# Análisis de datos proyecto ESPINA

Franklin Steven De la Cruz Paucar

[1]Yachay Tech University, Urcuquí, Ecuador
franklin.de@yachaytech.edu.ec

**Abstract.** En este estudio se examinó el efecto de la profesión de los padres (no agricultores no floricultores, solo agricultor, solo floricultor, y agricultor y floricultor ) en el desarrollo corporal y nutricional de adultos. Se evaluaron diferentes variables, como el índice de masa corporal, la altura, entre otros indicadores de salud. Los resultados sugieren que la profesión de los padres tiene un impacto no significativo en el desarrollo corporal y nutricional de los participantes. En general, se encontró que aquellos cuyos padres tenían relación con algún área de la floricultura en cuestión de altura no presentaban diferencia con aquellos cuyos padres no tenián relacíon con una área de la floricultura. Sin embargo, para los resultados nutricionales tales como el grado de anémia, los participantes que tenían anemia leve pertenecían en su mayoría a aquellos cuyos padres tenían algura relación con la floricultura. Estos hallazgos pueden ser útiles para informar políticas públicas y programas de intervención dirigidos a mejorar la salud y el bienestar de la población adulta en cuya niñez sus padres tuvieron un oficio relacionado a la producción de flores.

**Keywords:** anemia · floricultura · nutrición

## 1 Descripción de la base de datos

La base de datos contiene información recolectada en tres periodos de tiempo: 2008, 2016 y 2022. Cada base de datos contiene información acerca de la profesión de los padres de los niños, su estado nutricional, sus medidas antropométricas, entre otros.

## 2 Metodología

Para el primer análisis se utulizará la base de datos del 2022 que contiene la altura de los participantes y se unirá con la base de datos del 2008 donde se especifica la profesión de los padres de los participantes para comprobar a que categoría pertenecen aquellos participantes cuya altura es menor a la promedio en Ecuador, 167cm para hombres y 154 cm para mujeres.

El segundo análisis propuesto es similiar al anterior mencionado, se utilizará la base de datos del 2016 donde existe información acerca del estado nutricional de los participantes dividio en: 1severe anemia, 2moderate anemia, 3mild anemia, y 4normal. Se unirá con la base de datos del 2008 para comprobar si su estado nutricional tiene relación con la profesión de sus padres en la niñez.

# 3 Resultados

## 3.1 Primer análisis

Para este análisis se toma como referencia la altura promedio de hombres y mujeres en Ecuador.

```
import pandas as pd
import matplotlib.pyplot as plt

# Leer archivo de Excel en un objeto de DataFrame
df = pd.read_excel('archivo.xlsx')

# Contar el total por cada categor a
total_por_categoria = df.groupby('a_categocup').count()

# Crear gr fico de pastel
plt.pie(total_por_categoria['a_noflonoagr'], labels=
                                total_por_categoria.index,
                                autopct='%1.1f%%')
plt.title('Total por categor a')
plt.show()
```

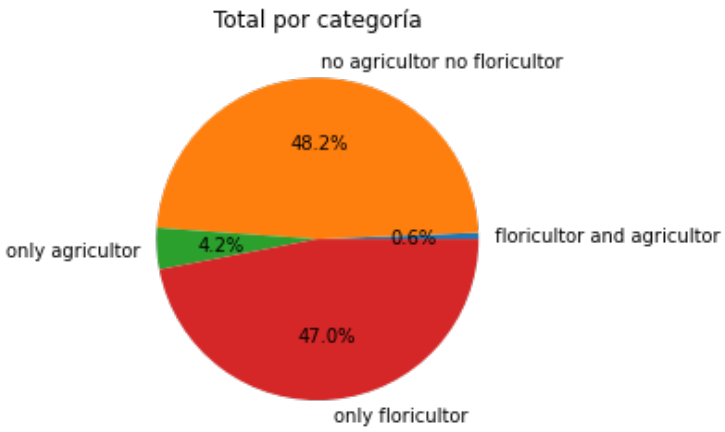


**Fig. 1.** Distribución de la base de datos del 2008 con la relación de los padres de los participantes con el área de floricultura

