ASIGNATURA: ESTADÍSTICA (11539) - GRADO DE INGENIERÍA INFORMÁTICA **versión 13/02/2023**

TRABAJO ACADÉMICO DE ESTADÍSTICA curso 2022-2023

INDICACIONES GENERALES

- Consistirá en desarrollar un proyecto, como si se tratara de un informe sobre el análisis estadístico de unos datos, que una empresa encarga a un ingeniero. Fundamentalmente se trata de realizar cálculos con apoyo de Statgraphics (u otro programa estadístico), presentando los resultados en forma de tablas o gráficas, y discutiendo convenientemente estos resultados. El objetivo fundamental es extraer información útil que se deduce de datos reales. Se pretende que el trabajo abarque la mayor parte del temario de la asignatura.

- Constará de varias partes que los alumnos deberán entregar a lo largo del curso, en las fechas que el profesor señalará con antelación suficiente. El objetivo es que los alumnos vayan desarrollando este trabajo conforme se avance en las clases de teoría.

- El trabajo podrá ser individual, por parejas o grupos de tres (a elección de los alumnos). Dado que el trabajo contará un porcentaje considerable en la nota final, el profesor podrá hacer preguntas sobre el mismo en sesiones on-line (a través de *Teams* o similar) cuando el trabajo se entregue en su versión final, antes de ser calificado, si lo considera oportuno, para verificar que el alumno ha interiorizado todos los contenidos reflejados en el documento. En caso de trabajos realizados por parejas o grupos de tres, si el profesor detectase que un alumno ha contribuido notablemente más que su compañero, podrá asignar una nota distinta a cada uno.

- Al final de este documento están las instrucciones para descargarse el *Statgraphics Centurion* en castellano o inglés. Este software permite *copiar* → *pegar* texto o tablas directamente a un documento de Word. La información que ofrece el programa en el *StatAdvisor* puede resultar útil, pero no se debe copiar directamente ya que es tarea del alumno la interpretación de los resultados.
También se pueden *copiar* → *pegar* figuras directamente a un documento de Word. No obstante, si se desea editar las figuras (cambiar el color de fondo, grosor de las líneas, etc.) se puede realizar:

a) Botón derecho del ratón → opciones gráficas.
b) Copiar la figura, abrir Power Point→ *pegado especial* → pegar como: "imagen (meta archivo mejorado". Esto permite desagrupar las figuras, y realizar cualquier modificación que se desee.

INSTRUCCIONES PARA CONSEGUIR LOS DATOS

En PoliformaT (Recursos → *carpeta "información general y sistema de evaluación"*) está disponible una hoja Excel con enlaces (links) para acceder a bases de datos (datasets) útiles para realizar el trabajo. Es requisito que no haya dos trabajos que utilicen el mismo dataset (excepto si se usan variables distintas). Algunos enlaces adicionales con muchísimos conjuntos de datos son:

https://www.kaggle.com/
Azure Open Datasets (https://azure.microsoft.com/es-es/services/open-datasets/#overview)
Repositorio de datasets variados (https://github.com/awesomedata/awesome-public-datasets)
Datasets de Google (https://research.google/tools/datasets/)
UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php)
DrivenData: Competiciones de datos con un carácter humanitario
(https://www.drivendata.org/competitions/)

Instituto Nacional de Estadística (www.ine.es)
Base de datos abiertos del Ayuntamiento de Valencia
(http://www.valencia.es/ayuntamiento/DatosAbiertos.nsf/)
Datos abiertos Generalitat valenciana y de entidades locales que publican datos abiertos
http://portaldadesobertes.gva.es/va/comunitat-oberta
Portal de datos Abiertos de Esri España
https://opendata.esri.es/

La idea es partir del *dataset* original y seleccionar las variables más interesantes, para disponer de una **matriz con las siguientes características**:

- N° de observaciones **recomendado** (filas): entre 150 y 1000. No es requisito necesario que la matriz tenga menos de mil filas; el problema es que si se trabaja con miles de observaciones, los test de inferencia habitualmente tienden a salir estadísticamente significativos y esto puede dificultar la interpretación de los resultados.
    - Recomendación en caso de matrices muy grandes: si contiene por ejemplo 3000 observaciones y una variable cualitativa con dos variantes (ej: hombres y mujeres), un alumno puede trabajar con los datos para hombres, y otro compañero con el resto. Otra alternativa sería realizar una selección aleatoria de observaciones, pero los profesores recomendamos la opción mencionada.
    - Para realizar una selección aleatoria de 200 observaciones, por ejemplo, se puede proceder del siguiente modo: con Excel, crear una columna de valores aleatorios, con: *=aleatorio()* ; después, consolidar los valores (copiar y pegar en otra columna con: pegado especial -> pegar valores). Después ordenar las filas según los valores de la variable aleatoria, y seleccionar las 200 primeras filas.
    - Las observaciones tienen que ser independientes entre sí: no sirven series temporales (es decir, secuencias de valores medidos a lo largo del tiempo) ya que este tipo de análisis no se han visto en la asignatura. No sirven variables del tipo: "n° de infectados por una epidemia a lo largo de los distintos días".
- Los datos deben constituir una muestra de una población, no la población entera.
    - No sirven, por ejemplo, variables económicas de los países de la Unión Europea, o de todos los municipios de Valencia, etc., ya que si se dispone de todos los datos de la población, no tiene sentido realizar un test de inferencia.
    - Recomendamos no utilizar, por ejemplo: "películas más taquilleras de 2019", o "canciones más descargadas de Spotify", pues no son ejemplos de muestras aleatorias extraídas de una población, lo cual dificulta la interpretación de los test de inferencia.
- A ser posible, las variables deben contener todos los datos para todas las observaciones, es decir, hay que evitar datos faltantes (o como mucho, que falten pocos), puesto que si hay una proporción relevante de datos faltantes, la interpretación de resultados puede ser un poco liosa.
- Tanto las variables como los individuos deben tener una interpretación física, es decir, no pueden ser datos simulados. Dado que se trata de interpretar los resultados, es muy importante que se entienda qué es lo que se mide con cada una de las variables.

