Análisis Inferencial FLN y ALA

Equipo de investigación

2025-10-15

Introducción

Se cargan los datos combinados del cuestionario 2019–2023.

```
datos <- readr::read_csv(
  "nuevote_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"
"P36_4"
##  [9] "factor" "year"
```

```
dim(datos)
```

```
## [1] 5924   10
```

El análisis se centra en la relación entre el fomento a la lectura en la niñez (FLN) y la actividad lectora actual (ALA), entendida como los minutos continuos que una persona dedica a leer en una sesión (P26). La pregunta de investigación se formula así: **¿El FLN contribuye a mantener sesiones de lectura más prolongadas en la adultez?**

El objetivo principal es evaluar si las personas leen menos tiempo seguido y, en particular, determinar si contar con estímulos lectores tempranos amortigua esa reducción. El documento expone una estrategia inferencial para contrastar la influencia del FLN en la retención lectora.

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.
- **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.
- **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.

```
# 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),
    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

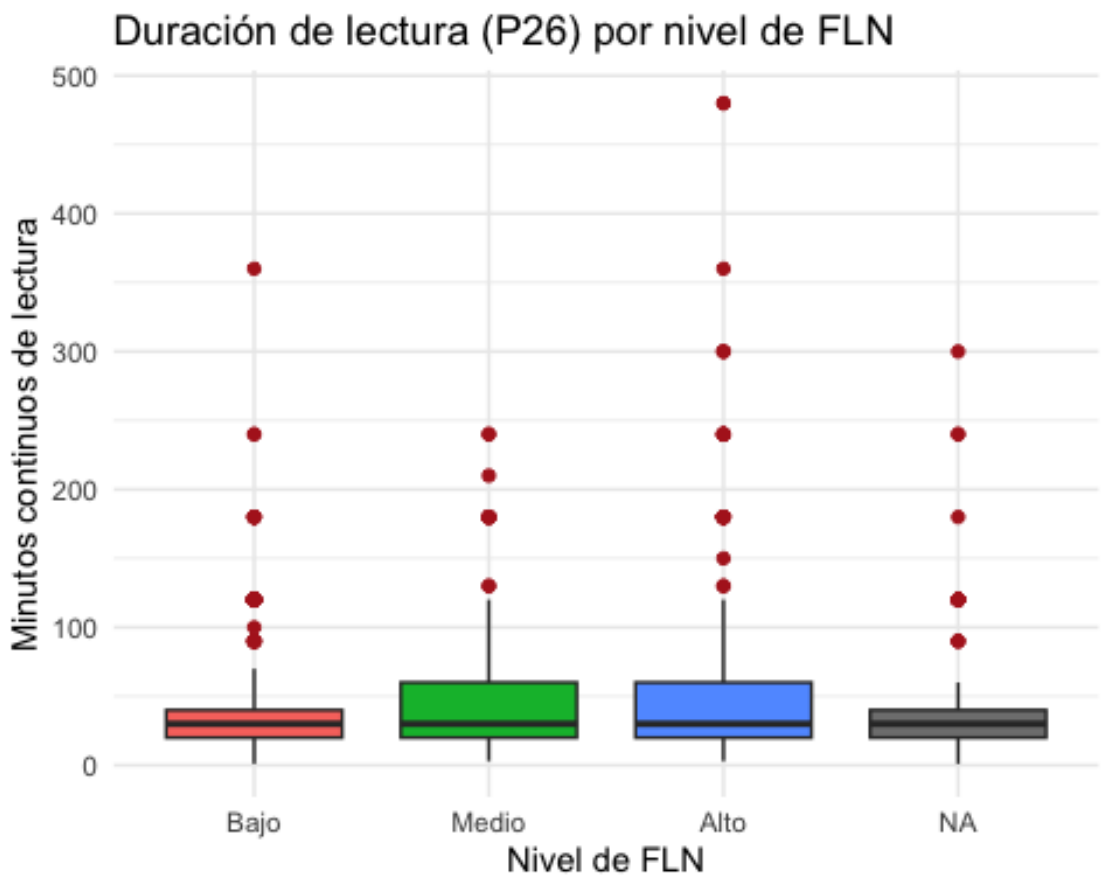
La exploración inicial resume el comportamiento de P26 por nivel de FLN y visualiza la relación entre ambas variables.

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

