Estadística I & Estadística Descriptiva

Coordinación de Estadística - UFME

Actualizado al 03-02-2023

Índice general

Presentación del curso

La asignatura Estadística I & Estadística Descriptiva, es el primer curso estadístico de la carrera de Ingeniería Comercial e Ingeniería en Control de Gestión respectivamente, los cuales, entregan las herramientas necesarias para realizar análisis descriptivos de datos, incluyendo formas simples de modelar relaciones entre variables con la intención de facilitar e iluminar la toma de decisiones económicas, de negocios, entre otros ámbitos. Al mismo tiempo, entrega los fundamentos básicos de la teoría de probabilidades para la modelación de fenómenos con base probabilística.

Esta asignatura aspira a enseñar estadística de forma aplicada, haciendo uso de herramientas modernas de programación, situando al estudiante en un rol de analista dentro de una unidad organizacional.

Modalidad de trabajo

Unidad 1

Tópicos básicos de estadística

Para los ejemplos y ejercicios de esta unidad se hará uso la base de datos *Tasa+euro+dolar+historica2023.csv* cuando corresponda. La base de datos contiene el registro diario histórico de la tasa de cambio del Euro a Dólar, el detalle de las columnas es el siguiente:

- Date: Fecha de medición (yyyy-mm-dd), desde enero del 2003 hasta enero del 2023.
- Open: tasa de apertura.
- High: tasa más alta alcanzada en el día.
- Low: tasa más baja alcanzada en el día.
- Close: tasa de cierre del día.
- Adj Close: tasa de cierre ajustada del día (precio de cierre sin dividendos).

El código para cargar la base de datos en R es:

```
datos =
 ↪  read.csv("https://raw.githubusercontent.com/Dfranzani/Bases-de-datos-para-cursos/main/2023-1/
```

1.1. Conceptos

En esta sección repasaremos algunos conceptos claves de la estadística que están asociados a las ciencias cognitivas. Luego, se ahondará en las técnicas básicas de visualización para el estudio de estos.

1.1.1. Datos

El dato es la unidad básica de la estadística. Esta unidad es cualquier evento o hecho que no ha sido dotado de significado, es decir, un hecho del cual no se puede dar interpretación alguna (Brachman and Levesque, 2004).

Un ejemplo de este concepto, es cuando tratamos de responder la pregunta ¿por qué al caminar nos detenemos al encontrarnos con un semáforo en rojo? ¿Cuál es el dato? ¿Cuál es el significado?

1.1.2. Información

Información = Datos + Significado

Por otro lado, los datos existen independiente de quien observa, y cuando una persona adquiere datos y los dota de significado, estos se convierten en información (Brachman and Levesque, 2004). Otra forma de entenderlo es:

Información = Datos + Reglas para decodificar

En el ejemplo anterior, el decodificador es la persona que va caminando, y el significado (reglas para decodificar) que le damos al semáforo al estar en rojo, viene de las reglas sociales que indican como actuar en determinadas situaciones.

En estadística, mediante el uso de distintas herramientas (gráficos, tablas, entre otras), dotaremos de significado a los datos, para así generar información de utilidad en distintos fenómenos de estudio.

1.1.3. Tipos de variables

El concepto de datos está fuertemente ligado a su naturaleza, es decir, el contexto de estudio que que los rodea. En este sentido, los datos están asociados a lo que llamamos variable ("naturaleza del dato", "los tipos de valores que adquiere el dato"), las cuales, se pueden clasificar la siguiente manera (Anderson et al., 2008, página 7):

- **Cualitativas** (Nominales y Ordinales): variables no numéricas que pueden o no llevar un orden, respectivamente.
- **Cuantitativas** (Discretas y Continuas): variables numéricas que pueden o no ser enumeradas, respectivamente.

Ejercicio

1. Determinar la clasificación de las siguientes variables: tiempo, dinero, altura, cantidad de vecinos en el lugar donde vivo, grado de conformidad (conforme, medianamente conforme, nada conforme) respecto a un servicio, color de pelo de un grupo de personas.

1.1.4. Población y Muestra

Los ingenieros y científicos constantemente están expuestos a la recolección de hecho o datos, tanto en sus actividades profesionales como en sus actividades diarias. La disciplina de estadística proporciona métodos para organizar y resumir datos y de sacar conclusiones basadas en la información contenida en datos.

Una investigación típicamente se enfocará en una colección bien definida de objetos que constituyen una **población** de interés. Cuando la información deseada está disponible para todos lo objetos de la población, se tienen lo que se llama un **censo**. Las restricciones de tiempo, dinero y otros recursos escasos casi siempre hacen que un censo sea infactible. En su lugar, se selecciona un subconjunto de la población, una **muestra**, de manera prescrita (Devore, 2008, página 2).

- **Población**: La población es el conjunto de todos los sujetos de interés en un estudio.
- **Muestra**: La muestra es un subconjunto de la población a través de los cuales el estudio recoge los datos.

Ejercicio:

1. Determine la población y muestra de los siguientes enunciados:

- Se realiza un sondeo para determinar los rubros con mayor inflación de venta de mercado en Santiago, para ello se estudia el rubro con mayor ingreso líquido de ventas, en algunas de las comunas de Santiago.
- La encuesta ENUSC elabora anualmente un informe respecto a la seguridad ciudadana, para ello, se contacta a una cantidad de personas determinadas de cada región del país, dando así, resultados a nivel nacional y regional.

1.1.5. Parámetros y Estadísticos

Ambos conceptos están fuertemente ligados a los de población y muestra de la siguiente manera (Anderson et al., 2008, página 83):

- **Parámetros**: corresponde a una característica de resumen de la población.
- **Estadísticos**: corresponde a una característica de resumen de la muestra.

En la figura 1.1 se observa un ejemplo de círculos rojos y azules tanto para la población como para una muestra de esta. Dado que la población contiene todos los datos (censo), es posible apreciar todos los círculos con sus colores. Por otro lado, la muestra es solo una pequeña parte de la población, es decir, seleccionan algunos de los círculos al "azar" con sus respectivos colores.

Un ejemplo de los conceptos explicados es **la proporción de círculos rojos**. En caso de que estuviésemos interesados en dicha característica en la población,

se hablaría de un parámetro, mientras que, si se está interesado en la muestra se hablaría de estadístico.

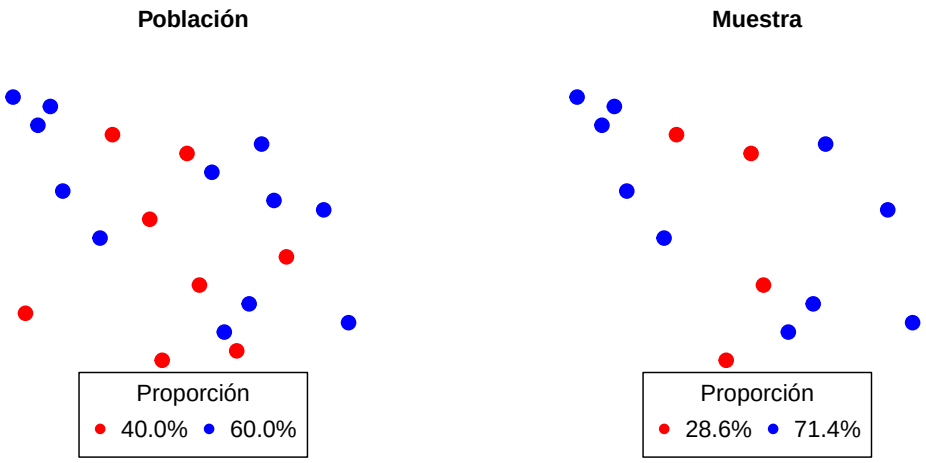


Figura 1.1: Parámetro y estadístico

1.1.6. Estimador y Estimación

Una extensión de los conceptos de parámetro y estadístico, son los de estimador y estimación, para los cuales, se hace la siguiente distinción:

- **Estimador**: Un estimador es un estadístico usado para aproximar (incertidumbre) el valor de un parámetro. Usualmente no cambia la técnica entre la población y la muestra, por ejemplo, si deseo aproximar la proporción de bolitas rojas en la población, se usaría la proporción de bolitas rojas en la muestra.
- **Estimación**: Una estimación es el número que resulta de aplicar el estimador a una muestra particular. Esto difiera levemente de la definición anterior, ya que en términos estrictos, el estimador solo es la "fórmula", y la estimación es el valor resultante al aplicar la fórmula. Sin embargo, hoy en día es muy común encontrar textos en donde el estimador se considera tanto para la fórmula como para el valor obtenido.