- **Dos variables deben contener información cualitativa**, con un número de variantes aconsejado entre 2 y 10 (por ejemplo, si la variable es "color", debe contener al menos dos colores distintos y 10 como máximo). A estas variables las vamos a llamar **F_1 y F_2** (pues serán los factores del ANOVA); F_1 será la que menos variantes tenga de las dos. También pueden emplearse variables discretas, con un rango de valores menor de diez.

  Se requiere que haya al menos 5 observaciones para cada una de las combinaciones de las dos variables cualitativas (ya que, de lo contrario, la interpretación de la interacción doble en ANOVA puede ser problemática). Este requisito se puede verificar con Statgraphics: *describir → datos categóricos → tabulación cruzada* (seleccionar F_1 y F_2).

- **Cuatro variables deben** ser continuas (o bien discretas con un rango de valores elevado, superior a 10), que llamaremos **X_1**, **X_2**, **X_3** y **X_4**. Al menos una de ella debería tener una distribución más o menos normal (o con una ligera asimetría), o bien que se ajuste a un modelo normal aplicando algún tipo de transformación (raíz cuadrada o logaritmo). A esta variable es a la que llamaremos **X_1**.

Muchas veces las bases de datos contienen muchas variables continuas y pocas cualitativas. En ese caso, se puede transformar una continua en cualitativa, de la siguiente forma: todos los datos menores a la mediana se codifican como nivel "bajo", y los superiores a la mediana como nivel "alto". O bien, valores < percentil 33, como nivel "bajo", entre percentil 33 al 66 como nivel "medio" y valores superiores como nivel "alto". De esta forma se soluciona el problema.

LISTA DE OBJETIVOS (ítems evaluables)

Hay que desarrollar unos **objetivos** concretos que se presentan a continuación, numerados secuencialmente según se corresponde con el temario de la asignatura. Todos estos ítems serán evaluados por el profesor, de acuerdo a la puntuación que se indica (no es la misma para todos pues depende del grado de dificultad y esfuerzo requerido). Para facilitar al profesor la corrección, hay que desarrollar el trabajo con la misma secuencia señalada, e indicando también el nº de ítem. No hace falta copiar el enunciado de cada ítem en el documento, pero es necesaria una redacción tal que se entienda el texto sin necesidad de recurrir al enunciado. En caso de desarrollar algo "extra" no contemplado en el guion, deberá señalarse como EXTRA.

1) Portada [1 punto]: en la que se indicará el nombre del alumno (o los dos si se hace por parejas), grupo y el título del informe estadístico (ejemplo: "análisis estadístico de variables socioeconómicas de municipios de Valencia"). En este ítem se valorará la adecuación del título elegido y la calidad (estética) de la presentación del documento, puntuándose favorablemente si las páginas están numeradas, la estética del encabezamiento de página, el uso de logotipos institucionales, imágenes en la portada, etc.
Nota: no es necesario incluir un índice del trabajo. En principio, no está previsto evaluarlo si se incluye, salvo que el profesor indique lo contrario.

2) Descripción de la base de datos (dataset) [1.5 puntos] (15-45 líneas).
- Indicar el link a partir del cual se han obtenidos los datos de partida.
- Indicar el nº de observaciones y variables en el dataset original.
- Contextualizar el estudio, indicando qué son las observaciones y, a grandes rasgos, qué mide el conjunto de variables. Si es posible, indicar cómo se obtuvo esta información.

- Si no se utilizan todas las observaciones (filas) del dataset original, ¿cuántas se han utilizado para el trabajo? ¿Con qué criterio se han seleccionado?
- Si no se utilizan todas las variables (columnas) del dataset original, ¿cuántas se han utilizado para el trabajo? ¿Con qué criterio se han seleccionado?
- Describir las variables utilizadas, con una pequeña explicación de lo que mide cada una (conviene indicar las unidades de medida). Si resulta conveniente, puede definirse una abreviatura para las variables (Ej: *el peso del vehículo en lo sucesivo se denominará variable "peso"*; es preferible esta denominación a lo largo del trabajo, en lugar de llamarla variable X_1).
- Indicar si hay datos faltantes en alguna de las variables. En caso afirmativo, mencionar la proporción.

3) Objetivos particulares [1 punto] (3-9 líneas)
En función del contexto, debe hacerse un esfuerzo por plantear seis objetivos, como mínimo, que previsiblemente pueden abordarse con el estudio estadístico. Ejemplos:

- Si de un conjunto de titulados universitarios tenemos la nota lograda en el título y el salario mensual después de dos años, cabe plantearse si existe correlación entre ambos.
- Si tenemos notas de los alumnos de dos colegios distintos, podemos plantear si existen diferencias estadísticamente significativas entre ellas.
- Dado que hay 4 variables continuas, se pueden estudiar como máximo 6 parejas de estas variables. Se puede indicar cuáles de estas parejas, a priori, parece que su estudio es más interesante por sospechar que puede haber correlación entre ellas.

4) Discusión de la muestra y población: **[0.5 puntos]** Indica cuál es la muestra y cuál es la población objeto de estudio. ¿Consideras que tu muestra (es decir, el conjunto de individuos con los cuales has realizado el trabajo) es representativa de la población? Justifica la respuesta (3-9 líneas).

