

Actividad 2: Análisis descriptivo e inferencial

Enunciado

Análisis Estadístico - Semestre 2025.1

Índice

1. Esperanza de vida	3
1.1. Análisis descriptivo	3
1.2. Hipótesis	3
1.3. Test	4
1.4. Aplicación del test	4
1.5. Desarrollo	4
1.6. Resultados	4
1.7. Interpretación	4
2. Desigualdad a lo largo del tiempo	4
2.1. Análisis descriptivo	4
2.2. Hipótesis	5
2.3. Test	5
2.4. Aplicación del test	5
2.5. Desarrollo	5
2.6. Resultados	5
2.7. Interpretación	5
3. Desigualdad según el tipo de régimen	5
3.1. Preparación de los datos	6
3.2. Hipótesis	6
3.3. Test	6
3.4. Aplicación del test	6
3.5. Desarrollo	6
3.6. Resultados	6
3.7. Interpretación	6

4. Proporción de países que cumplen con el objetivo de no agotamiento de bosques	6
4.1. Análisis descriptivo	7
4.2. Hipótesis	7
4.3. Test	7
4.4. Aplicación del test	7
4.5. Desarrollo	7
4.6. Resultados	7
4.7. Interpretación	7
5. Conclusiones y comunicación de resultados	8

Introducción

En esta actividad aplicaremos un análisis descriptivo e inferencial sobre el conjunto de datos Global Sustainability que hemos preprocesado en la actividad anterior.

Recordamos que el conjunto de datos “Global Sustainability” (<https://www.kaggle.com/datasets/trucue/worldsustainabilitydataset/data>) recopila información sobre el desempeño de 173 países en relación con diversos indicadores de sostenibilidad a lo largo de un periodo de 19 años. El conjunto de datos se ha generado a partir de la fusión de múltiples fuentes internacionales, entre ellas el World Bank DataBank. Incluye un total de 54 variables, que abarcan diferentes dimensiones de la sostenibilidad:

- Económica: Se proporcionan indicadores como el porcentaje de exportaciones del país, producto nacional bruto, crecimiento del ingreso nacional bruto ajustado per cápita, ahorro neto ajustado, etc.
- Social: Contiene indicadores como la esperanza de vida al nacer, índice de Gini (desigualdad de ingresos), educación, etc.
- Ambiental: Indicadores que tienen en cuenta la degradación de recursos naturales y los daños ambientales.

Este conjunto de datos nos ofrece múltiples posibilidades para explorar los indicadores vinculados a los objetivos de desarrollo sostenible (ODS) de las Naciones Unidas. Debido a la extensión del conjunto de datos, nos centraremos sólo en algunas preguntas de investigación concretas:

1. PR1: ¿La esperanza de vida es mayor en los países con alto nivel de ingresos que en el resto?
2. PR2: ¿La desigualdad se reduce en un período de 10 años?
3. PR3: ¿La proporción de países con mayor desigualdad es diferente entre los países con régimen democrático y aquellos con régimen autocrático?
4. PR4: ¿La proporción de países que cumplen con el objetivo de no agotamiento neto de bosques es mayor que 50 %?

Salvo que se indique lo contrario, utilizaremos un nivel de confianza del 95 %.

Se adjunta la descripción del conjunto de datos (Data Dictionary.xls) y los datos preprocesados (WorldSustainabilityDS_clean.csv). Debido a la cantidad de datos con valores perdidos, éstos no han sido reemplazados o eliminados para preservar la información original del conjunto de datos. Cuando sea preciso, deberemos tratar estos valores perdidos, como se indica más adelante.

A tener en cuenta para realizar la actividad:

- Es necesario entregar el archivo Rmd y el archivo de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Para facilitar la corrección, los archivos deben entregarse por separado en el aula virtual. Es decir, hay que subir cada archivo por separado.
- Se debe respetar la misma numeración de los apartados que en el enunciado.
- No se pueden incluir listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificultaría la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden utilizar las funciones **head** y **tail**, que solo muestran unas líneas del archivo de datos.
- Se valorará la precisión de los términos utilizados (es necesario utilizar de forma precisa la terminología estadística).
- También se valorará la concisión en la respuesta. No se trata de realizar explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos, de forma clara y concisa.
- No está permitido copiar código de actividades anteriores. El código que presentéis debe ser de creación propia.
- Para dudas técnicas sobre R y el entorno de programación, tenéis a vuestra disposición el aula Laboratorio R.