Si consideramos un ejemplo similar al anterior (Figura 1.2), y establecemos que el **parámetro** a estudiar es la proporción de círculos rojos, es natural pensar que en la muestra (**estadístico**) el comportamiento debería ser similar. La intención de decir "usaremos la proporción de círculos rojos en la muestra para deducir como es la proporción de círculos rojos en la población" corresponde al **estimador** (otro tema es argumentar si esto es correcto o no), mientras que, el cálculo del estimador (cálculo de la proporción de círculos rojos en la muestra) lleva el nombre de **estimación**.

Respecto a lo anterior:

- ¿Cuál sería la estimación de los círculos rojos?
- Si observamos la muestra de la figura 1.1 y 1.2, ¿cuándo diríamos que una estimación es buena?

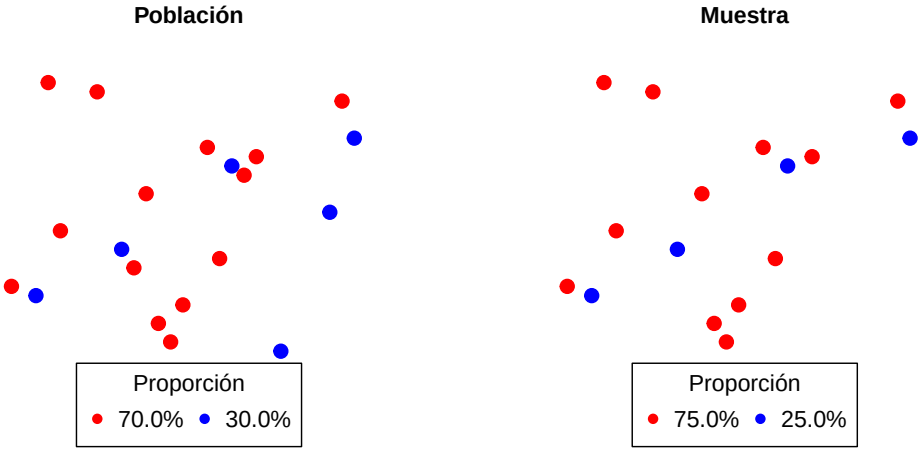


Figura 1.2: Estimador y estimación

1.1.7. Variabilidad muestral

Efectivamente, la estimación de un parámetro está determinada por la muestra con la que se trabaja. La forma en la que se elige una muestra es azarosa (que no se puede intencionar en su totalidad), por lo que es imposible saber de antemano si la estimación será buena o mala respecto al parámetro (error de estimación). En estadística, la forma en la que se elige o genera una muestra puede llegar a ser muy compleja, siendo un tema que está fuera del alcance de este curso.

El concepto detrás de esto es la **variabilidad muestral**, el cual, indica que dependiendo de la muestra que se obtenga de la población, esta se comportará distinto en relación al estadístico (igualmente para el valor del estimador: estimación). Para ilustrar esto, observemos la figura 1.3.

¿Cuál es la proporción de círculos rojos en la población relfejada en la figura ?

Luego,

¿qué podríamos inferir sobre el color predominante en la población en base a la muestra de la figura 1.4**?**

¿Y ahora? (Figura 1.5**)**

Efectivamente, diferentes muestras se comportan de manera diferente, es decir, la estimación depende de la selección de la muestra. Esto se denomina como

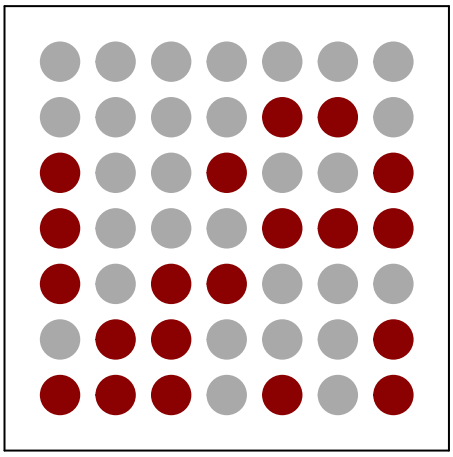


Figura 1.3: Población

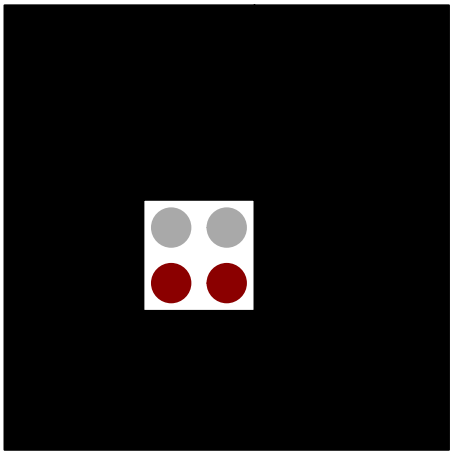


Figura 1.4: Muestra 1

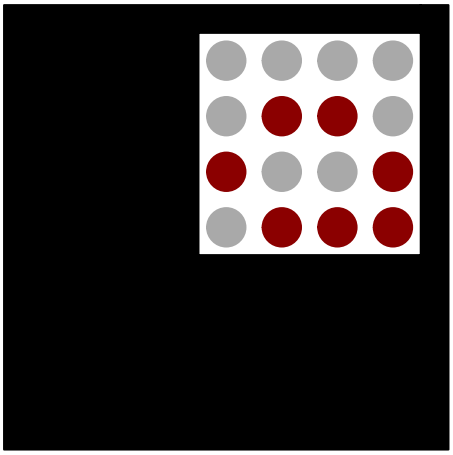


Figura 1.5: Muestra 2

variabilidad muestral.

1.1.8. Representatividad y sesgo de la muestra

Ambos conceptos se usan con frecuencia en la vida cotidiana, y a su vez están mal empleados. El sesgo no es una propiedad de la muestra sino que del estimador (concepto avanzado de estadística).

Por otro lado, la representatividad no es un concepto válido matemáticamente (no existe tal definición).

1.1.9. Medidas de localización

Los resúmenes visuales de datos son herramientas excelentes para obtener impresiones y percepciones preliminares. Un análisis de datos más formal a menudo requiere el cálculo e interpretación de medidas de resumen numéricas. Es decir, de los datos se trata de extraer varios números resumidos, números que podrían servir para caracterizar el conjunto de datos. Las tres medidas de resumen más utilizadas son la media, la medianda y la moda.

Media

Para un conjunto dado de números $x_1, x_2, x_3, \ldots, x_n$, la medida más conocida y útil es la **media** o promedio aritmético. Usualmente se asume que los números x_i hace parte de una muestra, por lo que a este promedio se le connota como **media muestral** y se denota con por $\bar{x}$.

