Análisis de datos (general)

Verdadero o Falso

- ***Dada una matriz de datos de 1800 registros x 7 variables cada uno, la matriz de varianzas y covarianzas tiene 49 valores a estimar.***
  FALSO. Tendría 28 porque es una matriz simétrica entonces $X_{ij}=X_{ji}$ (7+6+5+4++3+2+1, media matriz incluida la diagonal).

- ***Se dice que un valor es un outlier, cuando se encuentra a más de 2 desvíos estándar de la media de la variable.***
  FALSO. Los outliers moderados quedan de 1.5 a 3 *distancias intercuartil* del primero o tercer cuartil y los severos más de 3 distancias intercuartil. Además, en caso de haber outliers, la media no es representativa por lo cual no tiene sentido usarla.

- ***Si una base de datos tiene dimensión n x p, la matriz de varianzas y covarianzas tiene dimensión p x p.***
  VERDADERO. Si dijese n x n se debería contestar que no depende del total de observaciones sino del total de variables.

- ***El vector de medias y la matriz de varianzas y covarianzas resumen toda la información disponible en un conjunto de datos.***
  FALSO. Son importantes pero no alcanzan. Falta información sobre outliers.

- ***Los outliers sólo pueden ser detectados a través de métodos univariados o bivariados.***
  FALSO. También pueden detectarse con métodos multivariados y MCD.

- ***El vector de medias de una matriz de datos de n x p tiene n componentes.***
  FALSO. n x p es filas x columnas. Tiene p componentes, la cantidad de variables.

Preguntas a desarrollar

- ***¿Que permite observar el gráfico de mosaicos? ¿Y el gráfico de Caritas de Chernoff? ¿Y el de coordenadas paralelas?***
  Mosaicos: Se utiliza para representar distribuciones conjuntas multivariadas.
  Chernoff: Cada variable es representada por una parte de la cara. Esta representación permite rápidamente hacer asociaciones y detectar diferencias.
  Coordenadas paralelas: se utiliza para variables cuantitativas. Consiste en ubicar un eje por cada variable que se quiere analizar; luego se traza una línea que toque cada eje en la altura que corresponde. Se debe trazar una línea por cada individuo.

- ***¿Con qué propósito se realizan transformaciones por fila y por columna?***
  Las transformaciones por fila se aplican con el objeto de hacer comparables los valores de los distintos individuos, neutralizando tendencias muy extremas (ej: juez). Las transformaciones por columna se utilizan para neutralizar el efecto de las unidades. Estas trans-

formaciones 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 la mediana y la desviación intercuartil o la mediana y el MAD.

- ***¿Qué propiedades son importantes de la matriz de varianzas y covarianzas? ¿Y de la matriz de correlación?***
  La matriz de covarianzas:
  - Es simétrica
  - Es semidefinida positiva, es decir que todos sus autovalores son mayores o iguales a cero.
  - La traza, que es la suma de los autovalores, indica de alguna forma la magnitud del problema.

  La matriz de correlación:
  - La traza nos da la cantidad de variables involucradas en el problema.
  - El módulo de cada componente es menor o igual 1.
  - Si r_{ik}= 1 significa que los datos caen sobre una línea recta de pendiente positiva.
  - Si r_{ik}= -1 significa que los datos caen sobre una línea recta de pendiente negativa.
  - Si 0 < r_{ik} < 1 significa que los datos caen alrededor de una línea recta de pendiente positiva.
  - Si -1 < r_{ik} < 0 significa que los datos caen alrededor de una línea recta de pendiente negativa.
  - Si r_{ik}= 0 indica que no hay una asociación lineal entre los datos de las variables.

- ***¿Qué es un outlier? ¿Cómo debe procederse con ellos?***
  Los outliers son valores atípicos que se presentan en las muestras, son valores extremos. Se los debe analizar, estudiar y si se deben a un error de medición o tipeo se los debe eliminar. Si no son errores, pueden utilizarse métodos robustos que le otorgan menos peso a los datos alejados.

- ***¿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 medición.

