Actividad de foro evaluable UC1

En este documento se describen las instrucciones para la actividad correspondiente a la UC1, de la asignatura 01GIIN Estadística

1. Objetivo

Utilizar las herramientas que se han explicado en clase para analizar el conjunto de datos.

IMPORTANTE: Esta actividad no es para evaluar vuestros conocimientos de programación o uso de softwares. No es obligatorio utilizar Excel, SPSS o RStudio para realizarla, si conocéis y preferís otro software (Python, Matlab,…) está permitido. Debe indicarse qué softwares se han utilizado para efectuar los cálculos/gráficos mostrados.

Realizamos una clase donde se ha usado RStudio, con la intención de que conozcáis la herramienta y podáis utilizarla para realizar el análisis, si así lo deseáis, también hemos visto cómo usar SPSS o Excel. La tarea se puede realizar con la herramienta con la que más cómodos os encontréis, incluso sin usar ningún software realizándola manualmente, pero mi recomendación, por vuestra comodidad, es que os apoyéis en alguno de los softwares vistos.

Se recomienda:

2. Instrucciones

- Ver las sesiones de las clases relacionadas con el análisis descriptivo.
- Descripción de los datos que se utilizarán para realizar el foro evaluable de la asignatura:

Se ha descargado el conjunto original de datos de países de: https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles . La tabla se ha tratado para facilitar al alumnado la comprensión y el tratamiento de los datos.

El conjunto contiene datos de diferentes indicadores estadísticos de países, extraídos de la web UNData de las Naciones Unidas en la que se recopilan datos de más de 20 fuentes internacionales. Los datos comprenden indicadores generales, económicos, sociales, ambientales y de infraestructura y se corresponden con 2017.

A continuación se presentan las características principales del conjunto de datos:

- El fichero de datos datosPaisesDelMundo.csv es de tipo “valores separados por comas”. Así, es un fichero de texto plano en el que cada observación corresponde a una línea y cada variable está separada de las demás por comas.
- El carácter separador de decimales es un punto, de manera que no se confunda con el separador entre variables.
- Se utilizan los valores ‘-99’, ‘…’, NA como entradas no válidas en la tabla.

  El conjunto de datos consta de 42 variables, descritas a continuación:

  - [1] "Pais": Nombre del país
  - [2] "Region": Región en la que se encuentra
  - [3] "Superficie_km2": Superficie en kilómetros cuadrados.
  - [4] "Poblacion_Miles_2017": Población en miles
  - [5] "Poblacion_Denskm2_2017." : Densidad de población, por km cuadrado
  - [6] "SexoProporcionMpor100F_2017.": Cantidad de hombres por cada 100 mujeres.
  - [7] "PIB_millonesUSD": Producto interior bruto en millones de dólares.
  - [8] "PIB_crecimientoAnual": Crecimiento anual del PIB.
  - [9] "PIB_PerCapita". PIB por cada habitante.
  - [10] "Economia_PorcentajeAgricola": Porcentaje de la economía basado en la agricultura.
  - [11] "Economia_PorcentajeIndustria": Porcentaje de la economía basado en la industria.
  - [12] "Empleados_PorcentajeAgricola": Porcentaje de la fuerza laboral dedicado a la agricultura.
  - [13] "Empleados_PorcentajeIndustria": Porcentaje de la fuerza laboral dedicado a la industria.
  - [14] "Desempleo_PorcentajePob": Porcentaje de desempleo
  - [15] "IndiceProduccionAgricola": Índice de producción agrícola. Indica la producción agrícola de cada año en relación con el período base 2004-2006. Incluye todos los cultivos excepto los forrajeros
  - [16] "IndiceProducionAlimentaria": Índice de producción alimentaria. Indica la producción de todos los productos alimentarios (comestibles y que contienen nutrientes) en relación con el período base 2004-2006
  - [17] "ComercioIntExportacion_millonesUSD": Volumen de exportaciones en dólares.
  - [18] "ComercioIntImportacion_millonesUSD": Volumen de importaciones en dólares.
  - [19] "CrecimientoPoblacion_TasaAnual": Tasa de crecimiento anual de la población en tanto por ciento.

