Estadísticas

Daniel Franzani

Actualizado al 04-03-2024

Índice general

Presentación

La asignatura Estadísticas, es el único curso estadístico de la carrera de Auditoría. El curso tiene un enfoque práctico con un fuerte énfasis en el estudio descriptivo de datos. Este primer documento, se concentra en gráficos descriptivos, medidas de resumen, funciones de probabilidad y distribuciones muestrales, haciendo uso del entorno de software R.

Enlace al documento del segundo curso de Estadística: Estadística II & Inferencia Estadística.

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 fueron extraídos de la bibliografía obligatoria, sin embargo, a algunos de estos se ha incorporado el **uso de R** como programa de análisis estadístico. El desarrollo de los ejercicios por parte del estudiante se recomienda que sea en Google Colab R. Esta plataforma cuenta con una opción de configuración interna para R (desde Google Drive: Nuevo -> Más -> Google Collaboratory, dentro del archivo: Entorno de ejecución -> Cambiar tipo de entorno de ejecución -> Tipo de entorno de ejecución -> R -> Guardar). En el siguiente enlace se puede acceder a un documento con una configuración preestablecida para este lenguaje. El archivo generado se guardará automáticamente en la cuenta de Gmail predeterminada; otra opción en caso de no querer modificar su cuenta predeterminada (si es que debiese hacerlo) es descargar el archivo y cargarlo manualmente en la carpeta de Drive que estime conveniente. Los aspectos relacionados con el uso de R serán abordados en el Taller Introductorio.

3. Se cuenta con **talleres de práctica**, lo cuales, cuentan con ejercicios propuestos para desarrollar en clases y ejercicios para el trabajo independiente del estudiante.

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

    - (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, se añaden citas que refuerzan el contenido presentando, las cuales se encuentra en detalle al final de cada sección.

5. Las **bases de datos** a utilizar en el curso se encuentran disponibles en un repositorio web público.

Unidad 1

Tópicos básicos de estadística

Las bases de datos que se trabajarán en esta unidad son las siguientes:

- Tasa Euro/Dólar: Contiene el registro diario histórico de la tasa de cambio del Euro a Dólar durante el 2023. Las columnas de la base de datos son las siguientes:
  - 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).
- Precios de electricidad: Un conjunto de datos históricos que contiene el precio por hora de la electricidad para Bélgica. Las columnas de la base de datos son las siguientes:
  - MTU: Hora de inicio (formato fecha y hora) del coste de la electricidad.
  - EUR_MWh: Precio por hora (Euros por MWh).
- Pacientes: Contiene datos respecto a los ataques al corazón de distintos pacientes hospitalarios. El detalle de algunas de las columnas de la base de datos que utilizaremos son las siguientes:
  - age: edad del paciente (en años).

- sex: sexo del paciente (Hombre: 1 y Mujer: 0).
- cp: Tipo de dolor en el pecho, Valor 1: angina típica, Valor 2: angina atípica, Valor 3: dolor no anginoso, Valor 4: asintomático.
- trtbps: presión arterial en reposo (en mm Hg).
- chol: nivel de colestorol (en mg/dl).
- fbs: azúcar en sangre en ayunas > 120 mg/dl (V = 1; F = 0).
- thalachh: frecuencia cardíaca máxima alcanzada (en latidos por minuto).
- oldpeak: tiempo de duración del último ataque al corazón (en minutos).

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.

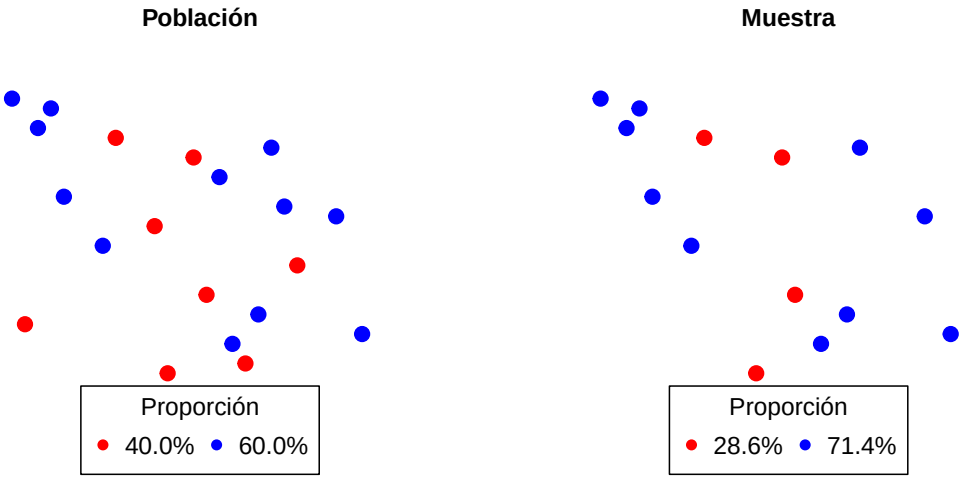


Figura 1.1: Parámetro y estadístico

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?

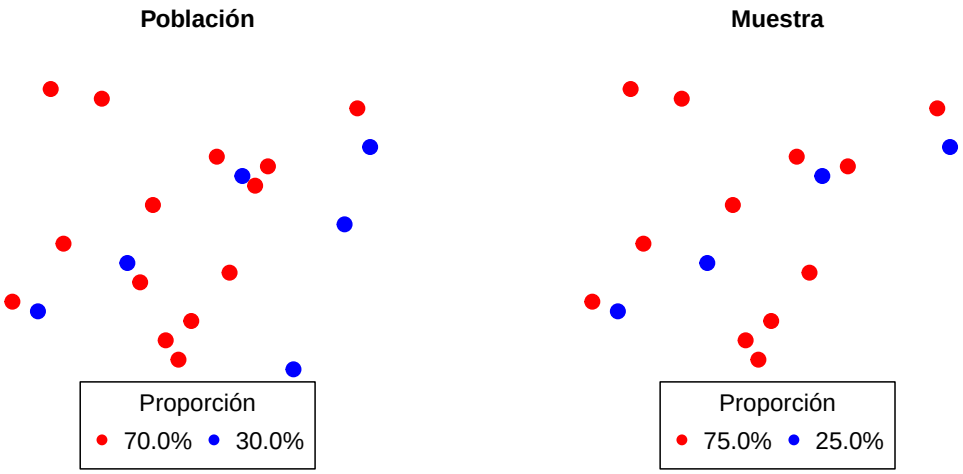


Figura 1.2: Estimador y estimación

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 ?

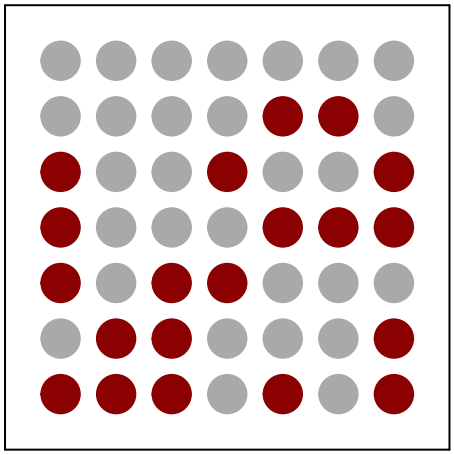


Figura 1.3: Población

Luego,

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

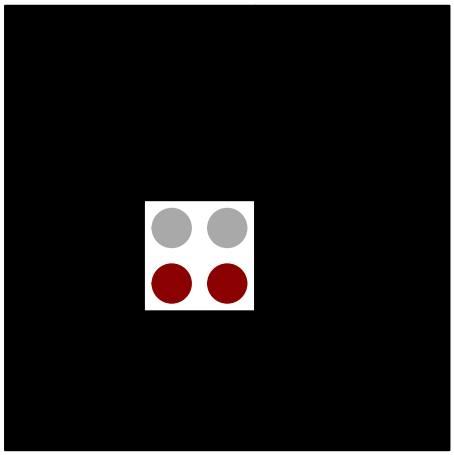


Figura 1.4: Muestra 1

¿Y ahora? (Figura 1.5**)**

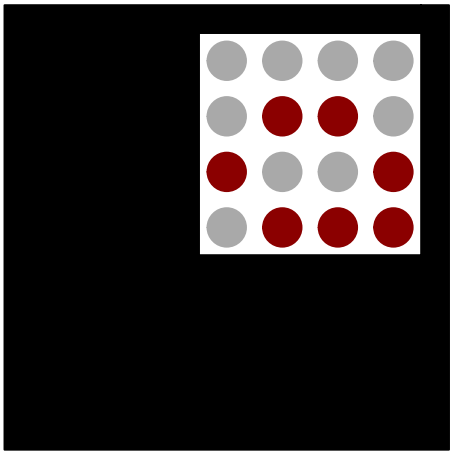


Figura 1.5: Muestra 2

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.

1.2. 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.

1.2.1. 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. Respecto a la base Tasa Euro/Dólar, 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.

Ejercicio 1.4. Obtenga de la media del precio por hora de la electricidad, en la base Precios de electricidad. Interprete.

Ejercicio 1.5. Considerando la base de datos Pacientes, determine la media del nivel de colesterol por sexo. Interprete.

1.2.2. 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)
```

```r
# 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 1.6. Obtenga la mediana de las distintas variables (cuando corresponda) de la base de datos Pacientes. Interprete.

1.2.3. 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:

```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 Tasa Euro/Dólar (sin imprimir la tabla).

   ```r
   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 1.7. Considerando la base de datos Pacientes, determine la moda de las variables cualitativas. Interprete.

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).

1.3. 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	–	–

1.3.1. 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 Tasa Euro/Dólar.

```
# 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
```

El valor máximo siempre estará en la segunda posición y le mínimo en la segunda.

Ejercicio 1.8. Utilizando la base de datos Paciente, determine el rango de

las variables cuantitativas. Interprete.

Ejercicio 1.9. Determine el rango de la fechas de medición en los precios de electricidad en la base Precios. Interprete.

1.3.2. Varianza y desviación estándar

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

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

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 Tasa Euro/Dólar, 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. Similarmente, se observa que la tasa de cierre tiene menor variabilidad que la tasa de apertura.

```
c(var(datos$Open), var(datos$Close))
```

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

```
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 1.10. Utilice la varianza directamente para comparar la variabilidad de la variables *chol*, *age*, *trtbps* y *oldpeak* entre hombres y mujeres, en la base Pacientes. Interprete.

Ejercicio 1.11. Utilice la varianza directamente para comprar la variabilidad de los precios de la electricidad entre las primeras y últimas 1000 mediciones, en la base Precios. Interprete.

1.3.3. 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 ejemplo 1.3, 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 1.12. Compare la variabilidad de la presión arterial en reposo y el nivel de colesterol de los pacientes registrados en la base de datos Pacientes. Repita el estudio diferenciado por sexo. Interprete.

Ejercicio 1.13. Utilizando la base de datos Pacientes, compare la variabilidad de la edad de los paciente, según el tipo de dolor de pecho que presentan. 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).

1.4. 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

1.5. 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).

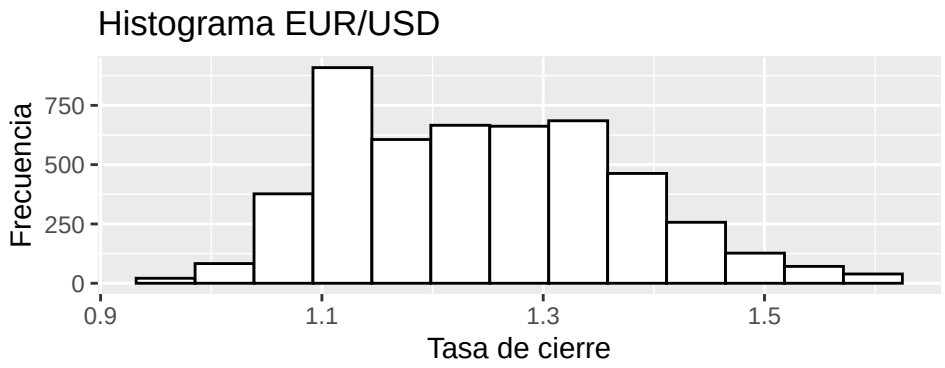
1.5.1. Histograma

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 (se toma como ejemplo la base de datos Tasa Euro/Dólar):

```
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 = round(1 + log2(dim(datos)[1])), # Cantidad de
     ↪  intervalos del histograma: regla de Sturges
    color = "black", # Color del borde de las barras del
     ↪  histograma
    fill = "white", # 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
```



Para interpretar un histograma, basta con indicar los siguientes aspectos:

- Forma visual de la distribución de las barras (en general).
- Mencionar si existe una concentración fuera del comportamiento general, y en dónde se encuentra.

En el caso del histograma de la tasa de cierre del EUR/USD, se observa una forma de campana centrada cerca del 1.3, además, se evidencia la presencia de una frecuencia superior al resto que se encuentra la izquierda del gráfico cerca del 1.1.

Cabe mencionar, que existen otros aspectos que son posibles mencionar, para ello consulte la bibliografía del curso.

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: regla de
          ↪  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 1.14. Utilizando la base de datos de Precios, elabore un histograma de los precios de la electricidad. Interprete.

1.5.2. Gráfico de 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.

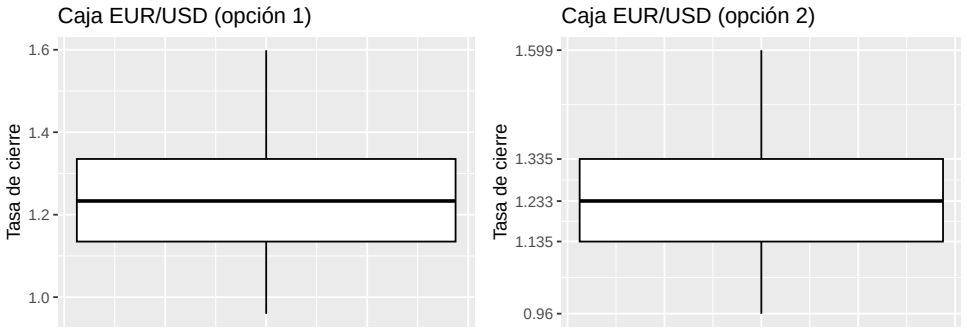
Los pasos para elaborar un gráfico de caja son los siguiente (Anderson et al., 2008, página 106):

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 (se toma como ejemplo la base de datos Tasa Euro/Dólar):