De lo anterior, la media muestral ($\bar{x}$) de una conjunto de datos $x_1, x_2, x_3, \ldots, x_n$ está dada por (Devore, 2008, página 25)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1.1}$$

En R, para obtener el promedio aritmético de los datos se hace uso de la función **mean()**. A continuación, un ejemplo.

```
# Un conjunto de datos cualquiera
x = c(1,2,3,6,1,-4,-2,6,0,10,-20)
# Promedio de los datos
mean(x)
```

```
## [1] 0.2727273
```

El promedio ($\bar{x}$) representa el valor central de las observaciones incluidas en una muestra. Sin embargo, esta medida puede llegar a ser inapropiada en algunas circunstancias, específicamente cuando existen valores extremos. Un ejemplo de esto, es el promedio de los ingresos (el caso de Chile), ya que, es común que unos cuantos afortunados ganen cantidades astronómicas, por lo que el uso del ingreso promedio como medida de resumen puede ser engañoso (otro ejemplo, es la valorización de BitCoin al dólar estadounidense).

A pesar de lo anterior, esta medida sigue siendo ampliamente utilizada, en gran medida porque existen muchas poblaciones para las cuales un valor extremo en la muestra sería altamente improbable (ejemplo: tipo de cambio del dólar y el euro).

Ejercicio:

1. Utilizando la base de datos de la unidad, obtenga la media de la variable (columna) **Open**. Interprete.
2. Utilice el comando **colMeans()** para obtener la media de todas las variables asociadas a la tasa de conversión (ignore la columna asociada a la variable fecha). Interprete

Mediana

La palabra mediana es sinónimo de "medio" y la mediana muestral es en realidad el valor medio una vez que se ordenan las observaciones de la más pequeña a la más grande (Devore, 2008, página 26).

La mediana muestral se obtiene ordenando primero las observaciones de la más pequeña a la más grande. Por lo tanto,

- Si la cantidad de datos es impar, entonces, la mediana es igual al número en la posición $\frac{n+1}{2}$.
- Si la cantidad de datos es par, entonces, la mediana es el promedio entre los números ubicados en las posiciones $\frac{n}{2}$ y $(\frac{n}{2} + 1)$.

Para poder calcular la mediana en R, se debe hacer uso del comando **median()**, tal como se muestra a continuación.

```
# Conjunto de datos (cantidad impar)
x = c(1,2,3,4,5,6,7,-3,-1,-2,5.4,9.3,0)
# Mediana del conjunto de datos
median(x)
```

```
## [1] 3
```

```
# Conjunto de datos (cantidad par)
x = c(1,2,3,4,5,6,7,-3,-1,-2,5.4,9.3)
# Mediana del conjunto de datos
median(x)
```

```
## [1] 3.5
```

En ambos casos, se entiende que, ordenando los datos de menor a mayor (en una recta real), tanto a la derecha como izquierda de la mediana se encuentra la misma cantidad de datos.

Ejercicio:

1. Utilizando la base de datos de la unidad, determine la mediana de cada una de las variables presentes en la base (ignore la columna asociada a la variable fecha). Interprete.

Moda

La moda es la medida más intuitiva de las tres, ya que simplemente corresponde al valor que se presenta con mayor frecuencia (Anderson et al., 2008, página 85). Para ilustrar esto, veamos el siguiente código en R:

```
# El siguiente vector contiene la información de la cantidad de
 ↪  hermanos
# de un determinado grupo de personas
hermanos = c(1,2,3,1,2,3,3,3,4,1,7,1,0,0,1,0,2)
# Utilizando el comando table podemos obtener la frecuencia de
 ↪  cada una
# de las distintas observaciones
table(hermanos)
```

```
## hermanos
## 0 1 2 3 4 7
## 3 5 3 4 1 1
```

```
# Como resultado se aprecia que la cantidad de hermanos que más
 ↪  se repita dentro
# del grupo de personas es de 5
```

Ejemplo:

1. Cree un objeto que guarde la tabla de frecuencias de la variable Open de la base de datos de la unidad (sin imprimir la tabla).

```
tabla = table(datos$Open)
```

2. Ya que es imposible buscar manualmente la frecuencia más alta, utilice el comando **which.max()** para encontrar la posición en la que se ubica esta, ingresando como argumento la tabla anteriormente guardada. Guarde este valor en un objeto.

```
(posicion = which.max(tabla))
```

```
## 1.336005
##     3067
```

3. Finalmente, consulte de manera directa en la tabla en valor de la frecuencia en la posición calculada en el paso anterior. Interprete.

```
tabla[posicion]
```

```
## 1.336005
##        6
```

Esto quiere decir, que el valor de apertura de la tasa EUR/USD que más se repite históricamente es 1.336005 con una frecuencia de 6.

Nota: En caso de que existan dos o más valores con las frecuencias más altas, el programa solo reporta la primera, según el orden lexicográfico de las columnas.

Ejercicio:

1. Replique los anterior para el resto de variables presentes en la base de datos de la unidad (ignore la columna asociada a la variable fecha).

Nota: en el documento se usará simplemente el nombre de la medida de localización (media, moda, mediana) para referirse a la medida de localización muestral. En casos determinados se hará la distinción entre el caso muestral y poblacional, según corresponda (ejemplo: media poblacional, media muestral).

1.1.10. Medidas de escala

Al momento de reportar la media solo se obtiene información parcial sobre el un conjunto de datos. Diferentes muestras o poblaciones pueden tener medidas idénticas de localización y aún diferir entre sí en otras importantes maneras. La tabla 1.1 muestra las notas obtenidas por los alumnos de 2 dos cursos con la misma media, aunque el grado de **dispersión** (variabilidad) en torno a esta es diferente para ambas muestras, es decir, en el Curso 1 las se observan notas más bajas y altas que el Curso 2.

Tabla 1.1: Notas por curso

Curso 1	Curso 2	Curso 3
3.0	4.0	3.0
4.5	5.0	7.0
5.0	6.0	–
5.5	–	–
7.0	–	–

Rango

La medida más simple de variabilidad en una muestra es el **rango**, el cual es la diferencia entre los valores muestrales más grande y más pequeño (Devore, 2008, página 32). El rango de las notas del curso 1 en la tabla 1.1 es más grande que el del curso 2, lo que refleja más variabilidad en la primer muestra que en la segunda. Un defecto del rango, no obstante, es que depende de solo las dos observaciones más extremas y hace caso omiso de las posiciones de los valores restantes. Los cursos 1 y 3 tienen rangos idénticos, aunque cuando se toman en cuenta las observaciones entre los dos extremos, existe mucho menos variabilidad o dispersión en la tercera muestra que en la primera.

Ejemplo:

Obtengamos el rango de la tasa de apertura histórica del EUR/USD de la base de datos de la undiad.

```
# Utilizando el comando range() se obtienen los valores mínimo y
 ↪  máximo
# de la variable en cuestión
(rango = range(datos$Open))
```

```
## [1] 0.959619 1.598184
```

```
# Luego calculamos la diferencia entre el valor máximo y mínimo
rango[2] - rango[1]
```

```
## [1] 0.638565
```

```
# Nota: El valor máximo siempre estará en la segunda posición y
 ↪  le mínimo en la segunda.
```

Ejercicio:

1. Obtener el rango del resto de variables de la base de datos de la undiad (ignore la columna asociada a la variable fecha).