- [20] "PoblacionUrbana_Porcentaje". Porcentaje de población urbana.
- [21] "CrecimientoPobUrbana_TasaAnual" Tasa de crecimiento de la población urbana en tanto por ciento.
- [22] "TasaFertilidac_NacVivosMujer": Tasa de fertilidad descrita por la cantidad de nacimientos vivos por mujer.
- [23] "MortalidadInfantil_porMilNac": Tasa de mortalidad infantil, mortalidad por cada 1000 nacimientos vivos.
- [24] "GastoSalud_PIB": Gasto total del PIB en salud.
- [25] "medicos_porMilpoblacion": Cantidad de médicos por cada mil habitantes.
- [26] "GastoEducacion_PIB": Gasto total del PIB en educación.
- [27] "participacionLaboralFem": Participación laboral de la población femenina en porcentaje.
- [28] "participacionLaboralMasc": Participación laboral de la población masculina en porcentaje.
- [29] "EducPrimariaTasaFem": Tasa bruta de ingreso, educación primaria, mujeres. Corresponde al número total de niñas que ingresan por primera vez al primer grado de educación primaria, independientemente de su edad, expresado como porcentaje de la población de niñas en edad oficial de ingreso. Esta tasa puede ser superior a 100% debido a que hay niñas o niños que ingresan a la primaria antes o después de la edad oficial.
- [30] "EducPrimariaTasaMasc": Tasa bruta de ingreso, educación primaria, hombres.
- [31] "EducSecundariaTasaFem": Tasa bruta de ingreso femenino en educación secundaria.
- [32] "EducSecundariaTasaMasc": Tasa bruta de ingreso masculino en educación secundaria.
- [33] "EducTerciariaTasaFem": Tasa bruta de ingreso femenino en educación superior.
- [34] "EducTerciariaTasaMasc": Tasa bruta de ingreso masculino en educación superior.
- [35] "AccesoAguaMejorPobUrbana": Porcentaje de población urbana con acceso a agua potable tratada.
- [36] "AccesoAguaMejorPobRural"; Porcentaje de población rural con acceso al agua potable tratada.
- [37] "EsperanzaVidaFem": Esperanza de vida al nacer de la población femenina en años.
- [38] "EsperanzaVidaMasc": Esperanza de vida al nacer de la población masculina en años.
- [39] "PorcentPoblacionJoven": Porcentaje de la población total entre 0 y 14 años.
- [40] "PorcentPoblacionVieja": Porcentaje de la población total mayor de 60 años.
- [41] "MigracionNum": Población residente en el extranjero

- [42] "MigracionPorcentaje" Porcentaje de población residente en el extranjero.

- Hacer como mínimo una contribución al foro, aplicando herramientas de las dadas en clase a alguna o varias de las variables para profundizar en el conocimiento del conjunto de datos. Se puede estudiar la distribución de alguna de las variables, haciendo un análisis descriptivo, representar datos gráficamente, se puede estudiar la relación entre dos variables, comparar las distribuciones, estudiar si hay valores atípicos…(es decir, hacer un análisis estadístico, basándonos en lo visto en la UC1 o de lo que podáis investigar por vuestra cuenta).
- Se valorará hacer más de una contribución e interaccionar con los compañeros sobre cómo hacer un análisis individual más completo.
- En el anexo, se pueden ver ejemplos de contribuciones aceptables (aunque mejorables). Se trata de reflexionar acerca de los valores que nos dan las herramientas tratadas en clase y no de meramente aplicarlas.

3. Entrega

- La entrega de la actividad serán las contribuciones realizadas en el **foro evaluable** del aula de la asignatura.
- La estética y formato de las contribuciones es libre, debiéndose escribir en español correcto y coherente.
- Deben **explicarse las respuestas e indicarse con qué software** se han obtenido, en el caso de haberse usado alguno. Imágenes o valores sueltos no se considerarán una respuesta válida.
- **NO ES NECESARIO ENTREGAR CÓDIGO ESCRITO**
- La fecha límite está indicada en la guía docente.
- La actividad es individual. Está terminantemente prohibido cualquier tipo de fraude, incluyéndose copia y plagio. En caso de detectarse, la actividad será suspensa y se reportará a la universidad para que tomen las medidas disciplinarias pertinentes.

