



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

REDES DE INVESTIGACIÓN EN PROBABILIDAD Y  
ESTADÍSTICA EN LATINOAMÉRICA: DESARROLLO  
Y ANÁLISIS ESTADÍSTICO DE UNA BASE DE DATOS  
CON PERSPECTIVA DE GÉNERO.

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

**Matemático Aplicado**

PRESENTA:

**Jorge Salvador Martínez Villafan**

TUTORA:

Alma Saraí Hernández Torres

COTUTORA:

Laura Clementina Eslava Fernández

Ciudad Universitaria, CD. MX., 2023





1. Datos del alumno

Martínez

Villafan

Jorge Salvador

7353590747

Universidad Nacional Autónoma de México

Facultad de Ciencias

Matemáticas Aplicadas

418003839

2. Datos del tutor

Dra.

Alma Saraí

Hernández

Torres

3. Datos del sinodal 1

Dra.

Laura Clementina

Eslava

Fernández

4. Datos del sinodal 2

Dra.

Lizbeth

Naranjo

Albarrán

5. Datos del sinodal 3

Dra.

María Fernanda Gil

Leyva

Villa

6. Datos del sinodal 4

Act.

Luis Enrique

Serrano

Gutiérrez

7. Datos del trabajo escrito

Redes de investigación en probabilidad y estadística en Latinoamérica: desarrollo y análisis estadístico de una base de datos con perspectiva de género.

116 páginas

2023



# Agradecimientos

---

Quiero agradecer a todas las personas que de una u otra forma han contribuido en la realización de esta tesis. A mis padres por haberme brindado la oportunidad de estudiar y apoyarme a lo largo de mis estudios. A mis amigos, por ser mi red de apoyo de manera académica y no académica. A la DGAPA por haberme otorgado la beca PAPIIT con clave de proyecto IN102822 y al Dr. Sergio I. López por su tiempo y ayuda con los trámites de esta beca. A mis profesores por todas sus enseñanzas a lo largo de mi carrera.

En especial, quiero expresar mi agradecimiento a mis tutoras, la Dra. Laura y la Dra. Saraí. Agradezco a la Dra. Laura por confiar en mí y permitirme ser parte de este proyecto que sé que es tan importante para ellas como lo es ahora para mí. A la Dra. Saraí le agradezco su constante apoyo, paciencia y guía durante todo el proceso de investigación. Sin su ayuda este trabajo no habría sido posible.

Finalmente, quiero agradecer a la suerte. Con suerte me refiero a todos aquellos factores externos que estuvieron fuera de mi control, pero favorecieron al desarrollo de este proyecto. El mayor ejemplo de cómo la suerte jugó un papel fundamental en esta tesis, es que aprendí a programar en R en el curso ofrecido por el Taller de Matemáticas de la Facultad de Ciencias, cuyo cupo lo gané en una rifa.

Pero sobre todo, menciono a la suerte porque, en palabras del Dr. Derek Muller, gran parte de la suerte que tenemos se debe a las oportunidades que crearon las personas que estuvieron antes que nosotros. Tomar conciencia de nuestras circunstancias afortunadas nos permite sentir gratitud. Esta gratitud nos hace querer retribuir para mejorar la suerte de los demás mediante la creación de oportunidades. Y crear y mejorar las oportunidades de los y las demás es uno de los objetivos principales del proyecto que la Dra. Laura y la Dra. Saraí encabezan.



## Declaración de autenticidad

---

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Jorge Salvador Martínez Villafan. Ciudad Universitaria, CD. MX., 2023

*„Wir müssen wissen, wir werden wisse.“*  
*“Debemos saber, sabremos.”*

*David Hilbert*





# Resumen

---

Este trabajo documenta aspectos de la brecha de género existente entre investigadores e investigadoras en las áreas de probabilidad y estadística.

El punto de partida es la creación de una base de datos que almacena datos cualitativos y cuantitativos recolectados del 3 de junio al 30 de septiembre de 2022 y del 1 de octubre al 4 de noviembre de 2022, respectivamente. Esta información está disponible públicamente en internet y corresponde a investigadores e investigadoras en las áreas de probabilidad y estadística, que trabajan en Latinoamérica y el Caribe.

Al obtener la base de datos, se realizan análisis de los datos recopilados. Para estudiarlos, primero se utilizan métodos exploratorios con el fin de entender el comportamiento de las variables de nuestra muestra, que identifican el método estadístico más adecuado para realizar una prueba. A su vez, se utilizan herramientas de estadística descriptiva como tablas de contingencia, estadísticas de resumen y gráficos con el objetivo de resumir y describir los registros almacenados en nuestra base de datos.

# Índice general

<b>Agradecimientos</b>	<b>v</b>
<b>Resumen</b>	<b>xI</b>
<b>Índice de figuras</b>	<b>xIV</b>
<b>Índice de tablas</b>	<b>xVI</b>
<b>Introducción</b>	<b>1</b>
<b>1 Fundamentos y marco teórico</b>	<b>3</b>
1.1 Redes de información . . . . .	3
1.2 Redes sociales . . . . .	4
1.3 Brecha de género . . . . .	7
1.4 Cuantificación de la productividad científica . . . . .	7
<b>2 Metodología</b>	<b>11</b>
2.1 Recopilación de datos . . . . .	11
2.1.1 Recolección de datos cualitativos . . . . .	12
2.1.1.1 México . . . . .	12
2.1.1.2 Brasil . . . . .	17
2.1.1.3 Argentina, Chile, Colombia, Cuba, Perú y Uruguay . . . . .	20
2.1.2 Recolección de datos cuantitativos . . . . .	22
<b>3 Análisis y representación de los datos</b>	<b>27</b>
3.1 Descripción de métodos y pruebas estadísticas . . . . .	27
3.1.1 Bootstrap . . . . .	27
3.1.2 Gráfico Q-Q . . . . .	27
3.1.3 La prueba de Shapiro . . . . .	28
3.1.4 La prueba de rango con signo de Wilcoxon de una muestra . . . . .	28
3.1.5 La prueba W-Mann-Whitney . . . . .	28
3.1.6 La función de distribución acumulada . . . . .	28

3.1.7	K-means . . . . .	29
3.1.8	Coeficiente de correlación de Spearman . . . . .	29
3.2	Grado académico . . . . .	30
3.2.1	Dependencia del grado doctorado con el género (inferido) . . . . .	33
3.3	Análisis de género en departamentos . . . . .	36
3.3.1	Diferencias entre la cantidad de hombres y mujeres en departamentos . . . . .	37
3.3.2	Porcentaje de investigadoras por país . . . . .	44
3.4	Independencia del género con las citas recibidas . . . . .	47
3.5	Clústeres . . . . .	51
3.5.1	Clústeres con datos únicamente de mujeres . . . . .	60
3.5.2	Comparación de promedios en artículos y citas por instituciones divididos por género . . . . .	62
3.6	Correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento . . . . .	66
3.7	Diferencias significativas entre los datos de hombres y mujeres . . . . .	70
<b>4</b>	<b>Resultados, alcances y conclusiones</b>	<b>75</b>
4.1	Resultados . . . . .	75
4.2	Alcance de los resultados . . . . .	77
4.3	Conclusiones . . . . .	78
<b>A</b>	<b>Código R</b>	<b>81</b>
<b>B</b>	<b>Fuentes</b>	<b>91</b>
<b>C</b>	<b>Datos crudos del Capítulo 3</b>	<b>95</b>
<b>D</b>	<b>Resultados de la consola de RStudio</b>	<b>101</b>
	<b>Siglas institucionales</b>	<b>103</b>
	<b>Glosario</b>	<b>107</b>

# Índice de figuras

1.1	Red de investigadores(as) de PyE compuesta de tres clústeres que representan distintos departamentos en la UNAM y las aristas indican la relación entre la institución donde se realiza el doctorado y el lugar de trabajo. . . . .	5
1.2	Ejemplo de las conexiones entre algunos investigadores dentro de la red de Erdős, en donde las aristas significan colaboración [10]. . . . .	6
1.3	Ejemplo de cómo se calcula el índice $h$ , inspirado en [25] . . . . .	10
2.1	Red de investigadores(as) de PyE compuesta de tres clústeres que representan distintos departamentos en Brasil y las aristas indican la relación entre la institución donde se realiza el doctorado y el lugar de trabajo. . . . .	18
3.1	Gráfica del recuento de personas cuyo máximo grado es el indicado, donde NA significa desconocido. . . . .	30
3.2	Gráfica de densidad de las 500 réplicas del bootstrap. . . . .	32
3.3	Densidad nula cuando “Doctorado” es independiente de “Género”. . . . .	34
3.4	Investigadores con doctorado y no doctorados por género. . . . .	35
3.5	Grados académicos por género. . . . .	36
3.6	Porcentaje de mujeres por departamento. . . . .	37
3.7	Distribución de hombres y mujeres en los departamentos. . . . .	38
3.8	Función de distribución acumulada de la proporción de mujeres. . . . .	39
3.9	Histograma de investigadores en las instituciones según género. . . . .	40
3.10	Gráfico Q-Q de cantidad de hombres y mujeres. . . . .	41
3.11	Gráfica de la prueba W-Mann-Whitney. . . . .	42
3.12	Comparación de los porcentajes de mujeres observados con los porcentajes de mujeres teóricos sin desigualdad. . . . .	43
3.13	Resultado de la prueba Wilcoxon para determinar si la mediana de los porcentajes de mujeres difiere estadísticamente del 50 %. . . . .	44
3.14	Proporción de hombres y mujeres dentro de los países. . . . .	45

3.15 Proporción de hombres y mujeres dentro de las instituciones en México. . . . .	46
3.16 Categorías de citas por género. . . . .	49
3.17 Categorías “Alto” y “NoAlto” divididas por género. . . . .	50
3.18 Densidad nula cuando “Citas” y “Género” son independientes. . . . .	51
3.19 Densidad del promedio de citas de las 41 instituciones. . . . .	53
3.20 Gráfica de distancias. . . . .	54
3.21 Recomendaciones de dos métodos sobre el número óptimo de clústeres. . . . .	55
3.22 Clústeres de Instituciones según su nivel en citas y artículos. . . . .	56
3.23 Gráficas de la función de probabilidad acumulada de cada clúster. . . . .	58
3.24 Comparación de los porcentajes de mujeres entre clústeres. . . . .	59
3.25 Gráfica de distancias entre instituciones, únicamente con datos de mujeres. . . . .	60
3.26 Clústeres de instituciones según su nivel en citas y artículos, únicamente con datos de mujeres. . . . .	61
3.27 Distancia entre instituciones, con datos de instituciones divididas por género. . .	63
3.28 Clústeres de instituciones según su nivel en citas y artículos, con datos de instituciones divididas por género. . . . .	65
3.29 Gráfico de dispersión de medias de citas y artículos, con datos de instituciones divididas por género. . . . .	65
3.30 Gráfico de dispersión con datos de investigadores e investigadoras individualmente.	66
3.31 Promedio de citas por universidad. . . . .	67
3.32 Promedio de artículos por universidad. . . . .	68
3.33 Matriz de correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento. . . . .	69
3.34 Gráfica Q-Q de citas. . . . .	70
3.35 Comparación de citas de investigadores masculinos y femeninos utilizando la Prueba W-Mann-Whitney. . . . .	71
3.36 Comparación de artículos de investigadores masculinos y femeninos utilizando la Prueba W-Mann-Whitney. . . . .	73
D.1 Resumen de 30 métodos. . . . .	101
D.2 Resumen de 30 métodos, únicamente con datos de mujeres. . . . .	102
D.3 Resumen de 30 métodos, con datos de instituciones divididas por género. . . . .	102

# Índice de tablas

1.1 Ejemplo de cómo se calcula el índice $h$ . . . . .	10
2.1 Información recopilada en la base de datos. . . . .	12
2.2 Número de investigadores(as) por universidad en México. . . . .	13
2.3 Número de investigadores(as) por universidad en Brasil. . . . .	19
2.4 Número de investigadores(as) por país. . . . .	20
2.5 Número de investigadores(as) por universidad en Argentina. . . . .	20
2.6 Número de investigadores(as) por universidad en Chile. . . . .	21
2.7 Número de investigadores(as) por universidad en Colombia. . . . .	21
2.8 Número de investigadores(as) por universidad en Cuba. . . . .	21
2.9 Número de investigadores(as) por universidad en Perú. . . . .	22
2.10 Número de investigadores(as) por universidad en Uruguay. . . . .	22
2.11 Número de investigadores con datos numéricos conocidos por universidad en México. .	24
2.12 Número de investigadores con datos numéricos conocidos por universidad en Brasil. .	25
3.1 Número de investigadores según su grado de estudios . . . . .	30
3.2 Proporciones generadas en las primeras 10 réplicas de bootstrap. . . . .	31
3.3 Intervalos de confianza por grado académico. . . . .	32
3.4 Resultados de las dos primeras replicas. . . . .	34
3.5 Primeros 10 resultados de los 500 conjuntos generados. . . . .	34
3.6 Comparación de género entre los niveles de grado “Doctorado” y “Nodoctorado”. .	35
3.7 Comparación de género en los diferentes grados académicos: licenciatura, maestría y doctorado. . . . .	36
3.8 Cantidad de investigadores e investigadoras dentro de la base de datos. . . . .	37
3.9 Muestra de instituciones con su respectiva cantidad de hombres y mujeres, ordenados longitudinalmente. . . . .	39

## ÍNDICE DE TABLAS

---

3.10 Resultado de la prueba Shapiro para la variable “Cantidad” en ambos géneros. . . . .	41
3.11 Porcentaje de hombres y mujeres por país en la base de datos. . . . .	45
3.12 Porcentajes de investigadores por género en instituciones en México. . . . .	47
3.13 Muestra de categorización de investigadores según el número de citas obtenidas (Bajo, Medio, Medio alto, Alto). . . . .	48
3.14 Comparación de género en las distintas categorías según su cantidad de citas. . . . .	48
3.15 Comparación de género en los niveles de citas “No alto” y “Alto”. . . . .	49
3.16 Diferencia entre 10 conjuntos de datos generados. . . . .	50
3.17 Muestra de 15 instituciones. . . . .	52
3.18 Muestra de promedios de citas y artículos. . . . .	52
3.19 Datos centralizados. . . . .	53
3.20 Datos centralizados para instituciones A y B. . . . .	54
3.21 Instituciones divididas en clústeres según su nivel en citas y artículos. . . . .	55
3.22 Instituciones del clúster número 1 y sus porcentajes de mujeres y hombres. . . . .	57
3.23 Instituciones del clúster número 2 y sus porcentajes de mujeres y hombres. . . . .	57
3.24 Instituciones divididas en clústeres según su nivel en citas y artículos, únicamente con datos de mujeres. . . . .	61
3.25 Muestra centralizada de promedios de citas y artículos con transformación logarítmica, con datos de instituciones divididas por género. . . . .	62
3.26 Instituciones divididas en clústeres según su nivel en citas y artículos, con datos clasificados por género. . . . .	64
3.27 Muestra de instituciones con sus promedios de citas y artículos y sus porcentajes de mujeres y hombres respectivamente. . . . .	67
3.28 Resultado de la prueba Shapiro para la variable “Citas” en ambos géneros. . . . .	70
 B.1 Universidades en Brasil y sus fuentes de información correspondientes. . . . .	91
B.2 Universidades en Argentina y sus fuentes de información correspondientes. . . . .	92
B.3 Universidades en Chile y sus fuentes de información correspondientes. . . . .	92
B.4 Universidades en Colombia y sus fuentes de información correspondientes. . . . .	93
B.5 Universidades en Cuba y sus fuentes de información correspondientes. . . . .	93
B.6 Universidades en Perú y sus fuentes de información correspondientes. . . . .	93
B.7 Universidades en Uruguay y sus fuentes de información correspondientes. . . . .	93
 C.1 Porcentaje de hombres y mujeres dentro de cada departamento. . . . .	95

## ÍNDICE DE TABLAS

---

C.4	Instituciones del clúster número 1 y sus porcentajes de mujeres y hombres. . . . .	96
C.5	Instituciones del clúster número 2 y sus porcentajes de mujeres y hombres. . . . .	97
C.2	Promedios de citas y artículos de cada universidad, únicamente con datos de mujeres.	99
C.3	Muestra centralizada de promedios de citas y artículos de universidades con transformación logarítmica, únicamente con datos de mujeres. . . . . . . . . .	100

# Introducción

---

Según datos de la UNESCO, se estima que solo el 30% de las personas que realizan investigación académica en el mundo son mujeres [1]. Este porcentaje varía dependiendo del campo de investigación. Por ejemplo, desde 1990, las mujeres constituyen más de la mitad de todos los científicos en áreas como sociología y psicología [2]. No obstante, en los campos relacionados con la ciencia, tecnología, ingeniería y matemáticas (STEM por sus siglas en inglés), las mujeres solo representan alrededor del 28% [3]. Aunque el número de mujeres en estas áreas casi se ha duplicado desde 1993 al 2010 [2], la brecha de género sigue estando presente.

Algunos factores clave, que perpetúan la brecha de género en estas áreas, son los estereotipos de género, el ambiente dominado por hombres y los pocos modelos femeninos a seguir en estas áreas [3]. Aunque en la educación básica los hombres y las mujeres reciben clases de matemáticas y ciencias por igual, el número de mujeres estudiando una carrera relacionada con las STEM se reduce drásticamente [4]; una de las consecuencias de este hecho es que existan pocas mujeres realizando investigación en ciencias exactas.

Entre los esfuerzos internacionales para tratar de reducir la brecha de género en las áreas STEM, la UNESCO ha creado el Proyecto SAGA (STEM and Gender Advancement), el cual tiene como objetivo aumentar la visibilidad, la participación y el reconocimiento de las contribuciones de las mujeres en Ciencia, Tecnología, Ingeniería y Matemáticas [5]. A su vez, la ONU en 2015 declaró al 11 de febrero como el “Día internacional de las Mujeres y las Niñas en la Ciencia” como reconocimiento al trabajo que las mujeres desempeñan en la ciencia.

Esta tesis forma parte del proyecto “Redes orgánicas entre las mujeres probabilistas en Latinoamérica y el Caribe”. Este proyecto inició en el evento “Latin American and Caribbean Workshop on Mathematics and Gender”, en Casa Matemática Oaxaca. Como lo indica su nombre, uno de los objetivos de ese proyecto es la construcción de redes basadas en conexiones, auténticas y duraderas, entre mujeres interesadas en realizar una carrera en investigación en los campos de la probabilidad y la estadística.

Como señalamos en los párrafos anteriores, la desigualdad entre hombres y mujeres en las áreas STEM es un problema que ha persistido a lo largo del tiempo y aún es una preocupación importante en la actualidad, por lo que es fundamental estudiarlo desde una perspectiva específica. En este sentido, esta tesis se enfoca en las áreas de probabilidad y estadística en nuestra región, lo que representa un enfoque novedoso en el tema que no ha sido explorado antes, hasta lo que sabemos.

## ÍNDICE DE TABLAS

---

En esta tesis se describe la recolección de datos cualitativos y cuantitativos, realizada del 3 de junio al 30 de septiembre de 2022 y del 1 de octubre al 4 de noviembre del mismo año, respectivamente, de investigadores en probabilidad y estadística trabajando en Latinoamérica, con el fin de crear una base de datos. Se analiza la base de datos con pruebas estadísticas de distintos tipos para documentar las diferencias estadísticas entre hombres y mujeres de acuerdo a los datos disponibles en Internet. Se analiza la proporción de grados académicos dentro de la base de datos, la independencia del género con tener un doctorado, la independencia del género del autor(a) de un artículo y el número de citas que recibe, el porcentaje de hombres y mujeres dentro de los departamentos, el promedio de artículos y citas para cada género, la correlación de la proporción de cada género, dentro de cada institución, con la productividad en términos de artículos y citas. Además, se cuantifica y analiza el impacto de los artículos considerados en estas áreas de investigación. Los resultados son presentados en esta tesis.

En el Capítulo 1 presentamos conceptos básicos en redes e igualdad de género, los cuales fundamentan el resto del trabajo.

En el Capítulo 2 se detalla la metodología utilizada para recolectar datos en algunos países de Latinoamérica con el fin de construir la base de datos. Durante este proceso, se hizo uso de las redes sociales y del fenómeno matemático de la aparición de clústeres o comunidades en redes, el cual resultó útil para identificar los departamentos de matemáticas y estadística en los países sudamericanos.

En el Capítulo 3 se presenta el análisis estadístico de la base de datos recolectada. Este análisis incluye una exploración descriptiva de las variables relevantes y una parte de inferencia estadística.

Finalmente, se presentan los resultados y conclusiones en el Capítulo 4.

Con el propósito de simplificar las referencias, se utilizará la expresión “código R en A.X” para señalar el código empleado en el Apéndice A.X, donde X representa el número asignado a cada uno de los apéndices, que contienen los códigos utilizados en este estudio.

La continuación del proyecto “Redes orgánicas entre las mujeres probabilistas en Latinoamérica y el Caribe” constará de dos fases. En la primera, se creará una página web donde se mostrarán los resultados obtenidos en esta tesis y se presentará la información disponible en nuestra base de datos. Esta página web tiene el objetivo de facilitar la creación de redes orgánicas entre aquellas que desean ser investigadoras o continuar con su trabajo de investigación. Además, servirá como un medio para visibilizar el trabajo de las mujeres investigadoras en probabilidad y estadística en América Latina y el Caribe, y concientizar sobre los retos que aún debemos superar en cuanto a la brecha de género en la investigación en estas áreas de estudio.

En la segunda fase del proyecto se llevará a cabo un programa de lectura, el cual consistirá en vincular a estudiantes de posgrado con estudiantes de licenciatura, que tengan interés en explorar más a fondo algún campo relacionado con probabilidad o estadística. Tanto el programa de lectura como la información que estará disponible en la página web, junto con las áreas de investigación, universidad de trabajo, país, tendrán el propósito de crear y fortalecer de redes de investigación entre mujeres, equiparando así las existentes entre hombres, ya que es sabido que los hombres tienden a colaborar más entre hombres [6].

---

## Capítulo 1

# Fundamentos y marco teórico

---

Los dos ejes de este trabajo son las redes y la brecha de género. El término red, siguiendo [7], se define como un conjunto de puntos unidos en pares mediante líneas. A dichos puntos nos referimos como nodos y a las líneas que los unen como aristas. Las redes están presentes en una gran variedad de los campos de estudio. Respecto a su aplicación, podemos dividirlas en cuatro categorías: redes tecnológicas, redes de información, redes sociales y redes biológicas [7]. En nuestro trabajo haremos uso de las redes de información y las redes sociales. A continuación, en las Secciones 1.1 y 1.2 describiremos cómo aparecen las redes de información y las redes sociales en este trabajo. En la Sección 1.3 introduciremos conceptos de género y en la Sección 1.4 veremos cómo las redes de investigación valúan el trabajo de los investigadores e investigadoras que las integran.

### 1.1. Redes de información

Las redes de información son estructuras que describen la organización de los cuerpos de información. Un claro ejemplo es la World Wide Web, en donde las páginas actúan como nodos y los hipervínculos, es decir, los enlaces que sirven para ir de una página a otra, como aristas.

Las redes de información que estudiamos en este trabajo son las redes de citas, las cuales se crean al momento de que los investigadores hacen referencia al trabajo de otros académicos. En las redes de citas, los investigadores actúan como nodos y las referencias como aristas. Se define que investigadores A y B están conectados entre sí, si la investigadora A hace referencia al trabajo de otro investigador B.

Al momento de trabajar con redes de citas, es común encontrar el fenómeno de agrupamiento dentro de una misma red, al cual nos referimos por “clúster” o comunidades en redes [7]. Podemos pensar en el clustering como la formación de redes menores dentro de una red mayor. La ocurrencia

## 1. FUNDAMENTOS Y MARCO TEÓRICO

---

de clustering, dentro de las redes de citas, puede ocurrir debido a distintos factores. Por ejemplo, la agrupación de citas entre artículos que abordan el mismo tema o la agrupación de citas entre investigadores que pertenecen a una misma institución o departamento. Dentro del agrupamiento por temas relacionados pueden ocurrir otros dos fenómenos: la co-citación y el acoplamiento bibliográfico. El primer caso ocurre cuando dos artículos son citados por un tercer artículo, mientras que el acoplamiento bibliográfico ocurre si dos artículos citan a la misma lista de referencias.

Otro ejemplo de la ocurrencia de clustering sucede cuando investigadores de una misma área, en este caso probabilidad y estadística (PyE), pertenecen a un mismo departamento o institución.

En la Figura 1.1<sup>1</sup> se muestra una representación de algunos de los investigadores y las investigadoras en probabilidad y estadística dentro de la UNAM. Cada clúster representa un departamento académico y cada nodo representa la universidad de su último grado de estudio<sup>2</sup>, mientras que las aristas representan la conexión entre la universidad de estudio y el departamento. Es decir, cada arista representa a un investigador que estudió en determinada universidad y trabaja en determinado departamento. Por ejemplo, el nodo UNAM tiene dos aristas que la relacionan con el IMATE. Lo que significa que dos investigadores del IMATE realizaron sus estudios de doctorado en el posgrado de la UNAM<sup>3</sup> (código R en A.7).

### 1.2. Redes sociales

Una red social es un concepto teórico-social, el cual representa el tejido de contactos y sus relaciones entre los miembros de una misma comunidad [8], y representa a un grupo de personas que pertenecen a una misma organización. En nuestro caso, estas organizaciones son departamentos e institutos. Formalmente, podemos pensar en una red social donde los investigadores actúan como nodos y las aristas corresponden a la existencia de cierta interacción social entre ellos.

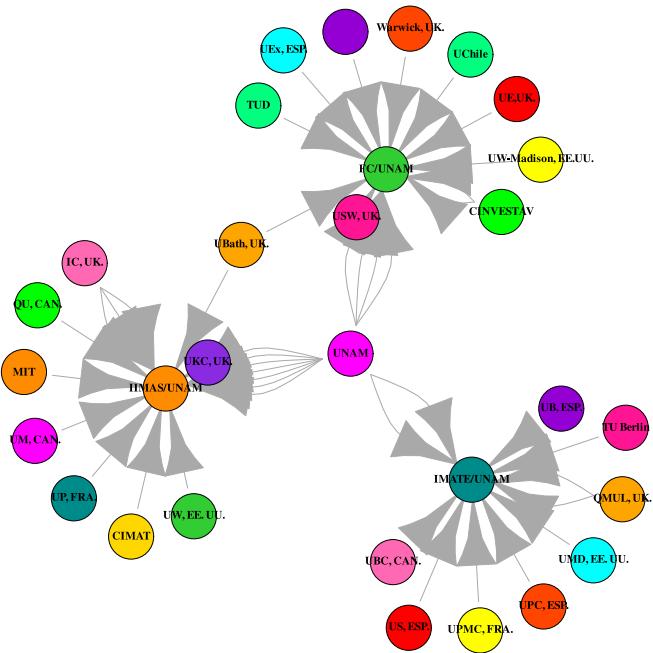
En nuestro trabajo, nos centraremos en las relaciones de colaboración en trabajos académicos, asesorías y en la relación que dos o más investigadores tienen por pertenecer a una misma organización. Decimos entonces que dos investigadores están relacionados si colaboran para realizar un artículo académico; dicha relación es simétrica, ya que la investigadora A está relacionada con el investigador B mediante la vinculación de colaboración, y el investigador B está relacionado con la investigadora A mediante el mismo vínculo. De manera similar, dos o más investigadores están relacionados si trabajan para un mismo instituto o departamento. En esta situación la relación también se considera simétrica. Por último, decimos que dos investigadores están vinculados mediante el asesoramiento si un investigador es asesor de otro, en este caso la relación es asimétrica, ya que una investigadora A puede ser asesora de un investigador B, pero el investigador B no puede ser asesor de la investigadora A. (véase las Secciones 2.1.1.2 y 2.1.2, donde esta relación resulta útil para la construcción de la base de datos).

---

<sup>1</sup>El significado de las siglas se encuentra en la Sección Siglas institucionales.

<sup>2</sup>Los nodos sin etiquetas indican que la universidad de estudio es desconocida.

<sup>3</sup>Por cuestiones de privacidad se decidió utilizar el nombre de la universidad de estudio en lugar del nombre del investigador(a)

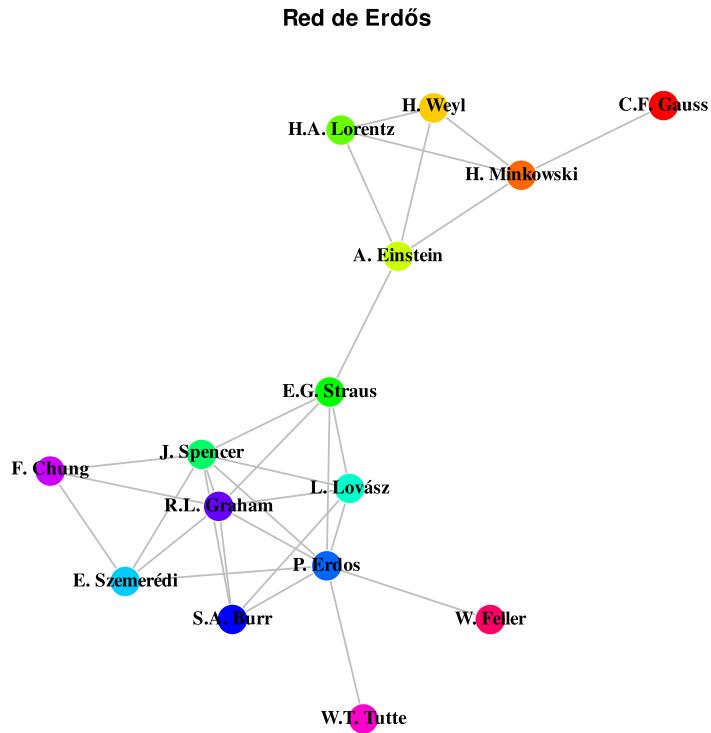


**Fig. 1.1:** Red de investigadores(as) de PyE compuesta de tres clústeres que representan distintos departamentos en la UNAM y las aristas indican la relación entre la institución donde se realiza el doctorado y el lugar de trabajo.

## 1. FUNDAMENTOS Y MARCO TEÓRICO

---

Para ejemplificar el concepto de red de colaboración académica consideremos la “Red de Erdős”, la cual es una famosa red de colaboración científica. Lleva su nombre en honor al matemático húngaro Paul Erdős, quien escribió alrededor de 1500 artículos. La mayoría de estos artículos fueron en coautoría con 509 colaboradores directos [9]. Dentro de dicha red, el número de Erdős es una medida de la distancia académica entre cualquier matemático y Paul Erdős, respecto a la red de colaboración. El matemático húngaro tiene un número de Erdős igual a 0, quienes han coescrito con Paul Erdős tienen un número Erdős igual a 1, quienes han colaborado con ellos, pero no con Erdős, reciben un número de Erdős igual a 2, y así sucesivamente. En la Figura 1.2 vemos a algunos investigadores famosos dentro de la red de Erdős (código R en A.8).



**Fig. 1.2:** Ejemplo de las conexiones entre algunos investigadores dentro de la red de Erdős, en donde las aristas significan colaboración [10].

### 1.3. Brecha de género

Para poder definir la «brecha de género» es importante comenzar definiendo el término género. Aunque coloquialmente sexo y género son utilizados como sinónimos, estos dos términos se refieren a conceptos distintos. El término sexo se refiere a las características cromosómicas, hormonales, anatómicas y fisiológicas de las personas [11]. Por su parte, el término género es una construcción social a partir de sexo [11]. Esta construcción social es un conjunto de ideas, creencias y atribuciones sobre las características que una persona debe poseer a partir de su sexo biológico, “resaltando culturalmente sus diferencias – habilidades y/o aptitudes –, las cuales impulsan o inhiben comportamientos y conductas en el conjunto de cada sexo” (Rendón, 2003).

Al hablar de brecha de género nos referimos a la disparidad existente, en el mundo laboral, político, etcétera, entre hombres y mujeres, en cuanto a recursos, derechos u oportunidades [12] (es importante mencionar que también existen personas que se identifican como no-binarias). Esta disparidad tiene distintos orígenes y agravantes tales como la clase social, la etnia, entre otros. Como mencionamos anteriormente, a las personas se les asignan distintos roles sobre lo que deben ser y cómo comportarse de acuerdo a su sexo biológico [13]. Esta asignación de características provocan que algunos oficios o profesiones sean feminizados o masculinizados. Es decir, existen algunos trabajos que socialmente son atribuidos a determinado género, como la enfermería para las mujeres o las ingenierías para los hombres. Entre los trabajos masculinizados también se encuentran las ciencias exactas [14].

La brecha de género presente en el mundo de la investigación científica puede tener varios orígenes, como la feminización o masculinización de las profesiones [15] o incluso la brecha de género en la educación, ya que, según datos de la UNESCO, en 2014 el 53 % de los graduados de licenciatura y maestría fueron mujeres, pero a nivel de doctorado la proporción de mujeres graduadas se reduce al 44 % [16].

El estudio de las causas de la brecha de género en el mundo de la investigación académica [17] está fuera de la naturaleza matemática de este proyecto. Por lo cual, el objetivo de este trabajo es investigar si existen diferencias significativas entre el trabajo de las mujeres y los hombres en investigación en las áreas de probabilidad y estadística en los países latinoamericanos en nuestra base de datos.

### 1.4. Cuantificación de la productividad científica

En México, la práctica científica profesional es relativamente reciente. Fue en 1970 que se fundó la institución antecesora del Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT, anteriormente CONACYT), la Academia de la Investigación Científica y del Instituto Nacional de la Investigación Científica. En 1984 se creó el Sistema Nacional de Investigadores (SNI) para apoyar a los investigadores mexicanos y favorecer que su trabajo se realice en México [18].

Originalmente, el SNI contaba con 1,396 miembros, mientras que para el año 2014 el número de investigadores había aumentado a 21,359. Los investigadores dentro del SNI se distribuyen en

## **1. FUNDAMENTOS Y MARCO TEÓRICO**

---

tres categorías de reconocimiento: como candidato a investigador nacional, investigador nacional en nivel I, II, III o como investigador nacional emérito [19].

La evaluación de los investigadores en el Sistema Nacional de Investigadores se basa en la evaluación por pares. Las comisiones dictaminadoras, compuestas por expertos en la materia miembros del SNI, evalúan el nivel de cada investigador a través del análisis de sus productos de investigación, su currículum y su reputación. Para lograr la mayor objetividad y evitar dictámenes subjetivos, numerosas comisiones optan por cuantificar el desempeño de los científicos a través de la cantidad de artículos publicados y la cantidad de citas recibidas. Esta práctica da mayor jerarquía a aquellas personas que publican más artículos y son más citadas. Esto da como resultado que los investigadores tienen más incentivos para publicar, con el fin de mejorar su posición en el sistema, que para comprender un fenómeno y contribuir a la solución de problemas importantes. El uso de estadísticas bibliométricas para evaluar el desempeño de todos los científicos del país se considera por algunos un “problema de una evaluación que privilegia la cantidad sobre la calidad del trabajo” [18].

Este es un problema presente en varios países que genera preocupación entre los investigadores, como por ejemplo en Brasil, donde investigadores han expresado al Consejo Nacional de Desarrollo Científico y Tecnológico (CNPq) y a la Coordinación de Perfeccionamiento de Personal de Nivel Superior (CAPES, en sus siglas en portugués) la necesidad de mejorar los procesos de evaluación en la producción científica [20]. Se destaca la importancia de valorar la calidad por encima de la cantidad, ya que el excesivo énfasis en el número de artículos publicados puede conducir a la creación de colaboraciones “artificiales” con el fin de aumentar la cantidad de publicaciones [21].

La necesidad de publicar para avanzar en las categorías de reconocimiento plantea un problema para aquellos que tienen dificultades para producir contenido con regularidad, ya sea por razones personales, sociales o de situación laboral.

Las mujeres enfrentan obstáculos adicionales en la investigación, como la responsabilidad familiar, la dificultad para obtener grados académicos, entre otros [19]. Esto puede limitar su producción de artículos y, consecuentemente, su avance entre las categorías de reconocimiento. Si analizamos los datos del 2014 en México, de los 21,359 investigadores, 34% son mujeres. Dentro de las mujeres en el SNI, 58% se posicionaban en el nivel I; mientras que el 21% de estas eran candidatas, el 15% tenían el nivel II y por último, solo el 5% se posicionaba en el nivel III [19]. Para el año 2022, del total de investigadores en los niveles SNI II, SNI III y Emeritazgo, solo el 20% eran mujeres [22].

