



EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO®

Instituto Tecnológico de Aguascalientes

ESTADISTICA INFERENCIAL II PROYECTO-EXAMEN UNIDAD 4

PRESENTA:
GARCIA MEDINA LUIS FERNANDO

CARRERA:
INGENIERIA EN GESTION EMPRESARIAL

Nombre del maestro: José Cruz

06/12/2025

Introducción

El rendimiento académico de los estudiantes es un fenómeno complejo influido por múltiples factores personales, familiares, escolares y socioeconómicos. Comprender cómo interactúan estas variables resulta fundamental para identificar oportunidades de mejora educativa y generar estrategias que fortalezcan el aprendizaje. En este proyecto se analizan diversos elementos que pueden incidir en el puntaje obtenido en los exámenes, tales como las horas de estudio, la asistencia, el nivel de motivación, la participación parental, el acceso a recursos, el tipo de escuela, la calidad del profesorado, entre otros.

A través de un análisis estadístico basado en ANOVA y pruebas de comparación múltiple de Tukey, se busca determinar qué variables ejercen una influencia significativa en el rendimiento académico y cuáles no presentan efectos estadísticamente relevantes. El propósito central es aportar evidencia que permita comprender de manera más clara qué condiciones favorecen o limitan el desempeño escolar, ofreciendo así un panorama integral que pueda utilizarse como base para futuras investigaciones o propuestas educativas.

Marco Teórico

Para interpretar los resultados del análisis se utilizaron conceptos clásicos de estadística inferencial y técnicas de análisis comparativo. A continuación, se definen y explican los conceptos empleados:

ANOVA (Análisis de Varianza)

Es una prueba estadística que permite comparar las medias de tres o más grupos para determinar si existen diferencias significativas entre ellos. Su objetivo es evaluar si la variación observada en la variable dependiente (puntaje de examen) se debe al efecto de las variables independientes analizadas o simplemente al azar. En este estudio, ANOVA permitió identificar qué factores influyen realmente en el rendimiento académico.

Modelo lineal (OLS – Ordinary Least Squares)

Es un método de regresión que estima la relación entre una variable dependiente y una o más variables independientes. Sirve como base para la construcción del modelo utilizado en el ANOVA. El modelo lineal aplicado incluyó variables numéricas y categóricas para evaluar su impacto sobre el puntaje del examen.

Prueba de Tukey (Tukey HSD – Honest Significant Difference)

Es una prueba post hoc utilizada después del ANOVA. Permite comparar todas las parejas posibles de grupos dentro de una variable categórica para identificar entre qué grupos existen diferencias significativas. En este estudio se aplicó para variables como participación parental, nivel de motivación, acceso a internet, entre otras, ayudando a determinar qué categorías presentan mejores o peores puntajes.

Variable dependiente

Es la variable principal que se desea explicar o predecir. En este estudio, la variable dependiente es el puntaje del examen, que representa el rendimiento académico del estudiante.

Variables independientes

Son las variables que se analizan para determinar si influyen en la variable dependiente. En el análisis se incluyeron variables individuales (horas de estudio, asistencia, horas de sueño), familiares (participación parental, ingreso familiar, nivel de estudios de los padres), escolares (tipo de escuela, calidad del profesorado, influencia de compañeros) y socioeconómicas (acceso a internet, recursos educativos, distancia al centro educativo).

Variables categóricas (C())

Son aquellas que representan grupos o categorías (por ejemplo: “bajo, medio, alto”). En el modelo se transformaron mediante la función C() para que el software estadístico pudiera incluirlas correctamente en el ANOVA. Estas variables permiten comparar el desempeño entre diferentes niveles o grupos.

Nivel de significancia ($\alpha = 0.05$)

Es la probabilidad máxima aceptable de cometer un error al rechazar una hipótesis nula verdadera. En este estudio se utilizó un nivel de significancia estándar del 5%, lo que significa que una variable se considera significativa si su valor p es menor a 0.05.

Valor p (p-value)

Indica la probabilidad de que los resultados observados ocurran si la hipótesis nula fuera cierta. Valores p pequeños (menores a 0.05) indican que la diferencia entre grupos es estadísticamente significativa. Por ejemplo, horas de estudio y asistencia presentaron valores p de 0.000, lo que confirma su alta influencia en el rendimiento.

Grados de libertad (df)

Representan la cantidad de información disponible para estimar los parámetros del modelo. En el ANOVA, los grados de libertad se dividen entre los factores evaluados y el residuo, indicando la variabilidad explicada y no explicada.

Intervalos de confianza (IC)

En las gráficas de Tukey, los intervalos muestran el rango dentro del cual se espera que esté la diferencia real entre medias. Cuando el IC no incluye el cero, significa que existe una diferencia significativa entre grupos.

Datos y librerías utilizadas

Para realizar el análisis se empleó un dataset con más de 6,300 observaciones y las librerías:

- **pandas** (manejo y limpieza de datos)
- **statsmodels** (ANOVA y Tukey)
- **matplotlib** (gráficos)
- **ols** (construcción del modelo lineal)

Metodología

La base de datos utilizada contenía más de 6,300 registros, cada uno correspondiente a información académica, personal, familiar y escolar de distintos estudiantes. Antes de realizar cualquier análisis estadístico, se llevó a cabo un proceso de limpieza para eliminar registros incompletos o con valores faltantes, garantizando así la calidad y consistencia del conjunto de datos.

Posteriormente, se identificaron las variables de interés. La variable dependiente fue el puntaje del examen, mientras que las variables independientes incluyeron horas de estudio, asistencia, nivel de motivación, participación parental, acceso a internet, calidad del profesorado, ingreso familiar, distancia, actividades extracurriculares, entre otras. Las variables categóricas fueron codificadas adecuadamente mediante la función `C()` para ser incorporadas en el modelo.

Una vez depurado el dataset, se construyó un modelo lineal (OLS) que serviría como base para el Análisis de Varianza (ANOVA). Dicho modelo permitió estimar el efecto de cada variable independiente sobre el puntaje académico. Se utilizó un nivel de significancia de 0.05, criterio estándar para determinar si una variable ejerce un impacto estadísticamente significativo.

Después de obtener la tabla ANOVA, se aplicaron pruebas de comparaciones múltiples Tukey HSD para cada una de las variables categóricas del estudio. Estas pruebas tuvieron como objetivo identificar entre qué grupos específicos existían diferencias significativas en el puntaje de examen. De este modo, fue posible distinguir, por ejemplo, si los estudiantes con “alta motivación” rendían significativamente mejor que aquellos con “motivación baja”.

Durante el análisis también se generaron gráficas de intervalos de confianza para visualizar las diferencias entre categorías y detectar patrones relevantes. Asimismo, se revisaron las distribuciones, promedios y comportamiento general de las variables numéricas con el fin de identificar posibles valores atípicos o inconsistencias que pudieran afectar el análisis.

