DESCRIPCIONS PROBABILÍSTIQUES I ESTADÍSTIQUES:
SEGONA ENTREGA: VARIABLES ALEATÒRIES

Guillem Garcia Dausà (NIU: 1636279)
Martí Llinés Viñals (NIU: 1637804)

Grau en Enginyeria de Dades (UAB)
Data d'entrega: 21/11/2022

Índex de continguts

Índex de taules

Índex de diagrames

1. PRIMER EXERCICI

1.1. APARTAT A

Si obrim el fitxer “enquesta.csv” i filtrem les dades per veure el nombre de fumadors, entre el total de les persones que s’han enquestat (120 persones), 44 d’aquestes fumen. Així, podem veure que la probabilitat de triar una persona a l’atzar que fumi, es calcula a partir de la definició de probabilitat:

$$p(X)=\frac{Casos\ favorables}{Casos\ possibles} \rightarrow p(fumador)=\frac{44}{120}=0.366=36.6\,\%$$

Fórmula 1: Probabilitat de triar una persona aleatòria que fuma

La probabilitat de triar una persona fumadora és del 36,6% aproximadament.

Aquesta probabilitat és més alta que de la gent que fa esport com a oci, ja que veient amb el mateix procediment les persones que realitzen esport (atribut “OCI” igual a 4), 30 són el total de persones de la mostra que fan esport.

$$p(esport)=\frac{30}{120}=25\,\%$$

Fórmula 2: Probabilitat de triar una persona que fa esport

Es veu després de realitzar el càlcul que aquesta probabilitat és menor a l’anterior calculada.

1.2. APARTAT B

Primerament, el que hem fet és veure el nombre de dones en què es composa la nostra base de dades, que són en total 57. Sobre aquestes dones, veiem quantes en tenen una altura més alta de 160 cm. Aquest càlcul dóna 43 dones (tenint en compte que és estrictament més gran que 160 cm). Per tant podem realitzar els següents càlculs:

$$p(dona \wedge altura>160)=\frac{43}{57}=75.44\,\%$$

Fórmula 3: Probabilitat de triar una dona amb altura més gran a 160cm

Es pot veure a partir de la fórmula que només s’ha tingut en compte el nombre de dones de la mostra perquè ja sabem de base que s’ha triat una dona a l’atzar. Per tant, hi ha un 74,44% de triar una dona a l’atzar que sigui més altra que 160 cm.

Amb la mateixa premissa de triar una dona a l'atzar, al veure la unió de les dones que tenen com oci l'ordinador o com oci la televisió, veiem que dona un total de 24. Per tant la probabilitat serà de:

$$p(dona \wedge (oci = ordinador \vee oci = televisió)) = \frac{24}{57} = 42.10\,\%$$

Fórmula 4: Probabilitat de triar una dona a l'atzar i que tingui com oci l'ordinador o la televisió

1.3. APARTAT C

Inicialment calculem el nombre d'homes que han contestat a l'enquesta (que són 63 homes). A continuació filtrem els homes que fumen (atribut "TABAC" igual a 1), que dóna com a resultat 23. Aleshores:

$$p(home \wedge fumador) = \frac{23}{63} = 36.51\,\%$$

Fórmula 5: Probabilitat de triar un home a l'atzar i que sigui fumador

L'explicació és la mateixa que per l'apartat anterior, però aquest cop s'han tingut només en compte els homes de la mostra per veure quants d'ells fuma.

En l'apartat "a" ja s'havia mostrat les persones que fumen (que són 44), d'aquestes 44, veurem quants són homes i a la vegada fan esport. El resultat donat és de 6 homes que fumen i fan esport, al que correspon a:

$$p(home \wedge fumar \wedge oci = esport) = \frac{6}{44} = 13{,}63\,\%$$

Fórmula 6: Probabilitat de triar un fumador a l'atzar que sigui home i esportista

1.4. APARTAT D

De totes les persones enquestades, n'hi ha més persones que es dediquen el seu temps lliure a l'ordinador o a la música/lectura que persones que fan esport i no fumen. En el primer cas hem vist que hi ha 22 persones que fan esport i no fumen, i en el segon cas n'hi ha 27 persones que es dediquen el seu temps lliure a l'ordinador o a la música/lectura (5 més que l'anterior).

