Statistics in R – Modus Paraguayensis

Explorando la Estadística desde la Perspectiva de R

Trabajo #2

"Statistics in R – Modus Paraguayensis" consiste en una serie de cuadernos técnicos que condensan un esfuerzo sistemático por digitalizar y potenciar el análisis estadístico tradicional. Propone abordar una metodología de migración desde el tratamiento estadístico convencional hacia el lenguaje de programación R, priorizando la reproducibilidad científica y la interpretación de los datos en el contexto de la economía y las ciencias sociales.

Autor: Mag. Econ. Marcos Gómez Hermosa
Facultad de Ciencias Económicas (FACE)
Universidad Católica Campus Itapúa (UCI)
Encarnación, Paraguay – 2026

Nota Editorial

Esta serie de cuadernos está concebida exclusivamente con fines pedagógicos. El objetivo apunta al desarrollo de competencias en la metodología de trabajo con **R** y la interpretación básica de datos aplicados a la economía y las ciencias sociales.

Las bases de datos fueron construidas bajo simulación, de modo a colaborar con la progresión del proceso analítico desde la aplicación de los métodos estadísticos hacia su interpretación en contexto, buscando establecer una conexión entre la teoría de los manuales clásicos y la práctica programada.

Este material está dirigido a estudiantes de cursos básicos e iniciales de estadística que buscan integrar la programación en **R** con las nociones fundamentales de estadística descriptiva e inferencial. El material se enfoca en la interpretación funcional de los resultados, delegando las discusiones sobre la profundidad del rigor técnico o validaciones metodológicas avanzadas para instancias posteriores de formación o literatura académica especializada.

Análisis de un índice para la Articulación Académico-Profesional de estudiantes en situación de empleo

Un ejemplo de aplicación de análisis descriptivo estadístico de datos ***no agrupados***

Contents

1. Introducción

En el Trabajo #1 (disponible en el repositorio Github: https://github.com/elprofemarcosgh/Statistics-in-R-Modus-Paraguayensis/blob/main/trabajo%231/trabajo1_statistics_in_r_modus_paraguayensis_para_pdf.pdf), se desarrolló la estructura clásica de la estadística descriptiva para datos agrupados, priorizando la ejecución procedimental ante la discusión de su validez metodológica. Si bien en la actualidad el agrupamiento de datos puede considerarse una técnica "analíticamente obsoleta" frente a la capacidad de procesamiento de la computación moderna, su dominio sigue siendo menester para el analista en formación. Comprender la mecánica subyacente a la agrupación y el cálculo de sus estadísticos es fundamental para desarrollar una lectura estadística básica antes de estudiar la precisión del algoritmo.

En este segundo cuaderno, el Trabajo #2, se explorará de vuelta el constructo hipotético denominado "IAAP" (Índice de Articulación Académica-Laboral), que consiste en un intento de síntesis mediante la reducción de datos, a partir de la técnica de Análisis de Componentes Principales (PCA), de 8 variables originales (descriptas en el Trabajo #1) con el fin de otorgar una puntuación que "caracterizaría" la situación de los estudiantes en empleo (entiéndase esto como estudiantes que cursan y trabajan al mismo tiempo). La lógica del índice sigue siendo la misma: se genera a partir de la raíz de la suma ponderada de los cuadrados de los valores de los vectores; bajo esta premisa, a mayor puntuación, se asume una mejor articulación académica-laboral del estudiante.

El lector familiarizado con el Análisis de Componentes Principales (PCA) o métodos de validación de escalas por supuesto tendrá sus válidas objeciones respecto al tratamiento y la robustez del constructo. No obstante, el propósito de este análisis es centrar la atención en el cálculo de estadísticos para datos no agrupados.

La motivación de este material es acompañar a los estudiantes en su transición de la estadística convencional hacia la computacional. El objetivo final es contrastar ambos enfoques, analizar la divergencia metodológica y estudiar las causas subyacentes a las discrepancias entre ambos métodos: el cálculo de estadísticos por datos agrupados y los no agrupados.

```
library(readr)
# link "Raw" al archivo de datos
url_raw <- paste0("https://raw.githubusercontent.com/",
                  "elprofemarcosgh/Statistics-in-R-Modus-Paraguayensis/",
                  "main/trabajo_2/datos_para_tabla_est.csv")

df_indices <- read_csv(url_raw)

df_indices <- df_indices %>%
  mutate(IAAP = sqrt(Indice_Capacidad_General^2 + Indice_Especializacion^2))
# head() para mostrar solo los primeros 10 registros
# select() para organizar las columnas de forma lógica para el lector
library(dplyr)
vista_previa <- df_indices %>%
  select(Indice_Capacidad_General, Indice_Especializacion, IAAP) %>%
  head(10)

knitr::kable(vista_previa,
             caption = "Muestra de los primeros 10 registros de la base de datos
              ↪  (Variables clave e Índices)",
             digits = 2)
```

Table 1: Muestra de los primeros 10 registros de la base de datos (Variables clave e Índices)

Indice_Capacidad_General	Indice_Especializacion	IAAP
-1.03	0.25	1.06
-0.14	1.24	1.25
-0.25	-0.48	0.54
-1.34	-0.01	1.34
-1.36	1.27	1.86
1.04	-0.02	1.04
-1.56	0.20	1.57
-0.58	0.31	0.66
-0.15	-1.35	1.36
-2.07	0.02	2.07

Siguiendo lo expuesto en el Trabajo #1, los vectores de Capacidad General y Especialización corresponden a puntuaciones factoriales, centradas en cero, donde los signos indican la posición relativa del estudiante respecto a la media de la muestra, y la magnitud representa la distancia en desviaciones estándar. Un valor negativo en estos vectores se interpreta como una posición por debajo de la media del grupo. Por el contrario, los valores positivos indican un desempeño o una especialización superior al promedio de sus pares. La magnitud del número indica qué tan alejado se encuentra el estudiante de ese centro "gravitacional" estadístico.

Así, el Índice de Articulación Académica Profesional (IAAP) sintetiza ambos vectores en una sola métrica ponderada. El IAAP intenta medir la intensidad y la armonía con la que el estudiante integra su formación académica (Capacidad General) con su desempeño laboral (Especialización). Un IAAP elevado sugeriría una sinergia entre ambas esferas. El IAAP trata de estudiar qué tan alineados están los dos pilares del desarrollo de los estudiantes en situación de empleo, el laboral y el académico.

2. Disponibilidad de Código y Datos

El código fuente para este análisis, incluyendo los scripts de procesamiento en R, las funciones personalizadas y la base de datos utilizada, está disponible en el repositorio oficial de "Statistics in R - Modus Paraguayensis":

https://github.com/elprofemarcosgh/Statistics-in-R-Modus-Paraguayensis

Nota: Los materiales específicos de este capítulo se encuentran en la carpeta /trabajo_2.

3. Medidas de tendencia central

Media Aritmética

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$\bar{x}$: Media aritmética del índice, x_i: Valor individual del IAAP para cada estudiante, n: Tamaño de la muestra, $\sum$: Sumatoria de todos los valores observados.

Media Geométrica

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln(x_i)\right)$$

$\bar{x}_G$: Media geométrica del índice, $\prod$: Productoria (multiplicación de todos los valores), ln: Logaritmo natural de cada valor individual, n: Tamaño de la muestra, exp: Función exponencial para revertir el logaritmo y obtener el promedio final.

Mediana

$$\tilde{x} = x_{(\frac{n+1}{2})} \text{ (para } n \text{ impar)}$$

$\tilde{x}$: Mediana de la distribución, $x_{(\frac{n+1}{2})}$: El valor que ocupa la posición tras ordenar los datos de forma ascendente, n: Tamaño de la muestra.