```
## # A tibble: 4 × 4
##   FLN_cat     n media    sd
##   <fct>   <int> <dbl> <dbl>
## 1 Bajo     1148  36.3  31.5
## 2 Medio    1117  39.2  30.6
## 3 Alto     1481  46.0  39.8
## 4 <NA>      341  37.2  32.9
```

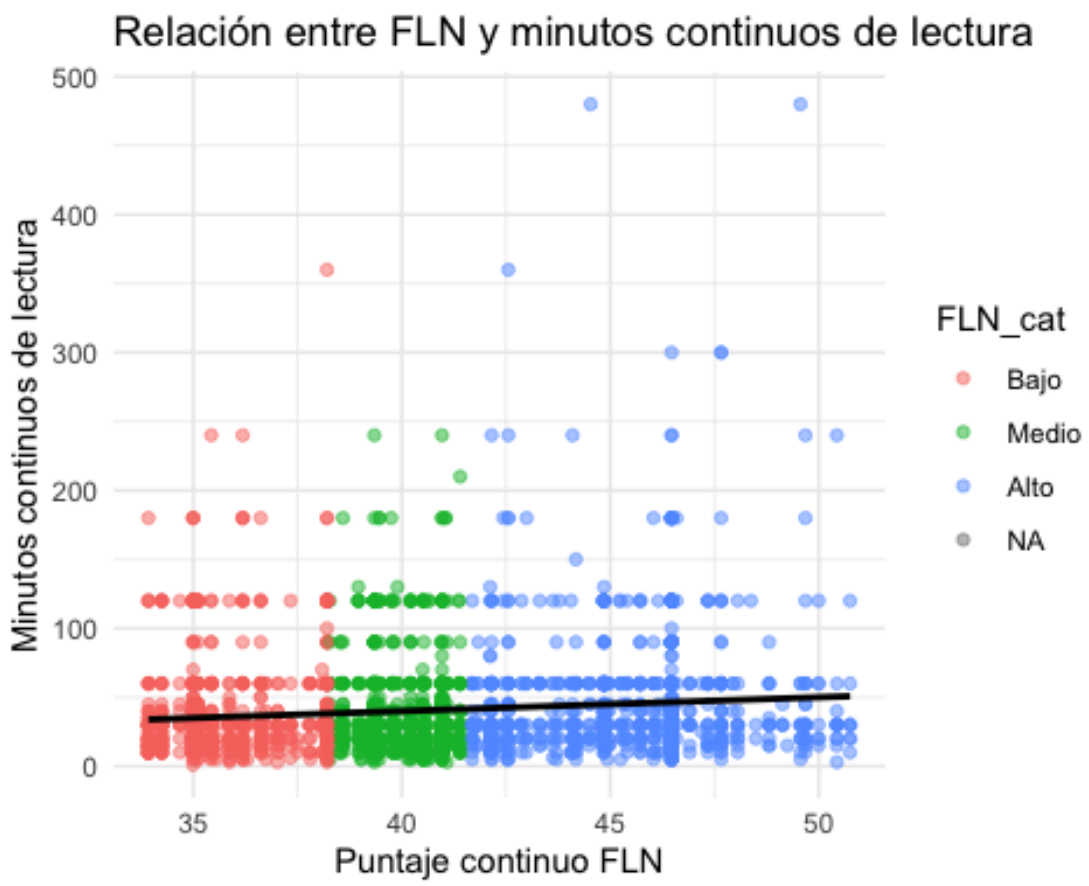
El nivel de FLN se asocia con cambios observables en la media y la dispersión del tiempo de lectura.

