Unidad 2

Análisis exploratorio de datos

IU Digital de Antioquia

INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA

ORD No 74 de 2017 · VIGILADA MinEducación

www.iudigital.edu.co

Resumen:

La estadística, como disciplina, se dedica al análisis y búsqueda de relaciones entre datos, procesándolos y analizándolos para entender su significado y comportamiento.

Este proceso es conocido como *caracterización estadística*, y nos permite identificar la estructura de los datos a partir de su tipología. Esto con el fin de seleccionar adecuadamente los algoritmos que debemos emplear en nuestros proyectos.

Para aprovechar el potencial de la información contenida en los datos, en este curso tendrás la oportunidad de aplicar conceptos y técnicas para clasificar, modelar e interpretar datos en diferentes contextos. Lo cual te será de gran utilidad para la toma de decisiones en cualquier tipo de organizaciones.

A través del desarrollo de las actividades de aprendizaje de cada unidad, podrás comprender conceptos básicos como *Población* y *Muestra*. Además, conocerás algunas herramientas de gran ayuda para entender y representar de manera gráfica el comportamiento de los datos.

Gracias a los contenidos dispuestos en este curso, entenderás que los datos son algo más que un cúmulo de textos y números. De hecho, si se caracterizan correctamente, pueden ser un instrumento útil para comprender el comportamiento de cualquier sistema u organización.

Presentación general del curso:

En el contexto de los procesos de análisis de la información en grandes volúmenes, existe una premisa bastante común: "*cuando a un sistema entra basura, sale basura".* Esta expresión básicamente nos advierte que, independiente del volumen de los datos que ingresan, lo relevante es la calidad.

Así se pone en evidencia el aporte de la estadística como disciplina, enfocada en analizar y buscar relaciones entre los datos, acopiando, procesando, analizando e interpretándolos para entender su comportamiento.

La caracterización estadística de los datos nos permite identificar su estructura y, a partir de su tipología, en estadios superiores, se podrá identificar qué tipo de algoritmos podrán ser utilizados. De manera que el papel que juega la estadística en la representación de los datos, de manera visual y descriptiva, permite aprovechar el potencial de la información contenida en

los mismos, a través de la modelación con técnicas como, por ejemplo, el aprendizaje automático.

Por otro lado, es importante señalar que la estadística y el *Big Data* comparten los mismos objetivos: identificar la calidad de los datos y representarlos visualmente para establecer si son claros. Además, apoyan la toma de decisiones a partir de la comprensión del comportamiento de los datos.

Este curso nos permitirá desarrollar habilidades para la representación y comprensión de datos, toda vez que los volúmenes de información son cada vez mayores, y requieren de análisis acordes a las complejidades y retos a los que se enfrentan las organizaciones.

Objetivo de aprendizaje:

Aplicar conceptos y técnicas para clasificar, modelar e interpretar datos en diferentes contextos.

Resultado de aprendizaje:

Al finalizar el curso, el estudiante estará en capacidad de:

- Aplicar los conceptos estadísticos sobre diversos conjuntos de datos.

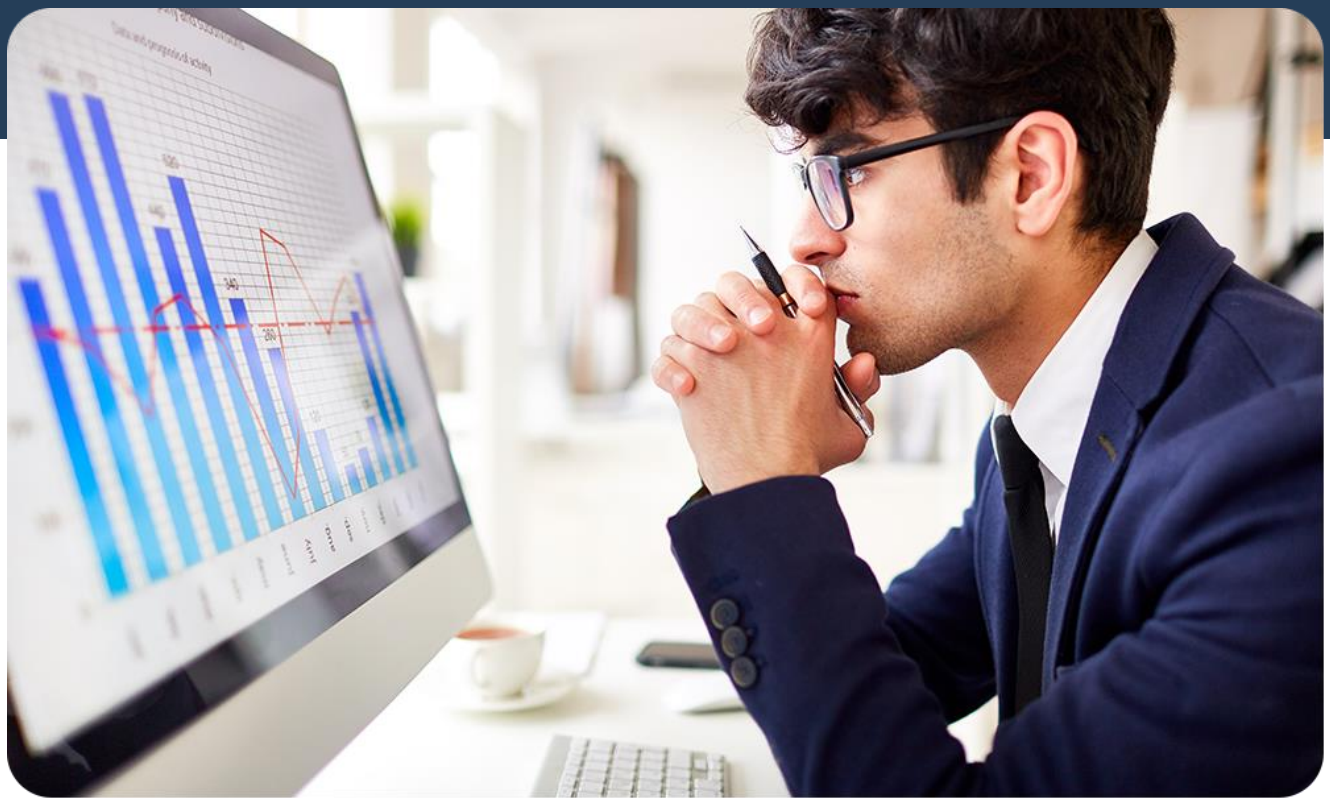
Pregunta orientadora:

Hasta hace algunos años, era sorprendente la curva de crecimiento que tenía la tecnología, ligada al aumento en el uso de celulares, tabletas y todo tipo de dispositivos electrónicos. Además, el auge de las redes sociales y las páginas web suponían un fenómeno sin precedentes.

En la actualidad, todas estas interacciones derivadas de la tecnología son tan comunes que han aumentado exponencialmente la información que se genera cada segundo.
De hecho, con la cantidad de información que se produce en las redes sociales, es posible perfilar el comportamiento, opiniones y preferencias de los usuarios.

¿En un mundo donde las tecnologías digitales, las redes sociales y el *Internet de las Cosas* configuran contextos en los que se producen y almacenan grandes cantidades de datos, es posible que, a través de ellos, sea posible comprender algunos fenómenos o realidades?

Unidad 2:
Análisis exploratorio de datos



Introducción a la unidad 2

Al interior del Gobierno, las empresas, las comunidades o, incluso, en el contexto personal, las decisiones están ligadas a la información con la que cada individuo u órgano cuenta. Dicho de otra manera, los datos influyen en la mayoría de las decisiones que se toman en todos los niveles organizacionales, desde las empresas globales hasta los pequeños núcleos familiares.

Los datos nos pueden resultar de mucha utilidad porque, de acuerdo con sus valores, es posible identificar su comportamiento para establecer la relación que tienen con las dinámicas más comunes de la vida. Esto nos permite entender cómo operan las dinámicas sociales, de mercado, de seguridad, de producción y demás, que hacen parte de un universo tan extenso como la cantidad de datos que se puedan encontrar.

Así pues, en esta unidad abordaremos nociones básicas para interpretar diferentes fenómenos y dinámicas a partir de los datos.

Nuestro recorrido empezará por conocer la definición de tamaño y las formas de selección de una muestra, además de su relación con el concepto de población. Igualmente, identificaremos elementos asociados al error de estimación y el nivel de confianza, y aprenderemos a representar el comportamiento de los datos, algo que resulta muy útil para comunicar de manera intuitiva tales conductas.

Disfruta este contenido y aprovéchalo para mejorar la forma en que interpretas el mundo, tanto en el contexto laboral y empresarial, como en el contexto personal.

Objetivos de aprendizaje de la Unidad 2

- Clasificar datos según los conceptos de muestra y población.
- Interpretar conjuntos de datos a partir de las medidas estadísticas de tendencia central, posición y dispersión.
- Representar visualmente el comportamiento de datos a partir de gráficos estadísticos.

Unidad 2.
Actividad de aprendizaje 1: Explorando los datos

En esta actividad de aprendizaje abordaremos los conceptos de: muestra, población, medidas estadísticas y representación gráfica del comportamiento de los datos.

Por consiguiente, los temas que estudiaremos a partir de este momento son:

- Tipos de muestreo
- Elección y tamaño de la muestra
- Cálculo de medidas
- Gráficos estadísticos

Al finalizar el estudio de los temas propuestos, estarás en capacidad de resolver el problema que se ha planteado en la parte final, mediante el cual podrás aplicar todos los conceptos abordados durante la Unidad 2.

Desarrollo de los temas de la actividad de aprendizaje 1 de la Unidad 2

1. Tipos de muestreo

Antes de comenzar a desarrollar este tema, es preciso reconocer inicialmente un conjunto de definiciones que nos permitan comprender los procesos de identificación y selección de una muestra.

Población

Conjunto completo de individuos, objetos, eventos (podemos llamarlos: observaciones) que tienen una característica en común.

Criterios de selección

Se refiere a las características que deben tener los individuos y que los define como parte de la población. Se pueden entender también como las condiciones de inclusión o exclusión de las observaciones, las cuales se pueden considerar de manera independiente.