Las herramientas empleadas para realizar todo el procedimiento fueron las librerías `pandas` (gestión y limpieza de datos), `statsmodels` (ANOVA, OLS y Tukey), y `matplotlib`

(visualización gráfica). Estas herramientas permitieron realizar un análisis robusto, replicable y organizado, asegurando la validez de los resultados obtenidos.



```
In [5]: import pandas as pd
url = "https://raw.githubusercontent.com/luissfergarcia/Estadistica2025B/refs/
df = pd.read_csv(url)
df = df.dropna()
df
#LUIS FERNANDO GARCIA MEDINA
```

```
Out[5]:
```

	horas_estudio	asistencia	participacion_parental	acceso_a_recursos	actividad
0	23	84	bajo	elevado	
1	19	64	bajo	medio	
2	24	98	medio	bajo	
3	29	89	bajo	medio	
4	19	92	medio	bajo	
...
6602	25	69	elevado	medio	
6603	23	76	elevado	medio	
6604	20	90	medio	bajo	
6605	10	86	elevado	elevado	
6606	15	67	medio	bajo	

6378 rows × 20 columns

Tabla ANOVA

```
In [6]: from statsmodels.formula.api import ols
import statsmodels.api as sm

# Nivel de significancia = 0.05
# H0: Todas las medias son iguales
# H1: Al menos una media es distinta

# Realizar el modelo ANOVA de dos vías
modelo_lineal = ols(f'puntaje_examen ~ horas_estudio + asistencia + C(participacion_parental)')
tabla_anova = sm.stats.anova_lm(modelo_lineal, typ=2)
tabla_anova
#LUIS FERNANDO GARCIA MEDINA
```


Out[6]:

	sum_sq	df	F	PR(>F)
C(participacion_parental)	2934.495355	2.0	333.048013	2.839802e-13
C(acceso_a_recursos)	2614.915722	2.0	296.777599	6.082330e-12
C(actividades_extracurriculares)	474.735679	1.0	107.759431	4.809221e-2
C(nivel_motivacion)	869.647999	2.0	98.699948	6.136556e-4
C(acceso_internet)	374.929769	1.0	85.104660	3.779018e-2
C(ingreso_familiar)	1053.895937	2.0	119.611008	1.019015e-5
C(calidad_profesorado)	524.860600	2.0	59.568600	2.340878e-2
C(tipo_escuela)	1.940406	1.0	0.440449	5.069297e-0
C(influencia_companieros)	967.444992	2.0	109.799333	1.320573e-4
C(discapacidad_aprendizaje)	449.456984	1.0	102.021464	8.303936e-2
C(nivel_estudio_padres)	949.107969	2.0	107.718188	9.877729e-4
C(distancia)	681.248315	2.0	77.317688	6.663308e-3
C(genero)	3.291714	1.0	0.747181	3.874030e-0
horas_estudio	19675.585241	1.0	4466.127064	0.000000e+0
asistencia	33347.323947	1.0	7569.451388	0.000000e+0
horas_suenio	0.336359	1.0	0.076349	7.823158e-0
calificaciones_previas	3191.145017	1.0	724.352488	3.516636e-15
sesiones_tutoria	2395.146284	1.0	543.670112	1.894735e-11
actividad_fisica	244.181730	1.0	55.426389	1.098955e-1
Residual	27975.013805	6350.0	NaN	Na

Participación parental

```
In [7]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['participacion_parental'],
                           alpha=nivel_de_significancia)

# Mostrar los resultados
print(tukey)

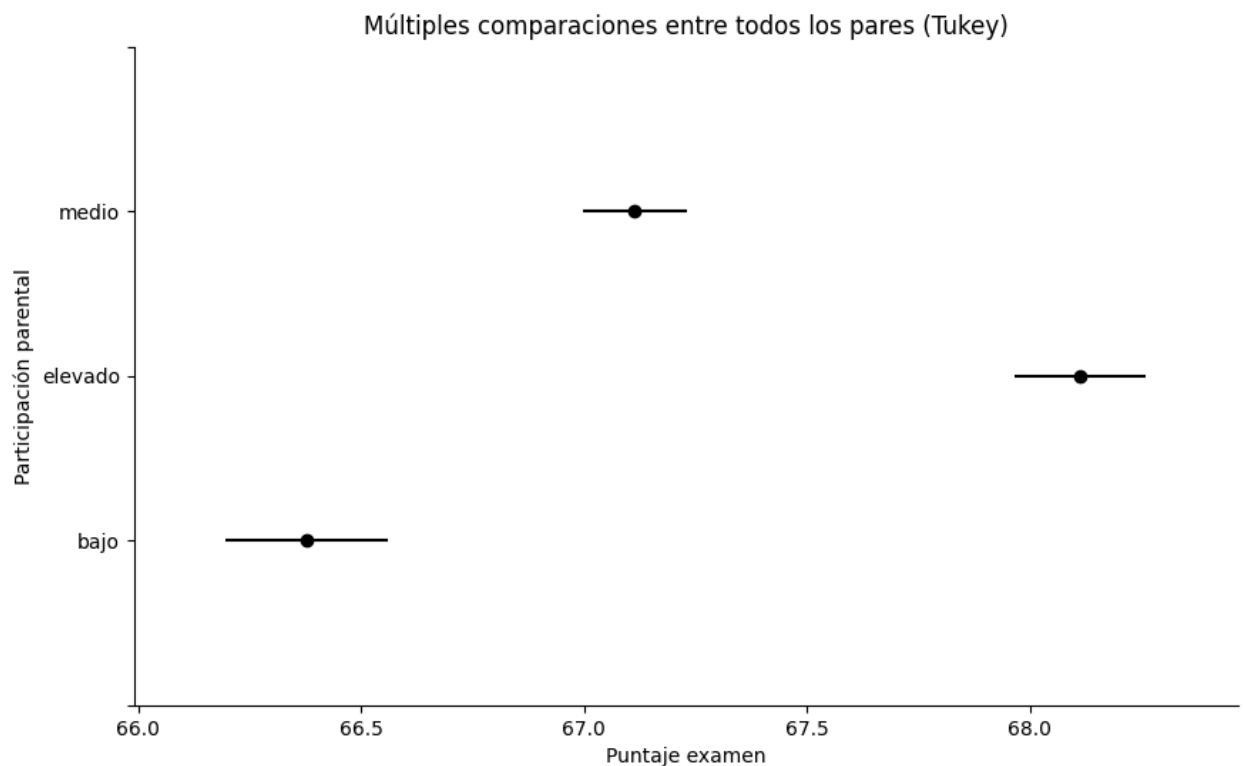
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Participación parental", xlabel="Puntaje examen")
```

```
plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False) # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
bajo elevado 1.7334 0.0 1.4042 2.0626 True
bajo medio 0.7344 0.0 0.4363 1.0326 True
elevado medio -0.999 0.0 -1.2636 -0.7344 True
-----
```

Out[7]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Acceso a recursos

```
In [8]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['acceso_a_recu

# Mostrar los resultados
print(tukey)
```