4. Bibliografía

Se propone a los alumnos consultar, en caso de dudas, el siguiente material:

Juan Antonio González González, (2009) *Manual básico para SPSS*
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwj728eyufGEAxUJVaQEHS6yDi8QFnoECBMQAQ&url=https%3A%2F%2Fwww.fibao.es%2Fmedia%2Fuploads%2Fmanual_basico_spss_universidad_de_talca.pdf&usg=AOvVaw1yzpYFSeNf_EBE0OThZCFk&opi=89978449

Juan Carlos Vergara Schmalbach & Víctor Manuel Quesada Ibarguen, *Estadística básica con aplicaciones en ms Excel*. ISBN: 978-4-690-5503-8.
https://ebevidencia.com/wp-content/uploads/2017/05/estadistica-basica-con-excell.pdf

Pujol Jover, M. & Pujol Jover, M. (2017). *Análisis cuantitativo con R: matemáticas, estadística y econometría.*. Editorial UOC. https://go.exlibris.link/17PzkZlr

Tutoriales del Dr. Francisco Charte (U. Jaen):
Introduccion a R/Rstudio: https://fcharte.com/tutoriales/20170106-IntroR/
Importancia de visualizar: https://fcharte.com/tutoriales/20170110-ImportanciaVisualizacion/
Visualización (utiliza otro paquete distinto al que se ve en clase):
https://fcharte.com/tutoriales/20170112-VisualizacionBasica/
Análisis exploratorio: https://fcharte.com/tutoriales/20170108-AnalisisExploraR/

ANEXO:

Ejemplos de intervenciones en el foro.

Se plantea un ejemplo de foro en el que se puede ver contribuciones aceptables para la actividad. Se trata de un ejemplo orientativo, ya que cada conjunto de datos será diferente y en ocasiones se podrá o deberá contribuir en diferentes hilos. Se considera un conjunto de datos similar al trabajado en la sesión de RStudio, que con datos de los estudiantes de un curso de matemáticas y que contiene las siguientes columnas:

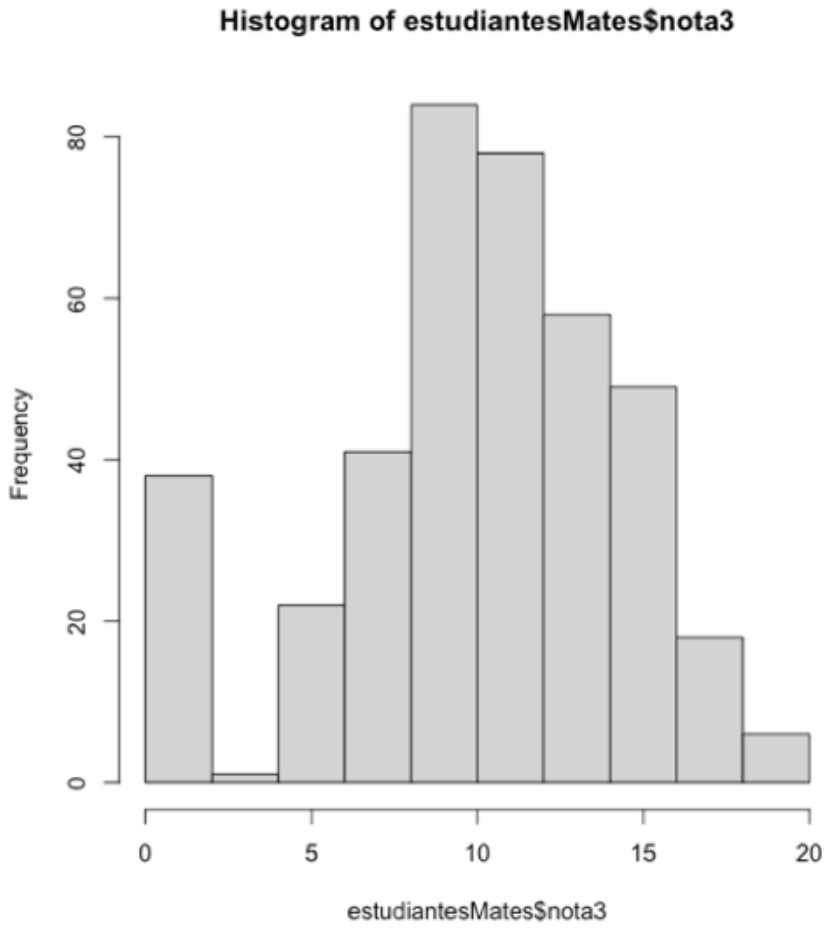
- X: índice (numérico)
- Edad: valor numérico con la edad en años de los estudiantes
- Sexo: factor: sexo de los estudiantes, puede tomar los valores 'M' (masculino) o 'F' (femenino.
- mediaAnterior: número entre el 0 y el 20 con la nota media en pruebas anteriores.
- Nota3: número entre el 0 y el 20 con la nota de la prueba final.

