Desarrollo de contenido

Unidad 1

Estadística II

Ingeniería de Software y Datos

IU Digital de Antioquia
INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA

Unidad 1. Probabilidad

Introducción a la Unidad 1

La estadística es el área derivada de las matemáticas que se encarga de la obtención, sistematización, representación y análisis de los datos. También es el puente que permite transformar los datos brutos en información útil para explicar y predecir diferentes fenómenos de interés.

En *Estadística I*, nos centramos en todas aquellas técnicas necesarias para describir un fenómeno o proceso específico utilizando para ello las medidas de tendencia central, dispersión y algunas representaciones tabulares y gráficas específicas que nos permiten explorar el fenómeno de estudio.

En *Estadística II*, aprenderemos a generar modelos explicativos y predictivos que nos ayudarán a anticiparnos al futuro o a conocer el posible comportamiento de una población, basados en una pequeña porción de ella (muestra). Al igual que en *Estadística I*, estaremos haciendo predicciones sobre algo que no conocemos. De hecho, el profesional en ciencia de datos debe asignar probabilidades a esos escenarios predichos ya que el azar puede jugar un papel importante y puede que el desenlace predicho no sea 100 % seguro.

Es allí donde el desarrollo de esta primera unidad tiene una validez e importancia trascendental, pues lo que buscamos es desarrollar tus capacidades para estimar y comprender el concepto de probabilidad y que la veas aplicada en cada una de las decisiones de tu vida con el fin de mejorar la toma de decisiones.

¡Que sea este el momento para asombrarte y motivarte en el aprendizaje sobre probabilidad!

Muchos éxitos en este proceso de aprendizaje.

Objetivo de aprendizaje de la Unidad 1

Reconoce y aplica los principales conceptos de la probabilidad, probabilidad conjunta, condicional y teorema de Bayes a la solución de problemas relacionados con la ciencia de datos.

Cronograma de actividades de la Unidad 1

Actividad de aprendizaje*	Evidencia de aprendizaje**	Semana***	Ponderación
AA1. Actividad evaluativa no calificable de conocimientos previos	EA. Actividad de conocimientos previos	Semana 1	0 %
AA1. Análisis de caso aplicando tablas de contingencia y teorema de Bayes en el contexto de la ciencia de datos	EA1. Informe escrito en PDF que dé cuenta de la solución a una pregunta de investigación específica a partir del análisis de dos variables de naturaleza cualitativa y el uso de probabilidades	Semana 3	33.33 %
Total			**33.33 %**

Unidad 1. Actividad de aprendizaje 1. Instalación del software estadístico R, Rcmdr y Jamovi para el análisis de datos

La probabilidad es el cálculo matemático que evalúa las posibilidades que existen de que una cosa suceda cuando interviene el azar. Por eso, en esta unidad estudiaremos temas como el espacio muestral, los eventos, las probabilidades de un evento, las reglas aditivas, la regla de Bayes y la introducción a las distribuciones de probabilidad.

Para el desarrollo del curso, es fundamental el uso de las herramientas informáticas que permitan los desarrollos en estadística y ciencia de datos. Por eso, en esta unidad, usaremos el software R y la interfaz Rcommander (Rcmdr). Además, realizaremos el trabajo con datos reales utilizando Jamovi.

A continuación, se presenta brevemente las instrucciones de instalación para los softwares:

1. Instalación del software R y Rcmdr (Aplica para sistemas operativos Windows y IOS de Mac)

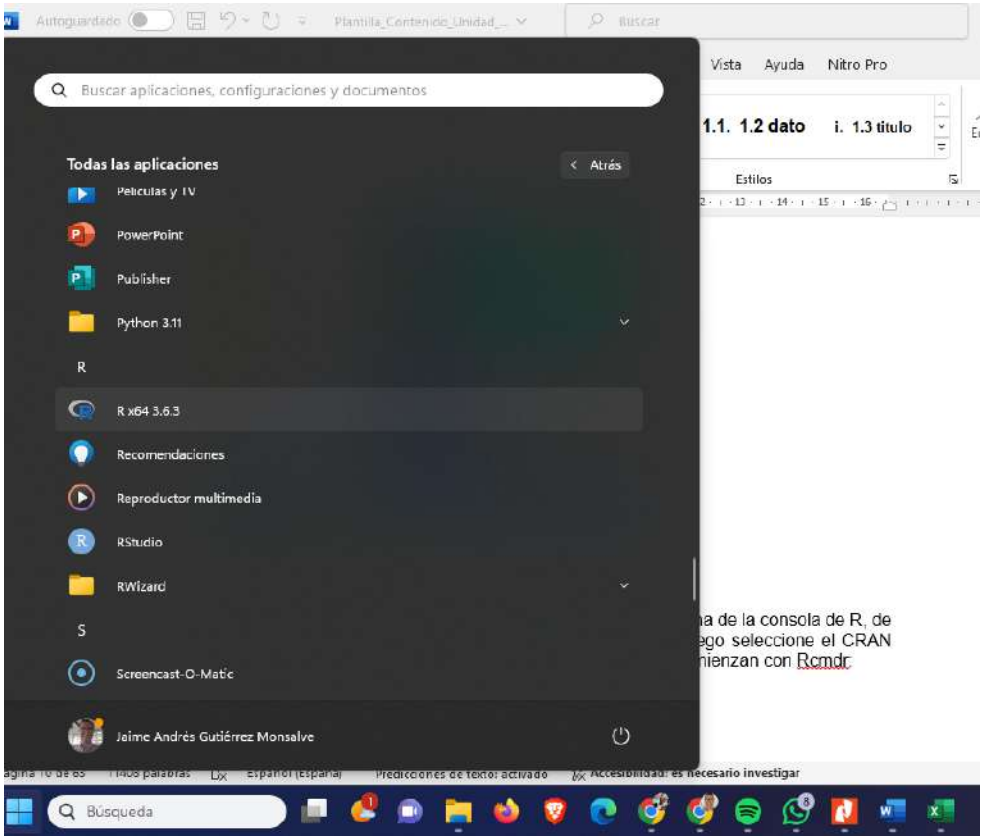
Este proceso de instalación aplica para los sistemas operativos *Windows* y *IOS* de Mac.

a) Descarga el instalador de R versión 4.3.2 de acuerdo con el tipo de sistema operativo del cuál dispongas.
   - En el caso de Windows, puedes descargarlo aquí: https://cran.r-project.org/bin/windows/base/
   - Si tu computador es Mac, puedes descargar la versión de IOS aquí: https://cran.r-project.org/bin/macosx/

b) Instala el software desde la carpeta de descargas. Para ello, debes hacer clic derecho al instalador y luego instalarlo como administrador. Por favor selecciona sí a todas las opciones de instalación que aparezcan en las ventanas emergentes.

c) Si el software quedó adecuadamente instalado este deberá aparecer en la opción de Inicio (en los programas de tu sistema operativo). En la siguiente figura, te mostramos cómo se visualiza en el sistema operativo de Windows.

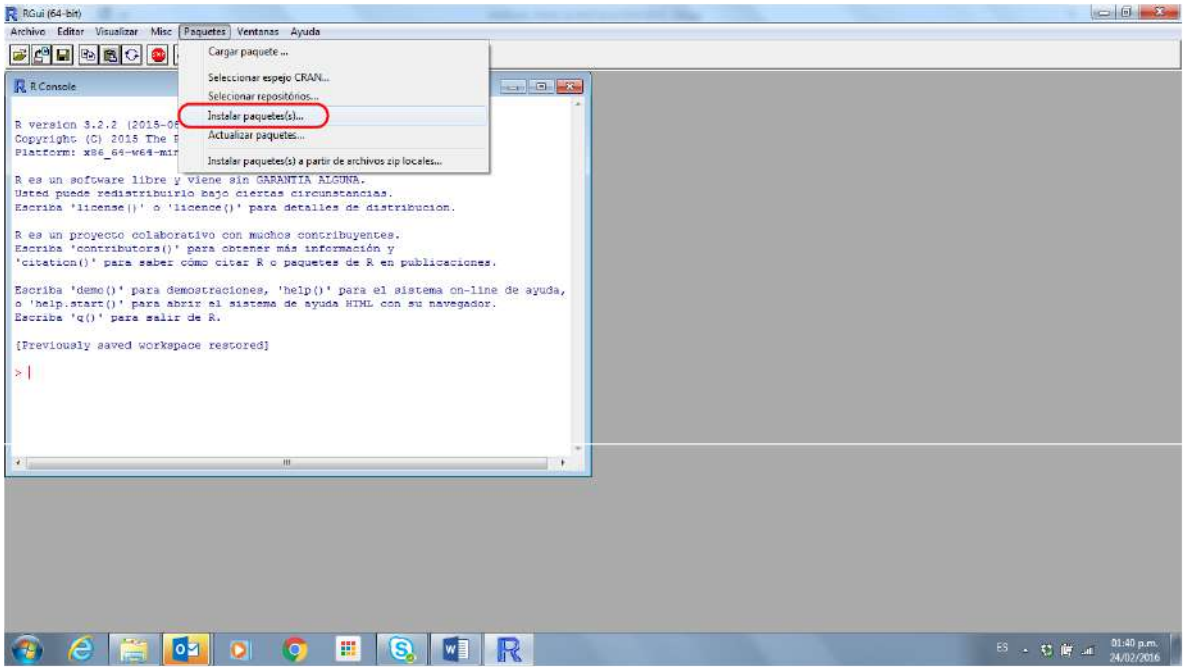
Figura 1. Visualización de la instalación del software R 4.3.2 en la ventana de programas de Windows



Luego, debes abrir el software haciendo clic en el ícono Rx64 4.6.3. A veces aparece otro ícono relacionado con la versión de 32 *bit*, por favor no des clic en él y más bien desinstálalo con el fin de no generar conflicto con la versión de 64 *bit*.

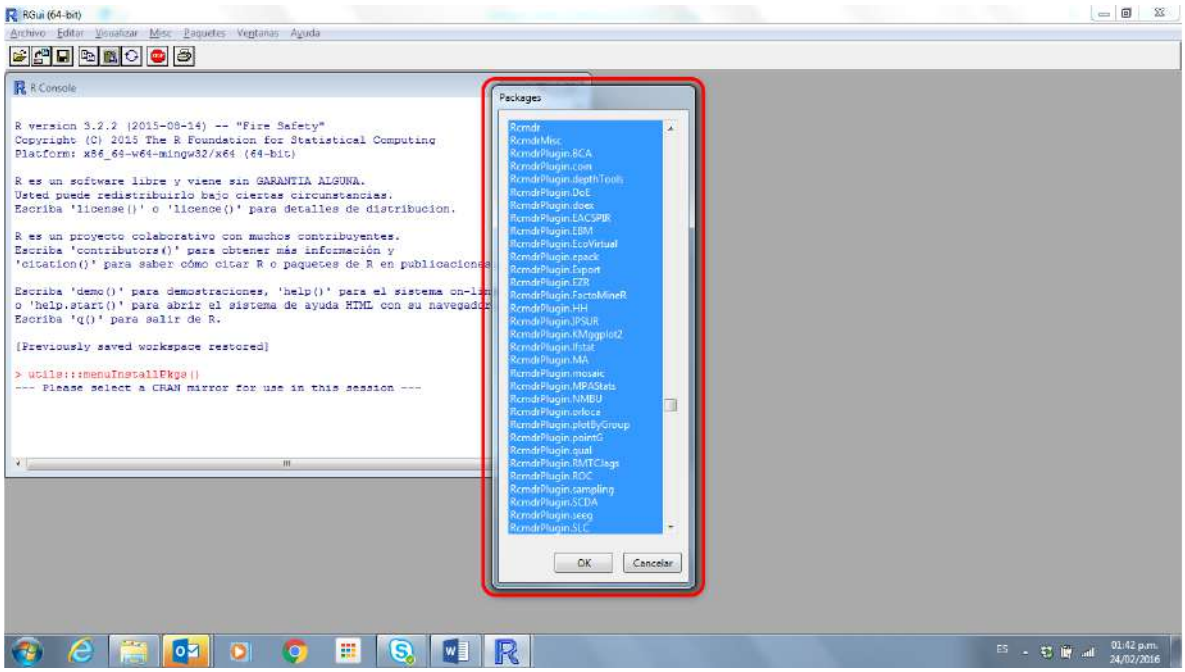
d) Al iniciar el programa R x 64 3.6.3 te aparecerá la ventana de la consola de R. Haz clic en Paquetes y selecciona la opción Instalar paquetes, tal como lo muestra la siguiente figura.

Figura 2. Instalar paquetes



Luego, selecciona el CRAN (Espejo) Austria, y busca e instala todos los paquetes que comienzan con Rcmdr.

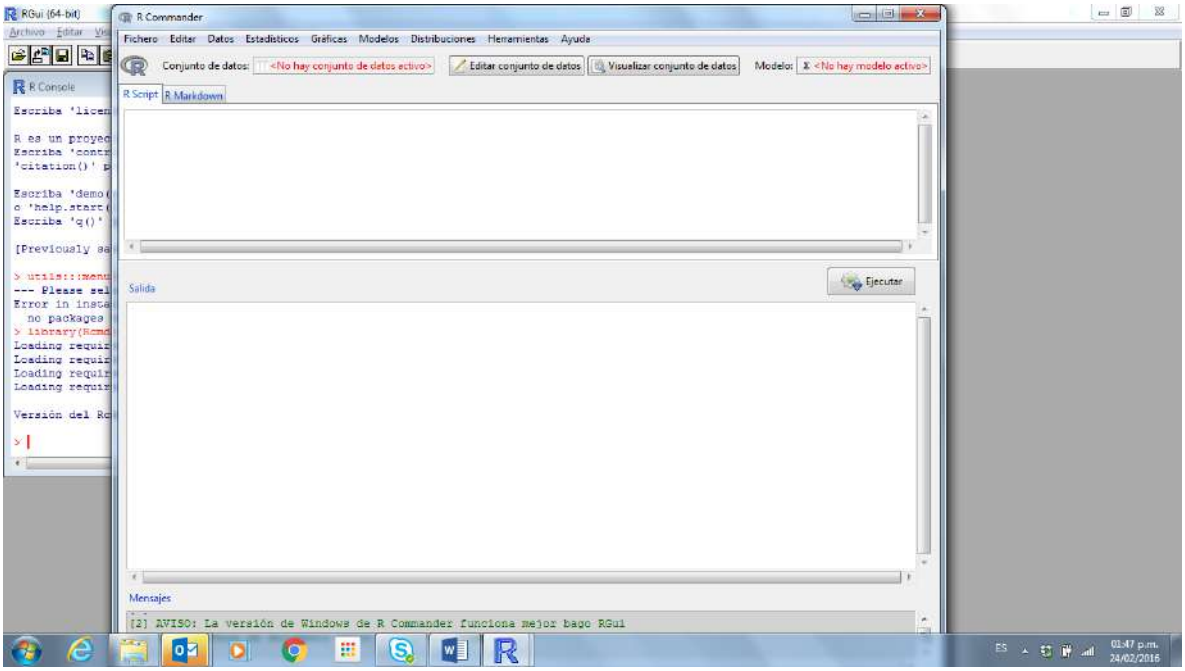
Figura 3. Paquetes Rcmdr



Cuando termines, debes hacer clic en Ok y esperar hasta que instale todos los paquetes de Rcmdr, este procedimiento puede durar de 30 minutos a 1 hora, dependiendo de la velocidad de la red.

e) Al finalizar el proceso de instalación, abre Rcmdr digitando en la línea de comando: library(Rcmdr) en la consola de Rx64_4.3.2 respetando mayúsculas y minúsculas. Luego de que ingreses esta orden te aparecerá un cuadro en donde te avisará que faltan algunos paquetes por instalar, debes hacer clic en Ok y el sistema seguirá instalando los paquetes faltantes.
   En caso de que te aparezca un letrero que diga “falta el paquete lillie.test”, por ejemplo, o cualquier otro, debes ir nuevamente a paquetes, seleccionar el espejo Austria e instalar el paquete que te hace falta, es decir, seguir el mismo proceso del paso anterior.
f) El proceso ha finalizado cuando te aparece la siguiente ventana:

Figura 4. Ventana de visualización de la instalación exitosa de Rcmdr



Nota: cada vez que desees llamar el programa debes abrir Rx64 6.6.3 y poner en la línea de comando la instrucción library(Rcmdr) e inmediatamente te aparecerá la ventana de Rcmdr que es la herramienta que utilizaremos.