### **Índice *h***

Una métrica de la productividad científica es el índice *h*. El objetivo del índice *h* es medir tanto la productividad como el impacto del trabajo publicado de un científico o académico. Este índice fue propuesto por J.E. Hirsch en 2005 [23].

El índice *h* es una métrica cuantitativa basada en el análisis de datos de publicaciones y citas para proporcionar una estimación de la importancia, el significado y el impacto acumulado de las contribuciones de investigación de un científico o académico.

La interpretación del índice *h* es la siguiente, según Hirsch [24]:

## 1.4 Cuantificación de la productividad científica

---

- Un índice  $h$  de 20 después de 20 años de actividad científica indica a un científico reconocido.
- Un índice  $h$  de 40 después de 20 años de actividad científica, caracteriza a científicos destacados, que probablemente se encuentren en universidades prestigiosas o en los principales laboratorios de investigación.
- Un índice  $h$  de 60 después de 20 años, o 90 después de 30 años, caracteriza individuos verdaderamente únicos.

A continuación enunciaremos sus ventajas, desventajas y controversias.

### Ventajas

- Mide el impacto acumulativo de la producción y el desempeño académico de un autor; mide tanto la cantidad como la calidad, comparando las publicaciones con las citas.
- Corrige el peso desproporcionado de publicaciones muy citadas o publicaciones que aún no han sido citadas.
- Varios recursos, como Google Scholar y ResearchGate, calculan automáticamente el índice  $h$  como parte de los informes de citas para los autores.

### Desventajas y controversias

- Evalúa todo el cuerpo de producción académica de un autor; no está destinado a un período de tiempo específico.
- Es insensible a las publicaciones que rara vez se citan, como los resúmenes de reuniones, y a las publicaciones que se citan con frecuencia, como las reseñas.
- Los problemas de variantes en el nombre del autor y las múltiples versiones del mismo trabajo plantean desafíos para establecer datos automáticos de citas precisos para un autor específico. Este problema sucede frecuentemente en Latinoamérica, cuyas convenciones en los nombres son distintas a las anglosajonas.
- No proporciona el contexto de las citas.
- No penaliza las citas que un autor hace a su propio trabajo o las citas innecesarias entre colegas que pueden sesgar el índice  $h$ .
- Variará entre las referencias y sitios en línea —como ResearchGate o Google Scholar— según los datos de publicación que se incluyan en el cálculo del índice.
- Ignora la clasificación de los autores y las características de los coautores en las publicaciones.

### Cálculo del índice $h$

Se calcula ordenando de mayor a menor los artículos científicos según el número de citas recibidas, siendo el índice  $h$  el valor en el que coinciden el número de orden con el número de

## 1. FUNDAMENTOS Y MARCO TEÓRICO

---

citas. Por ejemplo, un investigador tiene un índice  $h$  igual a 5 si tiene 5 publicaciones con al menos 5 citas cada uno. Se muestra un ejemplo en la Tabla 1.1 y la manera gráfica de calcularlo en la Figura 1.3 (código R en A.26).

Artículos ordenados por número de citas	Número de citas
1	15
2	10
3	10
4	8
5	5
6	4
7	4
8	2
9	2
10	1

Tabla 1.1: Ejemplo de cómo se calcula el índice  $h$ .

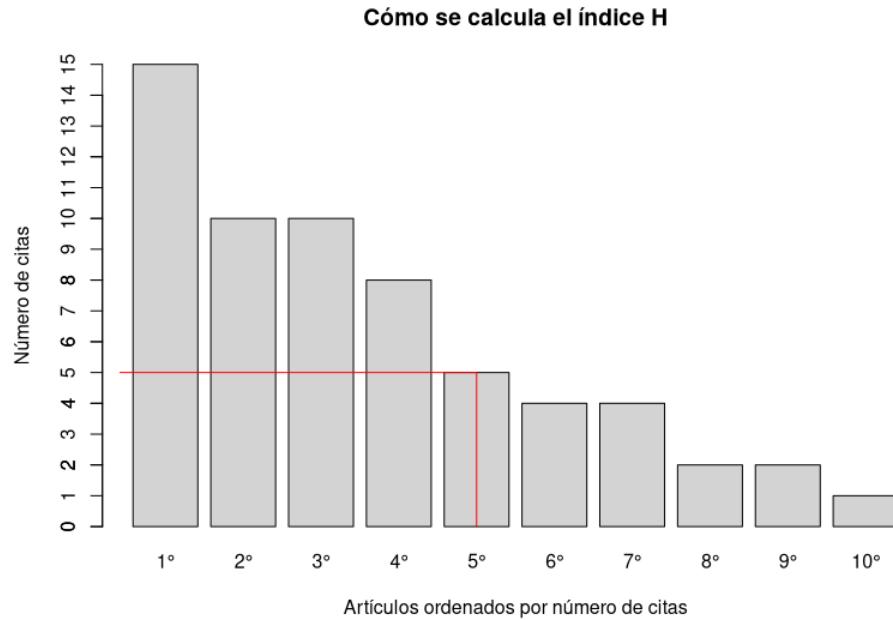


Fig. 1.3: Ejemplo de cómo se calcula el índice  $h$ , inspirado en [25]

### Índice $i_{10}$

El índice  $i_{10}$  es una variación del índice  $h$  que únicamente considera las publicaciones que se han citado al menos 10 veces.

---

## Capítulo 2

# Metodología

---

El trabajo de esta tesis consiste en dos partes, la recopilación de datos para la creación de una base de datos y el análisis estadístico de los datos obtenidos. La parte de estadística estará comprendida, a su vez, de dos partes principales: la estadística inferencial y la estadística descriptiva.

La estadística inferencial nos permite extraer conclusiones acerca de la población a partir del análisis e interpretación de nuestra muestra. En este trabajo, la población corresponde a los investigadores e investigadoras cuya información es accesible por Internet. Para ello se utilizarán distintos métodos estadísticos, como el bootstrap, las pruebas de hipótesis como la prueba Mann-Whitney, entre otros. Cabe señalar que hacemos un análisis exploratorio utilizando métodos gráficos y estadísticos, como la prueba Shapiro-Wilks que nos permite medir la “normalidad” de una variable, todo esto con el fin de determinar qué prueba estadística será conveniente utilizar. Por otra parte, en la estadística descriptiva, usamos tablas, gráficos y medidas descriptivas, con las cuales resumiremos y describiremos los datos de nuestra muestra poblacional.

### 2.1. Recopilación de datos

En la primera parte del proyecto se recolectaron datos cualitativos y cuantitativos disponibles públicamente en Internet, de investigadores enfocados en las áreas de probabilidad y estadística que laboran en departamentos de Latinoamérica y el Caribe. La información de los investigadores que consideramos relevante de almacenar se muestra en la Tabla 2.1.

La recolección de datos se dividió en dos partes: recolección de datos cualitativos y recolección de datos cuantitativos. Discutiremos la metodología de la recolección de datos cualitativos en la Sección 2.1.1 y la metodología correspondiente a los datos cuantitativos en la Sección 2.1.2.

## 2. METODOLOGÍA

---

	Información
1	Nombre(s)
2	Primer apellido
3	Segundo apellido
4	Último grado académico registrado (solo tomando en cuenta licenciatura, maestría o doctorado)
5	Especialidad en la que realizó el último grado de estudio
6	Áreas o líneas de investigación
7	Género (inferido)
8	País de origen (inferido)
9	Estado
10	País de residencia (inferido por la universidad o institución donde labora)
11	Universidad del último grado de estudio
12	Universidad o institución donde trabaja actualmente
13	Página web personal
14	Página web institucional
15	Número de citas totales
16	Índice <i>h</i> (véase la Sección 1.4)
17	Índice i10 (que indica las publicaciones que se han citado al menos 10 veces)
18	Número de artículos publicados en revistas científicas indexadas
19	Enlace a perfil de Google Scholar, ResearchGate u ORCID

**Tabla 2.1:** Información recopilada en la base de datos.

### 2.1.1. Recolección de datos cualitativos

Se decidió iniciar recolectando datos cualitativos, ya que éstos suelen mantenerse más estables a lo largo del tiempo. Es decir, datos como el nombre o apellido, en general, no sufren cambios. Por otra parte, aunque algunos datos categóricos como el “Grado Académico” puede presentar cambios, estos toman años en actualizarse. Además, una vez que alguien obtiene un doctorado, usualmente no se estudia otro grado. La recopilación de datos cualitativos se llevó a cabo del 3 de junio al 30 de septiembre de 2022.

#### 2.1.1.1. México

El registro de los datos cualitativos se inició con las universidades de México. En nuestro país, se recolectaron datos de 18 dependencias, dando un total de 200 perfiles de investigadores laborando en algún instituto mexicano. Estos datos desglosan en la Tabla 2.2 (código R en A.2).

## 2.1 Recopilación de datos

---

	Universidades e institutos	Investigadores(as)
1	Centro de Investigación en Matemáticas	51
2	Centro de Investigación y de Estudios Avanzados	19
3	Colegio de Postgraduados	20
4	Departamento de Matemáticas/Universidad de Guanajuato	2
5	ENES	1
6	Facultad de Ciencias/UNAM	15
7	Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas/UNAM	18
8	Instituto de Matemáticas/UNAM	11
9	Instituto Tecnológico Autónomo de México	18
10	Tecnológico de Monterrey	3
11	Universidad de Sonora	8
12	Universidad Autónoma Chapingo	5
13	Universidad Autónoma de Aguascalientes	10
14	Universidad Autónoma de Guerrero	1
15	Universidad del Mar	1
16	Universidad Iberoamericana	2
17	Universidad Juárez Autónoma de Tabasco	1
18	Universidad Veracruzana	13
19	NA	1
		200

**Tabla 2.2:** Número de investigadores(as) por universidad en México.

El desglose de la metodología para encontrar los datos es el siguiente:

**Centro de Investigación en Matemáticas:** Se visitaron los sitios web [26] y [27] de las sedes del CIMAT (código R en [A.3](#)).

Los investigadores del CIMAT se reparten entre sus sedes de la siguiente manera.

	Sedes	<i>n</i>
1	Aguascalientes	1
2	Ciudad de México	5
3	Guanajuato	27
4	Nuevo León	11
5	Desconocido	7
		51

Posteriormente, se revisó el perfil de los investigadores registrados. Si un investigador tenía como área o línea de investigación un tema relacionado con probabilidad o estadística entonces fue añadido a nuestra base de datos.

Algunos otros perfiles fueron recolectados mediante el buscador Google Scholar. En este caso, para ser añadido a la base de datos debían cumplir con dos requisitos: que tuvieran una dirección de correo electrónico verificado de *cimat.mx* y que en sus palabras clave apareciera que

## 2. METODOLOGÍA

---

realizan investigación en temas relacionados con probabilidad y estadística. Para este caso, debido a que la recolección fue hecha en Google Scholar y que, al momento en que se realizó la búsqueda las páginas del CIMAT estaban siendo actualizadas, no fue posible encontrar en qué sede laboran algunos investigadores.

### CINVESTAV

	Sedes	<i>n</i>
1	Ciudad de México	12
2	Coahuila	1
3	Guanajuato	2
4	Jalisco	3
5	Querétaro	1
		19

Para el CINVESTAV se visitó la página del departamento de Matemáticas del CINVESTAV [28]. De ahí, en la Sección de “Investigadores”, se extrajo la información de aquellos que realizan investigación en campos relacionados, directa o indirectamente, con probabilidad y estadística (en este último caso, consideramos a las personas que trabajan en Ciencias de Datos o Aprendizaje de Máquina).

De la misma forma, se añadieron más perfiles que fueron encontrados en el buscador de Google Scholar que tuvieran una dirección de correo verificada de *cinvestav.mx* y en sus etiquetas áreas relacionadas con probabilidad y estadística.

### Colegio de Postgraduados

	Sedes	<i>n</i>
1	Estado de México	20
		20

Para el Colegio de Postgraduados se extrajo la información de los integrantes del núcleo académico de los posgrados en estadística; es decir, [29] y [30].

La información se completó con los perfiles en Google Scholar de investigadores registrados con una dirección de correo verificada de *colpos.mx* y en su campo de investigación en probabilidad o estadística.

### Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas/UNAM

	Sedes	<i>n</i>
1	Ciudad de México	18
		18

La información fue obtenida en la página del Departamento de Probabilidad y Estadística

## 2.1 Recopilación de datos

---

[31].

### **Instituto de Matemáticas/UNAM**

	Sedes	<i>n</i>
1	Ciudad de México	6
2	Morelos	4
3	Oaxaca	1
		11

La información de los investigadores fue obtenida de la página oficial del Instituto de Matemáticas [32]. De dicha página se extrajo la información de los investigadores en las áreas interesadas, de las distintas sedes, las cuales son Ciudad Universitaria, Unidad Cuernavaca y Unidad Oaxaca.

### **ITAM**

	Sedes	<i>n</i>
1	Ciudad de México	18
		18

La parte del ITAM fue llenada con datos extraídos de [33], junto a investigadores en probabilidad o estadística registrados en Google Scholar con correo institucional *itam.mx*.

### **Universidad de Sonora**

	Sedes	<i>n</i>
1	Sonora	8
		8

Los datos fueron obtenidos del Departamento de Matemáticas enfocado en el área de probabilidad y estadística [34].

### **Universidad Autónoma Chapingo**

	Sedes	<i>n</i>
1	Estado de México	5
		5

Los datos de la Universidad Autónoma Chapingo fueron obtenidos de Google Scholar. Se tomaron los datos de los investigadores con dirección de correo verificada de *chapingo.mx* y que realizan investigación en probabilidad o estadística.

## 2. METODOLOGÍA

---

### **Universidad Autónoma de Aguascalientes**

	Sedes	<i>n</i>
1	Aguascalientes	10
		10

Los datos de los investigadores de la Universidad Autónoma de Aguascalientes fueron obtenidos de la página [35] siguiendo el orden: Centros Académicos, Centro de Ciencias Básicas, Estadística, Docentes.

### **Universidad Iberoamericana**

	Sedes	<i>n</i>
1	Ciudad de México	2
		2

La información se extrajo de la página oficial de Investigación de la Universidad Iberoamericana [36], de ahí se tomaron a los investigadores en las áreas de probabilidad o estadística.

### **Universidad Veracruzana**

	Sedes	<i>n</i>
1	Veracruz	13
		13

La información sobre los investigadores fue obtenida de la página [37] y completada de la lista de investigadores en probabilidad o estadística en Google Scholar registrados con dominio *uv.mx*.

### **Facultad de Ciencias, UNAM**

	Sedes	<i>n</i>
1	Ciudad de México	15
		15

La información sobre los investigadores fue obtenida del directorio de la Facultad de Ciencias, UNAM [38].

### **Otras instituciones**

Los datos de los investigadores de la siguiente lista se obtuvieron vía comunicación personal a través de la Dra. Eslava.

## 2.1 Recopilación de datos

---

	Sedes	<i>n</i>
1	Departamento de Matemáticas/Universidad de Guanajuato	2
2	ENES	1
3	Tecnológico de Monterrey	3
4	Universidad Autónoma de Guerrero	1
5	Universidad del Mar	1
6	Universidad Juárez Autónoma de Tabasco	1
7	Desconocido	1
		10

### 2.1.1.2. Brasil

La recolección de datos categóricos para el país de Brasil fue la más extensa durante la creación de la base de datos. Sin embargo, no fue una recolección difícil debido a que Brasil cuenta con páginas institucionales y personales, de fácil acceso, donde la información se encuentra organizada. La mayoría, por no decir todos los investigadores, cuentan con un currículum llamado “Currículo Lattes” que es creado por el Consejo Nacional de Desarrollo Científico y Tecnológico [39] el cual es un órgano del Ministerio de Ciencia, Tecnología e Innovación de Brasil para promover la investigación.

En el Currículo Lattes, los datos de la Tabla 2.1 están disponibles, aunque en muchas ocasiones el uso del Currículo Lattes no fue necesario, ya que las páginas institucionales nos brindaban tales datos.

Por razones geográficas, el nombre de las universidades e institutos del país no eran tan conocidos al autor como sus equivalentes en México, pero este inconveniente fue solucionado de la siguiente manera: La Dra. Laura Eslava realizó una lista, en el cual venían nombres de algunos investigadores en probabilidad y estadística de Brasil. Entonces se aplicó el siguiente algoritmo de búsqueda:

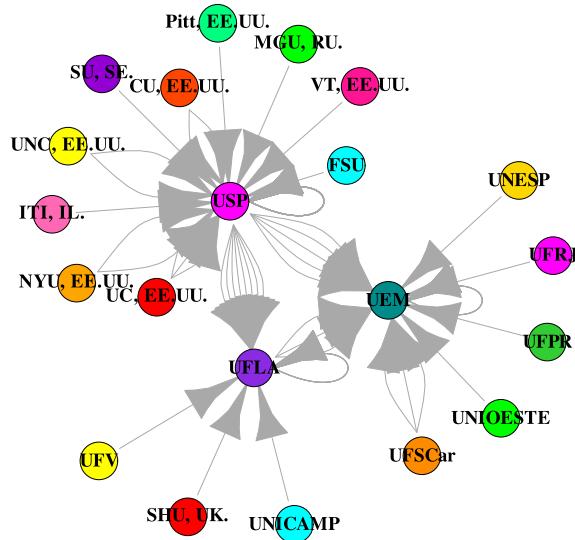
1. Se buscó la información de cada uno de estos investigadores.
2. Si se detectaba que un investigador trabajaba en cierto departamento, dicho departamento era analizado y se extraía la información de los miembros del departamento.
3. Mientras se recopilaba la información, si se detectaba que algún investigador realizó sus estudios de posgrado en una universidad distinta, significaba que esta universidad podría tener un departamento de estadística.
4. Se buscaba al departamento de esta segunda universidad y los datos de los miembros se agregaban a la base de datos.

A continuación exemplificamos la aplicación del algoritmo de búsqueda anterior. El Departamento de Estadística de la Universidad Estatal de Maringá cuenta con dos investigadores y una investigadora que realizaron sus estudios en la Universidad Federal de Lavras. Esto significa

## 2. METODOLOGÍA

---

que la Universidad Federal de Lavras cuenta con un departamento de estadística. Al recolectar los datos de la Universidad Federal de Lavras nos percatamos que, entre sus investigadores e investigadoras, se encuentran ocho personas que realizaron sus estudios en la Universidad de São Paulo. Es decir, que la Universidad de São Paulo también cuenta con un departamento de estadística, así que los miembros de dicho departamento fueron anexados a nuestra base de datos. En la Figura 2.1<sup>1</sup> observamos una representación visual de este hecho<sup>2</sup> (código R en A.7).



**Fig. 2.1:** Red de investigadores(as) de PyE compuesta de tres clústeres que representan distintos departamentos en Brasil y las aristas indican la relación entre la institución donde se realiza el doctorado y el lugar de trabajo.

Lo anterior es un claro ejemplo de cómo las redes sociales fueron útiles en la recolección de datos.

Definimos entonces una red social en la cual los investigadores son los *nodos* y cada *arista* es la vinculación que tienen los miembros por realizar investigación en las áreas de probabilidad y estadística (véase la Sección 1.2). A su vez, mediante la vinculación que un investigador tiene con una universidad por haber realizado sus estudios de posgrado en dicha universidad, nos fue posible encontrar la presencia de clústeres. Es decir, detectar las agrupaciones de investigadores que pertenecen a una misma organización.

---

<sup>1</sup>Se omitieron a los investigadores cuya universidad de estudio es desconocida.

<sup>2</sup>El significado de las siglas se encuentra en la Sección Siglas institucionales.

## 2.1 Recopilación de datos

---

Otra técnica fue la de utilizar el buscador de Google Maps y poner en el recuadro de búsqueda “Departamentos de Estadística en Brasil”. Los resultados de la búsqueda dieron una lista de departamentos y su página institucional.

En total se recolectaron 588 perfiles de investigadores enfocados en probabilidad o estadística en Brasil. El desglose de los datos se muestra en la Tabla 2.3 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en el Apéndice B.1.

	Universidades e institutos	Investigadores(as)
1	Escuela Nacional de Ciencias Estadísticas	34
2	Instituto de Matemática Pura y Aplicada	10
3	Pontificia Universidad Católica de Río de Janeiro	2
4	Universidad de Brasilia	50
5	Universidad de São Paulo	41
6	Universidad Estatal de Campinas	27
7	Universidad Estatal de Feira de Santana	3
8	Universidad Estatal de Londrina	13
9	Universidad Estatal de Maringá	22
10	Universidad Estatal de Río de Janeiro	19
11	Universidad Federal de ABC	2
12	Universidad Federal de Amazonas	22
13	Universidad Federal de Bahía	33
14	Universidad Federal de Ceará	20
15	Universidad Federal de Juiz de Fora	16
16	Universidad Federal de Lavras	16
17	Universidad Federal de Minas Gerais	30
18	Universidad Federal de Paraíba	27
19	Universidad Federal de Paraná	25
20	Universidad Federal de Pernambuco	26
21	Universidad Federal de Río Grande del Sur	11
22	Universidad Federal de Santa Catarina	12
23	Universidad Federal de São Carlos	26
24	Universidad Federal de São Joao del Rei	14
25	Universidad Federal de Viçosa	16
26	Universidad Federal del Rio de Janeiro	28
27	Universidad Federal Fluminense	28
28	Universidad Federal Rural de Pernambuco	15
		588

**Tabla 2.3:** Número de investigadores(as) por universidad en Brasil.

Los criterios para recolectar la información de los docentes fueron similares a los de México: se extrajeron a todos los miembros de los departamentos especializados en estadística o probabilidad, si los departamentos también eran de matemáticas u otras áreas, se recopilaban solo a los investigadores que se enfocaran en las áreas de probabilidad o estadística. Es importante aclarar que solo se guardó como institución a la entidad mayor, que contiene al departamento o instituto donde trabajan.

## 2. METODOLOGÍA

---

### 2.1.1.3. Argentina, Chile, Colombia, Cuba, Perú y Uruguay

En la recolección de datos cualitativos también incluyó a los siguientes países: Argentina, Chile, Colombia, Cuba, Perú y Uruguay.

La información disponible sobre los institutos y universidades en estos países resultó ser más difícil de acceder, a diferencia de la información para México y Brasil. También, en este caso, sólo se guardó como institución a la entidad mayor que contiene al departamento o instituto donde trabajan.

El desglose de los países y sus universidades aparece en la Tabla 2.4 (código R en A.4).

	Universidades e institutos	Investigadores(as)
1	Argentina	147
2	Chile	125
3	Colombia	54
4	Cuba	2
5	Perú	49
6	Uruguay	69
		446

**Tabla 2.4:** Número de investigadores(as) por país.

#### Argentina

El desglose de las universidades en Argentina se muestra en la Tabla 2.5 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en el Apéndice B.2.

	Universidades e institutos	Investigadores(as)
1	Universidad de Buenos Aires	5
2	Universidad Nacional de Córdoba	13
3	Universidad Nacional de La Plata	17
4	Universidad Nacional de Rosario	44
5	Universidad Nacional de Tres de Febrero	22
6	Universidad Nacional de Tucumán	34
7	Universidad Nacional del Comahue	12
		147

**Tabla 2.5:** Número de investigadores(as) por universidad en Argentina.

#### Chile

El desglose de las universidades en Chile se muestra en la Tabla 2.6 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en B.3.

## 2.1 Recopilación de datos

---

	Universidades e institutos	Investigadores(as)
1	Pontificia Universidad Católica de Chile	19
2	Pontificia Universidad Católica de Valparaíso	20
3	Universidad Adolfo Ibáñez	1
4	Universidad Católica de Temuco	1
5	Universidad Católica del Norte	1
6	Universidad de Antofagasta	9
7	Universidad de Atacama	1
8	Universidad de Chile	1
9	Universidad de Concepción	10
10	Universidad de La Frontera	39
11	Universidad de Talca	3
12	Universidad de Tarapacá	3
13	Universidad del Bío-Bío	11
14	Universidad Técnica Federico Santa María	4
15	Universidad Tecnológica Metropolitana	2
		125

**Tabla 2.6:** Número de investigadores(as) por universidad en Chile.

### Colombia

El desglose de las universidades en Colombia se muestra en la Tabla 2.7 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en el Apéndice B.4.

	Universidades e institutos	Investigadores(as)
1	Pontificia Universidad Javeriana	6
2	Universidad de los Andes	3
3	Universidad del Norte Barranquilla	12
4	Universidad Nacional de Colombia	32
5	Desconocido	1
		54

**Tabla 2.7:** Número de investigadores(as) por universidad en Colombia.

### Cuba

El desglose de las universidades en Cuba se muestra en la Tabla 2.8 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en el Apéndice B.5.

	Universidades e Institutos	Investigadores(as)
1	Universidad de Oriente Cuba	2
		2

**Tabla 2.8:** Número de investigadores(as) por universidad en Cuba.

## 2. METODOLOGÍA

---

### Perú

El desglose de las universidades en Perú se muestra en la Tabla 2.9 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en el Apéndice B.6.

	Universidades e institutos	Investigadores(as)
1	Universidad Nacional Mayor de San Marcos	37
2	Universidad Nacional San Agustín	12
		49

**Tabla 2.9:** Número de investigadores(as) por universidad en Perú.

### Uruguay

El desglose de las universidades en Uruguay se muestra en la Tabla 2.10 (código R en A.2). Le sugerimos al lector consultar las fuentes de información utilizadas en el Apéndice B.7.

	Universidades e institutos	Investigadores(as)
1	Media Maren	5
2	Universidad de la República	64
		69

**Tabla 2.10:** Número de investigadores(as) por universidad en Uruguay.

### 2.1.2. Recolección de datos cuantitativos

Aunque la recopilación de datos cualitativos incluyó a más países, la recolección de datos cuantitativos como el número de citas, índice  $h$ , índice  $i10$  y número de artículos se centró en México y Brasil, donde la información es más amplia.

Esta recopilación de datos se llevó a cabo del 01 de octubre al 4 de noviembre de 2022. La recolección de datos cuantitativos tuvo como fuente principal a la plataforma “Google Scholar”, seguida por “ResearchGate”.

El mayor obstáculo al momento de recabar los datos numéricos fue encontrar los perfiles de los investigadores registrados en nuestra base de datos. Esto se debió a que la mayoría de los investigadores en Latinoamérica siguen la línea establecida por los investigadores en países angloparlantes [40], utilizando solo un nombre y un apellido para identificarse, e incluso en algunos casos, solo iniciales. Esta falta de nombres completos o información detallada dificultaba la identificación precisa de los investigadores en la base de datos. Como consecuencia, fue necesario inferir cuáles eran los nombres, apellidos o iniciales que utilizaban en sus perfiles. Para abordar este inconveniente, se implementó la siguiente metodología:

- Método 1: Búsqueda por institución en Google Scholar (en lugar de realizar una búsqueda de cada investigador). En la parte de “Perfiles” se buscaba a la institución completa, ya sea por su nombre o el dominio de sus correos electrónicos. De la lista de resultados se recabaron los datos de los investigadores que aparecían tanto en nuestra base de datos como en los resultados de la búsqueda. Por ejemplo, para el llenado de los datos numéricos de

## 2.1 Recopilación de datos

---

los investigadores del IIMAS, se buscó a los perfiles que decían laborar en el “Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas” y a los perfiles verificados con una dirección de correo electrónico verificado de *sigma.iimas.unam.mx*.

- Método 2: Búsqueda por coautores o citas. Una vez encontrado el perfil de algún investigador, trabajando en un determinado departamento, se revisaban a sus coautores y a los investigadores que había citado que, generalmente, resultaban ser de la misma institución.
- Método 3: Búsqueda por nombre en Google Scholar. Si alguno de los perfiles de los investigadores no fueron encontrados por ninguno de los dos métodos anteriores, se realizaba una búsqueda por su nombre completo. Aunque, en la mayoría de las veces esta búsqueda resultaba ser infructífera; existieron casos en los que los investigadores tenían todos sus nombres y apellidos. Al encontrar a un investigador mediante este método, se revisaba a sus coautores y citas aplicando el Método 2.
- Método 4: Búsqueda por nombre en ResearchGate. Si ninguno de los tres métodos anteriores funcionaba para encontrar a ciertos perfiles, se buscaba por nombre completo en ResearchGate. La elección de ResearchGate se debe a que el algoritmo de búsqueda de esta plataforma encuentra a los perfiles de manera más efectiva, incluso si los investigadores no utilizan su nombre completo. Si alguno de los investigadores era encontrado, y entonces conocíamos su nombre como autor, se regresaba a Google Scholar a buscarlo con dicho nombre. Si era encontrado entonces se aplicaba nuevamente el Método 2. De lo contrario, se extraían los datos encontrados en ResearchGate.

Cabe señalar que, en los Métodos 1 y 2, el principal objetivo era detectar la presencia de clústeres. Para el Método 1 la intención fue encontrar los agrupamientos de investigadores que pertenecen a una misma institución o departamento. En el Método 2 se encontraban los agrupamientos que se forman al citar. Es necesario mencionar que en una red de citas las aristas están dirigidas hacia atrás en el tiempo. Es decir, desde los artículos más recientes hasta los más antiguos [7]. Esto representa un inconveniente, ya que hace más difícil encontrar a investigadores que recién empiezan a publicar.

En general, estos fueron los métodos utilizados para encontrar los datos que necesitábamos. Si ninguno de estos cuatro métodos funcionaba, descartamos la búsqueda del investigador en cuestión y dejábamos a dicho investigador con datos vacíos.

En México y Brasil, se contaba con un total de 788 perfiles. De este conjunto, se recolectaron datos numéricos para 619 perfiles, lo que equivale aproximadamente al 78 % del total. La repartición de estos datos se pueden ver en las Tablas 2.11 y 2.12 (código R en A.5).

## 2. METODOLOGÍA

---

### México

	Universidades e Institutos	Investigadores(as)
1	CIMAT	47
2	CINVESTAV	19
3	Colegio de Postgraduados	16
4	DEMAT/Universidad de Guanajuato	2
5	ENES	1
6	FC/UNAM	15
7	IIMAS/UNAM	17
8	IMATE/UNAM	9
9	ITAM	16
10	Tecnológico de Monterrey	3
11	UNISON	8
12	Universidad Autónoma Chapingo	4
13	Universidad Autónoma de Aguascalientes	9
14	Universidad Autónoma de Guerrero	1
15	Universidad del Mar	1
16	Universidad Iberoamericana	2
17	Universidad Juárez Autónoma de Tabasco	1
18	Universidad Veracruzana	13
19	Desconocido	1
		185

**Tabla 2.11:** Número de investigadores con datos numéricos conocidos por universidad en México.

## 2.1 Recopilación de datos

---

### Brasil

	Universidades e Institutos	Investigadores(as)
1	Escuela Nacional de Ciencias Estadísticas	21
2	Instituto de Matemática Pura y Aplicada	8
3	Universidad de Brasilia	32
4	Universidad de São Paulo	36
5	Universidad Estatal de Campinas	24
6	Universidad Estatal de Feira de Santana	2
7	Universidad Estatal de Londrina	11
8	Universidad Estatal de Maringá	19
9	Universidad Estatal de Río de Janeiro	15
10	Universidad Federal de ABC	2
11	Universidad Federal de Amazonas	9
12	Universidad Federal de Bahía	22
13	Universidad Federal de Ceará	17
14	Universidad Federal de Juiz de Fora	10
15	Universidad Federal de Lavras	14
16	Universidad Federal de Minas Gerais	28
17	Universidad Federal de Paraíba	15
18	Universidad Federal de Paraná	20
19	Universidad Federal de Pernambuco	21
20	Universidad Federal de Río Grande del Sur	9
21	Universidad Federal de Santa Catarina	8
22	Universidad Federal de São Carlos	20
23	Universidad Federal de São Joao del Rei	10
24	Universidad Federal de Viçosa	15
25	Universidad Federal del Rio de Janeiro	18
26	Universidad Federal Fluminense	16
27	Universidad Federal Rural de Pernambuco	12
		434

**Tabla 2.12:** Número de investigadores con datos numéricos conocidos por universidad en Brasil.



## Análisis y representación de los datos

---

En este capítulo, se presentan los análisis y representaciones realizados durante la investigación.

En primer lugar, se describen los métodos, pruebas estadísticas y gráficos utilizados para el procesamiento y análisis de los datos recolectados. Posteriormente, se presenta el análisis y la representación de nuestros datos, con el fin de investigar la brecha de género en la investigación académica en las áreas de probabilidad y estadística. Se exploran diferentes aspectos, que consideran la distribución de grados académicos entre investigadores e investigadoras, la proporción de género en los departamentos y las posibles disparidades en el número de citas y publicaciones entre hombres y mujeres.

### 3.1. Descripción de métodos y pruebas estadísticas

- **3.1.1. Bootstrap**

El bootstrap es un método estadístico de remuestreo no paramétrico que se utiliza para evaluar la precisión de los estadísticos de interés, como la media o la mediana, a partir de una única muestra de datos. El método se basa en la idea de que una muestra de datos es una estimación ruidosa de la distribución de probabilidad subyacente de la población. El bootstrap aborda esta incertidumbre al generar múltiples muestras de la muestra original mediante remuestreo con reemplazo, lo que permite construir una distribución empírica de los estadísticos de interés [41]. Fue introducido por Efron en 1979 [42] y Efron y Tibshirani en 1993 [43]. Véase [44] para una referencia en español.

- **3.1.2. Gráfico Q-Q**

El gráfico Q-Q, también conocido como gráfico de probabilidad normal o Q-Q Plot Normal, es una técnica gráfica utilizada para probar la normalidad de un conjunto de datos. Consiste en comparar la distribución empírica de los datos con la distribución normal a través de los cuantiles de la normal estándar. Si los puntos en el gráfico se aproximan a una línea recta, se puede suponer que los datos tienen una distribución normal [45].

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

#### ■ 3.1.3. La prueba de Shapiro

La prueba de Shapiro es un método utilizado para determinar si un conjunto de datos sigue una distribución normal. Este test se utiliza principalmente cuando el tamaño de la muestra es menor a 50 observaciones. Este test se basa en el gráfico de probabilidad normal, que compara la distribución empírica de los datos con la distribución normal esperada. La línea diagonal del gráfico es la línea de ajuste perfecto y cualquier desviación de esta línea indica una desviación de la normalidad. Aunque la prueba de Shapiro se recomienda para muestras pequeñas, Royston sugirió una extensión que permite su aplicación en muestras grandes y que se implementa en algunos programas software especializados estadísticos [46] (véase la documentación en [47]).

#### ■ 3.1.4. La prueba de rango con signo de Wilcoxon de una muestra

Fue introducido por Frank Wilcoxon en 1945 [48]. Existen dos versiones de esta prueba: la prueba de suma de rangos de Wilcoxon, que se utiliza para comparar dos muestras independientes, y la prueba de rango con signo de Wilcoxon, que se emplea para comparar dos muestras relacionadas o pareadas. La prueba de rango con signo de Wilcoxon de una muestra es un caso particular de esta última, la cual es una prueba no paramétrica utilizada para comparar una muestra con un valor específico. Es decir, se utiliza para determinar si un grupo es significativamente diferente de un valor conocido o hipotético en la variable de interés [49]. En esta tesis para realizar la prueba se utiliza la función *gghistostats* del paquete “*ggstatsplot*” (véase la documentación en [50]).

#### ■ 3.1.5. La prueba W-Mann-Whitney

Fue introducido por Mann y Whitney en 1947 [51]. La prueba W-Mann-Whitney se utiliza cuando no se supone normalidad ni ningún otro supuesto sobre la varianza de las variables aleatorias. También conocida como la prueba de suma de rangos de Wilcoxon, la prueba W-Mann-Whitney es la alternativa no paramétrica a la prueba *t* de Student para dos grupos independientes. Mientras que la prueba *t* de Student compara las medias, la prueba W-Mann-Whitney compara las diferencias entre dos medianas. Por lo cual, la prueba W-Mann-Whitney se basa en rangos [52]. En esta tesis se utiliza la función *ggbetweenstats* del paquete “*ggstatsplot*” (véase la documentación en [53]) para comparar las medianas de dos conjuntos con distribución no normal. Aunque es una prueba ampliamente utilizada, es importante considerar que su uso es discutido en algunos ámbitos [54].

