Estadística Descriptiva e Inferencial

Maestría en Ciencias Computacionales

Gabriel Mantilla Saltos

2022-02-09

Índice

Resumen

Este repositorio contiene un set de datos obtenido de la plataforma Kaggle. La tarea consiste en utilizar la herramienta R Studio para presentar un análisis de los datos. El archivo con los datos es **StudentsPerformance.csv**, Las variables de interés en este dataset son las que miden el rendimiento académico del estudiante son:

* math score.

* reading score.

* writing score.

* El trabajo consiste en usar la estadística descriptiva para presentar un resumen gráfico de los datos y la inferencial para responder a las siguientes preguntas:

* ¿Qué factores afectan más el rendimiento del estudiante? ¿Cómo?

El análisis debe ser hecho usando R Studio y R Markdown (o un Notebook de R), es decir en este repositorio deben incluirse archivos tipo Rmd que al compilarse generen el reporte.

* Estadística descriptiva: El reporte debe hacer un resumen de los datos usando gráficos relevantes (diagramas de caja, dispersión, densidad, etc.) y estadísticas de dispersión y tendencia central.

* Estadística inferencial: El reporte debe hacer pruebas de hipótesis y/o regresión lineal múltiple para justificar sus conclusiones. Las pruebas de hipótesis deben demostrar lo que asumen (por ejemplo normalidad en los datos) y presentar claramente las hipótesis nulas y alternativas con el respectivo p-value y alpha.

Estadística Descriptiva

Analizamos estadísticas descriptivas para las notas de matemáticas por las siguientes variables categóricas: género, raza étnica, nivel de educación, tipo de lunch y la verificación de completado del test:

Género.

```
##   gender   n  mean    sd min Q1 median Q3 max percZero mode
## 1 female 518 63.63 15.49   0 54     65 74 100     0.19   65
## 2   male 482 68.73 14.36  27 59     69 79 100     0.00   62
```

Raza Étnica.

```
##      race   n  mean    sd min    Q1 median Q3 max percZero mode
## 1 group A  89 61.63 14.52  28 51.00   61.0 71 100     0.00   58
## 2 group B 190 63.45 15.47   8 54.00   63.0 74  97     0.00   62
## 3 group C 319 64.46 14.85   0 55.00   65.0 74  98     0.31   65
## 4 group D 262 67.36 13.77  26 59.00   69.0 77 100     0.00   69
## 5 group E 140 73.82 15.53  30 64.75   74.5 85 100     0.00   79
```

Nivel de educación

```
##            education   n  mean    sd min    Q1 median Q3 max percZero mode
## 1 associate's degree 222 67.88 15.11  26 57.00   67.0 80 100     0.00   65
## 2  bachelor's degree 118 69.39 14.94  29 61.00   68.0 79 100     0.00   65
## 3        high school 196 62.14 14.54   8 53.75   63.0 72  99     0.00   62
## 4    master's degree  59 69.75 15.15  40 55.50   73.0 81  95     0.00   80
## 5       some college 226 67.13 14.31  19 59.00   67.5 76 100     0.00   69
## 6   some high school 179 63.50 15.93   0 53.00   65.0 74  97     0.56   62
```

Lunch

```
##          lunch   n  mean    sd min Q1 median Q3 max percZero mode
## 1 free/reduced 355 58.92 15.16   0 49     60 69 100     0.28   61
## 2     standard 645 70.03 13.65  19 61     69 80 100     0.00   69
```

Test

```
##        test   n  mean    sd min Q1 median    Q3 max percZero mode
## 1 completed 358 69.70 14.44  23 60     69 79.00 100     0.00   65
## 2      none 642 64.08 15.19   0 54     64 74.75 100     0.16   62
```

Para el Género la nota promedio más alta (**68**) es para los hombres, aunque su moda sea de **62**. Mientras que, si analizamos por raza, el promedio más alto es de **73.8** para el grupo E, con una moda de **79**, incluso los datos ordenados hasta el 75 % para el mismo grupo se mantiene más alto (**85**) que los demás. El nivel de educación con la nota más alta es para los universitarios con pregrado, con **69.7** siendo su moda de **80**. Algo interesante es que la nota promedio para los que tienen un lunch estándar es de **70** con respecto a los free/reduce de **58** siendo significaativo el tipo de lunch en la nota. Por último la cantidad de alumnos que completaron el examen fue de **358** de los **1000** que lo dieron.

Analicemos las estadísticas descriptivas de las notas de lectura por las variables categóricas mencionadas anteriormente:

Género.

```
##   gender   n  mean    sd min    Q1 median Q3 max percZero mode
## 1 female 518 72.61 14.38  17 63.25     73 83 100        0   72
## 2   male 482 65.47 13.93  23 56.00     66 75 100        0   70
```

Raza Étnica.

```
##      race   n  mean    sd min    Q1 median    Q3 max percZero mode
## 1 group A  89 64.67 15.54  23 53.00     64 74.00 100        0   67
## 2 group B 190 67.35 15.18  24 56.00     67 79.75  97        0   65
## 3 group C 319 69.10 14.00  17 60.00     71 78.50 100        0   72
## 4 group D 262 70.03 13.90  31 60.25     71 79.00 100        0   72
## 5 group E 140 73.03 14.87  26 63.00     74 84.00 100        0   75
```

Nivel de educación

```
##            education   n  mean    sd min   Q1 median    Q3 max percZero mode
## 1 associate's degree 222 70.93 13.87  31 61.0   72.5 81.00 100        0   76
## 2  bachelor's degree 118 73.00 14.29  41 63.0   73.0 82.75 100        0  100
## 3        high school 196 64.70 14.13  24 54.0   66.0 74.25  99        0   62
## 4    master's degree  59 75.37 13.78  42 65.5   76.0 84.50 100        0   81
## 5       some college 226 69.46 14.06  23 60.0   70.5 79.75 100        0   72
## 6   some high school 179 66.94 15.48  17 56.5   67.0 79.00 100        0   76
```

Lunch

```
##          lunch   n  mean    sd min Q1 median Q3 max percZero mode
## 1 free/reduced 355 64.65 14.90  17 56     65 75 100        0   58
## 2     standard 645 71.65 13.83  26 63     72 82 100        0   64
```

Test

```
##        test   n  mean    sd min Q1 median Q3 max percZero mode
## 1 completed 358 73.89 13.64  37 65     75 84 100        0   74
## 2      none 642 66.53 14.46  17 57     67 76 100        0   72
```

En conclusión, para el Género, las mujeres obtuvieron en promedio una mayor nota (**72**) con respecto a los hombres, aunque su moda es similar (**72** mujeres ~ **70** hombres). Por otro lado, la raza étnica del grupo E obtuvo la mayor nota con **73** puntos en promedio, siendo un tercer cuartil de **84**. El nivel de educación en cambio indica que los que tienen maestría obtuvieron una mayor nota (**75**) con respecto a los demás, aunque la moda de los estudiantes de pregrado obtuvo **100**. Del mismo modo los estudiantes que tienen un lunch estándar obtuvieron una nota más alta (**71**). Y por último los que completaron el examen obtuvieron una nota más alta de **73.8** con respecto a los que no en su totalidad **66**.

