# 8. If and basic statistics procedures

**Daniel Alconchel Vázquez**

**Exercise 1.** Create a data set which contains observations 4 i 20 from *sashelp.air*. Use IF, value of variable _N_ from PDV-vector and OUTPUT instruction.

```
data air_subset;
    set sashelp.air;
    if _N_ >= 4 and _N_ <= 20 then output;
run;
```

The sentence `if *N* >= 4 and *N* <= 20 then output;` checks if the currento bservation number (N) is between 4 and 20 (inclusive), and if so, outputs that observation to the new data set.

**Exercise 2.** Print (at least) five different descriptive statistics for groups given by variable SMOKING_STATUS from SASHELP:HEART for variables: WEIGHT, DIASTOLIC, SYSTOLIC, MRW (Use BY in PROC MEANS). Create also an output table containing these statistics.

```
proc sort data=sashelp.heart out=heart;
    by smoking_status;
run;
proc means data=heart max min mean var std;
    by smoking_status;
    var weight diastolic systolic mrw;
run;
```

- `proc means data=sashelp.heart max min mean var std;` starts the PROC MEANS procedure and specify the statistics that are going to be observed.

- `by smoking_status;` groups the data by this variable.

- `var weight diastolic systolic mrw;` specifies the variables for which to calculate the descriptive statistics.

### Smoking Status=' '

| Variable | Etiqueta | Máximo | Mínimo | Media | Varianza | Desv. est |
|---|---|---|---|---|---|---|
| Weight | | 209.0000000 | 117.0000000 | 152.8888889 | 749.0730159 | 27.3691983 |
| Diastolic | | 136.0000000 | 68.0000000 | 86.7777778 | 224.0634921 | 14.9687505 |
| Systolic | | 234.0000000 | 94.0000000 | 138.5833333 | 854.6500000 | 29.2343975 |
| MRW | Metropolitan Relative Weight | 170.0000000 | 85.0000000 | 117.2222222 | 360.4634921 | 18.9858761 |

### Smoking Status=Heavy (16-25)

| Variable | Etiqueta | Máximo | Mínimo | Media | Varianza | Desv. est |
|---|---|---|---|---|---|---|
| Weight | | 300.0000000 | 85.0000000 | 154.7629063 | 810.3820103 | 28.4672094 |
| Diastolic | | 138.0000000 | 55.0000000 | 83.8527725 | 144.3342842 | 12.0139204 |
| Systolic | | 270.0000000 | 89.0000000 | 133.3575526 | 411.2729908 | 20.2798666 |
| MRW | Metropolitan Relative Weight | 250.0000000 | 73.0000000 | 115.3565966 | 315.0086024 | 17.7484817 |

### Smoking Status=Light (1-5)

| Variable | Etiqueta | Máximo | Mínimo | Media | Varianza | Desv. est |
|---|---|---|---|---|---|---|
| Weight | | 239.0000000 | 72.0000000 | 146.7664360 | 756.6541921 | 27.5073480 |
| Diastolic | | 160.0000000 | 50.0000000 | 83.7754750 | 169.5827193 | 13.0223930 |
| Systolic | | 236.0000000 | 82.0000000 | 134.1485320 | 496.2962392 | 22.2777072 |
| MRW | Metropolitan Relative Weight | 198.0000000 | 75.0000000 | 118.3044983 | 412.6766175 | 20.3144436 |

### Smoking Status=Moderate (6-15)

| Variable | Etiqueta | Máximo | Mínimo | Media | Varianza | Desv. est |
|---|---|---|---|---|---|---|
| Weight | | 237.0000000 | 83.0000000 | 144.5860870 | 728.6785517 | 26.9940466 |
| Diastolic | | 138.0000000 | 50.0000000 | 82.6076389 | 155.2301329 | 12.4591385 |
| Systolic | | 290.0000000 | 90.0000000 | 131.7135417 | 498.5560598 | 22.3283689 |
| MRW | Metropolitan Relative Weight | 184.0000000 | 76.0000000 | 113.4104348 | 299.2005878 | 17.2974156 |