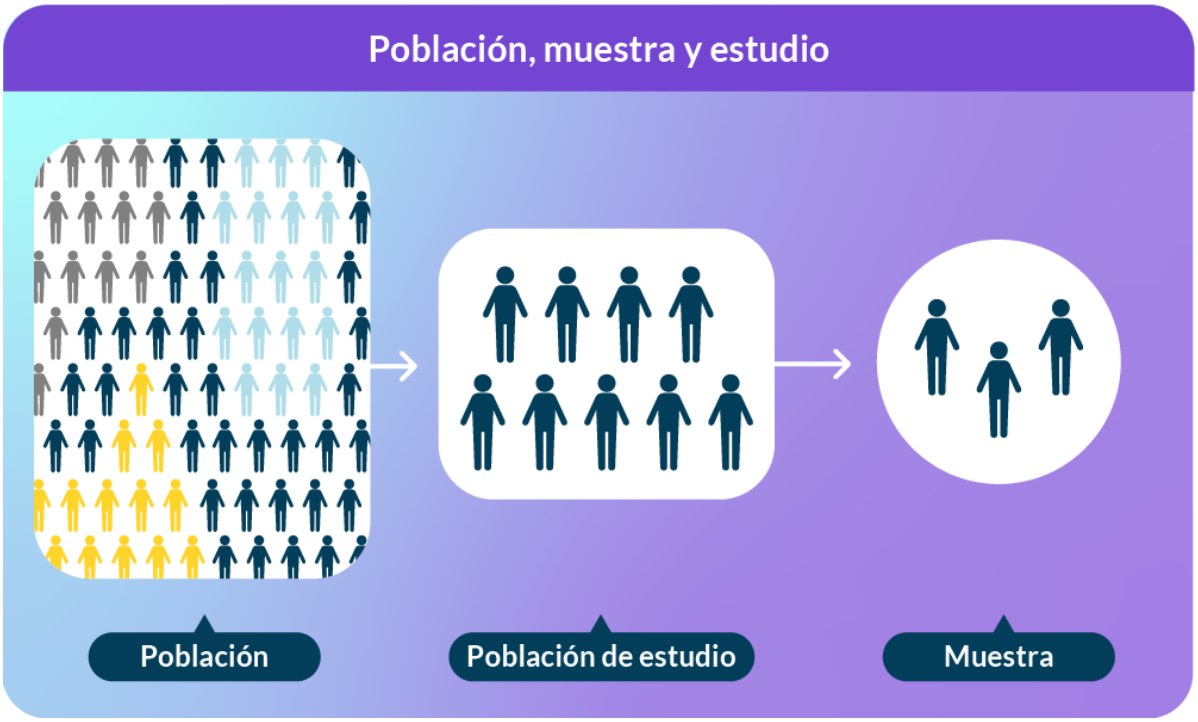
Dicho de otra forma, los criterios de selección actúan como un filtro para obtener la población de estudio.

El propósito de estos criterios es estandarizar las variables que se emplearán para el análisis de los datos, de manera que sea posible controlar los factores de confusión u otras que pueden afectar el resultado que se espera. Dichas variables o factores de confusión distorsionan la medida de asociación, entre otras variables del fenómeno de estudio.

Población de estudio

Hablemos ahora de población y muestra, conceptos que debemos tener claros para desarrollarlos en el artículo de investigación.

Figura 1. Población, muestra y población de estudio



La ***figura 1*** ejemplifica los tres conceptos que recién mencionamos. Representa una *población* sobre la cual se aplican los criterios de selección con los que se busca obtener la *población de estudio.* Finalmente, a partir de esta última, se define la *muestra*.

Aunque el término de *población* se utiliza comúnmente para referirse a individuos humanos, también puede incluir objetos, entidades, eventos, observaciones, etc.

Por otro lado, la *población de estudio* corresponde al conjunto de unidades que cumplen con los criterios de selección para el análisis de los datos, lo cual nos permitirá entender ciertos aspectos de su comportamiento.

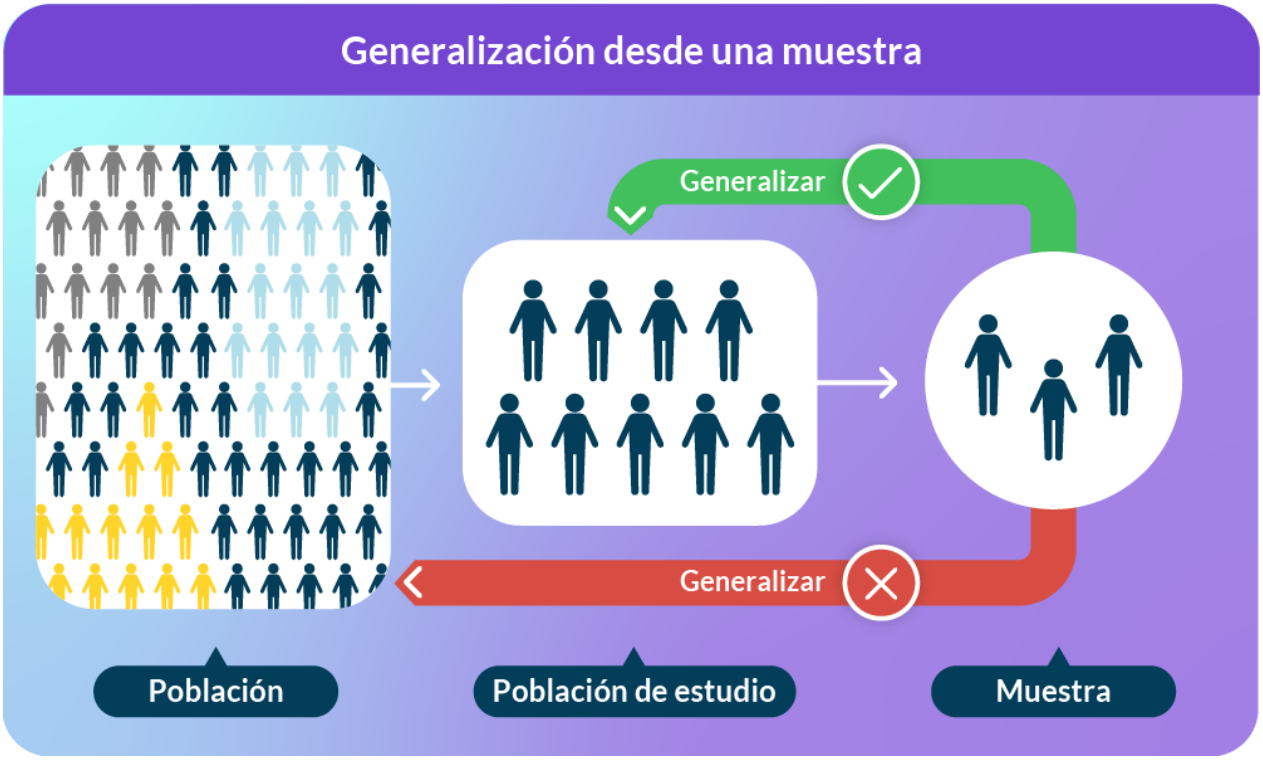
Si bien, en un ambiente ideal, lo mejor sería analizar toda la *población de estudio* para extrapolar los resultados a este mismo grupo, lo cierto es que no es recomendable hacerlo debido a los altos costos que esto podría ocasionar (en tiempo, dinero y energía). Por esta razón, es más práctico definir una *muestra*.

Por supuesto, tal definición obedece a un proceso riguroso, que depende de la fórmula de tamaño muestral y el tipo de muestreo que elijamos, cuyo propósito es obtener una muestra representativa.

Asimismo, los conceptos de *población* y *muestra* son importantes para delinear la muestra representativa, la cual se puede considerar como un conjunto o subconjunto de los datos que representará la *población de estudio,* de tal manera que valide las conclusiones que se puedan obtener.

Una muestra representativa implica utilizar y aplicar una fórmula de tamaño muestral y un tipo de muestreo, los cuales permitirán finalmente generalizar los resultados.

Figura 2. Generalización desde una muestra



Es importante señalar que sí es posible generalizar los resultados obtenidos de la muestra a la *población de estudio*. Sin embargo, generalizar estos resultados a toda la *población* es un error, tal como se ilustra en la imagen.

Recuerda que la muestra se obtiene a partir de unos criterios muy específicos que describen a la *población de estudio*.

2. Elección y tamaño de la muestra

Si consideramos el conjunto definido por la *población general*, los elementos como individuos, eventos u observaciones deben tener la misma probabilidad de ser escogidos para integrar una muestra, una característica estrechamente relacionada con el tipo de muestreo aleatorio simple.

Para entender mejor lo anterior, a lo largo de este tema, aprenderemos cómo obtener el tamaño de una muestra cuando las variables observadas son discretas o continuas.

Debemos recordar que el objetivo del muestreo es conocer las características de una población de grandes dimensiones (que eventualmente podría ser infinita, según el fenómeno abordado) a través de un grupo más manejable (o población finita). Además, la definición de *población* hace referencia a un conjunto de observaciones (individuos, eventos, objetos) que son susceptibles de ser medidos.

Asimismo, la *muestra* puede referirse a una colección de elementos no solapados, es decir, que no están repetidos en la población.