ESTADÍSTICA DESCRIPTIVA

5) [1 p.] Representa el gráfico de barras y el gráfico de tartas de la variable F_1 (Statgraphics: *Describir → datos categóricos→ tabulación*). ¿Las categorías tienen una frecuencia similar? (3-9 líneas).

6) [1 p.] Calcula la tabla de frecuencias de la variable F_2
- ¿Qué se calcula en cada una de las columnas de la tabla?
- Comenta los resultados más relevantes (3-9 líneas).

7) [1 p.] Construye una tabla de frecuencias cruzadas con las variables F_1 y F_2 (4-12 líneas)
- Con el botón derecho → opciones de ventana: hay que elegir entre "porcentajes por fila" o "por columna": decidir cuál de estas dos opciones aporta más información (justificando la respuesta). Inserta la tabla con una de estas dos opciones.
  - Explica la diferencia entre frecuencia absoluta y relativa.
  - Explica la diferencia entre frecuencia marginal y condicional.
- ¿Hay relación entre las variables F_1 y F_2? justifica la respuesta a partir de la tabla cruzada, explicando en caso afirmativo cómo es dicha relación.

Nota: se recomienda elegir convenientemente los datos para que ninguna frecuencia absoluta sea cero, ya que en ese caso no es posible estudiar la interacción de los dos factores con ANOVA (preguntas 31 a 36). En caso de que existan ceros, se recomienda eliminar variantes o agruparlas en categorías más generales.

8) **[1 p.]** Indica en una única tabla los principales estadísticos de las 4 variables continuas: rango, rango intercuartílico, media, mediana, varianza, desviación típica, coeficiente de variación, coeficiente de asimetría estandarizado y coeficiente de curtosis estandarizado.
- Indica cuáles de estos son parámetros de posición, dispersión o de forma.

9) **[1 p.]** Realiza un papel probabilístico normal para cada una de las variables $\mathbf{X_i}$, indicando con ellos y con los coeficientes de asimetría y curtosis estandarizados cuál/es de esas variables es la/s que puedes tomar como $\mathbf{X_1}$.

10) **[1 p.]** Representa los cuatro histogramas de las variables $\mathbf{X_i}$, primero con el número de intervalos que proporciona el Statgraphics y, si es necesario, con un número de intervalos más adecuado, indicando el porqué del cambio.

11) **[1 p.]** Representa los cuatro diagramas de caja y bigotes de las variables $\mathbf{X_i}$, indicando similitudes y diferencias entre las variables.

12) **[1 p.]** Coloca en una misma figura: un histograma de la **variable $\mathbf{X_2}$** y a su derecha, un gráfico de caja-bigotes múltiple, en función del factor F_1. Si la distribución es muy asimétrica, puede ser conveniente insertar adicionalmente otro gráfico con la variable transformada, empleando raíz cuadrada o logaritmo.
- Comenta en qué consiste el gráfico de caja-bigotes múltiple (2-6 líneas).
- Calcula los coeficientes de asimetría y curtosis estandarizados de X_2 en función de las distintas variantes de F_2. Inserta esta información en una tabla, indicando también el nº de datos.

13) **[1 p.]** Comenta las diferencias observadas entre las distintas variantes de F_1 a partir de la información del ítem anterior (variable $\mathbf{X_2}$):

- Diferencias en cuanto a la posición: ¿cuál tiene mayor media o mediana?
- Diferencias en cuanto a la dispersión: ¿cuál tiene mayor intervalo intercuartílico?
- diferencias en cuanto a la forma (simétrica o asimétrica):
  - En caso de simetría, ¿para qué variantes podría asumirse un modelo normal?
  - En caso de asimetría, comentar su signo e intensidad.

- Comenta si hay datos claramente anómalos que deberían descartarse del estudio. En caso afirmativo, estos deberán descartarse, fundamentalmente para los estudios de inferencia.

14) **[1 p.]** Respecto a la pregunta anterior, si tuviéramos que comparar los datos de $\mathbf{X_2}$ en función de las distintas variantes del factor F_1, indica qué parámetros de posición y dispersión serían más adecuados en este caso para describir la pauta de variabilidad. Justifica la respuesta. Calcula dichos parámetros.

15) **[1 p.]** Para describir gráficamente la pauta de variabilidad de la **variable $\mathbf{X_3}$**, elige el gráfico que consideres que aporta más información (histograma, papel probabilístico normal o gráfico de caja-bigotes).
- ¿Por qué has elegido este gráfico? Justifica la respuesta (1-3 líneas).
- Comenta la información más relevante que se deduce del gráfico (3-9 líneas).
- Comenta si hay datos claramente anómalos que deberían descartarse del estudio. ¿Por qué deberían descartarse?

16) **[1 p.]** Crea una nueva variable uniendo (no sumando) los datos de las variables X_1, X_2, X_3 y X_4 (es decir, copiando los datos de X_2 debajo de los de X_1, luego los de X_3 debajo de los anteriores, y los de X_4 debajo). A continuación, construye un histograma de la nueva variable. Si la distribución resulta muy asimétrica puede ser conveniente aplicar alguna transformación. ¿Qué se aprecia en la gráfica? ¿Por qué? (4-12 líneas).

DISTRIBUCIONES DISCRETAS Y CONTINUAS

17) **[1 p.]** Elige una variable discreta (que sea el resultado de contar algo, ej: nº de veces que sucede una avería, etc.), y cuyo rango de valores sea inferior a 100. ¿Su distribución sigue alguno de los modelos discretos estudiados en la asignatura? Ajustar un modelo teórico de distribución con Statgraphics: *describir → ajuste de distribuciones → ajuste de datos no censurados*.

- Inserta el gráfico del mejor ajuste y comenta los resultados (2-6 líneas).
- Si el ajuste es razonablemente bueno, indica el valor de los parámetros del modelo.