### Smoking Status=Non-smoker

| Variable | Etiqueta | Máximo | Mínimo | Media | Varianza | Desv. est |
|---|---|---|---|---|---|---|
| Weight | | 300.0000000 | 67.0000000 | 153.7422969 | 857.2169893 | 29.2782682 |
| Diastolic | | 155.0000000 | 50.0000000 | 86.9100360 | 176.2843032 | 13.2772099 |
| Systolic | | 300.0000000 | 86.0000000 | 140.3782487 | 658.1336707 | 25.6541161 |
| MRW | Metropolitan Relative Weight | 268.0000000 | 67.0000000 | 123.8987595 | 437.7107230 | 20.9215373 |

### Smoking Status=Very Heavy (> 25)

| Variable | Etiqueta | Máximo | Mínimo | Media | Varianza | Desv. est |
|---|---|---|---|---|---|---|
| Weight | | 256.0000000 | 96.0000000 | 164.0810235 | 758.1472673 | 27.5344742 |
| Diastolic | | 144.0000000 | 54.0000000 | 85.6666667 | 160.7120567 | 12.6772259 |
| Systolic | | 294.0000000 | 94.0000000 | 136.0000000 | 410.1361702 | 20.2518189 |
| MRW | Metropolitan Relative Weight | 193.0000000 | 84.0000000 | 119.4925373 | 309.2590254 | 17.5857620 |

**Exercise 3:**

**1.** You have to create one-dimensional table of frequency (using PROC FREQ) for variable Category from the set np_species.sas7bdat (directory

/home/u48457320/my_shared_file_links/u48457320/data). Use ORDER to sort data in the report in descending way.

```
proc freq
data="/home/u48457320/my_shared_file_links/u48457320/data/np_species.sas7bdat"
order=freq;
    tables Category / out=freq_table;
run;
```

It seems I do not have authorization:

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69           proc freq data="/home/u48457320/my_shared_file_lir
ERROR: User does not have appropriate authorization level for
70              tables Category / out=freq_table;
ERROR: No data set open to look up variables.
71          run;
```

But in case of having access `order=freq` would sort the frequency table in descending order by frequency count and `tables Category / out=freq_table;` requests a one-dimensional frequency table for the variable "Category" and outputs the results to an output data set called "freq_table".

**2.** Create a crossing-table for variables Category and Nativeness from the previous set np_species.sas7bdat. Do not display cumulative values.

As before, I do not have access to the set, but in case I had the solutions should be:

```
proc freq
data="/home/u48457320/my_shared_file_links/u48457320/data/np_species.sas7bdat"
;
    tables Category*Nativeness / nocum;
run;
```

The `nocum` option suppress the printing of cumulative counts and `*` operator allows us to create a crossing table.

**Exercise 4.** Display extreme observations, moments and histogram for variable Salary from sashelp.baseball.

```
/* Display extreme observations */
proc univariate data=sashelp.baseball;
    var salary;
    output out=stats pctlpts=1 5 10 25 50 75 90 95 99 pctlpre=percentile_;
run;

proc print data=stats;
    var percentile_:;
run;



/* Calculate moments */
proc means data=sashelp.baseball n mean std skew kurt;
    var salary;
run;

