Estadística I & Estadística Descriptiva

Unidad de Formación en Matemática y Estadística (UFME)

Coordinación de Estadística

Actualizado al 05-03-2023

Índice general

Presentación

La asignatura Estadística I & Estadística Descriptiva, es el primer curso estadístico de la carrera de Ingeniería Comercial e Ingeniería en Control de Gestión respectivamente. Estos son el primero de una serie de dos cursos introductorios de estadística, los cuales tienen un enfoque práctico con un fuerte énfasis en el estudio descriptivo de datos. Los cursos, se concentran en gráficos descriptivos, medidas de resumen, funciones de probabilidad, distribuciones muestrales, pruebas de hipótesis e intervalos de confianza. La metodología de aprendizaje se basa en clases interactivas - participativas, las que están basadas en el uso de R como programa estadístico para el análisis de datos.

Modalidad

La modalidad de trabajo consta de los siguientes elementos:

1. El **documento** web cuenta con el desarrollo de todos los tópicos de curso, además de ejemplificaciones y ejercicios.

2. En su mayoría, los ejemplos y ejercicios presentes en el documento hacen **uso de R** como programa de análisis estadístico. El desarrollo de los ejercicios por parte del estudiante serán en Google Colab R. Esta plataforma no cuenta con una opción de configuración interna para R, sin embargo, en el siguiente enlace se puede acceder a un documento con una configuración preestablecida para este lenguaje. El archivo que se genera se guardará automáticamente en la cuenta de Gmail predeterminada, por lo que se recomienda que dicha cuenta corresponda a la asociada a la UDP; otra opción en caso de no querer modificar su cuenta predeterminada, es descargar el archivo y cargarlo manualmente en la carpeta de Drive que estime conveniente. Los aspectos relacionados con el uso de Google Colab R serán abordados en el **Taller Introductorio**.

3. Se cuenta con **talleres de práctica**, lo cuales se desarrollarán en de Google Colab R. Estos talleres cuentan con tres elementos: ejercicios desarrollados, ejercicios propuestos para desarrollar en clases y ejercicios para el trabajo independiente del estudiante. Para estos últimos, **NO** habrá pauta, ya que se espera que el estudiante sea capaz de evaluar críticamente la solución obtenida.

4. El curso cuenta con **bibliografía** obligatoria y suplementaria:

    - (Obligatoria) *"Estadística para Administración y Economía"* (Anderson et al., 2008)
    - (Obligatoria,) *"Probabilidad y Estadística para Ingeniería y Ciencias"* (Devore, 2008)
    - (Complementaria) *"R Programming for Data Science"* (Peng, 2016)
    - (Complementaria) *"The R Software: Fundamentals of Programming and Statistical Analysis"* (de Micheaux et al., 2013)

- (Complementaria) *"ggplot2: Elegant Graphics for Data Analysis"* (Wickham, 2009)

Además, a lo largo del documento se añaden citas que refuerzan el contenido presentando. Al final de cada sección se encuentra el detalle de cada una de ellas.

5. Las **bases de datos** a utilizar en el curso se encuentran disponible en un repositorio web de libre acceso.

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 1.1. 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 1.2. 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.

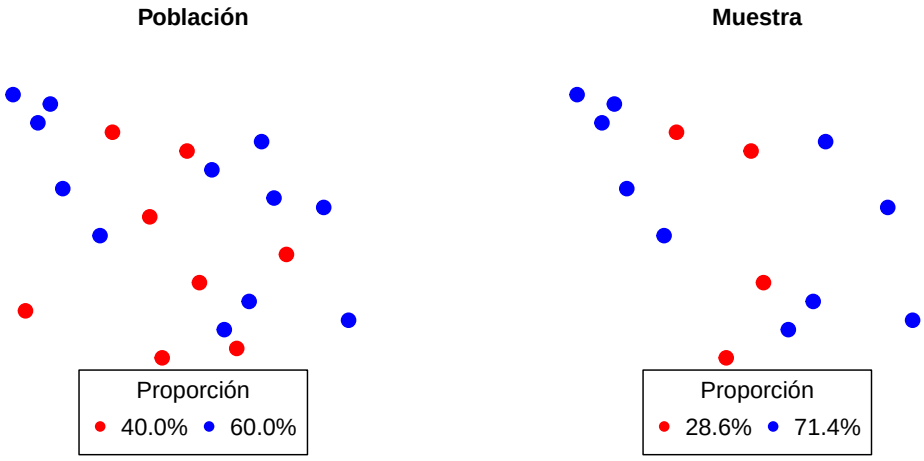


Figura 1.1: Parámetro y estadístico

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?

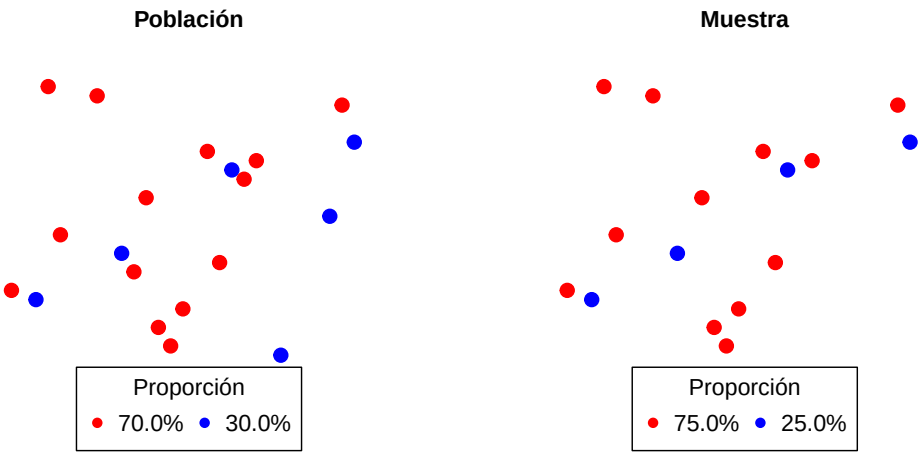


Figura 1.2: Estimador y estimación

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 reflejada en la figura ?

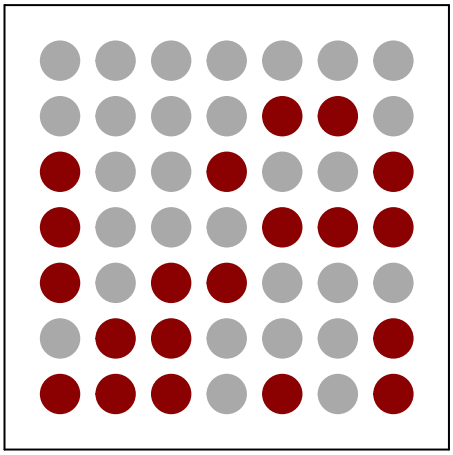


Figura 1.3: Población

Luego,

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

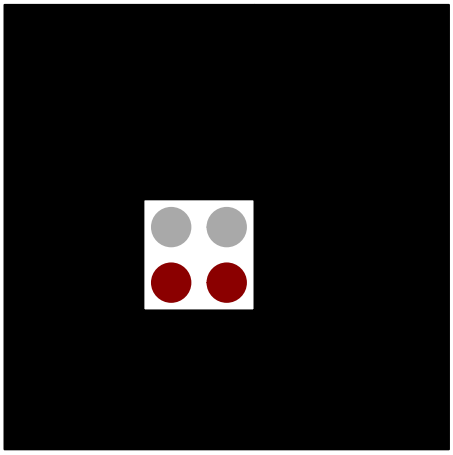


Figura 1.4: Muestra 1

¿Y ahora? (Figura 1.5**)**

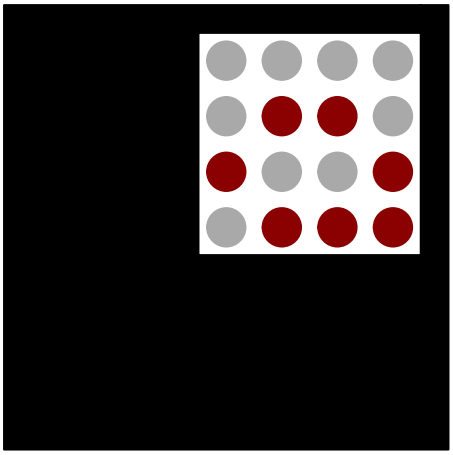


Figura 1.5: Muestra 2

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

1.1.8. Representatividad y sesgo de la muestra

- **Representatividad**

  Comúnmente se escucha hablar de que una muestra debe ser representativa respecto de la población, algo muy similar a lo presentado en la sección 1.1.7. Sin embargo, este concepto no tiene sustento matemático, ya que, para poder verificar que una muestra es representativa se debe conocer a toda la población (la característica de estudio), lo cual en la práctica no ocurre. Y en caso de que se conociesen todos los datos de la población, sería absurdo calcular la estimación de un parámetro, ya que podría calcularse directamente el valor del parámetro en cuestión.

- **Sesgo**

  Hay personas utilizan la siguiente frase "*la muestra está sesgada*", lo cual es incorrecto en su totalidad en estadística. El concepto de sesgo no es únicamente propio de la estadística, sin embargo, en esta área, corresponde a una propiedad de los estimadores. Se dice que un estimador es insesgado cuando el valor esperado de este es igual al parámetro. Y al igual que el concepto anterior, no es posible verificarlo en la práctica, aunque si tiene un sustento matemático por detrás.

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 mediana y la moda.

Media

Para un conjunto dado de números $x_1, x_2, x_3, \dots, x_n$, 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 $x_1, x_2, x_3, \dots, x_n$ 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 1.3.

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 1.4. 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 1.1.

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 1.5. 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 1.2. Obtener el rango de la tasa de apertura histórica del EUR/USD de la base de datos de la unidad.

```
# 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 1.6. Obtener el rango del resto de variables de la base de datos de la unidad (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 1.3. 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 1.7. 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 1.4. 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,
# 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 1.8. 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

Tabla 1.2: Notación de parámetros y estadísticos

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

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).

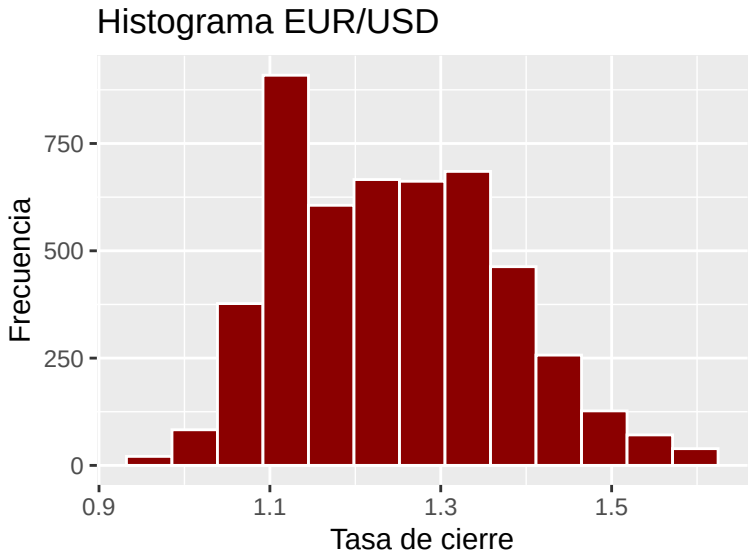
1.2.1. Histograma

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) # Librería de ggplot2
ggplot( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
       aes( # Comandos estéticos
         x = Close)) + # Eje X y variable asociada
  geom_histogram( # Objeto a graficar: histograma
    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( # Títulos
    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 1.9. 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.

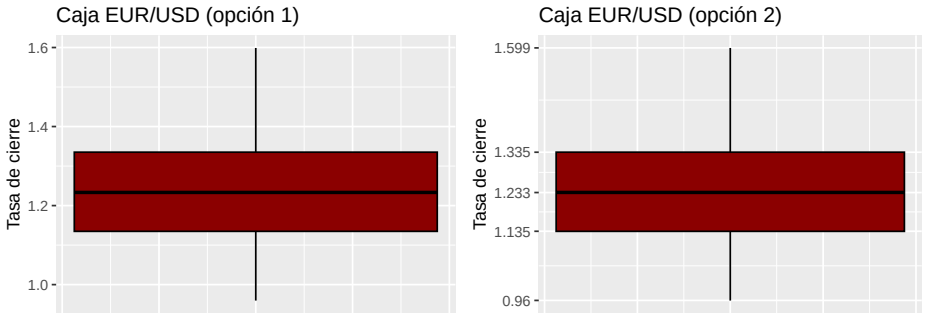
Los pasos para elaborar un gráfico de caja son los siguiente (Anderson et al., 2008, página 106):

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( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comandos estéticos
    y = Close)) + # Eje Y y variable asociada
  geom_boxplot( # Objeto a graficar: gráfico de caja
    color = "black", # Color del borde del gráfico
    fill = "darkred") + # Color de relleno del gráfico
  labs( # Títulos
    title = "Caja EUR/USD (opción 1)", # Título del gráfico
    x = "", # Título del eje X
    y = "Tasa de cierre") + # Título del eje Y
  theme( # Aspectos visuales del gráfico
    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( # Modificar el eje Y
    breaks = values, # Modificamos los puntos a considerar en el
     ↪  eje Y
    labels = values) + # Modificamos los valores mostrados en el
     ↪  eje Y
  labs( # Títulos
    title = "Caja EUR/USD (opción 2)") # Título del gráfico
library(gridExtra) # Librería para juntar gráficos de ggplot2
grid.arrange(g, # Gráfico
             g1, # Gráfico
             ncol = 2) # Despliegue en a dos columnas
```



