

# Análisis Inferencial FLN y ALA

Equipo de investigación

2025-10-15

## Introducción

El presente análisis busca determinar si las personas con alto Fomento a la Lectura en la Niñez (FLN) conservaron su capacidad de atención lectora (medida como minutos continuos de lectura por sesión, variable P26) durante y después de la pandemia, en comparación con quienes tuvieron bajo FLN.

Se analizan los periodos prepandemia (2019), durante la pandemia (2021) y postpandemia (2023), con el fin de identificar si el tiempo de lectura disminuyó y si el FLN actuó como un factor protector frente a dicha caída.

Se cargan los datos combinados del cuestionario 2019–2023 para sostener el análisis inferencial. El nivel de significancia adoptado en todas las estimaciones es  $\alpha = 0.04$ , equivalente a un 96% de confianza.

```
datos <- readr::read_csv(
  "nuevotec_molec_light_2019_2021_2023.csv",
  locale = readr::locale(encoding = "Latin1"),
  show_col_types = FALSE
)

datos <- datos %>%
  rename(
    P26 = p26,
    P34_1 = p34_1,
    P34_2 = p34_2,
    P34_3 = p34_3,
    P36_1 = p36_1,
    P36_2 = p36_2,
    P36_3 = p36_3,
    P36_4 = p36_4
  )

names(datos)

## [1] "P26" "P34_1" "P34_2" "P34_3" "P36_1" "P36_2" "P36_3"
## [8] "P36_4"
## [9] "factor" "year"

dim(datos)

## [1] 5924 10
```

El análisis se centra en seguir la trayectoria del tiempo de lectura (ALA = P26) antes, durante y después de la pandemia para los grupos con FLN alto y bajo. La pregunta de investigación se formula así: **¿El fomento lector temprano permitió conservar sesiones de lectura prolongadas durante todo el periodo pandémico en comparación con quienes tuvieron un FLN bajo?**

El objetivo principal es comparar los cambios temporales en la duración de las sesiones de lectura y, en particular, determinar si contar con estímulos lectores tempranos amortigua la reducción que pudo haber ocurrido durante la emergencia sanitaria. El documento actualiza la estrategia inferencial previa para incorporar el eje temporal y contrastar la interacción entre FLN y periodo.

## Construcción de variables

- **FLN (Fomento a la Lectura en la Niñez).** Se modela a partir de los reactivos que describen experiencias lectoras en el hogar y la escuela mediante un modelo de mínimos cuadrados ordinarios con P26 como variable dependiente. El puntaje ajustado (FLN\_score) resume el aporte del conjunto de ítems al tiempo de lectura actual en los tres años observados.
- **Validación alternativa.** Como verificación robusta se propone un Análisis de Componentes Principales (PCA) sobre los indicadores de FLN para confirmar la consistencia del constructo. Si los componentes convergen con el puntaje del modelo lineal, se mantiene la solución; de lo contrario, se reevalúa la ponderación.
- **ALA (Actividad Lectora Actual).** Se define como P26, la cantidad de minutos continuos de lectura durante la sesión más reciente y funciona como proxy de la atención lectora.
- **Periodo.** Se crea una variable categórica Periodo con tres niveles —“Pre-pandemia” (2019), “Durante pandemia” (2021) y “Post-pandemia” (2023)—; cada uno representa el levantamiento del cuestionario antes de la emergencia sanitaria, durante las restricciones sanitarias y tras la reapertura gradual, respectivamente.
- **Clasificación de FLN.** Se construye una variable categórica FLN\_cat con tres niveles —Bajo, Medio, Alto— usando los cuantiles 0.33 y 0.66 del FLN\_score. Esta segmentación permite contrastar diferencias de tiempo de lectura entre grupos comparables en cada periodo.

```
# puntaje continuo a partir del modelo lineal
modelo_fln <- lm(P26 ~ P34_1 + P34_2 + P34_3 + P36_1 + P36_2 + P36_3 +
P36_4, data = datos)

datos <- datos %>%
  mutate(
    FLN_score = predict(modelo_fln, newdata = datos),
    Periodo = factor(
      dplyr::case_when(
```

```

    year == 2019 ~ "Pre-pandemia",
    year == 2021 ~ "Durante pandemia",
    year == 2023 ~ "Post-pandemia",
    TRUE ~ NA_character_
  ),
  levels = c("Pre-pandemia", "Durante pandemia", "Post-pandemia")
),
FLN_cat = cut(
  FLN_score,
  breaks = quantile(FLN_score, probs = c(0, 1 / 3, 2 / 3, 1), na.rm =
TRUE),
  include.lowest = TRUE,
  labels = c("Bajo", "Medio", "Alto")
)
)

