TRABAJO APLICADO ESTADÍSTICA INFERENCIAL

Santiago Cubillos Cruz

Universidad del Rosario

Facultad de Administración de empresas

Inferencia Estadística

Docente: Adriana Paola pachón Gutierrez

TRABAJO APLICADO ESTADÍSTICA INFERENCIAL

Introducción

Esta muestra es resultado de la iniciativa de estudiantes de psicología de la universidad de antioquia para conocer conceptos básicos sobre los estudiantes que entran a la universidad basándose en la prueba de estado presentada antes de entrar a la universidad y los resultados del primer semestre . Se realiza una selección de 11 variables de las 22 variables en total recogidas.Las variables escogidas se dan de manera natural en el entendimiento ,y relación de los valores de un examen de estado y la dedicación durante el semestre .Durante el el presente trabajo expondremos las principales características del conjunto de datos y pondremos en juicio algunas de las preguntas presentadas por el equipo psicológico que tomó la muestra.

Planteamiento del Problema

¿Existe una relación intrínseca entre las notas obtenidas por los estudiantes en las pruebas de estado y el promedio del primer semestre?

Objetivo General

-Identificar las variables que intervienen activamente en la obtención del puntaje promedio en primer semestre

Objetivo Especifico

-Analizar las variables y sus relaciones para poder saber si existe una correlación entre las notas , otras variables y el promedio como variable respuesta .

Definición de Variables

Variables Cualitativas

Escala Nominal:

- Genero :F-femenino ,M-masculino
- ¿Realiza actividades extracurriculares? :Si , NO

Escala Ordinal:

- Estrato social: "1","2","3","4","5","6"
- ¿Repasa los temas vistos en clase? : Nunca, Pocas Veces,Algunas Veces,Casi siempre , Siempre

Variables Cuantitativas

Discreta :

- Edad (en años) : mínimo 15 años y máximo 25 años [razón]
- Tiempo de estudio semanal (horas) : mínimo 30 horas y máximo 55 horas [razón]
- ¿A cuántas tutorias asistió durante el semestre? : mínimo 1 y máximo 10 [razón]

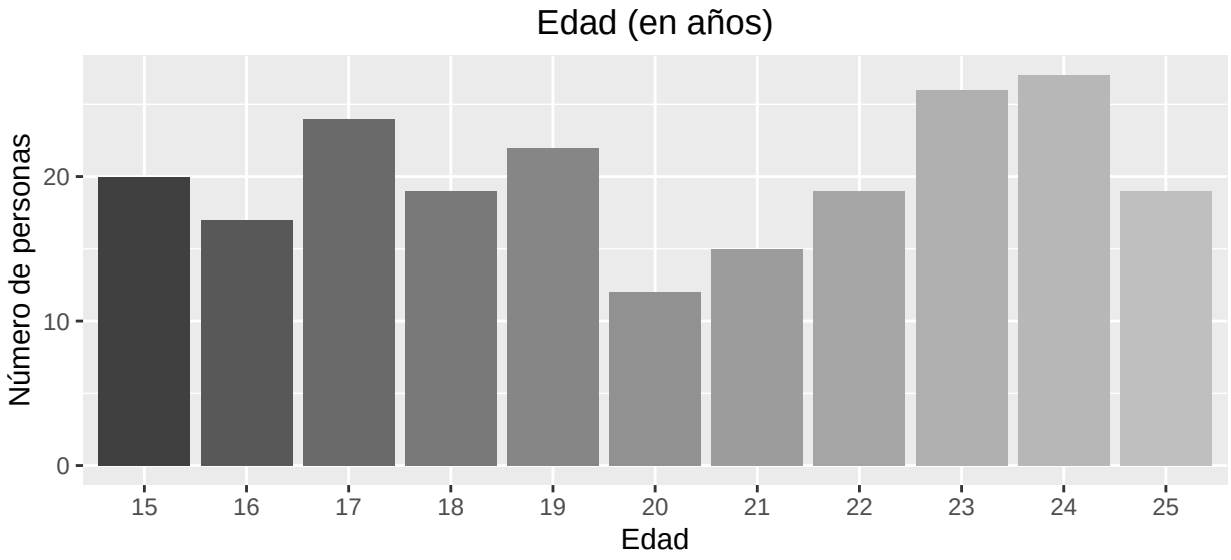
Continua :

- PUNTAJE MATEMÁTICAS : entre 32 y 55 [razón]
- PUNTAJE C. NATURALES : entre 47 y 73 [razón]
- PUNTAJE INGLES : entre 37 y 63 [razón]
- Promedio Primer Semestre : entre 3.0 y 4.9 [razón]

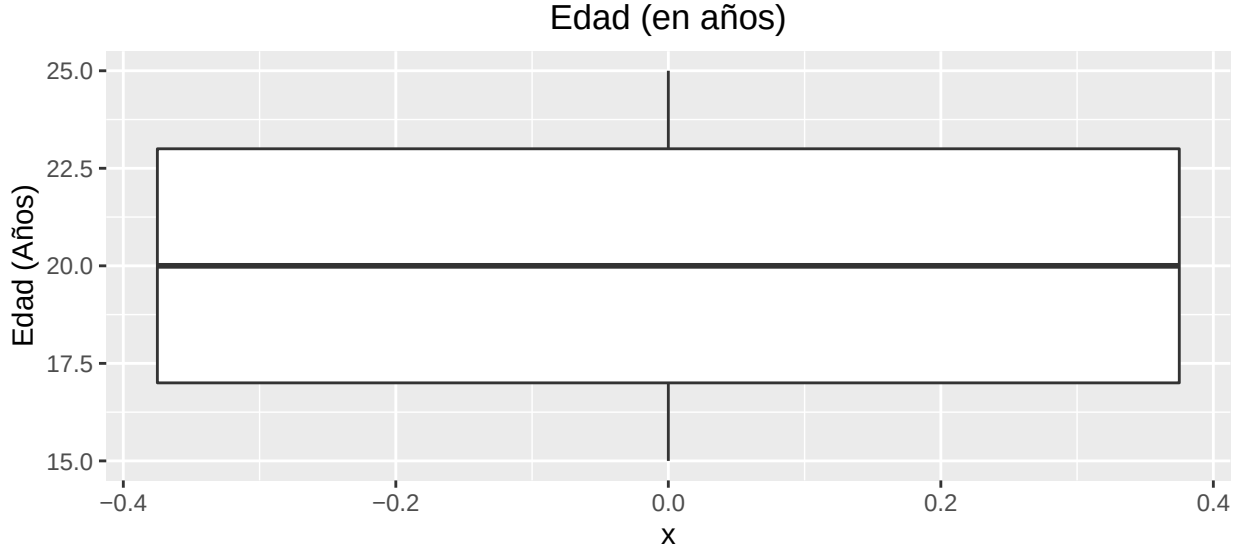
Análisis de datos Cuantitativos

Se establecerán las principales características de las variables cuantitativas encontradas en la investigación .

Edad (en años)



Se observa uniformidad en la cantidad de estudiantes por cada edad además de que al ser universitarios están entre los 15 y 25 años.



la distribución de los datos parece no tener datos atípicos .

Tabla 1 : Descripción numérica para la variable Edad.

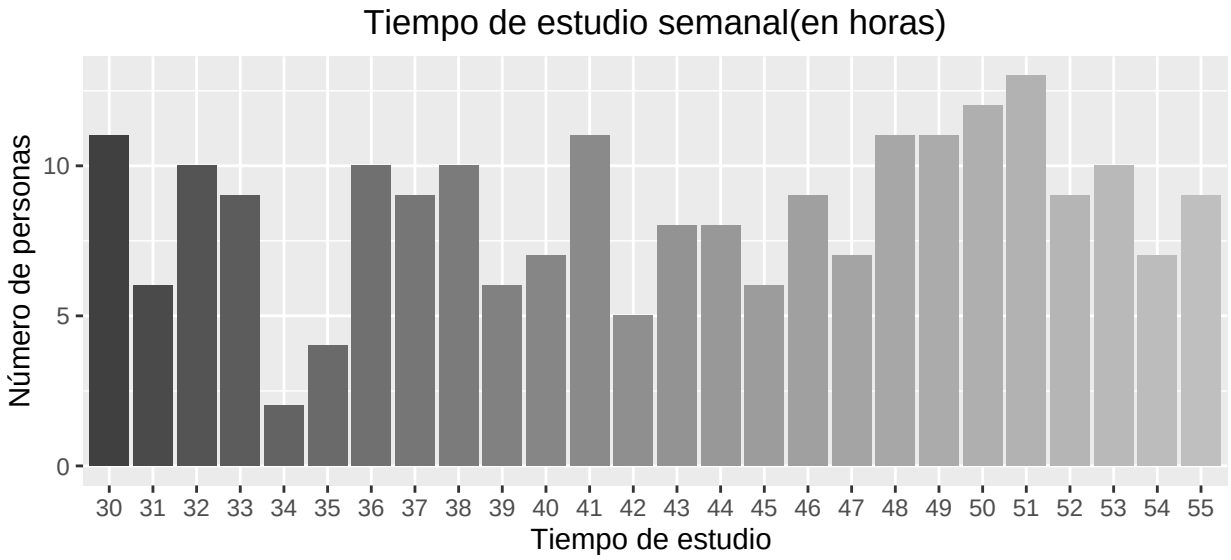
	Datos
Promedio	20.15
Mediana	20.00
Varianza	10.56
SD	3.25
CV	16.12
Q1	17.00
Q2	20.00
Q3	23.00
Minimo	15.00
Maximo	25.00

CV es menor al 20% la media será una buena medida descriptiva de la centralidad edad de los estudiantes universitarios.

