Debreceni Egyetem
Informatika Kar

STATISZTIKAI PROBLÉMÁK MEGOLDÁSA SZÁMÍTÓGÉP SEGÍTSÉGÉVEL

Témavezető:
Dr. Baran Sándor
egyetemi tanár

Készítette:
Máté Zsolt
gazdaságinformatikus Bsc

Debrecen
2010

Szeretnék köszönetet mondani Dr. Baran Sándornak, aki támogatott az elképzeléseimben és hozzásegített szakdolgozatom megfelelő színvonalú elkészítéséhez.

Tartalomjegyzék

Bevezetés

A statisztikai tevékenység népszámlálás, vagyonfelmérés formájában már időszámításunk előtt is megjelent, önálló tudományként való alkalmazása pedig a XVIII. században indult el. Eredetileg a statisztika az államról (az elnevezés is a latin status szóból ered), annak felépítéséről, berendezkedéséről, állapotáról átfogó képet adó ismeretek összeségét jelentette, fokozatosan azonban kiterjedt az emberi tevékenység valamennyi területére, és egyben tudományos módszertanná nőtte ki magát.

Magát a statisztika szót többféle értelemben használjuk. Statisztikának nevezzük a tömegjelenségek adatait, az ún. statisztikai számanyagot (statisztika a foglalkoztatottságról és a munkanélküliségről, a népesedési folyamatokról stb.). De azt a tevékenységet is statisztikának hívjuk, amely az adatok gyűjtését, rendezését, tömörítését, elemzését foglalja magába. A másik értelmezés pedig a módszertan, ami a statisztikai gyakorlati tevékenység, illetve a statisztikai következtetések elméletével, módszereivel foglalkozik.

A statisztikai módszertannak többféle ágát lehet megkülönböztetni. Mi most a két ágat említünk meg, a leíró vagy deskriptív statisztikát és a következtető vagy más szóval induktív statisztikát. A **leíró statisztika** a vizsgálat tárgyát képező jelenség tömör, számszerű jellemzését adja az adatok rendezése és elemzése alapján a sokaság egészére vonatkozóan. Nem lép túl a megfigyelés körén, de a megfigyelt adatok legjobb megértésére, bemutatására, összefoglaló jellemzésére törekszik. A megfigyelt adatok sokoldalú jellemzéséhez gazdag elemzési eszköztárt (például grafikonok, táblázatok, középértékek) kínál a leíró statisztika. A **következtető statisztika** a sokaság egy kiválasztott részéből, a mintából következtet a sokaság egészére, azaz általánosítást jelent. Minőségellenőrzés során például meghatározott számú, és meghatározott módon kiválasztott termék vizsgálata alapján következtetni lehet arra, hogy a termékek összessége megfelel-e az előírt követelményeknek.

A szakdolgozatomat is ezek alapján két részre osztottam: leíró statisztikák és hipotézisvizsgálat. A hipotézisvizsálat témakörén belül megtalálhatók mind a paraméteres próbák, illetve a hozzájuk tartozó nem-paraméteres megfelelőjük is. Az itt közölt próbák, a teljesség igénye nélkül, inkább a hipotézisvizsgálathoz kapcsolodó fogalmak begyakorlását teszi lehetővé, semmint vállalati környezetben való alkalmazását.

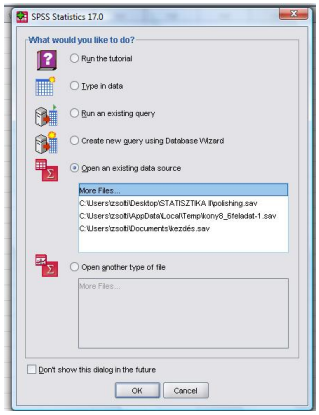
A szakdolgozatomban található feladatok megoldásához két különböző, a felsőoktatásban gyakran használt számítógépes programcsomagot használtam: Az egyik programcsomag **SPSS** néven vált ismertté. Maga a név a Statistical Package for the Social Sciences rövidítése. Az első verziója már 1968-ban megjelent. Manapság a legelterjedtebb statisztikai elemző szoftver lett a piacon, amellyel nagyméretű, összetett adatbázist lehet feldolgozni gyorsan és hatékonyan. Már sok hazai felsőoktatási intézményben a különböző statisztikai, vagy statisztikát használó tárgyak oktatásának alapja. A programot, a piacon elért sikerei láttán 2009 októberében meg is vette az IBM és összehangolta a meglévő adatbázis és elemző eszközeivel. A másik programcsomag a **MATLAB** nevet viseli. Maga a név a MATrix LABoratory rövidítése. A MATLAB története a 70-es években kezdődött, amikor Cleve Moler és munkatársai kifejlesztettek egy FORTRAN szubrutin könyvtárat (LINPACK, EISPACK).A 70-es évek végén Cleve Moler a New Mexikói Egyetem Számítástudományi Tanszékének volt a tanszékvezetője és lineáris algebrát oktatott. Azért, hogy a LINPACK és az EISPACK csomagok használatához hallgatóinak ne kelljen a FORTRAN programozási nyelvet megtanulniuk, szabadidejében hobbiból elkezdett egy olyan programot írni, amely egy interaktív hozzáférést biztosít a LINPACK és az EISPACK-hez. Később, 1983 elején John Little mérnök felismerte a MATLAB mérnöki alkalmazásának lehetőségét. Little, Moler és Steve Bangert egy fejlesztő teamet alakított, hogy megalkossa a MATLAB C-ben írt professzionális verzióját. Mára már világszerte elterjedt rendszerré vált. Számos egyetemen tanítják és alkalmazzák egyes tárgyak segédeszközeként, ugyanakkor megtalálható ipari környezetben is, ahol mérnöki és matematikai feladatok megoldására használják.

1. fejezet

A programok leírása

1.1. Az SPSS használata

A program indításakor megjelenik egy párbeszédablak:



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

Ha egy üres adatbázist szeretnénk, egyszerűen nyomjuk meg a Cancel gombot. Most a **Data Editor**-ban (Adatszerkesztő) járunk. Itt az adatbázisunk sorokból (rekordokból) és oszlopokból (változókból) áll. Maga az adatszerkesztő két lapból áll: **Data View** (adatbeviteli nézet) és **Variable View** (változó definiálási nézet), amelyeket a bal alsó sarokban található fülekre kattintva, illetve a CRTL+T billentyűkombinációval lehet váltogatni.

A **Variable View**-ban lehet beállítani a változók paramétereit:

- **Name**: a vátozó rövid neve. Ha nem adjuk meg, akkor automatikusan VAR000n n=1-től folyamatosan növekszik.

- **Type**: a változó típusa; számunkra fontos formátumok:

  - Numeric: numerikus típus; a számokat a legegyszerűbb formátumában jeleníti meg (pl. 12345,67)
  - Comma: a tizedesvesszőt ponttal, az ezres helyiértéket pedig vesszővel jelöli (pl. 12,345.67)
  - Dot: tizedesvesszőt vesszővel, az ezres helyiértéket pedig ponttal jelöli (pl. 12.345,67)
  - String: szöveges adatok beviteléhez használjuk

Az SPSS minden adatállományt mátrix formátumnak tekint. A mátrix oszlopaiban a **változók** (variables) helyezkednek el, azaz az egy oszlopban lévő adatok egyneműek (azonos dimenziójúak) és általában független megfigyeléseket tartalmaznak. Az egy sorban lévők az **esetek** (cases) vagy megfigyelések (observations), amelyek általában független mérésből erednek és többnyire különneműek.
Pl. az *Employee data.sav* nevű fájlban egy bank dolgozóira vonatkozó adatok vannak. Minden sor egy dolgozó személyi adatait (nem, iskolai végzettség, kezdő fizetés stb.) tartalmazza. Ezek az esetek. Minden oszlopban valamilyen egynemű adatnak (pl. beosztás (jobcat)) a bank összes dolgozóira vonatkozó esetei állnak. Ezek a változók.

A bevitt adatok elemzése az **ANALYZE** menüpont alatt történik. Itt megtalálható az összes elemzési eszköz, ami számunkra fontos lehet:

- Descriptives Statistics (leíró statisztika)
  - Frequencies (gyakoriságok)
  - Descriptives (leíró statisztikák)
  - P-P Plots (egy változó empirikus eloszlásfüggvényét rajzolja ki egy megadott elméleti eloszlásfüggvénnyel)
  - Q-Q Plots (egy változó empirikus kvantiliseit ábrázolja egy megadott elméleti eloszlás kvantiliseivel)
- Compare Means (paraméteres próbák)
  - One-Sample T Test (egy mintás t-próba)
  - Independent-Samples T Test (független mintás t-próba)
  - Paired-Samples T Test (páros mintás t-próba)
  - One-Way ANOVA (egyszempontú szórásanalízis)
- Nonparametric Tests (nemparaméteres próbák)
  - Chi Square (χ^2 próbák)
  - Binomial (binomiális próba)
  - 1-Sample K-S (egymintás Kolmogorov-Szmirnov próba)
  - 2 Independent Samples... (homogenitás-vizsgálat)
- Graphs (grafikák)
  - Bar (oszlopdiagrammok)
  - 3D Bar (háromdimenziós oszlopgrafikonok)
  - Pie (körgrafikonok)
  - Boxplot ("doboz-ábra")
  - Scatter/Dot (pontdiagramok)
  - Histogram (egy változó eloszlását szemléltető hisztogram)

A **TRANSFORM/COMPUTE VARIABLE** menüpont egy nagyon sokszor használt menüpont, amelynek segítségével egy új változó, vagy egy már létező változó eseteinek értékei számolhatók ki, a többi változó esetei értékeinek különféle függvényeiként. Lehetőség van arra is, hogy valamilyen logikai feltételt is beállítsunk.

A következőkben néhány olyan feladatok megoldását mutatom be, amik később hasznosak lehetnek az SPSS gördülékenyebb használatához.

1.1.1. Példa Adatmanipulálás az Employee data.sav állományban.

- Számoljuk ki a jelenlegi fizetés (salary) és a kezdő fizetés (salbegin) változók különbségét egy új változóba! (Transform/Compute Variable... menüpontban a Numeric Expression mezőbe be kell írni salary-salbegin kifejezést, a Target Variable mezőbe egy általunk kitalált új változónevet kell beírni)
- Rendezzük át az állományt jelenlegi fizetés (salary) szerint növekvő sorrendbe! (Data/Sort Cases... menüpontban beállítjuk a Sort by mezőnek a salary változót, a Sort order lehetőségnél az Ascending-et jelöljük be)
- Állítsuk át a gender nevű változó szélességét 3-ra és írjuk át a benne lévő értékeket a magyar megfelelőjükre (f-nő, m-ffi)!
  (Variable View-ban a gender változó Width értéke legyen 3, majd Transform/Redode into Same Variables... menüpontban kiválasztjuk a gender változót, majd átvisszük a String Variables... mezőbe, utána megnyitjuk az Old and New Values... lehetőséget, ahol beállítjuk értelemszerűen az Old és a New Variable párokat)
- Jelenítsük meg egy új adatmátrixon a biztonságiakat (custodial)! (data/Select Cases... menüpontban kiválasztjuk az If condition is statisfied lehetőséget, ahol a Numeric Expression mezőbe a jobcat=2 kifejezést írjuk, majd visszalépve a Select Cases menübe, kiválasztjuk a Copy selected to a new dataset lehetőséget és megadunk neki egy tetszőleges fájlnevet)

1.1.2. Példa Adott eloszlású véletlenszámok generálása.
Ha 100 darab standard normális eloszlású véletlenszámot szeretnénk generálni, akkor azt következő módon lehet megtenni.

- definiáljunk egy eset nevű változót a Variable View ablakban
- váltsunk át Data View ablakba és a változó oszlopában lévő első sorba írjuk be az 1-et, majd a Page Down billentyűvel haladjunk lejebb, ameddig a rendszer engedi. Ott ismét írjuk be egy tetszőleges számot, majd ismét haladjunk lejebb Page Down billentyűvel. Ezt az eljárást ismételgetve jussunk el 100-ig, ahová végül ismét irjunk be egy tetszőleges számot
- a Transform/Compute Variable... menüpont választással megjelenő Numeric Expression mezőbe írjuk be a $CASENUM kifejezést, a Target Variable mezőbe pedig azt, hogy eset
- szintén a Transform/Compute Variable... menüpontban Numeric Expression mezőbe írjuk be a RV.NORMAL(0,1) kifejezést, a Target Variable mezőbe, hogy normal

1.1.3. Példa Empirikus eloszlásfüggvény kirajzoltatása.
Az előző példában generált standard normális veletlen számsorozatnak számoljuk ki az empirikus eloszlásfüggvényét és rajzoltassuk is ki! Listázzuk ki a megfelelő elméleti eloszlásfüggvényt is!

- Transform/Rank Cases... menüpontban a Variable(s) mezőbe húzzuk be a normal változót, ezáltal létrejön egy új, Rnormal nevű változónk, ami rangsorolja a normal változóban lévő eseteket
- Transform/Compute Variable... menüpontban a Numeric Expression mezőbe írjuk be a Rnormal/100 kifejezést
- Data/Sort Cases... menüpontban a Sort by mezőbe húzzuk be a normal változót és állítsuk be növekvő (ascending) sorrendbe
- Transform/Compute Variable... menüpontban a Numeric Expression mezőbe írjuk be a CDF.NORMAL(normal,0,1) kifejezést

- Graphs/Legacy Dialogs/Scatter/Dots... menüpontban válasszuk az Overlay Scatter lehetőséget, majd állítsuk be az empir-normal illetve a theor-normal értékpárokat

1.1.4. Példa Kockadobás-sorozat szimulálása.
Szimuláljunk egy 200 dobásból álló kockadobás-sorozatot! Készítsük el a keletkező változó oszlopdiagramját!

- 1-től 200-ig futó esetszám változó létrehozása
- Transform/Compute Variable... menüpontban a Numeric Expression mezőbe írjuk be az RND.(RV.UNIFORM(0,1)*6-0,5)+1 kifejezést és a Target Variable-nek adjuk meg a kocka nevet
- Graphs/Legacy Dialogs/Bar/Simple-ben a Category Axis legyen a kocka

1.2. A MATLAB használata

A MATLAB egy kifejezés típusú nyelv, azaz a beírt kifejezéseket a program értelmezi, majd kiértékeli. A MATLAB utasítások általában a következő alakúak:

```
>> változó = kifejezés, vagy
>> kifejezés
```

A kifejezések általában műveleti jelekből, függvényekből és változókból állnak. A kifejezés kiértékelésének eredménye egy mátrix, amely megjelenik a képernyőn illetve a későbbi felhasználás céljából egy változóhoz kapcsolódik. Ha a változó név és az egyenlőségjel hiányzik, automatikusan létrejön egy **ans** (válasz) nevű változó és az eredményt ez tartalmazza. A MATLAB az utasítás, függvény és változó nevek esetében a kis és nagy betűket megkülönbözteti.

A **who** parancs felsorolja a munkaterületen található változókat. Egy változó törlésére a munkaterületről a **clear változónév** parancs szolgál. A **clear** parancs önmagában valamennyi nem statikus változót törli.