/* Create histogram */
proc univariate data=sashelp.baseball;
    var salary;
    histogram / normal(mu=est sigma=est);
run;
```

## Procedimiento UNIVARIATE (continuación — primera ventana)

**Observaciones extremas**

| Inferior | | Superior | |
|---|---|---|---|
| Valor | Obs | Valor | Obs |
| 67.5 | 41 | 1940.00 | 230 |
| 68.0 | 213 | 1975.00 | 83 |
| 70.0 | 260 | 2127.33 | 218 |
| 70.0 | 110 | 2412.50 | 164 |
| 70.0 | 93 | 2460.00 | 101 |

**Valores ausentes**

| Valor ausente | Conteo | Porcentaje de | |
|---|---|---|---|
| | | Todas las observaciones | Observaciones ausentes |
| . | 59 | 18.32 | 100.00 |

| Obs | percentile_1 | percentile_5 | percentile_10 | percentile_25 | percentile_50 | percentile_75 | percentile_90 | percentile_95 | percentile_99 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 70 | 86.5 | 100 | 190 | 425 | 750 | 1050 | 1350 | 2127.33 |

**Procedimiento MEANS**

Analysis Variable : Salary 1987 Salary in $ Thousands

| N | Media | Desv. est | Asimetria | Curtosis |
|---|---|---|---|---|
| 263 | 535.9258821 | 451.1186807 | 1.5889674 | 3.0589647 |

**Procedimiento UNIVARIATE**
Variable: Salary (1987 Salary in $ Thousands)

**Momentos**

| N | 263 | Sumar pesos | 263 |
|---|---|---|---|
| Media | 535.925882 | Observ suma | 140948.507 |
| Desviación std | 451.118681 | Varianza | 203508.064 |