- Hilo 1: Calificaciones de los estudiantes en la prueba final
  - Contribución 1 ALUMNO A: (Medidas básicas de centralización y dispersión): Para empezar a conocer el desempeño de los estudiantes del curso vamos a realizar una descripción básica de la variable nota3 utilizando indicadores numéricos. Tenemos 390 valores de nuestra variable que presenta los siguientes indicadores numéricos de medidas de dispersión y centralización.

Indicador	Valor	Indicador	Valor
MEDIA	10,42	DESVIACIÓN TÍPICA	4,58
MEDIANA	11	RANGO	20

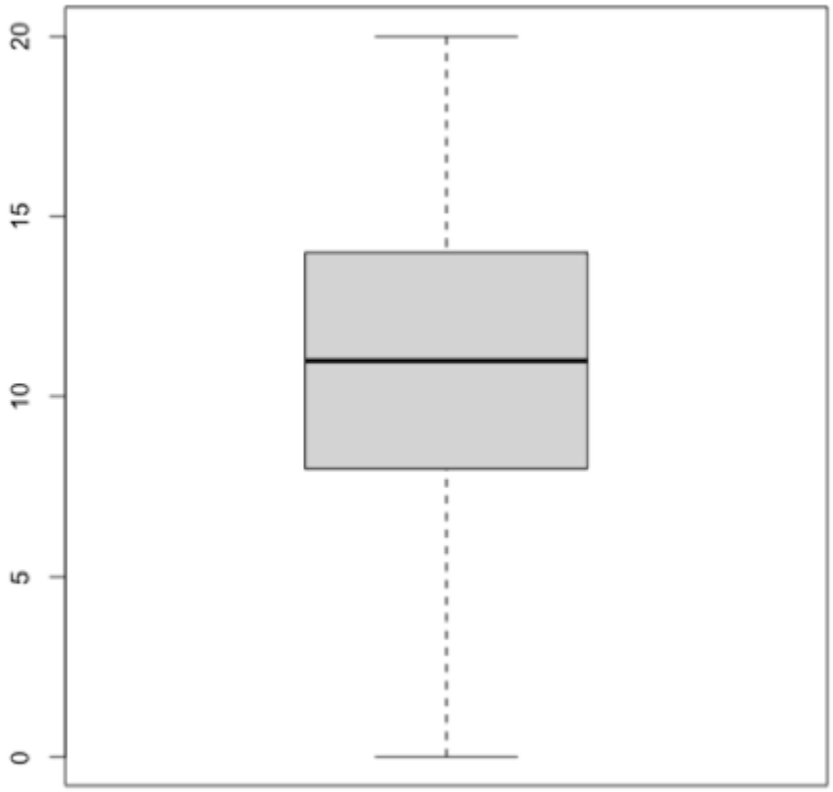
La media de las calificaciones es 10.42, que está aproximadamente en la mitad de su rango total, que va de 0 a 20 y representaría un aprobado. Este rango nos indica que se alcanzan tanto el mayor como el menor valor posible de notas. La desviación típica es de 4,58, que es un valor grande considerando los valores del rango y de la media. Teniendo esto en cuenta, se puede considerar que el valor de la mediana, de 11 está próximo al de la media, lo que apunta a una simetría global de la distribución, que habría que comprobar en detalle más adelante.

- Contribución 2 ALUMNO B: (Forma de la distribución)
  Para explorar con más detalle la distribución de calificaciones finales, se realiza un histograma, que se presenta en el siguiente gráfico.



Se observa el pico de la distribución en torno a un valor de 10 puntos, lo cual es consistente tanto con la media como con la mediana calculadas anteriormente. Llama la atención un pico con calificaciones entre cero y dos que contrasta con las calificaciones entre cero y cuatro, que son muy escasas. Si todas las calificaciones de ese intervalo son ceros, podría tratarse de alumnos que no se han presentado al examen, pero solo con la información del histograma no es posible saberlo. Esta anomalía, rompe la simetría global de la distribución que se podía plantear de los valores numéricos presentados por el compañero.

- Contribución 3 ALUMNO C: (Análisis de outliers)
  Se evalúa ahora si aparecen outliers o valores atípicos en las calificaciones finales de los alumnos. En primer lugar se presenta un boxplot de la variable nota 3, en el que la línea marca la mediana de la distribución, la caja los cuartiles y los bigotes, en este caso en máximo y mínimo valores, ya que no existen datos anómalos, no hay outliers en la distribución, como se puede observar en el boxplot.