```

## Estadística descriptiva

La exploración inicial resume el comportamiento de P26 cruzando los niveles de FLN y los tres periodos (pre, durante y post pandemia) para identificar patrones temporales en la duración de las sesiones de lectura.

```

resumen_p26 <- datos %>%
  group_by(Periodo, FLN_cat) %>%
  summarise(
    n = sum(!is.na(P26)),
    media = mean(P26, na.rm = TRUE),
    sd = sd(P26, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(Periodo, FLN_cat)
resumen_p26

```

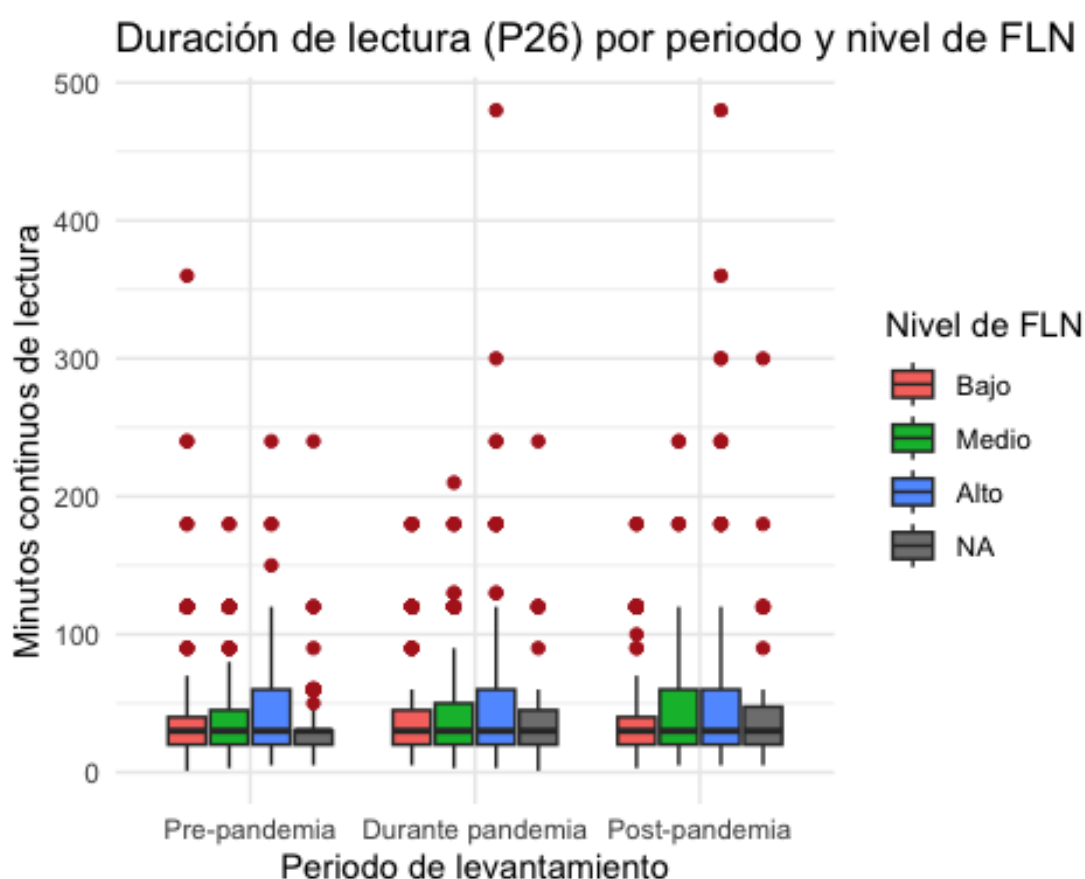
```

## # A tibble: 12 × 5
##   Periodo      FLN_cat      n media    sd
##   <fct>      <fct>    <int> <dbl> <dbl>
## 1 Pre-pandemia Bajo      431  35.6  33.3
## 2 Pre-pandemia Medio     410  37.9  27.2
## 3 Pre-pandemia Alto      467  42.4  29.2
## 4 Pre-pandemia <NA>     116  33.3  28.8
## 5 Durante pandemia Bajo     355  38.5  32.4
## 6 Durante pandemia Medio     353  39.6  32.0
## 7 Durante pandemia Alto     536  46.9  40.5
## 8 Durante pandemia <NA>     110  37.6  31.6
## 9 Post-pandemia Bajo     362  35.0  28.4
## 10 Post-pandemia Medio     354  40.2  32.7
## 11 Post-pandemia Alto     478  48.5  47.3
## 12 Post-pandemia <NA>     115  40.8  37.5

