Universitat de Barcelona

Treball de camp: Anàlisi i conclusions.

Introducció a la Inferència

Ferran Pérez Anton
Marc Casas Llacer
Laura Julià Melis
31/05/2017

Índex

1. Introducció.

En el present informe s'inclouen els principals aspectes d'un estudi referent a dues habilitats: la capacitat de càlcul i la capacitat física. En aquest estudi es durà a terme una comparació entre els àmbits de la salut i les ciències socials i humanitats per observar si existeixen diferències entre les dues capacitats.

El primer pas d'aquesta investigació fou la elaboració d'un qüestionari que permetés mesurar les habilitats a estudiar. Després va tenir lloc la recollida de dades mitjançant enquestes, realitzades a estudiants escollits de manera aleatòria en el carrer, i la introducció de les dades al suport informàtic (a través d'un formulari de Google). Un cop disponibles les dades dels 641 enquestats en un full de càlcul Microsoft Office Excel, va ser necessària la depuració i revisió de la base de dades, explicada en el següent apartat.

A continuació s'explicaran també la metodologia emprada i es presentaran els resultats obtinguts en l'anàlisi, diferenciant cadascuna de les capacitats, i adjuntant taules i gràfics. Finalment, tindrà lloc la redacció de les conclusions més rellevants a les que s'arribaran.

2. Depuració de la base de dades.

Per començar, s'eliminaren de la base de dades les variables referides a l'agudesa visual i a la memòria (habilitats mesurades en l'enquesta i disponible a la base de dades però que en aquest anàlisi no s'estudiaran), així com també les variables *Marca temporal* i *Identificador de grup* i els registres de l'Àmbit de ciències, per no ser necessaris, deixant la base amb 473 registres i 15 variables.

A continuació, es va procedir a canviar el nom de les variables: les quatre variables referides a la capacitat física seran A1, A2, A3 i A4, mentre que les de càlcul seran B1, B2, B3 i B4. També fou necessària la recodificació d'algunes variables. La variable A4 (amb quina freqüència fas esport?) podia prendre els valors "Cap cop", "1 o 2 dies", "3 o 4 dies" i "5 o més dies", els quals es substituïren per 0, 1, 2 i 3, respectivament. D'altra banda, la variable B2 (Quin nombre falta en les següents sèries numèriques?) recollia múltiples formats: text, nombre real, una sèrie de nombres, etcètera; per això, es va decidir tenir en compte el nombre de sèries completades correctament (nombre real entre 0 i 4), i tots els valors impossibles (hi havia valors que no es podien interpretar, com per exemple Si/No o nombres al voltant del centenar) s'eliminaren.

Finalment, per a la variable que recull el nombre de salts a la corda que es capaç de fer l'enquestat en un minut, es va trobar que cinc dades prenia valors molt extrems (1,4,7,7, i 9) tractant-se d'una variable amb una mitjana de 64,8 salts, però possibles al cap i a la fi.

3. Mètodes emprats.

Primerament, amb les variables numèriques contínues, s'ha emprat la prova chi-quadrat de Pearson per comprovar la normalitat de les dades. com que aquestes no seguien la distribució normal en cap cas, s'ha procedit a realitzar la prova de Mann-Whitney-Wilcoxon per a dues mostres independents per tal de comprovar si provenien d'una mateixa distribució (mateixa mediana).

Quant a les variables discretes, s'ha realitzat la mateixa prova (test chi-quadrat de Pearson) per validar la distribució que seguien. També s'han portat a terme tests F de Fisher per a la comparació de variàncies, i proves t de Student per a la comparació de mitjanes entre les dues poblacions: salut i ciències socials.

Finalment, amb les variables categòriques, s'han elaborat taules de contingència i realitzat també el test chi-quadrat per comprovar si les proporcions de les dades estaven relacionades.

4. Resultats.

A. Capacitat física.

1. Temps que triga el cor en fer 20 pulsacions.

Es realitzarà una comparació de variàncies. Les hipòtesis platejades són:

H_o: Les variàncies són iguals.

H_1: Les variàncies són diferents.

Es realitza un test F de Fisher, amb un nivell de significació del 5%.

```
        F test to compare two variances

data:  A1 by Àmbit
F = 0.673, num df = 256, denom df = 215, p-value = 0.002404
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5196978 0.8689857
sample estimates:
ratio of variances
         0.6729989
```

Arribem a la conclusió de que com que l'1 no està a dins de l' interval [05196, 0.8689], rebutgem la hipòtesi nul·la, per tant les variàncies són diferents.