Ejercicio 1.10. 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:

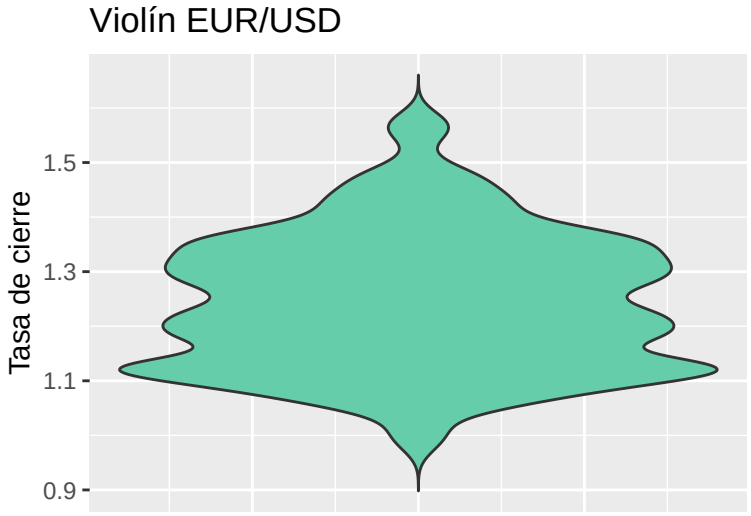
1. Dibujo de la traza de densidad: la traza de densidad se dibuja sobre el eje vertical en el gráfico de violín (“forma suavizada del histograma”).

2. Creación de la sección central simétrica: se crea una sección central simétrica que representa la mitad de la traza de densidad.

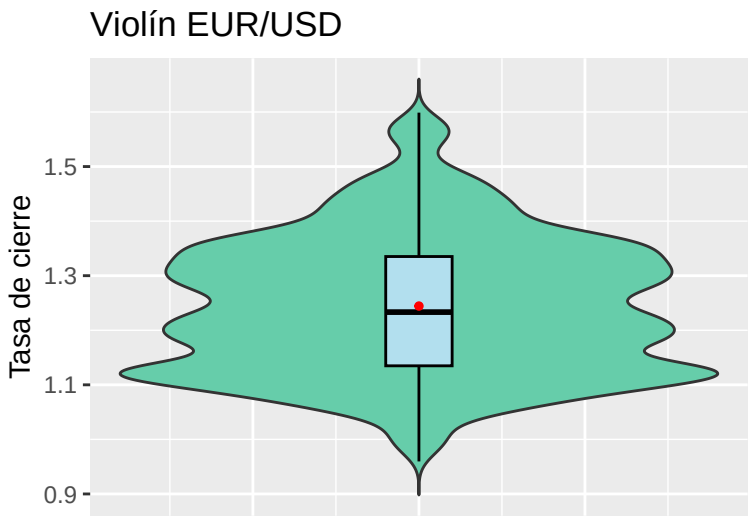
Adicionalmente, es común agregar un gráfico de caja junto al de violín con el fin de incorporar la visualización de las medidas de posición.

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( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comandos estéticos
    x = 1, # Se fija el valor horizontal del gráfico (a elección)
    y = Close)) + # Eje Y y variable asociada
  geom_violin( # Objeto a graficar: violín
    trim = F, # Modifica las terminaciones visuales superior e
     ↪  inferior
    fill = "aquamarine3") + # Color de relleno del gráfico
  labs( # Títulos
    title = "Violín EUR/USD", # Título del gráfico
    x = "", # Título del eje X
    y = "Tasa de cierre") + # Título del eje Y
  theme( # Aspectos visuales del gráfico
    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
```



```r
# Agregamos otros elementos al gráfico guardado
g + geom_boxplot( # Objeto a graficar: gráfico de caja
  width = 0.1, # Anchura proporcional del nuevo gráfico de caja
  color = "black", # Color de borde del gráfico
  fill = "lightblue2") + # Color de relleno del gráfico
  stat_summary( # Función para agregar información de resumen
    fun = mean, # Tipo de información: promedio
    geom = "point", # Forma visual
    size = 1, # Tamaño
    color = "red") # Color
```

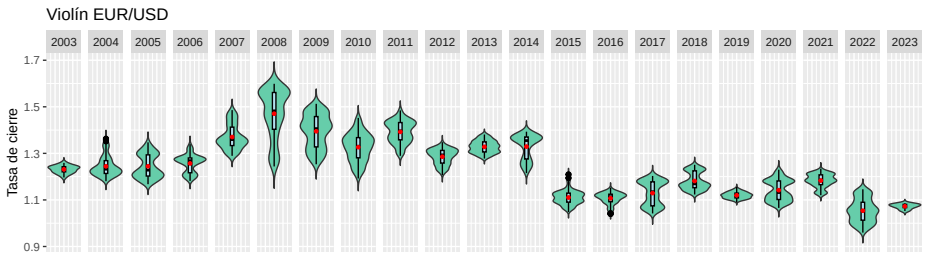


Ejercicio 1.11. Utilizando las variables **Low** y **Open** de la base de datos de la unidad, elabore un gráfico violín incluyendo un gráfico de caja y el promedio para cada una. Interprete.

Ejemplo 1.5. 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.

```r
# Extraemos el año de la variable Date, y la guardamos en un
 ↪  nueva columna
datos$Ano = substr(datos$Date, 1, 4)
ggplot( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comandos estéticos
    x = 1, # Se fija el valor horizontal del gráfico (a elección)
    y = Close)) + # Eje Y y variable asociada
  geom_violin( # Objeto a graficar: violín
```

```
    trim = F, # Modifica las terminaciones visuales superior e
     ↪  inferior
    fill = "aquamarine3") + # Color de relleno del gráfico
  geom_boxplot( # Objeto a graficar: gráfico de caja
    width = 0.1, # Anchura proporcional del nuevo gráfico de caja
    color = "black", # Color de borde del gráfico
    fill = "lightblue2") + # Color de relleno del gráfico
  stat_summary( # Función para agregar información de resumen
    fun = mean, # Tipo de información: promedio
    geom = "point", # Forma visual
    size = 1, # Tamaño
    color = "red") + # Color
  labs( # Títulos
    title = "Violín EUR/USD", # Título del gráfico
    x = "", # Título del eje X
    y = "Tasa de cierre") + # Título del eje Y
  theme( # Aspectos visuales del gráfico
    axis.ticks.x = element_blank(), # Elimina las regletas del
     ↪  eje X
    axis.text.x = element_blank()) + # Elimina los números del
     ↪  eje X
  facet_wrap( # Segregación del gráfico
    vars(Ano), # Variable que se utiliza para segregar el gráfico
    nrow = 1) # Disposición visual: una 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 1.12. 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.

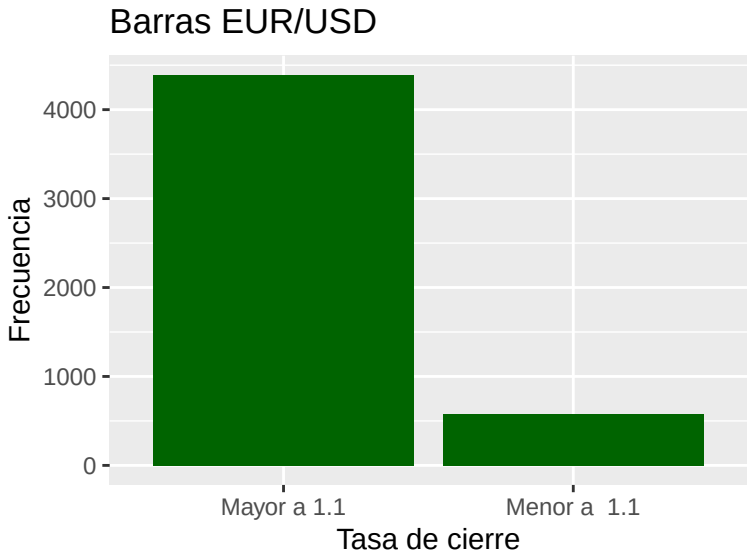
1.2.4. Barras

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:

```
# 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( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comandos estéticos
    x = Close2)) + # Eje Y y variable asociada
  geom_bar( # Objeto a graficar: gráfico de barras
    fill = "darkgreen") + # Color de relleno
  labs( # Títulos
    title = "Barras EUR/USD", # Título del gráfico
    x = "Tasa de cierre", # Título del eje X
    y = "Frecuencia") # Título del eje Y
```



Ejercicio 1.13. 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).

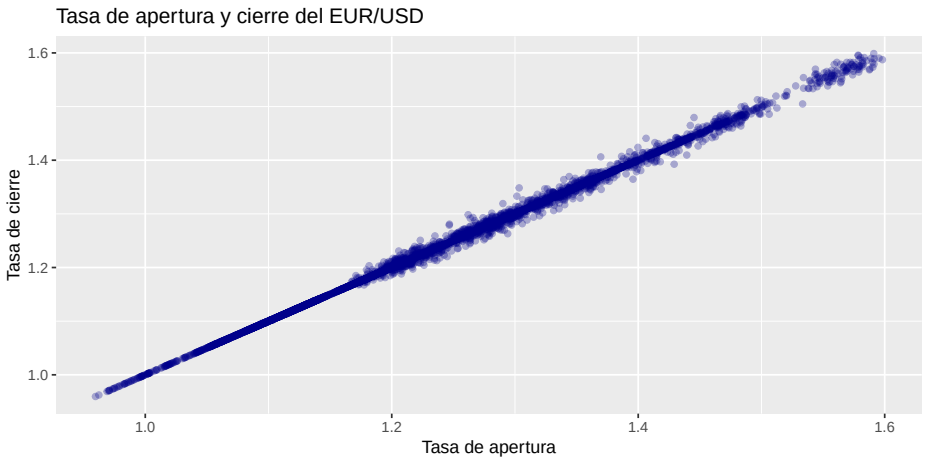
Los pasos para elaborar un gráfico de caja son los siguiente (Healy, 2019):

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( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comando estéticos
    x = Open, # Eje X y variable asociada
    y = Close)) + # Eje Y y variable asociada
  geom_point( # Objeto a graficar: Gráfico de dispersión
    color = "darkblue", # Color
    alpha = 0.3) + # Opacidad
  labs( # Títulos
    title = "Tasa de apertura y cierre del EUR/USD", # Título del
     ↪  gráfico
```

