S05_T01_Sampling

April 5, 2022

1 S05_T01_Sampling

1.0.1 Ex1: Agafa un conjunt de dades de tema esportiu que t'agradi. Realitza un mostreig de les dades generant una mostra aleatòria simple i una mostra sistemàtica

[26]:
```
#Importem llibreries necessàries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import random
from random import randint
```

[27]:
```
data_df= pd.read_csv("statistics_players.csv", sep= ";")
data_df.head()
```

[27]:

	name	ranking	puntos	partidos_jugados	partidos_ganados	partidos_perdidos	efectividad	racha_victorias	compañero	posicion	…	dieciseisavos 2021
0	Juan Lebrón Chincoa	1	15450	312	199	113	63,78	18	Alejandro Galán Romo	Revés	…	0
1	Alejandro Galán Romo	2	14830	297	206	91	69,36	18	Juan Lebrón Chincoa	Revés	…	0
2	Francisco Navarro Compán	3	10970	416	312	104	75	15	Martín Di Nenno	Revés	…	0
3	Fernando Belasteguín	4	10815	423	377	46	89,13	63	Carlos Daniel Gutiérrez	Revés	…	0
4	Carlos Daniel Gutiérrez	5	10770	439	342	97	77,9	13	Fernando Belasteguín	Drive	…	0

```
  partidos jugados 2020 partidos ganados 2020 efectividad 2020  campeon 2020  \
0                  42.0                  37.0             88,1           6.0
1                  42.0                  37.0             88,1           6.0
2                  31.0                  22.0            70,97           1.0
3                  34.0                  26.0            76,47           2.0
4                  32.0                  22.0            68,75           1.0

   finalista 2020 semifinalista 2020  cuartos 2020  octavos 2020  \
0             2.0                3.0           0.0           0.0
1             2.0                3.0           0.0           0.0
2             2.0                4.0           3.0           0.0
3             2.0                2.0           4.0           0.0
4             2.0                5.0           1.0           2.0

   dieciseisavos 2020
0                 0.0
1                 0.0
2                 0.0
3                 0.0
4                 0.0

[5 rows x 32 columns]
```

[42]:
```
data_df.tail()
```

[42]:
```
                         name  ranking  puntos  partidos_jugados  \
44      Paula Josemaría Martín        7    9235               143
45  Beatriz González Fernández       10    6465               134
46  Patricia Llaguno Zielinski       11    6170               285
47         Delfina Brea Senesi       14    3390               118
48      Tamara Icardo Alcorisa       23    1828               137

    partidos_ganados efectividad              compañero posicion  \
44                81       56,64  Ariana Sánchez Fallada    Drive
45                78       58,21     Lucía Sainz Pelegri    Revés
46               200       70,18    María Virginia Riera    Revés
47                71       60,17  Tamara Icardo Alcorisa    Drive
48                58       42,34     Delfina Brea Senesi    Revés

      lugar nacimiento fecha nacimiento  … partidos jugados 2021  \
44  Moraleja (Caceres)       31/10/1996  …                    4
45              Málaga       23/11/2001  …                    4
46           Cartagena       25/02/1985  …                    3
47        Buenos Aires       05/12/1999  …                   50
48            Valencia       10/10/1995  …                    4

   partidos ganados 2021  efectividad 2021  campeon 2021 finalista 2021  \
```

```
44                    4              100            1             0
45                    3               75            0             1
46                    2            66,67            0             0
47                   35               70            2             1
48                    3               75            0             0

    partidos jugados 2020  partidos ganados 2020  efectividad 2020  \
44                   33.0                   24.0             72,73
45                   48.0                   33.0             68,75
46                   35.0                   24.0             68,57
47                   34.0                   19.0             55,88
48                   13.0                    4.0             30,77

    campeon 2020 finalista 2020
44           1.0            3.0
45           2.0            0.0
46           0.0            1.0
47           1.0            0.0
48           0.0            0.0

[5 rows x 22 columns]
```

