív hipotézis:	- bal oldali:	Pl. a férfiak aránya nagyobb
	- jobb oldali:	Pl. a férfiak aránya kisebb

A hipotézisvizsgálat során tehát a sokaság valamely ismérvéről megfogalmazott állításunk igaz voltát ellenőrizzük úgy, hogy a sokaságból vett véletlen minta statisztikáit összehasonlítjuk ún. tesztstatisztikákkal. De a véletlen minta statisztikáinak hipotézisvizsgálata során kétféle hibát is elkövethetünk:

2. táblázat. A hipotézisvizsgálat lehetséges hibái

		Tény	
		Igaz	Hamis
Döntés	Elfogadjuk	Helyes döntés (1-α)	Másodfajú hiba (β)
	Elvetjük	Elsőfajú hiba (α)	Helyes döntés (1- β)

Forrás: Szűcs, 2004

Elsőfajú hiba (α-val jelöljük): a nullhipotézis igaz, de mi elutasítjuk. Az elsőfajú hiba elkövetésének valószínűségét **szignifikanciaszintnek** nevezzük. A másodfajú hibát (β) akkor követjük el, ha a nullhipotézis nem igaz, de mi elfogadjuk.

2. A szignifikanciaszint (α) megválasztása a kutató szubjektív döntése, a legáltalánosabban elfogadott szignifikanciaszint az 5%. Ettől az értéktől természetesen eltérhetünk lefelé vagy felfelé, ezáltal fokozva, illetve lazítva a hipotézisvizsgálat szigorát.

A statisztikai programcsomagokban vagy beállítható a szignifikanciaszint, vagy a teszt eredménye egy ún. **empirikus szignifikanciaszinttel (p érték) számol, ami az a legkisebb valószínűség, amely mellett a H_0 elvethető H_1-gyel szemben.**
Például ha az SPSS-programmal vizsgáljuk két változó átlagának azonosságára vonatkozó nullhipotézist és az eredmény P = 0.02, akkor 98%-os biztonsági szinten állíthatjuk, hogy a két változó átlaga különbözik. Másképp fogalmazva, ha H_0-t elutasítjuk, akkor 2% az esélye, hogy hibásan döntöttünk, és 98% a valószínűsége, hogy helyes döntést hoztunk.

Figyelembe véve, hogy a legáltalánosabban elfogadott szignifikanciaszint az 5%, megfogalmazhatjuk a kvantitatív elemzések leggyakrabban igénybe vett, legnépszerűbb **aranyszabályát: ha az empirikus szignifikanciaszint kisebb, mint 5% ($p < 0.05$), akkor állíthatjuk, hogy a vizsgált két érték nem azonos, eltér egymástól!**

3. A próbafüggvény értékének és a kritikus értéknek a kiszámítása

A mintabeli változó adatai alapján kiszámoljuk az ún. **próbafüggvény** értékét. A hipotézistől függően az alábbi próbafüggvényeket alkalmazzuk a leggyakrabban:[5]

- a változó középértékére (átlagára) vonatkozó próbák: t-próba és z-próba
- szórások összehasonlítására vonatkozó próba: F-próba
- eloszlásokra vonatkozó próba: Khi-négyzet próba

4. A kritikus érték kikeresése a megfelelő táblázatból

A próbafüggvény értékének kiszámítása után egy táblázatból kikeressük a kritikus értéket, amit a próbafüggvény, a választott szignifikanciaszint és a minta elemszámának ismerete alapján egyértelműen meghatározhatunk. A mintaelemszámot nem közvetlenül használjuk, hanem egyes próbafüggvényeknél – mint például a t-próba, F-próba – ez alapján számoljuk ki az úgynevezett szabadságfokot[6] (df). A különböző próbafüggvények esetében eltérő módon számoljuk ki a szabadságfok értékét, a legegyszerűbb esetben eggyel csökkentjük a mintaelemszámot (df = n-1).

5. Döntés a statisztikai hipotézis elfogadásáról vagy elvetéséről

A szignifikanciaszint meghatározása, a próbafüggvény értékének kiszámolása és a kritikus érték táblázatból való kikeresése után a döntés már automatikus, a két értéket összehasonlítva fogadjuk vagy utasítjuk el a nullhipotézist. Mérlegelnünk akkor kell, ha az előre meghatározott szignifikanciaszinten el kell utasítanunk a nullhipotézist, de a szignifikanciaszint csökkentésével már elfogadhatóvá válik. Ez azt jelenti, hogy nem 95%-os biztonsági szinten utasítjuk el a nullhipotézist, hanem csak 90%-osan.

A statisztikai programcsomagok előbb ismertetett empirikus szignifikanciaszintje megkímél attól, hogy az újabb szignifikanciaszinteknek megfelelő kritikus értéket újra és újra kikeressük addig, amíg megtaláljuk azt a legkisebb valószínűséget, amely mellett a nullhipotézis elutasítható.

6. A szakmai következtetés levonása

A statisztikai hipotézisvizsgálat után a statisztika területéről visszajutunk arra a tudományterületre, amelyre a kutatási témánk vonatkozik. Megvizsgáljuk, hogy a kutatási hipotézisünk elfogadása vagy elutasítása alapján az adott kutatási témához

[5] A próbafüggvények eszköztára bővülhet aszerint, hogy egy- vagy kétmintás próbákról beszélünk.
[6] A szabadságfokot angol elnevezése – degree of freedom – alapján általában df-nek rövidítjük.

kapcsolódóan milyen szakmai következtetéseket fogalmazhatunk meg. Lehetséges, hogy a végeredményhez jutottunk és már csak az eredmények prezentálására kell figyelnünk, de az is lehet, hogy az eredmények újabb kutatási hipotéziseket vetnek fel.

Szakdolgozatpélda. A szakirodalom szerint városunkban az X bank lakossági ügyfélkörének az online bankinget igénybe vevők aránya az országos átlag alatt van. Az okok kereséséhez a következő további kérdésekre kell választ találnunk:

– mekkora az X bank online bankoló ügyfeleinek országos aránya? Ezzel a külső, nem a kutatásból származó információval azt a hipotézist ellenőrizzük, hogy az X banknak általában országosan is kevesebb internetes szolgáltatást igénybe vevő ügyfele van, mint a többi banknak.

– ha igazolódott, hogy nem a bank online banking stratégiájával van a probléma, akkor a csíkszeredai ügyfélkör az országos átlag alatti keresletére kell magyarázatot találnunk:

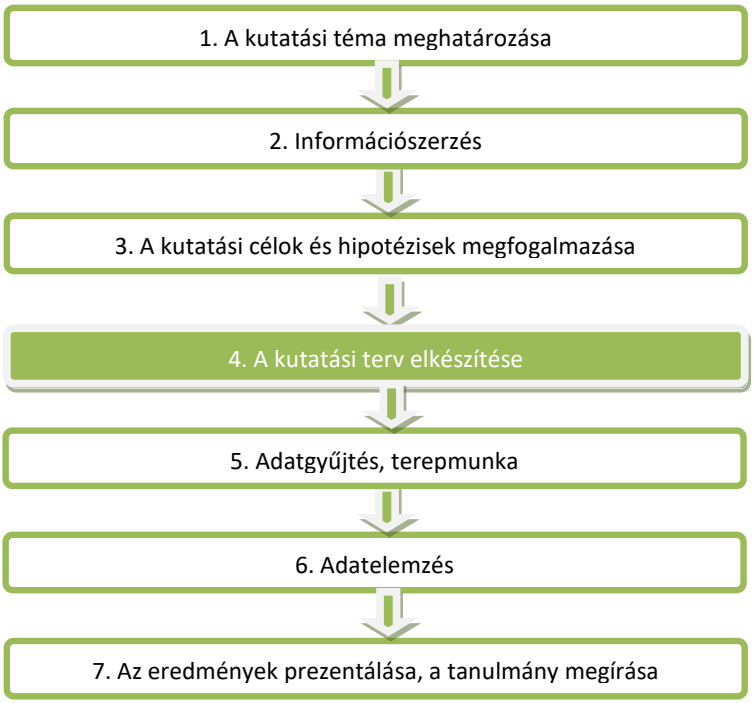
– a csíkszeredai internetkapcsolattal rendelkező háztartások aránya kisebb, mint az országos átlag.

– a csíkszeredai X bankfiók ügyfélkörének más az – online bankinget is meghatározó – demográfiai összetétele, mint az országos átlag.

– hasonló internet-penetráció és a banki ügyfélkör hasonló demográfiai összetétele mellett is előfordulhat, hogy az online banking „még nem érkezett" meg, nem jutott el a fogyasztók tudatába.

Megjegyzendő, hogy a kutatási hipotéziseken túl a statisztikai hipotézisvizsgálat részleteit nem minden kutató szokta feltüntetni egy nagyobb kutatás eredményeit bemutató tanulmányában, de a kutatás módszertani felkészültségéről bizonyságot tevő egyetemi vagy doktori hallgatóknak ez mindenképp ajánlott.

2. A KUTATÁSI TERV ELKÉSZÍTÉSE



A kutatási hipotézisek konkretizálása során olyan kérdések is felmerülnek (kikre vonatkozik a kutatás?; a teljes alapsokaságot vizsgálom, vagy csak egy mintát belőle?; milyen változókkal tudom számszerűsíteni a kutatási hipotézist?), amelyek megválaszolása átvezet bennünket a kutatási folyamat következő szakaszába: a kutatási terv elkészítéséhez. **A kutatási terv a kutatási hipotézisek vizsgálatához szükséges adatok megszerzésének és elemzésének a részleteit adja meg.**

A kutatási terv legfontosabb részei:

- a kutatási módszer(ek) kiválasztása
- mintavételi terv
- kérdőívszerkesztés
- adatelemzési terv.

A kutatási terv alapján – nem utolsósorban – képesek leszünk a kutatás idő- és költségigényét is meghatározni.

2.1 Kutatási módszerek tipológiája

A kutatás részletes tervezési szakaszában az első feladatunk a kutatási módszer vagy akár módszerek kiválasztása. A következőkben a különféle kutatási módszereket, típusokat mutatjuk be a korábban ismertetett tipológiának megfelelően. A kutatási témánk meghatározása, az információszerzés és a kutatási célok megfogalmazása után rátérünk a kutatási módszer kiválasztására.

1. Feltáró jellegű kutatás: a kutató az előzetes tájékozódás után is csak kevés információval rendelkezik az adott kutatási témáról, nem tud pontos kutatási hipotéziseket megfogalmazni. Az információszükségletet csak nagy vonalakban lehet meghatározni. A kiválasztott minta kicsi és nem megfelelően reprezentálja a sokaságot, egyféle betekintést nyújt a sokaság elemeibe. A feltáró kutatásokat rugalmasság és módszertani változatosság jellemzi, formális kutatási tervet nem alkalmaznak – szakértői megkérdezés.

2. Leíró jellegű kutatás: fő célkitűzése valamely gazdasági vagy társadalmi jellemzőknek vagy funkcióknak a leírása. Előre megfogalmazott specifikus kutatási hipotézisek és a probléma pontos megfogalmazása jellemzi, általában nagy reprezentatív mintákon alapul, előre tervezett és jól strukturált kutatási tervvel meghatározza az információforrásokat és az adatgyűjtés módját.

3. Ok-okozati jellegű kutatás: ok és hatás (ok-okozat) kapcsolatáról való bizonyosság megszerzésére használják. Hasonlóan a leíró kutatásokhoz, az ok-okozati kutatásokat is előre tervezett módon és strukturáltan kell felépíteni. Célja annak megértése, hogy mely változók az okozó (független) és melyek az okozatok (függő változók), illetve a független változók közötti kapcsolat természetének meghatározása.

Mint említettük, a gyakorlati kutatások során ezek együttesen is megjelennek, ezért nem élesen elkülönülő kutatástípusokról, hanem kutatási jellegről beszélünk. Például összetett, jól felépített kutatásoknál gyakori, hogy a kutatási téma előzetes megismerésénél alkalmazott szekunder vagy kvalitatív kutatásban egyértelműen a feltáró jelleg dominál, ezt követi egy kvantitatív kutatás, amelyben egyaránt megtaláljuk a leíró és az ok-okozati részeket. Ma már inkább minőségi szinteket jelölnek ezek, piackutatóktól nem fogadnak el olyan elemzést, amelyben leírják ugyan a gazdasági jelenséget, de nem keresnek magyarázatot, ok-okozati összefüggést a leírt jelenségekre.

A kutatási módszereket tehát a kutatási adatok származása, jellege és az időbelisége alapján csoportosítjuk. Az alábbi (3.) táblázatban e három dimenzió mellett megjelenítettük azt is, hogy a különböző módszerek mennyire jellemzőek az üzleti, a tudományos és az államvizsga-dolgozat megírása céljából végzett kutatások körében.

3. táblázat. Kutatástípusok osztályozása

	Kvalitatív	**Kvantitatív**
Szekunder adat	-	**1. szekunder-kvantitatív-keresztmetszeti** - üzleti: gyakori - tudományos: gyakori - államvizsga: gyakori
		2. szekunder-kvantitatív-longitudinális: - üzleti: gyakori - tudományos: nagyon gyakori - államvizsga: gyakori
Primer adat	**3. primér-kvalitatív-keresztmetszeti**: - üzleti: nagyon gyakori - tudományos: nagyon gyakori - államvizsga: gyakori	**5. primer-kvantitatív-keresztmetszeti**: - üzleti: nagyon gyakori - tudományos: nagyon gyakori - államvizsga: gyakori
	4. primér-kvalitatív-longitudinális: - üzleti: ritka - tudományos: nagyon ritka - államvizsga: -	**6. primer-kvantitatív-longitudinális** - üzleti: gyakori - tudományos: ritka - államvizsga: nagyon ritka

Forrás: saját szerkesztés

A kutatási módszer kiválasztásának egyik fontos szempontja, hogy a kutatási sokaságról mi gyűjtjük össze a kutatási adatokat (primer kutatás) vagy más felmérések, kutatások adataiból és eredményeiből új következtetéseket vonunk le (szekunder kutatás). Nézzük előbb a szekunder kutatások jellemzőit!

2.2 Szekunder kutatás

A szekunder kutatás tehát **olyan kutatás, amelynek adatait más, nem az adott kutatási probléma megoldása céljából gyűjtötték** (Malhotra, 2001). Hangsúlyoznunk kell, hogy annak ellenére, hogy más kutatás során létrehozott adatokat használunk, a kutatásunk nem más kutatás reprodukálása, hanem új kutatási eredmények megfogalmazását jelenti. Nem arról van szó, hogy a szekunder kutatás során nincs adatgyűjtés – általában a teljes kutatási idő arányában több időt kell fordítanunk adatgyűjtésre, mint a primer kutatásnál –, hanem a tapasztalati megfigyeléseket nem mi számszerűsítjük, nem mi mérjük. A piackutatásban a szekunder kutatást a terepmunka hiánya miatt „íróasztal-kutatásnak" (*desk research*-nek) nevezik.

A szekunder kutatás **előnyei** a primer kutatáshoz képest:

- Idő-, munka- és pénzmegtakarítás. Valóban a primer kutatás sok közvetlen és közvetett költsége nem jelenik meg egy szekunder kutatásnál, de a munkaidő igénye a vártnál nagyobb lehet. Paradox módon a szekunder kutatás alapját képező adatforrások túláradó bősége növelheti a kutatás időigényét. Például a lehető legegzaktabban megfogalmazott internetes keresés ellenére is több ezer oldal „salak" közül kell kiemelnünk az újracsiszolandó gyémántunkat. Ennek ellenére – ha rendelkezésünkre állnak a megfelelő adatok – a szekunder kutatás jóval gyorsabb és olcsóbb a primernél.
- Olyan adatokhoz való hozzáférés, amelyeket primer adatgyűjtéssel nem tudnánk összegyűjteni. Legfontosabb indoka a szekunder kutatásoknak az, ha a kutatási témánk szempontjából meghatározó adatok olyan jellegűek, hogy a primer adatgyűjtés lehetetlen a kutatásunk térbeli, időbeli és erőforráskorlátai miatt. Például a nemzetközi vagy akár országos összehasonlításhoz szükséges adatok vagy kutatásunk olyan megfigyelési egységekre vonatkozik, amelyek mérésére nincs lehetőségünk, pl. vállalati értékesítési, termelési adatok.

A szekunder kutatás **hátrányai:**

- Régi adatok. Még az olyan kutatásnál is, ahol a kutatás alapvető célja a kutató felkészültségének a bizonyítása (államvizsga-dolgozat), nem ajánlott egy-két évesnél régebbi adatokat használni, inkább váltsunk kutatási témát vagy egy kisebb primer kutatással szerezzük be a szükséges adatokat.

- Pontatlan, tudománytalan források – az adatok hitelességére, pontosságára nincs garancia, ezért törekedjünk neves intézményektől, kutatócégektől szerezni az adatokat.
- Az adatok szerkezete nem felel meg az adatigényünknek. A leggyakoribb és legbosszantóbb problémája a szekunder kutatásnak, ha a kutatási céljainknak megfelelő friss adatokat találunk, de olyan szerkezetben, hogy nem tudjuk az elvárásainknak megfelelően átstrukturálni.

A szekunder kutatást tehát egyaránt tekinthetjük kényszerűségnek a primer adatgyűjtés kivitelezhetetlensége miatt, és lehetőségnek is, ha találunk a kutatási témánk szempontjából releváns adatforrásokat, és nem szükséges a költségesebb primer kutatást lefolytatnunk. A gyakorlati tapasztalat alapján legtöbbször az a helyzet áll elő, hogy a szekunder adatok nem teljes mértékben felelnek meg a kutatási céljainknak, és ekkor vagy a célok kompromisszumos újratervezésére van szükség, vagy primer kutatásra törekszünk. Ezért a kutatás megtervezésének nagyon fontos eleme, hogy tisztában legyünk a kutatási célok eléréséhez szükséges adatok jellegével és begyűjtési lehetőségével.

Adatgyűjtésünket is felgyorsítja, áttekinthetőbbé teszi, ha rendszerezzük a szekunder információ-forrásokat:

- az adathordozó,
- az adatok tartalmi jellege,
- az adatközlő alapján.

1. A szekunder kutatás információforrásait az **adathordozók** alapján csoportosíthatjuk:

- internet,
- elektronikus média,
- nyomtatott média.

Nyilvánvalóan a három közül a legfontosabb az internet, ma már nem nagyon van olyan fontos információ, amelynek ne lenne elektronikus verziója és azt ne tennék fel az internetre, de a **nyomtatott** sajtóban, könyvekben, statisztikai kiadványokban is találhatunk kutatásunk szempontjából lényeges információkat. Az **elektronikus médiára** jellemző példa az üzleti kutatások alapsokasági nyilvántartását jelentő cégadatbázisok. Ezek a CD-n elérhető adatbázisok általában egy multinacionális szakosított cég termékei, és meglehetősen nagy pontossággal

tartalmazzák egy országban bejegyzett valamennyi cég alapvető adatait. Ezeket az információkat leggyakrabban direkt marketing célra használják, de a vállalati szegmensre vonatkozó kutatások során a piackutató cégek, kutatók számára a mintavétel alapja. Hasonló adatbázisokat még a statisztikai hivatalok is szolgáltatnak.

De a szekunder adatok legkézenfekvőbb és legtöbb adatot tartalmazó „tárhelye" az **internet**. Adatgyűjtési szempontból fontos megkülönböztetnünk a **fizetős** és az **ingyenes** oldalakat.

- **Fizetős adatforrások.** A szekunder kutatásunk sokkal gyorsabb lehet és a kutatási céljainknak pontosabban megfelelő adatokat találhatunk, ha a különböző adatbázisok elérését lehetővé tevő internetes oldalak (portálok) szolgáltatásait vesszük igénybe.
- **Ingyenes adatforrások**. Az internetet nem véletlenül nevezhetjük az emberiség legdemokratikusabb „intézményének", ma is nagyrészt érvényes az induláskor megfogalmazott alapelv, hogy bárki tartalmat (adatokat) tölthessen fel vagy le. Az ingyenesen elérhető adatforrásokat leginkább a bőség zavarával jellemezhetjük, nagyon időigényes megtalálni a keresett adatot és könnyen meglehet, hogy nem találjuk meg. A keresési folyamat gyorsításának és egyszerűsítésének lehetőségét kínálja a minél pontosabban fogalmazott internetes keresés. A keresőmotorok használatakor ajánlott a részletes keresés-beállítási (Advanced Search) lehetőségeket maximálisan kihasználni. Gyakran igénybe vett opciók: pontos kifejezés keresése (egy mondatrész egészét keresi, nem csak a szavakat külön-külön), fájltípus, dátum, nyelv, származási hely, a keresett kifejezés megjelenési helye a honlapon.

Példa: tételezzük fel, hogy a könyv második felében ismertetett SPSS adatelemzési program gyakorlásához szeretnénk adatokat gyűjteni. Az SPSS adatfájlok megtalálásához a Google keresőjét úgy állítjuk be, hogy „.sav" kiterjesztésű fájlokat keressen, azaz beírjuk a filetype: sav kifejezést. Szűkítsük tovább a keresést úgy, hogy biztosak legyünk abban, hogy az adatfájl gazdasági, üzleti információkat is tartalmaz. A fájltípus beállítása mellett írjuk be a keresőbe az árbevétel kifejezést angolul (*turnover*), így a keresés jóval kevesebb találatot eredményez, ami már kényelmesen áttekinthető.

2. Az **adatok tartalmi jellegük** alapján nagyon sokfélék lehetnek, a kutatásunk tervezett témája határozza meg, hogy makrogazdasági, kereskedelmi, demográfiai, vállalati szintű, háztartási, egyéni fogyasztói stb. adatokra van szükségünk.

3. Az **adatközlők** alapján a következő felosztást tehetjük:

- **nemzetközi intézmények** publikus adatai. Az ENSZ, EU szervezeteinek (IMF, IBRD, Eurostat stb.) nagyon sok kiadványa, statisztikai adata elérhető az interneten;
- **belföldi országos hatáskörű intézmények** publikus adatai (statisztikai hivatal, nemzeti bank, államigazgatási intézmények, kutató intézetek stb.), **egyéb belföldi intézmények**: egyetemek, főiskolák, szakkönyvtárak és más nonprofit intézmények publikus adatai;
- **professzionális adatközlő**knek nevezzük az adatgyűjtést, elemzést és közlést üzleti alapon végző cégeket. Legfontosabbak a szaklapok és azok tematikus mellékletei, mivel valamennyi eredményük nyilvános, sőt alapvető érdekük, hogy minél könnyebb legyen a hozzáférés. Ezenkívül az adatgyűjtést és kutatást üzleti alapon végző piac-, marketing- és közvélemény-kutató cégek, reklámügynökségek, pénzügyi intézetek és egyéb, az üzleti szolgáltatások területén tevékenykedő cégek is marketincélból nyilvánossá tesznek értékes adatokat, részeredményeket;
- **vállalati publikus** adatok. Törvényi kötelezettsége valamennyi bejegyzett cégnek nyilvánossá tenni a mérlegét, ez megtekinthető például a pénzügyminisztérium honlapján (www.mfinante.ro);
- **vállalati belső** adatok. Ha kutatásunk egy cég megbízásából történik, akkor természetesen rendelkezésünkre állhatnak a szükséges, de nem nyilvános belső adatok. Külső kutató számára ezek már csak a vállalatvezetéssel kötött egyezség alapján válnak elérhetővé.

2.2.1 Szekunder keresztmetszeti kutatás

A szekundér adatforrások áttekintése után nézzük az **adatelemzés módját**. A kutatástípusok osztályozásánál (3. táblázat) láthattuk, hogy a szekunder kutatások körében csak a kvantitatív módszereket tüntettük fel. Viszonylag szűk körű speciális kutatások során használnak kvalitatív módszereket is (pl. dokumentumelemzés), de ezek bemutatása nem tartozik könyvünk tematikájába.
Az adatgyűjtés részfolyamatának lezárása után már nincs lényeges különbség a szekunder-kvantitatív és a primer-kvantitatív kutatások között az adatelemzés tekintetében, de a kutatás időbelisége alapján megkülönböztetett keresztmetszeti és

a longitudinális kutatások már jelentősen meghatározzák az elemzés módját. A kvantitatív keresztmetszeti kutatások adatelemzési módszereiről a későbbiekben lesz szó, most a longitudinális, az időtényezőt figyelembe vevő kutatási módszer sajátosságait vizsgáljuk.

2.2.2 Szekundér longitudinális kutatás – idősorelemzés

Az **idősorelemzés** a longitudinális kutatások egy típusa, amely **a megfigyelni kívánt társadalmi-gazdasági jelenségek változását, fejlődését az idő függvényében mutatja be.** Általában szekunder információforrásból származó adatokkal végezünk idősorelemzést, államvizsga-dolgozat készítése céljából nagyon gyakran végzett kutatás.

Az idősorok jellemzői (Ertsey in Szűcs, 2004):

- az idősorok lehetőleg minél hosszabbak legyenek, minél hosszabb időintervallumot fogjanak át.
- az adatfelvételek időpontja közötti időintervallumok hossza legyen azonos (pl. nem lehet az adatsorunk egyik részében heti, másikban havi adataink).
- az adatok tartalma azonos kell legyen mindegyik időpontban, nem változhatnak az osztályozási rendszerek vagy a mértékegységek.
- az adatoknak azonos megfigyelési típusból kell származniuk, nem keverhetők össze az alapsokasági adatfelvételek a különböző mintavételből származó adatokkal vagy becslésekkel.

Az idősorelemzés modelljei két, egymástól lényegesen különböző modellből, illetve ezek kombinációiból állnak (Hunyadi et al., 1996):

- a **determinisztikus** idősorelemzés feltételezi, hogy az idősorok előre determinált hosszú távú pályát követnek. Az elemzés fő célja, hogy ezt a pályát meghatározza, elemeire bontsa és ezek segítségével hosszabb távon is előre jelezze. A véletlen hatását az idősorokra szükséges rossznak tekinti, és igyekszik kiszűrni.
- a **sztochasztikus** idősorelemzés felfogása szerint a véletlen szerves alkotórésze az idősornak, ezért a valószínűségek vizsgálata a modellezés alapja.

A sztochasztikus idősorelemzés modelljeinek ismerete túlmutat jegyzetünk tartalmán, a következőkben a determinisztikus modellt részletezzük. Az idősorelemzés alapvető célja, hogy az idősort felbontsuk négy összetevőjére, komponensére. Ezt az eljárást nevezzük **dekompozíciónak**.

- **Trend** – az idősorban tartósan, hosszú távon érvényesülő tendencia a fejlődés irányát és mértékét meghatározó legfontosabb komponens.
- **Szezonalitás** – szabályos, rövid távú, rendszeresen ismétlődő ingadozás.
- **Ciklikus komponens** – szabálytalan, hosszabb távú ingadozás.
- **Véletlen ingadozás** – a zavaró hatásokat leíró véletlen változó, az előbbi három determinisztikus komponens által nem megmagyarázott érték.

Az idősor additív (összegző) modellje a következő

$$Y = \hat{Y} + S + C + \varepsilon \qquad (1)$$

alakban írható fel, ahol $\hat{Y}$ a trend, S a szezonalitás, C a ciklikus komponens, ε a véletlen ingadozás. Az additív modellen kívül létezik multiplikatív modell is, amelyben a komponensek összeszorzódnak.

Példa. Nézzünk egy rövid példát a dekompozíció jelentőségére. A csíki sörgyárak (az igazi és a még igazibb) számára elengedhetetlen a napi értékesítési adatok éves szintű idősorának felbontása e négy komponensre. Könnyen belátható, hogy ez az értékvesztés nélkül sokáig nem tárolható termék értékesítési előrejelzései alapvető fontosságúak a termelésirányítás, illetve a cég több funkcionális területe számára. Empirikus tapasztalataink alapján feltételezhetjük, hogy a sörfogyasztásban jelentős szezonalitás van, nyáron sokkal többet fogyasztva. Kérdés, hogy a szezonhatást lebontva, több év átlagában milyen trend érvényesül, növekszik vagy csökken az értékesítés? Érdemes-e termelésbővítésbe beruháznunk, vagy csak a nyári szezonnak tulajdonítható, hogy nem tudjuk kielégíteni a keresletet, és a fogyasztó a konkurens termékekre fanyalodik? És ha kivételesen nem csak egy napra esik a csíki nyár, akkor a nyári szezonhatáson túl további keresletnövekedést okozhat a harmadik komponens valamilyen turistacsalogató eseménye (csíksomlyói búcsú, városnapok). Mindezek figyelembe vétele mellett is a véletlen meglephet egy kiadós jégesővel, ami téli szintre csökkentheti az aznapi sörfogyasztást.

1. Trendszámítás

Az idősorelemzés alapvető célja tehát meghatározni az idősor egészén uralkodó trendet. Erre alapvetően két modell áll rendelkezésünkre, mindkettőt érdemes megismernünk, mert a dekompozíció során mindkettőre szükségünk lesz.

1.1 Mozgó átlagolás

Az idősor adataiból láncszerűen továbbhaladó átlagolással egy újabb idősort képezünk, amely értékei jelentik a trend értékeit. A mozgó átlagolású trendszámítás lényege, hogy az idősor t-edik eleméhez úgy rendelünk trendértéket, hogy átlagoljuk az idősor t-edik elemének bizonyos környezetében lévő elemeket (Hunyadi et al., 1996). Legegyszerűbb esetben a t-edik elemet megelőző és követő értékeket vesszük figyelembe:

$$\hat{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3} \qquad (2)$$

ahol $\hat{y}_t$ a trend, y_t pedig az idősor t-edik eleme. A mozgó átlagolás folyamata:

- meghatározzuk a mozgó átlag tagszámát. Az előbbi esetben 3 tagú mozgó átlag (m=3) képletét írtuk fel. A tagszám lehet páros vagy páratlan is;
- kiszámoljuk a trend értékeit, 3 tag esetén a következő módon:

$$\hat{y}_2 = \frac{y_1 + y_2 + y_3}{3},\ \hat{y}_3 = \frac{y_2 + y_3 + y_4}{3},\ldots,\hat{y}_{n-1} = \frac{y_{n-2} + y_{n-1} + y_n}{3} \qquad (3)$$

Megállapíthatjuk, hogy az idősor első és utolsó értékére nem tudunk trendértéket számolni, mivel nincs azt megelőző, illetve követő érték. A trend adatsora példánkban két értékkel rövidebb, mint az eredeti idősor, és a különbség változhat a tagszám értékétől függően.

A mozgóátlagolás lényege, hogy **kisimítja az idősor szezonális és ciklikus hullámzásait.** Minél nagyobb tagszámot választunk, azaz minél nagyobb időintervallumot helyettesítünk egyetlen értékkel az átlagával, annál simább, kiegyenlítettebb lesz az idősorunk. Ennek azonban ára van, a tagszám növelése csökkenti a trend intervallumát.

Ha az idősor grafikus képe alapján feltételezésünk lehet a szezonalitás hullámhosszára (pl. heti, havi, negyedéves stb.), akkor a tagszámot tegyük egyenlővé ezzel az értékkel. Ebben az esetben jól kisimítjuk a szezonhatást, ellenkező esetben vagy nem simít eléggé a trendünk, vagy csak „eltoljuk" a szezonális hullámokat. Megfelelően kisimított szezonhatás és elég hosszú távú ciklikus komponens esetén a mozgó átlagolású adatsorunk nemcsak a trendet, hanem a ciklikus komponenst is tartalmazza ($\hat{Y}_{m.a.} = \hat{Y} + C$).

1.2 Analitikus trendszámítás

Az idősor trendjét valamilyen jól illeszkedő függvénytípussal fejezzük ki. Az idősor tényleges értékeire a legkisebb négyzetek módszerével illesztjük a függvényt úgy, hogy az idősor értékei és a függvényértékek közötti távolság a lehető legkisebb legyen. Leggyakrabban alkalmazott függvénytípusok: lineáris, exponenciális, hiperbolikus, polinomiális, logisztikus. A függvényillesztésről további részleteket a jegyzet SPSS-program használatáról szóló részben, az 5.6 alfejezetben találunk.

2. A szezonális ingadozás mérése

A szezonális ingadozás azt mutatja, hogy **az idényhatás következtében az idősor értékei átlagosan milyen mértékben térnek el a trendtől.** Az idősor additív alapmodellje

$$Y = \hat{Y} + S + C + \varepsilon \qquad (4)$$

és az alapján, hogy a mozgó átlagolású trendünk nemcsak a trendet, hanem a ciklikus komponenst is tartalmazza ($\hat{Y}_{m.a.} = \hat{Y} + C$), ha kivonjuk az idősorból a mozgóátlagolású trend értékeit, akkor a szezonális és a véletlen ingadozás összegét kapjuk:

$$S = Y - \hat{Y}_{m.a.} + \varepsilon \qquad (5)$$

A véletlen hatását úgy zárjuk ki, vagy legalábbis csökkentjük, hogy átlagoljuk a szezonális ingadozásokat. Ha előzetesen az idősor grafikus képe alapján felismertük, hogy mekkora a szezonális ingadozás hullámhossza,[7] akkor ezen az időintervallumon átlagoljuk a szezonális ingadozás értékeit.

$$s_j = \frac{\sum_{i=1}^{n}\left(y_{ij} - \hat{y}_{ij(m.a.)}\right)}{n} \qquad (6)$$

ahol j a szezonális ingadozások száma adott időintervallumon belül. Feltételezésünk, hogy a szabályos, szezonális ingadozások kiegyenlítik egymást, tehát az összegük, illetve az átlaguk nulla kell legyen. Ezért ezeket az értékeket **nyers szezonális ingadozásnak** nevezzük, és ha nem felelnek meg az említett feltételnek, akkor szükség van a **korrigált szezonális ingadozás** kiszámolására:

[7] Ezt használtuk a mozgó átlagolás tagszámának a meghatározására is.

$$\hat{s}_j = s_j - \frac{\sum_{i=1}^{m} s_j}{m} \qquad (7)$$

ahol m a szezonális ingadozások száma.

3. A ciklikus komponens meghatározása.
A hosszú távú, szabálytalan ciklikus komponenst már könnyen meghatározhatjuk az

$$\hat{Y}_{m.a.} = \hat{Y} + C \qquad (8)$$

összefüggésből, ahol az $\hat{Y}_{m.a.}$ a mozgóátlagolású és az $\hat{Y}$ az analitikusan meghatározott trend, a ciklikus komponens pedig a kettő különbsége.

2.3 Primer kutatás

A kutatási adatok jellege szerint két nagy csoportra bontjuk a primer kutatásokat: **kvalitatív és kvantitatív** kutatásokra. A módszertana alapján ez a két kutatás típus jól elkülöníthető, a gyakorlati kutatások nagy többsége jól besorolható az egyik vagy másik típusba.

2.3.1 Kvalitatív kutatás

A kvalitatív kutatás feltáró jellegű, a probléma megértését szolgáló kutatási módszer. Lényeges különbség a kvantitatív kutatással szemben, hogy a kutatási adatokat kis mintából gyűjtjük, és a **minta nem reprezentatív, vagyis az eredményeket nem általánosíthatjuk a teljes alapsokaságra**. Kvalitatív kutatást olyankor célszerű alkalmazni, amikor a kutatási témánk olyan kérdéseket tartalmaz, amelyekre **az emberek közvetlenül valószínűleg nem tudnak, vagy nem akarnak válaszolni.** Ilyenek lehetnek például az emberek egészségi állapotára, higiéniájára, különböző attitűdjeire, véleményére (pl. fajgyűlölet), vagyoni helyzetére, luxuscikkek vásárlására, márkahűségük okaira stb. vonatkozó kérdések.
Az emberek **értékeire, motivációira, érzelmi mozgatórugóira vagyunk kíváncsiak,** és ezek feltárására nem, vagy csak részben alkalmasak a kvantitatív kutatás direkt kérdései.

A kvalitatív kutatás **jellemzői**:

- kis, nem reprezentatív minta, de kis alapsokaság esetén lehet teljes körű is;
- módszertani rugalmasság és változatosság;
- az adatok elemzése nem statisztikai módszerekkel történik, az eredmények gyakran szubjektív értelmezésen alapulnak;
- mélyebb, nem nyilvánvaló ok-okozati összefüggések feltárására is alkalmas;,
- nehezen definiálható kutatási célok esetén megalapozhat egy kvantitatív kutatást;
- bizalmas vagy bonyolultabb témák vizsgálatára is alkalmas.

4. táblázat. A kvalitatív és a kvantitatív kutatás összehasonlítása

	Kvalitatív	**Kvantitatív**
Célkitűzés	A mögöttes okok és motivációk minőségi megértése	Az adatok számszerűsítése és az eredmények általánosítása a mintáról az alapsokaságra
Minta	Kisszámú, nem reprezentatív	Nagyszámú, reprezentatív
Adatgyűjtés	Nem strukturált	Strukturált
Adatelemzés	Nem statisztikai	Statisztikai

Forrás: Malhotra alapján, 2001

A lényeges különbségek ellenére vagy épp emiatt a két kutatástípusra nem egymást helyettesítő, hanem kiegészítő módszerekként kell tekintenünk. Egy komplex kutatási téma esetében, ha a kutatási idő- és költségkeret megengedi, akkor ideális esetben a kvalitatív kutatás megalapozza a kvantitatívot, pontosítja, hogy a kutatási témán belül milyen konkrét hipotéziseket fogalmazzunk meg és teszteljünk statisztikai módszerekkel.

Azonban – mint már említettük – a kutatási témánk lehet olyan vagy tartalmazhat olyan részeket, ami főképp vagy kizárólag kvalitatív módszereket igényel. Ennek eldöntése a kutatás megtervezésének alapját jelenti, mivel ez a döntés meghatározza a kutatás egészének további folyamatát.

Az előző fejezetben ismertetett **kutatási folyamat alapvető szakaszai megegyeznek mindkét kutatástípusnál, de lényeges különbségek vannak a kutatás tervezése során a mintavétel és a kérdőívkészítésnél, továbbá az adatgyűjtés és az adatelemzés során**.

A kvalitatív kutatás általunk vizsgált típusai az egyéni mélyinterjú és a fókuszcsoport.[8]

[8] Ezeken kívül több, speciális kvalitatív módszer létezik (lásd Malhotra, 2001), amelyeket itt nem részletezünk. Ilyen például a projektív technikák csoportja, amelyek elfedik a kutatás valódi célját és a válaszadót nem arra kérik, hogy saját magatartását írja le, hanem hogy mások magatartását értelmezze.

Egyéni mélyinterjú

Közvetlen, személyes interjú, amelyben **egy képzett kérdező az interjúalany motivációit, nézeteit, attitűdjeit vizsgálja**. Az interjú során nem a kvantitatív kutatások során használt strukturált kérdőívet alkalmazzuk, hanem egy olyan **interjú vezérfonalat, ami tartalmazza az előzetesen kigondolt kutatási kérdéseket, de lehetőséget nyújt új gondolatok, összefüggések feltárására**. Az egyéni mélyinterjút gyakran használjuk az előzetes tájékozódás során a kutatási téma alaposabb feltárására, de jelentheti a kutatás fő módszerét is. Főképp kis költségvetéssel rendelkező kutatók vagy egyetemi hallgatók körében népszerű, de nem a költségvetési korlátok, hanem mindenekelőtt a kutatási témának kell megindokolnia az alkalmazását.
Az interjú alanyai két fő csoportból származhatnak: a kutatási téma szakértőiből és/vagy a megfigyelni kívánt alapsokaságból. A kétféle interjúalany értelemszerűen kétféle megközelítést, kétféle strukturálatlan kérdőívet igényel ugyanannál a témánál is.

A mélyinterjú jellemzői:

- felkészültséget, a kutatási téma nagyfokú ismeretét igényli,
- az interjú hossza általában fél - egy óra,
- személyesebb vagy bonyolultabb témák is megbeszélhetők,
- nincs a kvantitatív kutatásokhoz hasonló strukturált kérdőív, hanem csak egy vázlat, ami kizárólag nyílt kérdéseket tartalmaz vagyis a kérdező nem határozza meg előzetesen a kérdésre adható válaszlehetőségeket,
- a kérdőív vázlat ellenére a kérdések megfogalmazását és sorrendjét a válaszadó feleletei befolyásolják.

Fókuszcsoport

A fókuszcsoport strukturálatlan és közvetlen interjú, amelyben **egy jól felkészült kérdező (moderátor) beszélget a kutatási célsokaság egy csoportjával**. Módszertanában nagymértékben hasonlít az egyéni mélyinterjúhoz, de kihasználja a csoportos interjú adta többletlehetőségeket. Fontos különbség, hogy **a csoportos beszélgetés dinamikájának köszönhetően olyan váratlan eredmények merülhetnek fel, amire a kutatók nem is számítottak**.
A csoportinterjú a kutató által ellenőrzött és az interjúalanyok által kellemesnek mondható környezetben zajlik. Videofelvétel készül az interjú során, ami segíti a moderátort az interjú után a kutatási eredmények megfogalmazásában, lehetővé teszi, hogy az elhangzottakon túl figyelembe vegye az interjúalanyok hangulatát,

érzelmeit, metakommunikációs jelzéseit is. Piackutató cégeknél a fókuszcsoport-terem legtöbbször olyan tükörfallal rendelkezik, ami lehetővé teszi, hogy más kutatók (pl. a moderátor segítsége) vagy a kutatást megrendelő cég képviselője kívülről figyelje a történéseket, sőt be is avatkozhatnak elektronikus üzeneteket küldve a moderátor számítógépére.

Piackutatók körében annyira elterjedt ez a kutatási technika, hogy sokan a kvalitatív kutatás szinonimájaként használják (Malhotra, 2001). A kutatócégek, illetve a moderátorok specializálódnak egy-egy kutatási területre (pl. fiatalok, háziasszonyok, vállalati közép- és felsővezetők körében végzett fókuszcsoportok, vagy téma szerinti specializáció: távközlési technológiák és szolgáltatások, egészségügy stb.).

Az előzőkben említett infrastrukturális igények miatt a tudományos célú kutatások egyéni kutatói vagy a kutatói felkészültséget bizonyító hallgatók kevésbé tudják igénybe venni ezt a technikát, de egy szerényebb körülmények között lefolytatott csoportinterjú is jó eredményekkel kecsegtethet. A fizikai feltételek megteremtése mellett komoly feladat a 8-12 interjúalany meggyőzése és közös időpont egyeztetése.

A fókuszcsoport alkalmazása:

- a kutatási téma előzetes megismerésénél említettük, hogy a fókuszcsoport alkalmazható a kutatási téma alaposabb kidolgozásához, szempontokat nyújthat a kvantitatív kérdőívek szerkesztéséhez, a kvantitatív kutatási hipotézisek megfogalmazásához és korábbi kvantitatív kutatások eredményeinek értelmezéséhez;
- a kutatási célkitűzéseket fókuszcsoport alkalmazásával kívánjuk elérni. A piac- és marketingkutatások során a következő célokra használják rendszeresen a fókuszcsoportot:
    - valamilyen termékkel kapcsolatos fogyasztói percepciók, preferenciák és magatartás megértése;
    - reklám kreatív koncepcióinak és reklámszövegeknek a kialakítása;
    - termékinnovációk tesztelése. Ma már a termékinnovációs folyamat része a koncepció vagy a kész prototípus tesztelése a célpiacról származó fókuszcsoportban;
    - adott marketingmixszel kapcsolatos fogyasztói reakciók előzetes megismerése.

Szakdolgozatpélda. Esettanulmányunkat folytatva, az államvizsgára készülő hallgatónk alkalmazhatja a fókuszcsoport-technikát. A helyi bankfiókok szakembereivel és/vagy külső szakértőkkel folytatott csoportinterjún a következő alapvető kérdésekről beszélhetnek: hogyan látják a lakossági bankszektor fejlődését, az online banking várható keresletét, melyek a legnépszerűbb szolgáltatások és ezek igénybe vehetők-e interneten keresztül.

Emellett egy másik fókuszcsoportra is szükség lenne, sőt talán az előbbinél is fontosabb információkat szolgáltathat a keresleti oldal vizsgálata. A potenciális fogyasztók köréből származó csoportban azt az alapvető kérdéskört kellene vizsgálni, hogy hogyan vélekednek a potenciális fogyasztók a közgazdászok és programozók által létrehozott ezen új szolgáltatásról.

A fókuszcsoport jellemzői:

- a fókuszcsoport moderátora az adott témából jól felkészült szakember, de azonkívül jó megfigyelő, kapcsolatteremtő és kommunikációs képességekkel rendelkezik;
- az előzetesen szelektált 8-12 fős csoportok demográfiai, társadalmi-gazdasági jellemzők alapján homogének kell legyenek;
- a csoportinterjú 1-3 óra időtartamú;
- videofelvétel segíti az interjú utáni feldolgozást.

Az eddigiek alapján is egyértelmű, hogy a **fókuszcsoportos kutatás folyamata** sok részletében különbözik a kvantitatív kutatásétól (Malhotra 2001 nyomán):

1. **A fókuszcsoport által megválaszolandó kérdések meghatározása** – az eddig definiált kutatási célok és a kutatási technika (fókuszcsoport) ismeretéből le kell vezetnünk és meg kell fogalmaznunk egy részletes listában, hogy milyen kérdésekre keressük a választ.
2. **Az interjúalanyokat kiválasztó szűrő kérdőív megírása és toborzás** – a résztvevők során alkalmazott szűrő kérdőívvel biztosítjuk a csoport demográfiai szempontból homogén jellegét.
3. **A moderátor interjú-vezérfonalának összeállítása** – ez a moderátor és a kutató(k) közötti szoros együttműködésen kell alapuljon. A moderátornak alaposan ismernie kell az adott témát, és hogy melyik kérdéssel milyen kutatási célt ér el.
4. **A fókuszcsoport lebonyolítása** – az interjú elején a moderátor bemutatkozik és bemutatja a résztvevőket, ismerteti a csoportvita

szabályait, meghatározza a célokat, megpróbál vitát generálni, összefoglalja a válaszokat, ha a csoporttagok egyetértésre jutottak.

5. **A felvételek visszajátszása, az adatok elemzése** – a csoportinterjú után a moderátor és általában még egy kutató az emlékeik és a videofelvétel alapján leírják a kutatási eredményeket. A megfogalmazott vélemények mellett figyelemmel vannak a vélemény erősségét kifejező verbális és metakommunikációs jelzésekre (arckifejezések, gesztusok), felismerik az új, a moderátori vezérfonalban nem érintett, de releváns gondolatokat, meghatározzák a csoportot legjobban összetartó és megosztó kérdéseket. Mindezek alapján próbálják megrajzolni azt az összképet, ami a csoport véleményét a kutatási kérdésekkel kapcsolatban legjobban kifejezi.
6. **A tanulmány megírása** során a kis minta miatt nem számszerűsítjük az adatokat, nem használunk gyakorisági eloszlást vagy átlagokat. Például ha a tízes csoportból hatan valamilyen kérdésben egyetértettek, akkor nem a csoport 60%-áról beszélünk, hanem tipikusan olyan megfogalmazást használunk, hogy „a válaszadók több mint fele" vagy „a résztvevők szűk többsége".

A fókuszcsoport hátrányai:

- a kisminta alapján nem vonhatunk le általános következtetéseket az alapsokaságra nézve;
- a csoportos környezet bátorítóan, de fékezően is hathat a véleményalkotásban;
- fennáll a veszélye annak, hogy a csoport tagjai hasonulnak egy domináns résztvevő véleményével;
- kevés a jól képzett moderátor;
- az adatok rendezetlensége. A csoportbeszélgetés során elhangzott vélemények nem strukturáltak, elemzésük és értelmezésük nehéz feladat;
- nagy szerepet kap az elemző szubjektivitása, a kutatási adatokat könnyebben félre lehet értelmezni, mint a kvantitatív kutatások során.

2.3.2 Kvantitatív kutatás

Kvantitatívnak nevezzük az olyan kutatást, amelyben **a számszerűsített adatok elemzése és az eredmények az alapsokaságra való általánosítása során statisztikai módszereket használunk**. A megfigyelni kívánt alapsokaságról előre megszerkesztett, strukturált kérdőív alkalmazásával gyűjtünk adatokat, amelyet – a kvalitatív kutatástól eltérően – a kutatási folyamat közben nem szabad módosítanunk.
A valóság, a tapasztalati tények számszerűsítése és elemzése sokféleképp történhet, mivel a megfigyelni kívánt valóság is nagyon változatos. Ezért a formalizált kvantitatív kutatások többféleképp osztályozhatók:

- a kérdőíves megkérdezés módja alapján;
- a kutatás időbelisége alapján.

1. A kvantitatív kutatási módszerek osztályozása **a kérdőíves megkérdezés módja alapján** gyakorlati szempontból nagyon lényeges:

1.1 Személyes kérdezés:[9] A személyes kérdezést nevezhetjük a klasszikus értelemben vett interjúnak, amelynek során a kérdezőbiztos személyesen találkozik az interjúalannyal. A személyes kérdezés a helyszín alapján lehet otthoni, irodai, utcai vagy – újabban – bevásárlóközponti kérdezés. Irodai vagy munkahelyi interjúkat olyan kutatások során folytatunk, amikor a célsokaság az intézményi vagy a vállalati szegmensből kerül ki. Az utcai kérdezés egyre inkább kezd visszaszorulni a bevásárlóközponti javára, ahol sokkal megfelelőbb körülmények adottak. Ez a típusú kutatás nyújt lehetőséget a leghosszabb és a legbonyolultabb vagy személyesebb témájú megkérdezésre.

1.2 Telefonos kérdezés során az interjúalanyt otthonában, üzleti kutatások esetén munkahelyén vagy mobiltelefonon hívja a kérdezőbiztos. A telefonos kérdezések rövidebbek, mint a személyesek, a kérdezés hossza általában nem lehet hosszabb, mint 15 perc, és a kutatás témája is csak kevésbé bonyolult vagy személyes lehet.

1.3. Internetes kérdezés alatt webes felületen, honlapon elhelyezett és kitöltött kérdőívet értünk. Az internetes kérdezések teret nyertek az utóbbi években az internethasználat terjedésének és az internetes kutatás gyorsaságának és

[9] A kutatási szaknyelvben gyakran használják az angol megnevezését is: *face to face (F2F)*.

olcsóságának köszönhetően. Egyre gyakoribbak a specializált kutatócégek, amelyek **internetes paneleket** próbálnak kiépíteni, azaz a potenciális válaszadók olyan csoportját, akik hajlandóak időközönként – egy felkérő e-mail hatására – a kutatócég honlapján kitölteni a kérdőívet. A paneltagok toborzása és megtartása ajándékok, jutalmak kisorsolásával biztosítható.

1.4 Postai úton történő kérdezés során az interjúalany az önkitöltő kérdőívet postán vagy e-mailen keresztül kapja, és kitöltése után visszaküldi a feladóhoz. A szakirodalom érthető módon az e-mailben elküldött kérdőívet az internetes kérdezések közé sorolja, azonban sokkal nagyobb különbség van a webes és az e-mailes kérdezés között kérdezéstechnikai és egyéb kutatási szempontból, mint aközött, hogy hagyományos módon vagy elektronikusan küldjük a levelet.

A postai úton történő kérdezéseket egyszerűségek és olcsóságuk ellenére ritkán használják kutatócégek, mivel nagyon kicsi a válaszadási hajlandóság, az elküldött kérdőíveknek csak töredéke érkezik vissza, ami lehetetlenné teszi a kutatás időbeni lefolytatását. Ilyen típusú adatgyűjtést a statisztikai hivatalok használnak, akik nagyobb arányban, de szintén csak részlegesen kapják vissza az elküldött kérdőíveket, annak ellenére, hogy a felmérésbe bevont cégeknek, intézményeknek törvény által előírt adatbeszámolási kötelezettségük van. Üzleti kutatások körében újságok, lapok alkalmazzák ezt a módszert, rövid, maximum egyoldalas kérdőívet csatolva a laphoz. A kutatás sikerességét, a válaszadási hajlandóságot az olvasótábor lojalitása mellett jutalmak kisorsolása is befolyásolja.

A **módszerek közötti választást** a következő táblázatban (5.) feltüntetett szempontok alapján dönthetjük el. Ritka alkalmazásuk miatt kihagytuk a postai úton keresztüli és a hagyományos telefonos kérdezést, ezek használata egyéni kutatók nagyon szűk anyagi keretekkel rendelkező kutatásai során fordulhatnak elő. A személyes kutatási formák közül külön jellemezzük az otthoni és a bevásárlóközponti, illetve a számítógéppel segített (*Computer Aided Personal Interview*) kérdezést.

5. táblázat. A kvantitatív kutatási módszerek összehasonlítása

	Telefonos CATI	Otthoni személyes	Bevásárló-központi	CAPI	Internet
Az adatgyűjtés rugalmassága	közepes	magas	magas	magas	közepes
A kérdések változatossága	alacsony	magas	magas	magas	közepes, magas
Fizikai ingerek alkalmazása	alacsony	közepes	magas	magas	közepes
A minta elérhetősége	közepes	magas	közepes	közepes	alacsony, közepes
A kérdezés környezetének ellenőrzése	közepes	közepes, magas	magas	magas	alacsony
A terepkutatók ellenőrzése	közepes	alacsony	közepes	közepes	magas
Az adatok, kérdések mennyisége	alacsony	magas	közepes	közepes	közepes
Válaszadási arány	közepes	magas	magas	magas	alacsony
Kényes kérdések lehetősége	magas	alacsony	alacsony	alacsony	magas
A kérdezőbiztosi torzítás lehetősége	közepes	magas	magas	alacsony, közepes	nincs
Gyorsaság	magas	közepes	közepes, magas	közepes, magas	nagyon magas
Költségek	közepes	magas	közepes, magas	közepes, magas	alacsony

Forrás: Malhotra nyomán, 2001

A szempontok közül vizsgáljuk meg a kevésbé egyértelműeket. Az **adatgyűjtés rugalmasságát** a kérdezőbiztos és az interjúalany együttműködési lehetősége határozza meg. Személyes kutatásoknál összetett, bonyolultabb kérdőíveket is le lehet kérdezni, mivel a kérdezőbiztos megmagyarázhatja a nehezebb kérdéseket. **Fizikai ingerek alkalmazása** alatt termékek, reklámfilmek, promóciós anyagok bemutatását vagy fizikai ingerek (pl. ízteszt) alkalmazását értjük.

A **válaszadási arány** fontos mutatója az adatgyűjtés hatékonyságának, a sikeres interjúk arányát mutatja százalékban kifejezve a megkezdett interjúk számához képest. Legmagasabb arányt a bevásárlóközponti személyes megkérdezéseknél érnek el, általában 100 megszólított potenciális interjúalanyból több mint 80-nal sikerül befejezni az interjút.

A **kérdezőbiztosi torzítás lehetősége** sajnos többféle is lehet: nem az előírásoknak megfelelően választja ki az interjúalanyt, nem megfelelően teszi fel a

kérdéseket (a későbbiekben látni fogjuk, hogy milyen a jó kérdezés) vagy hibásan rögzíti a válaszokat.
Mindezeket a potenciális hibaforrásokat kiküszöbölik a CAPI és az internetes kutatások leprogramozott kérdőívei. Végül, de nem utolsósorban a piackutatók számára talán két legfontosabb szempont a **gyorsaság** és a **költségek**; tudományos célú kutatások során talán kevésbé fontosak.

2. A kutatás időbelisége alapján is több típusú primer, kvantitatív kutatást különböztetünk meg:

1. Keresztmetszeti kutatás: ezen belül megkülönböztetjük az egyszeri és a többszöri keresztmetszeti kutatást.
2. Longitudinális kutatás: ugyanazon a rögzített mintán ismételten végeznek méréseket.

