Lectura DS1.

Presentación de resultados.
Análisis descriptivo univariado

Índice

1. Recolección de datos
2. Análisis descriptivo univariado
   - Definiciones
   - Descriptores numéricos y gráficos
3. Material de estudio extra

1. RECOLECCIÓN DE DATOS

Definiciones

Datos: medidas u observaciones de interés.

Unidad muestral: elemento o individuo de la muestra.

Variable: característica de interés de un elemento.

- **Numérica**: el valor de la variable es un número.
  - **Continua**: la variable puede tomar cualquier valor en un intervalo.
  - **Discreta**: la variable puede tomar una cantidad numerable de valores numéricos diferentes.
- **Categórica**: el valor de la variable no es un número.
  - **Ordinal**: valores discretos que tienen un orden claro; excelente, bueno, mediocre, malo.
  - **Nominal**: valores discretos sin un orden específico; blanco, negro, rojo, verde, azul.

Tabla o matriz de datos: tabla donde las unidades muestrales corresponden a las filas y las variables corresponden a las columnas.

Población: conjunto de todos los individuos o elementos de interés.

Muestra: subconjunto de la población, idealmente se espera que la muestra sea representativa de la población.

Censo: muestra correspondiente a toda la población; típicamente es muy caro, en algún sentido, trabajar con un censo.

Tipos de muestreo

- **Muestreo no probabilístico**: La representatividad depende de cosas difíciles de controlar.
  - **Muestreo por conveniencia**: se toman los datos que se tienen, sin preocuparse por su procedencia. La representatividad de la muestra depende de múltiples factores fuera del control del investigador, no es fácil saber qué impacto tiene la conveniencia de la muestra.
  - **Muestreo por juicio**: un experto dictamina que elementos de la población son seleccionados para la muestra, la representatividad de la muestra depende del juicio del experto.
- **Probabilístico**: La representatividad depende del azar, el tamaño de la muestra y las suposiciones de estructura.
  - **Muestreo aleatorio simple (MAS)**: se seleccionan al azar los elementos de la población que participan en la muestra. Todas las muestras posibles deben tener la misma probabilidad.
  - **Muestreo estratificado**: La población está particionada en grupos disjuntos homogéneos de acuerdo a algunas características. Se hace muestreo aleatorio simple en cada estrato, el número de elementos a tomar en cada estrato es, frecuentemente, proporcional al tamaño del estrato.
  - **Muestreo por conglomerados**: La población está particionada en conglomerados disjuntos, similares entre sí. Primero se toma una muestra aleatoria simple de los conglomerados y luego se toma una muestra aleatoria simple de cada conglomerado seleccionado.

Sesgo

Al tomar una muestra de una población, podríamos obtener un conjunto de elementos que no sea representativo de la población.

A las distintas formas en las que la representatividad de una muestra se desvía de la población se les conoce como sesgos y al diseñar un muestreo se tratan de evitar los diversos sesgos. Un sesgo común es el sesgo de selección, que tiene diversos sub tipos.

Por ejemplo se puede cometer sesgo de selección cuando:

- Se toma una muestra de voluntarios y el objeto de estudio está conectado al voluntariado.
- Se toma una muestra por conveniencia y el objeto de estudio está relacionado a lo que hace que los individuos sean accesibles.

Un ejemplo interesante de sesgo de selección es el sesgo de supervivencia.

Los sesgos pueden alterar significativamente los resultados de un estudio, hasta el punto de hacerlo inválido.

Comandos

Estos son algunos comandos útiles, ejecute todos los comando y comente los resultados de al menos 3 de ellos (estos pueden ser ejemplos de aplicaciones, recomendaciones o cosas interesantes del comando).

