

ANOVA y Análisis de Componentes Principales

Unidad 2: Estadística Básica y Aplicada

Nicolás Sidicaro

Octubre 2025

Parte 1: ANOVA

Analysis of Variance

Motivación: El problema de las comparaciones

Situación: Queremos comparar salarios entre 5 industrias diferentes.

Opción ingenua: Hacer múltiples t-tests

- Industria A vs B
- Industria A vs C
- Industria B vs C
- ... (10 comparaciones en total)

Problema: Con cada test hay 5% de probabilidad de error tipo I

$$P(\text{al menos un error}) = 1 - (0.95)^{10} = 0.40$$

Con 10 tests, hay **40% de probabilidad** de encontrar al menos una diferencia falsa.

ANOVA: La solución

Análisis de Varianza (ANOVA) resuelve este problema con un **único test omnibus**:

- H_0 : Todas las medias grupales son iguales ($\mu_1 = \mu_2 = \dots = \mu_k$)
- H_1 : Al menos una media es diferente

Ventajas:

- Controla el error tipo I familiar-wise
- Un solo p-valor para la pregunta global
- Descompone la varianza total en componentes interpretables

¿Cuándo usar ANOVA?

- Comparar 3+ grupos en una variable continua
- Diseño experimental con múltiples tratamientos
- Datos que cumplen supuestos paramétricos

La lógica de ANOVA

ANOVA descompone la **varianza total** en dos fuentes:

$$SS_{Total} = SS_{Between} + SS_{Within}$$

Varianza Between-groups (SS_B):

- ¿Qué tan diferentes son las medias grupales?
- Atribuible al factor de interés

Varianza Within-groups (SS_W):

- ¿Qué tan variables son los datos dentro de cada grupo?
- Error aleatorio, variabilidad natural

Estadístico F: Compara ambas fuentes

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{SS_B / (k - 1)}{SS_W / (n - k)}$$

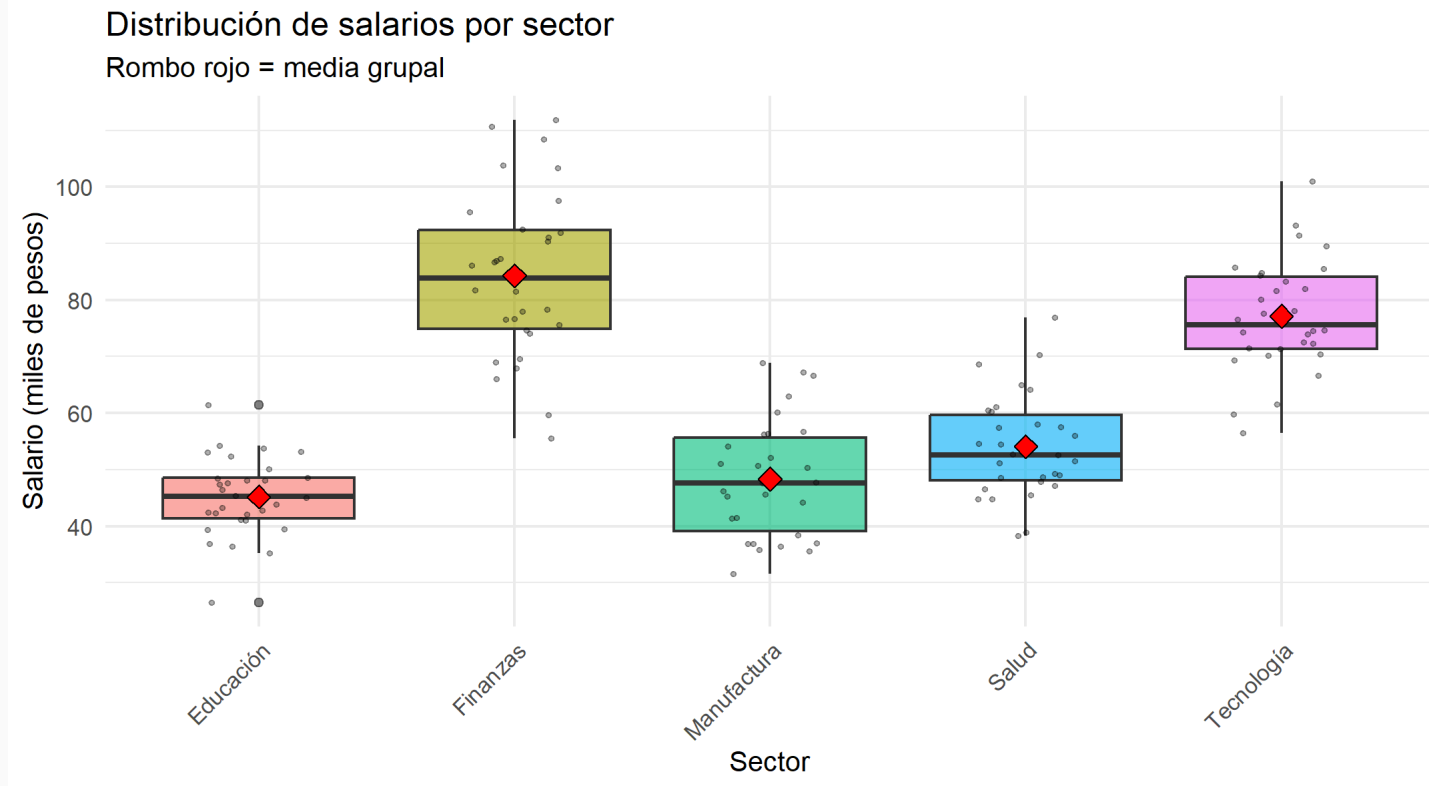
Si F es grande \rightarrow la variación entre grupos es mucho mayor que dentro de grupos \rightarrow evidencia contra H_0