2.1. Keresztmetszeti kutatás: az adatgyűjtés az alapsokaság elemeiből egyszeri alkalommal vett valamely mintán alapul. Az **egyszeri** keresztmetszeti kutatásban az alapsokaságból csakis egyetlen mintát vesznek, és az információgyűjtés ebből az egyetlen mintából és csak egyszeri alkalommal történik. A **többszöri** keresztmetszeti kutatás során két vagy több mintát különböző időpontban vesznek az alapsokaságból.
Az üzleti kutatások túlnyomó többsége egyszeri, keresztmetszeti kutatás. A kutatás alapvető célja nyilvánvalóan meghatározza a módszert, például kutatók, egyetemi vagy doktori hallgatók által végzett, a módszertani felkészültséget igazoló primer kutatások szinte kizárólag egyszeri keresztmetszeti kutatások. Cégek vagy állami intézmények által végzett/megrendelt kutatások között azonban gyakoriak a célpiac fogyasztói magatartásának, társadalmi-gazdasági tényezők változására irányuló többszöri keresztmetszeti kutatások.
Marketingkutatók az ugyanarra az alapsokaságra vonatkozó, de mindig más mintát vizsgáló, rendszeres időközönként (akár hetente) végzett többszöri kutatásokat **tracking**-nek nevezik.

2.2. Longitudinális kutatás: ugyanazon a rögzített mintán ismételten végeznek méréseket. A primer longitudinális kutatás abban különbözik a keresztmetszeti kutatástól, hogy **a minta nem változik az idő folyamán,** ugyanazok a résztvevők alkotják. Az ilyen mintát **panelnek** nevezzük.

Előnye a keresztmetszetivel szemben, hogy felfedi a változásokat, mert ugyanazokat az ismérveket ugyanazon a mintán ismételten méri.[10]
Hátránya, hogy nagyon nehéz fenntartani a minta változatlanságát, kiküszöbölni a résztvevők különböző okok miatti lemorzsolódását.

Példa: szinte az egész társadalmat naponta befolyásoló kereskedelmi televíziók üzleti modellje a nézettségre vonatkozó audiométernek nevezett longitudinális kutatásokon alapul. A reklámidő beárazásához mind a TV-társaság, mind a reklámidőt vásárló cég marketingeseinek ismernie kell, hogy adott időben hányan és milyen demográfiai ismérvekkel jellemezhető személyek nézik a tévét. A kutatás e célja alapján inkább többszöri keresztmetszeti kutatásról beszélhetnénk, mivel a hangsúly nem a változáson van, hanem meghatározott időintervallumokban végzett felmérésekre van szükségünk. Azonban az adatgyűjtés technikája változatlan háztartási panelt indokol. A mérés ugyanis a háztartással kötött megállapodás alapján egy, a tévékészülékre szerelt műszer segítségével történik, ami rögzíti, hogy a család melyik tagja melyik csatornát mennyi időn keresztül nézi. A családtagok azonosítása egy speciális távirányító segítségével történik, amelyen minden családtagnak külön gombja van.

[10] A mintavételre vonatkozó részben ismertetett reprezentatív mintavétellel ez a hátrány jórészt kiküszöbölhető.

2.4 Mintavétel

A kutatási módszer kiválasztása után mutatjuk be a minta kiválasztásának folyamatát, de a gyakorlatban a kutatás tervezésénél sokszor párhuzamosan kell gondolkodnunk ezekről, a mintavételi lehetőségeink befolyásolhatják a kutatási módszer kiválasztását. Az empirikus kutatások elsődleges, már a témaválasztásnál megválaszolandó döntési kérdése, hogy kit vizsgáljon, kire/kikre legyenek érvényesek a kutatási eredmények. Mindenekelőtt eldöntendő, hogy a teljes célsokaságot (populációt) vizsgáljuk, vagy annak csak egy részét, azaz mintát veszünk az alapsokaságból.

Határozzuk meg a mintavételhez kapcsolódó legfontosabb fogalmakat.

- **Célsokaság/alapsokaság**: azon személyek, szervezetek stb. összessége, akikre vonatkoztatni akarjuk a kutatás eredményeit, akikre vonatkozóan állításokat akarunk megfogalmazni.
- **Minta**: az alapsokaság egy reprezentatív része. **Reprezentativitás**: a kutatás szempontjából fontos minőségi/mennyiségi jellemzők (kulcsváltozók) mintabeli megoszlása megegyezik az alapsokaságéval.
- **Teljes körű kutatás** során a kutatás mindenkire, a teljes célsokaságra, illetve minden szituációra kiterjed. Ezzel szemben a **mintavétel** a populáció egy meghatározott részsokaságának bevonása a kutatásba. Célja, hogy a mintabeli információk alapján a teljes alapsokaságra tegyünk megállapításokat, vonjunk le következtetéseket.

2.4.1 A mintavétel folyamata

1. a célsokaság meghatározása;
2. a mintavételi keret meghatározása;
3. a mintavételi technika kiválasztása;
4. a mintanagyság meghatározása;
5. a mintavétel kivitelezése.

1. A célsokaság meghatározása

A célsokaságot a kutatási téma önmagában nagymértékben meghatározza, de nem mindig magától értetődő, hogy kire vonatkozzék a kutatás. A téma komplexitása mellett a rendelkezésünkre álló kutatási erőforrások, lehetőségek is befolyásolhatják a kérdést, megtörténhet, hogy nem az eredeti kutatási témánk szerinti célsokaságot választjuk. Ekkor kell eldöntenünk azt is, hogy kisebb

célsokaság esetén teljes körű felmérést végzünk, vagy mintát veszünk a célsokaságból. Teljes körű felmérés esetén nem kell számolnunk a mintavételi véletlen hibával, de így sokkal költségesebb lesz a kutatás.[11]

2. A mintavételi keret meghatározása

A **mintavételi keret** a célsokaság elemeinek megjelenítése, egy lista vagy a sokaság beazonosítását szolgáló irányadás. Például egy város háztartásainak jegyzéke, mobil- vagy vezetékes telefonkönyv, cégadatbázisok stb. Ha nem rendelkezünk a célsokaságra vonatkozó listával, akkor a véletlenszerű kiválasztást próbáljuk biztosítani valamilyen irányadással. Például egy háztartásban azzal a felnőttel készítenek interjút, akinek a születésnapja a legközelebb esik az aznapi dátumhoz, vagy a háztartások véletlenszerű kiválasztására gyakran használt egy magyar származású amerikai statisztikus Leslie Kish algoritmusa.

Ha a mintavételi keretből a sokaság néhány eleme kimarad, ez hibához, a mintavételi keretből eredő hibához vezet. Ilyen esetben a mintavételi keretnek megfelelően újra kell definiálnunk a célsokaságot.

3. A mintavételi technika kiválasztása

Többféle mintavételi technika létezik, könvünkben nyolc módszert két nagy csoportba sorolva mutatjuk be:

I. Valószínűségi mintavételi technikák

Az alapsokaságból véletlenszerűen választjuk ki a mintát. Feltétele, hogy az alapsokaságról legyen nyilvántartásunk, és az alapsokaság sorrendjében ne legyen semmi szisztematikusság.

- **Egyszerű véletlen mintavétel** során az alapsokaság minden egyede egyforma valószínűséggel kerülhet a mintába. Feltételei, hogy a populáció a kutatási szempontból fontos jellemzők szerint bizonyos mértékben homogén[12] legyen, és az alapsokaságról legyen olyan nyilvántartásunk, amiből véletlenszerűen kiválaszthatjuk a minta elemeit.

Példa. A Sapientia hallgatóit tartalmazó adatbázisból véletlenszerűen – például véletlenszám-generátor alkalmazásával – veszünk egy 100-as mintát.

[11] Ezért a nemzeti statisztikai hivatalok általában csak tízévenként végeznek népszámlálást, úgynevezett *cenzust*. A közbeeső években különböző becslésekkel korrigálják az adatokat.

[12] Előzetesen nem definiálható, hogy mennyi az a „bizonyos mértékű" homogenitás. Tökéletes homogenitás, vagyis azonosság esetén nincs mit vizsgálnunk, de nagyfokú heterogenitás, eltérések esetén nem alkalmas az egyszerű véletlen mintavételi technika.

• **Szisztematikus mintavétel**: az ismert alapsokaság valamennyi eleme kap egy sorszámot, majd egy véletlenszerűen kiválasztott kezdőpontból kiindulva minden k-adik elemet betesszük a mintába. A „k"-t mintavételi intervallumnak nevezzük, és úgy határozzuk meg, hogy a populáció elemszámát elosztjuk a tervezett mintaelemszámmal.

Példa. A Sapientia csíkszeredai hallgatóit tartalmazó 828 elemű adatbázisból valamelyik véletlenszerűen kiválasztott hallgatótól kiindulva minden kilencedik hallgatót kiválasztva veszünk egy 100-as mintát.

• **Rétegzett mintavétel**: heterogén alapsokaság esetén az alapsokaságot valamelyik ismérv alapján homogén részsokaságokra bontjuk, majd ezeken belül egyszerű véletlen vagy szisztematikus mintát veszünk. A rétegképző ismérvet a kutatási téma szempontjából fontos ismérvek közül kell kiválasztani, ilyen lehet például a nem, a végzettség, az életkor, az ágazat stb. A rétegek mintabeli aránya meg kell egyezzen az alapsokasági aránnyal.

Példa. Kutatási cél a Hargita megyei cégek HR-menedzsment jellemzőinek a leírása. Mintavétel: a cégek alkalmazottainak száma alapján legalább három réteget (kis-, közép- és nagyvállalatok) képezzünk.

• **Csoportos mintavétel**: ha nincs vagy nehezen kivitelezhető az alapsokasági nyilvántartásunk, de vannak listáink az alapsokaságba tartozó különböző csoportokról. Először el kell készíteni az alapsokasági csoportok listáját, majd a csoportokból mintát venni.

II. Nem valószínűségi mintavételi technikák

Közös jellemzőjük, hogy az alapsokaságból nem véletlenszerűen választjuk ki a mintát.

• **Önkényes:** a válaszadókat önkényesen, a lehető legegyszerűbben választjuk ki. Nem reprezentatív, és ezért csak a feltáró jellegű kutatásoknál ajánlott használni.

• **Szakértői**: a válaszadókat valamilyen szakértői szempontoknak megfelelően választják ki, és szintén nem biztosítja a minta reprezentativitását. Főképp kvalitatív kutatások, fókuszcsoportok mintavételénél alkalmazhatók.

• **Kvótás mintavétel.** Alkalmazásának feltétele, hogy ha nem is rendelkezünk alapsokasági nyilvántartással, akkor ismerjük a kutatás célja szempontjából fontos alapsokasági jellemzők együttes megoszlását.

Például kutatási célunk az ország valamennyi háztartásának tartós fogyasztási cikkek iránti keresletének vizsgálata. Az előzetes tájékozódás nyomán a téma szempontjából fontos kérdésnek tartjuk, hogy a háztartás az ország melyik régiójában és milyen típusú településén található (főváros, 50 ezer fő feletti nagyváros, kisváros, falu). A kvótás mintavételhez tehát tudnunk kell e két ismérv együttes eloszlását, például azt, hogy a teljes alapsokaság hány százaléka él például a központi régió kisvárosaiban.

Első lépésben tehát meghatározzuk a két fontosnak tartott ismérv együttes megoszlását az alapsokaságon belül, majd ezt az arányt a mintanagyság ismeretében leképezzük a mintára.

Példánkat folytatva tételezzük fel, hogy a népszámlálási adatok alapján az összes háztartás 3,2%-a található a központi régió kisvárosaiban, és ha a tervezett mintaméret 1000 háztartás, akkor a mintánkban pontosan 32 ilyen háztartás körében végezzük el a kutatást. Az alapsokaságnak egy ilyen homogén csoportját nevezzük **kvótának**. A kvótából már egyszerű véletlen mintavétellel választjuk ki a mintaelemeket.

- **Hólabda mintavételt** olyankor használhatunk, ha nagyon kevés információval rendelkezünk az alapsokaságról. Egy vagy néhány interjúalannyal kezdjük az interjúkat, majd ők ajánlanak további interjúalanyokat, és ezáltal hólabdaszerűen gördül tovább a mintavétel.

Pl. ha a csíkszeredai könyvvizsgálók körében akarunk kutatást végezni, akkor a hólabda mintavétel nagy valószínűséggel alkalmazható.

4. A mintanagyság meghatározása

A mintavételi folyamat negyedik szakasza a minta nagyságának meghatározása. Ennek pontos meghatározása nemcsak módszertani szempontból, az eredmények érvényessége szempontjából fontos, hanem jelentős kihatással van a kutatás költség- és időigényére is.

A gyakorlatban nagyon különböző mintanagyságokkal találkozhatunk: különböző kísérleteknél, orvosi kutatásoknál a minta nem tudja meghaladni a 20-30-as elemszámot, míg a felső határ általában 1100-1200. A szakmai felkészültség igazolása céljából lefolytatott, az **államvizsga-dolgozathoz kapcsolódó kutatások esetében a 100-as minta ajánlott**, de problémásabb célsokaság, kutatási körülmények miatt még elfogadható a minimum 60-70 elemből álló minta is. Ez alatti mintaelemszám már olyan virtuális kapcsolatokat eredményezhet az

ismérvek (változók) között, amelyek ténylegesen nem léteznek, ezért félrevezetheti az elemzést, tönkreteheti az egész kutatást.
A részletek ismerete nélkül is gyanítjuk, hogy a nagyobb, a célsokaságot minél jobban lefedő minta jobb, pontosabb becsléseket eredményez. Mi határozza meg mégis a minta nagyságát? Elméleti felső korlát természetesen a célsokaság nagysága, a gyakorlatban pedig a legmeghatározóbb tényező a kutatás költségvetése és időigénye. Az említett 1100-1200-as mintanagyság azért felső határ, mert efölé már nem érdemes növelni az elemszámot, a költségek növekedése csak kis mértékű pontosságnövekményt eredményez.

A szakirodalom sokféle, többé-kevésbé bonyolult megközelítését tárgyalja a mintanagyság meghatározásának, mi az alapsokasági statisztikák (pl. valamely változó átlaga vagy aránya) **becslési pontosságából** indulunk ki.[13] A mintavétel nagyságának egyfajta optimalizálása, ha azt a minimális méretű mintát keressük, amelyik még szakmailag elfogadható pontosságú becsléseket eredményez. Ez a megközelítés – akárcsak a többi – tartalmaz egy fontos feltételezést: a különböző ismérvek alapsokasági és mintabeli eloszlása megközelítően normális eloszlású vagy legalábbis egymóduszú kell legyen. A legtöbb ismérvnek szerencsére ilyen eloszlása van, ha pedig ezt ellenőrizni akarjuk, akkor a könyv második felében bemutatott SPSS-program segítségével ezt megtehetjük.

Ha tehát nem az egész alapsokaságot vizsgáljuk, hanem mintát veszünk, akkor a mintavételnek egyenes következménye a mintavételi véletlen hiba. A **mintavételi véletlen hiba** nevével ellentétben nem olyan értelemben hiba, amelyet egy felkészültebb kutató kiküszöbölhet a kutatás folyamán, hanem egzakt módon meghatározható értéket jelent, amit egy alapsokasági érték becslésekor figyelembe kell vennünk.
A mintavételi véletlen hiba legfőbb eleme a **standard hiba,** ami abból adódik, hogy nem az egész alapsokaságot vizsgáljuk, hanem egy reprezentatív mintabeli érték alapján következtetünk az alapsokaságra. Egy alapsokasági ismérv egy statisztikájának (pl. arány) értékét a mintából becsüljük, és a becslés pontosságát fejezi ki a standard hiba.

[13] Gyakorló kutatók vagy módszertant tanuló hallgatók számára nem sok értelme van a szakirodalomban gyakori „feltételezzük, hogy ismert a σ alapsokasági szórás" kezdetű megközelítésnek.

Egy arány standard hibáját a következő képlettel (9.) számoljuk ki:

$$s.e._p = \sqrt{\frac{p \cdot (1-p)}{n-1}} \qquad (9)$$

ahol a p egy mintabeli változó aránya, az n a mintaelemszám.

Alaposan megfigyelve a fenti képletet megállapíthatjuk, hogy nem számol az alapsokaság nagyságával (N). A standard hiba, végső soron a mintavételi hiba nagysága független lenne az alapsokaság nagyságától?! Itt szükséges megkülönböztetnünk a „véges" és a „végtelen" alapsokaságot. A „véges" alapsokaság a mintamérethez viszonyítva viszonylag kis alapsokaságot jelent, míg az ellentéte nem végtelent, hanem viszonylag nagyot, például egy ország lakosainak vagy háztartásainak számát. Pontosítható ezt az arány: **ha a tervezett mintanagyság nem haladja meg az alapsokaság 10%-át (n/N ≤ 0,1), akkor a standard hibát a fenti képlettel számoljuk, ami nem veszi figyelembe az alapsokasági méretet.** Mivel a mintaméret ritkán haladja meg az 1200-1500-as elemszámot, ezért a több mint 12-15 ezer megfigyelési egységből álló – a kutatási téma szempontjából homogén – alapsokaságot már végtelennek tekintjük.

Ellenben ha a minta tervezett nagysága meghaladja az alapsokaság 10%-át, akkor a következő szorzóval kell **korrigálnunk a standard hibát**:

$$s.e._{kp} = s.e._p \cdot = \sqrt{\frac{p \cdot (1-p)}{n-1}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad (10)$$

A mintavételi hiba ismeretéhez a standard hiba alapján kiszámoljuk egy becslés **pontossági szintjét**. A pontossági szint meghatározza, hogy az alapsokasági érték milyen intervallumba esik.

Például a Sapientia csíkszeredai hallgatói körében kívánjuk vizsgálni a hallgatók internetezési szokásait. A kutatás során kiválasztottunk egy 300 fős mintát, ami reprezentatív a 880 fős alapsokaságra. Azt az eredményt kapjuk, hogy a hallgatók 31,5%-a naponta internetezik, és az ehhez az arányhoz tartozó korrigált standard hiba a fenti képlet alapján 2,2. Kutatási eredményünk úgy fogalmazható meg, hogy 95%-os biztonsággal állítható, hogy az összes csíkszeredai hallgató 27,5% és 36,5% közötti aránya naponta internetezik.

Ezt az intervallumot nevezzük **konfidenciaintervallumnak,** és úgy számoljuk ki, hogy a mintabeli statisztikához (átlaghoz, arányhoz) hozzáadjuk, illetve kivonjuk a standard hiba és egy adott megbízhatósági szinthez tartozó érték szorzatát.

$$p_{1,2} = p \pm z \cdot s.e._p \qquad (11)$$

ahol a p egy változó mintabeli aránya, a $p_{1,2}$ a konfidencia intervallum alsó és felső határa, a z egy választott megbízhatósági szinthez tartozó érték, az $s.e._p$- pedig továbbra is a p arányhoz tartozó standard hiba. Láthatjuk, hogy a pontossági szint két változó szorzata ($z \cdot s.e._p$), tehát nagyságát mind a megbízhatósági szint, mind a standard hiba mértéke befolyásolja.

A kutatón múlik, hogy milyen megbízhatósági szintet választ (pl. 90%, 95%, 99% stb.). A képlet alapján beláthatjuk, hogy ha adott konfidenciaintervallumot nagyobb megbízhatósági szinten akarunk meghatározni, akkor csökkentenünk kell a standard hibát, következésképp nagyobb mintára van szükségünk. A legáltalánosabban elfogadott a 95%-os biztonsági szint, aminek a z értéke 1,96. Ajánlott tehát elfogadni ezt a biztonsági szintet, és innentől kezdve konstansnak (1,96) tekinteni ezt a z értéket, ezáltal leegyszerűsítve a mintavételi hiba operacionalizálásának kissé bonyolult folyamatát. A konfidenciaintervallum alsó és felső határát tehát a következő egyszerű képlettel (12.) számoljuk:[14]

$$p_{1,2} = p \pm 1{,}96 \cdot \sqrt{\frac{p \cdot (1-p)}{n-1}} \qquad (12)$$

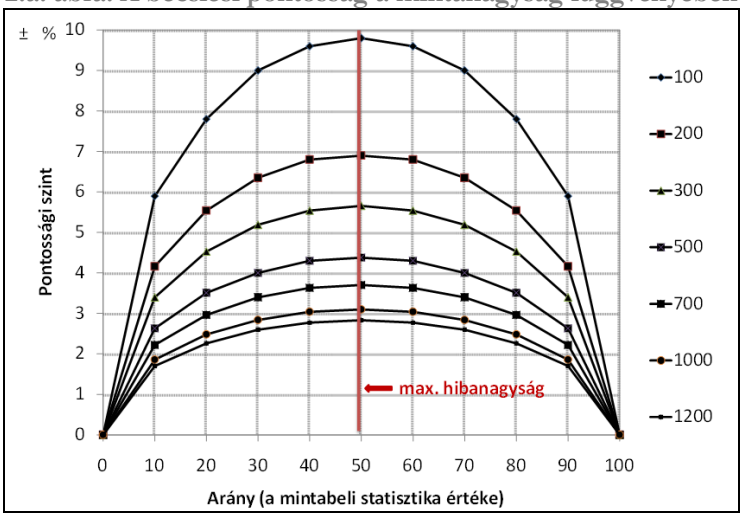
Érdemes ezt a képletet egy Excel-fájlban rögzítenünk, így könnyedén meghatározhatjuk a számunkra elfogadható hibahatárt eredményező mintanagyságot. A fenti képletből természetesen kifejezhetjük közvetlenül is a mintanagyságot (n), de amíg a mintanagyság meghatározására csak egyszer van szükségünk a kutatás során, a fenti képletet sokszor használjuk a részletes kutatási eredmények bemutatásánál.

Itt szükséges pontosítanunk egy korábbi kijelentésünket: nem a mintának van egy egységes hibája, hanem eltérő pontossági szintjei vannak a különböző mintabeli statisztikák becsléseinek. Ezért ha a kutatás rövid bemutatására van szükség, akkor **a minta lehetséges maximális hibájáról beszélünk.**[15] A normális eloszlásnak

[14] Véges alapsokaság esetén a korrigált standard hibát használjuk (10. képlet).
[15] Ilyen megjegyzéseket gyakran hallhatunk, olvashatunk a kutatási eredményeket bemutató médiákban.

köszönhetően akkor maximális a standard hiba (és ez alapján a mintavételi véletlen hiba), ha a mintabeli statisztika (arány) értéke 50%. Például, ha az ország lakosságára vonatkozó közvélemény-kutatás során az 1100 elemből álló minta egyik eredménye szerint a megkérdezettek 50%-a támogatja az államelnököt valamilyen politikai kérdésben, akkor a pontossági szint ± 3%, ha pedig egy másik eredmény 30%, akkor a pontossági szint ±2,7%. Megállapítottuk, hogy „végtelen" alapsokaság esetén, ha egy alapsokasági arányt a mintából kívánjuk becsülni, akkor **a becslés pontosságát a minta nagysága és az arány értéke határozza meg**. A következő (2.a.) ábrán a pontossági szint mértékét e két változó rögzített értékei mellett tüntettük fel:

2.a. ábra. A becslési pontosság a mintanagyság függvényében



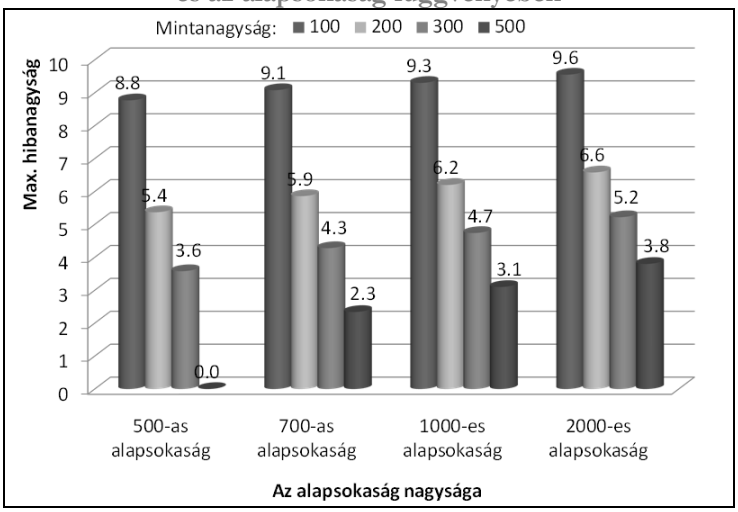
Forrás: saját szerkesztés

A grafikon értelmezését egy példával mutatjuk be: ha egy 500-as elemű mintában egy változó értéke 30%, akkor a becslés pontossága ±4,0%, vagyis az alapsokasági érték 26% és 34% közötti konfidenciaintervallumban vehet fel értéket, de a legvalószínűbb a 30%. Láthatjuk, hogy adott mintanagyság mellett az 50%-os értéknek van a legnagyobb hibája, ezalatt és efölött pedig szimmetrikusan csökken nulláig.
Megállapíthatjuk, hogy a mintaelemszám növekedésével nem lineárisan csökken a max. hibanagyság, és azt is, hogy az 500-as mintaelemszám alatt olyan mértékű lehet a hiba, hogy használhatatlanná teheti az egész kutatást.

„Véges" alapsokaságnál a becslés pontosságát az alapsokaság mérete is befolyásolja. A becslés pontosságát befolyásoló három tényező (az alapsokaság, a

minta nagysága és a becsült statisztika értéke) közül az alábbi ábrán – könnyen belátható okokból – csak az előbbi két dimenziót ábrázoltuk, illetve az 50%-os arányhoz tartozó maximális hibanagyságot.

2.b. ábra. Becslési pontosság a mintanagyság és az alapsokaság függvényében



Forrás: saját szerkesztés

A 2.b. ábrát a következőképp értelmezzük: 500-as alapsokaság (pl. egy falu háztartásainak a száma) és 100-as minta esetén az 50%-os arányhoz tartozó max. hiba ±8,8%. Tehát ha például a 100-as mintában 50 háztartásnak van vezetékestelefon-kapcsolata, akkor a teljes alapsokaságra vonatkozóan megállapíthatjuk, hogy a falu háztartásainak 50%-a (41,2%-58,8%) rendelkezik ilyen kapcsolattal. Ekkora standard hiba és az ebből számolt konfidenciaintervallum láttán arra gondolhatunk, hogy hasunkra ütve sem sokkal kevésbé pontos becslésre lennénk képesek. Ez egy gyakran megfogalmazott vélemény olyanok részéről, akik először szembesülnek a viszonylag kis mintákból számolt becslések pontosságával. A két-háromszáz, vagy annál kevesebb elemből álló minták olyan mértékű mintavételi hibát tartogatnak számunkra, ami első ránézésre jelentősen csökkenti a kutatási eredmények értékét. Ha nem vagyunk tudatában, nem számszerűsítjük pontosan a mintavételi hibát, akkor könnyen előfordulhat, hogy az alapsokasági becsléseink vagy egymásnak, vagy külső, nem a kutatásból származó információknak[16] ellentmondanak. Ezért a mintanagyság

[16] Az ugyanarra a tartalomra vonatkozó összehasonlításra alkalmas információt *benchmark*-nak nevezzük.

meghatározása olyan kérdés, ami befolyásolja a kutatási módszer megválasztását és a kutatási eredmények értelmezését.
Nagy mintavételi hiba esetén a kutatás feltáró jellegét próbáljuk kihasználni, összefüggéseket, okokat keresni és kevésbé a leíró kutatások alapsokasági becsléseire tegyük a hangsúlyt. A megoldás nem az, hogy elutasítjuk vagy megkérdőjelezzük a primer kutatások eredményeit, hanem pontosan számszerűsítve a mintavételi hibát, értelmezzük az eredményeket.

2.4.2 Reprezentatív, kvótás mintavétel a gyakorlatban

Ha sikerült megbirkóznunk a mintavételi hiba és a reprezentativitás fogalmával, akkor időszerű, hogy megnézzünk egy reprezentatív mintavételt egy gyakorlati példán keresztül. A felsorolt nyolc mintavételi technika közül a kvótás mintavételt mutatjuk be részletesen, mivel ez az egyik leggyakrabban használt módszer egyetemi hallgatók kutatásai során. Ezt kiegészítjük az adattábla súlyozásának bemutatásával, egy olyan nem mintavételi módszerrel, ami javíthatja a mintánk reprezentativitását.

Példa. Tegyük fel, hogy egy országosan reprezentatív kutatást szeretnénk a közösségi oldalak mobiltelefonon keresztüli használatával kapcsolatban. Témavezetőnkkel megállapodtunk a kutatási terv több részletében; a tervezett mintanagyság 400 fő lenne, kvótás mintavétel, online kérdezés professzionális kérdőívfelületen.

A mintavétel folyamatában az első lépés **a célsokaság meghatározása**, eldöntöttük, hogy ez Magyarország felnőtt lakossága lesz. Csak az a kérdés, hogy hány éves kortól kerüljenek az interjúalanyok a mintánkba, hiszen már a tizenöt évesek is igencsak aktívak a kutatási témánk szempontjából. A kérdés megválaszolását egy kicsit elodázzuk, amíg hozzájutunk az alapsokasági adatokhoz, vagyis az ország lakosságának életkor szerinti eloszlásához.

Második lépés a **mintavételi keret meghatározása**, de alaposabb megfontolás nélkül is érezzük, hogy az ország valamennyi lakosának címét tartalmazó adatbázis megszerzése utópia lenne. Ilyen esetben a mintavételi keret helyett az alapsokaság főbb jellemzőinek, változóinak az eloszlására van szükségünk. Emlékezünk eddigi tanulmányainkból, hogy – a népszámlálásoknak köszönhetően – elég sok lakossági ismérv aggregált adatai, eloszlásai hozzáférhetőek a statisztikai hivatal honlapján.
Eldöntendő, hogy közülük **melyik változó alapján biztosítsuk a minta reprezentativitását**. Olyan változót válasszunk, ami – a szakirodalom feltárásából származó ismereteink szerint – meghatározza a kutatási témánkat. Választásunk az

életkor változóra esik, mivel tapasztalatunk alapján jelentős eltérés van a közösségi oldalak használatában életkor szerint.

Az *életkor* alapsokasági eloszlásának a vizsgálatánál egy kicsit a bőség hoz zavarba, a KSH oldalán akár demográfiai korfát is találunk, ami az életkor folytonos változóként való figyelembe vételét is lehetővé tenné. Vagyis pl. a 22 évesek mintabeli aránya megegyezne az alapsokaságéval, a 23 éveseké szintén és így tovább. Ezt a lehetőséget azért nem választjuk, mivel valószínűleg a mintabeli életkor változónkban több év nem lenne képviselve, és ez további bonyodalmakat okozna. Szintén a KSH oldalán találjuk az életkor 2016-os mikrocenzus szerinti eloszlását, az életkorcsoportok (10–11, 12–14, 15–19, 20–24,..., 70–74, 75+) gyakorisága abszolút számban van megadva, ráadásul különbontva nemek és iskolai végzettség szerint.

Terveink szerint nem tartoznak kutatásunk célsokaságába a 10–14 évesek, de úgy ítéltük meg, hogy a 15–19 éves kategória már releváns válaszokat tud adni a kutatási témánkkal kapcsolatban. Lakossági célsokaságra irányuló kutatások során gyakran felső korhatárt is meghatároznak, de ezt most nem tartjuk indokoltnak, figyelembe véve a szépkorúak növekvő közösségioldal-használatát. Ellenben az életkor-kategóriák 5 éves terjedelme (pl. 20–24) túl sok kategóriát, kategóriánként pedig kis elemszámot eredményezne a mintánkban, ami már a súlyozásnál is problémát okozhat, de az adatelemzésnél mindenképp. Úgy döntünk, hogy elég részletes lesz a 15–19, 20–29, 30–39, ..., 70+ kategorizálás. Ennek megfelelően kell az eltérő korcsoportok, nemek és iskolai végzettség szerint bontott alapsokasági gyakoriságokat összegeznünk, és kiszámolnunk a relatív gyakoriságot (lásd a 6. táblázat első két oszlopát).

6. táblázat. A kvóták meghatározása az életkor változó eloszlása alapján

életkor kategóriák	alapsokasági relatív gyakoriság	kvóták (egytizedes pontossággal)	kvóták (egész értékkel)
15–19	6.1%	24.4	24
20–29	14.6%	58.4	58
30–39	18.3%	73.2	73
40–49	19.1%	76.4	77
50–59	14.5%	58	58
60–69	15.4%	61.6	62
70–	11.9%	47.6	48
	100.0%	400	400

Forrás: saját szerkesztés

Miután sikerült meghatározni a kutatási tervünknek megfelelő értékekkel (kategóriákkal) rendelkező életkor változó alapsokasági eloszlását, rávetítjük a tervezett mintanagyságra. **Összeszorozva a relatív gyakoriságokat a mintaelemszámmal megkapjuk a kvótákat**, vagyis az adott életkor kategóriába tartozó interjúalanyok számát (lásd 6. táblázat). Mivel értelemszerűen az interjúalanyok száma csak egész érték lehet, ezért a táblázat utolsó oszlopában úgy kerekítjük az értékeket, hogy az összeg, a tervezett mintaelemszám ne módosuljon.
Amennyiben sikerül lefolytatnunk az adatgyűjtést a tervezett kvóták szerint, úgy kijelenthető, hogy a **mintánk reprezentatív életkor szerint az országos 15+ lakosságra.**

Azonban a gyakorlatban ritkán sikerül határidőre (!) valamennyi kvótát „betölteni", azaz pont a kvótának megfelelő interjúalanyt lekérdezni. Példánkban nagyon valószínű, hogy a 70+ kategória 48 fős kvótáját csak nagyon nehezen, lassan tudnánk az online, önkitöltős kérdőívünkkel lekérdezni. Kismértékben ugyan, de szintén torzítja a mintánk reprezentativitását, hogy a tizedes pontossággal kiszámolt kvótákat kerekítenünk kellett.
A reprezentativitás kismértékű torzításának a problémájára is van megoldásunk: az adattábla súlyozása!

2.4.3 A minta súlyozása

A súlyozás során **az adattábla valamennyi változójának értékeit szorozzuk egy előre meghatározott súlyváltozó értékeivel.**

Ha szigorúan értelmezzük a reprezentativitás definícióját, akkor valamennyi változó mintabeli eloszlása meg kell egyezzen, vagy legalábbis nem térhet el szignifikánsan az alapsokasági eloszlástól. Ennek kivitelezése a gyakorlatban szinte lehetetlen, de elméletileg is okafogyottá válna a mintavétel, ha úgyis minden kutatási változónk alapsokasági eloszlását ismerjük. Akkor hogyan lehet elfogadható reprezentativitást elérni?

A kompromisszum abban áll, hogy rosszabb esetben egy, **leggyakrabban két-három változó szerint biztosítjuk a minta reprezentativitását.** Ennél több változó esetén a súlyok meghatározásának komplexitása exponenciális mértékben növekszik, olyan iteratív folyamattá válik, amely több pontján is a kutató szubjektív döntését igényelheti.

A súlyozás folyamata egy változó esetén:

1. Kiválasztjuk a reprezentativitást biztosító változót, és meghatározzuk az alapsokasági eloszlását.
2. Meghatározzuk a reprezentativitást biztosító változó valamennyi értékére a célsokasági és a mintabeli gyakoriságának a hányadosait, a súlyokat.
3. A súlyokat hozzárendeljük az adattáblához; az SPSS adattáblában létrehozzuk a *suly* változót.
4. Aktiváljuk az adattáblában a súly változót, majd ellenőrzés céljából a reprezentativitást biztosító változó súlyozott, mintabeli gyakorisági eloszlását összehasonlítjuk az alapsokasági eloszlással.

Példánkon keresztül nézzük a súlyozás folyamatát. Kvótás mintavétellel, az *életkor* változó szerint terveztük biztosítani a 400 fős mintánk reprezentativitását.

Egyénekre vonatkozó primer kutatásoknál a reprezentativitást biztosító változók legtöbbször *demográfiai* változók (nem, életkor stb.), mivel a legtöbb kutatásban előfordulnak, illetve az alapsokasági eloszlásuk is a népszámlálásoknak, mikrocenzusoknak köszönhetően rendelkezésünkre áll.

1. A súlyozás folyamatának első lépése a reprezentativitást biztosító változó kiválasztása és alapsokasági eloszlásának meghatározása példánkban már megtörtént, lásd a 6. táblázatot. A gyakorlatban a változó kiválasztását attól tesszük függővé, hogy melyik változó alapsokasági eloszlása áll a rendelkezésünkre.
2. Képezzük az *életkor* változó valamennyi értékére (kategóriájára) a célsokasági és a mintabeli gyakoriságának a hányadosait, a súlyokat.

$$súly = célsokasági\ érték\ /\ mintabeli\ érték \qquad (13)$$

Az életkor mintabeli eloszlását, vagyis relatív gyakoriságát az SPSS-ben, a későbbiekben részletezett FREQUENCY paranccsal kapjuk meg.[17] Az adott életkor-kategóriához tartozó célsokasági és mintabeli értékek hányadosaként számítjuk ki a súlyokat.[18]

[17] Példánk jobban megérthető, ha előbb megismerkedünk a könyv második felében bemutatott SPSS-alapokkal.

[18] Ezt legegyszerűbben egy Excel-táblában tehetjük meg.

7. táblázat. A súlyok kiszámítása

életkor-kategóriák	célsokasági eloszlás	mintabeli eloszlás	súlyok
15–19	6.1%	5.2%	**1.18**
20–29	14.6%	17.5%	**0.84**
30–39	18.3%	24.1%	**0.76**
40–49	19.1%	21.5%	**0.89**
50–59	14.5%	12.6%	**1.15**
60–69	15.4%	11.8%	**1.30**
70–	11.9%	7.3%	**1.63**
	100.0%	100.0%	

Forrás: saját szerkesztés

3. A súlyokat hozzárendeljük az adattáblához: **létrehozzuk az SPSS adattáblában a *suly* nevű[19] változót**. A *suly* változó az adattábla valamennyi sorához (esetéhez) hozzárendeli az életkor kategóriának megfelelő súlyértéket.

   Az új változó létrehozására alkalmas COMPUTE parancs leírását a későbbiekben láthatjuk, de itt a kiváló alkalom, hogy megismerkedjünk a haladó SPSS-felhasználók játszóterével, az SPSS Syntax felületével. Az SPSS alapfokú ismerete nélkül ez még korainak tűnik, de a súly változó meghatározását egyszerűbb módon le sem tudnánk írni, mint az SPSS syntax parancssoraival.

   Nyissuk meg a syntax felületet az SPSS-menüből: File–New–Syntax. A súlyváltozó létrehozásához írjuk be a következő, értelemszerű parancssorokat, jelöljük ki egyszerre valamennyit, és futtassuk (RUN-parancs).[20]

```
if korkat=1 suly = 1.18.
if korkat=2 suly = 0.84.
if korkat=3 suly = 0.76.
if korkat=4 suly = 0.89.
if korkat=5 suly = 1.15.
if korkat=6 suly = 1.30.
if korkat=7 suly = 1.63.
execute.
```

[19] Bárhogy nevezhetjük a változót.

[20] A syntax használatának előnyei a menüből választott parancsokkal szemben itt talán még nem egyértelműek, de majd az lesz, ha nem csak egy változó szerint biztosítanánk a reprezentativitást.

4. **Aktiváljuk az adattáblában a súlyt**, és ellenőrzés céljából a reprezentativitást biztosító változó súlyozott mintabeli gyakorisági eloszlását összehasonlítjuk az alapsokasági eloszlással. A létrehozott *suly* nevű változónk aktiválásához az SPSS adattáblamenüjében válasszuk a Data menüből a WEIGHT CASES parancsot, majd a felugró ablakban a Weight cases by utasítást kiválasztva a bal oldali mezőből vigyük át a *suly* változót a jobb oldali, Frequency Variable mezőbe. Mindaddig súlyozott lesz az adattáblánk, amíg az előbbi ablakban a Do not weight cases opciót nem választjuk. A súlyozás sikerességének ellenőrzéséhez kérjük az életkor változónk (*korkat*) gyakorisági eloszlását (FREQUENCY), és összehasonlítjuk az alapsokasági eloszlással.

Megállapíthatjuk, hogy az egy változó szerinti súlyozás nem is annyira bonyolult, de azért figyelnünk kell a súlyozás lehetséges **problémáira**:

- A súlyváltozónak ne legyenek túl nagy vagy túl kicsi értékei. Egy 3-as értékű súly azt jelenti, hogy az adott interjúalany(ok) valamennyi kérdésre adott válaszait 3-szorosan vesszük figyelembe. Viszonylag széles körben elfogadott gyakorlati hüvelykujjszabály, hogy a **súlyváltozó elfogadható, ha a súlyok 90%-a 0,7 és 1,3 közötti értéket vesznek fel**. Ha nem, akkor a reprezentativitást biztosító változó kiugró értékeit tartalmazó eseteket töröljük az adattáblából, amennyiben a mintaelemszám megengedi. Rosszul kivitelezett mintavétel kijavítására a súlyozás nem alkalmas.
- Komplexebb statisztikai modellek és tesztek a súlyozás egyedi megoldásait igényelhetik, ezt az adott módszer alkalmazásánál vegyük figyelembe.[21]

Miután láttuk a súlyozás folyamatát és logikáját egy változón keresztül, megállapíthatjuk, hogy nagyfokú egyszerűsítés mindössze egy változó szerint „biztosítani" a minta reprezentativitását. Márpedig egy üzleti célú kutatásnál a megbízó nem a kompromisszum-készségünkért fizet.

2.4.4 Több változó együttes eloszlása szerinti súlyozás

A több változó szerinti reprezentativitás biztosítását három változó esetére példázzuk, ez alapján könnyen megvalósíthatjuk a kétváltozó szerinti súlyozást.

[21] A könyvben nem tárgyalunk ilyen módszereket.

Példa. A mintánkat az interjúalanyok *életkora* mellett *iskolai végzettség* és *nem* szerint is tegyük reprezentatívvá, mivel rendelkezésünkre állnak alapsokasági eloszlásaik (8. táblázat).

8. táblázat. A 15+ életkorú lakosság alapsokasági eloszlása nem, életkor és iskolai végzettség szerint

férfi	**alap-fok**	**közép fok**	**felső-fok**			**nő**	**alap-fok**	**közép fok**	**felső-fok**	
15–19	8.4%	0.6%	-	9.1%		**15–19**	3.0%	0.6%	-	3.6%
20–29	4.8%	6.4%	1.2%	12.4%		**20–29**	7.6%	5.9%	3.0%	16.6%
30–39	10.6%	6.4%	2.9%	19.8%		**30–39**	8.8%	6.5%	1.8%	17.1%
40–49	11.6%	5.6%	2.4%	19.6%		**40–49**	9.7%	6.3%	2.7%	18.7%
50–59	8.8%	3.8%	1.6%	14.3%		**50–59**	8.1%	4.7%	1.9%	14.7%
60–69	8.8%	3.9%	1.7%	14.4%		**60–69**	9.5%	5.0%	1.6%	16.1%
70–	6.2%	2.8%	1.4%	10.4%		**70–**	8.8%	3.3%	1.1%	13.2%
	59.3%	29.5%	11.2%	100%			55.5%	32.4%	12.1%	100%

Forrás: saját szerkesztés

A súlyok meghatározásához ugyanezt a táblázatot elő kell állítanunk az adattáblánkban is. Kereszttábla-elemzés[22] segítségével (CROSSTABS), a sorba az *életkor* változót, oszlopba a *végzettséget* és a Layer mezőbe a *nem* változót beállítva. A Cells ablakban, ha csak a Total percentages-t kérjük, akkor a kereszttáblánk szerkezete teljesen megegyezik az alapsokasági eloszlások táblázatával.

Kiterjedt táblázatunkban a *nem*, az *életkor* és az *iskolai végzettség* együttes eloszlása összesen 2x3x7=42 cellában jelenik meg, tehát ennyi értéke lesz a súlyváltozónknak. A súlyokat az előbbiekben bemutatott módon számoljuk ki, az alapsokasági és a mintabeli együttes eloszlások (a 8. táblázatban a szürke háttérrel kiemelt cellák) hányadosaként. Egymás mellé rendezett kereszttáblák esetén ezt Excel-ben, a képletek másolásával könnyen megtehetjük.

A súlyok hozzárendelése az adattáblához. A súlyváltozót meghatározó SPSS syntax most egy kicsit összetettebb és 42 sorból áll, ezért az alábbiakban nem jelentetjük meg valamennyit.[23]

[22] Lásd a Kereszttábla-elemzés alfejezetet.

[23] Itt már nyilvánvaló az SPSS syntax használatának előnye, nem kell 42-szer a menüből választanunk a COMPUTE parancsot.

```
if (nem=1 and korkat=1 and isk=1) suly = 1.08.
if (nem=1 and korkat=1 and isk=2) suly = 0.87.
...
if (nem=2 and korkat=7 and isk=2) suly = 0.95.
if (nem=2 and korkat=7 and isk=3) suly = 0.92.
```

Nincs más hátra, mint aktiválni a súly változót az adattáblában (WEIGHT CASES) az egyváltozós súlyozásnál bemutatott módon, és a súlyozott adattáblában lefuttatni a három változó gyakorisági eloszlását (FREQUENCY), hogy leellenőrizhessük az alapsokasági eloszlással való egyezőségüket.

Hárm változó szerinti súlyozás esetén nagy valószínűséggel előfordul, hogy a mintánkban néhány cellában nincsen érték, amihez súlyt rendelhetünk. Ilyen esetben az üres cella gyakoriságát „arányosan osszuk szét" a többi cella között.

2.4.5 Peremeloszlások szerinti iteratív (RIM) súlyozás

Többváltozós súlyozás esetén gyakori helyzet, hogy **nem ismerjük a változók közötti együttes eloszlást, csak a változók egyenkénti eloszlását, az úgynevezett peremeloszlást.**

Ebben a részben azt vizsgáljuk, hogyan biztosíthatjuk a reprezentativitást a peremgyakoriságokkal, és ez miben különbözik az eddig tanult, együttes eloszlás szerinti súlyozástól. Előbb nézzük, hogy mi válthatja ki ezt a helyzetet. Az online kérdőívek gyorsan növekvő használatát sokszor nem követi a minta reprezentativitását biztosító technikák ismerete és alkalmazása. A szükséges tudás hiányán túl, ennek oka lehet a megfelelő alapsokasági statisztikák hiánya. Kisebb kutatások, államvizsga-dolgozathoz szükséges kutatások ritkán vonatkoznak az egész országra, a többé-kevésbé pontosan meghatározott földrajzi érvényesség területére (pl. Székelyföld) nem találunk demográfiai alapadatokat sem.

Az alapsokasági eloszlások hiánya adódhat abból is, hogy nem demográfiai, hanem valamely, **a kutatási témához tartozó változók** (pl. a mobiltelefon-használati jellemzők) alapján szeretnénk biztosítani a reprezentativitást. Könnyen belátható, hogy egy ilyen „tematikus" változó szerinti reprezentatitvitás sokkal relevánsabb a kutatásunk számára, mint egy demográfiai változó szerinti.
A kutatási problémából levezetett legfontosabb változók alapsokasági eloszlása biztosan nem áll rendelkezésünkre – hiszen épp ezért szeretnénk kutatni az adott témát –, de korábbi kutatási eredményekből találhattunk a reprezentativitás biztosításához figyelembe vehető adatokat. Ezek hiányában, kevésbé kutatott

témánál vagy a szakdolgozati kutatás során elfogadható a demográfiai változók szerinti reprezentativitás.
Akkor van szükség tehát a peremeloszlások szerinti súlyozásra (RIM), ha nem állnak rendelkezésünkre a reprezentativitást biztosító változók együttes eloszlásai. Több változó egymástól független eloszlásai szerinti súlyozás során **iteratív eljárással**, a súlyozási folyamatot addig ismételjük, amíg valamennyi, a reprezentativitást biztosító változó mintabeli eloszlása a **lehető legjobban illeszkedik** az alapsokasági eloszlásra.

Példa. Példánkban ez azt jelentené, hogy ismerjük külön-külön a *nem*, az *életkor* és az *iskolai végzettség* célsokasági eloszlásait, de az együttes eloszlást nem. A 8. táblázatban a szürkével kiemelt cellák jelentik az együttes eloszlást, a táblázat szélein az összesítő sor és oszlop mutatnák a peremeloszlást, de most feltételezzük, hogy csak a peremeloszlást ismerjük.

Kezdjük a súlyozás folyamatát:

1. A *nem*, az *életkor* és az *iskolai végzettség* változók célsokasági eloszlását gondosan átmásoljuk az adatforrásunkból (pl. KSH honlap) egy Excel-táblába.

2. A mintabeli eloszlások meghatározásához az SPSS-adattáblánkban gyakorisági eloszlást (FREQUENCY-t) futtatunk a *nem*, az *életkor* és az *iskolai végzettség* változókra. Bemásoljuk az Excel-táblába, majd az alapsokasági és a mintabeli eloszlások hányadosaiként kapjuk meg a súlyokat mindhárom változóra.

3. A súlyokat hozzárendeljük az adattáblához, egymás után mindhárom változó esetében. Menüből COMPUTE-paranccsal vagy syntax-szal az alábbi módon:

```
if nem=1 suly = 1.16.
if nem=2 suly = 0.84.
if korkat=1 suly = 1.07.
if korkat=2 suly = 0.96.
...
if korkat=7 suly = 0.75.
if isk=1 suly = 0.84.
if isk=2 suly = 1.24.
if isk=3 suly = 0.92.
```

4. Aktiváljuk a *suly* változót (WEIGHT CASES parancs), majd a súlyozott adattáblán gyakoriságot (FREQUENCY) futtatunk a három változóra.

Összehasonlítjuk a három változó súlyozott mintabeli gyakorisági eloszlását a célsokasági eloszlással, és megdöbbenve kérdezzük, hogy miért nem egyeznek. A válasz és a nagy különbség az előzőekben tárgyalt, az együttes eloszlásra alapozó

súlyozási módszerrel szemben a 3. pontban, a súlyok hozzárendelésében áll. Alapos, nagyon alapos olvasóink észrevehették, hogy **a súlyok „átfedik" egymást**. A két nem súlyainak hozzárendelése után a súlyokat felülírjuk az életkor kategóriák súlyaival, majd ezeket az iskolai végzettség súlyaival. Az együttes eloszlások szerinti súlyozásnál külön súllyal rendelkeztünk a három változó mind a 72 kombinációjára, de ez most nem áll rendelkezésünkre.
Itt jön a képbe az **iteráció, a súlyozási folyamatot (2–4 szakaszait) addig ismételjük, amíg mindhárom súlyváltozó mintabeli eloszlása hasonló lesz**. Teljesen egyező egyik változó szerint sem lesz, de megközelítően jó eloszlást elérhetünk mindhárom változóra.

Új súlyokat kell tehát képeznünk, és beállítanunk a már súlyozott mintabeli gyakoriságok és a célsokasági gyakoriságok hányadosaiként, majd újra ellenőriznünk az új súlyok szerinti gyakoriságokat.
Meddig folytatjuk ezt az iterációt? Az előzőekben említettük, hogy a súlyváltozó elfogadható, ha a súlyok 90%-a 0,7 és 1,3 közötti értéket vesznek fel, de ez a kritérium arra vonatkozik, hogy lehetőleg ne legyenek túl nagy (és kicsi) súlyok, ami egyes esetek (interjúalanyok) szerepét aránytalanul felnagyítaná (vagy csökkentené). Ezúttal olyan kritériumra van szükségünk, amely alapján ki tudjuk jelenteni, hogy a célsokasági és a mintabeli eloszlások kisebb eltérései ellenére a mintánk elfogadhatóan reprezentatív. Ilyen általános érvényű, analitikusan levezetett kritérium nincs, **nekünk kell eldöntenünk, hogy hány százalékpontos eltéréseket fogadunk el a két eloszlás értékei között**.
Legfontosabb, hogy a tanulmányunkban részletesen írjuk le a mintánk reprezentativitására vonatkozó információkat, szem előtt tartva, hogy tökéletesen reprezentatív minta nem létezik.

A peremeloszlások szerinti súlyozás (RIM) fáradságos és időigényes folyamat, ezért a tapasztalt kutatók, kutatócégek (pl. piackutatók) programokat használnak a súlyozási algoritmusok lefuttatására. A RIM súlyozó programok általában a következő jellemzőkkel rendelkeznek:

- több, akár 5-6 változó együttes eloszlása szerint biztosítja a reprezentativitást;
- meghatározható a súlyozott minta és a célváltozók eloszlása közötti elfogadható eltérés mértéke. Pl. ha két változó közötti együttes céleloszlás 30%, akkor 0,1%-os toleranciaszint mellett a két változó súlyozott mintabeli eloszlása 29,9% – 30,1% között lehet;

- képes kell legyen a súlyok automatikus újraszámolására, ha új esetek kerülnek az adattáblába, és ezáltal megváltoznak a változók súlyozatlan eloszlásai.

Miután végigdolgoztuk magunk a súlyozás folyamatán, kijelenthetjük, hogy a legjobb, ha nincs szükségünk súlyozásra, hanem már a mintavétellel biztosítjuk a reprezentativitást. Mint említettük, rosszul kivitelezett mintavétel kijavítására a súlyozás nem alkalmas, de kisebb korrekciókra igen.

2.4 Kérdőívszerkesztés

A kvantitatív kutatás kérdőívének szerkesztését még az adatgyűjtés elkezdése előtt véglegesítjük, és ezt a kutatás során nem szabad módosítanunk. Ha kifelejtjük valamelyik kutatási hipotézisünk teszteléséhez szükséges kérdés(eke)t, vagy nem megfelelő mérési skálákat használtunk, akkor a kérdőívszerkesztés befejezése után már nincs alkalmunk módosítani, az adott részeredményekről le kell mondanunk, szerencsétlen esetben hiábavaló lesz az egész kutatásunk. Ezért a kérdőívszerkesztés a kutatási folyamat kiemelt fontosságú része. A következőkben csak a kvantitatív kutatás kérdőívéről lesz szó, nem értjük ide a mélyinterjú strukturálatlan kérdőívét és a fókuszcsoport moderátori vezérfonalát.
A kérdőív jellemzői:

- A kérdőív **strukturált**, tehát a kutató által meghatározott szerkezetben vizsgálja az interjúalany véleményét.
- A **kérdések logikusan követik egymást**, tartalom szerint **témacsoportokba** sorolva, és ezek a témacsoportok is logikusan tagolják a kérdőívet.
- **Bemutatkozó szöveggel** kezdődik, amelyben a kérdezőbiztos bemutatkozik, majd közli az interjúalannyal az interjú célját és várható időtartamát.
- A **kérdéseknek a lehető legegyszerűbbeknek**, közérthetőnek kell lennie, szakkifejezéseket, homályosságot lehetőleg el kell kerülni. A kérdőívet szerkesztő kutató nem tévesztheti szem elől, hogy nem a témáról alkotott szakvéleményét kell közölnie az interjúalannyal, hanem az interjúalany véleményét mérni, rögzíteni.
- Az interjúalany **válaszait véglegesnek** kell tekinteni, hibák vagy hiányosságok esetén csak nagyon ritkán adódik az interjú kiegészítésére lehetőség.
- A kérdőívnek a lehető **legrövidebbnek** kell lennie, nem tartalmazhat ismétléseket, redundanciát.
- **Egységes** valamennyi interjúalany számára, nem módosíthatjuk az adatgyűjtés során.
- A kérdőív változhat a válaszok alapján. Bizonyos **szűrőkérdésekre** adott válaszok kizárhatják a valamilyen feltételnek nem megfelelő válaszadókat egészen vagy részlegesen az interjúból. A kiszűrt interjúalanyokat a megfelelő kérdőívrészre irányító utasítást nevezzük **ugrásnak.**

- A kérdezés során nincs lehetőségünk arra, hogy a válaszoló feleleteit **megfigyelési adatokkal kiegészítsük**, ilyen adatigény esetén a kvantitatív kutatásunkat ki kell egészíteni szekunder adatgyűjtéssel.

