A2 - Anàlitica descriptiva i inferencial

Xavier Vizcaino Gascon

6 de abril, 2022

Contents

1. Lectura del fitxer i preparació de les dades

Llegiu el fitxer CensusIncome_clean.csv i guardeu les dades en un objecte amb identificador denominat cens. A continuació, verifiqueu que les dades s'han carregat correctament.

Carreguem el fitxer de dades amb la següent comanda i generem un *dataset* que anomenarem **cens**:

```
cens <- read.csv("CensusIncome_clean.csv", stringsAsFactors=TRUE)
```

S'utilitza *read.csv* ja que el separador és la coma (,). També s'utilitza la opció *stringsAsFactors=TRUE*, doncs permet tenir una primera visió dels *strings* repetits en diferents registres.

Examinem el tipus de dades amb que R ha interpretat cada variable, per fer-ho apliquem a través de *sapply()* la funció *class()* en tot el *dataset*.

```
sapply(cens,class)
```

```
##          CS_ID            age      workclass  education_num marital_status
##       "factor"      "integer"       "factor"      "integer"       "factor"
##   relationship     occupation           race         gender hours_per_week
##       "factor"       "factor"       "factor"       "factor"      "numeric"
##         income  education_cat
##      "numeric"       "factor"
```

A continuació examinem els valors resum, de cada tipus de variable amb la funció *summary()* aplicada a tot el *dataset*:

```
summary(cens)
```

```
##      CS_ID            age                workclass     education_num
##  CS1    :    1   Min.   :17.00   Government   : 4349   Min.   : 1.00
##  CS10   :    1   1st Qu.:28.00   Other/Unknown: 1855   1st Qu.: 9.00
##  CS100  :    1   Median :37.00   Private      :22692   Median :10.00
##  CS1000 :    1   Mean   :38.55   Self-Employed: 3657   Mean   :10.08
##  CS10000:    1   3rd Qu.:48.00                         3rd Qu.:12.00
##  CS10001:    1   Max.   :80.00                         Max.   :16.00
##  (Other):32547
##  marital_status         relationship          occupation
##  D: 4442        Husband       :13192   Blue-Collar  :10060
##  M:15413        Not-in-family : 8303   Other/Unknown: 1850
##  S:10680        Other-relative:  981   Professional : 4139
##  W:  993        Own-child     : 5067   Sales        : 3649
##  X: 1025        Unmarried     : 3444   Service      : 5021
##                 Wife          : 1566   White-Collar : 7834
##
##                  race       gender    hours_per_week      income
##  Amer-Indian-Eskimo:  311   f:10767   Min.   : 1.00   Min.   : 0.10
##  Asian-Pac-Islander: 1039   m:21786   1st Qu.:40.00   1st Qu.:43.22
##  Black             : 3123             Median :40.50   Median :49.71
##  Other             :  271             Mean   :40.31   Mean   :48.75
##  White             :27809             3rd Qu.:45.00   3rd Qu.:54.32
##                                       Max.   :80.00   Max.   :68.37
##
##            education_cat
##  Postuniversitaria: 2712
##  Primaria         : 2644
##  Secundaria       :12106
##  Universitaria    :15091
##
##
##
```

Observem que només la variable *CS_ID* té un nombre elevat de *levels*, indicatiu que aquest tipus no és el més adequat i hauriem de canviar el tipus de variable a *char*.

```
#Modificació
cens$CS_ID<-as.character(cens$CS_ID)

#Comprovació
class(cens$CS_ID)
```

```
## [1] "character"
```

```
summary(cens$CS_ID)
```

```
##    Length     Class      Mode
##     32553 character character
```

2. Edat

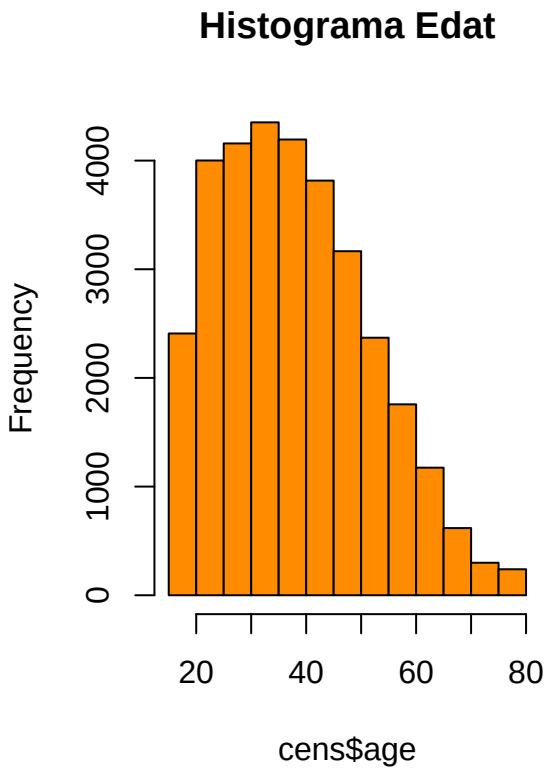
Calculeu l'interval de confiança de la mitjana d'edat. Seguiu els passos que s'especifiquen a continuació.

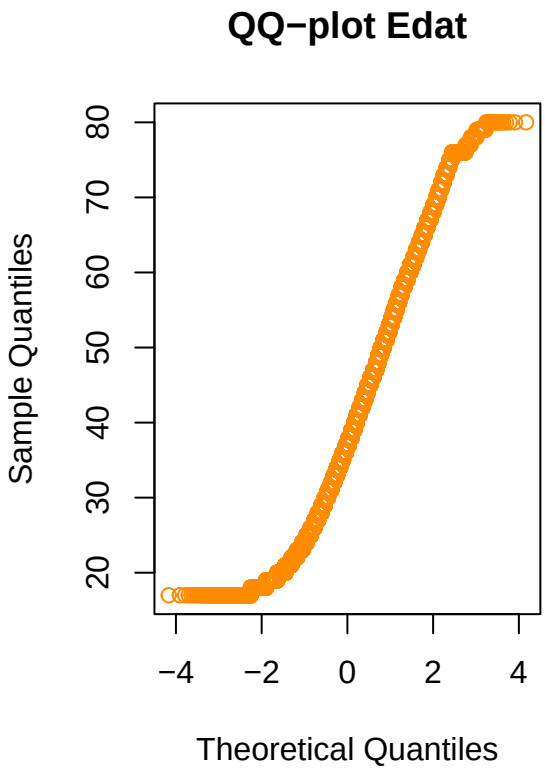
2.1. Distribució d'edats

Visualitzeu gràficament la distribució de l'edat. Escolliu el gràfic que sigui més apropiat, considerant que es vol conèixer la distribució de la variable i si aquesta segueix una distribució normal.

