**SESSIÓ 7 – ESTADÍSTICA DESCRIPTIVA I**

\*--------------------------------------------------------------------------;

\*I PROC FREQ (VARIABLES CATEGORICAS)

DISTRIBUCIONES DE FREQ UNIV Y BIVARIANTES

Y MULTIPLES CONTRASTES RELACIONADOS CON TABLAS BIVARIANTES;

\*--------------------------------------------------------------------------;

\*PROC FREQ (sintaxi minima);

title1 'Distribucio univariant de la variable sex';

title2 'Opcio missprint';

**proc** **freq** data = sashelp.cars;

table origin;

**run**;

title 'Tablas de contingencia';

**proc** **freq** data = sashelp.cars;

table origin\*type;

**run**;

title 'Tablas de contingencia nomes amb % de la taula';

**proc** **freq** data = sashelp.cars;

table origin\*type/norow nocol nofreq;

**run**;

title 'Tablas de contingencia només absoluts';

**proc** **freq** data = sashelp.cars;

table origin\*type/norow nocol nopercent;

**run**;

\* (ASSOCIACIÓ PERFECTA);

title 'Tablas de contingencia con comando weight';

**data** ex\_pes;

input a b pes @@;

datalines;

1 1 100 1 2 0 2 1 0 2 2 100

;

**proc** **freq** data = ex\_pes;

table a\*b /chisq;

weight pes;

**run**;

\* (NO HI HA ASSOCIACIÓ);

title 'Tablas de contingencia con comando weight';

**data** ex\_pes2;

input a b pes @@;

datalines;

1 1 100 1 2 100 2 1 100 2 2 100

;

**proc** **freq** data = ex\_pes2;

table a\*b /chisq;

weight pes;

**run**;

\* (ASSOCIACIÓ);

title 'Tablas de contingencia con comando weight';

**data** ex\_pes4;

input a b pes @@;

datalines;

1 1 100 1 2 100 2 1 100 2 2 180

;

**proc** **freq** data = ex\_pes4;

table a\*b /chisq;

weight pes;

**run**;

\*associacio ente atributs i test de la chi;

**DATA** TAULA1;

INPUT GENERE $ FUMA $ CASOS;

DATALINES;

h Si 100

h No 2

d Si 3

d No 110

;

/\*OPCIO CHISQ DEL COMANDAMANT TABLE I COMANDAMENT WEIGHT EN EL CAS QUE

LES DADES LES TINGUME JA EN FORMA DE TAULA\*/

**PROC** **FREQ** DATA = TAULA1;

TABLE GENERE\*FUMA /CHISQ;

WEIGHT CASOS;

**RUN**;

/\*EL VALOR DE LA CHI ES 195.44 amb un p-value Practicamente 0

rebutgem h0: de no exist d'assocciació, i per tant hi ha assoc

\*valores missing en tablas cruzadas;

data miss;

input x1 x2;

cards;

1 1

1 .

2 2

2 3

2 .

2 .

1 3

1 .

;

proc freq data = miss;

table x2\*x1; /\*no es contemplen les observacions miss\*/

table x2\*x1/missing; /\*no es contemplen les observacions miss\*/

**run**;

\*--------------------------------------------------------------------------;

\*II PROC MEANS (VARIABLES NUMÉRICAS)

ANÁLISIS DESCRIPTIVOS Y ALGUN TIPO DE CONTRASTE PARA VARIABLES CONTINUAS;

\*--------------------------------------------------------------------------;

\* PROC MEANS (sintaxi minima);

**PROC** **MEANS** DATA = SASHELP.CLASS;

**RUN**;

\* CLASS (Descriptiva por grupos);

**PROC** **MEANS** DATA = SASHELP.CLASS PRINTALLTYPES;

CLASS SEX;

VAR HEIGHT;

**RUN**;

\* INTERVALS DE CONFIANÇA AL 90% (con o sin class);

**PROC** **MEANS** DATA = SASHELP.CLASS ALPHA = **.10** CLM MEAN maxdec = **2** ;

VAR HEIGHT;

**RUN**;

\*2.7 INTERVALS DE CONFIANÇA AL 90% POR GRUPOS;

**PROC** **MEANS** DATA = W2.hosp01 ALPHA = **.10** CLM MEAN maxdec = **2** PRINTALLTYPES;

VAR HEIGHT;