```
ggplot(datos, aes(x = FLN_cat, y = P26, fill = FLN_cat)) +
  geom_boxplot(outlier.color = "firebrick") +
  labs(
    title = "Duración de lectura (P26) por nivel de FLN",
    x = "Nivel de FLN",
    y = "Minutos continuos de lectura"
  ) +
  theme_minimal() +
  guides(fill = "none")
```



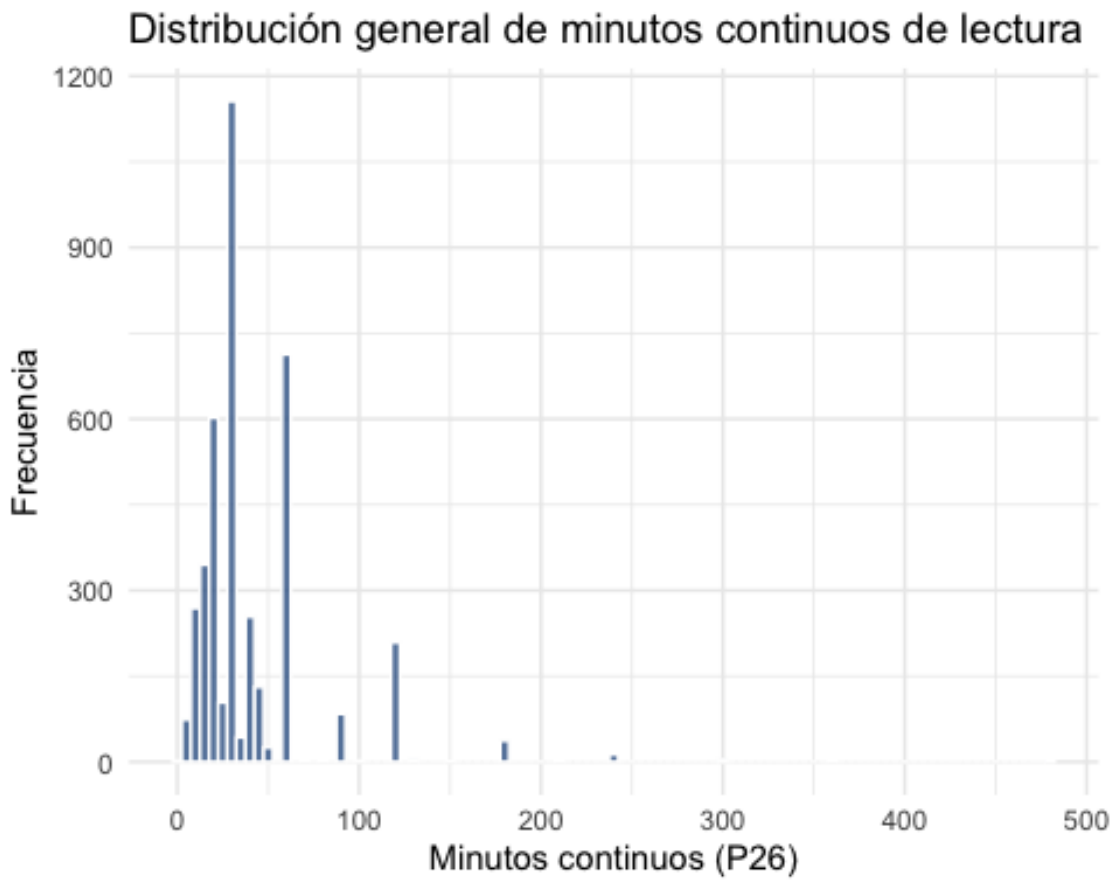
Los percentiles y la mediana sugieren si los niveles altos de FLN concentran tiempos de lectura más extensos o si existen asimetrías significativas en cada estrato.

```r
ggplot(datos, aes(x = FLN_score, y = P26, color = FLN_cat)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(
    title = "Relación entre FLN y minutos continuos de lectura",
    x = "Puntaje continuo FLN",
    y = "Minutos continuos de lectura"
  ) +
  theme_minimal()
```



La recta de tendencia permite identificar si el FLN ejerce un efecto lineal positivo sobre el tiempo de lectura, así como la magnitud de la variabilidad residual.

```r
ggplot(datos, aes(x = P26)) +
  geom_histogram(binwidth = 5, fill = "#366092", color = "white", alpha =
0.8) +
  labs(
    title = "Distribución general de minutos continuos de lectura",
    x = "Minutos continuos (P26)",
    y = "Frecuencia"
  ) +
  theme_minimal()
```



El histograma muestra si la distribución de P26 es unimodal, sesgada o multimodal, información clave para validar los supuestos del análisis inferencial.

Intervalos de confianza (α = 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.

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.

3. **Varianza de P26 (χ^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 χ^2 estima la variabilidad poblacional del tiempo de lectura, útil para contrastar escenarios de dispersión creciente.

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; se puede ajustar el umbral según la definición institucional de lectura corta.

Pruebas de hipótesis (α = 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`

- H_0: la media de P26 es igual en los tres niveles de FLN.
- H_1: al menos un nivel muestra una media distinta.

```
modelo_anova <- aov(P26 ~ FLN_cat, data = datos)
glance(modelo_anova)
```

```
## # A tibble: 1 × 6
##    logLik    AIC    BIC deviance  nobs r.squared
##     <dbl>  <dbl>  <dbl>    <dbl> <int>     <dbl>
## 1 -18610. 37229. 37254. 4532104.  3746    0.0143
```

El valor crítico proviene de la distribución *F* con gl_1 = k − 1 y gl_2 = n − k. Si el estadístico F supera el valor crítico al α = 0.04 o el p-value < 0.04, se rechaza H_0 y se concluye que el FLN influye en la duración de lectura. En caso contrario, no se evidencia diferencia significativa.

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
```

El estadístico *t* se compara con el valor crítico bilateral a α = 0.04. Un p-value menor que el nivel de significancia sugiere que la exposición alta al FLN se asocia con sesiones de lectura más extensas.