Escollim un histograma i un gràfic QQ-plot per validar si les dades segueixen una distribució normal

```
par(mfrow=c(1,2))
hist(cens$age, main="Histograma Edat", col = mypalette[1])
qqnorm(cens$age, main="QQ-plot Edat", col = mypalette[1])
```





2.2. Normalitat

Podem assumir normalitat per calcular l'interval de confiança de la mitjana d'edat? Justifiqueu la vostra resposta.

Com es pot observar en el gràfic QQ-plot, els registres segueixen una línia completament recta en l'interval entre 25 i 75 anys; per tant podem assumir que les dades segueixen una distribució normal en aquest interval. Tanmateix, com es pot observar en l'histograma, la freqüència més alta la trobem en el rang entre 30 i 35 anys, per tant podem deduir que la mitjana d'edat es trobarà dins de l'interval en que les dades tenen una distribució normal. Així doncs, podem assumir normalitat per a calcular l'interval de confiança de la mitjana d'edat.

2.3. Interval de confiança

Calculeu manualment l'interval de confiança de la mitjana de la variable age. Per fer-ho, definiu una funció IC que rebi la variable, la confiança, i que retorni un vector amb els valors de l'interval de confiança.

Funció IC per a calcular l'interval de confiança:

```
IC<-function(x, NC){
  alfa<-1-NC
  sd<-sd(x)
  n<-length(x)
  SE<-sd/sqrt(n)
  # Distribució t-student doncs no coneixem la variança poblacional
  z<-qt(alfa/2, df=n-1, lower.tail = FALSE)
  L<-mean(x)-z*SE
  U<-mean(x)+z*SE
  round(c(L,U),3)
}
```

2.4. Càlculs

Calculeu l'interval de confiança al 90% i 95 %. Compareu els resultats.

Apliquem la funció definida anteriorment a la variable *cens$age* amb els nivells de confiança sol · licitats:

```
R2a<-IC(cens$age, 0.90)
R2b<-IC(cens$age, 0.95)
R2a; R2b
```

```
## [1] 38.426 38.673
```

```
## [1] 38.403 38.697
```

L'interval de confiança al 90% és [38.426, 38.673] i l'interval de confiança al 95% és [38.403, 38.697]. A partir d'aquests resultats es pot observar que un increment en el nivell de confiança, comporta un increment en l'amplitud de l'interval de confiança calculat.

2.5. Interpretació

Expliqueu com s'interpreta l'interval de confiança a partir dels resultats obtinguts.

La interpretació dels resultats indica que el NC% de les mostres aleatòries obtingudes de la població donen lloc a un interval que conté el valor real de la mitjana poblacional. Anant al cas concret del primer càlcul, podem afirmar que en el cas que obtinguéssim infinites mostres de la població, el **90%** de les mostres, contindrien el valor real de la **mitjana poblacional** en l'interval [38.426, 38.673].

3. Salari

Ara investigarem el salari de la població. En particular, ens preguntem si en promig, el salari de les persones Self-Employed és inferior al de la resta de modalitats. Seguiu els passos que s'especifiquen a continuació.

3.1. Pregunta de recerca

Formuleu la pregunta de recerca.

La mitja de salari de les persones de la categoria *Self-Employed* és inferior a la resta de modalitats?

Considerarem dues opcions (o definicions més precises) de la pregunta de recerca:

- **Opció 1**: La mitja de salari de les persones de la categoria *Self-Employed* és inferior a la mitja de salaris del complementari, és a dir, de les persones *no Self-Employed*?
- **Opció 2**: La mitja de salari de les persones de la categoria *Self-Employed* és inferior a la mitja de salaris de les persones de cada una de les altres categories (*Government*, *Other/Unknown*, *Private*) ?

3.2. Hipòtesi

Escriviu les hipòtesis (hipòtesi nul · la i hipòtesi alternativa).

$$H_0 : \mu_{self} = \mu_i$$
$$H_1 : \mu_{self} < \mu_i$$

On μ_i fa referència a:

- La mitja de la mostra *no Self-Empoyed* en la **opció 1**.
- La mitja de la mostra per cada categoria (*Government*, *Other/Unknown*, *Private*) en la **opció 2**.

3.3. Test a aplicar

Expliqueu quin tipus de test podem aplicar atesa la pregunta de recerca plantejada i les característiques de la mostra. Justifiqueu la vostra elecció.

El test a aplicar és per a **dues mostres independents, sobre la mitjana amb variàncies desconegudes**.

En aquest moment, però, no sabem si les variàncies son desconegudes però iguals o bé, desconegudes i diferents. Per aquest motiu, aplicarem inicialment un test d'**igualtat de variàncies**

Test d'igualtat de variàncies

$$H_0 : \sigma^2_{self} = \sigma^2_i$$
$$H_1 : \sigma^2_{self} \neq \sigma^2_i$$

Funció var_test per a calcular el test de variança:

```
var_test<-function(m1, m2, NC){
  alfa<-1-NC
  meanX<-mean(m1); meanY<-mean(m2)
  nX<-length(m1); nY<-length(m2)
  sX<-sd(m1); sY<-sd(m2)

  fobs<-sX^2/sY^2
  fcritL<-qf(alfa/2, df1 = nX-1, df2 = nY-2)
  fcritU<-qf(1-alfa/2, df1 = nX-1, df2 = nY-2)
  pvalue<-2*min(pf(fobs, df1 = nX-1, df2 = nY-2, lower.tail = FALSE),
                pf(fobs, df1 = nX-1, df2 = nY-2))
  return(data.frame(fobs, fcritL, fcritU, pvalue, nX, nY))
}
```