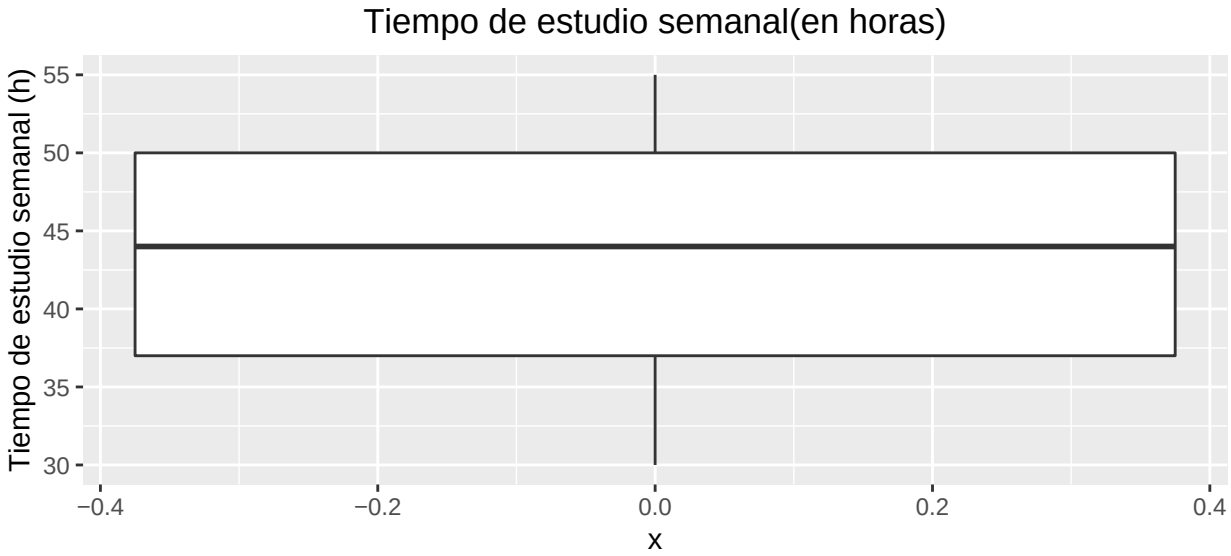
Tabla 2 : Tabla de frecuencias para la edad

Lower	Upper	Main	Frequency	Percentage	CF	CPF
15	16	15.5	37	16.8	37	16.8
16	17	16.5	24	10.9	61	27.7
17	18	17.5	19	8.6	80	36.4
18	19	18.5	22	10.0	102	46.4
19	20	19.5	12	5.5	114	51.8
20	21	20.5	15	6.8	129	58.6
21	22	21.5	19	8.6	148	67.3
22	23	22.5	26	11.8	174	79.1
23	24	23.5	27	12.3	201	91.4
24	25	24.5	19	8.6	220	100.0

Tiempo de estudio semanal (en horas)



Se observa uniformidad en la cantidad de tiempo dedicado al estudio por los estudiantes .



la distribución de los datos parece no tener datos atípicos .

Tabla 3 : Descripción numérica para la variable Tiempo de estudio semanal.

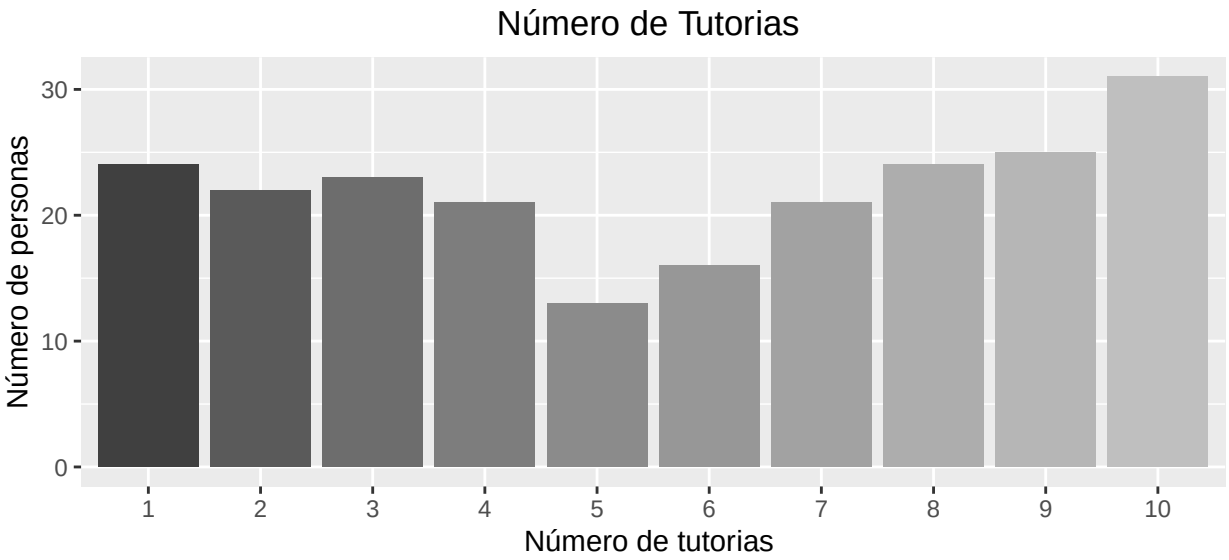
	Datos
Promedio	43.18
Mediana	44.00
Varianza	58.01
SD	7.62
CV	17.64
Q1	37.00
Q2	44.00
Q3	50.00
Minimo	30.00
Maximo	55.00

CV es menor al 20% la media será una buena medida descriptiva de la centralidad del tiempo de estudio semanal de los estudiantes universitarios.

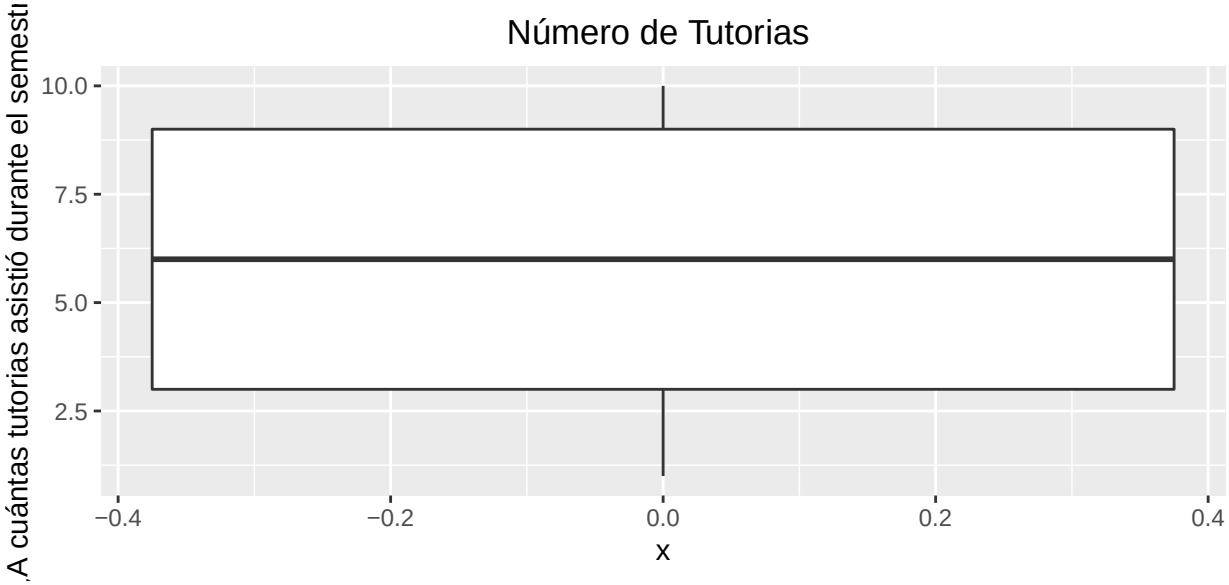
Tabla 4 : Tabla de frecuencias para el tiempo de estudio semanal (en horas)

Lower	Upper	Main	Frequency	Percentage	CF	CPF
30	35	32.5	42	19.1	42	19.1
35	40	37.5	42	19.1	84	38.2
40	45	42.5	38	17.3	122	55.5
45	50	47.5	50	22.7	172	78.2
50	55	52.5	48	21.8	220	100.0

Número de tutorias



Se observa uniformidad en la cantidad de tiempo dedicado al estudio por los estudiantes .



la distribución de los datos parece no tener datos atípicos .