```
    x = "Tasa de apertura", # Título del eje X
    y = "Tasa de cierre") # Título del eje Y
```

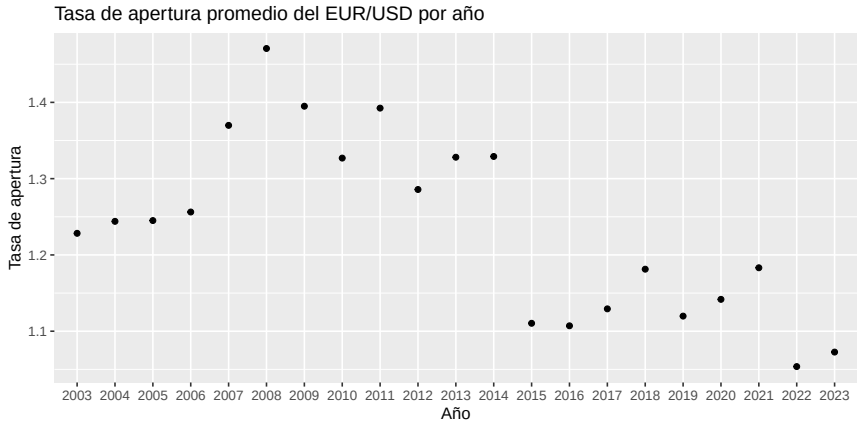


Ejercicio 1.14. Utilizando las variables **High** y **Low** de la base de datos de la unidad, elabore un gráfico de dispersión. Interprete.

Ejemplo 1.6.

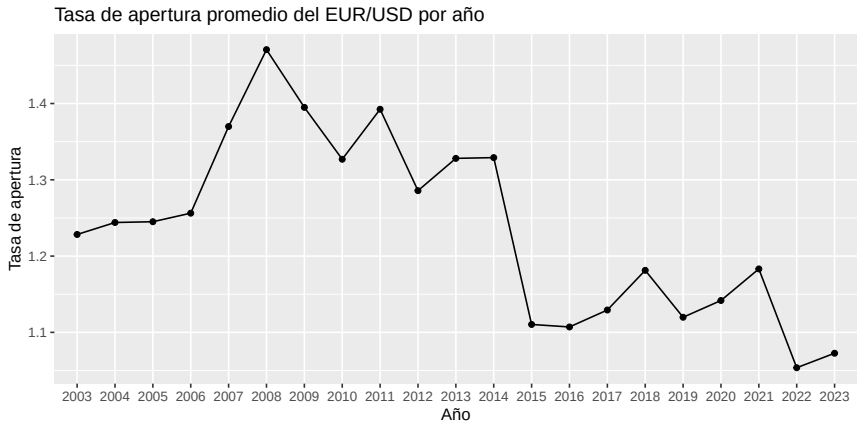
1. Es posible utilizar el gráfico de dispersión con variables que en su naturaleza son discretas. En este sentido, elabore un gráfico de dispersión entre el año de medición y el valor promedio de tasa de apertura del EUR/USD (guarde el gráfico en una variable).

```
g = ggplot( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comando estéticos
x = Ano, # Eje X y variable asociada
y = Open, # Eje Y y variable asociada
group = 1)) + # Comando únicamente necesario para la
 ↪  pregunta 2
  geom_point( # Objeto a graficar
stat = "summary", # Tipo de información a graficar: resumen
fun = "mean") + # Tipo de resumen: promedio de la variable Y
  labs( # Títulos
title = "Tasa de apertura promedio del EUR/USD por año", #
 ↪  Título del gráfico
x = "Año", # Título del eje X
y = "Tasa de apertura") # Título del eje Y
g # Desplegamos el gráfico guardado
```



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

```
g = g + # Añadimos otro gráfico
  geom_line( # Objeto a graficar: lineas
stat = "summary", # Tipo de información a graficar: resumen
fun = "mean") # Tipo de resumen: promedio de la variable Y
g # Desplegamos el gráfico guardado
```

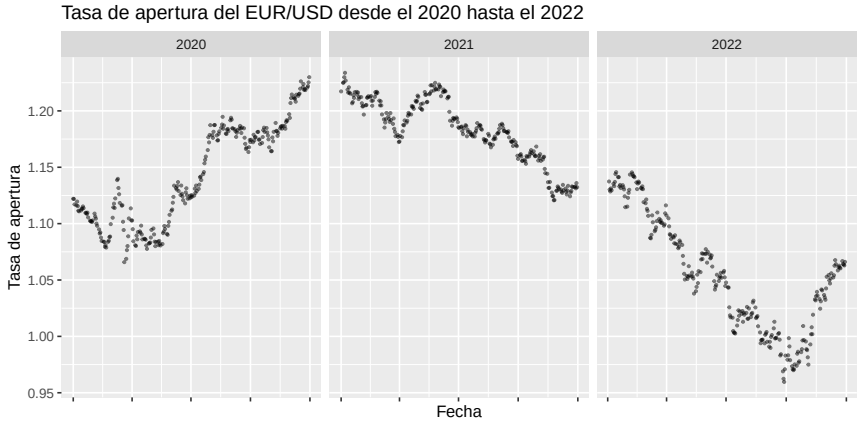


Hasta el 2008 la tasa promedio de apertura estuvo en alza, posteriormente, la tasa decayó a un valor inferior a 1.2.

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

```
datos$Date = as.Date(datos$Date) # Fechas en formato fecha
 ↪  de R
g = ggplot( # Ambiente gráfico
  data = datos[datos$Ano %in% 2020:2022,], # Datos de los
   ↪  años 2020 al 2022
```

```
  aes( # Comando estéticos
x = Date, # Comandos estéticos: Eje X y variable asociada
y = Open)) + # Eje Y y variable asociada
  geom_point( # Objeto a graficar
alpha = 0.5, # Opacidad
size = 0.6) + # Tamaño
  theme( # Aspectos visuales del gráfico
axis.text.x = element_blank()) + # Eliminamos el texto del
 ↪  eje X
  facet_wrap( # Segregación del gráfico
vars(Ano), # Variable que se utiliza para segregar el
 ↪  gráfico
nrow = 1, # Disposición visual: una fila
scales = "free_x") + # La escala del eje X es independiente
 ↪  para gráfico
  labs( # Títulos
title = "Tasa de apertura del EUR/USD desde el 2020 hasta el
 ↪  2022", # Título del gráfico
x = "Fecha", # Título del eje X
y = "Tasa de apertura") # Título del eje Y
g # Desplegamos el gráfico guardado
```

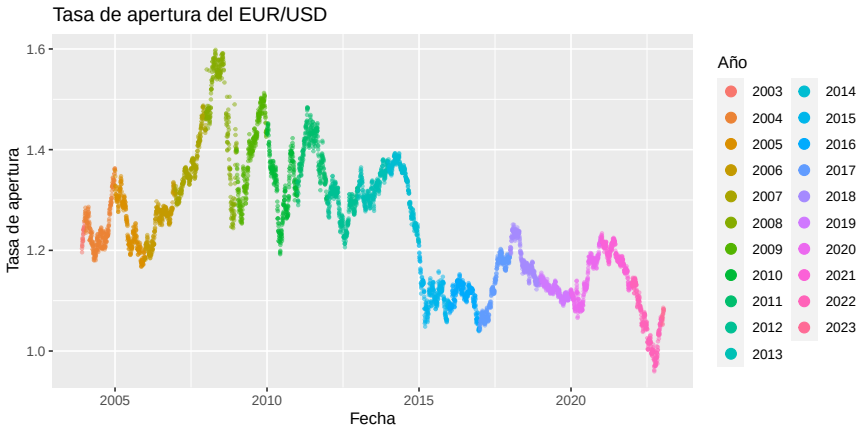


Durante los 3 años consecutivos, se observa que únicamente en el 2020 la tendencia de la tasa de apertura es al alza, mientra que para los otros dos años hubo un decaimiento en el valor de esta.

4. Grafique el valor de la tasa de apertura del EUR/USD diferenciando por año. Interprete.

```
g = ggplot( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
```

```
  aes( # Comando estéticos
x = Date, # Comandos estéticos: Eje X y variable asociada
y = Open, # Eje Y y variable asociada
color = Ano)) + # Color según el año
  geom_point( # Objeto a graficar
alpha = 0.5, # Opacidad
size = 0.7) + # Tamaño
  labs( # Títulos
color = "Año", # Título de la leyenda
title = "Tasa de apertura del EUR/USD", # Título del gráfico
x = "Fecha", # Título del eje X
y = "Tasa de apertura") + # Título del eje Y
  guides( # Edición de escalas
color = guide_legend( # Escala de color de la leyenda
  override.aes = list( # Comando estéticos asociados
    alpha = 1, # Opacidad de los puntos
    size = 3))) # Tamaño de los puntos
g # Desplegamos el gráfico guardado
```

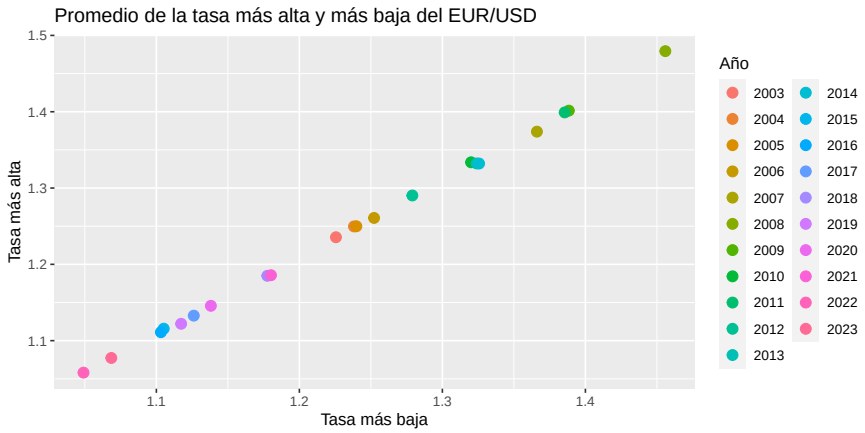


Al observar la evolución histórica de la tasa de apertura diferenciada por año, se aprecia que el periodo 2008 - 2010 es aquel con predominancia de valores más altos. Por otro lado, desde el 2016, se registraron pro primera vez valores menores a 1.1. En años posteriores, no ha observado que la tasa supere los 1.3 puntos.

5. 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( # Ambiente gráfico
  data = aggregate( # Función para resumir una base de datos
    ↪  (crea una nueva base)
x = datos, # Indicar la base de datos a modificar
```

```
by = list(Año = datos$Ano), # Variable por la cual se genera
 ↪  el resumen, y nombre asignado
FUN = mean), # Resumen que se genera en las columnas de la
 ↪  base de datos
  aes( # Comandos estéticos
x = Low, # Eje X y variable asociada
y = High, # Eje Y y variable asociada
color = Año)) + # Color según variable
  geom_point( # Objeto a graficar: gráfico de dispersión
size = 3) + # Tamaño de los puntos
  labs( # Títulos
title = "Promedio de la tasa más alta y más baja del
 ↪  EUR/USD", # Título del gráfico
x = "Tasa más baja", # Título el eje X
y = "Tasa más alta") + # Título el eje Y
  guides( # Edición de escalas
color = guide_legend( # Escala de color de la leyenda
  override.aes = list( # Comando estéticos asociados
    size = 3))) # Tamaño de los puntos
g # Desplegamos el gráfico guardado
```



El gráfico da cuenta de una tendencia positiva entre la tasa promedio más baja y más alta registrada en los diversos años, siendo el año 2008 aquel que destaca con las tasas promedio más altas.

Ejercicio 1.15.

1. Elabore un gráfico de dispersión entre el año de medición y el valor promedio de tasa de cierre del EUR/USD (guarde el gráfico en una variable).
2. Añadir al gráfico un formato de líneas entre los puntos. Interprete.
3. Grafique el valor de la tasa de cierre del EUR/USD desde el 2010 hasta el 2013 separadamente. Interprete.

4. Grafique el valor de la tasa de cierre del EUR/USD diferenciando por año. Interprete.
5. Grafique el valor el valor promedio de la tasa de cierre versus la más baja del EUR/USD para cada uno de los años. Interprete.

Unidad 2

Probabilidad y variables aleatorias

2.1. Elementos de probabilidad

Los elementos de probabilidad son los conceptos fundamentales que se utilizan en la teoría de la probabilidad para describir y analizar eventos aleatorios. Algunos de ellos son: espacio muestral, eventos, función de probabilidad, variable aleatoria, distribución de probabilidad, entre otros.