Opció 1

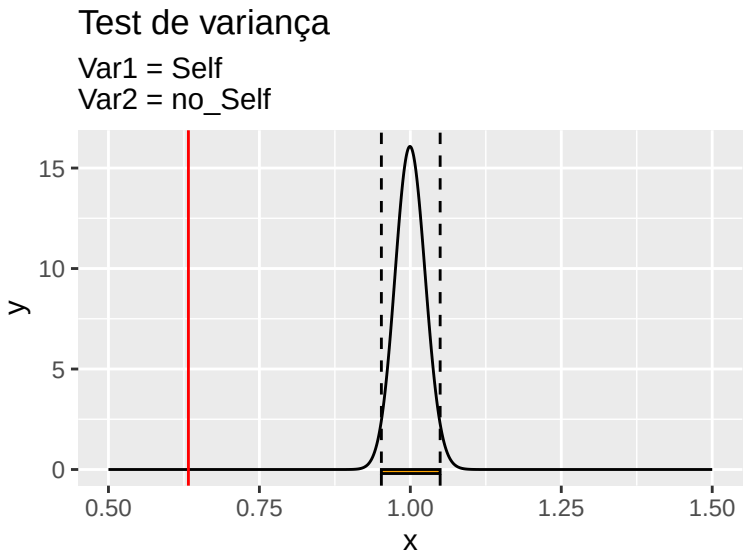
$$H_0 : \sigma^2_{self} = \sigma^2_{no\_self}$$
$$H_1 : \sigma^2_{self} \neq \sigma^2_{no\_self}$$

Aplicació de la funció definida anteriorment:

```
self<-cens$income[cens$workclass=="Self-Employed"]
no_self<-cens$income[!cens$workclass=="Self-Employed"]

r3a<-var_test(self,no_self, 0.95)
r3a
```

```
##        fobs    fcritL   fcritU       pvalue   nX    nY
## 1 0.6324152 0.9520603 1.049377 9.150353e-68 3657 28896
```

Graficant els valors resultat obtenim:



NOTA GENERAL

En el transcurs de l'informe, sempre que es representi gràficament el resultat d'un test trobarem:

- La distribució utilitzada en el càlcul, en negre sòlid.
- El valor o valors crítics, depenent si es un test unilateral o bilateral, representats amb una o dues línies verticals discontinues (f_{critL}, f_{critU}).
- L'interval d'acceptació d'H_0 marcat amb un petit rectangle taronja.
- El valor observat (Z_{obs}) representat amb una línia en vermell.

Així doncs, en el cas concret del test de variança obtenim tant numèricament com gràficament, que f_{obs} està fora de l'interval d'acceptació d'H_0, anàlogament, el valor p es inferior al nivell de significança; per tant rebutjem la hipòtesi nul · la en favor de l'alternativa i conseqüentment podem dir que les variàncies son diferents.

Opció 2

$$H_0 : \sigma^2_{self} = \sigma^2_{worklass_i}$$
$$H_1 : \sigma^2_{self} \neq \sigma^2_{worklass_i}$$

Aplicació de la funció definida anteriorment:

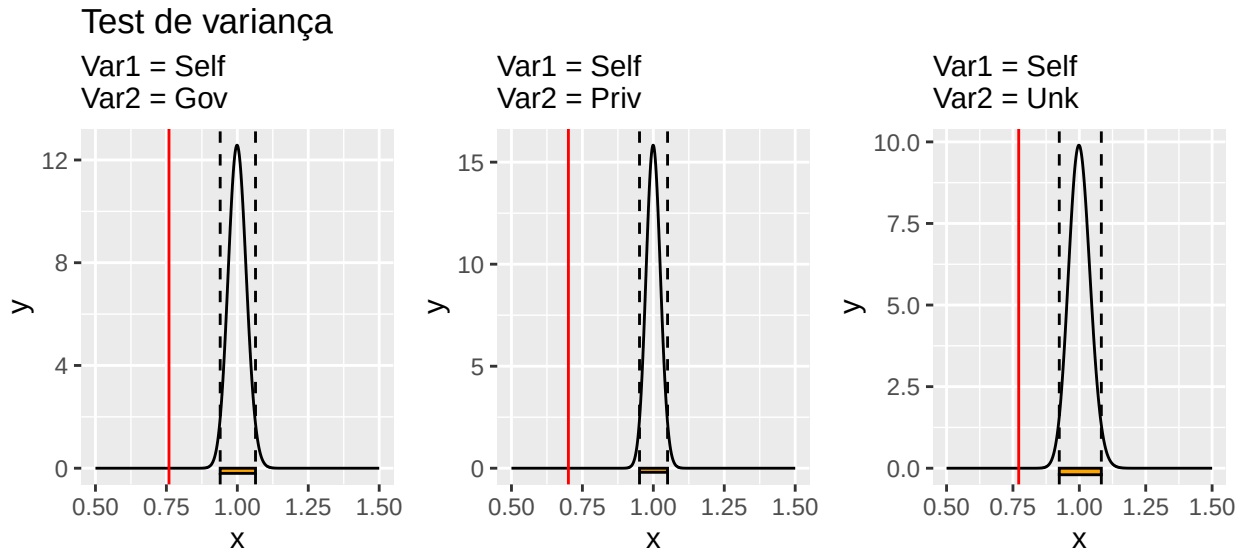
```
gov<-cens$income[cens$workclass=="Government"]
priv<-cens$income[cens$workclass=="Private"]
unk<-cens$income[cens$workclass=="Other/Unknown"]

r3b1<-var_test(self,gov,0.95)
r3b2<-var_test(self,priv,0.95)
r3b3<-var_test(self,unk, 0.95)
r3b1; r3b2; r3b3
```

```
##        fobs    fcritL   fcritU       pvalue   nX   nY
## 1 0.7591727 0.9396068 1.064095 5.982695e-18 3657 4349
```

```
##        fobs    fcritL   fcritU       pvalue   nX    nY
## 1 0.7007223 0.9513722 1.050174 1.358562e-41 3657 22692
```

```
##        fobs    fcritL   fcritU       pvalue   nX   nY
## 1 0.7717228 0.9244619 1.082833 7.368674e-11 3657 1855
```

Gràficament:



En aquest cas, també observem que la f_{obs} està fora de l'interval d'acceptació d'H_0, i que el valor p és inferior al nivell de significança en tots els casos d'estudi; per tant rebutjem la hipòtesi nul · la en favor de l'alternativa. I per tant, confirmem que en tots els casos les variàncies son diferents.

Finalment podem concloure que el test a aplicar es per a **dues mostres independents, sobre la mitjana amb variàncies desconegudes i diferents**.

3.4. Càlcul

Calculeu el test usant una funció pròpia. Implementeu una funció que realitzi el càlcul del test i que pugueu utilitzar amb diferents valors de nivell de confiança. Calculeu el test per a un nivell de confiança del 95% i del 90 %. Mostreu els resultats (valor observat, crític i valor p) en una taula.