#### ■ 3.1.6. La función de distribución acumulada

La función de distribución acumulada de una variable aleatoria es una función matemática que evalúa la probabilidad de que la variable aleatoria sea menor o igual a un determinado valor. Es decir, la función de acumulación de probabilidad o función de probabilidad acumulada de la variable aleatoria  $X$  evaluada en un número  $x$  cualquiera es  $\mathbb{P}(X \leq x)$ . En el caso discreto, se calcula sumando todos los valores positivos de la función de probabilidad evaluada en aquellos números menores o iguales a  $x$ . Esto se expresa matemáticamente como:

$$F(x) = \sum_{u \leq x} p(u), \quad (3.1)$$

donde  $p$  es la función de masa de probabilidad. En el caso absolutamente continuo, se calcula integrando la función de densidad de probabilidad en el intervalo  $(-\infty, x]$ , es decir

$$f(x) = \int_{-\infty}^x f(u)du. \quad (3.2)$$

Ver [55][Capítulo 2].

- **3.1.7. K-means**

El método de clustering K-means fue introducido por MacQueen en 1967 [56]. Es uno de los más populares y utilizados debido a su eficacia y sencillez de implementación. Este algoritmo se encarga de clasificar un conjunto de objetos en un número predeterminado de clústeres,  $K$ , siguiendo un procedimiento sencillo. Su nombre K-means es debido a que cada clúster se representa por su centroide, que es la media (o media ponderada) de sus puntos. Esta representación mediante centroides tiene la ventaja de tener una interpretación gráfica y estadística clara. Cada clúster se caracteriza por su centro o centroide, el cual se encuentra en el medio o centro de los elementos que lo componen [57]. En esta tesis los clústeres se realizan utilizando la función  $kmeans()$  que se encuentra dentro del paquete base de R (véase la documentación en [58]).

- **3.1.8. Coeficiente de correlación de Spearman**

El coeficiente de correlación de Spearman fue introducido por el psicólogo Charles Spearman en 1904 [59]. Se considera la alternativa no paramétrica al comúnmente utilizado coeficiente de Pearson [60]. Mientras que la correlación de Pearson toma en cuenta relaciones lineales de variables continuas con distribución normal, la correlación de Spearman considera relaciones dadas por funciones monótonas de variables ordinales organizadas en rangos o jerarquías (las funciones lineales son, en particular, monótonas) [61].

La correlación de Spearman es una medida estadística que permite conocer el grado de asociación entre dos variables. A través del cálculo del coeficiente de correlación de Spearman, se determina la fuerza de asociación y dirección que toma esta relación, el cual puede variar en el intervalo  $[-1, +1]$ . Entre más cercano a 1 sea el coeficiente de correlación de Spearman, mayor la fuerza de asociación. Cuando una relación es escasa o no existe, el coeficiente tiende a cero. Es importante tener en cuenta que la correlación de Spearman solo mide la tendencia de dos variables a ir juntas, a lo que también se denomina covarianza; y no implica necesariamente una relación de causa y efecto [61]. Por otro lado, la covarianza es una medida que proporciona la dirección de la relación lineal entre dos variables. No está acotada ni estándarizada, lo que significa que sus valores pueden variar ampliamente. En contraste, la correlación de Spearman incluye tanto la dirección como la intensidad de la relación entre las variables. Los valores de correlación se encuentran en el rango de  $[-1, 1]$ , independientemente de la escala de las variables. Además, la correlación está estandarizada, lo que facilita la comparación entre diferentes pares de variables. Dentro de la tesis, para

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

realizar la correlación de Spearman se utiliza la función “pais.panels” del paquete “psych” (véase la documentación en [62]).

## 3.2. Grado académico

Nuestra base de datos cuenta con 21 columnas y 1234 filas, en cada columna se encuentra información, descrita en la Tabla 2.1, de 1234 investigadores.

De momento trabajaremos con la columna “Grado”, esta variable es de tipo “Factor” y cuenta con 3 niveles: Licenciatura, Maestría y Doctorado. El recuento de grados académicos en nuestra base de datos se encuentra en la Tabla 3.1, de esta tabla obtenemos la Figura 3.1 (código R en A.13).

	Grado	Cantidad
1	Doctorado	866
2	Maestría	148
3	Licenciatura	58
4	Desconocido	162

Tabla 3.1: Número de investigadores según su grado de estudios

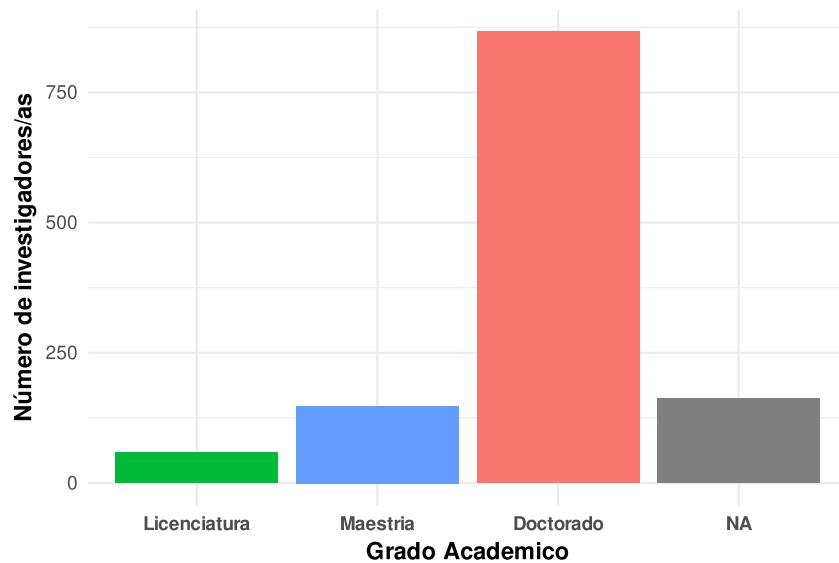


Fig. 3.1: Gráfica del recuento de personas cuyo máximo grado es el indicado, donde NA significa desconocido.

Para los siguientes análisis no tomaremos en cuenta los datos desconocidos en la variable “Grado”.

### 3.2 Grado académico

---

Calculamos las proporciones del grado académico, sin considerar los datos desconocidos, observamos que alrededor del 80.78 % de los investigadores tienen el grado de doctorado, el 13.80 % su máximo grado es de maestría y el 5.410 % solo tienen el grado de licenciatura.

Decir que el 80 % de los investigadores cuentan con un doctorado es una estimación razonable del porcentaje real de investigadores que son doctores, pero dado que nuestra base de datos no representa a la población total de investigadores, no aseguramos que el 80 % de investigadores tengan dicho grado. Para capturar la incertidumbre en nuestra estimación creamos un intervalo de confianza. Para consultar un ejemplo similar, le sugerimos al lector visitar la Sección “The General Social Survey” del curso “Inference for Categorical Data in R” en DataCamp [63].

Definimos a la proporción de doctores en nuestra base de datos como  $\hat{p} = 0.807$

Para encontrar el error estándar utilizamos la técnica de bootstrap (véase la Subsección 3.1.1 y el código R en A.6). En esta técnica se realiza lo siguiente:

1. Nos centramos en una variable específica, en nuestro caso “Grado”.
2. Extraemos una muestra con reemplazo del mismo tamaño de nuestro conjunto de datos original, esto con el fin de recrear la variación aleatoria que aparece cuando se extrae una muestra de una población.
3. Repetimos esto varias veces para crear conjuntos de datos replicados, en este caso 500 repeticiones.
4. A cada réplica le calculamos un estadístico muestral, en este caso la proporción de investigadores con doctorado.

De esta manera obtenemos una recopilación de estadísticas de remuestreo de nuestro conjunto de datos. Es decir, obtenemos un data frame con la proporción de doctores que tuvo cada conjunto generado. En la Tabla 3.2 vemos las primeras 10 iteraciones (código R en A.6) y creamos el gráfico de densidad de estas estadísticas muestrales que se muestra en la Figura 3.2<sup>1</sup>.

replica	1	2	3	4	5	6	7	8	9	10
stat	0.81	0.82	0.81	0.79	0.80	0.81	0.83	0.82	0.81	0.80

**Tabla 3.2:** Proporciones generadas en las primeras 10 réplicas de bootstrap.

La desviación estándar de esta distribución es una buena estimación del error estándar, por lo que

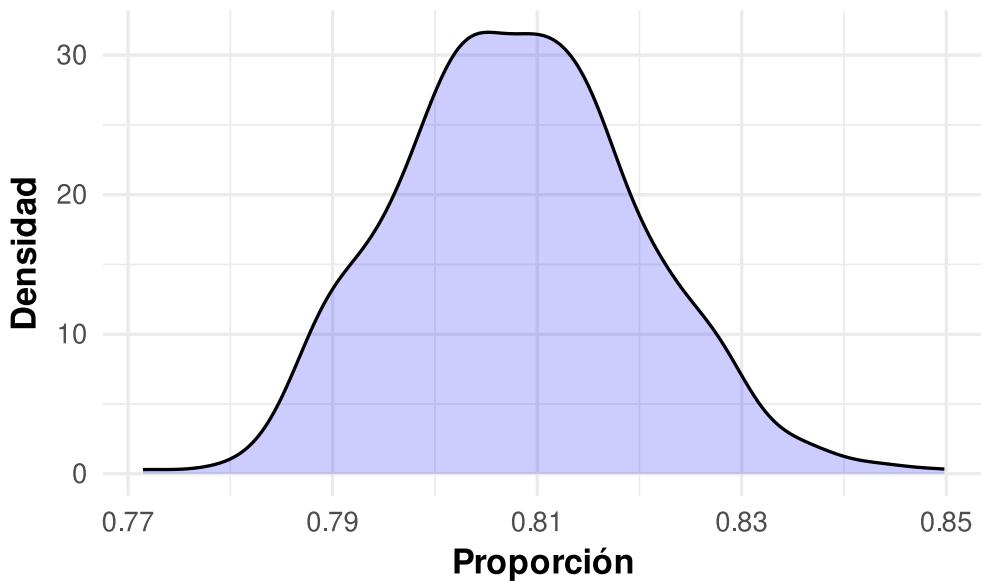
$$\text{Error estándar} = 0.011$$

Calculamos nuestro intervalo de confianza

$$(\hat{p} - 2 \times SE, \hat{p} + 2 \times SE)$$
$$(0.807 - 2 \cdot (0.012), \quad 0.807 + 2 \cdot (0.012)).$$

---

<sup>1</sup>Es importante aclarar que solo tenemos los datos disponibles en Internet en la fecha de la búsqueda.



**Fig. 3.2:** Gráfica de densidad de las 500 réplicas del bootstrap.

Concluimos que el intervalo de confianza es  $(0.784, 0.831)$ .

Entonces estimamos que entre el 78 % y 83 % de investigadores, que trabajan en probabilidad y estadística en los países estudiados, cuentan con doctorado.

De manera análoga se utilizó el método bootstrap para hacer estimaciones sobre los investigadores con grado académico máximo de maestría y licenciatura. Para el caso de los investigadores registrados con grado académico maestría se tiene un intervalo de confianza de  $(0.116, 0.160)$  y para los de grado académico licenciatura un intervalo de confianza de  $(0.041, 0.067)$ .

Este análisis también se realizó para México y Brasil como conjunto. Así mismo para México y Brasil, individualmente. Los resultados se muestran en la Tabla 3.3.

Grado académico	Intervalo de confianza		
	México y Brasil	México	Brasil
Doctorado	$(0.888, 0.931)$	$(0.870, 0.957)$	$(0.884, 0.931)$
Maestría	$(0.065, 0.104)$	$(0.033, 0.116)$	$(0.063, 0.112)$
Licenciatura	$(0, 0.010)$	$(0, 0.027)$	$(0, 0.008)$

**Tabla 3.3:** Intervalos de confianza por grado académico.

### 3.2.1. Dependencia del grado doctorado con el género (inferido)

Como el 80 % de los investigadores cuentan con doctorado, nos enfocaremos en este grado para conocer si existe alguna relación entre ser un investigador con doctorado y el género<sup>2</sup>, por lo que nuestra variable grado se convertirá en la variable “Doctorado” con dos niveles: “Doctorado” y “NoDoctorado”, en donde se agrupan los investigadores que pertenecían a los niveles “Maestría” o “Licenciatura” de la variable “Grado”. Para consultar un ejemplo similar, le sugerimos al lector visitar la Sección “Intervals for differences” del curso “Inference for Categorical Data in R” en DataCamp [64].

Para averiguar si la variable “Doctorado” y “Género” son independientes, formularemos la hipótesis nula de que la diferencia entre proporciones de hombres y mujeres con doctorado es cero. Es decir,

$$H_0 : P_{Hombres} - P_{Mujeres} = 0;$$

donde  $P_{Hombres}$  denota la proporción de hombres y  $P_{Mujeres}$  de mujeres.

La hipótesis alternativa sería que la proporción es distinta de cero. Comenzaremos calculando la proporción de investigadores con doctorado por género (código R en A.15).

Proporción de hombres con doctorado

$$\frac{\text{Hombres con doctorado}}{\text{Total de hombres}} = \frac{611}{711} = 0.859$$

Proporción de mujeres con doctorado

$$\frac{\text{Mujeres con doctorado}}{\text{Total de mujeres}} = \frac{255}{361} = 0.706$$

Calculamos la diferencia entre ellas.

$$P_{diferencia} = 0.152$$

Para saber si esta estadística es consistente con la hipótesis nula de que la variable “Doctorado” es independiente de la variable “Género”, generaremos conjuntos de datos mediante el uso del método bootstrap (véase la Subsección 3.1.1 y el código R en A.15), tomando a la hipótesis nula como cierta. Para la generación de datos tomaremos a la variable “Género” como una variable explicativa, ya que buscamos explicar el doctorado en función del género. Como tomamos la independencia de las variables como cierta, fijamos la variable “Género” y permutamos los datos de la variable “Doctorado”.

En la Tabla 3.4 se muestran los primeros 10 resultados de las primeras dos permutaciones. Los datos presentados en la Tabla 3.4 son similares al original, pero generados como si fueran de un mundo donde no existe asociación entre la variable “Doctorado” y “Género”.

Generamos 500 conjuntos de datos y calculamos la diferencia de proporción entre el género

---

<sup>2</sup>Este estudio es realizado para los datos disponibles en internet.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

masculino y femenino para cada uno. En la Tabla 3.5 la variable “stat” muestra la diferencia de proporciones en cada réplica. A partir de estos datos, construimos nuestra distribución nula. Además, generamos un intervalo de confianza del 95 %, cuyos límites, calculados a través de los percentiles 2.5 y 97.5, se sitúan en (-0.047, 0.044).

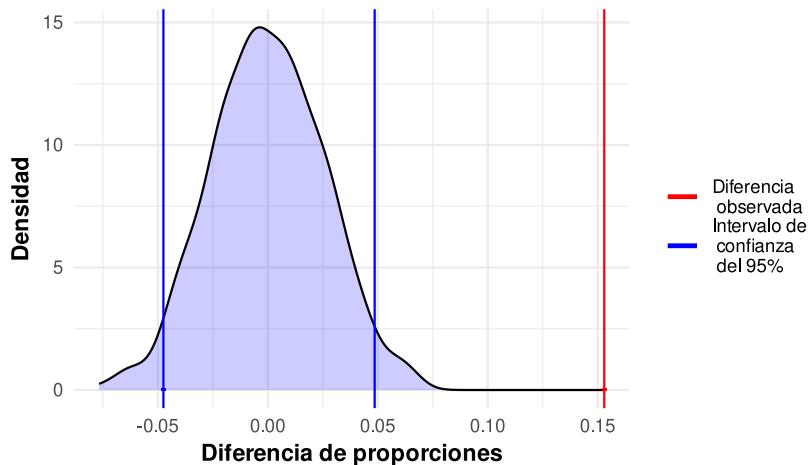
a Réplica 1			b Réplica 2				
	Doctorado	Género	replica		Doctorado	Género	replica
1	Doctorado	Masculino	1	1	Doctorado	Masculino	2
2	Doctorado	Femenino	1	2	NoDoctorado	Femenino	2
3	Doctorado	Masculino	1	3	Doctorado	Masculino	2
4	Doctorado	Masculino	1	4	Doctorado	Masculino	2
5	Doctorado	Masculino	1	5	NoDoctorado	Masculino	2
6	Doctorado	Masculino	1	6	Doctorado	Masculino	2
7	NoDoctorado	Femenino	1	7	Doctorado	Femenino	2
8	Doctorado	Masculino	1	8	Doctorado	Masculino	2
9	Doctorado	Masculino	1	9	NoDoctorado	Masculino	2
10	NoDoctorado	Masculino	1	10	NoDoctorado	Masculino	2

**Tabla 3.4:** Resultados de las dos primeras replicas.

replica	1	2	3	4	5	6	7	8	9	10
stat	-0.01	-0.04	0.01	0.05	0.00	-0.01	0.00	-0.02	0.01	-0.01

**Tabla 3.5:** Primeros 10 resultados de los 500 conjuntos generados.

Cuando graficamos nuestra estadística observada, es decir,  $P_{diferencia} = 0.152$  contra la distribución nula en la Figura 3.3, vemos que la diferencia observada es mayor que el tipo de diferencias que veríamos si las variables fueran independientes. Y queda fuera del intervalo de confianza del 95 % (-0.047, 0.044). Por lo que estos datos constituyen evidencia de que la variable “Doctorado” no es independiente de la variable “Género”.



**Fig. 3.3:** Densidad nula cuando “Doctorado” es independiente de “Género”.

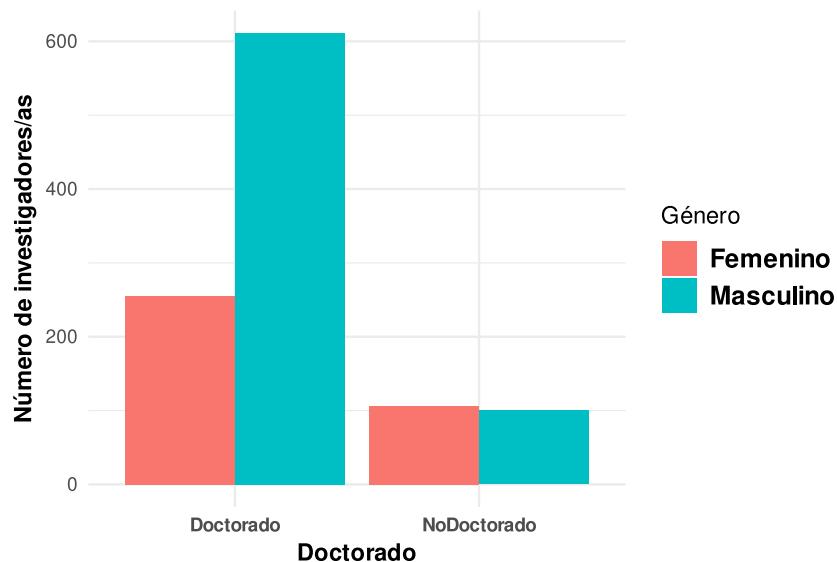
### 3.2 Grado académico

---

Visualizamos la disparidad entre géneros en la Tabla cruzada 3.6, de esta tabla obtenemos la Figura 3.4 (códigos R en [A.14](#) y [A.17](#)).

	Variable Doctorado		
Variable Género	Doctorado	NoDoctorado	Fila Total
<b>Femenino (N)</b>	255	106	361
N/total de la fila	0.706	0.294	0.337
N/total de la columna	0.294	0.515	
N/total de la tabla	0.238	0.099	
<b>Masculino (N)</b>	611	100	711
N/total de la fila	0.859	0.141	0.663
N/total de la columna	0.706	0.485	
N/total de la tabla	0.570	0.093	
Columna Total	866	206	1072
Proporción	0.808	0.192	

**Tabla 3.6:** Comparación de género entre los niveles de grado “Doctorado” y “NoDoctorado”.



**Fig. 3.4:** Investigadores con doctorado y no doctorados por género.

Como observamos, el número de investigadores del género masculino es más del doble que el de sus homólogas femeninas.

En la Tabla 3.6, al observar la relación entre la variable “Doctorado” y la variable “Género”, notamos que las 255 doctoras solo representan el 29.5 % de los investigadores con doctorado, mientras que los hombres representan el 70.5 %. En el caso de los investigadores sin el grado de doctorado, las mujeres tienen una ligera ventaja, representando el 50.5 % de los casos.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

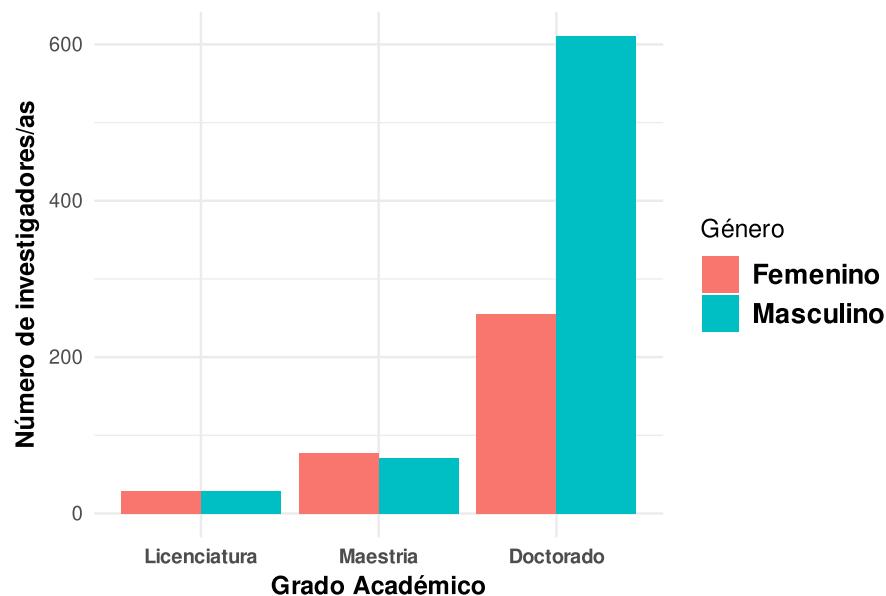
---

#### 3.3. Análisis de género en departamentos

Profundizamos en este análisis al realizar la Tabla cruzada 3.7 con todos los grados, de esta tabla obtenemos la Figura 3.5 (códigos R en A.14 y A.17).

	Variable Grado			
Variable Género	Licenciatura	Maestría	Doctorado	Fila Total
<b>Femenino (N)</b>	29	77	255	361
N/total de la fila	0.080	0.213	0.706	0.337
N/total de la columna	0.500	0.520	0.294	
N/total de la tabla	0.027	0.072	0.238	
<b>Masculino (N)</b>	29	71	611	711
N/total de la fila	0.041	0.100	0.859	0.663
N/total de la columna	0.500	0.480	0.706	
N/total de la tabla	0.027	0.066	0.570	
Columna Total	58	148	866	1072
Proporción	0.054	0.138	0.808	

**Tabla 3.7:** Comparación de género en los diferentes grados académicos: licenciatura, maestría y doctorado.



**Fig. 3.5:** Grados académicos por género.

Como notamos no existe disparidad entre los investigadores con grado licenciatura, para el caso de la maestría las mujeres superan en 6 unidades a los hombres. En total, de los 1072 datos analizados, 361 corresponden a mujeres y 711 a hombres, representando el 33.7% y 66.3% respectivamente del total.

### 3.3 Análisis de género en departamentos

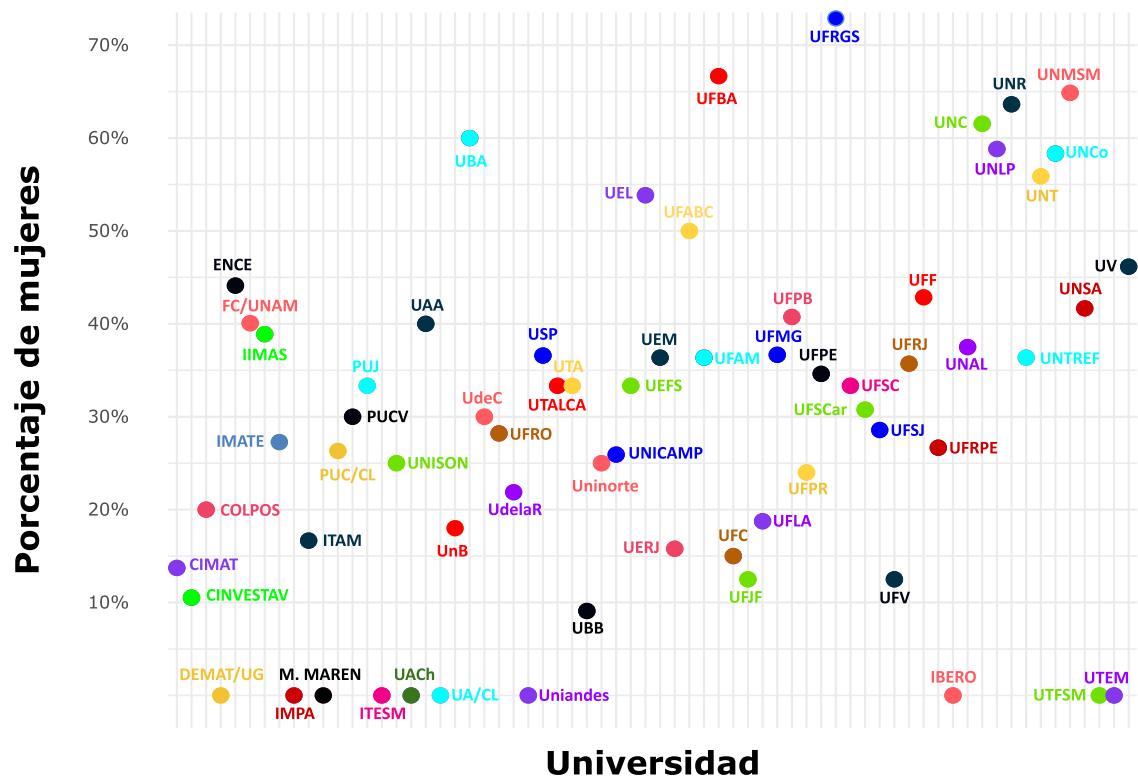
Esto es una buena aproximación, pero recordemos que solo hemos estado trabajando sobre los datos de investigadores cuyo grado académico conocemos. Para realizar un estudio más certero en relación con el número de investigadores de género femenino y masculino utilizaremos los 1234 datos que representan el total de nuestra muestra. De modo que las proporciones de mujeres y hombres son, más precisamente, 33.4 % y 66.6 % respectivamente.

Femenino	Masculino
412	822

**Tabla 3.8:** Cantidad de investigadores e investigadoras dentro de la base de datos.

### **3.3.1. Diferencias entre la cantidad de hombres y mujeres en departamentos**

Para realizar un análisis más detallado calcularemos el porcentaje de hombres y mujeres dentro de cada departamento registrado en nuestra base de datos. En este caso solo utilizaremos los departamentos en los que al menos tengamos los datos de dos investigadores. Los resultados se muestran en la Figura 3.6<sup>3</sup>.



**Fig. 3.6:** Porcentaje de mujeres por departamento.

Uno de los aspectos más notables es la ausencia de mujeres en algunos departamentos, ya que están constituidos únicamente por hombres, mientras que para las mujeres el porcentaje

<sup>3</sup>El significado de las siglas se encuentra en la Sección Siglas institucionales.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

máximo es del 73 %. Para conocer los porcentajes de hombres y mujeres dentro de cada universidad, le sugerimos al lector consultar el Apéndice C.1.

En la Figura 3.7 mostramos la manera en como se distribuyen los datos (código R en A.18).

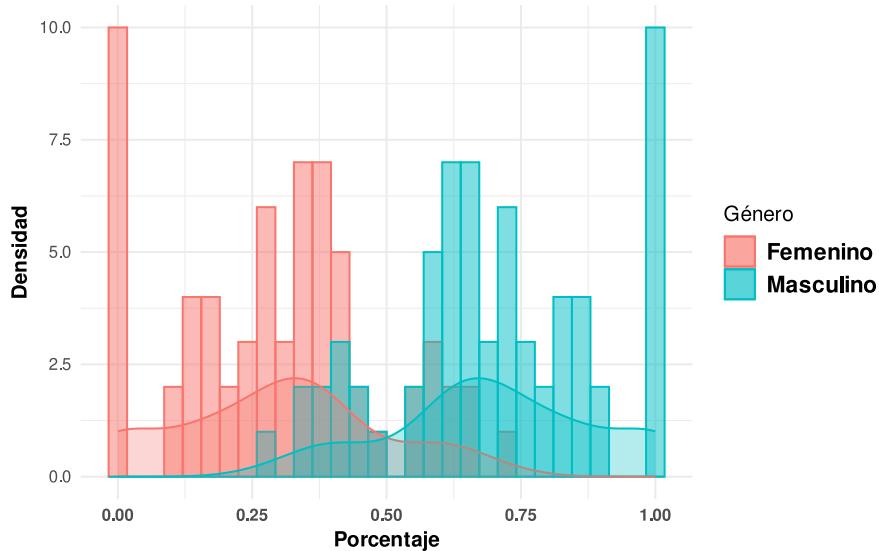


Fig. 3.7: Distribución de hombres y mujeres en los departamentos.

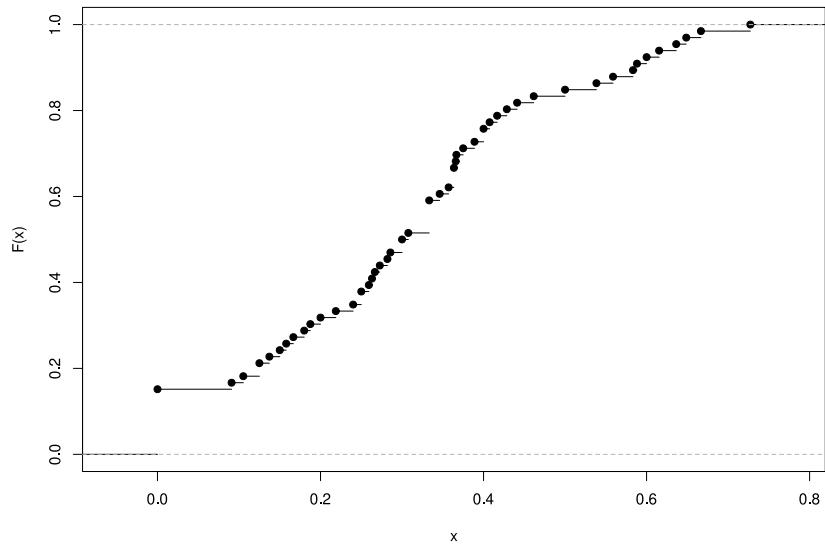
Para determinar la probabilidad de que un departamento seleccionado uniformemente al azar tenga como máximo una determinada proporción de mujeres entre sus investigadores, se utiliza la función de distribución acumulada. Esta función permite una representación clara de la información, indicando el porcentaje de instituciones que tienen un porcentaje máximo específico de mujeres investigadoras. Es decir  $\mathbb{P}(X \leq x)$ , donde  $X$  representa la proporción de instituciones y  $x$  el porcentaje máximo. En la Figura 3.8 se visualiza la función de distribución acumulada (véase la Subsección 3.1.6 y el código R en A.20).

Una interpretación basada en una observación simple es que en la mayoría de los departamentos analizados, la proporción de mujeres investigadoras es menor al 50 %, mientras que en el caso de los hombres es poco común encontrar un departamento con una proporción de investigadores masculinos inferior al 50 %. Para conocer si esta diferencia es estadísticamente significativa realizaremos una prueba de contraste de hipótesis. Es decir, se interpretará cuantitativamente los datos para determinar si esta diferencia es significativa desde un punto de vista estadístico.

El primer paso es utilizar los valores numéricos de cada departamento y ordenar nuestros datos para realizar dicha prueba. En la Tabla 3.9 vemos 10 entradas de la manera en como se organizaron los datos. En la Figura 3.9 se muestra el histograma de investigadores femeninos y masculinos dentro de las instituciones presentes en nuestra base de datos<sup>4</sup> (código R en A.28).

---

<sup>4</sup>El significado de las siglas se encuentra en la Sección Siglas institucionales.



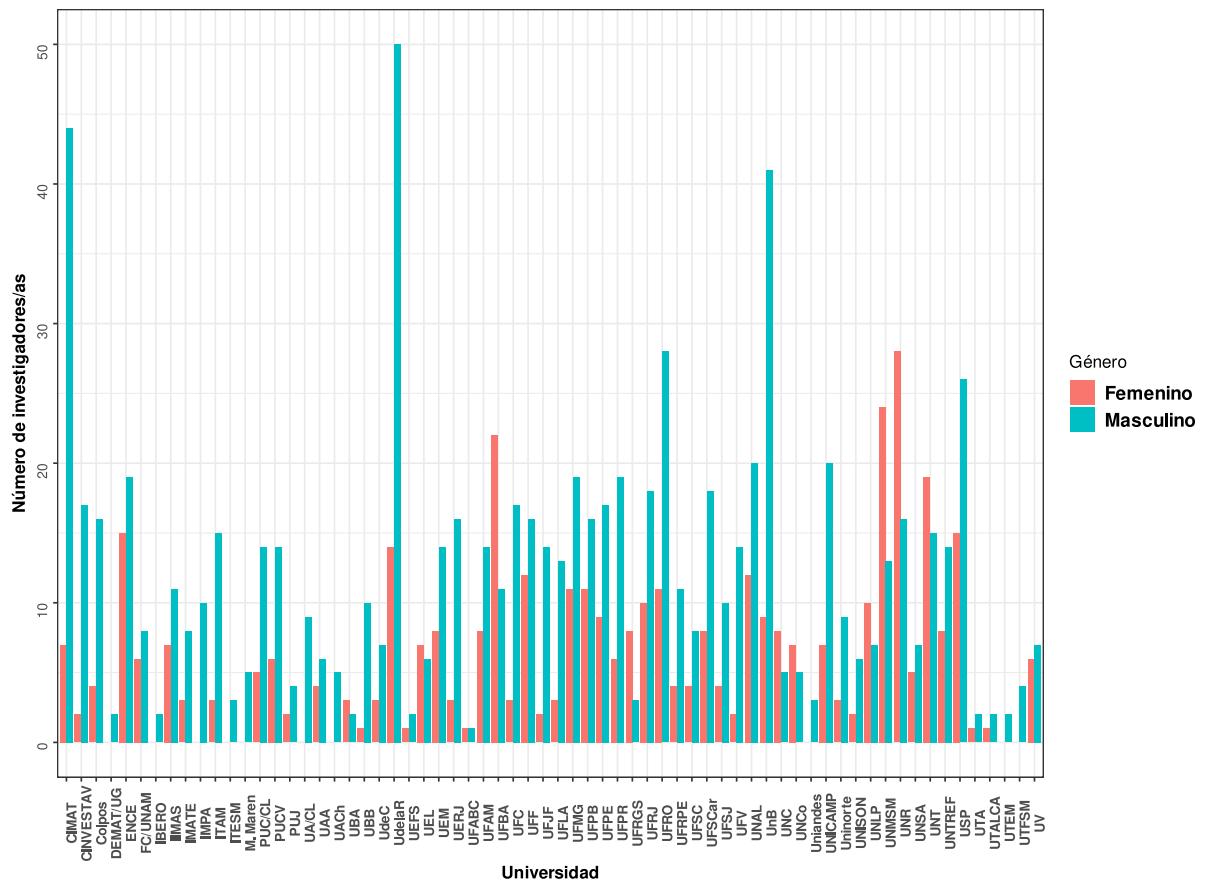
**Fig. 3.8:** Función de distribución acumulada de la proporción de mujeres.