A partir d'aquesta informació realitzem un test t de mitjanes de mostres independents i variàncies diferents.

H_0: Les mitjanes són iguals.

H_1: Les mitjanes són diferents.

```
        Welch Two Sample t-test

data:  A1 by Àmbit
t = 1.5032, df = 415.39, p-value = 0.1336
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.139609  1.047068
sample estimates:
mean in group Salut mean in group Social i Humanitats
          17.78938                          17.33565
```

S'obté amb un nivell de confiança del 95% l' interval [-0.139609, 1.047068], en el qual podem trobar el valor 0, en conseqüència acceptem la hipòtesi nul·la, és a dir, les mitjanes són iguals.

2. Salts a la corda.

Es planteja si la mitjana de salts que fan en un minut a la corda és la mateixa en ambdós grups. Però, abans de poder realitzar el test de diferència de les mitjanes mitjançant una t de Student s'ha de conèixer si podem suposar que les variàncies dels dos camps són iguals. Plantegem la hipòtesi nul·la i l'alternativa:

$H_{0:}$ Les variàncies són iguals
$H_{1:}$ Les variàncies són diferents

Es realitza un test F de Fisher amb R per comparar les dues variàncies amb un nivell de significació del 5%

```
        F test to compare two variances

data:  A2 by Àmbit
F = 1.0497, num df = 256, denom df = 215, p-value = 0.7142
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8105865 1.3553802
sample estimates:
ratio of variances
          1.049694
```

Com que el *p-value* és mes gran que el nivell de significació (0.7142 > 0.05), acceptem la hipòtesi nul·la i s'accepta que les variàncies són iguals.

A partir d'aquesta informació es procedeix a realitzar un test t de mitjanes de mostres independents i variàncies iguals. D'aquesta forma també es trobarà l' interval que ens demana, ja que es quasi bé el mateix procediment.

$H_{0:}$ Les mitjanes són iguals
$H_{1:}$ Les mitjanes són diferents

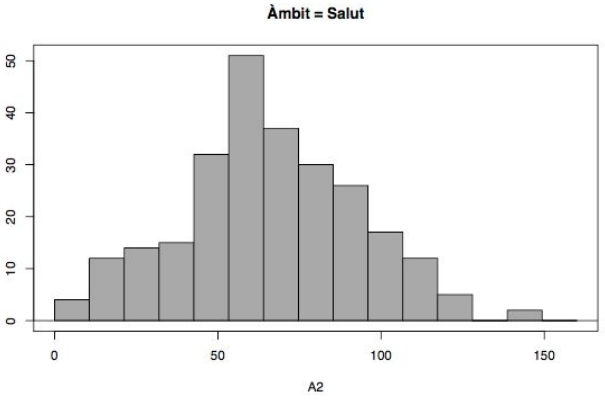
Es realitza un Test T de variables independents amb R, amb un nivell de significació del 5% i considerant les variàncies iguals :

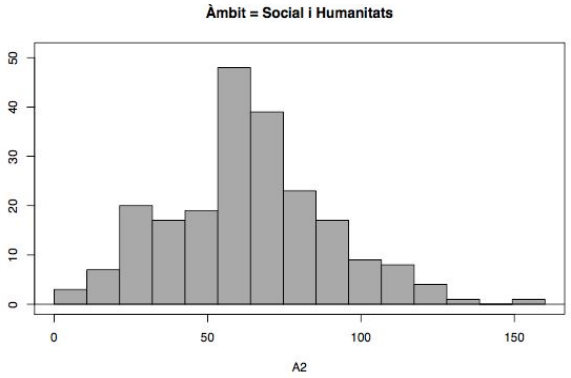
```
        Two Sample t-test

data:  A2 by Àmbit
t = 1.1149, df = 471, p-value = 0.2655
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.071260  7.504283
sample estimates:
              mean in group Salut mean in group Social i Humanitats
                        66.05447                           63.33796
```

Un cop introduïdes les dades d'aquestes dues mostres a l'R, s'obté amb un nivell de confiança del 95% l' interval [-2.071, 7.504], en el qual podem trobar el valor zero a dins, per tant podem acceptar la hipòtesi nul·la i podrà dir que les mitjanes són iguals. Es pot arribar a

la mateixa conclusió mirant el *p-value*, el qual és superior que el nivell de significació, per tant s'accepta H_0.