CLASS SEX;

**RUN**;

/\*

statistic-keyword(s)

specifies which statistics to compute and the order to display them in the output. The available keywords in the PROC statement are

Descriptive statistic keywords

CLM NMISS

CSS RANGE

CV SKEWNESS|SKEW

KURTOSIS|KURT STDDEV|STD

LCLM STDERR

MAX SUM

MEAN SUMWGT

MIN UCLM

MODE USS

N VAR

Quantile statistic keywords

MEDIAN|P50 Q3|P75

P1 P90

P5 P95

P10 P99

P20 P30

P40 P60

P70 P80

Q1|P25 QRANGE

Hypothesis testing keywords

PROBT|PRT T

Default:N, MEAN, STD, MIN, and MAX

\*/

\*--------------------------------------------------------------------------;

\* III PROC CORR (VARIABLES NUMÉRICAS)

OBTENER MATRICES DE CORRELACIONES

VARIANZAS Y COVARS Y CONTRASTES;

\*--------------------------------------------------------------------------;

\*PROC CORR (sintaxi minima);

**PROC** **CORR** DATA = SASHELP.CLASS;

VAR HEIGHT WEIGHT;

**RUN**;

TITLE 'Comandament by';

**PROC** **SORT** DATA = SASHELP.CLASS OUT = CLASS\_OR;

BY SEX;

**RUN**;

**PROC** **CORR** DATA = CLASS\_OR ;

VAR HEIGHT WEIGHT;

BY SEX;

**RUN**;

TITLE 'Opció cov';

**PROC** **CORR** DATA =SASHELP.CARS cov;

VAR CYLINDERS WEIGHT HORSEPOWER ;

**RUN**;

\*4.4;

TITLE 'Comandament with';

**PROC** **CORR** DATA =SASHELP.CARS cov;

VAR CYLINDERS WEIGHT;

WITH HORSEPOWER;

**RUN**;

\*4.5 Per introduir pesos;

TITLE 'Comandament FREQ';

**data** correla;

input x1 x2 p;

cards;

1 7 200

2 6 120

3 5 230

4 4 450

;

**PROC** **CORR** DATA = correla cov;

VAR x1 x2;

FREQ p;

**RUN**;

\*--------------------------------------------------------------------------;

\* EXERCICI 1

\*--------------------------------------------------------------------------;

/\* Llegir dades del estu96.dat per crear labase estu96.sas7bdat

Llegir c7 i c8 com a cadena per conservar els possibles zeros que tenen davant

i després obtenir la distribució de freq de la variable nombre pulsacions per minut agrupada

menys de 60, 60 però menys de 80 i 80 o més i tambe una anàlisi de les vars pes i altura

segons genere\*/

**data** estu96;

infile 'd:\estu96.dat';

input x1 $ x2-x6 (x7 x8) (:$3.) x9 x10 $10. ;

label x1 = Gènere x2 = Edat x3 = Pes x4 = Alçada

x5 = Pulsacions x6 = Germans x7 = **3** nombres aleatoris x8 = **3** xifres DNI

x9 = Com escriu? x10 = Esport;

if x8 = '9' then x8 = ' ';

if x2 = **99** then x2 = **.**;

if x5 = **999** then x5 = **.**;

**run**;

**proc** **contents** data = estu96;

**run**;

\*descriptiu per veure si hi alguns valors "anomals" en les "numèriques";

options nocenter nodate nonumber

title;

**proc** **means** data = estu96;

**run**;

/\*x2 (edat molts o pocs i decidir q fem) x5 (pulsacions) i x8 (dni) ???\*/

**proc** **print** data=estu96;

**run**;

**proc** **format**;

value x2f low-<**40** = 'menys de 40'

other = '40 o més de 40';

value x5f **48**-**120** = 'normal' ;

**run**;

**proc** **freq** data = estu96;

table x2 x5 x8;

format x2 x2f. x5 x5f. ;

**run**;

/\*pel que fa a les pulsacions el 999 era un missing d'usuari'\*/

/\* obtindrem la distrb de freq de la variable nombre pulsacions per minut agrupada segons:

menys de 60, 60 però menys de 80 i 80 o més\*/

**proc** **format**;

value x5f low-<**60** = menys de **60**

**60**-<**80** = '[60-80)'

**80**-high = **80** o més;

**run**;

**proc** **freq** data = estu96;

table x5 ;

format x5 x5f.;