- `help`: muestra la ayuda sobre un comando, parámetros, opciones y detalles.
  `help(ls)` muestra la ayuda sobre el comando ls.
  `help("library")` muestra la ayuda sobre el comando library
  `help("Arithmetic")` muestra la ayuda sobre las operaciones aritméticas.
- `help("<-")` muestra la ayuda sobre la asignación de variables.
- `install.library`: instala una librería específica, disponible en los repositorios de R.
  `install.library("readr")` instala la librería readr.
- `library`: carga un conjunto de objetos y funciones disponibles, de una librería instalada, al ambiente de trabajo.
  `library(readr)` carga la librería readr, que ofrece diversas funciones para cargar datos al ambiente de trabajo.
- `ls`: da la lista de los objetos disponibles en el ambiente de trabajo.
  `ls()` muestra todos los objetos disponibles en el ambiente de trabajo.
- `rm`: elimina objetos del ambiente de trabajo.
  `rm(list = ls())` elimina todos los objetos cargados al ambiente de trabajo.
- `sample`: sirve para tomar una muestra aleatoria de una población, con o sin reemplazo.
  `sample(1:10, 3, replace = FALSE)` toma 3 elementos de los números del 1 al 10, sin reemplazo.
  `sample(1:10, 3, replace = TRUE)` toma 3 elementos de los números del 1 al 10, con reemplazo.