Estos elementos son esenciales para el estudio de la probabilidad y su aplicación en la estadística y en muchas áreas de la ciencia, incluyendo la economía, la biología, la física, entre otras.

2.1.1. Experimento y Espacio muestral

En el contexto de la probabilidad, un **experimento** es definido como un proceso que genera resultados definidos. Y en cada una de las repeticiones del experimento, habrá uno y solo uno de los posibles resultados experimentales (Anderson et al., 2008, página 143).

Ejemplo 2.1.

Tabla 2.1: Experimentos y resultados

Experimento	Resultado experimental
Lanzar una moneda	Cara, cruz
Tomar una pieza para inspeccionarla	Con defecto, sin defecto
Realizar una llamada de ventas	Hay compra, no hay compra
Lanzar un dado	1, 2, 3, 4, 5, 6
Jugar un partido de fútbol	Ganar, perder, empatar

Al especificar todos los resultados experimentales posibles, está definiendo el **espacio muestral** de un experimento. En otras palabras, el espacio muestral de un experimentos es el conjunto de todos los resultados experimentales. Se usa la letra omega mayúscula (Ω) para referirnos a este conjunto. Un elemento genérico de Ω se denota como ω.

Ejemplo 2.2. Conduciendo hacia su trabajo, una persona debe pasar por tres semáforos. En cada cruce la persona puede detenerse (D) o continuar (C), de acuerdo con el color de la luz. ¿Cuál es el espacio muestral del experimento?

$$\Omega = \{CCC, DDD, CCD, CDD, CDC, DCD, DDC, DDD\}$$

Ejercicio 2.1. Un fabricante de ropa deportiva produce pantalones deportivos en dos colores (azul y gris) y en cuatro tamaños diferentes (pequeño, mediano, grande y extra grande). ¿Cuál es el espacio muestral del experimento de elegir al azar un pantalón deportivo de la línea de producción de la empresa?

Ejercicio 2.2. Un restaurante ofrece tres opciones de menú para el almuerzo: menú A, menú B y menú C. Además, cada menú se puede pedir con carne o con pescado. ¿Cuál es el espacio muestral del experimento de elegir al azar un menú para el almuerzo en este restaurante?

Ejercicio 2.3. Una compañía de seguros de autos ofrece pólizas de seguro con dos niveles de cobertura (básico y completo) y dos tipos de franquicia (alta y baja). ¿Cuál es el espacio muestral del experimento de elegir al azar una póliza de seguro de auto de la compañía?

2.1.2. Eventos aleatorios

En principio, un **evento aleatorio** (o simplemente evento) es algún subconjunto del espacio muestral Ω. Los eventos se anotan con una letra mayúscula a elección (Anderson et al., 2008, página 153).

A modo de ejemplo, consideren el experimento de lanzar un dado de 6 caras.

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Luego, el evento correspondiente a obtener un número par estaá dado por:

A: obtener un número par $\rightarrow A = \{2, 4, 6\}$

Ejercicio 2.4. Experimento aleatorio: Lanzamiento de un dado. Evento aleatorio: Obtener un número par. ¿Cuál es el conjunto correspondiente al evento aleatorio?

Ejercicio 2.5. Experimento aleatorio: Elegir una carta al azar de una baraja de 52 cartas. Evento aleatorio: Obtener una carta roja. ¿Cuál es el conjunto correspondiente al evento aleatorio?

Ejercicio 2.6. Experimento aleatorio: Lanzar dos monedas. Evento aleatorio: Obtener dos caras. ¿Cuál es el conjunto correspondiente al evento aleatorio?

Ejercicio 2.7. Experimento aleatorio: Elegir un estudiante al azar de una clase de 30 estudiantes. Evento aleatorio: Elegir a un estudiante que tenga una altura superior a 1,75 metros. ¿Cuál es el conjunto correspondiente al evento aleatorio?

Ejercicio 2.8. Experimento aleatorio: Lanzar un dardo a una diana circular. Evento aleatorio: Obtener un lanzamiento dentro del círculo exterior de la diana. ¿Cuál es el conjunto correspondiente al evento aleatorio?

2.1.3. Probabilidad de un evento

El concepto de probabilidad está asociado a la ocurrencia de un evento Sin embargo, el número que determina que tan factible es que dicho evento ocurra puede ser difícil de calcular. En este aspecto, como introducción, se hará uso de la definición clásica de probabilidad:

$$\textit{Probabilidad de que ocurra un evento} = \frac{\text{Casos favorables}}{\text{Casos totales}}$$

Por ejemplo, la probabilidad de obtener un número par al lanzar un dado una vez es:

A: obtener un número par. $\rightarrow A = \{2, 4, 6\}$

$$P(A) = \frac{\text{Casos favorables}}{\text{Casos totales}} = \frac{\{2, 4, 6\}}{\{1, 2, 3, 4, 5, 6\}} = \frac{3}{6} = \frac{1}{2}$$

Nota: La probabilidad de cualquier evento siempre estará entre 0 y 1. Los casos extremos suceden cuando los casos favorables son inexistentes o son la totalidad de casos posibles respectivamente.

Ejercicio 2.9. En una tienda de ropa hay 10 camisas rojas, 15 camisas azules y 20 camisas verdes. ¿Cuál es la probabilidad de que al escoger una camisa al azar, sea de color verde?

Ejercicio 2.10. En un mercado hay 200 vendedores, de los cuales el 70 % son hombres y el 30 % son mujeres. Si se elige al azar un vendedor, ¿cuál es la probabilidad de que sea mujer?

Propiedades

A continuación se mencionan algunas propiedades relacionadas con probabilidades (Anderson et al., 2008, página 157).

1. **Complemento de un evento:** Dado un evento A, el complemento de A se define como el evento que consta de todos los casos muestrales que **no** están en A, y se denota por A^c. Por ejemplo, si consideramos el experimento de lanzar el dado, y el evento de obtener un número par (A), entonces, el complemento corresponde a obtener un número que no sea par (A^c). De lo anterior se tiene que

$$P(A) + P(A^c) = 1 \tag{2.1}$$

   Ejemplo 2.3. Considere el caso de un administrador de ventas que, después de revisar los informes de ventas, encuentra que el 80 % de los contactos con clientes nuevos no producen ninguna venta. Si A denota el evento **hubo venta**, entonces A^c corresponde al evento de **no hubo venta**. Si el administrador tiene que $P(A^c) = 0.8$, mediante la ecuación (2.1) se ve que

$$P(A) = 1 - P(A^c) = 1 - 0.8 = 0.2$$

   La conclusión es que la probabilidad de una venta en el contacto con un cliente nuevo es de 0.2.

2. **Unión de dos eventos:** La unión de dos eventos A y B es el evento que contiene todos los casos muestrales que pertenecen a A o B o ambos. La unión se denota $A \cup B$.

3. **Intersección de dos eventos:** Dados dos eventos A y B, la intersección de A y B es el evento que contiene los casos muestrales que pertenecen tanto a A como a B. La intersección se denota $A \cap B$.

4. **Ley de la adición:** La ley de la adición proporciona una manera de calcular la probabilidad de que ocurra el evento A o el evento B o ambos. En otras palabras, esta ley se emplea para calcular la probabilidad de la unión de dos eventos. La ley de la adición se expresa de la siguiente manera.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.2}$$

   Ejemplo 2.4. Considere el caso de una pequeña empresa de ensamble en la que hay 50 empleados. Se espera que todos los trabajadores terminen su trabajo a tiempo y que pase la inspección final. A veces, alguno de los empleados no satisface el estándar de desempeño, ya sea porque no termina a tiempo su trabajo o porque no ensambla bien una pieza. Al final

del periodo de evaluación del desempeño, el jefe de producción encuentra que 5 de los 50 trabajadores no terminaron su trabajo a tiempo. 6 de los 50 trabajadores ensamblaron mal una pieza y 2 de los 50 trabajadores no terminaron su trabajo a tiempo y armaron mal una pieza.

Sea

$$L : \text{No se termino el trabajo a tiempo}$$
$$D : \text{Se armó mal la pieza}$$

La información de las frecuencias relativas lleva a las probabilidades siguientes.

$$P(L) = \frac{5}{50} = 0.1$$
$$P(D) = \frac{6}{50} = 0.12$$
$$P(L \cap D) = \frac{2}{50} = 0.04$$

Después de analizar los datos del desempeño, el jefe de producción decide dar una calificación baja al desempeño de los trabajadores que no terminaron a tiempo su trabajo o que armaron mal alguna pieza; por tanto, el evento de interés es $L \cup D$. ¿Cuál es la probabilidad de que el jefe de producción de a un trabajador una calificación baja de desempeño?

Esta pregunta se refiere a la unión de dos eventos. En concreto, se desea hallar $P(L \cup D)$, usando la ecuación (2.2) se tiene

$$P(L \cup D) = P(L) + P(D) - P(L \cap D)$$

Como conoce las tres posibilidades del lado derecho de la expresión, se tiene

$$P(L \cup D) = 0.1 + 0.12 - 0.04 = 0.18$$

Estos cálculos indican que la probabilidad de que un empleado elegido al azar obtenga una calificación baja por su desempeño es 0.18.

Ejemplo 2.5. Considere un estudio reciente efectuado por el director de personal de una empresa de software. En el estudio encontró que el 30 % de los empleados que se van de la empresa antes de dos años, lo hacen por estar insatisfechos con el salario, 20 % se van de la empresa por estar descontentos con el trabajo y 12 % por estar insatisfechos con las dos cosas, el salio y el trabajo. ¿Cuál es la probabilidad de que un empleado que se

vaya de la empresa en menos de dos años lo haga por estar insatisfecho con el salario, con el trabajo o con las dos cosas?

Sea

$$S : \text{El empleado se va de la empresa por insatisfacción con el salario}$$
$$W : \text{El empleado se va de la empresa por insatisfacción con el trabajo}$$

Se tiene $P(S) = 0.3$, $P(W) = 0.2$ y $P(S \cap W) = 0.12$. Al aplicar la ecuación (2.2), de la ley de la adición, se tiene

$$P(S \cup W) = P(S) + P(W) - P(S \cap W) = 0.3 + 0.2 - 0.12 = 0.38$$

Así, la probabilidad de que un empleado se vaya de la empresa por el salario o por el trabajo es 0.38.

5. **Eventos mutuamente excluyentes:** Se dice que dos eventos son mutuamente excluyentes si, cuando un evento ocurre, el otro no puede ocurrir. Por lo tanto, para que A y B sean mutuamente excluyentes, se requiere que su intersección sea nula, es decir,

$$\text{Si } A \cap B = \emptyset, \text{ entonces, } P(A \cap B) = 0. \tag{2.3}$$

6. **Ley de la adición para eventos mutuamente excluyentes:** En caso de que se cumplan las condiciones mencionadas en la ecuación (2.3), se tiene el siguiente resultado para la ecuación (2.2).

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - 0 \\ P(A \cup B) &= P(A) + P(B) \end{aligned} \tag{2.4}$$

Ejercicio 2.11. Suponga que tiene un espacio muestral con cinco resultados experimentales que son igualmente posibles: E_1, E_2, E_3, E_4 y E_5. Sean

$$\begin{aligned} A &= \{E_1, E_2\} \\ B &= \{E_3, E_4\} \\ C &= \{E_2, E_3, E_5\} \end{aligned}$$

a. Calcular $P(A)$, $P(B)$ y $P(C)$.

b. Calcular $P(A \cup B)$. ¿A y B son mutuamente excluyentes?
c. Determinar A^c, C^c, y calcular $P(A^c)$, $P(C^c)$.
d. Determinar $A \cup B^c$, y calcular $P(A \cup B^c)$.
e. Calcular $P(B \cup C)$.