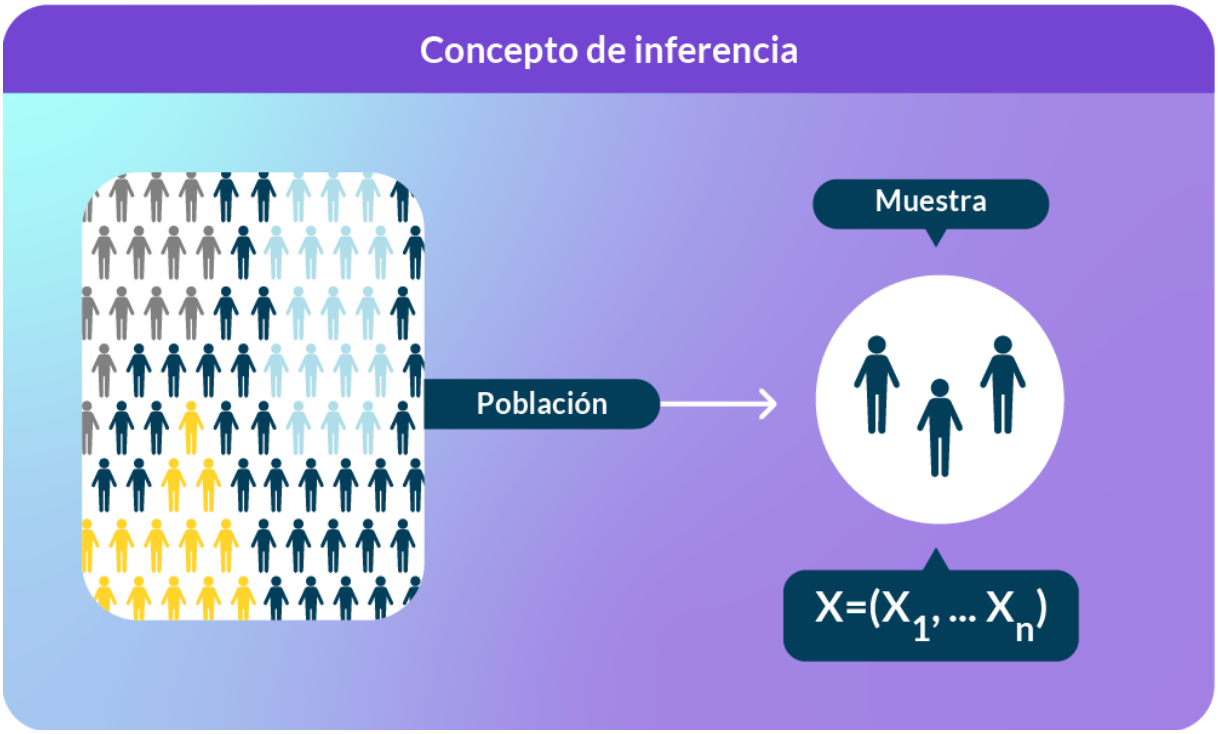
¿En qué condiciones se debe realizar un muestreo?

- Cuando el tamaño de la población sea tan grande que exceda las restricciones físicas, temporales y/o económicas del proceso/proyecto.
- Cuando la población sea uniforme con respecto a una característica que se espera medir.
- Cuando el proceso de medida sea destructivo como en los fenómenos de control de calidad, evaluación industrial, estudio de fármacos, entre otros.
- Por razones asociadas a la degeneración de la calidad.

En general, cuando se estimen menos errores al estudiar una muestra que en el estudio exhaustivo de la población.

Luego de calcular y seleccionar la muestra de la población de estudio, nos encontramos con el problema de extraer la información de los datos. A este proceso de extracción lo denominamos *inferencia*.

Figura 3. Concepto de inferencia



Un problema de inferencia estadística es aquel en el que, a partir de un conjunto de datos generados de acuerdo con una ley de probabilidad desconocida, se trata de averiguar cuál ha sido el mecanismo que los ha generado.

$$\alpha = \frac{\sigma^2}{\varepsilon^2} = \textbf{\textit{Nivel de significancia}} = \frac{\textbf{\textit{Variabilidad del error}}}{\textbf{\textit{Error}}}$$

La selección del tamaño de la muestra en una población permitirá controlar en todo momento el error cometido. En este modelo llamaremos:

- **ε (Épsilon)** al error cometido en la estimación.
- **α (Alfa)** al nivel de significancia.

¿Cómo se determina el tamaño de una muestra?

En cualquier proceso que pretenda conocer el comportamiento de una variable, observada en un fenómeno de estudio, es necesario establecer el tamaño muestral. Para ello, se deben considerar diferentes aspectos, no solo los estadísticos, antes de hacer el cálculo.

- ### Tamaño de la población

  Una definición común de población se refiere a una colección precisa de eventos, objetos, individuos u observaciones que poseen características similares.

  Otra representación bastante conocida es la población teórica, que se caracteriza por la multiplicidad de sus características.

  Por otro lado, la población de estudio es aquella a la que se tiene acceso y sobre la cual se hará la generalización de los resultados.

- ### Margen de error o Intervalo de confianza

  El margen de error es un estadístico que representa el valor del error de muestreo aleatorio en los resultados de un experimento.

  Dicho de otra manera, puede entenderse como la medida estadística del número de veces de cada 100, en los que se espera que los resultados se encuentren dentro de un rango específico.

  - **Nivel de confianza**
    Se refiere al conjunto de intervalos aleatorios utilizados para delimitar un valor con una determinada probabilidad alta.

    Por ejemplo, podemos contar con un intervalo de confianza de 95%. Esto significa que los resultados de un experimento podrán tener los valores esperados el 95% de las veces.

  - **Desviación estándar**
    Es un índice numérico que representa la dispersión (difusión) de un conjunto de datos (valores de las observaciones).

    Se deduce así que, entre más grande sea la desviación estándar, mayor será la dispersión de estos valores.

Cálculo de la muestra con tamaño de población desconocido

El siguiente es el modelo matemático que permite calcular el tamaño de la muestra cuando no se cuenta con el tamaño de la población:

$$n = \frac{Z_{\alpha}^{2} * p * q}{d^{2}}$$

En donde:

- Z_{α}: Nivel de confianza
- p: Probabilidad de éxito (proporción esperada)
- q: Probabilidad de fracaso
- d: Precisión (valor del error máximo admisible en términos de proporción)

El ***nivel de confianza*** es la probabilidad de que el parámetro a estimar se encuentre en el intervalo de confianza. El valor del nivel de confianza (p) se obtiene mediante el cálculo de 1 – α, y se suele expresar en forma de una razón o porcentaje.

$1 - \alpha$		$Z\alpha$
0.90	90%	1.645
0.95	95%	1.96
0.99	99%	2.575

El nivel de confianza o seguridad prefijado arroja un coeficiente Z_{α}, tal como se ilustra en la tabla.

Nota: estos valores provienen de las tablas de la distribución normal *Z*.

- Para una seguridad del 95%, su porcentaje es $Z_{\alpha} = 1.96$
- Para una seguridad del 99%, su porcentaje es $Z_{\alpha} = 2.57$

Se debe considerar que ***p*** y ***q*** son complementarios, y teniendo en cuenta que se expresan como una razón o porcentaje, su suma es igual a la unidad: $p + q = 1 = 100\%$.

Cuando se habla de máxima variabilidad, y no se cuenta con información preliminar ni antecedentes o publicaciones sobre el fenómeno, se asumen valores de variabilidad de $p = q = 0.5$.

Problema

¿Cuántas empresas tendríamos que estudiar para conocer la preferencia del mercado por una marca de equipos de cómputo para oficina, considerando que se desconoce la cantidad de empresas que los comprarán?

Para el cálculo del tamaño muestral se considerarán las siguientes restricciones:

- Seguridad = 95%
- Precisión = 3%;
- Proporción = 5% (probabilidad de éxito)

Solución

- $\boldsymbol{Z_\alpha}$: 1.96 (corresponde al 95%)
- $\boldsymbol{p}$: 0.05 (5%)
- $\boldsymbol{q}$: 0.95 (1-p = 1 - 0.05)
- $\boldsymbol{d}$: 0.03 (3%)

$$n = \frac{Z_\alpha^2 * p * q}{d^2} = \frac{1.96^2 * 0.05 * 0.95}{0.03^2} = 203$$

Es posible concluir, entonces, que se deben estudiar 203 empresas para tener una seguridad del 95% en cuanto a la preferencia por dicha marca de equipos.

Cálculo de la muestra con tamaño de población conocido

El siguiente es el modelo matemático que permite calcular el tamaño de muestra en los casos en los que sí se cuenta con el tamaño de la población:

$$n = \frac{N * Z_{\alpha}^{2} * p * q}{d^{2} * (N-1) + Z_{\alpha}^{2} * p * q}$$

En donde:

- N: Tamaño de la población
- Z_{α}: Nivel de confianza
- p: Probabilidad de éxito (proporción esperada)
- q: Probabilidad de fracaso
- d: Precisión (valor del error máximo admisible en términos de proporción)

Problema

¿Cuántas empresas tendríamos que estudiar para conocer la preferencia del mercado por una marca de equipos de cómputo para oficina, considerando que se desconoce la cantidad de empresas que los comprarán?

Para el cálculo del tamaño muestral se considerarán las siguientes restricciones:

- Seguridad = 95%
- Precisión = 3%;
- Proporción = 5% (probabilidad de éxito)

Solución

- Z_{α}: 1.96 (corresponde al 95%)
- p: 0.05 (5%)
- q: 0.95 (1-p = 1 - 0.05)
- d: 0.03 (3%)

$$n = \frac{Z_{\alpha}^{2} * p * q}{d^{2}} = \frac{1.96^{2} * 0.05 * 0.95}{0.03^{2}} = 203$$

¿Cuántas empresas tendríamos que estudiar para conocer la preferencia del mercado por una marca de equipos de cómputo para oficina, si la población está definida por 1200 empresas en proceso de compra?

Para el cálculo del tamaño muestral se considerarán las siguientes restricciones:

- Seguridad = 95%
- Precisión = 3%
- Proporción = 5% (probabilidad de éxito)

Solución

- $\boldsymbol{N}$: 1.200
- $\boldsymbol{Z_\alpha}$: 1.96 (corresponde al 95%)
- $\boldsymbol{p}$: 0.05 (5%)
- $\boldsymbol{q}$: 0.95 (1-p = 1 - 0.05)
- $\boldsymbol{d}$: 0.03 (3%)

$$\boldsymbol{n = \frac{N * Z_\alpha^2 * p * q}{d^2 * (N-1) + Z_\alpha^2 * p * q}}$$

$$\boldsymbol{= \frac{1200 * 1.96^2 * 0.05 * 0.95}{0.03^2 * (1200-1) + 1.96^2 * 0.05 * 0.95}}$$

$$\boldsymbol{= 173.56 \cong 174}$$

Es posible concluir, entonces, que se deben estudiar 174 empresas como mínimo, para tener una seguridad del 95% en cuanto a la preferencia por la marca de equipos en proceso de adquisición.

Tipos de muestreo

Ya hemos aclarado que la importancia del muestreo radica fundamentalmente en su utilidad para definir qué parte de una población debe ser analizada cuando no es posible hacerlo en su totalidad.

Ese ejercicio puntual de selección se realiza mediante muestreo probabilístico o no probabilístico, y depende de la naturaleza y objetivos del experimento.

- Muestreo probabilístico

En este tipo de muestreo es imperativo el cumplimiento de dos condiciones:

- Que todas las observaciones o elementos de la población sean susceptibles de ser elegidos, es decir, tienen una probabilidad superior a cero de ser parte de la muestra.
- Que el valor de probabilidad de inclusión de cada uno de los individuos u observaciones en la muestra se conoce de forma precisa.