- ***¿Si se desea detector outliers en forma multivariada cómo conviene hacerlo?***
  Distancia de Mahalanobis, vector de medianas, MVE (*minimum volume ellipsoid),* MCD (*minimum covariance determinant*).

Análisis de componentes principales (ACP)

Verdadero o Falso

- ***ACP solo es válido si se ha rechazado la hipótesis de independencia.***
  FALSO. La hipótesis de independencia no está relacionada, además ACP se aplica si las variables son dependientes, ya que explica la asociación entre ellas.

- ***El nombre "biplot" en el contexto de ACP, obedece al hecho de que permite ver al mismo tiempo las observaciones y los componentes.***
  FALSO. Se denomina *biplot* porque se muestran las variables y los valores de cada individuo en pares de componentes principales.

- ***Una componente principal se dice de tamaño cuando sus coeficientes (loadings) tienen todos del mismo signo.***
  VERDADERO.

- ***En ACP, algunas variables explican el comportamiento de otras.***
  FALSO. En ACP se pueden ver las asociaciones entre variables.

- ***En un biplot, si 2 variables están negativamente correlacionadas entonces las flechas que las identifican aparecen ortogonales (perpendiculares) entre sí.***
  FALSO. Cuando 2 variables están negativamente correlacionadas, las flechas tienen direcciones opuestas (forman un ángulo de 180º aprox.). Las variables cuyas flechas aparecen ortogonales no tienen relación.

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

- ***El ACP se basa en el análisis de los autovalores de la matriz de datos.***
  FALSO. Se basa en la matriz de correlación, que tiene los datos estandarizados.

- ***Para determinar el número de componentes a considerar existen sólo dos criterios: el bastón roto y el porcentaje de variabilidad explicada.***
  FALSO. Existen también el Criterio de Kaiser y el Test de esfericidad de Barlett.

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

- ***En el ACP siempre la suma de los autovalores de la matriz de varianzas y covarianzas es igual a la suma de las varianzas.***
  VERDADERO. La diagonal de la matriz son las varianzas de cada variable.

- ***El ACP se basa en el análisis de la matriz de varianzas y covarianzas.***
  FALSO, o verdadero parcialmente. Podría basarse en la matriz de correlación.

- ***Uno de los criterios para determinar con cuántas componentes quedarse en ACP consiste en tomar un número proporcional a la cantidad de variables en el problema.***
  FALSO. No hay ningún criterio basado en esto (sí hay basados en la variabilidad explicada).

- ***Hacer ACP con la matriz de correlaciones es lo mismo que hacerlo con la matriz de covarianzas utilizando variables estandarizadas.***
  VERDADERO. La matriz de correlación es equivalente a la de varianza con las variables estandarizadas.

Preguntas a desarrollar

- ***Explique el criterio de bastón roto y ejemplifique.***
  Si la proporción de variabilidad explicada por las componentes calculadas se estabiliza a partir de un cierto valor *m*, entonces aumentar la dimensión no aportaría cambios significativos. En un punto del gráfico la pendiente se suaviza pareciéndose a una meseta; la sugerencia de este criterio es seleccionar las componentes previas a la zona de acumulación de sedimentos.

  ¿Qué puede observarse/cómo debe interpretarse un biplot?
  Un Biplot es una representación gráfica de datos multivariados. Se observa que:
  - 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 ángulo que forman los vectores que la representan: menor ángulo gran asociación.
  - La longitud de los vectores de las variables indica la variabilidad de cada una.
  - La proyección que tengan sobre los ejes implica cómo son los *loadings* de esa variable en cada componente (todos del mismo signo son tamaño, de distinto signo de forma).
  - Permite observar si los individuos respecto a los componentes están muy posicionados arriba o abajo.
  - Cuando los vectores de dos variables son perpendiculares, no están correlacionadas.
  - Los individuos que están más cerca del centro son los más cercanos al promedio.
  - Los individuos que están en los extremos son opuestos con respecto a las componentes.

- ***¿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 (λ/traza).