```
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 = "white") + # 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
```



Para interpretar un gráfico de caja es recomendable utilizar la opción 2 mostrada anteriormente, ya que, se debe mencionar uno de los puntos relevantes del gráfico. En el ejemplo recién dado, se observa que, el primer, segundo y tercer cuartil están en 1.135, 1.233 y 1.335 respectivamente, mientras que el valor mínimo y máximo están en 0.96 y 1.559 respectivamente. Adicionalmente, se puede mencionar que los datos superiores en comparación a los inferiores, se encuentran más alejados de la mediana.

Ejercicio 1.15. Utilizando la base de datos de Precios, elabore un gráfico de caja de los precios de la electricidad. Interprete.

1.5.3. Gráfico de 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:

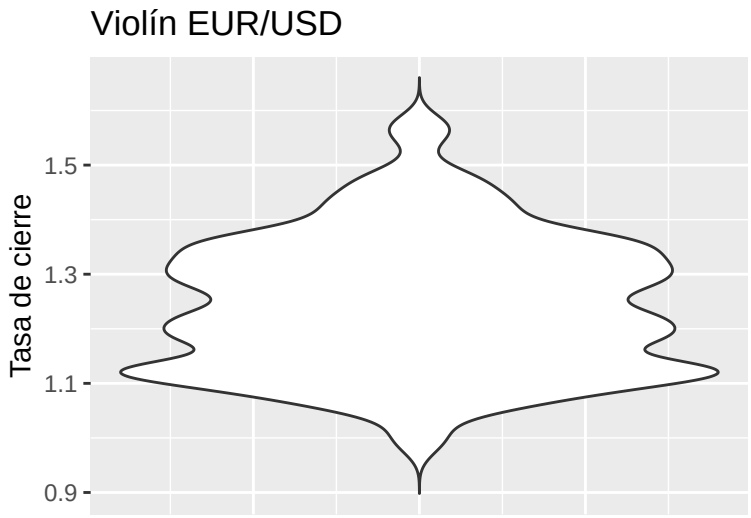
1. Dibujo de la traza de densidad: la traza de densidad se dibuja sobre el eje vertical en el gráfico de violín ("forma suavizada del histograma").
2. Creación de la sección central simétrica: se crea una sección central simétrica que representa la mitad de la traza de densidad.

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

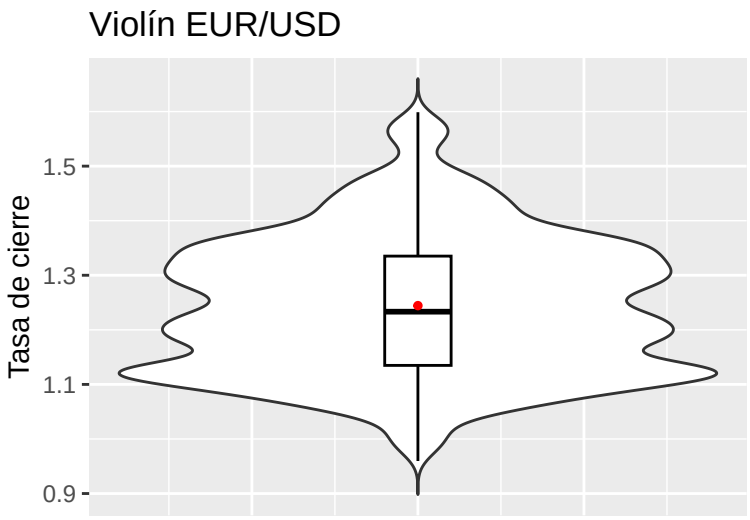
Para generar un gráfico de violín en R a partir de un conjunto de datos, se utiliza el siguiente código (se toma como ejemplo la base de datos Tasa Euro/Dólar):

```
# 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 = "white") + # 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
```




```
# 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 = "white") + # 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
    orientation = "x") # Orientación
```



Para interpretar un gráfico de violín con caja y promedio se deben mencionar tres aspectos relevantes:

- Ubicación de la(s) mayor(es) concentración(es) de datos, utilizando como referencia los cuartiles.
- Ubicación del promedio respecto a la mediana.
- Posibles razones por las cuales se explica la ubicación anteriormente mencionada del promedio respecto a la mediana.

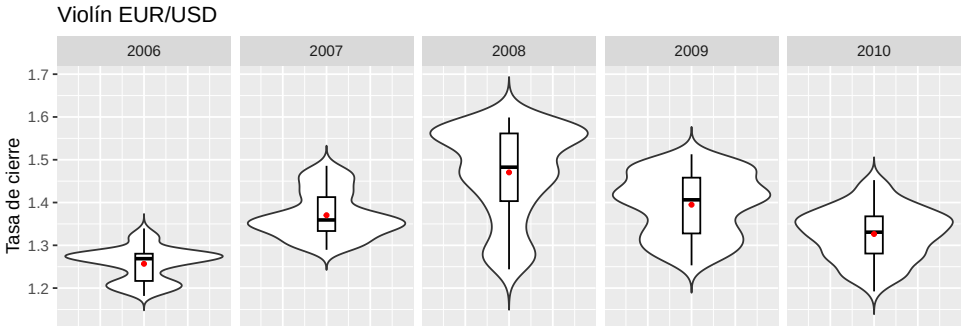
En el ejemplo anterior, la principal concentración se encuentra por debajo del primer cuartil, aunque destacan otras dos concentraciones que están por debajo del segundo cuartil y alrededor del tercer cuartil respectivamente. El promedio se encuentra sutilmente por encima de la mediana, esto se puede explicar debido a que los datos superiores del gráfico se encuentra más lejos de la mediana en comparación a los datos inferiores.

Ejercicio 1.16. Utilizando la base de datos de Precios, elabore un gráfico de violín con caja y promedio de los precios de la electricidad. 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** en el periodo de años 2006-2010, diferenciando por año.

```
# 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[datos$Ano %in% 2006:2010,], # 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 = "white") + # 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 = "white") + # 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
```



Para interpretar este tipo de gráficos, se debe realizar un interpretación uno a uno, siguiendo la recomendación antes dada. También, es posible comprar los gráficos a través de la media y coeficiente de variabilidad. Para este ejemplo, queda como trabajo del estudiante realizar esta comparación.

Ejercicio 1.17. Utilizando la base de datos de Paciente Realice un gráfico de violín con caja y promedio del nivel de colesterol de los paciente, diferenciando por el nivel de azúcar en sangre en ayunas. Interprete.

Ejercicio 1.18. Agregue una diferenciación por sexo a lo realizado en el ejercicio 1.17. Entienda que para cada nivel de azúcar en sangre se debe ver un desglose por sexo. Interprete.

Ejercicio 1.19. Utilizando la base de datos de Precios, realice un gráfico de violín con caja y promedio para el precio de la electricidad, diferenciado por año. Interprete.

1.5.4. Gráfico de Barras

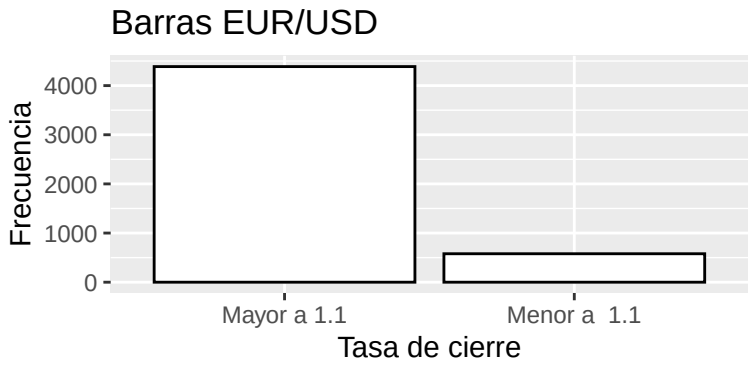
Una gráfico de barras, es una gráfica para representar los datos cualitativos de una distribución de frecuencia. El procedimiento de construcción es el siguiente (Anderson et al., 2008, página 29).

1. En uno de los ejes de la gráfica (por lo general en el horizontal), se especifican las etiquetas empleadas para las clases (categorías).
2. Para el otro eje de la gráfica (el vertical) se usa una escala para frecuencia, frecuencia relativa o frecuencia porcentual.
3. 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 (se toma como ejemplo la base de datos Tasa Euro/Dólar):

```
# 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 = "white", color = "black") + # Color de relleno y
     ↪  borde
  labs( # Títulos
    title = "Barras EUR/USD", # Título del gráfico
    x = "Tasa de cierre", # Título del eje X
    y = "Frecuencia") # Título del eje Y
```



La interpretación de este tipo de gráfico (gráfico de barras no apiladas) es bastante intuitiva, ya que consiste en mencionar cuál categoría tiene un mayor frecuencia. En caso de graficar una variable con más de 2 categorías, se debe mencionar a que altura del eje vertical se encuentra la altura de cada una de las barras. Para obtener mayor detalles respecto a las frecuencias por categoría, suele apoyarse con un tabla de frecuencias absolutas o relativas porcentuales.

En el ejemplo, la frecuencia de la cantidad de tasas de cierre que se encuentran por sobre 1.1 es mayor a las que se encuentran por debajo, con una frecuencia sobre 4000 y menor a 1000 respectivamente.

Ejercicio 1.20. Utilizando la base de datos Pacientes, elabore un gráfico de barras de la variable *fbs*. Interprete, apoyado de una tabla de frecuencias absolutas.

Ejercicio 1.21. Utilizando la base de datos Pacientes, elabore un gráfico de barras (no apiladas) de la variable *cp*, diferenciado por sexo. Interprete, apoyado de una tabla de frecuencias relativas porcentuales. Haga el contraste visual con el gráfico de barras apiladas.

1.5.5. Gráfico de 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).

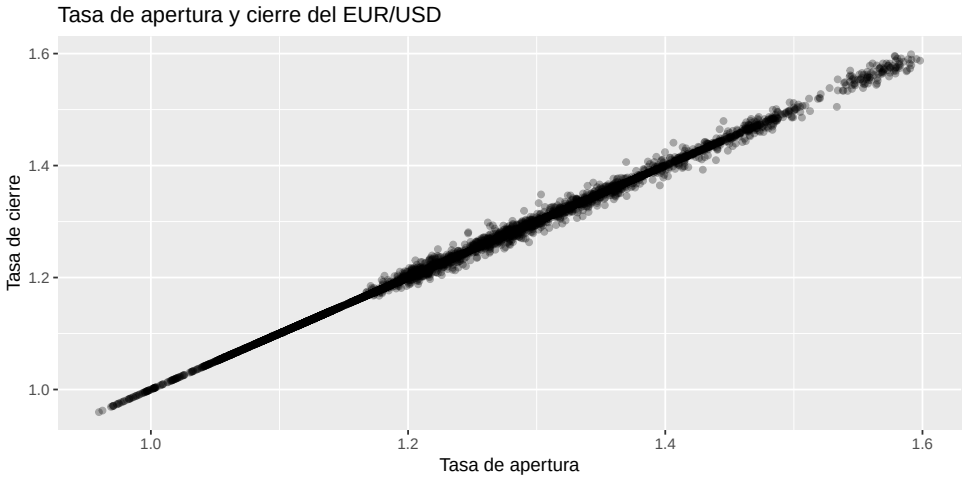
Los pasos para elaborar un gráfico de caja son los siguiente (Healy, 2019):

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 (se toma como ejemplo la base de datos Tasa Euro/Dólar):

```
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 = "black", # 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
```



Tal como se menciona al inicio, la interpretación de este tipo de gráficos radica en describir la tendencia de los puntos. En el ejemplo anterior, el gráfico muestra una tendencia al alza, es decir, que cuando la tasa de apertura del EUR/USD aumenta, entonces, la tasa de cierre tiende a aumentar.

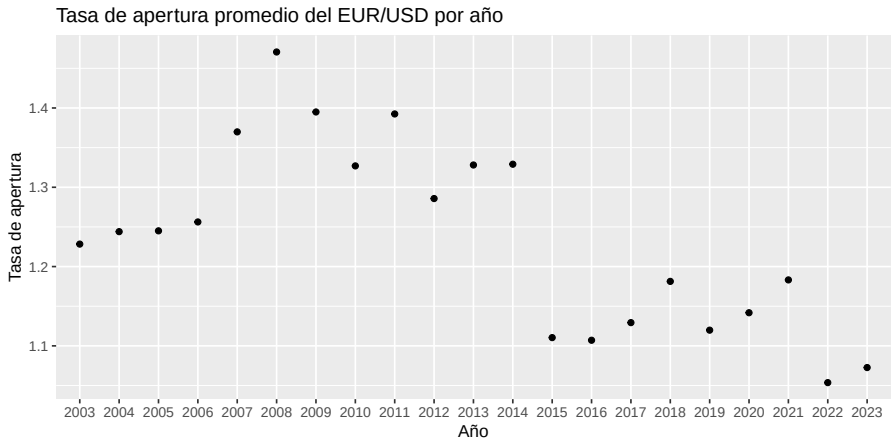
Ejercicio 1.22. Utilizando la base de datos Pacientes, realice un gráfico de dispersión entre la variable *age* (eje X) y la variable *thalachh* (eje Y). Interprete.

Ejemplo 1.6. En los siguiente gráficos se toma como ejemplo la base de datos Tasa Euro/Dólar.

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).

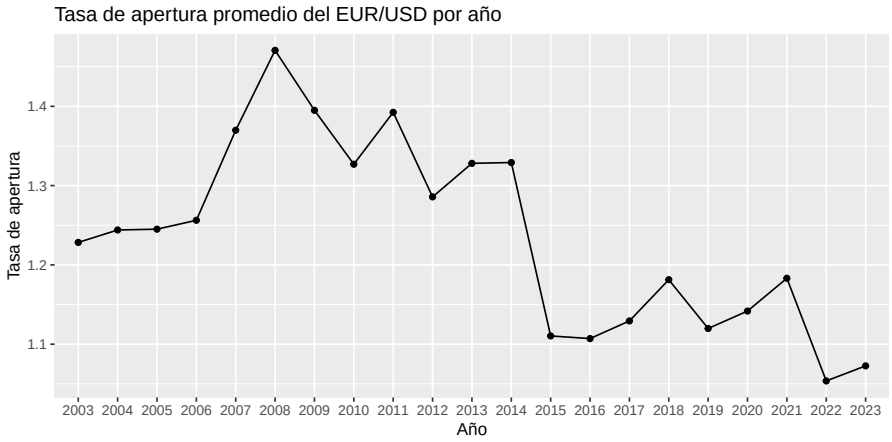
```
g = ggplot( # Ambiente gráfico
  data = datos, # Base de datos a utilizar
  aes( # Comando estéticos
x = Ano, # Eje X y variable asociada
y = Open, # Eje Y y variable asociada
group = 1)) + # Comando únicamente necesario para la
 ↪  pregunta 2
```

```
  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
```

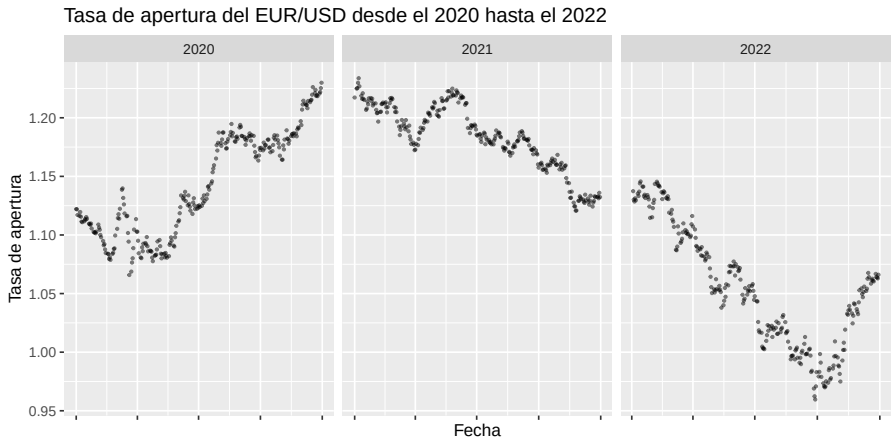


Hasta el 2008 la tasa promedio de apertura estuvo en alza, posteriormente, la tasa decayó a un valor inferior a 1.1.

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
```

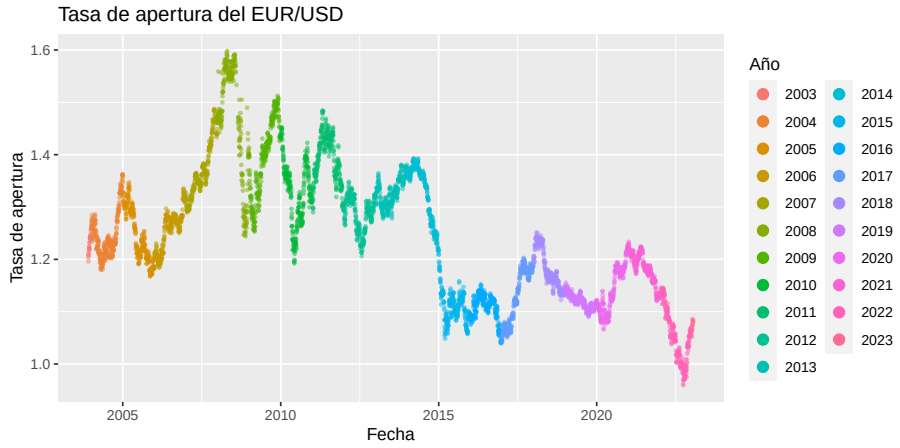


Durante los 3 años consecutivos, se observa que únicamente en el 2020 la tendencia de la tasa de apertura es al alza, mientra que para los otros dos años hubo un decaimiento en el valor de esta.

