

Taller de práctica Prueba 1

Operaciones en R (Taller Introductorio)
Conceptos básicos de estadística (sección 1.1)
Medidas de Localización (sección 1.2)
Medidas de escala (sección 1.3)

Parte 1: Operaciones en R

El conjunto de datos diabetes.csv (copie el link para cargar la base en Colab) proviene originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es estudiar de forma diagnóstica si un paciente tiene diabetes, en función de ciertas medidas de diagnóstico incluidas en el conjunto de datos. Se impusieron varias restricciones a la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de ascendencia indígena Pima. Las columnas de la base de datos con las siguientes:

- Pregnancies: Para expresar el número de embarazos.
- Glucose: Para expresar el nivel de Glucosa en sangre (mg/dL).
- BloodPressure: para expresar la medida de la presión arterial distólica (mm Hg).
- SkinThickness: Para expresar el grosor de la piel (mm).
- Insulin: para expresar el nivel de insulina en sangre (mg/dL).
- BMI: Para expresar el índice de masa corporal.
- DiabetesPedigreeFunction: Para expresar el porcentaje de Diabetes.
- Age: Para expresar la edad en años.
- Outcome: Para expresar el resultado final de tener diabetes, 1 es Sí y 0 es No.

A continuación, haciendo uso de la base de datos mencionada, explique cada uno de los siguientes códigos asociados a la base de datos. Considere que las variables se muestran en el mismo orden en el que están ordenadas.

```
1. # Considere que la base de datos se cargo con el nombre "datos"  
datos = datos[,c(8,4,9)]
```

```
2. datos = datos[datos$Age < 41 | datos$SkinThickness <= 30,]
```

```
3. datos = datos[datos$SkinThickness %in% 40:50 & datos$Age > 36, c(1,3)]
```

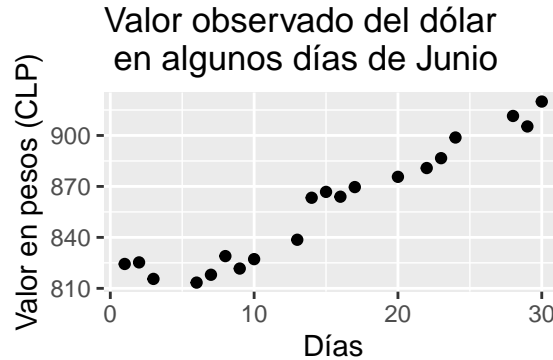
```
4. datos = datos[1:365,]
```

```
5. datos = datos[datos$Age > 55 & datos$Glucose < 120,]
```

```
6. datos = datos[datos$Age %in% 20:45 | datos$BMI >= 30.1, c(1,2)]
```

Parte 2: Conceptos básicos de estadística

1. Observe el siguiente gráfico, y explique conceptualmente, cuál es el dato y cuál sería la información. ¿Qué se necesita para transitar de un concepto a otro (¿y en qué dirección se transita?)?



2. Realice un esquema o mapa conceptual que relacione los conceptos de población, muestra, parámetro, estadístico, estimador y estimación.
3. Suponga, que está interesado en predecir la variación porcentual del IPC para el siguiente mes considerando un rezago a un mes (es decir, respecto al mes anterior), para ello, usted sabe que este valor será similar al promedio de las variaciones porcentuales del IPC a rezago de un mes de los meses anteriores (históricamente), sin embargo, utiliza los últimos doce meses, dando como resultado un 3%. A continuación, especifique, la población, la muestra, el parámetro, el estadístico, el estimador y la estimación del estudio. ¿Qué sucede si se considerarán para el cálculo los últimos 24 meses o los últimos 36 meses?, explique el concepto estadístico detrás de esta situación.
4. Suponga, que está interesado en estudiar el promedio anual histórico del porcentaje de empleo femenino poblacional, para ello, usted aproxima este valor mediante el cuadrado del promedio del empleo femenino poblacional anual del periodo 2001 - 2012, obteniendo un resultado del 37.2%. A continuación, elabore un listado o mapa conceptual para especificar la población, la muestra, el parámetro, el estimador y la estimación del estudio. ¿Este valor es una buena aproximación?, explique.
5. ¿Por qué es incorrecto hablar de una muestra representativa? Justifique.

Parte 3: Medidas de Localización y Escala

1. ¿Cuál es la desventaja del promedio como medida de resumen? Explique.
2. Mencione las ventajas y desventajas de utilizar el rango, la varianza (o desviación estándar) y el coeficiente de variación para comparar la variabilidad de los datos de dos más muestras.
3. La siguiente base de datos contiene registros del Índice de Confianza del Consumidor (ICC). Este indicador de confianza del consumidor proporciona una indicación de la evolución futura del consumo y el ahorro de los hogares. Un indicador por encima de 100 señala un aumento en la confianza de los consumidores hacia la situación económica futura, como consecuencia de la cual son menos propensos a ahorrar y más inclinados a gastar dinero en compras importantes en los próximos 12 meses. Los valores por debajo de 100 indican una actitud pesimista hacia la evolución futura de la economía, lo que posiblemente resulte en una tendencia a ahorrar más y consumir menos.

Las variables que contiene la base de datos son las siguientes:

- Locacion: lugar en donde se mide el ICC (FRA = Francia, POL = Polonia, OECD = OCDE, ESP = España, BEL = Bélgica, ITA = Italia, DEU = Alemania).
- Mes: corresponde al mes en el que se realiza la medición del índice.
- Ano: corresponde al año en el que se realiza la medición del índice.
- ICC: valor del índice de confianza del consumidor.

La base de datos al cargarse en R se ha guardado con el nombre de `datos`.

a) Interprete las siguientes salidas de código.

```
mean(datos$ICC[datos$Locacion == "ESP"])
## [1] 100.342
mean(datos$ICC[datos$Locacion == "DEU"])
## [1] 100.7433
median(datos$ICC[datos$Ano %in% c(2017:2020) & datos$Mes %in% c(2:5)])
## [1] 100.8683
median(datos$ICC[datos$Locacion == "OECD" & datos$Mes %in% c(3,6,7)])
## [1] 100.5179
```

b) Interprete las siguientes salidas, y compárelas según sea posible.

```
var(datos$ICC[datos$Locacion == "ESP"])
## [1] 9.03605
var(datos$ICC[datos$Locacion == "DEU"])
## [1] 1.605925
sd(datos$ICC[datos$Locacion == "ESP"])/abs(mean(datos$ICC[datos$Locacion == "ESP"]))*100
## [1] 2.995757
sd(datos$ICC[datos$Locacion == "DEU"])/abs(mean(datos$ICC[datos$Locacion == "DEU"]))*100
## [1] 1.257901
```