- ***Defina los conceptos de traza y determinante en función de los autovalores de una matriz. Explique qué relación puede establecerse con el ACP y AFC.***
  Traza: sumatoria de autovalores.
  Determinante: producto de autovalores.

En ACP la traza de la matriz de varianzas y covarianzas indica la variabilidad total. La traza de la matriz de correlación indica la cantidad de variables involucradas. En AFC, la traza de la matriz Chi cuadrado es la inercia total.

Contrastes de independencia y homogeneidad

Verdadero o Falso

- ***Si se rechaza H_0 de una prueba de independencia significa que la variable tiene una distribución diferente en cada una de las poblaciones.***
  FALSO. Al rechazar H_0 ("las variables son independientes") en una prueba de independencia implica que hay dependencia entre las variables, es decir que existe evidencia en contra de la hipótesis nula. Además, en una prueba de independencia hay solo una población y varias variables.

- ***Si en una tabla de contingencia no se ha rechazado H_0 del test de independencia de Chi Cuadrado es válido realizar un análisis de correspondencia.***
  FALSO. Si no se rechaza H_0 ("las variables son independientes") implica que no se puede afirmar que haya dependencia, con lo cual no tiene sentido hacer el análisis de correspondencia.

- ***El test de Chi cuadrado puede aplicarse tanto para homogeneidad como para independencia en cualquier conjunto de observaciones.***
  FALSO. No es para cualquier conjunto de datos porque se tienen que cumplir los supuestos. Para que sea válida la aplicación del test de Chi cuadrado, es necesario que todas las frecuencias esperadas resulten superiores a 1 y a lo sumo el 20% de las mismas inferiores a 5. Cuando no puede aplicarse el test de Chi cuadrado, una alternativa disponible es el test exacto de Fisher.

- ***La prueba de independencia testea la existencia de asociación entre variables categóricas. Rechaza la hipótesis de independencia cuando el estadístico toma un valor grande (p-valor menor que 0.05).***
  Esta afirmación es verdadera sólo si alfa es 0.05. En otro caso, es falsa, alfa no tiene por qué valer eso.

Preguntas a desarrollar

- ***¿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 dos o más variables categóricas o continuas. El test de homogeneidad se usa para comparar una variable en dos o más poblaciones.

- ***¿Para qué sirven los residuos en un test de Independencia?***
  Cuando la hipótesis nula se rechaza ("las variables son independientes"), debe suponerse que las variables son dependientes pero no se sabe en qué sentido están asociadas. Si deseamos indagar al respecto se pueden estudiar los residuos del modelo para saber qué tipo de dependencia existe entre ellas. Cuando un residuo tiene un valor absoluto superior a 2 en una celda hay que prestar atención a esa asociación porque esa casilla tiene un aporte importante al Chi Cuadrado.

- ***Establezca dos similitudes y dos diferencias entre los test de homogeneidad y los de Independencia.***
  Similitudes: Se calculan igual las frecuencias y valores esperados. La variable pivote es la misma. La zona de rechazo es la misma.
  Diferencias: Las hipótesis son distintas. En un caso hay *k* poblaciones y en otra una sola.

Análisis factorial de correspondencia (AFCS y AFCM)

Verdadero o Falso

- ***El AFCM (análisis de correspondencia múltiple) se puede aplicar solo en el caso que la inercia sea grande.***
  FALSO. Se puede aplicar cuando la cantidad de variables a analizar sea mayor a 2 (si son 2 se realizaría un análisis de correspondencia simple) y cuando se cumplen los supuestos para que se pueda aplicar el Test de Chi Cuadrado (las frecuencias esperadas resulten superiores a 1 y a lo sumo el 20 % de las mismas inferiores a 5).

- ***El AFC representa las distancias euclídeas entre los perfiles observados.***
  FALSO. Representa la distancia Chi Cuadrado que es la distancia euclídea entre los perfiles estandarizados.