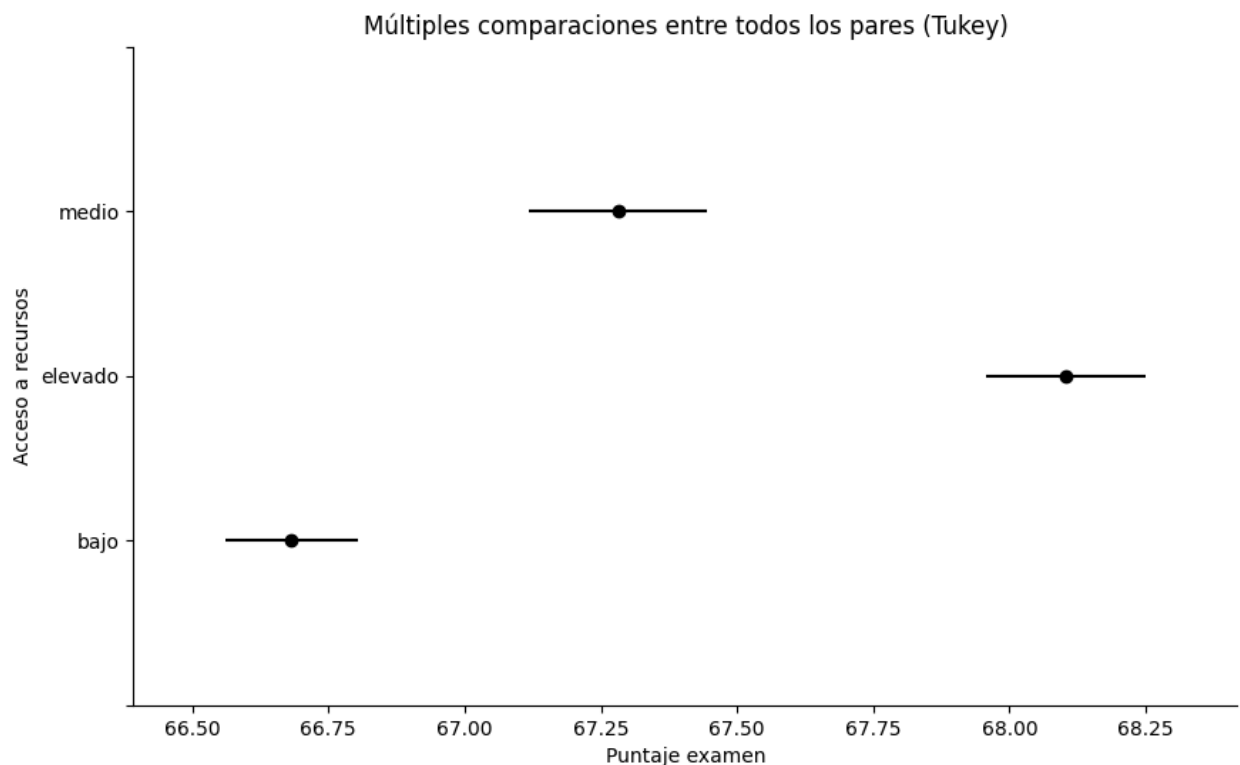
```
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Acceso a recursos", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
bajo elevado 1.4212 0.0 1.1539 1.6885 True
bajo medio 0.6001 0.0 0.3157 0.8845 True
elevado medio -0.8211 0.0 -1.1309 -0.5113 True
-----
```

Out[8]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Actividades extracurriculares

```
In [9]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['actividades_e
```

```

# Mostrar los resultados
print(tukey)

# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Actividades extracurriculares", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA

```

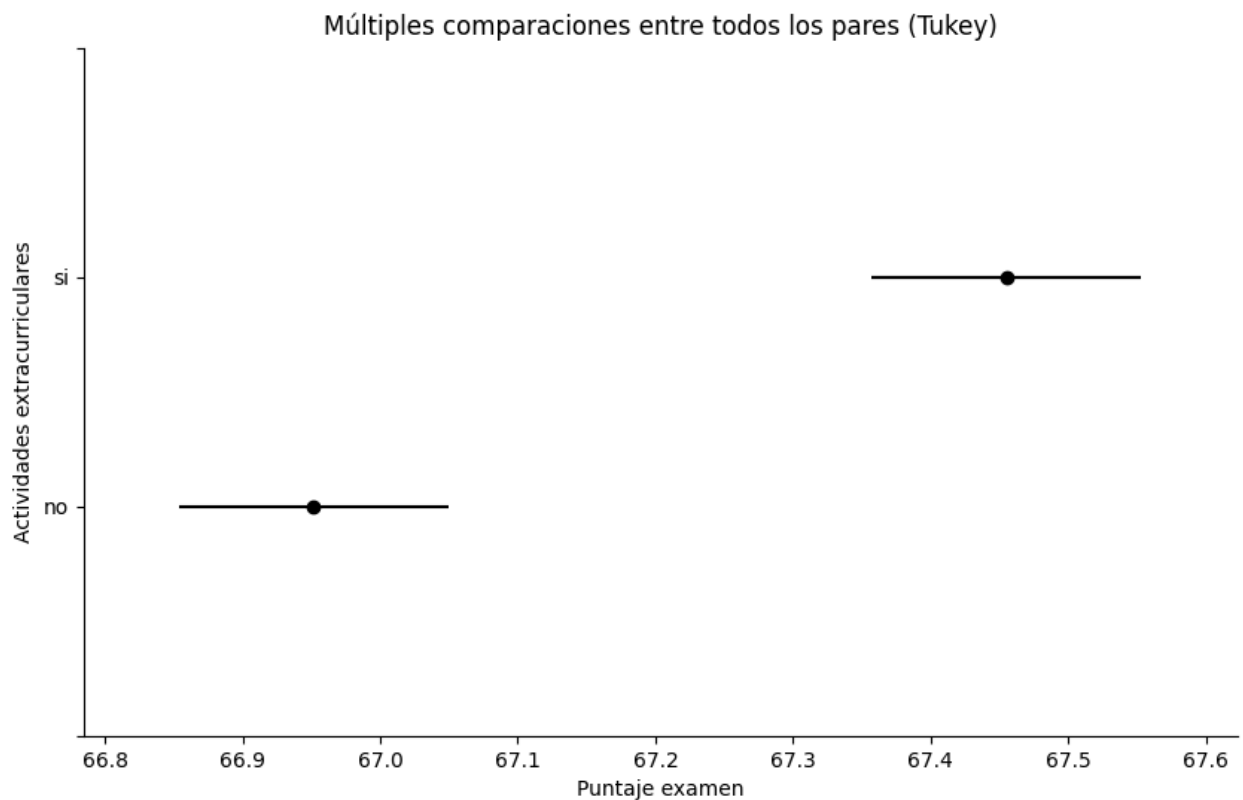
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```

=====
group1 group2 meandiff p-adj lower upper reject
-----
no      si     0.5032   0.0 0.3077 0.6987   True
-----

```

Out[9]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Nivel de motivación

```

In [10]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

```

```
# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['nivel_motivacion'])

# Mostrar los resultados
print(tukey)

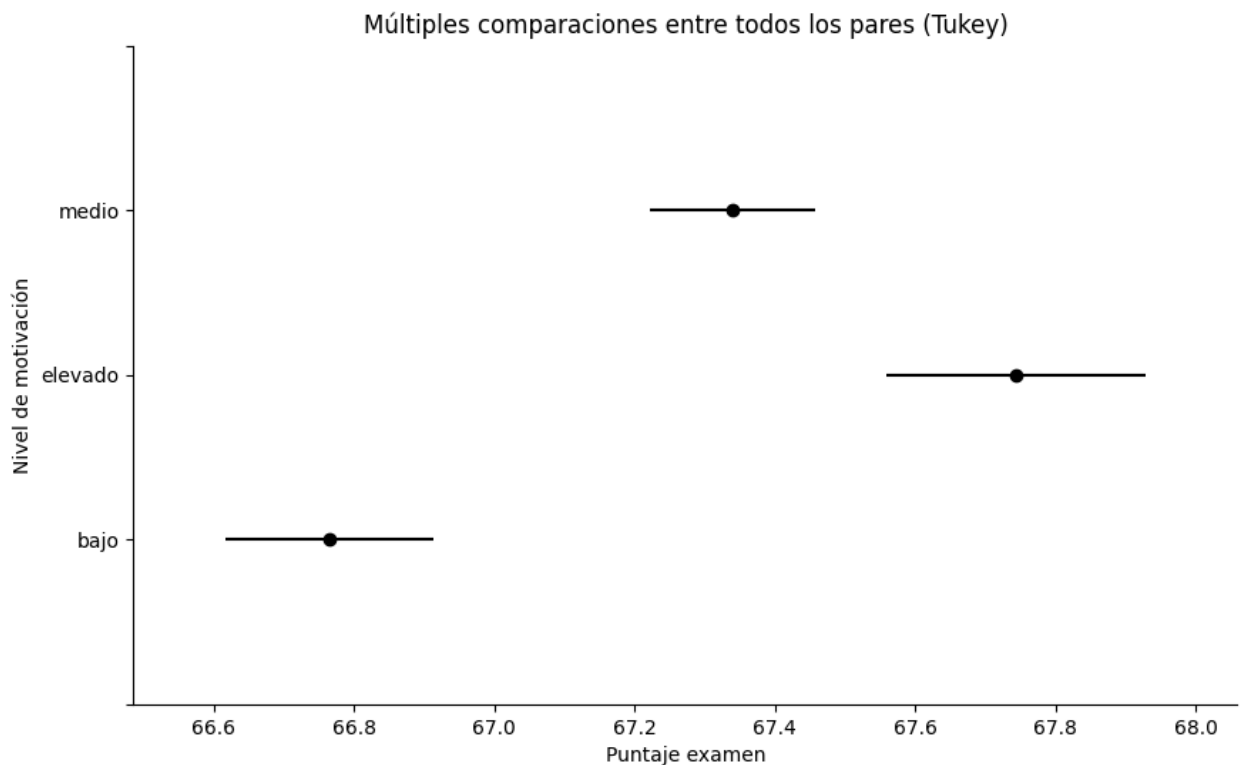
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Nivel de motivación", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
bajo	elevado	0.9794	0.0	0.6474	1.3115	True
bajo	medio	0.5744	0.0	0.3086	0.8402	True
elevado	medio	-0.405	0.0048	-0.7071	-0.103	True