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.

Ejercicio 1.23. Utilizando la base de datos de la Tasa Euro/Dólar:

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.

Ejercicio 1.24. Utilizando la base de datos de Precios de electricidad, elabore un gráfico de dispersión del precio de la electricidad a través del tiempo (considere año, mes y día). Interprete.

Ejercicio 1.25. Utilizando la base de datos de Pacientes:

1. Elabore un histograma del nivel de colesterol de los pacientes. Interprete.
2. Elabore un gráfico de caja del nivel de colesterol de los pacientes, diferenciando por sexo. Interprete y compare.
3. Elabore un gráfico de violín (más caja y promedio) del nivel de colesterol de los pacientes, diferenciado por tipo de dolor en el pecho. Interprete y compare.
4. Elabore un gráfico de dispersión entre la edad (eje X) y el nivel de colesterol (eje Y) de los pacientes, coloreando por sexo. 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, se 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?

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

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 está 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 impar. ¿Cuál es el conjunto correspondiente al evento aleatorio?

Ejercicio 2.5. Experimento aleatorio: Elegir una carta al azar de una baraja inglesa 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 (el centro y 6 sectores circulares). 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?

2.1.3.1. 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 : No se termino el trabajo a tiempo

D : 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 : El empleado se va de la empresa por insatisfacción con el salario

W : 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 %. Nueve de los fondos tienen un rendimiento anual arriba de 50 %, siete de los fondos a cinco años lo tienen 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 $0, 1, 2, ...$ 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 continuas

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)	$150 \leq x \leq 212$

Ejercicio 2.14. 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 caseta de peaje
c.	Revisar 50 declaraciones de impuestos	Número de declaraciones que tienen algún error
d.	Observar trabajar 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 requieren 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 uno 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 2.6. 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 a lo más dos bombas estén en servicio es

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

Por otro lado, la probabilidad de que entre 2 y 4 bombas (inclusive) estén en servicio es

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

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

```
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( # Ambiente gráfico
  data = df, # Base de datos a utilizar
  aes(x = x, # Variable del eje X
      y = p)) + # Variable del eje Y
  geom_point() + # Tipo de gráfico
  geom_segment( # Añadimos segmentos
    aes(x = x, # Coordenada X de los puntos de inicio
        y = rep(0,length(x)), # Coordenada Y de los puntos de
         ↪ inicio
        xend = x, # Coordenada X de los puntos de llegada
        yend = p)) + # Coordenada Y de los puntos de llegada
  labs( # Edición de títulos
    title = "Probabilidades de cada valor", # Título del
     ↪ gráfico
    x = "Valores del experimento (x)", # Título de eje X
    y = "Probabilidades") # Título del eje Y
```

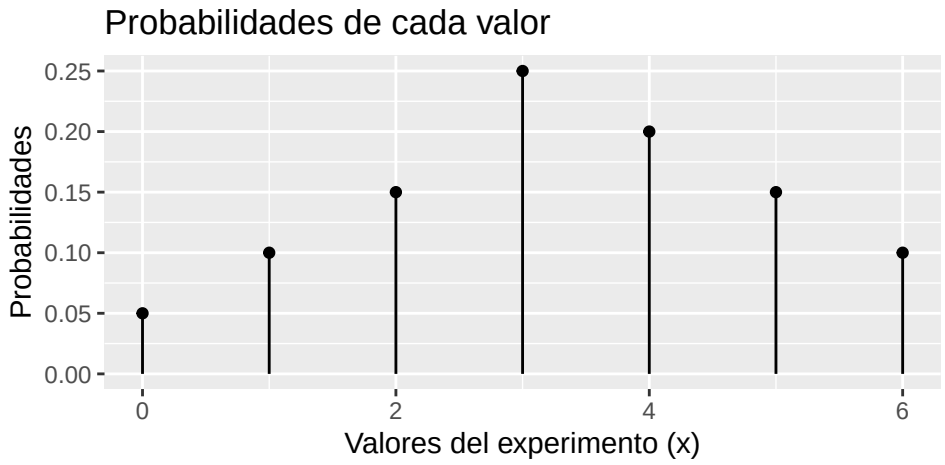


Figura 2.1: Función de masa

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 2.15. 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 componentes 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 2.16. Una empresa de ventas en línea dispone de seis líneas telefónicas. Sea X el número de líneas en uso en un tiempo especificado. 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 lo 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 2.7. 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 muestras, una de cada individuo hasta que se 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 $(1, 2, 3, 4)$:

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

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

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

La siguiente figura muestra la 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( # Ambiente gráfico
  data = df, # Base de datos
  aes(x = y, # NOmbre de la variable a graficar en el eje X
      y = p)) + # Nombre de la variable a graficar en el eje Y
  geom_segment( # Generar segmentos
    aes(x = c(min(y),y[-length(y)]), # Coordenada X de puntos
     ↪  de inicio de los segmentos
        y = c(min(y),p[-length(y)]), # Coordenada Y de puntos
         ↪  de inicio de los segmentos
        xend = c(min(y),y[-1]), # Coordenada X de puntos
         ↪  finales de los segmentos
        yend = c(min(y),p[-length(y)]))) + # Coordenada X de
         ↪  puntos finales de los segmentos
  geom_segment( # Código que genera la flecha dentro del
   ↪  gráfico
    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")), # Argumento que
     ↪  permite dar la forma de flecha al último segmento
    alpha = 0.4) +
  geom_point(col = "darkred") + # Punto de incio de los
   ↪  segmentos
  scale_y_continuous(limits = c(0,1)) + # Edición de los
   ↪  límites del eje Y
  labs( # Edición de títulos
```

```
    title = "Probabilidad acumulada", # Título del gráfico
    x = "Valores de y", # Título del eje X
    y = "F(y)") # Título del eje Y
```

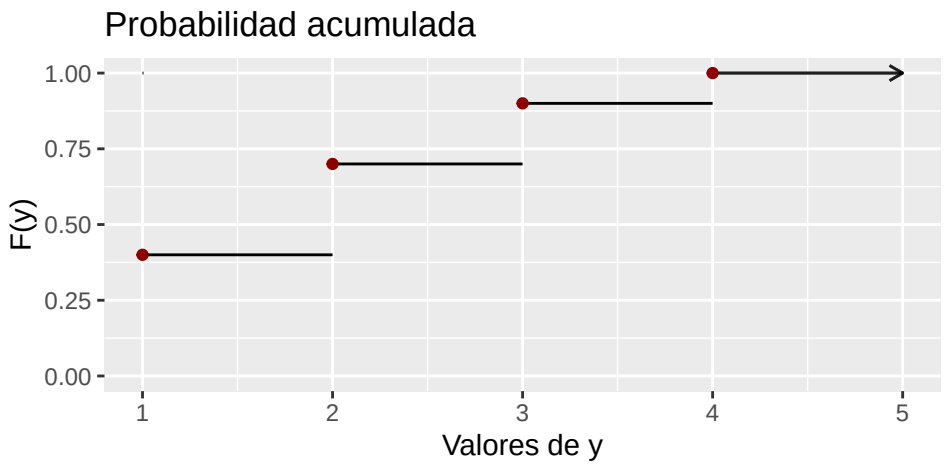


Figura 2.2: Función de distribución acumulada

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 2.17. Remítase al ejercicio 2.16 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 2.18. 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.

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

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.

Ejercicio 2.19. En una fábrica de productos electrónicos, se sabe que la probabilidad de que un artículo sea defectuoso sigue una distribución de probabilidad de masa con los siguientes valores:

Número de defectos	Probabilidad
0	0.50
1	0.30
2	0.02
3	0.08
4	0.10

A continuación:

1. Determine la función de masa de probabilidad.
2. Determine la función de distribución acumulada.
3. Si se selecciona un artículo al azar. Utilizando la función de distribución acumulada calcule:
   a. La probabilidad de que tenga cuando mucho de 2 defectos.
   b. La probabilidad de que tenga más de 0.4 defectos.
   c. La probabilidad de que tenga entre 1 y 4 defectos.
   d. La probabilidad de que tenga cuando menos 2.7 defectos.

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 estudiante a informarse de ellas en caso de que lo requiera.

2.3.3.1. 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

$n =$ número de valores que puede tomar la variable aleatoria.

Ejemplo 2.8. 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).

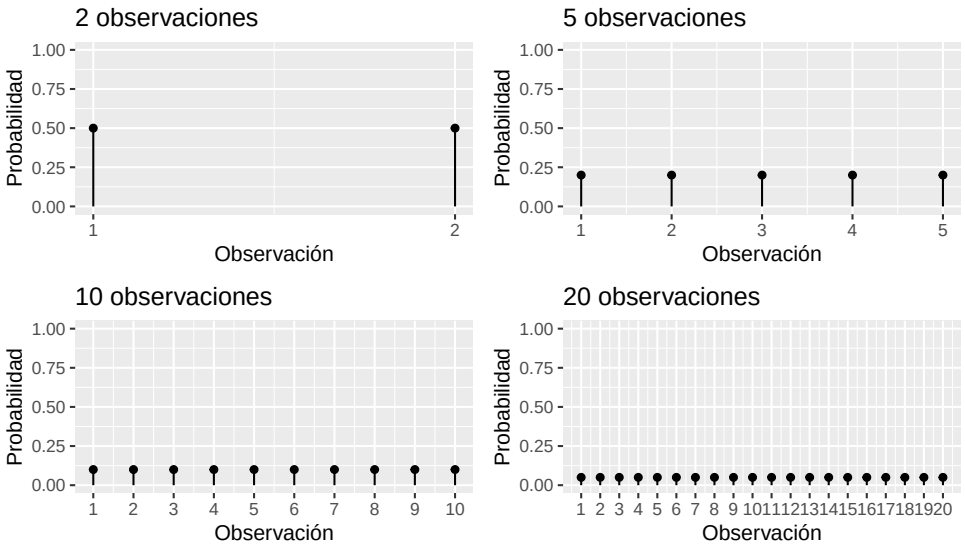


Figura 2.3: Simulación de la distribución Uniforme discreta

2.3.3.2. 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).

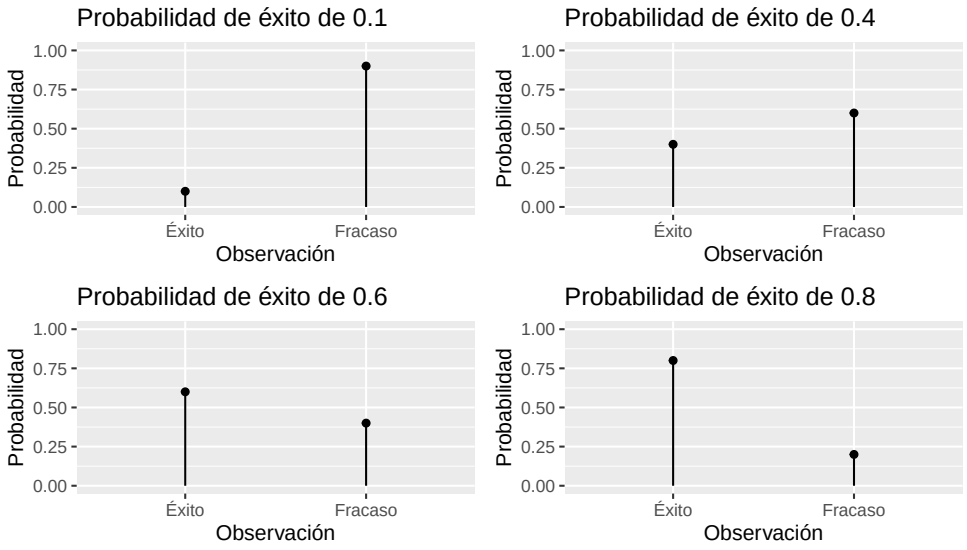


Figura 2.4: Simulación de la distribución Binomial

Ejemplo 2.9. 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 2.10. 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:

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

2.3.3.3. 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, paǵina 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 2.11. 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).

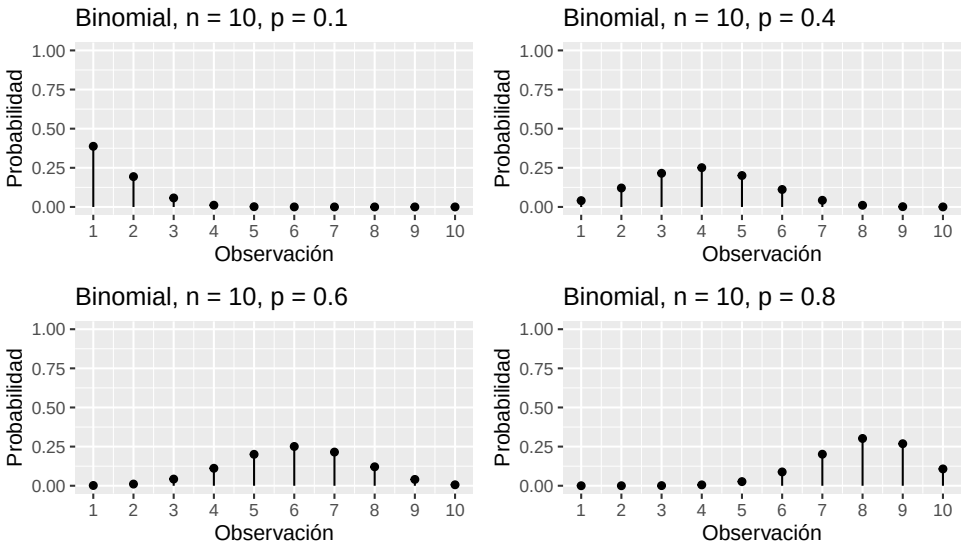


Figura 2.5: Simulación de la distribución Binomial

Ejemplo 2.12. Considere una distribución Binomial con $n = 7$ y $p = 0.2$.

a. Escriba la función de masa de probabilidad asociada.

$$P(X = x) = \binom{7}{x} 0.2^x (1 - 0.2)^{7-x}$$

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 = 0.2, # Probabilidad de éxito
)
```

```
## [1] 0.028672
```

Por lo tanto, la probabilidad de obtener 4 resultados exitosos de 7 ensayos es de 0.028672.