- ***La inercia cuantifica el grado de adecuación del AFC a este conjunto de datos.***
  FALSO. Una forma de cuantificar la magnitud de las diferencias entre lo observado y lo esperado (grado de adecuación) es el estadístico $\chi 2$ de Pearson. La inercia, en cambio, es el cociente entre el estadístico Chi Cuadrado y el total de observaciones: éste cuantifica el grado de dependencia/asociación entre las variables.

- ***El AFC requiere que las variables sean independientes para que tenga sentido aplicarse.***
  FALSO. Requiere que NO sean independientes, es para explicar la asociación.

- ***En AFC, la suma de los autovalores es igual al valor de Chi cuadrado.***
  FALSO. La suma de los autovalores es igual a la inercia (Chi^2/ cantidad de observaciones).

- ***El análisis de correspondencias se aplica solamente cuando los perfiles fila y columna son idénticos o paralelos.***
  FALSO. Esto implicaría que sean independientes, el análisis de correspondencia es para explicar asociación.

Preguntas a desarrollar

- ***¿En qué caso se quedaría con una sola componente en AFC?***
  En el caso que ese componente tenga toda la inercia.

- ***Defina los conceptos de traza y determinante en función de los autovalores de una matriz. Explique qué relación puede establecerse con el ACP y AFC.***
  Traza: sumatoria de autovalores
  Determinante: producto de autovalores.
  En ACP la traza de la matriz de varianzas y covarianzas indica la variabilidad total. La traza de la matriz de correlación indica la cantidad de variables involucradas. En AFC, la traza de la matriz Chi cuadrado es la inercia total.

- ***Explique la utilidad de la matriz de Burt y de la distancia Chi Cuadrado.***

La matriz de Burt es aquella que contiene los resultados de variables categóricas disjuntos. Contiene submatrices cuyas trazas indican el N de la muestra. Si alguno es menor indica que faltan datos.
La distancia Chi cuadrado es la distancia euclídea normalizada de los perfiles fila o columna. Posee una propiedad muy importante que cuando se colapsan filas o columnas no afecta la exposición de la información.

- ***Explique el concepto de inercia ¿Qué utilidad tienen los perfiles?***
  La inercia es igual a la distancia Chi cuadrado sobre el número de la muestra. La media de las distancias al cuadrado de cada punto de fila al centro de gravedad se conoce como inercia de filas, o inercia de columnas cuando se trata de las columnas, e inercia total de la nube de puntos cuando se consideran todos los elementos de la tabla. Una inercia baja significa que todos los productos están situados muy cerca del centro de gravedad y que en consecuencia son muy similares, mientras que altos valores de inercia en determinadas categorías implican grandes diferencias del perfil medio de las filas o las columnas.

- ***¿Para qué se utiliza la distancia Chi cuadrado? ¿Por qué se elige esa distancia?***
  La distancia Chi Cuadrado es la distancia entre los perfiles, la distancia euclídea entre los perfiles estandarizados. Sirve para neutralizar la diferencia de las totales filas y totales de columna: cuando está 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.

- ***Cuándo se utiliza la distancia Chi cuadrado? ¿Qué propiedad particular la hace útil en ese contexto?***
  Se utiliza en el análisis de correspondencia. La propiedad que lo hace útil 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.

- ***¿Qué cuantifica la inercia en un análisis de correspondencias?***
  Cuantifica el grado de asociación. La Inercia Total (medida análoga a la variación total en el caso de las componentes principales) cuantifica el grado de dependencia entre las variables.

- ***¿Cómo deben interpretarse los biplot simétricos de AC?***
  La asociación entre filas. Cuáles son más regulares.

- ***Explique el concepto de perfil fila y perfil medio en el contexto de AFCS.***
  Las frecuencias relativas de las filas y las frecuencias relativas de las columnas son los perfiles fila y perfiles columna, respectivamente.
  Las distancias entre perfiles no se miden entre dos filas o dos columnas sino con relación al perfil medio de fila o columna, es decir, con relación al promedio de las coordenadas de esa fila (o columna) ponderada por su masa (peso proporcional a su importancia en el conjunto). Este perfil medio aparecerá situado en el origen de coordenadas y es conocido como centro de gravedad.