Analicemos ahora las estadísticas descriptivas de las notas de escritura, para las variables categóricas:

Género.

```
##   gender   n  mean    sd min Q1 median    Q3 max percZero mode
## 1 female 518 72.47 14.84  10 64     74 82.00 100        0   70
## 2   male 482 63.31 14.11  15 53     64 73.75 100        0   68
```

Raza Étnica.

```
##      race   n  mean    sd min    Q1 median    Q3 max percZero mode
## 1 group A  89 62.67 15.47  19 51.00     62 73.00  97        0   54
## 2 group B 190 65.60 15.63  15 55.25     67 78.00  96        0   68
## 3 group C 319 67.83 14.98  10 57.00     68 79.00 100        0   74
## 4 group D 262 70.15 14.37  32 61.00     72 80.00 100        0   74
## 5 group E 140 71.41 15.11  22 62.00     72 80.25 100        0   70
```

Nivel de educación

```
##            education   n  mean    sd min   Q1 median Q3 max percZero mode
## 1 associate's degree 222 69.90 14.31  35 58.0   70.5 80 100        0   73
## 2  bachelor's degree 118 73.38 14.73  38 62.5   74.0 83 100        0   90
## 3        high school 196 62.45 14.09  15 52.0   64.0 73 100        0   68
## 4    master's degree  59 75.68 13.73  46 67.0   75.0 85 100        0   86
## 5       some college 226 68.84 15.01  19 60.0   70.0 79  99        0   70
## 6   some high school 179 64.89 15.74  10 54.0   66.0 77 100        0   78
```

Lunch

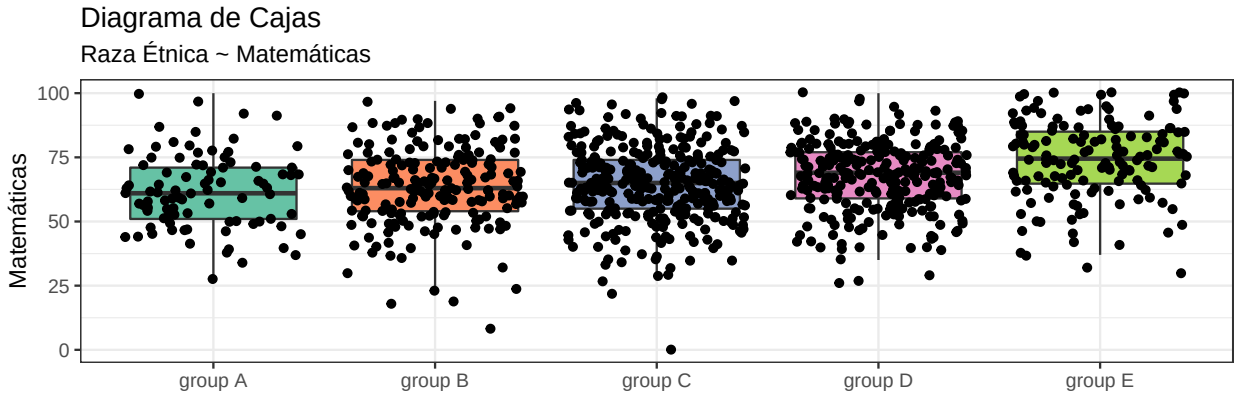
```
##          lunch   n  mean    sd min Q1 median Q3 max percZero mode
## 1 free/reduced 355 63.02 15.43  10 53     64 74 100        0   54
## 2     standard 645 70.82 14.34  22 62     72 81 100        0   74
```

Test

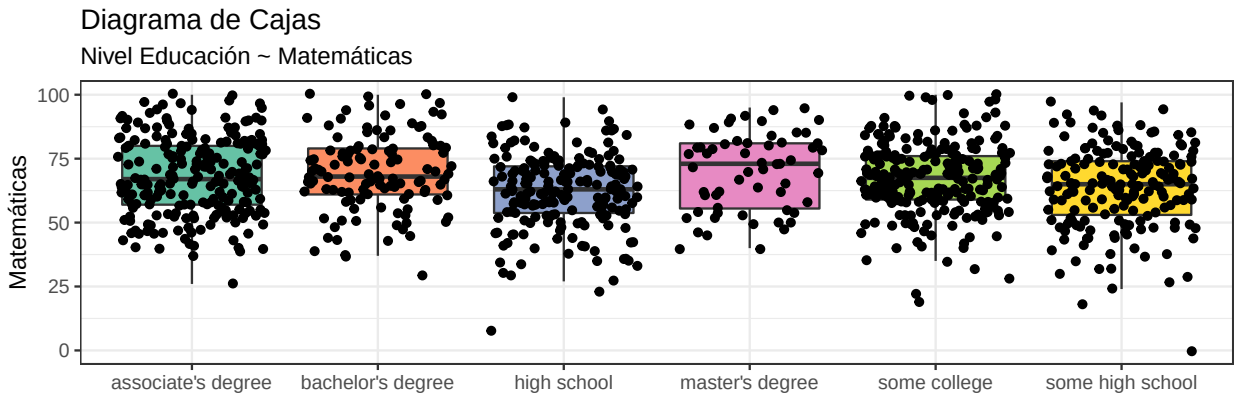
```
##        test   n  mean    sd min Q1 median Q3 max percZero mode
## 1 completed 358 73.89 13.64  37 65     75 84 100        0   74
## 2      none 642 66.53 14.46  17 57     67 76 100        0   72
```