Moda

$$Mo = \text{valor con mayor frecuencia absoluta } (n_i)$$

Mo: Moda de la variable IAAP, n_i: Cantidad de veces que se repite un valor específico dentro del conjunto de datos.

```
library(dplyr)
# Calcula las 3 medidas principales para el IAAP
media_arit_iaap   <- mean(df_indices$IAAP, na.rm = TRUE)
mediana_iaap <- median(df_indices$IAAP, na.rm = TRUE)

# Para la moda usa la librería modeest
# mfv devuelve un vector (por si hay más de una moda)
moda_iaap    <- mfv(df_indices$IAAP)

# Media Geométrica. Se calcula como la exponencial de la media de los logaritmos
media_geom_iaap <- exp(mean(log(df_indices$IAAP[df_indices$IAAP > 0]), na.rm = TRUE))

# Se crea una tabla simple para mostrar los resultados
tabla_tendencia <- data.frame(
  Medida = c("Media Aritmética","Media Geométrica", "Mediana", "Moda"),
  Valor = c(media_arit_iaap, media_geom_iaap, mediana_iaap, moda_iaap[1])
)

knitr::kable(tabla_tendencia,
             caption = "Medidas de Tendencia Central para el IAAP. Datos no agrupados",
             digits = 4)
```

Table 2: Medidas de Tendencia Central para el IAAP. Datos no agrupados

Medida	Valor
Media Aritmética	1.2761
Media Geométrica	1.1020
Mediana	1.2854
Moda	0.0536

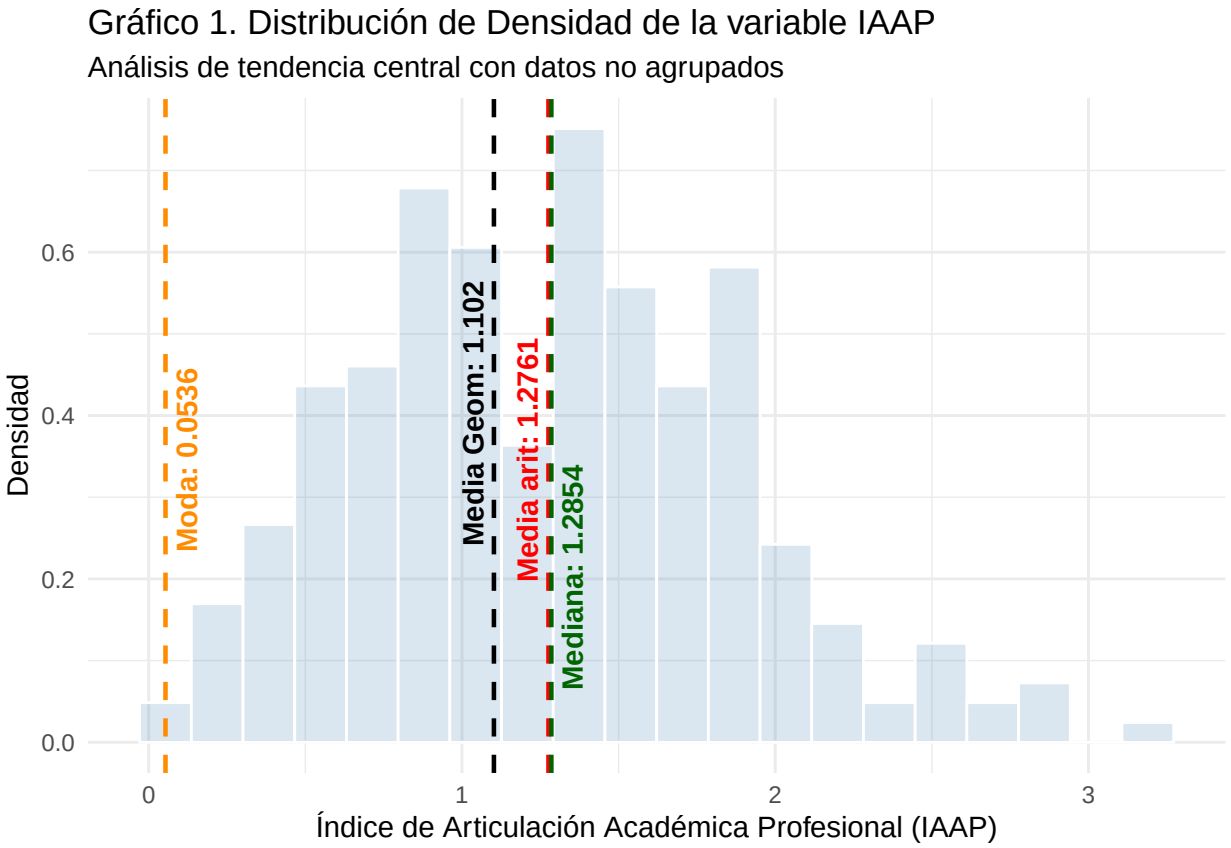
El análisis de los estadísticos de tendencia central revela una dinámica para el índice de articulación académico-profesional de los estudiantes. La Media Aritmética, situada en 1.2761 sugiere un nivel de “IAAP”

aceptable en términos generales. Sin embargo, al contrastarla con la Mediana de 1.2854, el promedio está siendo traccionado hacia la izquierda por valores inferiores, lo que le resta capacidad para representar fielmente al estudiante típico. Al observar la Media Geométrica de 1.1020, cuyo valor es sensiblemente menor, se advierte sobre la existencia de disparidades y una volatilidad interna en los componentes del índice IAAP. El dato más disruptivo lo aporta la Moda de 0.0536, que señala que el escenario más frecuente entre los sujetos analizados es una integración casi nula, se podría interpretar como estudiantes que no logran conciliar sus esferas académica y laboral a pesar de lo que sugieren los promedios globales.

Esta jerarquía numérica, donde la Moda es sensiblemente inferior a la Media y esta a su vez es superada por la Mediana, permite anticipar (solo anticipar, no asegurar aún) una distribución con asimetría negativa o sesgada a la izquierda. El gráfico de densidad muestra una curva que no se comporta como una campana de Gauss tradicional, sino que presenta una silueta bimodal con una acumulación de densidad en el extremo inferior. Esta relación entre los estadísticos anticipa una forma de distribución donde conviven dos grupos distintos: un grupo mayoritario con niveles de articulación medios-altos y un segmento que genera esa "cola" hacia los valores bajos.

```
# Se define la altura máxima de la densidad para ubicar las etiquetas
densidad_max <- max(density(df_indices$IAAP)$y)

ggplot(df_indices, aes(x = IAAP)) +
  # 1. El Histograma de fondo (suave)
  geom_histogram(aes(y = ..density..), bins = 20, fill = "steelblue", alpha = 0.2, color
   ↪  = "white") +

  # . Líneas de Tendencia Central
  geom_vline(xintercept = media_arit_iaap, color = "red", linetype = "dashed", size =
   ↪  0.8) +
  geom_vline(xintercept = mediana_iaap, color = "darkgreen", linetype = "dashed", size =
   ↪  0.8) +
  geom_vline(xintercept = media_geom_iaap, color = "black", linetype = "dashed", size =
   ↪  0.8) +
  geom_vline(xintercept = moda_iaap[1], color = "darkorange", linetype = "dashed", size =
   ↪  0.8) +

  #3 . Etiquetas (Ajustadas a la altura de la densidad)
  annotate("text", x = media_arit_iaap, y = densidad_max * 0.6,
           label = paste("Media arit:", round(media_arit_iaap, 4)),
           color = "red", angle = 90, vjust = -0.5, fontface = "bold") +

  annotate("text", x = mediana_iaap, y = densidad_max * 0.35,
           label = paste("Mediana:", round(mediana_iaap, 4)),
           color = "darkgreen", angle = 90, vjust = 1.5, fontface = "bold") +

  annotate("text", x = media_geom_iaap, y = densidad_max * 0.7,
           label = paste("Media Geom:", round(media_geom_iaap, 4)),
           color = "black", angle = 90, vjust = -0.5, fontface = "bold") +

  annotate("text", x = moda_iaap[1], y = densidad_max * 0.6,
           label = paste("Moda:", round(moda_iaap[1], 4)),
           color = "darkorange", angle = 90, vjust = 1.5, fontface = "bold") +

  # Estética y límites
  labs(title = "Gráfico 1. Distribución de Densidad de la variable IAAP",
       subtitle = "Análisis de tendencia central con datos no agrupados",
       x = "Índice de Articulación Académica Profesional (IAAP)",
```

```
      y = "Densidad") +
  theme_minimal()
```



4. Medidas de posicionamiento

Cuartiles (Q_k)

Medida estadística de posición que divide un conjunto de datos ordenados en cuatro partes iguales. Cada parte representa el 25% de la muestra total.

$$Q_k = x_{(i)} \quad \text{donde} \quad i = \frac{k(n+1)}{4}$$

Q_k: Valor del cuartil buscado ($k = 1, 2, 3$), k: Número del cuartil (25%, 50% o 75% de la distribución), n: Tamaño de la muestra, $x_{(i)}$: Valor del IAAP en la posición "i".

Deciles (D_k)

Medida estadística de posición que divide un conjunto de datos ordenados en diez partes iguales. Cada parte representa el 10% de la muestra total.

$$D_k = x_{(i)} \quad \text{donde} \quad i = \frac{k(n+1)}{10}$$

D_k: Valor del decil buscado ($k = 1, 2, \ldots, 9$), k: Número del decil (desde el 10% hasta el 90%), n: Tamaño de la muestra, $x_{(i)}$: El valor del registro en la posición "i".