	Institución	Género	Cantidad
1	CIMAT	Femenino	7
2	CINVESTAV	Femenino	2
3	Colegio de Postgraduados	Femenino	4
4	DEMAT/Universidad de Guanajuato	Femenino	0
5	Escuela Nacional de Ciencias Estadísticas	Femenino	15
6	FC/UNAM	Femenino	6
7	IIMAS/UNAM	Femenino	7
8	IMATE/UNAM	Femenino	3
9	Instituto de Matemática Pura y Aplicada	Femenino	0
10	ITAM	Femenino	3
11	CIMAT	Masculino	44
12	CINVESTAV	Masculino	17
13	Colegio de Postgraduados	Masculino	16
14	DEMAT/Universidad de Guanajuato	Masculino	2
15	Escuela Nacional de Ciencias Estadísticas	Masculino	19
16	FC/UNAM	Masculino	9
17	IIMAS/UNAM	Masculino	11
18	IMATE/UNAM	Masculino	8
19	Instituto de Matemática Pura y Aplicada	Masculino	10
20	ITAM	Masculino	15

**Tabla 3.9:** Muestra de instituciones con su respectiva cantidad de hombres y mujeres, ordenados longitudinalmente.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

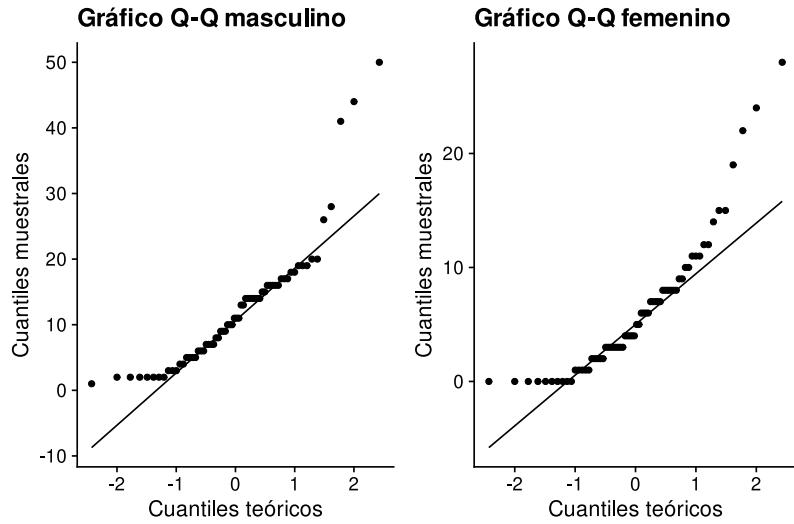


**Fig. 3.9:** Histograma de investigadores en las instituciones según género.

### 3.3 Análisis de género en departamentos

---

Lo siguiente será conocer si la variable “Cantidad” sigue una distribución normal en los datos de los investigadores de género femenino y masculino, para ello en la Figura 3.10 realizamos las gráficas Q-Q para cada género (véase la Subsección 3.1.2 y el código R en A.19).



**Fig. 3.10:** Gráfico Q-Q de cantidad de hombres y mujeres.

De acuerdo a la prueba Q-Q nuestra variable no parece que se distribuya de manera normal. Para tener mayor certeza de ello, realizaremos una prueba Shapiro a los datos de los investigadores femeninos y masculinos por separado. Los resultados se muestran en la Tabla 3.10 (véase la Subsección 3.1.3 y el código R en A.27).

**Tabla 3.10:** Resultado de la prueba Shapiro para la variable “Cantidad” en ambos géneros.

	Masculino		Femenino	
	Statistic	p-value	Statistic	p-value
Test	0.830	3.141e-07	0.855	1.736e-06

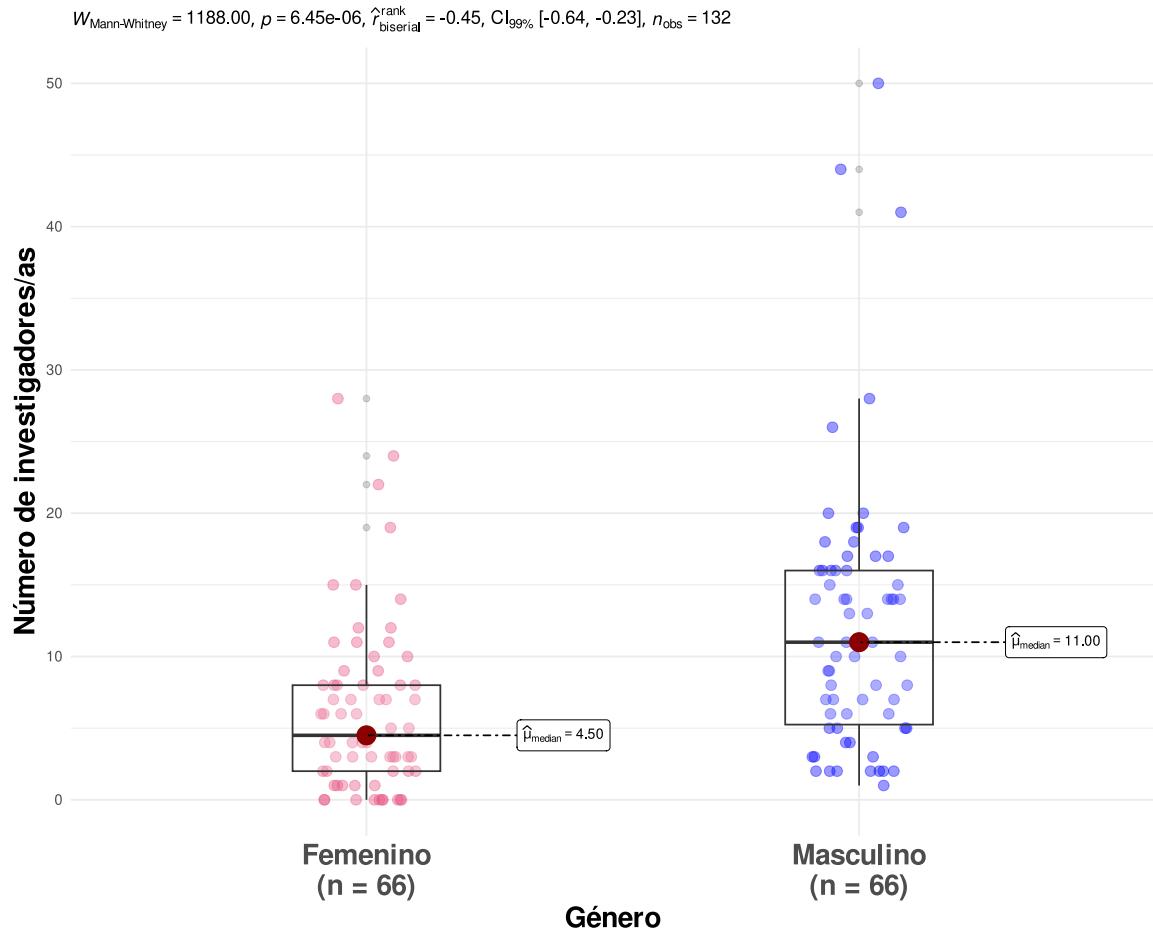
Ahora que sabemos que la probabilidad de que nuestra variable “Cantidad” se aproxime a una normal es muy baja, optaremos por utilizar una prueba no paramétrica.

En la Figura 3.11 se presentan los resultados de la prueba W-Mann-Whitney (véase la Subsección 3.1.5 y el código R en A.9). Para consultar un ejemplo similar, le sugerimos al lector visitar el vídeo [65].

Como resultado obtenemos un  $p-value = 6.45e - 06$  lo que indica que la diferencia observada en los grupos es estadísticamente significativa, es decir, existe una diferencia estadísticamente significativa entre el número de investigadores e investigadoras en los departamentos.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---



**Fig. 3.11:** Gráfica de la prueba W-Mann-Whitney.

### 3.3 Análisis de género en departamentos

Adicionalmente, realizamos un análisis para determinar si la mediana de los porcentajes de mujeres dentro de los departamentos presenta una diferencia estadísticamente significativa en comparación con el valor de referencia del 50 %. Se utiliza el valor del 50 % como referencia, debido a que en un escenario hipotético donde no existiera desigualdad entre el número de mujeres y hombres se esperaría que el porcentaje de mujeres fuera del 50 %. Para consultar un ejemplo similar, le sugerimos al lector visitar la página web [66].

Para comenzar, realizamos un resumen estadístico de los porcentajes de hombres y mujeres dentro de los departamentos (código R en A.10).

Género	Recuento	Media	Mediana	Desviación estándar	Rango intercuártil
Femenino	66	0.298	0.304	0.193	0.240
Masculino	66	0.702	0.696	0.193	0.240

A continuación, en la Figura 3.12 se muestran los porcentajes de mujeres presentes en los departamentos de nuestra base de datos y los porcentajes hipotéticos que se esperarían si no existiera una desigualdad (código R en A.29).

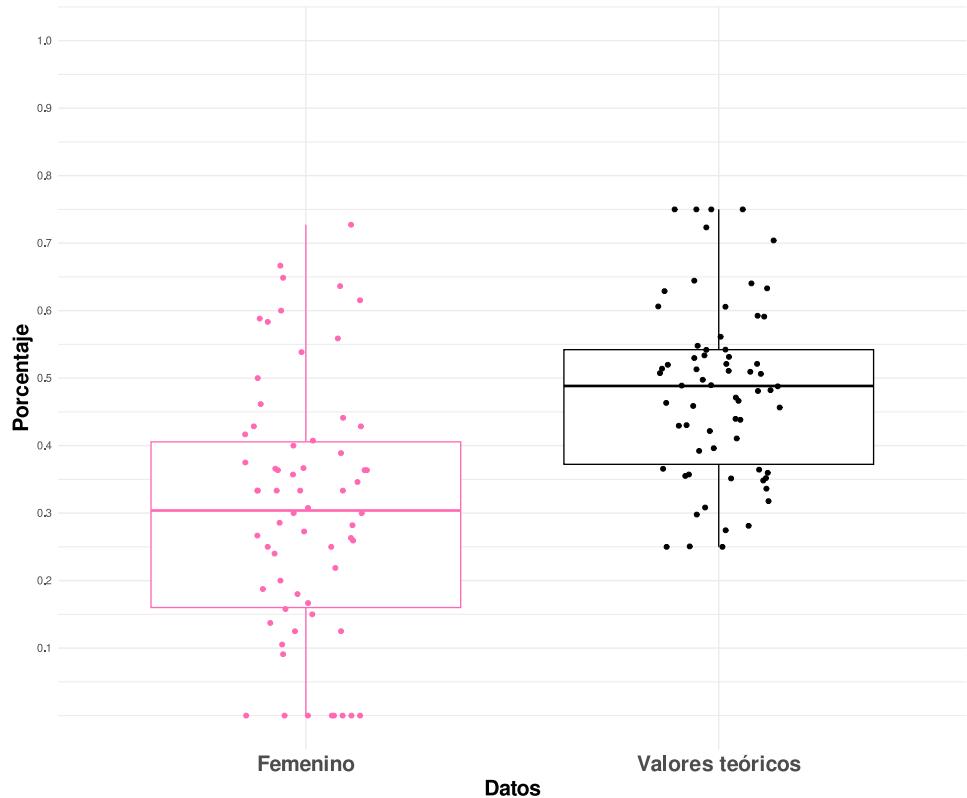
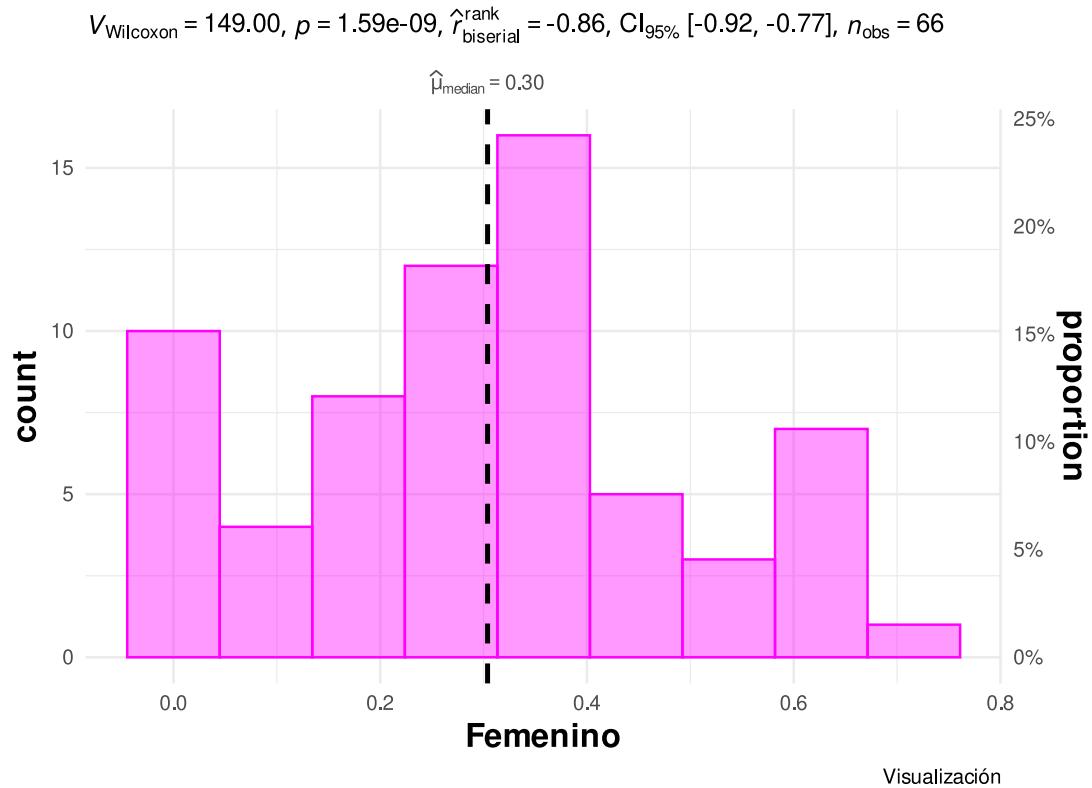


Fig. 3.12: Comparación de los porcentajes de mujeres observados con los porcentajes de mujeres teóricos sin desigualdad.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

Para determinar si la mediana de los porcentajes de mujeres observados dentro de los departamentos, que es igual al 30.4 %, difiere estadísticamente del 50 % realizamos una prueba de rango con signo de Wilcoxon de una muestra considerando un valor hipotético de mediana  $\mu = 0.50$  como referencia para la comparación. El resultado se muestra en la Figura 3.13 (véase la Subsección 3.1.4 y el código R en A.30).



**Fig. 3.13:** Resultado de la prueba Wilcoxon para determinar si la mediana de los porcentajes de mujeres difiere estadísticamente del 50 %.

Como obtenemos un  $p - value = 1.59e - 09$  concluimos que existe una diferencia estadísticamente significativa entre los porcentajes de mujeres observados dentro de los departamentos con el hipotético 50 % que habría si no existiera desigualdad.

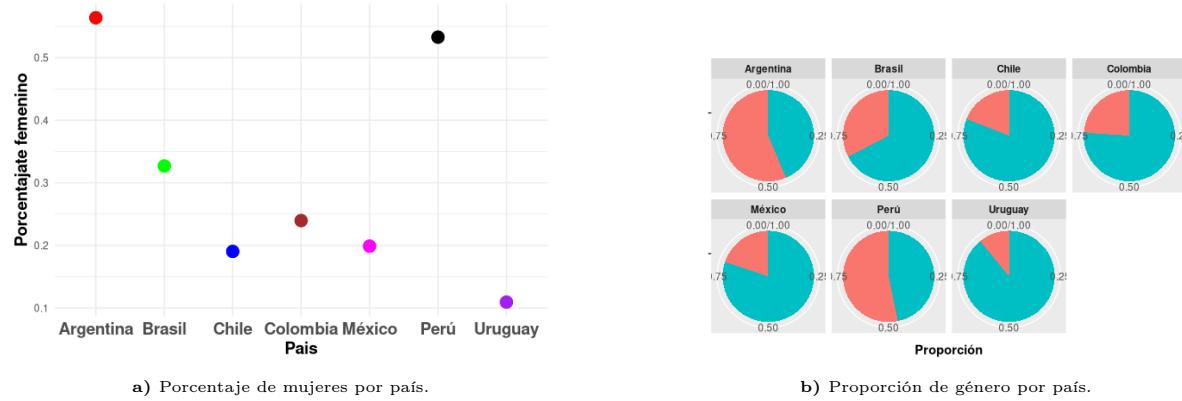
#### 3.3.2. Porcentaje de investigadoras por país

De manera análoga, analizamos el porcentaje de hombres y mujeres realizando investigación en probabilidad o estadística, agrupados por país. Los detalles se exponen en la Tabla 3.11. En las Figuras 3.14b) y 3.14a) se muestra el porcentaje de mujeres en cada país en nuestra base de datos (códigos R en A.22 y A.23).

### 3.3 Análisis de género en departamentos

	País	% Femenino	% Masculino
1	Argentina	0.56	0.44
2	Brasil	0.33	0.67
3	Chile	0.19	0.81
4	Colombia	0.24	0.76
5	México	0.20	0.80
6	Perú	0.53	0.47
7	Uruguay	0.11	0.89

**Tabla 3.11:** Porcentaje de hombres y mujeres por país en la base de datos.



**Fig. 3.14:** Proporción de hombres y mujeres dentro de los países.

Observamos que, entre los registros en nuestra base de datos, México es uno de los países con menor proporción de mujeres investigadoras. Debido a que estamos ubicados en México y considerando esta situación, decidimos enfocarnos en los datos correspondientes a nuestro país.

Para conocer más acerca de la distribución de mujeres investigadoras en México realizamos la Tabla 3.12, que muestra el porcentaje de mujeres investigadoras dentro de las instituciones del país, de acuerdo a nuestra base de datos. Además, realizamos una gráfica de pastel, la cual se muestra en la Figura 3.15, para cada departamento en México con el objetivo de visualizar que todos los departamentos están constituidos mayoritariamente por hombres (código R en A.23).

A continuación, resumimos algunas estadísticas para México (código R en A.10).

Género	Recuento	Media	Mediana	Desviación estándar	Rango intercuártil
Femenino	14	0.199	0.183	0.168	0.334
Masculino	14	0.801	0.817	0.168	0.334

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---



**Fig. 3.15:** Proporción de hombres y mujeres dentro de las instituciones en México.

### 3.4 Independencia del género con las citas recibidas

---

	Institución	% Femenino	% Masculino
1	CIMAT	0.14	0.86
2	CINVESTAV	0.11	0.89
3	Colegio de Postgraduados	0.20	0.80
4	DEMAT/Universidad de Guanajuato	0.00	1.00
5	FC/UNAM	0.40	0.60
6	IIMAS/UNAM	0.39	0.61
7	IMATE/UNAM	0.27	0.73
8	ITAM	0.17	0.83
9	Tecnológico de Monterrey	0.00	1.00
10	UNISON	0.25	0.75
11	Universidad Autónoma Chapingo	0.00	1.00
12	Universidad Autónoma de Aguascalientes	0.40	0.60
13	Universidad Iberoamericana	0.00	1.00
14	Universidad Veracruzana	0.46	0.54

Tabla 3.12: Porcentajes de investigadores por género en instituciones en México.

### 3.4. Independencia del género con las citas recibidas

En los siguientes análisis utilizaremos la variable numérica “Citas”, por lo que solo se utilizarán los datos de los 619 investigadores en la base de datos cuyo número de citas conocemos.

La variable “Citas” se caracteriza por las siguientes estadísticas de resumen (código R en [A.10](#)):

Género	Recuento	Media	Mediana	Desviación estándar	Rango intercuártil
Femenino	174	734	141	2267	586
Masculino	445	985	248	2806	677

Nuestro análisis busca relaciones entre el número de citas que un investigador recibe y su género. Para ello comenzaremos calculando los extremos y cuartiles respecto al total de investigadores que tenemos de la variable género (código R en [A.16](#)).

0 %	25 %	50 %	75 %	100 %
0	31.0	222	688	28642

Ahora, con base en los cuartiles, añadiremos una variable nueva llamada “Categoría de citas”, que constará de 4 niveles, uno para cada cuartil, que van desde: “Bajo”, “Medio”, “Medio alto” y “Alto”. Se puede ver una muestra en la Tabla 3.13.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

	Investigador	Citas	Categoría Citas
1	InvestigadorA	152	Medio
2	InvestigadorB	29	Bajo
3	InvestigadorC	362	Medio alto
4	InvestigadorD	1386	Alto
5	InvestigadorE	2	Bajo
6	InvestigadorF	55	Medio
7	InvestigadorG	1012	Alto
8	InvestigadorH	136	Medio
9	InvestigadorI	11	Bajo
10	InvestigadorJ	731	Alto
11	InvestigadorK	8264	Alto
12	InvestigadorL	483	Medio alto
13	InvestigadorM	141	Medio
14	InvestigadorN	416	Medio alto
15	InvestigadorÑ	20	Bajo

**Tabla 3.13:** Muestra de categorización de investigadores según el número de citas obtenidas (Bajo, Medio, Medio alto, Alto).

Realizamos la Tabla cruzada 3.14 entre el género y la categoría de citas (código R en A.14), de esta tabla obtenemos la Figura 3.16 (código R en A.17). Notemos que en todos los niveles, las mujeres solo representan entre el 22 % y 32 %. Por ejemplo, para la categoría “Alto”, las mujeres solo representan el 25 %.

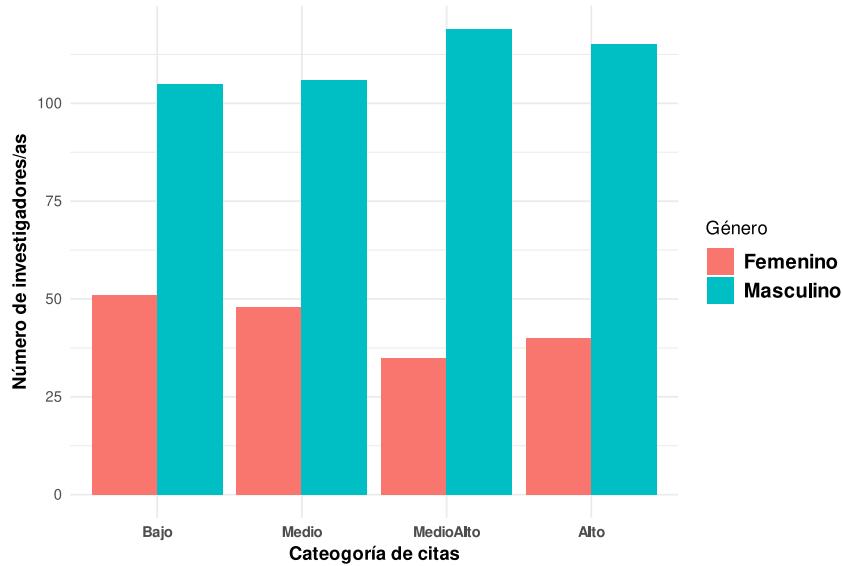
Variable Género	Variable Categoría de citas				
	Bajo	Medio	Medio alto	Alto	Fila Total
<b>Femenino (N)</b>	51	48	35	40	174
N/total de la fila	0.293	0.276	0.201	0.230	0.281
N/total de la columna	0.327	0.312	0.227	0.258	
N/total de la tabla	0.082	0.078	0.057	0.065	
<b>Masculino (N)</b>	105	106	119	115	445
N/total de la fila	0.236	0.238	0.267	0.258	0.719
N/total de la columna	0.673	0.688	0.773	0.742	
N/total de la tabla	0.170	0.171	0.192	0.186	
Columna Total	156	154	154	155	619
Proporción	0.252	0.249	0.249	0.250	

**Tabla 3.14:** Comparación de género en las distintas categorías según su cantidad de citas.

Para averiguar si existe alguna relación entre tener un nivel alto de citas y el género realizaremos una prueba estadística. Comenzamos dividiendo nuestra variable entre “Alto” y “No alto”, este segundo nivel se creará colapsando los niveles “Bajo”, “Medio” y “Medio alto”. Después de agruparlos entre “Alto” y “No alto”, los datos se muestran en la Tabla 3.15 y en la Figura 3.17 (códigos R en A.14 y A.17).

### 3.4 Independencia del género con las citas recibidas

---



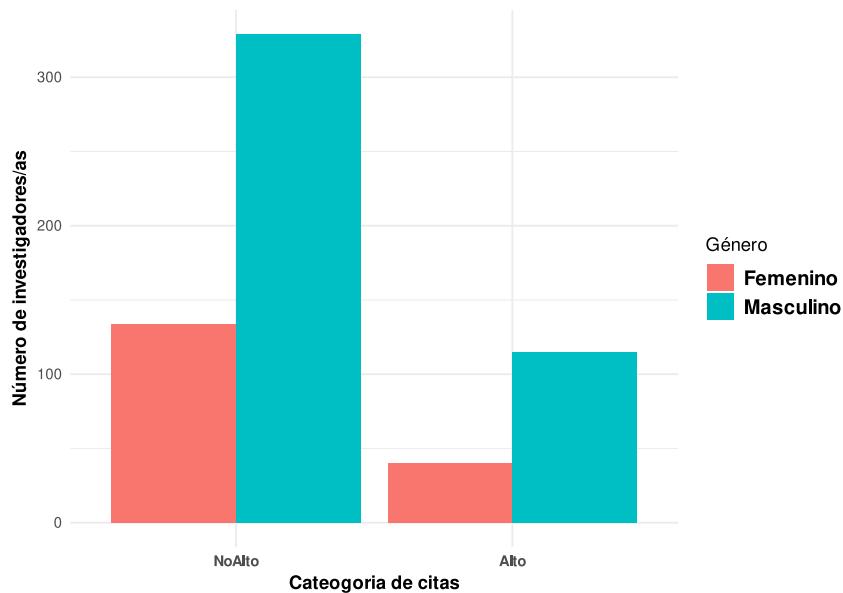
**Fig. 3.16:** Categorías de citas por género.

	Variable Categoría de citas		
Variable Género	No alto	Alto	Fila Total
<b>Femenino (N)</b>	134	40	174
N/total de la fila	0.770	0.230	0.281
N/total de la columna	0.289	0.258	
N/total de la tabla	0.216	0.065	
<b>Masculino (N)</b>	330	115	445
N/total de la fila	0.742	0.258	0.719
N/total de la columna	0.711	0.742	
N/total de la tabla	0.533	0.186	
Columna Total	464	155	619
Proporción	0.750	0.250	

**Tabla 3.15:** Comparación de género en los niveles de citas “No alto” y “Alto”.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---



**Fig. 3.17:** Categorías “Alto” y “NoAlto” divididas por género.

A continuación, se realiza un procedimiento análogo al utilizado en la Sección 3.2.1. En este caso, generamos conjuntos de datos tomando como cierta la hipótesis de independencia de las variables “Género” y la variable “Categoría de citas”, por lo que fijaremos a la variable género y permutaremos los datos de la variable “Categoría de citas”.

Utilizamos la técnica bootstrap para generar 500 conjuntos de datos en los cuales las variables son independientes. A cada conjunto de datos le calculamos el estadístico muestral, es decir, la diferencia de proporciones entre los géneros. Para consultar un ejemplo similar, le sugerimos al lector visitar la Sección “Intervals for differences” del curso “Inference for Categorical Data in R” en DataCamp [64].

En la Tabla 3.16 vemos 10 de las 500 réplicas generadas por la técnica bootstrap y la diferencia de proporciones de cada conjunto de datos generados (véase la Subsección 3.1.1 y el código R en A.15).

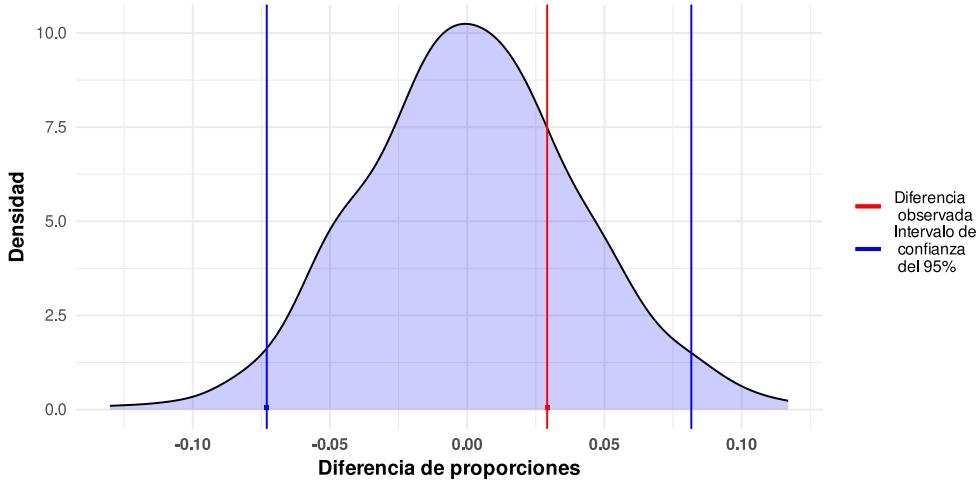
replica	1	2	3	4	5	6	7	8	9	10
stat	0.01	-0.02	0.05	-0.01	-0.06	-0.01	0.00	0.01	-0.01	0.04

**Tabla 3.16:** Diferencia entre 10 conjuntos de datos generados.

Calculamos la diferencia observada en nuestro conjunto de datos original:

$$\text{Diferencia observada} = \text{proporción hombres} - \text{proporción mujeres} = \frac{115}{445} - \frac{40}{174} = 0.028.$$

En la Figura 3.18, trazamos una línea roja en la gráfica de la distribución nula que representa la diferencia observada en el conjunto de datos original. Notamos que la diferencia observada es muy parecida al tipo de diferencias que veríamos si las variables fueran independientes y se encuentra dentro del intervalo de confianza del 95 % cuyos límites (-0.073, 0.081) se determinan mediante el cálculo de los percentiles 2.5 y 97.5 de la variable “stat”. Por lo que estos datos constituyen evidencia de que no existe ninguna asociación entre tener un número alto de citas y el género<sup>5</sup>.



**Fig. 3.18:** Densidad nula cuando “Citas” y “Género” son independientes.

## 3.5. Clústeres

En los siguientes análisis haremos uso de un método de machine learning llamado K-means para agrupar a las instituciones en función de su nivel de citas y artículos (véase la Subsección 3.1.7 y el código R en A.12). Para este propósito, a cada institución se le ha asignado un identificador único (ID) para facilitar su representación en estos análisis.

Comenzaremos calculando la media de citas y artículos para los investigadores(as) de 41 instituciones. Esto se muestra en la Tabla 3.17.

Si nos centramos en la variable “Citas” y ordenamos las instituciones de mayor a menor, observamos en la Subtabla 3.18a que existe una gran diferencia entre las dos primeras instituciones en comparación con la tercera. Es decir, la institución “AJ” tiene un promedio de citas que supera a la institución “AA” por 1789.07, y a su vez, la institución “T” la supera por 1343.39. A partir del cuarto sitio las medias se reducen de manera menos pronunciada.

---

<sup>5</sup>También se llevó a cabo este análisis separando a los investigadores en probabilidad y estadística en grupos individuales y en ningún grupo se encontró dependencia entre las variables “Citas” y “Género”.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

	Institución	Citas	Artículos
1	A	828.02	28.53
2	B	1206.79	65.53
3	C	1183.56	84.12
4	D	63.50	18.00
5	E	268.62	39.05
6	F	214.33	24.86
7	G	400.06	29.00
8	H	372.22	38.89
9	I	4090.25	149.38
10	J	347.19	39.81
11	K	101.00	17.00
12	L	147.62	16.75
13	M	204.00	44.75
14	N	154.56	26.89
15	O	706.69	38.81

**Tabla 3.17:** Muestra de 15 instituciones.

Para evitar que la distribución de las medias de las citas nos cause problemas, realizaremos una transformación logarítmica a la variable “Citas”. Como sabemos, en una escala logarítmica la distancia entre 1 y 10, 10 y 100, 100 y 1000 es la misma. Esto nos ayuda a que los datos pequeños se expandan y los mayores se compriman. Los resultados se muestran en la Subtabla 3.18b. Notamos que después de aplicar la transformación logarítmica el orden se preserva.

a Muestra del promedio de citas y artículos de instituciones.

Institución	Citas	Artículos
AJ	4535.93	411.53
I	4090.25	149.38
AA	2746.86	201.50
AE	1863.95	92.90
AF	1671.67	84.89
P	1588.83	109.28
T	1575.37	49.37
Q	1209.92	70.42
B	1206.79	65.53
C	1183.56	84.12
AB	1151.50	68.86
AK	883.44	38.06
A	828.02	28.53
AD	785.50	64.95
O	706.69	38.81
AM	537.00	76.00

b Muestra del promedio de citas y artículos con transformación logarítmica.

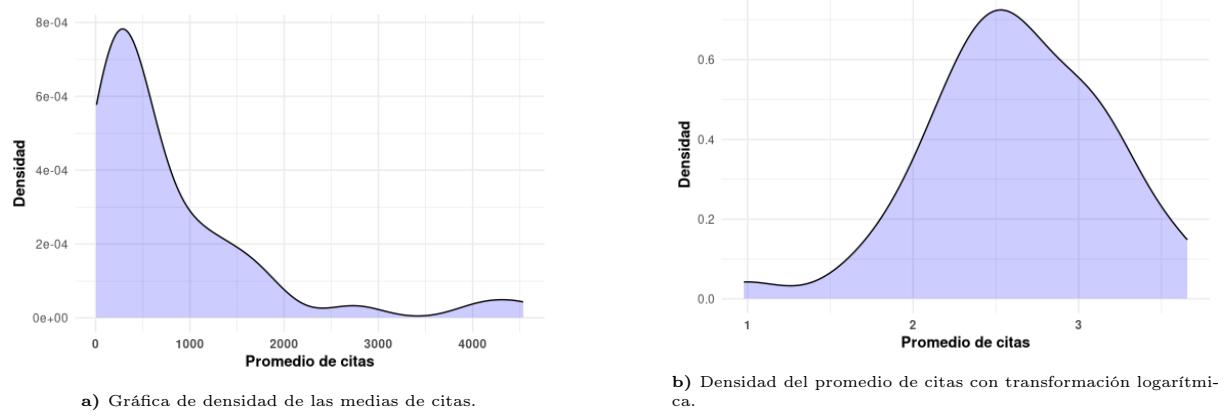
Institución	Citas	Artículos
AJ	3.66	2.61
I	3.61	2.17
AA	3.44	2.30
AE	3.27	1.97
AF	3.22	1.93
P	3.20	2.04
T	3.20	1.69
Q	3.08	1.85
B	3.08	1.82
C	3.07	1.92
AB	3.06	1.84
AK	2.95	1.58
A	2.92	1.46
AD	2.90	1.81
O	2.85	1.59
AM	2.73	1.88

**Tabla 3.18:** Muestra de promedios de citas y artículos.

### 3.5 Clústeres

---

En la Figura 3.19a) se muestra la gráfica de densidad de los promedios de citas originales, mientras que en la Figura 3.19b) se presenta la gráfica de densidad de las medias de las citas con transformación logarítmica. Notamos que los datos se comportan de mejor manera en esta última representación.



**Fig. 3.19:** Densidad del promedio de citas de las 41 instituciones.

Dado que se observan grandes diferencias en las medias de los artículos entre las instituciones, también se aplicó una transformación logarítmica a estos datos.

De esta manera continuaremos con el método K-means. Para consultar un ejemplo similar, le sugerimos al lector visitar el vídeo [67]. Lo siguiente será centralizar nuestros datos, ya que es un requisito del método.

Institución	Citas	Artículos
AJ	1.91	3.01
I	1.83	1.67
AA	1.51	2.07
AE	1.20	1.05
AF	1.11	0.93
P	1.07	1.26
T	1.07	0.21
Q	0.85	0.68
B	0.85	0.59
C	0.84	0.92
AB	0.81	0.65
AK	0.60	-0.13
A	0.55	-0.51
AD	0.51	0.58
O	0.42	-0.10
AM	0.20	0.78

**Tabla 3.19:** Datos centralizados.

A continuación calculamos las distancias entre las instituciones empleando la distancia euclíadiana, que para los puntos  $(x_2 - x_1), (y_2 - y_1) \in \mathbb{R}$  se define por la fórmula

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}. \quad (3.3)$$

Para ejemplificar este proceso, consideremos el cálculo de la distancia euclíadiana entre los institutos “A” y “B”. Cuyos valores correspondientes se muestran en la Tabla 3.20.

Institución	Citas	Artículos
A	0.55	-0.51
B	0.85	0.59

Tabla 3.20: Datos centralizados para instituciones A y B.

Para conocer dicha distancia, realizamos el siguiente cálculo

$$\sqrt{(0.55 - 0.85)^2 + (-0.51 - 0.59)^2} \approx 1.14 \quad (3.4)$$

En la Figura 3.20 se muestran las distancias entre instituciones. En esta representación, las distancias más cortas se muestran en color negro, las distancias intermedias se muestran en blanco y las distancias más largas se representan en rojo.