Para tener en cuenta

R es el software estadístico de libre distribución más utilizado en el mundo de la ciencia de datos, posee más de 15.000 paquetes y tiene una comunidad estimada en 10 millones de usuarios. Por tanto, en el desarrollo de la Ingeniería en Ciencia de Software y Datos, será un software de continuo uso.

Si deseas conocer más información sobre la instalación del software R, te invitamos a revisar el siguiente material:

Video

- **Autor:** Universidad Miguel Hernández de Elche.
- **Título:** INSTALAR R COMMANDER Y PRIMEROS PASOS.
- **URL:** https://www.youtube.com/watch?v=0Yz4mtfGOtY

Para aprender más

- **Título:** R commander an Introduction.
- **Autor:** Karp (2010).
- **URL:** https://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf

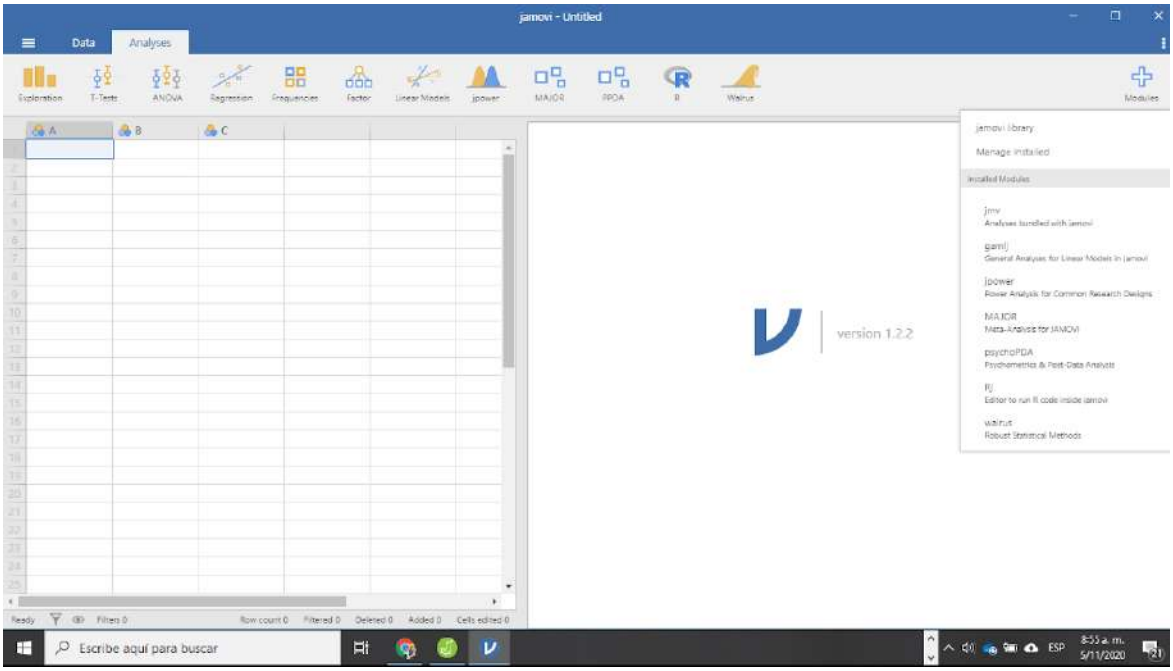
2. Instalación del software Jamovi

Este proceso de instalación aplica para los sistemas operativos *Windows* y *IOS* de Mac. Cuando instales el software R y Rcmdr, puedes proceder con la instalación de Jamovi. Para ello, debes ingresar a la página web de Jamovi: https://www.jamovi.org/download.html, seleccionar el instalador de acuerdo con tu sistema operativo, esperar a que descargue e instalarlo como administrador en tu PC. Recuerda que para instalar como administrador debes hacer clic derecho en el instalador e instalar como administrador en tu PC. Esto se debe a que el software debe quedar con permisos para acceder a los directorios de trabajo de tu PC.

Te recomendamos siempre instalar la versión solid de Jamovi, ya que corresponde a la que ha sido probada y por tanto está libre de fallos. Este software basado en R permite realizar todo tipo de análisis estadísticos univariados, bivariados y multivariados a partir de bases de datos estructuradas que se encuentren en Excel, SQL, Stata, RData, entre otros.

Una vez instalado Jamovi, te aparecerá en la lista de programas de tu PC y al abrirlo te saldrá la siguiente pantalla.

Figura 5. Interfaz de Jamovi



Tema 1. Espacio muestral

Cuando hablamos de espacio muestral hacemos referencia a la teoría de conjuntos. El espacio muestral puede definirse como el conjunto de elementos que conforma un problema de estudio. Se encuentra delimitado de tal manera que los elementos que no pertenecen a dicho objeto de estudio no pueden pertenecer al espacio muestral. Para comprender adecuadamente el espacio muestral es importante considerar algunos elementos que lo conforman: en este caso tenemos el evento y asociado al evento tenemos la probabilidad.

- Un **evento** puede ser definido como el posible resultado, valor u objeto perteneciente a un espacio muestral.
- La **probabilidad**, por su parte, es una medida dada en proporción o porcentaje, la cual representa la posibilidad de que un evento específico ocurra o esté presente en ese espacio muestral**.** Del mismo modo, la probabilidad es la proporción o porción de elementos de un tipo en todo el espacio muestral**.**

Para entender y operativizar estos conceptos, te invitamos a leer y estudiar atentamente los siguientes ejemplos:

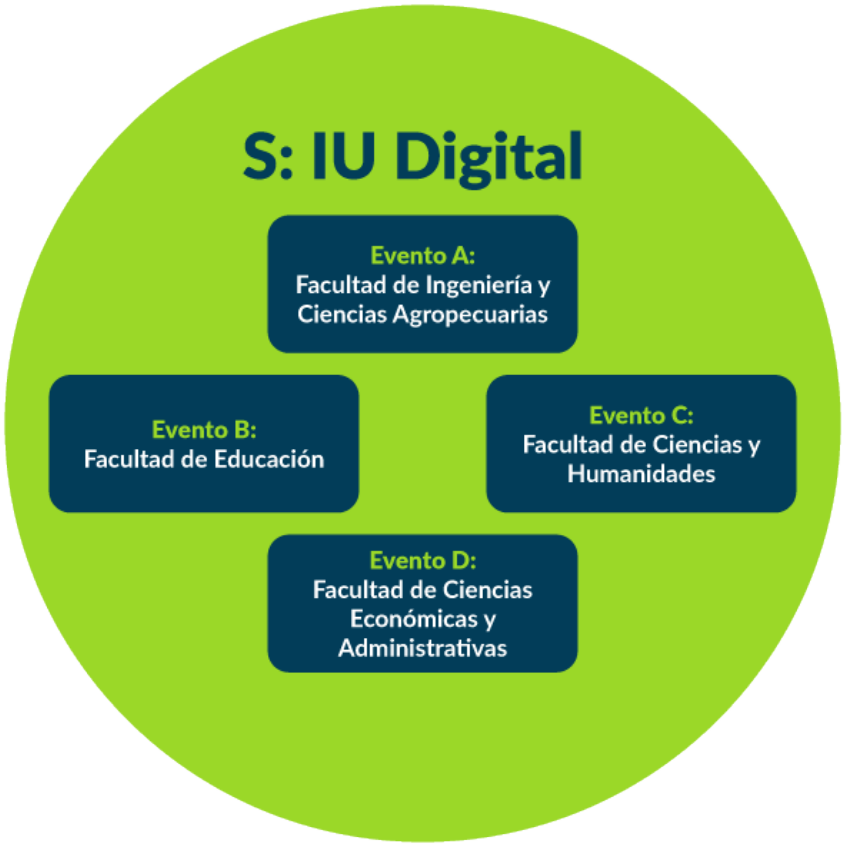
- **Ejemplo 1:** un investigador del programa Ingeniería de Software y Datos está interesado en estudiar la percepción que tienen las diferentes facultades de la IU Digital sobre las políticas educativas del Gobierno de turno. Con base en este enunciado identifica el espacio muestral y los eventos que pueden estar asociados a ellos.

Respuesta:
Espacio muestral: todas las facultades que hacen parte de la IU Digital. En notación probabilística el espacio muestral se denota como S y al interior se enuncian los eventos, así:

S = {Facultad de Ingeniería y Ciencias Agropecuarias, Facultad de Educación, Facultad de Ciencias y Humanidades, Facultad de Ciencias Económicas y Administrativas}
En el caso de los eventos, en probabilidad, estos se suelen definir por letras. En este espacio muestral tenemos cuatro eventos de esta manera:

Sea A el evento de pertenecer a la Facultad de Ingeniería y Ciencias Agropecuarias.
Sea B el evento de pertenecer a la Facultad de Educación.
Sea C el evento de pertenecer a la Facultad de Ciencias y Humanidades.
Sea D el evento de pertenecer a la Facultad de Ciencias Económicas y Administrativas

Figura 6. Espacio muestral y eventos



- **Ejemplo 2:** un estudiante quiere estudiar las probabilidades asociadas a sacar una carta negra en un mazo que contiene cartas rojas y negras. Establezca el espacio muestral y los eventos asociados a este problema de estudio.

Respuesta:
El objeto de estudio aquí son cartas rojas y negras en un mazo de cartas, por lo tanto, el espacio muestral se define como sigue: S = {Cartas rojas, Cartas negras}
Los eventos para este objeto de estudio pueden enunciarse de la siguiente manera:
Sea R el evento de ser una carta roja en el mazo de estudio.
Sea N el evento de ser una carta negra en el mazo de estudio.

Figura 7. Espacio muestral y eventos

Teniendo en cuenta que el objetivo del estudio de este curso consiste en poder integrar los conocimientos vistos en *Estadística I* con probabilidad, esta se asume como el puente entre la estadística descriptiva y la inferencial, es decir, el modelado explicativo y predictivo. De esta manera, el espacio muestral suele asociarse con las variables, y los eventos con los posibles valores que pueden tomar dichas variables.

¿Sabías qué?

En estadística, una variable es una característica particular de un objeto o fenómeno de estudio. Por ejemplo, en el caso de un estudio demográfico en la ciudad de Medellín, interesará conocer la distribución por sexo biológico de la población. En este caso la **variable** es **sexo** y esta puede tomar como valores o **eventos**: masculino o femenino.

Espacio muestral = variable, es decir, el sexo biológico: S = {Femenino, Masculino}
Eventos:
Sea F que una persona sea de Sexo Femenino.
Sea M que una persona sea de Sexo Masculino

Al hablar de eventos, también es importante reconocer que en probabilidad es común encontrar un término que se denomina "El complemento" o "Evento complementario". En notación probabilística, el complemento o evento complementario se señala con una comilla superior y se define como todos los elementos del espacio muestral que no pertenecen al evento que se enuncia. Veamos un ejemplo:

- **Ejemplo:** con el fin de identificar los sectores poblacionales más atendidos por la IU Digital en Colombia, te contratan como estadístico para indagar sobre la distribución por estrato socioeconómico de los estudiantes matriculados. Define el espacio muestral, los eventos y sus complementos. Utiliza la notación probabilística.

Respuesta:
La variable o el espacio muestral para este ejercicio son el estrato social de los estudiantes matriculados en la IU Digital:
S = {E1, E2, E3, E4, E5, E6}

Los eventos son:

Sea E1 el evento de que los estudiantes pertenezcan al Estrato 1.
Sea E2 el evento de que los estudiantes pertenezcan al Estrato 2.
Sea E3 el evento de que los estudiantes pertenezcan al Estrato 3.
Sea E4 el evento de que los estudiantes pertenezcan al Estrato 4.
Sea E5 el evento de que los estudiantes pertenezcan al Estrato 5.
Sea E6 el evento de que los estudiantes pertenezcan al Estrato 6.

Como ejemplo de los eventos complementarios vamos a proponer algunos:

(E1)' = Conjunto de todos los estudiantes menos los que pertenecen al Estrato 1.
(E2)' = Conjunto de todos los estudiantes menos los que pertenecen al Estrato 2.

En el primer caso de eventos complementarios se definen los estudiantes complementarios al E1 o Estrato 1 complementario. Y, en el segundo caso, Estrato 2 complementario o el complemento del E2.

Tema 2. Probabilidad de un evento

Como lo anticipamos, la probabilidad es un valor numérico que define la posibilidad de ocurrencia de un evento. Desde el punto de vista operativo, se define como el cociente entre el número de formas en las cuáles puede ocurrir un evento, dividido, el total de resultados posibles (total de elementos en el espacio muestral).

$$P = \frac{\#\ formas\ en\ las\ que\ ocurre\ un\ evento}{Total\ de\ resultados\ posibles\ en\ el\ espacio\ muestral}$$

De esta manera la probabilidad solo puede tomar valores que van desde 0 a 1 o desde 0 % a 100 %. En notación estadística, la probabilidad la denotamos con P () y dentro del paréntesis ponemos el evento al cual le estamos hallando la probabilidad. Miremos algunos ejemplos para comprender mejor el concepto.

- **Ejemplo 1:** ¿Cuál es la posibilidad de elegir una carta roja en un mazo que contiene 26 cartas rojas y 2 negras? ¿Cuál es la probabilidad de que, habiendo sacado una carta roja en el primer intento, selecciones una negra?

Respuesta:
Para resolver un problema como este, te sugerimos que siempre utilices un protocolo de planteamiento y solución.

1. Definir el espacio muestral o la variable de estudio y sus eventos.
2. Plantear correctamente en notación de probabilidades lo que se pide y definir numéricamente las probabilidades.
3. Enunciar la respuesta, para ello, se pueden utilizar porcentajes.

