

Análisis Económico en función del otorgamiento de la Beca de Manutención

Universidad Nacional Autónoma de México

FES Acatlán

García Mata Luis Ángel
Gayosso Serrano Oscar Omar
Meleza Ortega José
Ramírez González Marco Antonio
Rodríguez Mondragón Jessica Fernanda

INDICE

1. Presentación del trabajo.....	1
2. Introducción.....	1
3. Marco Teórico.....	3
3.1 Metodología.....	4
4. Resultados.....	7
4.1 Análisis de los datos.....	7
5. Conclusiones.....	8
6. Bibliografía	9

1. Presentación del trabajo

Objetivo

Analizar el comportamiento del gasto correspondiente a los alumnos que son beneficiarios del apoyo económico de manutención o de excelencia, en comparación con los alumnos que no fueron seleccionados beneficiarios de dichos apoyos, y comprobar que sus gastos tienden a comportarse de una manera distinta.

Hipótesis

Los alumnos de Actuaría de la Facultad de Estudios Superiores Acatlán de las generaciones 2017, 2018 y 2019, beneficiarios de la beca de Manutención, presentan un gasto mayor en actividades extraescolares y de entretenimiento a diferencia de los alumnos no beneficiados.

Población Objetivo

Nuestra principal población son los estudiantes de la carrera de Actuaría que se imparte en la Facultad de Estudios Superiores Acatlán, que actualmente se encuentran estudiando y que pertenecen a las generaciones 2017, 2018 y 2019.

Marco de Muestreo

La coordinación del programa de Actuaría nos proporcionó la lista de los alumnos que están inscritos en la carrera de Actuaría por generaciones, la cual tomamos como nuestro marco de muestreo.

2. Introducción

La Universidad Nacional Autónoma de México (UNAM), brinda a sus estudiantes acceso a programas de apoyo económico y de sustento de los estudios (BECAS), los cuales otorgan a los alumnos una remuneración monetaria como impulso para la no deserción y posterior conclusión de la licenciatura, en un panorama general, la comunidad universitaria asociada a la UNAM, tienen acceso generalizado a muchos programas de apoyo económico, fundamentados en el trabajo que desarrolla el Departamento de Becas de la UNAM.

Focalizando la generalización de los programas de apoyo para la comunidad universitaria, el de principal interés es la Beca de Manutención (pues en general es el apoyo al que todos tenemos alcance), cuyo objetivo resulta: fomentar que las y los estudiantes tengan acceso, continúen y concluyan oportunamente con sus estudios en el tipo superior, evitando así la deserción escolar. En términos del objetivo de la Beca de Manutención, denotamos que la situación que los acreedores de la misma viven, está dada por efectos de la insolvencia económica para el seguimiento de sus estudios, sin embargo, el conocimiento generalizado de la comunidad universitaria, nos supone un comportamiento desviado del objetivo principal de la beca, en pocas palabras, plantea una diferencia esencial entre los acreedores y no acreedores del apoyo, determinada por el uso que le dan.

En el presente proyecto, analizaremos la incidencia de la Beca de Manutención como factor clave de diferenciación en el gasto monetario que presentan los alumnos de Licenciatura en Actuaría de la Facultad de Estudios Superiores Acatlán (FES-A), en las generaciones 2017, 2018 y 2019. Abordamos la situación en torno a los gastos que la población en estudio presenta, basándonos fundamentalmente, en un margen generalizado por dos variables principales: el gasto en actividades extraescolares y el gasto en actividades escolares.

Observaremos si la presencia del apoyo económico provoca un cambio entre las poblaciones divididas en: los alumnos seleccionados por el programa y los no seleccionados, analizando la incidencia de las variables principales propuestas.

En la trayectoria estudiantil a lo largo de la carrera, los factores relevantes en el incremento de los gastos extraescolares, son inducidos por una tendencia a la afiliación del programa de Manutención, es decir, tomando análisis de los alumnos que incurren frecuentemente a los lugares de entretenimiento circundantes a la FES-A, denotamos que la observación de este hecho plantea que la población seleccionada por el programa de Manutención, intuitivamente, tiene un gasto mayor en actividades extraescolares, dicho de otra manera, el destino monetario del apoyo económico se inclina hacia las actividades de ocio y entretenimiento, tales como, el cine, las fiestas, salidas con amigos, visitas a lugares de reunión social, aplicaciones de películas y series (Netflix y sus homólogos), entre otros.