Varianza y desviación estándar

Las medidas principales de variabilidad implican las **desviaciones de la media**,

$$x_1 - \bar{x}, x_2 - \bar{x}, x_3 - \bar{x}, \dots, x_n - \bar{x}. \tag{1.2}$$

Es decir, las desviaciones de la media se obtienen restando $\bar{x}$ de cada una de las n observaciones muestrales. Una desviación será positiva si la observación es más grande que la media (a la derecha de la media sobre la recta real) y negativa si la observación es más pequeña que la media (a la izquierda de la media sobre la recta real). Si todas las desviaciones son pequeñas en magnitud, entonces todos los valores de la muestra son cercanos a la media y hay poca variabilidad. Alternativamente, si algunas de las desviaciones son grandes de magnitud, entonces algunos de los de los valores de la muestra están lejos de la media (sobre la recta real) lo que sugiere una mayor variabilidad.

Una forma de resumir las desviaciones sería sumando todas ellas. Sin embargo, es una mala idea, ya que la suma siempre es igual a cero (1.3), ¿alguna idea del por qué?

$$\text{Suma de las desviaciones en una muestra} = \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \tag{1.3}$$

En este sentido, para poder resumir las desviaciones de una muestra evitando el problema mencionado, se elaboran dos expresiones (Devore, 2008, página 32):

- Varianza (muestral):

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1.4}$$

- Desviación estándar (muestral):

$$S = \sqrt{S^2} \tag{1.5}$$

Las unidades correspondientes a la **varianza** suele causar confusión. Como los valores que se suman para calcular la varianza, $(x_i - \bar{x})^2$, están elevados al cuadrado, las unidades correspondientes a la varianza muestral también están elevadas al cuadrado. Las unidades al cuadrado de la varianza dificulta la compresión e interpretación intuitiva de los valores numéricos de la varianzas. Lo recomendable es entender la varianza como una medida útil para comparar la variabilidad de dos o más variables. Al comparar variables, la que tiene la varianza mayor, muestra más variabilidad. Otra interpretación del valor de la varianza suele ser innecesaria (Anderson et al., 2008, página 94).

La **desviación estándar** es la raíz cuadrada de la varianza, pero, ¿qué se gana con esto? Al calcular la desviación estándar, las unidades de esta son iguales a de la variable original, por lo que es más fácil de interpretar. Sin embargo, estas dos medidas tiene ciertas limitantes a la hora de comprar la variabilidad de dos variables:

1. Es ideal que ambas variable tengan la misma media.
2. Las variables deben tener la misma unidad de medida.

No seguir estas recomendaciones puede generar un falsa sensación en la comunicación de resultados.

Ejemplo:

Compare la variabilidad entre la tasa de apertura y la tasa de cierre histórica del EUR/USD presentes en la base de datos de la unidad, para ello:

1. Verifique la media de ambas variables la misma

```
mean(datos$Open) # Promedio de la tasa de apertura
```

```
## [1] 1.244338
```

```
mean(datos$Close) # Promedio de la tasa de cierre
```

```
## [1] 1.244363
```

```
# Las tasas son similares hasta el tercer decimal, se asumirá que
 ↪  las medias son iguales
```

2. Ya que tienen la misma unidad de medida, calcule la varianza y desviación estándar de cada una. Interprete.

```
# Al calcular la varianza muestral, se observa que la tasa de
 ↪  cierre es levemente menor
# variabilidad que la tasa de apertura.
c(var(datos$Open), var(datos$Close))
```

```
## [1] 0.01562596 0.01562404
```

```
# Al calcular la desviación estándar muestral, se observa que la
 ↪  tasa de cierre tiene menor
# variabilidad que la tasa de apertura.
c(sd(datos$Open), sd(datos$Close))
```

```
## [1] 0.1250038 0.1249962
```

```
# ¿Por qué es más clara la interpretación (primer decimal
 ↪  distinto) al utilizar
# la desviación estándar?
```

Ejercicio:

1. Utilice la varianza directamente para comparar la variabilidad entre la tasa de apertura y la tasa más alta histórica del EUR/USD presentes en la base de datos de la unidad.

Coeficiente de variación

Para subsanar el problema de las limitaciones de la varianza y desviación estándar, se encuentra la medida llamada **coeficiente de variación** (1.6).

$$CV = \left(\frac{S}{|\bar{x}|} \right) \cdot 100\,\% \tag{1.6}$$

Cuando el valor del coeficiente de variación es cercano a 100 % se habla de mayor dispersión (heterogéneo), mientras que un valor cercano a 0 % indica menor dispersión (homogéneo), además, se debe considerar que el porcentaje calculado corresponde a la variabilidad respecto a la media de los datos. Sin embargo, no es recomendable usar esta medida cuando el valor de la media es cercano a cero, ya que el CV pierde su significado al tomar valores muy grandes, lo que daría una falsa sensación de dispersión de los datos (Anderson et al., 2008, página 95).

Ejemplo:

En el ejercicio anterior, se utilizó la varianza para comprar directamente la variabilidad entre la tasa de apertura y la tasa más alta histórica del EUR/USD. Sin embargo, si calculamos las medias de ambas variables se puede verificar que son distintas. Utilice el CV para comprar la variabilidad de ambas variables.

```
# Claramente la media de la tasa más alta es mayor a la media de
 ↪  la tasa de apertura.
c(mean(datos$Open), mean(datos$High))
```

```
## [1] 1.244338 1.249022
```

```
# Al verificarse una de las dos limitantes mencionadas,
 ↪  procedemos
# a calcular el CV de ambas variables
CV_Open = sd(datos$Open)/abs(mean(datos$Open))*100
CV_High = sd(datos$High)/abs(mean(datos$High))*100
c(CV_Open, CV_High)
```

```
## [1] 10.04581 10.06323
```

```
# Se puede observar que el coeficiente de variabilidad de la tasa
 ↪  más alta (10.06%)
# es mayor a la de la tasa de apertura (10.04%). Por lo tanto,
```

Tabla 1.2: Notación de parámetros y estadísticos

	Poblacional	Muestral
Media	μ	$\bar{x}$
Varianza	σ^2	S^2
Desviación estándar	σ	S

```
# la variabilidad (dispersión) de los datos es más homogénea para
 ↪  la tasa de
# apertura. Sin embargo, la diferencia es muy pequeña, por lo que
 ↪  la dispersión en
# relación a la media es similar entre ambas variables.
```

Ejercicio:

1. Comparar la variabilidad entre la tasa de más baja y de cierre histórica del EUR/USD presentes en la base de datos de la unidad. Interprete.

Nota: en el documento se usará simplemente el nombre de la medida de escala (rango, varianza, desviación estándar y CV) para referirse a la medida de escala muestral. En casos determinados se hará la distinción entre el caso muestral y poblacional, según corresponda (ejemplo: varianza poblacional, varianza muestral).

Notación poblacional y muestral

1.2. Gráficos descriptivos

En este apartado, se considera la representación de un conjunto de datos por medio de técnicas visuales. A continuación, se hará mención de algunas de las técnicas más útiles y pertinentes a la estadística de descriptiva. Los ejemplos presentados en esta sección hacen uso de la base de datos de la unidad (sección 1).