En caso de no disponer de variables discretas, realizar una de estas dos opciones:

a) Simula una variable de tipo Poisson que tenga la misma media que X_1 (*describir → ajuste de distribuciones →distribuciones de probabilidad*). Inserta el gráfico de la función de masa/densidad y, a la derecha, el histograma de X_1. ¿Qué se deduce a la vista de ambos gráficos? (2-6 líneas).

b) Toma una variable continua y construye una nueva variable con la fórmula: 20·(X-min)/max ; luego redondea los valores resultantes [se puede usar la función de excel: *redondear(_ _;0)*]; Luego responde a lo indicado en este ítem.

18) **[1 p.]** Para dos de las variables continuas que has escogido cuya distribución no sea normal, estudia la bondad de ajuste a distintos modelos de distribuciones continuas: uniforme, exponencial, log-normal o triangular. De estos cuatro modelos, inserta el gráfico del mejor ajuste (*describir → ajuste de distribuciones → ajuste de datos no censurados*).

- Comenta las conclusiones derivadas de este estudio.
- Si el ajuste es razonablemente bueno, indica los valores estimados de los parámetros del modelo obtenidos a partir de Statgraphics.
- Opcional: esta opción de Statgraphics ofrece pruebas de bondad de ajuste que, aunque no se explican en la asignatura, pueden comentarse si se considera conveniente.

19) **[1 p.]** Para dos de las variables continuas X_i con distribución asimétrica positiva, conviene estudiar si alguna transformación es capaz de "normalizar" los datos, es decir, hacer que su distribución se asemeje a una normal. Si los datos son todos negativos hay que cambiarlos a positivos. Representar los datos de X_i sobre un papel probabilístico normal, y probar a continuación distintas transformaciones hasta encontrar una que normalice lo mejor posible los datos:

- En caso de asimetría positiva se aconseja probarlas en este orden: raíz cuadrada (X_i^0.5), (X-min)^0.5, raíz cuarta (X_i^0.25), (X-min)^0.25, logaritmo, y finalmente log (X - a), siendo "a" una constante cercana al mínimo.
- En caso de asimetría negativa se aconseja probar las siguientes: (max-X)^0.5, (max-X)^0.25

- Hay que insertar el gráfico del papel probabilístico de la variable original, y el papel probabilístico normal con la transformación que mejor haya funcionado para "normalizar" los datos. Justifica la respuesta (3-9 líneas).
- A la vista de los resultados, ¿puede decirse que alguna variable sigue una distribución log-normal?

- En caso de que ninguna transformación consiga una buena normalización, explica los motivos (por ejemplo, en caso de presentarse una mezcla de poblaciones).

20) **[1 p.]** ¿Cuáles de las variables que tienes en tu conjunto de datos siguen una distribución razonablemente normal? ¿Cuáles son sus parámetros? Justifica tu respuesta (3-7 líneas).
En caso de que no haya ninguna variable normal, explica los motivos en el contexto del estudio.

21) **[1 p.]** Genera aleatoriamente 100 valores de una distribución N(media=15, □=4) y otros 100 de una distribución N(media=3, □ =3). Statgraphics: *Describir* → *ajuste de distribuciones* → *distrib. de probab.* → (modelo normal) → (indicar los parámetros) → en *Tablas y gráficos*, activar "números aleatorios" → click en botón con forma de diskette. Genera una nueva variable sumando los datos por parejas.
- Describe brevemente las operaciones realizadas (2-4 líneas).
- A partir del valor de los coeficientes de asimetría y curtosis estandarizados, estudia si la variable suma se ajusta razonablemente a un modelo normal. En caso negativo, justifica las posibles razones.
- Calcula con Statgraphics la media y desviación típica de la variable suma.
- Calcula de modo teórico (con las fórmulas y estadísticos pertinentes) cuál sería la media y desviación típica que cabría esperar de esta variable.
- ¿Por qué no coinciden exactamente los parámetros teóricos con los observados?

Fin de la 1ª parte del trabajo (temario que otros años abarca el 1er parcial)	**Total: 21 puntos**

SEGUNDA PARTE DEL TRABAJO

DISTRIBUCIONES EN EL MUESTREO - INFERENCIA SOBRE UNA POBLACIÓN

22) **[1 p.]** Elegir la variable X_i con distribución más similar a la normal (o en su caso, utilizar la transformación que mejor normaliza la variable). Asumir que ésta sigue un modelo normal X ≈ N(m; □), de media m igual a la media muestral, y desviación típica □ igual al valor muestral. Si se tomara aleatoriamente una muestra de 10 datos de esta población y se calculase su media, calcular el intervalo de confianza que comprendería el 95% de estos valores. Nota: resolver el problema de modo teórico, justificando los cálculos (4-10 líneas).

23) **[1 p.]** Elegir la variable X_i con distribución más similar a la normal. Asumiendo que esta variable sigue una distribución normal, calcular con Statgraphics el percentil 52 de la distribución (Z_{52}).
- Resuelve con las fórmulas pertinentes (justificando los cálculos) el contraste de hipótesis H_0: m = Z_{52} frente a la alternativa: H_1: m ≠ Z_{52}. Considera como tamaño de la muestra el número de observaciones (individuos) del dataset utilizado. Considera un nivel de significación del 10%.
- Resuelve el mismo contraste de hipótesis con Statgraphics y verifica que se obtiene el mismo valor del estadístico de contraste, y las mismas conclusiones (4-12 líneas).

