Guia de pràctiques de mostreig estadístic

Montserrat Guillén

Dept. d'Econometria, Estadística i Economia Espanyola, RFA-IREA
Universitat de Barcelona

Mónica Bécue

Dept. d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Manuela Alcañiz

Dept. d'Econometria, Estadística i Economia Espanyola, RFA-IREA
Universitat de Barcelona

Guia de pràctiques de mostreig estadístic

Agraïm l'ajut rebut de la Generalitat de Catalunya dins la convocatòria d' Ajuts per al finançament de projectes per a la millora de la qualitat docent a les universitats catalanes *per a l'any 2009, MQD-00166.*
Les autores agraeixen als estudiants de l'assignatura Tècniques de Mostreig la seva col·laboració en millorar aquests exercicis pràctics.

Dipòsit legal:

ISBN: 978-84-615-0779-5

Impressió: Gràficas Rey, S.L.

Índex

Presentació **9**

Pràctiques **13**
Pràctica 1 13
Part 1 Extracció de mostres aleatòries simples 14
Part 2 Anàlisi bàsica de mostres aleatòries simples 17
Part 3 Exercicis individuals 18
Pràctica 2 21
Part 1 Selecció de mostres amb probabilitats desiguals 21
Part 2 Inferència en mostres aleatòries amb probabilitats desiguals 24
Part 3 Exercicis individuals 26
Pràctica 3 27
Part 1 Selecció de mostres amb estratificació 27
Part 2 Inferència en mostres aleatòries estratificades 31
Part 3 Exercicis individuals 33
Pràctica 4 35
Part 1 Anàlisi de mostres amb conglomerats 35
Part 2 Exercicis individuals 42
Pràctica 5 45
Part 1 Regressió amb dissenys mostrals complexos. Simulació 45
Part 2 Exercicis individuals 49
Pràctica 6 51
Part 1 Mètode bootstrap 51
Part 2 Exercicis individuals 52
Solucions 53
Pràctica 1 53
Pràctica 2 54
Pràctica 3 55
Pràctica 4 56
Pràctica 5 57
Pràctica 6 58

Bibliografia **61**

Presentació

La present publicació inclou el contingut aplicat de l'assignatura Tècniques de Mostreig, que s'ofereix al màster interuniversitari d'Estadística i Investigació Operativa, de la Universitat de Barcelona i la Universitat Politècnica de Catalunya.

La voluntat de la publicació no és esdevenir un recull exhaustiu d'exercicis relacionats amb els diferents mètodes de mostreig estadístic que es presenten a l'assignatura, sinó recollir el conjunt de pràctiques informàtiques, per resoldre amb el programa SAS©, que inclou l'assignatura en qüestió.

L'estructura del document és la següent: es reprodueixen els enunciats de les sis pràctiques[1] de l'assignatura (amb les pertinents indicacions de la sintaxi a utilitzar en cadascun dels exercicis); i, a continuació, a la secció Solucions, es recullen els resultats numèrics dels exercicis que ho requereixen. Finalment, una breu secció de referències bibliogràfiques clou la publicació.

Els objectius que es persegueixen a través de l'assignatura Tècniques de Mostreig i, per extensió, amb la present publicació són:

- Donar una visió actualitzada de les tècniques de mostreig estadístic i insistir sobre els desenvolupaments més recents. A més, es vol dedicar molta atenció als problemes que sorgeixen en la pràctica del mostreig, com ara el tractament de les no-respostes o les tècniques de mostreig indirecte a utilitzar en cas de no disposar d'un marc mostral.
- Mostrar com s'utilitza el programari de mostreig, en particular al programari especialitzat SAS©.
- Consolidar la noció que el disseny del mostreig estadístic és una etapa bàsica de qualsevol estudi i que cal tenir-la en compte en l'anàlisi posterior.

Barcelona, abril de 2011

[1] La realització de les pràctiques requereix disposar de diferents bases de dades, que estan a disposició de l'estudiant en el Campus Virtual. En cas contrari, cal contactar amb la professora Montserrat Guillén Estany (mguillen@ub.edu) per poder-ne disposar.

Pràctiques

Pràctica 1

Disposem de dues bases de dades SAS que trobareu en el Campus Virtual (ATENEA). Els arxius es diuen: `prob1.zip` i `prob2.zip`.

1. Creeu la carpeta `C:\TM` i descomprimiu els arxius anteriors.
2. Veurem el contingut de les bases de dades. Executeu:

```
libname prac 'C:\TM\';
proc contents data=prac.prob1; run;
```

L'arxiu `prob1` és un marc mostral i conté les variables de la taula 1 sobre clients d'una empresa d'assegurances.

Taula 1. Contingut de les dades `prob1`

Nom	Etiqueta
antig	Seniority of the customer
claim_b1	Claim within past 2 monts
claim_b2	Claim between 2 and 6 months ago
claim_b3	Claim between 6 and 12 months ago
claim_b4	Claim between 12 and 24 months ago
claim_b5	Claim more than 2 years ago
corecust	Core customer status
edat	Age of the customer when lapse occur
gend	Gender

L'arxiu `prob2` és un marc mostral i conté les variables de la taula 2 sobre visitants d'un museu.

Taula 2. Contingut de les dades `prob2`

Nom	Etiqueta (no inserida)
entrada	Hora de validació de l'entrada
sortida	Hora de validació de la sortida
d	Dia de la setmana (1 = dilluns, ..., 7 = diumenge)
t	Temps en minuts de durada de la visita
s	Individu selecionat per a enquesta
Obra	Obra preferida (O1, ..., O5), només per als enquestats

Part 1 Extracció de mostres aleatòries simples

Treballarem amb la base de dades `prob1`. Desarem les comandes a l'arxiu `CognomNom.sas`, que enviareu a ATENEA. Poseu `* EXERCICI num ;` per mostrar les comandes de cada apartat.

Exercici 1 Extraieu una mostra de mida 100 i deseu-la com a base de dades SAS al directori `C:\TM` amb el nom `assr1`.

```
Data temp;
 Set prac.prob1;
 alea= ranuni (12345);
 run;

Proc sort data = temp; by alea;

Data prac.assr1;
 set temp (obs = 100);
 drop alea;
 run;
```

Exercici 2 Feu el mateix amb PROC SURVEYSELECT i deseu-ho com a base de dades SAS al directori `C:\TM` amb el nom `assr2`. S'han seleccionat les mateixes unitats que a l'exercici 1?

```
Proc surveyselect data = prac.prob1
     method = SRS n = 100 out = prac.assr2;
 run;
```

Exercici 3 Extraieu 3 mostres independents amb selecció aleatòria simple sense reposició de mida $n = 10$ i deseu-les juntes com a base de dades SAS al directori `C:\TM` amb el nom `assr3`. Quina variable es crea automàticament i identifica cada mostra?

```
Proc surveyselect data=prac.prob1
     out=prac.assr3 sampsize=10 method=srs rep=3 seed=12345 stats;
 run;
```

Exercici 4 Seleccioneu 3 mostres aleatòries simples independents de mida $n = 100$ i deseu-les com a tres bases de dades SAS al directori `C:\TM` amb el nom `assr4a1`, `assr4a2` i `assr4a3`.

```
%Macro repeticions(numero);
 %do i=1 %to &numero;
 Proc surveyselect data = prac.prob1
      method = SRS n = 100 out = prac.assr4a&i;
 run;
 %end;
%mend;

%repeticions(3);
```

Exercici 5 Seleccioneu una mostra de mida 200 i deseu-la com a base de dades `assr5`, deseu-hi també el número de l'observació del marc original amb el nom `k`.

```
Data temp;
 set prac.prob1;
 k=_N_;
 run;

Proc surveyselect data = temp
     method = SRS n = 200 out = prac. assr5;
 run;
```

Exercici 6 Seleccioneu 50 elements i 2 reserves diferents per a cadascun. Deseu-ho tot junt com a base de dades `assr6`. Poseu a l'individu principal $res = 0$; al primer reserva, $res = 1$ i al tercer, $res = 2$.

```
Proc surveyselect data=prac.prob1  method = srs sampsize = 150
     rep=1 seed=12345 out=prac.assr6;
 run;

Data prac.assr6;
 set prac.assr6;
 res=mod((_N_)-1, 3);
 run;
```

Exercici 7 Seleccioneu 500 elements amb reposició. Deseu-los sense repetició com a base de dades `assr7a1`. Feu un taula de `numberhits` i creeu una base de dades `assr7a1` on es repeteixin tantes vegades com s'hagin seleccionat els elements que han estat seleccionats més d'una vegada.

```
Proc surveyselect data=prac.prob1  method = urs sampsize = 500
     rep=1 seed=12345 out=prac.assr7a1;
 run;

Proc freq data = prac.assr7a1;
 table numberhits;
 run;

Data prac.assr7a2;
 set prac.assr7a1;
 do i = 1 to numberhits;
   output;
 end;
 drop i;
 run;

Proc freq data = prac.assr7a2;
 table numberhits;
 run;
```

Exercici 8 Calculeu la mida de mostra necessària en un disseny aleatori simple per a una població de 16000 individus, suposant que volem un marge d'error del 10% en una proporció $p = 50\%$. Suposem un nivell de confiança del 95%.

```
Data dades;
 input alfa p Npob merror;
 cards;
 0.05 0.5 16000 0.10
 ;

Data mostra;
  set dades;
  z=probit(1-alfa/2);
  v=(merror/z)*(merror/z);
  n=(1/(p+Npob*v-p*p))*(-Npob*p*p+Npob*V+Npob*p);
  put 'mida mostra ' n;
  run;
```

Exercici 9 Feu una selecció aleatòria simple "per unitat", recollida a la variable `antig`. Fixem-nos amb les variable `SelectionProb` i `SamplingWeight`. Deseu-ho com a base de dades SAS al directori `C:\TM` amb el nom `assr9`.

```
Proc surveyselect data=prac.prob1 method=pps
     sampsize=100 seed=12345 out=prac.assr9;
     size antig;
 run;
```

Exercici 10 Extraieu una mostra de mida 20 per un mètode seqüencial, deseu-la com a base de dades SAS al directori `C:\TM` amb el nom `assr10` i imprimiu els números d'observació seleccionats del marc mostral.

```
Data temp;
 set prac.prob1;
 k=_N_;
 run;

Proc surveyselect data = temp
     method = seq n = 20 out = prac.assr10;
 run;

Proc print data=prac.assr10;
 var  k;
 run;
```