1. Esperanza de vida

En primer lugar, atendiendo al ODS 3 (garantizar una vida sana y promover el bienestar para todos en todas las edades, y asegurar así el desarrollo sostenible), analizamos la esperanza de vida de los distintos países.

En concreto, nos preguntamos:

PR1: ¿La esperanza de vida es mayor en los países con alto nivel de ingresos que en el resto?

Para dar respuesta a la pregunta, usaremos análisis descriptivo e inferencial.

Seguid los pasos que se indican a continuación.

1.1. Análisis descriptivo

Antes de realizar el análisis inferencial, realizaremos un análisis descriptivo de este dato. Mostrad un gráfico comparando la esperanza de vida según el nivel de ingresos. Podéis usar un gráfico tipo boxplot.

Usad las variables: SP.DYN.LE00.IN (LE), Income y Region. Para hacer el código más comprensible, podéis renombrar como LE (Life Expectancy) la variable de esperanza de vida.

1.2. Hipótesis

Para el análisis inferencial, comparad los países con ingresos altos (“High income”) en relación con el resto. **Usad como muestra el último año del conjunto de datos (2018).**

Escribid las hipótesis estadísticas a partir de la pregunta de investigación.

1.3. Test

¿Qué tipo de test estadístico es apropiado para dar respuesta a la pregunta? Justificad vuestra elección.

1.4. Aplicación del test

Usad una librería de R para aplicar el test seleccionado.

Nota: Descartad temporalmente los valores ausentes en la variable esperanza de vida, sin eliminar los datos del conjunto original. Así se preservan el resto de datos para posteriores análisis. Para ello, cread un conjunto de datos que contenga solo el año 2018 y los datos existentes de esperanza de vida e ingresos.

1.5. Desarrollo

Desarrollad un código propio que implemente la prueba que habéis usado en el apartado anterior y verificad que obtenéis el mismo resultado. Podéis usar las funciones *qnorm*, *pnorm*, *qt*, *pt*, etc.

1.6. Resultados

Usando vuestra función, mostrad los resultados, incluyendo medias, valor observado, estadístico de contraste y valor p.

1.7. Interpretación

A partir de los resultados anteriores, dad respuesta a la pregunta de investigación planteada.

2. Desigualdad a lo largo del tiempo

En esta sección, examinaremos el ODS 10 (reducir la desigualdad dentro y entre países). En concreto, nos fijamos en el índice GINI, el cual mide la desigualdad en la distribución del ingreso en un país, donde 0 representa igualdad perfecta (todas las personas tienen el mismo ingreso) y 100 desigualdad máxima (una persona recibe todo el ingreso).

La pregunta de investigación es:

PR2: ¿La desigualdad se reduce en un período de 10 años?

Para ello, tomaremos el último año del conjunto de datos, 2018, y como año inicial, el 2008. Considerar, primero todos los países del mundo y luego, solo los países de la región “Europe and Northern America”.

Para hacer el código más comprensible renombrar la variable `SI.POV.GINI` como `GINI`.

Seguid los pasos que se indican a continuación.

2.1. Análisis descriptivo

Para orientar la posible respuesta, mostrad en un gráfico la desigualdad de los países en el año 2008 y 2018. Podéis usar un gráfico tipo boxplot. Debéis realizar un gráfico incluyendo todos los países y otro gráfico con los países de la región “Europe and Northern America”.

Asimismo, mostrad una tabla con los cinco países con más desigualdad y los cinco países con menos desigualdad (en todo el mundo y considerando solo la región “Europe and Northern America”).

2.2. Hipótesis

Escribid las hipótesis estadísticas a partir de la pregunta de investigación.

2.3. Test

¿Qué tipo de test estadístico es apropiado para dar respuesta a la pregunta? Justificad vuestra elección.

2.4. Aplicación del test

Usad una librería de R para aplicar el test seleccionado.

Nota: copiad en un conjunto de datos temporal sólo los datos necesarios para este análisis y eliminad los datos que tengan valores perdidos en las variables de interés.

2.5. Desarrollo

Desarrollad un código propio que implemente la prueba que habéis usado en el apartado anterior y verificad que obtenéis el mismo resultado.

Nota: No se debe repetir código en el documento. En este sentido, si procede, podéis usar código propio desarrollado en esta actividad. Podéis usar las funciones *qnorm*, *pnorm*, *qt*, *pt*, etc.

2.6. Resultados

Usando vuestra función, mostrad los resultados, incluyendo medias, valor observado, estadístico de contraste y valor p.

Nota: Debéis mostrar el resultado incluyendo todos los países, así como sólo los países de la región Europa y América del Norte.