Cumplir con ambos criterios garantiza que los resultados obtenidos no estén sesgados cuando se estudie la muestra, y ayuda a determinar el grado de incertidumbre propio del proceso de muestreo.

Eventualmente, los resultados no sesgados necesitan el uso de técnicas de ponderación *(weighting)*, que utilizan el valor de probabilidad de que cada individuo sea seleccionado en la muestra. En los experimentos, a las muestras que se generan de esta forma se les denomina *probabilísticas*.

En general, lo que se espera de un muestreo de este tipo es que todos los individuos de la muestra seleccionada tengan las mismas probabilidades de ser elegidos para garantizar que esta sea representativa.

Entre las diferentes técnicas de muestreo probabilístico, se pueden mencionar:

- Muestreo aleatorio simple
- Muestreo sistemático
- Muestreo estratificado
- Muestreo por conglomerados

Características

- No hay afectación (por la subjetividad) de quien aplica el experimento.
- La selección de las observaciones se hace a través de reglas mecánicas.
- Se considera el error muestral.
- Se conoce la probabilidad de inclusión.

- ## Muestreo no probabilístico

Esta técnica de muestreo es adecuada para el desarrollo de estudios exploratorios, aunque no son de gran ayuda para hacer generalizaciones.

La selección de observaciones o individuos se realiza a partir de diferentes criterios relacionados con las características del experimento. Es relevante considerar que no todas las observaciones tienen la misma probabilidad de ser seleccionadas, pues quien aplica el experimento suele determinar de manera subjetiva la población objetivo.

Las técnicas de muestreo no probabilístico más comunes son:

- Muestreo por conveniencia
- Muestreo secuencial
- Muestreo por cuotas
- Muestreo discrecional
- Muestreo por bola de nieve

Características:

- La muestra es a discreción de quien ejecuta el experimento.
- Las observaciones se seleccionan por conveniencia y no por criterios equiprobables.
- No hay error muestral o no se puede calcular.
- No se conoce la posibilidad de inclusión.

3. Cálculo de medidas

Las medidas estadísticas, también conocidas como parámetros estadísticos, son valores representativos de una colección de datos que resumen, en unos pocos valores, la información completa. Estas medidas permiten obtener información sobre la localización, dispersión y otros patrones de comportamiento de un conjunto de datos, para tener una idea de su estructura y relación.

Las medidas estadísticas más importantes son:

- **Tendencia central:** indican el valor medio de los datos.
- **Dispersión:** miden la variabilidad de los datos respecto a los parámetros de centralización.
- **Forma:** relacionada con su simetría y apuntamiento, las cuales indican la forma en que se distribuyen los datos.

Medidas de tendencia central

Las medidas de tendencia central están compuestas por los valores localizados en el centro del conjunto de datos de interés, los cuales se encuentran ordenados según su magnitud. Aunque existen varias medidas de tendencia central, las más utilizadas son: media, mediana y moda.

- Media

  También conocida como media aritmética o promedio simple. Se identifica con la letra griega $\boldsymbol{\mu}$ (mu) cuando se trata de una población, o con $\bar{x}$ para una muestra. La media se puede calcular sumando todos los elementos del conjunto de datos para luego dividirlos entre el número de elementos.

$$\boldsymbol{\mu} \text{ ó } \bar{\boldsymbol{x}} = \frac{\sum_{i=1}^{N} \boldsymbol{X_i}}{\boldsymbol{N}} = \frac{\boldsymbol{x_1 + x_2 + x_3 + \cdots + x_{N-1} + x_N}}{\boldsymbol{N}}$$

  Aunque la media es la más común de las medidas de tendencia central, los valores atípicos pueden afectarla fácilmente, lo cual supone una gran desventaja.

- Mediana (*Me*)

  La mediana es el valor localizado en la posición central del conjunto de datos que se están analizando. Para encontrar adecuadamente esta ubicación, los valores deben estar ordenados de manera que el cálculo permita encontrar fácilmente la mitad de las observaciones.

  Así es posible percatarse que la mitad de los valores es menor que el valor de la mediana, mientras que la otra mitad es mayor.

  Por esta razón, y para atenuar un poco los efectos de la perturbación que producen valores extremos, esta medida se utiliza en lugar de la media.

- Moda (*Mo*)

  Este valor representa al dato o datos que más veces se repite dentro de un conjunto. Cuando es un solo dato es que se repite, se denomina unimodal. Si se encuentran dos valores que se repiten el mismo número de veces se conoce como bimodal; y más de dos veces, se nombra multimodal.

  A diferencia del cálculo de la media, la moda no se afecta por la ocurrencia de valores extremos en el conjunto de datos.

Problema

La siguiente matriz de datos contiene los reportes de incidentes en el uso de una aplicación por día. Para un proceso de control de calidad, se ha propuesto analizar los estadísticos de tendencia central con el fin de tener una primera perspectiva del problema.

22	22	24	22	24
25	19	19	22	22
17	25	22	17	23
24	21	17	20	20
19	23	25	19	17
24	25	24	25	24
20	21	24	25	22
24	23	25	20	17
18	19	25	17	23
20	20	24	18	25

Se espera una solución en ***R***, usando la infraestructura de ***R Studio Cloud***.

Solución

Para resolver el problema en ***R***, lo primero que debemos hacer es importar la librería que permite leer archivos de Excel.

```
library(readxl)
```

Nota: si estás usando una versión de escritorio de ***R***, es necesario instalar esta librería con la instrucción: `install.packages("readxl")`

Podemos crear un archivo de Excel que denominaremos ***"data.xlsx"***. Luego, organizamos los datos en una hoja de cálculo en una sola columna, cuyo encabezado titularemos como ***"incidentes"***. Este archivo lo importaremos a ***R Studio Cloud***.

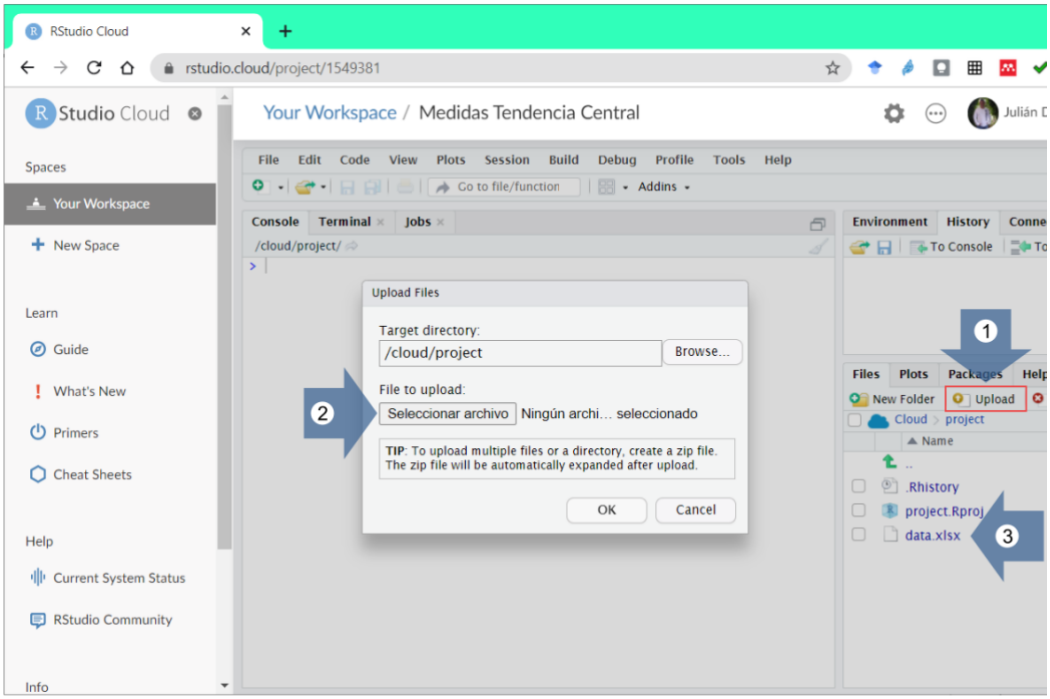


Figura 4. Interfaz de proyecto de *R Studio Cloud*

En la interfaz de *R Studio Cloud*:

1. Haz clic en ***Upload*** para cargar el archivo.
2. Cuando aparezca la ventana emergente de ***Upload Files***, haz clic en ***Seleccionar archivo***.
3. Después de haber seleccionado el archivo del computador, deberá aparecer en la lista de ***Archivos del proyecto*** en *R Studio*.

Ya con el archivo cargado, almacena en un *dataset,* que llamaremos ***dataReporte*,** el resultado de aplicar la función **read_excel.** A esta se envían los parámetros del nombre del archivo y el tipo de datos de las columnas (en este caso, solo hay una columna con datos de tipo numérico).

```
dataReporte <- read_excel("data.xlsx", col_types = c("numeric"))
```

Para verificar si el *dataset* quedó bien creado, puedes visualizar su contenido:

```
View(dataReporte)
```

Utilizamos la función ***mean*** para calcular la media de ese conjunto de datos, a la que debe enviarse el nombre del *dataset* (***dataReporte***) y el nombre del campo o columna del *dataset* (***$incidentes***).

```
mean(dataReporte$incidentes)
```

Si es necesario almacenar este resultado en una variable para utilizar más adelante en algún cálculo, se puede anotar la siguiente instrucción:

```
mediaIncidentes = mean(dataReporte$incidentes)
```

Y para imprimirlo, la instrucción sería:

```
print(mediaIncidentes)
```

La función ***median*** es útil para calcular la mediana, con un manejo de parámetros similar al de ***mean***.

```
medianaIncidentes = median(dataReporte$incidentes)
print (medianaIncidentes)
```

Para calcular la moda, debemos utilizar una librería externa, ya que ***R*** no cuenta con un método propio que permita hacerlo. Dicha librería se llama ***"modeest",*** y de ella se debe usar la función ***"mlv"*** para retornar un vector a valores numéricos.
Lo primero que debemos hacer es instalar el paquete:

```
install.packages("modeest")
```

Luego, activamos el paquete en ***R***:

```
library(modeest)
```

Finalmente, ejecutamos la función ***mlv*:**

```
mlv(dataReporte$incidentes, method = "mfv")
```

Podemos copiar y pegar las instrucciones como aparecen a continuación, las cuales se ejecutarían una tras otra:

```
install.packages("modeest")
library(modeest)
mlv(dataReporte$incidentes, method = "mfv")
```

Y el resultado debería ser similar al que se aprecia en la imagen:

Figura 5. Resultado de la instalación del paquete "***modeest***" en *R Studio Cloud*

Allí podemos observar que la primera línea muestra en rojo el proceso de instalación. La tercera línea devuelve el vector de resultados que, en este caso, contiene los dos valores de la moda, ya que el conjunto de datos es bimodal.