Part 2 Anàlisi bàsica de mostres aleatòries simples

Exercici 11 Utilitzeu la base de dades inicial `prob1` i calculeu el total de la variable `corecust` i la mitjana de la variable `edat`.

```
Proc means data=prac.prob1 sum n;
 var corecust;
 run;

Proc means data=prac.prob1 mean;
 var edat;
 run;
```

Exercici 12 Utilitzeu la base de dades inicial `prob1` i calculeu l'error estàndard de la variable `corecust`. Interpreteu-lo.

```
Proc means data=prac.prob1 sum n std;
 var corecust;
 run;
```

Exercici 13 Utilitzeu la mostra aleatòria simple desada a `assr1` i la desada a `assr2` per estimar el total de la variable `corecust` a la població.

```
Proc surveymeans data=prac.assr1 total=20309 sum;
 var corecust;
 run;

Proc surveymeans data=prac.assr2 total=20309 sum;
 var corecust;
 run;

*falta multiplicar per N=20309;
```

Exercici 14 Utilitzeu la mostra desada a `assr2` per estimar el total de la variable `corecust` directament i l'error estàndard de l'estimació.

```
Data temp;
 set prac.assr2;
 sw=20309/100;
 run;

Proc surveymeans data=temp total=20309 sum;
 var corecust;
 weight sw;
 run;
```

Exercici 15 Utilitzeu la mostra aleatòria simple desada a `assr2` per estimar el total de la variable `corecust` i un interval de confiança al 95%.

```
Proc surveymeans data=temp total=20309 sum clsum;
 var corecust;
 weight sw;
 run;
```

Exercici 16 Utilitzeu la mostra aleatòria simple desada a `assr1` per estimar la mitjana d'edat de la població, l'error estàndard de l'estimació i un interval de confiança al 95%.

```
Proc surveymeans data=temp total=20309 mean clm;
 var edat;
 weight sw;
 run;
```

Exercici 17 Fixeu-vos en els resultats de PROC SURVEYFREQ per fer una taula de freqüències de la variable `corecust`. Compareu el resultat amb un procediment PROC FREQ. Useu la base de dades `temp`.

```
Proc surveyfreq data=temp total=20309;
  table corecust;
  weight sw;
  run;

Proc freq data=temp ;
  table corecust;
  weight sw;
  run;
```

Part 3 Exercicis individuals

Exercici 18 Extraieu una mostra aleatòria simple de 180 individus de la base de dades `prob2` i deseu-la al directori `C:\TM` amb el nom `assr18`. Useu `seed` = 12345. Anoteu el número de l'última observació seleccionada a la mostra, utilitzant la numeració original del marc mostral.

Exercici 19 Estimeu la mitjana de durada (variable `t`) i un interval de confiança al 95%, a partir de l'anterior mostra `assr18`.

Exercici 20 A partir de la mostra `assr18`, estimeu el total de visitants que hi ha cada dia de la setmana. Compareu el resultat amb el nombre real que surt en el marc mostral `prob2`.

Exercici 21 Es vol fer una enquesta a un nombre suficient de visitants per conèixer la proporció dels que visiten el museu que no viuen en el mateix municipi. Una enquesta anterior estimava la proporció de visitants d'altres municipis en un 75%. Usant aquesta informació i població infinita, proposeu una mida de mostra suficient per garantir que l'estimació tindrà un interval de confiança amb un marge d'error inferior al 5%. Escriviu el programa per fer aquest càlcul.

Exercici 22 A l'arxiu del marc mostral `prob2` hi ha una variable `s` que indica que l'individu ha estat seleccionat per a una enquesta. Calculeu la mitjana de durada de la visita (variable `t`) i un interval de confiança al 95%, amb aquesta mostra. Compareu els resultats amb els de l'exercici 19.

Exercici 23 Utilitzeu la mostra identificada per la variable `s` en el marc mostral `prob2`, per dir quina és l'obra preferida dels visitants. Doneu un interval de confiança al 95%, per a la proporció de visitants que la prefereixen.

Exercici 24 Utilitzeu la mostra identificada per la variable `s` en el marc mostral `prob2`, i trobeu la proporció de visitants que prefereixen cadascuna de les obres (O1,O2,O3,O4,O5) amb el seu corresponent interval de confiança al 95%. Què es pot concloure sobre les preferències de les obres del museu?

Exercici 25 Utilitzeu la mostra identificada per la variable `s` en el marc mostral `prob2` i busqueu un model de regressió que expliqui la durada de la visita `t`. Seleccioneu entre els possibles regressors, les variables `obra` o categoritzacions de `d`.

Pràctica 2

Part 1 Selecció de mostres amb probabilitats desiguals

Recordem que treballem a la carpeta `C:\TM` amb els marcs mostrals `prob1` i `prob2`. Executeu:

```
libname prac 'C:\TM\';
```

Inicialment treballem amb la base de dades `prob1`.

Exercici 1 Extraiem una mostra de mida 100 i amb probabilitat de selecció proporcional a la variable `antig`. Desem-la com a base de dades SAS al directori `C:\TM` amb el nom `p2e1`.

```
Proc surveyselect data=prac.prob1 method=pps
     sampsize=100 seed=12345 out=prac.p2e1;
     size antig;
 run;
```

Exercici 2 Fem una gràfica de la probabilitat de selecció resultat de l'exercici anterior i la variable `antig`, per observar la relació lineal entre ambdues. Comprovem que la suma dels `SamplingWeight` no és igual a 1, ni a la mida de la població.

```
Proc gplot data = prac.p2e1;
  plot SelectionProb*antig;
 run;

Proc means data = prac.p2e1 sum;
  var SamplingWeight;
 run;
```

Exercici 3 Fem el mateix que a l'exercici 1, però ara afegim `outsize` en les opcions. Observem el contingut de les variables `TotalSize` i `SampleSize`. Desem només la variable `antig` a la base de dades de la mostra. Desem els resultats com a base de dades SAS al directori `C:\TM` amb el nom `p2e3`.

```
Proc surveyselect data=prac.prob1 method=pps outsize
     sampsize=100 seed=12345 out=prac.p2e3;
     size antig;
     id antig;
 run;
```

Exercici 4 Ara volem tornar a extreure una nova mostra de mida 1000, proporcional a l'antiguitat, però els clients que tenen una antiguitat superior a 10, volem que hi siguin amb tota seguretat. Desem els resultats com a base de dades SAS al directori `C:\TM` amb el nom `p2e4`. Comprovem el resultat amb una gràfica com a l'exercici anterior.

```
Proc surveyselect data=prac.prob1
     out=prac.p2e4 sampsize=1000 method=pps seed=12345
     certsize=10;
     size antig;
 run;

Proc gplot data = prac.p2e4;
  plot SelectionProb*antig;
 run;
```

Exercici 5 Ara veurem els diferents mètodes de selecció amb probabilitats desiguals. Desem les bases de dades SAS al directori `C:\TM` amb el nom `p2e5a`, `p2e5b`, `p2e5c`, etc. Usem un marc mostral petit amb només 10 registres per veure millor els resultats; la variable `x` és la referència per a les probabilitats desiguals. Agafem mostres de mida 5.

```
data prac.p2e5;
 input id x;
 cards;
 1 1
 2 1
 3 1
 4 2
 5 2
 6 2
 7 3
 8 3
 9 3
 10 4
 ;
proc print data=prac.p2e5;
 run;
```

a) Mètode amb probabilitats desiguals bàsic. La probabilitat de selecció és proporcional a la variable `x`.

```
Proc surveyselect data=prac.p2e5
  sampsize=5 method=pps out=prac.p2e5a seed=12345;
  size x;
  run;
Proc print data=prac.p2e5a;
 run;
```

b) Mètode Sampford. Selecciona sense reposició, perquè a partir de la segona unitat ja hi ha reposició; però, si es repeteixen unitats, es torna a extreure la mostra.

```
Proc surveyselect data=prac.p2e5
  sampsize=5 method=pps_sampford out=prac.p2e5b
  seed=12345;
  size x;
  run;
Title 'Sampford';
Proc print data=prac.p2e5b;
 run;
```

c) Mètode Chromy. Selecció seqüencial amb reposició mínima. Vol dir que el nombre de repeticions d'un registre pot ser igual a la part entera del nombre esperat de repeticions (calculat segons la mida del marc mostral i de la mida de la mostra) o, com a molt, el número enter següent.

```
Proc surveyselect data=prac.p2e5
  sampsize=5 method=pps_seq out=prac.p2e5c
  seed=12345;
  size x;
  run;
Title 'Chromy';
Proc print data=prac.p2e5c;
 run;
```

d) Mètode sistemàtic i amb probabilitats desiguals. Selecció seqüencial, però hi pot haver repetició.

```
Proc surveyselect data=prac.p2e5
  sampsize=5 method=pps_sys out=prac.p2e5d seed=12345;
  size x;
  run;
Title 'Prob. desiguals i sistematic';
Proc print data=prac.p2e5d;
 run;
```

e) Mètode amb probabilitats desiguals i reposició. Repetim-ho també amb mida de mostra 8.

```
Proc surveyselect data=prac.p2e5
  sampsize=5 method=pps_wr out=prac.p2e5e seed=12345;
  size x;
  run;
Title 'Prob. desiguals i reposicio';
Proc print data=prac.p2e5e;
 run;
```

Exercici 6 Usant la base de dades petita `prac.p2e5`, seleccionem una **mostra aleatòria simple sense reposició** amb una mida mínima de 3 donant una fracció de mostreig del 0, 1. Desem-la com a base de dades `p2e6`.

```
Proc surveyselect data = prac.p2e5
     method = srs nmin = 3 samprate=0.1
     out = prac.p2e6  seed=12345;
 run;
Proc print data=prac.p2e6;
 run;
```

Exercici 7 Usant la població `prob1`, seleccionem una **mostra aleatòria simple sense reposició** amb una fracció de mostreig del 0,05. Desem-la com a base de dades `p2e7`. La utilitzarem a l'apartat següent.

```
Proc surveyselect data = prac.prob1
     method = srs samprate=0.05 out = prac.p2e7
     seed=12345;
 run;
```