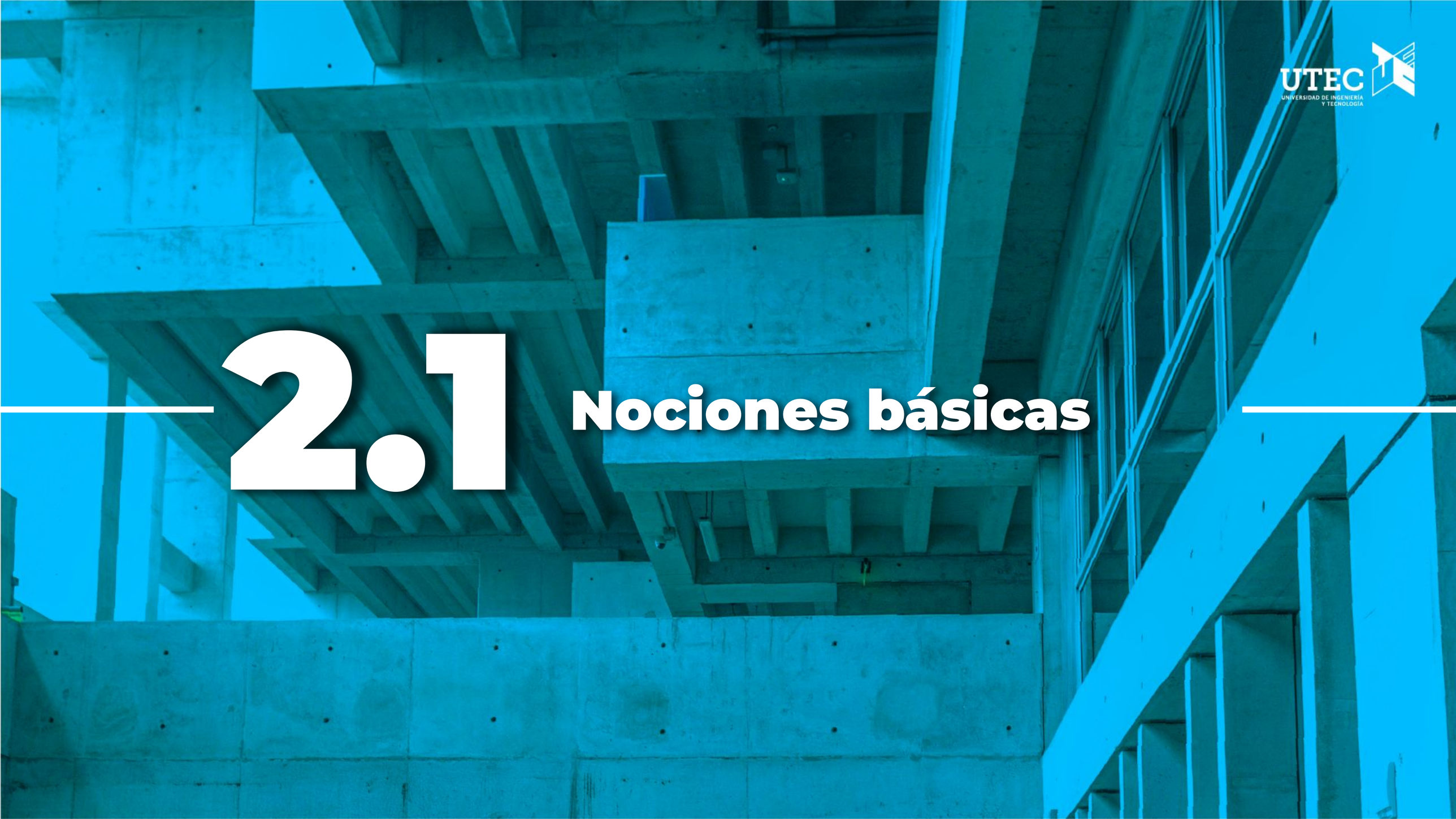
Recomendaciones y ejercicios

Abra un archivo en formato R Notebook y ejecute las siguientes instrucciones, intercalando los fragmentos de código con narrativa para dejar claro qué hace cada fragmento de código.

El documento debe reflejar que hizo y las razones por las que lo hizo.

- Lea la ayuda de los comandos `nrow` y `ncol`.
- Identificar el directorio de trabajo.
- Ejecute los ejemplos que se ofrecen en la lámina anterior.
- Lea la ayuda disponible en la opción *Help | Markdown Quick Reference* del menú de ayuda de RStudio.
- Instale y cargue las librerías readr y dplyr.
- Lea la ayuda del comando `read_csv` de la librería readr.
- Descargue el archivo compartido Datos Diversos en formato CSV y colocarlo en el directorio de trabajo.
- Cargue a la variable `DF` el contenido del archivo que colocó en su directorio de trabajo.
- Determine el número de observaciones y el número de variables del archivo.

2. Análisis descriptivo univariado



2.1 Nociones básicas

Justificación

Si una tabla de datos tiene muchas observaciones o muchas variables, *entender* el conjunto de datos puede ser difícil.

Muchos trabajos involucran múltiples tablas de datos, relacionadas entre sí de maneras complejas; el propósito del análisis descriptivo es simplificar el conjunto de datos de manera que se puedan entender las relaciones entre las distintas variables y la estructura de las mismas.

Para lograr esto es fundamental entender que todos los descriptores que podamos imaginar, debido a que, por diseño deben simplificar, va a terminar introduciendo errores e imprecisiones. En el fondo queremos unos pocos números, tablas o gráficas, que puedan destilar la esencia de los datos.

Claramente el propósito del estudio debe guiar los errores e imprecisiones que estamos dispuestos a tolerar.

Nos concentramos en descriptores para una sola tabla de datos, generalizando de la manera obvia a un mayor número.

Entre los descriptores más sencillos están:

- Cantidad y tipo de cada variable: número de columnas en la tabla, las columnas pueden ser de tipos distintos.
- Cantidad y descripción de cada observación: número de filas en la tabla, tamaño de la muestra.
- Número de datos faltantes: total de valores por variable y en la tabla, que de alguna manera faltan.
- Número de observaciones completas: total de observaciones que no tienen datos faltantes.

Consecuencias

Los descriptores anteriores son relativamente sencillos, podríamos denominarlos como descriptores de tipo y descriptores de cantidad. Los primeros, normalmente se incluyen en lo que se denomina la libreta de código de una tabla o *codebook* en inglés. Allí se especifican cosas como el tipo de la variable, restricciones a sus valores y detalles similares.

Los descriptores de cantidad hablan directamente sobre la complejidad de la tabla y pueden especificar detalles como cobertura de la muestra. Por ejemplo una muestra enorme donde muy pocas observaciones estén completas podría ser problemática si los datos faltantes están acumulados en las partes que se desean estudiar o podría estar bien si los datos faltantes están repartidos a lo largo de las observaciones sin sesgos particulares.

En general, mientras más datos hay, potencialmente más complejo es el problema.

Es importante entender que las distinciones entre variables de distintos tipos deben ser útiles, por ejemplo:

- El peso de una persona, habitualmente se considera como una variable continua, pero si se registra el peso en kilogramos enteros, se vuelve claramente discreta; en ambos casos la variable es positiva.
- La distancia típicamente es una variable continua, pero si se registra, en metros, con precisión de una billonésima de milímetro, entonces estamos en presencia de una variable técnicamente discreta, pero que desde el punto de vista práctico es continua.
- El DNI de una persona es claramente un número entero, pero se maneja como variable categórica ordinal.