Así, la población complementaria, tiene una tendencia hacia la contención del gasto en actividades como las presentadas con anterioridad, referimos entonces que la población que no cuenta con el apoyo de manutención, presenta una diferencia en el gasto con la población que si lo tiene, puesto que, la frecuencia de visitas a los lugares mencionados, es menor en el contexto de los individuos que no son acreedores del apoyo económico.

3. Marco Teórico

En la actualidad, hablar sobre la asignación de apoyos económicos a estudiantes representa un debate cotidiano que conlleva al enfrentamiento de posiciones e ideales. Por una parte, se piensa que esta asignación debería realizarse en función al promedio de los estudiantes como recompensa de su desarrollo académico, sin embargo, otro pensamiento que predomina en el diálogo es que la asignación debería tomar en consideración las condiciones económicas de los aspirantes.

La beca de manutención otorgada a los estudiantes de la Universidad Nacional Autónoma de México (UNAM) está dirigida para estudiantes que son desfavorecidos por sus condiciones económicas y que en muchas ocasiones estas influyen en un buen desarrollo académico, limitando la capacidad para comprar los materiales necesarios para sus actividades académicas, así como, la incapacidad para cubrir los gastos de transporte a la institución.

Esta beca tiene como objetivo fomentar que las y los estudiantes que tengan acceso a ella, continúen y concluyan oportunamente con sus estudios de superior, evitando así la deserción escolar tomando como criterios de selección que los aspirantes que provengan de municipios rurales, indígenas de alto y muy alto grado de marginación, así como provenir de un hogar

cuyo ingreso mensual per cápita sea igual o menor a \$3,081.10 pesos (Secretaría de Educación Pública, 2019).

Según el Portal de Estadística Universitaria, en 2018, 47 mil 507 alumnos en la UNAM fueron beneficiarios del apoyo de manutención a nivel licenciatura que cumplieron los requisitos para solicitarla, es decir, que son desfavorecidos económicamente o socialmente para continuar y culminar con sus estudios superiores, por lo que se espera apoyar a sus gastos escolares con el otorgamiento de esta beca.

3.1 Metodología

Para efectos del trabajo de investigación y con el propósito de extraer una muestra que sea capaz de representar a los estudiantes de las generaciones 2017, 2018 y 2019 de Actuaría en la FES Acatlán, hicimos uso de técnicas de muestreo, se realizó un muestreo aleatorio estratificado (MAE). El MAE es una técnica que mitiga el riesgo de obtener muestras no representativas, en donde la población objetivo está dividida en H subpoblaciones los cuales serán los estratos E_i .

Dado lo anterior podemos particionar nuestra población objetivo en tres subpoblaciones (estratos), las cuales deben ser homogéneas internamente y heterogéneas externamente, estas se establecieron como las generaciones 2017, 2018 y 2019 respectivamente.

Una vez que tenemos particionada nuestra muestra, se extrae una muestra aleatoria simple independiente de cada estrato, es decir, sobre la lista proporcionada por generaciones, se deberá aplicar un Muestreo Aleatorio Simple de tamaño n a cada estrato. Para obtener el tamaño n existen 3 estimadores, los cuales se presentan en la siguiente tabla.