2.7. Interpretación

A partir de los resultados anteriores, dad respuesta a la pregunta de investigación planteada.

3. Desigualdad según el tipo de régimen

En este apartado nos preguntamos si la proporción de países con altos niveles de desigualdad es diferente según el régimen. Para ello, abordamos la pregunta de investigación siguiente:

PR3: ¿La proporción de países con mayor desigualdad es diferente entre los países con régimen democrático y aquellos con régimen autocrático?

Consideramos “mayor desigualdad” los países con un índice GINI superior a 40. Escoged una muestra del año 2018.

Seguid los pasos que se indican a continuación.

3.1. Preparación de los datos

Cread una variable en el conjunto de datos “HighInequality” la cual toma el valor 1 si el índice GINI es superior a 40 y 0 en el caso contrario. Agrupad los países según si el régimen es democrático o autocrático. Mostrad una tabla que muestre la proporción de países con alta desigualdad según si el tipo de régimen es democrático o autocrático. Escoged como muestra el año 2018.

Nota: al igual que en apartados anteriores, trabajad con un conjunto de datos temporal que contenga sólo los datos necesarios y eliminad los registros que contengan valores perdidos.

3.2. Hipótesis

Escribid la hipótesis nula y alternativa.

3.3. Test

¿Qué tipo de test estadístico es apropiado para dar respuesta a la pregunta? Justificad vuestra elección.

3.4. Aplicación del test

Usad una librería de R para aplicar el test seleccionado.

3.5. Desarrollo

Desarrollad un código propio que implemente la prueba que habéis usado en el apartado anterior y verificad que obtenéis el mismo resultado.

3.6. Resultados

Usando vuestra función, mostrad los resultados, incluyendo los valores calculados, valor observado, estadístico de contraste y valor p.

3.7. Interpretación

A partir de los resultados anteriores, dad respuesta a la pregunta de investigación planteada.

4. Proporción de países que cumplen con el objetivo de no agotamiento de bosques

En esta sección nos centramos en el ODS número 13 (acciones para combatir el cambio climático y su impacto). Para ello, nos fijaremos en el índice “Net forest depletion”, que se podría traducir como “índice de agotamiento neto de bosques”. Este índice mide el valor económico de la pérdida de bosques de un país cuando la extracción de madera supera el crecimiento natural, expresado como porcentaje del ingreso nacional bruto, indicando si el uso forestal es sostenible o no.

Concretamente, queremos dar respuesta a la pregunta:

PR4: ¿La proporción de países que cumplen con el objetivo de no agotamiento neto de bosques es mayor que 50 %?

Por “no agotamiento de bosques” nos referimos al índice de agotamiento de bosques (NY.ADJ.DFOR.GN.ZS) igual o inferior a 0.

Para dar respuesta a esta pregunta, tomaremos los datos del año 2018. Seguid los pasos que se indican a continuación.

4.1. Análisis descriptivo

Mostrad un gráfico de vuestra elección para analizar visualmente el índice de pérdida de bosques de los países.

Para hacer el código más descriptivo podéis renombrar la variable *NY.ADJ.DFOR.GN.ZS* por *FOR*.

4.2. Hipótesis

Escribid las hipótesis estadísticas a partir de la pregunta de investigación.

4.3. Test

¿Qué tipo de test estadístico es apropiado para dar respuesta a la pregunta? Justificad vuestra elección.

4.4. Aplicación del test

Usad una librería de R para aplicar el test seleccionado.

Nota: al igual que anteriormente, cread un conjunto de datos que sólo contenga los datos necesarios para este análisis y eliminad los registros con valores perdidos.

4.5. Desarrollo

Desarrollad un código propio que implemente la prueba que habéis usado en el apartado anterior y verificad que obtenéis el mismo resultado.

4.6. Resultados

Mostrad los resultados del test, incluyendo el valor obtenido de la muestra, el valor observado y el valor p.

4.7. Interpretación

A partir de los resultados anteriores, dad respuesta a la pregunta de investigación planteada.

5. Conclusiones y comunicación de resultados

Resumid en una tabla los resultados del análisis. Incluid en esta tabla la pregunta de investigación, el tipo de test, valor p obtenido y una breve interpretación del resultado.

Puntuación de la actividad

- Apartado 1 (20 %)
- Apartado 2 (20 %)
- Apartado 3 (20 %)
- Apartado 4 (20 %)
- Apartado 5 (10 %)
- Calidad del informe dinámico (10 %)