Generación de datos: Salarios por sector

```
salarios <- tibble(  
  sector = rep(c("Finanzas", "Tecnología", "Educación", "Salud", "Manufactura"),  
              each = 30),  
  salario = c(  
    rnorm(30, mean = 85, sd = 15), # Finanzas  
    rnorm(30, mean = 75, sd = 12), # Tecnología  
    rnorm(30, mean = 45, sd = 8),  # Educación  
    rnorm(30, mean = 55, sd = 10), # Salud  
    rnorm(30, mean = 50, sd = 9)   # Manufactura  
  )  
)
```

Sector	N	Media	SD
Educación	30	45.2	7.0
Finanzas	30	84.3	14.7
Manufactura	30	48.3	10.3
Salud	30	54.1	9.1
Tecnología	30	77.1	10.0

Visualización de los datos



ANOVA en R: Implementación

```
# ANOVA de un factor
```

```
modelo_anova <- aov(salario ~ sector, data = salarios)
summary(modelo_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sector         4  37735     9434   85.07 <2e-16 ***
## Residuals    145  16080       111
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretación:

- Si $F_{value} < \alpha \rightarrow$ Rechazamos H_0
- Al menos un sector tiene salario promedio diferente
- El sector explica una parte significativa de la variabilidad salarial

Post-hoc: ¿Qué grupos difieren?

ANOVA solo responde "hay diferencias", no indica **cuáles**.

Solución: Comparaciones post-hoc con corrección por comparaciones múltiples.

```
# Test de Tukey (controla error familiar-wise)
comparaciones ← emmeans(modelo_anova, pairwise ~ sector, adjust = "tukey")
```

Comparación	Diferencia	p-valor	Sig.
Educación - Finanzas	-39.10	0.0000	***
Educación - Manufactura	-3.15	0.7743	
Educación - Salud	-8.87	0.0119	***
Educación - Tecnología	-31.94	0.0000	***
Finanzas - Manufactura	35.95	0.0000	***
Finanzas - Salud	30.23	0.0000	***
Finanzas - Tecnología	7.15	0.0701	
Manufactura - Salud	-5.71	0.2252	

ANOVA vs Regresión: Son equivalentes

```
# Opción 1: ANOVA  
modelo_anova ← aov(salario ~ sector, data = salarios)  
# Opción 2: Regresión con dummies (EQUIVALENTE)  
modelo_regresion ← lm(salario ~ factor(sector), data = salarios)
```

```

## Analysis of Variance Table
##
## Response: salario
##           Df Sum Sq Mean Sq F value    Pr(>F) ##
## factor(sector)    4   37735    9433.8   85.07 < 2.2e-16 ***
## Residuals       145   16080     110.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Call:
## lm(formula = salario ~ factor(sector), data = salarios)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.7927  -6.8327  -0.6807   6.6443  27.5103
## ---
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      45.195      1.923   23.507 < 2e-16 ***
## factor(sector)Finanzas      39.098      2.719   14.380 < 2e-16 ***
## factor(sector)Manufactura     3.152      2.719    1.159  0.24820
## factor(sector)Salud           8.866      2.719    3.261  0.00139 **
## factor(sector)Tecnología     31.945      2.719   11.749 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.53 on 145 degrees of freedom
## Multiple R-squared:  0.7012,    Adjusted R-squared:  0.693
## F-statistic: 85.07 on 4 and 145 DF,  p-value: < 2.2e-16