Part 2 Inferència en mostres aleatòries amb probabilitats desiguals

Exercici 8 Utilitzem la població coneguda de la base de dades inicial `prob1` i calculem el total de la variable `corecust` i la mitjana de la variable `edat`.

```
Proc means data=prac.prob1 sum n;
 var corecust;
 run;

Proc means data=prac.prob1 mean;
 var edat;
 run;
```

Exercici 9 Utilitzem ara la **mostra aleatòria simple sense reposició** de la base de dades inicial `prob1` que tenim desada com a base de dades `p2e7` i calculem l'estimació del total de la variable `corecust` i l'error estàndard de l'estimació.

```
Data prac.p2e9;
 set prac.p2e7;
 sw=20309/1016; *compte! posar midamostra;
 run;

Proc surveymeans data=prac.p2e9 total=20309 sum;
 var corecust;
 weight sw;
 run;
```

Exercici 10 Utilitzem una mostra aleatòria amb probabilitats desiguals proporcionals a la variable `edat`, que desem a `prac.p2e10` i amb mida de 1016, i després **ignorem les ponderacions**. Tornem a estimar el total de la variable `corecust`. Trobem evidència del biaix que tenim? Es deu a les probabilitats desiguals amb què s'ha obtingut la mostra o al fet d'haver ignorat les ponderacions?

```
Proc surveyselect data = prac.prob1
     method = pps sampsize=1016 out = prac.p2e10
     seed=12345;
     size antig;
 run;

Data temp;
 set prac.p2e10;
 sw=20309/1016;
 run;

Proc surveymeans data=temp total=20309 sum;
 var corecust;
 weight sw;
 run;
```

Exercici 11 Utilitzem la mostra aleatòria amb probabilitats desiguals que havíem desat a `prac.p2e10` i **ara no** ignorem les ponderacions. Tornem a estimar el total de la variable `corecust`. Comparem-ho amb els resultats anteriors.

```
Proc surveymeans data=prac.p2e10 total=20309 sum;
 var corecust;
 weight SamplingWeight;
 run;
```

Exercici 12 Mirem d'obtenir el resultat de l'exercici 11 amb els procediments estàndards usant la base de dades amb probabilitats desiguals `prac.p2e10`.

a) Inicialment comprovem el resultat de fer:

```
Proc means data = prac.p2e12 sum;
  var corecust;
  weight Samplingweight;
 run;
```

*b*1) Calculem la suma de `SamplingWeight`.

```
Proc means data = prac.p2e10 sum;
  var SamplingWeight;
  output out=prac.sumap2e12 sum=sumasw;
 run;
Proc print data=prac.sumap2e12;
 run;
```

*b*2) Recalculem pesos a partir de `SamplingWeight`.

```
Data prac.sumap2e12b;
  set prac.sumap2e12;
  do i=1 to 1016;
   output;
   end;
   keep sumasw;
   run;
Data prac.p2e12;
   merge prac.sumap2e12b prac.p2e10 ;
   pes=20309*SamplingWeight/sumasw;
   run;
```

*b*3) Utilitzem un procediment d'estimació del total estàndard amb pesos `pes` que sumen 20309.

```
Proc means data = prac.p2e12 sum;
  var corecust;
  weight pes;
 run;
```

Part 3 Exercicis individuals

Exercici 13 Repetiu els procediments anteriors (exercicis del 9 al 12) analitzant ara l'estimació de mitjana de la variable `edat`. Esperem que l'estimació de la mitjana d'edat segueixi sent esbiaixada quan utilitzem una mostra amb probabilitats desiguals proporcionals a la variable `edat`?

Exercici 14 Obteniu una **mostra aleatòria simple sense reposició** de la base de dades inicial `prob2` de mida 250 i calculeu la proporció de visitants que van al museu els diumenges amb un error estàndard de l'estimació. Compareu-ne el resultat amb el valor de la proporció real a la població `prob2`. Es té en compte la fracció de mostreig?

Exercici 15 Obteniu la **mostra aleatòria amb probabilitats desiguals sense reposició** de la base de dades inicial `prob2`, usant com a variable de referència la variable `t`, que indica la durada de la visita. La mostra ha de tenir una mida 250. Calculeu la proporció de visitants que van al museu els diumenges amb un error estàndard de l'estimació, tenint en compte les probabilitats desiguals. Compareu-ne el resultat amb el valor de la proporció real a la població `prob2` i amb l'obtinguda a l'exercici anterior.

Pràctica 3

Part 1 Selecció de mostres amb estratificació

Recordem que treballem a la carpeta `C:\TM` amb els marcs mostrals `prob1` i `prob2`. Executeu:

```
libname prac 'C:\TM\';
```

Inicialment, treballem amb la base de dades `prob1`.

Exercici 1 Extraiem una mostra de mida 100, estratificada per sexes (variable `gend`) i amb una mida 50 a cada estrat. Desem-la com a base de dades SAS al directori `C:\TM` amb el nom `p3e1`. Prèviament ordenem el marc mostral segons el sexe i fem una taula de freqüències per veure la mida poblacional dels estrats. Veurem que hi ha 625 casos que mostrejarà de l'estrat que té els casos sense sexe.

```
Proc sort data=prac.prob1;
 by gend;
 run;

Proc freq data=prac.prob1;
 table gend;
 run;

Proc surveyselect data=prac.prob1 method=srs n=50
  seed=12345 out=prac.p3e1;
  strata gend;
  run;

Proc freq data=prac.p3e1;
 table gend / missing ;
 run;
```

Exercici 2 A continuació mostregem amb mides diferents a cada estrat. Per exemple, 25 amb `gend` = 0, 75 amb `gend` = 1 i 2 dels "Missings". Desem la mostra a la base de dades de sortida al directori `C:\TM` amb el nom `p3e2`.

```
Proc surveyselect data=prac.prob1 method=srs n=(2 25 75)
  seed=12345 out=prac.p3e2;
  strata gend;
  run;

Proc freq data=prac.p3e2;
 table gend / missing;
 run;
```

Exercici 3 Si no volem cap individu a la mostra que tingui un valor "Missing" de la variable `gend`, aleshores hem d'eliminar abans aquest estrat del marc mostral. Desem el nou marc mostral amb el nom `prob1sm` a la llibreria `prac`. Després podem seleccionar una

mostra de 25 amb `gend` = 0, 75 amb `gend` = 1 i desar-la com a base de dades de sortida al directori `C:\TM` amb el nom `p3e3`.

```
Data prac.prob1sm;
 set prac.prob1;
 if (gend NE ".");
 run;

Proc surveyselect data=prac.prob1sm method=srs n=(25 75)
  seed=12345 out=prac.p3e3;
  strata gend;
  run;

Proc freq data=prac.p3e3;
 table gend;
 run;
```

Exercici 4 Observem que, a la mostra anterior, els individus seleccionats que pertanyen a estrats diferents tenen probabilitat de selecció diferent i pes mostral diferent segons l'estrat, mitjançant una taula de creuament.

```
Proc freq data=prac.p3e3;
 table gend*SelectionProb gend*samplingWeight / list;
 run;
```

Exercici 5 Suposem que volem una mostra del 10% de casos de cada estrat de sexe (variable `gend`). Treballem sobre el marc mostral que no conté "Missings". Desem el resultat en una base de dades al directori `C:\TM` amb el nom `p3e5` i fem una taula de freqüències per veure quants n'ha seleccionat de cada estrat.

```
Proc surveyselect data=prac.prob1sm method=srs
  samprate=0.10 seed=12345 out=prac.p3e5;
  strata gend;
  run;

Proc freq data=prac.p3e5;
 table gend;
 run;
```

Exercici 6 Podem donar les mides de mostra dels estrats també com a base de dades externa. Demanem ara una mostra de 25 amb `gend` = 0, 75 amb `gend` = 1. Desem la mostra resultant com a base de dades `p3e6`.

```
Data prac.mida;
 input gend _nsize_;
 datalines;
  0 25
  1 75
  ;
```

```
Proc surveyselect data=prac.prob1sm method=srs n=prac.mida
  seed=12345 out=prac.p3e6;
  strata gend;
  run;

Proc freq data=prac.p3e6;
 table gend;
 run;
```

Exercici 7 També podem demanar una fracció de mostreig diferent per a cada estrat. Per exemple, aquí podem demanar un 5% amb `gend` = 0 i un 1% amb `gend` = 1. Desem el resultat com a base de dades `p3e7` i l'usem tot seguit.

```
Data frac;
 input gend _rate_;
 cards;
  0 0.05
  1 0.01
  ;

Proc surveyselect data=prac.prob1sm method=srs
  samprate=frac seed=12345 out=prac.p3e7;
  strata gend;
  run;

Proc freq data=prac.p3e7;
 table gend;
 run;
```

Exercici 8 Ara, a part de l'estratificació, tenim en compte l'antiguitat del client; per tant, fem que la mostra també estigui ben distribuïda segons aquesta variable. Mantenim les fraccions de mostreig de l'exercici 7 i controlem per la variable `antig`. Desem el resultat com a base de dades `p3e8` i l'usem tot seguit. Poseu atenció que ara fem servir `method=sys`.

```
Proc sort data=prac.prob1sm;
 by gend antig;
 run;

Proc surveyselect data=prac.prob1sm method=sys
  samprate=frac seed=12345 out=prac.p3e8;
  strata gend;
  control antig;
  run;

Proc freq data=prac.p3e8;
 table gend;
 run;
```

Exercici 9 Mirem ara si hi ha diferències entre les mitjanes de la variable `antig` per al marc mostral `prob1sm`, a la mostra que no havia controlat per `antig` i que havíem desat com a `p3e7`, i a la mostra que tenia posat el control, anomenada `p3e8`.

```
Proc means data=prac.prob1sm mean  min max;
 var antig;
 run;

Proc means data=prac.p3e7 mean  min max;
 var antig;
 run;

Proc means data=prac.p3e8 mean  min max;
 var antig;
 run;
```

Exercici 10 Aparentment, no sembla que a l'exercici 8 haguem fet cap millora quan controlem per la variable `antig`, però si fem una anàlisi per estrats podem verificar què passa.

```
Proc means data=prac.prob1sm mean  min max;
 class gend;
 var antig;
 run;

Proc means data=prac.p3e7 mean  min max;
 class gend;
 var antig;
 run;

Proc means data=prac.p3e8 mean  min max;
  class gend;
 var antig;
 run;
```

Exercici 11 Usant al marc mostral `prob1sm`, seleccionem dues unitats per estrat de la variable `gend` amb el mètode de Brewer que utilitzi probabilitat de selecció proporcional a l'antiguitat (variable `antig`). Fixem-nos que, si desem la base de dades de sortida com a `p3e11` i la imprimim, podem veure la probabilitat de selecció conjunta.

```
Proc surveyselect data=prac.prob1sm method=pps_brewer
  seed=12345 out=prac.p3e11;
  size antig;
  strata gend;
  run;

proc print data=prac.p3e11;
 var gend SelectionProb SamplingWeight JtSelectionProb;
 run;
```

Per acabar amb aquesta part, podeu repetir l'exercici 11 fent servir el mètode de Murthy, poseu `method=pps_murthy`, i posant una variable a l'opció `size` o bé el mètode

de Sampford (`method=pps_sampford`) i en aquest últim cas podeu extreure més de dues unitats per estrat. Cal indicar també l'opció `size`, però en cas de voler més de dues unitats per estrat, aleshores es desen tantes probabilitats de selecció conjunta com possibles combinacions de parells d'observacions hi hagi, i es van desant a les variables `JtProb_1, JtProb_2,...`. **Atenció:** només funciona amb SAS 9.1.

