ASIGNATURA: ESTADÍSTICA (11539) - GRADO DE INGENIERÍA INFORMÁTICA **(GRUPO 1B y 1C)**

TRABAJO ACADÉMICO DE ESTADÍSTICA curso 2024-2025

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.

- El trabajo podrá ser individual o por parejas (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 presenciales u 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, si el profesor detectase que un alumno ha contribuido notablemente más que su compañero, podrá asignar una nota distinta a cada uno.

- El software *Statgraphics Centurion* 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

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

- Nº 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.
    - 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: "nº 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 una prueba 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 las pruebas 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 con 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.

0.1) Portada [0.25 puntos]: 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.

0.2) Descripción de la base de datos (conjunto de datos) [0.75 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 conjunto de datos 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 conjunto de datos original, ¿cuántas se han utilizado para el trabajo? ¿Con qué criterio se han seleccionado?
- Si no se utilizan todas las variables (columnas) del conjunto de datos 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.

0.3) Objetivos particulares [0.5 puntos] (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.

0.4) 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, explicando cómo has obtenido la muestra.

ESTADÍSTICA DESCRIPTIVA

1) [1 punto] 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.

- ¿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 xx a xx). En caso de que existan ceros, se recomienda eliminar variantes o agruparlas en categorías más generales.

2) [1.5 puntos] Realiza los diagramas de caja y bigotes y los papeles probabilísticos normales de las 4 variables X_i e indica en una única tabla sus principales estadísticos: rango, rango intercuartílico, media, mediana, varianza, desviación típica, coeficiente de variación, coeficiente de asimetría estandarizado y coeficiente de curtosis estandarizado. A partir de este ejercicio, justifica e indica cuál/cuáles de las variables X_i puedes tomar como X_1.

3) [1.5 puntos] Comenta las diferencias observadas en la variable $\mathbf{X_1}$ entre las distintas variantes de $\mathbf{F_1}$:

- 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.
- Indica y calcula qué parámetros de posición y dispersión serían más adecuados en este caso para describir la pauta de variabilidad. Justifica la respuesta.

DISTRIBUCIONES DISCRETAS Y CONTINUAS

4) [1 punto] 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.

5) [1 punto] 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 y log-normal. De estos tres 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.

6) [1 punto] 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).

7) [1 punto] ¿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.

Fin 1ª parte del trabajo (Estadística descriptiva y distribuciones de probabilidad) Total: 8 puntos

INFERENCIA DE UNA POBLACIÓN

8) [1,5 puntos] Obtén tres intervalos de confianza con un nivel de confianza del 95%: para la media, la desviación típica y la varianza de tu variable X_1. Indica lo que representa cada uno de ellos.

9) [1,5 puntos] Para un nivel de significación del 5%, realiza los siguientes contrastes de hipótesis para tu variable X_1, explicando en cada caso el resultado obtenido:
a) H_0: m = 100, H_1: m ≠ 100
b) H_0: $\sigma^2 = 4$, H_1: $\sigma^2 \neq 4$
c) H_0: $\sigma = 1$, H_1: $\sigma \neq 1$

INFERENCIA DE DOS POBLACIONES

10) [1,5 puntos] Si tienes dos variables normales o casi normales, obtén el intervalo de confianza para la diferencia de sus medias con un nivel de confianza del 90%. Si no tienes dos variables normales o casi normales, divide tu variable X_1 según la variable F_1 y compara los dos subgrupos cuyo tamaño sea más grande. ¿Qué puedes decir respecto a las dos medias poblacionales?

11) [1 punto] Si tienes dos variables normales o casi normales, realiza el contraste de hipótesis: H_0: $\sigma_1^2 = \sigma_2^2$, H_1: $\sigma_1^2 \neq \sigma_2^2$. Si no tienes dos variables normales o casi normales, divide tu variable X_1 según la variable F_1 y, con los dos subgrupos cuyo tamaño sea más grande, realiza el contraste de hipótesis indicado. ¿Cómo interpretas el resultado del contraste?

12) [1 punto] Realiza un test de independencia para las variables F_1 y F_2, indicando cuál es la conclusión.

ANOVA

13) [0,5 puntos] Realiza un ANOVA para estudiar el efecto de la variable F_2 en la variable X_1. ¿Qué puede decirse de la significatividad de F_2? ¿Qué indican los intervalos LSD?

14) [1 punto] Realiza un ANOVA para estudiar el efecto tanto de la variable F_1 como de la variable F_2 en la variable X1. ¿Qué puede decirse de la significatividad de F_1 y de F_2? ¿Qué indican los intervalos LSD? ¿Ha cambiado el p-valor de F_2 respecto a la pregunta anterior? ¿A qué crees que se debe?

15) [1,5 puntos] Realiza un ANOVA para estudiar el efecto de la variable F_1, de la variable F_2 y de su interacción en la variable X1. ¿Qué conclusiones extraes a partir de la tabla del ANOVA y de las gráficas de los intervalos LSD?

Fin 2ª parte del trabajo (Inferencia en una población, inferencia en dos poblaciones y ANOVA)
Total: 9,5 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, **viernes 23 de mayo de 2025**; los alumnos deberán responder a los **objetivos nº 8 a 15 (incluido)**.

- El trabajo se enviará **en formato PDF** a través de Tareas de PoliformaT (se avisará cuando esté creada la tarea correspondiente a la 1ª entrega del trabajo). **No hay que enviarlo por correo electrónico ni subirlo al espacio compartido**. El nombre del documento llevará vuestros apellidos. Ejemplo: proyecto realizado por José García Pérez y Antonio Molina Hernández, el nombre sería: 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 19.

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 **nº de serie** (no hay que copiar y pegar el nº de serie directamente desde el documento PDF ya que puede incluir información del formato que haga que el nº 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 nº de serie, se envía por e-mail el código de activación.