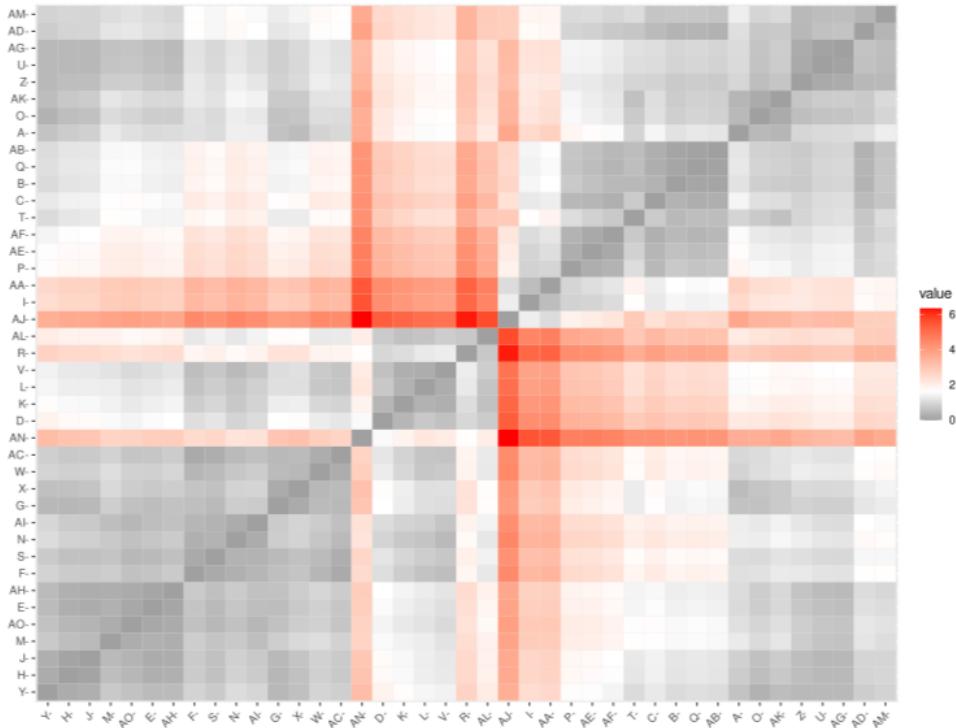
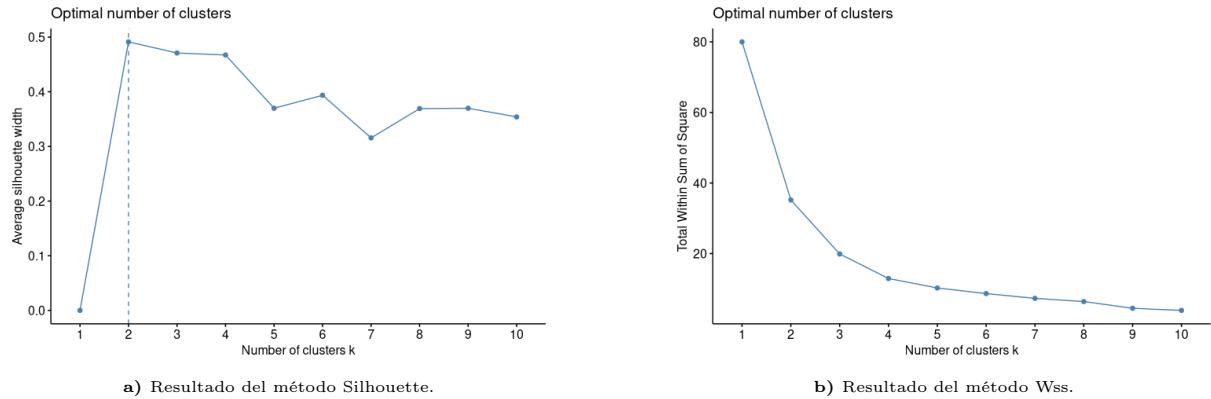


Fig. 3.20: Gráfica de distancias.

### 3.5 Clústeres

---

Lo siguiente será conocer el número óptimo de clústeres que debemos hacer. En la Figura 3.21a), el método Silhouette sugiere 2 clústeres, mientras que en la Figura 3.21b), el método Wss recomienda 2 o 3 clústeres.



**Fig. 3.21:** Recomendaciones de dos métodos sobre el número óptimo de clústeres.

Debido a que existen varios métodos y generalmente varían en el resultado que arrojan, utilizamos la función NbClust, que nos dará un resumen de 30 métodos. Los resultados de la función indican que el número óptimo de clústeres es 2. Los resultados de la consola de RStudio se pueden consultar en el Apéndice D.1.

Al momento de realizar los clústeres, las instituciones se organizan como se muestran en la Tabla 3.21.

B	C	I	A	D	E	F
1	1	1	2	2	2	2
O	P	Q	G	H	J	K
1	1	1	2	2	2	2
T	U	Z	L	M	N	R
1	1	1	2	2	2	2
AA	AB	AD	S	S	W	X
1	1	1	2	2	2	2
AE	AF	AG	Y	AC	AH	AI
1	1	1	2	2	2	2
AJ	AK	AM	AL	AN	AO	
1	1	1	2	2	2	

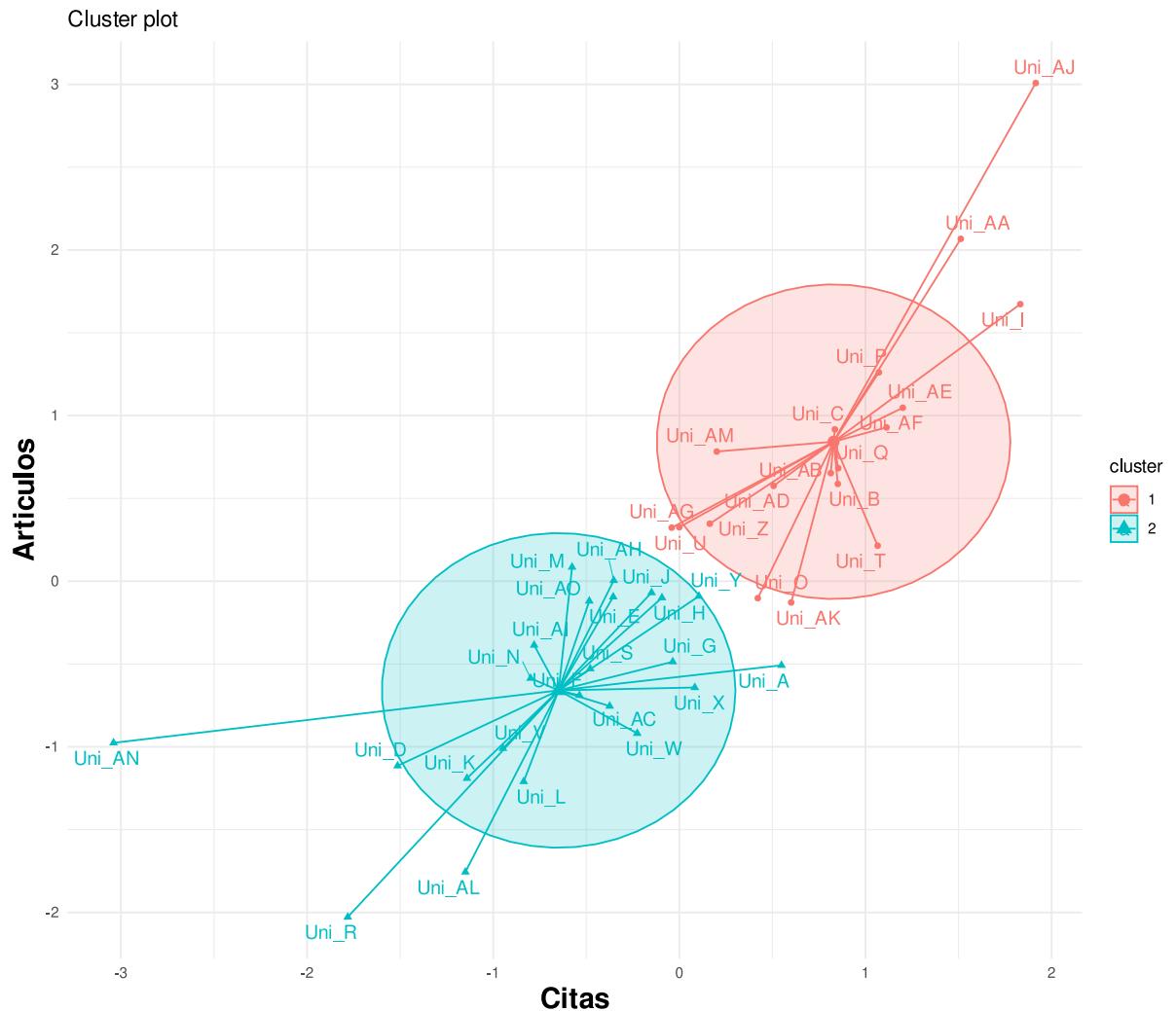
**Tabla 3.21:** Instituciones divididas en clústeres según su nivel en citas y artículos.

Esto se visualiza con la función *fvizcluster* en la Figura 3.22 (código R en A.12).

En la Figura 3.22 observamos que en el clúster número 1 el punto de la institución “I” es uno de los que sobresale. Si calculamos la proporción de género dentro de la universidad, nos percatamos que la institución “I” está conformado únicamente por hombres.

Institución	%Femenino	%Masculino
I	0	100

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS



**Fig. 3.22:** Clústeres de Instituciones según su nivel en citas y artículos.

### 3.5 Clústeres

---

Lo cual nos hace preguntarnos si lo mismo sucede en las instituciones del clúster número 1. Es decir, si las instituciones que tienen un alto nivel de citas y artículos están constituidas mayoritariamente por investigadores masculinos.

Para ello, agrupamos a las instituciones por clúster y calcularemos el porcentaje de mujeres y hombres que tiene cada institución. En las Tablas 3.22 y 3.23 se presentan una muestra de los resultados. Para conocer los datos completos de ambos clústeres, le sugerimos al lector consultar los Apéndices C.4 y C.5.

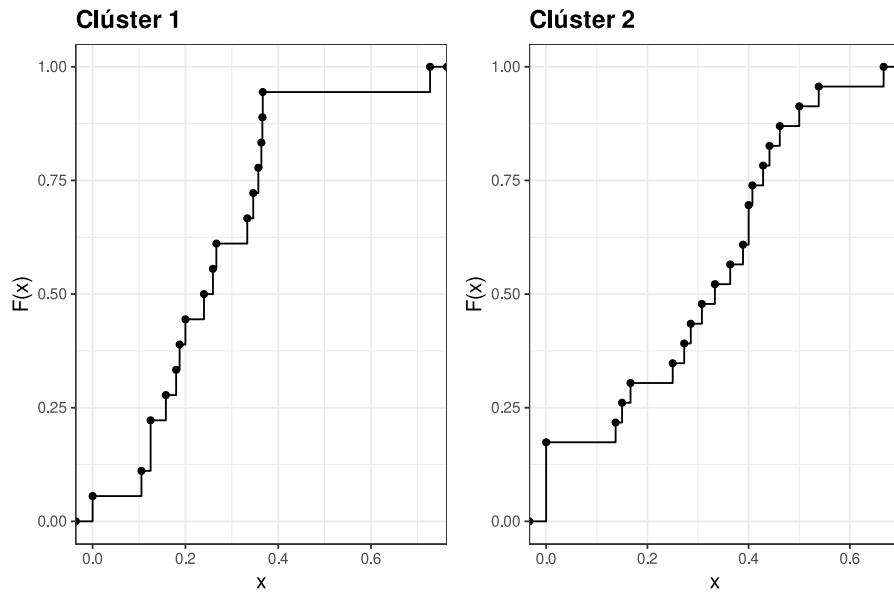
	Institución	Clúster	%Femenino	%Masculino	Promedio de citas	Promedio de artículos
1	B	1	0.11	0.89	1206.79	65.53
2	C	1	0.20	0.80	1183.56	84.12
3	I	1	0.00	1.00	4090.25	149.38
4	O	1	0.18	0.82	706.69	38.81
5	P	1	0.37	0.63	1588.83	109.28
6	Q	1	0.26	0.74	1209.92	70.42
7	T	1	0.36	0.64	1575.37	49.37
8	U	1	0.16	0.84	417.93	53.73
9	Z	1	0.12	0.88	512.60	54.60
10	AA	1	0.19	0.81	2746.86	201.50

**Tabla 3.22:** Instituciones del clúster número 1 y sus porcentajes de mujeres y hombres.

	Institución	Clúster	%Femenino	%Masculino	Promedio de citas	Promedio de artículos
1	A	2	0.14	0.86	828.02	28.53
2	D	2	0.00	1.00	63.50	18.00
3	E	2	0.44	0.56	268.62	39.05
4	F	2	0.40	0.60	214.33	24.86
5	G	2	0.39	0.61	400.06	29.00
6	H	2	0.27	0.73	372.22	38.89
7	J	2	0.17	0.83	347.19	39.81
8	K	2	0.00	1.00	101.00	17.00
9	L	2	0.25	0.75	147.62	16.75
10	M	2	0.00	1.00	204.00	44.75

**Tabla 3.23:** Instituciones del clúster número 2 y sus porcentajes de mujeres y hombres.

En la Figura 3.23 se muestran las gráficas de la función de probabilidad acumulada para cada clúster con el fin de visualizar el porcentaje de instituciones que tienen como máximo un determinado porcentaje de mujeres investigadoras. Es decir  $\mathbb{P}(X \leq x)$ , donde  $X$  representa la proporción de instituciones y  $x$  el porcentaje máximo (véase la Subsección 3.1.6 y el código R en A.21). Esta comparación nos permite identificar posibles diferencias en la distribución de mujeres investigadoras entre los clústeres.



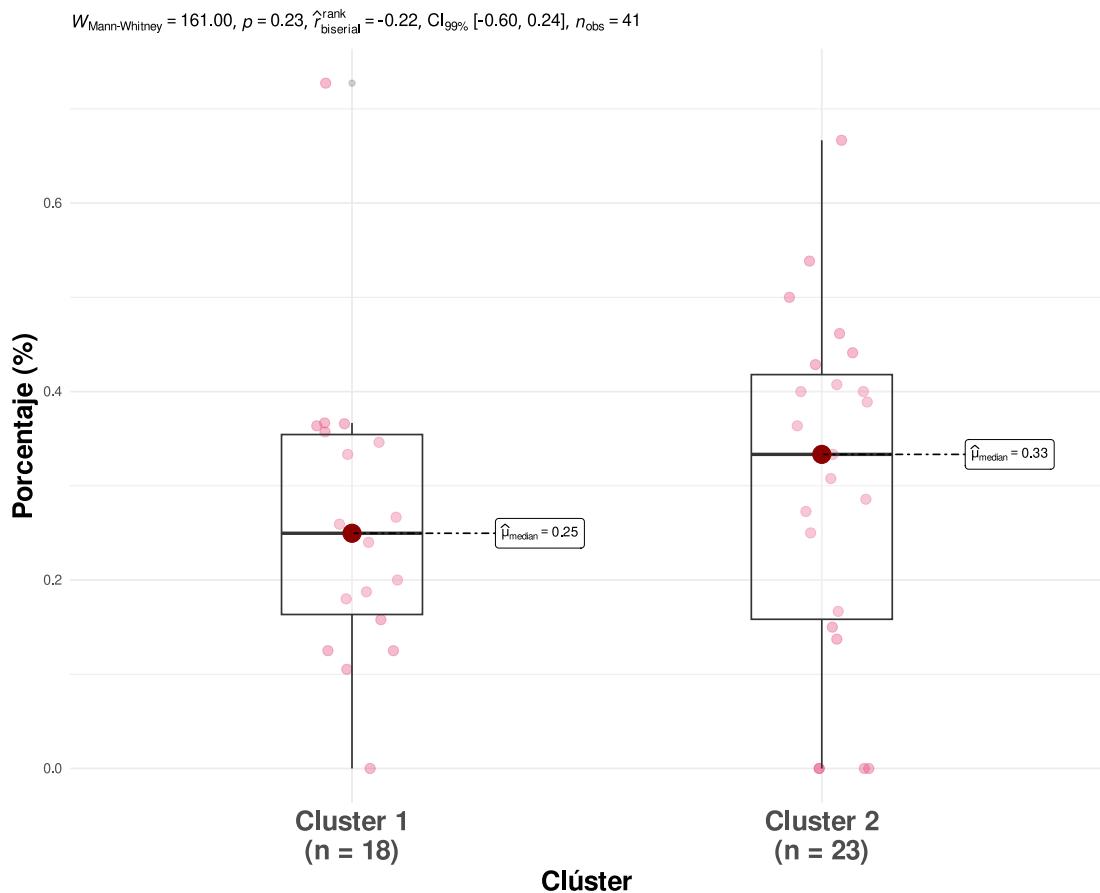
**Fig. 3.23:** Gráficas de la función de probabilidad acumulada de cada clúster.

A continuación, realizamos una prueba W-Mann-Whitney (véase la Subsección 3.1.5 y el código R en A.9) para determinar si hay una diferencia estadísticamente significativa entre los porcentajes de mujeres en el clúster de instituciones con un promedio alto de artículos y citas, y los porcentajes de mujeres en el clúster de instituciones con un promedio menor de artículos y citas. Los resultados se muestran en la Figura 3.24.

Obtenemos un  $p-value = 0.23$ , por lo que no rechazamos la hipótesis nula, es decir que no contamos con evidencia estadística suficiente para concluir que exista una diferencia significativa entre los porcentajes de mujeres en los clústeres.

El resumen estadístico es el siguiente (código R en A.10).

Clúster	Recuento	Media	Mediana	Desviación estándar	Rango intercuártil
1	18	0.261	0.250	0.159	0.191
2	23	0.300	0.333	0.187	0.260



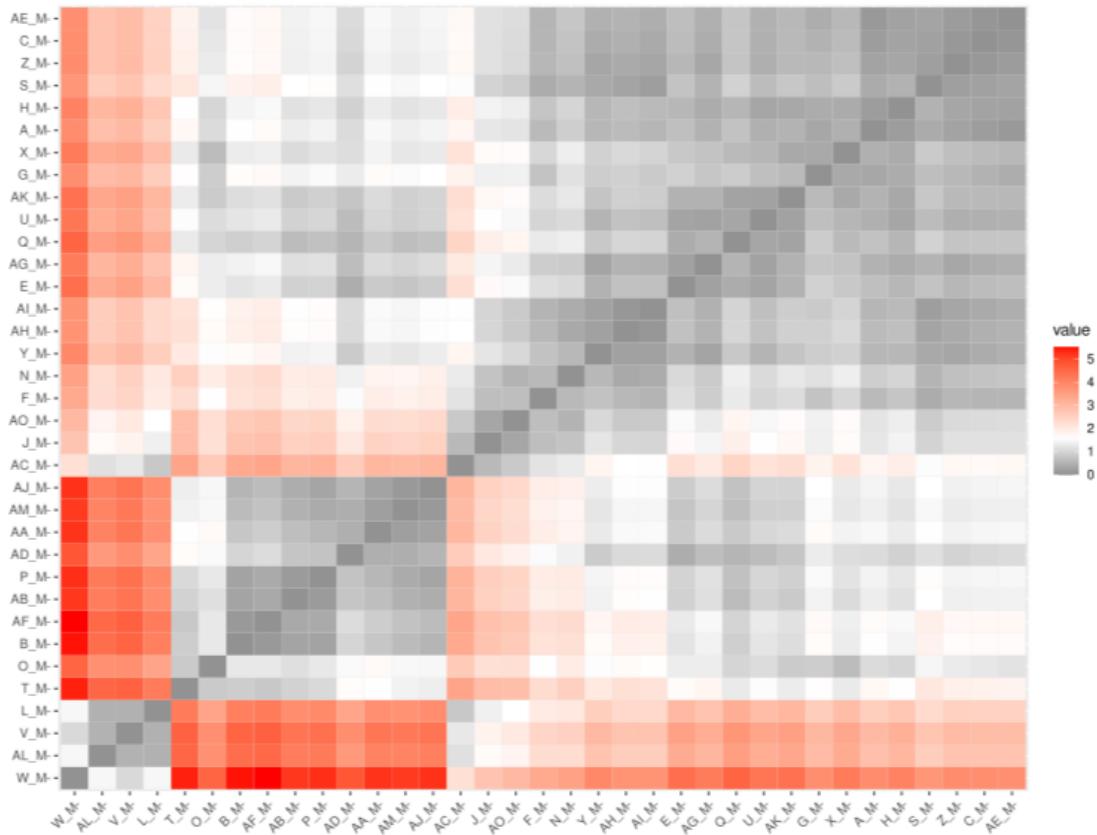
**Fig. 3.24:** Comparación de los porcentajes de mujeres entre clústeres.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

#### 3.5.1. Clústeres con datos únicamente de mujeres

Ahora nos enfocaremos en analizar el comportamiento de los datos exclusivamente para el grupo de mujeres. Para ello, filtramos los datos, seleccionando únicamente a la población femenina de cada universidad. Posteriormente, calculamos el promedio de citas y artículos asociados a este grupo. Asimismo, aplicamos una transformación logarítmica a los datos y los centralizamos. Los resultados se muestran en los Apéndices C.2<sup>6</sup> y C.3 (véase la Subsección 3.1.7 y el código R en A.12).

La matriz de las distancias euclidianas se muestra en la Figura 3.25, donde el negro representa una distancia corta, el blanco una distancia mediana y el rojo una distancia grande.



**Fig. 3.25:** Gráfica de distancias entre instituciones, únicamente con datos de mujeres.

El resumen de 30 métodos realizado por la función NbClust nos indica que el número óptimo de clústeres es 2. Los resultados de la consola de RStudio se pueden consultar en el Apéndice D.2.

<sup>6</sup>Notemos que el número de instituciones se redujo a 35.

### 3.5 Clústeres

Al realizar los clústeres, las instituciones quedan ordenadas como se muestran en la Tabla 3.24. Para consultar un ejemplo similar, le sugerimos al lector visitar el vídeo [67]. Los clústeres los visualizamos en la Figura 3.26.

A <sub>1</sub> M	B <sub>1</sub> M	C <sub>1</sub> M	E <sub>1</sub> M	F <sub>1</sub> M	G <sub>1</sub> M	H <sub>1</sub> M
N <sub>1</sub> M	S <sub>1</sub> M	T <sub>1</sub> M	U <sub>1</sub> M	X <sub>1</sub> M	Y <sub>1</sub> M	AL <sub>1</sub> M
Z <sub>1</sub> M	AA <sub>1</sub> M	AB <sub>1</sub> M	AE <sub>1</sub> M	AD <sub>1</sub> M	AF <sub>1</sub> M	AK <sub>1</sub> M
AM <sub>1</sub> M	AG <sub>1</sub> M	AH <sub>1</sub> M	AI <sub>1</sub> M	AJ <sub>1</sub> M	O <sub>1</sub> M	Q <sub>1</sub> M
P <sub>1</sub> M	J <sub>2</sub> M	V <sub>2</sub> M	W <sub>2</sub> M	AC <sub>2</sub> M	L <sub>2</sub> M	AO <sub>2</sub> M

Tabla 3.24: Instituciones divididas en clústeres según su nivel en citas y artículos, únicamente con datos de mujeres.

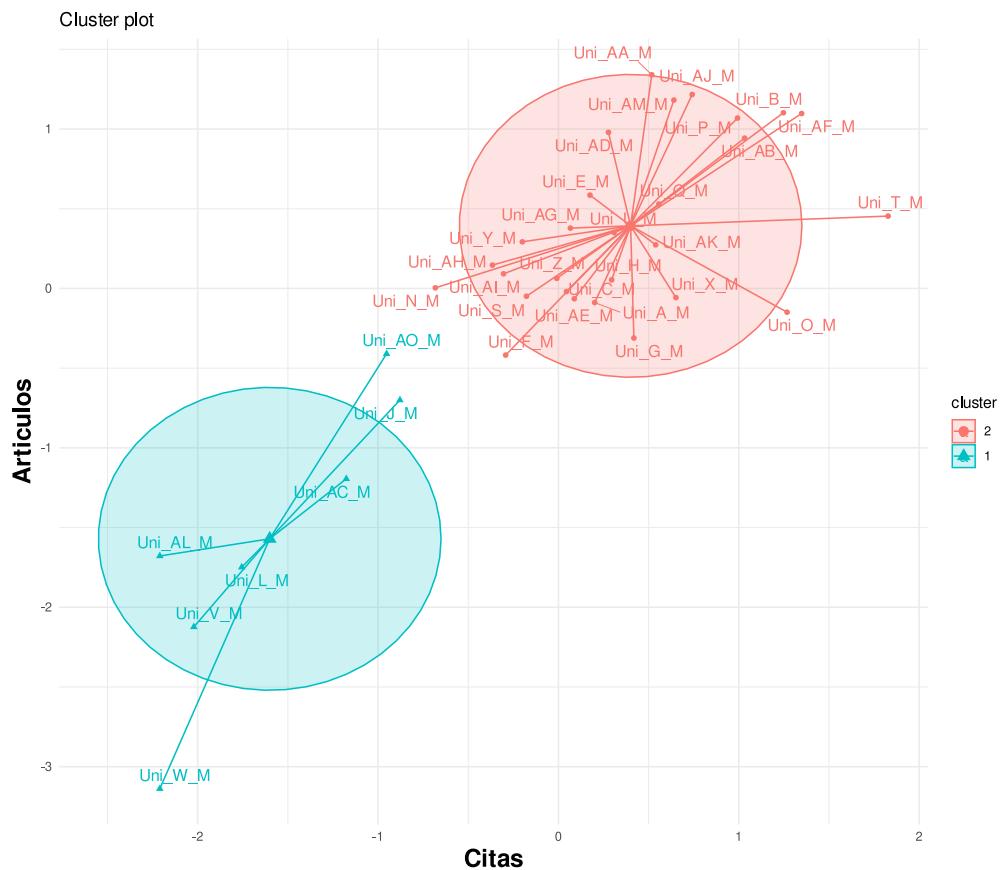


Fig. 3.26: Clústeres de instituciones según su nivel en citas y artículos, únicamente con datos de mujeres.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

En la Figura 3.26 notamos que si analizamos únicamente los datos de las mujeres, el clúster número 2 se reduce y se convierte en un subconjunto del clúster 2 presentado en el análisis anterior.

#### 3.5.2. Comparación de promedios en artículos y citas por instituciones divididos por género

Buscamos conocer si este patrón se mantiene al agregar los datos de los hombres. Para ello, dividimos a las instituciones entre poblaciones de hombres y mujeres, y utilizamos sus respectivos datos para el análisis. Para evitar complicaciones con la distribución del promedio de citas y artículos, también aplicamos la transformación logarítmica a cada variable. Posteriormente centralizamos los datos.

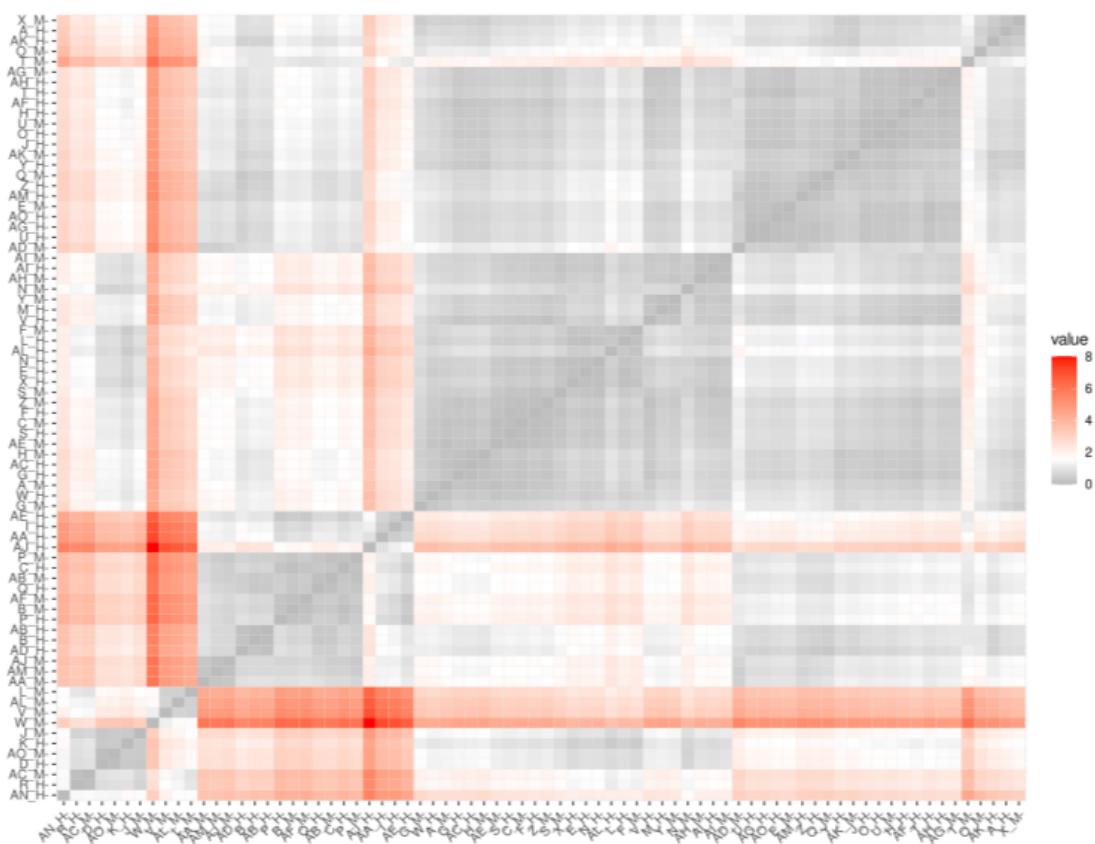
En la Tabla 3.25 se presenta una muestra del resultado, y en la Figura 3.27 graficamos la matriz de las distancias euclidianas.

Institución	Citas	Artículos
A_H	0.73	-0.26
B_H	0.89	0.56
C_H	1.08	1.10
D_H	-1.21	-0.77
E_H	-0.37	-0.49
F_H	-0.17	-0.24
AN_H	-2.59	-0.65
G_H	0.03	-0.11
A_M	0.05	-0.32
B_M	1.21	1.06
C_M	-0.13	-0.24
E_M	0.02	0.46
F_M	-0.50	-0.71
G_M	0.29	-0.58
H_M	0.15	-0.16
J_M	-1.15	-1.04

**Tabla 3.25:** Muestra centralizada de promedios de citas y artículos con transformación logarítmica, con datos de instituciones divididas por género.

El resumen de 30 métodos realizado por la función NbClust nos indica que el número óptimo de clústeres es 2. Los resultados de la consola de RStudio se pueden consultar en el Apéndice D.3. En la Tabla 3.26, se muestra el desglose de las instituciones en los clústeres, y en la Figura 3.28 se puede apreciar la representación gráfica de estos clústeres.

Debido a que el exceso de puntos dificulta la visualización, para conocer qué lugar ocupan los promedios de citas y artículos de las mujeres en comparación con los promedios de los hombres, realizamos el gráfico de dispersión que se muestra en la Figura 3.29 (código R en A.24). Además, en la Figura 3.30 se lleva a cabo este análisis utilizando los datos de los investigadores de manera individual (código R en A.24).



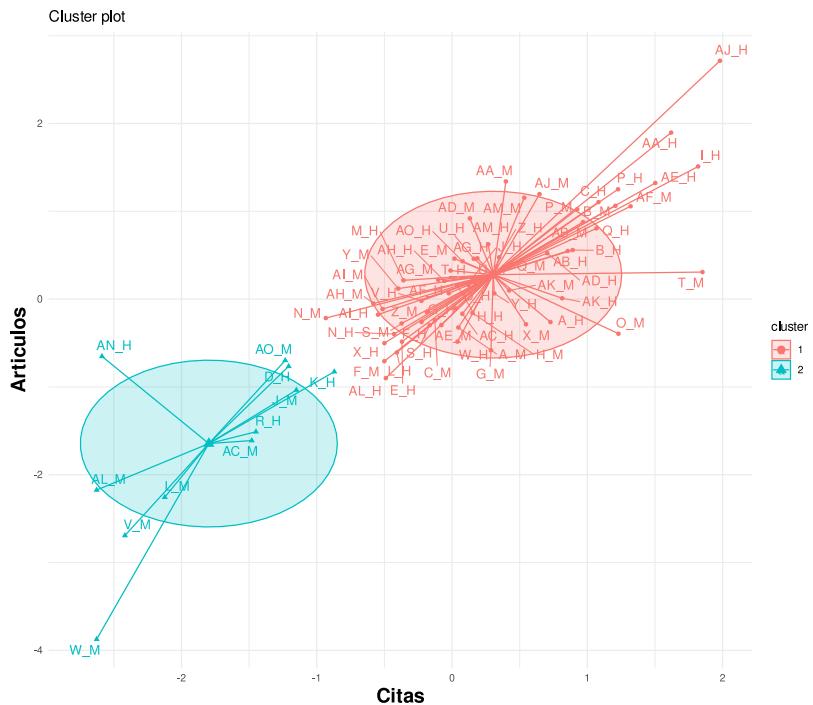
**Fig. 3.27:** Distancia entre instituciones, con datos de instituciones divididas por género.

A_H 1	B_H 1	C_H 1	E_H 1	F_H 1	G/_H 1	H/_H 1
I_H 1	J_H 1	N_H 1	M_H 1	S_H 1	T_H 1	U_H 1
V_H 1	W_H 1	X_H 1	Y_H 1	UFF_H 1	Z_H 1	AA_H 1
AB_H 1	AC_H 1	AE_H 1	AD_H 1	AF_H 1	AK_H 1	AM_H 1
AG_H 1	AH_H 1	AI_H 1	AJ_H 1	O_H 1	Q_H 1	L_H 1
P_H 1	UV_H 1	A_M 1	B_M 1	C_M 1	E_M 1	F_M 1
G_M 1	H_M 1	N_M 1	S_M 1	T_M 1	U_M 1	X_M 1
Y_M 1	Z_M 1	AA_M 1	AB_M 1	AE_M 1	AD_M 1	AF_M 1
AK_M 1	AM_M 1	AG_M 1	AH_M 1	AI_M 1	AJ_M 1	O_M 1
Q_M 1	P_M 1					
D_H 2	AN_H 2	K_H 2	R_H 2	J_M 2	V_M 2	W_M 2
AL_M 2	AC_M 2	L_M 2	AO_M 2			

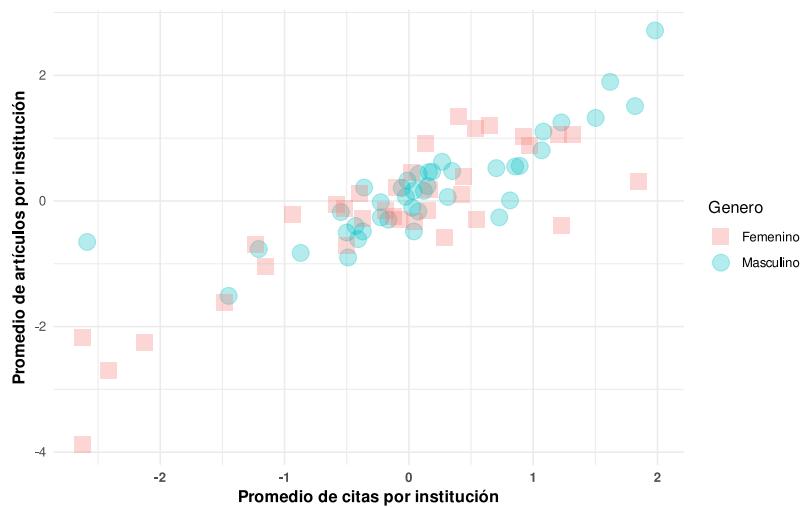
Tabla 3.26: Instituciones divididas en clústeres según su nivel en citas y artículos, con datos clasificados por género.

### 3.5 Clústeres

---



**Fig. 3.28:** Clústeres de instituciones según su nivel en citas y artículos, con datos de instituciones divididas por género.



**Fig. 3.29:** Gráfico de dispersión de medias de citas y artículos, con datos de instituciones divididas por género.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

Notamos que, tanto en la Figura 3.29 como en la Figura 3.30, en el número de citas no parece existir mucha diferencia entre el promedio de mujeres y el de los hombres. En el caso de artículos, se observa una ligera ventaja en el promedio de los hombres en comparación con el de las mujeres.