1.2.1. Histograma

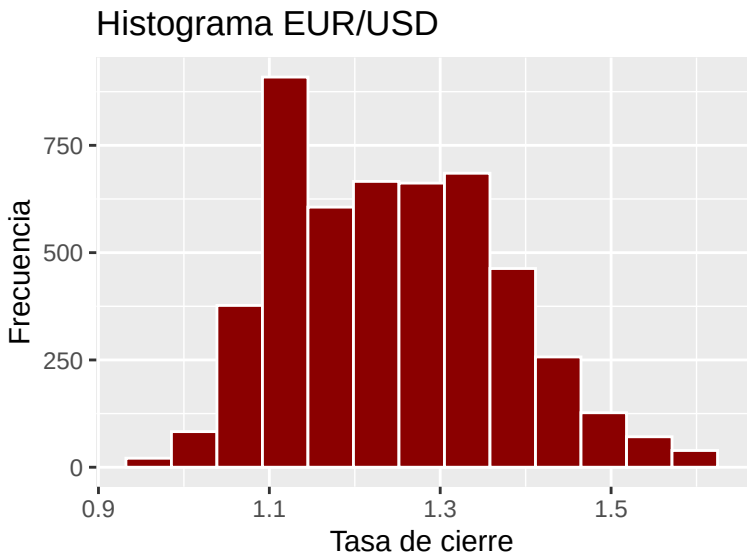
Algunos datos numéricos se obtienen contando para determinar el valor de una variable (cuántas veces se repite un hecho), mientras que otro datos se obtienen tomando mediciones (peso, altura, tiempo de reacción). Usualmente, este tipo de gráfico se utiliza con datos continuos (aunque tiene una versión para datos discretos), para lo cual, se debe hacer lo siguiente (Devore, 2008, página 12):

1. Subdividir los datos en **intervalos de clase** o **clases**, de tal manera que cada observación quede contenida en exactamente una clase. Para esto, se hace uso de la regla de Sturges (1926), la cual, consiste en calcular la

expresión $1 + \log_2(n)$, aproximando hacia el entero más próximo, dónde n corresponde a la cantidad de datos (existen otra variedad de técnicas).
2. Determinar la frecuencia y la frecuencia relativa de cada clase, es decir, cuántas observaciones hay en cada uno de los intervalos.
3. Se marcan los límite de clase sobre el eje horizontal del plano cartesiano.
4. Se traza un rectángulo cuya altura es la frecuencia absoluta (o relativa) correspondiente a cada intervalo de clase.

Para generar un histograma en R a partir de un conjunto de datos, se utiliza el siguiente código:

```
library(ggplot2)
ggplot(data = datos, # Base de datos a utilizar
       aes(x = Close)) + # Comando estético: eje y variable
  geom_histogram(bins = 13, # Cantidad de intervalos del
   ↪  histograma
                 color = "white", # Color del borde de las barras
                  ↪  del histograma
                 fill = "darkred", # Color de relleno de las
                  ↪  barras
                 closed = "left") + # Tipo de intervalo del
                  ↪  histograma
  labs(title = "Histograma EUR/USD", # Título del gráfico
       x = "Tasa de cierre", # Título del eje X
       y = "Frecuencia") # Título del eje Y
```



Es útil recordar que el histograma está asociado a una tabla de frecuencia por intervalos. Para obtener la tabla asociada a un histograma se puede utilizar el siguiente código.

```
# Datos del histograma guardados
h = hist(datos$Close, # Datos a graficar en el histograma
         breaks = 13, # Cantidad de intervalos (Sturges)
         right = F, # Cerrado por la izquierda
         plot = F) # No desplegar el gráfico en consola
library(agricolae) # Librería para generar la tabla de
 ↪  frecuencias
print(table.freq(h)) # Imprime en consola la tabla de frecuencias
```

```
##    Lower Upper  Main Frequency Percentage   CF   CPF
## 1   0.95  1.00 0.975        46        0.9   46   0.9
## 2   1.00  1.05 1.025        89        1.8  135   2.7
## 3   1.05  1.10 1.075       444        8.9  579  11.7
## 4   1.10  1.15 1.125       839       16.9 1418  28.6
## 5   1.15  1.20 1.175       591       11.9 2009  40.5
## 6   1.20  1.25 1.225       634       12.8 2643  53.2
## 7   1.25  1.30 1.275       614       12.4 3257  65.6
## 8   1.30  1.35 1.325       654       13.2 3911  78.8
## 9   1.35  1.40 1.375       510       10.3 4421  89.0
## 10  1.40  1.45 1.425       257        5.2 4678  94.2
## 11  1.45  1.50 1.475       166        3.3 4844  97.5
## 12  1.50  1.55 1.525        38        0.8 4882  98.3
## 13  1.55  1.60 1.575        84        1.7 4966 100.0
```

Ejercicio:

1. Utilizando la variable **Low** de la base de datos de la unidad, elabore un histograma y obtenga la tabla de frecuencias asociada. Interprete.

1.2.2. Caja

El gráfico de caja se utiliza para describir las siguiente características de un conjunto de datos (Devore, 2008, página 35):

- El centro.
- La dispersión.
- El grado y naturaleza de cualquier alejamiento de la simetría.
- La identificación de las observaciones "extremas" (atípicas) inusualmente alejadas del cuerpo principal de los datos.

Los pasos para elaborar un gráfico de caja son los siguiente (Anderson et al., 2008, página 106):

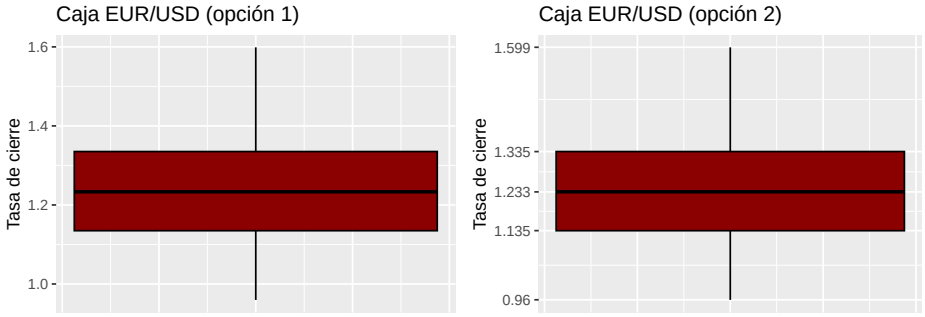
1. Se dibuja una caja cuyos extremos se localicen en primer y tercer cuartiles. Esta caja contiene 50 % de los datos centrales.
2. En el punto donde se localiza la mediana se traza una linea horizontal.
3. Usando el rango intercuartílico ($RIC = Q_3 - Q_1$), se localizan los límites. En un gráfico de caja los límites se encuentra a 1,5RIC abajo y arriba de

Q_1 y Q_3 respectivamente. Los datos que quedan fuera de estos límites se consideran observaciones atípicas (Tukey, 1977). La razón por la cual se considera 1.5 veces el rango intercuartílico es convencional, no obstante, hay argumento relacionados a la cantidad de datos dentro de los limites inferior y superior, los cuales indican que debe ser de 99.7 % (James et al., 2013).

4. Las lineas que se extienden verticalmente desde la caja se les llama *bigotes*. Los bigotes van desde los extremos de la caja hasta los valores menor y mayor de los límites calculados en el paso 3.
5. Mediante puntos se indica la localización de las observaciones atípicas.

Para generar un gráfico de caja en R a partir de un conjunto de datos, se utiliza el siguiente código:

```
g = ggplot(data = datos, # Base de datos a utilizar
       aes(y = Close)) + # Comando estético: eje y variable
  geom_boxplot(color = "black", # Color del borde del gráfico
               fill = "darkred") + # Color de relleno del gráfico
  labs(title = "Caja EUR/USD (opción 1)", # Título del gráfico
       x = "", # Título del eje X
       y = "Tasa de cierre") + # Título del eje Y
  theme(axis.ticks.x = element_blank(), # Elimina las regletas
    ↪  del eje X
        axis.text.x = element_blank()) # Elimina los números del
         ↪  eje X
info = unlist(ggplot_build(g)[[1]]) # Guardamos los valores del
 ↪  gráfico
values = round(as.numeric(info[1:5]), 3) # Extraemos los valores
 ↪  de construcción
g1 = g + # Creamos un nuevo gráfico a partir del anterior
  scale_y_continuous(breaks = values, labels = values) + #
    ↪  Modificamos los valores mostrados en el eje Y (valores de
    ↪  construcción)
  labs(title = "Caja EUR/USD (opción 2)")
library(gridExtra)
grid.arrange(g, g1, ncol = 2) # Desplegamos lo gráficos en la
 ↪  misma ventana
```



Ejercicio:

1. Utilizando la variable **Open** de la base de datos de la unidad, elabore un gráfico de caja. Interprete.

1.2.3. Violín

El gráfico de violín proporciona una representación más completa y precisa de la distribución de los datos que las técnicas anteriores, ya que muestra tanto la forma de la distribución como su concentración (Hintze and Nelson, 1998). La utilidad de este gráfico recae en la comparación de la distribución de los datos entre distintos grupos y/o categorías.