24) **[1 p.]** Elegir la variable X_i con distribución más similar a la normal. Asumir que esta variable sigue una distribución normal X ≈ N(m; □), de media igual a la media muestral, y desviación típica igual al valor muestral. Si se toma aleatoriamente una muestra de 10 datos de esta población y se calcula su varianza, calcular el intervalo de confianza que comprendería el 95% de estos valores. Nota: hay que resolver el ejercicio de modo teórico, justificando convenientemente todos los cálculos (4-10 líneas).

25) **[1 p.]** Elegir la variable X_i con distribución más similar a la normal. Asumir que esta variable sigue un modelo normal X ≈ N(m; □) de media igual a la media muestral, y desviación típica igual al valor muestral. Si se tomara aleatoriamente una muestra de 12 datos de esta población y se calculase su varianza, ¿cuál es la probabilidad de que sea superior a 3·□? (3-9 líneas).

26) **[1 p.]** Elegir la variable X_i con distribución más similar a la normal. Si se toman dos muestras al azar de 14 valores de esta variable, ¿cuál es la probabilidad de que la varianza de la segunda muestra sea más del triple que la primera? Nota: hay que resolver el ejercicio de modo teórico (3-9 líneas).

27) **[1 p.]** Obtener con Statgraphics un intervalo de confianza para la media de X_1 a nivel poblacional, con un nivel de confianza del 99%.
- ¿Qué sucedería en caso de que X_1 no se ajustara a una distribución normal? (2-6 líneas)
- Comenta qué opinas sobre la siguiente afirmación: "Si se toma cualquier valor perteneciente al intervalo de confianza obtenido y se realiza un test de hipótesis sobre la media, la conclusión de dicho test siempre será la misma considerando □=1%" ¿Es cierta o falsa? ¿Por qué? (3-9 líneas).

28) **[1 p.]** Elegir la variable X_i con distribución más similar a la normal. Obtener con Statgraphics un intervalo de confianza para la desviación típica de X_i a nivel poblacional, con un nivel de confianza del 95%. Calcular el intervalo también con un nivel de confianza del 99%.
- ¿Qué interpretación tiene en la práctica los intervalos obtenidos? (2-4 líneas).
- ¿Qué intervalo (95% o 99%) te parece más adecuado? ¿De qué depende esta decisión? (3-9 líneas).

29) **[1 p.]** Coloca en una tabla la varianza obtenida para la variable X_i seleccionando el subconjunto de observaciones que presenten una determinada alternativa de F_2 (por ejemplo: varianza de ventas para el subconjunto de discos de Jazz). A continuación calcula de nuevo la varianza de X_i para las observaciones en la que F_2 presente otra alternativa diferente a la anterior (ej: varianza de ventas para los discos de Rock). Incluye también en la tabla el número de observaciones (*n*) implicadas en el cálculo de cada varianza (ver tabla inferior como ejemplo). Si se toma la mayor de estas varianzas y la menor (siempre que *n* > 10), ¿puede afirmarse que las diferencias observadas resultan estadísticamente significativas? Considera el □ que consideres más adecuado en este caso (3-9 líneas).
Hay que resolver el cálculo de modo "manual" (como si fuera una pregunta de examen), justificando todos los pasos involucrados en la resolución. Ejemplo:

F_2: GÉNERO_MUSICAL	X_1: VENTAS	N
Jazz	$s^2 = 16$	78
RocK	$s^2 = 12$	53

ANÁLISIS DE LA VARIANZA

Nota: se trata de realizar básicamente el mismo análisis con X_1, X_2 y X_3, pero siguiendo un orden algo distinto con cada variable para comprender mejor la técnica y evitar repetir básicamente lo mismo en cada variable. Para interpretar los resultados del ANOVA recomendamos consultar los exámenes resueltos, pues en todos los del 2° parcial y final hay una pregunta sobre ANOVA, y habitualmente se pide que se interpreten los resultados.

30) **[2 p.]** Realizar un análisis de la varianza (ANOVA) para estudiar el efecto de F_2 en la variable $\mathbf{X_1}$.
- En caso de asimetría positiva fuerte, hay que utilizar la transformación más adecuada.

- Utiliza el nivel de significación que consideres más adecuado en este caso (justifica tu elección).
- Mostrar la tabla resumen del ANOVA del modelo.
- Mostrar el gráfico de medias con intervalos LSD (eligiendo el considerado más adecuado):
  - Indica si todos los intervalos tienen la misma amplitud o no y por qué (1-4 líneas).
  - ¿Las conclusiones que se derivan del gráfico son coherentes con la tabla resumen del ANOVA?
- Interpreta el análisis y explica las principales conclusiones derivadas del mismo (2-9 líneas).
- Muestra el papel probabilístico normal de los residuos. A partir de este gráfico, indica si existen residuos anómalos que deberían descartarse del modelo. En caso afirmativo, repite el análisis y explica los cambios en las conclusiones si los hubiera.

31) [1 p.] Incorpora al modelo anterior el factor F_1 y la interacción doble:
- Mostrar la tabla resumen del ANOVA inicial, y muestra y justifica también la tabla resumen ANOVA del modelo definitivo.
- Mostrar el gráfico de medias con intervalos LSD del factor F_1. ¿Las conclusiones que se derivan del gráfico son coherentes con la tabla resumen del ANOVA?
- Mostrar el gráfico de la interacción doble con intervalos LSD. ¿Qué información útil aporta en este caso? (2-9 líneas).
- A partir del modelo definitivo analizado, guarda los residuos y represéntalos sobre un papel probabilístico normal. ¿Qué se deduce? (1-3 líneas)

32) [1 p.] Objetivo: realizar un ANOVA para estudiar el efecto simple de los factores F_1 y F_2 en la variable $\mathbf{X_2}$ (no incluir la interacción doble). Trabaja con la variable original (sin transformar los valores en caso de asimetría).
- Utiliza el nivel de significación que consideres más adecuado en este caso (justifica tu elección).
- Mostrar la tabla resumen del ANOVA inicial con los dos efectos simples (sin la interacción).
- Mostrar el gráfico de medias con intervalos LSD para cada uno de los dos factores.
- Interpretar las principales conclusiones prácticas derivadas del análisis.
- Indica con qué valor de alfa se alcanza la significación. Explica el resultado y si no se alcanza con valores razonables, aportar posibles motivos, en base a los datos (en el contexto del estudio) que puedan justificar este hecho (incluir también la tabla resumen y gráficos LSD).