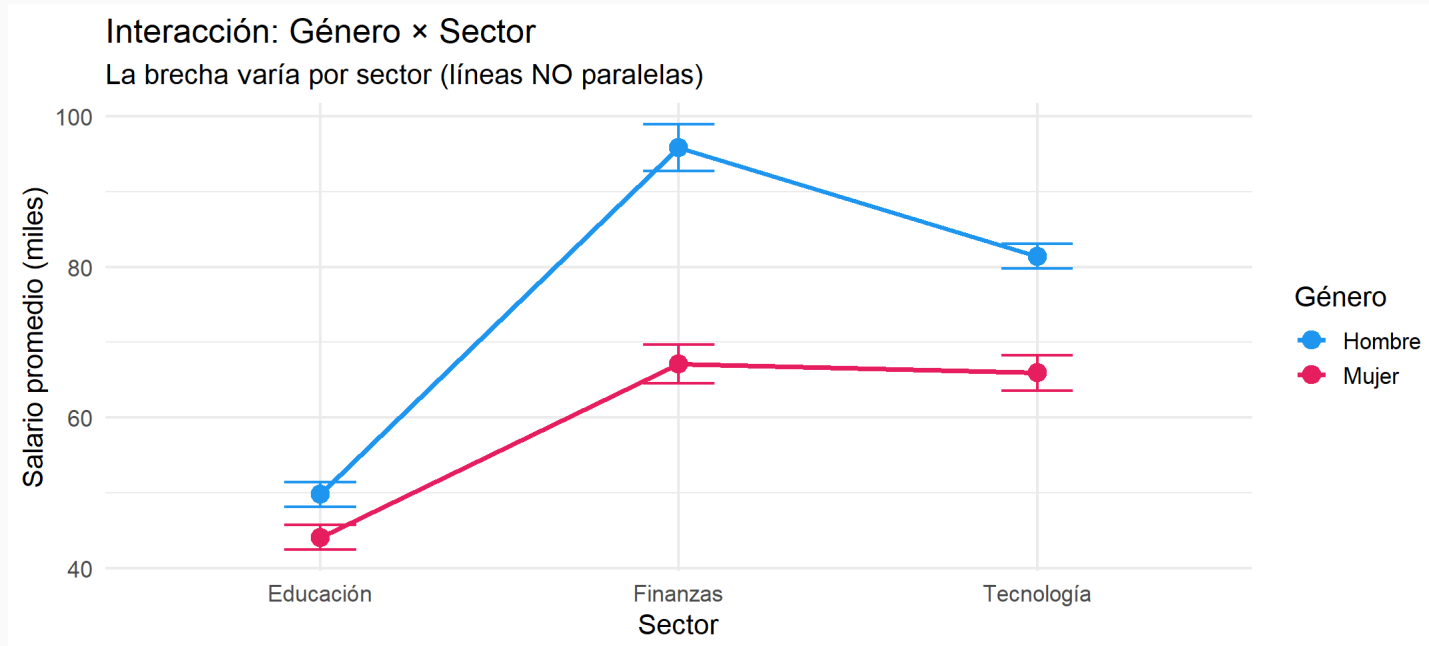
```

¿Entonces por qué usar ANOVA?

1. **Framework conceptual:** Pensar en "grupos" vs "predictores"
2. **Diseño experimental:** Lenguaje natural para A/B/C/D testing
3. **Post-hoc integrado:** Tukey, Bonferroni ya implementados
4. **Más fácil y rápido:** Directamente devuelve si hay diferencias o no las hay, sin depender de cuál es la categoría base.

Two-Way ANOVA: Interacciones

Pregunta: ¿La brecha salarial de género varía por sector?



Two-Way ANOVA en R

```
# ANOVA con dos factores e interacción
```

```
modelo_2way <- aov(salario ~ genero * sector, data = salarios_genero)
summary(modelo_2way)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## genero         1    8286     8286   82.65 3.62e-15 ***
## sector         2   26171    13085  130.52 < 2e-16 ***
## genero:sector   2    2658     1329   13.26 6.67e-06 ***
## Residuals     114   11429      100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretación:

- **genero**: Efecto principal significativo (brecha promedio existe)
- **sector**: Efecto principal significativo (sectores difieren)
- **genero:sector**: Interacción significativa (la brecha **varía** por sector)

Supuestos de ANOVA

ANOVA requiere verificar tres supuestos:

1. **Independencia:** Las observaciones son independientes

- Crítico, violación grave
- Verificar por diseño del estudio

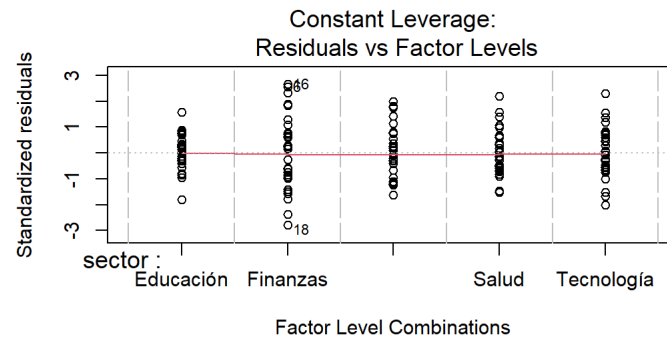
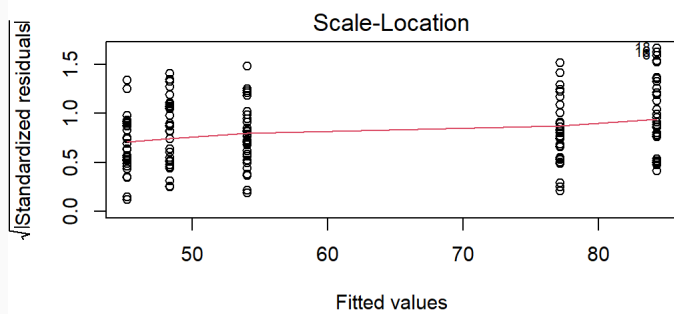
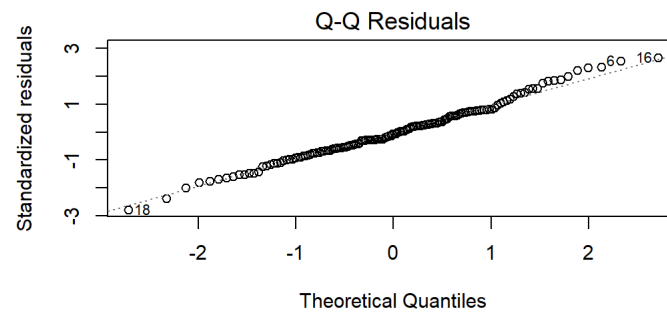
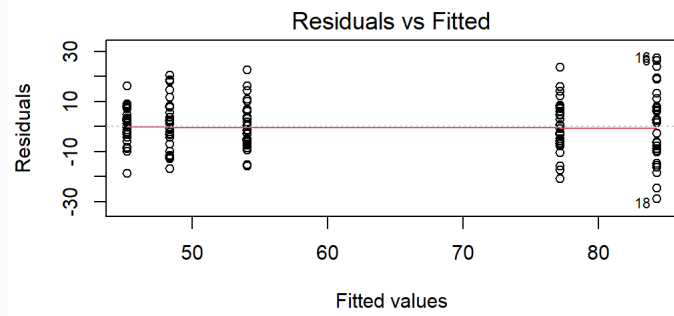
2. **Normalidad:** Los residuos siguen distribución normal

- Test de Shapiro-Wilk, Q-Q plots
- Robusto con n grande (TCL)

3. **Homogeneidad de varianzas:** Varianzas iguales entre grupos

- Test de Levene
- Si se viola: Welch ANOVA

Verificación de supuestos en R



Test de Levene: Homogeneidad de varianzas

```
# Test de Levene
leveneTest(salario ~ sector, data = salarios)

## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group     4  4.5581 0.001697 **
##           145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretación:

- Si $p > 0.05 \rightarrow$ No rechazamos H_0
- Las varianzas son homogéneas
- ANOVA estándar es apropiado

Si se violara ($p < 0.05$): Usar `oneway.test()` con corrección de Welch

```
# Alternativa robusta a heterogeneidad
oneway.test(salario ~ sector, data = salarios, var.equal = FALSE)
```

Alternativa no paramétrica: Kruskal-Wallis

ANOVA no paramétrico: Test de Kruskal-Wallis

```
# Alternativa cuando normalidad no se cumple
kruskal.test(salario ~ sector, data = salarios)