[28]:
```
data_df.describe()
```

[28]:
```
           ranking        puntos  partidos_jugados  partidos_ganados  \
count    49.000000     49.000000         49.000000         49.000000
mean     97.530612   3455.530612        154.102041         96.775510
std      91.798489   4825.290069        124.994541         96.900779
min       1.000000     25.000000          7.000000          1.000000
25%      11.000000     69.000000         41.000000         14.000000
50%      53.000000   1174.000000        118.000000         69.000000
75%     180.000000   6170.000000        260.000000        165.000000
max     255.000000  15450.000000        439.000000        377.000000

       partidos_perdidos  racha_victorias  partidos jugados 2021  \
count          49.000000        49.000000              49.000000
mean           57.326531         7.612245               9.408163
std            37.406766         9.517126              16.755495
min             5.000000         1.000000               0.000000
25%            29.000000         2.000000               0.000000
50%            54.000000         5.000000               3.000000
75%            80.000000         8.000000               8.000000
max           138.000000        63.000000              72.000000

       partidos ganados 2021  campeon 2021  finalista 2021  …  octavos 2021  \
count              49.000000     49.000000       49.000000  …     49.000000
mean                5.244898      0.122449        0.061224  …      0.163265
```

```
std                 11.287033      0.389051        0.242226  …      0.472005
min                  0.000000      0.000000        0.000000  …      0.000000
25%                  0.000000      0.000000        0.000000  …      0.000000
50%                  1.000000      0.000000        0.000000  …      0.000000
75%                  4.000000      0.000000        0.000000  …      0.000000
max                 52.000000      2.000000        1.000000  …      2.000000

       dieciseisavos 2021  partidos jugados 2020  partidos ganados 2020  \
count           49.000000              47.000000              47.000000
mean             0.204082              19.851064              11.659574
std              0.706505              13.834391              11.955127
min              0.000000               0.000000               0.000000
25%              0.000000               8.000000               2.500000
50%              0.000000              19.000000               5.000000
75%              0.000000              32.500000              22.000000
max              4.000000              48.000000              37.000000

       campeon 2020  finalista 2020  semifinalista 2020  cuartos 2020  \
count     47.000000       47.000000           47.000000     47.000000
mean       0.787234        0.531915            0.808511      1.042553
std        1.640999        0.952139            1.393124      2.074246
min        0.000000        0.000000            0.000000      0.000000
25%        0.000000        0.000000            0.000000      0.000000
50%        0.000000        0.000000            0.000000      0.000000
75%        1.000000        0.500000            1.500000      1.500000
max        6.000000        3.000000            5.000000     11.000000

       octavos 2020  dieciseisavos 2020
count     47.000000           47.000000
mean       0.787234            1.361702
std        1.531356            2.470922
min        0.000000            0.000000
25%        0.000000            0.000000
50%        0.000000            0.000000
75%        1.000000            1.000000
max        8.000000            8.000000

[8 rows x 22 columns]
```

[29]:
```
data_df.shape
```

[29]:
```
(49, 32)
```

[30]:
```
data_df.columns
```

[30]:
```
Index(['name', 'ranking', 'puntos', 'partidos_jugados', 'partidos_ganados',
       'partidos_perdidos', 'efectividad', 'racha_victorias', 'compañero',
```

```
       'posicion', 'lugar nacimiento', 'fecha nacimiento', 'altura',
       'residencia', 'partidos jugados 2021', 'partidos ganados 2021',
       'efectividad 2021', 'campeon 2021', 'finalista 2021',
       'semifinalista 2021', 'cuartos 2021', 'octavos 2021',
       'dieciseisavos 2021', 'partidos jugados 2020', 'partidos ganados 2020',
       'efectividad 2020', 'campeon 2020', 'finalista 2020',
       'semifinalista 2020', 'cuartos 2020', 'octavos 2020',
       'dieciseisavos 2020'],
      dtype='object')
```

[31]:
```
#eliminem columnes que no necessitarem
data_df = data_df.drop(['partidos_perdidos','racha_victorias','semifinalista␣
↪2021','cuartos 2021', 'octavos 2021',
      'dieciseisavos 2021','semifinalista 2020', 'cuartos 2020', 'octavos␣
↪2020',
      'dieciseisavos 2020'], axis = 1)
```

[32]:
```
data_df.describe()
```

[32]:

	ranking	puntos	partidos_jugados	partidos_ganados	\
count	49.000000	49.000000	49.000000	49.000000	
mean	97.530612	3455.530612	154.102041	96.775510	
std	91.798489	4825.290069	124.994541	96.900779	
min	1.000000	25.000000	7.000000	1.000000	
25%	11.000000	69.000000	41.000000	14.000000	
50%	53.000000	1174.000000	118.000000	69.000000	
75%	180.000000	6170.000000	260.000000	165.000000	
max	255.000000	15450.000000	439.000000	377.000000	

	partidos jugados 2021	partidos ganados 2021	campeon 2021	\
count	49.000000	49.000000	49.000000	
mean	9.408163	5.244898	0.122449	
std	16.755495	11.287033	0.389051	
min	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	
50%	3.000000	1.000000	0.000000	
75%	8.000000	4.000000	0.000000	
max	72.000000	52.000000	2.000000	

	finalista 2021	partidos jugados 2020	partidos ganados 2020	\
count	49.000000	47.000000	47.000000	
mean	0.061224	19.851064	11.659574	
std	0.242226	13.834391	11.955127	
min	0.000000	0.000000	0.000000	
25%	0.000000	8.000000	2.500000	
50%	0.000000	19.000000	5.000000	
75%	0.000000	32.500000	22.000000	

```
max          1.000000              48.000000              37.000000
```

```
       campeon 2020  finalista 2020
count     47.000000       47.000000
mean       0.787234        0.531915
std        1.640999        0.952139
min        0.000000        0.000000
25%        0.000000        0.000000
50%        0.000000        0.000000
75%        1.000000        0.500000
max        6.000000        3.000000
```

[34]:
```
data_df.shape
```

[34]:
```
(49, 22)
```

[36]:
```
#generem mostra aleàtoria simple
simple_sample_df=data_df.sample(20)
```

[37]:
```
simple_sample_df.head(10)
```

[37]:

	name	ranking	puntos	partidos_jugados	partidos_ganados	efectividad	compañero	posicion	lugar nacimiento	fecha nacimiento	…	partidos jugados 2021
27	Francisco Arenas Gualda	180	69	41	14	34,15	Rubén Sánchez Sánchez	Drive	Madrid	27/02/1993	…	24
7	Arturo Coello Manso	31	1783	72	39	54,17	Miguel Lamperti	Drive				
43	Marta Marrero Marrero	5	10350	288	224	77,78	Marta Ortega Gallego	Revés				
24	Cayetano Rocafort Lores	173	75	118	63	53,39	Simon Vasquez	Revés				
37	Javier Bravo Álvarez	255	25	11	3	27,27	Carlos Pérez Cabeza	Drive				
1	Alejandro Galán Romo	2	14830	297	206	69,36	Juan Lebrón Chincoa	Revés				
23	Carlos Mora Íscar	169	78	41	12	29,27	Marc Pou Serra	Revés				
29	Adrián Corona Roldán	192	60	60	16	26,67	Alfonso Sánchez Arriaga	Drive				
13	Maximiliano Grabiel Martínez	49	1203	293	170	58,02	Adrián Blanco Antelo	Drive				
11	Javier Garrido Gómez	38	1532	171	93	54,39	Juan Cruz Belluati López	Revés				

```
7                    Valladolid       08/03/2002  …                    39
43  Las Palmas de Gran Canaria       16/01/1983  …                     1
24                       Málaga       04/10/1994  …                     0
37                       Málaga       15/01/1990  …                     0
1                        Madrid       15/05/1996  …                     0
23            Palma de Mallorca       05/08/2000  …                     1
29             Medina del Campo       15/12/1991  …                    21
13                     La Plata       12/08/1976  …                     2
11                      Córdoba       26/10/2000  …                     0

   partidos ganados 2021  efectividad 2021  campeon 2021 finalista 2021  \
27                     7             29,17             0              0
7                     20             51,28             0              0
43                     0                 0             0              0
24                     0                 0             0              0
37                     0                 0             0              0
1                      0                 0             0              0
23                     0                 0             0              0
29                     7             33,33             0              0
13                     1                50             0              0
11                     0                 0             0              0

    partidos jugados 2020  partidos ganados 2020  efectividad 2020  \
27                   15.0                    6.0                40
7                    30.0                   18.0                60
43                   33.0                   24.0             72,73
24                    5.0                    3.0                60
37                    3.0                    2.0             66,67
1                    42.0                   37.0              88,1
23                   13.0                    4.0             30,77
29                   21.0                    5.0             23,81
13                   13.0                    3.0             23,08
11                   11.0                    2.0             18,18

    campeon 2020 finalista 2020
27           0.0            0.0
7            0.0            0.0
43           1.0            3.0
24           0.0            0.0
37           0.0            0.0
1            6.0            2.0
23           0.0            0.0
29           0.0            0.0
13           0.0            0.0
11           0.0            0.0

[10 rows x 22 columns]
```

[38]:
```
simple_sample_df.describe()
```

[38]:

	ranking	puntos	partidos_jugados	partidos_ganados	\
count	20.000000	20.000000	20.000000	20.000000	
mean	100.300000	3248.900000	157.650000	96.750000	
std	91.010757	4823.789315	122.260841	94.207987	
min	1.000000	25.000000	11.000000	3.000000	
25%	18.500000	69.750000	49.250000	15.500000	
50%	54.000000	1136.500000	144.500000	78.000000	
75%	178.500000	3717.500000	253.500000	166.250000	
max	255.000000	14830.000000	439.000000	342.000000	

	partidos jugados 2021	partidos ganados 2021	campeon 2021	\
count	20.000000	20.00000	20.000000	
mean	6.750000	3.10000	0.100000	
std	11.570721	5.24053	0.307794	
min	0.000000	0.00000	0.000000	
25%	0.000000	0.00000	0.000000	
50%	1.500000	0.50000	0.000000	
75%	4.250000	4.00000	0.000000	
max	39.000000	20.00000	1.000000	

	finalista 2021	partidos jugados 2020	partidos ganados 2020	\
count	20.0	20.000000	20.000000	
mean	0.0	19.050000	10.750000	
std	0.0	12.521455	10.977849	
min	0.0	0.000000	0.000000	
25%	0.0	10.000000	2.750000	
50%	0.0	17.000000	5.000000	
75%	0.0	30.500000	19.000000	
max	0.0	42.000000	37.000000	

	campeon 2020	finalista 2020
count	20.000000	20.000000
mean	0.550000	0.500000
std	1.468081	0.945905
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.250000
max	6.000000	3.000000

[45]:
```
#generem mostra aleàtoria sistemàtica
#calculem step(k) = (poblacio)49/(mostra)24=> k=2
#el valor d'inici es tria com un aleatori entre els primers k elemmets de la␣
 ↪llista(h)
k=2
```

```
h=np.random.randint(1,k)
l=len(data_df)
sistem_aleat_sample_df= data_df[h:l:k]
sistem_aleat_sample_df
#en el dataset que he triat, hi ha poc nombre de files (mostres), per tant, la␣
 ↪mostra aleàtoria sistemàtica
#no tindria molt sentit realitzar-la, però a mode d'exercici serveix!
```

[45]:

	name	ranking	puntos	partidos_jugados	partidos_ganados	efectividad	compañero	posicion
1	Alejandro Galán Romo	2	14830	297	206	69,36	Juan Lebrón Chincoa	Revés
3	Fernando Belasteguín	4	10815	423	377	89,13	Carlos Daniel Gutiérrez	Revés
5	Jorge Nieto Ruiz	21	2900	205	120	58,54	Juan Martín Díaz Martínez	Revés
7	Arturo Coello Manso	31	1783	72	39	54,17	Miguel Lamperti	Drive
9	Gonzalo Rubio Pérez	36	1574	260	122	46,92	Christian Fuster Simarro	Drive
11	Javier Garrido Gómez	38	1532	171	93	54,39	Juan Cruz Belluati López	Revés
13	Maximiliano Grabiel Martínez	49	1203	293	170	58,02	Adrián Blanco Antelo	Drive
15	Miguel Yanguas Diez	54	1166	99	62	62,63	Iván Ramírez Del Campo	Drive
17	Javier Valdés González	64	986	221	141	63,8	Simon Vasquez	Drive
19	Simon Vasquez	150	94	14	3	21,43	Javier Valdés González	Drive
21	Pablo José Díaz Romero	157	89	71	31	43,66	Jairo Jose Bautista Ortiz	Drive
23	Carlos Mora Íscar	169	78	41	12	29,27	Marc Pou Serra	Revés
25	Ricardo Martins	178	70	52				
27	Francisco Arenas Gualda	180	69	41				
29	Adrián Corona Roldán	192	60	60				
31	Hugo García Martínez	206	52	8				
33	Ramiro Pereyra	234	34	14				
35	Daniele Cattaneo	236	33	13				
37	Javier Bravo Álvarez	255	25	11				
39	Ariana Sánchez Fallada	1	12390	243				
41	Alejandra Salazar Bengoechea	3	11750	275				
43	Marta Marrero Marrero	5	10350	288				
45	Beatriz González Fernández	10	6465	134				
47	Delfina Brea Senesi	14	3390	118				

25	18	34,62	David Antolín Solla	Revés
27	14	34,15	Rubén Sánchez Sánchez	Drive
29	16	26,67	Alfonso Sánchez Arriaga	Drive
31	2	25	Antón Márquez Larrea	Revés
33	7	50	Martín Andornino	Revés
35	4	30,77	Simone Cremona	Revés
37	3	27,27	Carlos Pérez Cabeza	Drive
39	165	67,9	Paula Josemaría Martín	Revés
41	221	80,36	Gemma Triay Pons	Drive
43	224	77,78	Marta Ortega Gallego	Revés
45	78	58,21	Lucía Sainz Pelegri	Revés
47	71	60,17	Tamara Icardo Alcorisa	Drive

	lugar nacimiento	fecha nacimiento	…	partidos jugados 2021	\
1	Madrid	15/05/1996	…	0	
3	Pehuajo	19/05/1979	…	4	
5	Madrid	18/12/1998	…	2	
7	Valladolid	08/03/2002	…	39	
9	Sevilla	22/02/1991	…	0	
11	Córdoba	26/10/2000	…	0	
13	La Plata	12/08/1976	…	2	
15	Málaga	18/03/2002	…	69	
17	Santiago de Chile	27/06/1996	…	1	
19	Angelholm - Suecia	21/12/1992	…	11	
21	Sevilla	16/12/1990	…	0	
23	Palma de Mallorca	05/08/2000	…	1	
25	Lisboa - Portugal	27/08/1991	…	29	
27	Madrid	27/02/1993	…	24	
29	Medina del Campo	15/12/1991	…	21	
31	Móstoles	19/08/1998	…	8	
33	Mar del Plata - Arg	15/05/1996	…	9	
35	Vimercate - Italia	10/09/1989	…	6	
37	Málaga	15/01/1990	…	0	
39	Reus	19/07/1997	…	4	
41	Madrid	31/12/1985	…	3	
43	Las Palmas de Gran Canaria	16/01/1983	…	1	
45	Málaga	23/11/2001	…	4	
47	Buenos Aires	05/12/1999	…	50	

	partidos ganados 2021	efectividad 2021	campeon 2021	finalista 2021	\
1	0	0	0	0	
3	4	100	1	0	
5	1	50	0	0	
7	20	51,28	0	0	
9	0	0	0	0	
11	0	0	0	0	
13	1	50	0	0	

15	50	72,46	0	0
17	0	0	0	0
19	3	27,27	0	0
21	0	0	0	0
23	0	0	0	0
25	13	44,83	0	0
27	7	29,17	0	0
29	7	33,33	0	0
31	2	25	0	0
33	5	55,56	0	0
35	1	16,67	0	0
37	0	0	0	0
39	4	100	1	0
41	2	66,67	0	0
43	0	0	0	0
45	3	75	0	1
47	35	70	2	1

	partidos jugados 2020	partidos ganados 2020	efectividad 2020	\
1	42.0	37.0	88,1	
3	34.0	26.0	76,47	
5	22.0	11.0	50	
7	30.0	18.0	60	
9	13.0	3.0	23,08	
11	11.0	2.0	18,18	
13	13.0	3.0	23,08	
15	26.0	11.0	42,31	
17	19.0	7.0	36,84	
19	3.0	0.0	0	
21	6.0	3.0	50	
23	13.0	4.0	30,77	
25	19.0	5.0	26,32	
27	15.0	6.0	40	
29	21.0	5.0	23,81	
31	NaN	NaN	NaN	
33	2.0	1.0	50	
35	4.0	2.0	50	
37	3.0	2.0	66,67	
39	36.0	28.0	77,78	
41	36.0	28.0	77,78	
43	33.0	24.0	72,73	
45	48.0	33.0	68,75	
47	34.0	19.0	55,88	

	campeon 2020	finalista 2020
1	6.0	2.0
3	2.0	2.0

```
5            0.0            0.0
7            0.0            0.0
9            0.0            0.0
11           0.0            0.0
13           0.0            0.0
15           0.0            0.0
17           0.0            0.0
19           0.0            0.0
21           0.0            0.0
23           0.0            0.0
25           0.0            0.0
27           0.0            0.0
29           0.0            0.0
31           NaN            NaN
33           0.0            0.0
35           0.0            0.0
37           0.0            0.0
39           3.0            2.0
41           3.0            2.0
43           1.0            3.0
45           2.0            0.0
47           1.0            0.0

[24 rows x 22 columns]
```

1.0.2 Ex2: Continua amb el conjunt de dades de tema esportiu i genera una mostra estratificada i una mostra utilitzant SMOTE (Synthetic Minority Oversampling Technique)

Stratified Sampling

Assume that we need to estimate the average number of votes for each candidate in an election. Assume that the country has 3 towns: Town A has 1 million factory workers, Town B has 2 million workers, and Town C has 3 million retirees. We can choose to get a random sample of size 60 over the entire population but there is some chance that the random sample turns out to be not well balanced across these towns and hence is biased causing a significant error in estimation. Instead, if we choose to take a random sample of 10, 20 and 30 from Town A, B and C respectively then we can produce a smaller error in estimation for the same total size of the sample.

from sklearn.model_selection import train_test_split X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.25)

Oversampling using SMOTE

In SMOTE (Synthetic Minority Oversampling Technique) we synthesize elements for the minority class, in the vicinity of already existing elements

from imblearn.over_sampling import SMOTE

smote = SMOTE(ratio='minority') X_sm, y_sm = smote.fit_sample(X, y)

[ ]:

[ ]:

[ ]:

[ ]:

1.0.3 Ex3: Continua amb el conjunt de dades de tema esportiu i genera una mostra utilitzant el mètode Reservoir sampling

Reservoir Sampling

import random def generator(max): number = 1 while number < max: number += 1 yield number #Create as stream generator stream = generator(10000) #Doing Reservoir Sampling from the stream k=5 reservoir = [] for i, element in enumerate(stream): if i+1<= k: reservoir.append(element) else: probability = k/(i+1) if random.random() < probability: # Select item in stream and remove one of the k items already selected reservoir[random.choice(range(0,k))] = element print(reservoir)

[1369, 4108, 9986, 828, 5589]