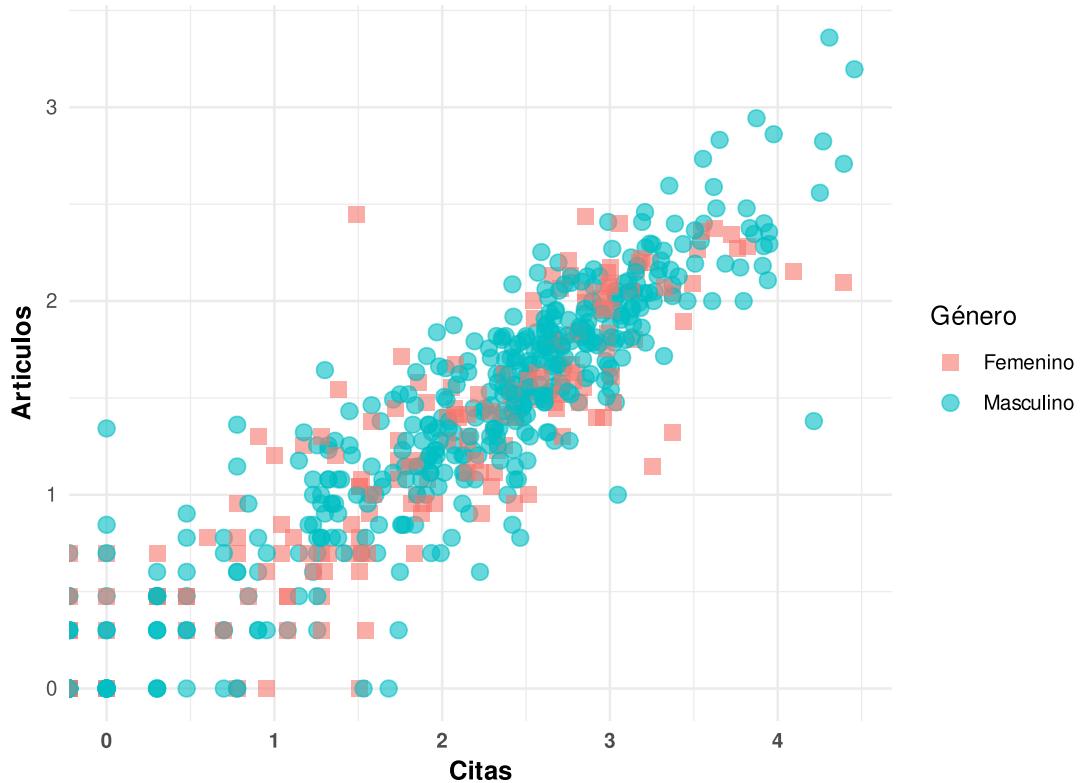


Fig. 3.30: Gráfico de dispersión con datos de investigadores e investigadoras individualmente.

#### 3.6. Correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento

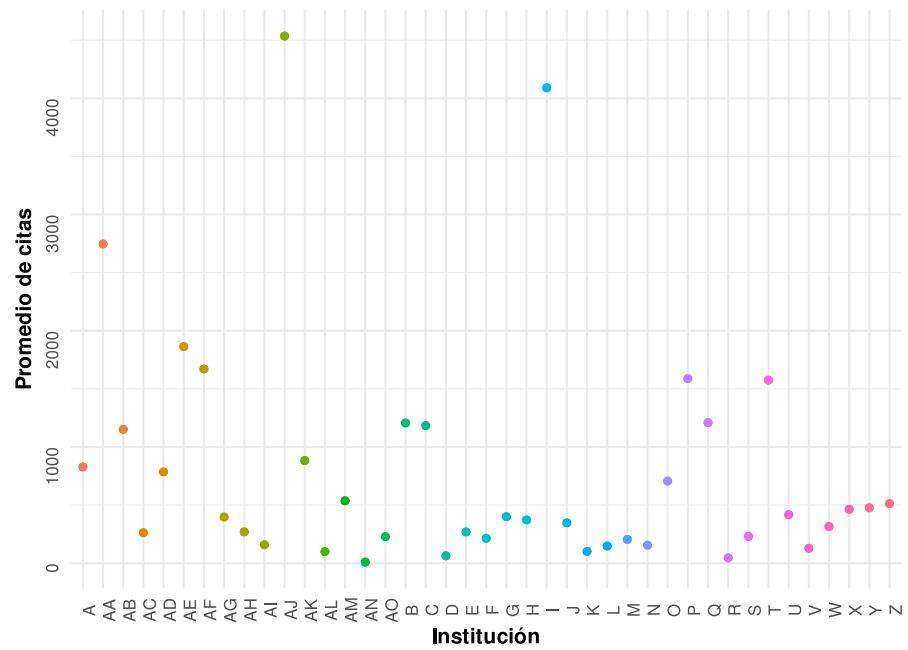
Para tener certeza estadística sobre la aportación de mujeres en artículos y citas dentro de cada departamento, ordenaremos los datos como se muestran en la Tabla 3.27.

Con estas cuatro variables realizaremos una matriz de correlación. Para consultar un ejemplo similar, le sugerimos al lector visitar el video [68]. Para este análisis se utilizará el coeficiente de correlación de Spearman. Utilizamos la correlación de Spearman porque tanto la variable “Citas” como “Artículos” presentan valores extremos. Observamos esto en las Figuras 3.31 y 3.32, los cuales afectarían mucho al coeficiente de correlación de Pearson (código R en A.25).

### 3.6 Correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento

Institución	% Femenino	% Masculino	Promedio de citas	Promedio de artículos
A	0.14	0.86	828.02	28.53
D	0.00	1.00	63.50	18.00
E	0.44	0.56	268.62	39.05
F	0.40	0.60	214.33	24.86
G	0.39	0.61	400.06	29.00
H	0.27	0.73	372.22	38.89
J	0.17	0.83	347.19	39.81
K	0.00	1.00	101.00	17.00
L	0.25	0.75	147.62	16.75
M	0.00	1.00	204.00	44.75
N	0.40	0.60	154.56	26.89
R	0.33	0.67	45.50	9.00
S	0.54	0.46	230.36	28.09

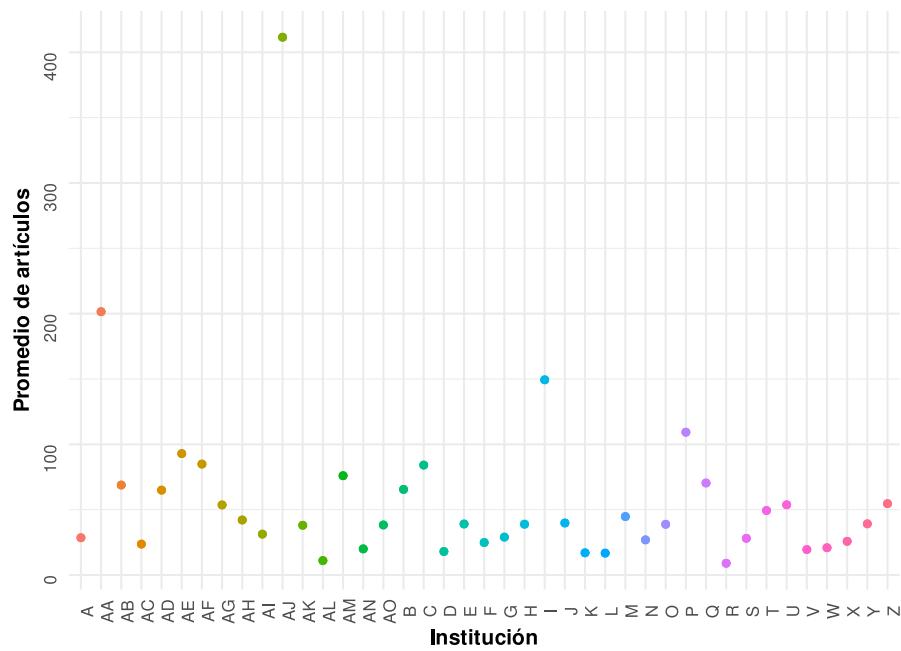
**Tabla 3.27:** Muestra de instituciones con sus promedios de citas y artículos y sus porcentajes de mujeres y hombres respectivamente.



**Fig. 3.31:** Promedio de citas por universidad.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---



**Fig. 3.32:** Promedio de artículos por universidad.

Realizamos la matriz de correlación de Spearman. Los resultados se muestran en la Figura 3.33 (véase la Subsección 3.1.8 y el código R en A.11).

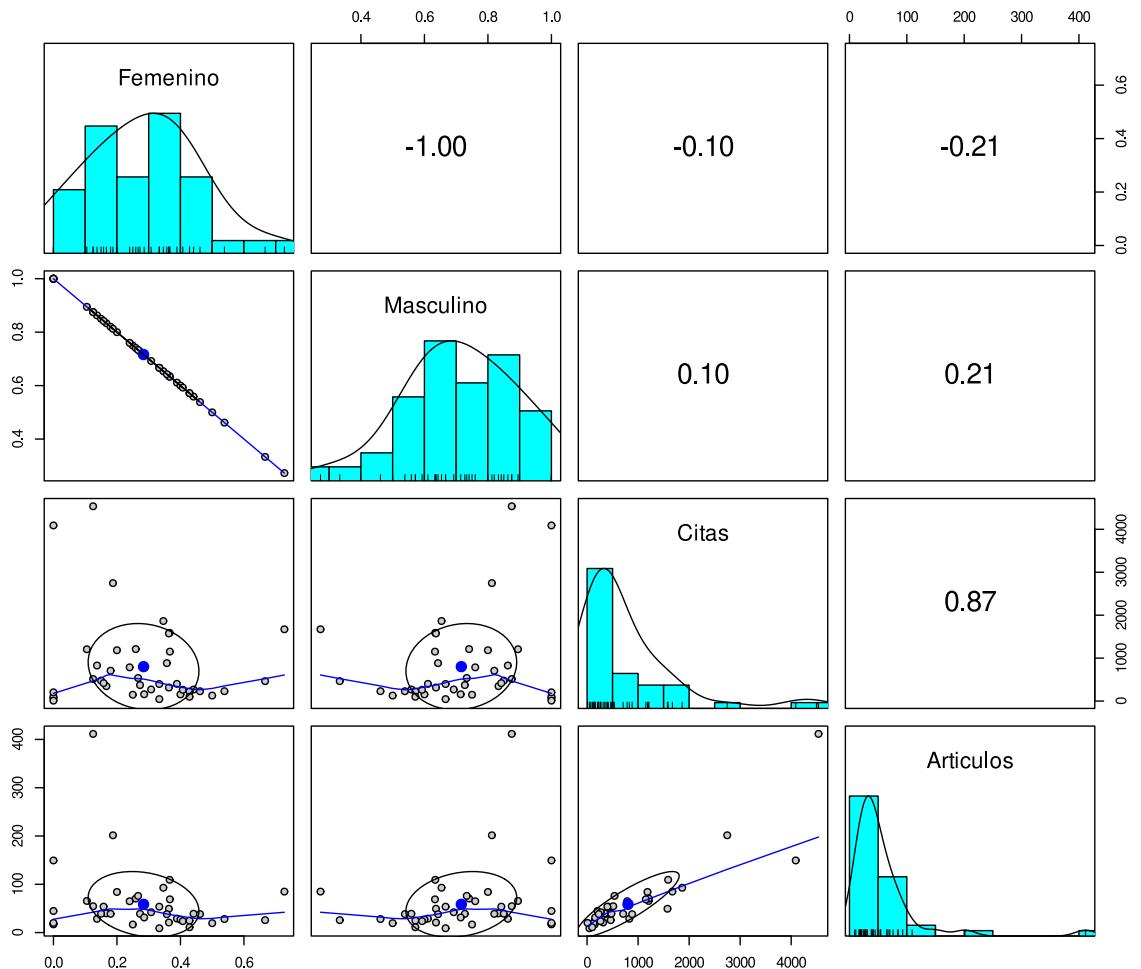
En la diagonal de la Figura 3.33 encontramos el nombre de la variable, en la parte inferior a la diagonal se encuentran las gráficas de cada correlación y en la parte superior a la diagonal aparece el coeficiente de correlación de Spearman.

En el primer caso, la Figura 3.33 nos muestra la correlación entre el porcentaje de mujeres y hombres dentro de los departamentos. Este coeficiente representa a una correlación perfectamente negativa. Lo cual tiene todo el sentido, ya que entre mayor sea el porcentaje de hombres, menor será el porcentaje de mujeres y viceversa.

En el siguiente recuadro de la Figura 3.33 nos encontramos el coeficiente de correlación entre la variable “Femenino” y “Citas”, con un valor de -0.10, caso contrario al coeficiente que resulta en la correlación de “Masculino” con “Citas”, el cual tiene un valor de 0.10. Tanto el coeficiente de las mujeres con la variable citas como el de hombres con citas es bastante cercano a cero, por lo que concluimos que la correlación del género con la variable citas es escasa o nula.

Al continuar con el análisis de los datos de nuestra matriz de correlación, se observa una correlación negativa débil de -0.21 entre el porcentaje de mujeres y el promedio de artículos por departamento; mientras que en el caso del porcentaje de hombres, la correlación es ligeramente positiva con un coeficiente de 0.21. Por lo que parece existir una ligera correlación entre el número de mujeres u hombres dentro de un departamento con la producción de artículos académicos.

### 3.6 Correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento



**Fig. 3.33:** Matriz de correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

---

En el último recuadro observamos el coeficiente de correlación entre las variables “Citas” y “Artículos”, el cual es de 0.87. Es decir, que existe una correlación positiva fuerte entre estas dos variables. Por lo que concluimos que entre más artículos produzca un departamento, más citas recibirá.

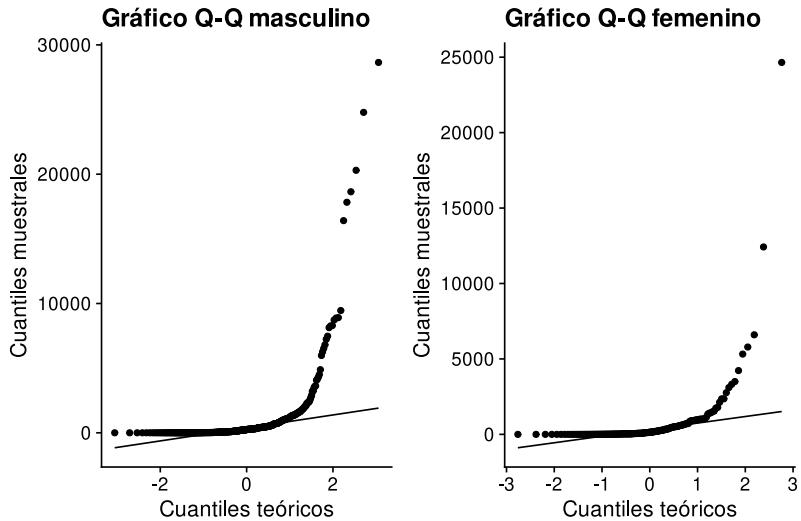
### 3.7. Diferencias significativas entre los datos de hombres y mujeres

El objetivo de esta sección es determinar si existe una diferencia estadísticamente significativa entre el número de citas que reciben los investigadores masculinos y femeninos. Aunque en los análisis anteriores se agruparon a los investigadores masculinos y femeninos por departamento, en estos análisis podremos hacer uso de todos los datos numéricos que tenemos disponibles.

Comprobamos la normalidad de la variable “Citas” para ambos géneros mediante la prueba Shapiro y el gráfico Q-Q. Los resultados se muestran en la Tabla 3.28 y en la Figura 3.34 respectivamente (véanse las Subsecciones 3.1.3 y 3.1.2, y los códigos R en A.27 y A.19). Notamos que no se distribuye de una manera normal, por lo que optaremos por hacer una prueba no paramétrica.

**Tabla 3.28:** Resultado de la prueba Shapiro para la variable “Citas” en ambos géneros.

	Masculino		Femenino	
	Statistic	p-value	Statistic	p-value
Test	0.342	<2.2e-16	0.302	<2.2e-16



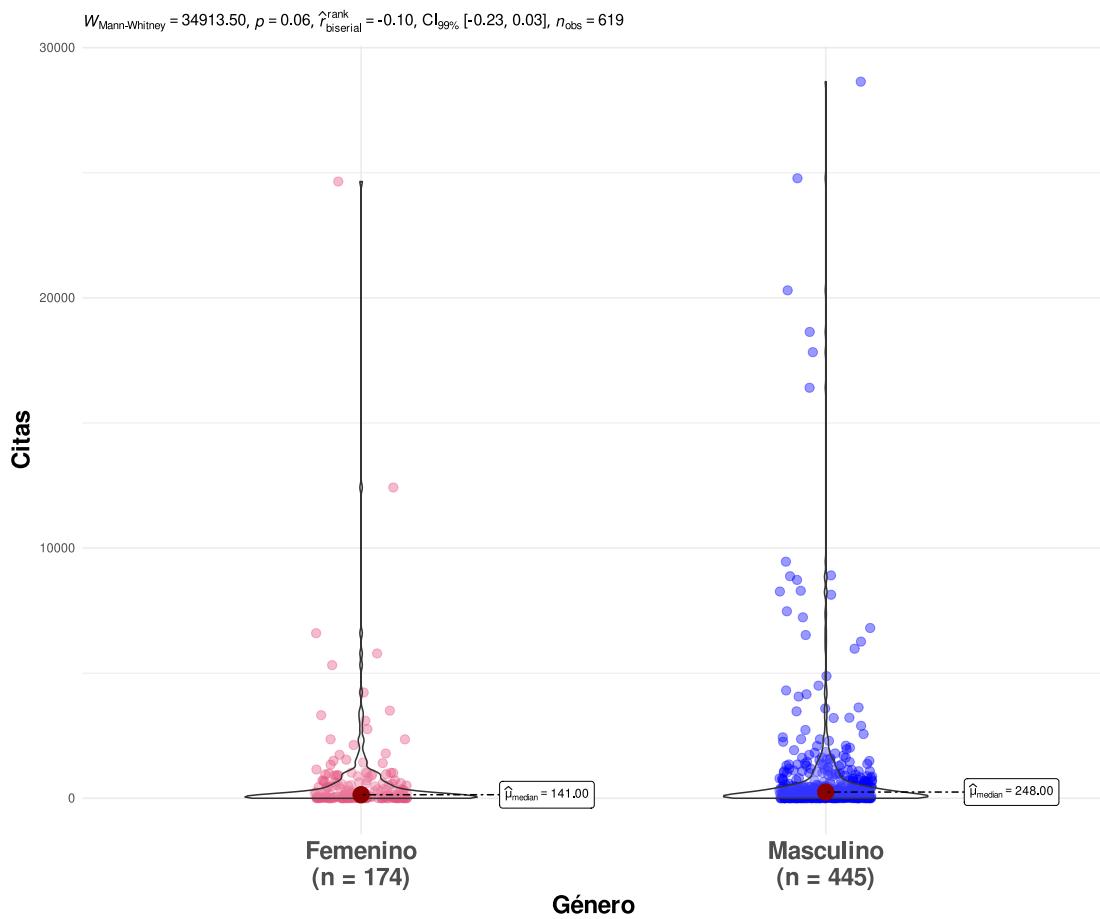
**Fig. 3.34:** Gráfica Q-Q de citas.

Obtenemos algunas estadísticas de resumen de la variable “Citas” (código R en A.10).

### 3.7 Diferencias significativas entre los datos de hombres y mujeres

Género	Recuento	Media	Mediana	Desviación estándar	Rango intercuártil
Femenino	174	734	141	2267	586
Masculino	445	985	248	2806	677

Realizamos una prueba W-Mann-Whitney para conocer si existe una diferencia estadísticamente significativa entre el número de citas y el género de los investigadores. Para consultar un ejemplo similar, le sugerimos al lector visitar el vídeo [65]. Los resultados se muestran en la Figura 3.35 (véase la Subsección 3.1.5 y el código R en A.9).



**Fig. 3.35:** Comparación de citas de investigadores masculinos y femeninos utilizando la Prueba W-Mann-Whitney.

Obtenemos un  $p-value = 0.06$  por lo que no rechazamos la hipótesis nula, es decir que no contamos con evidencia estadística suficiente para concluir que exista una diferencia significativa entre las medianas.

### 3. ANÁLISIS Y REPRESENTACIÓN DE LOS DATOS

#### **Artículos**

Como la variable “Artículos” tampoco se distribuye normalmente también realizaremos una prueba W-Mann-Whitney. Para consultar un ejemplo similar, le sugerimos al lector visitar el vídeo [65].

El resumen estadístico de la variable “Artículos” se muestra a continuación (código R en [A.10](#)).

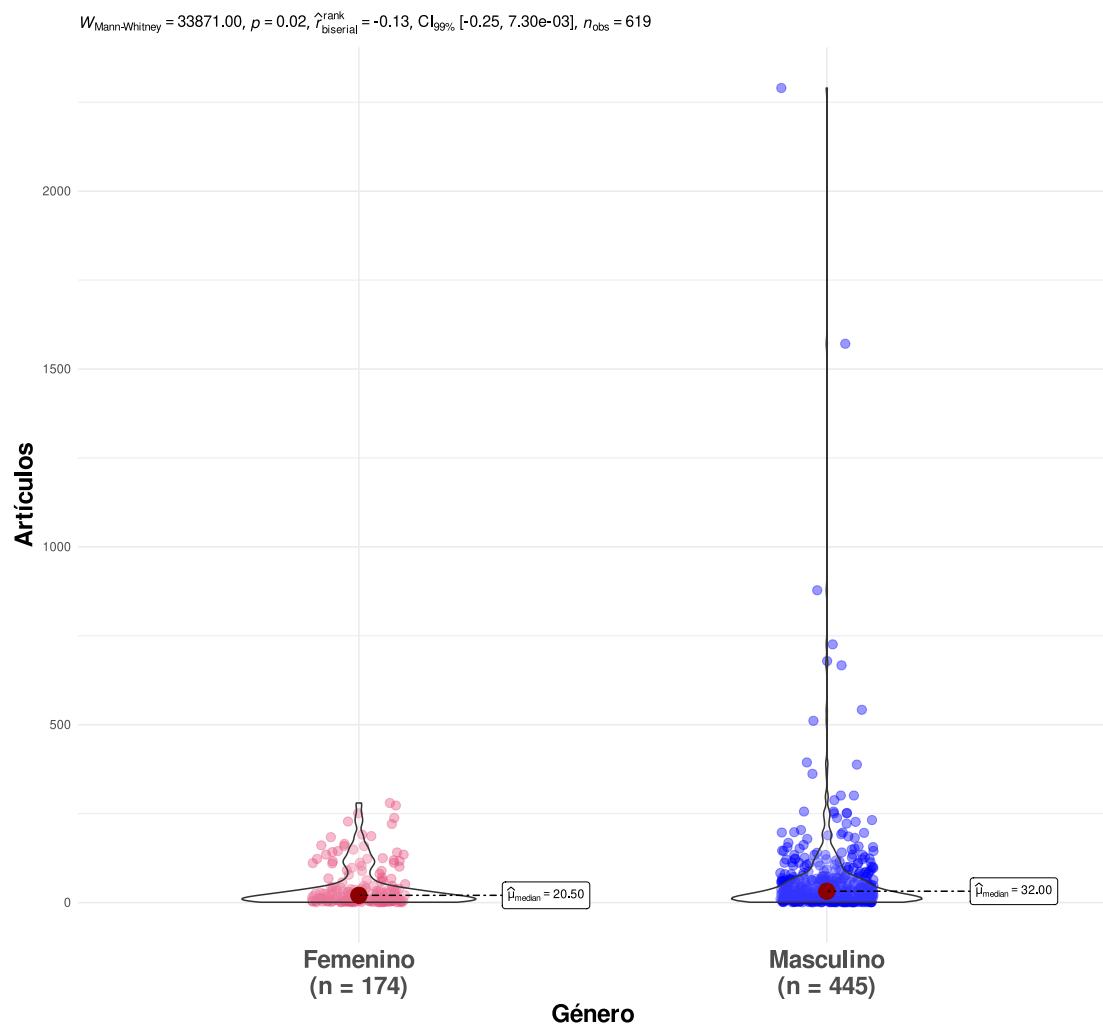
Género	Recuento	Media	Mediana	Desviación estándar	Rango intercuártil
Femenino	174	45.9	20.5	60.2	50.5
Masculino	445	68.9	32	159	59

Realizamos una prueba W-Mann-Whitney (véase la Subsección [3.1.5](#) y el código R en [A.9](#)) para conocer si existe una diferencia estadísticamente significativa entre el número de artículos y el género de los investigadores. Los resultados se muestran en la Figura [3.36](#) (código R en [A.9](#)).

Obtenemos un  $p - value = 0.02$ . Por lo tanto, concluimos que existe una diferencia estadísticamente significativa entre el número de artículos de hombres y mujeres.

### 3.7 Diferencias significativas entre los datos de hombres y mujeres

---



**Fig. 3.36:** Comparación de artículos de investigadores masculinos y femeninos utilizando la Prueba W-Mann-Whitney.



---

## Capítulo 4

# Resultados, alcances y conclusiones

---

En este capítulo se presentan los resultados de las pruebas estadísticas realizadas para evaluar la brecha de género en la investigación académica en las áreas de probabilidad y estadística. Además, se discuten las limitaciones que deben ser consideradas al interpretar los hallazgos, así como las conclusiones obtenidas a partir de resultados y las limitaciones discutidas.

### 4.1. Resultados

#### Grado académico

En la Sección 3.2 se empleó el método bootstrap para calcular intervalos de confianza de la proporción de cada grado académico presente en los departamentos de todos los países de nuestra base de datos. Adicionalmente, se calcularon los intervalos de confianza para México y Brasil, tanto en conjunto como de manera independiente.

Los resultados obtenidos indican que entre el 78 % y 83 % de los investigadores (incluidos en nuestra base de datos) en las áreas de probabilidad o estadística tienen el grado académico de doctorado.

Para verificar si existe una posible relación entre el género y el grado académico, se realizó una prueba de independencia en la Subsección 3.2.1 “Dependencia del grado doctorado con el género (inferido)”. El resultado de la prueba indica que existe una asociación negativa entre ser mujer y tener un doctorado, ya que la diferencia entre las proporciones de hombres y mujeres con doctorado observada es mayor de lo que se esperaría si las variables fueran independientes. Este resultado es consistente con los estudios de la UNESCO sobre la brecha de género en el doctorado [16].

#### Análisis de género en departamentos

En la Sección 3.3 se realizó un análisis para evaluar la proporción de hombres y mujeres en los departamentos de los países incluidos en nuestra base de datos. En la Subsección 3.3.1

## 4. RESULTADOS, ALCANCES Y CONCLUSIONES

---

“Diferencias entre la cantidad de hombres y mujeres en departamentos” se utilizó una prueba W-Mann-Whitney para comparar el número investigadores e investigadoras en cada departamento. En los resultados se obtuvo un  $p - value = 5.93e - 06$  lo que indica que existe una diferencia estadísticamente significativa en el número de mujeres y hombres por departamento.

Además, se realizó una prueba Wilcoxon para comparar la mediana obtenida de los porcentajes de mujeres dentro de los departamentos con un valor hipotético del 50 %, que representa el porcentaje esperado si no existiera desigualdad en la distribución de hombres y mujeres. El resultado muestra que existe una diferencia estadísticamente significativa entre la mediana (30.4 %) de los porcentajes de mujeres observados dentro de los departamentos y el 50 % esperado. Estos resultados son consistentes con los estudios de la UNESCO sobre la brecha de género en la investigación académica [1], la cual es más grande en las áreas STEM [3].

Finalmente, en la Subsección 3.3.2 “Porcentaje de investigadoras por país” se realizó un análisis de los datos para evaluar la proporción de mujeres en cada país. Los resultados indican que México es el segundo país con menor proporción de mujeres en los departamentos de estadística o probabilidad incluidos en nuestra base de datos.

### Independencia del género con las citas recibidas

En la Sección 3.4 se utilizó el método bootstrap para determinar si el género del investigador tiene alguna relación con pertenecer a la categoría de citas más alta, definida como el cuartil más alto(Q3) del número de citas respecto al total de investigadores en nuestra base de datos. El resultado indica que el género y el número de citas son independientes, ya que la diferencia observada de 0.028 es muy similar a la que se esperaría si las variables fueran independientes. Esto sugiere que el género no tiene una influencia significativa en el número de citas recibidas cuando éste es alto.

### Clústeres

En la Sección 3.5 se utilizó el método *kmeans* para agrupar a las universidades según sus promedios de artículos y citas en dos grupos. El primero compuesto por universidades que tienen un número alto de artículos y citas en promedio y el segundo grupo compuesto por universidades con un menor promedio de artículos y citas. Se realizó un análisis comparativo utilizando la prueba W-Mann-Whitney para determinar si existe una diferencia significativa en el porcentaje de mujeres en ambos grupos de universidades. Los resultados no mostraron una diferencia significativa en el porcentaje de mujeres entre los dos grupos.

En la Subsección 3.5.1 “Clústeres con únicamente datos de mujeres” se filtraron los datos para utilizar solamente información de mujeres en la creación de los clústeres. Se observó que la mayoría de universidades se agruparon en el clúster de universidades con un promedio alto de artículos y citas. Para verificar qué sucedía al incluir datos de hombres, en la Subsección 3.5.2 “Comparación de promedios en artículos y citas por departamentos divididos por género” se incorporaron datos masculinos y se realizó una comparación. Los resultados mostraron que, en general, los investigadores masculinos tenían una mayor producción de artículos, mientras que el

## 4.2 Alcance de los resultados

---

número de citas era similar entre hombres y mujeres.

### **Correlación entre el porcentaje de mujeres y el promedio de citas y artículos por departamento**

En la Sección 3.6 se calculó una correlación de Spearman para evaluar si existe alguna relación entre el porcentaje de hombres y mujeres en los departamentos y el promedio de artículos y citas que tienen. Los resultados indicaron que no se encontró una correlación significativa entre el porcentaje de hombres y mujeres en los departamentos y el promedio citas que tienen. No obstante, se encontró que el promedio de artículos estaba negativamente correlacionado con el porcentaje de mujeres dentro de los departamentos.

### **Diferencias significativas entre los datos de hombres y mujeres**

En la Sección 3.7 se realizaron pruebas W-Mann-Whitney para evaluar si existen diferencias estadísticamente significativas en el número de citas y artículos entre hombres y mujeres. Los resultados indicaron que no existe una diferencia estadísticamente significativa en el número de citas entre hombres y mujeres. Sin embargo, se encontró una diferencia significativa en el número de artículos entre hombres y mujeres.

## **4.2. Alcance de los resultados**

A continuación, se exponen algunas limitaciones que deben ser consideradas al interpretar los resultados de este trabajo.

- a) La información de la base de datos está sesgada. Para la base de datos, solo se utilizó información disponible en Internet, por lo que se debe tener en cuenta que la información puede estar incompleta o desactualizada. Además, es posible que algunas instituciones no tengan presencia en línea o no hayan sido incluidas en la base de datos. Por lo tanto, los resultados obtenidos están sesgados y no son representativos de la totalidad de la población académica en las áreas de probabilidad y estadística.
- b) La disponibilidad de información detallada sobre el personal docente puede variar según la institución y algunas instituciones optan por no proporcionar información completa sobre su personal. Además, se debe considerar que la falta de uniformidad en el tipo de nombramiento o la categoría que reciben los académicos, según su institución, puede afectar los resultados de esta investigación.
- c) La base de datos utilizada en esta tesis se recopiló en un período específico de tiempo y no refleja necesariamente la situación actual o futura. Además, debido a la pandemia de COVID-19, es posible que se hayan producido cambios significativos en la investigación académica en las áreas de probabilidad y estadística. Sería valioso continuar monitoreando

## **4. RESULTADOS, ALCANCES Y CONCLUSIONES**

---

la evolución de esta brecha en el futuro y analizar el impacto que eventos externos, como la pandemia, pueden tener en la dinámica de género en la investigación.

- d) El género de los investigadores y las investigadoras incluidos en la base de datos se infirió a partir de las convenciones culturales en Latinoamérica y al uso del género gramatical en Español y Portugués como “investigador/investigadora” y “pesquisador/pesquisadora”. Esta metodología limita la inclusión de disidencias de género.

### **4.3. Conclusiones**

La brecha de género presente en la investigación académica es un tema complejo que varía dependiendo del área de estudio. Aunque en algunas áreas de investigación, como en la psicología, esta brecha se ha ido cerrando de manera efectiva [2], en el mundo de la investigación en las ciencias exactas, como matemáticas, aún queda mucho por hacer.

En nuestro estudio nos pudimos concientizar que, según los registros en nuestra base de datos, existe una diferencia estadísticamente significativa entre el número de mujeres y hombres que realizan investigación en probabilidad o estadística (véase la Sección 3.3). Además, se observa una correlación negativa entre el porcentaje de mujeres y la producción de artículos de un departamento (véase la Sección 3.6). Esto conlleva consecuencias negativas en la producción de conocimiento y en la búsqueda de soluciones a problemas, debido a que la brecha de género limita que se tomen en cuenta a todas las perspectivas en la generación de ideas y soluciones. Por otro lado, es importante destacar que encontramos una correlación nula o escasa entre el porcentaje de mujeres en un departamento y la cantidad de citas recibidas en sus publicaciones (véase la Sección 3.6). Esto sugiere que la calidad de publicaciones entre hombres y mujeres es similar y que el género del autor no afecta la calidad o importancia de la investigación. Los resultados del análisis realizado en la Sección 3.4 respaldan nuestra conclusión anterior. En este caso, al agrupar a los investigadores con “Alto” y “No alto” número de citas, el género de los autores era independiente de tener un número alto de citas. En resumen, el éxito en términos de citas de una publicación no está relacionado con el género del autor, sino más bien con la calidad del trabajo en sí.

La igualdad de género es clave para el progreso de la ciencia y la generación de bienestar para toda la sociedad. Pese a que en los últimos años se ha realizado un avance importante para cerrar la brecha de género, es importante que se siga estudiando y documentando la situación para identificar los problemas y brindar soluciones adecuadas, con el fin de avanzar hacia una sociedad más igualitaria.

En complemento, la participación activa de las universidades, instituciones académicas, sociedad y los propios investigadores es esencial para lograr una reducción significativa de la brecha de género en la investigación académica.

Una forma de participación activa para abordar este problema es considerar alternativas a la evaluación bibliométrica de citas y artículos. Debido a que este método de evaluación puede mostrar parcialidad hacia el género masculino, ya que se basan en la cantidad de publicaciones y citas. Lo que resulta perjudicial para las mujeres, dado que a menudo enfrentan obstáculos

#### 4.3 Conclusiones

adicionales al publicar, como las dobles jornadas de trabajo debido a las responsabilidades familiares, la falta de esquemas institucionales que otorguen apoyos, entre otras [19].

Por lo tanto, es importante examinar estas medidas y considerar alternativas que reconozcan la calidad y el impacto real del trabajo académico, para garantizar que se valore adecuadamente el trabajo de las mujeres investigadoras.



---

## Apéndice A

# Código R

---

Los paquetes utilizados para en esta tesis se muestran en el apéndice A.1. Para consultar más detalles sobre la función utilizada nos referimos a [69].

```
1 #Funcion ipak: instala y carga multiples paquetes R.
2 #Comprueba si los paquetes estan instalados. Los instala si no lo estan, luego
  los carga en la sesion R.
3 ipak <- function(pkg){
4   new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
5   if (length(new.pkg))
6     install.packages(new.pkg, dependencies = TRUE)
7   sapply(pkg, require, character.only = TRUE)
8 }
9
10 #Uso
11 packages <- c("ggpubr", "factoextra", "NbClust", "tidyR", "foreign", "apaTables", "PerformanceAnalytics", "psych", "corrr", "rstatix", "haven", "apa", "dplyr", "BayesFactor", "ggstatsplot", "gvlma", "car", "ggfortify", "lmtest", "rvg", "reshape2", "kableExtra", "xtable", "purrr", "maps", "igraph", "cowplot", "ggpubr", "cluster", "gmodels", "broom", "ggplot2", "dplyr", "magrittr", "tidyverse", "stringr", "forcats", "readxl", "openxlsx", "infer", "tidyR", "nortest", "readr")
12
13 ipak(packages)
```

Código A.1: Función para cargar paquetes.

```
1 data_frame %>% filter(Pais == "Pais1") %>% group_by(Universidad) %>% count()
```

Código A.2: Desglose por país.

```
1 data_frame %>% filter(Universidad=="Universidad1") %>% group_by(Estado) %>%
  tally()
```

Código A.3: Desglose por sedes.

```
1 data_frame %>% group_by(Pais) %>% count()
```

Código A.4: Desglose por países.

```
1 data_frame %>% filter(!is.na(Citas) & !is.na(Articulos) & Pais == "Pais1") %>%
  group_by(Universidad) %>% count()
```

Código A.5: Universidades con datos numéricos.