Si haces esto con todos los problemas relacionados con la probabilidad, tu probabilidad de errar se disminuye de manera importante.
Apliquémoslo al problema:

1. Define el espacio muestral o la variable de estudio y sus eventos:

   El espacio muestral o la variable de estudio es el mazo, el cual contiene cartas rojas y negras: S = {Cartas Rojas, Cartas Negras}

   Sea R que una carta sea Roja.
   Sea N que una carta sea Negra Primero: define el espacio muestral o la variable de estudio y sus eventos

2. Plantea correctamente en notación de probabilidades lo que se te pide. Igualmente, define numéricamente las probabilidades. Recuerda que la clave son los conteos o el número de elementos que se tienen en el evento sobre el total de elementos del espacio muestral.

$$P(R) = \frac{26(Corresponde\ al\ número\ de\ cartas\ Rojas)}{28\ (Corresponde\ al\ total\ de\ cartas\ en\ el\ espacio\ muestral)} = 0.9295$$

3. Enuncia tu respuesta, para ello, puedes utilizar porcentajes. La probabilidad de sacar una carta roja en un mazo de 28 cartas es del 92.95 %. También puedes enunciarlo de la siguiente manera: El 92,95 % de las cartas contenidas en el mazo son rojas y el 7,15 % de las cartas contenidas son negras.

Para el segundo caso en el cual se nos pregunta que, sabiendo que hemos obtenido en el primer intento una carta roja, cuál es la probabilidad de que saquemos en el segundo intento una carta negra.

Como ya hemos definido nuestro espacio muestral y cada uno de los eventos, procedemos a la estimación de la probabilidad desde el punto de vista numérico:

$$P(N) = \frac{2\ (corresponde\ al\ número\ de\ cartas\ Negras)}{27\ (corresponde\ al\ total\ de\ cartas\ en\ el\ espacio\ muestral)} = 0.0741$$

Es decir, habiendo sacado una carta roja en el primer intento, la probabilidad de que sacar una carta negra en el segundo intento es de 7.41 %.

Observa que el espacio muestral se disminuyó en una carta, pues como sabemos salió una roja en el primer intento.

- **Ejemplo 2:** ¿Cuál es la probabilidad de sacar un número par al lanzar dos dados de 6 caras cada uno? Ten en cuenta que no importa el dado, es decir, si en el dado uno sacas 4 y en el dado dos sacas 1, es lo mismo que si en el dado uno sacas 1 y en el dado dos sacas 4. Este es un caso típico de personas que les gusta jugar parqués.

Ante esta pregunta, la mayoría responderíamos que la probabilidad es 50 %, puesto que en el imaginario se tiene que sacar par corresponde a la mitad de las posibilidades que se pueden presentar al lanzar dos dados a la vez.

¿Será esto cierto?, sigamos el protocolo propuesto y demostrémoslo, usando el protocolo.

1. Define el espacio muestral, la variable y los eventos asociados:

   En este estudio, el espacio muestral se configura sobre los posibles valores que pueden obtenerse al lanzar dos dados al mismo tiempo. Esta es la variable, los posibles valores que pueden tomar el lanzamiento de dos dados:

S = {2,3,4,5,6,7,8,9,10,11,12}

Sin embargo, en este estudio en particular interesa un caso específico, sacar par o impar. Por lo tanto, podemos configurar los eventos como siguen:

Sea P, sacar par al lanzar dos dados, es decir, 2, 4, 6, 8, 10 y 12.
Sea I, sacar impar al lanzar dos dados, es decir, 3, 5, 7, 9 y 11.

2. Plantea correctamente en notación probabilística lo que se pide. Igualmente, identifica numéricamente los conteos para poder plantear la fórmula de probabilidad correctamente.

$$P(P) = \frac{Número\ de\ formas\ de\ sacar\ par\ al\ lanzar\ dos\ dados}{El\ número\ total\ de\ valores\ que\ se\ pueden\ obtener\ al\ lanzar\ dos\ dados}$$

En este caso, para quienes no somos muy diestros en el juego de los parqués, nos es difícil establecer el número de formas en las cuales se puede sacar par al lanzar dos dados, igualmente nos es difícil estimar el número total de valores que se pueden obtener al lanzar los dos dados. Cuando pasa este tipo de situación, es mejor recurrir a una simulación. En la siguiente tabla, te mostramos el total de valores que se pueden obtener al lanzar dos dados, veamos.

Tabla 1. Valores que se obtienen al lanzar dos dados

Dado 1	Dado 2	Dado 1	Dado 2	Dado 1	Dado 2	Dado 1	Dado 2	Dado 1	Dado 2	Dado 1	Dado 2
1	1	2	2	3	3	4	4	5	5	6	6
1	2	2	3	3	4	4	5	5	6		
1	3	2	4	3	5	4	6				
1	4	2	5	3	6						
1	5	2	6								
1	6										

Sin importar el orden de los dados, en la tabla se muestra el total de resultados que se pueden obtener al lanzar dos dados. En verde, el total de resultados pares. De esta manera, al lanzar dos dados es posible obtener par de 12 maneras diferentes. Del mismo modo el total de valores que se pueden obtener al lanzar dos dados es de 21. Por tanto, la probabilidad queda esbozada como sigue:

$$P(P) = \frac{12\ formas\ de\ sacar\ par\ al\ lanzar\ dos\ dados\ \text{sin}\ importar\ el\ orden}{21\ \ valores\ que\ se\ pueden\ obtener\ al\ lanzar\ dos\ dados\ \text{sin}\ importar\ el\ orden} = 0.5714\%$$

3. Procede a responder a la pregunta del problema. Existe una probabilidad del 57,14 % de sacar un número par al lanzar dos dados sin tener en cuenta el orden en el que salen los números en los dados.

Como podemos apreciar, la probabilidad no es del 50 % como intuitivamente muchos pudimos haber pensado. En conclusión al problema, al lanzar dos dados, es más fácil sacar par que impar.

Comprendiendo los conceptos básicos de probabilidad, podemos estimar probabilidades simples en fenómenos mucho más complejos.

Tema 3. Reglas aditivas y teorema de Bayes

Una vez comprendido el concepto de probabilidad de un evento, es posible, desde la teoría de probabilidades, relacionar dos o más eventos con sus respectivas probabilidades. Este enfoque puede ser abordado desde dos perspectivas:

1. Una perspectiva abstracta y analítica, la cual se basa en el uso de teoremas y fórmulas para hallar soluciones a los problemas propuestos y estimar la probabilidad a partir de demostraciones.
2. Una perspectiva más práctica, la cual utiliza tablas de contingencia y con base en ello permite estimar las probabilidades y presentar informes relacionando variables de naturaleza cualitativa y las probabilidades asociadas a la ocurrencia de eventos.

Para comprender el concepto de probabilidad aditiva y conjunta, revisemos el siguiente ejemplo:

Ejemplo: un día cualquiera antes de amanecer, un estudiante de la Ingeniería en desarrollo de Software y Datos desea estimar la probabilidad de que llueva en ese día. Para ello, usa uno de los siguientes instrumentos: termómetro, anemómetro (para medir la velocidad y dirección del viento) o humidímetro (para medir la humedad relativa del aire). Con base en la información que le arroja el instrumento, el estudiante estima que la probabilidad es:

P(Llueva) = 0.6.

Imaginemos al mismo estudiante, con los mismos instrumentos y viendo que encima de su casa hay una nube negra. ¿Qué crees que pasará con la probabilidad inicial estimada? La solución es sencilla: aumentará, en este caso.

P(Llueva | Nubes.Negras) = 0.9

Por el contrario, en caso de que el estudiante se levante y vea que en el cielo no hay nubes negras, sino que está totalmente claro y despejado, con toda seguridad su estimación de la probabilidad de que llueva disminuirá.

P(Llueva | No.hay.nubes.negras) = 0.3

Aquí se evidencia el uso de la probabilidad condicional, por ejemplo, como puedes ver se usa el signo | que indica que el evento posterior a este es lo conocido o dado. De esta manera P(Llueva | Nubes.negras) = 0.9 se interpreta como: "la probabilidad de que llueva dado que en el cielo hay nubes negras es de 0.9".

En el caso P(Llueva | No.hay.nubes.negras) = 0.3 se interpreta como: "la probabilidad de que llueva dado que no hay nubes negras es de 0.3". También es posible enunciar esta probabilidad de la siguiente manera: "Dado que no hay nubes negras, la probabilidad de que llueva hoy es de 0.3".

Como nos encontramos frente a una perspectiva más práctica que teórica y analítica, apropiaremos el concepto de probabilidad aditiva, condicional y teorema de Bayes a partir de las tablas de contingencia.

3.1 Probabilidades simples, aditivas y condicionales a través de las tablas de contingencia

¿Qué es una tabla de contingencia?

Es un arreglo tabular que presenta la frecuencia de ocurrencia de dos eventos. Es una manera elegante de presentar el análisis de dos variables cualitativas que no cuenten con más de 8 niveles o categorías en cada uno. A partir de la tabla de contingencia, es posible estimar cualquier tipo de probabilidad asociada a la ocurrencia de cualquier evento de interés.

Las tablas de contingencia son la mejor forma para representar los sucesos de dos variables cualitativas. Se componen de una variable fila y otra variable columna. En la combinación entre filas y columnas, denominada celda, se representa la frecuencia de ocurrencia de casos entre los eventos relacionados.

Para entender mejor estas tablas, te invitamos a explorar el siguiente ejemplo:

Ejemplo: en el Centro de Recursos para el Aprendizaje y la Investigación de la IU Digital de Antioquia (CRAI), la bibliotecóloga está interesada en saber si, a medida en que los estudiantes avanzan en su formación profesional, visitan con mayor frecuencia la biblioteca. Para ello, elabora una encuesta donde indaga a cada uno de los visitantes sobre el semestre en el que se encuentra y la frecuencia con la cuál ha visitado la biblioteca durante los últimos 3 meses. Los resultados tabulados se presentan en la siguiente tabla de contingencia:

Tabla 2. Resultados de encuesta de visitas a la biblioteca

	Semestre cursado					
Asistencia a la Biblioteca	**1**	**2**	**3**	**4**	**5**	**Total**
Nunca	50	45	30	35	10	170
Menos de 4 veces	20	18	40	50	55	183
4 veces o más	10	30	35	42	49	166
Total	80	93	105	127	114	519

Como se puede apreciar en la *Tabla 2*, el total de estudiantes que fueron encuestados es de 519. Al considerar los estudiantes por semestre (suma totalizada por columna), encontramos que los estudiantes del primer semestre corresponden a 80, del segundo semestre a 93, del tercer semestre a 105, del cuarto semestre a 127 y del quinto semestre a 114. Por otro lado, entre los estudiantes encuestados, 170 nunca han ido a la biblioteca, 183 han ido menos de cuatro veces y 166 han ido 4 veces o más. ¿Identificas estos valores en la tabla de contingencia? ¡Corresponde a los valores totales!

En la tabla también podemos identificar intersección de eventos o regla **Y**, por ejemplo:

- ¿Cuántos estudiantes nunca han ido a la biblioteca **y** se encuentran en primer semestre? La respuesta son 50.
- ¿Cuántos estudiantes se encuentran en quinto semestre **y** han ido 4 veces o más a la biblioteca? En este caso miramos la celda que corresponde a la intersección de *Semestre 5* y *Asistencia a la biblioteca 4 veces o más*, en este caso, el valor es 49.

En la tabla también podemos utilizar la expresión o regla aditiva **O.** En ese sentido se puede preguntar:

- ¿Cuántos estudiantes se encuentran en primero y segundo semestre? Para responder, basta únicamente con sumar los del primer y el segundo semestre: 80+93 = 173 estudiantes, lo cuales están en primero o en segundo semestre.
- ¿Cuántos estudiantes nunca han visitado la biblioteca **o** están en tercer semestre? En este caso basta con sumar el total de estudiantes que nunca han ido a la biblioteca con el total de estudiantes que están en tercer semestre: 170 + 105 = 275 estudiantes.

Como observaste, la tabla de contingencia nos permite extraer muchos resultados asociados a la relación entre dos variables. Partiendo del acercamiento al manejo de estas tablas, continuemos con la estimación de probabilidades asociadas a la tabla de contingencia. Para ello, frente a un problema de este tipo, te invitamos a seguir el siguiente protocolo:

a) Comprende muy bien el enunciado de tu problema: identifica las variables y los eventos o valores asociados a cada una de las variables, Además, define con letras o símbolos cada uno de los eventos relacionados con las variables de interés. Te recomendamos que siempre los escribas para que no se te vaya a olvidar alguno.
b) Construye la tabla de contingencia, para esto puedes utilizar Excel o R, te sugerimos el uso de estos softwares pues facilitan enormemente la estimación de las probabilidades.
c) Utiliza notación probabilística para expresar la probabilidad que te piden.
d) Calcula la probabilidad utilizando sumas, conteos y divisiones de acuerdo con lo que se te pide.

Para comprender mejor la forma en que se operativiza una tabla de contingencia, retomaremos el ejemplo anterior sobre las visitas a la biblioteca. Recordémoslo.

Continuemos con el ejemplo: en el Centro de Recursos para el Aprendizaje y la Investigación de la IU Digital de Antioquia (CRAI), la bibliotecóloga está interesada en saber si, a medida en que los estudiantes avanzan en su formación profesional, visitan con mayor frecuencia la biblioteca. Para ello, elabora una encuesta donde indaga a cada uno de los visitantes sobre el semestre en el que se encuentra y la frecuencia con la cuál ha visitado la biblioteca durante los últimos 3 meses. Los resultados tabulados se presentan en la siguiente tabla de contingencia:

Tabla 2. Resultados de encuesta de visitas a la biblioteca

	Semestre cursado					
Asistencia a la Biblioteca	**1**	**2**	**3**	**4**	**5**	**Total**
Nunca	50	45	30	35	10	170
Menos de 4 veces	20	18	40	50	55	183
4 veces o más	10	30	35	42	49	166
Total	80	93	105	127	114	519

A partir del ejemplo:

Determina la probabilidad de que un estudiante que visite el CRAI sea de Semestre 1, también que sea de Semestre 2. Igualmente estima la probabilidad de que nunca haya ido

a la biblioteca, y la probabilidad de que haya ido 4 veces o más. (Todas estas se denominan probabilidades simples o de un solo evento).

Respuesta: a continuación, te mostraremos el paso a paso para resolver este problema, aplicando el protocolo.

a) Dado que el enunciado del problema es claro, al igual que el propósito que este encierra, procedamos con la identificación de variables y eventos enmarcados en el estudio.

   Tal y como se observa, en este estudio hay dos variables, la primera consiste en la "Asistencia a la Biblioteca" y esta variable en su espacio muestral asume los valores S = {Nunca, Menos de 4 veces, 4 veces o más}.

   Definiendo cada uno de los eventos para esta variable tenemos:

   Sea N: que un estudiante nunca visite la biblioteca.
   Sea 4-: que un estudiante visite la biblioteca menos de cuatro veces.
   Sea 4+: que un estudiante visite la biblioteca cuatro veces o más.
   Por otro lado, se encuentra la variable "Semestre cursado", la cual asume en su espacio muestral los siguientes eventos S = {1,2,3,4,5}.