Se observa una base de datos bien distribuída, haciendo incapié en que la mitad del mismo hace referencia a padres que tienen relación con el área de floricultura, y padres que no. Una vez observada la distribución de la base de datos, se procede a unir los resultados de altura en 2022 y la profesión de los padres de los participantes en el 2008.

```
import pandas as pd

# Leer los dos archivos de Excel en objetos de DataFrame
df1 = pd.read_excel('/content/ESP 22 Base de datos de
                                antropometria 230322.xlsx')
df2 = pd.read_excel('/content/archivo.xlsx')

# Seleccionar las columnas necesarias del archivo 1
df1 = df1.loc[:, ['nid','d_gender', 'd_height']]
# Seleccionar las columnas necesarias del archivo 2
df2 = df2.loc[:, ['nid', 'a_categocup']]

# Combinar los dos DataFrames utilizando el ID como clave de
                                uni n
df_combinado = pd.merge(df1, df2, on='nid', how='left')

# Guardar el DataFrame combinado en un nuevo archivo de Excel
df_combinado.to_excel('archivo_combinado.xlsx', index=False)
```

Para esta parte de la experimentación se encontró que hay varios participantes del 2022 que no constan en la base da datos del 2008, por lo que se procede a colocar en la profesión de sus padres la palabra 'no assigned´.

```
import pandas as pd

# Leer el archivo Excel en un objeto de DataFrame
df = pd.read_excel('archivo_combinado.xlsx')

# Reemplazar los valores vac os en la columna a_categocup
                                con la cadena 'no assigned'
df['a_categocup'] = df['a_categocup'].fillna('no assigned')

# Guardar el DataFrame actualizado en un nuevo archivo de
                                Excel
df.to_excel('archivo_actualizado.xlsx', index=False)
```

Se separa por género, y se seleccionan aquellos particpantes hombres con una estatura menor a 167cm y a que categoría pertenecen en relación a la profesión de sus padres.

```
import pandas as pd
import matplotlib.pyplot as plt

# Leer el archivo Excel en un objeto de DataFrame
df = pd.read_excel('hombres.xlsx')

# Filtrar las filas donde la columna 'estatura' es menor a
                                167.0
df_filtrado = df.loc[df['d_height'] < 167.0]
```

```
# Guardar el DataFrame filtrado en un nuevo archivo de Excel
df_filtrado.to_excel('hombres_estatura_profesion.xlsx', index
                                =False)

print(df_filtrado)

#Se grafica el total de hombres con estatura menor a 167.0 en
                                 relaci'on a la profesi n de
                                sus padres

# Leer archivo de Excel en un objeto de DataFrame
df = pd.read_excel('hombres_estatura_profesion.xlsx')

# Contar el total por cada categor a
total_por_categoria = df.groupby('a_categocup').count()

# Crear gr fico de pastel
plt.pie(total_por_categoria['d_gender'], labels=
                                total_por_categoria.index,
                                autopct='%1.1f%%')
plt.title('Total por categor a menores a 167.0')
plt.show()
```

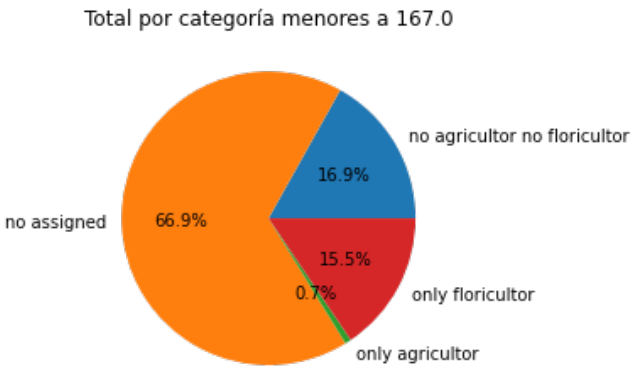


**Fig. 2.** Total dividido por relación de padres con la floricultura y la estatura de los particpantes hombres menor al promedio