A kérdőívszerkesztést jelentősen befolyásolja, a **kérdőív típusa**. Alapvető különbség van a típusok között a kérdőív **kitöltőjének személye** szerint:

- Segített kérdőív, az interjúkészítő (kérdezőbiztos) által kitöltött kérdőív – az interjút egy kérdezőbiztos vezeti, ő olvassa a kérdéseket, és szükség szerint segít az értelmezésükben, illetve rögzíti a válaszokat.
- Önkitöltő kérdőív – csak képletesen beszélhetünk „interjúról", mert az interjúalany egymaga olvassa, értelmezi, és jó esetben megválaszolja a kérdéseket.

A segített kérdőíveken belül a kvantitatív kutatások **kérdezési mód szerinti osztályozása** meghatározza a kérdőívet is:

1. Személyes kérdezés kérdőíve. Mivel a személyes kérdezés során maximális módon adva van a közvetlen párbeszéd, a kérdezés során felmerülő problémák megoldásának lehetősége, ezért a személyes kérdőív a többi típusnál hosszabb lehet, és bonyolultabb, személyesebb témákat tartalmazhat. Az interjú helyszíne (otthoni, utcai, bevásárlóközponti) is kisebb-nagyobb mértékben meghatározhatja a kérdőív jellemzőit.
A kérdőívszerkesztést a helyszínnél jobban befolyásolja, hogy a kérdőív **nyomtatott** formában van vagy számítógéppel támogatott kérdezés (CAPI) során egy kisebb laptopon le van **programozva**. Fontos különbség, hogy a kérdőív a számítógépen nem dokumentumként, hanem egy speciális szoftver segítségével programozva található.

2. Telefonos kérdezés során az interjúalanyt otthonában, üzleti kutatásoknál munkahelyén vagy mobiltelefonon hívja a kérdezőbiztos.

3. Internetes kérdezés alatt weblapon elhelyezett kérdőív kitöltését értjük. Azért soroljuk a segített kérdezések közé, mivel a programozás lehetővé teszi, hogy hibás válasz esetén figyelmeztessük az interjúalanyt, illetve a kérdések közötti logikai kapcsolatokat, ugrásokat automatikusan követi.

Az **önkitöltő kérdőíveken** belül az adathordozó alapján megkülönböztethetjük a **postai úton** eljuttatott nyomtatott, és az elektronikus postán, **e-mail**-en keresztül eljuttatott elektronikus kérdőíveket. Kérdőívszerkesztés szempontjából nincs nagy jelentősége a megkülönböztetésnek (a terepmunka szervezése szempontjából viszont igen), mindkét típus ugyanazokkal az előnyökkel, de főképp korlátokkal rendelkezik. A kérdezőbiztos segítsége és az általa nyújtott motiváció nélkül kitöltendő kérdőívek csak rövidek lehetnek és viszonylag egyszerű, egyértelmű témákat érinthetnek.

A kérdőív **szerkesztésének** lépései:

- Meg kell határozni a szükséges információk körét – a kutatási céloknak megfelelő kérdések meghatározása és felsorolása. Ezt követően a kérdéseket logikailag, tartalmilag összetartozó csoportokba, blokkokba rendezzük.
- Meg kell határozni a kérdezés módját és a kérdőív típusát – ideális esetben a pénz, idő vagy egyéb erőforráskorlátok nem determinálják a kérdezés módját, hanem a témának leginkább megfelelő kérdezési módot választhatjuk.
- Kérdések tematizálása, a témacsoportok sorrendjének kialakítása.
- Kérdések formába öntése: szövegezés, skálák, táblázatok, kártyafüzet.
- Formai és tartalmi ellenőrzés. A formai ellenőrzés során jó, ha nem a kérdőív készítője ellenőrzi a kérdések sorszámait, a zárt kérdések válaszlehetőségeinek kódjait, a logikai ugrások helyességét, a NT/NV válaszlehetőségek meglétét.
- Próbakérdezés (pilot interjú), véglegesítés – egy próbakérdezés során teszteljük, lekérdezzük a kérdőívet, lehetőleg a célcsoportba tartozó személlyel folytatva az interjút.

Kérdéstípusok

A kérdőívszerkesztés alapelveinek és folyamatának bemutatása után vizsgáljuk meg közelebbről a „kérdést". A jó kérdés jellemzői:

- nem fölöslegesen részletező vagy túlságosan specifikus, de nem is túl általános;
- nem sugalmazó;
- nem haladja meg a válaszadók elvárható tudását, tapasztalatait;

- nem vonatkozik direkt módon jövedelemre, fizetésre;
- hangneme udvarias, semleges, ne legyen vizsgáztató jellegű;
- közérthető legyen, ne tartalmazzon homályos, idegen kifejezéseket vagy sztereotípiákat;
- egy kérdéssel mindig csak egy dologra kérdezzünk;
- a lehető legrövidebb legyen, de ez ne rontsa az érthetőségét;
- és mindenekelőtt legyen indokolt, ne kérdezzünk feleslegesen.

A kérdéseket osztályozhatjuk a funkciójuk és a válaszlehetőségek szerint.

I. A funkciójuk szerint:

- Fő kérdések – közvetlenül a kutatási témára vonatkoznak.
- Kiegészítő kérdések – segítik a fő kérdésekből származó eredmények értelmezését, illetve a kérdőív elfogadhatóságát (pl. demográfiai adatokra vonatkozó kérdések, felvezető kérdések).

II. A válaszlehetőségek szerint:

- nyitott kérdés: szabad válaszkifejtés, a válaszlehetőségek nincsenek meghatározva (mit gondol róla?, mi a véleménye? stb.);
- egyválaszos zárt kérdés – egy válasz lehetséges az előre meghatározott lehetőségek közül;
- igen-nem válaszlehetőség;
- egyéb válaszlehetőség;
- skálán jelezhető a válaszerősség;
- többválaszos zárt kérdés - több válaszlehetőség közül egynél többet is választhat a kérdezett.

A nyitott-zárt kérdések aránya egy kérdőíven belül általában 30-70% és 20-80% intervallumok között mozog.

Valamennyi kérdésnél fontos, hogy a **válaszmegtagadást** is rögzítsük. A válaszmegtagadásnak két alapvető oka lehet; az interjúalany nem tud vagy nem akar válaszolni. Ezt a két lehetőséget csak nagyon ritkán indokolt megkülönböztetni, ezért a válaszmegtagadást általában **NT/NV** rövidítéssel jelöljük. Ha ezt nem jelölnék, akkor az adatelemzés során nem tudnák megkülönböztetni, hogy adott kérdés nem volt kérdezve az interjúalanytól, vagy valamilyen okból megtagadta a választ. Ennek jelentősége leginkább az adattisztítás, a kérdőív és az adattábla összhangjának az ellenőrzésénél van.

Széleskörűen elfogadott gyakorlat tehát, hogy valamennyi kérdés esetében van NT/NV válaszlehetőség, amit általában 9-essel kódolnak. Előfordulnak olyan próbálkozások, hogy a kutató szándékosan nem ír a kérdőívbe ilyen válaszlehetőséget, hogy kvázi kikényszerítse a választ az interjúalanyból. Ez a trükk azonban csak nagyon nyitott, segítőkész és a kutatási témát jól ismerő interjúalanyok körében működik. Rosszabb esetben kitöltetlen kérdőívet vagy irreleváns válaszokat kapunk.

A kérdésekre adott válaszokat a következő négy típusú **mérési skála** valamelyikének segítségével számszerűsíthetjük.

1. Nominális skála. Értékeinek nincs számszerű jelentésük, hanem kódként szerepelnek, és a változó értékeinek, kategóriáinak a jelentését a kódokhoz rendelt címkék (*label*-ek) adják.
Például: a cég jogi formája (1 – egyéni és családi vállalkozás; 2 – kft.; 3 – rt.; 4 – egyéb), a fogyasztott üdítőital márkája (1 – Coca-Cola; 2 – Pepsi-Cola; 3 – egyéb), az interjúalany neme (1 – nő; 2 – férfi) .

2. Sorrendi skála. A sorrendi skálaértékeknek sincs számszerű jelentésük, jelentésüket a címkék (*label*-ek) adják. Az különbözteti meg a nominális skálától, hogy a skálaértékek között rangsorbeli különbség van. Például a település típusa (1 – főváros; 2–100 ezer lakosúnál nagyobb város; 3-50-100 lakosú város; 4 – kisváros; 5 – falu, község), a fogyasztói elégedettség mértéke (1 – rossz; 2 – közepes; 3 – jó), a gazdaság fejlődése a következő évben (1 – csökkenés; 2 – stagnálás; 3 – növekedés).

3. Intervallumskála. Az intervallumskála értéke egy szám, amelynek önmagában jelentése van, nincs szükség címkére az értelmezéséhez. A skálaértékek kezdőpontja viszonylagos, nem egyértelmű, ezért a skála értékei összegezhetők, kivonhatók, de nem képezhetünk arányokat.
Legjellemzőbb példa a hőmérséklet Celsius-fokban kifejezve, ahol a 0 fok megegyezéses (nem ugyanannyi, mint Fahrenheitben), és nem mondhatjuk, hogy az 5 fokos „meleg" ötször melegebb, mint az 1 fokos.

4. Arányskála. Az arányskála értékei is számok, továbbá az különbözteti meg az intervallumskálától, hogy az értékekkel bármilyen numerikus műveletet végezhetünk.
Például az életkor években kifejezve elmondható, hogy egy 25 éves ötször idősebb, mint egy 5 éves. További példák: a cég árbevétele, az előállított termékek

darabszáma, a szolgáltatással kapcsolatos elégedettség mértéke 1-től 5-ig terjedő skálán stb.

A fenti skálákat a *mérési szint* növekvő sorrendjében tüntettük fel. Ez azt jelenti, hogy adatelemzési szempontból például az arányskála magasabb rendű, mint a nominális skála.[24] Ennek egyik jó példája, hogy *skálatranszformációval* a magasabb rendű skálákból képezhetünk alacsonyabb rendűt, de fordítva nem. Például a pontos évvel kifejezett életkorból (arányskála) képezhetünk életkor-kategóriákat (nominális skála): 1–18, 19–25, 26–35, 36–45, 46–60 és 60 felett.

A gyakorlat-orientáltságunknak megfelelően a következőkben bemutatjuk a kérdéstípusokat egy professzionális kérdőívszerkesztő, a *QuestionPro* segítségével. A 3. ábrán a kérdőívszerkesztő alap *(Basic)* és haladó szintű *(Advanced)* kérdéstípusait láthatjuk, amelyek közül a teljesség igénye nélkül kiemeljük a leggyakrabban használt, illetve a legérdekesebb kérdéstípusokat.

A Multiple Choice menüben a leggyakrabban használt kérdéseket találjuk, amelyeknél több válaszlehetőség közül választhatunk ki egyet *(Select One)* vagy többet *(Select Many)*. A Drop-down Menu-vel egy legördülő menüben kínálhatjuk fel a válaszlehetőségeket. A nyitott kérdések válaszait a Text opcióval, egy vagy több sorban is megkaphatjuk.
A Graphical Rating opcióival az ordinális skálájú válaszlehetőségeket értelmezését segíthetjük ikonokkal, képekkel. A Text Slider-rel az ordinális változó értékei külön kategóriák, amelyekhez címkét is rendelhetünk, a Numeric Slider már átvezet a numerikus skálákhoz, 1-től 100 pontig számszerűsíthetjük a véleményünket a skála két végpontja között.
Az Image Chooser kérdéstípusai nem különböznek kutatásmódszertani szempontból az eddig megismert egy- és többválaszos kérdésektől és ordinális, értékelési skálától, hanem képek segítségével teszik értelmezhetőbbé a kérdést és a válaszlehetőségeket.
A Rank Order opcióval sorba rendezzük a válaszlehetőségeket, a Constant Sum kérdéstípussal arra kérjük az interjúalanyt, hogy 100 pontot osszon szét az alternatívák között, azok fontossága szerint. A kérdőívszerkesztő természetesen ellenőrzi, hogy a szétosztott összeg pont 100 legyen. A Drag and Drop annyiban

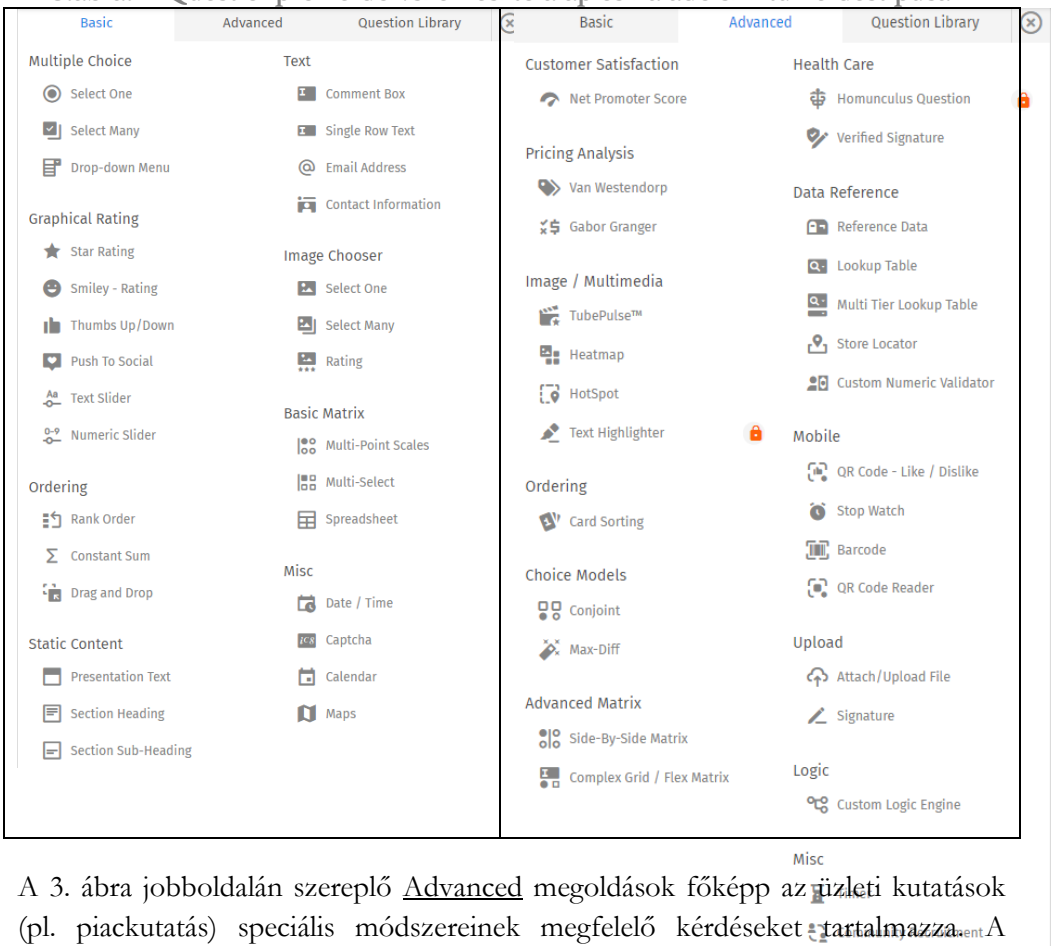
[24] Megjegyzendő, hogy a különböző nem lineáris és nem paraméteres módszerek kifejlesztésével és elterjedésével ez a megközelítés egyre inkább idejét múlt, egy nominális skála is lehet ugyanolyan „hasznos", mint egy arányskála.

különbözik a Rank Order-től, hogy nem sorszámot adunk a válaszlehetőségeknek, hanem „fizikailag" megragadva rendezzük sorba.

A Basic Matrix kérdéseivel egyszerre több szempontot értékelhetünk ordinális skálával a Multi-Point Scales opcióval szempontonként (soronként) egy válasz, a Multi Select-tel több válasz lehetséges. A Spreadsheet-tel szöveges válaszokat vihetünk be a sorok és oszlopok által meghatározott mátrix-cellákba, olyan ordinális értékelésnél, amikor a számszerűsítést nem találjuk megfelelőnek.

A Static Content-tel nem kérdéseket, hanem magyarázó, kérdéshez nem közvetlenül kapcsolódó, „statikus" szövegeket helyezhetünk el a kérdőívben. A Misc vegyes kínálatában dátumformátumot, ellenőrző *captcha*-t vagy térkép alapú kérdéseket találunk.

3.ábra. A Questionpro kérdőívszerkesztő alap és haladó szintű kérdéstípusai



A 3. ábra jobboldalán szereplő Advanced megoldások főképp az üzleti kutatások (pl. piackutatás) speciális módszereinek megfelelő kérdéseket tartalmazza. A Customer Satisfaction megfelel egy ordinális skálának (pl. Text Slider), de a Pricing Analysis menüben például az árkutatások egyik gyakran használt módszerét, a Van

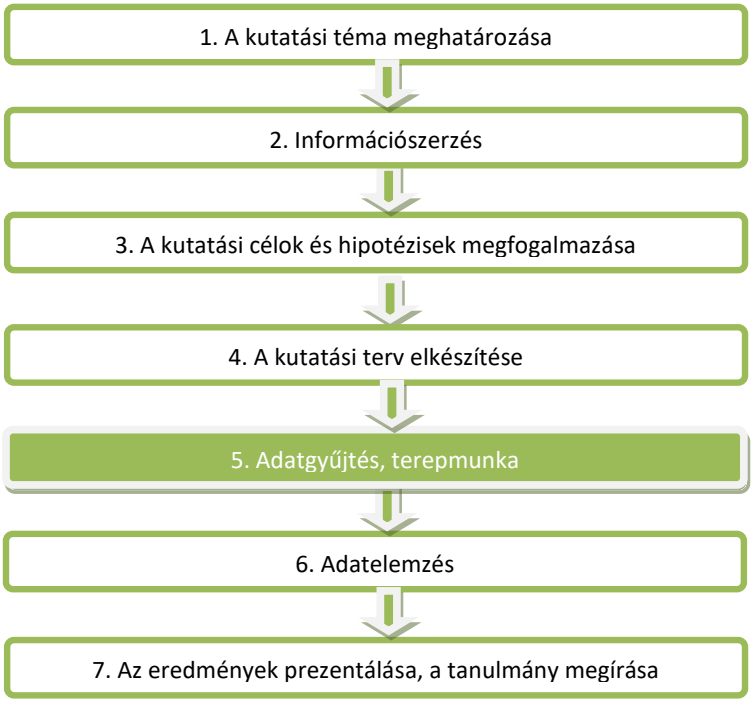
Westendorp modell meghatározott kérdéseit találjuk. Ugyancsak gyakran használtak a Conjoint modellek, amelyekkel egy termék árát annak még néhány jellemzőjével együtt (pl. márka, minőség, funkcionális jellemzők stb.) vizsgálhatjuk. Az interjúalany, a potenciális fogyasztó a számára fontos termékjellemzők kombinációi közül választja a számára legkedvezőbbet és megfizethetőt.
Ezeket a modelleket itt nem mutatjuk be részletesen, általánosabban használhatjuk az Advanced Matrix ordinális skáláit, amelyekkel egyszerre látja az interjúalany a több szempont szerinti értékeléseit, így jobban viszonyítja tudja egymáshoz az ordinális értékeket.

Az online, professzionális kérdőív előnyei:

- A kérdezés témáját, a kérdést szükség esetén értelmezhetőbbé tehetjük **egy kis film bemutatásával** vagy más multimédiás megoldással. A hagyományos, nyomtatott kérdőívet használó személyes kutatásoknál hasonló funkciót tölt be a kérdőív mellékleteként szereplő kártyafüzet, amelyben képeket lehet mutatni az interjúalanynak.
- A válaszlehetőségek közötti választást befolyásolhatja azok **sorrendje**. Például ha az interjúalanyt arra kérjük, hogy egy kérdésnél tíz válaszlehetőség közül válassza ki a három legfontosabbat. Főképp szubjektív (pl. attitűd) kérdéseknél lényeges a válaszlehetőségek sorrendje. A professzionális kérdőívszerkesztők esetén a válaszlehetőségek megjelenését véletlen sorrendre lehet állítani, így interjúalanyonként különböző (lehet) a sorrend.
- **Bonyolultabb logikai ugrásokat**, szűrőfeltételeket fogalmazhatunk meg, nem kell attól tartanunk, hogy a kérdezőbiztos nem tudja pontosan követni az utasításokat.
- Nem kérdőív szerkesztési szempont, de nagymértékben meghatározza az adatminőséget, a terepmunka gyorsaságát és költségét, hogy a programozott kérdőív során közvetlenül adatbázisba kerülnek az adatok, tehát **nincs szükség adatrögzítésre**.

2.5 Adatgyűjtés, terepmunka

A kutatási terv és a kérdőív szerkesztésének befejezése után indulhat a kutatás legidőigényesebb része, a terepmunka, vagy az angolból átvett nevén *field*. A kutatási folyamatnak ebben a szakaszában a kutató általában háttérbe szorul, átengedve a terepmunkát a terepkutatóknak, kérdezőbiztosoknak.



A terepmunka folyamata a következő szakaszokból áll (Malhotra, 2001): a terepkutatók kiválasztása, képzése, ellenőrzése, a munka elfogadása, értékelése.
A terepkutatók, kérdezőbiztosok képzése során megkülönböztethetjük az általános kapcsolatteremtő képesség fejlesztését adott kutatáshoz kapcsolódó ismeretek oktatásával. Olyan kutatócégeknél, ahol a folyamatos kiválasztás és képzésnek köszönhetően tapasztalt kérdezőbiztosok állnak a kutatás rendelkezésére, a terepmunka a kérdőív bemutatásával, elmagyarázásával kezdődik. Ilyenkor a kutatás vezetője:

- röviden bemutatja a kutatás célját;
- elmagyarázza a speciálisabb témákat;
- kérdésről kérdésre részletesen bemutatja a kérdőívet;
- a válaszok rögzítésének módját.

Miután a kutató bemutatta a kérdőívet a kérdezőbiztosoknak, a kutatás egészének szempontjából kulcsfontosságú momentum következik: rá tudja-e venni a kérdezőbiztos az interjúalanyt a kérdőív kitöltésére? Van néhány olyan technika, amivel ennek valószínűségét növelni tudjuk. A **sikeres interjú feltételei:**

- Az interjút néhány bevezető mondattal kell kezdeni, ami a kérdezés céljára, a kérdések jellegére és a kérdezés időtartamára vonatkozik. Utaljunk az anonimitásra és a válaszadás önkéntességére.
- Lehetőleg udvarias, barátságos légkört kell teremteni.
- A kérdezőbiztos ismerje jól a kérdőívet.
- Az általános kérdésektől haladjunk a speciálisabb kérdések felé.
- Érdeklődést kell mutatni, sohasem szabad meglepetéssel vagy helytelenítőleg reagálni a válaszra.
- Minden kérdést pontosan úgy kell feltenni, ahogyan megfogalmazták.

Ezek a technikák természetesen nem alkalmazhatók az önkitöltő kérdőív esetében. A **sikeres önkitöltő kérdőív** feltételei:

- Kísérőlevelet csatoljunk a kérdőív mellé, ami a kutatás céljait írja le, vagy a kérdőív elején írjuk le ugyanezt. Ebben udvarias megszólítást használjunk, tartalmazzon egy rövid témabemutatás, és sorolja fel a kérdőív kitöltésével járó „előnyöket".
- Részletes útmutató a kitöltéshez.
- Legyen minden maximálisan egyértelmű.

Az interjú **megtagadása** esetén a következőket tehetjük:

- Az elérési arány növelése érdekében: az el nem ért interjúalanyok többszöri felkeresése.
- A visszautasítás csökkentése érdekében:
  - – előzetes bejelentkezés, általában telefonon;
  - – az interjúalanyok érdeklődésének felkeltése;
  - – ösztönzők, ajándékok használata;
  - – a kérdőív tartalmi és szerkesztési módja.

Mint minden üzleti tevékenységnél, folyamatnál, itt is szükség van egy menedzsmentfunkcióra, a **kontrollra**. A kérdezőbiztosok ellenőrzésére a következő lehetőségek adottak:

- Minőség-ellenőrzés: Átnézzük a kitöltött kérdőívet, hogy az utasításoknak megfelelően van-e kitöltve.
- Mintavételi ellenőrzés. Személyes kérdezéseknél azt vizsgáljuk, hogy a mintavételi tervnek megfelelően választotta-e ki a kérdezőbiztos az interjúalanyt. Telefonos, internetes és postai kérdezéseknél különböző automatizmusok biztosítják ezt.
- Csalás kizárása. Előfordulhat, hogy a kérdezőbiztos a legközelebbi kávéautomata mellett „tölti ki" a kérdőíveket. Ennek ellenőrzése leginkább úgy történik, hogy szúrópróba-szerűen felhívják az interjúalanyokat, és rákérdeznek az interjúra.
- Az adatok alapján történő ellenőrzés. A terepmunka lezárása és az adattábla létrehozása után a kutatónak lehetősége van az adatok konzisztenciája és az eltérések vizsgálata alapján következtetni a csalás valószínűségére. (Ehhez az adattáblának tartalmaznia kell a kérdezőbiztos nevét vagy kódját.)

Az adatelőkészítés a terepmunka lezárásával, a kitöltött kérdőívek visszaérkezésével kezdődik.

Az adatelőkészítés folyamata a következő lépésekből áll:

1. A kérdőív ellenőrzése során a nem megfelelő válaszokat próbáljuk kezelni: a kérdőív visszaküldése, hiányzó értékek hozzárendelése, a rossz adatok kizárása által.
2. A kódolás nagyrészt a kérdőívszerkesztéskor megtörtént, de a nyílt kérdések kódolására ekkor kerül sor.
3. Adatbevitel – hagyományos személyes vagy telefonos kérdezésnél és a postai úton történő lekérdezésnél szükség van a kérdőíven rögzített adatok számítógépes adatbevitelére. A legpontosabban dolgozó adatrögzítőnél is történnek elütések, ezeket a hibákat is a következő lépésben próbáljuk kiszűrni.
4. Adattisztítás alatt a hibás adatok kijavítását, vagy ha erre nincs lehetőség, akkor a törlését értjük.
5. Statisztikai adatkiigazítás: súlyozás, a változó újradefiniálása, skálatranszformáció.
6. Adatelemzési tervet a marketingkutatási folyamat korábbi szakaszai (mindenekelőtt a kutatási célok), az adatok ismert jellemzői, a statisztikai módszerek tulajdonságai és a kutató felkészültsége és felfogása határozza meg.

3. BEVEZETÉS AZ SPSS PROGRAM HASZNÁLATÁBA

Az SPSS *(Statistical Package for Social Science)* egy olyan statisztikai program, amely kvantitatív adatok elemzésére specializálódott. A jegyzet írásakor az SPSS26 verziót használjuk, de a bemutatott módszerek és beállításaik szinte teljesen azonosak a különböző verziók esetében.
A tanulást különböző szerkesztési megoldásokkal is segíteni kívántuk:

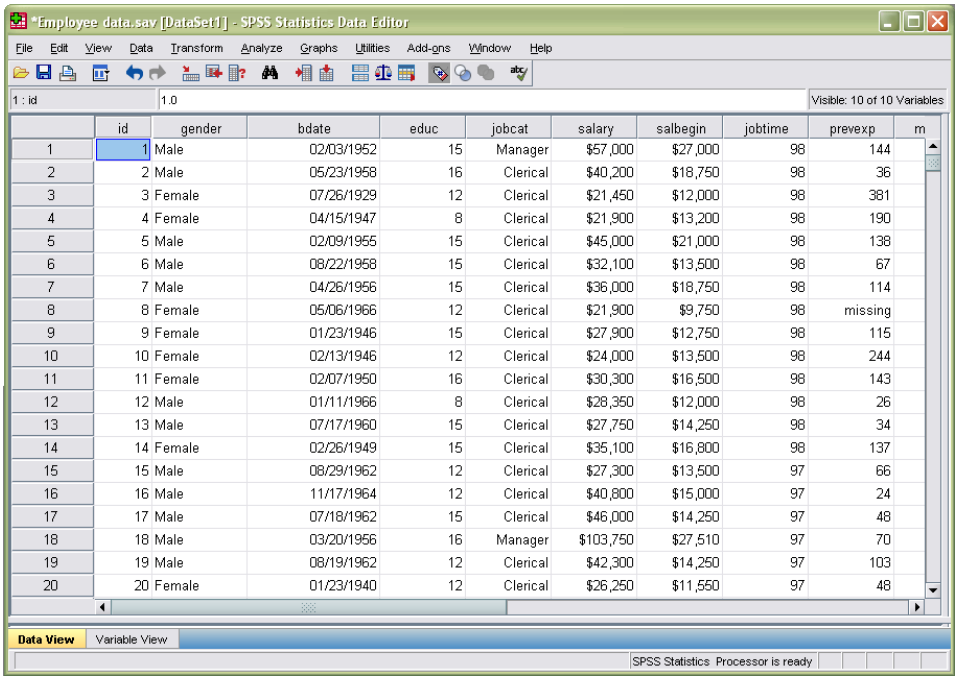
- valamennyi módszert egy **gyakorlati példán** keresztül mutatunk be. Ennek során nem használunk olyan adattáblákat, amihez az olvasó nem juthat hozzá, vagy az SPSS példa adattáblái[25] közül választunk, vagy megadjuk az elérhetőségét.
- a program számára kiadott utasításainkat a következőképp tüntetjük fel: **Analyze→DescriptiveStatistics→Frequency**. Az utasításokat a grafikus felületnek köszönhetően a menüsorból választjuk ki, nincs szükség a parancsok begépelésére. Ha csak az utasításra hivatkozunk, akkor azt nagybetűvel tesszük: FREQUENCY.
- a program ablakaiban, menüjében megjelenő beállítási lehetőségeket ábrákon is mutatjuk, szövegben mindig aláhúzva tüntetjük fel.
- minden módszer bemutatásánál a parancs futtatása után az **Eredmények értelmezése** következik.

Kezdjük el tehát az ismerkedést az SPSS-szel. A program több ablakot, felületet is tartalmaz, amelyek külön fájlként menthetők, kezelhetők. Alapbeállításban indításkor két ablakot nyit meg az SPSS:

1. Data Editor (adatszerkesztő ablak) – az adatokat tartalmazza, itt tudjuk az adatokat bevinni, módosítani. Más adatkezelő programokhoz hasonlóan (pl. Excel) az adatok mátrix formában vannak elrendezve, **a sorokban vannak a megfigyelési egységek** (pl. az interjúalanyok válaszai), **és az oszlopok e megfigyelési egységek ismérveit, változóit jelentik.**
Az adattábla fájlnak „.sav" kiterjesztése van. Az adatszerkesztő ablaknak két nézete van, amelyek között az ablak alján levő fülekre kattintva válthatunk: az egyik a tényleges adattáblanézet (Data View), a másik pedig a változók különböző jellemzőinek beállításait lehetővé tevő változónézet (Variable View).

[25] Ezeket nagy valószínűséggel a C:\Program Files\IBM\SPSS\Statistics\26\Samples mappában találjuk.

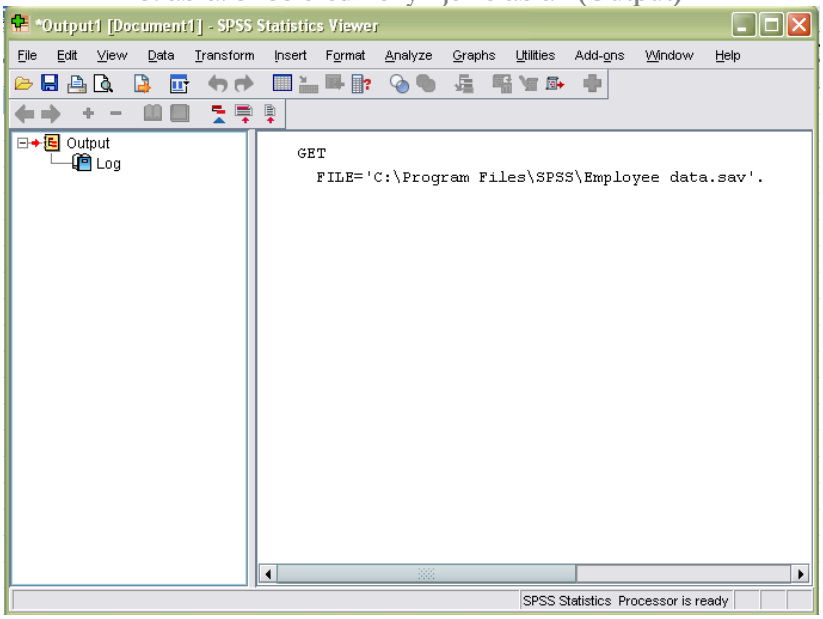
4. ábra. SPSS-adattábla



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	m
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	144	
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	36	
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	381	
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	190	
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	138	
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	67	
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98	114	
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98	missing	
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98	115	
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98	244	
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98	143	
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98	26	
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98	34	
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98	137	
15	15	Male	08/29/1962	12	Clerical	$27,300	$13,500	97	66	
16	16	Male	11/17/1964	12	Clerical	$40,800	$15,000	97	24	
17	17	Male	07/18/1962	15	Clerical	$46,000	$14,250	97	48	
18	18	Male	03/20/1956	16	Manager	$103,750	$27,510	97	70	
19	19	Male	08/19/1962	12	Clerical	$42,300	$14,250	97	103	
20	20	Female	01/23/1940	12	Clerical	$26,250	$11,550	97	48	

2. Az Output Viewer (eredménykijelző ablak) segítségével a feldolgozott statisztikai eredmények jelennek meg táblázatok, grafikonok formájában. A fájlnak .spo kiterjesztése van.

5. ábra. SPSS eredménykijelző ablak (Output)



Beállítástól függ, hogy a program indításakor automatikusan megnyílik-e az eredménykijelző ablak is, de bármilyen elemzési módszerre (Analyze menü) vagy grafikon készítésre (Graphs menü) vonatkozó parancs kiadása után az eredményekkel együtt szükségszerűen megjelenik.

A két ablak menüje megközelítőleg ugyanaz, viszont a menüsorban lévő ikonok különböznek. További programablakokat és funkciókat vehetnek igénybe a haladó felhasználók, például az utasítások parancssorait tartalmazó Syntax ablakot. Az SPSS legelső verziói nem tartalmaztak grafikus felületet, a parancsokat meghatározott szintakszis szerint kellet begépelni. Ez a lehetőség a későbbi verzióknál is opcionálisan megmaradt, főképp gyakran ismétlődő vagy egymással összefüggő utasítássorozatok esetén jelentősen felgyorsítja az elemzők munkáját. Másik gyakran igénybe vett funkciókat tartalmazó ablak az SPSS grafikonjainak szerkesztését lehetővé tevő Chart ablak. Ezt úgy tudjuk elindítani, ha duplán kattintunk az Output-ban megjelenő grafikonra. E két programablakba tartozó funkció részletes ismertetése nem célja a jegyzetnek, de perspektívaként szívesen ajánljuk az olvasó figyelmébe.

3.1 A változók típusai

A kvantitatív adatelemzés kiindulópontja, hogy felismerjük az adatok, a változók típusát. A mérési skálák leírásánál láthattuk, hogy négy fő típusa van a mérési skáláknak: nominális, ordinális, arány- és intervallumskála. Ennek megfelelően a változók (ismérvek) értékei e négyféle skála valamelyike közül kerülnek ki. Az adatelemzés kikerülhetetlen alapja, hogy felismerjük egy változó típusát, mivel ez meghatározza, hogy milyen statisztikai műveleteket, modelleket alkalmazhatunk.

1. Nominális skála → nominális változó. A változó értékeinek nincs számszerű jelentésük, hanem kódok, és a változó kategóriáit a kódokhoz rendelt címkék (*label*-ek) azonosítják. Pl. a cégforma a következő értékekkel: 1 – egyéni és családi vállalkozás; 2 – kft.; 3 – rt.; 4 – egyéb, vagy a fogyasztott üdítőital márkája: 1 – Coca-Cola; 2 – Pepsi-Cola; 3 – egyéb.
A kétértékű nominális változót dichotómnak nevezzük (pl. a nem változó 1 – nő, 2 – férfi értékekkel), és ennek egy típusa a *dummy* változó, amelynek két értéke 0 és 1 (pl. a nem változó 0 – nő, 1 – férfi értékekkel).

2. Sorrendi skála → ordinális változó. Az ordinális változó értékeinek sincs számszerű jelentésük, a változó értékeinek jelentését a címkék (*label*-ek) adják. Az különbözteti meg a nominális változótól, hogy a változó értékei, kategóriái között rangsorbeli különbség van.
Például az iskolai végzettség (1 – kevesebb, mint 10 osztály; 2 – szakiskola; 3 – érettségi; 4 – főiskola; 5 – egyetemi végzettség), elégedettség mértéke (1 – rossz; 2 – közepes; 3 – jó), a gazdaság fejlődése a következő évben (1 – csökkenés; 2 – stagnálás; 3 – növekedés).

3. Az arány- és az intervallumskála→numerikus/metrikus változó. Az arány- és az intervallumskála között olyan kismértékű gyakorlati különbség van, hogy az SPSS-ben történő adatelemzés során ezeket nem különböztetjük meg, hanem numerikus változótípusba soroljuk mindkettőt. A numerikus[26] változó értéke egy szám, amelynek önmagában jelentése van, nincs szükség címkére az értelmezéséhez.
Például az életkor évben kifejezve, a cég árbevétele, az előállított termékek darabszáma, a szolgáltatással kapcsolatos elégedettség mértéke 1-től 5-ig terjedő skálán stb.

Ez utóbbi – elégedettségre vonatkozó – példa előfordulása mind az ordinális, mind a numerikus változók között jelzi, hogy **nem mindig egyértelmű a változó típusának besorolása**. Amennyiben a kérdőívben és az interjúalany válaszában inkább a kategóriák címkéin van az értelmi hangsúly, akkor tekintsük ordinálisnak, ha pedig az 1-től 5-ig terjedő elégedettségi skálának[27] csak a két végpontja van címkézve, akkor numerikus. Nem követünk el hibát, ha mindkét változótípusra alkalmazható statisztikai módszereket alkalmazzuk az ilyen változókra.

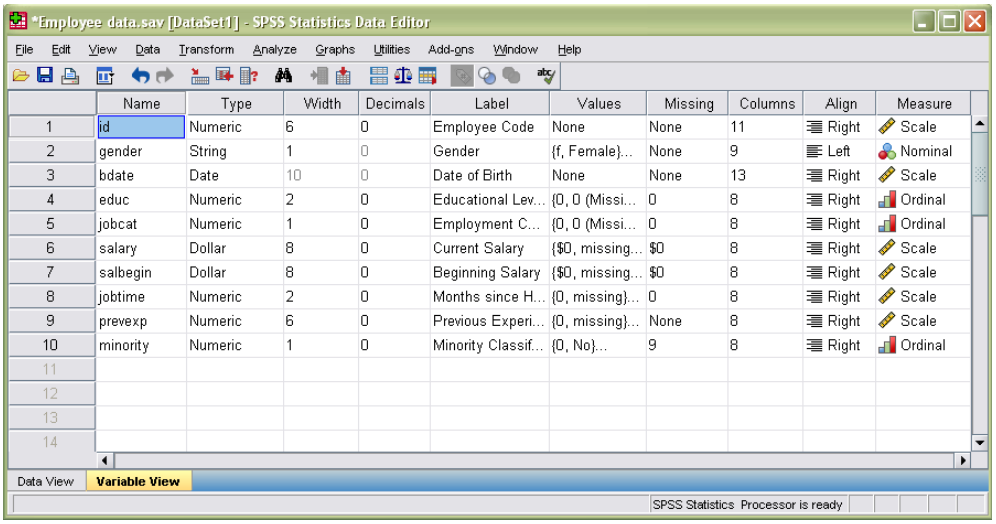
A mérési skálák nominális-ordinális-metrikus sorrendjét növekvőnek tartjuk, a metrikus a **legmagasabb mérési skála**, mivel a legtöbb matematikai műveletet lehet elvégezni az értékeivel.

3.2 A változók jellemzői

Mielőtt tovább haladunk az adatelemzési módszerek megismerése felé, nézzük meg, hogyan állíthatjuk be a változók különböző jellemzőit. Az adatszerkesztő ablak (Data Editor) két nézete közül már megismerkedtünk az adattáblanézettel (Data View), most a változónézet (Variable View) következik.

[26] A numerikus és metrikus változó fogalmakat szinonimaként használjuk.
[27] Az ilyen skálát *Likert*-skálának is nevezzük.

6. ábra. Az adattábla változónézete



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	6	0	Employee Code	None	None	11	Right	Scale
2	gender	String	1	0	Gender	{f, Female}...	None	9	Left	Nominal
3	bdate	Date	10	0	Date of Birth	None	None	13	Right	Scale
4	educ	Numeric	2	0	Educational Lev...	{0, 0 (Missi...	0	8	Right	Ordinal
5	jobcat	Numeric	1	0	Employment C...	{0, 0 (Missi...	0	8	Right	Ordinal
6	salary	Dollar	8	0	Current Salary	{$0, missing...	$0	8	Right	Scale
7	salbegin	Dollar	8	0	Beginning Salary	{$0, missing...	$0	8	Right	Scale
8	jobtime	Numeric	2	0	Months since H...	{0, missing}...	0	8	Right	Scale
9	prevexp	Numeric	6	0	Previous Experi...	{0, missing}...	None	8	Right	Scale
10	minority	Numeric	1	0	Minority Classif...	{0, No}...	9	8	Right	Ordinal
11										
12										
13										
14										

Name – a változó neve karakterekből és számokból állhat. Korábbi SPSS-verziók csak nyolc karakter hosszúságú nevet engedtek, most csak az a feltétel, hogy ne számmal kezdődjön, ne legyen benne szóköz, illetve a Windows-os fájlnevekben általában tilos karakterek (+-:?*, stb.) itt sem használhatók. A kutatók általában két lehetőség között választanak: vagy a változó tartalmára utaló rövidítést (pl. iskolai végzettség) használnak, vagy a kérdőív sorszámát egy karakterrel kiegészítve (pl. q1). Inkább az utóbbit ajánlhatjuk, főképp sok, 25-30-nál több változó esetén.

Type – a változó típusa. Megtévesztő elnevezés, mivel itt nem az előbb ismertetett, a mérési skáláknak megfelelő változó típusokról van szó, hanem adatbázis-programozási értelemben választhatunk a numerikus, dátum, a vesszővel, illetve ponttal elválasztott tizedes, a pénzösszeg (dollár), szövegformátumok között. A gyakorlatban nem sokat kell törődnünk ezekkel a beállításokkal, a változóink túlnyomó többsége – ebben az értelemben – numerikus. Néha előfordul, hogy a változó alapbeállításban nem mutatja a szám tizedeseit, de szükség van a megjelenítésére, vagy szöveget csak akkor tudunk bevinni az adattáblába, ha előzőleg a változót szövegtípusúra *(string)* állítottuk.

Width – a változó értékeinek karakterhossza.

Decimals – a tizedesek száma, az előbbi teljes karakterhosszúságon belül.

Label (változócímke) – a változó címkéje írható ebbe a mezőbe. Ez a címke alapján egyértelműsíthetjük, hogy pontosan mire vonatkozik a változó, milyen adatokat tartalmaz. A változó neve ugyanis csak nyolc karakter hosszúságot enged, ez ritkán

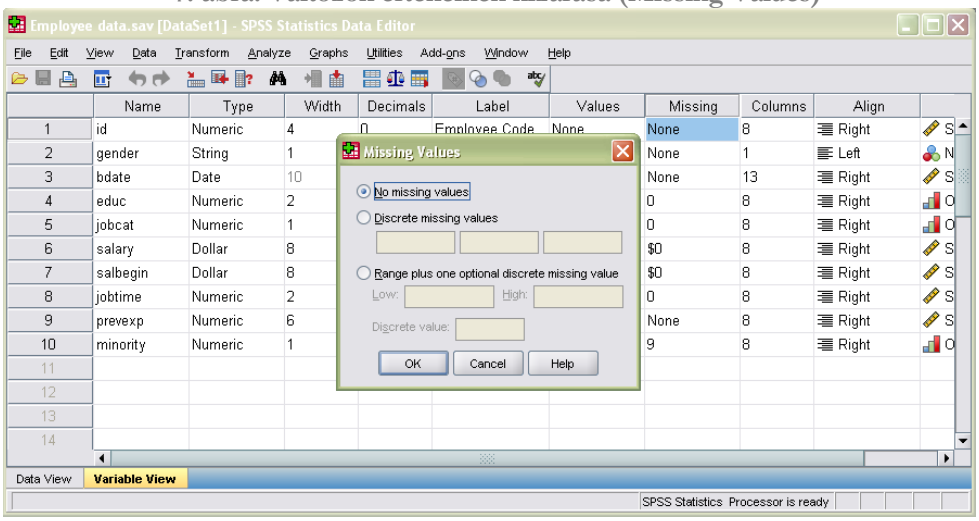
elég arra, hogy bemutassa a tartalmat. Kutatók gyakran a kérdőív kérdését másolják be (copy-paste) címkeként.

Values (értékcímke) – itt a változó értékeinek címkéit határozzuk meg. Természetesen csak a nominális és az ordinális változók értékei kódok, amelyek értelmezéséhez szükség van a szöveges címkékre, a numerikus változó értékeinek nincsenek címkéi.

Missing (hiányzó érték) **– a legfontosabb beállítási opció**, helytelen használata téves kutatási eredményekhez vezet. Az SPSS-ben két típusú hiányzó érték van:

- System Missing – nincs semmilyen érték az adatcellában. Ennek alapvetően két oka lehet: vagy valamilyen hiba történt (kérdezési, adatrögzítési, véletlenül kitöröltük az adattáblából), vagy a kérdőív logikai ugrásait követve nem is szabad adatnak lennie. (Például az internetezési szokásokat csak azoktól kérdezzük, akik egy megelőző kérdésnél azt állították, hogy interneteznek.) A hiányzó értékek vizsgálata nagyon fontos az adattisztítás során, amikor a kérdőív és a „nyers" adattábla összhangját vizsgáljuk.
- Missing Value - ebben az esetben a kutató határozza meg, hogy adott változó mely értékeit ne vegye figyelembe a program, gyakorlatilag hiányzónak tekinti. Tipikusan ilyen érték a Nem tudom/Nincs válasz (NT/NV) kódja, amely a változó karakterszámától függően a 9, 99, 999 stb. értéket szokta kapni. **A NT/NV kódját a gyakorisági eloszlásnál általában bennhagyjuk, de bármilyen változók közötti kapcsolatot vizsgáló módszer alkalmazásánál a NT/NV kódot szigorúan hiányzó értéknek kell tekinteni.** Szakzsargonnal szólva „a NT/NV-t kitesszük missing-re". Ennek hiányában a program NT/NV kódjával valós értékként számol, például figyelembe veszi egy numerikus változó átlagának a kiszámolásánál. A NT/NV válaszok kizárásának módját a 11. ábrán láthatjuk.

7. ábra. Változók értékeinek kizárása (Missing Values)



A Missing oszlopban a vizsgált változó cellájára kattintva felugrik egy Missing Values nevű ablak, amelyben alapbeállítás szerint nincs semmilyen érték bejelölve (No missing values). Lehetőségünk van három diszkrét értéket megadni a Discrete missing values felirat alatti mezőkben (mint említettem, a NT/NV leggyakrabban használt kódja a 9-es), vagy egy -tól -ig terjedő terjedelmet a Range... mezőkben, kiegészítve egy diszkrét értékkel.

Itt fontos megjegyezni, hogy egy változó értékei közül nem csak a NT/NV kódját indokolt időszakosan kizárni. Előfordulhat, hogy nincs szükségünk valamely értékekre, például egy numerikus változó gyakorisági eloszlásának szélein levő nagyon nagy vagy nagyon kis értékeket nem akarjuk bevonni az átlagszámításba.

Column és Align – a változó oszlopszélességét és a cella értékeinek vízszintes elhelyezkedését állíthatjuk be, sok szép kutatást el lehet végezni anélkül, hogy ezekkel foglalkoznánk.

Measure – itt állíthatjuk be a változó típusát (nominális, ordinális, numerikus). Kezdő felhasználónak, aki még nem tudja ránézésre gyorsan és biztosan megállapítani egy változó típusát, ajánlott ezeket a beállításokat az adatelemzés elkezdése előtt megejteni. Még egyszer hangsúlyozzuk, hogy ezek a mérési skáláknak megfelelő változótípusok nem ugyanazok, mint a Type oszlopban található adatbázis programozási szempontból fontos típusok.

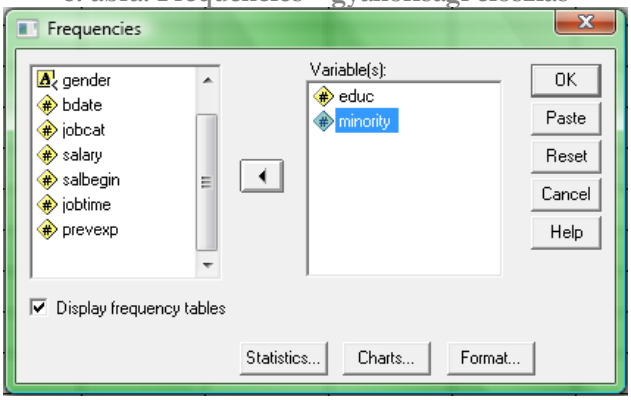
3.3 Gyakorisági eloszlás

Leggyakrabban használt SPSS-funkciók közé tartozik az egy vagy több változó értékeinek gyakorisági eloszlását eredményező FREQUENCY parancs. Nyissuk meg az SPSS példa-adattáblái közül, egy szervezet munkavállalói adatait tartalmazó employee data.sav fájlt.

Példa. Vizsgáljuk meg az employee data.sav adatfájlban a munkavállalók iskolai végzettségére (*educ*) és a kisebbségi státusára (*minority*) vonatkozó változók értékeinek gyakorisági eloszlását.

Analyze→Descriptive Statistics→ Frequency

8. ábra. Frequencies – gyakorisági eloszlás



A bal oldali ablak változólistájából kiválasztjuk, azaz kijelöljük, és a középen látható háromszögre kattintva átvisszük a minority változót a jobb oldali ablakba. Egyelőre ne törődjünk a többi beállítási lehetőséggel, hanem kattintsunk az OK gombra.

Eredmények értelmezése

A parancs futtatása után az eredmények azonnal megjelennek az Output-ablakban.

9. táblázat. A Frequency-be bevont változók

Statistics

		EDUC Education al Level (years)	MINORITY Minority Classifica tion
N	Valid	474	474
	Missing	0	0

Az output első táblázata (9. táblázat) megmutatja, hogy melyik változókat vontuk be a műveletbe, és ezeknek hány valós, illetve hiányzó értékük van.

A következő tábla (10. táblázat) az először megjelölt *educ* változó gyakorisági eloszlását tartalmazza. Az első oszlopban találjuk a változó értékeit, a legkisebb 8, a legnagyobb 21. Megállapíthatjuk, hogy az iskolai végzettség években kifejezett változója numerikus változó.

10. táblázat. Numerikus változó gyakorisági eloszlása

EDUC Educational Level (years)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	53	11.2	11.2	11.2
	12	190	40.1	40.1	51.3
	14	6	1.3	1.3	52.5
	15	116	24.5	24.5	77.0
	16	59	12.4	12.4	89.5
	17	11	2.3	2.3	91.8
	18	9	1.9	1.9	93.7
	19	27	5.7	5.7	99.4
	20	2	.4	.4	99.8
	21	1	.2	.2	100.0
	Total	**474**	**100.0**	**100.0**	

A második Frequency-oszlop az abszolút gyakorisági eloszlást tartalmazza, vagyis minden érték mintabeli előfordulásának számát. A Percent-oszlop a relatív gyakorisági eloszlás, az értékek mintabeli előfordulásának az aránya, egyszerűen úgy adódik, hogy elosztjuk (mármint az SPSS) az abszolút gyakoriságot a mintaelemszámmal. A Valid Percent értékei ebben a táblázatban nem különböznek a Percent-étől, mivel a vizsgált változó nem tartalmaz hiányzó értékeket. A Valid Percent ugyanis az a relatív gyakoriság, amely nem az adattábla teljes elemszámára arányosítja az abszolút gyakoriságot, hanem a változó valós, nem hiányzó értékeinek számára. A kumulált gyakoriság (Cumulative Percent) a Valid Percent-ben megjelenő relatív gyakoriság kumulálását, összegzését jelenti.

A következő táblázatnak két értéke van: a 0 jelöli a *Nem*-et, az 1-es pedig a kisebbségi sorba tartozást. Változónknak tehát két olyan értéke van, amit teljesen mértékben a címke (No vagy Yes) határoz meg, ezért ez egy **nominális változó**.

11. táblázat. Nominális változó gyakorisági eloszlása

MINORITY Minority Classification

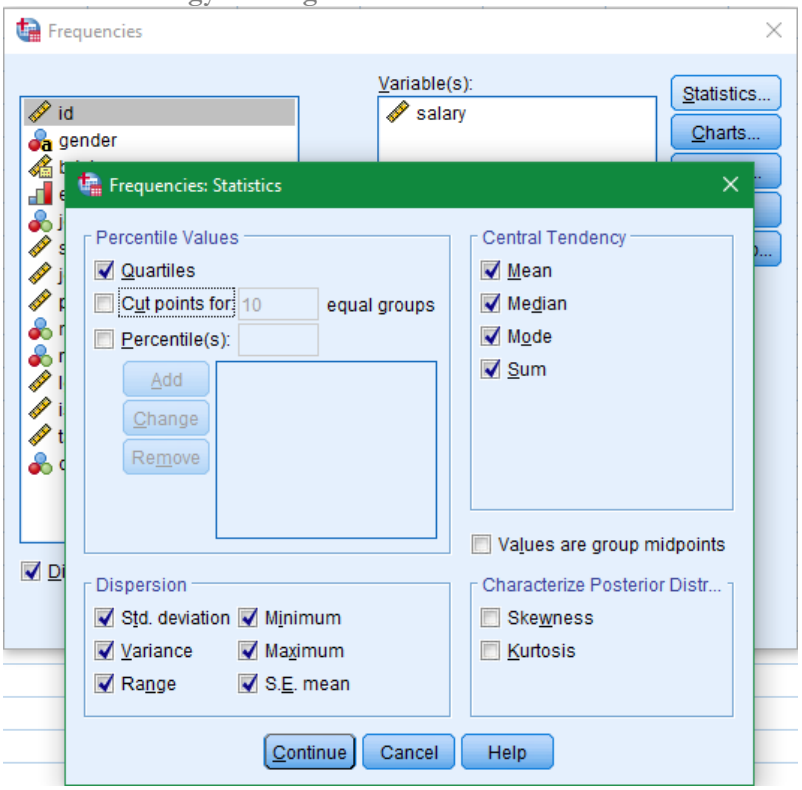
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 No	370	78.1	78.1	78.1
	1 Yes	104	21.9	21.9	100.0
	Total	**474**	**100.0**	**100.0**	

Ajánlott, hogy az adatelemzésünket mindig az adattáblában levő valamennyi változó gyakorisági eloszlásának a vizsgálatával kezdjük.