Out[10]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Acceso a internet

```
In [11]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt
```

```

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['acceso_internet'])

# Mostrar los resultados
print(tukey)

# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Acceso a internet", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA

```

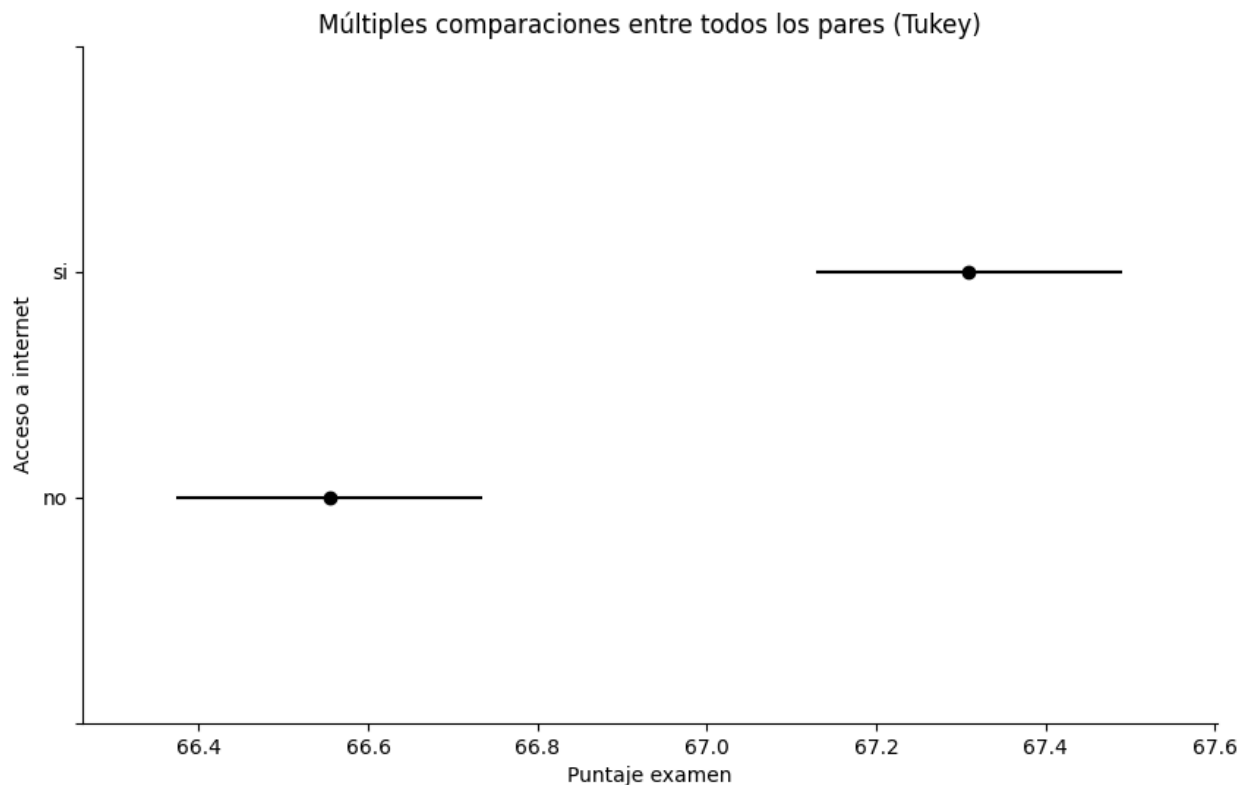
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```

=====
group1 group2 meandiff p-adj lower upper reject
-----
no      si     0.7549   0.0 0.3929 1.1169   True
-----

```

Out[11]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Ingreso familiar

```
In [12]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['ingreso_famil

# Mostrar los resultados
print(tukey)

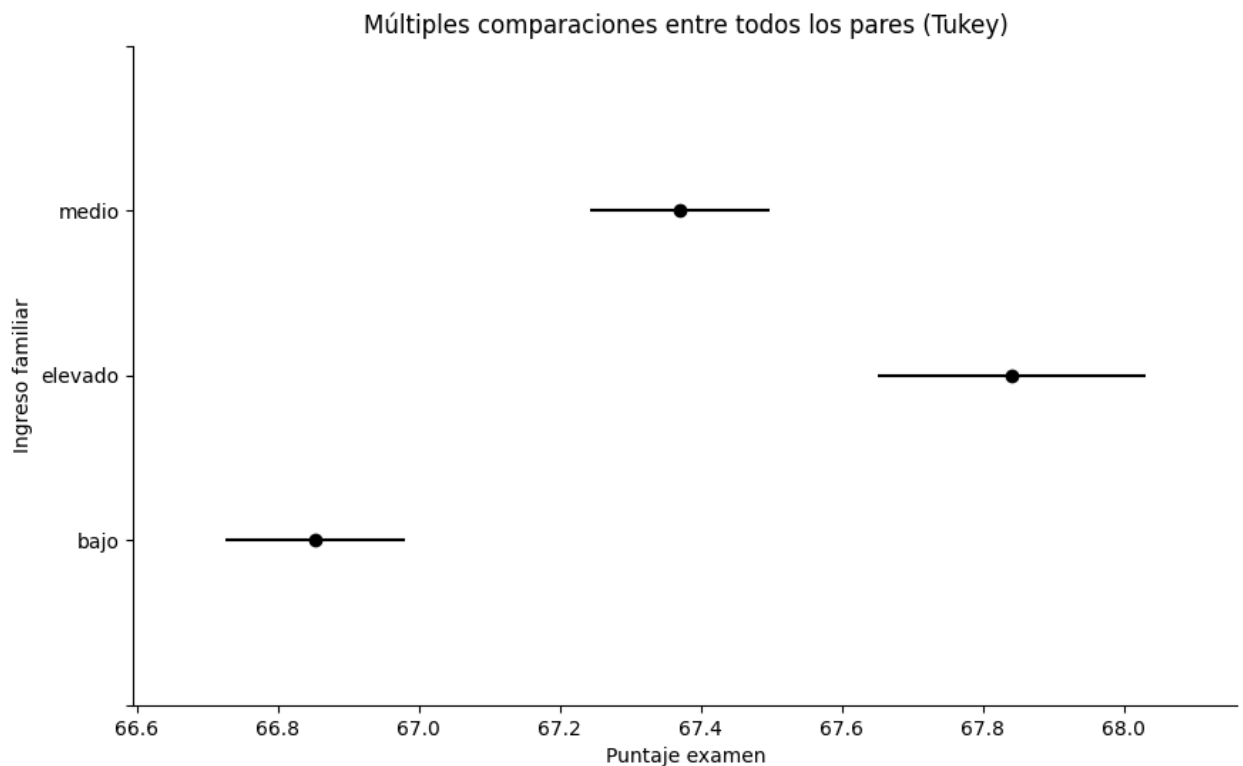
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Ingreso familiar", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
bajo elevado 0.9882 0.0 0.6717 1.3048 True
bajo medio 0.5178 0.0 0.2631 0.7725 True
elevado medio -0.4705 0.0015 -0.7873 -0.1536 True
-----
```