Part 2 Inferència en mostres aleatòries estratificades

Exercici 12 Mirem ara si hi ha diferències entre les mitjanes de la variable `antig` i la variable `edat` per al marc mostral `prob1sm`, a la mostra estratificada autoponderada amb fracció de mostreig 10% que havíem desat com a `p3e5`, a la mostra estratificada no autoponderada que no havia controlat per `antig` i que havíem desat com a `p3e7`, i a la mostra no autoponderada que tenia posat el control de la variable `antig`, anomenada `p3e8`. La inferència utilitza el disseny mostral estratificat, però no té en compte la grandària poblacional de cada estrat. Mirem de completar una taula com la següent. Els intervals de confiança són al 95%.

Taula de la variable `antig`

Base de dades	`prob1sm`	`p3e7`	`p3e8`	`p3e5`
			Estratificades	
Disseny mostral	Marc		Contr.	Autop.
Mitjana `antig`				
Límit inferior IC	–			
Límit superior IC	–			

Taula de la variable `edat`

Base de dades	`prob1sm`	`p3e7`	`p3e8`	`p3e5`
			Estratificades	
Disseny mostral	Marc		Contr.	Autop.
Mitjana `edat`				
Límit inferior IC	–			
Límit superior IC	–			

```
Proc means data=prac.prob1sm mean min max;
  var antig edat;
  run;

Proc surveymeans data=prac.p3e7;
  stratum gend;
  var antig edat;
  run;

Proc surveymeans data=prac.p3e8;
  stratum gend;
  var antig edat;
  run;
```

```
Proc surveymeans data=prac.p3e5;
  stratum gend;
  var antig edat;
  run;
```

Exercici 13 Analitzem què hauria passat en les mostres anteriors si no s'haguessin analitzat segons el disseny estratificat. Comproveu amb les taules anteriors i veieu que les fórmules que assumeixen SRS (assr) no són correctes en aquest cas i acaben donant pràcticament el mateix resultat anterior, on no es tenia en compte la grandària de l'estrat.

```
Proc means data=prac.prob1sm mean min max;
  var antig edat;
  run;

Proc means data=prac.p3e7 mean clm maxdec=2;
  var antig edat;
  run;

Proc means data=prac.p3e8 mean clm maxdec=2;
  var antig edat;
  run;

Proc means data=prac.p3e5 mean clm maxdec=2;
  var antig edat;
  run;
```

Exercici 14 Corregim ara l'exercici 12 per incloure la mida de l'estrat. I revisem les taules; aquestes són les correctes. Les taules són a continuació perquè en pugueu anar apuntant els resultats i veure'n les diferències.

```
Proc freq data=prac.prob1sm;
  table gend;
  run;

Data totestr;
  input gend _total_;
  cards;
  0 4698
  1 14986
;

Proc surveymeans data=prac.p3e7 total=totestr;
  stratum gend /list;
  var antig edat;
  run;

Proc surveymeans data=prac.p3e8 total=totestr;
  stratum gend /list;
  var antig edat;
  run;
```

```
Proc surveymeans data=prac.p3e5 total=totestr;
  stratum gend /list;
  var antig edat;
  run;
```

Taula de la variable `antig`

Base de dades	`prob1sm`	`p3e7`	`p3e8`	`p3e5`
			Estratificades	
Disseny mostral	Marc		Contr.	Autop.
Mitjana `antig`				
Límit inferior IC	–			
Límit superior IC	–			

Taula de la variable `edat`

Base de dades	`prob1sm`	`p3e7`	`p3e8`	`p3e5`
			Estratificades	
Disseny mostral	Marc		Contr.	Autop.
Mitjana `edat`				
Límit inferior IC	–			
Límit superior IC	–			

Part 3 Exercicis individuals

Exercici 15 Efecteu un mostreig estratificat (mostra autoponderada) de l'arxiu inicial `prob2`, utilitzant la variable d'estratificació `d`, que fa referència al dia de la setmana. Trieu una fracció de mostreig del 5% dels visitants i calculeu la durada mitjana de la visita (variable `t`) amb aquesta mostra, utilitzant el nombre de visitants total per a cada valor de la variable `d` com a mida poblacional de l'estrat.

Exercici 16 Obteniu una **mostra aleatòria simple sense reposició** de la base de dades inicial `prob2` de la mateixa mida que la que heu obtingut a l'exercici 15. Comproveu si la mitjana de la variable `t` amb aquesta mostra simple és millor o pitjor que l'estratificada.

Exercici 17 A la base de dades inicial `prob2` hi ha una mostra d'individus que no se sap com s'han seleccionat, però estan identificats amb `s = 1`. Mireu si aquesta mostra tindria una mitjana de la variable `t` menys esbiaixada que les anteriors, suposant que es tracta d'una mostra aleatòria simple.

Exercici 18 Per acabar, fem una estratificació doble; en aquest cas, cada vegada hem de tenir en compte els estrats formats pel dia de la setmana i hora d'entrada anterior a les 12.00. A continuació teniu com definir aquesta segona categorització, que anomenem `aviat` i val 1, si l'individu arriba abans de les 12.00, i 0 en cas contrari. Un cop definida, fem els estrats com a combinacions de `d` i `aviat` (només cal posar `d aviat`, separades per blancs on abans posàvem `gend`). A tots els llocs on posàvem una variable d'estrat ara en posem dues. Calculeu l'estimació de la mitjana de `t` amb aquesta mostra estratificada amb 14 estrats i recordeu que cal ordenar un altre cop les dades, com es mostra a continuació.

```
Data prac.prob2b;
 set prac.prob2;
 aviat= ((ksubstr(entrada, anydigit(entrada),2))<12);
 run;
Proc freq data=prac.prob2b;
 table d*aviat;
 run;
Proc sort data=prac.prob2b;
 by d aviat;
 run;
```

Pràctica 4

Part 1 Anàlisi de mostres amb conglomerats

Ara necessitareu utilitzar els arxius que teniu comprimits a `prob3`. Els arxius que hi trobareu són:

```
HOGARES.TXT
PERSONAS.TXT

HOGARES_LISTA.doc
PERSONAS_LISTA1.doc

HOGARES_DESC.doc
PERSONAS_DESC.doc
```

Bona part d'aquesta pràctica requereix familiaritzar-se amb les dades. Les fonts provenen de "Encuesta sobre Discapacidades, Deficiencias y Estado de Salud", INE (1999). El detall de totes les variables el teniu als arxius "LISTA", però nosaltres no usarem totes les variables. En aquesta mostra es fa una estratificació per zona (províncies) i mida del municipi. La unitat primària és la llar, on s'entrevisten totes les persones que hi conviuen i l'individu és la unitat secundària. Els estrats no són autorepresentats.

IMPORTANT: En els primers exercicis d'aquesta pràctica inicialment no s'usa el pes mostral. El pes mostral es diu `factor` i és el mateix tant en l'arxiu `hogares` com en l'arxiu `personas`, però en altres enquestes podria no coincidir.

Recordem que treballem a la carpeta `C:\TM` . Copieu-hi els arxius anteriors. Un cop al SAS, executeu:

```
libname prac 'C:\TM\';
```

Exercici 1 Llegiu les dades de l'arxiu `PERSONAS.TXT` Calculeu el nombre de persones que tenim a la mostra i deseu la base de dades SAS a la carpeta `C:\TM` , amb el nom `prac.PERSONAS`.

```
Data prac.personas;
   INFILE 'C:\TM\PERSONAS.TXT';
   INPUT
      IDENTPER   1-7
      N1ORDEN    8-9
      EDAD      10-11
      SEXO      12-12
      PROV      13-14
      ESTUDIO   27-27
      RESP1     30-31
      SITU_PR   44-44
      ACT_EST   46-46
      SS_SNS    47-47
      MP_SNS    48-48
      MP_MP     49-49
```

```
      MUCOL_O   50-50
      AFIL_PR   51-51
      OTRAS_C   52-52
      SC_ASSN   53-53
      SC_ASON   54-54
      NOSEGU    55-55
      PRE_AU    56-56
      CERT_MI   57-57
      TMUNI     58-58
      THOGAR    59-60
      FACTOR    61-71;
run;

proc contents data= prac.personas;
run;
```

Exercici 2 Volem saber el nombre de famílies que hi ha en l'arxiu anterior. La identificació de les famílies es fa mitjançant el codi que hi ha enregistrat a la primera part de la variable `IDENTPER` (els 5 primers dígits). Desem el codi a `codifam`. L'operació pot tardar uns segons perquè la base de dades és gran. Observeu el contingut de la base de dades `prac.p4e2`. Amb la sintaxi següent es crea una base de dades amb el codi de llar i el nombre de membres que hi conviuen i que seran entrevistats.

```
data prac.PERSONAS;
  set prac.PERSONAS;
  codifam=floor(identper/100);
  run;

proc freq data=prac.personas  noprint ;
  table codifam /sparse out=prac.p4e2;
  run;

data prac.p4e2;
 set prac.p4e2;
 if count>0;
 drop percent;
 run;

proc contents data=prac.p4e2;
 run;
```