```

Esta tabla permite observar si el tiempo de lectura disminuyó del periodo pre-pandemia hacia 2021 y si se recuperó o mantuvo en 2023, resaltando si la caída es menor entre quienes contaron con un FLN alto.

```
ggplot(datos, aes(x = Periodo, y = P26, fill = FLN_cat)) +
  geom_boxplot(outlier.color = "firebrick") +
  labs(
    title = "Duración de lectura (P26) por periodo y nivel de FLN",
    x = "Periodo de levantamiento",
    y = "Minutos continuos de lectura",
    fill = "Nivel de FLN"
  ) +
  theme_minimal()
```



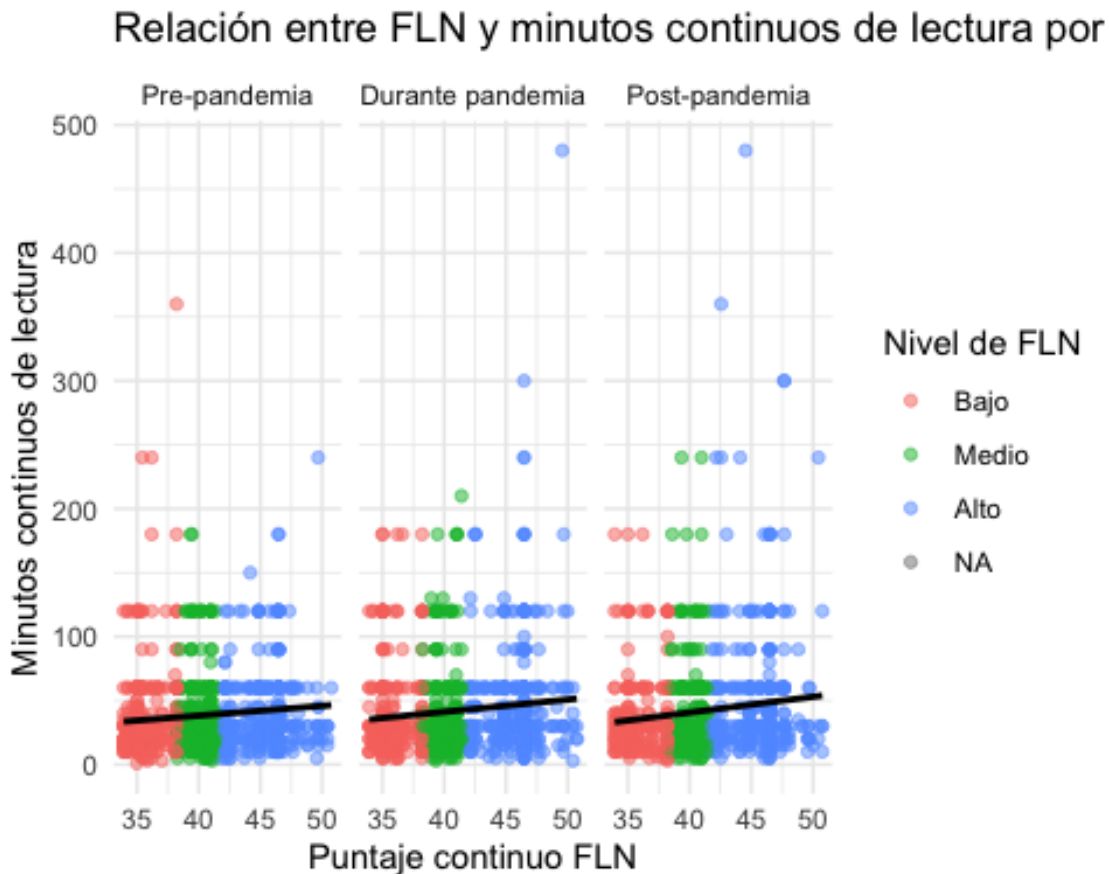
Los diagramas de cajas muestran la dispersión dentro de cada periodo y cómo cambian las medianas para FLN alto y bajo; una caída pronunciada en 2021 indicaría pérdida de retención lectora.

```
ggplot(datos, aes(x = FLN_score, y = P26, color = FLN_cat)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  facet_wrap(~Periodo) +
  labs(
```

```

title = "Relación entre FLN y minutos continuos de lectura por
periodo",
x = "Puntaje continuo FLN",
y = "Minutos continuos de lectura",
color = "Nivel de FLN"
) +
theme_minimal()