Per completar l'estudi d'aquest variable, s'afageixen dos histogrames,un per cada àmbit, a partir dels quals es poden observar gràficament els resultats obtinguts: les mitjanes són iguals.





3. Salts al peu coix.

Per aquesta qüestió, una variable dicotòmica, es fa una prova no paramètrica d'independència basada en la chi-quadrat. Primer es tabula en una taula de contingència la quantitat de persones de cada grup que sí recordava la successió i la quantitat de gent que no la recordava.

Es defineix com a hipòtesi nul·la que les dues proporcions són iguals, i com a hipòtesi alternativa que són diferents. S'està fent doncs, una prova d'independència per dues variables amb dues poblacions.

Es planteja la hipòtesi nul·la i l'alternativa de les proporcions:

H_0: Les proporcions són iguals

H_1: Les proporcions són diferents

La taula dels valors observats és la següent:

	SI	**NO**	**TOTAL**
Salut	226	31	257
Socials i humanitats	204	12	216
TOTAL	430	43	**473**

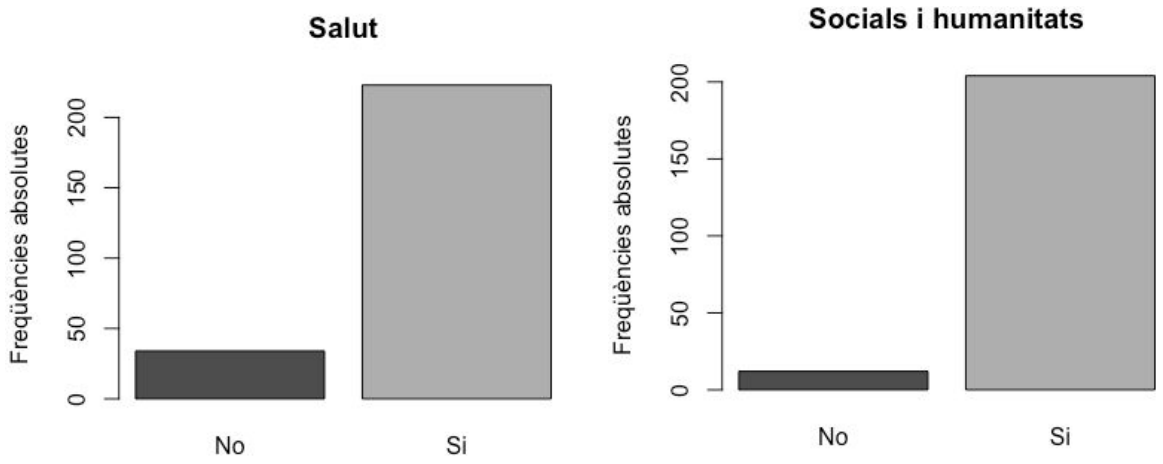
Ja des d'un primer moment es pot veure que els valors esperats i els valors observats no són gaire semblants per tant, ens porta a pensar que es rebutjarà la hipòtesi nul·la. Però per confirmar-ho es fa un test de la khi-quadrat amb R amb un nivell de significació del 5%.