En esta variable tenemos los siguientes eventos:

   Sea S1: que un estudiante se encuentre en el primer semestre.
   Sea S2: que un estudiante se encuentre en el segundo semestre.
   Sea S3: que un estudiante se encuentre en el tercer semestre.
   Sea S4: que un estudiante se encuentre en el cuarto semestre.
   Sea S5: que un estudiante se encuentre en el quinto semestre.

b) Construcción de la tabla de contingencia. En este ejercicio en particular nos presentan la tabla de contingencia, así que no tenemos que construirla.
c) Utilización de la notación probabilística para expresar la probabilidad que se nos pide y la estimamos. Recuerda que debemos utilizar también la definición de probabilidad:

$$P(S1) = \frac{80\ (corresponde\ al\ total\ de\ estudiantes\ del\ Semestre\ 1)}{519\ (Corresponde\ al\ total\ de\ estudiantes\ encuestados)} = 0.1541$$

Conclusión 1: la probabilidad de que un estudiante se encuentre en el primer semestre es de 0.1541. También es posible decir que el 15.41 % de los estudiantes pertenecen al Semestre 1.

$$P(S2) = \frac{93\ (corresponde\ al\ total\ de\ estudiantes\ del\ Semestre\ 2)}{519\ (Corresponde\ al\ total\ de\ estudiantes\ encuestados)} = 0.179$$

Conclusión 2: la probabilidad de que un estudiante se encuentre en el segundo semestre es de 0.179. También es posible decir que el 17.90 % de los estudiantes pertenecen al Semestre 2.

$$P(N) = \frac{170\ (corresponde\ al\ total\ de\ estudiantes\ que\ nunca\ han\ ido)}{519\ (Corresponde\ al\ total\ de\ estudiantes\ encuestados)} = 0.33$$

Conclusión 3: la probabilidad de que un estudiante nunca haya ido a la biblioteca es de 0.33. También es posible decir que el 33 % de los estudiantes nunca han ido a la biblioteca.

Y, así sucesivamente, se pueden calcular otras probabilidades asociadas a la tabla de contingencia.

¡Esperamos que todo te vaya quedando muy claro! Recuerda que para que logres un aprendizaje significativo debes practicar y realizar ejercicios, y no quedarte solo con la lectura o toma de notas de lo que vamos estudiando.

3.2 Probabilidad conjunta o aditiva: unión e intersección

Como ya explicamos, estas probabilidades se estiman para un evento (probabilidades simples), sin embargo, es posible estimar probabilidades relacionadas con dos o más eventos. En este caso podemos utilizar la conjunción **o** la cual en probabilidad se denota como **U**. Del mismo modo podemos utilizar el operador **y** el cual en probabilidad se expresa con el signo **∩,** o lo que es lo mismo: una U invertida.

Para comprender el uso de las probabilidades, continuemos con el ejemplo sobre las visitas a la biblioteca y apliquémoslo al caso concreto. Veamos los ejercicios que se proponen con dicho ejemplo:

a. Estima la probabilidad de que un estudiante sea del Semestre 5 o del Semestre 4. Para esto expresamos la probabilidad en términos de notación probabilística:

$$P(S4 \cup S5) = \frac{(127 + 114)(suma\ de\ los\ estudiantes\ del\ S4 + S5)}{519\ (suma\ total\ de\ participantes)} = 0.4643$$

Respuesta:
La probabilidad de que un estudiante pertenezca al Semestre 4 o Semestre 5 es de 0.4643. También es posible decir que el 46.43 % de los estudiantes incluidos en el estudio pertenecen al Semestre 4 o al Semestre 5.

b. Estima la probabilidad de que un estudiante haya ido a la biblioteca menos de 4 veces o esté en tercero o cuarto semestre. En este caso se mezclan tres eventos,

por lo que el enunciado de la probabilidad quedaría de la siguiente manera:

$$P(4 - \cup S3 \cup S4) = \frac{(183 + 105 + 127)(suma\ del\ total\ de\ todo\ los\ eventos)}{519\ (suma\ total\ de\ participantes)} = 0.80$$

Respuesta:
La probabilidad de que un estudiante haya ido cuatro o menos veces a la biblioteca o sea de Semestre 3 o sea de Semestre 4 es de 0.80. También es posible concluir que el 80 % de los estudiantes han ido menos de cuatro veces, o pertenecen al Semestre 3 o pertenecen al Semestre 4.

c. Para comprender el caso de la probabilidad asociada a la intersección o a **y,** veamos este ejercicio: ¿Cuál es la probabilidad de que un estudiante sea de Semestre 1 **y** nunca haya ido a la biblioteca? Para ello representamos la probabilidad utilizando la notación adecuada:

$$P(S1 \cap N) = \frac{(50)(intersección\ entre\ S1\ y\ nunca\ haber\ ido\ a\ la\ biblioteca)}{519\ (suma\ total\ de\ participantes)} = 0.096$$

Respuesta:
La probabilidad de que un estudiante nunca haya ido a la biblioteca y esté en Semestre 1 es de 0.096. También es posible concluir que el 9.6 % de los estudiantes se encuentran en Semestre 1 y nunca han ido a la biblioteca.

d. Utilizando la conjunción **y,** también podemos realizar este ejercicio: ¿Cuál es la probabilidad de que un estudiante se encuentre en Semestre 5 **y** nunca haya ido a la biblioteca? Enunciamos apropiadamente la probabilidad:

$$P(S5 \cap N) = \frac{(10)(intersección\ entre\ S1\ y\ nunca\ haber\ ido\ a\ la\ biblioteca)}{519\ (suma\ total\ de\ participantes)} = 0.019$$

Respuesta:
La probabilidad de que un estudiante nunca haya ido a la biblioteca y esté en Semestre 5 es de 0.019. También es posible concluir que el 1.9 % de los estudiantes se encuentran en Semestre 5 y nunca han ido a la biblioteca.

Conclusión del ejercicio: como puede apreciarse y, de acuerdo con las probabilidades halladas, parece que en la medida en que se avanza en el semestre, se disminuye la probabilidad de nunca haber ido a la biblioteca.

Para tener en cuenta

La probabilidad total, la probabilidad conjunta **o** y la probabilidad conjunta **y** tienen como denominador, para la estimación de la probabilidad, el total de elementos que conforman la muestra. Por esto a estas probabilidades se les conoce como **probabilidades totales**.

3.3 Probabilidad condicional y teorema de Bayes

Como tal vez deduzcas, en la medida en que conozcamos y reconozcamos los fenómenos de estudio, menor incertidumbre tenemos en la toma de decisiones basadas en probabilidad.

De esta manera, Bayes nos advierte sobre la probabilidad condicionada. El principio es sencillo, si tú miras por la ventana y observas nubes negras y rayos, la probabilidad de que vaya a llover es altísima; en contraposición con el hecho de que salgas de tu casa y observes un cielo azul claro y sin nubes. El evento dado aquí son las nubes negras y los truenos, los cuales modifican de manera importante la probabilidad de lluvia.

Para explorar el concepto de probabilidad condicional, retomaremos nuevamente el ejemplo sobre la frecuencia de visitas a la biblioteca, que hemos venido trabajando. Esperamos que lo que lo tengas a la mano. Revisa detenidamente los ejercicios que te proponemos a partir de dicho ejemplo:

a. Si un estudiante se encuentra en Semestre 1, ¿cuál es la probabilidad de que nunca haya ido a la biblioteca? Para resolver este reto, lo primero es especificar el problema utilizando notación probabilística. Recordemos que, para esto, el evento que se pone luego de | es el evento dado.

$$P(N|S1) = \frac{50(corresponde\ a\ la\ intersección\ entre\ N\ y\ S1)}{80(Corresponde\ al\ total\ del\ evento\ dado, S1)} = 0.625$$

Respuesta: dado que los estudiantes se encuentren en Semestre 1, el 62.5 % de ellos no ha ido a la biblioteca. Observa que en esta representación únicamente nos interesa el total de los estudiantes en el evento dado o conocido, por ello, es una probabilidad condicional.

b. En este ejercicio veremos que la utilidad práctica de este tipo de probabilidad se centra en poder realizar comparaciones objetivas entre los datos reportados en la tabla de contingencia. Así, si queremos determinar si el hecho de avanzar en los semestres disminuye la probabilidad de nunca haber asistido a la biblioteca,

podemos preguntarnos por la probabilidad de que un estudiante no haya asistido dado que se encuentra en Semestre 3.

$$P(N|S3) = \frac{30(corresponde\ a\ la\ intersección\ entre\ N\ y\ S1)}{105(Corresponde\ al\ total\ del\ evento\ dado, S1)} = 0.286$$

Respuesta: de los estudiantes que se encuentran en Semestre 3, el 28.6 % de ellos no ha ido a la biblioteca.
También nos interesará conocer la probabilidad de que un estudiante no haya asistido a la biblioteca dado que se encuentra en quinto semestre:

$$P(N|S5) = \frac{10(corresponde\ a\ la\ intersección\ entre\ N\ y\ S1)}{114(Corresponde\ al\ total\ del\ evento\ dado, S1)} = 0.088$$

Respuesta: de los estudiantes que se encuentran en Semestre 5, el 8.8 % de ellos no ha ido a la biblioteca.

Es de esta manera que la bibliotecóloga del ejemplo estará satisfecha al poder responder su pregunta inicial, es decir que, efectivamente, a medida en que los estudiantes van avanzando en su formación académica, el porcentaje de ellos que nunca han visitado la biblioteca se va reduciendo casi que una tercera parte por semestre.

Lo que acabamos de hacer es demostrar el teorema de Bayes, este se puede enunciar como sigue:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Y se puede comprender así: la probabilidad condicional de un evento sobre otro es igual a la probabilidad de la intersección, sobre la probabilidad total del evento que es dado.

Para demostrarlo utilizaremos nuevamente el ejemplo sobre la frecuencia de visitas a la biblioteca y nos enfocaremos en los estudiantes que han asistido 4 veces o más, para responder ¿cuál es la probabilidad de que pertenezcan al Semestre 1? Igualmente, pensando en los estudiantes que han asistido 4 veces o más, responderemos ¿cuál es la probabilidad de que sean del semestre 5? (Esto como comparativo).

A continuación, te el paso a paso de la demostración:

Paso 1: lo primero que hacemos es plantear la notación probabilística adecuadamente:

$$P(S1|4+) = \frac{10(corresponde\ a\ la\ intersección\ entre\ S1\ y\ 4+)}{166(Corresponde\ al\ total\ del\ evento\ dado, 4+)} = 0.0602$$

Conclusión: la probabilidad de que un estudiante se encuentre en Semestre 1 dado que ha asistido 4 veces o más a la biblioteca es de 0.06. Dicho de otro modo, el 6.02 % de los estudiantes que han asistido a la biblioteca 4 o más veces son del Semestre 1.
Si exponemos esta probabilidad a partir del teorema de Bayes, quedaría así:

$$P(S1|4+) = \frac{P(S1 \cap 4+)}{P(4+)}$$

Paso 2: ahora, para realizar adecuadamente la demostración, estimamos por aparte cada una de las probabilidades enunciadas:

$$P(S1 \cap 4+) = \frac{10(Corresponde\ al\ total\ de\ estudiantes\ en\ la\ intersección\ Sy\ y\ 4+)}{519(Corresponde\ al\ total\ de\ la\ muestra)} = 0.019$$

Conclusión: la probabilidad de que un estudiante se encuentre en el Semestre 1 y haya ido 4 o más veces a la biblioteca es de 0.019. También es posible afirmar que el 1.9 % de los estudiantes incluidos en el muestreo se encuentran en el Semestre 1 y han ido 4 o más veces a la biblioteca.

Estimemos la probabilidad relacionada con haber asistido cuatro o más veces a la biblioteca:

$$P(4+) = \frac{166(Total\ de\ personas\ que\ han\ asistido\ 4\ o\ más\ veces)}{519(Corresponde\ al\ total\ de\ la\ muestra)} = 0.32$$

Conclusión: la probabilidad de que los estudiantes hayan asistido 4 o más veces a la biblioteca es de 0.32. También es posible afirmar que el 32 % de los estudiantes considerados en el muestreo han asistido 4 o más veces.

Paso 3: habiendo estimado por separado cada probabilidad enunciada, procedamos con el cálculo que sugiere el teorema. Para ello, basta con dividir las probabilidades encontradas.

$$P(S1|4+) = \frac{P(S1 \cap 4+)}{P(4+)} = \frac{0.019}{0.32} = 0.060$$

En caso de que el resultado de esta división sea igual a la probabilidad condicional inicialmente enunciada (en el *Paso 1*), hemos concluido con la demostración.

Conclusión: la probabilidad de que un estudiante esté en Semestre 1 dado que ha visitado la biblioteca 4 o más veces es de 0.06. Esta probabilidad corresponde con la inicialmente estimada utilizando la tabla de contingencia en el *Paso 1*.

Paso 4: con el fin de complementar el análisis, se plantean las otras dos probabilidades de acuerdo con la definición del teorema de Bayes:

$$P(S5|4+) = \frac{P(S5 \cap 4+)}{P(4+)} = \frac{\frac{49}{519}}{\frac{166}{519}} = 0.295$$

Conclusión: de los estudiantes que han dicho asistir 4 o más veces a la biblioteca, el 29.5 % de ellos se encuentra en Semestre 5. Podemos comparar este resultado con la probabilidad de que los estudiantes estén en Semestre 1 dado que fueron 4 o más veces a la biblioteca, la cual fue de 6.0 %. En este sentido, es posible inferir que efectivamente entre mayor sea el avance de los estudiantes en el semestre, mayor será la probabilidad de que estos asistan 4 o más veces al recinto.

Reto formativo 1: Aplicación de software Rcmdr en ejercicio de probabilidad y teorema de Bayes

Planteamiento del ejercicio:
En este reto encontrarás una situación de la vida real, debes estudiarla y estimar las probabilidades que te solicitamos. Pero no estarás solo, te mostraremos el mostraremos el paso a paso para que lo repliques y realices el proceso a la par.

Instrucciones:

1. Lee detenidamente la situación y has un esfuerzo por entender lo que se te presenta a continuación:

   Un equipo de investigación pretende evaluar la capacidad de un nuevo procedimiento para identificar si un árbol de café se encuentra enfermo o no de un hongo que afecta la raíz y eventualmente es mortal. Al realizar el estudio se obtuvieron los siguientes resultados: a 436 árboles que presentaron la enfermedad, la prueba diagnóstica les dio positivo. Por su parte, a 5 árboles que no presentaron la enfermedad, la prueba diagnóstica dio positivo. Análogamente, la prueba dio negativo para 495 árboles que no tenían la enfermedad. Y, por último, la prueba dio negativo a 14 árboles que si tenían la enfermedad.

   Ten en cuenta que este ejercicio es utilizado en epidemiología para realizar una evaluación de las pruebas diagnósticas. Un uso muy común de este tipo de estudio

son el diseño de pruebas de embarazo, cada una de ellas tiene indicadores de sensibilidad, especificidad, positividad y negatividad.