```



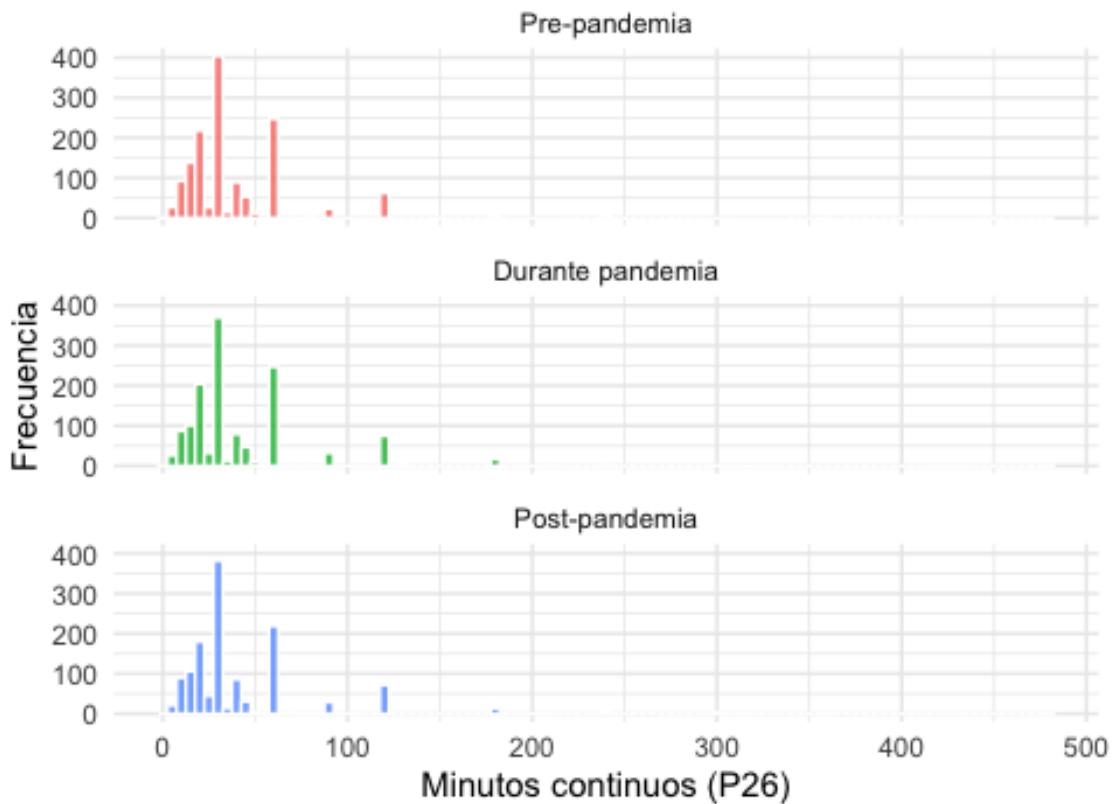
Las facetas por periodo permiten evaluar si el efecto lineal del FLN sobre la duración de lectura se mantiene estable a lo largo del tiempo o si se intensifica después de la pandemia.

```

ggplot(datos, aes(x = P26, fill = Periodo)) +
  geom_histogram(binwidth = 5, color = "white", alpha = 0.8) +
  facet_wrap(~Periodo, ncol = 1) +
  labs(
    title = "Distribución de minutos continuos de lectura por periodo",
    x = "Minutos continuos (P26)",
    y = "Frecuencia"
  ) +
  theme_minimal() +
  guides(fill = "none")

```

## Distribución de minutos continuos de lectura por period



El histograma facetado destaca si la distribución de P26 se desplazó a valores menores durante la pandemia y si hubo recuperación posterior.