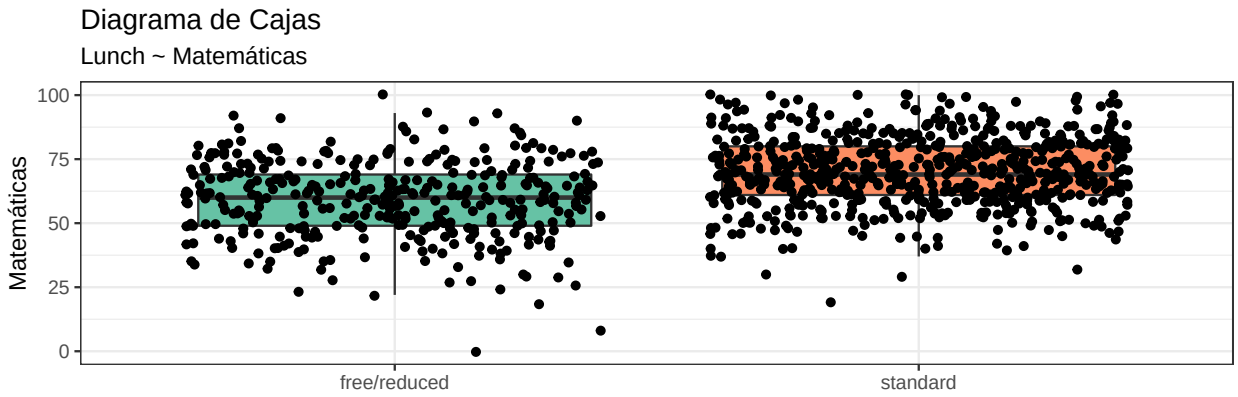
En conclusión, tenemos que las mujeres obtuvieron una mayor nota promedio de **72** con respecto a los hombres de **63**. La raza étnica del grupo E mantienen la punta con **71** puntos en promedio y un tercer cuartil de **80.2**. Por otro lado, el nivel de educación de los estudiantes que tienen una maestría, obtuvieron la mayor nota de **75.6** con una moda de **86**, mientras que el lunch mantiene la punta para los que toman un estándar, con un cuartil 1 (datos ordenandos hasta el 25 % de los datos) mayor de 62 con respecto a los **53** de los que toman un lunch free/reduced. Y por último los 358 estudiantes que completaron el examen obtuvieron una mayor nota de **73**, con respecto a los que no **66**.

Analicemos la dispersión de las notas por Género y raza Étnica. Podemos afirmar que la mayor cantidad de datos para ambos géneros se concentra en un rango entre **50** y **80** puntos, la dispersión en ambos géneros es similar y la posición de las cajas bordean una mediana de **70** puntos. Mientras que para la raza étnica, tenemos que el Grupo E sus puntos tienen una mediana mas alta (**75**).

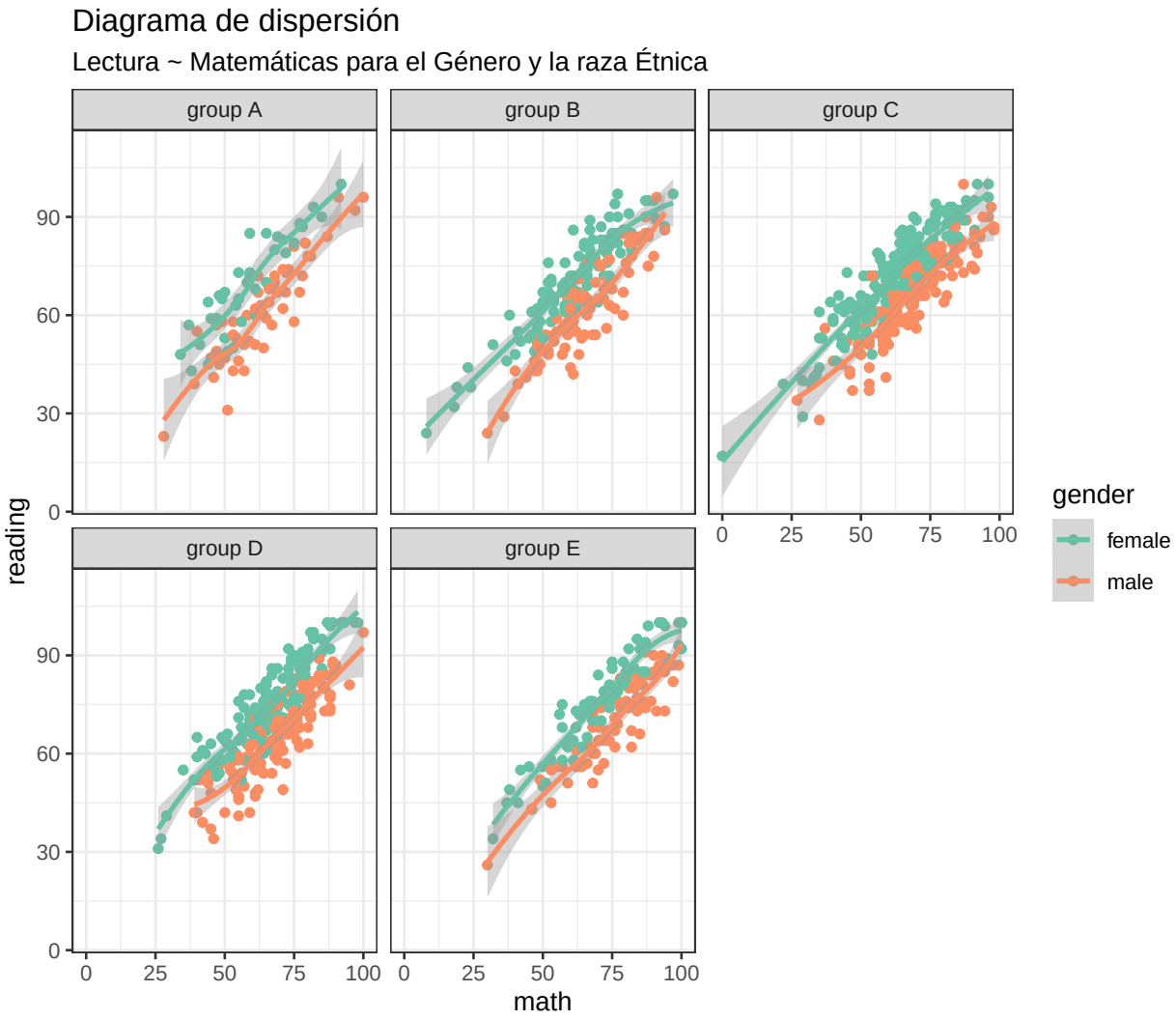


Analicemos la dispersión de las notas por Nivel de Educación y el Lunch . Podemos afirmar que la caja que más alta esta es la de los estudiantes de maestría con una mediana cercana a **75**, se observa un valor muy pequeño para some high school, probablemente alguien no dio el examen. Con respecto al lunch, se ve que la mayor cantidad de puntos del lunch estándar son más altos que free/reduced, explicando que los estudiantes que toman ese lunch mejoran sus notas de matemáticas.