Tabla 5 : Descripción numérica para la variable del número de horas de estudio

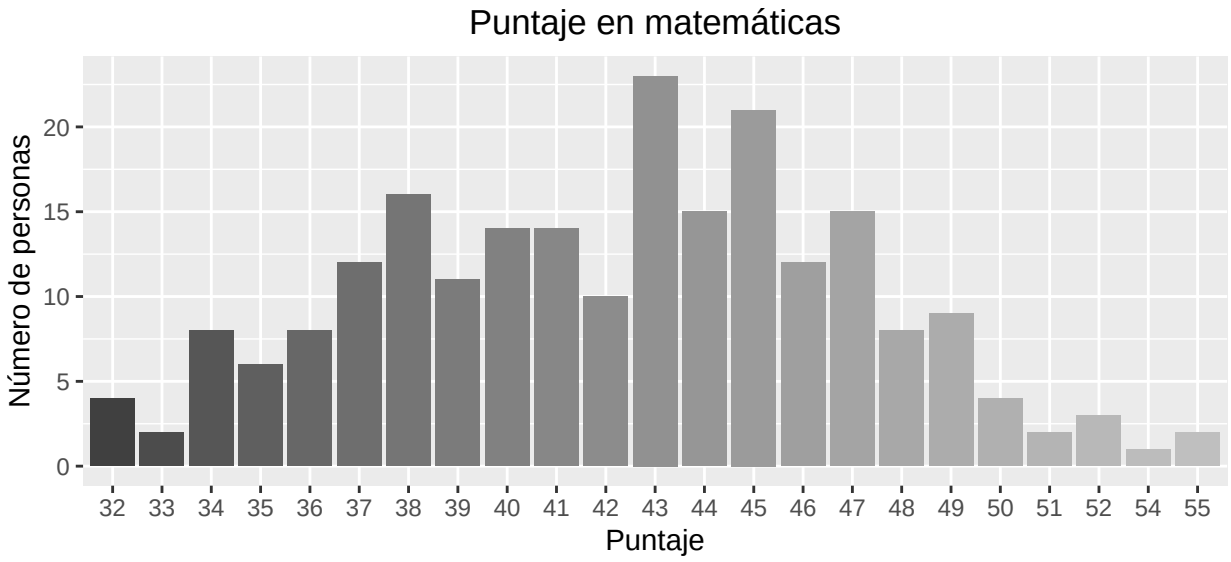
	Datos
Promedio	5.71
Mediana	6.00
Varianza	9.48
SD	3.08
CV	53.92
Q1	3.00
Q2	6.00
Q3	9.00
Minimo	1.00
Maximo	10.00

CV es mayor al 20% la media será tan buena medida descriptiva de la centralidad del número de tutorias que toman los estudiantes universitarios.

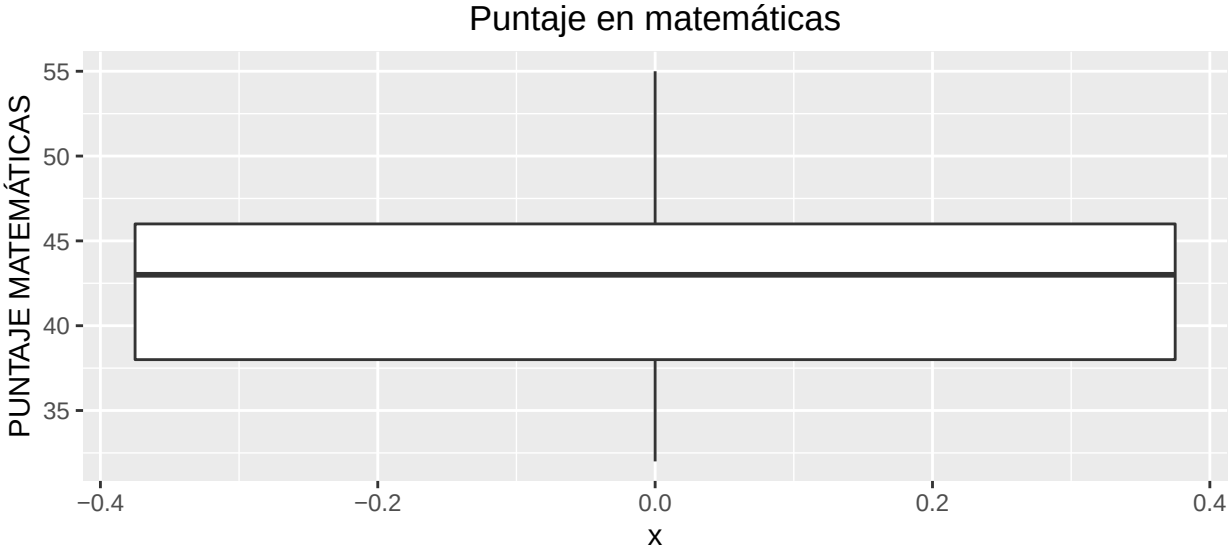
Tabla 6 : Tabla de frecuencias para el número de horas de estudio

Lower	Upper	Main	Frequency	Percentage	CF	CPF
1	2	1.5	46	20.9	46	20.9
2	3	2.5	23	10.5	69	31.4
3	4	3.5	21	9.5	90	40.9
4	5	4.5	13	5.9	103	46.8
5	6	5.5	16	7.3	119	54.1
6	7	6.5	21	9.5	140	63.6
7	8	7.5	24	10.9	164	74.5
8	9	8.5	25	11.4	189	85.9
9	10	9.5	31	14.1	220	100.0

Puntaje en Matemáticas



Es observable que el puntaje tiende a acercarse a la media haciendo las colas menos pesadas .



la distribución de los datos parece no tener datos atípicos .

Tabla 7 : Descripción numérica para el puntaje en matemáticas

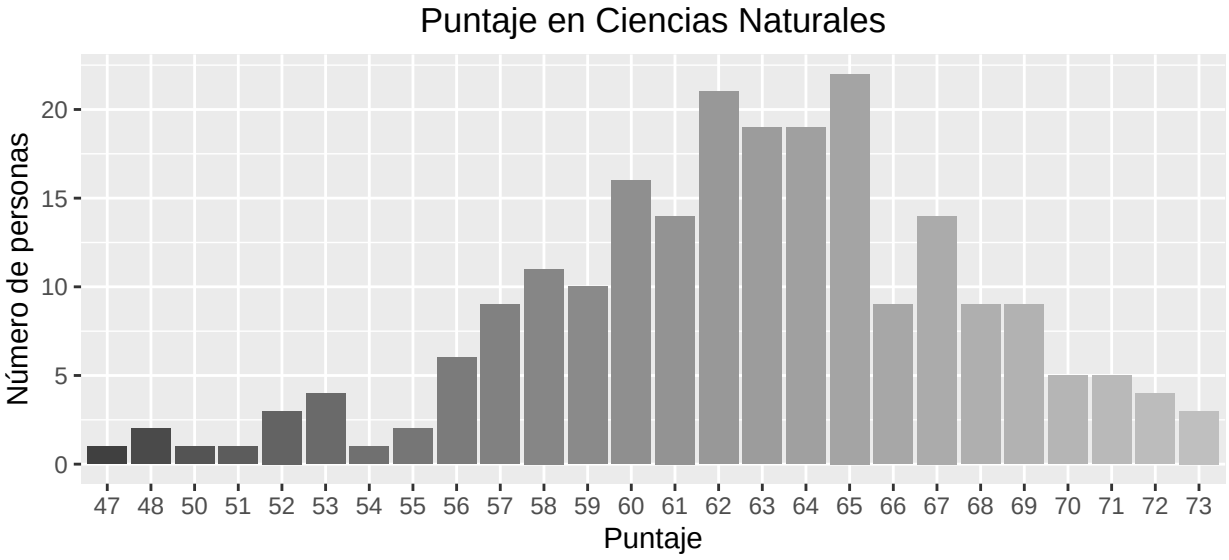
	Datos
Promedio	42.26
Mediana	43.00
Varianza	23.44
SD	4.84
CV	11.46
Q1	38.00
Q2	43.00
Q3	46.00
Minimo	32.00
Maximo	55.00

CV es menor al 20% la media será una buena medida descriptiva de la centralidad del puntaje en matemáticas de los estudiantes universitarios.

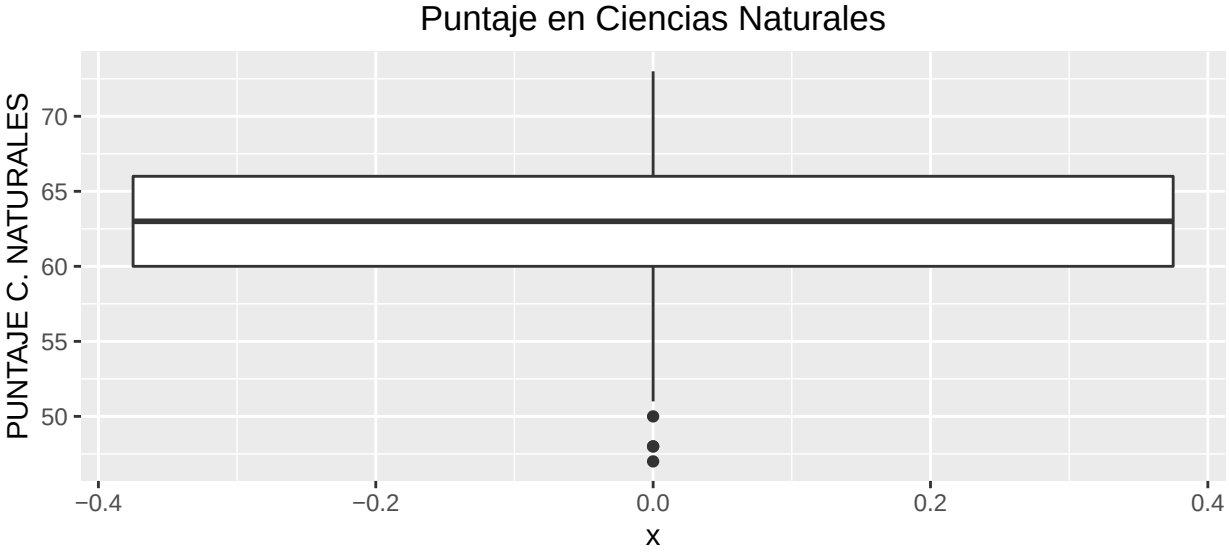
Tabla 8 : Tabla de frecuencias para el puntaje en matemáticas

Lower	Upper	Main	Frequency	Percentage	CF	CPF
30	35	32.5	20	9.1	20	9.1
35	40	37.5	61	27.7	81	36.8
40	45	42.5	83	37.7	164	74.5
45	50	47.5	48	21.8	212	96.4
50	55	52.5	8	3.6	220	100.0

Puntaje en Ciencias Naturales



Es observable que el puntaje tiende a acercarse a la media haciendo las colas menos pesadas .



la distribución parece ser apuntada a la media y posee datos atípicos .