Funció my_test_1 per a calcular el test sobre la mitjana de dues mostres independents amb variances desconegudes i independents:

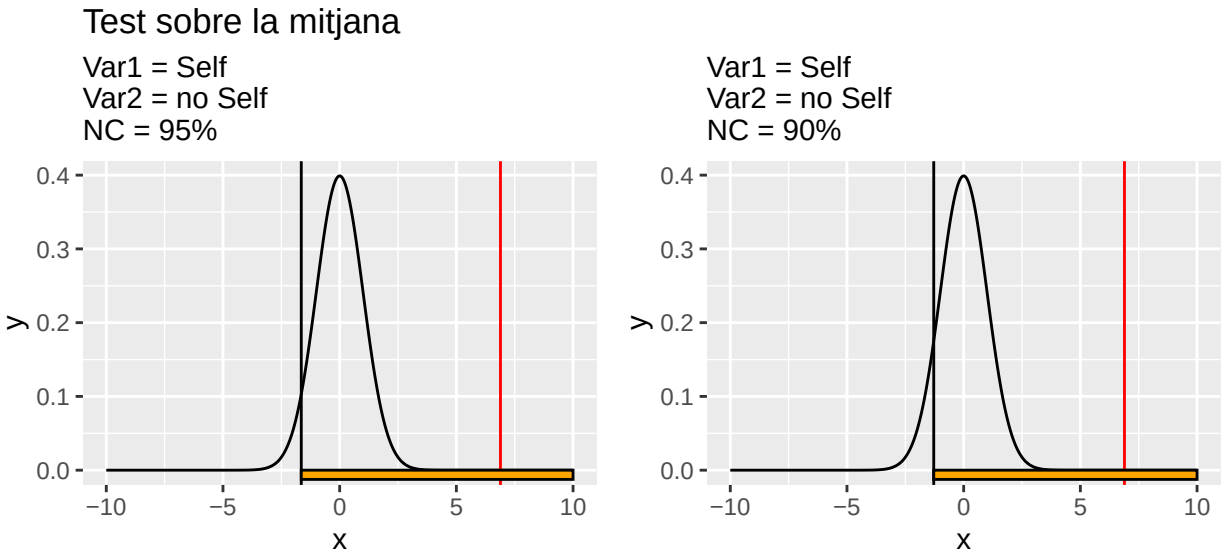
```
my_test_1<-function(m1, m2, NC){

  alfa<-1-NC
  meanX<-mean(m1); meanY<-mean(m2)
  nX<-length(m1); nY<-length(m2)
  sX<-sd(m1); sY<-sd(m2)

  v<-((((sX^2)/nX)+((sY^2)/nY))^2)/(((((sX^2)/nX)^2)/(nX-1))+((((sY^2)/nY)^2)/(nY-1)))

  tobs<-(meanX-meanY)/sqrt((sX^2/nX)+(sY^2/nY))

  tcritL<-qt(alfa, v)
  tcritU<-"INF"
  pvalue<-pt(tobs, df = v, lower.tail = TRUE)

  tobs<-round(tobs,4)
  tcritL<-round(tcritL,4)
  pvalue<-round(pvalue,4)

  return(data.frame(tobs,tcritL,tcritU,pvalue,v))
}
```

Opció 1

Aplicació de la funció definida anteriorment:

```
R3a1<-my_test_1(self, no_self, 0.95)
R3a2<-my_test_1(self, no_self, 0.90)
R3a<-rbind(R3a1,R3a2)
rownames(R3a)<-c("NC=95%", "NC=90%")
kable(R3a)
```

	tobs	tcritL	tcritU	pvalue	v
NC=95%	6.8897	-1.6451	INF	1	5239.124
NC=90%	6.8897	-1.2817	INF	1	5239.124



Opció 2

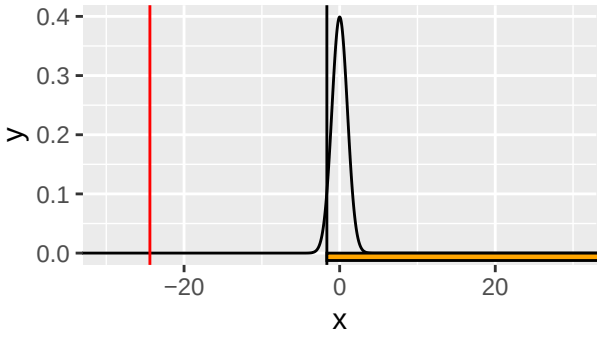
Aplicació de la funció definida anteriorment:

```
R3b1<-my_test_1(self,gov,0.95)
R3b2<-my_test_1(self,gov,0.90)

R3b3<-my_test_1(self,priv,0.95)
R3b4<-my_test_1(self,priv, 0.90)

R3b5<-my_test_1(self,unk,0.95)
R3b6<-my_test_1(self,unk, 0.90)

R3b<-rbind(R3b1,R3b2,R3b3,R3b4,R3b5,R3b6)
rownames(R3b)<-c("Gov NC=95%", "Gov NC=90%", "Priv NC=95%", "Priv NC=90%",
                 "Unk NC=95%", "Unk NC=90%")
kable(R3b)
```

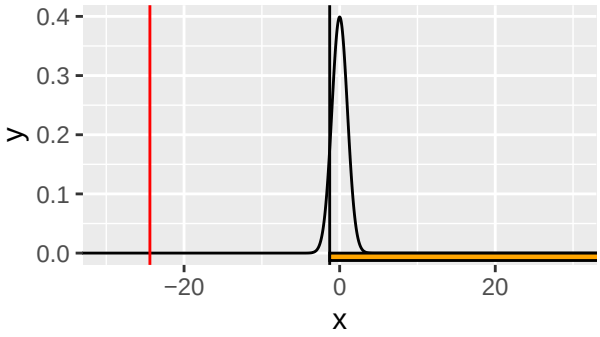
	tobs	tcritL	tcritU	pvalue	v
Gov NC=95%	-24.3927	-1.6450	INF	0	7993.913
Gov NC=90%	-24.3927	-1.2817	INF	0	7993.913
Priv NC=95%	8.6002	-1.6451	INF	1	5484.320
Priv NC=90%	8.6002	-1.2817	INF	1	5484.320
Unk NC=95%	44.5107	-1.6453	INF	1	3330.782
Unk NC=90%	44.5107	-1.2818	INF	1	3330.782