Kijelentkezéskor vagy kilépéskor a MATLAB összes változója elveszik. A kilépés előtt a **save** parancsot kiadva azonban az összes változó a **matlab.mat** nevű diszk fájlba menthető. A MATLAB-ot újra indítva, a **load** parancs visszatölti a munkaterület korábbi állapotát.

A MATLAB alapvetően egyetlen egy objektum típust használ, a **mátrix**ot, amelyben komplex számok is lehetnek. Egy mátrix létrehozásakor az oszlop elemeit szóközzel (lehet vesszővel is), a sorokat pedig pontosvesszővel választjuk el egymástól. Ne feledkezzünk meg a [ ] zárójelekről!

```
>> a = 1 (skalár vagy 1x1-es mátrix)
>> x = [1 5 13 6] (sorvektor vagy 1x4-es mátrix)
>> y = [1; 2; 3; 4] (oszlopvektor vagy 4x1-es mátrix)
>> A = [1 2 3; 4 5 6; 7 8 9] (3x3-as mátrix)
```

Néhány gyakran használt függvény és művelet:

```
>> help (beépített súgó)
>> help ztest (a z-próba súgója)
>> clc (törli a command view-t)
>> sort(x) (sorba rendezés)
```

```
>> lenght(x) (hossz)
>> size(A) (dimenzió)
>> sqrt(a) (négyzetgyök)
>> abs(a) (abszolútérték)
>> ones(2,3) (csupa 1-es elemeket tartalmazó, 2x3-as mátrix)
>> zeros(2,3) (csupa 0-kból álló 2x3-as mátrix)
>> rand(n) (n*n-es véletlen számokból felépített mátrixot hoz létre)
>> rand(m,n) (m*n véletlen számokból felépített mátrixot hoz létre)
```

Az ezeken kívül használt függvényeket a feladatok megoldásánál részletezem.

A MATLAB a lemezen, fájlokban tárolt utasítás sorozatokat is végrehajtja. Ezek az úgy nevezett **M-fájlok**, melyeknek a fájl név végén kötelezően az ".m" fájl típus szerepel. Az M-fájloknak két típusa van: a **parancs** (script) és a **függvény** (function) fájlok.

A parancs fájl a szokásos MATLAB utasítások sorozatát tartalmazza. Ha mondjuk a fájl neve **sajat.m**, akkor a **sajat** parancs hatására végrehajtásra kerülnek a fájlban lévő utasítások. A parancs fájlban található változók globálisak, azaz a környezetben lévő értékek megváltoznak.

A függvény fájlok biztosítják a MATLAB bővíthetőségét. Az általunk létrehozott speciális függvények a továbbiakban ugyanúgy használhatók mint a többi MATLAB függvény. A függvény fájlok változói lokálisak.
Például egy függvény fájl "belseje":

```
function [mean, stdev] = stat(x)
% STAT: Átlag és szórás számítása
% Egy x vektor esetén a stat(x) az x átlagát és szórását adja.
% Egy x mátrix esetén a stat(x) két sorvektort ad amelyek
% az egyes oszlopok átlagát ill. szórását tartalmazzák.
[m, n] = size(x);
if m == 1
m = n; % egy sorvektor kezelése
end
mean = sum(x)/m;
stdev = sqrt(sum(x.^2)/m - mean.^2);
```

A fentieket **stat.m** fájlba írva, az **[xmean, xdev] = stat(x)** parancs például az x vektor elemeinek átlagát és szórását átadja az **xmean** illetve **xdev** változóknak. A % jel azt jelzi, hogy a sor további része megjegyzés; a MATLAB a sor hátralévő részét nem veszi figyelembe. A M-fájlt dokumentáló első néhány megjegyzés sor azonban elérhető az on-line help segítségével, és megjelenik a képernyőn, ha például beírjuk a **help stat** parancsot. A **what** parancsot kiadva megjlenik a lemezen lévő M fájlok listája elérési útvonallal.

2. fejezet

Leíró statisztikák

Gyakori igény az, hogy egy adathalmazt elemei egyenkénti felsorolása helyett néhány jellemző tulajdonságának megadásával jellemezzünk. Ezeket az adatokból viszonylag könnyen kiszámítható paramétereket leíró statisztikáknak (vagy ritkán, de pontosabban: leíró statisztikai függvényeknek) nevezzük. (Matematikailag statisztikai függvénynek vagy röviden statisztikának neveznek minden olyan (rendszerint skaláris, olykor vektorértékű) függvényt, amelynek értelmezési tartománya a mintatér). Magyarul statisztika az, ami az adatainkból egy képlettel kiszámítható, vagy más módon meghatározható. Az említett leíró statisztikákon kívül igen fontosak még a hipotézis- vizsgálatoknál használt statisztikák (pl. z, t statisztika).

Sok ilyen paraméter van, két legfontosabb csoportjuk az ún. **elhelyezkedési** (measures of location or central tendency) és a **szóródást jellemző paraméterek** (measures of spread). Az elhelyezkedési paraméterek azt az értéket igyekeznek megadni, ami körül a mintánk elemei csoportosulnak (ilyen pl. átlag, medián) míg a szóródási paraméterek azt igyekeznek jellemezni, hogy értékeink mennyire szorosan vagy lazán helyezkednek el ekörül a pont körül (pl. szórás). Előfordul, hogy a minta elemeiről nem csak egyfajta adattal rendelkezünk. Kétféle adat esetén így összetartozó értékpárok jönnek létre (pl. emberek mintájában a testsúly és testmagasság). Az értékpárok közötti összefüggésről adnak információt a **kapcsolatot jellemző paraméterek** (measures of correlation).

A leíró statisztikák közül azok a legfontosabbak, amelyek a mintánkat adó populáció elméleti eloszlásfüggvényének valamelyik paraméterére adnak jó becslést a mintánkból. A leíró statisztikák gyakorlati alkalmazhatóságának ez az elméleti alapja. Itt csak annyit jegyzünk meg, hogy pl. a mintánkból meghatározott számtani átlag a populáció eloszlásfüggvényének várható értékére ad torzítatlan becslést. A mintából számított (ún. tapasztalati) szórás pedig a populáció eloszlásfüggvényét jellemző (ún. elméleti) szórás paraméter becslését adja.

A képet tovább bonyolítja, hogy a statisztikák a minta választásának esetlegessége miatt maguk is valószínűségi változók, melyeknek meghatározható az eloszlásfüggvénye, sőt ennek paraméterei becsülhetők, éspedig ismét valamilyen statisztikával. Például: nagyon gyakori, hogy összekeverik a mintából számított tapasztalati szórást (standard deviation, SD) az ugyancsak a mintából számítható 'átlag szórása' (standard error of the mean, SE) nevű paraméterrel. Sokan úgy gondolják, hogy a kettő lényegében ugyanaz, csak éppen az SE kisebb, mint az SD, ezért jobban fest a grafikonokon. Valójában az SE a mintaátlag (mint statisztika) elméleti eloszlásfüggvénye ismeretlen szórásparaméterének a becslése. Azt is mondhatjuk, hogy az SD egyszerű statisztika, az SE pedig egy statisztika statisztikája, tehát egy fokkal bonyolultabb fogalom.

Az adatok centrális helyzetét leíró statisztikák:

- **módusz** (mode): a változó esetei közül a leggyakrabban előforduló érték. Ha több ilyen is van az adatban, azok közül a legkisebb. Ordinális és intervallumskálás típusú adatoknál nem mindig van értelme.
- **medián** (median): páratlan mintaszám esetén a rendezett minta középső eleme: $x^*_{\frac{n+1}{2}}$, páros elem- számú minta esetén pedig a két középső elem átlaga: $\frac{x^*_{\frac{n}{2}}+x^*_{\frac{n}{2}+1}}{2}$.
- **átlag** (mean): az átlagérték. Ha $x_1, x_2 \cdots x_n$ jelöli az eseteket, akkor a számtani átlag a $\frac{\sum_{i=1}^{n} x_i}{n}$ érték. Néhány esetben a medián alkalmasabb a centrum kijelölésére, mert ha adathiba lépett fel, akkor az átlag nagyon elmozdulhat, míg a medián kevésbé érzékeny az adatvesztsére és a szélekre (pl. kiúgró értékekre (outliers)).

A szóródást (centrum körüli ingadozást) jellemző paraméterek:

- **terjedelem** (range): a legnagyobb és legkisebb adat különbsége, azaz $x_n^* - x_1^*$.

- **variancia** (variance): az adatoknak az átlagtól való négyzetes eltéréseinek átlaga ("kvadratikus középérték"). Torzítatlan becslése n elem esetén a négyzetes eltérések összege (n-1)-el elosztva: $s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, amit korrigált empirikus szórásnégyzetnek neveznek . (Torzítatlan egy becslés, ha a becslés elméleti középértéke minden mintaelemszám esetén éppen a keresett paraméter).

- **szórás** (standard deviation): a variancia négyzetgyöke: s_n^*. Fontos tudnunk , hogy értéke függ adataink mértékegységétől, így két minta szórása csak akkor hasonlítható össze, ha ugyanazt a mértékegységet használtuk.

- **relatív szórás** (coefficient of variation): e mérőszám a szóródás relatív nagyságát méri. A minta szórását a minta átlagához méri: $V = \frac{s_n^*}{\overline{x}}$. Dimenzió nélküli szám, kiszűri az értékek nagyságrendjét, ezáltal eltünteti az átlagok esetleges nagy eltéréséből fakadó hatást is. Egyben azt is megmutatja, hogy az egyes értékek az átlagtól relatíve átlagosan mennyivel (hány százalékkal) térnek el: $V = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{x_i - \overline{x}}{\overline{x}})^2}$.

- **standard hiba** (standard error v. standard error of mean, mintaátlag becsült szórása): a minta átlagának a várható értéktől való eltérését jellemző adat (a mintából nyert átlag mennyire pontosan becsüli a "valódi átlagot"). Jellemzően megegyezik a korrigált empirikus szórásnégyzet/mintanagyság négyzetgyökével: $\frac{s_n^*}{\sqrt{n}}$, ha nem ismert annak a populációnak a szórásnégyzete, amiből a minta származik.

- **kvantilis** (quantile): az $x_1, x_2, \cdots x_n$ minta p-kvantilise az a legnagyobb K szám, amelynél a minta legfeljebb p %-ka kisebb K-nál. A 0.5-kvantilis a medián, a 0.25-kvantilis az **alsó kvartilis**, a 0.75-kvantilis a **felső kvartilis**. (**interkvartilis terjedelem**: a felső és alsó kvartilis különbsége)

További jellemzők a minta eloszlására:

- **ferdeség** (skewness): azt fogalmazza meg, hogy a minta eloszlása mennyire nem szimetrikus. Képlete: $\beta_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})^3}{s_n^{*3}} \right)$ (harmadik centrális momentum/szórás^3 módon definiált mennyiség becslése). A szimmetrikushoz képest jobbra "elnyúló" eloszlás $\beta_1 > 0$, a balra "elnyúló" esetén pedig $\beta_1 < 0$. Az aszimmetria felmérésének egy igen egyszerű módja az, ha összehasonlítjuk a minta átlagát és mediánját: ha az átlag nagyobb, mint a medián, pozitív ferdeségről beszélünk, ha kisebb, akkor a ferdeség negatív, ha a két statisztika értéke egyenlő, az eloszlás szimmetrikus.

- **lapultság** (kurtosis): azt fogalmazza meg, hogy a minta sűrűségfüggvényének "csúcsossága" vagy "laposssága" hogyan viszonyul a normális eloszláséhoz. Kiszámítási módja: $\beta_2 = \frac{n(n-1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{(x_i - \overline{x})^4}{s_n^{*4}} \right) - 3\frac{(n-1)^2}{(n-2)(n-3)}$ (negyedik centrális momentum / szórás^4 -3 módon definiált mennyiség becslése). A haranggörbénél "csúcsosabb" eloszlásokra $\beta_2 > 0$, a "laposabbakra" pedig $\beta_2 < 0$.

- **hisztogram** (histogram): a minta eloszlását szemléltető olyan grafikon, amikor a minta terjedelme egymástól egyenlő távolságra lévő részintervallumaira fel van osztva, és az intervallumokba esés relatív gyakoriságainak megfelelő magasságú oszlopok állítódnak.

- **dobozábra** (boxplot): segítségével egyszerűen szemléltethető egy minta értékeinek elhelyezkedése és szóródása. A vízszintes tengelyen a különböző mintákat tüntetjük fel. Erre merőlegesen egy dobozt kell rajzolni, aminek alsó, illetve felső határa az első, illetve harmadik kvartilisnek megfelelően helyezkedik el. A dobozba egy vízszintes vonalat kell húzni a második kvartilisnek (medián) megfelelően. Könnyen felderíthetők vele az outlier értékek.

- **szár-és-levél ábra** (steam-and-leaf plot): a hisztogramnál informatívabb, de annál kevésbé látványos alakzat. A gyakoriságok nagyságának megfelelő hosszúságú stem oszlopok számokból vannak kialakítva, és azokról leolvasható, hogy a minta mely elemi estek konkrétan egy-egy részintervallumba.

2.0.1. Példa Egy újságárus valamely folyóiratból a naponta eladott mennyiséget 200 napon keresztül feljegyezte, és ebből az alábbi gyakorisági eloszlást készítette:

Eladott mennyiség	**Napok száma**
0	21
1	36
2	49
3	40
4	29
5	20
6	5
Összesen	200

Ábrázolja a gyakorisági sort! Állapítsa meg a módusz és az átlag értékét, és értelmezze őket!