| Asimetría | 1.58896735 | Curtosis | 3.05896473 |
| SC no corregida | 128857066 | SC corregida | 53319112.8 |
| Coef. variación | 84.1755727 | Media error std | 27.8171695 |

## Segunda ventana

**Tests para posición: Mu0=0**

| Test | Estadistico | | P valor | |
|---|---|---|---|---|
| T de Student | t | 19.26601 | Pr > \|t\| | <.0001 |
| Signo | M | 131.5 | Pr >= \|M\| | <.0001 |
| Rango con signo | S | 17358 | Pr >= \|S\| | <.0001 |

**Cuantiles (Definición 5)**

| Nivel | Cuantil |
|---|---|
| 100% Máx | 2460.00 |
| 99% | 2127.33 |
| 95% | 1350.00 |
| 90% | 1050.00 |
| 75% Q3 | 750.00 |
| 50% Mediana | 425.00 |
| 25% Q1 | 190.00 |
| 10% | 100.00 |
| 5% | 86.50 |
| 1% | 70.00 |
| 0% Mín | 67.50 |

**Observaciones extremas**

| Inferior | | Superior | |
|---|---|---|---|
| Valor | Obs | Valor | Obs |
| 67.5 | 41 | 1940.00 | 230 |
| 68.0 | 213 | 1975.00 | 83 |
| 70.0 | 260 | 2127.33 | 218 |
| 70.0 | 110 | 2412.50 | 164 |
| 70.0 | 93 | 2460.00 | 101 |

**Valores ausentes**

| Valor ausente | Conteo | Porcentaje de | |
|---|---|---|---|
| | | Todas las observaciones | Observaciones ausentes |
| . | 59 | 18.32 | 100.00 |

**Procedimiento UNIVARIATE**
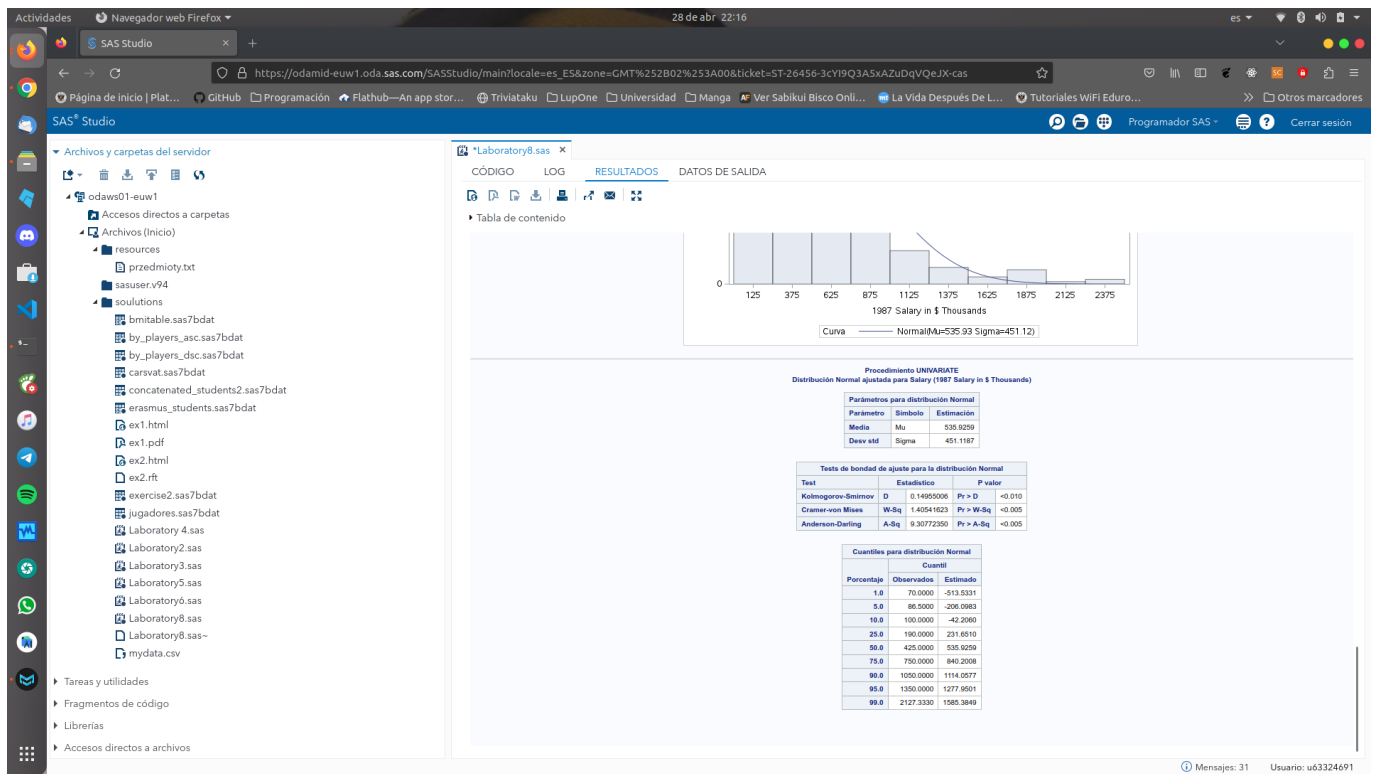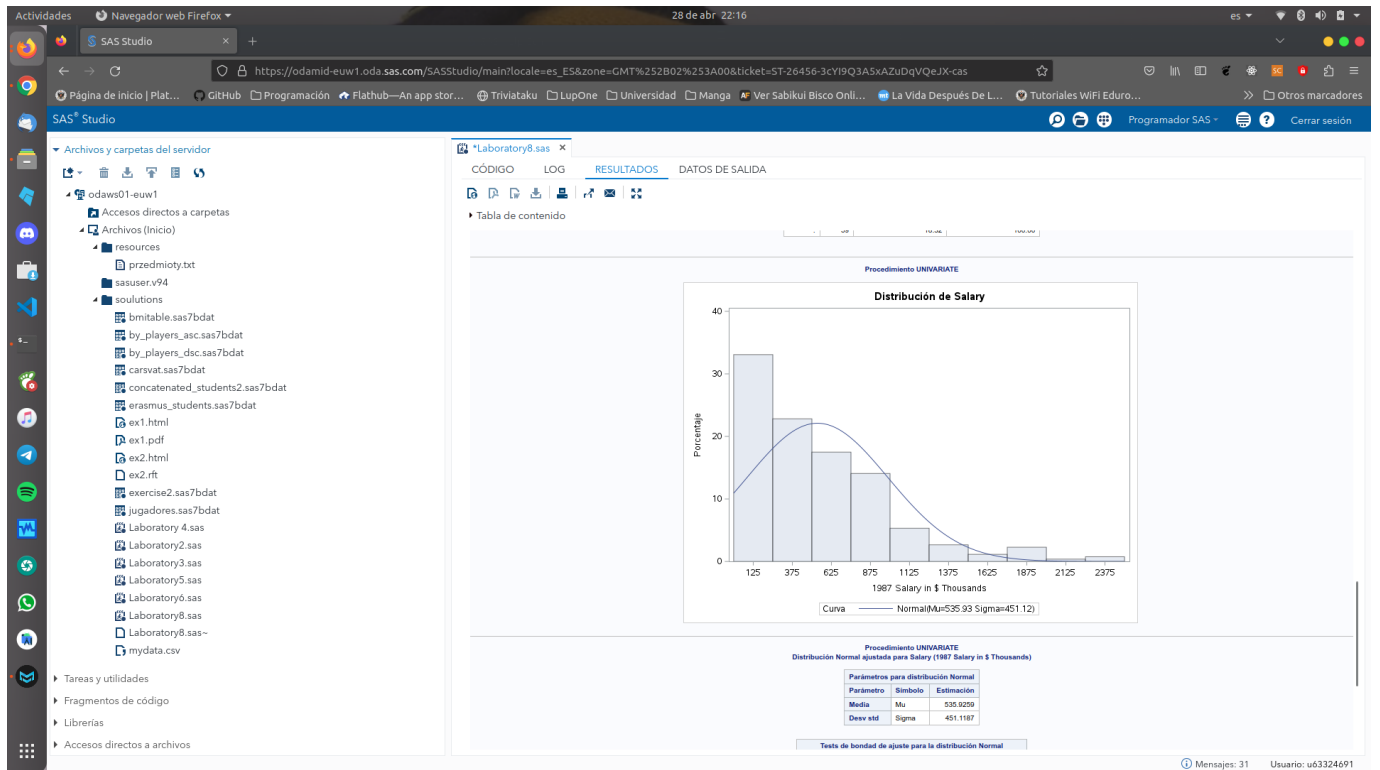
**Exercise 5.** From the following code