1.5. APARTAT E

És més probable agafar una persona que pesi més de 60 kilograms i tingui almenys 20 anys, ja que aquestes carecterístiques les compleixen 72 persones (que correspon al 60% de tota la mostra), i en canvi en l'altre supòsit les compleix 65 persones (54,16%).

2. SEGON EXERCICI

2.1. APARTAT A

2.2. APARTAT B

2.3. APARTAT C

2.4. APARTAT D

2.5. APARTAT E

3. TERCER EXERCICI

3.1. APARTAT A

A continuació es veurà la funció de densitat de la funció proporcionada per l'enunciat:

$$f(x)=\frac{1}{4}(x-1)^3, 1<x<3$$

Funció 1: Funció de densitat a treballar

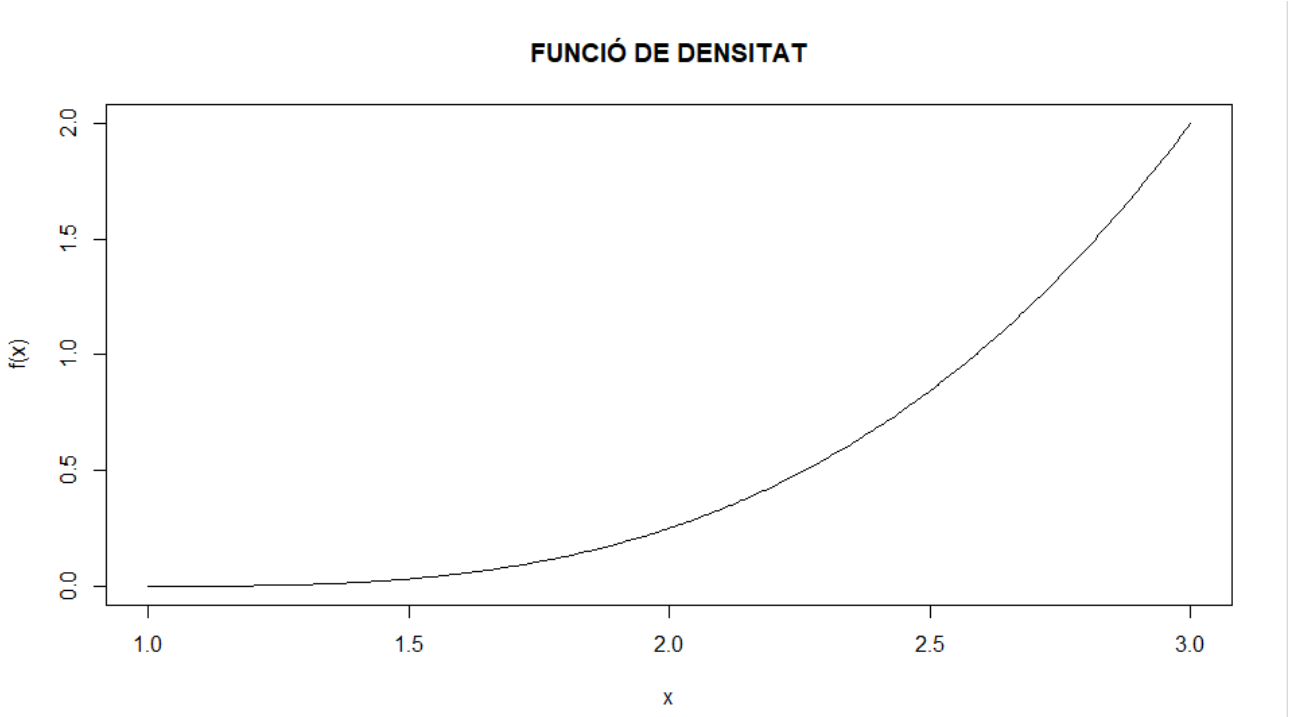


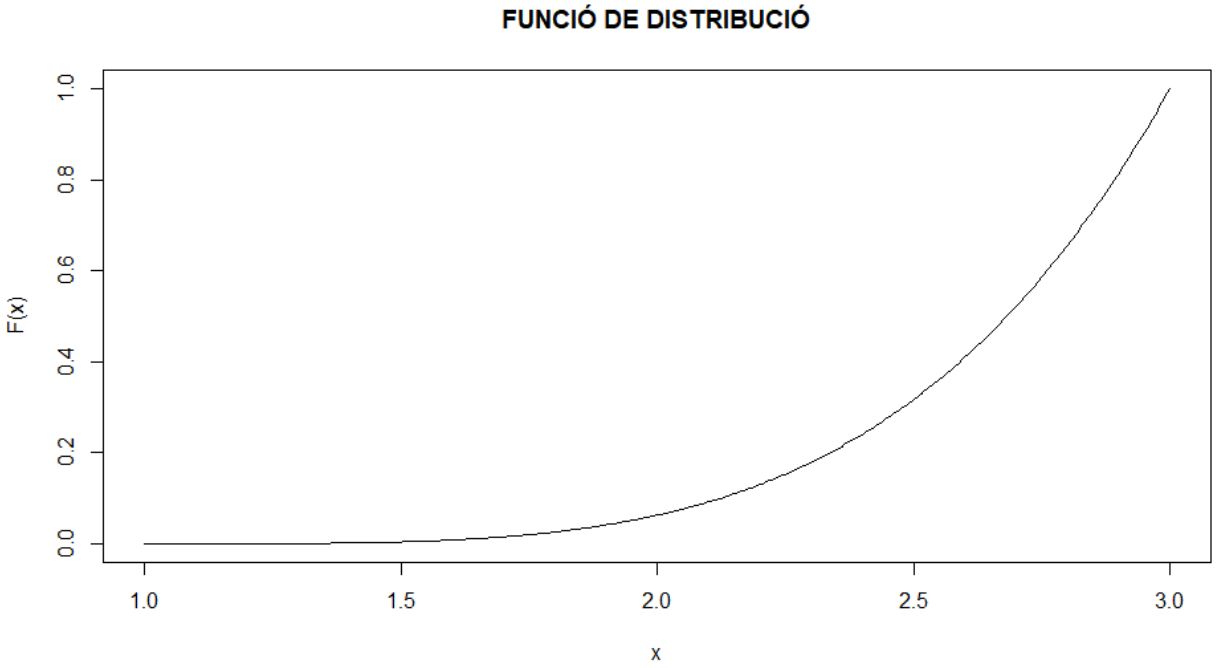
Diagrama 1: Funció de densitat de f(x)

La funció de distrubució s'ha calculat a partir d'integrar la funció de densitat definit en el punt "x= 1" i "x= 3". Aleshores, aquesta és la funció que queda després d'integrar:

$$F(x)=\frac{1}{16}(x-1)^4$$

Funció 2: Funció de distribució calculada a partir de l'integral de la densitat

Si representem aquesta funció com hem fet amb la densitat anteriorment, hauria de donar una representació gràfica d'aquesta manera:



Funció 3: Funció de distribució de f(x)

Veiem que aquesta corba generada és molt similar a la funció de densitat, i també podem comprovar que si avaluem del punt "x=1" al punt "x=3" de l'integral, l'àrea dintre de la corba és 1 (que és una manera de justificar que la representació és correcte).

3.2. APARTAT B

Per fer la representació del histograma amb el dibuix de la funció de densitat, primerament hem de generar els 80.000 valors uniformes aleatoris (comanda R "runif"), i calcular seguidament l'inversa de la funció de distribució. Aquest càlcul ens ha generat la funció següent:

$$F^{-1}(x)=16\,x^{\left(\frac{1}{4}\right)}+1$$

Funció 4: Funció de distribució inversa

Seguidament, amb aquesta funció dibuixem l'histograma i representem per sobre la funció de densitat proporcionada de l'enunciat. Per tant, així és com queda l'histograma finalment:

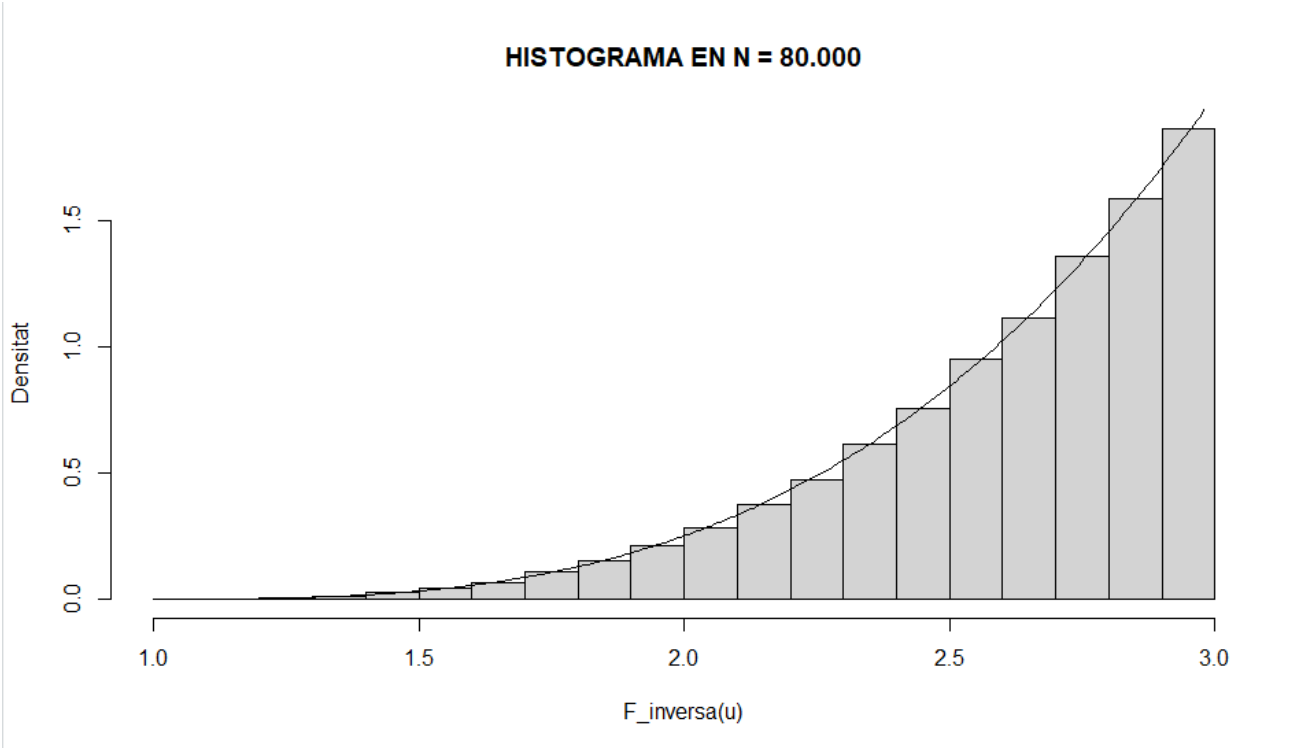


Diagrama 2: Histograma de la funció de distribució inversa

On el paràmetre "u" del diagrama indica que és cada valor de les 80.000 simulacions generades, i "F_inversa(u)" significa el resultat que dona de la funció de distribució inversa en funció d'aquesta variable "u". La densitat representa la freqüència dels resultats obtinguts per la funció de distribució inversa.

3.3. APARTAT C

La fórmula que hem usat per calcular l'esperança i la variància són les definides per les variables aleatòries en cas continu:

$$E(X)=\mu=\int_{(-\infty,\infty)} x f(x)\,dx.$$

Fórmula 7: Esperança d'una variable aleatòria contínua

$$Var(X)=\sigma^2=E(X^2)-(E(X))^2$$

Fórmula 8: Variància d'una variable aleatòria contínua

Hem desenvolupat les fórmules i aquests són els resultats teòrics:

$$E(x)=\int_{(1,3)}\left(x*\left(\frac{1}{4}\right)*(x-1)^3\right)=\frac{13}{5}=2.6$$

$$E(x^2)=\int_{(1,3)}\left(x^2*\left(\frac{1}{4}\right)*(x-1)^3\right)=\frac{103}{15}$$

$$var(x) = \frac{103}{15} - \left(\frac{13}{5}\right)^2 = 0.106$$

Empíricament, si calculem la mitjana i la variància (no corregida), veiem que ens dona aproximadament:

$$mean(x) = 2.599$$

$$var(x) = 0.1068$$

Veiem que els resultats són gairebé idèntics, i això ho trobem lògic perquè hem agafat una mostra bastant gran (n = 80.000) per tal de que s'assembli els resultats de la mostra a la funció de densitat original.

3.4. APARTAT D

Primerament i abans de desenvolupar l'exercici, el que hem fet és representar una distribució normal (N~(0,1)) amb una mostra n=80.000 per tal de veure com hauria de ser aproximadament el nostre resultat final aplicant el teorema central del límit.