Estimador	Tamaño de la muestra inicial	Tamaño de la muestra
\hat{t}_π	$n_0 = \frac{Z_{1-\alpha}^2 N^2 S_s^2}{\epsilon^2}$	$n = \frac{n_0}{1 + \frac{n_0}{N}}$
\bar{y}_s	$n_0 = \frac{Z_{1-\alpha}^2 S_s^2}{\epsilon^2}$	$n = \frac{n_0}{1 + \frac{n_0}{N}}$
\hat{p}_s	$n_0 = \frac{Z_{1-\frac{\alpha}{2}}^2 \hat{p}_s(1 - \hat{p}_s)}{\epsilon^2}$	$n = \frac{n_0}{1 + \frac{n_0}{N}}$

De donde:

\hat{t}_π = Estimador del total

ε = Error

\bar{y}_s = Estimador del promedio

S = Varianza de los estratos

\hat{p}_s = Estimador de la proporción

$Z_{1-\alpha}^2$ = Cuantil con un nivel $\alpha=95\%$

n_0 = Tamaño de la muestra inicial

N = Total de la población

n = Tamaño de la muestra

Para calcular el tamaño de la muestra utilizamos el estimador de la proporción \hat{p}_s , con la finalidad de obtener una muestra proporcional representativa de todas las generaciones. Para obtener la varianza de nuestros datos, primero se generó una encuesta piloto, en la cual se encuestaron a 10 personas de cada generación, preguntando a cerca de su gasto académico y no académico, además de 3 aspectos positivos, 3 negativos en los cuales gastan sus apoyos económicos.

Posteriormente se calculó la varianza total de las respuestas en función del gasto académico, y las variables restantes, a continuación, se muestra una tabla con los resultados obtenidos, asumiendo un error de 200 pesos.

Gasto Académico	
$Z_{95\%}$	1.644854
E	200
S^2	1,104,997
$Z_{95\%}^2$	2.706
E^2	40000
N	840
n_0	74.74
n	68.63

La generación 2017 cuenta con 279 alumnos, la generación 2018 con 281 y 2019 con 280 por lo que se calculó un factor de ponderación W_i por cada generación, los cuales resultaron como $w_1 = 0.333$, $w_2 = 0.3345$ y $w_3 = 0.3321$, y para obtener la muestra por generación, es decir n_i , se calculó de la siguiente manera:

$$n_i = \frac{N_i}{N}(n), n_1 = 23, n_2 = 23, n_3 = 23$$

Es decir, se requiere aplicar la encuesta a 23 alumnos por cada generación, dando un total de 69 registros.

Diseño de la Encuesta

Dado el fin de este proyecto, se definió como constructo “beca”, por lo que para aplicar la encuesta, se realizó una clasificación de las preguntas por dimensiones, las cuales son:

- Escuela
- Alimentos
- Transporte
- Vicios
- Entretenimiento
- Ropa
- Tecnología
- Apoyo Familiar
- Belleza e higiene personal
- Ahorro
- Deudas

Una vez establecidas las dimensiones se procedió a generar las preguntas, las cuales fueron las siguientes:

1. Género
2. Generación
3. Nombre
4. ¿Cuál es tu promedio?
5. ¿Cuál es el monto mensual de tu ingreso?
6. ¿En qué actividades gastas tu ingreso mensual?
7. ¿Cuentas con el apoyo económico de Manutención o de Excelencia?
 - 7.1 En un número aproximado, ¿Cuál es el monto que recibes mensualmente al ser beneficiario del apoyo?
 - 7.2 ¿Cual es el monto del apoyo que dedicas a actividades extra-escolares y de entretenimiento?
 - 7.3 ¿Cuánto dinero del apoyo dedicas a actividades escolares al mes?
8. El no ser beneficiario, ¿Afecta en tu economía?
9. ¿Cuánto dinero de tu ingreso mensual dedicas a actividades extra escolares y de entretenimiento al mes?
10. ¿Cuánto dinero de tu ingreso mensual dedicas a actividades escolares al mes?

4. Resultados

4.1 Limpieza de datos

Como en la mayoría de los casos, obtener los datos podría ser una tarea relativamente sencilla, sin embargo, al tratar de analizarlos, éstos se presentan de una manera muy “cruda”, por lo tanto, antes de empezar un análisis es muy oportuno hacer una limpieza a los resultados obtenidos.

Apoyados del lenguaje de programación Python, lo primero que hicimos fue la extracción de columnas que solo fueron útiles para el control de los encuestados y que no aportarían nada al estudio, en nuestro caso el tiempo en que se registraron las encuestas, pues todas se realizaron en el tiempo acordado. Otra variable que se extrajo fue el nombre de los encuestados, que fue una variable para el control del equipo al momento de realizar la encuesta. Posteriormente, se etiquetaron las columnas que se iban a analizar, para un manejo óptimo de los datos.