Ejercicio 2.12. Datos sobre las 30 principales acciones y fondos balanceados proporcionan los rendimientos porcentuales anuales y a 5 años para el periodo que termina el 31 de marzo de 2000 (*The Wall Street Journal*, 10 de abril de 2000). Suponga que considera altos un rendimiento anual arriba de 50 % y un rendimiento a cinco años arriba de 300 % y cinco de los fondos tienen tanto un rendimiento anual arriba de 50 % como un rendimiento a cinco años arriba de 300 %.

a. ¿Cuál es la probabilidad de un rendimiento anual alto y cuál es la probabilidad de un rendimiento a cinco años alto?
b. ¿Cuál es la probabilidad de ambos, un rendimiento anual alto y un rendimiento a cinco años alto?
c. ¿Cuál es la probabilidad de que no haya un rendimiento anual alto ni un rendimiento a cinco años alto?

Ejercicio 2.13. La oficina de Censos de Estados Unidos cuenta con datos sobre la cantidad de adultos jóvenes entre 18 y 24 años, que viven en casa de sus padres. Sea

$$M = \text{Aldulto joven que vive en casa de sus padres}$$
$$F = \text{Aldulta joven que vive en casa de sus padres}$$

Si toma al azar un adulto joven y una adulta joven, los datos de dicha oficina permiten concluir que $P(M) = 0.56$ y $P(F) = 0.42$. La probabilidad de que ambos vivan encasa de sus padres es 0.24.

a. ¿Cuál es la probabilidad de que al menos uno de dos adultos jóvenes seleccionados viva en casa de sus padres?
b. ¿Cuál es la probabilidad de que los dos adultos jóvenes seleccionados vivan en casa de sus padres?

2.2. Variable aleatoria

Una variable aleatoria proporciona un medio para describir los resultados experimentales utilizando valores numéricos, es decir, una variable aleatoria asocia un valor numérico a cada uno de los resultados experimentales. Una variable aleatoria puede ser *discreta* o *continua*, depende del tipo de valores numéricos que asuma. (Anderson et al., 2008, página 187)

- Una variable aleatoria se denomina **discreta** si asume un número finito de valores o una sucesión infinita de valores tales como $0, 1, 2, \ldots$. Consideremos el siguiente experimento como ejemplo: un contador presenta el

examen para certificarse como contador público. El examen tiene cuatro partes. Defina una variable aleatoria X como $X =$ *número de partes del examen aprobadas*. Esta es una variable aleatoria discreta porque puede tomar el número finito de valores $0, 1, 2, 3$ o 4. Otros ejemplos se pueden apreciar en la tabla 2.2.

Tabla 2.2: Ejemplos de variables aleatorias discretas

Experimento	Variable aleatoria (X)	Valores posibles para la variable aleatoria
Llamar a cinco clientes	Número de clientes que hacen un pedido	0,1,2,3,4,5
Inspeccionar un envío de 50 radios	Número de radios que tienen algún defecto	0,1,2,...,49,50
Hacerse cargo de un restaurante durante el día	Número de clientes	0,1,2,3,...
Vender un automóvil	Sexo del cliente	0 si el hombre, 1 si es mujer

- Una variable aleatoria se denomina **continua** si puede tomar cualquier valor numéricos dentro de un intervalo. Los resultados experimentales basados en escalas de medición tales como tiempo, peso, distancia y temperatura puede ser descritos por variables aleatorias continuas. Consideremos el siguiente experimento como ejemplo: observar las llamadas telefónicas que llegan a la oficina de atención de una importante empresa de seguros. La variable aleatoria que interesa es $X =$ *tiempo en minutos entre dos llamadas consecutivas*. Esta variable aleatoria puede tomar cualquier valor en el intervalo $[0, \infty)$. En efecto, x puede tomar un número infinito de valores, entre los cuales se encuentra valores como 1.25 minutos 3.4562 minutos, 4.33333 minutos, etc. En la tabla 2.3 aparecen otros ejemplos de variables aleatorias continuas.

Tabla 2.3: Ejemplos de variables aleatorias discretas

Experimento	Variable aleatoria (X)	Valores posibles para la variable aleatoria
Operar un banco	Tiempo en minutos entre la llegada de los clientes	$x \geq 0$
Llenar una lata de bebida (máximo 12.1 onzas)	Cantidad de onzas	$0 \leq x \leq 12.1$
Contruir una biblioteca	Porcentaje del proyecto terminado en seis meses	$0 \leq x \leq 100$
Probar un proceso químico nuevo	Temperatura a la que tiene lugar la reacción deseada (mín. 150 grados F, máx. 212 grados F)	$150 \leq x \leq 212$

Ejemplo 2.6. A continuación se da una serie de experimentos y su variable

aleatoria correspondiente. En cada caso determine qué valores toma la variable aleatoria y diga si se trata de una variable aleatoria discreta o continua.

	Experimento	Variable aleatoria (X)
a.	Hacer un examen con 20 preguntas	Número de preguntas contestadas correctamente
b.	Obervar los automóviles que llegan a una caseta de peaje en 1 hora	Número de automóviles que llegan a la casea de peaje
c.	Revisar 50 declaraciones de impuestos	Número de declaracionex que tienen algún error
d.	Observar trabar a un empleado	Número de horas no productivas en una jornada de 8 horas
e.	Pesar un envío	Número de libras

2.3. Variables aleatorias discretas (v.a.d)

La distribución de probabilidad de una variable aleatoria discreta describe como se distribuyen las probabilidades entre los valores de la variable aleatoria. En el caso de una variable aleatoria discreta x, la distribución de probabilidad está definida por una función de probabilidad o también llamada **función de masa de probabilidad** (fmp) (Devore, 2008, página 90).

2.3.1. Función de masa de probabilidad

Consideremos el siguiente ejemplo, una empresa acaba de adquirir cuatro impresoras láser y sea X el número entre estas que requieres servicio durante el periodo de garantía. Los posibles valores de X son entonces 0, 1, 2, 3 y 4. La distribución de probabilidad diría cómo está subdividida la probabilidad de 1 entre los cinco posibles valores: cuánta probabilidad está asociada con el valor 0 de X, cuánta está adjudicada con 1 de X y así sucesivamente. Se utiliza la siguiente notación para las probabilidades:

$$p(0) = \text{la probabilidad del valor 0 de } X = P(X = 0)$$

$$p(1) = \text{la probabilidad del valor 1 de } X = P(X = 1)$$

y así sucesivamente. En general, $p(x)$ denotará la probabilidad asignada al valor de x.

Ejemplo 2.7. Una cierta gasolinera tiene seis bombas. Sea X el número de bombas que están bajo servicio a una hora particular del día. Suponga que la distribución de probabilidad de X es como se detalla en la siguiente tabla;

la primera fila de la tabla contiene los posibles valores de X y la segunda la probabilidad de dicho valor.

x	0	1	2	3	4	5	6
$p(x)$	0.05	0.1	0.15	0.25	0.2	0.15	0.1

Ahora, utilizando propiedades de probabilidad elemental (revisar las propiedades mencionadas en la sección 2.1.3) es posible calcular otras probabilidades de interés. Por ejemplo, la probabilidad de que cuando dos gasolineras estén en servicio es

$$P(X \leq 2) = P(X = 0 \text{ o } 1 \text{ o } 2) = p(0) + p(1) + p(2) = 0.05 + 0.1 + 0.15 = 0.3$$

Por otro lado, la probabilidad de que entre 2 y 5 gasolineras inclusive estén en servicio es

$$P(2 \leq X \leq 4) = P(X = 2 \text{ o } 3 \text{ o } 4) = p(2) + p(3) + p(4) = 0.15 + 0.25 + 0.2 = 0.6$$

La figura 2.1 está reproducida mediante R, con la finalidad de visualizar la función de masa asociada al ejemplo.

```
df = data.frame("x" = 0:6, # Valores de X
                "p" = c(0.05,0.10,0.15,0.25,0.20,0.15,0.10)) #
                 ↪ Probabilidades asociadas
ggplot(
  data = df,
  aes(x = x,
      y = p)) +
  geom_point() +
  geom_segment(
    aes(x = 0:6,
        y = rep(0,7),
        xend = 0:6,
        yend = p)) +
  labs(
    title = "Probabilidades de cada valor",
    x = "Valores del experimento (x)",
    y = "Probabilidades")
```

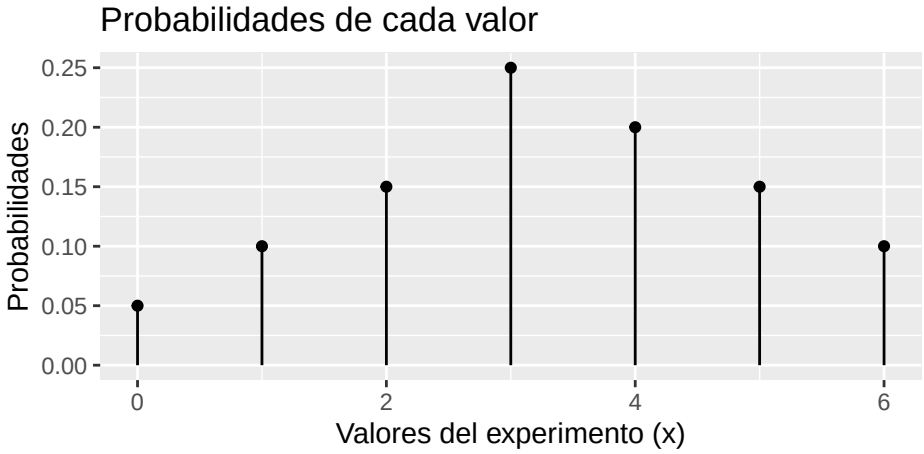


Figura 2.1: Función de masa

Cabe mencionar que cualquier función de masa de probabilidad requiere cumplir las siguientes condiciones

1. $p(x) > 0, \forall x \in X$
2. $\sum_{\text{todas las x posibles}} p(x) = 1$

para que se válida.

Ejercicio 2.14. Seis lotes de componentes están listos para ser enviados por un proveedor. El número de componentes defectuosos en cada lote es como sigue:

Lote	1	2	3	4	5	6
Número de defectuosos	0	2	0	1	2	0

Uno de estos lotes tiene que ser seleccionado al azar para ser enviado a un cliente particular. Sea X el número de defectuosos en el lote seleccionado. Los tres posibles valores de X son 0, 1 y 2.

a. Determine los la probabilidad para cada uno de los valores de X. Interprete.
b. Verifique las condiciones de la función de masa de probabilidad asociada al experimento.
c. Grafique la función de masa asociada.

Ejercicio 2.15. Una empresa de ventas en línea dispone de seis líneas telefónicas. Sea X el número de líneas en uso en un tiempo epsecificado. Suponga que la función de masa de probabilidad de X es la que se da en la tabla adjunta.

x	0	1	2	3	4	5	6
$p(x)$	0.1	0.15	0.2	0.25	0.2	0.06	0.04

Grafique la función de masa asociada, y luego calcule la probabilidad de cada uno de los siguientes eventos.

a. Cuando mucho tres líneas están en uso.
b. Menos de tres líneas están en uso.
c. Por o menos tres líneas están en uso.
d. entre dos y cinco líneas, inclusive, están en uso.
e. Entre dos y cuatro líneas, inclusive, no está en uso.
f. Por lo menos cuatro líneas no están en uso.