**run**;

/\*tambe una anàlisi de les vars pes (x3) i altura (x4) segons genere (x1)\*/

**PROC** **MEANS** DATA = estu96;

var x3 x4;

class x1;

**RUN**;

/\*analisi de la variable pes pero només pels que pesen menys de 60 Kg sense classificar\*/

**PROC** **MEANS** DATA = estu96 (where = (x3<**60**)) ;

var x3 ;

**RUN**;

/\*analisi dels pes i alt pero només pel grup de les dones\*/

**PROC** **MEANS** DATA = estu96 (where = (x1 = 'd')) ;

var x3 x4 ;

**RUN**;

\*--------------------------------------------------------------------------;

/\*utilitzacio de bucles amb sas COMANDAMENT DO en un bloc data\*/

\*--------------------------------------------------------------------------;

\*per a crear una bdd "a partir de la nada";

\*exemple generar una sèrie de nombres aletoris 200 OBS D'1 N(0,1);

**DATA** ALEAT (drop = i);

DO i = **1** to **200**;

z = rannor(**2565598**);/\*el valor de llavor\*/

OUTPUT;

END;

**RUN**;

\*Lectura de més d'una observ per línia;

/\*6 observacions de 2 individus\*/

**DATA** mult;

input genere edat @@;

datalines;

1 23 2 34 1 35 2

66

1 45 1 22

;

**data** notes;

input alu $ n1-n3;

datalines;

alu1 7 4 8

alu2 8 8 9

alu3 4 1.5 9

;

**proc** **print** data = notes;

**run**;

\*Observacions repetides passar a un fitxer amb 3 variable

alumnes numero\_test nota\_test ;

**data** notes2;

input alu $ @;

do test = **1** to **3**;

input nota @@;

output;

end;

datalines;

alu1 7 4 8

alu2 8 8 9

alu3 4 1.5 9

;

**proc** **print** data = notes2;

**run**;

\*--------------------------------------------------------------------------;

\* EXERCICI 2

\*--------------------------------------------------------------------------;

/\*fitxer padro.txt \*/

/\*Exercici:

Llegir les dades del fitxer PADRO.TXT

Són dades del padró d’un municipi, amb informació d’altes” i baixes” de l’any 2008

(primera inscripció el 2 de gener de 2008 fins última del 28 de desembre de 2008)

Cada inscripció consta de:

data\_inscripció

codi\_unitat\_familiar

nucli

barri

num\_persones\_inscrites(k): k vegades dni

gènere (1=home / 2=dona)

data\_neixement

Alta\_o\_Baixa(A=alta B=baixa)

DATA padro;

INFILE 'd:\padro.txt';

INPUT data\_insc DDMMYY8. codi\_uf nucli barri num ;

DO i=1 to num;

INPUT dni sexe data\_naix DDMMYY8. A\_B $ @@;

dies = data\_insc - data\_naix;

edat = dies/365;

edat\_anys\_r = round(edat,1);

edat\_anys\_i = int(edat);

OUTPUT;

END;

FORMAT data\_insc DDMMYY8. data\_naix DDMMYY8.;

run;

PROC PRINT data = padro;

RUN;

\*Indiqueu:

- Taules de freqüències de les variables nucli, barri, gènere i Alta\_o\_Baixa;

PROC FREQ data = padro;

TABLES sexe A\_B;

TABLES sexe\*A\_B;

RUN;

PROC FORMAT;

VALUE fsexe 1='home' 2='dona';

run;

PROC FREQ data = padro;

FORMAT sexe fsexe.;

TABLES nucli barri sexe A\_B;

RUN;

\*- Taula creuada de gènere i Alta\_o\_Baixa

És independent la distribució d’altes i baixes segons el gènere?;

PROC FREQ data = padro;

FORMAT sexe fsexe.;

TABLES sexe\*A\_B/chisq;

RUN;

\*- Taula creuada de gènere i barri

És independent la distribució per barris segons el gènere?;

PROC FREQ data = padro;

FORMAT sexe fsexe.;

TABLES sexe\*barri/chisq;

RUN;

\*- Mitjana i desviació de l’edat en les altes i en les baixes;

PROC MEANS DATA=padro;

VAR edat\_anys\_i;

CLASS A\_B;

RUN;