Uno de los problemas más comunes al tener los resultados de una investigación, se presenta al encontrar espacios en blancos por la naturaleza de la encuesta. En nuestra tabla de resultados se presentaron estos espacios continuamente, por lo tanto, se binarizaron esos espacios y las casillas con contenido, siempre y cuando a la variable se le pudiera dar un tratamiento binario, para así evitar sesgos en las variables que no tuvieran este carácter, por ejemplo, los intervalos de gastos en actividades extraescolares y escolares.

Para finalizar la etapa de limpieza, debemos ver cómo se encuentra la distribución de nuestro evento de interés el cuál es la beca, para poder ajustar un modelo que explique los resultados. En nuestra encuesta el 33.3% de la población encuestada resultó ser beneficiaria de algún apoyo económico, mientras que el 66.6% informó no tener alguna beca. A Partir de estos resultados se pueden empezar el proceso de análisis.

4.2 Análisis de datos

Como primer análisis, se comprobó si existe alguna diferencia entre la distribución del gasto para recreación y el gasto escolar, discriminando a partir de la variable que determina si son beneficiados de alguna beca o no. En este análisis no se contemplan aun, que variables influyen en la diferencia de la distribución. Para esto, se crearon dos conjuntos, uno de becados y otro de no becados, en los cuales resultaron un total de 23 alumnos becados y un total de 46 alumnos no becados.

Para analizar si existía un comportamiento diferente entre los alumnos no becados y los que cuentan con beca, se realizó un contraste de hipótesis utilizando Kolmogorov Smirnov, donde la hipótesis nula se planteó como un comportamiento similar entre los dos conjuntos. Como resultado de este contraste y respondiendo al objetivo de esta investigación, se puede concluir que el gasto escolar es completamente distinto entre los alumnos que cuentan con beca y aquellos que no, a un nivel de confiabilidad al 95% por el planteamiento del contraste.

Una vez que se concluyó que la distribución entre las dos poblaciones es distinta, es de nuestro interés analizar cuáles con las principales variables que determinan este comportamiento. Para ello se realizó un análisis de varianzas ANOVA, en el cual se analizó el gasto en actividades recreativas o extraescolares, actividades escolares y en gasto en general.

Con resultado del análisis de varianzas se puede concluir que las variables que estadísticamente son más significativas en el gasto escolar son la generación a la que pertenecen y contar con beca.

Las variables que destacan en el gasto de recreación son los gastos de entretenimiento, así como los vicios o apuestas de los jóvenes y el ingreso que reciben mensualmente.

Como se puede apreciar en la imagen, para el gasto general y como esperábamos, las variables que influyen en su distribución del gasto es que los alumnos cuenten con un beneficio económico otorgado por la institución, seguida de la generación a la que pertenecen.

	Columna	Statistic	P-value
14	scholarship	11.376377	0.001240
15	ecnmy_affected	5.986762	0.004080
1	gen	5.001410	0.009497
2	income	2.235873	0.074896
10	home	2.607648	0.111049
5	transport	1.761201	0.188980
7	entertainment	0.824394	0.367153
9	technology	0.738448	0.393223
11	personal_care	0.316054	0.575865
12	savings	0.288005	0.593280
6	vices	0.182636	0.670488
13	debts	0.051741	0.820753
4	food	0.049971	0.823794
3	school	0.014615	0.904139
0	sex	0.006748	0.934774
8	outfit	0.000046	0.994616

5. Conclusiones

Después de observar y analizar la información, es natural pensar de manera rebuscada que los diferentes apoyos otorgados por la UNAM a estudiantes están siendo concedidos a quienes desde un punto de vista muy objetivo tal vez no los merecen, puesto que los resultados de la muestra arrojan ciertos indicadores que así lo señalan. Es ese sentido es fácil conjeturar que tal vez se pueden mejorar los filtros para ceder los apoyos, pues en los resultados observamos que la mayoría de los encuestados que contestaron tener algún apoyo económico, lo gastan en actividades extraescolares, es decir no se le da un uso académico que a primera instancia se pensaría que esa es su finalidad.

