Diplomatura AACSyA 2018 - FaMAF – UNC

Análisis y visualización de datos

Laboratorio N° 2

Exploración de datos del Sistema Nacional de Estadísticas sobre Ejecución de la Pena – SNEEP

INFORME ANUAL 2016

Tabla de Contenido

Introducción

En el siguiente informe se hará una conclusión del análisis que se obtuvo de los datos pertenecientes al Sistema Nacional de Estadísticas sobre Ejecución de la Pena (SNEEP) del año 2016.

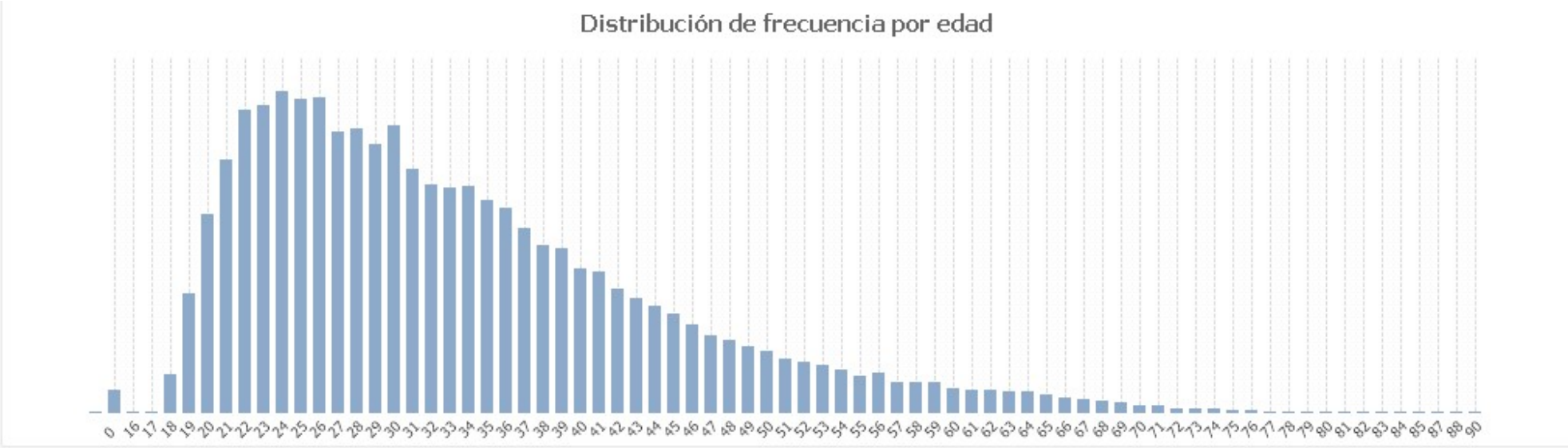
Analizaremos del conjunto de datos la edad de la población, sexo, tipo de condena, cantidad de años y la reincidencia teniendo en cuenta si participa o no en un programa educativo.

Edad de la Población detenida.

Para analizar la edad de la población se calcularon distintos estadísticos (media, mediana y moda) de los cual obtuvimos los siguientes resultados:

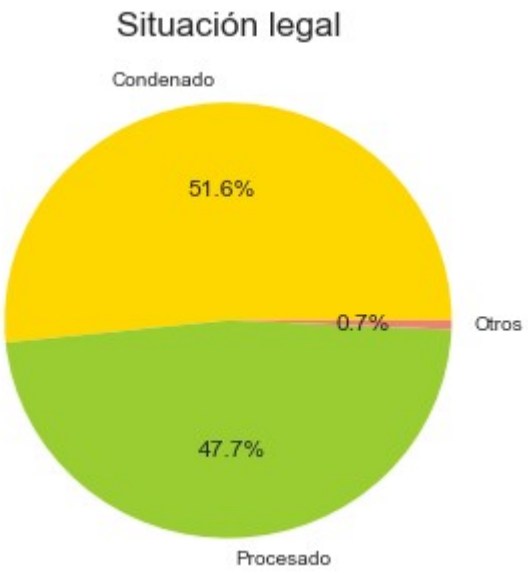
	Media	Mediana	Moda
Calculos Estadisticos	33.428054	31.0	24.0

Viendo el grafico que se muestra a continuación, determinamos que el valor de la Moda es el correcto, ya que muestra claramente que la edad de la población se centra en los 24 años.

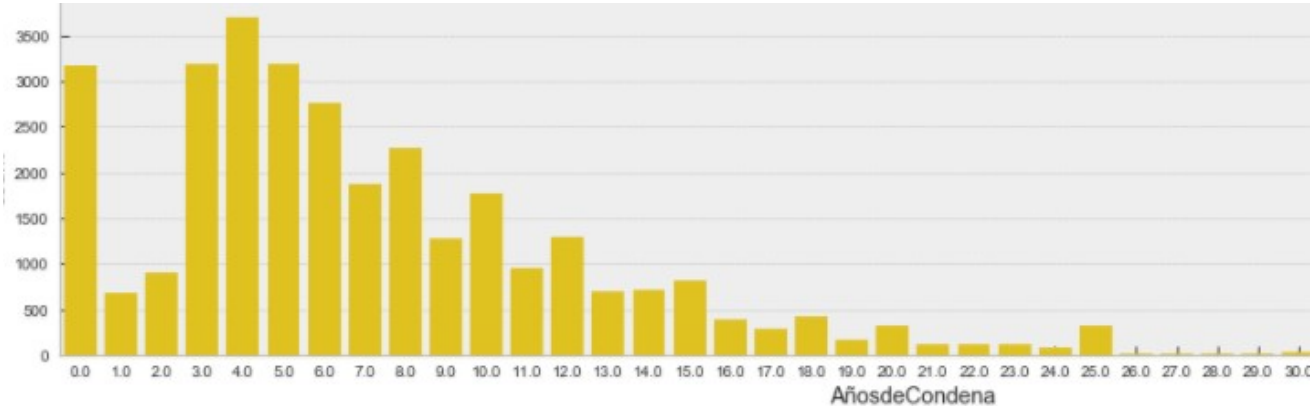


Situación Legal de la Población detenida

Antes de hacer el análisis de los años de condena vemos primero el % que existe según la situación penal de la Población:



De la cual analizaremos solo los **Condenados** para hacer el análisis por años de condena.



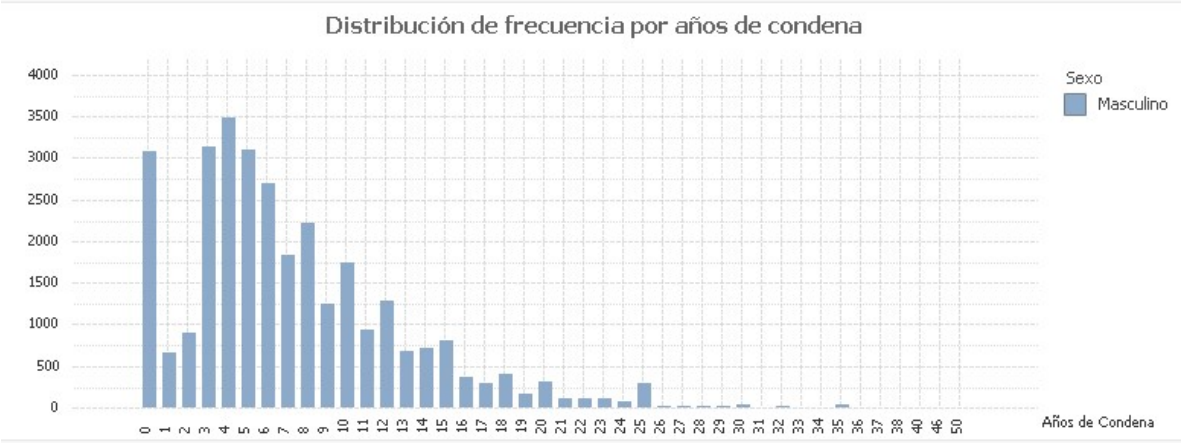
Volvemos a concluir que el estadístico que muestra mejor la frecuencia de los datos es la Moda.

	Media	Mediana	Moda
Calculos Estadisticos	7.215239	6.0	4.0

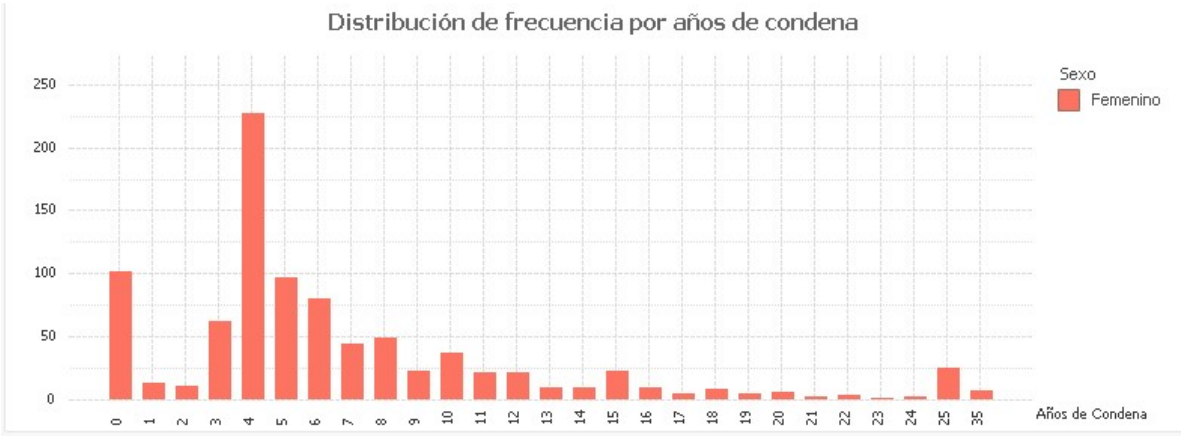
Análisis de la variable años de condenas por género de la población¶

Relacionando Sexo y los años de condena de aquellos que tiene una situación legal **Condenado** vemos la siguiente distribución.

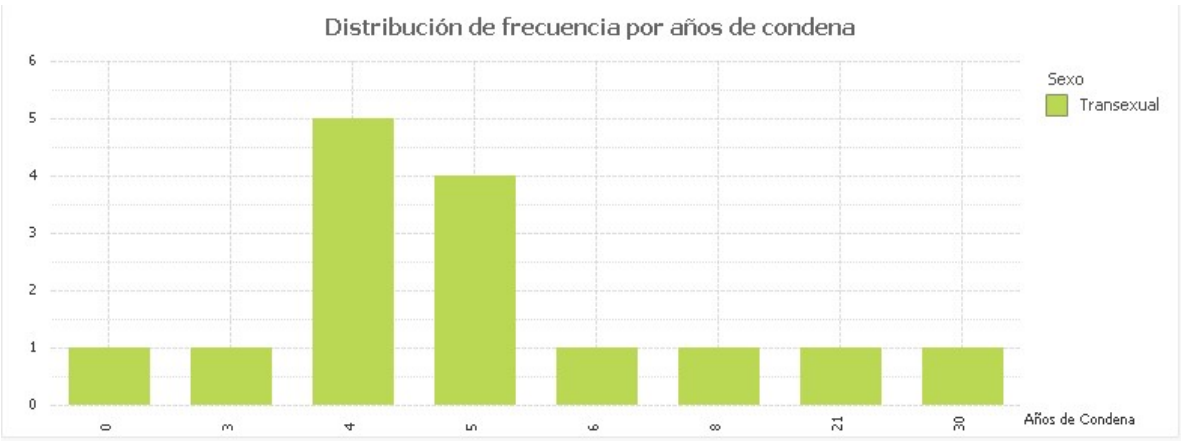
Sexo= Masculino



Sexo = Femenino



Sexo = Transexual



Si calculamos los estadísticos relacionando estas variables vemos que por género los valores son similares.

	Media	Mediana	Moda
Masculino	7.228325	6.0	4.0
Femenino	6.767699	5.0	4.0
Transexual	7.200000	5.0	4.0

Chi-Cuadrado de variables categóricas

Para hacer el cálculo de la relación de las variables utilizaremos las variables categóricas estado civil y situación legal de la población.

Las hipótesis son:

H0="las dos variables en estudio son independientes"

H1="las dos variables en estudio están relacionadas"

Análisis de las variables: Situación legal y Estado Civil

situacion_legal_descripcion	Condenado	Otros	Procesado	All
estado_civil_descripcion				
Casado	3386	65	3175	6626
Concubino	3709	23	2729	6461
Separado de hecho	213	1	200	414
Separado o divorciado	593	7	595	1195
Soltero	30730	397	28903	60030
Viudo	439	3	387	829
All	39070	496	35989	75555

Acá vamos a ver que no es lo mismo estar casado y en concubinato

Se puede ver rápidamente que los solteros condenados tienen el mayor número de la población

Para tener un análisis certero debemos calcular las frecuencias marginales ya que no tenemos la misma población para los casados y lo que están en concubinato.

La frecuencia esperada:

$$fe_{ij} = \frac{(\text{total fila i-ésima})*(\text{total columna j-ésima})}{\text{gran total}}$$

Recalculando nos queda

```
array([[3.42634928e+03, 4.34980610e+01, 3.15615266e+03],
       [3.34102667e+03, 4.24148766e+01, 3.07755845e+03],
       [2.14082192e+02, 2.71780822e+00, 1.97200000e+02],
       [6.17942558e+02, 7.84488121e+00, 5.69212560e+02],
       [3.10419178e+04, 3.94082192e+02, 2.85940000e+04],
       [4.28681490e+02, 5.44218119e+00, 3.94876329e+02]])
```

Observamos que para 10 grados de libertad (los correspondientes para nuestra tabla de contingencia => (6-1)*(3-1)=10)

```
Chi2 = 111.50175491453389
P-value = 2.6558791836114227e-19
DoF = 10
```

Para un analizar con el nivel de confianza del 90% vemos si aceptamos o rechazamos **H0**

```
def calc_conf(confianza):
    """
    Retorna el nivel de significancia.
    """
    return (100 - confianza) / 100

# Queremos confianza al 90%
if p < calc_conf(90):
    print("Rechazo H0 ==> Las Variables Estan Correlacionadas")
else:
    print("Acepto H0 ==> Las Variables Son Independientes")
```

```
Rechazo H0 ==> Las Variables Estan Correlacionadas
```

Análisis de las variables: Situación legal y Tipo de infracción

situacion_legal_descripcion	Condenado	Otros	Procesado
tipo_infraccion_disciplinaria_descripcion			
Faltas graves	3385	21	2551
Faltas leves	625	2	487
Faltas media	2470	4	2495
No cometió Infracción disciplinaria	27874	317	26592

Calculando la frecuencia esperada nos queda

```
array([[3.06252006e+03, 3.06662077e+01, 2.86381373e+03],
       [5.72712330e+02, 5.73479191e+00, 5.35552878e+02],
       [2.55458489e+03, 2.55800548e+01, 2.38883506e+03],
       [2.81641827e+04, 2.82018946e+02, 2.63367983e+04]])
```

Observamos que para 6 grados de libertad (los correspondientes para nuestra tabla de contingencia => (4-1)*(3-1)=6)

Para un analizar con el nivel de confianza del 90% vemos si aceptamos o rechazamos **H0**

```
def calc_conf(confianza):
    """
    Retorna el nivel de significancia.
    """
    return (100 - confianza) / 100

# Queremos confianza al 90%
if p < calc_conf(95):
    print("Rechazo H0 ==> Las Variables Estan Correlacionadas")
else:
    print("Acepto H0 ==> Las Variables Son Independientes")
```

```
Rechazo H0 ==> Las Variables Estan Correlacionadas
```

Chi-Cuadrado con muestra aleatoria de datos

Se genera una muestra aleatoria de datos de las variables situación legal y tipo de infracción para ver el chi cuadrado con la muestra y obtenemos lo siguiente:

```
np.random.seed(10)
# Sample data randomly at fixed probabilities
situacion_legal = np.random.choice(a= ['Procesado','Condenado','Otros'],
                             p = [0.48, 0.51 ,0.01],
                             size=1000)

# Sample data randomly at fixed probabilities
tipo_infraccion = np.random.choice(a= ['Faltas graves','Faltas leves','Faltas media','No cometió Infracción'],
                             p = [0.09, 0.02, 0.07, 0.82],
                             size=1000)
relacion = pd.DataFrame({"Situacion":situacion_legal,
                    "Tipo_infraccion":tipo_infraccion})
relacion_tab = pd.crosstab(relacion.Tipo_infraccion,relacion.Situacion, margins = True)
relacion_tab.columns = ['Procesado','Condenado','Otros',"row_totals"]

relacion_tab.index = ['Faltas graves','Faltas leves','Faltas media','No cometió Infracción',"col_totals"]

observed = relacion_tab.ix[0:5,0:4]
relacion tab
```

	Procesado	Condenado	Otros	row_totals
Faltas graves	45	0	36	81
Faltas leves	9	0	7	16
Faltas media	32	1	43	76
No cometió Infracción	417	8	402	827
col_totals	503	9	488	1000

Calculando la frecuencia esperada nos queda

```
crit = stats.chi2.ppf(q = 0.95, # Find the critical value for 95% confidence*
                      df = 6)   # *

print("Critical value")
print(crit)

p_value = 1 - stats.chi2.cdf(x=chi_squared_stat,  # Find the p-value
                             df=6)
print("P value")
print(p_value)
```

```
Critical value
12.591587243743977
P value
0.6790755309935894
```

Y el Chi – Cuadrado obtenido es

```
relacion_tab = pd.crosstab(relacion.Tipo_infraccion,relacion.Situacion)
stats.chi2_contingency(observed= relacion_tab)
```

```
(3.9822726897588017,
 0.6790755309935894,
 6,
 array([[4.07430e+01, 7.29000e-01, 3.95280e+01],
        [8.04800e+00, 1.44000e-01, 7.80800e+00],
        [3.82280e+01, 6.84000e-01, 3.70880e+01],
        [4.15981e+02, 7.44300e+00, 4.03576e+02]]))
```

```python
def calc_conf(confianza):
    """
    Retorna el nivel de significancia.
    """
    return (100 - confianza) / 100

# Queremos confianza al 90%
if p_value < calc_conf(90):
    print("Rechazo H0 ==> Las Variables Estan Correlacionadas")
else:
    print("Acepto H0 ==> Las Variables Son Independientes")
```

```
Acepto H0 ==> Las Variables Son Independientes
```

Las variables son independientes y no como se podía observar no como veíamos con la población.