χ^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
```

El estadístico χ^2 se contrasta con el valor crítico correspondiente a `(k - 1)(m - 1)` grados de libertad. Un resultado significativo implica que los niveles de FLN influyen en la probabilidad de mantener sesiones cortas.

Verificación de supuestos

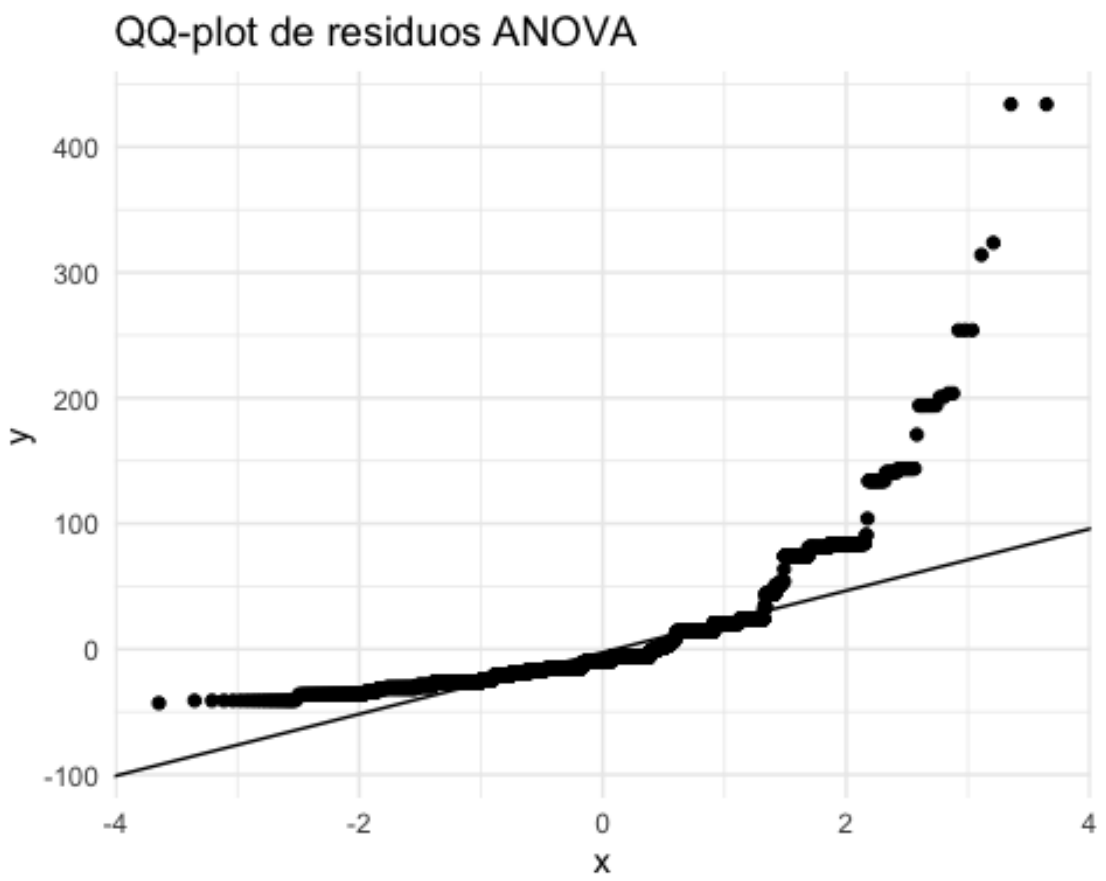
1. **Normalidad.** Evaluada con Shapiro-Wilk o gráficos QQ en los residuos del modelo ANOVA y en las distribuciones de `P26` por nivel de FLN.

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

##
##  Shapiro-Wilk normality test
##
## data:  residuos_anova
## W = 0.71053, 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.

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

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    2  13.149 2.039e-06 ***
##       3743
## ---
## 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 χ^2.** Se verifican que todos sean superiores a 5 para garantizar la validez asintótica.

```
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

El análisis inferencial propuesto permite determinar si el tiempo de lectura ha disminuido y si el FLN funciona como un factor protector frente a la reducción de minutos continuos. Dependiendo de los resultados:

- Si la media de P26 ha caído y la brecha entre FLN alto y bajo es significativa, se concluye que el fomento lector temprano ayuda a sostener sesiones más largas.
- Las pruebas de hipótesis y los intervalos ofrecen evidencia cuantitativa para orientar programas de intervención centrados en los hogares y escuelas.

Entre las limitaciones destacan el uso de variables auto-reportadas y la posible colinealidad entre indicadores de FLN. Los próximos pasos incluyen incorporar otras métricas de actividad lectora (como número de libros leídos, P4) y explorar modelos longitudinales que permitan seguir cohortes a lo largo del tiempo.