```
Out[12]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')
```



Calidad profesorado

```
In [13]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['calidad_profe

# Mostrar los resultados
print(tukey)

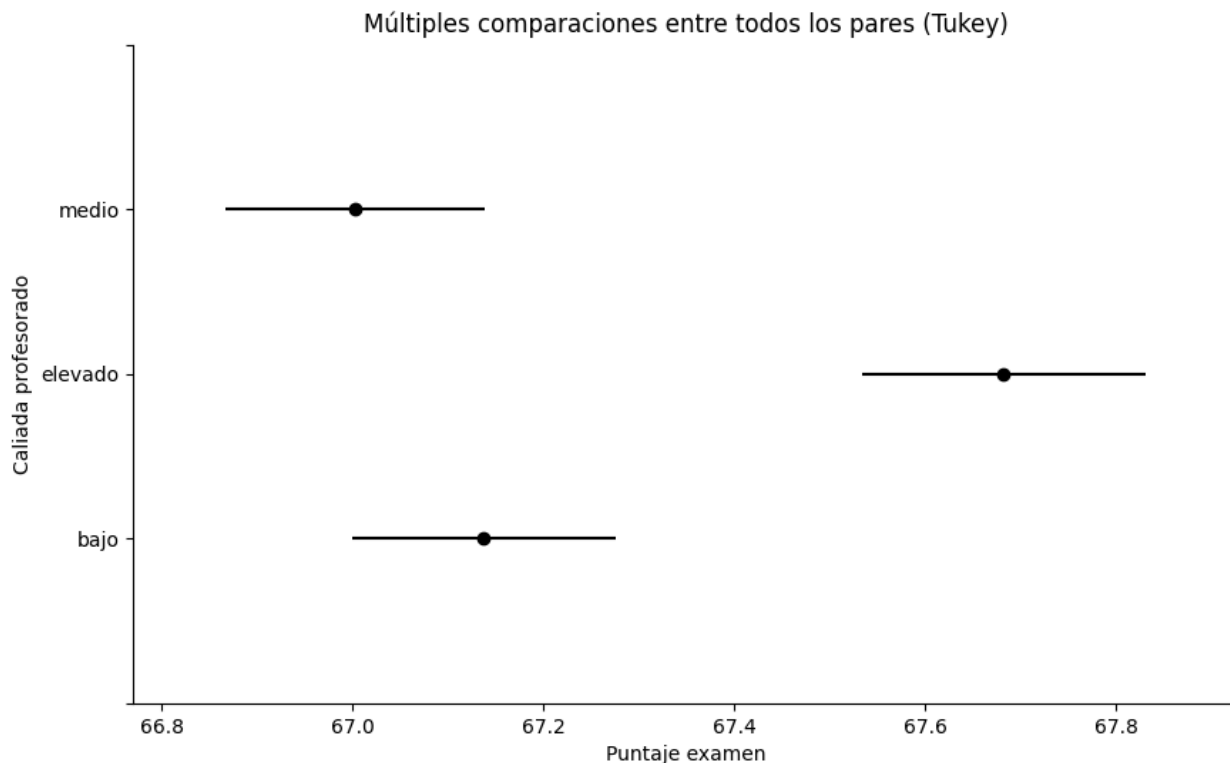
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Caliada profesorado", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```


Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
bajo	elevado	0.5447	0.0	0.2585	0.8309	True
bajo	medio	-0.1345	0.4821	-0.4082	0.1392	False
elevado	medio	-0.6792	0.0	-0.9638	-0.3947	True

Out[13]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Tipo de escuela

```
In [14]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['tipo_escuela'])

# Mostrar los resultados
print(tukey)

# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Tipo de escuela", xlabel="Puntaje examen")

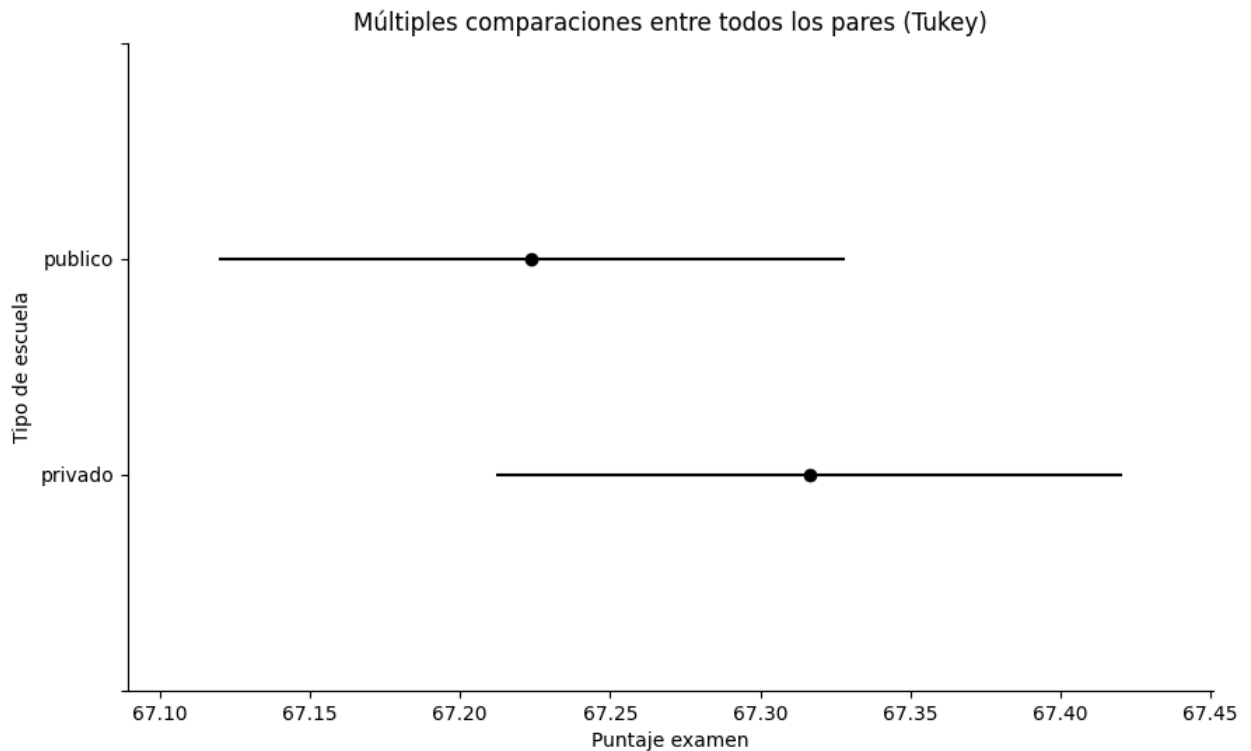
plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
```

```
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
privado publico -0.0924 0.3855 -0.3011 0.1163 False
-----
```