```
proc transpose data=sashelp.class out=class_stat;
    var weight age height;
run;



data class_statistics(drop=col1-col19);
    set class_stat;
    my_sum=sum(of col1-col19);
    my_mean=mean(of col1-col19);
    my_variance=var(of col1-col19);
    standard_deviation=std(of col1-col19);
    number_of_missing_data=nmiss(of col1-col19);
run;
```

Calculate the maximun and minimun for weight age and height.

We can modify the code as follows:

```
proc transpose data=sashelp.class out=class_stat;
    var weight age height;
run;

data class_statistics(drop=col1-col19);
    set class_stat;
    my_sum=sum(of col1-col19);
    my_mean=mean(of col1-col19);
    my_variance=var(of col1-col19);
    standard_deviation=std(of col1-col19);
    number_of_missing_data=nmiss(of col1-col19);
    minimun=min(of col1-col19);
    maximun=max(of col1-col19);
run;
```

- odaws01-euw1
  - Accesos directos a carpetas
  - Archivos (Inicio)
    - resources
      - przedmioty.txt
    - sasuser.v94
    - soulutions
      - bmitable.sas7bdat
      - by_players_asc.sas7bdat
      - by_players_dsc.sas7bdat
      - carsvat.sas7bdat
      - concatenated_students2.sas7bdat
      - erasmus_students.sas7bdat
      - ex1.html
      - ex1.pdf
      - ex2.html
      - ex2.rft
      - exercise2.sas7bdat
      - jugadores.sas7bdat
      - Laboratory 4.sas
      - Laboratory2.sas
      - Laboratory3.sas
      - Laboratory5.sas
      - Laboratory6.sas
      - Laboratory8.sas
      - Laboratory8.sas~
      - mydata.csv

Tareas y utilidades

Fragmentos de código

Librerías

Accesos directos a archivos

---

**\*Laboratory8.sas** ✕

CÓDIGO   LOG   RESULTADOS   DATOS DE SALIDA

Tabla: WORK.CLASS_STATISTICS ▾    Ver: Nombres de columna ▾    Filtrar: (ninguno)

Columnas

- ☑ Seleccionar todo
- ☑ △ _NAME_
- ☑ my_sum
- ☑ my_mean
- ☑ my_variance
- ☑ standard_deviation
- ☑ number_of_missing_data
- ☑ minimun
- ☑ maximun

| Propiedad | Valor |
|---|---|
| Etiqueta | |
| Nombre | |
| Longitud | |
| Tipo | |
| Formato | |
| Formato de lectura | |

Nº total de filas: 3  Nº total de columnas: 8      Filas 1-3

| my_variance | standard_deviation | number_of_missing_data | minimun | maximun |
|---|---|---|---|---|
| 518.65204678 | 22.773933494 | 0 | 50.5 | 150 |
| 2.2280701754 | 1.4926721594 | 0 | 11 | 16 |
| 26.286900585 | 5.1270752466 | 0 | 51.3 | 72 |