El proceso de construcción del gráfico es el siguiente:

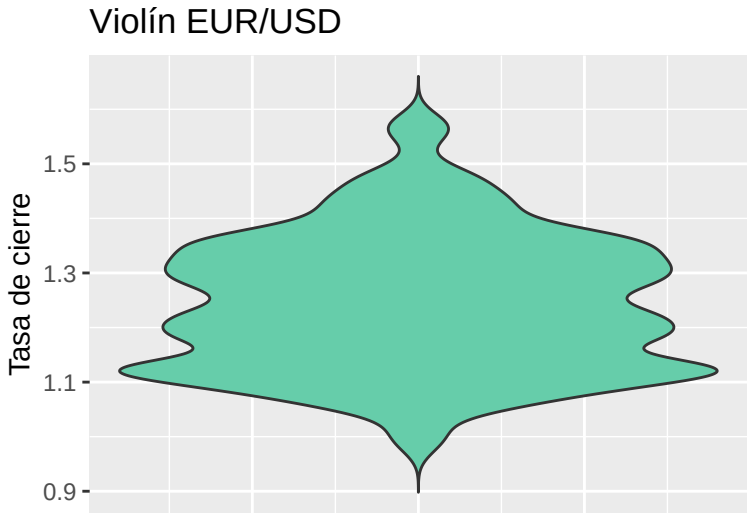
1. Dibujo de la traza de densidad: la traza de densidad se dibuja sobre el eje vertical en el gráfico de violín ("forma suavizada del histograma").
2. Creación de la sección central simétrica: se crea una sección central simétrica que representa la mitad de la traza de densidad.

Adicionalmente, es común agregar un gráfico de caja junto al de violín con el fin de incorporar la visualización de las medidas de posición.

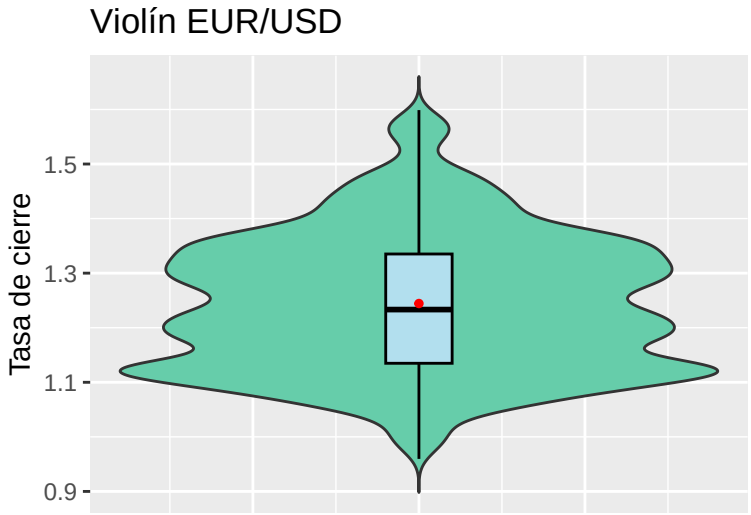
Para generar un gráfico de violín en R a partir de un conjunto de datos, se utiliza el siguiente código:

```
# Se guarda el gráfico en una variable para posteriormente
# integrar otros gráficos dentro de este.
g = ggplot(data = datos, # Base de datos a utilizar
           aes(x = 1, y = Close)) + # Comando estético: variables
            ↪  a graficar
  geom_violin(trim = F, # Modifica las terminaciones visuales
   ↪  superior e inferior
              fill = "aquamarine3") + # Color de relleno del
               ↪  gráfico
  labs(title = "Violín EUR/USD", # Título del gráfico
       x = "", # Título del eje X
       y = "Tasa de cierre") + # Título del eje Y
```

```
  theme(axis.ticks.x = element_blank(), # Elimina las regletas
   ↪  del eje X
        axis.text.x = element_blank()) # Elimina los números del
         ↪  eje X
g # Desplegamos el gráfico en el visualizador
```




```
# Agregamos otros elementos al gráfico guardado
g + geom_boxplot(width = 0.1, # Anchura del nuevo gráfico de caja
                 color = "black", # Color de borde del gráfico
                 fill = "lightblue2") + # Color de relleno del
                  ↪  gráfico
  stat_summary(fun = mean, # Función que indica el tipo de
   ↪  información (mean = media)
               geom = "point", # Forma visual
               size = 1, # Tamaño
               color = "red") # Color
```



Ejercicio:

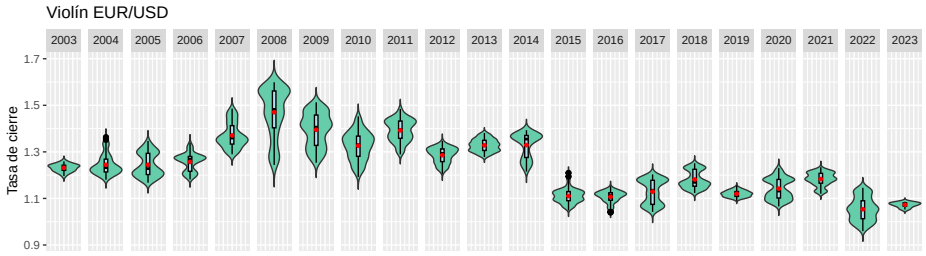
1. Utilizando las variables **Low** y **Open** de la base de datos de la unidad, elabore un gráfico violín incluyendo un gráfico de caja y el promedio para cada una. Interprete.

Ejemplo:

El siguiente código, crea una nueva columna en la base de datos que identifica el año en el que se realizó la medición de las tasas. A continuación, elabore un gráfico de violín (más gráfico de caja y promedio) de la variable **Close**, diferenciando por año. Interprete.

```
datos$Ano = substr(datos$Date, 1, 4)
ggplot(data = datos, # Base de datos a utilizar
       aes(x = 1, y = Close)) + # Comando estético: variables a
        ↪ graficar
  geom_violin(trim = F, # Modifica las terminaciones visuales
   ↪ superior e inferior
              fill = "aquamarine3") + # Color de relleno del
               ↪ gráfico
  geom_boxplot(width = 0.1, # Anchura del nuevo gráfico de caja
                 color = "black", # Color de borde del gráfico
                 fill = "lightblue2") + # Color de relleno del
                  ↪ gráfico
  stat_summary(fun = mean, # Función que indica el tipo de
   ↪ información (mean = media)
               geom = "point", # Forma visual
               size = 1, # Tamaño
               color = "red") + # Color
  labs(title = "Violín EUR/USD", # Título del gráfico
```

```
      x = "", # Título del eje X
      y = "Tasa de cierre") + # Título del eje Y
theme(axis.ticks.x = element_blank(), # Elimina las regletas
 ↪  del eje X
      axis.text.x = element_blank()) + # Elimina los números
       ↪  del eje X
facet_wrap(vars(Ano), # Variable que se utiliza para segregrar
 ↪  los gráficos
           nrow = 1) # Disposición visual: todo en la misma
            ↪  fila
```



Una vez obtenido el gráfico, podemos comparar las dispersiones de los datos en cada uno de los años. En específico, se aprecia que la dispersión está entre 1.1 y 1.3 para la mayoría de los años, a excepción del periodo 2007 - 2014, el cual, refleja valores superiores a este rango.