Se puede observar que la mitad de los participantes con una estatura menor al promedio se encuentran bien distriubuídos entre aquellos padres que en su niñez tenían alguna relación con la floricultura y a aquellos que no. Tenemos un margen de mejora que serían los participantes que no constan en la base de datos del 2008, por lo que los resultados podrían variar. Obviando los participantes de los

cuales no se tiene información sobre la relación de sus padres con la floricultura en su niñez, los resultados son los siguientes. Se realiza el mismo análisis para

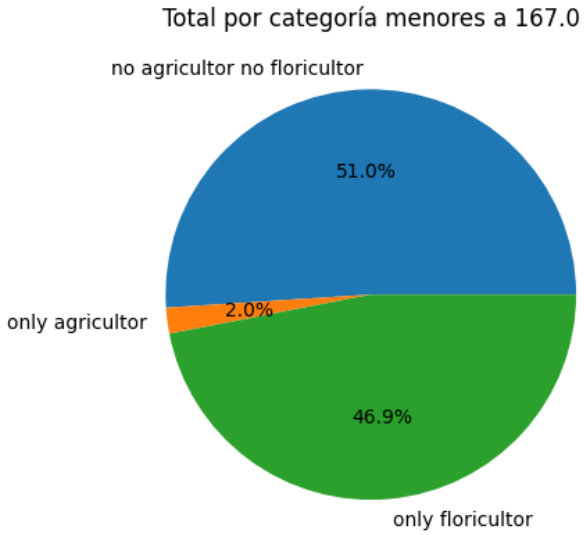


**Fig. 3.** Total dividido por relación de padres con la floricultura y la estatura de los particpantes hombres menor al promedio

las mujeres, y los reslutados son los siguientes:

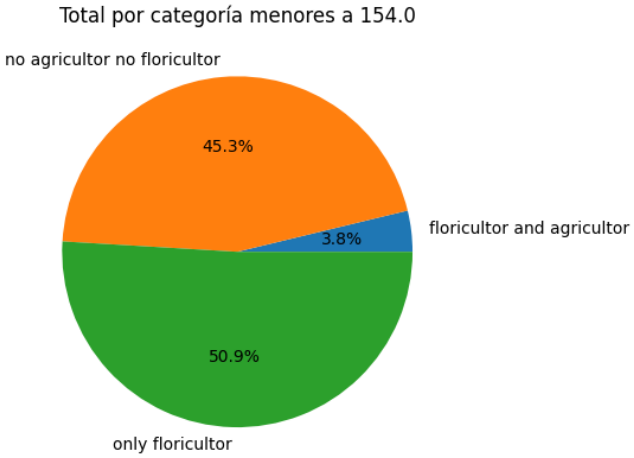


**Fig. 4.** Total dividido por relación de padres con la floricultura y la estatura de las particpantes mujeres menor al promedio

Se puede observar que la mayoría de mujeres con estatura menor al promedio se encuentran en la categoría de aquellos padres que tenían relación con el área de floricultura durantes su niñez, pero la diferencia no es significativa.

### 3.2 Segundo análisis

Teniendo en claro los procedimientos realizados anteriormente, la metodología se repitió para evaluar es estado de nutrición de los participantes del 2016 versus la relación que tenían sus padres con el área de la floricultura en 2008. Primero se observa la distribución de los resultados nutricionales de los participantes en el 2016.

Distribución total de datos de nutrición por categoría:

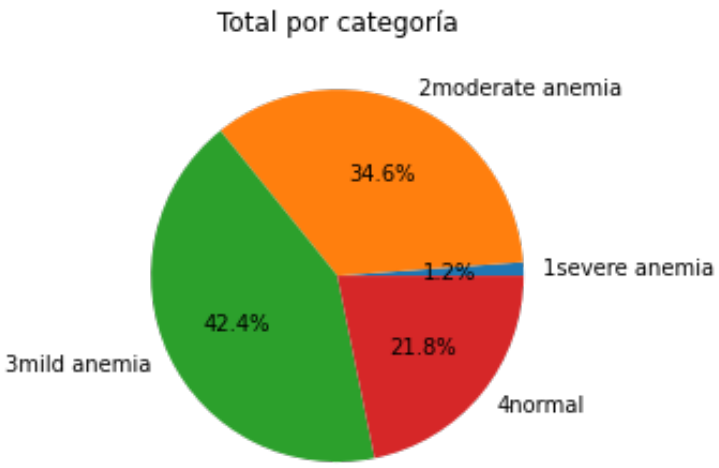


**Fig. 5.** Distribución de los datos de los participantes de acuerdo a su estado nutricional en 2016

Una vez con los datos, se hace una unión con la base de datos del 2008, para evidenciar de acuerdo a la categoría del estado nutricional de los participantes, la relación que tuvieron sus padres con la floricultura durante su niñez. Se obtiene los siguientes resultados dividio por estado nutricional.

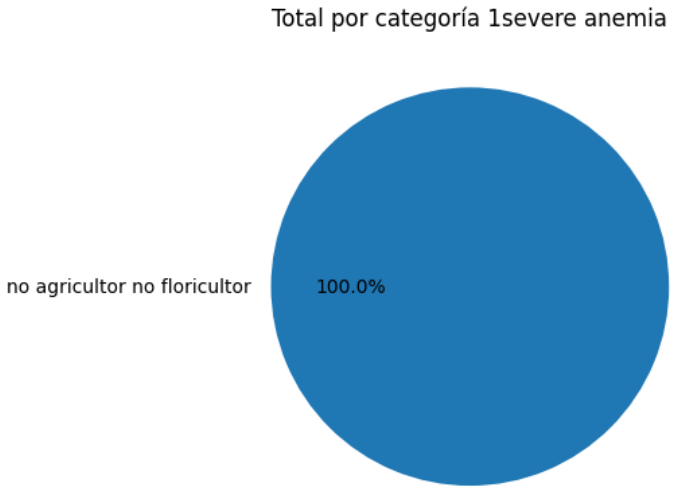


**Fig. 6.** Participantes con anemia severa versus la relación que tenían sus padres con la floricultura en su niñez

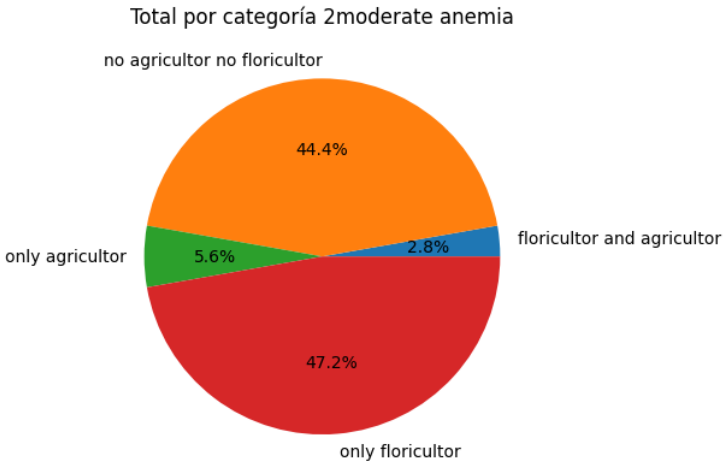


**Fig. 7.** Participantes con anemia moderada versus la relación que tenían sus padres con la floricultura en su niñez

Dado que la visualización por cada categoría varía mucho en cuanto al estado nutricional de los participantes, se procede a hacer un análisis chi cuadrado.