### Resumen temporal FLN × Periodo

El resumen de medias por periodo y nivel de FLN sirve como punto de partida para los contrastes inferenciales.

```
resumen_p26
```

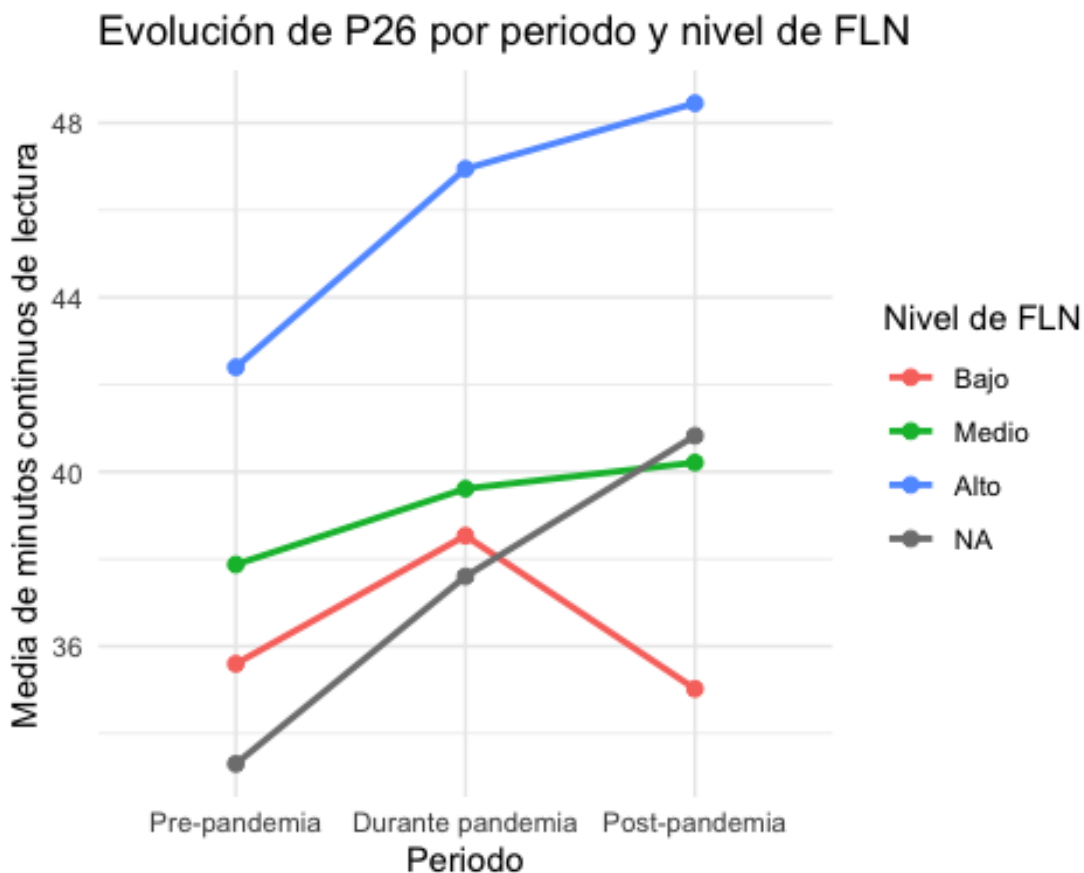
```
## # A tibble: 12 × 5
```

##	Periodo	FLN_cat	n	media	sd
##	<fct>	<fct>	<int>	<dbl>	<dbl>
##	1 Pre-pandemia	Bajo	431	35.6	33.3
##	2 Pre-pandemia	Medio	410	37.9	27.2
##	3 Pre-pandemia	Alto	467	42.4	29.2
##	4 Pre-pandemia	<NA>	116	33.3	28.8
##	5 Durante pandemia	Bajo	355	38.5	32.4
##	6 Durante pandemia	Medio	353	39.6	32.0
##	7 Durante pandemia	Alto	536	46.9	40.5
##	8 Durante pandemia	<NA>	110	37.6	31.6
##	9 Post-pandemia	Bajo	362	35.0	28.4
##	10 Post-pandemia	Medio	354	40.2	32.7

```
## 11 Post-pandemia      Alto      478  48.5  47.3
## 12 Post-pandemia      <NA>      115  40.8  37.5

lineas_perodo <- datos %>%
  group_by(Periodo, FLN_cat) %>%
  summarise(
    media = mean(P26, na.rm = TRUE),
    sd = sd(P26, na.rm = TRUE),
    n = sum(!is.na(P26)),
    .groups = "drop"
  )

ggplot(lineas_perodo, aes(x = Periodo, y = media, color = FLN_cat, group
= FLN_cat)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Evolución de P26 por periodo y nivel de FLN",
    x = "Periodo",
    y = "Media de minutos continuos de lectura",
    color = "Nivel de FLN"
  ) +
  theme_minimal()
```



La gráfica de líneas facilita interpretar la trayectoria 2019 → 2021 → 2023: una caída pronunciada en 2021 seguida de recuperación parcial en 2023 reflejaría el impacto de la pandemia, mientras que una pendiente más suave para FLN alto señalaría el efecto protector frente a la pérdida de concentración lectora.

