

## GUIA EXAMEN

### **PREGUNTA 1**

Se estudian 6 características de juego de 35 jugadores de tenis (derecho, revés, servicio, volea, retorno del servicio y el estado psíquico), que corresponden a notas entre 0 y 10. Los resultados del análisis en componentes principales sobre estos datos se encuentran en la tabla 1 y grafico 1.

- Interprete los valores propios adjuntos (tabla 1). Dé las proporciones de la varianza reproducida por cada componente principal. Expresé la primera componente principal en función de las 6 variables y comente. ¿Pueden expresar las 6 variables a partir de las componentes principales? ¿Cómo?
- A partir de las correlaciones adjuntas (tabla 1), haga un gráfico de las variables sobre las 2 primeras componentes principales. Interprete el gráfico. Deduzca los coeficientes de correlación aproximados entre las 6 variables.
- Interprete el gráfico 1. En particular, en que difieren Connors, Pecci, Solomon y Mc Enroe.
- Un nuevo jugador VILAS tiene como valores (centradas y reducidas) para las 6 variables: 1.1418 0.8038 -0.6381 -0.9754 0.8507 0.9692. Cuales son sus coordenadas en el plano principal de los 2 primeros factores. Describe su juego y ¿a quien se parece su juego?
- Se quiere hacer la regresión lineal de una nueva variable el "Smash" sobre las dos primeras componentes principales sabiendo que sus correlaciones con las dos primeras C. P. son 0.1982 y -0.9022. Dé los coeficientes de la regresión sabiendo que la desviación estándar de la variable "Smash" es 1.9462.
- Dé el coeficiente de correlación múltiple. ¿Este último aproxima bien el coeficiente de correlación múltiple de "Smash" sobre las 6 otras variables?

Tabla 1: Correlaciones entre variables antiguas y componentes principales

	Componentes principales					
	1	2	3	4	5	6
Valor propio	2.935	1.9598	0.4557	0.3547	0.1657	0.1290
Derecho	0.7962	0.2659	-0.4014	0.3596	-0.0687	-0.0134
Revés	0.9162	-0.0525	-0.07	-0.3095	-0.0821	0.2246
Servicio	0.0104	-0.9448	-0.1438	0.1124	0.2533	0.0993
Volea	-0.1033	-0.9422	-0.1379	-0.0799	-0.2490	-0.1189
Retorno servicio	0.9348	-0.0042	-0.0095	-0.2175	0.1569	-0.2328
Estado psíquico	0.7597	-0.3254	0.5	0.2515	-0.0594	0.0120

### PAUTA

- Los 6 valores propios suman 6 (la matriz R es de rango a lo más 6). Las dos primeras componentes principales reproducen 81,58% de la varianza.

	Componentes principales					
	1	2	3	4	5	6
Valor propio	2.935	1.9598	0.4557	0.3547	0.1657	0.1290
% acumulado	48,92	81,58	89,18	95,09	97,85	100,00

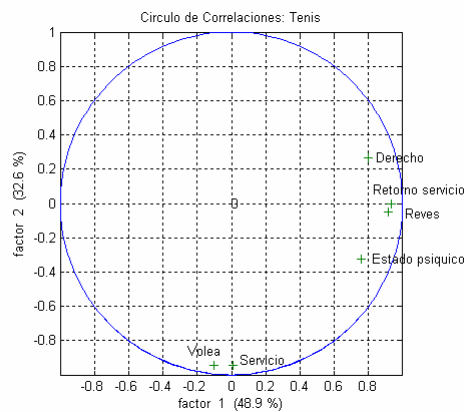
$$c_1 = \frac{0.79}{\sqrt{2.935}} x_1 + \frac{0.92}{\sqrt{2.935}} x_2 + \dots + \frac{0.759}{\sqrt{2.935}}$$

La primera C.P. tiene que ver con el derecho, revés, retorneo servicio y estado físico, mientras que la 2da con Servicio y Volea, es decir fuerza. Cada variable puede expresarse a partir de las 6 C. P.:

$$x_1 = \frac{0.79}{\sqrt{2.935}} c_1 + \frac{0.266}{\sqrt{1.96}} c_2 + \dots + \frac{-0.0134}{\sqrt{0.129}} c_6$$