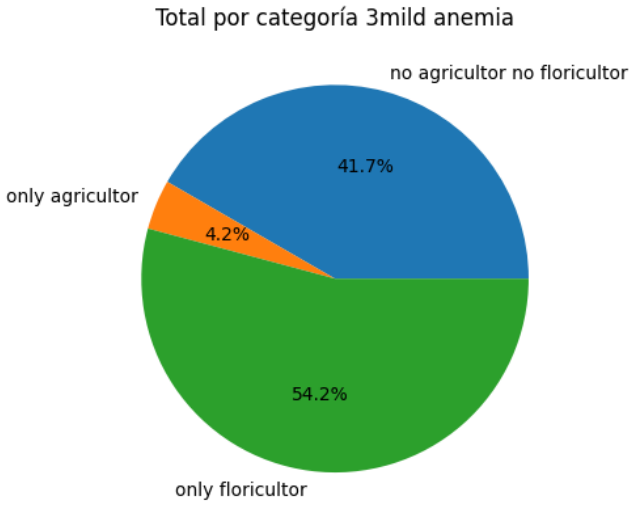


**Fig. 8.** Participantes con anemia leve versus la relación que tenían sus padres con la floricultura en su niñez

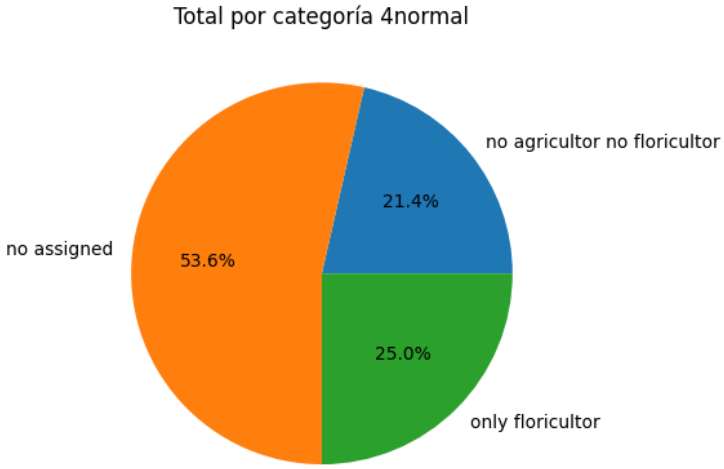


**Fig. 9.** Participantes con nutrición normal versus la relación que tenían sus padres con la floricultura en su niñez

Este análisis estadístico se utiliza para evaluar la relación entre dos variables categóricas. Esta técnica se utiliza comúnmente para analizar datos que se presentan en forma de tablas de contingencia. La idea básica detrás del análisis de chi cuadrado es comparar la distribución observada de frecuencias de las dos

variables categóricas con la distribución esperada si no hubiera relación entre ellas. Si la distribución observada es significativamente diferente de la distribución esperada, entonces se concluye que hay una relación significativa entre las dos variables. En este caso nuestras variables son la relación del padre con el área de floricultura y el estado de nutrición de los participantes.

```
import pandas as pd
from scipy.stats import chi2_contingency

# Leer los datos del archivo de Excel y convertirlos en un
                                dataframe de pandas
datos = pd.read_excel('inner_actualizado2016.xlsx')
print(datos)
# Crear una matriz de contingencia con los datos
matriz_contingencia = pd.crosstab(datos['a_categocup'], datos
                                ['b_hgb_status'])

# Realizar el an lisis de chi cuadrado y obtener los
                                resultados
estadistico, p_valor, grados_libertad, esperados =
                                chi2_contingency(
                                matriz_contingencia)

# Imprimir los resultados
print("Estad stico de chi cuadrado:", estadistico)
print("P-valor:", p_valor)
print("Grados de libertad:", grados_libertad)
print("Valores esperados:", esperados)
```

```
Estad stico de chi cuadrado: 6.296074373969111
P-valor: 0.7099590575117956
Grados de libertad: 9
```

El resultado del análisis de chi cuadrado obtenido indica que no hay una relación significativa entre las variables categóricas se está evaluando.

El estadístico de chi cuadrado obtenido es de 6.296 y los grados de libertad son 9. El p-valor obtenido es de 0.71, lo que indica que no hay suficiente evidencia estadística para rechazar la hipótesis nula, es decir, no hay una relación significativa entre las variables.

En resumen, el análisis de chi cuadrado sugiere que las variables no están relacionadas de manera significativa en el dataset.