Exercici 3 Calculem la mitjana d'edat de la base de dades `prac.personas` amb un interval de confiança al 95% com si es tractés d'una mostra aleatòria simple. Anoteu el resultat i comparem aquesta dada amb la mitjana d'edat de la població espanyola de l'any 1999 (que podem trobar a INE, Padró 1999 (s'ha de calcular). Anoteu un o dos decimals més per veure les diferències que hi ha entre aquest resultat oficial i els procediments que farem a continuació en els exercicis posteriors.
Enllaç:

`http://www.ine.es/jaxi/tabla.do?path=/t20/e245/p04/a1999/l0/&file=00000002.px&type=pcaxis`

Fixeu-vos que la mostra té biaix per moltes raons: estrats sobreponderats o infraponderats, persones que conviuen en famílies de més membres tenen més probabilitats de ser presents a la mostra, absència de persones que no viuen en llars, com les institucionalitzades en presons, residències o convents, etc.

```
proc means data= prac.personas n nmiss mean clm;
 var edad ;
 run;
```

Exercici 4 Mirem ara l'estimació i l'interval de confiança al 95% per a la mitjana d'edat dels individus, si sabem que és un mostreig amb conglomerats identificats per `codifam`.

```
proc surveymeans data= prac.personas mean  clm;
 var edad;
 cluster codifam;
 run;
```

Exercici 5 Mirem ara l'estimació i l'interval de confiança al 95% per a la mitjana d'edat, si sabem que és un mostreig amb conglomerats identificats per `codifam` i estratificats per mida del municipi `tmuni`. L'ordenació pot tardar una mica.

```
proc sort data=prac.personas;
 by tmuni codifam;
 run;

proc surveymeans data= prac.personas mean  clm;
 var edad;
 strata tmuni;
 cluster codifam;
 run;
```

Exercici 6 Mirem ara l'estimació i l'interval de confiança al 95% per a la mitjana d'edat, si sabem que és un mostreig amb conglomerats identificats per `codifam` i estratificats per mida del municipi `tmuni`, amb **pesos mostrals** continguts a la variable `factor`.

```
proc surveymeans data= prac.personas mean  clm;
 var edad;
 strata tmuni;
 cluster codifam;
 weight factor;
 run;
```

Exercici 7 Mirem ara l'estimació i l'interval de confiança al 95% per a la mitjana d'edat, si sabem que és un mostreig amb conglomerats identificats per `codifam` i estratificats per província `prov` i mida del municipi `tmuni`, amb pesos mostrals `factor`.

```
proc sort data=prac.personas;
 by prov tmuni;
 run;
```

```
proc surveymeans data= prac.personas mean  clm;
 var edad;
 strata prov tmuni;
 cluster codifam;
 weight factor;
 run;
```

Una raó per la qual aquesta estimació encara pot ser lleugerament inferior a la d'Espanya és perquè aquesta mostra no conté les persones institucionalitzades i, per tant, és esbiaixada.

Exercici 8 Ara llegim l'arxiu de dades de les llars (HOGARES.TXT). Calculem quantes famílies hi ha en aquest arxiu i comparem-ho amb el resultat de l'arxiu de persones.

```
data prac.HOGARES;
   INFILE 'C:\TM\HOGARES.TXT';
   INPUT
      IDENTHOG   1-5
      PROV       6-7
      IM_TRACP   8-8
      IM_TRACA   9-9
      IM_PCON    10-10
      IM_PNCON   11-11
      IM_SUBDE   12-12
      IM_PRHIJ   13-13
      IM_PRREG   14-14
      IM_RTAPC   15-15
      IM_OTROS   16-16
      CD_FUEN    17-17
      IM_MENS    18-18
      PRES_SS    19-19
      PRES_IAN   20-26
      FACTOR     27-37;
  run;
```

Exercici 9 Mirem els ingressos totals mitjans **mensuals** per llar (variable `im_mens`) i els ingressos **anuals** per prestacions (variable `pres_ian`) que hi ha a cada llar. Mireu les definicions de les variables a l'arxiu HOGARES_DESC.doc. Assignem la marca de classe als intervals de la primera variable (mireu la sintaxi que hi ha a continuació). Mirem també el quocient dels ingressos per prestacions sobre la renda mensual (multiplicat per 12), que anomenem `ratio` i que és el percentatge d'ingressos anuals per prestació. Treballem sense cap disseny mostral. Interpreteu una mica el resultat que obtenim.

```
proc means data=prac.hogares n nmiss mean;
 var im_mens pres_ian;
 run;

data prac.hogares;
 set prac.hogares;
  ing_men=0;
  if (im_mens=1) then ing_men=22000;
```

```
  if (im_mens=2) then ing_men=(65+44)*1000/2;
  if (im_mens=3) then ing_men=(130+65)*1000/2;
  if (im_mens=4) then ing_men=(195+130)*1000/2;
  if (im_mens=5) then ing_men=(260+195)*1000/2;
  if (im_mens=6) then ing_men=(325+260)*1000/2;
  if (im_mens=7) then ing_men=(390+325)*1000/2;
  if (im_mens=8) then ing_men=(650+390)*1000/2;
  if (im_mens=9) then ing_men=(650+(650-390))*1000;
  if pres_ian='.' then pres_ian=0;
  if (ing_men>0) then ratio=pres_ian/(ing_men*12);
  run;

proc means data=prac.hogares n nmiss mean;
 var im_mens pres_ian ing_men ratio;
 run;
```

Si volem posar-hi el disseny mostral de l'estratificació per província i la ponderació, aleshores hem de fer:

```
proc sort data=prac.hogares;
  by prov;
  run;

proc surveymeans data=prac.hogares mean;
 strata prov;
 weight factor;
 var im_mens pres_ian ing_men ratio;
 run;
```

Exercici 10 Ara volem assignar a cada individu de l'arxiu de `prac.personas` l'ingrés mensual i per prestació que li correspondria a ell, segons quants membres té la seva família i l'ingrés que té la família en conjunt per aquests conceptes. Anomenem la variable de l'ingrés individual `im_mens_pp` i l'arxiu que conté la informació `prac.fusio`.

```
proc sort data=prac.hogares;
 by identhog;
 run;

proc sort data=prac.personas;
 by codifam;
 run;

data prac.fusio;
   merge prac.personas(rename=(codifam=identhog))
         prac.hogares(drop=factor);
   by identhog;
   im_mens_pp= ing_men/thogar;
   keep identhog prov tmuni thogar
        im_mens_pp ing_men factor;
   run;
```

Exercici 11 Calculem el total d'ingressos de tots els seleccionats a través de l'arxiu de les famílies (arxiu `prac.hogares` i variable `ing_men`) i a través de l'arxiu dels individus (arxiu `prac.fusio` i variable `im_mens_pp`). No tingueu en compte el pes, inicialment, i introduïu-lo després. No calculeu l'interval de confiança.

```
proc means data=prac.hogares sum;
 var ing_men;
 run;

proc means data=prac.fusio sum;
 var im_mens_pp;
 run;

proc means data=prac.hogares sum;
 var ing_men;
 weight factor;
 run;

proc means data=prac.fusio sum;
 var im_mens_pp;
 weight factor;
 run;
```

Exercici 12 Calculeu, usant `proc surveymeans`, el mateix que en l'exercici anterior. En l'arxiu `hogares` i la variable `im_mens`, useu mostreig estratificat per `provi` `tmuni`, així com `factor` i en l'arxiu `prac.fusio`, amb la variable `im_mens_pp` i sabent que es tracta d'un mostreig estratificat (per `prov` i `tmuni`), amb pesos (`factor`) i conglomerats identificats per `codifam`. Fixeu-vos que l'arxiu `hogares` no té `tmuni`, per tant s'ha de buscar a l'arxiu `prac.personas` i assignar-li.

```
proc sort data=prac.personas;
 by codifam;
 run;

proc means data=prac.personas noprint;
 by codifam;
 var tmuni;
 output out=prac.p4e12 mean=tmuni;
 run;

proc sort data=prac.hogares;
 by identhog;
 run;
proc sort data=prac.p4e12;
 by codifam;
 run;

data prac.hogares;
  merge prac.p4e12(rename=(codifam=identhog))
        prac.hogares;
 by identhog;
 run;
```

```
proc sort data=prac.hogares;
 by prov tmuni;
 run;

proc surveymeans data=prac.hogares sum clsum;
 var ing_men;
 strata prov tmuni;
 weight factor;
 run;

proc sort data=prac.fusio;
 by prov tmuni;
 run;

proc surveymeans data=prac.fusio sum clsum;
 var im_mens_pp;
 strata prov tmuni;
 cluster identhog;
 weight factor;
 run;
```

Exercici 13 Comparem ara l'interval de confiança per a la mitjana d'ingressos familiars si usem l'arxiu `hogares` i la variable `im_mens` amb un mostreig aleatori simple (useu `factor`), i l'arxiu `prac.fusio`, amb la variable `im_mens_pp` i sabent que es tracta d'un mostreig aleatori amb conglomerats identificats per `codifam`. Usem la mitjana de la variable `thogar` per veure la relació que hi ha entre els dos resultats anteriors i els seus intervals de confiança respectius.

```
proc surveymeans data=prac.hogares mean clm;
 var ing_men;
 weight factor;
 run;

proc surveymeans data=prac.fusio mean clm;
 var im_mens_pp;
 cluster identhog;
 weight factor;
 run;
```

Exercici 14 Comparem ara l'interval de confiança per a la mitjana d'ingressos familiars si usem l'arxiu `hogares` i la variable `im_mens` amb un mostreig estratificat per `prov` i `tmuni`, i l'arxiu `prac.fusio`, amb la variable `im_mens_pp` i sabent que es tracta d'un mostreig estratificat (per `prov` i `tmuni`), amb pesos (`factor`) i conglomerats identificats per `codifam`. Usem la mitjana de la variable `thogar` per veure la relació que hi ha entre els dos resultats anteriors i els seus intervals de confiança respectius, i interpretem una mica els resultats que obtenim.

```
proc sort data=prac.hogares;
 by prov tmuni;
 run;
```

```
proc surveymeans data=prac.hogares mean clm;
 var ing_men;
 strata prov tmuni;
 weight factor;
 run;

proc sort data=prac.fusio;
 by prov tmuni;
 run;

proc surveymeans data=prac.fusio mean clm;
 var im_mens_pp;
 strata prov tmuni;
 cluster identhog;
 weight factor;
 run;
```

Per finalitzar comproveu que si no ens adonem que el mostreig és per conglomerats i, erròniament, ignorem aquesta característica, aleshores infraestimem l'error comès. Fem la prova eliminant la instrucció `cluster identhog`. Fixem-nos com disminueix l'error estàndard.

```
proc surveymeans data=prac.fusio mean clm;
 var im_mens_pp;
 strata prov tmuni;
 weight factor;
 run;
```

Part 2 Exercicis individuals

Exercici 15 Calculeu el percentatge de llars que tenen algun membre amb certificat de minusvalidesa (variable `CERT_MI`) i un interval de confiança al 95%, incloent en el disseny mostral els pesos, l'estratificació per província i mida de municipi, i el mostreig per conglomerats de famílies en la primera etapa.