Percentiles (P_k)

Medida estadística de posición que divide un conjunto de datos ordenados en cien partes iguales. Cada parte representa el 1% de la muestra total.

$$P_k = x_{(i)} \quad \text{donde} \quad i = \frac{k(n+1)}{100}$$

P_k: Valor del percentil Número del decil (desde el 1% hasta el 99%), k: Porcentaje acumulado de la muestra que queda por debajo de ese valor, n: Tamaño de la muestra, $x_{(i)}$: Valor observado en la posición "i".

```
# Cuartiles (Q1, Q2, Q3)
cuartiles_iaap <- quantile(df_indices$IAAP, probs = c(0.25, 0.50, 0.75))

# Deciles (D1 al D9)
deciles_iaap <- quantile(df_indices$IAAP, probs = seq(0.1, 0.9, by = 0.1))

# Percentiles (Seleccionamos los más representativos: 1, 5, 10, 90, 95, 99)
percentiles_iaap <- quantile(df_indices$IAAP, probs = c(0.01, 0.05, 0.10, 0.3, 0.5, 0.7,
 ↪  0.90, 0.95, 0.99))

# --- Presentación de resultados ---

# Tabla de Cuartiles (Para el reporte)
knitr::kable(as.data.frame(t(cuartiles_iaap)),
             caption = "Medidas de Posición: Cuartiles del IAAP.  Datos no agrupados",
             digits = 4)
```

Table 3: Medidas de Posición: Cuartiles del IAAP. Datos no agrupados

25%	50%	75%
0.8128	1.2854	1.6829

```
#Tabla de deciles
df_deciles <- data.frame(
  Decil = names(deciles_iaap),
  Valor = as.numeric(deciles_iaap)
)

# Renombra las filas para que se lea "D1, D2..." en la tabla
df_deciles$Decil <- paste0("D", 1:9, " (", df_deciles$Decil, ")")

# 4. Genera la tabla con kable
knitr::kable(df_deciles,
             col.names = c("Decil", "Valor IAAP"),
             caption = "Distribución por Deciles de la variable IAAP.  Datos no
              ↪  agrupados",
             digits = 4,
             align = 'lc')
```

Table 4: Distribución por Deciles de la variable IAAP. Datos no agrupados

Decil	Valor IAAP
D1 (10%)	0.5513
D2 (20%)	0.7350
D3 (30%)	0.8826
D4 (40%)	1.0533
D5 (50%)	1.2854
D6 (60%)	1.4263
D7 (70%)	1.5948
D8 (80%)	1.8153
D9 (90%)	1.9950

```
# Tabla de Percentiles clave

df_percentiles <- data.frame(
  Percentil = names(percentiles_iaap),
  Valor = as.numeric(percentiles_iaap)
)

knitr::kable(df_percentiles,
             caption = "Distribución por Percentiles de la variable IAAP.  Datos no
              ↪  agrupados",
             digits = 4)
```

Table 5: Distribución por Percentiles de la variable IAAP. Datos no agrupados

Percentil	Valor
1%	0.1548
5%	0.3600
10%	0.5513
30%	0.8826
50%	1.2854
70%	1.5948
90%	1.9950
95%	2.2752
99%	2.8563

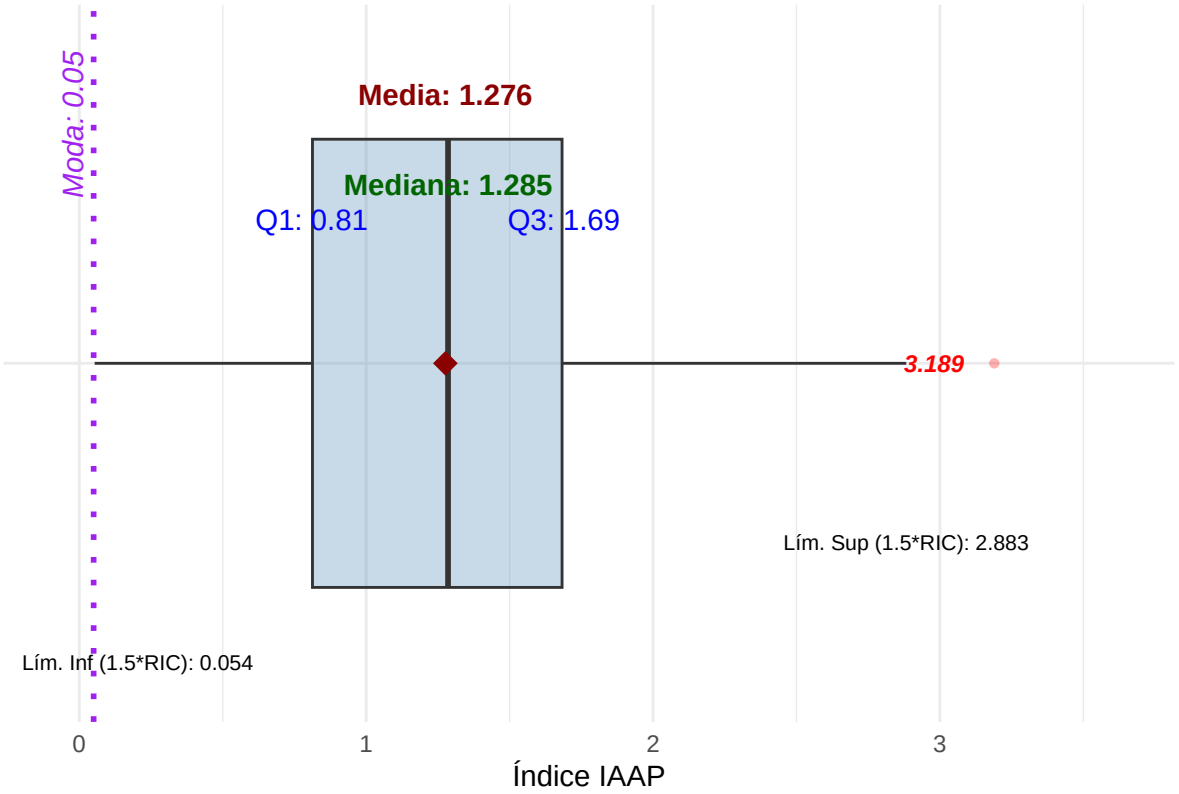
A medida que se recorre la distribución hacia los cuartiles, la brecha se hace más tangible a través del rango intercuartílico. El paso del Q_1 (0.81) al Q_3 (1.68) representa un salto de más del doble en el IAAP. Este salto estadístico nos dice que el 50% central de los estudiantes vive realidades diametralmente opuestas: mientras que el cuartil inferior se encuentra en valores bajos de sinergia académica laboral, el cuartil superior ya ha cruzado el umbral de la integración positiva. Esta dispersión es la que impide hablar de un "perfil único", ya que cualquier intervención diseñada tomando el promedio (1.27) resultaría demasiado avanzada para el primer cuartil y probablemente obsoleta o poco desafiante para el tercero.

Finalmente, el Noveno Decil ($D_9 = 1.9950$) desvela la existencia de los puntajes de IAAP. Este grupo, que representa apenas al 10% de los sujetos, ha logrado lo que el PCA identifica como una "desempeño sobresaliente". En este nivel, los vectores de Capacidad General y Especialización no solo se suman, sino que