2.3.2. Función de distribución acumulada

Para algún valor fijo de x, a menudo se desea calcular la probabilidad de que el valor observado de X sea cuando mucho x ($X \leq x$). Por ejemplo, consideremos la siguiente función de masa.

$$P(X = x) = p(x) = \begin{cases} 0.5 & x = 0 \\ 0.167 & x = 1 \\ 0.333 & x = 2 \\ 0 & \text{en otro caso} \end{cases}$$

La probabilidad de que X sea cuando mucho de 1 es entonces

$$P(X \leq 1) = p(0) + p(1) = 0.5 + 0.167 = 0.667$$

Asimismo,

$$P(X \leq 0) = P(X = 0) = 0.5.$$

La **función de distribución acumulada** (fda) $F(x)$ de una variable aleatoria discreta X con función de masa de probabilidad $P(X = x)$ se define como

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y) \tag{2.5}$$

Para cualquier número x, $F(X)$ es la probabilidad de que el valor observado de X será cuando mucho (como máximo) x. (Devore, 2008, página 95)

Ejemplo 2.8. Consideremos un grupo de cinco donadores de sangre potenciales, a, b, c, d y e, de los cuales solo a y b tienen sangre tipo O+. Se determinará en orden aleatorio el tipo de sangre con cinco muestra, una de cada individuo hasta

que identifique un individuo O+. Sea la variable aleatoria Y = *el número de exámenes de sangre para identificar un individuo O+*. Entonces la función de masa de probabilidad de Y es

y	1	2	3	4
$p(y)$	0.4	0.3	0.2	0.1

Para determinar la función de distribución acumulada $F(Y)$, lo primero es determinar el valor de $F(Y)$ para cada uno de los valores posibles del conjunto $(1, 2, 3, 4)$:

$$
\begin{aligned}
F(1) &= P(Y \leq 1) = P(Y = 1) = p(1) = 0.4 \\
F(2) &= P(Y \leq 2) = P(Y = 1 \text{ o } 2) = p(1) + p(2) = 0.7 \\
F(3) &= P(Y \leq 3) = P(Y = 1 \text{ o } 2 \text{ o } 3) = p(1) + p(2) + p(3) = 0.9 \\
F(4) &= P(Y \leq 4) = P(Y = 1 \text{ o } 2 \text{ o } 3 \text{ o } 4) = p(1) + p(2) + p(3) + p(4) = 0.1
\end{aligned}
$$

Ahora con cualquier otro número y, $F(Y)$ será igual al valor de F con el valor más próximo posible de Y a la izquierda de y. Por ejemplo, $F(2.7) = P(Y \leq 2.7) = p(Y \leq 2) = 0.7$ y $F(3.9999) = F(3) = 0.9$. La función de distribución acumulativa es por lo tanto

$$
F(y) = \begin{cases} 0 & \text{si } y < 1 \\ 0.4 & \text{si } 1 \leq y < 2 \\ 0.7 & \text{si } 2 \leq y < 3 \\ 0.9 & \text{si } 3 \leq y < 4 \\ 1 & \text{si } y \geq 4 \end{cases}
$$

La siguiente figura muestra una gráfica de $F(y)$.

```
df = data.frame("y" = 1:4, # Valores de Y
                "p" = c(0.4,0.7,0.9,1)) # Probabilidades
                 ↪  acumuladas asociadas
ggplot(
  data = df,
  aes(x = y,
      y = p)) +
  geom_segment(
    aes(x = c(y[1],y[-length(y)]),
        y = c(y[1],p[-length(y)]),
        xend = c(1,y[-1]),
        yend = c(1,p[-length(y)]))) +
```

```
  geom_segment(
    aes(x = y[length(y)],
        y = p[length(y)],
        xend = y[length(y)] + mean(y[2:length(y)]-
         ↪  y[1:length(y)-1]),
        yend = p[length(y)]),
    arrow = arrow(length = unit(0.2, "cm")),
    alpha = 0.4) +
  geom_point(col = "darkred") +
  scale_y_continuous(limits = c(0,1)) +
  labs(
    title = "Probabilidad acumulada",
    x = "Valores de y",
    y = "F(y)")
```

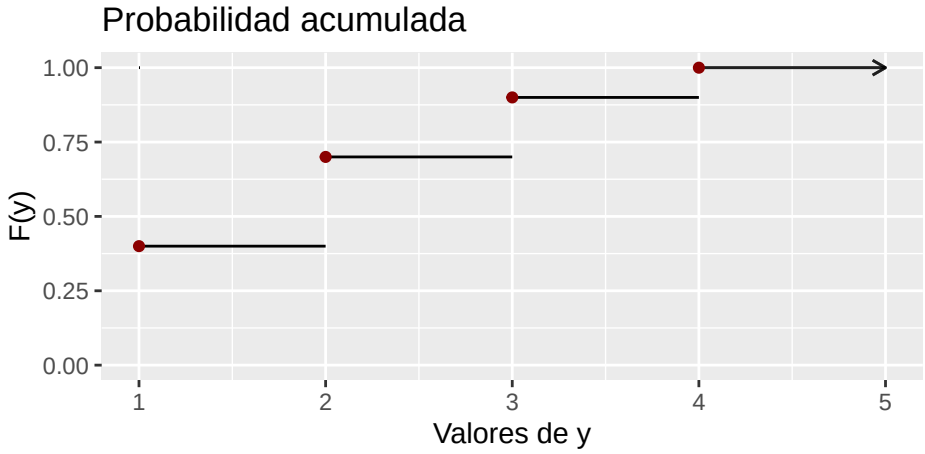


Figura 2.2: Función de distribución acumulada

Para una variable aleatorio discreta X, la gráfica de $F(X)$ mostrará un saltó con cada valor posible de X y será plana entre los valores posibles. Tal gráfica se conoce como función escalonada.

Una propiedad que surge de la función de distribución acumulada es que, para dos números cualesquiera a y b con $a \leq b$.

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) \tag{2.6}$$

En caso de que se desee calcular $P(a < X \leq b)$, la propiedad sería

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(a) - F(b) \tag{2.7}$$

De lo anterior se deduce, que dependiendo de los signos de desigualdad, cambiará la forma en la se escribe la propiedad.

Ejercicio 2.16. Remítase al ejercicio 2.15 y calcule y trace la gráfica de la función de distribución acumulada $F(X)$. Luego, utilícela para calcular las probabilidades de los eventos dados en los ítem a. y d. de dicho problema. Además, grafique la función de distribución acumulada.

Ejercicio 2.17. Una organización de protección al consumidor que habitualmente evalúa automóviles nuevos reporta el número de defectos importantes encontrados en un carro seleccionado al azar de cierto tipo. La función de distribución acumulativa de Y es la siguiente.

$$F(y) = \begin{cases} 0 & \text{si } y < 1 \\ 0.06 & \text{si } 0 \leq y < 1 \\ 0.19 & \text{si } 1 \leq y < 2 \\ 0.39 & \text{si } 2 \leq y < 3 \\ 0.67 & \text{si } 3 \leq y < 4 \\ 0.92 & \text{si } 4 \leq y < 5 \\ 0.97 & \text{si } 5 \leq y < 6 \\ 1 & \text{si } y \geq 6 \end{cases}$$

1. Calcule las siguientes probabilidades directamente con la función de distribución acumulada:
   a. $p(2)$, es decir, $P(Y = 2)$
   b. $P(Y > 3)$
   c. $P(2 \leq Y \leq 5)$
   d. $P(2 < Y < 5)$
2. ¿Cuál es la función de masa de probabilidad de X? Grafique la función de masa de probabilidad, y la función de distribución acumulada.

2.3.3. Distribuciones

A continuación se dan a conocer algunas de las distribución de probabilidad discreta más utilizadas. Cabe mencionar, que existen muchas otras distribuciones, por lo que se invita al lector a informarse de ellas en caso de lo requiera.

Uniforme

El ejemplo más sencillo de de una distribución de probabilidad discreta dada mediante una fórmula es la **distribución uniforme discreta** (Anderson et al., 2008, página 191). Su función de masa de probabilidad está definida por

$$P(X = x) = \frac{1}{n} \tag{2.8}$$

donde

$$n = \text{número de valores que puede tomar la variable aleatoria.}$$

Ejemplo 2.9. Consideremos el experimento de lanzar un dado de seis caras. Se define la variable aleatoria X como el número de puntos en la cara del dado que cae hacia arriba. En este experimento la variable aleatoria toma 6 valores posibles ($n = 6$). Por lo tanto, la función de masa de probabilidad de esta variable aleatoria uniforme discreta es

$$P(X = x) = 1/6,\ x = 1, 2, 3, 4, 5, 6$$

La figura 2.3, muestra una simulación de la función de masa de probabilidad de la distribución uniforme discreta, dependiendo del número de valores que puede tomar la variable aleatoria (n).

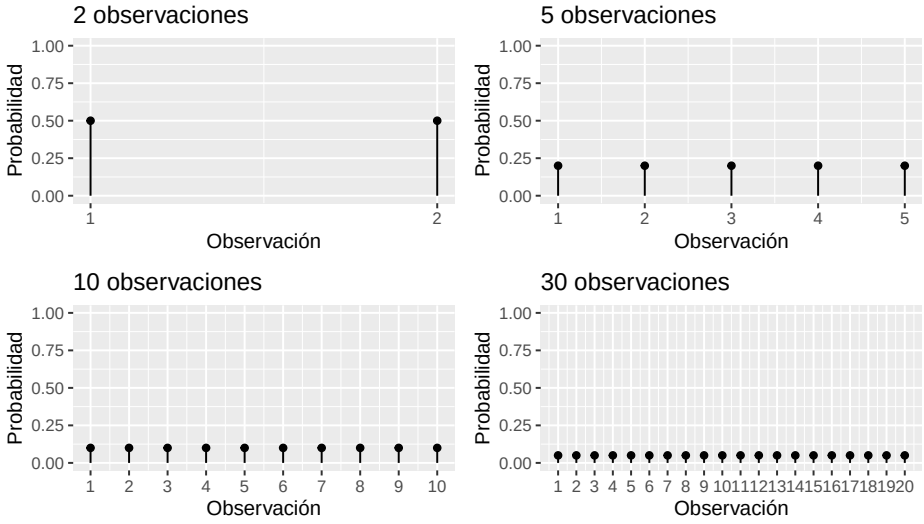


Figura 2.3: Simulación de la distribución Uniforme discreta

Bernoulli

La distribución Bernoulli es una distribución de probabilidad discreta que describe el resultado de un experimento de ensayo único que puede tener dos posibles resultados, a menudo etiquetados como éxito y fracaso, con una probabilidad de éxito de p y una probabilidad de fracaso de $q = 1 - p$. La función de masa de probabilidad de la distribución Bernoulli está dada por:

$$P(X = x) = p^x(1 - p)^{1-x} \tag{2.9}$$

donde x puede tomar únicamente los valores de 0 y 1 (Larsen and Marx, 2017, página 105).

La figura 2.4, muestra una simulación de la función de masa de probabilidad de la distribución Bernoulli, dependiendo de la probabilidad de éxito (p).

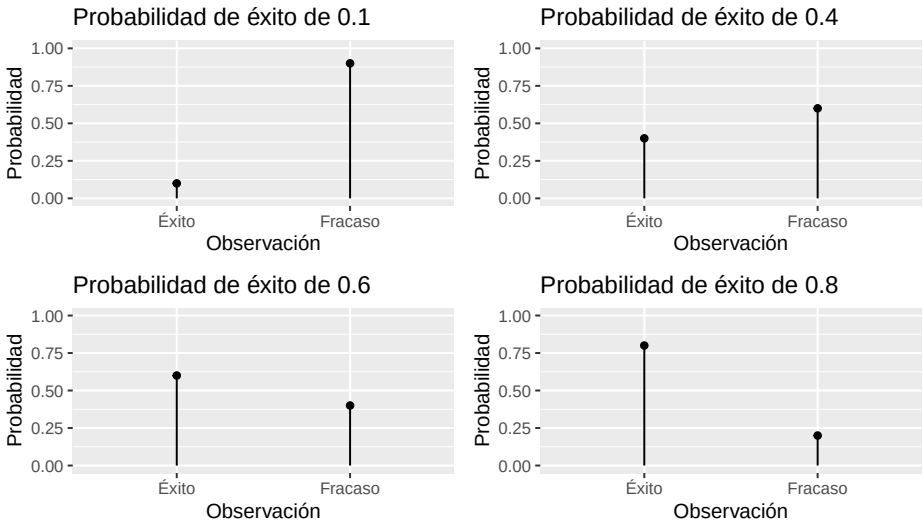


Figura 2.4: Simulación de la distribución Binomial

Ejemplo 2.10. En un experimento de lanzamiento de moneda, se puede modelar la probabilidad de obtener cara como una distribución Bernoulli. En este caso, si se define "éxito" como obtener cara y "fracaso" como obtener sello, entonces la probabilidad de éxito es $p = 0.5$ y la probabilidad de fracaso es $q = 1 - p = 0.5$. Entonces, la distribución Bernoulli para este experimento estaría dada por:

- $P(\text{Obtener cara}) = p = 0.5$
- $P(\text{Obtener sello}) = q = 0.5$

Ejemplo 2.11. En una campaña publicitaria en línea, se puede modelar la probabilidad de que un usuario haga clic en un anuncio como una distribución Bernoulli. En este caso, si se define "éxito" como un usuario que hace clic en el anuncio y "fracaso" como un usuario que no hace clic, entonces la probabilidad de éxito es p y la probabilidad de fracaso es $q = 1 - p$. Supongamos que la probabilidad de que un usuario haga clic en el anuncio es del 10 %, es decir, $p = 0.1$. Entonces, la distribución Bernoulli para este experimento estaría dada por:

- $P(\text{Hacer clic en el anuncio}) = p = 0.1$
- $P(\text{No hacer clic en el anuncio}) = q = 0.9$

Binomial

La distribución de probabilidad binomial es una distribución de probabilidad que tiene muchas aplicaciones. Está relacionada con un experimento de pasos múltiples al que se llama experimento binomial (Anderson et al., 2008, pagína 200).