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 6.0122, df = 1, p-value = 0.01421
```

El *p-value* obtingut dóna un valor més petit que el nivell de significació (0.01421 > 0.05). Per tant, estàvem en lo cert, i amb aquest resultat podem concloure que es rebutja la hipòtesi nul·la.

Els estudiants d'aquests dos àmbit presenten diferències rellevants en la resposta d'aquesta pregunta. Per tant es pot concloure que les dues variables: "ser d'un àmbit determinat" i "capaços de saltar a peu coix" tenen relació.

Gràficament, es pot veure això amb les freqüències d'estudiants que han respost correctament i incorrectament a la pregunta: s'observen més estudiants a l'àmbit de la salut que no han pogut saltar al peu coix 10 passes avançades.



4. Freqüència amb la que es fa esport.

Es tracta d'una variable categòrica ordinal amb quatre possibles valors. Volem establir si hi ha relació entre l'àmbit d'estudi (A) i la freqüència amb que es realitza alguna activitat física (B). Amb una mostra de 473 individus la taula de contingència resultant és:

Àmbit \ Freqüència	Cap cop	Un o dos cops	Tres o quatre cops	Més de cinc cops	TOTAL
Salut	55	104	78	20	257
Socials i humanitats	61	78	61	16	216
TOTAL	116	182	139	36	**473**

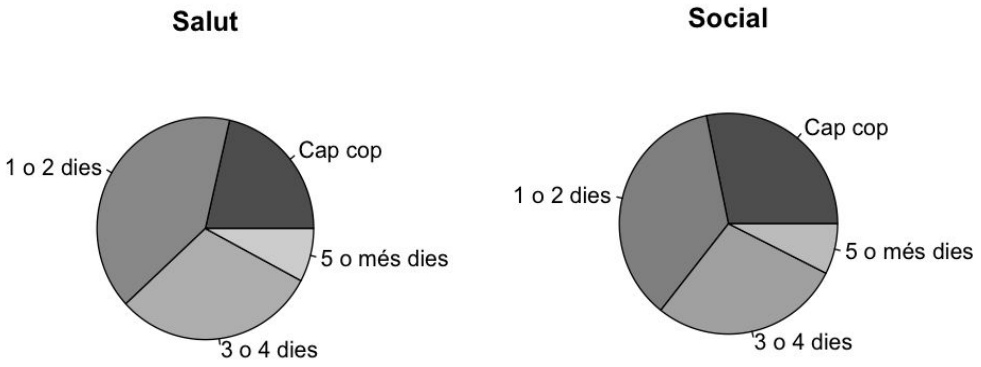
La prova d'hipòtesi en aquest cas és:
H_o: L'àmbit d'estudi i la freqüència amb que fan esport no estan relacionats, és a dir, A i B són independents.
H_1: L'àmbit d'estudi i la freqüència amb que fan esport estan relacionats, o sigui que A i B no són independents.

S'ha realitzat el test χ^2 de Pearson amb un nivell de significació del 5% i el resultat ha sigut

```
        Pearson's Chi-squared test
data:  .Table
X-squared = 3.017, df = 3, p-value = 0.389
```

Com que el valor observat χ^2 = 3.017 és menor que l'estadístic de Pearson χ_3^2 = 7.82, no rebutgem H_o i concluïm que, amb una significació del 5%, no hi ha cap relació entre l'àmbit d'estudi d'un estudiant, i la freqüència amb que realitza algun esport.

Es completa l'estudi d'aquesta variable amb un diagrama de sectors per a cada àmbit; en ells s'observen proporcions molt semblants en ambdós casos.



B. Capacitat de càlcul.

1. Piràmide de sumes.

Per aquesta qüestió, una variable dicotòmica, es fa una prova no paramètrica d'independència basada en la chi-quadrat. Primer es tabula en una taula de contingència la quantitat de persones de cada grup que sí recordava la successió i la quantitat de gent que no la recordava.

Es defineix com a hipòtesi nul·la que les dues proporcions són iguals, i com a hipòtesi alternativa que són diferents. . S'està fent doncs, una prova d'independència per dues variables amb dues poblacions.

Es planteja la hipòtesi nul·la i l'alternativa de les proporcions:

H_0: Les proporcions són iguals
H_1: Les proporcions són diferents

Les taula amb els valors observats és la següent:

	SI	**NO**	**TOTAL**
Salut	161	96	257
Socials i humanitats	152	64	216
TOTAL	313	160	**473**

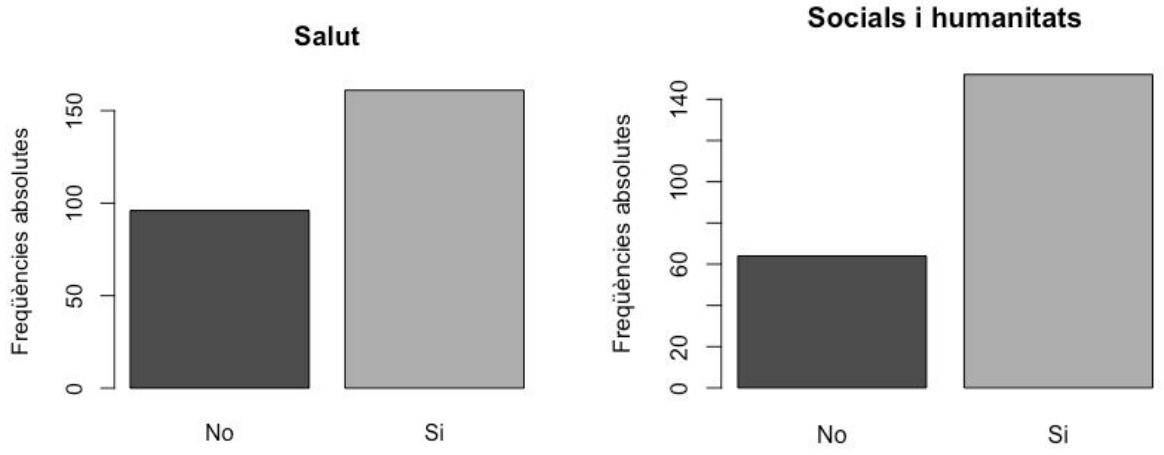
Ja des d'un primer moment es pot veure que els valors esperats i els valors observats no són gaire semblants per tant, ens porta a pensar que es rebutjarà la hipòtesi nul·la. Però per confirmar-ho es fa un test de la chi-quadrat amb R amb un nivell de significació del 5%.

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 3.1284, df = 1, p-value = 0.07694
```