se potencian mutuamente, alcanzando valores cercanos o superiores a 2.00, pero su escasa representatividad (solo 1 de cada 10) confirma que la excelencia en la articulación es, para esta muestra, una excepción y no la regla.

```
# 1. Cálculos base
media_val <- mean(df_indices$IAAP, na.rm = TRUE)
moda_val <- 0.05
estats <- boxplot.stats(df_indices$IAAP)$stats
q1 <- estats[2]; mediana <- estats[3]; q3 <- estats[4]
# Límites según la regla de 1.5 * RIC
lim_inf_ric <- estats[1]
lim_sup_ric <- estats[5]

# 2. Gráfico con ajustes de "aire" y nomenclatura técnica
ggplot(df_indices, aes(y = IAAP, x = "")) +
  geom_boxplot(fill = "steelblue", alpha = 0.3, outlier.color = "red", outlier.shape =
   ↪  16) +

  # Media (Diamante)
  stat_summary(fun = mean, geom = "point", shape = 18, size = 4, color = "darkred") +
  annotate("text", x = 1.45, y = media_val, label = paste0("Media: ", round(media_val,
   ↪  3)),
           color = "darkred", fontface = "bold") +

  # Moda (Línea y etiqueta inferior)
  geom_hline(yintercept = moda_val, linetype = "dotted", color = "purple", size = 1) +
  annotate("text", x = 1.4, y = moda_val, label = paste0("Moda: ", round(moda_val, 3)),
           color = "purple", fontface = "italic", angle = 90, vjust = -0.5) +

  # Cuartiles y Mediana - Desplazados hacia arriba (x = 1.3) para separarlos de la caja
  annotate("text", x = 1.3, y = q1, label = paste0("Q1: ", round(q1, 3)), color = "blue",
   ↪  vjust = 2.2) +
  annotate("text", x = 1.3, y = mediana, label = paste0("Mediana: ", round(mediana, 3)),
   ↪  color = "darkgreen", fontface = "bold") +
  annotate("text", x = 1.3, y = q3, label = paste0("Q3: ", round(q3, 3)), color = "blue",
   ↪  vjust = 2.2) +

  # Límites del RIC (Bigotes) - Etiquetas en la parte inferior (x = 0.7)
  annotate("text", x = 0.5, y = lim_inf_ric+0.15, label = paste0("Lím. Inf (1.5*RIC): ",
   ↪  round(lim_inf_ric, 3)), size = 2.8) +
  annotate("text", x = 0.7, y = lim_sup_ric, label = paste0("Lím. Sup (1.5*RIC): ",
   ↪  round(lim_sup_ric, 3)), size = 2.8) +

  # Outlier - Etiqueta a la IZQUIERDA del punto (hjust = 1.5)
  geom_text(aes(label = ifelse(IAAP > lim_sup_ric | IAAP < lim_inf_ric, round(IAAP, 3),
   ↪  "")),
            hjust = 1.5, color = "red", size = 3, fontface = "bold.italic") +

  labs(title = "Gráfico 2. Gráfico boxplot (cajas y bigotes)",
       y = "Índice IAAP", x = "") +
  theme_minimal() +
  scale_y_continuous(expand = expansion(mult = c(0.1, 0.2))) +
  coord_flip()
```

Gráfico 2. Gráfico boxplot (cajas y bigotes)



El Gráfico 2 sintetiza la distribución del IAAP mediante un diagrama de caja y bigotes integrado con medidas de tendencia central. El cuerpo central (la caja azulada) delimita el Rango Intercuartílico, extendiéndose desde el Primer Cuartil (Q_1 : 0.81) hasta el Tercer Cuartil (Q_3 : 1.69), con la Mediana (1.285) señalada por una línea interna vertical. En la parte superior del gráfico se han proyectado la Media (1.276), representada por un diamante rojo, y la Moda (0.05), indicada por una línea punteada púrpura. Hacia los extremos, los bigotes marcan los límites de dispersión normal basados en el $1.5 \times RIC$, situando el umbral inferior en 0.054 y el superior en 2.883. Finalmente, se identifica un valor atípico de 3.189 en el margen derecho, destacando como un caso de sinergia excepcional fuera de los parámetros comunes de la muestra.

Al analizar esta estructura, la amplitud de la caja revela una heterogeneidad significativa que confirma la inexistencia de un perfil estudiantil único, evidenciando que el IAAP es sumamente variable entre los sujetos. La posición de la mediana, ligeramente desplazada hacia la izquierda dentro de la caja, sugiere una asimetría positiva donde una parte considerable de los estudiantes se concentra en niveles de articulación moderados, mientras que unos pocos logran traccionar el índice hacia valores más altos. La presencia del valor atípico en 3.189 es particularmente reveladora, ya que actúa como un referente de lo que el sistema permite alcanzar en condiciones óptimas, pero su aislamiento gráfico subraya que tales niveles de sinergia entre lo académico y lo laboral son todavía la excepción y no la norma. En conjunto, el gráfico evidencia que, aunque la mayoría se sitúa en un rango positivo de integración, la brecha entre el mínimo y el máximo es lo suficientemente extensa como para demandar estrategias de gestión diferenciadas según el posicionamiento de cada alumno en este espectro.

La disposición visual de estos elementos sugiere la existencia de una asimetría negativa en la muestra, dado que la Moda (0.05) $<$ Media (1.276) $<$ Mediana (1.285). Esta configuración indicaría que, aunque la mayor concentración de estudiantes ("la joroba") se ubica en niveles de articulación positivos por encima de la unidad, existe una "cola" de alumnos en situación crítica hacia la izquierda que desplaza el promedio por debajo del valor central. La distancia entre la moda, donde se registra la mayor frecuencia de respuestas cercanas a la desconexión total, y la mediana revela una brecha estructural: el sistema educativo convive con un grupo de alto rendimiento (representado por el outlier y el Q_3) y un segmento de vulnerabilidad que,

al ser tan extremo, deforma la representatividad de la media aritmética. El gráfico demuestra que no existe un perfil estudiantil único, sino un espectro de integración condicionado por una fuerte disparidad interna.

5. Medidas de variabilidad

Varianza Muestral (S^2)

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

x_i: Valor del IAAP para cada estudiante, $\bar{x}$: Media aritmética del IAAP, n: Tamaño de la muestra.

Desviación Estándar (S)

$$S = \sqrt{S^2}$$

Coeficiente de Variación (CV)

$$CV = \left(\frac{S}{\bar{x}}\right) \times 100$$

CV: Coeficiente de Variación, S: Desviación Estándar Muestral, $\bar{x}$: Media Aritmética, 100: Es el factor de conversión para expresar el resultado final en términos porcentuales (%).

Error Estándar de la Media (EEM)

$$EEM = \frac{S}{\sqrt{n}}$$

S: Desviación estándar de la muestra, n: Tamaño de la muestra.

```r
# 0. Valores Base
min_iaap <- min(df_indices$IAAP)
max_iaap <- max(df_indices$IAAP)

# 1. Rango (Valor Máx - Valor Mín)
rango_iaap <- max(df_indices$IAAP) - min(df_indices$IAAP)

# 2. Varianza Poblacional y Muestral
# var() en R calcula la muestral (n-1)
varianza_muestral <- var(df_indices$IAAP)

# 3. Desviación Estándar
desv_std_iaap <- sd(df_indices$IAAP)

# 4. Coeficiente de Variación (CV)
cv_iaap <- (desv_std_iaap / media_arit_iaap) * 100

# 5. Error Estándar de la Media
error_est_iaap <- desv_std_iaap / sqrt(length(df_indices$IAAP))
# Consolida en una tabla para el reporte
tabla_variabilidad <- data.frame(
  Indicador = c("Valor Mínimo", "Valor Máximo","Rango", "Varianza (Muestral)",
    ↪  "Desviación Estándar", "C.V. (%)", "Error Estándar de la Media"),
```

```
  Valor = c(min_iaap, max_iaap, rango_iaap, varianza_muestral, desv_std_iaap, cv_iaap,
   ↪  error_est_iaap)
)
```

```
knitr::kable(tabla_variabilidad,
             caption = "Medidas de Variabilidad y Dispersión del IAAP",
             digits = 4)
```

Table 6: Medidas de Variabilidad y Dispersión del IAAP

Indicador	Valor
Valor Mínimo	0.0536
Valor Máximo	3.1895
Rango	3.1359
Varianza (Muestral)	0.3650
Desviación Estándar	0.6041
C.V. (%)	47.3408
Error Estándar de la Media	0.0382

El Valor Mínimo (0.0536) y el Valor Máximo (3.1895) definen un Rango de 3.1359, lo que indica que entre el estudiante con menor articulación y el de mayor desempeño existe una distancia de más de tres unidades, cubriendo casi toda la escala posible del índice IAAP. La gran amplitud del rango, impulsada por un valor mínimo que roza el cero y un máximo que se dispara como valor atípico, genera una desviación estándar elevada en comparación con el promedio. Esto significa que la media de 1.276 no representaría a un "estudiante promedio" inexistente, sino que es el resultado de promediar dos realidades opuestas.