Out[14]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Influencia compañeros

```
In [15]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['influencia_co

# Mostrar los resultados
print(tukey)

# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Influencia compañeros", xlabel="Puntaje examen

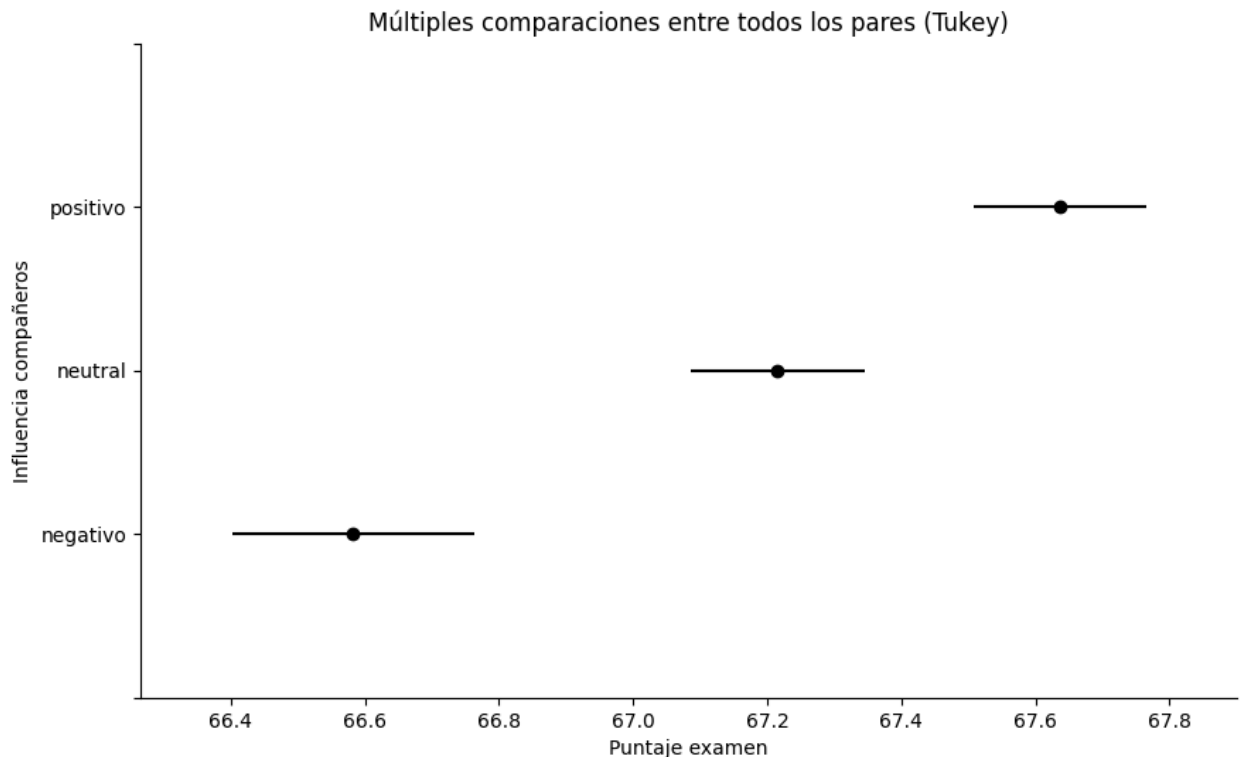
plt.gca().spines['right'].set_visible(False) # derecha
```

```
plt.gca().spines['top'].set_visible(False) # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1  group2  meandiff p-adj  lower  upper  reject
-----
negativo neutral  0.6329   0.0  0.3229  0.9429   True
negativo positivo  1.0538   0.0  0.745  1.3626   True
neutral positivo  0.4209  0.0004  0.1638  0.6779   True
-----
```

Out[15]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Discapacidad de aprendizaje

```
In [16]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['discapacidad_

# Mostrar los resultados
print(tukey)

# Gráfico de las diferencias entre grupos
```

```

tukey.plot_simultaneous(ylabel="Discapacidad de aprendizaje", xlabel="Puntaje

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA

```

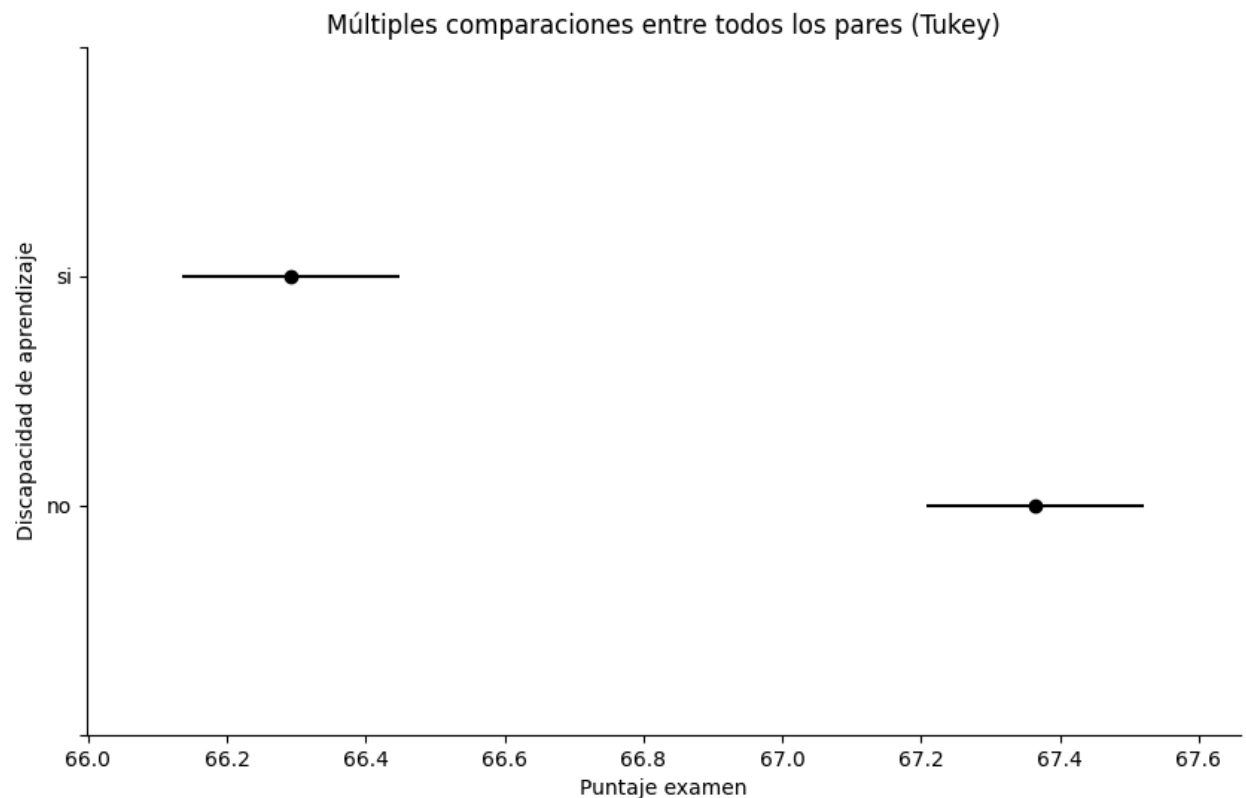
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```