El *p-value* obtingut dóna un valor més gran que el nivell de significació (0.07694 > 0.05). Per tant, no estàvem en lo cert, i amb aquest resultat podem concloure que no es rebutja la hipòtesi nul·la.

Els estudiants d'aquests dos àmbit no presenten diferències rellevants en la resposta d'aquesta pregunta. Per tant es pot concloure que les dues variables: "ser d'un àmbit determinat" i "completar la piràmide" no tenen cap relació.

Gràficament, es pot veure això amb els percentatges que representen els estudiants que han respòs correctament i incorrectament a la pregunta.



2. Sèries numèriques.

Per a aquesta variable quantitativa discreta, es desitja estudiar si la mitjana de nombres que falten a les series numèriques que fan son la mateixa en ambdós grups. Per aquest motiu, es vol realitzar el test de diferència de les mitjanes mitjançant una t de Student, però abans s'ha de conèixer si podem suposar que les variàncies dels dos camps són iguals. Plantegem la hipòtesi nul·la i l'alternativa:

$H_{0:}$ Les variàncies són iguals
$H_{1:}$ Les variàncies són diferents

Es realitza un test F de Fisher amb R per comparar les dues variàncies amb un nivell de significació del 5%

```
        F test to compare two variances

data:  B2 by Àmbit
F = 0.54337, num df = 225, denom df = 196, p-value = 1.038e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4136178 0.7119938
sample estimates:
ratio of variances
         0.5433709
```

Com que el *p-value* és mes petit que el nivell de significació (1.036e-05 > 0.05), rebutjem la hipòtesi nul·la i s'accepta que les variàncies són diferentes.

A partir d'aquesta informació es procedeix a realitzar un test t de mitjanes de mostres independents i variàncies iguals. D'aquesta forma també es trobarà l' interval que ens demana, ja que es quasi bé el mateix procediment.

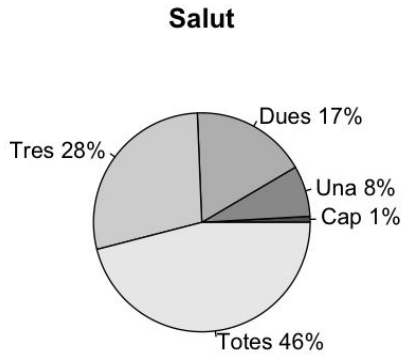
H_0: Les mitjanes són iguals
H_1: Les mitjanes són diferents

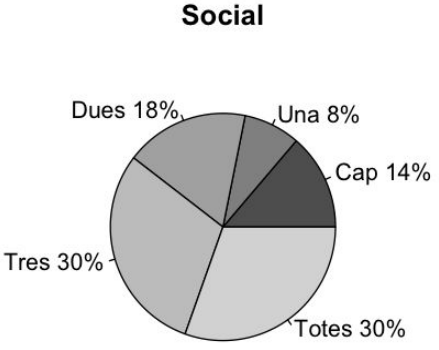
s realitza un Test T de variables independents amb R, amb un nivell de significació del 5% i considerant diferents les variàncies :

```
        Welch Two Sample t-test

data:  B2 by Àmbit
t = 4.737, df = 356.06, p-value = 3.139e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3259376 0.7887024
sample estimates:
             mean in group Salut mean in group Social i Humanitats
                        3.110619                           2.553299
```