3.3.1 Helyzetmutatók

A leíró statisztika a minta ismérveit kvantitatív módon mutatja be. Megkülönböztetjük az **induktív (következtetési) statisztikától**, amely adott jellemző mintabeli értéke alapján, a valószínűségszámítás eszközeit alkalmazva becsüli az alapsokasági értéket.

9. ábra. A gyakoriségi eloszlás statisztikáinak beállításai



Leggyakrabban használt leíró statisztikák az elhelyezkedési vagy **helyzetmutatók**:

- **Számtani átlag** *(Mean)*: az eloszlás **centruma**. Kiszámításához összeadjuk az összes adatot, és elosztjuk az adatok számosságával.
- **Módusz** *(Mode)*: a változó értékei közül a **leggyakoribb**at, a legtöbbször előfordulót jelenti. Egy eloszlás lehet több móduszú is (például bimodális, trimodális stb.).
- **Medián** *(Median)*: a változó sorba rendezett értékek közül az a **középső érték**, amelyikhez képest a sorba rendezett értékek egyik fele nagyobb, a másik fele kisebb. Páros számú megfigyelések, mintaelemszám esetén a két középső érték átlaga a medián. Úgy is definiálhatjuk, hogy a medián az a változó érték, amelyiknél a kumulált gyakoriság meghaladja az 50%-ot.
  A medián fontos tulajdonsága, hogy értékét nem befolyásolják a szélső értékek (a nagyon alacsony vagy magas értékek), ezért gyakran használják az átlag alternatívájaként egy eloszlás jellemzésére.
- **Kvartilisek** *(Quartiles)*: Speciális helyzetmutatók, a medián általánosításai. Osztópontok segítségével a növekvő sorrendbe állított adataink egyenlő gyakoriságú osztályokra bonthatók, a kvartilisek négy egyenlő részre bontják a változó értékeinek eloszlását.
  - alsó kvartilis az az érték, amely alatt a sokaság értékeinek 25%-a található.
  - a középső kvartilis az az érték, amely alatt a sokaság fele által felvett értékek találhatóak. Ez egybeesik a mediánnal.
  - felső kvartilis pedig az az érték, amely alatt a sokaság értékeinek 75%-a található.
- **Percentilis** *(Percentile)* – az az érték, amely alatt a sokaság értékeinek meghatározott aránya (%-a) található

3.3.2 Szóródási mutatók

- **Terjedelem** *(Range)* – a maximum és minimum érték közötti különbség, amelyben szóródnak a változó értékei. A terjedelem a legnagyobb szóródást mutatja, de félrevezető lehet, ha az értékek nagy része egy szűk tartományban szóródik és csak néhány értéknek tulajdonítható a nagy terjedelem. Ezért a szóródás mérésénél a szórást
- **Szórás** *(Standard deviation)* - azt mutatja hogy az értékek átlagosan mennyivel térnek el a számtani átlaguktól. Formálisan az egyes értékek számtani átlagtól vett eltéréseinek négyzetes átlaga.
- **Variancia** – a szórás négyzete.

3.3.3 Alakmutatók

Az **eloszlás alakjára** vonatkozó ferdeségi *(Skewness)* mutató azt fejezi ki, hogy az eloszlás milyen irányban és mértékben tér el a szimmetrikus eloszlástól. A csúcsossági *(Kurtosis)* mutató jelzi, hogy az eloszlás a normálhoz viszonyítva csúcsosabb (jobban tömörül) vagy laposabb (kevésbé tömörül). Mindkét alakmutatót a későbbiekben, a Normalitásvizsgálat alfejezetben (4.5) részletezzük.

12. táblázat. A gyakorisági eloszlás statisztikái

Statistics

salary Current Salary

N	Valid	474
	Missing	0
Mean		$34,419.57
Std. Error of Mean		$784.311
Median		$28,875.00
Mode		$30,750
Std. Deviation		$17,075.661
Variance		291578214.5
Range		$119,250
Minimum		$15,750
Maximum		$135,000
Sum		$16,314,875
Percentiles	25	$24,000.00
	50	$28,875.00
	75	$37,162.50

Mivel a medián szokott a legtöbb problémát okozni a hallgatók számára, ezért példánkban ezt értelmezzük: az alkalmazottak fele 28 875 $-nál kevesebbet keres, az elégedettebbik másik fele pedig többet. A módusz 30 750$, ennyit keresnek a legtöbben. Ha lefuttattuk a FREQUENCY parancsot, és megjelenítettük a gyakorisági táblát is, akkor láthatjuk, hogy 13-an keresnek ennyit. Annak is van információtartalma a vizsgált vállalat jövedelempolitikájára vonatkozóan, hogy az egyes értékeknek alacsony a gyakorisága: nem kategóriákból felépített bértábla, hanem egyedi – valószínűleg a teljesítmény által meghatározott – bérek jellemzőek. A terjedelem óriási, ha a kvartilisekhez viszonyítjuk, jelzi, hogy néhányan nagyon sokat keresnek.

3.4 Adattábla-műveletek

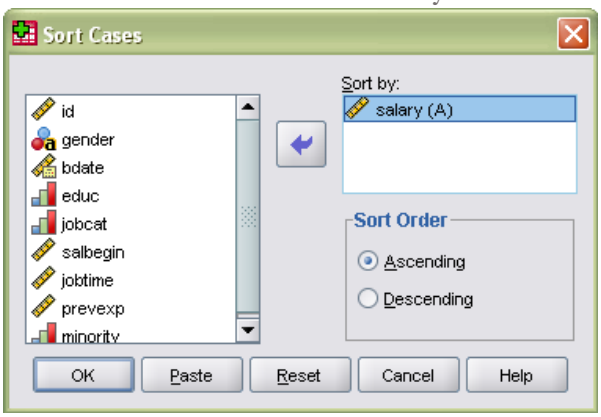
3.4.1 Az adattábla sorba rendezése és szelektálása

A Sort parancs az adattábla sorba rendezését eredményezi egy vagy több változó értékei szerint, növekvő vagy csökkenő sorrendben.

Példa: az adattábla sorba rendezése. Rendezzük sorba az employee data.sav adattáblát a fizetés nagysága szerinti növekvő sorrendbe.

Data→Sort Cases

10. ábra. Az adattábla sorba rendezése valamely változó értékei alapján



A változólistából kiválasztjuk a salary változót és a Sort Order résznél eldönthetjük, hogy növekvő (Ascending) vagy csökkenő (Descending) sorrendbe rendezzük az adattáblát.
Növekvő sorba rendezésnél az adott változó üres cellái, azaz hiányzó értékei (SYSTEM MISSING) a változó elejére kerülnek. E funkciónak a használatát nagyon gyakran a hiányzó vagy a hibás értékek, illetve egyéb adattisztítási műveletek indokolják.

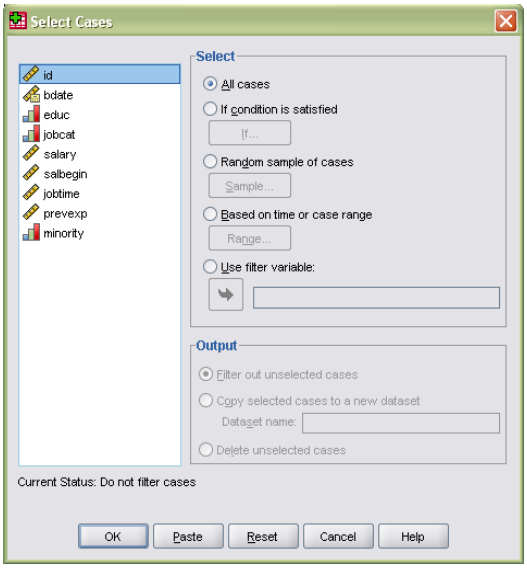
3.4.2 Az adattábla szűkítése, szelektálása

A Select paranccsal az adattábla esetei közül ideiglenesen vagy végérvényesen kizárhatjuk azokat az eseteket, amelyek nem felelnek meg egy bizonyos szűrőfeltételnek. A szűrőfeltételeket valamely változó vagy változók értékei alapján fogalmazhatjuk meg.

Példa: tételezzük fel, hogy egy cég munkavállalóinak adatait tartalmazó employee data.sav adatfájlban csak a kisebbséghez tartozó munkavállalók adataira van szükségünk. Ha megvizsgáljuk a *minority* változó értékeit – akár a változó ablakban, akár a változó gyakorisági eloszlását vizsgálva –, megállapíthatjuk, hogy az 1-es érték jelenti a kisebbségbe tartozást.

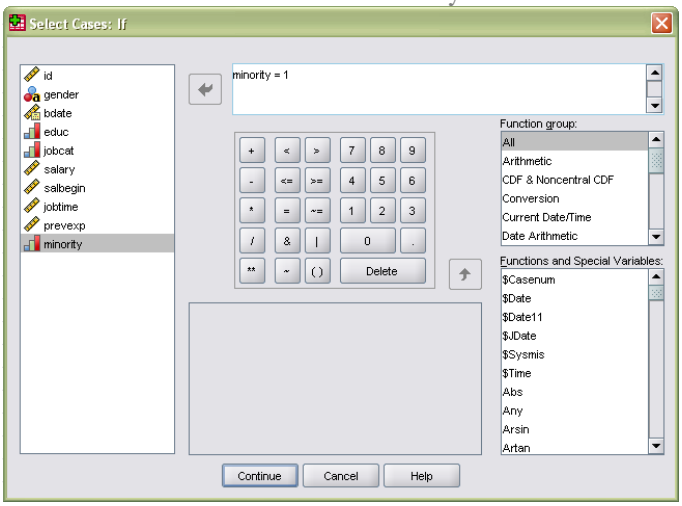
Data→Select Cases

11. ábra. Az adattábla szelektálása



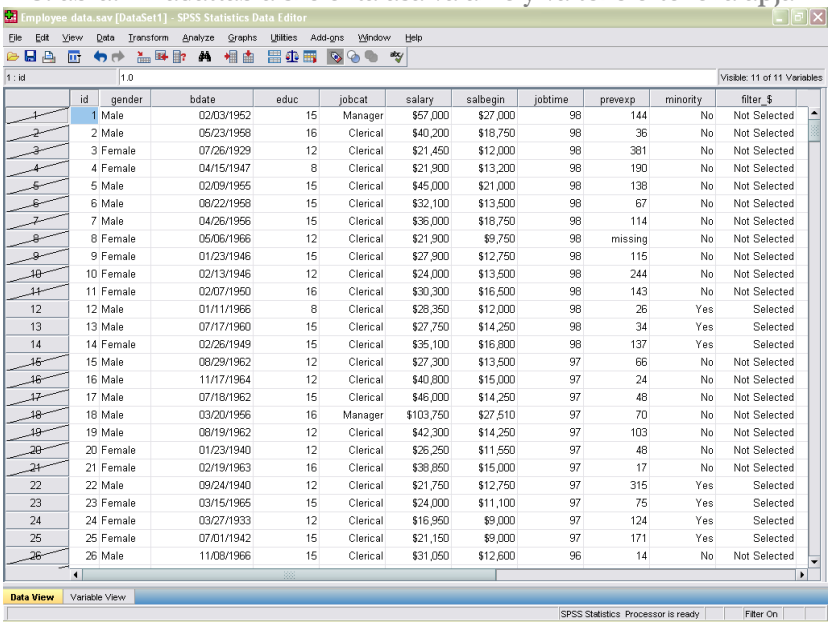
Az If condition is satisfied gombra, majd az If... -re kattintva felugrik a Select Cases: If ablak, amelyben beállíthatjuk a szűrőfeltételt.

12. ábra. Az adattábla szelektálása valamely változó értékei alapján



A baloldali ablak változólistájából kiválasztjuk, azaz kijelöljük, és a középen látható háromszögre kattintva átvisszük a *minority* változót a jobb oldali ablakba. A függvénylistában különböző függvények széles választéka található, de a *minority=1* feltétel beírásához erre most nincs szükségünk. Ezután a Continue, majd az OK gombbal zárjuk az utasítást.

13. ábra. Az adattábla szelektálása valamely változó értékei alapján



	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	filter_$
1	1	Male	02/03/1952	15	Manager	$57,000	$27,000	98	144	No	Not Selected
2	2	Male	05/23/1958	16	Clerical	$40,200	$18,750	98	36	No	Not Selected
3	3	Female	07/26/1929	12	Clerical	$21,450	$12,000	98	381	No	Not Selected
4	4	Female	04/15/1947	8	Clerical	$21,900	$13,200	98	190	No	Not Selected
5	5	Male	02/09/1955	15	Clerical	$45,000	$21,000	98	138	No	Not Selected
6	6	Male	08/22/1958	15	Clerical	$32,100	$13,500	98	67	No	Not Selected
7	7	Male	04/26/1956	15	Clerical	$36,000	$18,750	98	114	No	Not Selected
8	8	Female	05/06/1966	12	Clerical	$21,900	$9,750	98	missing	No	Not Selected
9	9	Female	01/23/1946	15	Clerical	$27,900	$12,750	98	115	No	Not Selected
10	10	Female	02/13/1946	12	Clerical	$24,000	$13,500	98	244	No	Not Selected
11	11	Female	02/07/1950	16	Clerical	$30,300	$16,500	98	143	No	Not Selected
12	12	Male	01/11/1966	8	Clerical	$28,350	$12,000	98	26	Yes	Selected
13	13	Male	07/17/1960	15	Clerical	$27,750	$14,250	98	34	Yes	Selected
14	14	Female	02/26/1949	15	Clerical	$35,100	$16,800	98	137	Yes	Selected
15	15	Male	08/29/1962	12	Clerical	$27,300	$13,500	97	66	No	Not Selected
16	16	Male	11/17/1964	12	Clerical	$40,800	$15,000	97	24	No	Not Selected
17	17	Male	07/18/1962	15	Clerical	$46,000	$14,250	97	48	No	Not Selected
18	18	Male	03/20/1956	16	Manager	$103,750	$27,510	97	70	No	Not Selected
19	19	Male	08/19/1962	12	Clerical	$42,300	$14,250	97	103	No	Not Selected
20	20	Female	01/23/1940	12	Clerical	$26,250	$11,550	97	48	No	Not Selected
21	21	Female	02/19/1963	16	Clerical	$38,850	$15,000	97	17	No	Not Selected
22	22	Male	09/24/1940	12	Clerical	$21,750	$12,750	97	315	Yes	Selected
23	23	Female	03/15/1965	15	Clerical	$24,000	$11,100	97	75	Yes	Selected
24	24	Female	03/27/1933	12	Clerical	$16,950	$9,000	97	124	Yes	Selected
25	25	Female	07/01/1942	15	Clerical	$21,150	$9,000	97	171	Yes	Selected
26	26	Male	11/08/1966	15	Clerical	$31,050	$12,600	96	14	No	Not Selected

A szűrőfeltételnek nem megfelelő, ideiglenesen **kizárt esetek sorszáma át van húzva**, illetve az adattábla végén egy új változó (filter_$) is jelzi, hogy melyik esetek lettek leszelektálva.

Ezek után valamennyi SPSS utasítás csak a szelektált adattáblán, a szűkített mintán lesz végrehajtva, amíg a Select Cases ablakban vissza nem állítjuk az All cases opciót. Amennyiben biztosak vagyunk abban, hogy nem lesz többet szükségünk a szűrőfeltételnek nem megfelelő esetekre szintén a Select Cases ablakban, az Unselected Cases Are résznél választhatjuk a Deleted lehetőséget.

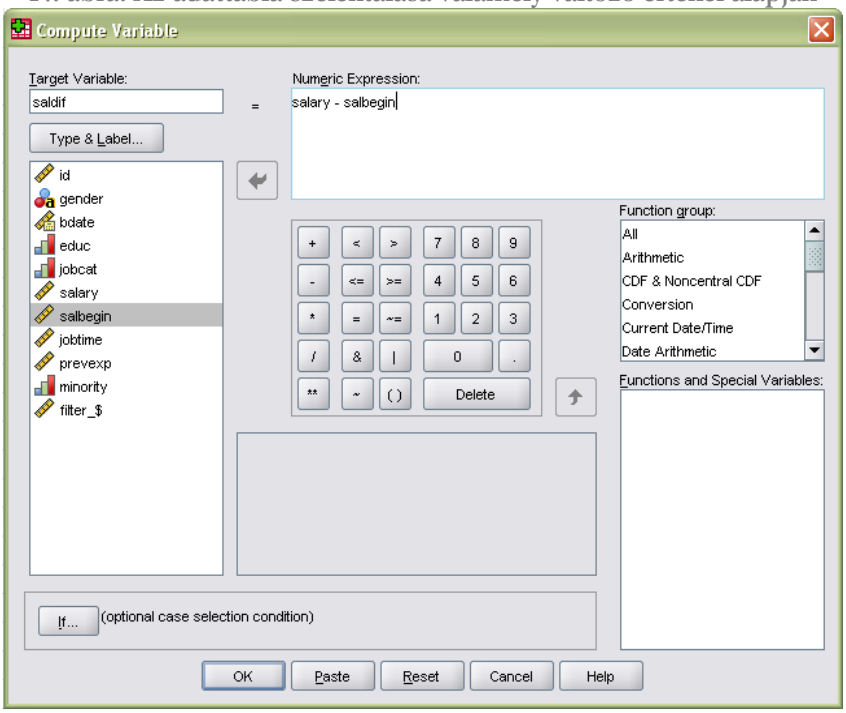
3.4.3 Új változó képzése

A **Compute** paranccsal egy új változót hozhatunk létre a meglévő változóink átalakítása révén.

Példa. Hozzunk létre egy új változót az employee data.sav adattáblában, ami az alkalmazottak fizetéseinek növekedését mutatja. Legyen az új *saldif* változó a jelenlegi (*salary*) és a kezdő fizetés (*salbegin*) különbsége.

Transform→Compute

14. ábra. Az adattábla szelektálása valamely változó értékei alapján



A Target Variable mezőben adhatunk nevet az újonnan létrehozandó változónak. A Numeric Expression mezőben határozhatjuk meg az új változót meghatározó algoritmust, amihez a változólistából hozzuk a régi változót, a számokat, matematikai műveleteket és/vagy függvényeket, vagy közvetlenül begépeléssel, esetleg az ablak menüjéből választjuk ki.

Jelenlegi példánkban nincs rá szükség, de az If... opciónál beállíthatunk olyan logikai feltételt, amelynek teljesülése esetén hajtódik végre csak a COMPUTE parancs. Ez a lehetőség funkcionalitásában megegyezik az előzőekben bemutatott SELECT paranccsal, a teljes adattáblát leszűkítjük egy valamilyen logikai feltételnek megfelelő rész-adattáblára. Például ha a jelenlegi és a kezdő fizetés közötti különbséget csak a legalább egy éve a cégnél dolgozókra akarjuk kiszámítani, akkor az If... ablakban a jobtime<12 szűrőfeltételt kell beírnunk.

A COMPUTE parancs végeredménye egy új változó, amely valamennyi munkavállalóra (hacsak nem állítottunk be szűrőfeltételt) meghatározza, hogy a cégnél mennyit nőtt a fizetése.

Egy kis házi feladat: az adatelemzések során az abszolút számoknál gyakran több információt hordoznak a fajlagos, valamilyen viszonyítási alapra vetített mutatók. Példánkban a fizetések növekedését nyilván nemcsak a munkavállaló teljesítményét honoráló fizetésemelések befolyásolják, hanem az is, hogy az illető mennyi ideje van a cégnél. Képezzünk egy olyan új változót, ami az egy hónapra vetített átlagos fizetésnövekedést mutatja, azaz képezzük a fizetésnövekedés (*saldif*) és a cégnél eltöltött idő hányadosát (*jobtime*).

3.4.4 Változók újrakódolása

Gyakran előfordul, hogy **egy változó értékei nem az adatelemzési tervünknek megfelelő formában találhatóak**. Például tételezzük fel, hogy a kérdezettek életkorára vonatkozó változónk nominális a következő 16–20, 21–25, 26–35, 36–50, 50+ életkor-kategóriákkal, de más eredményekkel való összehasonlíthatóság miatt 16–25 éves kategóriára lenne szükségünk.

Ilyen „kompatibilitási" probléma oka lehet, hogy adataink szekunder adatforrásból származnak, ezért már nem tudjuk befolyásolni az ismérvek kategóriáinak a kódolását, primer adatgyűjtés esetén nem voltunk elég figyelmesek a kérdőív készítésekor, vagy csak az eredeti adatelemzési terveinkhez képest újabb ötleteink adódtak.
A RECODE utasítással mindhárom típusú (nominális, ordinális, metrikus) változó értékeit átalakíthatjuk, vagy a magasabb mérési szintű változóból képezhetünk alacsonyabb szintűt (pl. a metrikusból ordinálist vagy nominálist). Két típusa van:

- Recode into Same Variables: az eredeti változók értékeit kódoljuk át úgy, hogy a régi értékek nem maradnak meg.
- Recode into Different Variables: az új kódok egy új változóban jelennek meg, ezáltal az eredeti változó is megmarad.

Példa. A változó értékeinek újrakódolása. FREQUENCY-t futtatva szintén az employee data adattábla *educ* változójára láthatjuk (13. táblázat), hogy az iskolai végzettséget az elvégzett évek számával mérik, vagyis egy numerikus változóval van dolgunk, amely 8 és 21 között vesz fel értékeket.

13. táblázat. Az iskolai végzettség változó gyakorisági eloszlása

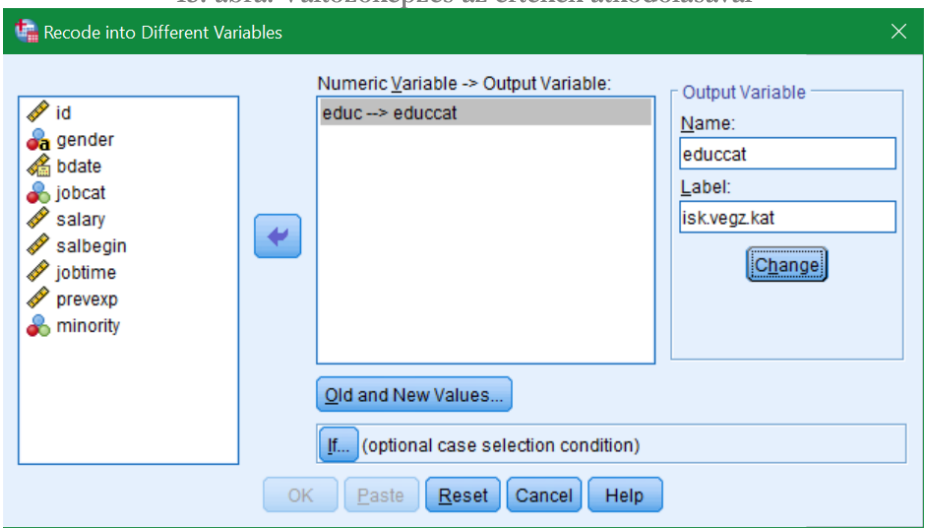
educ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	53	11.2	11.2	11.2
	12	190	40.1	40.1	51.3
	14	6	1.3	1.3	52.5
	15	116	24.5	24.5	77.0
	16	59	12.4	12.4	89.5
	17	11	2.3	2.3	91.8
	18	9	1.9	1.9	93.7
	19	27	5.7	5.7	99.4
	20	2	.4	.4	99.8
	21	1	.2	.2	100.0
	Total	474	100.0	100.0	

Képezzünk ebből egy új nominális változót a következő kategóriákkal: 8→1 „alapfokú", 12→2 „középfokú", 14–16→3 „főiskolai", 17–21→4 „egyetemi, posztgraduális". A felsősokú végzettség két kategóriába sorolását a 12 osztálynál többel rendelkezők nagy aránya indokolja (48.7%, lásd inverz kumulált gyakoriság).

Transform→Recode Into Different Variables

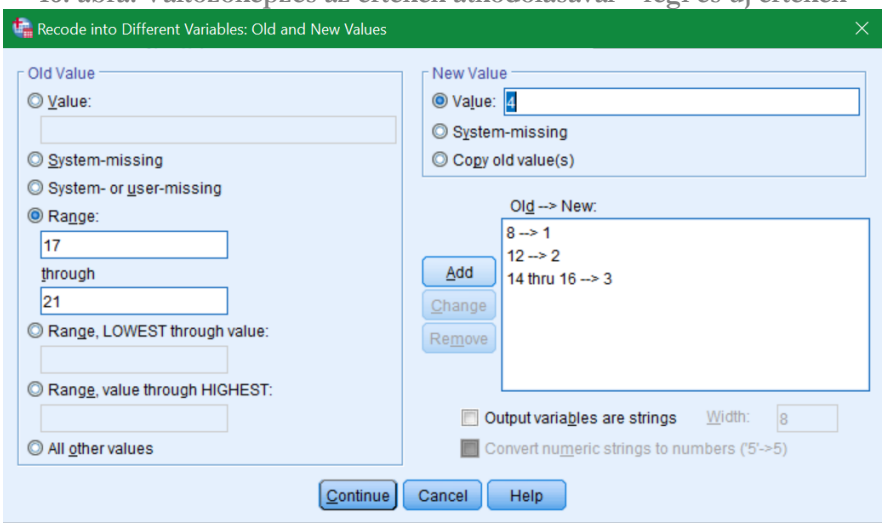
15. ábra. Változóképzés az értékek átkódolásával



A változólistából kiválasztjuk az educ változót, az Output Variable mezőbe pedig az új változó nevét kell beírnunk. Legyen az új név educcat jelezve, hogy kategóriákat tartalmazó nominális változóval van dolgunk, majd klikkeljünk a Change gombra, hogy a NumericVariable→Output Variable ablakban a régi

mellett megjelenjen az új változó neve is. Az átkódolni kívánt változó és az új változó nevének meghatározása után nézzük a többi beállítást. Az If... opció választásával szűkítő feltételt lehetne megfogalmazni, ebben az esetben – akárcsak a COMPUTE utasításnál – a RECODE parancs csak az adott feltételnek megfelelő esetekre lesz értelmezve.

16. ábra. Változóképzés az értékek átkódolásával – régi és új értékek



Az Old and New Values gombra kattintva megfeleltethetjük egymásnak a régi és új értékeket. Az Old Value ablakrészbe az eredeti *educ* változó értékeit írjuk be. Több opció közül választhatunk:

- egyetlen diszkrét érték esetén a Value gombra;
- hiányzó érték átkódolásakor a System-missing gombra kattintunk;
- ha pedig a változónk folytonos értékeire -tól -ig határok közötti terjedelmet akarunk meghatározni, akkor a Range három opciója közül választhatunk;
- az All other values-szal valamennyi másképp nem definiált értéket alakíthatjuk át.

Példánkban tehát beírjuk a 8-as értéket az Old Value ablakrész Value mezőjébe, az új 1-es kódot pedig a New Value ablakrész Value mezőjébe, majd az Add gombbal véglegesítjük a kódolást. (Ha mégis változtatni szeretnénk, akkor a Change, illetve Remove funkciókkal ez lehetséges.) Hasonlóképp járunk el a 12-es érték kódolásánál, a 14–16 és a 17–21 intervallumokat pedig a Range ablakokba visszük be. A végeredmény egy új *educcat* nevű, nominális változó négy értékkel.
Hasonlóképp kell eljárni a Recode into Same Variable utasításnál, itt kevesebb beállításra van szükségünk, mivel nem hozunk létre új változót.

4. VÁLTOZÓK KÖZÖTTI EGYDIMENZIÓS KAPCSOLATOK VIZSGÁLATA

Már a könyv elején hangsúlyoztuk, hogy a gazdasági jelenségek egymást meghatározó, befolyásoló kölcsönhatásban vannak. Ezek vizsgálata az adatelemzés során a változók közötti kapcsolatok vizsgálatát jelenti. Háromféle kapcsolattípusról beszélhetünk:

1. **Függetlenség**: a két változó független egymástól, ha az egyik változó értékeinek (eloszlásának) ismerete nem nyújt semmilyen információt a másik változó értékeiről.
2. **Sztochasztikus kapcsolat**: a két változó között sztochasztikus kapcsolat van, ha az egyik változó értékeinek ismerete kisebb-nagyobb mértékben, de nem teljesen egyértelműen magyarázza a másik változó értékeit.
3. **Determinisztikus kapcsolat**: a két változó között determinisztikus vagy függvényszerű a kapcsolat, ha az egyik változó értékeinek ismerete alapján biztosan következtethetünk a másik változó értékeire.

Empirikus kutatások adatai, ha nem függetlenek egymástól, akkor **leggyakrabban sztochasztikusan kapcsolódnak**. Determinisztikus kapcsolattal empirikus gazdasági vagy társadalomtudományi kutatásokban nagyon ritkán találkozunk, általában változó transzformáció során, vagy új, származtatott változók képzésekor definiálunk függvényszerű kapcsolatot két vagy több változó között.

Az alábbi táblázatban a sztochasztikus kapcsolatok vizsgálatára alkalmas módszereket tüntettük fel a változók típusa szerint.

14. táblázat. A változók közötti kapcsolatok vizsgálatára használt módszerek

	Nominális	**Ordinális**	**Numerikus**
Nominális	Kereszttábla	Kereszttábla ANOVA	t-próba ANOVA
Ordinális		Kereszttábla Rangkorreláció	t-próba ANOVA
Numerikus			Korrelációanalízis Regresszióanalízis

Forrás: saját szerkesztés

4.1 Kereszttábla-elemzés

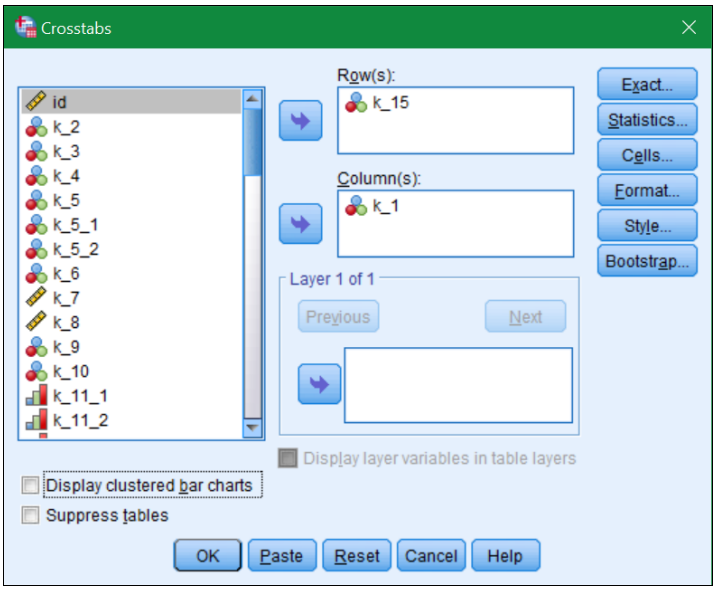
A kereszttábla-elemzés két nominális vagy ordinális változó közötti kapcsolat meglétének és a kapcsolat szorosságának vizsgálatára alkalmas.

4.1.1 Nominális változók közötti kereszttábla

Bemutató példánkhoz a GOBE2012.sav adatfájlt[28] hívjuk segítségül, azt szeretnénk megvizsgálni, hogy a GÓBÉ márka választásának indokai között vannak-e nemi különbségek. Megismerhetjük a két változó értékeit, ha Frequencyt futtatunk a két változóra. A GÓBÉ márka vásárlásának indokai (k_15 változó): jobb minőség (35,7%), ezzel támogatom a helyi termelőket (28,6%), mert ez saját, székely termék (18,7%), bio alapanyagokból készült (14,3%) és olcsóbb (2,7%). A nem változó (k_1) esetében nincs semmi meglepetés, dichotóm változó: nő (58,5%) és férfi (41,5%).
A kutatási hipotézisünk legyen az, hogy a férfiak és nők eltérő arányban említik az öt indok valamelyikét. A választott módszer, a kereszttábla-elemzés statisztikai nullhipotézise, hogy a két változó független egymástól.

Analyze → Descriptive Statistics → Crosstabs

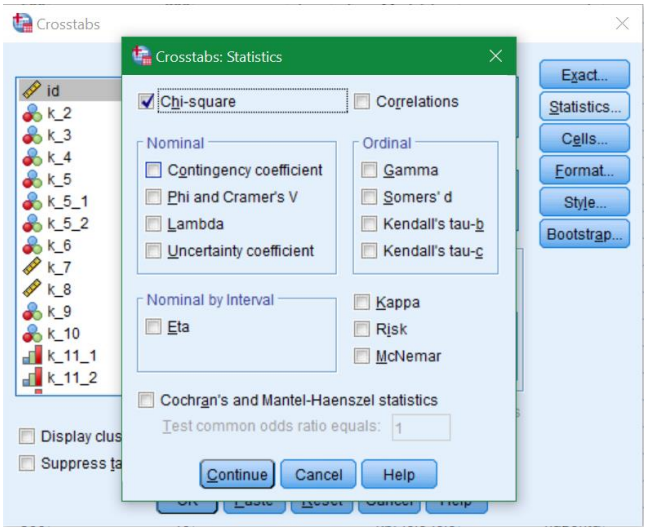
17. ábra. Kereszttábla: változók beállításai



[28] Letölthető a http://web.csik.sapientia.ro/kutatasmodszertan oldalról.

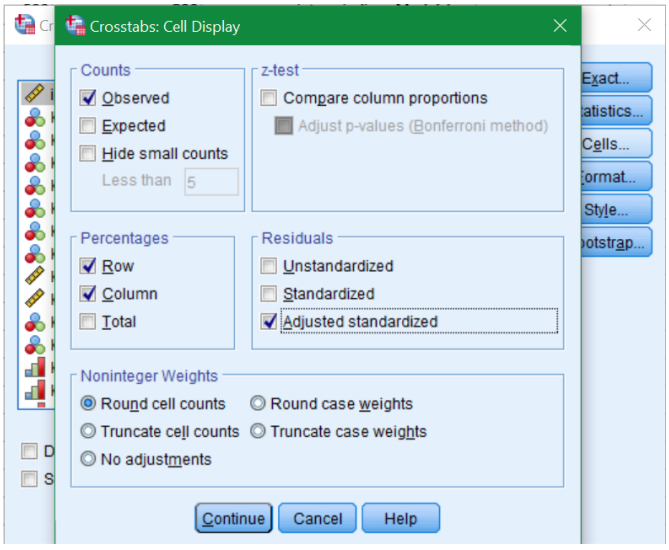
A Row(s) mezőbe bevisszük a GÓBÉ márka választásának indokait tartalmazó változót (k_15), a Column(s) mezőbe pedig a nem (k_1) változót. Ez a választás tetszőleges, a két változó akár helyet is cserélhet a kereszttáblában (matematikailag fogalmazva transzponálhatjuk a mátrixot).
Az alapvető (default) beállításokat egészítsük ki a következőkkel:

18. ábra. Kereszttábla: statisztikai beállítások



A Statistics ablakban állítsuk be a Chi-square opciót, ami a hipotézisünk ellenőrzésére alkalmas statisztikát fogja mutatni.

19. ábra. Kereszttábla: a cellák mutatóinak beállítása



A Cells ablakban a Percentages mezőben állítsuk be a Row, Column opciókat, ami a sor- és oszlopszázalékokat eredményezi, továbbá a Residuals mezőben az Adj. Standarized opciót is pipáljuk be.

Eredmények értelmezése. Az Output-ban a cím után következő első táblázat arról informál, hogy az összes esetből (Total) hány valós érték (Valid) lett bevonva a műveletbe, és hány hiányzó érték (Missing) maradt ki belőle.

15. táblázat. Kereszttábla: eredmények

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
k_15 Miért vásárolnak Góbé terméket? * k_1 Neme	364	100.0%	0	0.0%	364	100.0%

Az összesítő tábla után folytassuk az eredmények értelmezését az output harmadik, Chi-Square Test feliratú táblázatával, ami alapján a két változó közötti kapcsolatról dönthetünk.

16. táblázat. Kereszttábla: khi-négyzet próba

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	19.479[a]	4	.001
Likelihood Ratio	23.091	4	.000
Linear-by-Linear Association	8.192	1	.004
N of Valid Cases	364		

a. 1 cells (10.0%) have expected count less than 5. The minimum expected count is 4.15.

A khi-négyzet teszt Pearson Chi-Square értékének szignifikanciaszintje (az Asymptotic Significance (2-sided) oszlopban) mutatja, hogy a két változó egymástól nem független. **Ha a Pearson Chi-Square mutató (p-érték) szignifikanciaszintje kisebb, mint .05, akkor elutasítjuk a két változó függetlenségére vonatkozó hipotézisünket, azaz a két változó között kapcsolat van.** Ez a kapcsolat statisztikailag bizonyított, vagyis szignifikáns. A .05-nél kisebb szignifikanciaszint értelmezhető úgy is, hogy kevesebb, mint 5%-os biztonsággal fogadhatjuk el a nullhipotézisünket, vagyis állíthatjuk, hogy a két változó független.

Ezzel egyenértékűen úgy is fogalmazhatunk, hogy **több mint 95%-os biztonsággal jelenthetjük ki a két változó kapcsolatát**.
Amennyiben a p értéke .000, akkor ezt nem úgy értelmezzük, hogy a p érték nulla, hanem p< .0005, mivel ez a valószínűség sosem teljesen nulla.

A következő táblázat a tényleges kereszttábla, ami a két változó értékeinek együttes megoszlásait tartalmazza:

17. táblázat. Kereszttábla: kontingenciatáblázat

k_15 Miért vásárolnak Góbé terméket? * k_1 Neme Crosstabulation

			k_1 Neme		
			1 Nő	2 Férfi	Total
k_15 Miért vásárolnak Góbé terméket?	1 jobb minőségű	Count	84	46	130
		% within k_15 Miért vásárolnak Góbé terméket?	64.6%	35.4%	100.0%
		% within k_1 Neme	39.4%	30.5%	35.7%
		Adjusted Residual	1.8	-1.8	
	2 olcsóbb	Count	10	0	10
		% within k_15 Miért vásárolnak Góbé terméket?	100.0%	0.0%	100.0%
		% within k_1 Neme	4.7%	0.0%	2.7%
		Adjusted Residual	2.7	-2.7	
	3 bio alapanyagokból készül	Count	36	16	52
		% within k_15 Miért vásárolnak Góbé terméket?	69.2%	30.8%	100.0%
		% within k_1 Neme	16.9%	10.6%	14.3%
		Adjusted Residual	1.7	-1.7	
	4 ezzel támogatom a helyi termelőket	Count	48	56	104
		% within k_15 Miért vásárolnak Góbé terméket?	46.2%	53.8%	100.0%
		% within k_1 Neme	22.5%	37.1%	28.6%
		Adjusted Residual	-3.0	3.0	
	5 mert ez saját, székely termék	Count	35	33	68
		% within k_15 Miért vásárolnak Góbé terméket?	51.5%	48.5%	100.0%
		% within k_1 Neme	16.4%	21.9%	18.7%
		Adjusted Residual	-1.3	1.3	
Total		Count	213	151	364
		% within k_15 Miért vásárolnak Góbé terméket?	58.5%	41.5%	100.0%
		% within k_1 Neme	100.0%	100.0%	100.0%

A nem változó két értéke és a vásárlás indokai változó öt értéke egy tíz cellából álló mátrixot eredményez, ami kiegészül egy Total oszloppal és sorral. A cellákban a következő értékeket láthatjuk:

- Count: **abszolút gyakorisági eloszlás**, azt mutatja, hogy hány eset tartozik a két kategória közös részhalmazába. Pl. az adattáblánkban 84 olyan nő van, akik a jobb minőség miatt választják a GÓBÉ termékeket.
- % within k 15: **sorszázalék**, a 100% az adott sorban levő változó kategóriáját jelenti, sor mentén összegezve a százalékokat kapjuk a 100%-ot. Pl. azoknak, akik a jobb minőség miatt választják a GÓBÉ terméket, 64,6%-uk nő, és 35,4%-uk férfi.
- % within k 1: **oszlopszázalék**, a 100% az adott oszlopban lévő változókategóriát jelenti. Oszlopmentén összegezve az oszlopszázalékokat kapjuk a 100%-ot. Pl. a nők 39,4%-a a jobb minőség, 4,7%-a az olcsósága, 16,9%-a a bio alapanyagok, 22,5%-a a helyi termelők támogatása és 16,4%-a pedig a székely termék jellege miatt vásárol GÓBÉ terméket.
- Adjusted Residual: a két változó közötti szignifikáns kapcsolatot okozó értékek indikátora.

A khi-négyzet teszt alapján megállapítottuk, hogy a két változó kapcsolatban van, de még nem tudjuk, hogy a két változó kétszer öt „kapcsolódási pontjai", a kereszttábla tíz cellája közül melyek okozzák ezt a kapcsolatot. Ennek meghatározására szolgál a kereszttábla celláinak utolsó értéke, az adjusztált reziduum (Adjusted Residual). A nem túl szabadon fordított nevű mutató a következőképp jelzi a két változó kategóriái közötti kapcsolat létét:

- ha az adjusztált reziduum **abszolút értéke nagyobb vagy egyenlő 2-nél, akkor a két kategória között szignifikáns kapcsolat van,**
- értelemszerűen, ha az adjusztált reziduum abszolút értéke 2-nél, akkor a két kategória között nincs szignifikáns kapcsolat.

Az adjusztált reziduumok vizsgálatát a khi-négyzet próba vizsgálatával együtt kell végrehajtani. Viszonylag ritkán, de előfordul, hogy a khi-négyzet próba szignifikanciaszintje valamivel magasabb, mint 0,05, de a kereszttábla egy-két cellájának adjusztált reziduuma -2-nél kisebb, vagy 2-nél nagyobb. Ebben az esetben a khi-négyzet próba alapján mondjuk ki a két változó függetlenségét.

Példánkat folytatva az adjusztált reziduumok vizsgálata alapján megállapíthatjuk, hogy a két változó kategóriáinak tíz kapcsolódási pontja közül négyben szignifikáns a kapcsolat. A kutatási hipotézishez közelítő fogalmazással mondhatjuk, hogy a GÓBÉ termékek vásárlásának okai közül az olcsóság és a helyi termelők támogatásának említési aránya szignifikánsan eltér a két nem között.

A helyi termelők támogatása a férfiak által szignifikánsan nagyobb arányban említett szempont, a férfiak 37,1%-a, míg a nők 22,5%-a említette. Az olcsóság pedig a nők körében gyakrabban említett, 4,7%, szemben a férfiak 0%-ával. Fontos megértenünk, hogy ez a szempont **nem jellemző** a nőkre, mert a 4,7%-os említési gyakorisága csak az utolsó helyre elég az öt szempont közül, de szignifikánsan **jellemzőbb**, mint a férfiakra.
A másik három szempont említési gyakorisága – ahol az adjusztált reziduum abszolút értéke kisebb, mint 2 – nem különbözik szignifikánsan a nemek között. Például a jobb minőség 39,4%-os említése a nők körében nem különbözik statisztikailag a férfiak 30,5%-ától.

A kereszttábla-elemzés feltételei
Többértékű nominális változók kereszttáblájában és/vagy viszonylag kis minta esetén gyakran előfordul, hogy egy adott cellában az esetek száma (Count) túl alacsony. Általánosan elfogadott szabály szerint a kereszttábla összes cellájának 20%-ában lehet a gyakoriság kisebb, mint 5.
Ha ez a feltétel nem teljesül, akkor a változók kategóriáinak összevonásával tudjuk növelni a cellák esetszámát. Például egy részletesebb iskolai végzettség változó értékeit összevonjuk alap-, közép- és felsőfokú végzettségbe. A másik lehetőség kis mintánk esetére a **Fisher-teszt**, amelyet a későbbiekben tárgyalunk.

Célirányosabban: függő és független változó
Az eddig tanultak elégségesek ahhoz, hogy a kereszttábla-elemzés módszerével két nominális változó közötti szignifikáns kapcsolat létét megállapítsuk, és azt is meg tudjuk határozni, hogy a két változó kategóriái közül melyek okozzák a kapcsolatot. Azonban további információkat is nyerhetünk a vizsgált két változó kapcsolatáról, illetve megkönnyíthetjük, gyorsíthatjuk a kereszttábla értelmezését.

Egyszerűsíthetjük a kereszttábla értelmezését és további mutatók értelmezhetők, ha a kutatási téma kontextusában meghatározzuk, hogy logikailag melyik a függő és melyik a független változó. Ez a kutatási kérdések többségében nyilvánvaló, értelemszerű. Példánkban a nem változó a független, a nemi hovatartozást nem befolyásolja a GÓBÉ termék vásárlásának indoka.
Ennek megfelelően tegyük a független változót a kereszttábla sorába (Row), és a Cells ablakban a tényleges gyakoriság (Observed) mellé csak a sorszázalékot (Row) és az adjusztált reziduumot (Adjusted standardized) kérjük.

18. táblázat. Kereszttábla: kontingenciatáblázat

k_1 Neme * k_15 Miért vásárolnak Góbé terméket? Crosstabulation

			k_15 Miért vásárolnak Góbé terméket?					
			1 jobb minőségű	2 olcsóbb	3 bio alapanyagokból készül	4 ezzel támogatom a helyi termelőket	5 mert ez saját, székely termék	Total
k_1 Neme	1 Nő	Count	84	10	36	48	35	213
		% within k_1 Neme	39.4%	4.7%	16.9%	22.5%	16.4%	100.0%
		Adjusted Residual	1.8	2.7	1.7	-3.0	-1.3	
	2 Férfi	Count	46	0	16	56	33	151
		% within k_1 Neme	30.5%	0.0%	10.6%	37.1%	21.9%	100.0%
		Adjusted Residual	-1.8	-2.7	-1.7	3.0	1.3	
Total		Count	130	10	52	104	68	364
		% within k_1 Neme	35.7%	2.7%	14.3%	28.6%	18.7%	100.0%

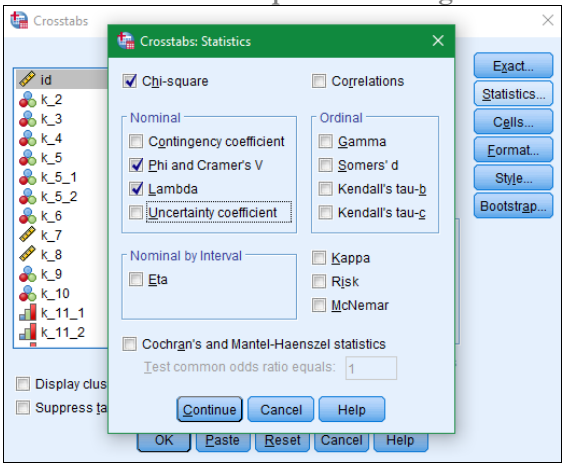
Ezzel az elrendezéssel célirányosabban meghatározhatjuk kereszttáblánk lényegi összefüggését: a férfiak szignifikánsan nagyobb arányban (37,1%) vásárolják a GÓBÉ terméket a helyi termelők támogatása miatt, mint a nők (22,5%). Fogalmazhatunk úgy is, hogy a férfiak **felülreprezentáltak**, a nők pedig **alulreprezentáltak** azok körében, akik a helyi termelők támogatása miatt vásárolják a GÓBÉ terméket.

Nyilvánvalóan vannak olyan kutatási kérdések, amikor nem értelemszerű, hogy melyik a függő és melyik a független változó. Például az interjúalanyok különböző attitűdjeire, véleményére vonatkozó változók. Ilyenkor is a kereszttábla statisztikai próbája, a khi-négyzet próba ugyanúgy működik, de a sor- és oszlopszázalékok alapos vizsgálatára van szükség a logikus értelmezés megfogalmazásához.

A kapcsolat szorossága

A kapcsolat erősségének a mutatóit a Statistics ablakban állíthatjuk be.

20. ábra. Kereszttábla: a kapcsolat erősségének mutatói



A Phi és Cramer V mutatókat egyszerre lehet beállítani, de általánosabb használhatósága miatt csak a Cramer V-t értelmezzük.[29]

19. táblázat. Kereszttábla: a Cramer V-mutató

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.231	.001
	Cramer's V	.231	.001
N of Valid Cases		364	

A Cramer V-féle asszociációs mérőszám 0,231-es értéke gyenge kapcsolatot jelez.

20. táblázat. Kereszttábla: a kapcsolat erősségének mutatói

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	.047	.046	1.011	.312
		k_15 Miért vásárolnak Góbé terméket? Dependent	.043	.042	.991	.321
		k_1 Neme Dependent	.053	.066	.785	.432
	Goodman and Kruskal tau	k_15 Miért vásárolnak Góbé terméket? Dependent	.013	.007		.001[c]
		k_1 Neme Dependent	.054	.019		.001[c]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

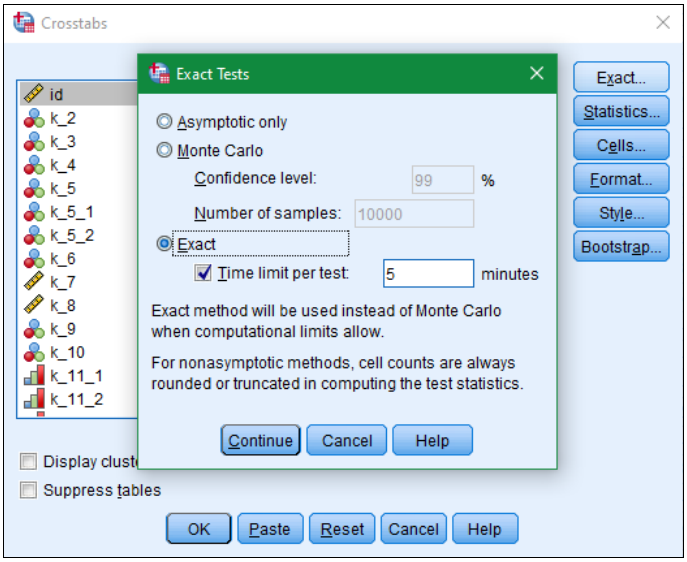
A lambda százalékos formában mutatja, hogy a független változó milyen mértékben magyarázza a függő változó értékét. Esetünkben a nem változó ismerete 4,3%-ban magyarázza, hogy valaki miért vásárol GÓBÉ terméket.

Kis minták esetén: a Fisher-teszt

Említettük a korábbiakban, hogy amennyiben nem teljesül az a feltétel, hogy a kereszttábla összes cellájának 20%-ában lehet a gyakoriság kisebb, mint 5, akkor alkalmazzuk a Fisher-tesztet. 2x2-es kereszttáblák esetében automatikusan megjelenik az Output-ban a Fisher's Exact Test értéke és szignifikanciaszintje, nagyobb kereszttáblák esetén – az eddig bemutatott beállítások mellett – a kereszttábla Exact ablakában választjuk az Exact opciót, meghagyva a Time limit per test: 5 minutes alapbeállítást.

[29] Ugyancsak nem részletezzük a kontingencia-együtthatót és a bizonytalansági együtthatót.

21. ábra. Kereszttábla: a Fischer-teszt beállítása



Eredmények értelmezése

21. táblázat. Kereszttábla: a Fischer-teszt szignifikanciája

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	19.479[a]	4	.001	.000		
Likelihood Ratio	23.091	4	.000	.000		
Fisher's Exact Test	20.255			.000		
Linear-by-Linear Association	8.192[b]	1	.004	.005	.002	.000
N of Valid Cases	364					

a. 1 cells (10.0%) have expected count less than 5. The minimum expected count is 4.15.

b. The standardized statistic is 2.862.

A Fisher's Exact Test szignifikanciaszintje alapján is kijelenthetjük, hogy a két változó között szignifikáns kapcsolat van.

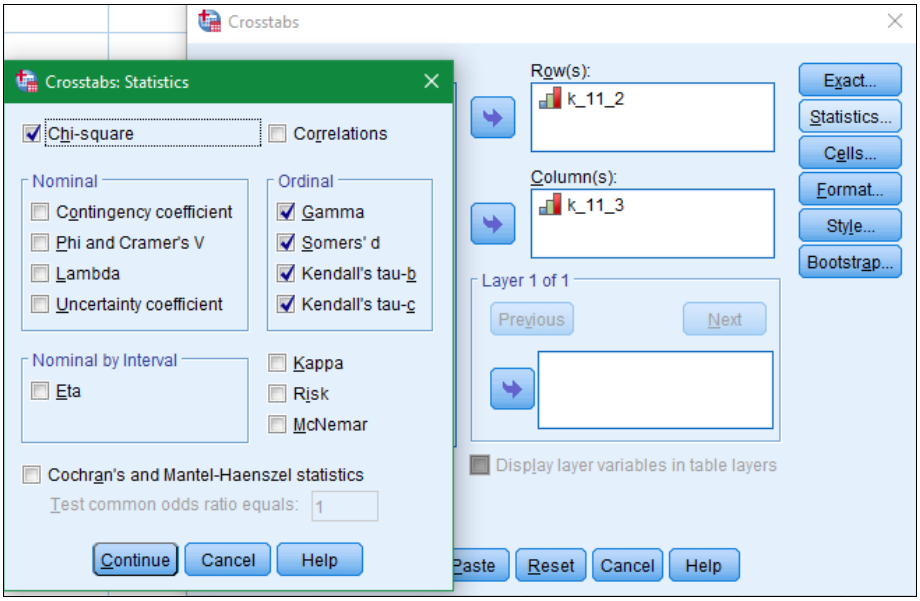
4.1.2 Ordinális változók közötti kereszttábla

Az ordinális változók közötti kapcsolatok a nominális változókhoz képest egy további információt hordoznak, a **változó értékeinek, kategóriáinak a rangsorát**.

Folytatva a GOBE2012.sav adatfájl elemzését, találunk hét, Likert-skálával mért változót, amelyek az élelmiszervásárlást meghatározó, a fogyasztói magatartást leíró szempontok fontosságára vonatkoznak. Vizsgáljuk meg ezek közül, hogy a primer kutatásunkban igazolható-e a fogyasztói magatartás tudományterületén törvényszerűségnek számító kapcsolat a minőség és a márka megítélése között

Analyze → Descriptive Statistics → Crosstabs…

22. ábra. Ordinális változók közötti kereszttábla beállítása



A Cells-ben ezúttal nem lesz szükségünk a sor- és oszlopszázalékokra, a tényleges elemszám (Observed) mellett csak az adjusztált reziduumot (Adj. standardized) állítjuk be. A kereszttábla Statistics ablakában az ordinális változókra vonatkozó mutatókat (Gamma, Somers'd, Kendall's tau-b, Kendall's tau-c) jelöljük.

Eredmények értelmezése

A Khi-négyzet teszt eredményét ezúttal nem ábrázoltuk, de a Pearson Khi-négyzet mutató szignifikanciaértéke, 0.000 egyértelművé teszi, hogy van szignifikáns kapcsolat a két változó között. A kereszttáblában (22. ábra) pedig azt látjuk, hogy a mátrix főátlójában és annak szomszédos celláiban találunk abszolút értékben 2-nél nagyobb vagy egyenlő adjusztált reziduum értéket.