## 4 Puntaje Z para antropometria

En esta sección se presentará los gráficos de los puntajes Z para la estatura de los participantes en 2022. Al observar una gráfica Z-score, podemos inferir varias cosas, entre ellas:

La forma de la distribución: Si los valores Z se distribuyen de manera normal alrededor de cero, entonces la distribución original también es una distribución normal. Si los valores Z no se distribuyen de manera normal, entonces la distribución original no es normal.

Valores atípicos o extremos: Los valores Z que se encuentran muy por encima o por debajo de cero (es decir, Z mayor 3 o Z menor -3) pueden indicar valores atípicos o extremadamente raros en la distribución.

La posición relativa de un valor en la distribución: Podemos ver si un valor particular está cerca del promedio o si está muy alejado del promedio en términos de desviaciones estándar.

La simetría o asimetría de la distribución: Podemos ver si la distribución es simétrica o asimétrica. Si la distribución es simétrica, los valores Z deben estar distribuidos de manera uniforme a ambos lados de cero. Si la distribución es asimétrica, los valores Z deben estar más concentrados en un lado de cero que en el otro.

```
Media:  159.10831683168317
Desviacion estandar:  8.82315417762786
```

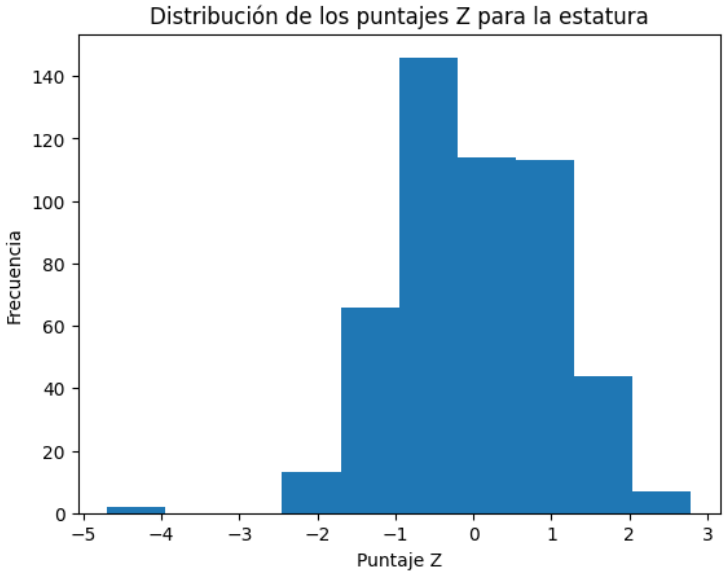


**Fig. 10.** Distribucion de puntaje Z para la estatura de todos los participantes

Teniendo en centa lo anteriormente mencionado, se puede observar que generalmente la distribución de los valores para la estatura de todos los participantes no es simétrica, y tiene valores atípicos, por lo tanto su distribución no es normal.

Por lo tanto vamos a dividir por género la gráfica de la distribución de los puntajes Z para visualizar de manera específica donde se encuentra el valor atípico. Si hay valores atípicos en los datos, esto puede afectar significativamente los resultados del análisis estadístico. Por ejemplo, un valor atípico puede aumentar o disminuir significativamente la media y la desviación estándar de la muestra, lo que puede afectar la interpretación de los resultados.

Además, si la distribución no es normal, puede haber problemas en la aplicación de ciertos análisis estadísticos que asumen una distribución normal de los datos. Esto puede llevar a resultados inexactos o a interpretaciones erróneas de los resultados.

Es importante tener en cuenta que la presencia de valores atípicos y una distribución no normal no necesariamente invalida los resultados del análisis estadístico, pero puede requerir la aplicación de técnicas estadísticas específicas para abordar estos problemas. Por ejemplo, existen pruebas estadísticas no paramétricas que no requieren que los datos sigan una distribución normal, y hay técnicas para detectar y manejar valores atípicos.