33) [1 p.] Incorpora al modelo anterior la interacción doble:
- Mostrar la tabla resumen del ANOVA al incorporar la interacción doble, y también la del modelo definitivo (tras eliminar los factores no significativos).
- Indica a partir de los resultados de la tabla si hay variaciones significativas en la interpretación de los mismos respecto a la cuestión previa y explica tus conclusiones.
- Justifica si tiene sentido o no estudiar el gráfico de la interacción doble.
- ¿Qué información adicional (a la obtenida en el ítem anterior) aporta en este caso concreto el estudio de la interacción doble? Es decir, indica cómo cambiaría la interpretación de los resultados (3-9 líneas).
- A partir del modelo en el que todos los efectos incluidos resultan estadísticamente significativos, guarda los residuos y represéntalos sobre un papel probabilístico normal. ¿Qué se deduce? En caso de observarse una asimetría positiva, ¿qué se recomendaría? (2-6 líneas).

34) [1 p.] Objetivo: estudiar con ANOVA el efecto de los factores F_1 y F_2, y el de su interacción doble, en la variable $\mathbf{X_3}$. Trabaja con la variable original (sin transformar los valores en caso de asimetría).
- Utiliza el nivel de significación que consideres más adecuado en este caso (justifica tu elección).

- Mostrar la tabla resumen del modelo inicial. ¿El efecto de la interacción doble es estadísticamente significativo?
    - En caso afirmativo: mostrar el gráfico de dicha interacción, con los intervalos LSD.
        - Botón derecho → *opciones de ventana* → *gráfica en eje*: activar "2° factor". Muestra también este gráfico. ¿A partir de cuál de los dos se deduce una información más clara?
        - Interpreta la información derivada de ambos gráficos (3-9 líneas).
        - Estudia los gráficos de medias de cada factor y si consideras que aportan información útil (por ejemplo, en caso de muchas variantes), muéstralos, explicando tus conclusiones.
    - Si la interacción no es significativa: mostrar el gráfico de medias con intervalos LSD para cada uno de los dos factores. Interpreta estos resultados (3-7 líneas).

35) [1 p.] Estudio de residuos del modelo anterior (variable X_3):
- A partir del modelo en el que todos los efectos incluidos resultan estadísticamente significativos, guarda los residuos y represéntalos sobre un papel probabilístico normal. Muestra dicho gráfico.
- ¿Qué se deduce? ¿Se observa una clara mezcla de poblaciones? En caso de detectarse datos claramente anómalos, elimínalos y repite el estudio. Si la existencia previa de los datos anómalos afectaba a las conclusiones, matiza lo que creas necesario con el resultado correcto sin datos anómalos. (2-9 líneas).

- Indica cuál de las dos afirmaciones es la correcta (justifica la respuesta):
    1) En caso de que la variable dependiente sea asimétrica positiva, es conveniente normalizarla para conseguir que los residuos del ANOVA se ajusten razonablemente a un modelo normal.
    2) En caso de que la variable dependiente sea asimétrica positiva, es preferible ajustar el modelo ANOVA y estudiar la distribución de los residuos; en caso de que estos sigan una distribución asimétrica, es conveniente probar distintas transformaciones en la variable dependiente hasta conseguir que los residuos se ajusten razonablemente a un modelo normal.
- En caso de que los residuos sigan una distribución asimétrica positiva:
    - Estudia cuál sería la transformación más adecuada que habría que aplicar a la variable dependiente hasta conseguir la normalidad de los residuos (2-9 líneas).
    - Indica qué conclusiones habría que modificar o matizar (2-9 líneas).

REGRESIÓN LINEAL

36) [0.5 p.] Obtener la matriz de varianzas-covarianzas de las variables X_1, X_2, X_3 y X_4. ¿Qué información útil aporta esta matriz? ¿Por qué es simétrica? ¿Cómo se interpretan los valores de la diagonal principal? (3-9 líneas).

37) [0.5 p.] Obtener la matriz de correlación de estas variables. Las celdas de esta matriz deben contener únicamente el coeficiente de correlación. En caso de asimetría positiva fuerte o muy fuerte es recomendable normalizar previamente las variables. O bien insertar dos matrices de correlación: con las variables originales, y con las variables normalizadas.
- ¿Qué se deduce de esta matriz? Es decir, comenta la correlación entre las variables (2-9 líneas).
- Explicar cuál es el valor de los elementos de la diagonal principal (2-6 líneas).

38) **[1 p.]** A partir de la matriz anterior, identifica la pareja de variables con mayor grado de correlación (a ser posible, que ésta no sea superior a 0.95). Realiza un gráfico de dispersión entre ambas (*graficar* → *gráficos de dispersión* → *gráfico X-Y*); inserta la figura en el trabajo. Utiliza un color distinto para cada variante del factor F_1 (botón derecho → opciones de ventana → indicar F_1 en "código de puntos"). Si alguna de las variables es muy asimétrica conviene utilizar transformaciones para normalizarlas.
- ¿Qué se deduce de este gráfico? (2-6 líneas)
- Describe la relación entre ambas variables (lineal o cuadrática, correlación positiva o negativa, fuerte/moderada/débil, etc.) (2-6 líneas).
- A la vista del gráfico, ¿la varianza de la distribución condicional de Y depende de X? Es decir, ¿puede asumirse la hipótesis de homocedasticidad?