Tabla 9 : Descripción numérica para el puntaje en ciencias naturales

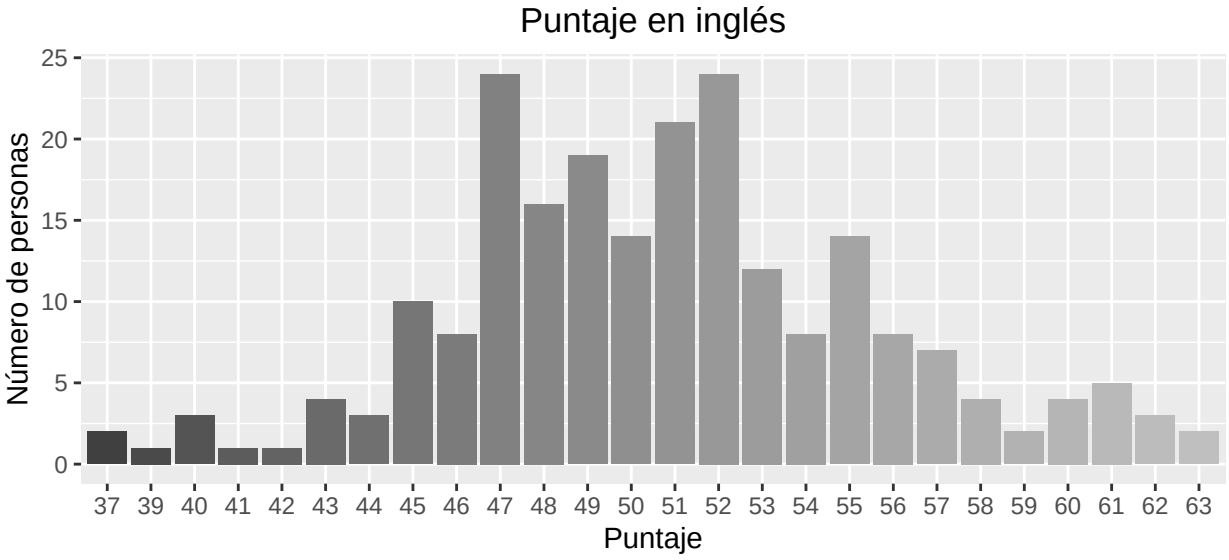
	Datos
Promedio	62.68
Mediana	63.00
Varianza	24.62
SD	4.96
CV	7.92
Q1	60.00
Q2	63.00
Q3	66.00
Minimo	47.00
Maximo	73.00

CV es menor al 20% la media será una buena medida descriptiva de la centralidad del puntaje en ciencias naturales de los estudiantes universitarios cabe notar la poca variación de los datos frente a la media aunque existen valores atípicos.

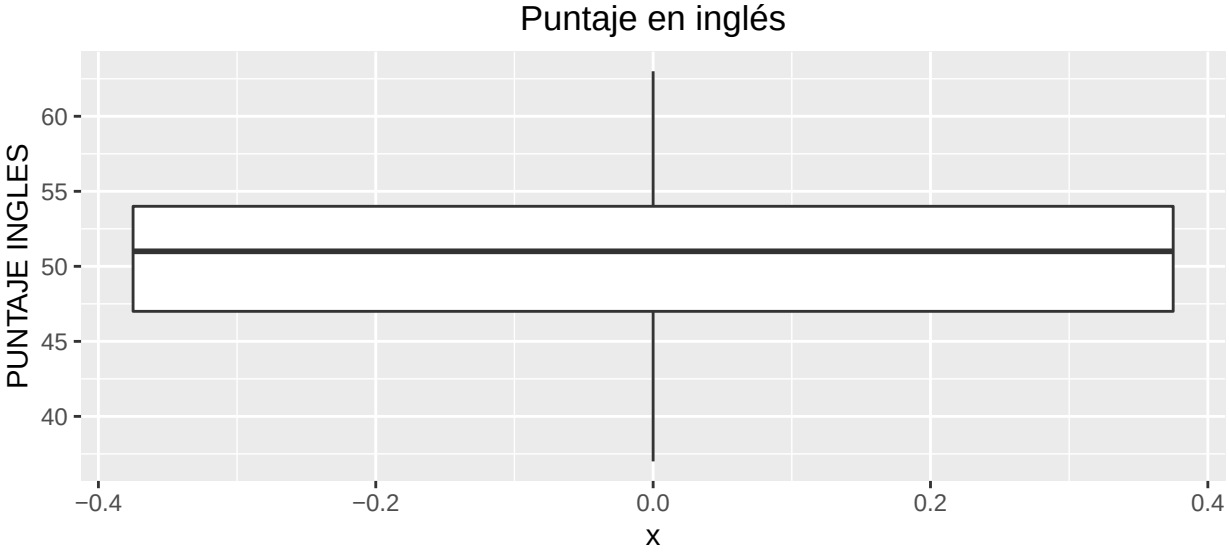
Tabla 10 : Tabla de frecuencias para el puntaje en ciencias naturales

Lower	Upper	Main	Frequency	Percentage	CF	CPF
45	50	47.5	4	1.8	4	1.8
50	55	52.5	11	5.0	15	6.8
55	60	57.5	52	23.6	67	30.5
60	65	62.5	95	43.2	162	73.6
65	70	67.5	46	20.9	208	94.5
70	75	72.5	12	5.5	220	100.0

Puntaje en Inglés



Es observable que e puntaje tiende a acercarse a la media haciendo las colas menos pesadas .



la distribución parece ser apuntada a la media y no posee datos atípicos. .

Tabla 11 : Descripción numérica para el puntaje en inglés

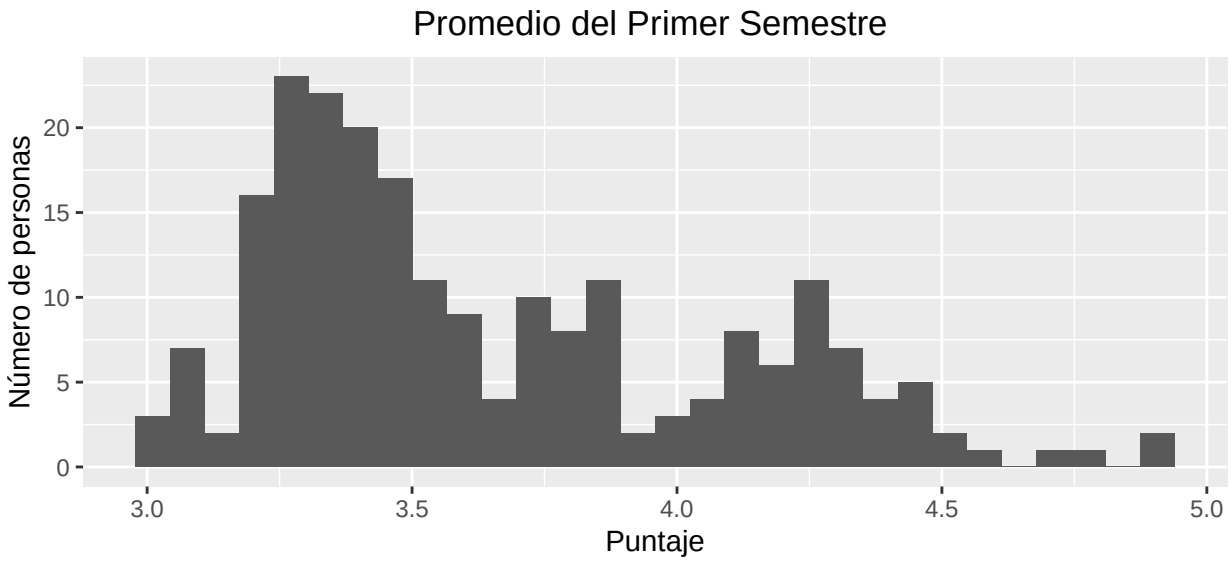
	Datos
Promedio	50.80
Mediana	51.00
Varianza	24.07
SD	4.91
CV	9.66
Q1	47.00
Q2	51.00
Q3	54.00
Minimo	37.00
Maximo	63.00

CV es menor al 20% la media será una buena medida descriptiva del puntaje en inglés de los estudiantes universitarios cabe notar la poca variación de los datos frente a la media .