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 $P(X \leq x)$. 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 estricto es del tipo mayor.

```
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 igual ($\geq$), lo más más recomendable es transformar la expresión a 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 2.13. 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.

```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.

```r
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 2.20. En una planta de revisión técnica, resulta rechazado el 42 % de los vehículos livianos. En la primera media hora de un día cualquiera se alcanzan a revisar 9 vehículos.

1. ¿cuál es la probabilidad de que más de 3 sean rechazados?
2. ¿cuál es la probabilidad de que a lo más 5 sean rechazados?
3. ¿cuál es la probabilidad de que menos de 2 sean rechazados?
4. ¿cuál es la probabilidad de no se rechacen todos los vehículos?

Ejercicio 2.21. 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 cuando mucho 85 empresas no aprueben la revisión.

Ejercicio 2.22. Un banco sabe que el 8 % de sus clientes no pagan a tiempo sus tarjetas de crédito. Si el banco emite 2500 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 9.2 %?
3. ¿Cuál es la probabilidad de que exactamente 250 de las tarjetas de crédito pertenezcan a clientes que no pagan a tiempo?
4. ¿Cuál es la probabilidad de que cuando mucho 210 de ellas pertenezcan a clientes que no pagan a tiempo?

2.3.3.4. 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, el 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 definida 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).

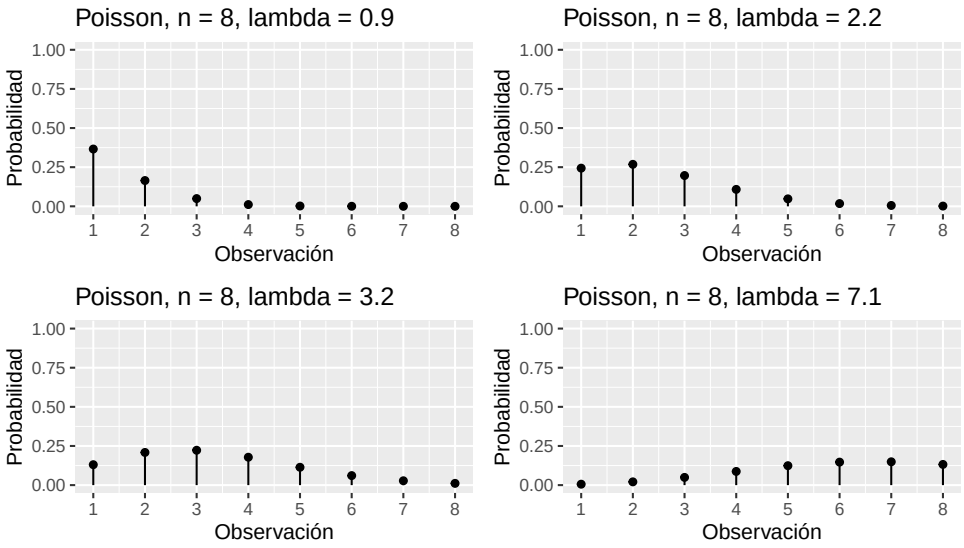


Figura 2.6: Simulación de la distribución Poisson

Ejemplo 2.14. 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 2.12, 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 = 2, # 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] 0.5768099
```

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 2.15. 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 2.23. 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. ¿Cuál es la probabilidad de que la compañía de seguros reciba como máximo 7 reclamaciones en un día?

3. ¿Cuál es la probabilidad de que la compañía de seguros reciba cuando menos 2 reclamaciones en un día?
4. ¿Cuál es la probabilidad de que la compañía de seguros reciba entre 2 y 5 reclamaciones en un día?
5. ¿Cuál es la probabilidad de que la compañía de seguros reciba de 1 a 3 reclamaciones en un día?

Ejercicio 2.24. 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 periodo de 90 minutos?
2. ¿Cuál es la probabilidad de que el banco reciba más de 4 solicitudes de préstamo hipotecario en un periodo de 80 minutos?
3. ¿Cuál es la probabilidad de que el banco reciba cuando mucho 5 solicitudes de préstamo hipotecario en un periodo de dos horas?

Ejercicio 2.25. 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 la tasa de enfermos que necesitan este medicamento es de 5 personas en promedio.

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 el número de enfermos sea de 3 a 6 por año?

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

Una diferencia fundamental entre las variables aleatorias discretas y las variables aleatorias continuas es cómo se calculan las probabilidades. En las variables aleatorias discretas la función de masa de probabilidad $P(X = x)$ da la probabilidad de que la variable aleatoria tome un valor determinado. En las variables aleatorias continuas, la contraparte de la función de probabilidad es la **función de densidad de probabilidad**, que se denota $f(x)$. La diferencia está en que la función de densidad de probabilidad no da probabilidades directamente. Si no que el área bajo la curva de $f(x)$ que corresponde a un intervalo determinado proporciona la probabilidad de que la variable aleatoria tome uno de los valores de ese intervalo. De manera que

cuando se calculan probabilidades de variables aleatorias continuas se calcula la probabilidad de que la variable aleatoria tome alguno de los valores dentro de un intervalo.

Como en cualquier punto determinado el área bajo la gráfica de $f(x)$ es cero, una de las consecuencias de la definición de la probabilidad de una variable aleatoria continua es que la probabilidad de cualquier valor determinado de la variable aleatoria es cero. (Anderson et al., 2008, página 227)

2.4.1. Función de densidad de probabilidad

Consideremos el siguiente ejemplo, supóngase que la variable X de interés es la profundidad de un lago en un punto sobre la superficie seleccionado al azar. Sea M = la profundidad máxima (en metros), así que cualquier número en el intervalo $[0, M]$ es un valor posible de X. Si se "discretiza" X midiéndola profundidad al metro más cercano, entonces los valores posibles son enteros no negativos menores que o iguales a M. La distribución discreta resultante de profundidad se ilustra con un histograma de probabilidad. Si se traza el histograma de modo que el área del rectángulo sobre cualquier entero posible k sea la proporción del lago cuya profundidad es (al metro más cercano) k, entonces el área total de todos los rectángulos es 1. En la figura 2.7a) aparece un posible histograma.

Si se mide la profundidad con mucho más precisión y se utiliza el mismo eje de medición de la figura 2.7a), cada rectángulo en el histograma de probabilidad resultante es mucho más angosto, aun cuando el área total de todos los rectángulos sigue siendo 1. En la figura 2.7b) se ilustra un posible histograma; tiene una apariencia mucho más regular que el histograma de la figura 2.7a). Si se continúa de esta manera midiendo la profundidad más y más finamente, la secuencia resultante de histogramas se aproxima a una curva más regular, tal como la ilustrada en la figura 2.7c). Como en cada histograma el área total de todos los rectángulos es igual a 1, el área total bajo la curva regular también es 1. La probabilidad de que la profundidad en un punto seleccionado al azar se encuentre entre a y b es simplemente el área bajo la curva regular entre a y b. Es de manera exacta una curva regular del tipo ilustrado en la figura 2.7c) la que especifica un distribución de probabilidad continua.

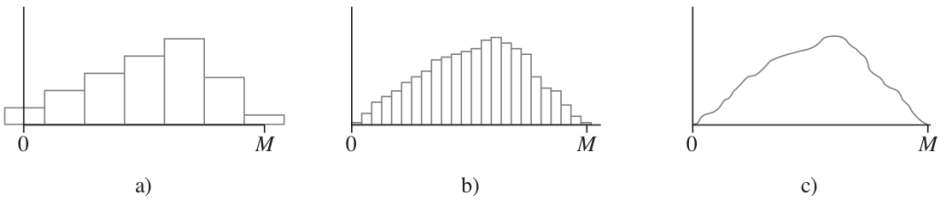


Figura 2.7: Histogramas de profundidad.

En este sentido, se obtiene la siguiente definición. Sea X una variable aleatoria continua. Entonces, una **función de densidad de probabilidad** (fdp) de X es una función $f(x)$ tal que para dos números cualesquiera a y b con $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Es decir, la probabilidad de que X asuma un valor en el intervalo $[a, b]$ es el área sobre este intervalo y bajo la gráfica de la función de densidad, como se ilustra en la figura 2.8. La gráfica de $f(x)$ a menudo se conoce como curva de densidad.

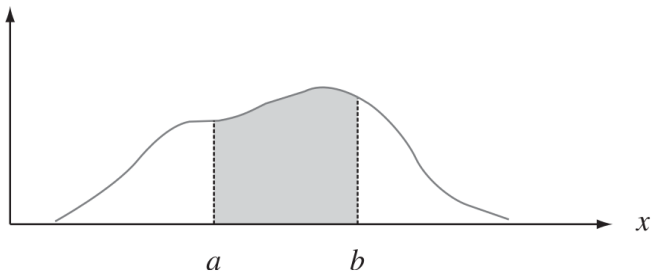


Figura 2.8: Área debajo de la curva de densidad.

Al igual que una distribución de masa de probabilidad, se deben cumplir condiciones para que la función de densidad de probabilidad sea legítima. Las condiciones son:

1. $f(x) \geq 0, \forall x$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Ejemplo 2.16. "Intervalo de tiempo" en el flujo de tránsito es el tiempo transcurrido entre el tiempo en que un carro termina de pasar por un punto fijo y el instante en que el siguiente carro comienza a pasar por ese punto. Sea $X =$ el intervalo de tiempo de dos carros consecutivos seleccionados al azar en una autopista durante un periodo de tráfico intenso. La siguiente función de densidad de probabilidad de X es en esencia el sugerido en "The Statistical Properties of Freeway Traffic" (*Transp. Res.* vol. 11: 221-228):

$$f(x) = 0.15e^{-0.15(x-0.5)},\ x \geq 0.5$$

La gráfica de $f(x)$ se da en la figura 2.9; no hay ninguna densidad asociada con intervalos de tiempo de menos de 0.5 y la densidad del intervalo decrece con rapidez (exponencial) a medida que x se incrementa a partir de 0.5. Claramente, $f(x) \geq 0$; para demostrar que la integral en todo el dominio de la función es igual a 1, se utiliza el siguiente resultado.

$$\int_a^{\infty} e^{-kx}dx = \frac{1}{k}e^{-ka}$$

Entonces,

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_{0.5}^{\infty} 0.15e^{-0.15(x-0.5)}dx = 0.15e^{0.075}\int_{0.5}^{\infty} e^{-0.15x}dx \\ &= 0.15e^{0.075}\frac{1}{0.15}e^{-0.075} = 1\end{aligned}$$

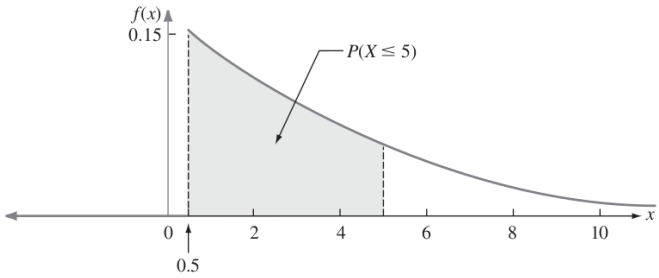


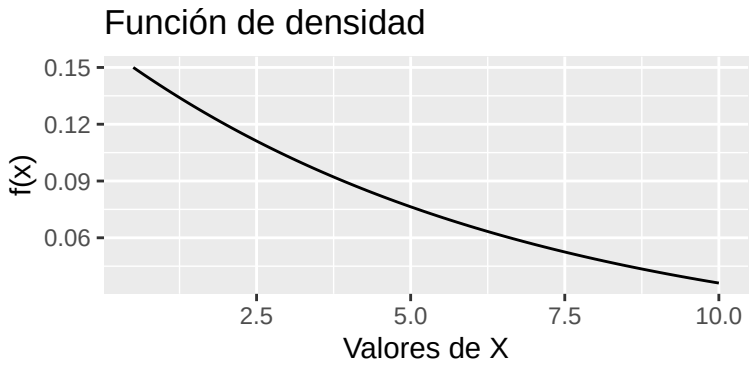
Figura 2.9: Curva de densidad del intervalo de tiempo del ejemplo 2.17.

En R, la gráfica de está función de densidad puede replicarse de la siguiente manera.

```r
fdp = function(x){
  f = 0.15*exp(-0.15*(x-0.5)) # Expresión de la función de
 ↪  densidad
  return(f)
}

x = seq(0.5, 10, by = 0.01) # Valores del dominio
f = fdp(x) # Valores del recorrido
df = data.frame("x" = x, "f" = f) # Data frame para poder usar
 ↪  ggplot

ggplot(data = df, aes(x = x, y = f)) + geom_line() +
  labs(title = "Función de densidad", x = "Valores de X", y =
   ↪  "f(x)")
```



Así, la probabilidad de que el intervalo de tiempo sea cuando mucho de 5 segundos es

$$\begin{aligned} P(X \leq 5) &= \int_{-\infty}^{5} f(x)dx = \int_{0.5}^{5} 0.15e^{-0.15(x-0.5)}dx \\ &= 0.15e^{0.075}\int_{0.5}^{5} e^{-0.15x}dx = 0.15e^{0.075}\frac{-1}{0.15}e^{-0.15x}\Big|_{x=0.5}^{x=5} \\ &= e^{0.075}\left(e^{-0.75} + e^{-0.075}\right) = 1.078(-0.472 + 0.928) = 0.491 \end{aligned}$$

En R para hacer el cálculo de la integral es posible utilizar el comando **integrate**:

```r
# Haciendo uso de la función utilizada para el gráfico
integrate(fdp, lower = 0.5, upper = 5)
```

```
## 0.4908436 with absolute error < 5.4e-15
```

Ejercicio 2.26. Sea X la cantidad de tiempo durante la cual un libro puesto en reserva durante dos horas en la biblioteca de una universidad es solicitado en préstamo por un estudiante seleccionado y suponga que X tiene la función de densidad

$$f(x) = 0.5x,\ 0 \leq x \leq 2$$

Calcule las siguientes probabilidades (manualmente y en R):

a. $P(X \leq 1)$.

b. $P(0.5 \leq X \leq 1.5)$.

c. $P(1.5 < X)$.