Un experimento binomial tiene las siguientes cuatro propiedades.

1. El experimento consiste en una serie de n ensayos idénticos.
2. En cada ensayo hay dos resultados posibles. A uno de estos resultados se le llama éxito y al otro se le llama fracaso.
3. La probabilidad de éxito, que se denota p, no cambia de un ensayo a otro. Por ende, la probabilidad de fracaso, que se denota $1-p$, tampoco cambia de un ensayo a otro.
4. Los ensayos son independientes.

Si se presentan las propiedades $2, 3$ y 4, se dice que los ensayos son generados por un proceso de Bernoulli. Si, además, se presenta la propiedad 1, se trata de un experimento binomial.

En un experimento binomial lo que interesa es el número de éxitos en n ensayos. Si X denota el número de éxitos en n ensayos, es claro que x tomará los valores $0, 1, 2, 3, ..., n$. Dado que el número de estos valores es finito, X es una variable aleatoria discreta. A la distribución de probabilidad correspondiente a esta variable aleatoria se le llama **distribución de probabilidad binomial**.

Ejemplo 2.12. Considere el experimento que consiste en lanzar una moneda cinco veces y observar si la cara de la moneda que cae hacia arriba es cara o cruz. Suponga que se desea contar el número de caras que aparecen en los cinco lanzamientos. ¿Presenta este experimento las propiedades de un experimento binomial? ¿Cuál es la variable aleatoria que interesa? Observe que:

1. El experimento consiste en cinco ensayos idénticos; cada ensayo consiste en lanzar una moneda.
2. En cada ensayo hay dos resultados posibles: cara o cruz. Se puede considerar cara como éxito y cruz como fracaso.
3. La probabilidad de éxito y la probabilidad de fracaso son iguales en todos los ensayos, siendo $p = 0.5$ y $1 - p = 0.5$.
4. Los ensayos o lanzamientos son independientes porque al resultado de un ensayo no afecta a lo que pase en los otros ensayos o lanzamientos.

Por tanto, se satisfacen las propiedades de un experimento binomial. La variable aleatoria que interesa es X = número de caras que aparecen en cinco ensayos. En este caso, X puede tomar los valores $0, 1, 2, 3, 4$ o 5.

La función de masa de probabilidad de la distribución Binomial está dada por:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2.10}$$

donde

$$
\begin{aligned}
P(X = x) &= \text{probabilidad de } x \text{ éxitos en } n \text{ ensayos} \\
n &= \text{ número de ensayos} \\
\binom{n}{x} &= \frac{n!}{x!(n - x!)} \\
p &= \text{probabilidad de un éxito en cualquiera de los ensayos} \\
1 - p &= \text{probabilidad de un fracaso en cualquiera de los ensayos}
\end{aligned}
$$

La figura 2.5, muestra una simulación de la función de masa de probabilidad de la distribución Binomial, dependiendo del número de ensayos (n) y de la probabilidad de éxito (p).

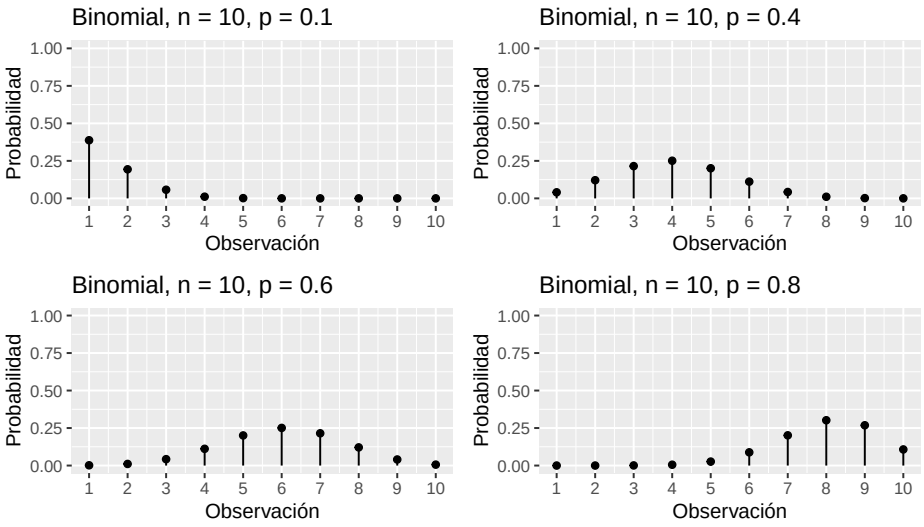


Figura 2.5: Simulación de la distribución Binomial

Ejemplo 2.13. Considere una distribución Binomial con $n = 7$ y $p = 0.4$.

a. Escriba la función de masa de probabilidad asociada.

$$P(X = x) = \binom{7}{x} 0.4^x (1 - 0.4)^{7-x}$$

b. Calcule $p(4)$.

Recordemos que $p(4) = P(X = 4)$. Para poder calcular probabilidades en un punto exacto (igual a 4) en R, se debe usar el prefijo d seguido de la

abreviatura de la distribución discreta, en este caso la abreviatura de la distribución Binomial es *binom*.

```
dbinom(
  x = 4, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  size = 7, # Cantidad de ensayos
  prob = 0.4, # Probabilidad de éxito
)
```

```
## [1] 0.193536
```

Por lo tanto, la probabilidad de obtener 4 resultados exitosos de 7 ensayos es de 0.19.

c. Calcule $P(X \leq 2)$.

En R para poder calcular probabilidades acumuladas es posible usar el prefijo *p* seguido de la abreviatura de la distribución discreta, en este caso la abreviatura de la distribución Binomial es *binom*.

Por defecto, R considera que las probabilidades acumuladas son del tipo $P(X \leq x)$, tal como se presenta en este enunciado.

```
pbinom(
  q = 2, # Se consideran valores MENORES o iguales a 2
  size = 7, # Cantidad de ensayos
  prob = 0.2, # Probabilidad de éxito
)
```

```
## [1] 0.851968
```

Por lo tanto, la probabilidad de obtener 2 o menos resultados exitosos de 7 ensayos es de 0.85.

d. Calcule $P(X < 5)$.

En este caso antes de calcular en la probabilidad en R, se debe transformar la expresión a la forma $P(X \geq x)$. Ya que estamos trabajando con eventos discretos, tenemos que

$$P(X < 5) = P(X \leq 4)$$

Luego, esta probabilidad la podemos calcular en R de la siguiente manera.

```
pbinom(
  q = 4, # Se consideran valores MENORES o iguales a 4
  size = 7, # Cantidad de ensayos
  prob = 0.2, # Probabilidad de éxito
)
```

```
## [1] 0.995328
```

Por lo tanto, la probabilidad de obtener menos de 5 resultados exitosos de 7 ensayos es de 0.99.

e. Calcule $P(X > 1)$.

R incluye un comando para aquellos casos en los que el signo de desigualdad no estricto está invertido si es esctricto, es decir, en vez de $\leq$ se tiene $>$.

```
pbinom(
  q = 1, # Se consideran valores MAYORES o iguales a 1
  size = 7, # Cantidad de ensayos
  prob = 0.2, # Probabilidad de éxito
  lower.tail = FALSE # En caso de que se tenga el signo
   ↪  mayor estricto
)
```

```
## [1] 0.4232832
```

Por lo tanto, la probabilidad de obtener más de 1 resultado exitoso de 7 ensayos es de 0.42.

f. Calcule $P(X \geq 1)$.

Para aquellos casos en que se tenga el signo de mayor no estricto ($\geq$), lo más más recomendable es transformar la expresión a no estricto ($>$) para así utilizar un código similar al del ejemplo *d.*. Ya que estamos trabajando con eventos discretos, tenemos que

$$P(X \geq 1) = P(X > 0)$$

Luego, esta probabilidad la podemos calcular en R de la siguiente manera.

```
pbinom(
  q = 0, # Se consideran valores MAYORES a 0
  size = 7, # Cantidad de ensayos
  prob = 0.2, # Probabilidad de éxito
  lower.tail = FALSE # En caso de que se tenga el signo
   ↪  mayor estricto
  )
```

```
## [1] 0.7902848
```

Por lo tanto, la probabilidad de obtener al menos 1 resultado exitoso de 7 ensayos es de 0.79.

g. ¿Para que valor de x, $P(X \leq x) = 0.6$?

Despejar esta ecuación puede llegar a ser engorroso. Sin embargo, R posee un argumento para determinar estos valores. Para el calculo se debe usar el prefijo q seguido de la abreviatura de la distribución discreta, en este caso la abreviatura de la distribución Binomial es *binom*.

```
qbinom(
  p = 0.6, # Valor resultante de la probabilidad
  size = 7, # Cantidad de ensayos
  prob = 0.2, # Probabilidad de éxito
  )
```

```
## [1] 2
```

Por lo tanto, para $x = 2$, la probabilidad de obtener a lo más x resultados exitosos es de 0.6.

Ejemplo 2.14. Un acusado va a ser declarado inocente o culpable por un jurado popular. Para ser condenado es necesario que al menos 7 personas de las 10 del jurado voten culpable. Dado que en los programas de televisión ya han dado muchos detalles del caso, los miembros del jurado están atendiendo *twitter* o leyendo el diario en vez de escuchar al fiscal y al abogado, porque van a decidir tirando una moneda al aire. ¿Cuál es la probabilidad de que el acusado sea declarado inocente?

La probabilidad de éxito (inocencia) es de $p = 0.5$. Sea X el número éxitos (votos de inocencia) en 10 ensayos (votos del jurado). Entonces, la probabilidad de ser declarado inocente esta dada por la siguiente expresión.

$$P(X \geq 4) = \sum_{k=4}^{10} \binom{10}{k} 0.5^k (1 - 0.5)^{10-k} = 0.82, \text{ o}$$

Antes de realizar el cálculo en R lo recomendable es transformar la expresión para utilizar el comando adecuado. En este caso, la expresión es:

$$P(X \geq 4) = P(X > 3)$$

Luego, en R.

```
pbinom(
  q = 3, # Se consideran valores MAYORES o iguales a 3 (es decir,
   ↪  mayor o igual a 4)
  size = 10, # Cantidad de ensayos
  prob = 0.5, # Probabilidad de éxito
  lower.tail = FALSE # TRUE: menor igual, FALSE: mayor estricto
)
```

```
## [1] 0.828125
```

Por otro lado, la probabilidad $P(X \geq 4)$ puede ser escrita como $1 - P(X \leq 3)$.

$$1 - P(X \leq 3) = \sum_{k=0}^{3} \binom{10}{k} 0.5^k (1 - 0.5)^{10-k} = 0.82$$

Es posible calcular esta expresión en R de la siguiente manera.

```
1 - pbinom(
  q = 3, # Se consideran valores MENORES o iguales a 3
  size = 10, # Cantidad de ensayos
  prob = 0.5 # Probabilidad de éxito
)
```

```
## [1] 0.828125
```

Por lo tanto, la probabilidad de que el acusado sea declarado inocente es de 0.82.

Ejercicio 2.18. En una planta de revisión técnica, resulta rechazado el 42 % de los vehículos livianos.