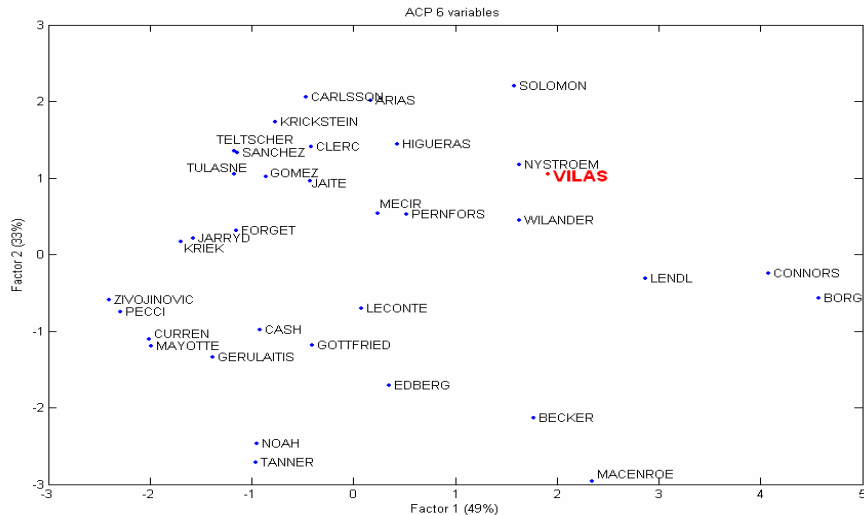
b) El círculo de correlaciones permite: Interpretar las componentes principales, ver la representatividad de las antiguas variables en los planos factoriales y representar la matriz de correlación R de las antiguas variables.

Los valores aproximados son los cosenos de los ángulos entre las variables en el círculo.



c) En el gráfico 1 se puede interpretar las proximidades entre los jugadores. En particular Connors tiene un muy buen derecho, revés, retorno de servicio y estado psíquico pero mediano en servicio y volea. Para Solomon es mediano salvo que es malo para el servicio y volea. McEnroe es bastante bueno en derecho, revés, retorno servicio y estado físico y muy bueno en servicio y volea. Finalmente Pecci es mediano en servicio y volea y malo en el resto.

d) VILAS: 1.9094 1.0460; Se parece a NYSTROEM.



- e) Los coeficientes de la regresión de "Smash" sobre las 2 primeras C.P. son los coeficientes de correlación divididos por la raíz del valor propio y multiplica par la desviación estándar de "Smash":

$$0.1982 * \sqrt{1.9462} / \sqrt{2.935} = 0.2252 \quad y$$

$$0.1982 * \sqrt{1.9462} / \sqrt{2.935} = 0.2252 .$$

- f) El coeficiente de correlación múltiple es:

$$\sqrt{0.1982^2 + 0.9022^2} = \sqrt{0.8532} = 0.9237 .$$

Las dos primeras componentes principales reproducen 81,58% de la varianza, 0.9237 aproxima probablemente bien el coeficiente de correlación múltiple de "Smash" sobre las 6 otras variables. ( A título indicativo, el valor real es: 0.96).

## PREGUNTA 2

Sean  $X$  una matriz de datos ( $n=200$  observaciones y  $p=4$  variables  $X_1, X_2, X_3, X_4$ ) que se supondrán centrados y reducidos. Sea  $R$  la matriz de correlación.

- 1.1 Dibuje el círculo de correlaciones a partir de la tabla 1.1. ¿Rango de la matriz  $R$ ?
- 1.2 El círculo de correlaciones tiene 3 funciones. Cítelas.
- 1.3 Se tiene una quinta variable  $Z$  cuyas correlaciones con los 3 factores son respectivamente: 0.75, 0.2 y 0.5. Dé el coeficiente de correlación múltiple de  $Z$  sobre las cuatro variables  $X_1, X_2, X_3, X_4$ .
- 1.4 Deduzca los coeficientes de la regresión lineal de  $Z$  (centrada y reducida) sobre las cuatro variables  $X_1, X_2, X_3, X_4$  (¡OJO! Con la varianza de los factores).