De igual manera, podemos aprovechar una herramienta de ***R*** que permite obtener un resumen de los estadísticos básicos de un conjunto de datos. Los parámetros son los mismos que ya hemos utilizado:

```
summary(dataReporte$incidentes)
```

Esta instrucción devuelve un arreglo con los siguientes resultados:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17.00   19.25   22.00   21.64   24.00   25.00
```

Con la instrucción ***summary*** en ***R,*** podemos obtener: el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo. Pero recuerda que este resumen no muestra el valor de la moda.

Medidas de dispersión

Estas medidas permiten analizar qué tan dispersos se encuentran los datos de la media. Entre más cerca de la media se encuentre agrupado el conjunto de datos, los valores de las medidas de dispersión son más bajos. Y en sentido contrario, mientras más alejados de la media se encuentren los valores del conjunto de datos, los valores de dispersión aumentan.

Las medidas de dispersión más utilizadas son:

- Varianza y cuasivarianza
- Desviación típica y cuasidesviación típica
- Desviación absoluta

Varianza y cuasivarianza

La varianza mide la dispersión de un conjunto de datos a partir de la comparación de cada uno de ellos con la media del conjunto. La varianza utiliza en su cálculo el promedio de las desviaciones al cuadrado.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

A menudo, debemos utilizar la varianza de una muestra como el valor estimado de la varianza de la población. Cuando esto suceda, podemos considerar un valor de error menor si se utiliza una *cuasivarianza* s^2 en lugar de emplear como estimador la varianza de la muestra.

El cálculo de la cuasivarianza es similar al de la varianza, solo que en el denominador se cambia n por $n - 1$.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1}$$

En los casos que se cuente con datos de toda la población, es recomendable utilizar n en lugar de $n - 1$.

A pesar de lo anterior, lo indicado es utilizar la cuasivarianza si consideramos que generalmente nuestros datos son una muestra obtenida de una población que, por supuesto, es de mayor tamaño.

En algunas bibliografías y documentos técnicos se encuentran referencias a la varianza que en realidad hablan del resultado del cálculo de la cuasivarianza. Programas estadísticos y algunos paquetes de estadística programados para ***R***, por ejemplo, solo incorporan funciones para el cálculo de cuasivarianza y no para la varianza.

- ## Desviación típica y cuasidesviación típica

  Evitar valores negativos es la razón por la cual en la varianza las diferencias se elevan al cuadrado. Ya que derivaría en una modificación de las unidades de medida de lo valores del conjunto de datos.

  Entonces, con el fin de contar con las mismas unidades para el valor de la medida de dispersión y la media, se utiliza la *desviación típica* (σ). En nuestro contexto, es más conocida como *desviación estándar*, la cual se calcula como la raíz cuadrada de la varianza.

$$\boldsymbol{\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}}}$$

  Al igual que la varianza, solo cuando el tamaño de la muestra es idéntico al tamaño de la población, se recomienda usar el valor de la *cuasidesviación típica* (s) en lugar de la desviación típica.

  Es decir, el cálculo de la *cuasidesviación típica* se obtiene al dividir el número de grados de libertad entre ($n - 1$), en lugar de hacerlo entre el total de datos (n).

$$\boldsymbol{s = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}}}$$

  Igualmente, es posible encontrar referencias a la *desviación típica* que en realidad hablan del resultado del cálculo de la *cuasidesviación típica*.

  Asimismo, en algunos programas estadísticos y paquetes de estadística programados para *R*, por ejemplo, solo se incorporan funciones para el cálculo *de cuasidesviación* típica y no para la *desviación estándar*.

- Desviación absoluta

  Esta medida de dispersión corresponde al promedio de los valores absolutos de las desviaciones. Mide las distancias a las que se encuentra cada valor alejado de la media y supone una mejor interpretación del grado de dispersión de los datos.

$$DAM = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{\bar{x}}|$$

Problema

El siguiente conjunto de datos corresponde a los reportes de incidentes en el uso de una aplicación por día. Para comprender el comportamiento de los datos, se debe contar con las medidas de dispersión más comunes.

22	22	24	22	24
25	19	19	22	22
17	25	22	17	23
24	21	17	20	20
19	23	25	19	17
24	25	24	25	24
20	21	24	25	22
24	23	25	20	17
18	19	25	17	23
20	20	24	18	25

Se espera una solución en ***R***, usando la infraestructura de ***R Studio Cloud***.

Solución

Ya que el conjunto de datos es el mismo del anterior problema, podemos reutilizar el script que permite la importación de los datos y la carga en un dataframe de R.

Recuerda que es importante llevar los datos de la matriz a un archivo de Excel y disponerlos en una sola columna.

En caso de que no hayas realizado todavía el ejercicio anterior, te recomendamos que veas por lo menos la primera parte de su solución, ya que en ella se muestra el proceso de carga del archivo en ***R Studio***.

A continuación, aparecen las dos líneas con las cuales iniciar. Recuerda que la primera de ellas carga la librería que permite trabajar con archivos de Excel, mientras que la segunda carga los datos del archivo en un *dataframe*.

```
library(readxl)
dataReporte <- read_excel("data.xlsx", col_types = c("numeric"))
```

Para usar los datos de la columna en el resto del código, es recomendable almacenarlos en un arreglo, que en el código se llamará *incidentes*.

```
incidentes = dataReporte$incidentes
```

En seguida, calcularemos el número de elementos del *dataset* de incidentes, de la *cuasivarianza* y de la *varianza* (recuerda que en se describe la diferencia entre ellas). Después, almacenamos los resultados en variables cuyos nombres indican de qué se trata.

```
n = length(incidentes)
cuasiVar = var(incidentes, na.rm = FALSE)
varianza = cuasiVar * ((n-1)/n)
```

Asimismo, haremos el cálculo de la *cuasidesviación estándar y de la desviación estándar* (el texto anterior también se describe la diferencia entre ellas). Además, con ayuda de la *varianza* y *cuasivarianza*, deduciremos la forma de almacenar los resultados en variables cuyos nombres indican de qué se trata.

```
cuasiDesv = sd(incidentes, na.rm = FALSE)
desvEstandar = cuasiDesv * ((n-1)/n)
```

Finalmente, hacemos el cálculo de la desviación absoluta de la media.

En la primera línea del siguiente código, se debe calcular la media de los valores del conjunto de datos de incidentes.

En la segunda línea, se obtiene la media de los valores absolutos de la diferencia entre cada valor y la media del conjunto de datos. Este resultado se almacena en la variable ***mediaAbsDif***. En la tercera línea se divide este valor entre el tamaño del conjunto de datos, dicho resultado corresponde a la desviación absoluta de la media, que llamaremos ***desAbsMedia***.

```
    mediaIncidentes = mean(incidentes)
    mediaAbsDif = mean(abs(incidentes - mediaIncidentes), na.rm =
TRUE)
    desAbsMedia = mediaAbsDif/n
```

Por último, para mostrar todos los resultados, es posible combinar un mensaje de texto con el valor de una variable, vamos a utilizar la función ***paste*** línea a línea.

```
    paste("Tamaño del conjunto de datos", n)
    paste("Valor de la CuasiVarianza", cuasiVar)
    paste("Valor de la Varianza", varianza)
    paste("Valor de la Cuasi-Desviación Estándar", cuasiDesv)
    paste("Valor de la Desviación Estándar", desvEstandar)
    paste("Valor de la Desviación Absoluta de la Media", desAbsMedia)
```

```
> paste("Tamaño del conjunto de datos", n)
[1] "Tamaño del conjunto de datos 50"
> paste("Valor de la CuasiVarianza", cuasiVar)
[1] "Valor de la CuasiVarianza 7.50040816326531"
> paste("Valor de la Varianza", varianza)
[1] "Valor de la Varianza 7.3504"
> paste("Valor de la Cuasi-Desviación Estándar", cuasiDesv)
[1] "Valor de la Cuasi-Desviación Estándar 2.73868730658783"
> paste("Valor de la Desviación Estándar", desvEstandar)
[1] "Valor de la Desviación Estándar 2.68391356045607"
> paste("Valor de la Desviación Absoluta de la Media", desAbsMedia)
[1] "Valor de la Desviación Absoluta de la Media 0.047552"
```

Figura 6. Visualización de resultados con ***paste*** en ***R***

Aunque también podemos intentar un bloque completo de impresión de resultados usando la función **c()**.