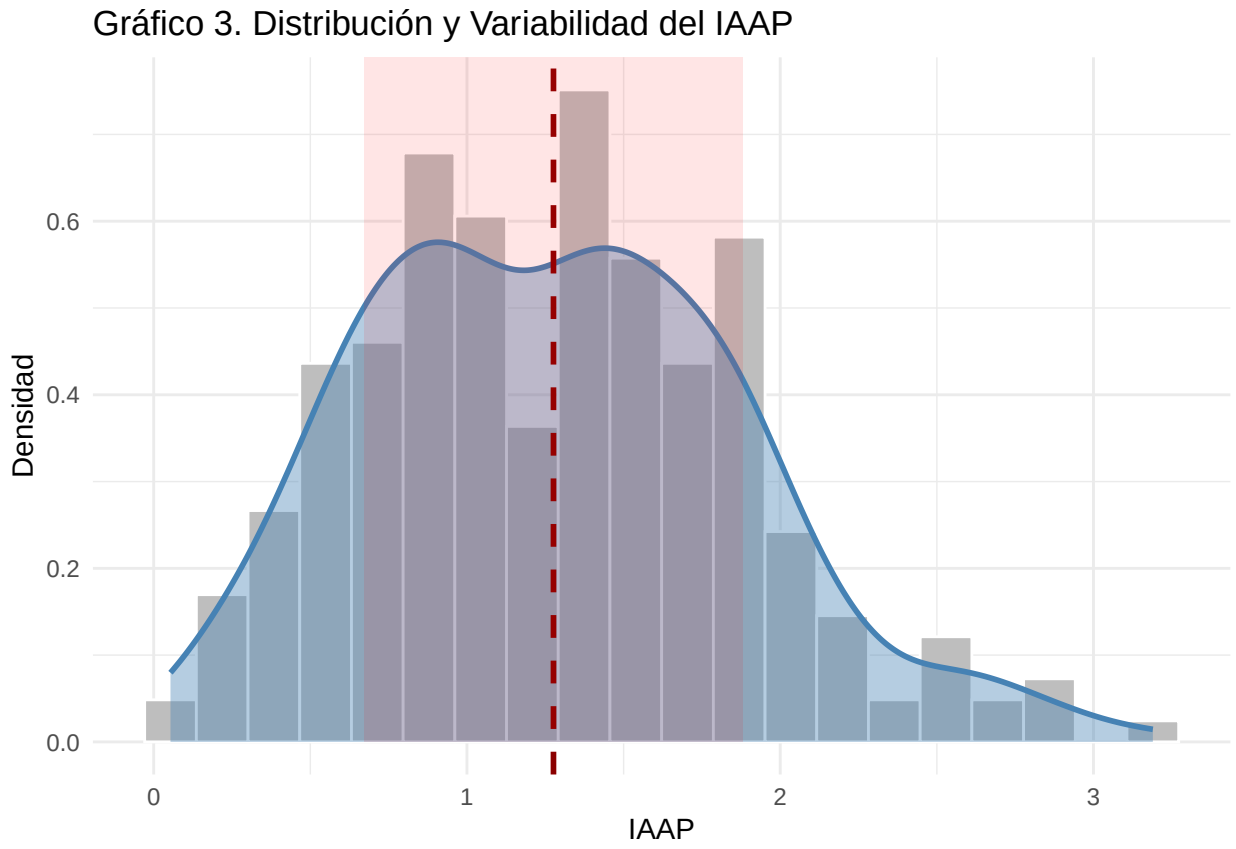
Al observar la dispersión promedio, la Varianza Muestral (0.3650) y la Desviación Estándar (0.6041) indicarían que, en promedio, las puntuaciones IAAP de los alumnos se alejan un poco más de medio punto respecto a la media de 1.276.

La interpretación integral de la estructura del IAAP se consolida mediante el Coeficiente de Variación (CV : 47.34%). Desde una perspectiva estrictamente matemática, este indicador permite validar la representatividad de la tendencia central en relación con la dispersión observada. Al superar el umbral técnico del 30%, el CV clasifica la distribución como altamente heterogénea, lo que implica que la desviación estándar representa casi la mitad del valor de la media aritmética.

Por otro lado, el alto $C.V.$ valida lo que se ve en el boxplot: existe una dispersión tan grande que sugiere múltiples realidades que requieren intervenciones diferenciadas, ya que casi la mitad de la variabilidad del índice está presente en cada medición.

```
ggplot(df_indices, aes(x = IAAP)) +
  # Histograma de fondo
  geom_histogram(aes(y = ..density..), bins = 20, fill = "gray", color = "white") +
  # Curva de densidad para ver la variabilidad suavizada
  geom_density(fill = "steelblue", alpha = 0.4, color = "steelblue", size = 1) +
  # Línea de la Media
  geom_vline(aes(xintercept = mean(IAAP)), color = "darkred", linetype = "dashed", size =
   ↪  1) +
  # Sombreado de la Desviación Estándar (opcional para ver la dispersión)
  annotate("rect", xmin = mean(df_indices$IAAP) - sd(df_indices$IAAP),
           xmax = mean(df_indices$IAAP) + sd(df_indices$IAAP),
           ymin = 0, ymax = Inf, alpha = 0.1, fill = "red") +
```

```
labs(title = "Gráfico 3. Distribución y Variabilidad del IAAP",
     x = "IAAP",
     y = "Densidad") +
theme_minimal()
```



El análisis visual del comportamiento del IAAP se fundamenta en la integración de un histograma de frecuencias y una curva de densidad suavizada, elementos que permiten contrastar la distribución real con la tendencia teórica de los datos. El eje horizontal delimita el recorrido del índice desde su valor mínimo hasta el máximo, mientras que el eje vertical de densidad permite identificar los puntos de mayor concentración sin depender del tamaño de la muestra. La morfología de la curva, caracterizada por una base extendida y una altura moderada, constituye la evidencia gráfica de la alta variabilidad detectada; una distribución homogénea presentaría una estructura estrecha y espigada, mientras que el achatamiento observado aquí confirma visualmente un coeficiente de variación elevado.

La línea vertical discontinua de color rojo se sitúa en el centro de la distribución, representando la media aritmética de 1.27 puntos. Alrededor de este eje, el área sombreada delimita el espacio ocupado por la desviación estándar, permitiendo dimensionar cuánta superficie de la muestra se aleja del promedio central. Es notable cómo las barras del histograma se desplazan considerablemente hacia los extremos del gráfico, lo que demuestra que una proporción importante de los datos se ubica fuera del entorno inmediato de la media. Este distanciamiento entre las barras y el eje rojo es la representación física de la dispersión de los datos.

Finalmente, la asimetría del gráfico se hace evidente al observar la extensión de la curva hacia el extremo derecho del eje horizontal. Mientras que la mayor acumulación de datos ocurre en valores bajos, la presencia de barras en niveles cercanos a 3 unidades genera una "cola" que estira la figura hacia valores positivos. Esta configuración visual anticipa el análisis de forma, sugiriendo que la distribución no es perfectamente simétrica y que los valores máximos ejercen una tracción constante sobre el promedio, alejándolo de la base de la pirámide donde se concentra la mayoría de las observaciones.

6. Medidas de Asimetría y Curtosis

Para los coeficientes de asimetría de Fisher, Pearson y Bowley, el objetivo es determinar hacia dónde se inclina la distribución. Si el resultado es cero, la distribución es simétrica (espejo perfecto). Un valor positivo indica una asimetría a la derecha, donde la cola se extiende hacia los valores altos del IAAP, sugiriendo que la media es mayor que la mediana. Por el contrario, un valor negativo indica una asimetría a la izquierda, con la cola apuntando hacia los valores bajos. La diferencia clave es que Pearson y Fisher consideran toda la serie de datos, mientras que Bowley nos dirá si el "corazón" de la muestra (el 50% central) también está sesgado o si la asimetría es causada solo por los valores extremos.

El coeficiente de Curtosis define el grado de concentración de los datos en la zona central de la distribución. Se interpreta tomando como referencia el valor 0 (distribución mesocúrtica o normal). Si el coeficiente es positivo, la distribución es leptocúrtica, lo que significa que hay una gran concentración de estudiantes con valores de IAAP muy similares entre sí, creando un pico muy alto. Si el coeficiente es negativo, la distribución es platicúrtica, lo que indica que los datos están muy repartidos y la curva es achatada, confirmando visualmente la alta dispersión que ya detectamos con el Coeficiente de Variación.

###Coeficiente de Asimetría de Fisher (g_1)

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3}$$

g_1: Coeficiente de asimetría de Fisher, x_i: Valor del IAAP para cada observación individual, $\bar{x}$: Media aritmética de la muestra, n: Tamaño de la muestra), $\sum(x_i - \bar{x})^3$: Sumatoria de las desviaciones respecto a la media elevadas al cubo.

Coeficiente de Asimetría de Pearson (A_p)

$$A_p = \frac{3(\bar{x} - Me)}{S}$$

A_p: Coeficiente de asimetría de Pearson, $\bar{x}$: Media aritmética, Me: Mediana de la distribución, S: Desviación estándar de la muestra

Coeficiente de Asimetría de Bowley (A_b)

$$A_b = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

A_b: Coeficiente de asimetría de Bowley, Q_1: Primer cuartil (percentil 25), Q_3: Tercer cuartil (percentil 75), Me: Mediana (percentil 50).

