Actividad. Análisis descriptivos en R

**Una captura de pantalla de un celular con letras

Descripción generada automáticamente**

Objetivos

Aplicarás los conceptos aprendidos sobre **bioestadística descriptiva**. Utilizarás un *dataset* que contiene información de la expresión de 46 genes en 65 pacientes, cada uno con distintos tipos de tratamiento y características tumorales. El objetivo principal, en primer lugar, se pretende comprobar la normalidad de la distribución de la expresión de los genes para asegurar los supuestos estadísticos necesarios. Luego, se calcularán los descriptivos de la base de datos (variables bioquímicas, síntomas, variables sociodemográficas y de los genes) en: 1) función del tratamiento y del tipo de tumor; 2) según la edad categorizando por la mediana. En la tabla descriptiva, tendrás que reflejar las medias y desviaciones estándar en caso de ser datos paramétricos; o medianas y rangos intercuartílicos si son datos no paramétricos. Por último, en la tabla se tendrá que reflejar el valor P que hace referencia al contraste de hipótesis correspondiente a cada una de las variables de la tabla descriptiva para determinar si existen diferencias significativas entre el tratamiento y del tipo de tumor.

Pautas de elaboración

Esta actividad consistirá en dos partes principales relacionado con datos de diferentes genes: 1) **elaboración de código** para que extraigas datos descriptivos y los reflejes en una tabla, y apliques los test de contraste de hipótesis en el documento HTML; y 2) **breve interpretación de los datos descriptivos** del *dataset* en el documento HTML.

A continuación verás: un apartado que explica el *dataset*, mientras que el siguiente son las preguntas que deberás de contestar.

***Dataset* de expresión de genes**:para la realización de esta parte de la actividad, deberás cargar el *dataset* de interés «Dataset expresión genes.csv», el cual se trata de una base de datos de 65 pacientes que contiene información de la expresión de 46 genes con diferentes funciones (para más información, ver el apartado de **Información de interés del dataset** después de la rúbrica). Además de estas variables, contiene otras variables de interés como el **tratamiento** (A o B) que siguen cada paciente, **tipo de tumor** que tienen (colorrectal, pulmón y mama) y la **extensión tumoral** (localizado, metastásico o regional). Por último, se recoge información de variables **bioquímicas**, **síntomas** y **otras variables** sociodemográficas.

Tras importarlo, deberás **responder a las siguientes cuestiones**, teniendo en cuenta todos los gráficos, datos descriptivos y de contraste de hipótesis generados mediante las librerías ***ggplot2, stats, car, nortest o gtsummary***:

1. Comprobar la normalidad de los genes y realizar una interpretación de los resultados obtenidos:
   * Utilizar pruebas estadísticas de normalidad (recuerda que Shapiro-Wilk funciona cuando la N es pequeña, si no puedes usar otras como addtest).
   * Refleja los valores p del contraste de hipótesis **en una tabla modelo 1** adjunta al final de documento en la sección “Extensión y formato”.
2. Calcular y analizar estadísticas descriptivas de los valores sin transformar (media + desviación estándar si son paramétricas, mediana + rango intercuartílico (p25-p75) si no lo son) de las variables bioquímicas, síntomas, sociodemográficas y de los genes:
   * En función del tratamiento (1er nivel) y del tipo de tumor (2º nivel), crea una tabla descriptiva (**tabla modelo 3**) únicamente con las variables de genes.
   * En función de la edad en 2 categorías según la mediana (**categoría 1**: edad <percentil 50 de edad; **categoría 2**: edad >=percentil 50 de edad), crea una tabla descriptiva (**tabla modelo 3**) únicamente con las variables de genes.
   * Consejos:
     + Las tablas tienen que ser legibles, entendibles, ordenadas y limpias. Si hay decimales, lo normal es poner 1, a excepción de valores muy bajos que puede extenderse a los que se consideren para poder entenderse el número. Si el valor numérico es muy pequeño, puede optarse por usar el formato científico (por ejemplo: 2\*10-6). Los valores P suele ponerse 3 decimales.
     + Lo más rápido es hacer las tablas descriptivas con la librería gtsummary. Para ello apóyate en lo visto en clase. Además, puedes ver aquí ejemplos y explicaciones: <https://www.danieldsjoberg.com/gtsummary/>. Si las generas con esta librería, no hace falta que generes las tablas modelo 2 y 3.
     + Si optas por gtsummary, como verás en la tabla 2, al haber 2 niveles de estratificación (1 tratamiento y 2 tipo tumor), tendrás que usar en primer lugar la función tbl\_strata con strat y .tbl\_fun; en segundo lugar tbl\_summary con by, statistics, type y digits; y en tercer lugar add\_p con testy pvalue\_fun.
     + Si no optas por gtsummary, mi recomendación es que crees *datasets* independientes y que saques los descriptivos para reflejarlo en la tabla del modelo 2.
     + Para la tabla 3 en la que tienes que generar una nueva variable categórica a partir de una variable numérica, utiliza la función quantile para sacar los percentiles y posteriormente usa la función cut.
3. Interpretar brevemente los resultados obtenidos de las tablas (cada tabla 150 palabras mínimo).