22. táblázat. Ordinális változók kontingenciatáblája

k_11_2 Minőség fontossága * k_11_3 Márka fontossága Crosstabulation

			k_11_3 Márka fontossága					
			1	2	3	4	5	Total
k_11_2 Minőség fontossága	1	Count	0	2	0	0	1	3
		% within k_11_2 Minőség fontossága	0.0%	66.7%	0.0%	0.0%	33.3%	100.0%
		Adjusted Residual	-.1	6.8	-.6	-1.2	-.8	
	2	Count	1	0	1	0	1	3
		% within k_11_2 Minőség fontossága	33.3%	0.0%	33.3%	0.0%	33.3%	100.0%
		Adjusted Residual	7.7	-.3	1.3	-1.2	-.8	
	3	Count	0	3	9	7	2	21
		% within k_11_2 Minőség fontossága	0.0%	14.3%	42.9%	33.3%	9.5%	100.0%
		Adjusted Residual	-.4	3.3	5.1	.2	-4.4	
	4	Count	1	1	14	28	25	69
		% within k_11_2 Minőség fontossága	1.4%	1.4%	20.3%	40.6%	36.2%	100.0%
		Adjusted Residual	1.1	-.7	3.1	1.9	-3.6	
	5	Count	0	4	13	78	173	268
		% within k_11_2 Minőség fontossága	0.0%	1.5%	4.9%	29.1%	64.6%	100.0%
		Adjusted Residual	-2.4	-2.4	-5.6	-1.3	5.8	
Total		Count	2	10	37	113	202	364
		% within k_11_2 Minőség fontossága	0.5%	2.7%	10.2%	31.0%	55.5%	100.0%

Az eddig tanultak alapján is meg tudjuk fogalmazni, hogy a minőség fontosságának megítélése együtt jár, egyenesen arányos a márka fontosságának a megítélésével. Nézzük, hogy milyen további információt nyerünk az ordinális változók kereszttáblájának statisztikáiból. A következő táblázatokban találjuk azokat a mutatókat, **amelyek a két ordinális változó rangsorának hasonlóságát mutatják**.

23. táblázat. Kereszttábla: ordinális változók kapcsolati erősségének mutatói

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Kendall's tau-b	.353	.048	6.568	.000
	Kendall's tau-c	.218	.033	6.568	.000
	Gamma	.603	.064	6.568	.000
N of Valid Cases		364			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

A fenti táblázatban a **Kendall's tau-b mutató** szimmetrikus kereszttáblák esetén értelmezhető, adott esetben ilyen 5x5-ös táblánk van, a **Kendall's tau-c** mutató nem szimmetrikus kereszttábláknál alkalmazható. Ezek a mutatók a két változó

értékeinek sorrendisége közötti hasonlóságot vizsgálja, értékei -1 és +1 között változnak, ahol az 1 azt jelenti, hogy a két változó értékei sorrendje minden esetben megegyeznek, -1 esetén pedig a lehető legnagyobb mértékben különböznek.

A **Gamma** mutató a két változó közötti kapcsolat erősségét fejezi ki – hasonlóképp a numerikus változók közötti korrelációs együtthatóhoz –, ahol a 0 érték a teljes függetlenséget, az 1 pedig a teljes egyezőséget jelenti a két változó között. A Gamma alkalmazható szimmetrikus és nem szimmetrikus kereszttáblák esetén is. Ezek a lényegesen eltérő számítási módok magyarázzák táblázatunkban a Kendall'tau-b és a Gamma mutatók értékei közötti jelentős különbséget.

A Kendall's tau mutatókhoz hasonlóan a **Sommers` delta** is -1 és 1 értéktartománnyal rendelkezik, ahol az 1-hez közeli érték a két változó értékeinek hasonló sorrendjét jelzi. Az eddigiekkel szemben a Somers`delta mutató egyedisége abban áll, hogy **a két ordinális változó közül az egyik értelmezhető függő, a másik pedig független változóként, és számszerűsíti a független változó hatását a függő változóra.**

24. táblázat. Kereszttábla: ordinális változók kapcsolati erősségének mutatói

Directional Measures

			Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Ordinal by Ordinal	Somers' d	Symmetric	.348	.048	6.568	.000
		k_11_2 Minőség fontossága Dependent	.299	.043	6.568	.000
		k_11_3 Márka fontossága Dependent	.417	.057	6.568	.000

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Amint a fenti (24.) táblázatban láthatjuk, a módszertani teljesség kedvéért és zavarunk fokozása érdekében az SPSS kiszámolja a Somers'd értékét úgy is, hogy szimmetrikusnak feltételezi a két változó közötti kapcsolatot. Ebben az esetben ugyanúgy értelmezendő, mint a Kendall's tau-b, példánkban a két mutató értéke nagyon közel áll egymáshoz (0.353 és 0.348).

A Somers`delta jelentősége azonban abban áll, hogy **számszerűsíti a független változó hatását a függő változór**a, példánk eredményét úgy is értelmezhetjük, hogy a márka fontosságának megítélését (függő változó) 41,7%-ban magyarázza a minőség fontosságának megítélése.

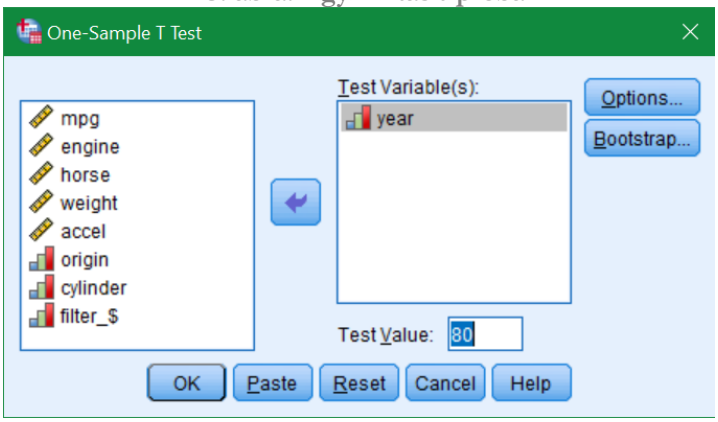
4.2 Egymintás t-próba

Az egymintás t-próbával nyitjuk azoknak a módszereknek a sorát, amelyek a metrikus változók vizsgálatát teszik lehetővé. Ezt a módszert akkor alkalmazzuk, ha **egy mintabeli numerikus változó átlagát akarjuk összehasonlítani egy külső forrásból származó átlaggal, értékkel**.

Példa: nyissuk meg a cars.sav adatfájlt az SPSS példa fájljai közül[30], amely autókra vonatkozó adatokat tartalmaz. Az autók ismérvei között van a gyártási év, ami egy metrikus változóban (*year*) van megjelenítve. Vizsgáljuk meg, hogy mennyi az autók átlagéletkora, pontosabban mennyi az átlagos gyártási év és hogy ez szignifikánsan különbözik-e az 1980-tól?

Analyze→Compare Means→One-Sample T Test

23. ábra. Egymintás t-próba



A bal oldali változólistából kiválasztjuk a year változót és a Test Value mezőbe beírjuk a 80-as értéket.[31]

Eredmények értelmezése

Az első eredménytábla az elemzésbe bevont változót (YEAR) és főbb statisztikáit mutatja: a változó valós értékeinek gyakoriságát (N), az átlagot, a szórást és a standard hibát.

[30] Ammenyiben az SPSS verziónk nem tartalmazza a cars.sav adatfájlt, szintén megtalálhatjuk a http://web.csik.sapientia.ro/kutatasmodszertan oldalon.

[31] Az autók gyártási évét tartalmazó *year* változó csak a két utolsó számjegyet tartalmazza, ezért írunk 80-at az 1980 helyett. Lényeges, hogy az átlaggal összehasonlítani kívánt érték ugyanolyan mértékegységben és formátumban legyen, mint a vizsgált változó.

25.táblázat. Egymintás t-próba: a bevont változó

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
year Model Year (modulo 100)	406	75.75	5.307	.263

A második táblázatban a t-próba eredményeit találjuk.

26. táblázat. Egymintás t-próba: a bevont változó alapstatisztikái

One-Sample Test

Test Value = 80

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
year Model Year (modulo 100)	-16.140	405	.000	-4.251	-4.77	-3.73

A szignifikancia szint (.000) alapján megállapítható, hogy az átlag (75.7) és a tesztelt érték (80.0) közötti különbség szignifikánsan különbözik nullától, vagyis az autóparkunk átlagos gyártási éve korábbi, mint 1980.

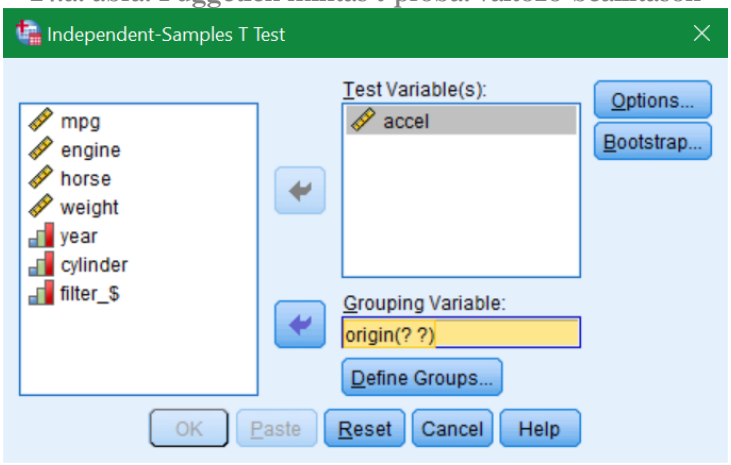
4.3 Független mintás t-próba

A független mintás t-próbával egy nominális vagy ordinális és egy vagy több numerikus változó közötti kapcsolatot vizsgálhatjuk. A módszer **a teljes mintát (adattáblát) a nominális/ordinális változó kategóriái szerint két, egymástól független részmintára bontja, majd összehasonlítja a vizsgált numerikus változó részmintabeli átlagait.**

Példa: folytassuk a cars.sav adatainak elemzését. Vizsgáljuk meg azt a hipotézist, hogy a mintánkban található amerikai autók gyorsulása jobb, mint az európai autóké. Az autók származásának ez a két kategóriája (amerikai, európai) egymástól független részmintákra osztja a teljes mintát, és e két részmintában található megfigyelési egységek (autók) egyik ismérvét (gyorsulását) kívánjuk összehasonlítani.

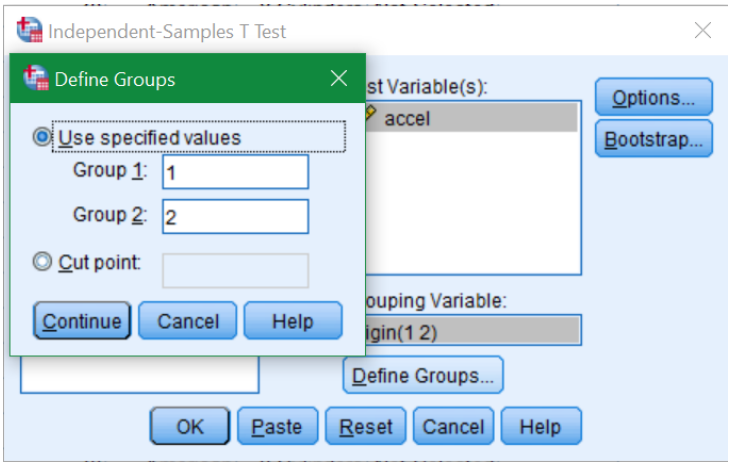
Analyze→Compare Means→Independent-Samples T Test

24.a. ábra. Független mintás t-próba: változó-beállítások



A Test Variable(s)[32] mezőbe a numerikus változót visszük be, adott esetben a gyorsulást *(accel)*. A Grouping Variable mezőben pedig a nominális változót definiáljuk, példánkban az autók eredetét *(origin)*. E változó neve után megjelenő kérdőjelek, illetve az OK gomb inaktivitása jelzi, hogy még további beállításokra van szükségünk. A Define Groups…-ra kattintva beállíthatjuk, hogy a nominális változó melyik két értéke határozza meg a két részmintát.

24.b. ábra. Független mintás t-próba: változó-beállítások



[32] A többes szám jelzi, hogy akár több numerikus változó átlagait egyszerre is vizsgálhatjuk.

Eredmények értelmezése

Az első táblázatban az alapvető statisztikákat találjuk, mindkét származási helyre vonatkozóan: N – elemszám, az átlagot, a szórást és az átlag standard hibáját.

27. táblázat. Független mintás t-próba: a változók átlaga és szórása

Group Statistics

	origin Country of Origin	N	Mean	Std. Deviation	Std. Error Mean
accel Time to Accelerate from 0 to 60 mph (sec)	1 American	253	14.93	2.801	.176
	2 European	73	16.82	3.011	.352

Az átlagok alapján megállapíthatjuk, hogy az amerikai autók gyorsulása jobb, kevesebb másodperc alatt érik el a 60 mérföld per órás (kb. 100 km/h) sebességet. Kérdés, hogy az átlagok közötti nem túl nagynak tűnő különbség statisztikailag szignifikáns-e? A második táblázatban (28.) erre találunk választ.

28. táblázat. Független mintás t-próba: a t-próba eredményei

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
accel Time to Accelerate from 0 to 60 mph (sec)	Equal variances assumed	.907	.342	-5.002	324	.000	-1.893	.379	-2.638	-1.149
	Equal variances not assumed			-4.806	110.482	.000	-1.893	.394	-2.674	-1.113

Az eredmények első két oszlopában levő Levene's teszt F próbája a **varianciák egyezőségére** vonatkozik, a táblázat többi részében (t-test for Equality of Means) pedig két sorban találjuk az **átlagok egyezőségére** vonatkozó t-tesztek eredményeit. A felső sorban a **standard t-teszt**, az alsó sorban a **Welch t-teszt** vizsgálja az átlagok egyezőségének nullhipotézisét.

A felső sor (Equal variances assumed) a két részminta egyenlő varianciája, az alsó pedig eltérő varianciák esetén érvényes.

Az első lépésben tehát a Levene's teszt alapján vizsgáljuk a varianciák egyezőségét. Ha a szignifikancia szint kisebb, mint .05 ($p < .05$), akkor elutasítjuk, ha nagyobb ($p > .05$), akkor elfogadjuk a varianciák homogenitására (egyezőségére) vonatkozó hipotézist. Példánkban az érték (Sig. .342) nagyobb, mint .05, ezért kijelenthetjük a varianciák egyezőségét. A táblázat második felében (t-test for Equality of Means) a felső sor t-tesztje érvényes az egyező varianciák miatt (Equal variances assumed), és a szignifikancia szint értéke (.000) alapján a nullhipotézist elvethetjük, vagyis a két átlag szignifikánsan különbözik.

4.4 Egyutas varianciaanalízis (ANOVA)

Az egyutas varianciaanalízis (One-way ANOVA[33]) **két vagy több, egymástól független csoport átlagainak szignifikáns különbségeinek a meghatározására alkalmas módszer**. Megkülönböztetjük a csoportképző, független változót és az átlagolandó, metrikus függő változót. A változók közötti kapcsolatok vizsgálatára alkalmas módszereket bemutató táblázatunk (14. táblázat) alapján úgy is fogalmazhatunk, hogy az ANOVA egy nominális típusú változó minden kategóriájára kiszámolja és összehasonlítja a metrikus változó átlagait.

Gyakorlati eredményeit tekintve annyiban különbözik a független mintás t-próbától, hogy nemcsak két kategória mentén hasonlítja össze az átlagokat, hanem a nominális változó valamennyi kategóriájára kiszámítja azokat. Az átlagok egyezőségére vonatkozó hipotézis tesztelésére azonban nem a t-próbát, hanem az F-próbát alkalmazza, aminek kiváló tulajdonsága, hogy elég nagy minták (>100 eset) esetén robusztus becslést[34] eredményez.
Ezeknek az előnyöknek köszönhetően a kutatók az átlagok összehasonlítására legtöbbször ANOVA-t használnak, kis (rész)minták esetén folyamodnak a t-próbához.

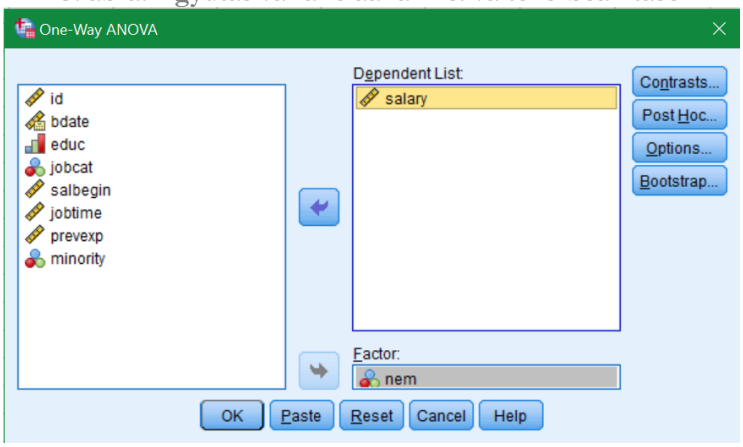
Példa. Példaként használjuk az SPSS példa adattáblái közül az employee data.sav fájlt. Ez a személyzeti nyilvántartás egy cég valamennyi alkalmazottjára (474) vonatkozó néhány fontos jellemzőt tartalmaz. Kutatási kérdésünk, hogy van-e szignifikáns különbség a férfiak és nők fizetésének nagysága között.

[33] A magyar nyelvű szakirodalomban az egyszempontú, az egytényezős varianciaanalízis, illetve egyszerűen csak ANOVA elnevezésekkel is találkozhatunk.
[34] Robusztus becslés esetén a statisztikai próba korlátozó feltételeinek részleges teljesülése nem rontja lényegesen az eredményt.

4.4.1 A varianciaanalízis beállításai

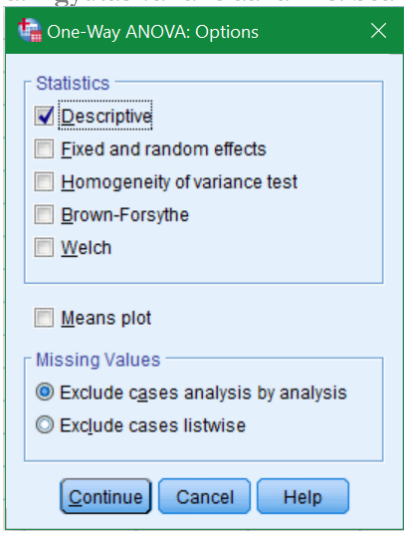
Analyze→Compare Means→One-way ANOVA

25. ábra. Egyutas varianciaanalízis: változó-beállítások



A Factor mezőbe tesszük a csoportképző, nominális változót, a Dependent List-be pedig az átlagolandó, metrikus változót vagy változókat. Ha egyszerre több metrikus változó átlagára vagyunk kíváncsiak, akkor többet is bejelölhetünk, az eredményeket egymás után jeleníti meg az SPSS Output.

26. ábra. Egyutas varianciaanalízis: beállítások



Az Options-ban kell beállítanunk a Descriptives funkciót, különben az átlagok nem jelennek meg, csak az F-próba és annak szignifikanciaszintje, majd Continue és OK.

Eredmények értelmezése

A Descriptives táblázatban találjuk a numerikus változó különböző statisztikáit, a nominális változó valamennyi kategóriája kiszámolva.

29. táblázat. Egyutas varianciaanalízis: a numerikus változó csoportonkénti statisztikái

Descriptives

salary Current Salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1 Female	216	$26,031.92	$7,558.021	$514.258	$25,018.29	$27,045.55	$15,750	$58,125
2 Male	258	$41,441.78	$19,499.214	$1,213.968	$39,051.19	$43,832.37	$19,650	$135,000
Total	474	$34,419.57	$17,075.661	$784.311	$32,878.40	$35,960.73	$15,750	$135,000

A táblázat első oszlopában láthatjuk a nominális változó (nem) két kategóriájának a kódjait és a címkéit, a második oszlopban (N) pedig a két kategóriába tartozó esetek számát. A Means oszlopban találjuk a nemek szerinti részátlagokat, illetve a teljes átlagát (Total). Megállapíthatjuk, hogy a nők (éves) fizetése átlagosan 26,031 dollár, a férfiaké pedig 41,441 dollár. A teljes minta átlaga (34,419 USD) nyilvánvalóan a két kategória részátlagainak súlyozott átlaga.

A két nem közötti bérkülönbség mértéke felülmúlja a legpesszimistább várakozásainkat is, de a tudományos kutatás módszertanát követve meg kell vizsgálnunk, hogy ez a különbség statisztikailag szignifikáns-e. Erre a kérdésre válaszol az Output következő, ANOVA feliratú táblázata:

30. táblázat. Egyutas varianciaanalízis: szignifikanciaszint

ANOVA

salary Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.792E+10	1	2.792E+10	119.798	.000
Within Groups	1.100E+11	472	233046530.5		
Total	1.379E+11	473			

Az F-próba értéke szignifikanciaszintjének vizsgálatakor ugyanaz az aranyszabály érvényesül, mint például a kereszttábla-elemzés Chi-négyzet próbájánál; a 0.05-nél kisebb érték esetén megállapíthatjuk, hogy szignifikáns különbség van a kategóriák részátlagai között.

Az átlagok után tekintsük át a Descriptives táblázat valamennyi eredményét:

- N – abszolút gyakoriság, azt mutatja, hogy az adott kategória hány értékéből számoltuk a részátlagot.
- Mean – átlag.
- Std. Deviation – szórás.
- Std. Error – standard hiba. Önmagában nem szükséges értelmeznünk, a konfidencia-intervallum meghatározásához szükséges.
- 95% Confidence Interval for Mean (95%-os konfidenciaintervallum) – az átlag körüli intervallum, egyenlő az átlag ±2× standard hiba (S.E.) mértékével. Ha a részátlagok szignifikánsan különböznek, akkor a konfidenciaintervallumok között nincs átfedés.
- Minimum és Maximum – a metrikus változó minimumát és maximumát mutatja az adott kategória esetében. Hasznos ezek vizsgálata, mivel így azonnal kiderül, ha valamilyen szélsőséges érték is bekerült az átlagba. Például ha az adott kérdésre nem válaszolók (NT/NV) 99-es kódját nem zártuk ki („nem tettük missing-re"), akkor jelentősen torzítjuk az átlagot.

Megállapítottuk eddig tehát az F-próba segítségével, hogy a két nem átlagbérei között szignifikáns különbség van. Kérdés, hogy kettőnél több csoport, a nominális változó több kategóriája esetén hogyan járunk el? Az ANOVA-tábla szignifikanciaszintje csak azt mutatja, hogy az átlagok között van szignifikáns eltérés, de nem pontosítja, hogy mely kategória átlagai között. A részátlagokat párosával összehasonlító módszereket post-hoc teszteknek nevezzük.
A kérdés vizsgálatához ne a *nem* változót vizsgáljuk (bár trendi lenne), hanem a *beosztás* (jobcat) változó három kategóriája (manager, szellemi, fizikai) mentén hasonlítsuk össze az átlagbéreket.

Analyze→Compare Means→One-way ANOVA

A Factor mezőben kicseréljük a nominális változót a *nemről* a *jobcat*-re, a Dependent List-ben meghagyjuk a *salary*-t, az Options-ban pedig a Descriptives. A Post Hoc gombra kattintva, a tesztek zavarba ejtő bőségével találjuk.

27. ábra. Egyutas varianciaanalízis: Post Hoc-tesztek beállítása

Válasszuk az első, nem túl biztató nevű LSD opciót, ami az ablakban felsorolt tesztek közül a legengedékenyebbnek tartott, olyan mértékű különbséget is szignifikánsnak tart, amit a többi próba még nem. A másik, leggyakrabban használt Tukey-próbát is bejelöljük. Continue, és ha a korábbi beállításaink még megvannak, akkor jöhet az OK.

Eredmények értelmezése

31. táblázat. Egyutas varianciaanalízis: csoportátlagok összehasonlítása

Descriptives

salary Current Salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1 szellemi	363	$27,838.54	$7,567.995	$397.217	$27,057.40	$28,619.68	$15,750	$80,000
2 fizikai	27	$30,938.89	$2,114.616	$406.958	$30,102.37	$31,775.40	$24,300	$35,250
3 Manager	84	$63,977.80	$18,244.776	$1,990.668	$60,018.44	$67,937.16	$34,410	$135,000
Total	474	$34,419.57	$17,075.661	$784.311	$32,878.40	$35,960.73	$15,750	$135,000

A menedzserek átlagfizetése (Mean oszlop) jóval meghaladja a másik két kategóriáét, okkal feltételezzük az átlagok közötti különbségek szignifikáns nagyságát, de arra a kérdésre, hogy a szellemi vagy a fizikai beosztottak keresnek-e többet, csak a következő táblázat vizsgálata alapján tudunk statisztikailag igazolt választ adni. A Multiple Comparisons nevű táblában a kategóriák részátlagai közötti különbségeket és azok szignifikanciaszintjeit (Sig.) találjuk, külön a Tukey HSD és külön az LSD-próba alapján.

32. táblázat. Egyutas varianciaanalízis: Post Hoc-tesztek értelmezése

Multiple Comparisons

Dependent Variable: salary Current Salary

	(I) jobcat beosztas	(J) jobcat beosztas	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1 szellemi	2 fizikai	-$3,100.349	$2,023.760	.277	-$7,858.50	$1,657.80
		3 Manager	-$36,139.258*	$1,228.352	.000	-$39,027.29	-$33,251.22
	2 fizikai	1 szellemi	$3,100.349	$2,023.760	.277	-$1,657.80	$7,858.50
		3 Manager	-$33,038.909*	$2,244.409	.000	-$38,315.84	-$27,761.98
	3 Manager	1 szellemi	$36,139.258*	$1,228.352	.000	$33,251.22	$39,027.29
		2 fizikai	$33,038.909*	$2,244.409	.000	$27,761.98	$38,315.84
LSD	1 szellemi	2 fizikai	-$3,100.349	$2,023.760	.126	-$7,077.06	$876.37
		3 Manager	-$36,139.258*	$1,228.352	.000	-$38,552.99	-$33,725.53
	2 fizikai	1 szellemi	$3,100.349	$2,023.760	.126	-$876.37	$7,077.06
		3 Manager	-$33,038.909*	$2,244.409	.000	-$37,449.20	-$28,628.62
	3 Manager	1 szellemi	$36,139.258*	$1,228.352	.000	$33,725.53	$38,552.99
		2 fizikai	$33,038.909*	$2,244.409	.000	$28,628.62	$37,449.20

*. The mean difference is significant at the 0.05 level.

Mindkét statisztikai próba szignifikanciaszintje (Sig. oszlop) alapján megállapíthatjuk, hogy a menedzserek bére szignifikánsan különbözik a másik két beosztás, kategória átlagbérétől, de a szellemi és fizikai munkások bére között már nincs szignifikáns különbség.

4.4.2 A varianciaanalízis alkalmazásának feltételei

Az ANOVA alkalmazásának van négy fontos korlátozó feltétele:

1. feltétel. A metrikus változónak (a modell függő változója) megközelítőleg **normális eloszlásúnak** kell lennie a nominális változó (a modell független változója) mindegyik kategóriájára. Az előbbiekben bemutatott példánkban a fizetés normál eloszlású kell legyen mind a férfiak, mind a nők körében.

Ennek ellenőrzésére szolgáló illeszkedés-vizsgálatra az SPSS a Shapiro–Wilk és a Kolmogorov–Smirnov-tesztet használja, amelyek leírását a *Normalitásvizsgálat* fejezetben találjuk meg. Csendben megjegyezzük, hogy ennek a feltételnek a nem teljesülése, vagyis a függő, metrikus változó nem normális eloszlása jóval gyakoribb probléma, mint ahogy azt a módszertani könyvek jelzik. A szakirodalom jó része, főképp gyakorlati, üzleti kutatók (pl. Sajtos—Mitev, 2007) szerint nem történik nagy probléma, ha a gyakorlatban ettől a matematikai-statisztikai feltételezés ellenőrzésétől eltekintünk. E kellő elbizonytalanítás után joggal kérdezhetjük, hogy mégis mit tehetünk, ha nem fogadjuk el a szakirodalomban gyakran előforduló, de nem túl szakszerű „nem nagy probléma" érvelést.

Két lehetőségünk van:

– Átalakítjuk, **transzformáljuk a változónkat** különböző algoritmusokkal úgy, hogy normál eloszlású legyen, és ezek a transzformációk ne befolyásolják a lényegi kérdést: az átlagok közötti különbségek szignifikanciájának a vizsgálatát (lásd Normalitásvizsgálat alfejezetet).

– A **nem parametrikus Kruskal–Wallis-**tesztet alkalmazzuk, amely nem feltételezi a normál eloszlást. E módszer leírását jelen jegyzet nem tartalmazza.

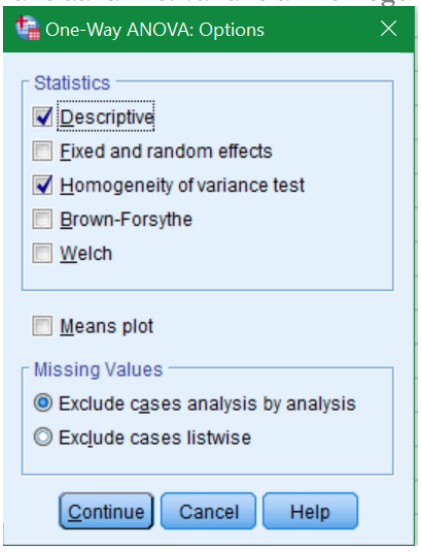
2. feltétel. A második korlátozó feltétel a varianciák homogenitására vonatkozik, vagyis a nominális változó kategóriái által képzett **csoportok szórásának is megközelítőleg azonosnak kell lenniük**. Ennek ellenőrzésére a független mintás t-próba bemutatásánál már említett Levene-teszt áll rendelkezésünkre, amelynek nullhipotézise az, hogy a két szórás azonos.

Futtassuk még egyszer a varianciaanalízist a *salary* és a *jobcat* változókkal.

Analyze→Compare Means→One-way ANOVA

Az eddig tanult beállításokon túl az Options-ban jelöljük be a Homogeneity of variance test opciót.

28. ábra. Egyutas varianciaanalízis: varianciák homogenitásának ellenőrzése



Eredmények értelmezése

33. táblázat. Egyutas varianciaanalízis: varianciák homogenitásának ellenőrzése

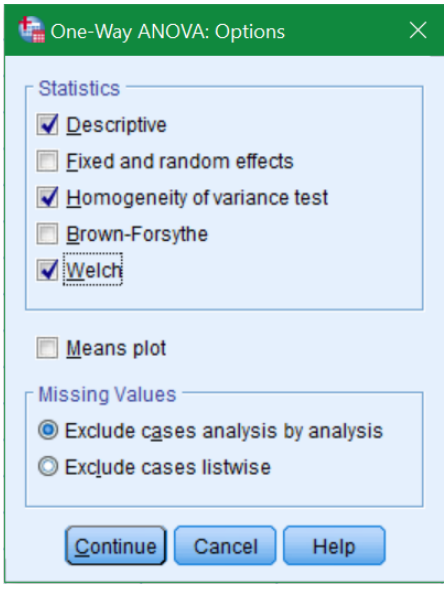
Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
salary Current Salary	Based on Mean	59.733	2	471	.000
	Based on Median	51.189	2	471	.000
	Based on Median and with adjusted df	51.189	2	240.176	.000
	Based on trimmed mean	56.201	2	471	.000

A Levene-teszt nullhipotézise a varianciák egyezősége, vagyis ha a szignifikanciaszint kisebb vagy egyenlő, mint 0.05, akkor nem teljesül a varianciák homogenitásának feltétele.

Ebben az esetben nem a standard ANOVA-t, hanem egy alternatív tesztet, a **Welch-ANOVA**-t, a csoportátlagok közötti különbségek meghatározására pedig egy alternatív post-hoc-tesztet, **a Games–Howell-**tesztet alkalmazhatjuk. Futtassuk újra az ANOVA-t.

29. ábra. Robusztus varianciaanalízis: a Welch-teszt



Eredmények értelmezése

34. táblázat. Robusztus varianciaanalízis: a Welch-teszt eredménye

Robust Tests of Equality of Means

salary Current Salary

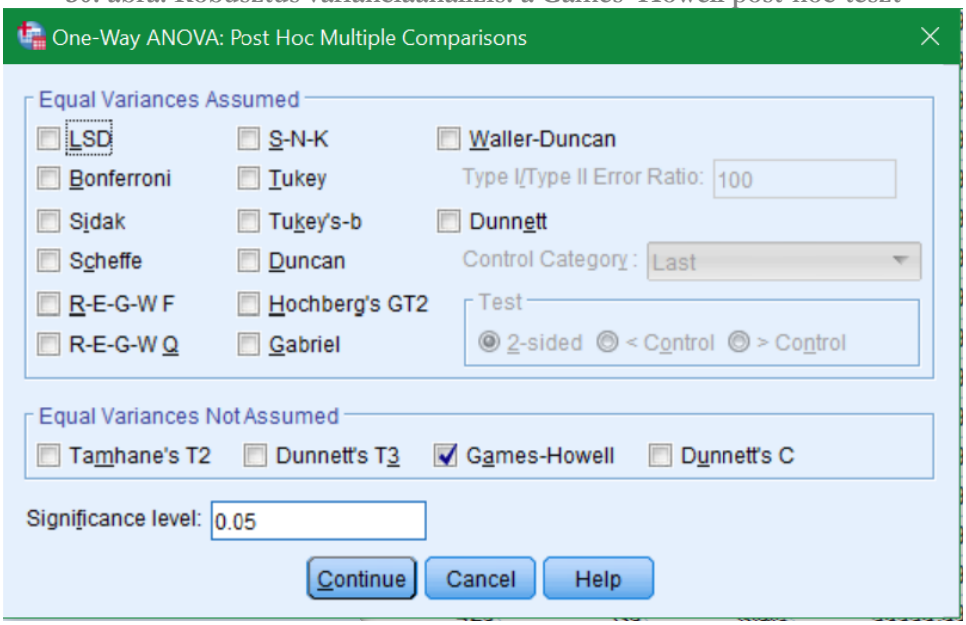
	Statistic[a]	df1	df2	Sig.
Welch	162.200	2	117.312	.000

a. Asymptotically F distributed.

A Welch-teszt statisztikailag szignifikáns (<0.05), vagyis legalább két csoport átlaga szignifikánsan különbözik. Ahhoz, hogy mindhárom munkahelyi beosztás átlagbérei közötti különbségeinek statisztikai szignifikanciáját megállapíthassuk, egy újabb post-hoc-tesztet alkalmazunk, a **Games–Howell-tesztet**.

Újrafuttatjuk az ANOVA-t az eddigi beállításokkal, kivéve, hogy a Post Hoc ablakában ezúttal a Games–Howell-tesztet jelöljük.

30. ábra. Robusztus varianciaanalízis: a Games–Howell post-hoc-teszt



Eredmények értelmezése

A Games–Howell-teszt eredményeit is úgy értelmezzük, mint a többi Post Hoc-tesztét, a szignifikanciaszint az átlagok egyezésének a valószínűségét mutatja.

35. táblázat. Robusztus varianciaanalízis: a Games–Howell post-hoc-teszt eredménye

Multiple Comparisons

Dependent Variable: salary Current Salary

Games-Howell

(I) jobcat beosztas	(J) jobcat beosztas	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 fizikai	2 szellemi	$3,100.349*	$568.679	.000	$1,745.88	$4,454.82
	3 manager	-$33,038.909*	$2,031.840	.000	-$37,881.37	-$28,196.45
2 szellemi	1 fizikai	-$3,100.349*	$568.679	.000	-$4,454.82	-$1,745.88
	3 manager	-$36,139.258*	$2,029.912	.000	-$40,977.01	-$31,301.51
3 manager	1 fizikai	$33,038.909*	$2,031.840	.000	$28,196.45	$37,881.37
	2 szellemi	$36,139.258*	$2,029.912	.000	$31,301.51	$40,977.01

*. The mean difference is significant at the 0.05 level.

A fizikai és szellemi munkások átlagbérei közötti különbséget vizsgálva rájövünk, hogy érdemes volt ez a további elemzés. Amíg nem vettük figyelembe az eltérő varianciákat, mindkét alkalmazott Post Hoc-teszt (LSD, Tukey HSD) a két beosztás átlagbéreinek egyezőségét mutatta, de a helyesen alkalmazott Games–Howell-teszt alapján kijelenthetjük, hogy a vizsgált cégnél a fizikai munkások szignifikánsan többet keresnek, mint a szellemi beosztottak.

3. Feltétel. Az előbbieknél jóval egyszerűbb feltétel, primer, keresztmetszeti kutatásoknál mondhatni értelemszerű, de a szakmai teljesség kedvéért megemlítjük a **megfigyelések függetlenségének** teljesülését. Vagyis ugyanaz a megfigyelési egység (pl. interjúalany) nem szerepelhet a nominális változó több kategóriájában, csak egyben. Általánosságban ajánlható, hogy egy megfigyelési egység csak egyszer szerepeljen az adattáblában, de kivételt képezhetnek olyan, könyvünkben nem tárgyalt speciális adatstruktúrák (pl. keresztmetszeti adatokból átalakított kvázi-longitudinális adattábla), amikor ez az általános feltétel nem teljesülhet. Ennek a feltételnek a teljesülését a kutatás, a mintavétel tervezésekor biztosíthatjuk.

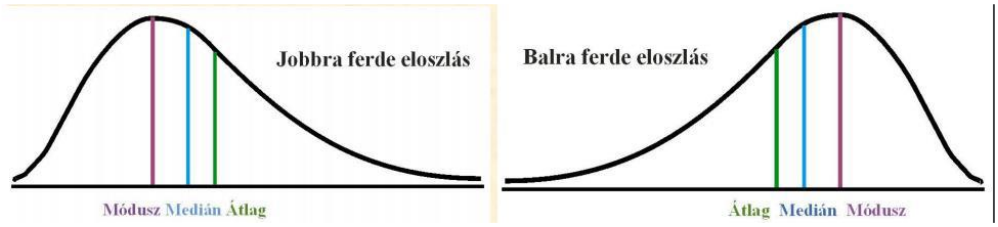
4. **Kiugró értékek kizárása**. A függő, metrikus változó olyan értékeit, amelyek jelentősen eltérnek (felfelé vagy lefelé) a változó többi értékétől, kiugró értékeknek (*outlier*) nevezzük. A kiugró értékek torzíthatják az ANOVA-teszt eredményeit, mert nagy befolyást gyakorolnak az adott csoport átlagára és szórására. Még jelentősebb ez a hatás, ha kisebb mintamérettel rendelkezünk.
A kiugró értékek kezelésének módjait a következő Normalitás-vizsgálat alfejezetben találjuk.

4.5 Normalitásvizsgálat

Normalitásvizsgálatnak nevezzük azt az eljárást, amely során megnézzük, hogy egy metrikus változó értékei **normál eloszlást követnek** vagy sem. Mint láttuk, több statisztikai módszer pontos működése feltételezi a magyarázott változó normál eloszlását (pl. ANOVA, t-próba, főkomponens-analízis). Közgazdasági, társadalmi jelenségek között sokkal ritkábban fordul elő a normál eloszlás, mint az szeretnénk, ezért a normalitásvizsgálatra, illetve a normál eloszlás hiányának kezelésére gyakran szüksége van a kutatónak. Ennek megfelelően részletesen tárgyaljuk ezt a kérdést.

A normál eloszlás **szimmetrikus**, ami azt jelenti, hogy az átlagtól ugyanakkora távolságra lévő értékek gyakorisága mindkét oldalon egyenlő, és a távolság növekedésével csökken.

31. ábra. Ferde eloszlások



Leggyakrabban a ferdeség okozza az eltérést a normál eloszláshoz képest, a hosszú, lapos, kis gyakoriságú részt a kiugró értékek alkotják. Láthatjuk (31. ábra), hogy az átlag, medián és módusz egymáshoz való viszonya összefügg az eloszlás formájával, a normál eloszlás legcsúcsosabb részénél, a legnagyobb gyakorisági értékeknél egybeesik az átlag, a módusz és a medián.

4.5.1 Grafikus módszerek

A grafikus módszerekkel képet kapunk a változó eloszlásáról és a normális eloszlásra való illeszkedéséről. Annak ellenére, hogy ezt nem egy tesztelhető mutatóval fejezi ki, mégis indokolt az elemzésünket a változó eloszlásának grafikus képének vizsgálatával kezdeni.

A **hisztogram** egy oszlopdiagram, amely a változó értékeit meghatározott osztályközökbe sorolja, és ezek gyakoriságát ábrázolja. Az oszlopok területe

arányos a gyakorisággal, a szélessége az osztályköz terjedelmét, magassága a gyakoriságot mutatja (lásd 33. ábra). Az SPSS automatikusan egyenlő terjedelmű osztályokat képez,[35] így az egyes oszlopok magassága a gyakoriságok arányát is érzékelteti. Akkor normális eloszlású a vizsgált változó, ha a hisztogram alakja megközelítőleg illeszkedik a haranggörbére.

A **Q-Q plot** (kvantilis-kvantilis ábra) a minta tapasztalati kvantiliseit veti össze az illesztett, azaz a standard normális eloszlás kvantiliseivel, a pontpárokat pedig ábrázolja (lásd 34. ábra). Az SPSS Q-Q ábráján minél jobban rásimulnak a pontok a 45 fokos egyenesre, annál jobban közelíti az eloszlás a normál eloszlást.
Az SPSS által alapból beállított **szár és levél** *(stem and leaf)* módszer régebb népszerű volt, de meglehetős bonyolultsága ellenére nem ad többlet információt a tanult módszerekhez képest, ezért mi nem tárgyaljuk.

4.5.2 Normalitástesztek az SPSS-ben

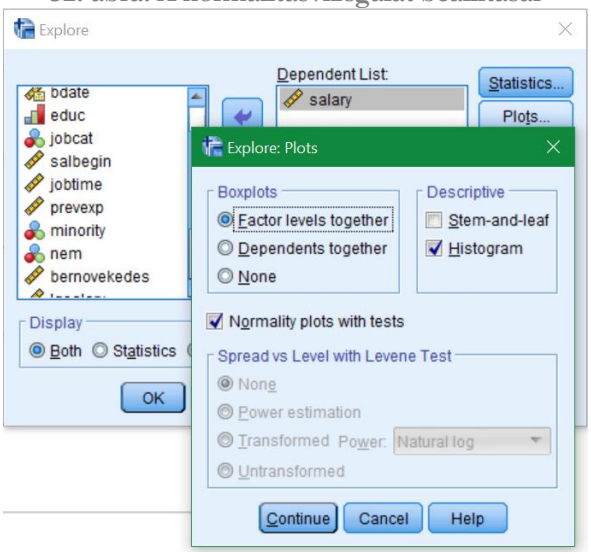
Az SPSS két, eltérő logikájú normalitástesztet futtat az EXPLORE paranccsal. A **Kolmogrov–Smirnov** illeszkedésvizsgálatát Lilliefors amerikai matematikus módosításával egészíti ki. A másik tesztet, **a Shapiro–Wilk**-próbát a legerősebbnek tartják más normalitástesztekkel összehasonlítva (Thode, 2002).

Példa. Vizsgáljuk meg az employee.sav adatfájl *salary* változójának eloszlását! Ismerve már a normál eloszlás sajátosságait, főképp a szimmetrikusságát, szkeptikusak vagyunk a példabeli cég fizetéseinek (*salary*) normál eloszlását illetően. Kevés tapasztalattal sem tartjuk valószínűnek, hogy egy cégnél megközelítően ugyanannyian és ugyanannyival keresnek az átlagnál kevesebbet, mint ahányan többet.

[35] Legalábbis a normalitásvizsgálatra vonatkozó EXPLORE-parancsban. A grafikonszerkesztő *(Chart Builder*) hisztogramjánál lehetőség van módosításra.

Analyze→Descriptive Statistics→Explore

32. ábra. A normalitásvizsgálat beállításai



Már rutinosan választjuk ki a *salary* változót a listából, majd a Plots ablakban eszközölünk néhány változtatást a 32. ábrának megfelelően. A Stem-and-leaf alapbeállítását kivesszük, cserébe bejelöljük a Histogram-ot és a Normality plots with tests opciókat. Ez utóbbi – a diszkrét megjelenése ellenére – nagyon fontos, ugyanis a Kolmogorov–Smirnov és a Shapiro–Wilk teszteket szolgáltatja. A Statistics ablakban válasszunk mindent is, a későbbiekben a kiugró értékek meghatározásánál szükségünk lesz a különböző statisztikákra.

Eredmények értelmezése

36. táblázat. Helyzet-, szóródás- és alakmutatók

Descriptives

			Statistic	Std. Error
salary Current Salary	Mean		$34,419.57	$784.311
	95% Confidence Interval for Mean	Lower Bound	$32,878.40	
		Upper Bound	$35,960.73	
	5% Trimmed Mean		$32,455.19	
	Median		$28,875.00	
	Variance		291578214.5	
	Std. Deviation		$17,075.661	
	Minimum		$15,750	
	Maximum		$135,000	
	Range		$119,250	
	Interquartile Range		$13,163	
	Skewness		2.125	.112
	Kurtosis		5.378	.224

A Descriptive tábla (36. táblázat) eredményei közül helyzetmutatókat (átlagtól a mediánig) és a szóródásmutatókat (varianciától a terjedelemig) már kedves ismerősként üdvözölhetjük, de újdonság az interkvartilis terjedelem *(Interquartile Range - IQR)*, a ferdeségi *(skewness)* és a csúcsossági *(kurtosis)* mutató.

Az **interkvartilis terjedelem** az első kvartilis (a kumulált gyakoriság 25%-a) és harmadik kvartilis (75%) közötti terjedelmet, a két kvartilis érték közötti különbséget (Q3-Q1) mutatja. Azt fejezi ki, hogy a változó értékeinek sorba rendezett középső 50%-a mekkora intervallumban szóródik. Grafikus megjelenítése a pénzügyekben, tőzsdei elemzéseknél gyakran használt dobozdiagram (36. ábra).

A **ferdeségi mutató** *(skewness)* azt fejezi ki, hogy az eloszlás milyen irányban és mértékben tér el a szimmetrikus eloszlástól. Többféle ferdeségi mutatót ismer a szakirodalom, az SPSS által számolt mutató a harmadik centrális momentum és a szórás köbének hányadosa. Normál eloszlás esetén a ferdeségi mutató értéke nulla, pozitív értékek esetén jobbra ferdének, negatív értékek esetén balra ferdének nevezzük az eloszlást, az értékeknek nincs alsó és felső határa.

A **csúcsossági** *(kurtosis)* mutató jelzi, hogy az eloszlás a normálhoz viszonyítva csúcsosabb (jobban tömörül) vagy laposabb (kevésbé tömörül). Ezúttal is a nulla a normál eloszlással megegyező értéket, a pozitív értékek viszonylag csúcsos, míg a negatív értékek viszonylag lapos elosztást jeleznek.

A helyzet-, szóródási- és alakmutatók nemcsak önmagukban, hanem az egymáshoz való viszonyuk által is információt nyújtanak az eloszlásról. Normál eloszlás esetén az **esetek középső 68%-a egy szórásnyi távolságra**, 95%-uk két szórásnyi távolságra, 99.7%-a pedig három szórásnyi távolságra vannak az átlagtól[36]. Példánkban ha lefuttatjuk a FREQUENCY paranccsal a gyakorisági táblát (frequency tables) nemcsak a statisztikákat, akkor a középső 68% felső határát ott találjuk, ahol a kumulált gyakoriság 84%-a van. Ennek az értéke 52 125$, amiből ha kivonjuk a szórást (17 108$), akkor az eredmény 35 017$, ami jóval nagyobb, mint 34 443$-os értéke. Ez már egy jobbra ferde eloszlást jelez, akárcsak a ferdeségi mutató pozitív értéke.

A **ferdeségi mutató és a saját standard hibájának az arányát** *(Std. error)*(36. táblázat) normalitás gyorstesztnek is tekinthetjük: nem normál az eloszlás, ha az

[36] A statisztikában ezt a 68–95–99.7 szabálynak is nevezik.

arány abszolút értékben nagyobb, mint kettőt. Példánkban a 2.125/0.112 arány jóval nagyobb kettőnél.

A leíró statisztikák megismerése után elérkeztünk a normál eloszlás egzakt, statisztikai tesztjeinek az értelmezéséhez, ami másodlagossá teheti számunkra a grafikus módszerekkel való szemlélődést.

37. táblázat. A normál eloszlás tesztjei

Tests of Normality

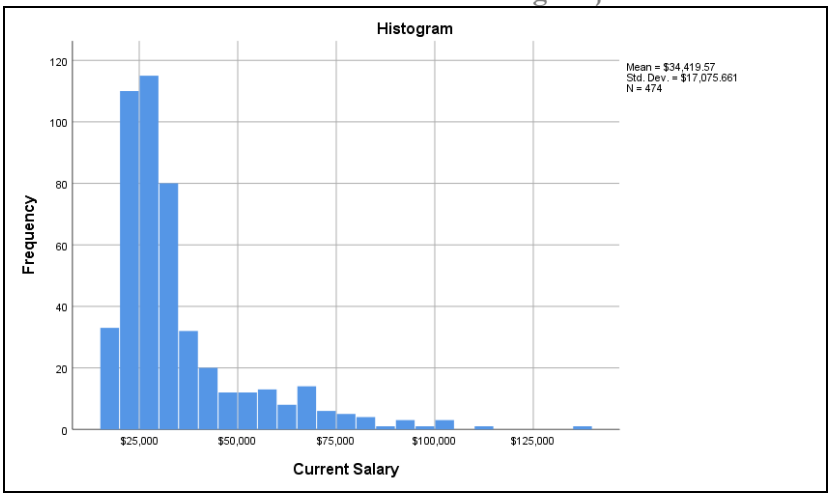
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
salary Current Salary	.208	474	.000	.771	474	.000

a. Lilliefors Significance Correction

Az SPSS a Kolmogor–Smirnov és a Shapiro–Wilk-teszttel vizsgálja a mintabeli változónk eloszlása és az elméleti normál eloszlás közötti hasonlóságot. **Mindkét teszt nullhipotézise az, hogy az eloszlás normális**, és a 37. táblázatban látható szignifikanciaszint szerint ennek nagyon kicsi a valószínűsége, vagyis a változónk nem normál eloszlású. Magas szignifikanciaszint ($p > .05$) esetén lenne a változó megközelítőleg normál eloszlású.

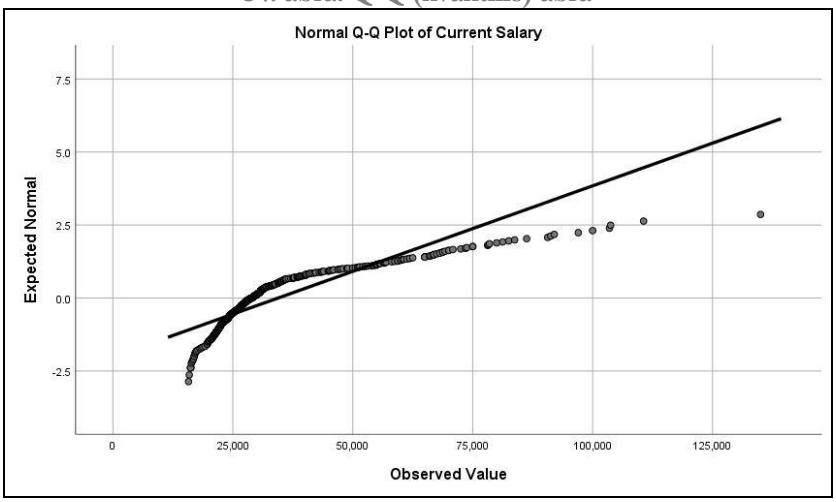
Ezt támasztja alá grafikus módon az eloszlás hisztogramja.

33. ábra. Az eloszlás hisztogramja



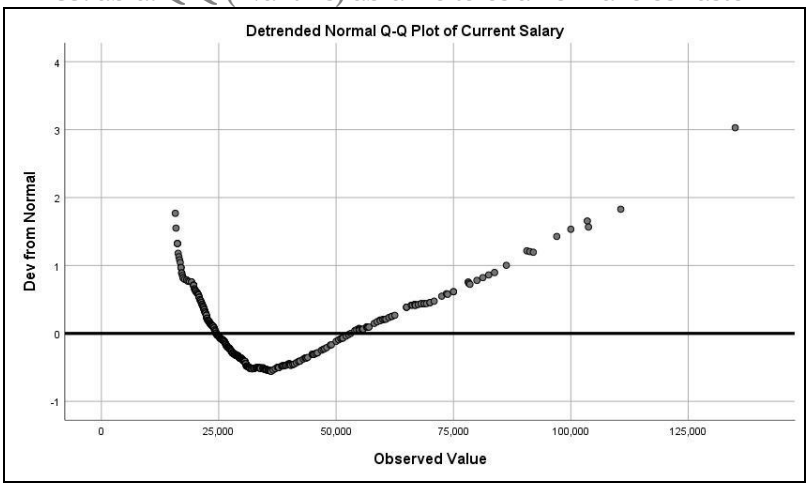
A hisztogram alapján láthatjuk, hogy a *jelenlegi fizetés* változónk eloszlása jobbra nagyon ferde, távol áll a normál eloszlás szimmetriájától.[37]

34. ábra. Q-Q (kvantilis) ábra



A Q-Q ábra akkor jelezné a megközelítőleg normál eloszlást, ha a pontokkal jelzett tényleges értékek rajta vagy közel lennének az elméleti normál eloszlást jelentő egyenesen. Az eltérés mértékét mutatja az SPSS által generált 35. ábra *(Detrended Normal Q-Q plot)*.

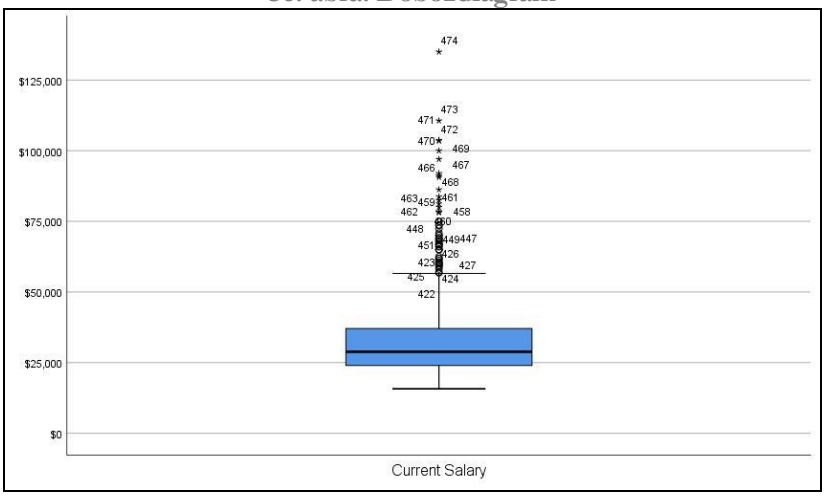
35. ábra. Q-Q (kvantilis) ábra – eltérés a normál eloszlástól



[37] Az SPSS EXPLORE utasítása érthetetlen módon nem jeleníti meg a hisztogramon az adott paraméterek szerinti normál eloszlás grafikonját. Ezt megtehetjük a FREQUENCY paranccsal, ha a Chart ablakban jelöljük a Histograms-ot és a Show normal curve on histogram opciót.

A Plots ablak Boxplots mezőjében a Factor levels together alapbeállítás meghagytuk, ezért az Output-ban megjelenik a következő dobozdiagram.