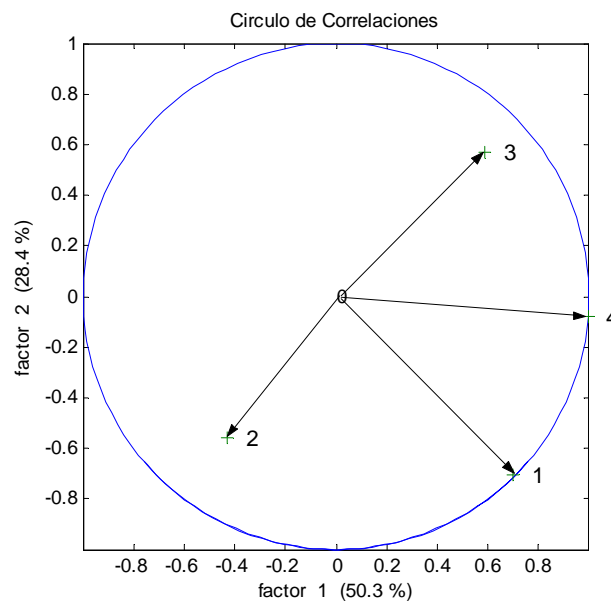
Tabla 1.1

	Factor 1	Factor 2	Factor 3
Valor propio	2.0145	1.1372	0.8483
Vector propio	0.4944	-0.6609	0.1134
	-0.3012	-0.5233	-0.7722
	0.4152	0.5327	-0.6238
	0.7018	-0.0742	-0.0423

## PAUTA

1.1 Se obtienen las correlaciones con  $\sqrt{\lambda_j} u_j$ :

Correlación	C1	C2	C3
Valor propio	2.0145	1.1372	0.8483
X1	0.7017	-0.7048	0.1045
X2	-0.4275	-0.5581	-0.7112
X3	0.5893	0.5681	-0.5745
X4	0.9961	-0.0791	-0.0389



Los 3 valores propios suman 4. La matriz R es de rango 3.

- 1.2 El circulo de correlaciones permiten:
- Interpretar las componentes principales.
  - Ver la representatividad de las antiguas variables en los planos factoriales.
  - Representar la matriz de correlación R de las antiguas variables.
- 1.3 El coeficiente de correlación múltiple es:  $\sqrt{0.75^2 + 0.2^2 + 0.5^2} = 0.8732$ .
- 1.4 Los coeficientes de la regresión de Z sobre los 3 factores son los coeficientes de correlación divididos por la norma de los factores si estos se obtienen como:

$$C_j = Xu_j = \sum_k u_{jk} X_k$$

$$Z = \frac{0.75}{\sqrt{\lambda_1}} C_1 + \frac{0.2}{\sqrt{\lambda_2}} C_2 + \frac{0.5}{\sqrt{\lambda_3}} C_3. \text{ Luego } Z = \frac{0.75}{\sqrt{\lambda_1}} Xu_1 + \frac{0.2}{\sqrt{\lambda_2}} Xu_2 + \frac{0.5}{\sqrt{\lambda_3}} Xu_3$$

$$\text{o sea } Z = \frac{0.75}{\sqrt{\lambda_1}} \left( \sum_k u_{1k} X_k \right) + \frac{0.2}{\sqrt{\lambda_2}} \left( \sum_k u_{2k} X_k \right) + \frac{0.5}{\sqrt{\lambda_3}} \left( \sum_k u_{3k} X_k \right)$$

$$Z = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \text{ con } \beta_k = \sum_j \frac{\text{cor}(C_j, Z)}{\sqrt{\lambda_j}} u_{jk}$$

$$Z = 0.1347X_1 - 0.6593X_2 - 0.1194X_3 + 0.2233X_4$$

### PROBLEMA 3

Un instituto agrícola quiere comparar el efecto de dos fertilizantes  $F_1$  y  $F_2$  sobre el rendimiento del cultivo de trigo. Con este propósito, diseña un experimento con tres grupos de parcelas: un grupo control sin fertilizante, un grupo con el fertilizante  $F_1$  y un grupo con el fertilizante  $F_2$ . En la tabla 1 son resumidos los resultados de la cosecha de trigo por unidad de superficie.