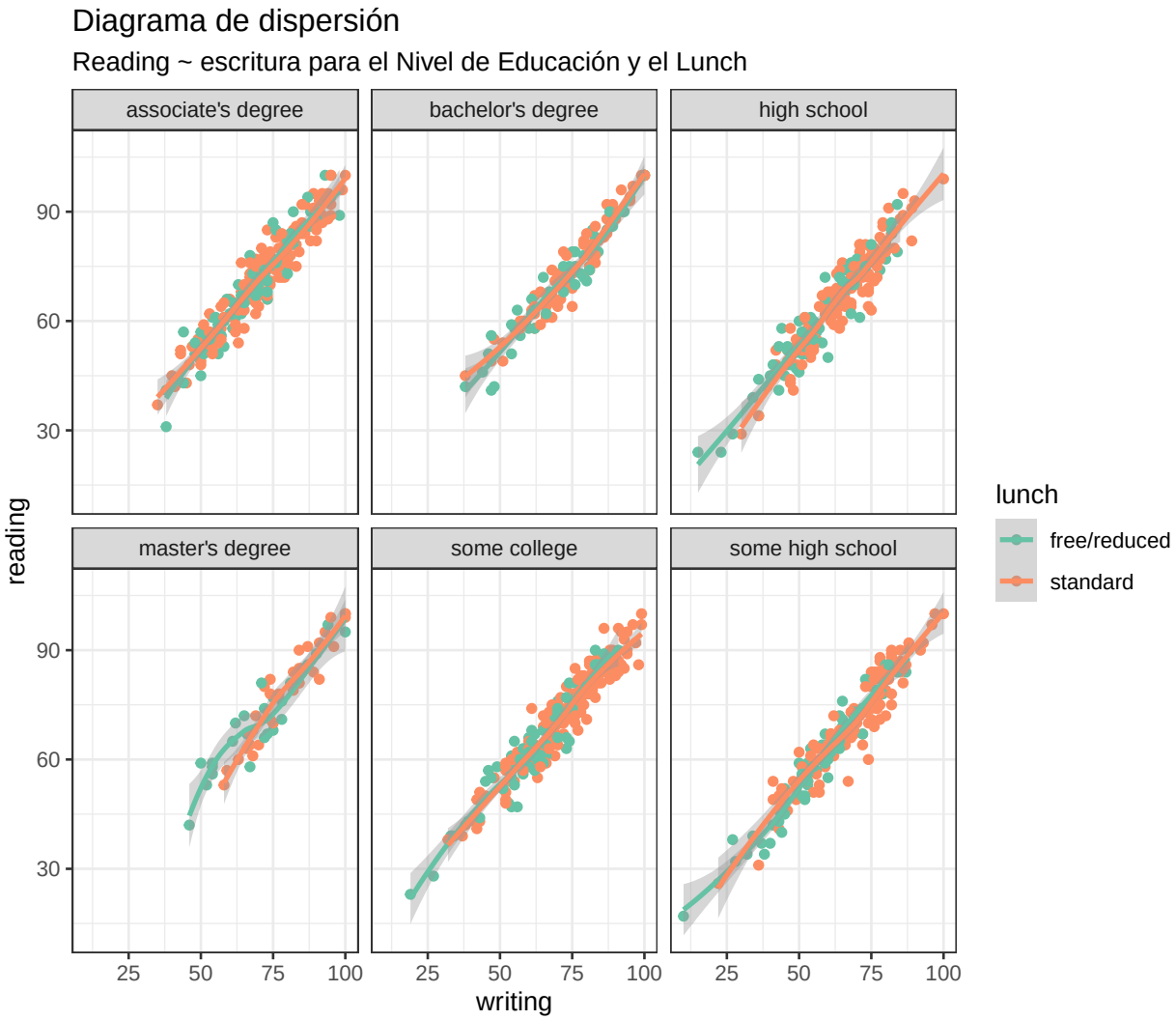




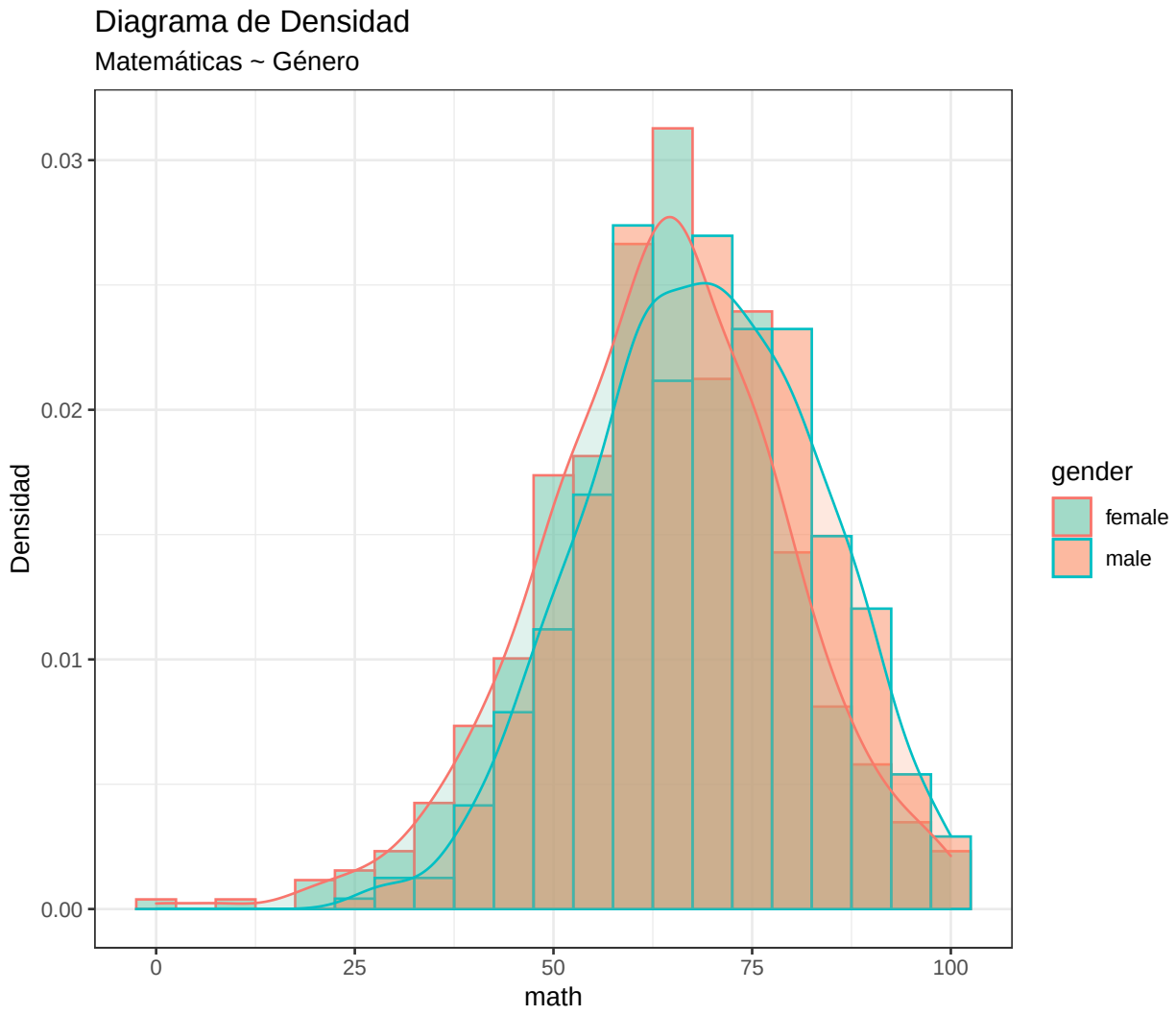
Ahora veamos gráficamente las dependencias del género entre las notas de lectura ~ matemáticas. Del siguiente gráfico podemos observar que la relación es lineal positivo entre las dos variables, y que las mujeres tienen mejores notas de lectura y que los hombres mejores notas de matemáticas en todos los grupos étnicos.