Test sobre la mitjana

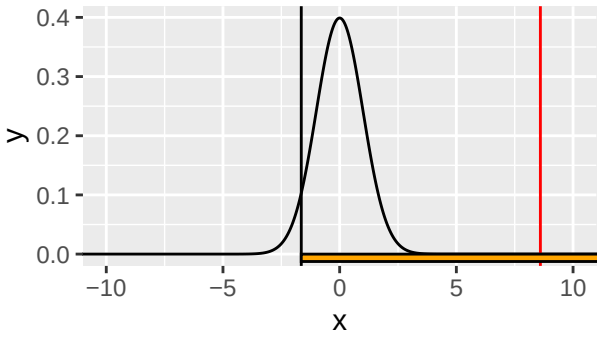
Var1 = Self
Var2 = Gov
NC = 95%



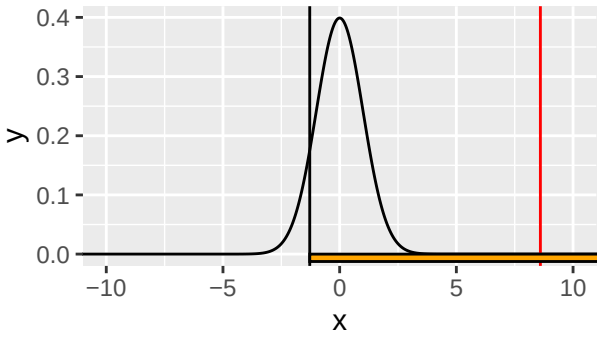
Var1 = Self
Var2 = Gov
NC = 90%



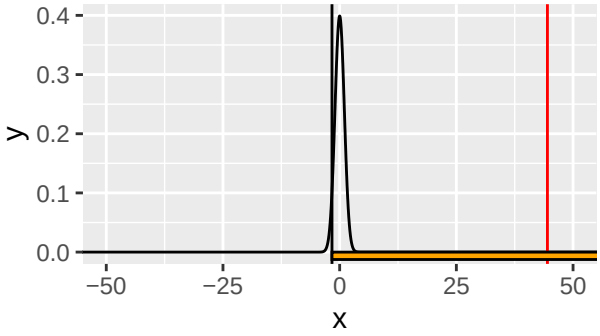
Var1 = Self
Var2 = Priv
NC = 95%



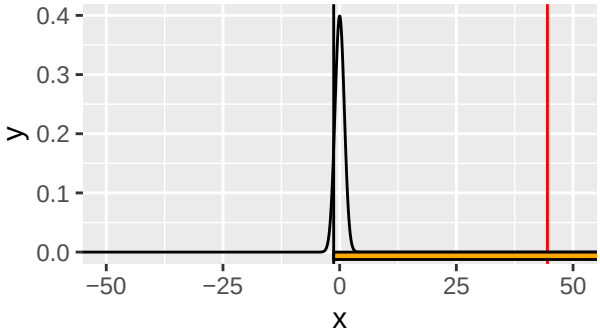
Var1 = Self
Var2 = Priv
NC = 90%



Var1 = Self
Var2 = Unk
NC = 95%



Var1 = Self
Var2 = Unk
NC = 90%



3.5. Conclusió

A partir dels resultats obtinguts, doneu resposta a la pregunta de recerca.

Opció 1

Amb els resultats obtinguts, podem concloure que a la població, la mitja de salari de les persones de la categoria *Self-Employed* **no** és inferior a la mitja de salari de les persones *no Self-Employed* amb un nivell de confiança del 90 o 95%, donat que la t_{obs} està dins de l'interval d'acceptació d'H_0. Exemple al 95% t_{obs}=6.8897, interval d'acceptació d'H_0=[-1.6451, INF].

Opció 2

1. En la població, la mitja de salari de les persones de la categoria *Self-Employed* **si** és inferior a la mitja de salari de les persones *Goverment* amb un nivell de confiança del 90 i del 95%, donat que la t_{obs} està fora de l'interval d'acceptació d'H_0 i per tant podem rebutjar la hipòtesi nul · la en favor de l'alternativa. Exemple al 95% t_{obs}=-24.3927, interval d'acceptació d'H_0=[-1.645, INF].
2. En la població, la mitja de salari de les persones de la categoria *Self-Employed* **no** és inferior a la mitja de salari de les persones *Private* amb un nivell de confiança del 90 i del 95%, donat que la t_{obs} està dins de l'interval d'acceptació d'H_0. Exemple al 95% t_{obs}=8.6002, interval d'acceptació d'H_0=[-1.6451, INF].
3. En la població, la mitja de salari de les persones de la categoria *Self-Employed* **no** és inferior a la mitja de salari de les persones *Other/Unknown* amb un nivell de confiança del 90 i del 95%, donat que la t_{obs} està dins de l'interval d'acceptació d'H_0. Exemple al 95% t_{obs}=44.5107, interval d'acceptació d'H_0=[-1.6453, INF].

4. Proporció de Self-Employed

Ens preguntem si el percentatge de Self-Employed a la població és superior al 10 %. Apliqueu el test necessari per donar resposta a aquesta pregunta. Seguiu els passos que s'indiquen a continuació.

4.1. Pregunta

Formuleu la pregunta de recerca que es planteja en aquesta secció.

El percentatge de persones *Self-Employed* a la població és superior al 10%?

4.2. Hipòtesi

Escriviu les hipòtesis (hipòtesi nul · la i hipòtesi alternativa).

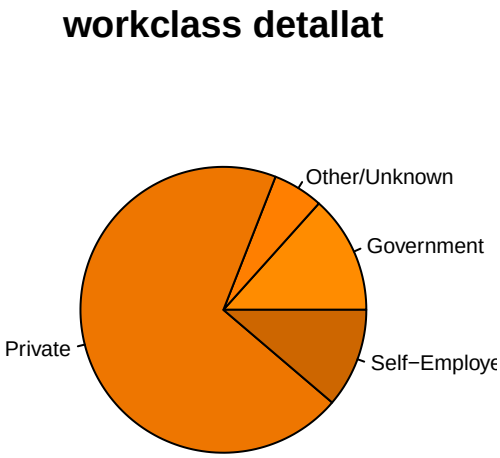
$$H_0 : p_{self} = 0.1$$
$$H_1 : p_{self} > 0.1$$

4.3. Anàlisi visual

Representeu de forma gràfica la proporció de Self-Employed a la mostra.