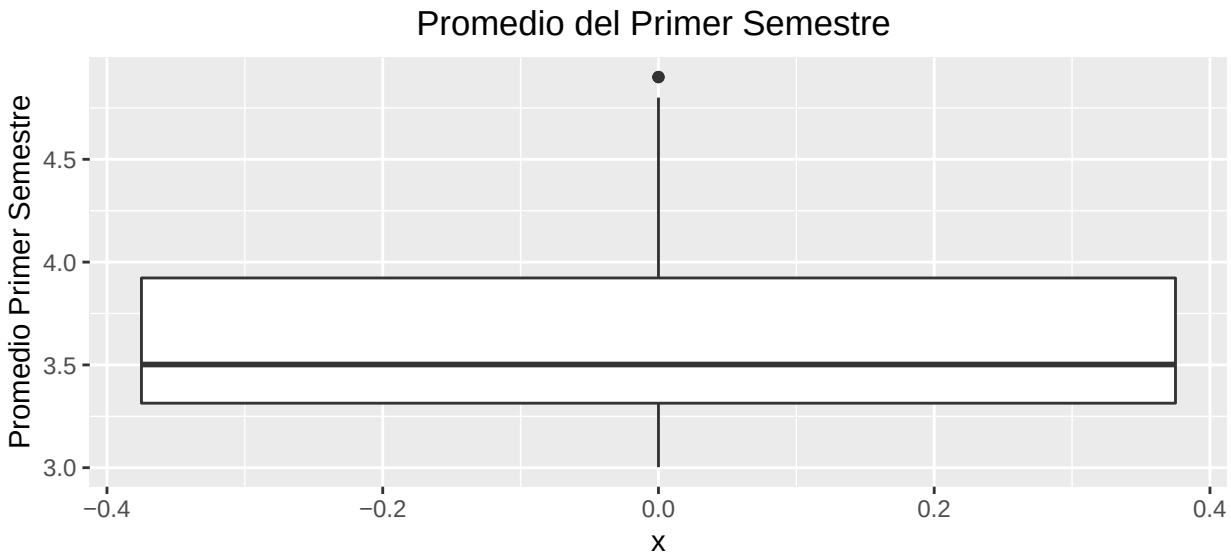
Tabla 12 : Tabla de frecuencias para el puntaje en inglés

Lower	Upper	Main	Frequency	Percentage	CF	CPF
35	40	37.5	6	2.7	6	2.7
40	45	42.5	19	8.6	25	11.4
45	50	47.5	81	36.8	106	48.2
50	55	52.5	79	35.9	185	84.1
55	60	57.5	25	11.4	210	95.5
60	65	62.5	10	4.5	220	100.0

Promedio del Primer Semestre



Es observable que el puntaje tiende a acercarse a la media haciendo las colas menos pesadas .



la distribución parece ser apuntada a dos puntos de acumulación y posee algunos valores atípicos.

Tabla 13 : Descripción numérica para el promedio del primer semestre

	Datos
Promedio	3.66
Mediana	3.50
Varianza	0.19
SD	0.43
CV	11.78
Q1	3.31
Q2	3.50
Q3	3.92
Minimo	3.00
Maximo	4.90

CV es menor al 20% la media será una buena medida descriptiva del promedio del primer semestre de los estudiantes universitarios cabe notar la poca variación de los datos frente a la media .

Tabla 14 : Tabla de frecuencias para el promedio del primer semestre

Lower	Upper	Main	Frequency	Percentage	CF	CPF
3.0	3.5	3.25	110	50.0	110	50.0
3.5	4.0	3.75	55	25.0	165	75.0
4.0	4.5	4.25	50	22.7	215	97.7
4.5	5.0	4.75	5	2.3	220	100.0

Intervalos De Confianza

Intervalos para la media

Edad (en años) : mínimo 15 años y máximo 25 años [razón]

```
## [1] 19.72274 20.58635
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una media de edad entre los 19.7 años y 20.6 años.

Tiempo de estudio semanal (horas) : mínimo 30 horas y máximo 55 horas [razón]

```
## [1] 42.16524 44.18930
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una media para el tiempo de estudio semanal entre 42 y 44 horas cabe aclarar que la distribución del tiempo semanal resulto aproximadamente uniforme .

¿A cuántas tutorias asistió durante el semestre? : mínimo 1 y máximo 10 [razón]

```
## [1] 5.300047 6.118135
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una media para el número de tutorias semanales entre 5 horas y 6 horas de tutorias

PUNTAJE MATEMÁTICAS : entre 32 y 55 [razón]

```
## [1] 41.61572 42.90246
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una nota media entre 41.61 y 42.25

PUNTAJE C. NATURALES : entre 47 y 73 [razón]

```
## [1] 62.01795 63.33660
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una nota media entre 62 y 63.33

PUNTAJE INGLES : entre 37 y 63 [razón]

```
## [1] 50.14811 51.45189
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una nota media entre 50.14 y 51.45

Promedio Primer Semestre : entre 3.0 y 4.9 [razón]

```
## [1] 3.598850 3.713283
## attr(,"conf.level")
## [1] 0.95
```

Interpretación: al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una nota media del promedio general del primer semestre entre 3.59 y 3.71

Intervalo de confianza para la proporción

Se desea obtener un intervalo de confianza para la proporción de los estudiantes que tienen mas de 4 , recordemos que :

$$IC_{1-\alpha}(p) = [\hat{p} \mp z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

entonces el intervalo será :

$$IC_{1-\alpha}(p) = [0.25 \mp 1.96 * \sqrt{\frac{0.25 * 0.75}{220}}] = (0.1927803, 0.3072197)$$

Interpretación:al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán una proporción para las personas que obtiene una nota por encima entre 19.28% y 30.72%

Intervalo de confianza para la varianza

estimamos un intervalo de confianza para la varianza de la nota promedio .

```
## [1] 0.1550600 0.2257113
## attr(,"conf.level")
## [1] 0.95
```

Interpretación:al extraer 100 veces la muestra cada una con 221 individuos , se espera que aproximadamente 95 de estas muestras tendrán varianza entre 0.155 y 0.2257

Pruebas de Bondad de Ajuste (Chi-cuadrado)

Para la prueba de independencia Chi-cuadrado se considerada la hipótesis nula:

$$\begin{cases} H_0: & \text{las variables son independendientes} \\ H_1: & \text{las variables son dependientes} \end{cases}$$

Género-Estrato Social

Se observa la relación entre el género y el estrato social

	1	2	3	4	5	6
F	23	21	17	19	16	15
M	18	20	18	17	21	15

Pearson's Chi-squared test

data: generoEstrato X-squared = 1.4314, df = 5, p-value = 0.9208

la prueba chi-cuadrado nos da un p-valor de 0.9208 , para un nivel de significancia del $\alpha = 0.05$ no hay evidencia estadística para rechazar H_o , es decir el género y el estrato social son independientes

Actividades extracurriculares - Repaso de temas

Se observa la relación entre el realizar actividades extracurriculares y la pregunta "¿repasa usted los temas vistos en clase?"

	NO	SI
Algunas veces	32	30
Casi siempre	23	33
Nunca	18	20
Pocas veces	5	18
Siempre	19	22

Pearson's Chi-squared test

data: extraRepaso X-squared = 6.5415, df = 4, p-value = 0.1622

la prueba chi-cuadrado nos da un p-valor de 0.1622 , para un nivel de significancia del $\alpha = 0.05$ no hay evidencia estadística para rechazar H_o , es decir el realizar actividades extracurriculares y la pregunta "¿repasa usted los temas vistos en clase?" son independientes

Estrato Social - actividades extracurriculares

Se observa la relación entre el estrato social y el realizar actividades extracurriculares

	1	2	3	4	5	6
NO	19	19	16	17	13	13
SI	22	22	19	19	24	17

Pearson's Chi-squared test

data: estratoExtra X-squared = 1.5599, df = 5, p-value = 0.9061

la prueba chi-cuadrado nos da un p-valor de 0.9064 , para un nivel de significancia del $\alpha = 0.05$ no hay evidencia estadística para rechazar H_o , es decir el estrato social y el realizar actividades extracurriculares son independientes.

Pruebas de Normalidad

Para la prueba de normalidad se considerada la hipótesis nula:

$$\begin{cases} H_0: & \text{La variable tiene distribución Normal} \\ H_1: & \text{las variables no tiene distribución normal} \end{cases}$$

puesto que la muestra es de 220 datos se usará el test mas potente que es el test de kolmogorov-smirnof en su variante el test de Lilie efors

Prueba para la edad

Lilliefors (Kolmogorov-Smirnov) normality test

data: Edad (Años) D = 0.13665, p-value = 9.794e-11

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable Edad no se distribuye de maneral normal .

Prueba para Tiempo de estudio semanal

Lilliefors (Kolmogorov-Smirnov) normality test

data: Tiempo de estudio semanal (h) D = 0.10943, p-value = 1.009e-06

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable Tiempo de estudio semanal no se distribuye de manera normal .

Prueba para el número de tutorias

Lilliefors (Kolmogorov-Smirnov) normality test

data: ¿A cuántas tutorias asistió durante el semestre? D = 0.13525, p-value = 1.665e-10

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable número de tutorias no se distribuye de manera normal .

Prueba para el Puntaje en matemáticas

Lilliefors (Kolmogorov-Smirnov) normality test

data: PUNTAJE MATEMÁTICAS D = 0.083536, p-value = 0.0007722

Shapiro-Wilk normality test

data: PUNTAJE MATEMÁTICAS W = 0.9874, p-value = 0.04939

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable Puntaje en matemáticas no se distribuye de manera normal .

Prueba para el Puntaje en Inglés

Lilliefors (Kolmogorov-Smirnov) normality test

data: PUNTAJE INGLES D = 0.089748, p-value = 0.0001899

Shapiro-Wilk normality test

data: PUNTAJE INGLES W = 0.9844, p-value = 0.016

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable Puntaje en Inglés no se distribuye de manera normal .

Prueba para el Puntaje en Ciencias Naturales

Lilliefors (Kolmogorov-Smirnov) normality test

data: PUNTAJE C. NATURALES D = 0.077534, p-value = 0.00267

Shapiro-Wilk normality test

data: PUNTAJE C. NATURALES W = 0.98188, p-value = 0.006352

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable Puntaje en Ciencias Naturales no se distribuye de manera normal .