Coeficiente de Curtosis (g_2)

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

g_2: Coeficiente de curtosis (exceso), $\sum(x_i - \bar{x})^4$: Sumatoria de las desviaciones respecto a la media elevadas a la cuarta potencia, 3: Valor de referencia de la distribución normal (mesocúrtica); al restarlo, una curtosis de 0 indica una forma normal.

```r
# 1. Asimetría de Fisher (G1)
# Usamos la librería moments
g1_fisher <- moments::skewness(df_indices$IAAP)

# 2. Asimetría de Pearson (Ap)
# Fórmula: (Media - Moda) / Desviación Estándar
# Usa la primera moda en caso de que sea multimodal
ap_pearson <- (media_arit_iaap - moda_iaap[1]) / desv_std_iaap

# 3. Asimetría de Bowley (Ab)
# Basada en cuartiles: (Q3 + Q1 - 2*Q2) / (Q3 - Q1)
# Es más robusta ante valores extremos
q1 <- cuartiles_iaap[1]
q2 <- cuartiles_iaap[2] # La mediana
q3 <- cuartiles_iaap[3]
ab_bowley <- (q3 + q1 - 2*q2) / (q3 - q1)

# 4. Curtosis (g2)
# El valor 3 representa la distribución Normal (Mesocúrtica)
g2_curtosis <- moments::kurtosis(df_indices$IAAP)

# Consolida para el reporte
tabla_forma <- data.frame(
  Medida = c("Asimetría de Fisher (G1)",
             "Asimetría de Pearson (Ap)",
             "Asimetría de Bowley (Ab)",
             "Curtosis (g2)"),
  Valor = c(g1_fisher, ap_pearson, ab_bowley, g2_curtosis)
)

knitr::kable(tabla_forma,
             caption = "Medidas de Forma: Asimetría y Curtosis del IAAP",
             digits = 4)
```

Table 7: Medidas de Forma: Asimetría y Curtosis del IAAP

Medida	Valor
Asimetría de Fisher (G1)	0.3847
Asimetría de Pearson (Ap)	2.0237
Asimetría de Bowley (Ab)	-0.0863
Curtosis (g2)	2.8723

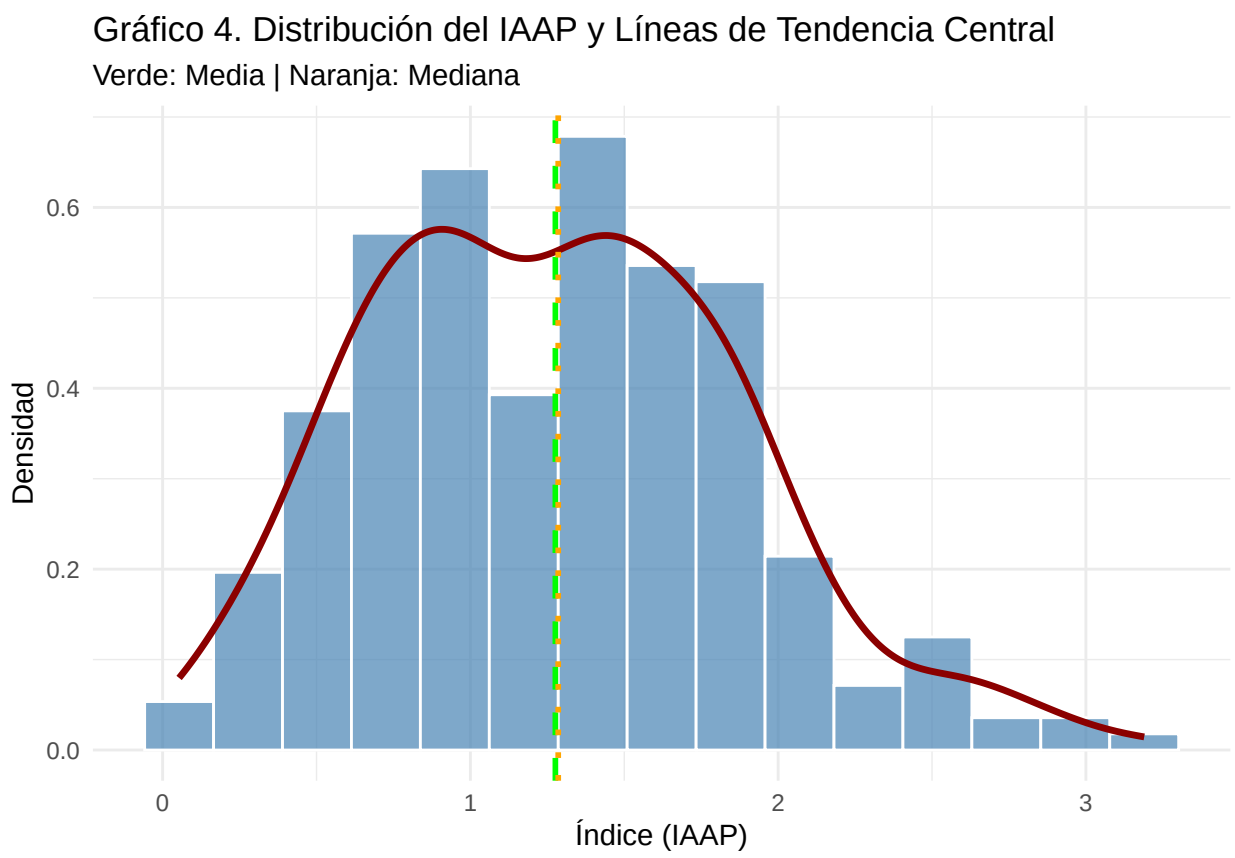
El Coeficiente de Asimetría de Fisher (0.3847) y el Coeficiente de Pearson (2.0237) arrojan valores positivos, lo que define técnicamente una distribución con sesgo a la derecha. Este fenómeno revela que el peso matemático de los valores máximos del IAAP es superior a la masa visual de los datos bajos; las desviaciones elevadas al cubo otorgan a los estudiantes con desempeños excepcionales (cercanos a 3.18) una "fuerza de tracción" desproporcionada que estira la cola de la distribución y desplaza la media por encima de la mediana.

Esta aparente contradicción con la percepción visual se resuelve al observar el Coeficiente de Bowley (-0.0863), el único que mantiene el signo negativo. Al basarse exclusivamente en la posición de los cuartiles y ser insensible a los valores extremos, Bowley confirma que el "corazón" de la muestra —el 50% central de los sujetos— presenta la inclinación negativa que se intuía originalmente. Esta discrepancia indica que la asimetría positiva general es provocada por outliers y no por un sesgo estructural en el bloque

central de estudiantes.En consecuencia, el IAAP presenta una dualidad morfológica: la distribución es estructuralmente negativa en su base (donde la mayoría tiende a agruparse en niveles superiores), pero técnicamente positiva en su totalidad debido a una minoría con niveles de articulación muy elevados que sesgan la estadística global.