2.2 Descriptores numéricos y gráficos

Descriptores

Algunas variables, a pesar de ser aleatorias, cumplen con una serie de restricciones o tienen algunas características que hacen que describir un conjunto numeroso de ellas sea relativamente sencillo, siempre que se entiendan cosas básicas respecto a la variable como puede ser la forma de su distribución y algunos valores clave.

Estos valores clave, se denominan **parámetros de la distribución** y pueden servir, junto con el tipo de variable, como herramientas poderosas para describir conjuntos numerosos de observaciones de esa variable.

Los parámetros de una distribución pueden verse reflejados en algunos de los descriptores numéricos a considerar.

Por ejemplo, si observamos suficientes lanzamientos de un dado justo, esperaríamos ver que el mínimo de estos lanzamientos sea 1 y el máximo 6, esperaríamos ver todos los enteros del 1 al 6 con aproximadamente la misma frecuencia y esperaríamos que el promedio de todos los lanzamientos fuese cercano a 3.5.

Los valores 1 y 6 juegan el papel de parámetros para la distribución de resultados de un dado justo.

No debería extrañarnos que un dado justo de 20 caras, tenga mínimo 1, máximo 20, y un promedio de 10.5.

En la segunda parte del curso veremos modelos parametrizados para variables aleatorias. Entenderlos puede servir para describir los conjuntos de datos.

Descriptores numéricos

Estos son algunos de los descriptores que aplican a diversas variables numéricas y tienen interpretaciones sensatas.

- **Descriptores de posición**: informan sobre la ubicación de los datos.
  - **Promedio**: los datos están alrededor del promedio, el promedio es el *centro de masa* de los datos.
  - **Mediana**: al menos la mitad de los datos son mayores o iguales a la mediana menores o iguales a la mediana. La mediana, en general, no es única.
  - **Cuantil q %**: al menos el q % de los datos es menor o igual al cuantil q %, al menos el (1-q) % de los datos es mayor o igual que el cuantil q %. El cuantil q %, en general no es único. La mediana es el cuantil 50 % o el segundo cuartil.
  - **Mínimo**: el valor del menor dato. Corresponde al cuantil 0 %.
  - **Máximo**: el valor del mayor dato. Corresponde al cuantil 100 %.
  - **Moda**: el o los datos que más se repiten.
- **Descriptores de dispersión**: informan sobre el nivel de variación de los datos.
  - **Rango**: diferencia entre el máximo y el mínimo, representa el máximo observado de variación de los datos. Rango grande, dispersión alta, rango pequeño, dispersión baja.
  - **Varianza**: promedio de los cuadrados de la diferencia el dato y el promedio. Poblacional o muestral.
  - **Desviación estándar**: raíz cuadrada de la varianza. Poblacional o muestral.
  - **Rango intercuartil**: diferencia entre el tercer y el primer cuartil.

Recomendaciones y ejercicios

Abra un archivo en formato R Notebook y cargue el archivo Datos Diversos en formato CSV en la tabla `DF`.

- Determine el tipo de cada variable de la tabla y establezca los valores teóricos de cada una, de ser posible.
- Determine el número de datos faltantes por variable y en total.
- Determine el número de observaciones y el número de observaciones completas.
- Describa cada una de las variables en términos de algunos de los descriptores mencionados.
- Interprete los resultados obtenidos por variable.
- Grafique la estatura en función del peso ¿Qué puede concluir?

Recuerde que el propósito de trabajar en el formato R Notebook es intercalar la narrativa con los resultados del procesamiento de los datos usando las diversas funciones de R.

Ejercicios adicionales de dificultad moderada:

- Construya una tabla de frecuencias relativa de los signos del zodíaco, ¿qué conclusiones puede sacar?
- El coeficiente de variación de una variable es el promedio de la variable dividido por su desviación estándar, determine los coeficientes de variación del peso y la estatura.