36. ábra. Dobozdiagram



A **dobozdiagram a kvartilisek alapján mutatja az eloszlást és a kiugró értékeket**. A besatírozott doboz az interkvartilis terjedelmet (IQR) mutatja, a sötét vonal a mediánt (50%). A kilógó alsó és felső „bajuszok" végei, korlátai általában a minimum és maximum értékeket jelenti, de az SPSS-ben – épp a kiugró értékek meghatározás miatt – mást jelent. Az SPSS Tukey-féle dobozdiagramja a doboz széleitől kivonva, illetve hozzáadva a doboz magasságának (interkvartilis terjedelem) másfélszeresét határozza meg az alsó és felső korlátot. Tehát:

- IQR = Q3 - Q1, ahol Q3 a harmadik, illetve Q1 az első kvartilis. Az interkvartilis terjedelem (IQR) értékét közli az SPSS (36. táblázat).
- a dobozdiagram alsó korlátja: F1 = Q1 - 1.5*IQR, vagy az eloszlás minimum értéke, amennyiben az nagyobb, mint az F1. Példánkban ez utóbbi érvényes (36. ábra).
- a dobozdiagram felső korlátja: F2 = Q3 + 1.5*IQR, vagy az eloszlás maximum értéke, amennyiben az kisebb, mint az F1.

A normalitásteszt eredményei között percentilisek, illetve a Tukey-módszerrel számolt kvartilisek *(Tukey`s Hinges)*[38] is fel vannak tüntetve (38. táblázat).

[38] A kerekítések miatt a dobozszélek (a Tukey-féle kvartilisek) értékei nagyon kis mértékben eltérhetnek az első és harmadik kvartilistől. További információért lásd https://www.ibm.com/support/pages/boxplots-hinges-and-quartiles (2022.08.).

38. táblázat. Percentilis értékek

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	salary Current Salary	$19,200.00	$21,000.00	$24,000.00	$28,875.00	$37,162.50	$59,700.00	$70,218.75
Tukey's Hinges	salary Current Salary			$24,000.00	$28,875.00	$37,050.00		

Most, hogy már értjük a Tukey-dobozdiagramot, használjuk az eredeti céljainkra; a normál eloszlás és a kiugró értékek vizsgálatára. **A dobozdiagramon** (36. ábra) **is a mediánt megjelenítő vonalhoz viszonyított szimmetriát keressük**, ezzel szemben aszimmetrikus diagramot és a kilógó értékek sorszámát találjuk. Az SPSS ponttal jelöli a kiugró és csillaggal az extrém kiugró értékeket.

4.5.3 A kiugró értékek meghatározása és kezelése

Szubjektív megítélés kérdése, hogy egy adott eloszlásban mi számít kiugró értéknek (*outlier*). A legáltalánosabban a kiugró érték olyan megfigyelés (adat), amely nem követi a többi megfigyelés mintázatát. Az SPSS által is használt definíció szerint, kiugró egy érték, ha **a dobozdiagram széleitől (harmadik illetve első kvartilis) számított távolsága legalább másfélszerese az interkvartilis terjedelemnek**, az extrém kiugró érték esetében pedig több mint háromszorosa.

A kiugró értékek megkeresésének egyik egyszerű módszere SPSS-ben, ha sorba rendezzük a változót (Sort cases), és az előbbiekben bemutatott Explore paranccsal megvizsgáljuk a dobozdiagramot. Ha sorba rendeztük a *salary* változót, akkor a legkisebb *outlier* sorszáma a legközelebb van a dobozhoz, így azonosítani tudjuk az adattáblában is. A másik, további felhasználást lehetővé tevő módszer, ha a definíciónak megfelelő algoritmussal képezünk egy új, outlier nevű változót, amelynek értéke 0 ha nem kiugró az adott érték, 1 ha igen, és 2 ha extrém kiugró. A Percentiles című táblázat (38. táblázat) *Tukey's Hinges* sorában megtaláljuk a Tukey-féle kvartiliseket, és kiszámíthatjuk ezekből az interkvartilis terjedelemet.

A következő algoritmust megfogalmazhatjuk menüből a COMPUTE paranccsal, vagy beírjuk a parancssorokat és futtatjuk a syntax ablakban:

```
if (salary-37050)/13050<1.5 outlier=0.
if ((salary-37050)/13050>=1.5 and (salary-37050)/13050<=3) outlier=1.
if (salary-37050)/13050>3 outlier=2.
exe[39].
```

[39] Ez az algoritmus a nagy kiugró értékek meghatározására alkalmas, a lefelé kiugró értékekhez az első kvartilisből kell kivonunk a *salary* változó értékeit.

Az eredmény egy új, *outlier* nevű változó, amit alaposan megvizsgálhatunk és eldönthetjük, hogy **mitévők legyünk a kiugró értékekkel**.

Könnyebb dolgunk van, ha a kiugró értékeket valamilyen **válaszadási, adatrögzítési hiba** okozta. Például az interjúalany egyéb jellemzői (iskolai végzettség, beosztás, kezdő fizetés) nem indokolják a kiugróan magas jövedelmét, vagy ha a jelenlegi jövedelem alacsonyabb, mint a kezdő fizetés. Az ilyen jellegű hibákat még az elemzés előtt, az adattisztítás szakaszában megpróbáljuk kiszűrni, de egy sokváltozós, nagy adattáblában gyakran marad olyan hiba, ami csak a mélyebb elemzésnél derül ki. Bizonyos esetekben egyértelmű a hibázás módja, és ennek megfelelően a javítás is, például egy nullával többet ütöttek az adatrögzítésnél. Azonban sokszor ez nem egyértelmű, ezért vagy helyettesítjük egy jellemzőnek vélt értékkel, leggyakrabban az átlaggal, vagy töröljük az adatcellát. Az átlaggal való helyettesítéskor érdemes az adott esetre (interjúalanyra) minél inkább hasonlító esetek (pl. nem, életkor, isk. végzettség szerinti) átlagát alkalmazni.[40]

Más a helyzet, ha a **kiugró értékek az eloszlás természetes részei**, és a jövedelem egy tipikusan ilyen változó. Elemezhetjük-e egy vállalat bérpolitikáját úgy, hogy ha kizárjuk a magas jövedelműeket? Lehet, hogy a fizetés változónk normális eloszlású lesz, de ez alapjaiban érintené az eredeti kutatási célunkat elérését. Akkor fogadható el a kiugró értékek kizárása, amikor nem valamiféle szisztematikusság határozza meg ezeket, hanem a véletlenszerűség vagy a kivételes egyediség. Ez a dilemma napjaink egyik fontos ismeretelméleti (*episztemológiai*) kérdésére is rávilágít, amit az eddig tanultak alapján teljességgel beláthatunk; a modern tudomány a többség, a fősodor jellemzői alapján fogalmaz meg törvényszerűségeket, ami nem magyarázza az eltérő jellemzőkkel rendelkező *outlier*-ek viselkedését.[41]

Elfáradva a sok mérlegeléstől, jobb híján huszárosan levágjuk a kiugró értékeket, remélve, hogy a *jövedelem* változónk elfogadható mértékben normál eloszlású lesz. A hisztogramon (37. ábra) értelmezve ez azt jelenti, hogy levágjuk az eloszlás jobb oldali, hosszú, lapos részét.

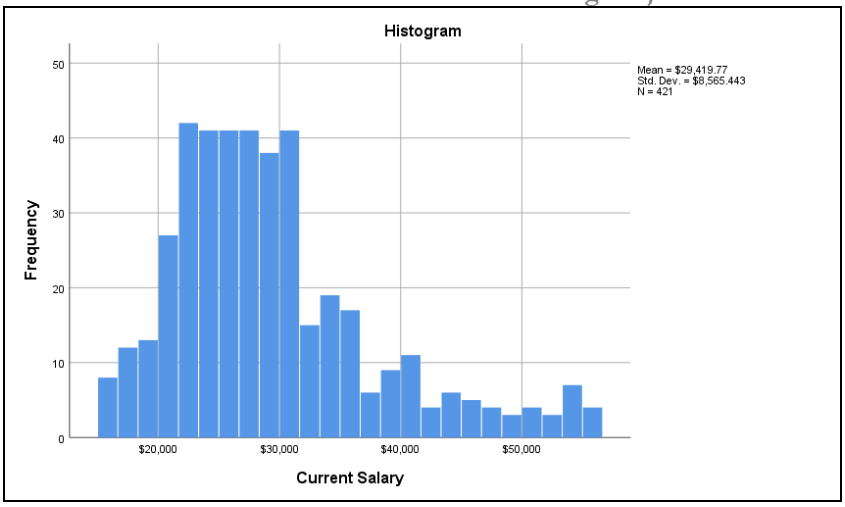
[40] Komplexebb statisztikai próbáknál ezt a lehetőséget a módszerbe beépítve, opcióként ajánlja az SPSS.

[41] Az internetre épülő új gazdaság sikerének egyik sarokköve, hogy a cégek képesek a fősodor mellett az egyedit is, az eloszlás hosszú végeit *(long tail)* figyelembe venni. Napjaink legértékesebb cégei az egyedi szegmentációt kínálják, nem csak néhány nagy szegmensét.

Példa. Próbáljuk ki tehát, hogy a kiugró értékek kizárása után normál eloszlású lesz-e a salary változónk. A dobozdiagramon látjuk, hogy a legkisebb kiugró érték sorszáma 422, az adattáblában kikeressük az ehhez tartozó fizetésértéket, ami 56,750$. A SELECT CASES paranccsal az ennél kisebb fizetésekre szűkítjük a mintát.

A részletes eredmények bemutatása nélkül közöljük, hogy a Shapiro–Wilk-teszt alapján a szűkített változó sem normális eloszlású. Ez jól látszik a következő, jobbra ferde hisztogramon is.

37. ábra. A szűkített változó hisztogramja



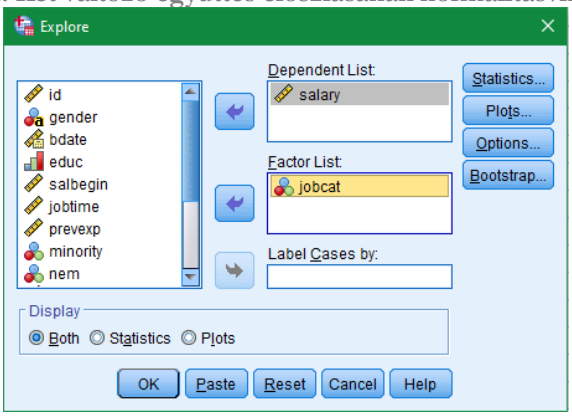
A legnagyobb értékek kizárása után is az új hisztogram is távol áll egy szimmetrikus haranggörbétől. Mélyrehatóbb változtatásra van szükségünk, és a motivációnk fenntartása érdekében közöljük, hogy a megoldást a **kétlépcsős transzformáció** fogja meghozni.

Előbb azonban nézzük meg a normalitásvizsgálatot olyan esetben, amikor két változó együttes eloszlását kell vizsgálni (pl. az ANOVA esetében).

4.5.4 Normalitásvizsgálat több változó esetén

Könyvünkben az ANOVA alkalmazásának feltételeinél fogalmaztuk meg, hogy a függő változó eloszlása a független, nominális változó kategóriái mentén is normális eloszlású kell legyen. Az együttes eloszlás vizsgálatához ugyanazt az EXPLORE-parancsot használjuk, mint az előbbiekben, de a Factor List mezőbe bevisszük nominális változót – az *Egyutas varianciaanalízis* alfejezetben leírtaknak megfelelően – a beosztás (*jobcat*) változót.

38. ábra. Két változó együttes eloszlásának normalitásvizsgálata



A többi beállítás ugyanaz, mint az előző, csak a függő változó normalitását vizsgáló esetben (lásd a 32. ábrát).

Eredmények értelmezése

39. táblázat. Két változó együttes eloszlásának normalitásvizsgálata – eredmények

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	jobcat beosztas	Statistic	df	Sig.	Statistic	df	Sig.
salary Current Salary	1 fizikai	.276	27	.000	.818	27	.000
	2 szellemi	.107	363	.000	.882	363	.000
	3 manager	.109	84	.016	.929	84	.000

a. Lilliefors Significance Correction

Az eredmények közül csak a tesztekét mutatjuk be (39. táblázat), ami alapján elmondhatjuk, hogy a *salary* változó eloszlása, a *jobcat* nominális változó kategóriáin belül sem normál eloszlású. A hisztogramok ferde eloszlása és a Q-Q plot ábrák grafikusan is megerősítik ezt az eredményt.

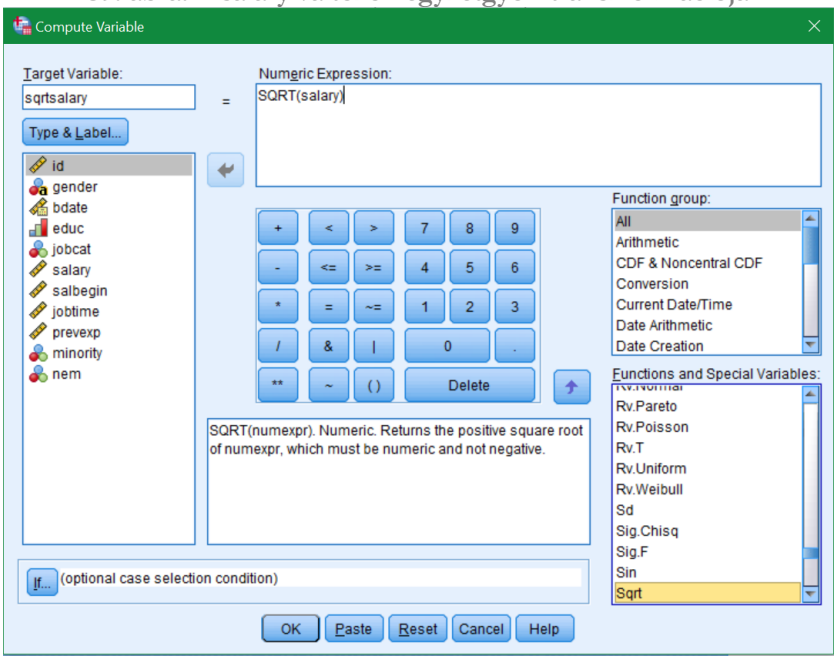
4.5.5 Transzformációk

Egy változó transzformálása alatt azt értjük, hogy **minden egyes adaton ugyanazt a matematikai műveletet hajtjuk végre**.
Egyszerű példa a hőmérsékleti adatok Celsius-fokról Fahrenheitre való alakítása: az eredeti adatokat megszorozzuk egy konstanssal (1,8-cal), és ehhez hozzáadunk egy másik számot (32-t). Azonban ez egy lineáris transzformáció, ami nem változtatja meg az eloszlás alakját, nem változtatja normálissá azt.

A szakirodalomban a transzformációra leggyakrabban használt függvények a logaritmus, négyzetgyök, az 1/x. Fontos szempont, hogy a transzformáció őrizze meg az értékek eredeti **sorrendiségét**, vagyis ha a>b, akkor f(a)>f(b).

Példa. A már megismert COMPUTE-paranccsal alkalmazzuk a négyzetgyök (sqrt) függvényt[42] a *salary* változónkra.

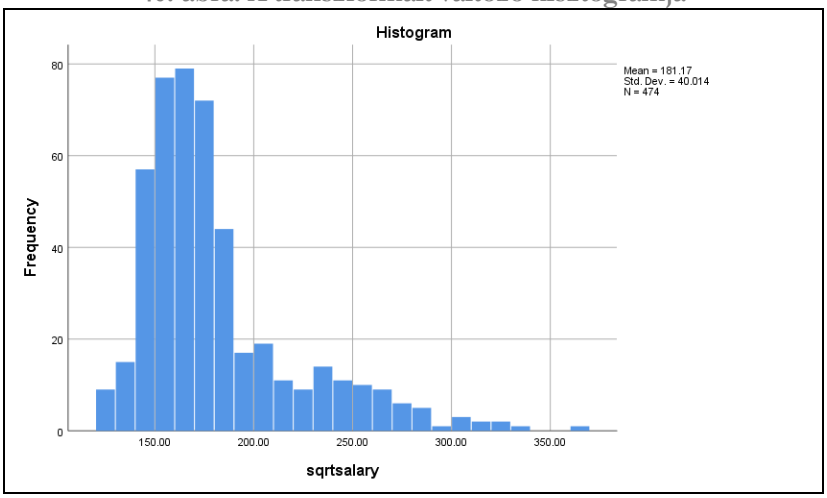
39. ábra: A salary változó négyzetgyök-transzformációja



A új, *sqrtsalary* nevű változónk normalitásvizsgálatát az eddig tanultaknak megfelelően elvégezve azt találjuk, hogy a transzformált változónk sem normál eloszlású, és ezt megerősíti a hisztogramja is.

[42] Angolul a négyzetgyök *square root*

40. ábra: A transzformált változó hisztogramja



Hasonlóan eredménytelen a tízes alapú logaritmikus (lg10) és a reciprok (1/x) függvény alkalmazása, de a következőkben bemutatott komplexebb, célirányosabb transzformáció eredményre vezet.

Kétlépcsős transzformáció

Az előző megoldás-kísérleteknél sokkal megbízhatóbb a G. Templeton (2011) által javasolt kétlépcsős eljárás.

1. lépés. Képezünk egy új, **egyenletes eloszlású változót**, ami az eredeti változó értékeinek **százalékos rangsorát** mutatja. A következő egyszerű képlettel számítunk:

$$Százalékos\ rangsor = 1 - rangsor(xi)/n$$

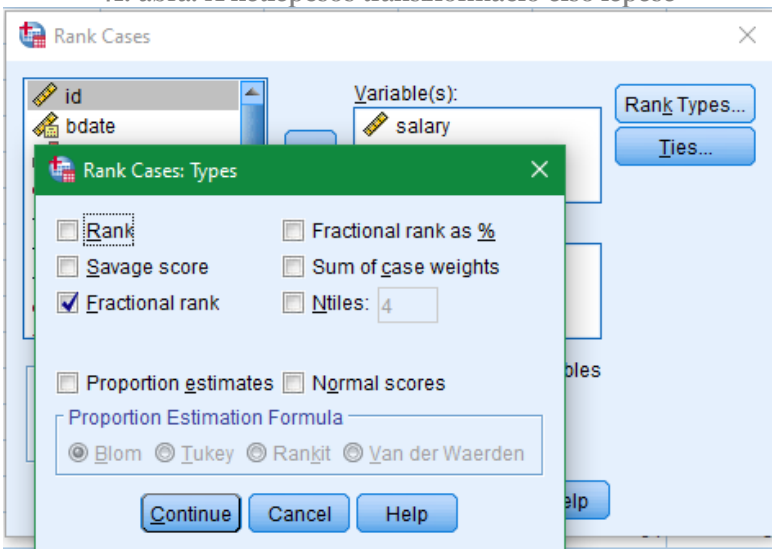
ahol a rangsor(xi) az xi érték rangsora, az n a mintaelemszám. Az eredmény egy 0 és 1 közötti értékeket felvevő változó, amely eléri az 1-et, de a 0-t nem.

Példánkban a legkisebb fizetés százalékos rangsora 1-1/474 = 0.0021, a legnagyobbé 1-474/474=1.

Transform→Rank cases

A Variables(s) mezőbe bevisszük a salary változót, majd a Rank Types-ra kattintva a Rank Cases: Types ablakban beállítjuk a Fractional rank opciót.

41. ábra: A kétlépcsős transzformáció első lépése



Az eredmény egy új változó az adattáblában, amit automatikusan *Rsalary* –nek nevezett a program.

2. lépés. Az egyenletes eloszlású változót **átalakítjuk normál eloszlásúvá, az inverz sűrűségfüggvény** alkalmazásával**.**

Transform→Compute variable

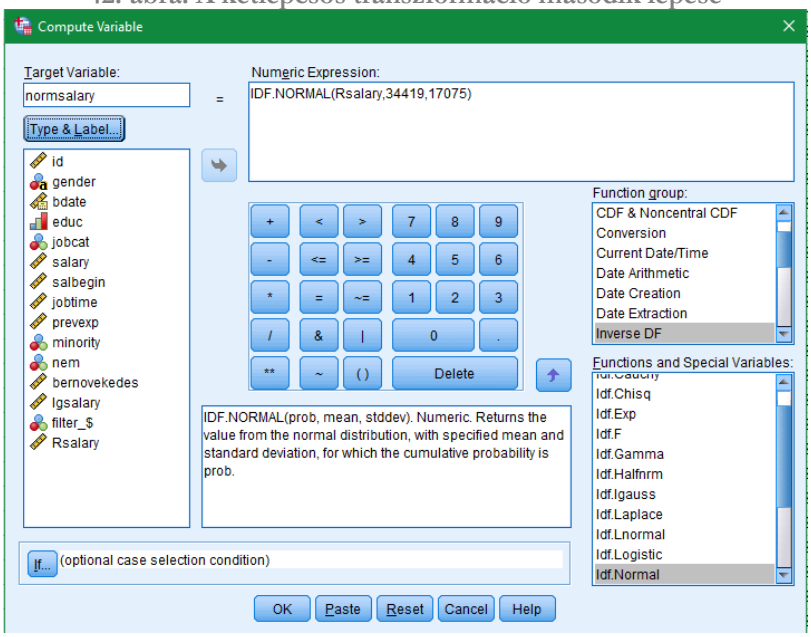
A COMPUTE-paranccsal létrehozzuk az új változót. A Target Variable mezőbe beírjuk az új változó nevét, legyen *normsalary,* és a következő lépésben a Numeric Expression mezőben beírjuk az új változót létrehozó algoritmust. Ez a normál eloszlás inverz sűrűségfüggvénye lesz, amihez a Function group mezőben kiválasztjuk az Inverse DF-t, majd a Function and Special Variables mezőben az Idf.Normal-t és a nyílra kattintva bekerül a Numeric Expression mezőbe.
Az inverz sűrűségfüggvényt a következő három paraméterrel határozzuk meg:

IDF.NORMAL (uniformizált változó, az eredeti változó átlaga és szórása)

Példánkban a sűrűségfüggvény utáni zárójelbe beírjuk az 1. lépésben létrehozott egyenletes eloszlású változót *(Rsalary)*, az eredeti változó (*salary*) átlagát, és vesszővel elválasztva a szórását.[43]

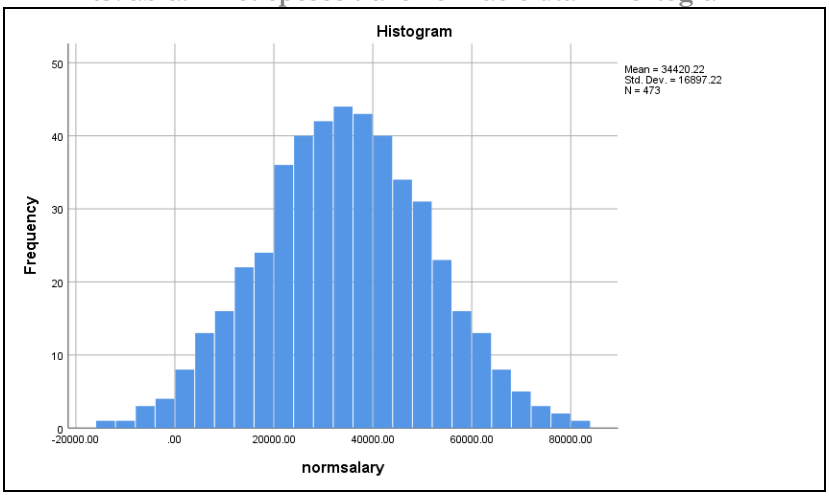
[43] Az átlaghoz és szóráshoz FREQUENCY-t futtatni már nem okoz nekünk gondot.

42. ábra: A kétlépcsős transzformáció második lépése



A puding próbája az evés, az eloszlásé a normalitásvizsgálat. Futtassuk a már jól ismert EXPLORE-parancsot a *normsalary* változóra.

43. ábra: A kétlépcsős transzformáció utáni hisztogram



A hisztogram már annyira normál eloszlást mutat, hogy szinte szép, akárcsak a normalitásvizsgálat két tesztje.

40. táblázat: A kétlépcsős transzformáció utáni normalitásteszt

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
normsalary	.014	473	.200*	.999	473	1.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Mind a Kolmogorov–Smirnov, mind a Shapiro–Wilk-teszt tökéletesen normál eloszlást mutat, ami nem meglepő, mivel a kétlépcsős transzformáció algoritmusa pontosan ezt célozza. Már csak az a nagy kérdés, hogy a transzformált *normsalary* változónk további elemzése mennyire felel meg az eredeti kutatási célkitűzésünknek, a fizetések *(salary)* elemzésének. .

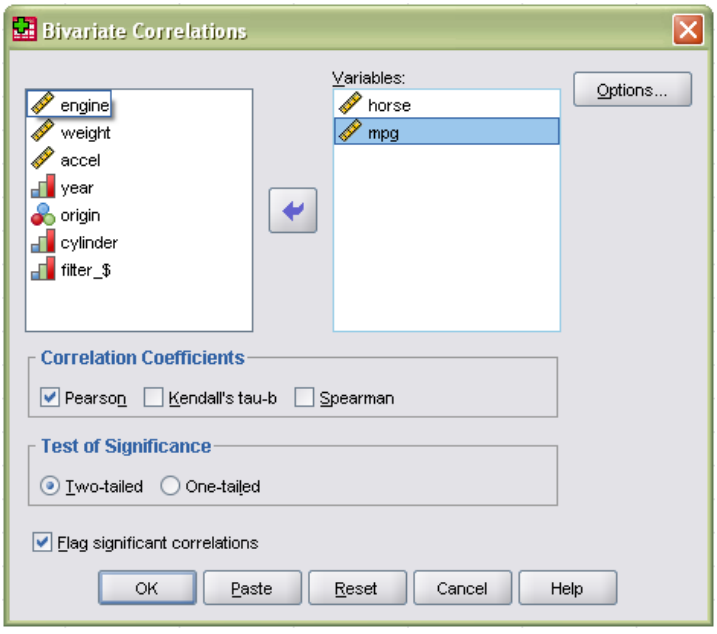
4.6 Korrelációanalízis

Két metrikus változó közötti kapcsolatot és annak szorosságát határozza meg a korrelációanalízis. A kapcsolat szorosságát kifejező Pearson-féle korrelációs együttható abszolút értéke 0 és 1 közötti értéket vehet fel, ahol a 0 érték a kapcsolat teljes hiányát jelzi, az 1-es pedig a két változó függvényszerű kapcsolatát jelenti. A kapcsolat szorossága mellett információt kapunk a kapcsolat irányáról is. Ha a korrelációs együttható előjele pozitív, akkor a két változó közötti kapcsolat egyenes arányú, negatív előjel esetén pedig fordított arányról beszélünk.

Példa: nyissuk meg újra az SPSS bemutató adattáblái közül a cars.sav adatfájlt. Teszteljük azt a nagyon valószerű hipotézist, hogy az autók teljesítménye és fogyasztása között egyenes arányú kapcsolat van. Az adattáblában a horse (lóerő) változó tartalmazza az autók teljesítményére vonatkozó információt, az mpg pedig a fogyasztását, konkrétan azt mutatja, hogy hány mérföldet lehet megtenni egy gallon (kb. 4 liter) benzinnel. (Az amerikai autók amerikai mértékegység szerint fogyasztanak, tehát a nagyobb érték kisebb fogyasztást jelent.)

Analyze→Correlate→Bivariate

44. ábra. Korrelációanalízis: változók bevitele



A Bivariate Correlations ablakban válasszuk ki a változólistából a horse és az mpg változókat, majd OK.

Eredmények értelmezése

41. táblázat. Korrelációanalízis: korrelációs együttható

Correlations

		HORSE Horsepower	MPG Miles per Gallon
HORSE Horsepower	Pearson Correlation	1.000	-.771**
	Sig. (2-tailed)	.	.000
	N	400	392
MPG Miles per Gallon	Pearson Correlation	-.771**	1.000
	Sig. (2-tailed)	.000	.
	N	392	398

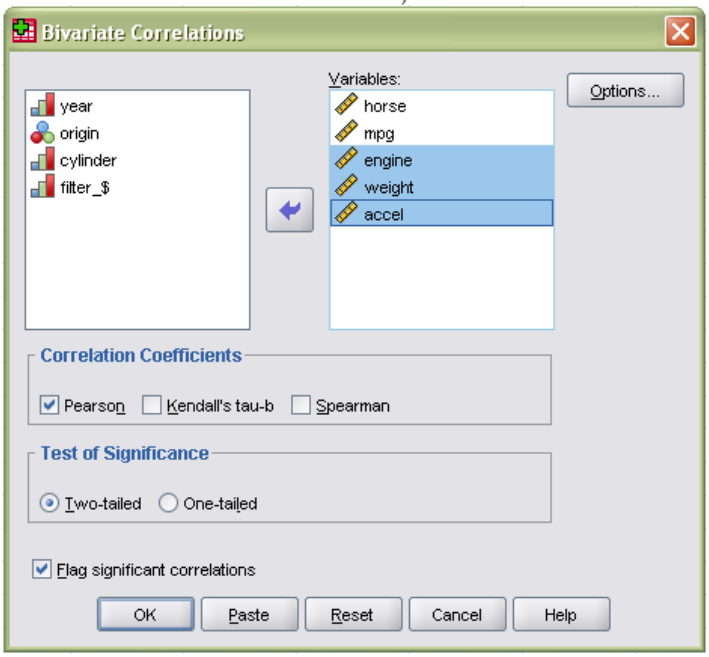
**. Correlation is significant at the 0.01 level (2-tailed).

Az eredmény egy szimmetrikus mátrix, amelynek főátlójában egy változó önmagával vett korrelációja látható, ezért az együttható értéke 1. A mátrix cellái három értéket tartalmaznak: az első a korrelációs együttható (Pearson Correlation), a második a szignifikanciaszint (Sig.(2-tailed)), a harmadik a korrelációanalízisbe bevont esetek száma (N). A szignifikanciaszint .05 alatti értéke alapján szignifikáns a kapcsolat a két változó között.
A korrelációs együttható negatív előjele azt jelzi, hogy ha nő az autók teljesítménye (lóerő), akkor csökken az egy gallon benzinnel megtehető mérföldek száma (magyarán nő a fogyasztás), a 0.771-es érték pedig szoros kapcsolatot mutat.

A korrelációanalízisbe kettőnél több numerikus változó is bevonható, de természetesen a változók közötti kapcsolatok páronként lesznek vizsgálva. Nézzük meg, hogy az adattáblában található, autókra vonatkozó többi numerikus változó hogyan viszonyul egymáshoz.
Vonjuk be tehát az elemzésbe az előző két változó mellé az ENGINE (motorűrtartalom köbhüvelykben), WEIGHT (súly fontban), ACCEL (gyorsulás, hány másodperc alatt éri el a 60 mérföld per órát) és a YEAR (gyártási év) változókat. Adjuk ki a következő parancsot:

Analyze→Correlate→Bivariate

45. ábra. Korrelációanalízis: újabb változók bevitele



Eredmények értelmezése

42. táblázat. Korrelációanalízis: korrelációs együtthatók

Correlations

		Miles per Gallon	Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)	Time to Accelerate from 0 to 60 mph (sec)
Miles per Gallon	Pearson Correlation	1	-.789**	-.771**	-.807**	.434**
	Sig. (2-tailed)	.	.000	.000	.000	.000
	N	398	398	392	398	398
Engine Displacement (cu. inches)	Pearson Correlation	-.789**	1	.897**	.933**	-.545**
	Sig. (2-tailed)	.000	.	.000	.000	.000
	N	398	406	400	406	406
Horsepower	Pearson Correlation	-.771**	.897**	1	.859**	-.701**
	Sig. (2-tailed)	.000	.000	.	.000	.000
	N	392	400	400	400	400
Vehicle Weight (lbs.)	Pearson Correlation	-.807**	.933**	.859**	1	-.415**
	Sig. (2-tailed)	.000	.000	.000	.	.000
	N	398	406	400	406	406
Time to Accelerate from 0 to 60 mph (sec)	Pearson Correlation	.434**	-.545**	-.701**	-.415**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.
	N	398	406	400	406	406

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

Az eredménytábla a bőség zavarával lep meg, valamennyi változó páronként szignifikáns korrelációs kapcsolatban van a másikkal. Legnagyobb korrelációs együtthatót (.933) az autó súlya és a motor hengerűrtartalma között találjuk, és erős, negatív korrelációs kapcsolat (-.807) van az autó súlya és az egy gallon benzinnel megtehető mérföldek száma (fogyasztás) között.

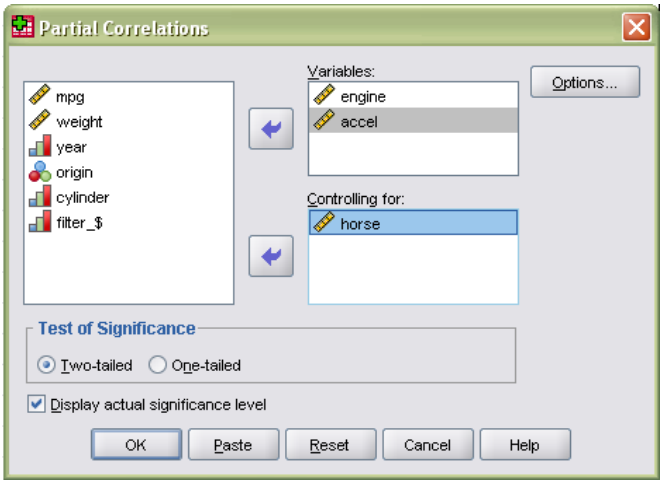
4.7 Parciális korrelációanalízis

Alaposan végiggondolva valamennyi eredményét az előző korrelációs táblázatnak, elbizonytalaníthat néhány dolog. Például a motor hengerűrtartalma (ENGINE) és a gyorsulás (ACCEL) változók közötti korrelációs együttható (-.545) azt mutatja, hogy nagyobb motortérfogat kevesebb idő alatti, azaz jobb gyorsulást jelent. Különösebb műszaki képzettség nélkül is sejtjük, hogy ez azért van, mert a nagyobb motor nagyobb teljesítményt, több lóerőt jelent. Ezt az összefüggést meg is találjuk a gyorsulás (ACCEL) és a lóerő (HORSE) ismérvek közötti korrelációs kapcsolatban (-.701). Kérdés, hogy az előbbi összefüggés csak a gyorsulás és a lóerő közötti összefüggésnek köszönhető-e, vagy azonos teljesítmény (lóerő) mellett is a nagyobb motortérfogat jobb gyorsulást eredményez? Úgy kellene vizsgálnunk ezt az összefüggést, hogy közben kontroll alatt tartjuk a teljesítmény indirekt hatását.

A parciális korrelációanalízis úgy vizsgálja két numerikus változó közötti összefüggést, hogy kontrollálja, kiküszöböli más változók hatását.

Analyze→Correlate→Partial

46. ábra. Parciális korrelációanalízis: változók bevitele



A Variables mezőbe bevisszük azt a két változót, amelyek között a korrelációs kapcsolatot kívánjuk vizsgálni, a Controlling for mezőbe pedig azt a változó(ka)t, amelynek a hatását kontroll alatt akarjuk tartani.

Eredmények értelmezése. A parciális korrelációs analízis outputja csak annyiban módosul a kétváltozós korrelációs mátrixhoz képest, hogy fel van tüntetve a kontrollváltozó (horse) és a cellákban a harmadik érték nem az elemszám (N), hanem a szabadságfok (df).

43. táblázat. Parciális korrelációanalízis: parciális korrelációs együttható

Correlations

Control Variables			Engine Displacement (cu. inches)	Time to Accelerate from 0 to 60 mph (sec)
Horsepower	Engine Displacement (cu. inches)	Correlation	1.000	.269
		Significance (2-tailed)	.	.000
		df	0	397
	Time to Accelerate from 0 to 60 mph (sec)	Correlation	.269	1.000
		Significance (2-tailed)	.000	.
		df	397	0

A teljesítmény indirekt hatását kiküszöbölve is, a motor hengerűrtartalma és a gyorsulás között szignifikáns, de pozitív korrelációs együtthatót (0.269) találunk. Ebből azt a némileg meglepő következtetést vonhatjuk le, hogy – azonos lóerejű teljesítmény mellett – a nagyobb motortérfogat hosszabb gyorsulási időt, azaz rosszabb gyorsulást jelent! Kontroll alatt tartva a teljesítmény hatását, a kétváltozós korrelációanalízis negatív korrelációs együtthatójával szemben (-0.545), most teljesen eltérő eredményt (0.269) kaptunk.

A parciális korrelációanalízissel az indirekt hatásoktól megtisztítva kapjuk meg a tényleges összefüggést két változó között. Egymással összefüggő változócsoportok elemzésekor az „egyszerű” Pearson-féle korrelációanalízis alkalmazása tapasztalatom szerint gyakran vezet téves eredményekhez, eltúlzott vagy alulbecsült összefüggésekhez. A többváltozós, egymással összefüggő és egymást meghatározó viszonyok elemzésére a **regresszióanalízis** az igazán alkalmas módszer.

5. VÁLTOZÓK KÖZÖTTI TÖBBDIMENZIÓS KAPCSOLATOK VIZSGÁLATA

5.1 Kétváltozós regresszióanalízis

A lineáris regresszió-analízis **úgy vizsgálja metrikus változók közötti kapcsolatot, hogy a különböző indirekt hatásokat kontroll alatt tartja**, de a korrelációval ellentétben a kapcsolat nem szimmetrikus, hanem az **egy vagy több független változó hatását vizsgáljuk egy függő változóra**[44].
A lineáris regresszió-analízis elsőfokú polinomiális függvénnyel fejezi ki a függő és a független változó közötti kapcsolatot:

$$y = b_0 + b_1 \cdot x \qquad (14)$$

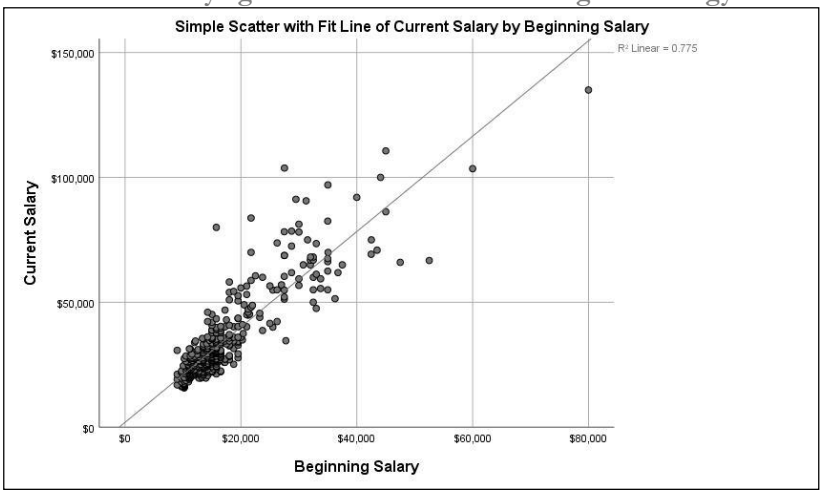
ahol y – függő változó, x – független változó, b_1– a független változó együtthatója, b_0 – konstans. Ez az általános formájú polinomiális függvényt akkor van meghatározva, akkor tudjuk felrajzolni a grafikus képét, ha ismerjük az együtthatók értékeit. Célunk tehát a b_0 konstans és a b_1 együttható (paraméter) meghatározása, erre a lineáris regresszió-analízis a **legkisebb négyzetek módszerét** alkalmazza. Ez a módszer úgy keresi a legjobban illeszkedő lineáris függvényt, hogy a tényleges megfigyeléseket jelentő pontok és az egyenes közötti távolságok négyzetösszege minimális legyen, magyarán mindegyik értékhez a lehető legközelebb legyen.

Példa: Folytassuk a korábbi példánkat, amelyben egy cég alkalmazottainak fizetését vizsgáljuk! Nyissuk meg Employee data.sav adatfájlt, és vizsgáljuk a legfontosabb kérdést, mitől függ a fizetés nagysága. Az adattáblában rendelkezésünkre álló lehetséges magyarázó változók között tallózva értelmesnek tűnik a kezdő fizetés *(salbegin)* változó bevonása, feltételezhetjük, hogy a jelenlegi fizetés mértéke függ az alkalmazáskori fizetés nagyságától.

A 47. ábrán az Employee data.sav adatfájlból a kezdő és a jelenlegi fizetés közötti kapcsolat grafikus képét egy pontdiagrammal ábrázoljuk[45].

[44] Szinonimákként használjuk a magyarázó illetve a magyarázott változó kifejezéseket is.
[45] Az ábra létrehozásának módját a későbbiekben tárgyaljuk.

47. ábra. A tényleges értékek és az illesztett regressziós egyenes

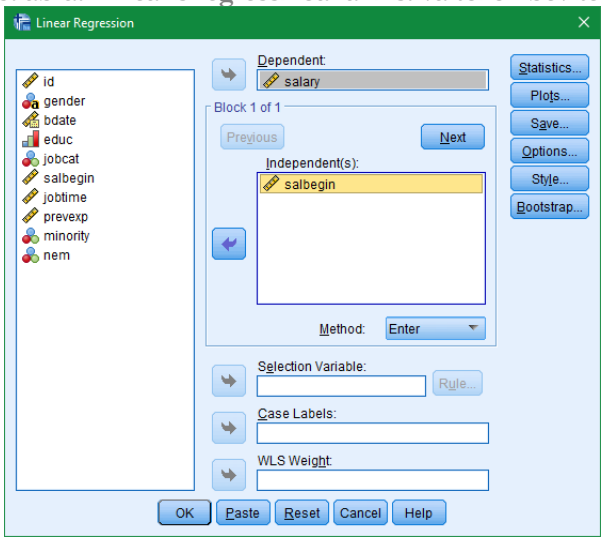


Az ábra pontjai a minta 474 esetét jelenítik meg, a hozzájuk tartozó fizetés és kezdő fizetés koordinátarendszerében. Feltüntettük az ábrán a legkisebb négyzetek módszerével illesztett lineáris regressziós egyenest is.

Analyze→Regression→Linear

Regressziós modellünk függő változóját, a jelenlegi fizetést *(salary)* vigyük be a Dependent mezőbe, magyarázó változónak (Independent(s)) pedig a kezdő fizetést (*salbegin*).

48. ábra. Lineáris regresszióanalízis: változók bevitele



A továbbiakban több lényeges beállítással is megismerkedünk, de most a minimális alapbeállításokkal kattintsunk az OK gombra.

Eredmények értelmezése. Az output első táblázata (44.) a modellbe bevont változókat sorolja.

44. táblázat. Lineáris regresszióanalízis: a modell változói

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	salbegin Beginning Salary[b]	.	Enter

a. Dependent Variable: salary Current Salary

b. All requested variables entered.

A következő táblázat fontos mutatója az R^2 *(R Square)* determinációs együttható.

45. táblázat. Lineáris regresszióanalízis: az R^2 determinációs együttható

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.880[a]	.775	.774	$8,125.224

a. Predictors: (Constant), salbegin Beginning Salary

Az **R négyzet determinációs együttható azt mutatja, hogy a modellbe bevont független változók milyen arányban magyarázzák a függő változó varianciáját**. Ez a 0 és 1 közötti értékeket felvevő mutató a modell jóságát, magyarázó erejét jelzi. Példánkban a .595-ös érték szerint elég nagy mértékben sikerült megmagyaráznunk a függő változót, egy kis egyszerűsítéssel úgy is fogalmazhatunk, hogy a kezdő fizetés majd 60%-ban megmagyarázza a jelenlegit.

46. táblázat. Lineáris regresszióanalízis: szignifikanciaszint

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.068E+11	1	1.068E+11	1618.073	.000[b]
	Residual	3.103E+10	470	66019261.63		
	Total	1.379E+11	471			

a. Dependent Variable: salary Current Salary

b. Predictors: (Constant), salbegin Beginning Salary

Az ANOVA feliratú táblázat (46.) a regressziós modell egészének szignifikanciáját mutatja, a modell valamennyi paraméterének **együttes szignfikanciáját** teszteli[46]. A regressziós eredmények értelmezését ajánlott ezzel kezdeni, mivel ha a modell nem szignifikáns, akkor felesleges a további vizsgálódás. Szignifikáns modell esetén azonban vigyázó szemeinket vessük a független változó együtthatójának (paraméterének) statisztikáit tartalmazó Coefficients nevű táblára (47. táblázat).

47. táblázat. Lineáris regresszióanalízis: a független változók paraméterei

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1938.907	890.412		2.178	.030
	salbegin Beginning Salary	1.910	.047	.880	40.225	.000

a. Dependent Variable: salary Current Salary

Az utolsó oszlopban (Sig.) a lineáris regressziós egyenlet konstans tagjának és a független változó (*salbegin*) becsült paraméterének a szignifikanciaszintjeit találjuk, amelyekre ugyanaz a már jól ismert küszöbérték ($p<.05$) érvényes.
A táblázat első oszlopában pedig a konstans és a független változó paraméterének értékeit találjuk (Unstandardized Coefficients, B). Ezek alapján **felírhatjuk a regressziós egyenletet**:

$$Y = 1938{,}9 + 1.91\ X \qquad (15)$$

ahol Y a jelenlegi és X a kezdő fizetés. Az X együtthatóját (b_1) úgy értelmezzük, hogyha egy egységgel nő az X, akkor 1.91-gyel az Y. Akinek modellünk szerint egy dollárral nagyobb volt a kezdő fizetése, annak 1.91-gyel nagyobb a jelenlegi. Úgy is fogalmazhatunk, hogy a vizsgált cégnél az alkalmazottak átlagosan körülbelül kétszer annyit keresnek, mint az alkalmazásukkor.
Könnyen érthető, egyszerű összefüggés vizsgálatát választottuk bevezetésképp a regressziós modellezésbe: a kezdő fizetéssel magyarázni a jelenlegi fizetést. Nyilvánvaló számunkra, hogy több más tényező (változó) is befolyásolja a fizetés nagyságát, ezt az eddigi elemzéseink során is láttuk (pl. az ANOVA alfejezetben). A következőkben szintet lépünk, a rendelkezésünkre álló további változók bevonásával a lehető legjobb magyarázó erejű modellt szeretnénk létrehozni.

[46] Ezt az F-próbát Wald-próbának is nevezik, mivel Wald Ábrahám (1902-1950) kolozsvári születésű matematikus dolgozta ki.

5.2 Többváltozós regresszióanalízis

A multidimenzionálisnak is hívott regresszióanalízis során **több magyarázó változót bevonunk az elemzésbe**, de a modell – a korrelációanalízistől eltérően – nem páronként vizsgálja a változók közötti kapcsolatokat, hanem **egyidejűleg valamennyi magyarázó változó a függő változóra gyakorolt hatását elemzi úgy, hogy kontroll alatt tartja a magyarázó változók közötti hatásokat**. Nagy jelentősége van a kvantitatív kutatások során a multidimenzionális elemzéseknek, mivel könnyen félrevezető eredményekhez jutunk, ha nem vesszük figyelembe a különböző indirekt hatásokat[47].

Példa: megértve a többváltozós elemzés felsőbbrendűségét a változók páronkénti összefüggés-vizsgálataival szemben, alkossunk végre teljes képet példabeli cégünk jövedelempolitikájáról. Az Employee data.sav adatfájl valamennyi metrikus változóját bevonjuk a modellbe, nemcsak azért, mert ez technikailag lehetséges, hanem mert logikus, hogy az iskolai végzettség *(educ)*, kezdő fizetés *(salbegin)*, régiség *(jobtime)*, előző régiség *(prevexp)* hatással lehet a fizetés nagyságára.

5.2.1 A többváltozós regressziós modell beállításai

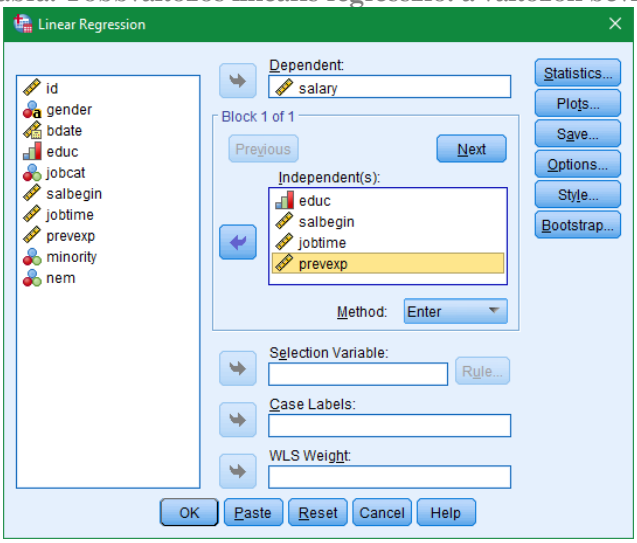
A többváltozós regresszióanalízis beállításait ugyanott eszközölhetjük, mint a kétváltozós esetben, csak most több változót vonunk be az Independent(s) mezőbe[48].

[47] A kétváltozós elemzések (pl korrelációanalízis) korlátaira vonatkozó vicces példák gyűjteményét találjuk a https://www.tylervigen.com/spurious-correlations oldalon.

[48] A modellezés előtt ajánlott a FREQUENCY paranccsal megismerni a független változók értékeinek eloszlását és főbb statisztikáit.

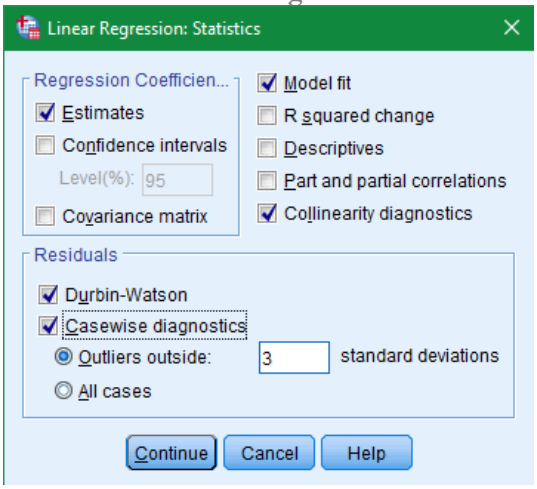
Analyze→Regression→Linear

49. ábra. Többváltozós lineáris regresszió: a változók bevitele



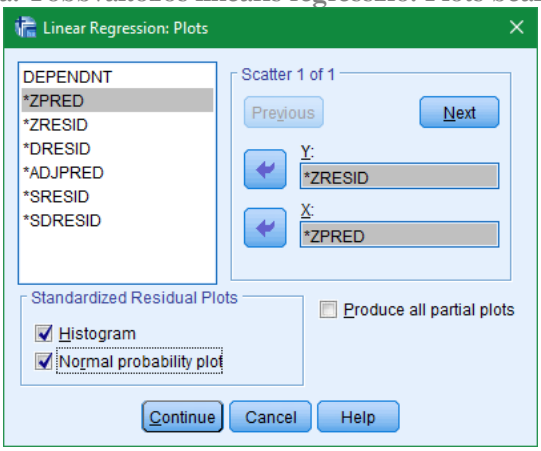
A Statistics ablakban az Estimates és a Model fit alapbeállítások mellé jelöljük még a Collinearity diagnostics, a Durbin-Watson és a Casewise diagnostics opciókat. Ez utóbbiakat a későbbiekben, a regressziós modell feltételeinek ellenőrzésénél részletezzük.

50. ábra. Többváltozós lineáris regresszió: Statistics beállítások



Szükségünk lesz továbbá a Plots ablakban egy fontos ábrára ezért az 51. ábrának megfelelően járjunk el.

51. ábra. Többváltozós lineáris regresszió: Plots beállítások



5.2.2 A többváltozós regressziós modell eredményeinek értelmezése

Az első táblázatot most nem tüntettük fel, a második táblázat már lényeges információt közöl.

48. táblázat. Többváltozós regresszió: az R^2 és a Durbin-Wastson teszt értékei

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.900[a]	.810	.809	$7,480.526	1.922

a. Predictors: (Constant), prevexp Previous Experience (months), jobtime Months since Hire, salbegin Beginning Salary, educ Educational Level (years)

b. Dependent Variable: salary Current Salary

A determinációs együttható (R Square) .810 értéke szerint nagymértékben, 81%-ban sikerült megmagyaráznunk, hogy milyen tényezők és milyen mértékben határozzák meg a fizetés mértékét. A közgazdasági kutatásokban meglehetősen ritka az ilyen nagy magyarázó erejű modell, ami valószínűleg kezdő fizetés változójának *(salbegin)* köszönhető ebben a modellben. Más kérdés, hogy a kutatás célja szempontjából nem sokat ér az az állítás, hogy „annak nagy jelenleg a fizetése, akinek az alkalmazásakor is nagy volt." Ezzel csak azt szeretnénk jelezni, hogy az R^2 értéke nem az egyedüli fontos kritérium a modell jóságának, érvényességének megítélése során.

A Durbin-Wattson teszt eredményét a későbbiekben, a regressziós modell feltételeinek ellenőrzésénél tárgyaljuk.

Ilyen magyarázó erő mellett természetes, hogy a modellünk szignifikáns, hiszen az ANOVA táblázat – amelyet ezúttal nem tüntettünk fel – F próbájának nullhipotézise, hogy a determinációs együttható, vagyis a modell magyarázó ereje nulla.

A regresszióanalízis legtöbb információját a következő, a regressziós egyenlet becsült paramétereire vonatkozó 49. táblázat tartalmazza.

49. táblázat. Többváltozós regresszió: a független változók paraméterei

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-16120.580	3275.580		-4.921	.000		
	educ Educational Level (years)	669.972	166.011	.113	4.036	.000	.516	1.937
	salbegin Beginning Salary	1.769	.059	.815	30.049	.000	.552	1.813
	jobtime Months since Hire	161.055	34.435	.095	4.677	.000	.993	1.007
	prevexp Previous Experience (months)	-17.249	3.542	-.105	-4.870	.000	.865	1.156

a. Dependent Variable: salary Current Salary

A táblázat vizsgálatát érdemes az együtthatók t-próbáinak szignifikanciaszintjével kezdeni, mivel csak a modell szignifkáns független változóival kell a továbbiakban foglalkoznunk, **csak ezek kerülnek be a regressziós egyenletbe**. Megállapíthatjuk, hogy a modellünkben valamennyi független változó szignifikáns hatással van a függő változóra.

A táblázat első oszlopa (Unstandardized Coefficients B) tartalmazza a **lineáris regressziós egyenletet meghatározó paramétereket**. Ezek közül csak a szignifikáns független változók együtthatóit kell vizsgálnunk, adott esetben valamennyit. Az iskolai végzettség *(educ)* független változó paraméterét (669.97) úgy értelmezzük, hogy modellünk alapján plusz egy év iskolai végzettség 670$ éves fizetéstöbbletet jelent, a többi változó azonos szintje mellett.
A kezdő fizetés *(salbegin)* együtthatója 1.77 (kéttizedesre kerekítve) kisebb, mint az egydimenziós modell 1.91-es együtthatója (47. táblázat). Ugyanis a többváltozós modell további változóinak (*educ, jobtime, prevexp*) hatásai közvetlenül jelennek meg a saját együtthatóikban, és nem indirekt hatásként a *salbegin* együtthatójában.
A *jobtime* együtthatója alapján megállapíthatjuk, hogy az adott cég jutalmazza a régiséget, átlagosan 161$-ral nő a jövedelem minden hónap után. Meglepő és tanulságos az előző régiség (*prevexp*) negatív együtthatója (-17.25$); vagyis a

kevesebb előző régiséggel rendelkezők – azonos iskolai végzettség, kezdő fizetés és régiség mellett - nagyobb fizetést kapnak. Megfogalmazhatjuk, hogy a vizsgált cég a kisebb régiséggel rendelkezőknek nagyobb fizetés ad, az azonos egyéb jellemzőkkel rendelkező jelöltek közül? Igen, amennyiben sikerül kontroll alatt tartanunk a beosztás (*jobcat*) hatását, amit nem tudtunk a modellbe bevonni, mivel nominális változó. Megtörténhet, hogy a magas fizetésű menedzsereket kis régiséggel, friss diplomával alkalmazzák, a fizikai és irodai beosztottakat pedig nagyobb tapasztalattal. Modellünk tökéletesítésének folyamatát, a **modellspecifikációt** ezzel folytatjuk, a nominális változók modellbe való bevonásával, de előbb ismerkedjünk meg a Coefficients tábla további adataival.