Exercici 16 Ens interessa quantificar les persones que viuen soles, amb el mateix disseny anterior. Hi ha diverses maneres d'aproximar-se a aquesta qüestió: veient quantes famílies tenen un sol membre (aleshores seria el percentatge de famílies d'un sol membre) o mirant quants individus viuen en una família que només té un membre (percentatge de persones que viuen soles). Si voleu podeu fer les dues. Justifiqueu si feu servir o no conglomerats i per què.

Exercici 17 Calculeu quin percentatge de llars tenen una ràtio d'ingressos per prestació superior al 50% i, per a les llars que reben prestacions, la mitjana de la ràtio. Poseu un interval de confiança al 95%.

Exercici 18 Estimeu el total de persones que tenen un certificat de minusvalidesa i viuen en una llar unifamiliar a Espanya. Doneu-ne el total amb un interval de confiança al 95%. A més, estimeu quantes d'aquestes persones tenen més de 75 anys (mitjançant la variable `EDAD`) i separeu-les segons si són homes o dones. En tots els casos, doneu l'interval de confiança.

Pràctica 5

Part 1 Regressió amb dissenys mostrals complexos. Simulació

Ara necessitareu utilitzar els arxius que teniu comprimits a `prob4`. Alternativament podeu generar les dades fent l'exercici 1. Recordem que treballem a la carpeta `C:\TM` . Copieu-hi els arxius anteriors. Un cop al SAS, executeu:

```
libname prac 'C:\TM\';
```

Exercici 1 Les dades simulades les hem generades manualment amb el programa següent. Si voleu, executeu el programa o, alternativament, copieu la base de dades `dades4` a la carpeta `C:\TM` . Fixeu-vos que la base de dades `prac.pesos4` té quatre variables anomenades `pes`, `altura`, `sexe` i `stra`.

```
proc iml;
a=20;
b=0.3;
c=10;
n1=5000;
n0=15000;
beta=a||b||c;
y1=normal(j(n1,1,12345))*12-78;
y0=normal(j(n0,1,12345))*9+160;
u=normal(j(n1+n0,1,12345))*3;
u2=normal(j(n1+n0,1,12345))*3;
xx=(y1//y0)||(j(n1,1,1)//j(n0,1,0));
x=j(n1+n0,1,1)||xx;
y=x*t(beta)+u;
stra=(y>0);
dades=y||xx||stra;
create prac.pesos4 from
  dades[colname={'pes', 'altura', 'sexe', 'stra'}];
append from dades;
quit;
```

Exercici 2 Mirem gràficament la forma de les dades veient com la variable dependent `pes` depèn de `sexe` i `altura`. Fixem-nos que, tal com ho hem generat, es compleix a nivell de la població que

$$pes_i = 20 + 0.3\_altura_i + 10\_sexe_i + u_i$$

```
proc sort data=prac.pesos4;
  by sexe;
  run;

proc gplot data= prac.pesos4;
  plot pes*altura=sexe;
  run;
```

Exercici 3 Verifiquem amb una regressió lineal que retrobem els valors dels paràmetres utilitzats en la generació de les dades. Anoteu el valor dels paràmetres estimats en la

regressió.

```
proc reg data=prac.pesos4;
 model pes=altura sexe;
 run;
```

Exercici 4 Ara seleccionem una mostra aleatòria simple de mida 200 de la població i estimem el model de regressió una altra vegada amb la mostra. Fem els càlculs sense i amb la correcció per mostra finita. Digueu què canvia en les dues estimacions i analitzeu-ne els resultats. Podem dir que la variable `sexe` té relació amb `pes`?

```
proc surveyselect data=prac.pesos4 method=srs n=200
 seed=12345 out=prac.p5e4;
 run;

proc reg data=prac.p5e4;
 model pes=altura sexe;
 title 'model 1 mostra aleatoria simple n=200';
 run;

proc surveyreg data=prac.p5e4 total=20000;
 model pes=altura sexe /solution;
 title 'model 1 mostra aleatoria simple n=200 N=20000';
 run;
```

Exercici 5 Repetiu l'exercici anterior amb una mida de mostra de 600 individus. Fem els càlculs sense i amb la correcció per mostra finita. Digueu què canvia en les dues estimacions i analitzeu-ne els resultats. Podem dir que la variable `sexe` no té relació amb `pes`?

```
proc surveyselect data=prac.pesos4 method=srs n=600
 seed=12345 out=prac.p5e5;
 run;

proc reg data=prac.p5e5;
 model pes=altura sexe;
 title 'model 2 mostra aleatoria simple n=600';
 run;

proc surveyreg data=prac.p5e5 total=20000;
 model pes=altura sexe /solution;
 title 'model 2 mostra aleatoria simple n=600 N=20000';
 run;
```

Exercici 6 Ara usem una mostra estratificada per `sexe` de mida 200 en el primer estrat i de 400 en el segon. No introduïm ponderacions (`Samplingweight`). Fem els càlculs sense i amb la correcció per mostra finita. Digueu què canvia en les dues estimacions i analitzeu-ne els resultats. Malgrat no haver recompost la mostra, hi ha biaix en les estimacions dels paràmetres?

```
proc sort data=prac.pesos4;
 by sexe;
 run;

proc surveyselect data=prac.pesos4 method=srs n=(200 400)
 seed=12345 out=prac.p5e6;
 strata sexe;
 run;

proc reg data=prac.p5e6;
 model pes=altura sexe;
 title 'model 3 mostra estratificada no ponderada';
 run;

data mida;
  input sexe _TOTAL_;
  * aqui sota poseu la mida de cada estrat on posa XXXXX;
  datalines;
  0 XXXXX
  1 XXXXX
  ;

proc surveyreg data=prac.p5e6 total=mida;
 model pes=altura sexe /solution;
 strata sexe;
 title 'model 3 mostra estrat. no ponderada N=20000';
 run;
```

Exercici 7 Ara usem la mateixa mostra estratificada per `sexe` de mida 200 en el primer estrat i de 400 en el segon, però ponderem utilitzant `Samplingweight`. Fem els càlculs sense i amb la correcció per mostra finita. Fixeu-vos si els resultats de l'estimació dels paràmetres milloren.

```
proc reg data=prac.p5e6;
 model pes=altura sexe;
 weight Samplingweight;
 title 'model 4 mostra estratificada ponderada';
 run;
```

```
data mida;
  input sexe _TOTAL_;
  * aqui sota, poseu la mida de cada estrat, on posa XXXXX;
  datalines;
  0 XXXXX
  1 XXXXX
  ;

proc surveyreg data=prac.p5e6 total=mida;
 model pes=altura sexe /solution;
  weight Samplingweight;
  strata sexe;
 title 'model 4 mostra estrat. ponderada N=20000';
 run;
```

La part final d'aquesta pràctica consisteix a usar una estratificació que depèn de la variable dependent.

Exercici 8 Ara usem una mostra estratificada per `stra` de mida 100 en el primer estrat i de 500 en el segon. No introduïm ponderacions (`Samplingweight`). Fem els càlculs sense i amb la correcció per mostra finita. Analitzeu-ne els resultats.

```
proc sort data=prac.pesos4;
 by stra;
 run;

proc surveyselect data=prac.pesos4 method=srs n=(100 500)
 seed=12345 out=prac.p5e8;
 strata stra;
 run;

proc reg data=prac.p5e8;
 model pes=altura sexe;
 title 'model 5 mostra estratificada no ponderada';
 run;

proc freq data=prac.pesos4;
 table stra;
 run;

data midastra;
  input stra _TOTAL_;
  * verifiqueu que aquesta sigui la mida dels estrats;
  datalines;
  0 3786
  1 6214
  ;
```

```
proc surveyreg data=prac.p5e8 total=midastra;
 model pes=altura sexe /solution;
 strata stra;
 title 'model 5 mostra estrat. no ponderada N=20000';
 run;
```

Exercici 9 Ara usem la mateixa mostra estratificada per `stra` de mida 100 en el primer estrat i de 400 en el segon, però ponderem utilitzant `Samplingweight`. Fem els càlculs sense i amb la correcció per mostra finita. Fixeu-vos si els resultats de l'estimació dels paràmetres milloren.

```
proc reg data=prac.p5e8;
 model pes=altura sexe;
 weight Samplingweight;
 title 'model 6 mostra estratificada ponderada';
 run;

data midastra;
  input stra _TOTAL_;
  datalines;
  0 3786
  1 6214
  ;

proc surveyreg data=prac.p5e8 total=midastra;
 model pes=altura sexe /solution;
  weight Samplingweight;
  strata stra;
 title 'model 6 mostra estratificada ponderada N=20000';
 run;
```

Part 2 Exercicis individuals

Exercici 10 Amb la base de dades de les entrades del museu havíem fet algunes regressions per explicar la durada de la visita en funció del dia d'entrada. Seleccioneu una mostra de mida 1000, estratificada amb afixació proporcional i no proporcional segons el dia de la setmana i executeu; repetiu les regressions emprant el disseny mostral i usant `proc surveyreg`. Comenteu els resultats i la diferència entre usar el disseny i no usar-lo, si féssim servir `proc reg`.