Ejercicio:

1. Utilizando las variables **Low** y **Open** de la base de datos de la unidad, elabore un gráfico violín incluyendo un gráfico de caja y el promedio para cada una, diferencia por año. Compara e interprete cada variable por separado.

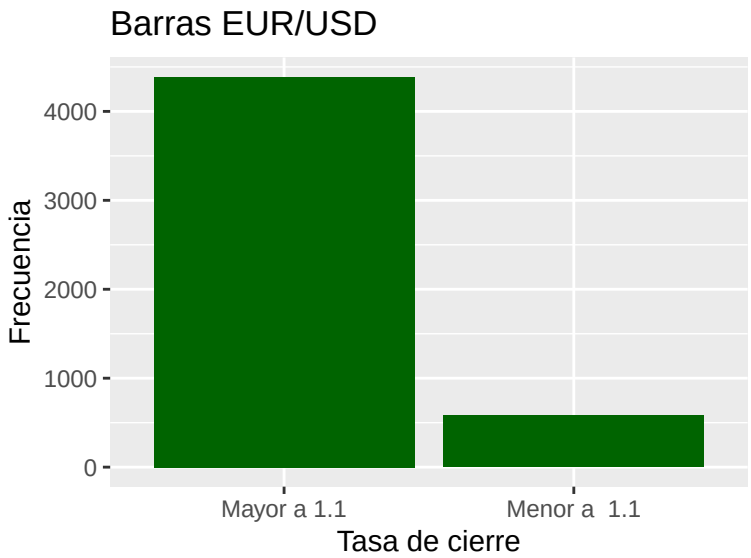
1.2.4. Barras

Una gráfico de barras, es una gráfica para representar los datos cualitativos de una distribución de frecuencia. El procedimiento de construcción es el siguiente (Anderson et al., 2008, página 29):

1. En uno de los ejes de la gráfica (por lo general en el horizontal).
2. Se especifican las etiquetas empleadas para las clases (categorías).
3. Para el otro eje de la gráfica (el vertical) se usa una escala para frecuencia, frecuencia relativa o frecuencia porcentual.
4. Finalmente, se emplea un ancho de barra fijo y se dibuja sobre cada etiqueta de las clases una barra que se extiende hasta la frecuencia de la clase (a diferencia del histograma, las barras deben estar separadas notoriamente).

Para generar un gráfico de barras en R a partir de un conjunto de datos, se utiliza el siguiente código:

```r
# Nueva variable para dicotomizar la tasa de cierre del EUR/USD
datos$Close2 = ifelse(datos$Close > 1.1, # Criterio
                      "Mayor a 1.1", # Valor asignado si se
                       ↪  cumple el criterio
                      "Menor a  1.1") # Valor asignado si no se
                       ↪  cumple el criterio
ggplot(data = datos, # Base de datos a utilizar
       aes(x = Close2)) + # Comando estético: eje y variable
  geom_bar(fill = "darkgreen") + # Color de relleno
  labs(title = "Barras EUR/USD", # Título del gráfico
       x = "Tasa de cierre", # Título del eje X
       y = "Frecuencia") # Título del eje Y
```



Ejercicio:

1. Utilizando las variables **High** y **Open** de la base de datos de la unidad, elabore un gráfico de barras para cada una. Interprete.

1.2.5. Dispersión

El gráfico de dispersión es útil para estudiar la relación entre dos variables continuas. Muestra cómo varía un variable en función de la otra y puede ayudar a identificar patrones y tendencias (Rowlingson, 2016).

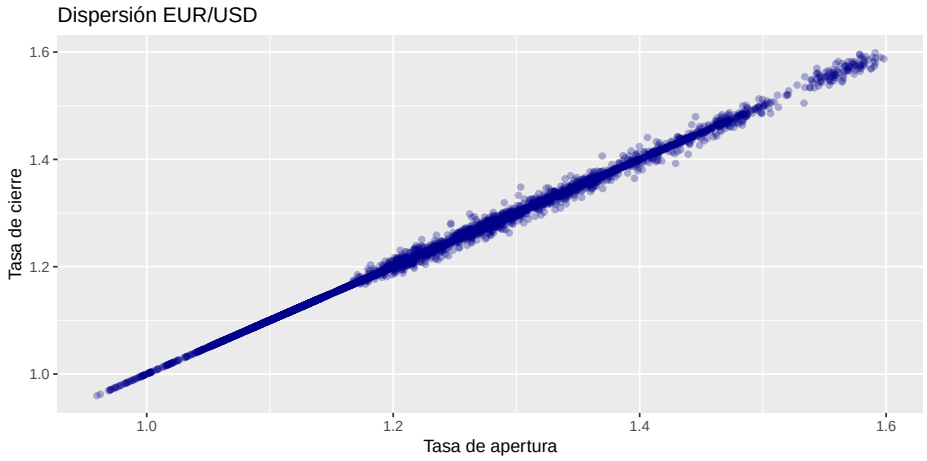
Los pasos para elaborar un gráfico de caja son los siguiente (Healy, 2019):

1. Elegir dos variables continuas de la base de datos a trabajar. Cada fila corresponde a una observación, por lo cual, hay una correspondencia entre los valores de una misma fila.

2. Elegir la variable estará en el eje X y Y.
3. Representar cada par ordenado con un punto.

Para generar un gráfico de dispersión en R a partir de un conjunto de datos, se utiliza el siguiente código:

```
ggplot(data = datos, # Base de datos a utilizar
       aes(x = Open, # Comandos estéticos: Eje X y variable
        ↪  asociada
           y = Close)) +  # Eje Y y variable asociada
  geom_point(color = "darkblue", # Color de los puntos
             alpha = 0.3) + # # Opacidad de los puntos
  labs(title = "Dispersión EUR/USD", # Título del gráfico
       x = "Tasa de apertura", # Título del eje X
       y = "Tasa de cierre") # Título del eje Y
```



Ejercicio:

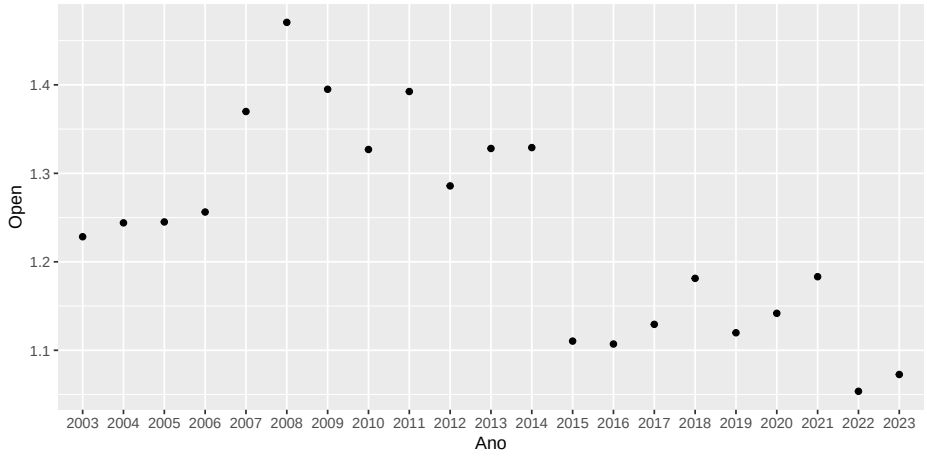
1. Utilizando las variables **High** y **Low** de la base de datos de la unidad, elabore un gráfico de dispersión. Interprete.