Az indirekt hatások kontroll alatt tartása mellett a többváltozós regressziós modell további előnyös tulajdonsága, hogy **fontossági sorba rendezhetjük a független változókat a függő változóra gyakorolt hatás alapján**. A becsült paraméterek (B) alapján tehetjük ezt meg, kérdés, hogyan hasonlítjuk össze az iskolai végzettség éveinek hatását a kezdő fizetés dollárjával? Az eddig megismert B standardizálatlan együttható nem alkalmas erre az összehasonlításra, mivel ez a független változók eredeti mértékegységét tartalmazza. Szükségünk van egy közös mértékegységre, pontosabban egy mértékegységtől független együtthatóra.
A Coefficients táblázat Beta oszlopában (Standardized Coefficients) található **standardizált béták** ilyen mértékegységtől független együtthatók, a szignifikáns **magyarázó változók hatásának az összehasonlítására alkalmasak**. Ezek alapján megállapíthatjuk, hogy a jelenlegi fizetésre legnagyobb hatást a kezdő fizetés (.815), az iskolai végzettség (.113), az előző régiség (-.105)[49] és a munkahelyi régiség (.095) gyakorolja.

A Coefficients Std. Error oszlop az együtthatók standard hibáit mutatja, a konfidencia intevallum kiszámításához használhatjuk[50], amennyiben alapsokasági becslére vagy előrejelzésre akarjuk használni modellünket.

Az utolsó két oszlop (49. táblázat) multikollinearitásra vonatkozó mutatóit (Tollerance és VIF) a későbbiekben, a regressziós modell feltételeinek ellenőrzésénél tárgyaljuk.

[49] A rangsor megállapításánál a beta előjeltől független, abszolút értékét vesszük figyelembe.
[50] Az együtthatók konfidencia intervallumainak kiszámítását kérhetjük a Statistics ablakban a Confidence intervals opcióval.

5.3 Nominális változók beépítése a modellbe

Bebizonyosodott az előzőekben, amit a regressziós modell ismerete nélkül is sejthettünk, hogy egy adott cég jövedelempolitikájának az elemzéséből kihagyhatatlan a beosztás, főképp a vezetői beosztás figyelembe vétele.

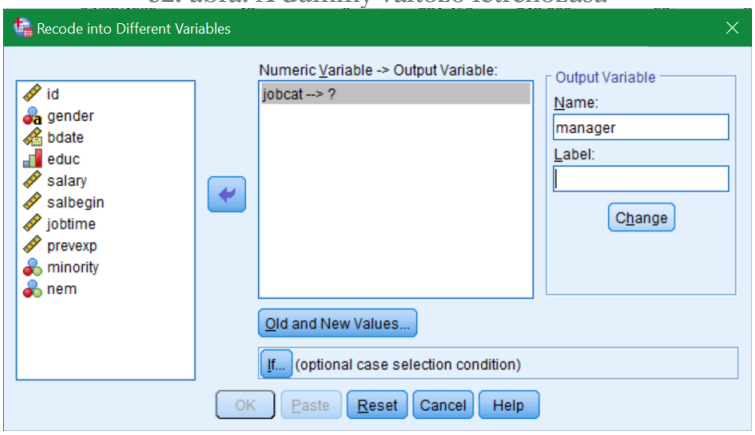
A lineáris regressziós modellbe lehetséges a kétértékű, nominális változók, az úgynevezett dummy változók bevonása. A dummy változó értékei tipikusan 1 – ha jellemző az adott ismérv, 0 – ha nem, tehát a regressziós egyenletben a dummy változó együtthatója $X_i=1$ esetén épp a vizsgált jellemző hatását mutatja, $X_i=0$ pedig nulla a hatás.

Az Employee data.sav két olyan nominális változót is tartalmaz, amelynek két értéke van, és kutatási szempontból indokolt a modellbe való bevonásuk. A *nem* változó bevonása azt a korábban, az ANOVA bemutatásánál vizsgált kutatási kérdés megválaszolását ígéri, hogy van-e nemi diszkrimináció az adott cégnél a fizetések terén? Hasonló kutatási kérdést fogalmazhatunk meg a *minority* változó kapcsán.

A beosztás változónk *(jobcat)* azonban három értékű: 1-fizikai, 2-szellemi, 3-menedzser. **Szaknyelven *dummy*-zásnak nevezzük azt a transzformációt, amivel a többértékű nominális változó kategóriájából külön dummy (0 és 1 értékű) változókat hozunk létre.**
A dummy változó értéke 1 - ha érvényes az adott jellemző és 0 – ha nem. Amennyiben a többértékű nominális változó valamennyi kategóriájából új változót szeretnénk képezni, akkor eggyel kevesebb dummy változót hozunk létre, mivel a meglévő dummy változók 0 értékei jelentik a „kihagyott" kategóriát. Tehát ha *k* számú értéke van a nominális változónknak, akkor *k-1* dummy változót képezhetünk, így elkerüljük az úgynevezett *dummy csapdát*.

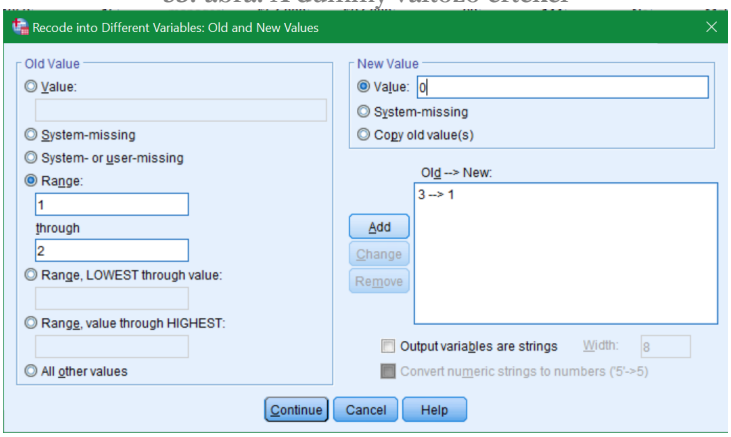
Mivel a beosztás változó három kategóriája közül a fizikai és a szellemi/irodai alkalmazottaknak a fizetése nem különbözik szignifikáns mértékben, ezért a beosztás változót úgy alakítjuk át, hogy egy kategória legyen a menedzser és egy másik a beosztottak (fizikai+szellemi). Ehhez elegendő lesz egy kétértékű (dummy) változót létrehoznunk a RECODE INTO DIFFERENT VARIABLES paranccsal.

52. ábra. A dummy változó létrehozása



A RECODE parancs alkalmazását láttuk a Változók újrakódolása című alfejezetben. A dummy változónk nevét *(manager)* beírjuk az Output Variable mezőbe, majd Change.
Az új változó értékeit az Old and New Values ablakban állítjuk be.

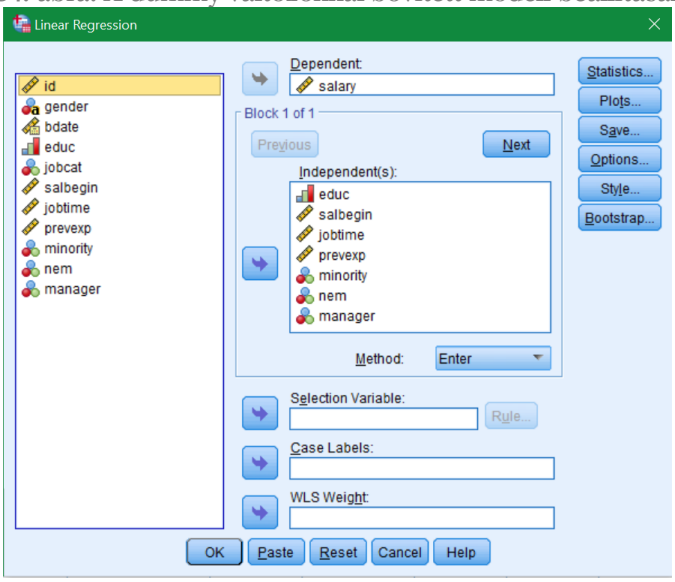
53. ábra. A dummy változó értékei



A *jobcat* változó 1 és 2 értékeiből lesz a *manager* 0 értéke, a 3-asból pedig az 1.

A létrejött *manager* változó mellett a modellbe bevonhatjuk a *nem* és a *minority* dummy változókat. Később bekapcsolódó olvasóinknak jelezzük, hogy korábban a *nem* változót az eredeti *gender*, szöveg változóból képeztük a Transform menü Automatic recode paranccsával.
Újrafuttatjuk regressziós modellünket már hét független változóval, és ugyanazokkal a beállításokkal, amit korábban alkalmaztunk (50. és 51. ábra).

54. ábra. A dummy változókkal bővített modell beállításai



Eredmények értelmezése

A modell összefoglaló táblázatából látjuk, hogy a determinációs .838-ra nőtt, ami egy nagyon jó magyarázó erejű modellt jelent. A korábbi (48. táblázat) .810-es értékhez viszonyított növekedés pedig a dummy változók bevonását igazolja.

50. táblázat. A dummy változókkal bővített modell összefoglalója

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.915[a]	.838	.835	$6,942.551	1.899

a. Predictors: (Constant), manager, jobtime Months since Hire, prevexp Previous Experience (months), minority Minority Classification, nem nem, educ Educational Level (years), salbegin Beginning Salary

b. Dependent Variable: salary Current Salary

A dummy változókkal bővített modell becsült paraméterei és a hozzájuk tartozó statisztikákat az 51. táblázatban találjuk.

51. táblázat. A dummy változókkal bővített modell paraméterei

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-9679.275	3179.332		-3.044	.002		
	educ Educational Level (years)	351.407	158.895	.059	2.212	.027	.486	2.060
	salbegin Beginning Salary	1.322	.074	.609	17.754	.000	.297	3.368
	jobtime Months since Hire	150.212	32.071	.088	4.684	.000	.986	1.014
	prevexp Previous Experience (months)	-16.336	3.396	-.100	-4.811	.000	.811	1.233
	minority Minority Classification	-855.755	802.182	-.021	-1.067	.287	.924	1.083
	nem nem	2867.188	753.036	.084	3.808	.000	.726	1.377
	manager	11327.362	1394.495	.254	8.123	.000	.359	2.786

a. Dependent Variable: salary Current Salary

Az előző modellben (49. táblázat) is tesztelt metrikus, független változók ezúttal is szignifikánsak, az együtthatók előjelei nem változtak, értékeik pedig kismértékben csökkent. A béták is értelemszerűen csökkentek, mivel az újonnan bevont változók közvetlenül jelenítik meg azok információtartalmát.

A dummy változók együtthatóinak értelmezéséhez ismernünk kell az értékeiket, a *nem* változó esetében 0 – nő, 1 – férfi. Modellünkben a *nem* együtthatóját úgy értelmezzük, hogyha a változó értéke 1, vagyis férfi, akkor a függő változó, vagyis a fizetés 2867 dollárral nő, a nem változó 0 értékéhez, vagyis a nők fizetéséhez képest. Mindez **a többi magyarázó változó azonos szintje mellett**!

A *manager* változó együtthatója 11327 dollár, tehát a – modell szerint – a menedzserek esetében ennyivel nő a fizetés a beosztottakhoz képest.

A modell további eredményeit, output-ját a következő résznél tárgyaljuk.

5.4 A lineáris regresszióanalízis alkalmazásának feltételei

Idáig viszonylag könnyen elsajátíthattuk ennek a nagyon hatásos adatelemzési módszernek az alkalmazását. Azonban a lineáris regressziós modellnek van néhány statisztikai feltétele, amelyek ha nem teljesülnek torzított becsléshez, következésképp téves kutatási eredményekhez jutunk.

5.4.1 Multikollinearitás

A multikollinearitás a többváltozós regressziómodellben **a független változók közötti erős korrelációs kapcsolatot** jelenti. Hangsúlyoztuk, hogy a regresszióanalízis fontos tulajdonsága, hogy úgy vizsgálja a függő és a független változók közötti kapcsolatot, hogy kiszűri, kontroll alatt tartja a többi, modellbe bevont független változó hatását. Ez azonban nem sikerülhet teljes mértékben, ha a független változók között determinisztikus (függvényszerű), vagy nagyon erős sztochasztikus kapcsolat van. Függvényszerű kapcsolat esetén tökéletes multikollinearitásról beszélünk, és ilyen esetben az SPSS automatikusan kizár egy vagy több változót a független változók közül, és hibajelzéssel figyelmeztet. Erős sztochasztikus kapcsolat esetén a kutatóra hárul a probléma súlyosságának és kezelésének a mérlegelése. Egy többé-kevésbé elfogadott hüvelykujjszabály szerint **részleges multikollinearitásról** beszélhetünk, ha bármely két független változó közötti **korrelációs együttható meghaladja a 0,7-es értéket.**
Az ilyen részleges multikollinearitás figyelmen kívül hagyása az alábbi következményekkel járhat (Ramanathan, 2003):

- csökkenti a független változók becsült paramétereinek a t-értékét, így az együtthatók kevésbé szignifikánsan lesznek, vagy a ténylegesen szignifikáns független változót nem annak tünteti fel,
- de nem ront a modell előrejelző képességén, esetenként még javíthatja is.

A multikollinearitás teszteléséhez tehát korrelációanalízissel vizsgáljuk a független változók páronkénti korrelációit, ezt nemcsak a már tanult módon, hanem a regresszióanalízis Statistics ablakában a Descriptive opció választásával is megtehetjük.

Két másik mutató is rendelkezésünkre áll a multikollinearitás tesztelésére (pontosabban egy és annak a reciproka). A Statistics ablakban beállítottuk a

Collinearity diagnostics opciót (50. ábra), ezért az eredmények között a Coefficients tábla a Collinearity Statistics (Tolerance és VIF) oszlopokkal bővült.

Egy adott független változó **VIF statisztikája** azt mutatja, hogy a változó becsült paraméterének varianciája hányszorosa annak, ami a multikollinearitás teljes hiányának esetén lenne. Értékelésekor azt mondhatjuk, hogy ha a VIF mutató 1 és 2 között van, akkor gyenge, ha 2 és 5 között van, akkor erős, zavaró, ha pedig 5 felett van, akkor nagyon erős, káros a multikollinearitás (Kovács, 2014). A **tolerancia mutató** a VIF reciproka, 0 és 1 között értékekkel, ahol a kisebb érték jelzi a nagyobb multikollinearitást.

Ha megállapítottuk a nagyon erős multikollinearitást, akkor a következő megoldási lehetőségeink vannak:

- **Kihagyjuk** a modellből az erős korrelációs kapcsolatban levő változókat. Ha csak két változó között van szoros korrelációs kapcsolat, akkor természetesen elég az egyik változót kihagynunk. Ez az egyszerű, de drasztikus megoldás akkor elfogadható, ha a kihagyott független változó nincs nagy hatással a függő változóra, vagyis a modell magyarázó ereje csak kis mértékben csökken.
- Ha több fontos független változó között van szoros korrelációs kapcsolat, akkor gyakran alkalmazott megoldás az ún. **főkomponens-analízis**. Ezzel a módszerrel a független változókból kevesebb számú, egymással nem korreláló új változót képezünk, majd ezeket építjük be a regressziós modellbe független változóként. E módszer ismertetése nem szerepel könyvünkben (ajánlott irodalom Székelyi-Barna 2002, Sajtos-Mitev 2007).

Esetünkben a *salbegin* változó 3.36-os VIF mutatója erős multikollinearitást jelez, ha megvizsgáljuk a Correlations táblát (újrafuttatva a modellt a Descriptive opcióval), akkor azt látjuk, hogy az *educ* változóval .633 és a *manager* változóval .782 a korrelációs együttható. Az utóbbi korreláció nehezen értelmezhető, mivel a *manager* egy dummy változó, de tudjuk, hogy jelentős jövedelemkülönbség van a menedzserek és a beosztottak között.
A szakirodalom alapján nem egyértelmű, hogy az ilyen mértékű multikollinearitás mekkora problémát jelent, a legjobb, ha mi járjuk körül azt, és kipróbáljuk a modellünket a *salbegin* változó kizárásával. Úgyis kicsit önmagát magyarázó tautológiának tűnt az alkalmazása a cég fizetési politikájának a modellezésénél, mivel a fizetési politika a kezdő fizetés megállapításánál is nyilvánvalóan érvényesül.

Kihagyjuk tehát a *salbegin* változót és újrafuttatjuk a modellt minden egyéb beállítást megtartva. Az eredmények közül csak a Coefficients táblát mutatjuk. A modell magyarázó ereje (a fel nem tüntetett Model summary táblázatban) a determinációs együttható alapján .728, ami meglepően jó.

52. táblázat. A multikollinearitást okozó változó nélküli modell paraméterei

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-5335.108	4103.405		-1.300	.194		
	educ Educational Level (years)	1266.338	194.570	.214	6.508	.000	.543	1.843
	jobtime Months since Hire	107.182	41.396	.063	2.589	.010	.992	1.009
	prevexp Previous Experience (months)	-2.962	4.286	-.018	-.691	.490	.853	1.173
	minority Minority Classification	-1917.011	1035.529	-.046	-1.851	.065	.929	1.077
	nem nem	6394.373	940.261	.186	6.801	.000	.780	1.281
	manager	27063.963	1393.582	.606	19.420	.000	.602	1.660

a. Dependent Variable: salary Current Salary

Jelentős változások történtek a modellünkben, a legjelentősebb, hogy a kezdő fizetés *(salbegin)* kihagyása után az előző munkahelyeken eltöltött régiség *(prevexp)* már nem szignifikáns, nem befolyásolja a jelenlegi fizetést. Legnagyobb magyarázó erejű változó a *manager* lett (.606-os bétával), ezt követi az *iskolai végzettség* és a *nem*.

5.4.2 A reziduumok normál eloszlása

A reziduumok a regressziós modell becsült értékei és a megfigyelt értékek közötti eltérés. Könnyen értelmezhetjük ezt a definíciót, ha a regressziós modell első ábrájára (47. ábra) tekintünk: a megfigyelt értékeket jelentő pontok és a becsült regressziós egyenes közötti távolságot jelenti.

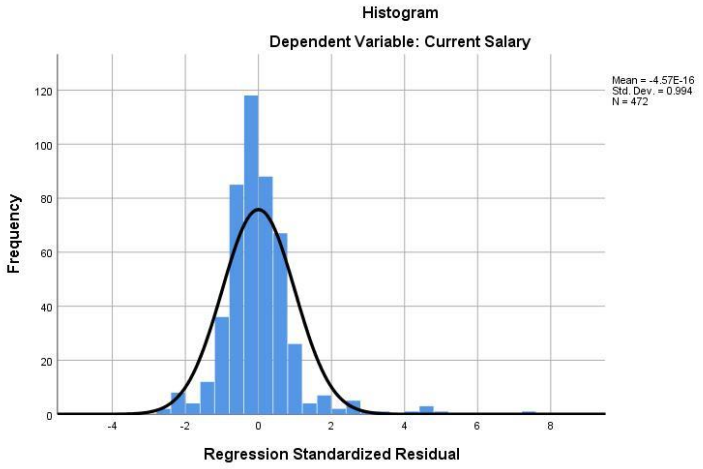
$$e_i = y_i - \hat{y}_i \,, \qquad (16)$$

ahol az y_i a megfigyelt és y_i a becsült érték. Az eltérések, a hibatagok nagyságát eredeti mértékegységük szerint nem tudjuk mérlegelni, ezért szükség van a **standardizálásukra**. Ez a reziduumok standard hibával való elosztását jelenti, így a standardizált reziduumok átlaga 0, a szórásuk egységnyi lesz. A **studentizált reziduum** esetében a reziduumokat egy korrigált standard hibával osztjuk.

A reziduumok normál eloszlására vonatkozó feltétel nem teljesülése a regressziós modell általánosíthatóságát, a statisztikai becslések pontosságát veszélyezteti. Két lehetőségünk is van reziduumok normál eloszlásának tesztelésére:

1. A **standardizált reziduum hisztogramjának és P-P ábrájának grafikus vizsgálata**, amelyeket a Plots ablakban már beállítottunk a regressziós modell futtatása előtt (51. ábra). A Plots ablak Standardezid Residual Plots mezőjében bejelöltük a Histogram és Normal probability plot opciókat, ezek eredménye a következő (55. és 56.) ábrák.

55. ábra. A standardizált reziduumok hisztogramja



Ránézésre megnyugtató képet mutat a hisztogram, amire meglehetősen jól illeszkedik **a normál eloszlás folytonos görbéje**. Megjegyezzük, hogy a hisztogram normalitásának grafikus vizsgálata félrevezető lehet, mivel az alakja nagymértékben függ az oszlopok, a program által automatikusan választott szélességétől.

56. ábra. A standardizált reziduumok P-P ábrája

A normál P-P diagram akkor jelzi az eloszlás megközelítőleg normális jellegét, ha a megfigyelt értékek pontjai közel vannak az egyeneshez. Ennél egzaktabb eredményt sajnos nem tudunk a grafikus módszerekkel megfogalmazni, ezért az alaposabb vizsgálathoz ajánljuk a Normalitásvizsgálat alfejezetben leírt teszteket.

2. Másik opció a **studentizált reziduumok normál eloszlásának** a vizsgálata. Ezt megtehetjük grafikus módszerekkel (hisztogram, Q-Q ábra) vagy a **normalitásteszteket** (Kolmogorov-Smirnov, Shapiro-Wilk[51]). Mindkettő részletes leírását a Normalitásvizsgálat alfejezetben találjuk, ezért itt már nem tárgyaljuk. Fontos megjegyeznünk, hogy a vonatkozó szakirodalom szinte kivétel nélkül „megközelítőleg" normál eloszlásról ír, hangsúlyozva, hogy a **legkisebb négyzetek módszere robusztus becslés a normalitás feltételének sérülésével szemben**.
Amennyiben jelentősen sérül ez a feltétel, a következőket tehetjük:

- transzformáljuk a függő változót (lásd a 4.5.5 alfejezetet), ami azonban megnehezíti a modell eredményeinek értelmezését;
- normál eloszlást nem feltételező regressziós modellt alkalmazunk, pl. a logisztikus regressziót;

[51] Érdekes, hogy maga az SPSS Tutorial nem ajánlja ezeket a túl szigorúnak tartott normalitásteszteket. *„However, we don't generally recommend these tests."* (https://www.spss-tutorials.com/spss-multiple-linear-regression-example/#multiple-regression-assumptions. 2022.08.)

- alkalmazzuk a lineáris regressziós modellt, és a következményeket megpróbáljuk figyelembe venni.

5.4.3 Független megfigyelések

A lineáris regressziómodell következő feltétele, hogy **a megfigyeléseink, az adattáblában szereplő eseteink függetlenek egymástól.** Ennek teljesülését már a kutatás tervezésénél tudjuk biztosítani. Keresztmetszeti, nem ismétlődő kutatásoknál, amennyiben az interjúalanyok függetlenek egymástól (pl. nem egy családból választottunk több interjúalanyt), nagy valószínűséggel nem lesz probléma a feltétel teljesülése, összetett szekunder vagy idősoros adatoknál már valószerű.

A feltételt úgy is megfogalmazhatjuk, hogy az adatok nem **autokorrelálnak,** azaz a különböző megfigyelésekhez tartozó **reziduumok között nincs függvényszerű kapcsolat**. Elsőrendű autókorreláció fennállása esetén egy reziduum függvényszerű kapcsolatban van a megelőző időponthoz tartozó reziduummal. Egzakt próbával állapíthatjuk meg az autókorreláció létét vagy hiányát: a Durbin-Watson próbával. A Durbin-Watson tesztstatisztika 0 és 4 között vehet fel értékeket, **ha 1.8 és 2.2 közötti intervallumba esik, akkor az autokorreláció nem jelent problémát** a modellünk számára. Amennyiben a DW mutató kívül esik az 1.8 – 2.2 tartományon vagyis az adataink autokorrelálnak, nem alkalmazhatjuk a lineáris regressziót, más módszerre van szükségünk, amit jelenleg nem részletezünk[52].

A regressziós modellünk beállításakor kértük a Durbin-Watson tesztstatisztikát a Statistics ablakban (50. ábra), az értékét a Model Summary táblázatban (50. tábla) láthatjuk: 1.899. Kijelenthetjük, hogy a modell teljesíti az adatok autokorrelálatlanságára vonatkozó feltételt.

Fontos megjegyeznünk, hogy a Durbin-Watson próba **függ az adatok sorba rendezésétől!** Példánkban ha az adattáblát sorba rendezzük a *salary* változó szerint a D-W mutató csak 1.321. Ezt a problémát úgy kerülhetjük el, ha az adattáblát egy véletlenszerű sorszám *(id)* szerint rendezzük.

[52]A Durbin–Watson-próba alkalmazási korlátainak, és további részletek megismerése céljából ajánljuk a Ramanathan (2003) könyvet.

5.4.4 A kiugró értékek, befolyásos esetek vizsgálata

Egy változó **kiugró értékei** *(outlier)* jelentősen befolyásolhatják az eloszlást, ezért ez a feltétel szorosan összefügg az előzővel. A Normalitásvizsgálat alfejezetben megfogalmazott definíciónk, miszerint a kiugró értékek eltérnek a többi adat „mintázatától" most még értelmezhetőbbé válik, ugyanis a regressziós egyenestől számított eltérés, a **reziduum** pontosan kifejezi azt.

A többváltozós regressziós modell beállítása során a Statistics ablakban jelöltük a Casewise diagnostics opciót (50. ábra). Itt a kiugró értékek úgy vannak definiálva, mint az az eset, **amelynek standardizált reziduuma meghaladja a szórás háromszorosát**. Mint láttuk, a standardizált reziduum átlaga 0, a szórása 1, tehát a következő táblázatban (53. táblázat) azon esetek sorszámai vannak feltüntetve, amelyek standardizált reziduuma nagyobb, mint három.

53. táblázat. A kiugró értékű esetek sorszámai és reziduumai

Casewise Diagnostics[a]

Case Number	Std. Residual	salary Current Salary	Predicted Value	Residual
18	5.911	$103,750	$62,793.45	$40,956.547
32	3.554	$110,625	$85,999.03	$24,625.971
103	3.479	$97,000	$72,892.57	$24,107.432
106	3.596	$91,250	$66,332.72	$24,917.285
205	-3.448	$66,750	$90,643.99	-$23,893.986
218	6.653	$80,000	$33,901.97	$46,098.025
274	4.377	$83,750	$53,423.19	$30,326.814
446	3.137	$100,000	$78,260.95	$21,739.055
454	3.722	$90,625	$64,832.44	$25,792.557

a. Dependent Variable: salary Current Salary

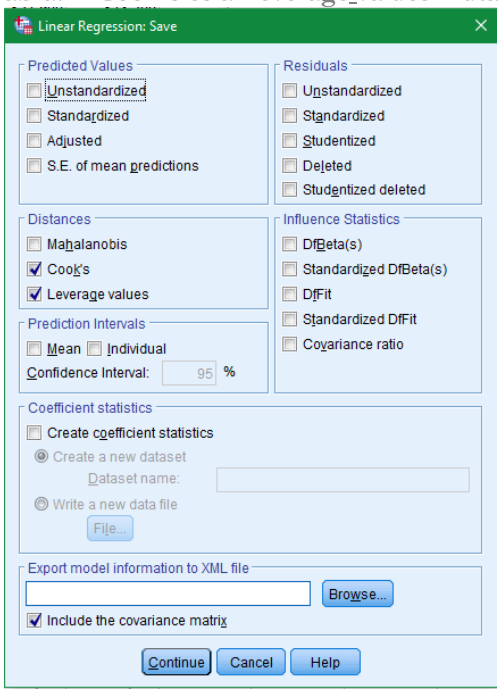
A feltüntetett sorszámokat az SPSS adattábla balszélén találjuk, szürke háttérrel kiemelve. Amennyiben az adattáblánkat valamely változó szerint sorba rendeztük (Data - Sort Cases) így, az SPSS „külső" sorszámai más eseteket jelölnek. A Casewise Diagnostics táblázat az adattábla legelső változója, az *id* szerint növekvő sorrendbe rendezett adattábla sorszámait mutatja.

A másik lehetőségünk a modellt torzító értékek kiszűrésére a **befolyásos esetek hatóerejének** *(leverage)* vizsgálata. Hatóerő alatt egy adott változó, adott értékének (x_{ij}) távolságát értjük a változó értékeinek átlagától. Minél nagyobb ez a távolság, annál nagyobb az adott érték hatóereje. Grafikusan értelmezve, a középponttól

távol eső pont „elhúzza” a pont irányába a regressziós felületet[53], ezzel „torzítva” a felület becslését (Zrínyi et al., 2012).

Két eljárás mutatunk be, amivel a nagyon eltérő vagy a nagy hatóerejű eseteket azonosíthatjuk. A lineáris regresszió Save ablakában állítsuk be a Cook`s és a Leverage values és opciókat.

57. ábra. A Cook`s és a Leverage values mutatók



Újrafuttatva a modellt az eredmény két új változó az adattáblában. Általános szabály szerint amennyiben a **hatóerő *(leverage)* változójának** (LEV_1) értékei kisebbek 0.2-nél biztonságos, 0.2 és 0.5 között kockázatos, 0.5 felett pedig veszélyes esetekkel van dolgunk. Ezeket az eseteket legegyszerűbben úgy szűrhetjük ki, ha sorba rendezzük az adattáblát a LEV_1 változó alapján. Példánkban csak egyetlen 0.2 feletti érték van (0.278), amely a legnagyobb fizetésű esethez tartozik. Hasonlóképp járunk el a **Cook távolság** mutató változójával, amelynél az 1 feletti értékek jelzik a modellünkre túl nagy hatást gyakorló eseteket. Példánkban nincs ilyen érték.
A kiugró vagy túl hatásos adatokat ajánlott kihagynunk, törölnünk a regressziós elemzésből.

[53] A többváltozós modell esetében már nem regressziós egyenesről beszélünk, hanem felületről.

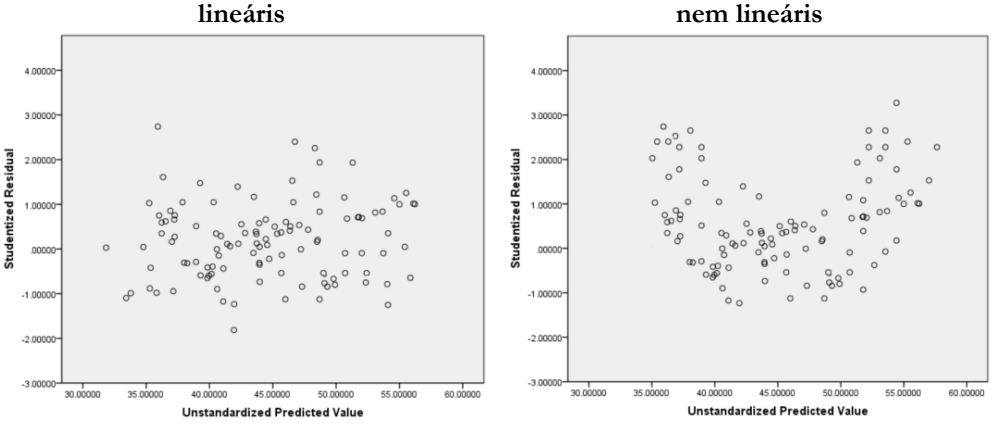
5.4.5 A változók közötti lineáris kapcsolat

A lineáris regresszióanalízis lényege, hogy **a függő és a független változó közötti kapcsolatot lineáris függvénnyel**, és annak grafikus képével egy egyenessel fejezi ki, amint láttuk a fejezet elején is a 47. ábrán. Többváltozós regressziós modell esetén kétféleképp értelmezhető és tesztelhető a függő és a több független változó közötti linearitás:

- **Együttes** lineáris kapcsolat a függő és valamennyi független változó között. Grafikus módon vizsgáljuk a linearitást, pontdiagrammal ábrázolva a studentizált reziduumokat a standardizálatlan becsült értékekkel szemben.
- **Külön** vizsgáljuk valamennyi független változó és a függő változó közötti linearitást, az úgynevezett parciális regressziós ábra segítségével.

Mindkét módszerrel nekünk kell döntenünk a grafikus kép alapján. Lineárisnak két változó közötti kapcsolatot, ha az együttes eloszlást jelző pontok jellemzően befférnek egy vízszintes sávba.

58. ábra. Lineáris és nem lineáris kapcsolatok

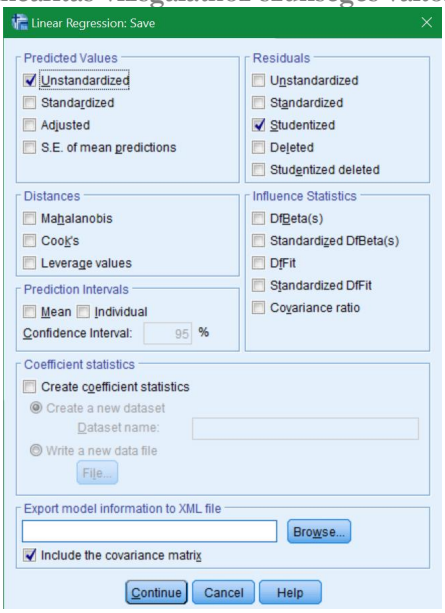


Forrás: LAERD Statistics

1. A linearitás együttes vizsgálata

A normalitás vizsgálat alfejezetnél már említett SPSS grafikonszerkesztőt, a Chart Builder-t alkalmazzuk a pontdiagram létrehozására, de ehhez szükségünk van a studentizált reziduumokra és a becsült értékekre. Ezért - megint csak - újrafuttatjuk a regressziós modellt, de a Save ablakban jelöljük az Unstandardized Predicted Values és a Studentized Residuals opciókat.

59. ábra. A linearitás-vizsgálathoz szükséges változók mentése

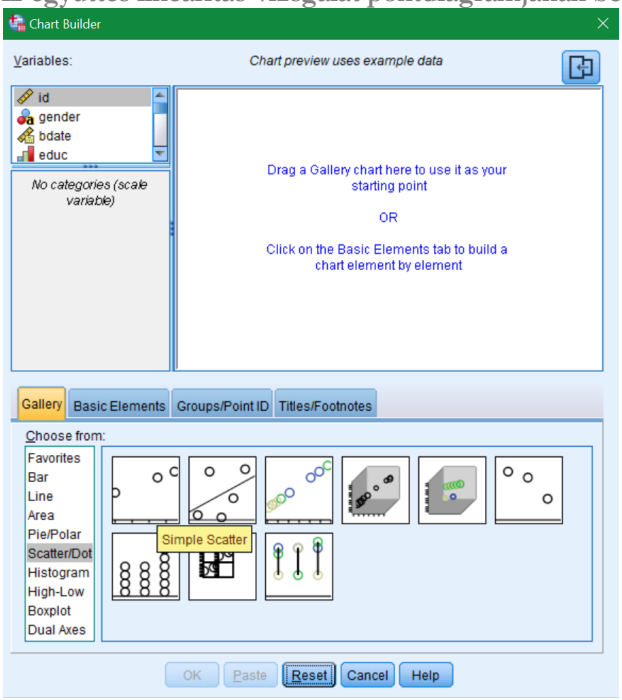


Ennek eredményeképp az adattáblában megjelennek a studentizált reziduumok *(SRE_1)* és a modell becsült értékeinek *(PRE_1)* változói. Következő lépésben létrehozzuk a két változó közötti pontdiagramot.

Graphs → Chart Builder

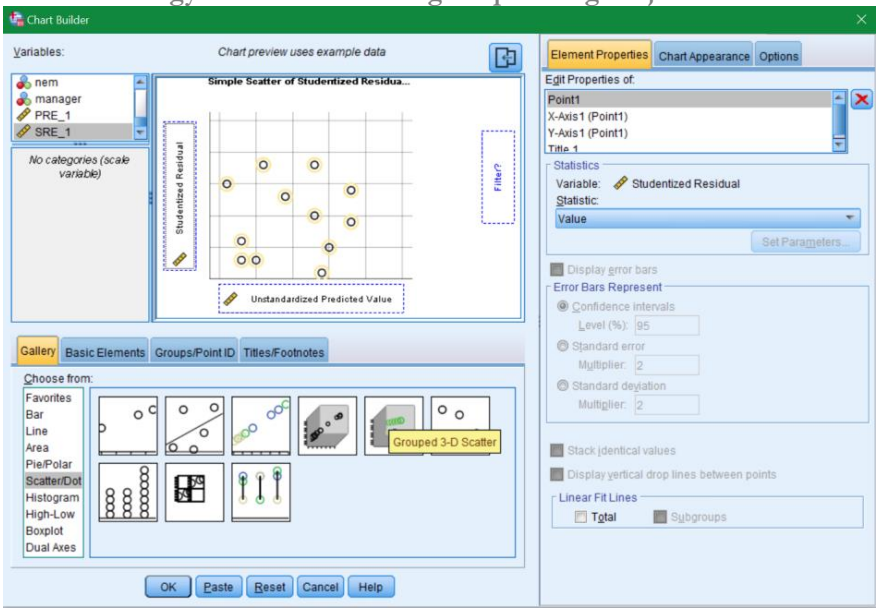
A Chart Builder ablak bal alsó részében, a diagramok listájából válasszuk ki a Scatter/Dot típust, majd az első, Simple Scatter altípust.

60. ábra. Az együttes linearitás-vizsgálat pontdiagramjának beállításai 1.



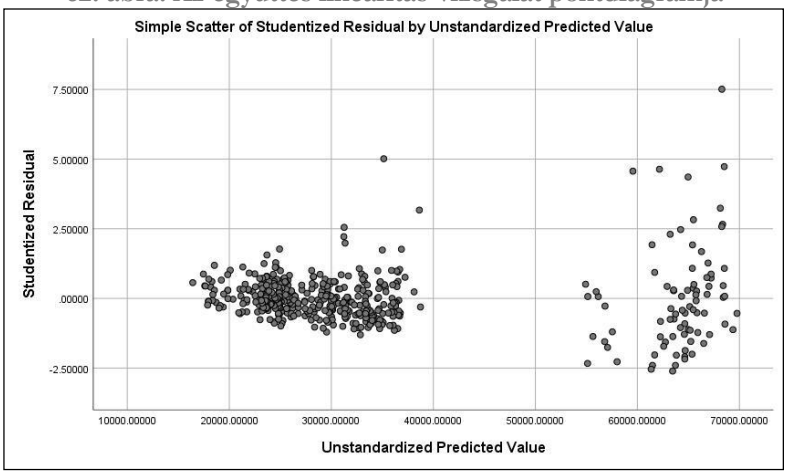
Rákattintva a Simple Scatter-re, az egerünkkel húzzuk be a fölötte levő üres mezőbe. Kibővül a Chart Builder ablak a következő ábrának (61.) megfelelően.

61. ábra. Az együttes linearitás-vizsgálat pontdiagramjának beállításai 2.



A bal felső mezőben levő változólistából behúzzuk a standardizálatlan becsült értékek változóját (PRE_1) a vízszintes tengelyre, a studentizált reziduumot (SRE_1) pedig a függőleges tengelyre, majd OK.

62. ábra. Az együttes linearitás-vizsgálat pontdiagramja



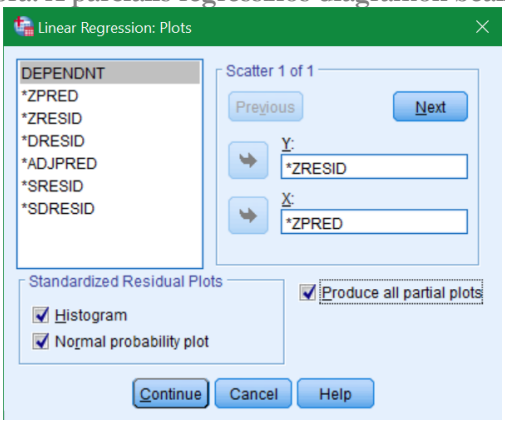
Láthatóan két, teljesen elkülönülő részből áll a ponthalmaz, amelyek külön-külön lineárisnak tekinthetők. Az eddigi elemzéseink alapján (pl. ANOVA) van egy sejtésünk, hogy a szellemi és fizikai munkások adatai képezik az egyik, és a menedzsereké a másik pontfelhőt. A linearitás mindkét ponthalmazban érvényesül, de ez az eredmény azt indokolja, hogy két külön regressziós modellel vizsgáljuk a menedzserek, illetve a beosztottak fizetéseit.

De nézzük előbb a linearitást valamennyi független változóra külön vizsgáló módszert.

2. A linearitás változónkénti vizsgálata

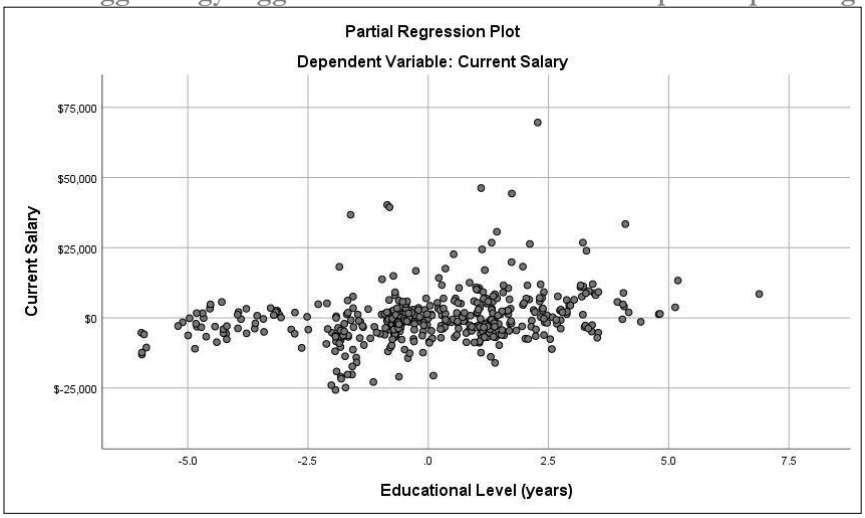
Az SPSS regresszióanalízise ábrázolja a **parciális regressziós diagramokat**. Ehhez az eddigi beállításainkat kiegészítjük azzal, hogy kérjük a Plots ablakban a Produce all partial plots opciót.

63. ábra. A parciális regressziós diagramok beállítása



Az SPSS legyártja valamennyi független változó és a függő változó közötti pontdiagramot, még a dummy változókkal is, amelyeket azonban nem kell figyelembe vennünk. Esetünkben a három releváns ábrából egyet tüntettünk fel, az iskolai végzettségét.

64. ábra. A függő és egy független változó közötti lineáris kapcsolat pontdiagramja

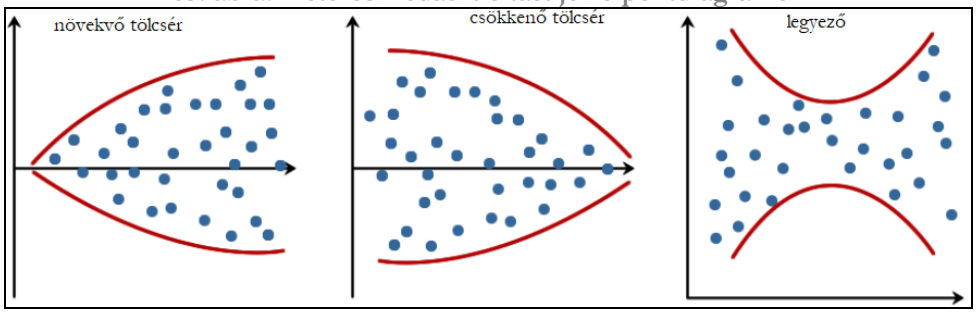


Az ábra alapján kimondhatjuk a jelenlegi fizetés és az iskolában töltött évek száma közötti lineáris kapcsolatot.

5.4.6 Homoszkedaszticitás

Ez a feltétel is a reziduumokra vonatkozik, a **hibatagok egyenlő szórását** jelenti. A feltétel nem teljesülése esetén **heteroszkedaszticitásról** beszélünk, ami a regressziós együtthatók rossz becslését eredményezi.
Magyarán úgy fogalmazhatjuk meg ezt a feltételt, hogy a regressziós modell egyformán jól magyaráz a függő változó valamennyi értéke esetén. Példánkban ez azt jelenti, hogy a lineáris regressziós modellünk azonosan jól magyarázza a kisebb, illetve a nagyobb fizetéseket. A **homoeszkadiszticitás grafikus tesztje** a standardizált reziddumokat vizsgálja a standardizált becsült értékekkel szemben, és akkor teljesül a feltétel, ha a pontok egy vízszintes sávba tömörülnek, vagyis nincsenek lényeges eltérések a ponthalmazban a vízszintes tengely mentén. A 65. ábrán látható példák esetében sérül ez a feltétel.

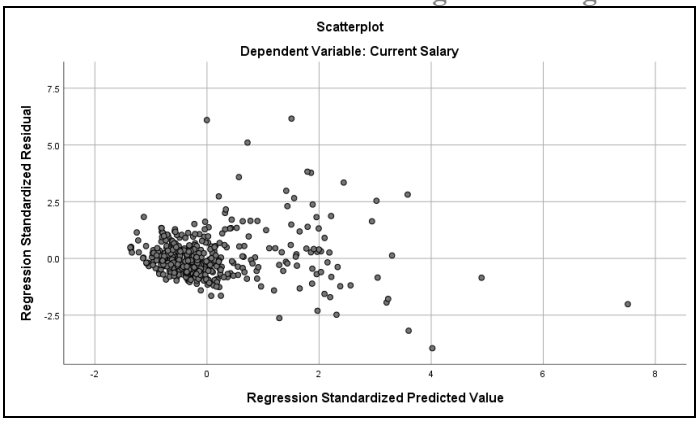
65. ábra. Heteroszkedaszticitást jelző pontdiagramok



Forrás: LAERD Statistics

Példánkban már a regressziós modell beállításánál kértük a Plots ablakban a standardizált becsült értékek változója (ZPRED) és a standardizált reziduumok változója (ZRESID) közötti pontdiagramot (51. ábra).

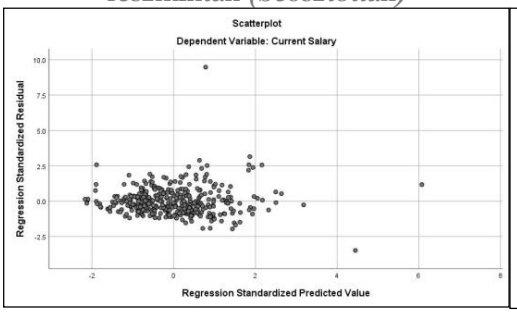
66. ábra. A homoszkedaszticitás grafikus vizsgálata



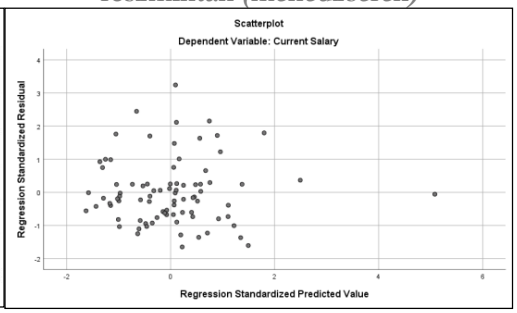
Bármennyire is nem szeretnénk észrevenni, de a pontfelhő bizony elég tölcsérszerű képet mutat, semmiképp sem egyenletes szóródást. Akárcsak a linearitás vizsgálat során (5.4.5 alfejezet) ezúttal is láthatóan két részből áll a ponthalmaz, kb. a vízszintes tengely nulla értékénél válik el a két rész. Ezúttal leellenőrizzük azt a sejtésünket, hogy a menedzserek és a beosztottak képezik a két ponthalmazt.
Kétfelé bontjuk tehát az adattáblát a Data - SELECT CASES (vagy a Data – SPLIT FILE) paranccsal, a korábban létrehozott *manager* változó használatával. Majd lefuttatjuk a regressziós modellünket külön-külön a beosztottak és a menedzserek körében, kivéve a *manager* változót a modell független változói közül.[54]

A homoeszkadiszticitás grafikus tesztje a beosztottak és a menedzserek körében.

67.a. ábra. Homoszkedaszticitás egy részmintán *(beosztottak)*



67.b. ábra. Homoszkedaszticitás egy részmintán *(menedzserek)*



A két részmintán lefuttatott regressziós modellek már megfelelnek a homoeszkadiszticitás feltételének (akárcsak a többi feltételnek). Láthatjuk, hogy a lineáris regresszió feltételeinek vizsgálata nemcsak egy statisztikai feltétel ellenőrzését jelenti, hanem **lényeges új információkat kaphatunk a vizsgált változóinkról, jobban megértjük az adatokat**. A továbbiakban mégis visszatérünk a teljes mintához, és a *manager* változó is visszakerül a magyarázó változók közé, hogy a feltételek nem teljesülése esetén alkalmazható megoldásokat bemutathassuk.

Ezúttal is vannak **analitikus módszerek** a homoszkedaszticitás tesztelésére, az egyik a nagymintákra ajánlott **Breusch-Pagan teszt**. Kisminták, nem normális eloszlások esetén a Koenker-tesztet alkalmazzuk (Daryanto, 2020).

[54] Ha erről megfeledkeznénk, a regressziós modell úgyis hibát jelez.

Mi a Breusch-Pagan tesztet részletezzük, ami viszonylagos egyszerűsége ellenére nem érhető el az SPSS menüből.[55] Megtaláljuk az interneten a tesztet SPSS syntax vagy SPSS makróként rögzítve, de a teszt lépéseit mi magunk is végig tudjuk vinni:

1. Futtatjuk a regressziós modellt, és mentjük a **standardizálatlan reziduumokat**.
2. Képezünk egy új változót a **reziduumok négyzetéből**.
3. Egy **új regressziós modellt**[56] futtatunk, amiben a független változó a reziduuom négyzete, a független változók pedig az eredeti modell független változói.
4. Kiszámoljuk a **Khi-négyzet statisztikát**:

$$\chi^2 = n\, R^2_{új}\,, \qquad (17)$$

ahol az n – mintaelemszám, $R^2_{új}$ – az új regressziós modell determinációs együtthatója.

5. Meghatározzuk a kiszámolt Khi-négyzet (χ^2) statisztika és adott szabadságfok mellett (egyenlő a független változók számával) a ***p* szignifikancia értéket**. Ezt megtaláljuk statisztikai táblázatokban, vagy használhatunk online kalkulátorokat. Pl. https://www.statology.org/chi-square-p-value-calculator/

A Breusch-Pagan teszt nullhipotézise a homoszkedaszticitás, tehát $p < 0.05$ esetén elutasítjuk a nullhipotézist, vagyis heteroszkedaszticitásról beszélhetünk (ha ki tudjuk ejteni). **Példánkban** a Breusch-Pagan teszt:

1. a lineáris regressziós modellt Save ablakában bejelöljük az Unstandardized Residuals opciót és futtatjuk a modellt.
2. Képezünk egy új változót (legyen *sqres*) a COMPUTE paranccsal.

   SQRES = RES_1 * RES_1
3. Új regressziós modellt futtatunk, amiben a független változó az sqres változó, a független változók pedig az eredeti modell független változói. Az új modell determinációs együtthatója $R^2=0.118$
4. Kiszámoljuk a Khi-négyzet statisztikát:

$$\chi^2 = n \cdot R^2_{új} = 474 \cdot 0.118 = 55.93$$

5. A Khi-négyzet 55.93, a szabadságfok 7 értékek mellett alkalmazva a kalkulátort (https://www.statology.org/chi-square-p-value-calculator/) a p

[55] Legalábbis a lineáris regresszió beállítási lehetőségei között, de megtalálható a General Linear Model Univariate eljárás alatt.
[56] A szakirodalom segédregressziónak is nevezi.

értéke 0.000000. Ennyi a nullhipotézis teljesülésének, vagyis a homoszkedaszticitásnak a valószínűsége.

Sikerült nem kevés munkával igazolnunk azt, ami a grafikus módszerrel is nyilvánvaló volt: nem teljesül a homoszkedaszticitás feltétele. Ha nem adjuk fel, akkor a következő lehetőségeink vannak:

- A függő változó transzformálása (pl. a Normalitásvizsgálat alfejezetben bemutatott módon).
- A modell átalakítása (pl. az előzőekben bemutatott módon külön modelleket specifikálunk a homogénebb részmintákra).
- Robusztus standard hibákat alkalmazunk – ezt az eljárást nem részletezzük.
- Robusztus regressziós modelleket alkalmazunk – ezeket a lehetőségeket sem részletezzük.
- A súlyozott legkisebb négyzetek módszerének alkalmazása.

5.5 Súlyozott legkisebb négyzetek módszere

A súlyozott legkisebb négyzetek módszerének lényege, hogy **nagyobb súllyal veszi figyelembe a pontosabb megfigyeléseket**, vagyis a kisebb varianciájú eseteket, és kisebb súlyt kapnak a nagyobb varianciájú esetek a regressziós együtthatók becslése során. Grafikusan értelmezhetjük úgy, hogy a varianciák inverzével való súlyozás a heteroszkedaszticitást jelző tölcsér alakú ábrát (62. ábra) kiegyenlítettebbé, a vízszintes tengellyel párhuzamosabbá változtatja.
Ehhez szükség van a modell egyik független változójának a varianciájára, egy olyan változóéra, amelyik a legnagyobb mértékben magyarázza a függő változó varianciáját. A súlyozott legkisebb négyzetek módszere két szakaszból áll:

- képezzük a súlyváltozót abból a független változóból, amelyik a legnagyobb hatással van a függő változónkra,
- újrafuttatjuk a regressziós modellünket a súlyváltozóval súlyozva.

Példánkban több gyakorlati lépésre bontjuk ezt a két szakaszt.

1. Az első lépésben **kiválasztjuk azt a változót**, amelynek varianciáit használjuk a súlyváltozóhoz. Könnyű dolgunk van, mert a modell független változói közül a legnagyobb magyarázóerővel, bétával rendelkező változót választjuk (lásd 51. táblázat). Esetünkben ez a kezdő fizetés (*salbegin*) változója.