## Intervalos de confianza ( $\alpha = 0.04$ )

Se construyen intervalos al 96% de confianza, especificando la distribución base y el criterio de interpretación.

### 1. Media de P26 (t de Student).

```
ic_media <- t.test(datos$P26, conf.level = 0.96)
ic_media

##
## One Sample t-test
##
## data:  datos$P26
## t = 74.566, df = 4086, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 96 percent confidence interval:
##  39.56036 41.80201
## sample estimates:
## mean of x
##  40.68118
```

La distribución *t* se utiliza porque la desviación estándar poblacional es desconocida; el intervalo sugiere el rango plausible de la media poblacional de minutos continuos considerando todas las mediciones de 2019 a 2023.

### 2. Diferencia de medias (FLN Alto vs Bajo, t de Welch).

```
p26_alto <- datos %>% filter(FLN_cat == "Alto") %>% pull(P26)
p26_bajo <- datos %>% filter(FLN_cat == "Bajo") %>% pull(P26)
ic_diferencia <- t.test(p26_alto, p26_bajo, conf.level = 0.96, var.equal = FALSE)
ic_diferencia

##
## Welch Two Sample t-test
##
## data:  p26_alto and p26_bajo
## t = 6.9485, df = 2625.8, p-value = 4.638e-12
## alternative hypothesis: true difference in means is not equal to 0
## 96 percent confidence interval:
##  6.813697 12.535519
## sample estimates:
## mean of x mean of y
##  45.99865  36.32404
```



Se emplea la versión de Welch al no asumir varianzas iguales. El intervalo indica si la brecha de minutos entre FLN alto y bajo es estadísticamente relevante a lo largo del periodo pandémico y posterior.

### 3. Varianza de P26 ( $\chi^2$ ).

```
n_total <- sum(!is.na(datos$P26))
s2 <- var(datos$P26, na.rm = TRUE)
chi_left <- qchisq(0.98, df = n_total - 1)
chi_right <- qchisq(0.02, df = n_total - 1)
ic_varianza <- c((n_total - 1) * s2 / chi_left, (n_total - 1) * s2 /
chi_right)
ic_varianza

## [1] 1163.035 1273.691
```

El intervalo basado en  $\chi^2$  estima la variabilidad poblacional del tiempo de lectura durante el trienio analizado, útil para contrastar si la dispersión aumentó en los años de pandemia.

### 4. Proporción de lectura corta (< 20 minutos).

```
lectura_corta <- mean(datos$P26 < 20, na.rm = TRUE)
n_corta <- sum(!is.na(datos$P26))
ic_prop <- prop.test(
  x = sum(datos$P26 < 20, na.rm = TRUE),
  n = n_corta,
  conf.level = 0.96
)
ic_prop

##
## 1-sample proportions test with continuity correction
##
## data:  sum(datos$P26 < 20, na.rm = TRUE) out of n_corta, null
probability 0.5
## X-squared = 1789, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 96 percent confidence interval:
##  0.1572554 0.1815792
## sample estimates:
##           p
## 0.1690727
```

La aproximación normal (con corrección de continuidad) estima la proporción poblacional que mantiene sesiones breves durante y después de la pandemia; se puede ajustar el umbral según la definición institucional de lectura corta.

## Pruebas de hipótesis ( $\alpha = 0.04$ )

Cada prueba reporta hipótesis nulas y alternativas, estadístico de prueba, valor crítico y p-value, complementando con conclusiones directas.