##
##      Kruskal-Wallis rank sum test
##
## data:  salario by sector
## Kruskal-Wallis chi-squared = 105.01, df = 4, p-value < 2.2e-16
```

Basado en rangos, no asume normalidad.

Resumen: ANOVA

Usar ANOVA cuando:

- Comparar 3+ grupos
- Diseño experimental (A/B/C/D testing)
- Se quieren descomponer las varianza explícitamente
- Necesitas comparaciones post-hoc con control de error (ej. para ver qué grupos difieren entre sí)

Pasar a regresión cuando:

- Predictores continuos
- Modelos complejos con muchas variables
- Necesitas regularización o ML

Recordar:

- ANOVA = regresión con dummies
- Verificar supuestos (especialmente con $n < 30$)
- Post-hoc con corrección (Tukey, Bonferroni)

Parte 2: PCA

Principal Component Analysis

Motivación: La maldición de la dimensionalidad

Problema frecuente en economía:

Datos de 30 países con 20 indicadores económicos:

- PIB per cápita, inflación, desempleo, Gini, esperanza de vida...
- Variables correlacionadas entre sí
- Difícil visualizar, interpretar y modelar

Desafíos:

1. **Multicolinealidad:** Variables redundantes generan inestabilidad
2. **Visualización:** Imposible graficar 20 dimensiones
3. **Interpretación:** ¿Qué patrones hay en los datos?
4. **Complejidad:** Modelos con muchos predictores tienden al overfitting

Solución: Reducir dimensiones preservando información

PCA: Idea intuitiva

Análisis de Componentes Principales busca nuevas variables (componentes) que:

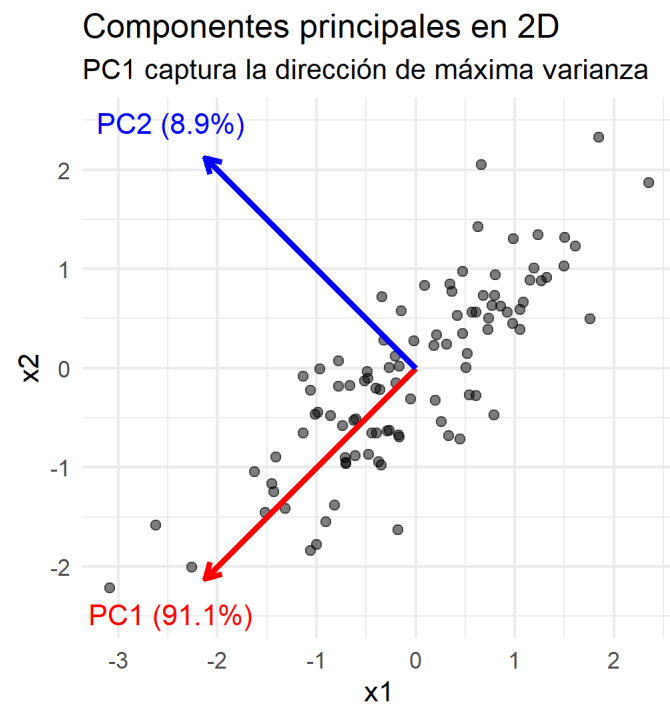
1. Son **combinaciones lineales** de las variables originales
2. Capturan la **máxima varianza** de los datos
3. Son **no correlacionadas** entre sí (ortogonales)
4. Están ordenadas por importancia (PC1 explica más que PC2, etc.)

Analogía: Imaginar una nube de puntos en 3D

- PC1: dirección de mayor dispersión (eje principal)
- PC2: segunda mayor dispersión, perpendicular a PC1
- PC3: tercera mayor dispersión, perpendicular a PC1 y PC2

PCA encuentra automáticamente estos ejes en p dimensiones.

Ejemplo visual: De 2D a 1D



Fundamento matemático

PCA busca una matriz de pesos W que maximiza la varianza de los datos proyectados:

$$\max_{w_1} \text{Var}(Xw_1) = \max_{w_1} w_1^T \Sigma w_1$$

sujeito a $\|w_1\| = 1$

Donde:

- X : matriz de datos centrados ($n \times p$)
- Σ : matriz de covarianza ($p \times p$)
- w_1 : vector de pesos del primer componente ($p \times 1$)

Solución: w_1 es el **eigenvector** asociado al **mayor eigenvalue** de Σ

Los siguientes componentes (w_2, w_3, \dots) son eigenvectors de eigenvalues decrecientes, con ortogonalidad:

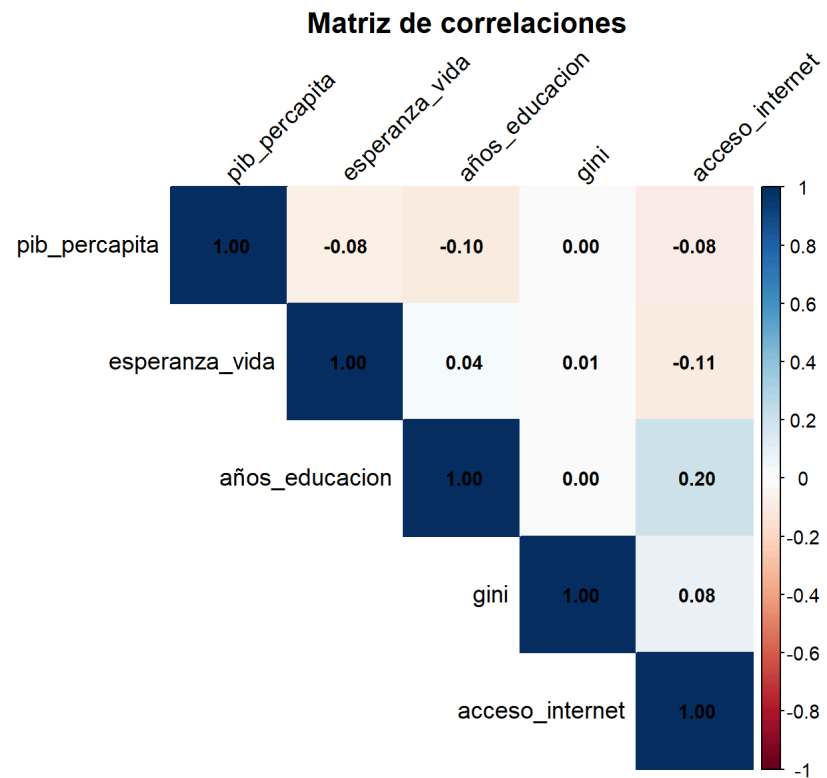
$$w_i^T w_j = 0 \text{ para } i \neq j$$

Implementación en R: Datos de países