39) **[1 p.]** Entre las 4 variables X_1 a X_4, elige aquella que podría considerarse como variable respuesta, es decir, que es función de alguna otra variable explicativa; a dicha variable la vamos a llamar Y. A partir de la matriz de correlación, identifica la variable con mayor correlación (positiva o negativa) con Y. En caso de que X o Y sigan una distribución asimétrica positiva fuerte, hay que normalizarlas convenientemente. Realiza un análisis de regresión lineal simple que permita predecir los valores de Y en función de X:

- Inserta el gráfico de dispersión de Y en función de X junto con la recta de regresión ajustada, y el intervalo de la predicción (con un nivel de confianza del 95%). Comenta la utilidad práctica de dicho intervalo.
- Inserta la tabla resumen del modelo obtenida con Statgraphics. A partir de la información de la tabla, ¿puede afirmarse que la correlación observada es estadísticamente significativa? (utiliza el nivel de significación que consideres más adecuado).
- Escribe la ecuación matemática del modelo: $Y = a + b \cdot X$. Comenta la significación estadística de ambos coeficientes, utilizando el nivel de significación que consideres más pertinente. ¿Qué utilidad práctica tiene la ecuación obtenida?

40) **[0.5 p.]** Respecto al ítem anterior:

- ¿Cuál es la interpretación práctica de los coeficientes "a" y "b" del modelo? ¿Tiene sentido esta interpretación? (2-5 líneas).
- Comenta la posible causalidad de la correlación: a partir de la interpretación física de las variables, ¿es posible sospechar que la correlación observada se debe a una relación causa-efecto, a una dependencia parcial, o bien a una interdependencia entre las variables? (2-6 lín.).

41) **[1 p.]** Estudio de los residuos: guarda los residuos del modelo anterior y represéntalos sobre un papel probabilístico normal. ¿Qué se deduce? En caso de observarse datos anómalos, ¿qué se recomendaría?
- Representa los residuos en función de X. ¿Se sospecha que pueda existir un efecto cuadrático?
- Explica cómo se puede verificar si dicho efecto resulta estadísticamente significativo. En caso afirmativo, interpreta dicho efecto (a partir de la figura y/o de los coeficientes de regresión estimados).
- Utilizando las fórmulas pertinentes y a partir de la información reflejada en la tabla resumen del modelo, calcula un intervalo de la predicción de Y cuando X vale su tercer cuartil (con un nivel de confianza del 95%). Justifica los cálculos. ¿Qué interpretación práctica tiene este resultado? (4-10 líneas)

42) Resumen **[1.5 puntos]** Consiste en resumir los aspectos más relevantes que se han puesto de manifiesto en todo el trabajo. En la última entrega del trabajo que se envíe al profesor, el trabajo constará de la portada, y a continuación se incluirá el resumen de todo el trabajo, escrito a modo de

"abstract" de un artículo científico, con una extensión entre 300 y 400 palabras (menos de una cara). Con él se pretende evaluar la capacidad del alumno para sintetizar los resultados más relevantes de modo concreto (no sirve decir "la variable X_1 está muy correlacionada con X_2..."). Puede ser útil leer: https://neoscientia.com/abstract-cientifico-ejemplos/

Nota: desde la pregunta 22 hasta el final (2ª parte del trabajo), las preguntas suman 21 puntos.

PUNTUACIÓN DEL TRABAJO

El profesor calificará el trabajo de 0 a 10 puntos en función de múltiples criterios: calidad de la redacción, lenguaje estadístico correctamente utilizado, existencia de errores interpretativos, omisión de los objetivos planteados, etc. A modo orientativo, se establece el siguiente criterio:

- No se alcanzará la puntuación máxima de cada apartado si no se escribe el número mínimo de líneas indicado.
- La nota = 5 corresponde a un trabajo que ha cumplido más o menos a los objetivos planteados y no contiene errores sustanciales, pero respondiendo muy escuetamente a lo que se pregunta, y con una interpretación deficiente de los resultados.
- Una nota menor a 5 correspondería a un trabajo que ha omitido algunos ítems, la interpretación es muy escueta y contiene algunos errores importantes.
- La nota máxima (10) correspondería a un trabajo excelente, muy bien presentado, sin errores, donde se aprecia que el alumno se ha esforzado por desarrollar convenientemente todos los ítems, interpretando correctamente los resultados.

El hecho de que un trabajo se realice por parejas no implica a priori una penalización en la puntuación (es decir, los dos alumnos pueden optar también a la nota máxima), con la salvedad de que el profesor, si lo considera oportuno, puede concertar una entrevista *on-line* con ambos alumnos para verificar si los dos han asimilado e interiorizado todos y cada uno de los ítems planteados. En esta entrevista se clarificarán dudas que al profesor le hayan podido surgir.

Aquellos alumnos que quieran "subir nota", se les concede la posibilidad de optar a un punto adicional (es decir, nota máxima de 11), el cual valorará todo aquello escrito en el trabajo que no se contempla en los ítems formulados, y que queda fuera del temario oficial de la asignatura. Para facilitar la corrección al profesor, conviene señalar claramente el texto como "EXTRA". Cada profesor podrá dar indicaciones concretas sobre este tema. No obstante, algunos de estos objetivos podrían ser los siguientes:

a) Realizar un ANOVA con más de dos factores e interpretar adecuadamente los resultados.
b) Estudiar la hipótesis de homocedasticidad en ANOVA (realizando un nuevo ANOVA tomando como variable respuesta los residuos al cuadrado).
c) Realizar un análisis de regresión lineal múltiple para predecir Y en función del resto de variables explicativas disponibles.
d) Realizar un contraste de independencia (test chi^2) en la tabla de frecuencias cruzadas.