Un cop introduïdes les dades d'aquestes dues mostres a l'R, s'obté amb un nivell de confiança del 95% l' interval [0.3259376, 0.7887024], en el qual no podem trobar el valor zero a dins, per tant hem de rebujtar la hipòtesi nul·la i podrà dir que les mitjanes són diferents. Es pot arribar a la mateixa conclusió mirant el *p-value*, el qual és inferior que el nivell de significació, per tant es rebutja H_0.

Ara, s'adjunten dos diagrames de sectors amb els percentatges, mitjançant els quals es pot observar com els enquestats de l'àmbit de la salut han obtingut millors resultats que els de socials i humanitats: un 74% dels de salut encertaren 3 o 4 sèries numèriques davant un 60% dels de socials i humanitats; de la mateixa manera, només un 1% de salut no saberen completar cap sèrie enfront un 14% per part de l'àmbit de les ciències socials.





3. Sèrie d'operacions fins esbrinar el nombre final.

A partir de les mesures del temps que van trigar els enquestats a realitzar un seguit d'operacions, es vol validar si les dades provenen d'una llei Normal(μ,σ^2) amb μ i σ^2 desconeguts mitjançant la prova chi-quadrat de Pearson. La prova d'hipòtesi i els resultats obtinguts, suposant 5 intervals, són els següent:

H_o: La variable es distribueix com una N(μ,σ^2).
H_1: La variable no es distribueix com una N(μ,σ^2).ç

Pearson chi-square normality test

data: B3
P = 22.275, p-value = 1.456e-05

Amb un nivell de significació $\boldsymbol{\alpha}$ = 0.05, la mesura de discrepància $\boldsymbol{\chi}_2^2$ és igual a 5.99. Com que hem obtingut $\boldsymbol{\chi}^2$ = 22.275, rebutgem la hipòtesi nul·la.

A continuació es procedeix a realitzar la comparació del temps que van trigar els estudiants de salut (A) i de ciències socials i humanitats (B). Com que són dues mostres clarament independents i no normals, es realitza la prova de Mann-Whitney-Wilcoxon. Plantegem la següent prova d'hipòtesi:

H_0: La mediana de temps d'A és igual a la mediana de temps de B.
H_1: La mediana de temps d'A és diferent a la mediana de temps de B.

Wilcoxon rank sum test with continuity correction

data: B3 by Àmbit
W = 30102, p-value = 0.1069
alternative hypothesis: true location shift is not equal to 0

Com que el p valor és superior a 0.05, no rebutgem H_o i afirmem que el temps que van trigar els enquestats en fer les operacions proposades és independent de l'àmbit dels estudis que realitzen.

5. Conclusions.

Un cop analitzades les qüestions assignades, ses pot afirmar, de manera general, que no hi ha gaire diferències entre la capacitat física i la de càlcul dels estudiants universitaris dels àmbits de la salut i les ciències socials i humanitats.

En el cas del la capacitat física, més concretament, els resultats dels tests han sigut que les mitjanes de les dues mostres són iguals; en la prova de proporcions, que no hi existeix cap relació entre les variables estudiades i l'àmbit d'estudi; de la mateixa manera que l'histograma i el diagrama de sectors han sigut molt semblants per als dos àmbits. Només en la prova de saltar al peu coix s'han pogut observar algunes petites diferències si s'observa el diagrama de barres: hi ha més estudiants incapaços de superar la prova en el cas de les ciències de la salut.

Pel que fa a la capacitat de càlcul, el resultat del test de Wilcoxon per a la pregunta 3 ha sigut que no hi ha hagut diferències entre els àmbits. D'altra banda, en l'exercici 1 (malgrat el test ens permeti afirmar que les proporcions són iguals) s'observa que hi ha hagut més estudiants de l'àmbit de la salut que no han sabut completar la piràmide. Tot i així, en l'exercici 2, les mitjanes han resultat ser diferents i amb el diagrama de sector s'ha observat que els estudiants de ciències socials han tingut una menor capacitat de càlcul; per la qual cosa, no sembla clar que els estudiants d'un àmbit tinguin major capacitat que els de l'altre.