Ejercicio 2.27. El error implicado al hacer una medición geográfica computarizada es una variable aleatoria continua X con función de densidad de probabilidad

$$f(x) = 0.09375(4 - x^2),\ -2 \leq x \leq 2$$

a. Bosqueje la gráfica de $f(x)$.

b. Calcule $P(X > 0)$.

c. Calcule $P(-1 < X < 1)$.

d. Calcule $P(X < 0.5 \text{ o } X > 0.5)$.

Ejercicio 2.28. Un profesor universitario nunca termina su disertación antes del final de la hora y siempre termina dentro de 2 minutos después de la hora. Sea $X =$ el tiempo que transcurre entre el final de la hora y el final de la disertación y suponga que la función de densidad de probabilidad de X es

$$f(x) = kx^2,\ 0 \leq x \leq 2$$

a. Determine el valor de k y trace en R la curva de densidad correspondiente.

b. ¿Cuál es la probabilidad de que la disertación termine dentro de un minuto del final de la hora?

c. ¿Cuál es la probabilidad de que la disertación continúe después de la hora durante entre 60 y 90 segundos.

d. ¿Cuál es la probabilidad de que la disertación continúe durante por lo menos 90 segundos después del final de la hora?

2.4.2. Función de distribución acumulada

La función de distribución acumulada $F(x)$ de una variable aleatoria discreta X da, con cualquier número especificado x, la probabilidad $P(X \leq x)$. Se obtiene sumando la función masa de probabilidad $p(y)$ a lo largo de todos los valores posibles y que satisfacen $y \leq x$. La función de distribución acumulada de una variable aleatoria continua da las mismas probabilidades $P(X \leq x)$ y se obtiene integrando la función de densidad de probabilidad $f(y)$ entre los límites $-\infty$ y x.

La función de distribución acumulada $F(x)$ de una variable aleatoria continua X se define para todo número x como

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Con cada x, $F(x)$ es el área bajo la curva de densidad a la izquierda de x. Esto se ilustra en la figura 2.10, donde $F(x)$ se incrementa con regularidad a medida que x se incrementa.

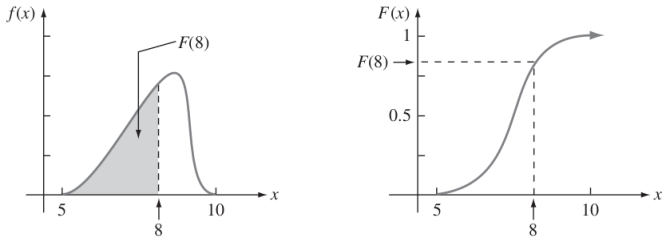


Figura 2.10: Función de densidad de probabilidad y distribución acumulada asociada

Ejemplo 2.17. Suponga que la función de densidad de probabilidad de la magnitud X de una carga dinámica sobre un puente (en newtons) está dada por

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \leq x \leq 2 \\ 0 & \text{en otro caso} \end{cases}$$

Para cualquier número x entre 0 y 2,

$$F(x) = \int_{-\infty}^{x} f(y)dy = \int_{0}^{x} \left(\frac{1}{8} + \frac{3}{8}y\right) dy = \frac{x}{8} + \frac{3}{16}x^2$$

Por lo tanto

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{8} + \frac{3}{16}x^2 & 0 \leq x \leq 2 \\ 1 & x > 2 \end{cases}$$

La importancia de la función de distribución acumulada en este caso, lo mismo que para variables aleatorias discretas, es que las probabilidades de varios intervalos pueden ser calculadas con una fórmula o una tabla de $F(x)$. En el caso de una variable aleatoria continua X con función de densidad de probabilidad $f(x)$ y función de distribución acumulada $F(x)$, se tiene que con cualquier número a,

$$P(X > a) = 1 - F(a)$$

y para dos números cualesquiera a y b con $a < b$.

$$P(a \leq X \leq b) = F(b) - F(a)$$

La figura 2.11 ilustra la probabilidad deseada es el área sombreada bajo la curva de densidad entre a y b, que es igual a la diferencia entre las dos áreas sombreadas acumulativas. Esto es diferente de lo que es apropiado para una variable aleatoria discreta de valor entero.

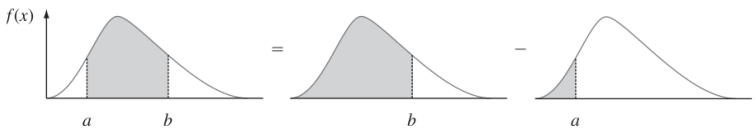


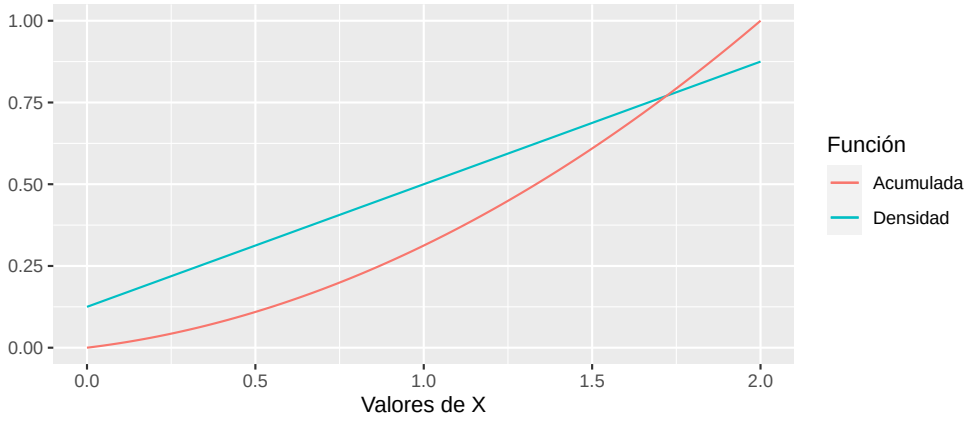
Figura 2.11: Cálculo de probabilidades acumulativas.

Ejemplo 2.18. Continuando con el ejemplo 2.17. Las gráficas de $f(x)$ y $F(x)$ son

```
densidad = function(x){ # Función de densidad
  return(1/8+3/8*x)
}
acumulada = function(x){ # Función de distribución acumulada
  return(x/8+3/16*x^2)
}

df = data.frame("x" = seq(0,2,0.01), # Valores de X
                "fx" = densidad(seq(0,2,0.01)), # Valores de
                 ↪ la densidad
                "Fx" = acumulada(seq(0,2,0.01))) # Valores
                 ↪ acumulados

# Gráfico de líneas de ambas funciones
ggplot(data = df) +
  geom_line(aes(x = x, y = fx, color = "Densidad")) +
  geom_line(aes(x = x, y = Fx, color = "Acumulada")) +
  labs(color = "Función", x = "Valores de X", y = "")
```



La probabilidad de que la carga esté entre 1 y 1.5 es

$$P(1 \leq X \leq 1.5) = F(1.5) - F(1)$$

Utilizando la función del gráfico para la función de distribución acumulada, el resultado es

```
acumulada(1.5) - acumulada(1)
```

```
## [1] 0.296875
```

Una vez que se obtiene la función de distribución acumulada, cualquier probabilidad que implique X es fácil de calcular sin cualquier integración adicional.

Ejercicio 2.29. Sea X la cantidad de tiempo durante la cual un libro puesto en reserva durante dos horas en la biblioteca de una universidad es solicitado en préstamo por un estudiante seleccionado. La función de distribución acumulativa del tiempo de préstamo X es

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x^2}{4} & 0 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Use esta para calcular lo siguiente:

a. $P(X \leq 1)$
b. $P(0.5 \leq X \leq 1)$
c. $P(X > 0.5)$

d. $f(x)$
e. Grafique $f(x)$ y $F(x)$.

Ejercicio 2.30. El error de medición de un proceso de control de gestión en la peligrosidad de residuos está dado por la siguiente función de distribución acumulada.

$$F(x) = \begin{cases} 0 & x < -2 \\ \frac{1}{2} + \frac{3}{32}\left(4x - \frac{x^3}{3}\right) & -2 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

a. Calcule $P(X < 0)$.
b. Calcule $P(-1 < X < 1)$.
c. Calcule $P(0.5 < X)$.
d. Calcule $f(x)$.
e. Grafique $f(x)$ y $F(x)$.

Ejercicio 2.31. El ejemplo 2.16 introdujo el concepto de intervalo de tiempo en el flujo de tránsito y propuso una distribución particular para $X =$ el intervalo de tiempo entre dos carros consecutivos seleccionados al azar (s). Suponga que en un entorno de tránsito diferente, la distribución del intervalo de tiempo tiene la forma

$$f(x) = \begin{cases} \frac{k}{x^4} & x > 1 \\ 0 & x \leq 1 \end{cases}$$

a. Determine el valor de k con el cual $f(x)$ es una función de densidad de probabilidad legítima.
b. Obtenga la función de distribución acumulada.
c. Use la función de distribución acumulada de (b) para determinar la probabilidad de que el intervalo de tiempo exceda de 2 segundos y también la probabilidad de que el intervalo esté entre 2 y 3 segundos.
d. Grafique la función de densidad de probabilidad y la función de distribución acumulada.

2.4.3. Distribuciones

A continuación, se dan a conocer algunas de las distribución de probabilidad continua más utilizadas. Cabe mencionar, que existen muchas otras distri-

buciones, por lo que se invita al estudiante a informarse de ellas en caso de que lo requiera.

2.4.3.1. Uniforme

Esta es una versión continua de la ya vista en modelos discretos, la diferencia radica en que la variable es del tipo continua valga la redundancia. La función de probabilidad asociada es:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } x \in (a,b) \\ 0 & \text{si } x \notin (a,b) \end{cases}$$

Notación: $U(a,b)$

Ejemplo 2.19. Imagine que es un analista financiero encargado de analizar los rendimientos diarios de una acción en particular en el mercado de valores. Los rendimientos diarios de esta acción se distribuyen de manera uniforme continua entre $-2\,\%$ y $2\,\%$, es decir, $U(a=-2, b=2)$.

En este caso, la distribución uniforme continua significa que cada valor dentro del rango dado tiene la misma probabilidad de ocurrir. Por lo tanto, la probabilidad de que el rendimiento diario de la acción sea cualquier valor entre $-2\,\%$ y $2\,\%$ es igual.

Supongamos que estás interesado en calcular la probabilidad de que el rendimiento diario de la acción sea mayor o igual a $1.2\,\%$. Para hacerlo, primero debemos calcular el ancho del rango en el cual el rendimiento se encuentra dentro o por encima del $1.2\,\%$. En este caso, el ancho del rango es $2\,\% - 1.2\,\% = 0.8\,\%$.

La probabilidad de que el rendimiento diario sea mayor o igual a $1.2\,\%$ es igual al ancho del rango ($0.8\,\%$) dividido por el ancho total de la distribución ($4\,\%$ en este caso), lo cual resulta en $0.8\,\%/4\,\% = 0.2$.

La figura 2.12, muestra una simulación de la función de densidad de probabilidad de la distribución uniforme continua.

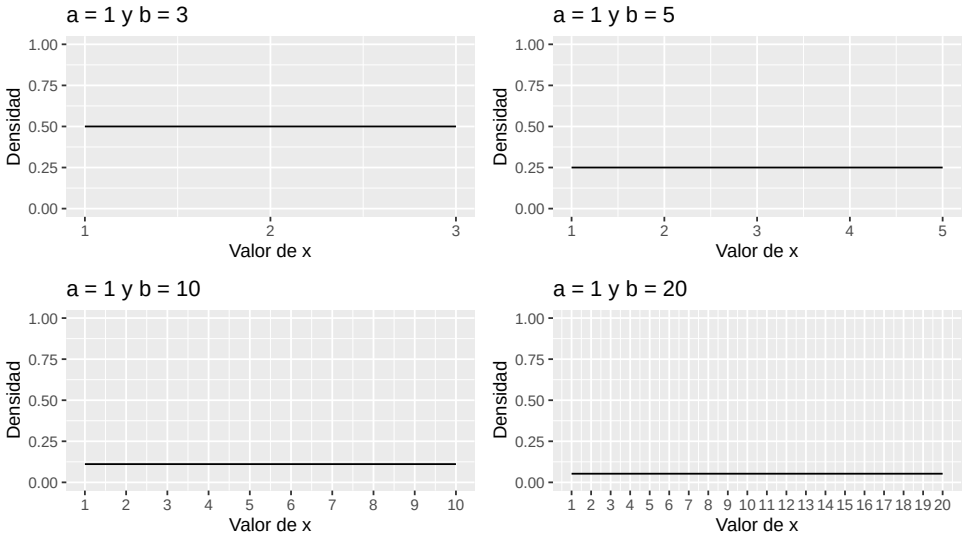


Figura 2.12: Simulación de la distribución Uniforme Continua

2.4.3.2. Exponencial

Esta es una de las pocas variables aleatorias continuas que se aplica a un contexto de determinado. Se utiliza para modelar tiempos de espera para la ocurrencia de un determinado evento (éxito). **¿Qué similitudes y diferencias tiene con la variable discreta Poisson?**.

La función de densidad asociada es:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

con $\lambda > 0$. ¿Qué interpretación tiene lambda?

Notación: Exp(λ)

La figura 2.13, muestra una simulación de la función de densidad de probabilidad de la distribución exponencial.

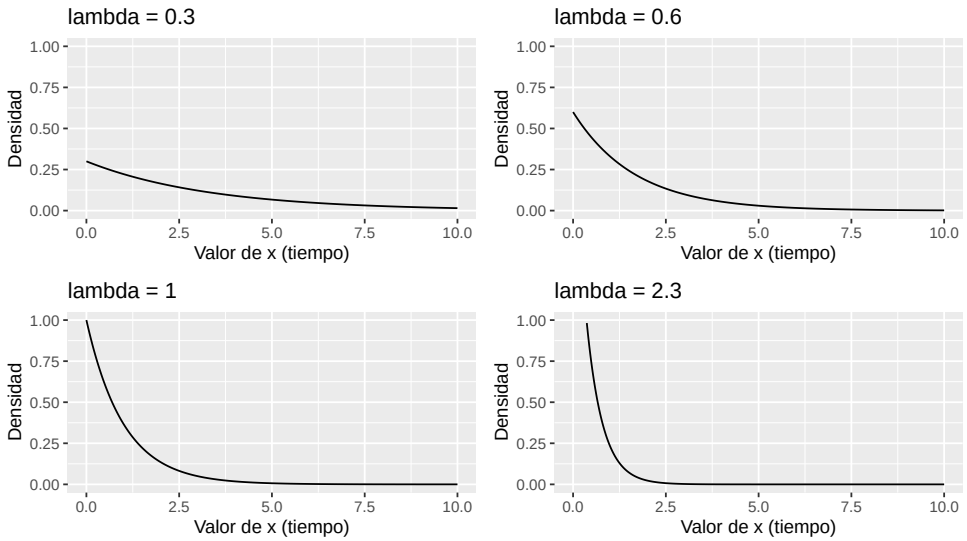


Figura 2.13: Simulación de la distribución Exponencial

Ejemplo 2.20. Se está analizando el tiempo de espera de los clientes en una tienda y se desea calcular la probabilidad de que un cliente tenga que esperar más de 15 minutos antes de ser atendido. Suponiendo que el tiempo promedio de espera es de 10 minutos, es posible utilizar la distribución exponencial para calcular esta probabilidad.