Tabla 1

Grupos	Media	Desviación típica	Frecuencia
Grupo control	4.8450	2.8409	120
Grupo $F_1$	5.3345	2.8964	80
Grupo $F_2$	9.0639	2.9386	75
Total	6.1380	3.4087	275

- 1.1 Construye la tabla ANOVA que permite decidir si se observan diferencias en el rendimiento de trigo entre los tres grupos. Interprete los resultados.
- 1.2 Realice los tres tests de comparación de medias sobre el rendimiento, considerando los tres pares de grupos. Precise los supuestos que hizo y las hipótesis planteadas. Concluye.
- 1.3 Si no cambian las medias y las desviaciones típicas de los dos primeros grupos y el tamaño del grupo control, como hay que modificar el tamaño del grupo  $F_1$  para que cambie el resultado del test de comparación del grupo control con el grupo  $F_1$ .

### PAUTA

- 1.1 La suma de los cuadrados debido a los grupos es  $270 \times \text{varianza intragrupos}$ :

$$120 * 2.8409^2 + 80 * 2.8964^2 + 75 * 2.9386^2 = 2312.615$$

La suma de los cuadrados debido a los residuos es  $270 \times \text{varianza intergrupos}$ :

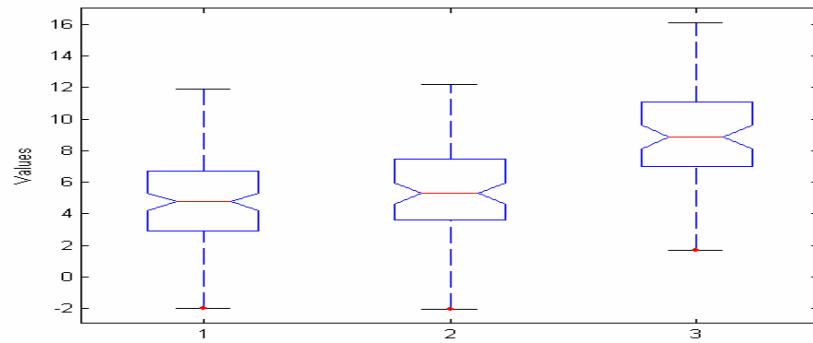
$$120 * (4.8450 - 6.1380)^2 + 80 * (5.3345 - 6.1380)^2 + 75 * (9.0639 - 6.1380)^2 = 894.346$$

Tabla ANOVA

Fuente de variación	Grados de libertad	Suma cuadrados	Cuadrado Medio	F	P_valor
Grupos	2	894.3459	447.1729	52.5946	0.000
Residuos	272	2312.615	8.5023		
Total	274	3206.9609			

Se concluye que hay diferencia entre los tres grupos.

1.2



Grupos	Hipótesis	Diferencia medias	Desv. Típica de los dos grupos	Grados de libertad	t	P_valor
Control / F <sub>1</sub>	H <sub>0</sub> : m <sub>0</sub> =m <sub>1</sub> H <sub>1</sub> : m <sub>0</sub> <m <sub>1</sub>	-0.4896	2.8922	198	-1.1728	0.1211
Control / F <sub>2</sub>	H <sub>0</sub> : m <sub>0</sub> =m <sub>2</sub> H <sub>1</sub> : m <sub>0</sub> <m <sub>2</sub>	-4.2189	2.9088	193	-9.8535	0.000
F <sub>1</sub> / F <sub>2</sub>	H <sub>0</sub> : m <sub>1</sub> =m <sub>2</sub> H <sub>1</sub> : m <sub>1</sub> <m <sub>2</sub>	-3.7294	2.9550	153	-7.8521	0.000

Las medias del grupo control con el grupo F<sub>1</sub> son significativamente diferentes. Las otras lo son. Este resultado está validado con el boxplot.

Los supuestos:

- Se asume la normalidad
- Se supone que la varianza en cada grupo es la misma

3.3 Basta aumentar suficientemente el tamaño del grupo F<sub>1</sub> para que el p-valor disminuya.

$$t = -0.4896 / (2.8922 * \sqrt{\frac{1}{120} + \frac{1}{80}}) = -1.1728$$

Por ejemplo con una muestra de 500 para el grupo F<sub>1</sub> obtenemos un p\_valor de 5%:

#### Problema 4

Un médico que realiza delicadas y costosas operaciones quirúrgicas ha oído hablar de un método estadístico que le permitiría estimar el tiempo de supervivencia de un paciente después de operarse. El cuenta con los resultados de los exámenes de los tests de hígado, enzima, coagulación y presión arterial, realizados a pacientes antes de ser operados. Además del tiempo de vida después de la operación de los mismos pacientes.