```
print(c(
"Tamaño del conjunto de datos:",n,
"CuasiVarianza:",cuasiVar,
"Varianza:",varianza,
"Cuasi-Desviación Estándar:",cuasiDesv,
"Desviación Estándar:",desvEstandar,
"Desviación Absoluta de la Media:",desAbsMedia
))
```

```
> print(c(
+     "Tamaño del conjunto de datos:",n,
+     "CuasiVarianza:",cuasiVar,
+     "Varianza:",varianza,
+     "Cuasi-Desviación Estándar:",cuasiDesv,
+     "Desviación Estándar:",desvEstandar,
+     "Desviación Absoluta de la Media:",desAbsMedia
+ ))
 [1] "Tamaño del conjunto de datos:"
 [2] "50"
 [3] "CuasiVarianza:"
 [4] "7.50040816326531"
 [5] "Varianza:"
 [6] "7.3504"
 [7] "Cuasi-Desviación Estándar:"
 [8] "2.73868730658783"
 [9] "Desviación Estándar:"
[10] "2.68391356045607"
[11] "Desviación Absoluta de la Media:"
[12] "0.047552"
```

Figura 7. Visualización de resultados con ***print*** en *R*

4. Gráficos estadísticos

La visualización juega un papel fundamental en la gestión de los datos y la modelación de los mismos como herramientas que permiten apoyar la toma de decisiones a partir del comportamiento de los mismos.

En este sentido, los gráficos son una herramienta adecuada para presentar de una forma sintáctica comportamientos, relaciones o el estado de los datos, ya que permiten a cualquier persona entender fácilmente estas propiedades.

Aunque hay una gran cantidad de tipos de gráficos, mencionaremos solo aquellos de uso frecuente. Por otro lado, te recomendamos explorar las opciones de visualización de datos que tiene *R*, y las librerías que amplían esas posibilidades.

Para las formas de gráfico estadístico que abordaremos, se utilizará el siguiente conjunto de datos. Te recomendamos crear un archivo .xls con estos datos para validar la aplicación de los conceptos y la utilización de los comandos de R Studio.

Semana	Tipo incidente	Cantidad incidentes	Modificación código
Semana 1	Acceso	22	3
Semana 1	Registro	17	1
Semana 1	Descarga	20	2
Semana 2	Acceso	23	5
Semana 2	Registro	17	0
Semana 2	Descarga	20	4
Semana 3	Acceso	18	1
Semana 3	Registro	17	1
Semana 3	Descarga	21	4
Semana 4	Acceso	21	4
Semana 4	Registro	18	2
Semana 4	Descarga	19	2
Semana 5	Acceso	23	3
Semana 5	Registro	20	4
Semana 5	Descarga	20	3
Semana 6	Acceso	22	5

Semana 6	Registro	22	4
Semana 6	Descarga	19	3
Semana 7	Acceso	20	4
Semana 7	Registro	18	1
Semana 7	Descarga	24	4
Semana 8	Acceso	19	2
Semana 8	Registro	21	3
Semana 8	Descarga	21	3
Semana 9	Acceso	20	4
Semana 9	Registro	20	3
Semana 9	Descarga	21	3
Semana 10	Acceso	18	0
Semana 10	Registro	19	2
Semana 10	Descarga	25	4
Semana 11	Acceso	21	4
Semana 11	Registro	23	5
Semana 11	Descarga	19	3
Semana 12	Acceso	18	2
Semana 12	Registro	22	4
Semana 12	Descarga	17	0
Semana 13	Acceso	20	4
Semana 13	Registro	17	0
Semana 13	Descarga	19	3
Semana 14	Acceso	25	5
Semana 14	Registro	17	2
Semana 14	Descarga	24	4
Semana 15	Acceso	25	5
Semana 15	Registro	22	3
Semana 15	Descarga	17	0
Semana 16	Acceso	23	3
Semana 16	Registro	24	3
Semana 16	Descarga	20	4
Semana 17	Acceso	17	1
Semana 17	Registro	21	5
Semana 17	Descarga	25	5
Semana 18	Acceso	18	1
Semana 18	Registro	17	2
Semana 18	Descarga	17	0

Semana 19	Acceso	25	4
Semana 19	Registro	22	5
Semana 19	Descarga	22	3
Semana 20	Acceso	20	2
Semana 20	Registro	22	3
Semana 20	Descarga	23	3

No te olvides de usar las instrucciones para importar el archivo .xls y la hoja en la que se encuentran los datos.

Esta matriz de datos se ha almacenado en el mismo archivo data.xls, en una segunda hoja denominada tipoIncidente, con el fin de motivar un desarrollo adecuado de la instrucción.

Recuerda que, primero, debemos activar la librería para importar archivos de Excel.

```
library(readxl)
```

Luego, cargamos a un dataframe, que llamaremos dataRegistro, la tabla de la hoja (sheet) tipoIncidente. En seguida, definimos los tipos de datos (col_types) de cada columna.

```
dataRegistro <- read_excel("data.xlsx", sheet = "tipoIncidente", col_types = c("text", "text", "numeric","numeric"))
```

Después podemos verificar el contenido del dataframe, visualizándolo.

```
View(dataRegistro)
```

Histogramas

Esta representación gráfica permite visualizar el comportamiento de una variable en forma de barras. Debido a que el histograma se utiliza para representar datos agrupados, puede emplearse en variables cuantitativas discretas y continuas, ya que en la parte central del gráfico se ubica la marca de clase.

Problema

A partir de los datos almacenados en el *dataframe* ***dataRegistro***, debemos construir un histograma.

Solución

Para facilitar la manipulación y el acceso a los datos, almacenaremos por separado, en un nuevo *dataframe*, los datos de la columna ***Cantidad Incidentes***.

```
cantIncidente = dataRegistro$"Cantidad Incidentes"
```

Luego, usando la función ***hist***, enviamos como argumentos el arreglo que acabamos de crear.

Indicamos que el valor de probabilidad **NO** se representará en el eje ***y***, sino el de la frecuencia.

En la etiqueta del eje ***y (ylab = "Frecuencia")*** usaremos el texto "*Frecuencia",* además, el color de las columnas será gris claro ***(col = "lightgrey")***.

Adicionalmente, los ejes serán visibles y se establecerá el texto del título principal ***(main)***.

Al final, en una nueva línea, indicaremos cuántos ejes serán visibles ***(axis(2))***.

El valor de 2 se utiliza para que aparezcan los ejes horizontal y vertical.

```
hist(cantIncidente,
        probability = FALSE,
        ylab = "Frecuencia",
        col = "lightgrey",
        axes = TRUE,
        main = "Histograma de Incidentes")
axis(2)
```

El resultado de la visualización que produce *R* se muestra en la siguiente imagen.

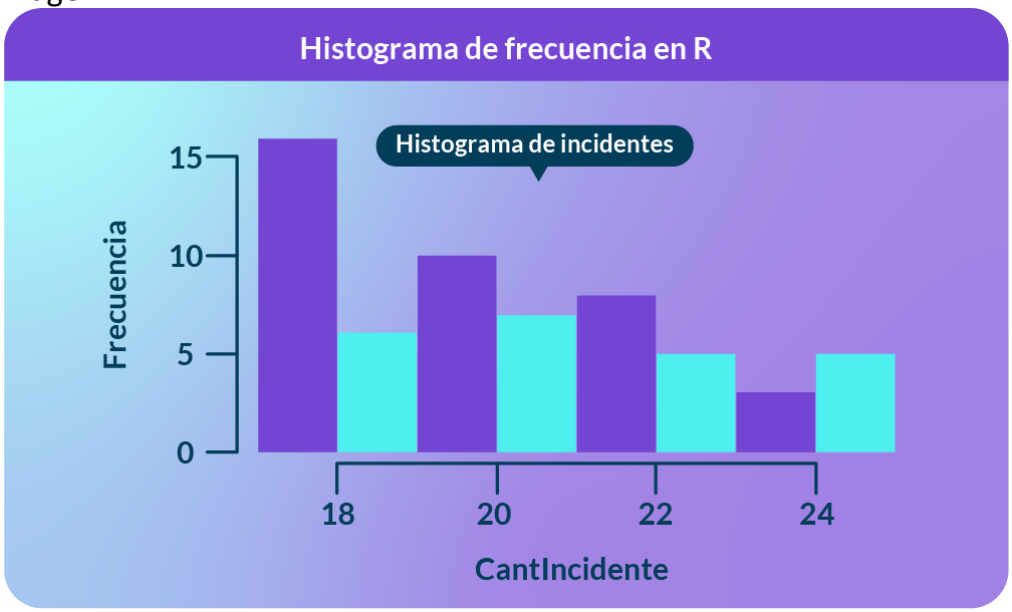


Figura 8. Histograma de frecuencia en *R*

Ahora, si debemos graficar la probabilidad en el histograma, el código cambia ligeramente.

En el argumento de la probabilidad indicamos que es verdadero ***(probability = TRUE)***, y, en consecuencia, cambiamos el texto de la etiqueta del eje ***y***.

```
hist(cantIncidente,
        probability = TRUE,
        ylab = "Probabilidad",
        col = "lightgrey",
        axes = TRUE,
```