## ANOVA: P26 ~ FLN\_cat \* Periodo

- $H_0$ : la media de P26 es igual entre todos los niveles de FLN y entre los tres periodos, sin interacción.
- $H_1$ : al menos un efecto principal o la interacción FLN  $\times$  Periodo modifica la media de P26.

```
modelo_anova <- aov(P26 ~ FLN_cat * Periodo, data = datos)
anova_summary <- summary(modelo_anova)
anova_summary

##              Df Sum Sq Mean Sq F value    Pr(>F)
## FLN_cat        2   65871    32936  27.236 1.81e-12 ***
## Periodo        2    7556     3778   3.124  0.0441 *
## FLN_cat:Periodo  4    5578     1394   1.153  0.3296
## Residuals     3737 4518971     1209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2178 observations deleted due to missingness

df_fln <- anova_summary[[1]]["FLN_cat", "Df"]
df_periodo <- anova_summary[[1]]["Periodo", "Df"]
df_interaccion <- anova_summary[[1]]["FLN_cat:Periodo", "Df"]
df_res <- anova_summary[[1]]["Residuals", "Df"]

fcrit_fln <- qf(1 - 0.04, df_fln, df_res)
fcrit_periodo <- qf(1 - 0.04, df_periodo, df_res)
fcrit_interaccion <- qf(1 - 0.04, df_interaccion, df_res)
c(
  Fcrit_FLN = fcrit_fln,
  Fcrit_Periodo = fcrit_periodo,
  Fcrit_Interaccion = fcrit_interaccion
)

##          Fcrit_FLN      Fcrit_Periodo Fcrit_Interaccion
##              NA          3.221650         2.509073
```

Los valores críticos se obtienen de la distribución  $F$  correspondiente a cada efecto (FLN, Periodo e interacción). Se rechaza  $H_0$  a  $\alpha = 0.04$  cuando el estadístico  $F$  supera el valor crítico o el  $p$ -value es menor al nivel de significancia, lo que indicaría que el FLN alto amortiguó la caída de minutos durante y después de la pandemia.

## t de Welch: P26 (FLN Alto vs Bajo)

- $H_0$ : las medias de P26 son iguales entre los niveles alto y bajo de FLN.
- $H_1$ : las medias de P26 difieren entre ambos niveles.

```
prueba_welch <- t.test(p26_alto, p26_bajo, alternative = "two.sided",
  conf.level = 0.96)
prueba_welch
```

```
##
## Welch Two Sample t-test
##
## data: p26_alto and p26_bajo
## t = 6.9485, df = 2625.8, p-value = 4.638e-12
## alternative hypothesis: true difference in means is not equal to 0
## 96 percent confidence interval:
##  6.813697 12.535519
## sample estimates:
## mean of x mean of y
## 45.99865 36.32404

df_welch <- prueba_welch$parameter
tcrit_welch <- qt(1 - 0.04 / 2, df = df_welch)
tcrit_welch

## [1] 2.05477
```

El estadístico  $t$  se contrasta con el valor crítico bilateral a  $\alpha = 0.04$  ( $tcrit\_welch$ ). Un p-value menor que el nivel de significancia sugiere que el FLN alto funcionó como un factor protector al sostener sesiones de lectura más extensas durante y después de la pandemia.

### $\chi^2$ de independencia: FLN\_cat vs lectura corta

- $H_0$ : FLN y lectura corta (< 20 minutos) son independientes.
- $H_1$ : existe asociación entre FLN y lectura corta.

```
datos <- datos %>%
  mutate(lectura_corta = factor(ifelse(P26 < 20, "Sí", "No")))

tabla_cont <- table(datos$FLN_cat, datos$lectura_corta)
prueba_chi <- chisq.test(tabla_cont, correct = FALSE)
prueba_chi

##
## Pearson's Chi-squared test
##
## data: tabla_cont
## X-squared = 34.43, df = 2, p-value = 3.339e-08

chi_critico <- qchisq(1 - 0.04, df = prueba_chi$parameter)
chi_critico

## [1] 6.437752
```

El estadístico  $\chi^2$  se contrasta con el valor crítico  $chi\_critico$ , calculado para  $(k - 1)(m - 1)$  grados de libertad. Un resultado significativo implicaría que los niveles de FLN modifican la probabilidad de mantener sesiones cortas durante y después de la pandemia.