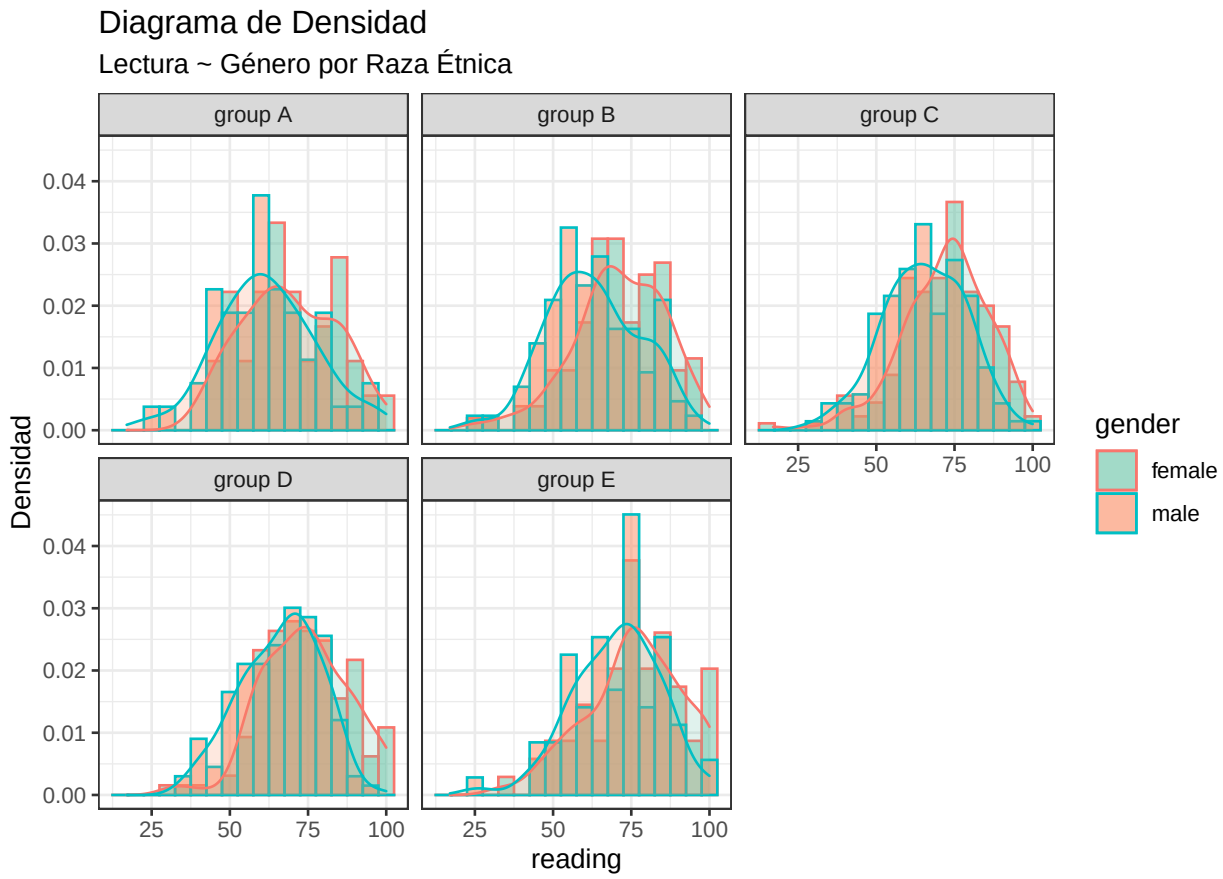
Si analizamos la relación entre las notas de escritura y lectura, para las variables categóricas nivel de educación y Lunch, podemos observar que la relación es positiva y en magnitud son muy similares dado que la nube de puntos del Lunch estándar y free/reduced son las mismas.