Puntuación aproximada: apartados a) y b) conjuntamente, máximo **+0.4** puntos; apartado c) máximo **+0.4** puntos. Apartado d) máximo **+0.2** puntos.

FECHAS DE ENTREGA E INSTRUCCIONES

Se pretende que los alumnos vayan avanzando el trabajo conforme se explican los distintos contenidos de la materia. Es un trabajo con muchos ítems al que hay que dedicar bastante tiempo. Los profesores de la asignatura han acordado las siguientes fechas de entrega (las mismas para todos los grupos):

- Primera entrega: **Fecha límite, domingo 30 de abril de 2023 a las 23 h;** los alumnos deberán responder a los objetivos **nº 1 a 21** (incluido).
- Segunda entrega: **Fecha límite, domingo 4 de junio de 2023 a las 23 h;** los alumnos deberán responder a los objetivos **nº 22 al 42** (ambos incluidos).
- Tercera entrega (optativo: cada profesor decidirá si establece esta entrega, o solamente las dos anteriores).

- En las distintas entregas hay que incluir la misma portada, para que conste el nombre del trabajo y de los alumnos.
- Las entregas no enviadas antes de la fecha límite serán puntuadas como cero, si no existe una causa extraordinaria y sobrevenida que justifique el retraso en la entrega.

- El trabajo se enviará a través de Tareas de PoliformaT (se avisará cuando esté creada la tarea correspondiente a la 1ª entrega del trabajo). Se ruega no enviarlo por correo electrónico.

- Cada profesor indicará si prefiere que los trabajos se entreguen en formato Word, PDF o cualquier otro. El nombre del documento debe empezar por la letra del grupo, seguido por vuestros apellidos. Ejemplo: proyecto realizado por José García Pérez y Antonio Molina Hernández del grupo G, el nombre sería: G_Garcia-Perez_Molina-Hernandez

- Hay que numerar correlativamente todas las tablas y figuras, con una breve descripción (1-2 líneas) para cada una. Ejemplo: *Tabla 1. Frecuencias de la variable tiempo*…. Esta leyenda generalmente se coloca antes (arriba) de la tabla, mientras que, para las figuras, se suele colocar debajo.

- Para insertar las tablas con Statgraphics Centurion: basta seleccionar el texto de la tabla, copiar y pegarlo directamente al fichero de Word.
- Para insertar las figuras: colocando el cursor encima de la figura, con el botón derecho del ratón: "copiar", y pegar directamente en Word. Una vez pegada, con el botón derecho del ratón se puede "editar imagen" por si se quieren realizar cambios de formato.

En formato Word, frecuentemente las figuras se "mueven de sitio" al introducir texto. Para evitar este problema, se recomienda: pinchar en la figura → botón derecho → *tamaño y posición*:

- *ajuste del texto → detrás del texto*
- *posición → mover objeto con texto*

INSTALACIÓN DEL PROGRAMA STATGRAPHICS Centurion:

Se puede acceder al programa a través de Polilabs: https://polilabs.upv.es Hay que identificarse en el sistema, seleccionar "aplicaciones con licencia de campus", y ejecutar Statgraphics (versión en castellano) o Statgraphics EN (versión en inglés). En ambos casos es la versión Centurion 18.

No obstante, la UPV tiene contratada una licencia que permite a todo el personal de la UPV (profesores y alumnos) instalar y usar legalmente el programa (en inglés o castellano), tanto en ordenadores de la UPV como en sus propios ordenadores domésticos, portátiles, etc. con sistema operativo Windows:

http://software.upv.es → identificarse → abrir carpeta "Software para Alumnos" → abrir carpeta "Statgraphics Centurion XVII" (o bien la versión XVIII).

• Hay un documento con las instrucciones de instalación, y otro con el **nº de serie**.

• Para instalar el programa en otros idiomas: ir a la carpeta “idiomas suplementarios” → disponible en inglés, francés, alemán e italiano.

Instrucciones de instalación:

- Tener una cuenta de administrador en Windows; si se tiene Windows Vista o Windows 7 hay que ejecutar el programa con el botón derecho del ratón → opción “**Ejecutar como administrador**”; esto hay que hacerlo al instalar y al ejecutar el programa por primera vez.
- Cuando el programa se inicia por primera vez, en la ventana de diálogo del “administrador de licencias” hay que pulsar “activar” y rellenar la pantalla de registro, introduciendo los datos, dirección de correo electrónico sin subdominio (por ejemplo: pepe@etsinf.upv.es se introducirá como pepe@upv.es) y **n° de serie** (no hay que copiar y pegar el n° de serie directamente desde el documento PDF ya que puede incluir información del formato que haga que el n° introducido no sea correcto).
- A continuación, pulsar la opción “2” (solicitar un código de activación por correo electrónico).
- El código de activación se envía por correo electrónico. Cuando se reciba el código, hay que escribirlo en la casilla del paso 3 y pulsar el botón “Activar”. Asegurarse de introducir el código correcto, lo mejor es copiar → pegar, asegurándose de que no se cuelen espacios en blanco al principio ni al final.
- La activación es válida durante un año como máximo. Pasado ese plazo, deberá repetir los pasos anteriores para solicitar un nuevo código.

INSTALACIÓN DE LA VERSIÓN EN INGLÉS

También se puede descargar de la página web: https://www.statgraphics.com/download18

El número de serie de la licencia educativa de la UPV es: `B4B0-9B0A-00E0-YK0E-DEM0`

Al introducir este n° de serie, se envía por e-mail el código de activación.