Representem les proporcions a la mostra en un diagrama de sectors.

```
cens$Self_Employed<-as.factor(ifelse(cens$workclass=="Self-Employed",
                                     "Self-Employed","no_Self-Employed"))
par(mfrow=c(1,2))
pie(summary(cens$Self_Employed),main = "workclass", col = mypalette, cex=0.7)
pie(summary(cens$workclass),main = "workclass detallat", col = mypalette, cex=0.7)
```



4.4. Contrast

Expliqueu quin tipus de contrast podem aplicar atesa la pregunta de recerca plantejada i les característiques de la mostra. Justifiqueu la vostra elecció.

Donada la hipòtesi alternativa, es tracta d'un test unilateral per la dreta d'una mostra sobre la proporció, l'estadístic de contrast sota la hipòtesi nul · la és:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

4.5. Càlcul

Calculeu el test usant una funció pròpia. Podeu crear una funció que rebi els paràmetres necessaris i el nivell de confiança. Després, calculeu el test, cridant aquesta funció amb nivell de confiança del 95 %. Mostreu els resultats (valor observat, crític i valor p) en una taula.

Funció my_test_2 per a calcular el test d'una mostra sobre la proporció:

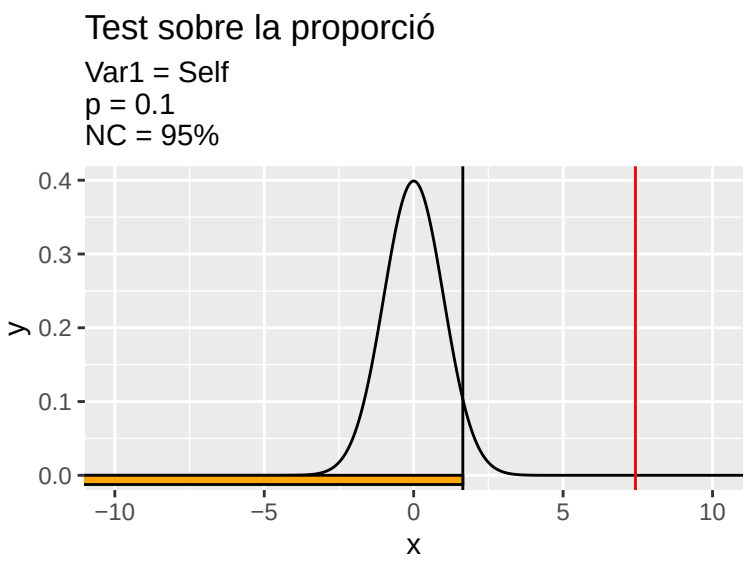
```
my_test_2<-function(m, cat, p, NC){

  alfa<-1-NC
  n<-length(m)
  pm<-length(m[m==cat])/n

  zobs<-(pm-p)/sqrt((p*(1-p))/n)

  zcritL<-"-INF"
  zcritU<-qnorm(1-alfa)
  pvalue<-pnorm(zobs, lower.tail = FALSE)
```

```
  zobs<-round(zobs,4)
  zcritU<-round(zcritU,4)
  pvalue<-round(pvalue,4)

  return(data.frame(zobs, zcritL, zcritU, pvalue))
}
```

Aplicació de la funció definida anteriorment:

```
R4<-my_test_2(cens$workclass, "Self-Employed", 0.1, 0.95)
kable(R4)
```

zobs	zcritL	zcritU	pvalue
7.4214	-INF	1.6449	0



4.6. Conclusió

A partir dels resultats obtinguts, doneu resposta a la pregunta de recerca.

Obtenim una z_{obs}=7.4214 fora de l'interval d'acceptació d'H_0=[-INF, 1.6449], anàlogament obtenim un valor p=0 molt inferior al nivell de significança (α=0.05); per tant podem rebutjar la hipòtesi nul · la en favor de l'alternativa i afirmar que el percentatge de persones *Self-Employed* a la població es superior al 10%.

5. Proporció de Self-Employed en dones i homes

Ens preguntem si la proporció de Self-Employed és menor entre les dones que entre els homes en la població. Per donar resposta a aquesta pregunta, seguiu els passos que s'indiquen a continuació.

5.1. Pregunta de recerca

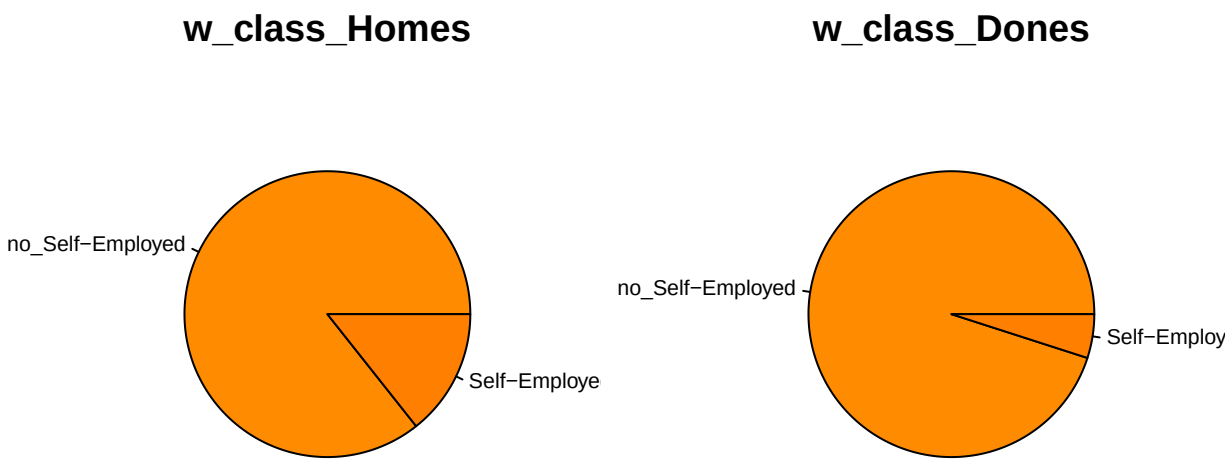
Formuleu la pregunta de recerca que es planteja en aquesta secció.

La proporció de persones *Self-Employed* és inferior entre dones que entre homes?