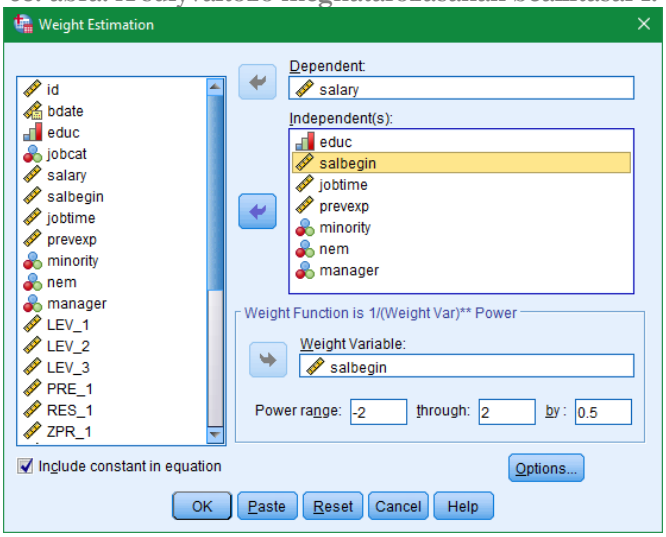
2. Létrehozzuk a súlyváltozót az **Analyze → Regression → Weight Estimation** eljárással. Ez az eljárás több transzformációt tesztel az optimális súlyváltozó meghatározásához. A súlyváltozó függvényének általános formája:

$$W = 1 \,/\, (salbegin)^{b}\,, \qquad (18)$$

ahol W – a súlyváltozó, b – hatványkitevő. A Weight Estimation eljárás egy általunk meghatározott terjedelemben teszteli a b hatványkitevőket, amíg megkapja a legjobb becslést eredményező függvényt.

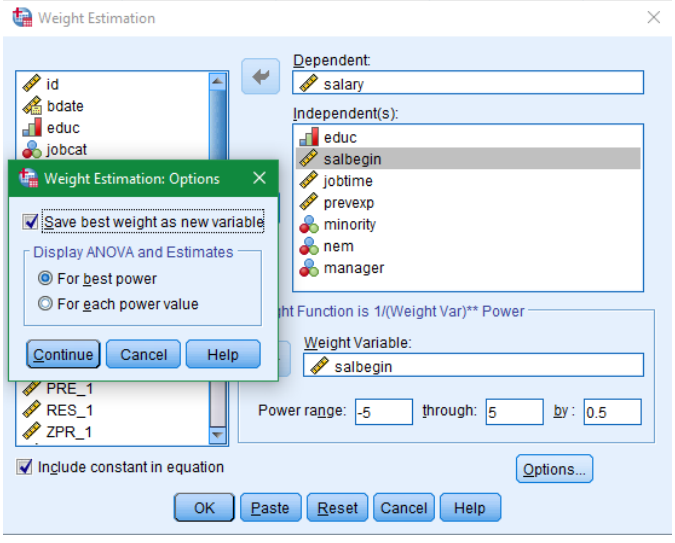
A Weight Estimation ablakban beállítjuk a regressziós modellünket a függő és független változóinkkal, a Weight Variable mezőbe pedig az előző lépésben kiválasztott változónk, a *salbegin* kerül. Ez a változó megmarad a független változónk, az Independent(s) mezőben is!

68. ábra. A súlyváltozó meghatározásának beállításai 1.



A Power range mezőkben beállíthatjuk azt a terjedelmet, amelyben teszteli az eljárás a különböző hatványkitevőket 0.5-ös léptékkel változtatva. Módosítsuk [-5, 5]-re a terjedelmet a következő ábrának megfelelően, hogy az eljárás több értéket teszteljen a legjobb súlyváltozó megtalálásához. (Közben hálát adva azért, hogy nem papír-ceruzával kell számolnunk.) Az Options ablakban beállítjuk a Save best weight as new variable-t.

69. ábra. A súlyváltozó meghatározásának beállításai 2.



Lefuttatva a WEIGHT ESTIMATION parancsot, az eredmény az adattáblánkban az új súlyváltozó (WGT_1 névvel), az Output-ban pedig a regressziós modell

eredményekhez hasonló táblázatokat találunk. Ezek közül az 54. táblázatot érdemes megtekintenünk, ami a súlyváltozót meghatározó függvény hatványkitevőjének a becslését mutatja.

54. táblázat. A transzformáló függvény hatványkitevőjének becslése

Log-Likelihood Values[b]

Power	-5.000	-5602.623
	-4.500	-5506.163
	-4.000	-5414.135
	-3.500	-5325.909
	-3.000	-5240.916
	-2.500	-5159.234
	-2.000	-5081.738
	-1.500	-5009.810
	-1.000	-4944.924
	-.500	-4888.314
	.000	-4840.779
	.500	-4802.610
	1.000	-4773.639
	1.500	-4753.367
	2.000	-4741.096
	2.500	-4736.030[a]
	3.000	-4737.365
	3.500	-4744.346
	4.000	-4756.301
	4.500	-4772.655
	5.000	-4806.573

a. The corresponding power is selected for further analysis because it maximizes the log-likelihood function.

b. Dependent variable: salary, source variable: salbegin

A táblázatban feltüntetett értékek közül az [a] betűvel jelölt, 2.5-ös hatványkitevő (*power*) eredményezi az optimális becslést. Nem részletezzük a log-likelihood függvényen alapuló becslési eljárást, csak annyiban, hogy a minimális értéke jelenti az optimumot. Láthatjuk, hogy érdemes volt a lehetséges hatványkitevők terjedelmét kiterjeszteni az alapbeállításról, mivel az optimum 2.5[57].

A súlyváltozót létrehozó függvény:

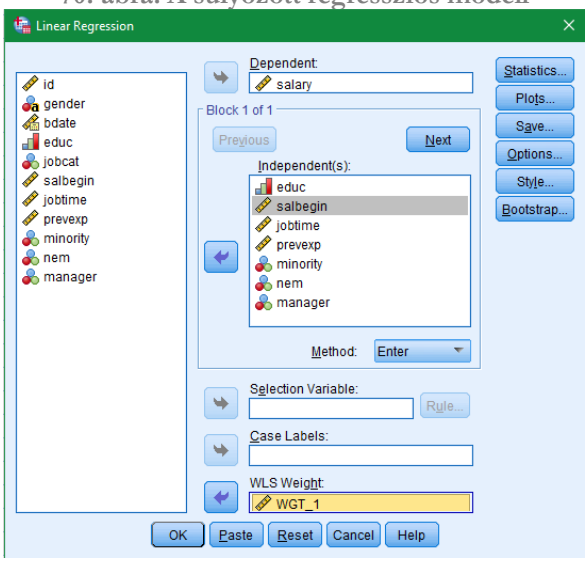
$$W = 1 / (salbegin)^{2.5}$$

Ez a táblázat a függvény meghatározásához nyújt információt, de további feladatunk nincs a függvénnyel, hanem az adattáblában lementett súlyváltozóval.

3. **Újrafuttatjuk regressziós modellünket a súlyváltozó alkalmazásával**. A lineáris regressziós modell ablakában az eddigi beállítások mellé bevisszük a WGT_1 súlyváltozót a WLS Weight mezőbe.

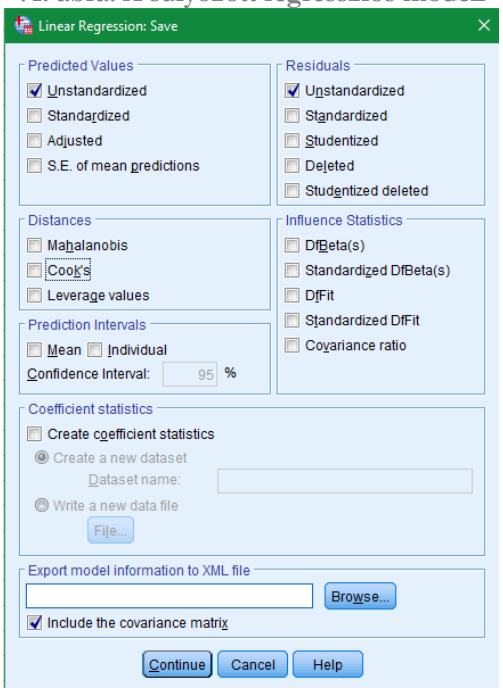
[57] Tovább finomíthatjuk az eljárást, ha például [2,3] terjedelemben 0.1-re állítjuk a léptéket.

70. ábra. A súlyozott regressziós modell



A Save ablakban ezúttal a standardizálatlan becsült értékeket (Predicted Values) és reziduumokat (Residuals) választjuk, amelyekre szükségünk lesz a homoszkedaszticitás grafikus ellenőrzéséhez. Ugyanis az SPSS súlyozott legkisebb négyzetek módszere nem hozza létre ezeket.

71. ábra. A súlyozott regressziós modell



Eredmények értelmezése

A Model Summary táblából láthatjuk, hogy az R^2 értéke kis mértékben csökkent, de még mindig nagyon jó a modell magyarázó ereje.

55. táblázat. A súlyozott modell főbb mutatói

Model Summary[b,c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.853[a]	.728	.724	$0.031	1.999

a. Predictors: (Constant), manager, jobtime Months since Hire, prevexp Previous Experience (months), minority Minority Classification, nem nem, educ Educational Level (years), salbegin Beginning Salary

b. Dependent Variable: salary Current Salary

c. Weighted Least Squares Regression - Weighted by WGT_1 Weight for salary from WLS, MOD_5 SALBEGIN** -2.500

A súlyozott modell együtthatóinak táblázata ugyanazokat a statisztikákat tartalmazza, mint a legkisebb négyzetek módszerével becsült modelleké. A következő (56.) táblázat együtthatóit nem indokolt értelmeznünk, mivel az előzőekben több feltétel (linearitás, normalitás) nem teljesült, csak az egységes példa kedvéért használtuk az eredeti modellt.

56. táblázat. A súlyozott modell együtthatói

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-7664.982	2337.692		-3.279	.001		
	educ Educational Level (years)	321.211	110.486	.086	2.907	.004	.673	1.486
	salbegin Beginning Salary	1.400	.108	.545	12.950	.000	.331	3.020
	jobtime Months since Hire	112.494	23.230	.121	4.843	.000	.940	1.063
	prevexp Previous Experience (months)	-10.485	2.191	-.125	-4.785	.000	.863	1.159
	minority Minority Classification	-1101.157	528.963	-.053	-2.082	.038	.910	1.098
	nem nem	2698.359	600.263	.140	4.495	.000	.602	1.662
	manager	11106.910	1590.654	.230	6.983	.000	.538	1.860

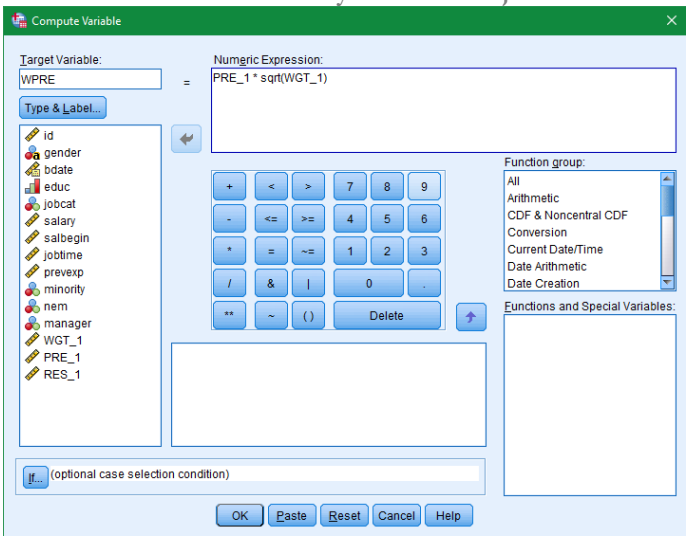
a. Dependent Variable: salary Current Salary

b. Weighted Least Squares Regression - Weighted by WGT_1 Weight for salary from WLS, MOD_5 SALBEGIN** -2.500

4. Végül **teszteljük a homoszkedaszticitást**, hiszen ez indokolta a súlyozott legkisebb négyzetek módszerének alkalmazását.

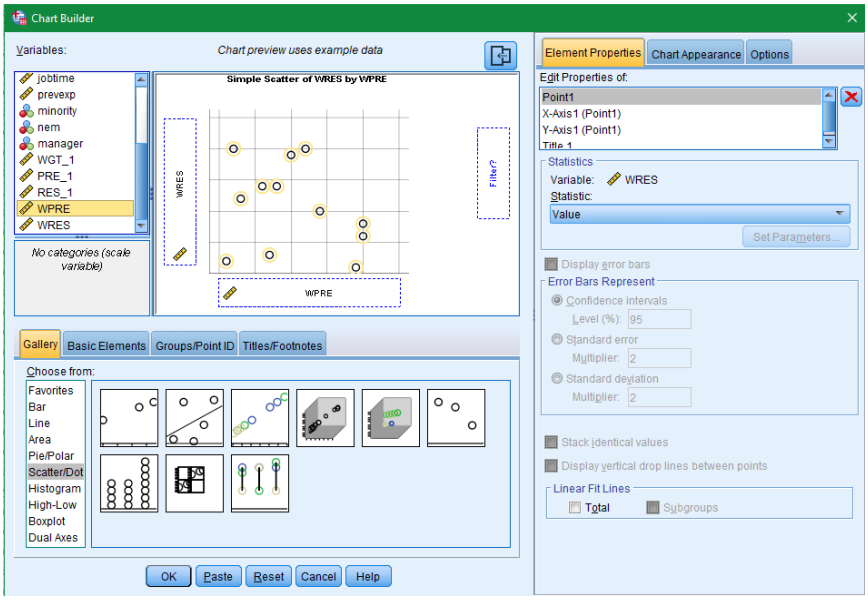
Ehhez szükségünk van a **standardizálatlan becsült értékek és reziduumok változóira, de súlyozott formában**. Ezért a COMPUTE paranccsal mindkettőt összeszorozzuk a súlyváltozó négyzetével. A 72. ábra a becsült értékek változójának átalakítását mutatja.

72. ábra. A becsült értékek súlyozott változójának létrehozása

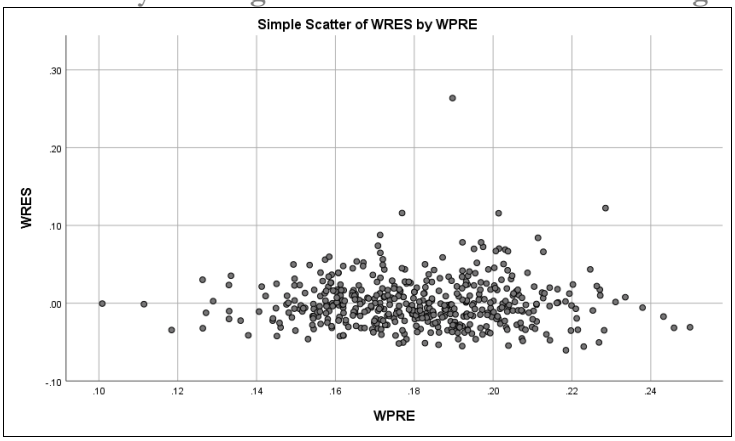


A két új változóval (nevezzük őket WPRE-nek és WRES-nek vagy bárhogy) létrehozzuk a homoszkedaszticitás grafikus ellenőrzésére alkalmas pontdiagramot, a Chart Builder diagramszerkesztővel.

73. ábra. A homoszkedaszticitás ellenőrzése



74. ábra. A súlyozott regresszió homoszkedaszticitás vizsgálata



A 74. ábra pontdiagramja megnyugtató képet mutat, néhány oulier-től eltekintve egy vízszintes sávban szóródnak a reziduumok, vagyis a becsült értékek nagyságától független varianciát igazoltan feltételezzük.

5.6 A regressziós modell egy gyakorlati alkalmazása: a keresleti függvény meghatározása

Az SPSS függvényillesztés-módszerével a kutatásból származó megfigyelési adatainkra függvényt illeszthetünk. Ez az eljárás nagymértékben hasonlít a regresszióanalízisnél alkalmazott legkisebb négyzetek módszeréhez, az adatok és az illesztett függvény közötti távolság négyzetösszegét minimalizálja. A lineáris függvény esetében gyakorlatilag ugyanazt az eredményt kapjuk, de lehetőségünk van több, nem lineáris függvényt is kipróbálni: másod-, harmadfokú polinomiálist, hatványkitevős, exponenciális, logisztikus stb. függvényeket. A függvényillesztésnek sokféle közgazdasági alkalmazása van: keresleti vagy termelési görbék illesztése és ezáltal az optimális ár vagy termelési mennyiség meghatározása különböző előrejelzésekre stb.

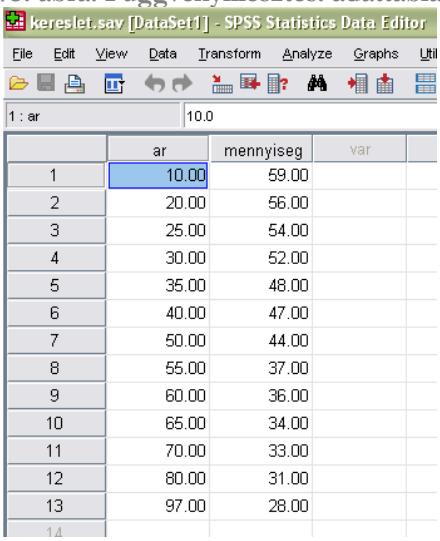
Példa: egy piackutatás során egy termékünk iránti keresletet vizsgáljuk, célunk a keresleti görbe meghatározása. Az 57. táblázatban feltüntettük, hogy adott áron a kérdezettek hány százaléka vásárolja meg a terméket. Szokatlannak tűnhet, hogy a keresletet a potenciális kereslet arányában fejezzük ki, de ezt könnyen átalakíthatjuk abszolút számokká.

57. táblázat. Keresleti adatok

Ár (RON)	1	2	2,5	3	3,5	4	5	5,5	6	6,5	7	8	9,7
Mennyiség (%)	59	56	54	52	48	47	44	37	36	34	33	31	28

Határozzuk meg az adatainkra legjobban illeszkedő keresleti görbét! Az elemzésünket kezdjük az adattábla létrehozásával, az árakat és a hozzájuk tartozó keresletet két változóba rögzítjük:

75. ábra. Függvényillesztés: adattábla



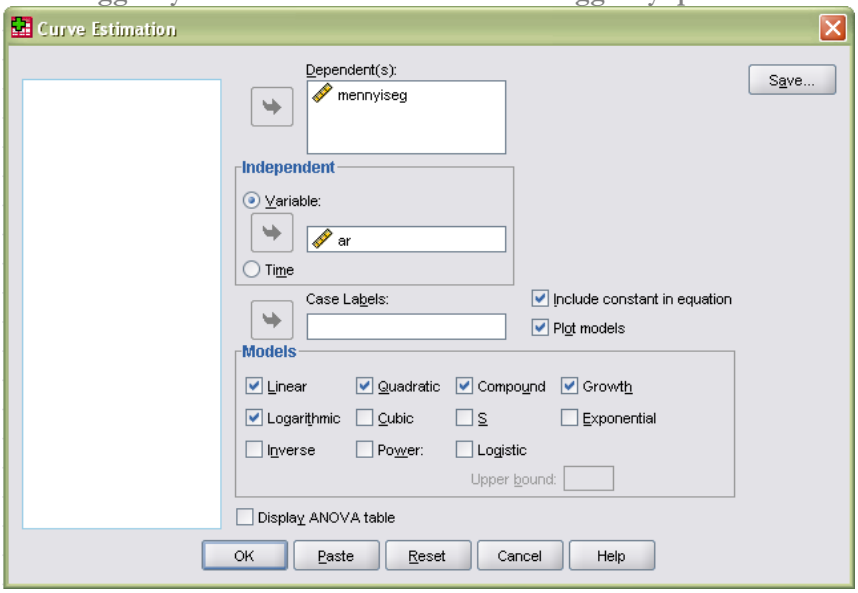
	ar	mennyiseg	var
1	10.00	59.00	
2	20.00	56.00	
3	25.00	54.00	
4	30.00	52.00	
5	35.00	48.00	
6	40.00	47.00	
7	50.00	44.00	
8	55.00	37.00	
9	60.00	36.00	
10	65.00	34.00	
11	70.00	33.00	
12	80.00	31.00	
13	97.00	28.00	

Az adattábla létrehozása után kezdhetjük a függvényillesztést.

Analyze→Regression→Curve estimation

A két változónk közül értelemszerűen a termékből eladott mennyiség az ár függvénye, ezért a mennyiseg változó kerül a Dependent(s), az ár pedig az Independent mezőbe.

76. ábra. Függvényillesztés: változók bevitele és a függvénytípusok kiválasztása



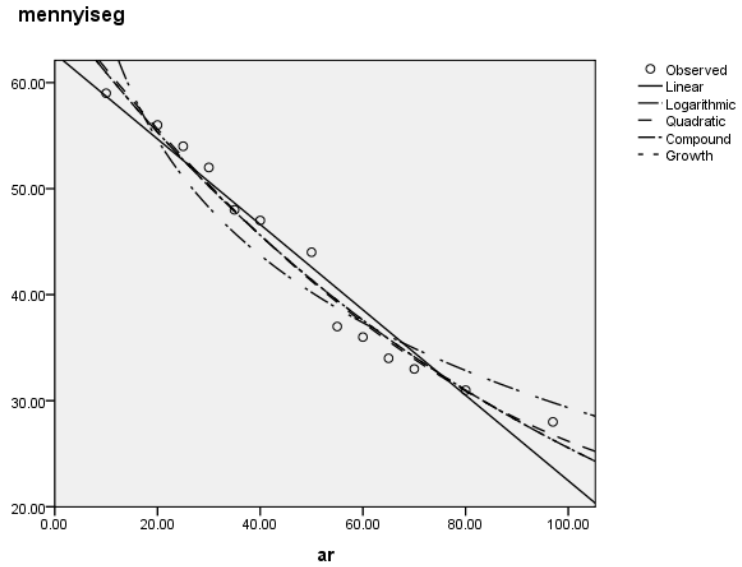
A függvényillesztés során 11 különféle függvénytípus közül választhatjuk ki az adatainkra legjobban illeszkedő függvényt. Egyelőre válasszuk ki az első ötöt, mivel valamennyi egyszerre áttekinthetetlenné tenné a grafikont. A kiválasztott függvények a következő általános alakkal rendelkeznek:

- Lineáris (Linear) $Y = b_0 + b_1 x$ *(19)*
- Másodfokú polinomiális (Quadratic) $Y = b_0 + b_1 x + b_2 x^2$ *(20)*
- Exponenciális (Compound) $Y = b_0 \cdot b_1^x$ *(21)*
- Növekedési (Growth, egyfajta exponenciális) $Y = e^{b_0 + b_1 x}$ *(22)*
- Logaritmikus (Logarithmic) $Y = b_0 + b_1 \ln x$ *(23)*

Eredmények értelmezése

Az adatsorunk és az öt típusú függvény grafikus ábrázolását az alábbi diagramon láthatjuk:

77. ábra. Függvényillesztés: az első öt függvénytípus



Karikával van jelölve az adatsorunk (Observed), és különböző módon az öt illesztett függvény. Ennek ellenére az ábra első ránézésre meglehetősen áttekinthetetlen, ezért nem ajánlott egyszerre két-három függvénytípusnál többet kipróbálni. Azonban a grafikus ábrázolás mellett egzakt mutató alapján

határozhatjuk meg, melyik a legjobban illeszkedő függvénytípus. A következő táblázatban az R-négyzet (R Square) a modell illeszkedésének jóságát mutatja 0 és 1 közötti értéktartománnyal, emellett az F-próba értékét és szignifikanciaszintjét, illetve a becsült paramétereket (Parameter Estimates) találjuk.

58. táblázat. Függvényillesztés: a függvénytípusok illeszkedésének jósága

Model Summary and Parameter Estimates

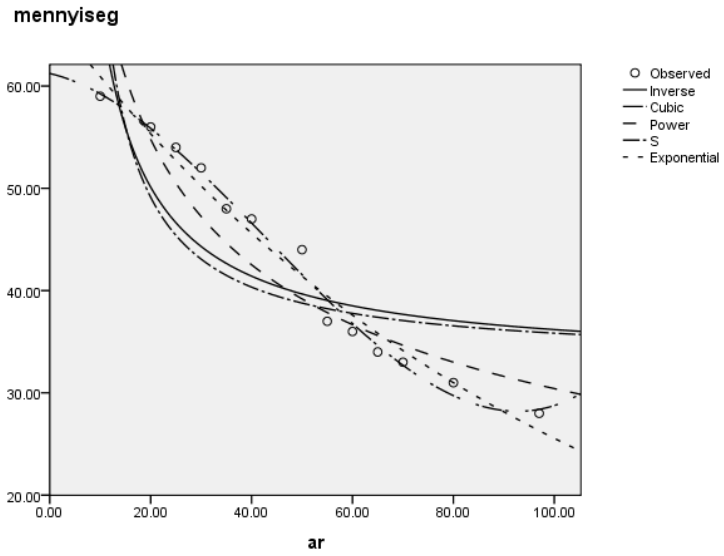
Dependent Variable:mennyiseg

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Linear	.958	248.388	1	11	.000	62.738	-.403	
Logarithmic	.911	112.635	1	11	.000	101.604	-15.690	
Quadratic	.977	210.334	2	10	.000	67.374	-.632	.002
Compound	.971	369.165	1	11	.000	67.108	.990	
Growth	.971	369.165	1	11	.000	4.206	-.010	

Az R-négyzetek vizsgálata alapján megállapíthatjuk, hogy valamennyi függvénytípus nagyon jól illeszkedik 0,90 feletti mutatóval, de a legjobb értéke a másodfokú polinomiális (Quadratic) függvénynek van (.977). Nézzük a következő általános formájú függvénytípusok illeszkedését:

- Harmadfokú polinomiális (Cubic) $Y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$ *(24)*
- S függvény (S, az exponenciális egy típusa) $Y = e^{b_0 + \frac{b_1}{x}}$ *(25)*
- Exponenciális (Exponential) $Y = b_0 \cdot e^{b_1 x}$ *(26)*
- Inverz (Inverse) $Y = b_0 + \frac{b_1}{x}$ *(27)*
- Hatványkitevős (Power) $Y = b_0 \cdot x^{b_1}$ *(28)*

78. ábra. Függvényillesztés: a második öt függvénytípus



A grafikon alapján is láthatjuk, hogy a harmadfokú polinomiális és az exponenciális függvények illeszkednek a legjobban. Ezt igazolja a következő táblázatban feltüntetett R^2 is:

59. táblázat. Függvényillesztés: a függvénytípusok illeszkedésének jósága

Model Summary and Parameter Estimates

Dependent Variable:mennyiseg

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Inverse	.652	20.622	1	11	.001	32.704	348.358		
Cubic	.988	257.368	3	9	.000	61.225	-.116	-.009	.000069
Power	.874	76.238	1	11	.000	163.916	-.366		
S	.590	15.817	1	11	.002	3.500	7.885		
Exponential	.971	369.165	1	11	.000	67.108	-.010		

A harmadfokú polinomiális függvény eredményezi a legjobb illeszkedést, a keresleti görbe tehát a következő formában írható fel.

$$Y = 0.000069X^3 - 0.009X2 - 0.116X + 61.2$$

A keresleti függvény a termékre vonatkozó egyik legfontosabb információ, az árpolitika alapja.

6. AZ EREDMÉNYEK PREZENTÁLÁSA, A TANULMÁNY MEGÍRÁSA

A kutatási folyamat legutolsó szakasza az eredmények bemutatás, prezentálása. Ennek az időben legutolsó szakasznak is nagy a jelentősége, akár a jó kutatási eredményeket tönkretehetjük egy rossz prezentálással. A kutatási eredmények közlésének alapvetően két formája van: a számítógépes diavetítéssel támogatott szóbeli prezentáció és az írásbeli tanulmány.[58] Leggyakrabban mindkettőre szükség van, az államvizsga-dolgozatot is követi egy tizenöt-húszperces prezentáció és védés. Mindkét esetben az eredményeket táblázatokban vagy grafikusan ábrázoljuk, ezért a következőkben a diagramkészítésről lesz szó.

6.1 Ábrák, táblázatok

Közgazdasági, társadalomtudományi dolgozat szinte elképzelhetetlen táblázatok, diagramok,[59] folyamatábrák vagy egyéb illusztrációk nélkül. Közismert, hogy egy jó ábra többet és gyorsabban mond egy féloldalnyi szövegnél. A szöveg tagolásában is megvan a szerepe, segíti az olvasót adott szövegegységen belüli struktúra áttekintésében.

Az **ábrakészítés alapelve, hogy úgy mutassunk be minél több adatot, hogy az jól áttekinthető, és a szövegkörnyezet nélkül is teljesen érthető legyen.**

Diagramkészítéskor sok lehetőségünk van egyéni megoldásokra, de van néhány szabály, amit ha nem tartunk be, hiányos vagy félreértelmezhető lesz a diagramunk:

- az adataink szerkezetének megfelelő diagramtípust használjunk;
- minden diagramnak tartalmaznia kell címet, az adatok mértékegységét a megfelelő tengely mellett, az adatok forrását és százalékok esetén a bázist, ami azt mutatja, hogy mi jelenti a 100%-ot, több adatsor esetén pedig a jelmagyarázatot;
- egy információ se szerepeljen egy ábra több helyén;
- egy prezentáció során ne használjunk többféle stílust a szövegeknél és legyünk következetesek a különféle beállításoknál;
- hasonlóképp lehetőleg ne használjunk 4-5-féle színnél többet;

[58] Természetesen más, kreatív formái is vannak az eredmények bemutatásának például a tudományos konferenciákon gyakran használt poszterek.

[59] A diagram, ábra, grafikon fogalmakat szinonimákként használjuk.

- az illusztrációkat típusonként (ábra, táblázat stb.) külön sorszámozzuk, az egész dolgozatban egységesen vagy fejezetenként. Az illusztráció sorszáma és típusa után pontot teszünk, pl. *1. ábra. A kutatás folyamata*;
- Amennyiben nem a saját kutatásból származó adatokkal dolgozunk, akkor mindig feltüntetjük a forrást is.

Diagramkészítésre többféle applikáció állhat rendelkezésünkre, egyetemi hallgatók leggyakrabban az Excel vagy a PowerPoint grafikonszerkesztőjét használják, ami gyakorlatilag ugyanaz. A kutatás folyamatába jobban illeszkedőnek tartjuk a PowerPoint használatát, mivel könnyebben lehet egységesíteni, kezelni az ábráinkat, és az adatokat amúgy sem Excel-ből, hanem az SPSS Output-ból másoljuk át. A következőkben röviden, a teljesség igénye nélkül felsorolunk néhány, a leggyakrabban használt diagramtípusokkal kapcsolatos megjegyzést, feltételezve, hogy az Olvasó többé-kevésbé gyakorlott az ábraszerkesztésben.

A **függőleges es vízszintes oszlopdiagramok** *(column, bar)* esetében három lényegesen eltérő altípus van:

- a csoportosított oszlop *(clustered)* – azonos kategória tengelyen különböző adatsorokat (pl. változók gyakoriságait) jelenít meg;
- halmozott *(stacked)* oszlop – olyankor használhatjuk, ha a különböző adatsorok nem csak külön, hanem összegezve is értelmezhetők;
- 100%-ig halmozott oszlop *(100% stacked)* – a halmozott adatsorok összege 100%, ezért az arányok bemutatására alkalmas. Gyakorlatilag megegyezik egy ábrán feltüntetett több kördiagrammal, de ezzel jobban össze tudjuk hasonlítani a kategóriákat. Kereszttábla eredményeinek bemutatására is használjuk, a sor- vagy az oszlopszázalékok feltüntetésével.

A függőleges és a vízszintes oszlopdiagram közötti választásnál néhány szempontra érdemes odafigyelni:

- idősor ábrázolásakor a dátumokat, időintervallumokat a vízszintes tengelyen ábrázoljuk, ezért függőleges oszlopdiagramot használunk;
- hasonlóképp a vízszintes tengelyen tüntetjük fel bármely ordinális vagy numerikus változó értékeit, vagy az azokból képzett kategóriákat;
- nominális változó esetén is használható a függőleges oszlopdiagram, de ha 4–5-nél több kategóriánk van, akkor praktikus szempontból a vízszintes oszlopdiagram ajánlott, mivel a vízszintes tengelyen nem fér el több kategória neve.

A vízszintes oszlopdiagramban az adatokat fentről lefele csökkenő sorrendben ábrázoljuk.

A **kör- vagy perecdiagrammal** nominális változó valamennyi kategóriájának relatív gyakoriságait ábrázoljuk. Áttekinthetőbb az ábra, ha a kördiagram értékeit is csökkenő sorrendben ábrázoljuk, a legnagyobb értéket (körcikket) a függőleges, 12-es óraállásnál kezdve, és az esetleges „Egyéb" vagy „NT/NV" kategóriákat a végére hagyva. Túl sok érték, körcikk esetén áttekinthetetlen lesz az ábránk, ilyenkor inkább vízszintes oszlopdiagramot vagy táblázatot válasszunk. Az egymásba ágyazott perecdiagramokkal jól szemléltethető a kategóriák eloszlásainak változása.

Pontdiagrammal *(scatter)* már sokszor találkoztunk ebben a jegyzetben: két numerikus változó együttes megoszlását egy koordináta rendszerben jelenítettük meg. Ilyen esetben a **buborék** *(bubble)* diagrammal egy harmadik változó értékeit is bevonhatjuk a buborék méreteként. Gyakran használjuk a pontdiagramot úgy is, hogy egy nominális változó kategóriái vannak a vízszintes tengelyen és a hozzájuk tartozó numerikus értékek a függőlegesen. Pl. egy adott év inflációja országonként.

A **vonaldiagramok** *(line)* talán a leggyakrabban használt diagramtípusok közgazdasági adatok ábrázolásánál, idősorok elsőszámú ábrája. Olyan kivételes esetben ne használjunk vonaldiagramot, amikor az idősor értékei között nincs értelemszerű átmenet, nincs folytonosság. Például egy cég negyedéves árbevétel adatainak ábrázolásakor használhatunk vonaldiagramot, mert a forgalom alakulásában van folytonosság – még ha valószínűleg nem úgy lineáris, mint az ábrán - , de a profit negyedéves (tőzsdei cégek esetében kötelezően nyilvános) adatai esetében már oszlopdiagram ajánlott, mert a folytonos működésből származó pozitív eredmény egy beruházási döntés hatására „egy nap alatt" veszteségesre fordulhat.

A **területdiagram** *(area)* funkciójában nem sokban különbözik a vonaldiagramtól, talán jobban kihangsúlyozza az adatsorok közötti különbséget, távolságot, és alkalmasabb az egymásra épülő adatsorok bemutatására is.

Sugárdiagrammal (radar vagy pókháló ábrának is nevezik, angolul *radar*) több metrikus változó statisztikáit (leggyakrabban átlagait) egyszerre hasonlíthatjuk össze. Több adatsor esetén is jól áttekinthető az ábra, könnyen meg tudjuk állapítani, hogy melyik dimenziók mentén van a legnagyobb különbség/hasonlóság az adatsorok között. Például egy termékkel kapcsolatos fogyasztói elégedettség több szempont szerint, férfiak és nők esetében.

Tőzsdézők jól ismerik az **árfolyam** *(stock)* diagramot, ami numerikus változók főbb statisztikáinak bemutatására használunk. A többféle altípus közös jellemzője, hogy

több eloszlás 4-5 statisztikáját hasonlítják össze. Például több részvény adott napi árfolyamait, vagy adott részvény árfolyamát több értékesítési napon.

A **kombinált** ábrák jól megfelelnek annak az alapelvnek, hogy minél több információt áttekinthető módon ábrázoljunk. Jellemző, hogy egy adatsor két-három típusú diagrammal is bemutatható, ez lehetővé teszi, hogy az összehasonlítandó adatsorokat eltérő típusú diagramokkal jelenítsünk meg. Pl. függőleges oszlopdiagram és vonaldiagram kombinációja.

6.2 Tanulmányírás

A kutatási folyamatban elérkeztünk a kutató számára legidőigényesebb szakaszhoz, a tanulmány, a dolgozat megírásához. A legtöbb kreativitásra is ekkor van szükség, ha a tanulmányírást és az adatelemzést szimultán végezzük. Nagyobb vagy bonyolultabb témájú kutatásoknál ajánlott az adatelemzés eredményét, az elfogadott vagy elutasított kutatási hipotézisekből származó információt azonnal megfogalmazni és leírni, mert ez további elemzési ötleteket generálhat. Csak nagyon egyszerű vagy rutinkutatásoknál ajánlott különválasztani az adatelemzést a tanulmányírástól.

A dolgozatnak/tanulmánynak a következő **struktúrát** kell követnie:

- Címoldal
- Tartalomjegyzék. Régebben a tartalomjegyzéket a szakkönyvek végén jelenítették meg, ma már egyértelműen az elején közöljük az olvasóval a dolgozat, tanulmány szerkezetét.
- Bevezetés. Leírjuk a kutatás célját, a kutatási hipotéziseket, a módszertant, a mintát és a mintavételt, vagyis a kutatási tervet az olvasó számára érthetően fogalmazva.
- Kidolgozás. A dolgozat gerince, ez tartalmazza a részletes kutatási eredményeket. Primer kvantitatív kutatásnál táblázatban vagy diagramban ábrázolja valamennyi kérdésre adott válaszok eredményeit.
- Összefoglalás. Az összefoglalás a következőket tartalmazza: a kutatási célok ismétlése, összegzése, a kutatási hipotézis igazolása, elvetése, az önálló tudományos eredmény hangsúlyozása, végül a kutatás távlatai, nyitott kérdések megfogalmazása.

Eltérő helyen van a struktúrában a szakdolgozat és az üzleti kutatás tanulmányának az összefoglalója. Amíg a tudományos cikkeknél, dolgozatoknál az összefoglalás a kidolgozás után van, addig az üzleti célú elemzések az elején tartalmazzák a két-három oldalas vezetői összefoglalókat.

- Bibliográfia – a későbbiekben részletezzük.
- Függelék/melléklet. A kutatás különféle eszközeit tartalmazza, amelyek elengedhetetlenek a kutatás részleteinek a megértéséhez, de feleslegesen terhelnék a főszöveget. Pl. kérdőív, moderátori vezérfonal, a minta demográfiai ismérvek szerinti jellemzése stb.

A dolgozat makroszerkezetén belül a tartalomnak kell meghatároznia a struktúrát, ezért vegyük figyelembe a következő **tartalmi szempontokat**:

- A fejezetek nagyságának arányai összhangban kell legyenek tartalmi fontosságukkal. Egy szakdolgozat esetében baj, ha az elméleti rész kapja a legnagyobb terjedelmet, a bíráló sokkal inkább a hallgató egyéni hozzáadott értékére kíváncsi.
- A fő- és alfejezetek tagolásának összhangban kell lennie a tartalmi terjedelemmel és a logikai felépítéssel. Ennek az összhangnak a megtalálása nem mindig magától értetődő, jelen jegyzet tartalomjegyzéke jó példa e probléma bemutatására. Az a szerzői szándék, hogy a könyv gerincét a kutatás folyamata adja, nagyon aránytalanná tenne egyes fejezeteket. Ha a folyamat szakaszai jelentenék a fejezeteket, akkor lenne 3 oldalas (előzetes tájékozódás) fejezetünk is, illetve 33 oldalas is (a kutatási terv elkészítése).
- Figyelni kell a dolgozat logikus felépítésére. Kvantitatív kutatásoknál gyakran a kérdőív kérdésblokkjainak a sorrendjét követik, de ez nem mindig célravezető, hiszen a kérdőív strukturálást kérdezéstechnikai szempontok is befolyásolják, amelyekre a tanulmányban nincs szükség.

A szöveg tagolása

Ha azt szeretnénk, hogy a dolgozatunkat valaki el is olvassa, vagy a kijelölt bíráló se szenvedjen nagyon, akkor a szövegünket ne egyetlen súlyos tömbbe öntsük, hanem formailag is olvashatóbbá kell tennünk a szöveg tagolása által. Szövegünket elsősorban tartalmi szempontok szerint tagoljuk, ezt tehetjük hangsúlyosabbá a különböző formai eszközökkel.

A **címrendszer** a dolgozat tartalmi, logikai szerkezetének legfőbb jelzője. A különböző szintű belső címek a folyószöveget összetartozó szövegegységekre bontják, illetve meghatározzák a szövegegységek közötti kapcsolatokat. Ne feledjük, hogy dolgozatunk bírálója először is a címrendszert összefoglaló tartalomjegyzéket vizsgálja meg, ebből alkotja meg az első benyomását a tartalomról és annak logikai szerkezetéről! Néhány tanács a címekkel kapcsolatban:

- ne adjunk hosszú címeket, a magyarázatra a szövegben lesz lehetőségünk;
- a belső címeket sorszámozzuk is, ezáltal is segítve az olvasót a szövegen belüli tájékozódásban;
- a címek hierarchiája tükrözze a szövegegységek egymáshoz való viszonyát; azonos rangú szövegrészeknek azonos fokozatú címeket adjunk, alárendelt szövegeknek alacsonyabb rendűt;
- legtöbbször elégséges a háromfokozatú címrendszer használata, speciális (pl. jogi) szövegek igényelhetnek mélyebb hierarchiát:
    1. rendű cím: fejezetcím;
    2. rendű cím: alfejezetcím;
    3. rendű cím: szakaszcím.
- ajánlott a szövegszerkesztők címrendszerének és automatikus tartalomjegyzékének a használata, amivel gombnyomásra átvezethetők a módosítások.

Bekezdés. Nagyon gyakori hiba kezdő dolgozatírók körében, hogy szinte minden mondatukat új bekezdésbe írják, sokszor még növelt sortávval is fokozva az új mondat megszületése feletti örömöt. A bekezdés egy új gondolat, téma kezdésének a jelzésére szolgál. Ennek megfelelően a bekezdés hossza a gondolatmenet hosszától függ, ezért nem lehet általában meghatározni, de az egy oldalnál hosszabb bekezdést próbáljuk az okfejtés valamely logikai pontján megszakítani.

Felsorolás. A felsorolásokat mindig jelöljük számokkal, betűkkel vagy valamilyen tipográfiai jellel. Sorszámok után pontot, betűk után kerek zárójelet teszünk, pont nélkül. Pl. 1. vagy a)
A felsorolások tartalma és hossza meghatározza a formai megjelenítést. Ha csak címeket vagy fontos fogalmakat sorolunk fel, nem kell írásjeleket használnunk, például nem kell pontot tennünk a végére, de teljes mondatok esetén igen.

Például:

1. termékpolitika
2. árpolitika
3. értékesítési utak politikája
4. promóciós politika

vagy például:

a) A kvantitatív kutatás során a kutatási adatokat számszerűsítve, statisztikai módszerekkel elemezzük.

b) A kvalitatív kutatási módszerek sajátossága a szubjektív értelmezés, kis mintán alapulnak és az eredmények statisztikai értelemben nem általánosíthatók a teljes alapsokaságra.

Idézet. Az idézetekre vonatkozó elvárások főképp Umberto Eco (2012) nyomán:

- mindig idézőjelek közé tesszük az idézett szöveget, mondatrész idézése esetén csak az eredeti szöveget tegyük idézőjelek közé;
- az idézet pontos legyen, még a szöveg szórendjét és helyesírását sem változtathatjuk meg;
- mindig meg kell adni a szerzőt és a szöveg forrását;
- az idézet olyan hosszú kell legyen, hogy alátámassza érvelésünket;
- több kiadás esetén lehetőleg a legutolsót idézzük.

Háromféle idézőjelet ismer a magyar helyesírási szabályzat, ezek közül az első a leggyakrabban használt:
„Macskaköröm”-mel jelezzük a pontosan idézett szöveget, vagy a szerző sajátos szóhasználatát. A »lúdlábat« ritkán használjuk; ha az idézett szövegen belül is idézőjelet kell tennünk. A Word szövegszerkesztőben az Alt0187 és Alt0171 billentyűkombinációval tudjuk behívni.
'Félidézőjellel' inkább filológiai munkákban találkozunk egy kifejezés jelentésének megadásakor, illetve sajátos szóhasználat jelzésekor.
Az idézőjelek és az idézett szöveg közé ne tegyünk szóközt. Hosszabb idézett szöveget jobban különítsük el a folyó szövegtől, például behúzással, kisebb sorközzel, szöveg előtti és utáni térközzel.

Hivatkozás. A szakirodalomból származó fontos kijelentések, megállapítások után zárójelben hivatkozzunk a szerzőre, az idézett könyv megjelenési dátumára vagy a folyóirat évfolyamára és számára. Ezt nevezzük Harvard-módszer (Gyurgyák, 2019), amely a magyar nyelvű szakirodalomban is kiszorította a korábbi, lábjegyzeteket alkalmazó hivatkozási módot.
A hivatkozások célja az olvasó figyelmének felkeltése a szakirodalom iránt, és a tanulmányunkban nem igazolt fontos megállapítások alátámasztása. Ugyanebből a célból indokolt feltüntetnünk a táblázatok és grafikonok adatainak a forrását. A hivatkozott szöveget lehetőleg a saját szövegünk kontextusában, lehetőleg

átfogalmazva írjuk, szó szerinti idézés esetén pedig az előbbiekben tárgyalt, az idézetre vonatkozó szabályok érvényesek.
Figyeljünk az idézetek és a hivatkozások megfelelő használatára, a kutatási témánk szempontjából lényeges állítások, eredmények, adatok hivatkozás nélküli feltüntetését **plágiumnak** nevezzük, és szellemi tulajdon lopását jelenti. Kezdő kutatóknál ez a vétség legtöbbször nem szándékos, szakdolgozatok esetében az egyetemek plágiumellenőrző szoftverekkel, a dolgozat javítási lehetőségével, és a témavezető tanár felelősségének a növelésével segítik az eredeti tanulmány létrejöttét.

Kiemelés. A szöveg tagolásának ezzel az eszközével fontosnak ítélt szavakat, mondatrészeket emelhetünk ki. Jól ismert módjai a kurzív, *dőlt* (italic), a **félkövér** (bold) betű vagy az aláhúzás használata. Macskakörömmel vagy félidézőjellel is kiemelhetünk egyes szavakat. Használatuk legfontosabb szabálya, hogy dolgozatunkban következetesen alkalmazzuk a kiemeléseket.

Az **utalások** a hivatkozásoktól eltérően nem más szövegre, hanem a saját szövegünk valamelyik részére vagy beillesztett táblázatára, ábrájára mutatnak. Az utalásokat a megfelelő helyen zárójelbe tesszük. Ha az aktuális szövegünk kiegészíti, részletezi az utalt szövegrészt, akkor a *lásd* vagy a *vesd össze* (vö.) kifejezéseket használhatjuk a zárójelbe helyezett utalásnál.

A dolgozat írása, **szövegezése** során figyeljünk:

- A terminológiára, a szakkifejezések használatára.
- Az **érvelés módjára**. Érvelésünk legyen következetes, objektív az eredmények értelmezésében, keressük a fontos megállapítások magyarázatát, összefüggéseit. Törekedjünk az elfogulatlan, értéksemleges következtetésekre, legalábbis az empirikus kutatási eredmények értelmezésénél mindenképp, a személyesebb véleményünk megjelenhet az összefoglalóban, esetleges ajánlásokban.
- **Következetesség**. Megállapításaink legyenek koherensek – a kutató számára nagy elégtétel, ha a részeredmények összhangban vannak, egymást igazolják. Nemcsak tartalmi, hanem formai szempontból is elvárt a következetesség, a végén mindig egységesítsük dolgozatunkban a hivatkozásokat, kiemeléseket, jegyzeteket stb.
- A **stílus** legyen tudományos, ami a szakkifejezések használatán túl azt jelenti, hogy például nem használunk közhelyeket, nem mesélősen,

terjengősen fogalmazunk, hanem szabatos és tömör, de egyúttal világos és érthető mondatokra törekszünk. A tudományos stílus ma több személyességet enged, mint korábban, mivel dolgozatunk célja nemcsak a kutatási problémát tárgyaló szakirodalom bemutatása, hanem elsősorban a saját válaszunk, kutatási eredményünk elfogadtatása az olvasóval.

- A leggyakrabban előforduló **stilisztikai hibák**: helytelen szóhasználat, nyelvhelyességi hibák, mondatszerkezeti hibák, szóismétlés, stílustévesztés.

Bibliográfia

Végül következzék az irodalomjegyzékre, a bibliográfiára vonatkozó néhány megállapítás. Mindenekelőtt ne feledjük, hogy a bibliográfia funkciója a hivatkozott irodalom azonosíthatósága, megtapasztalhattuk a kutatási folyamat elején a szakirodalmi tájékozódásban egy jó bibliográfia jelentőségét. A bibliográfia alapegysége a **bibliográfiai tétel**, amely részletezettsége alapján lehet egyszerűsített vagy bővített.

Könyvek esetében az egyszerűsített bibliográfia a következő elemeket kötelezően kell tartalmazza az adott sorrendben.

- Szerző neve – nemzetközileg elfogadott, hogy előbb a családnevet, majd a keresztnevet tüntetjük fel. Ez a sorrend a magyaron kívül nem sok nyelvre jellemző, ezért ha idegen nyelvű szakirodalomra hivatkozunk, akkor a családnév után vesszővel választjuk el a keresztnevet (pl. Greene, William). A szerző neve után kettőspont, és következik a cím.
- A szakirodalom címét *dőlt, kurzivált* betűkkel írjuk, majd pont.
- A kiadás helye, ideje és a kiadó neve. A kiadás helye és éve után vesszőt teszünk, a kiadó neve után, a bibliográfiai tétel végére pedig pontot.
  Pl. Gyurgyák János: *A tudományos írás alapjai*. Budapest, 2019, Osiris.

Több szerző esetén kötőjellel választjuk el a neveket.

Az internetes tartalom gyors és végtelenbe tartó növekedése felveti az internetes hivatkozások kérdését is, amelyben a következő konszenzus alakult ki. Először feltüntetjük a tartalomszolgáltató nevét, ami általában egy *domainnév*, de nyilván lehet egy *blogíró* tulajdonneve vagy álneve is. Ezt követi az írás címe, a visszakeresést lehetővé tevő *URL*-cím lehetőleg teljes hosszában, és a letöltés dátuma.

Ebben a tanulmányírásra vonatkozó részben igazából csak a tanulmány formai követelményeit fogalmaztuk meg néhány általános jótanács kíséretében. Szakdolgozat írás előtt ajánlott a hivatkozott sokkal részletesebb szakirodalom idevágó részeit átolvasni (leginkább Majoros Pál, Umberto Eco), továbbá az interneten is találhatunk nagyon jó gyakorlati tanácsokat a megfelelő szakdolgozatstílus kialakításához.
A legfontosabb azonban az, hogy merjünk írni, az előbbiekben felsoroltakat ne olyan korlátoknak tekintsük, amelyek szűkre szabják lehetőségeinket, hanem támpontokat adnak saját kutatási eredményeink közléséhez.

A modern tudományos megismerés folyamatának elsajátítása után, egy háromezer éves posztmodern igazsággal relatívizálhatjuk mindezek fontosságát:

„Ezeken felül, fiam fogadd meg az intést: a sok könyv írásának nincsen vége, és a sok gondolkodás elfárasztja a testet." (Biblia, Préd. 12,14).

IRODALOMJEGYZÉK

BABBIE, Earl: *A társadalomtudományi kutatás gyakorlata.* Budapest, 1996, Balassi.

BONCZ Imre: *Kutatásmódszertani alapismeretek*. Pécs, 2015, Pécsi Tudományegyetem.

DRÓTOS György: *Az információrendszerek perspektívái.* Budapest, 2000, Corvinus Egyetem, Phd értekezés.

ECO, Umberto: *Hogyan írjunk szakdolgozatot?* Budapest, 2005, Kairosz.

GREENE, William: *Econometric analysis.* Upper Saddle River, USA, 2003, Prentice Hall.

GRØNHAUG, Kjell – PERVEZ, Ghauri: *Kutatásmódszertan az üzleti tanulmányokban*. Budapest, 2016, Akadémiai Kiadó.

GYURGYÁK János: *A tudományos írás alapjai*. Budapest, 2019, Osiris.

HAJDU Ottó: *Többváltozós statisztikai módszerek.* Budapest, 2003, Aula.

HOFFMANN Mária - KOZÁK Ákos - VERES Zoltán: *Piackutatás*. Budapest, 2000, Műszaki Könyvkiadó.

HORNYACSEK Júlia: *A tudományos kutatás elmélete és módszertana.* Budapest, 2014, Nemzeti Közszolgálati és Tankönyv Kiadó Zrt.

HUNYADI László - MUNDRUCZÓ György - VITA László: *Statisztika*. Budapest, 1996, Aula.

HUNYADI László: *Statisztikai következtetéselmélet közgazdászoknak. Statisztikai módszerek a társadalmi és a gazdasági elemzésekben*. Budapest, 2001, Központi Statisztikai Hivatal.

HUNYADI László: *A heteroszkedaszticitásról egyszerűbben.* Statisztikai Szemle, 2006, 84. Évf., 1 szám.

HUZSVAI László – VINCZE Szilvia: *SPSS-könyv*. 2012, Seneca.

KŐRÖSI Gábor - MÁTYÁS László - SZÉKELY István: *Gyakorlati ökonometria*. Budapest, 1990, Közgazdasági és Jogi Könyvkiadó.

MAJOROS Pál: *A kutatásmódszertan alapjai*. Budapest, 2001, Perfekt Kiadó.

MALHOTRA, Naresh: *Marketingkutatás*. Budapest, 2001, Műszaki Könyvkiadó.

RAMANATHAN, Ramu: *Bevezetés az ökonometriába*. 2003, Budapest, Panem.

SAJTOS László - MITEV Ariel: *SPSS kutatási és adatelemzési kézikönyv*. Budapest, 2007, Alinea.

SZÉKELYI Mária - BARNA Ildikó: *Túlélőkészlet az SPSS-hez*. Budapest, 2002, Typotex.

SZŰCS István (szerk.): *Alkalmazott statisztika*. Budapest, 2004, Agroinform.

TEMPLETON, Gary F. (2011) *A Two-Step Approach for Transforming Continuous Variables to Normal: Implications and Recommendations for IS Research*. Communications of the Association for Information Systems: Vol. 28 , Article 4.

THODE, C. Henry: *Testing for normality,* New York, 2002, CRC Press.

TOMCSÁNYI Pál: *Általános kutatásmódszertan*. Budapest, 2000, Szent István Egyetem, Országos Mezőgazdasági Minősítő Intézet.

TÓTHNÉ LŐKÖS Klára: *Következtetés statisztika*. Gödöllő, 2008, GIK Kiadó.

A KÖNYV a Sapientia Erdélyi Magyar Tudományegyetem gazdasági szakos hallgatói számára készült jegyzet; a Kutatásmódszertan tárgy keretén belül kerül oktatásra. De az egyetemi hallgatókon, oktatókon túl az olvasói célcsoportba tartoznak azok a vállalati menedzserek, közgazdászok is, akik több, üzleti értéket jelentő információt szeretnének kinyerni a gyorsan növekvő vállalati adathalmazból vagy a piaci adatokból. Az információtechnológia fejlődésének, és ezzel párhuzamosan az adatelemzési módszertanok fejlődésének köszönhetően ma, a Big Data korszakában nemcsak lehetőség, hanem kihagyhatatlan szükségszerűség minden döntéshozó számára, hogy a belső-külső adatokat megragadja, elemezze és a döntési folyamataiba integrálja.

A közgazdasági kutatási folyamat jelenti a könyv vezérfonalát, az olvasó – ha elsajátítja a leírtakat – képes lesz végigvinni egy kutatást a probléma felismerésétől a tanulmányírás befejezéséig. Ennek megfelelően a könyv tartalmának nagy része az SPSS statisztikai programcsomag gyakorlati alkalmazását mutatja be.

SAPIENTIA
ERDÉLYI MAGYAR
TUDOMÁNYEGYETEM
Csíkszeredai Kar

ISBN 978-973-53-2906-8