Ejercicios adicionales difíciles:

- Invente un índice para el *grado de suciedad* de una variable categórica. Mientras más alto el valor del índice, más sucia está la variable. Calcule el valor de su índice para el sexo y signo del zodíaco.
- ¿Puede hacer algo similar para variables numéricas? ¿Cómo?

Tablas y descriptores gráficos

Dependiendo de los valores que pueden tomar los datos, estos se pueden agrupar en categorías y mostrar sus frecuencias absolutas o relativas.Estos son algunos de los descriptores que aplican a diversas variables numéricas y tienen interpretaciones sensatas.

Descriptores tabulares:

los datos se pueden agrupar en categorías y se presentan en forma tabular.

- **Tabla de frecuencias absolutas**: se totaliza el número de observaciones por categoría. Las categorías pueden ser las originales o agrupaciones de los valores originales.
- **Tabla de frecuencias relativas**: se normalizan los totales por el tamaño de la muestra.

Descriptores gráficos:

Se presentan los datos o algunos descriptores numéricos en un formato gráfico. Una imagen vale más que mil palabras.

- **Diagrama de puntos**: presenta los valores numéricos a lo largo de una línea.
- **Histograma** y **diagrama de barras**: son las versiones gráficas de las tablas de frecuencias, numérica y categórica.
- **Diagrama de caja y bigotes** o ***boxplot***: es una caricatura de un histograma, supone que los datos están alrededor de un *centro*.
- **Diagrama de dispersión**: Muestra la relación entre dos o variables.

Datos atípicos

Un valor de una variable es **atípico** si presenta algunas características que lo hacen sustancialmente diferente de los demás datos en la misma colección. Hay múltiples razones por las que un dato podría ser atípico, por ejemplo, el dato podría ser demasiado grande o demasiado pequeño en comparación con los otros datos, o el dato podría tener una parte decimal cuando los otros datos son todos enteros.

Hay que prestar especial atención a los datos atípicos pues podrían señalar errores de medición o transcripción, pero es fundamental entender que la designación de atípico es arbitraria en general y no necesariamente indica que un error ha ocurrido.

Definitivamente se debe entender que, eliminar los datos atípicos de un conjunto de datos, en general, no es una buena práctica ya que reduce el tamaño de la muestra y potencialmente sesga los resultados al eliminar observaciones que efectivamente ocurrieron.

En la mayoría de los casos, resulta interesante tratar de explicar los datos atípicos en vez de eliminarlos; esto lleva a una descripción más fiel de los datos y a mejores conclusiones basadas en la evidencia que estos proveen.

Algunas visualizaciones ofrecen un mecanismo para identificar los valores atípicos. Entre ellas estudiaremos el diagrama de caja y bigotes o *boxplot*, como se le conoce en inglés.

Robustez

Un descriptor es **robusto** si tolera la inclusión de datos atípicos, potencialmente erróneos, sin distorsionarse demasiado.

Por ejemplo, diremos que la mediana es más robusta que el promedio para determinar la centralidad de unos datos.

Supongamos que las longitudes, en centímetros, de los pétalos de 5 flores se han registrado correctamente como:

Longitud	5.7	9.2	8.5	6.9	1.9

Su promedio es 6.44 y su mediana 6.9. Esta es la realidad de la muestra suponiendo que no hemos cometido errores.

Supongamos que al momento de la transcripción se ha cometido un error, omitiendo el punto en el quinto valor:

Longitud	5.7	9.2	8.5	6.9	19

Promedio: 9.86 mediana: 8.5

En ambos casos, tanto la mediana como el promedio se han visto afectados, pero se puede observar que la variación de la mediana está contenida entre los valores que la encierran (es decir, modificando un solo valor, podemos hacer la mediana variar entre 5.7 y 8.5, mientras que el promedio puede hacerse arbitrariamente grande o pequeño dependiendo del tamaño del error cometido. En ambos casos, el quinto valor era atípico en algún sentido.

Comandos

Estos son algunos comandos útiles, ejecute todos los comandos y comente los resultados de cada comando.