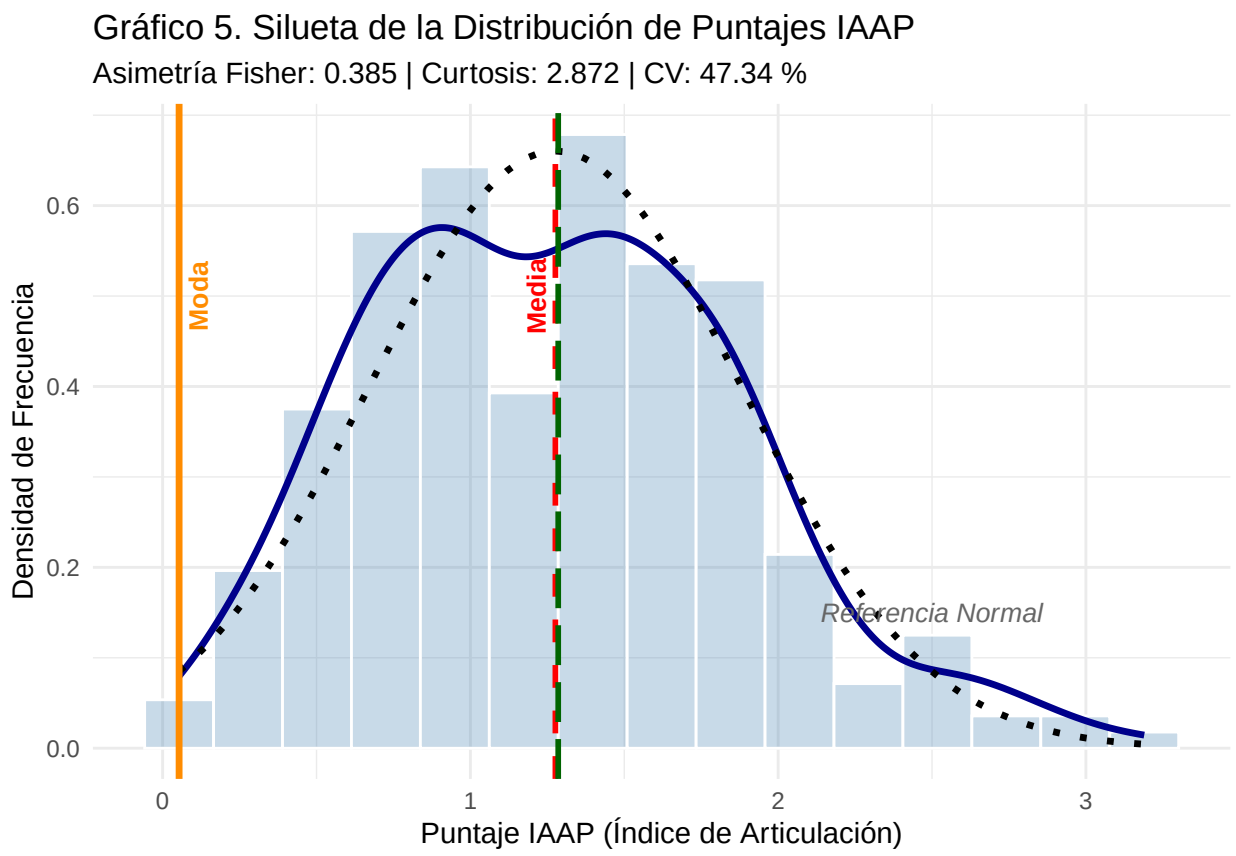
Finalmente, la Curtosis positiva clasifica la distribución como leptocúrtica. A pesar de la variabilidad detectada, los datos no se distribuyen de forma uniforme, sino que presentan un elevado grado de concentración en torno a la zona central, creando un pico más pronunciado que el de una distribución normal teórica, lo que visualmente se traduce en ese "pico" central que sobresale en el histograma..

```
ggplot(df_indices, aes(x = IAAP)) +
  geom_histogram(aes(y = ..density..), bins = 15,
                 fill = "steelblue", color = "white", alpha = 0.7) +
  geom_density(color = "darkred", size = 1.2) +
  geom_vline(aes(xintercept = media_arit_iaap), color = "green",
             linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = mediana_iaap), color = "orange",
             linetype = "dotted", size = 1) +
  labs(title = "Gráfico 4. Distribución del IAAP y Líneas de Tendencia Central",
       subtitle = "Verde: Media | Naranja: Mediana",
       x = "Índice (IAAP)", y = "Densidad") +
  theme_minimal()
```




```
ggplot(df_indices, aes(x = IAAP)) +
  # 1. Histograma
  geom_histogram(aes(y = after_stat(density)), bins = 15,
```

```
                fill = "steelblue", alpha = 0.3, color = "white") +

  # 2. Curva Kernel
  geom_density(color = "darkblue", size = 1.2) +

  # 3. Curva Normal Teórica (Referencia)
  # Usamos las variables originales: media_arit_iaap y desv_std_iaap
  stat_function(fun = dnorm,
                args = list(mean = media_arit_iaap, sd = desv_std_iaap),
                color = "black", linetype = "dotted", size = 1.2) +

  # 4. Líneas de referencia de Tendencia Central

  geom_vline(aes(xintercept = media_arit_iaap), color = "red", linetype = "dashed", size
   ↪  = 1) +
  geom_vline(aes(xintercept = mediana_iaap), color = "darkgreen", linetype = "longdash",
   ↪  size = 1) +
  geom_vline(aes(xintercept = moda_iaap[1]), color = "darkorange", size = 1.2) +

  # 5. Anotaciones de las líneas
  annotate("text", x = media_arit_iaap, y = 0.5, label = "Media", color = "red",
           angle = 90, vjust = -0.5, size = 3.5, fontface = "bold") +
  annotate("text", x = moda_iaap[1], y = 0.5, label = "Moda", color = "darkorange",
           angle = 90, vjust = 1.5, size = 3.5, fontface = "bold") +

  # Etiqueta para la curva normal de referencia
  annotate("text", x = 2.5, y = 0.15, label = "Referencia Normal",
           color = "grey40", size = 3.5, fontface = "italic") +

  # Estética y subtítulo dinámico
  labs(title = "Gráfico 5. Silueta de la Distribución de Puntajes IAAP",
       subtitle = paste("Asimetría Fisher:", round(g1_fisher, 3),
                        "| Curtosis:", round(g2_curtosis, 3),
                        "| CV:", round(cv_iaap, 2), "%"),
       x = "Puntaje IAAP (Índice de Articulación)",
       y = "Densidad de Frecuencia") +
  theme_minimal()
```



Aunque el análisis no pretende validar la normalidad de la variable, la superposición de una curva normal teórica funciona como una heurística visual para subrayar las desviaciones del IAAP. Al comparar la silueta real con la campana de Gaus se evidencia que los datos no siguen un patrón aleatorio uniforme. La curva real "sobresale" por encima de la normal en el centro, lo que ratifica visualmente la Curtosis leptocúrtica calculada; a su vez, el desplazamiento de las colas hacia la derecha respecto a la normal teórica expone la Asimetría positiva generada por los valores atípicos.

En este sentido, la normal no se presenta como un modelo de ajuste, sino como un marco de referencia. Permite al lector comprender que la "dualidad" mencionada (esa base negativa y totalidad positiva) es una desviación real y significativa respecto a lo que se esperaría de una distribución equilibrada. Esta comparación permanece en el plano ilustrativo, sirviendo para dar profundidad a la interpretación de los coeficientes de Fisher, Pearson y Bowley, sin pretender sustituir a las pruebas de bondad de ajuste

7. Comparativa con coeficientes obtenidos por datos agrupados

En el anterior material, en el Trabajo #1, los coeficientes obtenidos para datos agrupados fueron:

Table 8: Resumen de Estadísticos IAAP - Metodología de Datos Agrupados (T1)

Estadístico	Valor_Agrupado
Media Aritmética	1.2828
Media Geométrica	1.1102
Mediana (Interpolada)	1.2731
Moda (Marca de Clase)	0.9246
Varianza	0.3875
Desviación Estándar	0.6225
Coeficiente de Variación (%)	48.5266
Asimetría de Fisher	0.4378
Asimetría de Pearson	0.5754
Asimetría de Bowley	0.0663
Curtosis (g2)	-0.2105

A continuación, se presenta una tabla comparativa que sintetiza las diferencias operativas y analíticas entre ambas metodologías aplicadas en la serie del "IAAP":

Table 9: Comparación entre cálculo de datos agrupados y no agrupados

Indicador	Datos_Agrupados	Datos_no_agrupados	Diferencia
Media Aritmética	1.2828	1.2761	-0.0067
Media Geométrica	1.1102	1.1020	-0.0082
Mediana	1.2731	1.2854	0.0123
Moda	0.9246	0.0536	-0.8711
Varianza	0.3875	0.3650	-0.0226
Desviación Estándar	0.6225	0.6041	-0.0184
Coef. Variación (%)	48.5266	47.3408	-1.1858
Asimetría Fisher (g1)	0.4378	0.3847	-0.0531
Asimetría Pearson	0.5754	2.0237	1.4483
Asimetría Bowley	0.0663	-0.0863	-0.1525
Curtosis (g2)	-0.2105	2.8723	3.0828

Al analizar la tendencia central, se observa que la Media Aritmética (1.28 vs. 1.27) y la Geométrica (1.11 vs. 1.10) se mantienen notablemente estables, validando que el promedio es un indicador robusto ante la agrupación. Sin embargo, la Mediana y, especialmente, la Moda revelan la primera "grieta" metodológica profunda: mientras el método de Sturges situaba una moda de 0.92 (anclada a una marca de clase), el dato bruto la desplaza drásticamente a 0.05. Esta diferencia de -0.87 demuestra que el valor más frecuente es, en realidad, un extremo que el agrupamiento por intervalos suavizaba hasta hacerlo invisible.

En cuanto a la variabilidad, aunque la Varianza y la Desviación Estándar presentan cambios leves, el Coeficiente de Variación cae del 48.5% al 47.3%, confirmando que la agrupación del Trabajo #1 inflaba ligeramente la percepción de dispersión de los datos.

La morfología de la distribución es donde el contraste se vuelve incontestable. La Asimetría de Fisher se mantiene en rangos similares, pero la de Pearson se dispara de 0.57 a 2.02, consecuencia directa de haber identificado la moda real en los datos no agrupados. El Coeficiente de Bowley ofrece el hallazgo más disruptivo

al cambiar de signo (de 0.06 a −0.08), revelando que el núcleo de la muestra —el 50% central— posee un sesgo negativo que la agrupación ocultaba por completo.