=====
group1 group2 meandiff p-adj  lower  upper  reject
-----
no      si    -1.0725   0.0   -1.3852 -0.7598   True
-----

```

Out[16]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Nivel de estudios de los padres

```

In [17]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['nivel_estudio

# Mostrar los resultados
print(tukey)

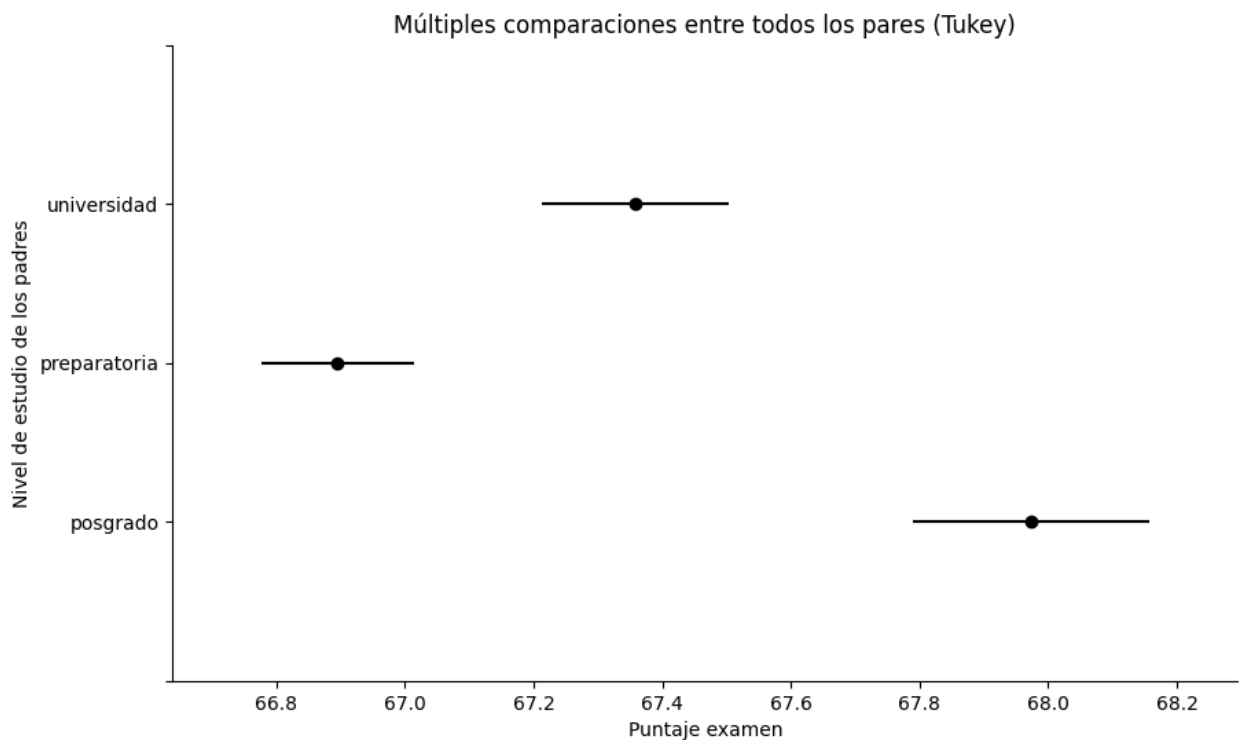
```

```
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Nivel de estudio de los padres", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
group1      group2      meandiff p-adj    lower    upper    reject
-----
posgrado preparatoria -1.0778    0.0 -1.3801 -0.7754    True
posgrado universidad  -0.6142    0.0 -0.9429 -0.2856    True
preparatoria universidad  0.4635 0.0001  0.2002  0.7268    True
-----
```

Out[17]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Distancia

```
In [18]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['distancia'],
```

```
# Mostrar los resultados
print(tukey)

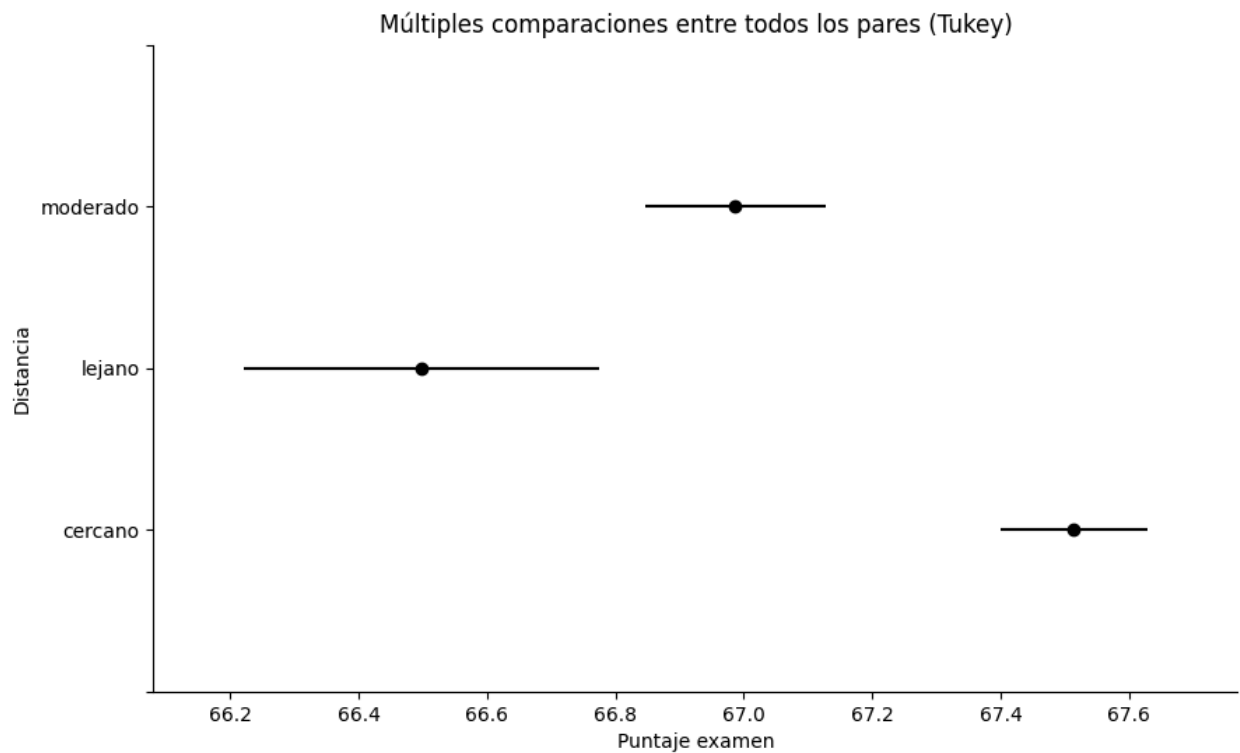
# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Distancia", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

```
=====
group1 group2 meandiff p-adj lower upper reject
-----
cercano lejano -1.0154 0.0 -1.407 -0.6237 True
cercano moderado -0.5272 0.0 -0.7822 -0.2722 True
lejano moderado 0.4882 0.017 0.0705 0.9059 True
-----
```