### 4.1 Puntaje Z estatura hombres

```
Media:  165.54979919678715
Desviacion estandar:  6.03096510687308
```

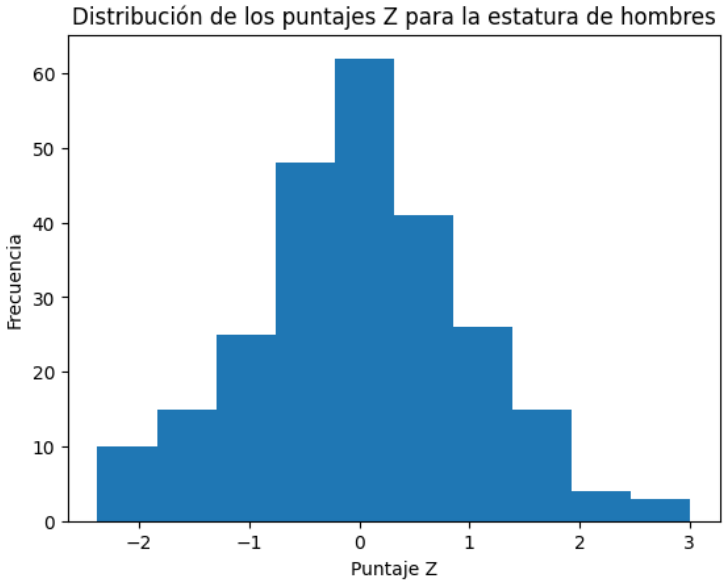


**Fig. 11.** Distribucion de puntaje Z para la estatura de todos los participantes hombres

La distribución de la estatura de los hombres se puede observar que tiene una distribución univorme alrededor de 0, por lo tanto tiene una forma normal, y sin valores atípicos.

### 4.2 Puntaje Z estatura mujeres

```
Media:  152.84296875
Desviacion estandar:  6.2110182286420175
```

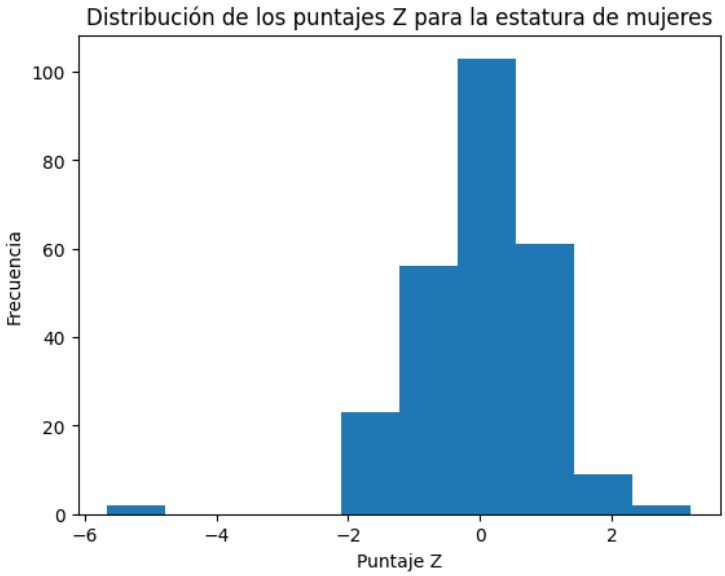


**Fig. 12.** Distribucion de puntaje Z para la estatura de todos los participantes mujeres

Se puede observar que en la distribución de estatura para las mujeres, hay un valor atípico claro, cuyo valor es menor a (-3)

## 5 Puntaje Z colinesterasa

Para esta sección se realizará el cálculo del puntaje Z para **d_Hgb_original** y **d_hgb_corregido**

```
Media:  13.787326732673266
Desviacion estandar:  1.5169790792588822
```

```
Media:  14.695209580838323
Desviacion estandar:  1.7781937633082308
```