- `is.na`: determina si un valor es un dato faltante, retornando `TRUE` es ese caso y `FALSE` en caso contrario.
- `complete.cases`: retorna verdadero para cada observación de la tabla que no tenga datos faltantes.
- `sum`: retorna la suma de los valores; en R los valores lógicos son numéricos, `TRUE` vale 1 y `FALSE` vale 0.
- `mean`: retorna el promedio de los valores; al colocar `na.rm = TRUE` descarta los valores faltantes.
- `median`: retorna una mediana de los valores; se usan convenciones para hablar de *la mediana*.
- `min`: retorna el mínimo de los valores; al colocar `na.rm = TRUE` descarta los valores faltantes.
- `max`: retorna el máximo de los valores; al colocar `na.rm = TRUE` descarta los valores faltantes.
- `var`: retorna la varianza muestral de los valores; tiene el parámetro `na.rm`.
- `sd`: retorna la desviación estándar muestral de los valores; tiene el parámetro `na.rm`.
- `table`: clasifica en una tabla de frecuencia absoluta los valores.
- `stripchart`: muestra una gráfica de puntos, puede ser útil para pocos valores; pueden apilarse los puntos.
- `hist`: muestra un histograma de los valores; clases automáticas o controladas con el parámetro `breaks`.
- `barplot`: muestra un diagrama de barras; se usa típicamente en conjunción con `table`.
- `boxplot`: muestra un diagrama de caja y bigotes; presentación gráfica de descriptores numéricos.
- `plot`: muestra un diagrama de dispersión; sirve para ver la relación entre dos variables numéricas.

Recomendaciones y ejercicios

Abra un archivo en formato R Notebook y cargue el archivo Datos Diversos en formato CSV en la tabla `DF`.

- Determine promedio y mediana de los tiempos de reacción, ¿qué significan los resultados?
- Determine el rango, rango intercuartil y desviación estándar de los tiempos de reacción.
- Utilice un *boxplot* para determinar si hay observaciones atípicas en la variable tiempos de reacción.
- ¿Cuántas observaciones atípicas hay en el tiempo de reacción?
- ¿Puede concluir algo sobre los tiempos de reacción atípicos?
- ¿Cuántos datos hay entre el primer y el tercer cuartil?
- ¿Cuál es la moda de los tiempos de reacción?
- haga un histograma de los tiempos de reacción ¿Es la distribución unimodal? ¿bimodal?

Repita el análisis anterior sobre la variable estatura.

Ejercicios adicionales de dificultad moderada:

- Algunos de los valores de estatura y tiempo de reacción son erróneos. ¿Qué valores cree que son erróneos?
- Elimine solo los valores erróneos de cada variable y repita el análisis anterior. ¿Qué puede concluir?

Ejercicios adicionales difíciles:

- Utilice la función `cut` para agrupar las estaturas originales en intervalos de 10 cm de ancho.
- Utilice una visualización apropiada para describir gráficamente los resultados de la operación anterior.
- ¿Cuál es la clase modal? ¿Es la distribución unimodal? Grafique el promedio de las estaturas en la visualización.

3. Material de estudio extra: viñetas

Viñetas

A partir de ahora, deben prestar especial atención a las viñetas que publicaremos.

Las viñetas son cuadernos de trabajo en formato R Notebook, con preguntas y sus respuestas. En los módulos podrán encontrar el código fuente (Rmd) y la viñeta tejida (html). La idea es que descarguen las viñetas y las modifiquen para sus propósitos.

En algunas viñetas resolveremos problemas no estructurados y en otras exploramos conceptos importantes para el curso; es importante que las descarguen y las usen, modificándolas para enriquecer sus apuntes de clase.

Los diversos descriptores numéricos y gráficos que veremos en las próximas clases están disponibles en las viñetas **Descripción de datos Parte 1** y **Descripción de datos Parte 2**. En la viñeta **Dispersión y consideraciones** encontrarán reflexiones sobre gráficos de dispersión. La viñeta **Análisis de datos del CDC** contiene un caso trabajado en detalle siguiendo la línea del libro *OpenIntro to Statistics*.

En todas las viñetas hay información sobre el uso de R para manipular tablas y hacer filtros sobre ellas, por favor no dejen de hacerlas.

UTEC

UNIVERSIDAD DE INGENIERÍA
Y TECNOLOGÍA