Con base en esta información estima las siguientes probabilidades:

a. La sensibilidad de la prueba: probabilidad de un resultado positivo de la prueba dada la presencia de la enfermedad.
b. La especificidad de la prueba: probabilidad de un resultado negativo de la prueba dada la ausencia de la enfermedad.
c. La positividad de la prueba de detección: probabilidad de que un individuo tenga la enfermedad, dado que el individuo presenta un resultado positivo en la prueba de detección.
d. La negatividad de la prueba: probabilidad de que un individuo tenga la enfermedad, dado que el resultado de la prueba fue negativo.

2. Define adecuadamente los eventos y asígnales letras o símbolos particulares.

3. Construye la tabla de contingencia de acuerdo con los eventos y las variables. Ingresa la tabla de contingencia en el software Rcmdr.

4. Expresa en correcta notación las probabilidades que te pide el ejercicio.

5. Utiliza el teorema de Bayes y el software estadísticos para encontrar las probabilidades que se te piden.

Desarrollo del reto o preguntas:

A continuación, encontrarás el proceso de solución de este ejercicio para que lo repliques y realices el proceso a la par:

Solución del Reto:

1. **Leamos la situación y comprendámosla.**
2. **Definamos las variables y los eventos:**

Como es una prueba diagnóstica, generalmente en este tipo de estudios siempre hay dos variables:

- Resultado de la prueba diagnóstica (asume valores positivo o negativo).
- Estado de la planta de café (asume valores enferma o no enferma).

Eventos para la variable prueba diagnóstica:

Sea + que la prueba diagnóstica dé un valor positivo.
Sea - que la prueba diagnóstica dé un valor negativo.

Eventos para la variable enfermedad:

Sea E que la planta de café se encuentre enferma.
Sea NE que la planta de café no se encuentre enferma.

3. **Construyamos la tabla de contingencia:**

En este caso pondremos la variable "enfermedad" en las columnas y la variable relacionada con la prueba diagnóstica en la fila. Basta entonces hacer uso de Excel y comprender muy bien los valores de la frecuencia para poder construir la tabla adecuadamente:

Tabla 3. Tabla de contingencia para identificar enfermedad de árbol de café

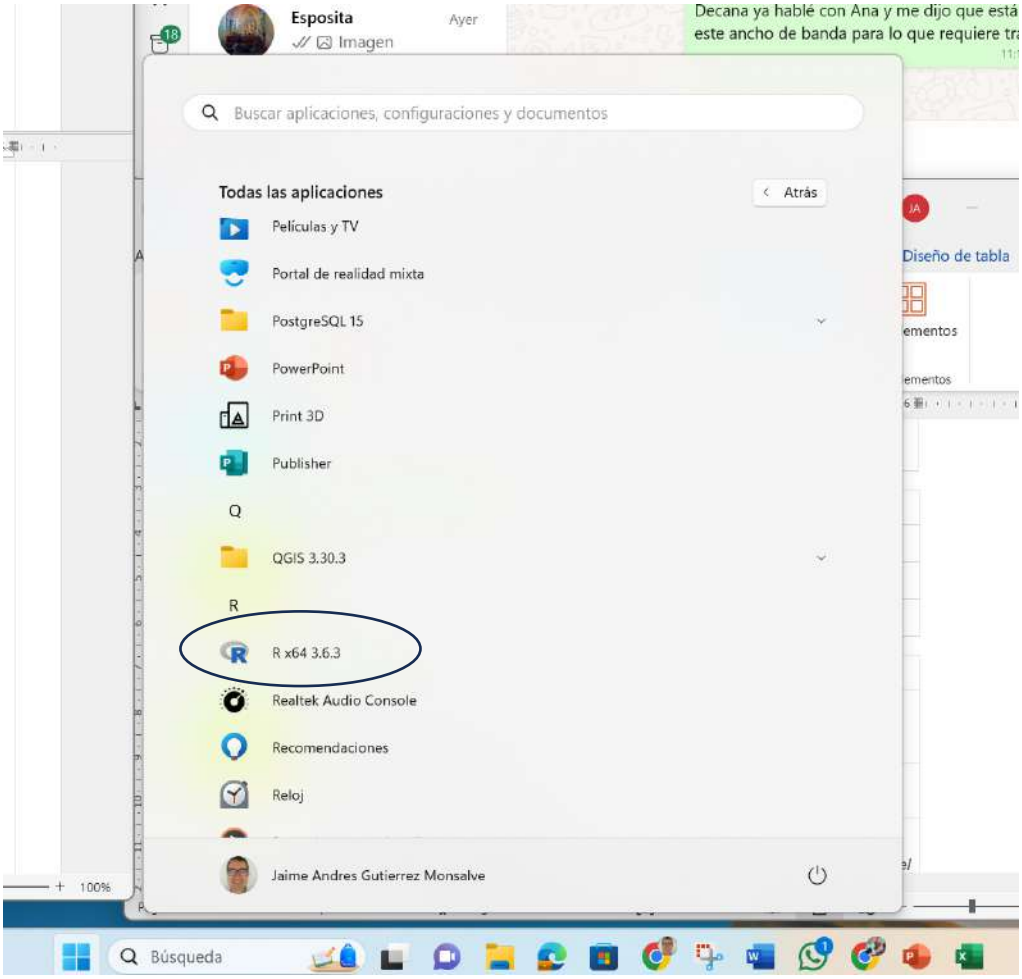
Resultado Prueba Diagnóstica	**Planta Enferma**		
	E	**NE**	**Total**
+	436	5	441
-	14	495	509
Total	450	500	950

Con la tabla de contingencia construida es posible estimar todas las probabilidades: la sensibilidad, la especificidad, la positividad y la negatividad de la prueba, tal y como lo hicimos en los ejercicios de la unidad. Sin embargo, en esta oportunidad haremos uso del software para estimar dichas probabilidades.

Trabajaremos con el Software Rcmdr, para ello, te invitamos a seguir el siguiente proceso para abrir el software e ingresar la tabla de contingencia, (está guía se propone usando *Windows*):

1) En el botón Inicio de tu equipo, abre el software R 4.3.2 x 64, el cual previamente debes tener instalado. Si aún no lo has hecho, estás a tiempo.

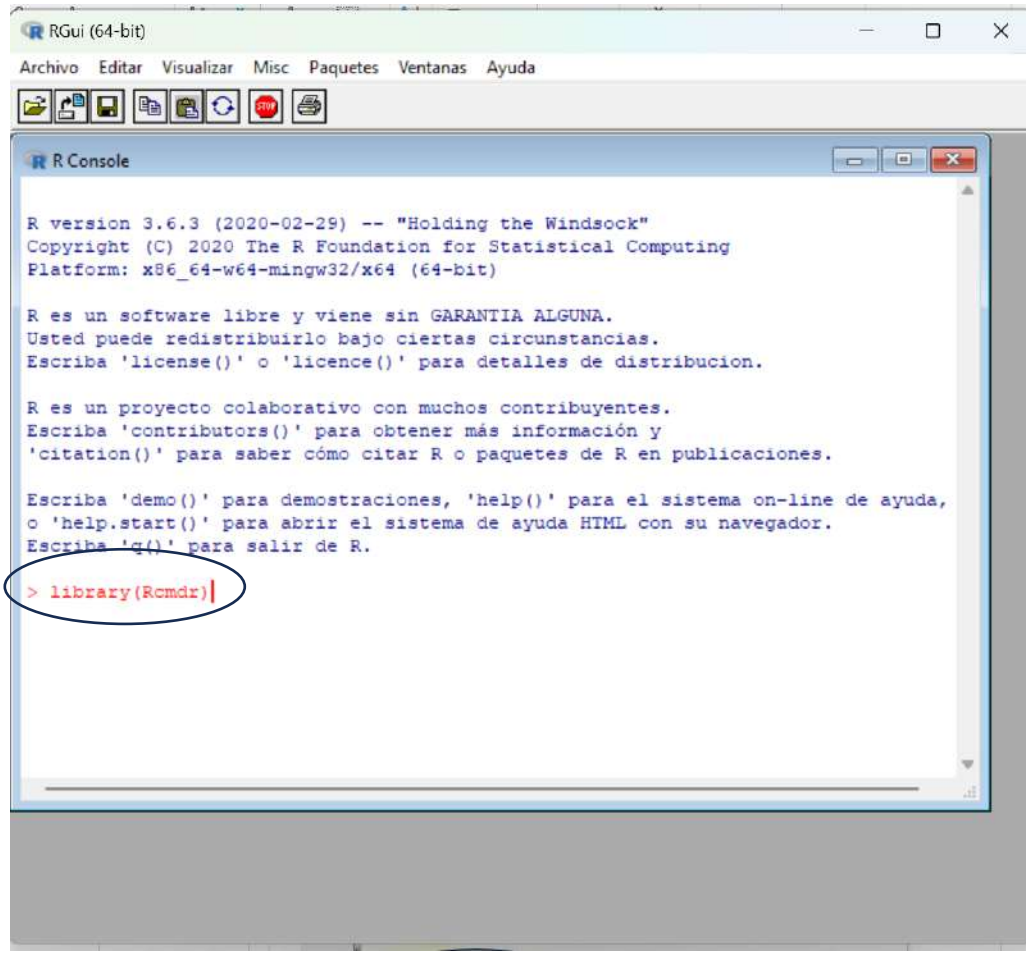
Figura 8. Abrir software R 4.3.2 x 64



2) Luego se abrirá el software y, para abrir Rcmdr, debes escribir en la consola el comando:

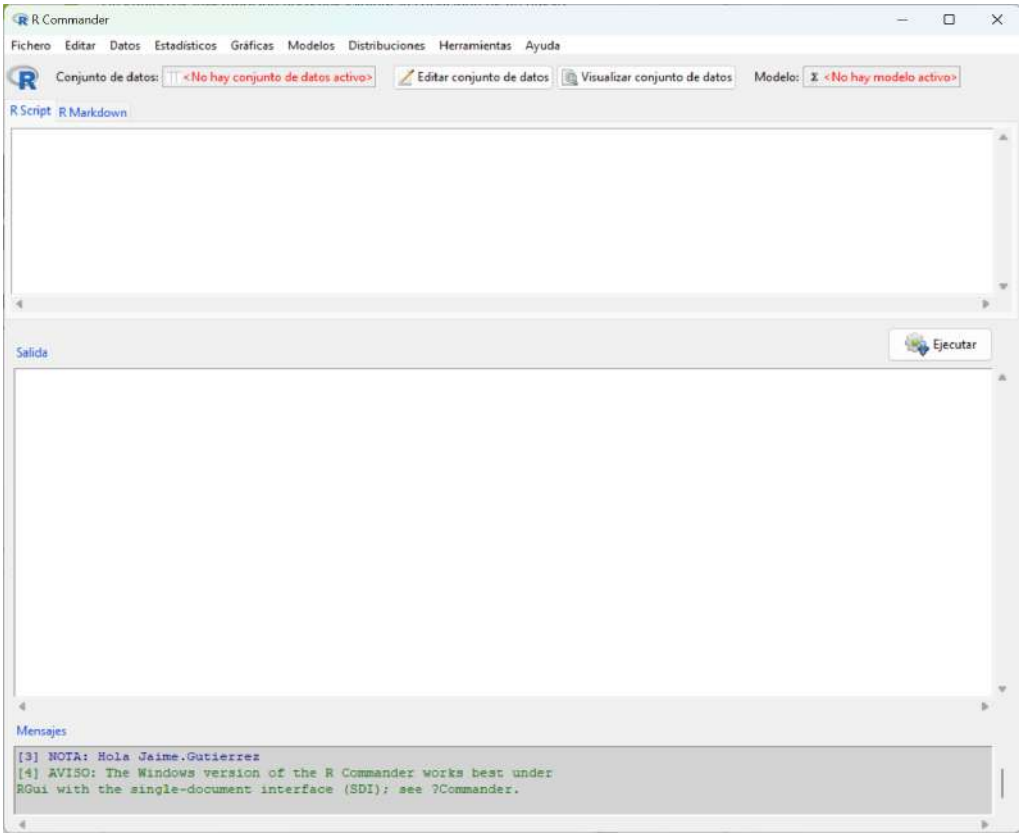
Library(Rcmdr)
###Esto lo haces porque es un paquete de R

Figura 9. Abrir Rcmdr



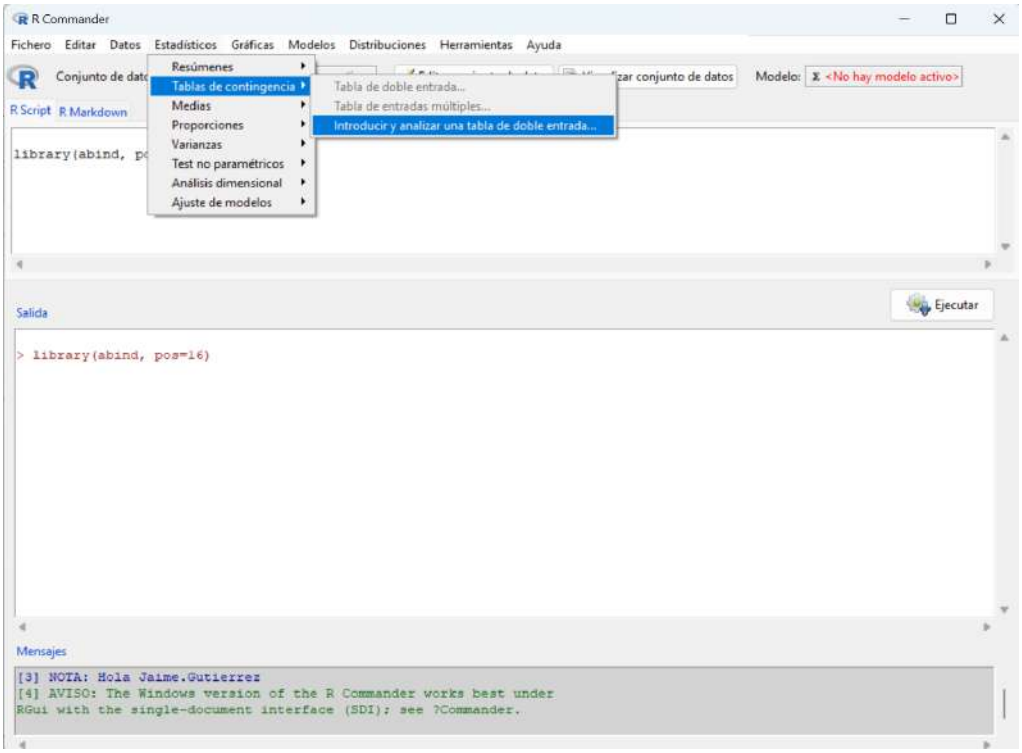
3) Pulsa Enter y se abrirá la interfaz de usuario Rcmdr. Deberá aparecer el siguiente recuadro:

Figura 10. Interfaz de usuario Rcmdr



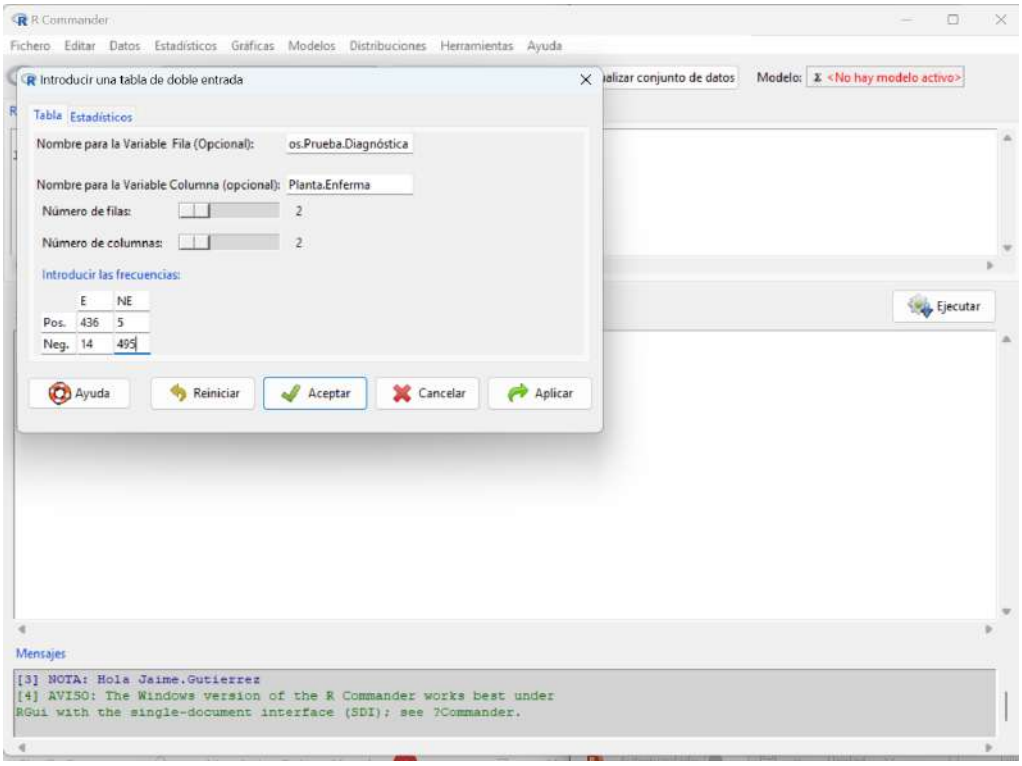
4) Para ingresar la tabla de contingencia, debes seguir la siguiente instrucción: Estadísticos >>> Tablas de contingencia >>> Introducir y analizar una tabla de doble entrada, así:

Figura 11. Ingreso de tabla de contingencia



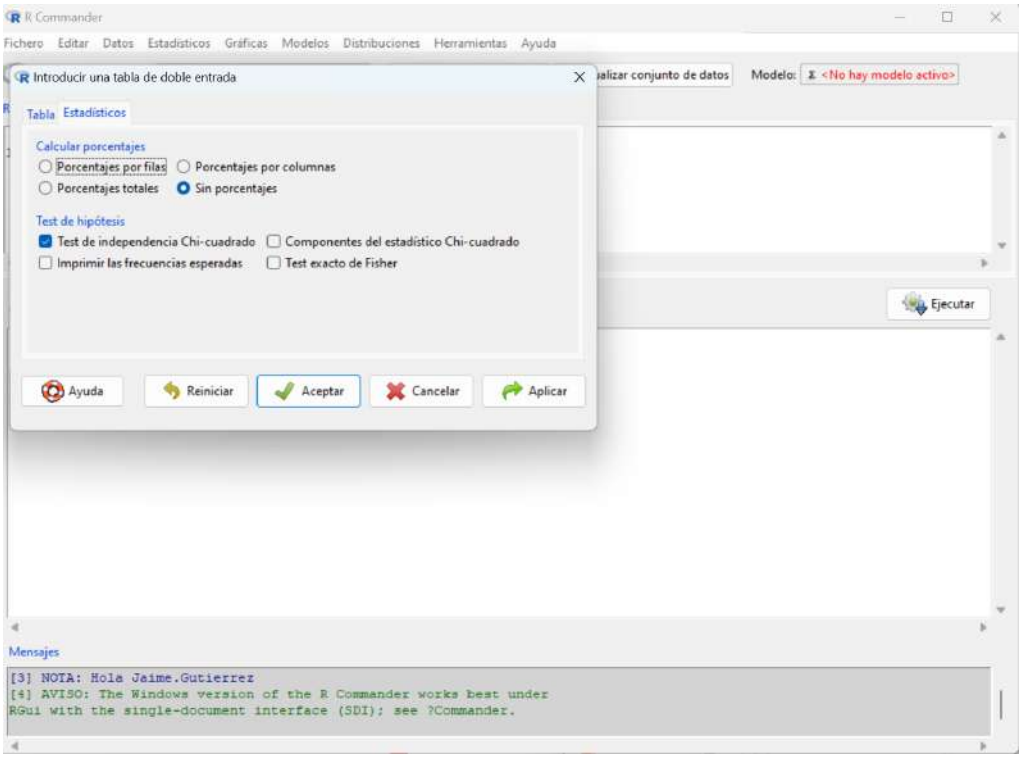
5) Luego aparecerá el siguiente recuadro, donde podrás ingresar la tabla de contingencia, tal y como la construiste en el numeral anterior:

Figura 12. Introducir una tabla de doble entrada



6) Debes hacer clic en la pestaña Estadísticos y configurar las opciones frente a los porcentajes:

Figura 13. Introducir una tabla de doble entrada – Estadísticos



Para las probabilidades condicionales, si la variable que condicionamos es la variable columna, le damos porcentaje por columna. Si la variable que condicionamos es la variable fila, le damos porcentaje fila. Las probabilidades totales las podemos estimar con la opción Porcentajes totales.

4. **Expresemos en correcta notación las probabilidades:**

Entendiendo el funcionamiento del software, especificaremos las probabilidades relacionadas con lo que se nos piden:

Sensibilidad:

$$Sensibilidad = P(+|P) = 0.969\ o\ 96.9\%$$

Dado que el evento que condiciona corresponde a la presencia de la enfermedad (variable columna), hacemos clic en Porcentaje por columnas y el software nos entrega la siguiente salida:

Figura 14. Resultados de Porcentaje por columnas

```
> .Table  # Counts
                              Planta.Enferma
Resultados.Prueba.Diagnóstica   E  NE
                         Pos. 436   5
                         Neg.  14 495

> colPercents(.Table) # Column Percentages
                              Planta.Enferma
Resultados.Prueba.Diagnóstica     E  NE
                         Pos.  96.9   1
                         Neg.   3.1  99
                         Total 100.0 100
                         Count 450.0 500
```

De esta manera observamos que en porcentaje columna, el 100 % se encuentra sobre los eventos de la columna. El valor solicitado corresponde a 96.9 %, es decir, la intersección entre Positivo (+) y Enfermo (E). La prueba tiene una sensibilidad del 96.9 %, es decir, la prueba es capaz de identificar como enfermas el 96.9 % de las plantas que verdaderamente están enfermas.

Especificidad:
En el caso de la especificidad, a continuación, se expresa la probabilidad*:*

$$Especificidad = P(-|NP) = 0.99\ o\ 99\%$$

Como la condición o el evento dado corresponde con la planta enferma, la opción Porcentaje por columnas es la adecuada. Resaltado en rojo se presenta la probabilidad solicitada. En este caso, la prueba diagnóstica es capaz de detectar correctamente y como negativo el 99 % de las plantas que no presentan la enfermedad*.*

Positividad y negatividad:
En el caso de la positividad, la variable que condiciona es la variable fila, por lo tanto, interesa elegir la opción Porcentaje por filas. Al hacer esto, los resultados se presentan a continuación:

Figura 15. Resultados de Porcentaje por filas

```
> .Table  # Counts
                                Planta.Enferma
Resultados.Prueba.Diagnóstica    E  NE
                          Pos. 436   5
                          Neg.  14 495

> rowPercents(.Table) # Row Percentages
                                Planta.Enferma
Resultados.Prueba.Diagnóstica    E   NE Total Count
                          Pos. 98.9  1.1   100   441
                          Neg.  2.8 97.2   100   509

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test
```

$$Positividad = P(E|+) = 0.989\ o\ 98.9\%$$
$$Negatividad = P(NE|-) = 0.972\ o\ 97.2\%$$

En las pruebas diagnósticas, existe una probabilidad que es crítica, esta consiste en la probabilidad de que la prueba entregue un falso negativo, es decir la probabilidad de un resultado negativo que de la prueba, dado que el árbol se encuentra enfermo*:*

$$Falso\ Negativo = P(-|E) = 0.031$$

Esto quiere decir que la prueba detecta como negativo el 3.1 % de los árboles que se encuentran enfermos.

En términos generales si la sensibilidad, la especificidad, la positividad y la negatividad se encuentran por encima del 98 %, se puede decir que la prueba diagnóstica en ciencias de la salud es adecuada. También se debe cumplir con que los falsos negativos no superen el 2 %. Por lo tanto, en esta prueba diagnóstica, todavía se deben hacer algunos ajustes para mejorar sobre todo los falsos negativos.

¿Seguiste los pasos y realizaste el ejercicio en tu equipo? ¡Esperamos que sí! Ya que es un ejercicio práctico muy enriquecedor.

Reto formativo 2: Aplicación de software Jamovi

Planteamiento:
En este reto encontrarás una situación con información cualitativa, debes estudiarla y construir las tablas de contingencia en el software Jamovi. Pero no estarás solo, te mostraremos el proceso de solución de este ejercicio para que lo repliques y realices el proceso a la par.

Instrucciones:

1. Lee detenidamente la situación:

   Se realizó un estudio con 1342 estudiantes con el fin de establecer la distribución por sexo y estado civil de los estudiantes matriculados en la Universidad de Malpelo. Para el desarrollo de este proyecto a usted se le presenta la siguiente base de datos (https://docs.google.com/spreadsheets/d/1PeH9UIMI5-ca6GZqnHPzsj90mBntWWVw/edit?usp=sharing&ouid=101484640100322140552&rtpof=true&sd=true) con las dos variables cualitativas.

2. Para el desarrollo de esta actividad sugerimos el uso del software Jamovi. Para ello, debes seguir la siguiente secuencia de pasos:
   a. Descarga la base de datos y guárdela en una carpeta que conozcas. Guárdala preferiblemente en formato de Excel.
   b. Abre Jamovi e ingresa correctamente la base de datos.
   c. Dirígete al Menú de Tablas de contingencia y construye las tablas de contingencia y los análisis que den cuenta del proyecto

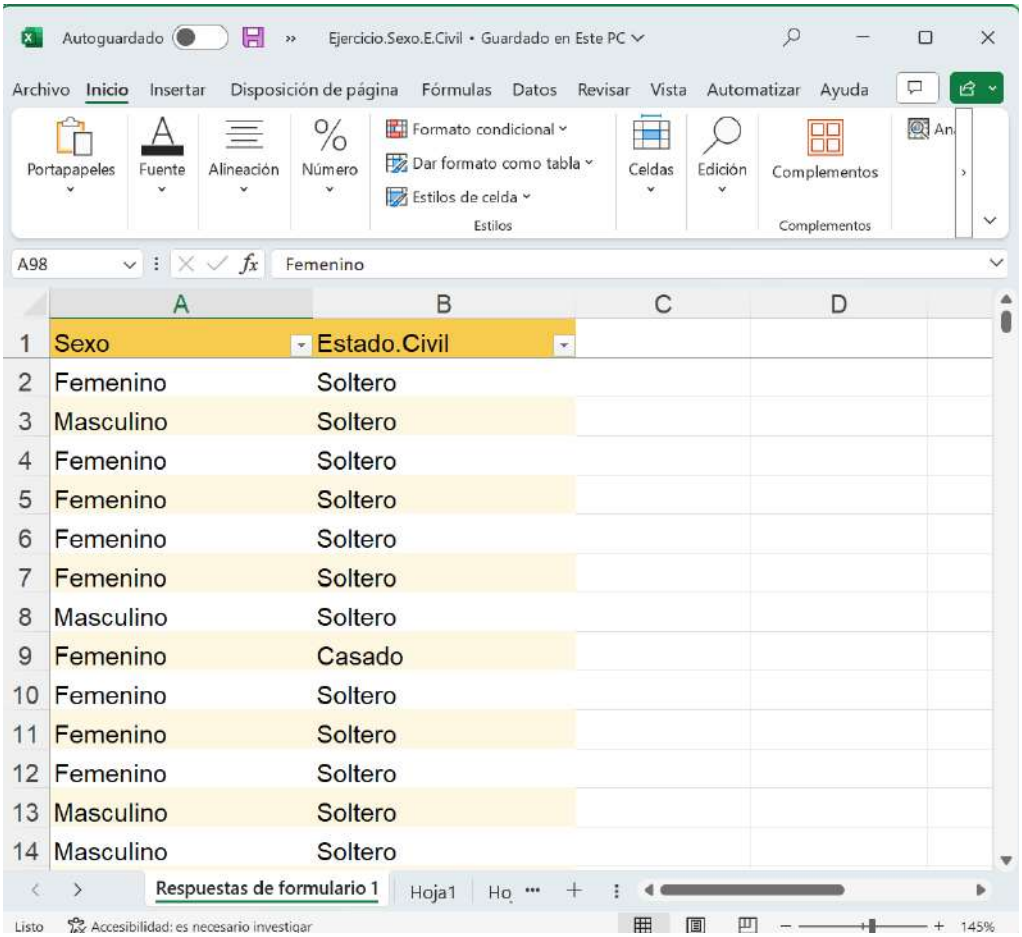
Desarrollo del reto o preguntas:

A continuación, encontrarás el proceso de solución de este ejercicio para que lo repliques y realices el proceso a la par:

Solución del Reto:

a. **Descargamos la base de datos y la revisamos:**
   La base de datos contiene dos variables de naturaleza cualitativa nominal. La primera es estado civil (eventos: casado, divorciado, unión libre, viudo y noviazgo) y la segunda es sexo (eventos: masculino y femenino).

