1. **Un outlier es un valor que supera en dos desvíos standard a la media.**

**FALSO**

Justificación 1: Los outliers moderados quedan de 1,5 a 3 distancias intercuartil del 1er. O 3er. cuartil y los severos mas de 3 distancias intercualtil.

Justificacion 2: En caso de haber outliers la media no es representativa por lo cual no tiene sentido usar la media.

1. **Si una base de datos tiene dimensión nxp, la matriz de varianzas y covarianzas tiene dimensión pxp**

**VERDADERO**

Si dijese n X n se debería contestar que no depende del total de observaciones sino del total de variables.

1. **El ACP es una técnica multivariada que se aplica para reducir la dimensión del problema**

**VERDADERO**

1. **El vector de medias y la matriz de varianzas y covarianzas resumen toda la información disponible en un conjunto de datos.**

**FALSO**

Justificación: Son importantes pero no alcanzan. Faltan si hay muchos o pocos outliers

1. **El AC representa las distancias euclídeas entre los perfiles observados.**

**FALSO:**

Justificación: Representa la distancia Chi Cuadrado que es la distancia euclidea entre los perfiles estandarizados.

1. **El ACP se basa en el análisis de los autovalores de la matriz de datos.**

**FALSO**

Justificación: Se basa en la matriz de Correlación que tiene los datos estandarizados.

1. **Para determinar el número de componentes a considerar existen sólo dos criterios: el bastón roto y el % de variablidad explicada**

**FALSO:** Criterio de Kaiser y Test de esfericidad de Barlett

1. **La inercia cuantifica el grado de adecuación del AC a este conjunto de datos.**

**FALSO**: Es decir que mide la distancia entre los perfiles fila (o columna) observados y los perfiles fila (o columna) esperados bajo la hipótesis de independencia. La inercia es el cociente entre el estadístico Chi Cuadrado y el total de observaciones.

1. **Si las variables son independientes no resulta adecuado aplicar ACP.**

**VERDADERO**

1. **El test de Chi cuadrado puede aplicarse tanto para homogeneidad como para independencia en cualquier conjunto de observaciones.**

**FALSO**

Justificación: No es para cualquier conjunto de datos porque se tienen que cumplir los supuestos.

2. Responda brevemente a las siguientes preguntas:

**a)** **¿En qué se diferencian los test de homogeneidad e independencia.**

En un test de independencia se selecciona una muestra en una sola población para estudiar en ella 2 o más variables categóricas o continuas

El test de homogeneidad se usa para comparar una variables en 2 o más poblaciones.

**b) ¿Para qué se utiliza la distancia Chi cuadrado? ¿Por qué se elige esa distancia?**

La distancia Chi Cuadrado es la distancia entre los perfiles. Es la distancia euclidea entre los perfiles estandarizados

Sirve para neutralizar la diferencia de los totales filas y totales de columna: cuando esta más representada una fila/columna que otra se estandariza por total fila/columna. Se usa Chi Cuadrado porque tiene una propiedad muy importante que es el **principio de equivalencia distribucional**, que implica que si dos filas tienen la misma estructura y las colapsamos en una nueva fila, las distancias entre las restantes filas permanecen invariables.

**c) ¿Para qué sirven los residuos en un test de Independencia?**

Cuando la hipótesis nula se rechaza, debe suponerse que las variables son **dependientes** pero no se sabe **en qué sentido están asociadas**.

Si deseamos indagar al respecto se puede estudiar los residuos del modelo para estudiar qué tipo de dependencia existe entre ellas. Cuando un residuo tiene un valor absoluto superior a 2 en una cela hay que prestar atención a esa asociación porque esa casilla tiene un aporte importante al Chi Cuadrado.

**d) ¿Qué puede observarse en un biplot?**

Un Biplot es una representación gráfica de datos multivariantes

Es un gráfico en el que se representan simultáneamente las variables y los valores de cada individuo en pares de componentes principales

El nivel de asociación lo muestra el angulo que forman los vectores que la representan: menor angulo gran asociación

La longitud de los vectores de las variables indicaban la variabilidad de cada una

La Proyeccion que tengan sobre los ejes implica como son los loadings de esa variable en cada componente (hacia el lado negativo, de forma o hacia el lado positivo, son tamaño)

Permite observar si los individuos respecto a los componentes están muy posicionados arriba o abajo

Cuando los vectores de 2 variables son perpendiculares, no están correlacionadas.

Los individuos que están mas cerca del centro son los mas cercanos al promedio

Los individuos que están en los extremos son opuestos con respecto a las componentes

**e) ¿Con qué propósito se realizan transformaciones por fila?**

Se aplican con el objeto de hacer comparables los valores de los distintos individuos, neutralizando tendencia muy extremas (ej: juez)

f) ¿Qué propiedades son importantes de la matriz de varianzas y covarianzas?¿ Y de la matriz de correlación?

La matriz de covarianzas es simétrica

La matriz de covarianzas es semidefinida positiva, es decir que todos sus autovalores son mayores o iguales a cero.

Luego la traza que es la suma de los autovalores indica de alguna forma la magnitud del problema.

Correlacion:

la traza nos da la cantidad de variables involucradas en el problema.

El modulo de cada componente es menor/igual 1

Si rik = 1 significa que los datos caen sobre una línea recta de pendiente positiva.

c) Si rik = -1 significa que los datos caen sobre una línea recta de pendiente negativa.

d) Si 0 < rik < 1 significa que los datos caen alrededor de una línea recta de pendiente

positiva.

e) Si -1 < rik < 0 significa que los datos caen alrededor de una línea recta de pendiente

negativa.

f ) Si rik = 0 indica que no hay una asociación lineal entre los datos de las variables.

g) ¿Es correcto eliminar los registros que se consideran outliers?

Se los debe analizar y solo se descartan los que se comprueban que son errores de tipeo o de medición

h) ¿Que cuantifica el cociente entre un autovalor y la traza de la matriz de covarianzas en ACP?

La proporción de la variabilidad que capta cada componente.

1. ¿Con qué propósito se realizan transformaciones por columna?

Se utilizan para hacer comparables distintas variables que serán usadas en análisis multivariados posteriores. Un ejemplo muy usual es la estandarización de las variables para neutralizar las distintas unidades de las variables. **Estas transformaciones tienen sentido en el caso en que la media y el desvío resulten una buena representación de la centralidad y la dispersión respectivamente.**

En caso contrario pueden considerarse en forma alternativa la mediana y la desviación intercuartil o la mediana y el MAD.