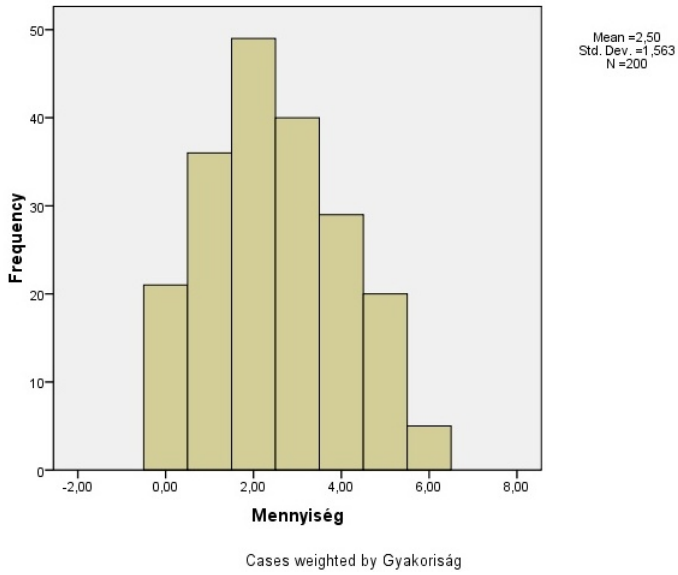
MATLAB:

```
>>x=[zeros(1,21), ones(1,36), 2*ones(1,49),
     3*ones(1,40), 4*ones(1,29), 5*ones(1,20), 6*ones(1,5)];
>>mean(x) (megadja az átlagot)
>>mode(x) (megadja a móduszt)
>>hist(x) (kirajzolja a hisztogrammot)
```

SPSS:

- Súlyozzuk a mennyiségeket a gyakoriságokkal: Data/Weight Cases.. menüpontban kiválasztjuk a Weight cases by lehetőséget, és Frequency Variable-nek megadjuk a gyakoriságokat tartalmazó változót

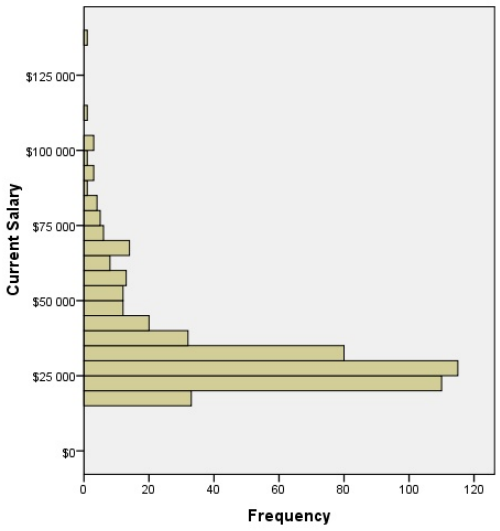
- Ábrázoljuk a gyakoriságokat: Graphs/Legacy Dialogs/ Histogram... menüpontban a Variable-nek megadjuk a mennyiségeket tartalamazó változót



A gyakorisági sor ábrázolása egy hisztogram megrajzolásával valósult meg. Az ábráról könnyen leolvasható a módusz értéke, mert a legmagasabb oszlophoz tartozó értéket kell keresni, ami jelen esetben 2 lap. Ez annyit jelent, hogy a leggyakoribb napi eladott mennyiség 2 lap az adott folyóiratból a vizsgált 200 nap alapján. A hisztogram ábrája mellet leolvasható az átlag értéke, ami jelen esetben 2,5 lap az adott folyóiratból. Tehát ha az összes eladott napi mennyiség helyébe a 2,5 lap értékét írnánk be, akkor az értékek összege nem változna.

2.0.2. Példa Jelenítsük meg Employee data.sav állományban lévő jelenlegi fizetéshez (salary) tartozó szár-és-levél ábrát, illetve hisztogramot (cseréljük fel majd a két tengelyt) !

Graph



Current Salary

Current Salary Stem-and-Leaf Plot

Frequency	Stem &	Leaf
33,00	1 .	5666667778899999
110,00	2 .	00000011111111112222222222222233333333444444444444444
115,00	2 .	55555555555666666666666777777777777778888888899999999999
80,00	3 .	00000000000000011111111122233333344444
32,00	3 .	555556667778899
20,00	4 .	00001233&
12,00	4 .	55678&
12,00	5 .	01244&
7,00	5 .	556
53,00	Extremes	(>=56750)

Stem width: 10000
Each leaf: 2 case(s)

& denotes fractional leaves.

Azért lett elforgatva a hisztogram, mert így közvetlenül összetvethető az ugyanezen az adatokra előállított szár-és-levél ábrával. A szár-és-levél ábra mint már említettem több információt hordoz a hisztogramhoz képest. A szár-és-levél ábrán látható, hogy a harmadik sorban (25000 és 30000 közötti tartományban) lesz a módusz, mert ez a leghosszabb sor, illetve a mellette lévő Frequency oszlop értéke itt a legmagasabb, de ezt a hisztogramból is le tudtuk volna olvasni. A szár-és-levél ábrából a leolvasható plusz információ az, hogy például 30000 és 35000 közé eső adatokon belül, 30000 és 31000 közé (11/27)*80, azaz kb. 33 adat esett.

3. fejezet

Hipotézisvizsgálat

3.1. Bevezetés, alapfogalmak

A statisztikában gyakran merülnek fel olyan problémák, ahol nem ismeretlen paraméterek becslése a feladat. Azt az eljárást, amelynek során a minta segítségével döntünk a hipotézisről (feltevésről), **statisztikai próbának** nevezzük. A hipotézis az alapsokaság valamilyen paraméterére vagy eloszlására vonatkozó feltevés. A feltevés helyességét a sokaságból vett minta alapján ellenőrizzük.
A vizsgálandó feltételezést **nullhipotézis**nek nevezzük, jele: H_0. Az ezzel ellentétes állítás az **alternatív hipotézis**, jele: H_1.

A nullhipotézisek különbözőek lehetnek. Vonatkozhatnak egy valószínűségi változó eloszlására, várható értékére, szórására, valószínűségi változók függetlenségére, korrelálatlanságára.

Legyen az X valószínűségi változó eloszlásfüggvénye $F_\vartheta(\mathrm{x})$, ahol ϑ az ismeretlen paraméter (skalár vagy vektor). Jelölje a θ a szóba jöhető paraméterek terét, tehát $\vartheta \in \theta$. Legyen θ_0 a Θ paraméter nemüres részhalmaza: $\theta_0 \subset \Theta$.

A **nullhipotézis** általános alakja: $H_0 : \vartheta \in \theta_0$.
Az **ellenhipotézis** általános alakja: $H_1 : \vartheta \in \theta - \theta_0$.

A nullhipotézis **egyszerű**, ha a θ_0 egy pontból álló halmaz, ellenkező esetben **összetett**. Hasonlóan az alternatív hipotézis is lehet egyszerű vagy összetett, a $\theta - \theta_0$ halmaz elemszámától függően.

$H_1 : \vartheta > \vartheta_0$ vagy $H_1 : \vartheta < \vartheta_0$ esetén **egyoldali ellenhipotézisről**, illetve **egyoldali próbáról** beszélünk. Ha $H_1 : \ \vartheta \neq \vartheta_0$, akkor **kétoldali ellenhipotézisről**, illetve **kétoldali próbáról** van szó.

Tekintsünk az X valószínűségi vátozóra vonatkozóan egy n elemű mintát: x_1, x_2, $..., x_n$. Az $\mathbb{R}^n$ teret tekinthetjük mintatérnek. A próba konstrukciója során a mintateret két diszjunkt halmazra bontjuk. Jelölje őket: C_0 és C_1. $C_0 \cap C_1 = \emptyset$. Ha a minta $(x_1, x_2, ..., x_n)$ realizációja a C_0 halmaz eleme, akkor elfogadjuk a nullhipotézist, ha $(x_1, x_2, ..., x_n) \in C_1$, akkor H_1 alternatív hipotézist fogadjuk el. A C_0 halmazt **elfogadási tartomány**nak, a C_1 halmazt **kritikus tartomány**nak nevezzük.

Ha H_0 igaz, és ennek ellenére elvetettük, akkor **elsőfajú hibát** követtünk el. Az elsőfajú hiba elkövetésének valószínűsége: $\mathsf{P}((x_1, ..., x_n) \in C_1 | H_0) = \alpha$. Ha a H_1 hipotézis az igaz és mégis elfogadjuk H_0-t, akkor **másodfajú hibáról** beszélünk. A másodfajú hiba elkövetésének valószínűsége: $P((x_1, ..., x_n) \in C_0 | H_1) = \beta$. Az elsőfajú hiba akkor következhet be, amikor ugyan jó a felvetésünk, de egy olyan szélsőséges mintát kapunk, ami adott valószínűséggel a felvetés ellen szól. A másodfajú hiba olyankor fordulhat elő, amikor a (rossz) feltevés elég közel esik az igazsághoz, így a mintából számolt próbafüggvény értéke egyaránt beleesik a tényleges és az általunk feltételezett sokasági paraméter köré szerkesztett elfogadási tartományba is.

A $P_\vartheta(C_1) \leq \alpha$ $(\vartheta \in \theta_0)$ relációt teljesítő α számot a **próba terjedelmének** (kritikus tartomány terjedelmének) nevezzük. A %-ban kifejezett értékére szokták a **szignifikancia szint** elnevezést használni. Az 1-α értéket a **próba megbízhatósági szintjének** nevezzük.

Az $\mathsf{P}((x_1, ..., x_n) \in C_1 | H_1) = 1 - \beta$ valósínűséget a C_1 kritikus tartományú **próba erejének** nevezzük. A próba ereje bizonyos értelemben a helyes döntés valószínűsége (H_1 igaz és a minta realizációja a C_1 tartományba esik). Akkor döntünk jól, ha az elsőfajú hiba elkövetésének kicsi a valószínűsége és ugyanakkor a próba ereje nagy.

	Elvetjük H_0-t	Nem vetjük el H_0-t
H_0 igaz	Elsőfajú hiba (α)	Helyes döntés (1-α)
H_1 igaz	Helyes döntés (1-β)	Másodfajú hiba (β)

p-érték: (vagy másnéven empirikus szignifikancia szint) annak eldöntésében segít, hogy mennyire nagy biztonsággal utasíthajuk el a nullhipotézist. A próbafüggvény mintából nyert értékéhez tartozó szignifikancia szint, ami mellett H_0 hipotézis már éppen elvethető. (elvetjük H_0-t, ha a p-érték $\leq \alpha$)

A próbák végrehajtásának általános feltételei:

1. A próbára vonatkozó alkalmazhatósági feltételek vizsgálata.
2. A szignifikanciaszint megválasztása.
3. A próbastatisztika értékének kiszámítása.
4. Kritikus tartomány kijelölése.
5. A nullhipotézisre vonatkozó döntés meghozatala.

3.2. Paraméteres próbák

Ha az eloszlás jellege ismert és a hipotézispár ezen eloszlás valamely paraméterére vonatkozik, akkor paraméteres próbáról beszélünk. Előnyeik, hogy az elméleti hátterük jól ismert és feltételeik teljesülése esetén a próák ereje viszonylag nagy. Hátrányuk a viszonylag szigorú feltételük, hogy a változók eloszlása az elméletileg megkövetelt legyen. Nominális és ordinális változókon használatuk nem ajánlott.

3.2.1. z-próba

Egymintás z-próba

- Alkalmazhatósági feltételek
  - a minta normális eloszlású
  - a populáció szórása ismert
- Hipotézisünk:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$
$$(H_1^j : \mu > \mu_0)$$
$$(H_1^b : \mu < \mu_0)$$

- A próbastatisztika:

$$z = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n} \sim N(0,1), \text{ ha } H_0 \text{ igaz.}$$

- Az elfogadási tartomány:

$$C_0 = (x_1, ..., x_n) : |z| < z_{1-\frac{\alpha}{2}}, \text{ ha kétoldali a próba}$$
$$C_0 = (x_1, ..., x_n) : z < z_{1-\alpha}, \text{ ha jobboldali a próba}$$
$$C_0 = (x_1, ..., x_n) : z > z_\alpha (= -z_{1-\alpha}), \text{ ha baloldali a próba}$$

3.2.1. Példa Egy teherautórakománnyi félliteres üdítőitalból 10 palackot véletlenszerűen kiválasztva és lemérve azok űrtartalmát az alábbi, milliliterben kifejezett értékeket kaptuk:

$$499,\ 525,\ 498,\ 503,\ 501,\ 497,\ 493,\ 496,\ 500,\ 495.$$

Ismert, hogy a palackokba töltött üdítőital mennyisége normális eloszlású 3 ml szórással. 95%-os döntési szintet használva vizsgálja meg a gyártó azon állítását, hogy a palackokba átlagosan fél liter üdítőitalt töltöttek!

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500$$
$$\alpha = 0.05$$

MATLAB:

```
>> h= ztest(x,m,sigma,alpha)
>> h= ztest(x,m,sigma)
>> [h,sig,ci,zval]=ztest(x,m,sigma,alpha,tail)
```

Az első parancs két oldali z-próbát hajt végre az x mintán annak eldöntésére, hogy a minta a sigma szórású és m várható értékű normális eloszlásból származik-e. A szignifikancia szint alpha. A h lehetséges értékei 0, ilyenkor nem utasítjuk el a nullhipotézist, illetve 1, ilyen esetben elutasítjuk a nullhipotézist. A második parancs ugyanazt hajtja végre. mint az első, de fix 5%-os szignifikancia szinttel. A harmadik parancs lehetőséget ad megadni a kétoldali ellenhipotézis tipusát: 0, ez a "default" kétoldali próba, az 1 érték jelöli a jobb oldali próbát, a -1 érték jelenti a bal oldali próbát. A ci érték az (1-α)*100% konfidencia intervallum az átlagra. A zval érték a próbafüggvény értéke.

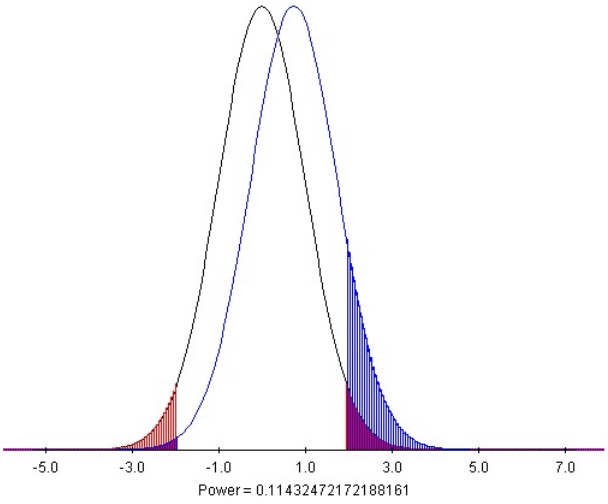
```
>> uditok = [499 525 498 503 501 497 493 496 500 495];
>> [h, sig, ci, zval] = ztest(uditok, 500, 3, 0.05)


Eredmény:
h = 0 (0-át kaptunk, ezért elfogadjuk a hipotézist)
sig = 0.4606 (p-érték, > 0,05)
ci = 498.8406  502.5594 (konfidencia-intervallum)
zval = 0.7379 (próbastatisztika értéke)
```

3.2.2. Példa Ellenőrizze le, hogy jól számolja-e ki a p-értéket a program! Számolja ki a próba erejét is!

- $\mathsf{P}(|Z| \geq 0.7379) = 2\mathsf{P}(Z \leq -0.7379) = 2\Phi(-0.7379)$

```
>> phi=normcdf(-0.7379);
>> p_ertek=2*phi
```

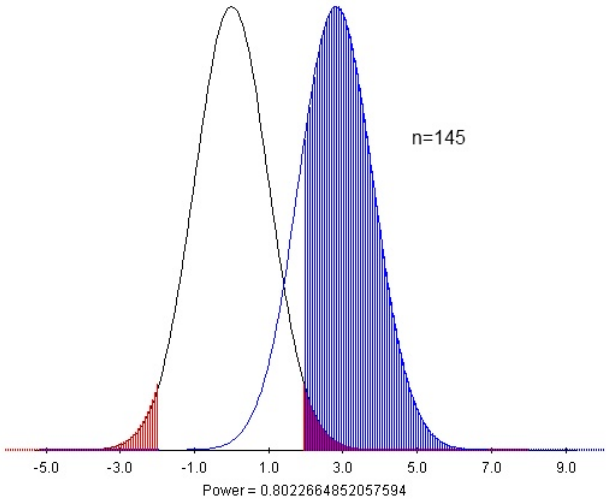
- $1 - \beta = 1 - \left(\Phi(z_{\frac{\alpha}{2}} - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma}) - \Phi(-z_{\frac{\alpha}{2}} - \frac{(\mu-\mu_0)\sqrt{n}}{\sigma})\right)$




```
>> error=3/sqrt(10);
>> also=500-norminv(0.975)*error;
>> felso=500+norminv(0.975)*error;
>> z_also=(also-mean(uditok))/error;
>> z_felso=(felso-mean(uditok))/error;
>> beta=normcdf(z_felso)-normcdf(z_also);
>> power=1-beta
```

Az ábrán a kék színű terület nagysága jelöli a próba erejét, ami jelen esetben 11.34%. Tehát 11.34% annak a valószínűsége, hogy elutasítjuk a rossz nullhipotézist, vagyis a másodfajú hiba elkövetésésnek az esélye 88.66%, feltéve

hogy H_1 igaz. Statisztikailag az a próba nevezhető erősnek, ami 80%-os vagy a feletti erővel rendelkezik. Jelen esetben ez nem teljesül, tehát ezt a próbát a nagyon gyenge jelzővel lehet illetni. Ahhoz, hogy a próbát erősebbé tegyük, a minta elemszámát kell növelnünk. Például n=50 esetén már 37.8%, n=145 esetén pedig 80.22% ennek a próbának az ereje!

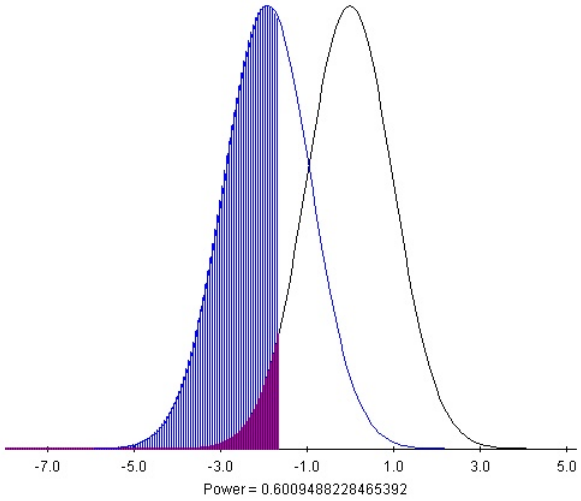


3.2.3. Példa Az *Ezt idd* teát 200 grammos dobozokban árulják, a csomagológép szórása 4 gramm. A Fogyasztóvédelmi Felügyelőség lemérte öt véletlenszerűen kiválasztott teásdoboz tömegét, melyekre az alábbi grammban kifejezett értékek adódtak:

$$196,\ 202,\ 198,\ 197,\ 190.$$

Hipotéziseit pontosan megfogalmazva és feltételezve, hogy a teásdobozok tömege normális eloszlást követ, döntsön 98%-os szinten, hogy az átlagos töltőtömeg tényleg 200 gramm, avagy kevesebb annál! Számolja ki a próba erejét is!

$$H_0 : \mu = 200$$
$$H_1 : \mu < 200$$
$$\alpha = 0.02$$

- ```
  >> teak=[196 202 198 197 190];
  >> [h,p,ci,zval]=ztest(teak,200,4,0.02,'left')
  ```
  ```

- ```
  >> error=4/sqrt(5);
  >> also=200-norminv(0.98)*error;
  >> z_also=(also-mean(teak))/error;
  >> beta=normcdf(z_also);
  >> power=1-beta
  ```



Egyoldali próbát hajtottunk végre, baloldali ellenhipotézissel, 2%-os szignifikancia szinten annak eldöntésére, hogy az átlagos töltő tömeg tényleg 200 gramm-e, avagy kevesebb. A próba végrehajtása után 98%-os megbízhatósági szinten azt tudjuk megállapítani, hogy el tudjuk fogadni a nullhipotézist, azaz hogy a teák átlagos töltőtömege 200 gramm. Látható a kép alapján, hogy viszonylag jól elkülönül a nullhipotézis az ellenhipotézistől, ezért a próba ereje (60,1 %) viszonylag magasnak mondható, feltéve hogy H_1 igaz.
  ```

3.2.2. t-próba

Egymintás t-próba

- Alkalmazhatósági feltételek
  - a minta normális eloszlású
  - a populáció szórása nem ismert
- Hipotézisünk:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$
$$(H_1^j : \mu > \mu_0)$$
$$(H_1^b : \mu < \mu_0)$$

- A próbastatisztika:

$$t = \frac{\bar{x}-\mu_0}{s_n^*}\sqrt{n} \sim t(n-1), \text{ ha } H_0 \text{ igaz.}$$

- Az elfogadási tartomány:

$C_0 = (x_1, ..., x_n) : |t| < t_{1-\frac{\alpha}{2}}(n-1)$, ha kétoldali a próba
$C_0 = (x_1, ..., x_n) : t < t_{1-\alpha}(n-1)$, ha jobboldali a próba
$C_0 = (x_1, ..., x_n) : t > t_\alpha(n-1)(= -t_{1-\alpha}(n-1))$, ha baloldali a próba

3.2.4. Példa Egy gabonaraktárban 60kg-os kiszerelésben búzát csomagolnak. A havi minőségellenőrzés során azt is megakarták vizsgálni, hogy a raktárból kikerülő zsákokban tényleg 60kg búza van-e, ezért lemértek tíz darab véletlenül kiválasztott zsákot. Eredményül a következőket kapták:

$$60.2,\ 63.4,\ 58.8,\ 63.6,\ 64.7,\ 62.5,\ 66.0,\ 59.1,\ 65.1,\ 62.0.$$

Hipotéziseit és az adatokra vonatkozó feltételeit pontosan megfogalmazva döntsön 95%-os szinten, a zsákok átlagos töltő tömege tényleg 60kg-e!

$$H_0 : \mu = 60$$
$$H_1 : \mu \neq 60$$
$$\alpha = 0.05$$

MATLAB:

```
>> h=ttest(x,m)
>> h=ttest(x,m,alpha)
>> [h,sig,ci,tstat]=ttest(x,m,alpha,tail)
```

Az első parancs kétoldali t-próbát hajt végre az x mintán annak eldöntésére, hogy a minta az m várható értékú normális eloszlásból származik-e. A második parancsnál a megszokott 5% szignifikancia szinttől eltérhetünk. A h változónak a lehetséges értékei 0, ilyenkor elfogadjuk a nullhipotézist, illetve az 1, ilyenkor elutasítjuk a nullhipotézist. A harmadik parancsnál lehetőségünk van egyoldali próbát végrehajtani: az 1 jelöli a jobboldali próbát, a -1 jelöli a baloldali próbát. A ci értékek az alsó illetve a felső korlát értékek az átlagra. A tstat 3 részből tevődik össze: az első a próbastatisztika értéke, a második a szabadsági fok, a harmadik pedig a korrigált empirikus szórásnégyzet.

```
>> zsakok=[60.2 63.4 58.8 63.6 64.7 62.5 66 59.1 65.1 62];
>> [h,sig,ci,tval]=ttest(zsakok,60)
```

```
Eredmény:
h=1 (elvetjük a nullhipotézist)
sig=0.0108 (p-érték,<0.05, ezért elvetjük a nullhipotézist)
ci=60.7454   64.3346 (a konfidencia-intervallum)
tval=tstat: 3.2017 (a próbafüggény értéke)
     df: 9 (szabadsági fok)
     sd: 2.5087 (a mintából számított szórás)
```

SPSS:

- Analyze/Compare Means/One Sample T-Test... menüpontban beállítjuk a Test Variable-nek a zsákokat tartalmazó változót
- a Test Value értéke legyen 60
- az Options-ben be lehet állítani a szignifikancia szintet, ami alapértelmezés szerint 5%

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
VAR00001	10	62,5400	2,50874	,79333

One-Sample Test

	Test Value = 60					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
VAR00001	3,202	9	,011	2,54000	,7454	4,3346

Az SPSS output első táblázatában található a próba végrehajtásához szükséges leíró statisztikák. Az első oszlopban található a minta elemszáma, a másodikban a mintában található elemek számtani átlaga ($\frac{\sum x_i}{n}$), a harmadikban a mintából számított korrigált empirikus szórás ($\sqrt{s_n^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$) és a negyedikben található az átlag szórása vagyis a standard hiba ($\frac{s_n^*}{\sqrt{n}}$).

A második táblázatban található a próba végrehajtása utáni állapot. Az első oszlopban látható a próbastatisztika értéke (3.202), a másodikban a szabadsági fok értéke (10-1=9), a harmadikban a p-érték (0.011), a negyedikben a minta átlagának

és a feltételezett eloszlás átlagának különbsége ($\mu_1 - \mu_0$=2,54) és az ötödik oszlopban található az 5%-os szignifikancia szinthez tartozó konfidenciaintervallum (62,65±2,62*(2,50874/$\sqrt{10}$)).

A próba p-értéke kisebb a megadott szignifikancia szintnél, ezért elvetjük a nullhipotézist, vagyis azt, hogy az átlagos töltőtömeg 60kg. Az SPSS mindig a p-érték alapján dönt, de mi tudunk dönteni p-érték ismerete nélkül is, mégpedig a próba értéke alapján. A t-próba értéke jelen esetben 3,202, az 5%-os szignifikancia szinthez tartozó 9 szabadságfokú t eloszlás értéke 2,262, ami most nekünk a felső korlátunk. A próba értéke nagyobb, mint a felső korlát, ezért is elvethetjük a nullhipotézist. A harmadik, ami alapján dönthetünk, az a konfidencia-intervallum (az Options-ben be lehet állítani más szignifikancia értéket az intervallumhoz). Ha a feltételezett eloszlás várható értéke bele esik az intervallumba, akkor elfogadjuk a nullhipotézist, különben nem. Jelen esetben az alsó korlátunk 62,54-0,754=61,786, ami nagyobb mint a mi feltételezett eloszlásunk várható értéke, ami 60. Tehát elvetjük a nullhipotézist.

Két független mintás t-próba

- Alkalmazhatósági feltételek
  - a 2 minta normális eloszlású
  - a populáció szórása nem ismert
  - a 2 minta független
  - a mintaelemszámok lehetnek különbözőek
- Hipotézisünk:

$$H_0 : \mu_x = \mu_y$$
$$H_1 : \mu_x \neq \mu_y$$
$$(H_1^j : \mu_x > \mu_y)$$
$$(H_1^b : \mu_x < \mu_y)$$

- A próbastatisztika:

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{(n_x-1)s_x^{*2}+(n_y-1)s_y^{*2}}{n_x+n_y-2}}\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \sim t(n_x+n_y-2), \text{ ha } H_0 \text{ igaz.}$$

- Az elfogadási tartomány:

$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : |t| < t_{1-\frac{\alpha}{2}}(n_x + n_y - 2)$, ha kétoldali a próba
$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : t < t_{1-\alpha}(n_x + n_y - 2)$, ha jobboldali a próba
$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : t > t_{\alpha}(n_x + n_y - 2)$, ha baloldali a próba

3.2.5. Példa Kétfajta instant kávé oldódási idejét tesztelték, melyekből minden alkalommal azonos mennyiséget tettek 1 dl forrásban lévő vízbe. A kísérletek eredményeit az alábbi táblázat tartalmazza:

Kávé	Oldódási idő (másodperc)							
Mokka Makka	8.2	5.0	6.8	6.7	5.8	7.3	6.4	7.8
Koffe In	5.1	4.3	3.4	3.7	6.1	4.7		

95%-os szinten vizsgáljuk meg azt az állítást, hogy a Mokka Makka kávé lassabban oldódik, mint a Koffe In!

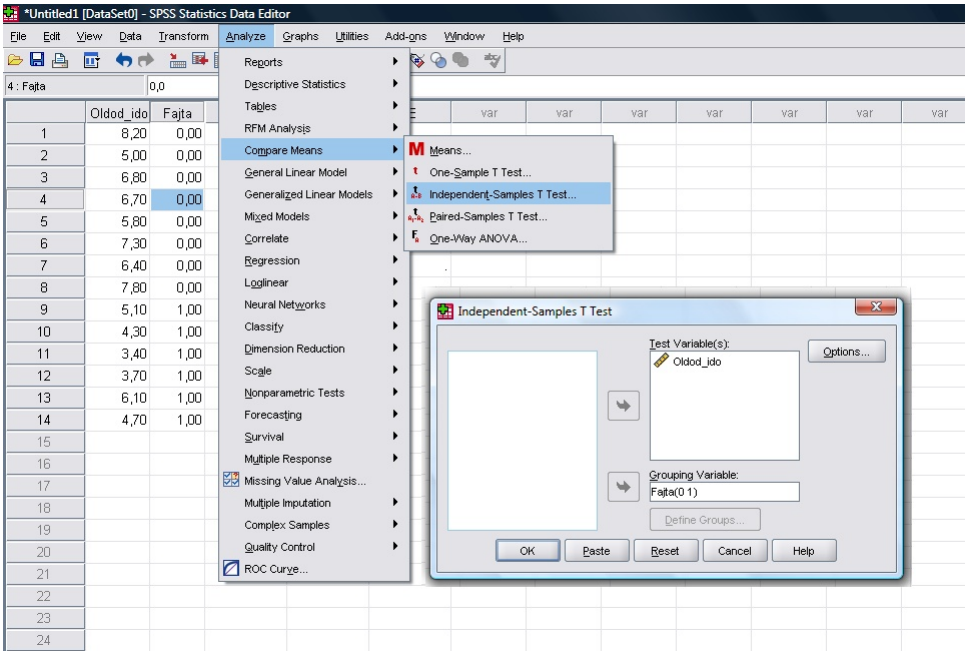
MATLAB:

```
>> h = ttest2(X,Y)
>> h = ttest2(X,Y,ALPHA)
>> [h,p,ci,stats] = TTEST2(X,Y,ALPHA,TAIL,VARTYPE)
```

Nagyon hasonlít az egymintás t-próbához, csak itt 2 változót kell megadni (x,y). A harmadik esetben a VARTYPE-nak egy stringet kell megadnunk, amivel megmondjuk, hogy a szórások egyenlőek-e vagy sem. Ha beírjuk az 'unequal' stringet, akkor azt állítjuk, hogy a szórások nem egyenlőek, nyilván ha elhagyjuk, akkor az alapértelmezett eset hajtódik végre, vagyis, hogy a szórások egyenlőek.

```
>> x=[8.2 5.0 6.8 6.7 5.8 7.3 6.4 7.8];
>> y=[5.1 4.3 3.4 3.7 6.1 4.7];
>> [h,p,ci,stats]=ttest2(x,y,0.05,'right')

Eredmény:
h=1 (elutasítjuk a nullhipotézist a megadott szignifikancia szinten)
p=8.7812e-004 (p-érték, <0.05)
ci =1.2202 Inf (konfidencia intervallum)
stats =tstat: 4.0017 (a próbastatisztika értéke)
       df: 12 (szabadsági fok)
       sd: 1.0180 (korrigált empirikus szórás)
```

SPSS:




```
T-TEST GROUPS=Fajta(0 1)
  /MISSING=ANALYSIS
  /VARIABLES=Oldod_ido
  /CRITERIA=CI(.95).
```

T-Test

```
[DataSet0]
```

Group Statistics

	Fajta	N	Mean	Std. Deviation	Std. Error Mean
Oldod_ido	,00	8	6,7500	1,04198	,36839
	1,00	6	4,5500	,98336	,40146

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
Oldod_ido	Equal variances assumed	,006	,939	4,002	12	,002	2,20000	,54976	1,00217	3,39783
	Equal variances not assumed			4,038	11,262	,002	2,20000	,54487	1,00415	3,39585

A fenti táblázat nem csak egyszerűen a t-próbát tartalmazza, hanem számos egyéb fontos dolgot is megtudhatunk benne. Mivel a t-próba csak akkor végezhető el ”tiszta lelkiismerettel”, ha a független minták szórása megegyezik, így adódik hogy ezt a Levene teszt F próbájával vizsgáljuk. A Levene teszt az egyetlen vizsgálat,

amelynél a szignifikancia szintet fordítva kell értelmezni, hiszen a H_0 a kedvező alternatíva, vagyis a magas érték a megfelelő számunkra. Ez egyszerűen a hipotézisek felállításából adódik, hiszen a Leven teszt F próbájánál a nullhipotézisben a szórásnégyzetek egyenlőségét fogalmazzuk meg, amit jelen esetben nem szándékozunk elvetni, hiszen számunkra ez jelenti azt, hogy a minták alkalmasak a t próbára.

A táblázatban az F értéke kicsi (0.006) a szignifikancia értéke pedig magas (0.939), tehát vizsgálhatjuk a t statisztikákat (vagyis az első sor tartalmazza a releváns értékeket, hiszen teljesül a varianciák egyenlőségének feltétele).

A t próba empirikus szingnifikancia szintje (0.002/2=0.001) az elfogadott 5% alá esik, így elvetjük a nullhipotézist, vagyis azt, hogy a Mokka Makka kávé lassabban oldódik, mint a Koffe In.

Két párosított (összetartozó) mintás t-próba

- Alkalmazhatósági feltételek
  - a 2 minta különbsége normális eloszlású
  - a populáció szórása nem ismert
  - a 2 minta párosított
- Hipotézisünk:

$$H_0 : \mu_x - \mu_y = \mu_d = 0$$
$$H_1 : \mu_x - \mu_y = \mu_d \neq 0$$
$$(H_1^j : \mu_d > 0)$$
$$(H_1^b : \mu_d < 0)$$

- A próbastatisztika:
  $(\overline{d} = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i))$

$$t = \frac{\overline{d}}{s_d^*}\sqrt{n} \sim t(n-1), \text{ ha } H_0 \text{ igaz.}$$

- Az elfogadási tartomány:

$$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : |t| < t_{1-\frac{\alpha}{2}}(n-1), \text{ ha kétoldali a próba}$$
$$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : t < t_{1-\alpha}(n-1), \text{ ha jobboldali a próba}$$
$$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : t > t_{\alpha}(n-1), \text{ ha baloldali a próba}$$

3.2.6. Példa Az Árelhajlásvizsgáló Hivatal összehasonlította két konkurens hipermarket élelmiszerárait. Tíz véletlenszerűen kiválasztott terméket vizsgáltak, melyek árait az alábbi táblázat tartalmazza:

Termék	A	B	C	D	E	F	G	H	I	J
Alfa Hipermarket	464	158	376	112	98	92	38	74	66	38
Beta Hipermarket	432	148	416	104	84	98	36	62	76	34

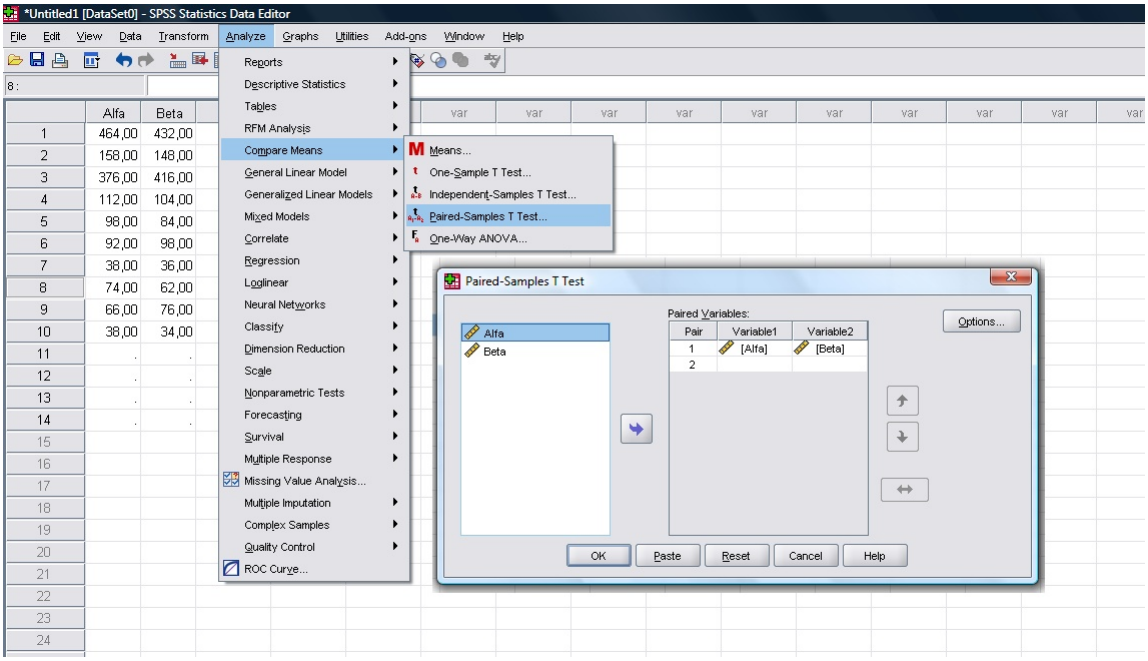
Az árkülönbségeket normális eloszlásúnak tételezve fel döntsön 95%-os szinten, van-e eltérés a két hipermarket élelmiszereinek árszintje között!

MATLAB:

```
>>Alfa=[463 158 376 112 98 92 38 74 66 38];
>>Beta=[432 148 416 104 84 98 36 62 76 34];
>> [h,p,ci,stats] = ttest(Alfa,Beta,0.05)
```

Mivel a páros mintás t-próbát úgy kell végrehajtani, mint az egymintás t-próbát, ezért csak egy második változót kell megadni a ”sima” t-próbának a MATLAB-ban.

SPSS:



```
T-TEST PAIRS=Alfa WITH Beta (PAIRED)
  /CRITERIA=CI(.9500)
  /MISSING=ANALYSIS.
```

T-Test

[DataSet0]

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Alfa	151,6000	10	147,24975	46,56446
	Beta	149,0000	10	148,74363	47,03686

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Alfa & Beta	10	,992	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Alfa - Beta	2,60000	18,90444	5,97811	-10,92342	16,12342	,435	9	,674

A páros mintás t-próba kiszámításához ki kellett számítani egy új változót, amely a páronkénti különbséget fejezi ki. Ennek az átlagát (2.6) és szórását (18.90444) látjuk a fenti táblázatban, és ezt teszteli a t-próba. Itt a szórás azért lett ilyen magas, mert különböző árszintű termékeket hasonlítottak össze.

Jelen esetben a t-próba p-értéke elég magas (0.674), így elfogajuk a nullhipotézist, vagyis azt, hogy a két szupermarket árai megegyeznek.

3.2.3. Szórásanalízis (ANOVA)

- Alkalmazhatósági feltételek:
  - A populáció amiből a csoportok származnak normális eloszlású.
  - A csoportok varianciái megegyeznek.
  - A csoportok egymástól függetlenek.
- Hipotézisünk:

$$H_0 : \mu_1 = \mu_2 \cdots = \mu_M$$
$$H_1 : \exists j, k : \mu_j \neq \mu_k (j \neq k)$$

- A próbastatisztika:

$$F = \frac{SS_K/(M-1)}{SS_B/(n-M)} = \frac{\sum_j n_j(\overline{x}_j - \overline{x})^2/(M-1)}{\sum_j \sum_i (x_{ij} - \overline{x}_j)^2/(n-M)} \sim F(M-1, n-M), \text{ ha } H_0 \text{ igaz.}$$

  Ahol:
  x_{ij}: j-edik csoport i-edik eleme
  $\overline{x}_j$: j-edik csoport átlaga
  n_j: j-edik csoport elemszáma
  $\overline{x}$: főtálag
  M: csoportszám
  n: minta elemszáma

- Az elfogadási tartomány:

$$C_0 = (x_1, ..., x_n) : F < F_{1-\alpha}(M-1, n-M)$$

A szórásanalízis olyan feladatokkal foglalkozik, amelyekben a vizsgált valószínűségi változó értéke egy, vagy több szisztematikus és (vagy) véletlen hatású, mennyiségi és (vagy) minőségi tényezőtől is függ a véletlen ingadozáson túl. Az analízis azt vizsgálja, hogy a tényezők valóban befolyásolják-e a valószínűségi változó értékét, vagy a tényezők különböző szintjei mellett mért értékek közötti eltérések csupán a véletlen ingadozásnak köszönhetőek.

Például nagyon sok esetben felmerülnek olyan kérdések, hogy hat-e a kezelés típusa a túlélési arányra egy bizonyos betegség esetén, vagy hogy hat-e a művelési mód a terméseredményekre. Az ilyen típusú kérdések esetén mindig felmerül az a gyanú, hogy a mért vagy megfigyelt különbséget nem az átalunk vizsgált effektus okozza. Lehet, hogy a beteg gyorsabb felépülése nem a kezelés típusától függ, hanem egyszerűen a jobb kondíciótól. Lehet, hogy a parcellán, amelyen a jobb eredményt érték el, a talaj minősége lényegesen jobb volt, mint a többin, így ez okozta a jobb terméseredményt.

Az ilyen típusú kérdések megválaszolására a varianciaanalízis módszere szolgál, amely tulajdonképpen a független mintás t-próba kiterjesztése több mintára (ha két mintánk van, akkor az egyszempontos ANOVA eredménye megegyezik a független mintás t-próba eredményével). Azt kell eldöntenünk, hogy a kettőnél több populáció átlagai azonosak-e vagy sem. Még ha átlagokat is hasonlítunk össze, a próbában varianciákat használunk, tehát az analízisnek nem célja, hanem eszköze a varianciák elemzése!

Felmerülhet az a kérdés, hogy miért nem alkalmazzuk a t-próbát páronként (két-két átlagot összehasonlítva egyszerre)? Azért, mert sok t-próbát kellene lefuttatni (minden lehetséges párra egyet). Például, ha 3 átlagot hasonlítunk össze, 3 t-próbára van szükség, 5 átlaghoz 10 t-próba, míg 10 átlaghoz 45 t-próba kell. Ekkor az igaz nullhipotézis elvetésének (elsőfajú hiba) esélye nő, hiszen az összes lehetséges páronkénti összehasonlítás nagy száma miatt véletlenül is kaphatunk szignifkáns eltéréseket.

Többféle varianciaanalízis létezik. Amennyiben a csoportok függetlenek, és csak egyetlen faktor (szempont) szerint különböznek (pl. többféle kezelést hasonlítunk

össze), akkor **egyszempontos** varianciaanalízissel (One-Way ANOVA) hasonlítjuk össze az átlagokat. Ha a csoportok függetlenek, de többféle faktor szerint is vizsgálhatóak (pl. nemek szerint), akkor **két- vagy többszempontos** varianciaanalízisről (Two-Ways ANOVA, Three-Ways ANOVA stb.) beszélünk. Az egyszempontos variancianalízis a független mintás t-próba általánosítása olyan esetekre, amikor több mint két minta átlagát szeretnénk összevetni. Ha két mintánk van, akkor az egyszempontos ANOVA eredménye megegyezik a független mintás t-próba eredményével.

Az analízis menete:
A populáció varianciájának kétféle becslését készítjük el. Az elsőt a **csoportok közötti varianciának** nevezik, és ez az átlagok szórásnégyzetét jelenti. A második a **csoportokon belüli variancia**, és ezt az összes adat alapján határozzuk meg. Ha nincs különbség az átlagok között,akkor a csoportok közötti és a csoportokon belüli varianciák egyenlőek és az F-próba értéke nagyjából 1. Amikor az átlagok lényegesen eltérőek, akkor a csoportok közötti variancia lényegesen nagyobb, mint a csoportokon belüli, és az F próbastatisztika értéke jóval nagyobb mint 1.

3.2.7. Példa Az Debreceni Egyetemen az egyik statisztika szemináriumvezető minden hétfőn, szerdán és pénteken autóval jár ki a Tócóskertből a város másik végén fekvő Kassai úti campusra. Otthonról mindig azonos időben indul el és ugyanazon az útvonalon autózik. Úgy érzi azonban, hogy a menetideje függ attól, hogy a hét melyik napján van órája. Ezért aztán márciusban, áprilisban és májusban véletlenszerűen kiválasztott 5-5 hétfőt, szerdát és pénteket és lejegyezte a menetidőket. Adatainak összegzését az alábbi táblázat tartalmazza:

Nap	Menetidő (x)					Összeg $\left(\sum x\right)$	Négyzet összeg $\left(\sum x^2\right)$
Hétfő	28	34	29	34	30	155	4837
Szerda	24	27	25	25	22	123	3039
Péntek	25	28	27	26	21	127	3255

Hipotéziseit pontosan megfogalmazva döntsön 99%-os szinten, igaz-e a szemináriumvezető sejtése!

H_0 : nincs különbség az átlagos menetidők között;

H_1 : van különbség.

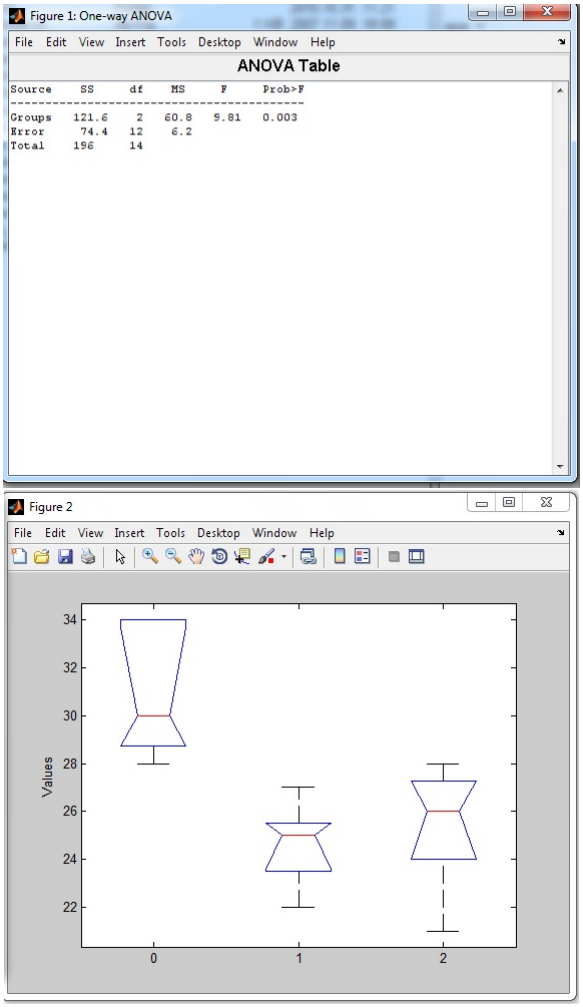
$\alpha = 0.01$.

MATLAB:

```
>> x=[28 34 29 34 30 24 27 25 25 22 25 28 27 26 21];
>> y=[0 0 0 0 0 1 1 1 1 1 2 2 2 2 2];
>> anova1(x,y)
```

EREDMÉNY:

Megkapjuk magát az ANOVA táblát és egy box-plot ábrát.



SPSS:

- felvisszük az egyik változóba az összes menetidőt, egy másik változóba pedig a menetidőkhöz tartozó csoport számát
- Analyze/One-Way-ANOVA... menüpontban a Depentdent List-nek megadjuk a menetidőket tartalmazó változót
- a factor a csoportosító változó legyen
- a Post Hoc-ban a szignifikancia szintet beállítjuk 0.01-re
- az Options-ben kipipáljuk a Descriptive-t és a Homogeneity-of-variance test-et

Descriptives

Menetido

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Hétfő	5	31,0000	2,82843	1,26491	27,4880	34,5120	28,00	34,00
Szerda	5	24,6000	1,81659	,81240	22,3444	26,8556	22,00	27,00
Péntek	5	25,4000	2,70185	1,20830	22,0452	28,7548	21,00	28,00
Total	15	27,0000	3,74166	,96609	24,9279	29,0721	21,00	34,00

Test of Homogeneity of Variances

Menetido

Levene Statistic	df1	df2	Sig.
,998	2	12	,397

ANOVA

Menetido

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	121,600	2	60,800	9,806	,003
Within Groups	74,400	12	6,200		
Total	196,000	14			

A boxplot ábrán látható 2 csoport átlaga(piros vízszintes vonalak) viszonylag egyenlőnek mondható, de a hétfői naphoz tartozó átlag sokkal magasabb, mint a szerdai és pénteki napok átlagai. A descriptives táblában leolvashatók a pontos átlagértékek. A második táblázat tartalmazza a Levene-teszt eredménye, aminek p-értéke alapján azt mondhatjuk, hogy a csoportok szórásai megegyeznek. A harmadik tábla tartalmazza magát a szórásfelbontó táblázatot, amiben megtalálható a csoportok közötti és csoportokon belüli eltérés-négyzetösszegek és a hozzájuk tartozó F-próba.

Mivel az F-próba p-értéke kisebb, mint az általunk megadott 1%-os szint, ezért a nullhipotézist elvetjük, és az alternatív hipotézist fogadjuk el, miszerint a napi menetidők különböznek.

3.2.8. Példa Vizsgáljuk meg, hogy az Employee data.sav állományban a kezdő fizetések (salbegin) egyenlőknek tekinthetők-e a három munkaköri kategóriában. Ugyanezt végezzük el a jelenlegi fizetésre (salary) is! (Nem tekinthetők ugyan normális eloszlásúaknak a három csoport kezdő és jelenlegi fizetései, de a szórásanalízis mégis elvégezhető ezekre az adatokra a normális feltételével szembeni viszonylagos robosztussága miatt.)

- Analyze/Compare Means/One-Way-ANOVA... menüpontban vigyük fel a salbegin és a salary változókat a Dependent List-be
- a factor (faktorváltozó) a jobcat legyen
- az Options-ben pipáljuk ki a Descriptive-t és a Homogeneity-of-variance test-et

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Current Salary	Clerical	363	$27,838.54	$7,567.995	$397.217	$27,057.40	$28,619.68	$15,750	$80,000
	Custodial	27	$30,938.89	$2,114.616	$406.958	$30,102.37	$31,775.40	$24,300	$35,250
	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
Beginning Salary	Clerical	363	$14,096.05	$2,907.474	$152.603	$13,795.95	$14,396.15	$9,000	$31,980
	Custodial	27	$15,077.78	$1,341.235	$258.121	$14,547.20	$15,608.35	$9,000	$15,750
	Manager	84	$30,257.86	$9,980.979	$1,089.014	$28,091.85	$32,423.86	$15,750	$79,980
	Total	474	$17,016.09	$7,870.638	$361.510	$16,305.72	$17,726.45	$9,000	$79,980

Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
Current Salary	59,733	2	471	,000
Beginning Salary	74,885	2	471	,000

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Current Salary	Between Groups	8,944E10	2	4,472E10	434,481	,000
	Within Groups	4,848E10	471	1,029E8		
	Total	1,379E11	473			
Beginning Salary	Between Groups	1,793E10	2	8,963E9	371,106	,000
	Within Groups	1,138E10	471	2,415E7		
	Total	2,930E10	473			

Az átlagok összehasonlításából láthajuk, hogy jelentős különbségek vannak mindkét változónál. A szórások is különbözőnek tűnnek, amit megerősít a második táblázat is, hiszen a szórások egyezésére vonatkozó Levene-teszt szignifikancia-szintje 0 mindkét változó esetében. Ezután nem meglepő, hogy az ANOVA-táblázatban azt olvashatjuk, hogy a csoportok közötti átlag négyzetösszeg jóval nagyobb, mint a csoportok közötti átlag négyzetösszeg. Ennélfogva a próbastatisztika nagy lett, a csoportok egyenlő várható értékére vonatkozó nulhipotézis empirikus szignifikancia-szintje mindkét változónál 0. Az alternatív hipotézist fogadjuk el, azaz a fizetések között lényeges különbségek vannak.

Post-Hoc tesztekről

Az ANOVA-táblázat csak azt mutatja meg, hogy van-e szignifikáns különbség, de azt nem, hogy pontosan melyik csoportok között. Ennek megállapítására több utóteszt is van, melyek a Post hoc fülre kattintva választhatóak ki. Nincs egyetlen általánosan elfogadott eljárás, amit mindenki használ, az egyes szempontokat mérlegelve kell kiválasztanunk a számunkra legmegfelelőbbnek tűnőt.

A post-hoc tesztek elsősorban aszerint vannak csoportosítva, hogy a szórásegyezés feltétele teljesül vagy sem. A próba kiválasztásánál két fontos szempontot kell figyelembe vennünk: mennyire könnyen lehet vele különbséget kimutatni (mennyire engedékeny), illetve mennyire megbízható. A két szempont között negatív összefüggés van, az engedékenyebb próbák kevésbé megbízhatók, és fordítva, a megbízhatók szigorúbbak. Az SPSS-ben a post-hoc tesztek e két szempont szerint vannak sorbarendezve, így például szórásegyezésnél a legelső felkínált próba, az LSD (Least Significant Difference), amellyel a legkorábban lehet különbséget kimutatni, ugyanakkor a megbízhatósága alacsony, továbbhaladva pedig nő a próbák megbízhatósága és szigorúsága.

A leggyakrabban használt post-hoc tesztek közé tartozik például a Tukey's b, illetve a Dunnett's T3, ha a szórások különböznek. (Mindegyik post-hoc tesztről rövid ismertetőt találunk az SPSS Help menüjében).

3.3. Nem-paraméteres próbák

Ha az alapsokaság eloszlása nem egy vagy több paraméterrel megadott, akkor nemparaméteres próbákat kell végeznünk. Ebben az esetben az előzetes feltevéseink nagyon általánosak, de természetesek. Például feltesszük, hogy a minta eloszlása folytonos, vagy, hogy a szórás véges, stb.. A próbák alkalmazása során nem szükséges a populáció paramétereinek (pl. átlag) becslése, illetve a paraméterekről szóló hipotézispár felállítása. Nem követelik meg, hogy a vizsgált változó valamely ismert elméleti eloszlást kövessen. Mivel kevesebb feltételt követelünk meg kiinduláskor, a következtetéseink levonásához nagyobb elemszámú mintákra lesz szükségünk, ami a mintavételezés költségeit növeli. A nemparaméteres próbákat szokták "eloszlásfüggetlen" próbáknak is nevezni.

Előnyeik, hogy kevesebb feltételük van, így hibás alkalmazásuk esélye kissebb. Nominális és ordinális változókon is használhatók. Próbastatisztikáik számítása sokszor egyszerűbb. Skálaérzéketlenek, azaz az adatok transzformálása nem befolyásolja a tesztek eredményét. Kevésbé érzékenyek a kiugró adatokra.

Hátrányaik, hogy erejük kisebb mint a paraméteres megfelelőiknek (azok feltételeinek teljesülése esetén), de ez sokszor nem jelentős (kb. 5%). Sok (főleg a komplikáltabb) parametrikus tesztnek nincs meg a nem-parametrikus megfelelője, főleg az elméleti háttér bonyolultabb volta miatt.

3.3.1. Binomiális próba

Ennél a próbánál a mintában lévő elemeket két csoportra osztjuk és teszteljük, hogy a két csoport megfigyelt relatív gyakoriságainak aránya megegyezik-e a megadott elméleti aránnyal.

3.3.1. Példa Az egyik élemiszerbolt-hálózat üzleteibe érkező import baracknak eddig átlagosan 15%-a sérült meg szállítás közben. Miután beszállítót váltottak, az új szállítmányból megvizsgáltak 50 barackot. Ezek között 3 sérültet találtak. 95%-os szinten döntsön abban a kérdésben, megérte-e lecserélni a régi beszállítót!

$$H_0 : p = 0.15;$$
$$H_1 : p < 0.15. \qquad \text{(egyoldali ellenhipotézis)}$$
$$\alpha = 0.05.$$

MATLAB:

```
>> binocdf(3,50,0.15)
A teszteléshez a binomiális eloszlást használjuk.
(Ezért binomiális próba a neve.)
```

SPSS:

- egy változóba fel kell vinni 47 db 0-t, és 3 db 1-est (1-essel kezdődjön a változó!)
- Analyze/Nonparametric Tests/Binomial menüpontot kell kiválasztani, ahol Test Variable-nek meg kell adni az 50 db megvizsgált barackot tartalmazó változót
- a Test Proportion-ba 0.15-öt kell beírni

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Barack	Group 1	1,00	3	,06	,15	,046[a,b]
	Group 2	,00	47	,94		
	Total		50	1,00		

a. Alternative hypothesis states that the proportion of cases in the first group < ,15.
b. Based on Z Approximation.

Látható, hogy a próba p-értéke 4.6 %, ami kisebb mint a megadott szignifikancia szint, ezért elvetjük a nullhipotézist. Tehát arra következtetésre jutottunk, hogy érdemes volt lecserélni a szállítót, mert a próba alapján a sérült barackok száma kevesebb lett 15 %-nál.
A táblázatban a Group1 jelöli a sérült barackokat, aminek mintában a megfigyelt relatív gyakorisága 6 %. Ez elegendően kevesebb 15 %-nál ahhoz, hogy 5 %-os szignifikancia szinten elvessük a nullhipotézist.

3.3.2. Előjelpróba

Az előjelpróba a páros mintás t-próba nem-paraméteres megfelelőjének tekinthető. A próba elvégzéséhez először képezzük a két minta különbségét, majd megszámoljuk a negatív és a pozitív különbségek számát (a nullákat kihagyjuk vagy ha a 0-k száma páros, akkor a felét az egyikbe, felét a másikba tesszük, vagy, ha a 0-k száma páratlan, akkor egy kimarad és a többit elosztjuk). Ha az eredeti két változó azonos eloszlású, akkor körülbelül azonos számú negatív és pozitív különbséget kapunk. Kis elemszámú minta ($n < 20$) esetében a binomiális eloszlás tulajdonságait használjuk fel, nagy elemszámú minta esetén ($n > 20$) az előjelek mintabeli eloszlásának megközelítésére a normális eloszlás felhasználható. Ezt a próbát egyszerűsége miatt általában gyors tájékozódás céljára használják.

Másik példa az előjelpróba használatára, amikor egy megfigyelés sorozat (minta) mediánját, nem pedig az átlagát kívánjuk egy ismert értékhez (ami lehet nulla, vagy egy jól megalapozott referencia érték) hasonlítani.

3.3.2. Példa Egy mozitulajdonos állítása szerint az egy-egy rajzfilmre hetente eladott gyermekjegyek mediánja 300. Állításának alátámasztására kiválasztott 8, a moziban vetített rajzfilmet, és feljegyezte, hogy egy-egy filmre egy adott héten mennyi gyermekjegyet váltottak. A következő eredményeket kapta:

$$412,\ 232,\ 197,\ 454,\ 251,\ 114,\ 256,\ 318.$$

Hipotéziseit pontosan megfogalmazva, az előjel próba segítségével döntsön 90%-os szinten, igaz-e a mozitulajdonos állítása!

$$H_0 : \mu = 300;$$
$$H_1 : \mu \neq 300.$$
$$\alpha = 0.1.$$

MATLAB:

```
>> p=signtest(x,m) (egymintás eset megadott mediánnal)
>> p=signtest(x,y) (párosított mintás eset)
>> [p,h,stat]=signtest(x,m,alpha,method)
```

A harmadik esetben a stat fogja tartalmazni a pozitív értékek számát (sign), illetve a z-próba értékét (zval), ha a method változóban ezt beállítjuk. A method esetén két lehetőségünk van: 'exact' (ilyenkor a binomiális eloszlás alapján számol) és 'approximation' (ilyenkor a normál eloszlás alapján számol).

```
>> jegyek=[412 232 197 454 251 114 256 318];
>> [p,h,stats]=signtest(jegyek,300,0.1)

Eredmény:
p=0.7266 (p-érték, >0.1)
h=0 (elfogadjuk a nullhipotézist)
stats=
    sign: 3 (pozitív előjelek száma)
```

SPSS:

- Analyze/Non Parametric Tests/Binomial... menüpontban megadjuk Test Variable-nek a gyermekjegyek értékeit tartalmazó változót
- a Define Dichtomy-nál a Cut Point-nak megadjuk a medián értékét, ami most 300
- a Test Proportion 0.5 legyen

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
VAR00003	Group 1	<= 300	5	,63	,50	,727
	Group 2	> 300	3	,38		
	Total		8	1,00		

A pozitív előjelek száma 3, a próba p-értéke 0.727, ami jóval nagyobb, mint a megadott szignifikancia szint, ezért a nullhipotézist elfogadjuk, vagyis azt, hogy a hetente eladott gyermekjegyek mediánja 300.

3.3.3. Példa Kétféle szójabab hozamát vizsgálva 12 parcellát megfeleztek, majd mindegyik parcella egyik felét az egyik, másik felét pedig a másik fajtával ültették be. A kilogrammban mért hozamokat az alábbi táblázat foglalja össze:

A fajta	142	124	133	151	121	127	135	141	149	150	151	132
B fajta	141	132	154	147	133	141	150	112	119	160	169	142

Hipotéziseit pontosan megfogalmazva döntsön 95%-os szinten, van-e különbség a két fajta hozama között!

$$H_0 : \mu_{A-B} = 0; \quad \text{(a különbségek mediánja nulla)}$$
$$H_1 : \mu_{A-B} \neq 0.$$
$$\alpha = 0.05.$$

MATLAB:

```
>> A=[142 124 133 151 121 127 135 141 149 150 151 132];
>> B=[141 132 154 147 133 141 150 112 119 160 169 142];
>> [p,h,stats]=signtest(A,B,0.05)

Eredmény:
p=0.3877 (p-érték, >0.05)
h=0 (elfogadjuk a nullhipotézist)
stats=
      sign: 4 (pozitív értékek száma)
```

SPSS:

- Analyze/Non Parametric Tests/2 Related Samples... menüpontban a Test Pairs-ben a Variable1-nek meg kell adni az A fajta értékeit, a Variable2-nek pedig a B fajta értékeit
- a Test Type-nak be kel pipálni a Sign-t

Sign Test

Frequencies

		N
VAR00001 - VAR00002	Negative Differences[a]	8
	Positive Differences[b]	4
	Ties[c]	0
	Total	12

a. VAR00001 < VAR00002
b. VAR00001 > VAR00002
c. VAR00001 = VAR00002

Test Statistics[b]

	VAR00001 - VAR00002
Exact Sig. (2-tailed)	,388[a]

a. Binomial distribution used.
b. Sign Test

A pozitív értékek száma 4, a p-érték 0.388, ami nagyobbb a szignifikancia szintnél, ezért elfogadjuk a nullhipotézist, miszerint nincs különbség a két fajta hozama között.

3.3.3. Wilcoxon-féle előjeles rangösszeg próba

A páros mintás t-próba nem-paraméteres másik alternatívája. A Wilcoxon-féle előjeles rangpróba nem csak az előjeleket, hanem a különbségek közötti nagyságrendeket is figyelembe veszi, így nagyobb erejű, mint az előjelpróba. A mintaelemek különbségeit (előjelüktől átmenetileg eltekintve) rangsorba állítjuk, és a különbségek helyébe azok rangsorát (rangszámát) írjuk (egyenlők esetén átlagosat, ezt kapcsolt rangnak, angolul "tie"-nak nevezik), majd a rangszámokat ellátjuk az eredeti különbségek előjelével. Ha a két minta azonos populációból származik, akkor az előjeles rangok összegének várható értéke 0.

Ugyanúgy lehet használni ezt a próbát egymintás esetben, mint ahogy azt a "sima" előjelpróbánál tettük.

3.3.4. Példa Döntsön a Wilcoxon-féle előjeles rangösszeg próba segítségével, igazat állít-e a 3.3.2-es Példában szereplő mozitulajdonos!

$$H_0 : \mu = 300;$$
$$H_1 : \mu \neq 300.$$
$$\alpha = 0.1.$$

```
>> [p,h,stats]=signrank(x,m,alpha) (egymintás eset)
>> [p,h,stats]=signrank(x,y,m,alpha,method) (kétmintás eset)

>> x=[412 232 197 454 251 114 256 318];
>> [p,h,stats]=signrank(x,300,0.025)
```

A pozitív rangok összege 14, a próba p-értéke 0.6406, ami nagyobb mint a szignifikancia szint, ezért elfogadjuk a nullhipotézist. Ha megnézzük, hogy az előjelpróba p-értéke mekkora, akkor összevetve ennek a próbának a p-értékével, látható, hogy ez a próba érzékenyebb, mivel kisebb a p-értéke ugyanarra a mintára végrehajtva.

3.3.5. Példa Egy kísérlet során azt vizsgálták, hogy a rendszeres sportolás milyen hatással van a gyerekek pulzusszámára. 16 gyereket vontak be a kísérletbe, akik közül 8 versenyszerűen sportol, a másik nyolc pedig nem rendszeresen sportoló egészséges gyermek. Ez utóbbiakat úgy választották ki, hogy minden sportoló gyereknek legyen egy nem sportoló párja, akinek nagyjából azonos a kora, testmagassága, tömege és testfelszíne. Az alábbi táblázat a mért pulzusszámokat tartalmazza:

Pár	1	2	3	4	5	6	7	8
Nem sportoló	90	85	75	120	95	105	100	95
Sportoló	95	75	75	85	80	80	85	75

A Wilcoxon-féle előjeles rangösszeg próba segítségével vizsgálja meg, igaz-e, hogy a sportoló gyerekek pulzusa lassabban ver, mint a nem sportoló társaiké! Döntsön 97.5%-os szinten!

MATLAB:

```
>> sportol=[90 85 75 120 95 105 100 95];
>> nem_sportol=[95 75 75 85 80 80 85 75];
>> [p,h,stats] = signrank(sportol,nem_sportol,'alpha',0.025,
   'method', 'approximate')
```

SPSS:

- Analyze/Nonparametric Tests/Two Related Samples... menüpontban Test Type-nak Wilcoxon-t kell bepipálni

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Sportol - NemSportol	Negative Ranks	6[a]	4,50	27,00
	Positive Ranks	1[b]	1,00	1,00
	Ties	1[c]		
	Total	8		

a. Sportol < NemSportol
b. Sportol > NemSportol
c. Sportol = NemSportol

Test Statistics[b]

	Sportol - NemSportol
Z	-2,201[a]
Asymp. Sig. (2-tailed)	,028

a. Based on positive ranks.
b. Wilcoxon Signed Ranks Test

A próba empirikus szignifikancia szintje 0.027, ami nagyobb mint a megadott szignifikancia szint (2.5 %), ezért elfogadjuk a nullhipotézist, miszerint a sportoló és nem sportoló gyermekek vérnyomása megegyezik. Ha az előírt szignifikancia szint a szokásos 5% lett volna, akkor szignifikáns különbséget mutatott volna ki a próba, és ezért elvetettük volna a nullhipotézist. (Ha az SPSS nem mutatta volna ki a próba p-értékét, akkor a felette lévő Z érték alapján ki tudtuk volna számolni azt a standard normális eloszlás táblázatának segítségével.)

3.3.4. Mann-Whitney-U próba

A független mintás t-próba nem-paraméteres alternatívája. A próba egy legalább ordinális változó mediánját hasonlítja össze két, egymástól független csoportnál. Intervallum változóknál is használhatjuk, például ha az eloszlás jelentősen eltér a normálistól. A próba végrehajtásának nincs előfeltétele, ezért lehet olyan magasabb mérési szintű változóknál is alkalmazni, ahol nem teljesül a szórásegyezés és/vagy a normális eloszlás előfeltétele. Ezt a próbát szokták Wilcoxon próbának is nevezni, mivel eredetileg Wilcoxon dolgozta ki, röviddel utána Mann és Whitney közölte ennek egy másik értelmezését.

A Mann-Whitney-U statisztika számítása két csoport elemeinek a párba állításán alapul. Az egyik csoport minden egyes elemét (x_i) párba állítjuk a másik csoport

minden egyes elemével (y_i), az így keletkezett párok száma $n_1 n_2$. Megvizsgáljuk, hogy a párok között hány olyan van, ahol az első szám kisebb, mint a másik ($x_i < y_i$). Ezeknek a pároknak a száma a Mann-Whitney-U-val jelölt statisztika (pontosabban, ha vannak a párok között egyenlők is, akkor az egyenlő párok számának a felét még hozzávesszük U-hoz). Ha a két populáció között nincs különbség, körülbelül egyforma számú olyan pár lesz, amelyekben $x_i < y_i$ mint amelyekben fordított a helyzet. Ha nagyon sok vagy nagyon kevés ilyen pár van, az arra utal, hogy a két populációban lévő számok nem egyformák egymáshoz viszonyítva. Az $\frac{U}{n_1 n_2}$ hányados annak a valószínűségnek a becslése, hogy egy, az első populációból véletlenszerűen választott új egyed értéke kisebb lesz, mint a másik populációból választott új egyedé.

Az U értéket az első csoportra számítjuk ki, és ha ez nagyobb, mint $\frac{n_1 n_2}{2}$, akkor U'=$n_1 n_2 - U$ értéket számoljuk ki. A W értéke megegyezik az első csoport rangszámösszegével, ha $U > \frac{n_1 n_2}{2}$, különben pedig a második csoport rangszámösszegével.

3.3.6. Példa A Csajágóröcsögei Vegyipari Kombinát gépkezelői közül néhányat továbbképzésre küldtek annak érdekében, hogy munkájuk során kevesebb hibát vétsenek. A tanfolyam eredményességét vizsgálandó 6, a tanfolyamot már elvégzett, és 13 még előtte álló gépkezelőnek ugyanazt a feladatot adták és feljegyezték a végrehajtás során vétett hibáik számát.

Tanfolyam után	11	9	4	7	6	2							
Tanfolyam előtt	3	17	12	13	21	29	5	1	15	19	16	14	10

Hipotéziseit pontosan megfogalmazva egy alkalmas nemparaméteres próba segítségével döntsön 95%-os szinten, volt-e haszna a tanfolyamnak!

$$H_0 : \mu_x = \mu_y; \quad \text{(hibák számának mediánjai megegyeznek)}$$
$$H_1 : \mu_x < \mu_y. \quad \text{(egyoldali ellenhipotézis)}$$
$$\alpha = 0.05.$$

MATLAB:

```
>> x=[11 9 4 7 6 2];
>> y=[3 17 12 13 21 29 5 1 15 19 16 14 10];
>> [p,h,stats]=ranksum(x,y,0.05)

Eredmény:
p=0.0462 (p-érték, kétoldali!!)
h=1 (elutasítjuk a nullhipotézist)
stats=
    ranksum: 37 (W értéke)
```

SPSS:

- Analyze/Nonparametric Tests/Two Independent Samples... menüpontban a Test Type-nak Mann-Whitney-t kell bepipálni

Mann-Whitney Test

Ranks

	Csoport	N	Mean Rank	Sum of Ranks
Hibak	,00	13	11,77	153,00
	1,00	6	6,17	37,00
	Total	19		

Test Statistics[b]

	Hibak
Mann-Whitney U	16,000
Wilcoxon W	37,000
Z	-2,017
Asymp. Sig. (2-tailed)	,044
Exact Sig. [2*(1-tailed Sig.)]	,046[a]

a. Not corrected for ties.

b. Grouping Variable: Csoport

Az első táblázat tartalmazza a rangokat és a hozzájuk tartozó átlagokat és összegeket a két csoportra lebontva. A második táblázat tartalmazza az U és W statisztikát illetve az empirikus szignifikancia szinteket.

A próba p-értéke $\frac{0.046}{2}$, ami kisebb mint az általunk megadott szignifikancia szint, ezért elvetjük a nullhipotézis, az ellenhipotézist fogadjuk el, miszerint a tanfolyam után vétett hibák mediánja kisebb, mint a tanfolyam elötti.

3.3.5. Khi-négyzet próbák

A Khi-négyzet teszt a minta elemeit kategóriákba rendezi, és utána számítja ki a statisztikát. A statisztika a megfigyelt gyakoriságok és a várható gyakoriságok közötti különbségek mértékét ítéli meg.

A próbastatisztika általánosan: $\sum \frac{(\text{megfigyelt gyakoriság-várt gyakoriság})^2}{\text{várt gyakoriság}}$.

Látható, hogy a várható gyakoriság szerepel a nevezőben, így ha ennek értéke túl kicsi, akkor a chi-négyzet értéke túl nagy lesz, ami hamis következtetések levonásához vezetne. De mi az a túl kicsi? Erre nézve a gyakorlatban elterjedt szabály az, hogy hogy az egy csoportba esés várható valószínűsége legalább 5 legyen. Ha ez nem teljesül, akkor szükségessé válhat a kis valószínűségű csoprotok összevonása.

Illeszkedésvizsgálat

Adott az x_1,x_2,...,x_n minta. Ellenőrizni akarjuk azt a feltevést, hogy a minta elméleti eloszlásfüggvénye éppen az F_0(x), az összes szóba jöhető eloszlásfüggvény között. Jelölje $p_1, ..., p_r$ az intervallumokba esés valószínűségeit az adott eloszlás fennállása esetén. Ha ezek a valószínűségek ismertek, **tiszta** illeszkedésvizsgálatról beszélünk. Ha nem ismerjük annak az eloszlásnak a paramétereit, amelyre a megfigyelt értékeket illeszteni szeretnénk, pusztán a típusát, akkor **becsléses** illeszkedésvizsgálatot végzünk. Ha H_0 igaz és n nagy, akkor a $\frac{k_i}{n}$ relatív gyakoriságok a p_i-k közelítései.
Ha a normális eloszláshoz való illeszkedés a kérdés, **normalitásvizsgálat**ról beszélünk. Általában azért akarjuk megvizsgálni, hogy az adatok eloszlása normális-e, mert ha igen, akkor alkalmazhatjuk rájuk a normális eloszlásra rendelkezésre álló statisztikai eljárásokat (z-próba, t-próba,...). A pozitív következtetés levonásánál nagyon óvatosan kell fogalmaznunk, mert ha nem túl sok adatunk van, akkor nagy a másodfajú hiba elkövetésének valószínűsége!

- Hipotézisünk:

$$H_0\text{: } \mathrm{P}(\mathrm{X} < \mathrm{x}) = F_0(\mathrm{x})$$
$$H_1\text{: } \mathrm{P}(\mathrm{X} < \mathrm{x}) \neq F_0(\mathrm{x})$$

- A próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(k_i - np_i)^2}{np_i} \sim \chi^2(r-b-1), \text{ ha } H_0 \text{ igaz.}$$

  Ahol b a p_i valószínűségek meghatározásához szükséges olyan paraméterek száma, amelyeket a mintából becsültünk.

- Az elfogadási tartomány:

$$C_0 = (x_1, ..., x_n) : \chi^2 < \chi^2_{1-\alpha}(r-b-1)$$

3.3.7. Példa Egyenletes eloszlásra történő illeszkedésvizsgálat.
Egy játékkockával 100 dobásból 12-szer 1-es, 20-szor 2-es, 14-szer 3-as, 15-szor 4-es, 18- szor 5-ös és 21-szer 6-os lett az eredmény. Ellenőrizzük 90%-os szignifikancia-szinten, hogy szabályos-e a dobókocka.

SPSS:

- súlyozzuk a gyakoriságokkal a dobásokat (Data/Weight Cases...)

- Analyze/Nonparametric Tests/Chi-Square... menüpontban a Test Variable-nek hozzáadjuk a dobásokat tartalmazó változót

- mivel egyenletes eloszlás, ezért All categories equal-t kell bepipálni (mindegyik dobás valószínűsége $\frac{1}{6}$)

Chi-Square Test

Frequencies

Dobások

	Observed N	Expected N	Residual
1,00	12	16,7	-4,7
2,00	20	16,7	3,3
3,00	14	16,7	-2,7
4,00	15	16,7	-1,7
5,00	18	16,7	1,3
6,00	21	16,7	4,3
Total	100		

Test Statistics

	Dobások
Chi-Square	3,800[a]
df	5
Asymp. Sig.	,579

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 16,7.

Mivel a p-érték jóval 10 % felett van, ezért nem vetjül el H_0-t, elfogadjuk, hogy a kocka szabályos (nincs elegendő bizonyítékunk arra, hogy nem szabályos).