Se plantea el modelo lineal: *tiempo supervivencia* =  $\beta_0 + \beta_1 \text{higado}$  (1)

- 2.1 Encuentre la expresión de los estimadores  $\hat{\beta}_0$  y  $\hat{\beta}_1$  de  $\beta_0$  y  $\beta_1$  respectivamente en función de los datos con los que cuenta el médico. (Hint: llame X e Y a las variables involucradas).
- 2.2 Al realizar la matriz de correlaciones de las variables se obtiene la tabla 2.
- ¿Dé la fórmula de cálculo de los coeficientes de la matriz?
  - ¿Qué rango de valores puede tomar cada coeficiente? ¿Cómo se interpreta?
  - ¿Qué cuidados se deben tener al interpretar cada coeficiente de la matriz?
  - Si usted pudiera observar solo un examen de un paciente y con este resultado predecir su tiempo de sobrevivencia ¿Cuál examen de los 4 exámenes utilizaría?

Tabla 2

Correlación	Coagulación	Presión	Enzima	Hígado	Tiempo
Coagulación	1.00000	0.09012	-0.14963	0.50242	0.37252
Presión	0.09012	1.00000	-0.02361	0.36903	0.55398
Enzima	-0.14963	-0.02361	1.00000	0.41642	0.58024
Hígado	0.50242	0.36903	0.41642	1.00000	0.72233
Tiempo	0.37252	0.55398	0.58024	0.72233	1.00000

- 2.3 Se plantea el siguiente modelo lineal (2):

$$\text{tiempo} = \beta_0 + \beta_1 \text{coagulación} + \beta_2 \text{presion} + \beta_3 \text{enzima} + \beta_4 \text{higado} \quad (2)$$

Obteniéndose al efectuar la regresión un  $R^2=0.8367$  y los resultados en las tablas 3 y 4. Escriba el modelo (2) en función de los estimadores  $\hat{\beta}_i$ . Interprete los valores obtenidos y justifique bien sus interpretaciones.

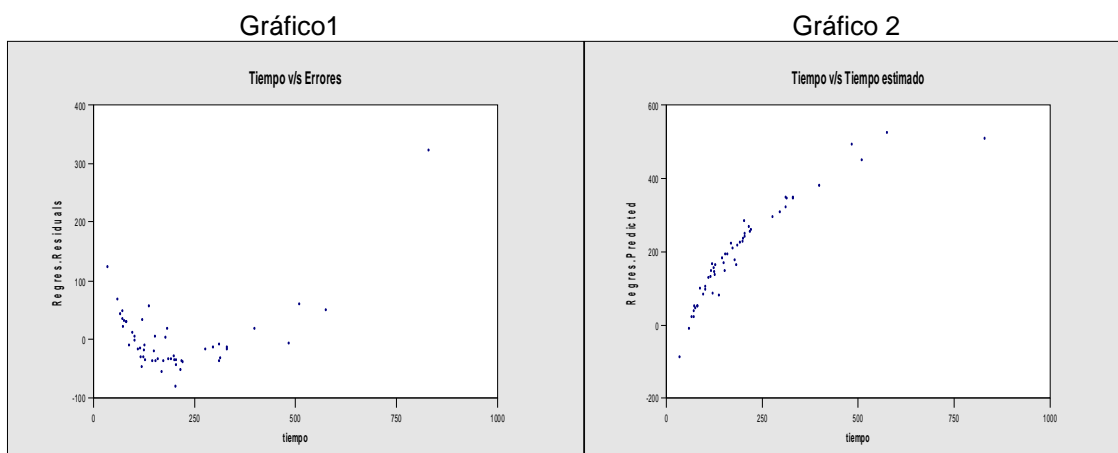
Tabla 3

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p_valor
Regresión	4	936264.538	234066.134	.787712	0.00000
Residuos	49	182666.962	3727.89719		
Total	53	1118931.5			

Tabla 4

Variable	Estimación parámetro	Error estándar	t Student	P_valor
Coagulación	33.164	7.017	4.726	0.000
Presión	4.272	0.563	7.582	0.000
Enzima	4.126	0.511	8.071	0.000
Hígado	14.092	12.525	1.125	0.266
Constante	-621.598	64.800	-9.592	0.000

- 2.4 ¿Su interpretación anterior contradice su respuesta de la parte 2.2(d)? ¿Cómo se explica que exista (o no exista) esta contradicción?
- 2.5 Al realizar el gráfico de los residuos y estimaciones de la regresión se obtienen los gráficos 1 y 2. Interprete los resultados obtenidos en función de los supuestos usuales del modelo lineal.