Ahora bien, no se está señalando que quienes son beneficiarios de alguna beca y gastan la misma en actividades que no son académicas, no debieran tener el apoyo, pues ese no es el objetivo de este trabajo. Sin embargo, queda la incertidumbre de que si los que no son favorecidos con algún sustento monetario le darían un uso “más adecuado”, es decir que si lo ocuparían para fines académicos.

Otro punto a resaltar en éste apartado, es que los alumnos que tienen beca presentan un gasto escolar totalmente distinto a los que no tienen, es decir, gastan de diferente manera su ingreso que perciben, lo cual de cierta manera es lógico puesto que se cree que quien tiene más gasta más. Situación que además se hace más notoria en cada generación, ya que en cada una se perciben ingresos y gastos similares internamente, pero que cambian entre cada una.

Por otro lado, el análisis de importancia de todas las variables presentes en el estudio, reveló que en la variable gasto extraescolar, la variable influyente en mayor grado fue la de tener o

no la beca, mientras en los otros dos tipos de gasto, la misma variable estuvo presente, en orden de importancia, en los primeros lugares.

Finalmente, el árbol de decisión nos ayuda a comprender mejor el comportamiento de los datos, siendo más específicos, nos ayuda a entender cómo es el proceder de los estudiantes entorno a cómo gastan su dinero condicionado a si tienen o no beca.

6. Bibliografía

Secretaría de Educación Pública. (2019). *Convocatoria. Beca de Manutención*. Obtenido en https://www.becarios.unam.mx/Portal2018/wp-content/uploads/2019/09/Manutenci%C3%B3n-UNAM-2020_compressed.pdf

Portal de Estadística Universitaria. (2018). Recuperado de: http://www.estadistica.unam.mx/series_inst/index.php (consultado 27 agosto 2018)

7. Anexo

Proyecto de Muestreo Noviembre 2019

Librerías:

In [1]:

```
import pandas as pd
pd.set_option('display.max_columns', 500)
```

A continuacion cargamos nuestros datos con las respuestas obtenidas

In [2]:

```
data = pd.read_csv('RESPUESTAS.csv', encoding = 'utf-8')
data.head(5) ## mostramos los 5 primero
```

Out[2]:

	Marca temporal	Género:	Generación:	Nombre:	¿Cuál es tu promedio? (Redondeado a 2 cifras)	¿Cuál es el monto de tu ingreso mensual?	¿En qué “actividades” gastas tu ingreso mensual? [Escuela (Libros, copias, útiles,etc)]
0	06/11/2019 13:23	Masculino	2017	Joshua David Ruiz Castañeda	7.90	2001–3000	Columna 1
1	06/11/2019 13:48	Masculino	2017	Erick Martínez Rodríguez	8.60	0–1000	NaN
2	06/11/2019 15:21	Masculino	2017	German	9.00	2001–3000	NaN
3	06/11/2019 15:51	Femenino	2017	Shirley Rabell	8.45	Más de \$4000	Columna 1
4	06/11/2019 16:38	Masculino	2017	Erick Ramón Reyes Valle	8.00	1001–2000	Columna 1

Limpieza de datos

In [3]:

```
data.columns = [x.strip() for x in data.columns]

data = data[[x for x in data.columns if x not in ['Marca temporal', 'Nombre:']]]
        .copy()

data = data.fillna(0)
```

Ya que tenemos los datos limpios, procederemos a renombrar las columnas

In [4]:

```
df = data.copy()
df.columns = ['sex', 'gen', 'grade', 'income', 'school', 'food', 'transport', 'v
ices',\
              'entertainment', 'outfit', 'technology', 'home', 'personal_care', '
savings',\
              'debts', 'scholarship', 'amount_of_shsp', 'recreation_amount', 'sch
olar_amount',\
              'ecnmy_affected', 'recreation_income_amount', 'scholar_income_ammoun
t']
```