```
        main = "Histograma de Incidentes")
axis(2)
```

Con estos argumentos, la salida será:

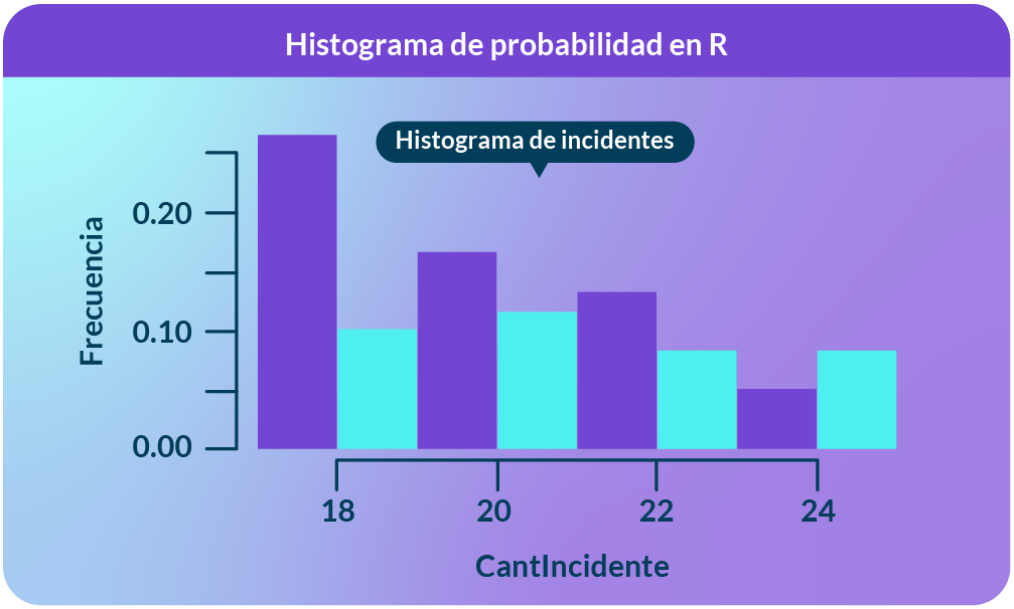


Figura 9. Histograma de probabilidad en *R*

Gráficos de cajas y bigotes

Los gráficos de cajas y bigotes son representaciones de las distribuciones de probabilidad. Estos poseen varias características, ya que las variables pueden ser continuas o discretas, por lo que algunos resultados derivan en gráficas simétricas o asimétricas, dependiendo todo del tipo de distribución.

Se utilizan para mostrar los datos numéricos a través de cuartiles. En esta representación, a las líneas que se extienden más allá de las cajas, se les conoce como *bigotes*, e indican la variabilidad fuera de los cuartiles superior e inferior.

Los elementos que componen un gráfico de bigotes se muestran en la siguiente ilustración.

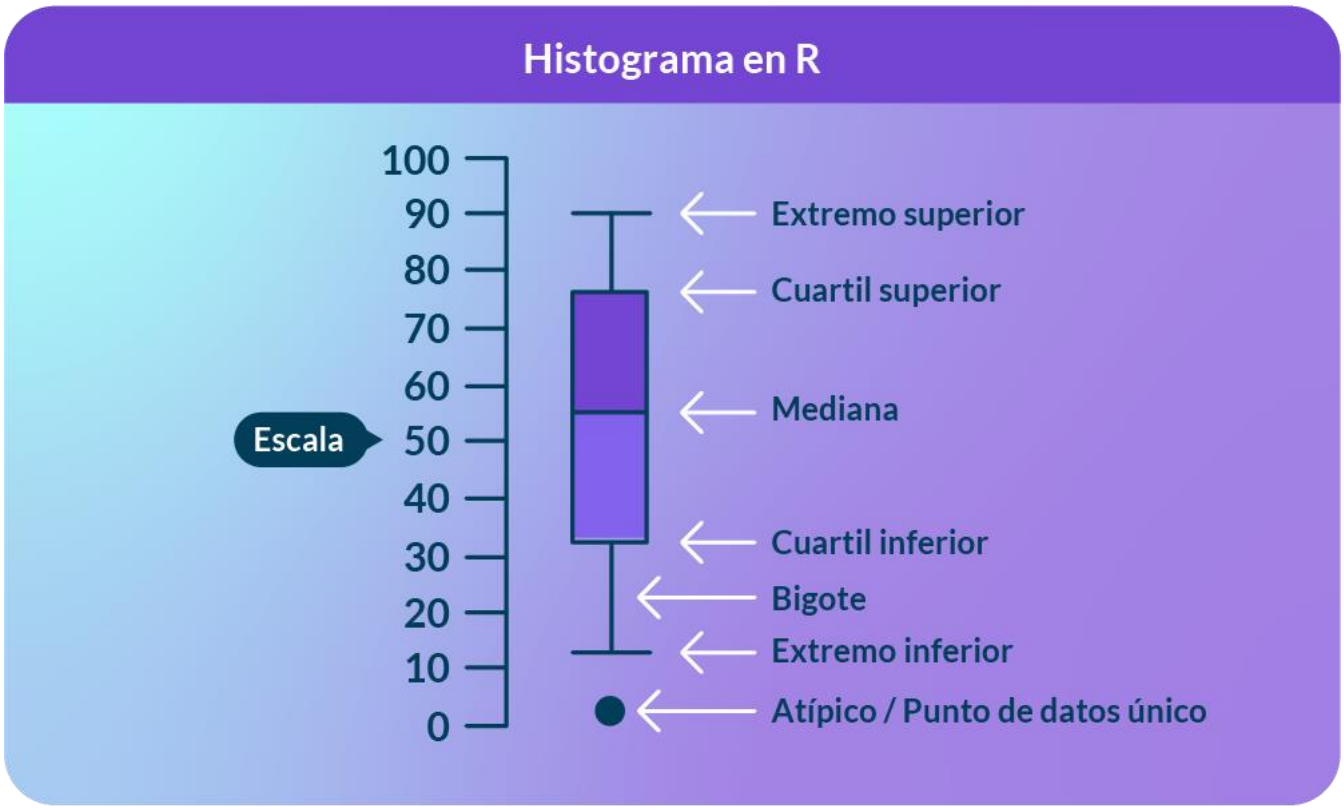


Figura 10. Histograma en *R*
Fuente: tomado de https://datavizcatalogue.com/ES/metodos/images/anatomy/SVG/diagrama_cajas_y_bigotes.svg.

Problema

A partir de los datos almacenados en el *dataframe* ***dataRegistro***, debemos construir una gráfica de cajas y bigotes.

Solución

La función para construir un gráfico de cajas es ***boxplot()*** y requiere de dos arreglos para funcionar. En este caso, una variable categórica almacena los tipos de incidentes, para facilitar su manejo y permitir con más propiedad su posición en los argumentos de la función.

Almacenaremos ambas por separado en *dataframes* independientes.

```
tipoIncidente = dataRegistro$"Tipo Incidente"
cantIncidente = dataRegistro$"Cantidad Incidentes"
```

A continuación, invocamos la función ***boxplot,*** enviando primero el *dataframe* que contiene los valores numéricos y luego los valores categóricos, separados por el signo **~**.

Adicionalmente, como ya lo hicimos anteriormente, se establecen los valores para los nombres de los ejes.

```
boxplot(cantIncidente ~ tipoIncidente,
     xlab = "Tipo de Incidente",
     ylab = "Frecuencia Cantidad de Incidentes")
```

La gráfica resultante es la siguiente:

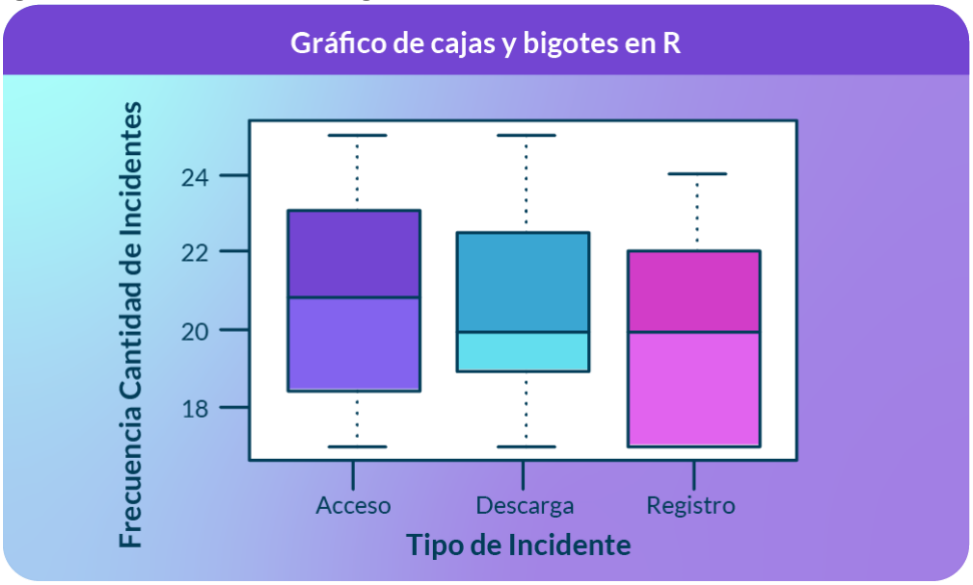


Figura 11. Gráfico de cajas y bigotes en ***R***

Si se prefiere una representación horizontal, basta con incluir en el parámetro ***horizontal*** el valor ***TRUE***. Vale la pena señalar que también es posible agregar color al gráfico.

```
boxplot(cantIncidente ~ tipoIncidente,
        xlab = "Tipo de Incidente",
        ylab = "Frecuencia Cantidad de Incidentes",
        horizontal = TRUE,
        col= 2:5)
```

Para un resultado como el siguiente:

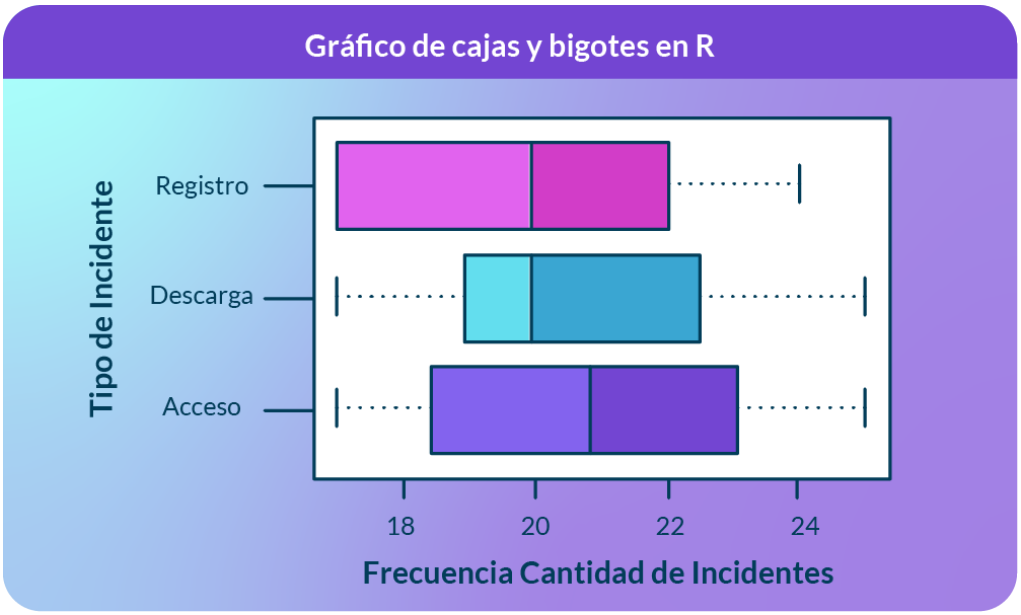


Figura 12. Gráfico de cajas y bigotes en *R*

En la gráfica son fácilmente identificables los elementos que corresponde a las posiciones de la mediana de los datos, y los valores mínimo y máximo entre otros. En este conjunto de datos se puede observar además que no existen valores atípicos.

Gráficos de dispersión

Estos gráficos generalmente se representan como una lluvia de puntos donde se relacionan dos o más variables. A esta relación se le denomina *correlación* y, de acuerdo con la caracterización de la lluvia, se dice que es:

- ***Fuerte***: si los puntos se proyectan hacia una misma dirección.
- ***Débil***: si algunos puntos se desvían de la dirección predominante.
- ***Nula***: si los puntos están dispersos y no se observa una relación.

Además, la correlación puede ser positiva si la dirección presenta un crecimiento; o negativa si la dirección tiene pendiente decreciente.

Problema

A partir de los datos almacenados en el *dataframe* **dataRegistro**, representemos gráficamente la correlación entre las modificaciones que se hacen al código y los incidentes reportados.

Solución

Al igual que lo hemos hecho en los problemas anteriores, se almacenarán por separado los valores de cada una de las columnas que se requieren, en *dataframes* diferentes.