### Problema 5

Se hizo un estudio sobre el nivel de desarrollo poblacional en el uso de las tecnologías de información, para lo cual se recolectó la siguiente información para 10 países:

Países	Area (Km2)	Población (millones)	Tasa Desempleo	Computadores por persona
E.E.U.U.	9629091	278,1	4,0	5,40
Namibia	825418	1,8	35,0	0,03
Francia	547030	59,6	9,7	0,89
Luxemburgo	2586	0,4	2,7	0,66
Finlandia	337030	5,2	9,8	1,20
Laos	236800	5,6	5,7	0,01
Chile	756950	15,3	8,0	0,56
Zimbawe	390580	11,4	50,0	0,03
Japan	377835	126,8	4,7	6,20
Kenia	582650	30,8	50,0	0,02

Para el estudio se utilizó un Análisis de Componentes Principales (ACP), del cual se presentan resultados en las tablas 5, 6 y 7):

Tabla 5

	Media	Desviación Estándar
Area	1368597,000	2763092,637
Población	53,504	83,497
Desempleo	17,960	18,252
Computadores	1,500	2,193

Tabla 6: Matriz de correlaciones

	Área	Población	Desempleo	Computadores
Área	1	0,895	-0,222	0,581
Población	0,895	1	-0,324	0,849
Desempleo	-0,222	-0,324	1	-0,471
Computadores	0,581	0,849	-0,471	1



Tabla 7: Valores y vectores propios normalizados

	$\lambda_1$	$\lambda_2$	$\lambda_3$	$\lambda_4$
Valor propio	2,753	0,874	0,353	0,020
	U1	U2	U3	U4
Área	0,517	0,393	-0,595	-0,473
Población	0,584	0,236	0,047	0,775
Desempleo	-0,318	0,881	0,346	-0,050
Computadores	0,539	-0,113	0,724	-0,415

- 3.1 Explique en qué consiste el análisis de componentes principales y cuáles son sus principales aplicaciones.
- 3.2 Interprete los resultados de las tablas 5 y 7.
- 3.3 Dibuje el círculo de correlaciones para los dos primeros ejes principales e interprete el resultado. Explique porqué es conveniente quedarse solo con estos.
- 3.4 Encuentre las componentes principales de cada país asociadas a los primeros ejes principales y grafique aproximadamente. Interprete los resultados.
- 3.5 Determine las contribuciones porcentuales de cada variable original en la construcción del eje.

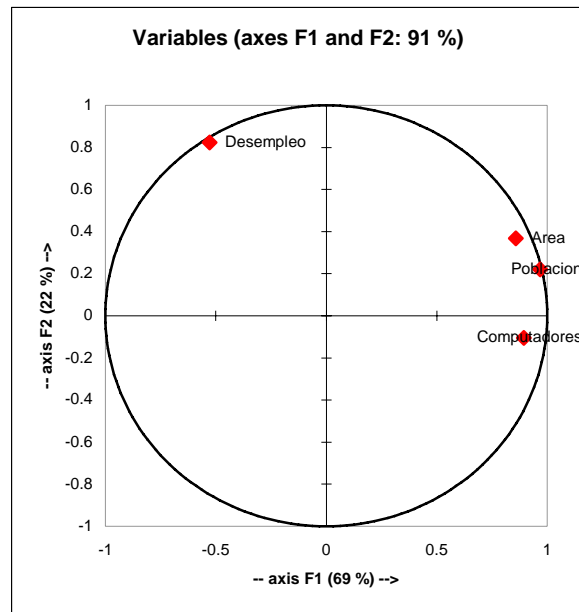
## PAUTA

Es técnica que permite describir un conjunto de información en función de un conjunto menor de variables no correlacionadas, tratando de preservar la mayor cantidad de variabilidad de los datos originales. Dentro de las principales aplicaciones del ACP, se encuentran: construcción de índices compuestos, reducción de multicolinealidad en modelos lineales y clasificación estadística entre otras.