```
# Datos de desarrollo económico (simulados)
set.seed(2025)
países <- tibble(
  país = paste("País", 1:50),
  pib_percapita = rnorm(50, mean = 25000, sd = 15000),
  esperanza_vida = rnorm(50, mean = 72, sd = 8),
  años_educacion = rnorm(50, mean = 12, sd = 3),
  gini = rnorm(50, mean = 40, sd = 10),
  acceso_internet = rnorm(50, mean = 60, sd = 25)
) %>%
mutate(across(where(is.numeric), ~pmax(., 0)))
```

país	pib_percapita	esperanza_vida	años_educacion	gini	acceso_internet
País 1	34311.4	75.0	14.1	35.2	55.1
País 2	25534.6	62.2	11.2	38.5	0.0
País 3	36597.3	70.7	15.1	36.6	42.4
País 4	44087.3	80.7	12.7	28.1	57.8
País 5	30564.6	74.9	8.0	37.8	55.6

Matriz de correlaciones



PCA paso a paso

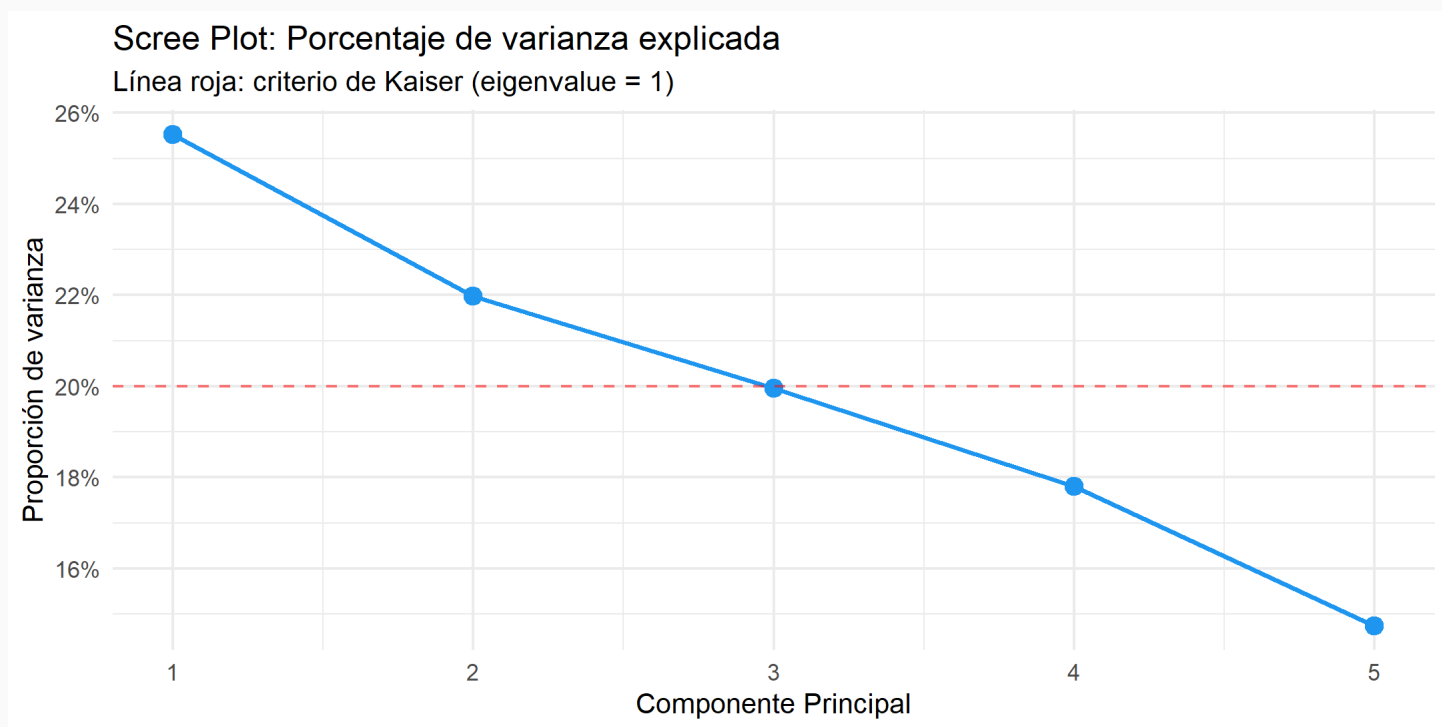
```
# PCA (scale = TRUE estandariza)
datos_pca <- paises %>% select(-pais)
pca_resultado <- prcomp(datos_pca, scale = TRUE)
```