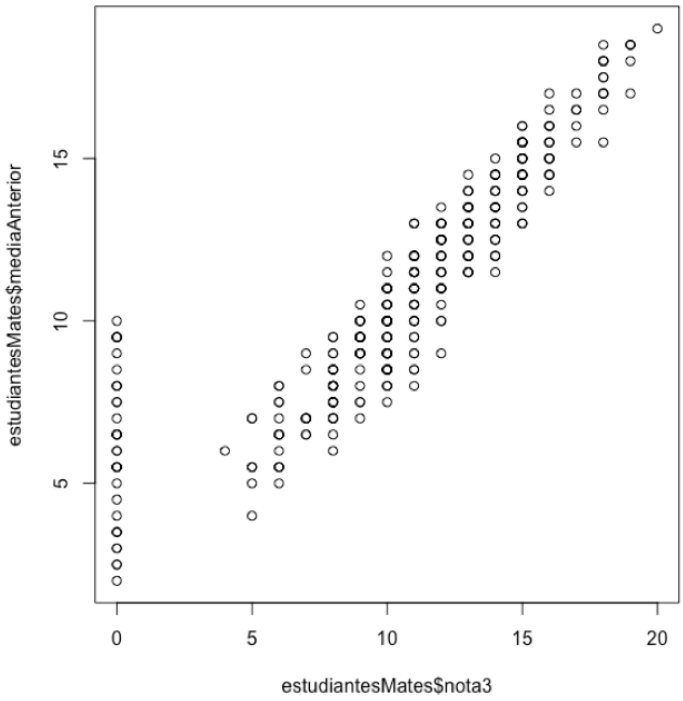


Sin embargo, considerando las contribuciones de los compañeros, nos podemos plantear que los valores menores que la media menos dos veces la desviación estándar pueden ser considerados atípicos, y en este caso, los valores menores que 1, que aparecerían en el pico de la distribución serían atípicos.

- Contribución 4 ALUMNO A: (Relación con las calificaciones previas)

Voy a considerar ahora si las calificaciones finales están relacionadas con las anteriores. (Esta intervención, ya sería una vez hayamos visto el tema 3).

En primer lugar visualizaremos una frente a otra :



Vemos una clara relación lineal y positiva entre el desempeño de los estudiantes a lo largo del curso, representado por la variable notaAnterior, y la calificación final del examen. Además, se confirma lo que ya indicaban anteriormente mis compañeros de una gran cantidad de alumnos cuya nota final es un cero, que muestran también un patrón diferente.
La relación lineal se confirma también con el coeficiente de correlación, que es alto y positivo en este caso con un valor de …

Otros hilos que podrían estudiarse, sobre este ejemplo dado…

- Hilo 2: Distribución de la edad....
- Hilo 3: Desempeño de la población femenina: Se separan los datos de estudiantes femeninas y se estudia su distribución...

Con estos ejemplos, tomados de intervenciones en foros de otros alumnos (con ello quiero decir, que pueden tener sus carencias o no), pero os podéis hacer una idea de lo que sería una intervención en el foro.

5. Bibliografía

Se propone a los alumnos consultar en caso de dudas el siguiente material:

Juan Antonio González González, (2009) *Manual básico para SPSS* https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwj728eyufGEAxUJVaQEHS6yDi8QFnoECBMQAQ&url=https%3A%2F%2Fwww.fibao.es%2Fmedia%2Fuploads%2Fmanual_basico_spss_universidad_de_talca.pdf&usg=AOvVaw1yzpYFSeNf_EBE0OThZCFk&opi=89978449

Juan Carlos Vergara Schmalbach & Víctor Manuel Quesada Ibarguen, *Estadística básica con aplicaciones en ms Excel*. ISBN: 978-4-690-5503-8. https://ebevidencia.com/wp-content/uploads/2017/05/estadistica-basica-con-excell.pdf

Pujol Jover, M. & Pujol Jover, M. (2017). *Análisis cuantitativo con R: matemáticas, estadística y econometría.*. Editorial UOC. https://go.exlibris.link/17PzkZlr

Tutoriales del Dr. Francisco Charte (U. Jaen):
Introduccion a R/Rstudio: https://fcharte.com/tutoriales/20170106-IntroR/
Importancia de visualizar: https://fcharte.com/tutoriales/20170110-ImportanciaVisualizacion/
Visualización (utiliza otro paquete distinto al que se ve en clase): https://fcharte.com/tutoriales/20170112-VisualizacionBasica/
Análisis exploratorio: https://fcharte.com/tutoriales/20170108-AnalisisExploraR/