```
modCodigo = dataRegistro$"Modificacion Codigo"
cantIncidente = dataRegistro$"Cantidad Incidentes"
```

A continuación, utilizamos la función ***plot()*** que contiene los parámetros ***x*** y ***y***, los cuales corresponden a las variables que debemos evaluar para determinar su nivel de correlación.

Como en otros gráficos, también se establecen los valores para los ejes y el título del gráfico.

```
plot(x = modCodigo, y = cantIncidente,
  xlab = "Num. de modificaciones al código",
  ylab = "Num. de Incidentes",
  main = "Correlación entre Modificaciones al código e Incidentes")
```

Por defecto, la salida para este gráfico debería verse así:

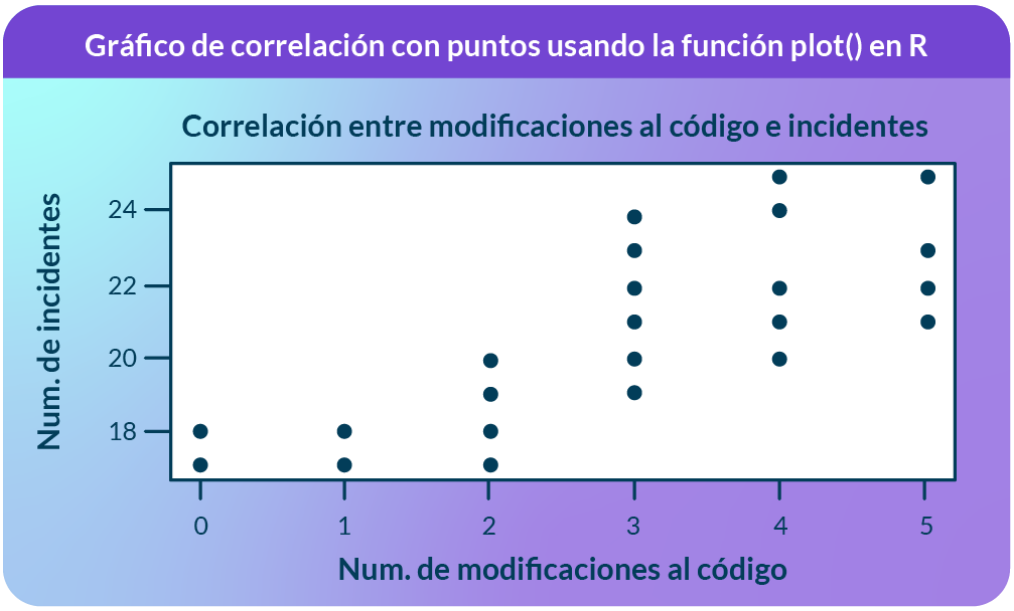


Figura 13. Gráfico de correlación con puntos usando la función *plot()* en *R*

En la gráfica se puede apreciar que, en efecto, existe una correlación entre los incidentes y las modificaciones que se hacen al código. De esta manera, es más fácil notar el comportamiento que se acentúa entre los números 3 y 4 de las modificaciones, además del aumento en los incidentes.

Si se quiere presentar la dispersión de puntos unidos con líneas, basta con agregar al parámetro ***type***, el argumento ***"l"***.

```
plot(x = modCodigo, y = cantIncidente,
  xlab = "Num. de modificaciones al código",
  ylab = "Num. de Incidentes",
  main = "Correlación entre Modificaciones al código e Incidentes",
  type = "l")
```

Lo anterior debe arrojar la siguiente gráfica.

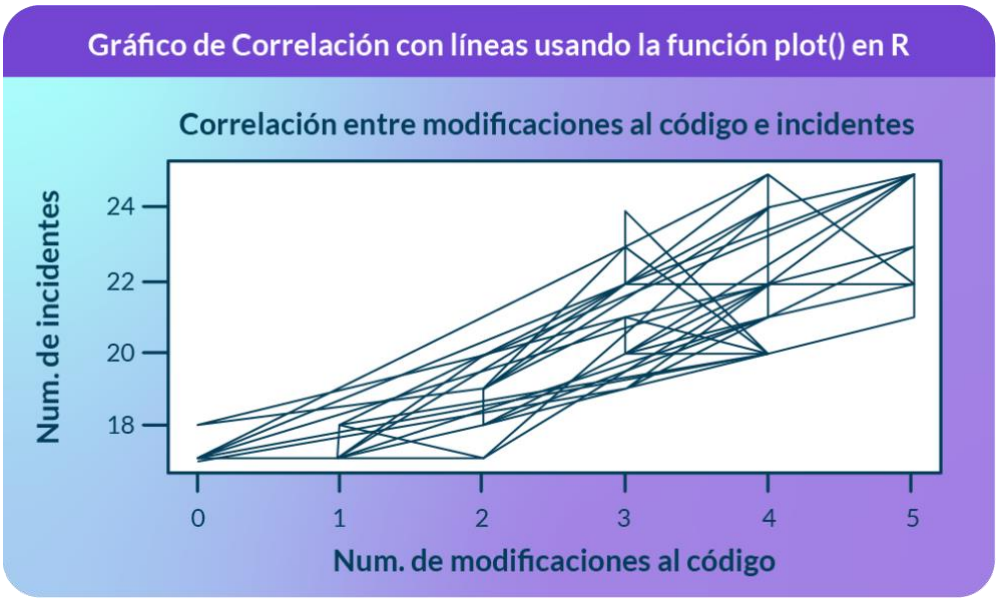


Figura 14. Gráfico de Correlación con líneas usando la función *plot()* en R

De esta manera, es más fácil notar cómo se acentúa la dirección de la correlación.

¡Felicitaciones!

Has finalizado con éxito la actividad de aprendizaje de la Unidad 2.

Los conocimientos que has adquirido te serán de gran utilidad al momento de gestionar e interpretar datos para obtener información valiosa y provechosa en tu trabajo, tus proyectos personales y en tu vida cotidiana.

Te invitamos a desarrollar la primera evidencia de aprendizaje del curso para poner en práctica cuanto has aprendido. Lee con atención el problema a resolver, identifica su contexto, requisitos y restricciones, y define cuál es la mejor forma de solucionarlo.

Evidencia de aprendizaje (EA) de la actividad 1:

Problema de satisfacción de clientes en las tiendas virtuales reportadas en el portal de datos abiertos del Gobierno.

Nombre de la evidencia de aprendizaje	Problema de satisfacción de clientes en las tiendas virtuales reportadas en el portal de datos abiertos del Gobierno.
Objetivo de la evidencia de aprendizaje	Analizar un conjunto de datos, calcular medidas de tendencia central, representar visualmente la frecuencia de los datos y analizar su resultado.
Contenidos	Los datos para analizar se deben descargar del portal de datos abiertos del gobierno nacional, en la dirección y con los criterios indicados.
Descripción de lo que debe hacer el estudiante	El estudiante debe leer las instrucciones para identificar y obtener la fuente de los datos, luego debe cargarlos en la plataforma de *R Studio* y allí hacer los análisis estadísticos que se solicitan, además de la representación gráfica y su análisis.
Especifique lo que debe entregar el estudiante	El estudiante debe crear un script en *R* usando la plataforma *R Studio*, luego descargar el archivo y subirlo como evidencia a la plataforma.

Referencias Bibliográficas

- Barrios, L. (2005). *Parámetros de dispersión*. Descartes 2D. http://recursostic.educacion.es/descartes/web/materiales_didacticos/unidimensional_lbarrios/pdispersion_est.htm

- Cañas, J. & Galo, J. (2015). Estadística. Proyectodescartes.org. https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/IntroduccionEstadisticaProbabilidad/3ESO/6_1RangoyDesviacionMedia.html

- Glosarios especializados. (2019). *Glosario de términos estadísticos*. https://glosarios.servidor-alicante.com/terminos-estadistica/

- López, J. (2017). *Desviación típica - Definición, qué es y concepto*. Economipedia. https://economipedia.com/definiciones/desviacion-tipica.html

- Navarro, J. (2015). *Definición de desviación*. Definición ABC. https://www.definicionabc.com/social/desviacion.php

- Real Academia Española. (2021). Biometría. En *Diccionario de la lengua española* [en línea]. https://dle.rae.es/biometr%C3%ADa

Lecturas y material complementario

- Diseño Gráfico: Steven Miranda Cardona. Dirección de Tecnología. IUD

- **Lectura**
  Autor: Carlos Ochoa
  Título: Muestreo probabilístico o no probabilístico
  URL: https://www.netquest.com/blog/es/blog/es/muestreo-probabilistico-o-no-probabilistico-ii

- **Lectura**
  Autor: Javier Gorgas & Nicolás Cardiel
  Título: Introducción a R
  URL: https://www.ucm.es/data/cont/docs/339-2016-09-29-Introduccion%20a%20R_v1617.pdf

- **Lectura**
  Autor: Cástor Guisande González, Antonio Vaamonde Liste, Aldo Barreiro Felpeto
  Título: Tratamiento de datos con R, statistica y spss
  URL: https://blog.utp.edu.co/estadistica/files/2017/09/TRATAMIENTO-DE-DATOS-CON-R-ESTATISTICA-Y-SPSS.pdf

- **Lectura**
  Autor: Carlos J. Gil Bellosta
  Título: R para profesionales de los datos: una introducción
  URL: https://www.datanalytics.com/libro_r/index.html

- **Lectura**
  Autor: Universidad Benito Juárez
  Título: Bienvenidos al análisis de datos con herramientas estadísticas
  URL: https://sites.google.com/site/estadisticadm/home

- **Video**
  Autor: UOC - Universitat Oberta de Catalunya
  Título: RStudio Cloud en la docencia online en asignaturas de metodología y estadística | Webinar UOC
  URL: https://youtu.be/Rw86uqi3o5E?t=900 (Minuto 15:00)

- **Video**
  Autor: Juan Correa
  Título: Tutorial para usar R studio cloud
  URL: https://youtu.be/zr63MPKl2kw?t=115 (Minuto 1:55)

- **Lectura**
  Autor: Juan Bosco Mendoza Vega
  Título: R para principiantes
  URL: https://bookdown.org/jboscomendoza/r-principiantes4/

- **Lectura**
  Autor: Gorgas, J. and Cardiel
  Título: Introducción a R
  URL: https://www.ucm.es/data/cont/docs/339-2016-09-29-Introduccion%20a%20R_v1617.pdf