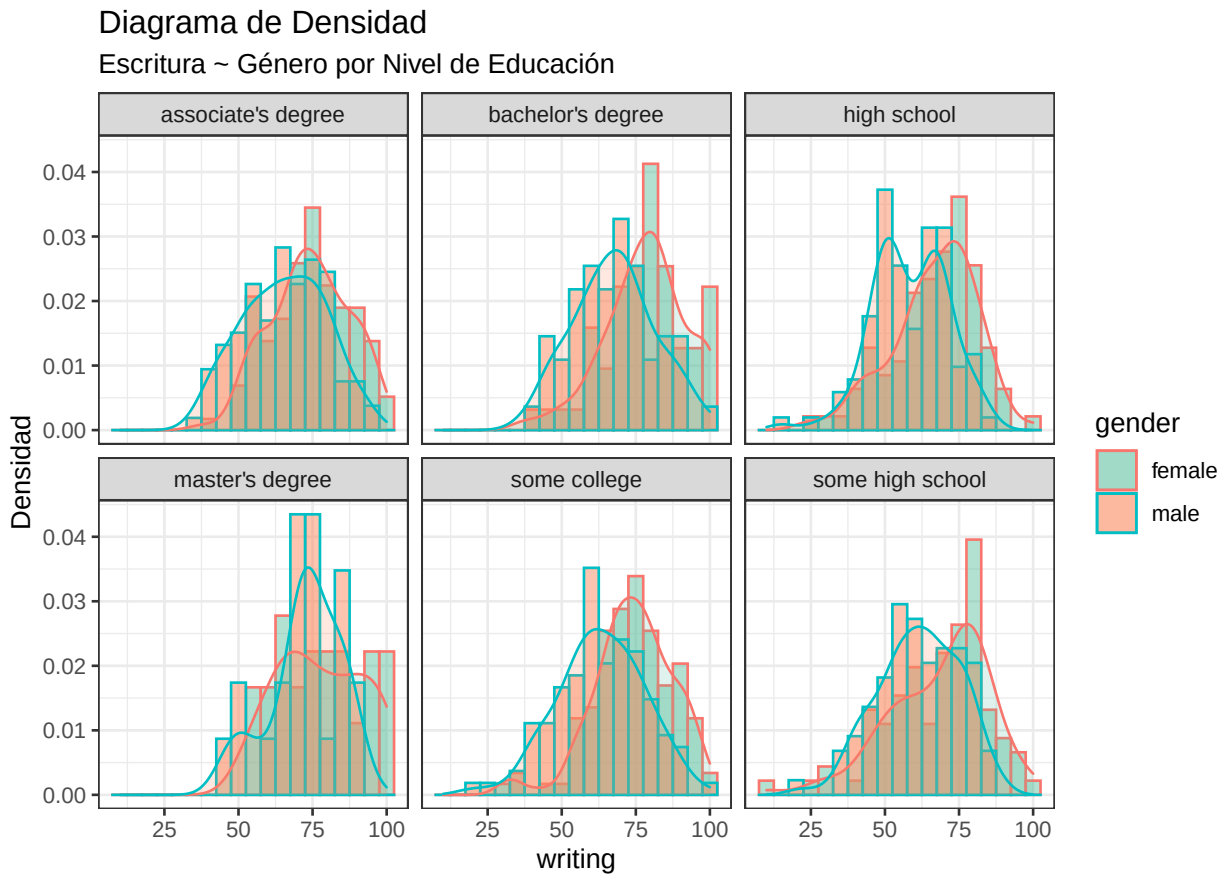
Si analizamos el diagrama de densidad de la variable matemáticas, en función del Género, podemos ver que la densidad de los datos es ligeramente mayor para las mujeres, y que el histograma de los hombres esta ligeramente más hacia la derecha, indicando notas más altas. También podemos observar que los datos tienden a una distribución normal para ambos géneros, esto lo podremos concluir más adelante.



Analizando el diagrama de Densidad de las notas de Lectura en función al Género ,por la raza Étnica, podemos observar que las densidad son diferentes para los distintos grupos étnicos, teniendo una ligera mayor densidad el grupo E y más hacia la derecha, indicando que las notas son mayores para este grupo. Las curvas no tienen el comportamiento de una distribución normal, solo observando los gráficos.



Analizando el diagrama de Densidad de las notas de Escritura en función al Género, por el Nivel de Educación, podemos observar que las densidades son ligeramente diferentes para los niveles de educación, los comportamientos más diferentes se observan en el grupo de High School para el género. El comportamiento de los histogramas para el grupo de estudiantes con maestrías, para el género masculino y femenino es significativamente variable según sus bandas.



Estadística Inferencial

Realizamos la prueba chi-cuadrado, que utiliza una aproximación a la distribución chi cuadrado de Pearson, para evaluar la discrepancia igual o mayor que la que exista entre los datos y las frecuencias esperadas. Para determinar si las variables son independientes con el nivel de significancia alpha, se rechaza la hipótesis nula, en base al valor p, si es menor o igual al nivel de significancia.

* H0: No hay diferencia en la proporción de estudiantes Hombres y Mujeres con respecto a su raza Étnica.
* H1: Si hay diferencia.

```
##
##  Pearson's Chi-squared test
##
## data:  table(DATA$gender, DATA$race)
## X-squared = 9.0274, df = 4, p-value = 0.06042
```

No Podemos concluir que las variables de Género y Raza Étnica son independientes, con un valor p mayor a 0.05, no tenemos suficinte evidencia estad+isticapara aceptar o rechazar la hipótesis nula.

Ahora estudiaremos la independencia entre el Género y el Nivel de Educación.

* H0: No hay diferencia en la proporción de estudiantes Hombres y Mujeres con respecto a su Nivel de Educación.

* H1: Si hay diferencia

```
##
##  Pearson's Chi-squared test
##
## data:  table(DATA$gender, DATA$education)
## X-squared = 3.3849, df = 5, p-value = 0.6409
```

Aceptamos la hipótesis Nula, de que no hay diferencias entre las proporciones de estudiantes hombres y mujeres con respecto a su nivel de educación.