1. En la primera media hora de un día cualquiera se alcanzan a revisar 9 vehículos ¿cuál es la probabilidad de que más de 3 sean rechazados?
2. Una mañana se rechazaron 25 vehículos de los cuales 8 fueron fabricados antes del año 2022. Un inspector revisa al azar los antecedentes de 5 vehículos rechazados ¿cuál es la probabilidad de que la mayoría de los vehículos revisados hayan sido fabricados antes del año 2022?

Ejercicio 2.19. Se sabe que la probabilidad de que una empresa no pase la revisión de fraude fiscal es de 0.21. De las siguientes 345 empresas que se revisan en búsqueda de fraude fiscal, calcule la probabilidad de que al menos 10 empresas no aprueben la revisión.

Ejercicio 2.20. Un banco sabe que el 3 % de sus clientes no pagan a tiempo sus tarjetas de crédito. Si el banco emite 5000 tarjetas de crédito,

1. ¿Cuál es la probabilidad de que al menos 200 de ellas pertenezcan a clientes que no pagan a tiempo?
2. ¿Qué pasaría con la probabilidad de que al menos 200 de las tarjetas de crédito pertenezcan a clientes que no pagan a tiempo si la tasa de incumplimiento del banco aumenta al 5 %?
3. Si el banco quisiera reducir la probabilidad de que al menos 200 de las tarjetas de crédito pertenezcan a clientes que no pagan a tiempo a menos del 0.05 %, ¿cuál debería ser el número máximo de tarjetas de crédito que debería emitir?
4. ¿Cómo cambiaría la probabilidad de que al menos 200 de las tarjetas de crédito pertenezcan a clientes que no pagan a tiempo si el banco decidiera aumentar el número de tarjetas de crédito que otorga a cada cliente?

5. ¿Cuál es la probabilidad de que exactamente 250 de las tarjetas de crédito pertenezcan a clientes que no pagan a tiempo?

Poisson

Una variable aleatoria discreta que se suele usar para estimar el número de veces que sucede un hecho determinado (ocurrencias) en un intervalo de tiempo o de espacio. Por ejemplo, número de reparaciones en un autopista o número de fugas en un tubería. Si se satisfacen las siguientes condiciones, el número de ocurrencias es una variable aleatoria discreta, descrita por la distribución de probabilidad de Poisson (Anderson et al., 2008, página 211).

1. La probabilidad de ocurrencia es la misma para cualesquiera dos intervalos de la misma magnitud.
2. La ocurrencia o no-ocurrencia en cualquier intervalo es independiente de la ocurrencia o no-ocurrencia en cualquier otro intervalo.

La función de masa de probabilidad de Poisson se define mediante la ecuación

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2.11}$$

en donde

$P(X = x) =$ probabilidad de x ocurrencias en un intervalo de tiempo
$\lambda =$ tasa de ocurrencias en un intervalo
$e \approx 2.71828$

Es importante observar, que el número de ocurrencias de x, no tiene límite superior. Esta es un variable aleatoria discreta que toma los valores de una sucesión infinita de números ($x = 0, 1, 2, 3, 4, ...$).

La figura 2.6, muestra una simulación de la función de masa de probabilidad de la distribución Poisson, dependiendo de la tasa en función del tiempo (λ).

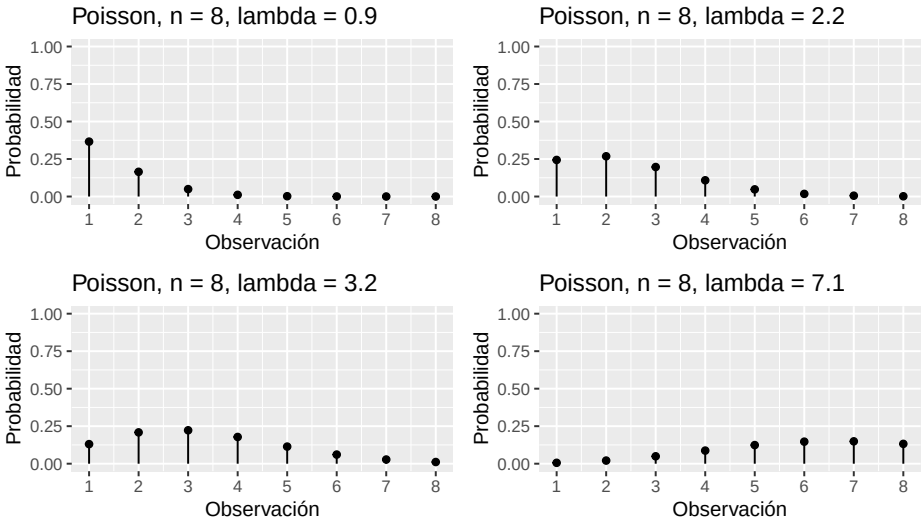


Figura 2.6: Simulación de la distribución Poisson

Ejemplo 2.15. Considere una distribución Poisson con $\lambda = 3$

a. Escriba la función de probabilidad asociada.

$$P(X = x) = \frac{3^x e^{-3}}{x!}$$

b. Calcule $p(2)$.

Considerando los comandos explicados en el ejemplo 2.13, solo es necesario modificar la abreviatura de la distribución de probabilidad, la cual, en este caso, corresponde a *pois*.

Recordemos que $p(2) = P(X = 2)$.

```
dpois(
  x = 2, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  lambda = 3 # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
  )
```

```
## [1] 0.2240418
```

Por lo tanto, teniendo una tasa de 3 por unidad de tiempo, la probabilidad de que ocurran dos sucesos es de 0.22.

c. Calcule $P(X \leq 3)$.

```
ppois(
  q = 3, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  lambda = 3 # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
  )
```

```
## [1] 0.6472319
```

Por lo tanto, teniendo una tasa de 3 por unidad de tiempo, la probabilidad de que ocurran a lo más tres sucesos es de 0.64.

d. Calcule $P(X < 2)$.

```
ppois(
  q = 1, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  lambda = 3 # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
  )
```

```
## [1] 0.1991483
```

Por lo tanto, teniendo una tasa de 3 por unidad de tiempo, la probabilidad de que ocurran menos de dos sucesos es de 0.19.

e. Calcule $P(X > 2)$.

```
ppois(
  q = 3, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  lambda = 3, # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
  lower.tail = FALSE # En caso de que se tenga el signo
   ↪  mayor estricto
  )
```

```
## [1] 0.3527681
```

Por lo tanto, teniendo una tasa de 3 por unidad de tiempo, la probabilidad de que ocurran más de dos sucesos es de 0.35.

f. Calcule $P(X \geq 5)$.

```
ppois(
  q = 4, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  lambda = 3, # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
```

```
  lower.tail = FALSE # En caso de que se tenga el signo
   ↪  mayor estricto
  )
```

```
## [1] 0.1847368
```

Por lo tanto, teniendo una tasa de 3 por unidad de tiempo, la probabilidad de que ocurran al menos cinco sucesos es de 0.18.

g. Calcule $P(X \leq x) = 0.1$.

```
qpois(
  p = 0.1, # Valor resultante de la probabilidad
  lambda = 3 # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
  )
```

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

Por lo tanto, para $x = 1$, la probabilidad de que ocurran a los más x sucesos es de 0.1, considerando una tasa de 3 por unidad de tiempo.

Ejemplo 2.16. Por un paradero pasan los buses de la línea A a una razón de 12 por hora y en forma independiente pasan los buses de la línea B a razón de 10 por hora. Un inspector observa la pasada de buses por el paradero.

1. ¿Cuál es la probabilidad de que en los primeros 10 minutos no pasen buses de la línea A?

   La tasa de llegada de buses de la línea A es de 12 por hora, por lo tanto, la tasa de llegada en 10 minutos es de 2 buses ($12/60 \cdot 10 = 2$). La probabilidad de que no pase ningún bus en esos 10 minutos esta dada por $P(X = 0)$. En R:

   ```
   dpois(
     x = 0, # Valor de X para el cual se desea calcular la
      ↪  probabilidad
     lambda = 2 # Tasa de ocurrencia por unidad de tiempo o
      ↪  espacio
   )
   ```

   ```
   ## [1] 0.1353353
   ```

   Por lo tanto, la probabilidad de que en los primeros 10 minutos no pase ningún bus de la línea A es de aproximadamente 0.13.

2. ¿Cuál es la probabilidad de que en los primeros 8 minutos pasen menos de 3 buses de la línea B?

   La tasa de llegada en 8 minutos es de aproximadamente 1.333 buses ($10/60 \cdot 8 = 1.333$). Entonces, la probabilidad de que pasen menos de 3 buses de

la línea B en esos 8 minutos esta dada por $P(X < 3) = P(X \leq 2)$. En R:

```
ppois(
  q = 2, # Valor de X para el cual se desea calcular la
   ↪  probabilidad
  lambda = 1.333 # Tasa de ocurrencia por unidad de tiempo o
   ↪  espacio
)
```

```
## [1] 0.8494467
```

La probabilidad de que en los primeros 8 minutos pasen menos de 3 buses de la línea B es de aproximadamente 0.84.

Ejercicio 2.21. Una compañía de seguros tiene un promedio de 4 reclamaciones de seguros de automóviles por día.

1. ¿Cuál es la probabilidad de que la compañía de seguros reciba menos de 3 reclamaciones en un día?
2. Si la compañía de seguros quiere asegurarse de tener suficientes empleados para manejar la carga de trabajo, ¿cuántos empleados deben tener en su equipo si quieren tener una probabilidad del 90 % de manejar cualquier cantidad de reclamaciones en un día?
3. Si la compañía de seguros quiere limitar su riesgo y sólo estar dispuesta a pagar por un número limitado de reclamaciones, ¿cuál es la cantidad máxima de reclamaciones que deberían estar dispuestos a pagar por día si sólo están dispuestos a aceptar un riesgo del 5 % de exceder ese límite?

Ejercicio 2.22. Un banco recibe en promedio 2 solicitudes de préstamos hipotecarios por hora.

1. ¿Cuál es la probabilidad de que el banco reciba exactamente 3 solicitudes de préstamo hipotecario en un período de 90 minutos?
2. Si el banco quiere asegurarse de tener suficiente personal para manejar una alta demanda de solicitudes de préstamo, ¿cuántos empleados deben tener en su equipo si quieren tener una probabilidad del 95 % de manejar cualquier cantidad de solicitudes en una hora?
3. Si el banco quiere limitar su exposición a riesgos y sólo estar dispuesto a aprobar un número limitado de préstamos, ¿cuál es la cantidad máxima de solicitudes que deberían estar dispuestos a aprobar por hora si sólo están dispuestos a aceptar un riesgo del 10 % de exceder ese límite?

Ejercicio 2.23. Un laboratorio farmacéutico encarga una encuesta para estimar el consumo de cierto medicamento que, elabora con el fin de controlar su producción. Se sabe que, a lo largo de un año cada persona tiene una posibilidad entre mil de necesitar el medicamento, y el laboratorio podrá vender en promedio 4000 unidades del producto por año.

1. ¿Cuál es la probabilidad de que el número de enfermos no exceda 4 por año?

2. ¿Cuál es la probabilidad de que el número de enfermos sea más de 2 por año?
3. ¿Cuál es la probabilidad de que haya 12 enfermos en un año?

2.4. Variables aleatorias continuas (v.a.c)

2.4.1. Función de densidad de probabilidad

2.4.2. Función de distribución acumulada

2.4.3. Distribuciones

Uniforme

Exponencial

Normal

T - Student

Ji - Cuadrado

2.5. Esperanza

2.5.1. v.a.d

2.5.2. v.a.c

2.6. Varianza

2.6.1. v.a.d

2.6.2. v.a.c

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.

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

de Micheaux, P. L., Drouilhet, R., and Liquet, B. (2013). *R and Its Documentation*, pages 141–150. Springer New York, New York, NY.

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.

Larsen, R. J. and Marx, M. L. (2017). *An Introduction to Mathematical Statistics and Its Applications*. Pearson, 6th edition.

Peng, R. D. (2016). *R programming for data science*. Leanpub, Victoria, BC, Canada.

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.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Use R! Springer, New York.