Lambda se calcula como el el recíproco del tiempo promedio, es decir, $\lambda = \frac{1}{10}$. Luego, se desea calcular

$$P(X > 15)$$

En R:

```
pexp(15,
     rate = 1/10, # Lambda
     lower.tail = F)
```

```
## [1] 0.2231302
```

En la tabla 2.4 se muestran los comandos para calcular probabilidades asociadas a esta distribución.

Tabla 2.4: Cálculo de probabilidades de la distribución exponencial en R.

Probabilidad	Comando
$P(X \leq x)$ o $P(X < x)$	pexp(q = , rate =)
$P(X \geq x)$ o $P(X > x)$	pexp(q = , rate = , lower.tail = F)
$P(X \leq q) = p$	qexp(p = , rate =)

Ejercicio 2.32. Un componente eléctrico tiene una vida útil media de 8 años. Si su vida útil se distribuye en forma exponencial. ¿Cuál debe ser el tiempo x de garantía que se debe otorgar, si se desea reemplazar a lo más el 15 % de los componentes que fallen dentro de este período?

Ejercicio 2.33. Las personas que solicitan créditos en un banco son sometidas a un estudio de morosidad. En general, el tiempo promedio de una persona para volverse morosa es de 4 años después de solicitar un crédito. Si el momento en que una persona se vuelve morosa distribuye de forma exponencial con x en años, ¿cuál es la probabilidad de que una persona se vuelva morosa posterior al quinto año después de solicitar un crédito?

Ejercicio 2.34. Las rutas de reparto de pedidos de Amazon en Chile son modificadas en promedio cada año. Si el tiempo de actualización de las rutas distribuye de forma exponencial con x en años, ¿cuál es la probabilidad de que las rutas de reparto se actualicen al cabo de 13 meses?

2.4.3.3. Normal

Esta es una de la variables continuas más usadas. Lamentablemente, no existe un característica fenomenológica que permita deducir cuando es adecuado utilizar esta variable. La función de desindad de probabilidad asociada es la siguiente.

$$f(x) = (2\pi\sigma^2)^{-1/2} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, x \in R$$

Notación: $N(\mu, \sigma^2)$

La figura 2.14, muestra una simulación de la función de densidad de probabilidad de la distribución normal.

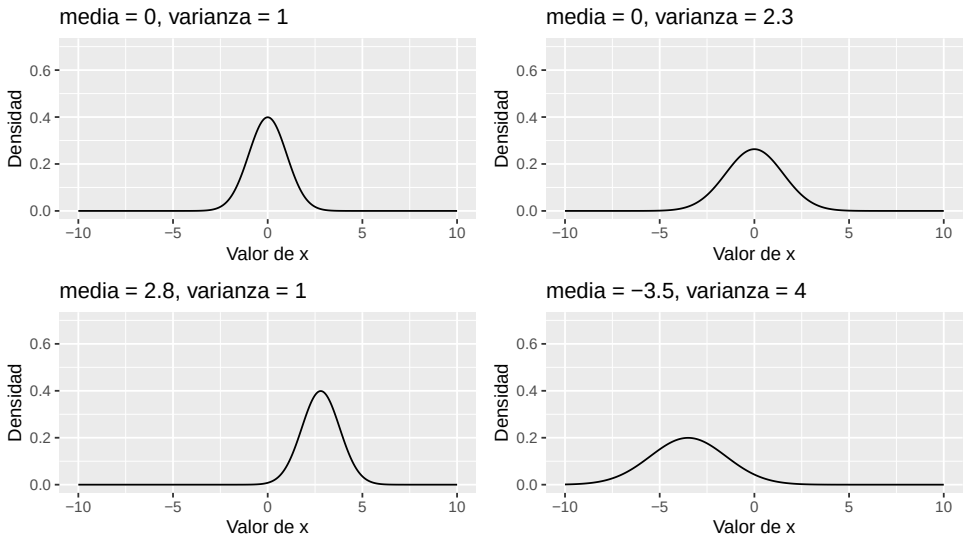


Figura 2.14: Simulación de la distribución Normal

Ejemplo 2.21. Se está analizando los ingresos mensuales de una empresa y se desea calcular la probabilidad de que los ingresos estén por debajo de \$6000 miles de dólares. Además, se sabe que los ingresos mensuales siguen una distribución normal con una media de \$5000 (miles de dólares)y una desviación estándar de \$1000 (miles de dólares).

La probabilidad a calcular es $P(X < 6000)$. En R:

```
pnorm(q = 6000, mean = 5000, sd = 1000)
```

```
## [1] 0.8413447
```

En la tabla 2.5 se muestran los comandos para calcular probabilidades asociadas a esta distribución.

Tabla 2.5: Cálculo de probabilidades de la distribución normal en R.

Probabilidad	Comando
$P(X \leq x)$ o $P(X < x)$	pnorm(q = , mean = , sd =)
$P(X \geq x)$ o $P(X > x)$	pnorm(q = , mean = , sd = , lower.tail = F)
$P(X \leq q) = p$	qnorm(p = , mean = , sd =)

Ejercicio 2.35. En un bar se ha instalado una máquina automática para la venta de cerveza. La máquina puede regularse de modo que la cantidad media de cerveza por vaso sea la que se desea. Además, se sabe que el proceso de llenado sigue una distribución normal, y que independiente de la cantidad a llenar la desviación estándar es 5.9 ml. Si el nivel medio se ajusta a 304.6 ml determine qué porcentaje de los vasos contendrá menos de 295.7 ml.

Ejercicio 2.36. Cada cierto periodo de tiempo, el Banco Central de Chile realiza una inyección de capital al mercado chileno, con el fin de mermar el efecto de la inflación. El dinero que destina el banco, sigue un distribución normal con media 15 (miles de millones) y una varianza de 4 (miles de millones). ¿Cuál es probabilidad el banco inyecte más de 15.6 miles de millones de pesos?

Ejercicio 2.37. La mayoría de las transacciones electrónicas en Chile están a cargo de la empresa Transbank. Además, se sabe que la cantidad de transacciones realizadas por Transbank sigue un distribución normal con una media de 20 millones.

- Si la varianza de las transacciones es de 25 millones, ¿cuál es la probabilidad de que Transbank esté a cargo de más de 30 millones de transacciones?
- Si la varianza de las transacciones es de 16 millones, ¿cuántas deben ser las transacciones para que la probabilidad de que Transbank se haga cargo de un número mayor sea de un 34 %?

Ejercicio 2.38. El tipo de cambio del euro ha fluctuado de manera similar al dólar en los últimos 4 años. Si valor observado del euro sigue un distribución normal con media \$912 y varianza \$144.

- ¿Cuál es la probabilidad de que el tipo de cambio del euro supere los

$920?

- ¿A cuánto debe estar el tipo de cambio para que la probabilidad de que el euro supere ese valor sea de un 21.9 %?

Ejercicio 2.39. Los residuos de vuelo de una determinad aerolínea son recolectados al final de cada semana, generando en promedio 48 toneladas de residuos. Además, la cantidad de residuos generados por los vuelos sigue una distribución normal.

- Si la varianza de los residuos generados es de 100, ¿cuál es la probabilidad de que la aerolínea genere más de 50 toneladas de residuos?
- Si la varianza de los residuos generados es de 120, ¿cuántos deben ser los residuos para que la probabilidad de que la aerolínea genere menos sea de un 87 %?

Ejercicio 2.40. El sistema logístico de lavado de autos de una determinada empresa ha optado por instalar una máquina automática para la atención de clientes. Se sabe que la cantidad de clientes que son atendidos por la máquina sigue una distribución normal con una varianza de 90 personas.

- Si el promedio de personas atendidas por la máquina es de 390, ¿cuál es la probabilidad de que la máquina atienda a menos de 385 personas?
- Si el promedio de personas atendidas por la máquina son de 360, ¿cuántas personas como mínimo que deben ir a la máquina para que la posibilidad de que sean atendidas sea de un 70 %?

2.4.3.4. T - Student y Ji - Cuadrado

Estas dos variables, al igual que la Normal son muy utilizadas. Sin embargo, la función de densidad de cada una de ellas es más compleja y engorrosa.

Los parámetros de cada una de ellas son lo siguientes:

- t - Student: $t(v)$, donde v son los denominados grados de libertad.
- χ^2: $\chi^2(v)$, donde v son los denominados grados de libertad.

La distribución t - Student y la Normal son muy parecidas cuando el tamaño de la muestra es grande. Además, en R se tiene que, la distribución t - Student siempre tiene media 0 y varianza fija > 1 (una especie de parámetros fijos).

La figura 2.15, muestra una simulación de la función de densidad de probabilidad de la distribución t - Student

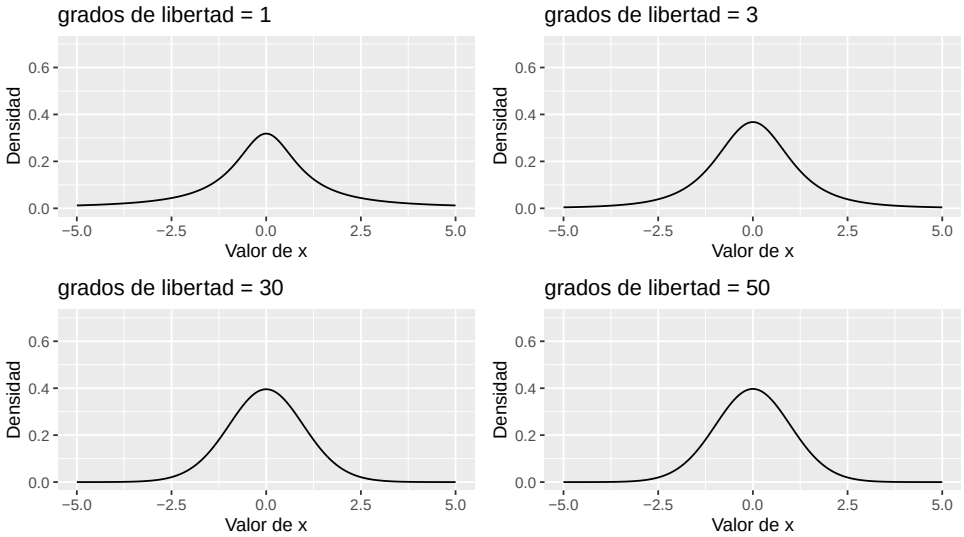


Figura 2.15: Simulación de la distribución t-Student

La figura 2.16, muestra una simulación de la función de densidad de probabilidad de la distribución Ji-Cuadrado

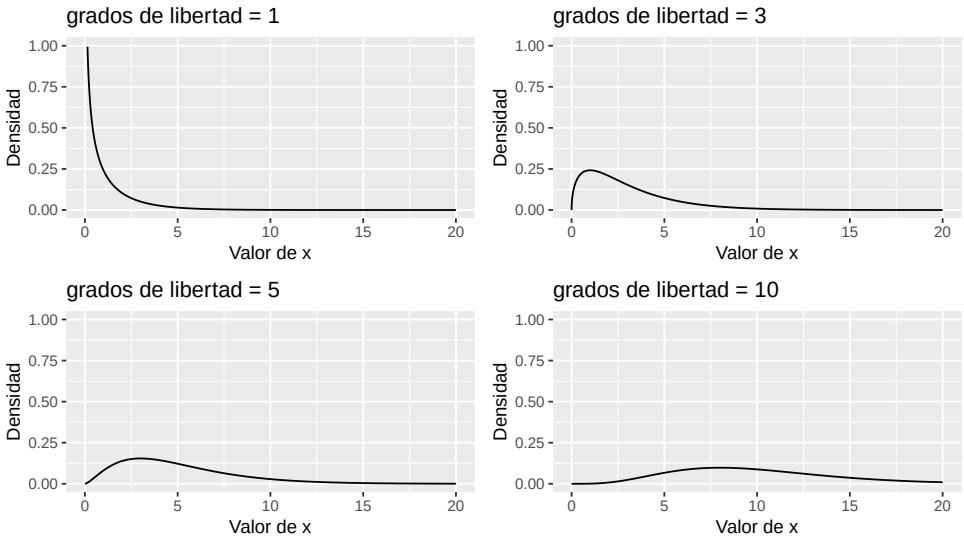


Figura 2.16: Simulación de la distribución Ji-Cuadrado

En la tabla 2.6 se muestran los comandos para calcular probabilidades asociadas a estas distribuciones.

Tabla 2.6: Cálculo de probabilidades de la distribución t-student y Ji-Cuadrado en R.

Probabilidad	T-Student	Ji-Cuadrado
$P(X \leq x)$ o $P(X < x)$	pt(q = , df =)	pchisq(q = , df =)
$P(X \geq x)$ o $P(X > x)$	pt(q = , df = , lower.tail = F)	pchisq(q = , df = , lower.tail = F)
$P(X \leq q) = p$	qt(p = , df =)	qchisq(p = , df =)

Ejercicio 2.41. Durante los últimos años, la cantidad de de personas que han solicitado avances en efectivo en cajeros automáticos ha ido en aumento. Si la cantidad de solicitudes de avances (en miles) siguen un distribución χ^2 con 34 grados de libertad, ¿cuál es probabilidad de que para este año la cantidad de avances realizados en cajero sea menor a 45 mil?

Ejercicio 2.42. La cantidad de fraudes financieros han ido en aumento a medida que la tecnología avanza. Si la cantidad de fraudes (en miles) sigue una distribución t - Student con 51 grados de libertad, ¿cuál es probabilidad de que para este año la cantidad de fraudes sea mayor a 0.7 mil?

Ejercicio 2.43. La empresa ROSEN tiene un sistema automático para el relleno de almohadas con plumas. Si el volumen de relleno de plumas de las almohadas sigue un distribución t - Student con 200 grados de libertad, ¿cuál es la probabilidad de que una almohada sea rellenada con menos de 0.5 mil plumas?

Por otro lado, la tabla 2.7 muestra un resumen de los comandos en R para los distintos casos de cálculo de valores en la función de densidad con distribuciones continuas.

Tabla 2.7: Cálculo de densidades de distribuciones continuas en R.