Ejemplo:

1. Es posible utilizar el gráfico de dispersión con variables que en su naturaleza son discretas. En este sentido, elabore un gráfico de dispersión entre el año de medición y el valor promedio de tasa de apertura del EUR/USD.

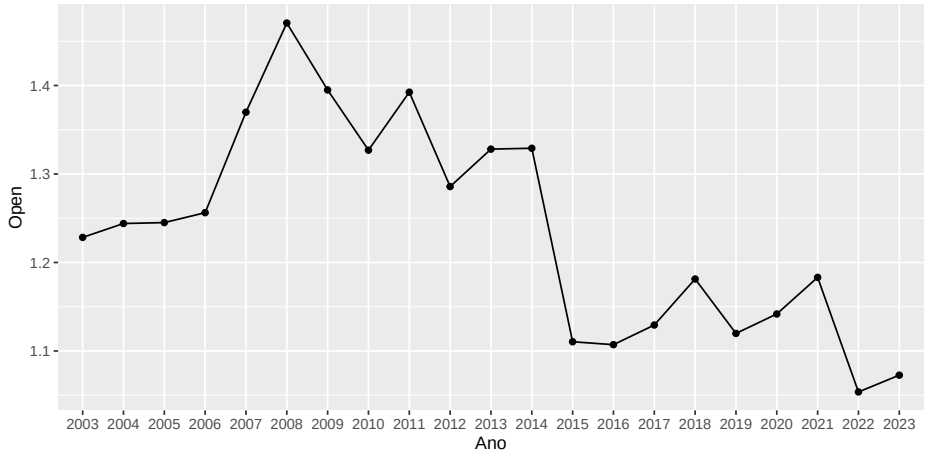
```
g = ggplot(data = datos,
           aes(x = Ano,
               y = Open,
               group = 1)) +
  geom_point(stat = "summary",
```

```
                fun = "mean")
g
```



2. Añadir al gráfico un formato de líneas entre los puntos. Interprete.

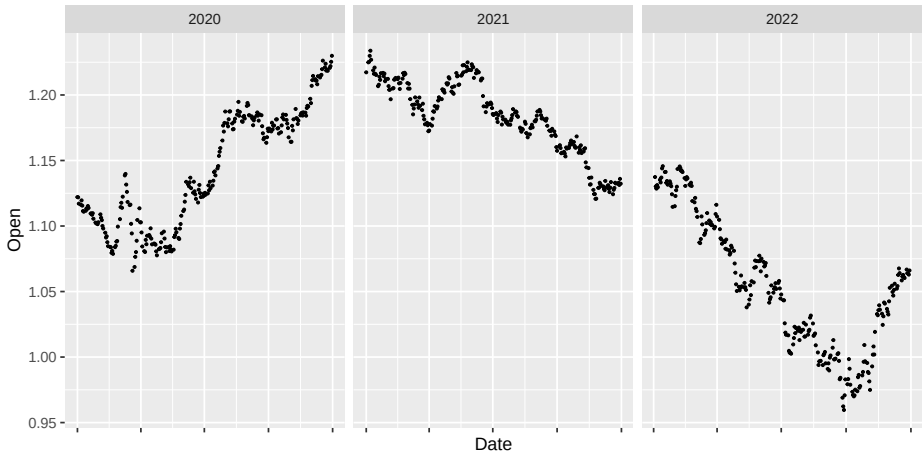
```
g = g +
  geom_line(stat = "summary",
            fun = "mean")
g
```



3. Grafique el valor de la tasa de apertura del EUR/USD desde el 2020 hasta el 2022 en orden temporal. Interprete.

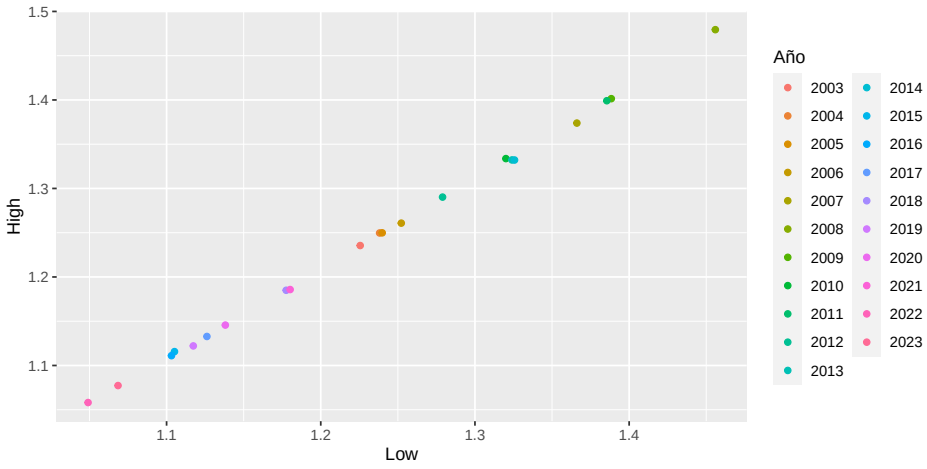
```
datos$Date = as.Date(datos$Date)
g = ggplot(data = datos[datos$Ano %in% 2020:2022,],
```

```
            aes(x = Date,
                y = Open)) +
  geom_point(size = 0.5) +
  theme(axis.text.x = element_blank()) +
  facet_wrap(vars(Ano),
             nrow = 1,
             scales = "free_x")
g
```



4. Grafique el valor el valor promedio de la tasa más alta versus la más baja del EUR/USD para cada uno de los años. Interprete.

```
g = ggplot(data = aggregate(x = datos, by = list(Año =
↪  datos$Ano), FUN = mean),
            aes(x = Low,
                y = High,
                color = Año)) +
  geom_point()
g
```


High
Low
1.5
1.4
1.3
1.2
1.1
1.1
1.2
1.3
1.4
Año
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

Unidad 2

Probabilidad y variables aleatorias.

2.1. Elementos de probabilidad

2.1.1. Espacio muestral

2.1.2. Función de probabilidad

2.2. Variable aleatoria

2.2.1. Función de probabilidad

2.2.2. Función de distribución

2.3. Variable aleatoria discreta

2.3.1. Uniforme

2.3.2. Bernoulli

2.3.3. Binomial

2.3.4. Poisson

2.4. Variable aleatoria continua

2.4.1. Uniforme

2.4.2. Exponencial

2.4.3. Normal

2.4.4. T - Student

2.4.5. Ji - Cuadrado

2.5. Esperanza

2.5.1. Variable aleatoria discreta

Unidad 3

Distribuciones muestrales y pruebas de hipótesis

Unidad 4

Intervalos de confianza

4.1. Una media

4.1.1. Bajo distribución normal

4.1.2. Asintótico

4.2. Diferencia de medias

4.3. Compraración de varianzas

4.4. Diferencia de proporciones

Bibliografía

Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2008). *Estadística para administración y economía*. Cengage Learning, México, 10a ed edition. OCLC: 893589801.

Brachman, R. J. and Levesque, H. J. (2004). *Knowledge representation and reasoning*. Morgan Kaufmann, Amsterdam ; Boston.

Devore, J. L. (2008). *Probability and statistics for engineering and the sciences*. Thomson/Brooks/Cole, Belmont, CA, 7th ed edition.

Healy, K. (2019). *Data Visualization: A Practical Introduction*. Princeton University Press.

Hintze, J. L. and Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Rowlingson, B. (2016). *Data Analysis with R*. Springer.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association*, 21(153):65–66.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.