Prueba para el Promedio en el Primer Semestre

Lilliefors (Kolmogorov-Smirnov) normality test

data: Promedio Primer Semestre D = 0.15422, p-value = 7.375e-14

Existe evidencia para rechazar H_0 , con un nivel de significancia de $\alpha = 0.05$ hay evidencia estadística para apoyar que la variable Promedio en el primer semestre no se distribuye de manera normal.

Planteamiento de hipótesis

Hipótesis estadística para la media:

El personal de estudios psicologicos desea saber si el promedio en matemáticas para los chicos de primer semestre es capaz de superar el humbral general que es de 42 puntos.

realizaremos la prueba de hipótesis para una media para la variable puntuación en matemáticas :

$$\begin{cases} H_0: & \mu \leq 42 \\ H_1: & \mu > 42 \end{cases}$$

los datos obtenidos son los siguientes:

$n = 220, \bar{X} = 42.26, S_x = 4.84$ para el estadístico de prueba se tendrá que :

$$T = \frac{\overline{X} - \mu_0}{\frac{S_X}{\sqrt{n}}}$$

```
##
##  One Sample t-test
##
## data:  PUNTAJE MATEMÁTICAS
## t = 0.79368, df = 219, p-value = 0.2141
## alternative hypothesis: true mean is greater than 42
## 95 percent confidence interval:
##  41.71986      Inf
## sample estimates:
## mean of x
##  42.25909
```

La prueba nos arroja un valor p de 0.2141 , con una significancia de $\alpha = 0.05$ no se reachaza H_0 es decir el valor de la media poblacional de la puntuación en matemáticas no supera el humbral.

Hipótesis estadística para la proporción :

el equipo psicologico de la universidad sugiere que tan solo el 25% de los estudiantes poseen una nota superior a 4.0

planteamos la hipótesis :

$$\begin{cases} H_0: & p \leq 0.25 \\ H_1: & p > 0.25 \end{cases}$$

donde p es la proporción de estudiantes que tiene 4 o mas en el semestre .

puesto que $np > 5$, $nq > 5$ y $n > 30$ es posible realizar la aproximación de la binomial a la normal .

la proporción muestral es de $\hat{p} = 0.253$ y $n = 220$ entonces usando el estadsítico de prueba

$$Z* = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = 0.10276$$

hallamos el valor p:

$$p - val = P(Z \geq Z*) = P(Z \geq 0.10276) = 0.4492328$$

para un p-valor menor o cercano no existe evidencia para rechazar H_0s , con una significancia del $\alpha = 0.05$ es decir el valor de la proporción es igual o menor a 0.25.

Hipótesis estadística para la diferencia de medias :

el equipo en psicologia desea saber si hay un diferencia significativa entre el promedio de primer semestre de las personas que si realizaron actividades curriculares y las que no .

planteamos la prueba de hipótesis :

$$\begin{cases} H_0: & \mu_{si} - \mu_{no} = 0 \\ H_1: & \mu_{si} - \mu_{no} \neq 0 \end{cases}$$

primero realizamos el test para comparar varianzas :

```
##
##  F test to compare two variances
##
## data:  mSi$‘Promedio Primer Semestre‘  and  mNO$‘Promedio Primer Semestre‘
## F = 1.0857, num df = 122, denom df = 96, p-value = 0.6768
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.739107 1.581031
## sample estimates:
## ratio of variances
##           1.085721
```

con un pvalor de 0.6768 no se rechaza H_0 , es decir las varianzas son iguales .

Realizamos el test de muestras no pareadas y varianzas iguales :

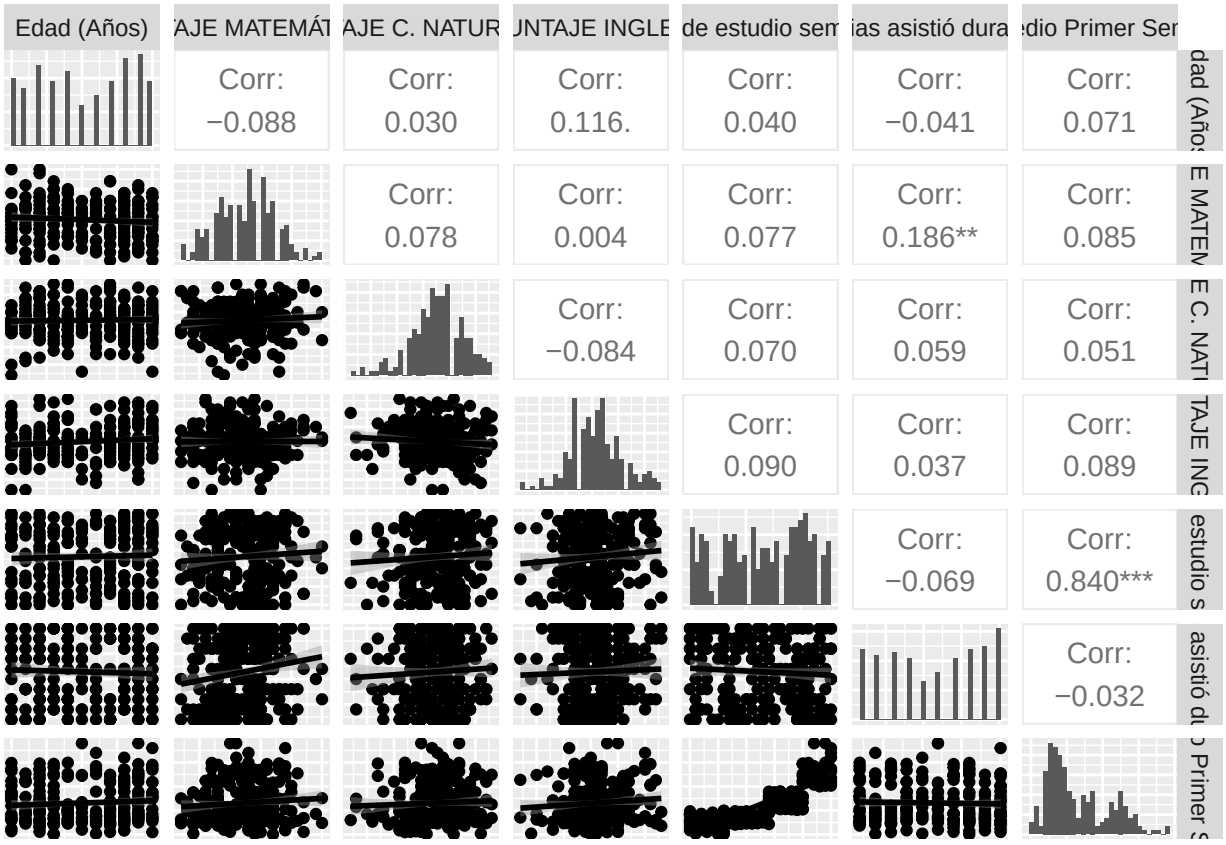
```
##
##  Two Sample t-test
##
## data:  mSi$‘Promedio Primer Semestre‘ and mNO$‘Promedio Primer Semestre‘
## t = -0.049351, df = 218, p-value = 0.9607
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1183993  0.1126148
## sample estimates:
## mean of x mean of y
##  3.654791  3.657683
```

Existe evidencia estadística para no rechazar H_0 , pues el p-valor es ded 0.8125 es demasiado alto para el valor de $\alpha = 0.05$ entonces no hay diferencia significativa entre las medias de las personas que realizan actividades extracurriculares .

Regresión Lineal Simple y Múltiple

Análisis de Correlación

Es deseable observar un diagrama de correlaciones para ver la posibles relaciones entre las varaibles y así mismo tener en cuenta las posibles predicciones.



existe una correlación leve entre el puntaje en matemáticas y el número de horas empleadas en el estudio esta es positiva .También encontramos una relación fuerte y positiva entre el número de horas de estudio y el promedio , esto es explicable ya que mayor número de horas se emplee mayor será la nota alcanzada .

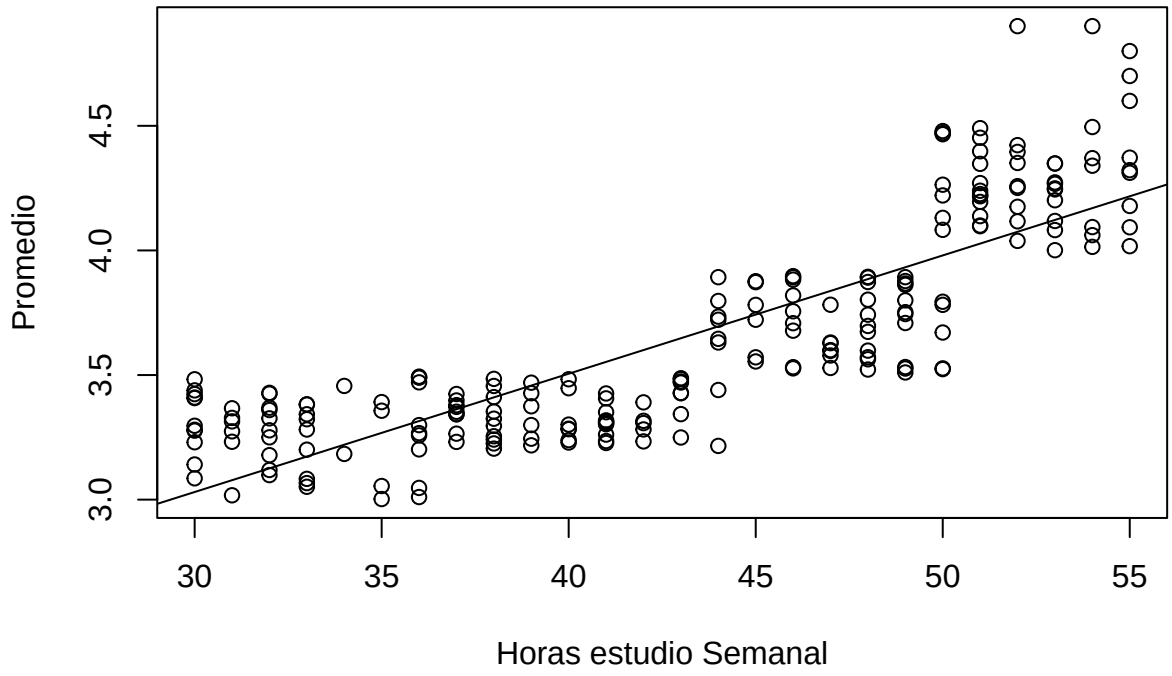
Regresión Lineal Simple.