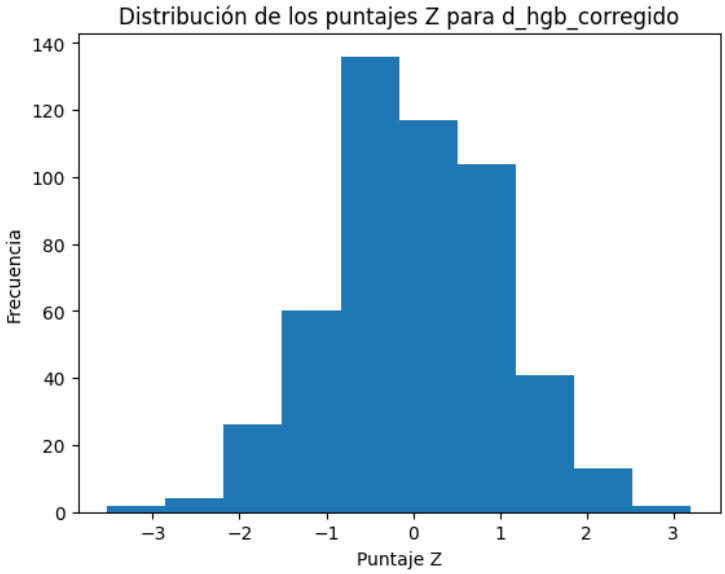


**Fig. 13.** Distribucion de puntaje Z para hgb corregido

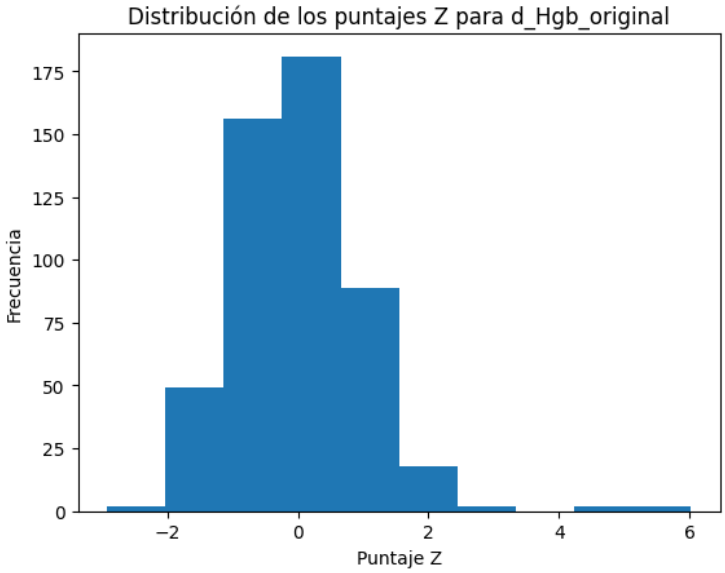


**Fig. 14.** Distribucion de puntaje Z para hgb corregido

## 6 Conclusiones

De acuerdo a la base de datos proporcionada, no se puede concluir de que la relación de los padres con el área de la floricultura en la niñez de los participantes haya tenido un impacto notable en su altura y nutrición. Sin embargo, hay que recalcar, que como se mencionó en la metodología, hay varios datos faltantes de participantes del 2016 y 2022 en relación a la profesión de sus padres. Por consiguiente, esta información faltante podría darnos un panoráma completamente diferente. Además que no hay información sobre su desarrollo intelectual para poder evaluar la correlacón de los particpantes con la exposición a plagicidas de manera directa o indirecta y su efecto en su desarrollo intelectual, tales como su memoria. Es importante tener en cuenta que la presencia de valores atípicos y una distribución no normal no necesariamente invalida los resultados del análisis estadístico, pero puede requerir la aplicación de técnicas estadísticas específicas para abordar estos problemas. Por ejemplo, existen pruebas estadísticas no paramétricas que no requieren que los datos sigan una distribución normal, y hay técnicas para detectar y manejar valores atípicos.

En resumen, al encontrarnos con valores atípicos y una distribución no normal en nuestros datos, es importante ser conscientes de estas limitaciones y aplicar técnicas estadísticas adecuadas para abordar estos problemas y obtener resultados precisos e interpretables.

Link de acceso al código presentado:

https://colab.research.google.com/drive/1HHxOusA5HYH-J3eRCs$_Zr9eFBhxR-VUr?usp = sharing$