5.2. Anàlisi visual

Representeu de forma gràfica la proporció de Self-Employed a la mostra d'homes i dones respectivament.

```
censM<-cens[cens$gender=="m",]
censF<-cens[cens$gender=="f",]

censM$Self_Employed<-as.factor(ifelse(censM$workclass=="Self-Employed",
                                      "Self-Employed","no_Self-Employed"))
censF$Self_Employed<-as.factor(ifelse(censF$workclass=="Self-Employed",
                                      "Self-Employed","no_Self-Employed"))
par(mfrow=c(1,2))
pie(summary(censM$Self_Employed),main = "w_class_Homes", col = mypalette, cex=0.7)
pie(summary(censF$Self_Employed),main = "w_class_Dones", col = mypalette, cex=0.7)
```



5.3. Hipòtesi

Escriviu la hipòtesi nul · la i la hipòtesi alternativa.

$$H_0 : p_{self}^{dones} = p_{self}^{homes}$$
$$H_1 : p_{self}^{dones} < p_{self}^{homes}$$

5.4. Test

Expliqueu quin tipus de test podem aplicar atesa la pregunta de recerca plantejada i les característiques de la mostra. Justifiqueu la vostra elecció.

Es tracta d'un test unilateral per l'esquerra de dues mostres independents (homes i dones) sobre la proporció.

5.5. Càlcul

Calculeu el test usant una funció pròpia. Igual que en apartats anteriors, es recomana definir una funció que faci el càlcul i que rebi els paràmetres necessaris. Calculeu el contrast per a un nivell de confiança del 97 %. Mostreu els resultats (valor observat, crític i valor p) en una taula.

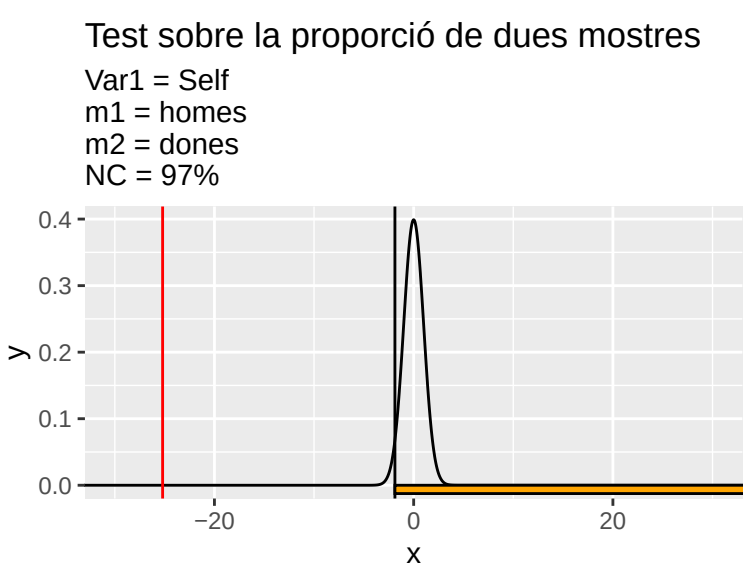
Funció my_test_3 per a calcular el test de dues mostres independents sobre la proporció:

```
my_test_3<-function(m1, m2, cat, NC){

  alfa<-1-NC
  n1<-length(m1); n2<-length(m2)
  pm1<-length(m1[m1==cat])/n1; pm2<-length(m2[m2==cat])/n2;
  p<-((n1*pm1)+(n2*pm2))/(n1+n2)

  zobs<-(pm1-pm2)/sqrt(p*(1-p)*((1/n1)+(1/n2)))

  zcritL<-qnorm(alfa)
  zcritU<-"INF"
  pvalue<-pnorm(zobs, lower.tail = TRUE)

  zobs<-round(zobs,4)
  zcritL<-round(zcritL,4)
  pvalue<-round(pvalue,4)

  return(data.frame(zobs, zcritL, zcritU, pvalue))
}
```

Aplicació de la funció definida anteriorment:

```
R5<-my_test_3(censF$workclass,censM$workclass,"Self-Employed", 0.97)
kable(R5)
```

zobs	zcritL	zcritU	pvalue
-25.202	-1.8808	INF	0



5.6. Conclusió

A partir dels resultats obtinguts, proporcioneu una resposta a la pregunta de recerca.

A partir dels càlculs anteriors obtenim una z_{obs}=-25.202 fora de l'interval d'acceptació d'H_0=[-1.8808, INF] i un valor p=0 molt inferior al nivell de significança (α=0.03), per tant podem rebutjar la hipòtesi nul · la en favor de l'alternativa i afirmar que el percentatge de persones *Self-Employed* és menor entre les dones que entre els homes en la població.

6. Dependència Gènere - Self-Employed

En aquesta secció es demana aplicar el test d´independència Chi quadrat per avaluar si les variables gènere i Self-Employed són independents. Seguiu els passos que s'indiquen a continuació..

6.1. Pregunta de recerca

Les variables *gender* i *workclass==Self-Employed* estan relacionades o son independents?

6.2. Hipòtesi

Escriviu la hipòtesi nul · la i alternativa.

- H_0: Les variables *gender* i *workclass==Self-Employed* són independents
- H_1: Hi ha una relació entre les variables *gender* i *workclass==Self-Employed*

6.3. Test

Descriviu breument en què consisteix el test Chi quadrat. Calculeu la matriu de contingència i mostreu-ne els valors.

El test Chi quadrat tracta de comparar les freqüències esperades si les variables no estiguessin relacionades, amb les obtingudes de la mostra. Quan les freqüències observades són molt diferent a les esperades, podem concloure que hi ha una relació entre les variables.

Càlcul de la matriu de contingència:

```
tc<-table(cens$gender,cens$Self_Employed)
tc<-cbind(tc,rowSums(tc))
colnames(tc)[3]<-"sum_row"
tc<-rbind(tc,colSums(tc))
rownames(tc)[3]<-"sum_col"
kable(tc)
```

	no_Self-Employed	Self-Employed	sum_row
f	10233	534	10767
m	18663	3123	21786
sum_col	28896	3657	32553

6.4. Càlcul

Realitzeu els càlculs del test Chi quadrat, implementant una funció pròpia. Calculeu el contrast per a un nivell de confiança de 97 %.