Densidad	Normal	Exponencial	T-Student	Ji-Cuadrado
$f(x)$	dnorm(x = , mean = , sd =)	dexp(x = , rate =)	dt(x = , df =)	dchisq(x = , df =)

2.5. Esperanza

La esperanza, esperanza matemática o valor medio, es un concepto que se aplica cuando un experimento es realizado muchas veces. Consideremos un dado honesto, es decir, la probabilidad de que cada una de las caras salga seleccionada en cada tiro es de 1/6. Si registramos cada uno de los resultados y luego los promediamos, ese resultado es el resultado final promedio. Sin embargo, nosotros podemos estar interesados en conocer ese valor sin la necesidad de tener que tirar el dado muchas veces, para ello, existen fórmulas que permiten calcular el valor promedio de los resultados al realizar el experimento muchas veces.

Si se trabaja con una variable aleatoria discreta, la esperanza matemática está dada por la siguiente expresión

$$\mu = E(X) = \sum_{i=1}^{n} x_i P(X = x_i)$$

Por otro lado, para una variable aleatoria continua la expresión es

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Ejemplo 2.22. Consideremos el lanzamiento del dado, del cual se habló anteriormente. La siguiente tabla, refleja la probabilidad de cada uno de los eventos

X	$P(X = x_i)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Haciendo uso de la fórmula de esperanza agregamos los valores de la expresión $x_i f(x_i)$.

X	$P(X = x_i)$	$x_i P(X = x_i)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	1

Finalmente, se deben sumar los valores de la tercera columna, obteniéndose como resultado

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i) = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 21/6 = 3.5$$

En base a este resultado, si una persona cobra 3 mil pesos por cada lanzamiento del dado, y devuelve 4 mil cada vez que el jugador obtiene un numero mayor a 2, ¿es rentable jugar?, ¿por qué?

Ejemplo 2.23. Sea X una variable aleatoria con distribución

$$P(X = x) = p(1-p)^{x-1}, x = 1, 2, ...$$

Determine $E(X)$.

$$\begin{aligned}
E(X) &= \sum_{i=1}^{\infty} x_i P(X = x_i) = \sum_{x=1}^{\infty} xP(X = x) = \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\
&= p\sum_{x=1}^{\infty} x(1-p)^{x-1} = p\sum_{x=1}^{\infty} \frac{\partial}{\partial p}\left(-(1-p)^x\right) \\
&= -p\frac{\partial}{\partial p}\left(\sum_{x=1}^{\infty}(1-p)^x\right) = -p\frac{\partial}{\partial p}\left(\sum_{x=0}^{\infty}(1-p)^x - 1\right) \\
&= -p\frac{\partial}{\partial p}\left(\frac{1}{1-(1-p)} - 1\right) = -p\frac{\partial}{\partial p}\left(\frac{1-p}{p}\right) \\
&= -p\frac{-p-(1-p)}{p^2} = \frac{1}{p}
\end{aligned}$$

Si X es una variable aleatoria y a, b y c son valores reales constantes, entonces, se tiene las siguientes propiedades de la espranza:

1. $E(k) = k$ si k es una constante.
2. $E(kX) = kE(X)$.
3. $E(aX + b) = aE(X) + b$.
4. Sea X una v.a.d, entonces, $E(h(X)) = \sum_{i=1}^{n} h(x_i)P(X = x_i)$.
5. Sea X una v.a.c, entonces, $E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx$.

Ejercicio 2.44. Un juego de azar consiste en lanzar dos dados y ganar tantos miles de pesos como valga la suma de las puntuaciones de los dados. ¿Cual es la cantidad media ganada cada vez que juego? Si me cobran 5 mil pesos por cada juego, ¿es interesante participar en este juego?

Ejercicio 2.45. Para una variable aleatoria discreta X la distribución de probabilidad viene dada por

$$P(X = x) = \begin{cases} kx & x = 1, 2, 3, 4, 5 \\ k(10 - x) & x = 6, 7, 8, 9 \end{cases}$$

Hallar el valor de la constante k y $E(X)$.

Ejercicio 2.46. (Verificación con R) Considere la variable aleatoria que representa el número que pudiese salir en una ruleta (del 1 al 36). Determine el valor medio del experimento asociado a jugar en la ruleta. Verifique su resultado generando 10, 20, 30, 100, 200, 500, 1000, 2000 y 5000 números aleatorios enteros (entre 1 y 36, inclusive ambos) en R, realice un gráfico de puntos con líneas entre la cantidad de números aleatorios generados y el promedio de ellos. Comente lo observado.

Ejercicio 2.47. Un distribuidor de enseres para el hogar vende tres modelos de congeladores verticales de 13.5, 15.9 y 19.1 pies cúbicos de espacio de almacenamiento, respectivamente. Sea $X =$ la cantidad de espacio de almacenamiento adquirido por el siguiente cliente que compre un congelador. Suponga que X tiene la función masa de probabilidad

x	13.5	15.9	19.2
$p(x)$	0.2	0.5	0.3

1. Calcule $E(X)$.
2. Calcule $E(X^2)$.

Ejercicio 2.48. Se selecciona al azar un individuo que tiene asegurado su automóvil con una compañía. Sea Y el número de infracciones de tránsito por las que el individuo fue citado durante los últimos 3 años. La función masa de probabilidad de Y es

y	0	1	2	3
$p(y)$	0.6	0.25	0.1	0.05

1. Calcule $E(Y)$.
2. Suponga que un individuo con Y infracciones incurre en un recargo de $\$100Y^2$. Calcule la cantidad esperada del recargo.

Ejercicio 2.49. Una barra de 12 pulgadas que está sujeta por ambos extremos se somete a una cantidad creciente de esfuerzo hasta que se rompe. Sea $Y =$ la distancia del extremo izquierdo al punto donde ocurre la ruptura. Suponga que Y tiene la función de densidad de probabilidad

$$f(y) = \begin{cases} \frac{y}{24}\left(1 - \frac{y}{12}\right) & 0 \le y \le 12 \\ 0 & \text{en otro caso} \end{cases}$$

Calcule $E(Y)$ y $E(Y^2)$.

Ejercicio 2.50. El tiempo X para la terminación de cierta tarea tiene una función de distribución acumulativa $F(x)$ dada por

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^3}{3} & 0 \le x < 1 \\ 1 - \frac{1}{2}\left(\frac{7}{3} - x\right)\left(\frac{7}{4} - \frac{3x}{4}\right) & 1 \le x \le \frac{7}{3} \\ 1 & x \ge \frac{7}{3} \end{cases}$$

Calcule $E(X)$ y $E(X^2)$.

2.6. Varianza

Al igual como se abordó en la unidad 1, la varianza mide la dispersión cuadrática de una variable respecto al promedio. Sin embargo la expresión que permite calcular la varianza de cualquier tipo de variable aleatoria es

$$\sigma^2 = Var(X) = E[(X-\mu)^2] \tag{2.12}$$

Utilizando las propiedades de la esperanza, es posible reducir la expresión (2.12) a una más sencilla.

$$\begin{aligned}
\sigma^2 = Var(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2X\mu + \mu^2] \\
&= E[X^2] - E[2X\mu] + E[\mu^2] \\
&= E[X^2] - 2\mu E[X] + \mu^2 \\
&= E[X^2] - 2\mu \cdot \mu + \mu^2 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - (E[X])^2 \\
\sigma^2 = Var(X) &= E[X^2] - (E[X])^2
\end{aligned} \tag{2.13}$$

Ejemplo 2.24. Suponga que se tiene una variable aleatoria continua X con función de densidad de probabilidad $f(x)$ dada por:

$$f(x) = 2x,\ 0 \leq x \leq 1$$

La media de X, se calcula como:

$$\mu = \int_0^1 x \cdot f(x)\,dx = \int_0^1 2x^2\,dx = \left[\frac{2}{3}x^3\right]\Big|_0^1 = \frac{2}{3}$$

La varianza de X se calcula como:

$$\sigma^2 = \int_0^1 (x-\mu)^2 \cdot f(x)\,dx$$

Sustituyendo $\mu = 2/3$, la fórmula se simplifica a:

$$\sigma^2 = \int_0^1 \left(x - \frac{2}{3}\right)^2 \cdot 2x \, dx$$

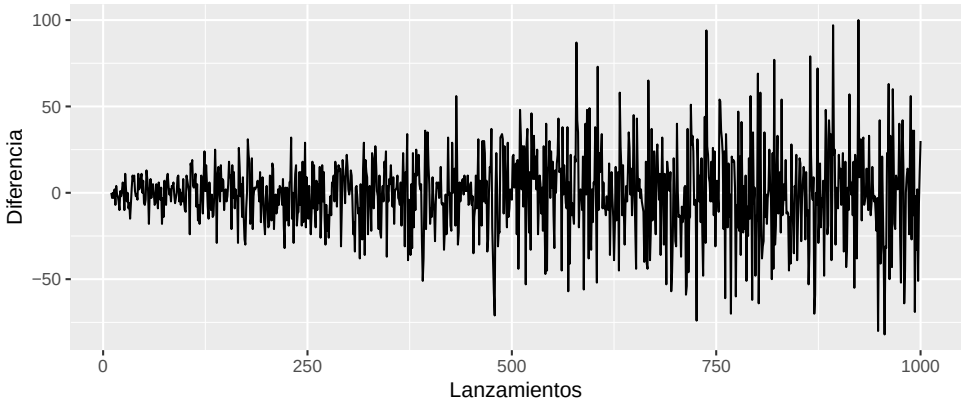
La solución de la integral para obtener la varianza es 1/18.

Ejemplo 2.25. (*La falacia del jugador*) La siguiente tabla muestra el registro de distintos experimentos asociados a tirar una moneda. En cada uno de ellos, se registra la cantidad de lanzamiento de la moneda, la cantidad de caras, la cantidad de sellos, la diferencia (número de caras - número de sellos) y la proporción (número de caras/número de sellos).

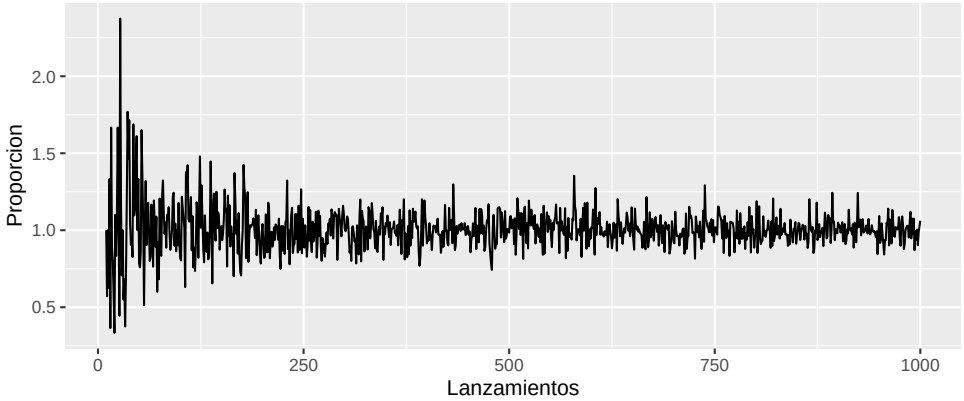
Tabla 2.8: Primeros 10 de 1000 instancias de lanzamientos de una moneda.

Lanzamientos	Caras	Sellos	Diferencia	Proporción
10	5	5	0	1.0000000
11	4	7	-3	0.5714286
12	6	6	0	1.0000000
13	5	8	-3	0.6250000
14	8	6	2	1.3333333
15	4	11	-7	0.3636364
16	10	6	4	1.6666667
17	8	9	-1	0.8888889
18	8	10	-2	0.8000000
19	7	12	-5	0.5833333

Fijémonos que ocurre si graficamos las diferencias entre caras y sellos según la cantidad de veces se lanza la moneda.



Por lo que se puede observar, la diferencia (amplitud vertical, variabilidad) aumenta a medida que también lo hace la cantidad de lanzamientos. Este resultado puede parecer extraño, ya que cuantas más veces se lanza la moneda, el números de caras y el de sellos debería tender a acercarse puesto que la cara y el sello son igual de probables. Sin embargo, para apreciar lo que realmente ocurre, se debe graficar las proporciones y no las diferencias.



Como se observa, la proporción de caras y sellos tiene a ser 1 cuántas más veces se arroje la moneda **(la diferencia también tenderá a cero, pero se necesitan muchos más intentos)**.

La conocida como falacia del jugador consiste en creer que, porque hayan salido de forma continuada un número de caras relativamente grande, en la siguiente jugada deberá salir sello puesto que los resultados deberán compensarse. Así, en una ruleta, si han salido 3 o 4 veces seguidas números de color rojo, pensar que en el siguiente movimiento de la ruleta es más probable que

salga negro es una falacia. Cada jugada es independiente de la anterior.

Otros planteamientos (incorrectos) equivalentes son: “Un resultado aleatorio tiene más probabilidades de ocurrir, si no ha ocurrido durante cierto periodo de tiempo”; o “Un resultado tiene menos probabilidades de ocurrir, si no ha ocurrido durante cierto periodo de tiempo”.

Ejercicio 2.51. Calcule la varianza de los siguientes ejercicios 2.45, 2.47, 2.48, 2.49 y 2.50. Realice el cálculo manualmente y confirme el resultado mediante R.

Ejercicio 2.52. Los ingresos de la empresa TechSolid están definidos por una variable aleatoria X (en miles de millones de dólares). Se sabe que la la función de densidad de probabilidad está dada por:

$$f(x) = \frac{1}{4}e^{-x/4}, x \geq 0$$

1. Determine el ingreso medio de la empresa.
2. Determine la varianza del ingreso de la empresa. Utilice expresamente $E[(X-\mu)^2]$.
3. Si los costos de la empresa se relacionan con los ingresos de la siguiente manera: $0.02X^2$. Determine los costos promedio de la empresa.

Ejercicio 2.53. Dada la siguiente función de densidad de probabilidad, determine la $Var(X)$.

$$f(x) = 4.5e^{-4.5x} \ , \ x \geq 0$$

Unidad 3

Distribuciones muestrales

Antes de iniciar esta sección es necesario aclarar el concepto de independencia. Para ello, consideremos dos variable aleatorias, y un evento de cada una, podemos entender la independencia como "la probabilidad de que ocurran los dos eventos al mismo tiempo es igual a la multiplicación de las probabilidades de cada evento por separado". Además, si las variables (que definen los eventos) tienen la misma función de distribución, se dice que las variables aleatorias son **independientes idénticamente distribuidas** (iid).

3.1. Distribución de muestreo de la media

3.1.1. Estandarización

La estandarización, es un proceso mediante el cual los datos originales de una variable se transforman en una nueva escala que tiene una media de cero y una desviación estándar de uno. Esto permite eliminar las diferencias de escala entre las variables y facilita la comparación y el análisis estadístico (Hair et al., 2013).