## A. CÓDIGO R

---

```
1 # Seleccionamos una columna especifica del data frame y la usamos como respuesta
  en una especificacion
2 data_frame %>%
3   specify(response = variable_resposta, success = "Valor_exitoso") %>%
4   # Generamos replicaciones usando bootstrap
5   generate(reps = 500, type = "bootstrap") %>%
6   # Calculamos la estadistica de interes (en este caso, la proporcion)
7   calculate(stat = "prop")
8
9 # Creamos una grafica de densidad para la estadistica calculada anteriormente
10 ggplot(data_frame, aes(x = stat)) +
11   geom_density(fill = "blue", alpha = 0.2) +
12   theme_minimal() +
13   labs(title = "Grafica de densidad", x = "Estadistica de interes", y =
14     "Densidad") +
15   theme(axis.title = element_text(face = "bold"))
16
17 # Calculamos la desviacion estandar de la estadistica
18 SE <- data_frame %>%
19   summarize(sd(stat)) %>%
20   pull()
21
22 # Imprimimos la desviacion estandar calculada
23 SE
24
25 # Calculamos el intervalo de confianza de la estadistica
26 c(stat - 2 * SE, stat + 2 * SE)
```

Código A.6: BootStrap para intervalo de confianza.

```
1 # Filtramos las observaciones del data frame que cumplan con ciertas condiciones
2 data_frame <- data_frame %>% filter(variable1 == valor1 | variable2 == valor2 |
  variable3 == valor3)
3
4 # Creamos un nuevo data frame a partir de columnas especificas del data frame
  original
5 nuevo_data_frame <- data.frame(origen = data_frame$columna1, destino = data_
  frame$columna2)
6
7 # Agrupamos el data frame por una columna especifica y contamos el numero de
  observaciones en cada grupo
8 data_frame %>%
9   group_by(columna_a_agrupar) %>%
10  count()
11
12 # Creamos un grafo a partir de un data frame con dos columnas que representan
  los nodos y los arcos
13 grafo <- graph_from_data_frame(data_frame, directed = TRUE)
14
15 # Agrupamos el data frame por dos columnas especificas y contamos el numero de
  observaciones en cada grupo
16 data_frame_agrupado <- data_frame %>%
17   group_by(columna1, columna2) %>%
18   summarize(cantidad = n())
19
20 # Creamos una secuencia de colores para los nodos del grafo
21 colores <- rainbow(nrow(data_frame_agrupado))
```

---

```

23 # Filtramos observaciones del data frame que cumplan con ciertas condiciones
24 nuevo_data_frame <- data_frame %>% filter(columna1 == valor1 & columna2 > valor2
    )
25
26 # Creamos un plot del grafo con un layout especifico y asignamos colores a los
    nodos segun su grupo
27 plot(grafo, layout = layout.kamada.kawai, vertex.color = colores[match(V(grafo)$
    name, data_frame_agrupado$columna1)])

```

Código A.7: Red de universidades.

```

1 # Crear el vector con los nombres
2 nombres <- c("C.F Gauss", "H. Minkowski", "H. Weyl", "A. Einstein", "H.A.
    Lorentz", "E.G. Straus", "J. Spencer", "L. Lovasz", "E. Szemerédi", "P.
    Erdős", "S.A. Burr", "R.L. Graham", "F. Chung", "W.T. Tutte", "W. Feller")
3
4 # Crear la matriz de adyacencia
5 matriz_ady <- matrix(0, ncol = length(nombres), nrow = length(nombres), dimnames
    = list(nombres, nombres))
6
7 # Establecer las conexiones
8 matriz_ady["W. Feller", "P. Erdős"] <- 1
9 matriz_ady["P. Erdős", c("S.A. Burr", "L. Lovasz", "J. Spencer", "E. Szemerédi",
    "R.L. Graham", "E.G. Straus", "W.T. Tutte", "W. Feller")] <- 1
10 matriz_ady["S.A. Burr", c("J. Spencer", "L. Lovasz", "R.L. Graham")] <- 1
11 matriz_ady["L. Lovasz", c("E.G. Straus", "J. Spencer", "R.L. Graham")] <- 1
12 matriz_ady["J. Spencer", c("E.G. Straus", "F. Chung", "R.L. Graham", "E.
    Szemerédi")] <- 1
13 matriz_ady["E. Szemerédi", c("F. Chung", "R.L. Graham")] <- 1
14 matriz_ady["R.L. Graham", c("E.G. Straus", "F. Chung")] <- 1
15 matriz_ady["E.G. Straus", "A. Einstein"] <- 1
16 matriz_ady["A. Einstein", c("H.A. Lorentz", "H. Minkowski", "H. Weyl")] <- 1
17 matriz_ady["H. Minkowski", c("H.A. Lorentz", "C.F Gauss", "H. Weyl")] <- 1
18 matriz_ady["H. Weyl", "H.A. Lorentz"] <- 1
19
20 # Crear el grafo
21 g <- graph.adjacency(matriz_ady, mode = "undirected", diag = FALSE)
22
23 # Establecer colores para los nodos
24 colores <- rainbow(length(V(g)))
25
26 # Definir la disposición de los nodos
27 layout <- layout_with_kk(g) #mds, kk o sugiyama
28
29 # Personalizar la visualización de la red
30 plot(g, layout = layout, vertex.size = 10, vertex.label.cex = 0.8, vertex.color
    = colores, vertex.frame.color = "white", vertex.label.color = "black", edge.
    color = "gray", edge.width = 1.5, main = "Red de Erdős")

```

Código A.8: Red de Erdős.

```

1 ggstatsplot::ggbetweenstats(
2   data = my_data,                      # Data frame que contiene los datos
3   x = my_group_var,                   # Nombre de la variable de agrupación o
    variable independiente
4   y = my_dependent_var,               # Nombre de la variable dependiente
5   xlab = "Variable X",                # Etiqueta para el eje X
6   ylab = "Variable Y",                # Etiqueta para el eje Y

```

## A. CÓDIGO R

---

```
7 type = "np",                      # Tipo de prueba estadistica ("p" para prueba
8   parametrica, "np" para prueba no parametrica, "r" para prueba robusta, "bf"
9   para factor bayesiano)
10 effsize.type = "g",                # Tipo de estimador de efecto ("d", "g", "r",
11   o la estimacion robusta por defecto)
12 conf.level = 0.99,                 # Nivel de confianza para intervalos de
13   confianza
14 plot.type = "violin",              # Tipo de grafico a generar ("box" para
15   diagrama de caja, "violin" para diagrama de violin)
16 outlier.tagging = TRUE,            # Indica si se debe etiquetar los valores
17   atipicos
18 outlier.coef = 1.5,                # Coeficiente para determinar los valores
19   atipicos (siguiendo la ley de Tukey)
20 outlier.label.args = list(color = "red"), # Argumentos para personalizar la
21   etiqueta de los valores atipicos
22 messages = FALSE,                  # Si se deben mostrar los mensajes de progreso
23 ggtheme = ggplot2::theme_gray(), # Tema de la grafica (consulte https://ggplot2
24   .tidyverse.org/reference/ggtheme.html para ver todos los temas disponibles)
25 package = "yarrr",                # Paquete de colores para utilizar (por
26   defecto "ggplot2")
27 palette = "info2",                # Paleta de colores dentro del paquete
28   especificado
29 title = "Prueba estadistica",     # Titulo de la grafica
30 caption = "Esta es una visualizacion de datos", # Leyenda de la grafica
31 + scale_color_manual(values=c("#E75480", "blue")) #colores de los puntos
32 )+
33   scale_color_manual(values=c("#E75480", "blue"))+
34   theme(axis.text.x = element_text(face = "bold"))+
35   theme(axis.title = element_text(face = "bold"))
```

Código A.9: Prueba Wilcoxon-Mann-Whitney con gráfica.

```
1 # Filtrar la base de datos y resumir por sexo
2 df_resumen <- data_frame %>%
3   filter(Pais == "PaisA") %>%
4   group_by(Sexo) %>%
5   summarise(
6     N = n(), # Recuento
7     Mean = mean(VariableA), # Media
8     Median = median(VariableA), # Mediana
9     SD = sd(VariableA), # Desviacion estandar
10    IQR = IQR(VariableA) # Rango intercuartil
11  )
```

Código A.10: Resumen estadístico.

```
1 pairs.panels(data_frame, #Base de datos
2 method = "spearman", # Metodo de correlacion utilizado
3 cex.cor = 1, # Tamaño de la letra para los coeficientes de correlacion
4 col = "blue", #Color de los puntos y la linea de tendencia
5 pch = 21, #Tipo de simbolo utilizado en la grafica
6 bg = "gray80") #Color del fondo de los puntos
```

Código A.11: Correlación.

```
1 datos <- data_frame %>%
2   mutate(Variable1 = log10(Variable1), # aplicar logaritmo a Variable1
3         Variable2 = log10(Variable2)) %>% # aplicar logaritmo a Variable2
4   scale(center = TRUE, scale = TRUE) # centralizar los datos en 0 y escalarlos
```

```

5 # calcular la matriz de distancia utilizando el metodo euclidian
6 matriz.distancia <- get_dist(datos, method = "euclidean")
7
8 # visualizar la matriz de distancia utilizando un gradiente de color
9 fviz_dist(matriz.distancia, gradient = list(low = "black", mid = "white", high =
10   "red"))
11
12 # determinar el numero ptimo de clusters utilizando el metodo silhouette
13 fviz_nbclust(datos, kmeans, method = "silhouette")
14
15 # determinar el numero ptimo de clusters utilizando el metodo within-cluster
16   sum of squares
17 fviz_nbclust(datos, kmeans, method = "wss")
18
19 # determinar el numero ptimo de clusters utilizando el metodo gap statistic
20 fviz_nbclust(datos, kmeans, method = "gap_stat")
21
22 # determinar el numero ptimo de clusters utilizando 30 indices diferentes
23 res.num.clust <- NbClust(datos, distance = "euclidean", min.nc = 2, max.nc = 10,
24   method = "kmeans", index = "alllong")
25
26 # realizar el algoritmo de clustering k-means con 2 clusters
27 k2 <- kmeans(datos, centers = 2, nstart = 25)
28
29 # mostrar los resultados del algoritmo k-means
30 k2
31
32 # visualizar los clusters utilizando un diagrama de dispersion y una ellipse
33 fviz_cluster(k2, data = datos, ellipse.type = "euclid", repel = TRUE, star.plot
34   = TRUE, main = "Cluster name")+
35   theme(axis.title = element_text(size = 17, face = "bold"))+
36   guides(fill = guide_legend(label.theme = element_text(size = 24, face = "bold"
37     )))

```

Código A.12: Clústeres.

```

1 # Crear un grafico de barras
2 ggplot(datos, aes(y = fct_rev(fct_infreq(Variable1)), fill = Variable1)) +
3   geom_bar() +
4   coord_flip() +
5   ylab("Variable1") + # Etiqueta del eje y
6   xlab("Variable2") + # Etiqueta del eje x
7   theme(axis.title.x=element_blank(), # Quitar la etiqueta del eje x
8         axis.text.x=element_blank(), # Quitar los valores del eje x
9         axis.ticks.x=element_blank())+ # Quitar las marcas del eje x
10  theme(axis.text.x = element_text(face = "bold"))+
11  theme(axis.title = element_text(face = "bold"))

```

Código A.13: Gráfico de barras.

```

1 # Aplicacion de la funcion CrossTable a dos variables
2 # con el fin de obtener una tabla cruzada de frecuencias
3 # donde se muestre la distribucion de los niveles de un
4 # atributo segun otro atributo.
5 CrossTable(data_frame$Variable1, data_frame$Variable2)

```

Código A.14: Tabla cruzada.

## A. CÓDIGO R

---

```
1 # Especificar las variables a analizar
2 variable1 <- data.frame(columna1)
3 variable2 <- data.frame(columna2)
4
5 # Calcular las proporciones
6 p_hats <- variable1 %>%
7   group_by(variable2) %>%
8   summarize(prop = mean(columna1 == "valor")) %>%
9   pull()
10
11 # Calcular la diferencia de proporciones
12 d_hat <- diff(p_hats)
13
14 # Crear la hipótesis nula de independencia
15 H_null <- variable1 %>%
16   specify(columna1 ~ variable2, success = "valor") %>%
17   hypothesize(null = "independence") %>%
18   generate(reps = 500, type = "permute") %>%
19   calculate(stat = "diff in props", order = c("valor1", "valor2"))
20
21
22 # Especificar el nivel de significancia (alpha)
23 alpha <- 0.05
24
25 # Calcular los límites del intervalo de confianza
26 lower <- H_null %>%
27   summarize(l = quantile(stat, probs = alpha / 2)) %>%
28   pull()
29 upper <- H_null %>%
30   summarize(u = quantile(stat, probs = 1 - alpha / 2)) %>%
31   pull()
32
33 # Verificar si la diferencia está dentro del intervalo de confianza
34 is_significant <- d_hat %>% between(lower, upper)
35
36 # Graficar la densidad de la distribución de la hipótesis nula y la diferencia
37   # observada
38 ggplot(H_null, aes(x = stat)) +
39   geom_density(fill = "blue", alpha = 0.2) +
40   geom_vline(xintercept = d_hat, color = "red") +
41   geom_vline(xintercept = lower, color = "blue") +
42   geom_vline(xintercept = upper, color = "blue") +
43   ylab("Densidad") + xlab("Diferencia de proporciones") +
44   theme_minimal() +
45   labs(title = "Grafica de densidad", x = "Diferencia de proporciones", y =
46     "Densidad", color = "") +
47   # Agregar etiquetas a las líneas verticales
48   geom_segment(aes(x = d_hat_dr, y = 0, xend = d_hat_dr, yend = max(H_nullDr$stat),
49     color = "Diferencia \n observada"), size = 1) +
50   geom_segment(aes(x = lower, y = 0, xend = lower, yend = max(H_nullDr$stat),
51     color = "Intervalo de \n confianza \n del 95%"), size = 1) +
52   scale_color_manual(values = c("Diferencia \n observada" = "red",
53     "Intervalo de \n confianza \n del 95%" = "blue")) +
54   theme(axis.title = element_text(face = "bold"))
```

Código A.15: Bootstrap para probar la independencia de dos variables.

```
1 quantile(df$Variable)
```

Código A.16: Cálculo de cuartiles y extremos.

```
1 ggplot(nombre_dataframe, aes(x = variable_X, fill = variable_Y)) +  
2   geom_bar(position = "dodge") + ylab("Nombre del eje Y") +  
3   theme(axis.text.x = element_text(face = "bold")) +  
4   theme(axis.title = element_text(face = "bold")) +  
5   guides(fill = guide_legend(title = "", label.theme = element_text(face = "bold")))
```

Código A.17: Gráfica de barras agrupados por característica.

```
1 ggplot(dataframe, aes(x = variable1, fill = variable2, color = variable2)) +  
2   geom_histogram(alpha = 0.5, position = "identity") +  
3   geom_density(alpha = 0.3) +  
4   ylab("y label") +  
5   theme(axis.text.x = element_text(face = "bold")) +  
6   theme(axis.title = element_text(face = "bold")) +  
7   guides(fill = guide_legend(label.theme = element_text(face = "bold")))
```

Código A.18: Histogramas por característica.

```
1 Data_frame %>%  
2   filter(Variablea == "condicion") %>% ggplot(aes(sample = Variable2)) +  
3   stat_qq() +  
4   stat_qq_line() + ggtitle("Titulo") +  
5   ylab("Cuantiles muestrales") +  
6   xlab("Cuantiles te ricos") +  
7   theme_cowplot()
```

Código A.19: Gráfica QQ.

```
1 plot(yb, main="Grafica de distribucion empirica cumulativa,", xlab="x", ylab="F(x)")
```

Código A.20: Gráfica de la función de distribución acumulada.

```
1 FPA1 <- ggplot(data = clus1, aes(x = Femenino)) +  
2   stat_ecdf(geom = "step", direction = "hv") +  
3   stat_ecdf(geom = "point") +  
4   theme_bw() +  
5   theme(plot.title = element_text(size = 14, face = "bold"),  
6         axis.title = element_text(size = 12)) +  
7   labs(title = "Cluster 1", x = "x", y = "F(x)")  
8  
9 FPA2 <- ggplot(data = clus2, aes(x = Femenino)) +  
10  stat_ecdf(geom = "step", direction = "hv") +  
11  stat_ecdf(geom = "point") +  
12  theme_bw() +  
13  theme(plot.title = element_text(size = 14, face = "bold"),  
14        axis.title = element_text(size = 12)) +  
15  labs(title = "Cluster 2", x = "x", y = "F(x)")  
16  
17 plot_grid(FPA1, FPA2)
```

Código A.21: Gráfica de la función de distribución acumulada para dos grupos.

## A. CÓDIGO R

---

```
1 data_frame %>% ggplot(aes(x= variable1, y= variable2, color = variable1, size  
2   =1)) +  
3   geom_point() +  
4   stat_summary(fun = mean, geom = "point") + scale_color_manual(values = c("red"  
5     , "green", "blue", "brown", "Magenta", "black", "purple")) +  
6   guides(color = FALSE, size = FALSE) + ylab("Porcentaje femenino") + theme_  
7   minimal() +  
8   theme(axis.text.x = element_text(size = 14, face = "bold"))+  
9   theme(axis.title = element_text(size = 14, face = "bold"))
```

Código A.22: Gráfica de dispersión por característica.

```
1 ggplot(data_frame, aes(x="", y=Variable1, fill=Variable1)) +  
2   geom_bar(stat="identity", width=1) +  
3   coord_polar("y", start=0) +  
4   labs(x = NULL) +  
5   facet_wrap(~Variable3)+  
6   xlab(NULL)+  
7   theme(  
8     strip.text = element_text(size = 20, face = "bold")  
9   )+  
10  theme(axis.title = element_text(size = 24, face = "bold"))+  
11  guides(fill = guide_legend(title = "", label.theme = element_text(size = 14,  
    face = "bold")))
```

Código A.23: Gráfica de pastel.

```
1 data_frame %>% ggplot(aes(x= Variable1, y= Variable2, color = Variable3))+  
2   geom_point(aes(fill = Variable3),  
3     alpha = 0.3,  
4     size = 5,  
5     shape = 22,  
6     color = 'black') + labs(x = "", y = "", fill = "Variable3")+ guides  
(color = FALSE, fill = FALSE)
```

Código A.24: Gráfica de dispersión para comparar dos grupos.

```
1 pC <- data_frame %>% ggplot(aes(x= Variable1, y= Variable2, color = Variable1))+  
2   geom_point() +  
3   stat_summary(fun = mean, geom = "point") + ylab("Promedio de citas") +  
4   guides(color = FALSE, size = FALSE)+ theme_minimal()  
5 pC + theme(axis.text = element_text(angle = 90))
```

Código A.25: Gráfica de dispersión.

```
1 # Crear un vector con los datos de las citas  
2 citas <- c(4, 2, 4, 1, 8, 5, 10, 2, 10, 15)  
3  
4 # Obtener el orden de los datos segun el numero de citas de mayor a menor  
5 orden_citas <- order(citas, decreasing = TRUE)  
6  
7 # Crear un vector con las etiquetas personalizadas del eje X  
8 etiquetas_x <- paste0(1:length(citas), " ")  
9  
10 # Crear el grafico de barras  
11 barplot(citas[orden_citas], names.arg = etiquetas_x, ylim = c(0, 15),  
12         ylab = "Numero de citas", xlab = "Articulos ordenados por numero de  
        citas",
```

---

```

13     main = "Como se calcula el indice H", col = "lightgray") + theme_minimal()
14
15 # Ajustar las etiquetas del eje Y para que muestren valores de 2 en 2
16 axis(2, at = seq(0, 15, by = 1))
17
18 # Agregar una linea vertical en x = 8, y = 0 que termine en y = 8, x = 8
19 #lines(c(9, 9), c(0, 8), col = "red")
20 lines(c(5.5, 5.5), c(0, 5), col = "red")
21 # Agregar una linea horizontal en y = 8, x = 0 que termine en x = 8, y = 8
22 lines(c(0, 5.5), c(5, 5), col = "red")

```

Código A.26: Índice h.

```

1 shapiro.test(data_frame$variable)

```

Código A.27: Prueba Shapiro.

```

1 ggplot(data_frame, aes(x = Variable1, y = Variable2, fill = Variable3)) +
2   geom_bar(stat = "identity", position = "dodge") +
3   labs(x = "etiqueta_eje_x",
4        y = "etiqueta_eje_y",
5        fill = "Variable3") +
6   theme_bw() +
7   theme(axis.text.x = element_text(face = "bold")) +
8   theme(axis.title = element_text(face = "bold")) +
9   guides(fill = guide_legend(title = "", label.theme = element_text(face = "bold")))

```

Código A.28: Histograma de frecuencia dividido por característica.

```

1 #Creamos una nueva columna en nuestro data frame
2 data_frame$Valores_teoricos <- rnorm(66, mean = 0.5, sd = 0.2 * abs(qnorm(0.25)))
3 )%>%
4   pmax(0.25) %>% pmin(0.75)
5
6 #Convertimos los datos a formato longitudinal usando la función melt
7 data_frame <- melt(data_frame, id.vars = "Universidad")
8 colnames(data_frame) <- c("Variable1", "Variable2", "Variable3")
9
10 #Creamos un grafico boxplot
11 data_frame %>%
12   ggplot(aes(x=Variable2,y=Variable3, color=Variable2)) +
13   geom_boxplot(outlier.shape=NA) +
14   geom_jitter(width=0.15) +
15   scale_y_continuous(limits=c(0,1), breaks=seq(0.1, 1, 0.1)) +
16   scale_color_manual(values=c("Caracteristica1"="#FF69B4", "Caracteristica2"="black")) +
17   theme(legend.position = "none") +
18   theme(axis.text.x = element_text(size = 14, face = "bold")) +
19   theme(axis.title = element_text(size = 14, face = "bold")) +
20   scale_x_discrete(labels=c("Valores_teoricos"="Valores te ricos"))

```

Código A.29: Gráfico boxplot.

```

1 gghistostats(
2   data = df, # dataframe
3   x = variable1, # variable
4   type = "nonparametric", # nonparametric = Wilcoxon, parametric = t-test

```

## A. CÓDIGO R

---

```
5 test.value = 0.50, # default value
6 centrality.line.args = list(color = "black", linewidth = 1, linetype = "dashed
  "), #color de la linea
7 bin.args = list(color = "magenta", fill = "magenta", alpha = 0.4), #color de
  las columnas
8 ggtheme = ggplot2::theme_gray(), #tema
9 caption = "Visualizaci n"
10 )+
11   theme(axis.title = element_text(size = 14, face = "bold"))
```

Código A.30: Prueba Wilxocon para comparar con un valor determinado.

---

## Apéndice B

# Fuentes

---

### Brasil

	Universidades e institutos	Estado	Fuente
1	Escuela Nacional de Ciencias Estadísticas	Rio de Janeiro	[70]
2	Instituto de Matemática Pura y Aplicada	Rio de Janeiro	[71]
3	Pontificia Universidad Católica de Río de Janeiro	Rio de Janeiro	[72]
4	Universidad de Brasilia	Brasilia	[73]
5	Universidad de São Paulo	São Paulo	[74]
6	Universidad Estatal de Campinas	Campinas	[75]
7	Universidad Estatal de Feira de Santana	Bahía	[76]
8	Universidad Estatal de Londrina	Paraná	[77]
9	Universidad Estatal de Maringá	Maringá	[78]
10	Universidad Estatal de Río de Janeiro	Río de Janeiro	[79]
11	Universidad Federal de ABC	São Paulo	[80]
12	Universidad Federal de Amazonas	Amazonas	[81]
13	Universidad Federal de Bahía	Bahía	[82]
14	Universidad Federal de Ceará	Ceará	[83]
15	Universidad Federal de Juiz de Fora	Minas Gerais	[84]
16	Universidad Federal de Lavras	Minas Gerais	[85]
17	Universidad Federal de Minas Gerais	Minas Gerais	[86]
18	Universidad Federal de Paraíba	Paraíba	[87]
19	Universidad Federal de Paraná	Paraná	[88]
20	Universidad Federal de Pernambuco	Recife	[89]
21	Universidad Federal de Río Grande del Sur	Porto Alegre	[90]
22	Universidad Federal de Santa Catarina	Santa Catarina	[91]
23	Universidad Federal de São Carlos	São Paulo	[92]
24	Universidad Federal de São Joao del Rei	Minas Gerais	[93]
25	Universidad Federal de Viçosa	Minas Gerais	[94]
26	Universidad Federal del Rio de Janeiro	Rio de Janeiro	[95]
27	Universidad Federal Fluminense	Rio de Janeiro	[96]
28	Universidad Federal Rural de Pernambuco	Pernambuco	[97]

**Tabla B.1:** Universidades en Brasil y sus fuentes de información correspondientes.

## B. FUENTES

---

### Argentina

	Universidades e institutos	Estado	Fuente
1	Universidad de Buenos Aires	Buenos Aires	[98]
2	Universidad Nacional de Córdoba	Córdoba	[99]
3	Universidad Nacional de La Plata	Buenos Aires	[100]
4	Universidad Nacional de Rosario	Santa Fe	[101]
5	Universidad Nacional de Tres de Febrero	Buenos Aires	[102]
6	Universidad Nacional de Tucumán	Tucumán	[103]
7	Universidad Nacional del Comahue	Comahue	[104]

**Tabla B.2:** Universidades en Argentina y sus fuentes de información correspondientes.

### Chile

	Universidades e institutos	Estado	Fuente
1	Pontificia Universidad Católica de Chile	Santiago	[105]
2	Pontificia Universidad Católica de Valparaíso	Valparaíso	[106]
3	Pontificia Universidad Católica de Chile	Santiago	[107]
4	Universidad Adolfo Ibáñez	Santiago	[108]
5	Universidad Católica de Temuco	Araucanía	[109]
6	Universidad Católica del Norte	San Pedro de Atacama	[110]
7	Universidad de Antofagasta	Antofagasta	[111]
8	Universidad de Atacama	San Pedro de Atacama	[112]
9	Universidad de Chile	Santiago	[38]
10	Universidad de Concepción	Concepción	[113]
11	Universidad de La Frontera	Temuco	[114]
12	Universidad de Talca	Maule	[115]
13	Universidad de Tarapacá	Arica y Parinacota	[116]
14	Universidad del Bío-Bío	Biobío	[117]
15	Universidad Técnica Federico Santa María	Valparaíso	[118]
16	Universidad Tecnológica Metropolitana	Santiago	[119]

**Tabla B.3:** Universidades en Chile y sus fuentes de información correspondientes.

---

## Colombia

	Universidades e institutos	Estado	Fuente
1	Pontificia Universidad Javeriana	Bogotá	[120]
2	Universidad de los Andes	Bogotá	[121]
3	Universidad del Norte Barranquilla	Barranquilla	[122]
4	Universidad Nacional de Colombia	Bogotá	[123]
5	Desconocido	Desconocido	[124]

**Tabla B.4:** Universidades en Colombia y sus fuentes de información correspondientes.

## Cuba

	Universidades e institutos	Estado	Fuente
1	Universidad de Oriente	Santiago de Cuba	[125]

**Tabla B.5:** Universidades en Cuba y sus fuentes de información correspondientes.

## Perú

	Departamento/Institución	Estado	Fuente
1	Universidad Nacional Mayor de San Marcos	Lima	[126]
2	Universidad Nacional San Agustín	Arequipa	[127]

**Tabla B.6:** Universidades en Perú y sus fuentes de información correspondientes.

## Uruguay.

	Departamento/Institución	Estado	Fuente
1	Media Maren		[128]
2	Universidad de la República	Montevideo	[129]

**Tabla B.7:** Universidades en Uruguay y sus fuentes de información correspondientes.



---

Apéndice C

## Datos crudos del Capítulo 3

---

	Universidad	% Femenino	% Masculino
1	CIMAT	0.14	0.86
2	CINVESTAV	0.11	0.89
3	Colegio de Postgraduados	0.20	0.80
4	DEMAT/Universidad de Guanajuato	0.00	1.00
5	Escuela Nacional de Ciencias Estadísticas	0.44	0.56
6	FC/UNAM	0.40	0.60
7	IIMAS/UNAM	0.39	0.61
8	IMATE/UNAM	0.27	0.73
9	Instituto de Matemática Pura y Aplicada	0.00	1.00
10	ITAM	0.17	0.83
11	Media Maren	0.00	1.00
12	Pontificia Universidad Católica de Chile	0.26	0.74
13	Pontificia Universidad Católica de Valparaíso	0.30	0.70
14	Pontificia Universidad Javeriana	0.33	0.67
15	Tecnológico de Monterrey	0.00	1.00
16	UNISON	0.25	0.75
17	Universidad Autónoma Chapingo	0.00	1.00
18	Universidad Autónoma de Aguascalientes	0.40	0.60
19	Universidad de Antofagasta	0.00	1.00
20	Universidad de Brasilia	0.18	0.82
21	Universidad de Buenos Aires	0.60	0.40
22	Universidad de Concepción	0.30	0.70
23	Universidad de La Frontera	0.28	0.72
24	Universidad de la República	0.22	0.78
25	Universidad de los Andes	0.00	1.00
26	Universidad de São Paulo	0.37	0.63
27	Universidad de Talca	0.33	0.67
28	Universidad de Tarapacá	0.33	0.67

Tabla C.1: Porcentaje de hombres y mujeres dentro de cada departamento.

### C. DATOS CRUDOS DEL CAPÍTULO 3

---

	Institución	Clúster	%Femenino	%Masculino	Promedio de citas	Promedio de artículos
1	B	1	0.11	0.89	1206.79	65.53
2	C	1	0.20	0.80	1183.56	84.12
3	I	1	0.00	1.00	4090.25	149.38
4	O	1	0.18	0.82	706.69	38.81
5	P	1	0.37	0.63	1588.83	109.28
6	Q	1	0.26	0.74	1209.92	70.42
7	T	1	0.36	0.64	1575.37	49.37
8	U	1	0.16	0.84	417.93	53.73
9	Z	1	0.12	0.88	512.60	54.60
10	AA	1	0.19	0.81	2746.86	201.50
11	AB	1	0.37	0.63	1151.50	68.86
12	AD	1	0.24	0.76	785.50	64.95
13	AE	1	0.35	0.65	1863.95	92.90
14	AF	1	0.73	0.27	1671.67	84.89
15	AG	1	0.33	0.67	397.25	53.62
16	AJ	1	0.12	0.88	4535.93	411.53
17	AK	1	0.36	0.64	883.44	38.06
18	AM	1	0.27	0.73	537.00	76.00

**Tabla C.4:** Instituciones del clúster número 1 y sus porcentajes de mujeres y hombres.

	Institución	Clúster	%Femenino	%Masculino	Promedio de citas	Promedio de artículos
1	A	2	0.14	0.86	828.02	28.53
2	D	2	0.00	1.00	63.50	18.00
3	E	2	0.44	0.56	268.62	39.05
4	F	2	0.40	0.60	214.33	24.86
5	G	2	0.39	0.61	400.06	29.00
6	H	2	0.27	0.73	372.22	38.89
7	J	2	0.17	0.83	347.19	39.81
8	K	2	0.00	1.00	101.00	17.00
9	L	2	0.25	0.75	147.62	16.75
10	M	2	0.00	1.00	204.00	44.75
11	N	2	0.40	0.60	154.56	26.89
12	R	2	0.33	0.67	45.50	9.00
13	S	2	0.54	0.46	230.36	28.09
14	V	2	0.50	0.50	128.50	19.50
15	W	2	0.36	0.64	315.44	20.89
16	X	2	0.67	0.33	463.64	25.77
17	Y	2	0.15	0.85	476.65	39.18
18	AC	2	0.41	0.59	262.27	23.67
19	AH	2	0.31	0.69	269.60	42.10
20	AI	2	0.29	0.71	158.30	31.30
21	AL	2	0.43	0.57	99.94	11.06
22	AN	2	0.00	1.00	9.50	20.00
23	AO	2	0.46	0.54	228.92	38.31

**Tabla C.5:** Instituciones del clúster número 2 y sus porcentajes de mujeres y hombres.

## C. DATOS CRUDOS DEL CAPÍTULO 3

---

	Universidad	% Femenino	% Masculino
29	Universidad del Bío-Bío	0.09	0.91
30	Universidad del Norte Barranquilla	0.25	0.75
31	Universidad Estatal de Campinas	0.26	0.74
32	Universidad Estatal de Feira de Santana	0.33	0.67
33	Universidad Estatal de Londrina	0.54	0.46
34	Universidad Estatal de Maringá	0.36	0.64
35	Universidad Estatal de Río de Janeiro	0.16	0.84
36	Universidad Federal de ABC	0.50	0.50
37	Universidad Federal de Amazonas	0.36	0.64
38	Universidad Federal de Bahía	0.67	0.33
39	Universidad Federal de Ceará	0.15	0.85
40	Universidad Federal de Juiz de Fora	0.12	0.88
41	Universidad Federal de Lavras	0.19	0.81
42	Universidad Federal de Minas Gerais	0.37	0.63
43	Universidad Federal de Paraíba	0.41	0.59
44	Universidad Federal de Paraná	0.24	0.76
45	Universidad Federal de Pernambuco	0.35	0.65
46	Universidad Federal de Río Grande del Sur	0.73	0.27
47	Universidad Federal de Santa Catarina	0.33	0.67
48	Universidad Federal de São Carlos	0.31	0.69
49	Universidad Federal de São Joao del Rei	0.29	0.71
50	Universidad Federal de Viçosa	0.12	0.88
51	Universidad Federal del Rio de Janeiro	0.36	0.64
52	Universidad Federal Fluminense	0.43	0.57
53	Universidad Federal Rural de Pernambuco	0.27	0.73
54	Universidad Iberoamericana	0.00	1.00
55	Universidad Nacional de Colombia	0.38	0.62
56	Universidad Nacional de Córdoba	0.62	0.38
57	Universidad Nacional de La Plata	0.59	0.41
58	Universidad Nacional de Rosario	0.64	0.36
59	Universidad Nacional de Tres de Febrero	0.36	0.64
60	Universidad Nacional de Tucumán	0.56	0.44
61	Universidad Nacional del Comahue	0.58	0.42
62	Universidad Nacional Mayor de San Marcos	0.65	0.35
63	Universidad Nacional San Agustín	0.42	0.58
64	Universidad Técnica Federico Santa María	0.00	1.00
65	Universidad Tecnológica Metropolitana	0.00	1.00
66	Universidad Veracruzana	0.46	0.54

---

	Universidad	Promedio de citas	Promedio de artículos
1	A_M	356.86	27.14
2	B_M	1763.00	98.50
3	C_M	281.00	29.25
4	E_M	343.00	56.30
5	F_M	168.00	19.00
6	G_M	497.33	21.33
7	H_M	412.00	31.67
8	J_M	68.67	14.00
9	N_M	92.67	30.00
10	S_M	200.33	28.33
11	T_M	4275.00	48.83
12	U_M	420.67	43.67
13	V_M	12.00	3.00
14	W_M	9.00	1.00
15	X_M	709.83	28.08
16	Y_M	193.50	41.00
17	AL_M	9.00	4.86
18	Z_M	259.00	32.00
19	AA_M	578.50	127.50
20	AB_M	1269.50	82.90
21	AC_M	43.60	8.20
22	AE_M	300.29	27.86
23	AD_M	401.00	86.25
24	AF_M	2056.86	98.00
25	AK_M	598.00	40.17
26	AM_M	697.67	107.33
27	AG_M	290.50	45.00
28	AH_M	150.20	35.00
29	AI_M	165.00	33.00
30	AJ_M	814.50	111.50
31	O_M	1818.14	25.43
32	Q_M	614.29	53.00
33	L_M	18.00	4.50
34	P_M	1195.31	95.00
35	AO_M	61.33	19.17

**Tabla C.2:** Promedios de citas y artículos de cada universidad, únicamente con datos de mujeres.

## C. DATOS CRUDOS DEL CAPÍTULO 3

---

Universidad	Citas	Artículos
A_M	0.20	-0.09
B_M	1.25	1.10
C_M	0.04	-0.02
E_M	0.18	0.59
F_M	-0.29	-0.42
G_M	0.42	-0.31
H_M	0.30	0.05
J_M	-0.88	-0.70
N_M	-0.68	0.00
S_M	-0.18	-0.05
T_M	1.83	0.45
U_M	0.31	0.35
V_M	-2.02	-2.12
W_M	-2.21	-3.14
X_M	0.65	-0.06
Y_M	-0.20	0.29
AL_M	-2.21	-1.68
Z_M	-0.01	0.06

**Tabla C.3:** Muestra centralizada de promedios de citas y artículos de universidades con transformación logarítmica, únicamente con datos de mujeres.