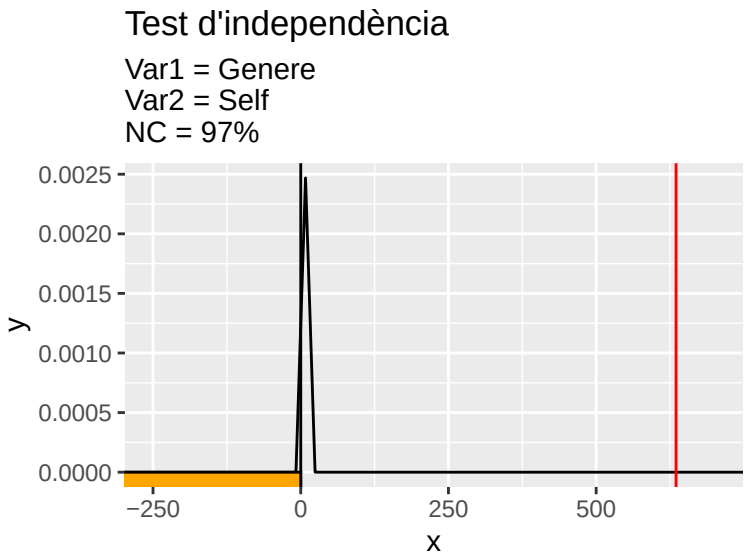
Funció my_test_4 per a calcular el test d'independència de dues variables:

```
my_test_4<-function(x, y, NC){
  alfa<-1-NC
  #taula de contingència
  tc<-table(x,y)
  tc<-cbind(tc,rowSums(tc))
  colnames(tc)[ncol(tc)]<-"sum_row"
  tc<-rbind(tc,colSums(tc))
  rownames(tc)[nrow(tc)]<-"sum_col"
  #valors esperats
  te<-matrix(1,2,2)
  for (i in 1:nrow(te)){
    for (j in 1:ncol(te)){
      te[i,j]<-round((tc[i,3]*tc[3,j])/tc[3,3],2)
    }
  }
  df<-(nrow(te)-1)*(ncol(te)-1)

  chisqobs<-(sum(((tc[1:2,1:2]-te)^2)/te))

  chicritL<-"-INF"
  chicritU<-qchisq(1-alfa, df, lower.tail = FALSE)
  pvalue<-pchisq(chisqobs,df, lower.tail = FALSE)

  return(data.frame(chisqobs, chicritL, chicritU, pvalue, df))
}
```

Aplicació de la funció definida anteriorment:

```
R6<-my_test_4(cens$gender,cens$Self_Employed, 0.97)
kable(R6)
```

chisqobs	chicritL	chicritU	pvalue	df
635.1357	-INF	0.0014144	0	1



6.5. Conclusió

Responeu la pregunta de recerca plantejada en aquest apartat. Relacioneu el resultat amb el contrast de la secció anterior, on es realitza un test sobre les proporcions.

A partir dels càlculs anteriors obtenim una χ^2_{obs}=635.1357441 fora de l'interval d'acceptació d'H_0=[-INF, 0.00141] i un valor p=$3.8181343 \times 10^{-140}$ molt inferior al nivell de significança (α=0.03), per tant podem rebutjar la hipòtesi nul · la en favor de l'alternativa i afirmar hi ha una relació entre les variables *gender* i *workclass==Self-Employed*.

Aquest mateix fet l'hem trobat en l'exercici anterior quan hem corroborat que $p^{dones}_{self} < p^{homes}_{self}$, fet que indica que existeix una relació entre les variables *gender* i *workclass==Self-Employed*. Si no existís relació entre aquestes variables esperaríem que la proporció de persones *Self-Employed* fos la mateixa tan per *gender=="f"* com per *gender=="m"*.

7. Resum i conclusions

Presenteu una taula amb els resultats principals de cada secció: la pregunta de recerca plantejada, els valors obtinguts del contrast i la conclusió obtinguda a cada apartat.

N	Pregunta	Resultat	Conclusió
2a	Interval de confiança de la mitjana d'edat al 95%	[38.4,38.7]	L'interval de confiança de la mitjana d'edat al 95% és [38.4,38.7]
2b	Interval de confiança de la mitjana d'edat al 90%	[38.43,38.67]	L'interval de confiança de la mitjana d'edat al 90% és [38.43,38.67]
3a.1	En promitg, el salari de Self-Employed és inferior a la resta (en conjunt) al 95%?	6.8897, -1.6451, INF, 1	En promitg, el salari de Self-Employed NO és inferior a la resta (en conjunt) al 95%
3a.2	En promitg, el salari de Self-Employed és inferior a la resta (en conjunt) al 90%?	6.8897, -1.2817, INF, 1	En promitg, el salari de Self-Employed NO és inferior a la resta (en conjunt) al 90%
3b.1	En promitg, el salari de Self-Employed és inferior a les Government al 95%?	-24.3927, -1.645, INF, 0	En promitg, el salari de Self-Employed SI és inferior a les Government al 95%

N	Pregunta	Resultat	Conclusió
3b.2	En promitg, el salari de Self-Employed és inferior a les Government al 90%?	-24.3927, -1.2817, INF, 0	En promitg, el salari de Self-Employed SI és inferior a les Government al 90%
3b.3	En promitg, el salari de Self-Employed és inferior a les Private al 95%?	8.6002, -1.6451, INF, 1	En promitg, el salari de Self-Employed NO és inferior a les Private al 95%
3b.4	En promitg, el salari de Self-Employed és inferior a les Private al 90%?	8.6002, -1.2817, INF, 1	En promitg, el salari de Self-Employed NO és inferior a les Private al 90%
3b.5	En promitg, el salari de Self-Employed és inferior a les Other/Unknown al 95%?	44.5107, -1.6453, INF, 1	En promitg, el salari de Self-Employed NO és inferior a les Other/Unknown al 95%
3b.6	En promitg, el salari de Self-Employed és inferior a les Other/Unknown al 90%?	44.5107, -1.2818, INF, 1	En promitg, el salari de Self-Employed NO és inferior a les Other/Unknown al 90%
4	El percentatge de persones Self-Employed a la població es superior al 10%?	7.4214, -INF, 1.6449, 0	SI, els percentatge de persones Self-Employed és superior al 10% en la població
5	La proporció de persones Self-EMployed és menor entre les dones que entre els homes a la població?	-25.202, -1.8808, INF, 0	SI, el percentatge de persones Self-Employed és menor entre les dones que entre els homes a la població
6	Hi ha una relació entre les variables Gender i Workclass==Self-Employed?	635.1357, -INF, 0.0014, 0	SI, existeix relació entre les variables Gender i Workclass==Self-Employed