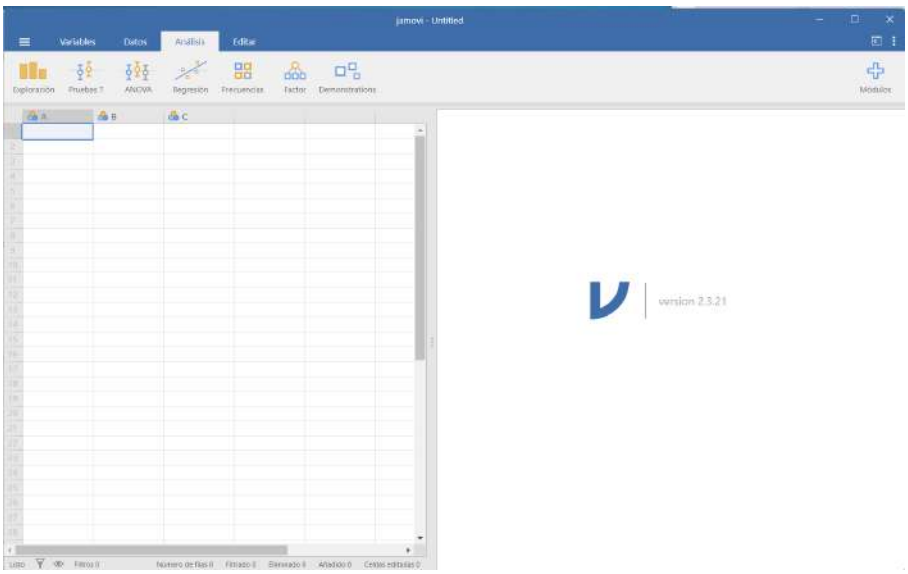
La representación de la base de datos se presenta como sigue:



	A	B	C	D
1	Sexo	Estado.Civil		
2	Femenino	Soltero		
3	Masculino	Soltero		
4	Femenino	Soltero		
5	Femenino	Soltero		
6	Femenino	Soltero		
7	Femenino	Soltero		
8	Masculino	Soltero		
9	Femenino	Casado		
10	Femenino	Soltero		
11	Femenino	Soltero		
12	Femenino	Soltero		
13	Masculino	Soltero		
14	Masculino	Soltero		

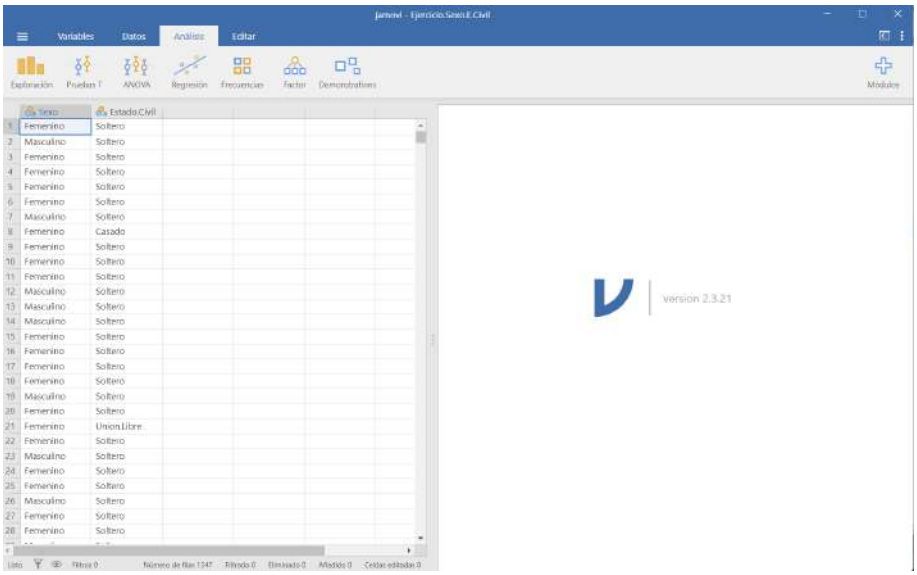
b. **Abrimos Jamovi e ingresamos la base de datos:**
Vamos al botón Inicio del equipo y abrimos Jamovi, la versión que descargamos e instalamos previamente (en esta guía estamos usando *Windows*). Sabrás que el software ha abierto adecuadamente si aparece la siguiente ventana:

Figura 17. Interfaz de Jamovi



Ahora, procedemos a cargar la base de datos, para ello, hacemos clic en las tres barras de la esquina superior izquierda >>> Abrir >>> Navegar y buscamos el archivo de Excel.
La base de datos quedará correctamente cargada cuando aparezcan los datos en el panel izquierdo:

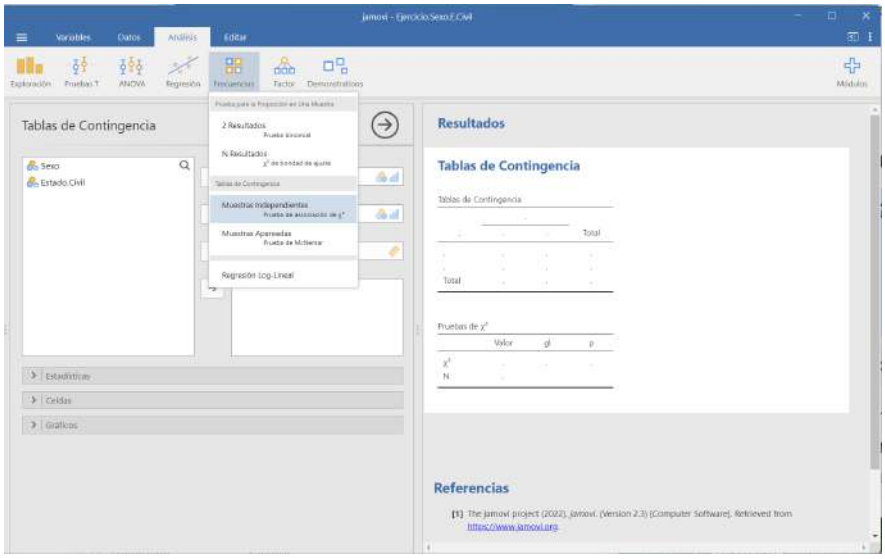
Figura 18. Base de datos cargada en Jamovi



c. **Construyamos las tablas de contingencia:**

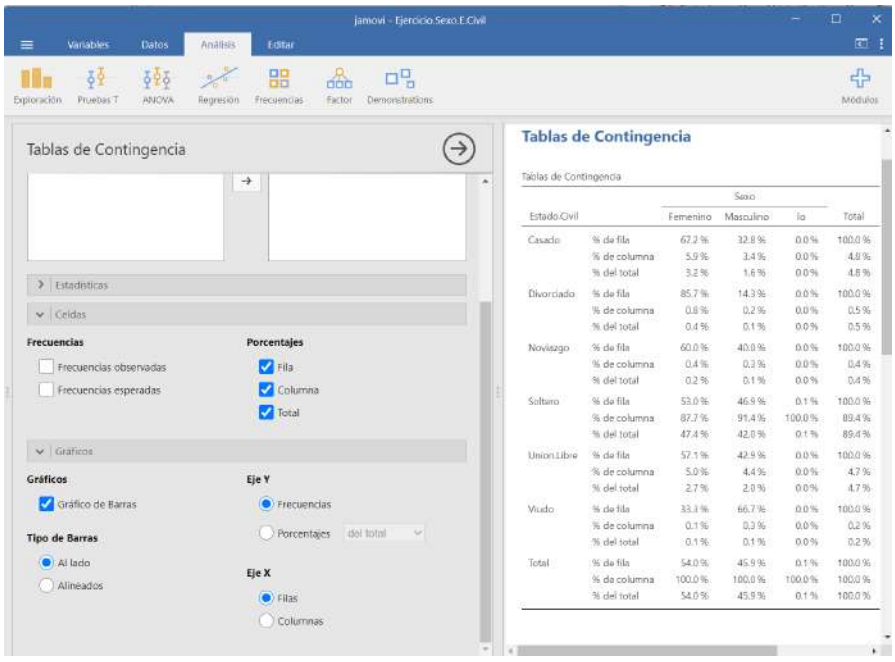
Hagamos clic sobre el menú Análisis, ubiquemos la opción Frecuencias, hagamos clic en Tablas de Contingencia y luego en Muestras independientes. Allí nos ofrece la opción de poner las variables en fila o columna según se deseen analizar:

Figura 19. Menú Análisis



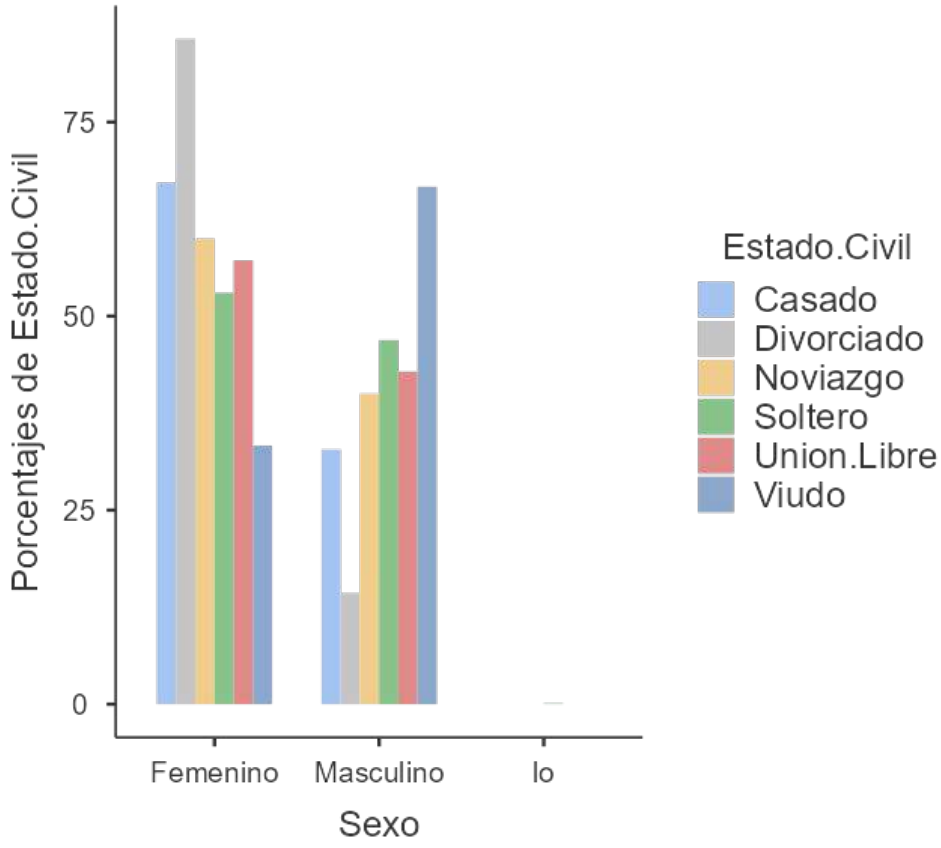
Para el análisis, sugerimos que la variable que tenga más eventos se ponga como variable fila y la de menor número de eventos se ponga como variable columna. De este modo, la variable Estado Civil se va hacia filas y la variable Sexo se va hacia columnas.
Luego, para reportar todas las posibles probabilidades, debemos desplegar la pestaña Celdas y, en Porcentajes, activar las casillas Fila, Columna y Total. Por último, en Gráficos, debemos activar Gráfico de Barras si queremos reportar un gráfico en conjunto con la tabla de contingencia. Tal como lo muestra la siguiente figura.

Figura 20. Porcentajes y gráficos



De esta forma se creará el gráfico:

Figura 21. Gráfico de barras



De esta manera es posible plantear un análisis desde la tabla de contingencia.

Por ejemplo: nos puede interesar analizar los porcentajes filas o las probabilidades fijando el Sexo. En ese sentido, de las mujeres es posible establecer que el 5.9 % de ellas están casadas, el 0.8 % dicen estar divorciadas, el 0.4 % está en relación de noviazgo, el 87.7 % dicen estar solteras y el 5.0 % dicen vivir en unión libre. Por su parte, considerando los hombres, el 3.4 % se encuentran casados, el 0.2 % están divorciados, el 0.3 % tienen una relación de noviazgo, el 91.4 % dice estar soltero y un 0.3 % es viudo.

Frente a la muestra presentada, del total de 1342 participantes, el 54 % son mujeres y el 46 % son hombres. Del total de la muestra el 4.8 % se encuentra casado, el 0.5 % está divorciado, el 0.4 % dice estar en una relación de noviazgo, el 89.4 % está soltero, el 4.7 % en unión libre y el 0.2 % dice estar viuda.

Como un ejercicio para realizar y verificar que, si entendiste los contenidos, te invitamos a poner todas las probabilidades enunciadas en este ejercicio en notación probabilística. ¡Inténtalo!

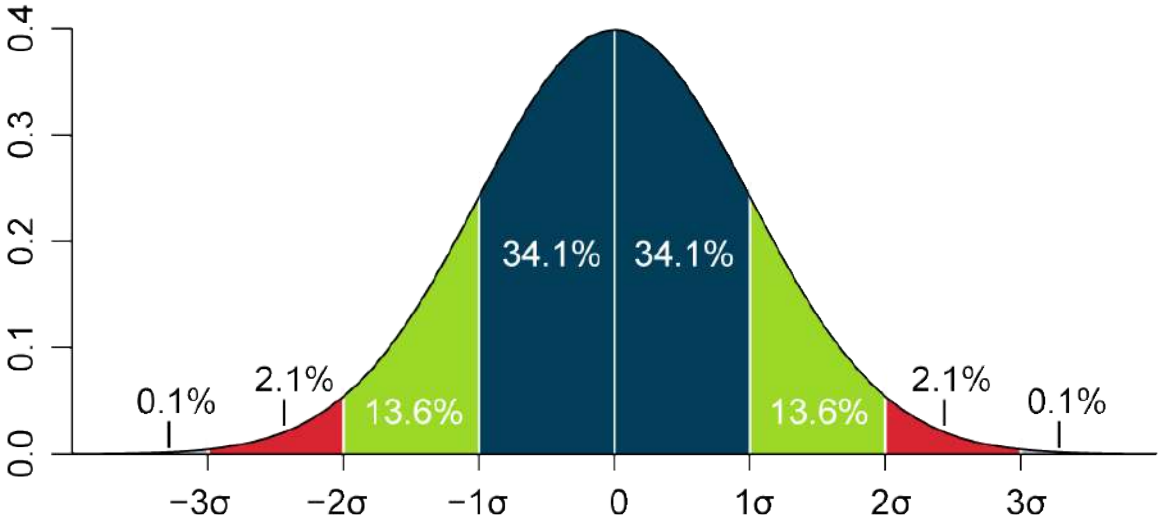
Tema 4. Introducción a las distribuciones de probabilidad

Para el abordaje de este último tema de la *Unidad 1*, revisaremos la definición de distribución de probabilidad a partir de un ejemplo aplicado y realizaremos varios ejercicios. En las siguientes unidades de este curso estudiaremos las distribuciones de probabilidad más utilizadas en los desarrollos de ciencia de datos. ¡Préstale mucha atención a esta parte introductoria!

¿Qué es una distribución de probabilidad?

Es una función, una gráfica o una tabla que relaciona la ocurrencia de un evento con su probabilidad característica. Nos permite estimar probabilidades conjuntas de manera muy ágil.

Figura 22. Distribución de probabilidad



Para comprender el concepto de distribución de probabilidad te proponemos el siguiente ejemplo de aplicación:

Ejemplo: en un artículo de la revista *American Journal of obstetrics and gynecology*, dos autores aseguran que durante 25 años se ha tomado mayor conciencia de los efectos potencialmente dañinos de los medicamentos y químicos en el desarrollo de los fetos. En la siguiente tabla se muestra la prevalencia del consumo de medicamentos prescritos y no prescritos durante el embarazo entre mujeres dadas de alta después del parto en un hospital del este de EUA, entre 1980 y 1982:

Tabla 4. Consumo de medicamentos durante el embarazo

No. Medicamentos	Frecuencias
0	1425
1	1351
2	793
3	348
4	156
5	58
6	28
7	15
8	6
9	3
10	1
11	1

Con base en esta información, construye la distribución de probabilidades y estima algunas.