Después de observar una relación fuerte entre el número de horas de estudio y el promedio se desea saber que tan buena es la predicción de las horas de estudio para el promedio general.

Se plantea el modelo de regresión de la siguiente manera

$$(promedio)_i = \beta_0 + \beta_1(Num.Horas)_i + \epsilon$$

donde la estimación de los valores será :




```
##
## Call:
## lm(formula = Base$‘Promedio Primer Semestre‘ ~ Base$‘Tiempo de estudio semanal (h)‘,
##     data = Base)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47946 -0.18797 -0.02077  0.16790  0.82512
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.60647    0.09118   17.62   <2e-16 ***
## Base$`Tiempo de estudio semanal (h)`  0.04747    0.00208   22.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2344 on 218 degrees of freedom
## Multiple R-squared:  0.705,  Adjusted R-squared:  0.7036
## F-statistic: 520.9 on 1 and 218 DF,  p-value: < 2.2e-16
```

entonces el modelo estimado será

$$(promedio)_i = 1.60647 + 0.04747(Num.Horas)_i$$

es decir por cada unidad que cambia en el número de horas la nota promedio aumenta en 0.04747 unidades y cuando se da que no hay que no hay horas de estudio la nota promedio estimada será de 1.6.

Es posbible observar la relación bien definida y ademas de causalidad donde el modelo propuesto es capaz de explicar el 70.36% de la variación en la nota promedio.

Regresión Lineal Múltiple

Observando las correlaciones entre las variables y las posibles explicaciones unilaterales

procederemos a hacer la selccción del mejor modelo por el proceso de Stepwise

```
## Start:  AIC=-621.01
## Base$‘Promedio Primer Semestre‘ ~ Género + ‘Estrato social‘ +
##     ‘Edad (Años)‘ + ‘PUNTAJE MATEMÁTICAS‘ + ‘PUNTAJE C. NATURALES‘ +
##     ‘PUNTAJE INGLES‘ + ‘Tiempo de estudio semanal (h)‘ + ‘¿Realiza actividades extrac
##     ‘¿Repasa los temas vistos en clase?‘ + ‘¿A cuántas tutorias asistió durante el se
##
##                                                 Df Sum of Sq    RSS
## - ‘¿Realiza actividades extracurriculares?‘      1    0.0034 11.518
## - Género                                         1    0.0049 11.519
## - ‘PUNTAJE C. NATURALES‘                         1    0.0050 11.519
## - ‘¿A cuántas tutorias asistió durante el semestre?‘ 1    0.0069 11.521
## - ‘Estrato social‘                               1    0.0077 11.522
## - ‘PUNTAJE INGLES‘                               1    0.0090 11.523
## - ‘PUNTAJE MATEMÁTICAS‘                          1    0.0151 11.529
## - ‘Edad (Años)‘                                  1    0.0288 11.543
## - ‘¿Repasa los temas vistos en clase?‘           4    0.3504 11.865
## <none>                                                      11.514
## - ‘Tiempo de estudio semanal (h)‘                1   27.5859 39.100
##                                                     AIC
## - ‘¿Realiza actividades extracurriculares?‘      -622.95
## - Género                                         -622.92
## - ‘PUNTAJE C. NATURALES‘                         -622.92
```

```
## - ‘¿A cuántas tutorias asistió durante el semestre?‘ -622.88
## - ‘Estrato social‘                                   -622.86
## - ‘PUNTAJE INGLES‘                                   -622.84
## - ‘PUNTAJE MATEMÁTICAS‘                              -622.72
## - ‘Edad (Años)‘                                      -622.46
## - ‘¿Repasa los temas vistos en clase?‘               -622.42
## <none>                                               -621.01
## - ‘Tiempo de estudio semanal (h)‘                    -354.05
##
## Step:  AIC=-622.95
## Base$‘Promedio Primer Semestre‘ ~ Género + ‘Estrato social‘ +
##     ‘Edad (Años)‘ + ‘PUNTAJE MATEMÁTICAS‘ + ‘PUNTAJE C. NATURALES‘ +
##     ‘PUNTAJE INGLES‘ + ‘Tiempo de estudio semanal (h)‘ + ‘¿Repasa los temas vistos en
##     ‘¿A cuántas tutorias asistió durante el semestre?‘
##
##                                                    Df Sum of Sq    RSS
## - Género                                            1    0.0048 11.522
## - ‘PUNTAJE C. NATURALES‘                            1    0.0064 11.524
## - ‘¿A cuántas tutorias asistió durante el semestre?‘ 1    0.0064 11.524
## - ‘Estrato social‘                                  1    0.0080 11.526
## - ‘PUNTAJE INGLES‘                                  1    0.0092 11.527
## - ‘PUNTAJE MATEMÁTICAS‘                             1    0.0166 11.534
## - ‘¿Repasa los temas vistos en clase?‘              4    0.3481 11.866
## - ‘Edad (Años)‘                                     1    0.0298 11.547
## <none>                                                      11.518
## - ‘Tiempo de estudio semanal (h)‘                   1   27.5827 39.100
##                                                        AIC
```

```
## - Género                                                -624.85
## - ‘PUNTAJE C. NATURALES‘                                -624.82
## - ‘¿A cuántas tutorias asistió durante el semestre?‘ -624.82
## - ‘Estrato social‘                                      -624.79
## - ‘PUNTAJE INGLES‘                                      -624.77
## - ‘PUNTAJE MATEMÁTICAS‘                                 -624.63
## - ‘¿Repasa los temas vistos en clase?‘                  -624.39
## - ‘Edad (Años)‘                                         -624.38
## <none>                                                  -622.95
## - ‘Tiempo de estudio semanal (h)‘                       -356.05
##
## Step:  AIC=-624.85
## Base$‘Promedio Primer Semestre‘ ~ ‘Estrato social‘ + ‘Edad (Años)‘ +
##     ‘PUNTAJE MATEMÁTICAS‘ + ‘PUNTAJE C. NATURALES‘ + ‘PUNTAJE INGLES‘ +
##     ‘Tiempo de estudio semanal (h)‘ + ‘¿Repasa los temas vistos en clase?‘ +
##     ‘¿A cuántas tutorias asistió durante el semestre?‘
##
##                                                        Df Sum of Sq    RSS
## - ‘¿A cuántas tutorias asistió durante el semestre?‘  1    0.0057 11.528
## - ‘Estrato social‘                                     1    0.0072 11.530
## - ‘PUNTAJE C. NATURALES‘                               1    0.0074 11.530
## - ‘PUNTAJE INGLES‘                                     1    0.0086 11.531
## - ‘PUNTAJE MATEMÁTICAS‘                                1    0.0188 11.541
## - ‘¿Repasa los temas vistos en clase?‘                 4    0.3437 11.866
## - ‘Edad (Años)‘                                        1    0.0318 11.554
## <none>                                                             11.522
## - ‘Tiempo de estudio semanal (h)‘                      1   27.8812 39.404
```

```
##                                                        AIC
## - ‘¿A cuántas tutorias asistió durante el semestre?‘ -626.74
## - ‘Estrato social‘                                   -626.72
## - ‘PUNTAJE C. NATURALES‘                             -626.71
## - ‘PUNTAJE INGLES‘                                   -626.69
## - ‘PUNTAJE MATEMÁTICAS‘                              -626.50
## - ‘¿Repasa los temas vistos en clase?‘               -626.39
## - ‘Edad (Años)‘                                      -626.25
## <none>                                               -624.85
## - ‘Tiempo de estudio semanal (h)‘                    -356.35
##
## Step:  AIC=-626.74
## Base$‘Promedio Primer Semestre‘ ~ ‘Estrato social‘ + ‘Edad (Años)‘ +
##     ‘PUNTAJE MATEMÁTICAS‘ + ‘PUNTAJE C. NATURALES‘ + ‘PUNTAJE INGLES‘ +
##     ‘Tiempo de estudio semanal (h)‘ + ‘¿Repasa los temas vistos en clase?‘
##
##                                       Df Sum of Sq    RSS     AIC
## - ‘PUNTAJE C. NATURALES‘               1    0.0068 11.535 -628.62
## - ‘Estrato social‘                     1    0.0068 11.535 -628.61
## - ‘PUNTAJE INGLES‘                     1    0.0096 11.538 -628.56
## - ‘PUNTAJE MATEMÁTICAS‘                1    0.0235 11.552 -628.30
## - ‘Edad (Años)‘                        1    0.0306 11.559 -628.16
## - ‘¿Repasa los temas vistos en clase?‘ 4    0.3607 11.889 -627.97
## <none>                                             11.528 -626.74
## - ‘Tiempo de estudio semanal (h)‘      1   28.0017 39.530 -357.65
##
## Step:  AIC=-628.62
```

```
## Base$‘Promedio Primer Semestre‘ ~ ‘Estrato social‘ + ‘Edad (Años)‘ +
##     ‘PUNTAJE MATEMÁTICAS‘ + ‘PUNTAJE INGLES‘ + ‘Tiempo de estudio semanal (h)‘ +
##     ‘¿Repasa los temas vistos en clase?‘
##
##                                       Df Sum of Sq    RSS     AIC
## - ‘Estrato social‘                     1    0.0074 11.542 -630.47
## - ‘PUNTAJE INGLES‘                     1    0.0113 11.546 -630.40
## - ‘PUNTAJE MATEMÁTICAS‘                1    0.0220 11.557 -630.20
## - ‘Edad (Años)‘                        1    0.0297 11.565 -630.05
## - ‘¿Repasa los temas vistos en clase?‘ 4    0.3580 11.893 -629.89
## <none>                                            11.535 -628.62
## - ‘Tiempo de estudio semanal (h)‘      1   28.0882 39.623 -359.13
##
## Step:  AIC=-630.47
## Base$‘Promedio Primer Semestre‘ ~ ‘Edad (Años)‘ + ‘PUNTAJE MATEMÁTICAS‘ +
##     ‘PUNTAJE INGLES‘ + ‘Tiempo de estudio semanal (h)‘ + ‘¿Repasa los temas vistos en
##
##                                       Df Sum of Sq    RSS     AIC
## - ‘PUNTAJE INGLES‘                     1    0.0117 11.554 -632.25
## - ‘PUNTAJE MATEMÁTICAS‘                1    0.0221 11.564 -632.05
## - ‘Edad (Años)‘                        1    0.0288 11.571 -631.93
## - ‘¿Repasa los temas vistos en clase?‘ 4    0.3524 11.895 -631.86
## <none>                                            11.542 -630.47
## - ‘Tiempo de estudio semanal (h)‘      1   28.1032 39.646 -361.00
##
## Step:  AIC=-632.25
## Base$‘Promedio Primer Semestre‘ ~ ‘Edad (Años)‘ + ‘PUNTAJE MATEMÁTICAS‘ +
```

```
##     ‘Tiempo de estudio semanal (h)‘ + ‘¿Repasa los temas vistos en clase?‘
##
##                                       Df Sum of Sq    RSS     AIC
## - ‘PUNTAJE MATEMÁTICAS‘                1    0.0224 11.577 -633.82
## - ‘¿Repasa los temas vistos en clase?‘  4    0.3441 11.898 -633.79
## - ‘Edad (Años)‘                        1    0.0338 11.588 -633.61
## <none>                                             11.554 -632.25
## - ‘Tiempo de estudio semanal (h)‘      1   28.3563 39.910 -361.54
##
## Step:  AIC=-633.82
## Base$‘Promedio Primer Semestre‘ ~ ‘Edad (Años)‘ + ‘Tiempo de estudio semanal (h)‘ +
##     ‘¿Repasa los temas vistos en clase?‘
##
##                                       Df Sum of Sq    RSS     AIC
## - ‘¿Repasa los temas vistos en clase?‘  4    0.3454 11.922 -635.35
## - ‘Edad (Años)‘                        1    0.0290 11.606 -635.27
## <none>                                             11.577 -633.82
## - ‘Tiempo de estudio semanal (h)‘      1   28.6916 40.268 -361.58
##
## Step:  AIC=-635.35
## Base$‘Promedio Primer Semestre‘ ~ ‘Edad (Años)‘ + ‘Tiempo de estudio semanal (h)‘
##
##                                     Df Sum of Sq    RSS     AIC
## - ‘Edad (Años)‘                      1    0.0583 11.980 -636.28
## <none>                                           11.922 -635.35
## - ‘Tiempo de estudio semanal (h)‘    1   28.4776 40.400 -368.86
##
```