Diccionario de variables

In [5]:

```
pd.DataFrame(dict(zip(data.columns, df.columns)), index = ['Label']).T
```

Out[5] :

	Label
Género:	sex
Generación:	gen
¿Cuál es tu promedio? (Redondeado a 2 cifras)	grade
¿Cuál es el monto de tu ingreso mensual?	income
¿En qué “actividades” gastas tu ingreso mensual? [Escuela (Libros, copias, útiles,etc)]	school
¿En qué “actividades” gastas tu ingreso mensual? [Comida]	food
¿En qué “actividades” gastas tu ingreso mensual? [Transporte (Transporte público y/o combustible)]	transport
¿En qué “actividades” gastas tu ingreso mensual? [Vicios (Alcohol, Apuestas, Cigarros,etc)]	vices
¿En qué “actividades” gastas tu ingreso mensual? [Entretenimiento (Fiestas, Netflix, Compras, Spotify, Cine, Conciertos,etc)]	entertainment
¿En qué “actividades” gastas tu ingreso mensual? [Ropa y/o calzado]	outfit
¿En qué “actividades” gastas tu ingreso mensual? [Tecnología (Teléfonos inteligentes, computadoras, lap top, videojuegos)]	technology
¿En qué “actividades” gastas tu ingreso mensual? [Apoyo familiar (Gasto, Renta, Pago de servicios, Despensa)]	home
¿En qué “actividades” gastas tu ingreso mensual? [Belleza e higiene personal]	personal_care
¿En qué “actividades” gastas tu ingreso mensual? [Ahorro]	savings
¿En qué “actividades” gastas tu ingreso mensual? [Deudas]	debts
¿Cuentas con el apoyo económico de Manutención o Beca de Excelencia?	scholarship
En un número aproximado, ¿Cuál es el monto que recibes (mensualmente) al ser beneficiario de este apoyo?	amount_of_shsp
¿Cuánto dinero del apoyo dedicas a actividades extra-escolares y de entretenimiento al mes?	recreation_amount
¿Cuánto dinero del apoyo dedicas a actividades escolares al mes?	scholar_amount
El NO ser beneficiario del apoyo de Manutención, ¿Afecta en tu economía?	ecnmy_affected
¿Cuánto dinero de tu ingreso mensual dedicas a actividades extra-escolares y de entretenimiento al mes?	recreation_income_amount
¿Cuánto dinero de tu ingreso mensual dedicas actividades escolares al mes?	scholar_income_amount

In [6]:

```
df['scholarship'].value_counts(True)
```

Out[6]:

```
No      0.666667
Si      0.333333
Name: scholarship, dtype: float64
```

In [7]:

```
df['ecnmy_affected'].head(10)
```

Out[7]:

```
0      Si
1      Si
2      No
3       0
4      No
5      Si
6      No
7      No
8      Si
9      Si
Name: ecnmy_affected, dtype: object
```

In [8]:

```
dict2 = {'Columna 1':1, 0:0}
var_bin = ['school', 'food', 'transport', 'vices','technology','entretainment',
'outfit',\
          'home', 'personal_care', 'savings','debts']
for v in var_bin:
    df[v] = df[v].map(dict2)
```

In [9]:

```
df['ecnmy_affected'] = df['ecnmy_affected'].replace(0,'Tiene Beca')
```

In [10]:

```
df['ecnmy_affected'].head(10)
```

Out[10]:

```
0      Si
1      Si
2     No
3  Tiene Beca
4     No
5      Si
6     No
7     No
8      Si
9      Si
Name: ecnmy_affected, dtype: object
```

In [11]:

```
df.head()
```

Out[11]:

	sex	gen	grade	income	school	food	transport	vices	entertainment	outfit	t
0	Masculino	2017	7.90	2001–3000	1	1	1	1	1	0	
1	Masculino	2017	8.60	0–1000	0	0	1	0	0	0	
2	Masculino	2017	9.00	2001–3000	0	1	1	0	0	0	
3	Femenino	2017	8.45	Más de \$4000	1	1	1	0	1	0	
4	Masculino	2017	8.00	1001–2000	1	1	0	0	1	0	

Para poder continuar con el análisis y seguir con la limpieza y el diseñado de datos, para ajustar algún modelo, debemos ver cómo se encuentra la distribución de nuestro evento de interés el cuál es la beca.