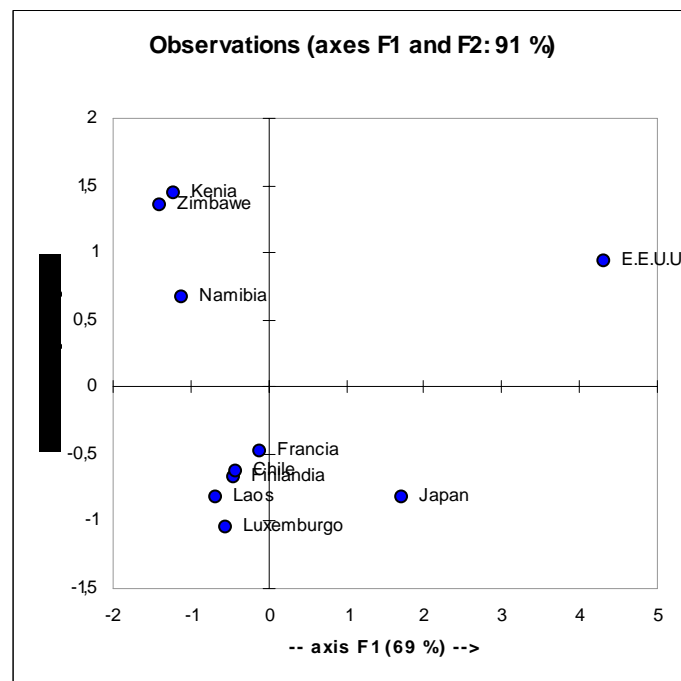
La tabla 5 muestra los valores promedio y desviaciones estándar de las variables originales, a partir de la cual se puede observar que los datos difieren bastante en escala, por lo que es importante trabajar con los datos estandarizados, y utilizar la matriz de correlaciones para llevar a cabo el ACP

Coordenadas de las variables en el círculo:

	U1	U2
Area	0,858	0,368
Poblacion	0,969	0,221
Desempleo	-0,528	0,824
Computadores	0,895	-0,105



Se puede observar que las variables Área, Población y Computadores están más vinculadas al primer eje principal, al cual podemos denominar tentativamente como “cobertura tecnológica”, mientras que el segundo eje está más relacionado al aspecto económico de los países, eje que podría denominarse “capacidad de trabajo” (leído en sentido inverso, es decir, de arriba hacia abajo). Las variables originales se encuentran cerca de la frontera del círculo, lo que les confiere un alto poder explicativo en la construcción de los ejes, es decir, se espera un elevado coeficiente de determinación. Debería esperarse que los países más pobres se encuentren cerca del II cuadrante, mientras que los países más desarrollados deberían encontrarse cerca del IV cuadrante.



El gráfico muestra precisamente lo que se intuye del círculo de correlaciones. Países como Japón y E.E.U.U. son los países con cobertura tecnológica de información (están más a la derecha en el primer eje principal), mientras que los países africanos tienen la menor puntuación, por ser los más poblados, con mayor desempleo y menor cantidad de computadores por persona, por ende tiene menor cobertura tecnológica y menor capacidad de trabajo. Ahora, E.E.U.U. se encuentra en el I cuadrante básicamente porque es un país altamente poblado y con mucha área geográfica. Los países intermedios como Chile, Francia y Luxemburgo tienen menor cobertura tecnológica que E.E.U.U. y Japón básicamente porque tienen una menor población y área geográfica; al igual que E.E.U.U. se diferencia de Japón por el mismo motivo.

Cosenos cuadrados de las variables:

	F1	F2
Area	0,735	0,135
Población	0,938	0,049
Desempleo	0,279	0,679
Computadores	0,801	0,011

Contribución (%):

	F1	F2
Area	26,711	15,460
Población	34,082	5,592
Desempleo	10,125	77,677
Computadores	29,082	1,272

Los cuadrados de las coordenadas de las variables originales en el plano principal determinan el aporte para cada eje, del cual se pueden deducir los aportes porcentuales de las mismas. Se puede observar que la población y el número de computadores son los principales responsables del valor obtenido en cobertura tecnológica, lo cual es bastante esperable desde el punto de vista intuitivo.