Como consejo y de forma opcional, se puede emplear *loops* para sacar las estadísticas descriptivas (ver más aquí: <https://rpubs.com/Miguel_Tripp/forloops>) o *lapply* (<https://r-coder.com/lapply-en-r/>).

Extensión y formato

Para la resolución de esta actividad, deberás entregar 3 archivos:

* Un único fichero **R Markdown** (*.Rmd*) con todo el código y texto Markdown que hayas generado. No existe un límite de extensión para este fichero. Hay que indicar los siguientes argumentos:
  + title: «Resolución Actividad 1 máster Bioinformática UNIR (2023)»
  + author: «Nombre y 2 apellidos»
  + date: «yyyy-mm-dd»
  + output: html\_document

Para crear código y generarlo en el fichero R Markdown poner siempre:

```{r}

empezar aqui con el código

```

* Un fichero **HTML** generado a partir de dicho archivo R Markdown. Asegúrate de que aparecen todas las figuras, el texto correspondiente y todos los cuadros de código R que hayas introducido en el archivo *.Rmd* anterior.
* Un fichero **Word** únicamente con las 3 tablas modelos. Asegúrate de que aparecen los títulos correspondientes de las tablas, las notas a pie de tabla, indicando en todo caso abreviaturas en orden alfabético, test utilizados para sacar las estadísticas, notas a pie de tabla que se desee reflejar para aclarar cualquier asunto relevante de la tabla, etc. Los valores numéricos se representan con 1 decimal, a excepción de los valores p que se representan con 3 decimales. Los datos descriptivos se representan en una misma celda la media y desviación estándar [por ejemplo: 3.4 (3)] al igual que la mediana y el RIQ [3.4 (1.3 a 9.8)]. El **formato de tablas que se debe de seguir** es el siguiente:

**Tabla 1.** {Aquí poner título}

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Test utilizado** | **Valor p** | **Interpretación** |
| Variable 1 | XX | XX | XX |
| Variable 2 | XX | XX | XX |
| … | XX | XX | XX |

{Aquí poner pie de tabla}

**Tabla 2**. {Aquí poner título}

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **tratA** | | | | **tratB** | | | |
|  | **CCR (n=XX)** | **CM (n=XX)** | **CP (n=XX)** | **p-value** | **CCR (n=XX)** | **CM (n=XX)** | **CP (n=XX)** | **p-value** |
| **Variable** | Media (SD) o  P50 (RIQ) | Media (SD) o  P50 (RIQ) | Media (SD) o  P50 (RIQ) |  | Media (SD) o  P50 (RIQ) | Media (SD) o  P50 (RIQ) | Media (SD) o  P50 (RIQ) |  |
| Variable 1 | XX (XX) | XX (XX) | XX (XX) | XXX | XX (XX) | XX (XX) | XX (XX) | XXX |
| Variable 2 | XX (XX) | XX (XX) | XX (XX) | XXX | XX (XX) | XX (XX) | XX (XX) | XXX |
| … | XX (XX) | XX (XX) | XX (XX) | XXX | XX (XX) | XX (XX) | XX (XX) | XXX |

{Aquí poner pie de tabla}

**Tabla modelo 3**. {Aquí poner título}

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Edad\_cat < percentil 50** | **Edad\_cat ≥ percentil 50** |  |
| **Variable** | Media (SD) o P50 (RIQ) | Media (SD) o P50 (RIQ) | **p-value** |
| Variable 1 | XX (XX) | XX (XX) | XXX |
| Variable 2 | XX (XX) | XX (XX) | XXX |
| … | XX (XX) | XX (XX) | XXX |

{Aquí poner notas a pie de tabla}

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Análisis descriptivos en R | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Pruebas | Construcción correcta de las pruebas y descriptivas tal y como se solicitan en el enunciado de la actividad | 3 | 30% |
| Justificación | Justificación detallada en la respuesta a las preguntas planteadas y en el análisis de los gráficos generados | 3 | 30% |
| Cuestiones | Resolución de las cuestiones planteadas de forma correcta | 2 | 20% |
| Originalidad | Generación de un informe R Markdown original, empleando temas o estilos CSS, índices, etc. | 1 | 10% |
| Orden | Exposición de lo que se pide en el enunciado de forma clara y ordenada | 1 | 10% |
|  |  | **10** | **100 %** |