---

## Apéndice D

# Resultados de la consola de RStudio

---

```
*** : The Hubert index is a graphical method of determining the number of clusters.  
      In the plot of Hubert index, we seek a significant knee that corresponds to a  
      significant increase of the value of the measure i.e the significant peak in Hubert  
      index second differences plot.  
  
*** : The D index is a graphical method of determining the number of clusters.  
      In the plot of D index, we seek a significant knee (the significant peak in Dindex  
      second differences plot) that corresponds to a significant increase of the value of  
      the measure.  
  
*****  
* Among all indices:  
* 6 proposed 2 as the best number of clusters  
* 6 proposed 3 as the best number of clusters  
* 3 proposed 4 as the best number of clusters  
* 2 proposed 6 as the best number of clusters  
* 1 proposed 7 as the best number of clusters  
* 6 proposed 9 as the best number of clusters  
* 3 proposed 10 as the best number of clusters  
  
***** Conclusion *****  
  
* According to the majority rule, the best number of clusters is 2  
  
*****
```

**Fig. D.1:** Resumen de 30 métodos.

## D. RESULTADOS DE LA CONSOLA DE RSTUDIO

---

```
*** : The Hubert index is a graphical method of determining the number of clusters.  
In the plot of Hubert index, we seek a significant knee that corresponds to a  
significant increase of the value of the measure i.e the significant peak in Hubert  
index second differences plot.  
  
*** : The D index is a graphical method of determining the number of clusters.  
In the plot of D index, we seek a significant knee (the significant peak in Dindex  
second differences plot) that corresponds to a significant increase of the value of  
the measure.
```

```
*****  
*****
```

- \* Among all indices:
- \* 10 proposed 2 as the best number of clusters
- \* 7 proposed 3 as the best number of clusters
- \* 1 proposed 4 as the best number of clusters
- \* 4 proposed 5 as the best number of clusters
- \* 5 proposed 9 as the best number of clusters
- \* 1 proposed 10 as the best number of clusters

```
***** Conclusion *****
```

- \* According to the majority rule, the best number of clusters is 2

```
*****  
*****
```

**Fig. D.2:** Resumen de 30 métodos, únicamente con datos de mujeres.

```
*** : The Hubert index is a graphical method of determining the number of clusters.  
In the plot of Hubert index, we seek a significant knee that corresponds to a  
significant increase of the value of the measure i.e the significant peak in Hubert  
index second differences plot.  
  
*** : The D index is a graphical method of determining the number of clusters.  
In the plot of D index, we seek a significant knee (the significant peak in Dindex  
second differences plot) that corresponds to a significant increase of the value of  
the measure.
```

```
*****  
*****
```

- \* Among all indices:
- \* 10 proposed 2 as the best number of clusters
- \* 8 proposed 3 as the best number of clusters
- \* 2 proposed 4 as the best number of clusters
- \* 2 proposed 5 as the best number of clusters
- \* 4 proposed 7 as the best number of clusters
- \* 1 proposed 8 as the best number of clusters
- \* 1 proposed 9 as the best number of clusters

```
***** Conclusion *****
```

- \* According to the majority rule, the best number of clusters is 2

```
*****  
*****
```

**Fig. D.3:** Resumen de 30 métodos, con datos de instituciones divididas por género.

## Siglas institucionales

---

**ABC** Academia Brasileña de Ciencias, Brasil.

**AU/CL** Universidad de Antofagasta, Chile.

**CIMAT** Centro de Investigación en Matemáticas, México.

**CINVESTAV** Centro de Investigación y de Estudios Avanzados, México.

**Colpos** Colegio de Postgraduados, México.

**DEMAT** Departamento de Matemáticas, Universidad de Guanajuato, México.

**ENCE** Escuela Nacional de Ciencias Estadísticas, Brasil.

**FC/UNAM** Facultad de Ciencias, UNAM, México.

**IBERO** Universidad Iberoamericana, México.

**IC, UK** Imperial College, Reino Unido.

**IIMAS/UNAM** Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, México.

**IMATE/UNAM** Instituto de Matemáticas, UNAM, México.

**IMPA** Instituto de Matemática Pura y Aplicada, Brasil.

**IPN** Instituto Politécnico Nacional, México.

**ITAM** Instituto Tecnológico Autónomo de México, México.

**ITESM** Tecnológico de Monterrey, México.

**MIT** Instituto Tecnológico de Massachusetts, Estados Unidos.

**PUC/CL** Pontificia Universidad Católica de Chile, Chile.

**PUCV** Pontificia Universidad Católica de Valparaíso, Chile.

**PUJ** Pontificia Universidad Javeriana, Colombia.

**PyE** Probabilidad y Estadística.

**QMUL, UK** Universidad Queen Mary de Londres, Reino Unido.

**QU, CAN** Universidad Queen's, Canadá.

**SNI** Sistema Nacional de Investigadores.

**TU Berlin** Universidad Técnica de Berlín, Alemania.

**UAA** Universidad Autónoma de Aguascalientes, México.

**UACh** Universidad Autónoma Chapingo, México.

**UB, ESP** Universidad de Barcelona, España.

**UBA** Universidad de Buenos Aires, Argentina.

**UBath, UK** Universidad de Bath, Reino Unido.

**UBB** Universidad del Bío-Bío, Chile.

**UBC, CAN** Universidad de British Columbia, Canadá.

**UChile** Universidad de Chile, Chile.

**UdeC** Universidad de Concepción, Chile.

**UdelaR** Universidad de la República, Uruguay.

**UE,UK** Universidad de Essex, Reino Unido.

**UEFS** Universidad Estatal de Feira de Santana, Brasil.

**UEL** Universidad Estatal de Londrina, Brasil.

**UEM** Universidad Estatal de Maringá, Brasil.

**UERJ** Universidad Estatal de Río de Janeiro, Brasil.

**UEx, ESP** Universidad de Extremadura, España.

**UFABC** Universidad Federal de ABC, Brasil.

**UFAM** Universidad Federal de Amazonas, Brasil.

**UFBA** Universidad Federal de Bahía, Brasil.

**UFC** Universidad Federal de Ceará, Brasil.

**UFF** Universidad Federal Fluminense, Brasil.

**UFJF** Universidad Federal de Juiz de Fora, Brasil.

**UFLA** Universidad Federal de Lavras, Brasil.

**UFMG** Universidad Federal de Minas Gerais, Brasil.

**UFPB** Universidad Federal de Paraíba, Brasil.

---

**UFPE** Universidad Federal de Pernambuco, Brasil.

**UFPR** Universidad Federal de Paraná, Brasil.

**UFRGS** Universidad Federal de Río Grande del Sur, Brasil.

**UFRJ** Universidad Federal del Rio de Janeiro, Brasil.

**UFRO** Universidad de La Frontera, Chile.

**UFRPE** Universidad Federal Rural de Pernambuco, Brasil.

**UFSC** Universidad Federal de Santa Catarina, Brasil.

**UFSCar** Universidad Federal de São Carlos, Brasil.

**UFSJ** Universidad Federal de São Joao del Rei, Brasil.

**UFV** Universidad Federal de Viçosa, Brasil.

**UKC, UK** Universidad de Kent, Reino Unido.

**UM, CAN** Universidad McGill, Canadá.

**UMCE** Universidad Metropolitana de Ciencias de la Educación, Chile.

**UMD, EE. UU.** Universidad de Maryland, Estados Unidos.

**UNAL** Universidad Nacional de Colombia, Colombia.

**UNAM** Universidad Nacional Autónoma de México, México.

**UnB** Universidad de Brasilia, Brasil.

**UNC** Universidad Nacional de Córdoba, Argentina.

**UNCo** Universidad Nacional del Comahue, Argentina.

**UNESCO** Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.

**Uniandes** Universidad de los Andes, Colombia.

**UNICAMP** Universidad Estatal de Campinas, Brasil.

**Uninorte** Universidad del Norte Barranquilla, Chile.

**UNISON** Universidad de Sonora, México.

**UNLP** Universidad Nacional de La Plata, Argentina.

**UNMSM** Universidad Nacional Mayor de San Marcos, Perú.

**UNR** Universidad Nacional de Rosario, Argentina.

**UNSA** Universidad Nacional San Agustín, Perú.

**UNT** Universidad Nacional de Tucumán, Argentina.

**UNTREF** Universidad Nacional de Tres de Febrero, Argentina.

**UP, FRA** Universidad de Paris, Francia.

**UPC, ESP** Universidad Politécnica de Catalunya, España.

**UPMC, FRA** Universidad de París VI, Francia.

**US, ESP** Universidad de Sevilla, España.

**USP** Universidad de São Paulo, Brasil.

**USW, UK** Universidad de Swansea, Reino Unido.

**UTA** Universidad de Tarapacá, Chile.

**UTALCA** Universidad de Talca, Chile.

**UTEM** Universidad Tecnológica Metropolitana, Chile.

**UTFSM** Universidad Técnica Federico Santa María, Chile.

**UV** Universidad Veracruzana, México.

**UW, EE. UU.** Universidad de Washington, Estados Unidos.

**UW-Madison, EE.UU.** Universidad de Wisconsin-Madison, Estados Unidos.

**Warwick, UK** Universidad de Warwick, Reino Unido.

# Glosario

---

**Artículo:** En este trabajo se sigue el estándar establecido por Google Scholar y un artículo académico se refiere a trabajos como: artículos de revistas, documentos de conferencias, informes técnicos o sus borradores, dissertaciones, pre-impresiones, post-impresiones o resúmenes [130].

**Citas académicas:** Se toma como una cita académica cuando un investigador, en su trabajo, hace referencia a otro artículo publicado por sí mismo u otro investigador.

**Correo electrónico verificado:** Dirección de correo electrónico con un dominio perteneciente a una universidad, departamento o instituto donde se realice investigación.

**DataFrame:** Es una estructura de datos bidimensional compuesto por filas y columnas, similar a una hoja de cálculo, que permite almacenar datos para su manipulación en R.

**Dato tipo factor:** Un dato factor se refiere a un dato categórico con un conjunto limitado y conocido de posibles valores.

**Departamento:** Unidad académica constituida por los docentes que realizan actividades de docencia, investigación, responsabilidad social, de gestión académico-administrativa, asesoría y tutoría de estudiantes. [131].

**Instituto:** Institución que se dedica de la investigación relacionada con matemáticas, probabilidad o estadística.

**Investigador:** Académico o académica que trabaja en alguna área relacionada con probabilidad y estadística, cuya información puede consultarse en alguna página institucional o en alguna plataforma de investigación disponible públicamente en internet.

**Niveles de un dato tipo factor:** Los niveles de un dato tipo factor son los únicos posibles valores que un dato puede tomar. Estos son útiles para especificar el orden de datos categóricos ordinales.

**Perfil:** Cuenta personal, creada por el investigador o generada automáticamente, en Google Scholar, ResearchGate o alguna otra plataforma utilizada en nuestro estudio, donde se muestra la información de la persona investigadora.

**Publicación:** Usualmente se utiliza como sinónimo de artículo. Sin embargo, una publicación indica que dicho artículo ha pasado por el proceso de publicación (en particular, de refleo por pares).

**Teórico social:** Un concepto teórico social es una herramienta que se utiliza para el análisis y la comprensión de fenómenos sociales. Con la intención de hacer predicciones y generalizaciones sobre el comportamiento humano y la interacción social.

## Bibliografía

---

- [1] M. C. Tapia1. La Participación de las Mujeres Investigadoras en México. [https://www.scielo.org.mx/scielo.php?script=sci\\_arttext&pid=S2448-76782015000200004#:~:text=Actualmente%20las%20mujeres%20forman%20parte,menores%20que%20los%20hombres%20en](https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2448-76782015000200004#:~:text=Actualmente%20las%20mujeres%20forman%20parte,menores%20que%20los%20hombres%20en), 2015 (citado en las págs. 1, 76).
- [2] S. E. Data y T. 2014. Has employment of women and minorities in SE jobs increased?, howpublished = "<https://nsf.gov/nsb/sei/edTool/data/workforce-07.html>", year = 2014, (citado en las págs. 1, 78).
- [3] AAUW. The STEM Gap: Women and Girls in Science, Technology, Engineering and Mathematics, howpublished = "<https://www.aauw.org/resources/research/the-stem-gap/>", year = 2023, (citado en las págs. 1, 76).
- [4] P. C. C. Catherine Hill. Why So Few? <https://www.aauw.org/app/uploads/2020/03/why-so-few-research.pdf>, 2010 (citado en la pág. 1).
- [5] UNESCO. STEM and Gender Advancement (SAGA). <https://en.unesco.org/saga>, 2010 (citado en la pág. 1).
- [6] B. Bozeman y M. Gaughan. How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research policy*, **40**(10):1393-1402, 2011 (citado en la pág. 2).
- [7] S. Yang. Networks: An Introduction by MEJ Newman: Oxford, UK: Oxford University Press. 720 pp., \$85.00. 2013 (citado en las págs. 3, 23).
- [8] R. CLARA. ¿Qué son y para qué sirven? <https://www.redclara.net/index.php/es/red/redes-de-investigacion-y-educacion/que-son-y-para-que-sirven#:~:text=Las%20redes%20de%20investigaci%C3%B3n%20y,serie%20de%20interconexiones%20de%20redes.>, 2022. [Accedido en línea en noviembre-2022] (citado en la pág. 4).
- [9] Wikipedia. Número de Erdős. [https://es.wikipedia.org/wiki/N%C3%BAmero\\_de\\_Erd%C5%91s](https://es.wikipedia.org/wiki/N%C3%BAmero_de_Erd%C5%91s), 2023. [Accedido en línea en enero-2023] (citado en la pág. 6).
- [10] Mathscinet. MR: Erdős, Paul. <https://mathscinet.ams.org/mathscinet/MRAuthorID/189017>, 2023. [Accedido en línea en enero-2023] (citado en la pág. 6).
- [11] ABC de la perspectiva de Género:Comisión Nacional de los Derechos Humanos Periférico Sur 3469, Col. San Jerónimo Lídice, C.P. 10200, Ciudad de México., author=Ana Luisa Nerio Monroy, year=2019, publisher=CNDH (citado en la pág. 7).
- [12] INMUJERES. “Brechas de género”. [https://crpd.cepal.org/3/sites/crpd3/files/presentations/panel2\\_marcelaeternod.pdf](https://crpd.cepal.org/3/sites/crpd3/files/presentations/panel2_marcelaeternod.pdf), 2018 (citado en la pág. 7).

## BIBLIOGRAFÍA

---

- [13] C. UNAM. Roles y estereotipos de género. <https://www.coursera.org/learn/genero-igualdad/lecture/9c2gV/roles - y - estereotipos - de - genero>, 2022 (citado en la pág. 7).
- [14] C. UNAM. Brecha de género. Feminización y masculinización de oficios y profesiones. <https://www.coursera.org/learn/genero-igualdad/lecture/fuR0y/brecha-de-genero-feminizacion-y-masculinizacion-de-oficios-y-profesiones>, 2022 (citado en la pág. 7).
- [15] D. F. Boeff y C. Cánovas. ESTRATÉGIAS DE MUJERES PROFESIONISTAS QUE TRABAJAN EN AMBIENTES MASCULINIZADOS. *Revista Ensino de Ciências e Humanidades-Cidadania, Diversidade e Bem Estar-RECH*, 4(1, jan-jun):26-44, 2020 (citado en la pág. 7).
- [16] UNESCO. Mujeres en la educación superior: ¿la ventaja femenina ha puesto fin a las desigualdades de género? <https://www.iesalc.unesco.org/wp-content/uploads/2021/03/Informe-Mujeres-ES-080321.pdf>, 2022 (citado en las págs. 7, 75).
- [17] V. López-Bassols, M. Grazzi, C. Guillard y M. Salazar. Las brechas de género en ciencia, tecnología e innovación en América Latina y el Caribe. *Resultados de una recolección piloto y propuesta metodológica para la medición*, 2018 (citado en la pág. 7).
- [18] J. A. de la Peña. Políticas públicas en ciencia. <http://universo.math.org.mx/2014-1/Politicas-publicas-en-ciencia/politicas-publicas.html>, 2014. [Accedido en línea en enero-2023] (citado en las págs. 7, 8).
- [19] A. M. Ramírez. *¿Legitimidad o reconocimiento? Las investigadoras del SNI. Retos y propuestas*. Ediciones La Biblioteca, SA de CV, 2015 (citado en las págs. 8, 79).
- [20] I. S. N. de Avaliação Científica. Recomendações às Agências de Fomento. [http://www.sbfisica.org.br/v1/arquivos\\_diversos/avaliacao-2010/Simposio\\_Avaliacao\\_Cientifica.pdf](http://www.sbfisica.org.br/v1/arquivos_diversos/avaliacao-2010/Simposio_Avaliacao_Cientifica.pdf), 2010. [Accedido en línea en enero-2023] (citado en la pág. 8).
- [21] D. Hirschfeld. Científicos de Brasil objetan criterios de evaluación. <https://www.scidev.net/america-latina/news/cient-ficos-de-brasil-objetan-criterios-de-evaluacion/>, 2010. [Accedido en línea en enero-2023] (citado en la pág. 8).
- [22] J. A. de la Peña. El Conacyt celebra el Día Internacional de la Mujer y la Niña en la Ciencia. <https://conacyt.mx/el-conacyt-celebra-el-dia-internacional-de-la-mujer-y-la-nina-en-la-ciencia/#:~:text=En%20este%20sentido%2C%20record%C3%B3que,y%20s%C3%B3lo%2038.2%20%25%20de%20mujeres.>, 2022. [Accedido en línea en enero-2023] (citado en la pág. 8).
- [23] Scopus. The Scopus h-index, what's it all about? Part I. <https://blog.scopus.com/posts/the-scopus-h-index-what-s-it-all-about-part-i>, 2022. [Accedido en línea en noviembre-2022] (citado en la pág. 8).
- [24] BERNADBECKER. Tools for Authors: What is the h index? <https://beckerguides.wustl.edu/authors/hindex>, 2022. [Accedido en línea en noviembre-2022] (citado en la pág. 8).
- [25] G. de la BUH. Evaluación de la Investigación: Índice H. <https://guiasbuh.uhu.es/c.php?g=655120&p=4605523#:~:text=%C3%8Dndice%20H%3A%20definici%C3%B3n&text=Se%20calcula%20ordenando%20de%20mayor,ver%20en%20la%20siguiente%20figura.>, 2022. [Accedido en línea en noviembre-2022] (citado en la pág. 10).
- [26] CIMAT. Investigadores Probabilidad y Estadística. [https://www.cimat.mx/investigadores\\_probabilidad\\_y\\_estadistica/](https://www.cimat.mx/investigadores_probabilidad_y_estadistica/), 2022. [Accedido en línea en julio-2022] (citado en la pág. 13).

## BIBLIOGRAFÍA

---

- [27] CIMAT. CIMAT. <https://www.cimat.mx/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 13).
- [28] CINVESTAV. CINVESTAV - Investigadores. <https://www.math.cinvestav.mx/investigadores>, 2022. [Accedido en línea en junio-2022] (citado en la pág. 14).
- [29] COLPOS. Maestría en Estadística. [https://www.colpos.mx/posgrado/psei/estadistica/i.n.a\\_m.php](https://www.colpos.mx/posgrado/psei/estadistica/i.n.a_m.php), 2022. [Accedido en línea en junio-2022] (citado en la pág. 14).
- [30] COLPOS. Doctorado en Estadística. [https://www.colpos.mx/posgrado/psei/estadistica/i.n.a\\_d.php](https://www.colpos.mx/posgrado/psei/estadistica/i.n.a_d.php), 2022. [Accedido en línea en junio-2022] (citado en la pág. 14).
- [31] IIMAS. Departamento de Probabilidad y Estadística. <http://www.dpye.iimas.unam.mx/>, 2022. [Accedido en línea en junio-2022] (citado en la pág. 15).
- [32] IMATE. Instituto de Matemáticas. <https://www.matem.unam.mx/>, 2022. [Accedido en línea en junio-2022] (citado en la pág. 15).
- [33] ITAM. Departamento Académico de Estadística. <http://estadistica.itam.mx/>, 2022. [Accedido en línea en junio-2022] (citado en la pág. 15).
- [34] UNISON. PROBABILIDAD Y ESTADÍSTICA. <https://www.mat.uson.mx/web/index.php/academias/probabilidad-y-estadistica/>, 2022. [Accedido en línea en junio-2022] (citado en la pág. 15).
- [35] U. A. de Aguascalientes. ESTADÍSTICA. <https://www.uaa.mx/portal/nuestra-universidad/centros-academicos-2/centro-de-ciencias-basicas/estadistica/siguiendo>, 2022. [Accedido en línea en junio-2022] (citado en la pág. 16).
- [36] Ibero. Investigadores. <https://investigacion.ibero.mx/investigadores>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 16).
- [37] U. Veracruzana. Portal de Académicos. <https://www.uv.mx/academicos/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 16).
- [38] U. de Chile. <https://www.uchile.cl/portafolio-academico/perfilAcademico.jsf?username=nhenriq>, 2022. [Accedido en línea en septiembre-2022] (citado en las págs. 16, 92).
- [39] Wikipedia. Consejo Nacional de Desarrollo Científico y Tecnológico. <https://www.gov.br/cnpq/pt-br>, 2023 (citado en la pág. 17).
- [40] Wikipedia. Nombres de nacimiento y de matrimonio. [https://es.wikipedia.org/wiki/Nombres\\_de\\_nacimiento\\_y\\_de\\_matrimonio#Pa%C3%ADses\\_angloparlantes](https://es.wikipedia.org/wiki/Nombres_de_nacimiento_y_de_matrimonio#Pa%C3%ADses_angloparlantes), 2023. [Accedido en línea en enero-2023] (citado en la pág. 22).
- [41] A. C. Davison y D. V. Hinkley. *Bootstrap methods and their application*, número 1. Cambridge university press, 1997 (citado en la pág. 27).
- [42] B. Efron. *Bootstrap methods: another look at the jackknife*. Springer, 1992 (citado en la pág. 27).
- [43] B. Efron y R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994 (citado en la pág. 27).
- [44] R. Ledesma. Introducción al Bootstrap. Desarrollo de un ejemplo acompañado de software de aplicación. *Tutorials in quantitative methods for psychology*, 4(2):51-60, 2008 (citado en la pág. 27).
- [45] S. Castillo-Gutiérrez y E. D. L. Aguilera. QQ Plot Normal. Los puntos de posición gráfica. *Iniciación a la Investigación*, (2):8, 2007 (citado en la pág. 27).
- [46] C. E. F. Tapia y K. L. F. Cevallos. PRUEBAS PARA COMPROBAR LA NORMALIDAD DE DATOS EN PROCESOS PRODUCTIVOS:: ANDERSON-DARLING, RYAN-JOINER,

- SHAPIRO-WILK Y KOLMOGOROV-SMIRNOV. *Societas*, **23**(2):83-106, 2021 (citado en la pág. 28).
- [47] R. C. Team. *shapiro.test: Shapiro-Wilk Test for Normality*. R package version 3.6.2. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/shapiro.test> (citado en la pág. 28).
- [48] F. Wilcoxon. Individual comparisons by ranking methods. En *Breakthroughs in Statistics: Methodology and Distribution*, páginas 196-202. Springer, 1992 (citado en la pág. 28).
- [49] R bloggers. One-sample Wilcoxon test in R. <https://www.r-bloggers.com/2022/07/one-sample-wilcoxon-test-in-r/>, 2022. Accedido en línea en octubre-2022 (citado en la pág. 28).
- [50] I. Patil. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, **6**(61):3167, 2021. DOI: [10.21105/joss.03167](https://doi.org/10.21105/joss.03167). URL: <https://doi.org/10.21105/joss.03167> (citado en la pág. 28).
- [51] H. B. Mann y D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*:50-60, 1947 (citado en la pág. 28).
- [52] A. R. Ríos y A. M. P. Peña. Estadística inferencial. Elección de una prueba estadística no paramétrica en investigación científica. *Horizonte de la Ciencia*, **10**(19):191-208, 2020 (citado en la pág. 28).
- [53] I. Patil. Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, **6**(61):3167, 2021. DOI: [10.21105/joss.03167](https://doi.org/10.21105/joss.03167). URL: <https://doi.org/10.21105/joss.03167> (citado en la pág. 28).
- [54] R. Turcios. Prueba de Wilcoxon-Mann-Whitney: mitos y realidades. *Rev Mex Endocrinol Metab Nutr*, **2**:18-21, 2015 (citado en la pág. 28).
- [55] L. Rincón. Introducción a la probabilidad, 2014 (citado en la pág. 29).
- [56] J. MacQueen. others. 1967. Some methods for classification and analysis of multivariate observations. En *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volumen 1 (citado en la pág. 29).
- [57] C. G. Cambronero e I. G. Moreno. Algoritmos de aprendizaje: knn & kmeans. *Inteligencia en Redes de Comunicación, Universidad Carlos III de Madrid*, **23**, 2006 (citado en la pág. 29).
- [58] R. C. Team. R documentation: kmeans. <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans>, 2019. Accedido en línea en octubre-2022 (citado en la pág. 29).
- [59] C. Spearman. The proof and measurement of association between two things. 1961 (citado en la pág. 29).
- [60] L. F. Restrepo y J. González. De Pearson a Spearman. *Revista Colombiana de Ciencias Pecuarias*, **20**(2):183-192, 2007 (citado en la pág. 29).
- [61] F. Mendivelso. Prueba no paramétrica de correlación de Spearman. *Revista Médica Sanitas*, **24**(1), 2021 (citado en la pág. 29).
- [62] W. Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.2.2. 2022. URL: <https://cran.r-project.org/package=psych> (citado en la pág. 30).
- [63] DataCamp. The General Social Survey. Inference for Categorical Data in R, 2022. URL: <https://app.datacamp.com/learn/courses/inference-for-categorical-data-in-r>. [Accedido en línea en octubre-2022 (citado en la pág. 31)].

## BIBLIOGRAFÍA

---

- [64] DataCamp. Intervals for differences. Inference for Categorical Data in R, 2022. URL: <https://app.datacamp.com/learn/courses/inference-for-categorical-data-in-r>. [Accedido en línea en octubre-2022 (citado en las págs. 33, 50)].
- [65] Pablo Vallejo Medina. [Rstudio] Cómo graficar T Student dos muestras [BONITO] [Chupito de R], jun. de 2020. URL: <https://www.youtube.com/watch?v=6Ds9L4NpRVQ&t=202s> (citado en las págs. 41, 71, 72).
- [66] A. Soetewey. One-sample Wilcoxon test in R. <https://statsandr.com/blog/one-sample-wilcoxon-test-in-r/>, 2022. Accedido en línea en octubre-2022 (citado en la pág. 43).
- [67] Pablo Vallejo Medina. [K means] Análisis de Clúster en R y Rstudio. [Chupitos de R], mayo de 2020. URL: <https://www.youtube.com/watch?v=7AFuL-1Q8eg&t=261s> (citado en las págs. 53, 61).
- [68] Pablo Vallejo Medina. Cómo hacer correlaciones en Rstudio y R. [Chupito de R], abr. de 2020. URL: <https://www.youtube.com/watch?v=uEcvj7C35ho&t=474s> (citado en la pág. 66).
- [69] S. Worthington. ipak.R. <https://gist.github.com/stevenworthington/3178163>, 2012. Accedido en línea en octubre-2022 (citado en la pág. 81).
- [70] E. N. de Ciencias Estadísticas. Cuerpo Docente. <https://ence.ibge.gov.br/index.php/portal-graduacao/portal-graduacao-corpo-docente>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [71] I. de Matemática Pura y Aplicada. IMPA - Instituto de Matemática Pura y Aplicada. <https://impa.br/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [72] P. U. C. de Río de Janeiro. Departamento de Matemáticas - PUC- Río. <http://www.mat.puc-rio.br/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [73] U. de Brasilia. Departamento de Estadística. <http://www.est.unb.br/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [74] U. de São Paulo. Departamento de Estadística. <https://www.ime.usp.br/mae/docentes/> <https://www.ime.usp.br/mae/docentes/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [75] U. E. de Campinas. Instituto de Matemáticas, Estadística y computación. <https://www.ime.unicamp.br/departamentos/estatistica/corpo-docente#>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [76] U. E. de Feira de Santana. <https://www.escavador.com/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [77] U. E. de Londrina. Departamento de Estadística e Informática. <http://www.deinfo.ufrpe.br/br/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [78] U. E. de Maringá. Cuerpo Docente - Departamento de Estadística. <http://www.des.uem.br/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [79] U. E. de Río de Janeiro. IME - Instituto de Matemáticas y Estadística. <https://www.ime.uerj.br/departamentos-ime/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [80] U. F. de ABC. UFABC Docentes. <https://www.ufabc.edu.br/ensino/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [81] U. F. de Amazonas. Departamento de Estadística. [https://www.icede.ufam.edu.br/index.php?option=com\\_content&view=article&id=89&catid=70](https://www.icede.ufam.edu.br/index.php?option=com_content&view=article&id=89&catid=70), 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).

## BIBLIOGRAFÍA

---

- [82] U. F. de Bahía. Cuerpo Docente | Departamento de Estadística. <https://est.ufba.br/pt-br/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [83] U. F. de Ceará. Profesores Dep. de Estadística y Matemática Aplicada. <https://dema.ufc.br/pt/professores/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [84] U. F. de Juiz de Fora. Cuerpo Docente - Departamento de Estadística/ICE. <https://www2.ufjf.br/estatistica/cursos/docencia/docentes/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [85] U. F. de Lavras. Departamento de Estadística de la Universidad Federal de Lavras - DES UFLA. <https://des.ufla.br/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [86] U. F. de Minas Gerais. Departamento de Estadística - ICEx - UFMG. <http://www.est.ufmg.br/portal/professores>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [87] U. F. de Paraíba. Estadística UFPR. <http://www.est.ufpr.br/docentes.html>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [88] U. F. de Paraná. Cuerpo Docente del Departamento de Estadística. <http://www.est.ufpr.br/docentes.html>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [89] U. F. de Pernambuco. Departamento de Estadística - UFPE. <https://www.ufpe.br/dep-estatistica>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [90] U. F. de Río Grande del Sur. NAE, Núcleo de Asesoría Estadística. <http://mat.ufrgs.br/~nae/equipe.htm>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [91] U. F. de Santa Catarina. UFSC - INE - Departamento de Informática y Estadística. <https://ine.ufsc.br/docentes/>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [92] U. F. de São Carlos. Docentes - Departamento de Estadística. <https://www.des.ufscar.br/departamento/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [93] U. F. de São Joao del Rei. Servidores de DEFIM. <https://ufsjiang.edu.br/defim/servidores.php>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [94] U. F. de Viçosa. Orientadores - Estadística Aplicada y Biometría. [https://ppestbio.ufv.br/docentes\\_ppestbio/](https://ppestbio.ufv.br/docentes_ppestbio/), 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [95] U. F. del Rio de Janeiro. Departamento de Métodos Estadísticos. <http://www.dme.im.ufrj.br/listarCorpoDocente.php>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [96] U. F. Fluminense. Departamento de Estadística de UFF. <http://est.uff.br>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [97] U. F. R. de Pernambuco. DOCENTES | Departamento de Estadística e Informática. <http://www.deinfo.ufrpe.br/br/docentes>, 2022. [Accedido en línea en julio-2022] (citado en la pág. 91).
- [98] U. de Buenos Aires. UBA - exactas Departamento de Matemática. <https://web.dm.uba.ar/index.php/institucional/integrantes/profesores/teacher/armendariz-ines>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [99] U. N. de Córdoba. Probabilidad y Estadística. <https://www.famaf.unc.edu.ar/investigaci%C3%B3n/%C3%A1reas-de-investigaci%C3%B3n/matem%C3%A1tica-ofi/probabilidad-y-estad%C3%ADstica/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).

## BIBLIOGRAFÍA

---

- [100] U. N. de La Plata. Matemática Aplicada Estudio de Series Temporales y Análisis Wa velet. <http://cmalp.mate.unlp.edu.ar/aplicada.html>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [101] U. N. de Rosario. FCEyE Facultad de Ciencias Económicas y Estadística UNR. <https://www.fcecon.unr.edu/web-nueva/investigacion/estadisticas-iitae/descripcion>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [102] U. N. de Tres de Febrero. CINEA: Centro de Investigaciones en Estadística Aplicada. <https://untref.edu.ar/instituto/cinea-centro-de-investigaciones-en-estadistica-aplicada>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [103] U. N. de Tucumán. Maestría en Estadística Aplicada. <https://face.unt.edu.ar/web-posgrados/posgrados/maestria-en-estadistica-aplicada/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [104] U. N. del Comahue. Maestría en Estadística Aplicada. [https://faeaweb.uncoma.edu.ar/oferta\\_academica/maestria-en-estadistica-aplicada/](https://faeaweb.uncoma.edu.ar/oferta_academica/maestria-en-estadistica-aplicada/), 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [105] P. U. C. de Chile. <https://www.mat.uc.cl/personas>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [106] P. U. C. de Valparaíso. Instituto de Estadística PUCV. <http://www.estadistica.cl/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [107] P. U. C. de Chile. Departamento de Matemática. <http://www.umce.cl/index.php/fac-ciencias-departamentos/fac-ciencias-depto-matematica/42-facultades/facultad-de-ciencias-basicas/d-matematica#armijo>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [108] U. A. Ibáñez. <http://postgrados.uantof.cl/academicos-magister-ciencias/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [109] G. Scholar. [https://scholar.google.com/citations?user=BQ7\\_uGoAAAAJ&hl=es](https://scholar.google.com/citations?user=BQ7_uGoAAAAJ&hl=es), 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [110] U. C. del Norte. <http://postgrados.uantof.cl/academicos-magister-ciencias/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [111] U. de Antofagasta. MAGÍSTER EN CIENCIAS MENCIÓN ESTADÍSTICA INDUSTRIAL, MENCIÓN MATEMÁTICA APLICADA. <http://postgrados.uantof.cl/academicos-magister-ciencias/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [112] U. de Atacama. Red de universidades del estado de Chile. <https://www.uestatales.cl/cue/?q=node/5745>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [113] U. de Concepción. Departamento de Estadística UdeC. <http://www.ing-estadistica.udec.cl/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [114] U. de La Frontera. Departamento de Matemáticas y Estadística DME. <https://dme.ufro.cl/el-departamento/academicos.html>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [115] U. de Talca. INSTMAT. <http://inst-mat. utalca.cl/html/index.php/acerca-del-imafi/17-acerca-de/profesores>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [116] U. de Tarapacá. Departamentos Facultad de Ciencias. <https://www.uta.cl/index.php/departamentos-facultad-de-ciencias/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).

## BIBLIOGRAFÍA

---

- [117] U. del Bío-Bío. Docentes - Departamento de Estadística. [http://estadistica.ubiobio.cl/?page\\_id=2](http://estadistica.ubiobio.cl/?page_id=2), 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [118] U. T. F. S. María. DEPARTAMENTO DE MATEMÁTICA. <http://matematica.usm.cl/personas/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [119] U. T. Metropolitana. UTEM. <https://www.utem.cl/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 92).
- [120] P. U. Javeriana. Estadística y Matemática Aplicada (EMAP). <https://www.javerianacali.edu.co/grupos-investigacion/estadistica-y-matematica-aplicada-emap>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [121] U. de los Andes. Departamento de Matemáticas. <https://matematicas.uniandes.edu.co/es>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [122] U. del Norte Barranquilla. Maestría en Estadística Aplicada. <https://www.uninorte.edu.co/web/maestria-en-estadistica-aplicada/nuestros-docentes>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [123] U. N. de Colombia. Departamento de Estadística. [bit.ly/3LFrbwU](https://bit.ly/3LFrbwU), 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [124] NA. [https://scienti.mincierias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod\\_rh=0000024660](https://scienti.mincierias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000024660), 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [125] U. de Oriente Cuba. Universidad de Oriente Cuba. [https://www.ecured.cu/Universidad\\_de\\_Oriente](https://www.ecured.cu/Universidad_de_Oriente), 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [126] U. N. M. de San Marcos. Find Profiles - Universidad Nacional Mayor de San Marcos. <https://siis.unmsm.edu.pe/en/persons/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [127] U. N. S. Agustín. Departamento Académico de Estadística. <https://fcnf.unsa.edu.pe/departamento-academico-de-estadistica-2/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [128] M. Maren. QUIÉNES SOMOS - MEDIA MAREN. <https://www.maren.cure.edu.uy/sobre-el-maren/>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [129] U. de la República. instituto de Matemática y Estadística Rafael Laguardia. <https://www.fing.edu.uy/es/imerl/docentes>, 2022. [Accedido en línea en septiembre-2022] (citado en la pág. 93).
- [130] G. Scholar. Inclusion Guidelines for Webmasters. <https://scholar.google.com/intl/es/scholar/inclusion.html#content>, 2022. [Accedido en línea en noviembre-2022] (citado en la pág. 107).
- [131] V. A. de Pregrado. Departamento Académico. [https://viceacademico.unmsm.edu.pe/?page\\_id=5589](https://viceacademico.unmsm.edu.pe/?page_id=5589), 2022. [Accedido en línea en noviembre-2022] (citado en la pág. 107).