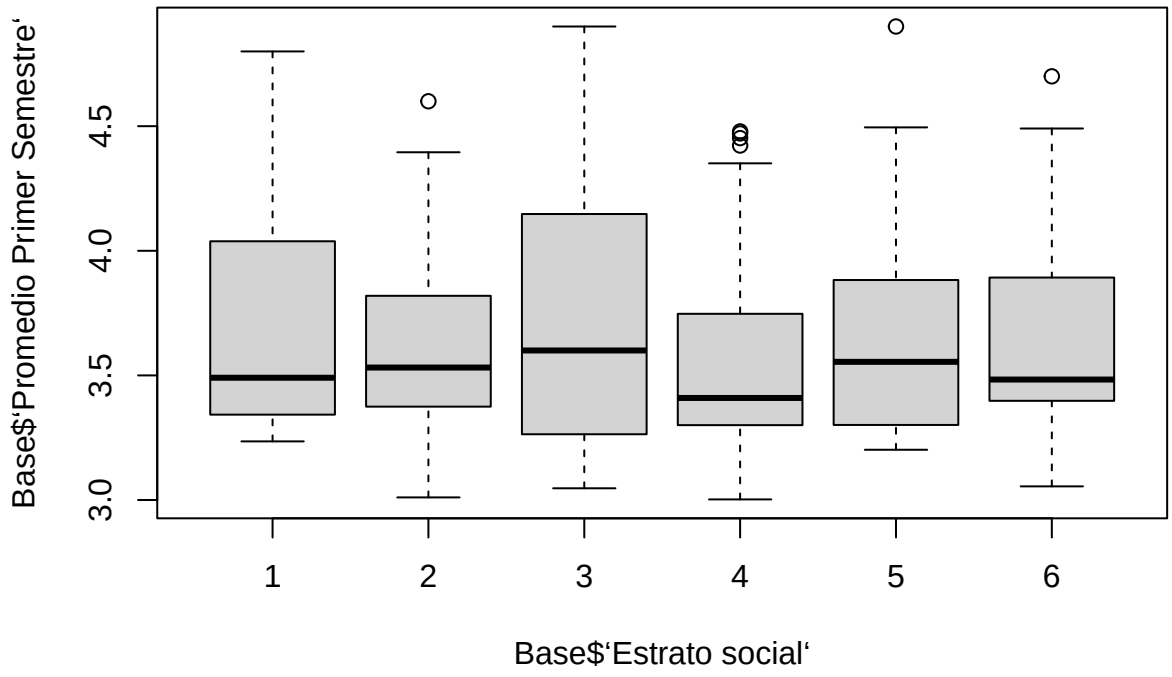
```
## Step:  AIC=-636.28
## Base$‘Promedio Primer Semestre‘ ~ ‘Tiempo de estudio semanal (h)‘
##
##                                   Df Sum of Sq    RSS     AIC
## <none>                                          11.980 -636.28
## - ‘Tiempo de estudio semanal (h)‘  1    28.627 40.607 -369.73

##
## Call:
## lm(formula = Base$‘Promedio Primer Semestre‘ ~ ‘Tiempo de estudio semanal (h)‘,
##     data = Base)
##
## Coefficients:
##                     (Intercept)  ‘Tiempo de estudio semanal (h)‘
##                         1.60647                          0.04747
```

Es posible observar que el mejor modelo que se puede tener es el modelo lineal planteado anteriormente.

ANOVA (Análisis de la varianza)

Primero observemos como se comporta la relación entre el promedio del primer semestre y el estrato social




```
## Call:
##    aov(formula = regreAnova)
##
## Terms:
##                 Base$‘Estrato social‘ Residuals
## Sum of Squares                0.02106  40.58594
## Deg. of Freedom                     1       218
##
## Residual standard error: 0.4314789
```

```
## Estimated effects may be unbalanced
```

podemos ver que se NO se rechaza la hipotésis nula de la igualdad de medias , por lo tanto el estrato social no influye en el promedio de manera significativa.

Conclusiones

Es posible concluir que las notas de las pruebas icfes no son útiles para precedir el promedio de los estudiantes que entran a a facultad de psicologia , sin embargo en la busqueda a la respuesta de este interrogante encontramos una relación entre el número de horas dedicacas al estudio y el incremento dela nota promedio de los estudiantes . Además , se encontró un modelo para poder precedicir con una acertividad del 70% aproximadamente.

References

Hernández, F. (2021, 26 julio). Manual de R. R manual. https://fhernanb.github.io/Manual-de-R/

Sancho, R. S. (2020). filter() | Programación en R. R. https://rsanchezs.gitbooks.io/rprogramming/content/chapter9/filter.html

RPubs - Contrastes de hipÃ3tesis en R. (2018, 25 abril). Hipotesis. https://rpubs.com/Jo_/contrastes_hipotesis_ttest

RPubs - Analisis_Univariable. (2014, 27 octubre). Analisis. https://rpubs.com/dsulmont/37910

RPubs - 3.3.2. Tablas de frecuencias agrupadas en R. (2021, 29 junio). tablas. https://rpubs.com/hllinas/R_Tablas_agrupadas

27 R Markdown | _main.utf8. (2020). Rmarkdown. https://es.r4ds.hadley.nz/r-markdown.html

Zhu, H. (2021, 19 febrero). Create Awesome HTML Table with knitr::kable and kableExtra. tabla. https://cran.r-project.org/web/packages/kableExtra/vignettes/awesome_table_in_html.html#Position