Pràctica 6

Part 1 Mètode bootstrap

Recordem que treballem a la carpeta `C:\TM` . Un cop al SAS, executeu:

```
libname prac 'C:\TM\';
```

Necessiteu utilitzar el procediment `proc surveyselect` per seleccionar diverses mostres d'una mateixa població i després aplicar el mètode bootstrap. Treballarem amb la població d'assegurats que havíem desat a l'arxiu `prac.prob1`.

Exercici 1 Seleccionem 50 mostres aleatòries simples de mida $n = 300$ del marc mostral `prac.prob1`. A continuació, i per a cada mostra, calculem la mediana d'edat dels assegurats que tenen 10 anys d'antiguitat o més. L'objectiu és donar un interval de confiança al 95%, per a aquest estadístic amb el mètode bootstrap.

```
libname prac 'C:\TM';

Proc surveyselect data=prac.prob1
     out=prac.p6e1 sampsize=300 method=srs rep=50
     seed=12345 stats;
 run;

Proc sort data=prac.p6e1;
 by replicate;
 run;

data prac.p6e2;
  set prac.p6e1;
  if (antig>10);
 run;

proc freq data=prac.p6e2;
 table replicate;
 run ;

proc means data=prac.p6e2 median;
 var edat;
 by replicate;
 output out=prac.p6e3 median=median;
 run;

proc print data=prac.p6e3;
 run;
```

```
proc iml;
 use prac.p6e3;
 read all var{replicate} into replicate;
 read all var{median} into median;
 print replicate median;
 n=nrow(median);
 print n;

start ordena(v);
  v0=v;
  v[rank(v)]=v0;
  rr=v;
  return(rr);
 finish;

 medianord=ordena(median);

i_limitinf=int(n*0.025);
i_limitsup=int(n*0.975);
print i_limitinf;
print i_limitsup;

me=sum(median)/n;
meinf=medianord[i_limitinf];
mesup=medianord[i_limitsup];
print 'mediana es' me;
print 'limit inferior es' meinf 'i superior es' mesup;

quit;
```

Part 2 Exercicis individuals

Exercici 2 Agafeu una mostra de mida $n = 1000$ del marc mostral que trobem a `prac.prob2` (visites a un museu). Mitjançant el mètode bootstrap (feu 10 submostres de mida 300), calculeu el mínim i el màxim, i un interval de confiança al 95% de la durada de la visita al museu per a cada dia de la setmana. Compareu el resultat amb el valor real del marc mostral.

Exercici 3 Mitjançant el mètode bootstrap (25 mostres de mida 1000), calculeu la variància de la ràtio prestacions per assistència dividit per ingressos mensuals a l'Enquesta de discapacitats (`prac.hogares`), per a aquelles llars que reben prestacions. Doneu un interval de confiança al 95% per a aquesta estimació.

Exercici 4 Mitjançant el mètode bootstrap (25 mostres de mida 100), calculeu la variància del percentil 25% dels ingressos familiars mensuals (recordem que havíem posat una marca de classe a cada nivell de renda) i dels ingressos individuals mensuals (valor que havíem assignat tenint en compte els ingressos familiars i el nombre d'individus de la llar). Doneu un interval de confiança al 95% per a ambdues estimacions.

Solucions

Pràctica 1

Part 1

Exercici 3. La variable en qüestió és `replicate`.

Part 2

Exercici 11. Total (`corecust`) = 10608; Mitjana (`edat`) = 47.2613.

Exercici 12. Std.Error = 0.4995.

Exercici 13. Total (`corecust`) = 11170.

Exercici 14. Total (`corecust`) = 11170, Std.Error = 1016.9469.

Exercici 15. Total (`corecust`) = 11170; IC(95%) = [9160.0436; 13180]

Exercici 16. Mitjana (`edat`) = 43.58, Std.Error = 1.6266; IC(95%) = [40.3524; 46.8076].

Part 3

Exercici 18. La darrera observació seleccionada a la mostra és la número 52240.

Exercici 19. Mitjana (`durada`) = 54.25; IC(95%) = [50.9754; 57.5246].

Exercici 20.

dia	Freq. a la mostra	Freq. a la població
1	292	439
2	876	800
3	584	856
4	2629	785
5	584	1450
6	14898	17800
7	32716	30450

Exercici 21. La mida de la mostra necessària és de $n = 1153$. La sintaxi necessària per obtenir-ho:

```
data dades;
 input alfa p Npob merror;
 cards;
  0.05 0.75 52580 0.05
 ;
run;
data mostra;
 set dades;
 z=probit(1-alfa/2);
 b=(merror/z)*(merror/z);
 n=(4*p*(1-p))/b;
 put 'mida mostra ' n;
run;
```

Exercici 22. Mitjana (`t`) = 52.1128; IC(95%) = [51.3161; 52.9095].

Exercici 23. L'obra preferida pels visitants és la número 3, en una proporció del 56.67%; IC (95%) = [55.007; 58.339].

Exercici 24.

Obra	Preferència	IC (95%)
O1	2.69%	[2.15 ; 3.24]
O2	19.08%	[17.76 ; 20.40]
O3	56.67%	[55.01 ; 58.34]
O4	17.11%	[15.84 ; 18.37]
O5	4.45%	[3.76 ; 5.14]

Pràctica 2

Part 1

Exercici 5.

a) Observacions incloses: 4,6,7,8,10
b) Observacions incloses: 4,7,8,9,10
c) Observacions incloses: 7,9,10,1,5
d) Observacions incloses: 2,5,7,8,10
e) Observacions incloses: 4,5,6,9,10

Exercici 6. Observacions incloses: 3,7,9.

Part 2

Exercici 8. Total (`corecust`) = 10608; mitjana (`edat`) = 47.2613.

Exercici 9. Total (`corecust`) = 10594, Std.Error = 310.3655.

Exercici 10. Total (`corecust`) = 11034.

Exercici 11. Total (`corecust`) = 10686.

Exercici 12.

a) Total (`corecust`) = 10685.98
b1) Suma (`SamplingWeight`) = 19918.06
b3) Total (`corecust`) = 10895.72

Part 3

Exercici 13.

Ex. 9. Mitjana (`edat`) = 47.6993, Std.Error = 0.5071
Ex. 10. Mitjana (`edat`) = 53.2756
Ex. 11. Mitjana (`edat`) = 47.4842
Ex. 12.a) Mitjana (`edat`) = 47.4842
Ex. 12.b1) Suma (`SamplingWeight`) = 19918.06
Ex. 12.b3) Mitjana (`edat`) = 47.4842

Exercici 14.

Proporció de visitants en diumenge

Estimador	Std. Error	Paràmetre poblacional
61.2%	3.08	57.91%

Exercici 15.

Proporció de visitants en diumenge

Estimador	Std. Error	Paràmetre poblacional
54.5656%	3.76	57.91%

Pràctica 3

Part 2

Exercici 12.

Taula de la variable `antig`

Base de dades	`prob1sm`	`p3e7`	`p3e8`	`p3e5`
		Estratificades		
Disseny mostral	Marc		Contr.	Autop.
Mitjana `antig`	9.54	9.41	9.85	9.09
Límit inferior IC	–	8.30	8.81	8.67
Límit superior IC	–	10.52	10.89	9.52

Taula de la variable `edat`

Base de dades	`prob1sm`	`p3e7`	`p3e8`	`p3e5`
		Estratificades		
Disseny mostral	Marc		Contr.	Autop.
Mitjana `edat`	47.08	48.07	48.37	46.44
Límit inferior IC	–	46.38	46.72	45.75
Límit superior IC	–	49.76	50.03	47.13

Exercici 14.

Taula de la variable `antig`

Base de dades	`p3e7`	`p3e8`	`p3e5`
	Estratificades		
Disseny mostral		Contr.	Autop.
Mitjana `antig`	9.43	9.85	9.10
Límit inferior IC	8.32	8.83	8.69
Límit superior IC	10.54	10.87	9.50

Taula de la variable `edat`

Base de dades	`p3e7`	`p3e8`	`p3e5`
	Estratificades		
Disseny mostral		Contr.	Autop.
Mitjana `edat`	48.07	48.38	46.44
Límit inferior IC	46.40	46.75	45.78
Límit superior IC	49.73	50.00	47.09

Part 3

Exercici 15. Mitjana (`durada`) = 51.7370; IC (95%) = $[51.5945; 51.8795]$.

Exercici 16. Seed = 12345: Mitjana (`durada`) = 51.68; IC (95%) = $[50.80; 52.56]$.

Exercici 17. Seed = 12345: Mitjana (`durada`) = 52.11; IC (95%) = $[51.29; 52.93]$.

Exercici 18. Seed = 12345: Mitjana (`durada`) = 51.7004; IC (95%) = $[51.5602; 51.8406]$.

Pràctica 4

Part 1

Exercici 1. $n = 218185$ persones a la mostra.

Exercici 2. $n = 70500$ llars a la mostra.

Exercici 3. Mitjana (`edat`) = 39.56; IC (95%) = [39.47; 39.66]. A l'INE, l'edat mitjana és de 39.14 anys.

Exercici 4. Mitjana (`edat`) = 39.56; IC (95%) = [39.44; 39.68].

Exercici 5. Mitjana (`edat`) = 39.56; IC (95%) = [39.44; 39.68].

Exercici 6. Mitjana (`edat`) = 38.85; IC (95%) = [38.71; 38.99].

Exercici 7. Mitjana (`edat`) = 38.85; IC (95%) = [38.71; 38.99].

Exercici 8. $n = 11092$ milions de llars.

Exercici 9. Mitjana (`im_mens`) = 4.0528; mitjana (`pres_ian`) = 178377.56.

Exercici 11. Total (`ing_men`) = 11092357000; total (`im_mens_pp`) = 11092357000.

Exercici 12. Total (`ing_men`) = $2.0927154E12$; total (`im_mens_pp`) = $2.0927154E12$.

Exercici 13.
Mitjana (`ing_men`) = 161657, IC (95%) = [160358; 162956];
Mitjana (`im_mens_pp`) = 52870, IC (95%) = [52443; 53297].