Finalmente, la Curtosis pasa de un estado casi plano (platicúrtico, −0.21) a un punzante 2.87 (leptocúrtico), evidenciando que los datos están mucho más concentrados y presentan picos mucho más agudos de lo que la visión agrupada permitía suponer.En conclusión, mientras que los datos agrupados entregaron para la distribución de datos de la IAAP una silueta suavizada y promediada, la analítica de datos brutos y no agrupados revela una realidad mucho más extrema, asimétrica y concentrada.

8. ¿Agrupar o no agrupar datos?

El contraste entre ambos métodos revela lo que en teoría de la información se conoce como la "entropía de agrupación". Al aplicar el Teorema de Shannon (1949) al análisis estadístico, se comprende por qué el proceso de agrupar datos brutos en intervalos actúa como un canal con ruido que limita la "capacidad de información" de la muestra. Según Shannon (1949), cualquier proceso de discretización de una señal continua implica una pérdida irreversible de resolución; al reducir observaciones individuales a intervalos, se disminuye el "ancho de banda" del dato.

Esta pérdida de fidelidad es la que Fisher (1922) describía como una "reducción de la eficiencia", donde la marca de clase se convierte en una representación "pixelada" del valor real. Siguiendo a Scott (1979), esta simplificación actúa como un filtro suavizador que elimina los picos de frecuencia. Esta distorsión es lo que Silverman (1986) define como el "sesgo de la estructura del histograma", reforzando la tesis de que la agrupación es una herramienta de visualización, pero un obstáculo para la precisión analítica. En su investigación, Cox (1957) lo define como "forma de pérdida de información por agrupamiento", y explica por qué la Media Aritmética y la Media Geométrica presentan variaciones leves, mientras que estadísticos de posición como la Moda y la Mediana sufren desplazamientos significativos al quedar "anclados" a marcas de clase artificiales.

En el caso de los Trabajos #1 y #2 desarrollados, el contraste más agresivo se evidencia en la Curtosis y la Asimetría. Según Scott (1979), el uso de anchos de clase fijos (como los derivados de la regla de Sturges) actúa como un filtro suavizador que elimina los picos de frecuencia, lo que justifica técnicamente por qué la curtosis agrupada resultó ser negativa (−0.2105), mientras que el dato bruto reveló una estructura leptocúrtica (2.8723), Asimismo, la inestabilidad en los coeficientes de Pearson y Bowley tras la agrupación confirma la advertencia de Kendall y Stuart (1977), quienes sostienen que los estadísticos de forma calculados mediante intervalos son meras "estimaciones burdas" que pueden incluso invertir el signo de la asimetría, como ocurrió en este estudio.

La recomendación principal de la American Statistical Association (ASA), así como de autores como Hyndman y Fan (1996), es que dada la capacidad computacional, el analista debe priorizar siempre el uso de datos no agrupados. La agrupación se considera hoy una técnica estrictamente de visualización y no una base de cálculo, ya que el cálculo sobre intervalos introduce un error innecesario que no tiene lugar en la ciencia de datos moderna.

Para los casos donde la presentación de datos agrupados sea mandatoria, la academia sugiere protocolos específicos para mitigar las discrepancias:

- Aplicación de las Correcciones de Sheppard: Para compensar la infraestimación de la variabilidad, se recomienda el uso de las correcciones de Sheppard (1898), que permiten ajustar la varianza de los datos agrupados restando un factor de corrección ($h^2/12$). No obstante, autores como Kendall y Stuart (1977) advierten que estas correcciones solo son efectivas si la distribución es aproximadamente normal en las colas, lo que refuerza la necesidad de validar primero con el dato bruto.
- Sustitución de Histogramas por Kernels: En lugar de depender de la rigidez de la regla de Sturges, la recomendación de Silverman (1986) y Wand y Jones (1995) es el empleo de la Estimación de Densidad por Kernel (KDE). Esta técnica permite visualizar la forma de la distribución —incluyendo la leptocurtosis y la asimetría— de manera continua, evitando los "escalones" artificiales que distorsionan la percepción visual y analítica del IAAP.

- Análisis de Sensibilidad de Intervalos: Si se opta por la agrupación, Scott (1979) recomienda realizar un análisis de sensibilidad variando el ancho del intervalo (h). Esto permite confirmar si la forma de la distribución (como la curtosis observada) es una propiedad real del fenómeno o simplemente un artefacto de cómo se cortaron los datos, un paso que en este trabajo se resolvió mediante el contraste directo con la muestra original.

9. Conclusiones

El presente Trabajo #2 (Datos no Agrupados) consiste en la continuación de lo desarrollado en el Trabajo #1 (Datos Agrupados). El Trabajo #1 desarrolló el cálculo de estadísticos tradicionales a partir de tablas de frecuencias generadas por agrupación de intervalos según la regla de Sturges. El objetivo del Trabajo #1 fue establecer una estructura visual y una primera aproximación al análisis de datos.

Por el contrario, el presente Trabajo #2, analizó la totalidad de las observaciones de forma individual, es decir, datos sin agrupación, buscando encontrar las diferencias que surgen al contrastar los resultados obtenidos mediante la técnica de agrupación. Este ejercicio permitió auditar el impacto de la pérdida de información que ocurre al agrupar los datos, verificando cómo la elección del tratamiento metodológico altera los estadísticos resultantes.

La evidencia recolectada sugiere que la aplicación de la regla de Sturges y la posterior agrupación en intervalos actúan como un filtro de "suavizado" que distorsiona la morfología de la muestra. A pesar de que la Media Aritmética (1.28) mostró estabilidad entre ambos métodos, el Coeficiente de Variación (47.34%) advierte que este promedio revela una profunda heterogeneidad y no la representación de una observación típica de la muestra.

En el Trabajo #1, la Varianza se calculó en 0.39, mientras que el tratamiento de datos no agrupados en el Trabajo #2 arroja un valor de 0.36. Esta diferencia de -0.02 indica que la técnica de intervalos introdujo una variabilidad adicional que no pertenece a los datos reales. Al asignar valores a una "marca de clase" (punto medio del intervalo), se genera un error de redondeo que sobreestima la dispersión total. Asimismo, la Desviación Estándar disminuyó de 0.62 a 0.60, lo que impacta en el Coeficiente de Variación, que desciende del 48.53% al 47.34%.

En cuanto a la simetría, en el Trabajo #1, el coeficiente de 0.07 sugería una distribución con una leve asimetría positiva. Sin embargo, al aplicar el tratamiento de datos no agrupados en el Trabajo #2, el valor se desplazó a -0.09 (una diferencia de -0.15). El contraste más severo se observó en el apuntamiento de la curva. La técnica de agrupación arrojó una curtosis de -0.21, lo que indicaba una distribución platicúrtica (achatada); por el contrario, el análisis de datos individuales en R reveló una estructura marcadamente leptocúrtica de 2.87. Esta diferencia de 3.08 sugiere que la población está mucho más concentrada en torno a valores específicos de lo que el método de intervalos permitía visualizar; la agrupación "aplanó" artificialmente la realidad, eliminando la rugosidad y los picos de frecuencia del dato original.

Finalmente, mientras la asimetría de Pearson en la agrupación arrojaba un valor de 0.58, el tratamiento de datos reales elevó este indicador a 2.02 (una diferencia de 1.45). Este incremento refuerza la tesis de que la estructura de la muestra es mucho más compleja y sesgada de lo que el modelo simplificado de Sturges permitía interpretar.

Desde la perspectiva de la Teoría de la Información de Shannon (1949), la agrupación de datos implica una pérdida de resolución que disminuye la potencia estadística. Por tanto, y en concordancia con las recomendaciones de la American Statistical Association (ASA), se concluye que:

- La agrupación de datos debe limitarse preferentemente a fines de presentación visual, evitando su uso como base para el cálculo de estadísticos de forma o variabilidad cuando se busca precisión analítica.
- El empleo de herramientas de programación como R favorece la reproducibilidad científica y permite mitigar los errores de aproximación que son frecuentes en el tratamiento de agrupación de datos.

- La toma de decisiones en el ámbito académico y social debería fijarse en el análisis de datos brutos (no agrupados) para evitar el sesgo de agregación que podría conducir a diagnósticos institucionales erróneos.

10. Referencias

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, 222, 309-368.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4), 361-365.

Kendall, M. G., & Stuart, A. (1977). *The Advanced Theory of Statistics*. Griffin.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3), 605-610.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Sheppard, W. F. (1898). On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant divisions of a scale. *Proceedings of the London Mathematical Society*, s1-29(1), 353-380.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Wand, M. P., & Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall/CRC