Estudiaremos una prueba para contrastar la normalidad en las variables de interés. La prueba de Hipótesis de Shapiro Wilk, plantea que si una muestra x1, . . . , xn proviene de una población normalmente distribuida. H0: La distribución es normal H1: La distribución no es normal,

```
##
##  Shapiro-Wilk normality test
##
## data:  DATA$math
## W = 0.99315, p-value = 0.0001455
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATA$reading
## W = 0.99292, p-value = 0.0001055
```

```
##
##  Shapiro-Wilk normality test
##
## data:  DATA$writing
## W = 0.99196, p-value = 2.922e-05
```

Podemos concluir que, para las 3 variables, el valor p es cercano a cero, es decir menor al nivel de significancia, por tanto, tenemos evidencia estadística para rechazar la idea de que las variables pertenecen a una población normalmente distribuida. De lo cual teniamos dudas de los histogramas en la sección anterior.

Dado que se no se asume normalidad en la variable Matemáticas, y que el tamaño de cada grupo de hombres y mujeres es mayor a 30, podriamos considerar usar la prueba t para comparacion de medias, siendo suficientemente robusto para obtener una conclusion mas acertada. Por otro lado por la proporcion de mujeres y hombres es necesario realizar una prueba de independencia para el Género, dado que la cantidad de estudiantes es similar (518 Mujeres, 482 Hombres).

* Ho: La proporción de estudiantes hombres es la misma que de la de mujeres.
* Ha: la proporción de estudiantes es diferente.

```
##
##  Chi-squared test for given probabilities
##
## data:  table(DATA$gender)
## X-squared = 1.296, df = 1, p-value = 0.2549
```

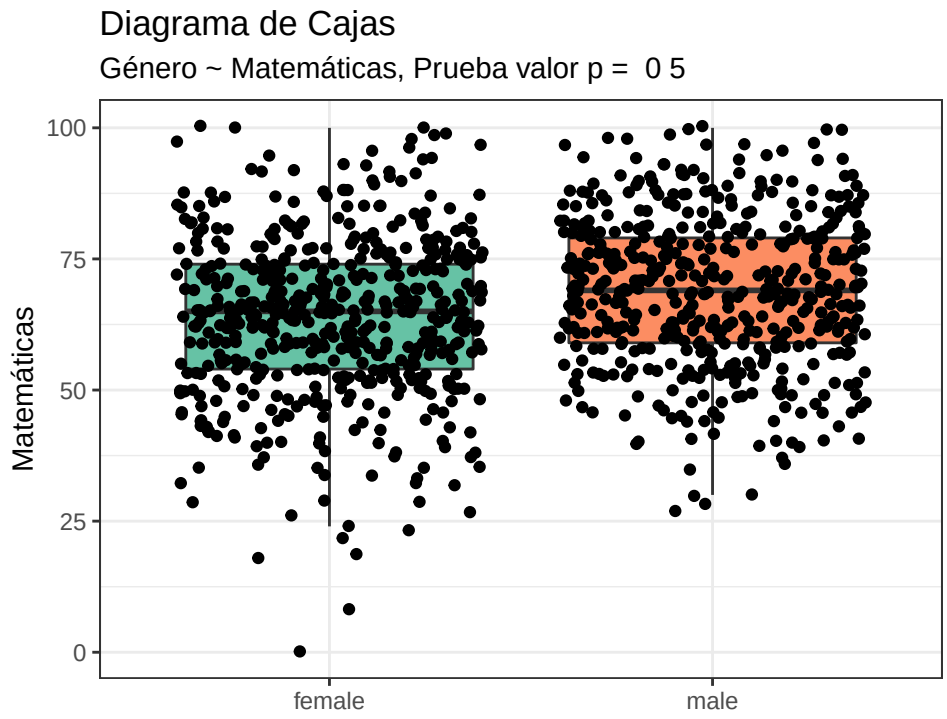
Podemos concluir que la proporción es la misma dado que el valor p es mayor que 0.05 y podemos aceptar la hipótesis nula. Ahora podemos obtener la diferencia entre las medias muestrales y suponer que la diferencia de medias es significativa (**5.09** puntos). Sin embargo, dado que el tamaño de cada grupo es mayor que 30 se puede considerar que el t-test que es suficientemente robusto para obtener una conclusión más clara.

La diferencia de medias muestrales es significativa para las notas de matemáticas entre los hombres y las mujeres es de:

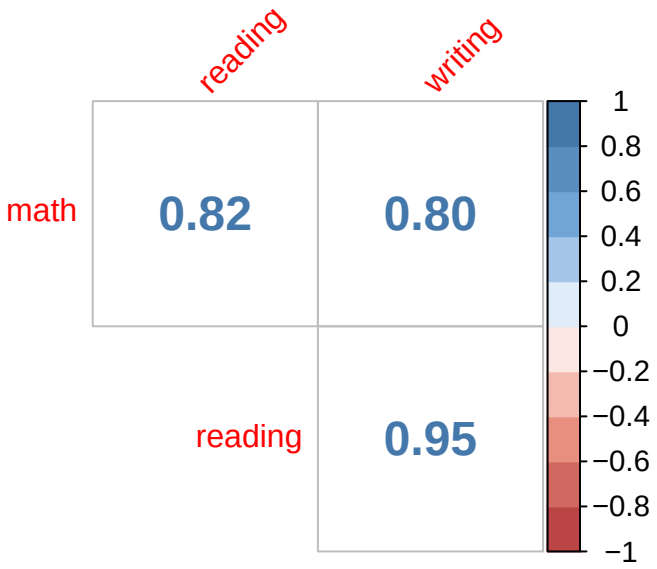
```
## [1] 5.095011
```

El valor p de la prueba es 8.41x10E^-18, que es menor que el nivel de significación alfa = 0.05. Podemos concluir que el la calificación promedio en matemáticas de los hombres es significativamente diferente de la calificación promedio de las mujeres.

```
##
##  Welch Two Sample t-test
##
## data:  hombres and mujeres
## t = 5.398, df = 997.98, p-value = 8.421e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.242813 6.947209
## sample estimates:
## mean of x mean of y
##  68.72822  63.63320
```



La matriz de correlación a continuación, nos muestra que existe entre las variables numéricas, una relación fuerte y positiva ($\mathbf{r > 0.8}$).



La prueba t asociada a la prueba de importancia del coeficiente estimado de las variables independientes, nos muestra cual es la probabilidad de observar cualquier valor igual o mayor al valor t. Un valor p pequeño indica que es probable que observemos una relación entre la variable dependiente y las independientes. El valor t del coeficiente es una medida de cuántas desviaciones estándar está lejos de 0 la estimación de nuestro coeficiente.