Exercici 14.
Mitjana (`ing_men`) = 161657, IC (95%) = [160389; 162925];
Mitjana (`im_mens_pp`) = 52870, IC (95%) = [52456; 53284];
Si no es tenen en compte els conglomerats, std.error (`im_mens_pp`) = 126.2534.

Part 2

Exercici 15. Percentatge de llars amb un o més membres amb certificat de minusvalidesa: 7.732%; IC (95%) = [7.507; 7.957].

Exercici 16.
Percentatge de llars unifamiliars: 14.08%, amb un IC (95%) = [13.7616; 14.3909];
Percentatge d'individus que viuen sols: 4.60%, amb un IC (95%) = [4.4889; 4.7184].

Exercici 17.
Percentatge de llars amb `ratio` superior a 0.5: 0.1135%, IC (95%) = [0.0825%; 0.1445%];
Mitjana `ratio` = 0.1245, IC (95%) = [0.1088; 0.1401].

Exercici 18.
Total de persones amb certificat de minusvalidesa i en llars unifamiliars = 84170, amb un IC (95%) = [75636; 92704];
Total de persones amb certificat de minusvalidesa i en llars unifamiliars, de més de 75 anys = 19416, amb un IC (95%) = [15516; 23317]: Homes - Total = 4986, IC (95%) = [3229; 6744]; Dones - Total = 14430, IC (95%) = [10942; 17918].

Pràctica 5

Part 1

Exercici 3. Constant: $\hat{\beta}_0 = 19.9192$, altura: $\hat{\beta}_1 = 0.3008$, sexe: $\hat{\beta}_2 = 10.1567$.

Exercici 4.

Sense correcció - Constant: $\hat{\beta}_0 = 21.5558$, altura: $\hat{\beta}_1 = 0.2906$, sexe: $\hat{\beta}_2 = 7.9222$;
Amb correcció - Constant: $\hat{\beta}_0 = 21.5558$, altura: $\hat{\beta}_1 = 0.2906$, sexe: $\hat{\beta}_2 = 7.9222$.

Exercici 5.

Sense correcció - Constant: $\hat{\beta}_0 = 19.1678$, altura: $\hat{\beta}_1 = 0.3052$, sexe: $\hat{\beta}_2 = 11.4606$;
Amb correcció - Constant: $\hat{\beta}_0 = 19.1678$, altura: $\hat{\beta}_1 = 0.3052$, sexe: $\hat{\beta}_2 = 11.4606$.

Exercici 6.

Sense correcció - Constant: $\hat{\beta}_0 = 17.8920$, altura: $\hat{\beta}_1 = 0.3135$, sexe: $\hat{\beta}_2 = 13.2694$;
Amb correcció - Constant: $\hat{\beta}_0 = 17.8920$, altura: $\hat{\beta}_1 = 0.3135$, sexe: $\hat{\beta}_2 = 13.2694$.

Exercici 7.

Sense correcció - Constant: $\hat{\beta}_0 = 18.0366$, altura: $\hat{\beta}_1 = 0.3126$, sexe: $\hat{\beta}_2 = 13.0550$;
Amb correcció - Constant: $\hat{\beta}_0 = 18.0366$, altura: $\hat{\beta}_1 = 0.3126$, sexe: $\hat{\beta}_2 = 13.0550$.

Exercici 8.

Sense correcció - Constant: $\hat{\beta}_0 = 33.9208$, altura: $\hat{\beta}_1 = 0.2287$, sexe: $\hat{\beta}_2 = -8.3123$;
Amb correcció - Constant: $\hat{\beta}_0 = 33.9208$, altura: $\hat{\beta}_1 = 0.2287$, sexe: $\hat{\beta}_2 = -8.3123$.

Exercici 9.

Sense correcció - Constant: $\hat{\beta}_0 = 19.2693$, altura: $\hat{\beta}_1 = 0.3021$, sexe: $\hat{\beta}_2 = 12.1019$;
Amb correcció - Constant: $\hat{\beta}_0 = 19.2693$, altura: $\hat{\beta}_1 = 0.3021$, sexe: $\hat{\beta}_2 = 12.1019$.

Part 2

Exercici 10.

Afixació proporcional:

Amb correcció - Categoria de referència (diumenge): $\hat{\beta}_0 = 69.9240$, dilluns: $\hat{\beta}_1 = 45.4093$, dimarts: $\hat{\beta}_2 = -31.2365$, dimecres: $\hat{\beta}_3 = -40.6299$, dijous: $\hat{\beta}_4 = -25.9907$, divendres: $\hat{\beta}_5 = -55.2812$, dissabte: $\hat{\beta}_6 = -45.5199$;

Sense correcció - Categoria de referència (diumenge): $\hat{\beta}_0 = 69.9240$, dilluns: $\hat{\beta}_1 = 45.4093$, dimarts: $\hat{\beta}_2 = -31.2365$, dimecres: $\hat{\beta}_3 = -40.6299$, dijous: $\hat{\beta}_4 = -25.9907$, divendres: $\hat{\beta}_5 = -55.2812$, dissabte: $\hat{\beta}_6 = -45.5199$.

Afixació no proporcional:

Amb correcció - Categoria de referència (diumenge): $\hat{\beta}_0 = 69.7568$, dilluns: $\hat{\beta}_1 = 49.8207$, dimarts: $\hat{\beta}_2 = -30.1230$, dimecres: $\hat{\beta}_3 = -39.8201$, dijous: $\hat{\beta}_4 = -24.9680$, divendres: $\hat{\beta}_5 = -55.2145$, dissabte: $\hat{\beta}_6 = -45.2849$;

Sense correcció - Categoria de referència (diumenge): $\hat{\beta}_0 = 69.7568$, dilluns: $\hat{\beta}_1 = 49.8207$, dimarts: $\hat{\beta}_2 = -30.1230$, dimecres: $\hat{\beta}_3 = -39.8201$, dijous: $\hat{\beta}_4 = -24.9680$, divendres: $\hat{\beta}_5 = -55.2145$, dissabte: $\hat{\beta}_6 = -45.2849$.

Comentari sobre els errors estàndards de les estimacions: en els models amb correcció que tenen en compte el mostreig dut a terme, els errors estàndards de les estimacions són considerablement menors que els errors estàndards de les estimacions en els models que no tenen en compte el mostreig.

Pràctica 6

Part 1

Exercici 1.
50 rèpliques: Mediana `edat` = 56.82, IC (95%) = [54; 60];
200 rèpliques: Mediana `edat` = 56.9825, IC (95%) = [53; 61].

Part 2

Exercici 3. Ràtio prestacions per assistència respecte a ingressos mensuals: Var(`ratio`) = 0.133, IC (95%) = [0.042; 0.224]. La sintaxi usada per a la resolució de l'exercici és:

```
proc surveyselect data=prac.hogares
 out=p6e3 sampsize=1000 method=srs rep=25
 seed=12345 stats;
run;
proc sort data=p6e3;
 by replicate;
run;

/*Seleccionem les llars que reben prestacions:*/
data p6e3;
 set p6e3;
 if (pres_ian>0);
run;
proc freq data=p6e3;
 table replicate;
run ;

/*Càlcul de l'estimació puntual*/
proc means data=p6e3 mean;
 var ratio;
 by replicate;
 output out=p6e3b mean=mean;
run;

/*Utilitzant el procediment de l'exercici 1, calculem l'IC (al 95%).*/
proc iml;
 use p6e3b;
 read all var{replicate} into replicate;
 read all var{mean} into mean;
 print replicate mean;
 n=nrow(mean);
 start ordena(v);
  v0=v;
  v[rank(v)]=v0;
  rr=v;
  return(rr);
 finish;
 meanord=ordena(mean);
 i_limitinf=int(n*0.025)+1;
 i_limitsup=int(n*0.975);
 me=sum(mean)/n;
 meinf=meanord[i_limitinf];
 mesup=meanord[i_limitsup];
quit;
```

Bibliografia

Ardilly, P.; Tillé, Y. *Sampling Methods: Exercises and Solutions*. Springer, 2005.

Fuller, C.H. *Weighting to adjust for survey nonresponse*. Public Opinion Quarterly, num. 38, 239-46, 1974.

Heeringa, S.G.; West, B.T.; Berglund, P.A. *Applied Survey Data Analysis*. CRC Press, 2010. Chapman & Hall/CRC Statistics in the Social and Behavioral Science.

Holt, D.; Elliot, D. *Methods of weighting for unit non-response*. The Statistician, num. 40, 333-42, 1991.

Lavallée P. *Le sondage indirect*. Editions de l'Université de Bruxelles, 2002.

Lohr, S. *Sampling: Design and Analysis*. Brooks/Cole, 2010.

Lohr, S. *Solutions Manual for Sampling: Design and Analysis* [en línia]. Brooks/Cole, 2010.

Lohr, S. *Computer Programs for Sampling: Design and Analysis* [en línia]. Brooks/Cole, 2010.

Lumley, T. *Complex Surveys: A Guide to Analysis Using R*. Wiley, 2010. Wiley Series in Survey Methodology.

Mandell, L. *When to weight: Determining nonresponse bias in survey data*. Public Opinion Quarterly, num.38, 247-51, 1974.

Pfeffermann, D. *The role of sampling weights when modeling survey data*. International Statistical Review, num. 61, 317-37, 1993.

Pfeffermann, D. *The use of sampling weights for survey data analysis*. Statistical Methods in Medical Research, num. 5, 239-61, 1996.

Särndal, C.-E.; Swensson, B.; Wretman, J. *Model assisted survey sampling*. Springer, 1997.

SAS Institute Inc. *SAS/OR® 9 User's Guide: Mathematical Programming*. SAS Institute Inc., 2002.

Skinner, C.J.; Holt, D.; Smith, T.M.F. *Analysis of complex surveys*. Wiley, 1989.

Tillé, Y. *Sampling Algorithms*. Springer, 2006.

Tillé, Y. *Teoría de muestreo*, traducció de *Théorie des sondages*.

Tillé, Y. *Théorie des sondages*. Dunod, 2001.

UCLA ATS Stat Consulting Group. *UCLA Stat Computing Portal: Survey Data Analysis Portal* [en línia]. UCLA ATS Stat Consulting Group [Data de consulta: 04/2011]. Disponible a: <http://statcomp.ats.ucla.edu/survey/>.