Para realizar el cálculo de las probabilidades que propondremos a continuación, podemos utilizar Excel, en ese sentido copiamos y pegamos la tabla de frecuencias en Excel. Igualmente, en una columna derecha en blanco podemos poner las probabilidades asociadas a cada uno de los medicamentos. Por ejemplo: *P(X=0)* es la probabilidad de que una mujer no haya consumido ningún medicamento en su embarazo y se estima dividiendo el valor asociado al número de mujeres que no consumieron medicamento (Xo = 1429) entre el total de participantes que corresponde a 4185.

A continuación, se presenta la distribución de probabilidad que sugerimos:

Tabla 5. Distribución de probabilidades asociadas

No. Medicamentos	Frecuencias	Notación	P(X=Xo)
0	1425	P(X=0)	0.34050179
1	1351	P(X=1)	0.32281959
2	793	P(X=2)	0.18948626
3	348	P(X=3)	0.08315412
4	156	P(X=4)	0.03727599
5	58	P(X=5)	0.01385902
6	28	P(X=6)	0.00669056
7	15	P(X=7)	0.00358423
8	6	P(X=8)	0.00143369
9	3	P(X=9)	0.00071685
10	1	P(X=10)	0.00023895
11	1	P(X=11)	0.00023895
Total	4185		1

Con este ejemplo se puede preguntar sobre diferentes probabilidades, conoce el cálculo de algunas que proponemos:

1. ¿Cuál es la probabilidad de que una mujer no haya consumido ningún medicamento?
   Respuesta: sea X los medicamentos que una mujer ha consumido durante su embarazo.

   $$P(X = 0) = 0.34$$

   La probabilidad de que una mujer no haya consumido ningún medicamento es del 34 %. Observa que solo basta con fijarse en la tabla para obtener la respuesta.

2. También podemos calcular la probabilidad de que una persona haya consumido tres o menos medicamentos.

   $$P(X \leq 3) = 0.34 + 0.32 + 0.19 + 0.083 = 0.933$$

   Respuesta: la probabilidad de que una mujer haya consumido 3 o menos medicamentos durante su embarazo es de 0.933. O, también, el 93.3 % de las mujeres ha consumido tres o menos medicamentos durante su embarazo.

3. En esta distribución de probabilidad podemos pedir una probabilidad con mayor. Por ejemplo, ¿cuál es la probabilidad de que una mujer haya consumido más de 5 medicamentos durante su embarazo? De la tabla de distribución de probabilidades tenemos:

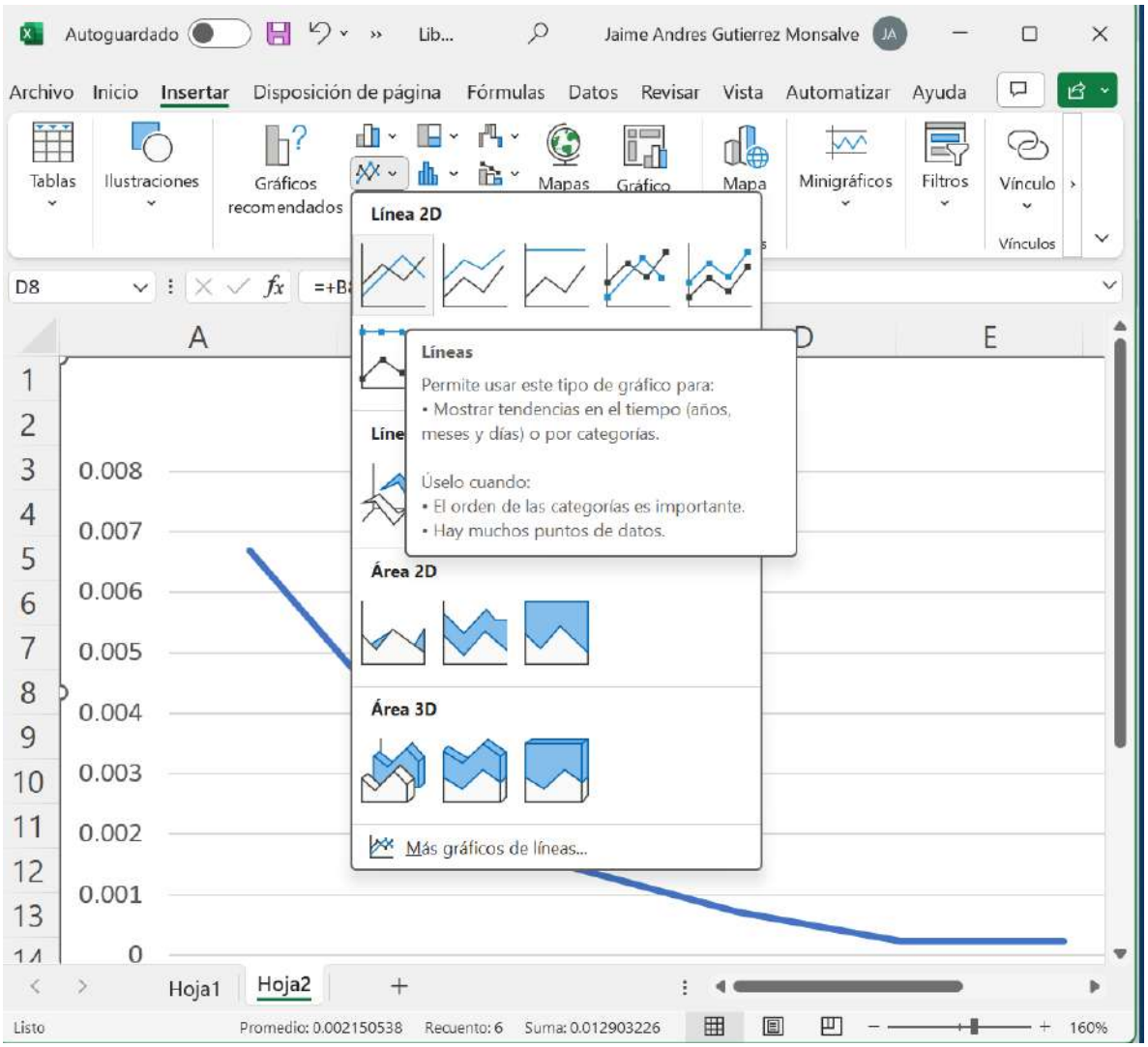
$$P(X > 5) = 0.067 + 0.358 + 0.143 + 0.00072 + 0.00024 + 0.00024 = 0.013$$

Observa que cuando se realiza la suma de las probabilidades no se incluye la probabilidad de que la mujer haya consumido exactamente 5 medicamentos.

Respuesta: la probabilidad de que una mujer haya consumido más de 5 medicamentos es de 0.013, o el 1.3 % de las mujeres incluidas en el estudio han consumido más de 5 medicamentos.

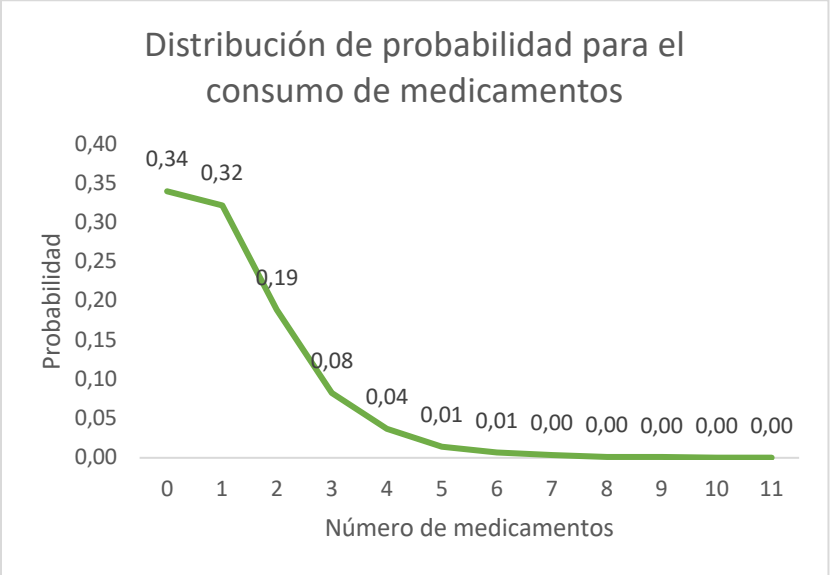
Utilizando Excel, también podemos graficar esta distribución de probabilidad. De esta manera la distribución quedaría de la siguiente manera:

Figura 23. Gráfica en Excel



La representación de la distribución de probabilidad puede representarse como sigue, con la ojiva de frecuencias relativas:

Figura 24. Distribución de probabilidad para el consumo de medicamentos



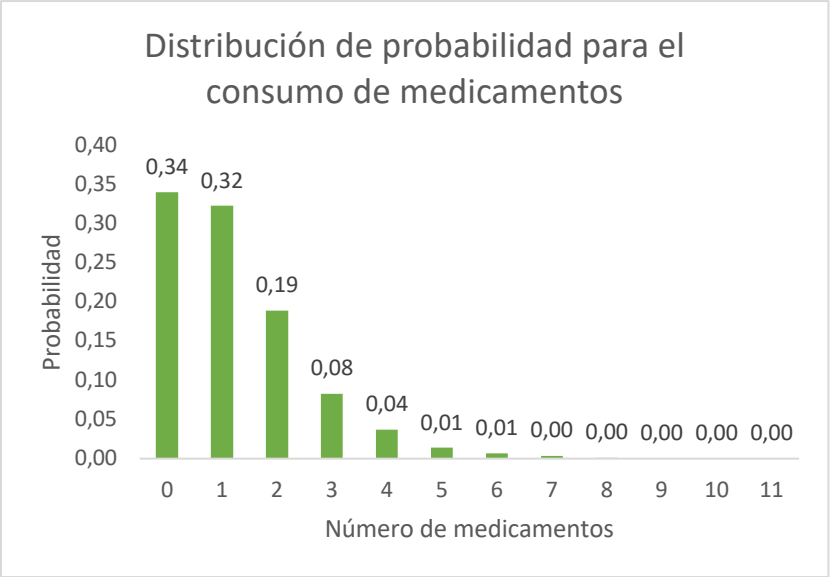
Observa que en el eje Y se presenta la probabilidad asociada a cada uno de los eventos. Así, por ejemplo, podemos estimar la probabilidad de que una mujer haya consumido 1 o menos medicamentos:

$$P(X \leq 1) = P(X = 0) + P(X = 1) = 0.34 + 0.32 = 0.66$$

En ese orden de ideas y observando los valores de la gráfica de distribución de probabilidad, el 66 % de las mujeres han consumido dos o menos medicamentos.

Frente a los gráficos, también pueden encontrarse gráficos con distribución de frecuencias utilizando barras, tal y como te mostramos en la siguiente figura:

Figura 25. Gráfica de barras



Cierre de la Unidad 1

¡Buen trabajo! Culminaste la *Unidad 1* del curso *Estadística II.* Esperamos que te hayas familiarizado con los principios básicos de la probabilidad de un evento, la probabilidad condicional, el teorema de Bayes y las distribuciones de probabilidad. Ya que estos son los insumos básicos para las demás unidades del curso.

Te invitamos a explorar la estadística inferencial a partir de la comprensión de la estimación y el cálculo del tamaño muestral, y a conocer otras distribuciones de variables aleatorias y continuas muy utilizadas en la industria.

Esperamos que hayas aprendido y disfrutado mucho esta unidad. Y que no te hayas quedado solo con la teoría, sino que hayas practicado con todos los ejercicios propuestos.

Referencias de imágenes

Figuras:
Figura 22. Estrategias de Trading (s.f.). Conceptos de estadística para traders – Distribución de probabilidad. https://estrategiastrading.com/conceptos-de-estadistica-para-traders-distribucion-de-probabilidad/

Bibliografía

- Ardila R. M. 2007. Fundamentos de estadística para investigadores en educación; un modelo de investigación científica en educación. Editorial Ecoe. 154p. Colombia.

- Angel G. J. 2007. Estadística general aplicada. Editorial Universidad Eafit. 652 p. Colombia.

- Briones G. 2002. Metodología cuantitativa para las ciencias sociales. Programa de especialización en teoría, métodos y técnicas de investigación social. ARFO Editores e impresos Ltda. 219 p. Colombia.

- Bello, León Darío. Estadística como apoyo a la Investigación. 2005. Editorial L. Vieco e Hijas Ltda.

- Blanco, Castañeda Liliana. 2004. "Probabilidad". Unibiblos. Universidad Nacional de Colombia.

- Box G. E., Hunter J. S. y Hunter W. G. 2008. Estadística para investigadores: diseño, innovación y descubrimiento. Edtorial Reverté. España.

- Castillo M. I. y Guijarro G. M. 2006. Estadística descriptiva y cálculo de probabilidades. Editorial Pearson Education. 425 p. España.

- Daniel, Wayne W. 2010. Bioestadística: Base para el análisis de las ciencias de la salud. Noriega editores, editorial Limusa. Cuarta edición. México.

- Domínguez F y Nieves A. 2010. Probabilidad y estadística para ingeniería, un enfoque moderno. Mc Graw Hill. 548p. México.

- Grisales Romero Hugo. 2002 Estadística Aplicada en Salud Pública: Estadística Descriptiva y Probabilidad. Editorial L-Vieco e Hijas.

- Kennet R. y Shelemyahu Z. 2000. “Estadística Industrial Moderna”. International Thomson Editores. México.

- Levin, J. y Levin William C. 2008. Fundamentos de Estadística en la Investigación social. Editorial Alfaomega. 305 p. México.
- Martínez B. C. 2012. Estadística y muestreo, editorial Ecoe. 874 p. Colombia.
- Martínez B. C. 2012. Estadística Básica Aplicada, editorial Ecoe. Cuarta edición. 388 p. Colombia.
- Milton, J. Susan, Estadística para la Biología y Ciencias de la Salud. Editorial McGraw-Hill. 2007. Tercera Edición ampliada. España.
- Montgomery D. y Runger G. 2009. Probabilidad y estadística aplicada a la ingeniería. Limusa Wiley. 817 p. México.
- Muñoz, Quevedo José María, 2002. "Introducción a la teoría de Conjuntos". Cuarta Edición. Panamericana, formas e impresos.
- Pérez T. H. 2008. Estadística para las ciencias sociales, del comportamiento y de la salud. Editorial Cengage Learning. 815 p. Mexico.
- Pineda A. L. 2009. Estadística. Editorial Pearson Educación. Decima edición, 866p.
- Triola, M. F. (2004). Probabilidad y estadística. Pearson educación.
- Wackerly, D. D. 2002. Estadística matemática con aplicaciones. Editorial International Thomson. Sexta edición. 853 p. México.

IU Digital de Antioquia

INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA