Statisztikai vizsgálatok az Openstat programmal

Várható értékek, variancia- és hipotézisvizsgálat

Készült a Matematika Szabadon kurzushoz

Készítette:

Bodnár Péter (bopnaat.sze)
SZTE TTK műszaki informatika
III. évf.

2007. november

Tartalomjegyzék:

Elméleti háttér

A statisztikai tudományos folyamat indukciós és dedukciós logikát, logikai szimbólumokat és matematikát használ. A folyamat menete:

- szisztematikus megfigyelések: Adott jelenség leírása, formalizálás, szimbólumok használata.
- lehetséges összefüggések kijelentése (hipotézis megfogalmazása)
- megfigyelések a hipotézis alapján, különbségek, egybeesések (kísérlet, a hipotézis tesztelése)
- a hipotézis elfogadása vagy elvetése a vizsgálat alapján (az eredmény értelmezése, hibakorlát figyelembe vétele)
- a kapcsolat vizsgálata (jóslás, következtetés későbbi eredményekre vizsgálat nélkül)
- a megfigyelt kapcsolatok kihasználása (a megfigyelt folyamat, jelenség irányítása, az adatok felhasználása)
- további kapcsolatok keresése és megfigyelése (teória, kutatás)

A hipotézisvizsgálat során a hipotézisvizsgálat lépéseit kell végrehajtani.

1. A tesztelni kívánt – nullhipotézisnek nevezett – feltételezés megfogalmazása.
2. Ezzel szemben mindig van egy alternatív hipotézis.
3. A nullhipotézist és a rendelkezésre álló információkat figyelembe véve a
4. próbafüggvény kiválasztása. Ezt a szoftverek – többnyire – elvégzik helyettünk.
5. A 0-hoz közeli α szignifikanciaszint kiválasztása, és a próbafüggvény értékkészletének elfogadási és kritikus tartományra bontása.
6. A próbafüggvény mintán felvett értékének megállapítása.
7. Döntés a nullhipotézis helyességének elfogadásáról-elvetéséről.

Mivel a statisztikai szoftverek jelentős hányada, az elemzések egy részénél azt feltételezi, hogy az adatállományunk egy minta jellemzői, ezért az elemzések során többnyire hipotézisvizsgálatot is folytatunk. A szoftverek segítségével semmit sem kell számolnunk. Azonban, egyrészt tudnunk kell, hogy az adott hipotézisvizsgálatnak mi a nullhipotézise, illetve az alternatív hipotézise, másrészt pedig, az eredményeket helyesen kell tudnunk értelmezni.

A statisztikai szoftverek hipotézisvizsgálat során az outputon megadnak egy értéket, a p-érték, p-value, Sig. jelölések valamelyikével. Ez az érték a nullhipotézis teljesülésének alószínűségét jelenti. Tehát, ha például a vizsgálataink során 5 százalékos szignifikanciaszintet használunk, akkor amennyiben a kapott érték 0,05-nál kisebb, akkor a nullhipotézist – ötszázalékos szignifikanciaszint mellett – elvetjük.

A nullhipotézis-vizsgálat elfogadására ill. elvetésére használt táblázat:

Possible Outcomes of an Experiment

Experimenter conclusion based on observed data		True State of Nature: H_O True	True State of Nature: H_O False
	accept H_O	$1 - \alpha$	ß Type II error
	reject H_O	Type I Error α	$1 - ß$

A hipotézisvizsgálat táblázatának magyarázata megtalálható a legtöbb statisztikával foglalkozó könyvben, például *Solt György: Valószínűségszámítás* könyvében, ezért a definíciókra, a hibaelemzésre, a próbafüggvények magyarázatára, eloszlásokra nem térek ki. A konkrét problémák megértéséhez ez nem is szükséges, csak a statisztikai vizsgálatok menetének pontos megértésében lenne szerepe, ami jóval meghaladná az esszé korlátait.

A következő táblázat összefoglalja a várható értékek vizsgálatának döntési táblázatát.

Várható értékek összehasonlítása	**Sorrendi mérési szintű változó (ordinális skála)**	**Metrikus változó (intervallumskála, arányskála)**
	Medián	**Átlag**
Várható érték összehasonlítása egy adott értékkel	WILCOXON-próba	Egymintás *z*-próba (ha az alapsokaság szórása ismert). Egymintás *t*-próba (ha az alapsokaság szórása nem ismert).
Két várható érték összehasonlítása egymással	MANN-WHITNEY próba	Kétmintás *z*-próba (ha az alapsokaságok szórása ismert). Kétmintás *t*-próba (ha az alapsokaságok szórása nem ismert. Ekkor külön tesztelnünk kell *F*-próbával a varianciák egyezőségét.)

Egymintás próbák esetén meg kell adnunk, hogy a várható értéket milyen konkrét értékkel szeretnénk összehasonlítani. Kétmintás próbák esetében, vagy csak a vizsgált két változót kell megadnunk, vagy pedig a várt eltérést is. Ennek az az oka, hogy nem csak a két várható érték egyezőségét, hanem adott értékkel való különbözőségét is lehet tesztelni. Ezeknél a próbáknál a nullhipotézis mindig a várható értékek egy adott értékkel történő, egymással való megegyezését, vagy különbségük adott értékkel való egyezőségét jelenti.

Kétmintás t-próba alkalmazására általában az alábbi három esetben kerül sor.

1. Kétmintás, páros t-próba. Ekkor az egyes megfigyelések egymással párba állíthatók. Tipikusan erről az esetről van szó, ha egy csoportot megfigyelünk valamilyen kísérlet előtt és után.
2. Kétmintás t-próba egyenlő varianciákkal. A próba előtt, F-próbával meg kell vizsgálnunk a varianciák egyezőségét.
3. Kétmintás t-próba nem egyenlő varianciákkal. A próba előtt, F-próbával meg kell vizsgálnunk a varianciák egyezőségét.

Az OpenStat programról

Az OpenStatot Bill Miller írta C++ban az Iowa Állami Egyetemen, a viselkedéstudományi intézet diákjai számára. Képes többváltozós statisztikai analízisre, nemparametrikus statisztikai műveletekre, folyamatirányításra és tartalmaz adatmanipulációs eszközöket is. Windows és Linux rendszereken egyaránt használható, ingyenes, nem kereskedelmi célú statisztikai szoftver.

Windows verzió:

http://www.statpages.org/miller/openstat/OpenStatSetup.exe

Linux alatt használható a WINE, vagy az OpenStat Linuxra gyártott verziója, a LinOStats. Ez a változat Kylix-ban íródott és Qt-t használ grafikus megjelenítésre.

Letölthető a bináris fájl:

http://www.statpages.org/miller/openstat/LINOSTATS.TGZ

vagy a forrás:

http://www.statpages.org/miller/openstat/LinOStatSOURCE.ZIP

A problémák bemutatásánál Windows verziót használtam, mert a szerző megjegyzése szerint nem minden funkció működik tökéletesen a Linuxra átvett változatban, és ezeket még idő hiányában nem tudta kijavítani. Főleg a Qt grafikus megjelenítőben lehetnek bugok, néha az eredményeket pontatlanul ábrázolja, például diagramoknál. A szerző arra bátorít, hogy ha jártasak vagyunk a pascal nyelvben, módosíthatjuk mi is a forrást.

A programhoz mellékeltek továbbá használati példafájlokat:

http://www.statpages.org/miller/openstat/OpenStatTabData.zip

Néhányat ennél az esszénél is hasznosítottam, a fájlok tartalmát mellékeltem.

Parametrikus analízis Openstattal

Az OpenStat statisztikai program lehetőséget nyújt, hogy mintáinkat vizsgálatoknak vessük alá. Analizálhatjuk az átlagot, a szórást, készíthetünk lineáris regressziót, és még számos lehetőség közül választhatunk. A következő oldalakon ezek közül a várható értékek (átlagok), és a varianciák (szórásnégyzetek, szórások) elemzésével foglalkozunk.

Minták várható értékeinek összehasonlítása

Egyik leggyakoribb kérdés a kutatásban: Különböznek-e a csoportok átlagai? Főleg ez akkor érdekes, ha a csoportok egy sokaságból származnak. A hipotézisvizsgálat akkor történik, amikor csak két minta van, és a két minta szórása megegyezik. Ha a különbség nulla, akkor elfogadjuk, hogy a két minta ugyanabból a sokaságból származik. Ha a hipotézist nem fogadjuk el, akkor a minták különböző sokaságból származnak. Gyakran a mintákat kísérleti alanyokból vesszük. Például egy orvos két különböző gyógymódot tesztel egy diagnózisra. Kiválaszt véletlenszerű egyedeket a sokaságból ugyanazzal a diagnózissal, azután méri a hatékonysági arányt a két csoporton. Ezután választhatja a hatékonyabb gyógymódot, vagy egyenlő hatékonyság mellett az olcsóbbikat. Ha különbözik a hatékonyság, felvetődik a kérdés, mennyire. A paramétereket és a mintavételi hibát figyelembe véve dönthetünk a hipotézis elfogadása vagy elutasítása felől.

A t-próba

Megmutatjuk a t-próbás hipotézisvizsgálatot egy demonstrációs fájlon, amit az Openstat disztribúcióhoz mellékeltek. A fájl neve *ancova3.tab.*

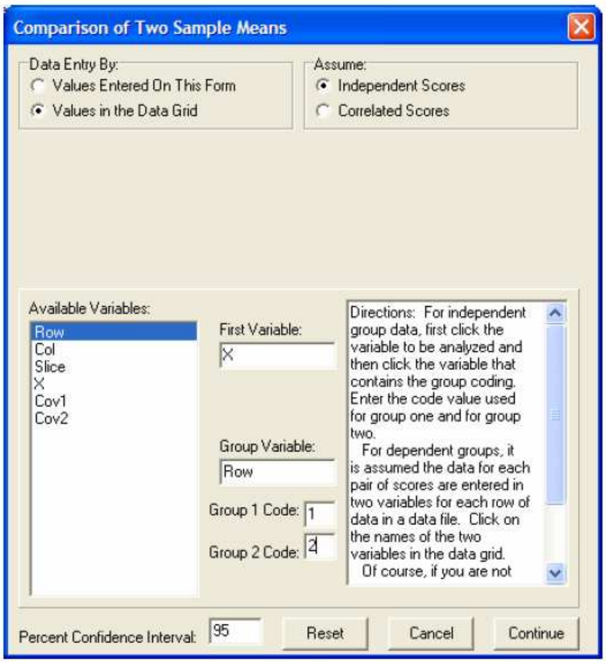
A bemeneti fájl tartalma:

Row	Col	Slice	X	Cov1	Cov2
1	1	1	1.00	3.00	1.00
1	1	1	2.00	1.00	3.00
1	1	1	3.00	4.00	3.00
1	1	2	2.00	6.00	1.00
1	1	2	3.00	5.00	5.00
1	1	2	4.00	3.00	4.00
1	1	3	3.00	4.00	6.00
1	1	3	4.00	1.00	3.00
1	1	3	5.00	7.00	1.00
1	2	1	6.00	5.00	2.00
1	2	1	5.00	4.00	4.00
1	2	1	4.00	6.00	4.00
1	2	2	5.00	3.00	5.00
1	2	2	4.00	3.00	3.00
1	2	2	3.00	4.00	4.00
1	2	3	4.00	1.00	6.00
1	2	3	3.00	2.00	6.00
1	2	3	2.00	6.00	4.00
2	1	1	1.00	4.00	5.00
2	1	1	2.00	5.00	5.00
2	1	1	3.00	5.00	2.00
2	1	2	4.00	8.00	3.00
2	1	2	5.00	5.00	4.00
2	1	2	6.00	4.00	5.00
2	1	3	1.00	3.00	2.00
2	1	3	2.00	3.00	3.00
2	1	3	3.00	5.00	4.00
2	2	1	4.00	2.00	5.00
2	2	1	5.00	4.00	4.00

2	2	1	6.00	6.00	6.00
2	2	2	5.00	4.00	8.00
2	2	2	6.00	5.00	6.00
2	2	2	7.00	3.00	3.00
2	2	3	7.00	3.00	4.00
2	2	3	8.00	4.00	5.00
2	2	3	9.00	1.00	2.00
1	1	1	1.00	3.00	1.00
1	1	1	2.00	1.00	3.00
1	1	1	3.00	4.00	3.00
1	1	2	2.00	6.00	1.00
1	1	2	3.00	5.00	5.00
1	1	2	4.00	3.00	4.00
1	1	3	3.00	4.00	6.00
1	1	3	4.00	1.00	3.00
1	1	3	5.00	7.00	1.00
1	2	1	6.00	5.00	2.00
1	2	1	5.00	4.00	4.00
1	2	1	4.00	6.00	4.00
1	2	2	5.00	3.00	5.00
1	2	2	4.00	3.00	3.00
1	2	2	3.00	4.00	4.00
1	2	3	4.00	1.00	6.00
1	2	3	3.00	2.00	6.00
1	2	3	2.00	6.00	4.00
2	1	1	1.00	4.00	5.00
2	1	1	2.00	5.00	5.00
2	1	1	3.00	5.00	2.00
2	1	2	4.00	8.00	3.00
2	1	2	5.00	5.00	4.00
2	1	2	6.00	4.00	5.00
2	1	3	1.00	3.00	2.00
2	1	3	2.00	3.00	3.00
2	1	3	3.00	5.00	4.00
2	2	1	4.00	2.00	5.00
2	2	1	5.00	4.00	4.00
2	2	1	6.00	6.00	6.00
2	2	2	5.00	4.00	8.00
2	2	2	6.00	5.00	6.00
2	2	2	7.00	3.00	3.00
2	2	3	7.00	3.00	4.00
2	2	3	8.00	4.00	5.00
2	2	3	9.00	1.00	2.00

A változók nevei *row, col, slice, X, cov1* és *cov2*. Az *X* változó a függő, ami a kísérlet kimenetének mérésére szolgál.

A *row* változó mutatja a két különböző gyógymódot a 72 alanyra, csoportonként 36 lesz. Az analízishez kiválasztjuk a *„test of two means”*-t az *Univariate* almenüből, ami az *Analyses* menüben található. Ezt a képet kell látnunk:



Vegyük észre hogy kiválasztottuk a *Values in the Data Grid*-et, tehát a bemenetet a program „*adatrácsáról*" kapjuk, és feltételezzük a másik rádiógomb-választóban (*Independent scores*), hogy a kiválasztott alanyok egymástól függetlenek. Először az X-változóra kattintunk a listán, azután a *Row*-ra, ami a különböző csoportokat szimbolizálja. A két csoport 1-es és 2-es számokkal van jelezve, tehát beírjuk *Group1 Code*-nak az 1-et, és hasonlóan a másikhoz 2-t. Végül, megadjuk a konfidencia-intervallumot 0,95-nek. Itt természetesen adhatunk mást is meg. A *Continue* gombra kattintva az alábbi kimenetet kapjuk:

```
COMPARISON OF TWO MEANS
Variable Mean Variance Std.Dev. S.E.Mean N
Group 1 3.50 1.63 1.28 0.21 36
Group 2 4.67 5.37 2.32 0.39 36
Assuming = variances, t = -2.646 with probability = 0.0101 and 70 degrees of freedom
Difference = -1.17 and Standard Error of difference = 0.44
Confidence interval = (-2.05, -0.29)
Assuming unequal variances, t = -2.646 with probability = 0.0106 and 54.44 degrees of
freedom
Difference = -1.17 and Standard Error of difference = 0.44
Confidence interval = (-2.05, -0.28)
F test for equal variances = 3.298, Probability = 0.0003
NOTE: t-tests are two-tailed tests.
```

Figyeljük meg, hogy az átlagok közti 1,17-es eltérés gyakorisága 0,0101 véletlen mintavételi hibánál, de csak akkor, ha a két csoport szórása megegyezik. Mégis, a teszt szerint a szórások látványosan különböznek (1,63 és 5,37), tehát olyan konzervatívabb tesztre lenne szükségünk, ami nem feltételez egyenlő szórást. Ez még mindig ahhoz a zavarhoz vezet, hogy a kezelések nem egyenlően hatékonyak, és a 2. csoportnak magasabb átlaga van. Ha a nagyobb átlag magasabb fehérvérsejt-számot jelentene például, akkor azt a következtetést vonhatnánk le, hogy az 1. gyógymód hatékonyabban fojtotta el a betegséget. Azt is megvizsgálhatjuk, miért lett nagyobb szórása a második gyógymódnak. Emlékezzünk rá, hogy a véletlenszerű minta különbséghez vezet, amire számítanunk kell a hibák elemzésénél.

Excel feladat megvalósítása OpenStatban

Ebben a fejezetben az Informatika II. jegyzetben megadott Excel mintafeladatot fogalmazzuk át az Openstat program számára, azután elvégezzük az elemzést.

Működési különbségek

Az Excel Adatelemzés bővítménye szolgál a statisztikai analízis elvégzésére. Választhatunk párosított vagy nem párosított kétmintás t-próba közül, illetve egyenlő és nem egyenlő szórásnégyzetekkel is tesztelhetünk. A kétmintás F-próba a szórásnégyzetre megvizsgálja a két minta varianciájának egyezőségét. Az Openstat képes egy mintát analizálni, vagy két mintás próbákat elvégezni egyetlen lépésben. További nagy előnye az Excellel szemben, hogy ingyenes, és konkrét statisztikai vizsgálatokra van specializálva.

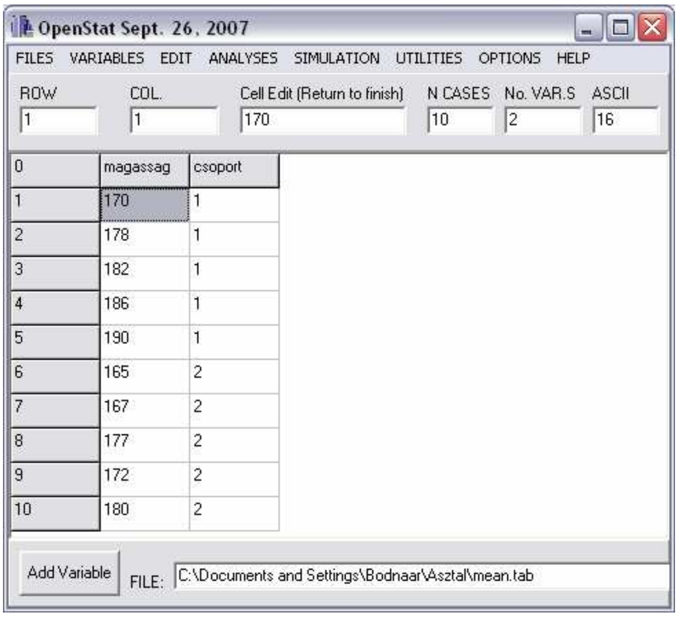
Az Excel példafeladat

Az alábbi táblázatban fiúk és lányok testmagassága van megadva. Vizsgáljuk meg ötszázalékos szignifikanciaszint mellett, hogy a lányok és a fiúk átlagos testmagassága azonosnak tekinthető-e!

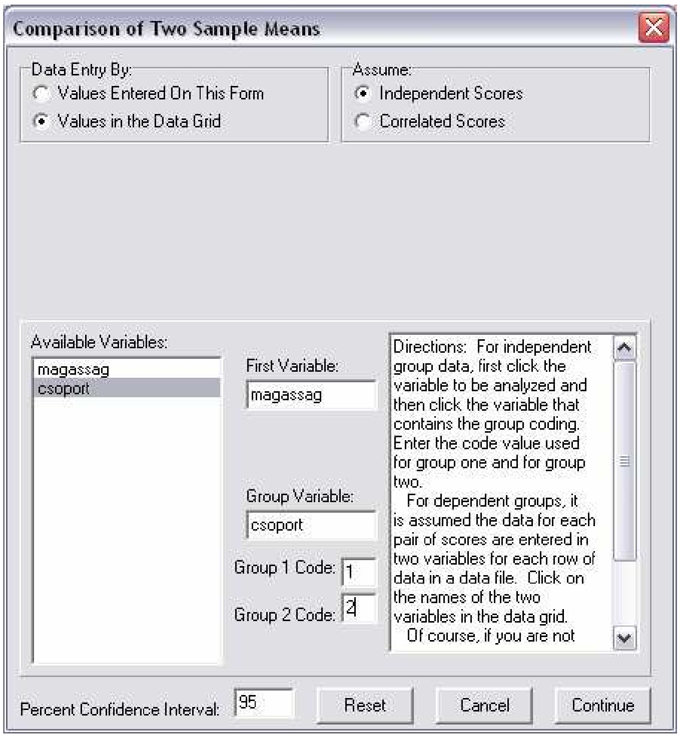
Lányok testmagassága (cm)	Fiúk testmagassága (cm)
165	170
167	178
177	182
172	186
180	190

A feladat megoldása Openstatban

Nyissuk meg az OpenStat-ot, ahol a *Data Grid*-et látjuk. Egyetlen változóval, a magassággal fogjuk feltölteni a táblázatot, mégpedig sorban, egymás alá. Ezután vegyünk fel egy új változót csoport névvel, ami majd megjelöli a fiúkat 1-el, a lányokat 2-vel.



Ezután az *Analyses* >> *Comparisons* >> *Comparison of Two Sample Means* menüpontot válasszuk ki. Inputnak adjuk meg a *Data Grid*-et, az *Assume*-nál hagyjuk meg független értékeknek. A változónk a *magassag* lesz, a group variable pedig a *csoport*. A csoportkódok fiúnál 1, lánynál 2. Legyen a konfidencia 95%, vagyis az α érték 0,05.



A kimenet:

```
COMPARISON OF TWO MEANS

Variable       Mean    Variance  Std.Dev.  S.E.Mean  N
Group 1       181.20     59.20      7.69      3.44   5
Group 2       172.20     40.70      6.38      2.85   5

Assuming equal variances,
t = 2.013 with probability = 0.0789 and 8 degrees of freedom
Difference = 9.00 and Standard Error of difference = 4.47
Confidence interval = (   -1.31,   19.31)

Assuming unequal variances,
t = 2.013 with probability = 0.0801 and  7.73 degrees of freedom
Difference = 9.00 and Standard Error of difference = 4.47
Confidence interval = (   -1.38,   19.38)

F test for equal variances = 1.455, Probability = 0.3627
```

Értelmezzük a kimenetet. Az első csoport (fiúk) átlagos magassága 181,2, szórása 7,69, a hibaátlag pedig 3,44. A program még megjegyzi a varianciát, illetve hogy 5 értékből állt a minta. Lányok magasságánál hasonlóan értelmezhetjük az adatokat. Ezután két értékelést ad a program, egyiket egyenlő szórást feltételezve, a másikat különböző szórást feltételezve. Az egyenlő szórásokat F-próbával teszteli, ennek értéke 0,3627(>0,05), ezt az adott szignifikancia-szinten elfogadjuk. Ezután már csak az egyenlő varianciát feltételező ágat kell néznünk. Megkapjuk a t-értéket és a valószínűséget. P=0,0789>0,05, tehát az adott nullhipotézist elfogadjuk.

Ezek szerint a fiúk és a lányok átlagos testmagassága azonosnak tekinthető. Vegyük észre, hogy például tíz százalékos szignifikancia-szint mellett más döntést hoztunk volna.

További vizsgálatok

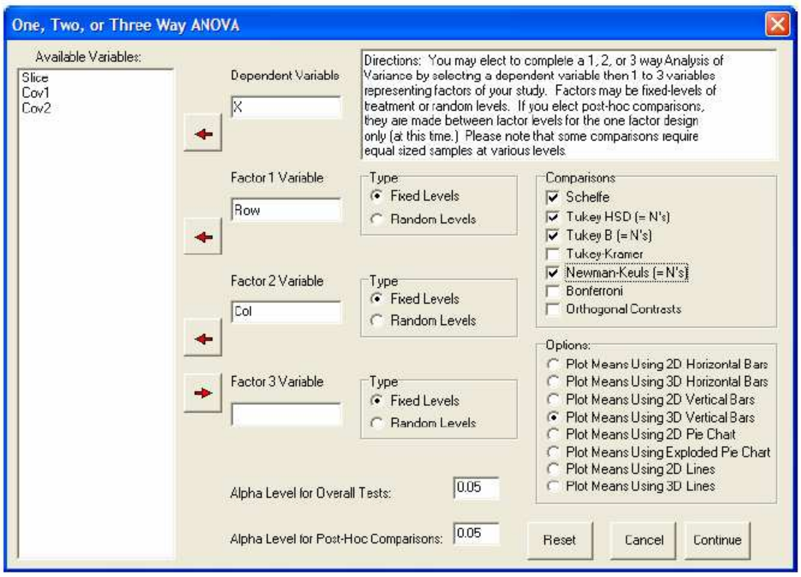
A hipotézisvizsgálat kapcsán nem csak az átlagok, de a szórások vizsgálata is érdekes lehet. Mélyebb betekintés nélkül nézzünk meg néhány példát parametrikus és nemparametrikus módszerekre az Openstat programmal.

Szórás parametrikus vizsgálata

Openstat: One, Two or Three Way ANOVA

Ha több, mint két csoportunk van, a t-próba nem a legjobb módszer az átlagok vizsgálatára. Ha több t-próbát végzünk, az elutasítás valószínűsége megnő. A kedveltebb módszer a szórások vizsgálata.

Folytatjuk a fenti orvosi példát: *ancova3.tab*.
Feltételezzük, hogy az orvos két különböző gyógyszerrel kezel egy fertőzést ami emelkedő fehérvérsejtszámot eredményez (X-változó), és a két 36-fős csoportokat további két csoportra osztjuk, a szerint, hogy kapnak-e lázcsillapítót, vagy sem. Ezek a csoportok 1-es és 2-es számot, illetve „*col*” címkét kapnak. A kutatónk azt teszteli, reakcióba lépnek-e ezek a gyógyszerek egymással, és van-e „szignifikáns” eset. Válasszuk a *„one, two or three way anova”*-t az *Analysis of Variance* menüből.



Figyeljük meg, hogy a függő változónk még mindig X, de két faktort választottunk ki független változóinknak. Ezek „*fixed-level*” értékőek, vagyis nem véletlenszerűen választjuk őket a kezelés fajtáihoz. Ábrázoljuk is az eredményt 3D oszlopdiagram segítségével. A szignifikanciát hagyjuk 5%-on. A futás eredménye:

Two Way Analysis of Variance
Variable analyzed: X
Factor A (rows) variable: Row (Fixed Levels)
Factor B (columns) variable: Col (Fixed Levels)
SOURCE D.F. SS MS F PROB.> F Omega Squared
Among Rows 1 24.500 24.500 12.250 0.001 0.083
Among Columns 1 84.500 84.500 42.250 0.000 0.304
Interaction 1 24.500 24.500 12.250 0.001 0.083
Within Groups 68 136.000 2.000
Total 71 269.500 3.796
Omega squared for combined effects = 0.470
Note: Denominator of F ratio is MSErr
Descriptive Statistics
GROUP Row Col. N MEAN VARIANCE STD.DEV.
Cell 1 1 18 3.000 1.412 1.188
Cell 1 2 18 4.000 1.412 1.188
Cell 2 1 18 3.000 2.824 1.680
Cell 2 2 18 6.333 2.353 1.534
Row 1 36 3.500 1.629 1.276
Row 2 36 4.667 5.371 2.318
Col 1 36 3.000 2.057 1.434
Col 2 36 5.167 3.229 1.797
TOTAL 72 4.083 3.796 1.948
TESTS FOR HOMOGENEITY OF VARIANCE

Hartley Fmax test statistic = 2.00 with deg.s freem: 4 and 17.
Cochran C statistic = 0.35 with deg.s freem: 4 and 17.
Bartlett Chi-square statistic = 7.26 with 3 D.F. Prob. larger value = 0.064

COMPARISONS AMONG COLUMNS WITHIN EACH ROW
ROW 1 COMPARISONS

Scheffe contrasts among pairs of means.
alpha selected = 0.05
Group vs Group Difference Scheffe Critical Significant?
Statistic Value
----
1 2 -1.00 2.12 2.074 YES

Tukey HSD Test for (Differences Between Means
alpha selected = 0.05
Groups Difference Statistic Probability Significant?
----
1 - 2 -1.000 q = 3.000 0.0376 YES

Tukey B Test for (Contrasts on Ordered Means
alpha selected = 0.05

Groups Difference Statistic d.f. Prob.>value Significant?
1 - 2 -1.000 3.000 2, 68 0.038 YES

Neuman-Keuls Test for (Contrasts on Ordered Means
alpha selected = 0.05
Group Mean
1 3.000
2 4.000
Groups Difference Statistic d.f. Probability Significant?
----
1 - 2 -1.000 q = 3.000 2 68 0.0376 YES

ROW 2 COMPARISONS

Scheffe contrasts among pairs of means.
alpha selected = 0.05
Group vs Group Difference Scheffe Critical Significant?
Statistic Value
----
1 2 -3.33 7.07 2.074 YES

Tukey HSD Test for (Differences Between Means
alpha selected = 0.05
Groups Difference Statistic Probability Significant?
----
1 - 2 -3.333 q = 10.000 0.0001 YES

Tukey B Test for (Contrasts on Ordered Means
alpha selected = 0.05

Groups Difference Statistic d.f. Prob.>value Significant?
1 - 2 -3.333 10.000 2, 68 0.000 YES

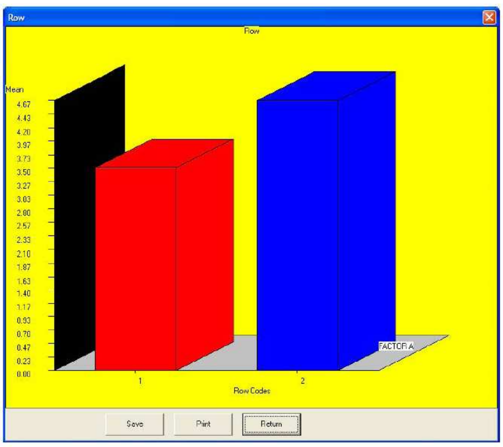
Neuman-Keuls Test for (Contrasts on Ordered Means
alpha selected = 0.05
Group Mean
1 3.000
2 6.333
Groups Difference Statistic d.f. Probability Significant?
----
1 - 2 -3.333 q = 10.000 2 68 0.0001 YES

COMPARISONS AMONG ROWS WITHIN EACH COLUMN
COLUMN 1 COMPARISONS

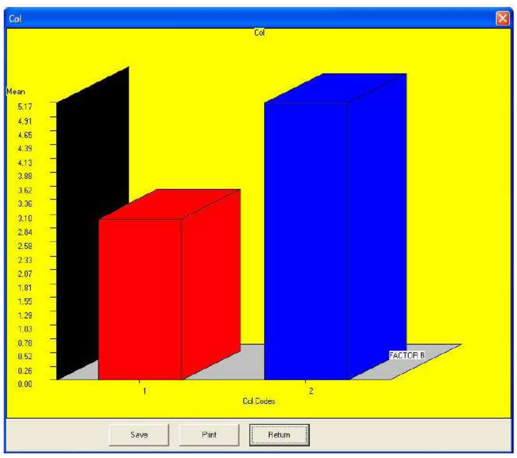
Scheffe contrasts among pairs of means.
alpha selected = 0.05
Group vs Group Difference Scheffe Critical Significant?
Statistic Value

```
1 2 0.00 0.00 2.074 NO
----------------------------------------------------------------
---------------------------------------------------------------
Tukey HSD Test for (Differences Between Means
alpha selected = 0.05
Groups Difference Statistic Probability Significant?
---------------------------------------------------------------
1 - 2 0.000 q = 0.000 1.0000 NO
---------------------------------------------------------------
---------------------------------------------------------------
Tukey B Test for (Contrasts on Ordered Means
alpha selected = 0.05
---------------------------------------------------------------
Groups Difference Statistic d.f. Prob.>value Significant?
1 - 2 0.000 0.000 2, 68 1.000 NO
----------------------------------------------------------------------
Neuman-Keuls Test for (Contrasts on Ordered Means
alpha selected = 0.05
Group Mean
1 3.000
2 3.000
Groups Difference Statistic d.f. Probability Significant?
----------------------------------------------------------------------
1 - 2 0.000 q = 0.000 2 68 1.0000 NO
----------------------------------------------------------------------
COLUMN 2 COMPARISONS
----------------------------------------------------------------
Scheffe contrasts among pairs of means.
alpha selected = 0.05
Group vs Group Difference Scheffe Critical Significant?
Statistic Value
----------------------------------------------------------------
1 2 -2.33 4.95 2.074 YES
----------------------------------------------------------------
---------------------------------------------------------------
Tukey HSD Test for (Differences Between Means
alpha selected = 0.05
Groups Difference Statistic Probability Significant?
---------------------------------------------------------------
1 - 2 -2.333 q = 7.000 0.0001 YES
---------------------------------------------------------------
---------------------------------------------------------------
Tukey B Test for (Contrasts on Ordered Means
alpha selected = 0.05
---------------------------------------------------------------
Groups Difference Statistic d.f. Prob.>value Significant?
1 - 2 -2.333 7.000 2, 68 0.000 YES
----------------------------------------------------------------------
Neuman-Keuls Test for (Contrasts on Ordered Means
alpha selected = 0.05
Group Mean
1 4.000
2 6.333
Groups Difference Statistic d.f. Probability Significant?
----------------------------------------------------------------------
1 - 2 -2.333 q = 7.000 2 68 0.0001 YES
----------------------------------------------------------------------
```

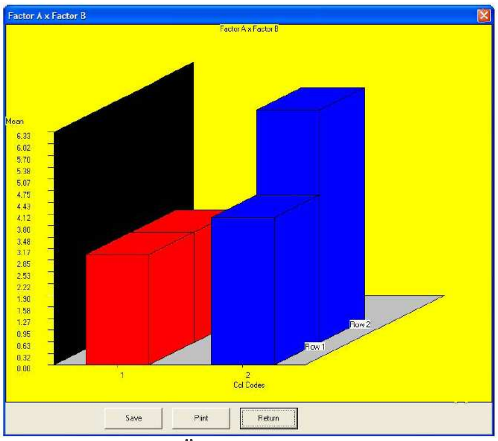
Néhány dolgot észrevehetünk a fenti vizsgálatban. Először is, találunk kiugró különbséget a sorok (anti-biotikum) és az oszlopok (lázcsillapító) között. Ráadásul az egymásra való hatásuk is „szignifikáns”, vagyis nem fogadhatjuk el a nullhipotézist.
Nézzük meg a diagramok eredményeit:



Row1 és Row2 átlagai



Column1 és Column2 átlagai



Összesített kép

Az utolsó ábra mutatja legjobban, a két hatóanyag együttes hatását. Figyeljük meg, hogy az *anti-inflammatory treatment* 2-ben az *anti-biotic treatment* 2 átlaga jóval nagyobb, mint a többi átlag. Ahogy az előző tesztek is mutatják, a minták átlaga nem különbözik a két sorban de a második oszlopban igen. Ez azt jelenti, az *Anti-biotic 1* a leghatékonyabb és az lép legkevésbé reakcióba mással.

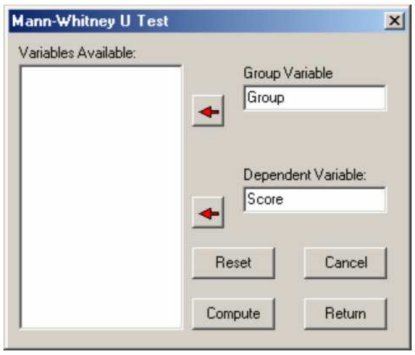
Szórás vizsgálata Mann-Whitney és Wilcoxon-próbákkal

Mann-Whitney Próba

Alternatíva a Student t-próbára ordinális skálával. Példafájl Openstatban: *MannWhitU.tab*.

Group	Score
1	13
1	12
1	12
1	10
1	10
1	10
1	10
1	9
1	8
1	8
1	7
1	7
1	7
1	7
1	7
1	6
2	17
2	16
2	15
2	15
2	15
2	14
2	14
2	14
2	13
2	13
2	13
2	12
2	12
2	12
2	12
2	11
2	11
2	10
2	10
2	10
2	8
2	8
2	6

Válasszuk a *Statistics / Non Parametric / Mann-Whitney U Test* menüpontot:



A példa kimenete:

```
Score Rank Group
6.00 1.50 1
6.00 1.50 2
7.00 5.00 1
7.00 5.00 1
7.00 5.00 1
7.00 5.00 1
7.00 5.00 1
8.00 9.50 1
8.00 9.50 2
8.00 9.50 2
8.00 9.50 1
9.00 12.00 1
10.00 16.00 1
10.00 16.00 2
10.00 16.00 2
10.00 16.00 2
10.00 16.00 1
10.00 16.00 1
10.00 16.00 1
11.00 20.50 2
11.00 20.50 2
12.00 24.50 2
12.00 24.50 2
12.00 24.50 2
12.00 24.50 2
12.00 24.50 1
12.00 24.50 1
13.00 29.50 1
13.00 29.50 2
13.00 29.50 2
13.00 29.50 2
14.00 33.00 2
14.00 33.00 2
14.00 33.00 2
15.00 36.00 2
15.00 36.00 2
15.00 36.00 2
16.00 38.00 2
17.00 39.00 2
Sum of Ranks in each Group
Group Sum No. in Group
1 200.00 16
2 580.00 23
No. of tied rank groups = 9
Statistic U = 304.0000
z Statistic (corrected for ties) = 3.4262, Prob. > z = 0.0003
```

Wilcoxon próba

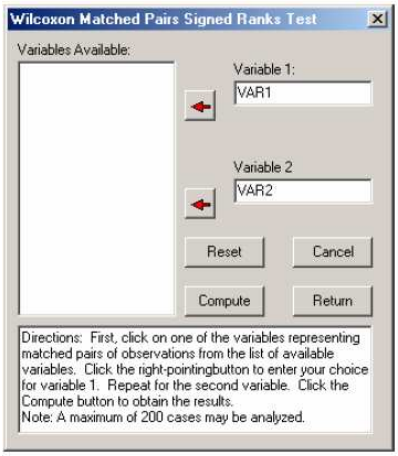
Ez a próba egy alternatívát ad a student t-próbára, ahol a párosításban a parametrikus t-próba nem adna megfelelő eredményt. Ennél a tesztnél minden párnál külön számolunk eltérést, például egy teszt előtti és teszt utáni álapotában a mintának. A negatív és a pozitív eltéréseket szummázzuk. A teszt statisztika a két summa kisebbike. Ha a nullhipotézis igaz, akkor a pozitív szumma nem különbözik a negatívtól a várt valószínűségnél jobban.

A példafájlt Opentatban a *Wilcoxon.tab*.

```
VAR1         VAR2
82           63
69           42
73           74
```

43	37
58	51
56	43
76	80
65	62

Töltsük ezt be, aztán a *Statistics / NonParametric / Wilcoxon Matched-Pairs Signed Ranks Test* -et válasszuk a menüben. Ezt fogjuk látni:

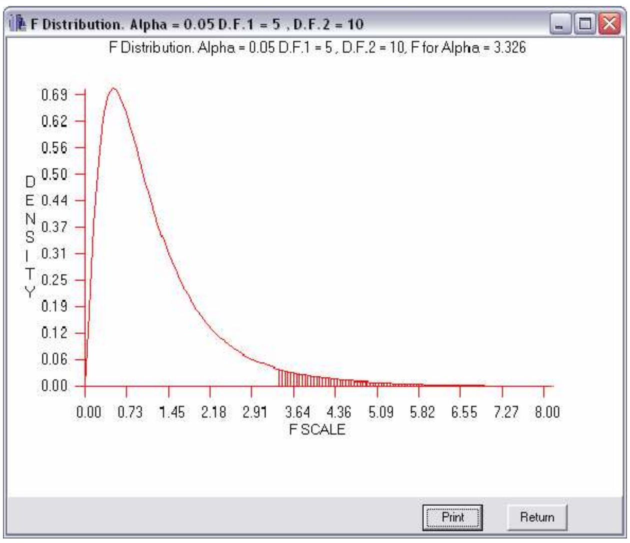


A próba kimenete itt látható:

```
The Wilcoxon Matched-Pairs Signed-Ranks Test
See pages 75-83 in S. Seigel's Nonparametric Statistics for
the Social Sciences
Ordered Cases with cases having 0 differences eliminated:
Number of cases with absolute differences greater than 0 = 8
CASE VAR1 VAR2 Difference Signed Rank
3 73.00 74.00 -1.00 -1.00
8 65.00 62.00 3.00 2.00
7 76.00 80.00 -4.00 -3.00
4 43.00 37.00 6.00 4.00
5 58.00 51.00 7.00 5.00
6 56.00 43.00 13.00 6.00
1 82.00 63.00 19.00 7.00
2 69.00 42.00 27.00 8.00
Smaller sum of ranks (T) = 4.00
Approximately normal z for test statistic T = 1.960
Probability (1-tailed) of greater z = 0.0250
NOTE: For N < 25 use tabled values for Wilcoxon Test
```

Az OpenStat további eszközei

A program lehetőséget ad néhány egyszerű számítás elvégzésére, egy *kcalc* vagy a windowsból ismert *calc* szintű számológépet biztosít erre. Tartalmaz egy képmegjelenítőt, ahol grafikonok, kimenetek, de egyszerű képfájlok böngészésére is lehetőségünk van. Pár soros, formázatlan jegyzet készítéséhez egy beépített szövegszerkesztőt lehet használni, ami a notepadnak felel meg. Megrajzolhatunk és elemezhetünk vele továbbá eloszlásokat, sűrűségfüggvényeket, tendenciákat. Példa egy 5%os hibahatárral rendelkező F-disztribúcióra:



Összefoglaló: OpenStat

Az OpenStat ingyenes statisztikai szoftver lehetőséget ad egyszerű és egészen komplex statisztikai problémák elemzésére, szimulációra, modellezésre, és tartalmaz néhány egyszerű eszközt is, például kalkulátort, képnézegetőt, szövegszerkesztőt is. Megfelelő segítség informatikusok, közgazdászok, pénzügyi elemzők, vagy matematikai statisztikával foglalkozók számára.

Források

- Kovács Péter: Informatika II. Számítógépes adatelemzés az Excel Adatelemzés… bővítményével (jegyzet)
- William G. Miller, PhD: Quick Guide to OpenStat
- William G. Miller, PhD: Statistics and Measurement Using OpenStat
- Openstat official example pack (ingyenes példák az Openstat programhoz)
- www.statpages.org/miller/openstat/