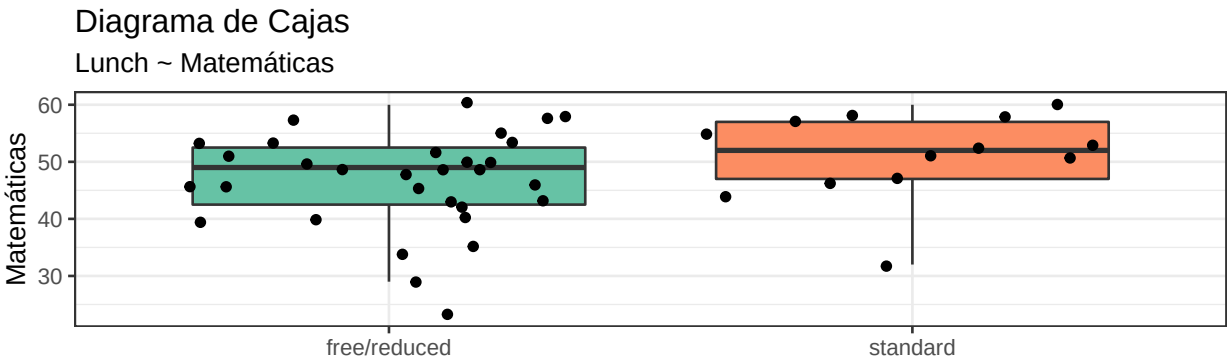
* H0: No existe relación entre la variable independiente y la dependiente
* H1: Si existe relación

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 7.524092   1.328226  5.6648 1.926e-08 ***
## reading     0.601290   0.063043  9.5378 < 2.2e-16 ***
## writing     0.249424   0.060573  4.1178 4.142e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

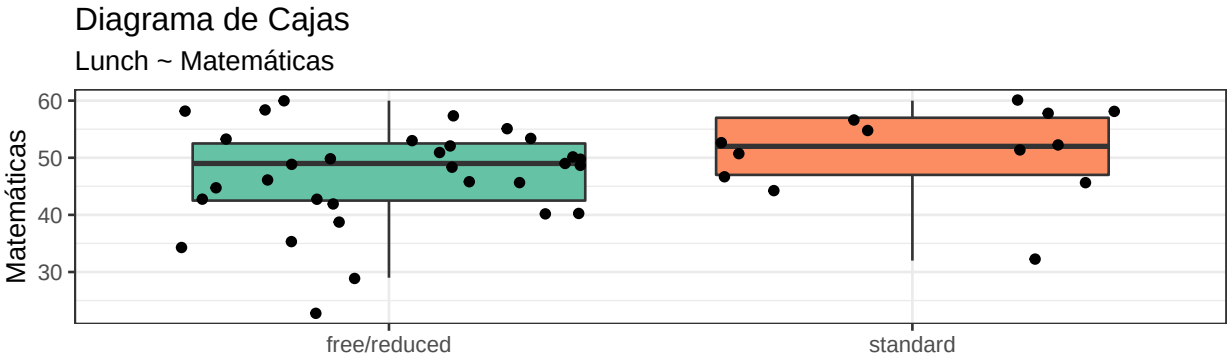
Podemos concluir que, por cada punto obtenido en lectura, obtendremos **0.6** puntos más en la nota de matemáticas y por cada punto obtenido en la nota de escritura, obtendremos **0.2** más en la nota de matemática. Los valores de p en ambas variables son cercanos a 0, es decir que las variables independientes son estadísticamente significativas para modelar el comportamiento de la nota de matemáticas.

¿Qué factores afectan más el rendimiento del estudiante? ¿Cómo?

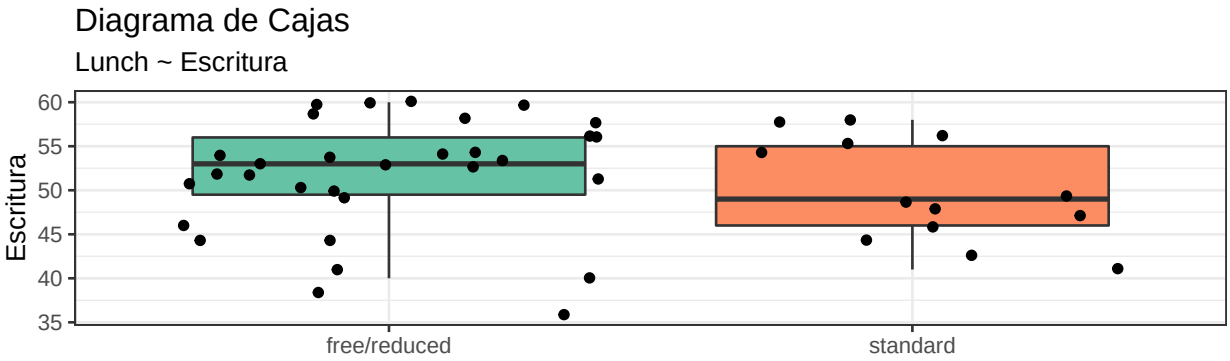
Para contestar esta pregunta, vamos a filtrar las notas obtenidas por los estudiantes en matemáticas, lectura y escritura, para aquellos que han completado el examen y han tenido notas menores a 50. Este resultado se lo puede cruzaar por el tipo de lunch que han tomado.



Analizando los gráficos de las 3 notas, en función del tipo de lunch, podemos observar que para las matemáticas los estudiantes obtuvieron menores notas para los que reciben un tipo de lunch Free/Reduced, aunque la dispersión es mayor, la mediana si sitúa 5 puntos abajo que para los que reciben un Lunch estándar. Para contestar esta pregunta, vamos a filtrar las notas obtenidas por los estudiantes en matemáticas, lectura y escritura, para aquellos que han completado el examen y han tenido notas menores a 50. Este resultado se lo puede cruzaar por el tipo de lunch que han tomado.



Para las notas de Lectura las cajas tienen aproximadamente la misma mediana, aunque la dispersión es mayor para el tipo de lunch free/reduced que para un lunch estándar.



Y por último para las notas de Escritura, los estudiantes que recibieron un tipo de lunch estándar, tienen una mediana menor que para los tuvieron un tipo de lunch/reduced, es decir que podríamos afirmar que la matemática y la escritura, es ligeramente dependiente del tipo de lunch que reciben los alumnos, porque se obtienen menores notas en matemáticas cunado toman un free/reduced, y menores notas en escritura si tomas uno estándar.