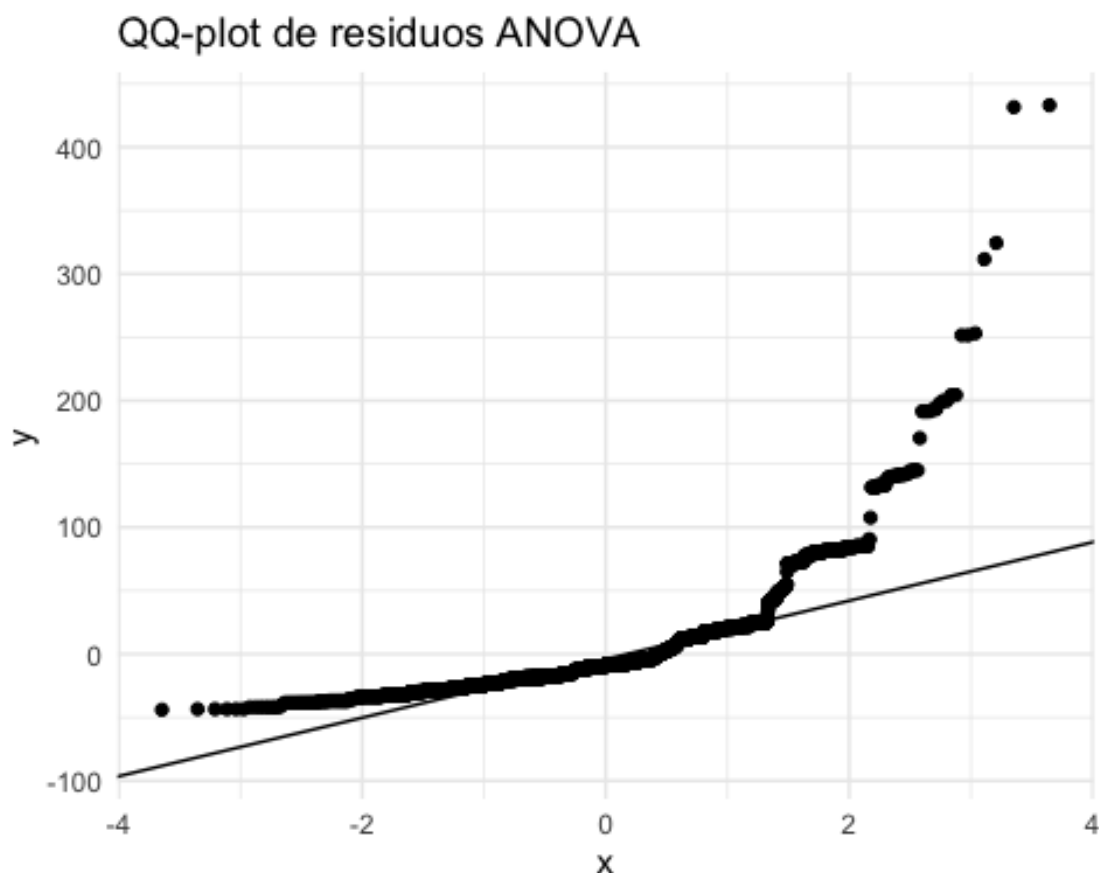
## Verificación de supuestos

1. **Normalidad.** Evaluada con Shapiro-Wilk o gráficos QQ en los residuos del modelo ANOVA factorial (FLN\_cat × Período) y en las distribuciones de P26 dentro de cada combinación.

```
residuos_anova <- residuals(modelo_anova)
shapiro.test(residuos_anova)

##
##  Shapiro-Wilk normality test
##
## data:  residuos_anova
## W = 0.71703, p-value < 2.2e-16

qqplot_anova <- ggplot(data.frame(residuos = residuos_anova), aes(sample
= residuos)) +
  stat_qq() + stat_qq_line() + theme_minimal() +
  labs(title = "QQ-plot de residuos ANOVA")
qqplot_anova
```



Si la normalidad no se cumple, se recomienda utilizar pruebas no paramétricas como Kruskal-Wallis o transformaciones logarítmicas.

2. **Homogeneidad de varianzas.** Contrastada con la prueba de Levene considerando los tres niveles de FLN y los tres periodos.

```
leveneTest(P26 ~ interaction(FLN_cat, Periodo), data = datos)

## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group      8  4.8046 6.707e-06 ***
##           3737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ante varianzas heterogéneas, se mantienen estimadores robustos (ANOVA de Welch) o se recurre a comparaciones no paramétricas (p. ej., prueba de Brown–Forsythe).

3. **Conteos esperados en  $\chi^2$ .** Se verifican que todos sean superiores a 5 para garantizar la validez asintótica al comparar FLN con lectura corta.

```
prueba_chi$expected

##
##           No      Sí
## Bajo  957.0753 190.9247
## Medio 931.2309 185.7691
## Alto 1234.6938 246.3062
```

Si los conteos esperados son bajos, se sugiere combinar categorías o emplear un test exacto de Fisher.

## Conclusiones

Los resultados permiten observar la evolución del tiempo de lectura durante la pandemia. Si bien el promedio general puede haberse reducido, el grupo con alto FLN muestra una caída menor, lo que sugiere un efecto protector del fomento lector temprano frente a la pérdida de concentración lectora.