Out[18]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Genero

```
In [19]: from statsmodels.stats.multicomp import pairwise_tukeyhsd
import matplotlib.pyplot as plt

nivel_de_significancia = 0.05
```

```

# Prueba de Tukey
tukey = pairwise_tukeyhsd(endog=df['puntaje_examen'], groups=df['genero'], alp

# Mostrar los resultados
print(tukey)

# Gráfico de las diferencias entre grupos
tukey.plot_simultaneous(ylabel="Genero", xlabel="Puntaje examen")

plt.gca().spines['right'].set_visible(False) # derecha
plt.gca().spines['top'].set_visible(False)   # superior
plt.title("Múltiples comparaciones entre todos los pares (Tukey)")
#LUIS FERNANDO GARCIA MEDINA

```

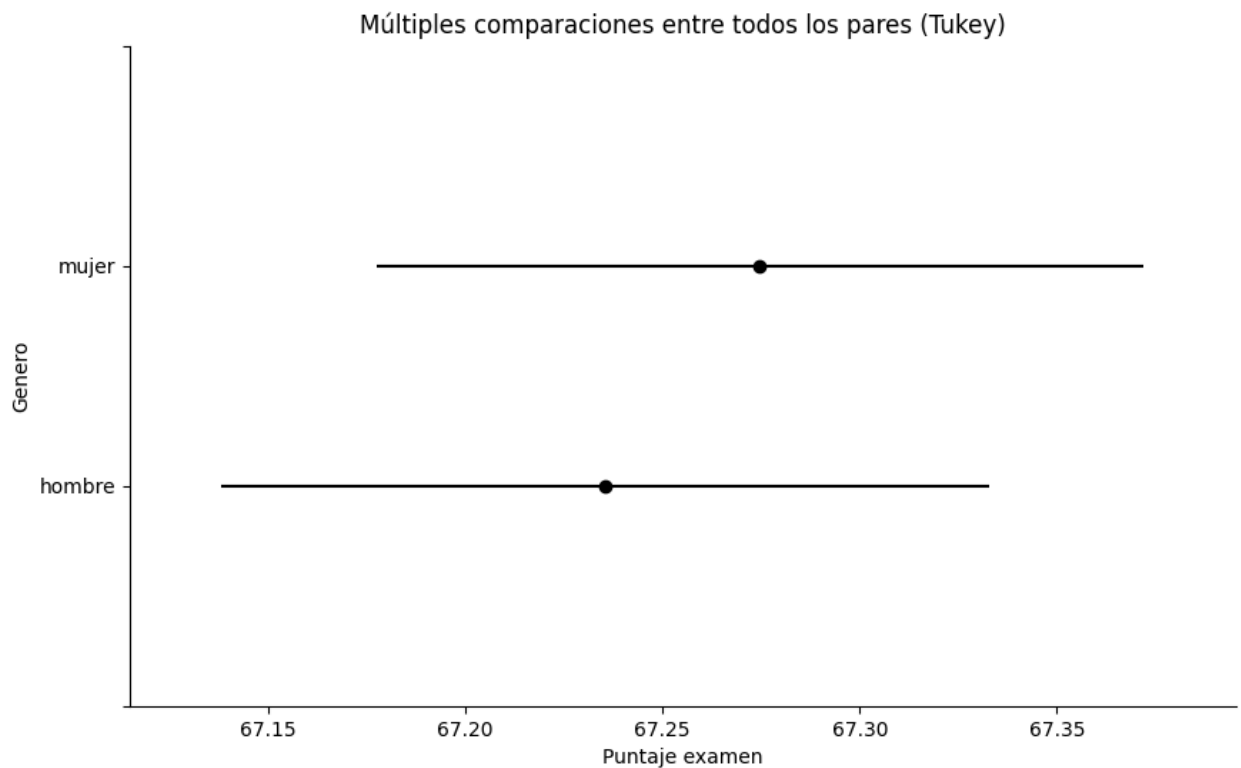
Multiple Comparison of Means - Tukey HSD, FWER=0.05

```

=====
group1 group2 meandiff p-adj  lower  upper  reject
-----
hombre  mujer   0.0391 0.6937 -0.1555 0.2337  False
-----

```

Out[19]: Text(0.5, 1.0, 'Múltiples comparaciones entre todos los pares (Tukey)')



Resultados

A partir del análisis estadístico realizado mediante el modelo lineal, la tabla ANOVA y las pruebas de Tukey, se obtuvieron los siguientes hallazgos:

Las pruebas ANOVA mostraron que varias variables cuantitativas, como horas de estudio, asistencia, calificaciones previas y sesiones de tutoría, presentan valores F extremadamente altos y valores p iguales a 0.0000. Esto indica que su influencia sobre el puntaje del examen es altamente significativa y que los estudiantes con mayor dedicación y preparación académica tienden a obtener mejores resultados.

En cuanto a las variables categóricas, los análisis de Tukey evidenciaron diferencias significativas entre grupos en factores como participación parental, acceso a recursos, nivel de motivación, ingreso familiar, influencia de compañeros y nivel de estudios de los padres. En todos estos casos, los grupos con mayores niveles de apoyo, recursos o motivación presentaron diferencias positivas en el puntaje promedio respecto a los niveles bajos. Esto sugiere un fuerte impacto del contexto familiar y socioeconómico en el desempeño académico.

Por otro lado, las comparaciones correspondientes a tipo de escuela y género mostraron valores p mayores a 0.05, lo que indica que no existe una diferencia significativa entre estudiantes de escuelas públicas vs. privadas, ni entre hombres y mujeres. Este hallazgo sugiere que el rendimiento académico no está definido por características demográficas básicas, sino por factores más vinculados al entorno de estudio y apoyo educativo.

Asimismo, factores como discapacidad de aprendizaje y distancia a la escuela arrojaron diferencias estadísticamente significativas, indicando que los estudiantes con limitaciones de aprendizaje o que viven más lejos presentan desventajas claras en su rendimiento, lo cual coincide con estudios previos sobre barreras educativas.

En conjunto, los resultados muestran que el rendimiento del estudiante depende principalmente de sus hábitos académicos, su motivación, su contexto familiar, la disponibilidad de recursos y la calidad de su entorno escolar. Las gráficas de

Tukey permitieron visualizar claramente estas diferencias en medias, reforzando la interpretación estadística obtenida a partir de la tabla ANOVA.

Conclusión

El análisis realizado permite concluir que el rendimiento académico es resultado de la interacción de múltiples factores personales, familiares y escolares, y que su influencia puede identificarse con precisión mediante técnicas estadísticas como ANOVA y las pruebas de Tukey. Los resultados evidencian que variables como las horas de estudio, la asistencia, las calificaciones previas, las sesiones de tutoría, el nivel de motivación, el acceso a recursos y la participación parental ejercen una influencia significativa y directa sobre el puntaje obtenido en el examen. Esto confirma que los hábitos de estudio, el compromiso académico y el apoyo del entorno constituyen elementos clave para favorecer el aprendizaje.

Asimismo, se observó que factores estructurales como el ingreso familiar, el nivel educativo de los padres, el acceso a internet y la influencia positiva de los compañeros también se relacionan con un mejor desempeño, lo que resalta la importancia de las condiciones socioeconómicas y del entorno escolar en el desarrollo académico del estudiante. En contraste, variables como el género y el tipo de escuela no mostraron diferencias significativas, lo que sugiere que los resultados académicos no dependen de características demográficas, sino de oportunidades, recursos y hábitos de estudio.

En conjunto, los hallazgos indican que fortalecer el rendimiento académico requiere estrategias integrales que incluyan apoyo familiar, acceso adecuado a recursos educativos, ambientes escolares positivos, así como el fomento de hábitos disciplinados de estudio. Este análisis proporciona una base sólida para futuros estudios y para el diseño de intervenciones que promuevan mejores condiciones de aprendizaje y mayor equidad educativa.