Consideremos X una variable aleatoria normalmente distribuida con media $E(X) = \mu$ y varianza $Var(X) = \sigma^2$. El proceso de estandarización de la variable aleatoria es:

$$\frac{X - \mu}{\sigma} \sim N(0, 1) \tag{3.1}$$

En el gráfico 3.1 se observa una ejemplificación de la estandarización. Al

estandarizar el punto más alto de la curva se centra al rededor del 0, a su vez que la curva se angosta.

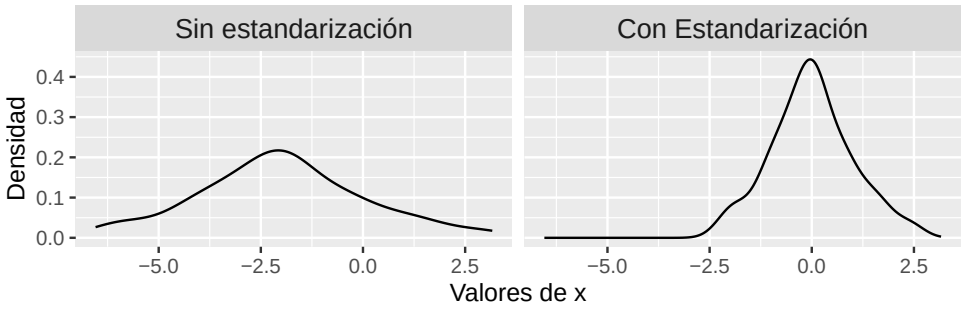


Figura 3.1: Estandarización de una variable con distribución Normal

3.1.2. Distribución de la media

Una de las medidas de resumen más importantes es la media (o promedio) de un conjunto de variables aleatorias independientes e idénticamente distribuidas.

Consideremos $X_1, \ldots, X_n$ un conjunto de n variables aleatorias (*iid*), tales que, $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$.

Entonces el estadístico

$$\bar{X} = (X_1 + \cdots + X_n)/n,$$

se define como la media de las n variables aleatorias *iid*

$$E(\bar{X}) = \mu$$
$$Var(\bar{X}) = \sigma^2/n$$

Una extensión de lo anterior es:

Sea $X_1, \ldots, X_n$ una conjunto de n variables aleatorias independientes normalmente distribuidas con medias $E(X_i) = \mu_i$ y varianzas $Var(X_i) = \sigma_i^2$ para $i = 1, \ldots, n$. Entonces la distribución de la media muestral es

$$\bar{X} \sim N(\mu, \sigma^2/n) \tag{3.2}$$

Con las propiedades mencionadas, es posible determinar la probabilidad de que ocurran determinados eventos que estén asociados al promedio, siempre y cuando las variables aleatorias tengan una distribución normal.

Ejemplo 3.1. El tipo de cambio del dólar a peso chileno cambia diariamente. Supóngase que este valor es una variable aleatoria distribuida normalmente con media \$720 y desviación estándar de \$10.

El gerente de finanzas de una empresa de *retail* a decidido reajustar mensualmente al alza todos los precios de sus productos, cuando el valor promedio del tipo de cambio del dólar sea superior a \$724. Considerando, que se ha tomado una muestra aleatoria de 22 mediciones del dólar del mes anterior, determine la probabilidad de que el gerente aplique un alza en los precios en el próximo mes.

En primer lugar, corresponde plantear la probabilidad que se desea calcular con relación a la aplicación en el alza de precios. En este caso, la variable de estudio tiene distribución normal, por lo cual, la estandarización del promedio de la variable tiene distribución $N(\mu = 0, \sigma^2 = 1)$.

$$\begin{aligned} P(\bar{X} > 724) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{724 - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{724 - 720}{10/\sqrt{22}}\right) \\ &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > 1.876\right) \end{aligned}$$

Luego, el cálculo en R es el siguiente:

```
pnorm(q = 1.876, mean = 0, sd = 1, lower.tail = F)
```

```
## [1] 0.03032764
```

Finalmente, la probabilidad de que el gerente aplique un alza en los precios en el próximo mes es de 0.03.

Ejercicio 3.1. La cantidad de operaciones bancarias que ocurren durante un día distribuye normal con media 75 (millones) y varianza 144 (millones). Considerando una muestra de tamaño 22 (días), ¿cuál es la probabilidad de que la media muestral de la cantidad de transacciones sea menor a 77.3 millones?

Ejercicio 3.2. El tiempo que demoran en hacerse efectivas las transferencias bancarias internacionales, distribuye normal con media 15 (segundos) y desviación estándar de 4 (segundos). Considerando una muestra de tamaño 29, ¿cuál es la probabilidad de que la media muestral de los tiempos de demora de transacción sea menor a 14.1 segundos? Mencione con rigurosidad todos los pasos de su solución.

3.1.3. Teorema del Límite Central

Considerando el ejemplo 3.1, **¿qué sucede cuándo la variable aleatoria del experimento no sigue una distribución normal?** Para estos casos se utiliza el Teorema del Límite Central (TLC) (Devore, 2008, página 215).

- **Anteriormente:**
  - (Condición) Las variables aleatorias del problema deben distribuir normal (*iid*).
  - (Resultado) Al estandarizar el promedio, este se transforma en una variable aleatoria normal con media 0 y varianza 1.
- **Con TLC:**
  - Se desconoce si las variable aleatorias del problema distribuyen normal (*iid*).
  - (Condición) La cantidad de datos debe ser grande. Por regla general se utiliza un tamaño mayor a 30, sin embargo, no hay investigaciones o fuentes que lo confirmen (Johnson et al., 1994).
  - (Resultado) Al estandarizar el promedio, este se transforma en una variable aleatoria normal con media 0 y varianza 1.

Ejemplo 3.2. Un accionista piensa comprar acciones de una determinada empresa chilena de tecnología. Dado los valores del mercado, se observa que el valor promedio de las acciones (iid) es de $4000 con una desviación estándar de $700. El accionista comprará acciones solo si una muestra aleatoria de 100 valores de acciones resulta en un valor promedio mayor a $4100. ¿Cuál es la probabilidad de que el accionista invierta en la empresa chilena?

Si bien no se especifica que los valores de la acciones distribuyen normal, es posible usar el TLC para determinar la probabilidad.

El planteamiento de la probabilidad es el siguiente:

$$
\begin{aligned}
P(\bar{X} > 4100) &= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{4100-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{4100-4000}{700/\sqrt{100}}\right) \\
&= P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > 1.428\right)
\end{aligned}
$$

El cálculo en R es el siguiente:

```
pnorm(q = 1.428, mean = 0, sd = 1, lower.tail = F)
```

```
## [1] 0.07664593
```

La probabilidad de que el accionista invierta en la empresa chilena es de 0.07

Ejercicio 3.3. La cantidad de operaciones bancarias que ocurren durante un día tienen una media de 75 (millones) y varianza 144 (millones). Considerando una muestra de tamaño 71 (días), y que se desconoce la distribución de los datos, ¿cuál es la probabilidad de que la media muestral de la cantidad de transacciones sea menor a 73.16 millones? Mencione con rigurosidad todos los pasos de su solución.

Ejercicio 3.4. El tiempo que demoran en hacerse efectivas las transferencias bancarias internacionales tienen una media de 15 (segundos) y desviación estándar de 4 (segundos). Considerando una muestra de tamaño 56, y que se desconoce la distribución de los datos, ¿cuál es la probabilidad de que la media muestral de los tiempos de demora de transacción sea menor a 15.8 segundos? Mencione con rigurosidad todos los pasos de su solución.

Ejercicio 3.5. Supóngase que el número de barriles de petróleo crudo que son importados diariamente es una variable aleatoria con una distribución no especificada. Si se observa la cantidad de barriles importados en 64 días, seleccionados en forma aleatoria , y si se sabe que la desviación estándar del número de barriles por día es $\sigma = 60$, determínese la probabilidad de que la media muestral se encuentre a no más de 700 barriles del verdadero valor de barriles importados.

3.2. Distribución de muestreo de la varianza

Si la media es una de las medidas de localización más usadas, entonces, la varianza sería aquella con más uso dentro de las medidas de escala. Y al igual que la media, también existe una distribución de muestreo asociada a la varianza muestral.

Consideremos la expresión utilizada para calcular la varianza de un conjunto de datos (muestra):

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n-1}$$

Se obtiene la siguiente propiedad. Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución normal con media μ y varianza σ^2. Entonces,

$$Y = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{3.3}$$

Ejemplo 3.3. Considere una medición del índice de pobreza (porcentual) proporcionada por un instrumento computacional, en donde el interés recae en la variabilidad de la lectura. Supóngase que, con base en la experiencia, la medición es una variable aleatoria normalmente distribuida con media 14.7 y desviación estándar igual a 1.2 unidades. Si se toma una muestra aleatoria procedente del proceso de medición computacional de tamaño 25, ¿cuál es la probabilidad de que el valor de la varianza muestral del índice de pobreza sea mayor a 1.1 unidades?

El planteamiento de la probabilidad es el siguiente.

$$\begin{aligned} P(S^2 > 1.1) &= P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)\cdot 1.1}{\sigma^2}\right) \\ &= P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{24\cdot 1.1}{1.2^2}\right) \\ &= P\left(\frac{(n-1)S^2}{\sigma^2} > 18.333\right) \end{aligned}$$

Considerando que esta nueva variable tiene distribución $\chi^2_{n-1} = \chi^2_{25-1} = \chi^2_{24}$, el cálculo en R es el siguiente:

```
pchisq(q = 18.333, df = 24, lower.tail = F)
```

```
## [1] 0.7865623
```

Finalmente, la probabilidad de que el valor de la varianza muestral del índice de pobreza sea mayor a 1.1 unidades cuadradas es de 0.7865.

Ejercicio 3.6. Se toma una muestra aleatoria de tamaño 67 proveniente de una población normal con desviación estándar $\sigma = 3.05$. Calcular la probabilidad de que la varianza muestral s^2 sea como mínimo 8.21? Utilice solo una de las siguiente tablas.

Tabla 3.1: Distribución Ji-Cuadrado, df = 66

Cuantil	Probabilidad
55.24886	0.1754
56.24816	0.2016
58.24886	0.2597
58.54886	0.2689
59.24886	0.2910

Tabla 3.2: Distribución Ji-Cuadrado, df = 65

Cuantil	Probabilidad
55.24886	0.1996
56.24816	0.2279
58.24886	0.2895
58.54886	0.2992
59.24886	0.3223

3.3. La distribución T-Student

Hasta el momento sabemos que:

1. $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$
2. $\dfrac{nS^2}{\sigma^2} \sim \chi^2_n$

Sin embargo, ¿qué sucede, cuándo no conocemos la varianza poblacional en la distribución de la media?

Para poder trabajar sobre la distribución de la media, reemplazamos la varianza poblacional por la muestral, dando origen a un nuevo estadístico:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \tag{3.4}$$

el cual, es un estadístico t-Student con $n-1$ grados de libertad.

Ejemplo 3.4. El Departamento de Protección al Medio Ambiente (DPMA) asegura que, para un automóvil compacto en particular, el consumo promedio de gasolina en carretera es de 45 kilómetros por litro. Una organización independiente de consumidores adquiere uno de estos autos y lo somete a prueba con el propósito de verificar la cifra proporcionada por el DPMA. El auto recorrió una distancia de 100 kilómetros en 25 ocasiones. En cada recorrido se anotó el número de litros necesarios para realizar el viaje. Para los 25 ensayos, la desviación estándar tuvo un valor de 2.5 kilómetros por litro. Si se supone que el número de kilómetros que se recorre por litro es una variable aleatoria distribuida normalmente, ¿cuál es la probabilidad de que el promedio de kilómetros recorridos por litro sea menor al 45.2?

A diferencia de los expuesto anteriormente, no se conoce la varianza poblacional de la variable de estudio (los kilómetros por litro que rinde el automóvil), por lo tanto, corresponde utilizar el estadístico t-Student, el cual hace uso de la varianza muestral. El planteamiento de la probabilidad es el siguiente:

$$\begin{aligned} P(\bar{X} < 45.2) &= P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} < \frac{45.2-\mu}{S/\sqrt{n}}\right) \\ &= P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} < \frac{45.2-45}{2.5/\sqrt{25}}\right) \\ &= P\left(\frac{\bar{X}-\mu}{S/\sqrt{n}} < 0.4\right) \end{aligned}$$

Recordando que esta nueva variable distribuye $t_{n-1} = t_{25-1} = t_{24}$, el cálculo en R es el siguiente:

```
pt(q = 0.4, df = 24)
```

```
## [1] 0.6536528
```

la probabilidad de que el promedio de kilómetros recorridos por litro sea menor al 45.2 es de 0.6536.

Ejercicio 3.7. La base de datos *dolar.csv* contiene el valor del dólar observado de algunos de los días de los meses de junio y julio, tomados por el el SII. A continuación:

1. Suponga que el valor del dólar distribuye $N(\mu = 880, \sigma^2 = 50^2)$. Determine probabilidad de que el valor promedio del dólar muestral supere los \$900.
2. Considerando los supuestos de la pregunta 2, ¿cuál es la probabilidad de que la varianza muestral no supere los $\$43^2$? Obtenga el valor de la varianza muestral en R y comente los resultados.
3. Si el valor del dólar distribuye $N(880, \sigma^2)$. Determine probabilidad de que el valor promedio del dólar muestral supere los \$900. Compare con lo obtenido en la pregunta 1, y comente.
4. Con los mismos supuestos de la pregunta 3, determine por separado la probabilidad de que en cada mes, el valor promedio de valor del dólar muestral no supere los \$890. Compare y comente.

Ejercicio 3.8. El conjunto de datos *diabetes.csv* proviene originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es estudiar de forma diagnóstica si un paciente tiene diabetes, en función de ciertas medidas de diagnóstico incluidas en el conjunto de datos. Se impusieron varias restricciones a la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de ascendencia indígena Pima. Las columnas de la base de datos con las siguientes:

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

A continuación,

1. Suponiendo que la edad sigue una distribución normal con media 70 y varianza 830. ¿Cuál es la probabilidad de que una persona de la base de datos tenga más de 73 años?
2. Suponiendo que el grosor de la piel sigue un distribución normal con varianza 414. ¿Cuál es la probabilidad de que la varianza de la muestra sea mayor a 406?

3. Suponiendo que la presión arterial sigue un distribución normal con media 69 y varianza 880. ¿Cuál es la probabilidad de que el promedio muestral de la presión arterial sea mayor a 70?
4. Responda la pregunta 3 suponiendo que desconoce la varianza poblacional.
5. Responda la pregunta 3 suponiendo que desconoce la distribución de los datos.

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. In *The R Software: Fundamentals of Programming and Statistical Analysis*, 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.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2013). *Multivariate data analysis*. Pearson Education.

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. Publisher: Taylor & Francis.

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

Johnson, R. A., Kotz, S., and Balakrishnan, N. (1994). *Sample Size Determination for Central Limit Theorem Applications*, volume 1. Wiley-Interscience.

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. Publisher: New York.

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

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