3.3.8. Példa Egy újonnan kifejlesztett müzli ötféle magot (A, B, C, D és E) tartalmaz, melyek százalékos megoszlása a terméken lévő tájékoztató szerint 35%, 25%, 20%, 10%, illetve 10%. Egy véletlenül kiválasztott zacskóban az alábbi mennyiségi megoszlást találtuk:

Összetevő	A	B	C	D	E
Szem (darab)	184	145	100	68	63

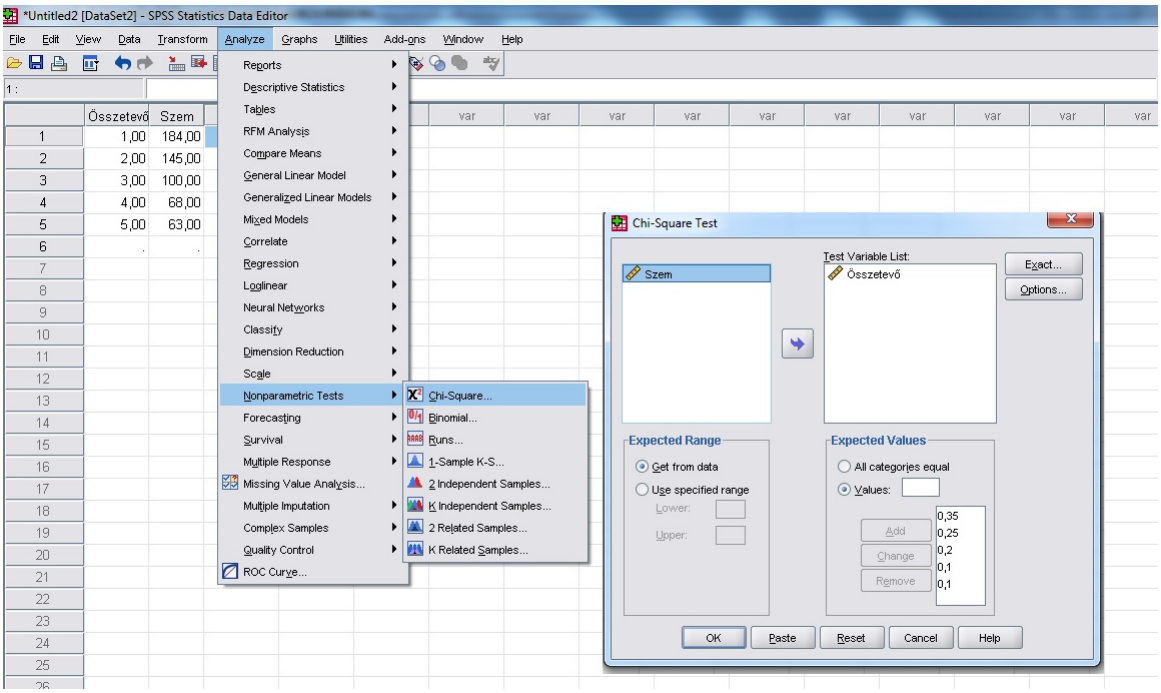
Döntsön 90%-os szinten, hogy a minta összetétele megfelel-e a csomagoláson feltüntetettnek!

H_0 : az összetétel megfelel a csomagoláson feltüntetettnek;

H_1 : az összetétel nem felel meg a csomagoláson feltüntetettnek.

$\alpha = 0.1$.

SPSS:



Chi-Square Test

Frequencies

Összetevő

	Observed N	Expected N	Residual
A	184	196,0	-12,0
B	145	140,0	5,0
C	100	112,0	-12,0
D	68	56,0	12,0
E	63	56,0	7,0
Total	560		

Test Statistics

	Összetevő
Chi-Square	5,645[a]
df	4
Asymp. Sig.	,227

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 56,0.

A próba empirikus szignifikancia szintje 0.227, ami jóval 5 % felett van, ezért elfogadjuk a nullhipotézist, vagyis azt, hogy a minta összetétele megfelel a csomagoláson feltüntetettnek.

Homogenitás-vizsgálat

A homogenitás-vizsgálat annak az eldöntésére szolgál, hogy két valószínűségi vátlozó azonos eloszlású-e, ugyanaz a függvény-e az eloszlásfüggvényük.

- Hipotézisünk:

$$H_0\text{: P(X} < \text{x)=P(Y} < \text{x)}$$
$$H_1\text{: P(X} < \text{x)}\neq\text{P(Y} < \text{x)}$$

- A próbastatisztika:

$$\chi^2 = n_y n_x \sum_{i=1}^{k} \frac{1}{n_y n_x}\left(\frac{n_{y_i}}{n_y} - \frac{n_{x_i}}{n_X}\right) \sim \chi^2(k-1), \text{ ha } H_0 \text{ igaz.}$$

- Az elfogadási tartomány:

$$C_0 = (x_1, ..., x_n), (y_1, ..., y_n) : \chi^2 < \chi^2_{1-\alpha}(r-1)$$

3.3.9. Példa Vizsgáljuk meg, hogy az Employee data.sav állományban homogén-e a férfiak és a nők jelenlegi fizetése és életkora?

- Transform/Compute Variable... menüpontban Traget Variable: eletkor, Numeric Expression: 2010-XDATE.YEAR(bdate)
- Transform/Recode/Into Difference Variable:gender-ből legyen sex változó, Old and new Values: f=2, m=1
- Analyze/Nonparamteric Tests/2-Independent Samples... menüpontban Test variable list: salary,eletkor, Grouping Variable: sex(1,2), Test Type: Kolmogorov-Smirnov Z

Two-Sample Kolmogorov-Smirnov Test

Frequencies

	sex	N
Current Salary	1,00	258
	2,00	216
	Total	474
eletkor	1,00	257
	2,00	216
	Total	473

Test Statistics[a]

		Current Salary	eletkor
Most Extreme Differences	Absolute	,523	,312
	Positive	,000	,182
	Negative	-,523	-,312
Kolmogorov-Smirnov Z		5,667	3,378
Asymp. Sig. (2-tailed)		,000	,000

a. Grouping Variable: sex

Az alacsony szingnifikancia szint alapján elvetjük azt a nullhipotézist, hogy a férfiak és nők fizetés és koreloszlása azonos lenne. Ha megnézzük a férfiak és nők átlagait, láthatjuk, hogy az átlagok mellett a szórások is jelentősen különböznek mindkét változóban.

Függetlenség-vizsgálat

Két ismérv valamely adott sokaságon belüli, egymástól való függetlenségének vizsgálata.

- Hipotézisünk:

$$H_0 : P_{ij} = P_i P_j (i = 1, \ldots, r; j = 1, \ldots, s)$$
$$H_1 : \exists i, j : P_{ij} \neq P_i P_j$$

Tegyük fel, hogy n számú kísérletet végeztünk, melynek eredményei két változó, X és Y értékeivel jellemezhetők. Feltesszük, hogy X és Y diszkrét valószínűségi változók, lehetséges értékeiket jelölje $x_1, x_2, ..., x_r$ és $y_1, y_2, ..., y_s$, melyek az A_1, A_2, ..., A_r és $B_1, B_2, ..., B_s$ események kimenetelei. Jelölje k_{ij} az (A_i,B_j) együttes bekövetkezésének gyakoriságát. Ezek a számok egy táblázatba rendezhetők, melyet gyakorisági táblázatnak vagy kontingencia táblázatnak nevezünk.

A sokaságot mindkét változó szerint csoportokba osztjuk, s a gyakoriságokat kontingencia táblázatban tüntetjük fel:

	B_1	B_2	$\cdots$	B_s	$\sum$
A_1	k_{11}	k_{12}	$\cdots$	k_{1s}	$k_{1.}$
A_2	k_{21}	k_{22}	$\cdots$	k_{2s}	$k_{2.}$
$\vdots$					
A_r	k_{r1}	k_{r2}	$\cdots$	k_{rs}	$k_{r.}$
$\sum$	$k_{.1}$	$k_{.2}$	$\cdots$	$k_{.s}$	N

A peremeken található számok:

$$k_{i.} = \sum_{j=1}^{s} k_{ij} \text{ (az } A_i \text{ esemény gyakorisága)}$$
$$k_{.j} = \sum_{i=1}^{r} k_{ij} \text{ (a } B_i \text{ esemény gyakorisága)}$$

- A próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - n^*_{ij})2}{n^*_{ij}} = n(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n^2_{ij}}{n_{i.} n_{.j}} - 1) \sim \chi^2((r-1)(c-1)), \text{ ha } H_0 \text{ igaz.}$$

- Az elfogadási tartomány:

$$C_0 = (x_1, ..., x_n) : \chi^2 < \chi^2_{1-\alpha}(r-1)$$

3.3.10. Példa Egy kutatócsoport azt vizsgálta, van-e összefüggés egy bizonyos betegség lefolyásának súlyossága és a betegek életkora között. A vizsgálat során 200 beteg adatait gyűjtötték össze, majd azokat csoportosították a betegség súlyossági foka és a paciens életkora szerint. Eredményül az alábbi táblázatot kapták:

		Életkor		
		40 alatti	40–60	60 fölötti
Lefolyás	enyhe	41	34	9
	közepes	25	25	12
	súlyos	6	33	15

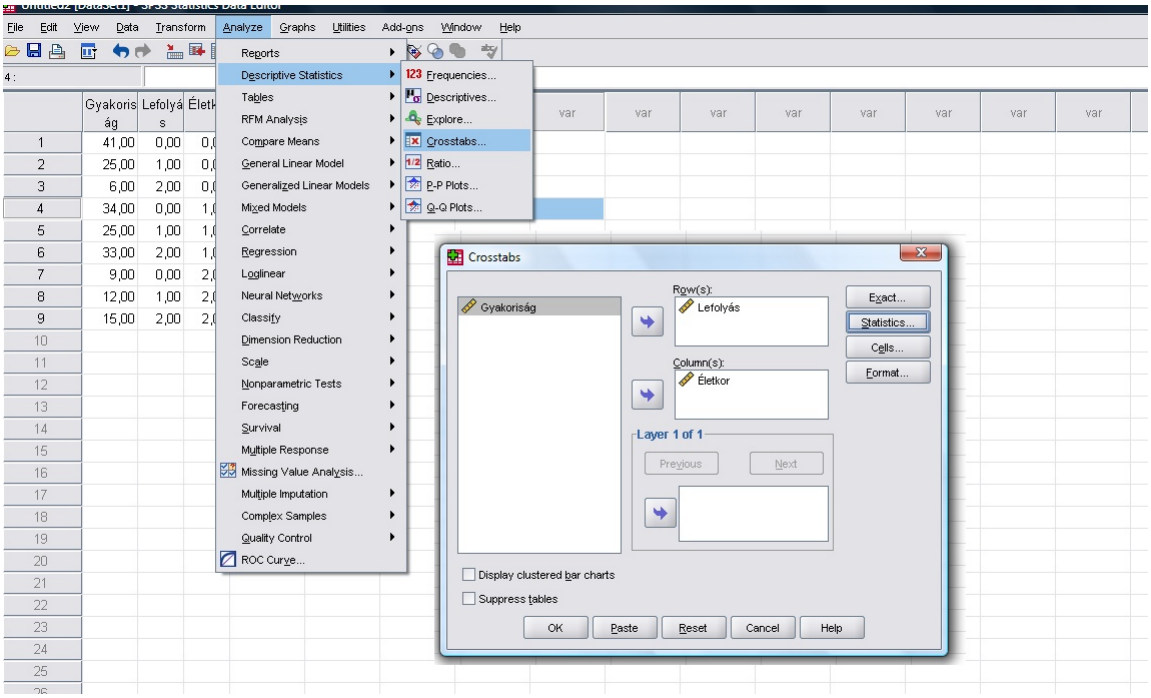
Hipotéziseit pontosan megfogalmazva döntsön 99%-os szinten, van-e összefüggés a betegek életkora és a betegség lefolyásának súlyossága között!

H_0 : nincs összefüggés;

H_1 : van összefüggés.

$\alpha = 0.01$.

SPSS:



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Lefolyás * Életkor	200	100,0%	0	,0%	200	100,0%

Lefolyás * Életkor Crosstabulation

			Életkor			
			40 év alatti	40-60	60 év feletti	Total
Lefolyás	enyhe	Count	41	34	9	84
		Expected Count	30,2	38,6	15,1	84,0
		Residual	10,8	-4,6	-6,1	
	közepes	Count	25	25	12	62
		Expected Count	22,3	28,5	11,2	62,0
		Residual	2,7	-3,5	,8	
	súlyos	Count	6	33	15	54
		Expected Count	19,4	24,8	9,7	54,0
		Residual	-13,4	8,2	5,3	
Total		Count	72	92	36	200
		Expected Count	72,0	92,0	36,0	200,0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22,523[a]	4	,000
Likelihood Ratio	25,405	4	,000
Linear-by-Linear Association	18,609	1	,000
N of Valid Cases	200		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,72.

A középső táblázat mutatja a kontingencia táblát. Megtalálhatók benne a megfigyelt gyakoriságokon kívül a várt gyakoriságok is, illetve a hozzájuk tartozó hibákat is kimutatja.

A harmadik tábla mutatja a Khi-négyzet próba eredményét, ahol a próba p-értéke 0 %, ezért elvetjük a nullhipotézist.

Összefoglalás

Az információ és az informatika korában élünk. A ránk zúduló információk özönéből nem könnyű kihámozni a számunkra hasznosat. A statisztika módszerei nagy tömegű adathalmazok kiértékelését teszik lehetővé. Egyre bővül a statisztikát felhasználók köre, akiknek a mindennapos tevékenységük során elengedhetetlenül fontos az, hogy az adatok tömegét gyorsan és helyesen fel tudják dolgozni. A közvélemény-kutató cégeknél például a felméréshez használt sokezer kérdőíveket, a szupermarketekben a vásárlók szokásait visszatükröző pénztárgépi adatokat, vagy a honlapok látogatóinak szokásait jellemző logfájlokat kell igen rövid idő alatt hatékonyan kiértékelni. Az ilyen és hasonló problémák megoldása nem képzelhető el valamilyen számítógépes statisztikai programcsomag nélkül.

A szakdolgozatomban közölt feladatokhoz a felsőoktatásban gyakran használt szoftvereket használtam. Léteznek ezeknek szabad felhasználású alternatívájuk is. Ilyen például: R, PSPP, OpenStat, Octave, stb.. Nagy előnyük még, hogy nem csak egyfajta operációs rendszeren futtathatóak, illetve letölthetőek hozzájuk különböző bővítő csomagok. Ha otthoni felhasználásban gondolkodunk és tanulás a célunk, akkor én mindenképp az ingyenesen elérhető programcsomagok közül választanék.

Irodalomjegyzék

[1] Baran Sándor: Feladatok a hipotézisvizsgálat témaköréből, mobiDIÁK könyvtár

[2] Ketskeméty László - Izsó Lajos: Bevezetés az SPSS programrendszerbe, ELTE Eötvös Kiadó, 2005

[3] Stoyan Gisbert: MATLAB: numerikus módszerek, grafika, statisztika, eszköztárak, Typotex, 2005

[4] Kerékgyártó Györgyné - L. Balogh I. - Sugár A. - Szarvas B.: Statisztikai módszerek és alkalmazásuk a gazdasági és társadalmi elemzésekben, AULA Kiadó, 2009

[5] Douglas C. Montgomery, George C. Runner: Applied Statistics and Probability for Engineers, Wiley, 2002

[6] SPSS Statistics 17.0 Algorithms: http://support.spss.com/ProductsExt/SPSS/Documentation

[7] http://www.tankonyvtar.hu/statisztika/biostatisztika-080904-92