## PROBLEMA 6

Se consideran el pulso de **200** pacientes antes y después de un ejercicio físico. Llamemos **x** al pulso antes del ejercicio e **y** al pulso después.

1.1 Suponiendo que **x** e **y** siguen distribuciones normales, dé la distribución de la diferencia: **d=y-x**.

1.2 Estime los parámetros de la distribución utilizando los datos de la tabla 1.

1.3 Construya un intervalo de confianza al 95% para la diferencia media  $\delta=E(d)$ . Interprete.

- 1.4 Efectué un test de hipótesis para  $H_0 : \delta = 0$  contra  $H_1 : \delta > 0$ . ¿Cambia significativamente el pulso?.
- 1.5 Grafique la función de potencia con una región crítica de 5% utilizando los valores  $\delta=5, 10, 15, 20, 25$ .

Tabla1

	Media	Desviación estándar	Correlación $r$
ANTES	70	4.6	0.61
DESPUÉS	75	6.6	

## PAUTA

- 1.1 La distribución de la diferencia es  $N(\delta, \sigma_d^2)$  con  $\delta = E(y) - E(x)$  y

$$\sigma_d^2 = \sigma_x^2 + \sigma_y^2 - 2 \text{cov}(x, y).$$

- 1.2  $\hat{\delta} = \bar{d} = \bar{y} - \bar{x} = 5$  y  $\hat{\sigma}_x = 4.6$ ,  $\hat{\sigma}_y = 6.6$ .

$$\text{Como } \text{cov}(x, y) = \text{cor}(x, y) * \sigma_x \sigma_y,$$

$$\text{cov}(x, y) = r * \hat{\sigma}_x \hat{\sigma}_y = 0.61 * 4.6 * 6.6 = 18.70$$

$$\hat{\sigma}_d = \sqrt{\hat{\sigma}_x^2 + \hat{\sigma}_y^2 - 2 \text{cov}(x, y)} = \sqrt{4.6^2 + 6.6^2 - 2 * 18.70} = 5.23.$$

- 1.3  $[\bar{d} - 1.96 \hat{\sigma}_d, \bar{d} + 1.96 \hat{\sigma}_d] = [-5.25, 15.25]$ . Este intervalo cubre el 0, lo que hace pensar que los pulsos promedio no son significativamente diferentes.

- 1.4 Para las hipótesis  $H_0 : \delta = 0$  contra  $H_1 : \delta > 0$ , la región crítica es de la forma:

$$\bar{d} > c \text{ con } P(\bar{d} > c | H_0) = \alpha \text{ o bien el p-valor } P(\bar{d} > 5 | H_0) = p.$$

$$\text{Como } \bar{d} \sim N(\delta, \frac{\sigma_d^2}{200}) \text{ y } \frac{200 \hat{\sigma}_d^2}{\sigma_d^2} \sim \chi_{199}^2, \text{ bajo } H_0 \text{ } \frac{\bar{d}}{\hat{\sigma}_d} \sim t_{199}. \text{ La región crítica para}$$

$$\alpha = 5\% \text{ es } \frac{\bar{d}}{\hat{\sigma}_d} > 1.64. \text{ Como en la muestra } \frac{\bar{d}}{\hat{\sigma}_d} = 0.956, \text{ no se rechaza que}$$

$$H_0 : \delta = 0. \text{ Esto está confirmado por el p-valor que es alto:}$$

$$P(t_{199} > 0.95) = 0.17.$$

- 1.5 La función de potencia es:

$$P(\bar{d} > 1.64 \hat{\sigma}_d | \delta) = P\left(\frac{\bar{d} - \delta}{\hat{\sigma}_d} > 1.64\right) = P\left(t_{199} > 1.64 - \frac{\delta}{\hat{\sigma}_d}\right).$$

$\delta$	5	10	15	20	25
$P(t_{199} > 1.64 - \frac{\delta}{\hat{\sigma}_d})$	0.25	0.61	0.89	0.98	0.999