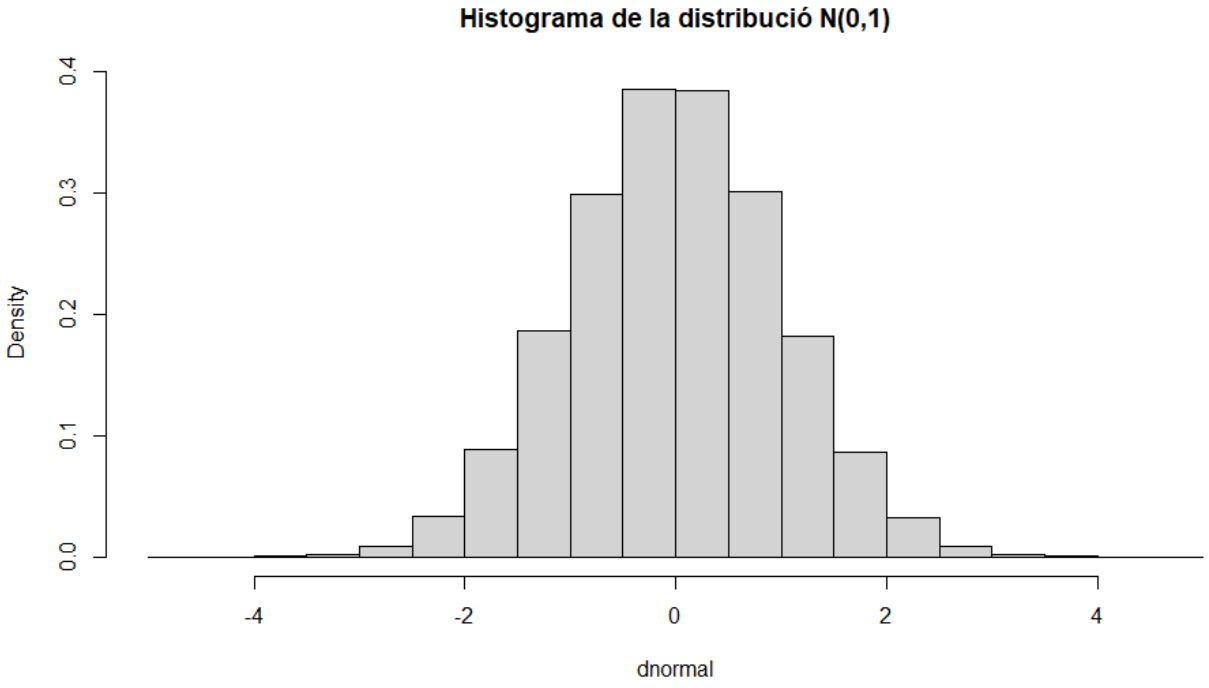


Diagrama 3: Distribució normal amb μ= 0 i σ = 1

Seguidament, hem aplicat el teorema central del límit, declarant els següents paràmetres:

Paràmetre	Valor
n (nombre de columnes matriu)	200

μ (Mitjana calculada empíricament)	2.599
σ (Arrel variància empírica)	0.3268

També s'ha declarat una matriu que conté els valors de la mostra amb un número de files igual a 400 i 200 columnes. El que es farà amb aquesta matriu serà fer la mitjana de cada fila per guardar-la en un vector i calcular el teorema central del límit amb cada valor d'aquest vector (correspon en el teorema com "Xn"). El teorema central del límit defineix el següent:

$$\frac{mean(Xn)-\mu}{\frac{\sigma}{\sqrt{(n)}}} \rightarrow N(0,1)$$

Fórmula 9: Teorema central del límit

Al calcular el teorema per cada valor del vector (on ho hem guardat dintre d'una variable anomenada "resultat"), hem representat el següent histograma:

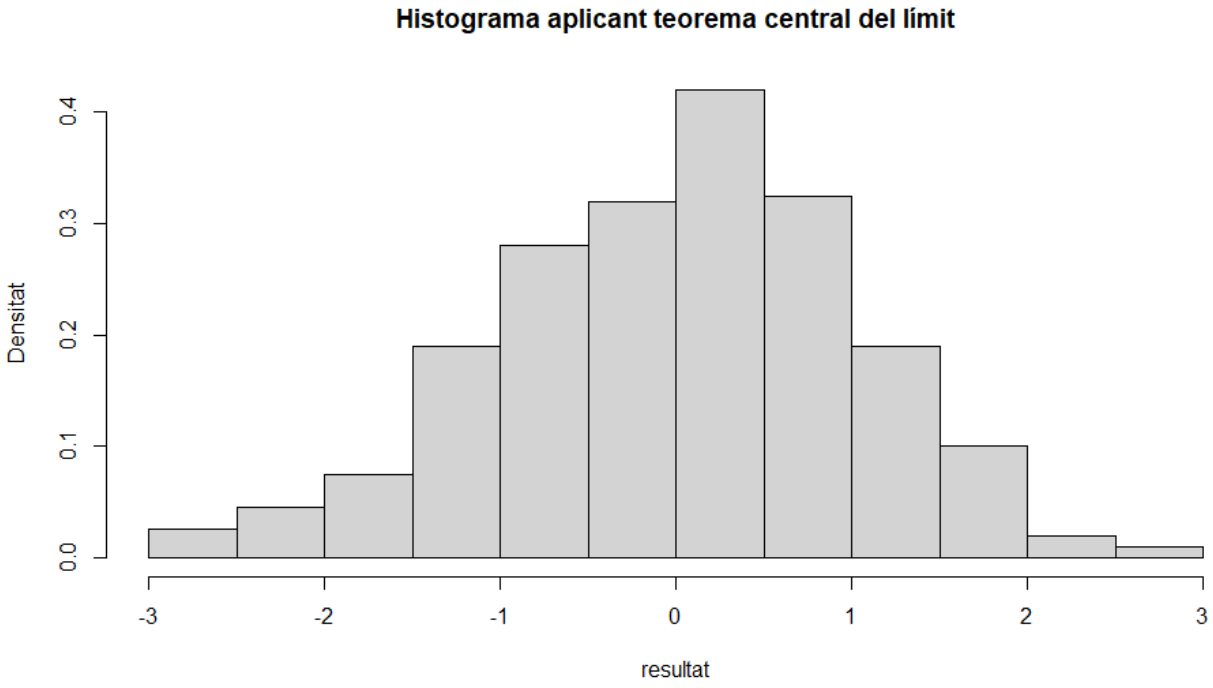


Diagrama 4: Histograma aplicant el teorema central del límit

Es pot veure gràficament que s'aproxima bastant a una distribució normal (N~(0,1)). També si fem certes comprovacions, veiem que l'esperança està a prop de 0 i la sigma també s'aproxima a 1(tal com hauria de ser).

3.5. APARTAT E

A partir de la funció quantila podem calcular la variable t que demana l'enunciat per tal de P(X<t) = 0.8. L'instrucció utilitzada és la següent:

```
#VEIEM QUIN ÉS EL VALOR T A PARTIR DE LA FUNCIÓ QUANTILA
valor <- qnorm(0.8, mitjana_empirica, varp_empirica)
valor
```

Comanda 1: Funció quantila per calcular la variable t

Aquesta variable ens ha donat aproximadament **2.692.** El que significa aquest resultat és que hi ha un 80% de probabilitats de que al generar una mostra aleatòria, el resultat de la funció és menor a 2.692.

Per veure el nombre de mostres empíriques que ens ha donat menor al resultat anterior, hem fet un bucle mirant cada resultat de la mostra i comptant quants és estrictament més petit que "t".

```
#VEIEM QUIN ÉS EL TOTAL DE MOSTRES QUE SON MENOR QUE EL VALOR
sumatori <- 0
for(i in seq_along(y)){
  if (y[[i]] < valor){
    sumatori <- sumatori + 1}
}
sumatori
```

Comanda 2: Iteració per veure el nombre de mostres que és més petit a t

Aquesta variable sumatori ens ha donat **40712** (una mica més de la meitat que el total de la mostra). En principi aquest resultat no ens el esperàvem, ja que imaginàvem que seria el 80% de les 80.000, o sigui més o menys 64.000 mostres. Però, si veiem l'histograma de l'apartat "b", veiem que la gran majoria de les mostres té com a resultat entre 2.5 i 3.

4. QUART EXERCICI

4.1. APARTAT A

4.2. APARTAT B

4.3. APARTAT C

4.4. APARTAT D

4.5. APARTAT E