In [12]:

```
df['scholarship'].value_counts(True)
```

Out[12]:

```
No      0.666667
Si       0.333333
Name: scholarship, dtype: float64
```

In [13]:

```
df['income'].value_counts(True)
```

Out[13]:

```
$1001 - $2000      0.333333
$2001 - $3000      0.260870
$3001 - $4001      0.173913
$0 - $1000         0.159420
Más de $4000       0.072464
Name: income, dtype: float64
```

Ahora revisaremos si existe alguna diferencia entre la distribución del gasto para recreación y escolar discriminando las poblaciones a partir de si son becadas o no.

Para esto, crearemos dos conjuntos, uno de becados y otro de no becados, esto se hará de la siguiente forma:

In [14]:

```
becados = df[df['scholarship']=='Si']
no_becados = df[df['scholarship']=='No']
```

Ahora revisaremos cómo se compara en cuanto a longitudes, es decir, cuánto tenemos de población en cada uno:

In [15]:

```
print('Tenemos un total de %s alumnos becados y un total de %s alumnos no becados.' % (len(becados), len(no_becados)))
```

Tenemos un total de 23 alumnos becados y un total de 46 alumnos no becados.

Ahora procederemos a realizar el contraste de hipótesis utilizando Kolmogorov Smirnov.

Prueba KS

In [16]:

```
from scipy.stats import ks_2samp
```

In [17]:

```
ks_2samp(becados['recreation_amount'], no_becados['recreation_income_amount'])
```

Out[17]:

```
Ks_2sampResult(statistic=0.15217391304347827, pvalue=0.8587491858195904)
```

In [18]:

```
ks_2samp(becados['scholar_amount'], no_becados['scholar_income_amount'])
```

Out[18]:

```
Ks_2sampResult(statistic=0.43478260869565216, pvalue=0.004956236799983138)
```

Utilizando ANOVA o análisis de varianza.

In [19]:

```
df.head(1)
```

Out[19]:

	sex	gen	grade	income	school	food	transport	vices	entertainment	outfit	t
0	Masculino	2017	7.9	2001–3000	1	1	1	1	1	0	

In [20]:

```
df['sch_expense'] = df['scholar_amount'] + df['scholar_income_amount']
df['rec_expense'] = df['recreation_amount'] + df['recreation_income_amount']
df['tot_expense'] = df['rec_expense'] + df['sch_expense']
```

In [21]:

```
var_c = ['grade', 'amount_of_shsp', 'recreation_amount', 'scholar_amount', 'recreation_income_amount', \
         'scholar_income_amount', 'sch_expense', 'rec_expense', 'tot_expense']
var_d = [x for x in df.columns if x not in var_c]
```


In [23]:

```
import scipy.stats as st

print('Esta tabla consta de los estadísticos y los p-values de la prueba de hipótesis de ANOVA.')
print('Con esto podemos observar qué variables es la que tien mayor impacto en el gasto escolar.')

c_hipo = []
for c in var_d:
    data = []
    se = set(df[c])
    for s in se:
        data.append(df[df[c] == s]['sch_expense'].astype(float))
    c_hipo.append([c, st.f_oneway(*data)[0], st.f_oneway(*data)[1]])
c_hipo_df = pd.DataFrame(c_hipo)
c_hipo_df.columns = ['Columna', 'Statistic', 'P-value']
c_hipo_df.sort_values(by='P-value')
```

Esta tabla consta de los estadísticos y los p-values de la prueba de hipótesis de ANOVA.
Con esto podemos observar qué variables es la que tien mayor impacto en el gasto escolar.