```
# Resumen
summary(pca_resultado)
```

```
## Importance of components:
```

##	PC1	PC2	PC3	PC4	PC5
## Standard deviation	1.1297	1.0481	0.9989	0.9435	0.8587
## Proportion of Variance	0.2552	0.2197	0.1995	0.1780	0.1475
## Cumulative Proportion	0.2552	0.4749	0.6745	0.8525	1.0000

Scree plot: ¿Cuántos componentes retener?



Criterios: (1) Eigenvalue > 1, (2) Codo, (3) Varianza acumulada 70-90%

Loadings: ¿Qué representa cada componente?

Variable	PC1	PC2	PC3
pib_percapita	-0.407	-0.485	-0.062
esperanza_vida	-0.035	0.773	-0.312
años_educacion	0.613	0.133	0.199
gini	0.203	-0.215	-0.926
acceso_internet	0.646	-0.322	0.046

Interpretación de PC1:

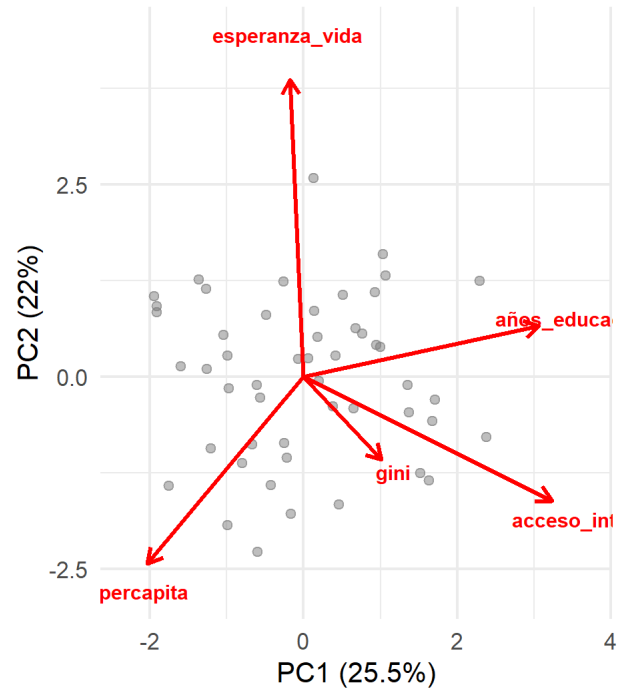
PC1 parece capturar **nivel de desarrollo**:

- Alto PC1 → mayor PIB, mayor esperanza de vida, más educación
- Bajo PC1 → lo opuesto

Biplot: Observaciones + Variables

Biplot: Países en espacio de PC1-PC2





Vectores rojos = variables originales



Tamaño muestral: Limitación crítica

Regla fundamental: $n \geq 5p$ (mínimo), $n \geq 10p$ (ideal)

Donde n = observaciones, p = variables

Ratio n:p	Evaluación	Acción
< 2:1	 Crítico	NO hacer PCA
2:1 - 5:1	 Problemático	PCA con advertencias
5:1 - 10:1	 Aceptable	Validar con KMO
> 10:1	 Ideal	PCA confiable

Además: $n < 50 \rightarrow$ evitar PCA, $n \geq 200 \rightarrow$ zona cómoda

Problema en economía: Frecuentemente tenemos pocos países/años pero muchas variables

Test de adecuación: KMO

Kaiser-Meyer-Olkin: Mide si PCA es apropiado para los datos