Out[23]:

	Columna	Statistic	P-value
1	gen	8.241669	0.000638
14	scholarship	10.564479	0.001805
15	ecnmy_affected	5.941853	0.004238
6	vices	2.404744	0.125679
10	home	2.292033	0.134743
5	transport	1.286026	0.260823
4	food	1.027895	0.314301
8	outfit	0.638159	0.427201
3	school	0.497028	0.483253
7	entertainment	0.432674	0.512933
2	income	0.689036	0.602209
11	personal_care	0.245771	0.621691
0	sex	0.211159	0.647348
13	debts	0.145056	0.704510
12	savings	0.029972	0.863077
9	technology	0.000005	0.998185

Podemos apreciar que las que tienen un impacto significativo son generación y beca, pues ecnmy_affected es el complemento de la de scholarship

In [24]:

```
print('Esta tabla consta de los estadísticos y los p-values de la prueba de hipótesis de ANOVA.')
print('Con esto podemos observar qué variables es la que tien mayor impacto en el gasto recreativo.')

c_hipo = []
for c in var_d:
    data = []
    se = set(df[c])
    for s in se:
        data.append(df[df[c] == s]['rec_expense'].astype(float))
    c_hipo.append([c, st.f_oneway(*data)[0], st.f_oneway(*data)[1]])
c_hipo_df = pd.DataFrame(c_hipo)
c_hipo_df.columns = ['Columna', 'Statistic', 'P-value']
c_hipo_df.sort_values(by='P-value')
```

Esta tabla consta de los estadísticos y los p-values de la prueba de hipótesis de ANOVA.
Con esto podemos observar qué variables es la que tien mayor impacto en el gasto recreativo.

Out[24]:

	Columna	Statistic	P-value
7	entertainment	9.317371	0.003253
6	vices	3.906986	0.052204
2	income	2.444259	0.055420
9	technology	2.675543	0.106590
8	outfit	2.039540	0.157901
4	food	1.954608	0.166702
3	school	1.073305	0.303925
0	sex	0.960163	0.330671
1	gen	0.956865	0.389358
12	savings	0.484382	0.488852
5	transport	0.200224	0.655984
14	scholarship	0.198127	0.657672
10	home	0.094930	0.758957
13	debts	0.064378	0.800483
15	ecnmy_affected	0.205214	0.814991
11	personal_care	0.027718	0.868276

Es curioso que para modelar este gasto la variable que realmente nos importa o interesa es la de tecnología, esta variable binaria nos indica que el gasto en tecnología se comporta de forma distinta, según el valor que tome.

In [25]:

```
print('Esta tabla consta de los estadísticos y los p-values de la prueba de hipótesis de ANOVA.')
print('Con esto podemos observar qué variables es la que tien mayor impacto en el gasto total.')

c_hipo = []
for c in var_d:
    data = []
    se = set(df[c])
    for s in se:
        data.append(df[df[c] == s]['tot_expense'].astype(float))
    c_hipo.append([c, st.f_oneway(*data)[0], st.f_oneway(*data)[1]])
c_hipo_df = pd.DataFrame(c_hipo)
c_hipo_df.columns = ['Columna', 'Statistic', 'P-value']
c_hipo_df.sort_values(by='P-value')
```

Esta tabla consta de los estadísticos y los p-values de la prueba de hipótesis de ANOVA.

Con esto podemos observar qué variables es la que tien mayor impacto en el gasto total.

Out[25]:

	Columna	Statistic	P-value
14	scholarship	11.376377	0.001240
15	ecnmy_affected	5.986762	0.004080
1	gen	5.001410	0.009497
2	income	2.235873	0.074896
10	home	2.607648	0.111049
5	transport	1.761201	0.188980
7	entertainment	0.824394	0.367153
9	technology	0.738448	0.393223
11	personal_care	0.316054	0.575865
12	savings	0.288005	0.593280
6	vices	0.182636	0.670488
13	debts	0.051741	0.820753
4	food	0.049971	0.823794
3	school	0.014615	0.904139
0	sex	0.006748	0.934774
8	outfit	0.000046	0.994616

De igual forma que en el caso anterior, tenemos que el tener o no una beca afecta directamente el comportamiento del gasto total. Seguido por una combinación de ecnmy_affected, que quedamos que es una transformación de la variable de beca y por último tenemos la de generación.

Árbol de decisión