```
# Calcular KMO  
kmo_resultado <- KMO(datos_pca)  
print(kmo_resultado$MSA)
```

```
## [1] 0.4944231
```

Interpretación:


- $KMO > 0.9$: Excelente
- $KMO > 0.8$: Bueno
- $KMO > 0.7$: Aceptable
- $KMO > 0.6$: Mediocre
- $KMO < 0.6$: Inaceptable para PCA

Test de Bartlett

```
# Test de esfericidad de Bartlett  
bartlett_test <- cortest.bartlett(cor(datos_pca), n = nrow(datos_pca))  
print(bartlett_test)
```

```
## $chisq  
## [1] 3.972002  
##  
## $p.value  
## [1] 0.9486012  
##  
## $df  
## [1] 10
```

Interpretación:

- H_0 : Matriz de correlación = identidad (variables independientes)
- $p < 0.05 \rightarrow$ Rechazamos $H_0 \rightarrow$ Variables correlacionadas \rightarrow OK para PCA 

Si $p > 0.05 \rightarrow$ Las variables son independientes \rightarrow PCA no tiene sentido

Limitaciones de PCA

1. Asume relaciones lineales

- PCA solo captura correlaciones lineales
- Alternativa: Kernel PCA, UMAP

2. Sensibilidad a outliers

- Outliers extremos distorsionan componentes
- Considerar PCA robusto

3. Interpretación no garantizada

- Los componentes son matemáticos, no necesariamente sustantivos
- Puede no haber interpretación económica clara

Limitaciones de PCA

1. Pérdida de interpretabilidad

- Variables originales tienen significado directo
- Componentes son combinaciones abstractas

2. No es feature selection

- PCA transforma variables, no las selecciona
- Alternativa: Lasso, Random Forest

Casos de uso en Economía

1. Índices compuestos

- Índice de desarrollo humano
- Índice de competitividad nacional
- Índice de riesgo país

2. Reducción dimensional para ML

- Feature engineering antes de clustering
- Evitar multicolinealidad en regresión
- Compresión de datos

Casos de uso en Economía

3. Visualización

- Graficar países en "espacio económico"
- Detectar patrones y outliers
- Identificar grupos naturales

4. Análisis exploratorio

- Entender estructura de correlaciones
- Identificar dimensiones latentes
- Detectar redundancia en variables

Ejemplo aplicado: Crear índice de desarrollo

```
# Usar PC1 como índice (si tiene interpretación válida)
países_con_indice ← países %>%
  mutate(
    indice_desarrollo = pca_resultado$x[, 1],
    indice_norm = scales::rescale(indice_desarrollo, to = c(0, 100))
  )

# Ranking de países
ranking ← países_con_indice %>%
  select(pais, indice_norm) %>%
  arrange(desc(indice_norm)) %>%
  mutate(ranking = row_number(), indice_norm = round(indice_norm, 1))
```

País	Índice (0-100)	Ranking
País 8	100.0	1
País 43	98.0	2
País 31	84.6	3
País 10	83.8	4
País 16	82.8	5

Resumen: PCA

Usar PCA cuando:

- Muchas variables correlacionadas
- $n \geq 5p$ (mínimo), idealmente $n \geq 10p$
- KMO > 0.6, Bartlett significativo
- Objetivo: reducción dimensional, no interpretación forzada

No usar PCA cuando:

- n muy pequeño relativo a p
- Variables no correlacionadas
- Interpretabilidad es crítica
- Necesitas selección de variables

Recordar:

- PCA es matemático, interpretación no garantizada
- Verificar tamaño muestral SIEMPRE
- Validar con KMO y Bartlett
- Considerar alternativas (Lasso, FA) según objetivo

Comparación final: ANOVA vs PCA

Aspecto	ANOVA	PCA
Tipo	Supervisado	No supervisado
Variables	1 continua	Múltiples continuas
Objetivo	Comparar grupos	Reducir dimensiones
Supuestos	Normalidad, homogeneidad	Linealidad, n/p
Output	p-valor, diferencias	Componentes, varianza
Interpretabilidad	Alta	Variable

Son herramientas complementarias: ANOVA para inferencia causal sobre grupos, PCA para exploración y reducción dimensional.

¿Preguntas?

Nicolás Sidicaro - FCE-UBA

Octubre 2025